Selected topics in Advanced Machine Learning

Lecture 03 – Anomalies

January 24, 2022

+ Objectives

- Nature of the data
- Data labels
- Types of Outlier/Anomaly techniques

+ 1. Nature of the Data: Input Data

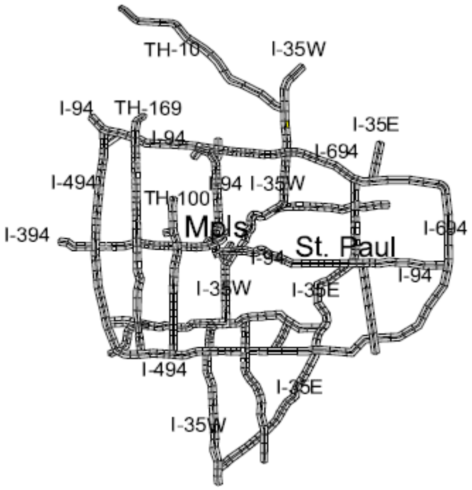
- Most common form of data handled by anomaly detection techniques is ***Record Data***
  - Univariate
  - Multivariate
- Nature of attributes
  - Binary
  - Categorical
  - Continuous
  - Hybrid

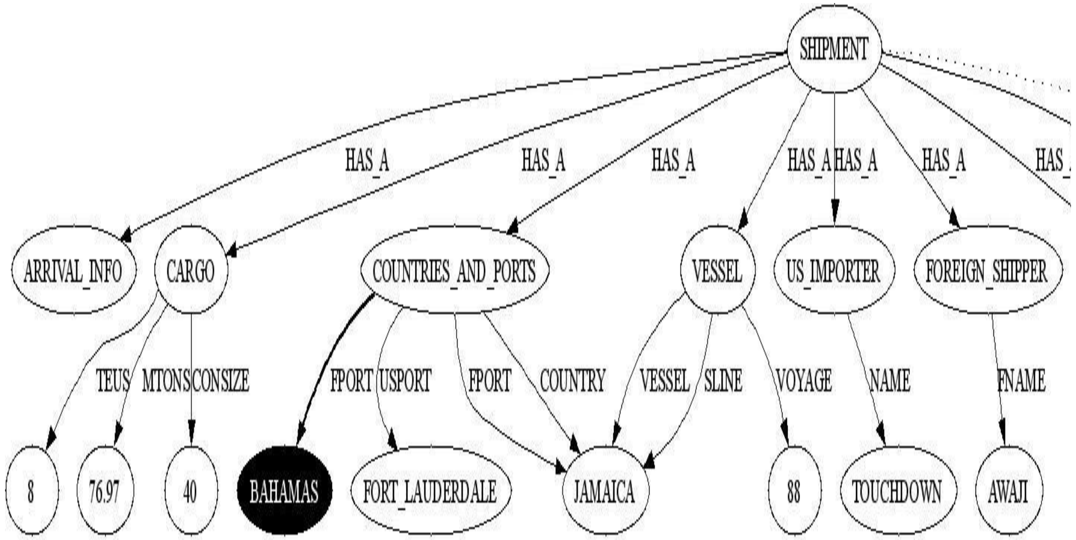
	categorical	continuous	categorical		continuous	Binary
Tid	SrcIP	Start time	Dest IP	Dest Port	Number	Attack
1	206.135.38.95	11:07:20	160.94.179.223	139	192	No
2	206.163.37.95	11:13:56	160.94.179.219	139	195	No
3	206.163.37.95	11:14:29	160.94.179.217	139	180	No
4	206.163.37.95	11:14:30	160.94.179.255	139	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

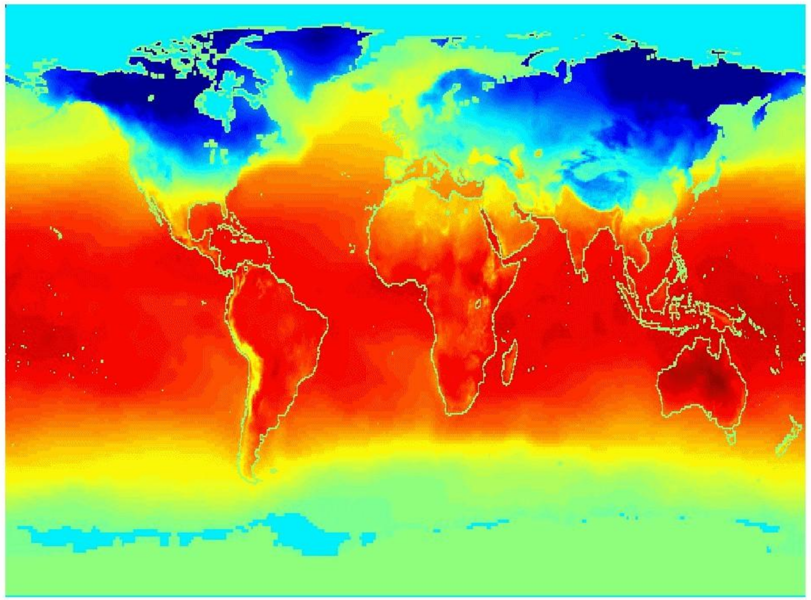
- Relationship among data instances
  - Sequential
    - Temporal
  - Spatial
  - Spatio-temporal
  - Graph

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





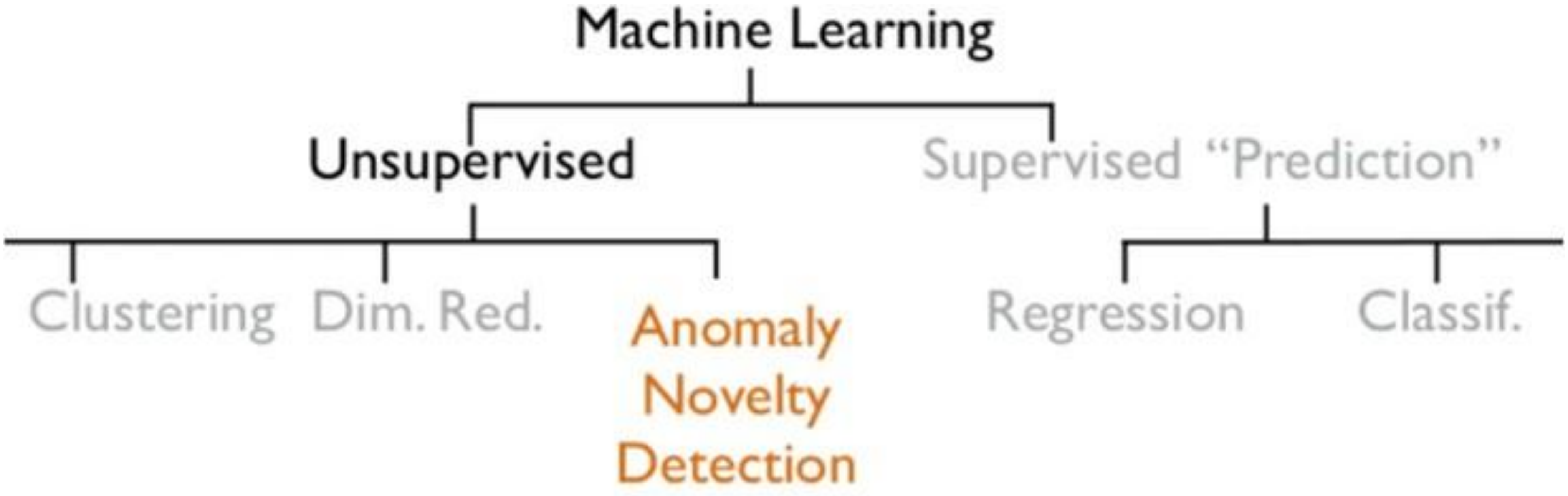
Jan



+ 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

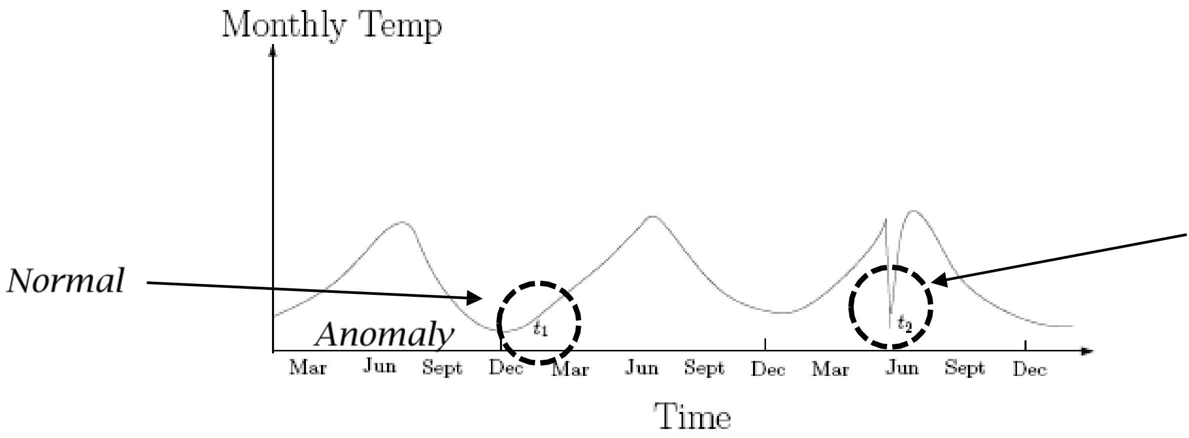
+ Global Anomalies

- Affect the entire system uniformly.
- Represent a sudden or consistent deviation across all data points.
- Example: All sensors in a factory show unusually high temperatures at the same time.

X

+ Contextual Anomalies

- An individual data instance is anomalous ***within a context***
- Requires a ***notion*** of ***context***
- Also referred to as ***conditional anomalies****



- Occur in a specific context (such as time, location, or condition).
- The data point is only abnormal under certain circumstances.
- Example: A temperature of 25°C is normal in summer but abnormal in winter.

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

+ Collective Anomalies

- Involve unusual patterns among a group of data points.
- Each point may look normal alone but abnormal together.
- **Example:** Multiple network devices suddenly show identical traffic spikes.

X

+ 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***

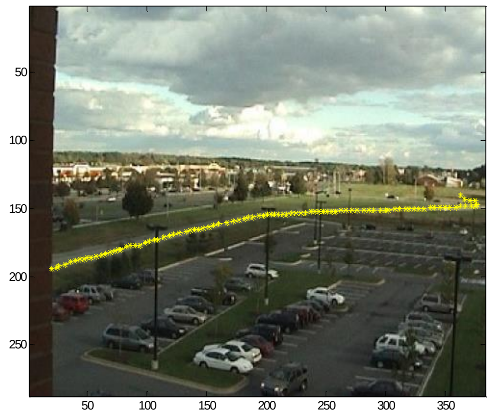
+ Applications of Anomaly Detection

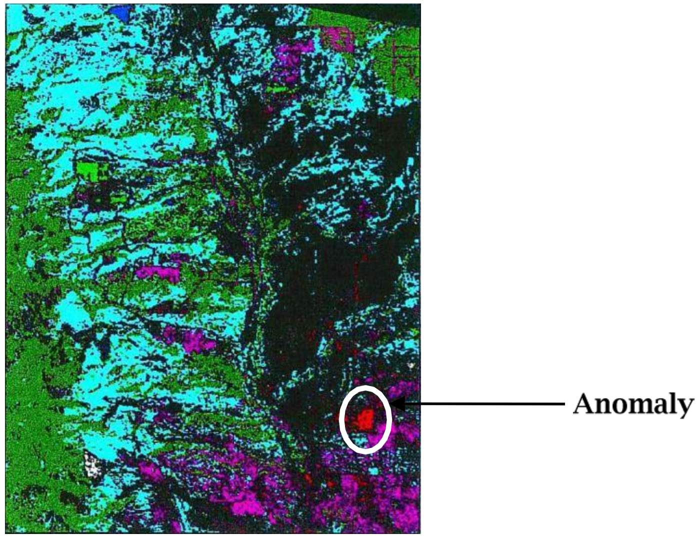
- ***Fraud detection:***
  - Fraud detection refers to detection of ***criminal activities*** occurring in ***commercial organizations***.
  - Malicious users might be the ***actual customers*** of the ***organization*** or might be ***posing*** as a ***customer*** (also ***known as identity theft***).
- ***Types of fraud***
  - Credit card fraud.
  - Insurance claim fraud
  - Mobile / cell phone fraud
  - Insider trading

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

+ 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*





+ 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

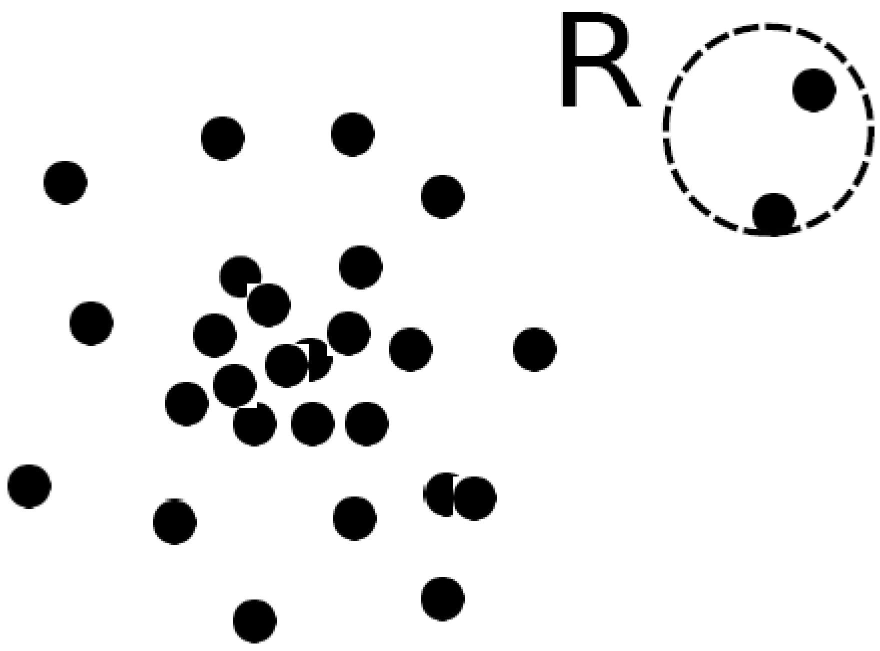
- Assume the normal objects are somewhat "***clustered***" into ***multiple groups***, each having some ***distinct features***
- An outlier is expected to be ***far away*** from any groups of normal objects
- ***Weakness***: ***Cannot detect collective outlier effectively***
  - Normal objects ***may 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***

+ 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.***



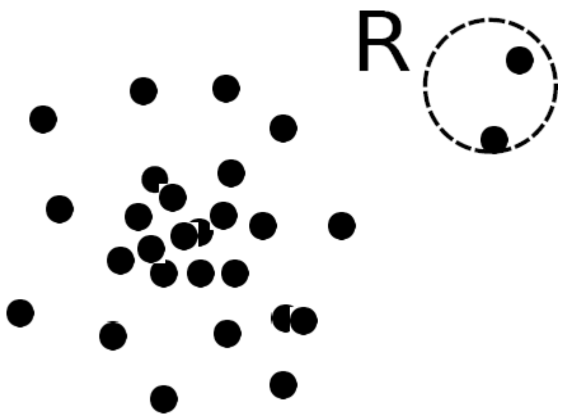
+ Challenges: Proximity based

- 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

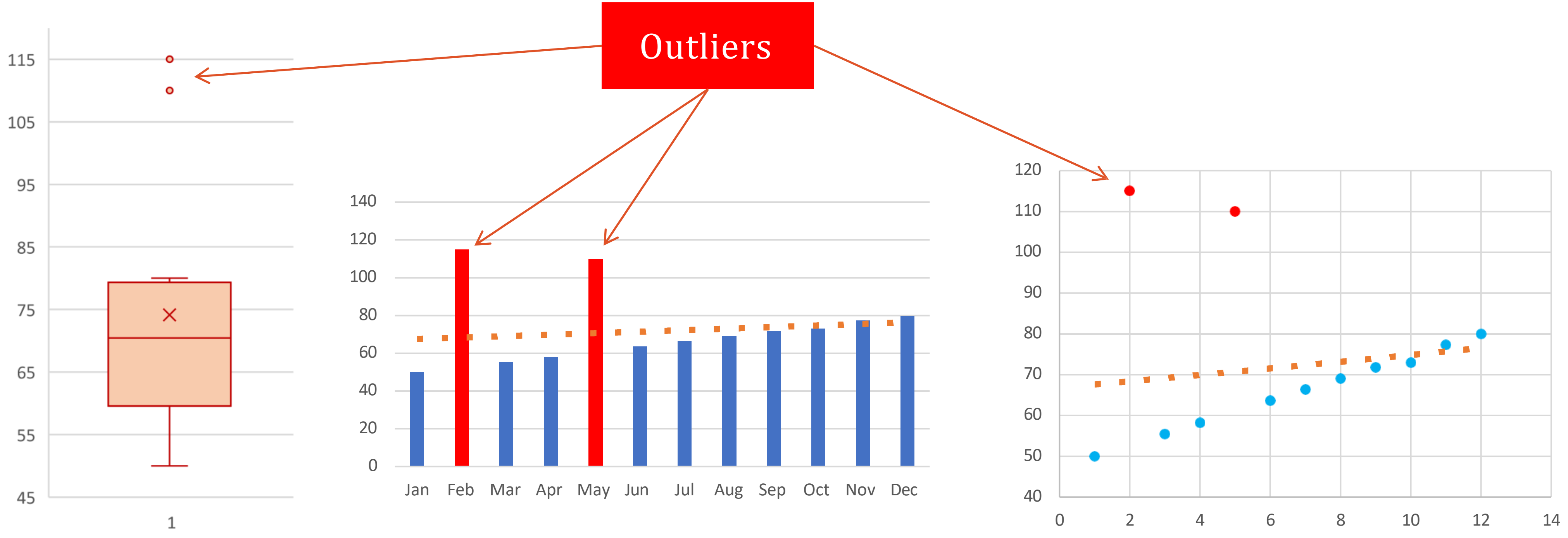


- *Challenges*
  - Clustering is expensive: straightforward adaption of a clustering method for outlier detection can be costly and does not scale up well for large data sets.

+ 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

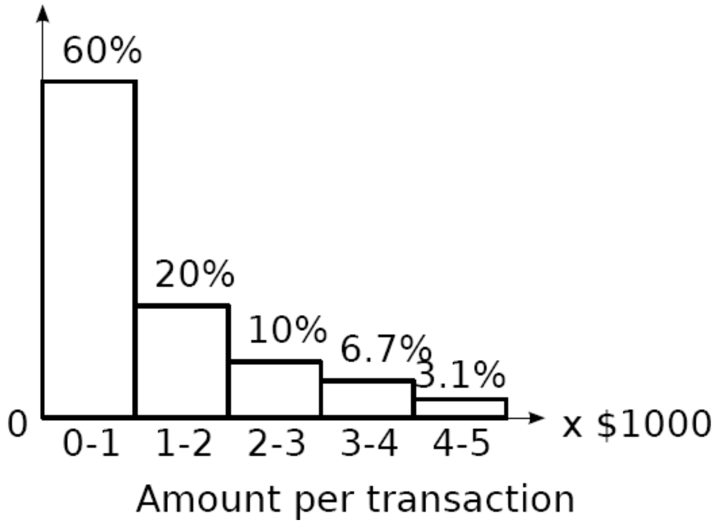
- Assumption: the normal data is generated by a ***parametric distribution*** with ***parameter θ.***

- 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

End of Lecture – 03