# Chap4: Machine Learning for Spam & Anomaly Detection

April 11, 2023

Devesh C Jinwala,
Professor, SVNIT and Adjunct Prof., CSE, IIT Jammu
Department of Computer Science and Engineering,
Sardar Vallabhhai National Institute of Technology, SURAT

# Topics to study in Chapter 1

- Machine learning for Anomaly Detection: Definition of an anomaly. Types of Anomalies or outliers in machine learning. Motivation for machine learning for anomaly detection.
  Data Visualization. Supervised, Unsupervised and Semi-supervised Learning methods for Anomaly Detection.
  Applications of Anomaly Detection: Intrusion detection, Fraud detection, Health monitoring, Defect detection, and lastly Spam detection. Intrusion Detection with Heuristics.Goodness-of-fit. Host Intrusion Detection. Network Intrusion Detection. Web Application Intrusion Detection.
  Overview of Machine learning Approaches for Anomaly Detection: Distance-based, Clustering-based and Model-based Approaches. Algorithms for Distance and Density-based approaches, Rank-based approaches, Ensemble Methods Algorithms for Time Series Data. Deep Learning for Anomaly Detection. Behavioural-based Anomaly Detection [8 hours]

# *Anomaly Detection: Background & Basics*

# Background: Scammers and Detection of attacks

- A central aspect of several instances of crime....

*...continued*

- A central aspect of several instances of crime....
  - scammers often, take advantage of a potential of humans to confuse the plausible and possible with the not improbable.

...*continued*

# Background: Scammers and Detection of attacks

- A central aspect of several instances of crime....
    - scammers often, take advantage of a potential of humans to confuse the plausible and possible with the not improbable.
    - e.g. ?? let us try to explore for examples ... ... ... ...

*...continued*

# Background: Scammers and Detection of attacks

- A central aspect of several instances of crime....
  - scammers often, take advantage of a potential of humans to confuse the plausible and possible with the not improbable.
  - e.g. ?? let us try to explore for examples ... ... ... ...
- Thus, one of the central focus in detection of attacks is .....*When network or the software attacks/other problems occur, how to identify them ?*

*...continued*

# Background: Scammers and Detection of attacks

- A central aspect of several instances of crime....
  - scammers often, take advantage of a potential of humans to confuse the plausible and possible with the not improbable.
  - e.g. ?? let us try to explore for examples ... ... ... ...
- Thus, one of the central focus in detection of attacks is .....*When network or the software attacks/other problems occur, how to identify them ?*
- the answer .....

*...continued*

# Background: Scammers and Detection of attacks

- A central aspect of several instances of crime....
  - scammers often, take advantage of a potential of humans to confuse the plausible and possible with the not improbable.
  - e.g. ?? let us try to explore for examples ... ... ... ...
- Thus, one of the central focus in detection of attacks is .....*When network or the software attacks/other problems occur, how to identify them ?*
- the answer .....
  - could be based on identifying the abnormalities in underlying behavior

*...continued*

# Background: Scammers and Detection of attacks

- A central aspect of several instances of crime....
  - scammers often, take advantage of a potential of humans to confuse the plausible and possible with the not improbable.
  - e.g. ?? let us try to explore for examples ... ... ... ...
- Thus, one of the central focus in detection of attacks is .....*When network or the software attacks/other problems occur, how to identify them ?*
- the answer .....
  - could be based on identifying the abnormalities in underlying behavior
  - in turn is based on separation of "normal" (non-attack) modes of behavior.

*...continued*

# Background: Scammers and Detection of attacks

- A central aspect of several instances of crime....
  - scammers often, take advantage of a potential of humans to confuse the plausible and possible with the not improbable.
  - e.g. ?? let us try to explore for examples ... ... ... ...
- Thus, one of the central focus in detection of attacks is .....*When network or the software attacks/other problems occur, how to identify them ?*
- the answer .....
  - could be based on identifying the abnormalities in underlying behavior
  - in turn is based on separation of "normal" (non-attack) modes of behavior.
- Thus, one would wish to analyze the broad approaches in detection. These include the following......

*...continued*

# Background: Scammers and Detection of attacks...

- The broad approaches in detection include the following:

# Background: Scammers and Detection of attacks...

- The broad approaches in detection include the following:
  - Look for damage after damage has occurred (e.g., look for unexpected transactions in a credit card bill);

# Background: Scammers and Detection of attacks...

- The broad approaches in detection include the following:
  - Look for damage after damage has occurred (e.g., look for unexpected transactions in a credit card bill);
  - Detect signs of damage (e.g., less data space available on hard drive than expected);

# Background: Scammers and Detection of attacks...

- The broad approaches in detection include the following:
  - Look for damage after damage has occurred (e.g., look for unexpected transactions in a credit card bill);
  - Detect signs of damage (e.g., less data space available on hard drive than expected);
  - Pattern-match against known "signatures" of malware, Internet site addresses, message routing paths, etc.; and

# Background: Scammers and Detection of attacks...

- The broad approaches in detection include the following:
  - Look for damage after damage has occurred (e.g., look for unexpected transactions in a credit card bill);
  - Detect signs of damage (e.g., less data space available on hard drive than expected);
  - Pattern-match against known "signatures" of malware, Internet site addresses, message routing paths, etc.; and
  - Anomaly detection: compare against expected or normal behaviors or data. Focus in this chapter.

# Background: Scammers and Detection of attacks...

- The broad approaches in detection include the following:
  - Look for damage after damage has occurred (e.g., look for unexpected transactions in a credit card bill);
  - Detect signs of damage (e.g., less data space available on hard drive than expected);
  - Pattern-match against known "signatures" of malware, Internet site addresses, message routing paths, etc.; and
  - Anomaly detection: compare against expected or normal behaviors or data. Focus in this chapter.

# Background: Scammers and Detection of attacks...

- The broad approaches in detection include the following:
  - Look for damage after damage has occurred (e.g., look for unexpected transactions in a credit card bill);
  - Detect signs of damage (e.g., less data space available on hard drive than expected);
  - Pattern-match against known "signatures" of malware, Internet site addresses, message routing paths, etc.; and
  - Anomaly detection: compare against expected or normal behaviors or data. Focus in this chapter.

• It is important to note that *relating to past experience is very vital and is of central focus...*
• e.g. sudden jailing of an acquaintance (and his sudden need for money)
• here, the probability of being scammed is much higher than the probability of one's acquaintance suddenly being jailed in a foreign country and needing money.

# Background: Scammers and Detection of attacks...

- It should be noted that following detection of an attack, further actions are required to be taken to recover from the attack.

# Background: Scammers and Detection of attacks...

- It should be noted that following detection of an attack, further actions are required to be taken to recover from the attack.
- These may include the following:

# Background: Scammers and Detection of attacks...

- It should be noted that following detection of an attack, further actions are required to be taken to recover from the attack.
- These may include the following:
    - patching software holes,

- It should be noted that following detection of an attack, further actions are required to be taken to recover from the attack.
- These may include the following:
  - patching software holes,
  - deleting undesirable files,

# Background: Scammers and Detection of attacks...

- It should be noted that following detection of an attack, further actions are required to be taken to recover from the attack.
- These may include the following:
    - patching software holes,
    - deleting undesirable files,
    - changing passwords,

# Background: Scammers and Detection of attacks...

- It should be noted that following detection of an attack, further actions are required to be taken to recover from the attack.
- These may include the following:
  - patching software holes,
  - deleting undesirable files,
  - changing passwords,
  - changing hardware,

# Background: Scammers and Detection of attacks...

- It should be noted that following detection of an attack, further actions are required to be taken to recover from the attack.
- These may include the following:
  - patching software holes,
  - deleting undesirable files,
  - changing passwords,
  - changing hardware,
  - etc.;

# Background: Scammers and Detection of attacks...

- It should be noted that following detection of an attack, further actions are required to be taken to recover from the attack.
- These may include the following:
  - patching software holes,
  - deleting undesirable files,
  - changing passwords,
  - changing hardware,
  - etc.;
- however, such recovery aspects are application-dependent, and hence out of the scope of this course.

# What is an anomaly? Its various definitions

Various definitions of anomalies and outliers....

- are rare items, events or observations which raise suspicions by differing significantly from the majority of the data.

[1] https://medium.com/@mehulved1503/outlier-detection-and-anomaly-detection-with-machine-learning-caa96b34b7f6

# What is an anomaly? Its various definitions

Various definitions of anomalies and outliers....

- are rare items, events or observations which raise suspicions by differing significantly from the majority of the data.
- are rare events that are statistically distant, and their early identification helps in avoiding biased results in analysis.

[1] https://medium.com/@mehulved1503/outlier-detection-and-anomaly-detection-with-machine-learning-caa96b34b7f6

Various definitions of anomalies and outliers....

- are rare items, events or observations which raise suspicions by differing significantly from the majority of the data.
- are rare events that are statistically distant, and their early identification helps in avoiding biased results in analysis.
- represent substantial variations from the norm.

[1] https://medium.com/@mehulved1503/outlier-detection-and-anomaly-detection-with-machine-learning-caa96b34b7f6

# What is an anomaly? Its various definitions

Various definitions of anomalies and outliers....

- are rare items, events or observations which raise suspicions by differing significantly from the majority of the data.
- are rare events that are statistically distant, and their early identification helps in avoiding biased results in analysis.
- represent substantial variations from the norm.
- are observations exhibiting abnormal behaviour compared to the majority of the samples.

[1] https://medium.com/@mehulved1503/outlier-detection-and-anomaly-detection-with-machine-learning-caa96b34b7f6

# What is an anomaly? Its various definitions

Various definitions of anomalies and outliers....

- are rare items, events or observations which raise suspicions by differing significantly from the majority of the data.
- are rare events that are statistically distant, and their early identification helps in avoiding biased results in analysis.
- represent substantial variations from the norm.
- are observations exhibiting abnormal behaviour compared to the majority of the samples.
- are linked to some sort of problem or rare event such as hacking, bank fraud, malfunctioning equipment, structural defects/infrastructure failures, or textual errors. [1]

[1] https://medium.com/@mehulved1503/outlier-detection-and-anomaly-detection-with-machine-learning-caa96b34b7f6

# What is an anomaly? Its various definitions

Various definitions of anomalies and outliers....

- are rare items, events or observations which raise suspicions by differing significantly from the majority of the data.
- are rare events that are statistically distant, and their early identification helps in avoiding biased results in analysis.
- represent substantial variations from the norm.
- are observations exhibiting abnormal behaviour compared to the majority of the samples.
- are linked to some sort of problem or rare event such as hacking, bank fraud, malfunctioning equipment, structural defects/infrastructure failures, or textual errors. [1]
- represent events, items, or observations which are suspicious because they differ significantly from standard behaviors or patterns.

[1] https://medium.com/@mehulved1503/outlier-detection-and-anomaly-detection-with-machine-learning-caa96b34b7f6

# What is an anomaly? Its various definitions

Various definitions of anomalies and outliers....

- are rare items, events or observations which raise suspicions by differing significantly from the majority of the data.
- are rare events that are statistically distant, and their early identification helps in avoiding biased results in analysis.
- represent substantial variations from the norm.
- are observations exhibiting abnormal behaviour compared to the majority of the samples.
- are linked to some sort of problem or rare event such as hacking, bank fraud, malfunctioning equipment, structural defects/infrastructure failures, or textual errors. [1]
- represent events, items, or observations which are suspicious because they differ significantly from standard behaviors or patterns.
- in data are also called standard deviations, outliers, noise, novelties, and exceptions.

[1] https://medium.com/@mehulved1503/outlier-detection-and-anomaly-detection-with-machine-learning-caa96b34b7f6

# Anomaly and Outliers ???

Anomalies or Outliers.....

- One definition on a blog says....

[1] https://medium.com/@mehulved1503/outlier-detection-and-anomaly-detection-with-machine-learning-caa96b34b7f6

# Anomaly and Outliers ???

Anomalies or Outliers.....

- One definition on a blog says....
  - Anomalies are patterns of different data within given data, whereas

[1] https://medium.com/@mehulved1503/outlier-detection-and-anomaly-detection-with-machine-learning-caa96b34b7f6

# Anomaly and Outliers ???

Anomalies or Outliers.....

- One definition on a blog says....
  - Anomalies are patterns of different data within given data, whereas
  - Outliers would be merely extreme data points within data.

[1] https://medium.com/@mehulved1503/outlier-detection-and-anomaly-detection-with-machine-learning-caa96b34b7f6

# Anomaly and Outliers ???

Anomalies or Outliers.....

- One definition on a blog says....
  - Anomalies are patterns of different data within given data, whereas
  - Outliers would be merely extreme data points within data.
- Thus, If not aggregated appropriately, anomalies may be neglected as outliers. [1]

[1] https://medium.com/@mehulved1503/outlier-detection-and-anomaly-detection-with-machine-learning-caa96b34b7f6

# Anomaly and Outliers ???

Anomalies or Outliers.....

- One definition on a blog says....
  - Anomalies are patterns of different data within given data, whereas
  - Outliers would be merely extreme data points within data.
- Thus, If not aggregated appropriately, anomalies may be neglected as outliers. [1]
- However, the standard textbook by Charu Aggarwal states the following:

[1] https://medium.com/@mehulved1503/outlier-detection-and-anomaly-detection-with-machine-learning-caa96b34b7f6

# Anomaly and Outliers ???

Anomalies or Outliers.....

- One definition on a blog says....
  - Anomalies are patterns of different data within given data, whereas
  - Outliers would be merely extreme data points within data.
- Thus, If not aggregated appropriately, anomalies may be neglected as outliers. [1]
- However, the standard textbook by Charu Aggarwal states the following:
  - An outlier is a data point that is significantly different from the remaining data.

[1] https://medium.com/@mehulved1503/outlier-detection-and-anomaly-detection-with-machine-learning-caa96b34b7f6

# Anomaly and Outliers ???

Anomalies or Outliers.....

- One definition on a blog says....
  - Anomalies are patterns of different data within given data, whereas
  - Outliers would be merely extreme data points within data.
- Thus, If not aggregated appropriately, anomalies may be neglected as outliers. [1]
- However, the standard textbook by Charu Aggarwal states the following:
  - An outlier is a data point that is significantly different from the remaining data.
- Hawkins defined an outlier as follows:

[1] https://medium.com/@mehulved1503/outlier-detection-and-anomaly-detection-with-machine-learning-caa96b34b7f6

# Anomaly and Outliers ???

Anomalies or Outliers.....

- One definition on a blog says....
  - Anomalies are patterns of different data within given data, whereas
  - Outliers would be merely extreme data points within data.
- Thus, If not aggregated appropriately, anomalies may be neglected as outliers. [1]
- However, the standard textbook by Charu Aggarwal states the following:
  - An outlier is a data point that is significantly different from the remaining data.
- Hawkins defined an outlier as follows:
  - *An outlier is an observation which deviates so much from the other observations as to arouse suspicions that it was generated by a different mechanism.*

[1] https://medium.com/@mehulved1503/outlier-detection-and-anomaly-detection-with-machine-learning-caa96b34b7f6

# Anomaly and Outliers ???

Anomalies or Outliers.....

- One definition on a blog says....
  - Anomalies are patterns of different data within given data, whereas
  - Outliers would be merely extreme data points within data.
- Thus, If not aggregated appropriately, anomalies may be neglected as outliers. [1]
- However, the standard textbook by Charu Aggarwal states the following:
  - An outlier is a data point that is significantly different from the remaining data.
- Hawkins defined an outlier as follows:
  - *An outlier is an observation which deviates so much from the other observations as to arouse suspicions that it was generated by a different mechanism.*
- Outliers are also referred to as abnormalities, discordants, deviants, or anomalies in the data mining and statistics literature.

[1] https://medium.com/@mehulved1503/outlier-detection-and-anomaly-detection-with-machine-learning-caa96b34b7f6

# Anomaly and Outliers ???

Anomalies or Outliers.....

- One definition on a blog says....
  - Anomalies are patterns of different data within given data, whereas
  - Outliers would be merely extreme data points within data.
- Thus, If not aggregated appropriately, anomalies may be neglected as outliers. [1]
- However, the standard textbook by Charu Aggarwal states the following:
  - An outlier is a data point that is significantly different from the remaining data.
- Hawkins defined an outlier as follows:
  - *An outlier is an observation which deviates so much from the other observations as to arouse suspicions that it was generated by a different mechanism.*
- Outliers are also referred to as abnormalities, discordants, deviants, or anomalies in the data mining and statistics literature.
- Thus, the terms outliers and anomalies are treated as synonyms further.

[1] https://medium.com/@mehulved1503/outlier-detection-and-anomaly-detection-with-machine-learning-caa96b34b7f6

- Many scientific and engineering fields are based on the assumption that

# Normal data implies Normal hypotheses, Abnormal data implies ??

- Many scientific and engineering fields are based on the assumption that
  - processes or behaviors exist in nature that follow certain rules or broad principles

# Normal data implies Normal hypotheses, Abnormal data implies ??

- Many scientific and engineering fields are based on the assumption that
  - processes or behaviors exist in nature that follow certain rules or broad principles
  - this results in the state of a system, manifested in observable data.

# Normal data implies Normal hypotheses, Abnormal data implies ??

- Many scientific and engineering fields are based on the assumption that
  - processes or behaviors exist in nature that follow certain rules or broad principles
  - this results in the state of a system, manifested in observable data.
  - it is possible to formulate hypotheses about the nature of the underlying process from the data, which can be verified upon observation of additional data.

# Normal data implies Normal hypotheses, Abnormal data implies ??

- Many scientific and engineering fields are based on the assumption that
  - processes or behaviors exist in nature that follow certain rules or broad principles
  - this results in the state of a system, manifested in observable data.
  - it is possible to formulate hypotheses about the nature of the underlying process from the data, which can be verified upon observation of additional data.
- these hypotheses describe the normal behavior of a system

# Normal data implies Normal hypotheses, Abnormal data implies ??

- Many scientific and engineering fields are based on the assumption that
  - processes or behaviors exist in nature that follow certain rules or broad principles
  - this results in the state of a system, manifested in observable data.
  - it is possible to formulate hypotheses about the nature of the underlying process from the data, which can be verified upon observation of additional data.
- these hypotheses describe the normal behavior of a system
- this is implicitly assuming that the data used to generate the hypotheses are typical of the system or are expected data of the system.

# Normal data implies Normal hypotheses, Abnormal data implies ??

- Many scientific and engineering fields are based on the assumption that
    - processes or behaviors exist in nature that follow certain rules or broad principles
    - this results in the state of a system, manifested in observable data.
    - it is possible to formulate hypotheses about the nature of the underlying process from the data, which can be verified upon observation of additional data.
- these hypotheses describe the normal behavior of a system
- this is implicitly assuming that the data used to generate the hypotheses are typical of the system or are expected data of the system.
- however, variations from the norm may occur in the processes, hence systems may also exist in abnormal states,

# Normal data implies Normal hypotheses, Abnormal data implies ??

- Many scientific and engineering fields are based on the assumption that
  - processes or behaviors exist in nature that follow certain rules or broad principles
  - this results in the state of a system, manifested in observable data.
  - it is possible to formulate hypotheses about the nature of the underlying process from the data, which can be verified upon observation of additional data.
- these hypotheses describe the normal behavior of a system
- this is implicitly assuming that the data used to generate the hypotheses are typical of the system or are expected data of the system.
- however, variations from the norm may occur in the processes, hence systems may also exist in abnormal states,
- this leads to observable data values that are different from the values observed when no such process/state variations occur.

## Anomaly detection

- serves to discover such variations (from the norm) in the observed data values,

## Anomaly detection

- serves to discover such variations (from the norm) in the observed data values,
- hence infer the variations in the underlying process.

# Normal data implies Normal hypotheses, Abnormal data implies ??...

## Anomaly detection

- serves to discover such variations (from the norm) in the observed data values,
- hence infer the variations in the underlying process.

# Normal data implies Normal hypotheses, Abnormal data implies ??...

## Anomaly detection

- serves to discover such variations (from the norm) in the observed data values,
- hence infer the variations in the underlying process.

However, the task of outlier detection is not as trivial as it sounds.....

# Outlier detection is non-trivial

However, the task of outlier detection is not as trivial as it sounds.....

[1] S. Papadimitriou, H. Kitagawa, P.B. Gibbons, C. Faloutsos, "Loci: Fast outlier detection using the local correlation integral," in Proceedings. 19th International Conference on Data Engineering, 2003 (IEEE, Washington, DC, 2003, pp. 315–326

# Outlier detection is non-trivial

However, the task of outlier detection is not as trivial as it sounds.....

## A fundamental problem

- that there is no simple unique definition that permits us to evaluate how similar are two data points, and hence how different is one data point from others in the data set.
- Papadimitriou et al. in [1]: ... *there is an inherent fuzziness in the concept of outlier and any outlier score is (only) an informative indicator than a precise measure.*

[1] S. Papadimitriou, H. Kitagawa, P.B. Gibbons, C. Faloutsos, "Loci: Fast outlier detection using the local correlation integral," in Proceedings. 19th International Conference on Data Engineering, 2003 (IEEE, Washington, DC, 2003, pp. 315–326

# Outlier detection is non-trivial

However, the task of outlier detection is not as trivial as it sounds.....

## A fundamental problem

- that there is no simple unique definition that permits us to evaluate how similar are two data points, and hence how different is one data point from others in the data set.
- Papadimitriou et al. in [1]: *. . . there is an inherent fuzziness in the concept of outlier and any outlier score is (only) an informative indicator than a precise measure.*

Let us try to understand this further......

[1] S. Papadimitriou, H. Kitagawa, P.B. Gibbons, C. Faloutsos, "Loci: Fast outlier detection using the local correlation integral," in Proceedings. 19th International Conference on Data Engineering, 2003 (IEEE, Washington, DC, 2003, pp. 315–326

## *Anomaly Detection*

Anomaly Detection

- refers to identifying unexpected intruders or breaches.

Anomaly Detection

- refers to identifying unexpected intruders or breaches.
- it takes tens of days for a system breach to be detected, an an average.

# Introduction: Anomaly Detection

Anomaly Detection

- refers to identifying unexpected intruders or breaches.
- it takes tens of days for a system breach to be detected, an an average.
- the nature of the attack could be data exfiltration, extortion through ransomware, adware, or advanced persistent threats (APTs).

Anomaly Detection

- refers to identifying unexpected intruders or breaches.
- it takes tens of days for a system breach to be detected, an an average.
- the nature of the attack could be data exfiltration, extortion through ransomware, adware, or advanced persistent threats (APTs).

# Introduction: Anomaly Detection

Anomaly Detection

- refers to identifying unexpected intruders or breaches.
- it takes tens of days for a system breach to be detected, an an average.
- the nature of the attack could be data exfiltration, extortion through ransomware, adware, or advanced persistent threats (APTs).

Verizon's 2022 Data Breach Investigations Report

Anomaly Detection

- refers to identifying unexpected intruders or breaches.
- it takes tens of days for a system breach to be detected, an an average.
- the nature of the attack could be data exfiltration, extortion through ransomware, adware, or advanced persistent threats (APTs).

Verizon's 2022 Data Breach Investigations Report

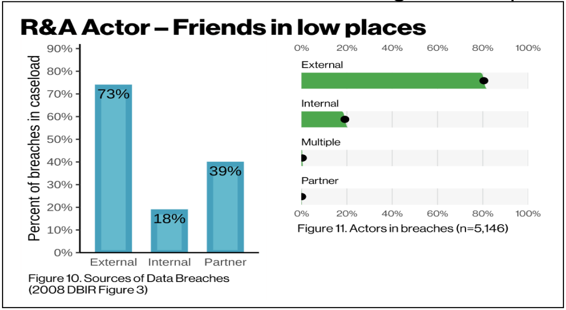


Figure: Sources of Data Breaches

:https://www.verizon.com/business/resources/reports/dbir/2022/results-and-analysis-intro/

Anomaly Detection

- is not confined to the context of security.

Anomaly Detection

- is not confined to the context of security.
- is any method for finding events that don't conform to an expectation.

Anomaly Detection

- is not confined to the context of security.
- is any method for finding events that don't conform to an expectation.
- applies to various applications....

Anomaly Detection

- is not confined to the context of security.
- is any method for finding events that don't conform to an expectation.
- applies to various applications....
  - identify early signs of system failure in system reliability - finding anomalies in the electrical power grid can potentially avoid expensive damage due to power surges

Anomaly Detection

- is not confined to the context of security.
- is any method for finding events that don't conform to an expectation.
- applies to various applications....
  - identify early signs of system failure in system reliability - finding anomalies in the electrical power grid can potentially avoid expensive damage due to power surges
  - identifying fraud in the financial industry requires to find an anomaly in a series of legitimate transactions.

Anomaly Detection

- Anomaly detection is different from misuse detection

Anomaly Detection

- Anomaly detection is different from misuse detection
  - the latter first defines the signature of abnormal behavior to indicate attacks whereas

# Introduction: Anomaly Detection & misuse detection

Anomaly Detection

- Anomaly detection is different from misuse detection
  - the latter first defines the signature of abnormal behavior to indicate attacks whereas
  - the former first defines a profile of normal behaviors, which reflects the health and sensitivity of a cyber-infrastructure.

# Introduction: Anomaly Detection & misuse detection

Anomaly Detection

- Anomaly detection is different from misuse detection
  - the latter first defines the signature of abnormal behavior to indicate attacks whereas
  - the former first defines a profile of normal behaviors, which reflects the health and sensitivity of a cyber-infrastructure.
- Correspondingly, an anomaly is defined as

# Introduction: Anomaly Detection & misuse detection

Anomaly Detection

- Anomaly detection is different from misuse detection
  - the latter first defines the signature of abnormal behavior to indicate attacks whereas
  - the former first defines a profile of normal behaviors, which reflects the health and sensitivity of a cyber-infrastructure.
- Correspondingly, an anomaly is defined as
  - a pattern in data that does not conform to the expected behaviors and include outliers, abbreviations, contaminants, and surprise, etc., in applications.

# Anomaly Detection: Anomalous vs Normal behaviour

How does anomaly detection work, broadly ?

- anomaly detection relies on a clear boundary between normal and anomalous behaviors

# Anomaly Detection: Anomalous vs Normal behaviour

How does anomaly detection work, broadly ?

- anomaly detection relies on a clear boundary between normal and anomalous behaviors
- the profile of normal behaviour is defined based on one or more of the following:

# Anomaly Detection: Anomalous vs Normal behaviour

How does anomaly detection work, broadly ?

- anomaly detection relies on a clear boundary between normal and anomalous behaviors
- the profile of normal behaviour is defined based on one or more of the following:
  - host/IP address or VLAN segment addresses

# Anomaly Detection: Anomalous vs Normal behaviour

How does anomaly detection work, broadly ?

- anomaly detection relies on a clear boundary between normal and anomalous behaviors
- the profile of normal behaviour is defined based on one or more of the following:
  - host/IP address or VLAN segment addresses
  - occurrence patterns of specific commands in application protocols,

# Anomaly Detection: Anomalous vs Normal behaviour

How does anomaly detection work, broadly ?

- anomaly detection relies on a clear boundary between normal and anomalous behaviors
- the profile of normal behaviour is defined based on one or more of the following:
  - host/IP address or VLAN segment addresses
  - occurrence patterns of specific commands in application protocols,
  - association of content types with different fields of application protocols,

# Anomaly Detection: Anomalous vs Normal behaviour

How does anomaly detection work, broadly ?

- anomaly detection relies on a clear boundary between normal and anomalous behaviors
- the profile of normal behaviour is defined based on one or more of the following:
  - host/IP address or VLAN segment addresses
  - occurrence patterns of specific commands in application protocols,
  - association of content types with different fields of application protocols,
  - connectivity patterns between protected servers and the outside world,

# Anomaly Detection: Anomalous vs Normal behaviour

How does anomaly detection work, broadly ?

- anomaly detection relies on a clear boundary between normal and anomalous behaviors
- the profile of normal behaviour is defined based on one or more of the following:
    - host/IP address or VLAN segment addresses
    - occurrence patterns of specific commands in application protocols,
    - association of content types with different fields of application protocols,
    - connectivity patterns between protected servers and the outside world,
    - rate and burst length distributions for all types of traffic

# Anomaly Detection: Anomalous vs Normal behaviour

How does anomaly detection work, broadly ?

- anomaly detection relies on a clear boundary between normal and anomalous behaviors
- the profile of normal behaviour is defined based on one or more of the following:
  - host/IP address or VLAN segment addresses
  - occurrence patterns of specific commands in application protocols,
  - association of content types with different fields of application protocols,
  - connectivity patterns between protected servers and the outside world,
  - rate and burst length distributions for all types of traffic
  - ... ... ...

# Anomaly Detection: Anomalous vs Normal behaviour

How does anomaly detection work, broadly ?

- anomaly detection relies on a clear boundary between normal and anomalous behaviors
- the profile of normal behaviour is defined based on one or more of the following:
  - host/IP address or VLAN segment addresses
  - occurrence patterns of specific commands in application protocols,
  - association of content types with different fields of application protocols,
  - connectivity patterns between protected servers and the outside world,
  - rate and burst length distributions for all types of traffic
  - ... ... ...
- Needless to say, whatever may be the data used in defining the profile, profiles based on a network

# Anomaly Detection: Anomalous vs Normal behaviour

How does anomaly detection work, broadly ?

- anomaly detection relies on a clear boundary between normal and anomalous behaviors
- the profile of normal behaviour is defined based on one or more of the following:
  - host/IP address or VLAN segment addresses
  - occurrence patterns of specific commands in application protocols,
  - association of content types with different fields of application protocols,
  - connectivity patterns between protected servers and the outside world,
  - rate and burst length distributions for all types of traffic
  - ... ... ...
- Needless to say, whatever may be the data used in defining the profile, profiles based on a network
  - must be adaptive and self-learning in complex and challenging network traffic

# Anomaly Detection: Anomalous vs Normal behaviour

How does anomaly detection work, broadly ?

- anomaly detection relies on a clear boundary between normal and anomalous behaviors
- the profile of normal behaviour is defined based on one or more of the following:
  - host/IP address or VLAN segment addresses
  - occurrence patterns of specific commands in application protocols,
  - association of content types with different fields of application protocols,
  - connectivity patterns between protected servers and the outside world,
  - rate and burst length distributions for all types of traffic
  - ... ... ...
- Needless to say, whatever may be the data used in defining the profile, profiles based on a network
  - must be adaptive and self-learning in complex and challenging network traffic

# Anomaly Detection: Anomalous vs Normal behaviour

How does anomaly detection work, broadly ?

- anomaly detection relies on a clear boundary between normal and anomalous behaviors
- the profile of normal behaviour is defined based on one or more of the following:
    - host/IP address or VLAN segment addresses
    - occurrence patterns of specific commands in application protocols,
    - association of content types with different fields of application protocols,
    - connectivity patterns between protected servers and the outside world,
    - rate and burst length distributions for all types of traffic
    - ... ... ...
- Needless to say, whatever may be the data used in defining the profile, profiles based on a network
    - must be adaptive and self-learning in complex and challenging network traffic

So, then what are the examples of anomalous behaviour?

# Anomaly Detection: What are anomalous behaviour?

Typical anomalous behaviour include any of the following

- segmentation of binary code in a user password, what is it ?

# Anomaly Detection: What are anomalous behaviour?

Typical anomalous behaviour include any of the following

- segmentation of binary code in a user password, what is it ?
- stealthy reconnaissance attempts, what is it ?

# Anomaly Detection: What are anomalous behaviour?

Typical anomalous behaviour include any of the following

- segmentation of binary code in a user password, what is it ?
- stealthy reconnaissance attempts, what is it ?
- backdoor service on a well-known standard port, what is it ?

# Anomaly Detection: What are anomalous behaviour?

Typical anomalous behaviour include any of the following

- segmentation of binary code in a user password, what is it ?
- stealthy reconnaissance attempts, what is it ?
- backdoor service on a well-known standard port, what is it ?
- natural failures in the network, what is it ?

# Anomaly Detection: What are anomalous behaviour?

Typical anomalous behaviour include any of the following

- segmentation of binary code in a user password, what is it ?
- stealthy reconnaissance attempts, what is it ?
- backdoor service on a well-known standard port, what is it ?
- natural failures in the network, what is it ?
- new buffer overflow attacks, what is it ?

# Anomaly Detection: What are anomalous behaviour?

Typical anomalous behaviour include any of the following

- segmentation of binary code in a user password, what is it ?
- stealthy reconnaissance attempts, what is it ?
- backdoor service on a well-known standard port, what is it ?
- natural failures in the network, what is it ?
- new buffer overflow attacks, what is it ?
- HTTP traffic on a nonstandard port, what is it ?

# Anomaly Detection: What are anomalous behaviour?

Typical anomalous behaviour include any of the following

- segmentation of binary code in a user password, what is it ?
- stealthy reconnaissance attempts, what is it ?
- backdoor service on a well-known standard port, what is it ?
- natural failures in the network, what is it ?
- new buffer overflow attacks, what is it ?
- HTTP traffic on a nonstandard port, what is it ?
- intentionally stealthy attacks, what is it ?

# Anomaly Detection: What are anomalous behaviour?

Typical anomalous behaviour include any of the following

- segmentation of binary code in a user password, what is it ?
- stealthy reconnaissance attempts, what is it ?
- backdoor service on a well-known standard port, what is it ?
- natural failures in the network, what is it ?
- new buffer overflow attacks, what is it ?
- HTTP traffic on a nonstandard port, what is it ?
- intentionally stealthy attacks, what is it ?
- variants of existing attacks in new environments, what is it ?

# Anomaly Detection: What are anomalous behaviour?

Typical anomalous behaviour include any of the following

- segmentation of binary code in a user password, what is it ?
- stealthy reconnaissance attempts, what is it ?
- backdoor service on a well-known standard port, what is it ?
- natural failures in the network, what is it ?
- new buffer overflow attacks, what is it ?
- HTTP traffic on a nonstandard port, what is it ?
- intentionally stealthy attacks, what is it ?
- variants of existing attacks in new environments, what is it ?
- unusual login time and source IP address, what is it?

# Anomaly Detection: What are anomalous behaviour?

Typical anomalous behaviour include any of the following

- segmentation of binary code in a user password, what is it ?
- stealthy reconnaissance attempts, what is it ?
- backdoor service on a well-known standard port, what is it ?
- natural failures in the network, what is it ?
- new buffer overflow attacks, what is it ?
- HTTP traffic on a nonstandard port, what is it ?
- intentionally stealthy attacks, what is it ?
- variants of existing attacks in new environments, what is it ?
- unusual login time and source IP address, what is it?
- ... ... ... ... and so on

# Anomaly Detection: What are anomalous behaviour?

Typical anomalous behaviour include any of the following

- segmentation of binary code in a user password, what is it ?
- stealthy reconnaissance attempts, what is it ?
- backdoor service on a well-known standard port, what is it ?
- natural failures in the network, what is it ?
- new buffer overflow attacks, what is it ?
- HTTP traffic on a nonstandard port, what is it ?
- intentionally stealthy attacks, what is it ?
- variants of existing attacks in new environments, what is it ?
- unusual login time and source IP address, what is it?
- ... ... ... ... and so on

# Anomaly Detection: What are anomalous behaviour?

Typical anomalous behaviour include any of the following

- segmentation of binary code in a user password, what is it ?
- stealthy reconnaissance attempts, what is it ?
- backdoor service on a well-known standard port, what is it ?
- natural failures in the network, what is it ?
- new buffer overflow attacks, what is it ?
- HTTP traffic on a nonstandard port, what is it ?
- intentionally stealthy attacks, what is it ?
- variants of existing attacks in new environments, what is it ?
- unusual login time and source IP address, what is it?
- ... ... ... ... and so on

Let us try to understand this with different usecases....

*Understanding anomalies: An overview of anomaly detection applications areas*

# Anomalies: IQ Test results example

- results of an IQ test expected to be around 100, standard deviation (s.d.) of 15,

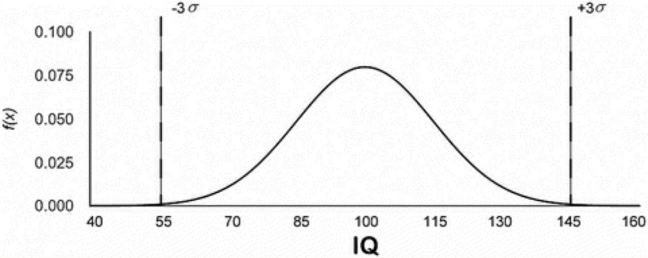


*Fig. 1.1* **Distribution of Intelligence Quotient (IQ) scores**—Mean of 100, Standard Deviation of 15

Figure: Results of an IQ Test

Src.:Mehrotra, Kishan G et al. Anomaly Detection Principles and Algorithms.... Springer International Publishing.

# Anomalies: IQ Test results example

- results of an IQ test expected to be around 100, standard deviation (s.d.) of 15,
- an IQ score of 115, varies from the norm, but not substantially.

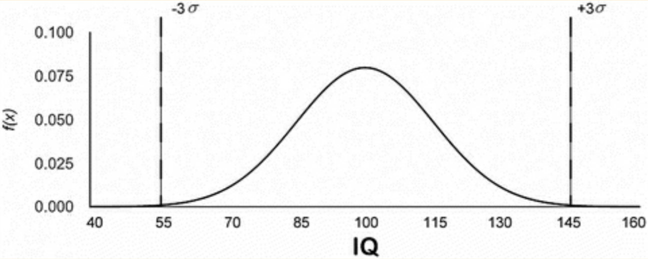


*Fig. 1.1* **Distribution of Intelligence Quotient (IQ) scores**—Mean of 100, Standard Deviation of 15

Figure: Results of an IQ Test

Src.:Mehrotra, Kishan G et al. Anomaly Detection Principles and Algorithms.... Springer International Publishing.

# Anomalies: IQ Test results example

- results of an IQ test expected to be around 100, standard deviation (s.d.) of 15,
- an IQ score of 115, varies from the norm, but not substantially.
- on the other hand a value of 145 could be substantial deviation.

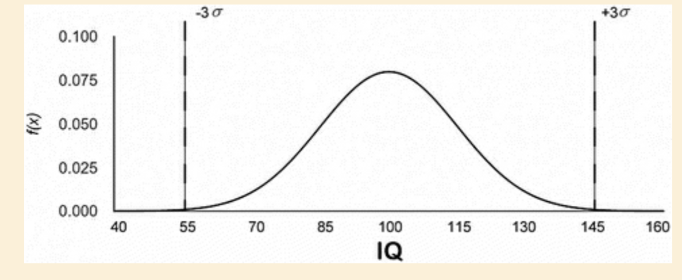


*Fig. 1.1* **Distribution of Intelligence Quotient (IQ) scores**—Mean of 100, Standard Deviation of 15

Figure: Results of an IQ Test

Src.:Mehrotra, Kishan G et al. Anomaly Detection Principles and Algorithms.... Springer International Publishing.

# Anomalies: IQ Test results example

- results of an IQ test expected to be around 100, standard deviation (s.d.) of 15,
- an IQ score of 115, varies from the norm, but not substantially.
- on the other hand a value of 145 could be substantial deviation.
- simple example - a single quantitative attribute (IQ score) with a unimodal distribution (with well-known statistics) to identify anomalies.

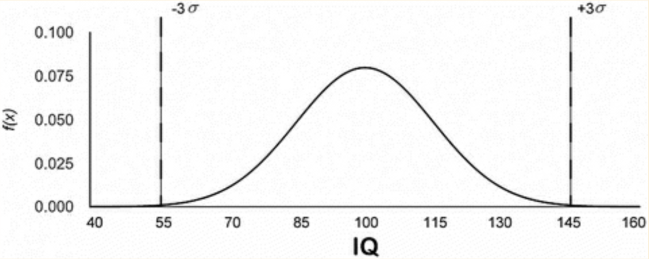


*Fig. 1.1* **Distribution of Intelligence Quotient (IQ) scores**—Mean of 100, Standard Deviation of 15

Figure: Results of an IQ Test

Src.:Mehrotra, Kishan G et al. Anomaly Detection Principles and Algorithms.... Springer International Publishing.

# Anomalies: Retailer Sales example

- Some values may seem anomalous, but are they really ?

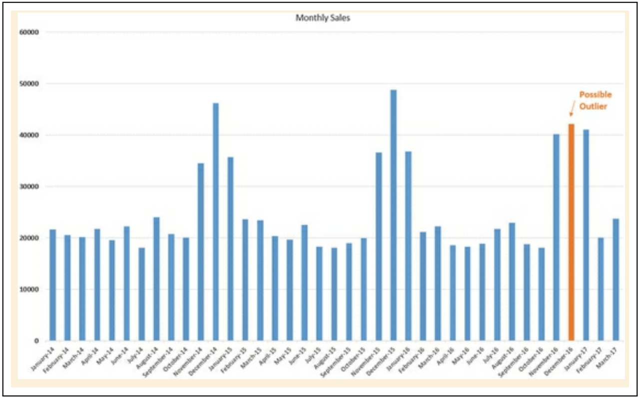


Figure: Monthly Sales for a retailer

Src.:Mehrotra, Kishan G et al. Anomaly Detection Principles and Algorithms.... Springer International Publishing.

# Anomalies: Retailer Sales example

- Some values may seem anomalous, but are they really ?
- therefore a model of the underlying process necessary.

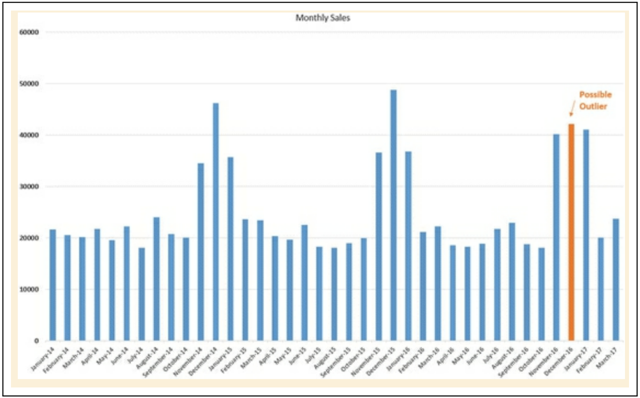


Figure: Monthly Sales for a retailer

Src.:Mehrotra, Kishan G et al. Anomaly Detection Principles and Algorithms.... Springer International Publishing.

Usecases

- Detecting anomalous log-on Patterns :

# Anomaly Detection in Cybersecurity: General

Usecases

- Detecting anomalous log-on Patterns :
  - e.g. an employee appears to have traveled an impossible distance between log-on attempts or accessing the data that they have never used before

# Anomaly Detection in Cybersecurity: General

Usecases

- Detecting anomalous log-on Patterns :
  - e.g. an employee appears to have traveled an impossible distance between log-on attempts or accessing the data that they have never used before
- Network Intrusion Detection

# Anomaly Detection in Cybersecurity: General

Usecases

- Detecting anomalous log-on Patterns :
  - e.g. an employee appears to have traveled an impossible distance between log-on attempts or accessing the data that they have never used before
- Network Intrusion Detection
  - resorting to user behavior-based anomaly detection when detecting threats in large heterogeneous environments, especially covert threats is difficult - especially in an organization having complex global nature

# Anomaly Detection in Cybersecurity: General

Usecases

- Detecting anomalous log-on Patterns :
    - e.g. an employee appears to have traveled an impossible distance between log-on attempts or accessing the data that they have never used before
- Network Intrusion Detection
    - resorting to user behavior-based anomaly detection when detecting threats in large heterogeneous environments, especially covert threats is difficult - especially in an organization having complex global nature
- Abnormal Finance Activities Detection

Usecases

- Detecting anomalous log-on Patterns :
  - e.g. an employee appears to have traveled an impossible distance between log-on attempts or accessing the data that they have never used before
- Network Intrusion Detection
  - resorting to user behavior-based anomaly detection when detecting threats in large heterogeneous environments, especially covert threats is difficult - especially in an organization having complex global nature
- Abnormal Finance Activities Detection
  - user behavior-based anomaly detection to detect and react quickly financial frauds due to social manipulation and blackmail or for personal gain.

# Anomaly Detection in Cybersecurity: General

Usecases

- Detecting anomalous log-on Patterns :
  - e.g. an employee appears to have traveled an impossible distance between log-on attempts or accessing the data that they have never used before
- Network Intrusion Detection
  - resorting to user behavior-based anomaly detection when detecting threats in large heterogeneous environments, especially covert threats is difficult - especially in an organization having complex global nature
- Abnormal Finance Activities Detection
  - user behavior-based anomaly detection to detect and react quickly financial frauds due to social manipulation and blackmail or for personal gain.
- Advanced Penetration Detection

# Anomaly Detection in Cybersecurity: General

Usecases

- Detecting anomalous log-on Patterns :
  - e.g. an employee appears to have traveled an impossible distance between log-on attempts or accessing the data that they have never used before
- Network Intrusion Detection
  - resorting to user behavior-based anomaly detection when detecting threats in large heterogeneous environments, especially covert threats is difficult - especially in an organization having complex global nature
- Abnormal Finance Activities Detection
  - user behavior-based anomaly detection to detect and react quickly financial frauds due to social manipulation and blackmail or for personal gain.
- Advanced Penetration Detection
  - detecting anomalous network traffic, unauthorized system access, unauthorized control of assets can be prevented by monitoring user logs/networks and tracking user behavior.

# Anomaly Detection: Isn't it a simple classification problem?

- At first sight, it appears that the problem is one of classification, i.e., separating data into two classes: anomalous and non-anomalous.

# Anomaly Detection: Isn't it a simple classification problem?

- At first sight, it appears that the problem is one of classification, i.e., separating data into two classes: anomalous and non-anomalous.
- However, is that really the case ? Then, why is anomaly detection studied as a head on its own ?

# Anomaly Detection: Isn't it a simple classification problem?

- At first sight, it appears that the problem is one of classification, i.e., separating data into two classes: anomalous and non-anomalous.
- However, is that really the case ? Then, why is anomaly detection studied as a head on its own ?
- It is tempting to address this problem using well-known ML classification algorithms, e.g., back-propagation neural networks, support vector machines, and decision trees.

# Anomaly Detection: Isn't it a simple classification problem?

- At first sight, it appears that the problem is one of classification, i.e., separating data into two classes: anomalous and non-anomalous.
- However, is that really the case ? Then, why is anomaly detection studied as a head on its own ?
- It is tempting to address this problem using well-known ML classification algorithms, e.g., back-propagation neural networks, support vector machines, and decision trees.
  - however, this will rarely be successful since there is a drastic imbalance between the two classes

# Anomaly Detection: Isn't it a simple classification problem?

- At first sight, it appears that the problem is one of classification, i.e., separating data into two classes: anomalous and non-anomalous.
- However, is that really the case ? Then, why is anomaly detection studied as a head on its own ?
- It is tempting to address this problem using well-known ML classification algorithms, e.g., back-propagation neural networks, support vector machines, and decision trees.
  - however, this will rarely be successful since there is a drastic imbalance between the two classes
  - anomalous data are much rarer than non-anomalous ones;

# Anomaly Detection: Isn't it a simple classification problem?

- At first sight, it appears that the problem is one of classification, i.e., separating data into two classes: anomalous and non-anomalous.
- However, is that really the case ? Then, why is anomaly detection studied as a head on its own ?
- It is tempting to address this problem using well-known ML classification algorithms, e.g., back-propagation neural networks, support vector machines, and decision trees.
    - however, this will rarely be successful since there is a drastic imbalance between the two classes
    - anomalous data are much rarer than non-anomalous ones;
    - thus, the results obtained will often result in too many false negatives (i.e., not recognizing anomalies).

# Anomaly Detection: Isn't it a simple classification problem?

- At first sight, it appears that the problem is one of classification, i.e., separating data into two classes: anomalous and non-anomalous.
- However, is that really the case ? Then, why is anomaly detection studied as a head on its own ?
- It is tempting to address this problem using well-known ML classification algorithms, e.g., back-propagation neural networks, support vector machines, and decision trees.
    - however, this will rarely be successful since there is a drastic imbalance between the two classes
    - anomalous data are much rarer than non-anomalous ones;
    - thus, the results obtained will often result in too many false negatives (i.e., not recognizing anomalies).
- other aspects.....*continued on next slide*

# Anomaly Detection: Isn't it a simple classification problem? ...

It is tempting to address this problem using well-known ML classification algorithms, e.g., back-propagation neural networks, support vector machines, and decision trees. However,.....

- Further, various anomalous cases may have very little in common.

It is tempting to address this problem using well-known ML classification algorithms, e.g., back-propagation neural networks, support vector machines, and decision trees. However,.....

- Further, various anomalous cases may have very little in common.
- Lastly, the occurrence of an anomaly may well be within the same bounds as those characterizing non-anomalous data

# Anomaly Detection: Isn't it a simple classification problem? ...

It is tempting to address this problem using well-known ML classification algorithms, e.g., back-propagation neural networks, support vector machines, and decision trees. However,.....

- Further, various anomalous cases may have very little in common.
- Lastly, the occurrence of an anomaly may well be within the same bounds as those characterizing non-anomalous data
  - and hence not distinguishable directly by attribute values, but may require analyses of their behavior with respect to subsets of neighbors or other data points

# Anomaly Detection: Isn't it a simple classification problem? ...

It is tempting to address this problem using well-known ML classification algorithms, e.g., back-propagation neural networks, support vector machines, and decision trees. However,.....

- Further, various anomalous cases may have very little in common.
- Lastly, the occurrence of an anomaly may well be within the same bounds as those characterizing non-anomalous data
  - and hence not distinguishable directly by attribute values, but may require analyses of their behavior with respect to subsets of neighbors or other data points
  - e.g., December 2016 using December 2015 and December 2014 in the earlier example of Retailer Sales analysis.

# Anomaly Detection: Isn't it a simple classification problem? ...

It is tempting to address this problem using well-known ML classification algorithms, e.g., back-propagation neural networks, support vector machines, and decision trees. However,.....

- Further, various anomalous cases may have very little in common.
- Lastly, the occurrence of an anomaly may well be within the same bounds as those characterizing non-anomalous data
  - and hence not distinguishable directly by attribute values, but may require analyses of their behavior with respect to subsets of neighbors or other data points
  - e.g., December 2016 using December 2015 and December 2014 in the earlier example of Retailer Sales analysis.
- Therefore, the need to develop carefully designed anomaly detection algorithms, along with an understanding of their applicability and limitations.

# Anomaly Detection in Cybersecurity: For Privacy

Examples of privacy leaks

- details of the bank accounts and medical records of celebrities leaked

# Anomaly Detection in Cybersecurity: For Privacy

Examples of privacy leaks

- details of the bank accounts and medical records of celebrities leaked
- selfies belonging to many celebrities shared publicly on the Internet - because owners' iCloud accounts have been hacked.

# Anomaly Detection in Cybersecurity: For Privacy

Examples of privacy leaks

- details of the bank accounts and medical records of celebrities leaked
- selfies belonging to many celebrities shared publicly on the Internet - because owners' iCloud accounts have been hacked.
- confidential data private to individuals (such as family details, and medical conditions) maybe non-financial but can eventually be used to gain unauthorized access to financial assets.

# Anomaly Detection in Cybersecurity: For Privacy

Examples of privacy leaks

- details of the bank accounts and medical records of celebrities leaked
- selfies belonging to many celebrities shared publicly on the Internet - because owners' iCloud accounts have been hacked.
- confidential data private to individuals (such as family details, and medical conditions) maybe non-financial but can eventually be used to gain unauthorized access to financial assets.
    - e.g. mother's maiden name, aadhar number, personal questions to be answered when forgetting password etc.

Examples of privacy leaks

- details of the bank accounts and medical records of celebrities leaked
- selfies belonging to many celebrities shared publicly on the Internet - because owners' iCloud accounts have been hacked.
- confidential data private to individuals (such as family details, and medical conditions) maybe non-financial but can eventually be used to gain unauthorized access to financial assets.
  - e.g. mother's maiden name, aadhar number, personal questions to be answered when forgetting password etc.
- Do the access control mechanisms not work in such cases ?

# Anomaly Detection in Cybersecurity: For Privacy

Examples of privacy leaks

- details of the bank accounts and medical records of celebrities leaked
- selfies belonging to many celebrities shared publicly on the Internet - because owners' iCloud accounts have been hacked.
- confidential data private to individuals (such as family details, and medical conditions) maybe non-financial but can eventually be used to gain unauthorized access to financial assets.
    - e.g. mother's maiden name, aadhar number, personal questions to be answered when forgetting password etc.
- Do the access control mechanisms not work in such cases ?
    - practically impossible to prevent inappropriate access (since individuals with access may unknowingly provide access to others),

# Anomaly Detection in Cybersecurity: For Privacy

Examples of privacy leaks

- details of the bank accounts and medical records of celebrities leaked
- selfies belonging to many celebrities shared publicly on the Internet - because owners' iCloud accounts have been hacked.
- confidential data private to individuals (such as family details, and medical conditions) maybe non-financial but can eventually be used to gain unauthorized access to financial assets.
    - e.g. mother's maiden name, aadhar number, personal questions to be answered when forgetting password etc.
- Do the access control mechanisms not work in such cases ?
    - practically impossible to prevent inappropriate access (since individuals with access may unknowingly provide access to others),
- anomaly detection algorithms can be used to monitor access to the data, and flag variations from the norm...

- the traditional signature detection based approach to detecting viruses and worms is not adequate

- the traditional signature detection based approach to detecting viruses and worms is not adequate
- focuses on after-the-fact recognition of their signature patterns and looking for such patterns within program code and data.

- the traditional signature detection based approach to detecting viruses and worms is not adequate
- focuses on after-the-fact recognition of their signature patterns and looking for such patterns within program code and data.
  - How to deal with the sudden influx of a new malware instance that does not match old malware signature patterns ?

- the traditional signature detection based approach to detecting viruses and worms is not adequate
- focuses on after-the-fact recognition of their signature patterns and looking for such patterns within program code and data.
  - How to deal with the sudden influx of a new malware instance that does not match old malware signature patterns ?
- an anomaly detection approach would instead monitor the appearance and behavior of malware to attempt to recognize variations from the norm is required.

Fraudulent Email Detection

- the net is periodically flooded with email messages that claim to originate from financial, information technology, or other service providers with whom individuals routinely conduct business.

Fraudulent Email Detection

- the net is periodically flooded with email messages that claim to originate from financial, information technology, or other service providers with whom individuals routinely conduct business.
- many individuals are lulled into clicking on apparently harmless websites, or providing their personal data to others, who misuse the data.

Fraudulent Email Detection

- the net is periodically flooded with email messages that claim to originate from financial, information technology, or other service providers with whom individuals routinely conduct business.
- many individuals are lulled into clicking on apparently harmless websites, or providing their personal data to others, who misuse the data.

Fraudulent Email Detection

- the net is periodically flooded with email messages that claim to originate from financial, information technology, or other service providers with whom individuals routinely conduct business.
- many individuals are lulled into clicking on apparently harmless websites, or providing their personal data to others, who misuse the data.

Anomaly detection algorithms can be used to guard against such attacks, at the individual level as well as by organizations protecting their users.

# Anomaly Detection in Finance: Credit Card Fraud

- Banks and credit card companies take many measures to facilitate the detection of potential anomalies in credit card usage,

# Anomaly Detection in Finance: Credit Card Fraud

- Banks and credit card companies take many measures to facilitate the detection of potential anomalies in credit card usage,
  - e.g., the use of a card in a location geographically distant from the specific user's normal usage region, or

# Anomaly Detection in Finance: Credit Card Fraud

- Banks and credit card companies take many measures to facilitate the detection of potential anomalies in credit card usage,
  - e.g., the use of a card in a location geographically distant from the specific user's normal usage region, or
  - to make purchases of unusual items such as electronic equipment using amounts of money that are unusual for that user.

# Anomaly Detection in Finance: Credit Card Fraud

- Banks and credit card companies take many measures to facilitate the detection of potential anomalies in credit card usage,
  - e.g., the use of a card in a location geographically distant from the specific user's normal usage region, or
  - to make purchases of unusual items such as electronic equipment using amounts of money that are unusual for that user.
- yet many early fraudulent usage instances go undetected, especially if they involve small amounts;

# Anomaly Detection in Finance: Credit Card Fraud

- Banks and credit card companies take many measures to facilitate the detection of potential anomalies in credit card usage,
  - e.g., the use of a card in a location geographically distant from the specific user's normal usage region, or
  - to make purchases of unusual items such as electronic equipment using amounts of money that are unusual for that user.
- yet many early fraudulent usage instances go undetected, especially if they involve small amounts;
- the small amounts may indicate a test, followed soon by a large fraudulent purchase using the same card.

# Anomaly Detection in Finance: Credit Card Fraud

- Banks and credit card companies take many measures to facilitate the detection of potential anomalies in credit card usage,
  - e.g., the use of a card in a location geographically distant from the specific user's normal usage region, or
  - to make purchases of unusual items such as electronic equipment using amounts of money that are unusual for that user.
- yet many early fraudulent usage instances go undetected, especially if they involve small amounts;
- the small amounts may indicate a test, followed soon by a large fraudulent purchase using the same card.
- Regular and periodic application of anomaly detection algorithms on recent purchase data would help prevent such problems to some extent.

# Anomaly Detection in Finance: Creditworthiness

- An important source of revenue for banks is the interest on the loans they make to individual consumers or businesses.

# Anomaly Detection in Finance: Creditworthiness

- An important source of revenue for banks is the interest on the loans they make to individual consumers or businesses.
- creditworthiness of an individual render the loans low risk for a bank.

# Anomaly Detection in Finance: Creditworthiness

- An important source of revenue for banks is the interest on the loans they make to individual consumers or businesses.
- creditworthiness of an individual render the loans low risk for a bank.
- useful if the risk of defaulting on a loan is substantially reduced at the time of making a loan;

# Anomaly Detection in Finance: Creditworthiness

- An important source of revenue for banks is the interest on the loans they make to individual consumers or businesses.
- creditworthiness of an individual render the loans low risk for a bank.
- useful if the risk of defaulting on a loan is substantially reduced at the time of making a loan;
- at the same time, qualified people should not be denied loans

# Anomaly Detection in Finance: Creditworthiness

- An important source of revenue for banks is the interest on the loans they make to individual consumers or businesses.
- creditworthiness of an individual render the loans low risk for a bank.
- useful if the risk of defaulting on a loan is substantially reduced at the time of making a loan;
- at the same time, qualified people should not be denied loans
- Accurate anomaly detection on the credit history and other datamfrom loan applicants is hence desirable.

# Anomaly Detection in Finance: Bankruptcy Prediction

- Risk is inherent in any entrepreneurial venture, and a significant number of companies file for bankruptcy every year

# Anomaly Detection in Finance: Bankruptcy Prediction

- Risk is inherent in any entrepreneurial venture, and a significant number of companies file for bankruptcy every year
- Detecting potential bankruptcy at an early stage would be very helpful to all the stakeholders.

# Anomaly Detection in Finance: Bankruptcy Prediction

- Risk is inherent in any entrepreneurial venture, and a significant number of companies file for bankruptcy every year
- Detecting potential bankruptcy at an early stage would be very helpful to all the stakeholders.
- Anomaly detection algorithms have been applied successfully to the task of analyzing company fundamentals (such as earnings) over time, to evaluate which companies are likely to go bankrupt.

# Anomaly Detection in Finance: Investing

- Stock prices fluctuate every day, and the investors' holy grail is to predict the future performance of a stock

# Anomaly Detection in Finance: Investing

- Stock prices fluctuate every day, and the investors' holy grail is to predict the future performance of a stock
- Accurate prediction is practically impossible, with unpredictable market shocks that affect stock prices.

# Anomaly Detection in Finance: Investing

- Stock prices fluctuate every day, and the investors' holy grail is to predict the future performance of a stock
- Accurate prediction is practically impossible, with unpredictable market shocks that affect stock prices.
- However, the performance of a stock over a short (recent) period of time can be compared with its prior performance, as well as the performance of other stocks of similar companies.

# Anomaly Detection in Finance: Investing

- Stock prices fluctuate every day, and the investors' holy grail is to predict the future performance of a stock
- Accurate prediction is practically impossible, with unpredictable market shocks that affect stock prices.
- However, the performance of a stock over a short (recent) period of time can be compared with its prior performance, as well as the performance of other stocks of similar companies.
- This can help identify anomalies that may signify that the company is outperforming or underperforming its competitors.

# Anomaly Detection in Finance: Investing

- Stock prices fluctuate every day, and the investors' holy grail is to predict the future performance of a stock
- Accurate prediction is practically impossible, with unpredictable market shocks that affect stock prices.
- However, the performance of a stock over a short (recent) period of time can be compared with its prior performance, as well as the performance of other stocks of similar companies.
- This can help identify anomalies that may signify that the company is outperforming or underperforming its competitors.
- Thus the application of anomaly detection algorithms can provide valuable information to potential and current investors in the company.

# Anomaly Detection in Healthcare: Diagnosis

- Practically every diagnostic effort is based on data that shows abnormalities in the patient's behavior or vitals (easily observable data characterizing the patient, e.g., blood pressure).

# Anomaly Detection in Healthcare: Diagnosis

- Practically every diagnostic effort is based on data that shows abnormalities in the patient's behavior or vitals (easily observable data characterizing the patient, e.g., blood pressure).
- In some cases, accurate analysis of quantitative data describing the patient is non-trivial, and would benefit from the application of anomaly detection algorithms.

# Anomaly Detection in Healthcare: Diagnosis

- Practically every diagnostic effort is based on data that shows abnormalities in the patient's behavior or vitals (easily observable data characterizing the patient, e.g., blood pressure).
- In some cases, accurate analysis of quantitative data describing the patient is non-trivial, and would benefit from the application of anomaly detection algorithms.
- A few examples are described below.

# Anomaly Detection in Healthcare: Diagnosis

- Practically every diagnostic effort is based on data that shows abnormalities in the patient's behavior or vitals (easily observable data characterizing the patient, e.g., blood pressure).
- In some cases, accurate analysis of quantitative data describing the patient is non-trivial, and would benefit from the application of anomaly detection algorithms.
- A few examples are described below.
  - Automatic detection of PVC from EKG data would significantly reduce the workload for cardiologists, and potentially increase the detection rate of PVC. Detection of Arrhythmias could potentially be performed more efficiently by anomaly detection algorithms.

# Anomaly Detection in Healthcare: Diagnosis

- Practically every diagnostic effort is based on data that shows abnormalities in the patient's behavior or vitals (easily observable data characterizing the patient, e.g., blood pressure).
- In some cases, accurate analysis of quantitative data describing the patient is non-trivial, and would benefit from the application of anomaly detection algorithms.
- A few examples are described below.
    - Automatic detection of PVC from EKG data would significantly reduce the workload for cardiologists, and potentially increase the detection rate of PVC. Detection of Arrhythmias could potentially be performed more efficiently by anomaly detection algorithms.
    - A similar issue is that of identifying possible evidence of epileptic seizures in patients, by examining electro-encephalogram (EEG) data for abnormal variations

# Anomaly Detection in Healthcare: Diagnosis

- Practically every diagnostic effort is based on data that shows abnormalities in the patient's behavior or vitals (easily observable data characterizing the patient, e.g., blood pressure).
- In some cases, accurate analysis of quantitative data describing the patient is non-trivial, and would benefit from the application of anomaly detection algorithms.
- A few examples are described below.
  - Automatic detection of PVC from EKG data would significantly reduce the workload for cardiologists, and potentially increase the detection rate of PVC. Detection of Arrhythmias could potentially be performed more efficiently by anomaly detection algorithms.
  - A similar issue is that of identifying possible evidence of epileptic seizures in patients, by examining electro-encephalogram (EEG) data for abnormal variations
  - Cancer Diagnosis: The classification of tumors as benign vs. malignant, from radiographic image data, has long been known to be a task that is particularly challenging because of the relatively small number of malignant cases.

# Anomaly Detection in Healthcare: Patient Monitoring

Patient Monitoring

- For patients being treated for some serious health disorders, it is very important to monitor progress and vital signs constantly, in order to determine the occurrence of unexpected side effects of medications or surgery

# Anomaly Detection in Healthcare: Patient Monitoring

Patient Monitoring

- For patients being treated for some serious health disorders, it is very important to monitor progress and vital signs constantly, in order to determine the occurrence of unexpected side effects of medications or surgery
- Unfortunately, due to the large number of patients in a hospital, signs of an abnormality are sometimes accidentally ignored.

# Anomaly Detection in Healthcare: Patient Monitoring

Patient Monitoring

- For patients being treated for some serious health disorders, it is very important to monitor progress and vital signs constantly, in order to determine the occurrence of unexpected side effects of medications or surgery
- Unfortunately, due to the large number of patients in a hospital, signs of an abnormality are sometimes accidentally ignored.
- Similarly, elderly and disabled individuals occasionally suffer from falls in their private residences, without receiving immediate care.

# Anomaly Detection in Healthcare: Patient Monitoring

Patient Monitoring

- For patients being treated for some serious health disorders, it is very important to monitor progress and vital signs constantly, in order to determine the occurrence of unexpected side effects of medications or surgery
- Unfortunately, due to the large number of patients in a hospital, signs of an abnormality are sometimes accidentally ignored.
- Similarly, elderly and disabled individuals occasionally suffer from falls in their private residences, without receiving immediate care.
- The application of anomaly detection algorithms to alert care providers, based on appropriate sensors, is hence essential.

# Anomaly Detection in Healthcare: Radiology

Radiology

- the field often involves searching for unusual data in X-ray, NMR, and other images.

# Anomaly Detection in Healthcare: Radiology

Radiology

- the field often involves searching for unusual data in X-ray, NMR, and other images.
- Anomaly detection algorithms can potentially assist in finding early phase tumor, facilitating early detection of cancer.

Epidemiology

- Viruses and bacteria evolve at a fast pace

Epidemiology

- Viruses and bacteria evolve at a fast pace
- hence, understanding them is vital for at least temporary success in the arms race between medications and pathogens.

# Anomaly Detection in Healthcare: Epidemiology

Epidemiology

- Viruses and bacteria evolve at a fast pace
- hence, understanding them is vital for at least temporary success in the arms race between medications and pathogens.
- the fields of genetics, proteomics, and metabolomics can be assisted by anomaly detection algorithms that may search for unusual mutations

Epidemiology

- Viruses and bacteria evolve at a fast pace
- hence, understanding them is vital for at least temporary success in the arms race between medications and pathogens.
- the fields of genetics, proteomics, and metabolomics can be assisted by anomaly detection algorithms that may search for unusual mutations
- these can signal specific diseases.

# Anomaly Detection in Healthcare: Epidemiology

Epidemiology

- Viruses and bacteria evolve at a fast pace
- hence, understanding them is vital for at least temporary success in the arms race between medications and pathogens.
- the fields of genetics, proteomics, and metabolomics can be assisted by anomaly detection algorithms that may search for unusual mutations
- these can signal specific diseases.
- it would also be beneficial to identifying points in time at which epidemiological data reveals the emergence of a drug-resistant mutation of the responsible pathogen.

Epidemiology

- Viruses and bacteria evolve at a fast pace
- hence, understanding them is vital for at least temporary success in the arms race between medications and pathogens.
- the fields of genetics, proteomics, and metabolomics can be assisted by anomaly detection algorithms that may search for unusual mutations
- these can signal specific diseases.
- it would also be beneficial to identifying points in time at which epidemiological data reveals the emergence of a drug-resistant mutation of the responsible pathogen.
- some such data can also be obtained from individual patients, whose response to a medication may follow an unusual path,

# Anomaly Detection in Healthcare: Epidemiology

Epidemiology

- Viruses and bacteria evolve at a fast pace
- hence, understanding them is vital for at least temporary success in the arms race between medications and pathogens.
- the fields of genetics, proteomics, and metabolomics can be assisted by anomaly detection algorithms that may search for unusual mutations
- these can signal specific diseases.
- it would also be beneficial to identifying points in time at which epidemiological data reveals the emergence of a drug-resistant mutation of the responsible pathogen.
- some such data can also be obtained from individual patients, whose response to a medication may follow an unusual path,
    - e.g., first appearing to improve, then degrading rapidly.

# Anomaly Detection in Healthcare: Epidemiology

Epidemiology

- Viruses and bacteria evolve at a fast pace
- hence, understanding them is vital for at least temporary success in the arms race between medications and pathogens.
- the fields of genetics, proteomics, and metabolomics can be assisted by anomaly detection algorithms that may search for unusual mutations
- these can signal specific diseases.
- it would also be beneficial to identifying points in time at which epidemiological data reveals the emergence of a drug-resistant mutation of the responsible pathogen.
- some such data can also be obtained from individual patients, whose response to a medication may follow an unusual path,
  - e.g., first appearing to improve, then degrading rapidly.
  - can be used to prevent the epidemic spread of a new drug-resistant pathogen.

Detecting unusual behaviors of people in public places

- such behaviour may indicate intent towards planned violent activities - may be observed using video-cameras with limited regions of surveillance, installed in multiple public areas, and monitored by security personnel.

Detecting unusual behaviors of people in public places

- such behaviour may indicate intent towards planned violent activities - may be observed using video-cameras with limited regions of surveillance, installed in multiple public areas, and monitored by security personnel.
- based on the frequent observed behaviors of people, normal patterns may be established with respect to how people tend to move...

Detecting unusual behaviors of people in public places

- such behaviour may indicate intent towards planned violent activities - may be observed using video-cameras with limited regions of surveillance, installed in multiple public areas, and monitored by security personnel.
- based on the frequent observed behaviors of people, normal patterns may be established with respect to how people tend to move...
    - variations from these patterns represent anomalies which can be detected and investigated.

# Anomaly Detection in Defense and Internal Security: Personnel Behaviour

Detecting unusual behaviors of people in public places

- such behaviour may indicate intent towards planned violent activities - may be observed using video-cameras with limited regions of surveillance, installed in multiple public areas, and monitored by security personnel.
- based on the frequent observed behaviors of people, normal patterns may be established with respect to how people tend to move...
  - variations from these patterns represent anomalies which can be detected and investigated.
- However, the application ofautomated anomaly detection algorithms to such video data is what is required and is a non-trivial task

Detecting unusual behaviors of people in public places

- such behaviour may indicate intent towards planned violent activities - may be observed using video-cameras with limited regions of surveillance, installed in multiple public areas, and monitored by security personnel.
- based on the frequent observed behaviors of people, normal patterns may be established with respect to how people tend to move...
  - variations from these patterns represent anomalies which can be detected and investigated.
- However, the application ofautomated anomaly detection algorithms to such video data is what is required and is a non-trivial task
- difficult to establish a clear narrative regarding each individual subject's behavior over an extended period of time.

Detecting unusual behaviors of people in public places

- such behaviour may indicate intent towards planned violent activities - may be observed using video-cameras with limited regions of surveillance, installed in multiple public areas, and monitored by security personnel.
- based on the frequent observed behaviors of people, normal patterns may be established with respect to how people tend to move...
  - variations from these patterns represent anomalies which can be detected and investigated.
- However, the application ofautomated anomaly detection algorithms to such video data is what is required and is a non-trivial task
- difficult to establish a clear narrative regarding each individual subject's behavior over an extended period of time.
  - there is also a potential for many false positive reports, since anomalous behavior may be symptomatic of illness or confusion on the part of the subjects, rather than violent intent.

Detecting unusual behaviors of people in public places

- such behaviour may indicate intent towards planned violent activities - may be observed using video-cameras with limited regions of surveillance, installed in multiple public areas, and monitored by security personnel.
- based on the frequent observed behaviors of people, normal patterns may be established with respect to how people tend to move...
  - variations from these patterns represent anomalies which can be detected and investigated.
- However, the application ofautomated anomaly detection algorithms to such video data is what is required and is a non-trivial task
- difficult to establish a clear narrative regarding each individual subject's behavior over an extended period of time.
  - there is also a potential for many false positive reports, since anomalous behavior may be symptomatic of illness or confusion on the part of the subjects, rather than violent intent.
  - in addition, individuals with violent intent may take extra care to appear 'normal' until an actual violent event begins to be carried out.

## Battlefield Behaviors

- efforts are made to infer the tactical intent of an opposing party from observable behaviors.

Battlefield Behaviors

- efforts are made to infer the tactical intent of an opposing party from observable behaviors.
- several classical battles have been won or lost when one party conducts unusual movements of forces which are not foreseen by the other party

Battlefield Behaviors

- efforts are made to infer the tactical intent of an opposing party from observable behaviors.
- several classical battles have been won or lost when one party conducts unusual movements of forces which are not foreseen by the other party
- Models would be constructed by each party regarding possible intentions of the opposing party, evaluating the most likely courses of action that may be taken.

Battlefield Behaviors

- efforts are made to infer the tactical intent of an opposing party from observable behaviors.
- several classical battles have been won or lost when one party conducts unusual movements of forces which are not foreseen by the other party
- Models would be constructed by each party regarding possible intentions of the opposing party, evaluating the most likely courses of action that may be taken.
- variations from these expectations should lead to questioning the model, and evaluation of alternative models to see if they would fit better with observed data.

Battlefield Behaviors

- efforts are made to infer the tactical intent of an opposing party from observable behaviors.
- several classical battles have been won or lost when one party conducts unusual movements of forces which are not foreseen by the other party
- Models would be constructed by each party regarding possible intentions of the opposing party, evaluating the most likely courses of action that may be taken.
- variations from these expectations should lead to questioning the model, and evaluation of alternative models to see if they would fit better with observed data.
- if none of the current models fit well with the data, closer monitoring and readiness to take actions would be required.

- Although there is broad agreement around the world on avoiding chemical and nuclear warfare, there have been some occasions where these expectations have been violated by individual governments or factions.

# Anomaly Detection in Defense & Internal Security: Unconventional Attacks

- Although there is broad agreement around the world on avoiding chemical and nuclear warfare, there have been some occasions where these expectations have been violated by individual governments or factions.
- the potential for their use hence cannot be ruled out.

- Although there is broad agreement around the world on avoiding chemical and nuclear warfare, there have been some occasions where these expectations have been violated by individual governments or factions.
- the potential for their use hence cannot be ruled out.
- this requires monitoring of signs indicating the accumulation of such agents as well as their use or experimentation...e.g.

- Although there is broad agreement around the world on avoiding chemical and nuclear warfare, there have been some occasions where these expectations have been violated by individual governments or factions.
- the potential for their use hence cannot be ruled out.
- this requires monitoring of signs indicating the accumulation of such agents as well as their use or experimentation...e.g.
    - byproducts of chemical reactions or radioactive decomposition.

- Although there is broad agreement around the world on avoiding chemical and nuclear warfare, there have been some occasions where these expectations have been violated by individual governments or factions.
- the potential for their use hence cannot be ruled out.
- this requires monitoring of signs indicating the accumulation of such agents as well as their use or experimentation...e.g.
    - byproducts of chemical reactions or radioactive decomposition.
    - data from chemical monitoring stations

- Although there is broad agreement around the world on avoiding chemical and nuclear warfare, there have been some occasions where these expectations have been violated by individual governments or factions.
- the potential for their use hence cannot be ruled out.
- this requires monitoring of signs indicating the accumulation of such agents as well as their use or experimentation...e.g.
  - byproducts of chemical reactions or radioactive decomposition.
  - data from chemical monitoring stations
  - data from satellite images

# Anomaly Detection in Defense & Internal Security: Unconventional Attacks

- Although there is broad agreement around the world on avoiding chemical and nuclear warfare, there have been some occasions where these expectations have been violated by individual governments or factions.
- the potential for their use hence cannot be ruled out.
- this requires monitoring of signs indicating the accumulation of such agents as well as their use or experimentation...e.g.
  - byproducts of chemical reactions or radioactive decomposition.
  - data from chemical monitoring stations
  - data from satellite images
  - evidence of unusual activities, such as the movements of individual scientists or high level military officers to unexpected locations.

- Although there is broad agreement around the world on avoiding chemical and nuclear warfare, there have been some occasions where these expectations have been violated by individual governments or factions.
- the potential for their use hence cannot be ruled out.
- this requires monitoring of signs indicating the accumulation of such agents as well as their use or experimentation...e.g.
  - byproducts of chemical reactions or radioactive decomposition.
  - data from chemical monitoring stations
  - data from satellite images
  - evidence of unusual activities, such as the movements of individual scientists or high level military officers to unexpected locations.
- these are all anomalies with respect to the usual behaviour.

Consumer Home Safety Anomaly detection algorithms

- assisted by sensors that are becoming ubiquitous in the home, can help with problems encountered in many households. e.g.

Consumer Home Safety Anomaly detection algorithms

- assisted by sensors that are becoming ubiquitous in the home, can help with problems encountered in many households. e.g.
  - occurrence of falls and other Problems for the disabled and senior citizens living alone or in environments with very little day-to-day human contact and direct monitoring. Some simple technologies, e.g., using accelerometers embedded in smart phones, can help detect some such situations. Anomaly detection algorithms that trigger alarms (with medical personnel or relatives) only when the sensed data indicates abnormal behavior, varying significantly from characteristics of data collected over a prior period of time during which no falls are known to occur.

# Anomaly Detection in Consumer Home Safety: Detecting Occurrence of Falls etc.

Consumer Home Safety Anomaly detection algorithms

- assisted by sensors that are becoming ubiquitous in the home, can help with problems encountered in many households. e.g.
    - occurrence of falls and other Problems for the disabled and senior citizens living alone or in environments with very little day-to-day human contact and direct monitoring. Some simple technologies, e.g., using accelerometers embedded in smart phones, can help detect some such situations. Anomaly detection algorithms that trigger alarms (with medical personnel or relatives) only when the sensed data indicates abnormal behavior, varying significantly from characteristics of data collected over a prior period of time during which no falls are known to occur.
    - the approach depends critically on the accuracy of the anomaly detection algorithms: too many false positives would be annoying since they result in excessive caregiver resource consumption and desensitization to true alarms.

Consumer Home Safety Anomaly detection algorithms

- assisted by sensors that are becoming ubiquitous in the home, can help with problems encountered in many households. e.g.
  - occurrence of falls and other Problems for the disabled and senior citizens living alone or in environments with very little day-to-day human contact and direct monitoring. Some simple technologies, e.g., using accelerometers embedded in smart phones, can help detect some such situations. Anomaly detection algorithms that trigger alarms (with medical personnel or relatives) only when the sensed data indicates abnormal behavior, varying significantly from characteristics of data collected over a prior period of time during which no falls are known to occur.
  - the approach depends critically on the accuracy of the anomaly detection algorithms: too many false positives would be annoying since they result in excessive caregiver resource consumption and desensitization to true alarms.
    - on the other hand, even a single false negative (undetected fall) would degrade confidence in the usefulness of the system, negating its utility.

- Signs of break-ins and burglaries are now routinely detected by sensors placed near windows and doors.

- Signs of break-ins and burglaries are now routinely detected by sensors placed near windows and doors.
- detection and handling of false alarms needs critical deployment.

# Anomaly Detection in Consumer Home Safety: Home Perimeter Safety

- Signs of break-ins and burglaries are now routinely detected by sensors placed near windows and doors.
- detection and handling of false alarms needs critical deployment.
- hence, the application of pattern recognition and anomaly detection algorithms, with data collected from “normal” and “abnormal” behavior that may be simulated experimentally.

- Signs of break-ins and burglaries are now routinely detected by sensors placed near windows and doors.
- detection and handling of false alarms needs critical deployment.
- hence, the application of pattern recognition and anomaly detection algorithms, with data collected from "normal" and "abnormal" behavior that may be simulated experimentally.
- additionally, in regions where the incidence of crime is relatively high, periodic monitoring of local traffic and pedestrian behaviors may reveal "normal" patterns of behavior within a specific neighborhood.

- Signs of break-ins and burglaries are now routinely detected by sensors placed near windows and doors.
- detection and handling of false alarms needs critical deployment.
- hence, the application of pattern recognition and anomaly detection algorithms, with data collected from "normal" and "abnormal" behavior that may be simulated experimentally.
- additionally, in regions where the incidence of crime is relatively high, periodic monitoring of local traffic and pedestrian behaviors may reveal "normal" patterns of behavior within a specific neighborhood.
- variations from which may trigger additional monitoring or deployment of law enforcement resources.

- many sensors currently exist to monitor the concentrations of carbon-monoxide and volatile organic compounds.

- many sensors currently exist to monitor the concentrations of carbon-monoxide and volatile organic compounds.
- in large cities, it is also important to monitor the levels of particulate air pollutants and sulfur-di-oxide.

- many sensors currently exist to monitor the concentrations of carbon-monoxide and volatile organic compounds.
- in large cities, it is also important to monitor the levels of particulate air pollutants and sulfur-di-oxide.
- the effective use of such sensors would be in collusion with anomaly detection algorithms that can sense potential problems before they actually occur, e.g.,

- many sensors currently exist to monitor the concentrations of carbon-monoxide and volatile organic compounds.
- in large cities, it is also important to monitor the levels of particulate air pollutants and sulfur-di-oxide.
- the effective use of such sensors would be in collusion with anomaly detection algorithms that can sense potential problems before they actually occur, e.g.,
  - using information from external sensors, weather data (e.g., wind velocity), and variations in pollutant density gradients, along with data relevant to "normal" conditions.

# Anomaly Detection in Manufacturing & Industry: Quality Control

- Statistical change detection algorithms have been used for a long time for quality control in manufacturing organizations, triggering alarms when sampled output characteristics of a product fall below expected quality constraints. In general, fluctuations in the underlying process may be detected by drastic changes in specific sensor data measurements. These may be considered to be simple anomaly detection algorithms. In addition, anomaly detection can be applied to data from multiple sensors located at various points in the monitored manufacturing environment. In addition to identifying problems after they have occurred, anomalies may be detected in unusual patterns of various sensor data, indicating possible locations in the manufacturing environment where faults or failures have occurred. For example, if the "normal" behavior of two adjacent sensors (possibly measuring different attributes or features of the system or its products) over time involves a linear relationship, with previously measured gradients, then a significant variation in this relationship may be detected as anomalous, triggering further investigation.

- Many retail organizations have to constantly monitor their revenues and earnings, to facilitate planning as well as to identify any potential disasters at an early stage. This involves constructing time series of sales data, and analyzing fluctuations in sales, comparing them to prior values, while factoring in various trends and relationships to periodic cycles and external events. Anomaly detection algorithms can play a useful role in this context, helping to separate insignificant fluctuations (noise) from potentially meaningful variations with significant implications for the future plans of the organization.

- Many retail and other organizations maintain inventories of various products and raw materials, and their effective management is an important factor influencing profitability, especially when demand fluctuates drastically over time. Difficulties arise when an organization does not have a product available for sale when there is a sudden surge in demand; conversely, maintaining extra inventory (in anticipation of a demand surge) can be expensive. Finally, some organizations are plagued by occasional occurrence of theft by employees or outsiders, which may only be detectable by effective inventory management and monitoring. Anomaly detection algorithms can play a role in this context, e.g., by enabling formulation of mathematical models or case data that describe "normal" behavior of inventory data (collected over time), and triggering warnings when such expectations are significantly violated.

- Most organizations are constantly striving to determine how best to allocate resources to maximize profit or revenue, by understanding customer behaviors. Anomaly detection algorithms can be considered to be one aspect of data mining efforts in this context. Past data can be analyzed to model customers' typical purchasing behaviors, and analyzing the subclasses of behaviors wherein purchasing increases or decreases with time, perhaps as a result of a change in store configuration or website. The application of anomaly detection algorithms can enable detecting variations from such models, triggering investigations into probable causes and possible remedies. In addition to purchasing, customer behaviors may also be relevant to identify potentially unacceptable actions or deception, e.g., in applications such as money laundering. Anomaly detection algorithms can then pursue three directions: comparing an individual's behavior to his own past behavior, or to the behavior of others in the group or category to which he belongs, or to the behavior of the entire collection of customers. Such monitoring may also reveal fraud being perpetrated by other individuals, who may have gained unauthorized access to a customer's account.

- Many organizations performing sensitive actions need to be extremely sensitive to the possible damage caused by a few errant employees who may have access to organizational resources in the course of normal performance of their everyday jobs. Indeed, some of the largest known fraudulent financial manipulations have been identified as occurring due to the unusual activities of a small number of individual employees, which could be detected by the passive monitoring of their actions using anomaly detection algorithms. External agencies can also apply anomaly detection algorithms to determine whether fraud is being perpetrated by the principals of an organization. For example, a famous Ponzi scheme could have been detected early by investigators if they had analyzed the promises of high guaranteed returns made by the organization to investors; such promises are well outside the norms of other stockbrokers and hedge funds, including even the most successful members of this group. The promise of high returns along with substantial opacity in the investing philosophy of the organization should have triggered warning bells. In addition to financial organizations, retail stores must monitor their employees to maintain productivity and integrity; this may again be assisted by anomaly detection algorithms.

# Anomaly Detection in Science

- The application of anomaly detection algorithms is ubiquitous in science. Indeed, according to one perspective, progress in science occurs due to paradigmatic revolutions caused by the discovery of anomalous data that contradict well-established models [78]. We consider below a few examples of anomaly detection in everyday scientific practice, rather than revolutions. The SETI project involves large scale efforts utilizing thousands of computers that have been launched to analyze electromagnetic data received by the earth, searching for anomalies that may indicate possible transmission of meaningful signals by intelligent extra-terrestrials. More successful have been efforts applied in the search for planets and stars with unusual behavior (compared to most other objects), revealing the existence of planets whose temperature and composition enables the occurrence of liquid water, hence presumed to be hospitable to life similar to that on earth. More routinely, the not-so-remote skies are periodically scanned with telescopes to discover any unusual objects that do not fall into the categories of known objects such as satellites; such monitoring is conducted to evaluate potential threats from other nations as well as natural objects in space that may be heading towards the earth. Even when they do not approach the earth, large natural

# Anomaly Detection: Topics

- What is anomaly detection? Understanding Anomaly Detection
- Anomaly Detection Metrics
- Dealing with the kind of Data? Old Problems vs New Problems
- Outliers in One-Dimensional Data & in Multi-Dimensional Data
- Overview of Anomaly Detection Approaches & Evaluation Criteria
  - Distance-based Anomaly Detection
  - Clustering-based Anomaly Detection
  - Model-based Anomaly Detection
- Anomaly Detection Algorithms
  - Distance & Density-based Approaches
  - Rank-based Approaches
  - Ensemble Methods
  - Algorithms for Time-series Data
- Research Paper Study1: Deep Learning Based Anomaly Detection for Multi-dimensional Time-series data
- Research Paper Study2: Anomaly Detection for Spam Filtering

# *Critical Aspects in Understanding Anomaly Detection*

# Formulation of Anomaly Detection Algorithms: Critical Questions

Critical questions relevant to the formulation of anomaly detection algorithms:

1. How is the norm characterized?

Critical questions relevant to the formulation of anomaly detection algorithms:

1. How is the norm characterized?
   - What if there are multiple and substantially different cases, all of which can be considered to be normal?

# Formulation of Anomaly Detection Algorithms: Critical Questions

Critical questions relevant to the formulation of anomaly detection algorithms:

1. How is the norm characterized?
   - What if there are multiple and substantially different cases, all of which can be considered to be normal?
   - What is considered to be a substantial variation, as opposed to a minor variation from a norm?

# Formulation of Anomaly Detection Algorithms: Critical Questions

Critical questions relevant to the formulation of anomaly detection algorithms:

1. How is the norm characterized?
   - What if there are multiple and substantially different cases, all of which can be considered to be normal?
   - What is considered to be a substantial variation, as opposed to a minor variation from a norm?
2. What are the metrics used for evaluation of the algorithms for anomaly detection?

# Formulation of Anomaly Detection Algorithms: Critical Questions

Critical questions relevant to the formulation of anomaly detection algorithms:

1. How is the norm characterized?
   - What if there are multiple and substantially different cases, all of which can be considered to be normal?
   - What is considered to be a substantial variation, as opposed to a minor variation from a norm?
2. What are the metrics used for evaluation of the algorithms for anomaly detection?
3. What kind of Data is used? Old Problems vs New Problems vis-á-vis Signature based detection OR anomaly based detection.

# Formulation of Anomaly Detection Algorithms: Critical Questions

Critical questions relevant to the formulation of anomaly detection algorithms:

1. How is the norm characterized?
   - What if there are multiple and substantially different cases, all of which can be considered to be normal?
   - What is considered to be a substantial variation, as opposed to a minor variation from a norm?
2. What are the metrics used for evaluation of the algorithms for anomaly detection?
3. What kind of Data is used? Old Problems vs New Problems vis-á-vis Signature based detection OR anomaly based detection.
4. How are Outliers handled in One-Dimensional Data ?

# Formulation of Anomaly Detection Algorithms: Critical Questions

Critical questions relevant to the formulation of anomaly detection algorithms:

1. How is the norm characterized?
   - What if there are multiple and substantially different cases, all of which can be considered to be normal?
   - What is considered to be a substantial variation, as opposed to a minor variation from a norm?
2. What are the metrics used for evaluation of the algorithms for anomaly detection?
3. What kind of Data is used? Old Problems vs New Problems vis-á-vis Signature based detection OR anomaly based detection.
4. How are Outliers handled in One-Dimensional Data ?
5. How are Outliers handled in MultiDimensional Data ?How do we address multi-attribute data?

# Formulation of Anomaly Detection Algorithms: Critical Questions

Critical questions relevant to the formulation of anomaly detection algorithms:

1. How is the norm characterized?
    - What if there are multiple and substantially different cases, all of which can be considered to be normal?
    - What is considered to be a substantial variation, as opposed to a minor variation from a norm?
2. What are the metrics used for evaluation of the algorithms for anomaly detection?
3. What kind of Data is used? Old Problems vs New Problems vis-á-vis Signature based detection OR anomaly based detection.
4. How are Outliers handled in One-Dimensional Data ?
5. How are Outliers handled in MultiDimensional Data ?How do we address multi-attribute data?
6. How do we address changes that happen over time?

# Formulation of Anomaly Detection Algorithms: Critical Questions

Critical questions relevant to the formulation of anomaly detection algorithms:

1. How is the norm characterized?
   - What if there are multiple and substantially different cases, all of which can be considered to be normal?
   - What is considered to be a substantial variation, as opposed to a minor variation from a norm?
2. What are the metrics used for evaluation of the algorithms for anomaly detection?
3. What kind of Data is used? Old Problems vs New Problems vis-á-vis Signature based detection OR anomaly based detection.
4. How are Outliers handled in One-Dimensional Data ?
5. How are Outliers handled in MultiDimensional Data ?How do we address multi-attribute data?
6. How do we address changes that happen over time?

# Formulation of Anomaly Detection Algorithms: Critical Questions

Critical questions relevant to the formulation of anomaly detection algorithms:

1. How is the norm characterized?
   - What if there are multiple and substantially different cases, all of which can be considered to be normal?
   - What is considered to be a substantial variation, as opposed to a minor variation from a norm?
2. What are the metrics used for evaluation of the algorithms for anomaly detection?
3. What kind of Data is used? Old Problems vs New Problems vis-á-vis Signature based detection OR anomaly based detection.
4. How are Outliers handled in One-Dimensional Data ?
5. How are Outliers handled in MultiDimensional Data ?How do we address multi-attribute data?
6. How do we address changes that happen over time?

These questions are first discussed in the following before discussing the anomaly detection approaches/algorithms.

# How is the norm characterized ?

Issues in Anomaly Detection

- As seen earlier, for the most real-life systems 100% correct detection is impossible.

# How is the norm characterized ?

Issues in Anomaly Detection

- As seen earlier, for the most real-life systems 100% correct detection is impossible.
- Hence, it is critical to devise mechanisms to minimize both false positives and false negatives.

# How is the norm characterized ?

Issues in Anomaly Detection

- As seen earlier, for the most real-life systems 100% correct detection is impossible.
- Hence, it is critical to devise mechanisms to minimize both false positives and false negatives.
  - therefore, rather than classifying any given point as whether it is anomalous or not, it is better to estimate the likelihood that it is an anomaly.

# How is the norm characterized ?

Issues in Anomaly Detection

- As seen earlier, for the most real-life systems 100% correct detection is impossible.
- Hence, it is critical to devise mechanisms to minimize both false positives and false negatives.
  - therefore, rather than classifying any given point as whether it is anomalous or not, it is better to estimate the likelihood that it is an anomaly.
- Another frequently used perspective is - to evaluate the relative anomalousness of different points.

# How is the norm characterized ?

Issues in Anomaly Detection

- As seen earlier, for the most real-life systems 100% correct detection is impossible.
- Hence, it is critical to devise mechanisms to minimize both false positives and false negatives.
    - therefore, rather than classifying any given point as whether it is anomalous or not, it is better to estimate the likelihood that it is an anomaly.
- Another frequently used perspective is - to evaluate the relative anomalousness of different points.
    - a possible scenario is to employ an algorithm to identify ten data points, rank-ordered, as being the most likely anomalous cases.

# How is the norm characterized ?

Issues in Anomaly Detection

- As seen earlier, for the most real-life systems 100% correct detection is impossible.
- Hence, it is critical to devise mechanisms to minimize both false positives and false negatives.
  - therefore, rather than classifying any given point as whether it is anomalous or not, it is better to estimate the likelihood that it is an anomaly.
- Another frequently used perspective is - to evaluate the relative anomalousness of different points.
  - a possible scenario is to employ an algorithm to identify ten data points, rank-ordered, as being the most likely anomalous cases.
- anomaly detection algorithms must account for the fact that the processes of interest are often neither deterministic nor completely random.

# How is the norm characterized ?

Issues in Anomaly Detection

- As seen earlier, for the most real-life systems 100% correct detection is impossible.
- Hence, it is critical to devise mechanisms to minimize both false positives and false negatives.
  - therefore, rather than classifying any given point as whether it is anomalous or not, it is better to estimate the likelihood that it is an anomaly.
- Another frequently used perspective is - to evaluate the relative anomalousness of different points.
  - a possible scenario is to employ an algorithm to identify ten data points, rank-ordered, as being the most likely anomalous cases.
- anomaly detection algorithms must account for the fact that the processes of interest are often neither deterministic nor completely random.
  - this is more so, in the cybersecurity applications - the observable behaviors are the result of the deliberate (non-random) actions of humans - not predictable.

# How is the norm characterized ?...

- Emphasizing that the anomalies or outliers are substantial variations from the norm, when an anomaly detection algorithm is applied, three possible cases need to be considered:

# How is the norm characterized ?...

- Emphasizing that the anomalies or outliers are substantial variations from the norm, when an anomaly detection algorithm is applied, three possible cases need to be considered:
  - Correct Detection: Detected abnormalities in data do correspond exactly to abnormalities in the process.

# How is the norm characterized ?...

- Emphasizing that the anomalies or outliers are substantial variations from the norm, when an anomaly detection algorithm is applied, three possible cases need to be considered:
  - Correct Detection: Detected abnormalities in data do correspond exactly to abnormalities in the process.
  - False Positives: The process continues to be normal, but unexpected data values are observed, e.g., due to intrinsic system noise.

# How is the norm characterized ?...

- Emphasizing that the anomalies or outliers are substantial variations from the norm, when an anomaly detection algorithm is applied, three possible cases need to be considered:
  - Correct Detection: Detected abnormalities in data do correspond exactly to abnormalities in the process.
  - False Positives: The process continues to be normal, but unexpected data values are observed, e.g., due to intrinsic system noise.
  - False Negatives: The process becomes abnormal, but the consequences are not registered in the abnormal data, e.g., due to the signal of the abnormality being insufficiently strong compared to the noise in the system.

# How is the norm characterized ?...

- Emphasizing that the anomalies or outliers are substantial variations from the norm, when an anomaly detection algorithm is applied, three possible cases need to be considered:
  - Correct Detection: Detected abnormalities in data do correspond exactly to abnormalities in the process.
  - False Positives: The process continues to be normal, but unexpected data values are observed, e.g., due to intrinsic system noise.
  - False Negatives: The process becomes abnormal, but the consequences are not registered in the abnormal data, e.g., due to the signal of the abnormality being insufficiently strong compared to the noise in the system.
- Thus, it is significant to understand the metrics of evaluation in anomaly detection.

# *Anomaly Detection: Metrics*

# Anomaly Detection Metrics: Precision, Recall & Rankpower

To evaluate the performance of the algorithms, three metrics are often used:

- Given a dataset $D$, suppose an outlier detection algorithm identifies $m > 0$ potential anomalies, of which $m_t$ ($\leq$m) are known to be true outliers. Then,

# Anomaly Detection Metrics: Precision, Recall & Rankpower

To evaluate the performance of the algorithms, three metrics are often used:

- Given a dataset $D$, suppose an outlier detection algorithm identifies $m > 0$ potential anomalies, of which $m_t$ ($\leq$m) are known to be true outliers. Then,
    - Precision: Precision, measures the proportion of true outliers in top $m$ suspicious instances, is:

$$P_r = m_t / m$$

      and equals 1.0 if all the points identified by the algorithm are true outliers.

## Anomaly Detection Metrics: Precision, Recall & Rankpower

To evaluate the performance of the algorithms, three metrics are often used:

- Given a dataset $D$, suppose an outlier detection algorithm identifies $m > 0$ potential anomalies, of which $m_t$ ($\leq$m) are known to be true outliers. Then,
  - Precision: Precision, measures the proportion of true outliers in top $m$ suspicious instances, is:

$$P_r = m_t / m$$

    and equals 1.0 if all the points identified by the algorithm are true outliers.
- If $D$, contains $d_t$ ($\geq m_t$) true outliers, then another important measure is Recall.

To evaluate the performance of the algorithms, three metrics are often used:

- Given a dataset $D$, suppose an outlier detection algorithm identifies $m > 0$ potential anomalies, of which $m_t$ ($\leq$m) are known to be true outliers. Then,
  - Precision: Precision, measures the proportion of true outliers in top $m$ suspicious instances, is:

$$P_r = m_t/m$$

    and equals 1.0 if all the points identified by the algorithm are true outliers.
- If $D$, contains $d_t$ ($\geq m_t$) true outliers, then another important measure is Recall.
  - Recall is defined as

$$R_e = m_t/d_t$$

    which equal 1.0 if all true outliers are discovered by the algorithm

# Anomaly Detection Metrics: Precision, Recall & Rankpower

To evaluate the performance of the algorithms, three metrics are often used:

- Given a dataset $D$, suppose an outlier detection algorithm identifies $m > 0$ potential anomalies, of which $m_t$ ($\leq$m) are known to be true outliers. Then,
    - Precision: Precision, measures the proportion of true outliers in top $m$ suspicious instances, is:
    $$P_r = m_t/m$$
    and equals 1.0 if all the points identified by the algorithm are true outliers.
- If $D$, contains $d_t$ ($\geq m_t$) true outliers, then another important measure is Recall.
    - Recall is defined as
    $$R_e = m_t/d_t$$
    which equal 1.0 if all true outliers are discovered by the algorithm
- e.g. for the credit card transaction fraud example, if the data set contains 1,000,000 transactions of which 200 are fraudulent, an algorithm which considers all data to be anomalous exhibits $P_r = 0.0002$ and $R_e = 1.0$,

# Anomaly Detection Metrics: Precision, Recall & Rankpower

To evaluate the performance of the algorithms, three metrics are often used:

- Given a dataset $D$, suppose an outlier detection algorithm identifies $m > 0$ potential anomalies, of which $m_t$ ($\leq$m) are known to be true outliers. Then,
  - Precision: Precision, measures the proportion of true outliers in top $m$ suspicious instances, is:
    $$P_r = m_t/m$$
    and equals 1.0 if all the points identified by the algorithm are true outliers.
- If $D$, contains $d_t$ ($\geq m_t$) true outliers, then another important measure is Recall.
  - Recall is defined as
    $$R_e = m_t/d_t$$
    which equal 1.0 if all true outliers are discovered by the algorithm
- e.g. for the credit card transaction fraud example, if the data set contains 1,000,000 transactions of which 200 are fraudulent, an algorithm which considers all data to be anomalous exhibits $P_r = 0.0002$ and $R_e = 1.0$,
  - whereas an algorithm with 8 true positives (anomalies) and 2 false positives exhibits $P_r = 0.8$ and $R_e = 10/200 = 0.05$

# Anomaly Detection Metrics: Precision, Recall & Rankpower...

To evaluate the performance of the algorithms, three metrics are often used:

- Precision and Recall are insufficient to capture completely the effectiveness of an algorithm, especially when comparing algorithms that result in different numbers of anomalies.

To evaluate the performance of the algorithms, three metrics are often used:

- Precision and Recall are insufficient to capture completely the effectiveness of an algorithm, especially when comparing algorithms that result in different numbers of anomalies.
- In particular, precision can take a small value just because m is large.

# Anomaly Detection Metrics: Precision, Recall & Rankpower...

To evaluate the performance of the algorithms, three metrics are often used:

- Precision and Recall are insufficient to capture completely the effectiveness of an algorithm, especially when comparing algorithms that result in different numbers of anomalies.
- In particular, precision can take a small value just because m is large.
- The values for the precision and recall remain the same even in case of opposing inferences from two algorithms.

# Anomaly Detection Metrics: Precision, Recall & Rankpower...

To evaluate the performance of the algorithms, three metrics are often used:

- Precision and Recall are insufficient to capture completely the effectiveness of an algorithm, especially when comparing algorithms that result in different numbers of anomalies.
- In particular, precision can take a small value just because m is large.
- The values for the precision and recall remain the same even in case of opposing inferences from two algorithms.
- When can an algorithm be considered more effective ?

# Anomaly Detection Metrics: Precision, Recall & Rankpower...

To evaluate the performance of the algorithms, three metrics are often used:

- Precision and Recall are insufficient to capture completely the effectiveness of an algorithm, especially when comparing algorithms that result in different numbers of anomalies.
- In particular, precision can take a small value just because m is large.
- The values for the precision and recall remain the same even in case of opposing inferences from two algorithms.
- When can an algorithm be considered more effective ?
  - if the true outliers occupy top positions and non-outliers are among the least suspicious instances.

# Anomaly Detection Metrics: Precision, Recall & Rankpower...

To evaluate the performance of the algorithms, three metrics are often used:

- Precision and Recall are insufficient to capture completely the effectiveness of an algorithm, especially when comparing algorithms that result in different numbers of anomalies.
- In particular, precision can take a small value just because m is large.
- The values for the precision and recall remain the same even in case of opposing inferences from two algorithms.
- When can an algorithm be considered more effective ?
    - if the true outliers occupy top positions and non-outliers are among the least suspicious instances.
- The RankPower metric captures this notion.

To evaluate the performance of the algorithms, three metrics are often used:

- Precision and Recall are insufficient to capture completely the effectiveness of an algorithm, especially when comparing algorithms that result in different numbers of anomalies.
- In particular, precision can take a small value just because m is large.
- The values for the precision and recall remain the same even in case of opposing inferences from two algorithms.
- When can an algorithm be considered more effective ?
  - if the true outliers occupy top positions and non-outliers are among the least suspicious instances.
- The RankPower metric captures this notion.
- Formally, if $R_i$ denotes the rank of the $i^{th}$ true outlier in the sorted list of most suspicious objects, then the RankPower is given by

$$RP = \frac{m_t(m_t + 1)}{2\Sigma_{i=1}^{m_t} R_i} \tag{1}$$

# Anomaly Detection Metrics: Precision, Recall & Rankpower...

Consider a dataset of size $n = 50$ that contains exactly 5 anomalies. Suppose that an anomaly detection algorithm identifies $m = 10$ data points as anomalous, of which $m_t = 4$ are true anomalies. In addition, let the true anomalies in occupy ranks equal to 1, 4, 5, and 8 in the sorted list of truly anomalous data points. Then, calculate precision, recall and rankpower?

To be calculated on board......

# How is the norm characterized ?...

- Statisticians have long used the notions of (arithmetic) mean, median, and mode to capture the norms associated with distributions.

# How is the norm characterized ?...

- Statisticians have long used the notions of (arithmetic) mean, median, and mode to capture the norms associated with distributions.
- Each of these is a single scalar or multidimensional vector

# How is the norm characterized ?...

- Statisticians have long used the notions of (arithmetic) mean, median, and mode to capture the norms associated with distributions.
- Each of these is a single scalar or multidimensional vector
- the distance of a point from the mean (or median or mode) has been used to assess the degree to which the point is abnormal.

# How is the norm characterized ?...

- Statisticians have long used the notions of (arithmetic) mean, median, and mode to capture the norms associated with distributions.
- Each of these is a single scalar or multidimensional vector
- the distance of a point from the mean (or median or mode) has been used to assess the degree to which the point is abnormal.
- But what distance ?

# How is the norm characterized ?...

- Statisticians have long used the notions of (arithmetic) mean, median, and mode to capture the norms associated with distributions.
- Each of these is a single scalar or multidimensional vector
- the distance of a point from the mean (or median or mode) has been used to assess the degree to which the point is abnormal.
- But what distance ?
- if the data distribution is normal, the Mahalanobis distance is used to get the estimate of the *abnormality* - the standard deviation along each dimension serves as a normalizing factor

# How is the norm characterized ?...

- Statisticians have long used the notions of (arithmetic) mean, median, and mode to capture the norms associated with distributions.
- Each of these is a single scalar or multidimensional vector
- the distance of a point from the mean (or median or mode) has been used to assess the degree to which the point is abnormal.
- But what distance ?
- if the data distribution is normal, the Mahalanobis distance is used to get the estimate of the *abnormality* - the standard deviation along each dimension serves as a normalizing factor
- Mahalanobis distance

# How is the norm characterized ?...

- Statisticians have long used the notions of (arithmetic) mean, median, and mode to capture the norms associated with distributions.
- Each of these is a single scalar or multidimensional vector
- the distance of a point from the mean (or median or mode) has been used to assess the degree to which the point is abnormal.
- But what distance ?
- if the data distribution is normal, the Mahalanobis distance is used to get the estimate of the *abnormality* - the standard deviation along each dimension serves as a normalizing factor
- Mahalanobis distance
  - measure has an important property: it is the square root of the log likelihood of a point belonging to the distribution, i.e., an indication of how "anomalous" is the point relative to the data.

# How is the norm characterized ?...

- Statisticians have long used the notions of (arithmetic) mean, median, and mode to capture the norms associated with distributions.
- Each of these is a single scalar or multidimensional vector
- the distance of a point from the mean (or median or mode) has been used to assess the degree to which the point is abnormal.
- But what distance ?
- if the data distribution is normal, the Mahalanobis distance is used to get the estimate of the *abnormality* - the standard deviation along each dimension serves as a normalizing factor
- Mahalanobis distance
  - measure has an important property: it is the square root of the log likelihood of a point belonging to the distribution, i.e., an indication of how "anomalous" is the point relative to the data.
- But does this straightforward interpretation apply with non-normal (and multi-modal) distributions ?

# How is the norm characterized ?...

With non-normal (and multi-modal) distributions,

- the "norm" is described as a set of points rather than a single point.

# How is the norm characterized ?...

With non-normal (and multi-modal) distributions,

- the "norm" is described as a set of points rather than a single point.
- For example, the norm may consist of

# How is the norm characterized ?...

With non-normal (and multi-modal) distributions,

- the "norm" is described as a set of points rather than a single point.
- For example, the norm may consist of
  - a collection of cluster centroids or alternatively the boundaries of clusters.

# How is the norm characterized ?...

With non-normal (and multi-modal) distributions,

- the "norm" is described as a set of points rather than a single point.
- For example, the norm may consist of
  - a collection of cluster centroids or alternatively the boundaries of clusters.
  - abnormal points must be substantially distant from each of the points in the norm.

# How is the norm characterized ?...

With non-normal (and multi-modal) distributions,

- the "norm" is described as a set of points rather than a single point.
- For example, the norm may consist of
  - a collection of cluster centroids or alternatively the boundaries of clusters.
  - abnormal points must be substantially distant from each of the points in the norm.
- Also, if the data set is characterized by variations in density over clusters in space, such variations must also be accounted for in determining whether a point is abnormal.

# How is the norm characterized ?...

With non-normal (and multi-modal) distributions,

- the "norm" is described as a set of points rather than a single point.
- For example, the norm may consist of
  - a collection of cluster centroids or alternatively the boundaries of clusters.
  - abnormal points must be substantially distant from each of the points in the norm.
- Also, if the data set is characterized by variations in density over clusters in space, such variations must also be accounted for in determining whether a point is abnormal.
- For instance,

# How is the norm characterized ?...

With non-normal (and multi-modal) distributions,

- the "norm" is described as a set of points rather than a single point.
- For example, the norm may consist of
  - a collection of cluster centroids or alternatively the boundaries of clusters.
  - abnormal points must be substantially distant from each of the points in the norm.
- Also, if the data set is characterized by variations in density over clusters in space, such variations must also be accounted for in determining whether a point is abnormal.
- For instance,
  - a larger distance may characterize an anomaly near a less dense cluster, whereas a smaller distance is reasonable for an anomaly near a more dense cluster.

# How is the norm characterized ?...

With non-normal (and multi-modal) distributions,

- the "norm" is described as a set of points rather than a single point.
- For example, the norm may consist of
  - a collection of cluster centroids or alternatively the boundaries of clusters.
  - abnormal points must be substantially distant from each of the points in the norm.
- Also, if the data set is characterized by variations in density over clusters in space, such variations must also be accounted for in determining whether a point is abnormal.
- For instance,
  - a larger distance may characterize an anomaly near a less dense cluster, whereas a smaller distance is reasonable for an anomaly near a more dense cluster.
- Local density, i.e., the number of data points in a unit (hyper)volume, then turns out to be a critical notion in identifying which points are more anomalous.

# *What Kind of Data? Old Problems vs New Problems*

# Old problems vis-á-vis New Problems & the kind of data

Signature-based detection

- Anomalies previously encountered in data, e.g., due to past attacks which were identified and not forgotten

Signature-based detection

- Anomalies previously encountered in data, e.g., due to past attacks which were identified and not forgotten
- Prior data analysis may then have identified signatures of patterns associated with such anomalies

# Old problems vis-á-vis New Problems & the kind of data

Signature-based detection

- Anomalies previously encountered in data, e.g., due to past attacks which were identified and not forgotten
- Prior data analysis may then have identified signatures of patterns associated with such anomalies
  - e.g. many viruses have been cataloged based on the effects they produce, or the occurrences of certain code fragments within the viruses

# Old problems vis-á-vis New Problems & the kind of data

Signature-based detection

- Anomalies previously encountered in data, e.g., due to past attacks which were identified and not forgotten
- Prior data analysis may then have identified signatures of patterns associated with such anomalies
  - e.g. many viruses have been cataloged based on the effects they produce, or the occurrences of certain code fragments within the viruses
- anti-malware software routinely applies the rules formulated to detect such occurrences and isolate potential malware.

# Old problems vis-á-vis New Problems & the kind of data

Signature-based detection

- Anomalies previously encountered in data, e.g., due to past attacks which were identified and not forgotten
- Prior data analysis may then have identified signatures of patterns associated with such anomalies
  - e.g. many viruses have been cataloged based on the effects they produce, or the occurrences of certain code fragments within the viruses
- anti-malware software routinely applies the rules formulated to detect such occurrences and isolate potential malware.
- thus the principal task is classification i.e. classification of data belonging to the "safe" category, and that belonging to each known malware category

Signature-based detection

- Anomalies previously encountered in data, e.g., due to past attacks which were identified and not forgotten
- Prior data analysis may then have identified signatures of patterns associated with such anomalies
  - e.g. many viruses have been cataloged based on the effects they produce, or the occurrences of certain code fragments within the viruses
- anti-malware software routinely applies the rules formulated to detect such occurrences and isolate potential malware.
- thus the principal task is classification i.e. classification of data belonging to the "safe" category, and that belonging to each known malware category
- Learning algorithms (such as Support Vector Machines and backpropagation-trained Neural Networks) have been used to develop models that enable analysts to perform this classification task.

# Old problems vis-á-vis New Problems & the kind of data

Signature-based detection

- Anomalies previously encountered in data, e.g., due to past attacks which were identified and not forgotten
- Prior data analysis may then have identified signatures of patterns associated with such anomalies
    - e.g. many viruses have been cataloged based on the effects they produce, or the occurrences of certain code fragments within the viruses
- anti-malware software routinely applies the rules formulated to detect such occurrences and isolate potential malware.
- thus the principal task is classification i.e. classification of data belonging to the "safe" category, and that belonging to each known malware category
- Learning algorithms (such as Support Vector Machines and backpropagation-trained Neural Networks) have been used to develop models that enable analysts to perform this classification task.
- What is the limitation of such approaches ?

# Old problems vis-á-vis New Problems & the kind of data...

What is the limitation of the signature-based/classification-oriented approaches ?

- can work only in detecting known problems of specific kinds,

# Old problems vis-á-vis New Problems & the kind of data...

What is the limitation of the signature-based/classification-oriented approaches ?

- can work only in detecting known problems of specific kinds,
- whereas the greatest damage in cyber-security applications is caused by unknown problems newly created by bad actors.

# Old problems vis-á-vis New Problems & the kind of data...

What is the limitation of the signature-based/classification-oriented approaches ?

- can work only in detecting known problems of specific kinds,
- whereas the greatest damage in cyber-security applications is caused by unknown problems newly created by bad actors.
- there exist problems whose manifestations in the data do not admit simple categorization; hence amelioratory actions cannot be performed rapidly enough to prevent a catastrophe

# Old problems vis-á-vis New Problems & the kind of data...

What is the limitation of the signature-based/classification-oriented approaches ?

- can work only in detecting known problems of specific kinds,
- whereas the greatest damage in cyber-security applications is caused by unknown problems newly created by bad actors.
- there exist problems whose manifestations in the data do not admit simple categorization; hence ameliorary actions cannot be performed rapidly enough to prevent a catastrophe
  - e.g., when an employee gives his password to someone over the telephone, and the security of a sensitive system is compromised.

# Old problems vis-á-vis New Problems & the kind of data...

What is the limitation of the signature-based/classification-oriented approaches ?

- can work only in detecting known problems of specific kinds,
- whereas the greatest damage in cyber-security applications is caused by unknown problems newly created by bad actors.
- there exist problems whose manifestations in the data do not admit simple categorization; hence amelioratory actions cannot be performed rapidly enough to prevent a catastrophe
  - e.g., when an employee gives his password to someone over the telephone, and the security of a sensitive system is compromised.
- in such cases, anomaly detection procedures and algorithms are called for.

# Old problems vis-á-vis New Problems & the kind of data...

What is the limitation of the signature-based/classification-oriented approaches ?

- can work only in detecting known problems of specific kinds,
- whereas the greatest damage in cyber-security applications is caused by unknown problems newly created by bad actors.
- there exist problems whose manifestations in the data do not admit simple categorization; hence amelioratory actions cannot be performed rapidly enough to prevent a catastrophe
  - e.g., when an employee gives his password to someone over the telephone, and the security of a sensitive system is compromised.
- in such cases, anomaly detection procedures and algorithms are called for.
- the basic assumption still is that : the observable data do reflect anomalies in the underlying process or behavior.

# Old problems vis-á-vis New Problems & the kind of data...

What is the limitation of the signature-based/classification-oriented approaches ?

- can work only in detecting known problems of specific kinds,
- whereas the greatest damage in cyber-security applications is caused by unknown problems newly created by bad actors.
- there exist problems whose manifestations in the data do not admit simple categorization; hence amelioratory actions cannot be performed rapidly enough to prevent a catastrophe
    - e.g., when an employee gives his password to someone over the telephone, and the security of a sensitive system is compromised.
- in such cases, anomaly detection procedures and algorithms are called for.
- the basic assumption still is that : the observable data do reflect anomalies in the underlying process or behavior.
- however the corresponding changes in the data are not expected to follow previously known patterns.

# Old problems vis-á-vis New Problems & the kind of data...

What is the limitation of the signature-based/classification-oriented approaches ?

- can work only in detecting known problems of specific kinds,
- whereas the greatest damage in cyber-security applications is caused by unknown problems newly created by bad actors.
- there exist problems whose manifestations in the data do not admit simple categorization; hence amelioratory actions cannot be performed rapidly enough to prevent a catastrophe
  - e.g., when an employee gives his password to someone over the telephone, and the security of a sensitive system is compromised.
- in such cases, anomaly detection procedures and algorithms are called for.
- the basic assumption still is that : the observable data do reflect anomalies in the underlying process or behavior.
- however the corresponding changes in the data are not expected to follow previously known patterns.
- hence, one may not have an explicit model, pattern or rule that describes the anomalies.

# Old problems vis-á-vis New Problems & the kind of data...

What is the limitation of the signature-based/classification-oriented approaches ?

- can work only in detecting known problems of specific kinds,

# Old problems vis-á-vis New Problems & the kind of data...

What is the limitation of the signature-based/classification-oriented approaches ?

- can work only in detecting known problems of specific kinds,
- whereas the greatest damage in cyber-security applications is caused by unknown problems newly created by bad actors.

# Old problems vis-á-vis New Problems & the kind of data...

What is the limitation of the signature-based/classification-oriented approaches ?

- can work only in detecting known problems of specific kinds,
- whereas the greatest damage in cyber-security applications is caused by unknown problems newly created by bad actors.
- there exist problems whose manifestations in the data do not admit simple categorization; hence amelioratory actions cannot be performed rapidly enough to prevent a catastrophe

What is the limitation of the signature-based/classification-oriented approaches ?

- can work only in detecting known problems of specific kinds,
- whereas the greatest damage in cyber-security applications is caused by unknown problems newly created by bad actors.
- there exist problems whose manifestations in the data do not admit simple categorization; hence amelioratory actions cannot be performed rapidly enough to prevent a catastrophe
  - e.g., when an employee gives his password to someone over the telephone, and the security of a sensitive system is compromised.

# Old problems vis-á-vis New Problems & the kind of data...

What is the limitation of the signature-based/classification-oriented approaches ?

- can work only in detecting known problems of specific kinds,
- whereas the greatest damage in cyber-security applications is caused by unknown problems newly created by bad actors.
- there exist problems whose manifestations in the data do not admit simple categorization; hence amelioratory actions cannot be performed rapidly enough to prevent a catastrophe
  - e.g., when an employee gives his password to someone over the telephone, and the security of a sensitive system is compromised.
- in such cases, anomaly detection procedures and algorithms are called for.

# Old problems vis-á-vis New Problems & the kind of data...

What is the limitation of the signature-based/classification-oriented approaches ?

- can work only in detecting known problems of specific kinds,
- whereas the greatest damage in cyber-security applications is caused by unknown problems newly created by bad actors.
- there exist problems whose manifestations in the data do not admit simple categorization; hence amelioratory actions cannot be performed rapidly enough to prevent a catastrophe
  - e.g., when an employee gives his password to someone over the telephone, and the security of a sensitive system is compromised.
- in such cases, anomaly detection procedures and algorithms are called for.
- the basic assumption still is that : the observable data do reflect anomalies in the underlying process or behavior.

# Old problems vis-á-vis New Problems & the kind of data...

What is the limitation of the signature-based/classification-oriented approaches ?

- can work only in detecting known problems of specific kinds,
- whereas the greatest damage in cyber-security applications is caused by unknown problems newly created by bad actors.
- there exist problems whose manifestations in the data do not admit simple categorization; hence amelioratory actions cannot be performed rapidly enough to prevent a catastrophe
    - e.g., when an employee gives his password to someone over the telephone, and the security of a sensitive system is compromised.
- in such cases, anomaly detection procedures and algorithms are called for.
- the basic assumption still is that : the observable data do reflect anomalies in the underlying process or behavior.
- however the corresponding changes in the data are not expected to follow previously known patterns.

# Old problems vis-á-vis New Problems & the kind of data...

What is the limitation of the signature-based/classification-oriented approaches ?

- can work only in detecting known problems of specific kinds,
- whereas the greatest damage in cyber-security applications is caused by unknown problems newly created by bad actors.
- there exist problems whose manifestations in the data do not admit simple categorization; hence amelioratory actions cannot be performed rapidly enough to prevent a catastrophe
    - e.g., when an employee gives his password to someone over the telephone, and the security of a sensitive system is compromised.
- in such cases, anomaly detection procedures and algorithms are called for.
- the basic assumption still is that : the observable data do reflect anomalies in the underlying process or behavior.
- however the corresponding changes in the data are not expected to follow previously known patterns.
- hence, one may not have an explicit model, pattern or rule that describes the anomalies.

# *Anomaly Detection Metrics: Numerical Problems*

# Anomaly Detection Metrics: Numerical Problem#1

Example: The "threads" currently running on a computer's processors may be observable. Some threads represent normal activity for a given time of day and for given users. But other threads may provide evidence that an unexpected computer program is currently executing, the "process" that resulted in the specific observable "data" (set of threads). Then, answer the following questions:

1. What are the *data points* in this case ?

# Anomaly Detection Metrics: Numerical Problem#1

Example: The "threads" currently running on a computer's processors may be observable. Some threads represent normal activity for a given time of day and for given users. But other threads may provide evidence that an unexpected computer program is currently executing, the "process" that resulted in the specific observable "data" (set of threads). Then, answer the following questions:

1. What are the *data points* in this case ?
2. What is an anomaly in this case ?

# Anomaly Detection Metrics: Numerical Problem#1

Example: The "threads" currently running on a computer's processors may be observable. Some threads represent normal activity for a given time of day and for given users. But other threads may provide evidence that an unexpected computer program is currently executing, the "process" that resulted in the specific observable "data" (set of threads). Then, answer the following questions:

1. What are the *data points* in this case ?
2. What is an anomaly in this case ?
3. What is the dimensionality of space in this case ? Especially, what set-size do the the *normal* cases may correspond to, in this example?

# Anomaly Detection Metrics: Numerical Problem#1

Example: The "threads" currently running on a computer's processors may be observable. Some threads represent normal activity for a given time of day and for given users. But other threads may provide evidence that an unexpected computer program is currently executing, the "process" that resulted in the specific observable "data" (set of threads). Then, answer the following questions:

1. What are the *data points* in this case ?
2. What is an anomaly in this case ?
3. What is the dimensionality of space in this case ? Especially, what set-size do the the *normal* cases may correspond to, in this example?

# Anomaly Detection Metrics: Numerical Problem#1

Example: The "threads" currently running on a computer's processors may be observable. Some threads represent normal activity for a given time of day and for given users. But other threads may provide evidence that an unexpected computer program is currently executing, the "process" that resulted in the specific observable "data" (set of threads). Then, answer the following questions:

1. What are the *data points* in this case ?
2. What is an anomaly in this case ?
3. What is the dimensionality of space in this case ? Especially, what set-size do the the *normal* cases may correspond to, in this example?

Further analysis to identify the anomalous cases here may be based on :

- The precise sequence in which certain processes are executed - may be important to signify malware.
- The time sequence of Internet nodes over which a message is routed, signifying whether it is anomalous.
- Behaviors of individuals: a single snapshot may not indicate anything wrong, but a series of observations of the same individual (over time) may indicate variations from the norm.

# *Anomaly Detection: Types of Approaches*

# Anomaly Detection Approaches: Types

The primary approaches for anomaly detection can be characterized as follows:

- Distance-based: Points that are farther from others are considered more anomalous.

# Anomaly Detection Approaches: Types

The primary approaches for anomaly detection can be characterized as follows:

- Distance-based: Points that are farther from others are considered more anomalous.
- Density-based: Points that are in relatively low density regions are considered more anomalous.

# Anomaly Detection Approaches: Types

The primary approaches for anomaly detection can be characterized as follows:

- Distance-based: Points that are farther from others are considered more anomalous.
- Density-based: Points that are in relatively low density regions are considered more anomalous.
- Rank-based:: The most anomalous points are those whose nearest neighbors have others as nearest neighbors.

# Anomaly Detection Approaches: Types

The primary approaches for anomaly detection can be characterized as follows:

- Distance-based: Points that are farther from others are considered more anomalous.
- Density-based: Points that are in relatively low density regions are considered more anomalous.
- Rank-based:: The most anomalous points are those whose nearest neighbors have others as nearest neighbors.

# Anomaly Detection Approaches: Types

The primary approaches for anomaly detection can be characterized as follows:

- Distance-based: Points that are farther from others are considered more anomalous.
- Density-based: Points that are in relatively low density regions are considered more anomalous.
- Rank-based:: The most anomalous points are those whose nearest neighbors have others as nearest neighbors.

For each of these approaches, the nature of the data may be supervised, semi-supervised, or unsupervised, as is usual.

# *Outliers in One-Dimensional Data*

# Outliers in One-Dimensional Data

- How do we deal with quantitative data i.e. how do we analyze quantitative data ?

| Location="Delhi" | | | | |
|---|---|---|---|---|
| Time (quarter) | item (type) | | | |
| | Egg | Milk | Bread | Biscuit |
| Q1 | 260 | 508 | 15 | 60 |
| Q2 | 390 | 256 | 20 | 90 |
| Q3 | 436 | 396 | 50 | 40 |
| Q4 | 528 | 483 | 35 | 50 |

Figure: 2-D Data

# Outliers in One-Dimensional Data

- How do we deal with quantitative data i.e. how do we analyze quantitative data ?
- Compute the statistics i.e. the parameters described earlier for the distribution of the quantitative data.

| Location="Delhi" | | | | |
|---|---|---|---|---|
| | item (type) | | | |
| Time (quarter) | Egg | Milk | Bread | Biscuit |
| Q1 | 260 | 508 | 15 | 60 |
| Q2 | 390 | 256 | 20 | 90 |
| Q3 | 436 | 396 | 50 | 40 |
| Q4 | 528 | 483 | 35 | 50 |

Figure: 2-D Data

# Outliers in One-Dimensional Data

- How do we deal with quantitative data i.e. how do we analyze quantitative data ?
- Compute the statistics i.e. the parameters described earlier for the distribution of the quantitative data.
- Which parameters?

| Location="Delhi" | | | | |
|---|---|---|---|---|
| Time (quarter) | item (type) | | | |
| | Egg | Milk | Bread | Biscuit |
| Q1 | 260 | 508 | 15 | 60 |
| Q2 | 390 | 256 | 20 | 90 |
| Q3 | 436 | 396 | 50 | 40 |
| Q4 | 528 | 483 | 35 | 50 |

Figure: 2-D Data

- How do we deal with quantitative data i.e. how do we analyze quantitative data ?
- Compute the statistics i.e. the parameters described earlier for the distribution of the quantitative data.
- Which parameters?
  - Remember again the IQ Test Score example.....

| Location="Delhi" | | | | |
|---|---|---|---|---|
| Time (quarter) | item (type) | | | |
| | Egg | Milk | Bread | Biscuit |
| Q1 | 260 | 508 | 15 | 60 |
| Q2 | 390 | 256 | 20 | 90 |
| Q3 | 436 | 396 | 50 | 40 |
| Q4 | 528 | 483 | 35 | 50 |

Figure: 2-D Data

# Outliers in One-Dimensional Data

But the quantitative data may have at least two cases/varieties :

- One, where each data point is a single number.

# Outliers in One-Dimensional Data

But the quantitative data may have at least two cases/varieties :

- One, where each data point is a single number.
- Other, where each data point represents multiple attributes

# Outliers in One-Dimensional Data

But the quantitative data may have at least two cases/varieties :

- One, where each data point is a single number.
- Other, where each data point represents multiple attributes
  - Multidimensional data is a data set with many different columns, also called features or attributes.

# Outliers in One-Dimensional Data

But the quantitative data may have at least two cases/varieties :

- One, where each data point is a single number.
- Other, where each data point represents multiple attributes
  - Multidimensional data is a data set with many different columns, also called features or attributes.
  - The more columns in the data set, the more likely one is to discover hidden insights.

# Outliers in One-Dimensional Data

But the quantitative data may have at least two cases/varieties :

- One, where each data point is a single number.
- Other, where each data point represents multiple attributes
  - Multidimensional data is a data set with many different columns, also called features or attributes.
  - The more columns in the data set, the more likely one is to discover hidden insights.

# Outliers in One-Dimensional Data

But the quantitative data may have at least two cases/varieties :

- One, where each data point is a single number.
- Other, where each data point represents multiple attributes
  - Multidimensional data is a data set with many different columns, also called features or attributes.
  - The more columns in the data set, the more likely one is to discover hidden insights.

| Location="Delhi" | | | | |
|---|---|---|---|---|
| | item (type) | | | |
| Time (quarter) | Egg | Milk | Bread | Biscuit |
| Q1 | 260 | 508 | 15 | 60 |
| Q2 | 390 | 256 | 20 | 90 |
| Q3 | 436 | 396 | 50 | 40 |
| Q4 | 528 | 483 | 35 | 50 |

Figure: 2-D Data

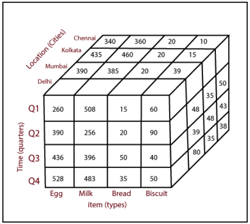


Figure: 3-D Data

# Distributions in Single-dimensional quantitative data

Following are the possible data distributions in the single-dimensional data:

- Uniform Distribution:

# Distributions in Single-dimensional quantitative data

Following are the possible data distributions in the single-dimensional data:

- Uniform Distribution:
- Normal Distribution:

# Distributions in Single-dimensional quantitative data

Following are the possible data distributions in the single-dimensional data:

- Uniform Distribution:
- Normal Distribution:
- Other Unimodal Distributions :

# Examples of normal/Gaussian distribution of data in real life

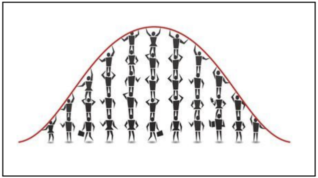

Figure: Height of a person

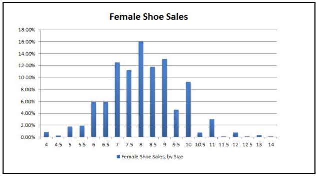


Figure: FemaleShoeSizes

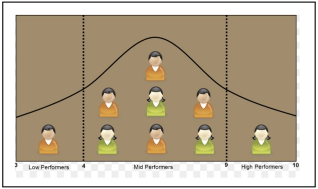


Figure: Performance in Exam [1]

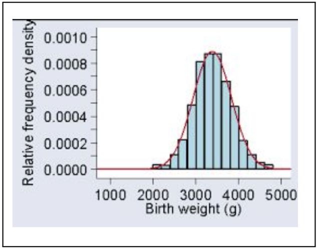


Figure: Weigth of a newborn

[1] https://studiousguy.com/real-life-examples-normal-distribution/

# Examples of Uniform distribution of data in real life

- The normal distribution is used to model phenomenon that tend to follow a "bell-curve" shape. e.g. Weight of a baby, Shoe-size, Height of a person etc.

# Examples of Uniform distribution of data in real life

- The normal distribution is used to model phenomenon that tend to follow a "bell-curve" shape. e.g. Weight of a baby, Shoe-size, Height of a person etc.
- Conversely, the uniform distribution is used to model scenarios where each potential outcome is equally likely.

# Examples of Uniform distribution of data in real life

- The normal distribution is used to model phenomenon that tend to follow a "bell-curve" shape. e.g. Weight of a baby, Shoe-size, Height of a person etc.
- Conversely, the uniform distribution is used to model scenarios where each potential outcome is equally likely.
  - e.g. rolling a die. If one rolls a die one time, the probability that it falls on a number between 1 and 6 follows a uniform distribution because each number is equally likely to occur - the probability that one rolls a 1 is $1/6$, one rolls a 2 is $1/6$, one rolls a 3 is $1/6$....OR

# Examples of Uniform distribution of data in real life

- The normal distribution is used to model phenomenon that tend to follow a "bell-curve" shape. e.g. Weight of a baby, Shoe-size, Height of a person etc.
- Conversely, the uniform distribution is used to model scenarios where each potential outcome is equally likely.
  - e.g. rolling a die. If one rolls a die one time, the probability that it falls on a number between 1 and 6 follows a uniform distribution because each number is equally likely to occur - the probability that one rolls a 1 is $1/6$, one rolls a 2 is $1/6$, one rolls a 3 is $1/6$....OR
  - if one walks up to a random person on the street, the probability that their birthday falls on a given date would follow a uniform distribution because each day of the year is equally likely to be their birthday OR

# Examples of Uniform distribution of data in real life

- The normal distribution is used to model phenomenon that tend to follow a "bell-curve" shape. e.g. Weight of a baby, Shoe-size, Height of a person etc.
- Conversely, the uniform distribution is used to model scenarios where each potential outcome is equally likely.
  - e.g. rolling a die. If one rolls a die one time, the probability that it falls on a number between 1 and 6 follows a uniform distribution because each number is equally likely to occur - the probability that one rolls a 1 is $1/6$, one rolls a 2 is $1/6$, one rolls a 3 is $1/6$....OR
  - if one walks up to a random person on the street, the probability that their birthday falls on a given date would follow a uniform distribution because each day of the year is equally likely to be their birthday OR
  - suppose one randomly selects a card from a deck. The probability that the card will be either a spade, heart, club, or diamond follows a uniform distribution because each suit is equally likely to be chosen.

Following are the possible data distributions in the single-dimensional data.

Normal Distribution:

- in this case, the density of points decreases substantially as we move away from the mean.

Following are the possible data distributions in the single-dimensional data.

Normal Distribution:

- in this case, the density of points decreases substantially as we move away from the mean.
- about 0.1% of the points are more than $3\sigma$ (three standard deviations) away from the mean, and only about 5 x $10^{-8}$% of the points are more than six standard deviations away from the mean.

Following are the possible data distributions in the single-dimensional data.

Normal Distribution:

- in this case, the density of points decreases substantially as we move away from the mean.
- about 0.1% of the points are more than $3\sigma$ (three standard deviations) away from the mean, and only about 5 x $10^{-8}$% of the points are more than six standard deviations away from the mean.
- hence it is often the case that a threshold (such as $3x\sigma$) is chosen, and points beyond that distance from the mean are declared to be anomalous.

Following are the possible data distributions in the single-dimensional data.

Normal Distribution:

- in this case, the density of points decreases substantially as we move away from the mean.
- about 0.1% of the points are more than $3\sigma$ (three standard deviations) away from the mean, and only about 5 x $10^{-8}$% of the points are more than six standard deviations away from the mean.
- hence it is often the case that a threshold (such as $3x\sigma$) is chosen, and points beyond that distance from the mean are declared to be anomalous.
- one contrary perspective is that the existence of some points far away from the mean is just a consequence of the fact that a variable is normally distributed.

Following are the possible data distributions in the single-dimensional data.
Normal Distribution:

- in this case, the density of points decreases substantially as we move away from the mean.
- about 0.1% of the points are more than $3\sigma$ (three standard deviations) away from the mean, and only about 5 x $10^{-8}$% of the points are more than six standard deviations away from the mean.
- hence it is often the case that a threshold (such as $3x\sigma$) is chosen, and points beyond that distance from the mean are declared to be anomalous.
- one contrary perspective is that the existence of some points far away from the mean is just a consequence of the fact that a variable is normally distributed.
- an inference hence is that a set of points (away from the mean) is anomalous if and only if their number is substantially higher than the number expected if the data were to be normally distributed

Following are the possible data distributions in the single-dimensional data.

Normal Distribution:

- in this case, the density of points decreases substantially as we move away from the mean.
- about 0.1% of the points are more than $3\sigma$ (three standard deviations) away from the mean, and only about 5 x $10^{-8}$% of the points are more than six standard deviations away from the mean.
- hence it is often the case that a threshold (such as $3x\sigma$) is chosen, and points beyond that distance from the mean are declared to be anomalous.
- one contrary perspective is that the existence of some points far away from the mean is just a consequence of the fact that a variable is normally distributed.
- an inference hence is that a set of points (away from the mean) is anomalous if and only if their number is substantially higher than the number expected if the data were to be normally distributed
- e.g., if 2% of the data points are found beyond the $3\sigma$ threshold.

Following are the possible data distributions in the single-dimensional data.

Uniform Distribution:

- When data is distributed uniformly over a finite range, the mean and standard deviation merely characterize the range of values.

Following are the possible data distributions in the single-dimensional data.

Uniform Distribution:

- When data is distributed uniformly over a finite range, the mean and standard deviation merely characterize the range of values.
- Two inferences

Following are the possible data distributions in the single-dimensional data.

Uniform Distribution:

- When data is distributed uniformly over a finite range, the mean and standard deviation merely characterize the range of values.
- Two inferences
  - No anomalous data points: if the neighborhood of any data point is as richly populated as any other point, it can be argued that there are no anomalous data points

Following are the possible data distributions in the single-dimensional data.

Uniform Distribution:

- When data is distributed uniformly over a finite range, the mean and standard deviation merely characterize the range of values.
- Two inferences
    - No anomalous data points: if the neighborhood of any data point is as richly populated as any other point, it can be argued that there are no anomalous data points
    - a small neighborhood contains substantially fewer or more data points than expected from a uniform distribution.

# Distributions in Single-dimensional quantitative data...

Following are the possible data distributions in the single-dimensional data.

Other Unimodal Distributions :

- Many unimodal distributions are not normal, e.g., when there is a strict lower bound for the range of data values. Examples include log-normal and Gamma distributions.

# Distributions in Single-dimensional quantitative data...

Following are the possible data distributions in the single-dimensional data.
Other Unimodal Distributions :

- Many unimodal distributions are not normal, e.g., when there is a strict lower bound for the range of data values. Examples include log-normal and Gamma distributions.
- As with the normal distribution, if the nature and characteristics of the distribution are known,

# Distributions in Single-dimensional quantitative data...

Following are the possible data distributions in the single-dimensional data.

Other Unimodal Distributions :

- Many unimodal distributions are not normal, e.g., when there is a strict lower bound for the range of data values. Examples include log-normal and Gamma distributions.
- As with the normal distribution, if the nature and characteristics of the distribution are known,
  - one may seek to find thresholds beyond which a relatively small number (e.g., 1%) of the data points are found.

# Distributions in Single-dimensional quantitative data...

Following are the possible data distributions in the single-dimensional data.
Other Unimodal Distributions :

- Many unimodal distributions are not normal, e.g., when there is a strict lower bound for the range of data values. Examples include log-normal and Gamma distributions.
- As with the normal distribution, if the nature and characteristics of the distribution are known,
  - one may seek to find thresholds beyond which a relatively small number (e.g., 1%) of the data points are found.
- then, a collection of points (in a small region of the data space) is anomalous if their number is larger than predicted by the statistics of the distribution.

# Distributions in Single-dimensional quantitative data...

Following are the possible data distributions in the single-dimensional data.
Other Unimodal Distributions :

- Many unimodal distributions are not normal, e.g., when there is a strict lower bound for the range of data values. Examples include log-normal and Gamma distributions.
- As with the normal distribution, if the nature and characteristics of the distribution are known,
  - one may seek to find thresholds beyond which a relatively small number (e.g., 1%) of the data points are found.
- then, a collection of points (in a small region of the data space) is anomalous if their number is larger than predicted by the statistics of the distribution.

# Distributions in Single-dimensional quantitative data...

Following are the possible data distributions in the single-dimensional data.
Other Unimodal Distributions :

- Many unimodal distributions are not normal, e.g., when there is a strict lower bound for the range of data values. Examples include log-normal and Gamma distributions.
- As with the normal distribution, if the nature and characteristics of the distribution are known,
  - one may seek to find thresholds beyond which a relatively small number (e.g., 1%) of the data points are found.
- then, a collection of points (in a small region of the data space) is anomalous if their number is larger than predicted by the statistics of the distribution.
- Note: the distributions for some data sets have multiple modes, discovered only when the data is closely examined.

# Distributions in Single-dimensional quantitative data...

Following are the possible data distributions in the single-dimensional data.
Other Unimodal Distributions :

- Many unimodal distributions are not normal, e.g., when there is a strict lower bound for the range of data values. Examples include log-normal and Gamma distributions.
- As with the normal distribution, if the nature and characteristics of the distribution are known,
  - one may seek to find thresholds beyond which a relatively small number (e.g., 1%) of the data points are found.
- then, a collection of points (in a small region of the data space) is anomalous if their number is larger than predicted by the statistics of the distribution.
- Note: the distributions for some data sets have multiple modes, discovered only when the data is closely examined.
- Heuristics such as the $3\sigma$ rule are not useful with such distributions.

# Distributions in Single-dimensional quantitative data...

Following are the possible data distributions in the single-dimensional data.

Other Unimodal Distributions :

- Many unimodal distributions are not normal, e.g., when there is a strict lower bound for the range of data values. Examples include log-normal and Gamma distributions.
- As with the normal distribution, if the nature and characteristics of the distribution are known,
  - one may seek to find thresholds beyond which a relatively small number (e.g., 1%) of the data points are found.
- then, a collection of points (in a small region of the data space) is anomalous if their number is larger than predicted by the statistics of the distribution.
- Note: the distributions for some data sets have multiple modes, discovered only when the data is closely examined.
- Heuristics such as the $3\sigma$ rule are not useful with such distributions.
- Instead it is more useful to think of the data as consisting of a collection of clusters of data points.

Clustering plays an important role in anomaly detection.

Other Unimodal Distributions :

- Obviously, points which do not belong to any cluster are candidates to be considered anomalous.

Clustering plays an important role in anomaly detection.

Other Unimodal Distributions :

- Obviously, points which do not belong to any cluster are candidates to be considered anomalous.
    - What about the points which are distant from neighboring clusters ?

# Distributions in Single-dimensional quantitative data: Clustering

Clustering plays an important role in anomaly detection.

Other Unimodal Distributions :

- Obviously, points which do not belong to any cluster are candidates to be considered anomalous.
  - What about the points which are distant from neighboring clusters ?
- How are the clusters formed ?

# Distributions in Single-dimensional quantitative data: Clustering

Clustering plays an important role in anomaly detection.

Other Unimodal Distributions :

- Obviously, points which do not belong to any cluster are candidates to be considered anomalous.
  - What about the points which are distant from neighboring clusters ?
- How are the clusters formed ?
  - That is, when can a collection of points should be considered a cluster, and what it means for a point to be considered sufficiently distant from a cluster (or multiple clusters)?

Clustering plays an important role in anomaly detection.

Other Unimodal Distributions :

- Obviously, points which do not belong to any cluster are candidates to be considered anomalous.
  - What about the points which are distant from neighboring clusters ?
- How are the clusters formed ?
  - That is, when can a collection of points should be considered a cluster, and what it means for a point to be considered sufficiently distant from a cluster (or multiple clusters)?
- Three alternatives followed here

# Distributions in Single-dimensional quantitative data: Clustering

Clustering plays an important role in anomaly detection.

Other Unimodal Distributions :

- Obviously, points which do not belong to any cluster are candidates to be considered anomalous.
    - What about the points which are distant from neighboring clusters ?
- How are the clusters formed ?
    - That is, when can a collection of points should be considered a cluster, and what it means for a point to be considered sufficiently distant from a cluster (or multiple clusters)?
- Three alternatives followed here
    - Density-based cluster identification

# Distributions in Single-dimensional quantitative data: Clustering

Clustering plays an important role in anomaly detection.

Other Unimodal Distributions :

- Obviously, points which do not belong to any cluster are candidates to be considered anomalous.
    - What about the points which are distant from neighboring clusters ?
- How are the clusters formed ?
    - That is, when can a collection of points should be considered a cluster, and what it means for a point to be considered sufficiently distant from a cluster (or multiple clusters)?
- Three alternatives followed here
    - Density-based cluster identification
    - Intra-group based cluster identification

# Distributions in Single-dimensional quantitative data: Clustering

Clustering plays an important role in anomaly detection.

Other Unimodal Distributions :

- Obviously, points which do not belong to any cluster are candidates to be considered anomalous.
  - What about the points which are distant from neighboring clusters ?
- How are the clusters formed ?
  - That is, when can a collection of points should be considered a cluster, and what it means for a point to be considered sufficiently distant from a cluster (or multiple clusters)?
- Three alternatives followed here
  - Density-based cluster identification
  - Intra-group based cluster identification
  - inter-group distances based cluster identification

# Distributions in Single-dimensional quantitative data: Clustering

Clustering plays an important role in anomaly detection.

Other Unimodal Distributions :

- Obviously, points which do not belong to any cluster are candidates to be considered anomalous.
    - What about the points which are distant from neighboring clusters ?
- How are the clusters formed ?
    - That is, when can a collection of points should be considered a cluster, and what it means for a point to be considered sufficiently distant from a cluster (or multiple clusters)?
- Three alternatives followed here
    - Density-based cluster identification
    - Intra-group based cluster identification
    - inter-group distances based cluster identification

Clustering plays an important role in anomaly detection.

Other Unimodal Distributions :

- Obviously, points which do not belong to any cluster are candidates to be considered anomalous.
    - What about the points which are distant from neighboring clusters ?
- How are the clusters formed ?
    - That is, when can a collection of points should be considered a cluster, and what it means for a point to be considered sufficiently distant from a cluster (or multiple clusters)?
- Three alternatives followed here
    - Density-based cluster identification
    - Intra-group based cluster identification
    - inter-group distances based cluster identification

We discuss these further.....

Density-based Clustering

- are very popular in identifying anomalous data.

Density-based Clustering

- are very popular in identifying anomalous data.
- If the relative number of points (per unit distance) is substantially higher in a small region than the entire data set, the points in that region can be considered as a cluster.

Density-based Clustering

- are very popular in identifying anomalous data.
- If the relative number of points (per unit distance) is substantially higher in a small region than the entire data set, the points in that region can be considered as a cluster.
- still not a strictly mathematical definition, since there are fuzzy phrases.

Density-based Clustering

- are very popular in identifying anomalous data.
- If the relative number of points (per unit distance) is substantially higher in a small region than the entire data set, the points in that region can be considered as a cluster.
- still not a strictly mathematical definition, since there are fuzzy phrases.
- the distribution of densities can itself be analyzed, and (if unimodal) we can identify the density threshold beyond which the density is high enough to consider a region to contain a cluster.

Density-based Clustering

- are very popular in identifying anomalous data.
- If the relative number of points (per unit distance) is substantially higher in a small region than the entire data set, the points in that region can be considered as a cluster.
- still not a strictly mathematical definition, since there are fuzzy phrases.
- the distribution of densities can itself be analyzed, and (if unimodal) we can identify the density threshold beyond which the density is high enough to consider a region to contain a cluster.
- Note that the same data set may contain one region of higher density and another region of lower density that may also be considered to be a cluster.

## Intra-group/Inter-group distance-based Clustering

- here, points form a cluster if they are substantially closer to each other (on average) than they are to the nearest points outside the cluster.

## Intra-group/Inter-group distance-based Clustering

- here, points form a cluster if they are substantially closer to each other (on average) than they are to the nearest points outside the cluster.
  - Unfortunately, there are some data distributions in which this definition leads to trivial cluster identifications (e.g., placing almost all the points in the same cluster).

## Intra-group/Inter-group distance-based Clustering

- here, points form a cluster if they are substantially closer to each other (on average) than they are to the nearest points outside the cluster.
  - Unfortunately, there are some data distributions in which this definition leads to trivial cluster identifications (e.g., placing almost all the points in the same cluster).
- Should every point be allocated to some cluster ?

## Intra-group/Inter-group distance-based Clustering

- here, points form a cluster if they are substantially closer to each other (on average) than they are to the nearest points outside the cluster.
  - Unfortunately, there are some data distributions in which this definition leads to trivial cluster identifications (e.g., placing almost all the points in the same cluster).
- Should every point be allocated to some cluster ?
  - some clustering algorithms allocate every point to some cluster, which is not necessary.

## Intra-group/Inter-group distance-based Clustering

- here, points form a cluster if they are substantially closer to each other (on average) than they are to the nearest points outside the cluster.
    - Unfortunately, there are some data distributions in which this definition leads to trivial cluster identifications (e.g., placing almost all the points in the same cluster).
- Should every point be allocated to some cluster ?
    - some clustering algorithms allocate every point to some cluster, which is not necessary.
    - other algorithms assume that the number of clusters is fixed (and predetermined or provided by the user); this is again not necessary.

Intra-group/Inter-group distance-based Clustering

- here, points form a cluster if they are substantially closer to each other (on average) than they are to the nearest points outside the cluster.
    - Unfortunately, there are some data distributions in which this definition leads to trivial cluster identifications (e.g., placing almost all the points in the same cluster).
- Should every point be allocated to some cluster ?
    - some clustering algorithms allocate every point to some cluster, which is not necessary.
    - other algorithms assume that the number of clusters is fixed (and predetermined or provided by the user); this is again not necessary.
- When clustering-based approaches are used for anomaly detection, points inside clusters of a minimal size are usually not considered to be anomalous.

Intra-group/Inter-group distance-based Clustering

- here, points form a cluster if they are substantially closer to each other (on average) than they are to the nearest points outside the cluster.
    - Unfortunately, there are some data distributions in which this definition leads to trivial cluster identifications (e.g., placing almost all the points in the same cluster).
- Should every point be allocated to some cluster ?
    - some clustering algorithms allocate every point to some cluster, which is not necessary.
    - other algorithms assume that the number of clusters is fixed (and predetermined or provided by the user); this is again not necessary.
- When clustering-based approaches are used for anomaly detection, points inside clusters of a minimal size are usually not considered to be anomalous.
    - this "minimal size" is again an externally specified parameter, such as a threshold based on the distribution of sizes of clusters in the data set.

# Anomaly Detection: Evaluation Criteria

- Every anomaly detection problem appears to have different characteristics, and an algorithm that *performs well* on one problem may not *perform well* on another.

# Anomaly Detection: Evaluation Criteria

- Every anomaly detection problem appears to have different characteristics, and an algorithm that *performs well* on one problem may not *perform well* on another.
- Three issues come up here

# Anomaly Detection: Evaluation Criteria

- Every anomaly detection problem appears to have different characteristics, and an algorithm that *performs well* on one problem may not *perform well* on another.
- Three issues come up here
  1. Can quantitative metrics be devised so that we can unambiguously say which of two data points in a given data set is "more anomalous"—without appealing to human intuition?

# Anomaly Detection: Evaluation Criteria

- Every anomaly detection problem appears to have different characteristics, and an algorithm that *performs well* on one problem may not *perform well* on another.
- Three issues come up here
  1. Can quantitative metrics be devised so that we can unambiguously say which of two data points in a given data set is "more anomalous"—without appealing to human intuition?
  2. Each anomaly detection algorithm answers the above question in a procedural way, e.g., based on distance to k nearest neighbors. Can this implicit choice be justified on mathematical or rational grounds?

# Anomaly Detection: Evaluation Criteria

- Every anomaly detection problem appears to have different characteristics, and an algorithm that *performs well* on one problem may not *perform well* on another.
- Three issues come up here
  1. Can quantitative metrics be devised so that we can unambiguously say which of two data points in a given data set is "more anomalous"—without appealing to human intuition?
  2. Each anomaly detection algorithm answers the above question in a procedural way, e.g., based on distance to k nearest neighbors. Can this implicit choice be justified on mathematical or rational grounds?
  3. In some cases, algorithms also appeal to the desire to compute the results in a "reasonable" amount of time, ruling out the search for optimal solutions. In such cases, can we say anything about the quality of the obtained solutions when compared to the optimal solutions?

# Anomaly Detection: Evaluation Criteria

- Every anomaly detection problem appears to have different characteristics, and an algorithm that *performs well* on one problem may not *perform well* on another.
- Three issues come up here
  1. Can quantitative metrics be devised so that we can unambiguously say which of two data points in a given data set is "more anomalous"—without appealing to human intuition?
  2. Each anomaly detection algorithm answers the above question in a procedural way, e.g., based on distance to k nearest neighbors. Can this implicit choice be justified on mathematical or rational grounds?
  3. In some cases, algorithms also appeal to the desire to compute the results in a "reasonable" amount of time, ruling out the search for optimal solutions. In such cases, can we say anything about the quality of the obtained solutions when compared to the optimal solutions?

# Anomaly Detection: Evaluation Criteria

- Every anomaly detection problem appears to have different characteristics, and an algorithm that *performs well* on one problem may not *perform well* on another.
- Three issues come up here
  1. Can quantitative metrics be devised so that we can unambiguously say which of two data points in a given data set is "more anomalous"—without appealing to human intuition?
  2. Each anomaly detection algorithm answers the above question in a procedural way, e.g., based on distance to k nearest neighbors. Can this implicit choice be justified on mathematical or rational grounds?
  3. In some cases, algorithms also appeal to the desire to compute the results in a "reasonable" amount of time, ruling out the search for optimal solutions. In such cases, can we say anything about the quality of the obtained solutions when compared to the optimal solutions?

Information theory provides a possible answer to these questions. That is, many real-life processes are amenable to succinct descriptions of their essence - so variation is considered an anomaly. We see this in an example....

*Next Chapter: Distance-Based Anomaly Detection Approaches & Algorithms*

*B l a n k*

*B l a n k*

*B l a n k*

*B l a n k*

*B l a n k*

*B l a n k*