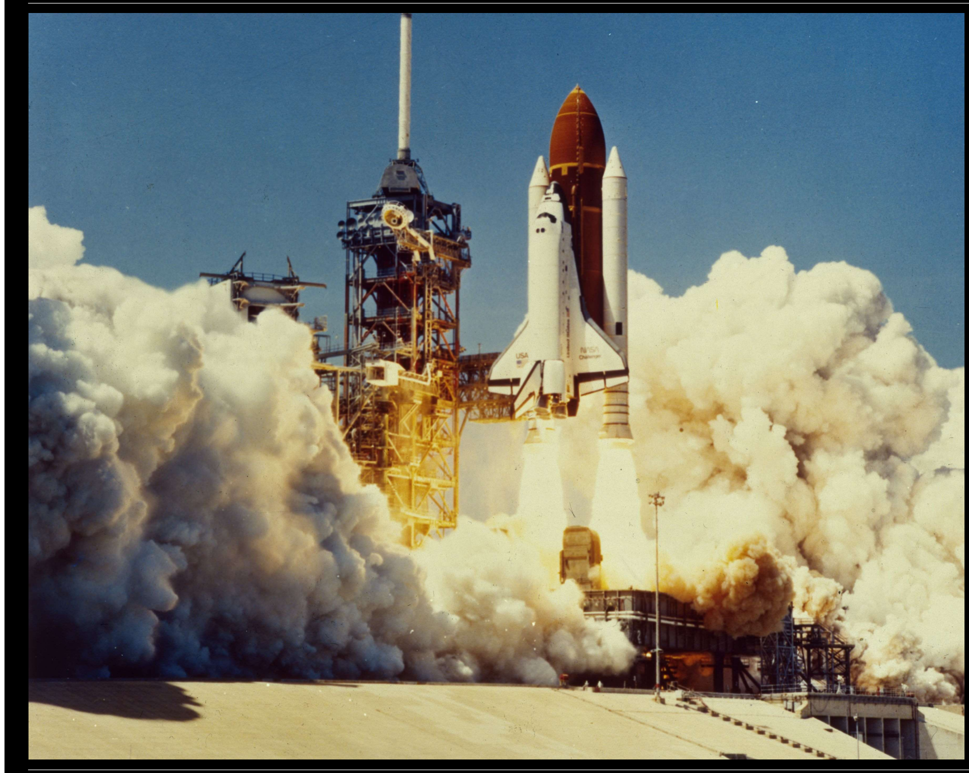
Master Degree in Artificial Intelligence for Science and Technology

Introduction to Anomaly Detection

Fabio Stella
Department of Informatics, Systems and Communication
University of Milan-Bicocca
fabio.stella@unimib.it

OUTLOOK

- Introduction
- Type of Attributes and Complex Data
- Types of Data
- Types of Anomalies
- Output of Anomaly Detection
- Applications of Anomaly Detection



The Callenger Space Shuttle

Kennedy Space Center in Florida

January 28, 1986

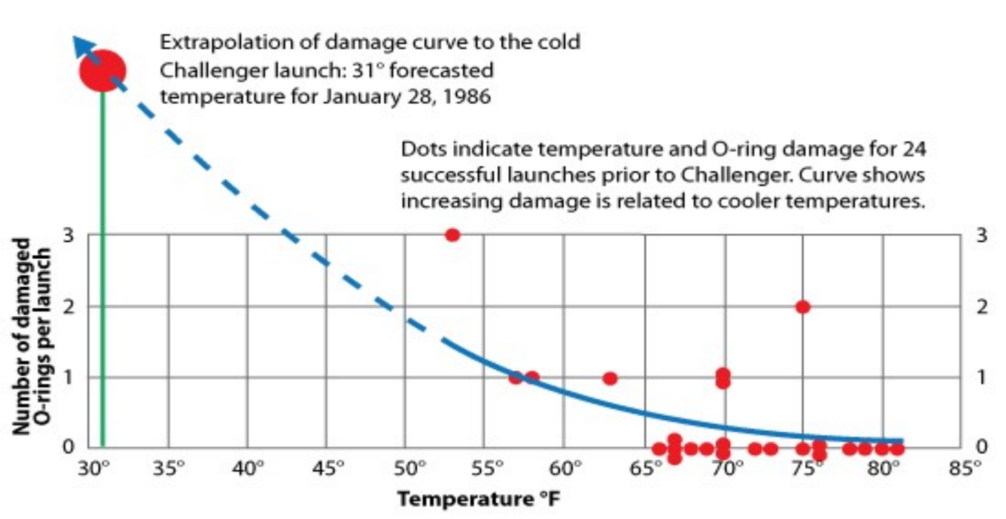
The space shuttle Challenger and its crew were destroyed in a fiery, catastrophic explosion.

NASA appointed members of the Rogers Commission to investigate the cause of the disaster.

When he was asked to be a part of this commission, Richard Feynman rather reluctantly accepted. Little did he know that he would be the one person to discover the exact cause of the explosion.

He learned many things from these people that would help him to discover the cause of the explosion; and also information that helped him realize what a risky business flying a shuttle really is. NASA officials said that the chance of failure of the shuttle was about 1 in 100,000; Feynman found that this number was actually closer to 1 in 100.

He also learned that rubber used to seal the solid rocket booster joints using O-rings, failed to expand when the temperature was at or below 32 degrees F (0 degrees C).

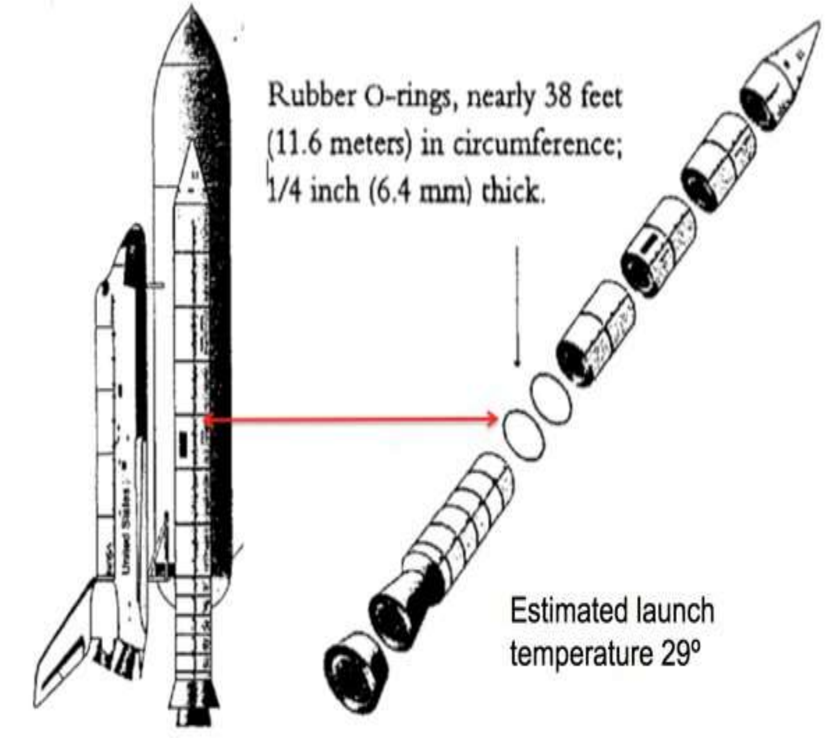

Extrapolation of damage curve to the cold Challenger launch: 31° forecasted temperature for January 28, 1986

Dots indicate temperature and O-ring damage for 24 successful launches prior to Challenger. Curve shows increasing damage is related to cooler temperatures.

Feynman now believed that he had the solution, but to test it, he dropped a piece of the O-ring material, squeezed with a C-clamp to simulate the actual conditions of the shuttle, into a glass of ice water. Ice, of course, is 32 degrees F.

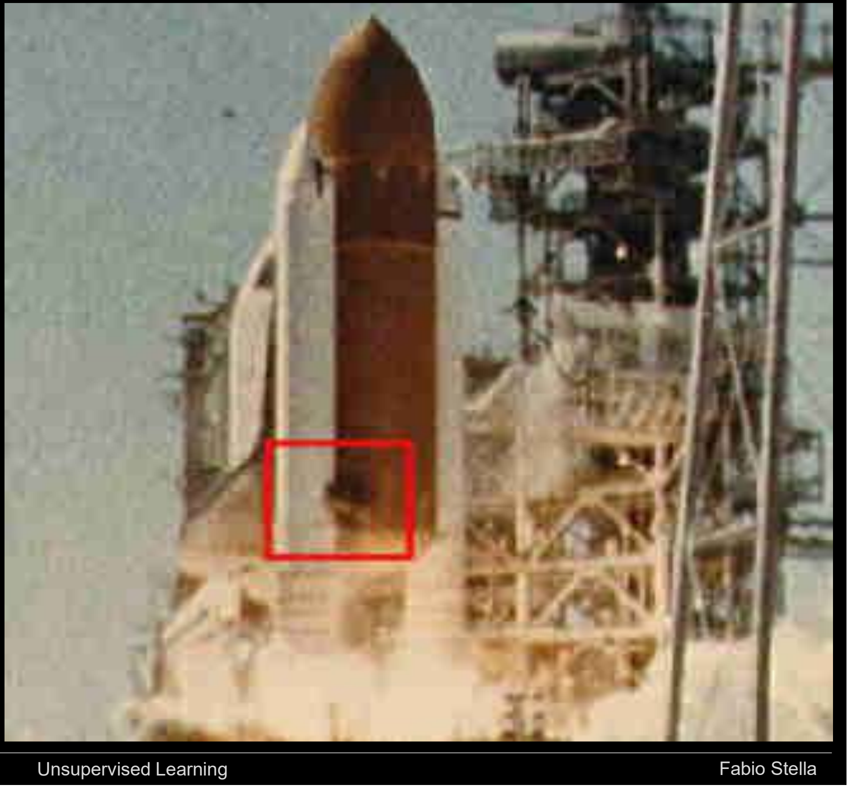
At this point one needs to understand exactly what role the O-rings play in the solid rocket booster (SRB) joints.

When the material in the SRB start to heat up, it expands and pushes against the sides of the SRB. If there is an opening in a joint in the SRB, the gas tries to escape through that opening (think of it like water in a tea kettle escaping through the spout.) This leak in the Challenger's SRB was easily visible as a small flicker in a launch photo.

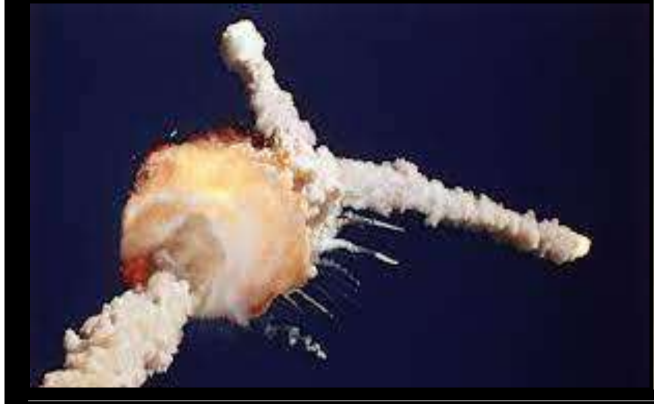

Rubber O-rings, nearly 38 feet (11.6 meters) in circumference; 1/4 inch (6.4 mm) thick.
Estimated launch temperature 29°

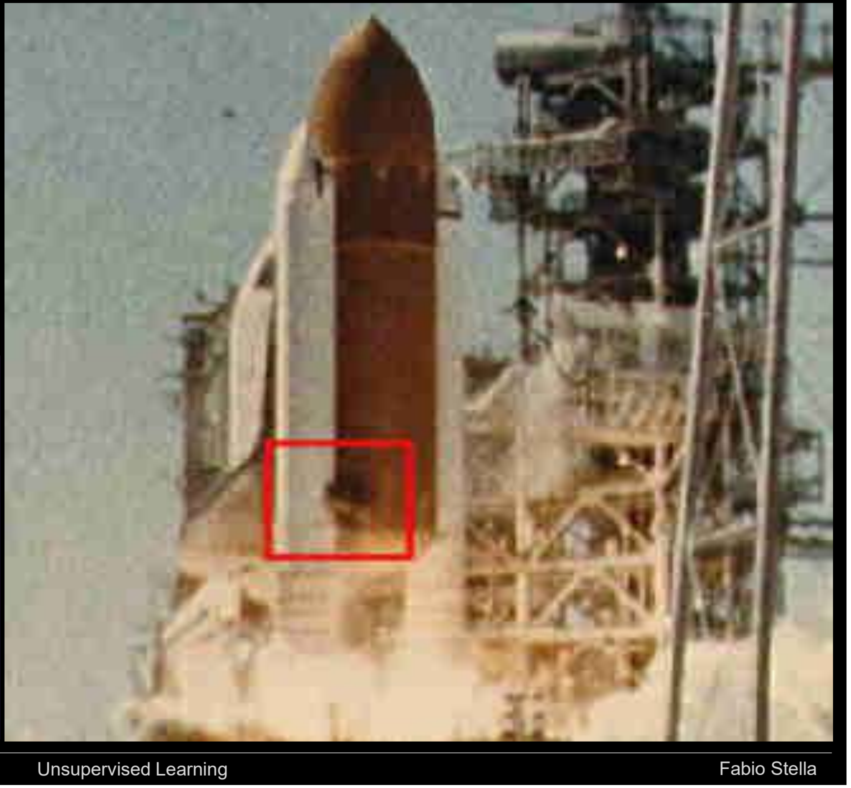
At this point one needs to understand exactly what role the O-rings play in the solid rocket booster (SRB) joints.

When the material in the SRB start to heat up, it expands and pushes against the sides of the SRB. If there is an opening in a joint in the SRB, the gas tries to escape through that opening (think of it like water in a tea kettle escaping through the spout.) This leak in the Challenger's SRB was easily visible as a small flicker in a launch photo.



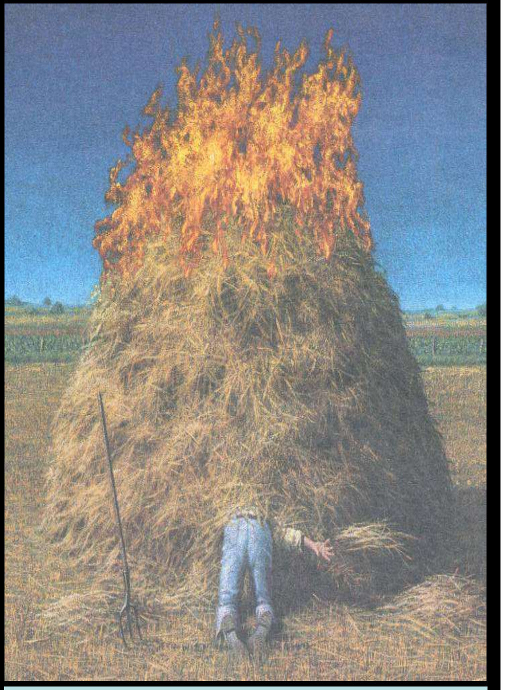
This flicker turned into a flame and began heating the fuel tank, which then ruptured. When this happened, the fuel tank released liquid hydrogen into the atmosphere where it exploded.





- What are anomalies/outliers?
  - The set of data points that are considerably different than the remainder of the data
- Natural implication is that anomalies are relatively rare
  - One in a thousand occurs often if you have lots of data
  - Context is important, e.g., freezing temps in July
- Can be important or a nuisance
  - Unusually high blood pressure
  - 200 pound, 2 year old
  - 80 years old and pregnant

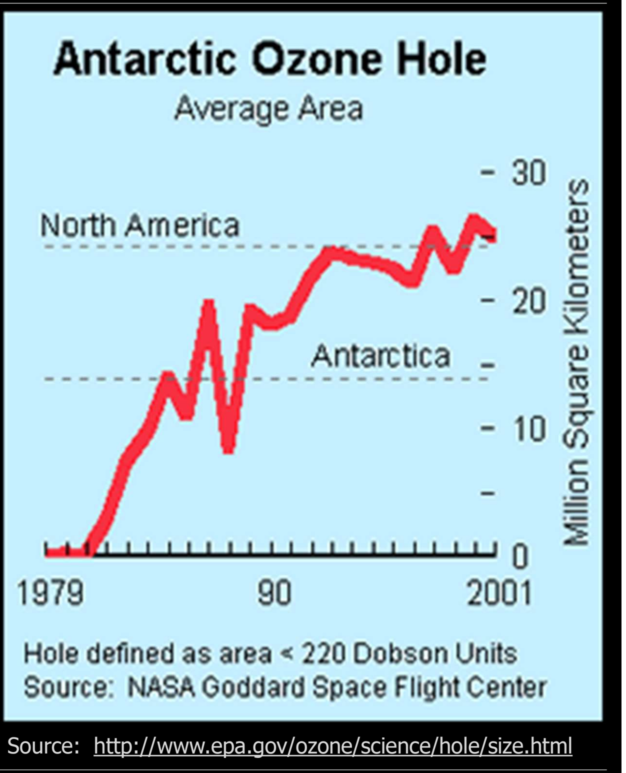
$$\underbrace{\overbrace{\underbrace{0C^\circ}_{\text{ok}} + \text{July } 30^{\text{th}}}^{?} + 3{,}000 \text{ Mts}}_{\text{ok}}$$



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

Ozone Depletion History

- In 1985 three researchers (Farman, Gardinar and Shanklin) were puzzled by data gathered by the British Antarctic Survey showing that ozone levels for Antarctica had dropped 10% below normal levels
- Why did the Nimbus 7 satellite, which had instruments aboard for recording ozone levels, not record similarly low ozone concentrations?
- The ozone concentrations recorded by the satellite were so low they were being treated as outliers by a computer program and discarded!



Source: http://www.epa.gov/ozone/science/hole/size.html

- Anomaly is a pattern in the data that does not conform to the expected behavior
- Also referred to as outliers, exceptions, peculiarities, surprises, etc.

Causes of Anomalies

- Data from different classes: measuring the weights of oranges, but a few grapefruit are mixed in
- Natural variation: unusually tall people
- Data errors: 200 pound 2 year old
- Cyber intrusions
- Credit card fraud
- Faults in mechanical systems

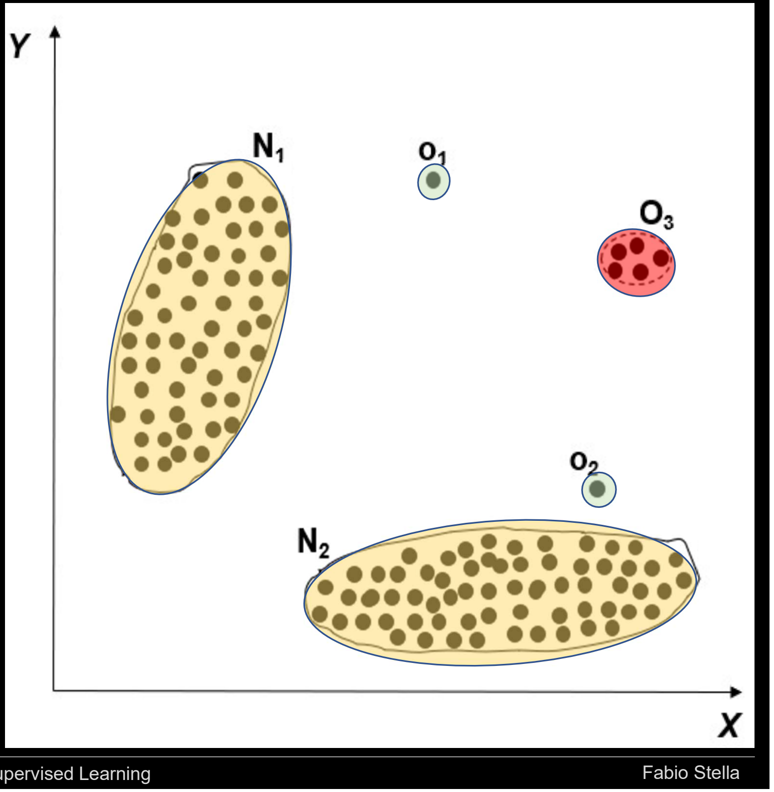
Real World Anomalies

- Credit Card Fraud
  - An abnormally high purchase made on a credit card

- Cyber Intrusions
  - A web server involved in ftp traffic

A Simple Example of Anomaly

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



Input Data

- Most common form of data handled by anomaly detection techniques is Record Data

Univariate

Engine Temperature
192
195
180
199
19
177
172
285
195
163

Multi-variate

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

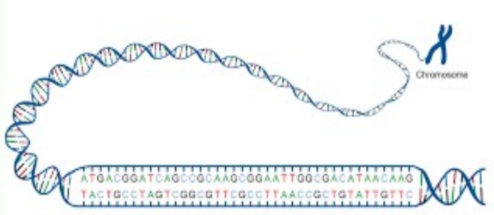
Type of Attributes

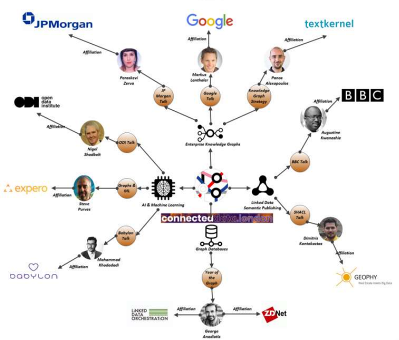
- Binary
- Categorical
- Continuous
- Hybrid

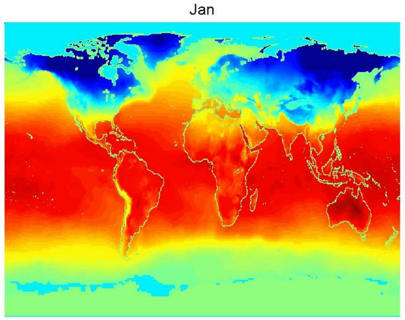
	categorical	continuous	categorical	continuous	binary
Tid	**SrcIP**	**Duration**	**Dest IP**	**Number of bytes**	**Internal**
1	206.163.37.81	0.10	160.94.179.208	150	**No**
2	206.163.37.99	0.27	160.94.179.235	208	**No**
3	160.94.123.45	1.23	160.94.179.221	195	**Yes**
4	206.163.37.37	112.03	160.94.179.253	199	**No**
5	206.163.37.41	0.32	160.94.179.244	181	**No**

INPUT DATA: COMPLEX DATA TYPES

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


Chromosome
ATGACGGATCAGCCGCAAGCGGAATTGGCGACATAACAAG
TACTGCCTAGTCGGCGTTCGCCTTAACCGCTGTATTGTTC





AVAILABILITY OF DATA LABELS

- Supervised Anomaly Detection
  - — Labels available for both normal data and anomalies
  - — Similar to rare class mining
- Semi-supervised Anomaly Detection
  - — Labels available only for normal data
- Unsupervised Anomaly Detection
  - — No labels assumed
  - — Based on the assumption that anomalies are very rare compared to normal data

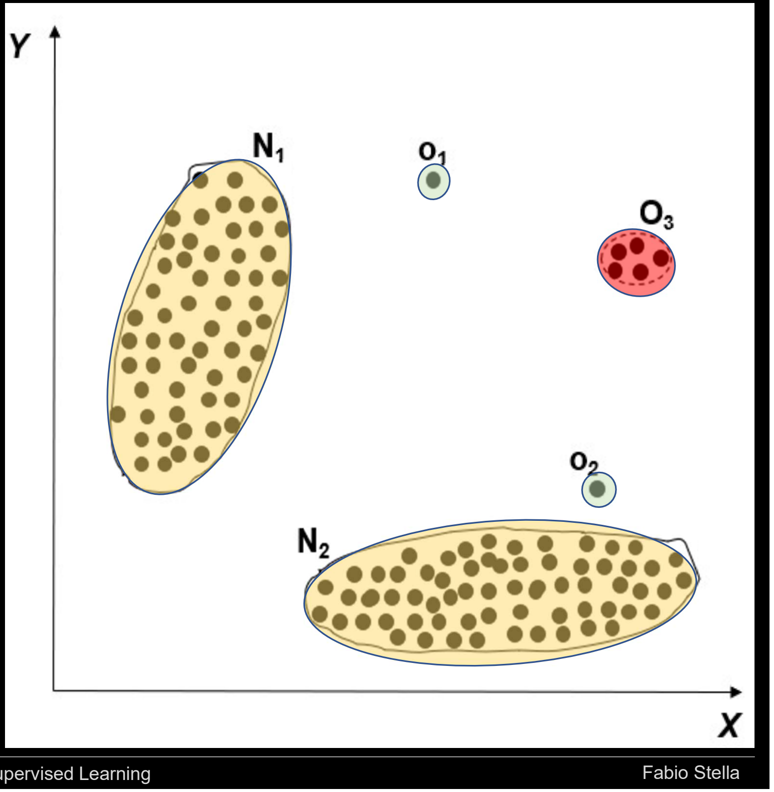
Types of Anomalies

- Point Anomalies
- Contextual Anomalies
- Collective Anomalies

TYPES OF ANOMALIES

- **POINT ANOMALIES**
- Contextual Anomalies
- Collective Anomalies

An individual data instance is anomalous w.r.t. the data



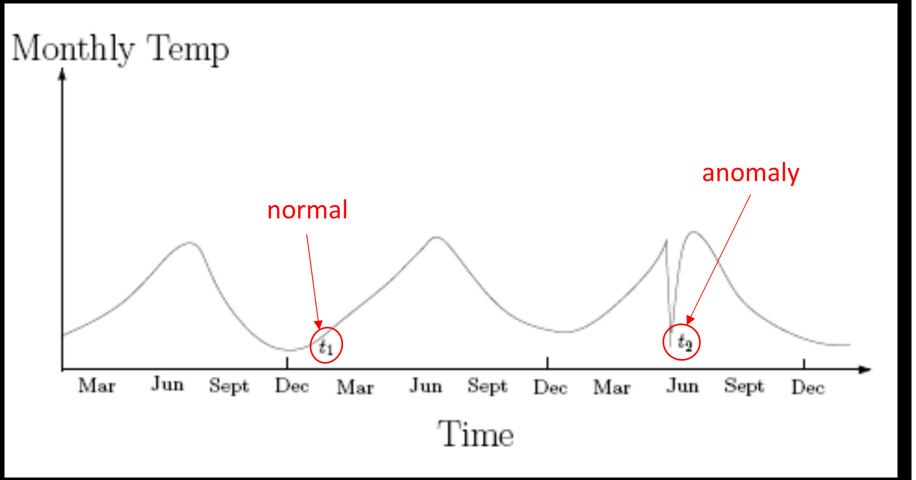
TYPES OF ANOMALIES

- Point Anomalies
- CONTEXTUAL ANOMALIES
- Collective Anomalies

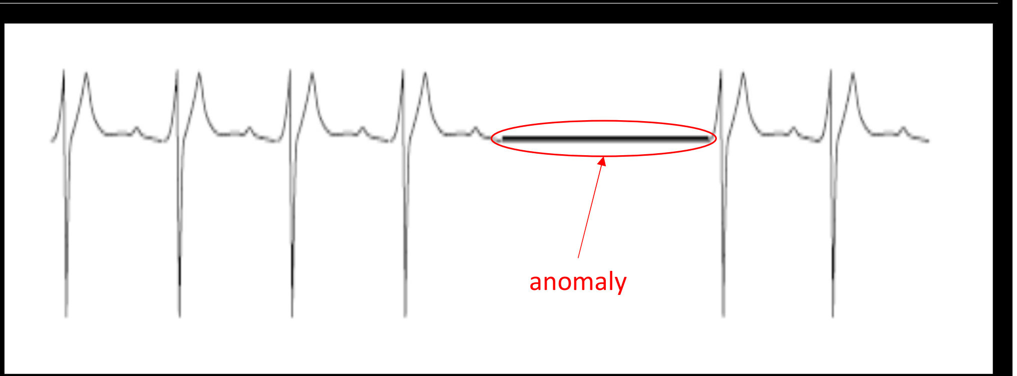
  — An individual data instance is anomalous within a context

  — Requires a notion of context

  — Also referred to as conditional anomalies



Types of Anomalies



- Point Anomalies
- Contextual Anomalies
- **Collective Anomalies**
  - A collection of related data instances is anomalous
  - Requires a relationship among data instances
    - Sequential Data
    - Spatial Data
    - Graph Data
  - The individual instances within a collective anomaly are not anomalous by themselves

Output of Anomaly Detection

- **LABEL**
  - Each test instance is given a normal or anomaly label
  - This is especially true of classification-based approaches

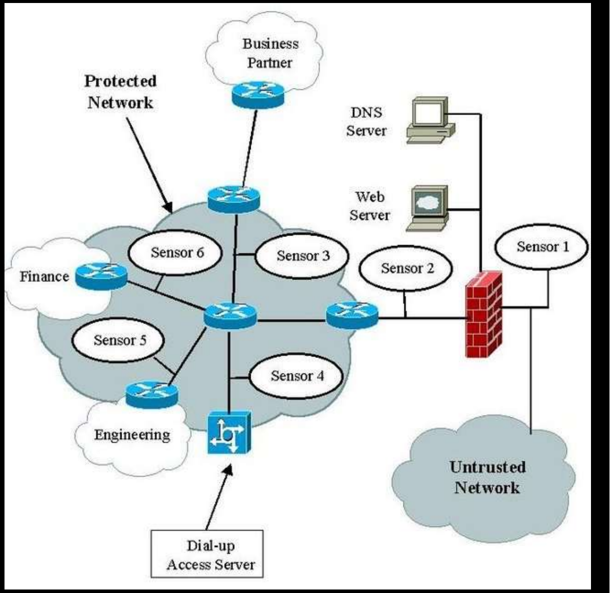
- **SCORE**
  - Each test instance is assigned an anomaly score
  - Allows the output to be ranked
  - Requires an additional threshold parameter

Applications of Anomaly Detection

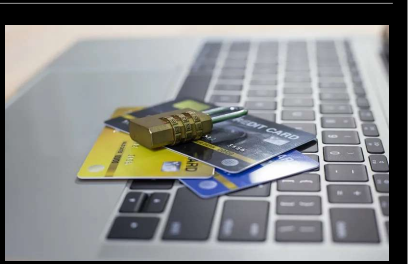
- Network intrusion detection
- Insurance / Credit card fraud detection
- Healthcare Informatics / Medical diagnostics
- Industrial Damage Detection
- Image Processing / Video surveillance
- Novel Topic Detection in Text Mining

APPLICATIONS OF ANOMALY DETECTION: NETWORK INTRUSION DETECTION

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



Applications of Anomaly Detection: Fraud Detection



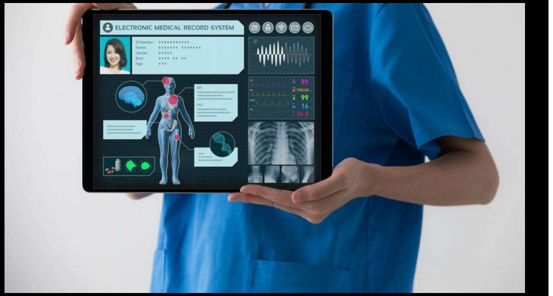
- Fraud detection refers to detection of criminal activities occurring in commercial organizations:
  - Malicious users might be the actual customers of the organization or might be posing as a customer (also known as identity theft).

- Types of fraud
  - Credit card fraud
  - Insurance claim fraud
  - Mobile / cell phone fraud
  - Insider trading

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

APPLICATIONS OF ANOMALY DETECTION: HEALTH INFORMATICS

- Detect anomalous patient records:
  - Indicate disease outbreaks, instrumentation errors, etc.
- Key Challenges
  - Only normal labels available
  - Misclassification cost is very high
  - Data can be complex: spatio-temporal



APPLICATIONS OF ANOMALY DETECTION: INDUSTRIAL DAMAGE DETECTION

- Industrial damage detection refers to detection of different faults and failures in complex industrial systems, structural damages, intrusions in electronic security systems, abnormal energy consumption, etc.
  - Example: Aircraft Safety
    - Anomalous Aircraft (Engine) / Fleet Usage
    - Anomalies in engine combustion data
    - Total aircraft health and usage management

- Key Challenges
  - Data is extremely huge, noisy and unlabelled
  - Most of applications exhibit temporal behavior
  - Detecting anomalous events typically require immediate intervention

Applications of Anomaly Detection: Image Processing

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

Model-based vs Model-free

- **Model-based Approaches**
  - Model can be parametric or non-parametric
  - Anomalies are those points that don’t fit well
  - Anomalies are those points that distort the model

- **Model-free Approaches**
  - Anomalies are identified directly from the data without building a model

- Often the underlying assumption is that most of the points in the data are normal

Point Anomaly Detection Techniques

- **Nearest Neighbor Based**
  - Anomalies are points far away from other points
- **Clustering Based**
  - Points far away from cluster centers are outliers
  - Small clusters are outliers
- **Statistical Approaches**
- **Reconstruction Based**

RECAP

- Introduction
- Type of Attributes and Complex Data
- Types of Data
- Types of Anomalies
- Output of Anomaly Detection
- Applications of Anomaly Detection