technische universität
dortmund

A COMPREHENSIVE SURVEY OF ANOMALY DETECTION TECHNIQUES FOR HIGH DIMENSIONAL BIG DATA

Presented by : Jaykumar Savani (230443)

Supervisor : Simon Klüttermann

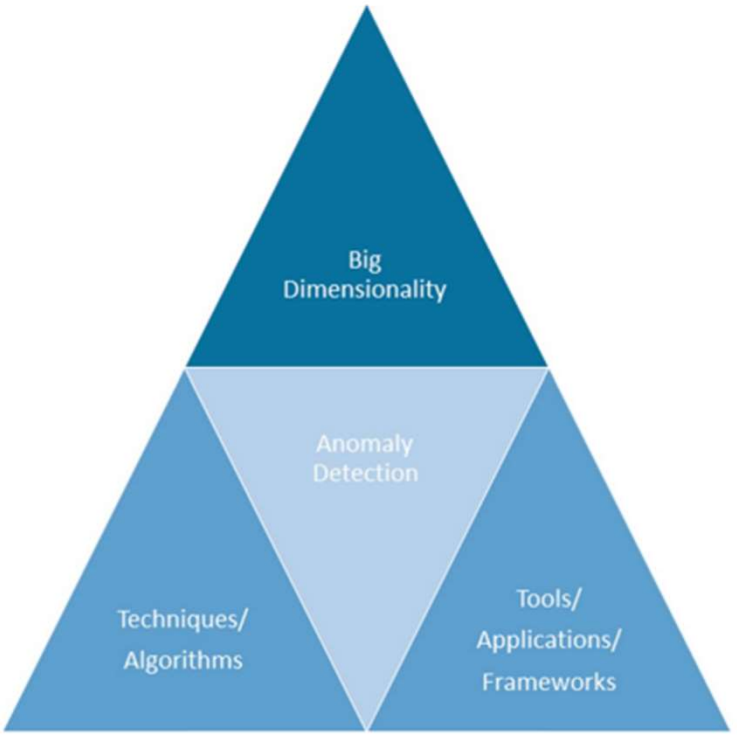
Professor : Prof. Dr. Emmanuel Müller

Introduction

What is the scope of the paper ?

To Understand :

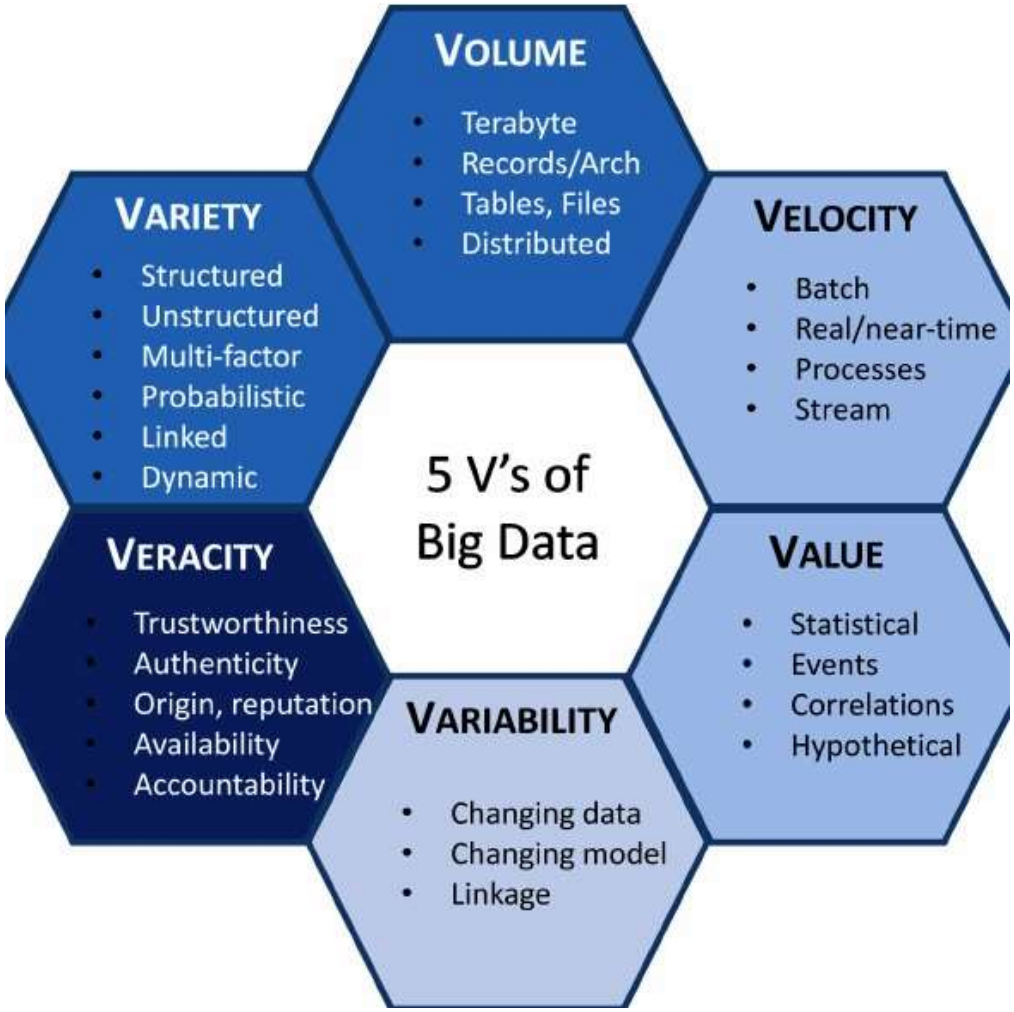
1. Big Dimensionality
2. Anomaly Detection
3. Techniques/ Algorithms
4. Curse of dimensionality
5. Tackling Methods
6. Tools/Application/Frameworks



1. Big data

What is Big Data?

- **Collection of datasets which are**
  - Large
  - Complex
  - Difficult to process using traditional data processing tools and techniques
- **Aim** : To solve new problems or old problems in a better ways
- **Characteristics** : (Most famous 5 - V's)
  - Value
  - Volume
  - Velocity
  - Variety
  - Veracity



Source : https://www.researchgate.net/figure/The-five-Vs-of-Big-Data-Adapted-from-IBM-big-data-platform-Bringing-big-data-to-the_fig1_281404634

2. Anomaly Detection

How to determine outlier-ness ?

- Two traditional ways :
  1. Real Valued outlier scores

     Quantifies tendency of data point by assigning a score or probability value
  2. Binary Labels

     Result of using threshold to convert scores to binary labels as "Inlier " or "Outlier"

- Depends on analyst's judgement.

3. Techniques and Algorithms

Techniques for anomaly detection

1. Statistical Techniques
   I. Parametric Techniques
      a. Gaussian model
      b. Regression based models
   II. Non Parametric Techniques
      a. Histogram based models
      b. Kernel based models
2. Proximity based models
   I. Cluster analysis
   II. Nearest Neighbour analysis
3. ensemble techniques
   I. Sequential ensembles
   II. Independent ensembles

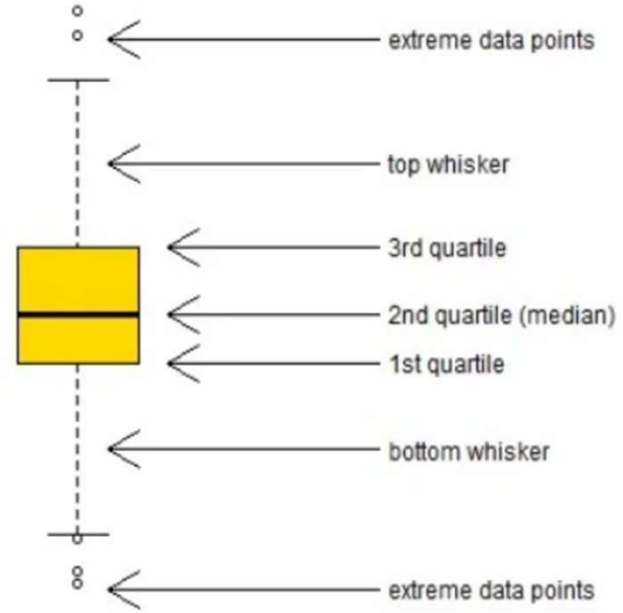
Statistical Techniques

- Fits statistical model to data
- Apply statistical inference test
- Assumption :
  - Distribution is known
  - Parameters for distribution can be estimated
- Data : depends on the technique – either parametric/non-parametric
- Result : whether given data point is anomaly or not ?

Parametric techniques - 1

- Aim : to determine point belongs to distribution or not
- Gaussian techniques :

1. Box plot:
   - Standard way of displaying variation of data – Five number summary
   - Outliers : Any data point does not between the min and max



Parametric techniques - 2

2. Chi Square test

- Performs a simple test for detecting outliers in univariate data
- Sample variance counts as estimator of variance
- Outliers can exist at both tails of the data

- $X^2 = \frac{\sum(observerd - expecte\ )^2}{expected}$

3. Grubb's test

- Performs a simple test for detecting outliers in univariate data
- But, assumed that data comes from normally distributed population
- H_o : No outliers
- H_1 : Exactly one outlier

- $G = \frac{\max_{i=1,..N} |Y_i - \bar{Y}|}{s}$

Parametric techniques - 2

Regression :

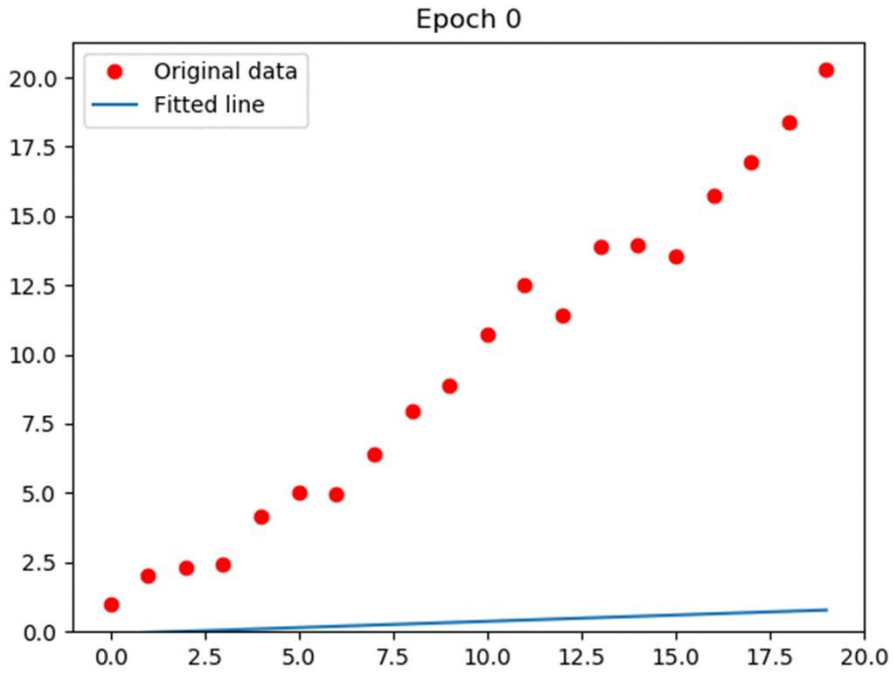
Summarize data

↓

Model parameterized by coefficients and constants

↓

Review magnitude of residuals → Anomalies

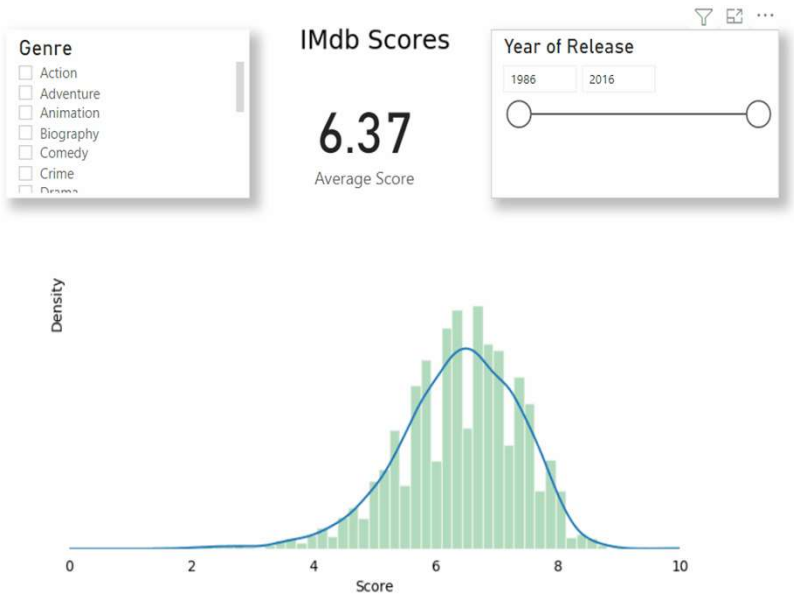


Source: https://github.com/arvention/linear-regression

Non-parametric techniques

- Assumption : Data distribution isn't known
- Aim : data point belongs to assumed normal model
- Traditional ways :

1. Histograms
2. Kernel Based



Source : https://medium.com/geekculture/advanced-python-visualizations-in-powerbi-histograms-and-frequency-plots-66f238684011

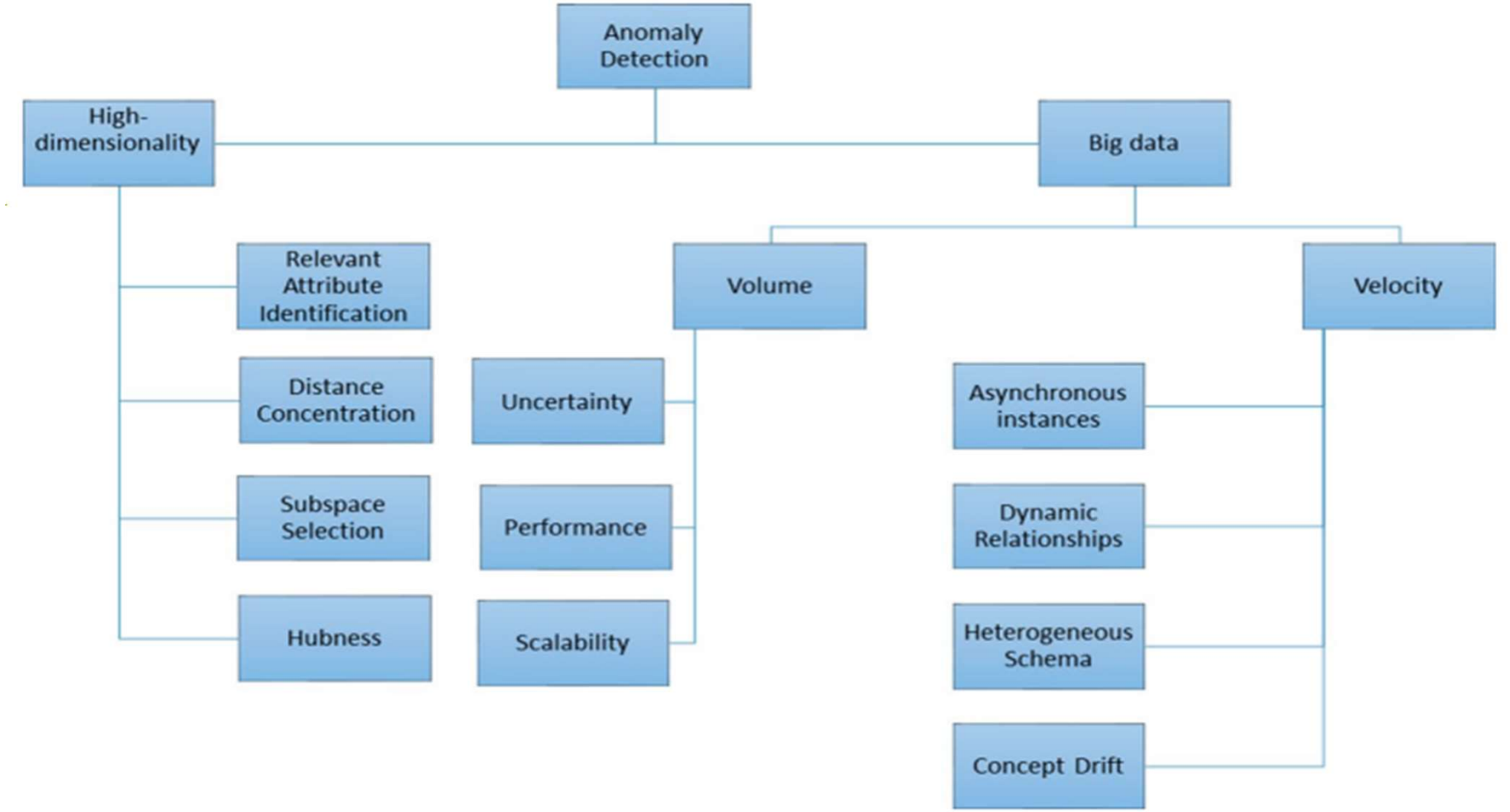
Proximity/Distance based techniques

- Use Case : Unsupervised Machine Learning
- Assumption : Anomalous data points are isolated from the rest of data groups
- Goal : Segmentation of data points to find anomalies
- Data : Multi-dimensional cross sectional data
- Algorithms / Techniques used for classification:
  - Clustering: K-Means
  - Nearest Neighbours : KNN algo, Local outlier factor score (LOF score)

Ensembles techniques

- Use Case : combining multiple algorithms to increase robustness of anomaly detection algorithms:
- Goal : Use ensembles to enhance the quality of anomaly detection
- Data : multi-dimensional, cross-sectional
- Algorithms/ Techniques :
  - Sequential ensembles – Boosting Methods (Ex. ADABOOST)
  - Independent ensembles – Random Forest

Challenges in context of high dimensional data



Source : https://journalofbigdata.springeropen.com/articles/10.1186/s40537-020-00320-x

Curse of Dimensionality - [by Richard E. Bellman (1961)]

Dimensionality increases

↓

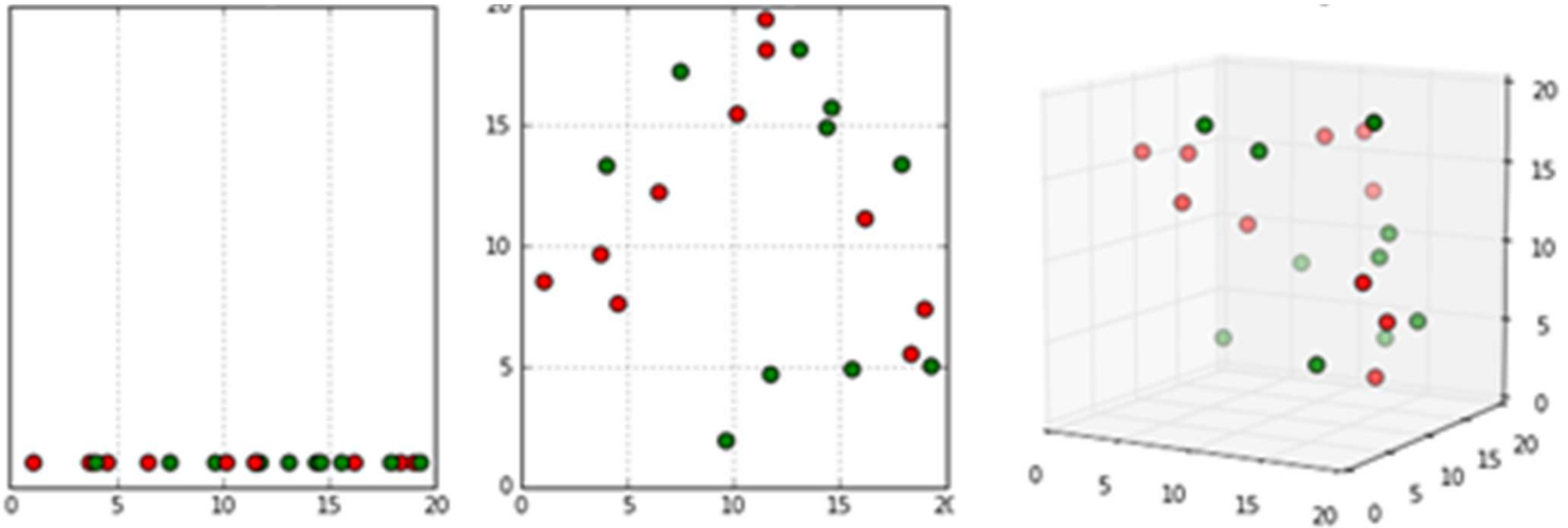
Volume of space increases

↓

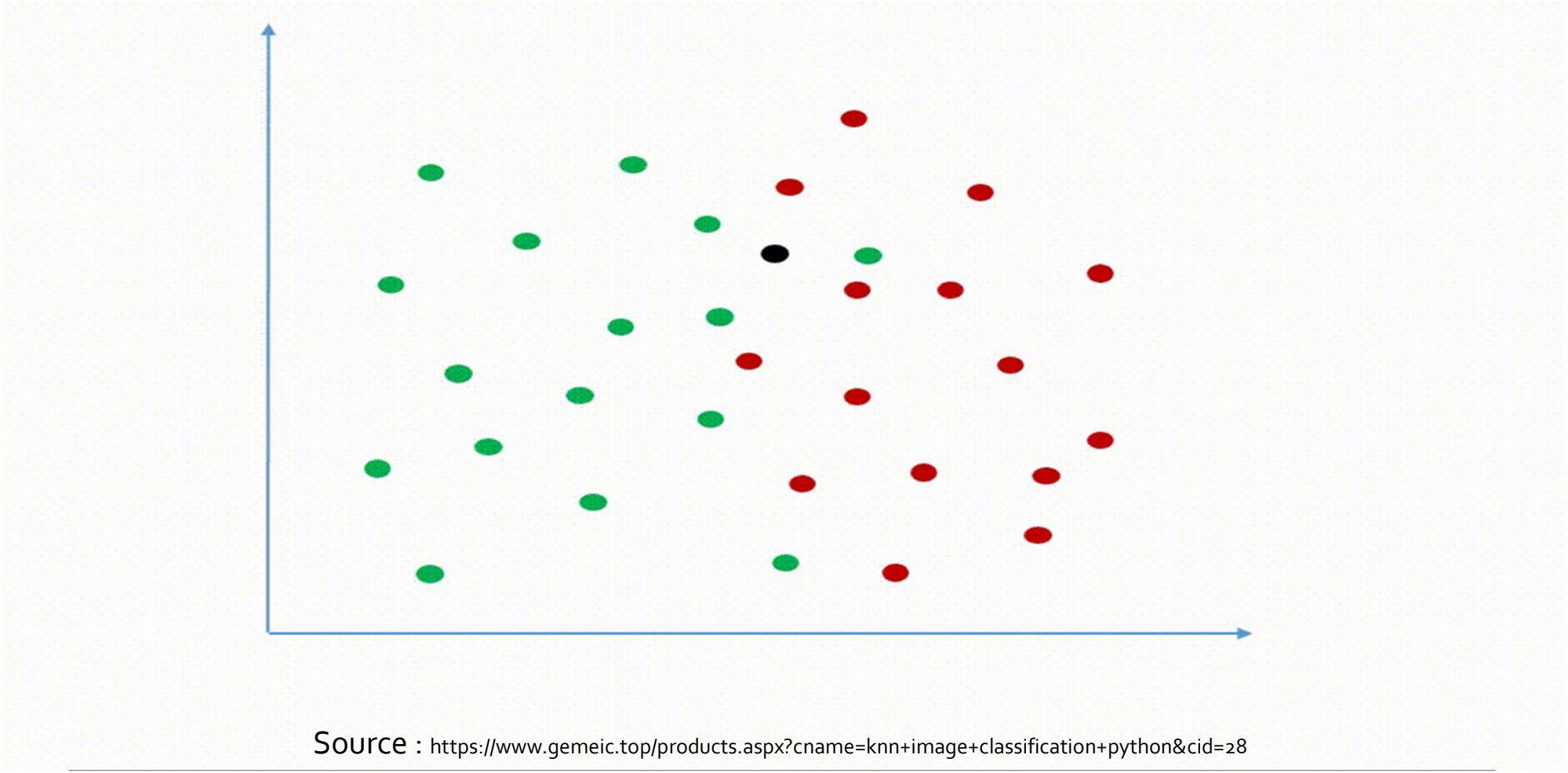
Data become sparse

Source : Conference of European statistics Stakeholders – Budapest, 2016

For Example : KNN



Source : https://deepai.org/machine-learning-glossary-and-terms/curse-of-dimensionality



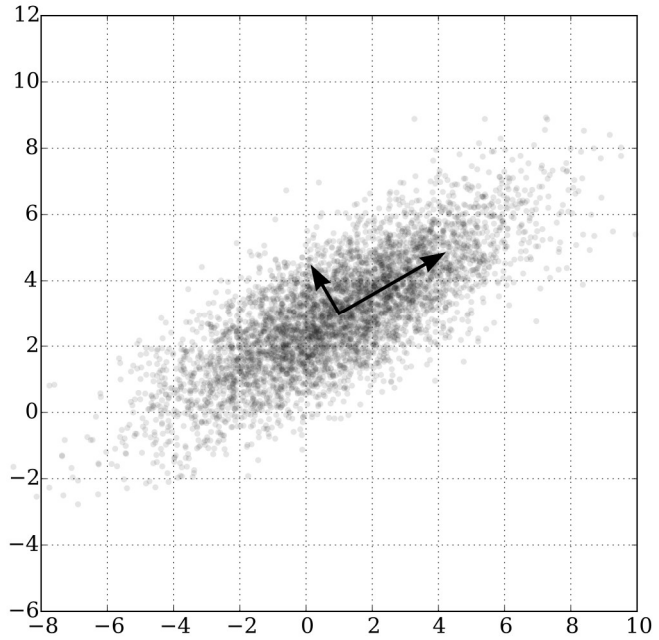
Source : https://www.gemeic.top/products.aspx?cname=knn+image+classification+python&cid=28

5. Tackling Methods

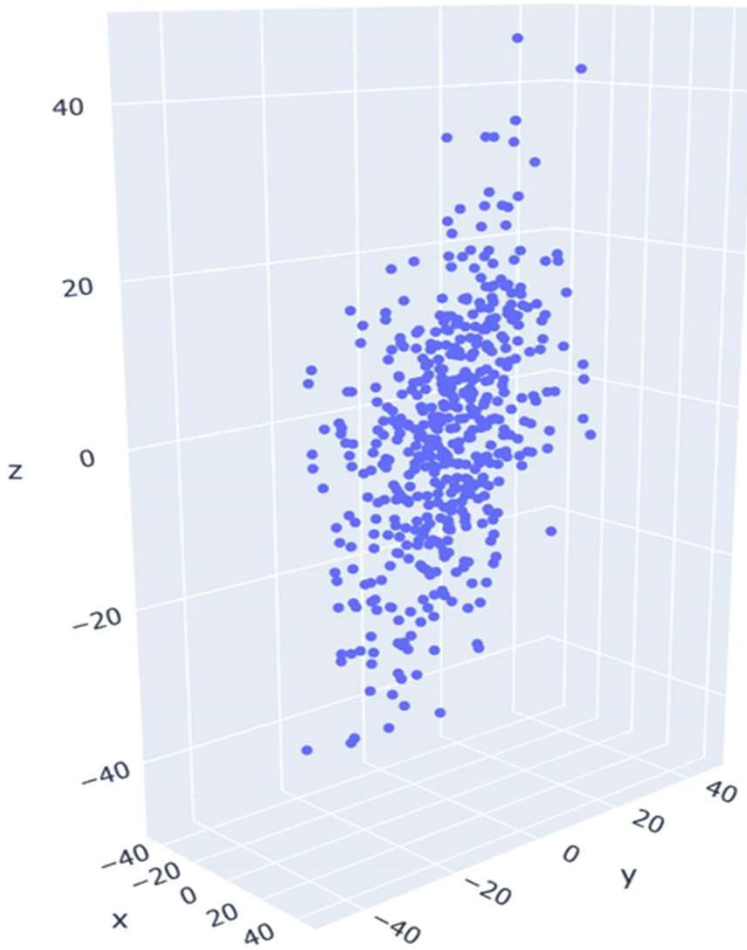
Principal Component Analysis (PCA)

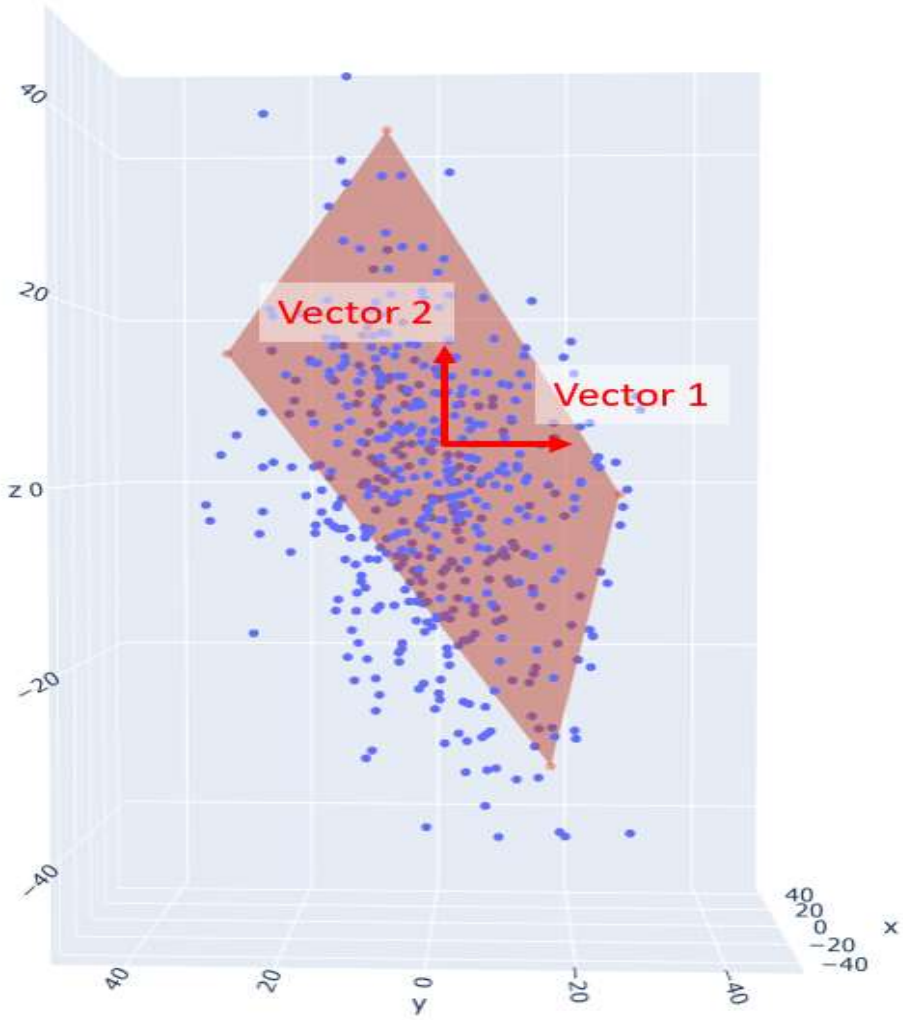
- Statistical procedure
- Use orthogonal transformation

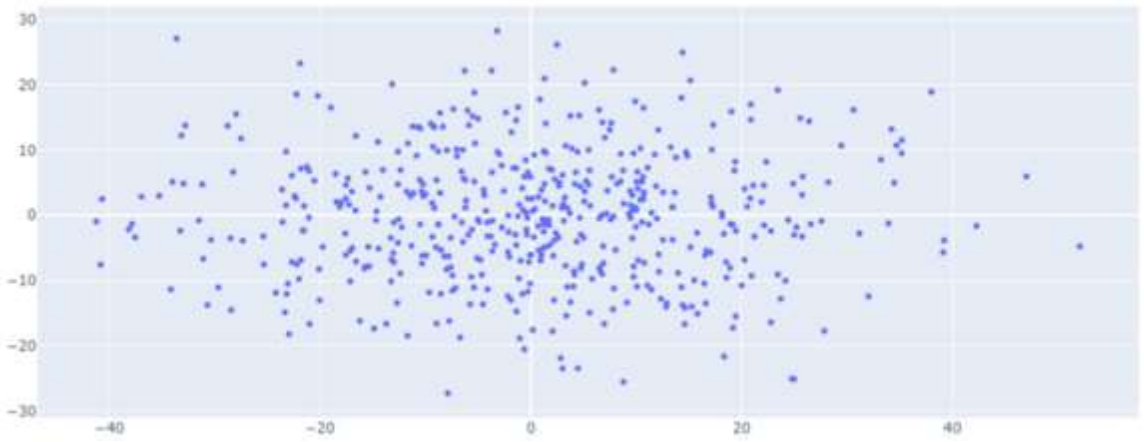
- How to implement for outlier analysis ?
  - Project most important dimensions
  - Combine selected dimensions and reduce size of data set
  - Simplify dataset by analysing by structure or observation and associated dimensions

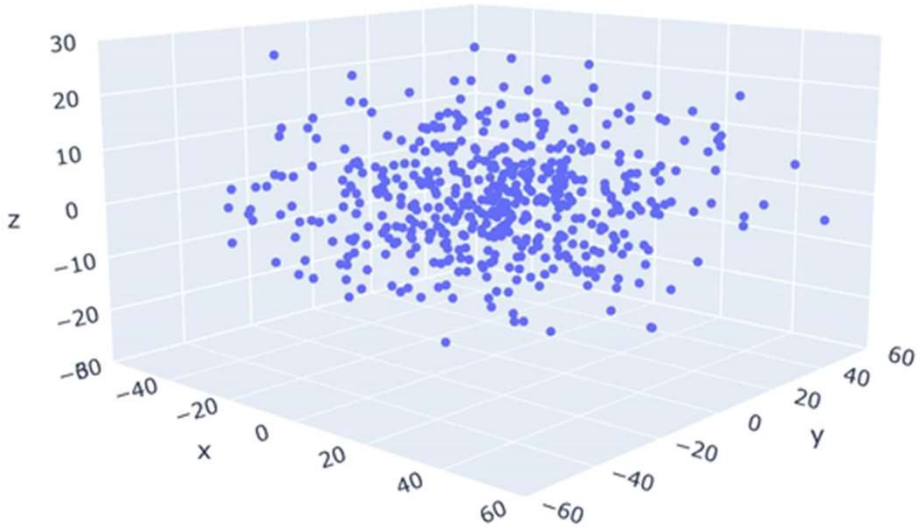


Source : https://en.wikipedia.org/wiki/Principal_component_analysis


40
20
0
−20
−40
z
−40
−20
0
20
40
x
−40
−20
0
20
40
y


Vector 2
Vector 1
40
20
0
−20
−40
z
40
20
0
−20
−40
y
40
20
0
−20
−40
x

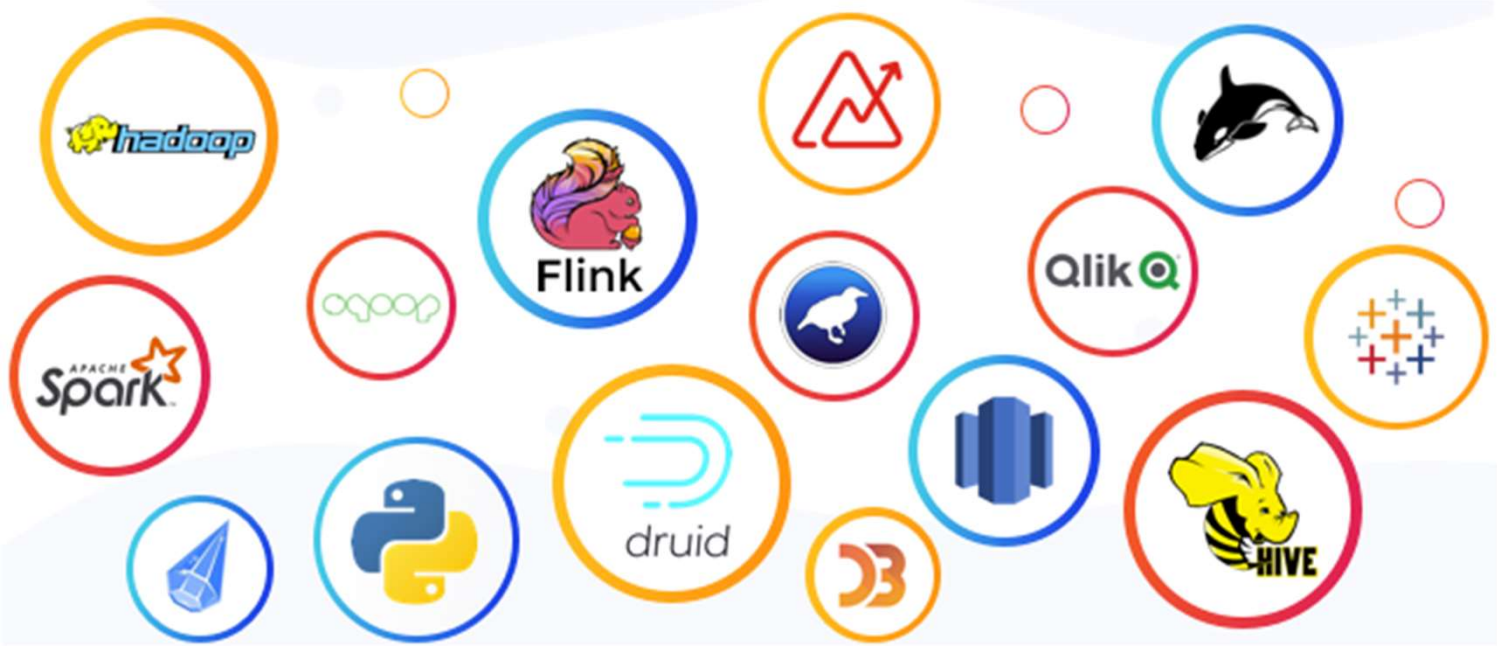

30
20
10
0
−10
−20
−30
−40
−20
0
20
40


z
30
20
10
0
−10
−20
−80
−40
−20
0
20
40
60
x
−60
−40
−20
0
20
40
60
y

5. Tools/ Application/ Frameworks

Tools Overview

- Apache Hadoop
- Apache Spark
- Apache Storm
- Apache Flink
- MXNet



Summary

- no generic approach for big data anomaly detection
- the problem of high dimensionality is inevitable in many application areas
- computationally more complex as the volume (and Velocity) of data increases

Suggestions:

- Improve balance between performance and accuracy of anomaly detection

Thank You !!!

!!! Q & A Session !!!