数据挖掘导论

Introduction to Data Mining

Outlier Detection

数据智能实验室
DATA INTELLIGENCE LABORATORY
D I-LAB

浙江大学
Zhejiang University

Agenda

- Concepts and Applications
- Statistical Methods
- Graphical Method
- Density-Based Methods
- Isolation Tree
- Summary

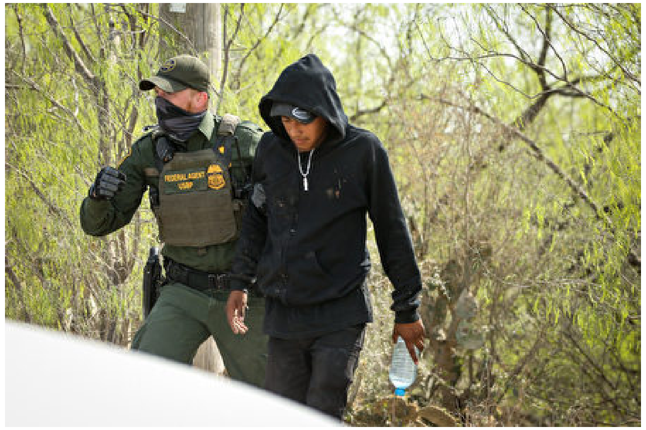
Outlier Detection

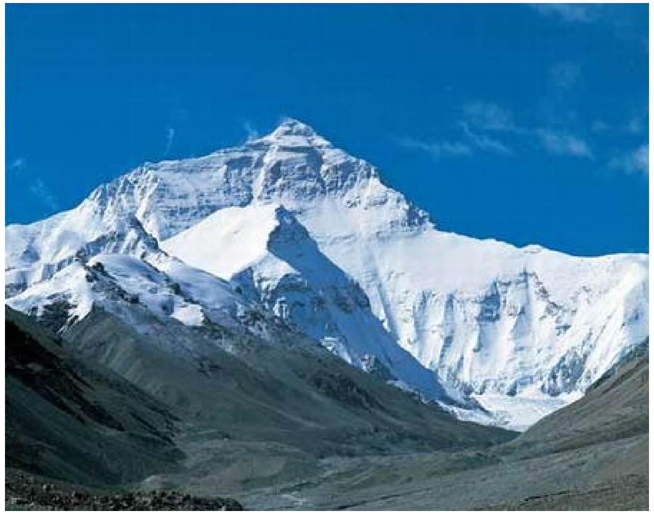
- **Also called anomaly detection**
  - ✓ Identify objects that are different from most other objects
- **Causes of anomalies**
  - ✓ Data from different class of object or underlying mechanism, e.g., rare disease, fraud detection
  - ✓ Natural variation e.g., tails on a Gaussian distribution
  - ✓ Data measurement or collection errors

Image source: https://blog.paperspace.com/anomaly-detection-isolation-forest/

Applications – Spam/Fraud Detection

Applications – Surveillance Outlier Event





Applications – Detour Detection

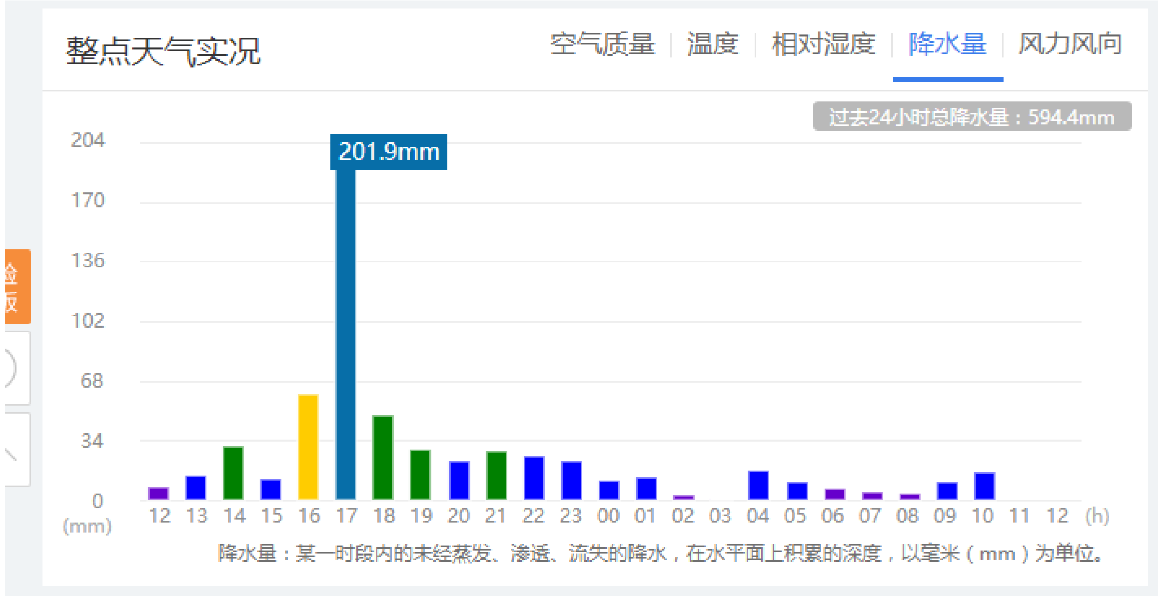
一个多小时行程结束以后，吴小姐傻眼了，滴滴系统发来的账单显示吴小姐要**支付540多万元的打车费用**，而打车时间显示的则是**1000多万分钟**。

图片来源

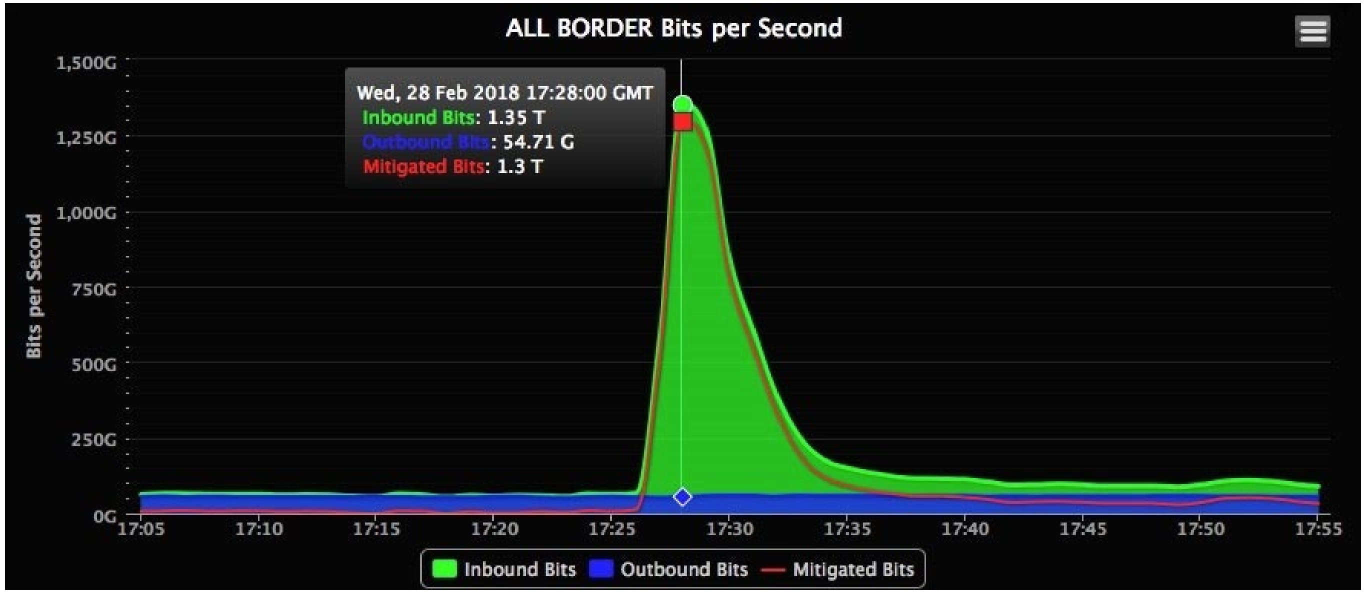
醉酒乘客打滴滴回家 司机疯狂绕路带其"环游"1小时

数据智能实验室 DATA INTELLIGENCE LABORATORY
Zhejiang University

Applications – Time Series Outlier Detection



国家防总：7月20日郑州最大小时降雨量达201.9毫米突破历史极值



GitHub Survived the Biggest DDoS Attack Ever Recorded

Point Anomaly

- An individual data instance is anomalous with respect to the data

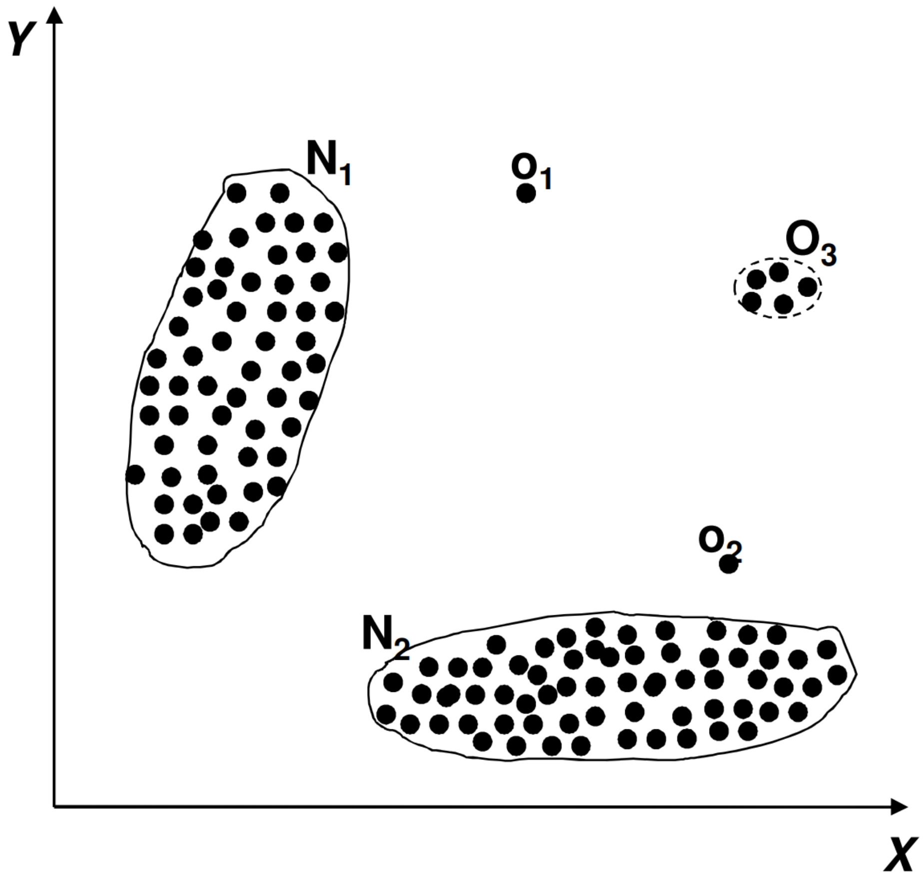
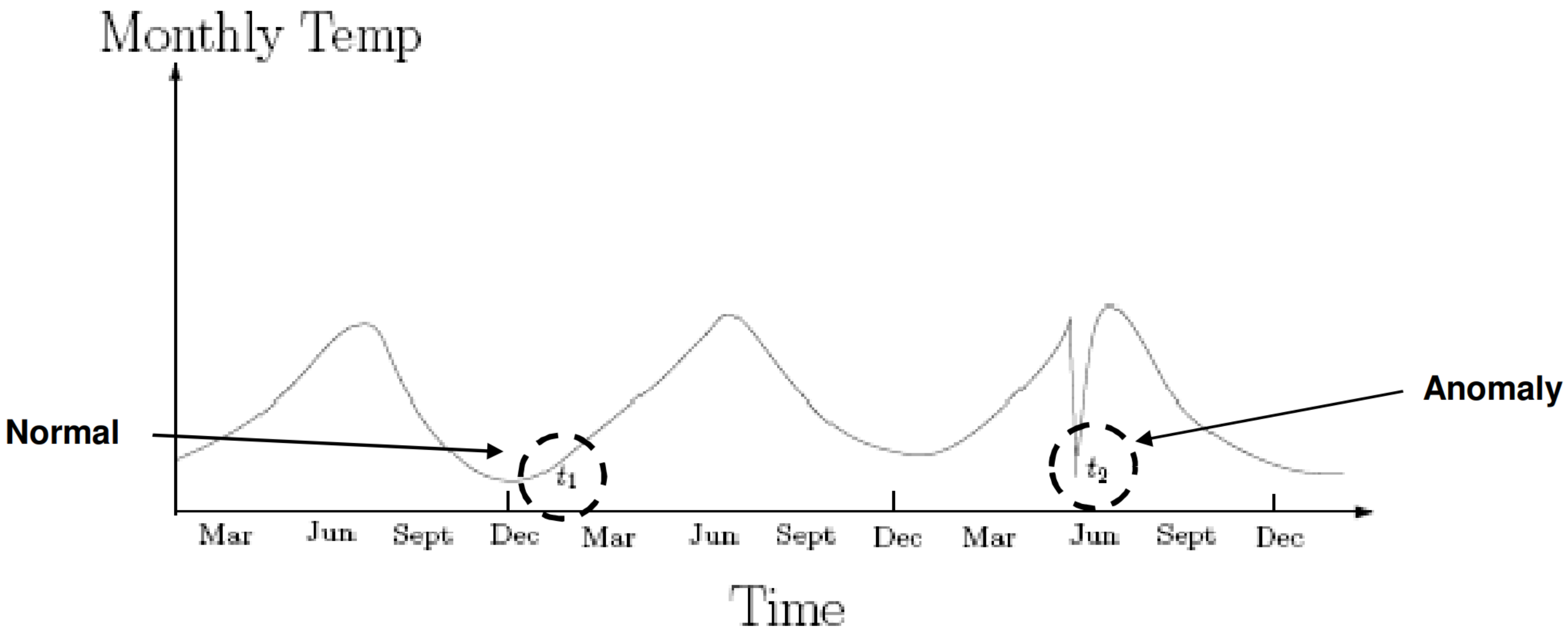


Image source: http://courses.washington.edu/css581/lecture_slides/18_anomaly_detection.pdf

Contextual Anomaly

- An individual data instance is anomalous within a context
- Dinosaurs are common in cretaceous, but anomalous nowadays



Agenda

- Concepts and Applications
- Statistical Methods
- Graphical Method
- Density-Based Methods
- Isolation Tree
- Summary

Statistical Outlier Detection

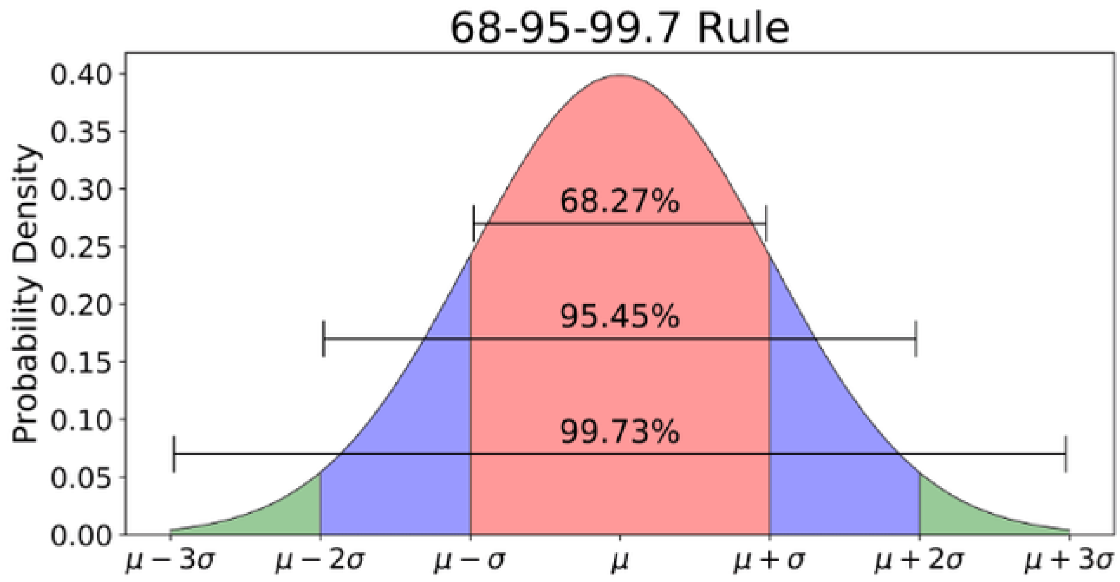
Core idea

- ✓ Fit the data with a probabilistic model
- ✓ Outliers are instances with low probability

Univariate Gaussian Distribution

- ✓ E.g., the height of students in ZJU
- ✓ Outlier defined by z-score > threshold

$$\boxed{z = \frac{x - \mu}{\sigma}}$$



The 68-95-99.7 Rule for a Normal Distribution

Grubbs' Test

- **Example: Given sampled data points 5, 10, 9.5, 9.8, 9.9, can we reject the outlier with 95% confidence?**
  - ✓ Find the mean ($\bar{x}$) and standard deviation of the data set ($\bar{x}$=8.84, σ=2.119)
  - ✓ Calculate z-score = |5-8.84|/2.119=1.812
  - ✓ Calculate the G-critical value, usually lookup from table (1.67 in this example)
  - ✓ Reject the outlier if the test statistic is greater than G-critical value

	Alpha				
N	0.1	0.075	0.05	0.025	0.01
3	1.15	1.15	1.15	1.15	1.15
4	1.42	1.44	1.46	1.48	1.49
5	1.6	1.64	1.67	1.71	1.75
6	1.73	1.77	1.82	1.89	1.94
7	1.83	1.88	1.94	2.02	2.1
8	1.91	1.96	2.03	2.13	2.22
9	1.98	2.04	2.11	2.21	2.32

Kernel Density Estimation

- **Estimate the PDF of a random variable in a non-parametric way**
  - ✓ It's related to a histogram but with a data smoothing technique.

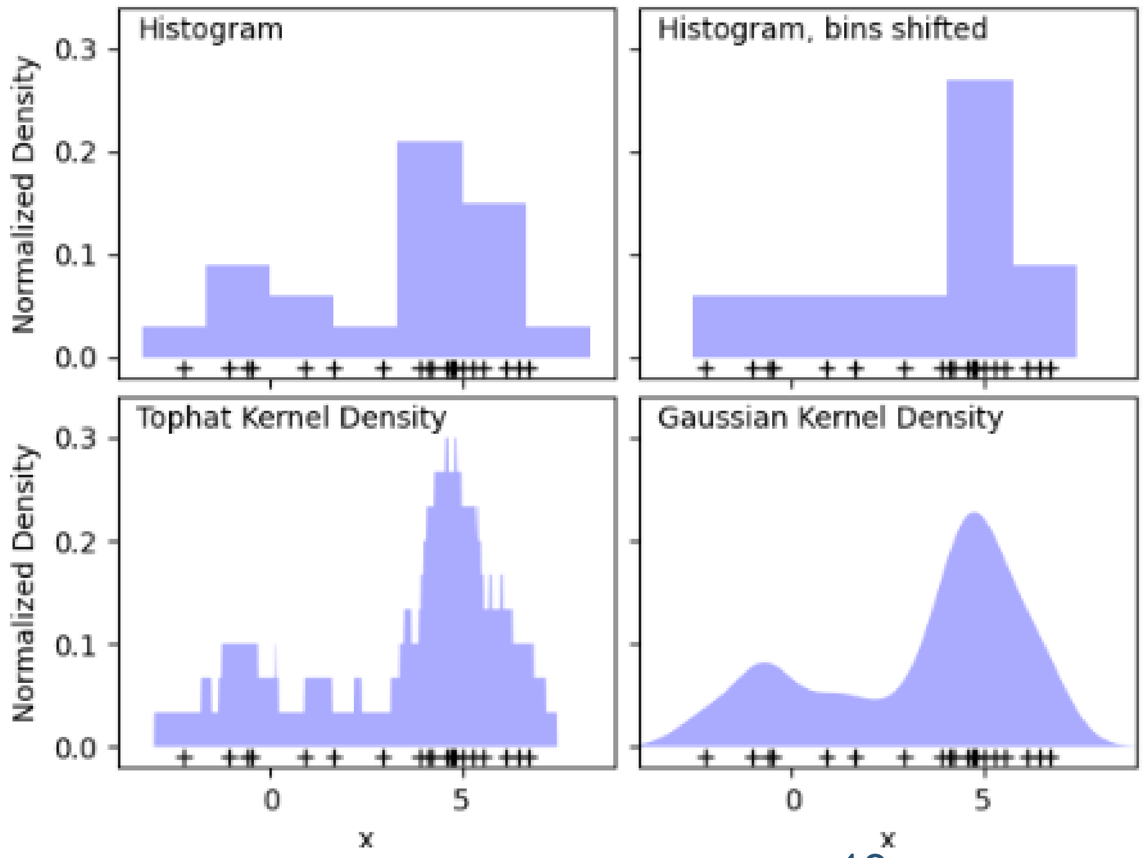


Image source: https://towardsdatascience.com/kernel-density-estimation-for-anomaly-detection-in-python-part-1-452c5d4c32ec

Kernel Density Estimation

On every data point x_i, we place a kernel function K. The kernel density estimate is

$$\hat{f}(x) = \frac{1}{N}\sum_{i=1}^{N} K(x - x_i).$$

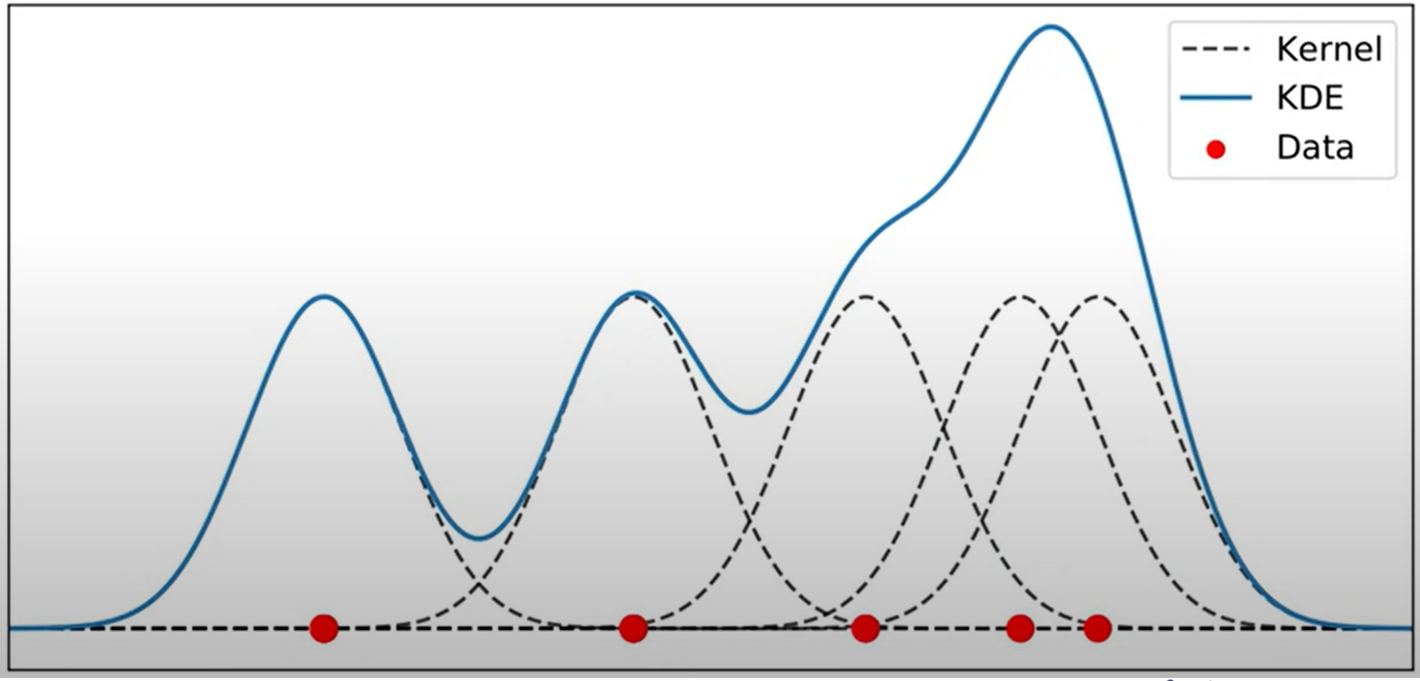


Image source: https://www.youtube.com/watch?v=x5zLaWT5KPs

Agenda

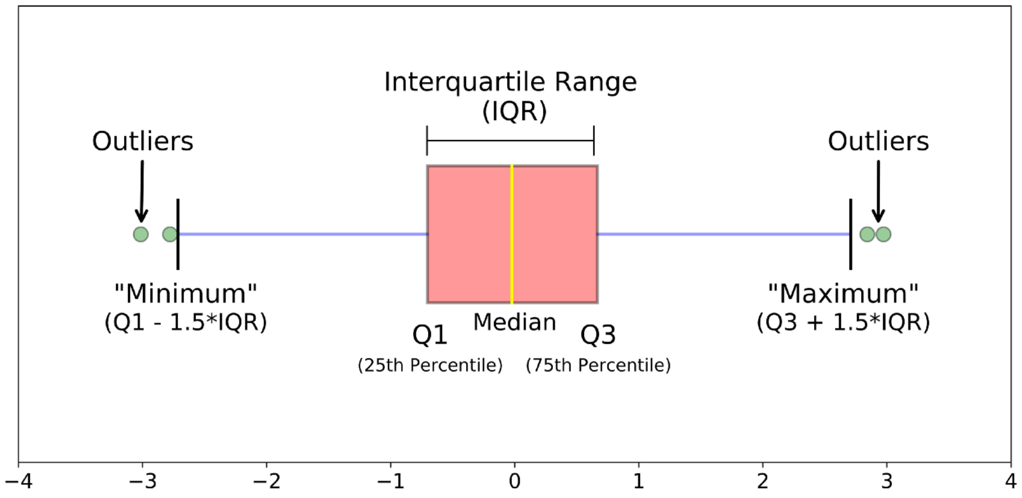
- Concepts and Applications
- Statistical Methods
- **Graphical Method**
- Density-Based Methods
- Isolation Tree
- Summary

Graphical Outlier Detection

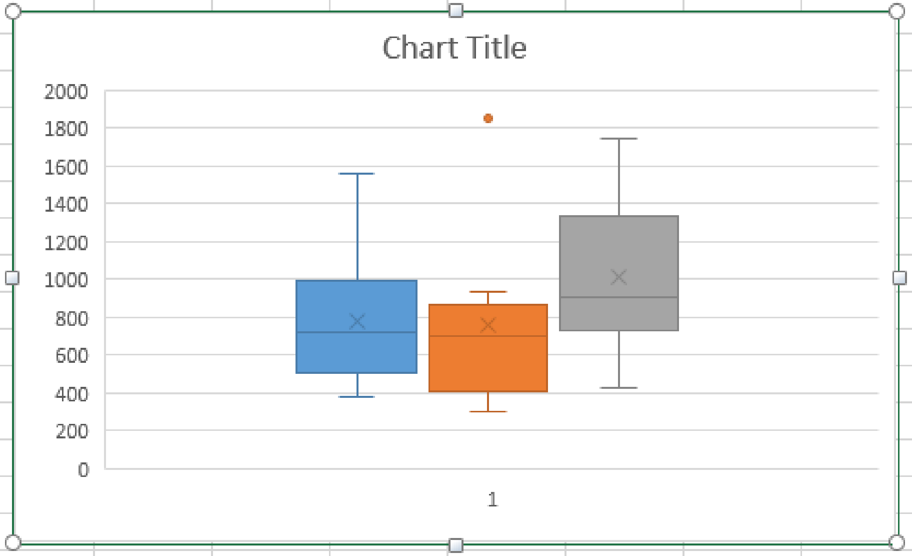
Graphical approach to outlier detection

- ✓ Look at a plot of the data
- ✓ Human decides if data is an outlier

Box plot



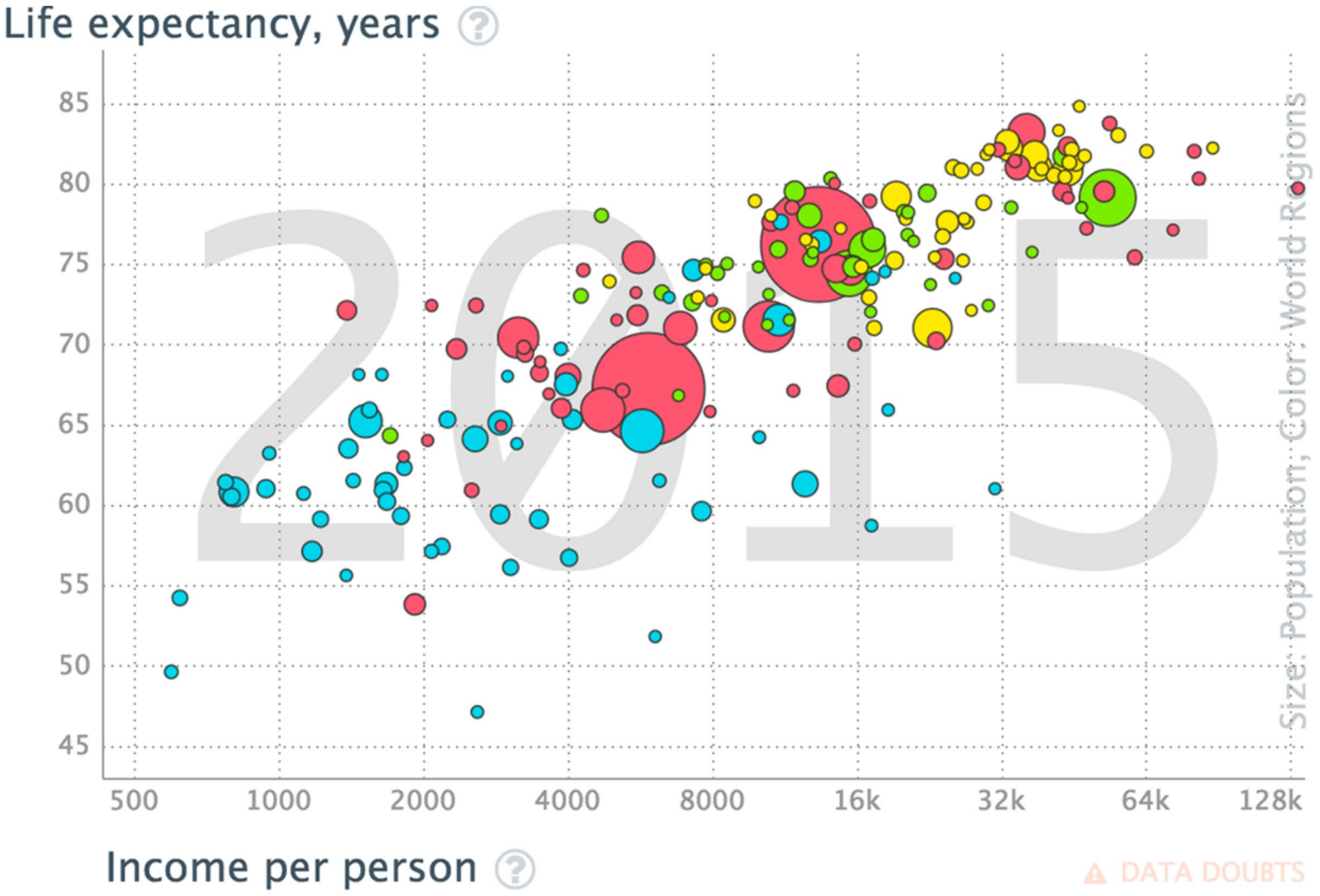
	Brand A	Brand B	Brand C
1	Brand A	Brand B	Brand C
2	1020	840	1430
3	1560	940	1750
4	560	780	870
5	780	650	920
6	990	720	1300
7	670	430	890
8	510	1850	740
9	490	300	720
10	380	360	430
11	880	690	1050



Graphical Outlier Detection

☐ Scatter Plot Array

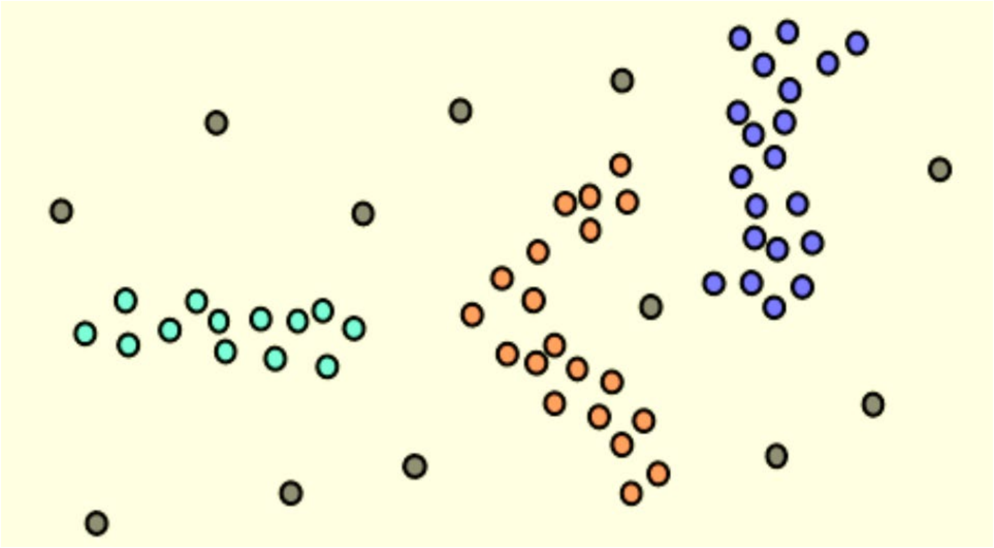
✓ Shows how multiple variables are related. The matrix can also identify outliers in multiple scatter plots

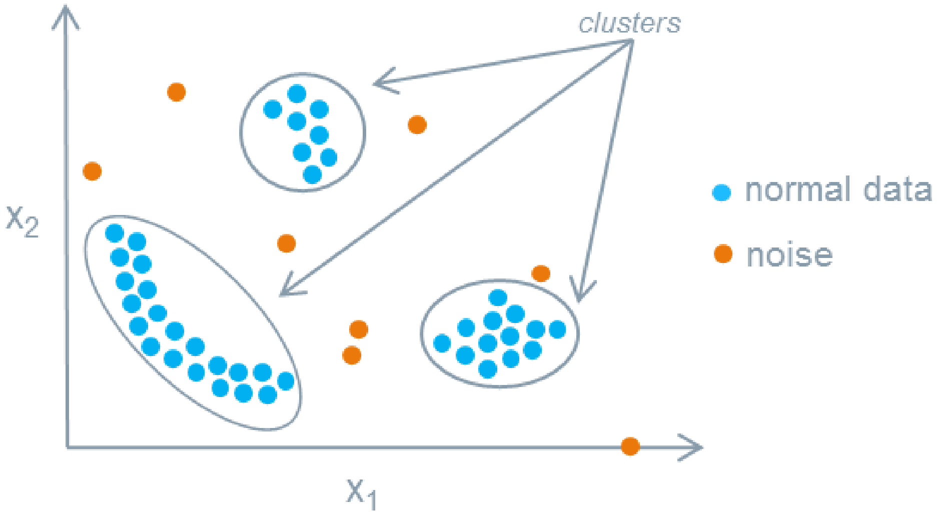


Agenda

- Concepts and Applications
- Statistical Methods
- Graphical Method
- **Density-Based Methods**
- Isolation Tree
- Summary

DBSCAN





Local Outlier Factor (LOF)

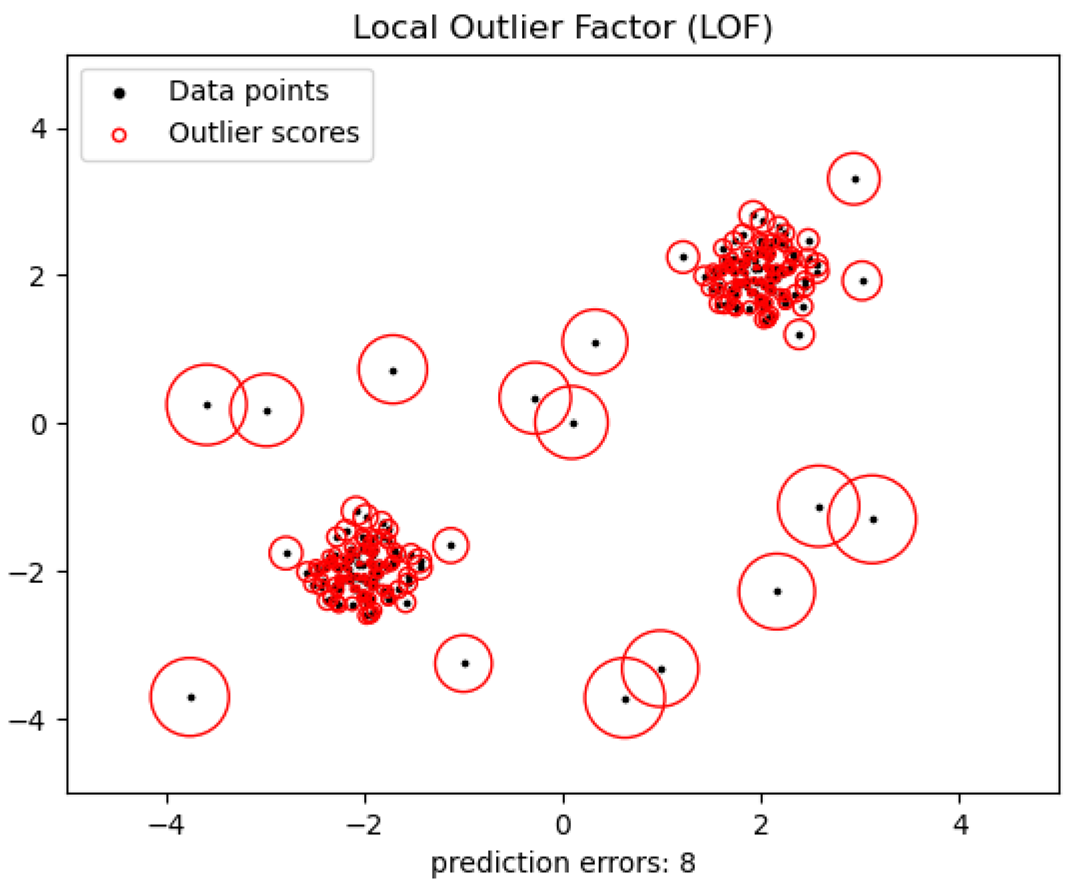


Image source: https://scikit-learn.org/stable/modules/outlier_detection.html

Local Outlier Factor (LOF)

- **K-distance(x)**
  - ✓ The distance of object x to its k-th nearest neighbor
- **Reachability Distance (RD)**
  - ✓ The maximum of the distance of two points and the k-distance of the second point
  - ✓ RD(a , b) = max{ k-distance (b) , distance (a , b) }
- **Local Reachability Density (LRD)**
  - ✓ It refers to how far we need to go from the point we are at to reach the next point or set of points.
  - ✓ LRD (a) = 1 / (sum (RD(a , n)) / k), where n is a member in the k nearest neighbors

Local Outlier Factor (LOF)

Local Outlier Factor

- ✓ Average ratio of LRD of neighbors of p and LRD of p
- ✓ LOF (p) =[(LRD (1st. neighbor) + LRD (2nd. neighbor) + ….. + LRD (kth. neighbor)) / LRD (p)] / k

Property of LOF

- ✓ A point is an outlier if its LOF >> 1

LOF(k) ~ 1 means **Similar density as neighbors,**

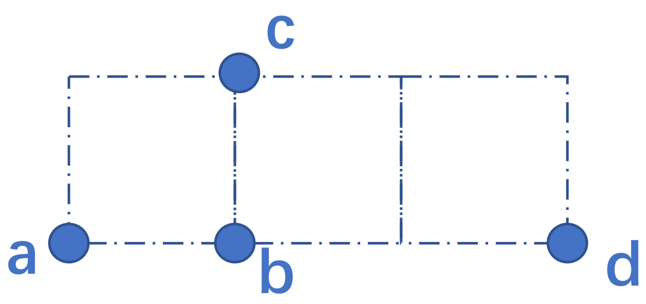
LOF(k) < 1 means **Higher density than neighbors**

LOF(k) > 1 means **Lower density than neighbors**

Local Outlier Factor (LOF)

Example of LOF calculation

- ✓ Given 4 points a(0,0), b(0,1), c(1,1) and d(3,0)
- ✓ Let k=2 and use Manhattan distance
- ✓ Calculate the LOF for each point and show the top-1 outlier

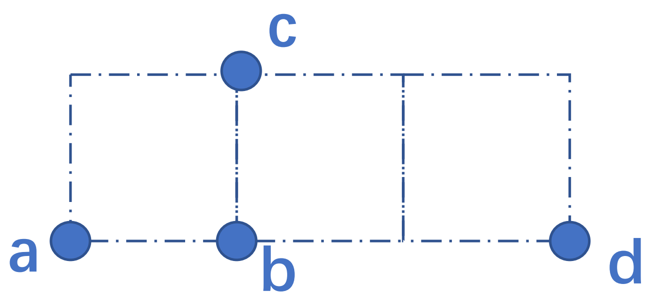


Example source: https://cse.hkust.edu.hk/~leichen/courses/comp5331/lectures/LOF_Example.pdf

Local Outlier Factor (LOF)

- Step1: calculate all the distances between each two data points

dist(a, b) = 1
dist(a, c) = 2
dist(a, d) = 3
dist(b, c) = 1
dist(b, d) = 3+1=4
dist(c, d) = 2+1=3



Local Outlier Factor (LOF)

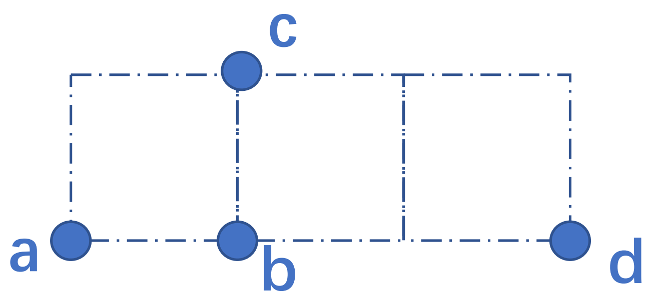
- Step2: calculate k-distance for each object

$N_2(a) = \{b, c\}$ $\quad$ $dist_2(a) = \mathbf{dist(a, c) = 2}$ *(c is the 2nd nearest neighbor)*

$N_2(b) = \{a, c\}$ $\quad$ $dist_2(b) = \mathbf{dist(b, a) = 1}$ *(a/c is the 2nd nearest neighbor)*

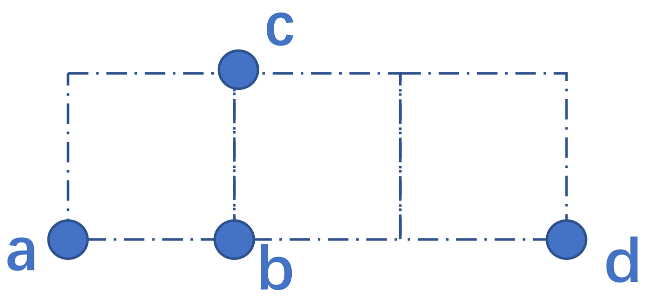
$N_2(c) = \{b, a\}$ $\quad$ $dist_2(c) = \mathbf{dist(c, a) = 2}$ *(a is the 2nd nearest neighbor)*

$N_2(d) = \{a, c\}$ $\quad$ $dist_2(d) = \mathbf{dist(d, a) = 3}$ *(a/c is the 2nd nearest neighbor)*



Local Outlier Factor (LOF)

- Step3: calculate RD and LRD for all the objects



RD(a,b)=1

RD(a,c)=2

LRD(a)=2/(1+2)=0.667

RD(b,a)=2

RD(b,c)=2

LRD(b)=2/(2+2)=0.5

RD(c,a)=2

RD(c,b)=1

LRD(c)=2/(1+2)=0.667

RD(d,a)=3

RD(d,c)=3

LRD(d)=2/(3+3)=0.333

Local Outlier Factor (LOF)

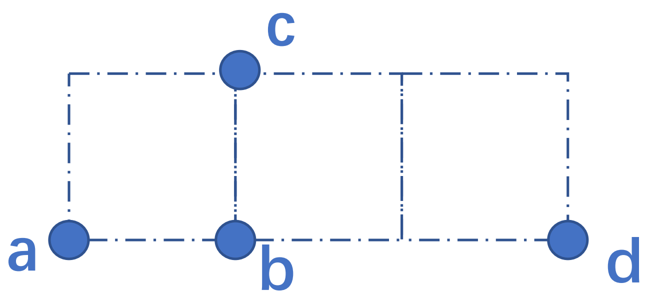
- Step4: calculate LOF for all the objects

LOF(a)=[(0.5+0.667)/0.667]/2=0.875

LOF(b)=[(0.667+0.667)/0.5]/2=1.334

LOF(c)=[(0.667+0.5)/0.667]/2=0.875

LOF(d)=[(0.667+0.667)/0.333]/2=2 → outlier



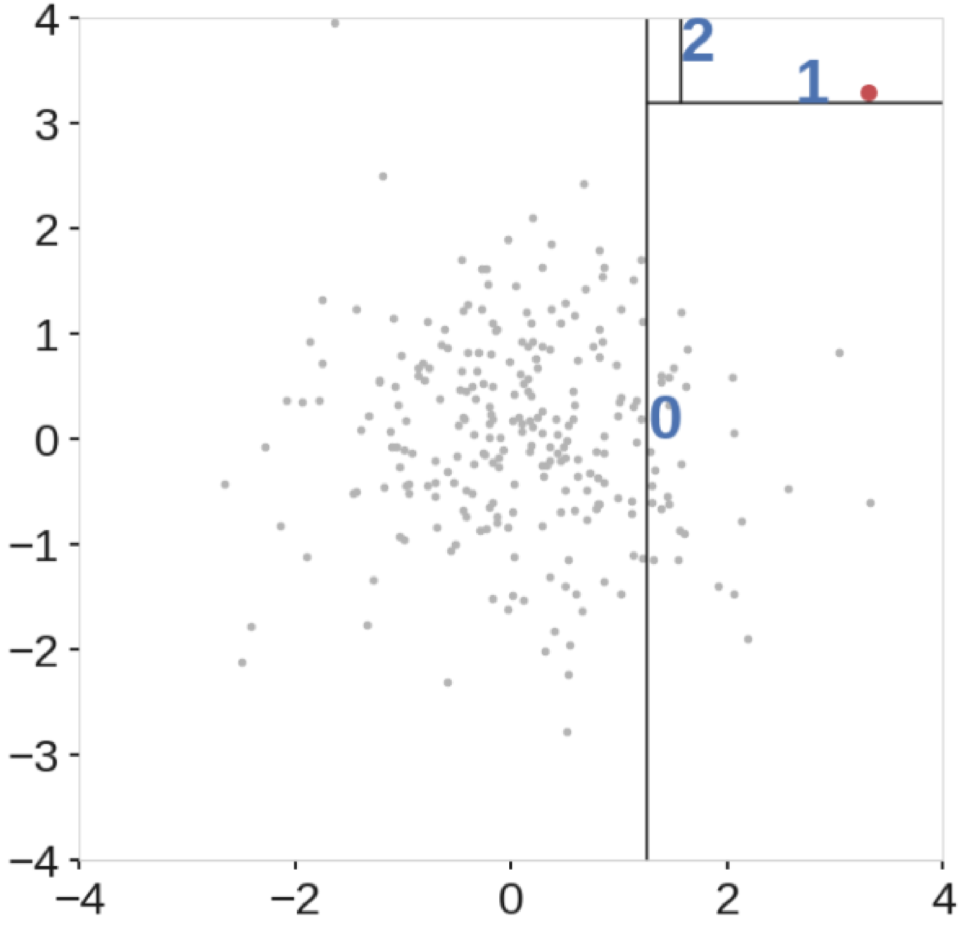
Agenda

- Concepts and Applications
- Statistical Methods
- Graphical Method
- Density-Based Methods
- Isolation Tree
- Summary

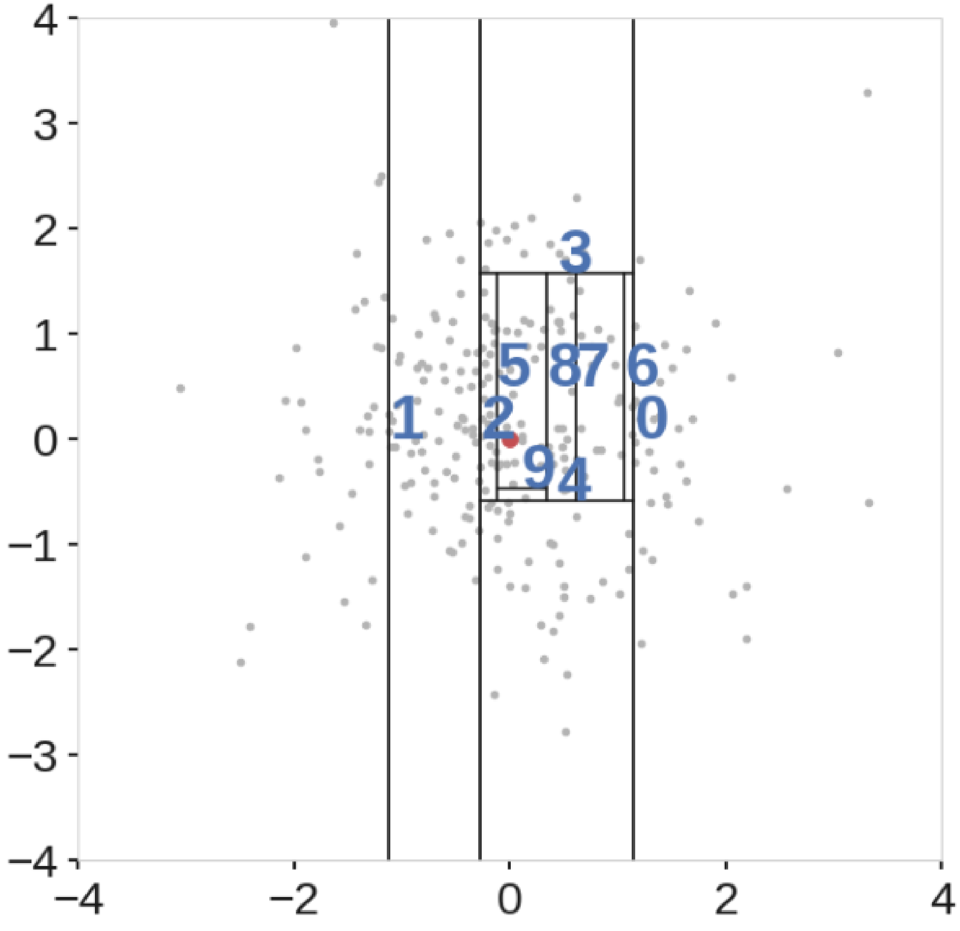
Isolation Tree

- **An unsupervised learning algorithm based on decision tree**
  - ✓ Builds decision trees with random partitioning
- **Build a decision tree with random partitioning**
  - ✓ Randomly select a feature from the given set of features
  - ✓ Randomly select a split value between the max and min values of that feature
- **Such random partitioning produces shorter paths in trees for the anomalous data points**
- **When a forest of random trees collectively produce shorter path lengths for particular samples, they are highly likely to be anomalies**

Isolation Tree



(a) Anomaly point



(b) Nominal point

Isolation Tree

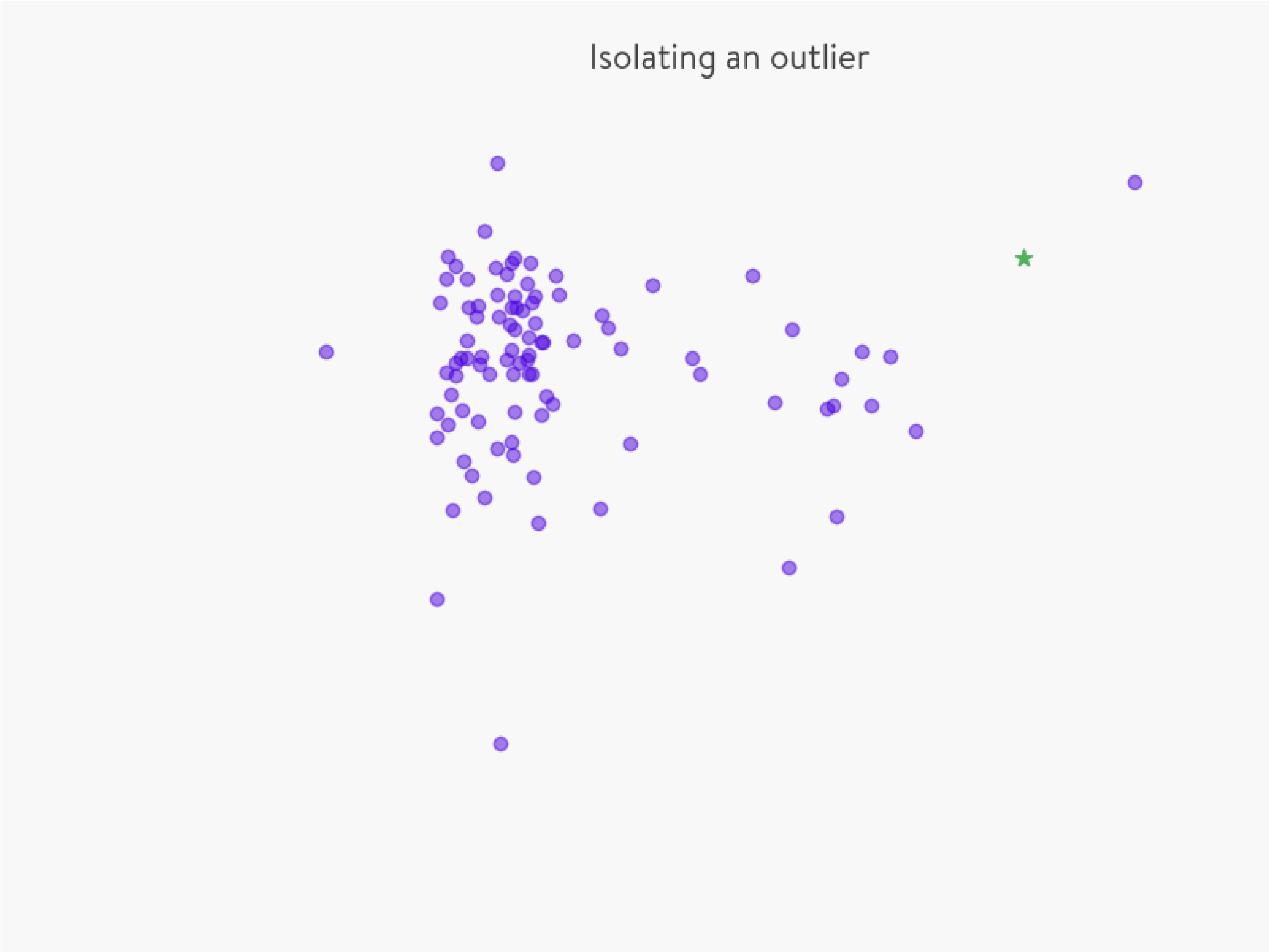


Image source: https://deepnote.com/@christopher-hui/Anomaly-Detection-Analysis-Isolation-Forest-c012da68-8081-4e2e-9bc8-8bc59a1c2d6c

Agenda

- ☐ Concepts and Applications
- ☐ Statistical Methods
- ☐ Graphical Method
- ☐ Density-Based Methods
- ☐ Isolation Tree
- ☐ Summary