# 异常检测
# Anomaly Detection

电子信息与通信学院 冯 镔
fengbin@hust.edu.cn

# 常见应用场景

- 金融领域
  - 信用卡欺诈、虚假信贷
- 网络安全
  - 网络入侵模式
- 电商领域
  - 羊毛党、恶意刷单
- 生态预警
  - 极端天气

# Anomaly Detection

- What are anomalies/outliers?
  - The set of data points that are considerably different than the remainder of the data
- Variants of Anomaly Detection Problems
  - Given a database D, find all the data points $x \in D$ with anomaly scores greater than some threshold t
  - Given a database D, find all the data points $x \in D$ having the top-n largest anomaly scores f(x)
- Working assumption:
  - There are considerably more "normal" observations than "abnormal" observations (outliers/anomalies) in the data

# 异常点类型

- 单点异常（**Global Outlier**）
  - 某个点与全局大多数点都不一样
- 上下文异常（**Contextual Outliers**）
  - 时间序列中的异常
- 集体异常（**Collective Outliers**）
  - 单独看某个个体可能并不存在异常，但这些个体同时出现，则构成了一种异常

# Anomaly Detection

- **Challenges**
  - **How many outliers are there in the data?**
  - **Method is unsupervised**
    - **Validation can be quite challenging (just like for clustering)**
  - **Finding needle in a haystack**

# Anomaly Detection Schemes

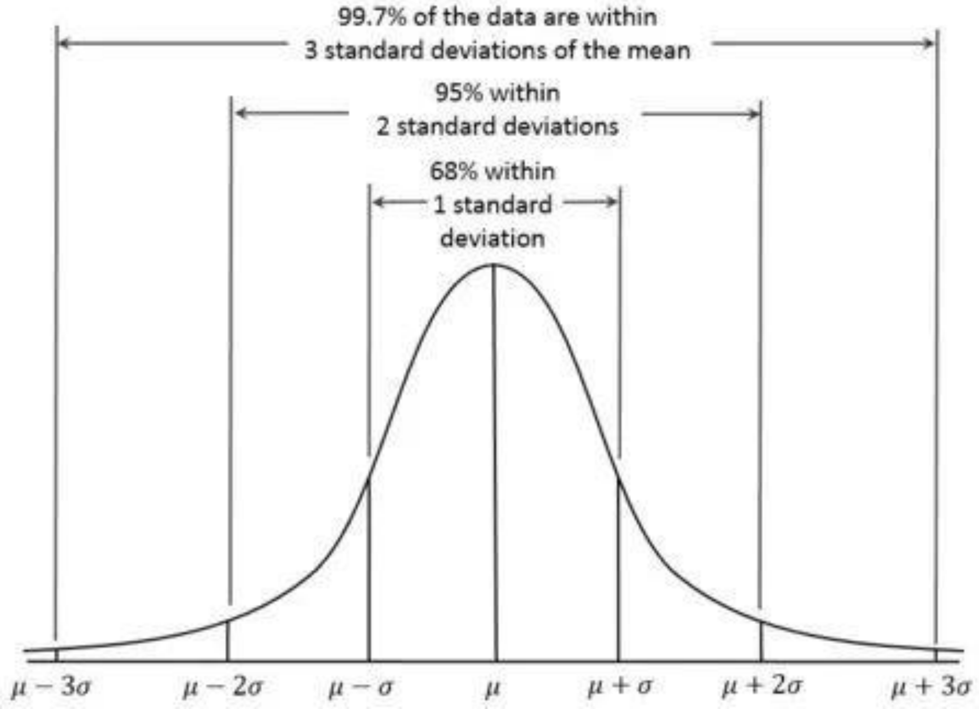
- **General Steps**
  - **Build a profile of the "normal" behavior**
    - **Profile can be patterns or summary statistics for the overall population**
  - **Use the "normal" profile to detect anomalies**
    - **Anomalies are observations whose characteristics differ significantly from the normal profile**

# Anomaly Detection Schemes

- Types of anomaly detection schemes
  - Statistical-based
  - Distance-based
  - Tree-based
  - Dimensionality reduction-based
  - Prediction-based

# Statistical Approaches

- **3sigma**
- 基于正态分布，认为超过3sigma的数据为异常点



# Statistical Approaches

- Z-score
- Z-score为标准分数，测量数据点A和平均值的距离
  - 若A与平均值相差2个标准差，Z-score为2
  - 当把Z-score=3作为阈值去剔除异常点时，便相当于3sigma

# Statistical Approaches

- Grubbs' Test
- Assume data comes from normal distribution
- Detects one outlier at a time, remove the outlier, and repeat
  - $H_0$: There is no outlier in data
  - $H_A$: There is at least one outlier
- Grubbs' test statistic: $$G = \frac{\max\left|X - \overline{X}\right|}{s}$$
- Reject $H_0$ if: $$G > \frac{(N-1)}{\sqrt{N}} \sqrt{\frac{t^2_{(\alpha/N, N-2)}}{N - 2 + t^2_{(\alpha/N, N-2)}}}$$

# Statistical Approaches

- Likelihood Approach
- Assume the data set D contains samples from a mixture of two probability distributions:
  - M (majority distribution)
  - A (anomalous distribution)
- General Approach:
  - Initially, assume all the data points belong to M
  - Let $L_t(D)$ be the log likelihood of D at time t
  - For each point $x_t$ that belongs to M, move it to A
    - Let $L_{t+1}(D)$ be the new log likelihood.
    - Compute the difference, $\Delta = L_t(D) - L_{t+1}(D)$
    - If $\Delta > c$ (some threshold), then $x_t$ is declared as an anomaly and moved permanently from M to A

# Limitations of Statistical Approaches

- Most of the tests are for a single attribute
- In many cases, data distribution may not be known
- For high dimensional data, it may be difficult to estimate the true distribution

# Distance-based Approaches

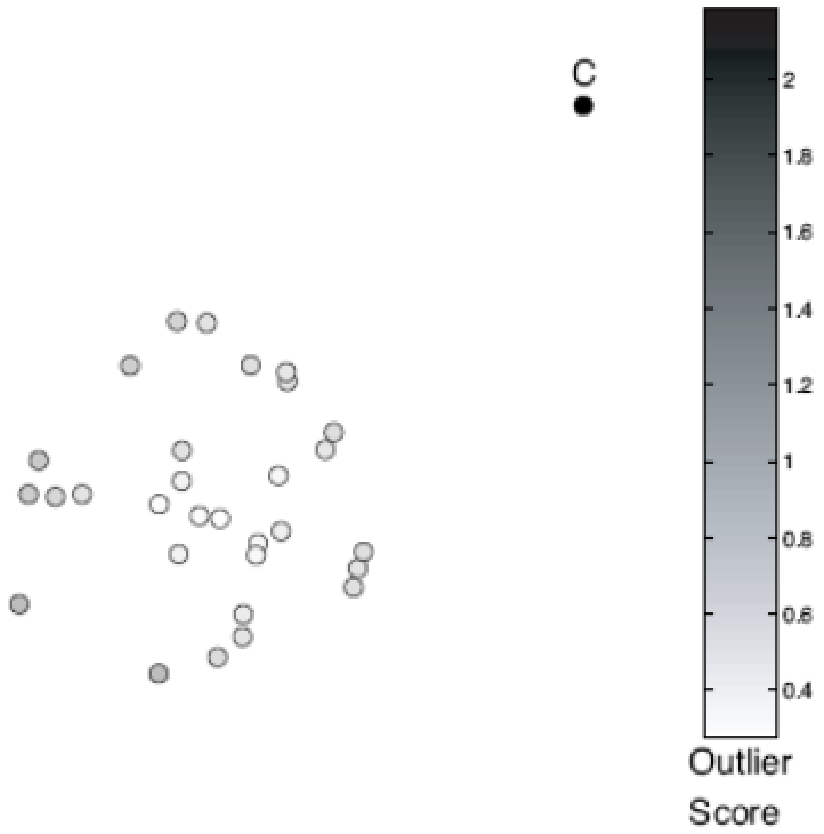
- Data is represented as a vector of features

- Three major approaches
  - Nearest-neighbor based
  - Density based
  - Clustering based

# Nearest-Neighbor Based Approach

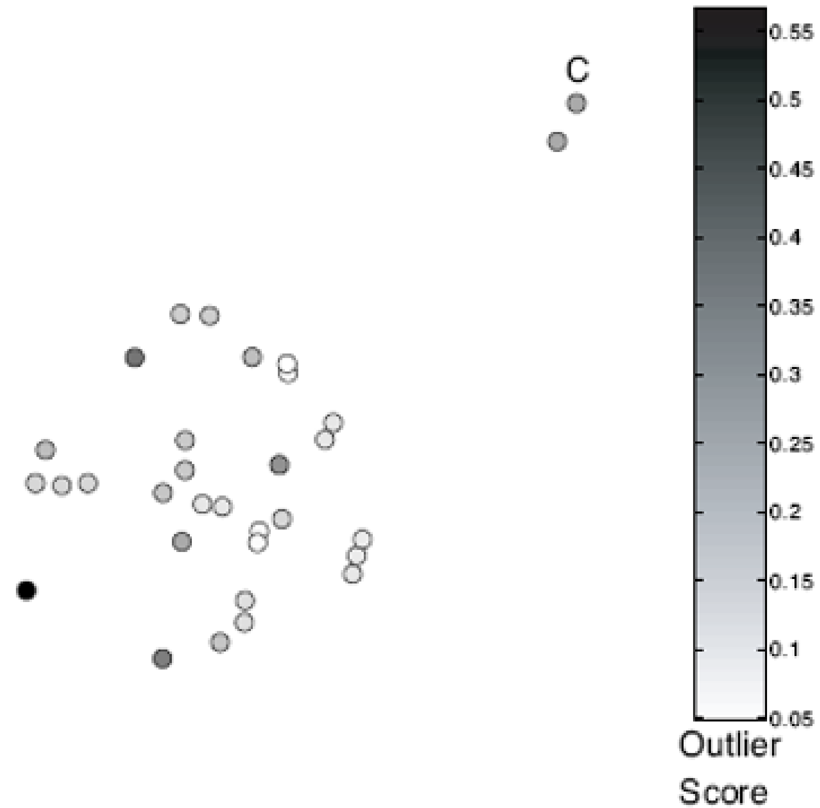
- **Approach:**
  - **Compute the distance between every pair of data points**
- **There are various ways to define outliers:**
  - **Data points for which there are fewer than *p* neighboring points within a distance *D***
  - **The top *n* data points whose distance to the *k*th nearest neighbor is greatest**
  - **The top *n* data points whose average distance to the *k* nearest neighbors is greatest**

# K Nearest-Neighbor Distance



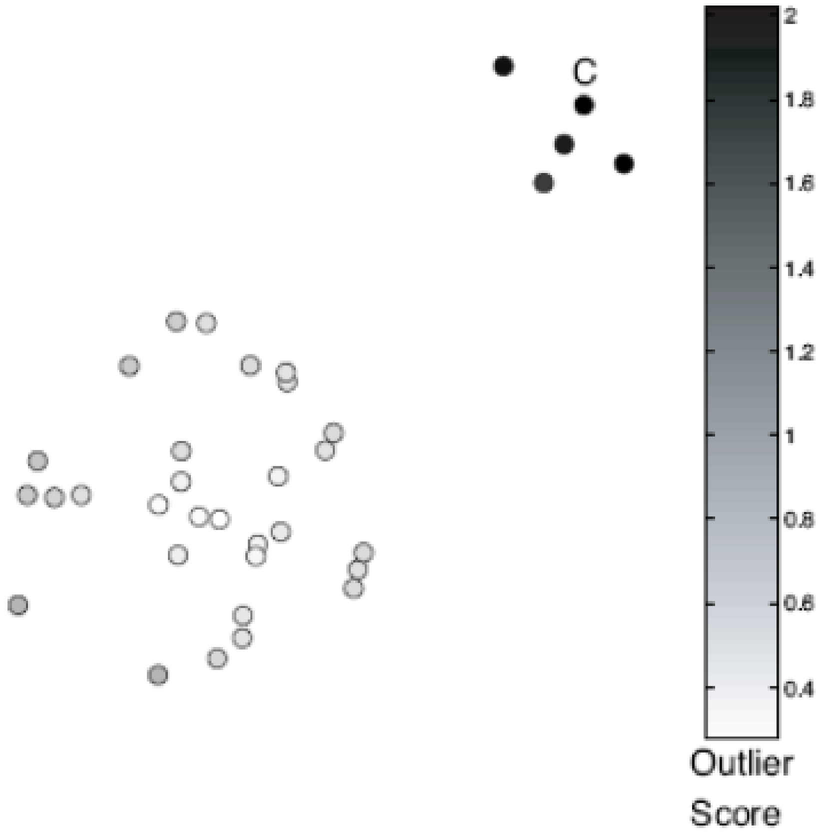

基于到第五个最近邻距离的离群点得分

# K Nearest-Neighbor Distance



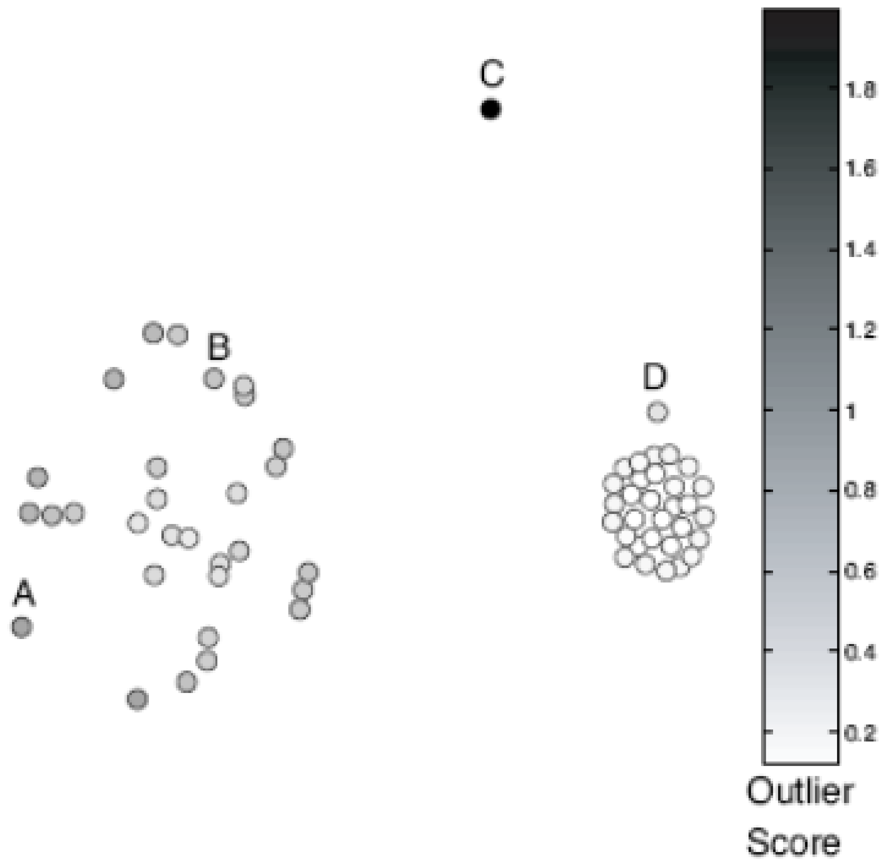

**1-最近邻导致较低的离群点得分**

# K Nearest-Neighbor Distance




**5-最近邻导致整簇变成离群点**

# K Nearest-Neighbor Distance




固定阈值不能处理不同密度簇的情况

# Outliers in Lower Dimensional Projection

- **In high-dimensional space, data is sparse and notion of proximity becomes meaningless**
  - **Every point is an almost equally good outlier from the perspective of proximity-based definitions**

- **Lower-dimensional projection methods**
  - **A point is an outlier if in some lower dimensional projection, it is present in a local region of abnormally low density**

# Outliers in Lower Dimensional Projection

- Divide each attribute into $\phi$ equal-depth intervals
  - Each interval contains a fraction $f = 1/\phi$ of the records
- Consider a *k*-dimensional cube created by picking grid ranges from *k* different dimensions
  - If attributes are independent, we expect region to contain a fraction $f^k$ of the records

# Outliers in Lower Dimensional Projection

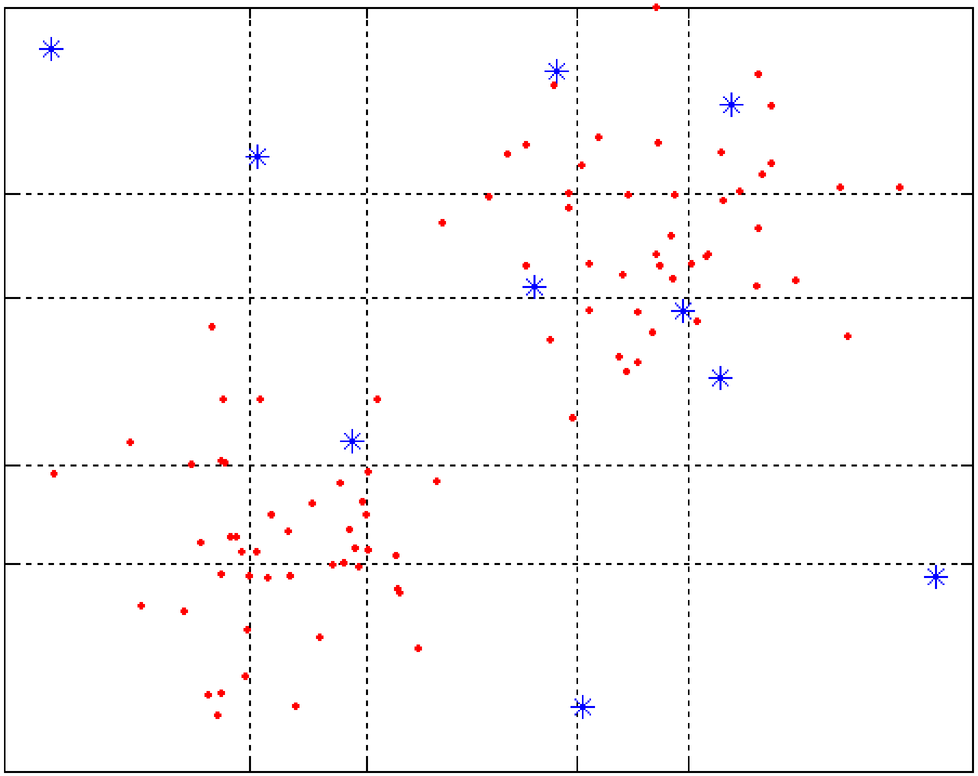
- **If there are N points, we can measure sparsity of a cube D as:**

$$S(\mathcal{D}) = \frac{n(D) - N \cdot f^k}{\sqrt{N \cdot f^k \cdot (1 - f^k)}}$$

- **Negative sparsity indicates cube contains smaller number of points than expected**

# Example

- N=100, $\phi$ = 5, f = 1/5 = 0.2, N × f² = 4

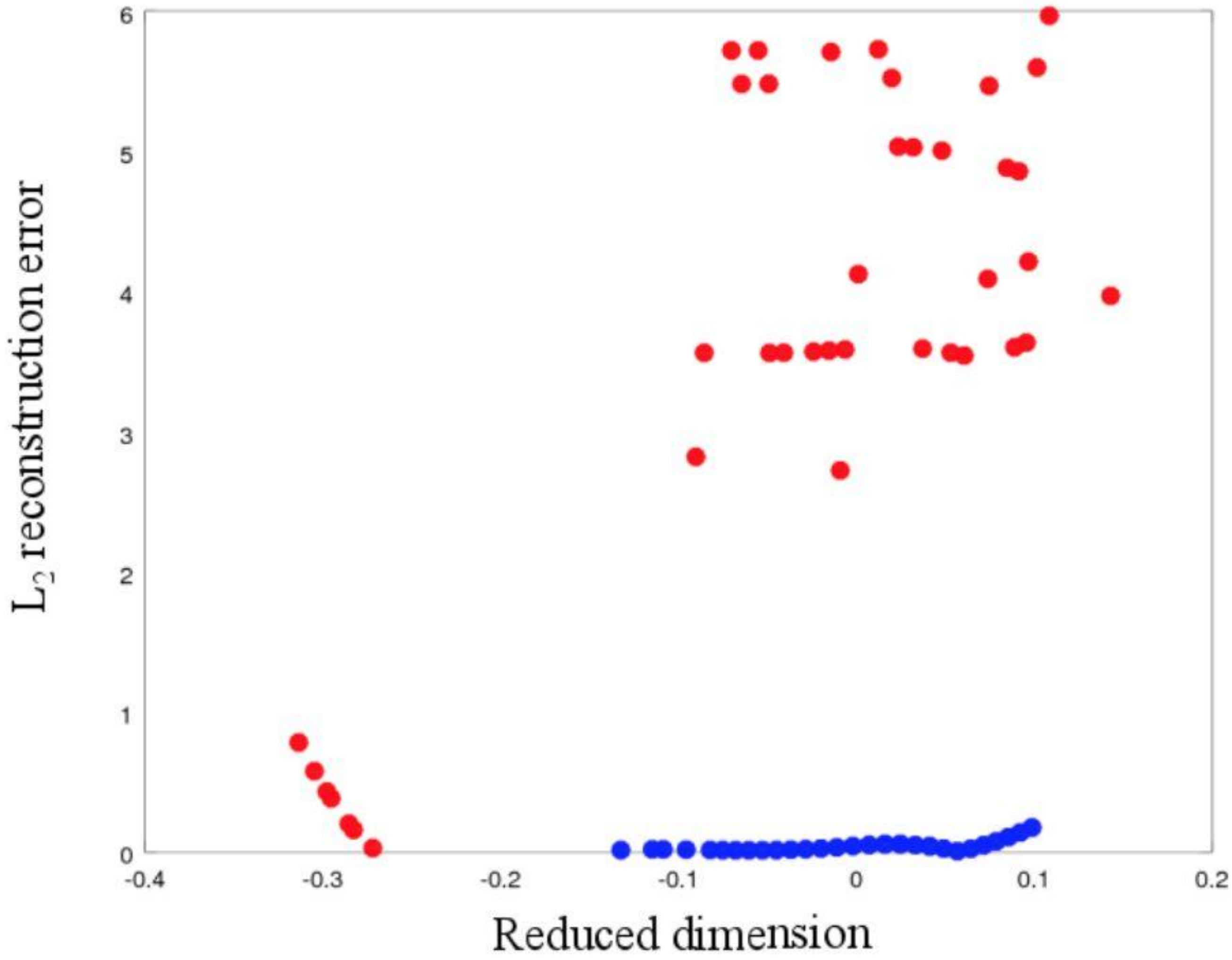


# Deep Autoencoder based Anomaly Detection

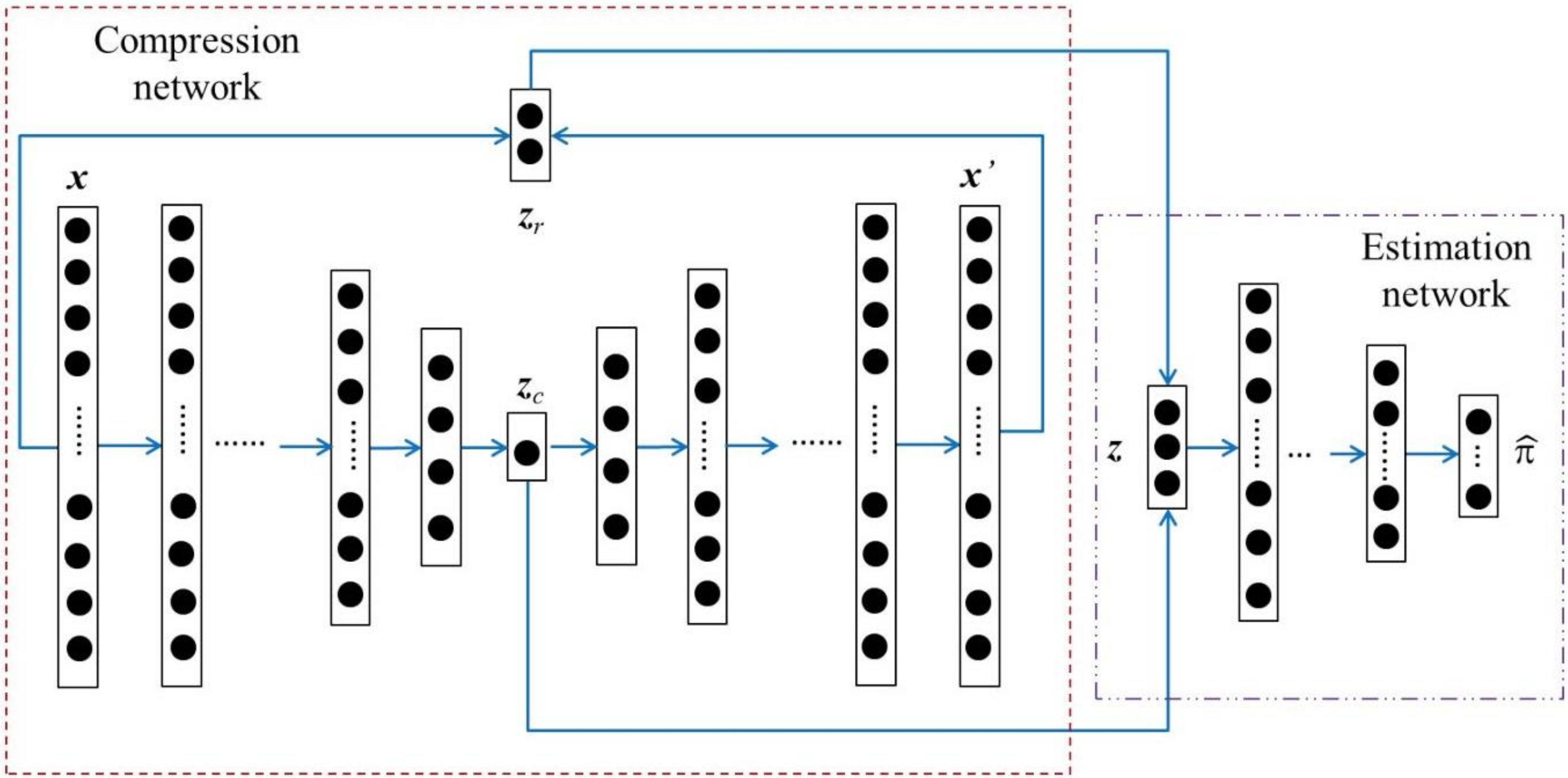
- 1. 将原始数据映射到低维特征空间，在其中评估每一个点跟其他数据点的偏离程度

- 2. 将原始数据映射到低维特征空间，然后由低维特征空间重新映射回原空间，尝试用低维特征重构原始数据，看重构误差的大小

# Deep Autoencoder based Anomaly Detection



# Deep Autoencoder based Anomaly Detection



# Density-based Approaches

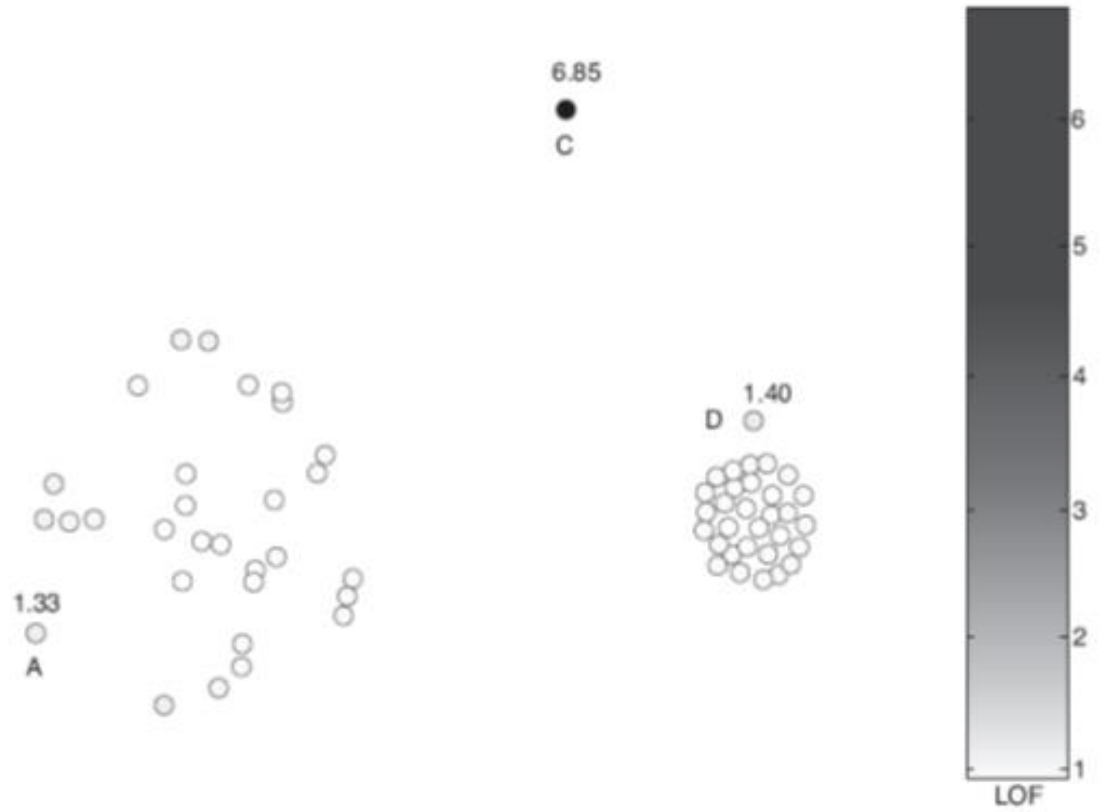
- 1. 一个对象的离群点得分是该对象周围密度的逆

$$\text{density}(\mathrm{x},\mathrm{k}) = \left(\frac{\sum_{y\in N(x,k)} distance(x,y)}{|N(x,k)|}\right)^{-1}$$

- 2. 一个对象周围的密度等于该对象指定距离*d*内对象的个数
- 3. 相对密度

$$\text{average relative density}(\mathrm{x},\mathrm{k}) = \frac{density(x,k)}{\sum_{y\in N(x,k)} density(y,k)/|N(x,k)|}$$

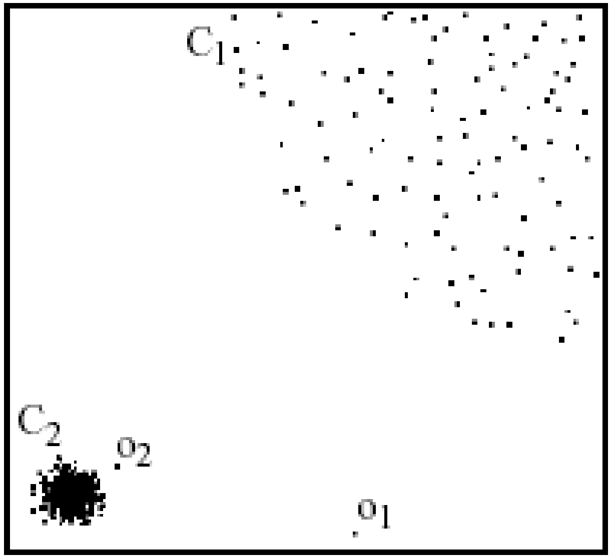
# Density-based Approaches




**LOF离群点得分**

# Density-based: LOF approach

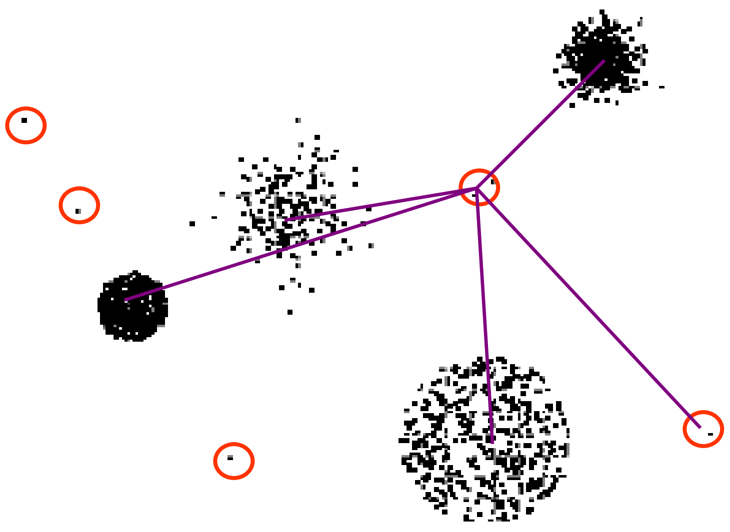
- **For each point, compute the density of its local neighborhood**
- **Compute local outlier factor (LOF) of a sample p as the average of the ratios of the density of sample p and the density of its nearest neighbors**
- **Outliers are points with largest LOF value**



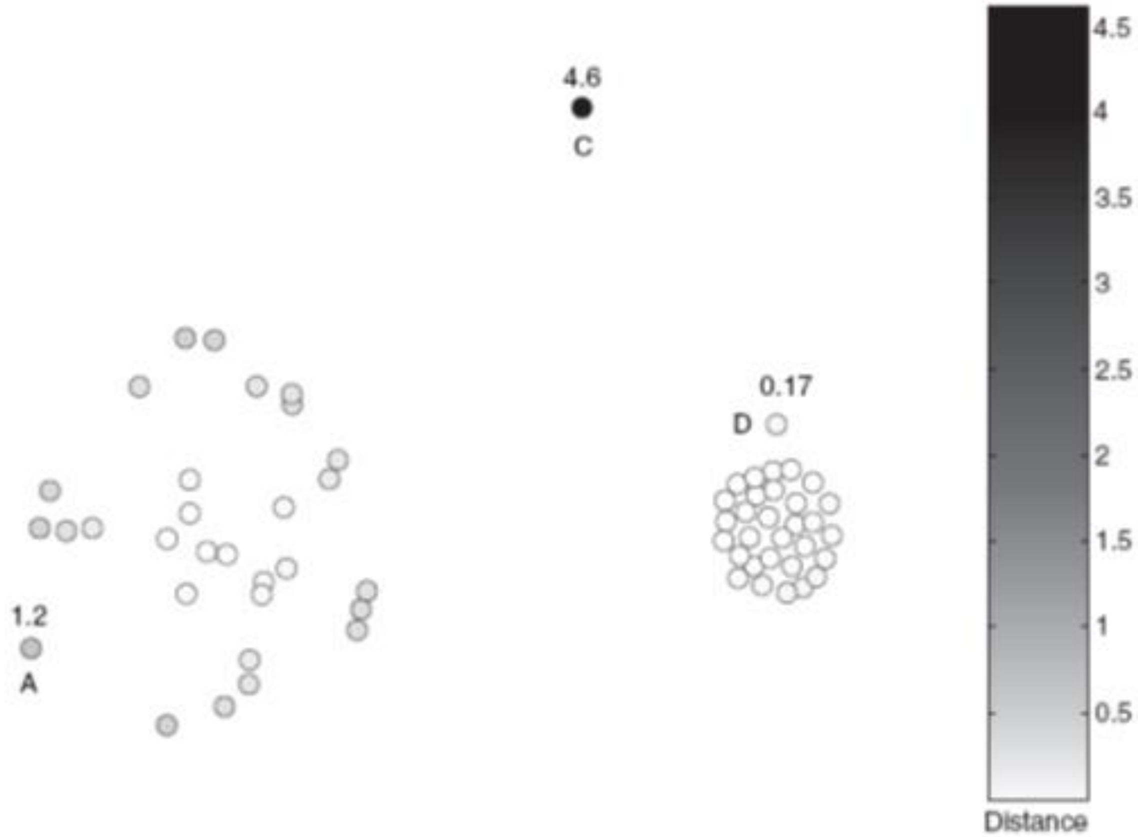

In the NN approach, $o_2$ is not considered as outlier, while LOF approach find both $o_1$ and $o_2$ as outliers

# Clustering-Based

- **Basic idea:**
  - **Cluster the data into groups of different density**
  - **Choose points in small cluster as candidate outliers**
  - **Compute the distance between candidate points and non-candidate clusters**
    - **If candidate points are far from all other non-candidate points, they are outliers**



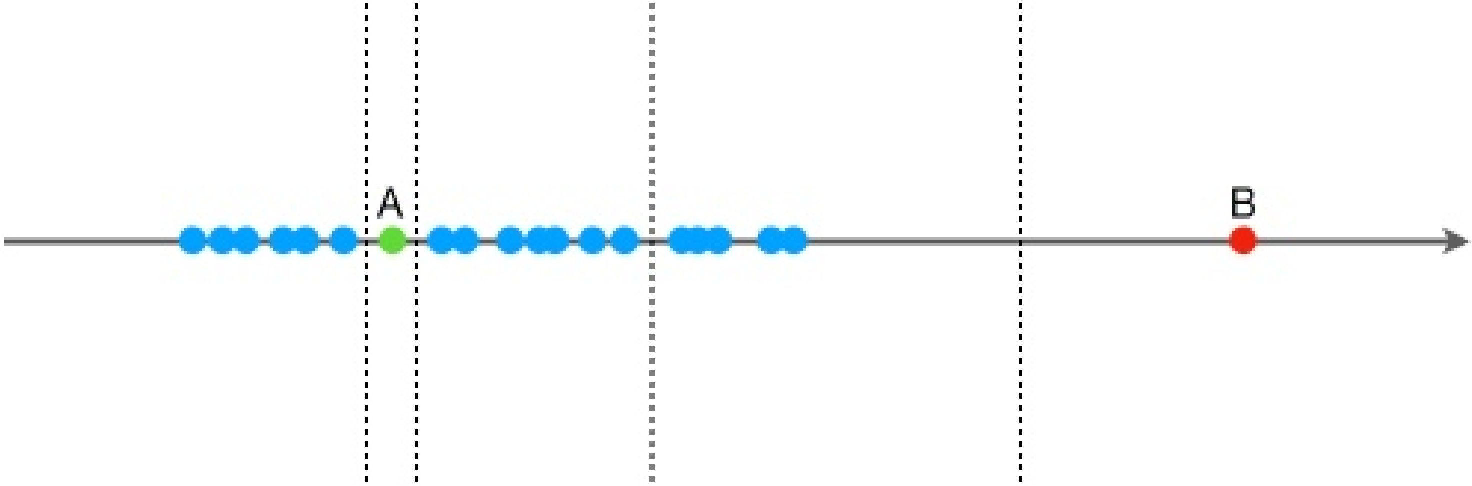
# Clustering-Based




点到最近质心的距离

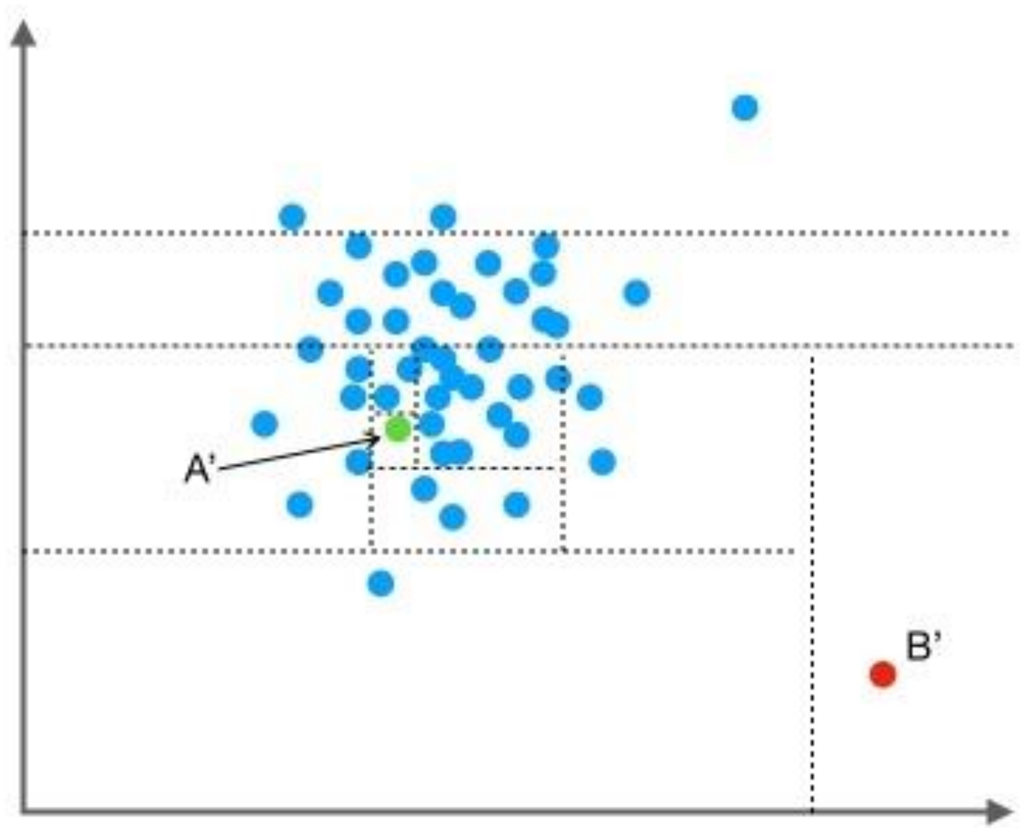
# Isolation Forest

- 如何从下面的数据中，将A和B单独分离出来？



# Isolation Forest

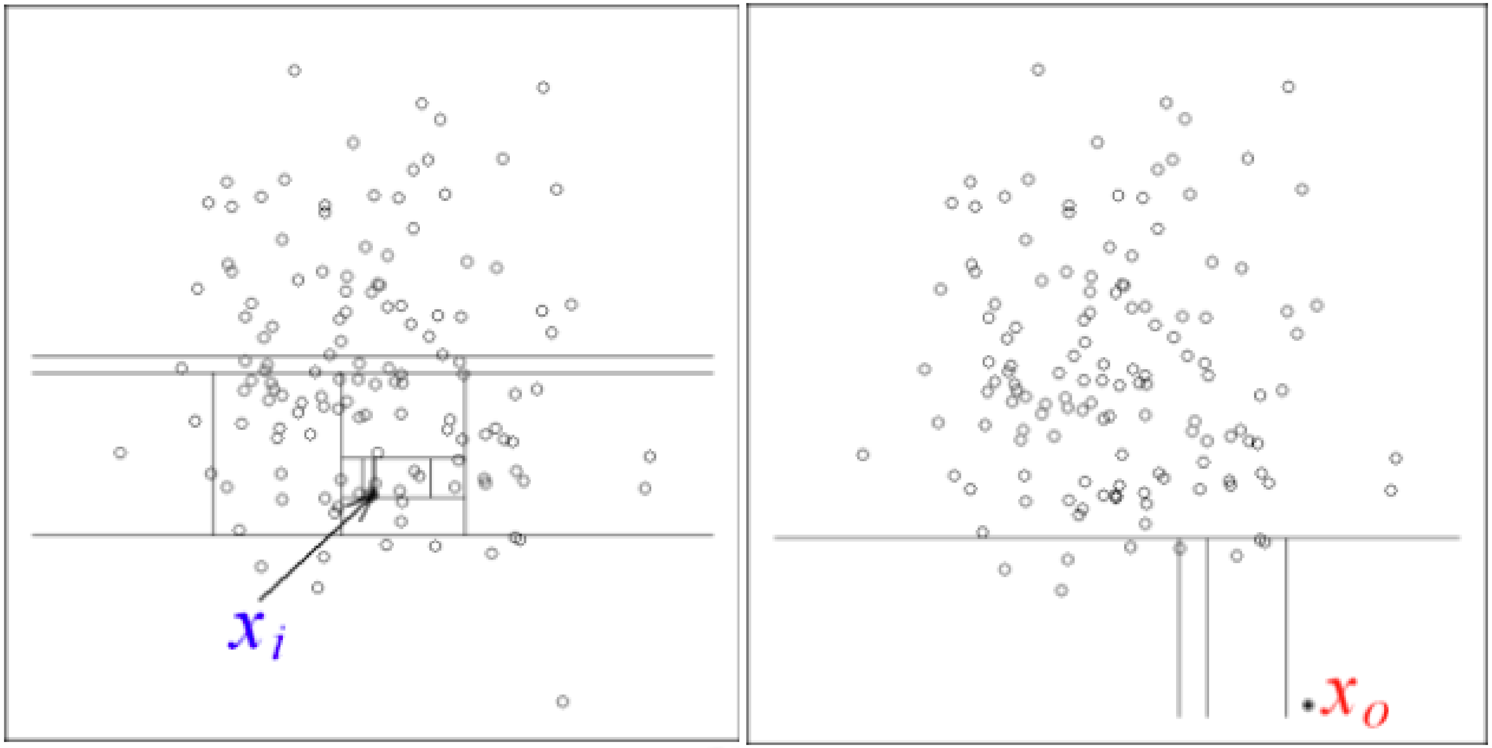
- 将数据拓展到二维



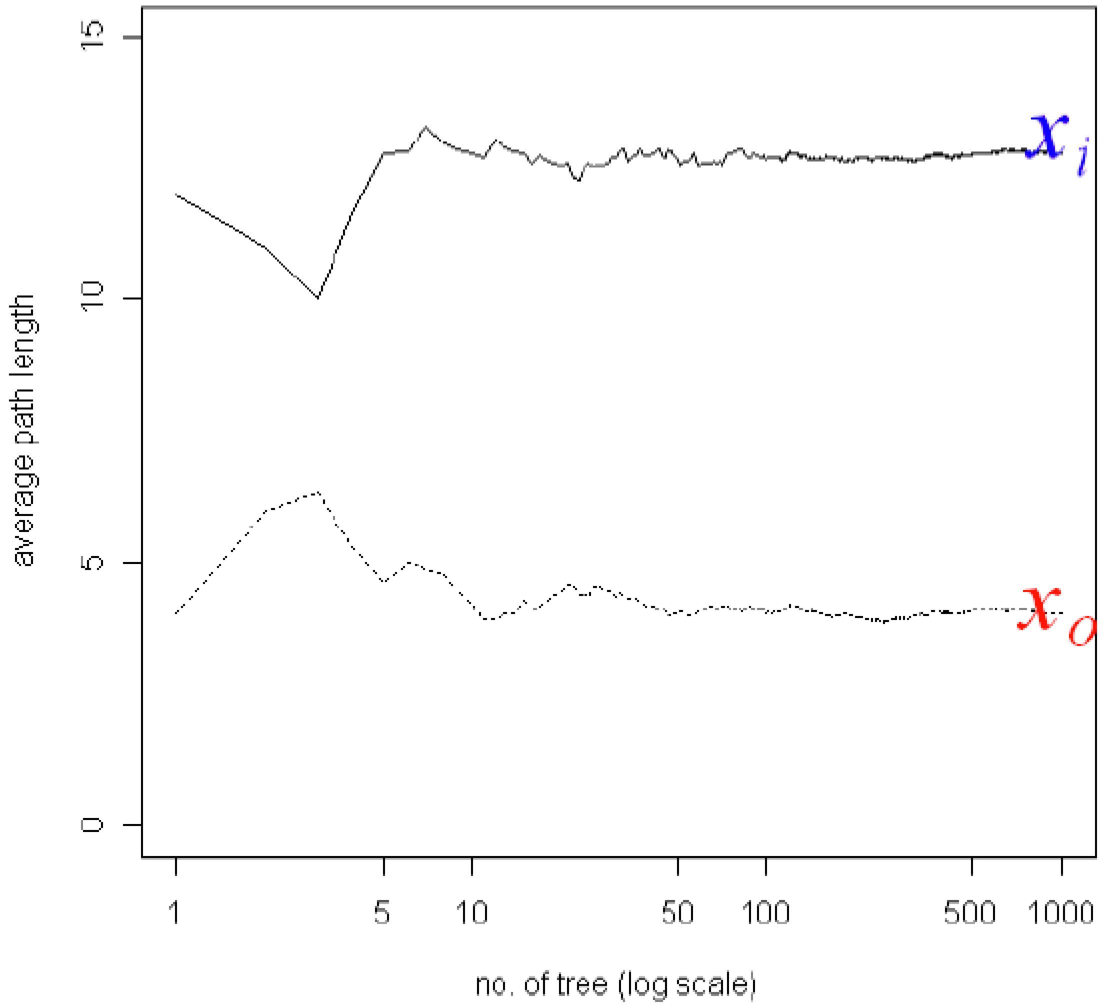
# Isolation Forest

- **Basic Idea**
- 异常数据由于与其他数据点较疏离，可能需要较少几次切分就可以将它们单独划分出来，而正常数据恰恰相反
- 采用二叉树对数据进行切分，数据点在二叉树中所处的深度反应了数据的“疏离”程度

# Isolation Forest



# Isolation Forest



# Isolation Forest

- 算法步骤：
- 1.训练：抽取多个样本，构建多棵二叉树（Isolation Tree，即 iTree）
- 2. 预测：综合多棵二叉树的结果，计算每个数据点的异常分值

# 训练过程

- （1）从全量数据中抽取一批样本
- （2）随机选择一个特征作为起始节点
- （3）在该特征值的范围内随机选择一个值，将样本中小于该取值的数据划到左分支，大于等于该取值的划到右分支
- （4）重复第（3）步，直到满足停止条件
  - 数据不可再分
  - 二叉树达到限定的最大深度

# 预测过程

- （1）估算数据x在每棵树 iTree 中的路径长度（深度）

$$h(x) = e + C(T.size)$$

- （2）综合多棵树的结果，得到数据x的异常分数

$$Score(x) = 2^{-\frac{E(h(x))}{C(\psi)}}$$

- 得分越接近1，表示越异常；越接近0，表示越正常

# Isolation Forest

- 总结
- 两个主要参数
  - 二叉树个数
  - 训练单棵二叉树时抽取的样本数目
- 是一种无监督的的异常检测算法
- 具有线性时间复杂度
- 不能违背异常数据比例低的基本假设

# Performance Comparison

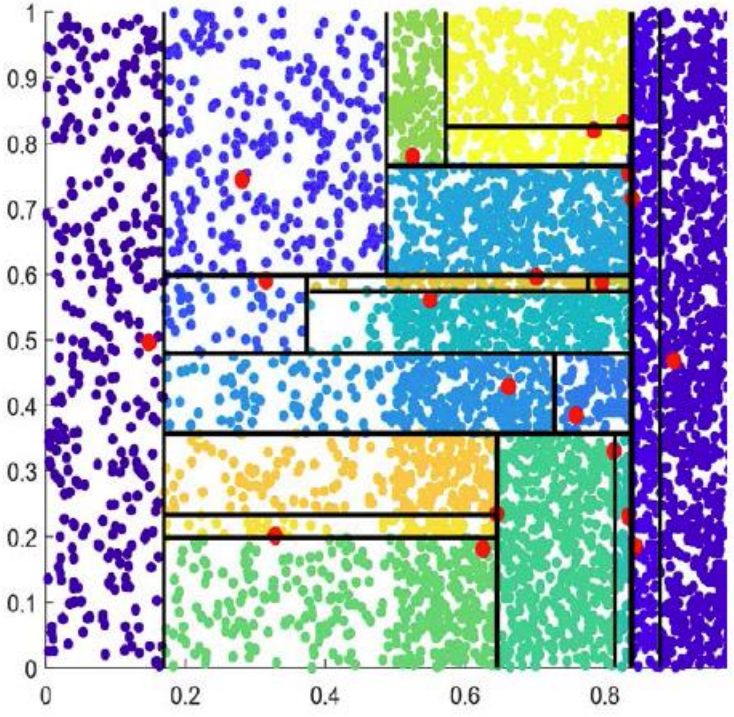
*conditional formatting is a row-wise comparison

ROC Performances (average of 10 independent trials)

| Data | #Samples | # Dimensions | Outlier Perc | ABOD | CBLOF | FB | HBOS | IForest | KNN | LOF | MCD | OCSVM | PCA |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| arrhythmia | 452 | 274 | 14.60 | 0.7688 | 0.7835 | 0.7781 | 0.8219 | 0.8005 | 0.7861 | 0.7787 | 0.779 | 0.7812 | 0.7815 |
| cardio | 1,831 | 21 | 9.61 | 0.5692 | 0.9276 | 0.5867 | 0.8351 | 0.9213 | 0.7236 | 0.5736 | 0.8135 | 0.9348 | 0.9504 |
| glass | 214 | 9 | 4.21 | 0.7951 | 0.8504 | 0.8726 | 0.7389 | 0.7569 | 0.8508 | 0.8644 | 0.7901 | 0.6324 | 0.6747 |
| ionosphere | 351 | 33 | 35.90 | 0.9248 | 0.8134 | 0.873 | 0.5614 | 0.8499 | 0.9267 | 0.8753 | 0.9557 | 0.8419 | 0.7962 |
| letter | 1,600 | 32 | 6.25 | 0.8783 | 0.507 | 0.866 | 0.5927 | 0.642 | 0.8766 | 0.8594 | 0.8074 | 0.6118 | 0.5283 |
| lympho | 148 | 18 | 4.05 | 0.911 | 0.9728 | 0.9753 | 0.9957 | 0.9941 | 0.9745 | 0.9771 | 0.9 | 0.9759 | 0.9847 |
| mnist | 7,603 | 100 | 9.21 | 0.7815 | 0.8009 | 0.7205 | 0.5742 | 0.8159 | 0.8481 | 0.7161 | 0.8666 | 0.8529 | 0.8527 |
| musk | 3,062 | 166 | 3.17 | 0.1844 | 0.9879 | 0.5263 | 1 | 0.9999 | 0.7986 | 0.5287 | 0.9998 | 1 | 1 |
| optdigits | 5,216 | 64 | 2.88 | 0.4667 | 0.5089 | 0.4434 | 0.8732 | 0.7253 | 0.3708 | 0.45 | 0.3979 | 0.4997 | 0.5086 |
| pendigits | 6,870 | 16 | 2.27 | 0.6878 | 0.9486 | 0.4595 | 0.9238 | 0.9435 | 0.7486 | 0.4698 | 0.8344 | 0.9303 | 0.9352 |
| pima | 768 | 8 | 34.90 | 0.6794 | 0.7348 | 0.6235 | 0.7 | 0.6806 | 0.7078 | 0.6271 | 0.6753 | 0.6215 | 0.6481 |
| satellite | 6,435 | 36 | 31.64 | 0.5714 | 0.6693 | 0.5572 | 0.7581 | 0.7022 | 0.6836 | 0.5573 | 0.803 | 0.6622 | 0.5988 |
| satimage-2 | 5,803 | 36 | 1.22 | 0.819 | 0.9917 | 0.457 | 0.9804 | 0.9947 | 0.9536 | 0.4577 | 0.9959 | 0.9978 | 0.9822 |
| shuttle | 49,097 | 9 | 7.15 | 0.6234 | 0.6272 | 0.4724 | 0.9855 | 0.9971 | 0.6537 | 0.5264 | 0.9903 | 0.9917 | 0.9898 |
| vertebral | 240 | 6 | 12.50 | 0.4262 | 0.3486 | 0.4166 | 0.3263 | 0.3905 | 0.3817 | 0.4081 | 0.3906 | 0.4431 | 0.4027 |
| vowels | 1,456 | 12 | 3.43 | 0.9606 | 0.5856 | 0.9425 | 0.6727 | 0.7585 | 0.968 | 0.941 | 0.8076 | 0.7802 | 0.6027 |
| wbc | 378 | 30 | 5.56 | 0.9047 | 0.9227 | 0.9325 | 0.9516 | 0.931 | 0.9366 | 0.9349 | 0.921 | 0.9319 | 0.9159 |
| | | | mean | 0.7031 | 0.7636 | 0.6767 | 0.7819 | 0.8179 | 0.7758 | 0.6792 | 0.8075 | 0.7935 | 0.7737 |
| | | | median | 0.7688 | 0.8009 | 0.6235 | 0.8219 | 0.8159 | 0.7986 | 0.6271 | 0.8135 | 0.8419 | 0.7962 |
| | | | sd | 0.2038 | 0.1890 | 0.1966 | 0.1866 | 0.1590 | 0.1764 | 0.1929 | 0.1738 | 0.1775 | 0.1916 |

# Isolation Forest

- 缺点
- 1. 不适用于特别高维数据
- 2. 仅对全局异常点敏感，不擅长处理局部的相对异常点
- 3. 划分边界平行于坐标轴



# iNNE

- iForest的改进算法
  - Isolation-based anomaly detection using nearest-neighbor ensembles
- 借鉴数据孤立机制，并结合最近邻距离计算
  - 采用多维超球体切割数据空间
  - 考虑了数据局部分布特性

# 训练阶段

- 1. 从训练数据中随机选择Ψ个样本点构成子空间，对于每个样本点都找到其在另外（Ψ-1）个点中距离最近的点（最近邻），以到该最近邻的距离为半径，自己为圆心画出Ψ个超球

- 2. 重复上一步t次，得到 t 组超球，每次样本点都是独立从所有原始数据中随机抽样产生

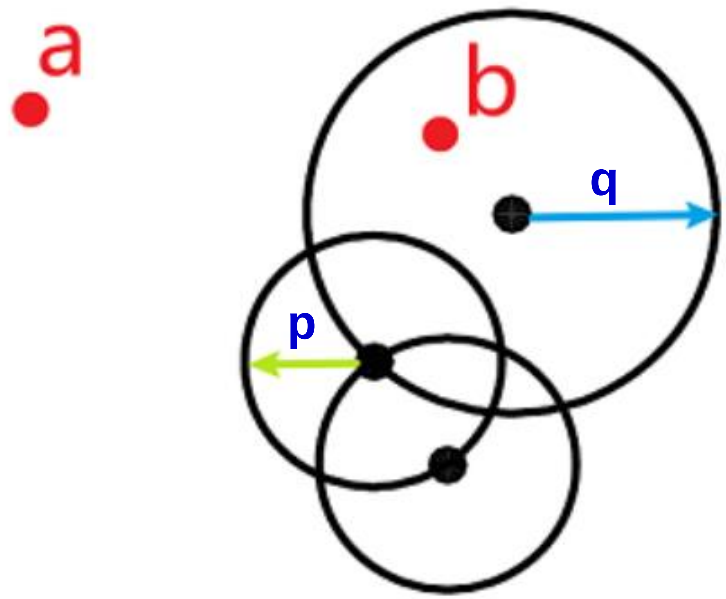
# 测试阶段

- 1. 如果测试数据不在任何一个超球内，则其异常值为1，判为孤立点

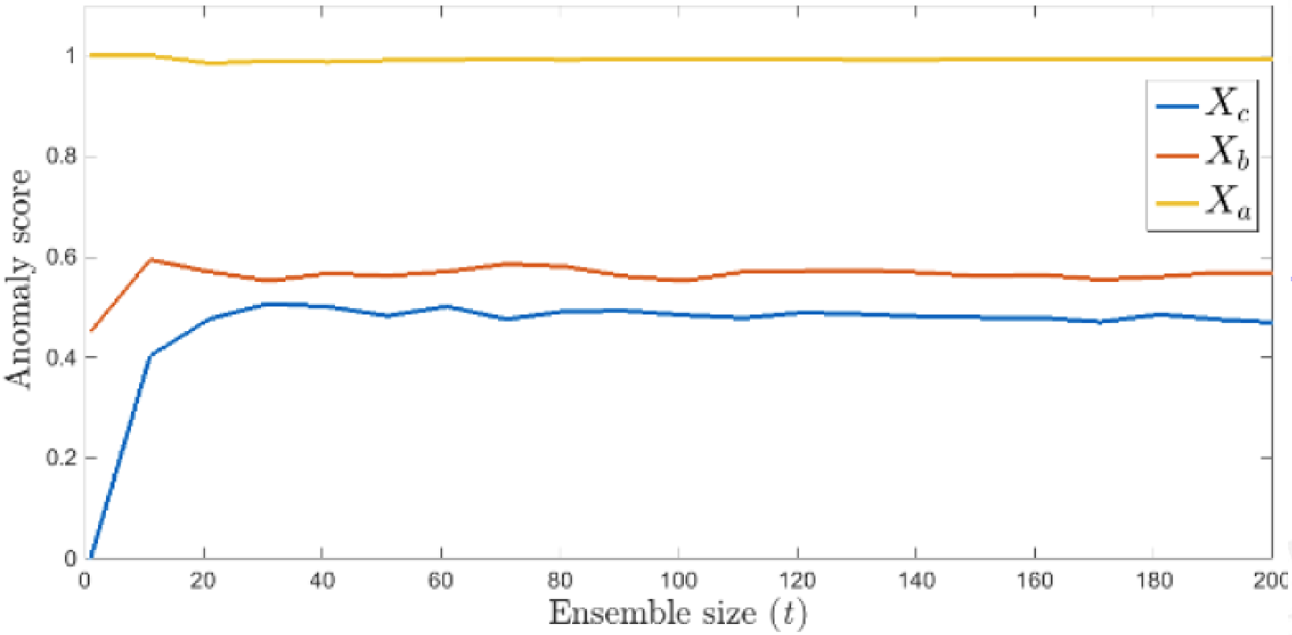
- 2. 如果该测试数据在某个超球Q的范围内，首先记录超球Q的半径为q，之后再找到离超球Q最近的超球P，记录其半径为p，该测试数据的异常值为 1 - p/q

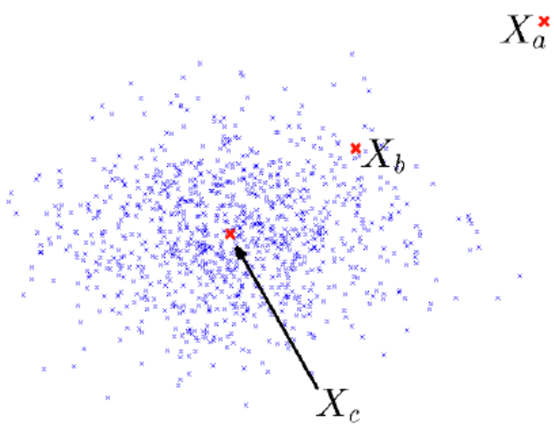
# 测试阶段

➢ **3. 将测试数据分别放进每组超球中进行评估，得出 t 个异常值，然后计算平均值做为测试数据最后的异常指标**



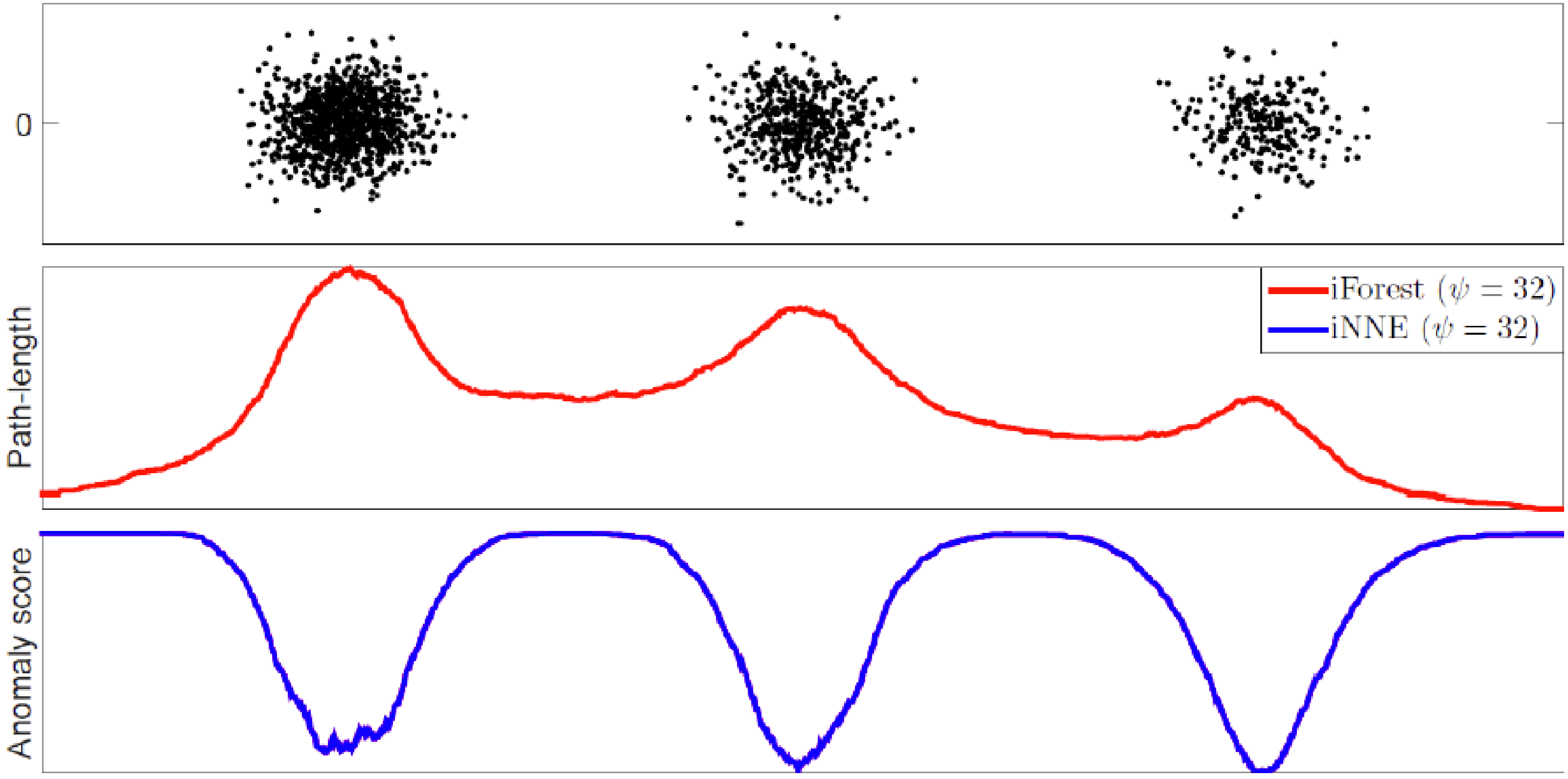
# Example


Anomaly score
Ensemble size ($t$)
0
0.2
0.4
0.6
0.8
1
0
20
40
60
80
100
120
140
160
180
200
$X_c$
$X_b$
$X_a$


$X_a$
$X_b$
$X_c$

# 性能对比

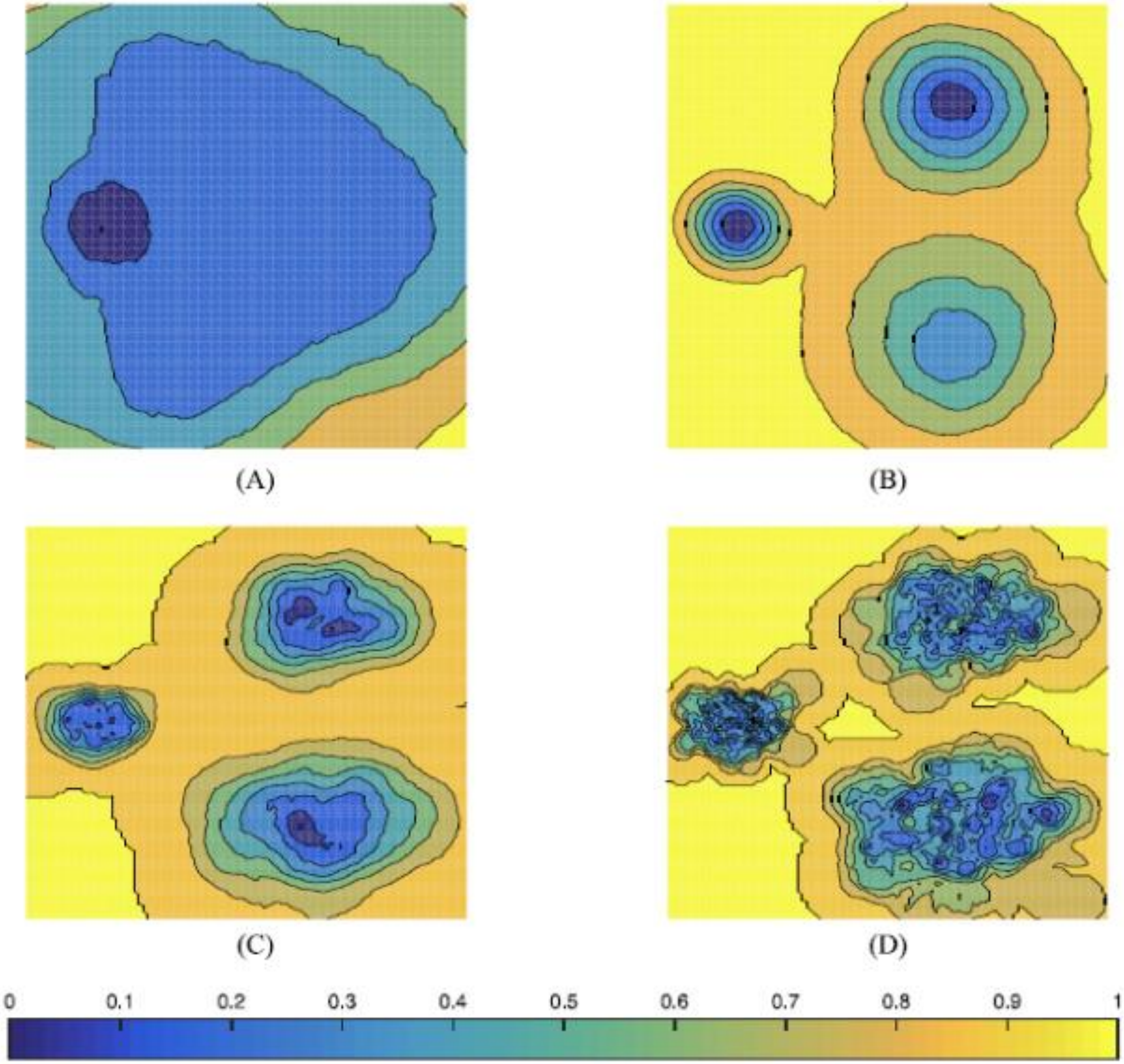
## ➢ iForest vs. iNNE



# 参数影响

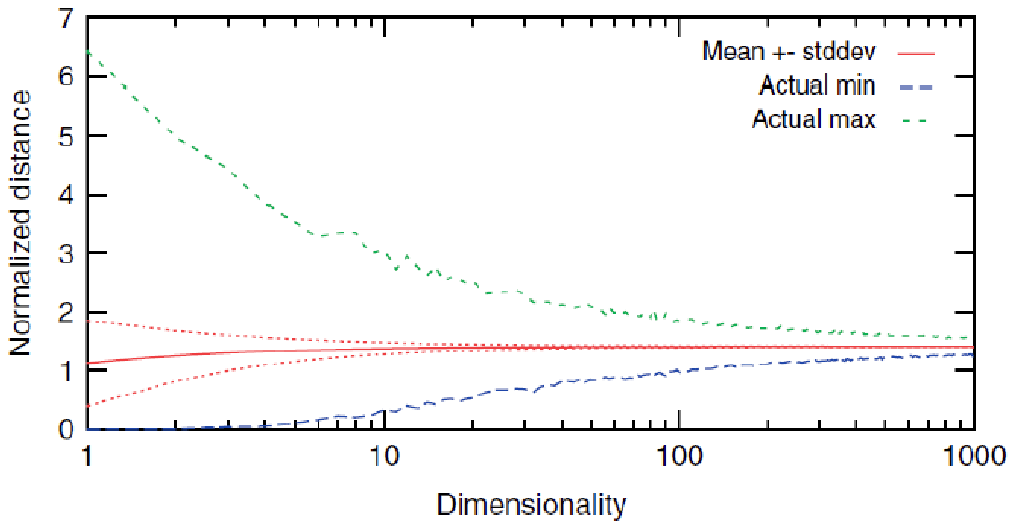
- 1. 重复抽样t次，得到t组超球
  - t 取值越大结果越稳定，运行时间越长
- 2.随机选择Ψ个样本点构成子空间
  - 根据数据中有多少个高密度区域来调整

# 参数影响



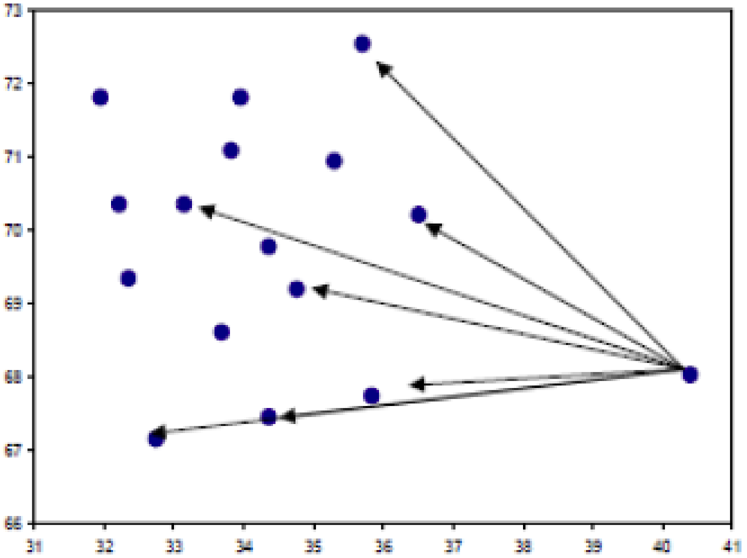
# 高维数据异常检测

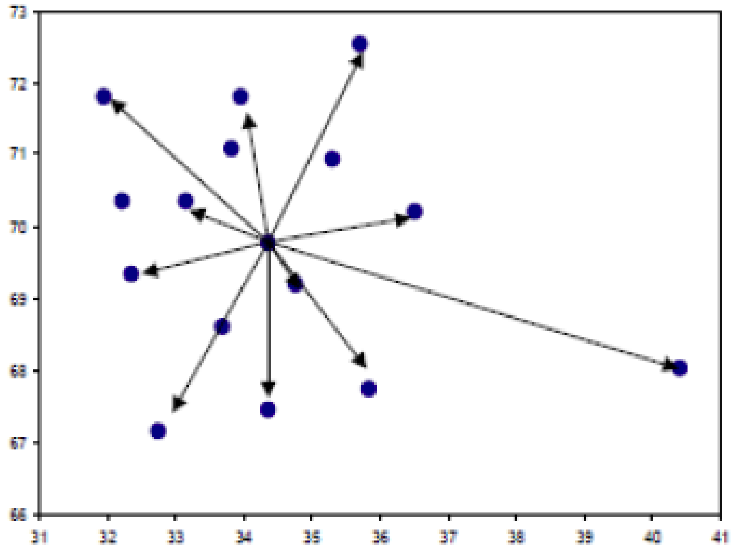
- 针对十万至百万级别维度的数据
- 维度灾难
  - 数据变的稀疏，异常点被遮挡
  - 数据两两之间的距离几乎相等（距离集中效应）
  - 最近邻（NN）概念无意义



# 基于角度的异常检测

- Angle-based Outlier Factor (ABOF)




- 基于角度的度量比距离更稳定
- 基于角度的度量对维灾难更鲁棒

# 时间序列异常检测

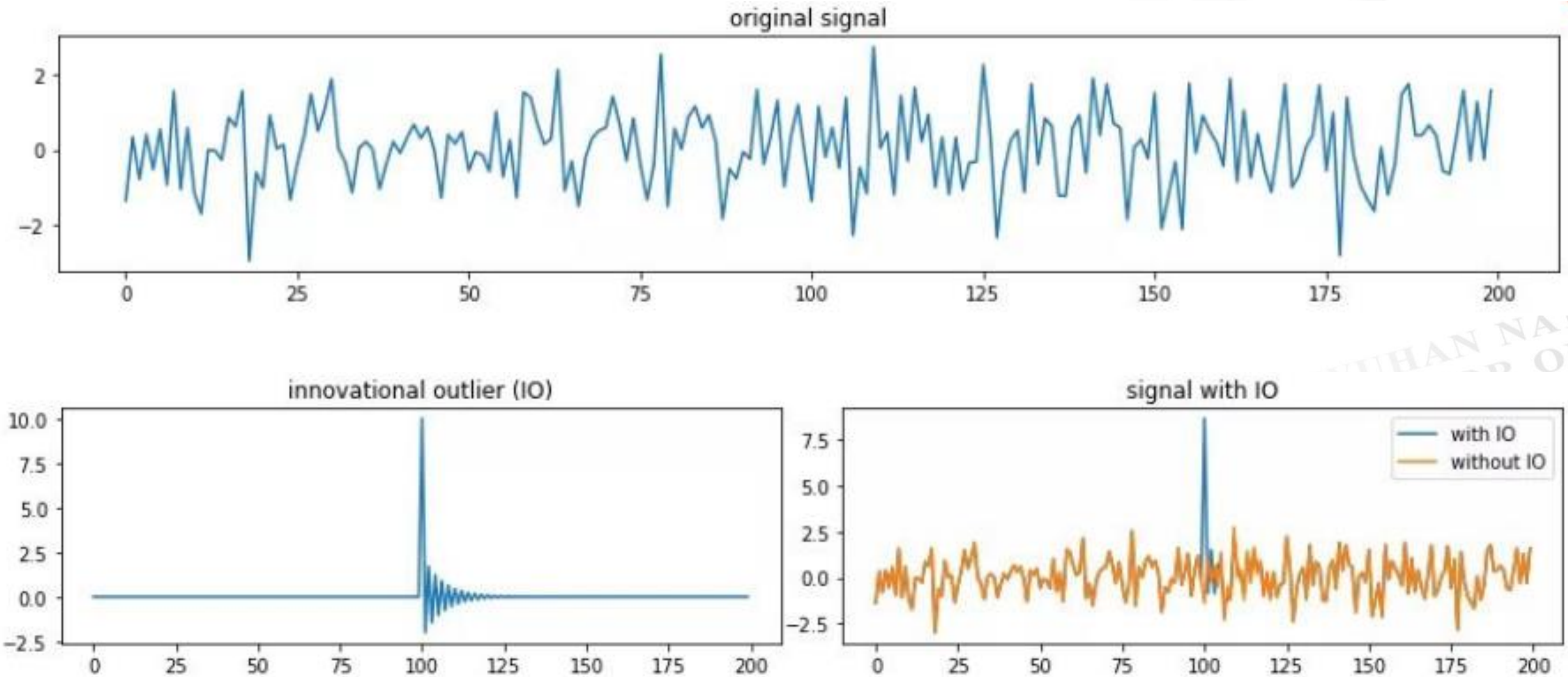
- 关注时间序列（如信号）的异常检测
  - 峰值
  - 趋势变化
  - 等级转换

# 时间序列异常类型

- 革新性异常：**Innovational Outlier (IO)**
  - 干扰不仅作用于X(T)，而且影响T时刻以后的序列
- 附加性异常：**Additive Outlier (AO)**
  - 只影响该干扰发生的时刻T上的序列值，不影响该时刻以后的序列值
- 水平移位异常：**Level Shift (LS)**
  - 持续影响T时刻以后的所有行为，往往表现出T时刻前后的序列均值发生水平位移
- 暂时变更异常：**Temporary Change (TC)**
  - 在T时刻干扰发生时具有一定初始效应，以后随时间呈指数衰减

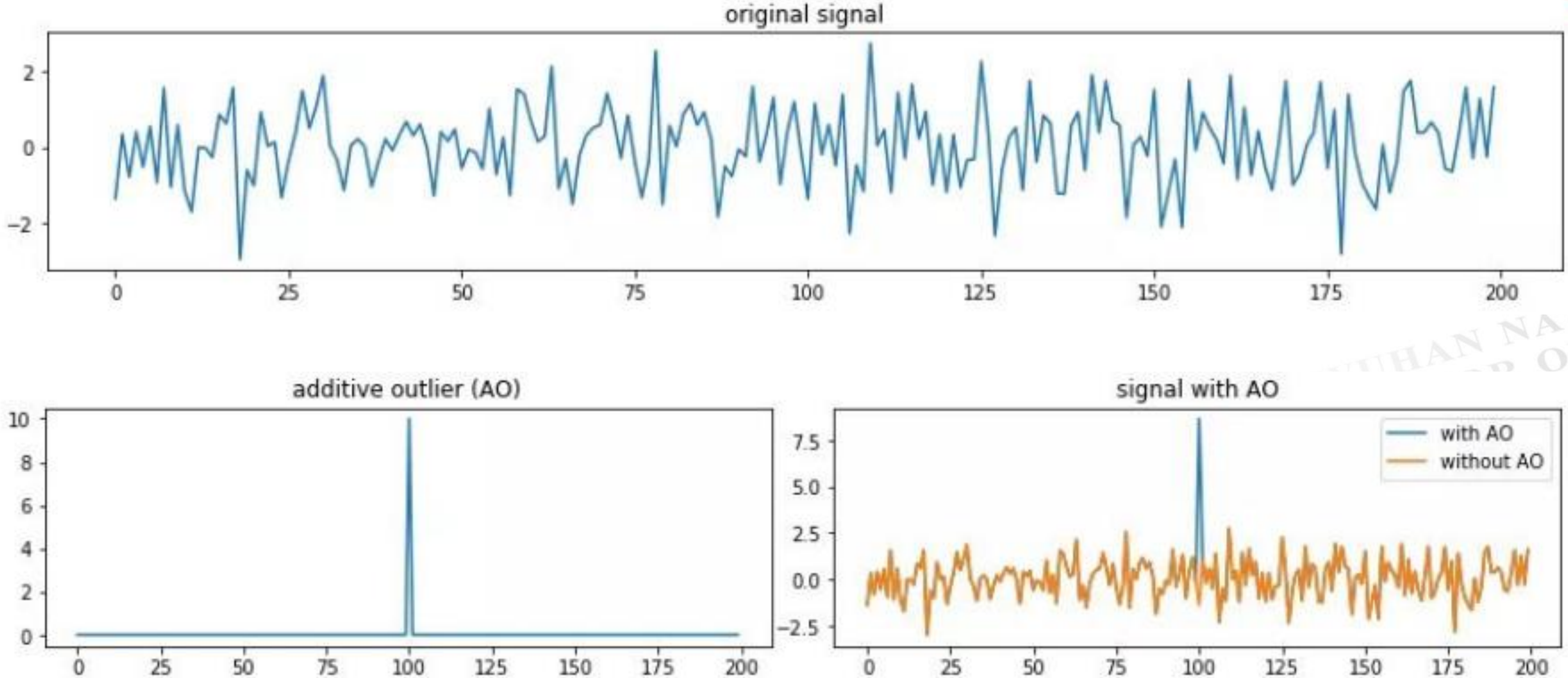
# 时间序列异常类型

## ➢ 革新性异常



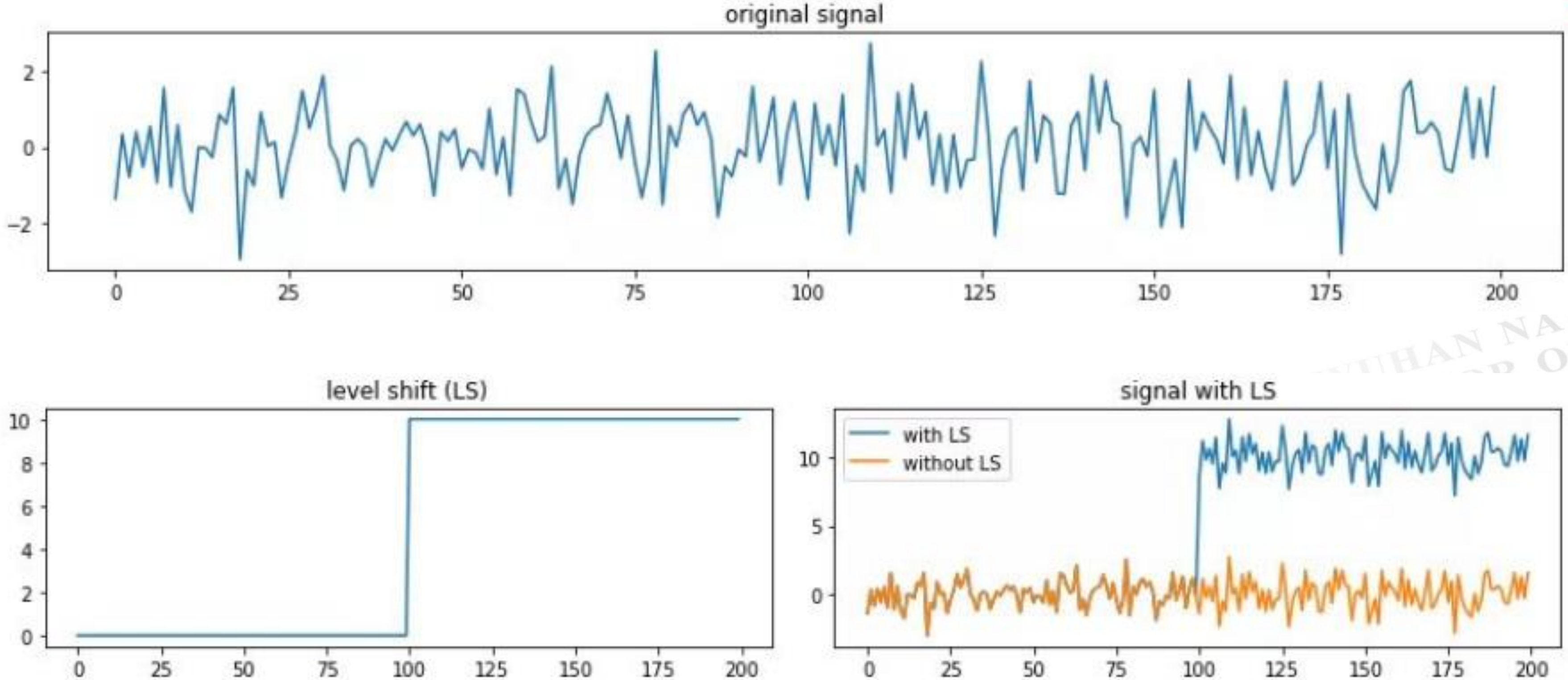
# 时间序列异常类型

## ➢ 附加性异常



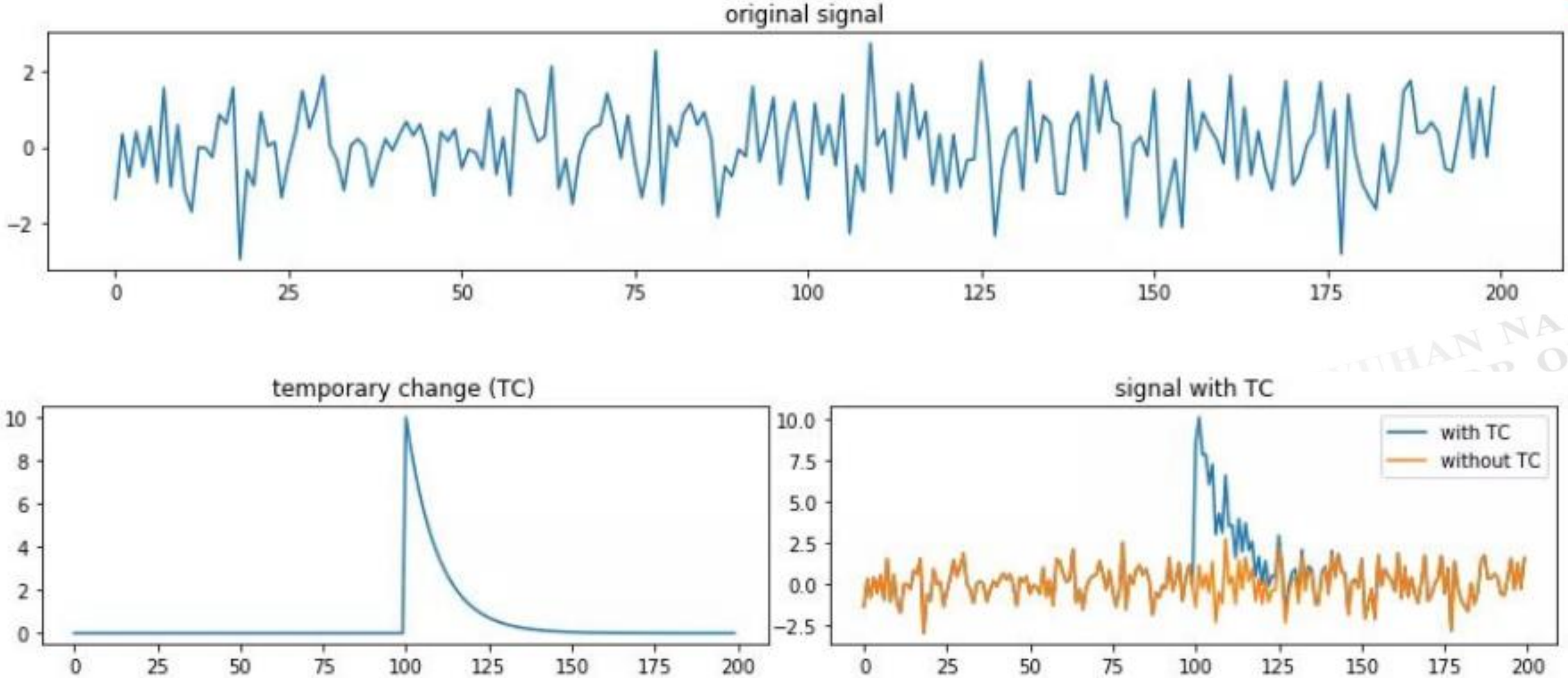
# 时间序列异常类型

## ➢水平移位异常



# 时间序列异常类型

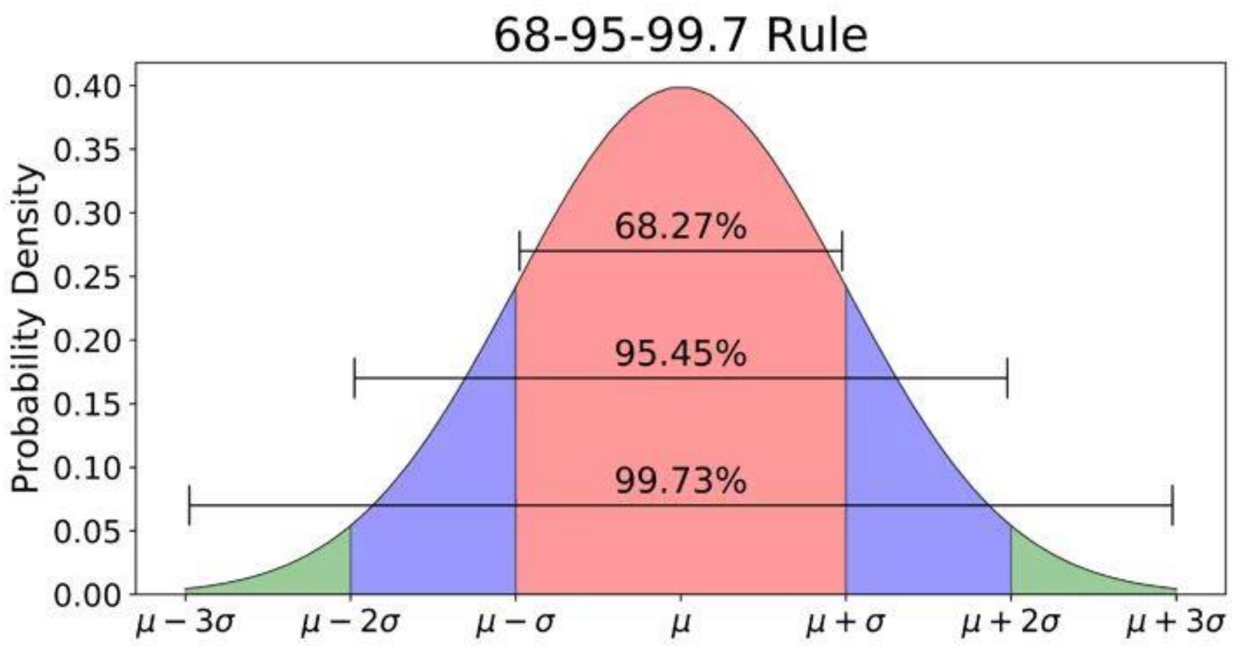
- 暂时变更异常



# 时序预测模型

## ➢3-Sigma

- ➢假设一组检测数据只含有随机误差，对原始数据进行计算处理得到标准差，然后按一定的概率确定一个区间，误差超过这个区间的就属于异常值

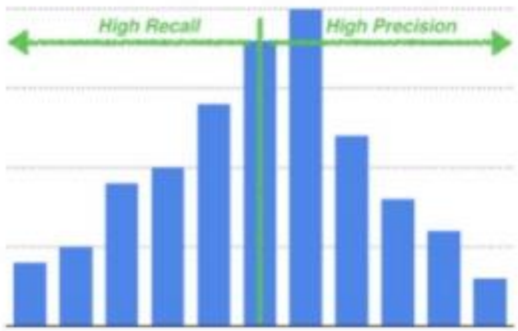


# 美团案例

- 基于形变分析模型的异常检测系统

- 现状：业务持续高速成长，业务迭代快，逻辑复杂，关联服务多

# 业务痛点




告警精确率与召回率难平衡

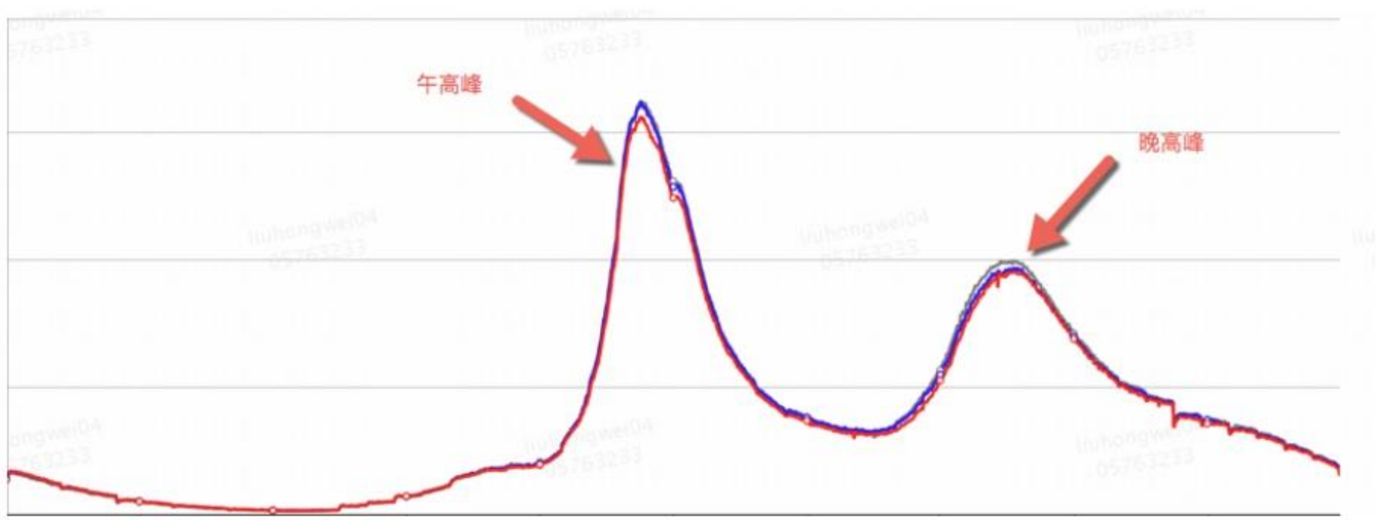
人工配置告警阈值成本高

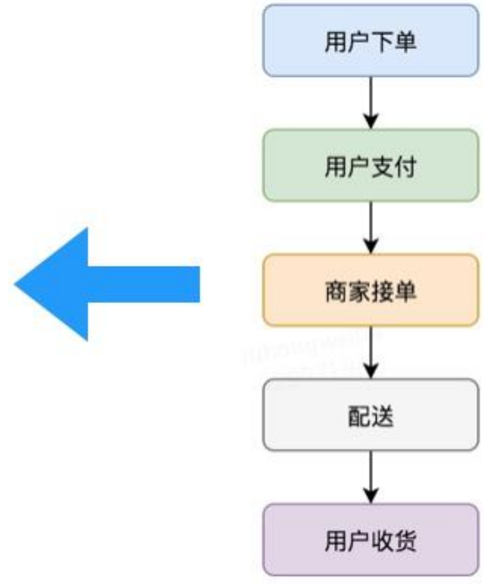

典型故障场景分析需要人工介入


重大事故时如何避免告警洪潮

# 数据特点

## ➤1. 有规律的时间序列



# 数据特点

- **2. 无规律的时间序列**

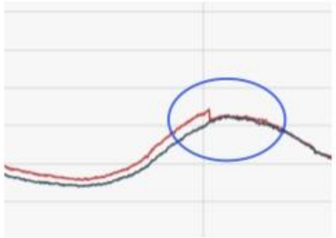

AVG、TP99等性能指标

# 形变分析模型

异常检测 → 形变分析 ＋ 相关性变点检测 ＋ 告警收敛策略

形变分析 → （1 - 余弦相关）x |实时当前值- 基线当前值|

相关性变点检测 →



针对当前值减去基线值趋近于0的超级变点，采用前一分钟数值减去当前值作为补充。

（1 - 余弦相关）x |前一分钟数值- 当前值|

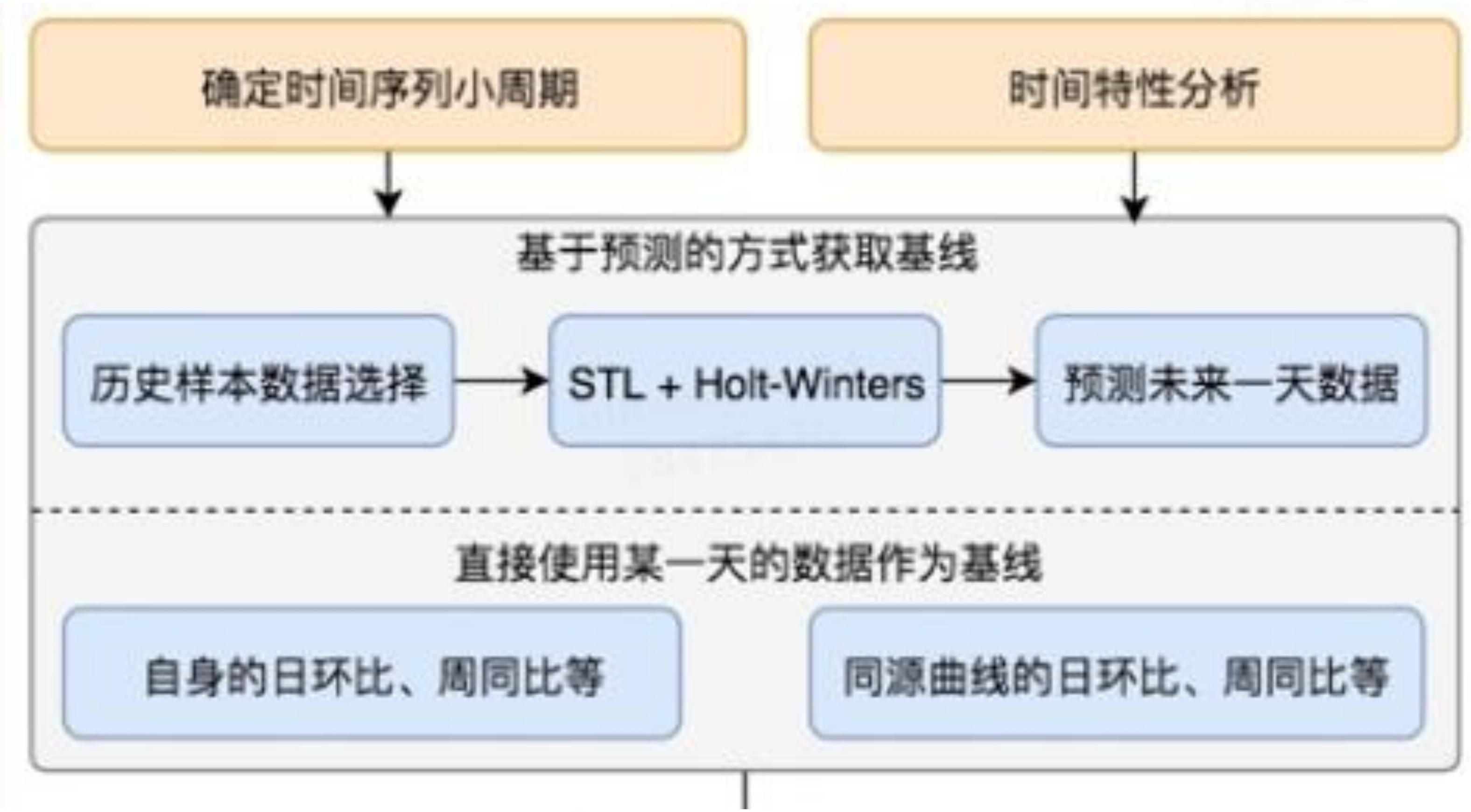
告警收敛策略 →

链路维度收敛：同一刻单链路多条曲线告警收敛为一条。

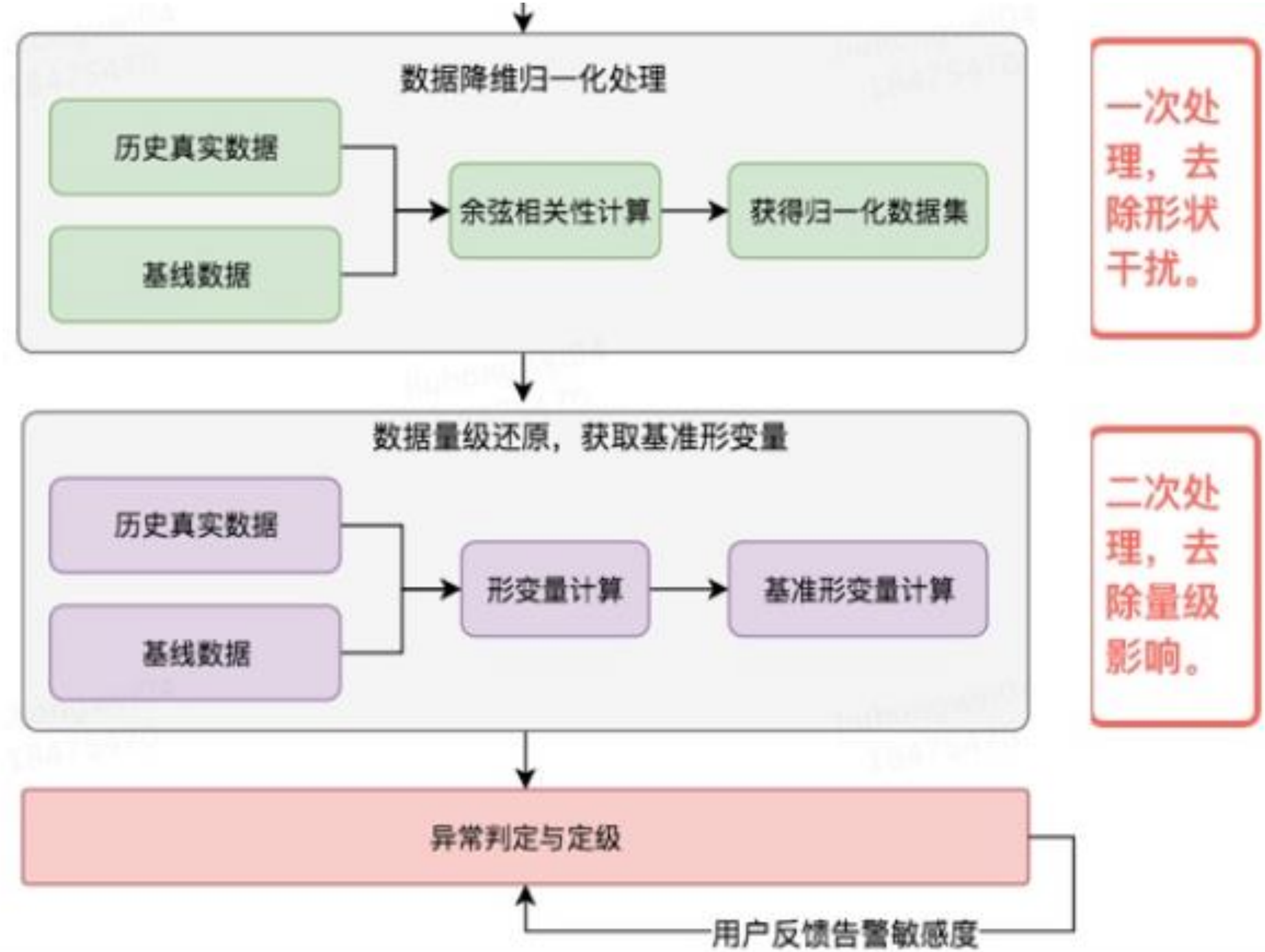
按照次数或时间桶进行收敛。

图形化告警信息，直观体现前后异常趋势。

# 模型分析过程

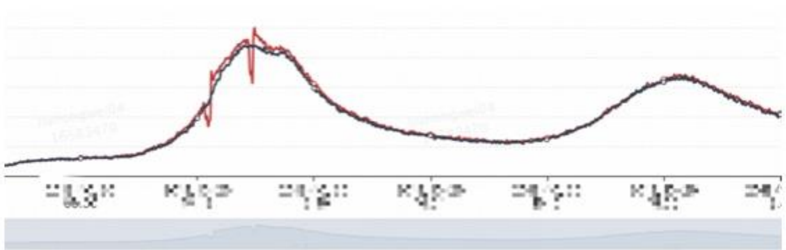

确定时间序列小周期
时间特性分析
基于预测的方式获取基线
历史样本数据选择
STL + Holt-Winters
预测未来一天数据
直接使用某一天的数据作为基线
自身的日环比、周同比等
同源曲线的日环比、周同比等

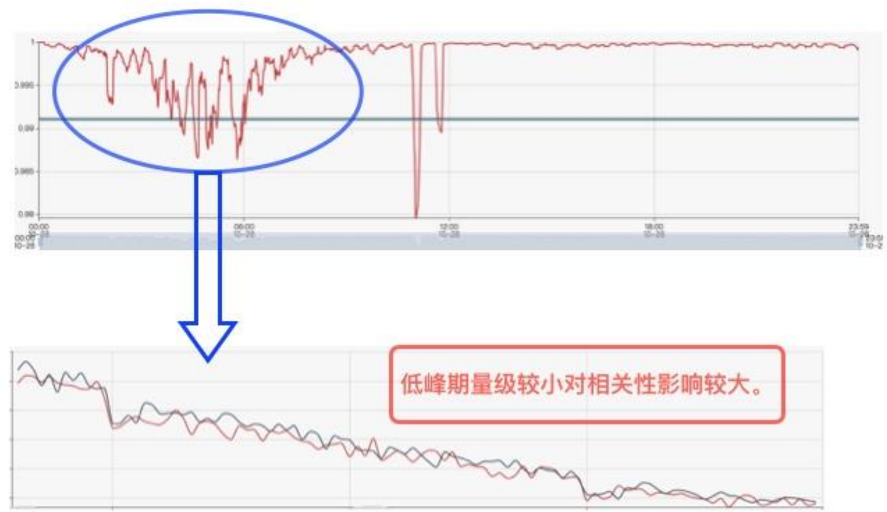
# 模型分析过程


数据降维归一化处理
历史真实数据
基线数据
余弦相关性计算
获得归一化数据集
一次处理，去除形状干扰。
数据量级还原，获取基准形变量
历史真实数据
基线数据
形变量计算
基准形变量计算
二次处理，去除量级影响。
异常判定与定级
用户反馈告警敏感度

# 二次处理

- 真实数据与基线数据进行归一化操作

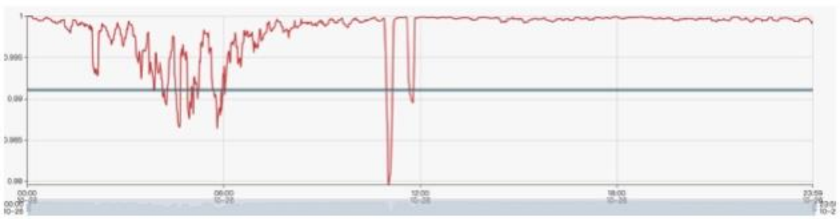





归一化互相关（余弦相关性）：

$$norm\_corr(x,y)=\frac{\sum_{n=0}^{n-1}x[n]*y[n]}{\sqrt{\sum_{n=0}^{n-1}x[n]^2*\sum_{n=0}^{n-1}y[n]^2}}$$

# 二次处理




- 形变量计算：

（1 - 余弦相关性）x |实时当前值- 基线当前值|

# 告警收敛策略

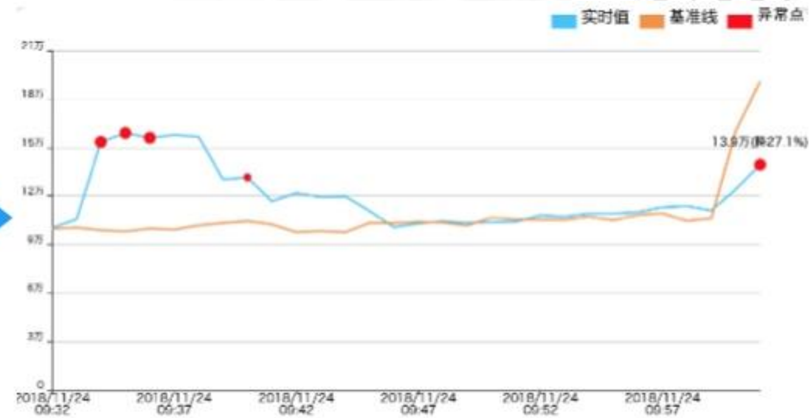
- 目标

- 1. 针对典型场景，快速给出简单直观的建议
- 2. 针对重大事故，避免出现告警洪潮

# 告警收敛策略

## ➢1. 简化告警内容，直观展示异常点与变化趋势

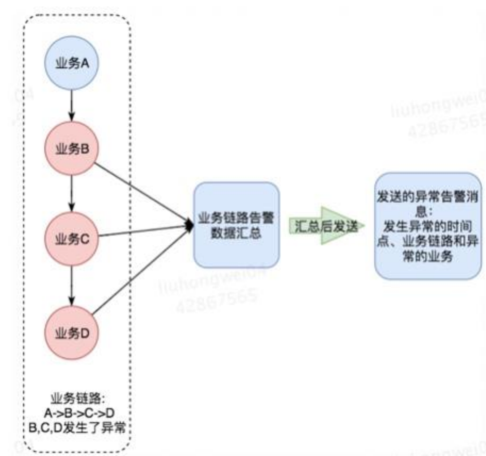
【P1业务告警】
【业务大盘】：
【业务图表】：
【业务指标】
【异常时间】：2018-11-24 12:33:00
【指标详情】：当前值:115436, 预测线值:144665, 下降:29229, 降幅:20.2%



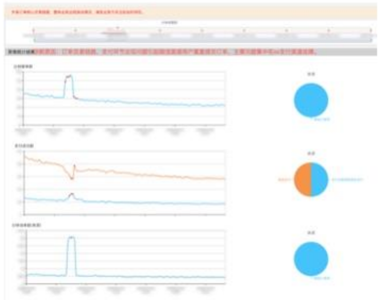
# 模型分析过程

## ➢2. 从时间和业务两个维度上避免告警洪潮

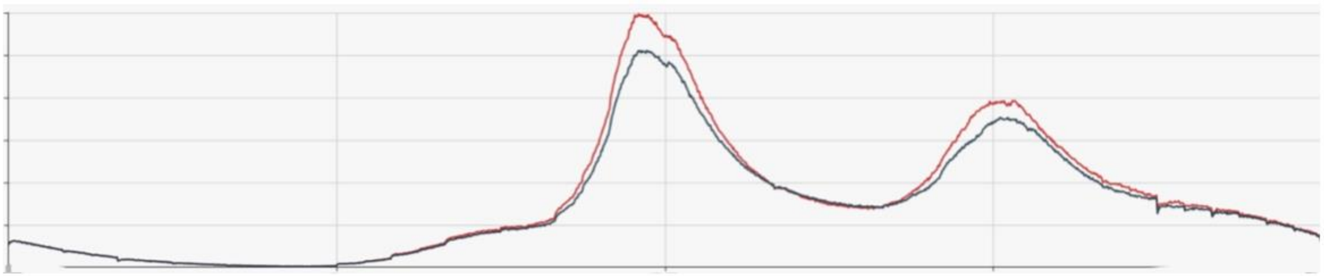

在事故持续时间较长时，每分钟都发送告警会对业务造成干扰，在连续三分钟发送异常告警之后，采用间隔3、5、7、7....... 直到判断异常恢复为止。




根据业务相关性，从强相关的业务链路上收集异常告警事件进行分析，从更高维度给出链路级分析报告。



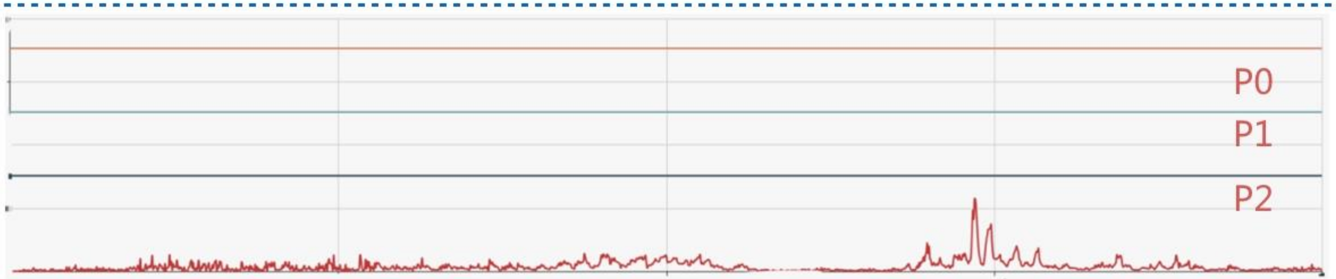
# 案例1



不应该被识别为异常的非事故案例

案例1：因为全国大范围出现恶劣天气，引起了午晚高峰整体抬升，这种情况不希望出现连续告警。

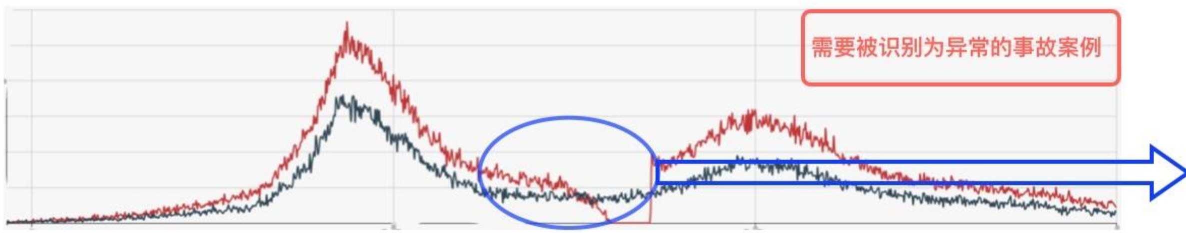
一次处理，将历史样本与基线进行归一化互相关计算，得到数据集看到在业务低峰期时，相关性波动很大，在午晚高峰时相关性较高。

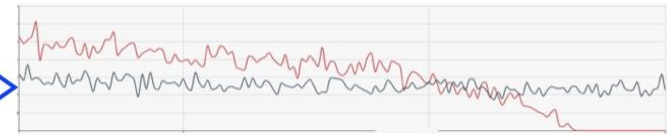



二次处理，还原量级，去除量级维度，并通过基准形变量计算出不同告警等级对应的形变量，我们发现没有任何点需要告警，符合预期。

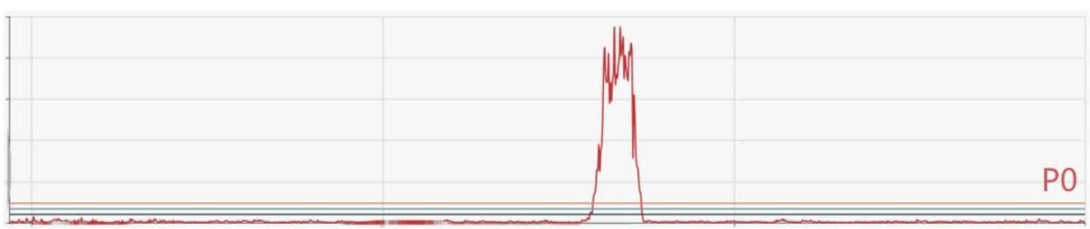
# 案例2

案例2：某一渠道出现问题引起整体流量缓慢下降，需要识别出异常点。



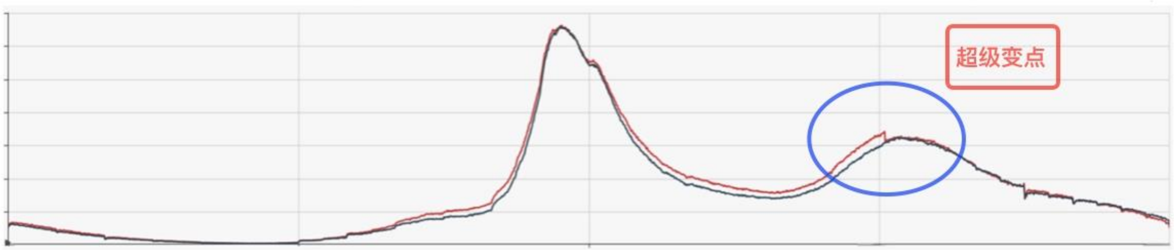



一次处理，将历史样本与基线进行归一化互相关计算，得到数据集看到在渠道异常期间相关性波动较大。




二次处理，还原量级，去除量级维度，并通过基准形变量计算出不同告警等级对应的形变量，在渠道异常期间引起的指标缓慢下降，逐渐超过相应等级的告警阈值，符合预期。

# 案例3




需要被识别为异常的事故案例

案例3：某服务入口流量因为某一渠道突然故障，引起整体流量陡降，之后曲线形状保持不变，陡降异常点需要被识别出。

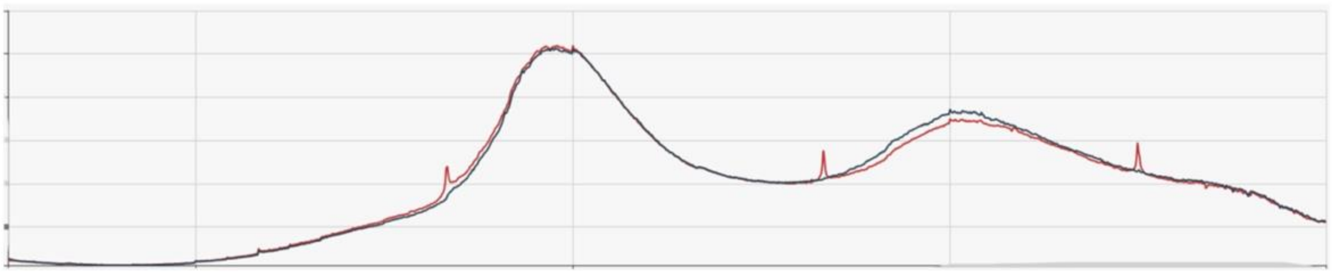
（1 - 余弦相关）x |实时当前值- 基线当前值|

一次处理，将历史样本与基线进行归一化互相关计算，因为故障之后曲线形状迅速恢复，相关性依然很高。


（1 - 余弦相关）x |前一分钟数值- 当前值|


二次处理，还原量级，去除量级维度，并通过基准形变量计算出不同告警等级对应的形变量，因为异常之后真实值与基线基本吻合，形变量计算在这种特例下无法识别，需要同时增加前后一分钟的形变量分析，在两个结果中任何一个超过对应等级的告警阈值则认为是异常点，符合预期。

# 案例4



需要被识别为异常的非事故案例

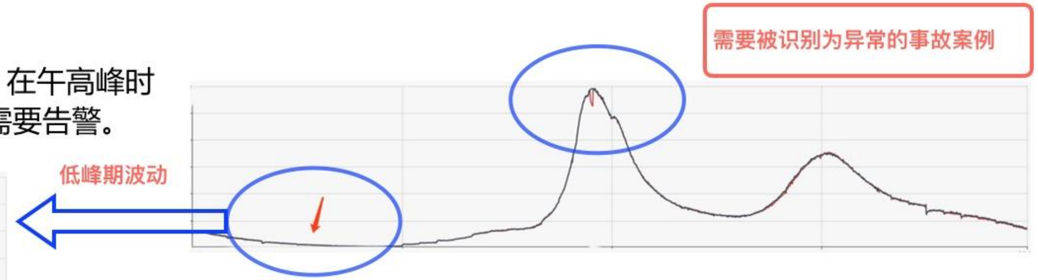
案例4：世界杯期间的营销活动不定时引起指标陡升，正常需要识别出异常点。

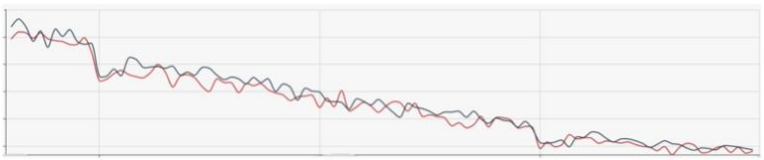
一次处理，将历史样本与基线进行归一化互相关计算，得到数据集看到在活动期间相关性波动较大


二次处理，还原量级，去除量级维度，并通过基准形变量计算出不同告警等级对应的形变量，三处活动期间引起的指标陡升超过告警阈值，符合预期。

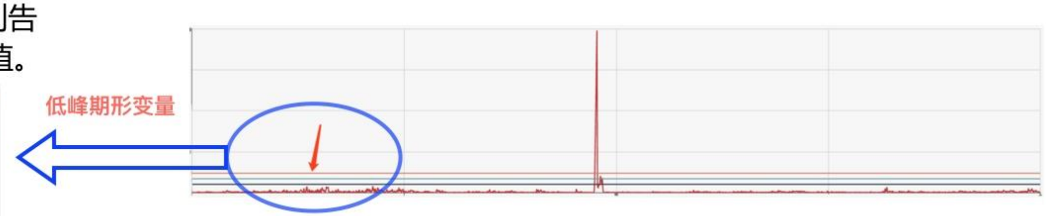
# 案例5

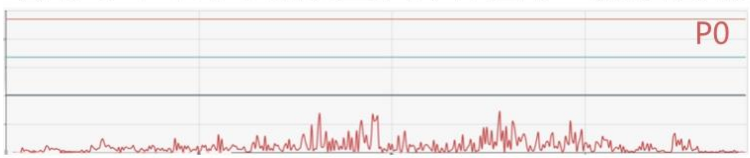
案例5：高峰期与低峰期都有跟基线相比下降5%的异常点，在午高峰时需要进行P0级别告警，低峰期波动经常超过10%可能都不需要告警。



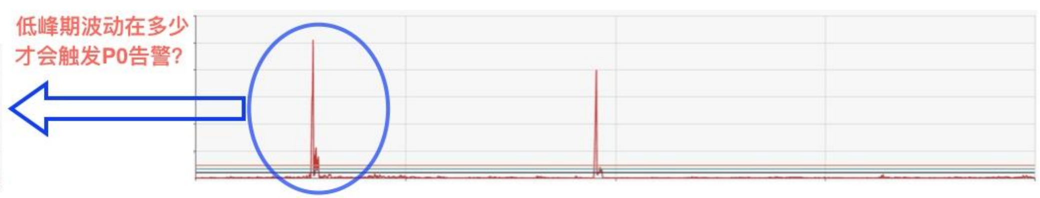



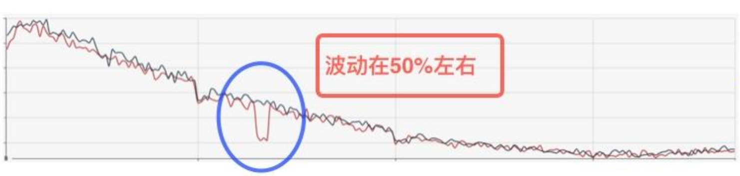
两次处理后，可以看到在低峰期时形变量非常小，达不到告警阈值。在午高峰时形变量非常大，达到P0级别告警阈值。




在低峰期如果想达到P0级别告警阈值，需要波动在50%左右，这个案例体现了形变分析模型在阈值判定上较好的适应性。



# 使用效果

- 覆盖美团外卖核心业务指标2400多个
- 单次异常检测流程时间可以控制在200ms
- 异常检测的精确率、召回率可以达到80%

# 异常检测面临的困难

- 数据没有标签，无法使用监督学习方法
- 噪声和异常点混杂，难以区分
- 不同类型的异常难以区分，无法定义
- 解决思路：无监督学习 + 专家经验