GIAC

主办方：msup | ARCHNOTES 架构

GIAC

全球互联网架构大会

GLOBAL INTERNET ARCHITECTURE CONFERENCE

阿里巴巴立体化智能监控策略的探索和实践

张译尹　　阿里巴巴 高级算法工程师

内容摘要

01

稳定性挑战和对策

- 应急监控策略挑战
- 立体化智能监控

02

立体化智能监控策略

- “分而治之”
  - 监控数据路由算法
  - 量级监控策略：智能基线
  - 成功率监控算法
  - 黄金指标联合判断
- 系统指标监控策略

03

智能监控策略探索及展望

- 基于 VAE 的单指标异常检测
- 多指标异常波动关联分析
- 智能监控未来的展望

阿里业务多样性和复杂性给稳定性带来的挑战

业务数量巨大

- 50+BU 数以万计应用程序

业务形态差异较大

- 电商、金融、云计算、物流、文娱、社交 ...

业务关联复杂

- 用户行为对业务的影响
- 应用程序之间的链路复杂

线上故障需要统一治理：阿里全球运行指挥中心（ GOC ）

心

业务故障统一发现

故障的影响面和根因
需要统一收口和推送

跨 BU 故障协同处理

故障快速恢复
需要统一机制

立体化监控策略的背景

站点
产品/产品线
业务功能/指标
应用/服务 | 中间件/基础服务 | 数据库

实例
虚拟机 | 容器
物理服务器
IDC | 网络

业务监控
线上稳定性
核心重保

应用 / 系统
指标监控
业务异常沉淀
底层指标，
帮助故障收敛

立体化智能监控算法策略

分钟级基本盘重保 + 秒级加速故障收敛

成功率监控算法模块	智能基线	系统指标监控算法策略
成功率指标	量级指标	应用/系统指标

基于深度学习的业务指标路由算法模块

业务特性决定了：线上稳定性依靠业务监控项的重保！

□急控策略挑□□□□

精准监控·少误报

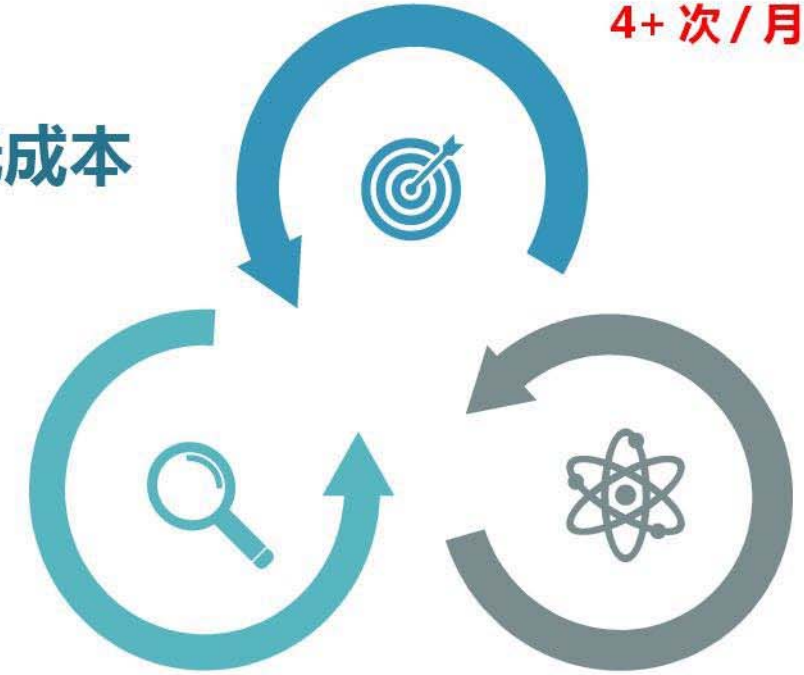
传统监控规则
漏报数量

4+ 次/月

高效监控·低成本

传统监控规则
维护次数

158 次/月



全面监控·少漏报

传统监控规则
误报率

45.29%

我们引入了

立体化智能监控算法

立体化智能监控算法策略

05 系统指标监控策略

01 基于深度学习的路由策略

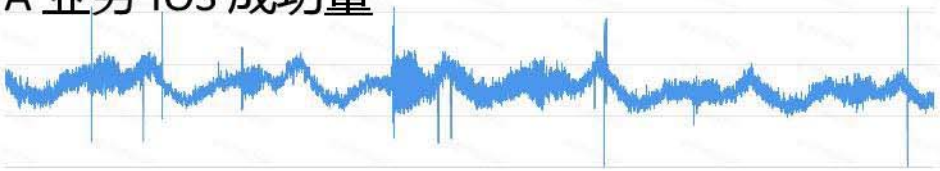
04 成功率智能监控策略

02 量级异常检测智能基线

03 增强版业务监控策略

01 基于深度学习的算法路由策略

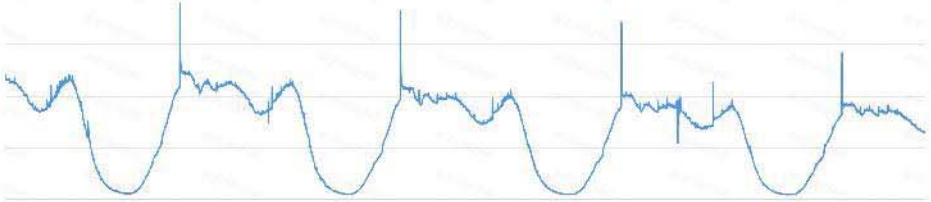
A 业务 IOS 成功量



A 业务 Andriod 成功量

B 业务总量

B 业务成功量

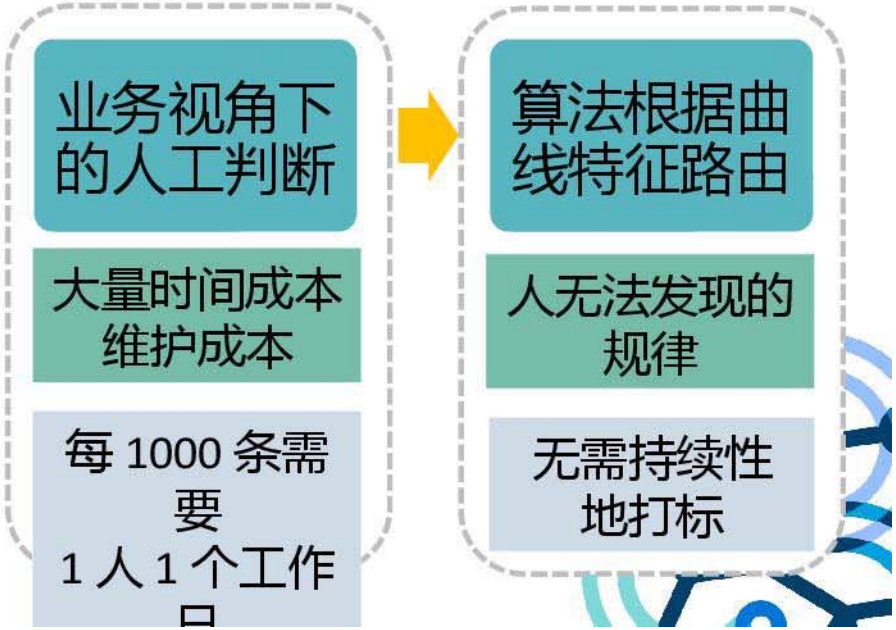


➤ 接入监控算法一共有 **1 万 +** 条业务监控时间序列，新增 **数百** 条 / 天。

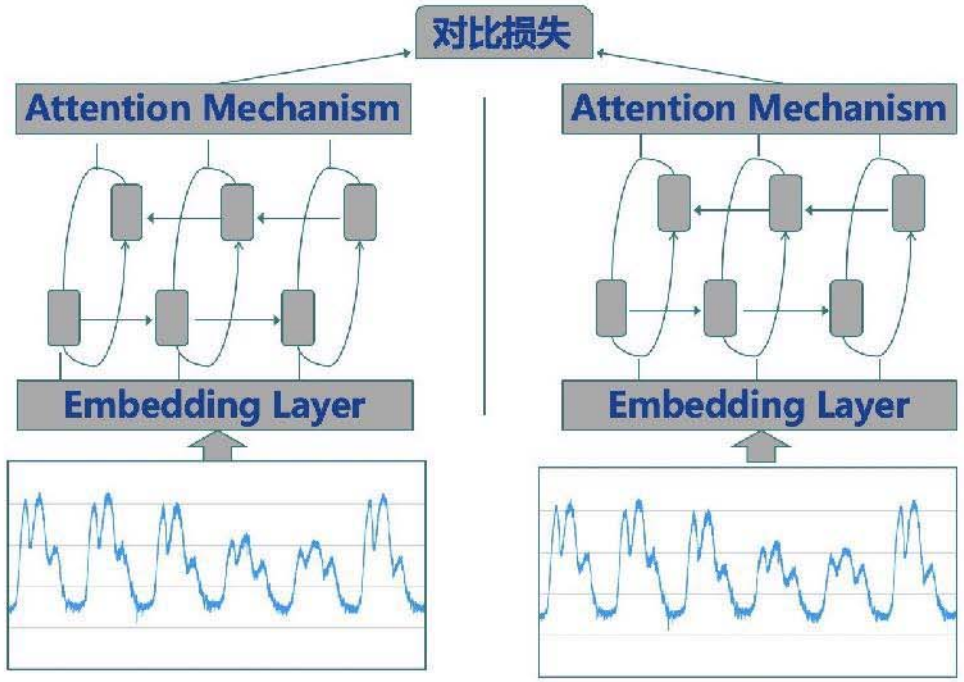
？ 一个“万金油”算法解决 **NO**

➤ 分而治之

□□□□□□□□□□□□□□□□□□□□□□□□□□□□□



01 基于深度学习的算法路由策略



相似性判断

孪生网络 (Siamese Network) [1] 。

时间序列聚合

固定输入长度，减少计算量，加速训练。

Bi-LSTM

同时学习时间序列的正向、逆向顺序关系。

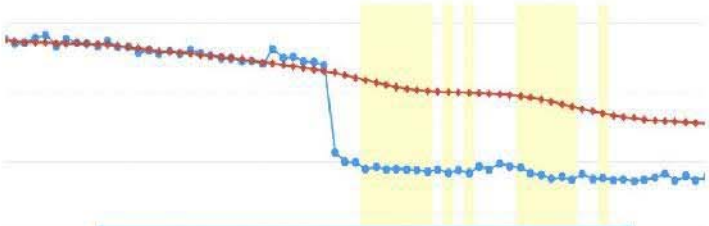
Attention 机制

动态分配权重系数。

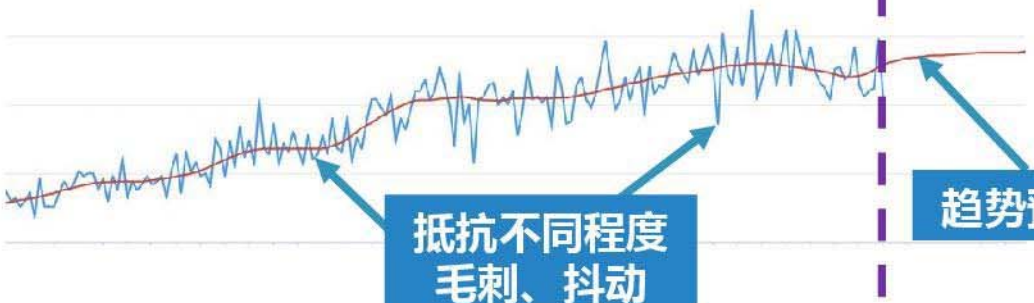
➤ 分类准确率：90%+

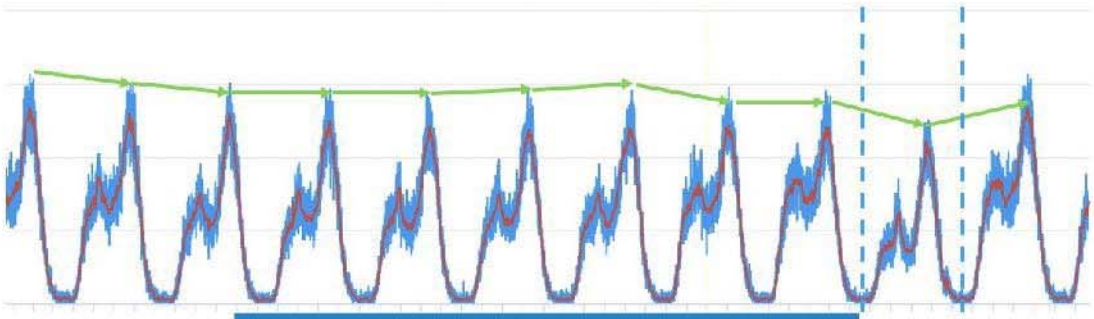
方法 / 效果	聚类效果衡量： Jaccard 系数
FFT + DBSCAN	0.52X
Euclidean Distance + DBSCAN	0.65X
DTW + DBSCAN	0.48X
BI-LSTM-Attention	1.0X

02 业务量监控算法核心：智能基线



智能基线异常检测实例

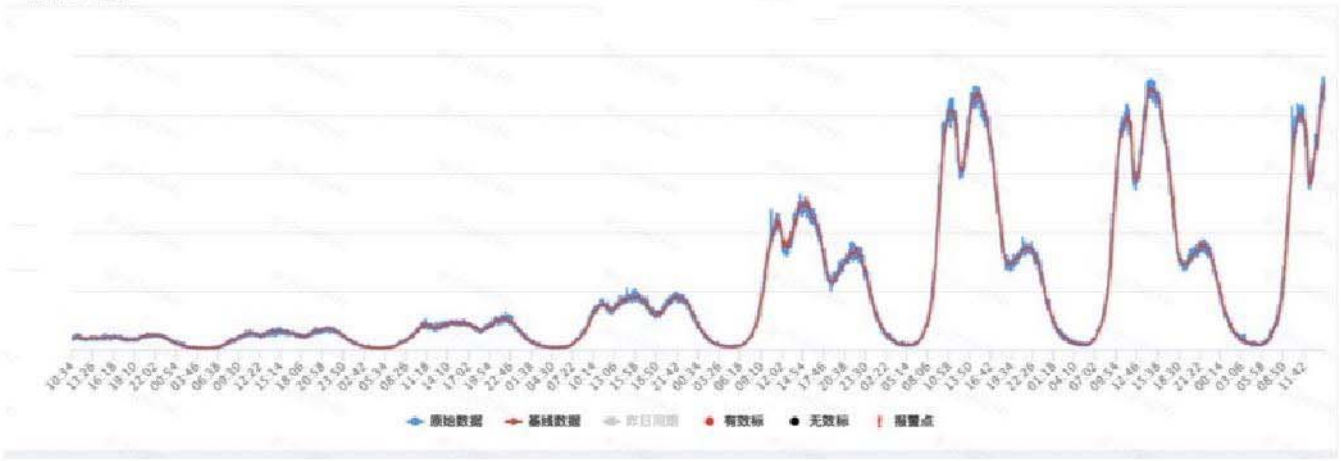




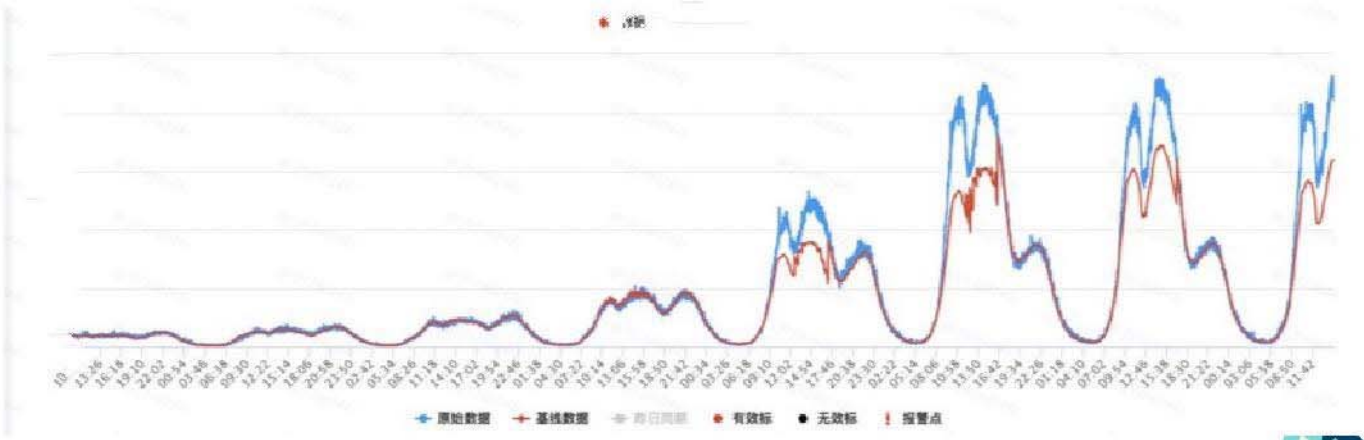
拟合周期性和业务宏观趋势

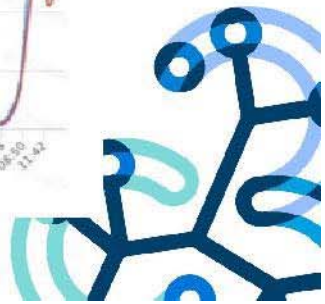
02 业务量监控算法核心：智能基线是怎样炼成的

【新基线】：



【旧基线】





02 业务量监控算法核心：智能基线是怎样炼成的

[难点] 基线问题难以定位

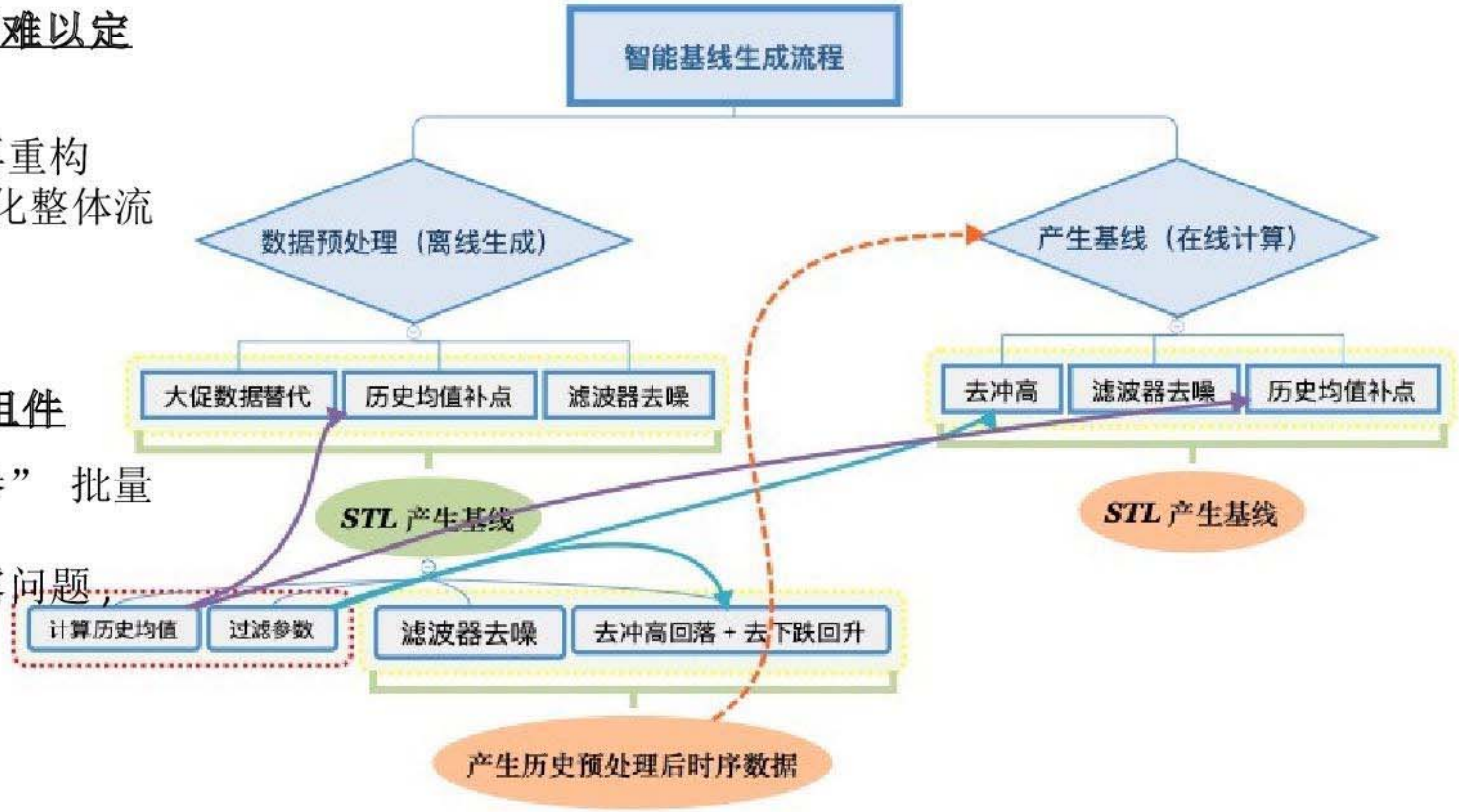
➤ 解构再重构
简化整体流程。

[难点] 众多算法组件

➤ “ 调参” 批量回溯
暴露问题，
逐一解决。

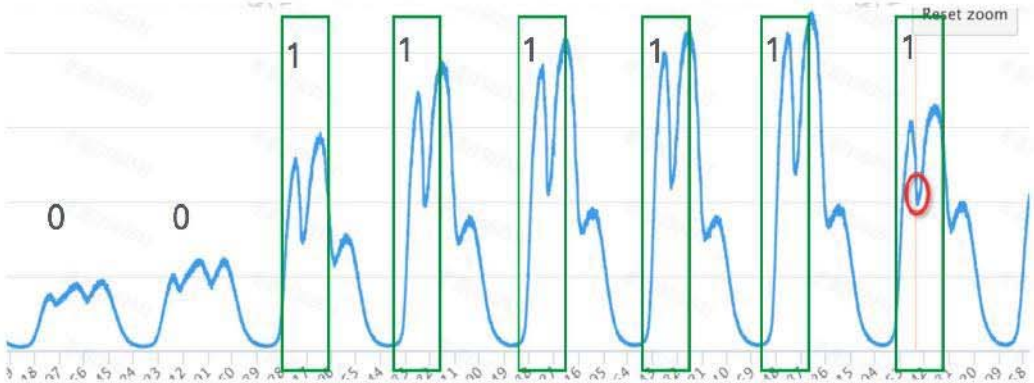
[难点] 耗时控制

➤ 离线预处理
结合在线计算。



03 增强版业务监控算法：异常检测的挑战

□□□□□□□□□□□□□□□□□□□□□□□□□□□□□

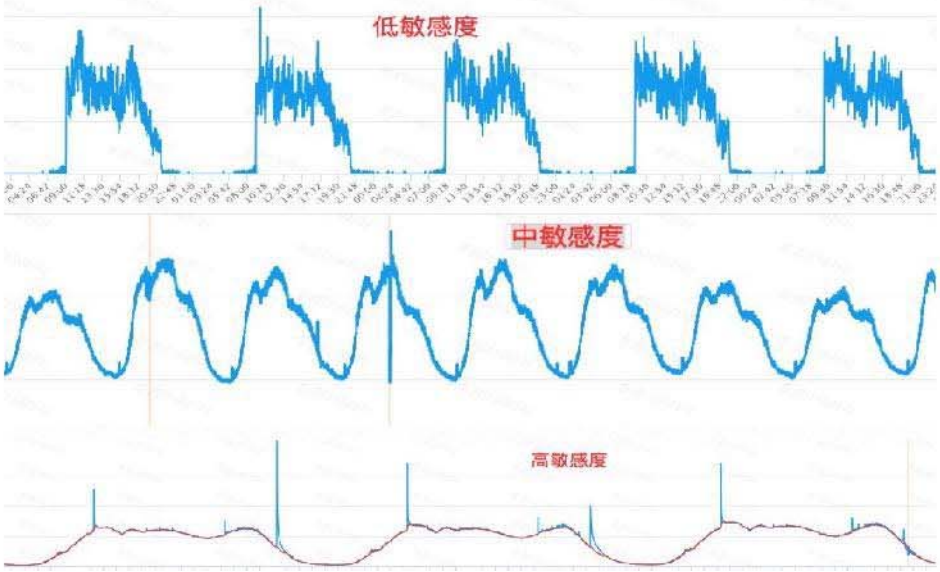


余弦相似度 ✚ DBSCAN

同期时序的夹角信息作为特征

聚类评估同期报警相似性

□□□□□□□□□□□□□□□□□□□□□□□□□□□□□



基线拟合特征 ✚ 时序频域特征

衡量时序数据的周期性

衡量时序数据的稳定性

03 增强版业务量监控策略

—— 让”盯屏”彻底成为历史

基线生成 & 特征工程

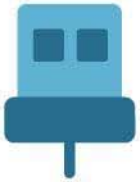
拟合特征：在线 + 离线基线生成局部特征 (LOESS)
异常特征：残差，余弦相似度

报警敏感度

基线质量分
残差分布
用户自定义

异常判别

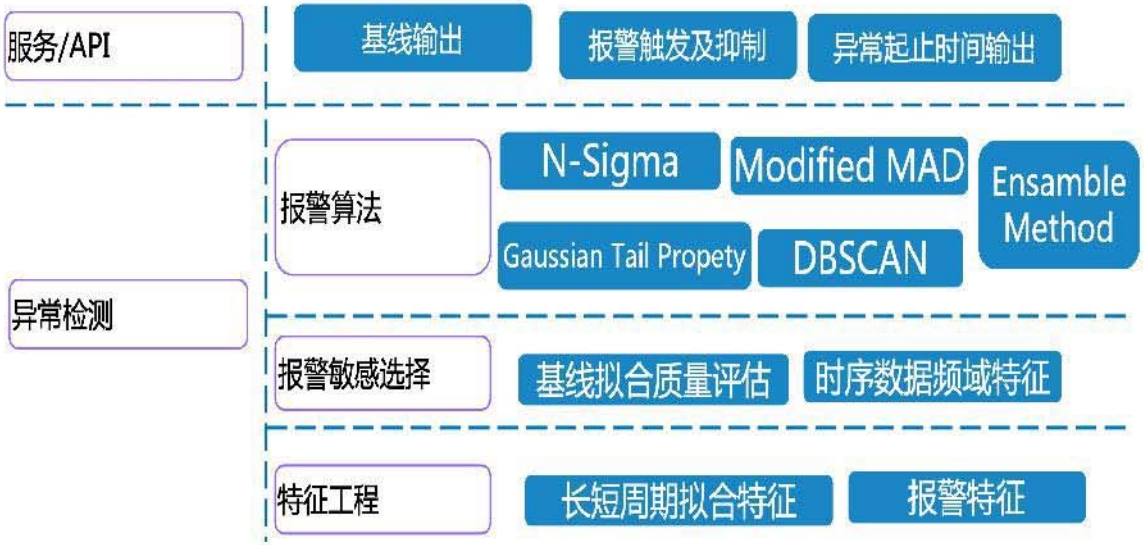
统计策略：N-sigma, MAD, Mutli-Gaussian tail
集成策略：串行



重复异常抑制

局部拟合特征
动态对齐相位差
DBSCAN

异常检测流程



业务成果

- 基线质量（鲁棒性）大幅度提升
- 故障发现召回率：80%+
- 故障发现准确率：90%+

- 秒级监控提效：<30s .vs. 传统分钟级策略 >60s
- 秒级全自动化故障通告时效性：<2min .vs. 传统通告 5min

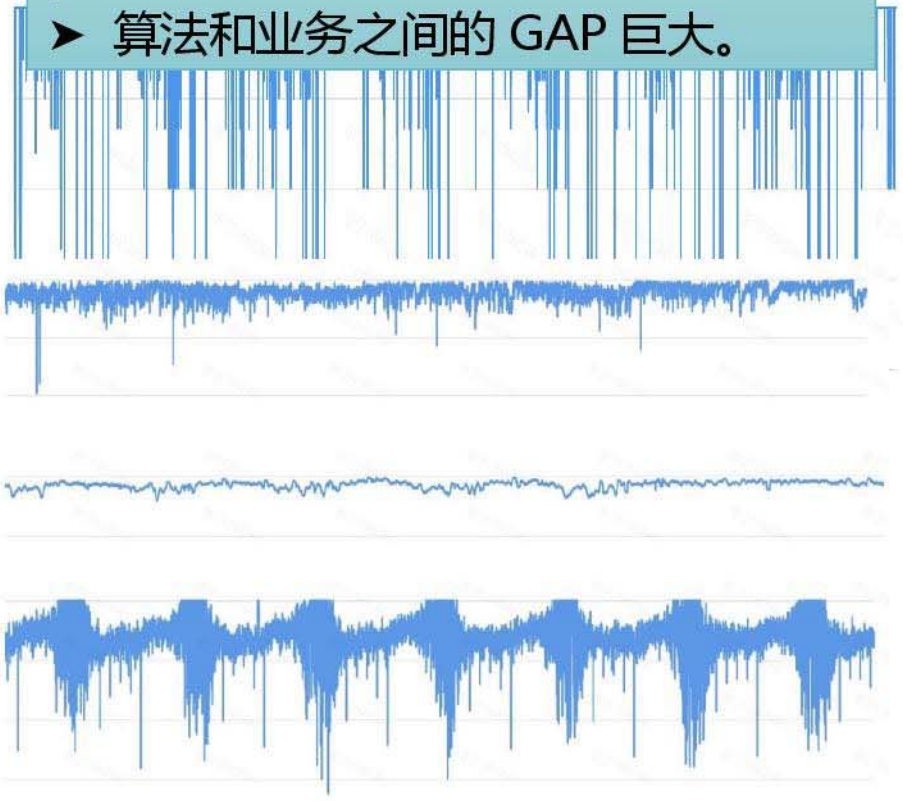
04 成功率监控策略

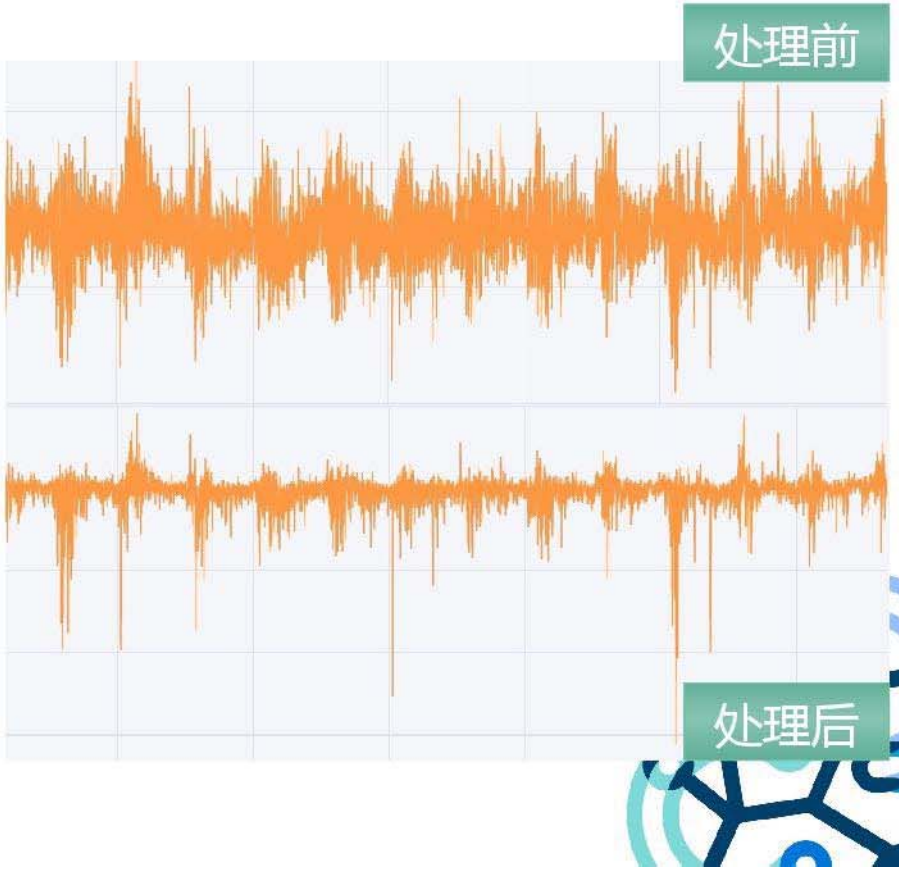
成功率监控？直接配置 <95% 报警不就可以了吗？

[成功率异常检测难点]

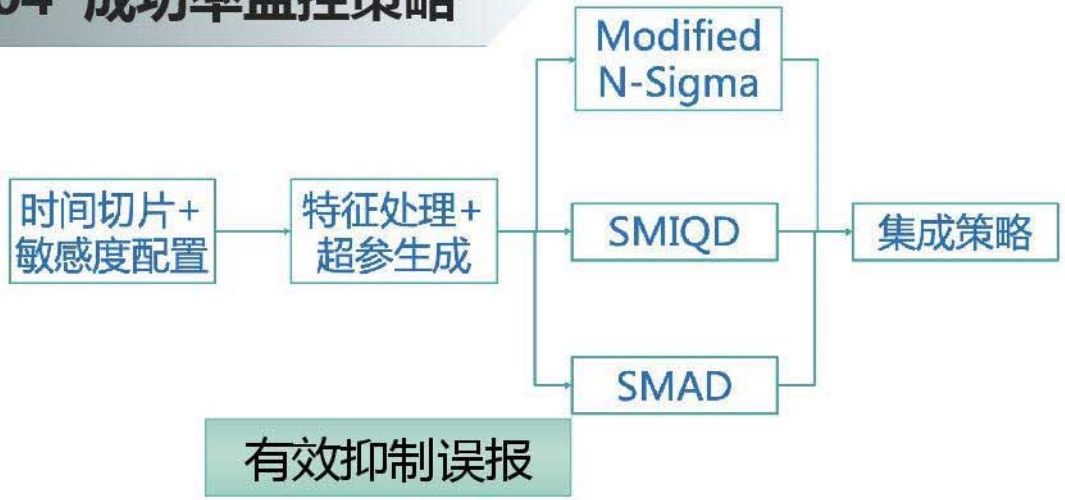
- 频繁持续抖动，并非稳定在100%。
- 不同时间片，成功率抖动幅度不同。
- 算法和业务之间的 GAP 巨大。

结合 exponential activation 对同环比残差数据进行噪声抑制





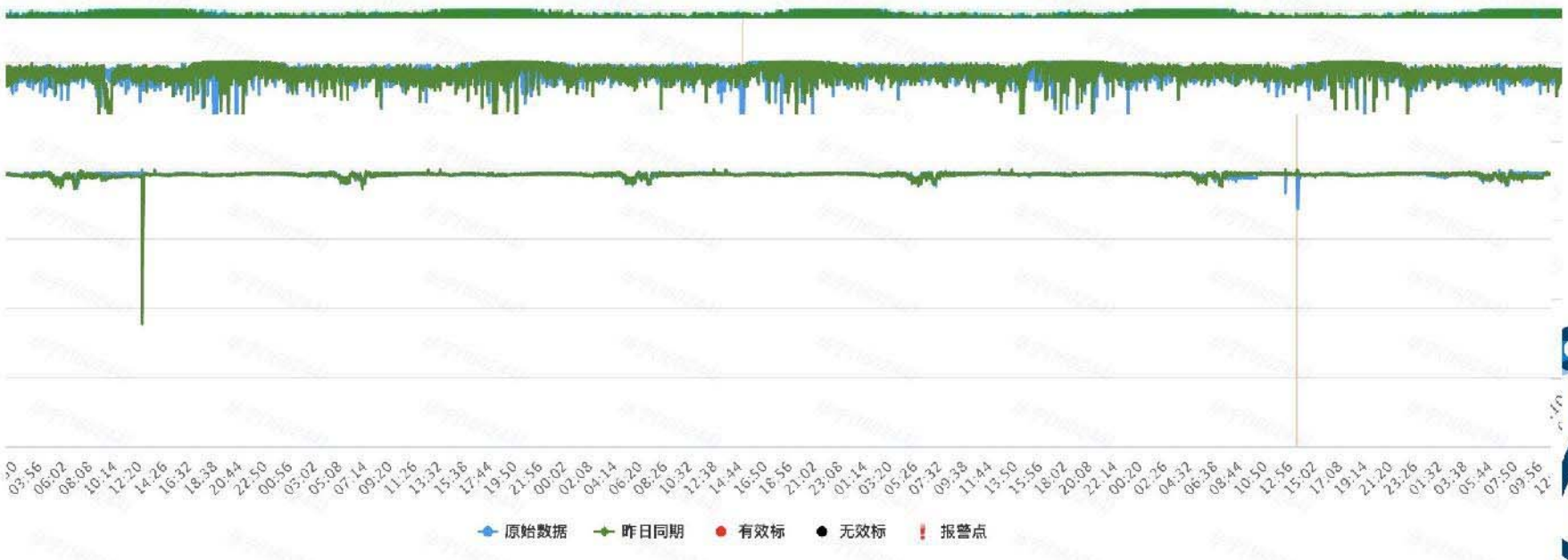
04 成功率监控策略



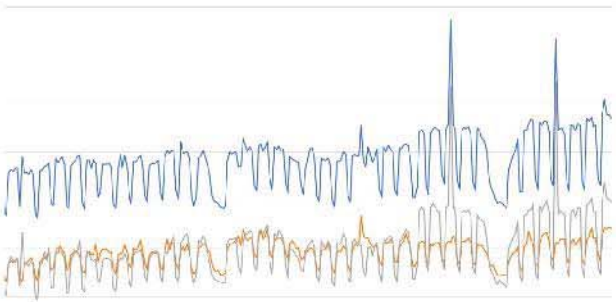
- 报警准确率：85%+
- 故障召回率：90%+

有效抑制误报

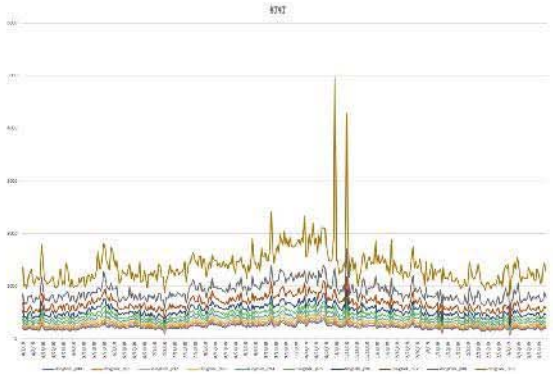
稳定异常召回



05 系统指标监控的困难与挑战



集团每天有 2W+ 同学接收报警



90% 的同学日报警量在 300+ 条

多数报警来自系统 / 应用级监控

报警数量大

报警处理不及时，而这背后的原因可能是大量报警造成的风暴

报警质量低

一线研发 / 运维同学往往用大量误报的代价换取少量召回

维护成本高

业务变化、应用迁移、混合部署等因素都可能导致监控规则需要重新配置，监控维护成本高

05 系统级指标无阈值智能监控策略

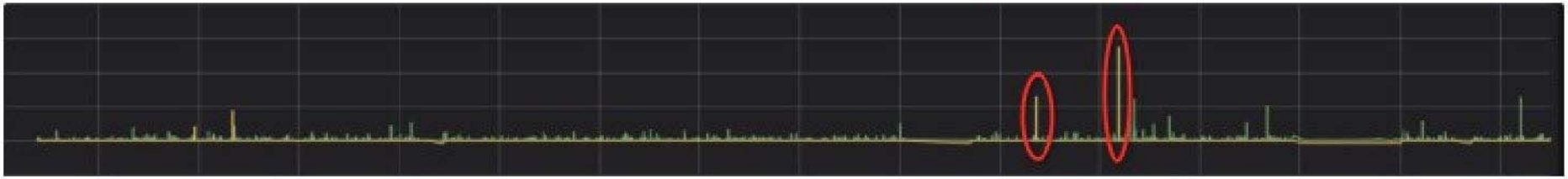
- 使用场景
  - 机器基础指标 — cpu, load, Disk I/O, memory, JVM
  - 应用监控指标 — 错误码检测
- 算法优势
  - 免阈值规则配置成本
  - 算法配置、后续维护成本低
  - 抑制大量误报
    - 抑制历史出现过的规律异常波动造成的误报
    - [Disk I/O场景] 异常召回保持90%，误报降低90%

算法定时学习前 7 天的数据获得模型，然后每分钟实时异常检测。

05 系统级指标无阈值智能监控策略

- 准确 / 召回 : 85%
+

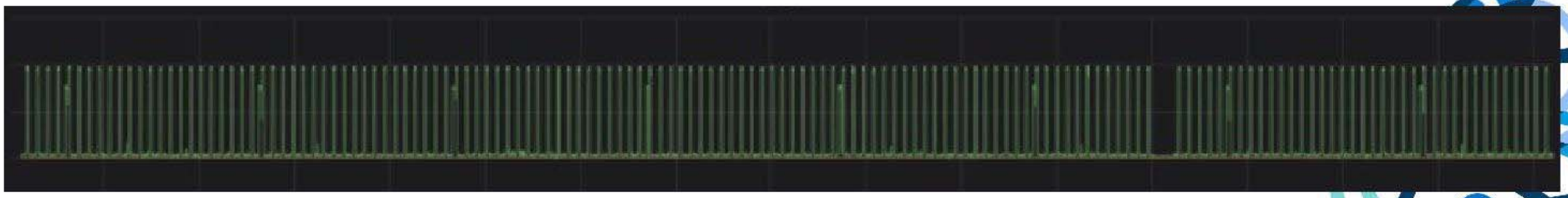
- 自动发现历史上不常见的异常波形。



- 系统波形特征变化后，算法会自适应学习。

- 对于持续时间内异常的判断

- 自动学习经常频繁出现的波形特征。



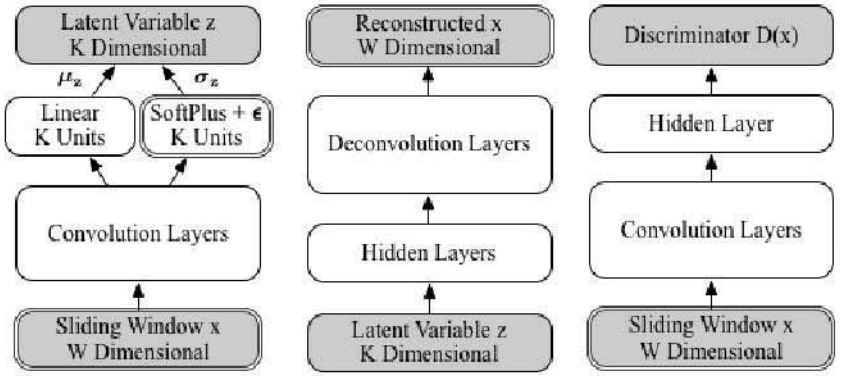
智能监控策略探索及展望

基于深度生成模型 VAE 的异常检测方案探索

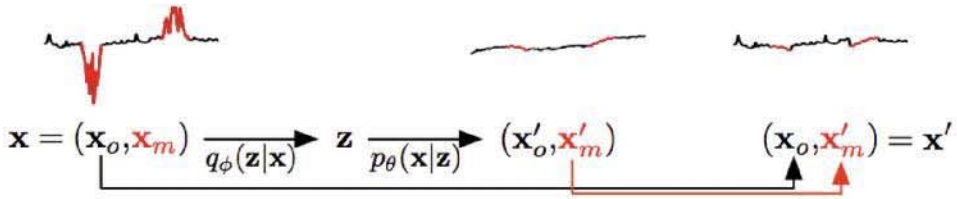
清华学术合作

Unsupervised Anomaly Detection via Variational Auto-Encoder for Seasonal KPIs in Web Applications
WWW 2018: The 2018 Web Conference

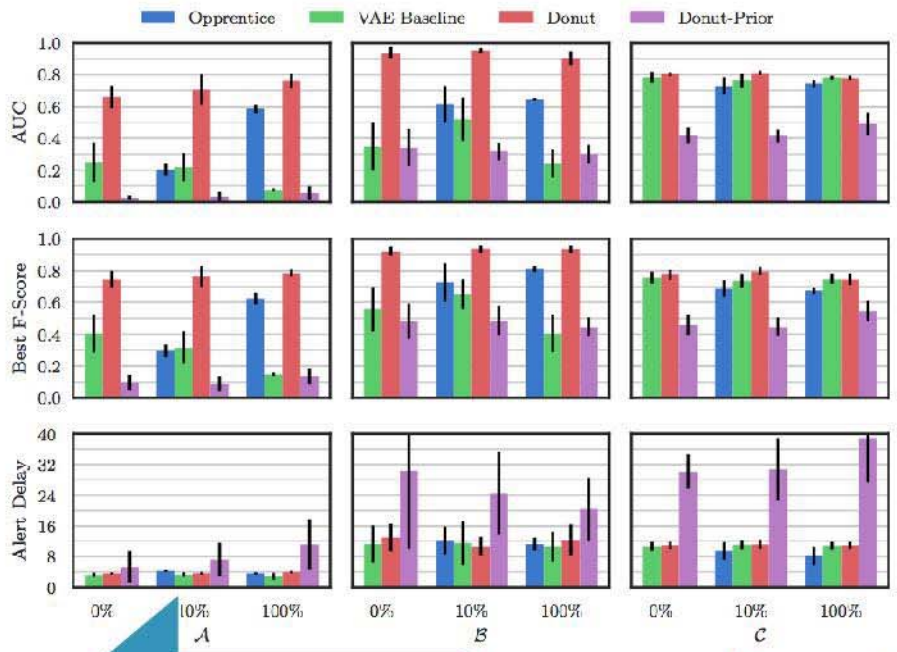
VAE-WGAN 框架



Variational　　Generative　　Discriminate



数据增强： MCMC



优势
- 无人工特征工程成本
- 模块召回异常能力强

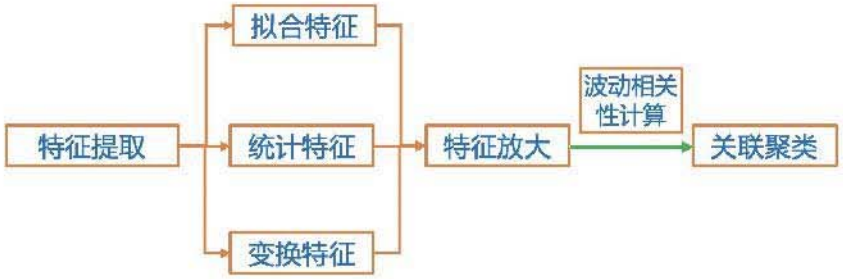
不足
- 泛化能力弱
- 异常判别参数确定成本高

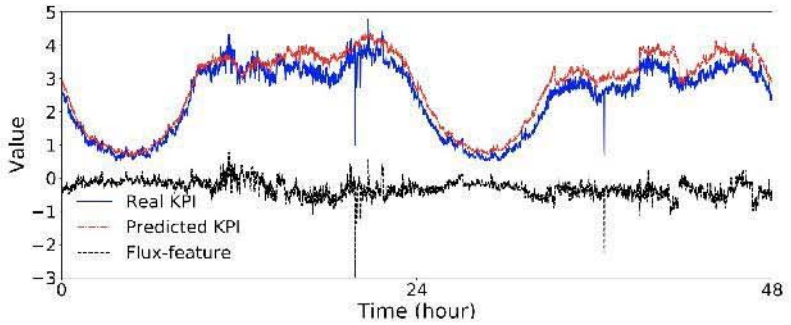
智能监控策略探索及展望

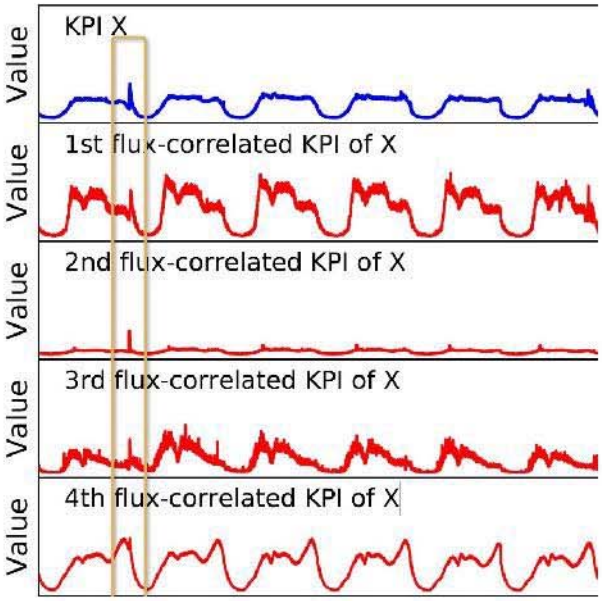
多个监控指标异常波动相关性的检测

CoFlux: Robustly Correlating KPIs by Fluctuations for Service Troubleshooting 清华学术合作

IEEE/ACM IWQoS 2019







- 时间序列的波动相关性：同时（同向 / 异向）波动、相继波动
- 异常片段的相关性比整体曲线的相关性更加需要关注；
- 异常波动相关性可用于：故障预警、根因分析。

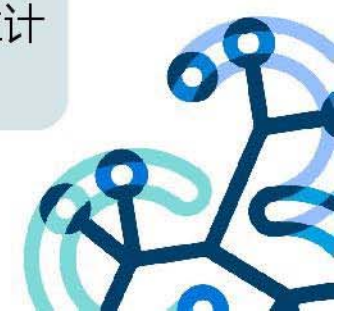
▻ 生产数据效果对比

- 80.51%
- 异常波动相关性计算准确率

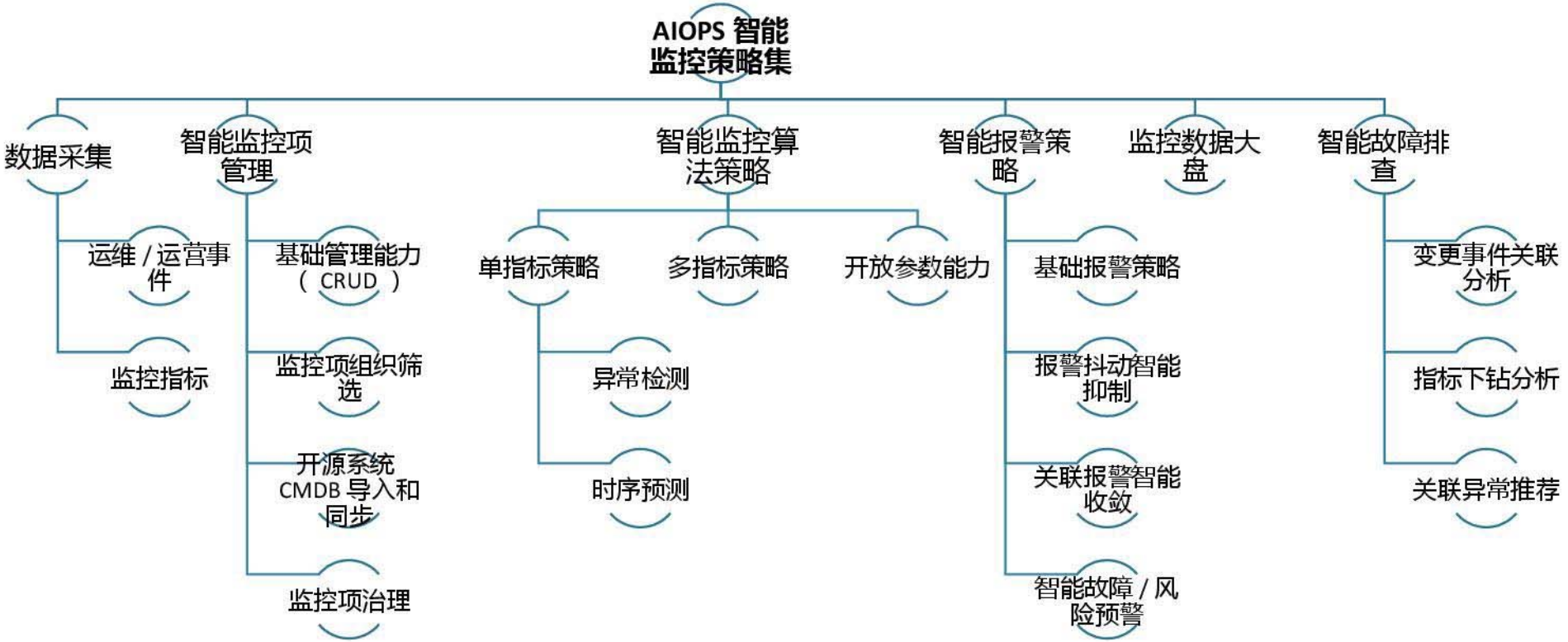
VS

- 55.94%
- 整体波形相关性计算准确率

▻ 局限： 特征工程成本 & 计算性能



智能监控策略未来展望


AIOPS 智能监控策略集
数据采集
运维 / 运营事件
监控指标
智能监控项管理
基础管理能力（ CRUD ）
监控项组织筛选
开源系统 CMDB 导入和同步
监控项治理
智能监控算法策略
单指标策略
异常检测
时序预测
多指标策略
开放参数能力
智能报警策略
基础报警策略
报警抖动智能抑制
关联报警智能收敛
智能故障 / 风险预警
监控数据大盘
智能故障排查
变更事件关联分析
指标下钻分析
关联异常推荐