

O'REILLY®

Velocity

CONFERENCE

BUILD RESILIENT SYSTEMS AT SCALE

velocity.oreilly.com.cn

#velocityconf

智能运维在监控中的探索

Baidu Intelligent Operation 

曲显平

quxianping@baidu.com

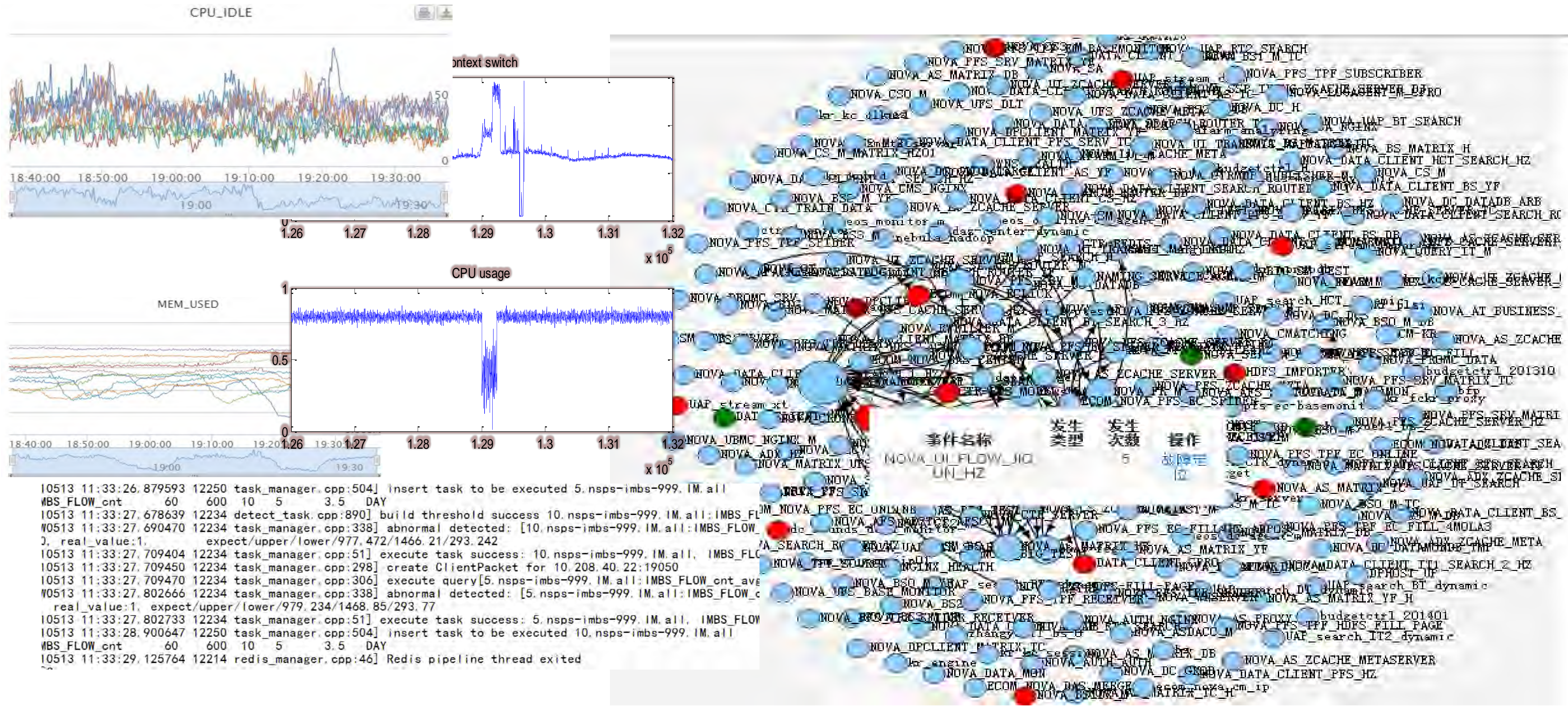
 Bai du 百度

背景：服务器爆发式增长

百度服务器数量变化趋势



监控和问题诊断很复杂



监控规模

■ 监控数据

- 这里只讲时序指标数据（不含日志）
- 服务器指标数量：>1亿
- 业务指标数量：>8千万
- 数据增长速度：50TB/日



问题发现

- 指标繁多
 - 管理配置监控复杂
 - 误报、漏报严重
- 报警风暴
 - 太多噪音
 - Oncall人员应接不暇

问题诊断

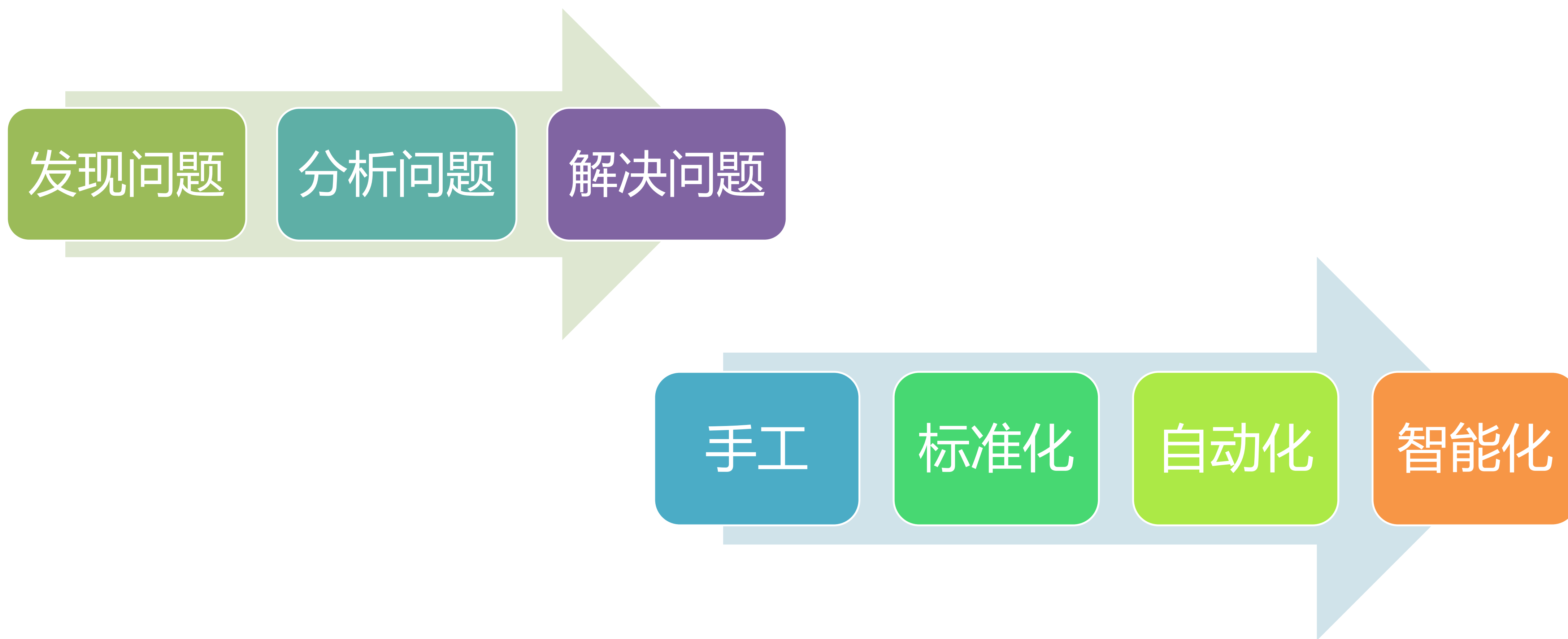
- 故障定位
 - 没有服务拓扑
 - 指标与事件之间缺乏关联关系
 - 基本靠人的经验

问题解决

- 止损
 - 不是服务自恢复
 - 基本靠人工操作

运维&监控的演变

- 运维人手不足，如何解决日益复杂化的运维问题？



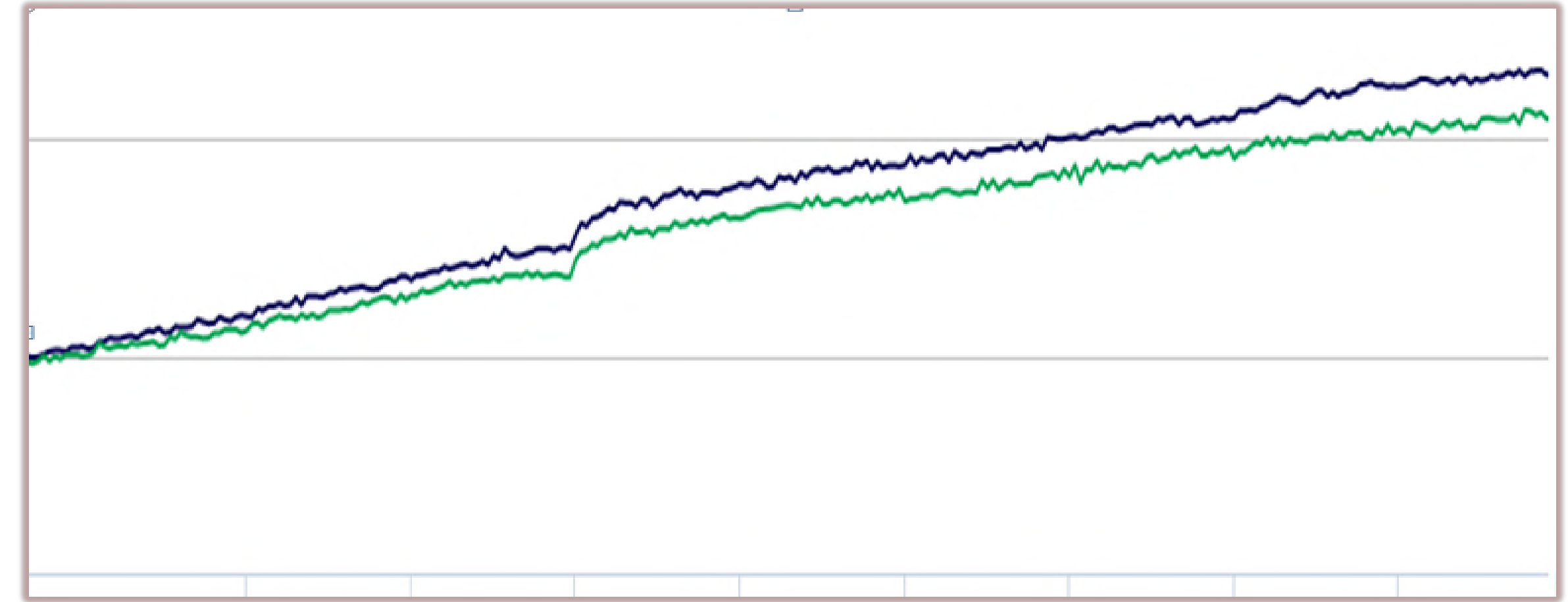
发现问题篇：异常自动检测

如何合理设定阈值？

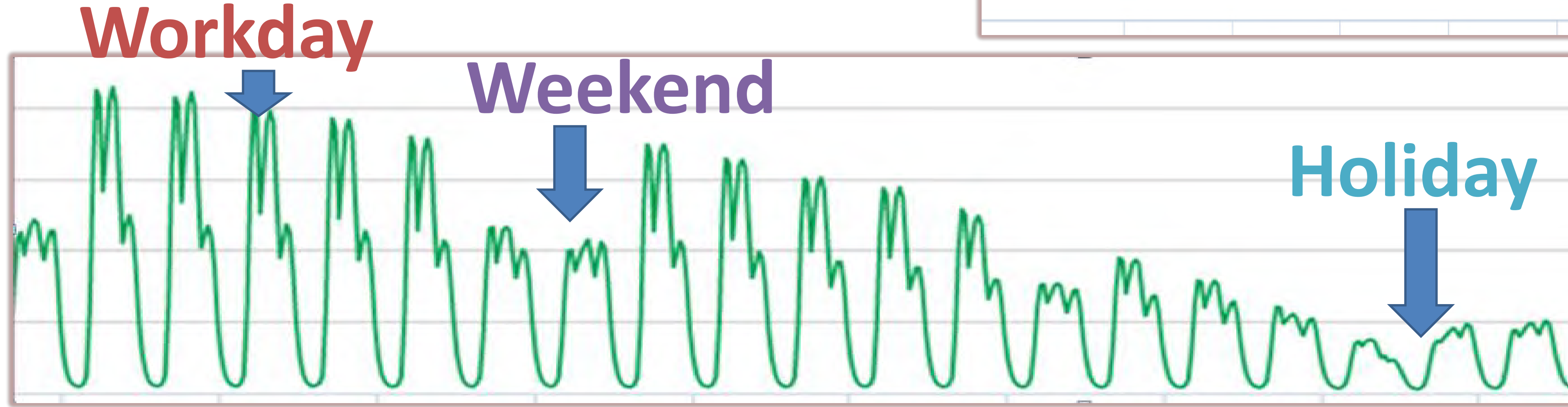
环比昨日的明显问题



持续偏离的明显问题



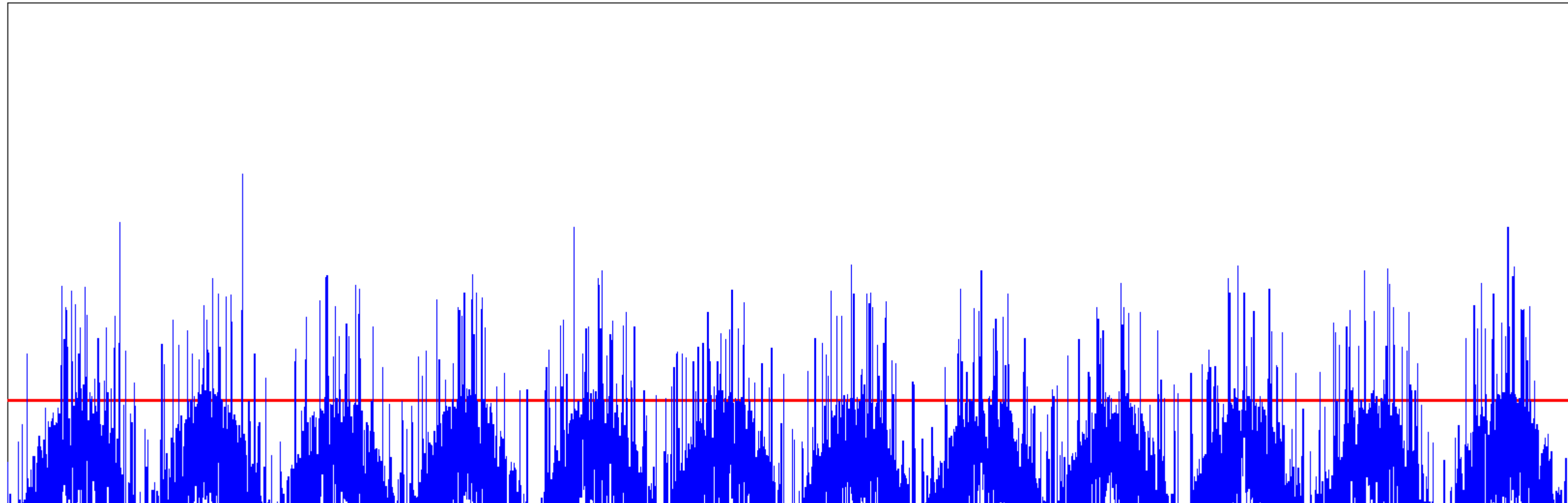
随时间周期漂移的指标数据



恒定阈值设定

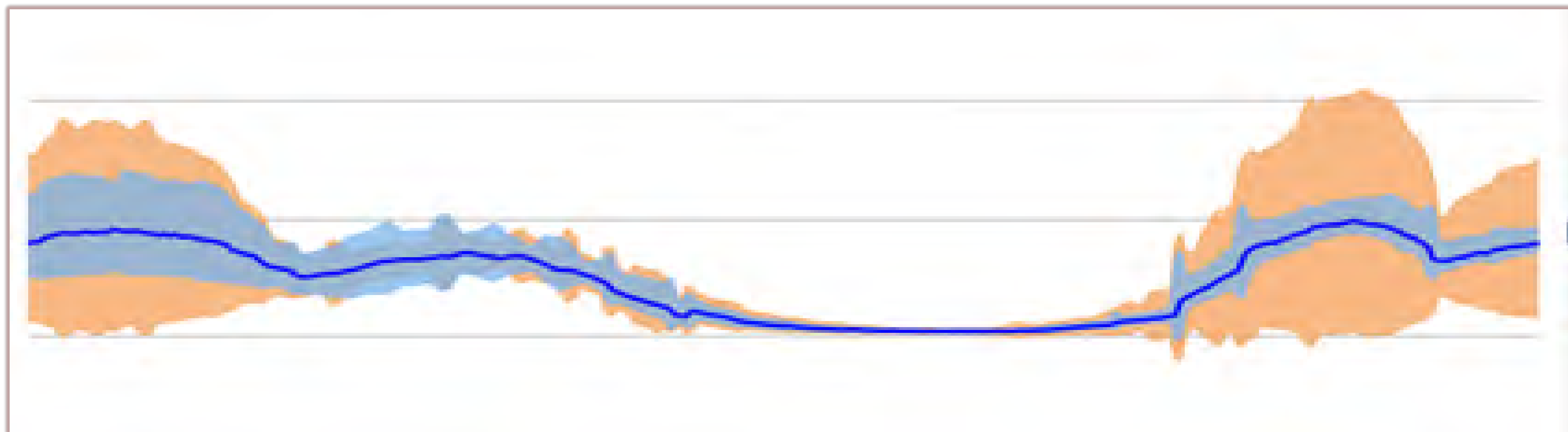
- 简单恒定阈值的计算
 - 基于历史数据统计
 - 假设正态分布
 - 3-sigma $\bar{x} \pm 3\sigma$

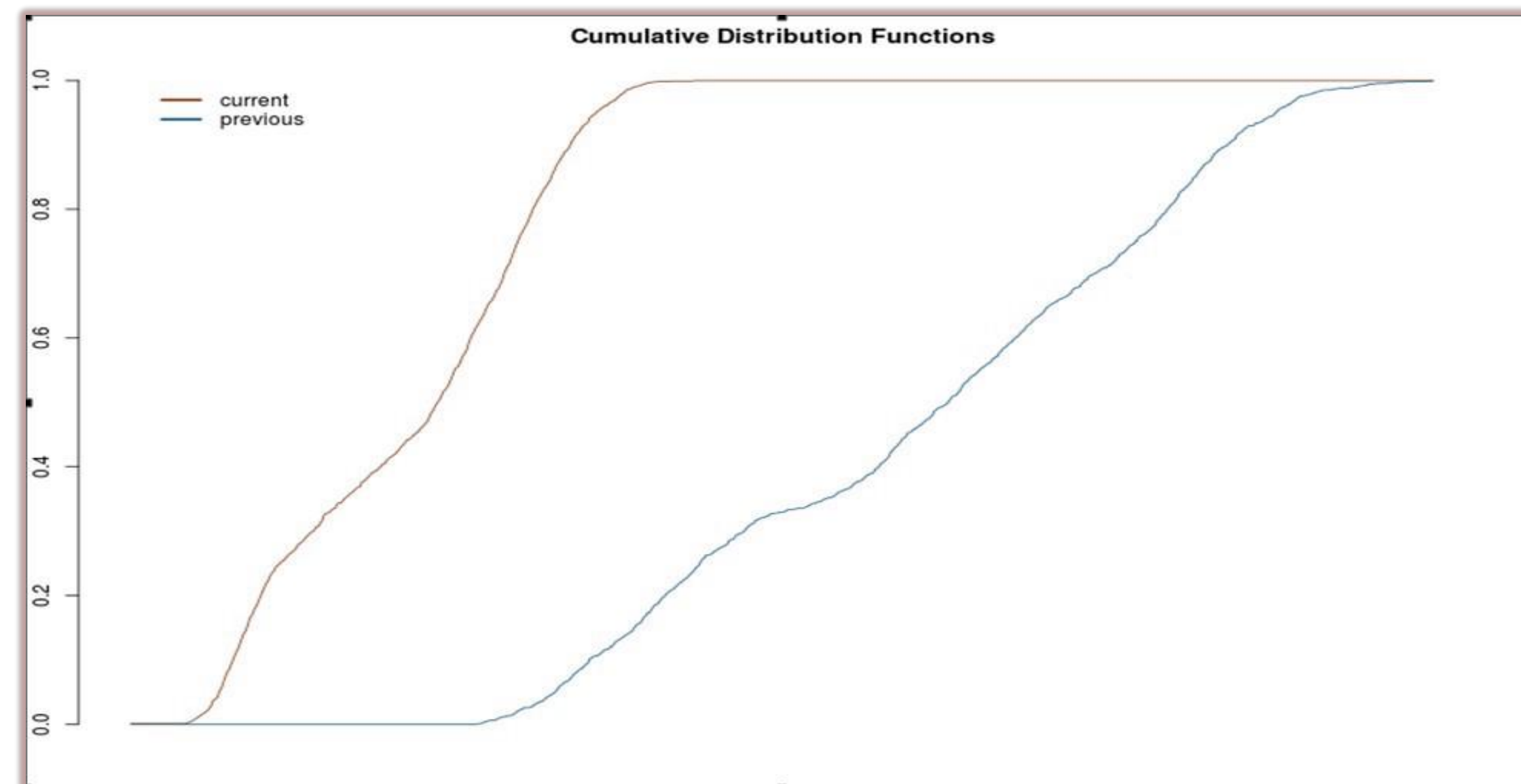
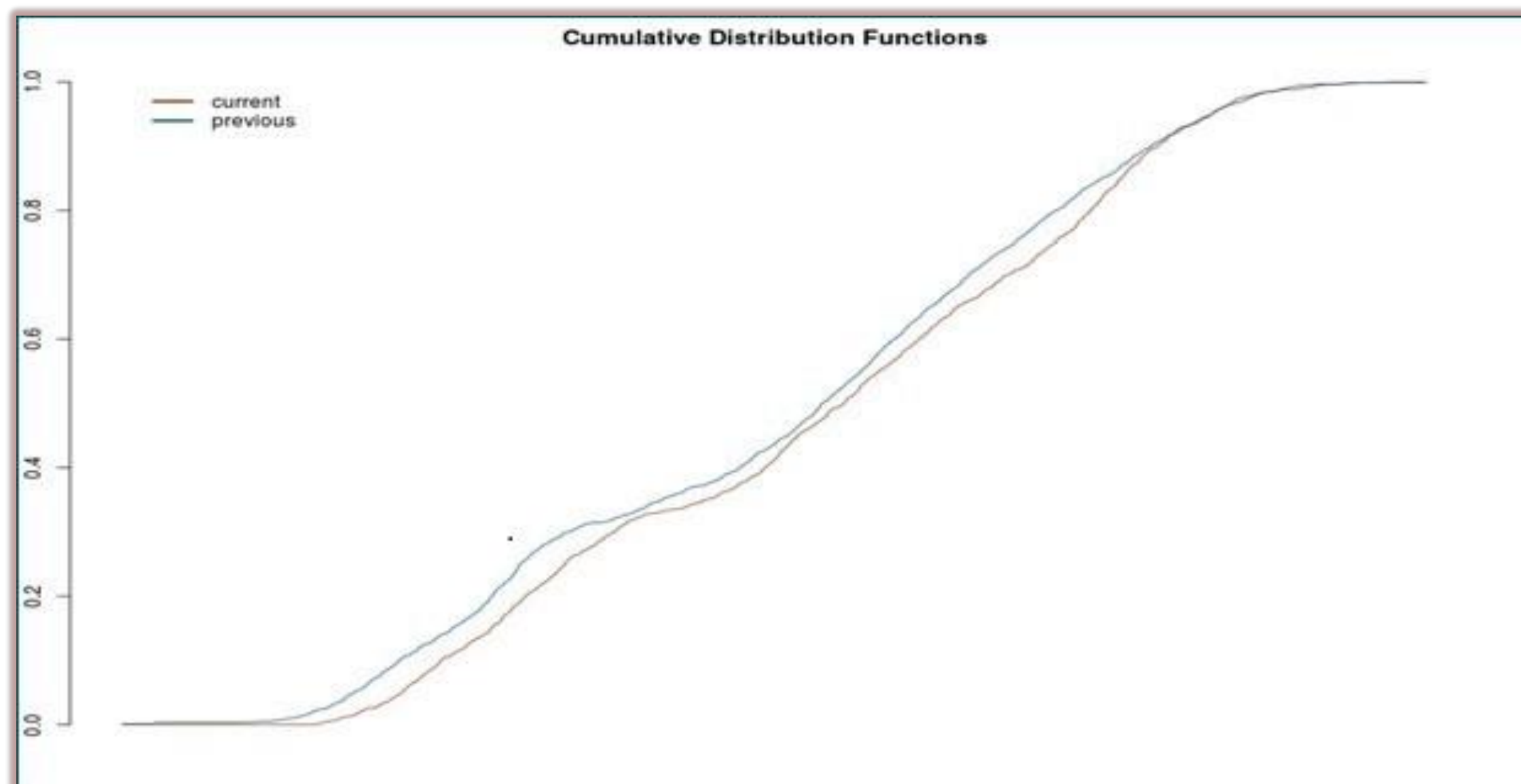
3 Sigma Rule



动态阈值的计算

- day vs night
- 多分布形式：将数据分段（ e.g. window=15min ）
- 按天同期计算统计阈值
- 分段3-sigma策略



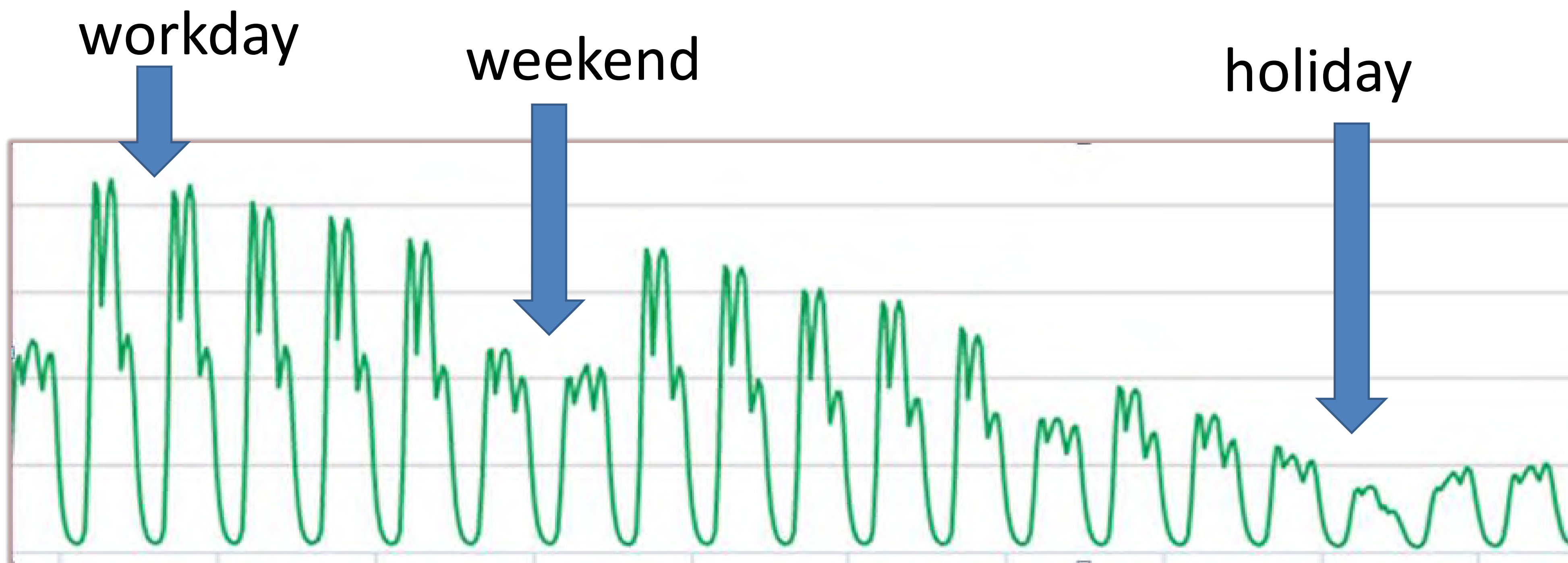


■ 区分数据是否为多分布：

- KS-test
- 选择恒定阈值法或动态阈值法
- 也可区分工作日/周末/长假

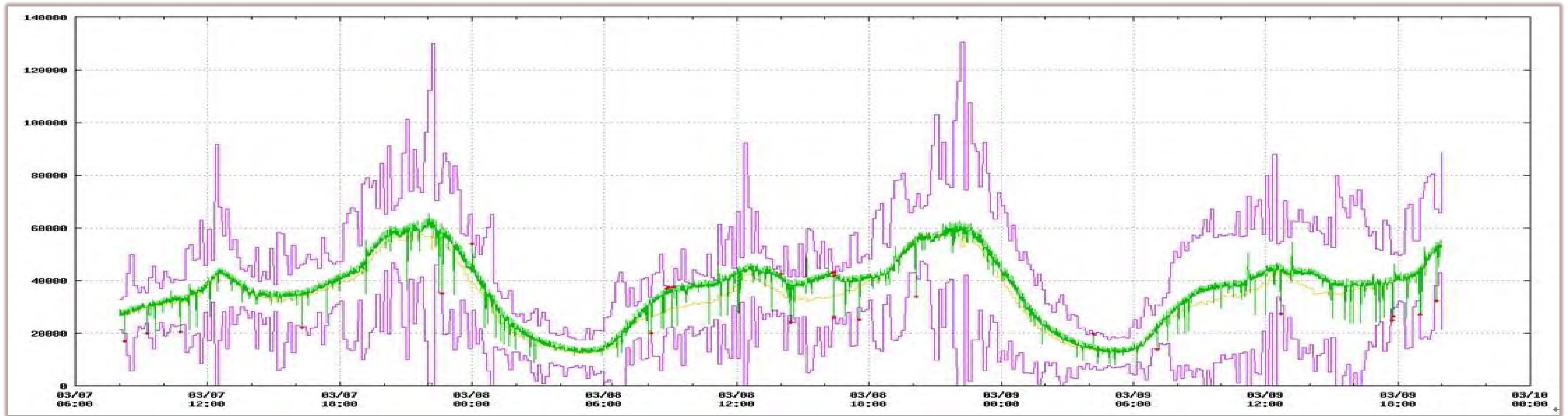
数据漂移

- 指标随时间周期性漂移
 - 工作日流量 vs. 周末流量 vs. 长假流量



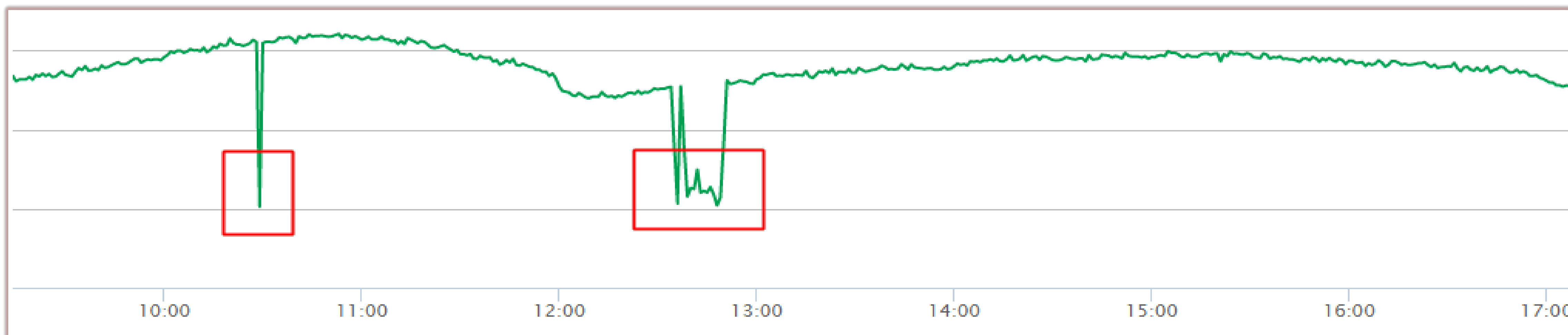
复杂动态阈值法

- Holt-winters & ARIMA
- 季节性和趋势性同时学习



突增突降检测

- 局部平滑
 - 局部回归：LOESS
- 速度



缓慢下降变化的判断

- 动态窗口累积计算
 - 监控系统精细采集的弊端
 - 10s, 30s => 15mins, 1hour, 24hours
 - 对动态累积窗口应用前述方法

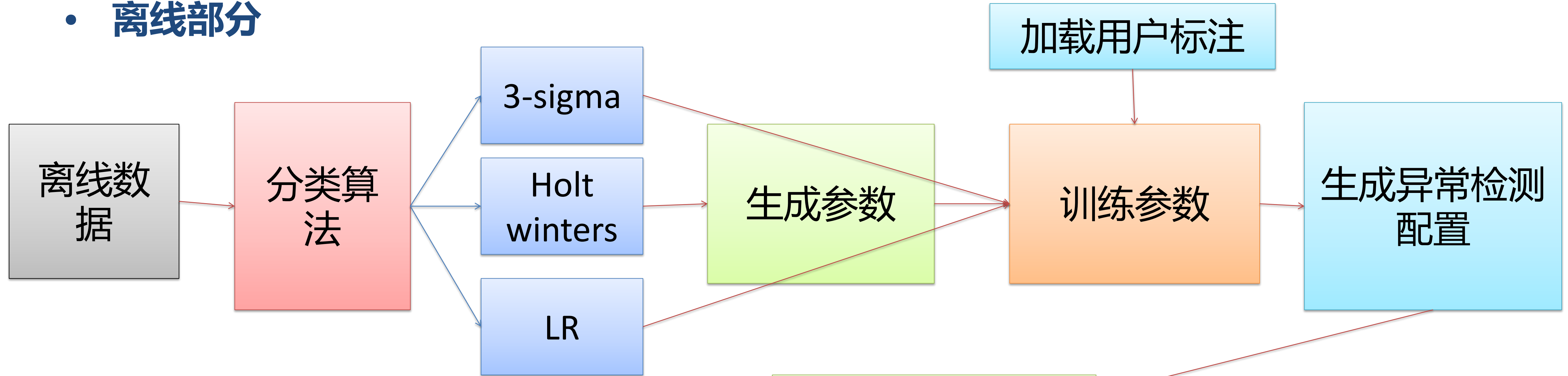
自动生成的阈值 vs. 业务需求

- 工程师标注
 - 修改参数
 - 标记未检测到的异常
 - 标记错误的报警

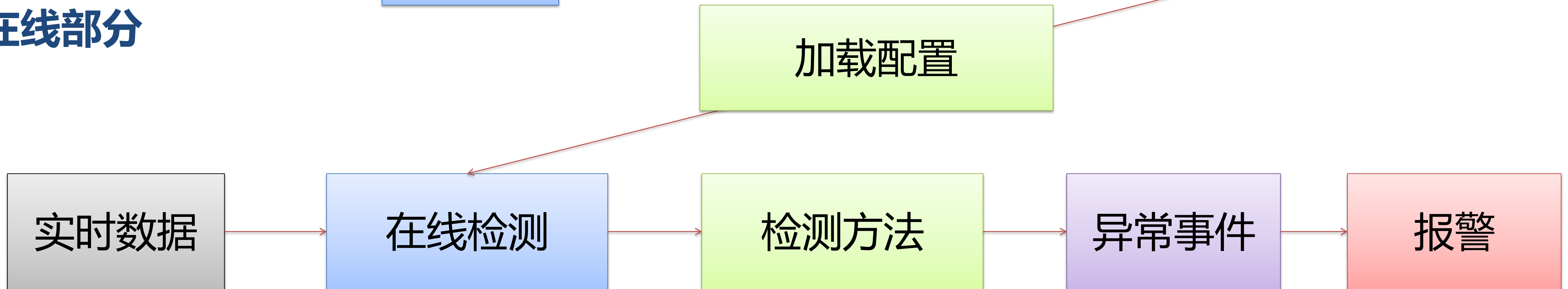
- 机器学习
 - 标注报警 \Rightarrow 参数训练 \Rightarrow + / -

异常检测系统

• 离线部分



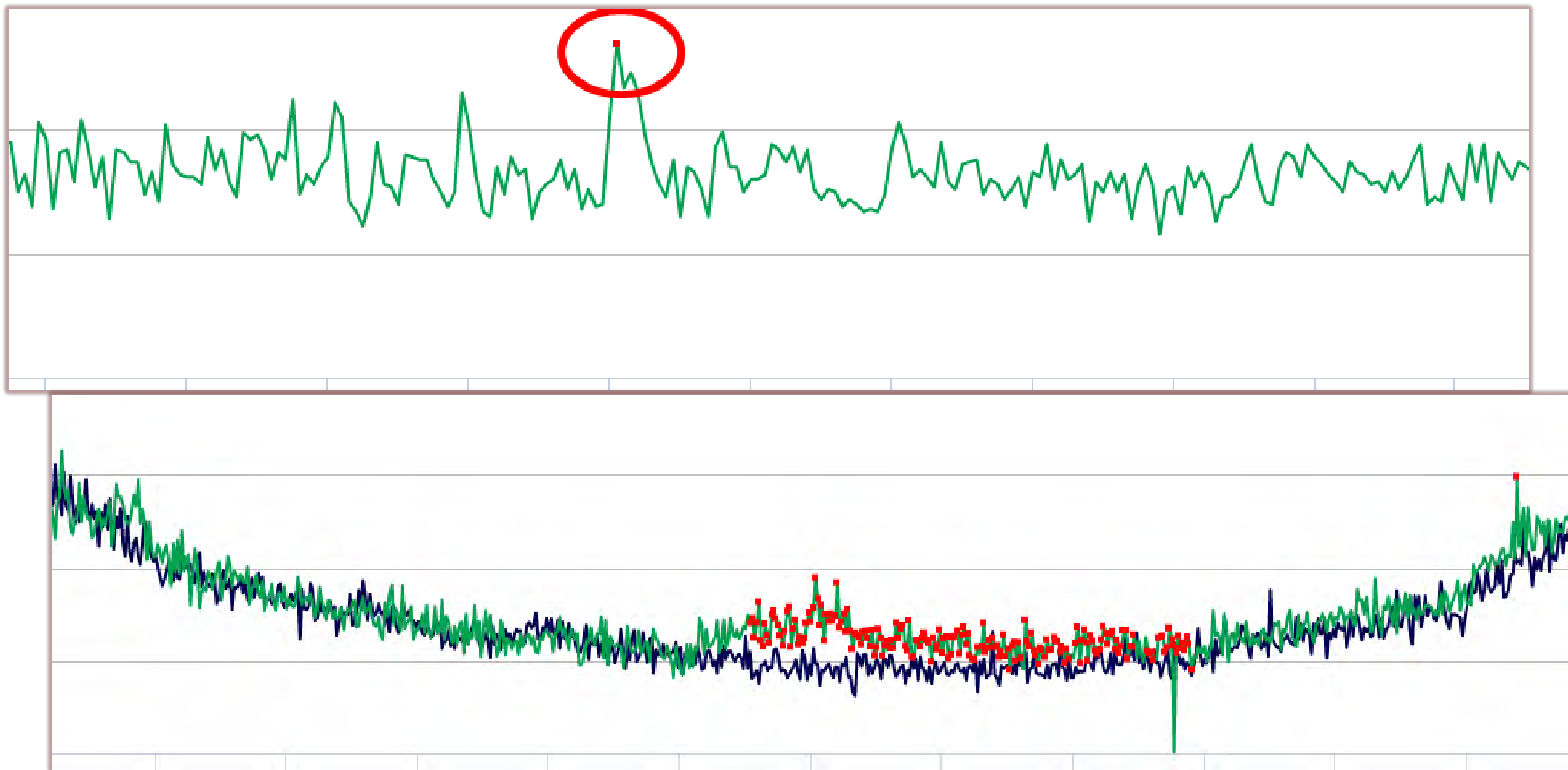
• 在线部分



发现问题篇：精准报警

报警风暴

- 是否每次异常都应该报警？



异常过滤

- Viterbi
- 将异常点转化为异常事件或状态



报警合并

■ 简单策略

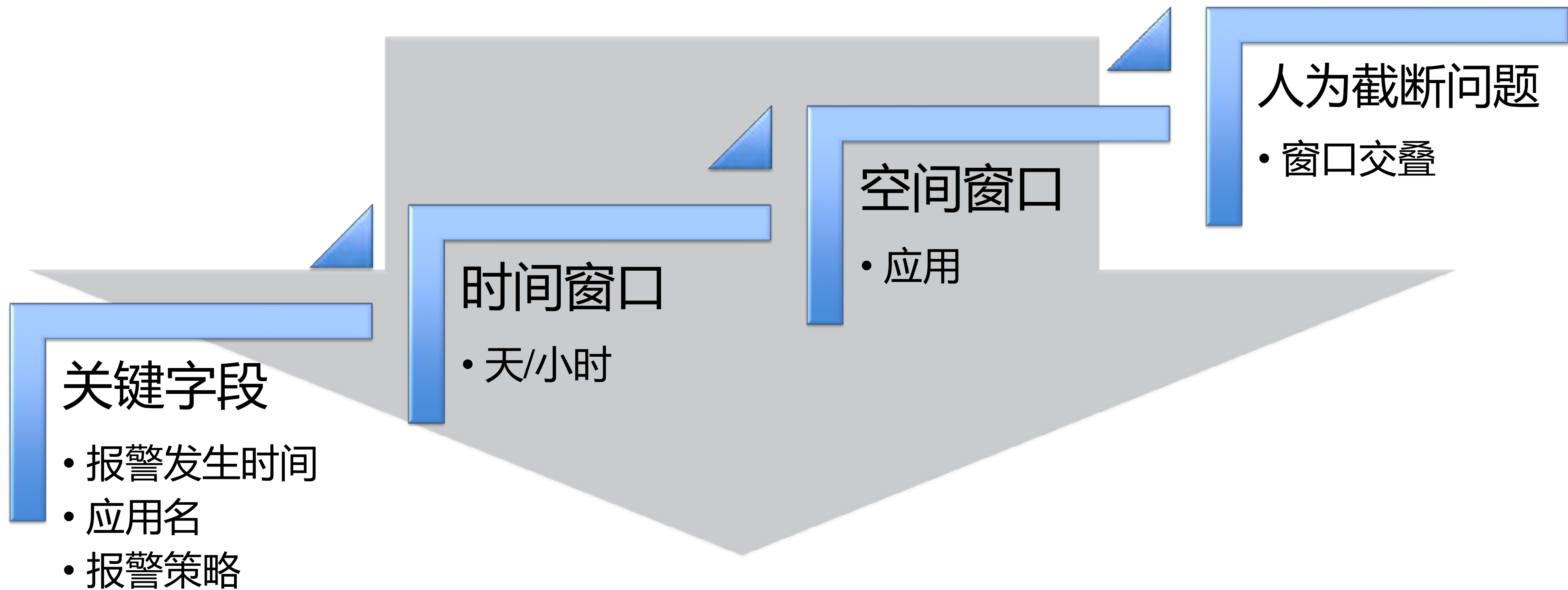
- 固定时间窗口
- 相同监控策略
- 相同监控对象

■ 复杂策略

- 关联挖掘
- 找到报警的频繁项集
- 自动合并置信度较高的频繁项集

报警的频繁项集挖掘

2014-11-27 21:33:25 , appui-81.pro.all , appui-81.pro.all:instance:conn_php_failed_cnt_gt_0
2014-11-27 22:43:25 , back-13.pro.all , back-13.pro.all:instance:runtime



pro_2014-11-26 appui-81.pro.all:instance:conn..., back-13.pro.all:instance:runtime..., ...
pro_2014-11-27 appui-81.pro.all:instance:conn..., back-13.pro.all:instance:runtime..., ...
pro_2014-11-28 appui-81.pro.all:instance:conn..., back-13.pro.all:instance:runtime..., ...

报警依赖&升级

- 报警依赖

- 策略依赖
- 异常依赖

- 报警升级

- 第一报警处理接收人错过报警时，通知更大范围处理人员
- 核心指标报警持续异常时间达到向上通报阈值时，通知更高level人员

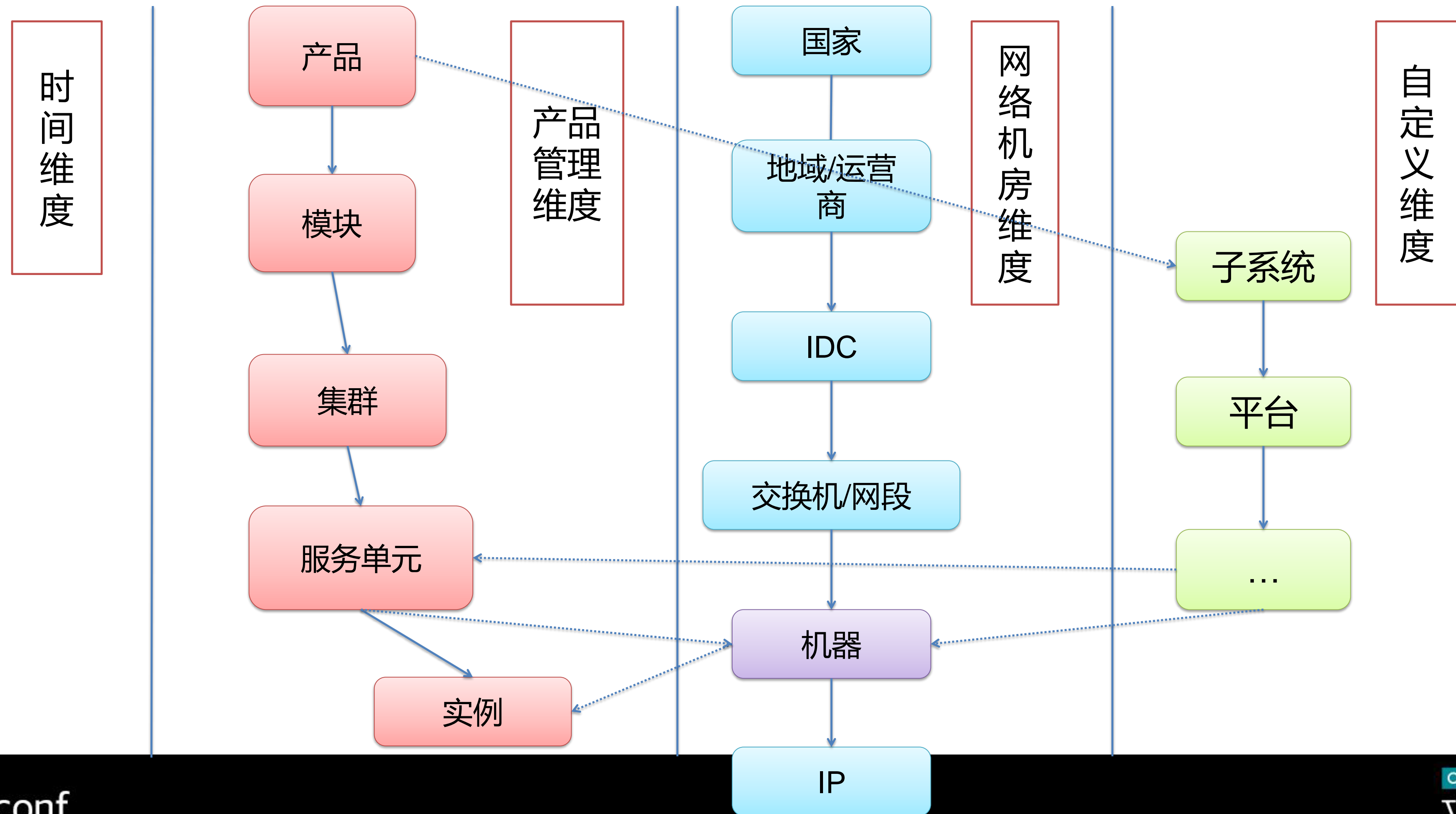
这样就够了吗？



- 分析问题篇
 - 建立关联分析
 - 常用于问题定位，迅速找到相关的指标

产品服务层级的关联关系

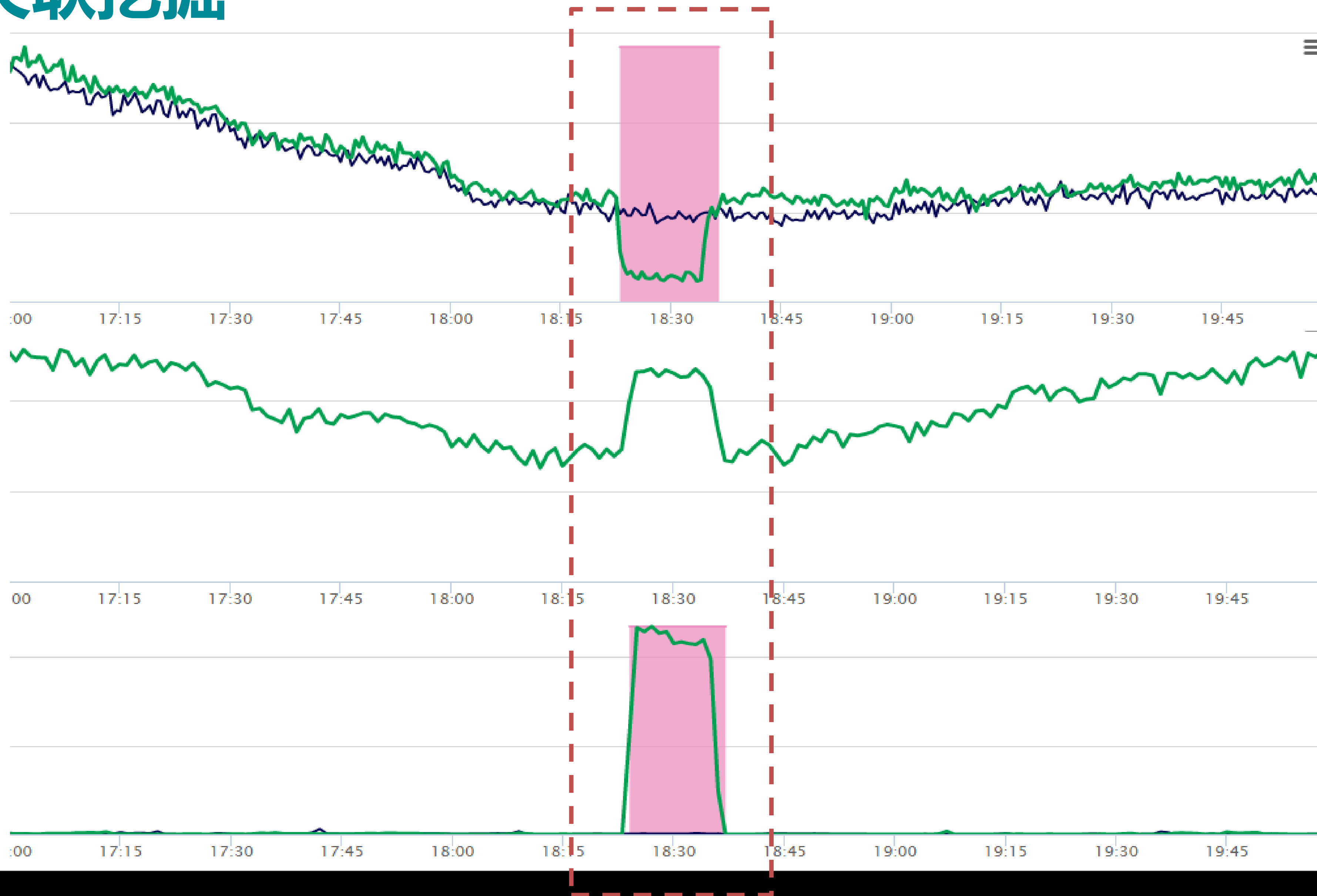
如何为复杂多样的运维数据建立关联？



关联挖掘

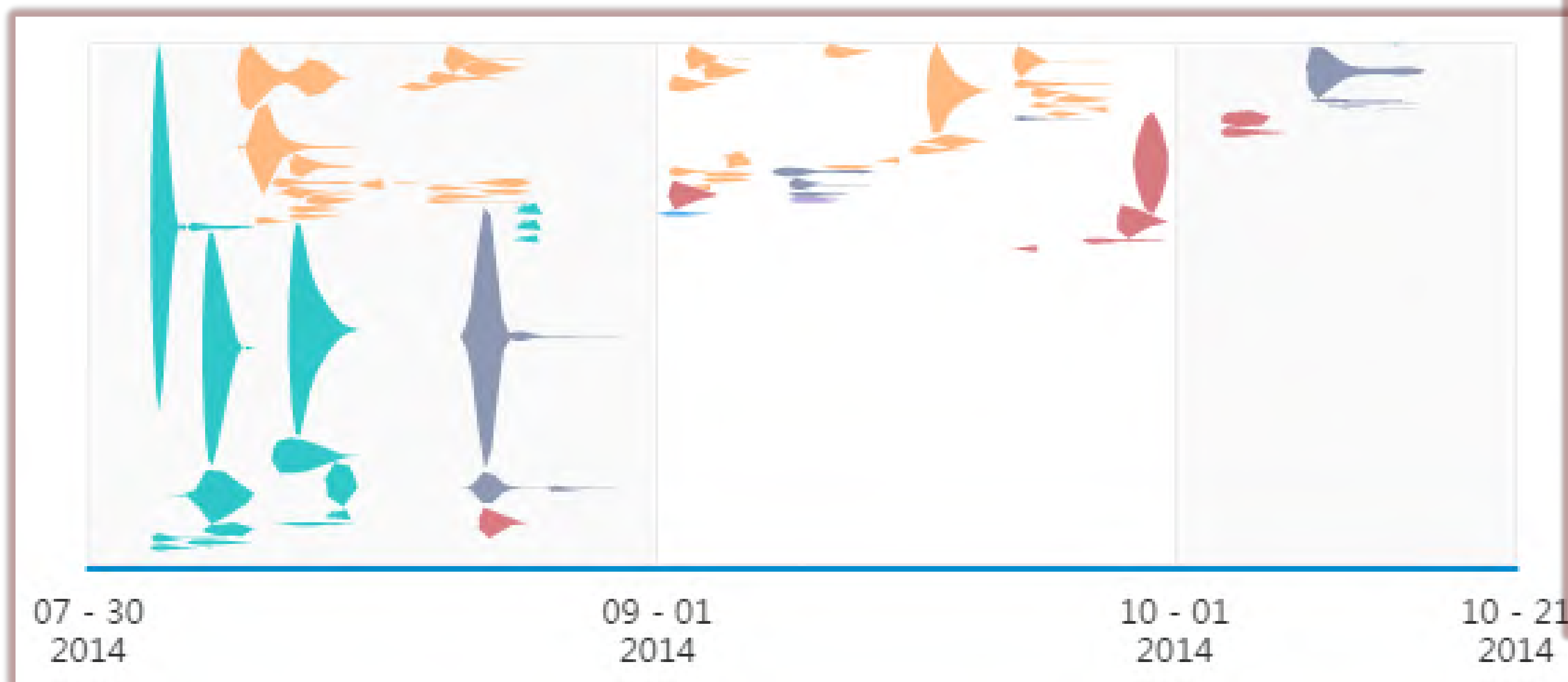
- 事件&事件
 - 频繁项集挖掘
 - 所有运维事件：报警、变更、调度任务...
- 事件&时序
 - 指标异常经常部署变更事件相伴发生
 - 问题诊断&故障定位

多时序关联挖掘



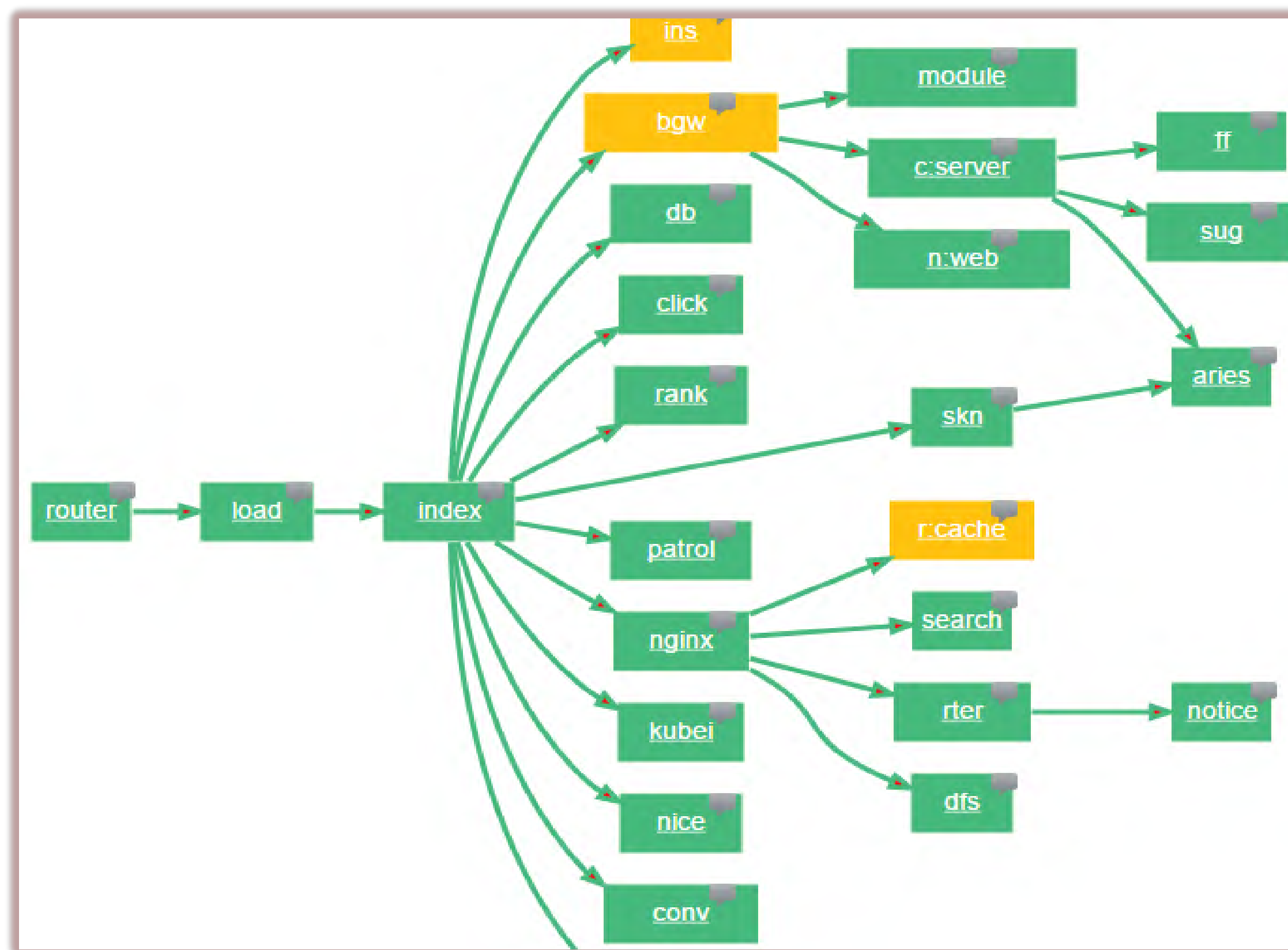
关联可视化

- 事件&事件关联
 - 事件流图
- 事件&时序关联
 - 指标趋势+变更事件展示



服务透视

- 模块调用关系
- 事件和模块关联



故障定位

- 找到了关联还不够
- 分析问题解决问题才是关键

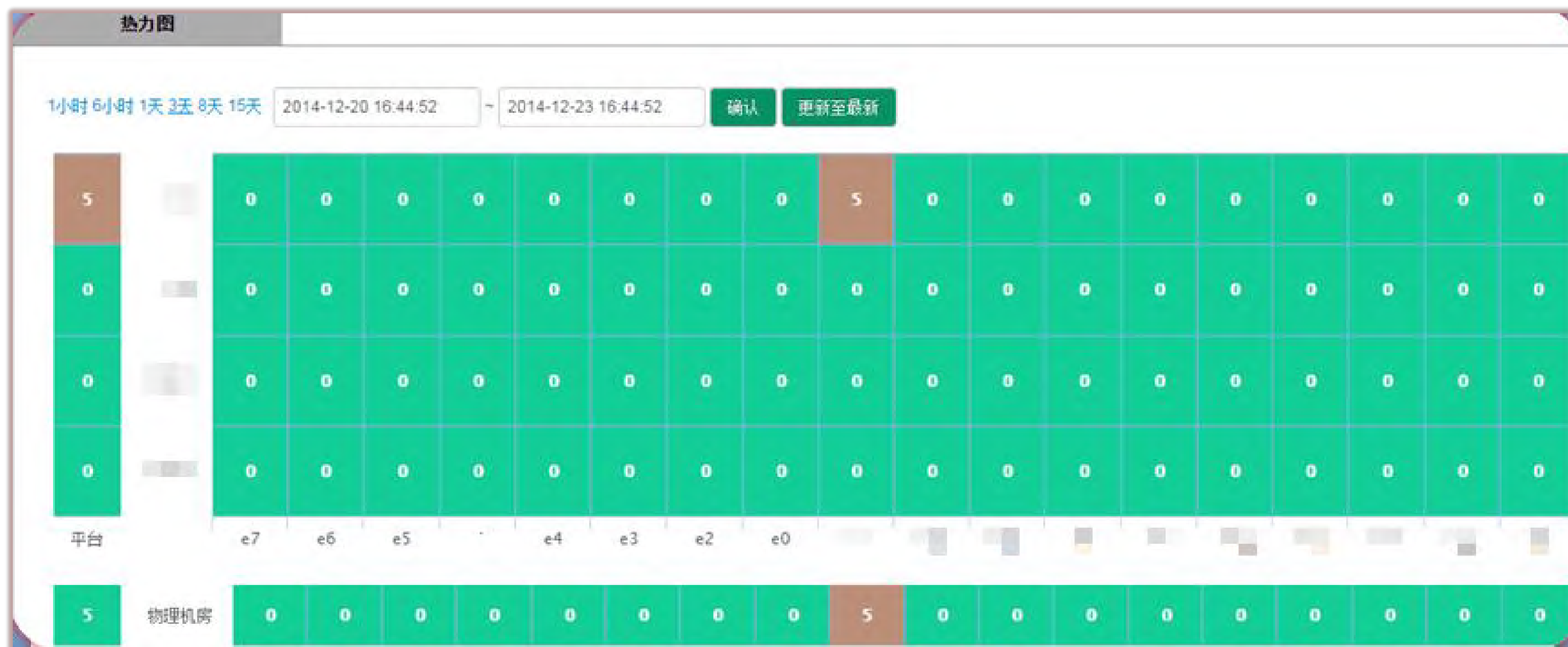
多维数据分析

- 总体维度与细分维度
- 举例: 总体流量vs.分地域流量
- 按细分维度影响权重排序

属性	当天值	占比	天同比	天变化量	周同比	周变化量	操作
[blurred]	[blurred]	0.56%	-11.68%	-465.14	0.97%	33.79	查看
[blurred]	[blurred]	1.55%	-3.87%	-390.12	19.47%	1,578.63	查看
[blurred]	[blurred]	0.28%	-12.24%	-240.55	-2.85%	-50.60	查看
[blurred]	[blurred]	0.50%	-4.39%	-144.71	7.21%	211.84	查看
[blurred]	[blurred]	0.87%	-1.68%	-92.70	-1.61%	-88.87	查看
[blurred]	[blurred]	0.23%	-4.82%	-72.90	30.80%	339.08	查看
[blurred]	[blurred]	0.04%	-20.74%	-69.28	-0.03%	-0.08	查看
[blurred]	[blurred]	0.04%	-22.99%	-67.92	7.53%	15.93	查看
[blurred]	[blurred]	0.01%	-52.96%	-57.22	-33.17%	-25.23	查看
[blurred]	[blurred]	0.79%	-0.83%	-41.00	2.01%	97.13	查看

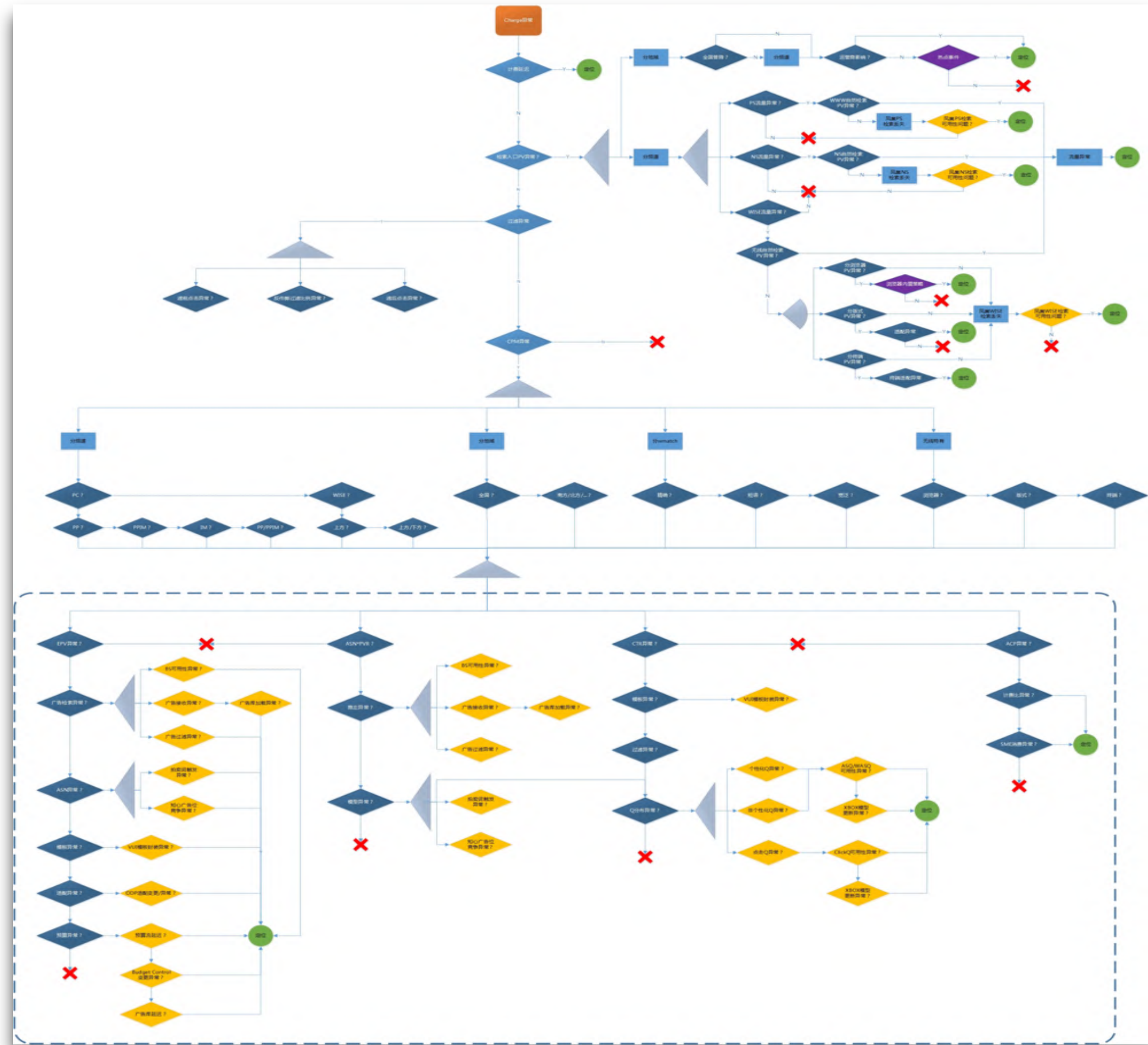
多维数据分析的应用

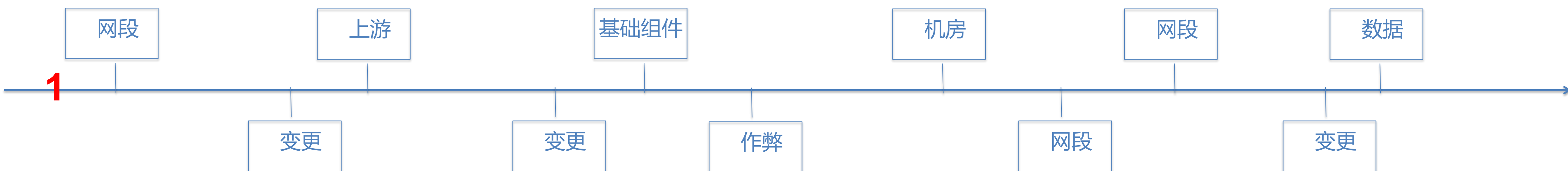
- 交叉维度展示异常情况
- 分地域、浏览器、运营商、数据中心...
- 热力图直接展现异常维度



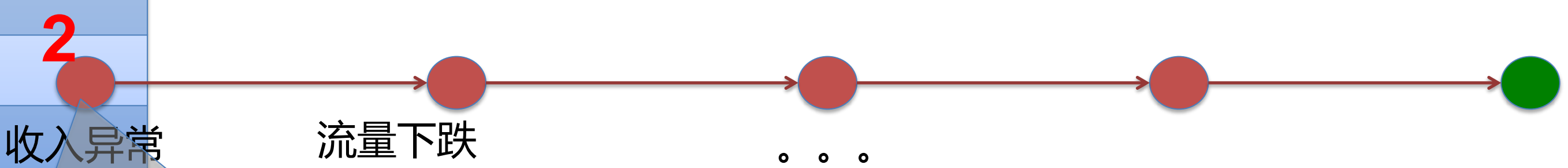
故障诊断树

- 领域专家知识
- 逻辑推导引擎
- 迅速找到问题根因





编号	时间	推导过程
1	2015年5月15日 12:20:09	
2		
3		



3

xx升级
xx报警
xx数据更新

监控“闭环”

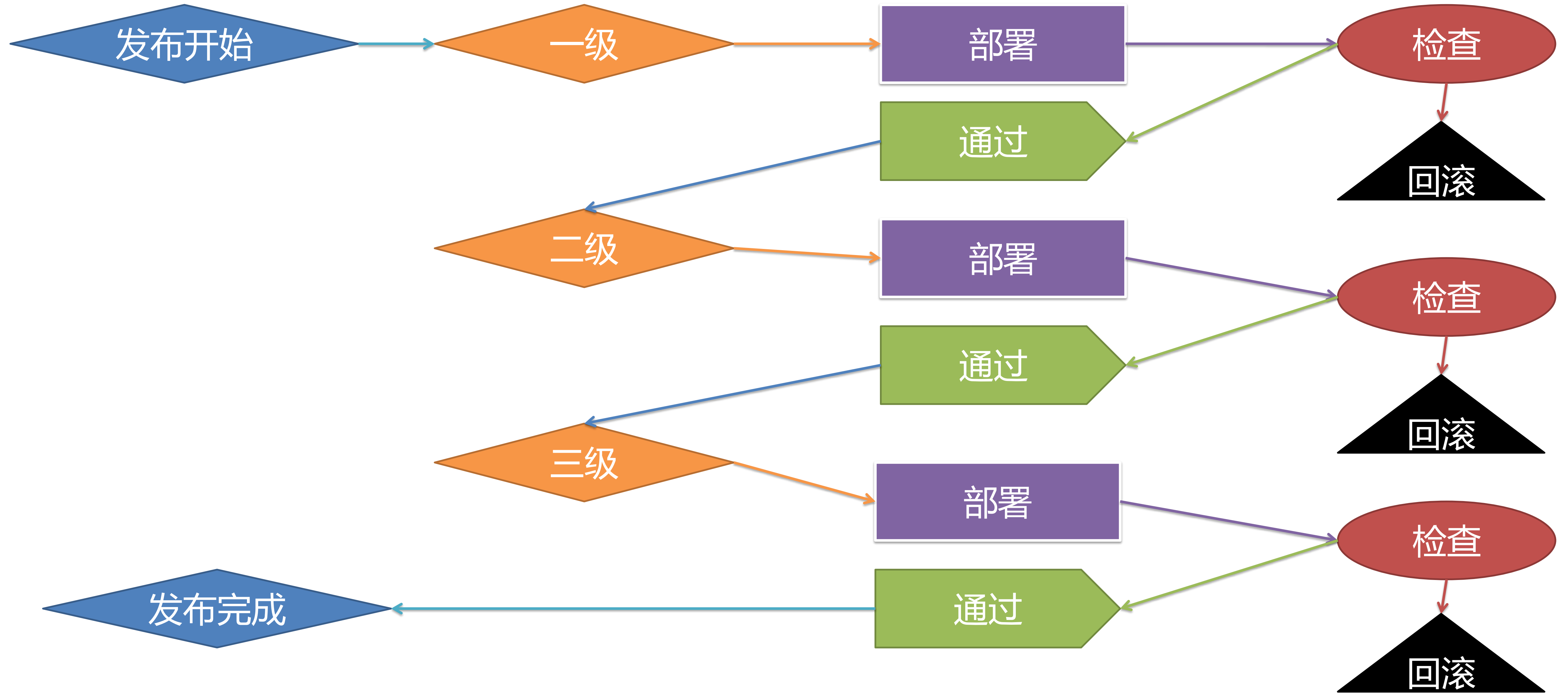
- 让监控系统的环闭合！
 - 问题发现 -> 分析决策 -> 处理 -> 循环往复
- 监控系统与部署调度系统结合
 - 监控系统产生决策
 - 部署调度系统执行



解决问题篇：单边故障自动止损

- 什么是单边故障？
 - 单个IDC故障、单个链路故障等
- 如何止损？
 - 实现自动冗余与调度
 - 智能监控系统负责动态决策
 - 部署调度系统复杂调度执行

灰度发布自动止损



监控闭环：发现 -> 分析决策 -> 解决

异常检测

报警收敛

关联分析

故障定位

自动处理

未来：变被动为主动

■ 全方位覆盖

- 在用户端（APP、浏览器等）、云端（机房、服务器、自身服务、第三方服务等）、管道（链路、运营商）等任何维度进行数据采集并进行异常自动检测

■ 大数据分析&可视化

- 分析运用已有数据，并把服务状态、问题影响分析等可视化

■ 让监控更聪明

- 自动学习并理解故障的趋势和模式
- 自动发现服务或依赖环境的变更

■ 预测故障

- 先于故障发生之前解决

谢谢

