在安全分析中如何规避“大”数据分析

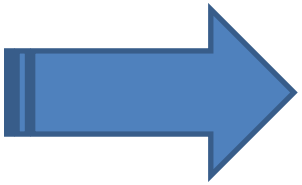
启明星辰 周涛

- **企业安全大数据概述**
  - 安全大数据的成因和特点
  - 安全大数据带来的挑战
- 解决方案1：利用内部威胁情报实现数据浓缩
  - 对威胁情报认知上的几个误区
  - 案例介绍
  - 从企业安全数据中提取威胁情报的几点建议
- 解决方案2：利用异常检测技术减量和降维
  - 异常检测技术的发展及现状
  - 提升异常检测准确率的技术路线
  - 实例分享
- 总结

企业级用户面临安全威胁的演变

历史：常规恶意代码

- 蠕虫
- 病毒
- 木马
- 僵尸网络
- ……



现状：由APT引发的data breach

- 针对特定目标
- 为获取特定政治或经济利益
- 有组织甚至是国家力量支持

应对：改变事件处置响应的被动模式，从更基础的数据中主动发现威胁！

安全大数据的特点

大量

- 大型机构年度汇总的安全数据容量可达PB级

高速

- HP[1]：每天1T条，平均1200万EPS
- EMC[2]：每天1.4 bn条，容量约1TB

多样

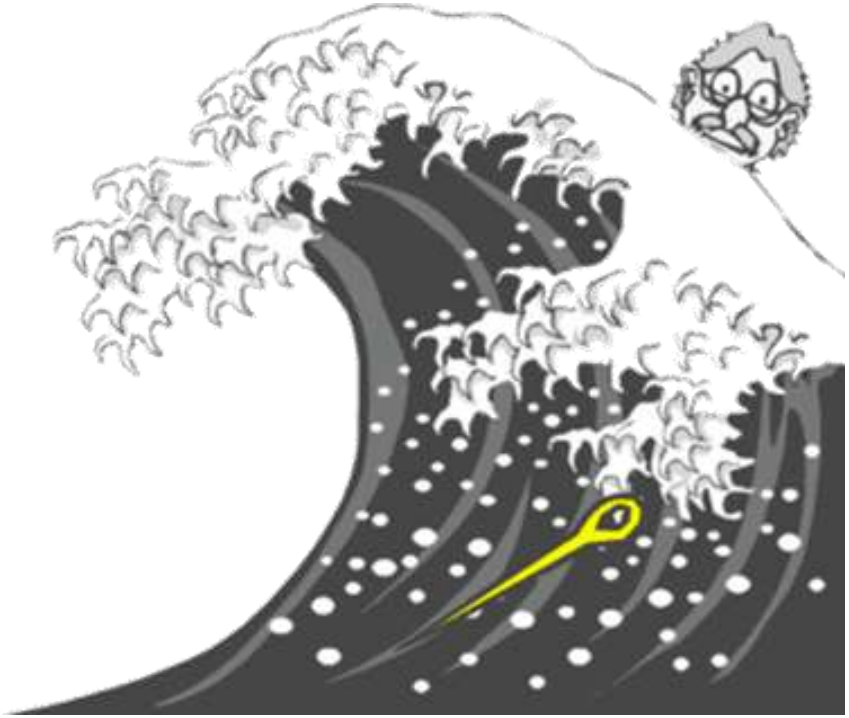
- 事件
- 日志
- 数据流
- 原始报文
- 样本文件
- 威胁情报
- ……

价值密度低

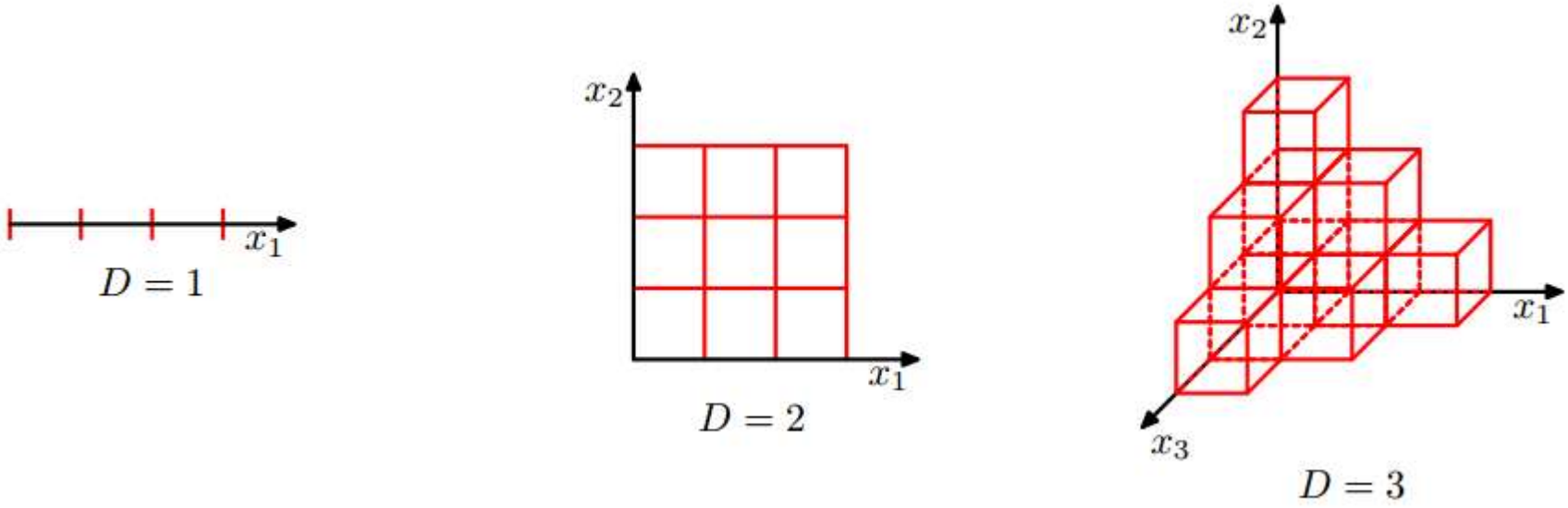
- 与攻击行为相关的数据占比低，有价值的数据淹没在大量背景噪声中

[1] CSA. Big Data Analytics for Security Intelligence. 2013.
[2] Yen T F, et al. Beehive: Large-scale log analysis for detecting suspicious activity in enterprise networks. 2013.

- 大量

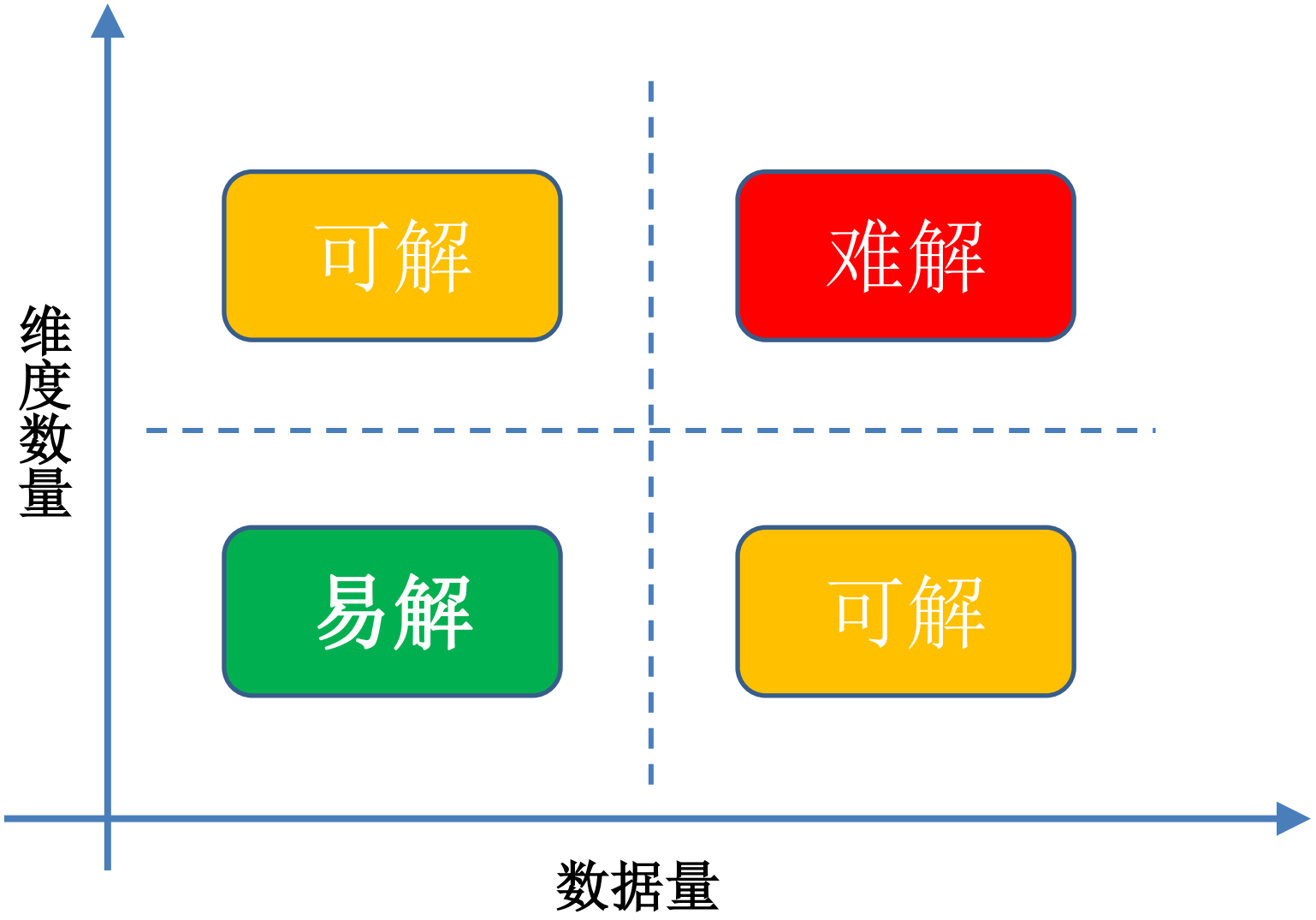


- 高维



图片来源： Bishop C M. Pattern recognition and machine learning. 2006.

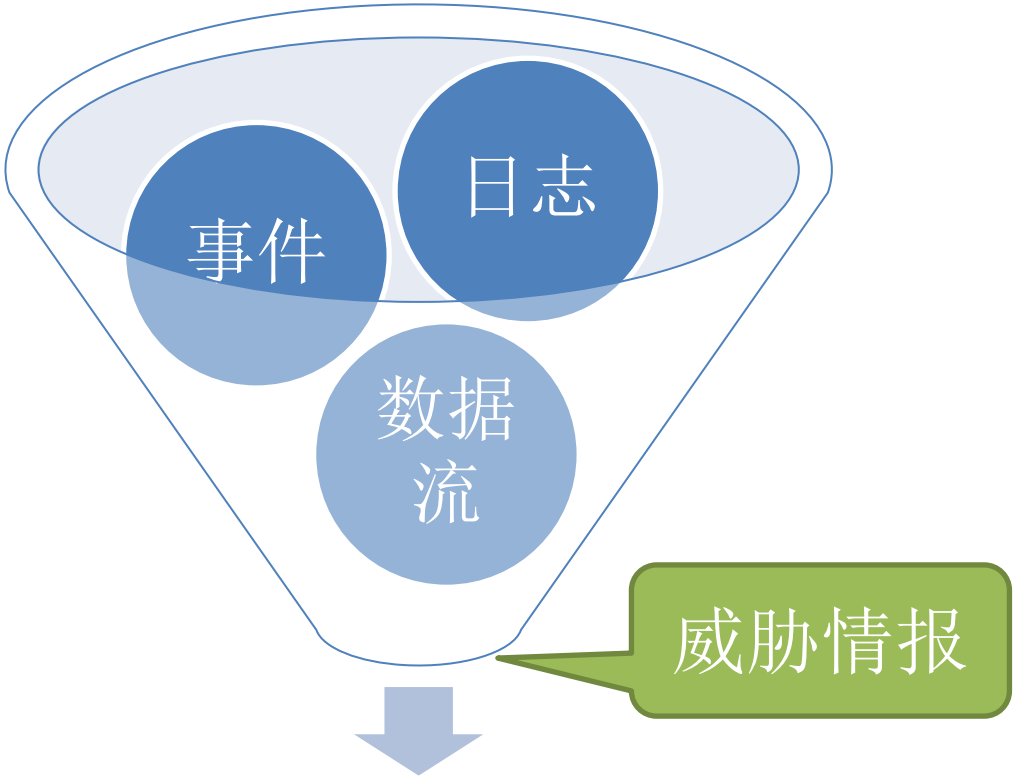
大数据带来的挑战



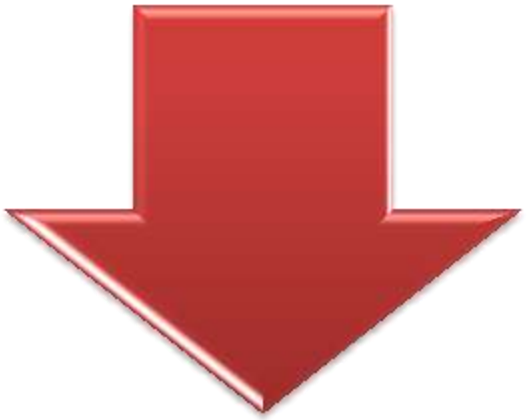
目录

威胁情报的作用

- 从“大数据”中浓缩有价值的“小数据”



关于威胁情报的几个误区



误区1：威胁情报只能依靠互联网公司或产业联盟来提供

真相：威胁情报有外部和内部之分；从自身数据中提取的情报更有价值

误区2：有了威胁情报就能对抗未知威胁

真相：情报是攻击的指征(Indicator)而非特征(Signature)，缺乏完善的数据支撑和基础检测分析能力，情报将难以发挥价值

误区3：威胁情报就是处理原始数据的滤网，对情报的使用方式是单向的

真相：要利用情报过滤数据，再从过滤的结果中发掘新情报，情报和数据间存在双向互动的过程

- 针对Lockheed Martin的APT攻击实例

Phase	Intrusion 1	Intrusion 2	Intrusion 3
Reconnaissance	[Recipient List] Benign PDF	[Recipient List] Benign PDF	[Recipient List] Benign PPT
Weaponization	Trivial encryption algorithm		
	Key 1		Key 2
Delivery	[Email subject] [Email body]	[Email subject] [Email body]	[Email subject] [Email body]
	dn...etto@yahoo.com		ginette.c...@yahoo.com
	60.abc.xyz.215	216.abc.xyz.76	
Exploitation	CVE-2009-0658 [shellcode]		[PPT 0-day] [shellcode]
Installation	C:\...\fssm32.exe C:\...\IEUpd.exe C:\...\IEXPLORE.hlp		
C2	202.abc.xyz.7 [HTTP request]		
Actions on Objectives	N/A	N/A	N/A

案例来源：Hutchins E M, et al. Intelligence-driven computer network defense informed by analysis of adversary campaigns and intrusion kill chains. 2011.

- 首例Redis-DDoS Botnet样本捕获过程

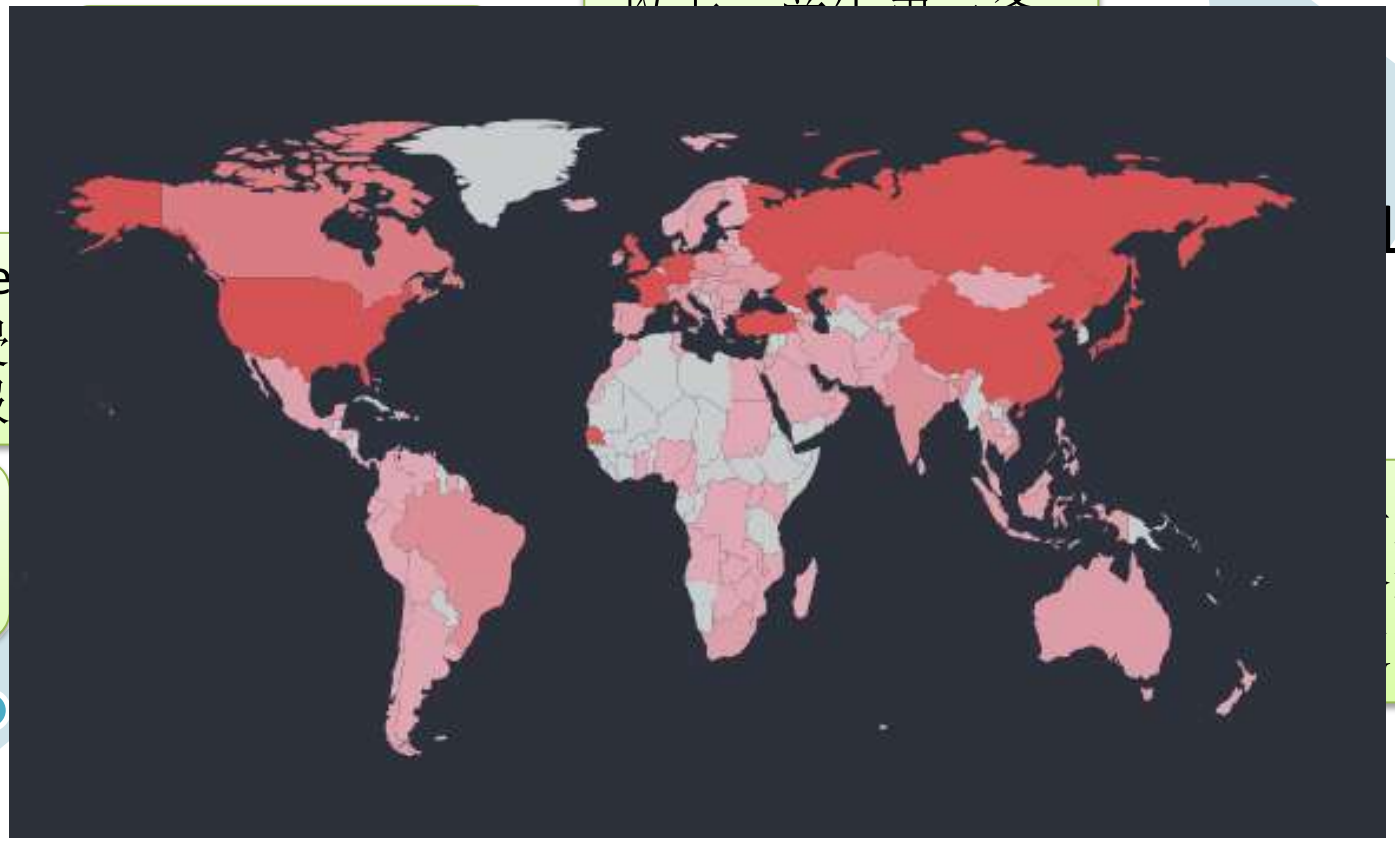
僵尸程序发起DDoS

某用户Re
器被入侵
第一条报

11-16

sebug网
站发布漏
洞通告

家
析、
告



图片来源： https://www.sebug.net/vuldb/ssvid-89715

提供威胁情报的指征定义

- 通过非80端口访问web服务器
- 未通过登录页面的新建HTTP连接
- Web服务器访问了内部非指定终端
- Web服务器主动访问了Internet
- 内部运维操作来自非指定地址
- 对后台数据库的访问未通过JDBC协议或标准服务端口

检测到指征时的应对措施

- 从触发规则起，记录该连接5分钟内的全流量数据

- 改变了攻防双方不对称性，使天平向防御方倾斜

攻击方：技战术优势

- 全面设防 VS 单点突破

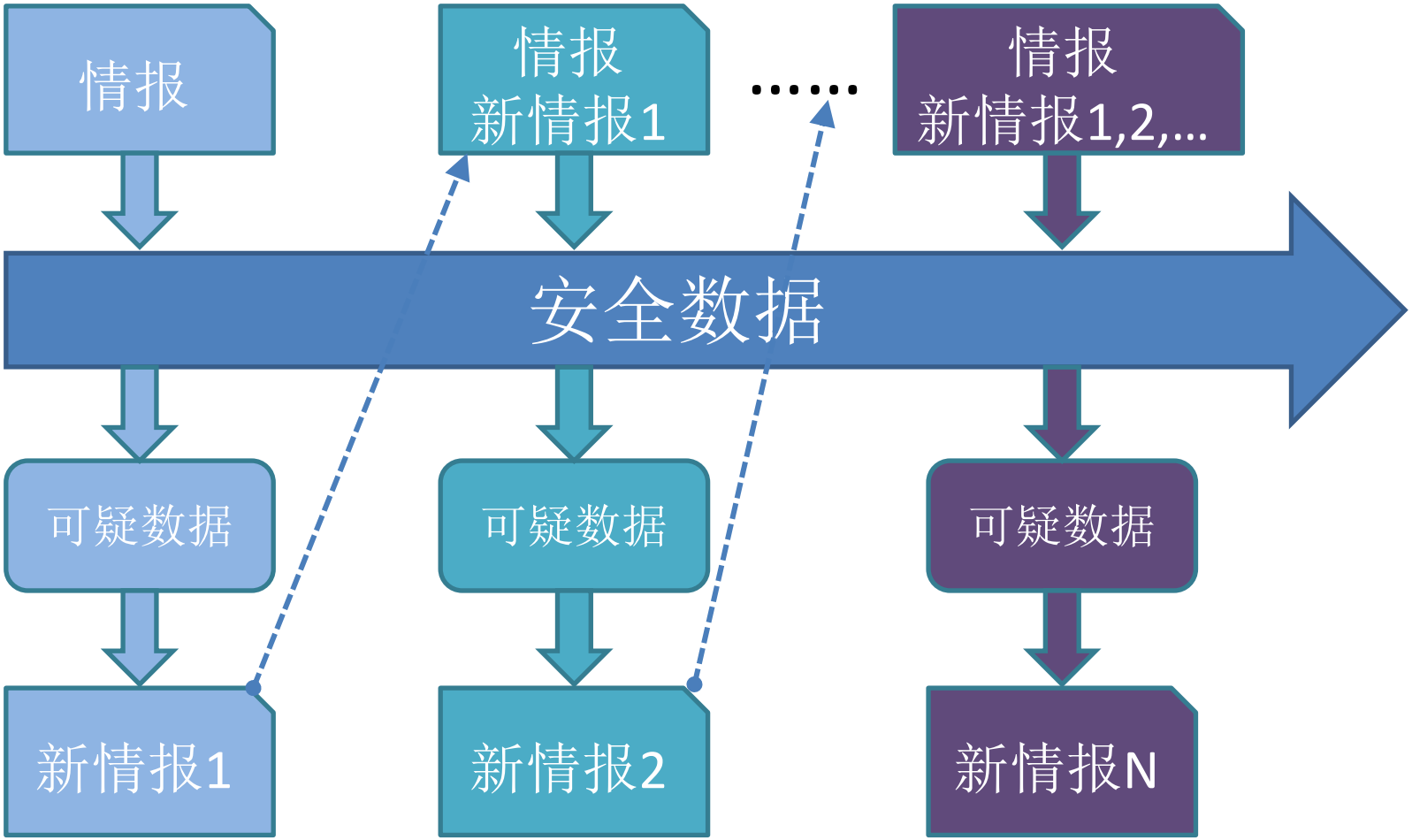
防御方：数据优势

- 利用情报提前布防，取得时间优势
- 增加攻击方绕过情报的成本

- 维度纵深：攻击场景还原
  - 参考攻击链等模型，获取完整的入侵行为数据描述

关于提取内部威胁情报的几点建议

- 时间纵深：对已有情报的持续跟踪和丰富机制



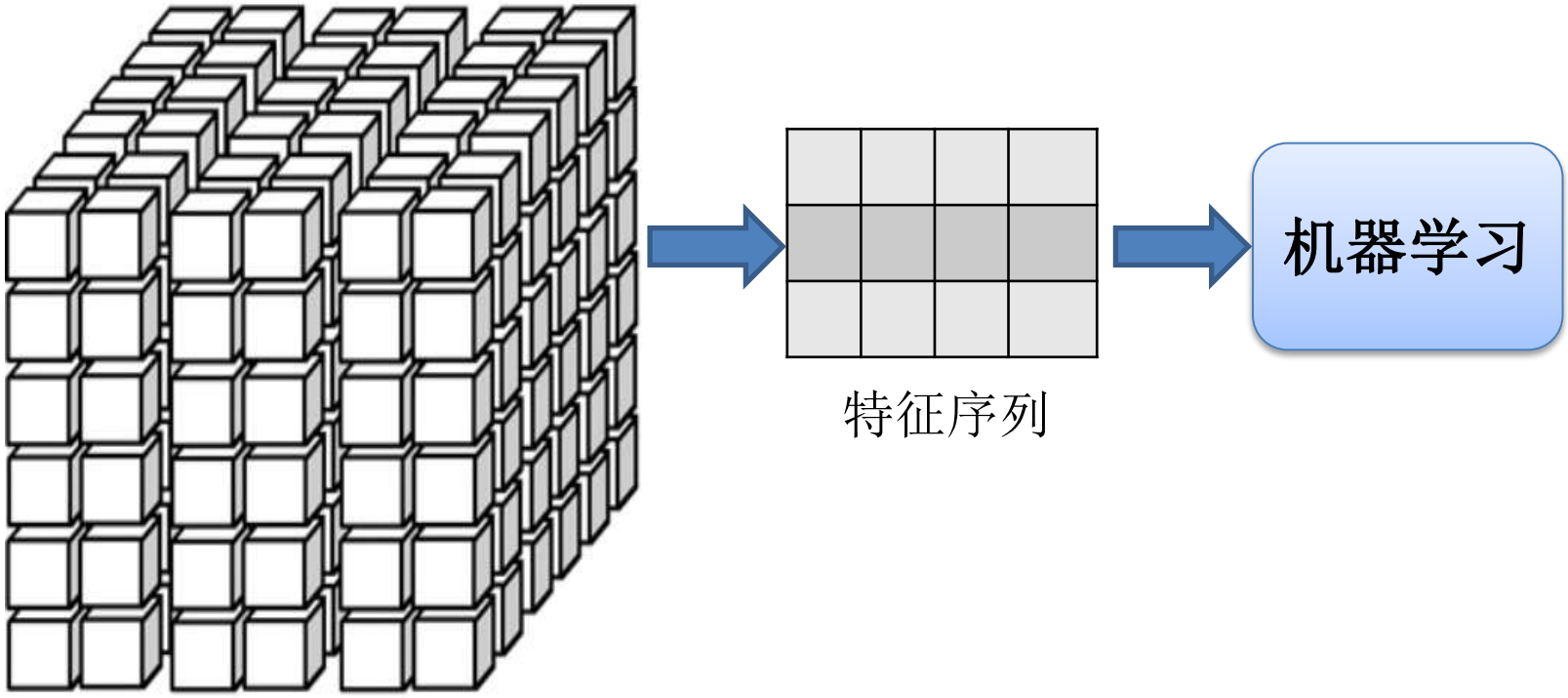
- 从各类异常行为中挖掘威胁情报
  - 关键：如何定义异常，可对正反两类行为分别描述

误用行为模型	对典型攻击过程及结果的描述
	如果匹配则存在异常
正常行为模型	对正常业务行为的描述
	如果偏离则存在异常

目录

异常检测的作用

- 同时降低大数据的量级和维度
- 具备未知威胁检测能力



回顾异常检测技术

起源

- 约1990s

特点

- 具备未知威胁检测能力
- 高漏报和高误报，未能成为商业化产品的主流技术

原因

- 受限于计算和存储能力
  - 模型粒度不够精细
  - 特征维度较低
  - 模型训练不够充分

异常检测技术的回归

内在需求

- 对高级威胁需要未知威胁检测能力

外部条件

- 以云计算、大数据为代表的ICT技术发展成熟
  - 更细的模型粒度
  - 更高维的特征选择
  - 更充分的模型训练

提升异常检测准确率的技术路线

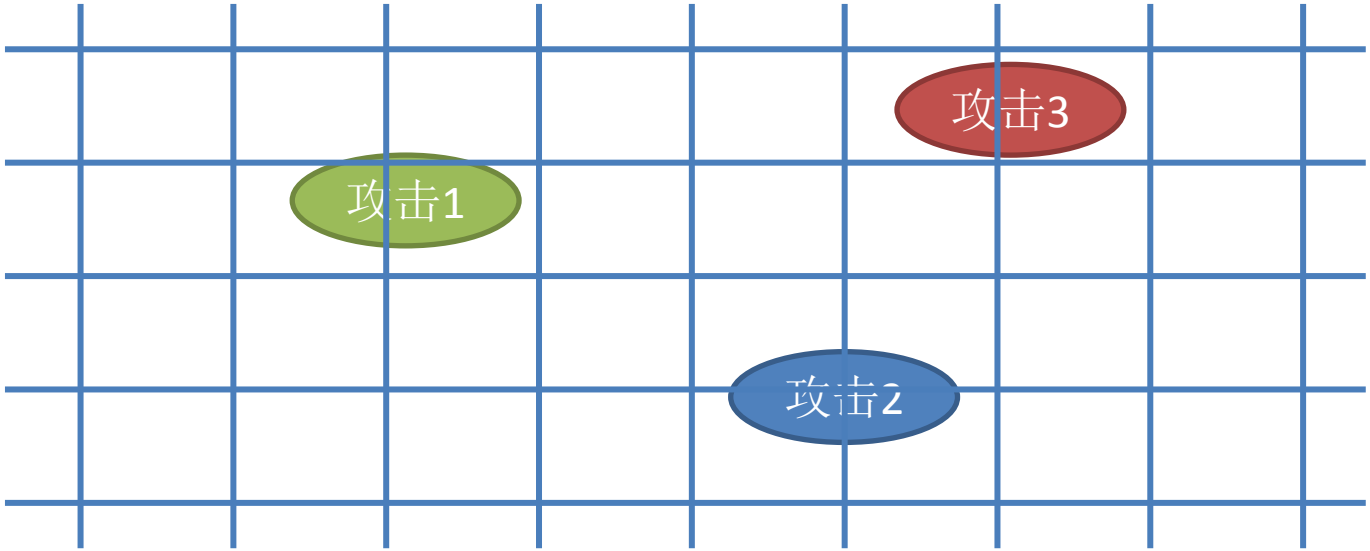
- 建立细粒度模型
  - 举例：主机画像、主机应用画像

优点
- 对异常的反映灵敏，可降低漏报率

缺点
- 可能会提升误报率

改进
- 利用积累的大数据做更长期的建模
- 自身纵向比较和域内横向比较相结合

- 高维特征(feature)提取
  - 根据行为概括足够丰富的特征参数，使得对于任何可描述的攻击行为，总能体现在一组特征参数的异常上



提升异常检测准确率的技术路线

- 提升(Boosting)算法
  - 参考机器学习领域的成功经验

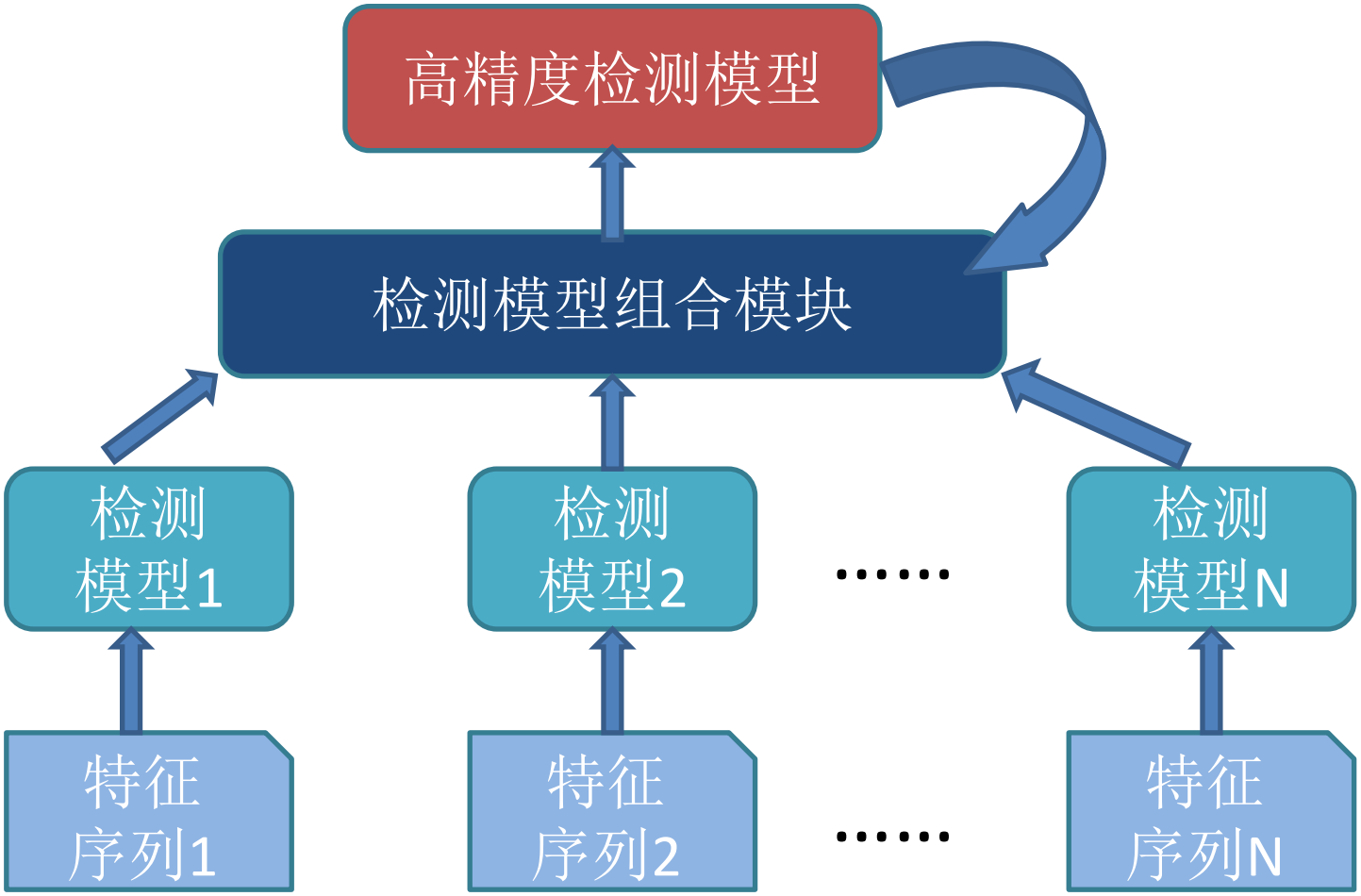
单个分类器的准确度有限

提升单一分类器的精度困难

将若干个简单的弱分类器组合，仍然有可能构造出强分类器

提升异常检测准确率的技术路线

- 基于提升算法的异常检测模型

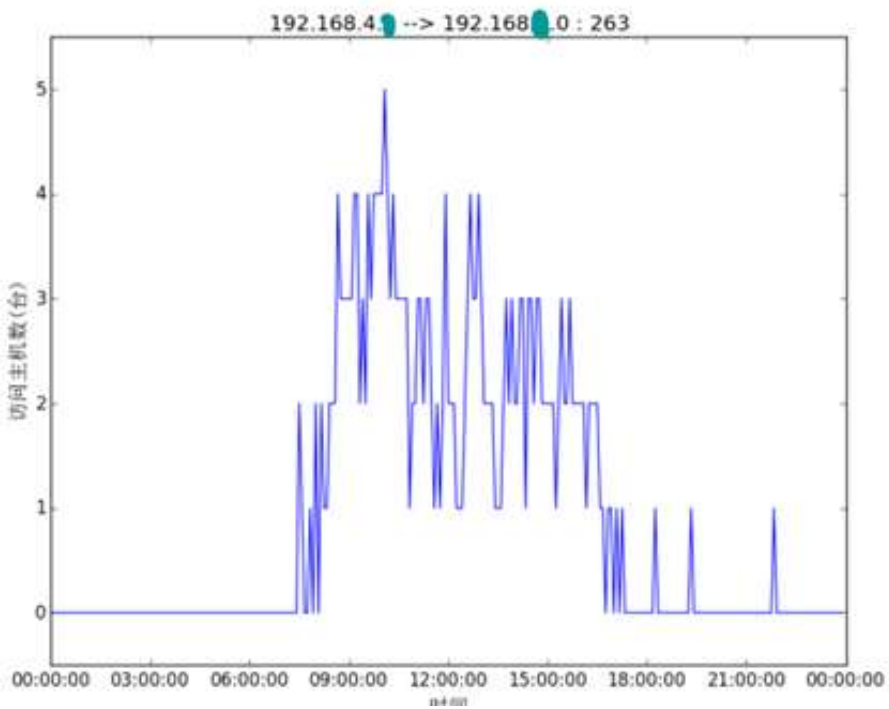


实例：异常访问流量检测

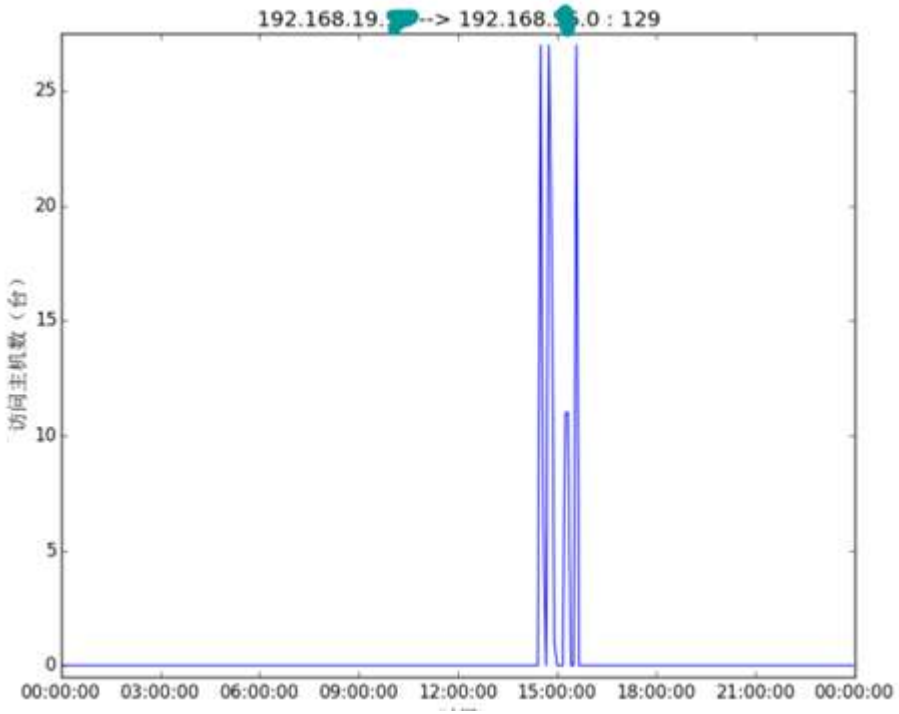
会话信息类指标	一台主机单位时间内不同协议类型的会话统计信息
应用分布类指标	一台主机单位时间内不同应用类型的访问统计信息
指示位标识类指标	一台主机单位时间内收发的含特定协议标识位的数据包数量及其比值
地址分布指标	一台主机单位时间内访问的IP地址网段分布、内外网分布等参数

实例：异常访问流量检测

- 慢速扫描试探行为检测



某正常主机行为序列



某攻击主机行为序列

目录

应对高级威胁：大数据安全分析

大数据带来的挑战：大量和高维

内部威胁情报：实现数据的浓缩

异常检测：多维分析和提升机制

谢谢！

zhoutao@venustech.com.cn