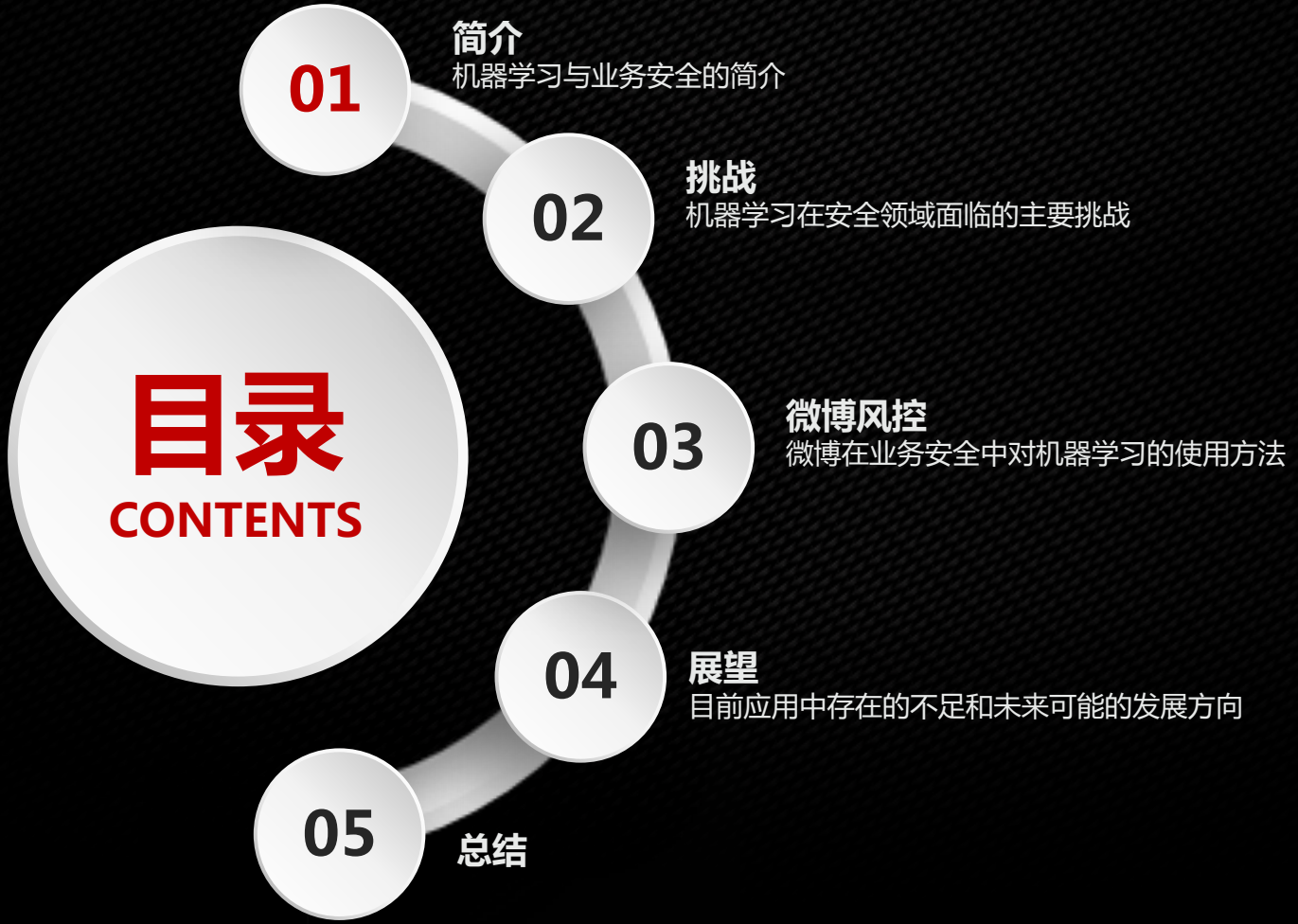
机器学习在微博业务安全中的定位

微博安全：何为舟

日期：2017.10.18


主办方 Geekbang 极客邦科技 InfoQ
ArchSummit
全球架构师峰会 2017
12月8-9日 北京·国际会议中心


AiCon
全球人工智能技术大会2018
助力人工智能落地
2018.1.13 - 1.14 北京国际会议中心
AiCon
扫描关注大会官网


目录
CONTENTS
01
简介
机器学习与业务安全的简介
02
挑战
机器学习在安全领域面临的主要挑战
03
微博风控
微博在业务安全中对机器学习的使用方法
04
展望
目前应用中存在的不足和未来可能的发展方向
05
总结

简介
Introduction

业务安全


信息泄漏
爬虫抓站
用户诈骗
虚假注册
批量关注
恶意营销
热门话题
信用卡套现
账号盗用
涉黄涉政
金融欺诈
活动作弊
垃圾内容
撞库扫号
黄牛刷单
羊毛党

业务安全

机器

OR

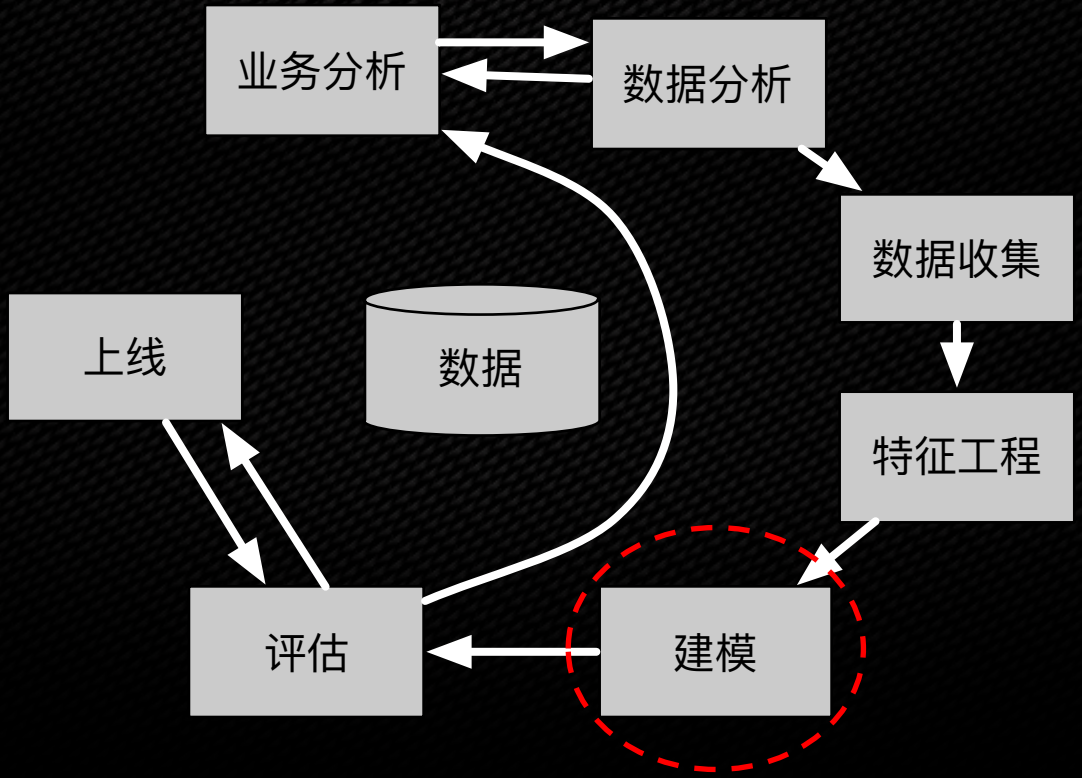
用户

恶意用户

OR

正常用户

业务安全



02

Part One

挑战

Challenge

机器学习

图像处理

- 内容识别、特征描述
- 人脸识别

自然语言处理

- 翻译（文本、音频）
- 语义分析

其他

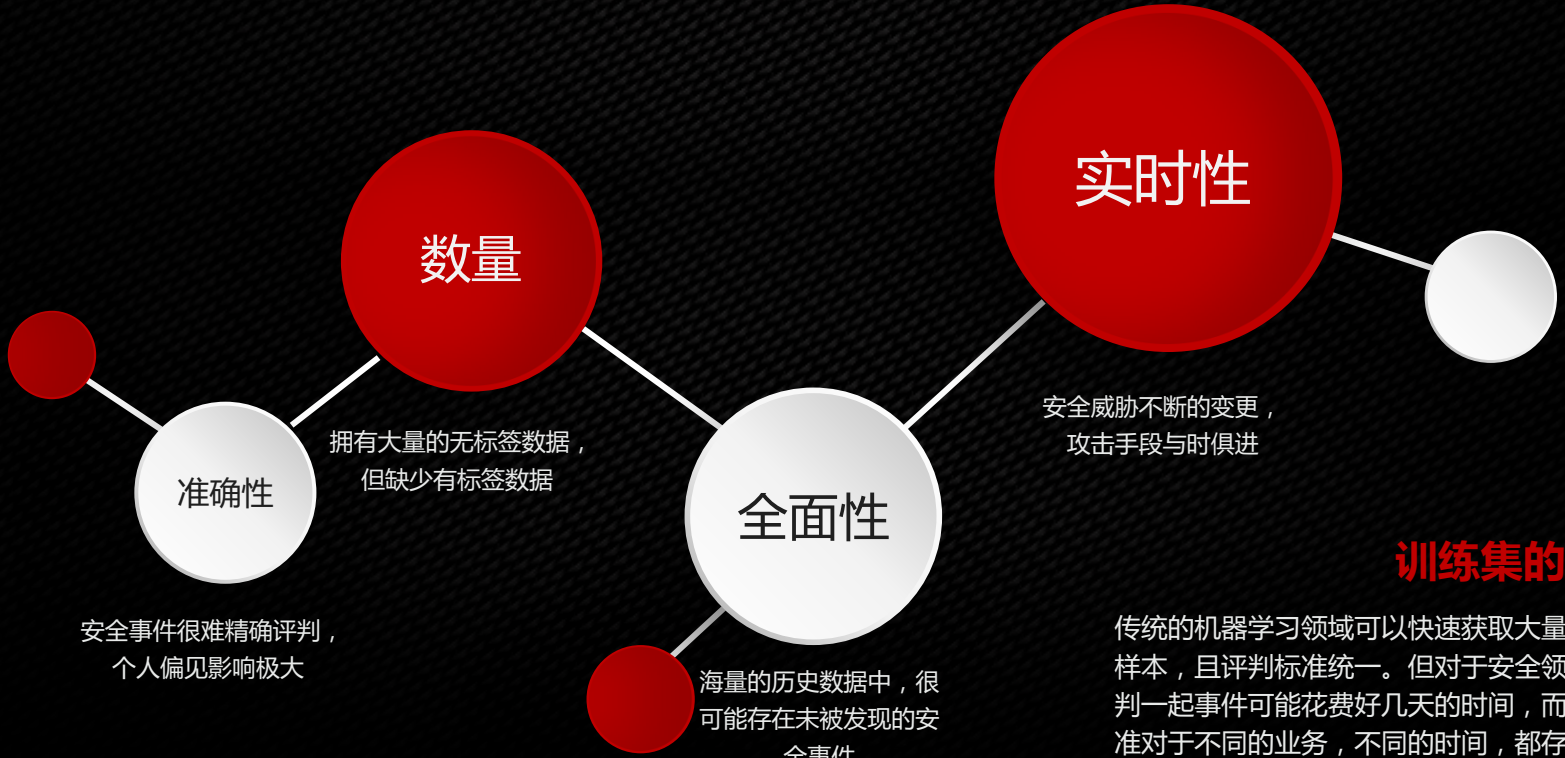
- 推荐
- 搜索

挑战

Challenge

训练集的缺陷

训练集的缺陷



训练集的难题

传统的机器学习领域可以快速获取大量的训练样本，且评判标准统一。但对于安全领域，评判一起事件可能花费好几天的时间，而评判标准对于不同的业务，不同的时间，都存在差异

训练集的缺陷

DARPA Intrusion Detection

- 引用次数 2000+
- 首次公布 1998
- 最后更新 2000

KDD 1999

- 引用次数 926
- 最后更新1999

Enron email dataset

- 引用次数 183
- 首次公布 2004
- 最后更新 2015

German Credit Data

- 最后更新 1994

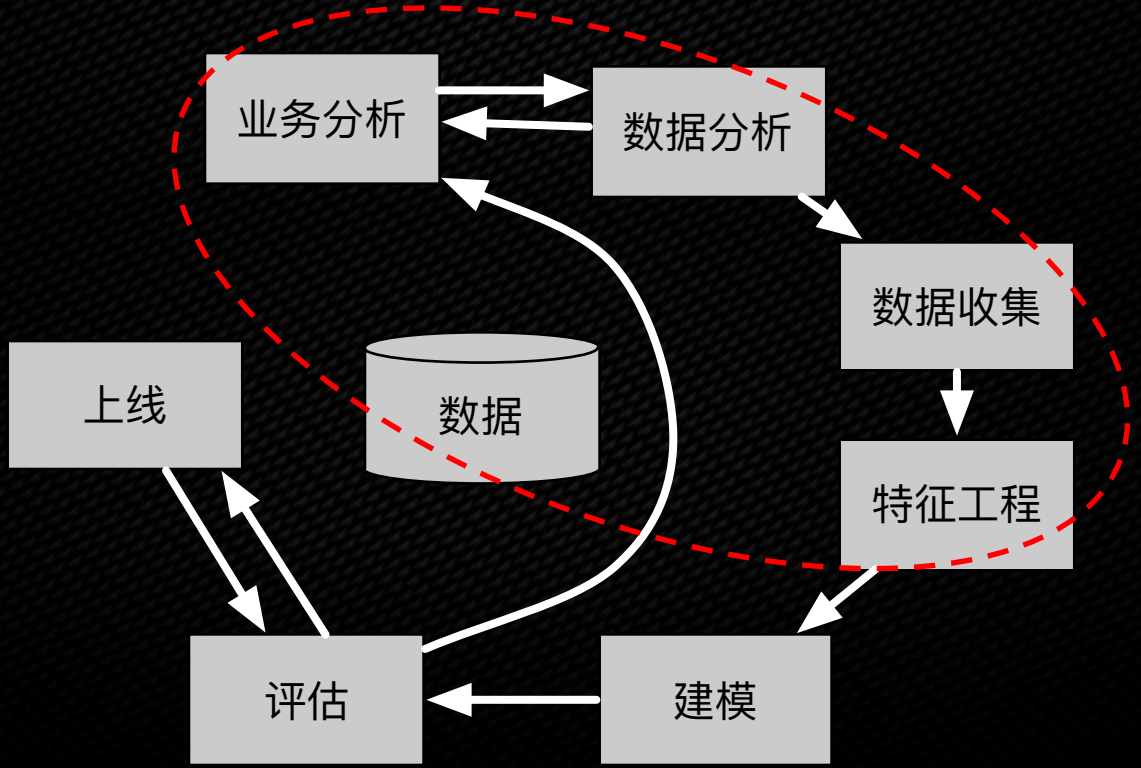
挑战

Challenge

训练集的缺陷

对人工的依赖

对人工的依赖



挑战

Challenge

训练集的缺陷

对人工的依赖

对抗攻击的能力

对抗攻击的能力

Original image
Output Label: **Teapot**

Original image
Output Label: **Property**

Original image
Output Label: **Airplane**

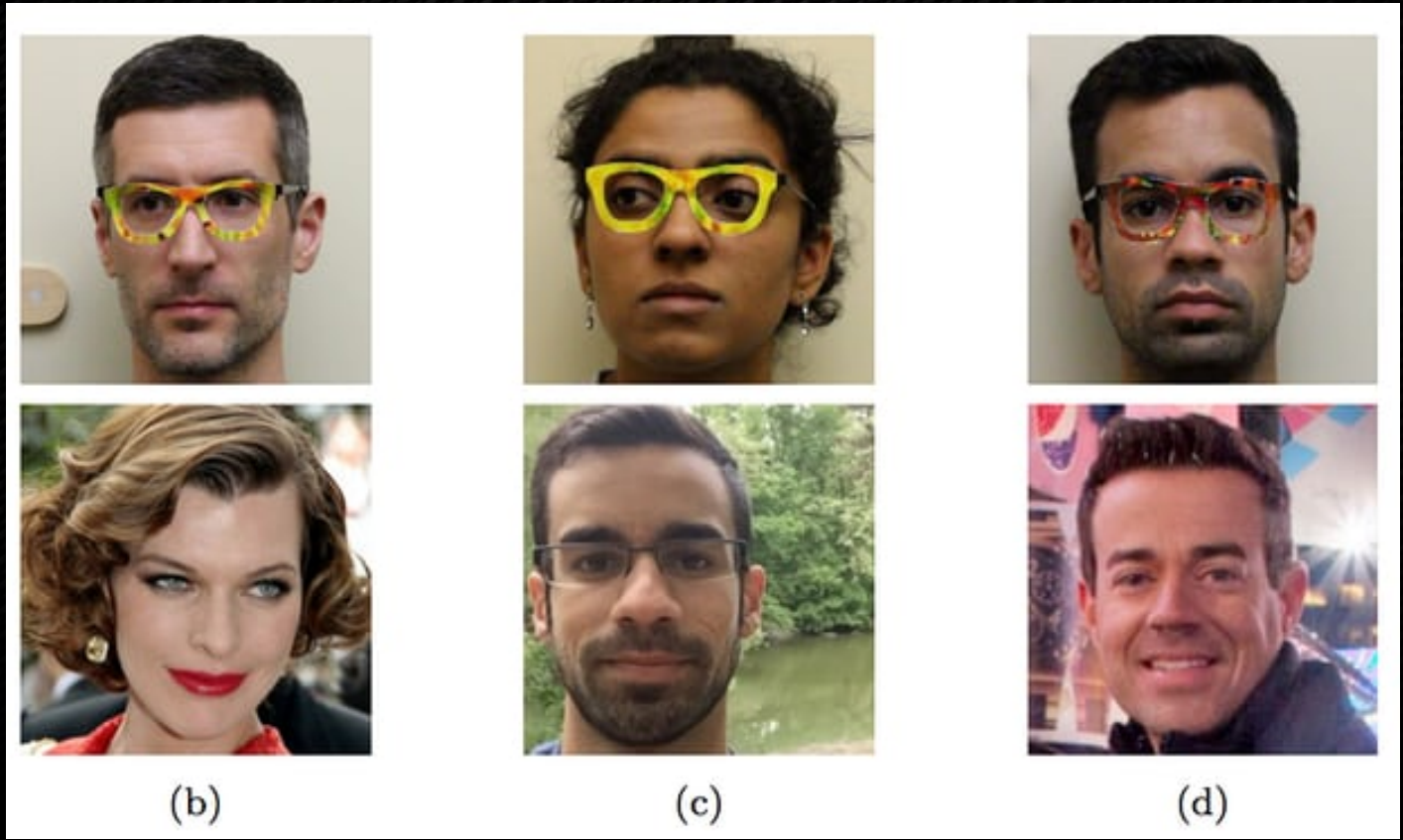
Noisy image (10% impulse noise)
Output Label: **Biology**

Noisy image (15% impulse noise)
Output Label: **Ecosystem**

Noisy image (20% impulse noise)
Output Label: **Bird**

Hosseini H, Xiao B, Poovendran R. Google's Cloud Vision API Is Not Robust To Noise[J]. arXiv preprint arXiv:1704.05051, 2017.

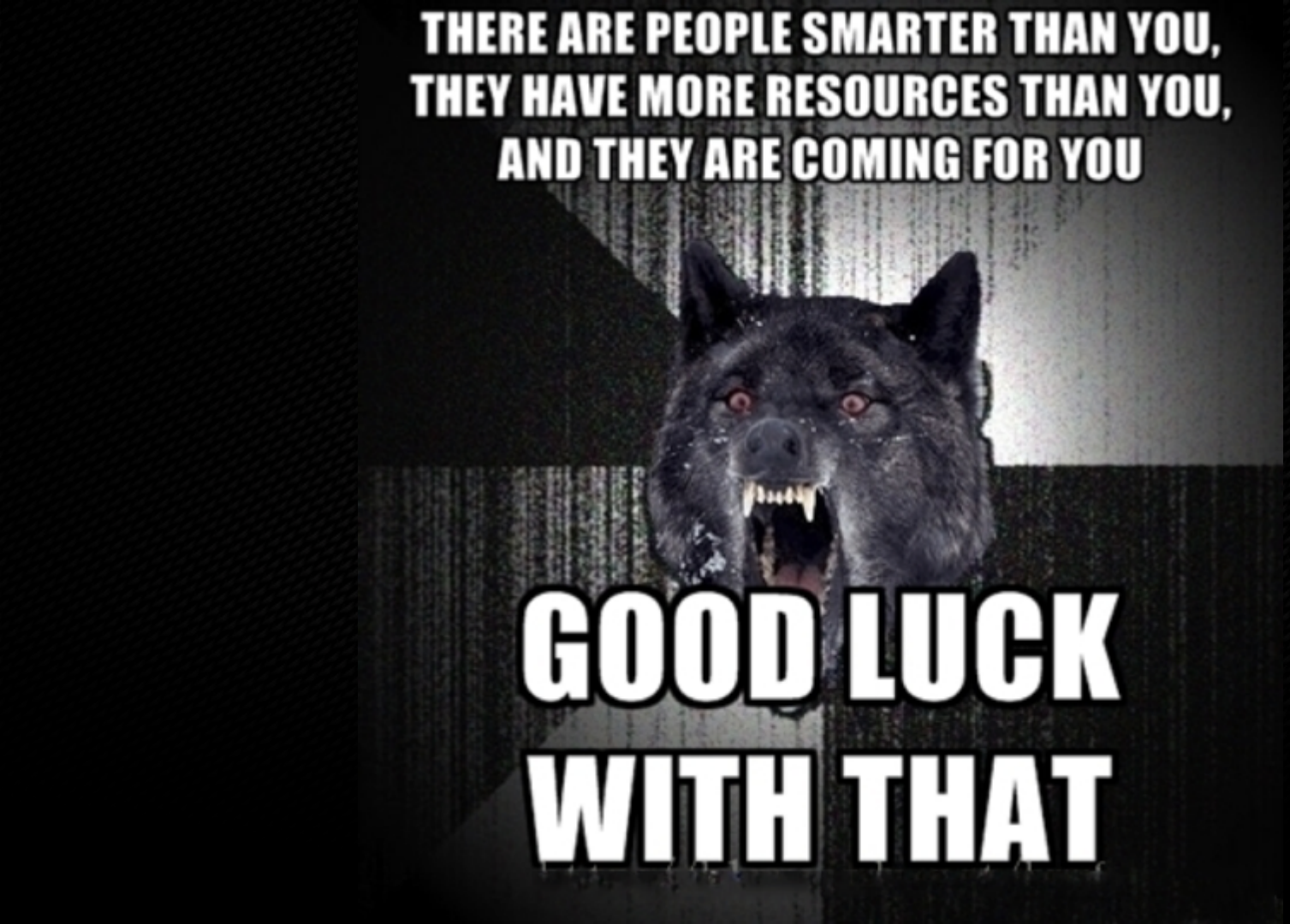
对抗攻击的能力



Sharif, M., Bhagavatula, S., Bauer, L., & Reiter, M. Accessorize to a Crime: Real and Stealthy Attacks on State-Of-The-Art Face Recognition. In Proc. CCS, 2016.

对抗攻击的能力

https://www.rbth.com/science_and_tech/2017/07/22/hiding-from-artificial-intelligence-in-the-age-of-total-surveillance_808692


THERE ARE PEOPLE SMARTER THAN YOU,
THEY HAVE MORE RESOURCES THAN YOU,
AND THEY ARE COMING FOR YOU
GOOD LUCK
WITH THAT

微博风控体系
Weibo Risk Control

需求分析

训练集

- 在少量训练集情况，能够取得较好的效果
- 能够简单的获得准确、实时的训练集

可控性

- 保证结果的绝对准确，尽可能做到零误伤
- 对每一个评判结果，需要给出合理的解释
- 当出现误伤时，能够立即纠正模型

更新

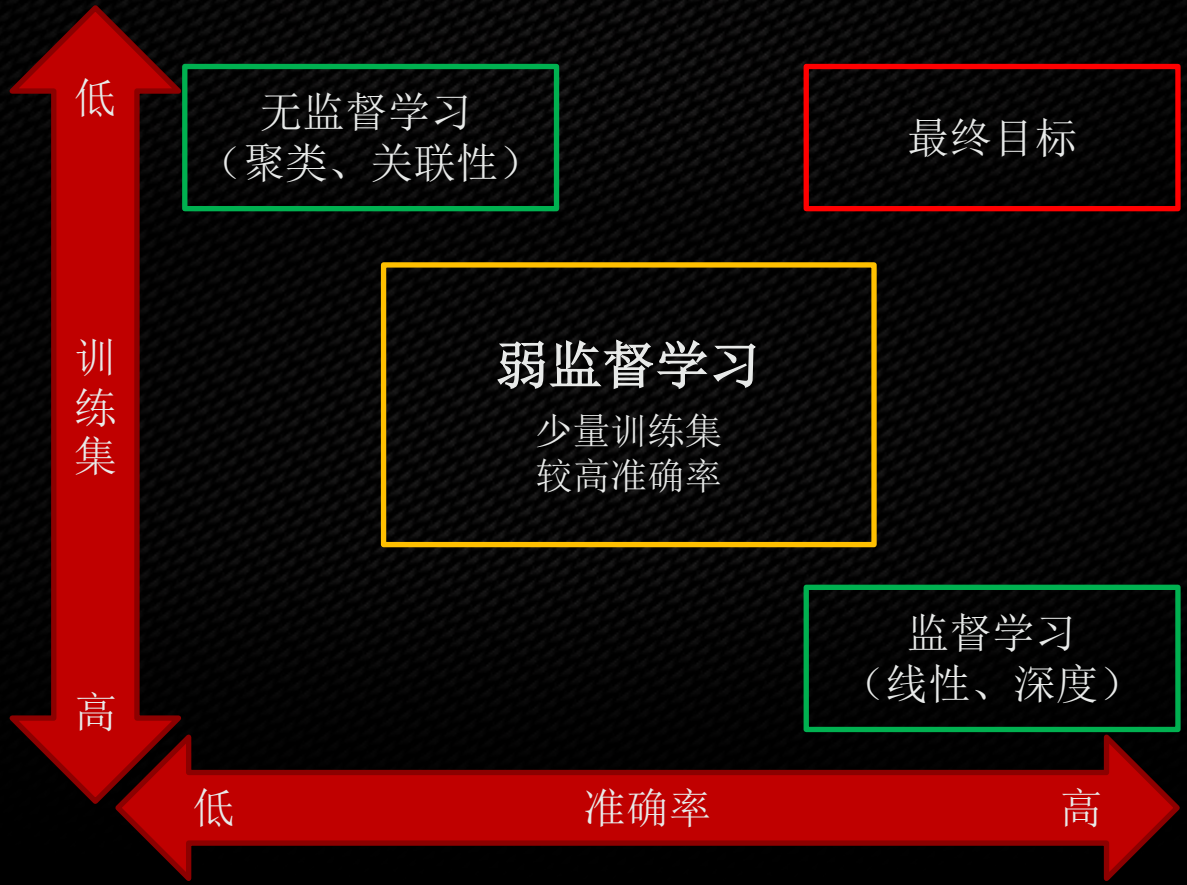
- 模型必须随着训练集的变化，实时更新

训练集

训练集难题

- 扩大来源
  - 客服对接
  - 用户反馈
- 提升质量
  - 弱监督学习

弱监督学习



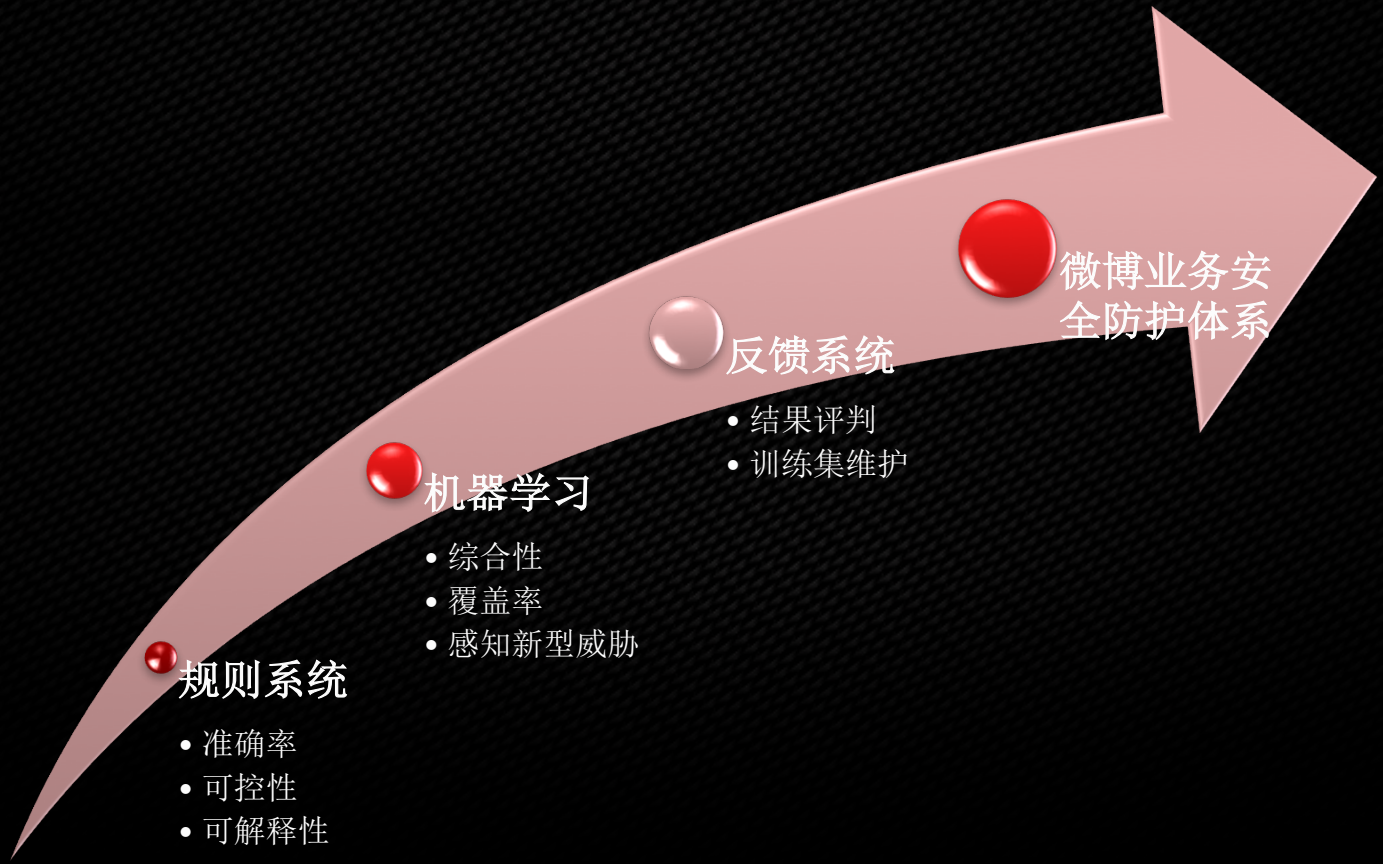
弱监督学习

- UU(Unlabeled-Unlabeled)分类
  - 二分聚类算法
- PU(Positive-Unlabeled) / NU(Negative-Unlabeled)分类
  - 利用某一类样本估计总体损失函数
- PUNU分类
  - 对PU和NU损失函数的综合
- PNU分类
  - 自训练模型
  - 聚类+分类(S3VMs)

训练集

可控性

整体架构


规则系统
准确率
可控性
可解释性
机器学习
综合性
覆盖率
感知新型威胁
反馈系统
结果评判
训练集维护
微博业务安全防护体系

训练集

可控性

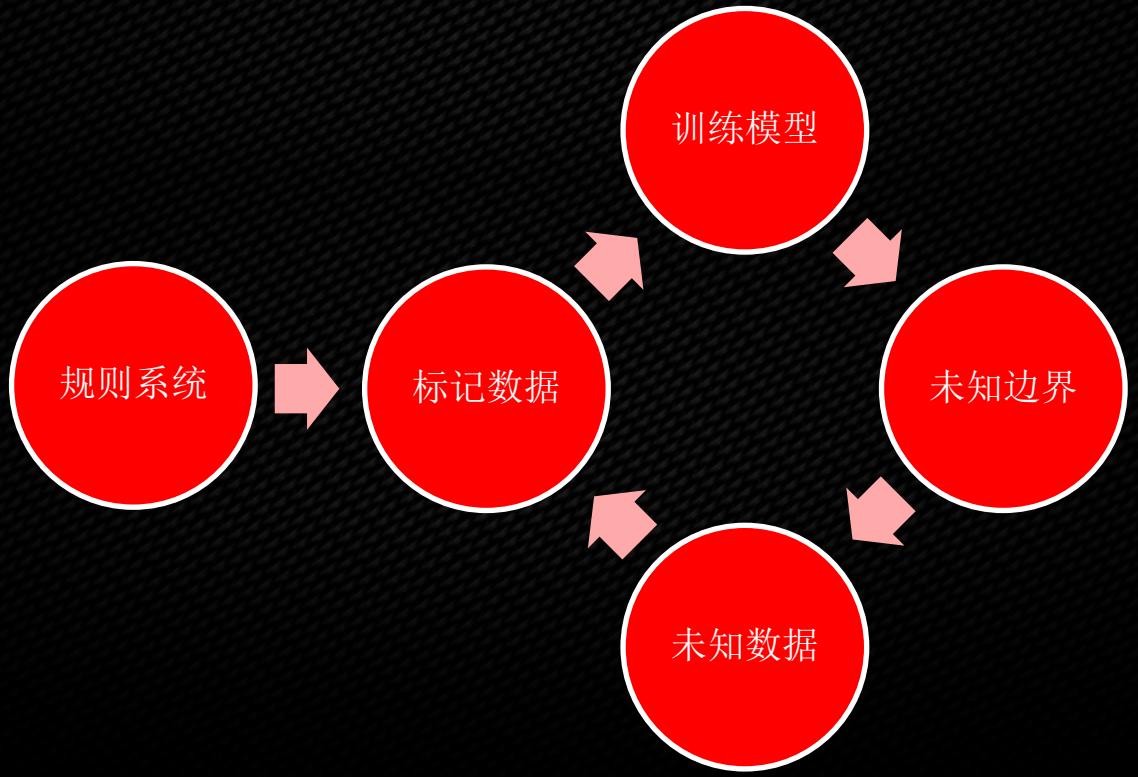
更新

规则更新

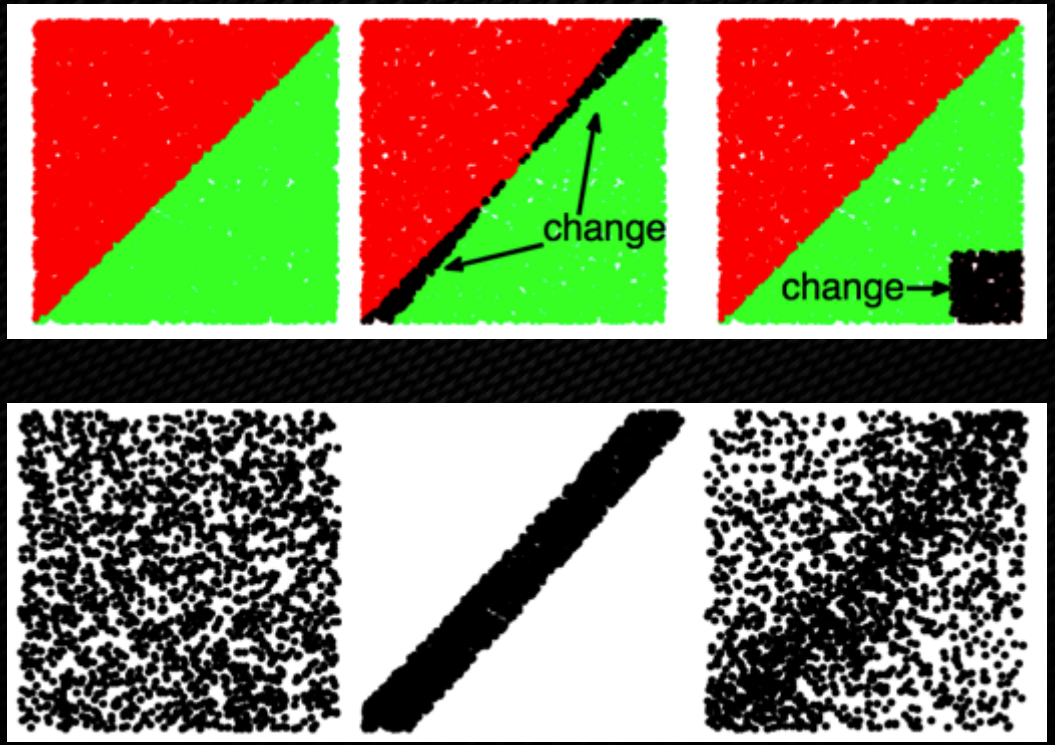
特征工程 → 归一化 → FP-growth → 频繁集判定 → 规则优化

	账号1	账号2	账号3	账号4	账号5	账号6	账号7	账号8
9月1日	A段IP注册	A段IP注册	A段IP注册	B段IP注册	B段IP注册	B段IP注册	B段IP注册	C段IP注册
9月2日	发微博	发微博	发微博	发微博	发微博	-	-	-
9月3日	-	-	-	-	-	发评论	发评论	发评论
9月4日	修改头像	-	修改头像	-	-	-	-	修改头像
9月5日	-	D段IP关注用户X	D段IP关注用户X	D段IP关注用户X	-	-	-	-
9月6日	-	-	-	-	-	E段IP关注用户Y	E段IP关注用户Y	

模型更新

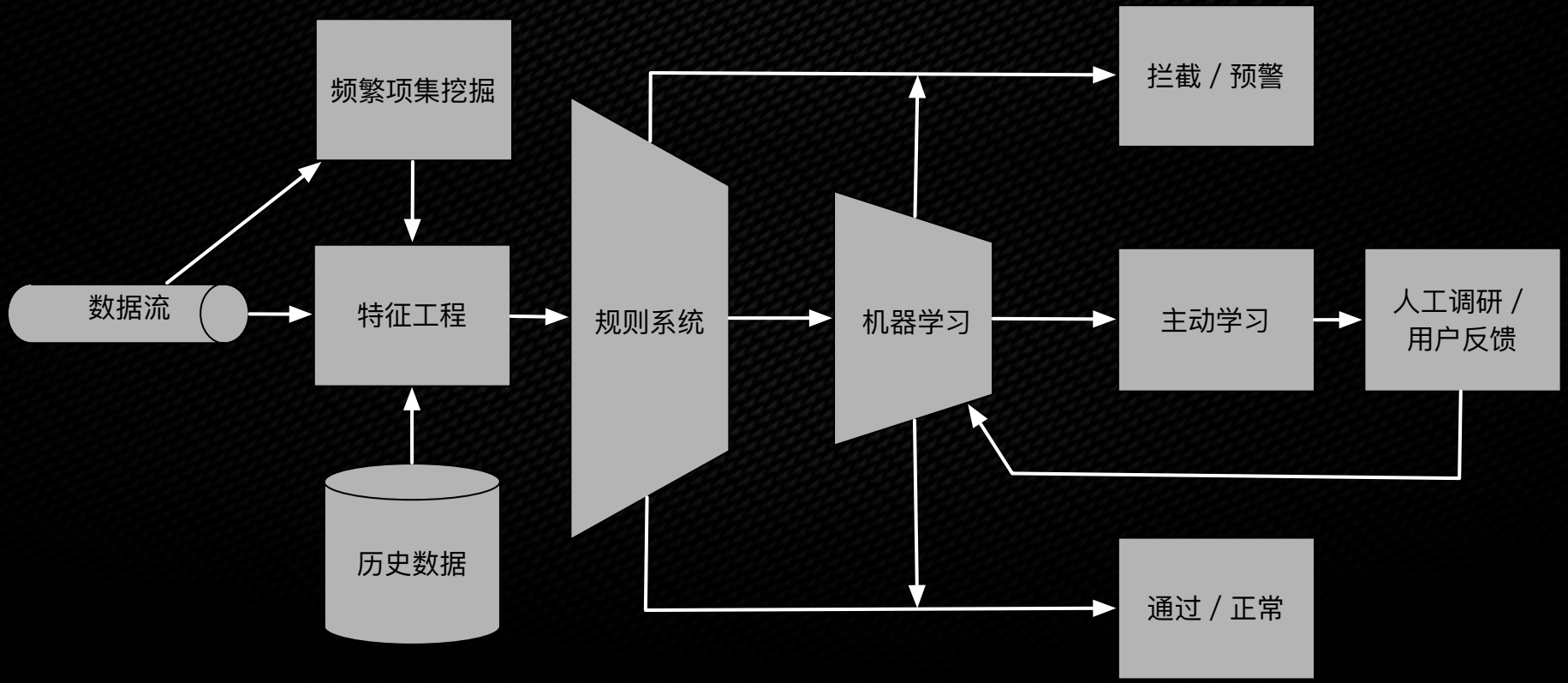


主动学习

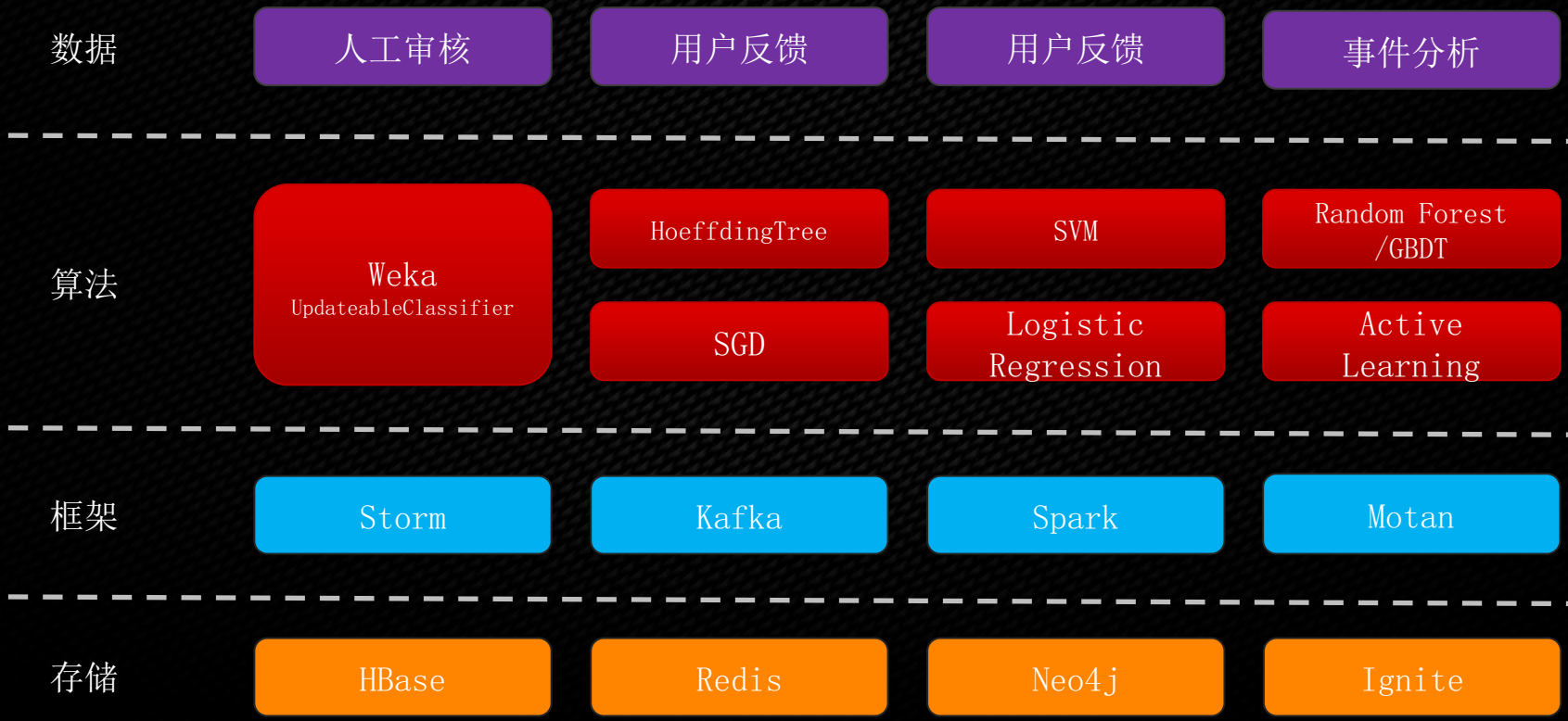


Zliobaite I, Bifet A, Pfahringer B, et al. Active learning with drifting streaming data[J]. IEEE transactions on neural networks and learning systems, 2014, 25(1): 27-39.

整体架构



整体架构


数据
人工审核
用户反馈
用户反馈
事件分析
算法
Weka
UpdateableClassifier
HoeffdingTree
SVM
Random Forest
/GBDT
SGD
Logistic
Regression
Active
Learning
框架
Storm
Kafka
Spark
Motan
存储
HBase
Redis
Neo4j
Ignite

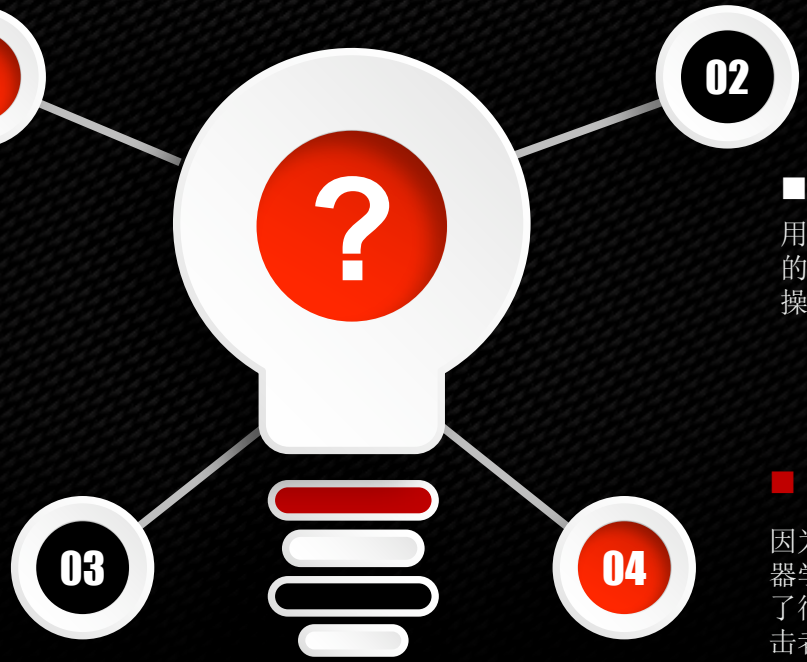
04

Part One

展望

Future

存在的不足



依赖特征工程

只能在特征工程的基础上去进行分析，一旦某种威胁无法通过现有的特征来发现，则可以规避检测

用户反馈准确度

用户反馈同样无法保障百分之百的准确度，可能因为记忆、后台操作等原因给出错误的反馈

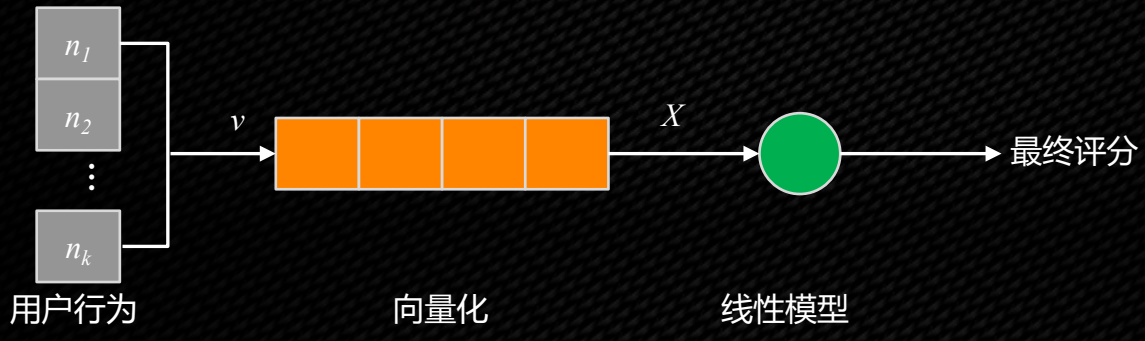
缺乏评估方式

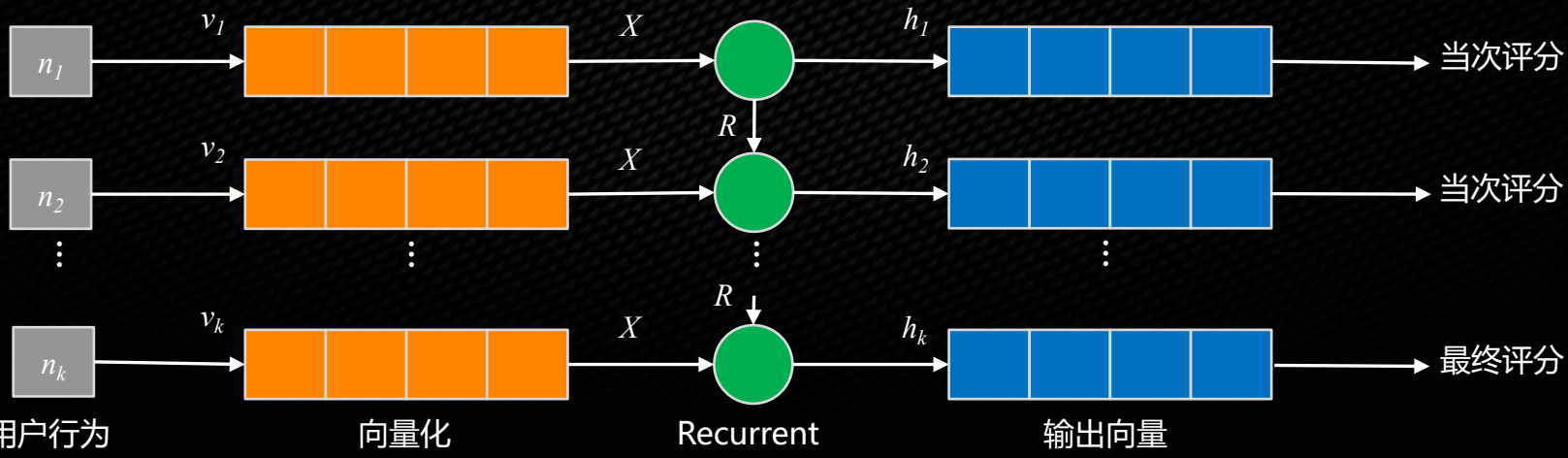
对于分类结果，无法直接计算准确率。只能通过抽样调查或者反馈间接进行估计。

潜在的攻击入口

因为用户反馈数据直接投入到机器学习系统的训练集。虽然加入了很多控制措施，但仍然给了攻击者一个可以影响模型的潜在入口

深度学习


n_1
n_2
n_k
v
X
最终评分
用户行为
向量化
线性模型


n_1
v_1
X
h_1
当次评分
R
n_2
v_2
X
h_2
当次评分
R
n_k
v_k
X
h_k
最终评分
用户行为
向量化
Recurrent
输出向量

05

Part One

总结

Conclusion


Silver Bullet
Just Ahead

谢 谢 观 赏

汇报人：何为舟　汇报时间：2017.10.18