机器学习&攻击检测

携程 / 岳良

目录

1 ML&恶意域名检测

2 ML&web攻击检测

3 Question

ML&恶意域名——检测需求的诞生

那一年应急带来的痛苦回忆

xshell，wannacry

依靠传统杀毒软件？

NO！还不如靠朋友圈

怎么解决

买威胁情报服务？

自研产品，猜测存在隐患的随机域名？

看下恶意域名的特点

yhxpqrhnhnwmvvdcwj.eu

bxedpvqwnbrqilaawj.com

提炼特点

不可读

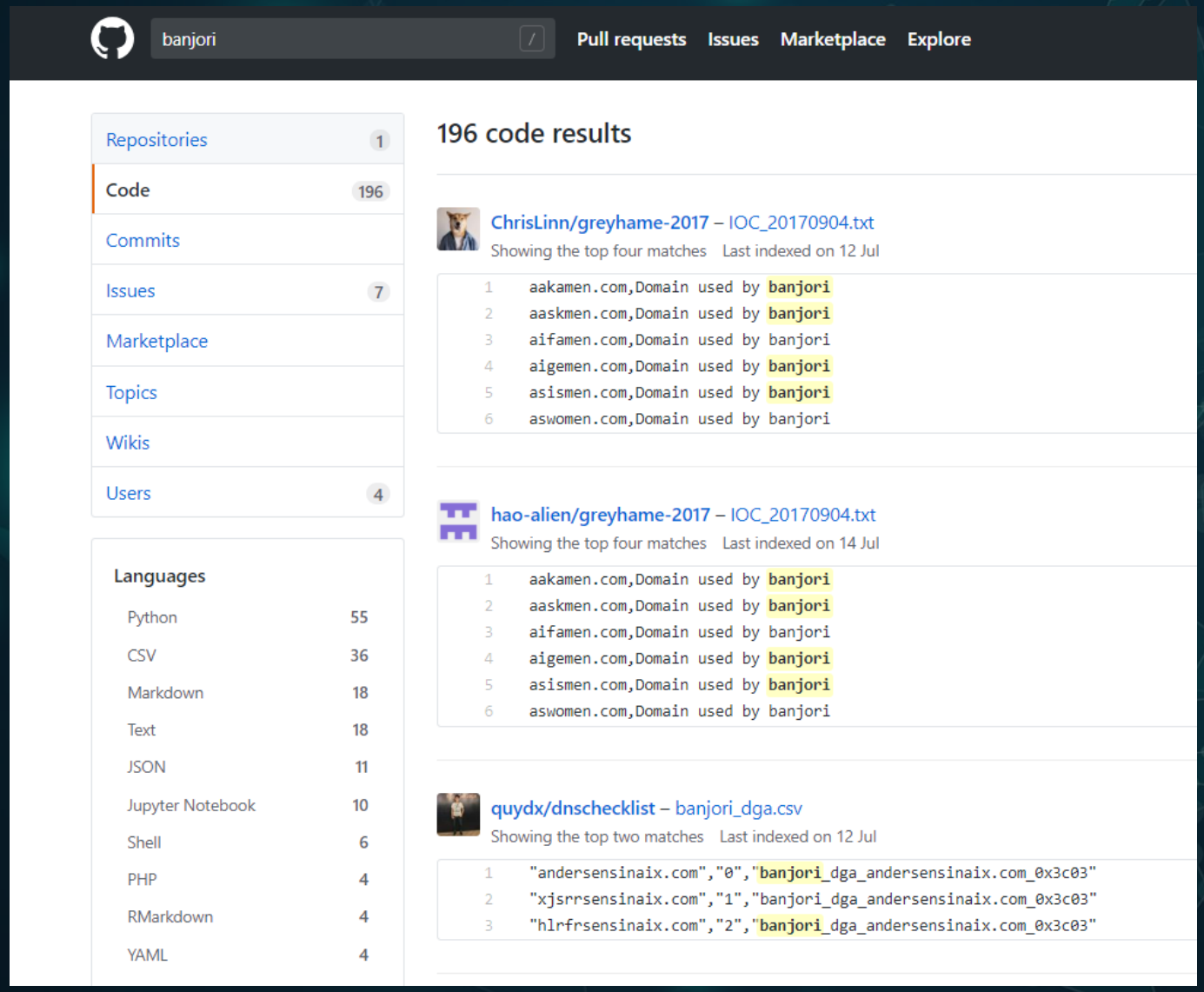
比较长

有的域名甚至都没注册

怎么检测

为什么人眼一看就知道哪些域名可疑

ML&恶意域名——DGA算法



ML&恶意域名——检测产品设计

字典匹配

携程内部的专有域名
alexa top 100万的域名
收集的100万黑域名

机器学习检测

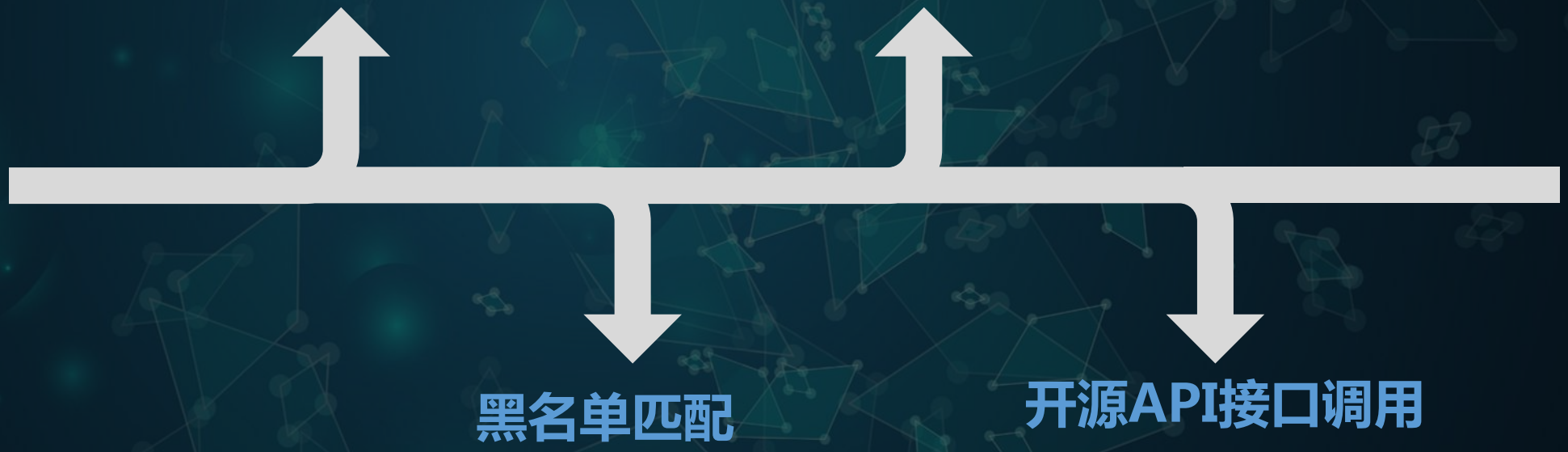
TFIDF提取特征训练模型

外部威胁情报API调用

调用某国外厂商接口做机器学习结果的二次验证

ML&恶意域名——检测产品设计

疑似为黑的展示在首页，判断为白的放过，结果暂存在缓存中



ML&恶意域名检测——ML步骤

步骤

- 定义问题
- 数据收集，清洗
- 特征工程
- 训练模型
- 评估模型效果，改进
- 上线调用模型

随机域名识别的二分类问题

外部各100万黑白样本

tfidf提取

xgboost算法

交叉验证，线上测试

线上RUN

ML&恶意域名检测——总结人脑的工作流程

问题？小学生A今年7岁。有一天拿到了一篇不带标题的微信文章，请问其是如何分辨这篇文章的主题是“世界杯”还是“娱乐”

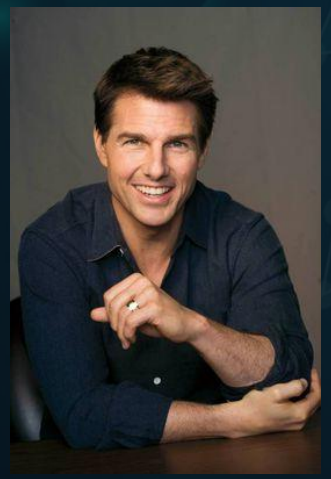
二分类问题

首先，如果小学生A家里没网没电视没报纸，那他肯定分辨不出来

没有样本

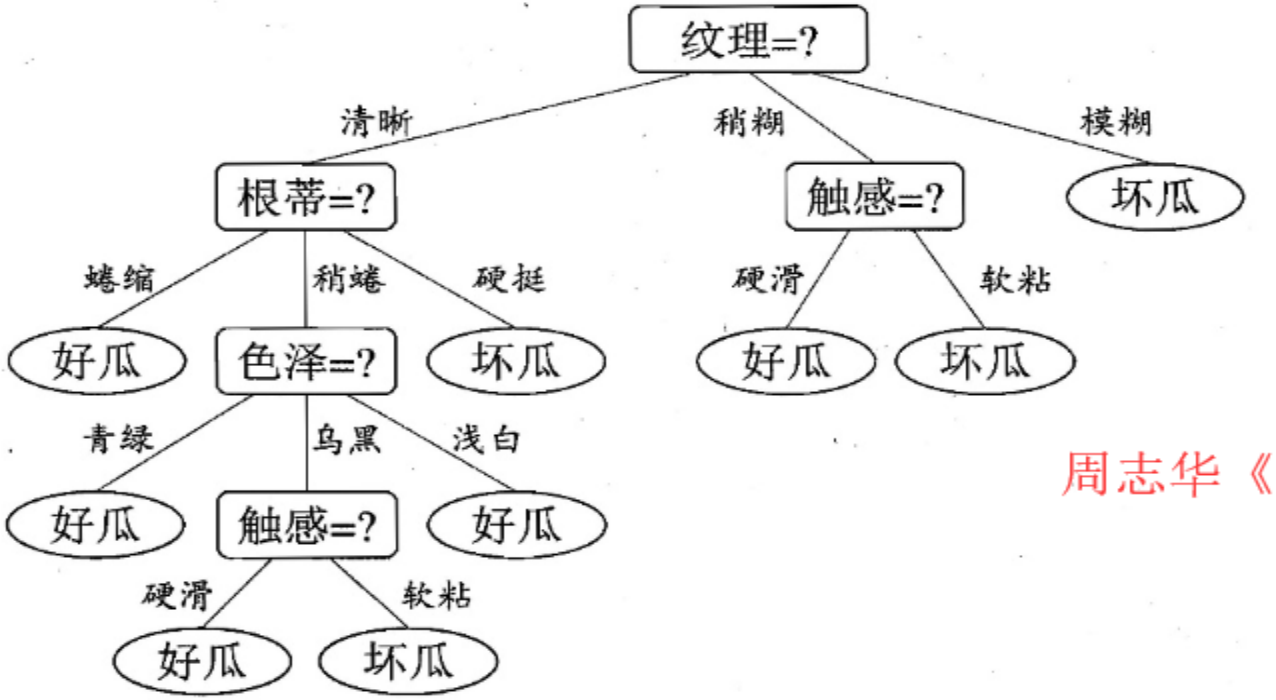
小学生A家里新装了宽带，但是A骨骼惊奇，将网上的世界杯和娱乐新闻全部转成了汉语拼音,然后得到结论，a ba fa de几个拼音出现多的就是世界杯

特征提取有问题




ML&恶意域名检测——总结人脑的工作流程

经过老师的指点，A重新改了特征（分词取汉语关键词出现频率），且大脑总结了一套算法，先看是否有一次“足球”，如果有再看有没有超过2次“梅西”，如果有就基本95%确定是世界杯新闻了 **训练并使用了“决策树”机器学习算法**



周志华《机器学习》决策树

ML&恶意域名检测——特征提取Tfidf

国际足联/裁判委员会/主席/科里纳/在/总结/48/场/小组赛/时/表示，共用/VAR/查看了/335/次/犯规，每场/比赛/接近/7/次/。这些/犯规/中，在/没有/VAR/的/情况/下，95%/的/判罚/是/正确/的，但/VAR/更正/了14/次/判罚，让/准确率/提升/到/99.3%。

在/今年/金马奖/官宣/前,巩俐/就/已经/先后/出任/过/第/50/届/戛纳/国际/电影节/主竞赛/单元/评委/、第/50届/柏林/国际/电影节/主席/、第/59/届/威尼斯/国际/电影节/主席。

ML&恶意域名检测——特征提取Tfidf

ctrip.com,ngram=2 ,按字符抽取

得到[ct, tr, ri,ip]

计算ct的词频TF
TF=ct在[ct, tr, ri,ip]出现的次数1/数组长度4=0.25

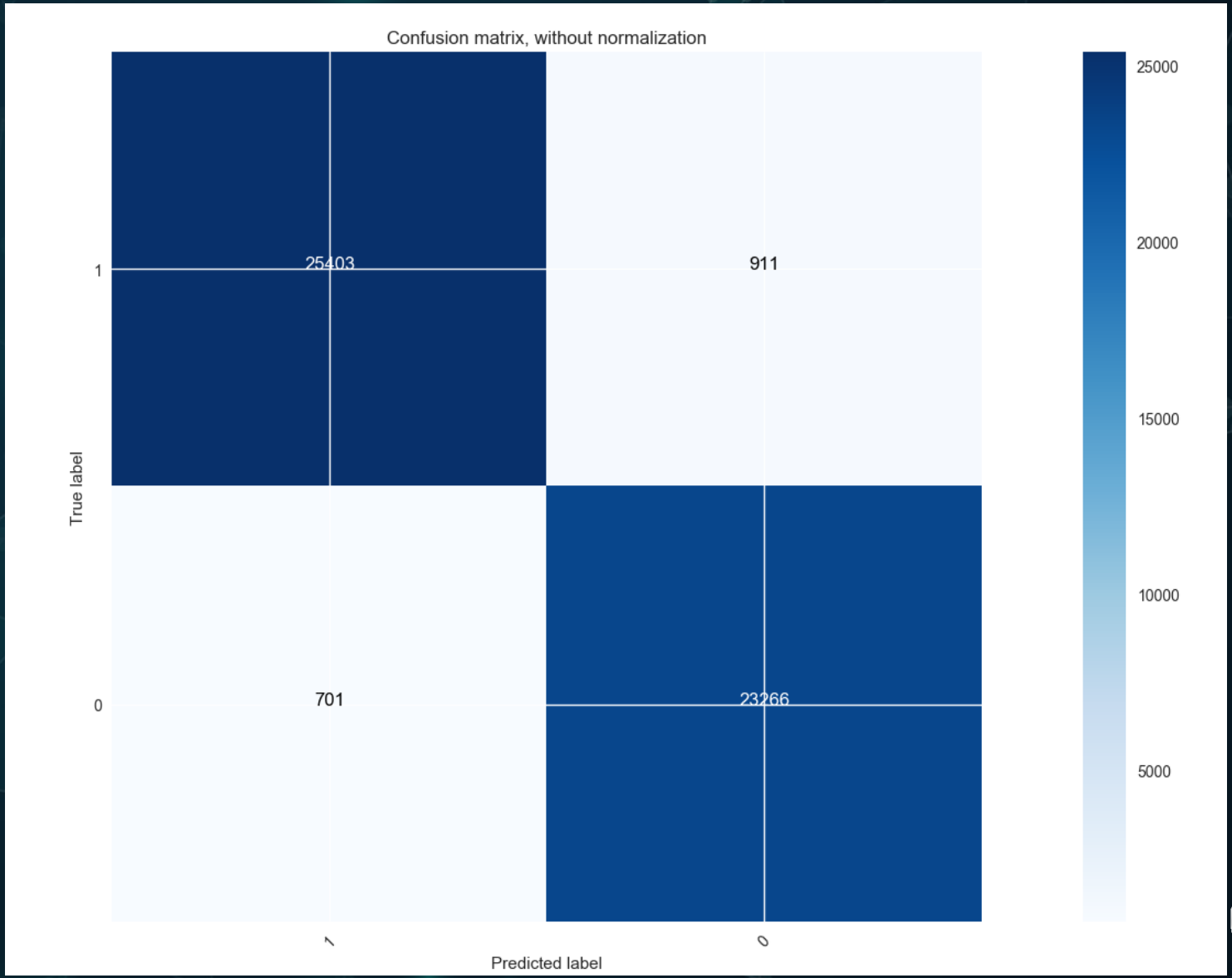
计算ct的逆文档词频IDF
IDF=log(训练语句总数/(含ct的语句个数+1))=log(10000/11)

TFIDF = TF * IDF

ML&恶意域名检测——模型效果

决策树
Precision: 0.933
Recall: 0.900
F1: 0.916

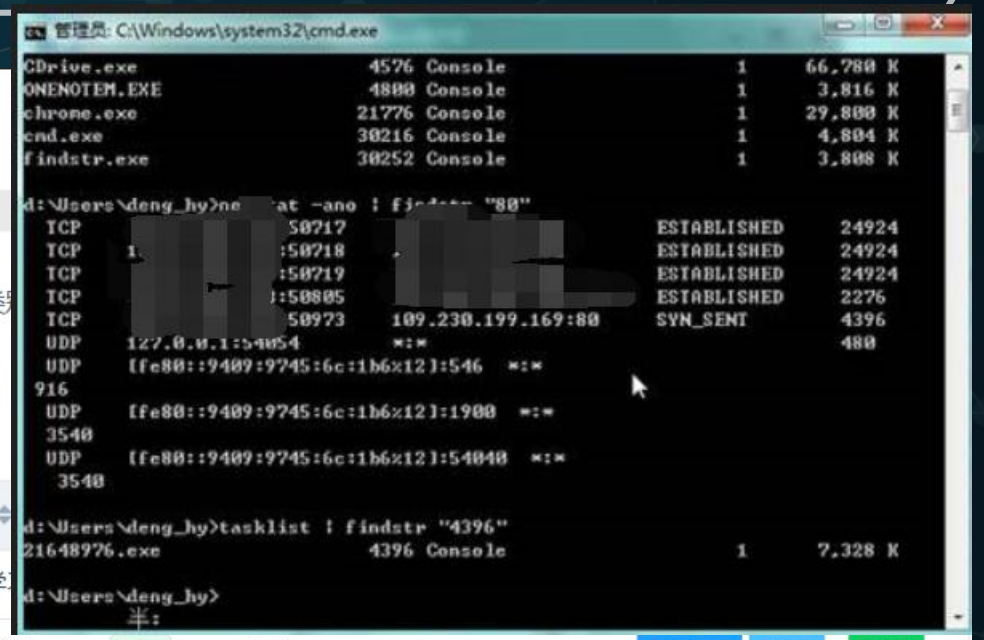
xgboost
Precision: 0.973
Recall: 0.965
F1: 0.969



说一下为什么3.5%的漏报率可以接受

sentqq.com
sexoqq.com
sexyos.net
goodfred431.com

ML&恶意域名检测——实际效果



识别结果 白名单 黑名单 帮助文档

时间 2018/04/07 - 2018/09/01 域名 f582e064be47 服务器类

类别 全部 状态 全部 搜索

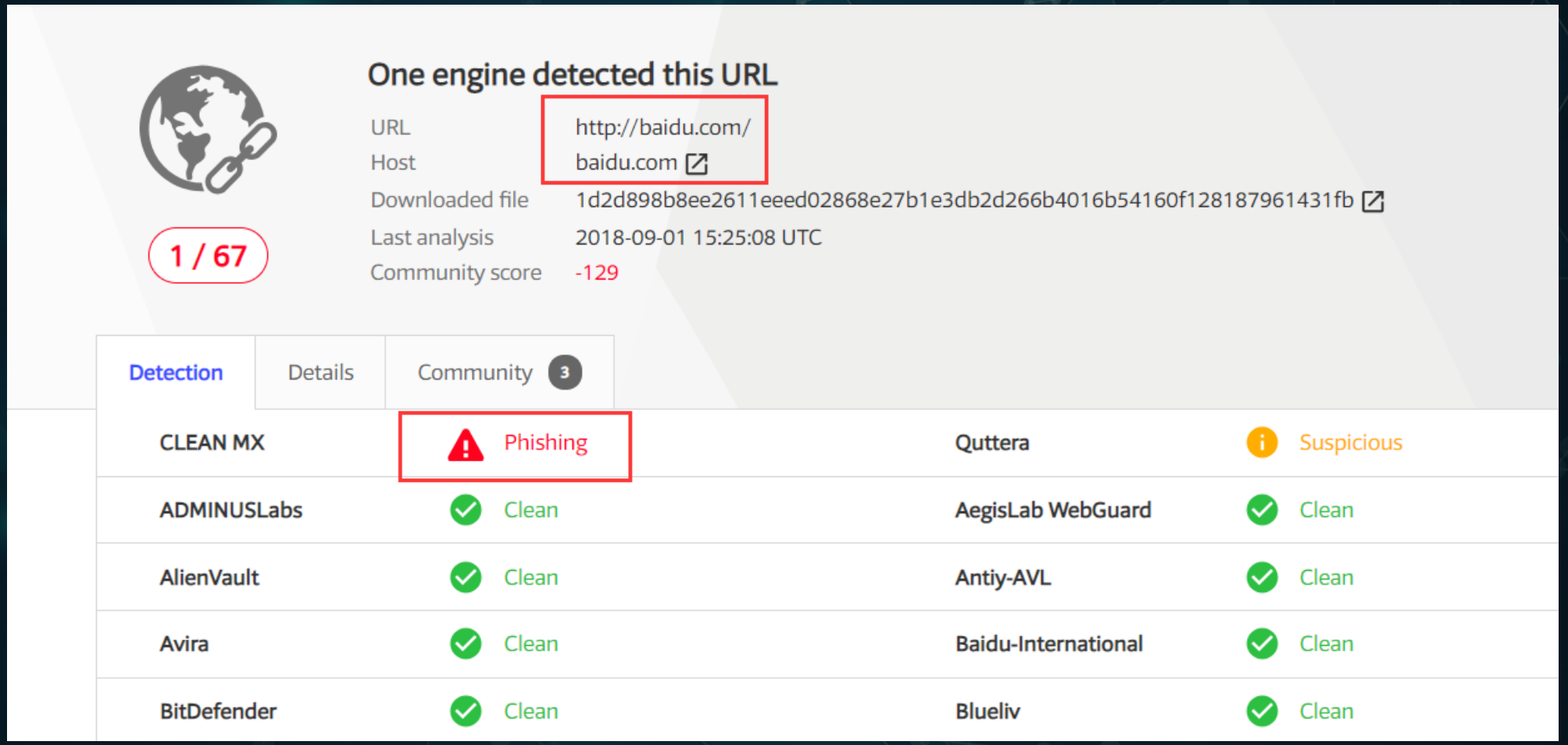
编号	时间	域名	次数	类别			
556967	2018/04/26 00:14:22	f582e064be47	90	机器学			
554110	2018/04/25 09:02:47	f582e064be47	79	机器学习	确认	N/A	标记 备注 推送
553646	2018/04/24 18:33:52	f582e064be47	6	机器学习	待定	N/A	标记 备注 推送
551548	2018/04/23 16:32:29	f582e064be47	3	机器学习	待定	N/A	标记 备注 推送
565017	2018/04/29 11:37:09	f582e064be47	1	样本数据	确认	http://camaradeideias.com.br/...	标记 备注 推送
569527	2018/05/03 13:46:35	f582e064be47	1	样本数据	确认	http://camaradeideias.com.br/...	标记 备注 推送
566202	2018/05/01 07:45:16	f582e064be47	1	样本数据	确认	http://camaradeideias.com.br/...	标记 备注 推送
566039	2018/04/30 21:01:20	f582e064be47	1	样本数据	确认	http://camaradeideias.com.br/...	标记 备注 推送

ML&恶意域名检测——后续改进

误报

http://www.jxskqyy.com/

ML&恶意域名检测——API校验注意点



One engine detected this URL

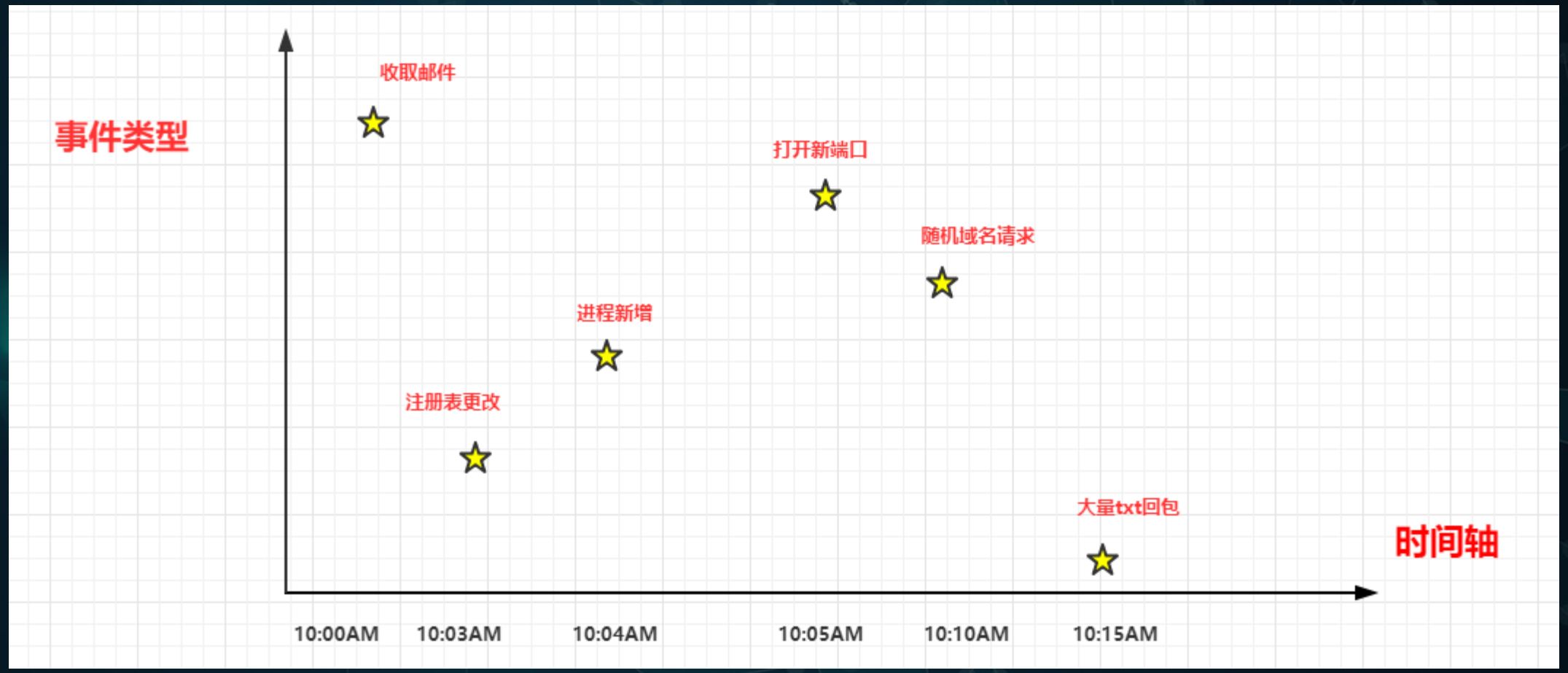
1 / 67

URL http://baidu.com/
Host baidu.com
Downloaded file 1d2d898b8ee2611eeed02868e27b1e3db2d266b4016b54160f128187961431fb
Last analysis 2018-09-01 15:25:08 UTC
Community score -129

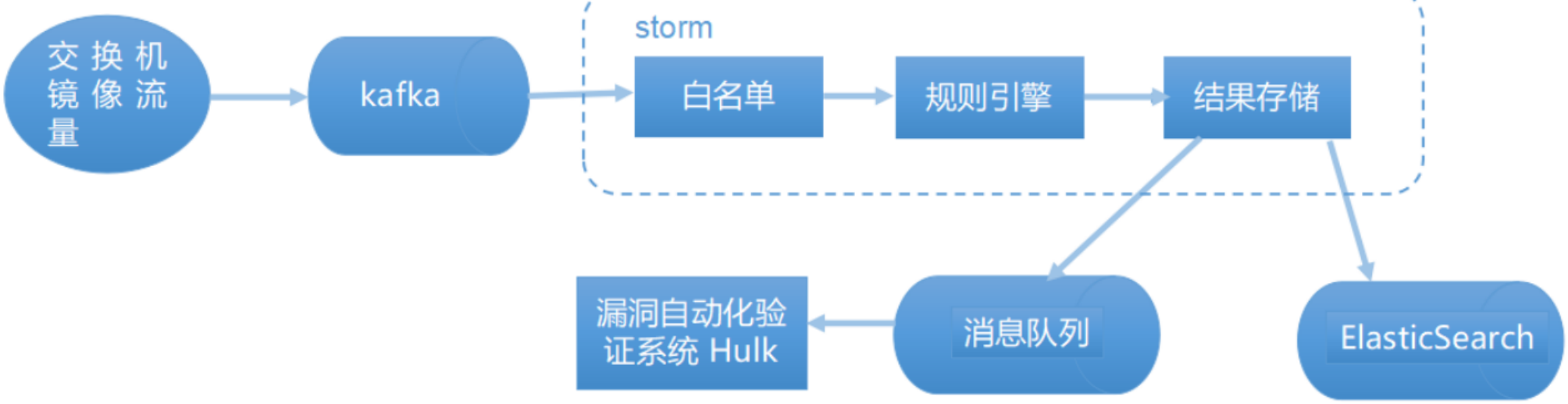
Detection | Details | Community 3

CLEAN MX	Phishing	Quttera	Suspicious
ADMINUSLabs	Clean	AegisLab WebGuard	Clean
AlienVault	Clean	Antiy-AVL	Clean
Avira	Clean	Baidu-International	Clean
BitDefender	Clean	Blueliv	Clean

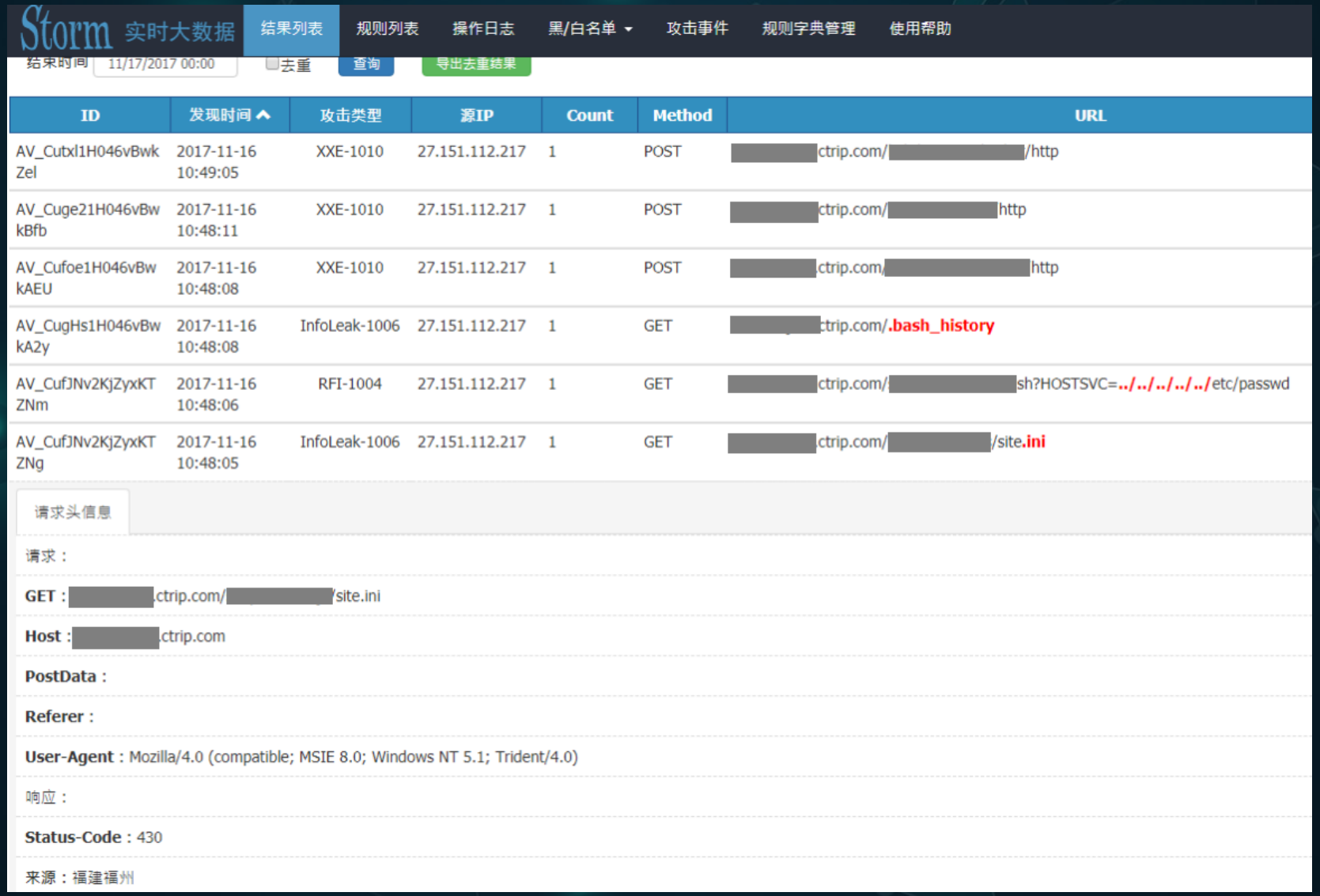
ML&恶意域名检测——产品的下一步融合



ML&web攻击检测——攻击复验系统介绍



ML&web攻击检测——攻击复验系统介绍



ML&web攻击检测

为什么想到用机器学习来做web攻击检测？

例如一条检测sql注入的正则语句如下：

```
String inj_str =
"'|and|exec|insert|select|delete|update|co
unt|*|%|chr|mid|master|truncate|char|decl
are|;|or|-|+|,";
```

新买的selected衬衫脏了！

ML&web攻击检测

为什么想到用机器学习来做web攻击检测？

pattern:[^\w\s\?\此处保密\!\@\(\'\~]{1,}

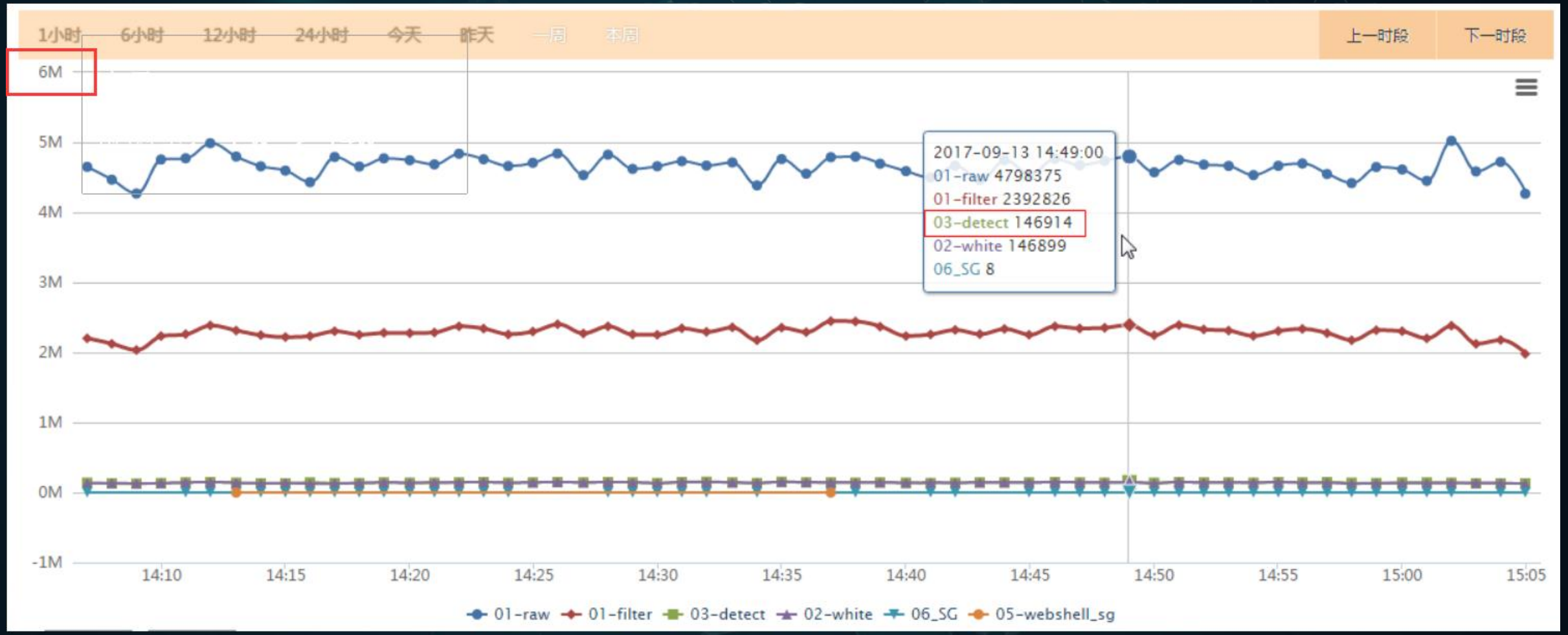
规则难以维护，自己写的正则自己都读不懂

ML&web攻击检测——未改进前

为什么想到用机器学习来做web攻击检测？

队列积压严重，根本消费不完，昨天的攻击今天还没检测

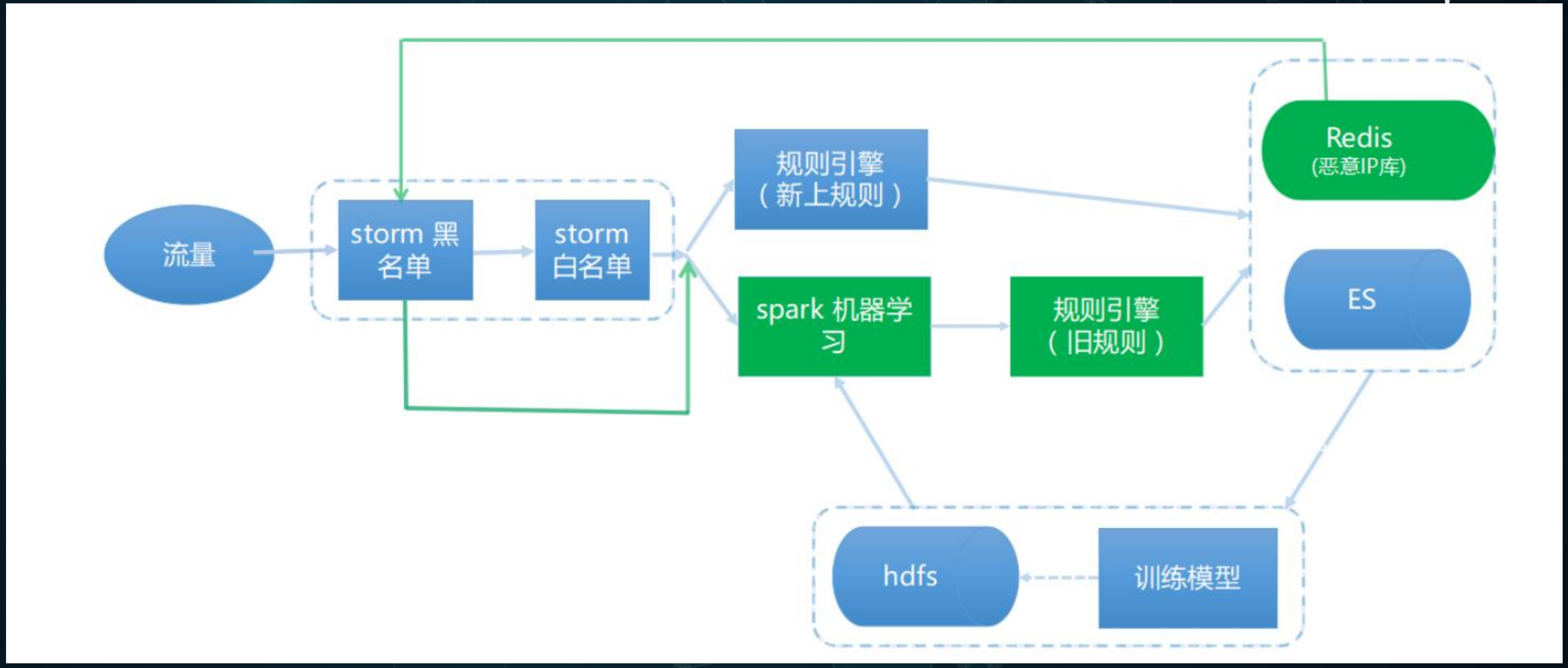
ML&web攻击检测——改进后性能效果



ML&web攻击检测——和正则对比效果

Time	is_ml	rule_result	postdata
July 3rd 2018, 15:34:04.669	true	white	out_trade_no=1122%00'&subject=0
July 3rd 2018, 15:34:01.091	true	white	Y2hlbmdzaGlzLmMjd=echo md5(zhimakaimen233333333);
July 3rd 2018, 15:33:02.170	true	white	second=1@%7Cecho%20%27%3C%3Fphp%20phpinfo%28%29%3B%3F%3E%27%20%3E%20dsanhssds.php
July 3rd 2018, 15:32:01.845	true	white	name=1@|echo '<?php phpinfo();?>' > ssfds.php
July 3rd 2018, 15:28:07.650	true	white	{"username":"' and 1=char(106) --"}
July 3rd 2018, 15:21:07.303	true	white	key=%' and 1=char(106)-- &method=GetDeptJSON&org=2&all=0&hr=0&dll=IBP_Core.dll&class=IBP_Core.OrgStructure
July 3rd 2018, 15:21:04.452	true	white	uid`%3d-42873%0bor%0b42873%3d42873%23=1
July 3rd 2018, 15:21:01.360	true	white	login=%df%27test&email=test@test.com
July 3rd 2018, 15:16:03.862	true	white	username=eye' and 1=char(106) --&password=sdasd&expires=&butOk=

ML&web攻击检测——第6版框架



ML&web攻击检测——模块解释

白名单

args参数值不带英文标点和控制字符的请求，全都算白名单，不过任何检测引擎

www.ctrip.com?目的地=**北京**

ML和正则引擎的关系

过完白名单的请求，进入ML引擎，ML预测为黑则继续进入正则，

否则打上ML白色结果

恶意IP库

命中正则规则超过20次的ip，进入黑名单，缓存1day，此ip之后不过ML引擎，

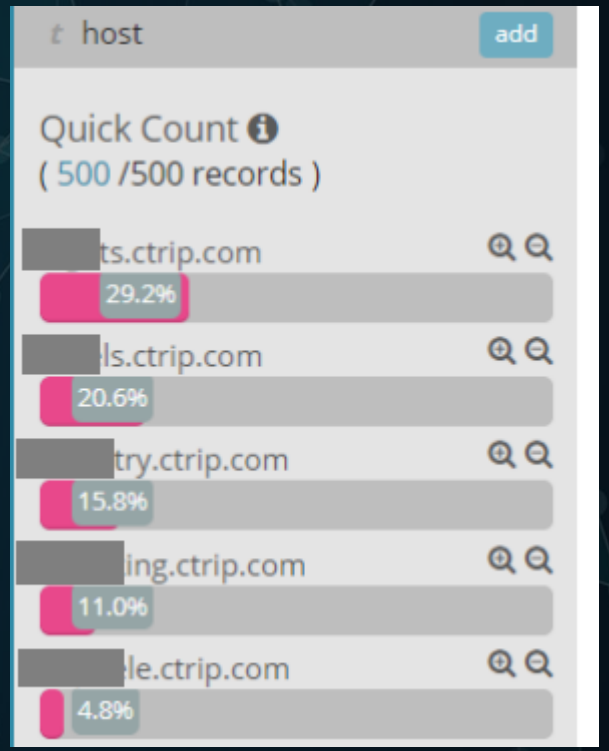
直接进正则

ML&web攻击检测——聊下训练数据的搜集

白日志来源

内网PPOBE日志

少部分外部白日志



难点：
1.大部分都是酒店机票的日志
2.大部分流量都不带任何标点和特殊字符
3.要达到黑白样本1:1

写脚本ES捞，手工洗，去重，日志尽量多样化
针对样本污染问题，使用正则关键字从白样本里洗掉黑色数据

ML&web攻击检测——Details

01 GET ,POST分开取args进行建模

http://test.ctrip.com/TrainBooking/Search.aspx?from=shanghai&to=beijing&day=2018-09-05

按args取训练素材，还方便使用网上直接搜集来的POC，因为不用考虑定制化

过拟合

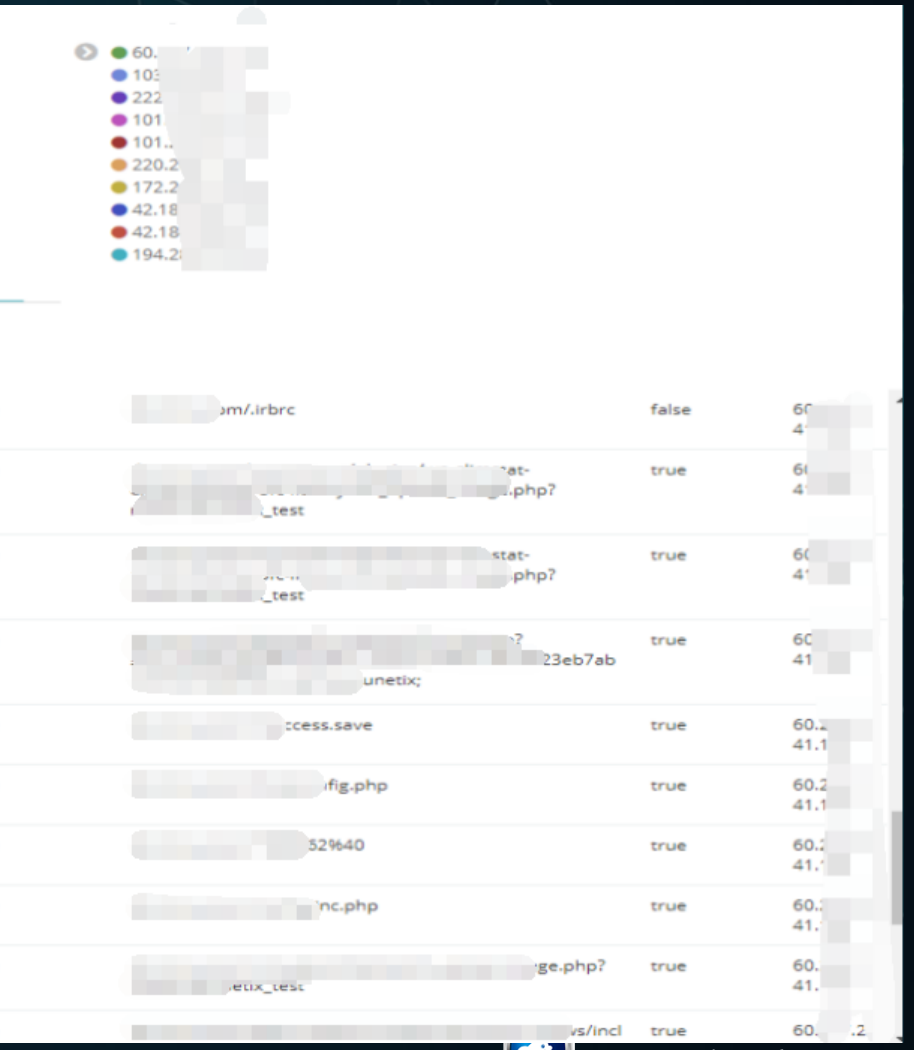
ML&web攻击检测——Details过拟合

02 图片上传，加密数据的怎么办？

正则检测不了的，机器学习也不能强求
多层防御，用其他方式来解决

ML&web攻击检测——动态黑名单弥补漏报

03 漏报了怎么办



04 误报了怎么办

误报的影响不大，直接交给正则，不同于WAF的使用场景

使用HMM做异常检测

HMM异常检测——问题抛出

世界杯开始前，C罗状态好的概率为90%

……

C罗状态好，葡萄牙赢球概率为90%

……

C罗这一场状态好，下一场状态好的概率为85%

….

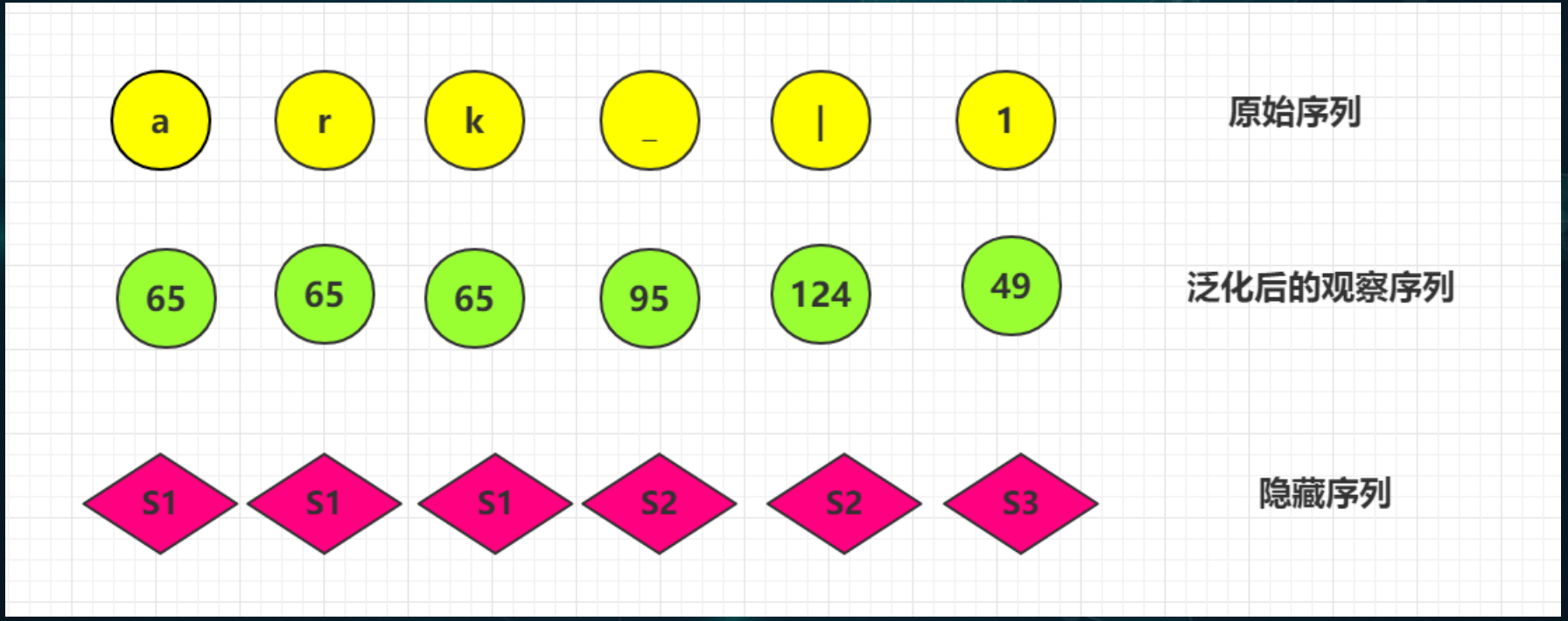
问葡萄牙世界杯开始3连跪的概率是多少

HMM异常检测——泛型化

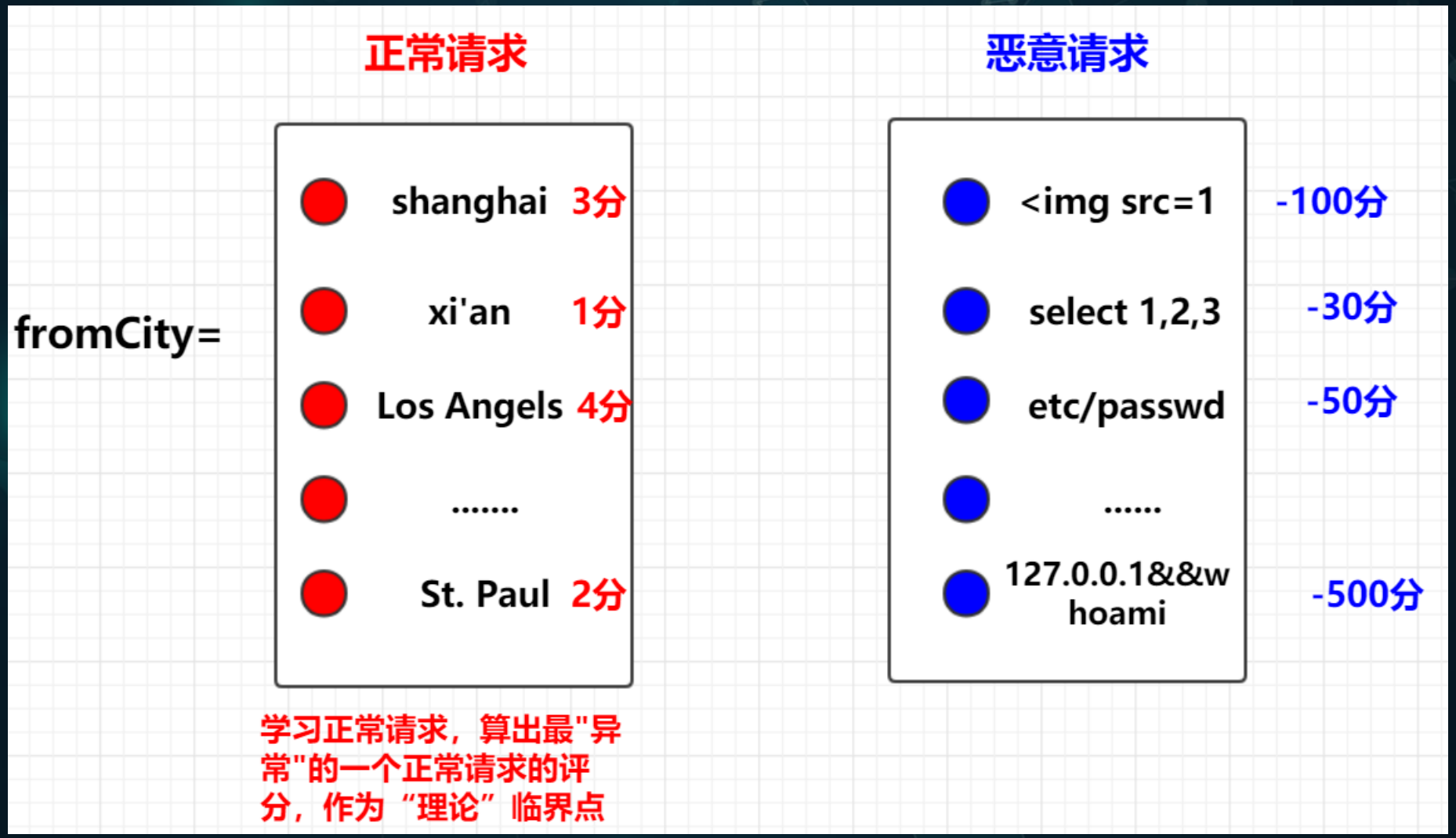
某参数正常的张这样，问题：来一个请求，判断是否是异常的

ark_bus_vivo|12308
ark_bus_xiaomi|12308
ark_android_jpskb|TY
ark_bus_hicloud|ky12308

首先做泛型：
字母-> ord(A) 数字-> ord('N') 中文-> ord('C') 其他->
取其ASCII码



HMM异常检测——计算观察序列概率



HMM异常检测——样本里的异常点和分布

deviceId
00277AD2-AFD1-4F79-AED9-E41[redacted]
00000000-0000-0000-0000-000000000000
001A7481-9EC4-42AB-97FA-7[redacted]
005325AA8D278EB3C5DC[redacted]

deviceId 00000000-0000-0000-0000-000000000000得分：-114.04052019890963

<script>alert(123)得分：43.74534140982247

样本收集最好做到分散：从不同的源ip收集，避免单个ip贡献过多样本，恶意ip库里面的数据坚决删除

监测模型是否待更新？或者干脆定期更新

Questions


THANKS