




# •安全数据科学分享

Via@91Ri-Team 隐形人真忙



# FrameWork

- 机器学习的基本概念
  - 机器学习在安全领域的应用
  - 如何入门和学习安全数据分析
- 

## 机器学习的基本概念

探究和开发一系列算法来实现如何使计算机不需要通过外部明显的指示，就可以自己通过数据来学习，建模，并且利用建好的模型和新的输入来进行预测的学科。


**传统应用：**

**语音识别、自动驾驶、语言翻译、计算机视觉、推荐系统、无人机、识别垃圾邮件**



## 一些术语

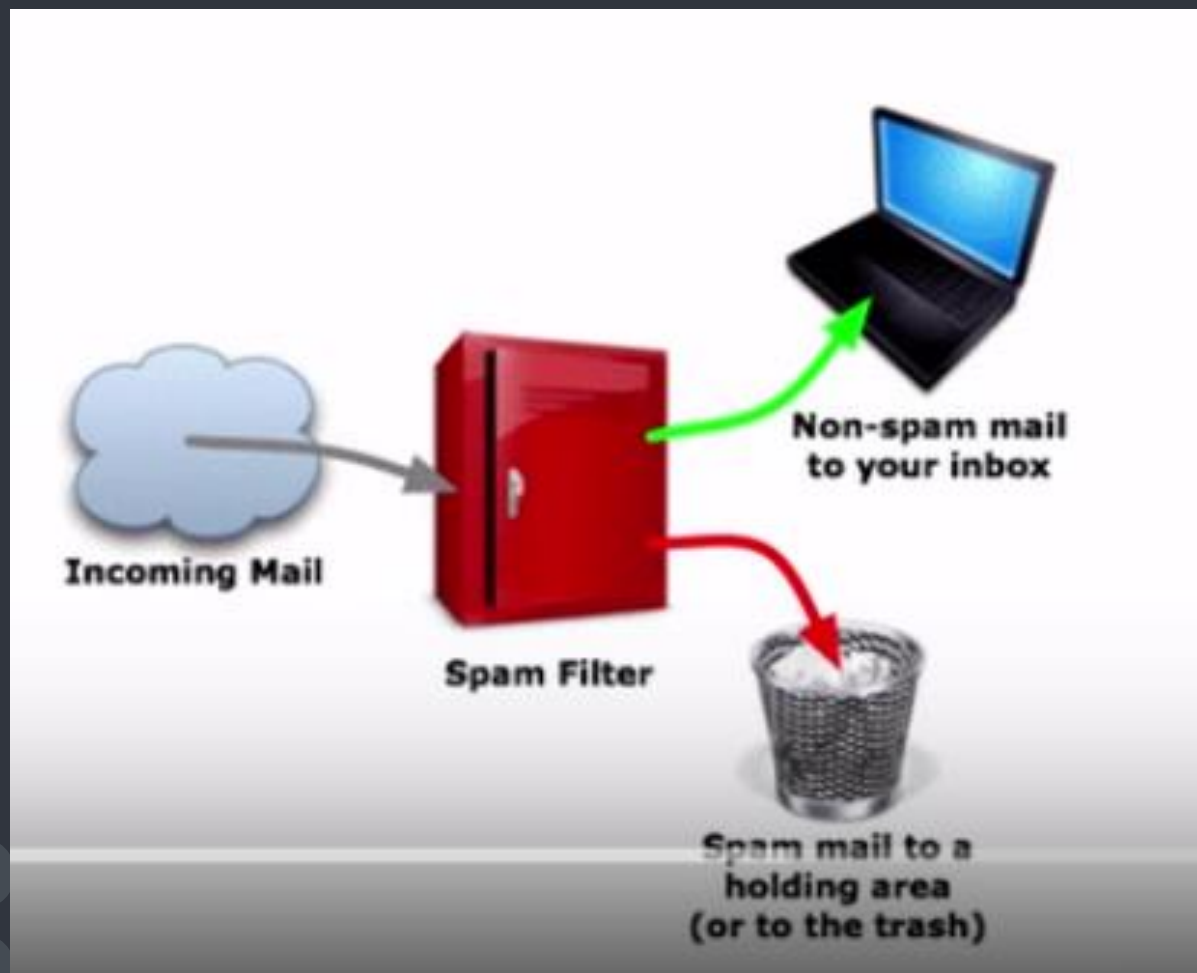
训练集、  
测试集、  
特征向量（特征工程）、  
监督学习、非监督学习、  
分类、聚类、  
回归、拟合



# 监督学习（分类）

例子：如何识别一封邮件是正常的还是异常的（垃圾邮件）

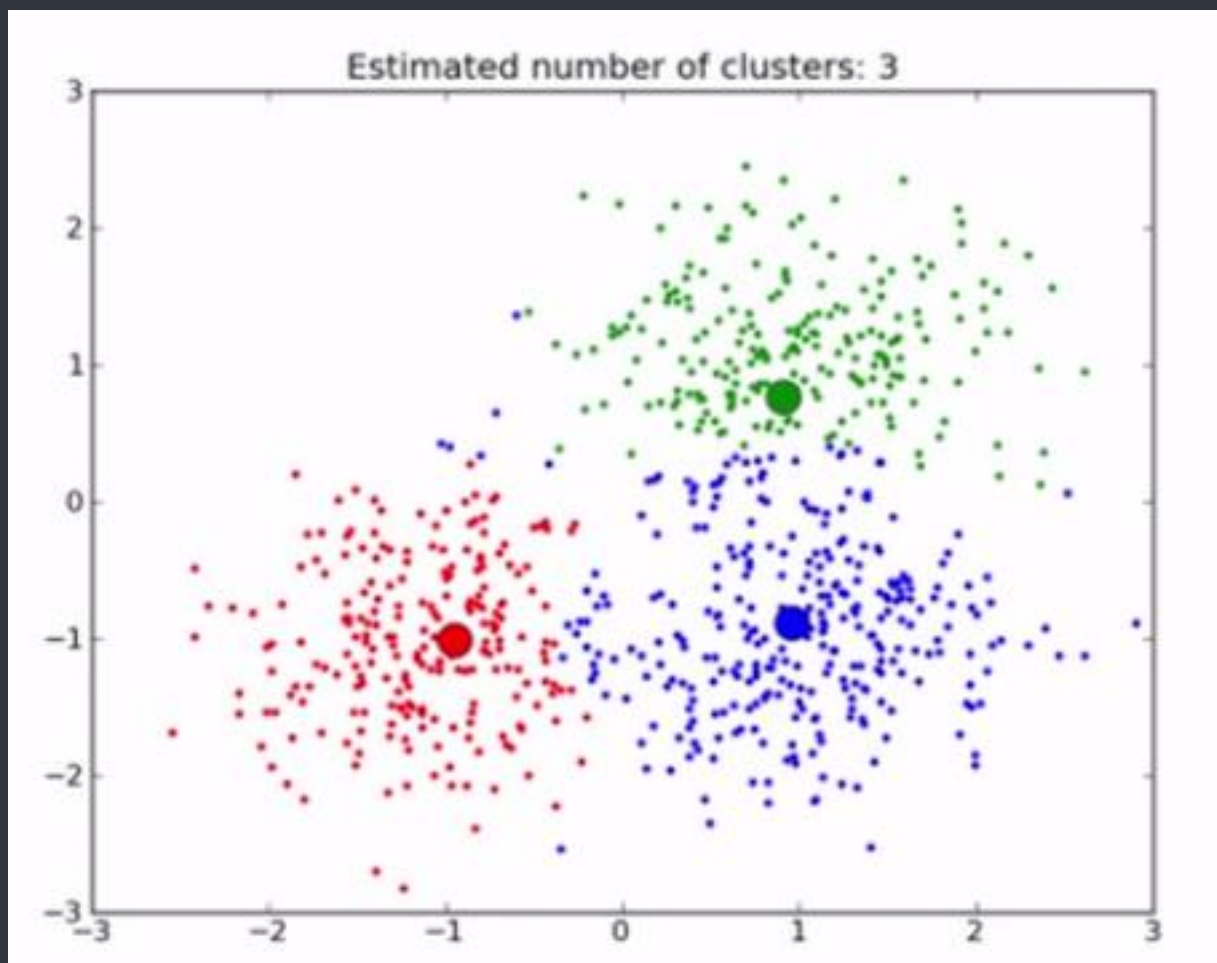
监督学习：有特定的输出，比如这里的（yes or no）



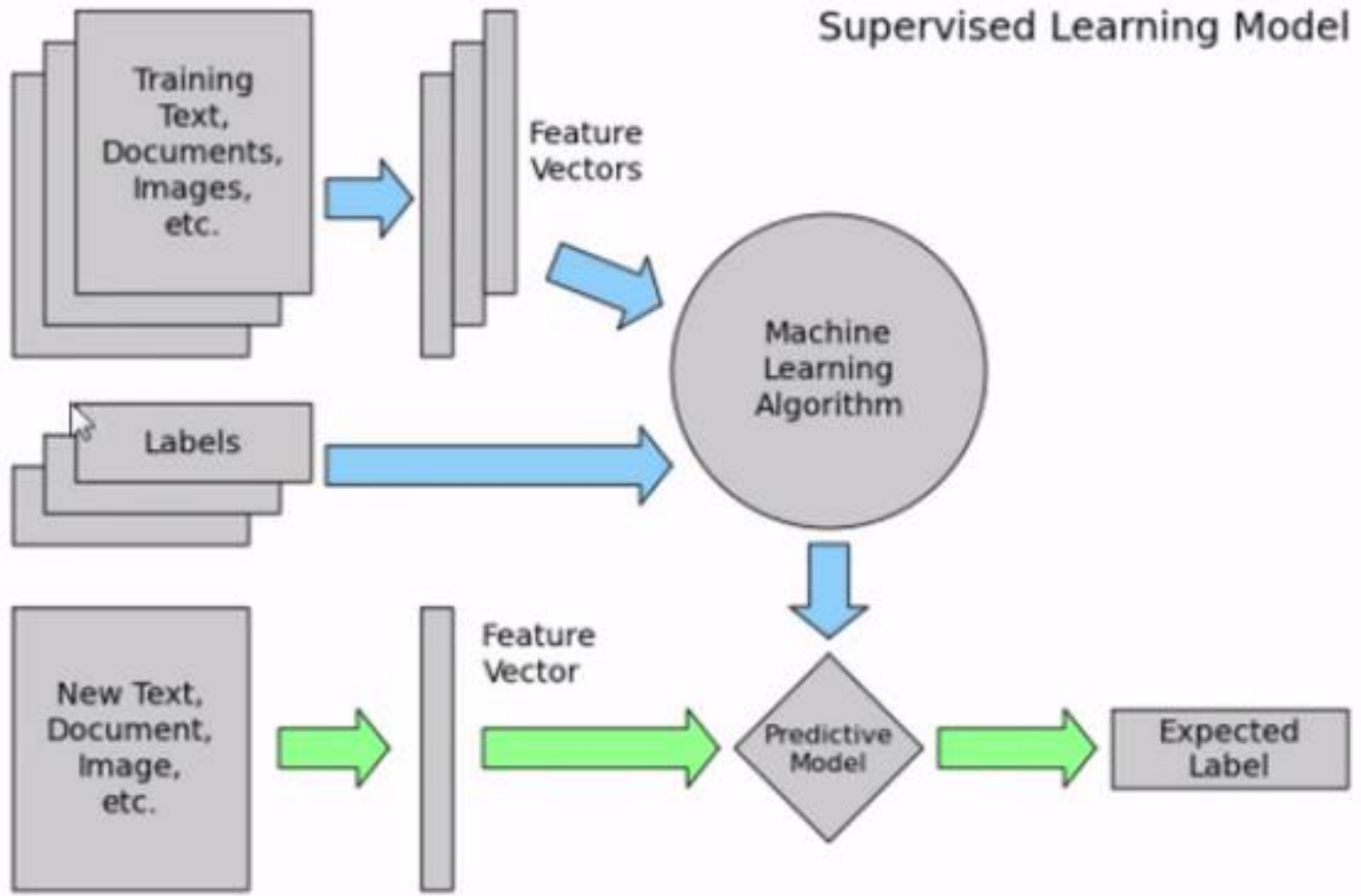
# 非监督学习（聚类）

例子：找出一堆文章中的热点词条，做成云标签

特点：没有实现固定的分类目标



# 测试集和训练集



# 机器学习在安全领域的应用案例

## 下一代的WAF、IDS、IPS（异常检测模型）


### User


 → `www.xxx.com/index.php?id=123`

 → `www.xxx.com/index.php?id=124`

 → `www.xxx.com/index.php?id=125`

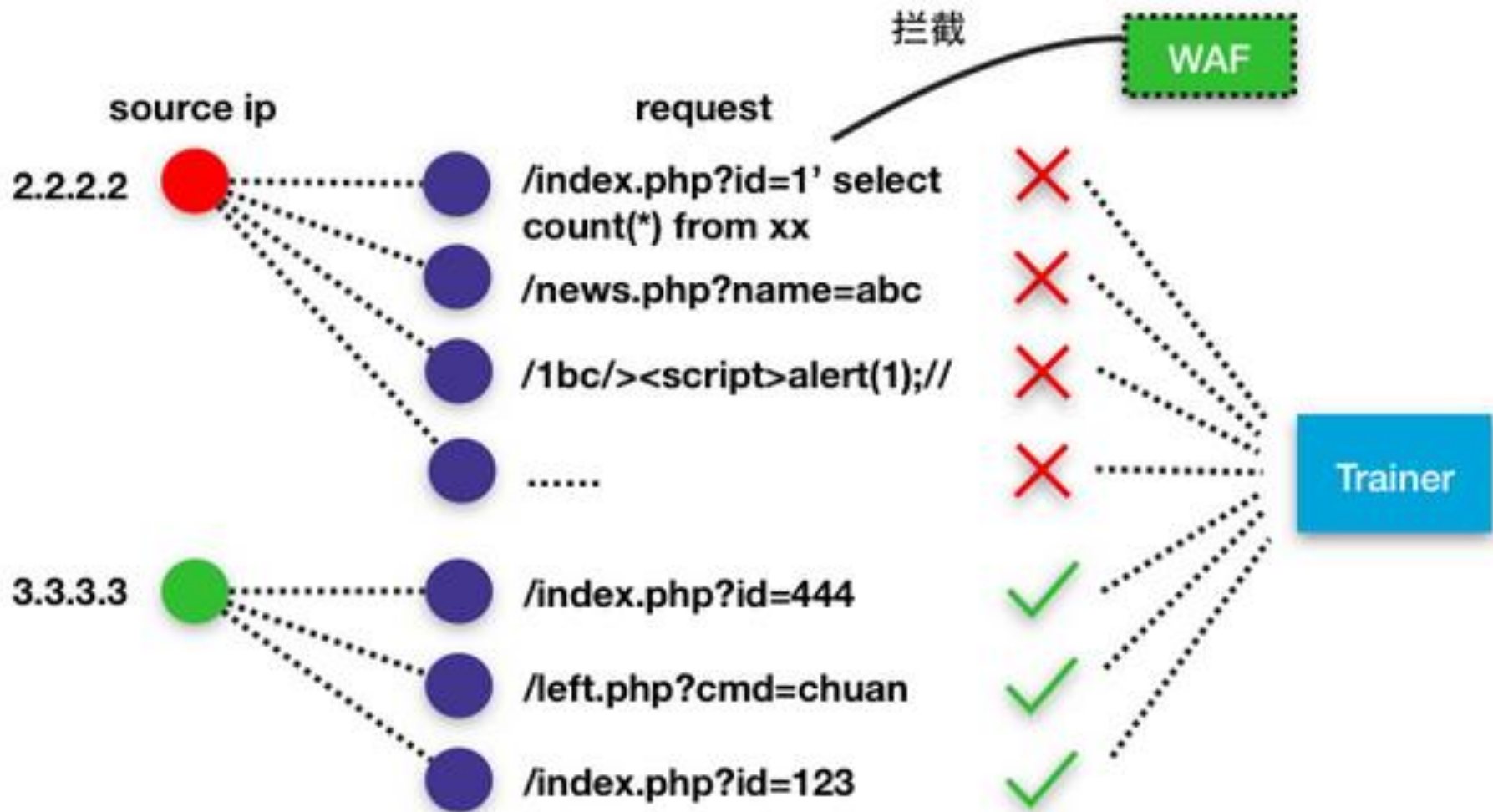
### Attacker

 → `www.xxx.com/index.php?id=123' union select xxx from xxx`

 → `www.xxx.com/index.php?id=%3Cscript%3Ealert('XSS')%3C`

 → `www.xxx.com/index.php?id=125$%7B@print(md5(123))%7D`





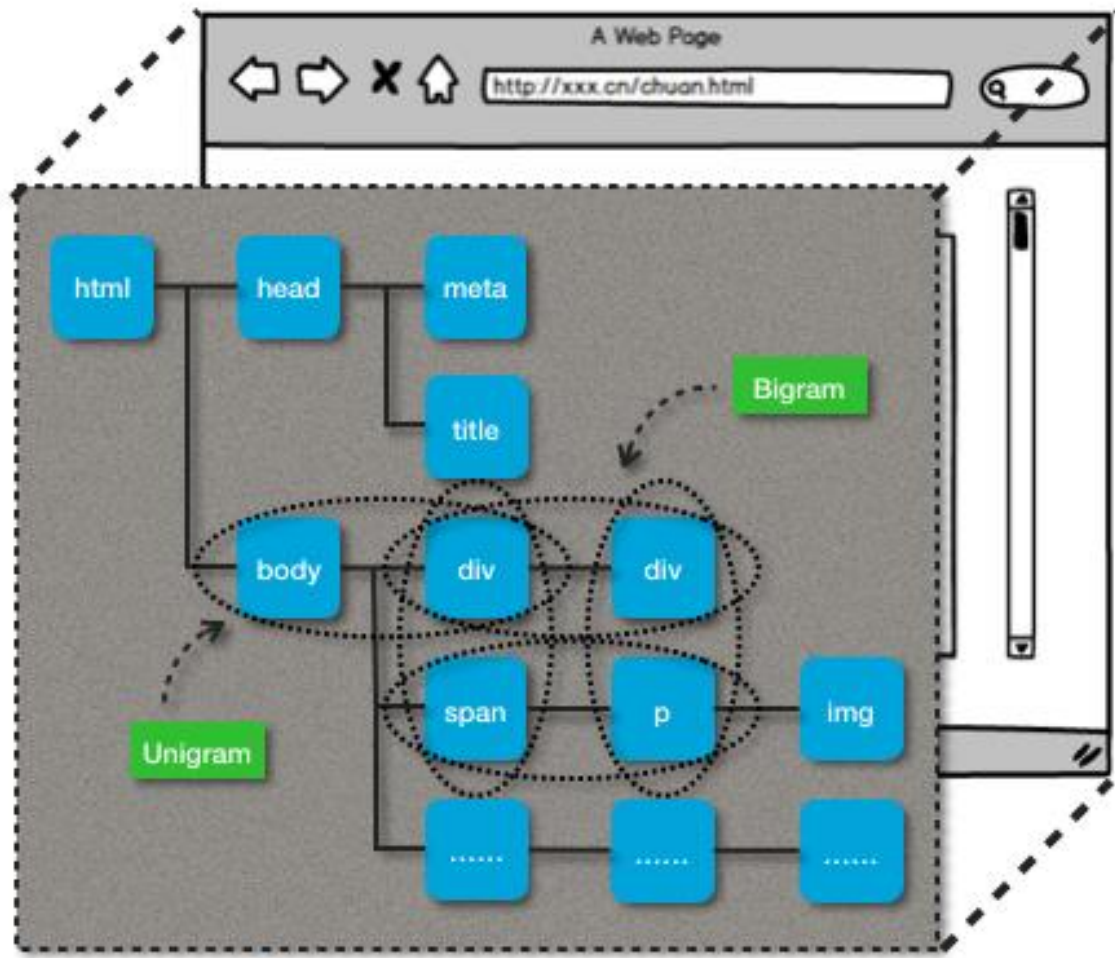
# Webshell识别

## 人工方法

- 1.从Http数据中发现一个url有点「异常」；
- 2.访问一下该uri，通过安全经验知识从返回页面识别出是webshell（大马）；
- 3.如果返回页面为空白（疑似一句话），从代码层面来识别。

## 机器学习思路

- 1.搜集大量的Webshell样本；
- 2.将原始的HTML转换为DOM-Tree
- 3.使用算法计算DOM-Tree之间的相似度
- 4.与阈值做比较，综合判断是否为webshell



提取向量



空间压缩



# 社交网络情报分析

twitter上挖出ISIS账号

QQ群泄露数据分析

分析舆论走势

.....



# 如何学习安全数据分析？

- 1、先学习基本的算法原理，补充数学知识——Coursera上的机器学习课程
- 2、学习Python的几个机器学习工具——pandas , numpy , seaborn , sklearn
- 3、去Kaggle上打比赛，学习特征工程和别人的代码
- 4、找project做
- 5、学习大规模数据处理——spark hadoop storm
- 6、.....



**Q & A**