聚·变

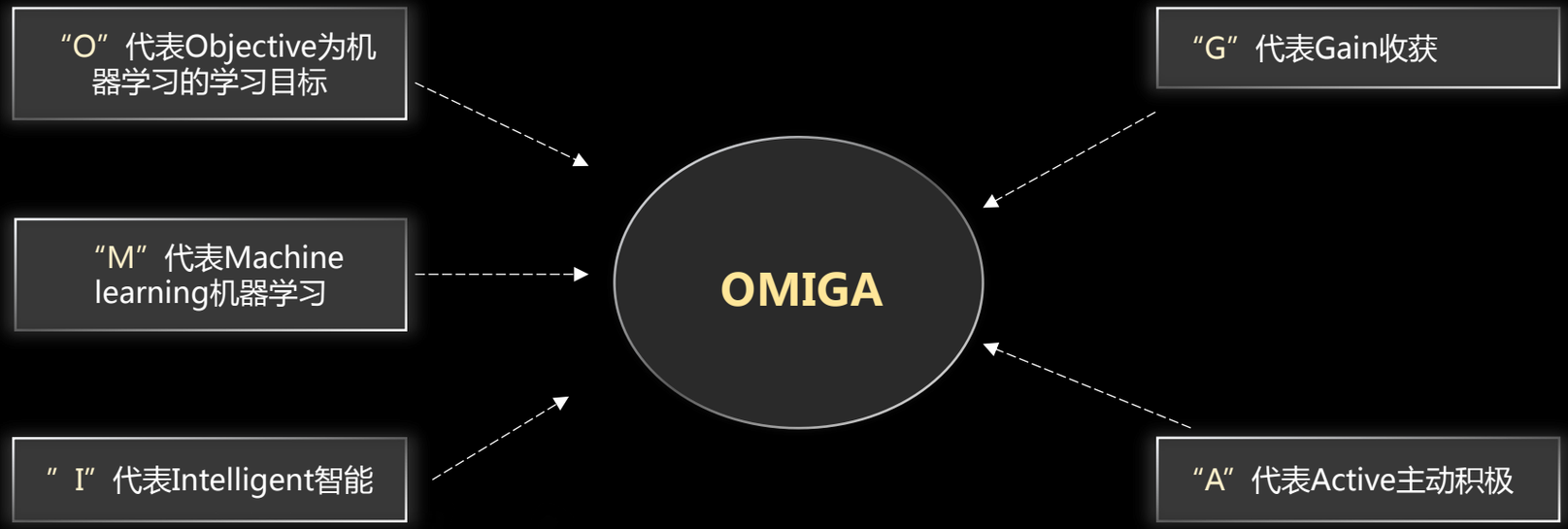
第二届顺丰信息安全峰会分论坛

AI 安全与隐私保护（1）

安全数据分析中对抗学习的研究与应用

张振海　　顺丰信息安全工程师

OMIGA


“O”代表Objective为机器学习的学习目标
“M”代表Machine learning机器学习
”I”代表Intelligent智能
OMIGA
“G”代表Gain收获
“A”代表Active主动积极

主要内容

1、安全数据分析中面临的问题

2、对抗学习

3、对抗学习在DGA检测上的应用

4、总结与未来工作

经常遇到的一个问题

大数据分析

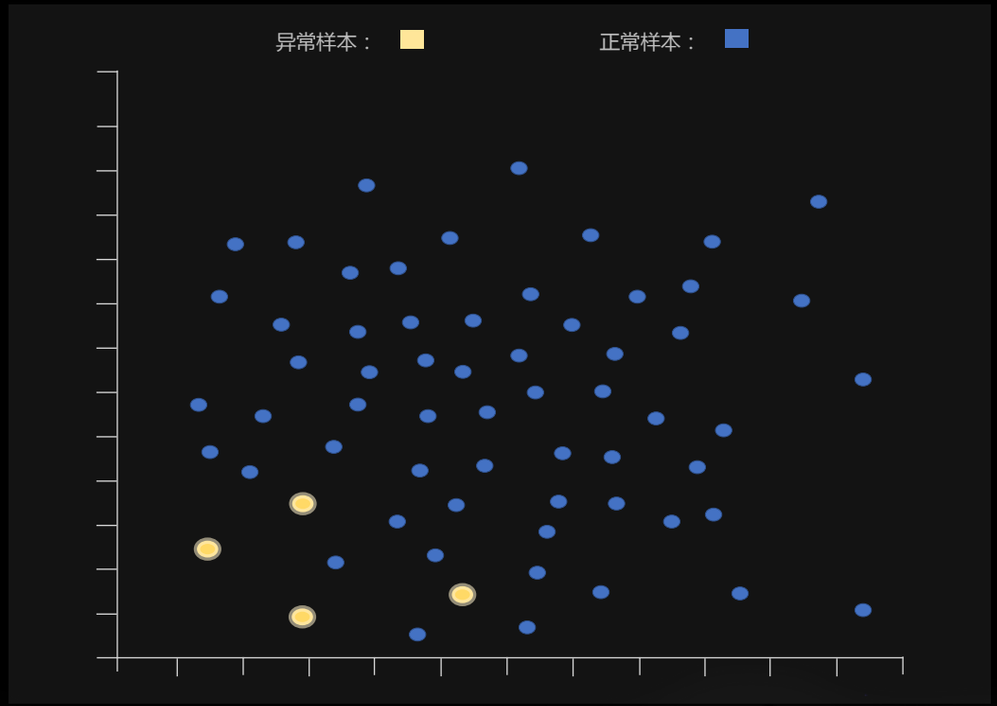
统计学

机器学习

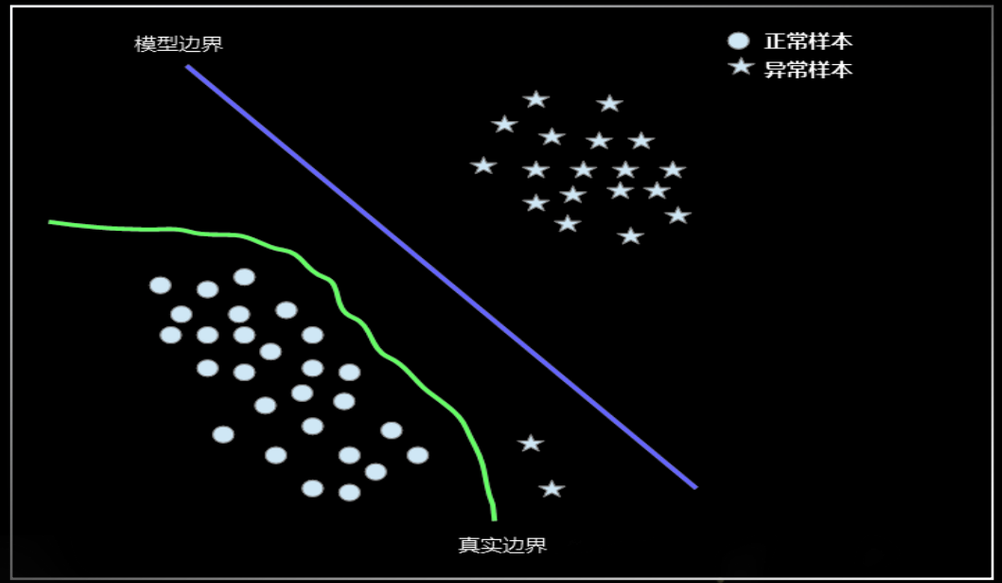
样本严重不均衡

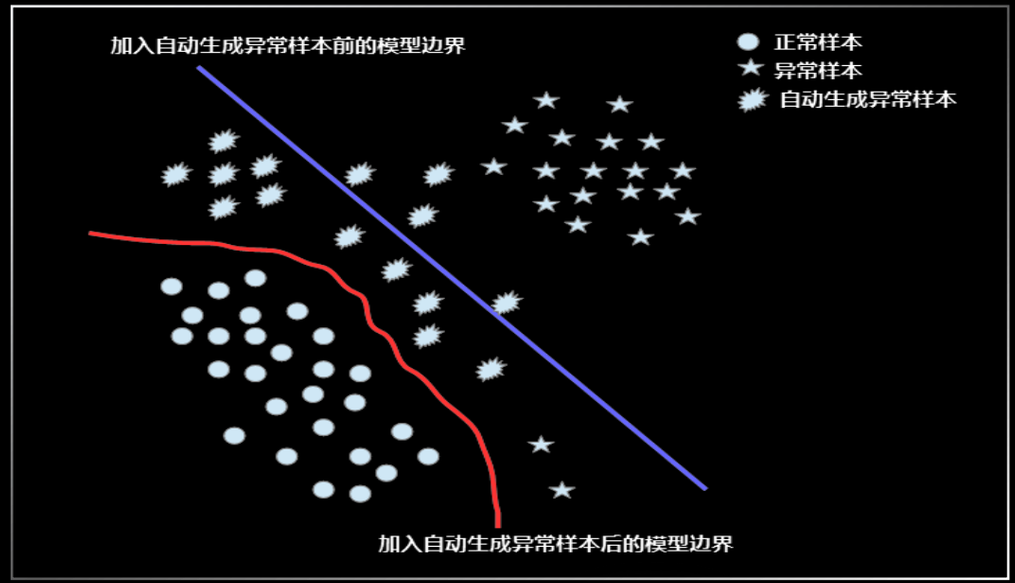
异常样本稀有

正常样本充足



能否自动生成异常样本，弥补异常样本的不足？





主要内容

1、安全数据分析中面临的问题

2、对抗学习

3、对抗学习在DGA检测上的应用

4、总结与未来工作

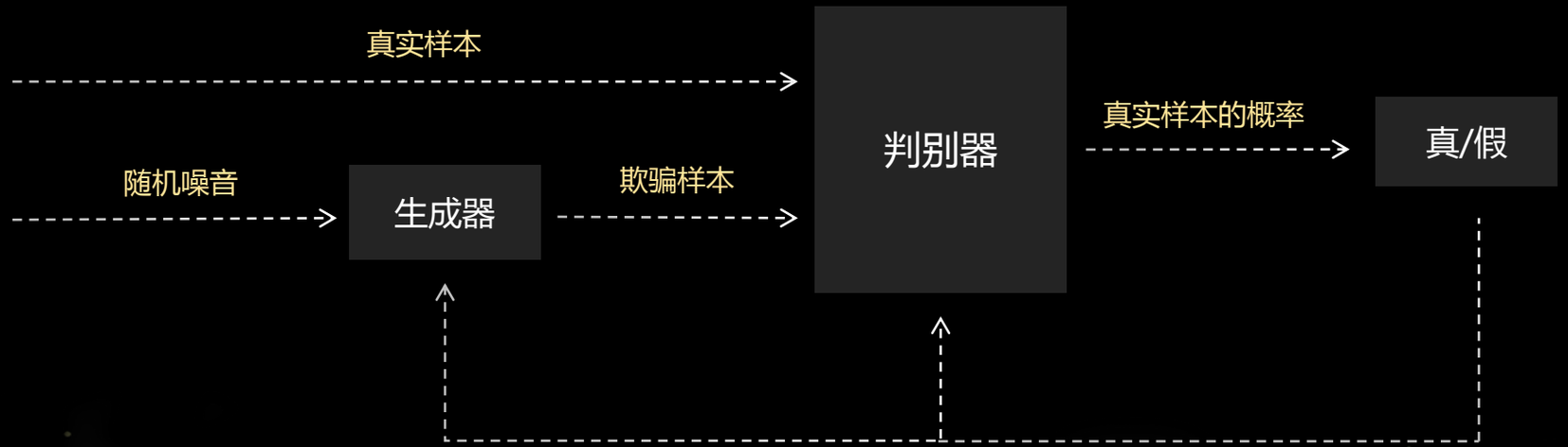
对抗学习是什么

生成式对抗网络

2014年由Goodfellow提出

DCGAN、AC-GAN、GAN-CLS和MalGAN(网络结构各异)

目标：判别器无法辨别真实样本与欺骗样本，即输出概率为0.5



假设真实样本的分布为 $p_{data}(x)$，生成器生成的样本分布为 $p_{model}(x)$

生成器的参数为 $\theta^{(G)}$，判别器的参数为 $\theta^{(D)}$，对抗网络的训练过程可以描述为一个极小化极大问题

$$arg \min_G \max_D (V(\theta^{(G)}, \theta^{(D)}))$$

$$V(\theta^{(G)}, \theta^{(D)}) = E_{x \sim p_{data}} \log D(x) + E_{x \sim p_{model}} \log(1 - D(x))$$

理论上收敛，得到一个全局最优解 $p_{model} = p_{data}$

主要内容

1、安全数据分析中面临的问题

2、对抗学习

3、对抗学习在DGA检测上的应用

4、总结与未来工作

为什么是DGA

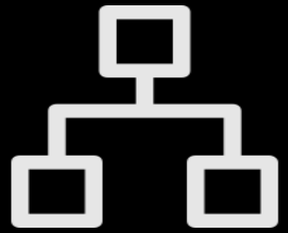
DGA(Domain generation algorithms)域名是一种通过算法自动生成的随机域名

1pcx96minkw591r8z9uo1jm1u7m.org
auwsleasuredehydratorysagp.com
kwmzazgwfqfpbhkprc.com
flswardenslavetusul.com
1hgv0n3qzeqpn1lrfgbmgz6i6d.biz
1i7yafmy0uqh0id43ipnlnzx1.net
dkm6lo2jzpkgeg6cwv8jdlu.org
1jkki72ti7gb51rx7u4xx9dsow.net
gkygvinskycattederifg.com
……

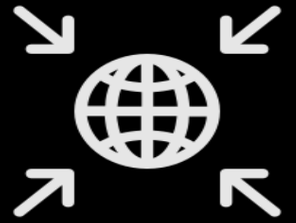
需要解决的问题

- 域名向量化

- 对抗网络的结构



- 对抗网络的训练



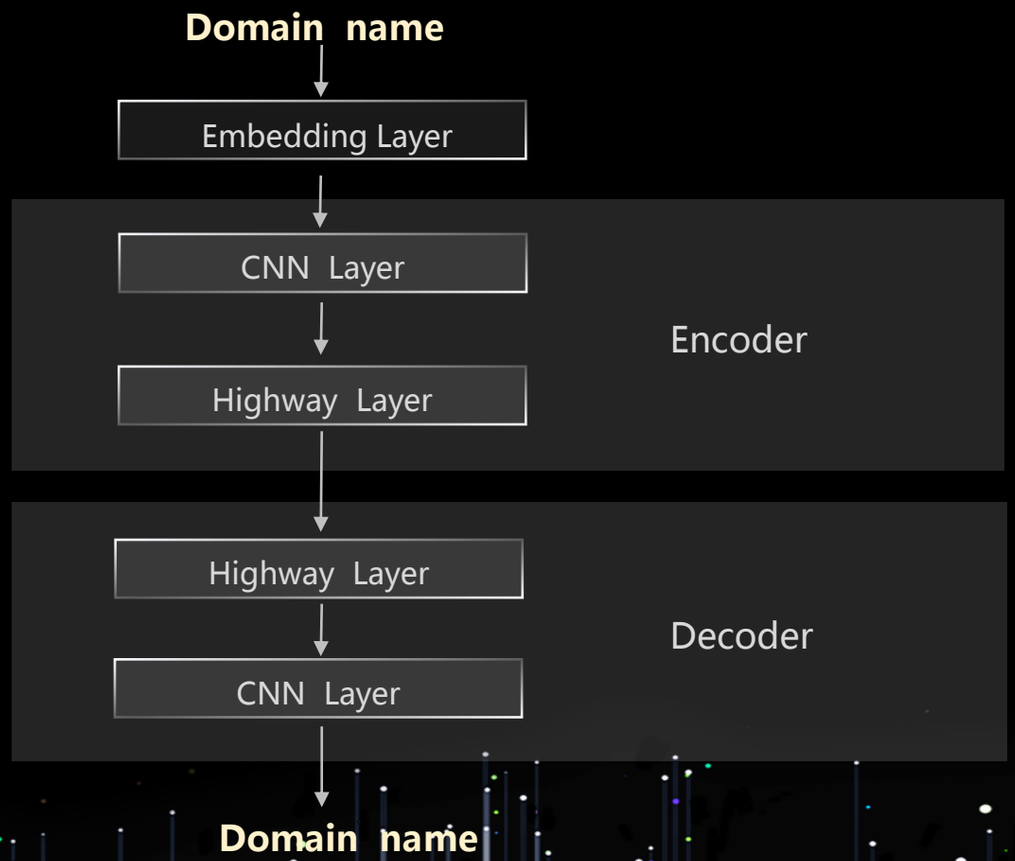
域名向量化-自编码器

- **Highway Network**

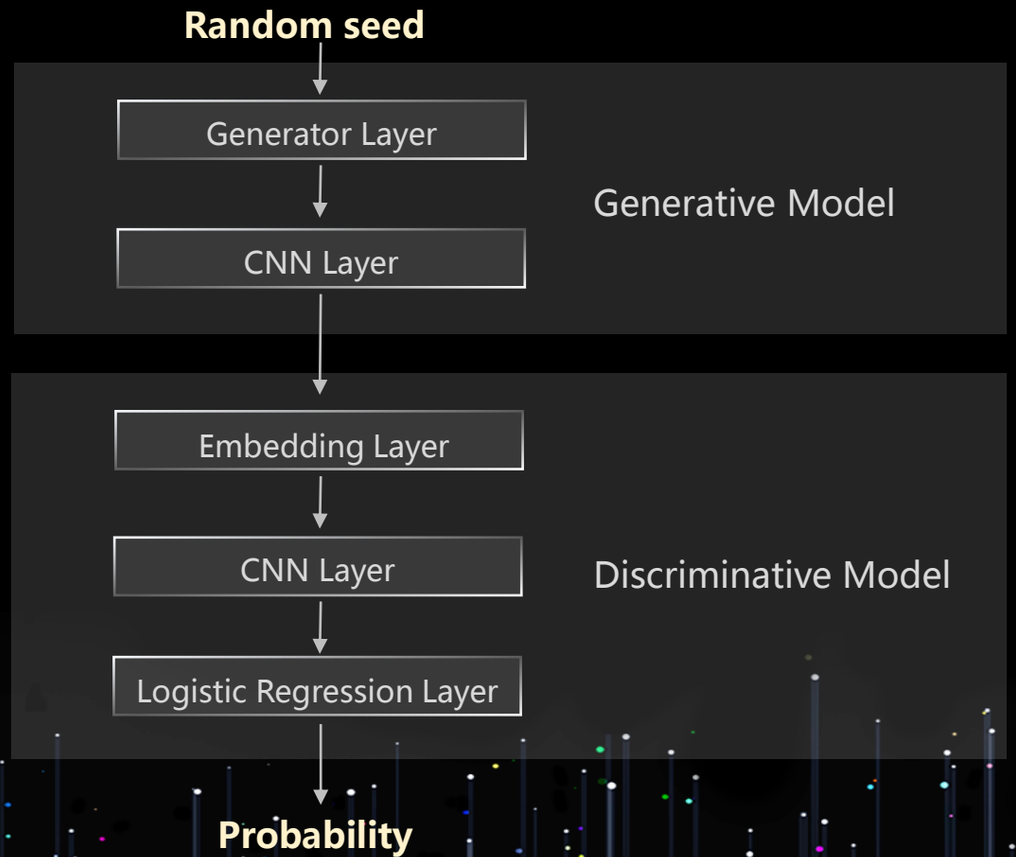
对于输入x, Highway Network的输出y为

$$y = t \cdot f(\mathbf{W}x + b) + (1 - t) \cdot x$$

其中，f为激活函数，$t \in [0,1]^d$



对抗网络结构



- **Generator Layer**

  Dense

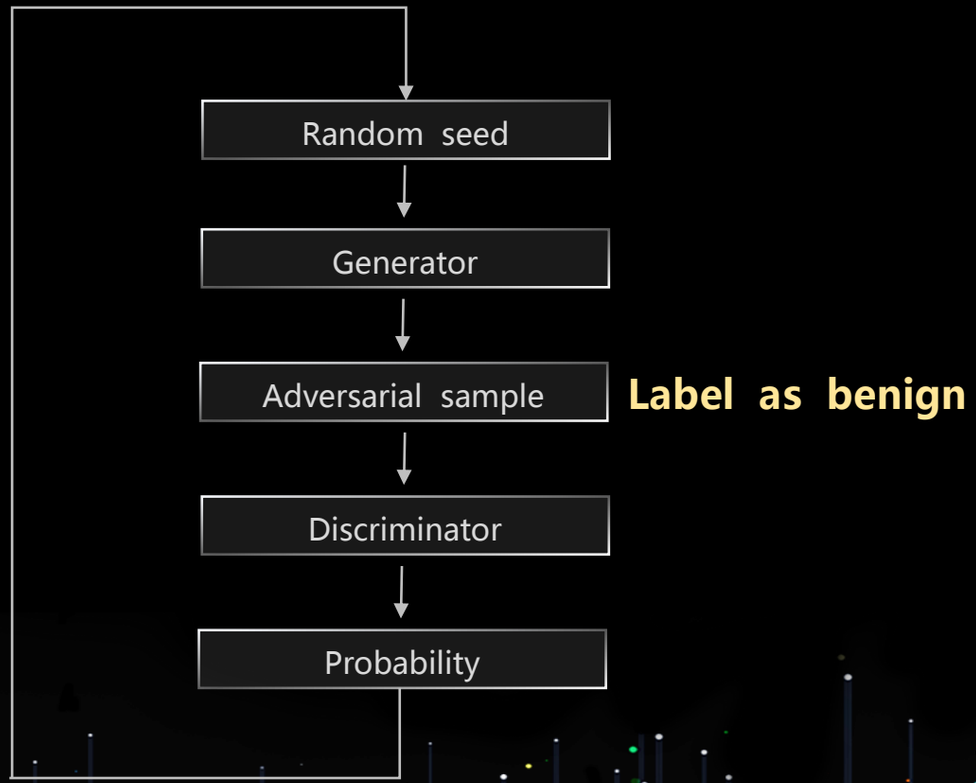
- **Embedding Layer**

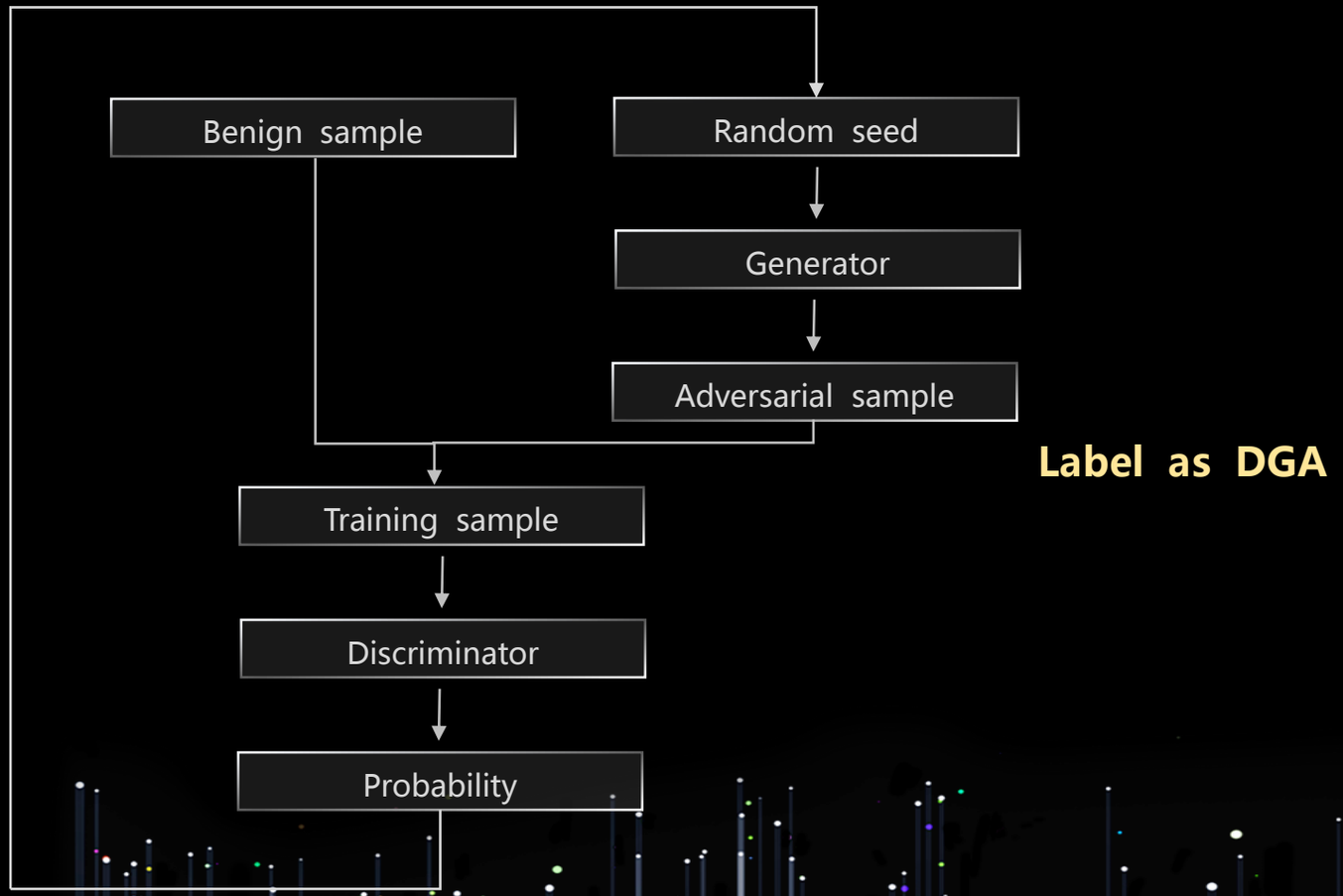
  AutoEncoder

- **Logistic Regression Layer**

  Dense (softmax)

对抗网络的训练


Random seed
Generator
Adversarial sample
Discriminator
Probability
Label as benign


Benign sample
Random seed
Generator
Adversarial sample
Label as DGA
Training sample
Discriminator
Probability

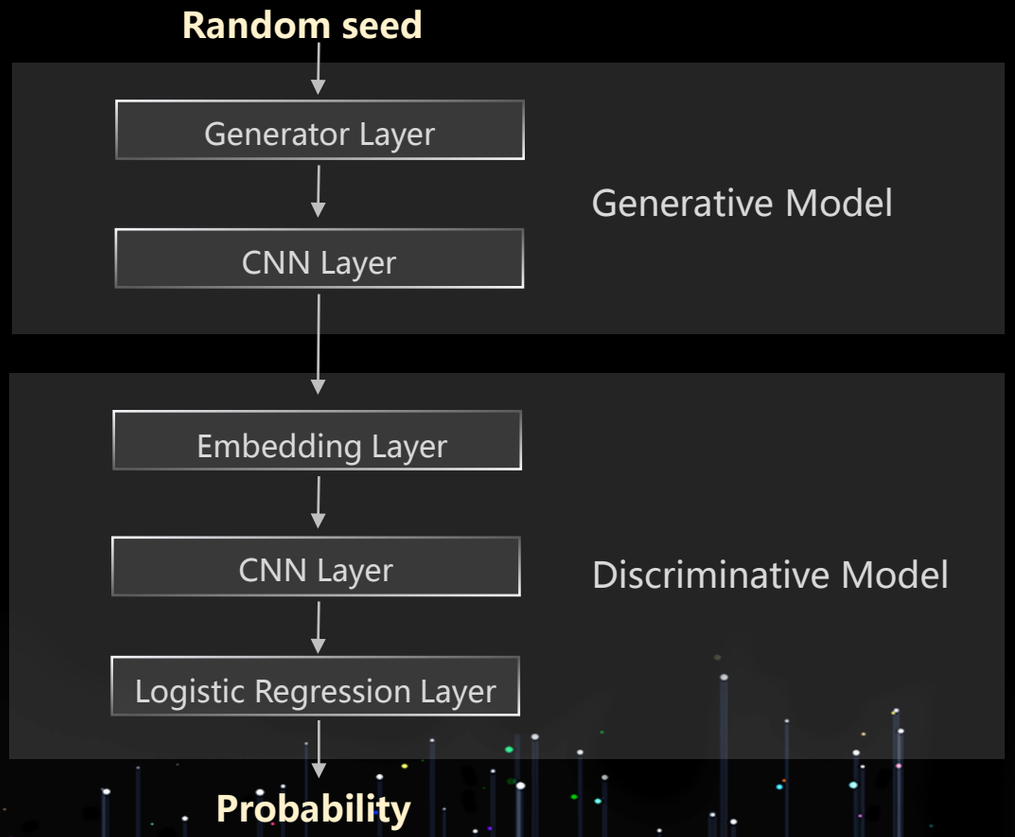
自编码器结果

输入	输出
0123moviescom	0224Qoviescom
m73lae5cpmgrv38com	m83lae5cpmgrv38com
w3schoolcomcn	w3svhoolcomcn
p30downloadcom	a40downloadcom
static1squarespacecom	static2squarespacecom
v2profitcom	b2profitcom
piratebay247net	piratebay248net

很不稳定的对抗网络

fhmgnzfpksffsphgkggdksdfudeeloddvQusdhglzcgn
fhmgnzfpksffsphgkgWgdksdfudeelodchdvQushglzcgn
fhmgfpksffsphgkgWgdksdfudeelodcmvQusdhglzcgn
fhmgnzfpksffsphgkgWgdksvdfudeelodchdvQusdhglzcrn
fhmgfpksffsphgkgWgdksvdfudeelodchdvQusdhglzcon
.....

对抗网络结构和训练调整



- Generator Layer

对抗网络结构和训练调整

生成器预训练

样本：随机噪声

标签：合法域名的独热编码向量

判别器预训练

样本：DGA域名和合法域名

标签：域名标签

调整后的对抗网络

5rpkvlgaakaceredawamookuelasmrcuc9oi
bdukly0dfuttrfefsv.ebmeoswvgrlpcsrome
csdmelv0eftfeptasoreu4asagnliesppee5uee
kta2zytq8l5ousammoponebetudsdmpmneca
ita6mbnsa4upkr2yocimaegauiybse6trttle
asnc9ipernuki.ar5eea20oaceulfemdtelrn
cor2sy8edli.0td2.xlinbobok4rguaxl7ie1tce
C3obtimiibtuzdsesp0atrlrj8hiqsakpao
Swh6fwt828aeesyfcxjgpudu6cppat1gan
.....

模型验证

分类器模型

- 逻辑回归算法
- 10万DGA，100万合法域名
- 特征：N-Gram（3）

测试集

- 10万DGA
- 10万对抗样本（识别率： 10^{-4} ）
- 实际采集的DNS数据（提升50%）

训练数据集	准确率
10万DGA，100万合法域名	0.8851
10万DGA复制9次，100万合法域名	0.8928
10万DGA，90万对抗样本，100万合法域名	0.9137

模型验证

分类器模型

- 逻辑回归算法
- 1万DGA，100万合法域名
- 特征：N-Gram（3）

测试集

- 10万DGA

训练数据集	准确率
1万DGA，100万合法域名	0.6942
1万DGA复制99次，100万合法域名	0.7531
1万DGA，99万对抗样本，100万合法域名	0.7685

主要内容

1、安全数据分析中面临的问题

2、对抗学习

3、对抗学习在DGA检测上的应用

4、总结与未来工作

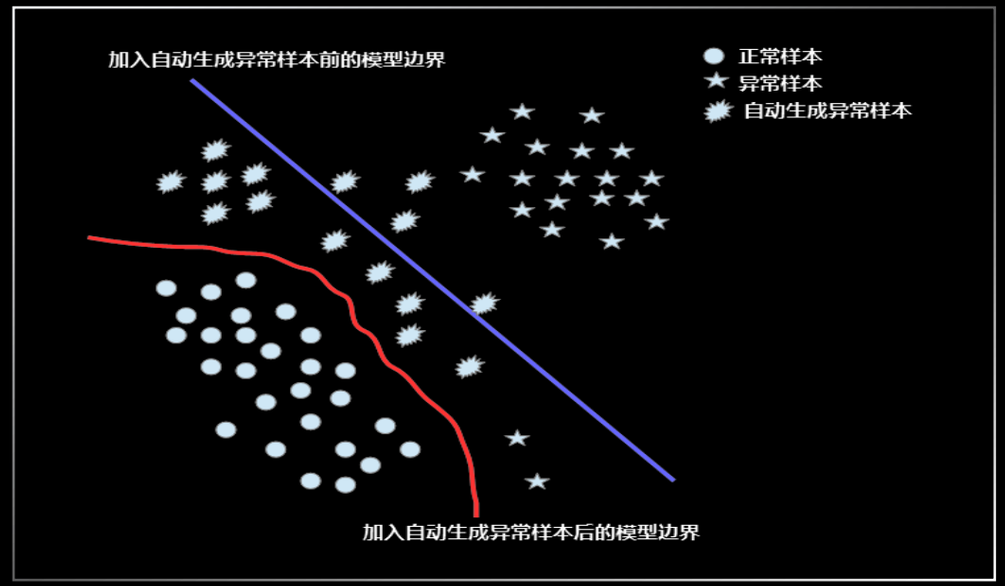
我们还需要努力

应用前提

- 2类样本中，某一类样本要足够多

未来工作

- 如何保证对抗样本是足够的
- 如何高效地生成对抗样本



THANK YOU