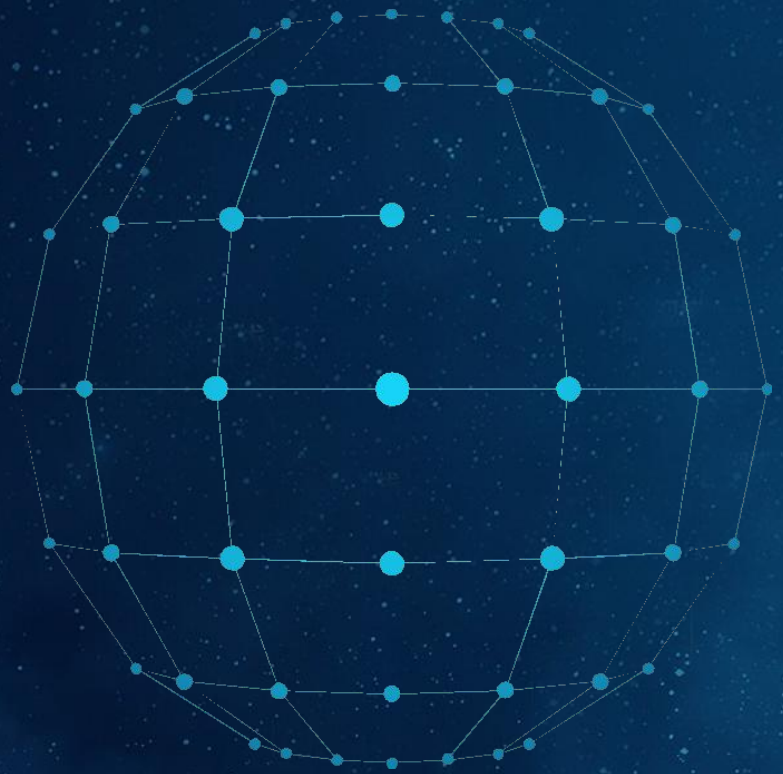

Ctrip
携程
2019
携程信息安全沙龙
畅谈——让安全无边界
Manufacturing
Supply chain
Product
Cargo
Customer
Delivery
Inventory
Management
Freight
Innovation
Branding
Solution
Marketing
Analysis
Ideas
Success
Management
22.02.35.2
80.08.1
SOCIAL NETWORK



统一内容检测服务

王乐 / 携程信息安全部/高级业务安全工程师

CONTENTS

CONTENTS

Ctrip 携程

背景

统一内容检测服务

- 以避免网站对外展示内容出现违规为出发点，而诞生的产品

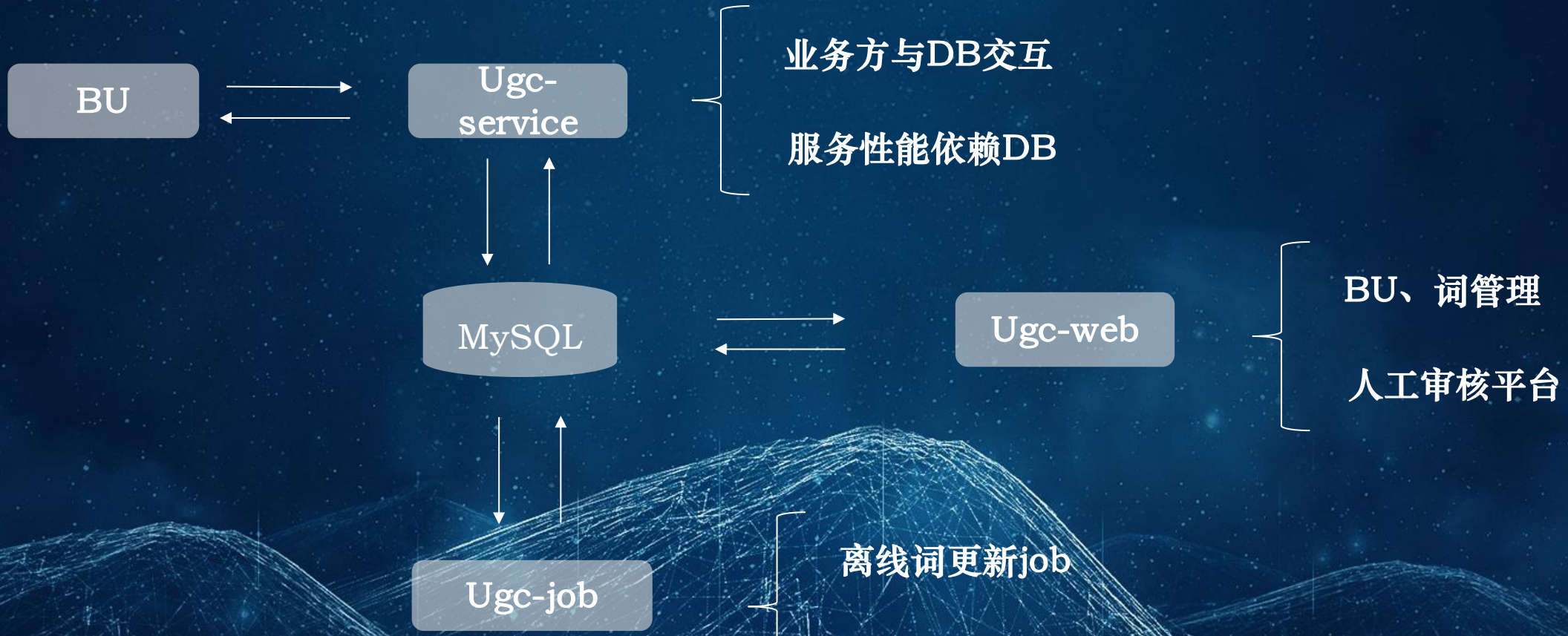
违规问题点

- 违反法规：黄赌毒、反动、违禁品等

Ctrip 携程

背景

UGC 1.0(旧服务)


BU
Ugc-service
业务方与DB交互
服务性能依赖DB
MySQL
Ugc-web
BU、词管理
人工审核平台
Ugc-job
离线词更新job
离线敏感词过滤job

背景

使用面临的问题

- 关键词覆盖率，维护困难
- 难以应对变形文字，抗符号干扰能力弱
- 误命中率较高

业务新需求

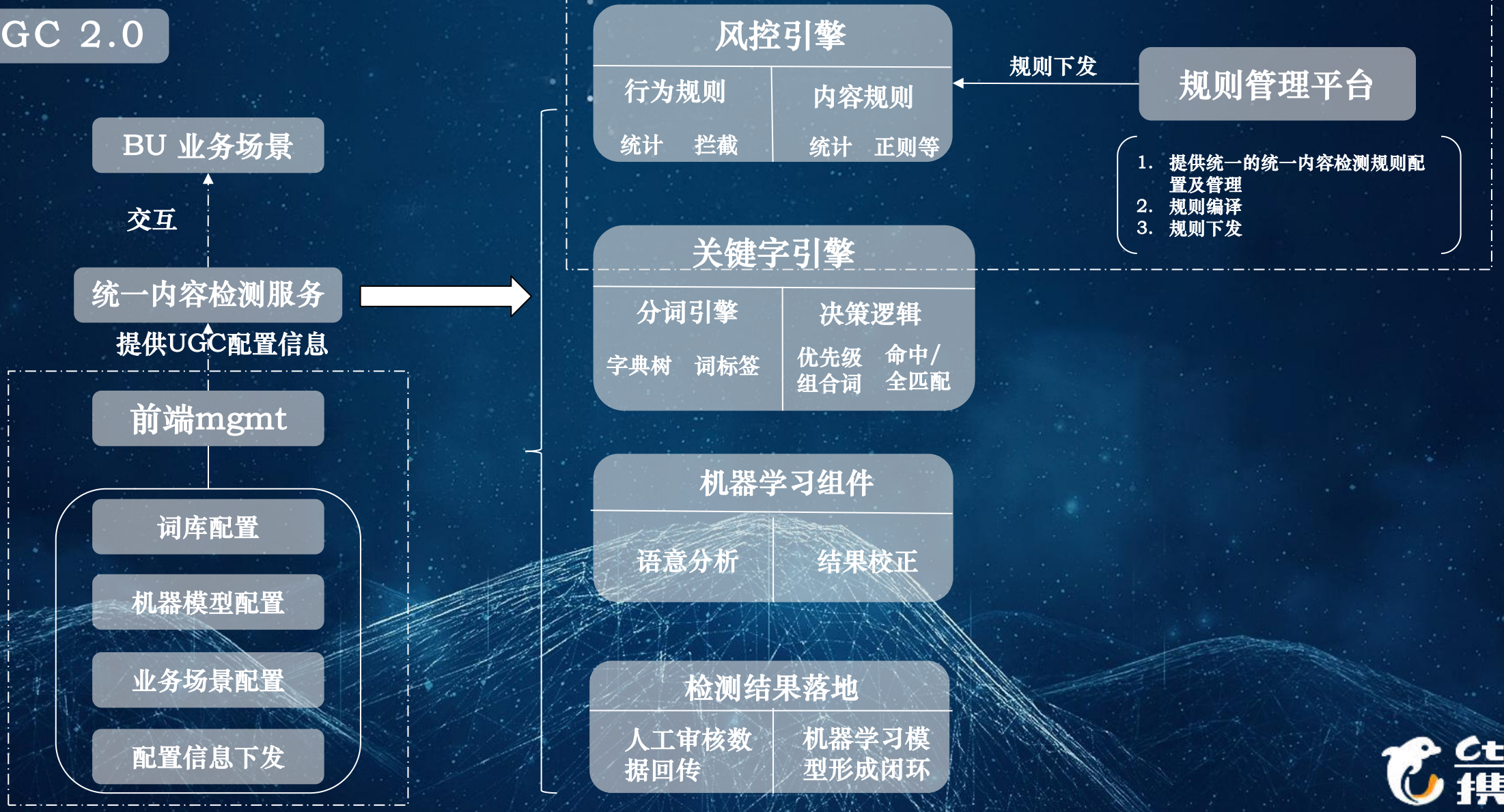
- 控制发布频率，定制化关键词返回，黑白名单

CONTENTS

Ctrip 携程

提升

UGC 2.0


BU 业务场景
交互
统一内容检测服务
提供UGC配置信息
前端mgmt
词库配置
机器模型配置
业务场景配置
配置信息下发
风控引擎
行为规则
统计 拦截
内容规则
统计 正则等
规则下发
规则管理平台
1. 提供统一的统一内容检测规则配置及管理
2. 规则编译
3. 规则下发
关键字引擎
分词引擎
字典树 词标签
决策逻辑
优先级组合词 命中/全匹配
机器学习组件
语意分析
结果校正
检测结果落地
人工审核数据回传
机器学习模型形成闭环

提升

实时风控

行为分析，多维度统计，
正则匹配

提升

关键词引擎

区分业务场景，关键词优先级，
全匹配打码，简繁体转换

提升

机器学习组件

特征判断，语义分析，
结果校正

CONTENTS

Ctrip 携程

模型探索

模型解决的问题

- 对命中敏感词的文本数据使用深度学习模型自动识别该内容是否符合规范。

入手方向

- 文本判别本质上是NLP(Natural Language Processing)中短文本二分类的问题，即根据短文本中语义的特征判断类别

语义识别：

- 情感识别：根据情感正负向进行二分类
- 内容识别：根据内容判断所属类别进行多分类

模型探索

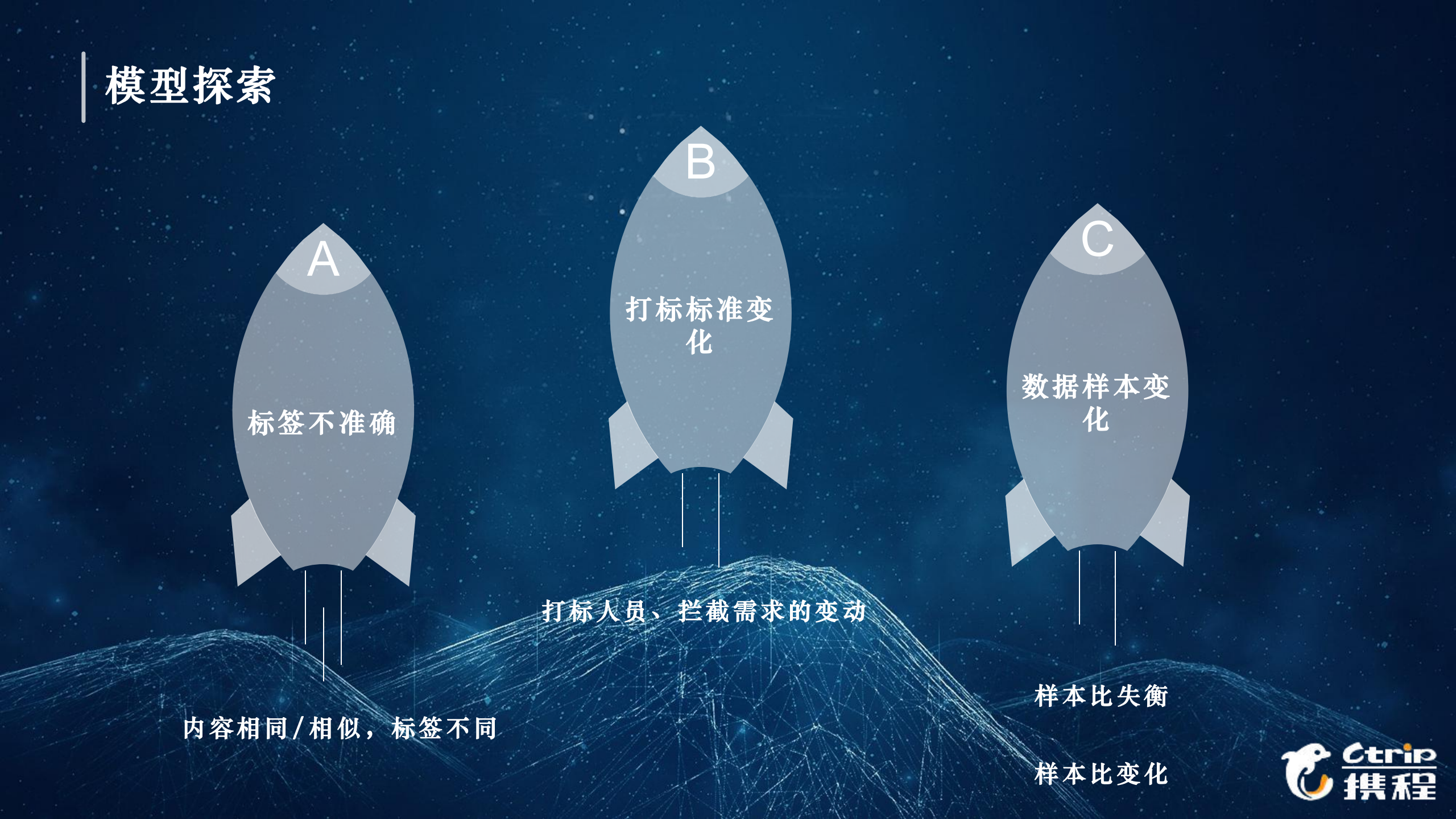
数据源问题：

- **打标标准复杂**

传统的文本识别是根据文本语义判断情感的正负向，实际文本打标标准按照业务和违禁关键词的规则集合

负向情感	景色不好，服务也很差，下次再也不来了。	✓
	傻*商家，他*的就知道骗钱，诅咒他******	✗
正向情感	清纯小姐姐在线发牌，给您至尊体验，XXX娱乐城。	✗

模型探索


A
标签不准确
内容相同/相似，标签不同
B
打标标准变化
打标人员、拦截需求的变动
C
数据样本变化
样本比失衡
样本比变化

模型探索

数据预处理：

- 标签扶正
  - 文本相同标签不同
  - 人工复核
- 数据清洗
  - 标点统一保留(英文版感叹号、问号、句号、逗号)、外文过滤、重复标点去重、emoji表情过滤，乱码过滤
- 词向量
  - 自训练词向量
- 规则清洗
  - 切客
  - 联系方式等

模型探索

规则清洗：

关键词	占总量	通过	不通过	主要判不通过原因
谢,您,网站	19.57%	53.72%	46.28%	联系电话、人名
优惠价	14.55%	53.73%	46.27%	乱字符、切客a
携程	10.18%	97.91%	2.09%	语言攻击、切客
携程,酒店	4.01%	95.15%	4.85%	语言攻击、切客
酒店,携程	3.45%	95.13%	4.87%	语言攻击、切客、联系电话

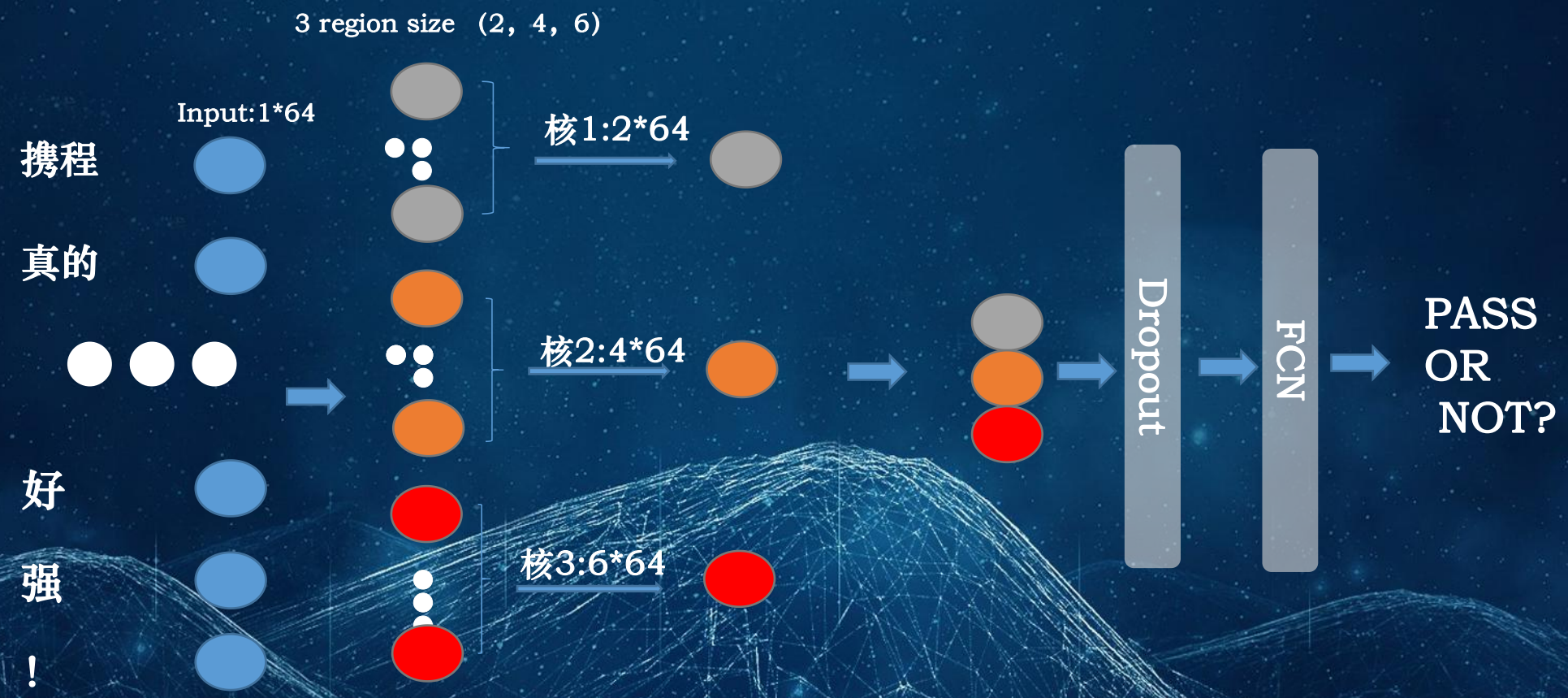
关键词	占比	通过	不通过	主要判不通过原因
政府、警察	11.31%	96.98%	3.02%	联系电话、切客
开,发票	6.52%	97.31%	2.69%	联系电话、人名
服务,热线	3.95%	40.72%	59.28%	联系电话、人名
人名	3.34%	94.80%	5.20%	人名、乱字符
提供,发票	3.15%	96.44%	3.56%	联系电话、人名

- 剔除可提取统一标准的违规数据（联系方式、切客）
- 无统一标准的数据（开发票、语言攻击），需要根据模型识别

Ctrip 携程

模型探索

模型1:TextCNN



模型探索

模型2:LSTM

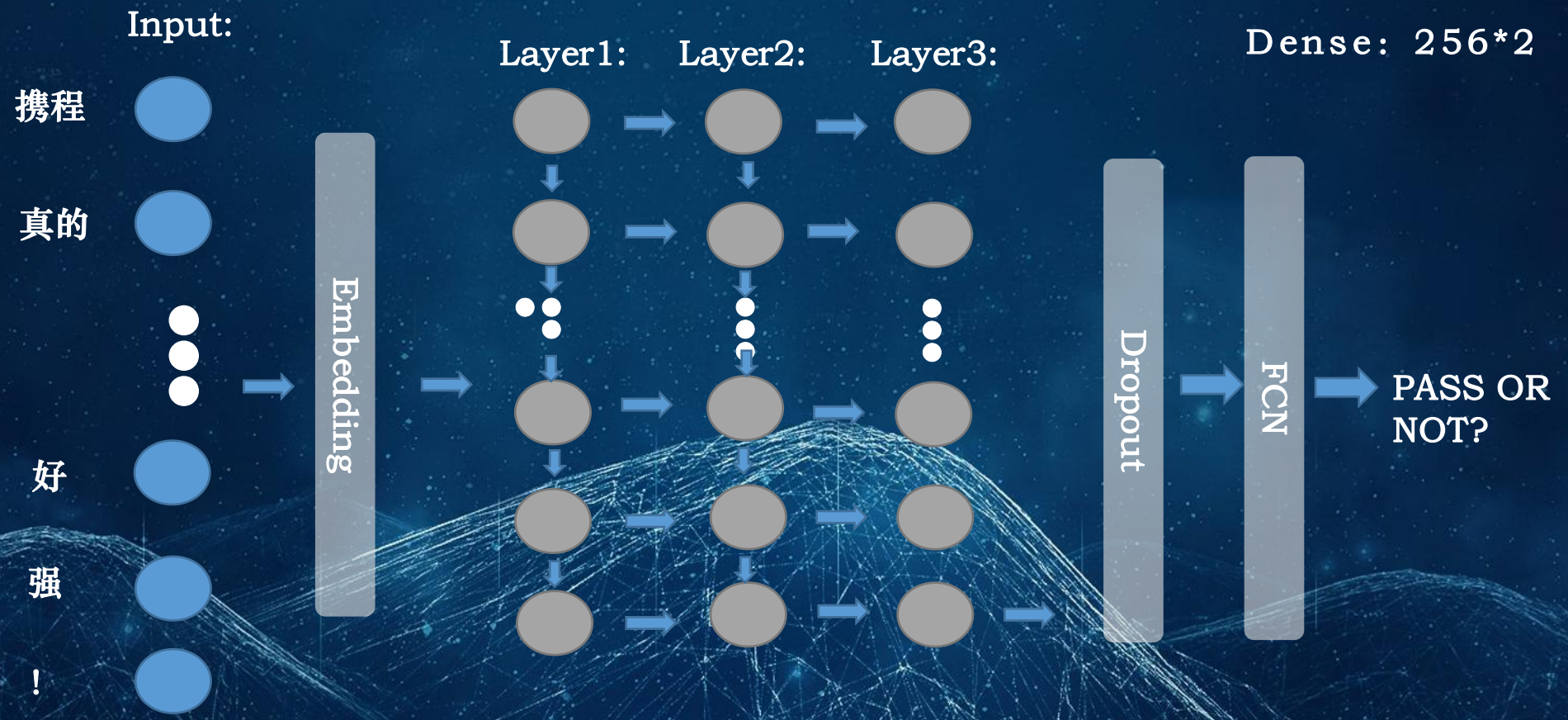
Input: 1*64

Layer1: (64+256)*256

Layer2: (64+256)*256

Layer3: (64+256)*256

Dense: 256*2



模型探索

模型对比

	textCNN	Lstm
特点	提取序列中类似n-gram的局部特征，特征间独立。适合分类任务。	提取序列中上下文的整体特征，适合语义分析任务。
效果对比	注重句子中的局部特征相关性。漏报率较低，误报率较高	注重整体句意。误报率较低，漏报率较高

TextCNN		LSTM		Textcnn + LSTM	
漏报率	误报率	漏报率	误报率	漏报率	误报率
0.85%	12.70%	1.20%	6.10%	0.51%	15.60%

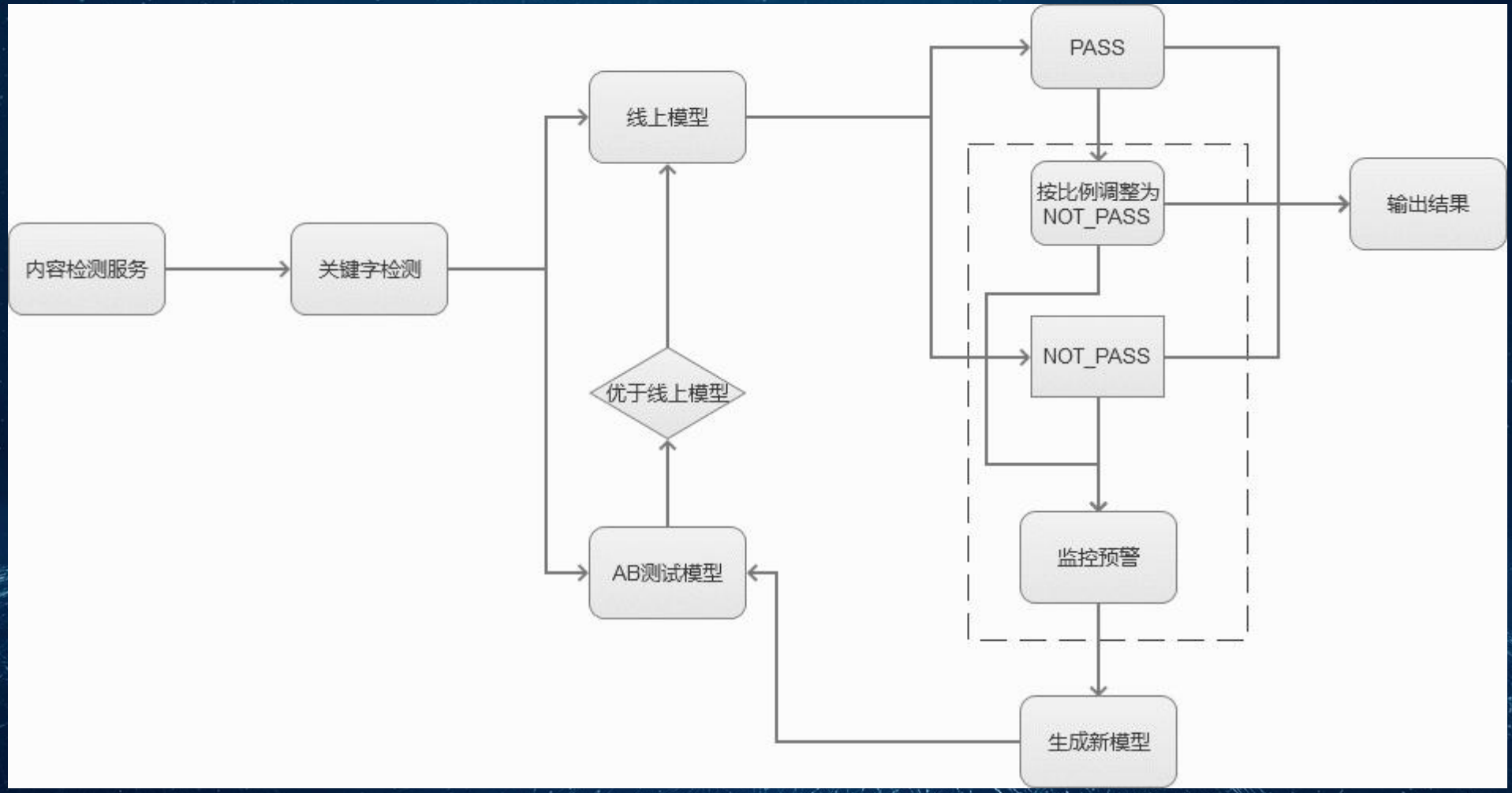
模型探索

部署对比

	django The web framework for perfectionists with deadlines.	TensorFlow Java
优点	Django+python部署，模型载入预测简单，精度无损失	Tensorflow+java部署，并发强，实时性高
缺点	并发差，时延长，难以满足实时性要求	模型精度略微有损失

模型探索

模型迭代：


内容检测服务
关键字检测
线上模型
AB测试模型
优于线上模型
PASS
按比例调整为
NOT_PASS
NOT_PASS
输出结果
监控预警
生成新模型

现场提问

扫码发送暗号

“2019”

即可加入交流群

扫码关注

携程安全应急响应中心

公众号

Ctrip 携程

Thanks

主办方：携程信息安全部