51CTO

WOTI

# 机器学习在电商领域的场景化应用

## “商品数据多维度优化”


**郑志彤**

京东商城基础平台部首席研究员

51CTO

WOTD

World Of Tech

2017年12月1-2日

# 全球软件开发技术峰会

[ 深圳站 ]

报名咨询： 010-68478816
议题提交： wot@51cto.com
市场合作： yangxh@51cto.com
商务合作： songjc@51cto.com
媒体合作： yankk@51cto.com
在线咨询（微信）：18401576051

团·购·享·受·更·多·优·惠

**5折** 优惠（截止8月31日）
现在报名，立省1400元 / 张

# 电商商品数据 – Multi-Modal & Noisy

**现状**：

- 商品数据录入难以有效监管，数据质量参差不齐
- 用户反馈数据没有得到有效使用



**目标**：

- 对于商家录入的商品数据进行清洗，提升数据准确率
- 对于原先没有得到有效利用的数据，整合抽取
- 为商家生态提供算法支持，从源头把控商品数据质量

# 目录

- 电商数据的**信息合规**
- 商品基本属性优化：**图文不一致校验**
- 电商短文本理解：**商品标题属性理解与重组**
- 智慧生态：**商品类目的自动识别**
- 富集商品数据：**多维度知识抽取**
- **构建商品知识图谱**

# 信息合规

## 违禁词理解

*Q：为何秒针对不准刻度？*

A：这在石英表里是普遍存在的现象，并非质量问题，石英表的步进马达本来就很难稳定推动秒针前进固定的间隔，这不会影响手表的计时精确度。卡西欧手表的精确度为+-20秒/月，已经是手表行业的最高质量标准。

*Q：我的日期为什么不走？调节方法？*

A：很多客户会存在一个12小时时间差的问题，如果日期不跳，就按照下列步骤重新调节一次：

第一步：把星期和日期都调前一天(如：今天3号星期五 就先调到2号星期四)；

第二步：来开表把两档，往时间增大的方向调节，一直到12点附近，日期和星期都跳了一格。这个时候代表凌晨时间！如果上午10点 就要往前调节到10点就OK了，如果是下午14:00，那就要先调节到中午12点，然后再往前调节2个小时到14:00。

**上线效果：**

无效审核下降73%：

漏掉率为7.2%：

**More！ 通用信息合规服务：**

**价格敏感词识别　　广告用语识别**

**联系方式识别　　低俗用语识别**

# 信息合规

## 自营店铺价格合规

**每天自动审核发现数千价格违规现象**


【京东超市】[redacted] 床品套件 高支高密英式田园 全棉斜纹床单四件套 凡尔赛花园 1.8米床

618 全民年中购物节

京 价 ¥249.00 降价通知

累计评价 1900+

促 销 满减 此价格不与营销优惠同时享受

配 送 至 北京大兴区亦庄经济开发区 有货

支持 99元免基础运费(50kg内) 货到付款 京准达

由 京东 发货 [redacted] 提供售后服务。15:00前下单,可预约今晚送达

重 量 2.47kg

选择颜色 凡尔赛花园 机洗 无荧光 实拍图 尤尼斯-咖 波斯卡纳

选择尺码 1.5/1.8米床 被罩200x230cm 1.8/2.0米床 被罩220*240cm

白条分期 不分期 ¥84.25×3期 ¥42.75×6期 ¥22×12期 ¥11.62×24期

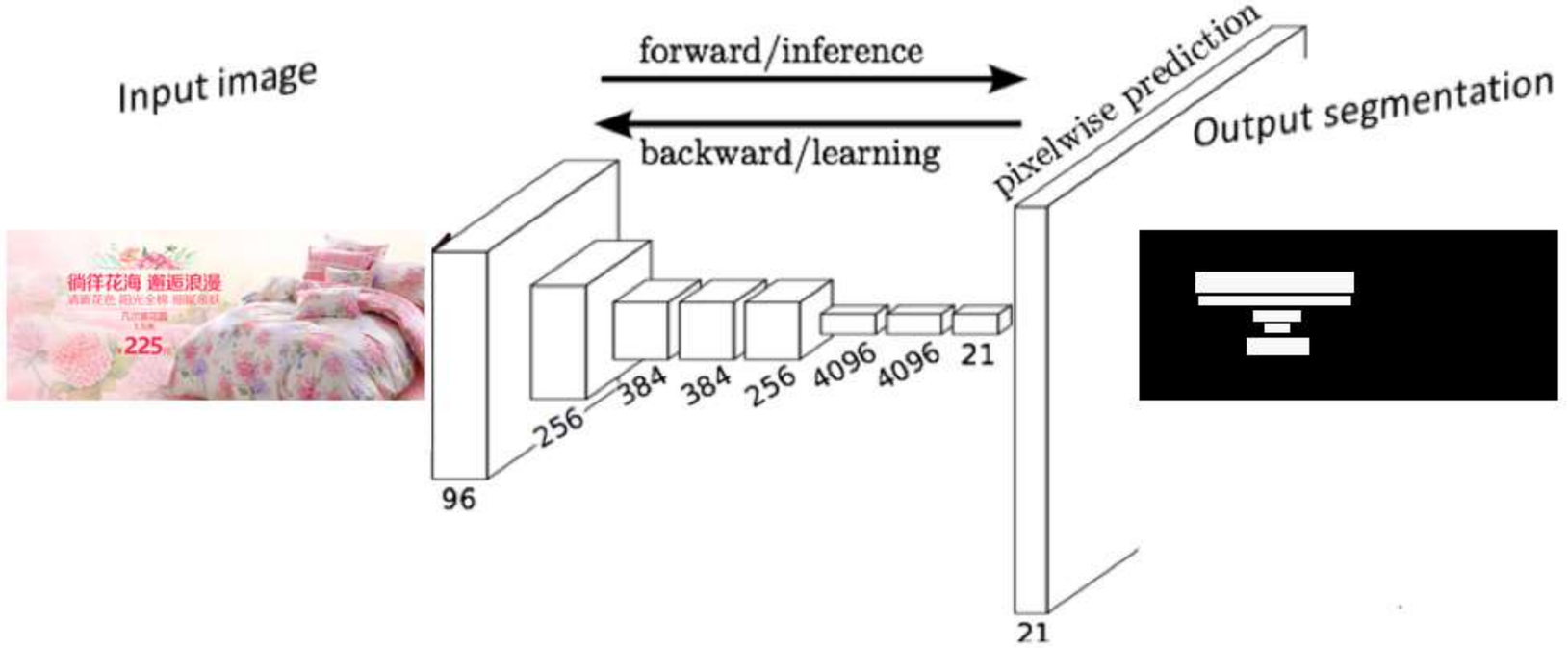
关注 分享 举报

# 信息合规

## 自营店铺价格合规

### 基于全卷积网络的文本行定位

由于商品详情图片的版面复杂，风格差异较大，传统的基于启发式规则的版面分析方法是行不通的。我们采用在图像语义分割领域十分有效的深度学习模型-全卷积网络（Fully Convolutional Networks），来实现端到端的可学习的图片版面分析和文字定位算法，达到区分文字和背景区域的目标，如下图所示。



# 信息合规

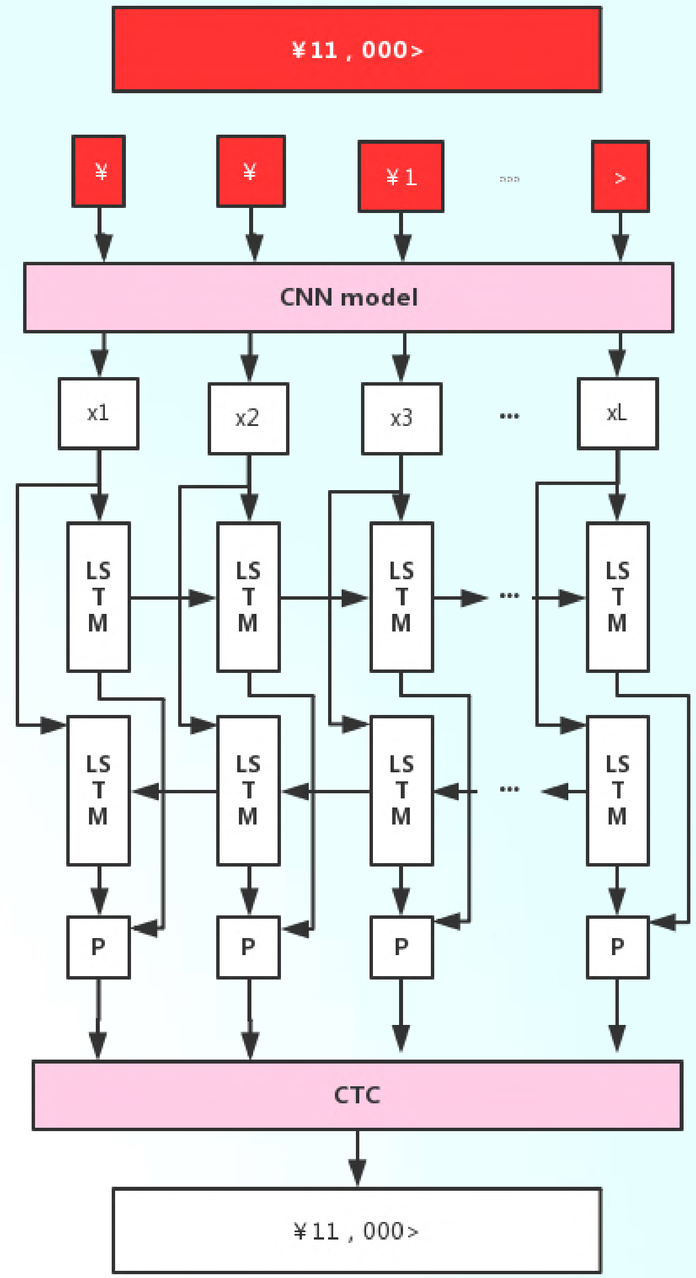
## 自营店铺价格合规

### 端到端的通用字符串识别系统

如图所示，通过CNN model获得图片的特征，与基于大规模语料数据训练循环神经网络（LSTM）的通用语言模型相结合，再通过基于时序分类（CTC）输出。

端到端的文本检测与识别算法克服了传统OCR鲁棒性不足的问题，即使对于京东网站上各种压缩失真和版面复杂的图片，也能得到很好的文字识别结果。

目前每天可以自动发现数千个价格不一致现象。同时图片文字识别出的语句在通过文本合规服务后，自动发现包含违禁语义的图片，净化京东生态

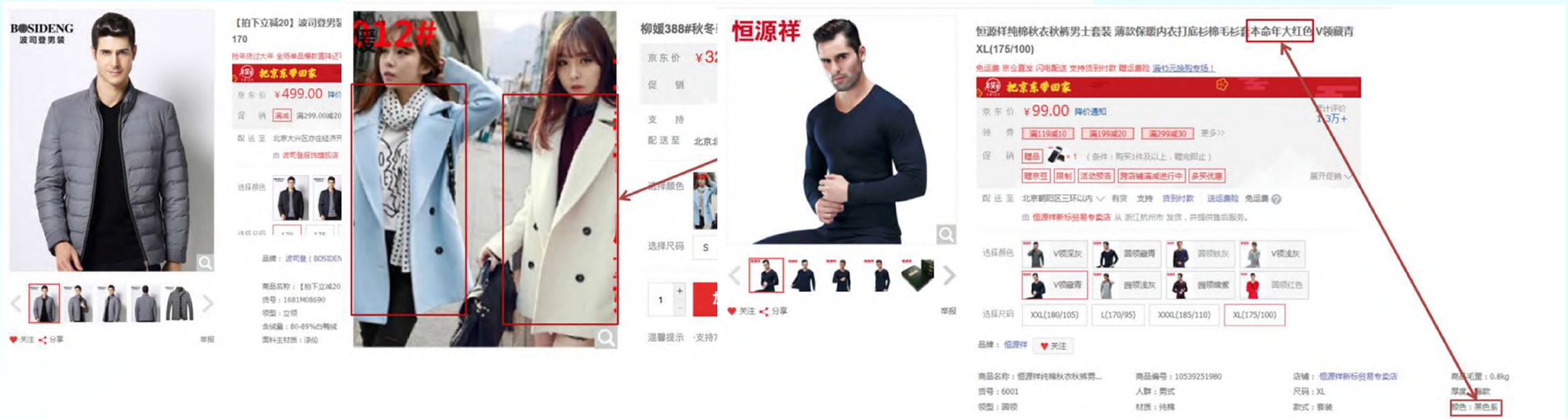


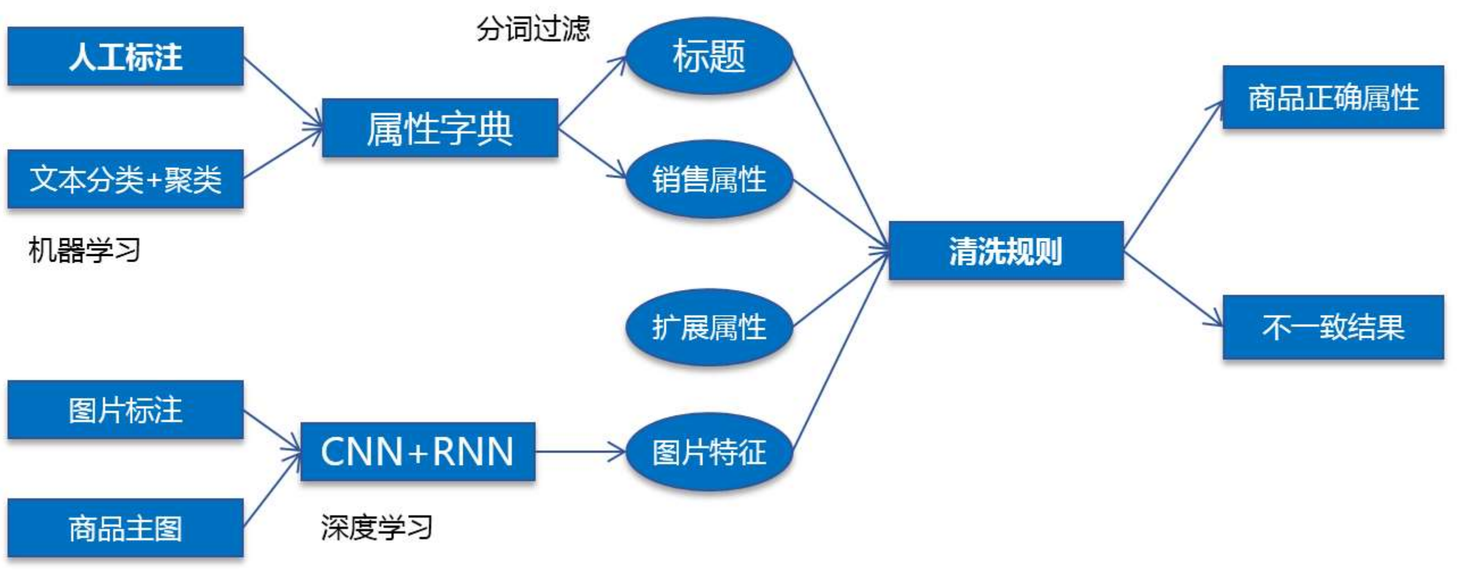
# 商品基本属性优化：**图文不一致校验**

# 图文不一致校验

## 商品基本属性优化

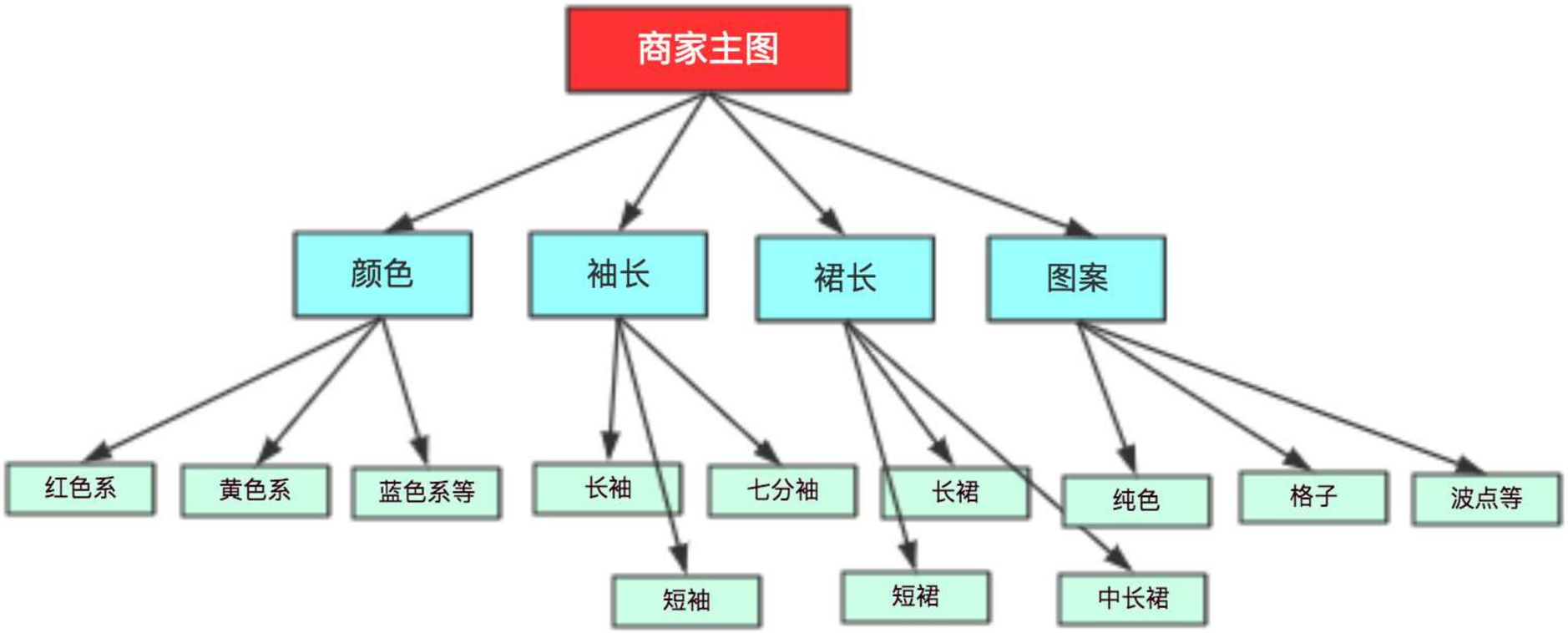
- 属性间的不一致对上层系统影响巨大，搜索、推荐调用错误数据，结果随之错误



# 图文不一致校验

## 商品基本属性优化



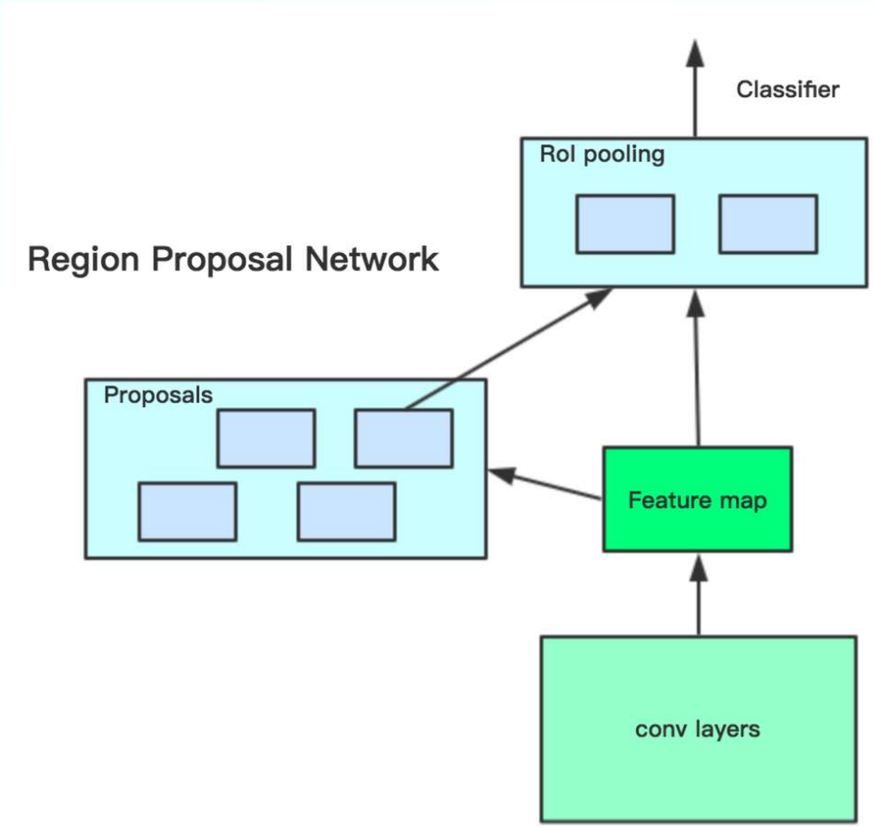

- 目前覆盖四个一级类目：运动户外、服饰内衣、鞋靴、礼品箱包
- 颜色识别覆盖61个三级品类，准确率95.65%
- 累计修正2亿条以上商品属性，1亿多条sku商品数

# 图文不一致校验

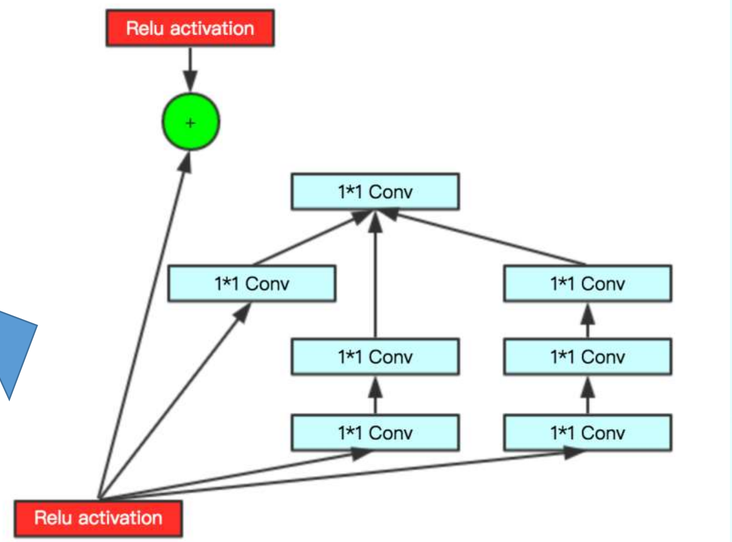
## 商品基本属性优化

对商城图片进行颜色检测时，如果将图片直接输入，去识别颜色的话，会导致很多问题。
比如一张图片里女model提着红色手包，穿着白色上衣，蓝色裤子，这种图片直接识别不能分别得到这三个主体的颜色分类。

**特征提取+ 主体颜色识别：**faster r-cnn
改进特征提取部分，加入reception和resnet结构以提高检测和分类准确率，实际过程中也提高了训练速度。




**reception和resnet结构**



# 电商短文本理解：**商品标题属性理解与重组**

## 商品标题堆砌、展示不全


弗洛米 iPhone7/7plus手机壳/保护套 苹果7plus超薄全包硅胶透...

¥34.50
3410条评价 97%好评　广告

邦克仕 (Benks)苹果 iPhone7 Plus手手机壳/保护套 苹果7手机套 磨...

¥25.00
自营 18420条评价 94%好评

亿色 (ESR) iPhone 7 plus手机壳 苹果7plus手机套 TPU 防摔软壳 ...

¥29.90
自营 29309条评价 94%好评

蒙奇奇 iPhone7手机壳手机套 防摔磨砂保护壳 适用于苹果 ipho...

¥32.00 商家免邮
京配 38998条评价 98%好评

138条评价 83%好评

**弗洛米 iPhone7/7plus手机壳/保护套 苹果7plus超薄全包硅胶透明电镀软壳 5.5英寸 炫亮黑☆炫亮电镀**


弗洛米 iPhone7/7plus手机壳/保护套 苹果7plus超薄全包硅胶透明电镀软壳 5.5英寸 宝石蓝☆炫亮电镀

【关注商品】送品牌钢化膜

¥ 34.50　降价通知


宾际运动鞋男鞋夏季男士网面跑步鞋女休闲板鞋透气韩版潮鞋情侣旅游大码网布鞋子 男 黑色 40

夏季新款透气网面运动鞋时尚跑步鞋，支持货到付款！！！

¥ 79.00　降价通知

**宾际运动鞋男鞋夏季男士网面跑步鞋女休闲板鞋透气韩版潮鞋情侣旅游大码网布鞋子 男 黑色 40**

# • 电商短文本理解

## 商品标题属性理解与重组

**搜索请求：iphone7手机壳**


**弗洛米 iPhone7/7plus手机壳/保护套 苹果7plus 超薄全包硅胶透明电镀软壳 5.5英寸 炫亮黑☆炫亮电镀**

↓ **分词**

**弗洛米 iPhone7/7plus 手机壳 / 保护套 苹果7plus 超薄 全包 硅胶 透明 电镀 软壳 5.5英寸 炫亮黑☆炫亮 电镀**

↓ **命名实体识别**

品牌词　产品词　产品词　产品词　产品词　风格　风格　材质　风格　工艺　款式　尺寸　颜色　工艺

↓ **短文本理解：依存分析**

**Head**：手机壳　**pure modifier**：超薄、透明、硅胶、炫亮黑　**constraint**：iphone7

↓ **应用： 标题重组**

**弗洛米 Iphone7 超薄电镀硅胶手机壳 炫亮黑**

# 智慧生态：**商品类目的自动识别**

# •类目自动识别

## 商品类目属性优化


**类目错绑严重：**

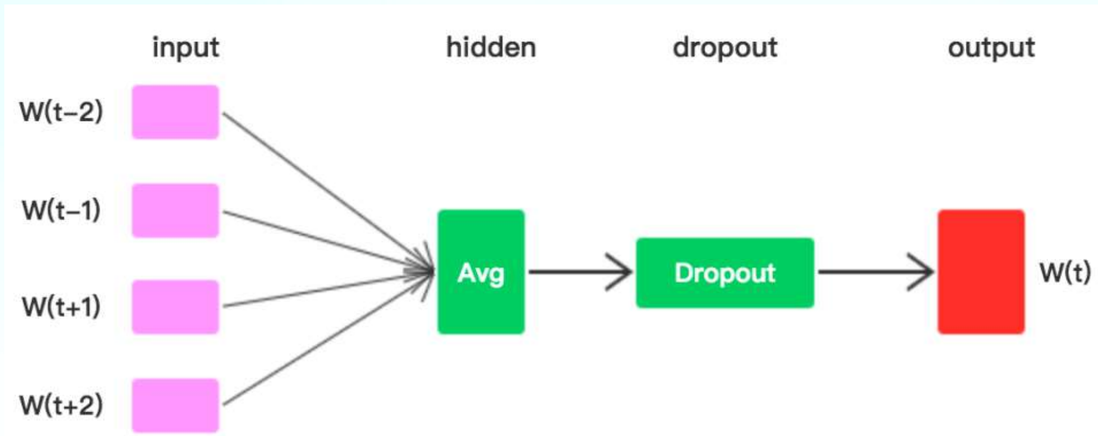
- 商品录入量大，难以管控：大型店铺sku数量数十万条
- 商品类目数多，精准录入难：三级分类数近4000条
- 主观理解商品类目划分错误：部分商品类目有重叠，难界定

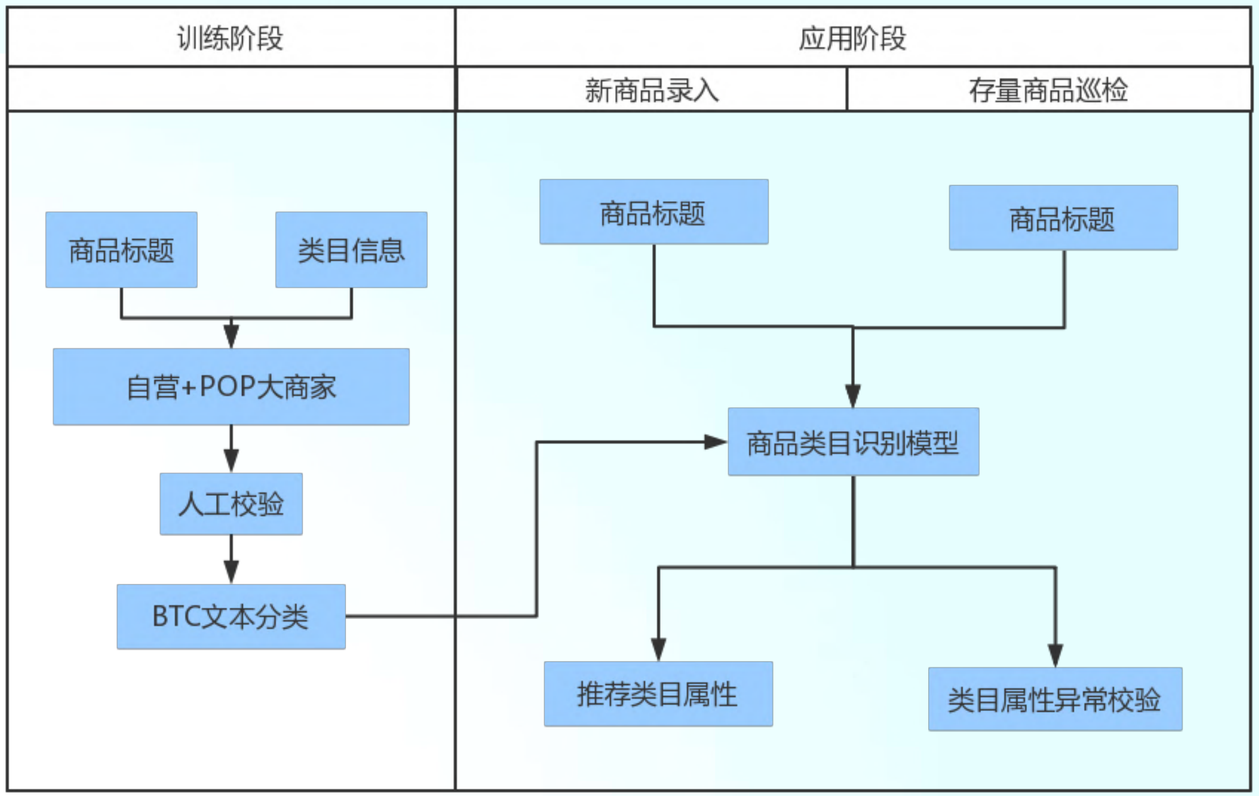
618大促前期，商品上新修正量每日达到5000万

# •类目自动识别

商品类目属性优化

## 构建模型，理解标题与类目之间的关系

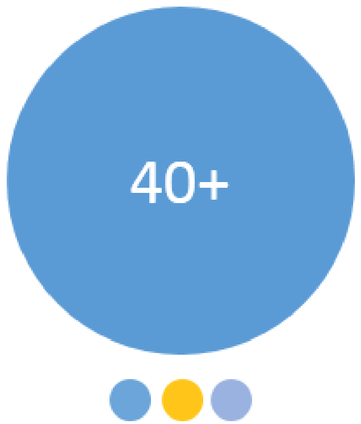



- 在多个试点一级类目实现分类准确率99%
- 修正上千万SKU类目错绑属性

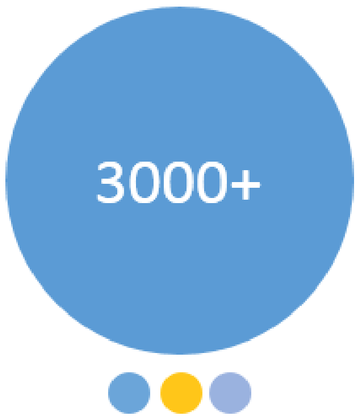
# •类目自动识别

商品类目属性优化




一级类目


二级类目




三级类目

- 部分类目间商品文本特征重合，
- 大量类目下商品数量千万级别，且具备细分条件

- **探索：利用机器学习来定义类目的合并与拆分**

# 富集商品数据：**多维度知识抽取**

# 多维度知识抽取

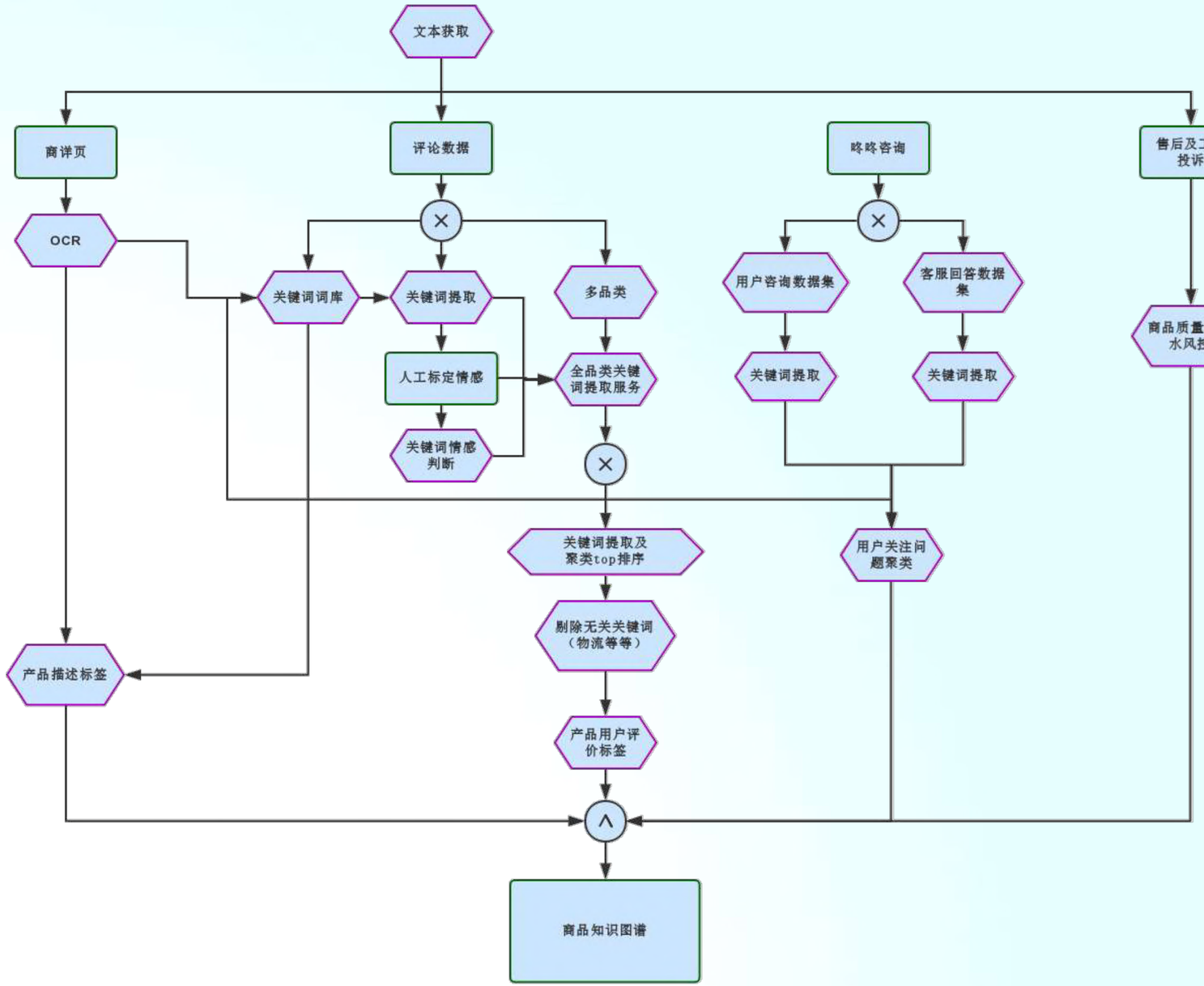
## 商品属性填充

**多场景文本信息获取，构建商品知识图式存储**

**商详页OCR**

**用户评论**

**客服聊天**



# 多维度知识抽取

## 商品属性填充

### 评论知识抽取

京东评价：

w***s

钻石会员

不怎么样，打死也不买了

500-600g 2015-08-01 21:58 举报 0 0

然而这五百元的差价，彻底被那个红铜色的锤子logo给征服了。当然还有nfc功能，我习惯这个功能了。

方方正正的形状，是我喜欢的风格。有人说，这个坚果pro是早产的t系列，完全认同。和手中的t1，风格上太接近了。来自于苹果4的经典三明治结构，双面玻璃，都是我喜欢的。

上一次发出感叹，是入手小米5，也是双面玻璃，尤其背面用了3d曲面屏。用过一段，发现背面尽管贴手，但不够硬朗，加之偶有发热，运行变慢，半年就停用了。

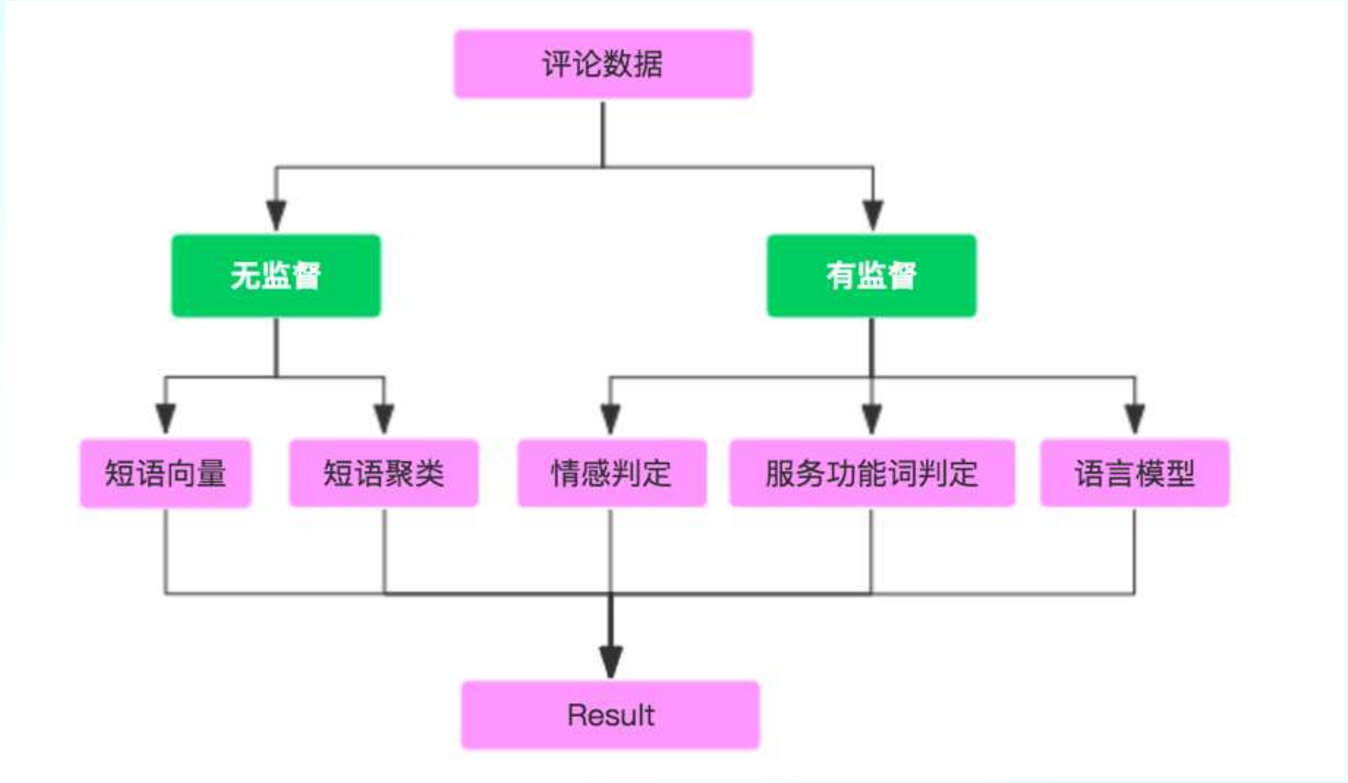

细红线特别版 128GB 官方标配 2017-05-11 11:43 举报 56 22

# 多维度知识抽取

## 商品属性填充

### 评论知识抽取

- 评论的情感分析
- 商品与评论相关性分析
- 评论的关键词抽取与聚类
- 商品评论标签




- 依据用户真实情感判定评论星级
- 折叠无意义评论
- 提供按照关键词查看评论功能
- 商品评论标签用于搜索推荐

| | 机器准确率 | 人工准确率 |
|---|---|---|
| 无意义检测 | **99.8%** | **99.91%** |
| 星级分析 | **89.2%** | **90.19%** |

```
1 画面比较清晰 140
2 看起来很大气 140
3 外形美观大方 108
4 质优价廉物美 102
5 大品牌值得信赖 72
6 看上去很漂亮 68
7 功能齐全 38
8 价格比较实惠 35
9 立体感很强 31
10 相信海信品牌 31
11 性价比比较高 29
12 做工精细 28
```

# 多维度知识抽取

## 商品属性填充

### 商详页图片信息抽取


标签：60、70年代怀旧风

袖型：荷叶袖

领型：飘带领

布料：数码雪纺印花

- **补齐商品属性**
- **富足商品标签**

# 构建商品知识图谱

# •总结：信息合规与多维度数据校验、信息抽取

## 图文属性校验


图片特征提取+NLP

↓

提升搜索精准性
试点类目颜色精准度提升16%

## 知识抽取

商详页OCR　用户评论　客服聊天

多场景文本信息获取，构建商品知识图式存储

## 电商短文本识别

宾际 运动鞋男士跑步鞋春季飞线休闲板鞋韩版透气男鞋子网布潮鞋潮流学生跑鞋 2017夏季新款 灰色(革面) 40

运动鞋、跑步鞋、休闲板鞋、潮鞋、跑鞋

↓

建立结构化电商短语层次化图谱
为商品标题降噪，优化商品生态

## 商品类目自动分类

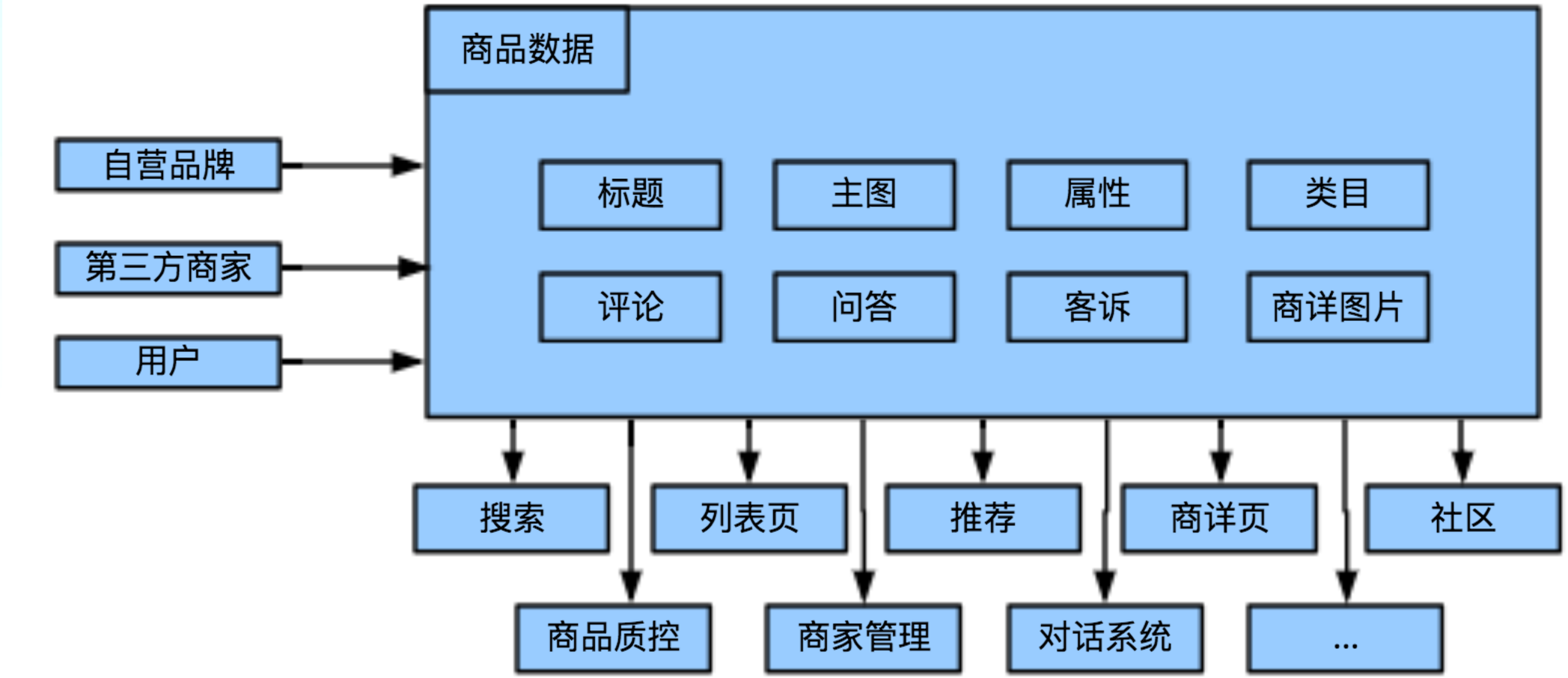
家用电器
手机 / 运营商 / 数码
电脑 / 办公
家居 / 家具 / 家装 / 厨具
男装 / 女装 / 童装 / 内衣
美妆个护 / 宠物
女鞋 / 箱包 / 钟表 / 珠宝
男鞋 / 运动 / 户外
汽车 / 汽车用品
母婴 / 玩具乐器
食品 / 酒类 / 生鲜 / 特产
医药保健 / 计生情趣
图书 / 音像 / 电子书
机票 / 酒店 / 旅游 / 生活
理财 / 众筹 / 白条 / 保险

由于历史遗留及搜索漏洞问题
京东大部分品类类目错绑现象严重

↓

利用深度学习模型
自动指定类目，营造智慧生态

# •总结：信息合规与多维度数据校验、信息抽取



# THANK YOU