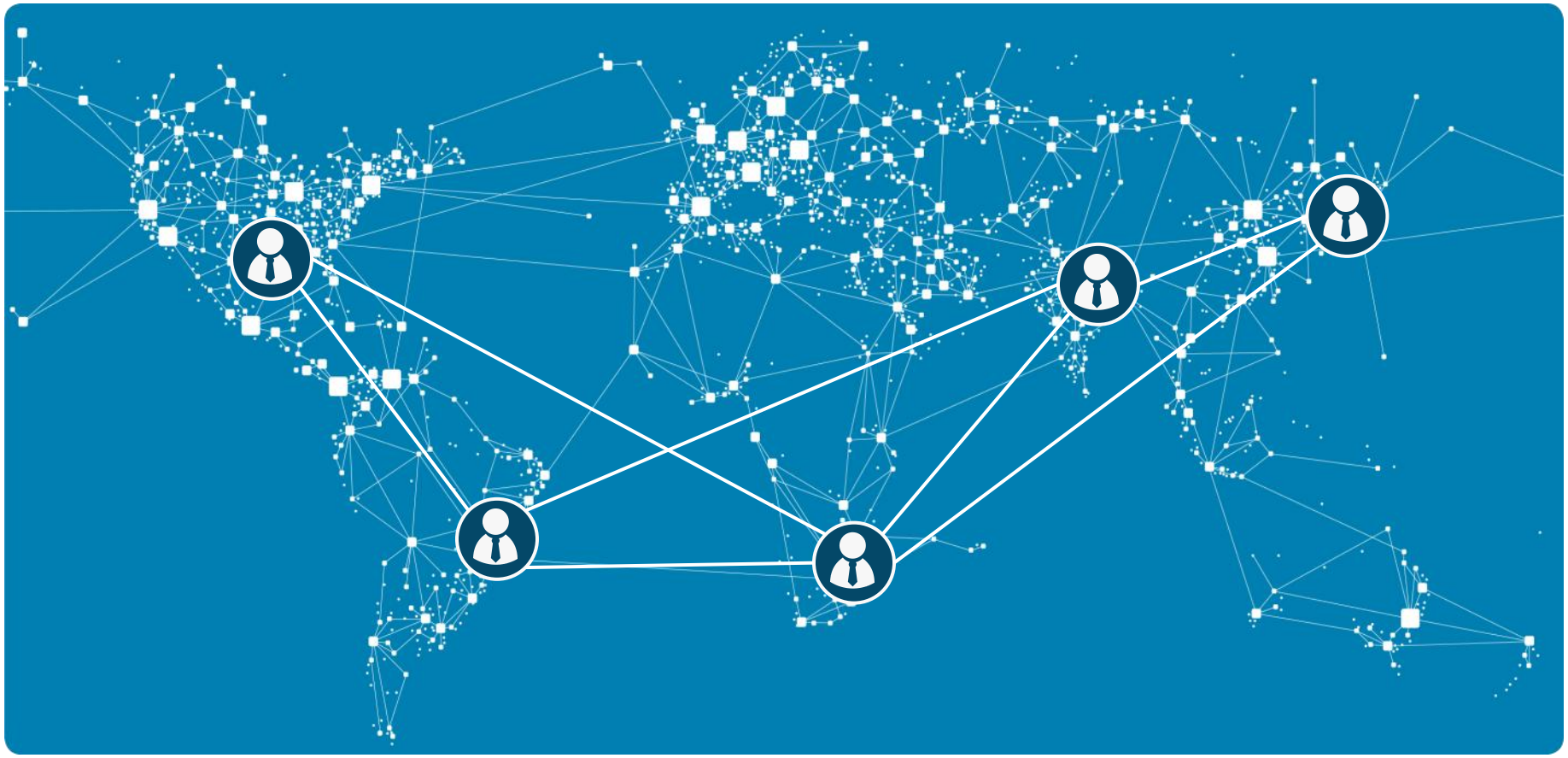
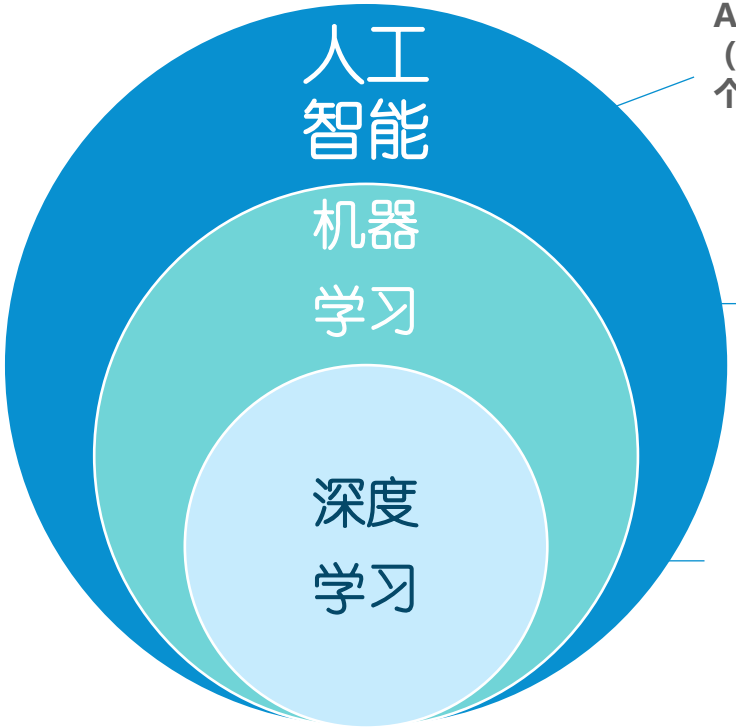


AI应用及广告营销前瞻

浅谈趋势及在广告行业的应用前景

人工智能的定义

- 人工智能(Artificial Intelligence,简称AI)，简而言之，就是“让机器学会如何自主思考”。
- 2016年3月，当Google Deepmind的AlphaGo击败韩国围棋九段李世石后，AI概念开始进入“大众视野”。媒体在报道中就混杂了人工智能、机器学习和深度学习等多种术语,但三者的关系其实如图所示。



AI在学术界可追溯到1950年阿兰•图灵提出的著名“图灵测试”（Turing Test); 1956年达特茅斯学术会议确定了人工智能的各个领域，包括：神经网络、自然语言处理、机器智能等。

机器学习也称为统计机器学习(Machine Learning)或数据挖掘Data Mining),是AI的一个分支，其基本思想是基于数据构建统计模型，并利用模型对数据进行分析和预测的一门学科。

深度学习是机器学习的一类具体算法，属于人工神经网络。在学术界：2006年，Geoffery Hinton首次在Science杂志的论文提出“深度信念网络”的概念，并在2012年ImageNet图片自动识别竞赛中取得分类错误率仅15%的成绩。在商业界：AlphaGo可谓是最为成功的“商业秀”。

AI在广告行业的具体应用及发展趋势

消费者深度洞察

- 以家庭路由器为核心的“单源样本组”(SSP, Single Source Panel) 提供了“机器学习”的“原料”
- 不同于传统的问卷调研，搜集大量消费者网络浏览或手机行为数据已成为可能，我们可以从行为特征推断消费者偏好

网络自然语言处理

- 让机器理解中文已经不是梦想，有成熟的算法处理“中文分词”，“词性标注”，“语义分析”等
- 这项技术早已成为DMP和搜索引擎的核心，如网页类型及主题的判断。
- 随着社交媒体的兴起（微信、微博），已经成为社交媒体及舆情分析的利器，并深刻影响广告行业的研究方法论

异常流量+可视曝光量

- 异常流量(NHT, Non-Human Traffic)是由程序（机器人）产生的活动，通常用于广告欺诈。AI技术用于识别并剔除虚假广告。
- 即使消除NHT问题，部分曝光量仍没有机会被受众看到，通常是因为库存位于用户未翻阅的网页版面。识别这部分流量依靠的是“图像识别”技术。
- 只有剔除了上述流量，才能衡量数字广告真实的效果，

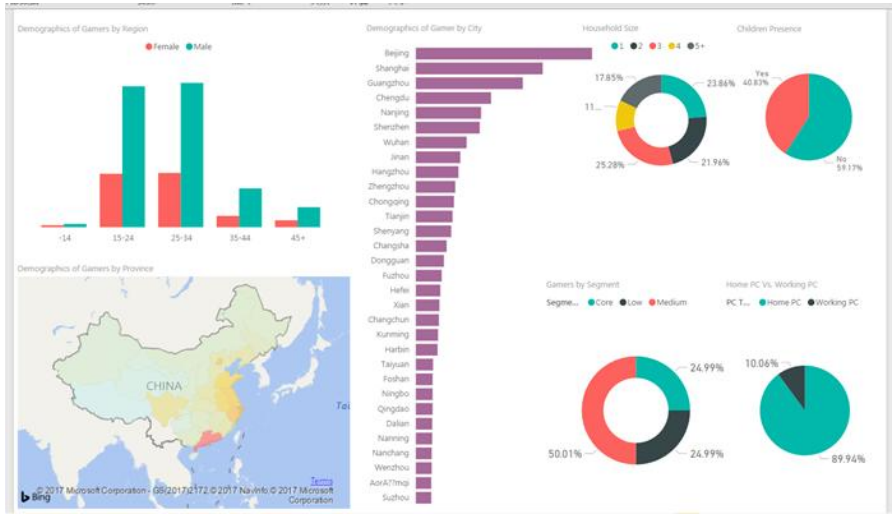
消费者深度洞察

原理及演示案例

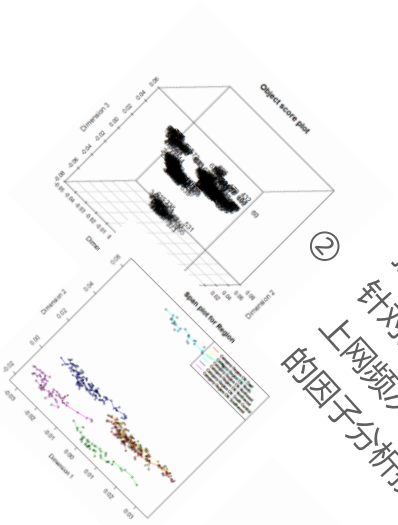
“大数据”为AI的知识发现提供了真正的“原料”

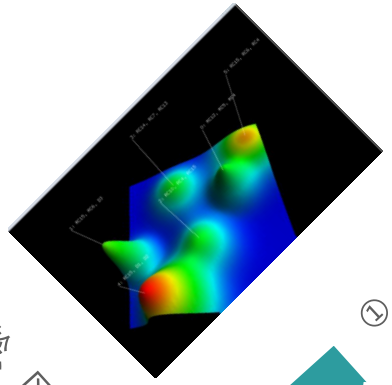
单源数据固定样本组（SSP）
供应商 AdMaster或XInsight

基于Meter监测的固定样本组
供应商 ComScore或ComRating



依据的机器学习算法

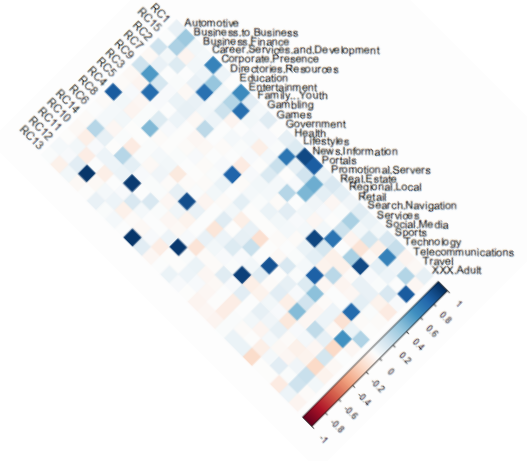




01特征变量降维

① 针对性别、年龄、区域等离散型变量采用多重对应分析技术转为连续性变量。

② 针对连续性变量如单天平均上网频次，则直接采用传统的因子分析技术



02 聚类分析

① 两大划分算法：经典的k-均值聚类(k-mean)和改进的k-中心点聚类(PAM, Partitioning Around Medoids)

② 三大层次算法：(1)反复二分法(Repeat Bisection)（2）直接方法，k路聚类同时找到k个簇；（2）凝聚方法，即采用自底向上的层次凝聚方法。

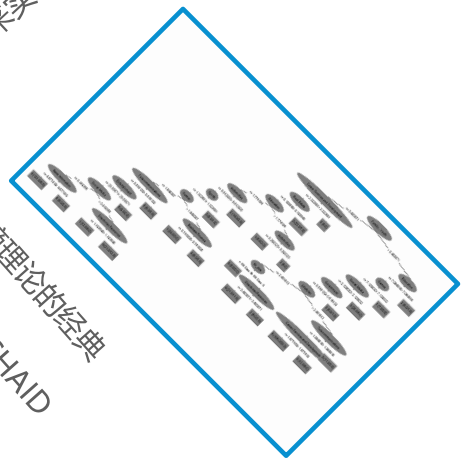
③ 图方法：这里主要通过最近邻图和最小割（mincut)图划分算法来实现

03 分类树

① 基于信息熵理论的经典C4.5算法

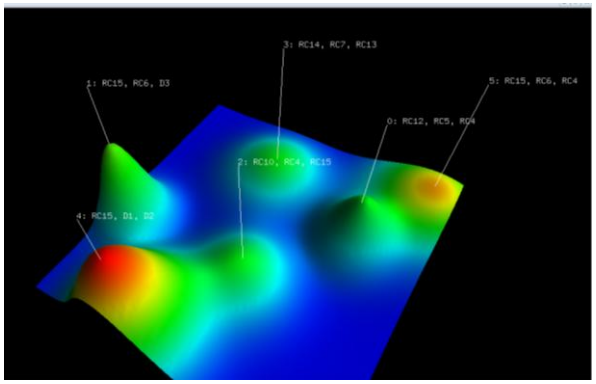
② 基于统计理论的CHAID算法

③ 基于不纯性度量(Gini指数)的CART算法

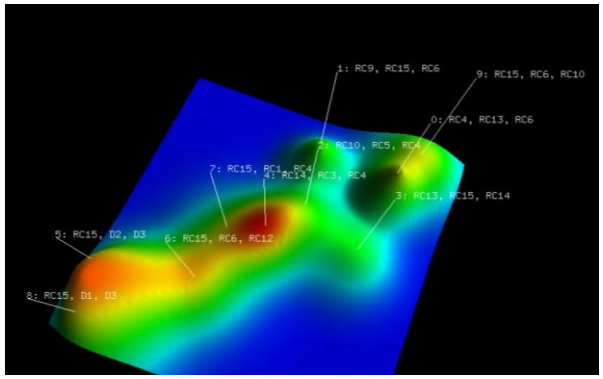


聚类分析-“物以类聚，人以群分”

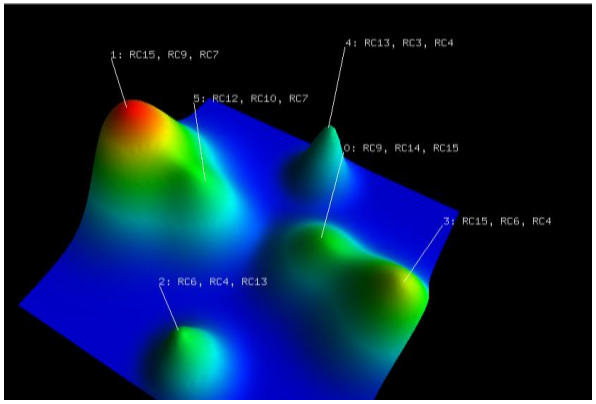
相关系数 图方法 （6类）



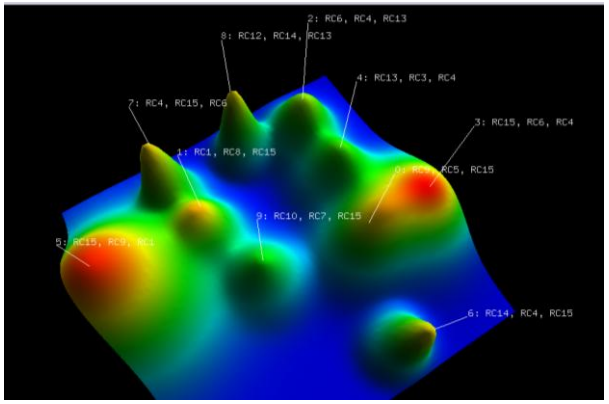
相关系数 图方法 （10类）



余弦相似度 凝聚方法 （6类）



余弦相似度 凝聚方法 （10类）



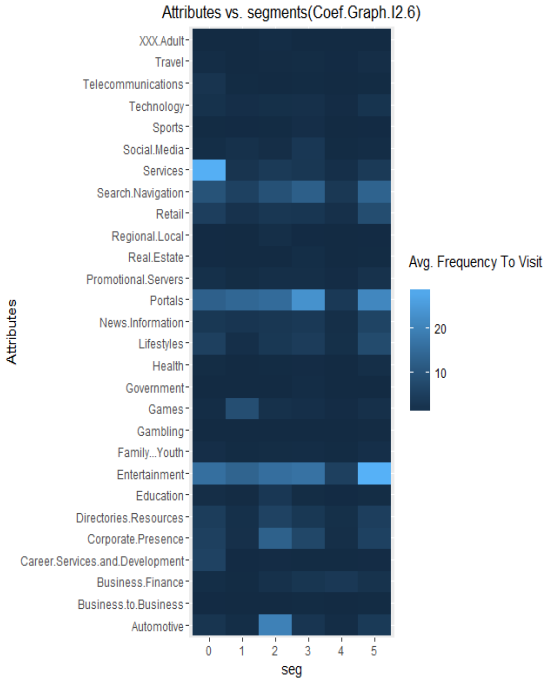
曲面可视化最为直观。

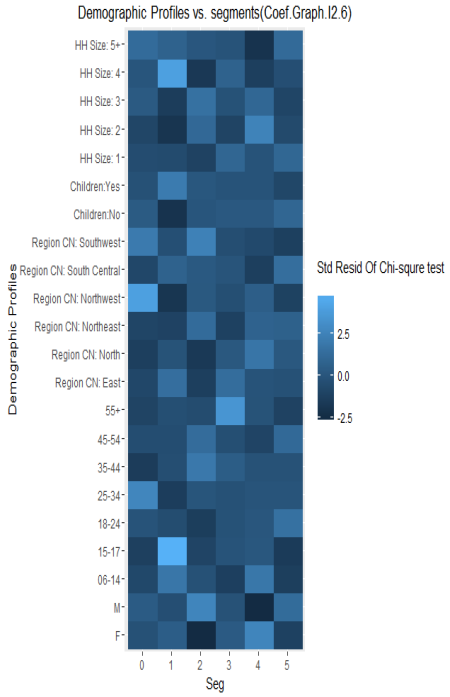
从本例可看出：
相关系数图方法以分6类更清晰。

如果要分10类，则建议采用基于余弦相似度的凝聚方法。

给我们的细分群体打上“标签”

在选定最优聚类方案后，再通过特征细节的具体比较，可为每个聚类簇贴上合适的“标签”，转化为营销角度的细分市场。
根据行业和营销目标，再找到投资回报率最高的目标市场。



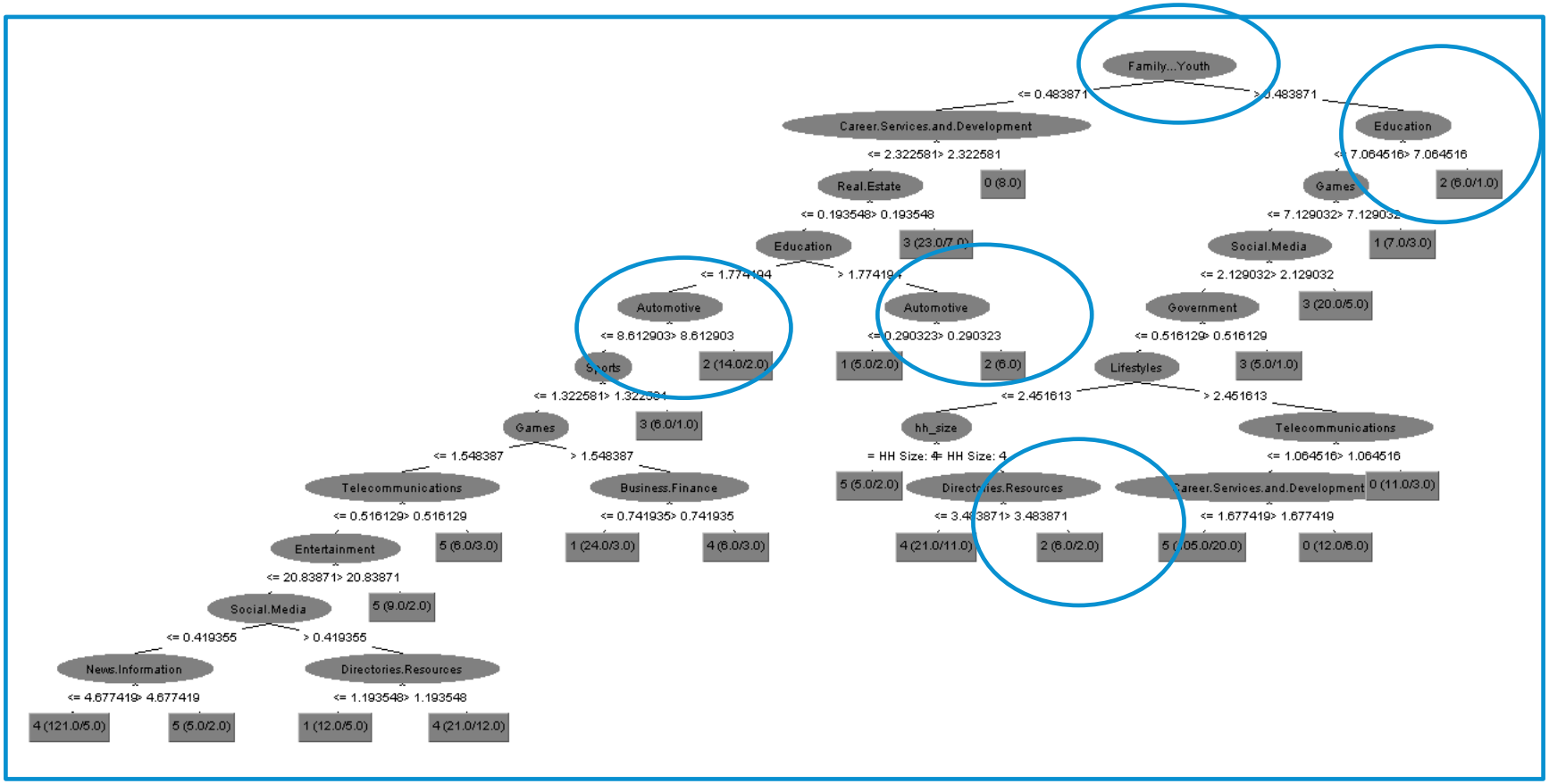


细分群体	上网习惯	背景特征
0:实用主义者	偏好服务类网站，有特定使用目的	25-34岁 北方和南方城市偏多
1:游戏爱好者	偏好游戏类网站	17岁以下，家庭成员数为四口的大家庭
2:汽车爱好者	偏好汽车类网站、教育培训类、信息导引类(地图)和家庭类网站	男性偏多，35~44岁，家庭成员数2~4为主
3:传统使用者	偏好门户类网站，传统搜索及娱乐	最显著特征是以55岁以上老年人为主
4:随意冲浪者	无特定偏好和目的	女性偏多，家庭成员数2~3为主
5.娱乐搜寻者	明显偏好娱乐类网站，多用门户类及搜索	男性，18~24岁，南部城市偏多

营销及广告策略的指导意义

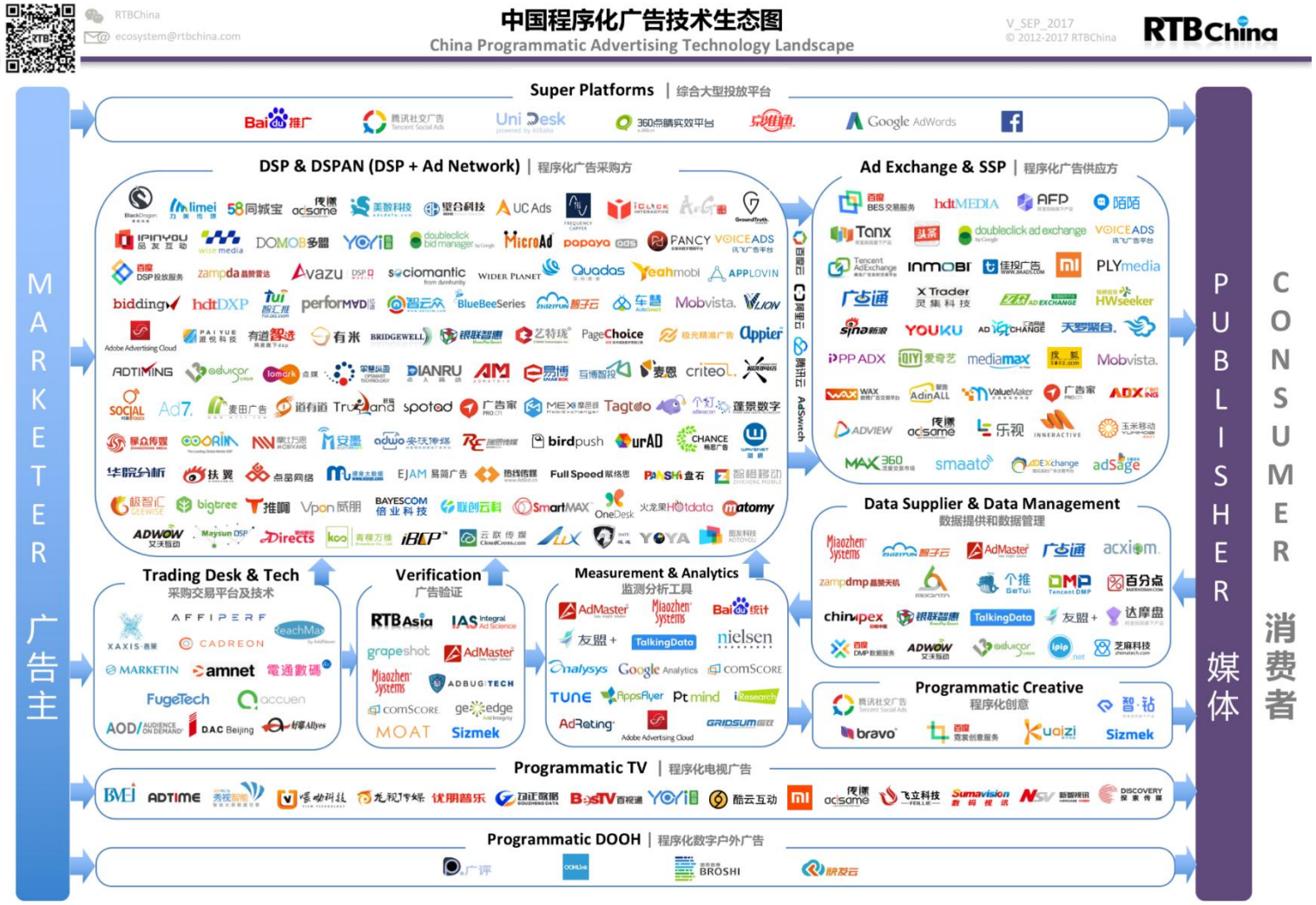
如何更有效地触达“汽车爱好者”是成功营销策略的关键:

- ✓ 此群体是“上有老下有小”的典型中产阶级，非常适合中等偏上家用车目标群体。
- ✓ 汽车垂直类网站是核心；教育培训类和信息导引类(地图)网站的广告投资是有效的组合
- ✓ 与家庭有关的网站（如儿童教育）广告是有效的手段，可有效触达家庭成员数大于4的目标群体。



程序化购买-让“消费者洞察”更为“实用”

“消费者洞察”已经广泛应用于程序化购买中。我们也可以添加自己的“标签”，并与DSP供应商打通（数据包方式）和“扩大”到业务级别（Looks-like算法）。



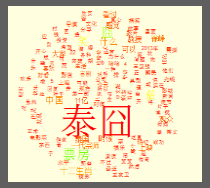
网络自然语言处理

原理及演示案例

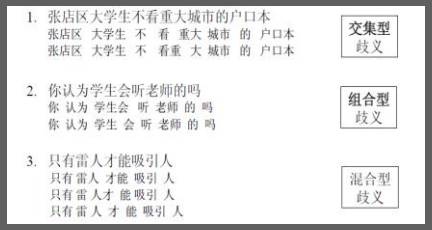
何谓“自然语言处理”技术？

第一步: 中文分词

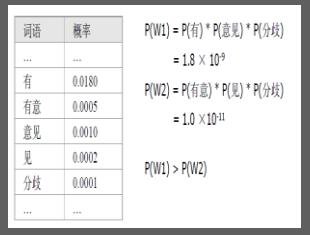
中文自动分词是理解语义的基础。没有迈出这一步，任何深入分析都是妄谈。



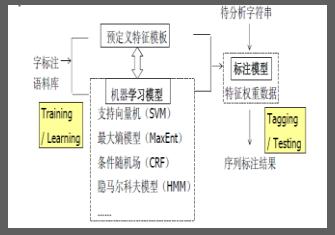
但中文自动分词在技术上有很大难度，主要困难是语义歧义性



解决方案一：
最大概率法分词

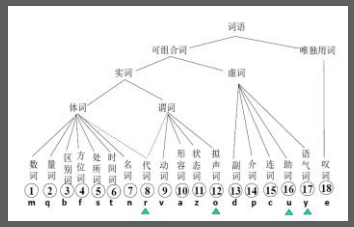


解决方案二：
基于字序列标注

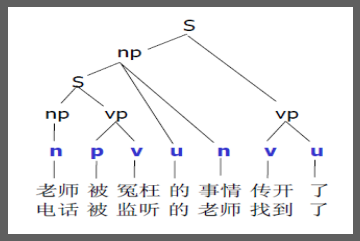


第二步: 语义分析

中文词性标注，即主谓宾分析



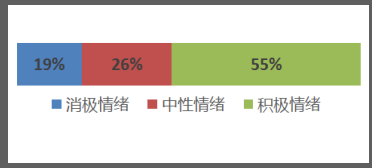
中文句法结构分析，算法自动解析完整句子的语法树，语义消歧，确定多义词在具体语境里的含义



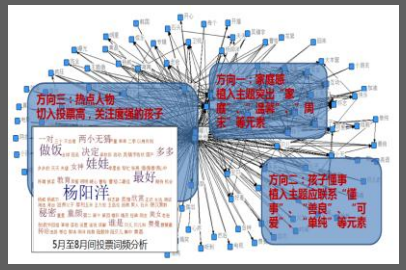
配价描述的内容	取值（具体的描述方式）
配价数	1. 动词跟 x 类名词有语义联系； 2. 动词周围有 x 个空位放置名词；（x 一般取值 0-3）
论旨角色	1. 跟动词有语义关联的 x 类名词分属哪些类型（T_1, ... T_i） 2. T_i 能够出现在动词周围的哪些空位上
动词对其论旨角色的选择限制	充当 T_i 的名词要满足哪些条件？ 1. 语法（形式）特征 | 语义属性（类别）特征 2. 包容性条件 | 排除性条件

第三步: 具体应用

情感度分析：微博或网页对节目线或明星/编导的具体评论的情感度得分，是正面评价还是负面



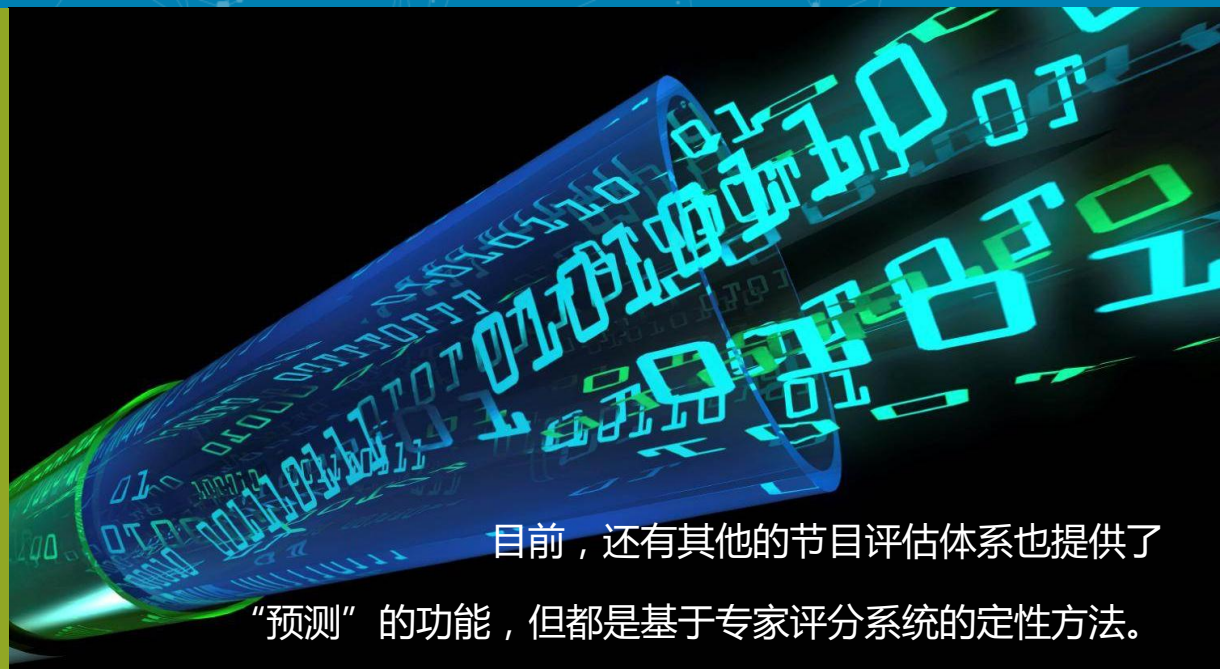
语义关联度分析： 导致情感的关键因素有哪些？



评论相似度/聚类分析：哪些评论具有相似特征？ 哪些评论具有“水军”特征？ 所有的评论大致可分为哪几个大类和方向？

此外，Predictor 2.0通过算法和大数据处理，对全新的节目做出预测，帮助广告主以及市场营销人员：

- 在节目开播之前就了解节目的潜力和潜在观众
- 将节目的内容元素进行分解，研究可能影响收视率的最重要元素，为品牌植入提供突破点



目前，还有其他的节目评估体系也提供了“预测”的功能，但都是基于专家评分系统的定性方法。

并且这些系统只提供了“推荐指数”，而不直接预测收视率。

Predictor2.0通过对过往节目信息（包括收视率、平台、明星等）以及社交网络热度、搜索指数、新闻热度等数据进行分析，从更加定量的角度，预测节目收视率，为广告主评估并建议内容机会。

结合社交舆论热点，提高新节目收视率预测精度

Predictor 2.0通过建立模型分析社交网络大数据（至少一千万贴），识别出的12个影响收视率的热点。

除此之外，收视率预测模型中考虑的其他因素还包括：电视频道、明星、制作团队、节目模式是否引进、播放时间、以及同时段是否有其他重要节目分流等等。这诸多因素，让Predictor2.0预测全新节目收视率成为可能。

异常流量及可视曝光量

原理及演示案例

何谓“异常流量”？

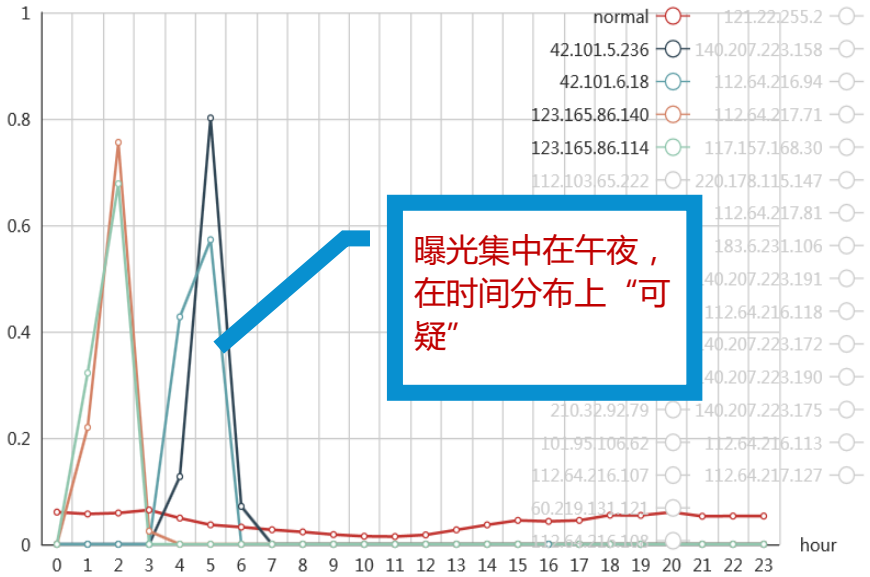
- 知识背景：互联网用户分为三个最大的“人群”：男，女，机器人；机器人会模仿人类的形态浏览网页、观赏视频、点击广告、撰写评论、投票点赞。机器人通常居住在“数据中心”。

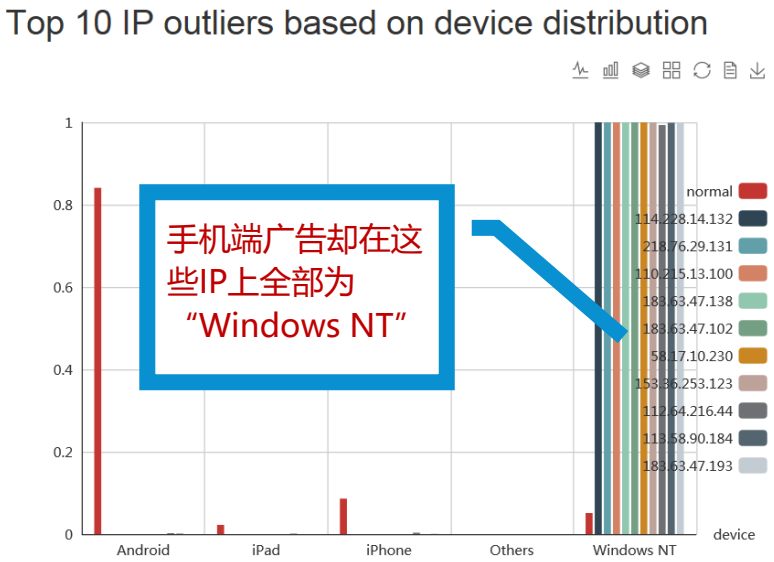
- 当IP场景为“数据中心”时，此IP发出的网页浏览行为大多数情况下属于NHT（Non Human Traffic，非人类的访问），可能是各种功能的机器人：搜索爬虫、内容采集器、舆情监控、网站性能监控、压力测试器、自动发帖机、安全检测软件等等，你懂的...

- 甄别“异常流量”也是AI的一个重要分支，属于“异常值检验”范畴。通过机器学习作出自动判别：“真人概率”的数值在50%以上，可以被认定为此IP的网页访问量基本由人类主动行为产生，分值愈高越真实。低于50%则有较高可能性是此IP的行为是机器人主导。

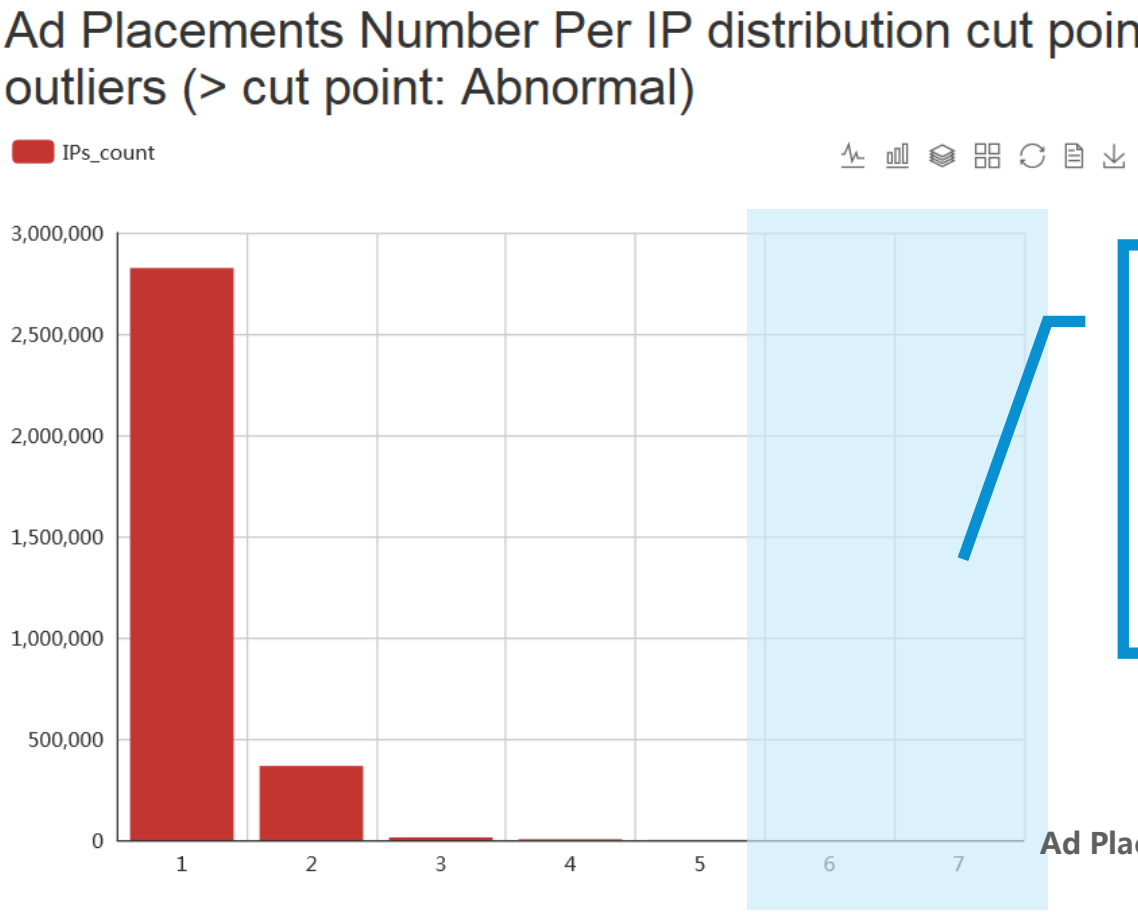
如何甄别“异常流量”NHT? –传统方法

传统方法主要通过上网行为特征来判断“异常流量”NHT。
但作弊“机器人”目前也越来越“智能”，完全可以更“逼真”地模拟人类的行为。





如何甄别“异常流量”NHT? –现代方法



基本逻辑: IP空间如此巨大，许多广告通过同一个IP曝光的概率微乎其微。

作弊广告的目的是非法盈利，作弊的IP必然有限。根据我们自己数据的研究，一个IP平均超过6次的广告位只有5%的发生概率。如果“小概率”事件发生了，IP本身就是可疑的。

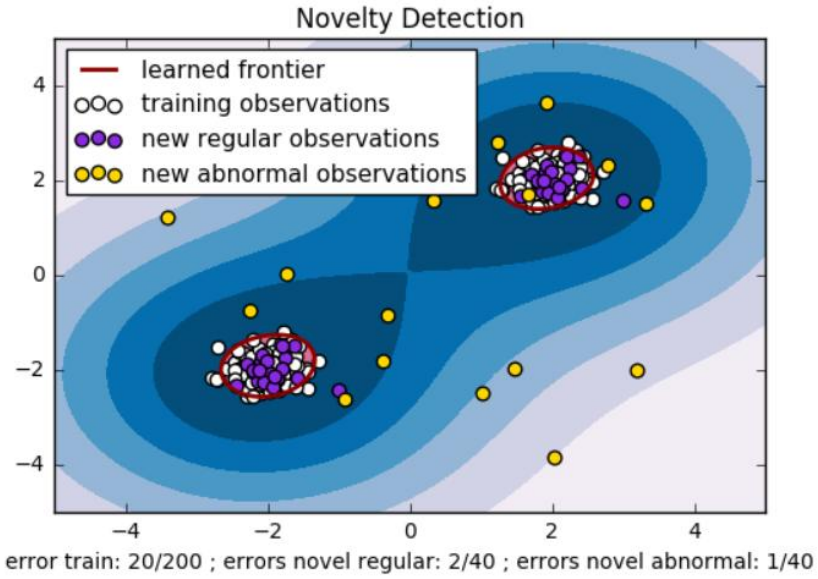
```
## [1] "Kolmogorov-Smirnov test: Null hypo- Ad Placements Number Per IP is in Exponential Distribution"
```

甄别“异常流量”的AI技术

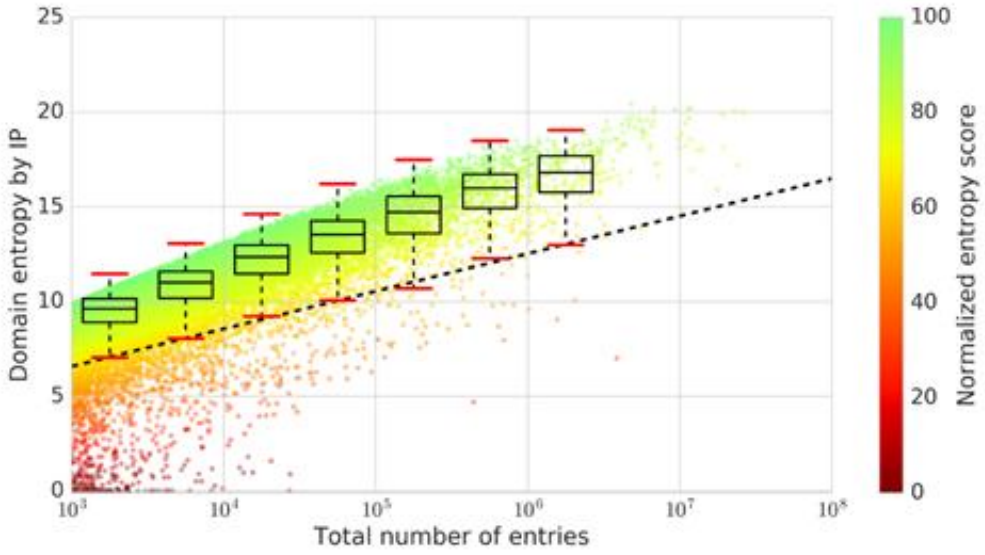
方法论上基本采用的是“机器学习”（也叫数据挖掘）的“异常值检验”。
如下是目前常用的技术。

在不久的未来，我们需要的是“行业标准”，如权威的IP黑名单和白名单，GIVT和SIVT标准（一般由行业协会牵头）。

基于“行为特征”的传统方法

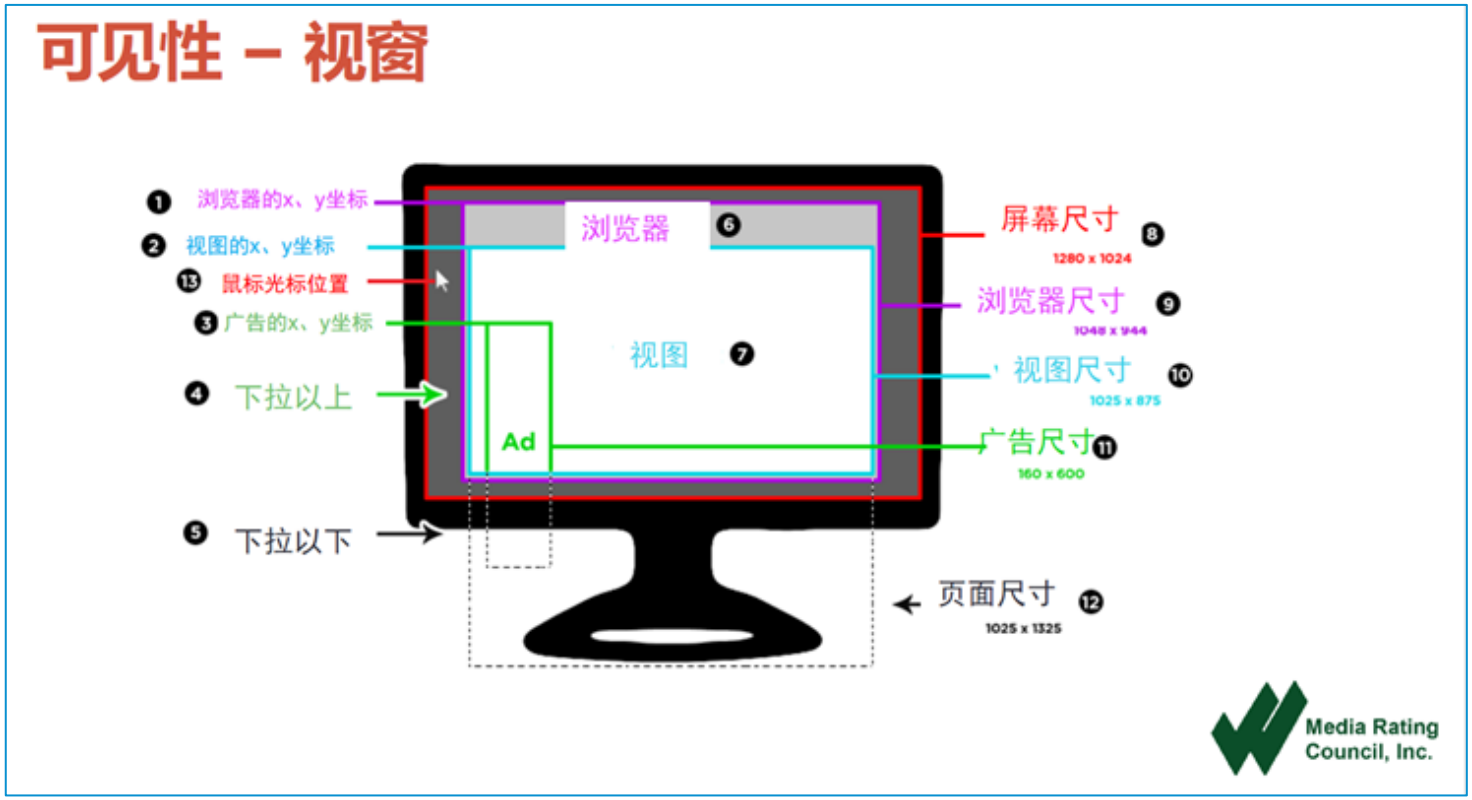


基于“IP与广告位或Reference页面关联性”的现代方法（如信息熵）



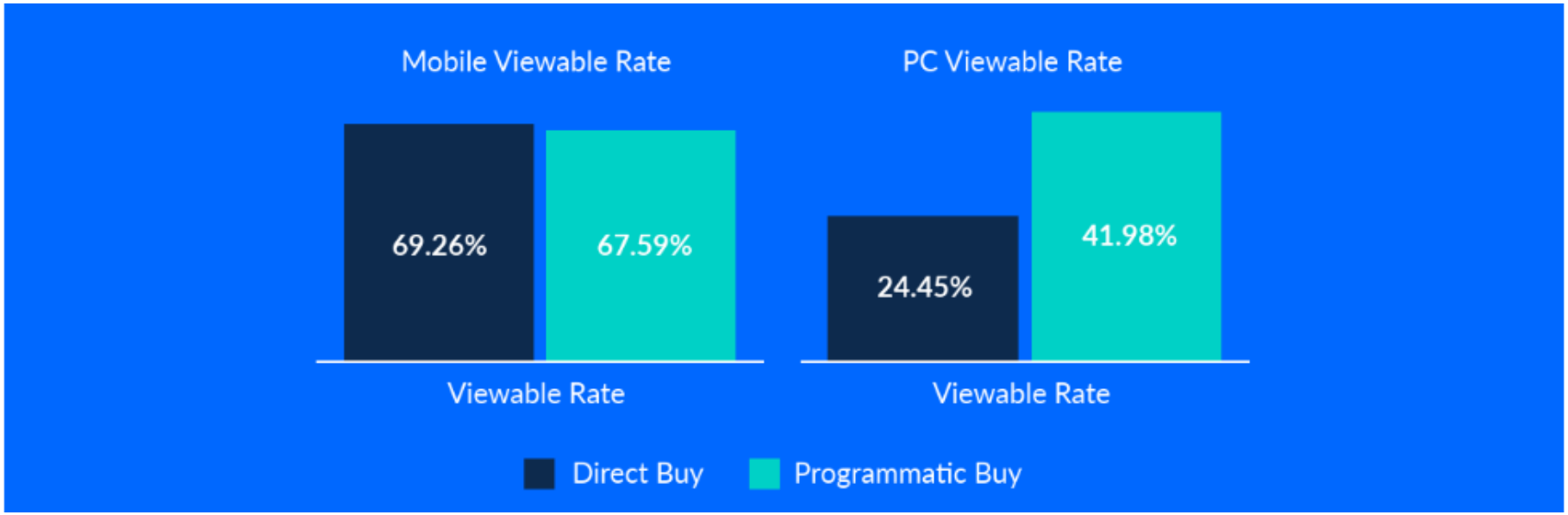
何谓“可视曝光量”（MRC的定义）

- 测量标准2014年发布台式电脑指南（2015年更新）；2016年发布移动设备指南。
- 在曝光可计数之前，确定可见广告的门槛值为：50% 的广告面积露出/1 秒钟展示，50%的广告面积露出/2秒视频。
- 通过要求最低可见参数，建立了数字广告的“机会可见”概念，类似于其他媒体上的“可见概率”概念（OTS, Opportunity To See)



我国的现状-基于Sizmek中国数字广告可见性报告2017年版

- Sizmek的报告是基于IAB的标准即50%的广告呈现在用户屏幕的可见区域上一秒或更多。可以看到移动端得益于便携性，屏幕尺寸以及广告形式，可视化曝光的比例都接近七成，远高于PC端，广告的投放效果更加直观。 同时程序化购买可以使得PC端广告的可见率提高将近一倍。



图像识别-“可见曝光量”测度的AI技术

- 最近很“火”的所谓“深度学习”最初就是源自于“图像识别”，比如自编码网络AE(Auto Encoder)算法压缩数值矩阵，再结合GPU等技术，使得监测实际需要的“大规模处理”成为可能。

页面几何：测量标识相关的矩形，并跟踪广告与页面、页面与视窗的相对位置

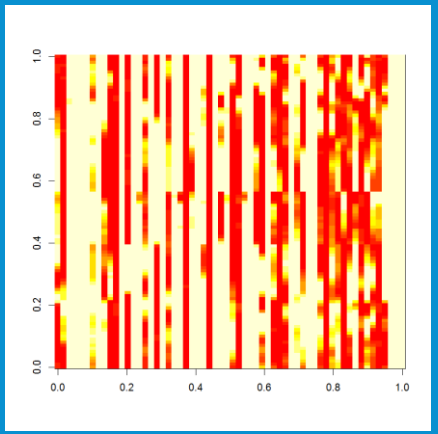
- 与页面相关的广告
- 与视窗相关的页面
- 与屏幕相关的页面

浏览器优化：测量相对于广告位置的特定位置识别，并监视计算机和浏览器对其的资源分配

- 所谓资源分配是围绕广告矩形周边布置的Flash像素
- 位置和像素数量因测量供应商而异

本质上就是基本的“图像压缩及识别”技术。

如下图为AE的例子：将图像转化为数值矩阵，压缩提取特征。然后将广告所在区域“划框”，识别在页面图像上的坐标位置。



谢谢！