大数据·小算法

-实用的用户行为研究方法

艾瑞 郝欣诚

2014.4

iResearch
艾瑞咨询集团

内容目录

企业数据分析的需求动因

客户：

今天请你们来主要是想...

案例一：商品定价数据分析

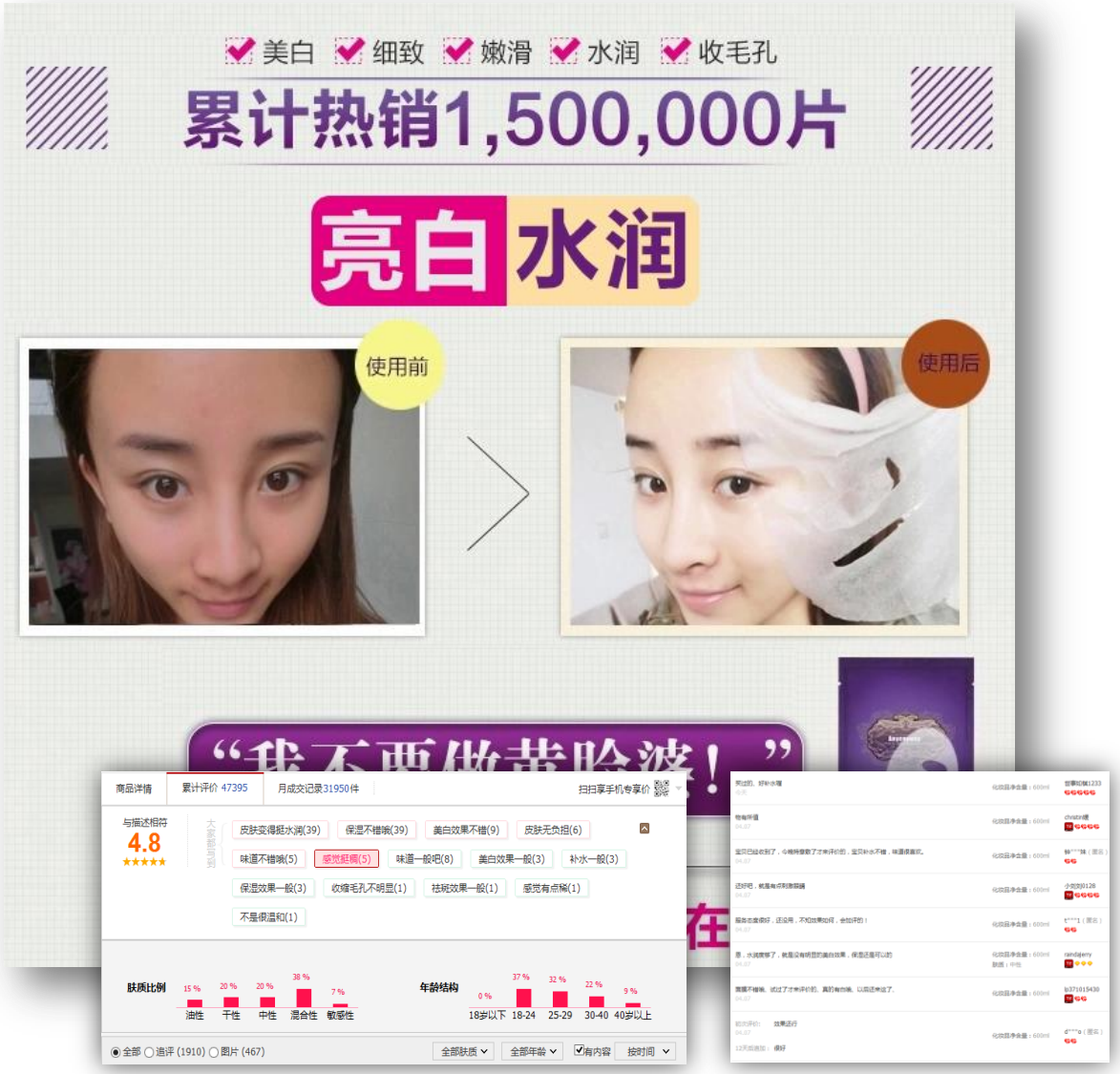
- 客户问题：
  - 儿童内衣主力淘宝定价在50-80元之间，并且竞争非常激烈，如何根据产品的特点，结合数据分析指导定价？

- 需求细节 ：
  - 60元新品亏损线
  - 套装组合及定价
  - 新特性产品及定价

案例二：商品特征驱动力分析

- 客户问题：
  - 美白面膜在高度饱和性竞争中，希望能对美白之外的附加特征中，寻找细分或潜在市场的驱动力

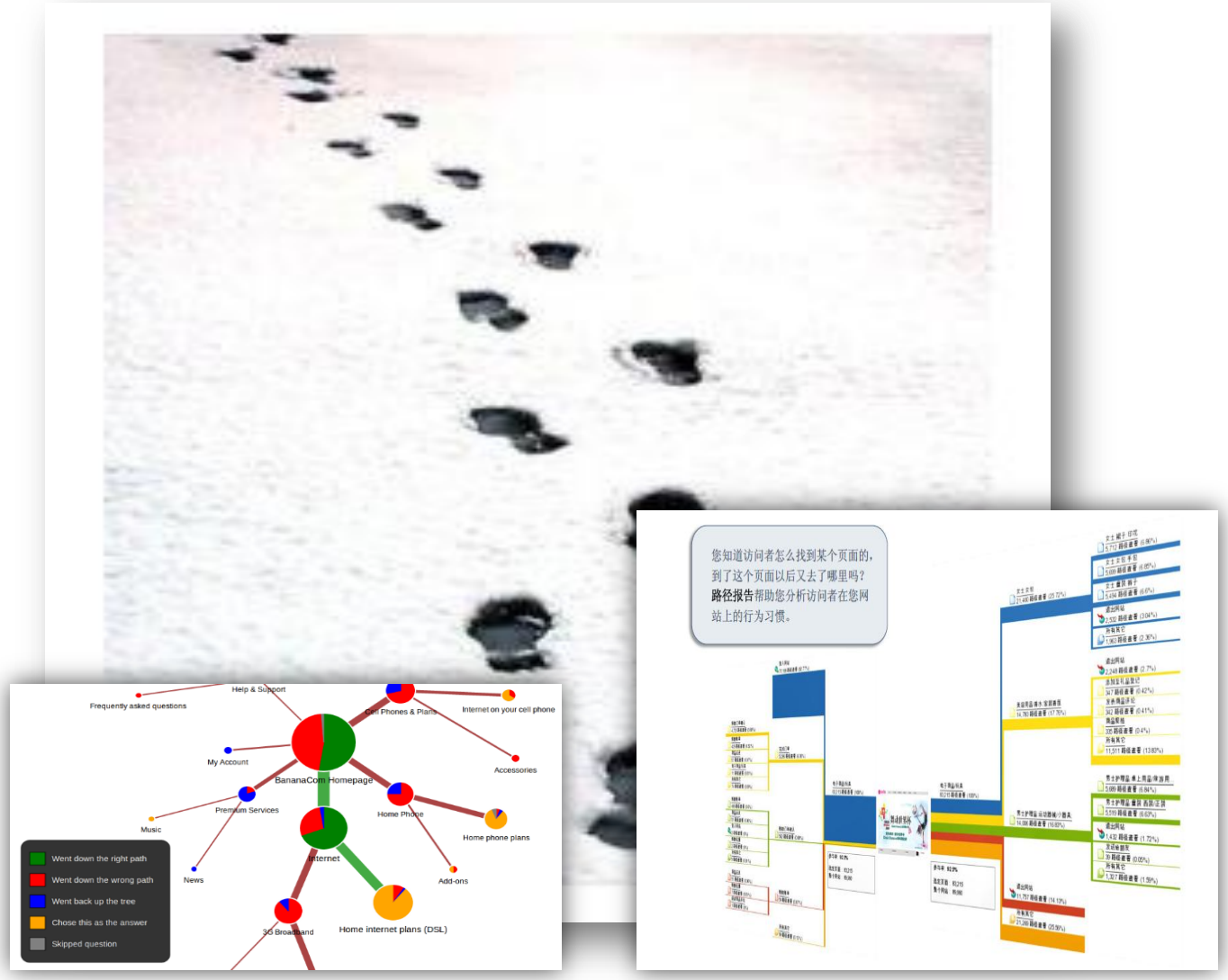
- 需求细节 ：
  - 需要有定量的结论
  - 对“舒服”进行细分



案例三：访客行为路径分析

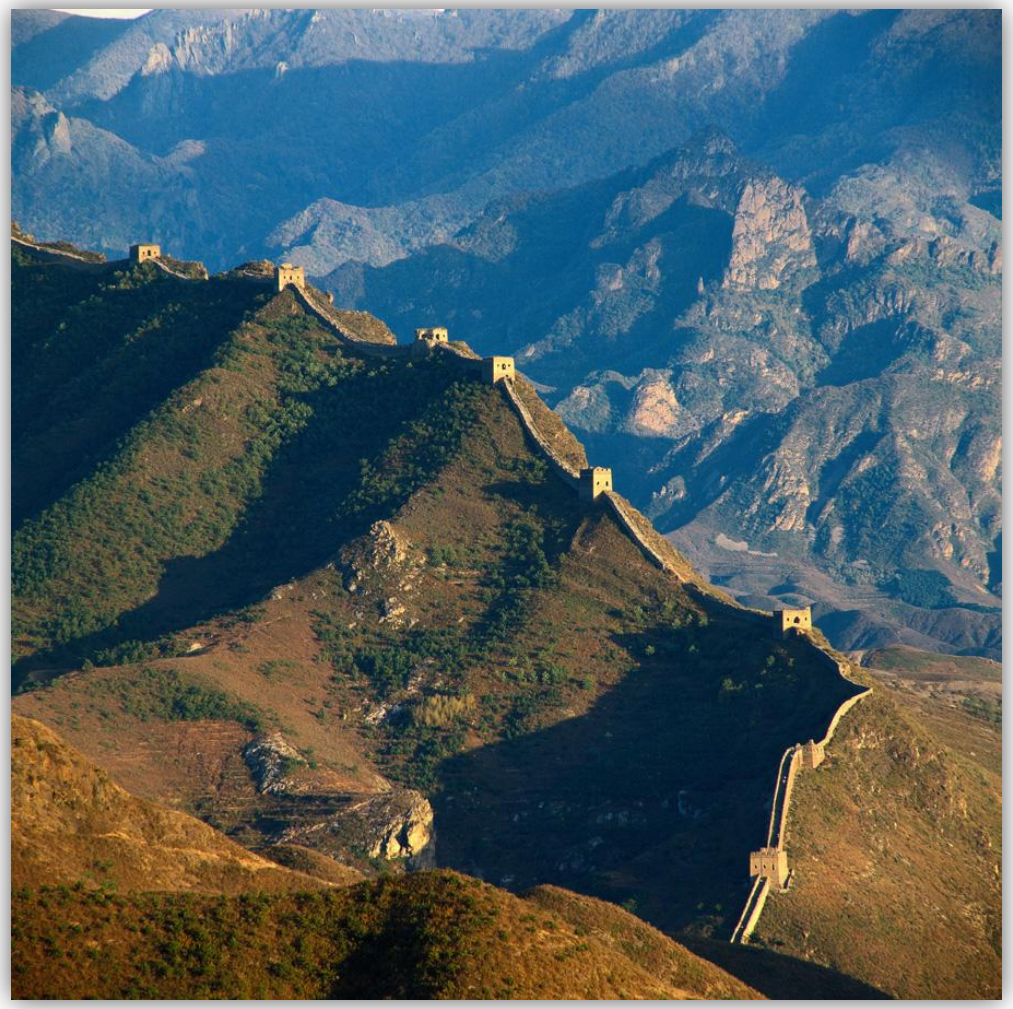
- 客户问题：
  - 网站访客，分析访问前中后行为，加强重要的引流渠道投入，指导内容编辑和竞品研究工作。

- 需求细节 ：
  - 数据实时性要求高
  - 热点情报准确
  - 与客户数据协同分析

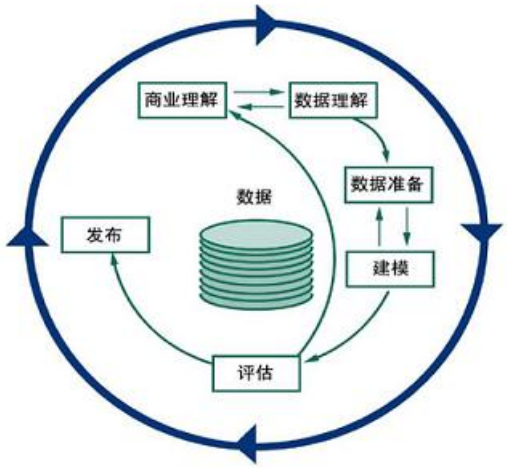


数据分析项目的成本收益

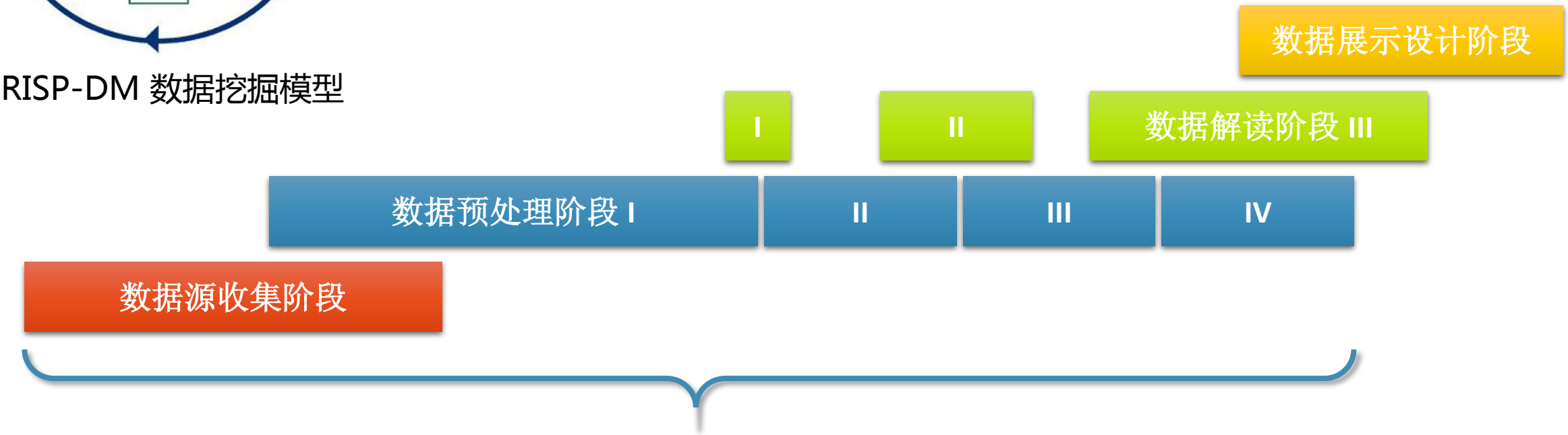
- 数据本身没有固定价值，处理数据的过程存在成本，应用数据的收益决定价值！
- 数据成本核算（TCO / MB）：
  - 建设成本，软硬IT系统投入及年度维持、升级成本
  - 运营成本，数据源成本及协调、分析、管理人员成本
- 数据质量核算（TNR / MB）：
  - 净结果集含量，最低维度结果集行数和列数
  - 净结果准确率，准确率2.5倍率衰减
- 数据回报率核算（ROI / RMB）：
  - 基准收益率 vs 优化收益率
  - 基准收益规模 vs 优化收益规模



数据项目周期分布特性



CRISP-DM 数据挖掘模型



为期6周的定制数据项目，90%的时间涉及处理数据

数据挖掘的九大定律

Business Goals Law：每个数据挖掘解决方案的根源都是有商业目的的。

Business Knowledge Law：数据挖掘过程的每一步都需要以商业信息为中心。

Data Preparation Law：数据挖掘过程前期的数据准备工作要超过整个过程的一半。

NFL Law：没有免费午餐，数据挖掘的任何一个过程都是来之不易的。

Watkins Law：数据总是有模式可循，找不到规律不是因为规律不存在，而是还没发现它。

Insight Law：数据挖掘可以把商业领域的信息放大。

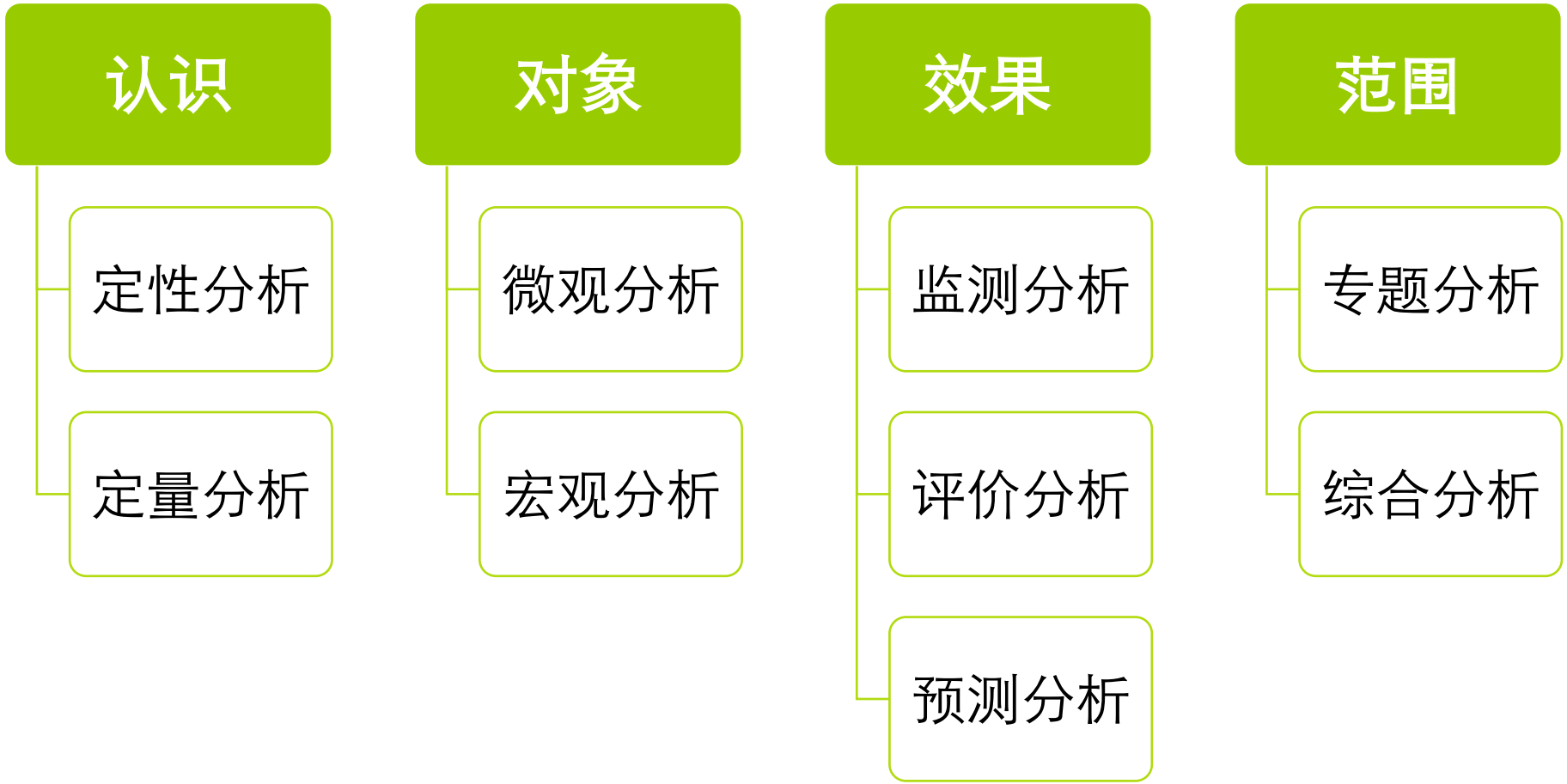
Prediction Law：预测可以为我们增加信息。

Value Law：数据挖掘模式的精准和稳定并不决定数据挖掘过程的价值

Law of Change：所有的模式都会变化。

统计方法四大维度

统计透过现象的数量表现，来认识事物的本质和发展变化的规律

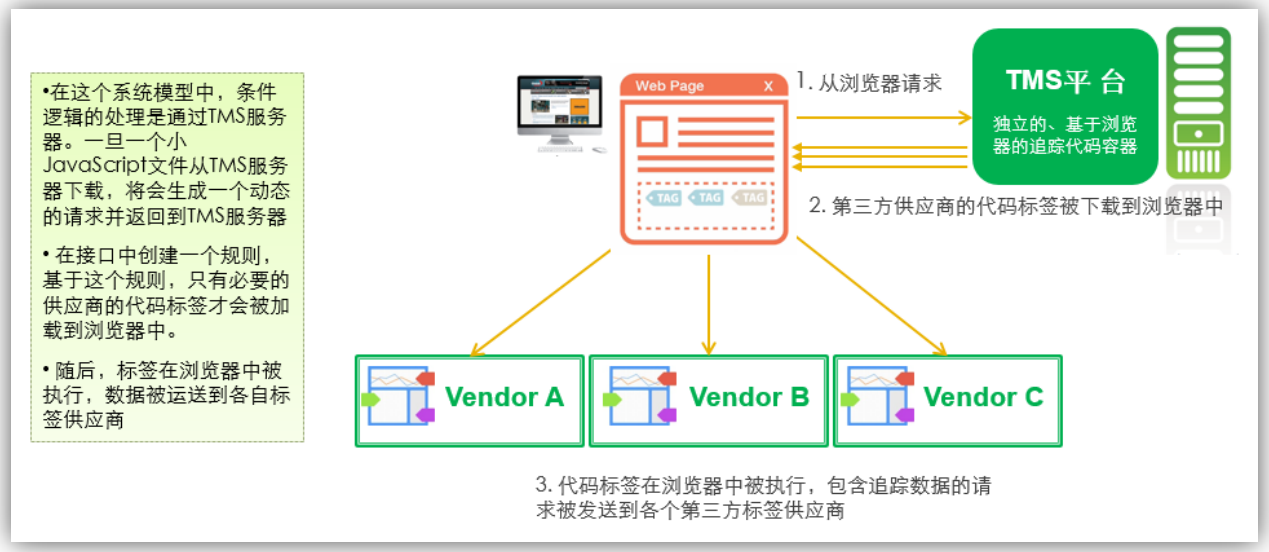


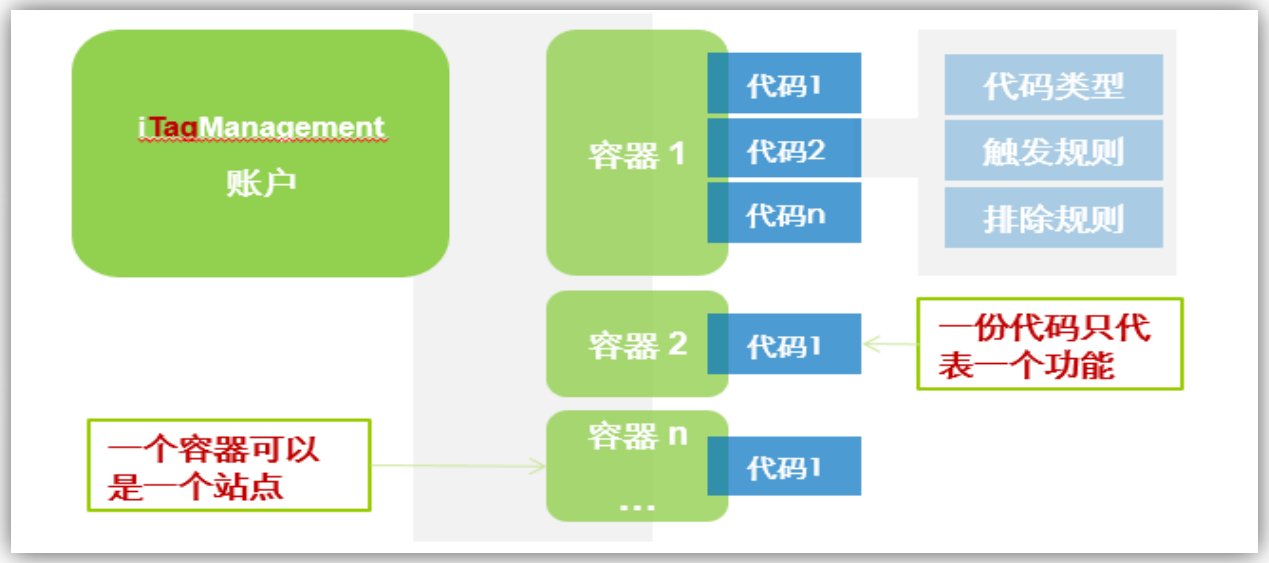
商业解决思路

“思路决定价值” 按规矩办事不会错

内容目录

TMS系统，行为分析项目利器

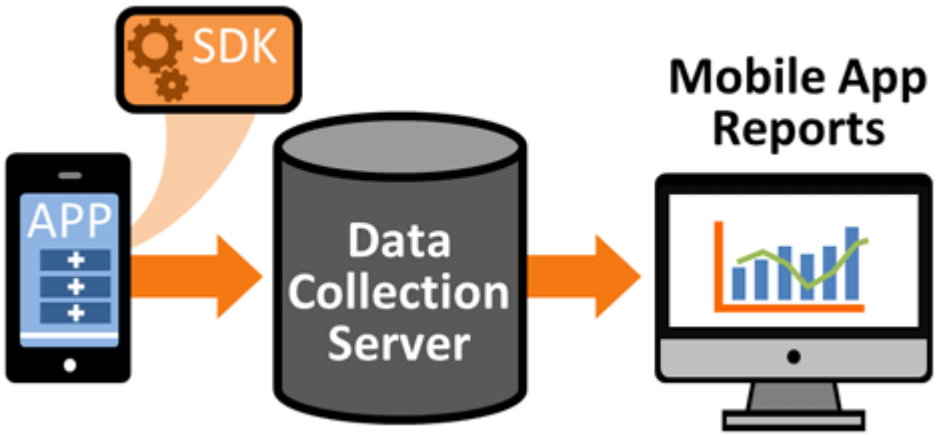




- 一次布码，全面支持各种数据代码
- 可以按十余种条件，激活不同代码或传递参数
- 大大缓解与客户协调技术布码的沟通成本
- 数据节点端点选择丰富，并且可以大幅降低无效数据量
- 目前Google有提供免费产品，部分国内企业也开始提供

移动端HTML5及SDK数据支持

- 移动端数据，更依赖客户自有数据源
- 跨屏数据打通，暂时没有好方案，停留在CRM层面
- 微信大建站之后，引发HTML5数据关注



服务器日志最简单也最难受

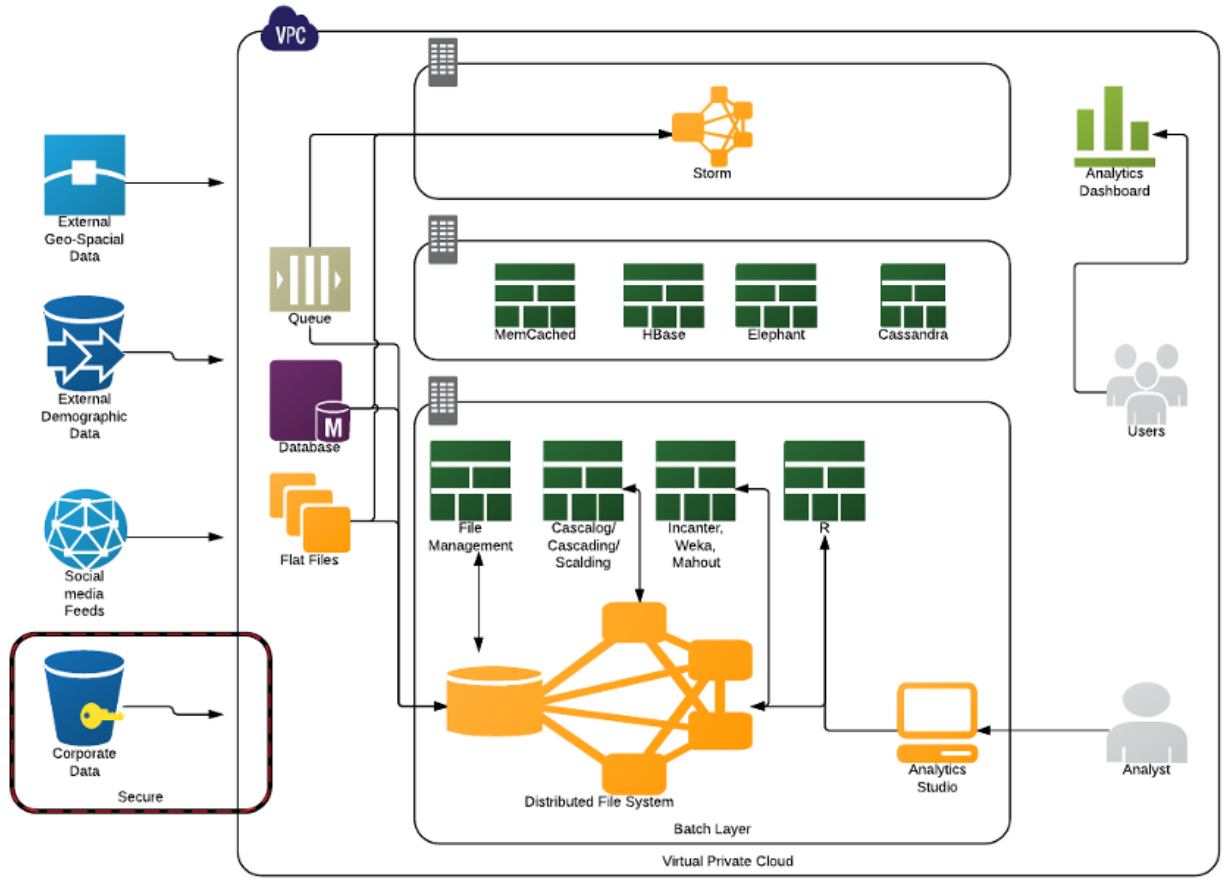
- 服务器日志节点单一，用于复杂逻辑分析非常不易
- 前端数据收集成本持续降低，成本优势已不明显
- 期待服务端分析系统，提供革命性的产品

数据扩维提升数据性价比

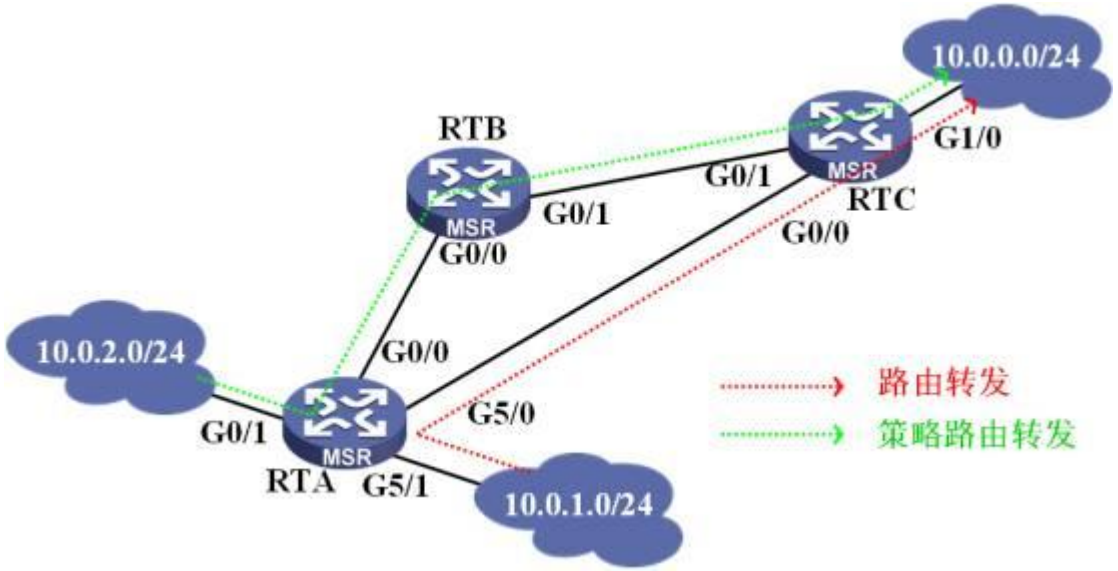
- URL数据本身含量固定，结合爬虫 TNR 可增加数倍
- 两个多列数据源关联，TNR 收益成基数增长
- 时间、地区、手机号，都可以通过成熟方案，简单扩展维度
- 逻辑关系组织，过滤异常数据，提升数据质量

扩展Reduce脚本简单实用

- 用堆栈处理个性化行为路径分析，可以分布式处理，性能好规则灵活
- 改造Reduce实现按规则喷射，让Storm做各种统计逻辑，在大型日志分析项目中，兼顾统计灵活、性能和可靠性
- 一直期待能完成上述两项工作的产品工具的出现



匿名身份识别处理问题

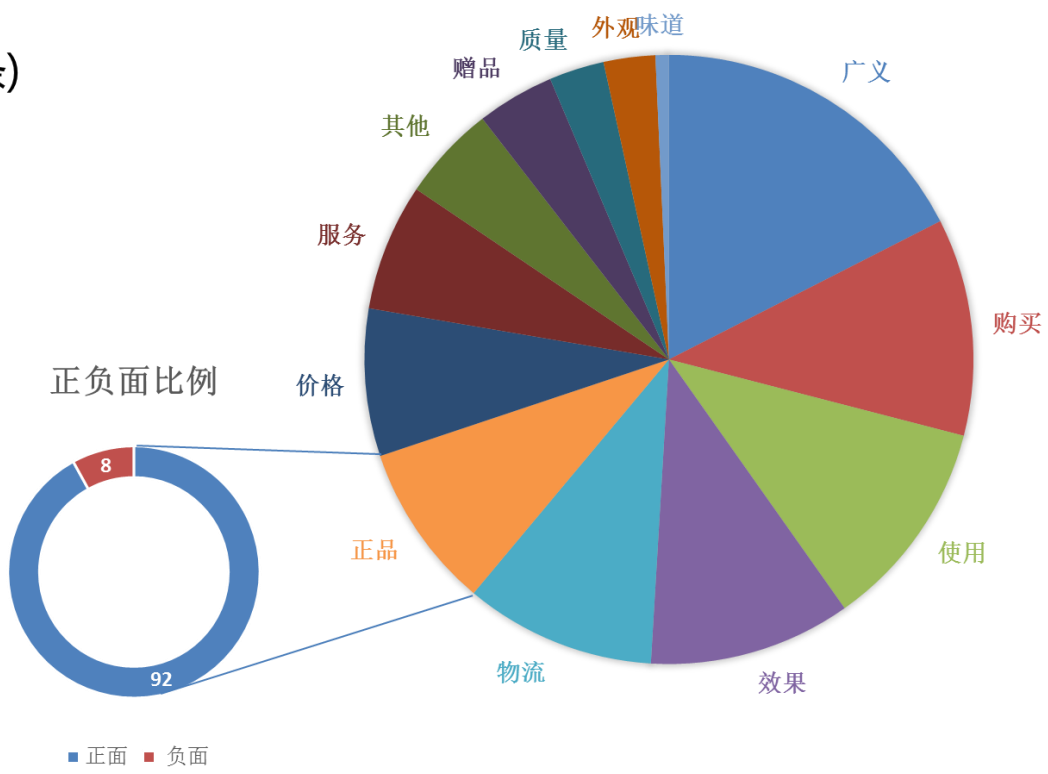


- 身份识别是行为分析项目的基础，即便是想统计出访问设备个数，也会受到诸多因素的影响
- 国内PC设备，10个小时内，UIP与UHID增量关系约1.214稳定，24小时误差为+-4%
- 策略路由的大量应用，单用户会对应多个IP地址

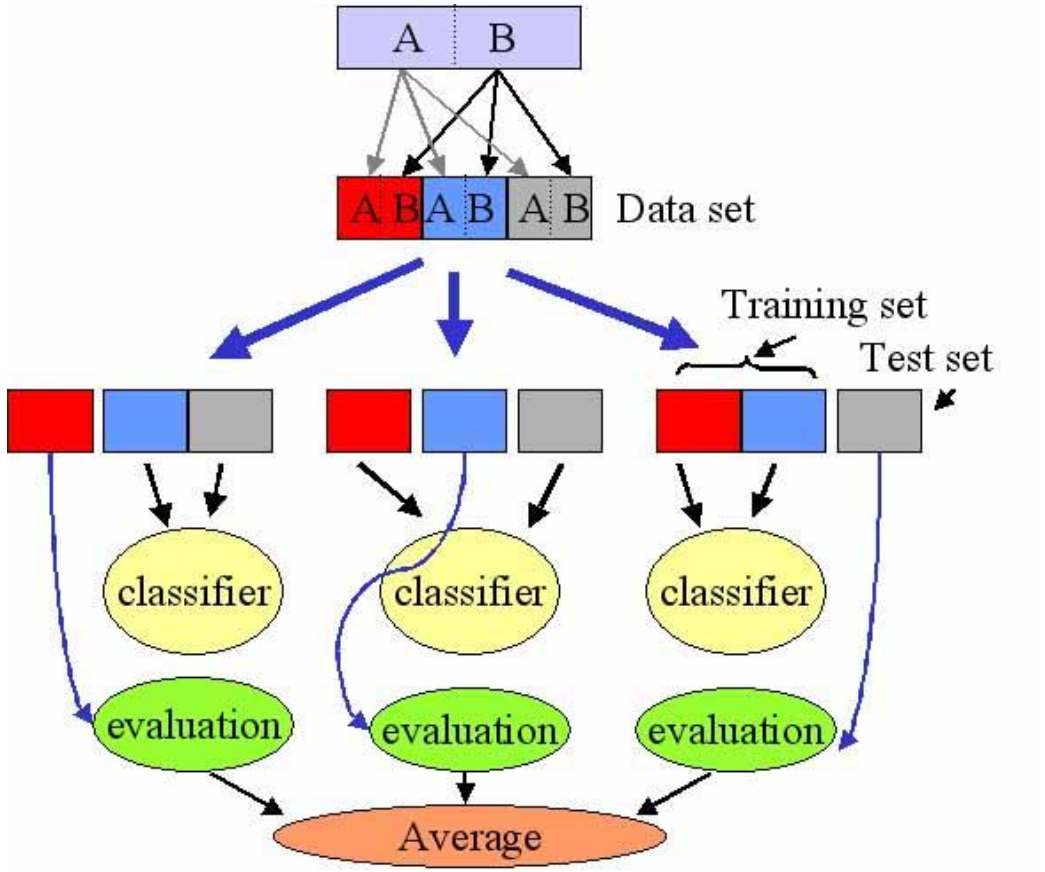
易学难精的文本挖掘

- 从最简单的关键词Like统计，到复杂的句义驱动、近义词聚类，最终到事物、情景和情绪识别都可以称为文本挖掘
- 识别精度和覆盖度两项都必须足够高，并且辅助功能完善，简单易用的产品非常少
- 这个领域我们一直希望能找到靠谱的技术或外包服务商

广义 12090 (条)
购买 7969
使用 7725
效果 7417
物流 7003
正品 6068
价格 5405
服务 4663
其他 3504
赠品 2844
质量 2020
外观 1894
味道 490



定制化分析项目如何验证数据



- 验证数据的方法很多，例如常见的十折验证，但在现实工作中，需求每天都在变化中
- 实事数据要具备最基本的条件统计分析能力支持验证工作，因为80%的错误都是计算过程出错
- 实践中难的用到十折这样严谨的工作流，但可以通过抽样条件，将数据分为三份与合计值对比
- 计算过程要自动化，并且保留每一步的中间数据，保证数据反复重算也能得到相同的结果

数据模型对解读数据的价值

- 模型对数据分析的意义，主要是将复杂逻辑隐藏的规律显现出来，使数据易于解读
- 最简单的指数应用，就是几个有逻辑关系的指标，计算组成
- 抽样和加权，其实也是建立数据模型的过程

可视化工具对数据解读的重要性

系统>图>表>数据

- 人类天生对图形接受能力强于文字，尤其是逻辑问题，因此数据解读的观点最好用图来表达
- 优秀的可视化工具，不仅限于仪表盘，要能与数据系统互动，进行条件查询和展示
- 现在企业对数据挖掘，也是随需应变的，可视化系统要能灵活组合应对

内容目录

来自艾瑞和客户的疑问

- 看似数据很多，其实有价值的数据很少，是方向问题还是方法问题？
- 大数据IT系统产品化程度不高，有没有Excel一样的工具产品？
- 统数据仓库技术是否已经淘汰？是否有必要迁往Hadoop？
- 大数据除了营销应用，是否还有其他的应用情景？

- 艾瑞作为第三方研究公司，是否应该有如此“重”的技术投入？
- 数据源和数据买家，是艾瑞商业逻辑的两端，管道化艾瑞准备好了么？

THANKS

选择艾瑞 选择可以信任的合作伙伴

技术副总裁
郝欣诚