

数据分析方法培训

目录

I

数据分析前的思考

II

案例分享

III

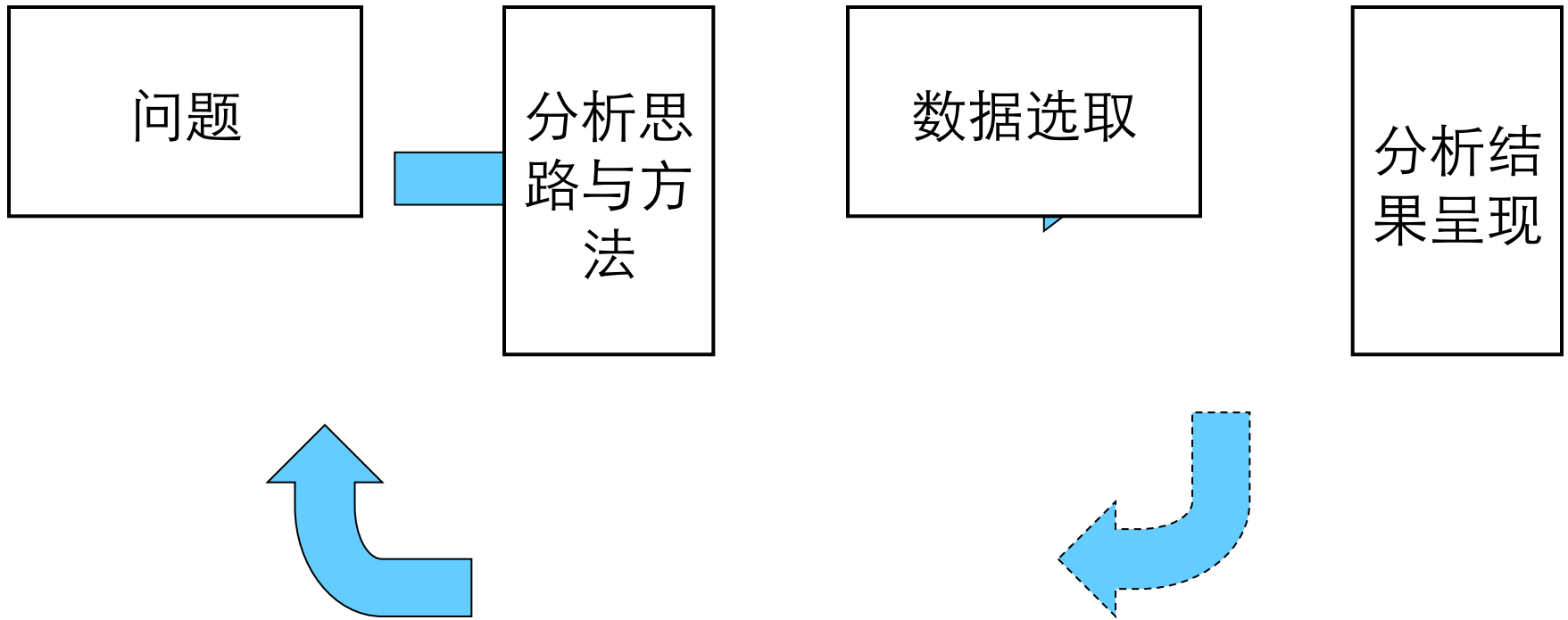
深层次数据分析

数据分析前，我们需要思考

《孙子兵法·谋攻篇》：故**上兵伐谋**，其次伐交，其次伐兵，其下攻城；攻城之法为不得已。

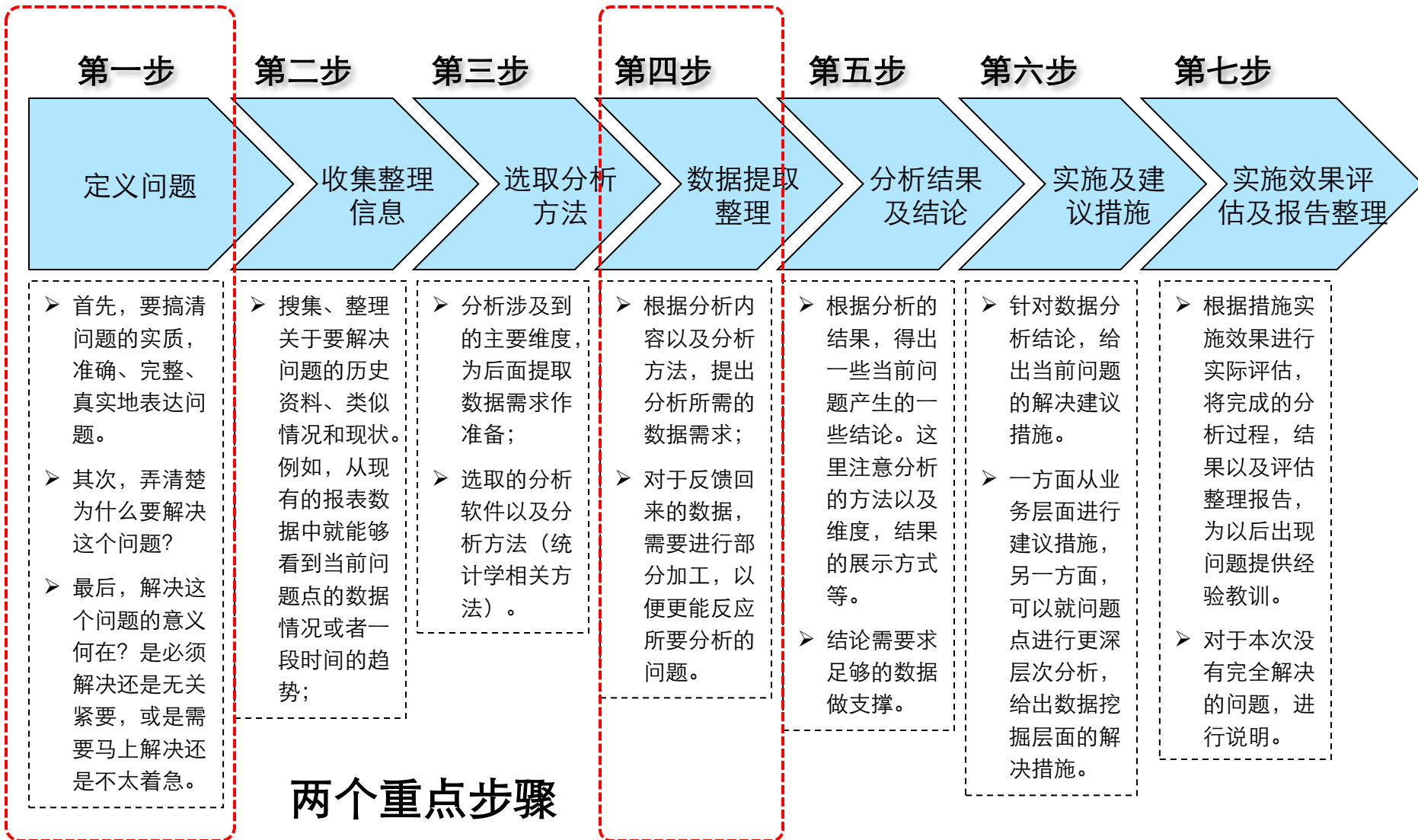
像一场战役的总指挥影响着整个战役的胜败一样，**数据分析师的思想**对于整体分析思路，甚至分析结果都有着关键性的作用。

数据分析前，我们怎么去思考？



每一个步骤可能面临的问题以及需要准备的东西???

分析问题和解决问题的思路



精确的陈述问题



爱因斯坦说：“精确的陈述问题比解决问题还来得重要”

5W2H法：

5W: What, When, Where, Who, Why;

2H: How及How many;

Where----哪里存在问题？

What-----存在的问题是什么？

Why-----原因在哪里？

When-----什么时候开始出现这样的问题？

Who-----与什么对象有关？

How many-----发生的次数和数量？

How much-----损失有多大？

使用这个方法

阿根廷队世界杯输球了，如果你是马拉多纳，你怎么去思考？



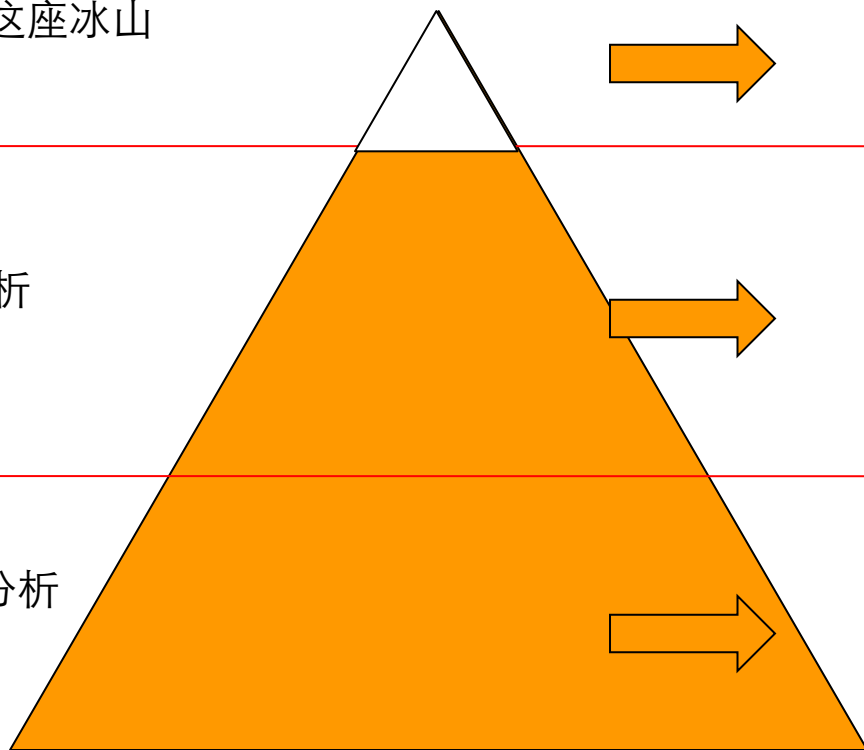
问题展现方式



问题的结构如同这座冰山

初步的问题分析

深层次的问题分析



治标

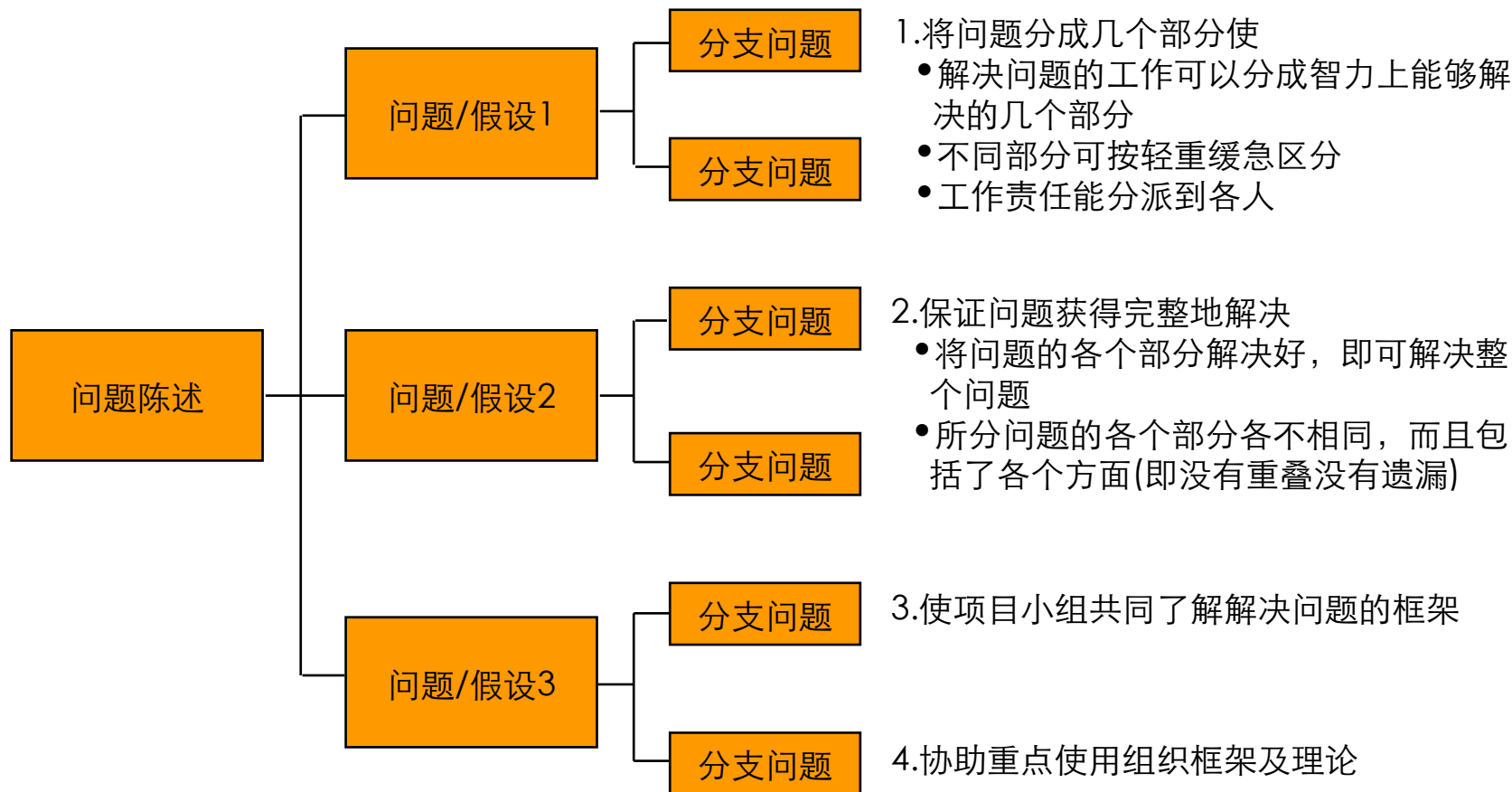
治本

问题结构是由现状、直接原因以及最终原因构成的。针对直接原因进行的叫初步问题分析，针对最终原因进行分析的叫深层次问题分析。

问题分解



为什么使用逻辑树?

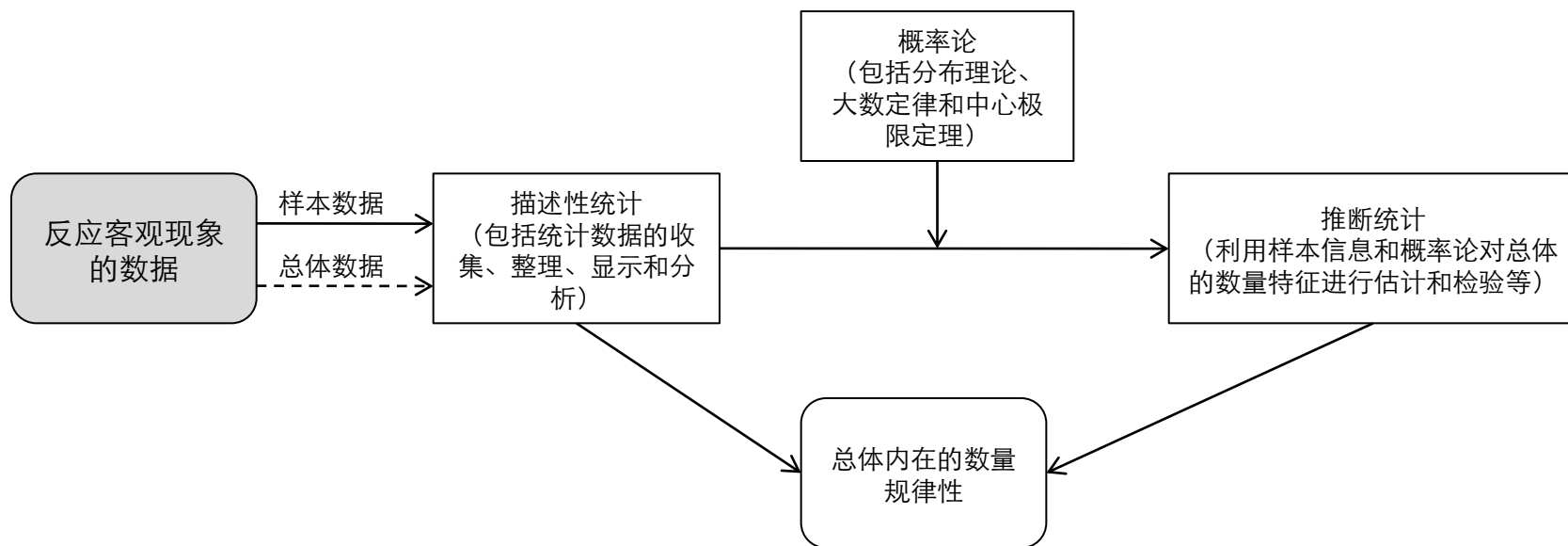


分析方法



统计方法的三大特性，用三句话来简单概括一下：

- **实用性**：除了实情，数据能证明一切；
- **丰富性**：统计就像比基尼，露出来的部分固然诱人，没露出来的部分才是最要命的；
- **公平性**：我们相信上帝，其它人请用数据说话。



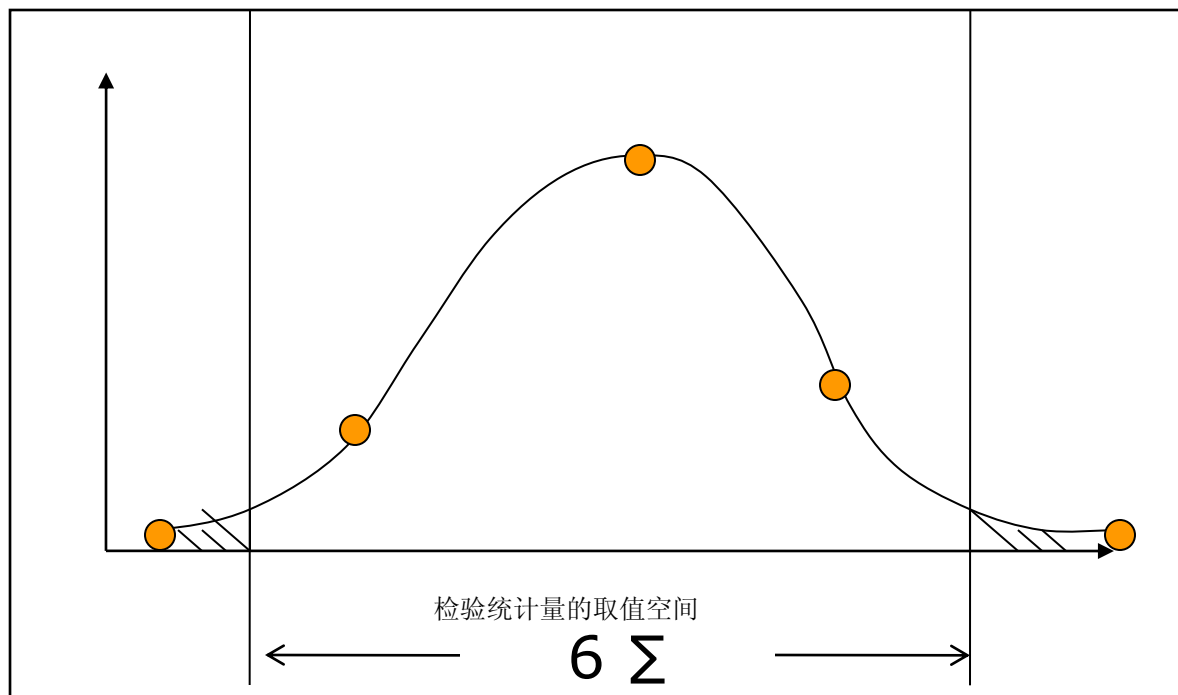
描述性统计分析



“五点法”：最小值，1/4分位数，均值，3/4分位数，最大值

“两度”：峰度，偏度

六西格玛：



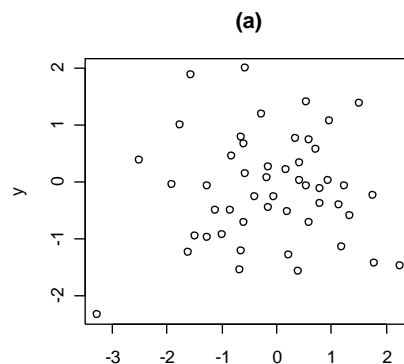
推断统计分析



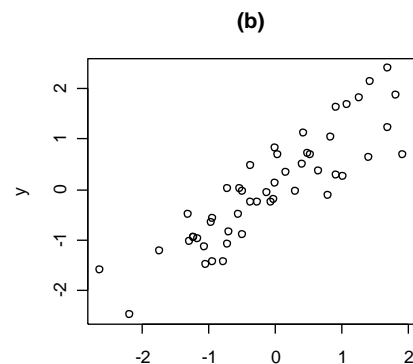
回归分析是统计分析思想中最基础、最集中的一个领域。

高斯、高尔顿

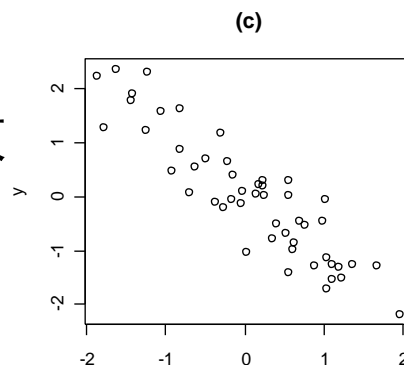
相关分析&回归分析



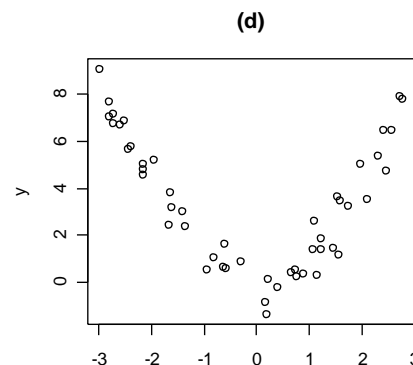
不相关



正相关



负相关



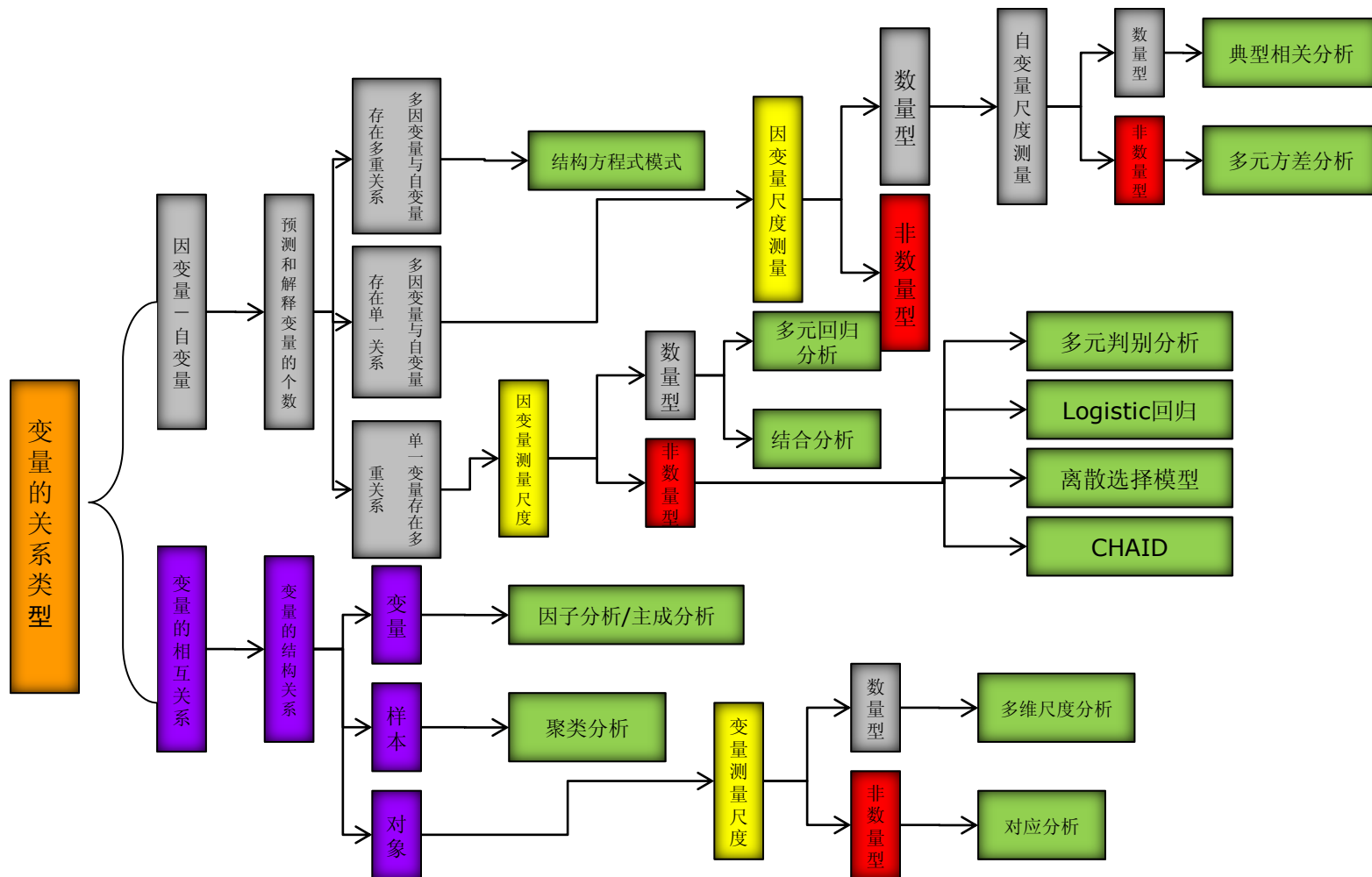
相关但非线性相关

变量的选取;

预测推断;

P值: 回归分析就是放 “P” , 放得好, 就合格。

变量分析方法选取



一张简单的图胜过千言万语！！！！

数据挖掘分析




按挖掘方法分类：包括统计方法，机器学习方法，神经网络方法和数据库方法，其中：

- 统计方法可分为：判别分析（贝叶斯判别、费歇尔判别、非参数判别等），聚类分析（系统聚类、动态聚类等），探索性分析（主成分分析等）等。
- 机器学习方法可分为：归纳学习方法（决策树、规则归纳等），基于范例学习，遗传算法等。
- 神经网络方法可以分为：前向神经网络（BP算法等），自组织神经网络（自组织特征映射、竞争学习等）。
- 数据库方法分为：多为数据分析和OLAP技术，此外还有面向属性的归纳方法。

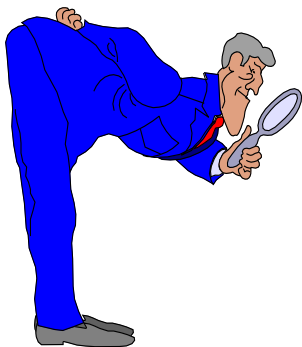
关联规则

关联规则反映一个事物与其它事物之间的相互依存性和关联性，如果两个事物或者多个事物之间存在一定的关联关系，那么其中一个事物就能够通过其他事物预测到。



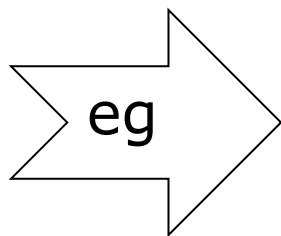
多元统计分析中的聚类分析有个**阈**值，用于确定分类的一个临界值，平时会遇到把它读成fá，误以为它是“**阀**”字。正确的应该是阈（念yù）值，而不是阈值。

选取分析所需的相关数据



数据提取时注意的几点问题。

海量的数据



- 经分数据
- BOSS数据
- 网管中心数据
- CRM数据
- 一经数据
- 第三方调查数据
-

制定数据提取需求



人口统计

- 性别
- 年龄
- 户籍
- 职业
- 婚姻状况
- 教育程度
- 收入
-

行为方式

- 通话时段
- 繁忙和非繁忙通话量
- 漫游服务
- 方便程度
- 行为方式的变化
- ...

态度

- 形象
- 价值观
- 生活方式
- 心理因素
- ...

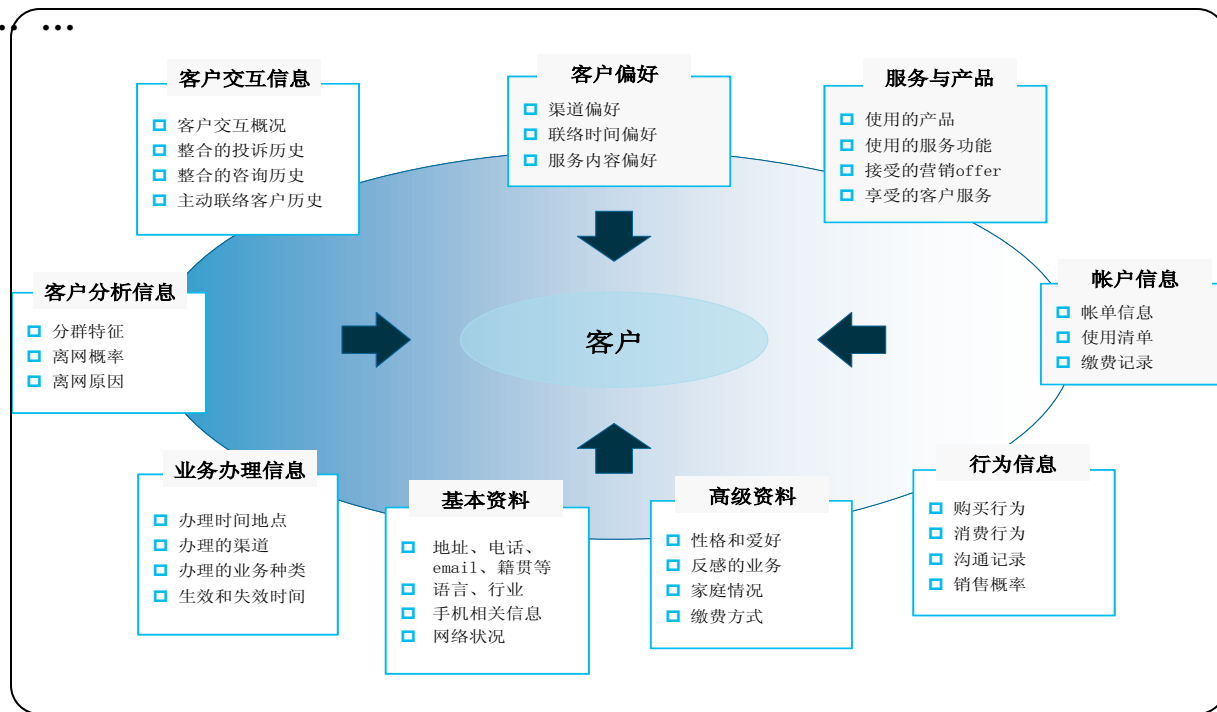
客户价值

- 高利润率
- 中等利润率
- 低利润率
- 负利润率
- ...



Microsoft Office
Excel 工作表

注意数据提取粒度



数据质量的评估



- 在现实社会中，存在着大量的“脏”数据

- > 不完整性（数据结构的设计人员、数据采集设备和数据录入人员）

- 缺少感兴趣的属性
- 感兴趣的属性缺少部分属性值
- 仅仅包含聚合数据，没有详细数据

- > 噪音数据（采集数据的设备、数据录入人员、数据传输）

- 数据中包含错误的信息
- 存在着部分偏离期望值的孤立点

- > 不一致性（数据结构的设计人员、数据录入人员）

- 数据结构的不一致性
- Label的不一致性
- 数据值的不一致性

- > 数据类型冲突

- 性别：string(Male、Female)、Char (M、F)、Integer (0、1)
- 日期：Date、DateTime、String

- > 数据标签冲突：解决同名异义、异名同义

- 学生成绩、分数

- > 度量单位冲突

- 学生成绩
 - 百分制：100 ~ 0
 - 五分制：A、B、C、D、E
 - 字符表示：优、良、及格、不及格

- > 概念不清

- 最近交易额：前一个小时、昨天、本周、本月？

- > 聚集冲突：根源在于表结构的设计

业务角度对于数据质量进行
初步评估！！！！

数据质量的评估



技术角度进行数据质量评估



数据的清洗处理



主要任务：

补充缺失数据

识别孤立点，平滑噪音数据

处理不一致的数据

处理方法：

分箱（Binning）的方法：

聚类方法：

检测并消除异常点

线性回归：

对不符合回归的数据进行平滑处理

人机结合共同检测：

由计算机检测可疑的点，然后由用户确认

... ..

怎样将分析的结果呈现出来？



- 指标分析与政策分析并重；
- 反应重点问题、实事求是；
- 材料、数据要真实，论据要有说服力。

切记……

- 分析角度：缺乏分析中心思想或主干线
- 文字表达：“一图二表三文字”
- 逻辑结构：论点、论据、论证

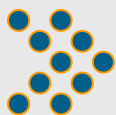
分析结果呈现基本原则



数据分析结果呈现准备工作:

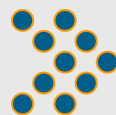
确定表达的主题

- 使用图形的目的: 将思想和观点形象化的表达, 加深读者或听众的印象。
- 使用图表时, 必须明确通过图表要表达的信息是什么。



确定对比关系

- 同一类别不同项目间的对比
- 不同类别不同项目间的对比
- 时间对比: 把时间作为项目分类的标准
- 频率对比: 以部分占整体的百分比为项目分类的标准
- 相关性对比: 按照项目之间的函数关系作为项目分类的标准
- 其他对比: 逻辑关系的对比(因果, 时间序列……)



选择图形

- 饼图
- 柱状图
- 百分比柱状图
- 堆积柱状图
- 线形图
- 雷达图
- 面积图
- 点图
- 气泡图
- 矩阵图
- 逻辑图

如何用图来表示数据?

定义问题

收集整理
信息

选取分析
方法

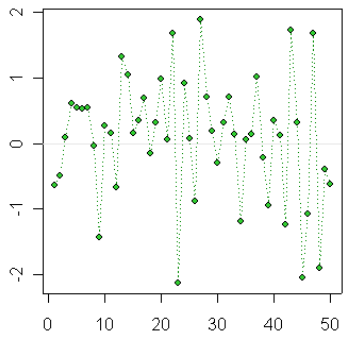
数据提取
整理

分析结果
及结论

实施及建
议措施

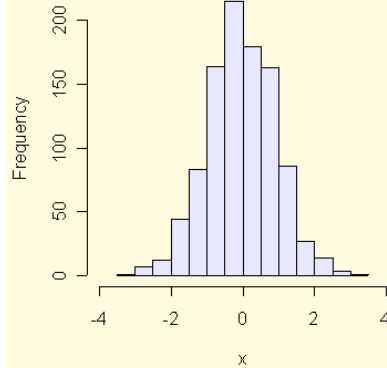
实施效果评
估及报告整理

Simple Use of Color In a Plot

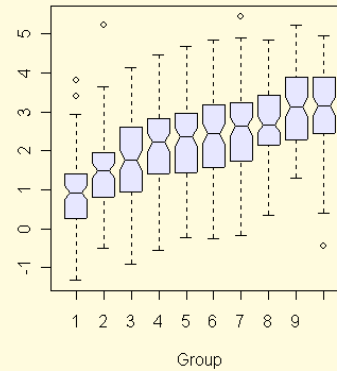


Just a Whisper of a Label

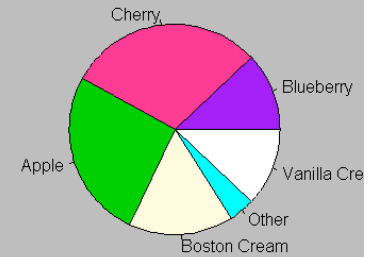
1000 Normal Random Variates



Notched Boxplots

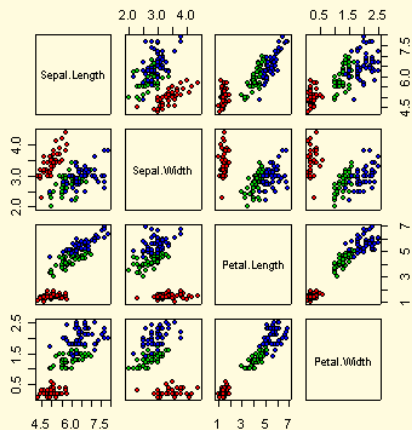


January Pie Sales



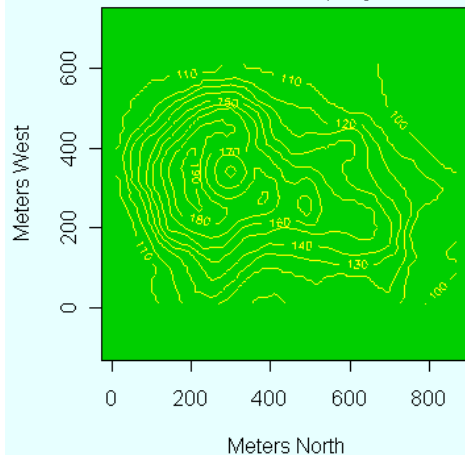
(Don't try this at home kids)

Edgar Anderson's Iris Data

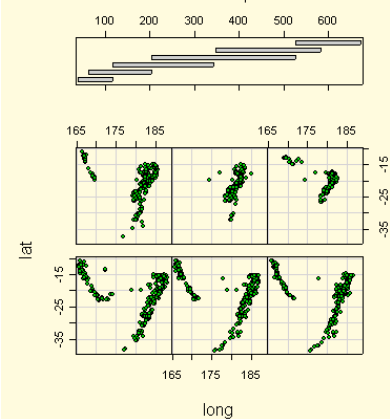


A Topographic Map of Maunga Whau

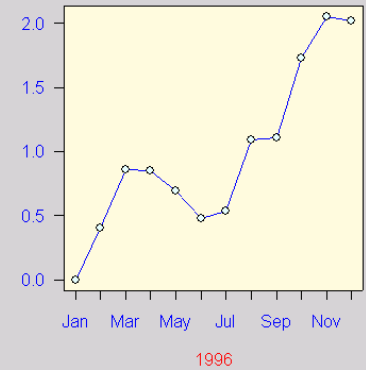
10 Meter Contour Spacing



Given : depth



The Level of Interest in R



定量数据的图表示



- 对于一个定量变量;
- 用图形来表示这个数据, 使人们能够看出这个数据的大体分布或“形状”的一个办法是画直方图(histogram)。

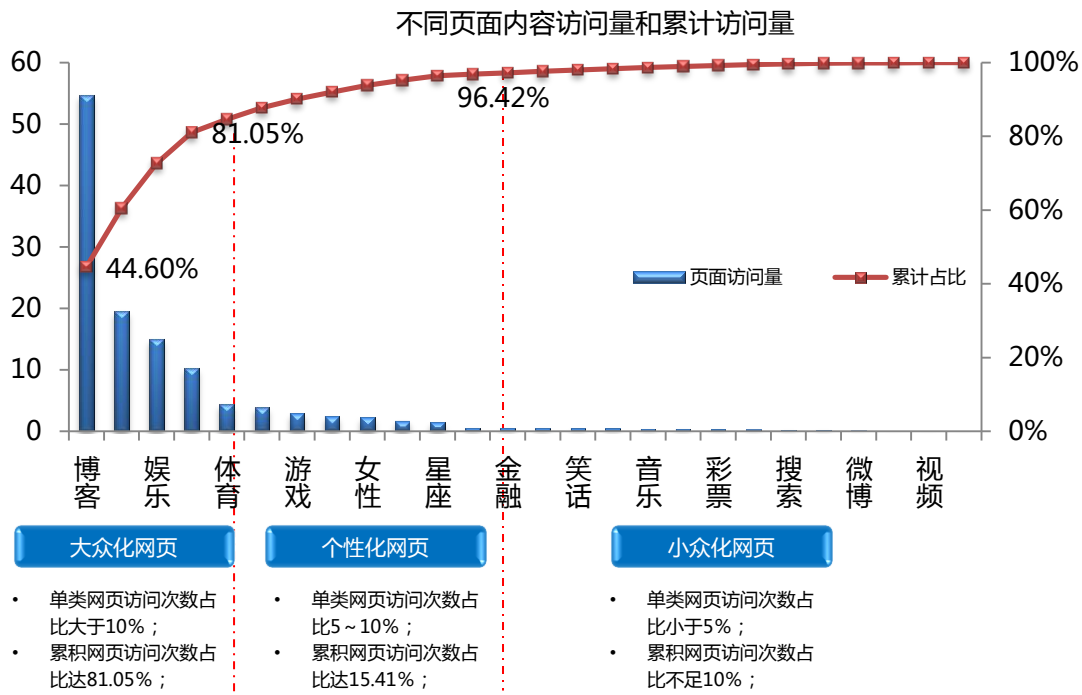


Microsoft Office
Excel 97-2003 工作簿

定性数据的图表示



定性变量（或属性变量，分类变量）不能点出直方图、散点图或茎叶图，但可以描绘出它们各类的比例。



Microsoft Office
Excel 97-2003 工作簿

常见的分析模式



内容决定形式，形式服务于内容，当形式经过实践考验被普遍接受后就固化成一种模式。

分析报告的模式主要包括：

- 金字塔式;
- 综合式;
- 三步曲;
- 专题式;
- 通报;
- 简报式;
- 工作汇报式.

分析总结及建议措施



建议措施分类

- 业务层面
- 数据挖掘

现状及问题

group by: Country



数据分析

总结

分析总结

针对问题1
建议措施

针对问题2
建议措施

针对问题3
建议措施



实施效果评估及报告整理



➤ 营销活动效果反馈数据，分析对于问题的解决程度；

- ✓ 活动历史响应数据的积累；

- ✓ 活动流程固化；

- ✓

➤ 业务模型优化提升；

- ✓ 对比组，显示模型本身的优越性；

- ✓ 营销活动数据对于模型的提升情况；

- ✓

回顾一下

分析前的思考？ ？ ？ ？



目录

I

数据分析前的思考

II

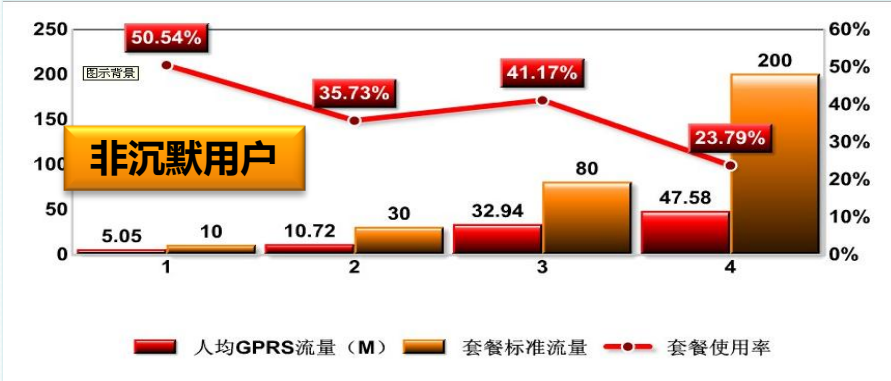
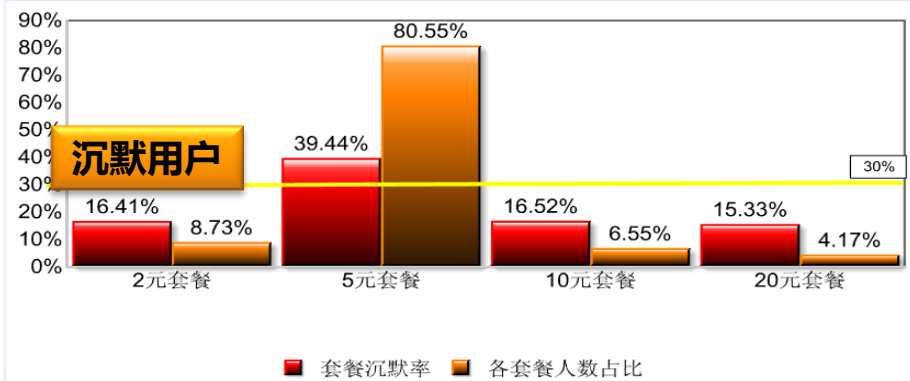
案例分享

III

深层次数据分析

手机上网当前遇到的问题——“一高两低”

按沉默用户和非沉默用户分析



注：沉默用户指套餐沉默用户，由于2元、5元、10元、20元这四大套餐用户占总套餐用户的85%（5月数据），故取四大套餐为研究对象。各套餐人数占比=套餐用户数/四大套餐用户总数；套餐使用率=人均套餐使用量/套餐包含的标准流量。

四大套餐沉默率高

占套餐用户总人数80.55%的5元套餐沉默率为39.44%，高于当前套餐沉默率指标30%。其它套餐虽然沉默率低，但总人数也低，故降低5元套餐沉默率是当前急需解决的问题。

怎样降低套餐沉默率

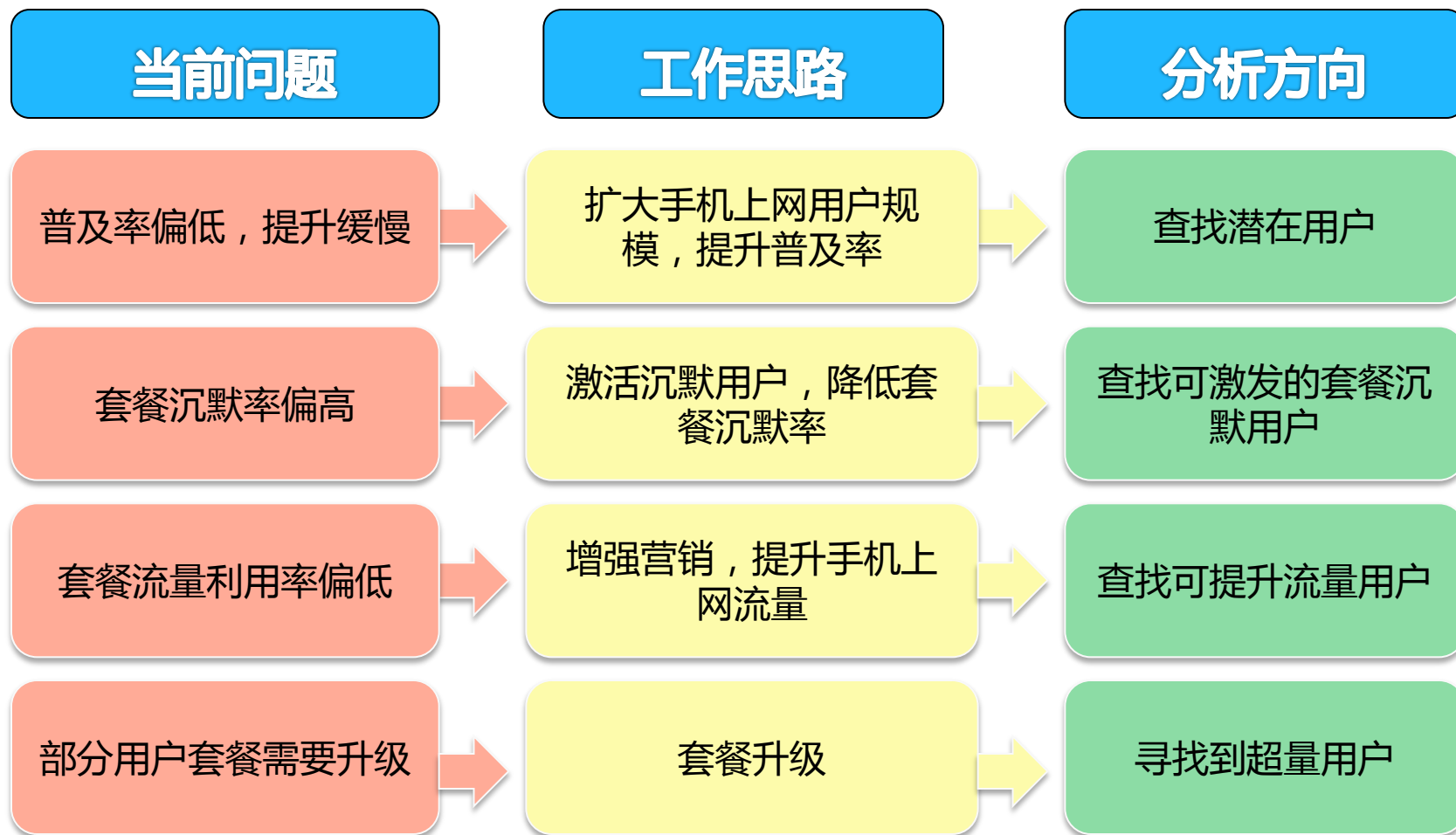


非沉默用户人均流量低

四大套餐非沉默用户人均流量均远低于套餐包含的标准流量，而人数最多的5元套餐（含30MGPRS流量）人均流量也只有10.72M，是套餐可使用量的35.73%。

怎样提升套餐均流量

手机上网问题分解及用户定义



相关分析数据字段提取

基本属性

- ◆手机号码
- ◆ 品牌
- ◆ 付费类型
- ◆imei号
- ◆终端是否支GPRS
- ◆入网时间
- ◆年龄
- ◆性别
- ◆ARPU

数据业务

- ◆点对点短信上行条数
- ◆ 梦网短信条数
- ◆点对点彩信上行量
- ◆短信计费量
- ◆新业务费用
- ◆是否是转转赢用户
- ◆是否是大赢家用户
- ◆是否飞信用户
- ◆是否无线音乐高级会员
- ◆定制手机报类型
- ◆彩铃主动下载次数
- ◆是否使用手机搜索

GPRS

- ◆GPRS套餐类型
- ◆GPRS流量
- ◆GPRS费用
- ◆CMWAP流量
- ◆CMNET流量
- ◆cmwap登陆次数
- ◆cmnet登陆次数

语音业务

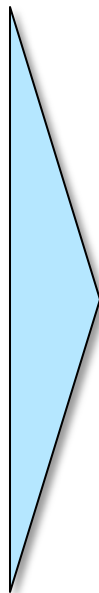
- ◆语音业务费用
- ◆ 本地通话时长
- ◆本地通话次数
- ◆总打入号码数
- ◆总打出号码数

查找潜在用户(略去)

沉默用户流量提升分析总结及建议措施

分析总结

- 手机上网整体普及率较低，仅33%；
- 近半年多来手机上网普及率提升较慢；
- 动感地带品牌对于手机上网接受程度最高；
- 手机上网业务粘性较差；
- 手机上网与特定业务订购有很强关联性；
- 5元套餐是提升重点；



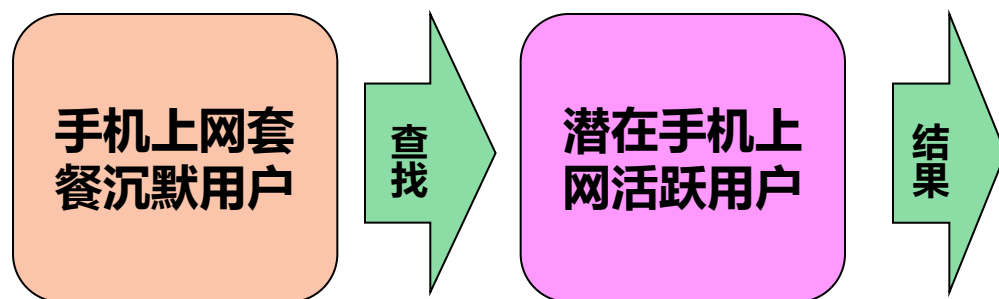
建议措施

手机上网潜在用户查找建议措施：

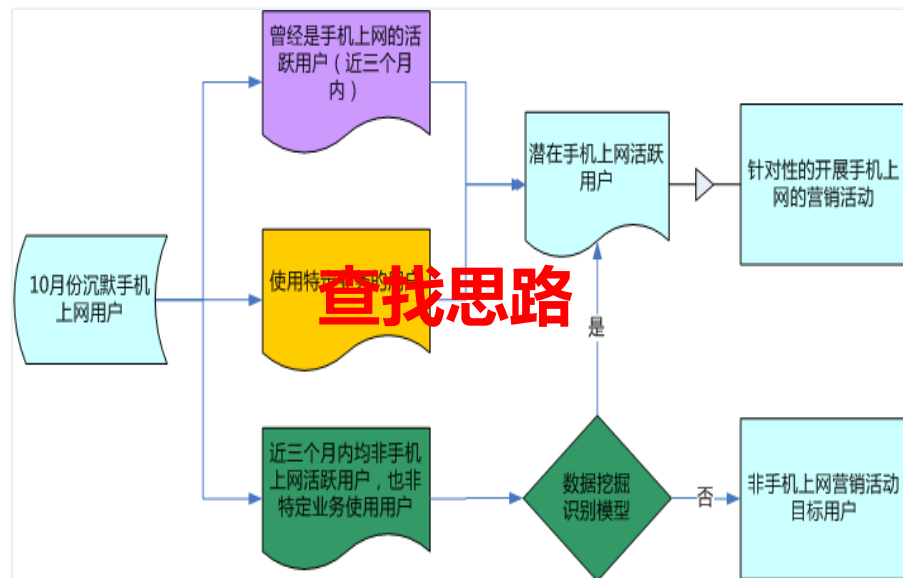
- 较低的普及率为手机上网潜在用户查找提供了上升的空间；
- 上月或上上月是手机上网活跃用户；
- 使用特定业务的手机上网沉默用户；
- 加强手机上网业务体验营销和手机上网助手业务的宣传；
- 加强手机上网用户主动偏好需求研究。

对于不满足建议查找条件的非手机上网用户，利用数据挖掘技术进行查找。

手机上网潜在活跃用户查找



非手机上网营销目标用户：
79.2%



目标用户提取规则

示例

目录

I

数据分析前的思考

II

案例分享

III

深层次数据分析

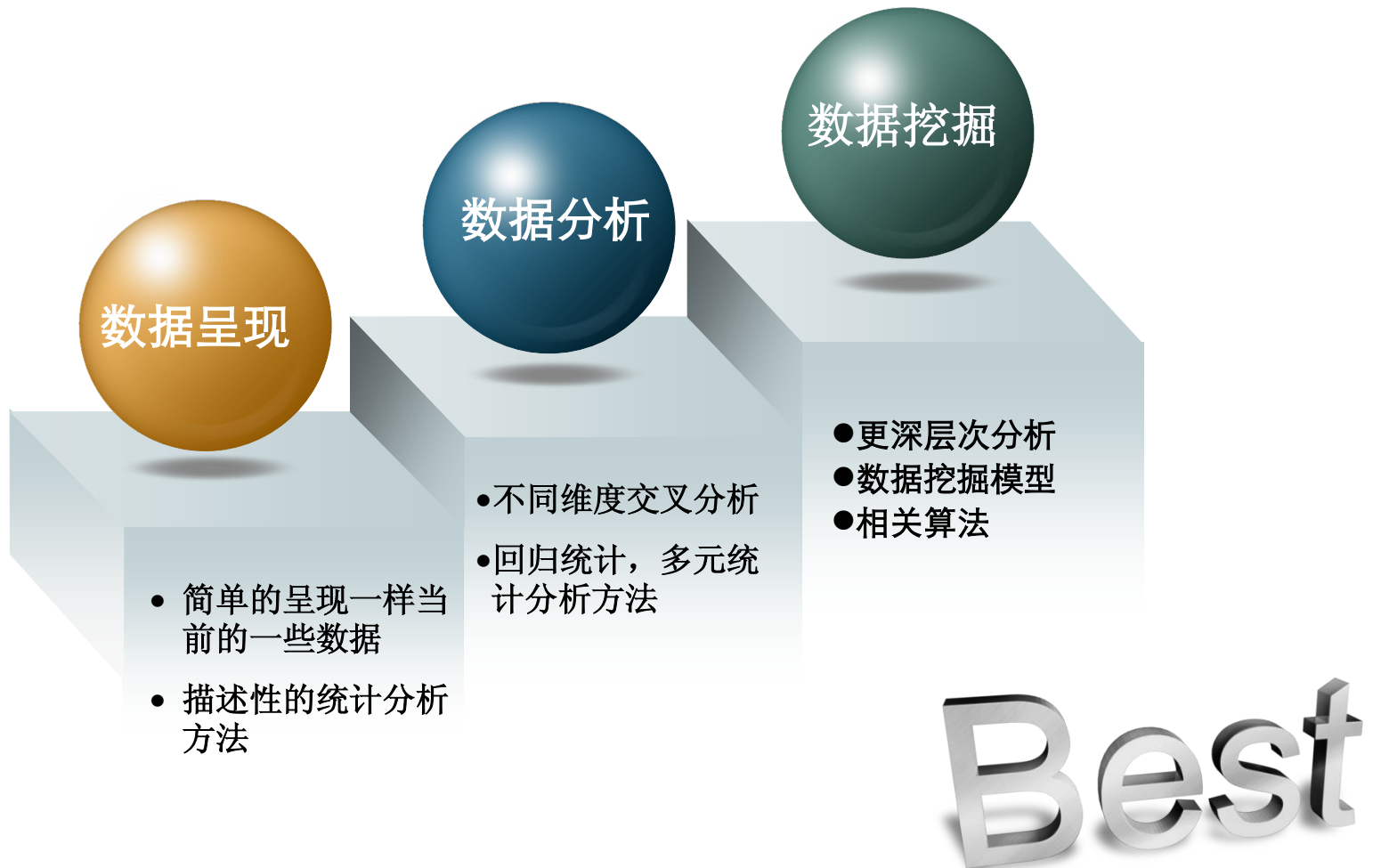
这样的客户需求，我们怎么处理？

用户和需求：对于复杂现象的简单结论

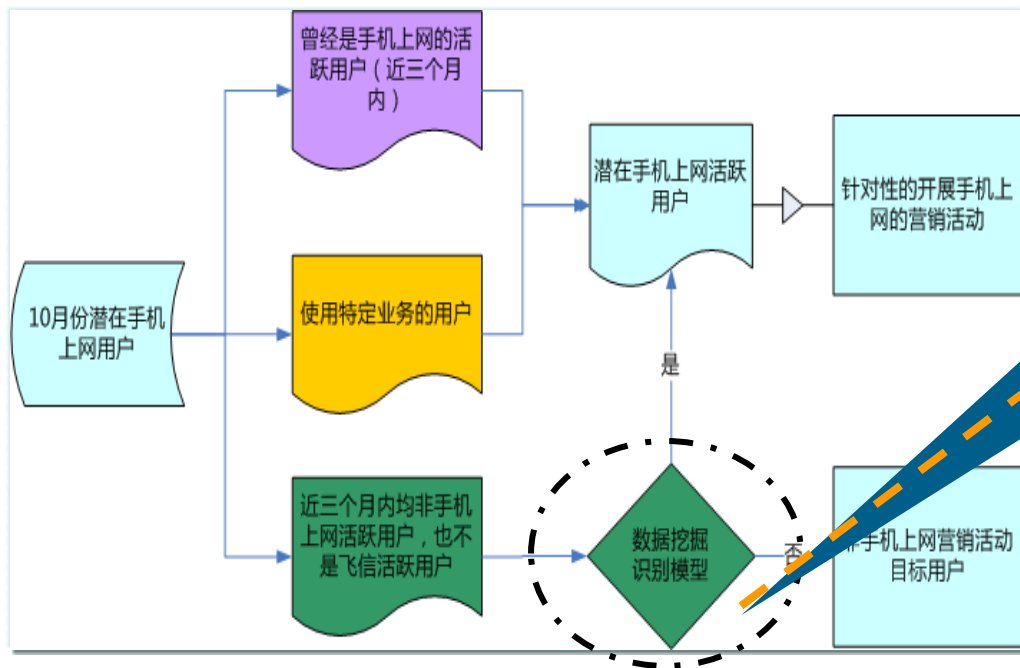
- 市场—谁将会购买这个产品？
- 预测—我们将面临何种需求？
- 忠诚度—谁最有可能流失？
- 信用—哪一类人群不还款的倾向严重？
- 欺诈—什么时候会发生？

当然这些问题，从业务角度，能够有一定的回答，但是，如果有更深层次的分析，会得到比业务层面更好的效果！！！！

数据分析与数据挖掘的关系



更深层次的分析

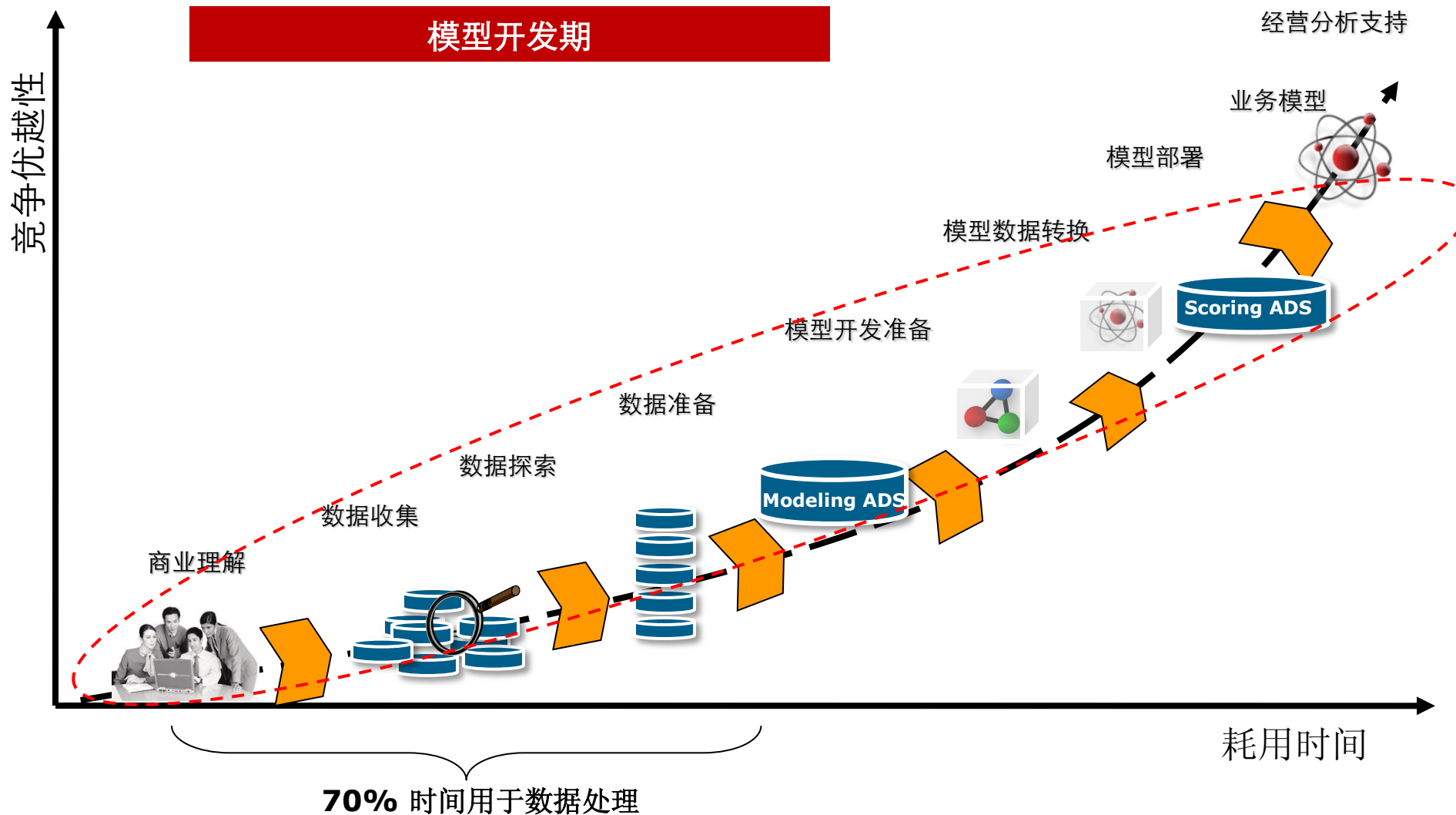


基于数据分析之上的数据挖掘

➤数据挖掘的重点和难点是什么？

数据挖掘是数据分析的一个环节，同时是解决实际问题的一个环节，当然，数据挖掘的应用就能体现出来！！！！

数据挖掘过程



Thank-you