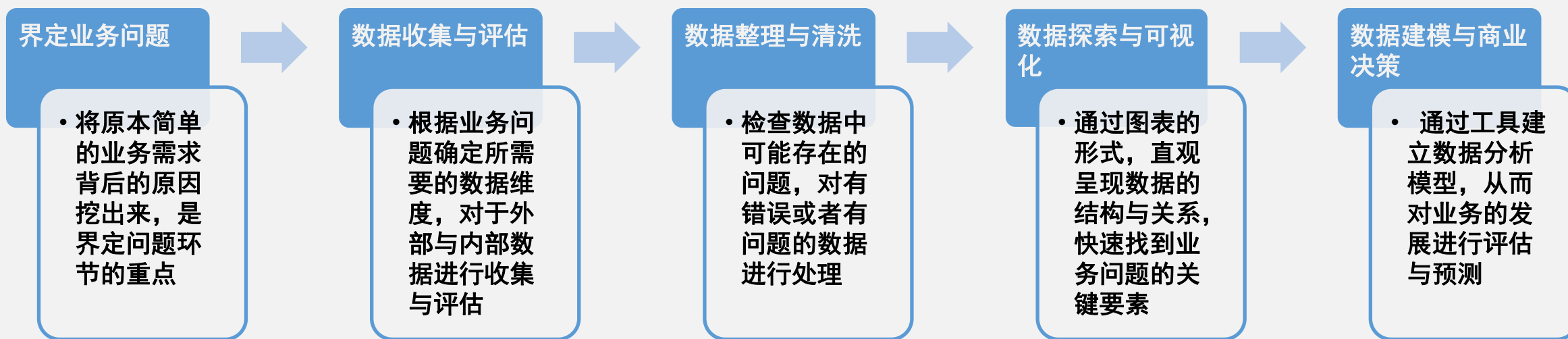




数据分析

第1课：3分钟了解数据分析工作全流程

数据分析工作最重要的5个环节



数据分析工作最重要的5个环节

Step 1 界定业务问题

将原本简单的业务需求背后的原因挖出来，是界定问题环节的重点

以宜家为例：

➤ WHAT（什么发生了？）

如：“我们这个月的收入和利润如何？”

➤ WHY（为什么会发生？）

如：“为什么顾客购买的平均单价在下降？”

➤ HOW（我们能做什么让它变得更好？）

如：“我们能让他们在宜家店里多花钱？”



数据分析工作最重要的5个环节

Step 2 数据收集与评估

根据业务问题确定所需要的数据维度，对于外部与内部数据进行收集与评估

- 我们需要哪些数据？
- 这些数据可以如何获得？
- 这些数据是否存在异常或者缺失



数据分析工作最重要的5个环节

Step 3 数据整理与清洗

检查数据的中可能存在的问题，对有错误或者有问题的数据进行处理

- 如何快速查询我所需要的数据？
- 如果数据存在缺失应该怎么做？
- 如果数据错误应该怎么做？



数据分析工作最重要的5个环节

Step 4 数据探索与可视化

通过图表的形式，直观呈现数据的结构与关系，快速找到业务问题的关键要素

- 使用怎样的图表呈现数据的趋势？
- 使用怎样的图表呈现数据的分布？
- 使用怎样的图表呈现数据的相关性？



数据分析工作最重要的5个环节

Step 5 数据建模与商业决策

通过工具建立数据分析模型，从而对业务的发展进行评估与预测

- 哪一种数据模型可以预测销售额未来的趋势？
- 哪一种数据模型可以预测消费者的购买行为？
- 哪一种数据模型可以构建用户画像？



宜家IKEA案例

宜家的挑战

行业角度：

- 宜家是在开创以平实价格销售自行组装家具的领导品牌
- 国内市场良莠不齐，宜家品牌优势收到冲击

消费者角度：

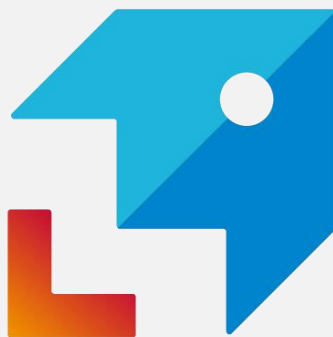
- 不同顾客群体的购买习惯和关注产品差异很大
- 价格敏感度也决定了他们是否认可宜家的品牌定位



用数据深度理解客户的态度和行为成了重要的竞争优势

- 宜家为代表的实体零售是最早累积数据的领域
- 宜家从创立之初就建立了会员体系，已经拥有了1800万的注册会员
- 线下门店和线上会员运营的O2O结合，为宜家这样的企业转型为
新零售模式提供了坚实的数据基础





数据分析

第2课： 什么是一个好的问题？

界定业务问题

“提出一个好的问题可以说问题解决了一半”

分析问题大体可以分为下面三类问题：

- WHAT类问题：通常围绕企业所关注的重点指标
- WHY类问题：关注业务现状背后的原因
- HOW类问题：通常有业务方面的需求



WHAT类问题

通常围绕企业所关注的重点指标

对于企业，会关注用户类指标与收入类指标

➤用户类指标

- 每月用户总量是多少？
- 每月有多少新增用户？
- 每月活跃用户有多少？

➤收入类指标

- 每月销售额总量是多少？
- 每月销售额增长了多少？



WHAT类问题

数据分析师应做的是：

- 观测这些指标的当前值，**监控**他们本周，本月，本年的表现，
以及与上个月或者上个季度或者去年同期的**比较**
- 将重要指标汇总到报表中，让系统可以自动化的**定期更新**，
帮助业务方和决策者可以随时了解企业的经营状况



WHY类问题

关注业务现状背后的原因

➤问题：

- 为什么这个月的新增用户下降了8%？
- 为什么这个月的销售额达成下降了5%？
- 为什么用户平均消费的单价变低了500元？

➤原因：

可能涉及到经济，市场和其他等因素



HOW类问题

当我们找到背后的原因，怎样能够让关键指标提升？

- 我们如何识别高价值用户（消费更多的用户）？
- 我们如何向高价值的用户定向推荐他们喜欢的产品？
- 我们如何让高价值的用户进行更多消费？



HOW类问题

针对HOW类问题，“我们能让他们在宜家店里多花钱？”

数据分析师可以做的是：

➤ 精准营销

业务问题实际上：

“识别高价值顾客，实施针对性的营销方案，发放专属优惠券促进顾客购买”

1. 定义高价值顾客，并从数据中识别高价值顾客的具体特征
2. 选择那些在收到优惠券之后使用的顾客，将优惠券推送至会员卡，
并跟踪分析后续的使用和购买情况



定义好问题和分析计划的步骤

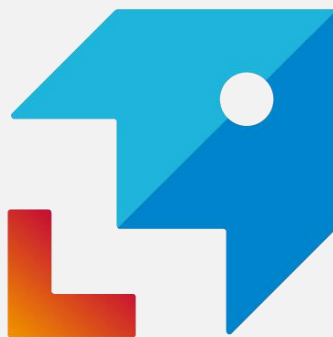
- 首先，与关键的业务人员一同参与，识别能用数据驱动业务发展的关键环节（WHAT→WHY→HOW）
- 其次，根据业务目标找到所有可能解决的技术方法与所需数据，选择分析计划
- 最后，明确衡量分析项目成功的标准



界定业务问题

	解决方法	交付方式
What类问题	以在数据库中抽取，拼接，聚合为主	Excel或者Tableau可视化报表
WHY类问题	从数据中找出洞见，在WHAT呈现的报表基础上做简单的数据探索和分析	以PPT或者文档方式撰写结论报告
HOW类问题	提出对应的分析模型解决	模型结论和实施方案





数据分析

第3课： 如何获取你想要的数据？

数据收集与评估

基于前面所设定的**数据分析问题和计划**，
在此步骤中，我们将**收集**后续分析所需的原始数据，
并进行基本的数据质量**评估**。

- 在收集过程中，需要注意各数据源的格式以及相关关系
- 在评估过程中，主要是数据的完整性，准确性和及时性



数据收集的分类

➤ 广义的数据收集

企业将有关自身利益的**各类内部和外部数据**纳入到数据库的系统流程

➤ 狭义的数据收集

从数据库的多个数据表中进行**抽取，拼接，聚合**的工作，
以形成解决问题所需的数据集的过程

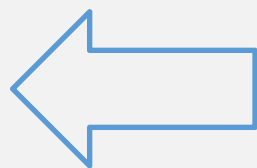


数据收集

要回答的问题：“如何找到高价值用户？”

什么是高价值的用户呢？

- 他今年多大了？他是男生还是女生？
- 他在我们这里总共花费了多少钱？
- 他最喜欢购买的产品是什么？
- 他最喜欢在哪家门店消费？



从数据表单中
寻找对应的数据维度
来还原高价值用户



数据收集

要回答的问题：“如何找到高价值用户？”

大体可以分为下面三个步骤：

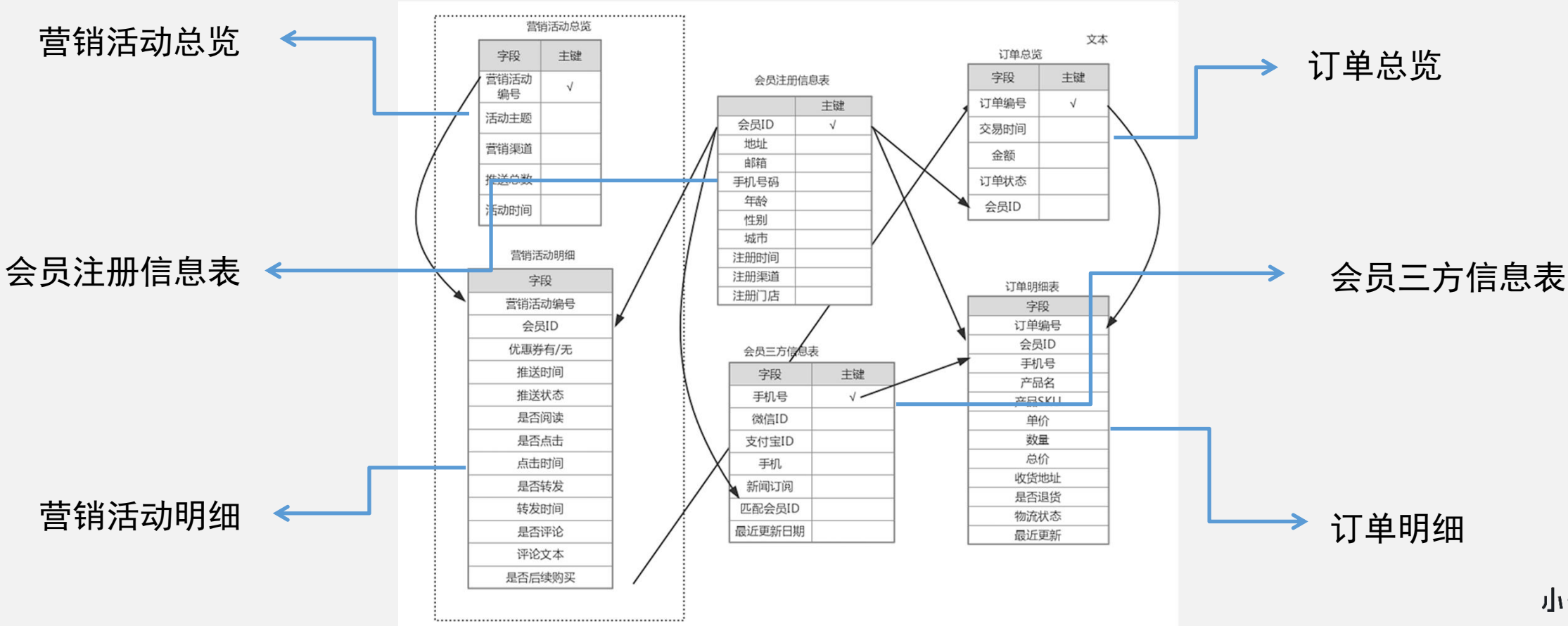
- **第一步：数据收取**，围绕你的问题，提取关键的数据维度
- **第二步：数据拼接**，在对应的表中找到需要的数据维度，
并找到连接表和表的key
- **第三步：数据聚合**，将你需要的数据进行拼接与聚合，
形成新的表



数据收集——数据库概述

➤ 以宜家数据库的简化版本为例

构成： 营销活动信息 会员信息 交易订单



数据收集

要回答的问题：“如何找到高价值用户？”

提取用户的年龄、性别、购买金额、产品偏好与门店偏好这几个关键维度，
从而建立如图表格结构

用户ID	性别	年龄	购买金额	产品偏好	门店偏好
1001					
1002					
1003					
1004					



数据拼接

要回答的问题：“如何找到高价值用户？”

在会员表、产品表与销售表中，找到对应的数据维度进行拼接

表和表之间的关联关系：

- ▶ 一对一
- 一对多
- 多对多

用户ID	性别	年龄	地区
1001			
1002	会员信息注册表		
1003			
1004			

用户ID	购买产品	购买金额
1001		
1002	会员表	
1003		
1004		

用户ID	性别	年龄	地区	购买产品	购买金额
1001					
1002	以会员ID进行拼接没有重复				
1003					
1004					

数据拼接

要回答的问题：“如何找到高价值用户？”

在会员表、产品表与销售表中，找到对应的数据维度进行拼接

表和表之间的关联关系：

一对一

一对多

▶ 多对多

订单ID	产品	金额	时间
1001	A	100	
1001	B	50	
1003			
1004			

订单信息表

营销活动	购买产品	订单ID
2001		1001
2004		1001
2005		
2006		

营销活动表

订单ID	营销活动ID	金额			
1001	2001	100			
1001	2004	50			
1001	2001	100			
1001	2004	50			

营销活动ID和订单ID
都不是各自表的唯一值
订单金额多次重复

避免多对多的情况

数据拼接

要回答的问题：“如何找到高价值用户？”

在会员表、产品表与销售表中，找到对应的数据维度进行拼接

表和表之间的关联关系：

一对一

▶ 一对多

多对多

订单ID	产品	金额	时间
1001	A	100	
1002	B	50	
1002	C	120	

订单表

用户ID	性别	年龄	地区
1001	M	35	
1002	F	50	

会员表

用户ID	性别	年龄	订单ID
1001	M	35	1001
1002	F	50	1002
1002	F	30	1002

一个会员ID对应多个订单ID



数据拼接

要回答的问题：“如何找到高价值用户？”

在会员表、产品表与销售表中，找到对应的数据维度进行拼接

表和表之间的关联关系：

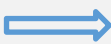
一对一

▶ 一对多

多对多



原因



会员会多次发生购买行为



解决

定义：第一次购买时间、第一次购买商品等数据维度

用户编号	第一次订单金额	第一次订单时间	累积订单金额	累积购买次数
1001	200	2019/02/12	500	2
1002	5000	2018/06/15	12500	4
1003	350	2018/10/2	3600	1
1004	60	2018/01/20	300	3



数据聚合

要回答的问题：“如何找到高价值用户？”

将你所需要的数据进行聚合，形成最后的数据表

用户编号	第一次订单金额	第一次订单时间	累积订单金额	累积购买次数
1001	200	2019/02/12	500	2
1002	5000	2018/06/15	12500	4
1003	350	2018/10/2	3600	1
1004	60	2018/01/20	300	3

基于以往工作经验或行业通用的分析框架，对数据进行聚合，
可固化在SQL代码中形成标签化工具

数据评估

在数据收集过程前，需要评估各个数据源的完整性和及时性

➤完整性检查：

原始数据不存在和已知业务常识之间的明显差距

➤及时性检查：

确保各数据源都反应的是相同时间窗口数据



数据评估

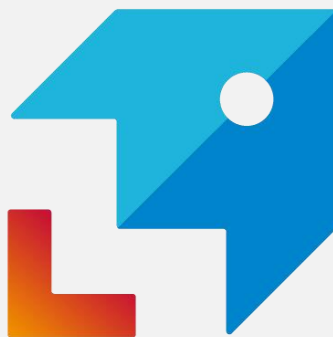
在分析数据集后，需要评估汇总数据的准确性和一致性

➤ 注意分析数据和原始数据的差异

➤ 对于关键的数据可以进行简单的统计查询

如：用户年龄的最大最小值，订单金额的最大值最小值和分布等





数据分析

第4课：如何识别数据中的缺失与异常？

你获取的数据，都有哪些问题？

数据缺失

数据不规范

客户ID	性别	收入	城市	累积购买总额	历史优惠券兑换记录
1001	M	75,000	北京市	5000	1
1002	F	-40,000	上海	6000	0
1003		60, 000	武汉	6000	0
1004	M	50,000	内蒙	8000	1
1005	F	99,999	成都市	20000	0
1006	M	30, 000	贵阳	15000	1
1007	F	25, 000	北京	8000	0
1008	M	19, 000	北京	6000	0
1009	F	10,000,000	上海	5000	1
1010	F	35, 000	上海	3000	0

数据异常

数据清洗与整理

在日常的数据分析工作中，发现数据中的问题，并将数据进行调整，就是非常重要的**数据清洗与整理**环节

➤ **数据清洗：将有问题的数据排除出去**（最需要时间和精力环节）

这里主要涉及对数据缺失，异常和其他问题的处理

➤ **数据整理：将数据转化成更有助于后续分析的样式**

如：将实际年龄转换成80后，90后这样有业务意义的分组



数据清洗

当发现数据中的缺失与异常值时进行数据处理

- **第一步：检查原表**——先检查原表是否是同样的数据
 - 该数据来自哪个数据表？在原表中也存在该问题吗？
- **第二步：确认问题**——检查此数据是如何收集而来，和业务团队确认
 - 如果在原表中也存在问题，是否数据收集过程发生了问题？
- **第三步：数据清洗**——从技术角度评估该数据是否缺失/异常，以及如何处理
 - 该数据是否存在问题？应该如何处理？



数据清洗——数据缺失

遇到数据缺失时，你可以依次思考以下问题：

➤ 缺失的信息来自于哪个数据表？在原表中它们也是缺失的吗？

遇到会员性别的缺失，首先查找一下会员表中的信息

➤ 如果在原表也缺失，那么是否是有收集信息的疏漏？

如果在原表中也存在问题，是否数据收集过程发生了问题？

➤ 如何处理？

当判定该数据为缺失值时，往往会通过python等工具进行缺失值填充



数据清洗——数据异常

数据异常的3种情况：

有违常识的数据异常

客户ID	性别	收入	城市	累积购买总额	历史优惠券兑换记录
1001	M	75,000	北京市	5000	1
1002	F	-40,000	上海	6000	0
1003		60, 000	武汉	6000	0
1004	M	50,000	内蒙	8000	1
1005	F	99,999	成都市	20000	0
1006	M	30, 000	贵阳	15000	1
1007	F	25, 000	北京	8000	0
1008	M	19, 000	北京	6000	0
1009	F	10,000,000	上海	5000	1
1010	F	35, 000	上海	3000	0

和其他数据不同

离群值：与其他数据在数值上差异较大

数据清洗——数据异常

针对收入99, 999这一异常数据的处理方式：

客户ID	性别	收入
1001	M	75,000
1002	F	-40,000
1003		60, 000
1004	M	50,000
1005	F	99,999
1006	M	30, 000
1007	F	25, 000
1008	M	19, 000
1009	F	10,000,000
1010	F	35, 000

➤ 第一步 检查原表

检查会员表，客户ID1005的收入

➤ 第二步 确认问题

如果收入也是99, 999，询问业务人员

➤ 第三步 数据清洗

判断为数据异常，进行数据清洗



数据整理

为了能够进一步进行数据分析，将进行数据整理环节

- 对数据进行统一的格式化和命名规则处理
- 对某些信息进行重新编码以满足后续分析需求



数据整理——数据标准化

➤ 对数据进行统一的格式化和命名规则处理

客户ID	性别	收入	城市	累积购买总额	历史优惠券兑换记录
1001	M	75,000	北京市	5000	1
1002	F	-40,000	上海	6000	0
1003		60, 000	武汉	6000	0
1004	M	50,000	内蒙	8000	1
1005	F	99,999	成都市	20000	0
1006	M	30, 000	贵阳	15000	1
1007	F	25, 000	北京	8000	0
1008	M	19, 000	北京	6000	0
1009	F	10,000,000	上海	5000	1
1010	F	35, 000	上海	3000	0

数据不规范

- 全名和简称的差异（北京，北京市）
- 内蒙作为省被放入了城市这一栏
- 等等



影响

用户画像分析

要对城市这一个数据进行规范化的标注

数据整理——数据编码

➤对某些信息进行重新编码以满足后续分析需求

订单ID	产品	金额	时间
1001	A	100	2019/2/20 15:48



订单ID	Day	Period	Hour
1001	Workday	Afternoon	2~4pm



对时间维度进行了重新编码

对此可以分析

- 在周末与工作日的销售情况是怎样的？
- 早上/下午/晚上的销售情况是怎样的？
- 哪一个时间段的销量最高？



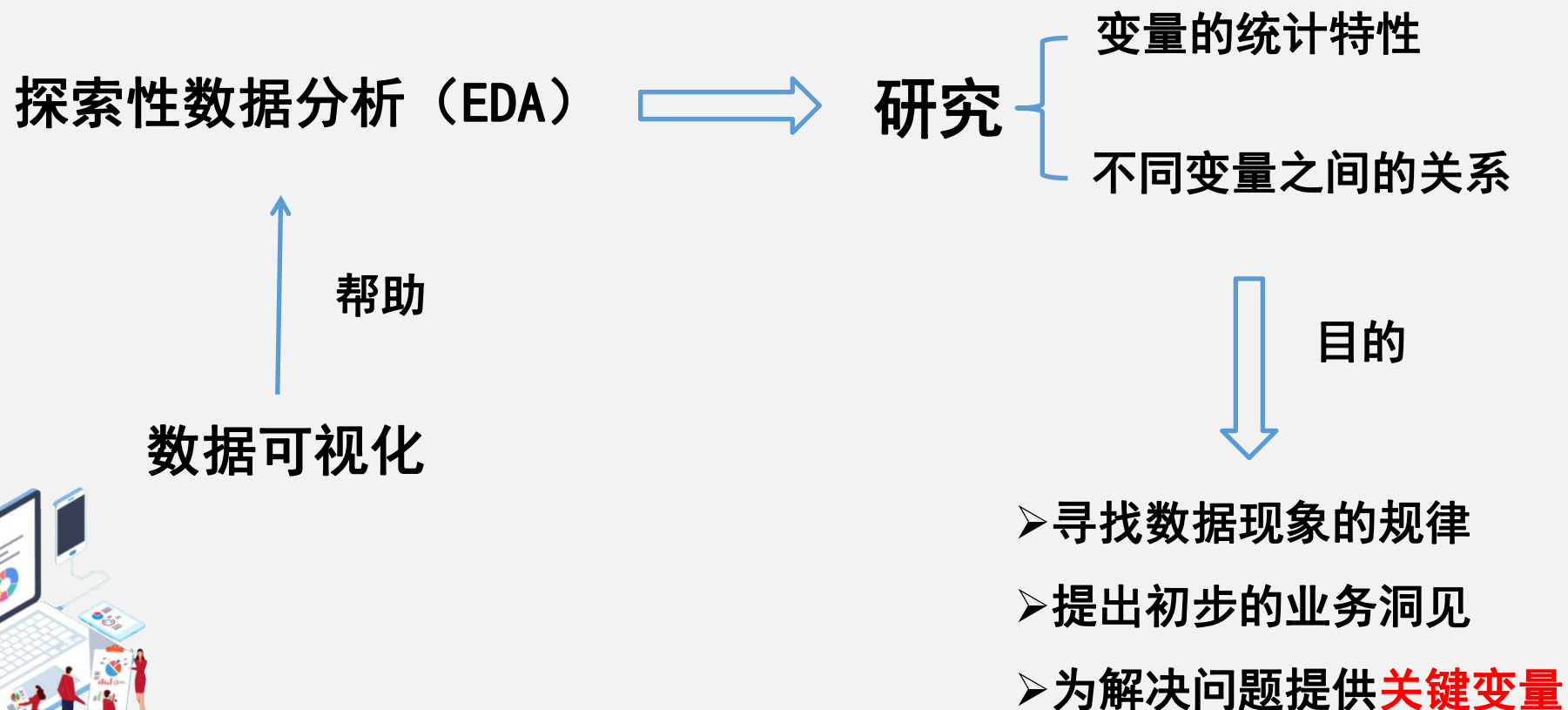


数据分析

第5课：如何用可视化图表呈现关键？

数据探索与可视化

在可视化工具的帮助下，找到数据的底层结构和规律，找到能帮助解决问题的关键因素



数据探索——寻找关键变量

问题：如何找到 “高价值顾客”

根据经验

■ 在宜家花了很多钱的顾客

■ 购买金额低但是频次高的顾客

问题：价值无法直接衡量和观测

解决方案：
在现有信息中找到关键变量来间接反映价值

“累积购买总额”

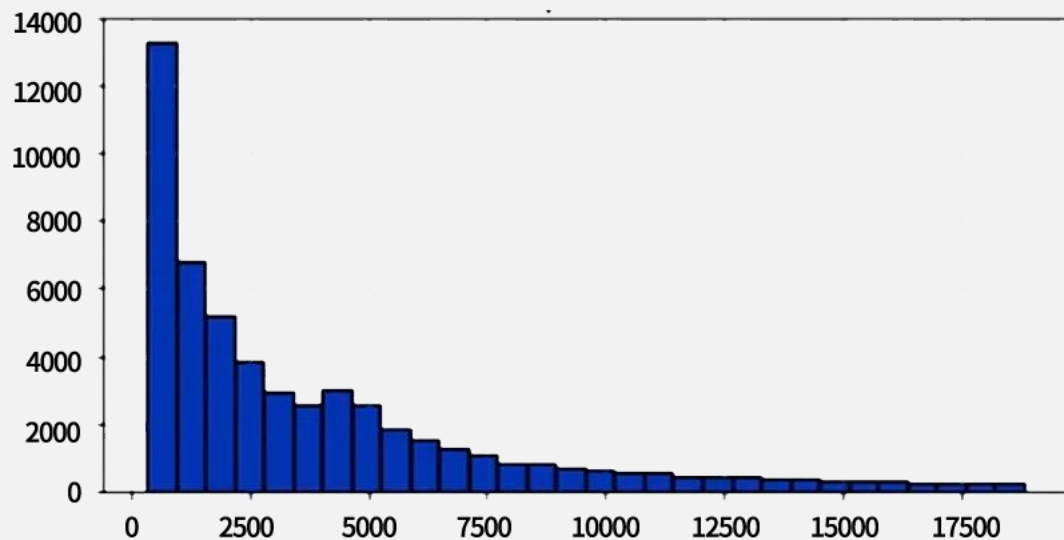
“最近一次购买距今的时间”



可视化分析

数据探索与可视化——变量分析

用图表可视化对单个变量进行分析——累积购买总额



➤ 创建直方图

横轴表示不同消费的分组

纵轴则表示对应的用户的数量



随着销售收入的增加，购买人数逐渐减少



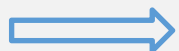
数据探索与可视化——变量分析

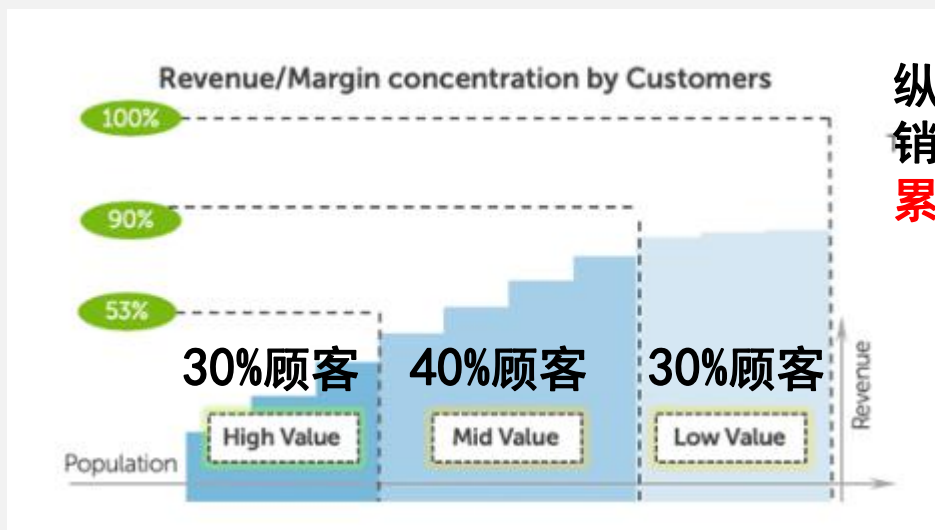
用图表可视化对单个变量进行分析——累积购买总额

➤ 第一步：抽样

- 从所有数据中抽样十万个顾客数据
- 将顾客的累计订单金额从高到底排列
- 每1万个顾客为1组

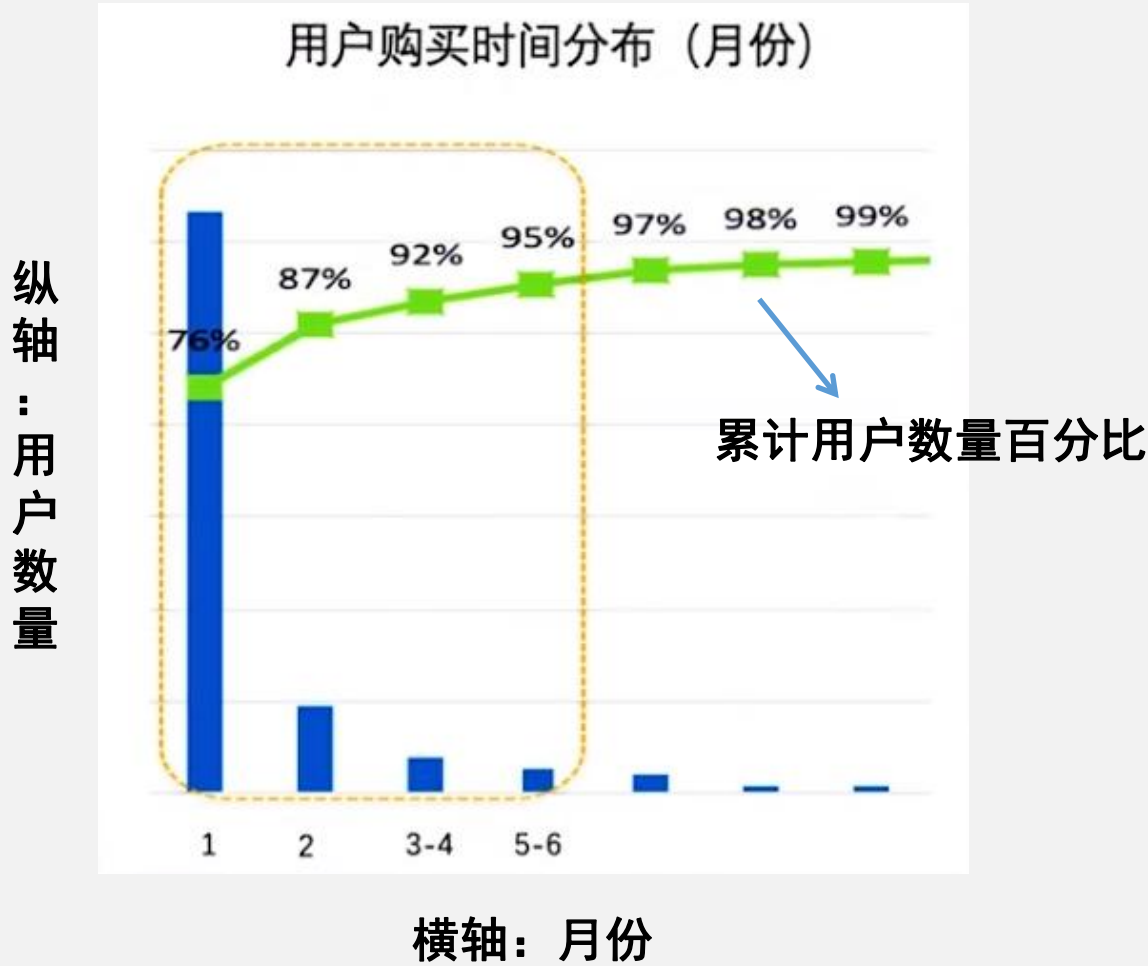
➤ 第二步：制作直方图

- 前30%贡献收入53%  高价值用户
- 中间40%贡献收入37% (90%-53%)
- 后30%贡献收入10% (100%-90%)



数据探索与可视化——变量分析

用图表可视化对单个变量进行分析——“最近一次购买距今的时间”



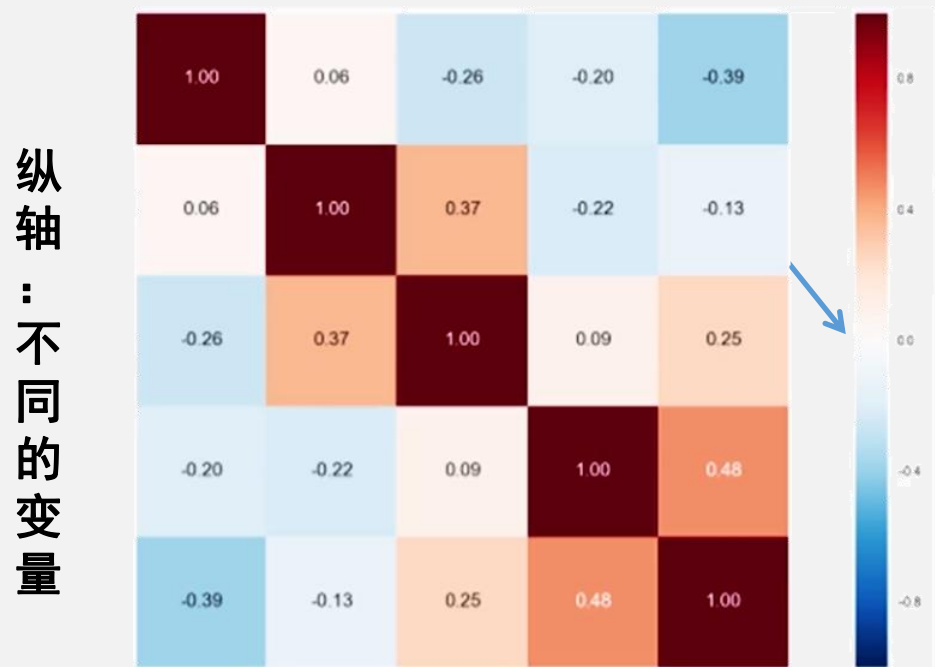
➤ 大部分用户（95%） ➡ 高价值用户
最近购买都集中在6个月以内

➤ 部分用户顾客（5%） ➡ 非高价值用户
6个月以内没有在宜家购买任何产品

可能流失到竞争对手
短期内他们也不应该被视为高价值用户，
曾经的高销售有可能是一次性装修等产生的需求

数据探索与可视化——变量分析

用图表可视化对多个变量进行相关性分析——“皮尔森相关系数”

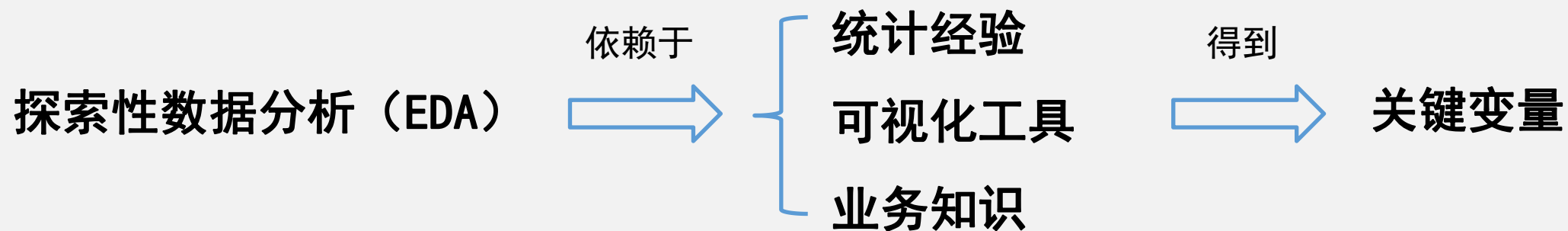


- 衡量的是变量之间的线性相关性
- 结果的取值区间为 $[-1, 1]$
 - -1表示完全的负相关
 - +1表示完全的正相关
 - 0表示没有线性相关
- 颜色的深浅表示数值的大小

➡ 从多个相关关系中锁定相关系数最强的那些



数据探索与可视化



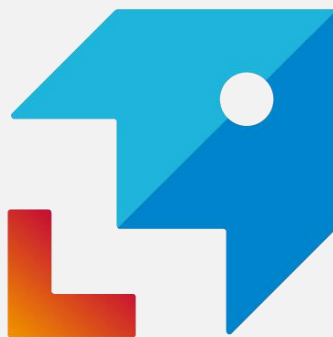
刚刚入行数据分析，
要面对几千上万个变量的分析研究

节省时间

Python这类工具

- 批量实现对所有变量的基础统计分析
- 生成各自对应的基础图表





数据分析

第6课：如何用数据进行商业决策？

数据建模与商业决策

将这些变量信息将被输入到分析模型中，
经过模型的选择和调整，最终给出能部署到业务中的数据分析结果



数据分析的模型建立，就像我们在学校里所学的函数：

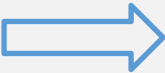
- 对自变量 x ，施与某一个规则 f ，得到了应变量 y
- y 与 x 之间的关系可以用 $y=f(x)$ 来表示

数据分析模型



数据建模与商业决策

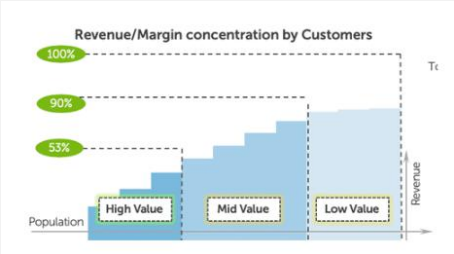
定义高价值用户，并从数据中识别高价值用户的具体特征



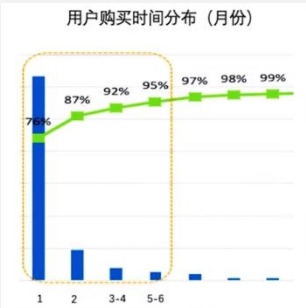
进行分群

累计购买总额 (X)

最近购买时间 (X)



累计购买总额高的为高价值用户



最近购买都集中在6个月以内
为高价值用户

问题：只有两个变量，不具有完全的科学，
损失了很多其他有价值的信息

解决方案：聚类分析
将所有变量都纳入到分析中，让机器计算出最佳分组



数据建模——聚类分析模型（在第四周的课程中会详细讲解）

定义高价值用户，并从数据中识别高价值用户的具体特征

聚类分析

- 定义：把相似的分析对象根据各自特征分成不同的组别的统计方法
- 应用场景：**客户分群(segmentation)**，并由此衍生出对客户画像工作
- 作用：能帮助我们更清楚的认识客户，识别用户特征

有助于我们回答下面这些观念建的问题

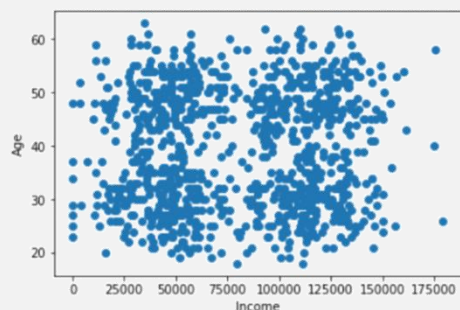
- 如何能够为不同客群提供定制化的产品或者服务
- 如何设定品牌的主要形象和定位
- 如何根据顾客需求，挖掘新的产品和服务机会



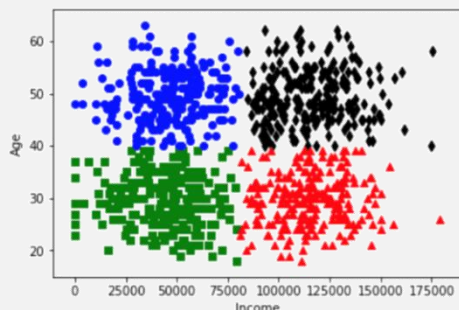
数据建模——聚类分析模型（在第四周的课程中会详细讲解）

定义高价值用户，并从数据中识别高价值用户的具体特征

聚类分析



聚类前



聚类后

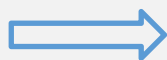
- 图中将年龄与收入两个维度进行模型的建立
- 你会发现
 - 各个顾客之间距离最近的个体合成一个小群体
 - 直到每个个体都存在于一个小群体
 - 在这里我们一共将所有顾客分成了4个小群体

真实数据建模时，获得的聚类分析模型如上图



数据建模——聚类分析模型（在第四周的课程中会详细讲解）

定义高价值用户，并从数据中识别高价值用户的具体特征



聚类分析



高价值用户



具体特征



➤ 城市新居住者

- 购买店面发生改变
- 送货地址发生改变
- 购买产品为生活日用类为主
- 注册手机号与所购买城市不同

➤ 新婚家庭

- 顾客为男性为主
- 购买产品为家具类为主
- 双人床和衣柜等高价格家具关注度高

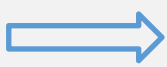
➤ 初为父母

- 顾客为女性为主
- 初次购买儿童/婴幼儿产品类型

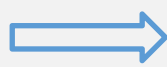


数据建模——预测分析模型（在第四周的课程中会详细讲解）

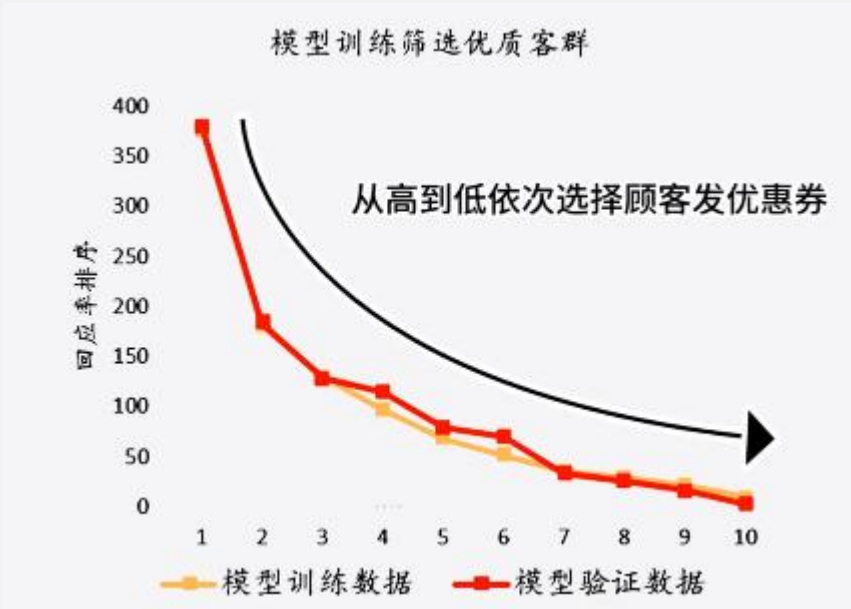
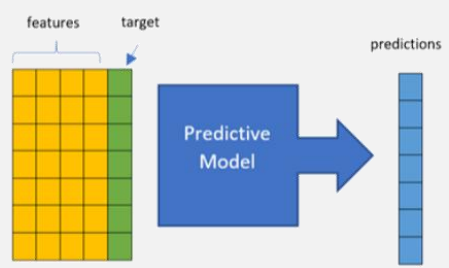
预测哪些用户会在收到优惠券后使用用于购买



预测模型

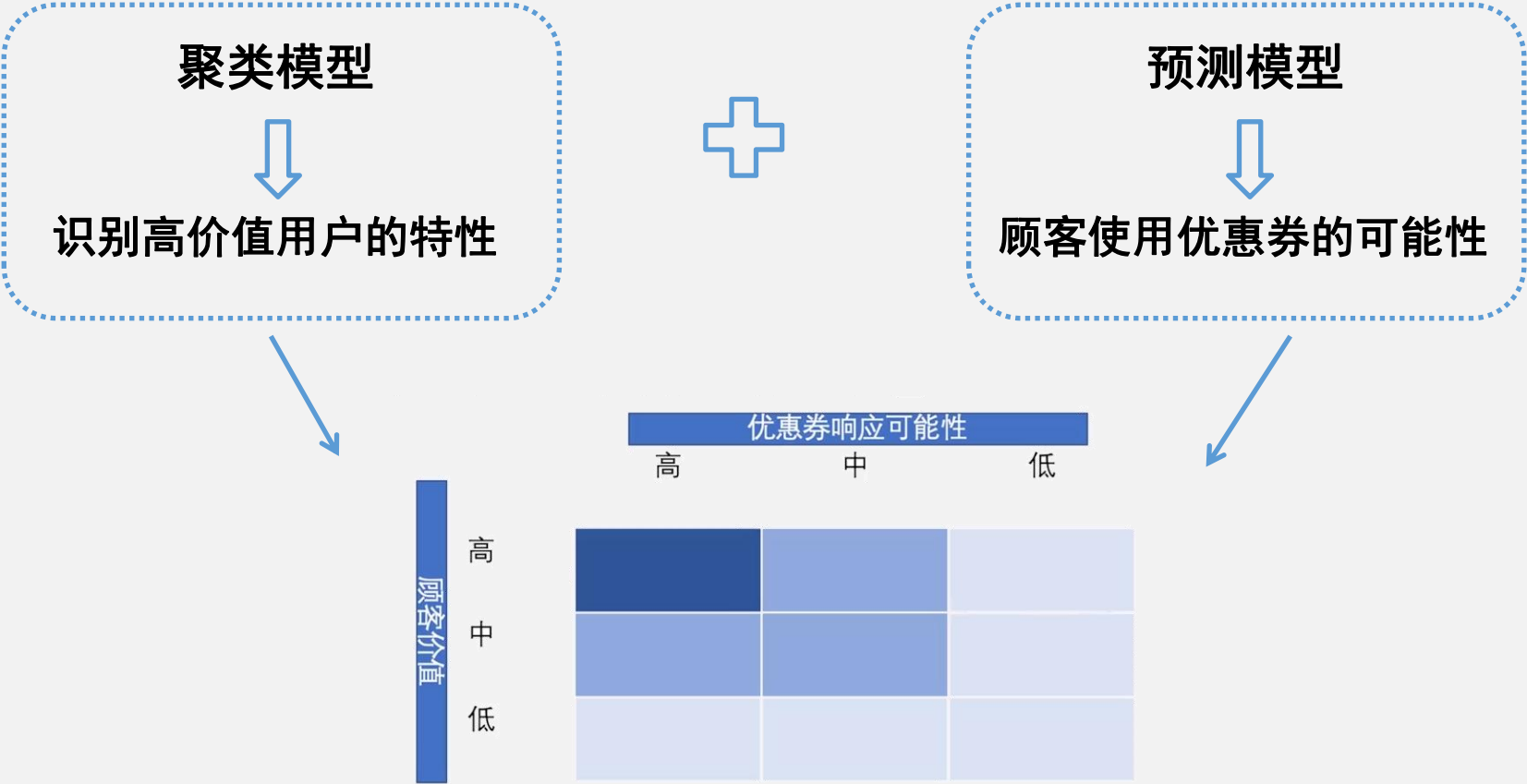


用户使用优惠券的可能性



数据建模与商业决策

在实际业务中：



优先选择那些潜在价值高，同时又会积极响应营销活动的客群

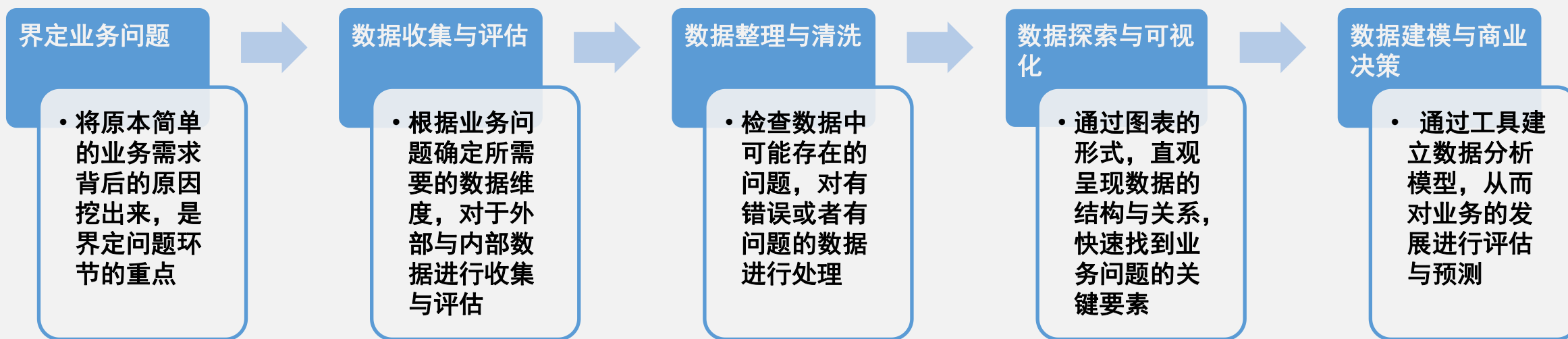




数据分析

总结

数据分析工作最重要的5个环节



数据分析工作最重要的5个环节

Step 1 界定业务问题

将原本简单的业务需求背后的原因挖出来，是界定问题环节的重点

针对HOW类问题，“我们能让他们在宜家店里多花钱？”

业务问题实际上：

“识别高价值顾客，实施针对性的营销方案，发放专属优惠券促进顾客购买”

1. 定义高价值顾客，并从数据中识别高价值顾客的具体特征
2. 选择那些在收到优惠券之后使用的顾客，将优惠券推送至会员卡，并跟踪分析后续的使用和购买情况



数据分析工作最重要的5个环节

Step 2 数据收集与评估

根据业务问题确定所需要的数据维度，对于外部与内部数据进行收集与评估

➤了解数据库 ➤数据收取 ➤数据拼接 ➤数据聚合

用户编号	第一次订单金额	第一次订单时间	累积订单金额	累积购买次数
1001	200	2019/02/12	500	2
1002	5000	2018/06/15	12500	4
1003	350	2018/10/2	3600	1
1004	60	2018/01/20	300	3

数据分析工作最重要的5个环节

Step 3 数据整理与清洗

检查数据的中可能存在的问题，对有错误或者有问题的数据进行处理

➤ **数据清洗：**将有问题的数据排除出去（最需要时间和精力的环节）

这里主要涉及对数据缺失，异常和其他问题的处理

➤ **数据整理：**将数据转化成更有助于后续分析的样式

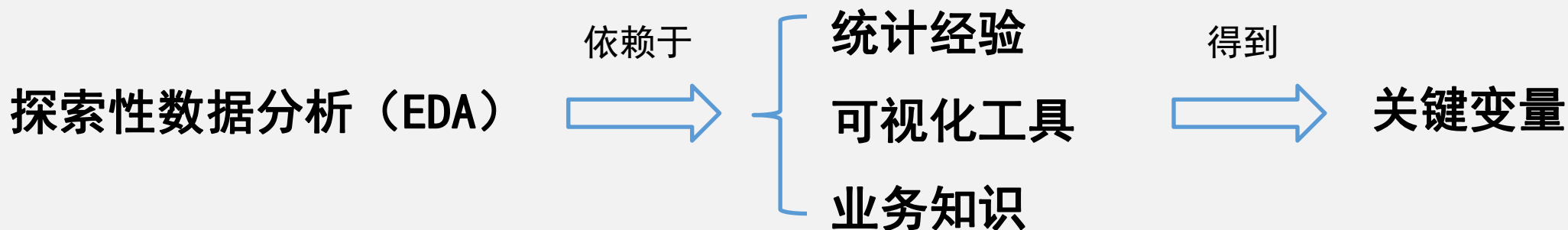
- 对数据进行统一的格式化和命名规则处理
- 对某些信息进行重新编码以满足后续分析需求



数据分析工作最重要的5个环节

Step 4 数据探索与可视化

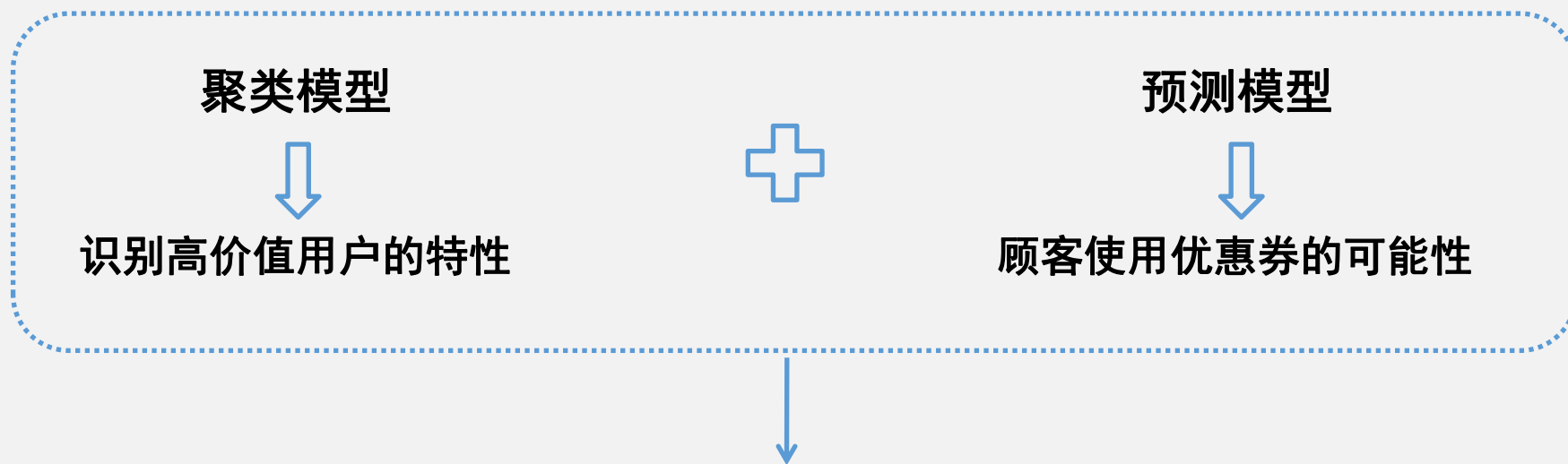
通过图表的形式，直观呈现数据的结构与关系，快速找到业务问题的关键要素



数据分析工作最重要的5个环节

Step 5 数据建模与商业决策

通过工具建立数据分析模型，从而对业务的发展进行评估与预测



优先选择那些潜在价值高，同时又会积极响应营销活动的客群

