第二章：数据

商务分析中的数据

- **数据（Data）**：通过一些测量过程获得的数字（或文字、图表）结果。
- **信息（Information）**：分析数据的结果，即从数据中抽取的可用于支持评估和决策制定的有意义的部分。

数据集

- **数据集**- 数据的集合.
  - 例:营销调研结果，历史股票价格，生产过程中抽取样本的尺寸
  - 一个数据集文件一般是一个二维表格，其中每一行代表一条记录或一个个体，每一列代表一个属性(fields, or attributes)或变量（variable）

例1.3：一个销售交易数据集

观测值

	A	B	C	D	E	F	G	H
1	Sales Transactions: July 14							
2								
3	Cust ID	Region	Payment	Transaction Code	Source	Amount	Product	Time Of Day
4	10001	East	Paypal	93816545	Web	$20.19	DVD	22:19
5	10002	West	Credit	74083490	Web	$17.85	DVD	13:27
6	10003	North	Credit	64942368	Web	$23.98	DVD	14:27
7	10004	West	Paypal	70560957	Email	$23.51	Book	15:38
8	10005	South	Credit	35208817	Web	$15.33	Book	15:21
9	10006	West	Paypal	20978903	Email	$17.30	DVD	13:11
10	10007	East	Credit	80103311	Web	$177.72	Book	21:59
11	10008	West	Credit	14132683	Web	$21.76	Book	4:04
12	10009	West	Paypal	40128225	Web	$15.92	DVD	19:35
13	10010	South	Paypal	49073721	Web	$23.39	DVD	13:26

记录、个体

属性、变量

大数据

- 大数据：传统数据处理应用软件不足以处理的大或复杂的数据集
  - 数据量大（Volume）--记录和变量特别多的数据集
  - 数据种类多(Variety)
  - 数据价值密度低（Value）
  - 数据产生和处理速度快（Velocity）
- 初始的分析基于小数据

数据测量尺度

- 每个变量的观测值需要用一些测量尺度来度量。
- 不同的测量尺度决定了数据中的信息量是不同的，并且需要用不同的方法去分析这些数据。

数据测量尺度

- **名义尺度** –数据只展示类别信息
- **顺序尺度** – 数据展示了顺序等级
- **间隔尺度**– 数值间的距离按某一固定度量单位显示，可比较（最常见的类型）
- **比率尺度** – 距离可比较，此外还有绝对零点的定距数据，数值之间的比率也有意义

分类型数据和数量型数据

- **分类型数据**-名义数据、顺序数据
  - 用分类型数据表示的变量为分类（型）变量
- **数量型数据**-间隔数据、比率数据
  - 用数量型数据表示的变量为数量（型）变量

例 1.4:数据分类

应付账款时长

	A	B	C	D	E	F	G	H	I	J
1	**Purchase Orders**									
2										
3	**Supplier**	**Order No.**	**Item No.**	**Item Description**	**Item Cost**	**Quantity**	**Cost per order**	**A/P Terms (Months)**	**Order Date**	**Arrival Date**
4	Hulkey Fasteners	Aug11001	1122	Airframe fasteners	$ 4.25	19,500	$ 82,875.00	30	08/05/11	08/13/11
5	Alum Sheeting	Aug11002	1243	Airframe fasteners	$ 4.25	10,000	$ 42,500.00	30	08/08/11	08/14/11
6	Fast-Tie Aerospace	Aug11003	5462	Shielded Cable/ft.	$ 1.05	23,000	$ 24,150.00	30	08/10/11	08/15/11
7	Fast-Tie Aerospace	Aug11004	5462	Shielded Cable/ft.	$ 1.05	21,500	$ 22,575.00	30	08/15/11	08/22/11
8	Steelpin Inc.	Aug11005	5319	Shielded Cable/ft.	$ 1.10	17,500	$ 19,250.00	30	08/20/11	08/31/11
9	Fast-Tie Aerospace	Aug11006	5462	Shielded Cable/ft.	$ 1.05	22,500	$ 23,625.00	30	08/20/11	08/26/11
10	Steelpin Inc.	Aug11007	4312	Bolt-nut package	$ 3.75	4,250	$ 15,937.50	30	08/25/11	09/01/11
11	Durrable Products	Aug11008	7258	Pressure Gauge	$ 90.00	100	$ 9,000.00	45	08/25/11	08/28/11
12	Fast-Tie Aerospace	Aug11009	6321	O-Ring	$ 2.45	1,300	$ 3,185.00	30	08/25/11	09/04/11
13	Fast-Tie Aerospace	Aug11010	5462	Shielded Cable/ft.	$ 1.05	22,500	$ 23,625.00	30	08/25/11	09/02/11
14	Steelpin Inc.	Aug11011	5319	Shielded Cable/ft.	$ 1.10	18,100	$ 19,910.00	30	08/25/11	09/05/11
15	Hulkey Fasteners	Aug11012	3166	Electrical Connector	$ 1.25	5,600	$ 7,000.00	30	08/25/11	08/29/11

名义　顺序　名义　名义　比率　比率　比率　比率　间隔　间隔

数据来源

- 一手数据：自己调查（访谈、询问、问卷等方式）得来的数据内容。
- 二手数据：从各有关方面（例如国家统计局等）间接得到的数据内容。

一手数据

- 探索性数据收集：
  - 目的：形成最初的预见和洞查，例如销量为什么下降了。
  - 方法：焦点小组，深度访谈

焦点小组

- 小组座谈、专题组座谈

项目	特征
小组规模	8-12人
小组组成	预先筛选的调查对象
物理环境	放松的、非正式的气氛
时间长度	1-3小时
记录	利用录音和录像
主持人	主持人的观察、互动与沟通技能

焦点小组

- 卡夫100卡路里包装饼干
- Corporate Apparel公司网站设计
- 拉斯维加斯豪华住宅设计

深度访谈

- 一对一
- 不受他人影响
- 深入性佳
- 费用高
- 费时
- 坦诚性好，可用于敏感话题

一手数据

- 描述性数据收集：
  - 目的：产生相关顾客群的特征的数据，例如顾客为我们的产品花了多少钱，为竞争对手的产品花了多少钱，买我们品牌产品的顾客有什么特征。
  - 方法：问卷调研

问卷调研

- 广泛使用在几乎所有500强公司中
- 用于收集顾客的态度、满意度、购买习惯
- 数据有助于对顾客进行分类、制定营销策略

测量量表

- 量表：测量工具
  - 定类量表（收集的数据类型为名义数据）
  - 定序量表（收集的数据类型为顺序数据）
  - 定距量表（收集的数据类型为间隔数据）
  - 定比量表（收集的数据类型为比率数据）

测量量表

- 定类
  - 下面的饮料中，你喜欢哪一个？（选出所有符合条件的）：

1. 可乐__
2. 雪碧__
3. 气泡水__
4. 果汁__

测量量表

- 定序
  - 根据你的喜好程度对这些饮料排序（最喜欢的饮料=1，最不喜欢的饮料=4）：

可乐__

雪碧__

气泡水__

果汁__

测量量表

- 定距
  - 请你在量表中合适的位置标出你对每种饮料的喜好程度

	非常讨厌	讨厌	喜欢	非常喜欢
可乐				
雪碧				
气泡水				
果汁				

测量量表

- 定距
  - 请移动鼠标至合适的位置，表明你对可乐的喜爱程度

测量量表

- 定比
  - 请将100分分配给下面的饮料，用来表示你对它们的喜好程度：

可乐__

雪碧__

气泡水__

果汁__

测量量表

量表	基本比较	例子	平均测量
定类	同一性	男-女、使用-不使用	众数
定序	有序性	品牌偏好、质量等级	中位数
定距	定距比较	对品牌态度	均值
定比	绝对数量比较	单位销售量、购买数量	几何平均数

问卷调研步骤

第0步，确定对象
第一步，明确所需信息
第二步，确定问卷的类型和实施方法
第三步，确定每个问题的内容
第四步，确定每个问题的回答形式
第五步，确定每个问题的措辞
第六步，确定问题的顺序
第七步，设计问卷的外观特性
第八步，检查步骤1-7并且进行修订
第九步，预测试问卷并修正问卷

问卷调研优缺点

- 问卷调研的优点：低成本 、易实施，是了解客户的很好的方式
- 问卷调研的缺点：难以获得无偏的回答；如何选择回答问卷的合适人群（可结合焦点小组）
- 参考书：
  - *《营销调研方法论基础》，北京大学出版社*
  - *《市场营销研究：应用导向》，电子工业出版社*

一手数据

- 因果性数据收集：
  - 目的：验证因果关系
    - 例如改变登录界面是否有助于更多的客户注册登录？
    - 把儿童食品包装盒设计得矮一些不易被小孩碰翻，是否能增进家长对此产品的态度评价？
  - 方法：试验
    - AB测试

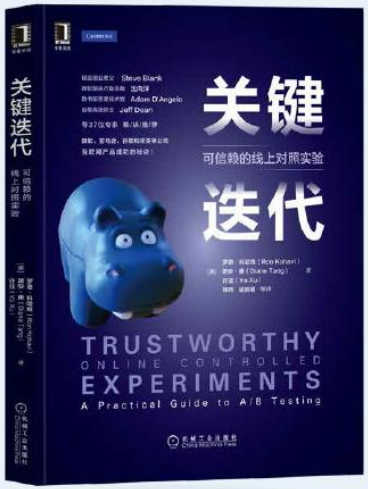
AB测试

微软、亚马逊、谷歌和领英等公司

互联网产品成功的秘诀！

线上对照实验「教父」领衔撰写

A/B测试领域「圣经」之作



汲取业内顶尖专家智慧

践行数据驱动决策

真实案例+常见陷阱+解决方案

为你提供切实可行的实践路线图

AB测试

- AB测试：对照实验。它将测试对照组(A)与实验 组(B)，基于关键指标来衡量哪一个版本更成功。
  - 爱彼迎（Airbnb）、亚马逊、谷歌、脸书、微软等企业每年运行成千上万个AB测试。
    - 必应改进搜索页广告标题陈列方式：将标题下方第一行文字移至标题同一行，以使标题变长。
      www.bing.com
    - 亚马逊信用卡推广从主页换到购物车。根据顾客购物车的商品显示个性化推荐。
    - 谷歌调整页面配色方案。

计费）或者因网页出错而导致只能看到广告。

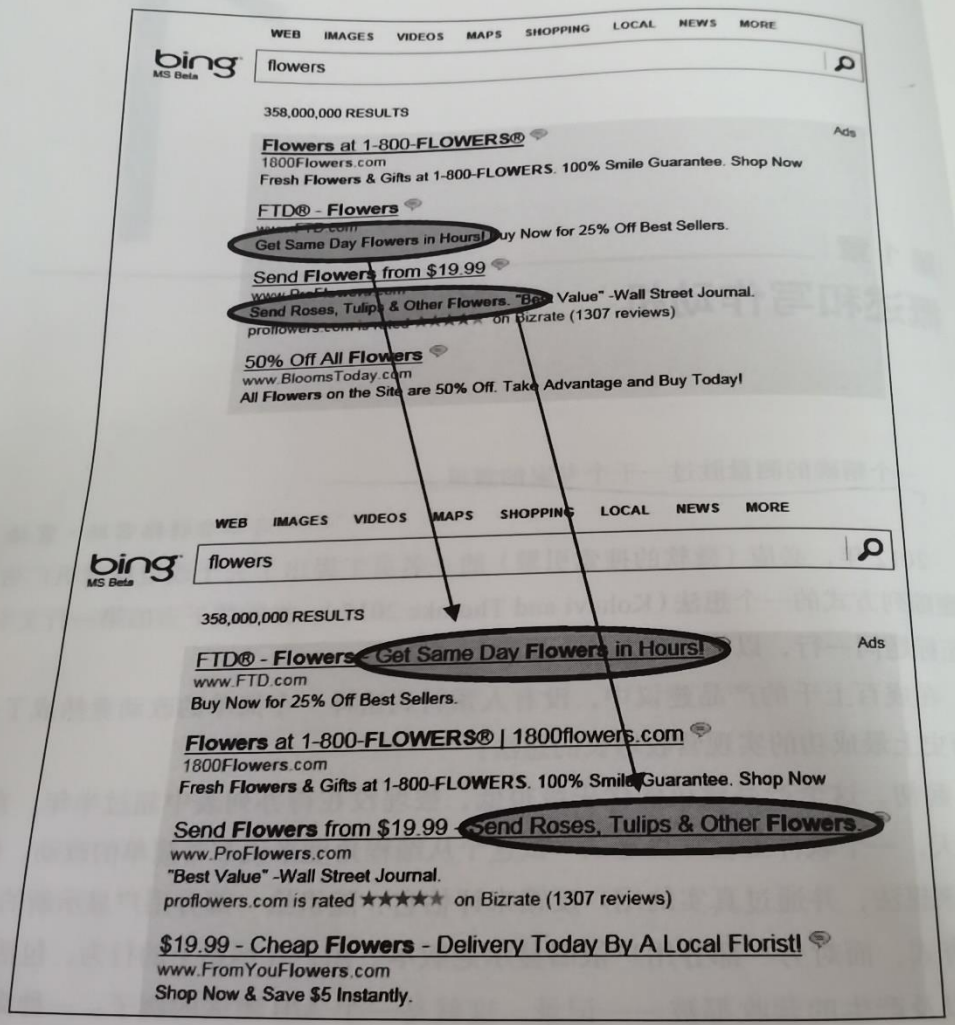


图 1.1　改进必应上广告陈列方式的实验

然而就这个实验而言，营收增长是真实有效的。在没有显著损害其他关键用

AB测试

- 实验的评估标准（overall evaluation criterion, OEC）
  - 你若不能测量一个事物，也就无法改进它。

    --彼得.德鲁克，现代管理学之父
  - 市场营销部门希望通过发送含打折优惠券的促销邮件提高销量。公司担心在结账页面增加优惠券会降低营收。
    - “画门法”
    - 人均营收：仅包括开始付款流程的用户？仅包括完成付款的用户？所有访问该网站的用户？

AB测试

- 实验的评估标准（overall evaluation criterion, OEC）
  - 你怎么测量我，我就怎么表现。

    --伊利雅胡·高德拉特，以色列管理学家
  - 必应使用两个组织层面的关键指标来评估进展：搜索份额和营收。
  - 当必应有一个排序漏洞时，导致实验组的用户看到一些很差的搜索结果。
  - 但搜索份额增加10%和营收增加30%。

AB测试

- 巧妙构建实验
- 网站载速过慢带来的后果：用户失望放弃、品牌声誉下跌、运营成本增加，以及营收相应减少。
  - 如果一个工程师可以将服务器提速10毫秒（眨眼时间的1/30），那么他/她的贡献就已经高于自身年薪了。每一毫秒都是有用的。
    - --Kohavi, Deng, Frasca, Walker, Xu and Pohlmann(2013)
- 减速实验

因果性数据收集

- AB测试不可行时，进行观察性因果研究
  - 实验单元数很少
  - 所测试的项目不合伦理（科学的进展超前于人类的伦理规范—查理.卓别林）
  - 。。。。。。
- 因果关系推断
  - 相关性
  - 时间顺序
  - 没有共同的驱动因素

观察性研究

- 假设你为爱奇艺工作，每月会员流失率为X。你决定引入一个新功能，观察到使用这个新功能的用户的流失率为X/2。得出结论：该新功能使流失率降低一半。
- 微软Office 365发现看到错误信息并遭遇系统崩溃的用户有较低的流失率。得出结论：使用户看到错误信息并遭遇系统崩溃，可以降低客户流失率。
- 影院小吃店发现看《战狼2》的顾客消费的饮料比看《唐人街探案》的顾客消费的饮料多。得出结论：电影类型会影响顾客饮料消费。

一手数据—总结

- 探索性数据收集
  - 研究问题偏探索性（ambiguous problems）
    - 产品的销量降低了，为什么？
- 描述性数据收集
  - 研究问题偏描述性（aware of problem）
    - 什么样的顾客在买我们的产品？哪些人在买竞争对手的产品？
- 因果性数据收集
  - 研究问题的因果性较强（problem clearly defined）
    - 如果我改变登录页面的设置，是否会有更多的人购买产品？

调查研究

- 企业中建立数据驱动文化，还是依赖HiPPO? (Highest Paid Person’s Opinion)

- 事实依据越少，观点就越强。

--Arnold Glasow

调查研究

- 毛泽东
  - “没有调查，就没有发言权”
  - 《湖南农民运动考察报告》

  https://zhuanlan.zhihu.com/p/466523404
- 习近平
  - 调查研究是谋事之基、成事之道，没有调查就没有发言权，没有调查就没有决策权。调查研究是我们做好工作的基本功。
  - 善于获取数据、分析数据、运用数据，是领导干部做好工作的基本功。

调查研究

- 上海财经大学“千村调查”
  - 以“三农”问题为研究对象的大型社会实践和社会调查研究项目
  - “走千村，访万户，读中国”
    - 农村养老问题现状调查
    - 中国农村创业现状调查
    - 农村文化状况调查研究
    - ……

二手数据

- **数据库** – 一系列相关数据集
  - 例：一家零售店有客户信息的数据集、每件产品销售量的数据集等
- 常见的数据库
  - 企业内部数据库：产品销售情况、运营指标、人力资源考核、网络点击数据（访客的地点、访问时间、停留长度、访问路径、搜索产品、浏览产品、点开广告、最终购买、阅读点评等）
  - 盈利机构数据库：淘系数据等

https://tianchi.aliyun.com/?spm=5176.12282016.J_3941670930.9.34946d92cQGW7I

https://www.daas-auto.com/supermarket_data_De/727.html

https://www.datatang.com/

  - 政府机构数据库：经济统计年鉴

https://sufe.libguides.com/az.php?t=23126

一手数据 OR 二手数据

- 一手数据获得成本较高
- 根据研究目的，如有二手数据可以使用时，优先使用二手数据