数据世界探秘

第二章　数据及数据来源

授课要点

- 一、什么是统计
- 二、研究设计
- 三、数据定义和类型
- 四、传统的数据来源
- 五、大数据数据来源

一、什么是统计

（一）什么是统计

（二）统计工作四步曲

（三）统计学基本概念

（一）什么是统计

1. 三种涵义

- 统计数据/统计资料
- 统计工作：对统计数据进行收集、整理和分析的过程。
- 统计学：一门研究总体数量特征的方法论科学。

2. 两重关系

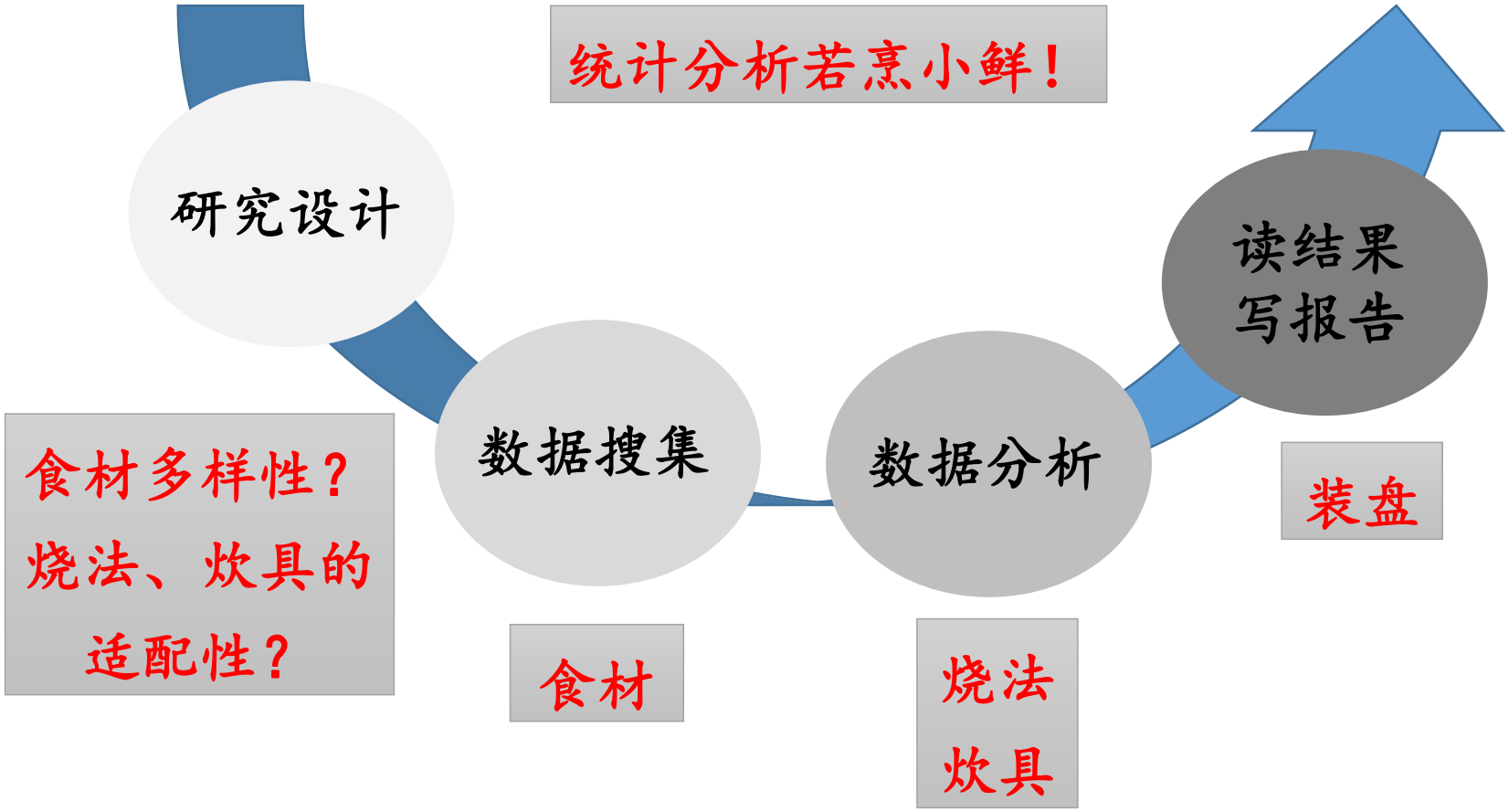
- 统计数据
- 统计工作

→ 工作与工作对象（成果）的关系

- 统计学

→ 实践与理论的关系

（二）统计工作四步曲

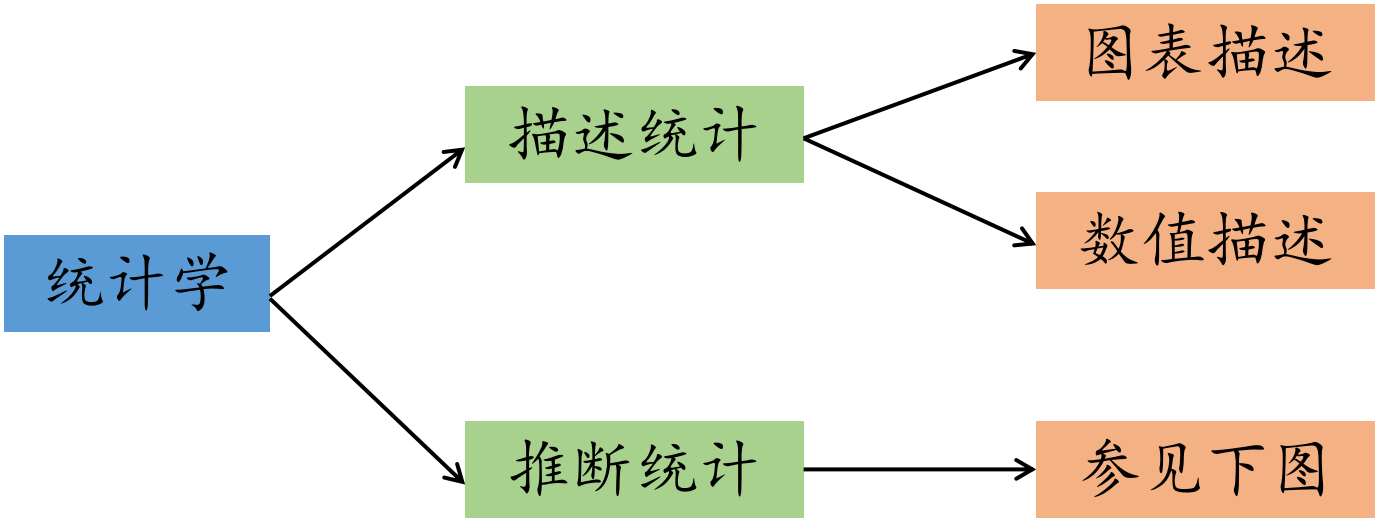


（三）统计学基本概念

1. 定义

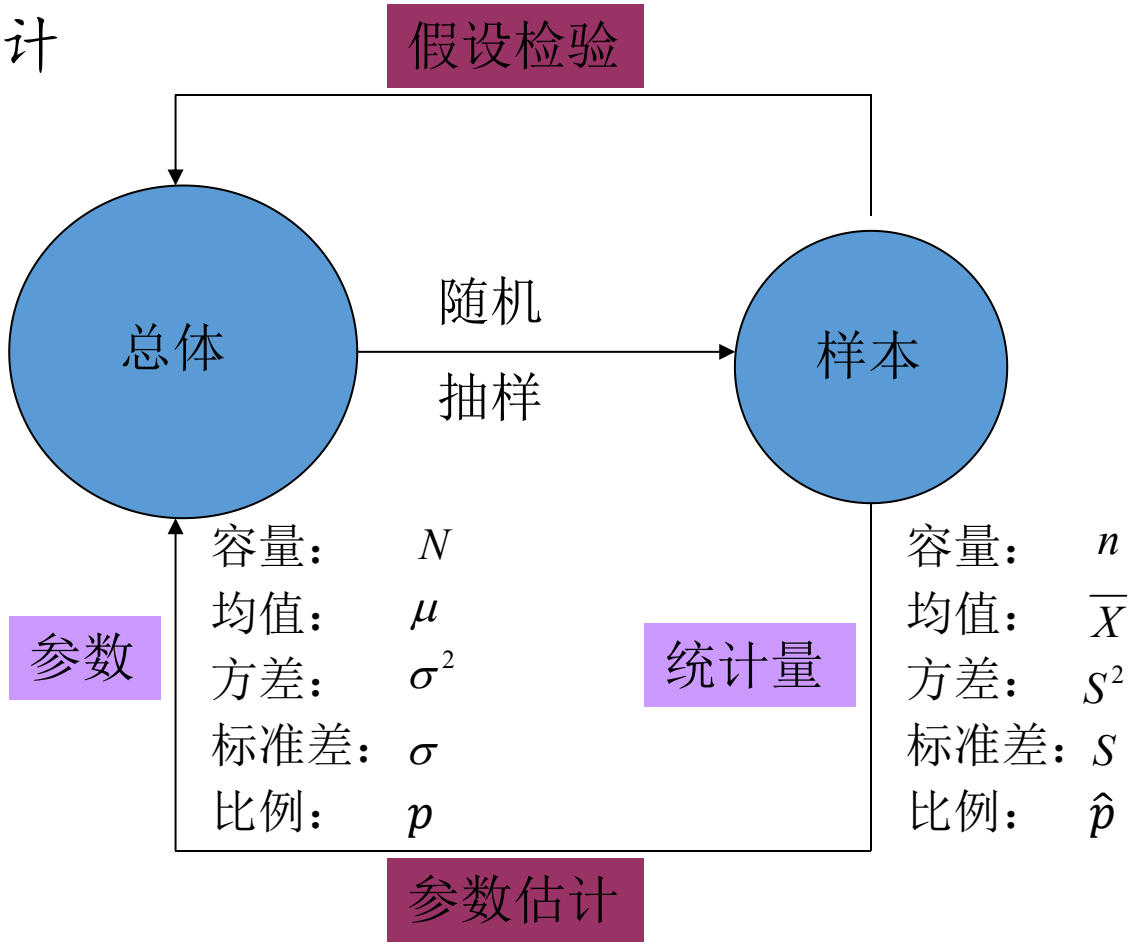
统计学（Statistics）是对统计数据进行搜集、整理分析，并得出结论的一整套方法论科学。

2. 基本框架



（三）统计学基本概念

推断统计

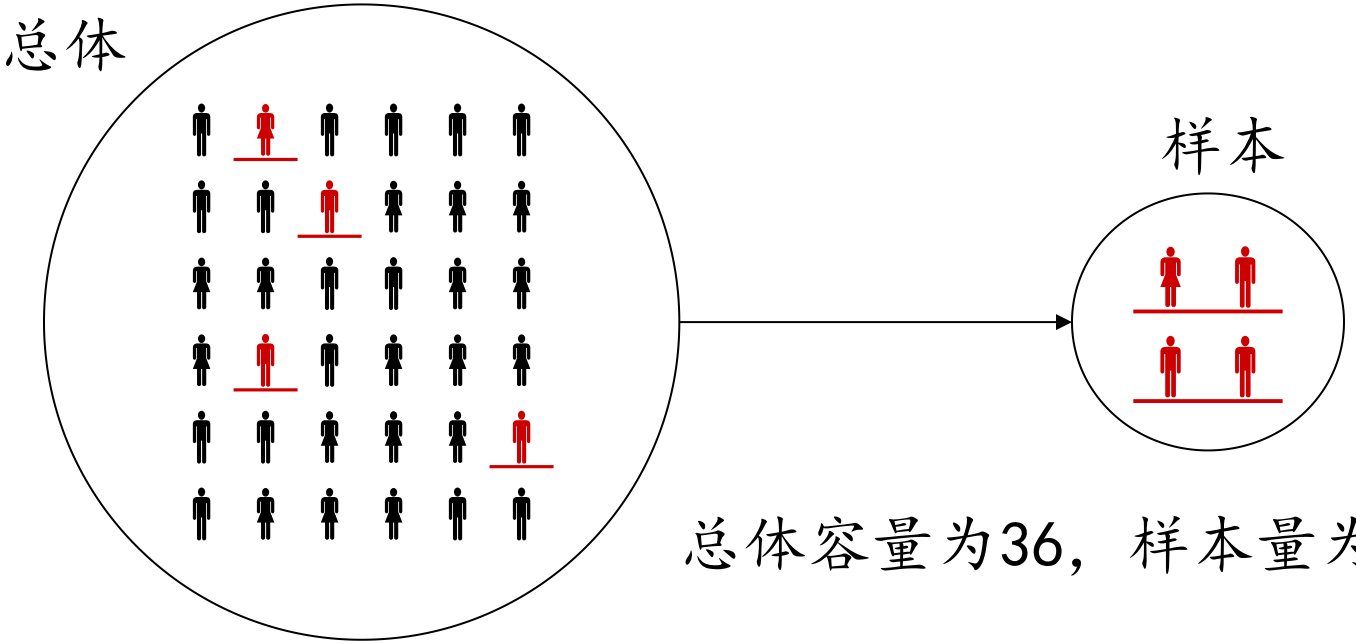


（三）统计学基本概念

上海财经大学 SHANGHAI UNIVERSITY OF FINANCE AND ECONOMICS

3. 重要术语

- 总体(Population)：根据一定目的确定的所要研究事物的全体。构成总体的每个成员称为个体。总体中个体数量称为总体容量。
- 样本（Sample）：从总体中随机抽取的n个个体构成的集合体。样本中的个体称为样品，n称为样本容量，简称样本量。



（三）统计学基本概念

- 参数（Parameter）：描述总体数量特征的量。常见的有：
  - ✓ 总体均值 μ
  - ✓ 总体方差 σ^2、总体标准差σ
  - ✓ 总体比例 p
- 统计量（Statistic）：描述样本数量特征的量。
  - ✓ 样本均值 $\overline{X}$
  - ✓ 样本方差 S^2、样本标准差S
  - ✓ 样本比例 $\hat{p}$

（三）统计学基本概念

➢ 思考题2.1

如果我们对统计学大一A班学生的情况感兴趣，已知班级共有学生50名，我们从统计学大一A班中随机抽取了10名同学。试回答以下问题：

（1）什么是总体？总体容量是多少？

（2）什么是样本？样本量是多少？

（3）试对该例的参数、统计量进行举例。

（三）统计学基本概念

➢ 解答：

- 总体：统计学大一A班的所有学生构成总体。
- 总体容量：50。
- 参数举例：

  统计学大一A班学生的平均身高（总体均值）

  统计学大一A班学生中女同学的比例（总体比例）
- 样本：被抽中的10名同学形成一个样本。
- 样本量：10。
- 统计量举例：

  被抽中的10名同学的平均身高（样本均值）

  被抽中的10名同学中女同学的比例（样本比例）

二、研究设计

（一）研究设计内容

1. 明确统计分析的问题

（1）把握三个点

- **注意点**：选择具有现实意义的问题，特别是与中心工作、全局性工作有密切联系的问题。
- **矛盾点**：选择影响比较大且争论比较多的问题。
- **发生点**：选择改革开放和国民经济发展中出现的新情况、新问题、新经验，或者带有苗头性、动向性、突发性的问题。

二、研究设计

- **注意点：**
  - ✓ “三新”经济:2018年8月14日，国家统计局印发《新产业新业态新商业模式统计分类（2018）》。2018年11月22日，国家统计局对外发布报告：《2017年我国“三新”经济增加值相当于GDP的比重为15.7%》。
  - ✓ 高质量发展：2017年中共十九大首次提出的新表述，表明中国经济由高速增长阶段转向高质量发展阶段。
  - ✓ 研究与试验发展（R&D）投入统计：2019年4月19日，国家统计局印发《研究与试验发展（R&D）投入统计规范（试行）》。

二、研究设计

- **矛盾点：**
  - ✓ 互联网金融曝雷与反欺诈
  - ✓ 网约车的进与退
  - ✓ 消费升级还是消费降级
- **发生点：**
  - ✓ 共享经济
  - ✓ 支持民营企业和中小企业发展
  - ✓ （上海）垃圾分类

二、研究设计

（2）问题的三种类型

- ✓ **描述型问题**——描述某种现象的问题
  - 中国城镇失业率是多少？
  - 某生产线的生产过程是否正常？
  - 某种疾病的患病人数有多少？
- ✓ **关系型问题**——比较研究对象的问题
  - 与美国失业率相比，中国失业率的情况是怎样的？
  - 各个生产企业的生产数量是否存在差异？
  - 某种疾病的患病人数在城镇和农村之间是否存在差异？
- ✓ **因果型问题**——寻找某种现象生成原因或关联性的问题
  - 为什么中国城镇失业率高于农村失业率？
  - 为什么生产企业A的生产数量是最高的？
  - 为什么某种疾病的患病人数，城镇比农村更高？

二、研究设计

2. 明确统计分析的目的

- 给谁用？（国家机关？企事业单位？其他对象？）
- 怎么用？（反映现状或动态？产生问题的原因分析？对策建议？）

二、研究设计

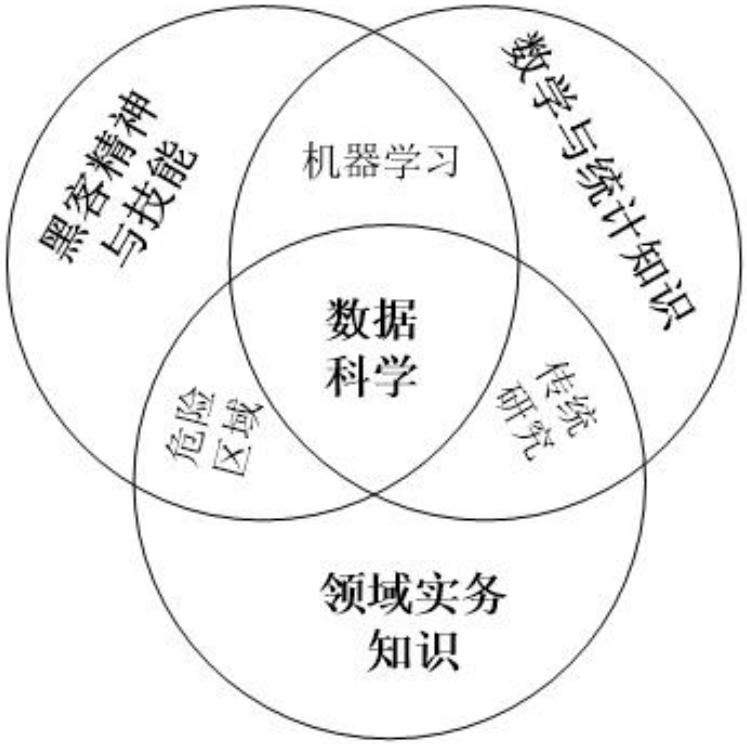
3. 查阅相关文献

- 理论研究文献
- 相关政策法规
- 实务现行做法（横向资料）

不以相关理论为基础的统计分析都是空中楼阁！

二、研究设计

2010年，Drew Conway提出了“数据科学维恩图”。根据图形可知，数据科学位于统计学、计算机科学和某一领域实务知识的交叉之处，具备较为显著的交叉型学科的特点。即数据科学是一门以统计学、计算机科学和领域知识为理论基础的新兴学科。



4. 搭建统计分析框架

- 分析目的分解：将分析目的分解成若干个不同的分析要点。
- 明确分析要点：依据相关理论，确定每个分析要点的逻辑思维和分析角度，确定分析指标，思考数据分析方法。

二、研究设计

（二）研究设计的重要性

- 研究设计是确保后续统计分析过程有效进行的先决条件。它为数据搜集、数据分析和撰写报告提供了清晰的指引方向。
- 研究设计很大程度上决定了统计分析的深度。

（三）研究设计的注意事项

- 有意义的选题
- 理论联系实际
- 框架的逻辑性
- 深度和前瞻性

二、研究设计

➢ 案例2.1：上海市民消费需求调研分析

研究背景

出口、投资和消费是拉动经济增长的“三驾马车”。长期以来，出口和投资一直是拉动我国经济增长的主要方式，但是在2009年国际金融危机的影响下，2010年我国出口大幅下降，对我国经济增长呈现出负作用。投资的目的是为了扩大再生产，而消费是生产的最终环节，商品只有进入到了消费阶段，生产过程才最终完成，商品的效用和价值才得以最终实现。为了促进我国国民经济持续健康地发展，首要任务是扩大内需，促进消费。

二、研究设计

上海要建设成为“四个中心”，需要扩大内需，促进消费。但是其长期明显偏低的消费率却并不利于上海市经济的长远发展。根据《上海市统计年鉴》，2000年至2007年，上海居民消费率一直维持在35%～37%的水平，而最终消费率则徘徊在47%～49%之间。2002年，世界平均消费率为80.1%，其中，低收入国家为80.7%，中等收入国家为74.3%，而高收入国家为81.0%。上海市的消费率远远低于世界平均水平，甚至与低收入国家相比，还存在较大的差距。

在此背景下，上海市政府发展研究中心设立了决策咨询研究项目《上海市民消费需求调研分析》。

二、研究设计

研究设计

- 明确统计分析的问题：上海市民消费需求分析
- 明确统计分析的目的：——为上海市政府提供决策依据
  - ✓ 了解上海市民的消费现状
  - ✓ 掌握制约消费的主要问题
  - ✓ 挖掘未来消费的热点
  - ✓ 为拉动内需，促进上海经济快速稳定发展提供决策依据
- 查阅相关文献资料：查阅与消费相关的理论和文献、消费现行政策，为确定消费制约因素、未来消费热点提供依据。

二、研究设计

上海财经大学 SHANGHAI UNIVERSITY OF FINANCE AND ECONOMICS

- 搭建统计分析框架
  - ✓ 了解上海市民的消费现状：通常而言，消费现状由收入水平、消费水平、消费倾向（消费支出占收入的比重，即消费率）、消费结构等加以反映。
  - ✓ 掌握制约消费的主要问题：通常而言，收入水平、物价水平、存钱意愿、投资意愿、消费意愿、消费时间和消费品供给是影响消费的主要因素。
  - ✓ 挖掘未来消费的热点：在购房、购车、耐用消费品消费、消费结构、消费方式（线上线下）等方面挖掘热点。
  - ✓ 提供决策依据：根据调研分析结果，为拉动内需，促进上海经济快速稳定发展提供决策依据。

三、数据定义和类型

（一）什么是数据

（二）数据的类型

（一）什么是数据

1. 常见回答

- 数据就是数字 ×

数字是最典型的传统数据，例如，GDP，股市指数，人的身高、体重、血压等都是数字，也都是数据。

数字是数据，但是数据却不完全是数字。

- 数据就是信息 √

数据就是信息，但是由于数据和信息都是非常抽象的概念，两者的相互定义并不好理解。

（一）什么是数据

上海财经大学 SHANGHAI UNIVERSITY OF FINANCE AND ECONOMICS

2. 常见概念

- Data is a set of values of subjects with respect to qualitative or quantitative variables.

—— *Wikipedia*

- Data are the facts and figures collected, analyzed, and summarized for presentation and interpretation. 数据是所搜集、分析、汇总，用以描述和解释的事实和数字。

—— David R. Anderson, Dennis J. Sweeney, Thomas A. Williams, *Statistics for Business and Economics*, Cengage Learnings

- 凡是能够通过必要的信息化技术和电子化手段进行记录的都是数据。

——王汉生，《数据思维——从数据分析到商业价值》，中国人民大学出版社

（一）什么是数据

产生数据的技术手段具有强烈的时代
技术手段越来越丰富，形成了大量新的数

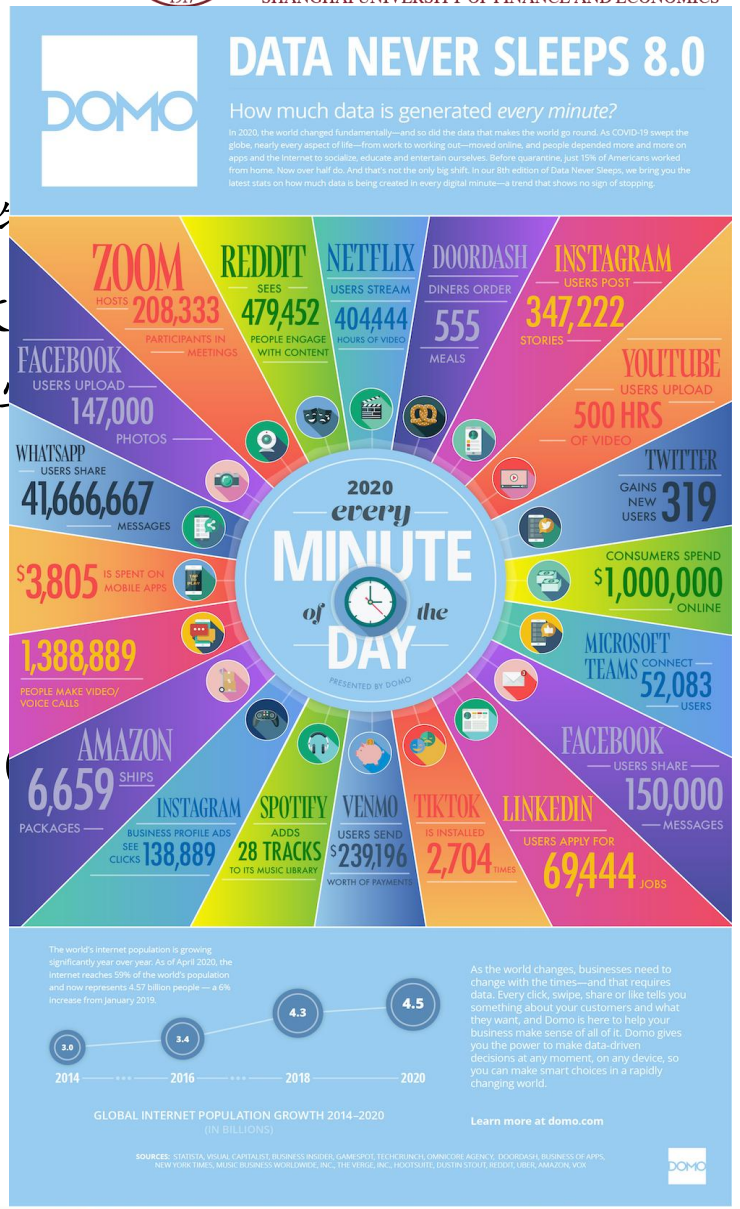
音频设备（采集声音，转化为音频数

数码成像技术——图像

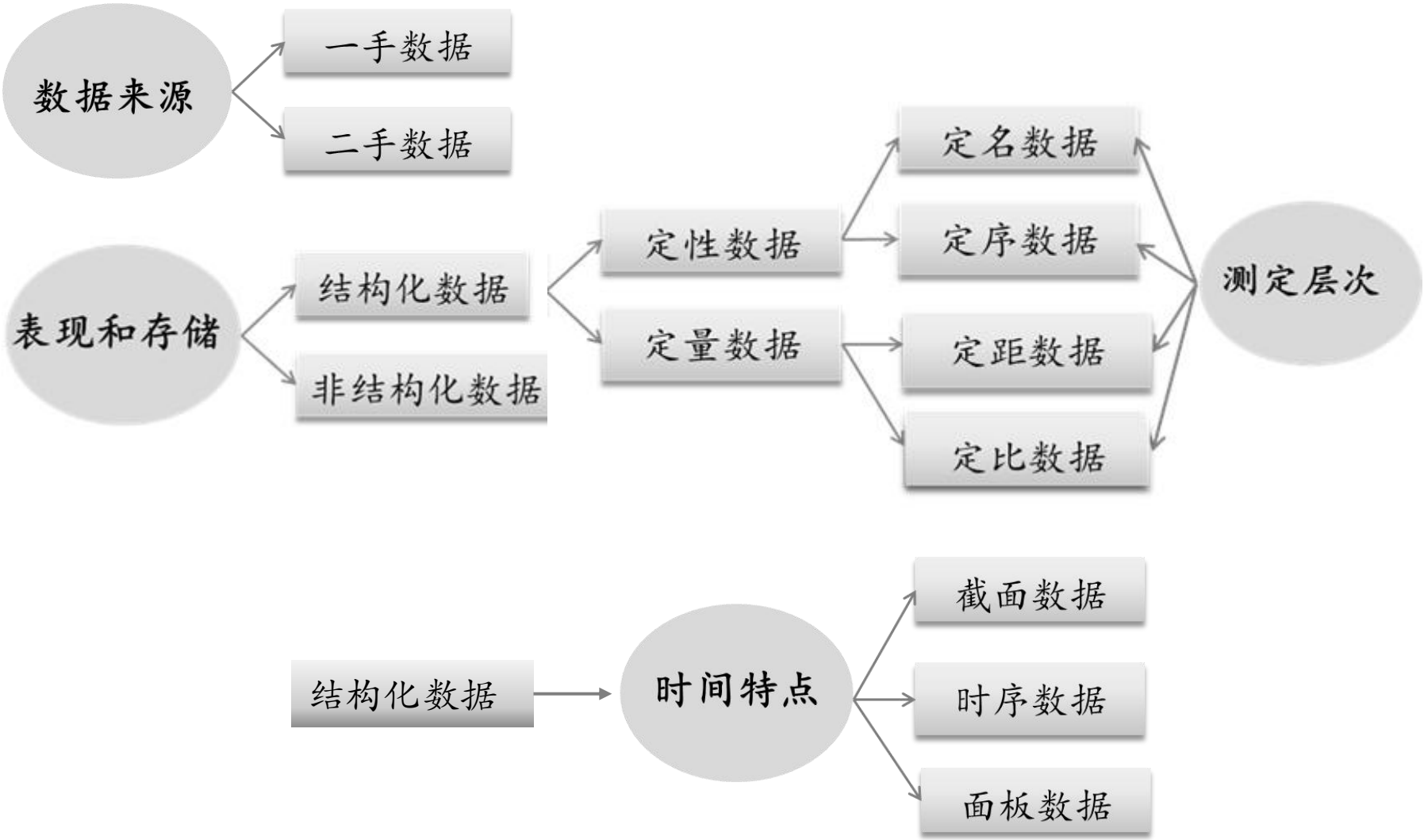
社交网络的兴起——社交链数据

物联网技术的成熟——车联网数据

生物信息技术的进步——微阵列数据



（二）数据的类型


数据来源
一手数据
二手数据
表现和存储
结构化数据
非结构化数据
定性数据
定量数据
定名数据
定序数据
定距数据
定比数据
测定层次
结构化数据
时间特点
截面数据
时序数据
面板数据

（二）数据的类型

1. 按照数据来源的数据分类

（1）一手数据(Primary Data)/原始数据(Raw Data)

- 概念：是指通过调查、测定等方式直接从各个调查单位收集的，尚未经过整理的数据。
- 举例：

  国家统计局开展人口普查获得我国人口的相关数据

  企业开展顾客满意度调查获得的数据

  上海财经大学暑期实践开展的千村调查获得的原始数据

（二）数据的类型

（2）二手数据(Secondary Data)/次级数据

- 概念：相对于一手数据而言的，指那些并非为正在进行的研究而是为其他目的已经收集并加工整理过的数据。二手数据往往是公开发表的数据。
- 举例：

① 从统计年鉴上获得我国人口普查的数据

② 从企业年报上取得的企业财务数据

③ 从教务处取得的老师授课的评教数据

④ 从中国健康与营养调查(CHNS)获得的健康和营养状况数据

（二）数据的类型

2. 按数据表现和存储的数据分类

（1）结构化数据(Structured Data)

- 概念：是指可以使用关系型数据库表示和存储，表现为二维形式的数据。一般特点是：数据以行为单位，一行数据表示一个观测对象的信息，一列表示一个变量。每一行数据的属性是相同的。

（二）数据的类型

- 结构化数据举例：

表 2.1　包含学生信息的样本数据集

学生 ID	年级	绩点（GPA）	…
⋮	⋮	⋮	⋮
1034262	四年级	3.24	…
1052663	二年级	3.51	…
1082246	一年级	3.62	…
⋮	⋮	⋮	⋮

一个观测对象（学生）的观测值

一个变量及其变量值

（二）数据的类型

- 几个相关概念：

✓ 数据集(Data Set)：是指用于特定研究而搜集的数据形成的集合。

✓ 观测对象(Element)：是指数据收集过程中被观测的单位，又被称为数据对象、记录、案例等。

✓ 变量(Variable)：指观测对象某个方面的特性，又被称为属性、特征等。

✓ 观测值(Observation)：指从一个观测对象收集来的所有信息（数据）。

✓ 变量值(Variable Value)：指所有观测对象在一个变量上的各种取值（数据）。

（二）数据的类型

表 2. 2　上市公司信息表

变量

观测值

Company	Stock Exchange	Annual Sales($M)	Earn/ Sh.($)
Dataram	AMEX	73.10	0.86
EnergySouth	OTC	74.00	1.67
Keystone	NYSE	365.70	0.86
LandCare	NYSE	111.40	0.33
Psychemedics	AMEX	17.60	0.13

观测对象

数据集

（二）数据的类型

上海财经大学 SHANGHAI UNIVERSITY OF FINANCE AND ECONOMICS

（2）非结构化数据(Unstructured Data)

- 概念：不适合由关系型数据库的二维形式来表现的数据。主要形式有：包括所有格式的办公文本、XML、HTML、图片、音频、视频等。

文本

图片

音频

视频

（二）数据的类型

- 特点：

✓ 数据量大，蕴含高价值。据国际数据公司（IDC）的一项调查报告中指出：企业中80%的数据都是非结构化数据，且每年都按指数增长60%。

✓ 格式多样化，缺乏统一的计算技术。非结构化数据五花八门，每类数据都有各自的计算处理手段，比如语音识别、图像比对、文本搜索、图结构计算等，不存在一种适用于所有非结构化数据的通用处理技术。

（二）数据的类型

（3）结构化数据和非结构化数据的分析

- 结构化数据的分析已经形成较成熟的流程和技术。根据变量（数据）类型不同，有不同的分析方法。
- 非结构化数据需要转化为结构化数据后再进行相应的分析。

（二）数据的类型

3. 数据的测定层次(Scales of Measurement)

（1）数据的测定层次是美国心理学家史蒂文斯（Stanley Smith Stevens）于1968年提出的数据分类方式。

（2）数据的测定层次是针对结构化数据提出的。数据测定层次的分类也是变量的分类。

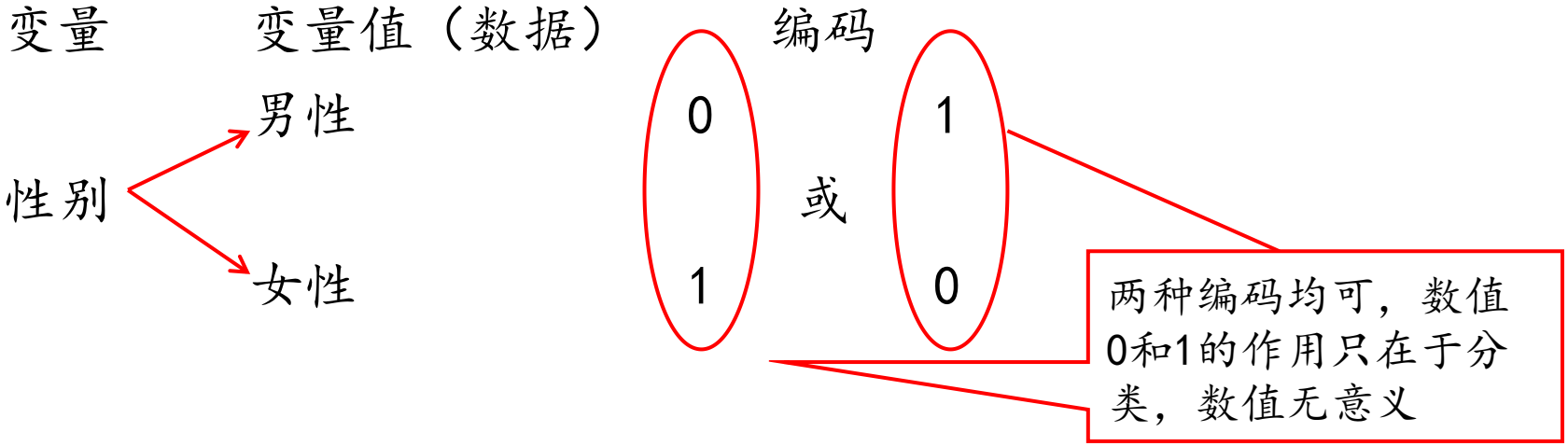
（3）数据的测定层次考虑了变量的特性和数学运算的功能特点。

（二）数据的类型

	数据测定层次	运算功能	特性	举例
定性数据	1、定名测定	计数	分类	案件类型、性别、眼球颜色、邮政编码
定性数据	2、定序测定	计数	分类	企业等级、矿石硬度、等级成绩
定性数据	2、定序测定	排序	排序	企业等级、矿石硬度、等级成绩
定量数据	3、定距测定	计数	分类	摄氏或华氏温度、日历日期
定量数据	3、定距测定	排序	排序	摄氏或华氏温度、日历日期
定量数据	3、定距测定	加、减	有基本的测量单位	摄氏或华氏温度、日历日期
定量数据	4、定比测定	计数	分类	商品数量、商品销售额、长度、年龄、质量、电流量等
定量数据	4、定比测定	排序	排序	商品数量、商品销售额、长度、年龄、质量、电流量等
定量数据	4、定比测定	加、减	有基本的测量单位	商品数量、商品销售额、长度、年龄、质量、电流量等
定量数据	4、定比测定	乘、除	有绝对零点	商品数量、商品销售额、长度、年龄、质量、电流量等

（二）数据的类型

- **定名测定**(Nominal Scale)/定名变量：定名测定只能按照事物的某种属性对其进行平行的分类或分组。每一类型都有特定的文字或数值编码进行标示，这种数值编码只是代号而无量的意义。定名测定需要遵循互斥原则（每一个观测对象只能划归到某一种类型中）和穷尽原则（所有观测对象都可归属到适当的类型中）。
- **举例**：



（二）数据的类型

- **定序测定**(Ordinal Scale)/定序变量：定序测定不仅具有定名测定的特点，将所有的观测对象按照互斥和穷尽的原则加以分类，而且各种类型之间具有某种意义的等级差异，从而形成一种确定的排序。不过，各种等级差异之间的间距大小不能具体测定。
- **举例：**

  企业按照经营效益划分为：一级企业、二级企业等。

  职工按照受正规教育划分为：大学毕业、中学毕业、小学毕业。

（二）数据的类型

- **定距测定**(Interval Scale)/定距变量：定距测定不仅能将事物区分为不同类型并进行排序，而且可以测定其间距大小。定距测定的量可以进行加或减的运算，但却不能进行乘或除的运算，因为在等级序列中没有固定的、有确定意义的零点。
- **举例：**

学生百分制成绩：学生甲得分为90分，学生乙得分1分，可以说甲比乙多得89分，不说甲的成绩是乙的90倍，因为“零”分并不意味着学生毫无知识。（值得商榷）。

温度：可以用摄氏度或者华氏度来表示，比如两地某时刻温度分别为30度和15度，但不能说30度是15度的2倍，因为没有绝对零度，而且用两种单位测量，比值也不一样，缺少实际含义。

（二）数据的类型

- **定比测定**(Ratio Scale)/定比变量：定比测定可以进行加减的运算，也可以进行乘除的运算。这是因为定比测定中存在绝对固定的、非任意的零点，零点是有实质意义的。几乎所有的物理量都是定比测定，同时，绝大多数的经济变量也可以进行定比测定。
- **举例：**

年龄：甲某人今年60岁，乙某人今年30岁，我们既可以说甲比乙年长30岁，又可以说甲的年龄是乙的2倍。

（二）数据的类型

上海财经大学 SHANGHAI UNIVERSITY OF FINANCE AND ECONOMICS

➢ 思考题2.2

指出以下数据是哪种测定层次？

变量	变量值	测定层次
行业	零售业、旅游业、汽车制造业等	定名数据
身高	1.7米、1.68米、1.55米、1.60米等	定比数据
产品质量等级	一等品、二等品、次品	定序数据
体重	55kg、45kg、60kg、75kg等	定比数据
顾客满意程度	非常满意、满意、一般、不满意、非常不满意	定序数据
收入分类	2000元以下；2000-5000元；5000-10000元；10000元以上	定序数据

（二）数据的类型

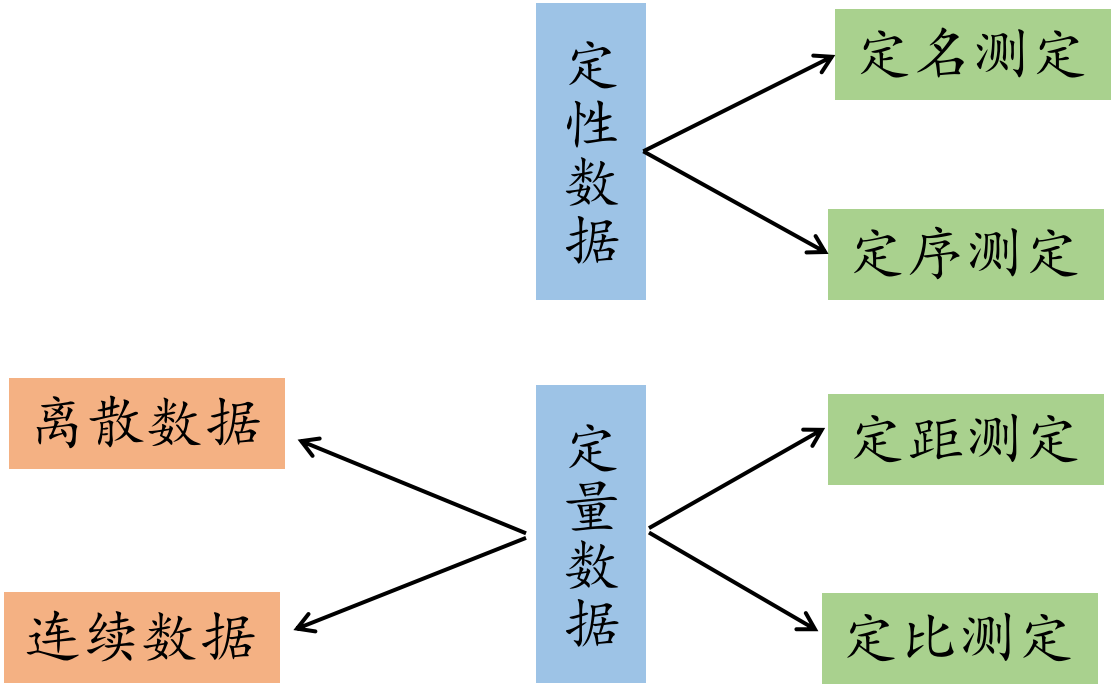
（4）**数据测定层次的重要性**

- 测定层次决定了数据包含的信息量：测定层次越高，包含的信息量越多。其中，定名测定包含的信息量最少；定比测定包含的信息量最多。
- 测定层次决定了后续数据分析使用的分析方法。测定层次越高，可以使用的分析方法越丰富。

（二）数据的类型

数据测定层次	数据分析举例	
	描述分析	推断分析
1、定名测定	频数、频率、众数	列联表χ^2检验
2、定序测定	百分位数、中位数、秩相关系数	游程检验、符号检验
3、定距测定	全距、均值、方差（标准差）、皮尔逊相关系数	t检验、F检验
4、定比测定	几何平均、调和平均	回归分析、因子分析

（二）数据的类型



（二）数据的类型

（5）**定性数据**(Qualitative Data)/定性变量：定性数据是用于反映事物的性质，或者规定事物类别的数据。定性数据采用定名测定或定序测定。定性数据一般表现为文字，或者通过编码转化为相应的数值，但是数值无量的意义。

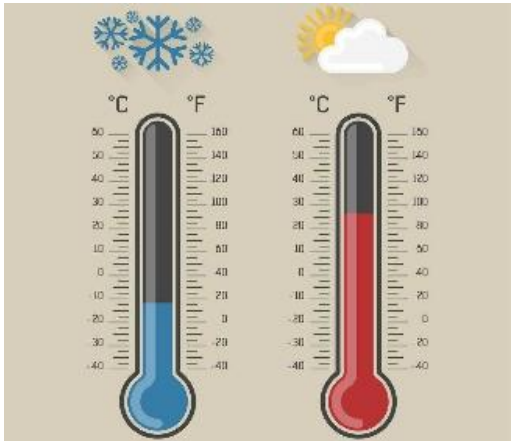
定量数据(Quantitative Data)/定量变量：定量数据用于反映事物数量上的差异。定量数据采用定距测定或定比测定，表现为数值，且数值存在量的意义。

（二）数据的类型

（6）定量数据分类

离散数据(Discrete Data)/离散变量：离散变量通过计数方式取得数据，只取有限个变量值或无限可列个变量值。例如：人数、顾客数量、商品数量等。

连续数据(Continuous Data)/连续变量：连续变量取值是一直叠加上去的，在一定区间内可以取任意值。例如：温度、体重、速度等。



（二）数据的类型

上海财经大学 SHANGHAI UNIVERSITY OF FINANCE AND ECONOMICS

➢ 案例2.1（续）：上海市民消费需求调研分析

消费支出的制约因素

Q14. 请您根据以下因素对您家庭消费支出的制约作用在 0 分~5 分之间进行打分，其中“0 分”表示“没有制约作用”，“5 分”表示“制约作用非常大”。在对应分值的[ ]内打勾选择。

制约因素	0 没有制约作用	1	2	3	4	5 制约作用非常大
收入水平	[ ]	[ ]	[ ]	[ ]	[ ]	[ ]
物价水平	[ ]	[ ]	[ ]	[ ]	[ ]	[ ]
存钱意愿	[ ]	[ ]	[ ]	[ ]	[ ]	[ ]
投资意愿	[ ]	[ ]	[ ]	[ ]	[ ]	[ ]
市场现有消费品满足需要的情况	[ ]	[ ]	[ ]	[ ]	[ ]	[ ]
用于消费的时间多少	[ ]	[ ]	[ ]	[ ]	[ ]	[ ]
是否愿意多花钱消费	[ ]	[ ]	[ ]	[ ]	[ ]	[ ]

通过打分题的设计，将定性变量转化为离散的定量变量。

（二）数据的类型

上海财经大学 SHANGHAI UNIVERSITY OF FINANCE AND ECONOMICS

收入、消费支出的基本情况
Q5. 您家庭的人均月收入（包括所有收入）为（　　）
①1000 元及以下　②1001～2000 元　③2001～3000 元　④3001～5000 元
⑤5001～10000 元　⑥10001～15000 元　⑦15001～20000 元　⑧20000 元以上

收入、年龄等敏感问题，通过设计为分段选择，降低被访者敏感度，降低拒访率。

通过分段选择题的设计，将定量变量转化为定性变量（定序变量）。

（二）数据的类型

4. 按照时间特点的数据分类——针对结构化数据

（1）截面数据(Cross-Sectional Data)

- 概念：是指搜集不同观测对象（个体、公司、地区、国家等）在同一时间的数据。
- 举例：某月份各个电厂的发电量
  世界各个国家今年的GDP
  我校各专业今年录取的本科生人数
- 分析：比较、分析各观测对象之间或各变量之间的关系（静态相关）。

（2）时序数据(Time Series Data)

- 概念：是指搜集同一观测对象（个体、公司、地区、国家等）在不同时间的数据。
- 举例：近10年来，某个电厂的年发电量
  近10年来，我国历年的GDP
  近1年来，某只上市公司股票的每日收盘价
  近10年来，我校统计学专业每年录取的本科生人数
- 分析：可用于时间序列预测：通过对历史时间序列的分析，探寻时间序列的动态相关性，并利用该动态相关，对时间序列的未来走向进行预测。也可用于回归分析，利用多个时间序列数据，探讨它们之间的相互作用关系。

（二）数据的类型

（3）面板数据(Panel Data)

- 概念：面板数据也既有截面又有时间序列的特点。在时间序列上取多个截面，在这些截面上的观测对象相同。
- 举例：下面的表格是美国150个城市与犯罪相关的2年的面板数据。

表1-3　　2年的城市犯罪相关的面板数据

观测值	城市	年份	谋杀数	人口数量	失业率	警察人数
1	1	1986	5	350000	8.7	440
2	1	1990	8	359200	7.2	471
3	2	1986	2	64300	5.4	75
4	2	1990	1	65100	5.5	75
⋮	⋮	⋮	⋮	⋮	⋮	⋮
299	150	1986	25	543000	4.3	520
300	150	1990	32	546200	5.2	493

（二）数据的类型

➢ 案例2.2：截面数据分析案例

——我国31个省、直辖市和自治区按照消费性支出的聚类分析

表2-3列出了1999年全国31个省、直辖市和自治区的城镇居民家庭平均每人全年消费性支出的八个主要变量数据。这八个变量分别是：

x_1：食品；　　x_2：衣着；　　x_3：家庭设备用品及服务；

x_4：医疗保健；　　x_5：交通和通讯；　　x_6：娱乐教育文化服务；

x_7：居住；　　x_8：杂项商品和服务.

对各地区按照消费支出特点进行聚类分析。

（二）数据的类型

上海财经大学 SHANGHAI UNIVERSITY OF FINANCE AND ECONOMICS

表2—3　消费性支出数据

地区	x_1	x_2	x_3	x_4	x_5	x_6	x_7	x_8
北京	2959.19	730.79	749.41	513.34	467.87	1141.82	478.42	457.64
天津	2459.77	495.47	697.33	302.87	284.19	735.97	570.84	305.08
河北	1495.63	515.9	362.37	285.32	272.95	540.58	364.91	188.63
山西	1406.33	477.77	290.15	208.57	201.5	414.72	281.84	212.1
内蒙古	1303.97	524.29	254.83	192.17	249.81	463.09	287.87	192.96
辽宁	1730.84	553.9	246.91	279.81	239.18	445.2	330.24	163.86
吉林	1561.86	492.42	200.49	218.36	220.69	459.62	360.48	147.76
黑龙江	1410.11	510.71	211.88	277.11	224.65	376.82	317.61	152.85
上海	3712.31	550.74	893.37	346.93	527	1034.98	720.33	462.03
江苏	2207.58	449.37	572.4	211.92	302.09	585.23	429.77	252.54
浙江	2629.16	557.32	689.73	435.69	514.66	795.87	575.76	323.36
安徽	1844.78	430.29	271.28	126.33	250.56	513.18	314	151.39
福建	2709.46	428.11	334.12	160.77	405.14	461.67	535.13	232.29
江西	1563.78	303.65	233.81	107.9	209.7	393.99	509.39	160.12
山东	1675.75	613.32	550.71	219.79	272.59	599.43	371.62	211.84

（二）数据的类型

上海财经大学 SHANGHAI UNIVERSITY OF FINANCE AND ECONOMICS

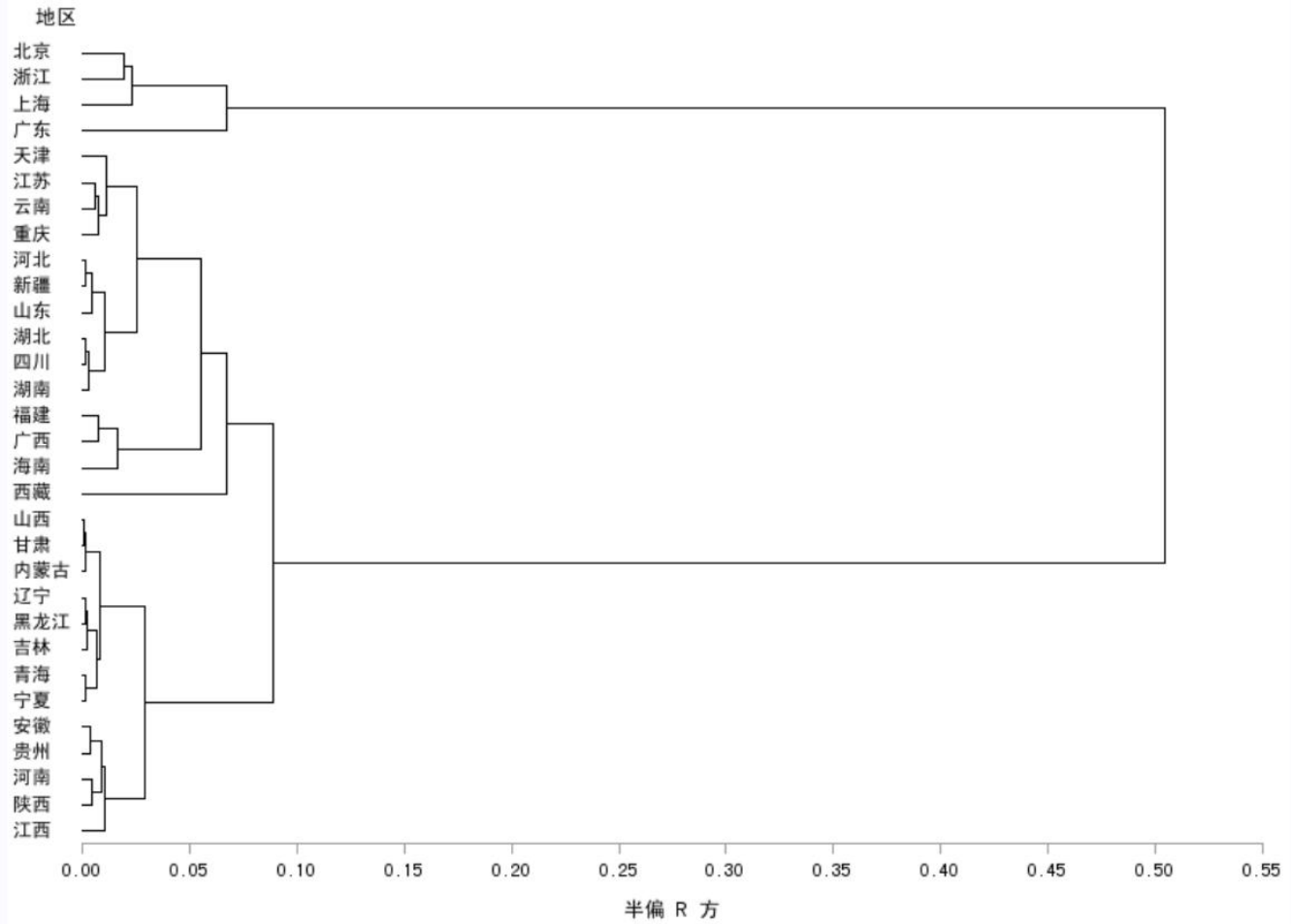
表2—3　消费性支出数据（续表）

地区	x_1	x_2	x_3	x_4	x_5	x_6	x_7	x_8
河南	1427.65	431.79	288.55	208.14	217	337.76	421.31	165.32
湖北	1783.43	511.88	282.84	201.01	237.6	617.74	523.52	182.52
湖南	1942.23	512.27	401.39	206.06	321.29	697.22	492.6	226.45
广东	3055.17	353.23	564.56	356.27	811.88	873.06	1082.82	420.81
广西	2033.87	300.82	338.65	157.78	329.06	621.74	587.02	218.27
海南	2057.86	186.44	202.72	171.79	329.65	477.17	312.93	279.19
重庆	2303.29	589.99	516.21	236.55	403.92	730.05	438.41	225.8
四川	1974.28	507.76	344.79	203.21	240.24	575.1	430.36	223.46
贵州	1673.82	437.75	461.61	153.32	254.66	445.59	346.11	191.48
云南	2194.25	537.01	369.07	249.54	290.84	561.91	407.7	330.95
西藏	2646.61	839.7	204.44	209.11	379.3	371.04	269.59	389.33
陕西	1472.95	390.89	447.95	259.51	230.61	490.9	469.1	191.34
甘肃	1525.57	472.98	328.9	219.86	206.65	449.69	249.66	228.19
青海	1654.69	437.77	258.78	303	244.93	479.53	288.56	236.51
宁夏	1375.46	480.89	273.84	317.32	251.08	424.75	228.73	195.93
新疆	1608.82	536.05	432.46	235.82	250.28	541.3	344.85	214.4

（二）数据的类型

上海财经大学 SHANGHAI UNIVERSITY OF FINANCE AND ECONOMICS

数据分析：采用系统聚类（离差平方和）方法对我国31个省、直辖市和自治区按照消费性支出特点进行聚类分析。得到树形图如下：



（二）数据的类型

分析结论：根据树形图可知，31个地区分为以下三类：

- ✓ 第Ⅰ类：北京、浙江、上海和广东。这些都是我国经济最发达、城镇居民消费水平最高的沿海地区。
- ✓ 第Ⅱ类：天津、江苏、云南、重庆、河北、新疆、山东、湖北、四川、湖南、福建、广西、海南和西藏。这些地区在我国基本上属于经济发展水平和城镇居民消费水平中等的地区。
- ✓ 第Ⅲ类：山西、甘肃、内蒙古、辽宁、黑龙江、吉林、青海、宁夏、安徽、贵州、河南、陕西和江西。这些地区在我国基本上属于经济较落后地区，城镇居民的消费水平也是较低的。

该聚类结果能够比较好地反映各地区真实的消费水平。

➢ 案例2.3：截面数据分析案例

——顾客满意度与顾客忠诚度的关系分析

按照传统的管理和营销理论，提高顾客满意度是培育忠诚顾客的有效途径。利用某电信公司2006年的调查数据，试分析顾客满意度与顾客忠诚度的作用机制。

该电信公司从其客户群中按照年龄、性别等指标等比例分层抽样，并采用电话调查的方式，共调查了1500份问卷，经过预处理后，实际有效问卷为1153份。

其中，顾客满意度用3个观测变量进行衡量：总体满意度（SATI_1）、对预期的满足（SATI_2）和与理想的差距（SATI_3）。顾客忠诚度由4个观测变量进行衡量：推荐他人使用的可能性（LOYA_1）、继续使用的可能性（LOYA_2）、增大使用量的可能性（LOYA_3）、使用其他业务的可能性（LOYA_4）。调查采用10级尺度李克特量表，顾客按照自身的满意程度/忠诚程度进行打分，其中，1表示根本不满意/根本不可能，10表示十分满意/十分可能。

（二）数据的类型

数据分析及分析结论：

（1）顾客按照满意程度进行聚类分析，结果如下：

表 2-4　顾客满意度分组后观测值个数、各观测变量的均值

组　别	组 1（不满意顾客）	组 2（满意顾客）	组 3（非常满意顾客）
观测值个数	10	319	824
观测值比例	0.87%	27.67%	71.47%
SATI_1	3.200	6.881	8.636
SATI_2	2.500	6.467	8.682
SATI_3	2.500	6.179	8.551

根据表2-4，按照顾客满意度的高低可将顾客分为3类：（1）组1为不满意顾客，记为SATI1，这类顾客在各个观测值上的得分均值都低于3.5分，得分很低；（2）组2为满意顾客，记为SATI2，这类顾客在各个观测值上的得分均值都在6分到7分之间，超过了量表的中点，即5分；（3）组3为非常满意顾客，记为SATI3，这类顾客在各个观测值上的得分均值均超过8.5分，得分偏高。

（二）数据的类型

（2）顾客按照忠诚程度进行聚类分析，结果如下：

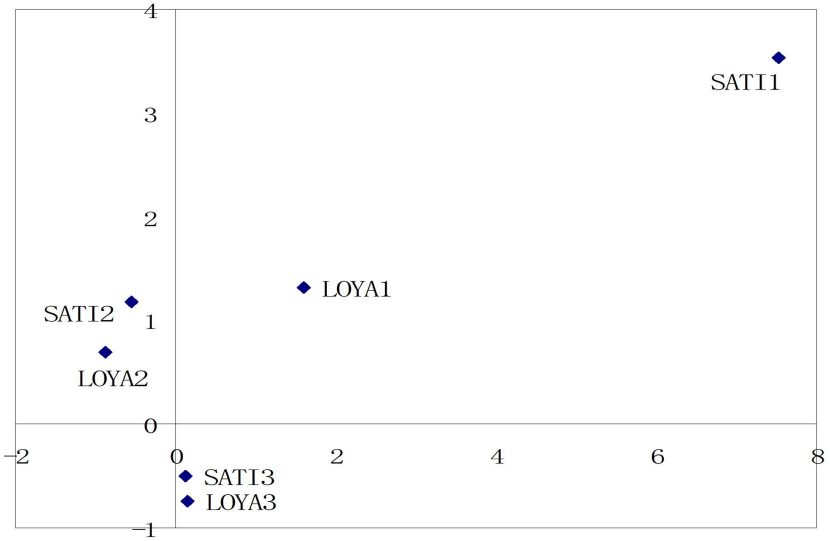
表 2-5　顾客忠诚度分组后观测值个数、各观测变量的均值

组　别	组 1（不忠诚顾客）	组 2（潜在忠诚顾客）	组 3（忠诚顾客）
观测值个数	145	385	623
观测值比例	12.58%	33.39%	54.03%
LOYA_1	3.262	4.846	8.727
LOYA_2	8.462	8.612	9.444
LOYA_3	2.227	6.849	8.637
LOYA_4	2.268	7.296	7.887

根据表2-5，按照顾客忠诚度的高低可将顾客分为3类：（1）组1为不忠诚顾客，记为LOYA1，这类顾客除了继续使用的可能性均值偏高以外，其他观测变量的均值都低于3.5分，得分很低；（2）组2为潜在忠诚顾客，记为LOYA2，这类顾客除了推荐他人使用可能性的得分均值低于5分以外，其他观测变量的均值都落在6.5分到9分之间；（3）组3为忠诚顾客，记为LOYA3，这类顾客在所有观测值上的得分均值均超过7.5分，得分偏高。

（二）数据的类型

（3）对顾客按照顾客满意度和顾客忠诚度分类后，形成两个分类变量，对这两个分类变量进行多重对应分析．以了解分类变量各类别之间的对应关系。



根据对应图可知，顾客满意度与顾客忠诚度之间存在着明显的对应关系。其中，非常满意顾客与忠诚顾客很接近；满意顾客与潜在忠诚顾客很接近；同时还可以看到，虽然不满意顾客与不忠诚顾客并不接近——这可能与不满意顾客人数仅为10人，人数过少有关——但是它们仍然落在同一区域之内，具有一定联系。

（二）数据的类型

表 2-6　　顾客满意度和顾客忠诚度的列联表

分组频数	不忠诚顾客	潜在忠诚顾客	忠诚顾客	小计
不满意顾客	9	1	0	10
满意顾客	62	169	88	319
非常满意顾客	74	215	535	824
小计	145	385	623	1153

两个变量之间的对应关系在列联表中也有所体现。从顾客满意度分类的角度来看，在不满意顾客中，不忠诚顾客比例最高，占90%；满意顾客中，潜在忠诚顾客的比例最高，占52.9%；非常满意顾客中，忠诚顾客的比例最高，占64.9%。从顾客忠诚度分类的角度来看，在忠诚顾客中，非常满意顾客占85.9%。非常满意顾客占比随着顾客忠诚度的下降，依次降低为55.8%和51.0%。

（二）数据的类型

➢ 案例2.4：时间序列数据分析案例

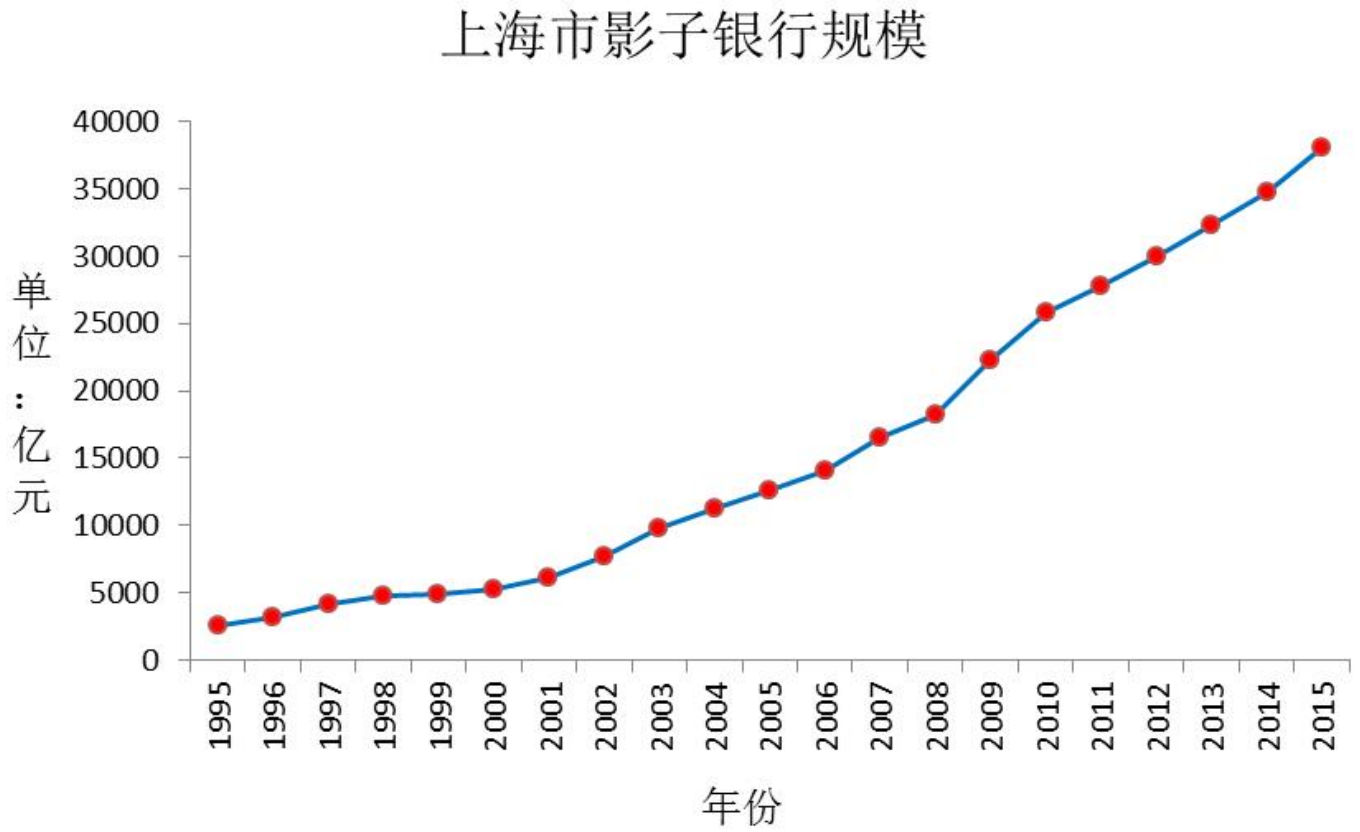
——利用ARMA模型预测上海市影子银行的规模

随着各国经济的不断发展，其金融业发展以及金融创新会造成影子银行体系的形成和扩张。影子银行作为一种新型贷款通道，降低了有需求的中小企业及自然人借贷融资难的问题，一定程度上促进了经济发展。

上海作为我国经济中心，影子银行具有怎样的规模？根据国民经济核算原理，测算了1995年至2015年上海市影子银行的规模。

利用1995年至2015年上海市影子银行规模的数据，建立合适的时间序列模型，对2016年至2018年上海市影子银行的规模进行预测。

（二）数据的类型


上海市影子银行规模
单位：亿元
40000
35000
30000
25000
20000
15000
10000
5000
0
1995
1996
1997
1998
1999
2000
2001
2002
2003
2004
2005
2006
2007
2008
2009
2010
2011
2012
2013
2014
2015
年份

（二）数据的类型

数据分析及分析结论：

（1）模型结果：在经过平稳检验以及模型精度等方面的考虑，1995年至2015年上海市影子银行时间序列可采用ARMA（1,1）模型进行拟合。即：

$$y_t = 1.014y_{t-1} + \varepsilon_t + 0.6083\varepsilon_{t-1}$$

（2）模型评价：为了评价ARMA（1,1）模型对上海市影子银行时间序列的拟合效果，利用ARMA（1,1）模型测算2013年至2015年上海市影子银行规模的拟合值，与实际规模进行比较。由于每年的误差均不超过5%，这说明ARMA(1,1)模型的拟合效果较好。

表2-7　2013-2015年上海市影子银行规模拟合值和实际值的比较

年份	测算值	预测值	误差率
2013	32303.49	32859.62	1.7%
2014	34710.97	36315.50	4.6%
2015	38013.56	38948.67	2.4%

（二）数据的类型

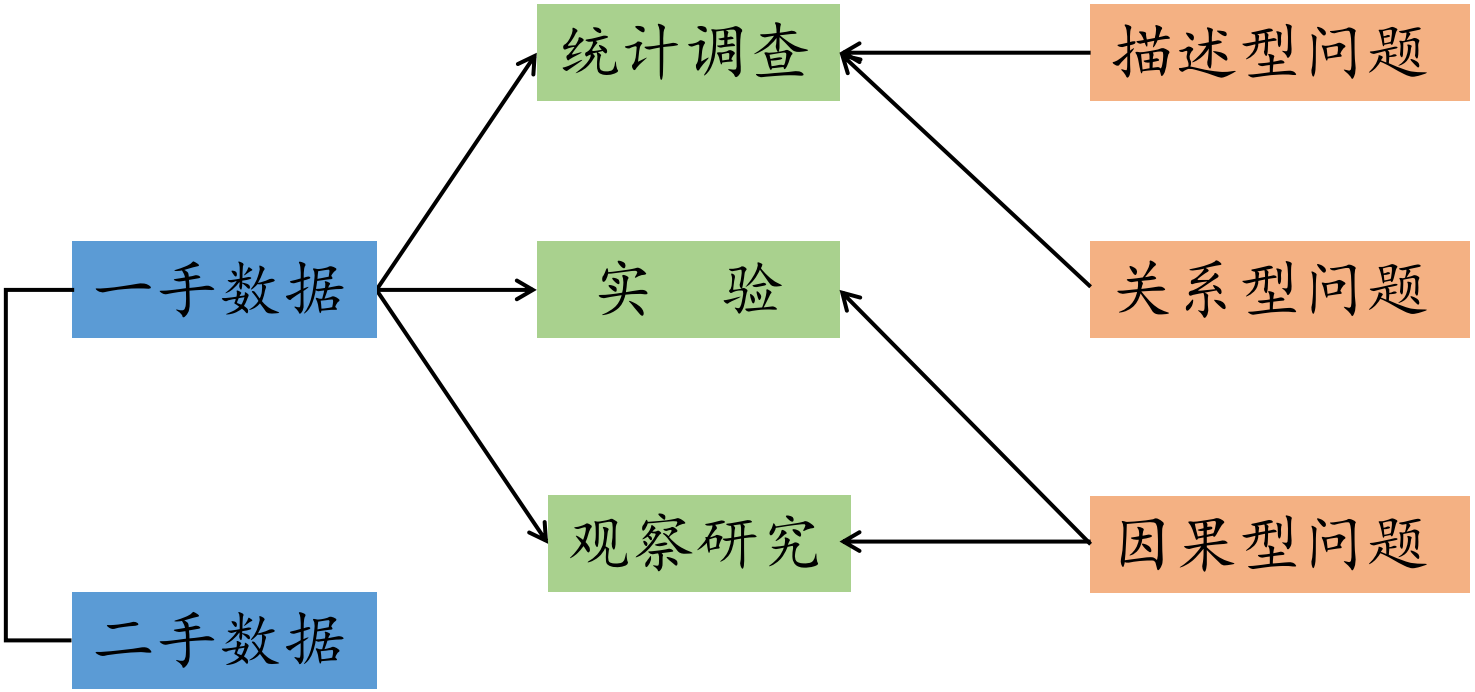
（3）模型预测：利用ARMA（1,1）模型对2016年至2018年上海市影子银行规模进行预测，预测结果如下：

表2-8　2016-2018年上海市影子银行规模预测值

上海市影子银行规模	2016 年	2017 年	2018 年
预测值（亿元）	43347.31	50312.06	58571.30

利用ARMA(1,1)预测上海市影子银行的规模，上海市影子银行规模将逐年增大，2016年超过4000亿元，2017年突破5000亿元，而2018年将逼近6000亿元。

四、传统的数据来源



四、传统的数据来源

上海财经大学
SHANGHAI UNIVERSITY OF FINANCE AND ECONOMICS

（一）统计调查

（二）实验

（三）观察研究

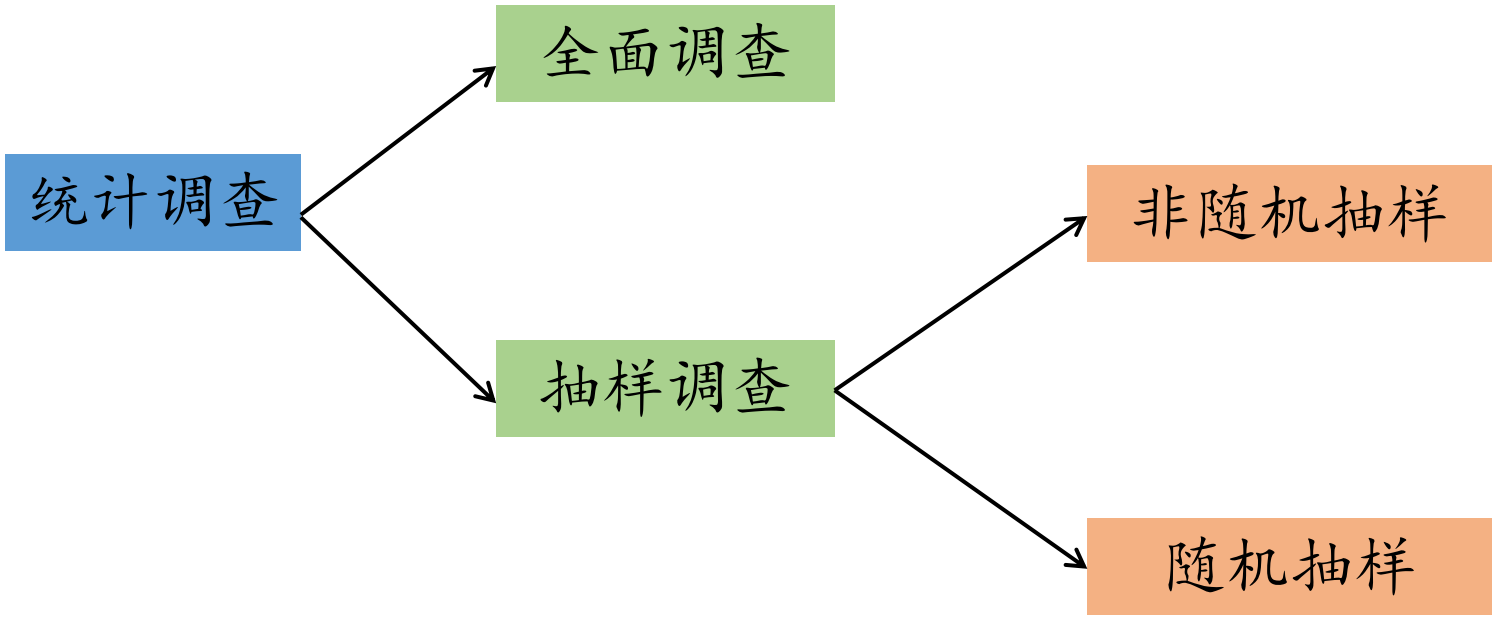
（一）统计调查

1. 定义

统计调查（Survey）是根据调查的目的与要求，运用科学的调查方法，有计划、有组织地搜集数据信息资料的过程。统计调查对调查对象的行为不进行任何控制，仅提出问题，然后对调查对象的回答结果进行整理、编码和分析。

（一）统计调查

2. 组织方式



（一）统计调查

（1）全面调查（Census Survey）

- 定义：指为了某些特定的目的而组织的、对总体中的全部个体都进行的调查。
- 优点：
  - ✓ 调查涵盖了所有个体
  - ✓ 精确性较高（不是所有情况精确性都较高）
- 缺点：
  - ✓ 耗费大量的人、财、物力
  - ✓ 可能无法涵盖所有的个体
  - ✓ 如果调查问卷设计得不好或者测量精度不够高，调查结果可能无法保证精确性
- 普查和统计报表制度属于全面调查。

（一）统计调查

- 普查
  - ✓ 定义：普查是为了某种特定目的而专门组织的一次性全面调查，用以搜集重要国情国力和资源状况的全面资料，为政府制定规划、方针政策提供依据。
  - ✓ 举例：10年一次的全国人口普查（2010年第六次全国人口普查）；5年一次的全国经济普查（2018年第四次全国经济普查）
  - ✓ 原则：
    - 规定统一的标准时点（2010年第六次全国人口普查的标准时点为2010年11月1日零时；2018年第四次全国经济普查的标准时点为2018年12月31日）
    - 规定统一的普查期限（普查的调查时间期限）
    - 规定普查的项目和指标（2010年第六次全国人口普查调查问卷分为四种调查表，均由国家统计局统一编制并供调查使用）

（一）统计调查

➢ 案例2.5：第六次全国人口普查调查表

第六次全国人口普查表共分为：《第六次全国人口普查表短表》（简称普查表短表）、《第六次全国人口普查表长表》（简称普查表长表）、《第六次全国人口普查表短表（供港澳台和外籍人员使用）》（简称境外人员普查表）和《第六次全国人口普查死亡人口调查表》（简称死亡人口调查表）四种表。

下面展示短表的调查表。

——参考 国家统计局网站. 网址链接：
http://www.stats.gov.cn/ztjc/zdtjgz/zgrkpc/dlcrkpc/dlcrkpcgg/201010/t20101025_70245.htm

（一）统计调查

上海财经大学
SHANGHAI UNIVERSITY OF FINANCE AND ECONOMICS

第六次全国人口普查表

短表封面

地址：

_____ 省（自治区、直辖市） □□

_____ 地（市） □□

_____ 县（市、区） □□

_____ 乡（镇、街道） □□□

_____ 普查区 □□□

_____ 普查小区 □□□

本普查小区短表登记

总户数_____□□□ 总人数_____□□□□ 总张数_____□□□

普查员（签字）：　　普查指导员（签字）：

（一）统计调查

上海财经大学 SHANGHAI UNIVERSITY OF FINANCE AND ECONOMICS

经国务院批准进行第六次全国人口普查
人口普查的标准时点为2010年11月1日零时
人口普查的原始资料不向任何单位和个人提供，仅供汇总使用
公民应履行如实申报普查项目的义务

第六次全国人口普查表短表

表　　号：表 6 0 1 表
制定机关：国家统计局
　　国务院第六次全国人口普查办公室
批准文号：国发［2009］23号
有效期至：2010年12月

本户地址和地址码：＿＿＿＿ 县、市区 □□ ＿＿＿＿ 乡、镇街道 □□□ ＿＿＿＿ 普查区 □□□ ＿＿＿＿ 普查小区 □□□ 建筑物编号 ＿＿＿＿ □□

H1. 户编号	H2. 户别	H3. 本户应登记人数		H4. 2009.11.1－2010.10.31 出生人口		H4. 2009.11.1－2010.10.31 死亡人口		H5. 本户住房建筑面积	H6. 本户住房间数
＿＿＿号 □□□	1. 家庭户 2. 集体户 □	2010年10月31日晚居住本户的人数 ＿＿人 □□	户口在本户，2010年10月31日晚未住本户的人数 ＿＿人 □□	男＿＿人 □	女＿＿人 □	男＿＿人 □	女＿＿人 □	＿＿＿平方米 □□□	＿＿＿间 □□

每　个　人　都　填　报

6周岁及以上（2004年10月31日以前出生）的人填报

R1. 姓名	R2. 与户主关系	R3. 性别	R4. 出生年月	R5. 民族	R6. 普查时点居住地	R7. 户口登记地	R8. 离开户口登记地时间	R9. 离开户口登记地原因	R10. 户口性质	R11. 是否识字	R12. 受教育程度
□	0. 户主 1. 配偶 2. 子女 3. 父母 4. 公父母或外公婆 5. 祖父母 6. 孙侯 7. 孙子女 8. 兄弟姐妹 9. 其他 □	1. 男 2. 女 □	出生于： ＿＿＿年 ＿＿＿月 □□□□	族 □□	1. 本普查小区 2. 本村（居）委会其他普查小区 3. 本乡（镇、街道）其他村（居）委会 4. 本县（市、区）其他乡（镇、街道） 5. 其他县（市、区），**请填写下面地址** 6. 港澳台或国外 □ □□ ＿＿＿＿省（区、市） □□ ＿＿＿＿地（市） □□ ＿＿＿＿县（市、区）	1. 本村（居）委会 2. 本乡（镇、街道）其他村（居）委会 3. 本县（市、区）其他乡（镇、街道） 4. 其他县（市、区），**请填写下面地址** 5. 户口待定→R11 □	1. 没有离开户口登记地→R10 2. 半年以下 3. 半年至一年 4. 一年至二年 5. 二年至三年 6. 三年至四年 7. 四年至五年 8. 五年至六年 9. 六年以上 □	1. 务工经商 2. 工作调动 3. 学习培训 4. 随迁家属 5. 投亲靠友 6. 拆迁搬家 7. 寄挂户口 8. 婚姻嫁娶 9. 其他 □	1. 农　业 2. 非农业 □	1. 是 2. 否 □	1. 未上过学 2. 小学 3. 初中 4. 高中 5. 大学专科 6. 大学本科 7. 研究生 □
□	1. 配偶 2. 子女 3. 父母 4. 公父母或外公婆 5. 祖父母 6. 孙侯 7. 孙子女 8. 兄弟姐妹 9. 其他 □	1. 男 2. 女 □	出生于： ＿＿＿年 ＿＿＿月 □□□□	族 □□	1. 本普查小区 2. 本村（居）委会其他普查小区 3. 本乡（镇、街道）其他村（居）委会 4. 本县（市、区）其他乡（镇、街道） 5. 其他县（市、区），**请填写下面地址** 6. 港澳台或国外 □ □□ ＿＿＿＿省（区、市） □□ ＿＿＿＿地（市） □□ ＿＿＿＿县（市、区）	1. 本村（居）委会 2. 本乡（镇、街道）其他村（居）委会 3. 本县（市、区）其他乡（镇、街道） 4. 其他县（市、区），**请填写下面地址** 5. 户口待定→R11 □	1. 没有离开户口登记地→R10 2. 半年以下 3. 半年至一年 4. 一年至二年 5. 二年至三年 6. 三年至四年 7. 四年至五年 8. 五年至六年 9. 六年以上 □	1. 务工经商 2. 工作调动 3. 学习培训 4. 随迁家属 5. 投亲靠友 6. 拆迁搬家 7. 寄挂户口 8. 婚姻嫁娶 9. 其他 □	1. 农　业 2. 非农业 □	1. 是 2. 否 □	1. 未上过学 2. 小学 3. 初中 4. 高中 5. 大学专科 6. 大学本科 7. 研究生 □

（超过五人的户，从第2张普查表起，户记录只填写“H1.户编号”）

（一）统计调查

- 统计报表制度
  - ✓ 定义：统计报表制度是按一定的表式和要求，自上而下统一布置，自下而上提供统计资料的一种统计调查方法。
  - ✓ 分类：
    - 按报表内容和实施范围不同，分为国家统计报表、部门统计报表和地方统计报表。
    - 按报送周期长短不同，分为日报、旬报、季报、半年报和年报。
    - 按填报单位不同，分为基层统计报表和综合统计报表。

（一）统计调查

- 案例2.6：北京市工业统计报表制度（2017年统计年报和2018年定期统计报表）

基层统计报表

年报

国家报表

二、报表目录

表号	报表名称	报告期别	统计范围	报送单位	报送时间及方式			页码
					报送单位	各区报市统计局	市统计局报国家	
（一）年报 1.单位基本情况统计								
101-1表	法人单位基本情况	年报	辖区内第二、第三产业规模（限额）以上法人单位（律师事务所视同法人单位）和其它有500万元以上在建项目的法人单位，在中关村国家自主创新示范区注册的第一产业法人单位及第二、第三产业规模（限额）以下法人单位	辖区内第二、第三产业规模（限额）以上法人单位（律师事务所视同法人单位），在中关村国家自主创新示范区注册的第一产业法人单位及第二、第三产业规模（限额）以下法人单位	2018年3月10日24时前网上填报	2018年3月23日18时前完成数据验收	2018年4月15日24时前	14

——参考 北京市政府信息公开专栏. 网址链接：

http://zfxxgk.beijing.gov.cn/110037/jctjbbzd53/2017-12/21/content_106ff970b4cc42ad9a8e4e147ef2f93d.shtml

(2) 抽样调查 (Sample Survey)

- 定义：指按一定的程序从总体中抽取样本进行调查。
- 优点：
  - ✓ 节省费用
  - ✓ 快速
  - ✓ 可以获得较高的精确性（主要指概率抽样）
- 缺点：
  - ✓ 可能对总体缺乏代表性（无偏样本vs有偏样本_Ch6）
  - ✓ 如果调查问卷设计得不好或者测量精度不够高，调查结果可能无法保证精确性

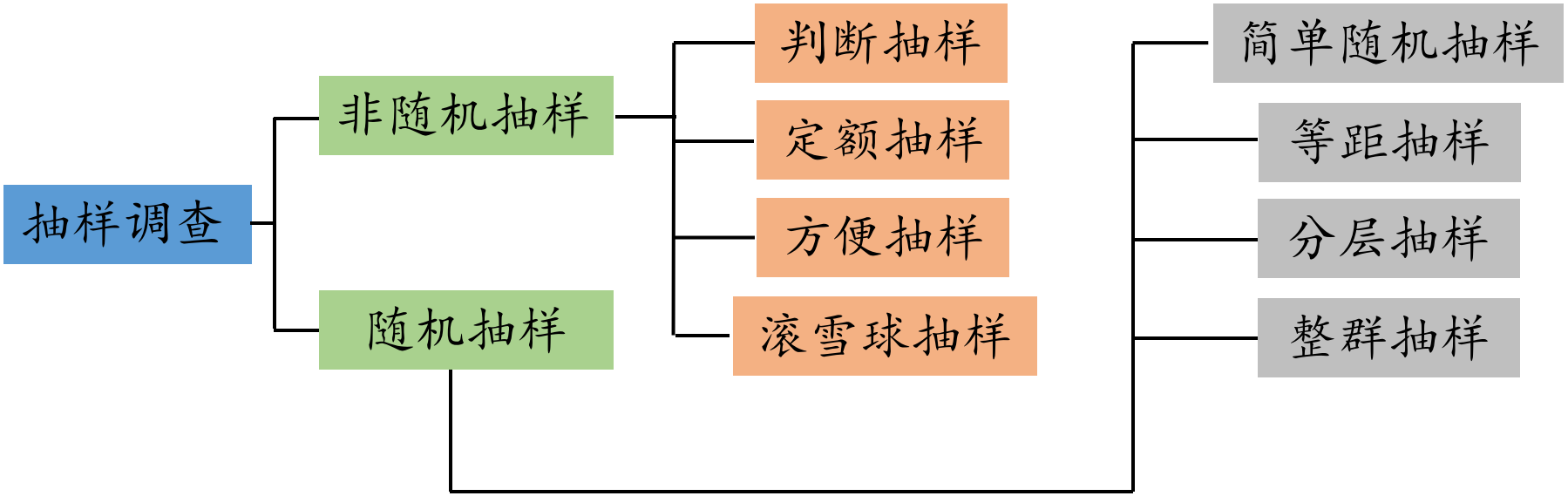
（一）统计调查

- 适用场合：
  - ✓ 某些不可能进行全面调查的情况。有些总体规模很大，或者是连续不断发生的过程。例如，对连续生产线上的产品质量的检查、海洋生物的研究、流动人口的调查等，只能进行抽样调查。有些检查是破坏性的，例如产品的寿命试验、破坏性试验、可靠性试验等，也只能进行抽样调查。
  - ✓ 虽然可能取得全面资料，但是不必进行全面调查的情况。例如，民意调查等。
  - ✓ 对全面调查资料进行验证和修正。例如，全国人口普查后的人口抽样调查。

（一）统计调查

上海财经大学
SHANGHAI UNIVERSITY OF FINANCE AND ECONOMICS

- 分类：



（一）统计调查

- 分类：
  - ✓ 非随机抽样（非概率抽样）
    - 不按照随机原则抽取样本。
    - 无法估计样本推断总体的精确度，不能利用样本对总体进行推断。
  - ✓ 随机抽样（概率抽样）
    - 按照随机原则抽取样本。每个样本都有一个事先确定的被抽中的概率，称为入样概率。
    - 可以估计样本推断总体的精确度，可以利用样本对总体进行推断。
    - 为了进行随机抽样，需要先列出总体中所有的抽样单元，并编上号码，即编制抽样框。

（一）统计调查

- **非随机抽样**：
  - ✓ 判断抽样（经验抽样）：抽样者根据自己的经验在总体中选择若干有代表性的单位组成样本进行调查。
  - ✓ 定额抽样：抽样者依据一定的标志将总体分为若干层，确定各层在总体中所占的比例。并按这些比例分配样本量在各层的数额，让调查员抽到每一层所需的“定额”为止。常用的定额标志是地理区域、性别、年龄等，定额抽样在民意测验和市场调查中用得比较多。
  - ✓ 方便抽样（随意抽样）：抽样者按照自己的方便，随意地抽取样本。例如：街头拦截调查。
  - ✓ 滚雪球抽样：往往用于对稀少群体的调查。抽样者先找出少数个体，通过这些个体了解其他个体，再由自己了解到的个体去发现更多的个体，以此类推，就像滚雪球一样，了解到的个体越来越多。

（一）统计调查

● 随机抽样

✓ 简单随机抽样（纯随机抽样）

- 定义：指对总体单位不作任何分类或排序，完全按随机原则逐个地抽取样本单位。
- 方法：抽签、使用随机数表、计算机产生随机数
- 适用：总体规模不大，内部差异不大。
- 具体抽样方式：放回抽样、无放回抽样

✓ 等距抽样（机械抽样、系统抽样）

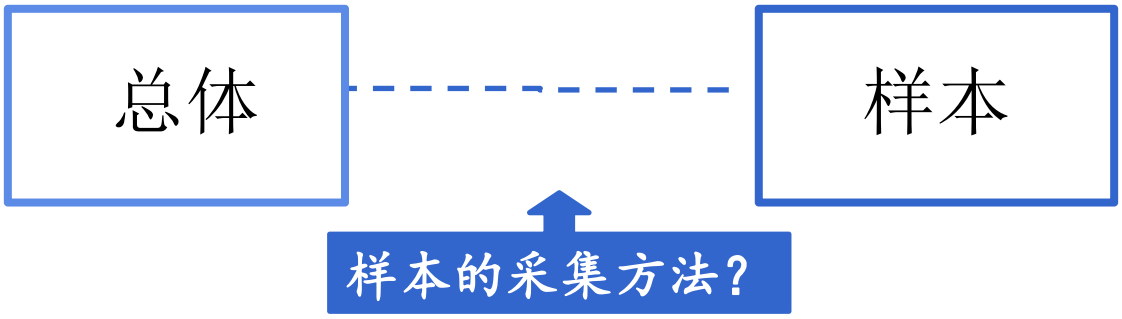
- 定义：将总体各单位按照某个标志顺序排列，然后按照一定的间隔抽取样本单位。
- 分类：**按排列依据的标志：**无关标志、有关标志；**按样本单位选取方式：**随机起点抽样、半距起点抽样、对称等距抽样。
- 缺点：**按照有关标志排列并采用随机起点抽样时，容易引起系统误差；半距起点抽样缺乏随机性。**

思政案例：统计调查

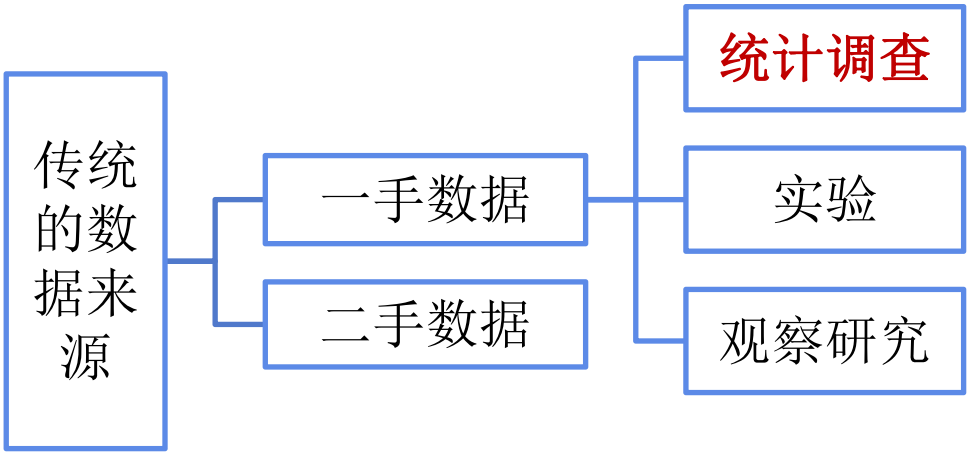
总体和样本

通过本小节学习，我们了解到，样本是从总体中抽取的部分观察单位，是总体中有代表性的一部分。样本能够一定程度上代表总体、通过样本的参数可以推断总体的统计量。

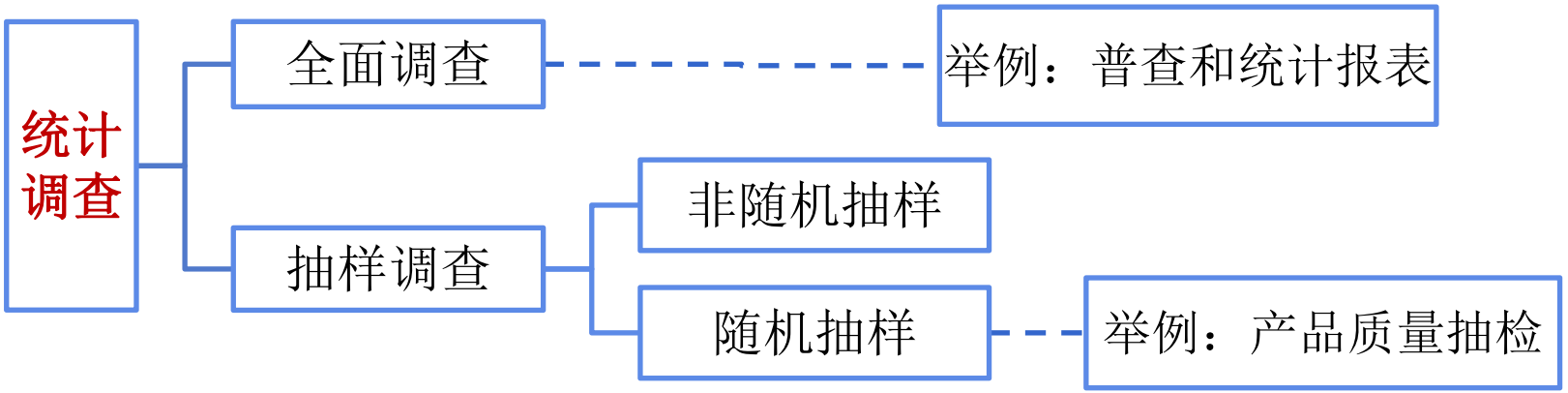
要使得推断可靠，对样本和抽样方法是有要求的。



思政案例：统计调查



统计调查（Survey）是根据调查的目的与要求，运用科学的调查方法，有计划、有组织地搜集数据信息资料的过程。统计调查对调查对象的行为不进行任何控制，仅提出问题，然后对调查对象的回答结果进行整理、编码和分析。



思政案例：统计调查

统计技术在我国社会调查中的起源和发展

- ✓ 中国有关统计方法、统计思想和工作可上溯到远古时代。《尚书·禹贡》把当时的中国分为九州，分别叙述了各地的物产、交通、植物等。

社会调查的“历史踪迹”

- ✓ 中国最早的调查著作：1927年《社会调查的原理及方法》
- ✓ 1928年，抽样理论的创立；30年代学者们积极主张将抽样调查的方法运用到中国的人口调查中。1936年《社会调查与统计学》(陈毅夫)
- ✓ 文革结束后，遭破坏的政府统计工作逐步恢复。统计调查方法作为一种社会研究模式，20世纪80年代初传入中国。
- ✓ 20世纪90年代以来，国家和地方统计部门完成了一系列重大国情国力普查，开展了一系列抽样调查和专项调查，并实施了新国民经济核算体系和新的国家统计报表制度。随着计算机技术、网络和通信技术等高新科技的发展，统计逐渐走向现代化。

思政案例：统计调查

全面调查（Census Survey）

- ✓ 定义：指为了某些特定的目的而组织的、对总体中的全部个体都进行的调查。普查和统计报表都属于全面调查。

普查

- ✓ 定义：普查是为了某种特定目的而专门组织的一次性全面调查，用以搜集重要国情国力和资源状况的全面资料，为政府制定规划、方针政策提供依据。

统计报表制度

- ✓ 定义：统计报表制度是按一定的表式和要求，自上而下统一布置，自下而上提供统计资料的一种统计调查方法。

思政案例：统计调查

小科普：

我国通过普查进行的统计调查内容和时间周期已经规范化、制度化，具体包括：

人口普查	• 每10年进行一次，逢“0”的年份进行 • 如2010年进行了中国第六次人口普查
农业普查	• 每10年进行一次，逢“7”的年份进行 • 如2017年进行了中国第三次农业普查
经济普查	• 为全面掌握我国第二、第三产业的发展规模、结构和效益，健全基本单位名录等进行的全面调查 • 每5年进行一次，除第一次经济普查(2004年)外，逢“3”和“8”的年份进行。如2018年进行了中国第四次经济普查

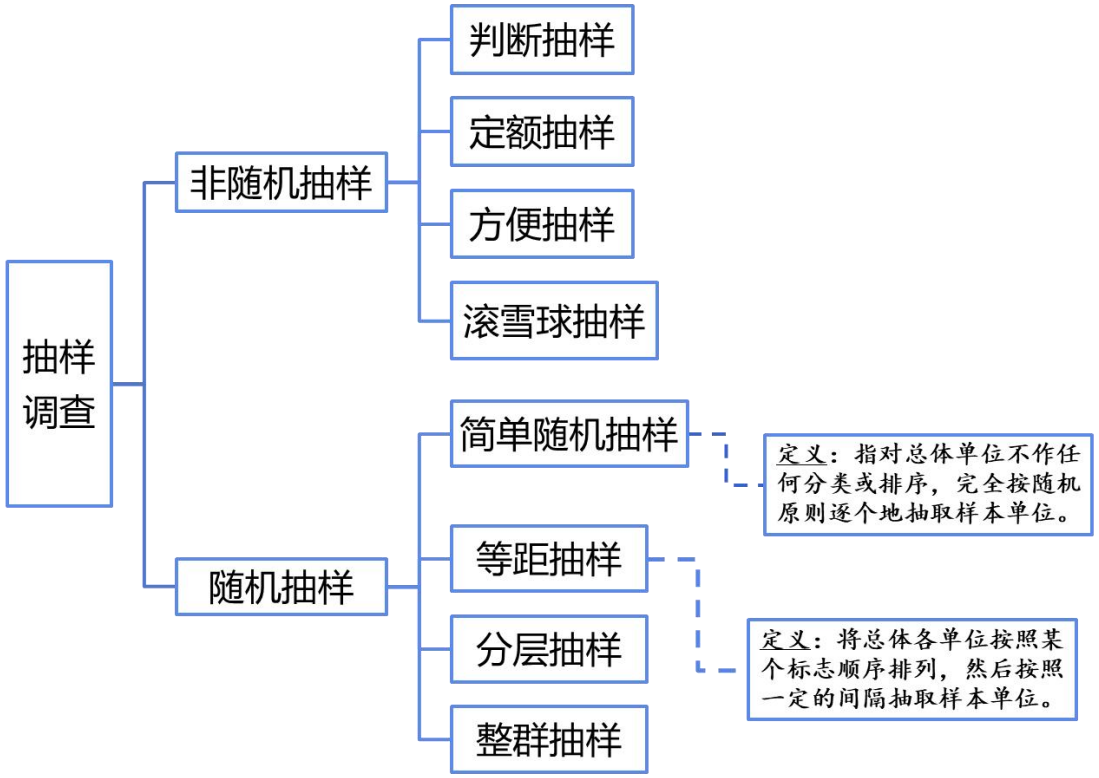
思政案例：统计调查

抽样调查（Sample Survey）

- ✓ 定义：指按一定的程序从总体中抽取样本进行调查。
- ✓ 优点：节省费用、快速、概率抽样可以获得较高的精确性
- ✓ 缺点：可能对总体缺乏代表性（无偏样本vs有偏样本）、对调查问卷的设计和测量精度要求较高

适用场合：

1. 某些不可能进行全面调查的情况。总体规模很大，或连续不断发生的过程。
2. 不必进行全面调查的情况。如民意调查。
3. 对全面调查资料进行验证和修正。



思政案例：统计调查

小科普：

我们身边的抽样调查：上海财经大学千村调查项目、上海市消费者信心指数、中国家庭金融调查专题——中国居民杠杆率和家庭消费信贷问题研究等等。

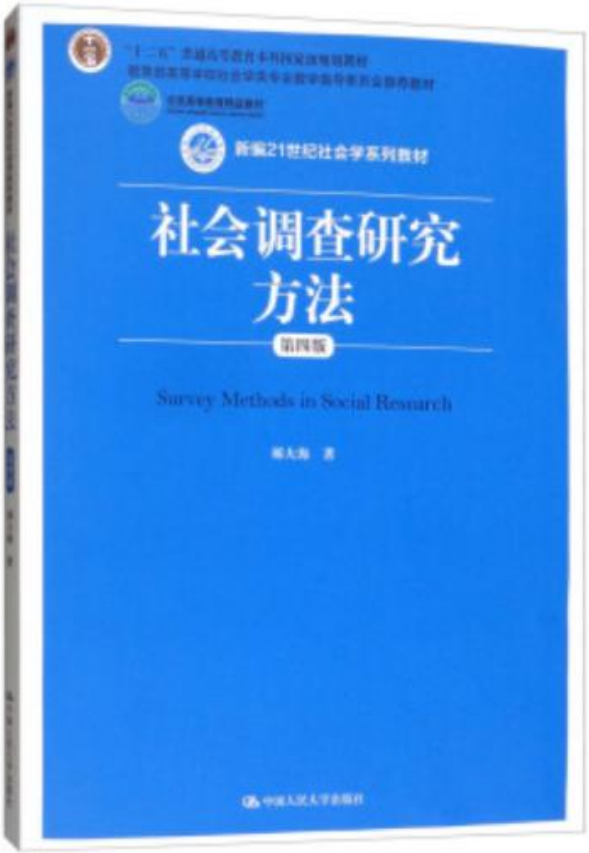
上海财经大学“211工程”三期创新人才培养项目，已成功实施十期。

“走千村，访万户，读中国”。“千村调查”项目是以“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，旨在通过专业的社会调查获得我国“三农”问题的数据资料，形成调查研究报告和决策咨询报告，供国家相关部门决策参考。

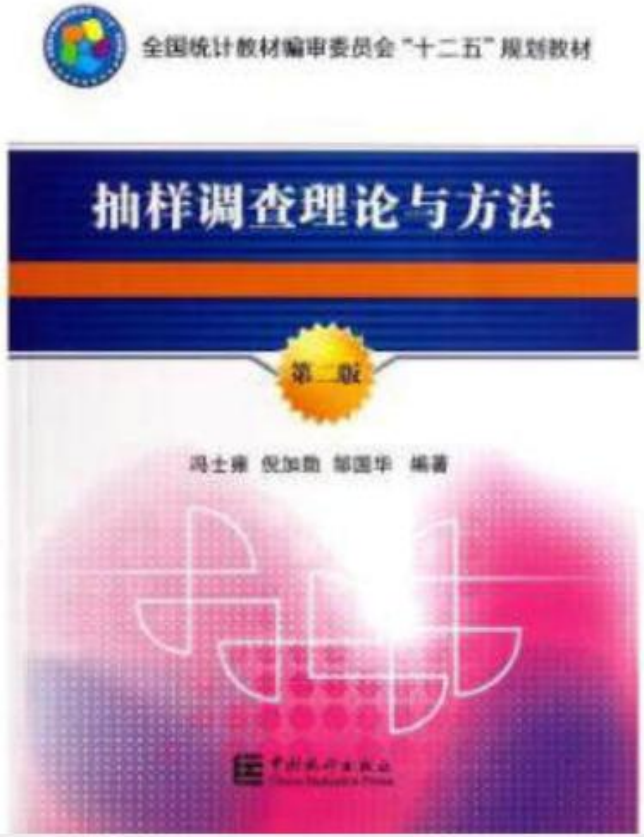
“千村调查”采用随机抽样定点调查和学生返乡调查相结合的方法。调查范围覆盖全国32个省（市、自治区）、近万个农户家庭。

思政案例：统计调查

推荐书目：



《社会调查研究方法》
郝大海
中国人民大学出版社



《抽样调查理论与方法》
冯士雍、倪加勋、邹国华
中国统计出版社

➢ 案例2.7：一个有偏的样本

某超市的货物柜有6层箱子高，最低层箱子摆放的是畅销商品，最高层箱子摆放的是滞销商品，而中间4层箱子摆放的是销售速度一般的商品，采用等距抽样的方法抽样调查商品的销售速度，假设抽样起点为第1个箱子，并且每隔2个箱子抽样一次，于是编号为：1、4、7、10、13、16、19和22号箱子被选中。

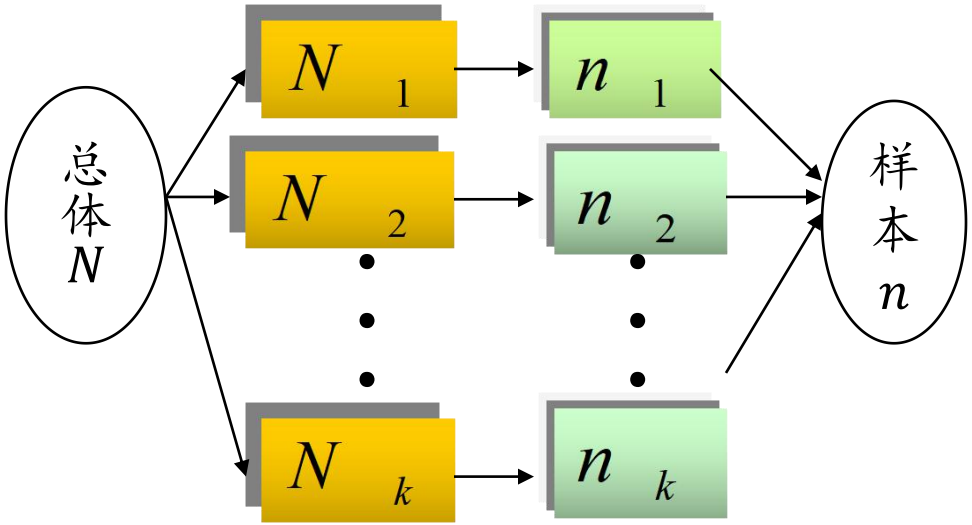
6	7	18	19	→ 滞销商品
5	8	17	20	
4	9	16	21	→ 一般商品
3	10	15	22	
2	11	14	23	
1	12	13	24	→ 畅销商品

很显然，该样本偏向滞销和畅销商品。

（一）统计调查

✓ 分层抽样（类型抽样）

- 定义：将全部总体单位按照某个标志分成若干个类型组，分别在各组中采用**简单随机抽样（或其他形式）**抽取样本单位。
- 分层原则：扩大层间差异，缩小层内差异。



$$n_1 = n_2 = \cdots = n_k$$

等额

$$n = \sum_{i=1}^{k} n_i$$

等比例 $n_i = \dfrac{N_i}{N} \cdot n$

最优 $n_i = \dfrac{N_i \cdot \sigma_i^2}{\sum N_i \cdot \sigma_i^2} \cdot n$

➢ 案例2.8：分层抽样中各层样品数量的确定

利用分层抽样的方法抽取一个样本量为50的样本，分别利用等额分配及等比例分配的方法计算各组所包含的样品数量。

组	资产盈利率	总体单位	等额分配	等比例分配
1	30%及以上	8	10	1
2	20%～30%	35	10	5
3	10%～20%	189	10	27
4	0%～10%	115	10	16
5	负数	5	10	1
	合计	352	50	50

总体分组后，如果某些组内包含的个体数量较少时，等额分配可能不具操作性。

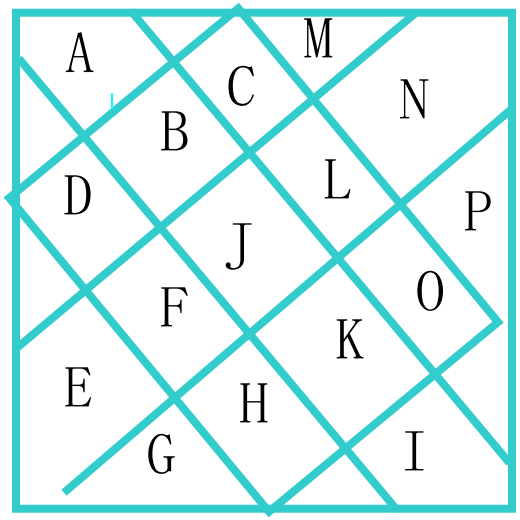
（一）统计调查

上海财经大学 SHANGHAI UNIVERSITY OF FINANCE AND ECONOMICS

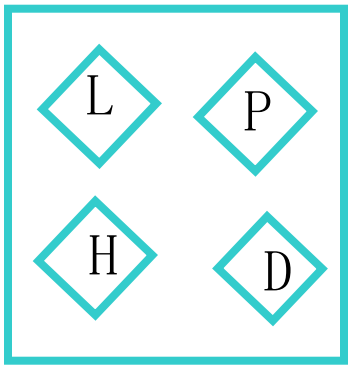
✓ 整群抽样

- 定义：从总体中成群地抽取样本单位，将若干个群组成样本，对抽中的群进行全数登记调查。
- 分组原则：扩大群内差异，缩小群间差异。

总体群数R＝16



样本群数r＝4



样本量

$$n=N_d+N_p+N_l+N_h$$

（一）统计调查

➢ 案例2.9：1948年盖洛普的民意测验

1948年，盖洛普的民意测验预测共和党候选人杜威(T. E. Dewey)将战胜民主党候选人杜鲁门(H. S. Truman)当选总统。1948年11月1日的《华盛顿邮报》报道：“盖洛普预测杜威将获49.5%的选票，杜鲁门将获44.5%。”。

11月2日，大选投票开始。《芝加哥每日论坛报》由于过分相信盖洛普的预测，为抢独家新闻，事先印好标题为“杜威击败杜鲁门”的报纸并发售。最后计票结果是：杜鲁门获24105812张选民票，约占总票数的49.5%，303张选举人票；杜威获21970065张选民票，约占总票数的45.1%，189张选举人票。杜鲁门战胜杜威成为美国总统。《芝加哥每日论坛报》成为流传至今的失实报道之范例。

选举后的第二天，新闻界和盖洛普的民调成了笑柄。杜鲁门手举印有“杜威击败杜鲁门”大幅通栏标题的《芝加哥每日论坛报》返回华盛顿。

（一）统计调查

杜鲁门手举刊登“杜威击败杜鲁门”新闻的《芝加哥每日论坛报》。

（一）统计调查

盖洛普利用定额抽样进行调查。盖洛普公司的访问人员过多地选择了共和党人作为样本，因为一般而言，共和党人比起民主党人富裕一些并受过较好的教育，他们更可能拥有电话，住在较好的地段，总而言之，访问他们比较容易。于是在定额抽样中产生了“共和党偏差”。虽然自1936年、1940年和1944年的三次总统选举预测中，盖洛普的调查同样存在“共和党偏差”，但是由于当时民主党领先优势十分明显而使这种偏差得到抑制，从而盖洛普的预测正确。但是到了1948年，民主党领先的优势变得微弱，这种微弱的优势被定额抽样中的“共和党偏差”压倒，于是盖洛普得出了共和党杜威将战胜民主党杜鲁门的错误预测。

——参考《抽样调查的理论和方法》，施锡铨主编，上海财经大学出版社1996年版。

——参考《“科学民调”隐含悖论：现代民意调查方法1948年遭遇首次“滑铁卢”》，网址链接：http://www.globalview.cn/html/societies/info_2722.html

——参考《盖洛普与民意测验》，《青年文摘》1986年11期上半月，网址链接：http://www.fx361.com/page/1986/1101/3947653.shtml

3. 数据收集方法

（1）访问调查：是调查者与被调查者通过面对面地交谈从而得到所需资料的调查方法。

（2）邮寄调查：是通过宣传媒体或邮寄等方式将调查表或调查问卷送至被调查者手中，由被调查者填写，然后将调查表寄回或投放到指定收集点的一种调查方法。

（3）电话调查：它是调查人员利用电话同受访者进行语言交流，从而获得信息的一种调查方式。电话调查具有时效快、费用低等特点。

（4）网上调查：在网络上开展的调查。

（一）统计调查

上海财经大学
SHANGHAI UNIVERSITY OF FINANCE AND ECONOMICS

问卷星，网址链接：

https://www.wjx.cn/?source=baidu&plan=%E9%97%AE%E5%8D%B7%E6%98%9F&keyword=%E9%97%AE%E5%8D%B7%E6%98%9FBH

（一）统计调查

上海财经大学
SHANGHAI UNIVERSITY OF FINANCE AND ECONOMICS

问卷网，网址链接：

https://www.wenjuan.com/?utm_source=baidu-ss&plan=renqundingxiang&keyword=wenjuanwangguanwang&device=pc&audience=184523&bd_vid=8281903923423296192

（一）统计调查

（5）座谈会：也称为集体访谈法，是将一组被调查者集中在调查现场，让他们对调查的主题（如一种产品、一项服务或其他话题）发表意见，从而获取调查资料的方法。

（6）个别深度访问：是一种一次只有一名受访者参加的特殊的定性研究。“深访”是一种无结构的个人访问，调查人员运用大量的追问技巧，尽可能让受访者自由发挥，表达其想法和感受。

（一）统计调查

➢ 案例2.10：目标总体、抽样框的确定

消费者信心指数已经成为市场发达国家预测经济，为经济运行测温的主要工具之一。

为了及时、准确地反映上海市消费者信心及其变化趋势，从而为政府宏观调控、社会各界和消费者把握上海市消费变化趋向和上海经济发展走势，并对上海市经济运行起到预警作用。上海财经大学自2007年开始编制上海市消费者信心指数。

上海市消费者信心指数的目标总体为上海市20～65岁的上海市常住居民，以及来沪一年以上的外来务工者和居民，在校学生除外，外籍人士除外。

（一）统计调查

上海市消费者信心指数采用分层抽样和定额抽样相结合的组织方式；采用电话调查的数据收集方法。

分层抽样中以上海12个区作为分层依据，按照各区上海市常住人口数量等比例分配各区的调查配额。同时，调查中按照上海市常住人口的年龄实施定额抽样。

电话调查的程序为：首先，与上海电信合作，由上海电信提供上海市固定电话号码中12个区的局号，形成12个区的局号框。在局号框基础上随机生成电话号码得到抽样框。

抽样框为：上海市固定电话号码（局号+随机生成电话号码）。

电话调查虽然操作简单、方便，但是容易形成系统性误差。抽样框遗漏了家中没有安装固定电话的上海市常住居民。

4. 调查方案设计

（1）确定调查目的：即明确调查所要解决的问题。

（2）明确调查对象和调查单位：调查对象是指需要进行调查的社会经济现象的总体。确定调查对象就是要明确目标总体的范围。调查单位是指个体，即需要进行调查的承担者。

（3）确定调查方法。确定调查组织方式和数据收集方法；确定样本量；编制抽样框等。

（4）拟定调查内容：即明确调查登记的具体内容或项目，设计调查问卷。

（5）确定调查时间：确定资料本身的时间（标准时点）和调查工作的起止时间。

（6）编制调查的组织计划。包括：调查的组织机构、参加调查的单位和人员、调查员培训、调查的工作地点、调查文件准备和费用预算等。

（一）统计调查

上海财经大学
SHANGHAI UNIVERSITY OF FINANCE AND ECONOMICS

➢ 案例2.11：第二次全国经济普查方案

第二次全国经济普查方案

中华人民共和国国家统计局

国务院第二次全国经济普查领导小组办公室

说　明

根据国务院《关于开展第二次全国经济普查的通知》（国发〔2007〕35 号文件）和《全国经济普查条例》，国务院第二次全国经济普查领导小组办公室在征求各地区、各部门意见的基础上，制定了《第二次全国经济普查方案》。本方案已经国务院第二次全国经济普查领导小组批准。

《第二次全国经济普查方案》由国务院第二次全国经济普查领导小组办公室负责解释。

2008 年 8 月

（一）统计调查

上海财经大学
SHANGHAI UNIVERSITY OF FINANCE AND ECONOMICS

目　　录

第一部分 总说明

一、普查对象和范围——1
二、普查时间——1
三、普查的主要内容——1
四、普查用标准目录——1
五、普查登记和报送原则——2
六、普查方法——2
七、普查的组织实施原则——2

第二部分 普查表式及主要审核关系

一、普查基层表式——4
二、普查主要指标审核用综合表式——62
三、普查表主要审核关系——100

第三部分 普查填表说明及指标解释

一、单位基本情况表指标解释及填报说明——125
二、财务、能源、信息化和产业活动单位基本情况附表指标解释——136
三、专业普查表指标解释——159

第四部分 普查单位划分的有关规定

一、法人单位、产业活动单位和个体经营户划分规定——214
二、普查单位划分有关问题的处理办法——216

第五部分 普查单位清查办法

一、单位清查的目的——221
二、单位清查的对象和范围——221
三、单位清查的原则和方法——221
四、单位清查的表式和内容——222
五、单位清查的时间——222
六、单位清查的标准——222
七、单位清查的实施步骤——223
附件 1：普查单位清查表式及填报说明——227
附件 2：单位清查底册整理及使用说明——244

第六部分 部门实施普查的办法

一、铁路运输业——247
二、银行及其他金融业——248
三、证券业——250
四、保险业——251
五、军队——253
六、武警——254
第七部分 水及能源消费重点调查方案——255
第八部分 普查工作实施细则

一、普查指导员、普查员的选聘和培训工作细则——257
二、普查区划分工作细则——259
三、普查区划分与绘图电子化操作工作细则——261
四、普查数据质量控制和验收工作细则——263
五、普查表管理工作细则——268
第九部分 普查工作进度表及业务流程图

一、普查工作进度表——282
二、普查业务流程——283
附录 1 国务院关于开展第二次全国经济普查的通知——284
附录 2　全国经济普查条例——286

（一）统计调查

上海财经大学 SHANGHAI UNIVERSITY OF FINANCE AND ECONOMICS

第一部分　总 说 明

根据《国务院关于开展第二次全国经济普查的通知》（国发〔2007〕35号文件）和《全国经济普查条例》，制定第二次全国经济普查方案。

一、普查对象和范围

第二次全国经济普查的普查对象是在我国境内从事第二产业和第三产业的全部法人单位、产业活动单位和个体经营户。

普查具体范围包括：采矿业，制造业，电力、燃气及水的生产和供应业，建筑业，交通运输、仓储和邮政业，信息传输、计算机服务和软件业，批发和零售业，住宿和餐饮业，金融业，房地产业，租赁和商务服务业，科学研究、技术服务和地质勘查业，水利、环境和公共设施管理业，居民服务和其他服务业，教育，卫生、社会保障和社会福利业，文化、体育和娱乐业，以及公共管理与社会组织等国民经济行业。

二、普查时间

普查时点为2008年12月31日24时，普查时期为2008年1月1日—12月31日。

第二条 经济普查的目的，是为了全面掌握我国第二产业、第三产业的发展规模、结构和效益等情况，建立健全基本单位名录库及其数据库系统，为研究制定国民经济和社会发展规划，提高决策和管理水平奠定基础。

（一）统计调查

三、普查的主要内容

第二次全国经济普查的主要内容包括：单位基本属性、财务状况、生产经营情况、生产能力、能源消耗、主要生产设备、信息化和科技活动情况等。其中：各类被调查单位必须填报的共性内容为：单位基本情况、财务状况、水及能源消费情况、信息化情况主要指标。

普查表分为以下三类：

1．普查通用表。包括单位基本情况普查表（包括法人和产业活动单位表）、企业普查表、非企业单位普查表、水及能源消费情况、信息化情况主要指标；

2．专业普查表。包括规模以上工业、资质内建筑业、限额以上批发和零售业、住宿和餐饮业、房地产开发业企业的普查表；

3．部门普查表。包括铁路运输业普查表、银行业及相关金融业（不包括典当业）、证券业、保险业的财务表。

各地区原则上不要扩充普查内容，如确有需要增加指标和内容，不得影响国家普查方案的完整性和准确性，不得变更国家普查指标的名称、解释和编码。

（一）统计调查

四、普查用标准目录

1. 全国组织机构代码编制规则（GB 11714-1997）
2. 统计上使用的行政区划代码结构及编制规则
3. 国民经济行业分类（GB/T4754-2002）
4. 关于划分企业登记注册类型的规定
5. 关于企业登记注册类型对应调整的说明
6. 关于统计上对公有和非公有控股经济的分类办法
7. 单位隶属关系代码（GB/T12404-1997）
8. 统计上单位划分的规定
9. 工业产品生产、销售、库存目录
10. 科技统计分类目录
11. 建筑业企业资质等级编码
12. 房地产开发企业资质管理规定
13. 物业服务企业资质管理办法
14. 零售业态分类标准（GB/T 18106-2004）
15. 批发和零售业商品类值目录
16. 商品交易市场类别目录

（一）统计调查

五、普查登记和报送原则

法人单位在其主要经营活动所在地进行普查登记，但建筑企业在法人单位注册地进行普查登记。各地区普查机构原则上按行政区域组织实施普查。

普查表的基层报送单位为法人单位；法人所属的产业活动单位的普查表由法人单位统一组织填报。

跨地区（省）的产业活动单位采取双重报送原则：跨地区的产业活动单位一方面要向法人单位报送产业活动单位普查表；另一方面，跨地区的产业活动单位还要按其活动所在地普查机构的要求向当地报送产业活动单位普查表，但产业活动单位所在地的普查机构不再将其上报上级普查机构。地方普查机构可以对本地法人在外地的产业活动单位，以及外地法人在本地的产业活动单位资料分别进行汇总。

六、普查方法

对法人单位和产业活动单位采用全面调查的方式；对个体经营户进行全面清查。

调查的工作地点

（一）统计调查

七、普查的组织实施原则

第二次全国经济普查按照“全国统一领导、部门分工协作、地方分级负责、各方共同参与”的原则组织实施。

为了加强对此项工作的组织和领导，国务院设立第二次全国经济普查领导小组，成员单位由各有关部门组成（详见附录1），负责普查的组织实施。

国务院第二次全国经济普查领导小组办公室具体负责普查的日常组织和协调。普查办公室内设的工作小组要分工协作，共同完成宣传协调、普查方案的设计、普查培训和布置、普查单位清查、普查数据的录入、审核和汇总、普查数据处理及资料开发等全部任务。

军队系统、武警系统的第二次经济普查工作由中国人民解放军、中国人民武装警察部队经济普查办公室负责组织完成；铁路运输业的第二次经济普查工作由铁道部经济普查办公室负责组织实施，铁路运输业活动以外，铁路系统从事第二、三产业活动的企业和单位的普查工作，由地方人民政府经济普查机构负责组织实施；银行、证券、保险及其他金融业（不包括典当业）的普查工作，由中国人民银行、中国银行业监督管理委员会、中国证券监督管理委员会、中国保险监督管理委员会与各级人民政府普查机构共同组织完成。国务院其他各有关部门，也要充分发挥各自的职能，各负其责、通力协作、密切配合第二次全国经济普查工作。

地方各级人民政府要设立相应的普查领导小组及其办公室，结合当地实际，组织好本地区的普查工作。

八、本方案由国务院第二次全国经济普查办公室负责解释。

调查的组织机构、参加调查的单位和人员

5. 调查问卷设计

（1）调查问卷组成部分

- 说明词：列于问卷的前面，用来说明调查目的、内容和要求，请求被访者给予合作等。
- 主题问句：即用来搜集资料的一系列问句，是问卷的主体。
- 作业记录：是问卷执行和完成情况的记录，由调查者和问卷审核人员进行填写。

（2）问题设计的原则

- 避免过于笼统的问题

【例】您对××品牌电视机有什么改进意见？

点评：问题欠具体，过于笼统。改为如下的提问：

您对××品牌电视机的外观设计有什么改进意见？

您认为××品牌电视机的质量还可以做哪些改进？

- 避免定义不清的问题

【例】您今年有多大年龄？____岁

点评：由于我国不同地区对年龄有不同的表达，比如虚岁和周岁，容易引起歧义。改为如下提问：

您今年有多大年龄？____周岁

（一）统计调查

- 避免使用多意语字眼

【例】您最近是否看过××报纸？

点评：每个人对“最近”有不同的理解。改为如下的提问：

您昨天是否看过××报纸？

- 避免使用模棱两可的问题

【例】您本周内是否看过《每周广播电视》或报纸？

点评：调查者到底想了解被访者是否看过《每周广播电视》还是报纸？

- 避免出现引导性问题

【例】现在多数电视观众喜欢看××节目？您也喜欢看吗？

点评：这种提问方式会让被访者产生一种自己不喜欢看××节目是不正确的误解。改为如下的提问：

您对××节目的看法是：

（1）喜欢　（2）一般　（3）不喜欢

（二）实验

1. 定义

实验（Experiment）是根据科学研究的目的，尽可能地排除外界的影响，突出主要因素并利用一些专门的仪器设备，而人为地变革、控制或模拟研究对象，使某一些事物（或过程）发生或再现，从而去认识自然现象、自然性质、自然规律。

2. 特点——随机分配（Random Assignment）

实验对象以随机分配的方式被分派到实验组和控制组（或各个不同的实验组）。在大样本情况下，按照随机抽样的原则，各个组的实验对象的构成、条件基本相同，从而**使实验结果凸显出处理（Treatment）的效果**。随机分配无须对实验对象的各种属性进行研究，应用方便，成为实验最常用的方法。但在小样本情况下，随机分配也会出现实验组和控制组实验对象不对称的情况。这时可采用配对和随机化相结合的方法即分块法（blocking），样本先按某关键变量配对，然后随机分配。分块后，尽管比随机化分配的情况要好，但是否要分块，取决于分块的复杂程度即其成本。

（三）观察研究

1. 定义

观察研究（Observational Study）是在自然状态下对研究对象的特征进行观察、记录，并对结果进行描述和对比分析。

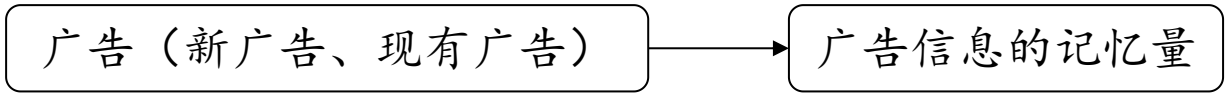
2. 特点

观察研究由于对研究对象不进行控制，因此处理和结果之间如果存在关系，不能把这种关系归结为因果关系。

（三）观察研究

➢ 案例2.12：实验还是观察研究

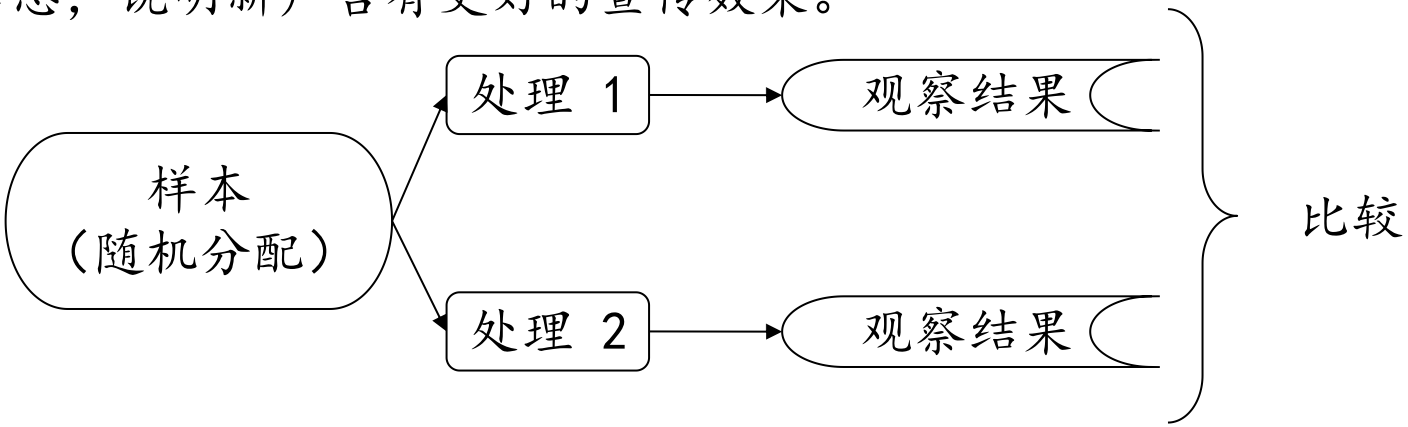
研究1：某产品即将推出一个新的广告，新广告是否比现有广告的效果好是市场部关心的问题。市场研究人员邀请了一批顾客作为测试对象，对每一个测试对象，随机分配其观看新广告还是现有广告，最后比较观看两种广告的测试对象记住了多少广告信息。



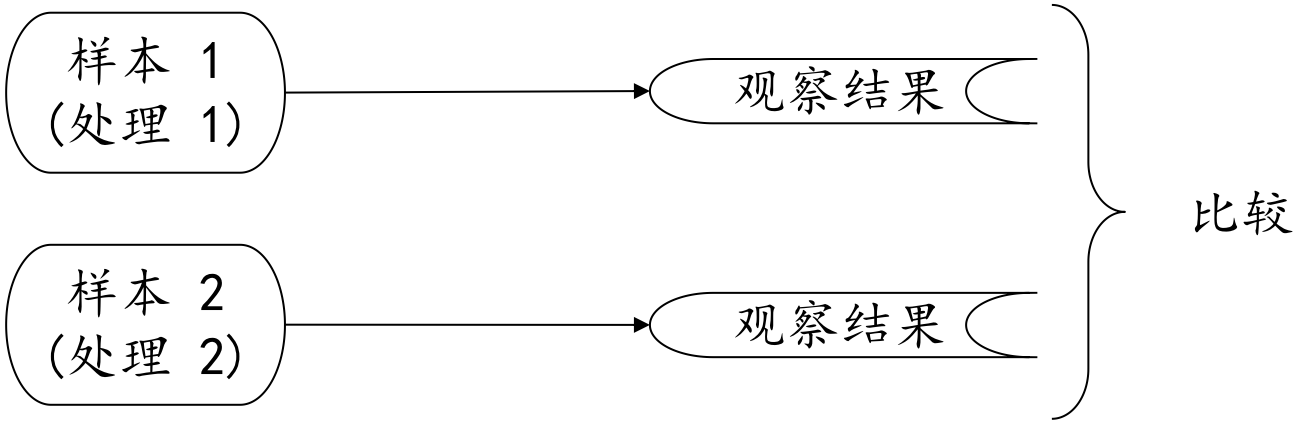
研究2：抽烟对得肺癌是否有影响？研究人员找到一些吸烟者和不吸烟的人，对这两组人做长达5年的观察，记录并比较5年后两组人员患肺癌的比例。

（三）观察研究

研究1采用实验搜集数据。如果观看新广告的对象能够记忆更多的广告信息，说明新广告有更好的宣传效果。



研究2采用观察研究搜集数据。如果吸烟者患肺癌概率更高，说明吸烟与患肺癌之间存在着一定关联关系。但不能认为吸烟导致肺癌。



五、大数据数据来源

（一）大数据

（二）大数据数据来源

（三）大数据分析举例

（一）大数据

1. 定义

- “大数据”（Big data）研究机构Gartner（高德纳是全球最具权威的IT研究与顾问咨询公司）

“大数据”是需要新处理模式才能具有更强的决策力、洞察发现力和流程优化能力的海量、高增长率和多样化的信息资产。

- 麦肯锡报告《Big Data: The Next Frontier for Innovation, Competition, and Productivity》

大数据指的是大小超出常规的数据库工具获取、存储、管理和分析能力的数据集。

（一）大数据

- Wikipedia

Big data is a term used to refer to data sets that are too large or complex for traditional data-processing application software to adequately deal with.

2. 性质

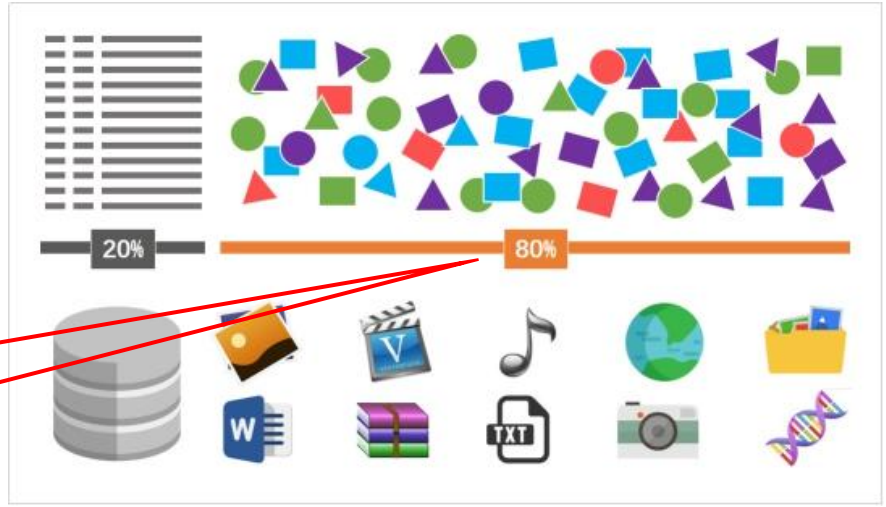
- 国际数据公司（IDC）

✓ Volume 海量的数据规模

✓ Velocity 快速的数据流转和动态的数据体系

✓ Variety 多样的数据类型（大数据可以是结构化数据和非结构化数据，但未来世界将是非结构化的）

✓ Value 巨大的数据价值



非结构化数据占数据总量的80%以上

（二）大数据数据来源

伴随着技术的进步，能够记录数据的技术手段越来越丰富，大数据的来源渠道越来越多，下面列举一些重要的，典型的大数据数据来源。

（二）大数据数据来源

1. 运营商数据

（1）概念：运营商数据就是指由移动通讯运营商所采集的数据。

（2）数据提供者：中国移动、中国联通、中国电信。

（3）数据特点

- 覆盖面广。每个运营商能够覆盖的用户数量非常巨大。据非权威数据统计，2016年，联通电信各自的用户规模都超过2亿，而中国移动的用户规模超过了8亿。
- 精确到个人。从2013年9月开始，我国对移动电话用户实施真实身份信息登记。到2017年底，我国所有通讯设备，基本完成实名登记。因此，相关数据分析结果以及相关决策，都可以精确到一个自然人身上。
- 内容丰富。运营商数据内容非常丰富。首先，因为实名登记制度，可以了解个人的身份证号码、年龄、籍贯等信息。另外，每个通话记录、通话位置、使用的APP及观看内容均可获悉。

➢ 案例2.13：运营商与银行的合作

近年来，中国移动注重大数据的对外变现。其中应用在征信领域代表性的案例是东莞移动与东莞某银行的合作。

东莞移动构建的基于电信大数据的个人征信评价体系由客观变量、评价指标及信用评分三部分组成。

客观变量包含ARPU（Average Revenue Per User，每用户平均收入）、流量消费、在网时长等用户业务数据，以及用户的性别、年龄、身份证号码或者实名制信息等用户基础数据共24个变量。

基于客观变量，将24个变量划分为身份特征、通信消费能力、信用历史、行为偏好以及人脉关系五类评价指标。

最终评分参考了美国个人信用评级法FICO评分，通过评分模型计算输出用户评级，分数越高代表用户的信用越好。

（二）大数据数据来源

上海财经大学
SHANGHAI UNIVERSITY OF FINANCE AND ECONOMICS

银行在发行信用卡或审核贷款时通常面临申请人资料是否属实和给予申请人信用卡贷款额度多少的问题。

为解决这两个问题，东莞某银行与东莞移动就大数据评分在信用卡授信方面联合建模，提出了应用方案：问题一可通过运营商实名制结果予以验证。问题二可通过综合信用评分予以解决。

该合作从2015年9月份开始正式运营，共受理银行信用卡授信请求1709笔，通过银行验证1075笔，应用效果良好。

——参考：梁杨、朱宏文、赵大海，
《大数据技术及其在电信行业的应用研究》，
《移动通信》，2017年第5期。

（二）大数据数据来源

2. 支付交易数据

（1）概念：是指普通消费者通过支付或者交易而产生的流水数据。

（2）支付数据提供者

- 支付结算系统（银联）：每天海量的刷卡交易数据，包括：户名、卡号、交易时间、商户、消费金额等。
- 互联网支付通道（支付宝、微信、拉卡拉等）：海量的支付流水数据。
- 银行：记录各家银行银行卡的交易行为，无法覆盖所有人群，覆盖到的人群也无法精确了解用户的所有消费行为。

（3）交易数据提供者

- 线上线下零售平台：以淘宝、天猫、京东等为代表，形成海量的交易数据，数据比较详细。既包括了交易本身；还包括了交易的场景，比如，时间地点、动态页面、实时广告促销等众多信息；以及对交易的评价信息。

（二）大数据数据来源

➢ 案例2.14：淘宝十年账单

2014年12月8日，支付宝发布了“淘宝十年账单”。自2004年支付宝成立，十年里，全国人民网络支出总计423亿笔。朋友圈里立即掀起一场“血雨腥风”的晒单，“败家”小伙伴纷纷表示，手指已经不够剁了。

其实，从2012年开始，支付宝就针对消费者发布年度对账单。年度对账单的出现，意味着支付宝已经具备描述单个消费者特征的能力，拥有消费者的消费偏好以及区域、联系方式等属性信息，更意味着巨大的商业价值。

——参考：百度百科“淘宝十年账单”

——参考：季鸿、张秀凤、柴林麟，《大数据在电信行业的应用展望》，《通信企业管理》，2014年第1期。

淘宝网
Taobao.com

（二）大数据数据来源

上海财经大学
SHANGHAI UNIVERSITY OF FINANCE AND ECONOMICS

（二）大数据数据来源

3. 手机数据

（1）概念：是指一大类来自于手机的数据。大概可以分为三种类型：

- 来自运营商的数据；×
- 来自手机制造商的数据；√
- 来自APP开发商的数据。√

（二）大数据数据来源

（2）来自手机制造商的数据

绝大部分的手机根据操作系统分为两大类：苹果和安卓。

其中，苹果公司控制了来自苹果手机的大量数据。这些数据包括但不局限于：身份信息、支付信息、地理位置、通话记录等。

如果手机是安卓系统，相应的数据被不同的手机制造商掌握。有媒体报道，华为作为一个重要的手机制造商，通过其荣耀Magic智能手机，手机用户活动信息，并通过对这些信息的分析做个性化推荐。收集的信息包括但不限于：短信内容、微信的聊天信息等。

手机制造商数据的**劣势**：与运营商相比，覆盖面相对较小。

手机制造商数据的**优势**：数据更加丰富：一是，覆盖不同的运营商；二是，可能覆盖非通讯类内容（例如不同APP的使用情况）。

（二）大数据数据来源

（3）来自APP开发商的数据

据媒体报道，微信的月活跃用户总量已经达到9亿左右，甚至超过了中国移动用户数量。这些用户，横跨不同的运营商、不同的手机制造商、不同地域（甚至海外），因此覆盖面极其完整。

当然，更多的APP装机量并不多，但是数据也非常有价值。

（4）数据特点：手机数据没有隐私

- 位置是暴露的
- 购物是暴露的
- 社交是暴露的

（二）大数据数据来源

➢ 案例2.15：高德大数据揭车主形象

2017年1月，高德地图联合交通运输部科学研究院发布了《2016年度中国主要城市交通分析报告》。报告显示了2016年全年国内交通拥堵城市排名以及对2016年全国驾驶行为的分析。

与此同时，高德地图此次发布的交通报告，还针对不同品牌车主的驾驶行为进行了分析归纳。其中，不同汽车品牌的车主在驾驶行为上也有不同表现。

——参考：《高德大数据揭车主形象：凯迪拉克和MINI“躺枪”》，网址链接：https://www.sohu.com/a/124121509_121861

（二）大数据数据来源

2016年度“野蛮”驾驶品牌排行榜，多数上榜品牌为中高档汽车。

MINI车主以80.68分排名榜首，成为“最野蛮驾驶员”，这或与MINI女车主众多有关。

如路虎，Jeep等SUV品牌也位列野蛮驾驶者行列。

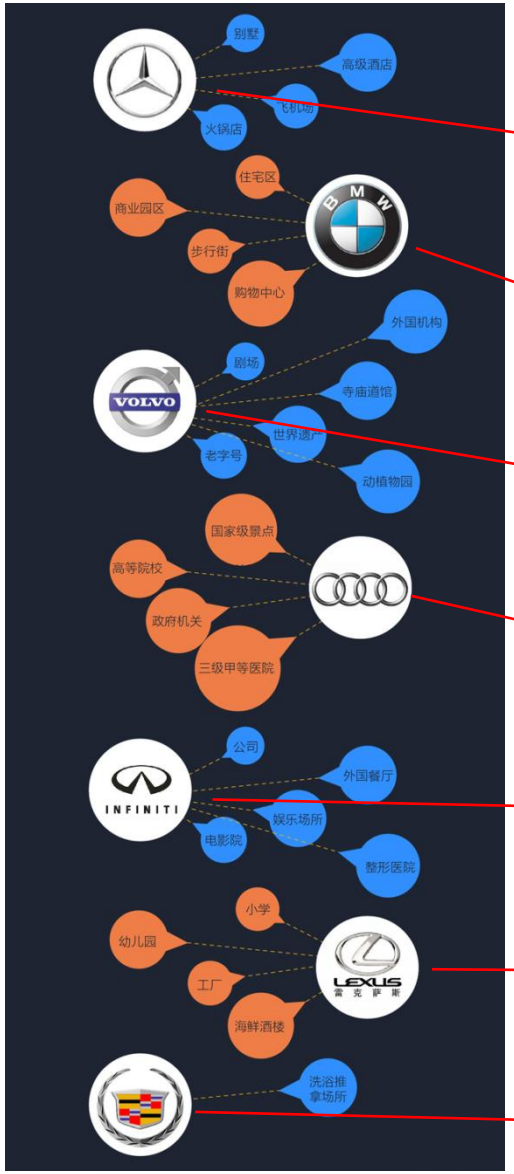
如凯迪拉克、宝马、保时捷等注重车型性能的品牌同样榜上有名。

（二）大数据数据来源

与野蛮相反的，2016年开车最文明的品牌全部被自主品牌包揽。

北汽幻速以88.7分的成绩位列第一。

（二）大数据数据来源



对部分高档汽车品牌的出行地分析，结果如下：

奔驰车主出行频率较高的目的地主要是别墅、飞机场、高级酒店、火锅店等高消费、商务场所。

宝马车主则更多去往步行街、购物中心、商业园区等，偏向于消费。

沃尔沃车主则出入寺庙、老字号、世界遗产等文化景点次数较多。

奥迪车主的目的地更多集中于高等院校、国家级景点、政府机关等。

英菲尼迪车主则频繁出入整容医院、电影院、娱乐场所等年轻化场所。

雷克萨斯车主则频繁出入幼儿园、小学、工厂、海鲜酒楼等场所。

凯迪拉克车主出行频率较高的独占洗浴推拿场所。

（二）大数据数据来源

4. 社交网络数据

（1）概念：是指借助于各种通讯网络、邮件网络、社交网站（例如Facebook、Twitter、微博、微信）等所记录的由社会个体集合及个体之间的连接关系构成的社会性结构数据。

（2）构成

- 描述社交网络的拓扑结构。一个社交网络的用户，看作是一个节点，而关注关系，看作是一条边。因此，社交网络的拓扑结构可以通过一系列的点和边描述出来。这些边可能是带有方向性的（例如：微博的关注关系，称为有向网络），也可能是对称的（例如：微信的好友关系，称为无向网络）。
- 附着在网络拓扑结构上的数据。有些数据是关于点的属性，例如：一个用户的性别、年龄、教育程度等。有些数据涉及到两个不同的节点，是关于边的属性，例如：一个用户给另一个用户点了多少赞，一个用户转发了另一个用户多少文章等。

（二）大数据数据来源

➢ 案例2.16：社交网络数据的应用

用于企业招聘

企业招聘过程中，可以通过利用社交网络数据更加深入地了解他们的招聘对象。招聘软件开发商Jobvite的一组数据显示，在使用Facebook的企业中，有三分之二的企业希望使用该网站的好友查询功能完成招聘工作。在使用Twitter的企业中，则有54%的企业希望能用该服务了解应聘者潜在的观点和兴趣。

社交网络上的许多社交数据反映出了应聘者的兴趣爱好、技能专长、价值取向、以及文笔风格、创意风格等。这些数据可以直接从应聘者的主页上收集到，也可以通过应聘者在人人网、新浪微博或者QQ空间中探讨的内容、转发的内容中间接获得。因此，通过大量的社交数据分析，HR就能够清楚应聘者适合哪个岗位，谁对产品开发感兴趣，谁对市场营销有见地。

——参考：《社交网络大数据应用》，网址链接：http://www.dataguru.cn/thread-478022-1-1.html

（二）大数据数据来源

用于社会化推荐

社会化推荐之所以受到很多网站的重视，主要因为下面的优点：

好友推荐可以增加推荐的信任度。美国著名的第三方调查机构尼尔森调查了影响用户相信某个推荐的因素。调查结果显示，90%的用户相信朋友对他们的推荐，70%的用户相信网上其他用户对广告商品的评论。可以看出，好友的推荐对于增加用户对推荐结果的信任度非常重要。用户往往不一定信任计算机的智能推荐，但会信任好朋友的推荐。

社交网络可以解决冷启动问题。当一个新用户通过社交网站登录时，可以从社交网站中获取用户的好友列表，然后给用户推荐好友在网站上喜欢的物品。从而在没有用户行为记录时就给用户提供较高质量的推荐结果，部分解决了推荐系统的冷启动问题。

——参考：《推荐系统浅谈系列（六） - 社交网络数据》，网址链接：https://www.jianshu.com/p/d8cf25dde131

（二）大数据数据来源

社交网络数据征信

2015年8月，国外社交巨头Facebook推出了涉足社交大数据征信的专利——当一个用户申请贷款的时候，贷款方会审查该用户社交网络好友的信用等级。只有这些好友的平均信用等级达到了最低的信用分要求，贷款方才会继续处理贷款申请。否则的话，该申请将被拒绝。

在此之前，阿里巴巴旗下蚂蚁金服推出的芝麻信用分也在使用人脉关系、消费行为作为评估信用水平的依据。

腾讯也进行了互联网征信建设的探索，该公司主要依靠大数据与人工智能技术，基于旗下微信、QQ等近十亿用户的社交数据来进行征信工作，通过把结构化数据，文本分类，基于位置服务（LBS）数据，社交网络传播扩散等挖掘之后形成用户画像刻画。

质疑：（1）社交网络数据和个人信用表现关联性不强，其在大数据征信中的作用有限。（2）社交网络数据的真实性有待提升。朋友圈、微博、空间的状态与评论互动大部分实质可归结为感性的“秀炫晒”。

——参考：《社交网络数据征信的作用你猜有多大？》，网址链接：

https://www.leiphone.com/news/201609/HbVtLMIut1jOAUK9.html

（二）大数据数据来源

5. 浏览日志的数据

（1）概念：指互联网的浏览日志。

（2）商业价值：

- 网络日志关乎网络安全。许多网站肩负着十分重要的公共任务。例如：重要的政府机构（例如：外交部）、重要的事业单位（例如：大学网站）、重要商业机构（例如：银行、运营商等）。通过在体量巨大的网络日志中挖掘如下信息：怎样的访问？来自什么位置？IP地址如何？什么浏览器？提出什么请求？浏览什么页面？做出什么动作？等，从而识别异常行为，探测攻击行为，学习攻击规律，寻找系统安全漏洞。
- 网络日志关乎消费者画像。以大学网站的浏览为例，一个用户来到某大学的网站后，做了什么搜索，看了什么内容，浏览时间长短，都记录在网络日志中。通过这些信息，可以推断：用户是一个学者还是MBA学员？用户对哪些教授、哪些课程、哪些研究感兴趣？从而帮助大学进一步完善其课程体系。

（三）大数据分析举例

非结构化数据需要转化为结构化数据后再进行相应的分析。

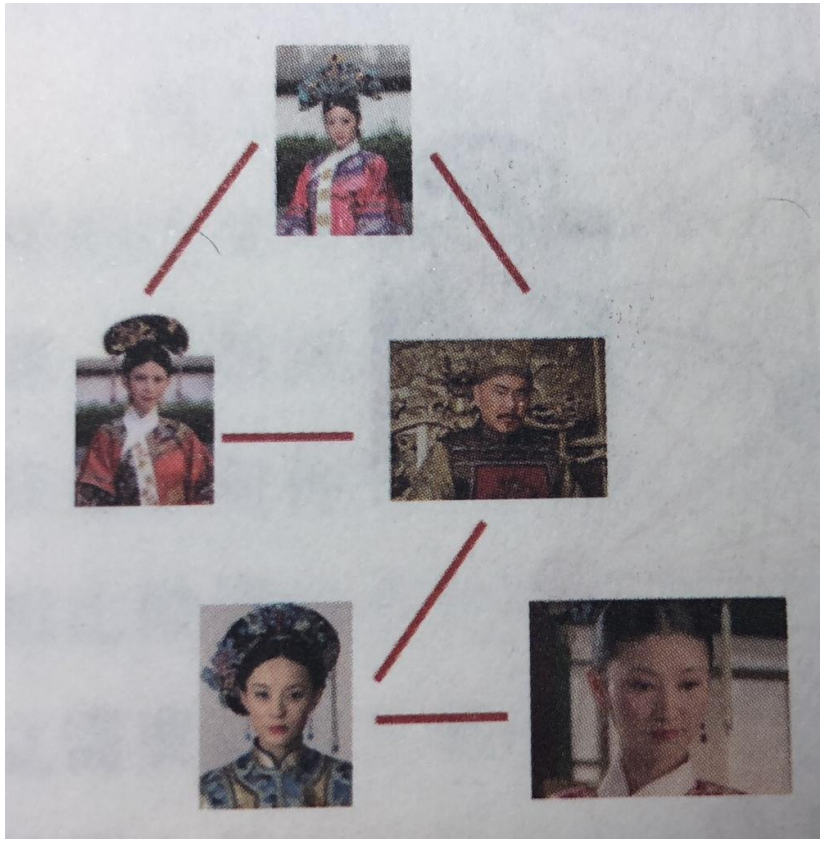
（三）大数据分析举例

➢ 案例2.17：社交网络数据的转换

——《甄嬛传》中的爱恨情仇

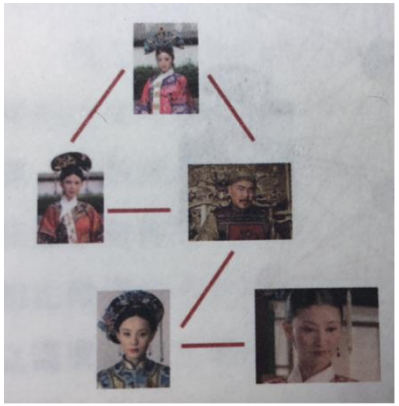
社交网络数据通常转换为邻接矩阵（adjacency matrix）。

以《甄嬛传》为例，在最开始的时候，皇帝、皇后、华妃、甄嬛和沈眉庄之间的关系可以用右图表示。此时，皇后、华妃还没有和甄嬛结怨，故不存在关系。注意这是一个无向网络。



（三）大数据分析举例

将该社交网络数据转换为邻接矩阵，如下表所示：



	1 皇帝	2 甄嬛	3 皇后	4 华妃	5 沈眉庄
1 皇帝	0	1	1	1	0
2 甄嬛	1	0	0	0	1
3 皇后	1	0	0	1	0
4 华妃	1	0	1	0	0
5 沈眉庄	0	1	0	0	0

（三）大数据分析举例

网络密度：网络中实际存在的边与可能存在的边的比例。

	1 皇帝	2 甄嬛	3 皇后	4 华妃	5 沈眉庄
1 皇帝	0	1	1	1	0
2 甄嬛	1	0	0	0	1
3 皇后	1	0	0	1	0
4 华妃	1	0	1	0	0
5 沈眉庄	0	1	0	0	0

网络密度$=\frac{10}{5\times4}=50\%$，这是一个很稠密的网络。

在现实网络数据中，由于社交网络的稀疏性，网络密度往往非常低。

稀疏性：网络数据中任何两个个体（节点）产生一条边的概率几乎为0，即$P(a_{ij}=1)\to0$，个体i与个体j产生关系的概率趋于0。

（三）大数据分析举例

出度：指由某一个节点向外发出的边的个数。

入度：是指向某一个节点的所有边的个数。

	1 皇帝	2 甄嬛	3 皇后	4 华妃	5 沈眉庄
1 皇帝	0	1	1	1	0
2 甄嬛	1	0	0	0	1
3 皇后	1	0	0	1	0
4 华妃	1	0	1	0	0
5 沈眉庄	0	1	0	0	0

出度＝3

入度＝1

皇帝的出度为3，说明他和甄嬛、皇后和华妃有联系。

沈眉庄的入度为1，说明只有甄嬛和她有联系。

出度和入度数值越大，越说明是一个“社交达人”。

（三）大数据分析举例

传递性：传递性能够帮助社交网络中的节点建立更多的关系。

	1 皇帝	2 甄嬛	3 皇后	4 华妃	5 沈眉庄
1 皇帝	0	1	1	1	0
2 甄嬛	1	0	0	0	1
3 皇后	1	1	0	1	0
4 华妃	1	0	1	0	0
5 沈眉庄	0	1	0	0	0

（三）大数据分析举例

互粉性：一旦给定个体i到个体j的关系，那么会大大增加从个体j到个体i建立关系的概率。

	1 皇帝	2 甄嬛	3 皇后	4 华妃	5 沈眉庄
1 皇帝	0	1	1	1	0
2 甄嬛	1	0	1	0	1
3 皇后	1	1	0	1	0
4 华妃	1	0	1	0	0
5 沈眉庄	0	1	0	0	0

——参考：《数据思维：从数据分析到商业价值》，王汉生编著，中国人民大学出版社

（三）大数据分析举例

➢ 案例2.18：文本的转换——ACME对网络上电话产品评论的分析

文本分析是指通过对文本数据进行表示、处理和建模来获得有用的见解。文本挖掘是文本分析的一个重要组成部分，是在大量的文本集合中发现关系和有趣模式的过程。

文本分析的数据源是典型的非结构化数据。

考虑一个生产电话和电子书阅读器的公司ACME，该公司生产的电话为bPhone。ACME与生产和销售类似产品的其他公司之间存在着激烈的竞争关系，为了在竞争中获得成功，ACME希望提升产品质量进而增加销量。

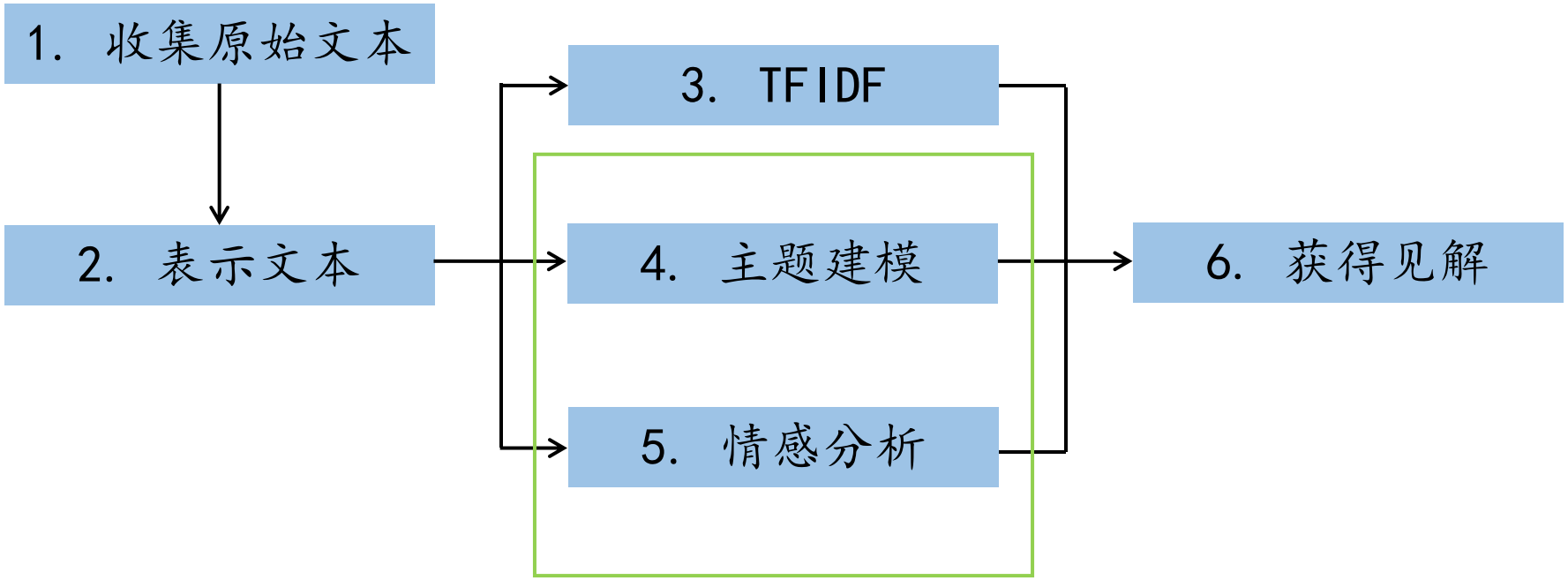
ACME公司一直注重收集网络上有关该公司产品的评论。希望通过对这些评论进行分析，从而回答如下问题：

（1）人们有没有提到ACME的产品？

（2）当提到ACME的产品时，都说了什么？对产品的评价是正面的还是负面的？如果是负面评价，原因是什么？例如，抱怨bPhone的电池续航能力不强等。

（三）大数据分析举例

ACME的文本分析流程表述如下：



（三）大数据分析举例

1. 收集原始文本

ACME的数据科学团队积极监控网站上用户的评价内容，包括：新闻门户和博客上的相关文章；在线商店或评论网页上关于ACME产品的评论；社交媒体上包含bPhone关键词的帖子等。

许多新闻门户和博客站点都能提供具有开放标准格式的数据源。

如果计划从在线商店和评论站点上收集用户评论，但是这些站点不提供应用程序编程接口（Application Programming Interface，API）或数据源，团队就需要编写网页爬虫来解析网页，然后从这些HTML文件中自动提取感兴趣的数据。

许多网站为第三方开发人员提供了公开的API来访问网站上的数据。比如，Twitter API允许开发人员获取Twitter上包含关键字bPhone或者bEbook的公开tweet（用户发到Twitter上的信息）。开发人员也可以实时读取特定用户的tweet或特定地点附件的tweet。

如果不打算自己收集数据，许多公司还可以提供数据收集服务。

（三）大数据分析举例

2. 表示文本

当收集到原始文本后，需要使用文本规范化技术（分词、大小写转换等）对原始文本进行转换，并用更加结构化的方法进行表示，供后续分析使用。

分词：是从文本正文中分离单词。分词后，原始文本转换为一组标记（token）的集合，每个标记通常是一个词。通常的方法是使用空格和标点符号进行文本分词（英文分词），但是在某些特定情况下基于标点符号分词并不合适（例如We’ll标记为We和ll；can’t标记为can和t等），因此，通常会将标准分词技术和查找表匹配使用。

大小写转换：将所有的字母都变成小写（或者都变成大写），如果大小写转换应用不正确，会造成歧义。例如，世界健康组织WHO变成了who。可以对不进行大小写转换的单词建立一个查找表。

（三）大数据分析举例

文本在通过分词和大小写转换规范化以后，需要以更结构化的方式来表示。

词袋法将文本当作一个无序的词的集合，以文本中出现的所有词及其出现的次数体现文档的特征。也就是说，词袋法统计文本中每个单词的词频（Term Frequency，TF）。通常文本中出现的单词很多，因此，文本分析的一个挑战是高维度。

更高级的方法还需要考虑词序、上下文、推论等因素。

（三）大数据分析举例

3. TFIDF

词频（TF）是仅仅考虑各个词在文本自身出现的次数。以广受欢迎的儿童书籍Green Eggs and Ham为例，虽然书中包含了804个单词，但是大量单词重复出现，只用了50个不同的单词，需要统计每个单词的词频。

下面通过一个例子简要介绍词频的计算。

考虑一个有10个单词的词袋向量空间：i、love、acme、my、bebook、bphone、fantastic、slow、terrible、terrific。如果从网站上下载了一条评论I love LOVE my bPhone. 在转换大小写与分词之后的相应词频向量为下页左表。

由于文本通常具有高维度特性，而高维度使得存储和解析难度增加，因此需要降低维度。通常移除**停止词**，即对理解文本没有多大帮助的词，例如冠词等。或者**只存储文本中至少出现过一次的词的词频**。

经过降维处理后的相应词频向量为下页右表。

（三）大数据分析举例

词语	频数
i	1
love	2
acme	0
my	1
bebook	0
bphone	1
fantastic	0
slow	0
terrible	0
terrific	0

词语	频数
i	1
love	2
my	1
bphone	1

（三）大数据分析举例

有时仅仅考虑词频是不够的。用语料库进一步解决无用词语的问题。这是基于如下的思考：词的重要性随着该词在语料库中出现频率的增加而减少。比如有个关于动物的各种文章的语料库，包含“动物”这个词的文章是很多的，那么“动物”这个词对于区分语料库中的文章有帮助吗？答案是显而易见的，并没有帮助。也就是说一个词在语料库的文章中出现的次数越少，作用才会越大。

IDF（Inverse Document Frequency，IDF）反映了语料库的文章中出现该词的文章频率，具体计算规则如下：

$$\text{文件频率（DF）}=\frac{\text{包含该词的文章数}}{\text{语料库中所有文章数}}\ ;$$

$$\text{逆向文件频率（IDF）}=\frac{\text{语料库中所有文章数}}{\text{包含该词的文章数}}$$

注意，有时TF和IDF都会对数或分母+1处理。

由此可知，某个词的IDF越大，说明该词在语料库各文章的普及率越小，作用越大。

（三）大数据分析举例

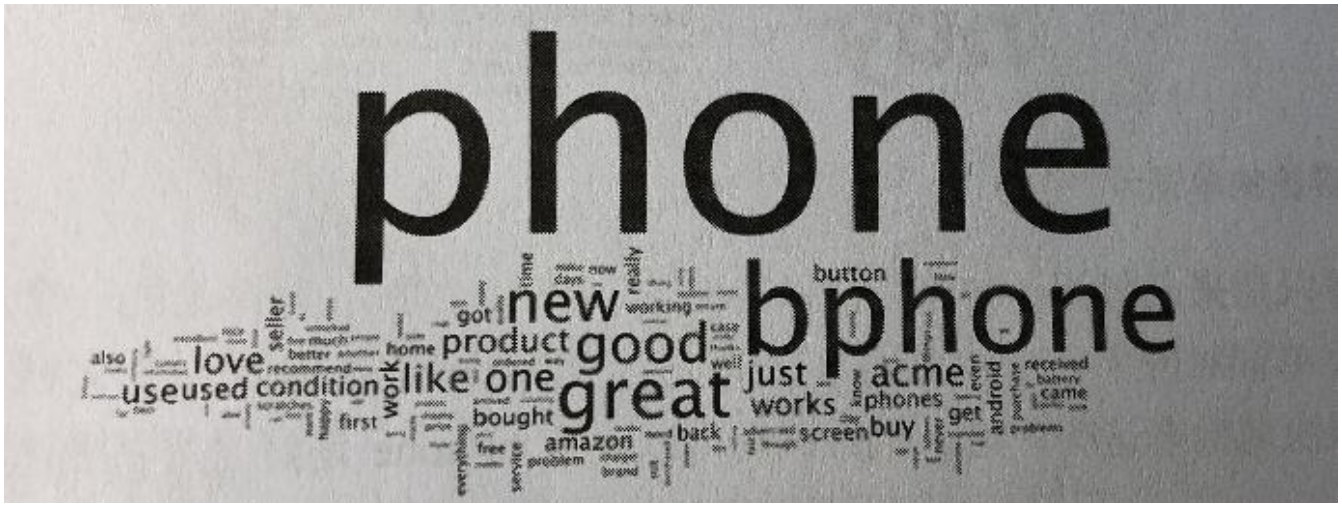
而TFIDF则是将词频和文件频率相结合的方法，即

TFIDF=TF×IDF

一个词的TFIDF数值越高，说明它在该文本中出现越频繁，但在语料库的所有文章中出现得越少。

6. 获得见解

ACME的数据科学团队用bPhone为关键词从网站上收集了300多条评论。在经过一系列前期数据处理滞后，将300条评论转化为词云图。

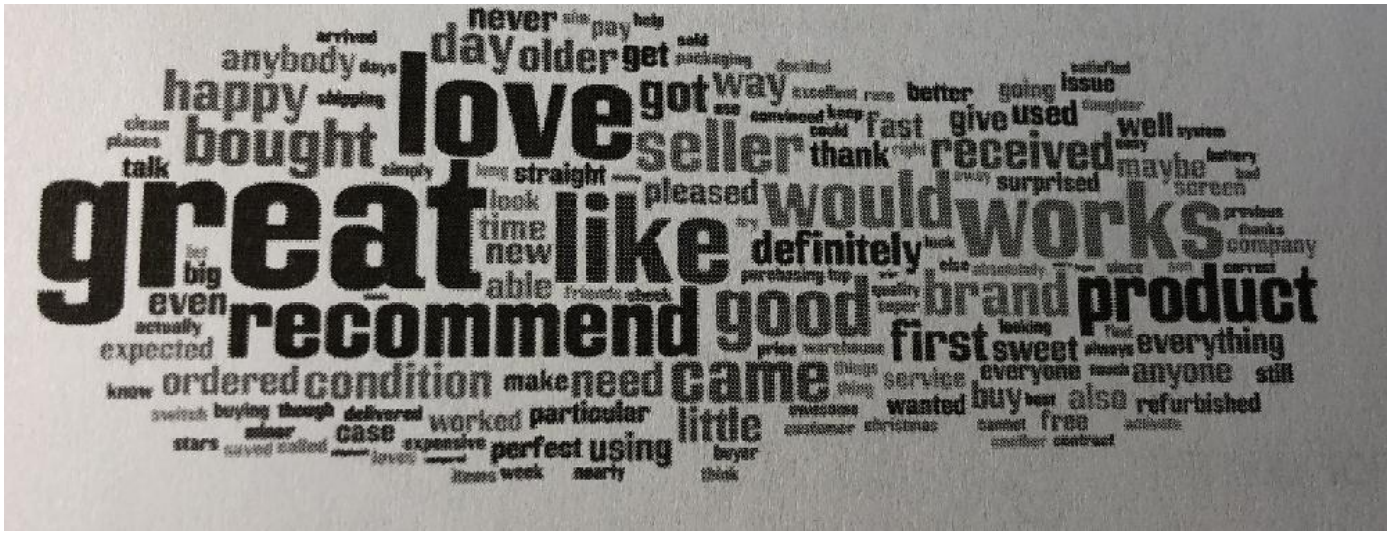


（三）大数据分析举例

由于图的大部分由phone和bPhone组成，这两个词频繁出现但并不具有信息量，因此团队可以删除诸如phone、bPhone和ACME这样的词，这类词称为特定领域的停止词。

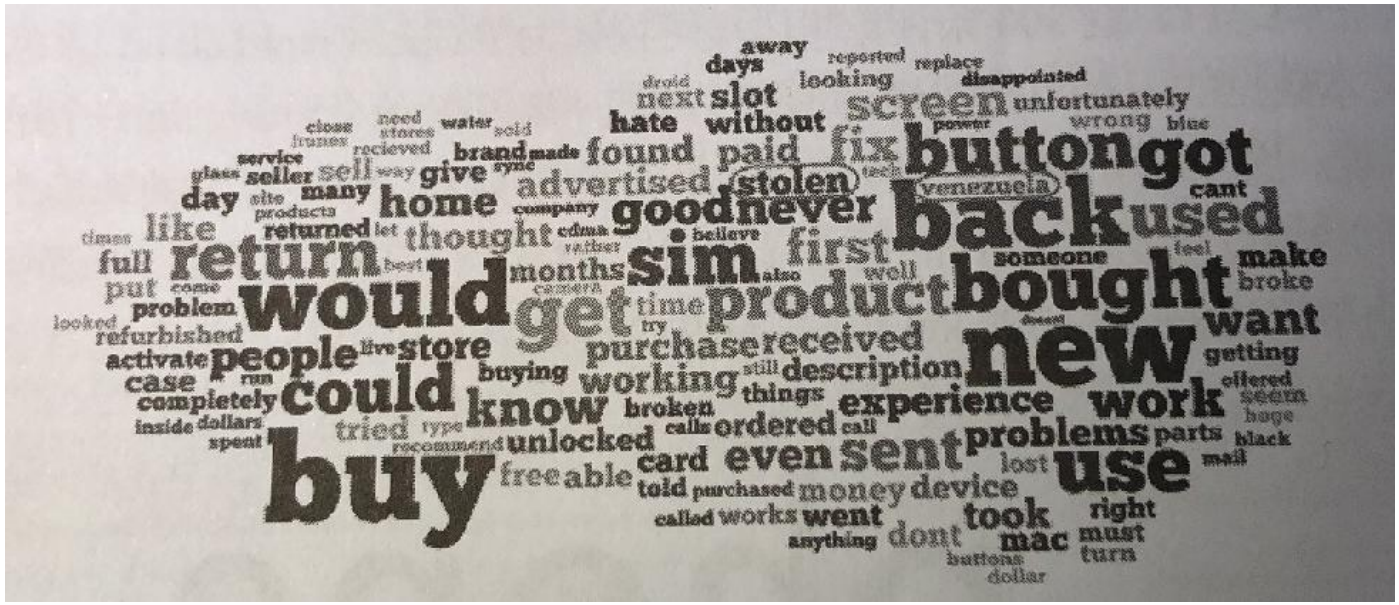
删除特定停止词后，按照用户1～5的打分，分别绘制5星评论的词云图和1星评论的词云图。

5星评论词云图如下所示。这说明用户对seller、brand和product满意，他们会向朋友或家人推荐（recommend）bPhone手机。



（三）大数据分析举例

1星评论词云图如下所示。其中，sim和button频繁出现，建议将包含这些词语的评论作为研究对象，进一步确定用户对sim卡和按键的看法。另外，词云图中还包含了stolen、Venezuela等单词，进一步研究表明，位于委内瑞拉的一些未经授权的销售商会出售偷来的bPhone。ACME公司需要针对这一点采取进一步的行动。



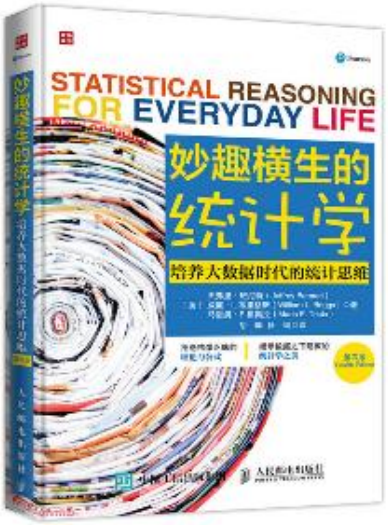
（三）大数据分析举例

——参考：《数据科学与大数据分析：数据的发现、分析、可视化与表示》，EMC Education Services著，曹於、刘文苗、李枫林译，人民邮电出版社

——参考：《机器学习笔记笔记之三——文本类型处理-词袋法、TF-IDF理解》，网址链接：https://blog.csdn.net/qq_35946969/article/details/84562104

六、推荐阅读

上海财经大学
SHANGHAI UNIVERSITY OF FINANCE AND ECONOMICS



谢 谢!

Thank You