第 2 讲 大数据预处理

海量数据计算研究中心

杨东华

大数据分析

- **什么是数据分析 ：**用适当的统计分析方法及工具对处理过的数据进行分析，将它们加以汇总、理解并消化，以求最大化地开发数据的功能，发挥数据的作用，形成有效结论的过程。
- **数据分析的目的 ：**把隐藏在一大批看似杂乱无章的数据背后的信息集中和提炼出来，总结出研究对象的内在规律。

大数据分析的流程

1 大数据的采集和存储

2 大数据预处理

3 大数据分析建模和大数据分析方法

4 大数据分析结果展示

1. 大数据的采集与存储

常用的大数据获取途径

（1）系统日志采集

可以使用海量数据采集工具，用于系统日志采集。

- Hadoop的Chukwa、Cloudera的Flume、Facebook的Scribe等;
- 这些工具均采用分布式架构，能满足大量的日志数据采集和传输需求。

1. 大数据的采集与存储

常用的大数据获取途径

（2）互联网数据采集

通过网络爬虫或网站公开API等方式从网站上获取数据信息。

- 该方法可以从网页中把数据抽取出来，将其存储为统一的本地数据文件，它支持图片、音频、视频等文件或附件的采集，附件与正文可以自动关联。
- 除了网站中包含的内容之外，还可以使用DPI(Deep Packet Inspection, 深度包检测) 或DFI (Deep Flow Inspection, 深度流检测)等带宽管理技术实现对网络流量的采集。

1. 大数据的采集与存储

常用的大数据获取途径

（3）APP移动端数据采集

APP是获取用户移动端数据的一种有效方法。

- APP中的SDK插件可以把用户使用APP的信息汇总给指定服务器，即便用户在没有访问时，也能够获知用户终端的相关信息，包括安装应用的数量和类型等。
- 单个APP用户规模有限，数据量有限；但数十万APP用户，获取的用户终端数据和部分行为数据会达到数亿的量级。

1. 大数据的采集与存储

常用的大数据获取途径

（4）与数据服务机构进行合作

- 数据服务机构通常具备规范的数据共享和交易渠道，用户可以在平台上快速、明确地获取自己所需要的数据。

- 对于企业生产经营数据或科学研究数据等保密性要求较高的数据，可以通过与企业或研究机构合作，使用特定系统接口等相关方式采集数据。

1. 大数据的采集与存储

小结

- 多个数据库接收发自客户端（Web、App或者传感器形式等）的数据，并且以不同的形式将这些数据存储在数据库中。用户可以通过这些数据库来进行简单的查询和处理工作。
- 但是，这些数据可以是结构化数据，也可以是非结构化数据，而且这些数据易受到噪声数据、数据值缺失、数据冲突等影响，因此需首先对收集到的大数据集合进行预处理，以保证大数据分析预测结果的准确性与价值性。

2. 大数据的预处理

数据在采集和存储之后，往往还需要进行必要的加工整理后才能真正用于分析建模。

- 数据归约是在不损害分析结果准确性的前提下降低数据集规模，使之简化。包括维归约、数据抽样等技术，这一过程有利于提高大数据的价值密度。

- 数据清理技术包括对数据的不一致检测、噪声数据的识别、数据过滤与修正等方面，有利于提高大数据的一致性、准确性。

3. 大数据分析模型与分析方法

按照一定的方法建立大数据分析模型，并且用合适的统计分析方法对收集来的规模巨大的量数据进行分析，提取有用信息和形成结论。

- 大数据分析模型用于描述输入和输出之间的关系，所讨论的问题是“从大数据中发现什么？”
- 大数据分析模型的建立是最基础也是最重要的步骤，我们经常听说的贝叶斯分类器、聚类、决策树都是大数据分析模型。

4. 大数据分析结果展示及评估

将分析所得的数据进行可视化处理，使用户能更方便地获取数据，更快更简单地理解数据。

- 将分析所得的数据进行可视化处理，以计算机图形或图像的直观方式显示给用户的，并可与用户进行交互式处理。
- 可以提高大数据分析结果的直观性，使得用户能更方便地获取数据，更快更简单地理解数据，有利于发现大量业务数据中隐含的规律性信息，以支持管理决策。

大数据分析的流程

1 大数据的采集和存储

※ 2 大数据预处理

※ 3 大数据分析建模和大数据分析方法

4 大数据分析结果展示

预处理的重要性

引言

对大数据进行加工处理:

(1) 通过缩减数据规模，将形式不同、内容不同的数据整理为形式和语义一致的数据;

(2) 对数据进行有效清理等，可以有效支持大数据分析，增加数据分析的有效性和准确性。

数据预处理的三个重要步骤

1. 数据抽样和过滤

通过缩减数据规模，挑选出对于分析有用的数据。这一过程有利于提高大数据的价值密度。

2. 数据标准化和归一化

通过基本描述统计量的计算、数据取值的转换、数据的正态化处理等，将形式不同、内容不同的数据整理为形式和语义一致的数据。

数据预处理的三个重要步骤

3. 数据清洗

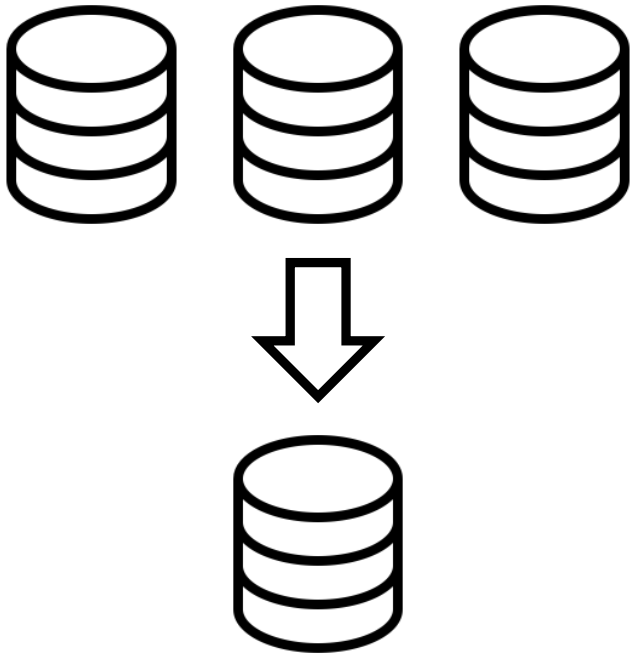
发现并修复数据中的错误，从而最小化数据中的错误对大数据分析结果的负面影响。

- ✓ 通过修复数据中的错误、数据缺失值处理等数据清洗操作为进一步深入的分析和建模建立基础。
- ✓ 数据清理技术包括对数据的不一致检测、噪声数据的识别、数据过滤与修正等方面，有利于提高大数据的一致性、准确性。

1. 数据抽样和过滤

目的

减少要处理的数据量，使当前的处理能力能够处理这些数据。



1.1 数据抽样

抽样的基本思想：

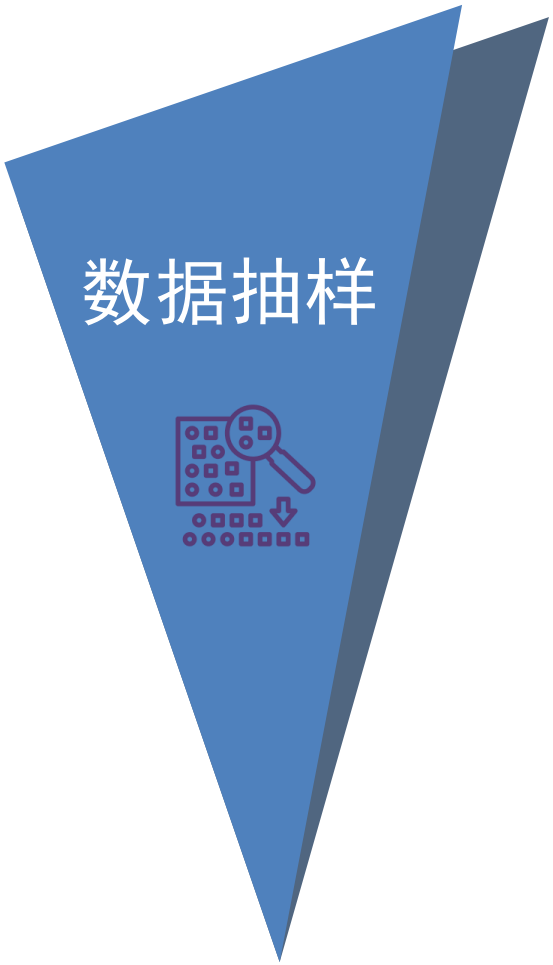
用样本估计总体，即通常不直接去研究总体，而是通过从总体中抽取一个样本，根据样本的情况去估计总体的相应情况。

要达到这个目的，就要求抽取的样本具有能够代表总体质量特征的性质。

我们就要采用合适、合理的，

能够使样本具有代表性特征的抽样方法来抽取子样。

1.1 数据抽样



1. 随机抽样
2. 系统抽样
3. 分层抽样
4. 整群抽样

1.1 数据抽样 — 随机抽样

简单随机抽样的定义

- 一般地，设一个总体含有N个个体，从中逐个不放回的抽取n个个体作为样本，如果总体内的每个个体被抽到的机会相等，均为n/N，就称这种抽样方法为简单随机抽样。
- 常常用于总体个数N较少的情况。

1.1 数据抽样 — 随机抽样

简单随机抽样的特点

(1) 样本的个数是有限的;

(2) 它是从总体中逐个进行抽取;

(3) 它是一种不放回抽样;

(4) 它是一种等可能的抽样。

1.1 数据抽样 — 随机抽样

简单随机抽样的常用方法

（1）抽签法(其过程简记为：编号、制签、搅匀、抽签、取个体)

把总体中的N个个体编号，把号码写在号签上，将号签放在一个容器中，搅拌均匀后，每次从中抽取一个号签，连续抽取n次，就得到一个容量为n的样本。

➢抽签法的优点是简单易行，缺点是当总体的容量非常大时，费时、费力，又不方便。如果标号的签搅拌不均匀，会导致抽样不公平。

1.1 数据抽样 — 随机抽样

简单随机抽样的常用方法

（2）随机数法(其过程简记为：编号、选数、读数、取个体)

即利用随机数表、随机数骰子或计算机产生的随机数进行抽样。

- 随机数法的优点是简单易行，缺点是当总体容量较大时，仍然不是很方便，但是比抽签法公平。

简单随机抽样总结

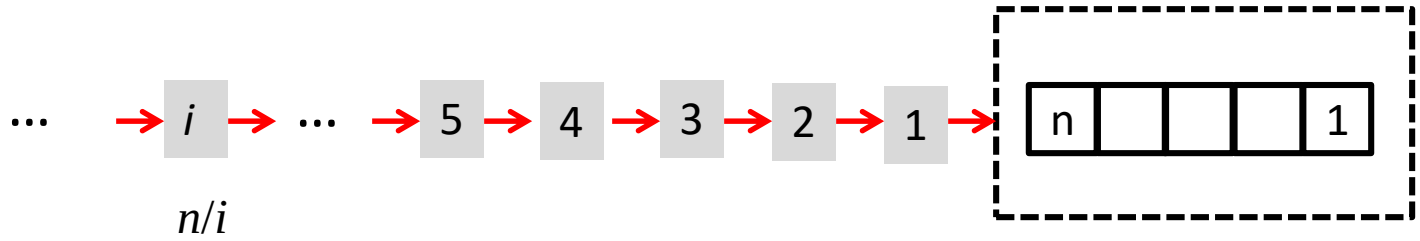
- ✓ 简单随机抽样事先要把研究对象编号，比较费时、费力；
- ✓ 当样本容量较小时，可能发生偏向，影响样本的代表性。
- ✓ 它只适用于总体单位数量有限的情况，否则编号工作繁重；
- ✓ 对于复杂的总体，样本的代表性难以保证；不能利用总体的已知信息等。
- ✓ 适用情景：在总体单位有限，或抽样的个体情况不明，或总体单位之间特性差异程度小时采用此法效果较好。

适合总体容量较少的抽样

1.1 数据抽样 — 随机抽样

简单随机抽样的例子

数据流介绍



- 在数据流处理中的一个常见问题就是数据采样问题。
- 希望从流中选择一个子集，以便能够对它进行查询并给出统计性上对整个流具有代表性的结果。

简单随机抽样的例子

- 具体问题就是要从数据流中随机抽取n个元素。如果数据流长度N事先已经知道，那这个问题就非常简单，每个元素以n/N的概率选取即可。

- 但是该问题中N是未知的。数据流的前n个元素依次加入到大小为n的窗口中。对于数据流第i个元素（$i>n$），以n/i的概率替换窗口中的某个元素。最终窗口的元素出现概率均为n/N。

总体要求：从N个元素中随机的抽取n个元素，其中N无法确定，保证每个元素抽到的概率相同。

1.1 数据抽样 — 随机抽样

数据流算法:

是指数据源源不断地到来，根据到来的数据返回相应的部分结果。适用于两种情况:

- 数据量非常大仅能扫描一次时，可以把数据看成数据流，把扫描看成数据到来。
- 数据更新非常快，不能把所有数据都保存下来再计算结果，此时可以把数据看成是一个数据流。

空间亚线性算法:

由于大数据算法中涉及到的数据是海量的，数据难以放入内存计算，所以一种常用的处理办法是不对全部数据进行计算，而只向内存里放入小部分数据，仅使用内存中的小部分数据，就可以得到一个有质量保证的结果。

1.1 数据抽样 — 随机抽样

水库抽样

一般应用在数据流的情况下，是一个典型的空间亚线性算法，问题可以描述为：

输入：一组数据，但大小未知

输出：这组数据的n个均匀抽样。

三点要求：

（1）仅允许扫描数据一次。

（2）扫描到数据的前i个数据时（$i>n$），保存当前已扫描数据的n个均匀抽样。

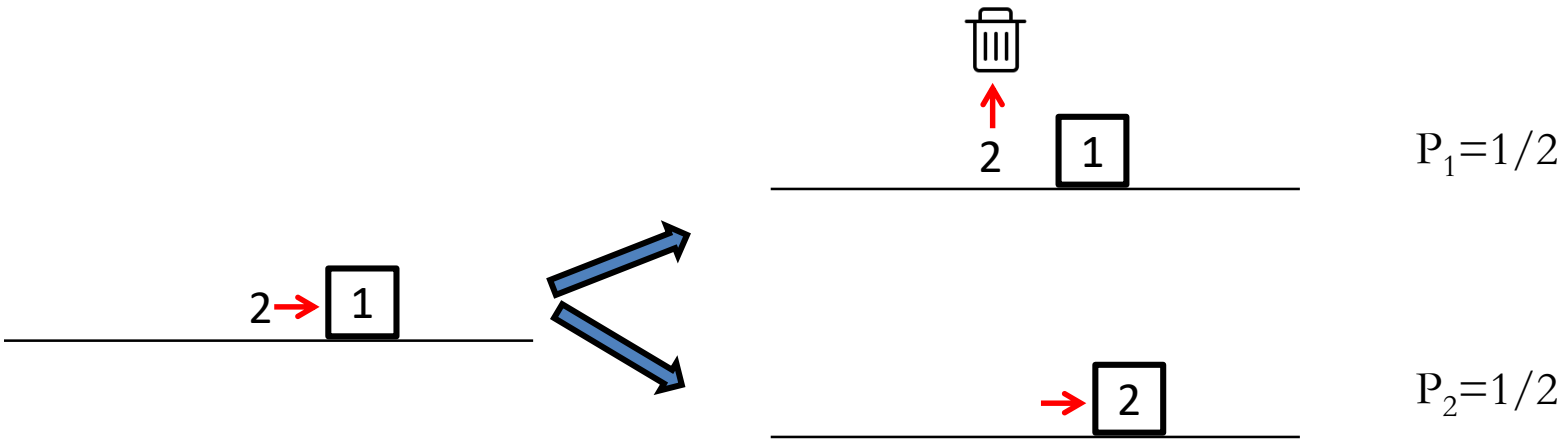
（3）空间亚线性算法，空间复杂度为$O(n)$。空间复杂度和抽样大小有关，而与整个数据的数据量无关。

1.1 数据抽样 — 随机抽样

水库抽样示例

$P_1=1$

1.当数据1到达时，我们将它保存下来。



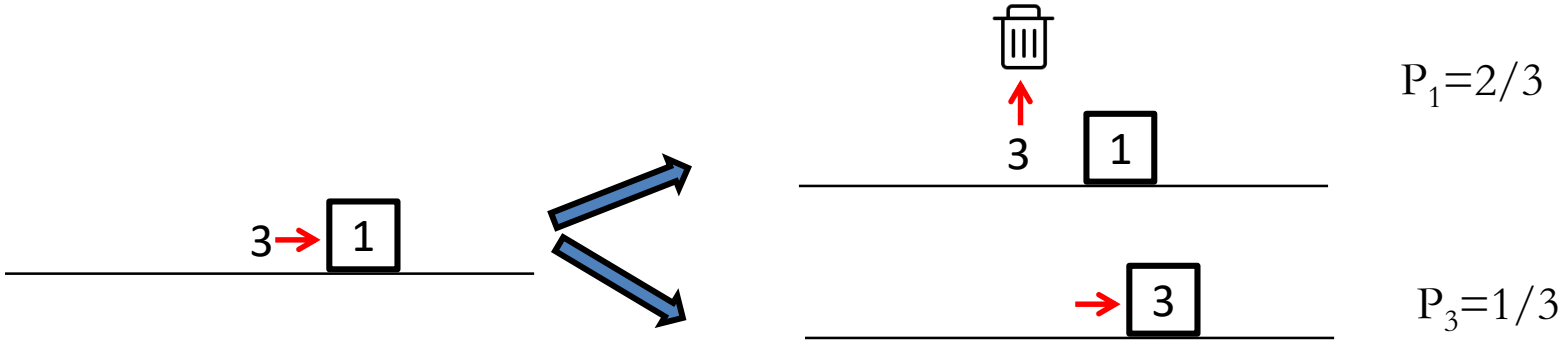
$P_1=1/2$

$P_2=1/2$

2. 当数据2到达时，以1/2的概率，舍弃数据2，保留数据1；以1/2的概率，舍弃数据1，保存数据2;

1.1 数据抽样 — 随机抽样

水库抽样示例



3. 当数据3到达时，以2/3的概率，舍弃数据3，保留数据1；以1/3的概率，舍弃原数据1，保存数据3。

数据1、数据2和数据3被留下的概率都为1/3。

数据1：$1\times(1/2)\times(2/3)=1/3$

数据2：$(1/2)\times(2/3)=1/3$

数据3：$(1/3)=1/3$

1.1 数据抽样 — 随机抽样

定理：水库抽样得到的采样是均匀的，在任何时候接收到大于n的i个数时，选出的这n个数一定都是它的一个均匀采样。

- 在接收第i+1个数时，第i个数还能保存在数组当中的概率是

$$(1-\frac{1}{i+1})。$$

因为在接收到第i+1个数时要以$\frac{n}{i+1}$的概率随机替换，而第i个数被选中的概率是$\frac{1}{n}$，它们相乘为$\frac{1}{i+1}$。$\frac{1}{i+1}$就是第i个数被换出数组的概率，所以$1-\frac{1}{i+1}$就是在接收第i+1个元素时第i个数在数组当中的概率。

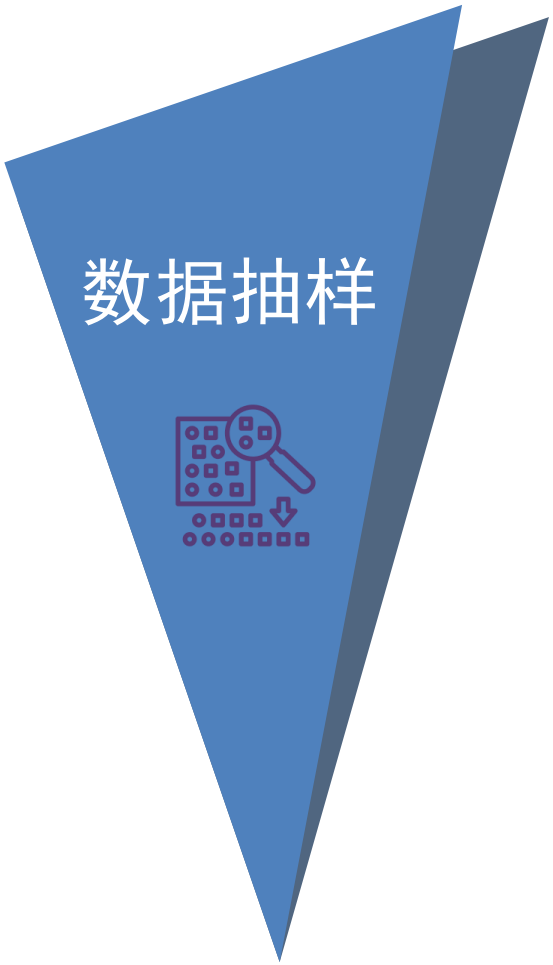
1.1 数据抽样 — 随机抽样

- 同理，在接收第i+2个数时，第i个数仍然保留在数组当中的概率是$1-\frac{1}{i+2}$。依此类推，当接收第N个数时，第i个元素保存在数组当中的概率是1-$\frac{1}{N}$。
- 如果这些事件都发生了，那么在接收第N个数时，第i个数字才能保留在数组当中。因此它保留在抽样当中的概率是发生这些事件的概率的积，就是。

$$\frac{n}{i}\times(1-\frac{1}{i+1})\times(1-\frac{1}{i+2})\times\ldots..\times(1-\frac{1}{N})=n/N$$

可以得到，对于任意一个元素 i，其被选入样本的概率均为n/N，证明其符合随机抽样。

1.1 数据抽样

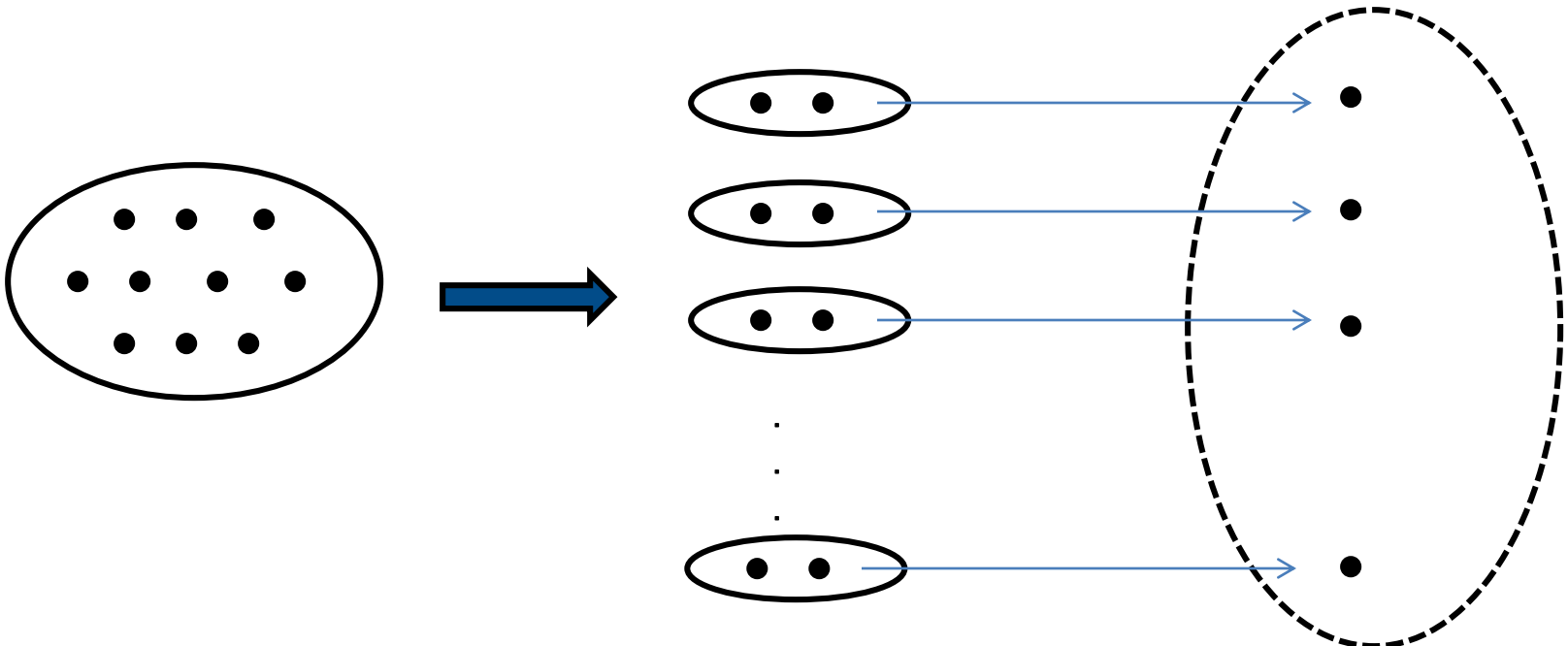


1. 随机抽样
2. 系统抽样
3. 分层抽样
4. 整群抽样

1.1 数据抽样 — 系统抽样

系统抽样

- 系统抽样也称为等距抽样、机械抽样。
- 将总体分成均衡的几个部分，然后按照预先定出的规则，从每一部分抽取一个个体，得到所需要的样本。



1.1 数据抽样 — 系统抽样

系统抽样的具体实现

1. 在系统抽样中，先将总体从1～N相继编号，并计算抽样距离$K=N/n$。式中N为总体单位总数，n为样本容量。
2. 然后在1～K中抽一随机数 k_1，作为样本的第一个单位;
3. 接着取k_1+K, k_1+2K, ……，直至抽够n个单位为止。

➢ 使用场景：主要针对按一定关系排好的数据。

系统抽样的示例

从503名学生中抽取50名学生作为样本。

1. 编号：随机剔除三名学生，对剩余的学生编号，依次为1, 2, 3, … ,500。

2. 确定抽样距离：$K=N/n=500/50=10$。（把500名学生分成50段，每段10名学生）

3. 确定随机数：从1~10中随机选取一个数字，如5。

4. 抽样：那么被抽取到的学生的编号分别为5, 15, 25, … ,495。

1.1 数据抽样 — 系统抽样

系统抽样的特点

- 用系统抽样抽取样本时，每个个体被抽到的可能性是相等的，个体被抽取的概率等于n/N。
- 系统抽样适用于总体中个体数较多，抽取样本容量也较大时;
- 系统抽样是不放回抽样。

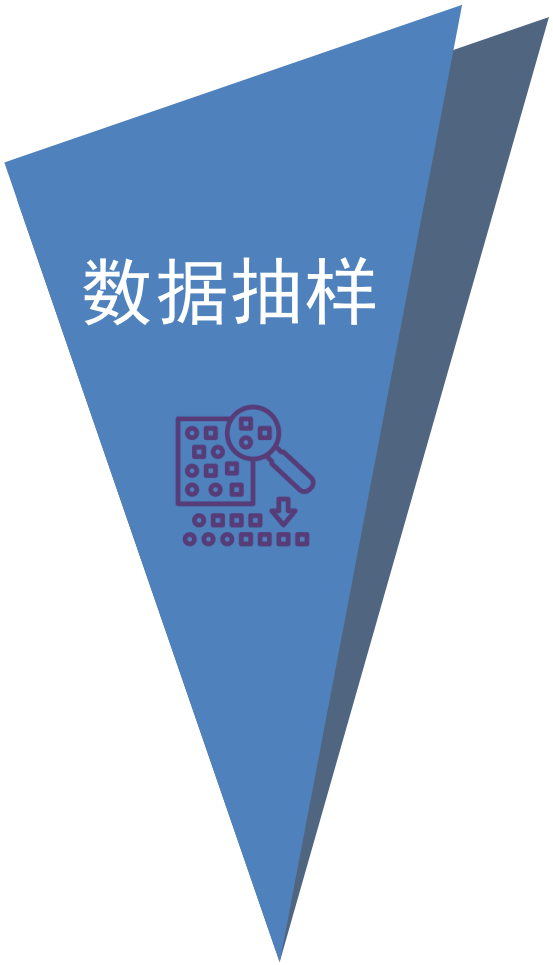
1.1 数据抽样

系统抽样与简单随机抽样的对比

- 系统抽样比简单随机抽样更容易实施，可节约抽样成本;
- 系统抽样的效果会受个体编号的影响，而简单随机抽样的效果不受个体编号的影响;
- 系统抽样所得样本的代表性和具体的编号有关，而简单随机抽样所得样本的代表性与个体的编号无关。
- 在系统抽样中，编号的个体特征随编号的变化呈现一定的周期性，可能会使系统抽样的代表性很差。

  例如，学号按照男生单号女生双的方法编排，那么，用系统抽样的方法抽取的样本就可全部男生或全部女生。

1.1 数据抽样

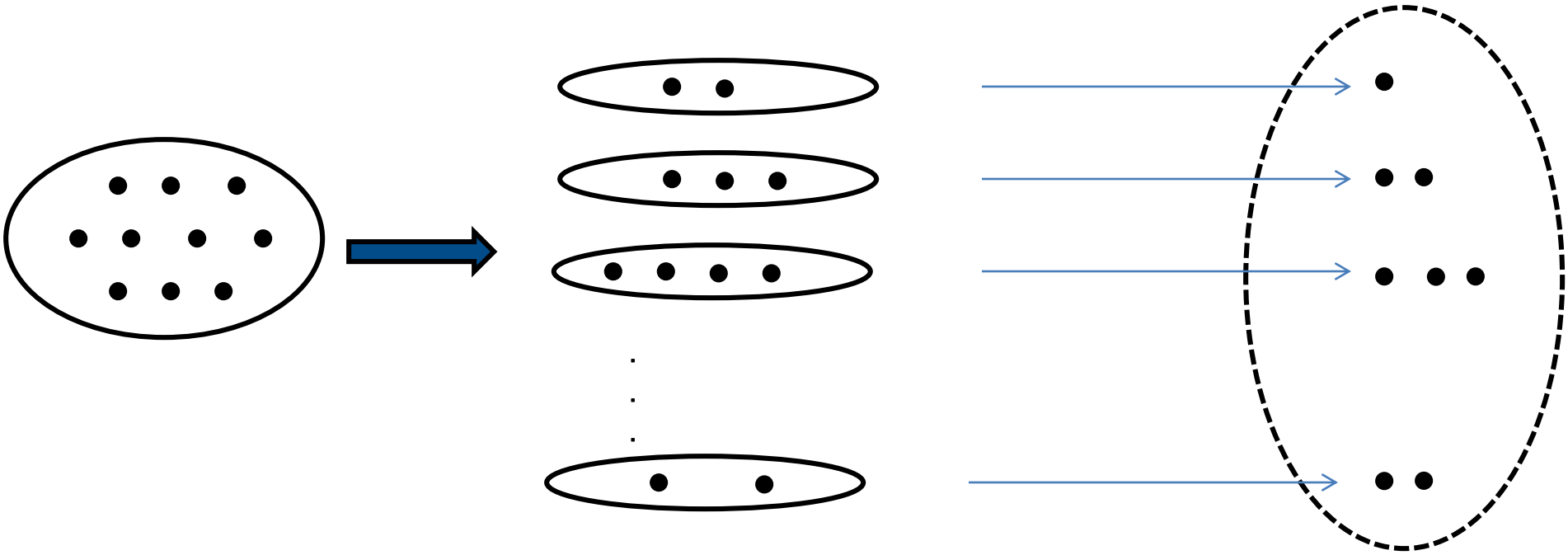


1. 随机抽样
2. 系统抽样
3. 分层抽样
4. 整群抽样

1.1 数据抽样 — 分层抽样

分层抽样

将总体分成互不交叉的层，然后按照一定的比例，从各层独立地抽取一定数量的个体，将各层取出的个体合在一起作为样本。



1.1 数据抽样 — 分层抽样

分层抽样的具体步骤

1. 分层

- 将相似的个体归入一类，即为一层。分层时要求每层的个体之间互不交叉，即遵循不重、不遗漏的原则。
- 分层的原则是增加层内的同质性和层间的异质性，即要求层内样本的差异要小，层与层之间的样本差异要大，且互不重叠。
- 常见的分层变量有性别、年龄、教育、职业等。

分层抽样的具体步骤

2. 确定抽样比和各层的样本数。为了保证每个个体等可能入样，所有层应采用同一抽样比等可能抽样。

各层样本数的确定方法有3种:

- 分层定比。即各层样本数与该层总体数的比值相等。例如，样本大小n=50，总体N=500，则n/N=0.1即为样本比例，每层均按这个比例确定该层样本数。

- 非比例分配法。当某个层次包含的个体数在总体中所占比例太小时，为使该层的特征在样本中得到足够的反映，可人为地适当增加该层样本数在总体样本中的比例。

1.1 数据抽样 — 分层抽样

分层抽样的具体步骤

- 奈曼法。即各层应抽样本数与该层总体数及其标准差的积成正比。

奈曼抽样单位分配(Neyman allocation of sampling units)简称“奈曼分配”。分层抽样中，在样本容量n固定的情形下，从每层抽选的单位个数，与该层的单位总数和标准差成比例的样本单位分配方法，由奈曼在1934年提出。设$N_i(i=1,2,\ldots,K)$是第i层的抽样单位总数，S_i是第i层(所考察统计标志)的标准差，则奈曼抽样单位分配从第i层抽选的单位数n_i为：

$$n_i = \frac{n \cdot N_i S_i}{\sum_{j=1}^{K} N_i S_i}, \qquad (i = 1, \ldots, K)$$

分层抽样的具体步骤

3. 抽取个体

每一层中，抽取时采用系统抽样或简单随机抽样、各层的抽取数之和应等于样本容量。

1.1 数据抽样 — 分层抽样

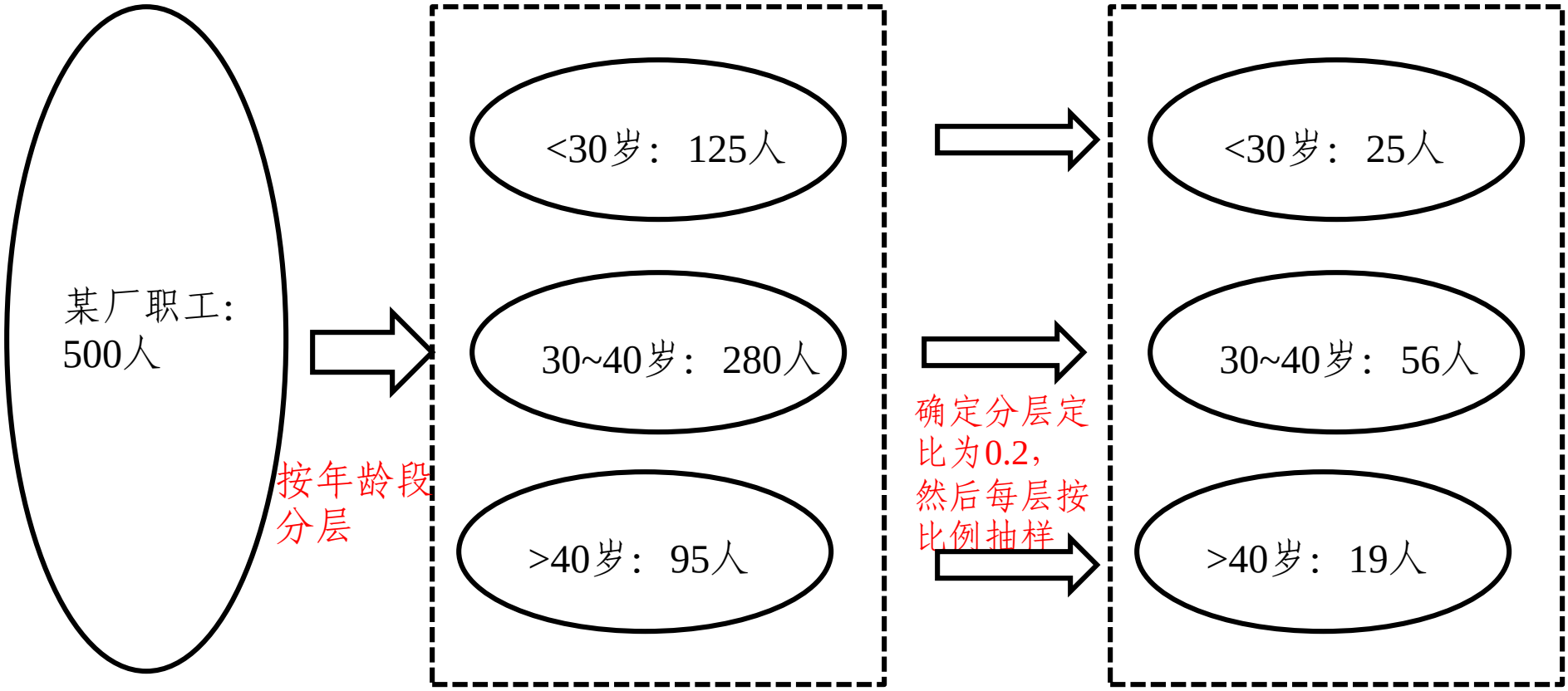
分层抽样的特点

- 分层抽样主要特征是分层按比例抽样，主要应用于总体中的个体有明显差异。每个个体被抽到的概率都相等n/N。
- 分层减小了各抽样层变异性的影响，抽样保证了所抽取的样本具有良好的代表性。

分层抽样示例

一个单位的职工有500人，其中不到30岁的有125人，30~ 40岁的有280人，40岁以上的有95人。为了了解该单位职工年龄与身体状况的有关指标，从中抽取100名职工作为样本，应该怎样抽取?

1.1 数据抽样 — 分层抽样

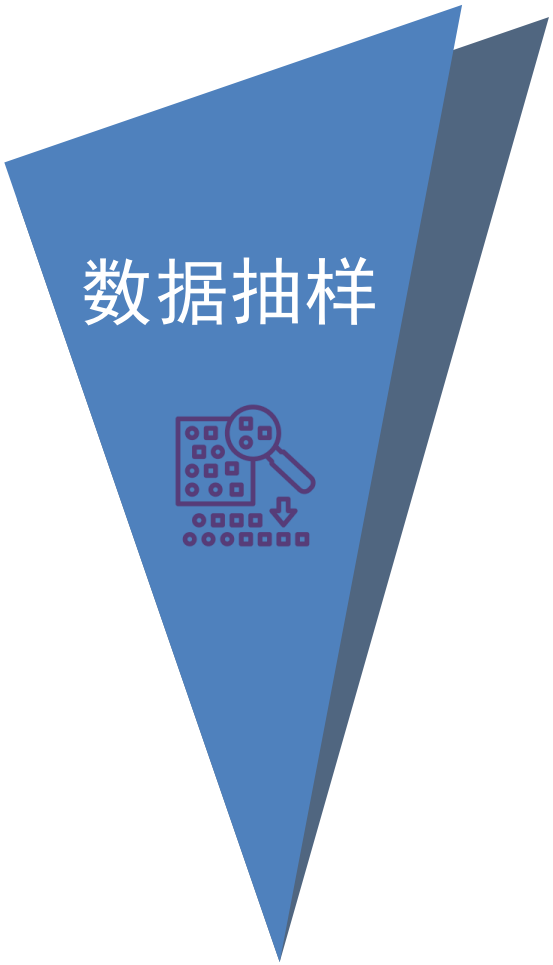

某厂职工:
500人
按年龄段
分层
<30岁： 125人
30~40岁： 280人
>40岁： 95人
确定分层定
比为0.2，
然后每层按
比例抽样
<30岁： 25人
30~40岁： 56人
>40岁： 19人

1.1 数据抽样 — 分层抽样

小结：

- 分层抽样是等概率抽样，在整个抽样过程中每个个体被抽到的概率均为n/N。
- 分层抽样是建立在简单随机抽样或系统抽样的基础上的，由于充分利用了已知信息，因此它获取的样本更具代表性，在实用中更为广泛。

1.1 数据抽样

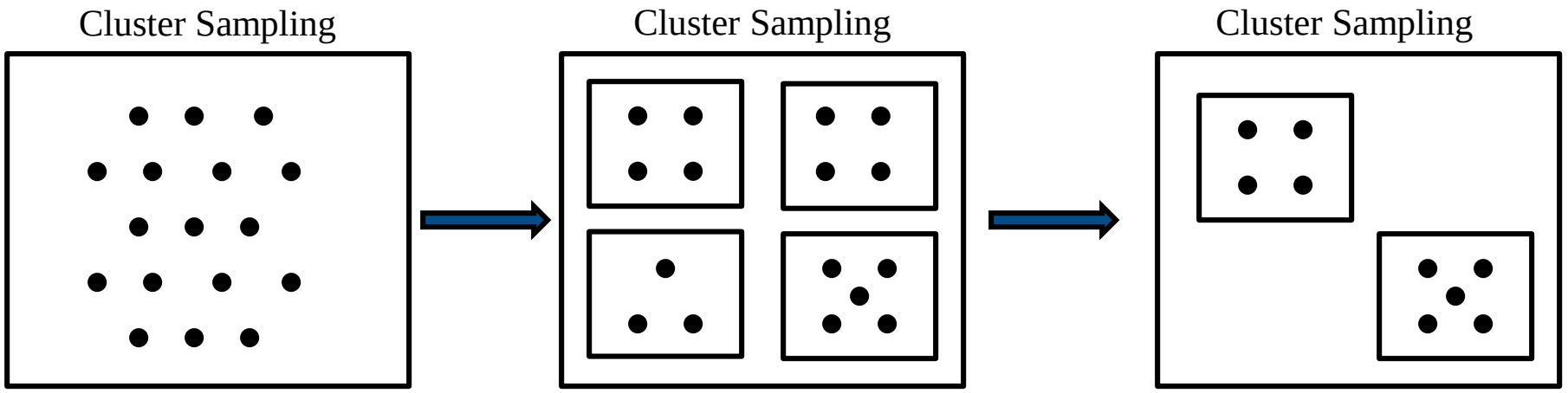


1. 随机抽样
2. 系统抽样
3. 分层抽样
4. 整群抽样

1.1 数据抽样 — 整群抽样

整群抽样

将总体中的个体归并成若干个互不交叉、互不重复的集合，称为群，然后以群为抽样单位抽取样本。



1.1 数据抽样 — 整群抽样

整群抽样的实现步骤

1、确定分群的标注。把总体（N）分成若干个互不重叠的部分，每个部分为一群。

2、根据样本量，确定应该抽取的群数。

3、采用简单随机抽样或系统抽样方法，从群中抽取确定的群。

1.1 数据抽样 — 整群抽样

整群抽样的特点

- 应用整群抽样时，要求各群有较好的代表性，即群内各单位的差异要大，群间差异要小。
- 整群抽样优点是实施方便、节省经费。
- 整群抽样的缺点是由于不同群之间的差异较大，由此而引起的抽样误差往往大于简单随机抽样、且样本分布面不广、样本对总体的代表性相对较差等。

1.1 数据抽样 — 整群抽样

分层抽样与整群抽样的比较

- 分层抽样要求各层之间的差异很大，层内个体差异小；
- 整群抽样要求群与群之间的差异比较小，群内个体差异大；

- 分层抽样的样本是按比例从每个层内抽取若干单元或个体构成；
- 整群抽样则是要么整群抽取，要么整群不被抽取。

1.2 数据过滤

数据过滤

可以通过设置限定条件来选择满足某种条件的数据，从而减少数据量。

在大数据处理过程中，数据过滤可以采用数据库的基本操作来实现，将过滤条件转换为选择操作来实现。

1.2 数据过滤

数据过滤举例

在电子商城图书的销售表中，对“小说”类别的图书的销量进行分析，就可以在整个销售表中选择出类别为“小说”的图书。

			类型						
			小说						
			小说						
			散文						
			散文						
			诗歌						
			诗歌						
			小说						
			小说						
			散文						
			诗歌						
			散文						
			小说						

1.2 数据过滤

数据过滤举例

在电子商城图书的销售表中，对“小说”类别的图书的销量进行分析，就可以在整个销售表中选择出类别为“小说”的图书。

			类型						
			小说						
			小说						
			散文						
			散文						
			诗歌						
			诗歌						
			小说						
			小说						
			散文						
			诗歌						
			散文						
			小说						

数据抽样和过滤 —小结

抽样和过滤的相同点

减少要处理的数据量，使得当前的处理能力能够处理这些数据。

抽样和过滤的区别

- 抽样主要依赖随机化技术，从数据中随机选出一部分样本。
- 而过滤依据限制条件仅选择符合要求的数据，参与下一步骤的计算。

1.1 数据抽样

例1: 下列抽取样本的方式是否属于简单随机抽样?

(1)从无限个个体中抽取100个个体作为样本。

(2)盒子里共有80个零件，从中选出5个零件进行质量检验。在抽样操作时，从中任意拿出一个零件进行质量检验后再把它放回盘子里。

(3)从20 件玩具中一次性抽取3件进行质量检验。

(4) 某班有56 名同学，指定个子最高的5名同学参加学校组织的篮球赛。

解: (1) 不是简单随机抽样。因为被抽取的样本总体的个体数是无限的，而不是有限的。

(2)不是简单随机抽样。因为它是放回抽样。

(3)不是简单随机抽样。因为这是“一次性”抽取， 而不是“逐个”抽取。

(4) 不是简单随机抽样。因为不是等可能抽样。

1.1 数据抽样

例2: 某校高一年级共有20个班，每班有50名学生。为了了解高一学生的视力状况，从这1000人中抽取一个容量为100的样本进行检查，应该怎样抽样?

- 通常先将各班平均分成 5组 ，再在第一组(1到 10号学生)中用抽签法抽取一个，然后按照 “ 逐次加 10(每组中个体个数)” 的规则分别确定学号为 11 到 20 、21 到 30 、31 到 40 、41 到 50 的学生代表。
- 将总体平均分成几个部分，然后按照预先定出的规则，从每个部分中抽取一个个体，得到所需的样本，这样的抽样方法称为系统抽样。

1.1 数据抽样

例3 :为了解1200名学生对学校教改试验的意见，打算从中抽取一个容量为30的样本，考虑采用系统抽样，则分段间隔是多少？

40

1.1 数据抽样

例4: 某商场新进3000袋奶粉，为检查其三聚氰胺是否超标，先采用系统抽样的方法从中抽取150袋检查，若第一组抽取号码是11，则第61组抽出的号码？

$11+60\times20=1211$

1.1 数据抽样

例5:某工厂生产产品，用传送带将产品送放下一道工序，质检人员每隔十分钟在传送带的某一个位置取一件检验，这种抽样方法是哪种抽样?

A. 简单随机抽样　B. 系统抽样

C. 分层抽样　D. 非上述答案

由于系统抽样间隔一样，故该题中的抽样是系统抽样。

1.1 数据抽样

例子7：某初级中学有学生270人，其中七年级108人，八、九年级各81人，现要利用抽样方法抽取10人，考虑选用简单随机抽样、分层抽样和系统抽样。使用简单随机抽样和分层抽样时，将学生按7、8、9年级依次统一编号为1，2, ..., 270，并将整个编号依次分10段，如果抽得的号码有以下四种情况

① 7,34,61 ,88,115,142,169,196,223,250

② 5,9,100,107,111,121,180,195,200,265

③ 11,38,65,92,119,146,173,200,227,254

④ 30,57,84,111 ,138,165,192,219,246,270

关于上述样本的下列结论中，正确的是(　　).

A. ②③都不能为系统抽样

B. ②④都不能为分层抽样

C. ①④都可能为系统抽样

D. ①③都可能为分层抽样

因为七、八、九年级的人数之比为108：81：81 = 4：3：3，又因为共抽取10人，根据系统抽样和分层抽样的特点可知:

①②③都可能为分层抽样，所以答案B不对;

②④不可能为系统抽样，所以答案C不对;

①③可能为系统抽样，所以答案A不对;

故选D.

2 数据标准化和归一化

1 数据抽样和过滤

2 **数据标准化与归一化**

3 数据清洗

2 数据标准化和归一化

引例：5 名新出生婴儿的体重(斤)资料为: 5、6、7、8、9; 同时又有5 名成年人的体重(斤)资料为: 130、131、132、133、134, 要求对比分析两组人员体重差异的大小。

从数据表面看, 两组人员体重的平均差异均是1斤,

两组人员体重的差异和程度相同?

- 两组人员的体重水平不在同一等级上, 即量纲不同。从外观上看, 婴儿的体重相差1斤就比较明显了, 而成年人体重相差1 斤则基本察觉不到。
- 这时比较两组人员体重上差异的大小, 不应该用平均差异, 而应该消除其量纲(即体重基本水平)上的不同。

2 数据标准化和归一化

解决办法：用相对数表示为：体重的平均差异值/平均体重，即用1/7和1/ 132进行比较、分析。

- 这种简单的对比分析的过程，表面上看是指标的选用问题，实际上则是指标数值无量纲化的处理问题。
- 为了消除指标之间的量纲和取值范围差异的影响，需要进行无量纲化处理。

2 数据标准化和归一化

无量纲化

- 无量纲化，也叫数据的标准化、规格化。它是通过简单的数学变换来消除各指标量纲影响的方法。
- 数据的标准化（normalization）是将数据按比例缩放，使之落入一个小的特定区间。
  - 在某些比较和评价的指标处理中经常会用到，去除数据的单位限制，将其转化为无量纲的纯数值，便于不同单位或量级的指标能够进行比较和加权。

2 数据标准化和归一化

数据标准化

- 例如，在多指标评价体系中(评价房价、预测某个人患病率等问题)，各评价指标的性质不同，通常具有不同的量纲和量纲单位。
  - 量纲：对于评价房价来说量纲包括面积、房间数、楼层等；对于预测某个人患病率来说量纲包括身高、体重等。
  - 量纲单位：例如：面积单位：平方米、平方厘米等；身高：米、厘米等。
- 当各指标间的水平相差很大时，如果直接用原始指标值进行分析，就会突出数值较高的指标在综合分析中的作用，相对削弱数值水平较低指标的作用。

2 数据标准化和归一化

数据标准化

- 因此，为了保证结果的可靠性，消除指标之间量纲的影响，需要进行数据标准化处理，以解决数据指标之间的可比性。

原始数据经过数据标准化处理后，各指标处于同一数量级，适合进行综合对比评价。

2 数据标准化和归一化

- 目前数据标准化方法有多种，归结起来可以分为直线型方法(如极值法、标准差法)、折线型方法(如三折线法)、曲线型方法(如半正态性分布)。不同的标准化方法，对系统的评价结果会产生不同的影响，
- 注意：在数据标准化方法的选择上，还没有通用的法则可以遵循。

2 数据标准化和归一化

1. 直线型无量纲化方法

指在指标实际值转化成不受量纲影响的指标值时，假定二者之间呈线性关系，指标实际值的变化引起标准化后数值一个相应的比例变化。线性无量纲化方法主要有：

(1) 极值法

(2) 标准差标准化法

2 数据标准化和归一化

1. 直线型无量纲化方法

(1) 极值法

利用指标的极值(极大值或极小值)计算指标的无量纲值x_i'

$$x'_i = \frac{x_i}{max(x_i)} \qquad x'_i = \frac{max(x_i) - x_i}{max(x_i)}$$

$$x'_i = \frac{x_i - min(x_i)}{x_i} \qquad x'_i = \frac{x - min(x_i)}{max(x_i) - min(x_i)}$$

2 数据标准化和归一化

1. 直线型无量纲化方法

(2) 标准差标准化法，其计算公式为：

$$x'_i = \frac{x_i - \bar{x}}{s} \qquad \text{其中，} s = \sqrt{\frac{1}{n}\sum (x_i - \bar{x})^2}$$

这种方法在原始数据呈正态分布的情况下，转化结果是比较合理的。

2 数据标准化和归一化

2. 折线型无量纲化方法

有些情况下，指标在不同水平、区域内的变化对综合分析结果的影响是不一样的。比如在多指标综合评价时，

- 若x小于某个数值时，x变化对综合水平影响较大，评价值也有较大的变化；
- 而当x大于该值时，x的变化对被评价对象综合水平的影响较小，则评价值的变化也较小。

2 数据标准化和归一化

2. 折线型无量纲化方法

这时，应采用折线型的无量纲化方法来处理。三折线公式如下：

$$x' = \begin{cases} 0 & x_i < a \\ \dfrac{x_i - a}{b - a} & a \leqslant x_i < b \\ 1 & x_i \geqslant b \end{cases}$$

2 数据标准化和归一化

3. 曲线型无量纲化方法

采用曲线型的无量纲化方法，意味着指标实际值与无量纲值之间不是等比例的变动，而是非线性关系。曲线型公式种类很多，如：

半正态型分布

$$x_i' = \begin{cases} 0 & 0 \leqslant x_i \leqslant a \\ 1 - e^{k(x-a)^2} & x_i > a \end{cases}$$

式中，k, a, b为曲线待定参数。

2 数据标准化和归一化

数据归一化

数据标准化中最典型的就是数据归一化，其目标是：

- 把数变为（0，1）之间的小数

  主要是为了数据处理方便提出来的，把数据映射到0～1范围之内处理，更加便捷快速，应该归到数字信号处理范畴之内。

2 数据标准化和归一化

数据归一化

- 把有量纲表达式变为无量纲表达式

  归一化是一种简化计算的方式，即将有量纲的表达式，经过变换，化为无量纲的表达式，成为纯量。

- 比如，复数阻抗可以归一化书写：$Z = R + j\omega L = R(1 + j\omega L/R)$，复数部分变成了纯数量了，没有量纲。

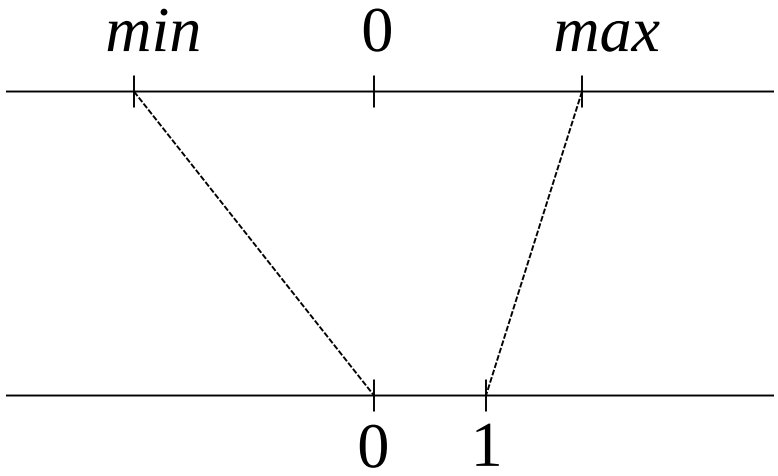
- 另外，在电路分析、信号系统、电磁波传输等，有很多运算都可以如此处理，既保证了运算的便捷，又能凸现出物理量的本质含义。

2 数据标准化和归一化

常用方法

（1）0-1标准化(*min-max*标准化)

也叫离差标准化，是对原始数据的线性变换，使结果落到[0,1]区间。



转换函数：

$$h(x) = \frac{x - min}{max - min}$$

2 数据标准化和归一化

常用方法

如果想要将数据映射到[-1,1]，则将公式换成：

$$x^* = \frac{x - mean}{max - min}$$

其中mean表示数据的均值。

- 缺陷：就是当有新数据加入时，可能导致max和min的变化，需要重新定义。

2 数据标准化和归一化

常用方法

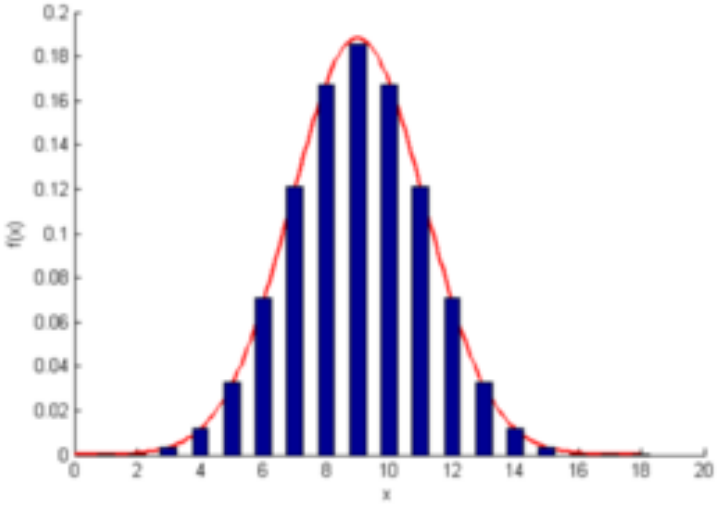
（2）Z-score标准化（ 0均值化的归一方法）

给予原始数据的均值（mean）和标准差（standard deviation）进行数据的标准化。经过处理的数据符合标准正态分布，即均值为0，标准差为1，转化函数为:

$$h(x)=(x-\mu)/\sigma$$

μ为所有样本数据的均值，

σ为所有样本数据的标准差。



2 数据标准化和归一化

注意：

- 一般来说z-score不是归一化，而是标准化。
- 归一化只是标准化的一种。

两点说明：

- z-score标准化方法适用于属性*A*的最大值和最小值未知的情况，或有超出取值范围的离群数据的情况。
- 该标准化方式要求原始数据的分布可以近似为高斯分布，否则效果会变得很糟糕。

2 数据标准化和归一化

归一化和标准化的异同：

- 归一化是将样本的特征值转换到同一量纲下把数据映射到[0,1]或者[-1, 1]区间内，仅由变量的极值决定。
- 标准化是依照特征矩阵的列处理数据，其通过求z-score的方法，转换为标准正态分布，和整体样本分布相关，每个样本点都能对标准化产生影响。
- 它们的相同点在于都能取消由于量纲不同引起的误差；都是一种线性变换，都是对向量X按照比例压缩再进行平移。

2 数据标准化和归一化

0-1标准化和Z-score的应用场景

1、在分类、聚类算法中，需要使用距离来度量相似性的时候、或者使用主成分分析(PCA)技术进行降维的时候，Z-score方法表现更好。

原因：使用0-1标准化方法(线性变换后)，其协方差产生了倍数值的缩放，因此这种方式无法消除量纲对方差、协方差的影响，对PCA分析影响巨大；同时，由于量纲的存在，使用不同的量纲、距离的计算结果会不同。

2 数据标准化和归一化

0-1标准化和Z-score的应用场景

2、在不涉及距离度量、协方差计算、数据不符合正态分布的时候，可以使用0-1标准化方法。比如图像处理中，将RGB图像转换为灰度图像后将其值限定在[0, 255]的范围。

原因：Z-Sroce方式中，新的数据由于对方差进行了归一化，这时候每个维度的量纲其实已经等价了，每个维度都服从均值为0、方差1的正态分布，在计算距离的时候，每个维度都是去量纲化的，避免了不同量纲的选取对距离计算产生的巨大影响。

2 数据标准化和归一化

0-1标准化和Z-score的应用场景

小结:

- 涉及距离度量（聚类分析）或者协方差分析（PCA、LDA等）的，同时数据分布可以近似为正态分布，应当使用Z-score方法。
- 其它应用中，根据具体情况选用合适的归一化方法。

2 数据标准化和归一化

例子：如对于一个178*13的数据集，每一行代表一瓶酒，每一列代表一种特征（一共13个特征）。而在这组数据中，特征5和特征13相对于其他数据，是两组奇异样本数据，所谓奇异样本，是相对于其他输入样本特别大或特别小的样本矢量。

178x13 double

	1	2	3	4	5	6	7	8	9	10	11	12	13
1	14.2300	1.7100	2.4300	15.6000	127	2.8000	3.0600	0.2800	2.2900	5.6400	1.0400	3.9200	1065
2	13.2000	1.7800	2.1400	11.2000	100	2.6500	2.7600	0.2600	1.2800	4.3800	1.0500	3.4000	1050
3	13.1600	2.3600	2.6700	18.6000	101	2.8000	3.2400	0.3000	2.8100	5.6800	1.0300	3.1700	1185
4	14.3700	1.9500	2.5000	16.8000	113	3.8500	3.4900	0.2400	2.1800	7.8000	0.8600	3.4500	1480
5	13.2400	2.5900	2.8700	21	118	2.8000	2.6900	0.3900	1.8200	4.3200	1.0400	2.9300	735
6	14.2000	1.7600	2.4500	15.2000	112	3.2700	3.3900	0.3400	1.9700	6.7500	1.0500	2.8500	1450
7	14.3900	1.8700	2.4500	14.6000	96	2.5000	2.5200	0.3000	1.9800	5.2500	1.0200	3.5800	1290
8	14.0600	2.1500	2.6100	17.6000	121	2.6000	2.5100	0.3100	1.2500	5.0500	1.0600	3.5800	1295
9	14.8300	1.6400	2.1700	14	97	2.8000	2.9800	0.2900	1.9800	5.2000	1.0800	2.8500	1045
10	13.8600	1.3500	2.2700	16	98	2.9800	3.1500	0.2200	1.8500	7.2200	1.0100	3.5500	1045
11	14.1000	2.1600	2.3000	18	105	2.9500	3.3200	0.2200	2.3800	5.7500	1.2500	3.1700	1510

2 数据标准化和归一化

- 在对数据进行操作前，先将数据归一化处理，使其无量纲化，避免较大数值的数据的变化掩盖掉小数值的变化。

89x13 double

	1	2	3	4	5	6	7	8	9	10	11	12	13
1	0.6262	-0.6450	0.1505	-0.4845	0.4074	0.2683	0.5153	-0.4000	0.4745	-0.1391	0.1546	0.9414	0.1227
2	-0.0156	-0.6147	-0.1613	-0.9381	-0.2593	0.1638	0.3482	-0.4800	-0.3176	-0.4172	0.1753	0.5604	0.1013
3	-0.0405	-0.3636	0.4086	-0.1753	-0.2346	0.2683	0.6156	-0.3200	0.8824	-0.1302	0.1340	0.3919	0.2939
4	0.7134	-0.5411	0.2258	-0.3608	0.0617	1	0.7549	-0.5600	0.3882	0.3377	-0.2165	0.5971	0.7147
5	0.0093	-0.2641	0.6237	0.0722	0.1852	0.2683	0.3092	0.0400	0.1059	-0.4305	0.1546	0.2161	-0.3481
6	0.6075	-0.6234	0.1720	-0.5258	0.0370	0.5958	0.6992	-0.1600	0.2235	0.1060	0.1753	0.1575	0.6719
7	0.7259	-0.5758	0.1720	-0.5876	-0.3580	0.0592	0.2145	-0.3200	0.2314	-0.2252	0.1134	0.6923	0.4437
8	0.5202	-0.4545	0.3441	-0.2784	0.2593	0.1289	0.2089	-0.2800	-0.3412	-0.2693	0.1959	0.6923	0.4508
9	1	-0.6753	-0.1290	-0.6495	-0.3333	0.2683	0.4708	-0.3600	0.2314	-0.2362	0.2371	0.1575	0.0942
10	0.3956	-0.8009	-0.0215	-0.4433	-0.3086	0.3937	0.5655	-0.6400	0.1294	0.2097	0.0928	0.6703	0.0942
11	0.5452	-0.4502	0.0108	-0.2371	-0.1358	0.3728	0.6602	-0.6400	0.5451	-0.1148	0.5876	0.3919	0.7575

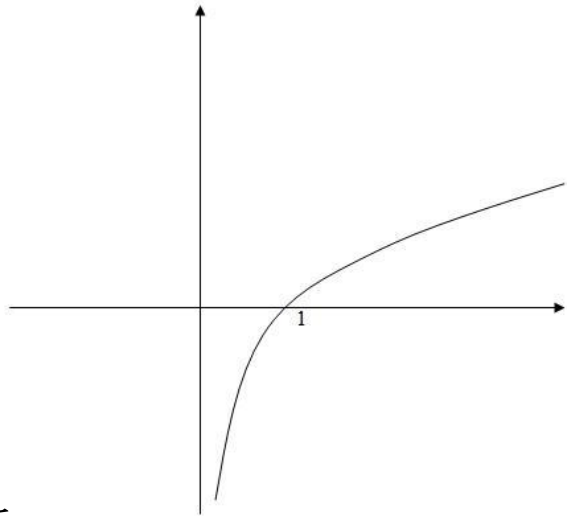
通过函数对图中的每一列数据进行归一化。归一化后的数据都被映射在[-1,1]范围内，对进一步的数据操作提供方便。

2 数据标准化和归一化

常用方法

（3）Log函数转换

$$h(x) = log_{10}\, x$$



Log函数转换

- $\log_{10}$(x)仅作了标准化，未进行归一化。

  因为这个结果并非一定落到[0,1]区间上。

$$h(x) = \frac{log_{10}\, x}{log_{10}\, max}$$

- 如果要进行归一化，应该还要除以$log_{10}(max)$, max为样本数据最大值，并且所有的数据都要大于等于1。

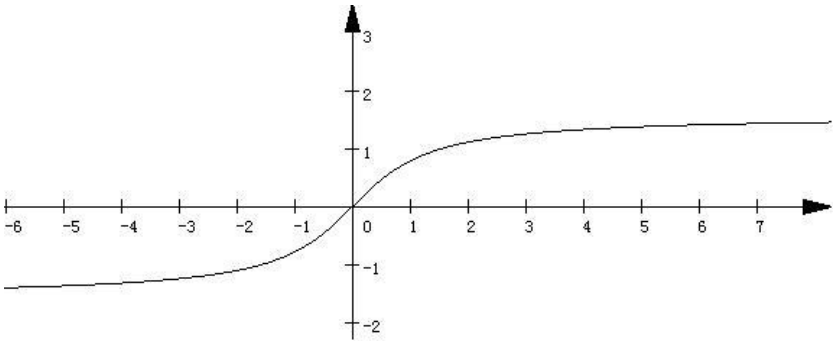
2 数据标准化和归一化

常用方法

（4）反正切函数转换

转换函数

$$h(x) = \frac{2arctanx}{\pi}$$



y=arctan x 函数图像

- 需要注意的是如果想映射的区间为[0,1]，则数据都应该大于等于0，小于0的数据将被映射到[-1,0]区间上。

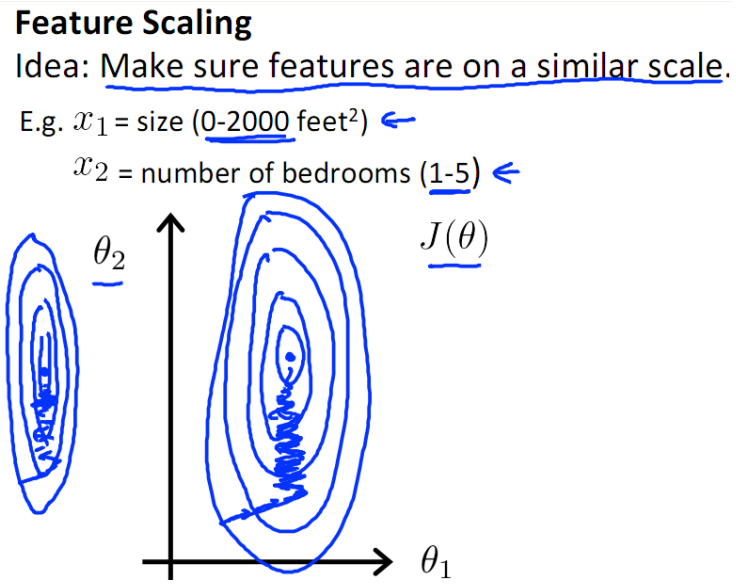
2 数据标准化和归一化

例子：

在使用梯度下降的方法求解最优化问题时， 归一化/标准化后可以加快梯度下降的求解速度，即提升模型的收敛速度。

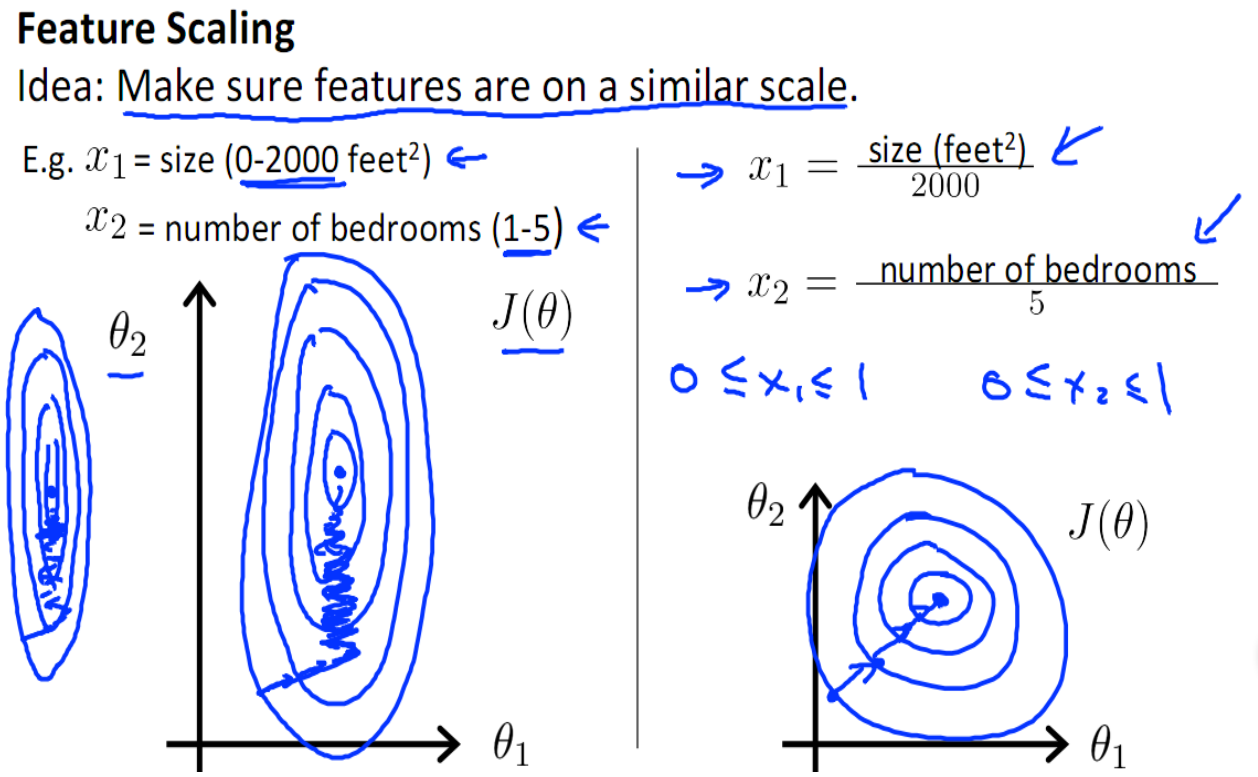
x_1的取值为0-2000，而x_2的取值为1-5，假如只有这两个特征，对其进行优化；

得到一个窄长的椭圆形，导致在梯度下降时，梯度的方向为垂直等高线的方向而走之字形路线，这样会使迭代很慢；



2 数据标准化和归一化

对两个特征进行归一化，对应的等高线就会变圆，在梯度下降进行求解时能较快的收敛。

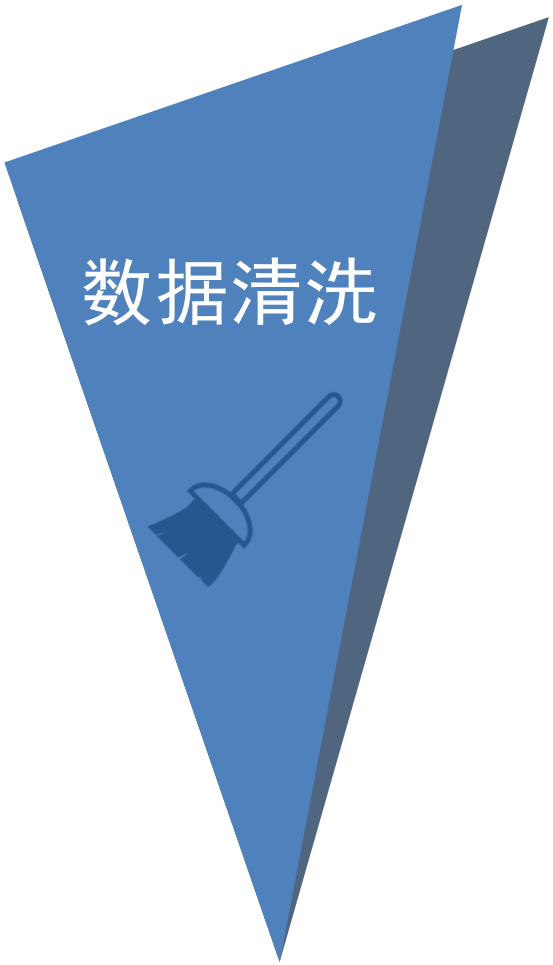


3 数据清洗

3 数据清洗

- 数据清洗的目的就是检测数据中存在的错误和不一致，对它们进行修改，保证数据的准确、一致、无冗余，从而提高数据质量，获得可信的和可用的数据。
- 数据清洗 是整个数据分析过程中不可缺少的一个环节，其结果的质量直接关系到模型效果和最终结论。
- 在实际操作中，数据清洗通常会占据分析过程50%~80%的时间。

3 数据清洗



1. 数据质量概述
2. 缺失值填充
3. 实体识别与真值发现
4. 错误发现与修复

3 数据清洗 —数据质量管理

什么是数据质量管理

指对数据生命周期（计划、获取、存储、共享、维护、应用、消亡）的每个阶段可能引发的各类数据质量问题，进行识别、度量、监控、预警等一系列管理活动，使数据质量获得进一步提高。

3 数据清洗 —数据质量管理

数据质量上存在的问题，在全球范围内已经造成了恶劣的后果，严重困扰着信息社会。

- 在医疗方面，美国由于数据错误引发的医疗事故每年导致达98,000名以上的患者死亡；
- 在工业方面，错误和陈旧的数据每年给美国的工业企业造成约6,110亿美元的损失；
- 在商业方面，美国的零售业中，每年仅错误标价这一种数据质量问题的诱因，就导致了25亿美元的损失。

数据质量管理的5个维度

1. 数据一致性

数据集合中，每个信息都不包含语义错误或相互矛盾的数据。

公司=“先导”
国码=“86”
区号=“10”
城市=“上海”

例如，数据（公司=“先导”，国码=“86”，区号=“10”，城市=“上海”）含有一致性错误，因为10是北京区号而非上海区号（21）。

数据质量管理的5个维度

2. 数据精确性

数据集合中，每个数据都能准确表述现实世界中的实体。

城市人口 4,130,465 城市人口 400万

例如，某城市人口数量为4,130,465人，而数据库中记载为400万。宏观来看，该信息是合理的，但不精确。

数据质量管理的5个维度

3. 数据完整性

数据集合中包含足够的数据来回答各种查询，并支持各种计算。

例如，某医疗数据库中的数据是一致且精确的，但遗失某些患者的既往病史，从而存在不完整性，可能导致不正确的诊断甚至严重医疗事故。

数据质量管理的5个维度

4. 数据时效性

信息集合中，每个信息都与时俱进，保证不过时。

例如，某数据库中的用户地址在2010年是正确的，但在2011年未必正确，即这个数据已经过时。

数据质量管理的5个维度

5. 实体同一性

在所有数据集合中，同一实体的标识必须相同而且数据必须一致。

姓名	性别	电话	出生年月
王小明	男	18277777777	1997.01
王晓明	男	18277777777	1997.01

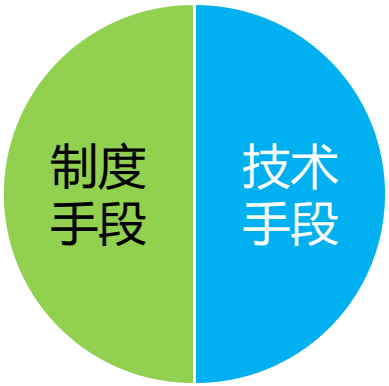
例如：企业的市场、销售和服务部门可能维护各自的数据库，如果这些数据库中的同一个实体没有相同的标识或数据不一致，将存在大量具有差异的重复数据，导致实体表达混乱。

3 数据清洗 —数据质量管理

数据质量管理的方法

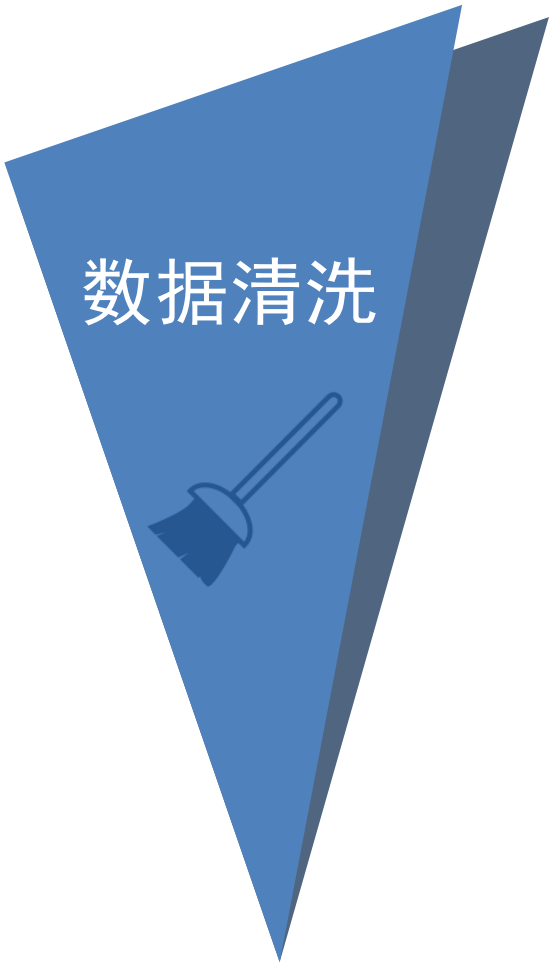
- 制度手段
  - ✓ 制定数据质量度量标准
  - ✓ 制定数据质量监管体系
  - ✓ 制定数据质量管理制度

- 技术手段
  - ✓ 面向大数据的数据质量描述
  - ✓ 大数据上错误自动发现
  - ✓ 大数据上错误自动清洗
  - ✓ 大数据上劣质容忍计算



劣质容忍计算是指在包含错误的信息和知识上完成精确或近似计算和推理。

3 数据清洗



1. 数据质量概述
2. 缺失值填充
3. 实体识别与真值发现
4. 错误发现与修复

3 数据清洗 —数据值缺失

造成数据缺失的原因是多方面的，主要可能有以下几种:

（1）有些信息暂时无法获取。例如在医疗数据库中，并非所有病人的所有临床检验结果都能在给定的时间内得到，就致使一部分属性值空缺出来。

（2）有些信息是被遗漏的。

- 可能是因为输入时认为不重要、忘记填写了;
- 对数据理解错误而遗漏;
- 由于数据采集设备的故障、存储介质的故障、传输媒体的故障、一些人为因素等原因而丢失了。

3 数据清洗 —数据值缺失

（3）有些对象的某个或某些属性是不可用的。也就是说，对于这个对象来说，该属性值是不存在的，如一个未婚者的配偶姓名、一个儿童的固定收入状况等。

（4）有些信息（被认为）是不重要的。如一个属性的取值与给定语境是无关的，或训练数据库的设计者并不在乎某个属性的取值（称为don’t-care value）。

（5）获取这些信息的代价太大。

（6）系统实时性能要求较高，即要求得到这些信息前需要迅速做出判断或决策。

3 数据清洗 —缺失值填充

缺失值填充的方法

1. 删除

最简单的方法是删除，直接删除属性或者删除样本。

- 如果大部分样本该属性都缺失，这个属性能提供的信息有限，可以选择放弃使用该维属性；
- 如果一个样本大部分属性缺失，可以选择放弃该样本。

姓名	性别	籍贯	电话	驾龄/年
王小明	男	湖南	18277777777	10
李刚	男		18266666666	1
张一	女		13634567890	5
王五	女			

姓名	性别	籍贯	电话	驾龄/年
王五	女			

3 数据清洗 —缺失值填充

说明:

- 方法简单易行，如果对象有多个属性的值缺失，并且对象与信息表中的数据量相比非常小，该方法是非常有效的。

- 该方法有很大的局限性。它是以减少历史数据来换取信息的完备，会造成资源的大量浪费，丢弃了大量隐藏在这些对象中的信息。

- 如果信息表本来就包含很少的对象，删除少量对象就足以严重影响到信息表信息的客观性和结果的正确性。比如，当遗漏数据所占比例较大，特别当遗漏数据非随机分布时，这种方法可能导致数据发生偏离，从而引出错误的结论。

3 数据清洗 —缺失值填充

2. 统计填充

- 对于缺失值的属性，尤其是数值类型的属性，根据所有样本关于这维属性的统计值进行填充。如使用平均数、中位数、众数、最大值、最小值等。
- 如果有可用类别信息，还可以进行类内统计，比如身高，男性和女性的统计填充应该是不同的。

姓名	性别	籍贯	电话	驾龄/年	年收入
王小明	男	湖南	18277777777	10	
李刚	男		18266666666	1	20万
张一	女		13634567890	5	10万
王五	女				15万

- 在商品推荐场景下填充平均值，15万。
- 借贷额度下填充最小值，10万。

3 数据清洗 —缺失值填充

3. 统一填充

- 对于含缺失值的属性，把所有缺失值统一填充为自定义值。
- 常用的统一填充值有：“空”、“0”、“正无穷”、“负无穷”等。
- 如果有可用类别信息，也可以为不同类别分别进行统一填充。

姓名	性别	电话	驾龄/年	本科毕业时间
王小明	男	18277777777	10	$+\infty$
李刚	男	18266666666	0	2000年

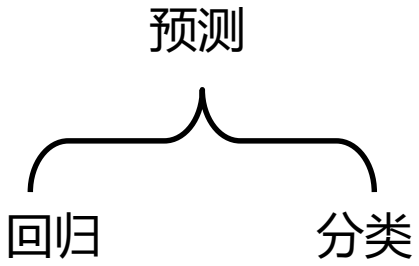
3 数据清洗 —缺失值填充

4. 预测填充

通过预测模型，利用不存在缺失值的属性来预测缺失值，也就是先用预测模型把数据填充，然后再做进一步的工作，如统计、学习等。

➢对于数值属性，可以采用回归的方法进行填充。通过建模，以模型预测值作为缺失值的估计值。

➢对于类别属性，可以使用分类方法进行填充，如朴素贝叶斯方法。



3 数据清洗 —缺失值填充

(1) 对于数值属性，可以采用回归的方法进行填充

- 基于完整的数据集，建立回归方程（模型）。
- 对于包含空值的对象，将已知属性值代入方程来估计未知属性值，以此估计值来进行填充。
- 如果变量之间的回归关系比较显著，那么通过回归模型得到的估计值往往更接近于真实值;
- 但构造和评估回归费时繁琐，需要对模型进行评价，因此多用于对重要变量缺失值的填充。

3 数据清洗 —缺失值填充

(2) 对于类别属性，可以使用分类方法进行填充，如朴素贝叶斯方法。

- 朴素贝叶斯的思想：根据某些先验概率来计算Y变量属于某个类别的后验概率。

朴素贝叶斯公式：

$$P(Y \mid X) = \frac{P(YX)}{P(X)} = \frac{P(X \mid Y)P(Y)}{P(X)}$$

- 如果要确定某个样本属于哪一类，则需要计算属于不同类的概率，再从中挑选出最大的概率。

数据清洗

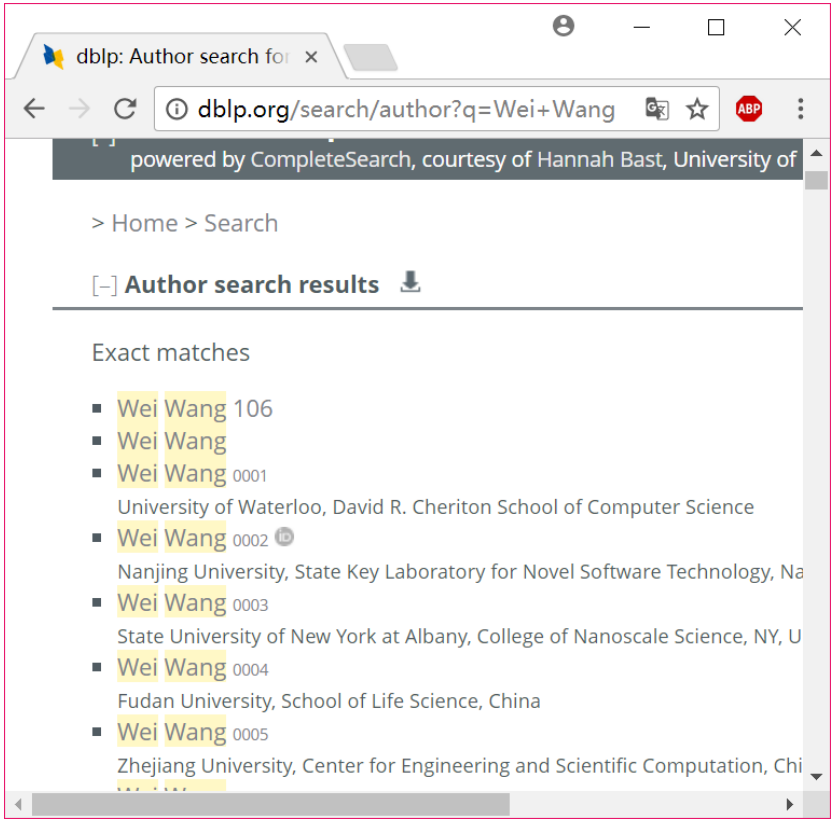
1. 数据质量概述
2. 缺失值填充
3. 实体识别与真值发现
4. 错误发现与修复

3 数据清洗 —实体识别

实体识别

引例

当DBLP中检索“Wei Wang”的文章时，会检索到14个“Wei Wang”的197篇文章。



3 数据清洗 —实体识别

什么是实体识别

在给定的对象集合中，正确发现不同的实体对象，并将其聚类，使得每个经过实体识别后得到的对象簇在现实世界中指代的是同一实体。

3 数据清洗 —实体识别

实体识别需要解决的两个问题

1. 冗余问题：同一类实体可能由不同的名字指代。

Name	affiliation
Wei Wang	National University of Singapore
Wang Wei	National University of Singapore

3 数据清洗 —实体识别

实体识别需要解决的两个问题

2. 重名问题：不同类的实体可能由相同的名字指代。

Name	affiliation
Wei Wang	National University of Singapore
Wei Wang	Fudan University, School of Life Science, China

3 数据清洗 —实体识别

实体识别的两类技术

1. 冗余发现

- 构造对象名称的相似性函数，计算对象之间的相似性，并与阈值比较，从而判定对象是否属于同一实体类。
- 冗余发现主要是利用对象名字的相似性来解决实体识别中的异名问题，这种方法的局限性在于不能有效够区分具有相同名字的不同实体。

3 数据清洗 —实体识别

实体识别的两类技术

2. 重名检测

- 使用基于聚类的处理技术，即利用对象每个属性值的相似性来判断对象之间的相似性，来决定对象是否属于同一实体类。
- 这种方法仅仅针对具有相同名字的实体识别，若直接用于识别具有不同名字却指代相同实体的数据对象，会导致结果的召回率降低。

3 数据清洗 —实体识别

真值发现

经过实体识别之后，描述现实世界中同一个实体的不同元组被聚到了一起，然而这些对象的相同属性值可能包含冲突值。

Name	affiliation	Age
Wei Wang	National University of Singapore	41
Wang Wei	National University of Singapore	47
Wang Wei	National University of Singapore	47

真值发现就是在这些冲突值中，发现真实的值。

3 数据清洗 —实体识别

真值发现的两个思路

（1）投票方法

- 采用基本投票的方式进行选择真值，认为被更多数据源描述的值更准确。
- 根据数据值的投票数来判断其准确性。

Name	affiliation	Age
Wei Wang	National University of Singapore	41
Wang Wei	National University of Singapore	47
Wang Wei	National University of Singapore	47

3 数据清洗 —实体识别

投票方法的说明：

- 因为数据源本身提供的数据有真有假，因此数据源本身就具有可靠性。这种方法对每个数据源同等对待，没有考虑不同数据源可信性的差异，而这种差异在实际中普遍存在。
- 另一方面，假数据可能通过拷贝等方式被广泛传播，尤其针对互联网这种传播性很强的环境而言，这种情况下假数据得票值可能超过真值的得票数，从而以假乱真。

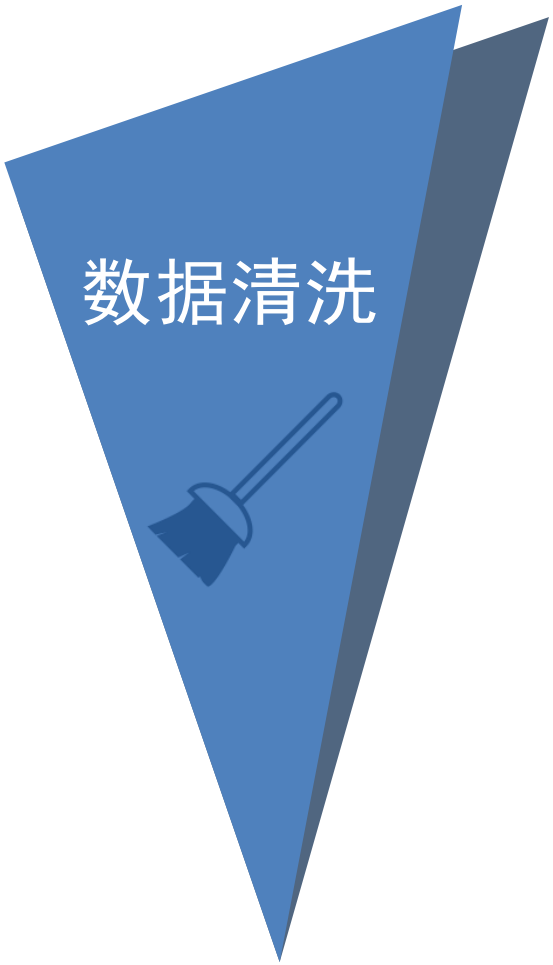
真值发现的两个思路

（2）考虑数据源精度的迭代方法

- 迭代地计算数据源的可信度，进而计算事实(数据值)的置信度。
- 由越多高可信性数据源提供的数据值其正确的可能性越大，提供越多正确数据的数据源其可信性也越高，以此对数据源的质量进行建模。进一步提高真值发现的准确性。

Name	affiliation	Age
Wei Wang	Naonal Unrsity of Singpe	47
Wang Wei	Natiol Uniety of Sinaore	47
Wei Wang	National University of Singapore	41
Wang Wei	National University of Singapore	41

3 数据清洗 —错误发现与修复



1. 数据质量概述
2. 缺失值填充
3. 实体识别与真值发现
4. 错误发现与修复

3 数据清洗 —错误发现与修复

- 如果数据是由系统日志而来，那么通常在格式和内容方面，会与元数据的描述一致。
- 如果数据是由人工收集或用户填写而来，则有很大可能性在格式和内容上存在一些问题。

3 数据清洗 —错误发现与修复

一、格式内容清洗

1. 显示格式不一致(时间、日期、数值、全半角等)

这种问题通常与输入端有关，在整合多来源数据时也有可能遇到，将其处理成一致的某种格式即可。

Date
2017/10/2
2017-10-2

3 数据清洗 —错误发现与修复

一、格式内容清洗

2. 内容中有非法字符或有不该存在的字符

➢某些属性值只允许包含一部分字符，如身份证只包含数字和X，中国人姓名是汉字(赵C这种情况还是少数)。

身份证号
3607281999010Y010X

➢最典型的就是头、尾、中间的空格，也可能出现姓名中存在数字符号、身份证号中出现汉字等问题。

➢这种情况下，需要以半自动校验半人工方式来找出可能存在的问题，并去除不需要的字符。

3 数据清洗 —错误发现与修复

一、格式内容清洗

3. 内容与该字段应有内容不符

可能是用户将本属于一个属性的数据填写到了另一个属性中。姓名写成了性别，身份证号写成了手机号等等，均属这种问题。

姓名	性别
小明	男
男	小刚

- 但该问题特殊性在于：并不能简单删除。因为有可能是人工填写错误，也有可能是前端没有校验，还有可能是导入数据时部分或全部存在列没有对齐的问题。
- 因此要详细识别问题类型。

二、逻辑错误清洗

去掉一些使用简单逻辑推理就可以直接发现问题的数据，防止分析结果走偏。

二、逻辑错误清洗

1.去重

去除重复信息，解决数据中存在的同名和异名问题。

- 去重通常通过实体识别技术来实现;
- 去重后出现的冲突值使用真值发现技术来消解。

二、逻辑错误清洗

2.去除不合理值

某比如年龄2000岁，月收入100,000万。

这类不合理的检测主要依靠属性值上的约束。

二、逻辑错误清洗

3.修正矛盾内容

有些字段是可以互相验证的。比如某用户电话的区号为“010”，但所在城市是“上海”；举例：身份证号是1101031980XXXXXXXXX，然后年龄填18岁，

- 根据字段的数据来源，来判定哪个字段提供的信息更为可靠，去除或重构不可靠的字段。

二、逻辑错误清洗的说明

- 逻辑错误除了以上列举的情况，还有很多未列举的情况，在实际操作中要酌情处理。
- 另外，这一步骤在之后的数据分析建模过程中有可能重复。因为即使问题很简单，也并非所有问题都能够一次找出，能做的是使用工具和方法，尽量减少问题出现的可能性，使分析过程更为高效。

3 数据清洗 —错误发现与修复

三、非需求数据清洗

通俗的说，就是把不要的字段删除。但实际操作起来，有很多问题，如:

- 可能删除看上去不需要但实际上对业务很重要的字段;
- 有时候觉得某个字段有用但又没想好怎么用，无法确定是否该删;

如果数据量没有大到不删字段就没办法处理的程度，那么能不删的字段尽量不删。

- 看走眼了，删错字段等。 勤备份数据

三、非需求数据清洗

解决办法

- 在出现错误的数据上增加错误标记;
- 针对存在的错误值，设计劣质容忍的数据分析算法以最小化错误对分析结果的影响。

总结：大数据的预处理

1 数据抽样和过滤

※
1. 随机抽样
2. 系统抽样
3. 分层抽样
4. 整群抽样

2 数据标准化与归一化

※
1. 0-1标准化
2. Z-score标准化
3. Log函数转换

3 数据清洗

※ 1. 缺失值填充
※ 2. 实体识别与真值发现
3. 错误发现与修复

习题和讨论

讨论1：有时两个元组指代的是同一实体，但其相似度却低于指向不同实体的元组?这是为什么?请举例说明。

例 8-1　表 8-2 中有 7 条元组，它们是 7 个名叫“wei wang”的论文作者。通过访问作者的个人主页，我们手动地将这 7 条元组分成三类。*id* 为 *o*11、*o*12 和 *o*13 的元组指代一个 UNC 大学的作者实体，记为 *e*1，*id* 为 *o*21 和 *o*22 的元组指代 UNSW 大学的一个作者实体，记为 *e*2，*id* 为 *o*31 和 *o*32 的元组指代复旦大学的一个作者实体，记为 *e*3。

表 8-2　论文作者元组

id	*name*	*coauthors*	*title*	*class*
*o*11	wei wang	Zhang	inferring...	*e*1
*o*12	wei wang	duncan, kum, pei	social...	*e*1
*o*13	wei wang	cheng, li, kum	measuring...	*e*1
*o*21	wei wang	lin, pei	threshold...	*e*2
*o*22	wei wang	lin, hua, pei	ranking...	*e*2
*o*31	wei wang	shi, zhang	picturebook...	*e*3
*o*32	wei wang	pei, shi, xu	utility...	*e*3

习题和讨论

讨论1：

因此，对于任意两个元组X和Y，用X和Y在属性***coauthors***上的相似度作为X和Y的相似度，记为$Sim(X, Y)$。由于Jaccard相似性测度常常被用来测量集合的相似度，因此可以定义两个元组X和Y的相似度为：

$$Sim(X, Y) = \frac{|coauthor(X) \cap coauthor(Y)|}{|coauthor(X) \cup coauthor(Y)|}$$

习题和讨论

讨论1：

由于$Sim(o11, o12)=0$且$Sim(o11, o31)=1/2$，所以$Sim(o11, o12) < Sim(o11, o31)$;

由于$Sim(o12, o13)=1/5$且$Sim(o12, o21)=1/4$，所以$Sim(o12, o13) < Sim(o12, o21)$。

表 8-2　论文作者元组

id	*name*	*coauthors*	*title*	*class*
*o*11	wei wang	Zhang	inferring...	*e*1
*o*12	wei wang	duncan, kum, pei	social...	*e*1
*o*13	wei wang	cheng, li, kum	measuring...	*e*1
*o*21	wei wang	lin, pei	threshold...	*e*2
*o*22	wei wang	lin, hua, pei	ranking...	*e*2
*o*31	wei wang	shi, zhang	picturebook...	*e*3
*o*32	wei wang	pei, shi, xu	utility...	*e*3

习题和讨论

讨论2： AbeBooks.com 上查询ISBN 为1555582184 的书《Digital Visual Fortran Programmer‘s Guide》的作者，返回12 个卖这本书的网上书店，其中7个认为Michael Etzel是书的作者，4个认为Michael Etzel 和Karen Dickinson两个人是书的作者，1个认为Karen Dickinson 是书的作者。

(1)利用简单投票的方式，会得出怎样的回答?

少数服从多数，所以Michael Etzel是作者。

习题和讨论

讨论2：

（2）书的作者是Michael Etzel 和Karen Dickinson 两个人，请查阅资料， 如何更好地从大量的冲突数据中找出真值?

考虑数据源精度的迭代方法，定义一系列的启发式规则，数据可信度和事实上的可信度是相互确定的，所以就可以使用迭代的方法来计算。当计算达到稳定状态时就停止。

习题和讨论

讨论3：

10. 表 8-4 为公司统计的培训表信息：

表 8-4　题 10 用表

姓名	出生日期	联系电话	培训日期	培训内容
Y 元芳	1995.08.29	132××××0569	2017.1.5	Hadoop
李连	1994-04-26	133××××2241	2016.12.27	Spark
周妍	1995.02.16	158××××07468	Hadoop	2017.1.5
李升	1996.2.29	177××××3152	2016.12.27	Spark
赵忠	1993.11.11		2017.1.13	HDFS

（1）表 8-4 中的数据有哪些错误？该如何修复？

第一行姓名属性应该改成汉字；第三行联系电话位数不对，培训日期应为数字，培训内容不应该是日期；第五行联系电话不能为空。