

数据挖掘之本——数据

张冬松

信阳学院
大数据与人工智能学院

dszhang@nudt.edu.cn



多选题

以下属于数据挖掘的是：

- ☐ A 从电话本中查找电话号码
- ☐ B 从搜索引擎中搜索“数据挖掘”相关信息
- ☐ C 发现某些名字在中国某些地区更流行
- ☐ D 根据内容对搜索引擎返回的文档进行聚类

提交

目录 content



第一节

数据挖掘回顾

第二节

数据类型

第三节

数据质量



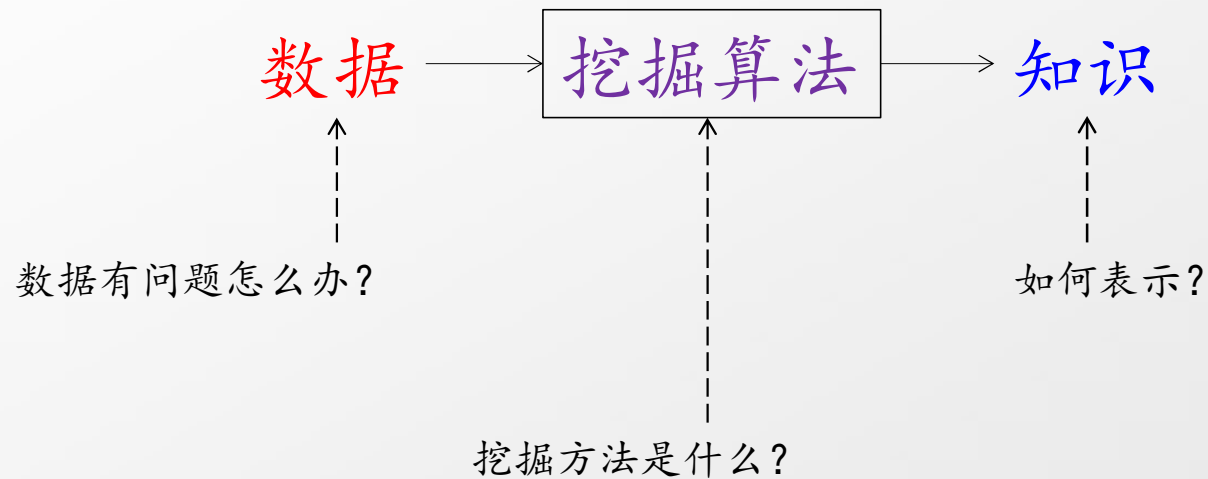
第一节 数据挖掘回顾

- 什么是数据挖掘
- 数据挖掘要解决的问题
- 数据挖掘建模过程



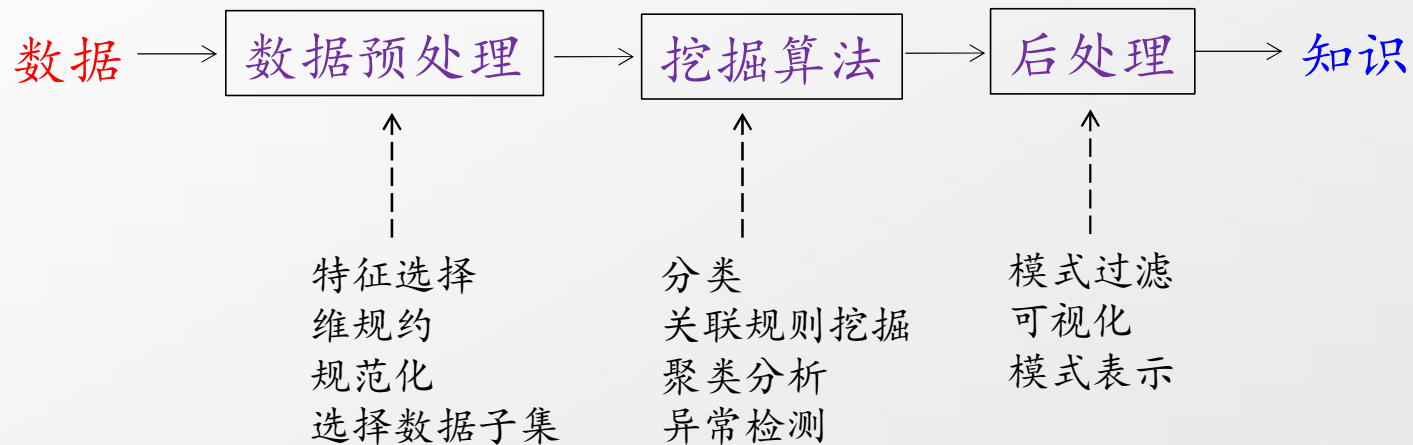
1.1 什么是数据挖掘

- 直观理解：利用计算机从数据中提取知识的过程



1.1 什么是数据挖掘

- **数据挖掘**：探查大型数据库，发现**前所未有的有用**模式或预测未来观测结果。





1.1 什么是数据挖掘

多选题

以下属于数据挖掘的是：

- ☐ A 从电话本中查找电话号码
- ☐ B 从搜索引擎中搜索“数据挖掘”相关信息
- ☒ C 发现某些名字在中国某些地区更流行
- ☒ D 根据内容对搜索引擎返回的文档进行聚类

提交



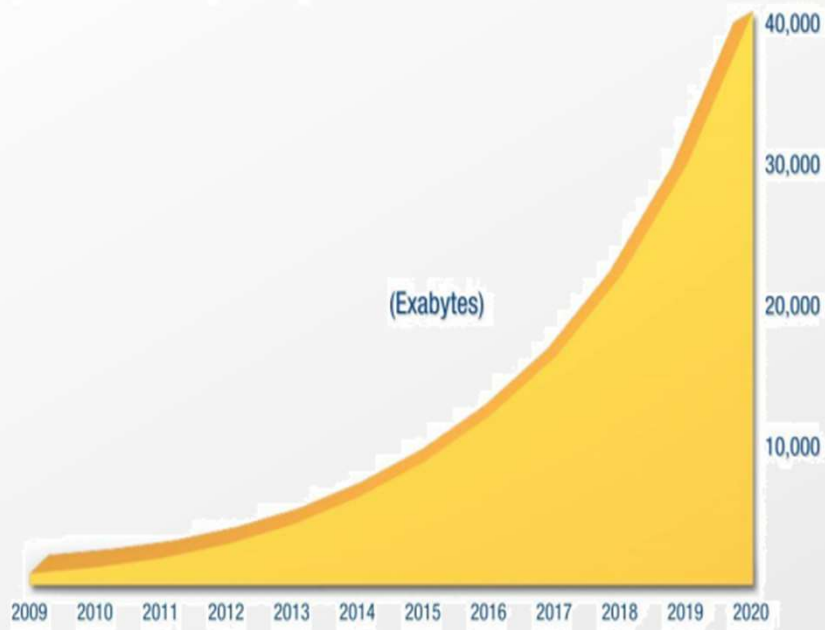
第一节 数据挖掘回顾

- 什么是数据挖掘
- 数据挖掘要解决的问题
- 数据挖掘建模过程



1.2 数据挖掘要解决的问题

A. 大规模数据处理



1KB=1024B

1MB=1024KB

1GB=1024MB

1TB=1024GB

1PB=1024TB

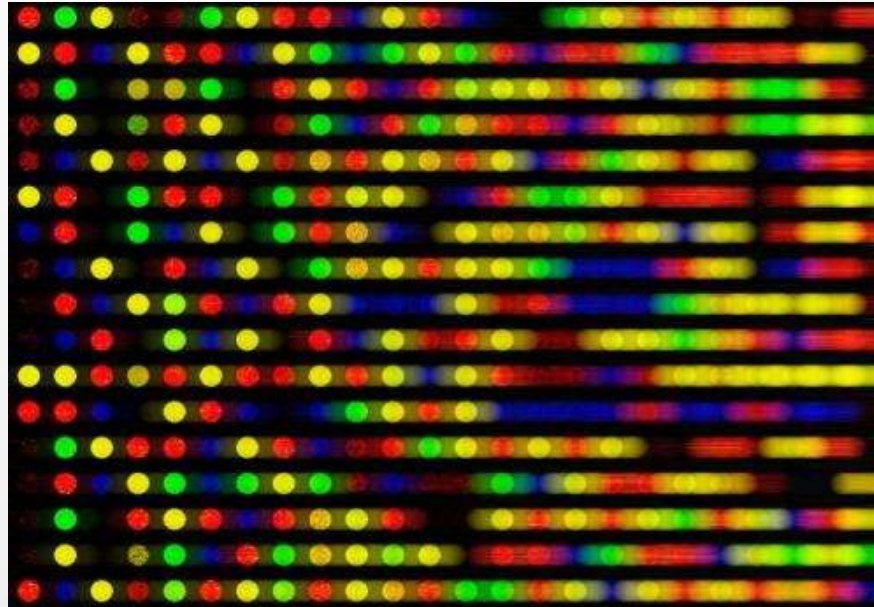
1EB=1024PB

1ZB=1024EB

1.2 数据挖掘要解决的问题

B.高维性

降低维度对算法复杂性的影响



生物信息学中的涉及数千特征的基因表达数据

1.2 数据挖掘要解决的问题

C. 异种数据和复杂数据

特殊数据的处理

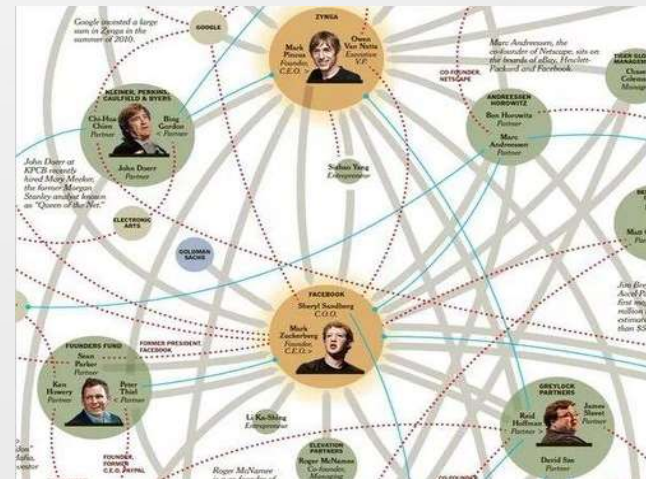
多种类型数据的关联



多模态数据采集



基因数据



社交网络数据



第一节 数据挖掘回顾

- 什么是数据挖掘
- 数据挖掘要解决的问题
- 数据挖掘建模过程



1.3 数据挖掘建模过程





第二节

数据类型

- 属性与度量
- 数据集的类型





2.1 属性与度量

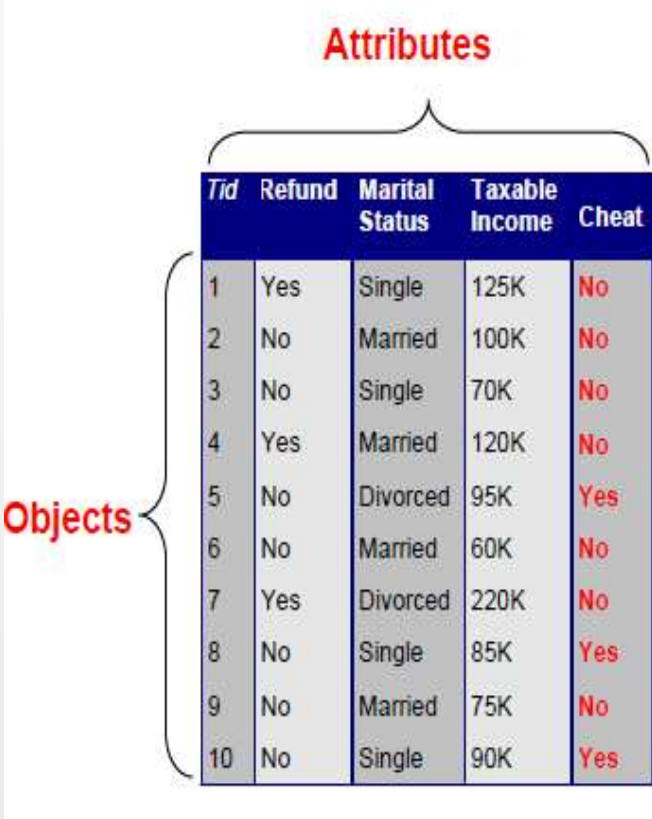
A. 了解数据

1. 数据有哪些列?
2. 每一列的物理意义是什么? 数据单位是什么? 取值范围是什么?
不同的数值含义是什么?
3. 数据质量如何?

2.1 属性与度量

B.数据类型

- 数据集：数据对象(Object)及其属性(Attribute) 的集合。
- 属性：用于刻画对象的性质或特征，因对象而异，随时间而变化
 - 如人的身高、性别等
 - 也称变量、特性、字段、特征或维等
- 数据对象：用一组属性刻画
 - 也称记录、点、向量、案例、样本、观测等



Attributes				
Tid	Refund	Marital Status	Taxable Income	Cheat
1	Yes	Single	125K	No
2	No	Married	100K	No
3	No	Single	70K	No
4	Yes	Married	120K	No
5	No	Divorced	95K	Yes
6	No	Married	60K	No
7	Yes	Divorced	220K	No
8	No	Single	85K	Yes
9	No	Married	75K	No
10	No	Single	90K	Yes



2.1 属性与度量

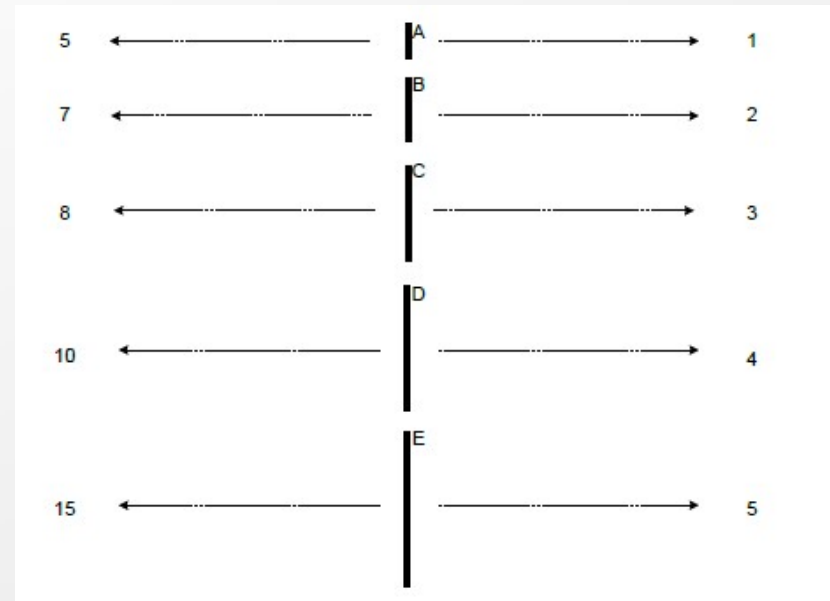
C.属性值

- 属性值：赋予属性的数字或符号
- 属性和属性值的区别
 - 相同的属性可以被赋予不同的属性值
 - 例如：小明和小兰的身高分别是1.7米和1.6米
 - 例如：同样的高度用英尺或米度量时数值不同
 - 不同的属性可以被赋予相同的属性值集合
 - 例如：ID和年龄的属性值范围都是正整数
 - 但是属性值取值各有特点
 - ✓ 例如, ID是没有限制的, 而年龄是最大最小值限制的

2.1 属性与度量

D.属性度量

- **属性度量**：将属性值与对象属性相关联的规则或函数
- 属性可以用一种不描述属性全部性质的方式度量
 - 左边：捕获长度的序性质
 - 右边：捕获长度的序性质和可加性



用不同的方法将线段长度映射到整数



2.1 属性与度量

E.属性的不同类型

■ 属性的不同类型

➤ 分类的或定性的属性

- 不具有数的大部分性质，即使用数表示，也应当像对待符号一样对待
- 例如，性别、邮政编码等
- 包括标称属性、序数属性

➤ 定量的或数值的属性

- 用数表示，并且具有数的大部分属性
- 例如，温度、年龄等
- 包括区间属性、比率属性



2.1 属性与度量

E.属性的不同类型

■ 属性的不同类型

➤ 标称

- 标称值只提供足够的信息以区别对象
- 例如，性别、邮政编码等

➤ 序数

- 序数值提供足够的信息确定对象的序
- 例如，矿石硬度{好、较好、最好}



2.1 属性与度量

E.属性的不同类型

■ 属性的不同类型

➤ 区间

- 值之间的差是有意义的，即存在测量单位
- 例如，日历日期

➤ 比率

- 值之间的差和比率都有意义
- 例如，货币量、年龄、质量等



2.1 属性与度量

E. 属性的不同类型

- 属性的类型取决于是否能够进行如下操作：
 - 相异性：=和 \neq
 - 序：<、 \leq 、 \geq 和>
 - 加法：+和-
 - 乘法：*和/

标称：相异性

序数：相异性、序

区间：相异性、序、加法

比率：相异性、序、加法、乘法



2.1 属性与度量

F.允许的类型变换

属性 类型	变换	注释
标称	任何一对一的变换	如果所有雇员的ID号都重新赋值,不会出现任何不同
序列	值得保序变换, 即新值= f (旧值) 其中 f 是单调函数	包括好、不好、最好的属性可以完全等价的用值{1,2,3}或用{0.5,1,10}表示
区间	新值= a *旧值+ b 其中 a 、 b 是常数	华氏和摄氏温度的零度的位置不同, 1度的大小 (即单位长度) 也不同
比率	新值= a *旧值	长度可以用米或英尺度量



第二节

数据类型

- 属性与度量
- 数据集的类型





2.2 数据集的类型

■ 记录数据

- 数据矩阵
- 事务数据

■ 基于图形的数据

- 带有对象之间联系的数据，如万维网
- 具有图形对象的数据，如蛋白质分子数据

■ 有序数据

- 空间数据、时序数据、序列数据、时间序列数据



2.2 数据集的类型

A. 记录数据

- 数据由有固定属性集合的记录组成

<i>Tid</i>	Refund	Marital Status	Taxable Income	Cheat
1	Yes	Single	125K	No
2	No	Married	100K	No
3	No	Single	70K	No
4	Yes	Married	120K	No
5	No	Divorced	95K	Yes
6	No	Married	60K	No
7	Yes	Divorced	220K	No
8	No	Single	85K	Yes
9	No	Married	75K	No
10	No	Single	90K	Yes

2.2 数据集的类型

B. 文档集合

- 每个文档表示成一个关于词的向量

	team	coach	play	ball	score	game	win	lost	timeout	season
Document 1	3	0	5	0	2	6	0	2	0	2
Document 2	0	7	0	2	1	0	0	3	0	0
Document 3	0	1	0	0	1	2	2	0	3	0



2.2 数据集的类型

C.事务数据

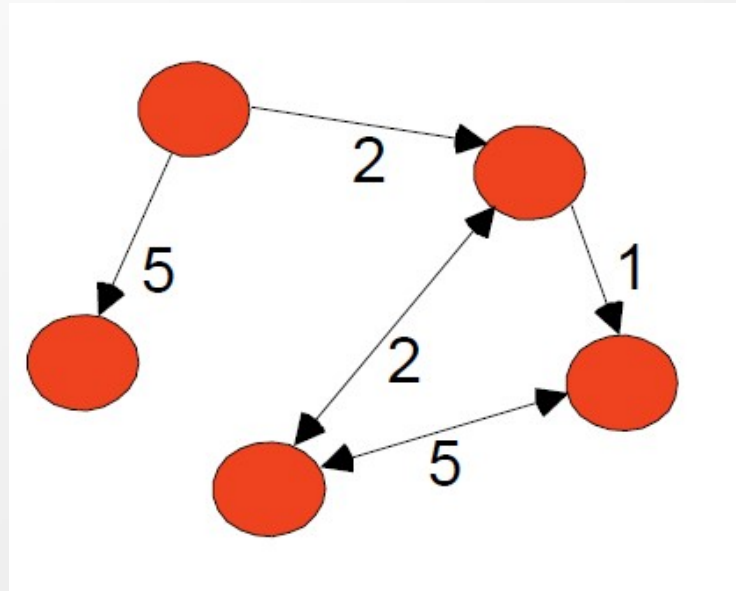
- 每条记录（事务）由一系列“项”构成

<i>TID</i>	<i>Items</i>
1	Bread, Coke, Milk
2	Beer, Bread
3	Beer, Coke, Diaper, Milk
4	Beer, Bread, Diaper, Milk
5	Coke, Diaper, Milk

2.2 数据集的类型

D. 基于图形的数据

- 节点代表数据对象，边代表连接，数字代表链接权重





2.2 数据集的类型

E.时序数据

- 每个记录包含一个与之关联的时间，记录之间通过时间产生关联

时间	顾客	购买的商品
t1	C1	A,B
t2	C3	A,C
t3	C1	C,D
t2	C2	A,D
t4	C2	E
t5	C1	A,E

顾客	购买时间与购买商品
C1	(t1:A,B) (t2:C,D) (t5:A,E)
C2	(t3:A,D) (t4:E)
C3	(t2:A,C)



2.2 数据集的类型

F. 序列数据

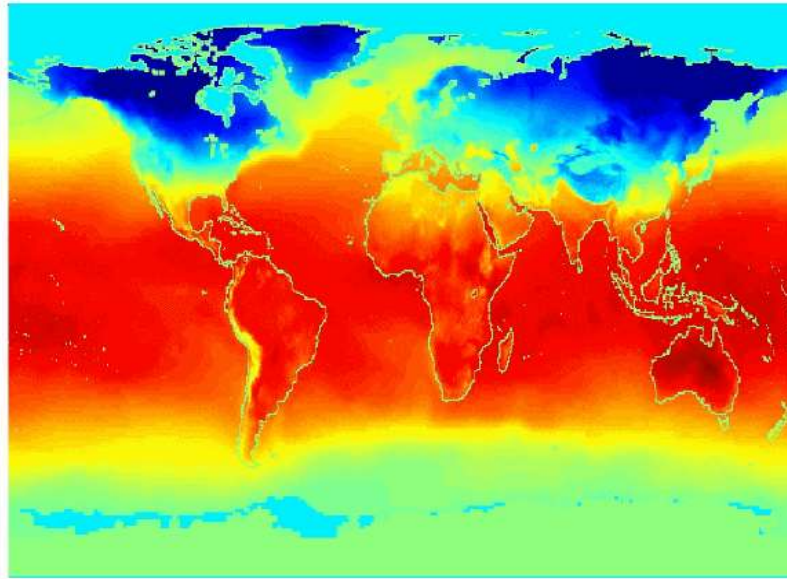
- 一条记录是各个实体的序列，如词或字母的序列

```
GGTTCCGCCTTCAGCCCCGCGCC  
CGCAGGGCCCGCCCCGCGCCGTC  
GAGAAGGGCCCGCCTGGCGGGCG  
GGGGGAGGCGGGGCCGCCCGAGC  
CCAACCGAGTCCGACCAGGTGCC  
CCCTCTGCTCGGCCTAGACCTGA  
GCTCATTAGGCGGCAGCGGACAG  
GCCAAGTAGAACACGCGAAGCGC  
TGGGCTGCCTGCTGCGACCAGGG
```

2.2 数据集的类型

G.空间数据

- 数据对象带有空间属性



第三节 数据质量

- 数据质量问题是如何产生
- 有哪些类型的数据质量问题
- 如何发现数据质量问题
- 如何处理这些问题





3.1 测量和数据收集问题

■ 测量误差

- 测量过程中导致的问题
- 例如，记录值与实际值不同

■ 数据收集错误

- 诸如遗漏数据或属性值，或不当地包含了其他数据对象等错误
- 例如，大熊猫的数据中混入了小熊猫的数据



3.1 测量和数据收集问题

A. 测量误差

- 在统计学和实验科学中，测量过程和结果数据的质量用**精度**和**偏倚**度量
 - **精度**：（同一个量的）重复测量值之间的接近程度
 - **偏倚**：测量值与被测量之间的系统的变差



3.1 测量和数据收集问题

A. 测量误差

■ 假设基本量的真实值为 x 反复测量，得到不同观测值 $\{x_1, x_2, \dots, x_n\}$ ，并使用均值 \bar{x} 作为实际值的估计

➤ **精度**：值集合的标准差

$$precision = \sqrt{\frac{(x_1 - \bar{x})^2 + (x_2 - \bar{x})^2 + \dots + (x_n - \bar{x})^2}{n - 1}}$$

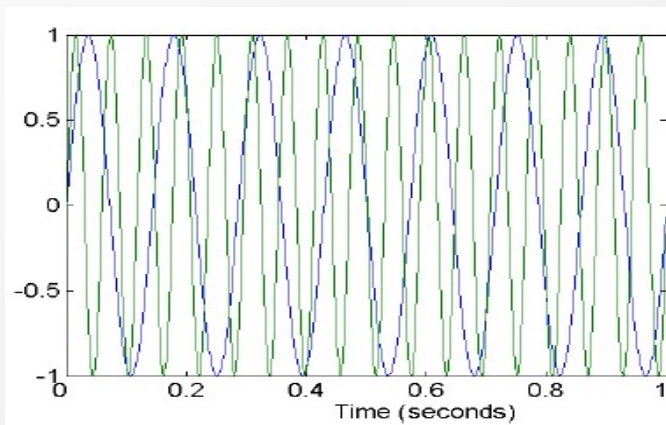
➤ **偏倚**：值集合的均值与真实值的差

$$bias = \bar{x} - x$$

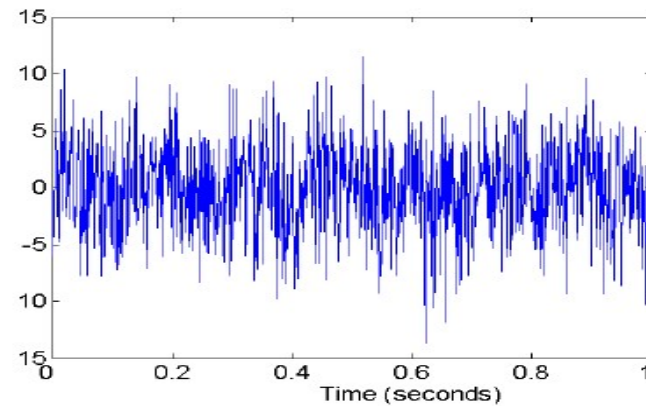
3.1 测量和数据收集问题

A. 测量误差

- 噪声：测量误差的随机部分。
 - 例如，通信质量差引起的人声失真
 - 通常用于包含时间、空间分量的数据
 - 采用信号或图像处理技术，从中发现可能淹没在噪声中的模式



人声数据



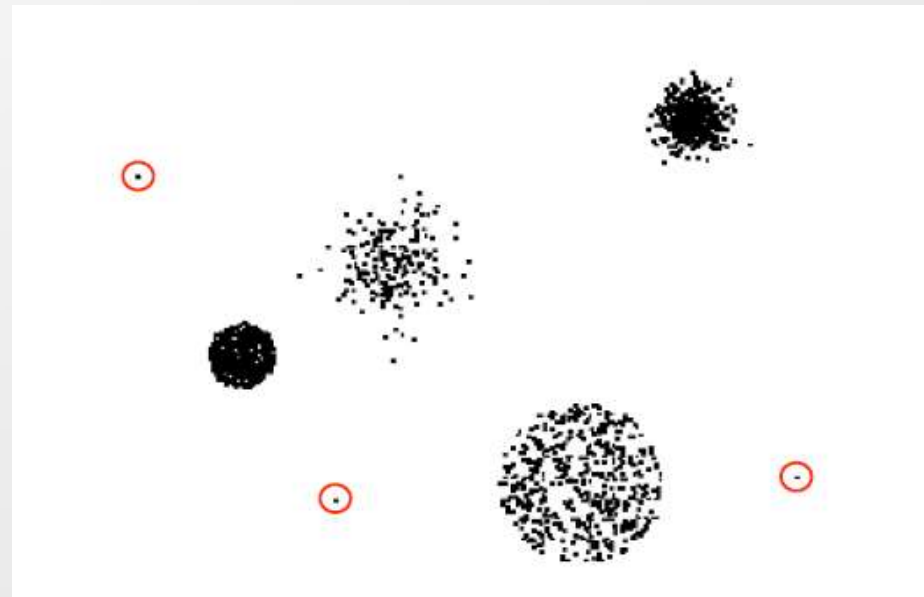
有噪音的人声数据

3.1 测量和数据收集问题

B.数据收集错误

■ 离群点（或异常值）：

- 具有不同于数据集中其它大部分数据对象的特征的数据对象，或是相对于该属性的典型值来说不寻常的属性值
- 可以使合法的数据对象或值





3.1 测量和数据收集问题

B.数据收集错误

■ 异常值检测

- 查看**最大值**和**最小值**，判断这个变量中的数据是不是超出了合理的范围
- 例如，身高的最大值为5米，则该变量的数据存在异常。



3.1 测量和数据收集问题

B.数据收集错误

■ 异常值的处理

异常值处理方法	方法描述
删除含有异常值的记录	直接将含有异常值的记录删除。
视为缺失值	将异常值视为缺失值，利用缺失值处理的方法进行处理。
平均值修正	可用前后两个观测值的平均值修正该异常值。
不处理	直接在具有异常值的数据集上进行挖掘建模。



3.1 测量和数据收集问题

B.数据收集错误

■ 遗漏值

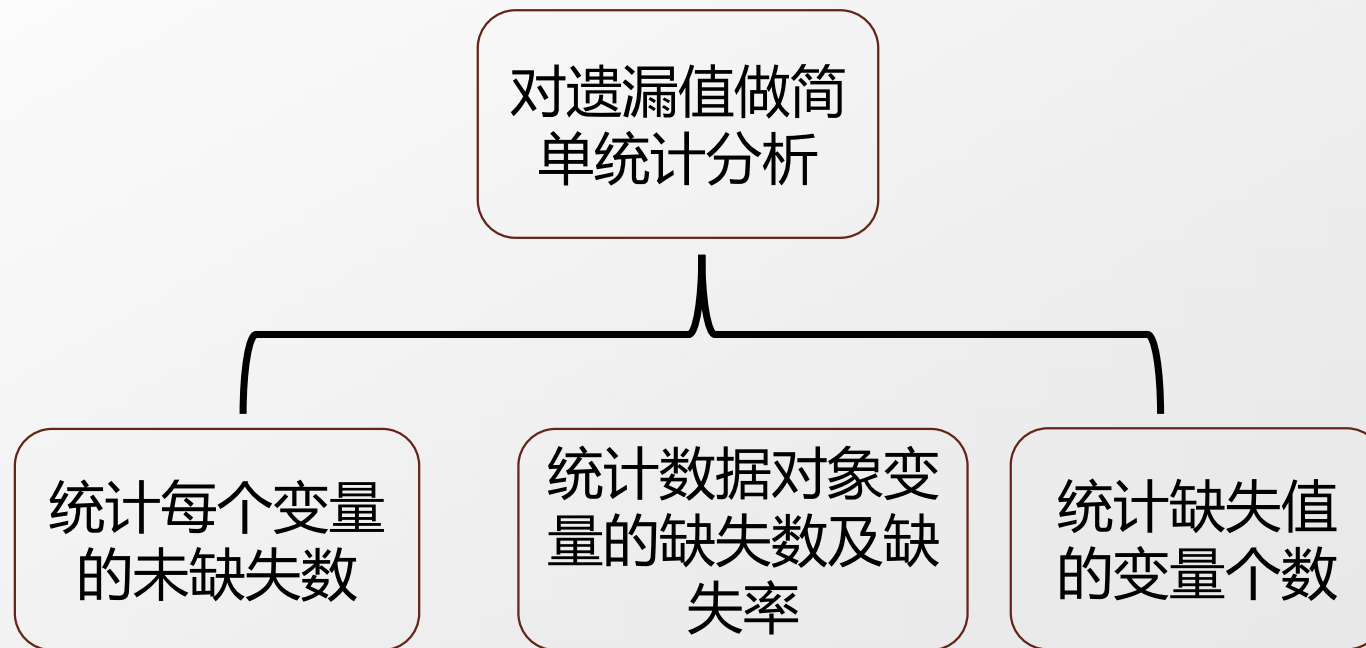
➤ 产生原因

- 信息收集不全，例如，用户拒绝提供年龄和体重
- 属性不适用于所有对象，例如，年收入不适合孩子

3.1 测量和数据收集问题

B.数据收集错误

■ 遗漏值的检测





3.1 测量和数据收集问题

B.数据收集错误

■ 遗漏值

➤ 处理手段

- 删除数据对象或属性
- 估计遗漏值
- 在分析时忽略遗漏值

3.1 测量和数据收集问题

B.数据收集错误

■ 不一致的值

- 与实际不符
- 多来源数据不一致

来源：网站1		单位不一致	来源：网站2	
美国总统	身高		美国总统	身高
特朗普	190	不一致	川普	1.90
奥巴马	185		肯尼迪	1.83
布什	180	不一致	尼克松	1.82
克林顿	188		布什	1.79
.....

3.1 测量和数据收集问题

提问

- 以下数据中存在哪些数据收集导致的质量问题？

安徽	518	河北	525
北京	548	湖南	517
福建	465	湖南	517
新加坡	490	河南	523
广东	508	台湾	-
广西	502	湖北	512
贵州	473	黑龙江	486
海南	602

3.1 测量和数据收集问题

提问

- 以下数据中存在哪些数据收集导致的质量问题？

安徽	518	河北	525	
北京	548	湖南	517	重复值
福建	465	湖南	517	
不一致的值 → 新加坡	490	河南	523	
广东	508	台湾	-	缺失值
广西	502	湖北	512	
贵州	473	黑龙江	486	
异常值 → 海南	602	



第四节 报告题目列表

- 报告题目4
- 报告题目5
- 报告题目6





3.1 报告题目4

4.互联网广告领域中的搜索广告系统分析与应用研究

结合当前互联网企业进行搜索广告系统的进展，分析主流搜索广告系统的设计方案？调研有哪些应用？



3.2 报告题目5

5.医院管理数据仓库系统的分析与设计研究

结合医院日常管理工作，进行医院管理数据仓库系统的分析与设计，列举完整的医院管理数据仓库系统的体系结构，功能组成，使用方式等。



3.3 报告题目6

6.基于Spark的分布式数据处理框架的关键技术与应用研究

结合当前主流的开源分布式框架Spark，分析并研究分布式数据处理框架的关键技术与应用方法，以了解掌握工业界当前大数据应用发展为目标。



张冬松
dszhang@nudt.edu.cn



谢谢! Q&A

THANKS FOR YOUR ATTENTION