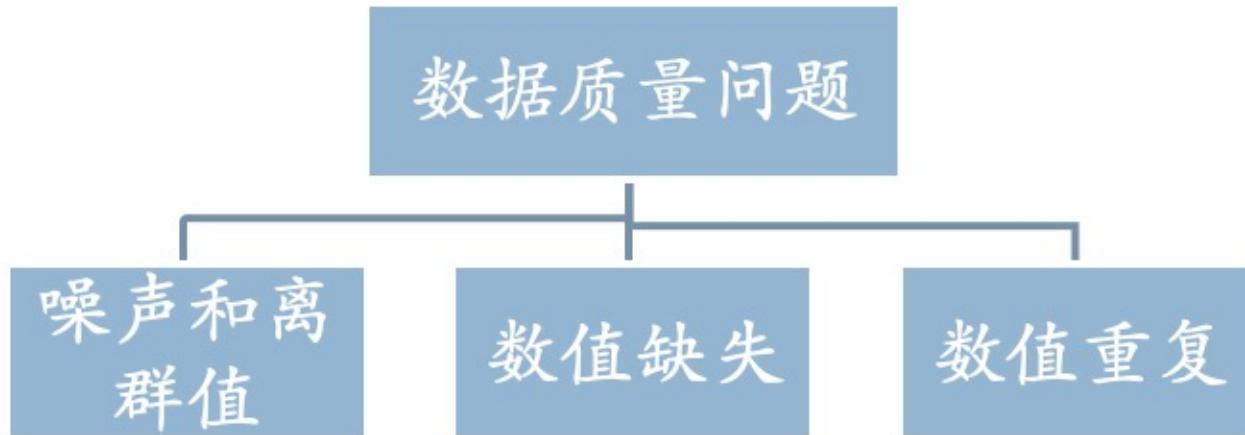


数据清洗（继续）



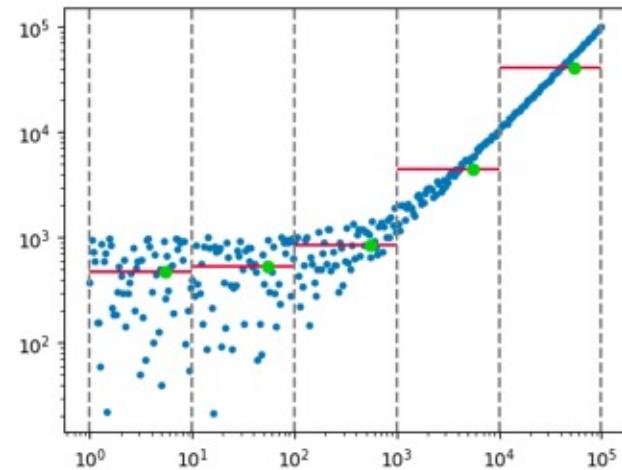
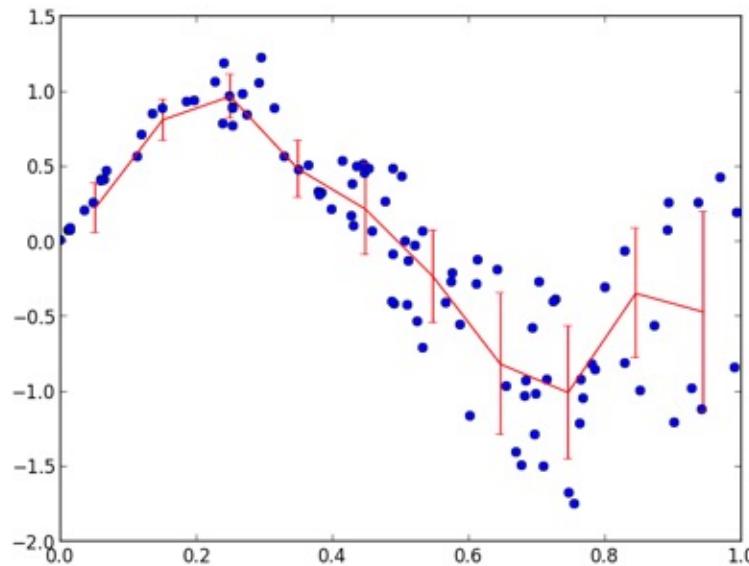
数据清洗

- 未经处理的原始数据中，通常包含大量的无效数据，这些数据在存储之前，就应该被过滤掉。



数据清洗：离群值处理

- 分箱法 (binning)
 - 适用于非正态分布的数据，通过考察相邻数据来确定最终值。
 - 按照属性的值把数据划分成不同的区间，再考察每一个区间中的数据
- 优点
 - 提高模型的稳健性和包容性
 - 还可用于平滑数据，在数据量极大的情况下有助于提高模型的运算效率



数据清洗：缺失值处理

- 缺失值：数据集中某个或某些属性的值是不完全的
 - 如果缺失值过多，说明在数据收集过程中存在着严重的问题。
 - 如果不对缺失值进行处理，会影响数据分析结果的有效性，在后续的数据分析步骤里引发未知风险。
- 出现缺失值的主要原因包括：
 - 信息未被记录
 - 某些属性不适用于所有对象

姓名	出生地	年龄	身高	本科院校	宠物	职业
丁一	北京	24	183	地大	狗	
杜二	河北	26	177	地大	狗	地质工程师
张三		23	180	北大	猫	教师
李四	北京		175	地大		公务员
王五	辽宁	25	168			导游
赵六	北京	26	172	林大		警察
沈七	北京	33		清华	狗	
陈八	山西	29	179			软件工程师
韩九	北京	17	166			高中生

数据清洗：缺失值处理

- 处理数据缺失的方法
 - 删除数据对象

年份	属性1	属性2	属性3
2001	33.42	63.2	46.52
2002	37.69	59.9	45.17
2003	39.4	55.11	42.79
2004	42.52	61.49	43.29
2005	45.31	65.7	45.41
2006	47.29	56.02	48.89
2007	50.02	47.53	
2008	50.33	61.94	50.29
2009	54.13	74.58	56.14
2010	56.38	66.38	55.38
2011	59.82	70.37	56.71
2012	62.1	72.08	54.26
2013	64.23	76.18	55.44
2014	67.83	45.03	
2015	70.21	73.56	58.89
2016	73.55	76.29	59.88
2017	77.81	77.66	58.6
2018	80.72	56.48	46.36
2019	84.35	63.42	
2020	87.11	74.81	59.19
2021	89.95	81.53	61.5
2022	92.8	52.15	45.49

数据清洗：缺失值处理

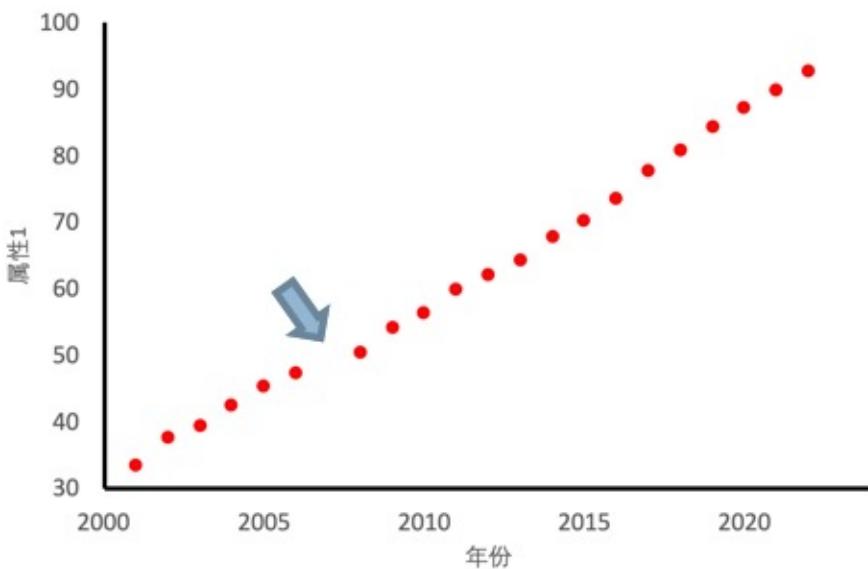
- 处理数据缺失的方法
 - 删除数据对象
 - 忽略缺失值

年份	属性1	属性2	属性3
2001	33.42	63.2	46.52
2002	37.69	59.9	45.17
2003	39.4	55.11	42.79
2004	42.52	61.49	43.29
2005	45.31	65.7	45.41
2006	47.29	56.02	48.89
2007		56.02	47.53
2008	50.33	61.94	50.29
2009	54.13	74.58	56.14
2010	56.38	66.38	55.38
2011	59.82	70.37	56.71
2012	62.1	72.08	54.26
2013	64.23	76.18	55.44
2014	67.83		45.93
2015	70.21	73.56	58.89
2016	73.55	76.29	59.88
2017	77.81	77.66	58.6
2018	80.72	56.48	46.36
2019	84.35	63.42	
2020	87.11	74.81	59.19
2021	89.95	81.53	61.5
2022	92.8	52.15	45.49

数据清洗：缺失值处理

- 处理数据缺失的方法

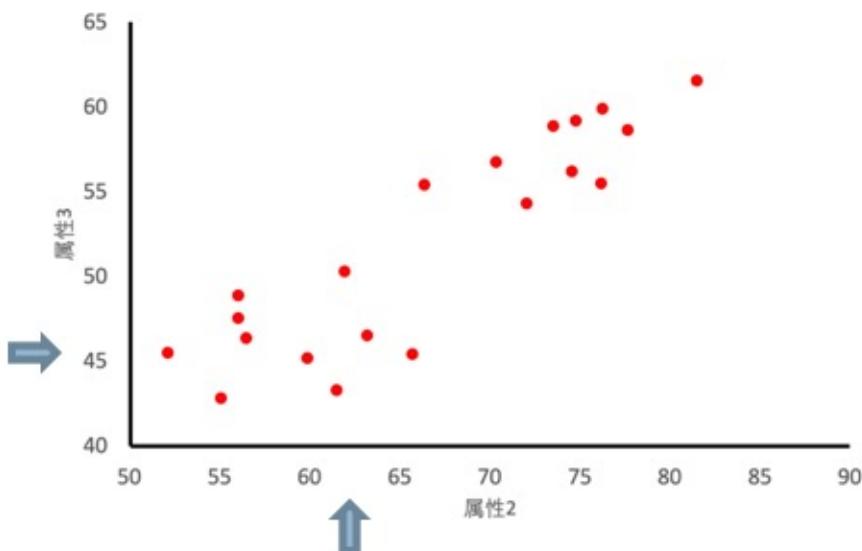
- 删除数据对象
- 忽略缺失值
- 插值法计算缺失值



年份	属性1	属性2	属性3
2001	33.42	63.2	46.52
2002	37.69	59.9	45.17
2003	39.4	55.11	42.79
2004	42.52	61.49	43.29
2005	45.31	65.7	45.41
2006	47.29	56.02	48.89
2007		56.02	47.53
2008	50.33	61.94	50.29
2009	54.13	74.58	56.14
2010	56.38	66.38	55.38
2011	59.82	70.37	56.71
2012	62.1	72.08	54.26
2013	64.23	76.18	55.44
2014	67.83		45.93
2015	70.21	73.56	58.89
2016	73.55	76.29	59.88
2017	77.81	77.66	58.6
2018	80.72	56.48	46.36
2019	84.35	63.42	
2020	87.11	74.81	59.19
2021	89.95	81.53	61.5
2022	92.8	52.15	45.49

数据清洗：缺失值处理

- 处理数据缺失的方法
 - 删除数据对象
 - 忽略缺失值
 - 插值法计算缺失值
 - 用概率模型、聚类算法来估算缺失值



年份	属性1	属性2	属性3
2001	33.42	63.2	46.52
2002	37.69	59.9	45.17
2003	39.4	55.11	42.79
2004	42.52	61.49	43.29
2005	45.31	65.7	45.41
2006	47.29	56.02	48.89
2007		56.02	47.53
2008	50.33	61.94	50.29
2009	54.13	74.58	56.14
2010	56.38	66.38	55.38
2011	59.82	70.37	56.71
2012	62.1	72.08	54.26
2013	64.23	76.18	55.44
2014	67.83		45.93
2015	70.21	73.56	58.89
2016	73.55	76.29	59.88
2017	77.81	77.66	58.6
2018	80.72	56.48	46.36
2019	84.35	63.42	
2020	87.11	74.81	59.19
2021	89.95	81.53	61.5
2022	92.8	52.15	45.49

数据清洗：重复值处理

- 重复数据是指关键字段发生重复的数据，主要来源是不同数据源的合并。
 - 重复数据不一定所有属性都一样
 - 录入时间不同，在此期间属性发生了变化
 - 不同数据源录入的标准不同
 - 录入或合并数据源时出现错误

Date	Transaction No.	Posting	Name	Memo/Desc Account	Split	Amount
2020-08-10 Invoice	1004	Yes	Alpha business	Accounts Receivable (A/R) Sales		\$500.00
2020-08-10 Payment		Yes	Alpha business	Cash and cash equivalents Accounts Rec		\$400.00
2020-08-24 Expense		Yes	Clean oceasn	Cash and cash equivalents Office expens		-\$35.00
2020-10-12 Invoice	1001	Yes	Alpha business	Accounts Receivable (A/R) -Split-		\$1,840.00
2020-10-05 Invoice	1002	Yes	Axiom	Accounts Receivable (A/R) Sales		\$3,000.00
2020-10-12 Payment		Yes	Alpha business	Cash and cash equivalents Accounts Rec		\$840.00
2020-10-12 Expense		Yes	Clean oceasn	Cash and cash equivalents Office expens		-\$51.00
2020-10-12 Expense		Yes	Data driven solutions	Cash and cash equivalents Legal and prc		-\$501.00
2020-10-22 Payment		Yes	Axiom	Cash and cash equivalents Accounts Rec		\$10,023.00
2020-10-22 Payment		Yes	Axiom	Cash and cash equivalents Accounts Rec		\$10,050.00
2020-09-15 Payment		Yes	Axiom	Cash and cash equivalents Accounts Rec		\$10,023.00
2020-10-22 Payment		Yes	Axiom	Cash and cash equivalents Accounts Rec		\$10,023.00
2020-10-12 Expense		Yes	Clean oceasn	Cash and cash equivalents Office expens		-\$51.00

数据清洗：重复值处理

- 处理方式
- 如果数值完全相同：删除重复项
- 如果数值发生了变化
 - 专业人员手动分辨
 - 取平均数、中位数或众数
 - 取最新（身高体重、考试成绩）或最旧（与回忆相关的信息）

	id	first_name	last_name	email
▶	1	Carine	Schmitt	carine.schmitt@verizon.net
	4	Janine	Labrune	janine.labrune@aol.com
	6	Janine	Labrune	janine.labrune@aol.com
	2	Jean	King	jean.king@me.com
	12	Jean	King	jean.king@me.com
	5	Jonas	Bergulfsen	jonas.bergulfsen@mac.com
	10	Julie	Murphy	julie.murphy@yahoo.com
	11	Kwai	Lee	kwai.lee@google.com
	3	Peter	Ferguson	peter.ferguson@google.com
	9	Roland	Keitel	roland.keitel@yahoo.com
	14	Roland	Keitel	roland.keitel@yahoo.com
	7	Susan	Nelson	susan.nelson@comcast.net
	13	Susan	Nelson	susan.nelson@comcast.net
	8	Zbyszek	Piestrzeniewicz	zbyszek.piestrzeniewicz@att.net

分公司	销量	录入年份
北京	170	2020
北京	185	2021
北京	189	2022
河北	101	2021
河北	97	2022
河南	173	2022
山东	111	2022
上海	150	2022
天津	161	2022
浙江	133	2022

数据清洗：目标

准确性

- 数据主题是否合适，离群值是否已经处理

有效性

- 数据来源和收集方式是否可信、是否真实合理
- 80%的数据来源于Nature和Science，20%的数据来源于百度贴吧？

一致性

- 数据的格式、单位、参考基准是否一致
 - 格式：经纬度哪个在前？
 - 单位：千米/英里，摄氏度/华氏度
 - 参考基准：海拔以黄海海面还是世界平均海面为基准？

完整性

- 数据是否仍然有缺失项或重复项

时效性

- 数据的适用范围是否和任务相符
 - 研究2022年的世界经济，是否混入了2012年的数据？

Pop Quiz

回答正确得2分
回答错误得1分
缺席不得分

1. 以下哪一项**不属于**大数据的“5V”基本特征之一?

- A. Value | B. Veracity | C. Viscosity | D. Volume

2. 以下关于某人的属性中，哪一项是**连续属性**?

- A. 体温 | B. 国籍 | C. 学历 | D. 头发颜色



3. 数据记录集包括三种主要形式，以下哪一项**不是**其中之一?

- A. 矢量化文档 | B. 栅格化文档 | C. 数据矩阵 | D. 事务处理数据

4. 以下哪一项是来源于**实验测量或实际观测**的数据?

- A. 有限元模拟 | B. 疾控中心流行病档案 | C. 社交网站上的景区打卡 | D. 古气候模型

5. 正态分布的情况下，**均值±3个标准差**能覆盖_____的数据点。

- A. 68.3% | B. 87.1% | C. 95.5% | D. 99.7%



数据预处理（数据精简）

数据预处理

- 大数据时代，数据常有高维度性、稀疏性和多尺度性。
- 直接对大数据进行可视化通常会产生杂乱无章的结果，这种现象被称为视觉混乱。
- 为了能够在有限的显示空间内，表达大量的复杂数据，我们需要进行数据精简，也叫数据预处理。



数据收集



数据清洗



数据预处理



数据可视化

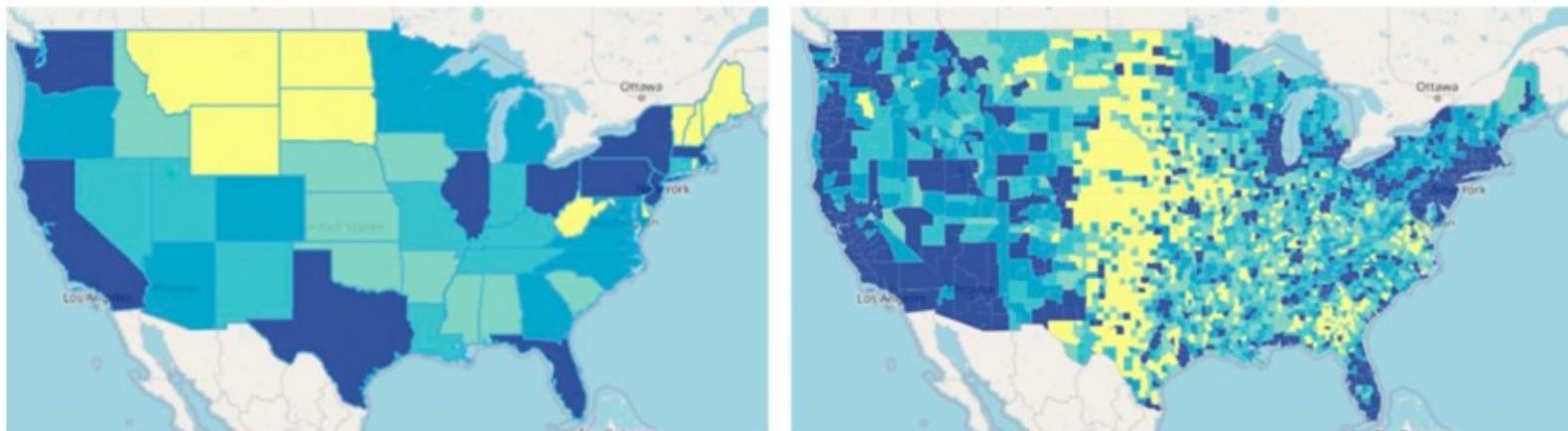
数据预处理

- 数据预处理的意义
 - 数据存储、分析层面
 - 降低数据复杂度，减少数据点数目，并同时保留数据中的基本特征，让数据更直观
 - 减少查询和处理时的资源开销，提高查询的响应性能
 - 在数据仓库或联机分析处理系统应用中
 - 提升复杂数据查询和管理的交互性
 - 当数据量极大时，如果分析和推理只需要定性的结果，可以采用近似解，从而提高效率。



数据预处理：合并法

- 合并法
 - 将两个以上的属性合并为同一个
 - 举例：将县级数据合并到省级；将多个用户的数据合并成一个数据点
- 优点：
 - 节省数据分析的时间成本、储存成本
 - 保护隐私、敏感数据脱密



数据预处理：抽样法

- 抽样法（采样法）在对数据的初步探索和最后的数据分析环节经常被采用。
- 使用原因：获取并处理全部数据集的成本太高。

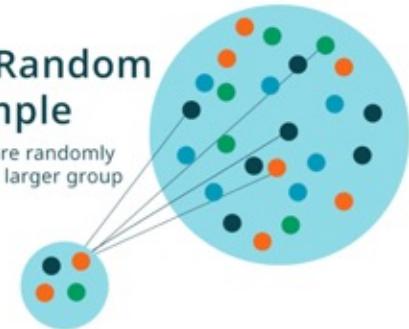


- 如果抽样结果大致具备原始数据的特征，那么这个抽样是具有代表性的。
- 随机抽样
 - 简单随机抽样：随机从数据集中选择数据项。
 - 分层随机抽样：先将数据集分为多份，然后在每份中随机采样。

GeoPoll

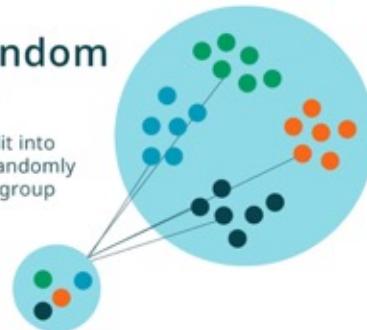
Simple Random Sample

Respondents are randomly selected from a larger group



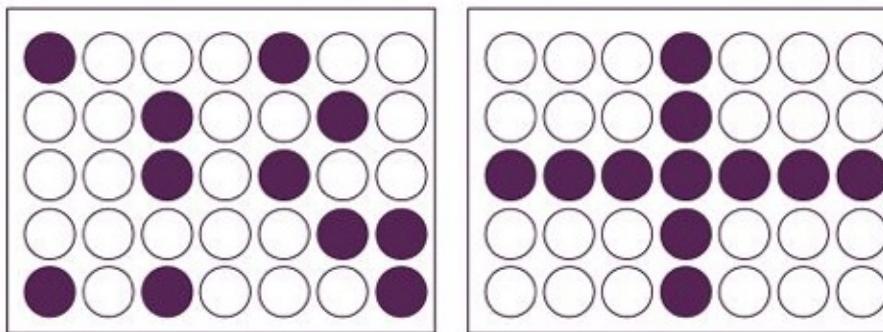
Stratified Random Sample

Respondents are split into sub-groups and then randomly selected from each group



数据预处理：抽样法

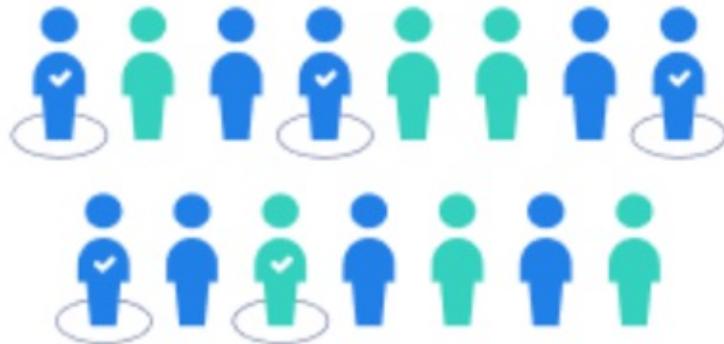
- 概率抽样 Probability Sampling
 - 依据概率论原理，按照随机化原则，从总体中抽取样本的方法。
 - 优点：抽取的样本具有一定的代表性，可以通过样本推断总体特征。
 - 缺点：操作复杂、成本较高。
- 非概率抽样 Non-Probability Sampling
 - 主要依据研究者的主观意愿和判断，从总体中抽取样本的方法。
 - 优点：快速、简易、廉价。
 - 缺点：所抽取的样本代表性较差，一般只会用于探索性分析。



Probability Sampling Vs Non-Probability Sampling

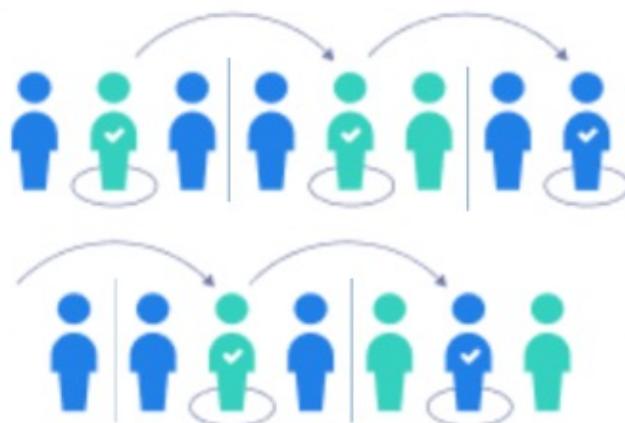
数据预处理：概率抽样

- 单纯随机抽样（简单抽签法）
 - 方法：从N个样本里不重复地随机抽取n个
 - 属于等概率抽样方法，过程是逐个进行抽取，而且样本不放回。
 - 所产生的样本具有代表性，因为所有个体被抽到的机会相等。
- 优点：抽样方法简单、易行。
- 缺点：当总体例数较多时，实施抽样比较困难，有时难以实现。



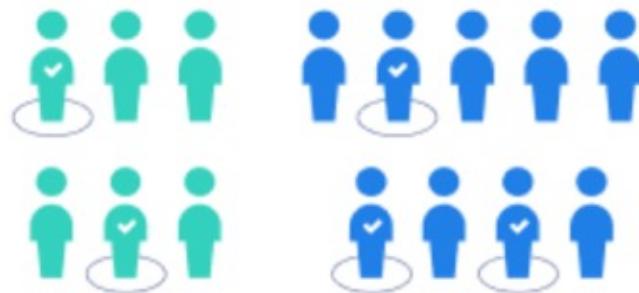
数据预处理：概率抽样

- 系统抽样（也叫等距抽样、机械抽样）
 - 将容量为N的总体，按某一顺序编号，并平均分成n个部分，每部分包含K个个体 ($K=N/n$)。
 - 从第一部分中随机抽取一个个体，依次用相等的间隔，机械地从每一部分中各抽取一个个体，共抽得n个个体组成样本。
 - 每个个体被抽到的可能性相等。
- 优点：操作简单，易得到一个按比例分配的样本，而且抽样误差比单纯随机抽样小，因为覆盖相对均匀
- 缺点：当总体内的数据有周期性，或为单调递增或递减时，将会产生明显的偏差



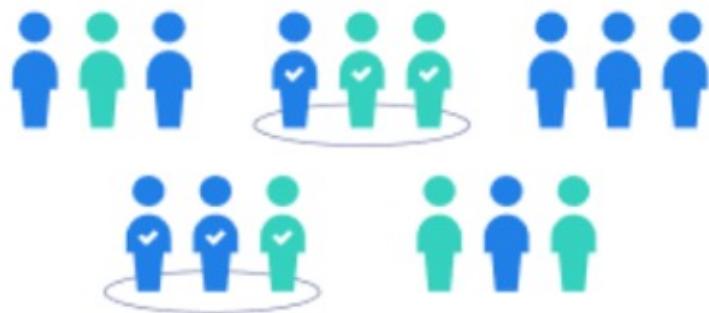
数据预处理：概率抽样

- 分层抽样
 - 将总体按某种特征分成若干个组，统计学上称为“层”(stratum/strata)。
 - 在各层里随机抽样，组成总体样本。
- 每一层抽取的数量
 - 按比例分配抽样：根据每一层内数据点的多少来决定。
 - 最优分配抽样：各层根据内部差异的多少采用不同的比例抽样，由各层内部的数据点数量及标准差共同决定。
- 优点：所得样本代表性较好，抽样误差较小。
- 缺点：抽样方法复杂。



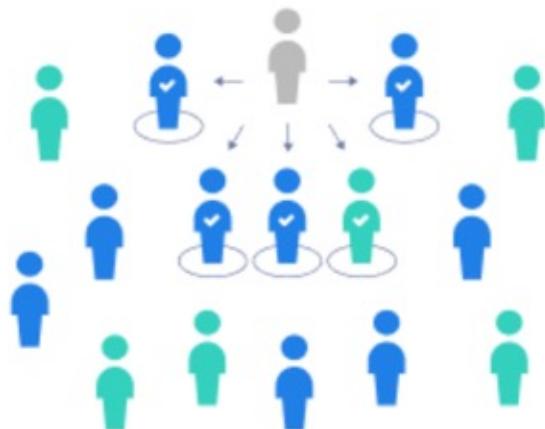
数据预处理：概率抽样

- 整群抽样
 - 先将总体按某种与研究目的无关的客观特征（如班级、城市等）划分为K个群，再从K个群中随机抽取n个群作为研究对象。
- 优点：在大规模调查中，整群抽样易于组织，容易控制调查质量。
- 缺点：各群组间差异可能较大，所以抽样误差较大。



数据预处理：非概率抽样

- 偶遇抽样（方便抽样）
 - 调查者于特定时间、特定地点随意选择个体（回答者）



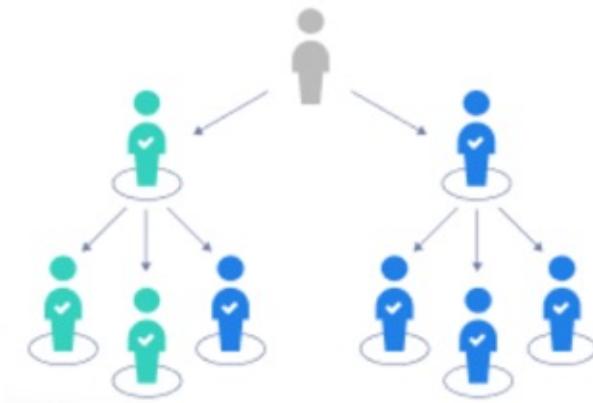
数据预处理：非概率抽样

- 定额抽样（配额抽样）
 - 调查者将调查总体样本按一定标准分类或分层，然后确定各类（层）单位的样本数额，并在配额内任意抽选样本
- 和分层抽样比较：分层抽样要考虑每层的数据量或内部差异，定额抽样不考虑每层（类）的具体特征。



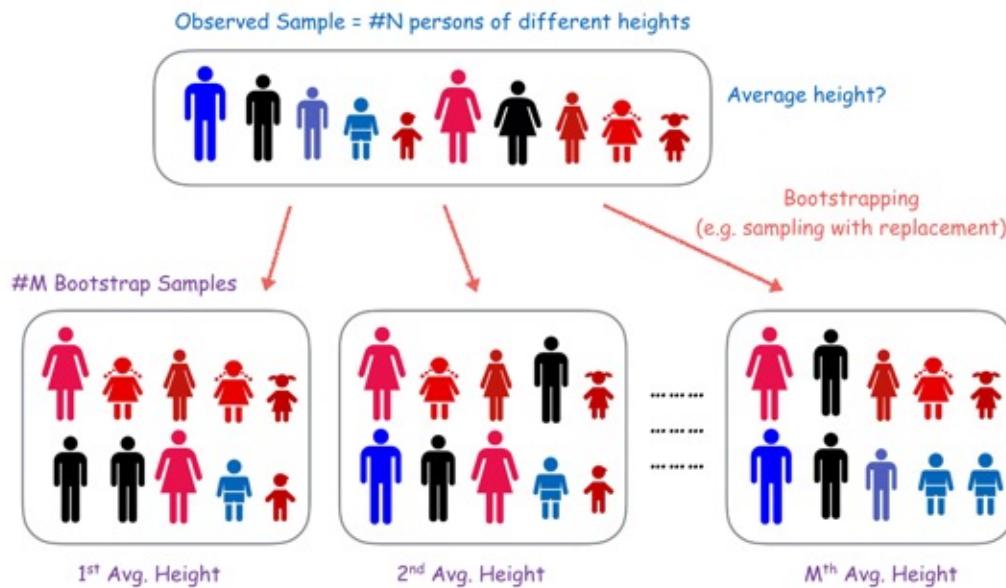
数据预处理：非概率抽样

- 滚雪球抽样（裙带抽样、推荐抽样）
 - 一种在稀疏总体中寻找受访者的抽样方法
 - 让受访者去招募新的受访者，使其成为样本的一部分。
- 优点：容易找到具有某些特征的人群，而这些人可能用常规方法难以找到（例如患有罕见疾病的人）。

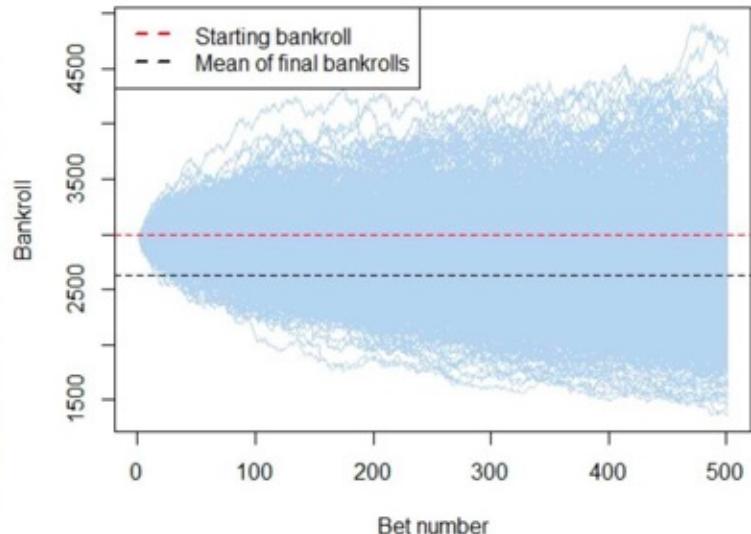


数据预处理：自举抽样

- 自举抽样（自助抽样，Bootstrap）
 - 在训练集中进行有放回地均匀抽样，通常适用于小样本数据集。
 - 被选中的个体允许等可能地被再次选中，并被再次添加到训练集中。
 - 通常需要结合蒙特卡洛模拟（Monte Carlo Simulation）。



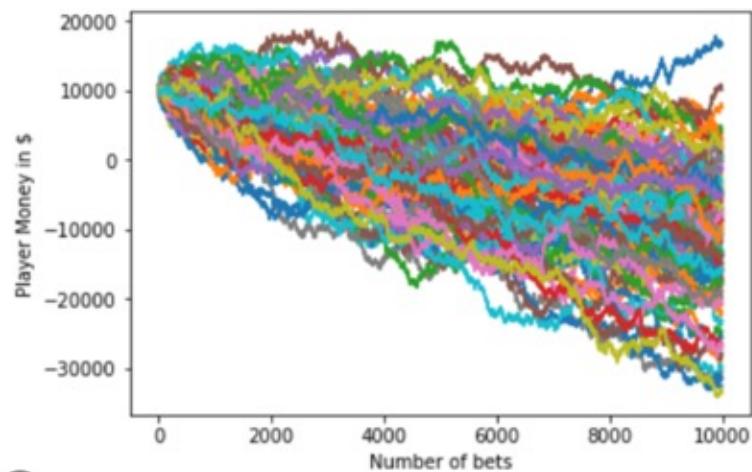
蒙特卡洛模拟



冯·诺依曼

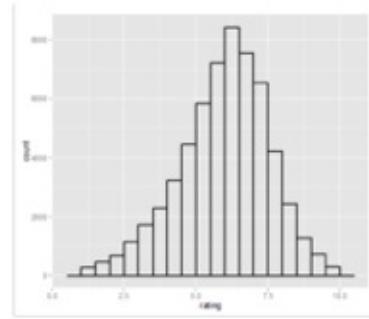
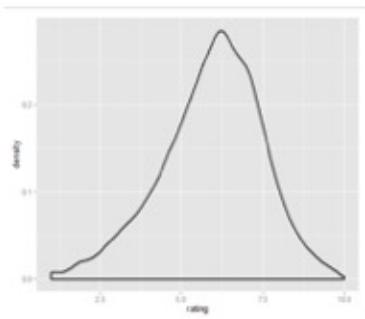


斯坦尼斯拉夫·乌拉姆



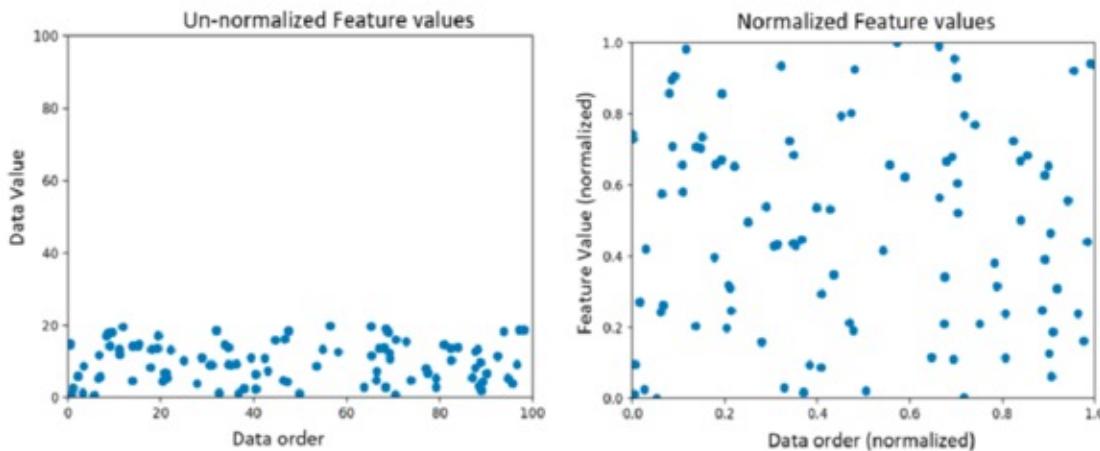
数据预处理：离散化

课程目标	评价标准			
	90-100分	75-89分	60-74分	0-59分
	优	良	中/及格	不及格
1.能够综合运用所学学科原理，针对数据科学与大数据技术应用复杂工程问题，结合数据分析与可视化的理论与技术，设计合适的研究方案。	具备数据分析与可视化的理论与技术能力；能用于复杂工程问题的分析。	掌握数据分析与可视化的理论与技术能力；熟悉复杂工程问题分析的基本流程。	掌握数据分析与可视化理论与技术相关文献阅读、分析、总结能力。	不具备数据分析与可视化理论与技术能力；不能用于复杂工程问题的分析。
2. 具备自主学习和终身学习的意识。	具备自主学习和终身学习的能力。	基本养成了自主学习和终身学习的意识。	了解主动学习专业相关知识的习惯，并部分养成了学习意识。	不具备自主学习和终身学习的意识。



数据预处理：归一化

- 把数据映射到 $[0, 1]$ 区间上。
- 归一化是特殊的缩放法（min-max scaling）。





Introduction to Data Analysis

数据分析基础

中国地质大学（北京）信息工程学院

叶山

yes@cugb.edu.cn

本节内容

- 数据的组织和管理
- 数据分析

数据的组织和管理



数据组织和管理

无关和错误数据的产生和引入原因

使用不同的数据源

数据处理阶段
的不当操作

计算资源的限
制



数据组织：采用自动控制，使数
据能够进行有效的组织、存储。

数据组织和管理

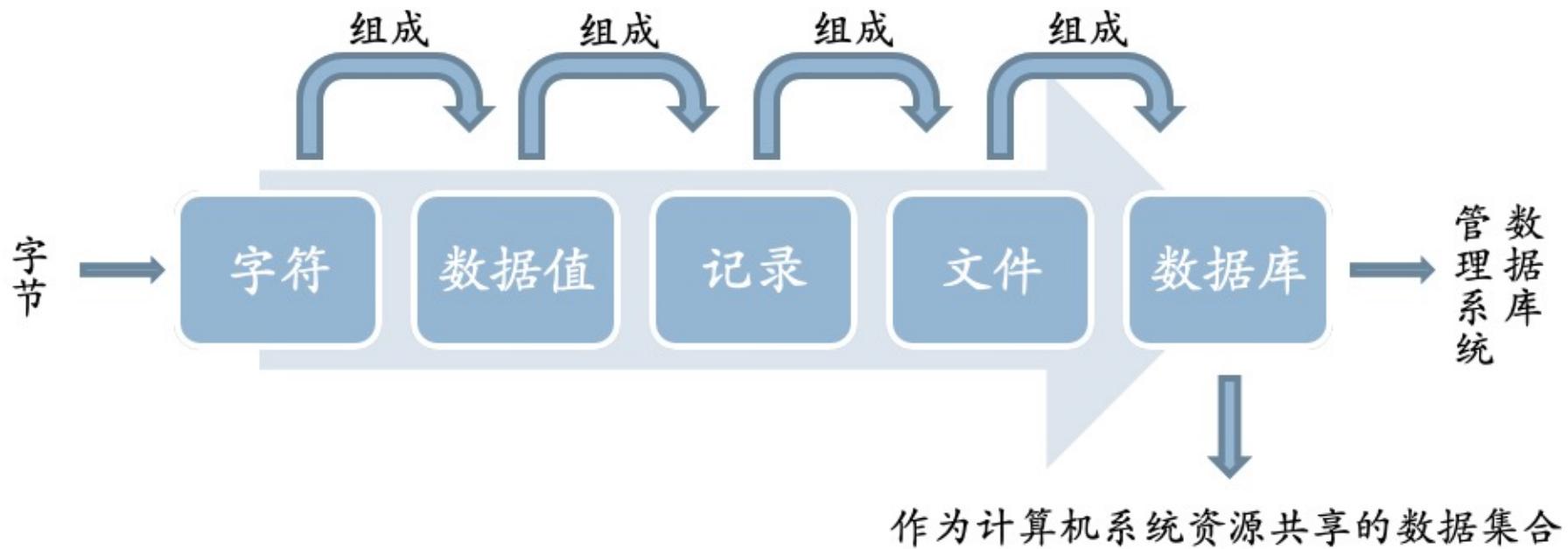
- 数据管理：对数据进行有效的收集、存储、处理和应用。
- 面向应用的数据管理对象



ETL：提取、转换、加载

数据组织和管理

数据组织的层次体系



数据组织和管理

数据可视化相关的数据组织



数据表单

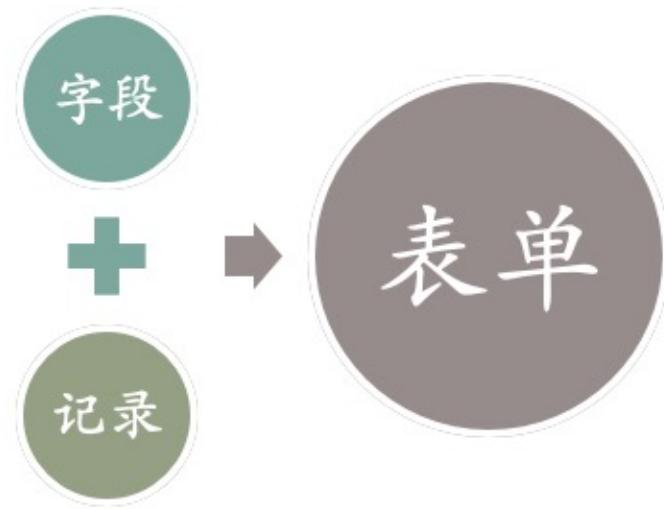
- 表单文件（Spreadsheet）是最简单的数据组织管理形式
 - 由字段（field/attribute）和记录（record）共同组成的数据队列。
- 数据表单的设计可以反映数据分析师的水平
 - 对数据表单的理解是数据分析工作能够进行的基础。
 - 数据表单的设计是否合理关系着后期数据分析的效率及效果。

	A	B	C	D	E	F	G	H	I	J
1	Sl #	Vendor Name	Vendor Address	Country	Project Code	Sale Date	Sale Value	Invoice Date	Invoice #	Due Date
2	1	Philips	x	India	P154	6/16/2019	65629	6/20/2019	P154_21	8/6/2019
3	2	Symphony	x	Polland	P117	5/19/2019	29569	5/25/2019	P117_17	7/8/2019
4	3	LG	x	India	P170	5/15/2019	21195	5/17/2019	P170_22	6/29/2019
5	4	Sain-Go-Bain	x	India	P190	7/15/2019	26845	7/19/2019	P190_22	8/25/2019
6	5	IBM	x	India	P179	7/27/2019	37924	8/1/2019	P179_23	9/13/2019
7	6	Infosys	x	USA	P106	6/29/2019	68237	7/4/2019	P106_23	8/20/2019
8	7	Blue Star	x	USA	P188	6/12/2019	71442	6/14/2019	P188_18	7/27/2019
9	8	Symphony	x	India	P186	7/16/2019	62320	7/20/2019	P186_14	8/25/2019
10	9	IBM	x	India	P169	7/10/2019	49379	7/12/2019	P169_19	8/23/2019
11	10	Infosys	x	Germany	P186	7/7/2019	24658	7/12/2019	P186_11	8/20/2019

数据表单

字段也叫属性，是事物或现象的某种特征，在统计学中称为变量。

序号	日期	销量	销售额	销售商品数	动销率
1	2018/9/9	1811	81944.8	55	54.46%
2	2018/9/8	713	30967.2	34	33.66%
3	2018/9/7	186	8892.1	31	30.69%
4	2018/9/6	240	10185.1	22	21.78%
5	2018/9/5	230	9347.2	21	20.79%
6	2018/9/4	467	19602.8	32	31.68%
7	2018/9/3	412	18180.2	34	33.66%



记录也叫数据点、数据值、数据元或变量值，是事物或现象某种特征的具体表现。

数据表单

设计原则

1. 数据表由标题行和数据构成。
2. 第一行为表的列标题（字段），列标题不能重复。
3. 从第二行起是数据部分，每一行数据称为一个记录，原则上不允许出现空自行和空白列。
4. 在数据表中不能出现跨越多行或多列的单元格。
5. 数据表单优先以一维的形式存储，但在实际工作中，所接触到的数据往往是以二维表格的形式存在的，这时应尽量将二维表转换为一维表的形式进行存储。

Date	Start Time	Lunch Start	Lunch End	End Time	Overtime	Hourly Rate (USD)	Total Hours Worked (h)	Amount Earned
4/1/2019	8:00am	1:00pm	1:30pm	5:00pm	0	50	8	400
4/2/2019	8:00am	1:00pm	1:30pm	5:00pm	0	50	8	400
4/3/2019	8:00am	1:00pm	1:30pm	5:00pm	2	50	10	500
4/4/2019	8:00am	1:00pm	1:30pm	5:00pm	0	50	8	400
4/5/2019	8:00am	1:00pm	1:30pm	5:00pm	0	50	8	400
4/6/2019	9:00am	2:00pm	2:30pm	6:00pm	0	50	8	400
4/7/2019	9:00am	2:00pm	2:30pm	6:00pm	0	50	8	400
4/8/2019	8:00am	1:00pm	1:30pm	5:00pm	1	50	9	450
4/9/2019	8:00am	1:00pm	1:30pm	5:00pm	1	50	9	450
4/10/2019	8:00am	1:00pm	1:30pm	5:00pm	1	50	9	450

AL	Montgomery	Huntsville	Dec 14, 1819	5,024,279	52,420	135,767
AK	Juneau	Anchorage	Jan 3, 1959	733,391	665,384	1,723,337
AZ	Phoenix		Feb 14, 1912	7,151,502	113,990	295,234
AR	Little Rock		Jun 15, 1836	3,011,524	53,179	137,732
CA	Sacramento	Los Angeles	Sep 9, 1850	39,538,223	163,695	423,967
CO	Denver		Aug 1, 1876	5,773,714	104,094	269,601

数据表单

- 一维表是指每条数据只有一个对应数值，每一列都是独立参数的数据表。一维表能够容纳更多的数据，让数据更丰富、更详细。

品类	季度	销售额 / 元
面膜	Q1	16800
面膜	Q2	16204
面膜	Q3	19262
面膜	Q4	19502
润肤乳	Q1	19344
润肤乳	Q2	19649
润肤乳	Q3	15496
润肤乳	Q4	15608
保湿霜	Q1	17754
保湿霜	Q2	16224
保湿霜	Q3	16572
保湿霜	Q4	17134
洁面霜	Q1	18360
洁面霜	Q2	18513
洁面霜	Q3	15556
洁面霜	Q4	19060

每个数据值只有一个维度：
要么是季度，要么是销售额

数据表单

- 二维表是指每个数据都有两个对应数值，每一列都是同类参数的数据表。二维表能够更直观地显示数据特点，让数据更直观、更明确

品类	Q1	Q2	Q3	Q4
面膜	16800	16204	19262	19502
润肤乳	19344	19649	15496	15608
保湿霜	17754	16224	16572	17134
洁面霜	18360	18513	15556	19060

每个数据值有两个维度：
季度（Q1到Q4）和销售额

一维表：存
储数据

二维表：展
示数据

数据表单

优点

- 灵活度好，处理数据时的自由度高
- 有齐全的配套软件
- 对用户的技术要求低
- 软件对电子表单文件的变种（CSV、TSV等格式）支持度好

缺点

- 数据冗余率高
- 不易访问调用、查错改错
- 难以添加约束
- 安全性低
- 缺少数据类型和元数据，需要语义解释