数据库系统概论新技术篇

大数据近似算法

魏哲巍

中国人民大学信息学院

2017年4月

大数据近似算法

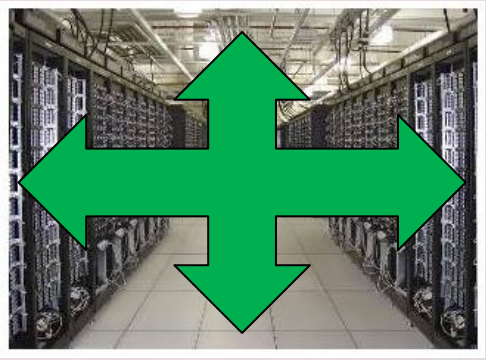
- 研究背景与计算模型
- 随机采样算法
- 基于计数的近似算法
- 基于哈希的近似算法
- 研究成果简介

大数据近似算法

- 研究背景与计算模型
- 随机采样算法
- 基于计数的近似算法
- 基于哈希的近似算法
- 研究成果简介

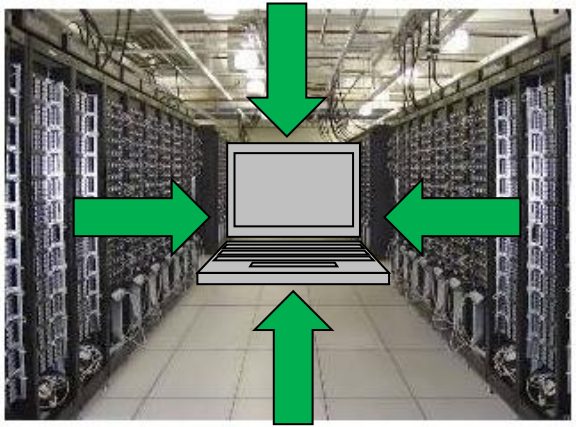
研究背景

- 处理大数据的两种思维模式：
  - 扩展计算能力：。
- 方案一：扩展计算能力
  - 超级计算机、分布式系统。。
  - 问题：成本昂贵、能源消耗



研究背景

- 处理大数据的两种思维模式：
  - 扩展计算能力：。
- 方案一：扩展计算能力
  - 超级计算机、分布式系统。。
  - 问题：成本昂贵、能源消耗
- 方案二：降低数据规模
  - 通过引入近似/允许误差，将大数据变为小数据
  - 优点：成本小，可与方案一结合
  - 缺点：需针对特定问题设计特定算法

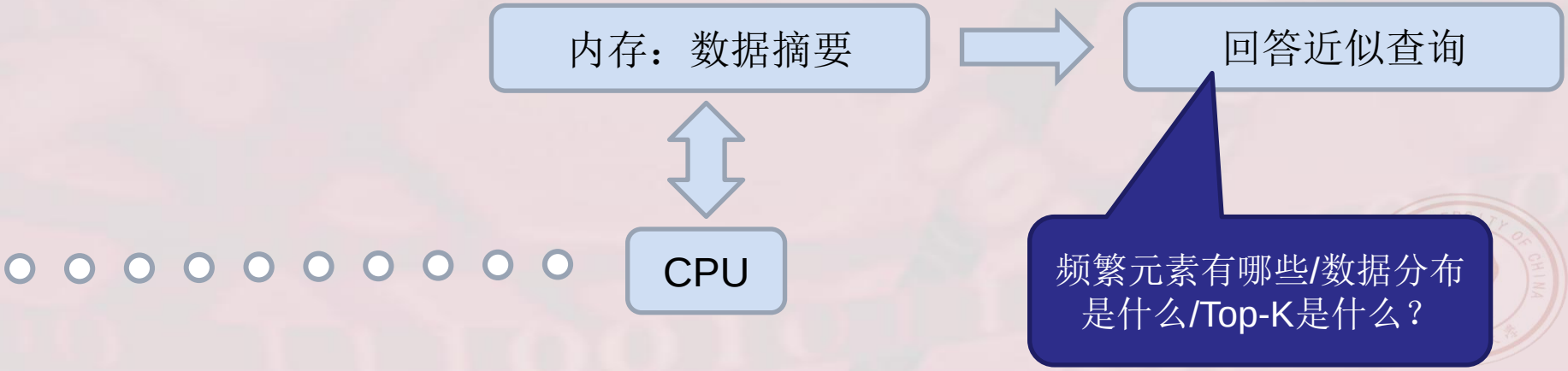


研究背景

- 大数据近似算法：利用采样(sampling)、略图(sketch)、摘要(summary)等技术，引入可控误差，解决由数据规模扩大带来的时间/空间/通讯量效率问题。
  - 大数据通常有冗余，有价值的数据量可能很小；
  - 统计量从宏观上能反映实际问题的特质；
  - 现有的数据采集系统和分析算法也不可避免地会产生误差。

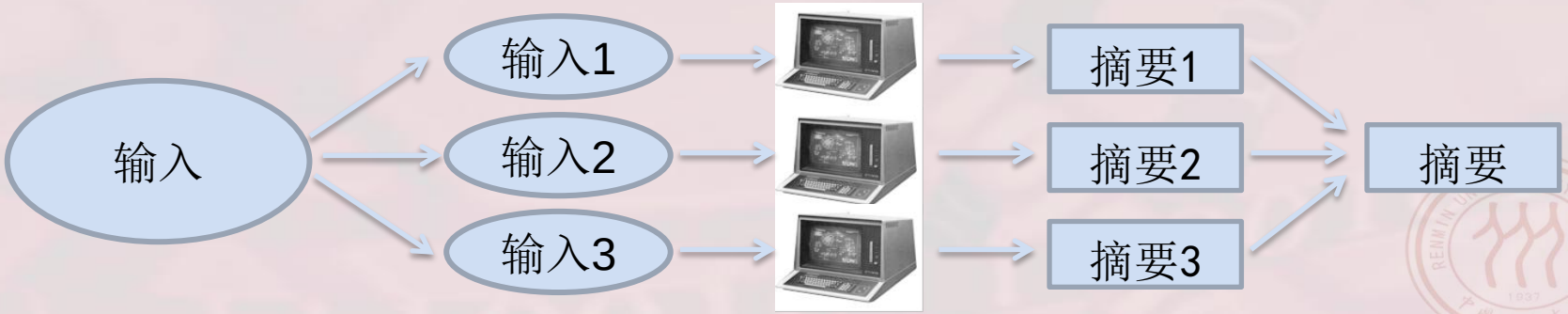
计算模型：数据流模型

- 数据流是一个由海量数据组成的数据序列
  - **Single Pass**: 每个数据最多访问一次
  - **Small Space**: 存储空间非常小
  - **Small time**: 更新(插入删除)速度快



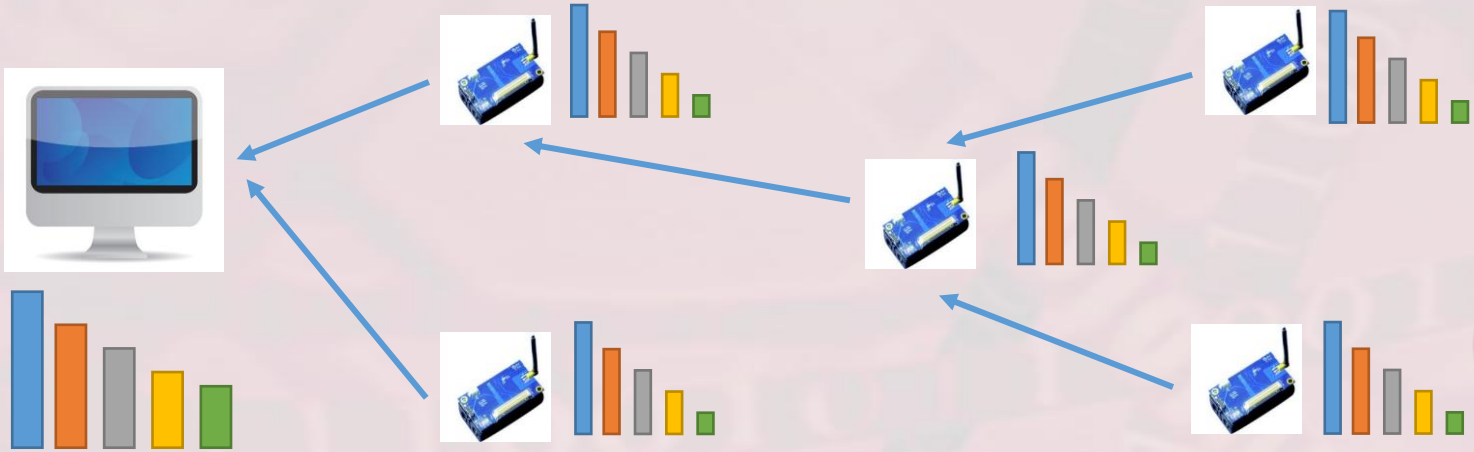
计算模型：分布式模型

- 针对MapReduce、Hadoop等分布式计算平台
  - 输入数据分布在多个节点
  - 每个节点基于其数据，独立计算摘要
  - 将多个摘要在主节点合并，回答关于原始输入数据的查询



计算模型：分布式模型

- 模拟传感器网络中的网络内聚合(In-network aggregation)
  - 每个传感器独立观测数据(如湿度、温度、车流量等)，并计算摘要
  - 摘要通过通讯依次传输合并
  - 最后在主节点合并所有摘要，形成对聚合数据的估计



大数据近似算法

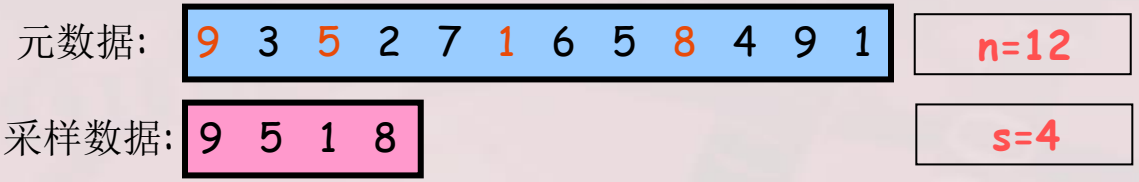
- 研究背景与计算模型
- 随机采样算法
- 基于计数的近似算法
- 基于哈希的近似算法
- 研究成果简介

随机采样算法

- 随机采样与大数定理
- 随机采样应用实例：**Online Aggregation**
- 水塘采样算法
- 随机采样的融合

随机采样

- 在元数据中随机选取一些数据作为“代表”



- 回答近似查询：
  - 平均值：近似avg=(9+5+1+8)/4=5.75，真实avg=60/12=5，误差=(5.75-5)/5=15%
  - 奇数个数：近似count(odd)=3*12/4=9，真实count(odd)=8，误差=(9-8)/8=12.5%

随机采样与大数定理

- 无偏估计：对于AVG, COUNT，SUM等查询，随机采样给出的查询结果的期望就是真实结果
- 需要多少个采样？
  - 根据大数定理，采样数越多，结果越精确，误差越小
- Probably approximately correct (PAC)，可能近似正确保证：
  - 以99.99%的概率/置信度，误差不超过0.01%
  - Chernoff不等式：当采样数达到$\frac{1}{\varepsilon^2}\log\frac{1}{\delta}$时，以$1-\delta$的置信度，保证误差不超过$\varepsilon$
    - 大数据->小数据：当误差ε和置信度δ确定后，采样数与元数据大小无关
    - 采样数和误差平方成反比，1%误差需要10000+个采样。

随机采样应用：Online Aggregation

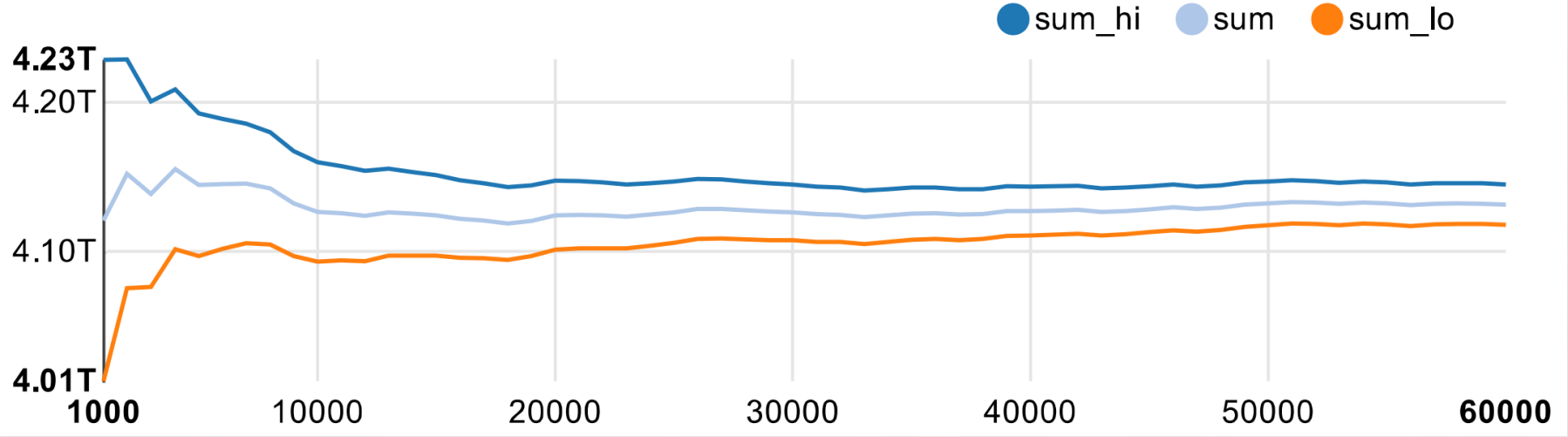
```
SELECT SUM(l_extendedprice * (1 - l_discount))
FROM customer, lineitem, orders, nation, region
WHERE c_custkey = o_custkey
  AND l_orderkey = o_orderkey
  AND l_returnflag = 'R'
  AND c_nationkey = n_nationkey
  AND n_regionkey = r_regionkey
  AND r_name = 'ASIA'
```

该查询返回某公司亚洲地区由于退货导致的利润缩减

随机采样应用：Online Aggregation

```
SELECT SUM(l_extendedprice * (1 - l_discount))
FROM customer, lineitem, orders, nation, region
WHERE c_custkey = o_custkey
  AND l_orderkey = o_orderkey
  AND l_returnflag = 'R'
  AND c_nationkey = n_nationkey
  AND n_regionkey = r_regionkey
  AND r_name = 'ASIA'
  Error 0.01 Confidence 95%
```

随机采样应用：Online Aggregation

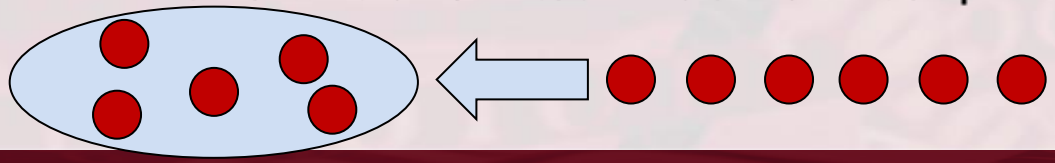


$$\Pr\left[\tilde{Y} - \varepsilon < Y < \tilde{Y} + \varepsilon\right] > 0.95$$

置信区间　　置信度

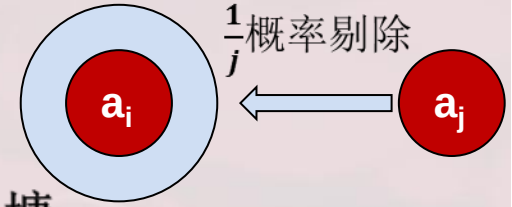
水塘采样算法

- 给定N个元素{a_1, a_2, ..., a_n}，希望随机选取k个元素，
  - 使得每个元素被选取的概率都是$\frac{k}{n}$？
- 数据流模型：只能扫描所有元素一遍，可用空间为k
- 水塘采样算法(Reservoir Sampling)
  - 选取前k个元素放入水塘
  - 对于i>k，当扫描到第i个元素a_i时，以$\frac{k}{i}$的概率选取a_i
  - 若a_i被选取，从水塘中随机剔除一个元素，将a_i加入水塘



水塘采样算法分析

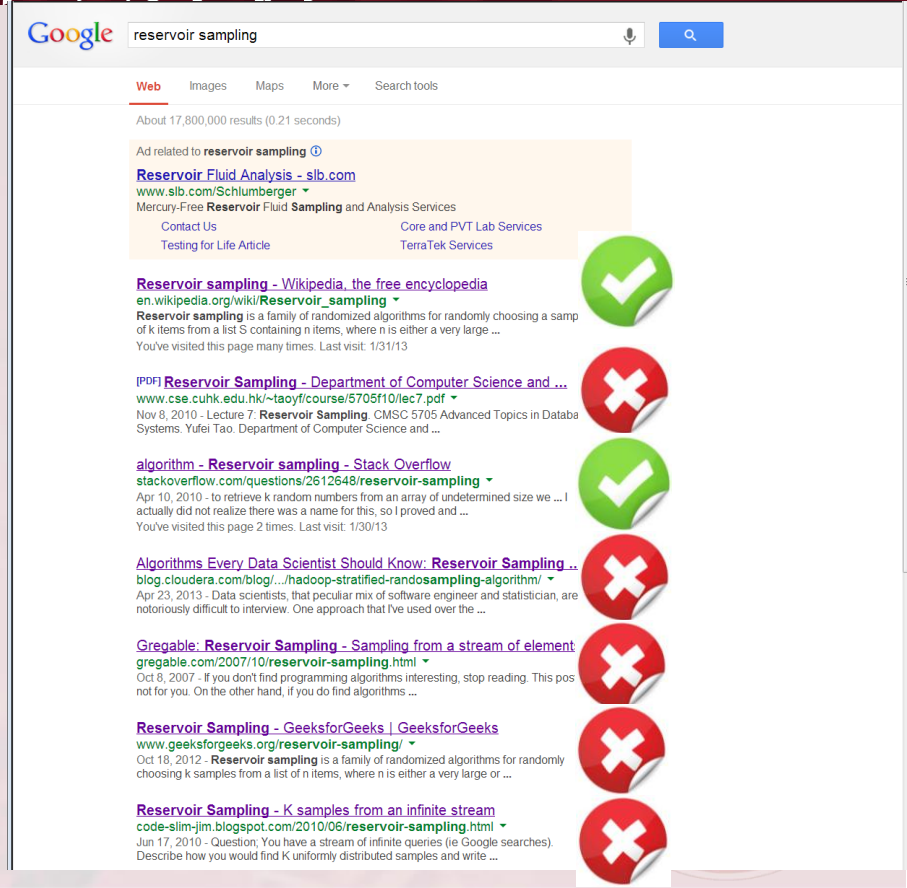
- 考虑k=1的简单情况
- 第i个元素a_i被采样的概率是$\frac{1}{n}$吗？
  - 首先，在扫描到a_i时，a_i必须被选取，概率为$\frac{1}{i}$；
  - 其次，在之后扫描$a_{i+1},\ldots,a_n$时，a_i没有被剔除出水塘。
    - 扫描第j个元素a_j时，a_i被剔除出水塘⇔a_j以$\frac{1}{j}$的概率被选取
    - 因此，a_i仍留在水塘的概率为$1-\frac{1}{j}=\frac{j-1}{j}$



- 因此a_i被采样的概率为：$\frac{1}{i}\times\frac{i}{i+1}\times\frac{i+1}{i+2}\times\cdots\times\frac{n-1}{n}=\frac{1}{n}$

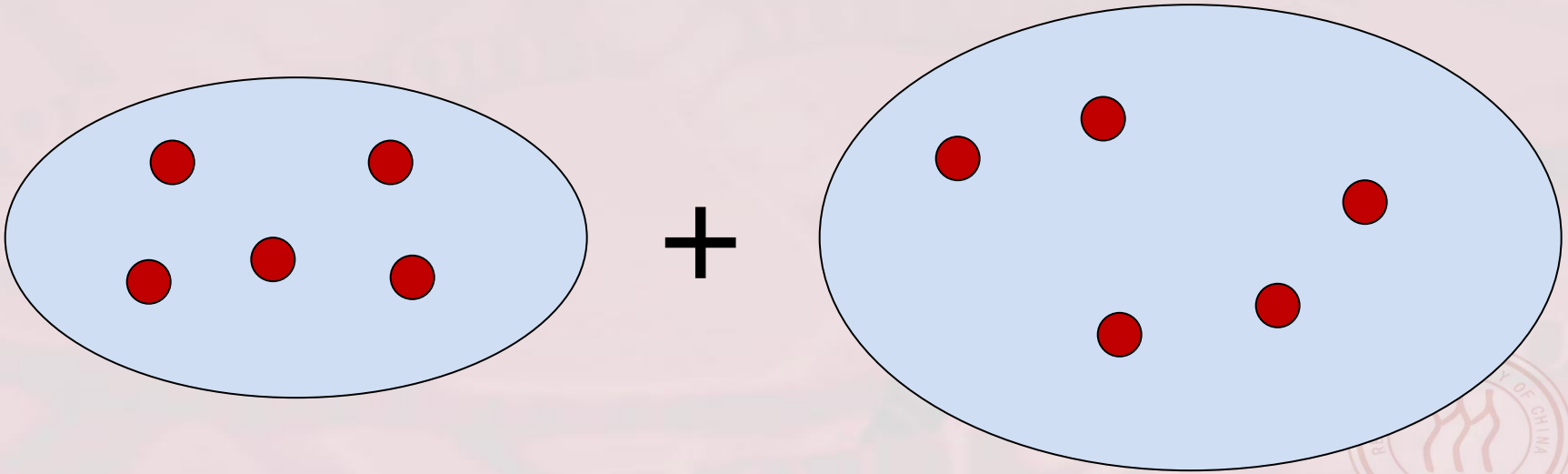
水塘采样算法分析

- 当k>1时，也可以证明采样的随机性。
- 需要注意的是，不仅仅需要证明a_i被选取的概率是$\frac{k}{n}$
- 还需要证明任意一个大小为k的子集被采样到的概率相等
- 正确的证明涉及到Fisher-Yates shuffle。

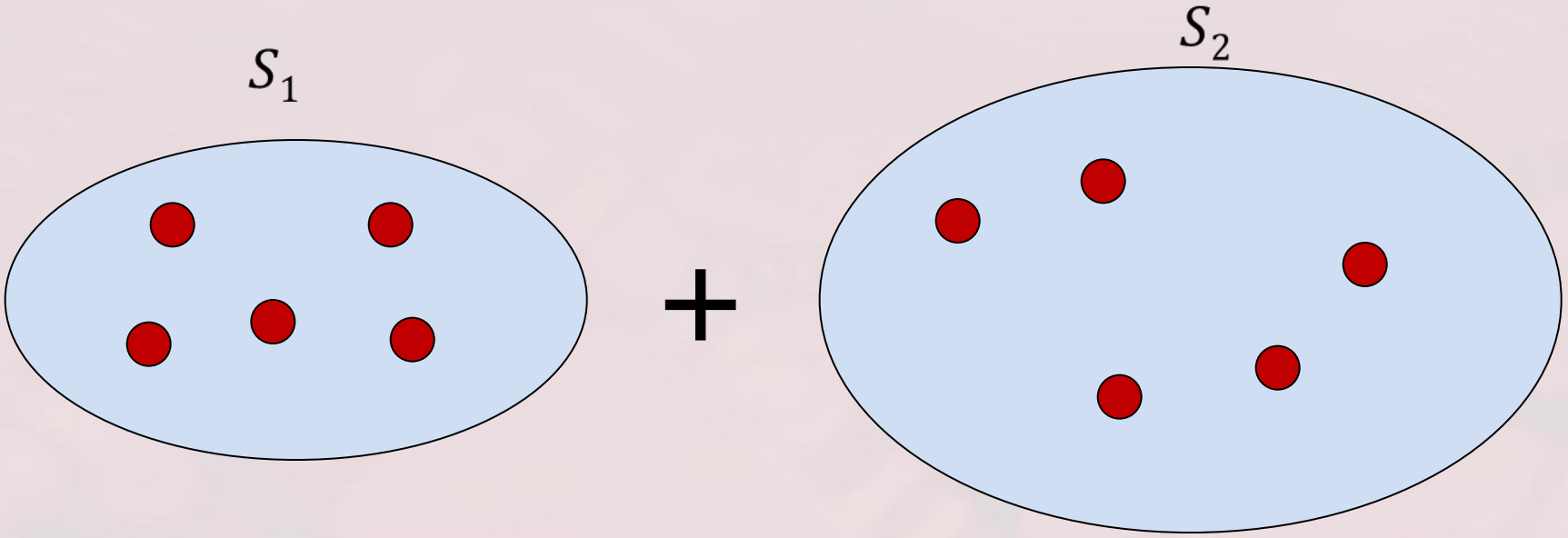


随机采样的合并算法

- 给定集合S_1的k个采样与集合S_2的k个采样，如何在不访问S_1和S_2的前提下，获得S_1与S_2合集的采样？



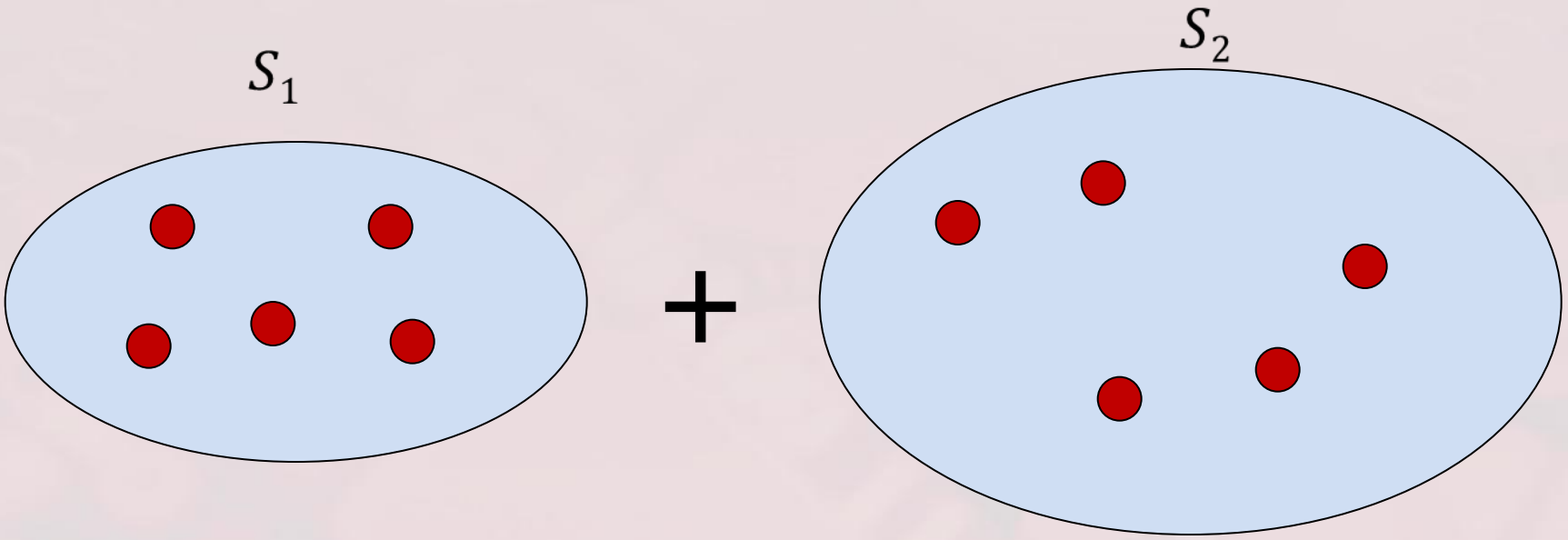
随机采样的合并



$N_1 = 10$

$N_2 = 15$

随机采样的合并

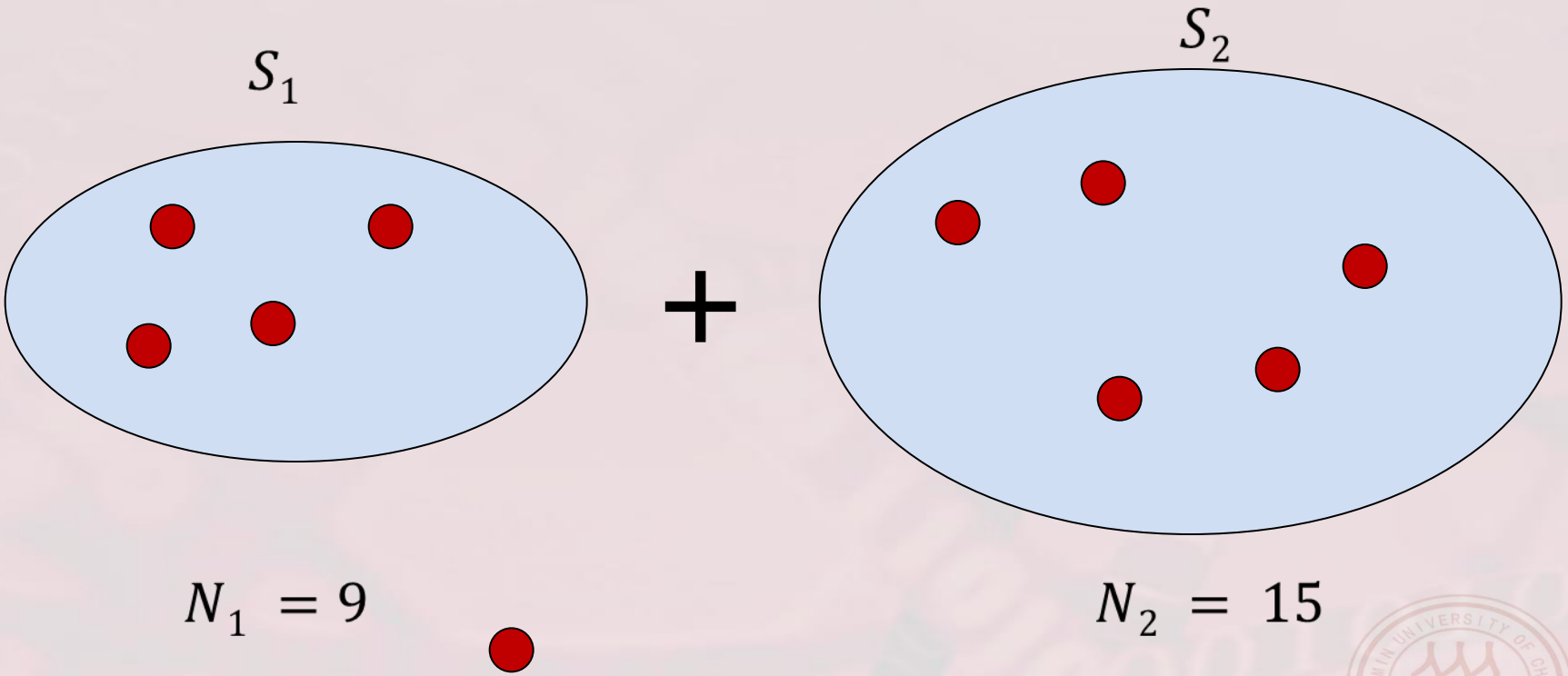


$N_1 = 10$

$N_2 = 15$

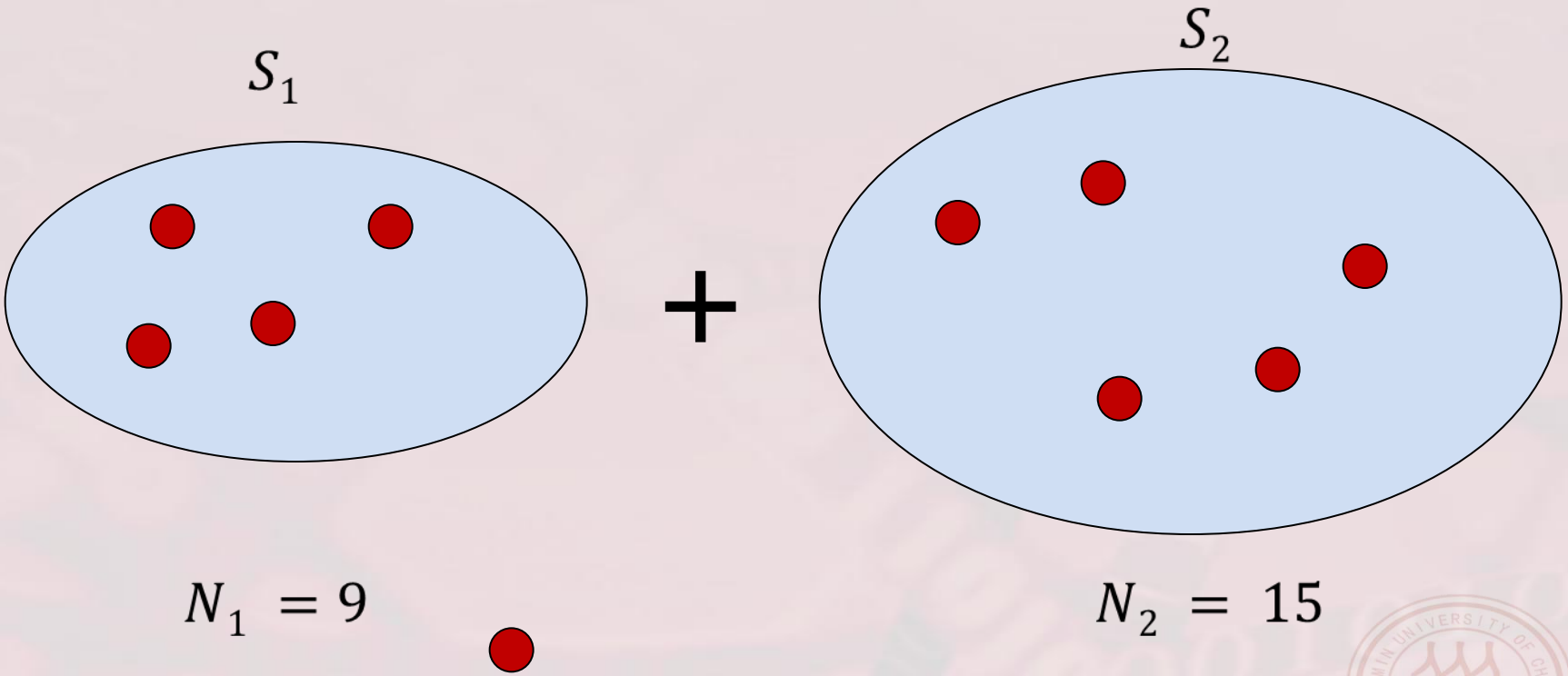
以概率$N_1/(N_1 + N_2)$, 从S_1选取一个采样

随机采样的合并



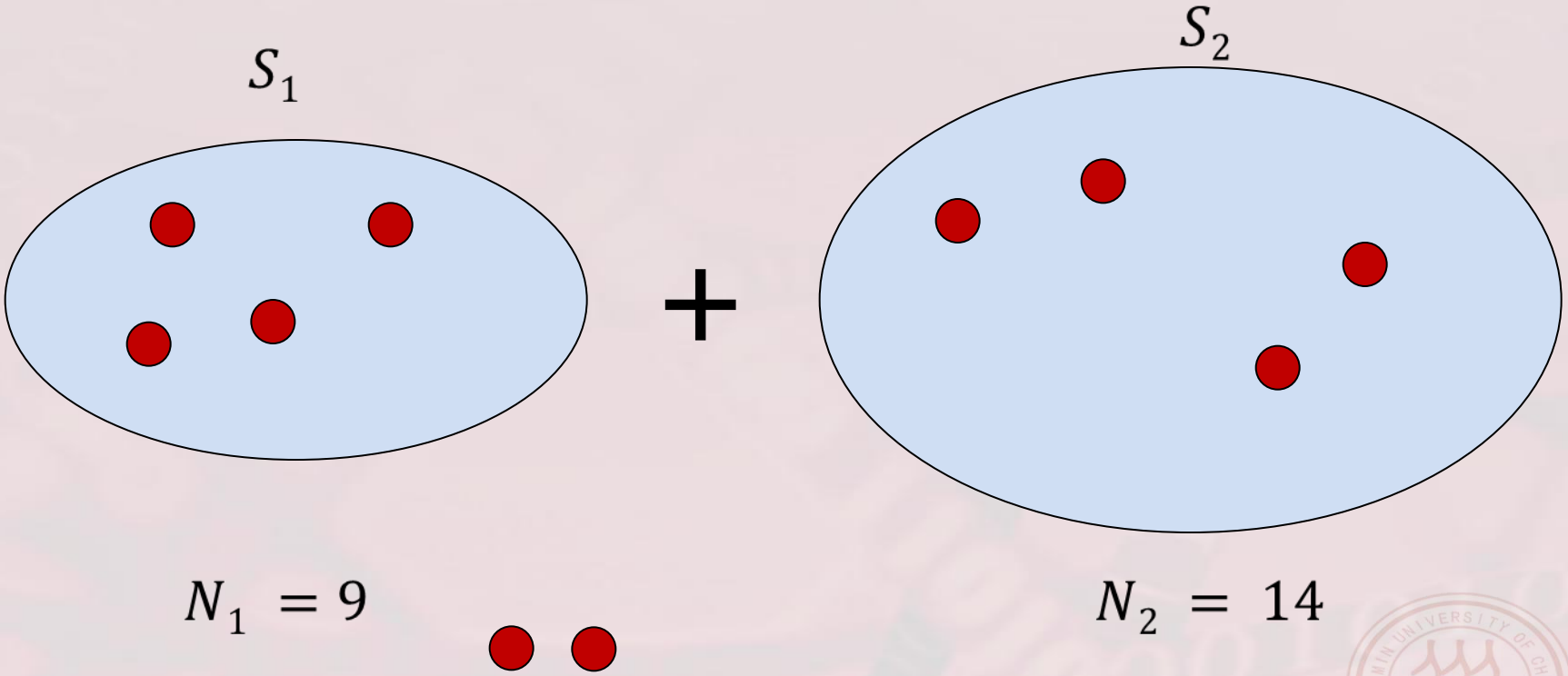
以概率$N_1/(N_1+N_2)$从S_1选取一个采样

随机采样的合并



以概率$N_2/(N_1+N_2)$从S_2选取一个采样

随机采样的合并



随机采样的合并