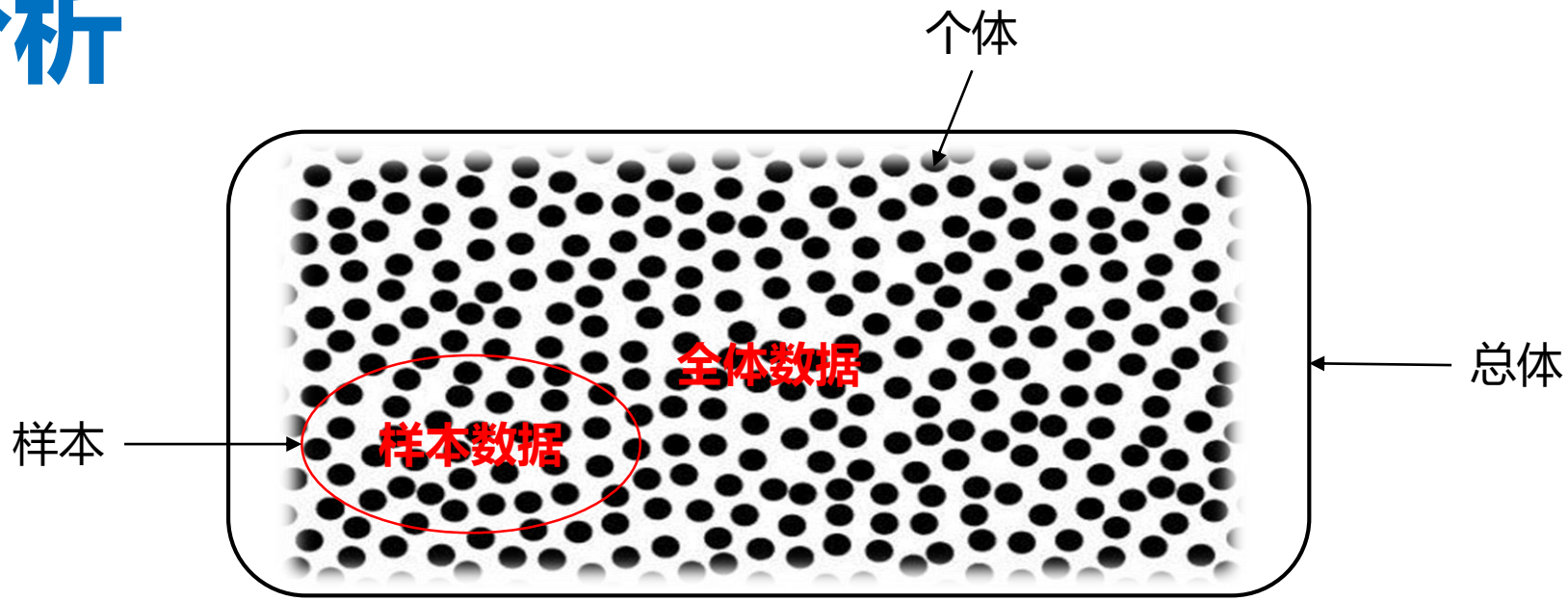
数据科学基础

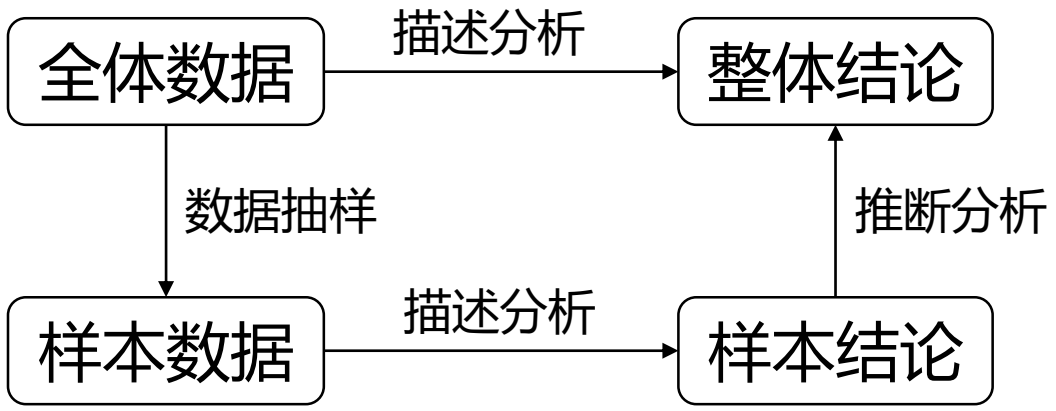
Foundations of Data Science

4.1 数据抽样

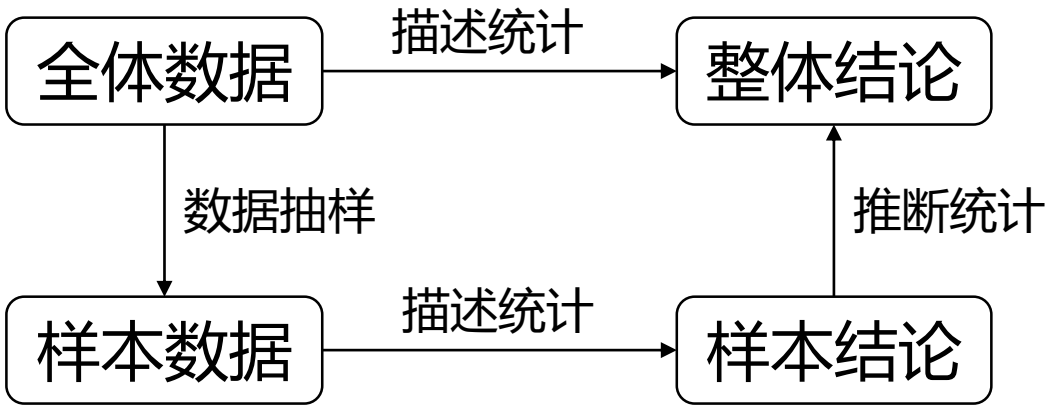
陈振宇
南京大学智能软件工程实验室
www.iselab.cn

数据分析





描述统计与推断统计



•描述统计学(descriptive statistics)通过描述性方法综合概括与分析得出反映客观现象的规律性数据特征。

•推断统计学（inferential statistic）在对样本数据进行描述的基础上，对统计总体的未知数据特征做出概率形式的推断。

抽样方法

抽样方法通常可分为概率抽样和非概率抽样。非概率的抽样方法主要依据研究目的、研究对象和调查资源的受限情况制定的方法，不依据某种随机原则。概率抽样方法依据某种随机原则，主要包括以下四大类：

- 简单随机抽样
- 系统抽样
- 整群抽样
- 分层抽样

(简单)随机抽样

简单随机抽样是指从一个数量为N(可能很大)的总体中逐个等概率无放回抽取个体，直至达到需要的n个个体为止。简单随机抽样是最简单的一种抽样方法，也是其它概率抽样方法的一个基础。

- 由于*N*很大，通常采用有放回抽样替代无放回抽样
- 在应用中，待抽样的研究对象可能是动态变化
- 如何准确有效的编码（映射）是随机抽样的前提

系统抽样

当总体中的个体数较多时，采用简单随机抽样显得较为费事。这时，可根据总体特征然后按照预先定出的规则抽取个体，得到所需要的样本，这种抽样叫做系统抽样。

- 系统抽样的主要特征是：规则+随机
- 等距规则是最常见的系统抽样规则，但也可以是非等距的
- 抽样规则跟系统特征匹配是实施的关键所在

整群抽样

整群抽样先对总体分组(称为群)，再随机抽取群(非个体)，被抽中的群的所有个体组成样本。整群抽样时只需要把群作为抽样框，而不需要把数量庞大的个体作为抽样框，因此能大大降低抽样的成本，提高抽样效率。

- 整群抽样抽取的是群（子集）而非个体，以提高抽样效率
- 当抽样成本较高（通常是物理世界）时，整群抽样较常使用
- 群之间存在差异，由此而引起的抽样误差往往大于简单随机抽样

分层抽样

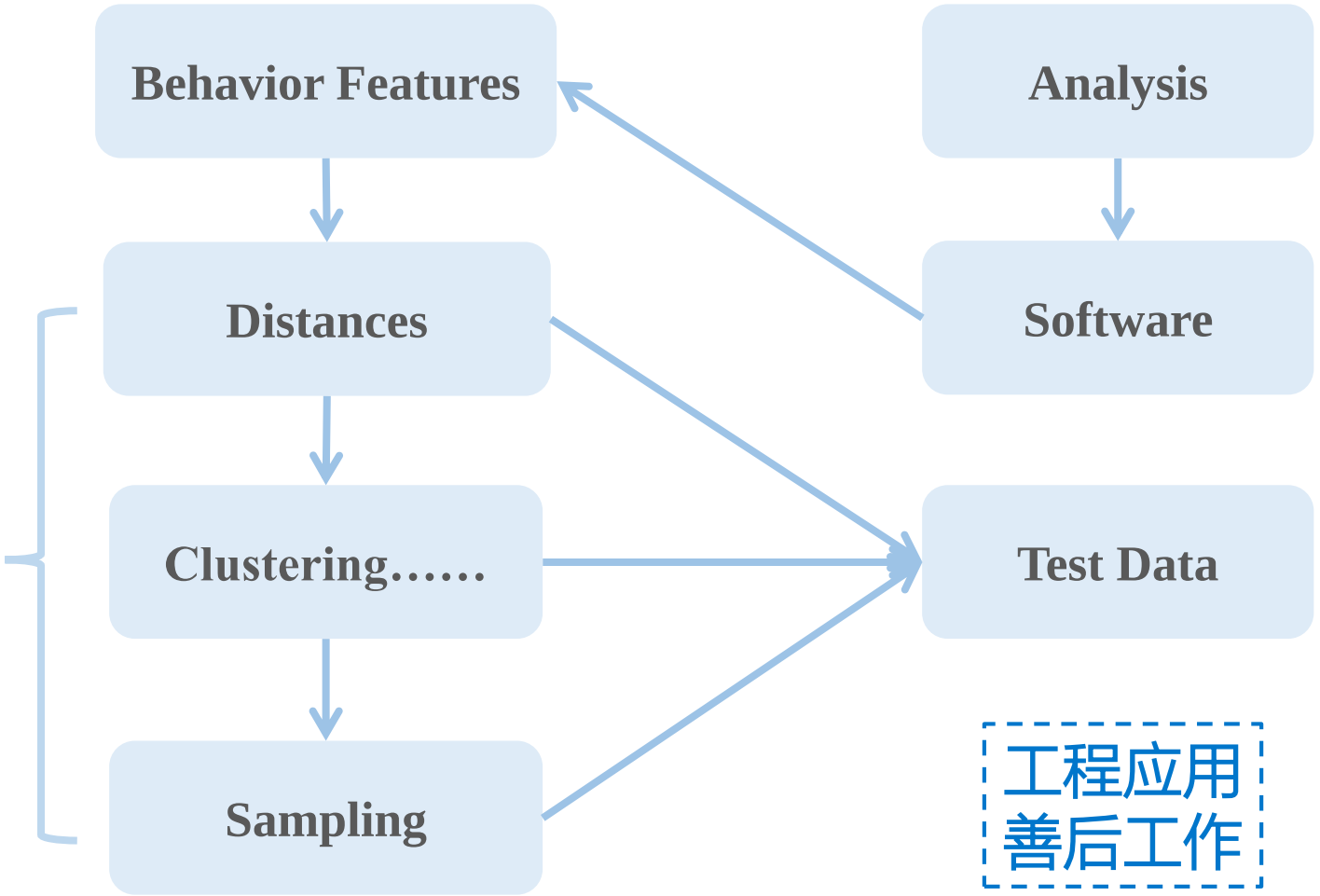
分层抽样先按对观察指标影响较大的某种特征，将总体分为若干个类别(称为层)，再从每一层内随机抽取一定数量的个体，组成样本。

- 研究对象分布不均匀而且特征明显时适合分层抽样
- 层（类别）的大小没有严格要求，依赖特征自动分层
- 分层的特征与研究对象和研究目标的贴切程度决定了分析结果

测试本质上是抽样

工程重点

研究重点



工程应用
善后工作

- 执行路径比覆盖更具行为表征意义
- 执行路径同时复合了代码权重
- 抽样方式更具多样性和资源弹性

更多详情，请咨询课程助教。

失效行为构建分层

执行路径距离示例

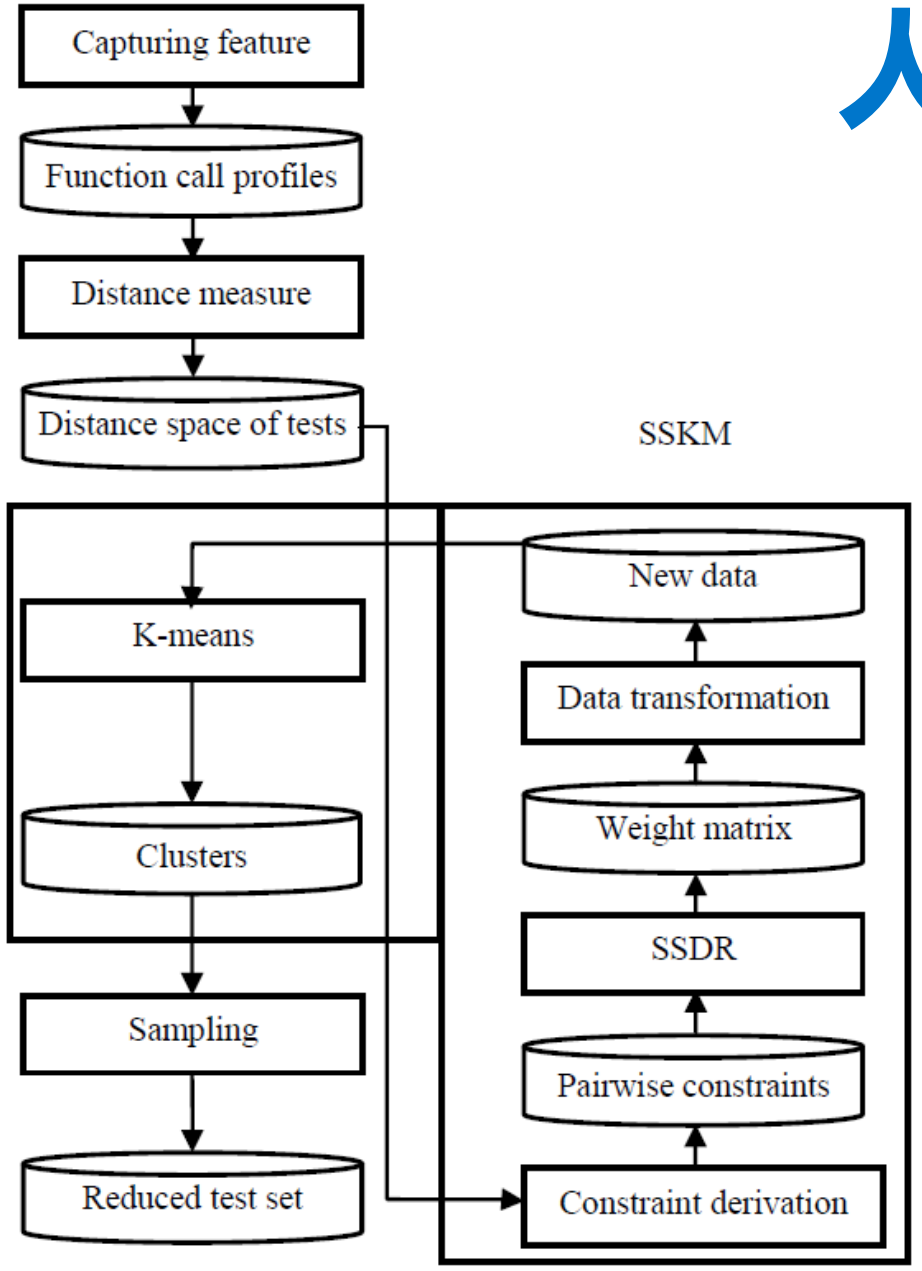
- 海明距离等
- **D(t1,t2)=4**
- **D(t1,t3)=3**
- **D(t1,t2)> D(t1,t3)**
- **[t1,t3] 比 [t1,t2] 更相似**

由于软件系统结构和缺陷产生机理的复杂性

执行路径并不能完全准确刻画软件失效行为

	t1	t2	t3	t4
s1	1	0	1	0
s2	1	1	0	1
s3	0	1	0	1
s4	1	1	1	0
s5	1	0	1	1
s6	0	0	1	1
s7	1	1	0	1

人机协同改进分层精度



标注信息

- Must-Link:相同软件行为 M
- Cannot-Link:不同软件行为 C

$$J(w)=\frac{1}{2n^2}\sum_{i,j}(w^Tx_i-w^Tx_j)^2$$

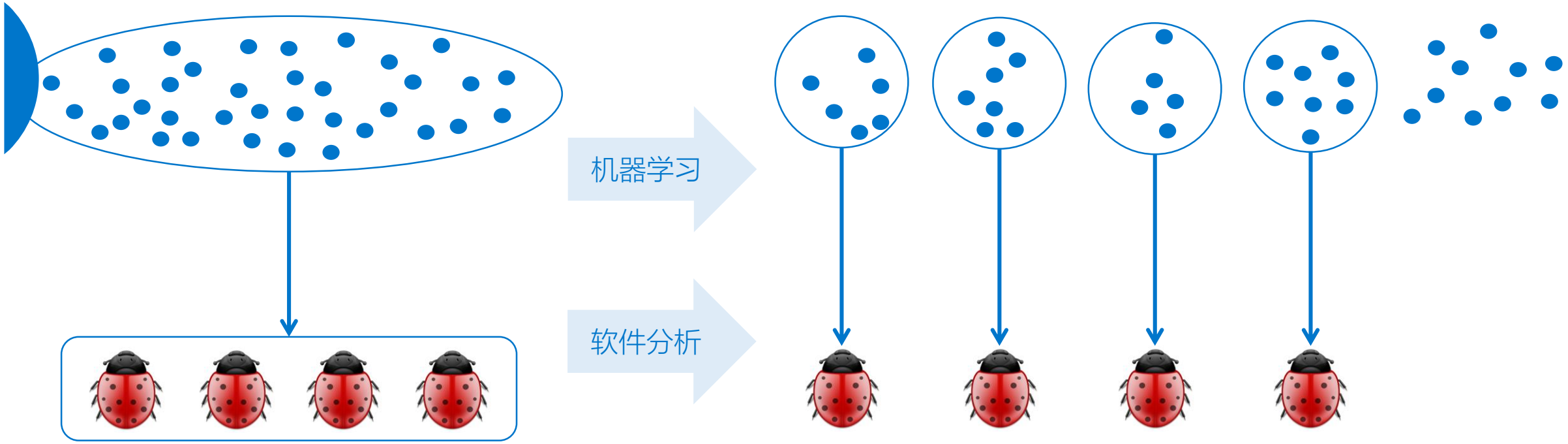
寻找行为空间转换矩阵***W***
极大化优化目标***J(W)***

$$+\frac{\alpha}{2n_C}\sum_{(x_i,x_j)\in C}(w^Tx_i-w^Tx_j)^2 \quad (3)$$

$$-\frac{\beta}{2n_M}\sum_{(x_i,x_j)\in M}(w^Tx_i-w^Tx_j)^2$$

Songyu Chen, **Zhenyu Chen**, Zhihong Zhao, Baowen Xu, Yang Feng. Using Semi-supervised Clustering to Improve Regression Test Selection Techniques ICST 2011

更多详情，请咨询课程助教。



动态抽样
ICST 2010

半监督
ICST 2011

逻辑测试
SCIS 2012

多标签
ICSE 2012

动态加权
JSS 2014

相似性排序
SQJ 2014

动静态融合
ISSTA 2014

众包报告
FSE 2015

更多详情，请咨询课程助教。

感谢上述五大科研合作伙伴的鼎力支持

更多详情，请咨询课程助教。