# 数据库系统概论新技术篇

# 大数据思维和方法

文继荣

中国人民大学信息学院

2017年4月

# 什么是大数据

# 大数据的通常定义

- 百度百科
  - 大数据（big data），指**无法在一定时间范围内**用**常规软件工具**进行捕捉、管理和处理的数据集合，是需要**新处理模式**才能具有更强的决策力、洞察发现力和流程优化能力的海量、高增长率和多样化的**信息资产**。

# 大数据的通常定义

- Wikipedia:
  - Big data is a term for data sets that are so large or complex that traditional data processing application software is inadequate to deal with them.

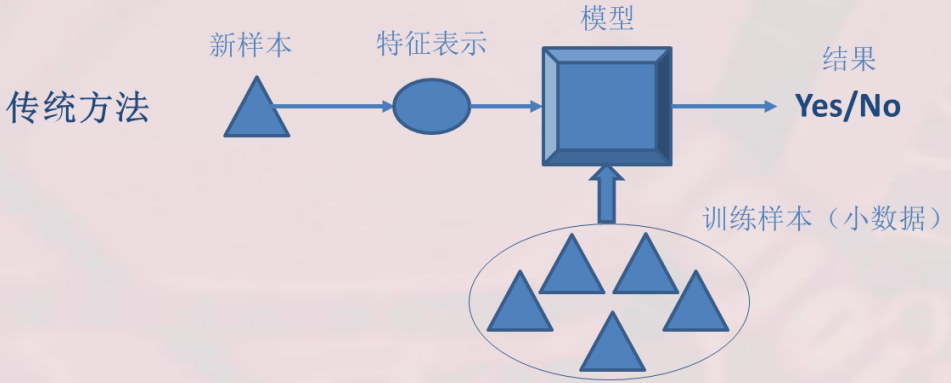
- 问题：始终围绕着“大”来定义，没有揭示背后的深层意义

# 人类的理性主义传统

- 经验收集和分享的困难
- 理性主义
  - 从特殊到一般：相信人能透过现象看到本质
  - 从一般到特殊
- 模型缓解了经验的不足
  - 从有限的个人经验中得到普遍性的规律
  - 泛化：从已知到未知

# 基于理性主义的传统方法

- 从理性或直觉中建立问题的模型，或通过少量样本数据的观察归纳出模型
- 通过模型判别新样本



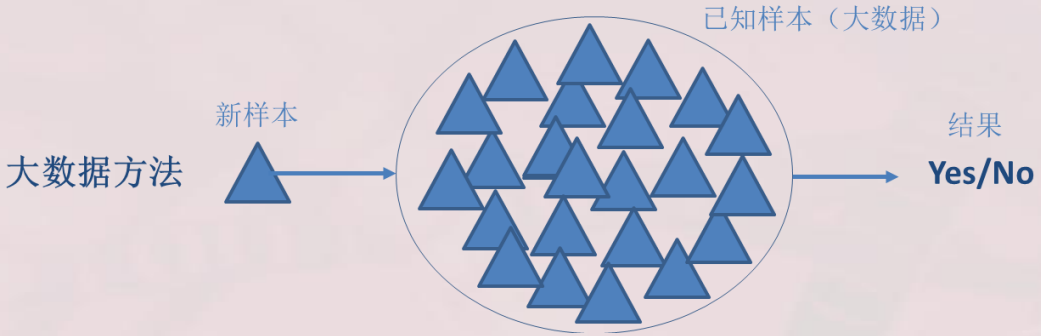
# 传统方法的内在困难

- 是否总是能从特殊推到一般？
- 复杂模型：股市预测

# 大数据时代

- 新技术使得经验数据的收集和分享变得容易
  - 互联网
  - 物联网
  - 移动设备
  - 穿戴设备
- 数据越多，就越不需要依靠模型的泛化能力

# 大数据方法




- 覆盖度：对所有或大部分情况，我们有样本来覆盖
- 精度：对常见情况，我们有足够多样本来提升精度

# 什么是大数据？

- 大数据是现代社会在掌握海量数据收集、存储和处理技术基础上所产生的一种以海量数据进行判断和预测的能力，代表了一种新经验主义方法。

- 特点
  - 经验主义 > 理性主义
  - 数据 > 模型

# 谢谢!