数据库系统概论新技术篇

大数据思维和方法

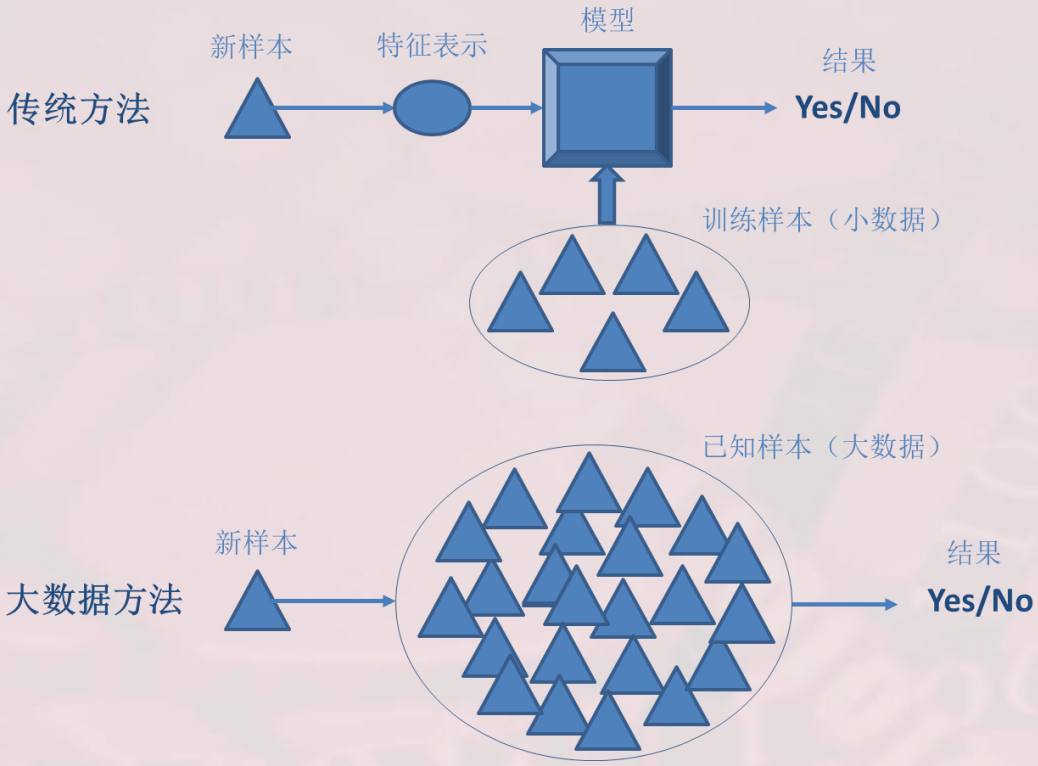
文继荣

中国人民大学信息学院

2017年4月

大数据方法及实例

传统方法 vs. 大数据方法



例子一：考试成绩换算

❖一次期末考试成绩

学生	基本分 x_1	附加分 x_2	调整后成绩 y
001	90	10	100
002	80	5	92
003	85	0	92
004	78	10	93
005	75	5	89
006	66	15	89
007	52	5	75
008	83	10	?

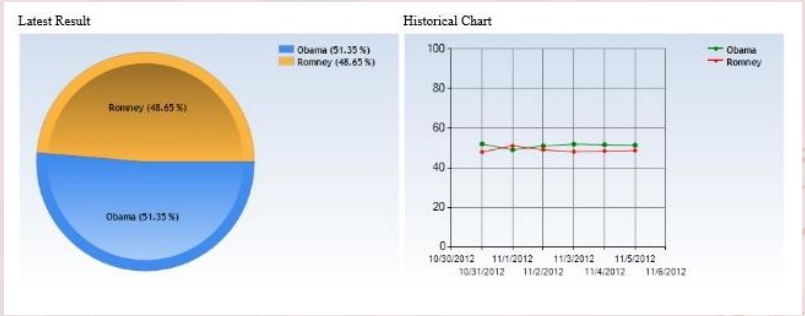
$$y = 10 \times \left(\sqrt{x_1} + \frac{x_2}{20}\right)$$

例子二：机器翻译

- 问题：将一种语言（如中文）自动翻译为另一种语言（如英文）
- 传统解决方法：语料库+翻译模型
- 大数据方法：平行语料挖掘
  - 从互联网上自动发现大量的双语语料
  - 统计词语、短语、甚至句子之间的对照关系
  - **非常显著的性能提升，目前最好的方法**

例子三：预测美国大选

- 问题：2012年美国大选，奥巴马和罗姆尼谁会赢
- 传统的解决方法：民意调查，专家意见，预测模型
- 大数据方法：网络数据舆情分析
  - http://research.microsoft.com/en-us/projects/websensor/election2012.aspx
  - 从公开的网络数据源（论坛，新闻评论，社交媒体）中收集大量相关数据
  - 分析和统计网民的民意
  - **与真实大选结果非常接近**



多大的数据是大数据？

- 当数据多到能对整个样本空间进行充分覆盖，这样的数据就足够“大”了
  - 对于第一个例子中的考试成绩换算问题，样本空间为300，因此均匀分布的300个样本就足够了
  - 对于机器翻译，样本空间的数量级就大很多：所有可能的句子

模型真的没有用吗？

- 数据总是不够
  - 样本空间太大
    - 机器翻译例子中所有可能的句子
  - 样本空间变化
    - 新的词语和新的含义在不停出现
- 模型需要和大数据结合，提供适当的泛化能力

谢谢!