基于大数据文科综合训练国家级虚拟仿真实验教学示范中心重点建设课程

大数据分析导论

文继荣、窦志成

http://playbigdata.com

中国人民大學
RENMIN UNIVERSITY OF CHINA
中国人民大学信息学院
SCHOOL OF INFORMATION RENMIN UNIVERSITY OF CHINA

大数据思维、方法及应用

文继荣

国家“千人计划”特聘专家
中国人民大学信息学院

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

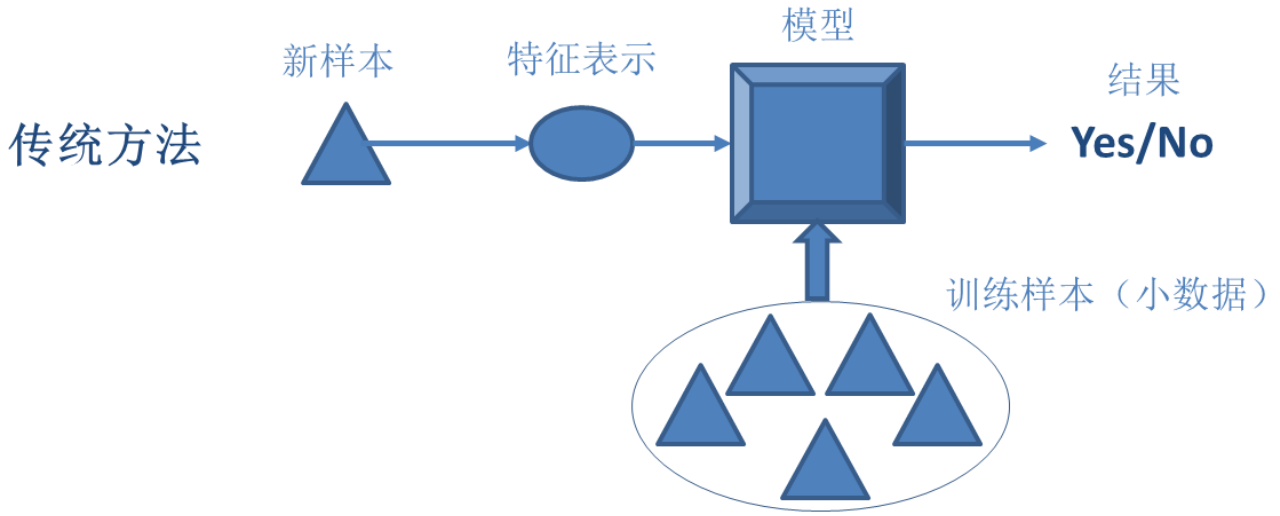
- Wikipedia:
  - Big data is a term for data sets that are so large or complex that traditional data processing application software is inadequate to deal with them.
- 问题：始终围绕着“大”来定义，没有揭示背后的深层意义

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



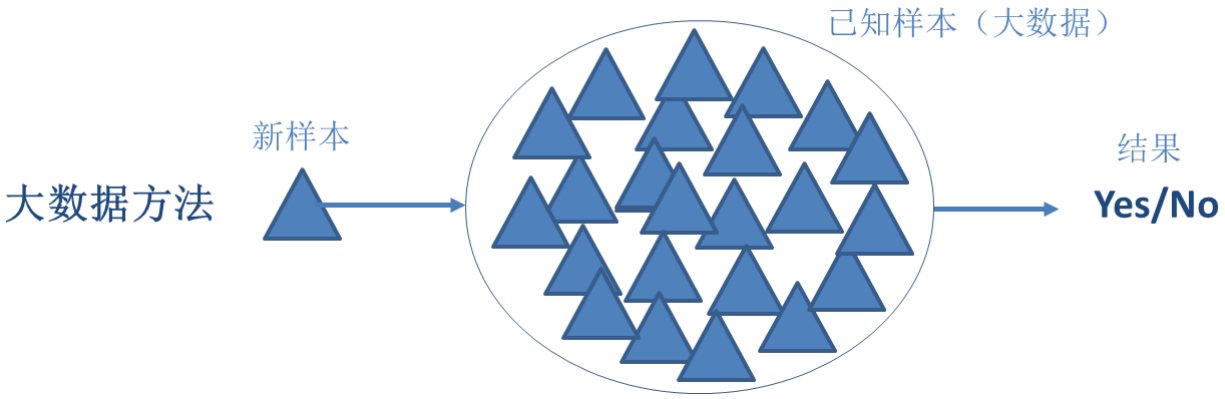
传统方法的内在困难

- 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

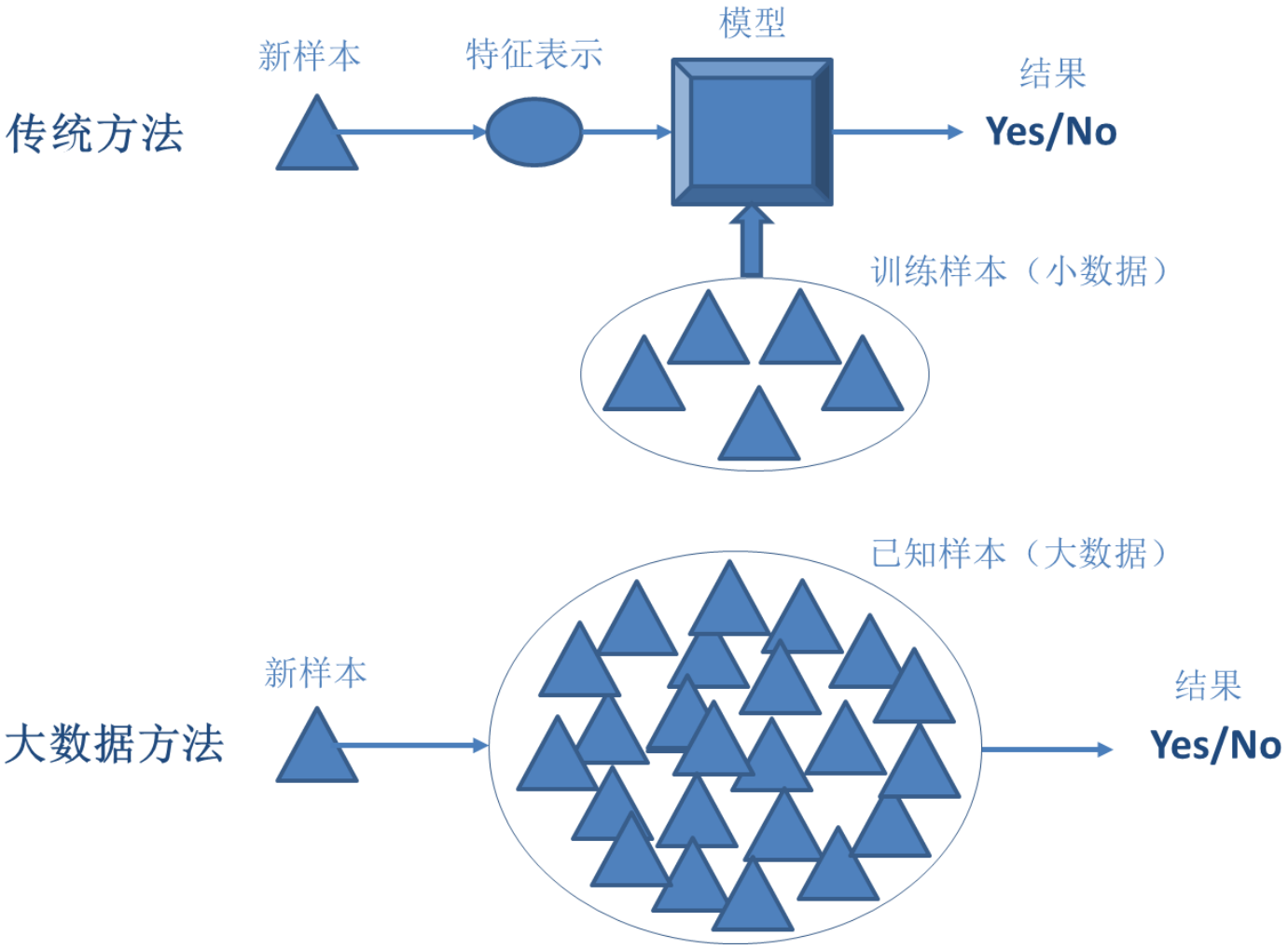
什么是大数据？

- 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

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**

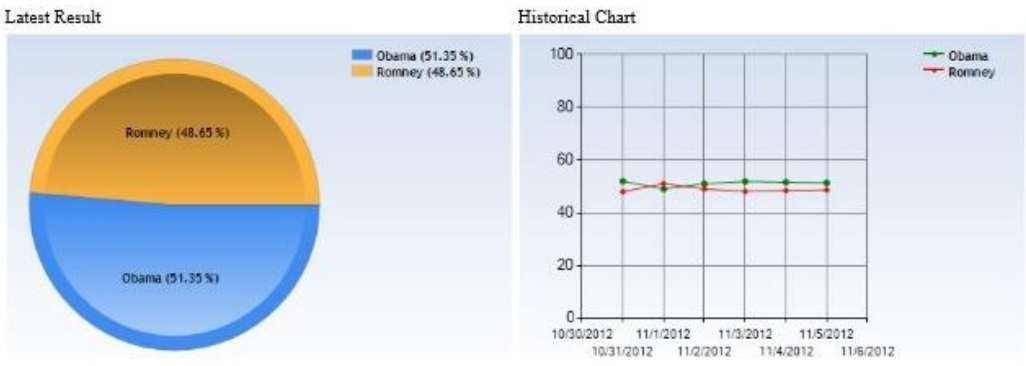
例子三：查询结果排序

- 问题：给定一个查询，对网页进行相关性排序
- 传统解决方法：排序模型
  - 概率模型
  - 语言模型
  - 神经网络（Learning to rank）
- 大数据方法：用户点击数据挖掘
  - 给定一个查询，根据用户对网页的点击率排序
  - 需要大量的数据：查询数*网页数
  - **效果明显提升：最强的feature**

例子四：微软小冰

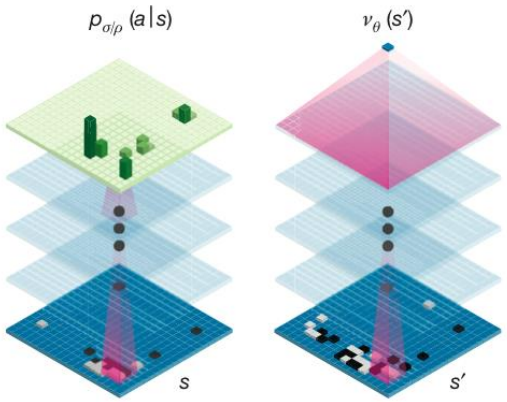
例子五：预测美国大选

- 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
- 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**



例子六：AlphaGo

AlphaGo



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
    - 查询结果排序例子中，新的查询和新的网页在不停出现
- 模型需要和数据结合，提供适当的泛化能力
  - 如何结合？

大数据应用

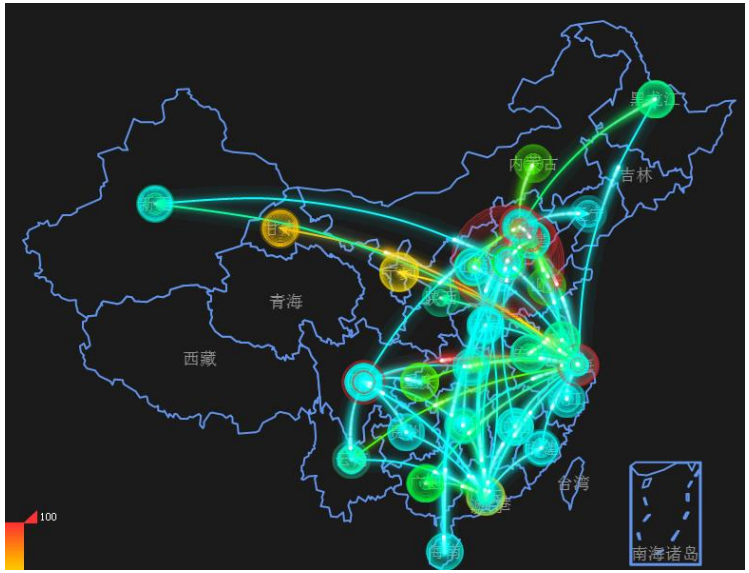
社会科学应用

- 探索大数据驱动的人文社会科学研究新模式

- 数据世界 vs. 物理世界
  - 大数据中蕴含着丰富的社会信息，是对真实社会的映射
- 经验 vs. 模型
  - 人类社会中的很多问题非常复杂，大数据方法的出现使得一些传统上利用理论模型无法有效表达的问题可以得到经验意义上的解决
- 大数据对于人文社会科学研究提供了新的思路、方法和工具

大数据+经济学

- **产业转移**：中国近年来各省之间产业转移的情况如何？
  - 没有直接统计数据
- 解决方法：
  - 新闻数据挖掘：y=产业转移数(时间，省份1，省份2)
  - 模式挖掘：y=频繁模式（产业转移数(时间，省份1，省份2)）



http://websensor.playbigdata.com/chanyeqianyi

大数据+社会学

- **社会意识形态分析：中国近年来民族主义意识在抬头吗？**
- 大数据分析
  - 1.7亿用户，27亿微博
  - 146个意见领袖对于不同国家不同事件的民族主义倾向

Figure 1: Distribution of Sentiment across Different Topics, 2013

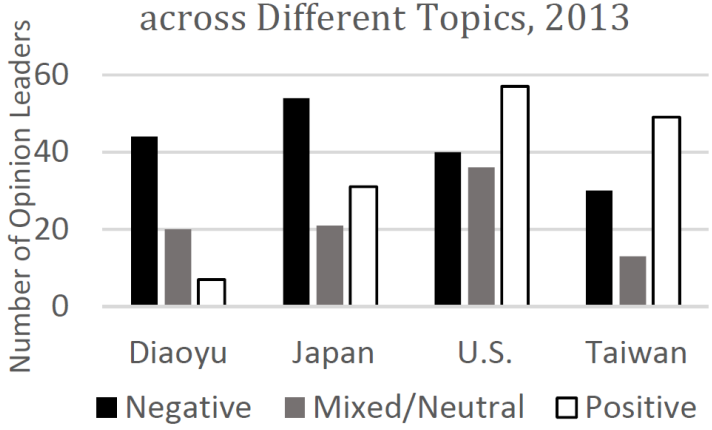
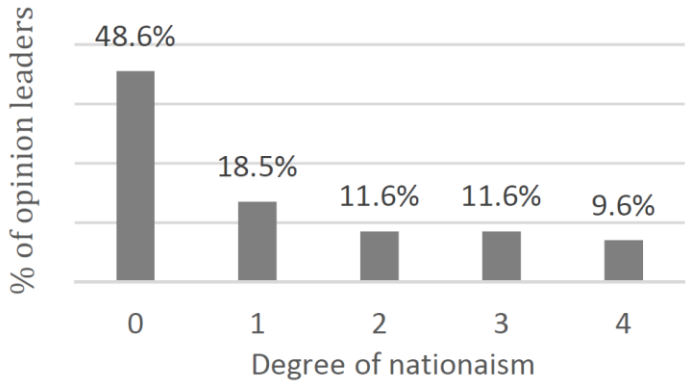


Figure 3: Different Levels of Nationalism among Opinion Leaders (N=146)



- 结论
  - 民族主义在中国并非想象的那么严重
  - 民族主义表现呈现多元化特点
  - 民众更关心国内事务和政策

大数据+历史学

- **中国历史上自然灾害大数据分析**
  - 史籍浩如烟海，尤其是明清史
  - 历史上的自然灾害记录散布在数量巨大的史料中
  - 人工分析已不可行
- 史料数据挖掘
  - y=自然灾害发生数（时间，地点，自然灾害类型）

大数据+金融

- 媒体关注与资产定价效率—基于互联网大数据的经验证据
  - 研究互联网媒体中关于上市公司的新闻报道的数量和情绪对股票收益率的影响
  - 策略组合和结论
    - 买入没有媒体报道的股票，卖出媒体关注度高的股票：获得无风险超额收益
    - 买入正面报道股票，卖出负面报道股票：获得无风险超额收益

中国人民大学攻读博士学位研究生

学位论文开题报告书

拟定学位论文题目：

媒体关注与资产定价效率

——基于互联网大数据的经验证据

院、系：商学院

专　业：财务学

姓　名：刘向强

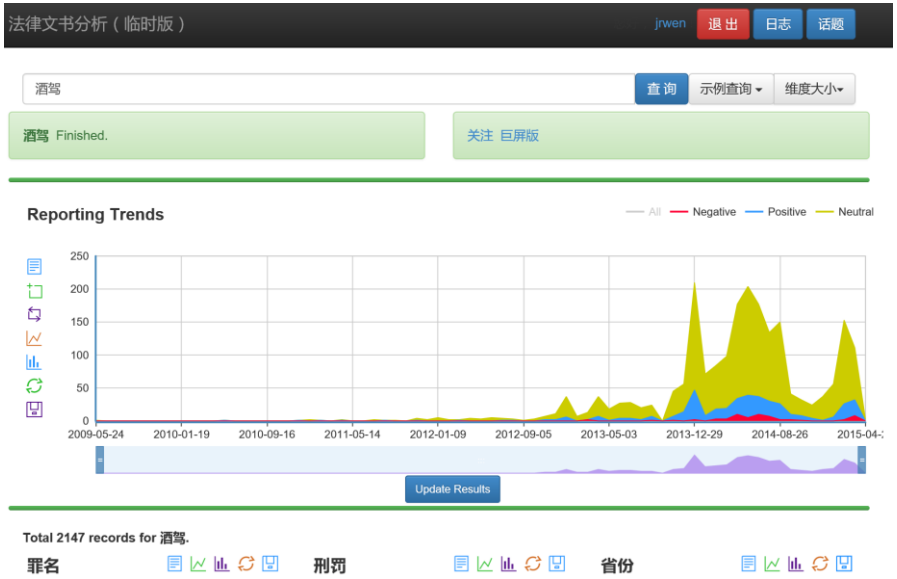
学　号：2013000511

研究方向：公司财务与资本市场

指导老师：王化成　教授

大数据+法学

- 英美法系：判例法
  - 基本原则是“遵循先例”，即法院审理案件时，必须将先前法院的判例作为审理和裁决的法律依据
- 大陆法系：制定法
  - 立法机关依照法定程序制定和公布的法律
  - 中国大部分法律为制定法
- 大数据
  - 更精细全面的判例法实施
  - 统计分析：频率=f(罪名，省份)



http://websensor.playbigdata.com/verdict2/Index.aspx

谢谢!