社会科学方法前沿（二）

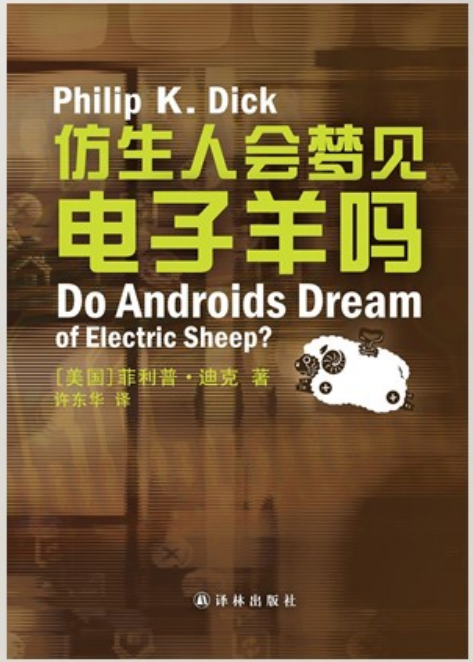
机器学习

林景 讲师

南京财经大学法学院

什么是机器学习?

- 仿生人是否会梦见电子羊吗?
  - *Do Androids Dream of Electric Sheep ?*
  - 人与仿生人的区别
    - Nexus-6
    - 沃伊-坎夫测试（ Voight-Kampff Test ）
    - 提出20-30个相互参照的、可能会引起情绪激动的问题，然后测试对方的一系列身体机能，包括呼吸、心率、脸红和眼球运动等
  - 情感化的行动VS机械化的行动



人的学习VS机器学习

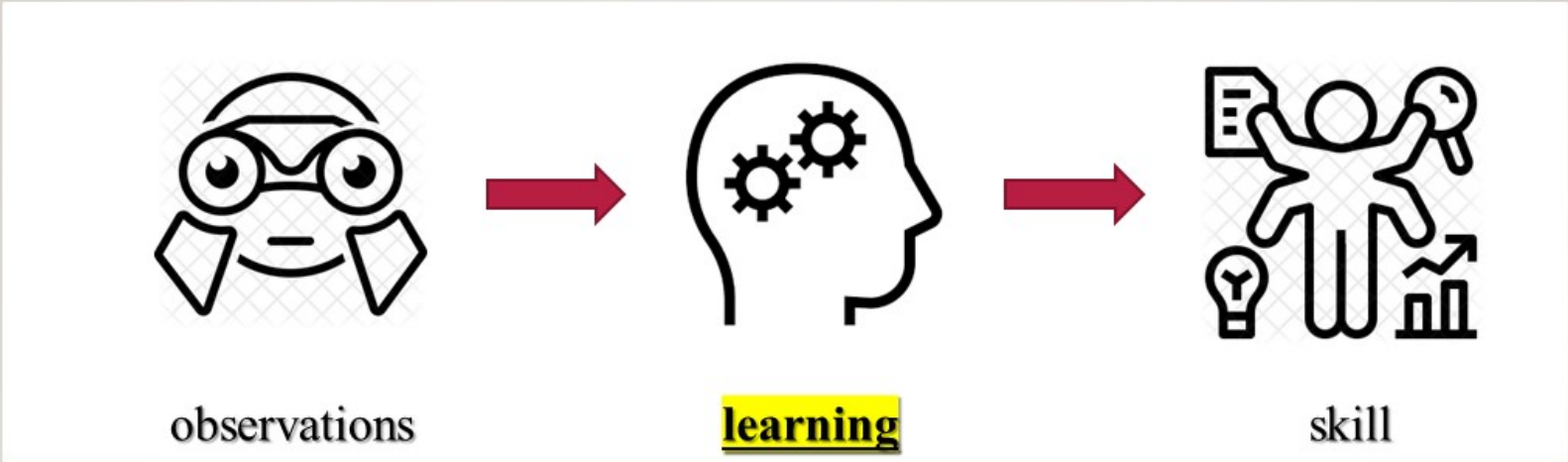
- 语音辨识能力超过人类
  - 机器辨识能力5.5% VS 人类辨时能力5.1%
  - *English Conversational Telephone Speech Recognition by Humans and Machines*
- 影像辨识能力超过人类
  - 对动植物种群的识别和分类
  - *Imagenet classification with deep convolutional neural networks*
- 阅读理解能力超过人类
  - 机器阅读得分87分 VS 人类阅读得分84分
  - *SQuAD: 100,000+ Questions for Machine Comprehension of Text*

人的学习VS机器学习

- 下围棋超过人类
  - AlphaGo：战胜围棋界大国手李世乭、柯洁
  - *Mastering the game of Go with deep neural networks and tree search*
- 打《星际争霸》超过人类
  - AlphaStar：超越99.8%活跃玩家
  - *Grandmaster level in StarCraft II using multi-agent reinforcement learning*
  - 能使用任一种族，包括人族、神族和虫族；
  - 和人类玩家一样的游戏视野；
  - 和人类玩家一样的操作频率；
  - 机器学习过程完全自动化。

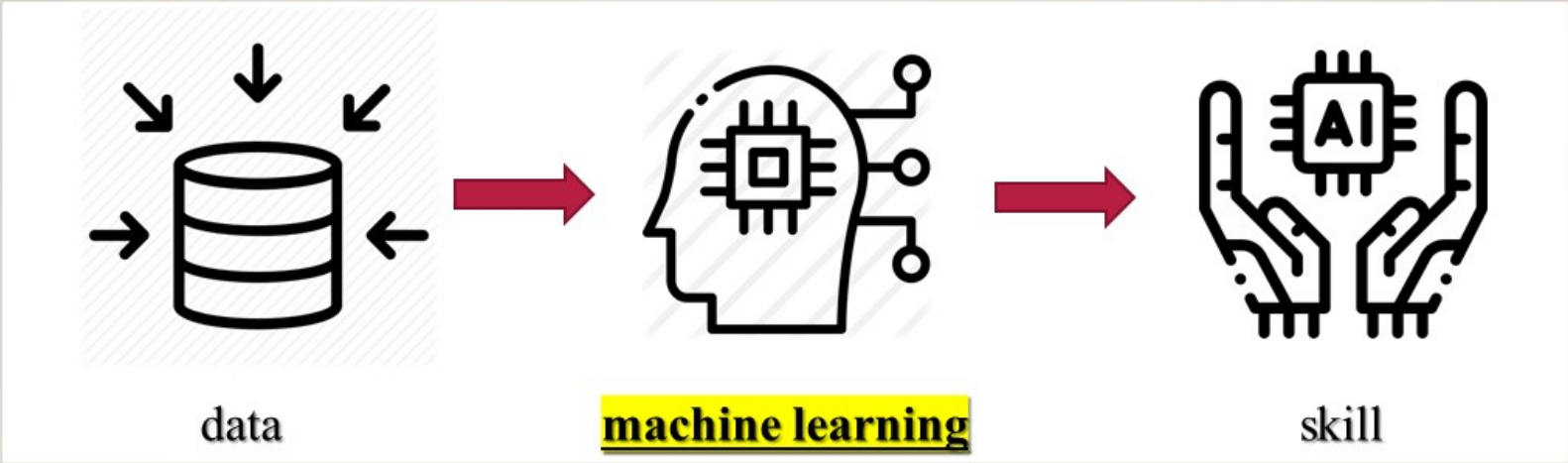
从人的学习到机器学习

- 人的学习（human learning）
  - 学习方式：从观察中累积经验，进而获得技能（acquiring skill with experience accumulated from observations）



从人的学习到机器学习

- 机器学习（machine learning）
  - 学习方式：从资料累积和计算中累积经验，进而获得技能（acquiring skill with experience accumulated/computed from data）

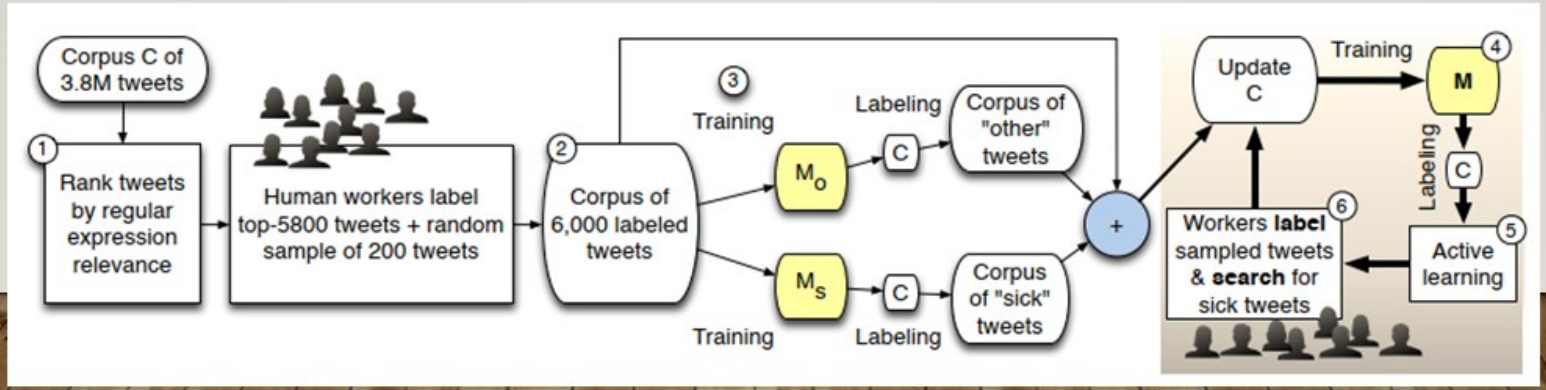


机器学习的基础概念

- 机器学习的概念界定
  - 利用从数据或资料中计算获得的经验来改进机器/计算机的某些性能，以帮助或辅助社会科学研究。这些性能包括数据或资料的处理、运算、分析和预测等。
- 机器学习对社会科学的推助
  - 当社会科学家无法快速做出决定时，如金融学高频交易研究；
  - 当社会科学家无法手动控制研究进程时，如谣言信息扩散研究；
  - 当社会科学家需要大规模面向重复劳动时，如海量数据的采集、编码和分析；
  - 当社会科学家期待综合既有条件做出最优预测时，如经济增长预测。

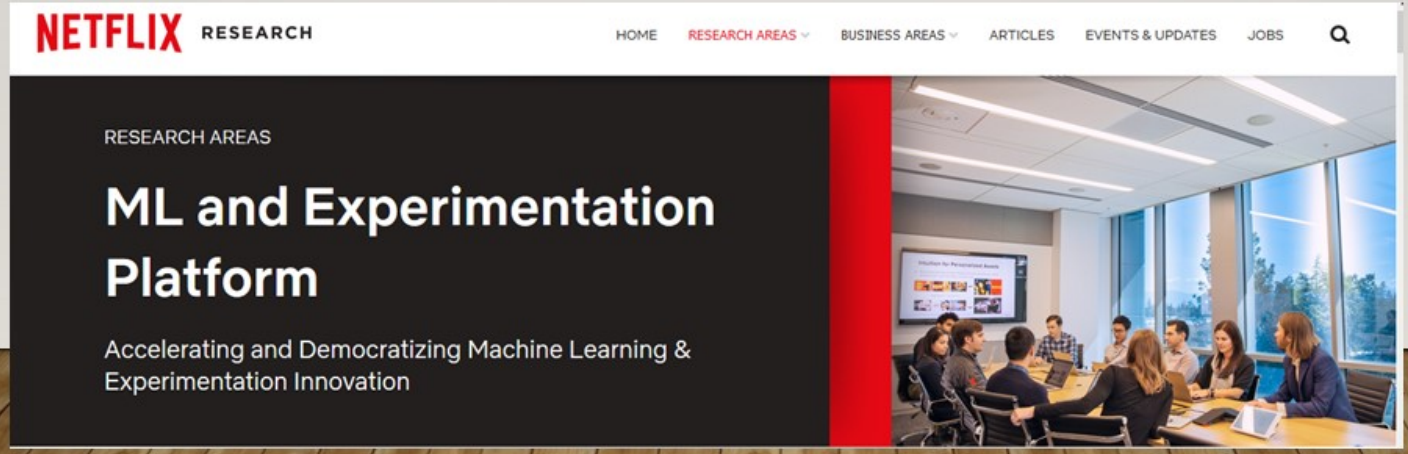
机器学习的应用场景

- 日常生活应用场景
  - 食物中毒分析
  - *nEmesis: Which restaurants should you avoid today?*
  - 数据来源：380万条Twitter数据（单词描述+餐馆地理定位）
  - 机器学习技能：分析人们在网上交流的语言，找到可能出现食物中毒的个体和餐馆，并正确计算每个餐馆可能存在的食品安全风险。

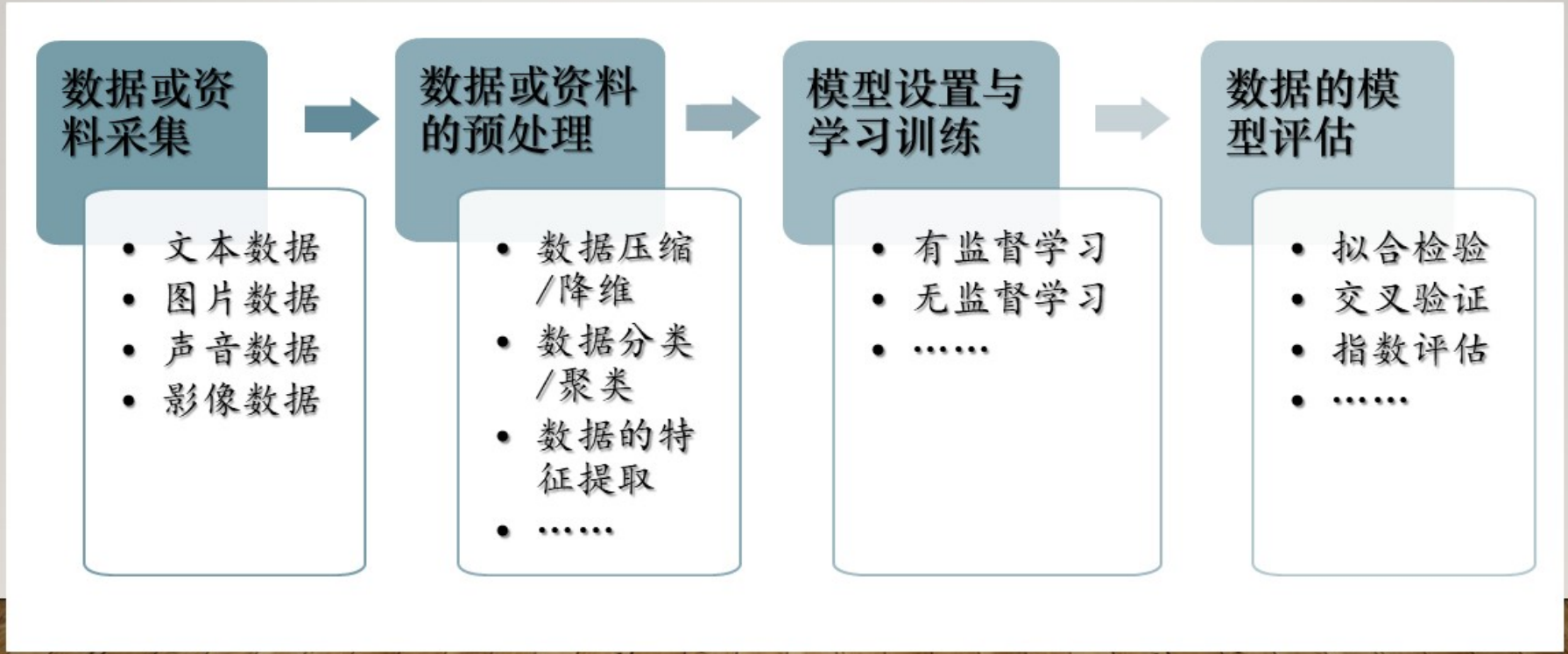


机器学习的应用场景

- 日常生活应用场景
  - 对即将上映电影的推荐
  - Netflix Machine learning and Experimentation Platform
  - 数据来源：48万名用户给1.7万部电影的评分数据
  - 机器学习技能：预测电影观众对即将上映的电影的评分



机器学习的操作步骤


数据或资料采集
文本数据
图片数据
声音数据
影像数据
数据或资料的预处理
数据压缩/降维
数据分类/聚类
数据的特征提取
……
模型设置与学习训练
有监督学习
无监督学习
……
数据的模型评估
拟合检验
交叉验证
指数评估
……

案例（一）：唐诗宋词创作

- 《如梦令》

（一）

无限长城古寺，剩有残碑废垒，荒草乱鸦啼，一片平芜烟里，凝睇，凝睇，望断碧云天际。

（二）

一寸柔肠千万，总是离愁难遣，何况鹿门前，只有鸳鸯湖畔，天远，天远，人在海棠花面。

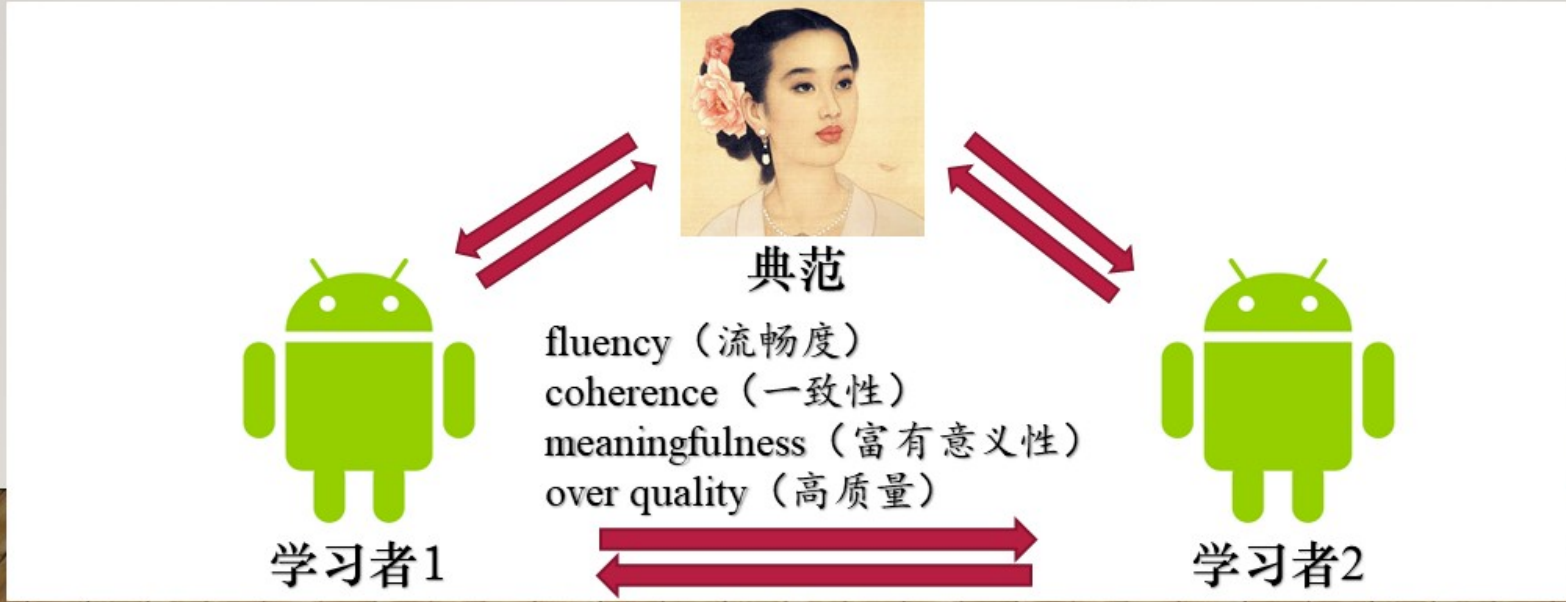
V.S.

案例（一）：唐诗宋词创作

- 九歌
  - https://jiuge.thunlp.cn/index.html
  - 人工智能诗歌写作系统，清华大学自然预言处理与社会人文计算实验室开发；
  - 数据来源：超过80万首人类诗词作者创作的诗歌；
  - 机器学习技能：采用深度学习（deep learning）方法，结合多个为诗歌生成专门设计的模型，目前可创作绝句、律师、藏头诗、词等作品。

案例（一）：唐诗宋词创作

- 九歌
  - *Automatic Poetry Generation with Mutual Reinforcement Learning*
  - 原理：相互强化学习模式（mutual reinforcement learning schema）



案例（二）：对央行行长经济口风的研究

- 林建浩等
  - 《如何测度央行行长的口头沟通信息》
  - 研究问题：央行行长的口头沟通信息通常被认为是中国经济基本状况的重要信号。但口头沟通信息作为一种非结构化信息，如何才能精确地度量和分析？
  - 数据来源：2003-2018年央行行长的口头沟通信息，包括演讲、访谈、采访、发布会报告等**非书面信息**。
  - 机器学习技能：通过对央行行长口头沟通信息的测度和学习，构建中国季度货币沟通指数和经济形势沟通指数，进而理解中央对经济政策的干预。

案例（二）：对央行行长经济口风的研究

- 数据来源
  - 百度新闻中检索“周小川”、“周小川+货币”、“周小川+经济形式”、“易纲”、“易纲+货币”、“易纲+经济形势”等关键词，获取媒体报道；
  - 搜集中国人民银行官方网页信息公开栏目；
  - 收集内部刊物《紫光阁》中记录的口头信息。

表 2　　**2003–2018** 年央行行长口头沟通渠道统计表

类别	讲话	采访	时报文章	新闻发布会
次数	334	90	22	20
占比	71.67%	19.31%	4.73%	4.29%

案例（二）：对央行行长经济口风的研究

- 机器学习步骤
  - 建立语料库
  - **货币政策议题**

表 3　央行行长口头沟通句子划分: 货币政策议题(部分句子)

宽松	2003/10/20: 要进一步加大公开市场操作力度，继续进行必要的窗口指导，保持贷款和货币供应量的适度增长。 2016/02/26: 鉴于中国经济和全球经济形势的看法，中国人民银行的货币政策是处于稳健略偏宽松的状态。
稳健	2003/10/20: 当前和今后一个时期，要继续执行稳健的货币政策，研究综合运用多种货币政策工具。 2014/06/11: 继续实施稳健的货币政策，不断完善调控方式和手段，增强调控的前瞻性、针对性和协同性。
从紧	2004/02/12: 从今年开始，中国将重点防止通货膨胀。 2008/04/12: 防止通货膨胀和经济过热仍然是重点工作，未来评估以利率，存款准备金率，以及公开市场操作等方式，来实行从紧的货币政策。

案例（二）：对央行行长经济口风的研究

- 机器学习步骤
  - 建立语料库
  - **经济政策议题**

表 4　央行行长口头沟通句子划分：经济形势议题（部分句子）

正面	2004/04/24：今年以来，中国经济继续保持快速增长的势头，投资、消费和出口强劲增长，工业企业利润大幅度增加，企业家和消费者对宏观经济走势信心持续增强。 2016/02/26：最新数据显示，1 月信贷需求强劲，春节黄金周全国销售同比增长 11.2%，表明中国经济出现积极迹象。
中性	2009/01/12：从生产角度来看，中国的表现仍很好，但需要确保采取有力政策以避免出现任何急剧放缓。 2016/07/24：上半年中国经济运行总体平稳，稳中有进，经济增速处于合理区间，物价和就业形势保持稳定。
负面	2012/04/22：目前看来，国际金融危机还没有平息，外部环境还要再予以观察，国内也面临经济趋缓和通胀压力并存问题。 2015/03/29：通货膨胀继续向下走，经济仍不景气。

案例（二）：对央行行长经济口风的研究

- 机器学习步骤
  - 生成计算词典

表 7　　短语频次统计表(部分短语)

短语	频次	货币政策			经济形势			其他	属性
		宽松	稳健	从紧	正面	中性	负面		
适度宽松的货币政策	18	15	0	0	0	0	0	3	宽松
降低社会融资成本	2	2	0	0	0	0	0	0	宽松
实施稳健的货币政策	32	0	31	1	0	0	0	0	稳健
优化信贷结构	7	1	4	1	0	0	0	1	稳健
实行从紧的货币政策	6	0	0	6	0	0	0	0	从紧
防止通货膨胀	7	0	1	6	0	0	0	0	从紧
较快增长	8	0	0	0	5	2	0	1	正面
扩大消费	7	0	0	0	7	0	0	0	正面
合理区间	6	0	0	0	2	4	0	0	中性
经济持续平稳发展	3	0	0	0	0	2	0	1	中性
下行	8	0	0	0	0	0	8	0	负面
放缓	23	2	0	1	1	5	14	0	负面

案例（二）：对央行行长经济口风的研究

- 机器学习步骤
  - 评估测度效果

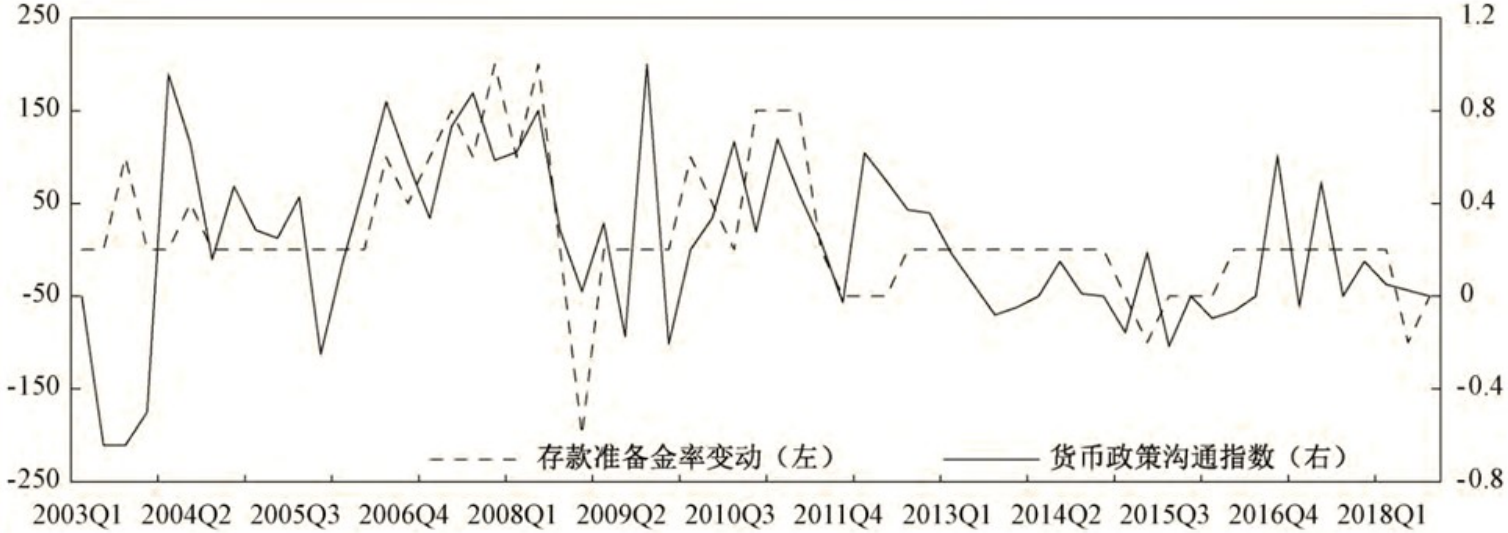


图 2　货币政策沟通指数与数量型货币政策干预对比图

案例（二）：对央行行长经济口风的研究

- 机器学习步骤
  - 评估测度效果

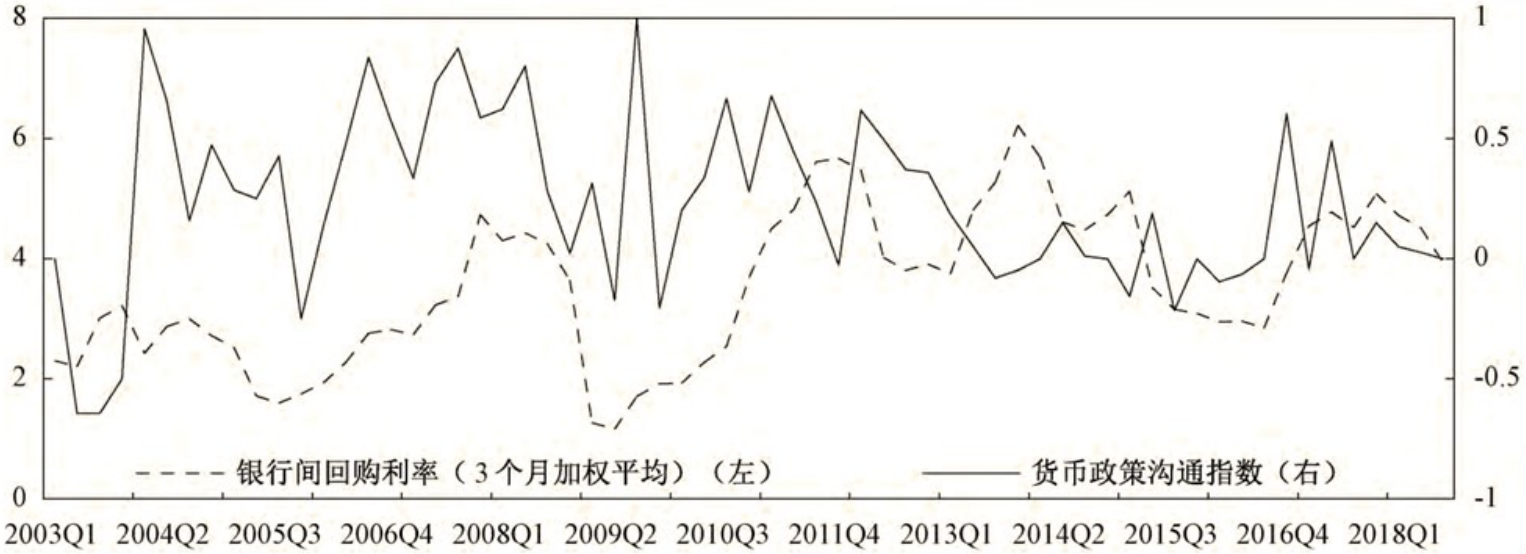


图 3　货币政策沟通指数与价格型货币政策干预对比图

案例（三）：国会议员与大众对社会议题意见的研究

- Pablo Barbara et al.
  - *Who Leads? Who Follows? Measuring Issue Attention and Agenda Setting by Legislators and the Mass Public Using Social Media Data*
  - 研究问题：过去理论认为，政治人物处理的问题大多是民众关心的问题。但对于具有立法功能性质的政治人物即国会议员而言，他们是追随民众意见，还是引导民众意见？是回应支持者意见，还是回应普罗大众意见？
  - 数据来源：美国113届国会议员在2013-2014年的65万条Twitter推文，包括参议院议员和众议院议员。
  - 机器学习技能：使用主题模型和向量自回归模型。

案例（三）：国会议员与大众对社会议题意见的研究

- 机器学习步骤
  - 训练数据集

Group	*N*	Avg	Min	Max	Tweets
House Republicans	238	1,215	70	8,857	267,311
House Democrats	207	1,177	113	5,993	222,491
Senate Republicans	46	1,532	73	6,627	67,412
Senate Democrats	56	1,616	150	10,736	87,307
Random sample	25k	465	1	8,926	11,316,396
Informed public	10k	948	100	5,861	9,487,382
Republican supporters	10k	1,091	100	8,804	10,911,813
Democratic supporters	10k	1,306	100	5,122	13,058,947
Media outlets	36	7,803	8	15,858	273,121

案例（三）：国会议员与大众对社会议题意见的研究

- 机器学习步骤
  - 社会议题的主题识别

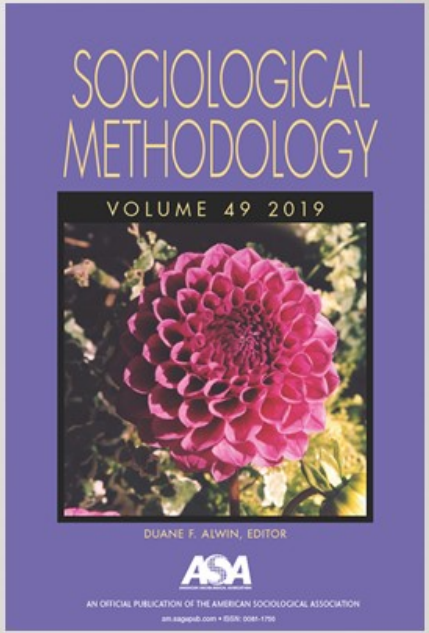
Topic Number	Label	Topic number	Label
3	Investigation of Benghazi attack	50	Climate change
7	100 days of #BringBackOurGirls campaign	51	Lame duck congress
9	Gender wage gap	53	Minimum wage
12	Republican issues *Spring 2013*	58	Affordable Care Act
14	Marriage equality	62	Border crisis in Texas
15	Gun violence	63	Obamacare (employer mandate)
16	Abortion (pro-life)	64	FAA furloughs cause flight delays
18	Veteran affairs delays scandal	66	Malaysia Airlines crash in Ukraine
20	NSA surveillance scandal	67	Comprehensive immigration reform
23	#BringBackOurGirls campaign	70	#MiddleClassFirst campaign
28	Employment Non-Discrimination Act	75	Military Justice Improvement Act
32	Islamic state	81	Poverty (SNAP program)
33	Use of military force in Syria	83	Twenty-first century cures initiative
36	Ebola	85	Unemployment insurance

案例（三）：国会议员与大众对社会议题意见的研究

- 最终识别结果
  - 国会议员倾向回应而非领导大众意见；倾向回应支持者而非普罗大众意见。

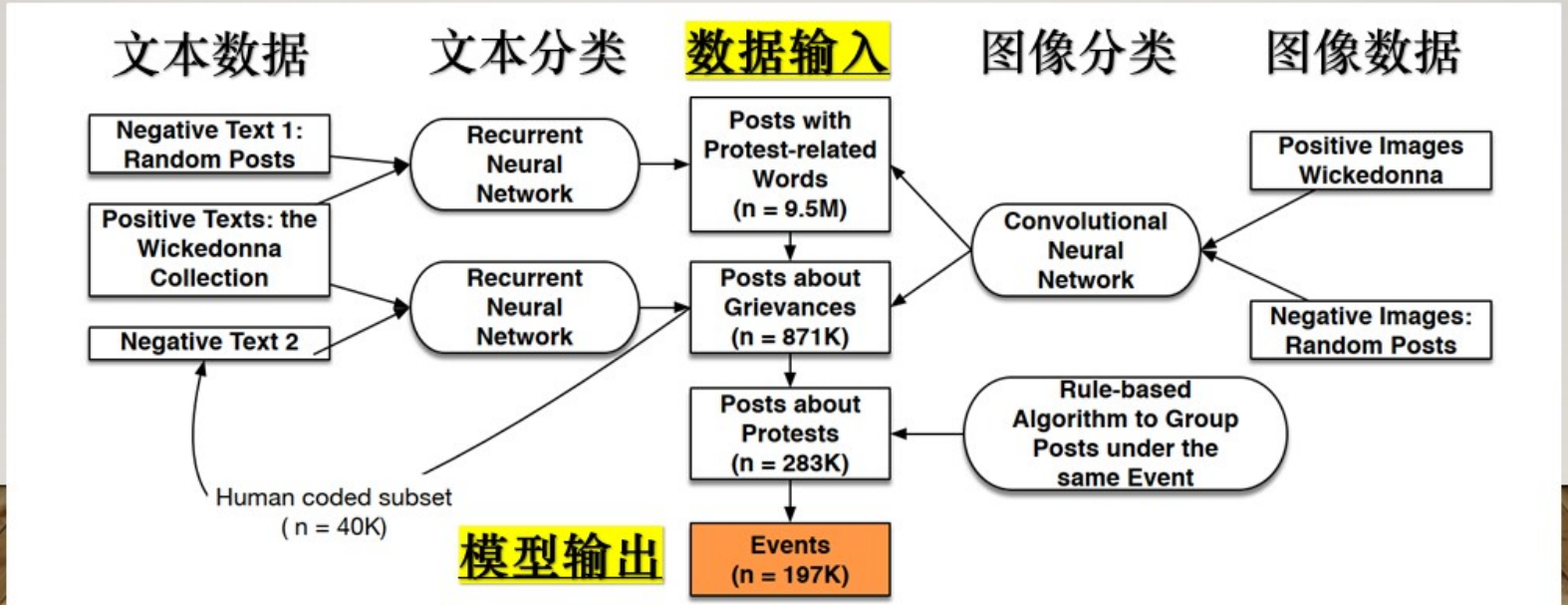
案例（四）：中国抗议性群体行动事件研究

- 张翰 & Jennifer Pan
  - *CASM: A Deep-Learning Approach for Identifying Collective Action Events with Text and Image Data from Social Media*
  - 研究问题：抗议性群体行动事件（protest collective action events）是一种常见的社会运动现象。但传统媒体对这类事件的态度往往带有偏见和偏差，更倾向于报道更大规模和更耸人听闻的事件。如何改善这些偏见造成的数据缺失？
  - 数据来源：2010-2017年新浪微博950万个帖子和360万张图片
  - 机器学习技能：利用回归神经网络（recurrent neural networks）和积卷神经网络（convolutional neural networks）识别超过10万个抗议性群体行动事件。



案例（四）：中国抗议性群体行动事件研究

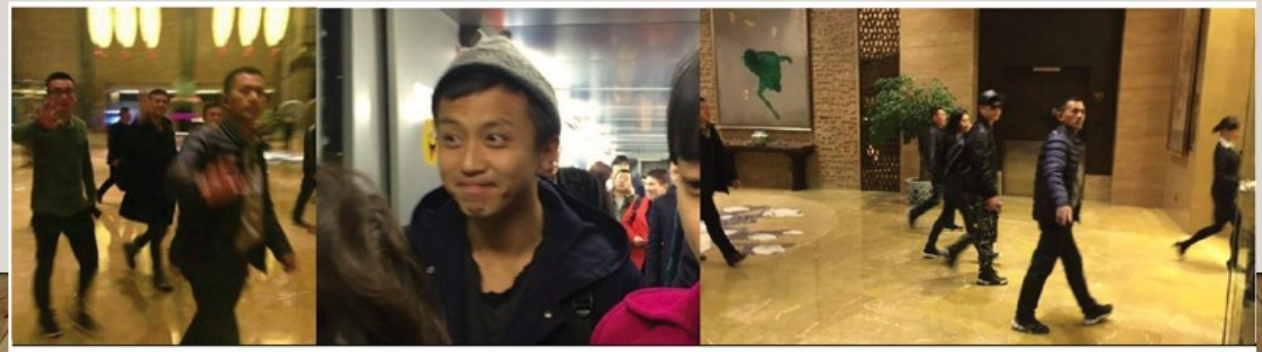
- 正向训练
  - 训练数据集：Yuyu Lu & Tingyu Li于2013年6月-2016年6月手工编码的抗议性群体行动事件数据库Wickedonna，包含24万条微博和23万张图片；



案例（四）：中国抗议性群体行动事件研究

- 负向训练
  - 正向训练的结果

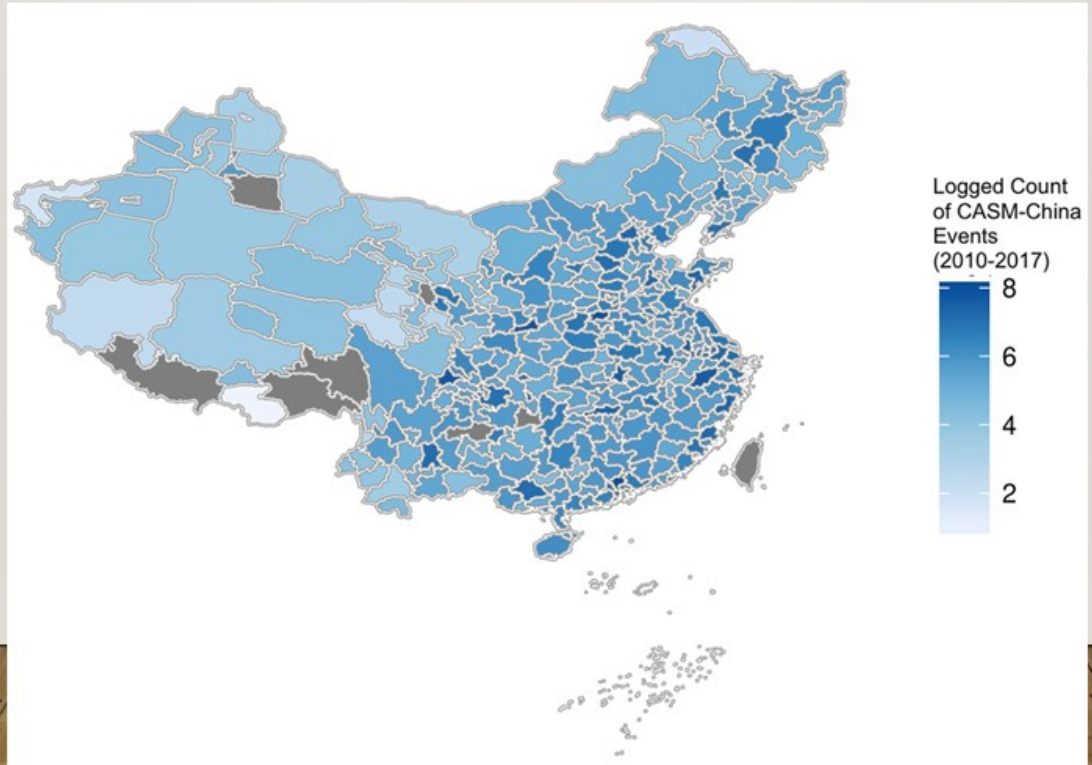
  - **用负向训练调整误识别问题**



奔跑吧兄弟 2 成都开录邓超 baby 李晨等遭围堵.

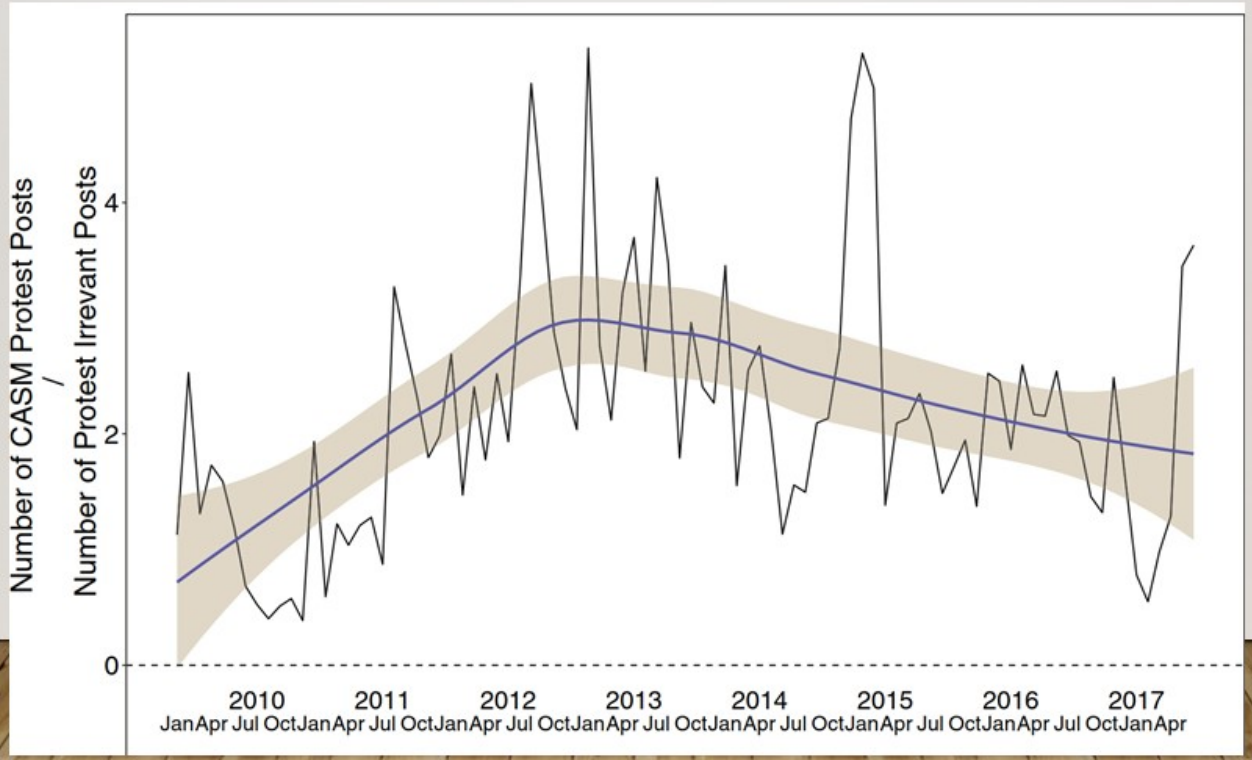
案例（四）：中国抗议性群体行动事件研究

- 最终识别结果



案例（四）：中国抗议性群体行动事件研究

- 最终识别结果



机器学习应用于社会科学的潜在问题