# 从舍恩伯格“大数据带来的三个转变”谈起

**刘　挺**

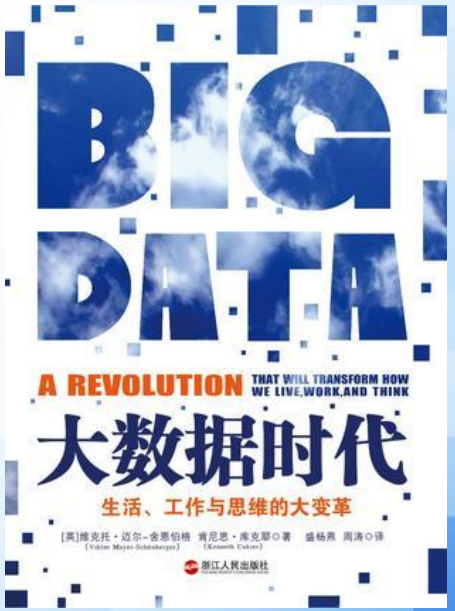
哈工大计算机科学与技术学院

2014年4月18日，贵阳

# 大数据带来三个转变

- 牛津大学教授舍恩伯格提出的三个转变
  - 第一，可获得并处理和某个特别现象相关的所有数据，而不再依赖于随机采样
  - 第二，不再热衷于追求精确度，而容忍混杂
  - 第三，不再热衷于寻找因果关系，转而寻找事物之间的相关关系

维克托 迈尔-舍恩伯格



# 对三个转变的疑问

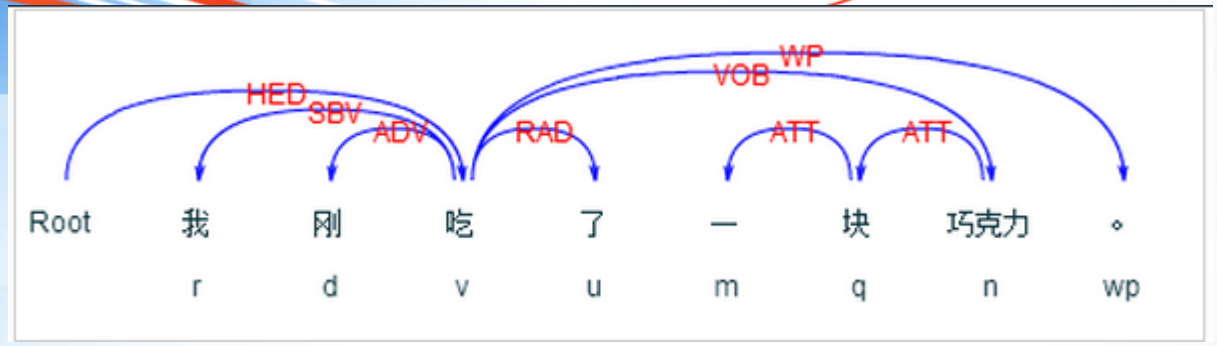
- 采样，还是全集?
- 精准，还是混杂?
- 因果，还是相关?

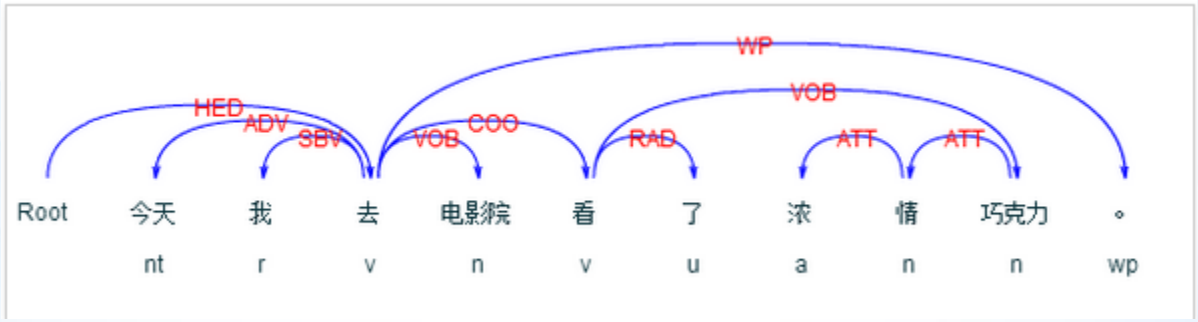
# 有全集吗？需要全集吗？

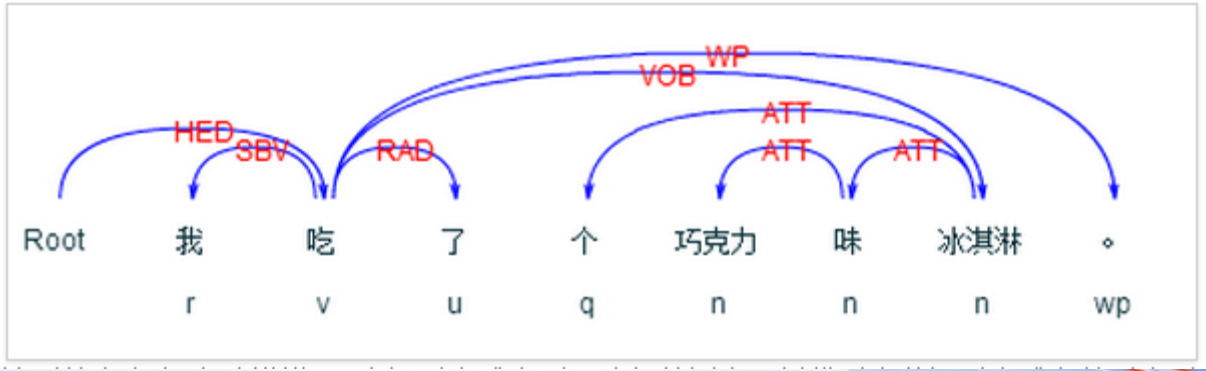
- 研究一个问题，有多大概率能够拿到全集？
- 数据不可避免的缺失
  - 比如，要采集“随着一个话题的传播，多少人取消了对一位“公知”的关注？”，对于过去的话题，已经无法采集到上述信息
- 有没有必要拿到全集？

# 计算社会语言学：基于社会媒体的民俗调查
# 八维饮食地图：http://ys.8wss.com

# 应用依存句法分析技术

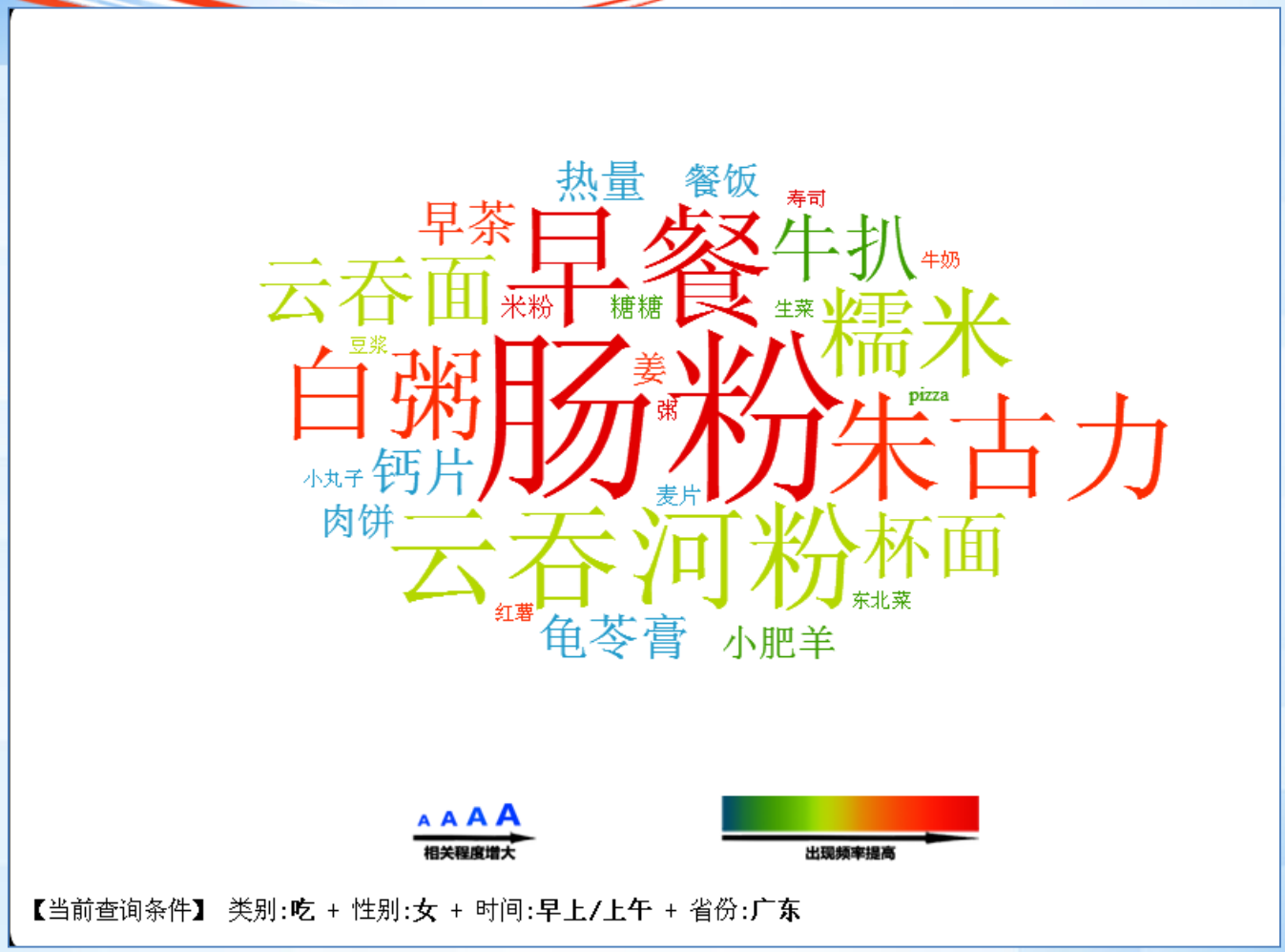

HED
SBV
ADV
RAD
WP
VOB
ATT
ATT
Root 我 刚 吃 了 一 块 巧克力 。
r d v u m q n wp


WP
VOB
HED
ADV
SBV
VOB
COO
RAD
ATT
ATT
Root 今天 我 去 电影院 看 了 浓 情 巧克力 。
nt r v n v u a n n wp


WP
VOB
ATT
ATT
ATT
HED
SBV
RAD
Root 我 吃 了 个 巧克力 味 冰淇淋 。
r v u q n n n wp

# “广东女人早上吃什么？”

## 采样偏置：这里的“广东女人”=“早晨爱发微博的广东女人”



# 样本误差和采样偏置

## 样本误差

- 样本误差是指一组随机选择的样本可能无法真实地反映全部现象
- 样本误差的幅度，会随着样本数量的增加而减小

## 采样偏置

- 样本可能根本就不是随机选择的
- 1936年，美国《读者文摘》预测总统大选失败，原因：从车辆注册信息和电话号码簿里选择问卷对象，该群体偏富裕阶层
- 2013年，美国的Twitter中年轻的，居住在大城市或者城镇的，黑色皮肤的用户比例偏高

## 结论：

- “很少但很有代表性”，比“很多但很偏”要好

# 几个方法论上的疑问

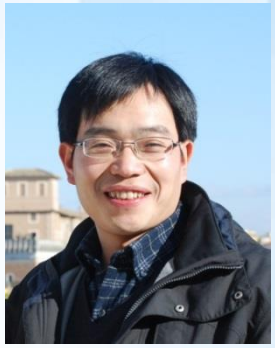
- 采样，还是全集?
- 精准，还是混杂?
- 因果，还是相关?

# 社会科学数据采集的主要方法

| 传统社会科学方法 | | 社会化计算方法 |
|---|---|---|
| 定量数据 | 定性数据 | |
| 抽样调查 | 个人或小组访谈 | 日志分析 |
| 心理学实验 | 观察或个案研究 | -- |
| 内容分析 | 文本分析 | 网页挖掘 |
| 约占三分之二 | 约占三分之一 | 约占1-2% |

香港城市大学
祝建华教授

# 社会学严格的数据需求



中国人民大学
冯仕政教授

- **代表性**：能够有效地代表所欲研究的社会总体
- **精确性**：是社会特征的真实反映，并且尽可能精确
- **系统性**：尽可能取得理论所关切的所有变量，减少缺失值
- **类型化（categorization）**：根据社会特征和理论关切将数据标准化，以便统计分析

# 社会媒体为社会学研究提供数据

| | 问卷调查 | 社会媒体采集 |
|---|---|---|
| 数据质量 | 高：精密、干净 | 低：富含噪声、混杂 |
| 完整性 | 完整 | 不完整 |
| 可信性 | 高 | 低 |
| 采集成本 | 很高 | 低 |
| 样本偏置 | 可控 | 比较严重 |
| 数据量 | 小 | 巨大 |
| 覆盖面 | 有限 | 广 |

社会媒体帮助社会学以更“广大精微”的视角观察社会

# 问卷调查就可靠吗?

- 问卷调查都能接触到什么人?
- 问题：谁是研究的重点?
  - 普罗大众?
  - 精英分子?
- 如果重点是精英分子，问卷调查有效吗?
- 历史是谁创造的?

# 现实瓶颈

- 难以取得符合上述要求的数据，以致很多社会学所关心的研究议题难以开展
- 困境的三个方面：
  - 技术上的
  - 法规上的
  - 伦理上的
  - 政治上的
    - 与“细胞”和“词汇”不同，“人和社会组织”会抗议

# 容忍误差?

- 多重误差
  - 一手数据本身的误差
  - 采集过程中的数据缺失
  - 自动分析工具不准确带来的误差（比如：倾向性分析）
- 与社科学者精耕细作的传统不符
  - 传统：准确率为95%
  - 网络数据自动处理：80%，能接受吗?

# 有误差的数据，能够带来正确的结论吗?


汽车口碑分析系统——爱搜车

数据有误差，处理有误差，

但结论是对的，可能吗?

1. 原始数据有5%的误差
2. 情感倾向性分析技术正确率70%
3. 但最终结论与用户的直观感觉一致：“**标致307的内饰比福克斯好**”

# Google根据用户的搜索词预测流感趋势

- Detecting influenza epidemics using search engine query data
- 2009.2，Nature
- 被引用1000余次
- 大数据的经典案例

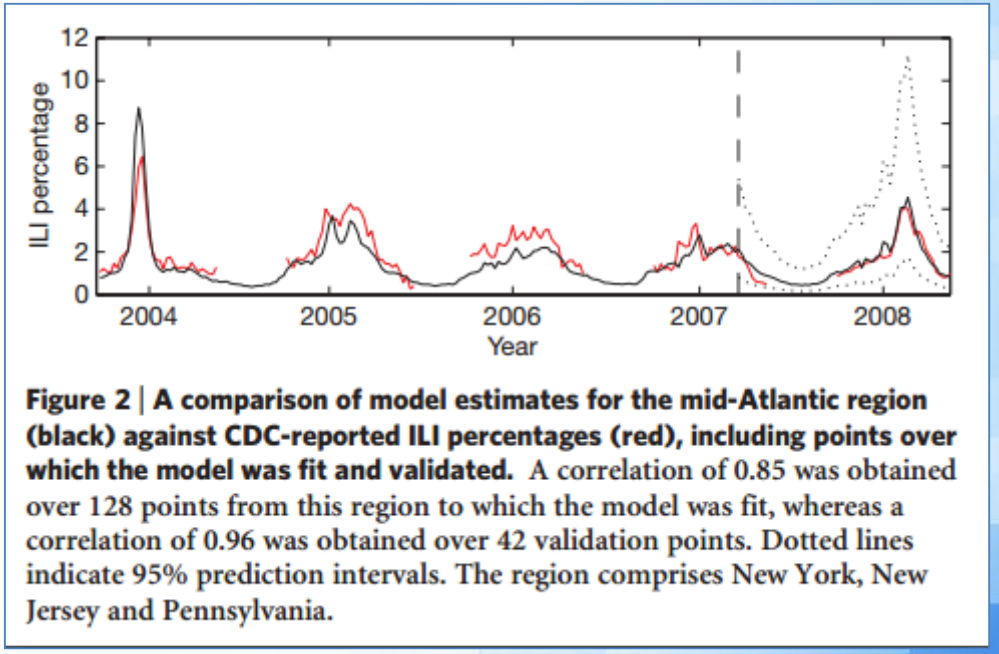
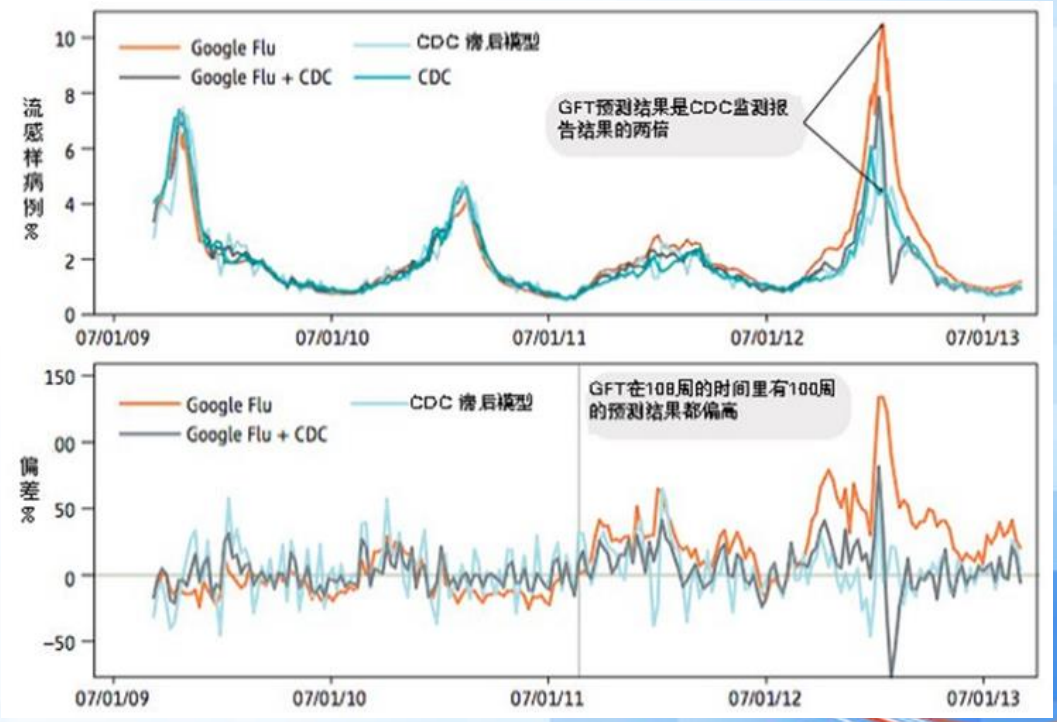



**Figure 2 | A comparison of model estimates for the mid-Atlantic region (black) against CDC-reported ILI percentages (red), including points over which the model was fit and validated.** A correlation of 0.85 was obtained over 128 points from this region to which the model was fit, whereas a correlation of 0.96 was obtained over 42 validation points. Dotted lines indicate 95% prediction intervals. The region comprises New York, New Jersey and Pennsylvania.

# Google预测的流感病例数超过了美国疾病预防控制中心预测结果的2倍

- 最近该项研究受到质疑
- The Parable of Google Flu: Traps in Big Data Analysis
- David Lazer
- 哈佛大学
- 2014.3



# Google流感预测失准的原因

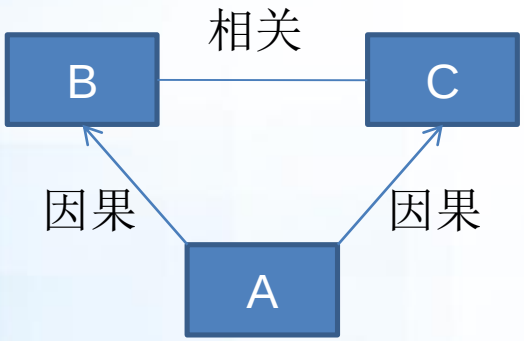
- “大数据傲慢”(Big Data Hubris)
  - 即认为大数据可以完全取代传统的数据收集方法，而非作为后者的补充
  - 很多关键词只是看似与流感相关，但实际上却并无关联
- 算法变化（Algorithm Dynamics）
  - 媒体上充斥着各种关于流感的骇人故事，看到这些报道之后，即使是健康的人也会跑到互联网上搜索相关的词汇
  - 人们输入病症时，Google推荐的一些诊断结果会影响用户的搜索行为，例如：搜索“发烧”时，会推荐相关词“流感”

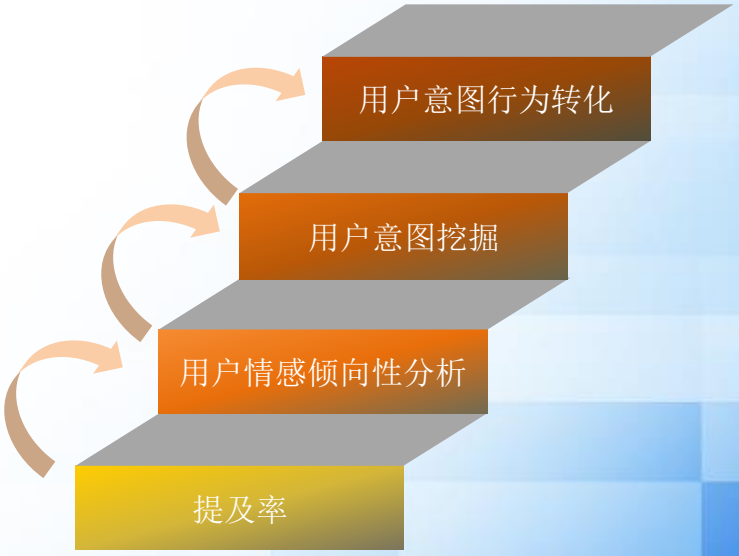
# 几个方法论上的疑问

- 采样，还是全集?
- 精准，还是混杂?
- 因果，还是相关?

# 电影票房预测

- 基于社会媒体的电影票房预测（相关）
  - 微博提及率、消费意图、消费意图转化率
  - 排片数
  - 导演知名度、主演知名度



# 消费意图正例

胖青Catharine：<一九四二>，考完试一定去看，有期待的感觉真好。

27分钟前　来自摩托罗拉智能手机　　转发 | 收藏 | 评论

核信-Lee：我想去看《一九四二》，看过原著才知道原来老家还有这样的一段历史，更想看电影，电影更直观。

思凡隊長請帶我神遊：mark一下想看的电影《一九四二》《一次别离》《少年派的奇幻漂流》《血滴子》~

今天15:22　来自新浪微博　　转发 | 收藏 | 评论

裳依依：看吴君如专访听她说陈可辛如何舍不得扔掉他和女儿小时候的东西，即使其他东西没地方放也不舍得扔掉那些没用的，是一个心思异于常人细腻的人。我明白了为什么有的导演能拍出那么震撼人心的电影，唐山大地震和一九四二这种只要看过就会被深深震撼的电影，有人说不太敢看一九四二，而我却以为就需要看看。

今天14:30　来自360安全浏览器　　转发 | 收藏 | 评论

十一月的宁静：#十一月电影抢先看# 11月有好多片子想看捏。梁家辉、郭富城、刘德华主演的《寒战》；冯小刚执导、陈道明张国立葛优出演的电影《一九四二》；还有《少年派的奇幻漂流》~~ http://t.cn/zl3JaXs

今天13:35　来自微话题　　转发 | 收藏 | 评论

蕾妈的碎碎念：从现在到2012年末，全心全意等待的电影只有一部：冯小刚的《一九四二》。

今天12:03　来自新浪微博　　转发 | 收藏 | 评论(2)

# 消费意图反例

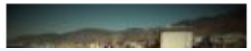
白宇IR：关注//@刘挺:《一九四二》7亿票房？ @water @丁效SCIR

@凌平V：【今年贺岁档本土电影票房预测】冯小刚《1942》7亿以上；成龙《十二生

LISA_两点一线：冯小刚执导的影片《一九四二》总投资额2.1亿元的经历了18年沉淀酝酿、8个月精心筹备、5个月艰难拍摄及8个月的后期制作后，于罗马时间11月11日22时迎来了它的第一批观众。11月29日，这部电影将在中国的8000块银幕同时上画。

今天14:04 来自新浪微博 转发 | 收藏 | 评论

小谢宇：在双廊休假带了两本书，一本@丁丁张 的人生 需要揭穿。一本刘震云精选集，里面有温故一九四二 两天争取看完



babulo紫蓟：≪一九四二≫罗马电影节首映，全体起立鼓掌。无国界的好才是真的好

今天11:19 来自Weico.iPhone 转发 | 收藏 | 评论(3)

弄十三：期待一代宗师，不期待一九四二，期待许巍新专辑，不期待周董新专辑。

今天10:40 来自新浪微博 转发 | 收藏 | 评论(2)

大小姐：期待《温故一九四二》！

@苏有朋北京后援会：@蘇有朋 宋子文，期待你给我们下一个惊喜！

# 电影预测失效的原因探究
# 八维票房预测：http://yc.8wss.com


- 《一九四二》
  - 首周预测1.5亿，实际1.38亿
  - 总票房预测5亿，实际不足4亿
- 原因分析
  - 在贺岁档上演灾难片，时机不对
  - 《少年派》和《泰囧》的前后夹击
  - 冯小刚爆粗口，羞辱观众
  - 3D银幕的争夺，引发《少年派》观众对《一九四二》不满

# 对因果分析的价值

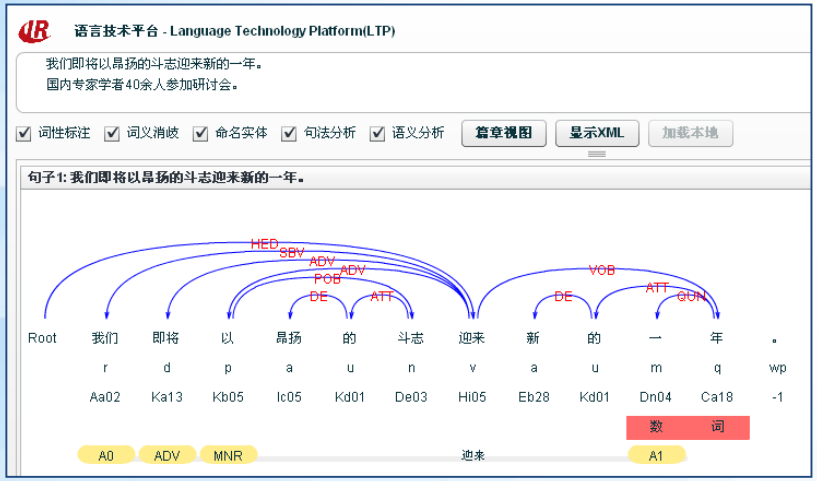
- 不知道相关性背后的原因，无法得知这种相关性在什么情况下会消失
- 更深入地理解社会现象，预测社会活动趋势
- 易于给出解决方案，干预社会活动的进程
- 克服数据稀疏
- 问题是：如何分析因果关系?

# 对三个疑问的一点儿结论

| 方法系列1 | 方法系列2 | 点评 |
| --- | --- | --- |
| 采样√ | 全集 | 很难拿到全集，采样仍然必要 |
| 精准 | 混杂√ | 必须接受混杂而真实的数据，同时必须提高数据分析精度 |
| 因果√ | 相关 | 对因果的探讨，很必要 |

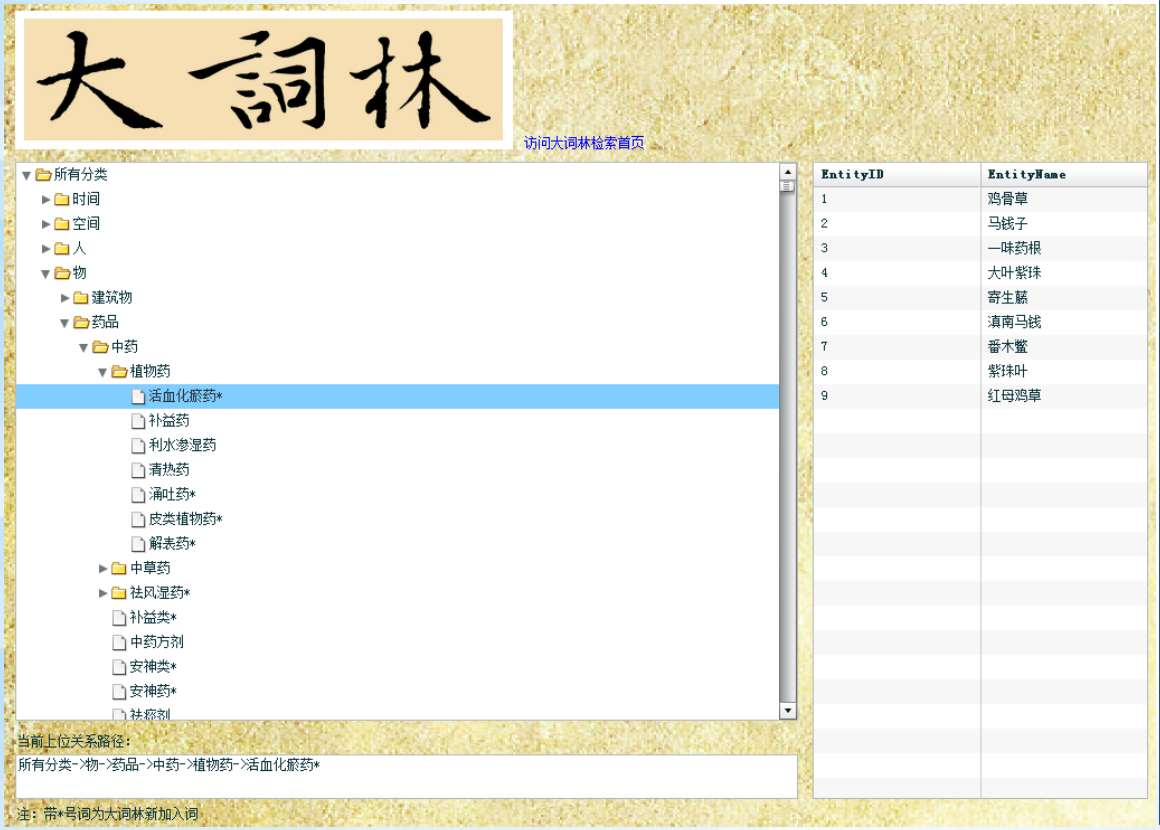
# 语言云(LTP-Cloud, Since 2013.9)

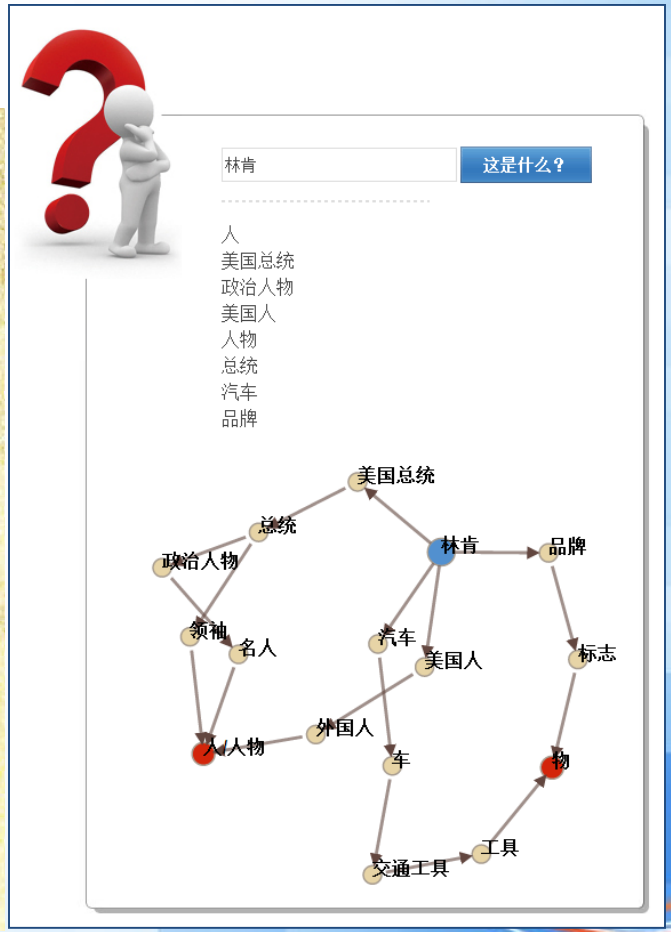
- 全称“语言技术平台云”
  - 基于云计算的中文自然语言处理服务平台
  - http://www.ltp-cloud.com/
  - 有来自世界各地1,100余名用户注册使用“语言云”
  - 日均处理请求近百万次。




# 大词林(http://www.bigcilin.com)

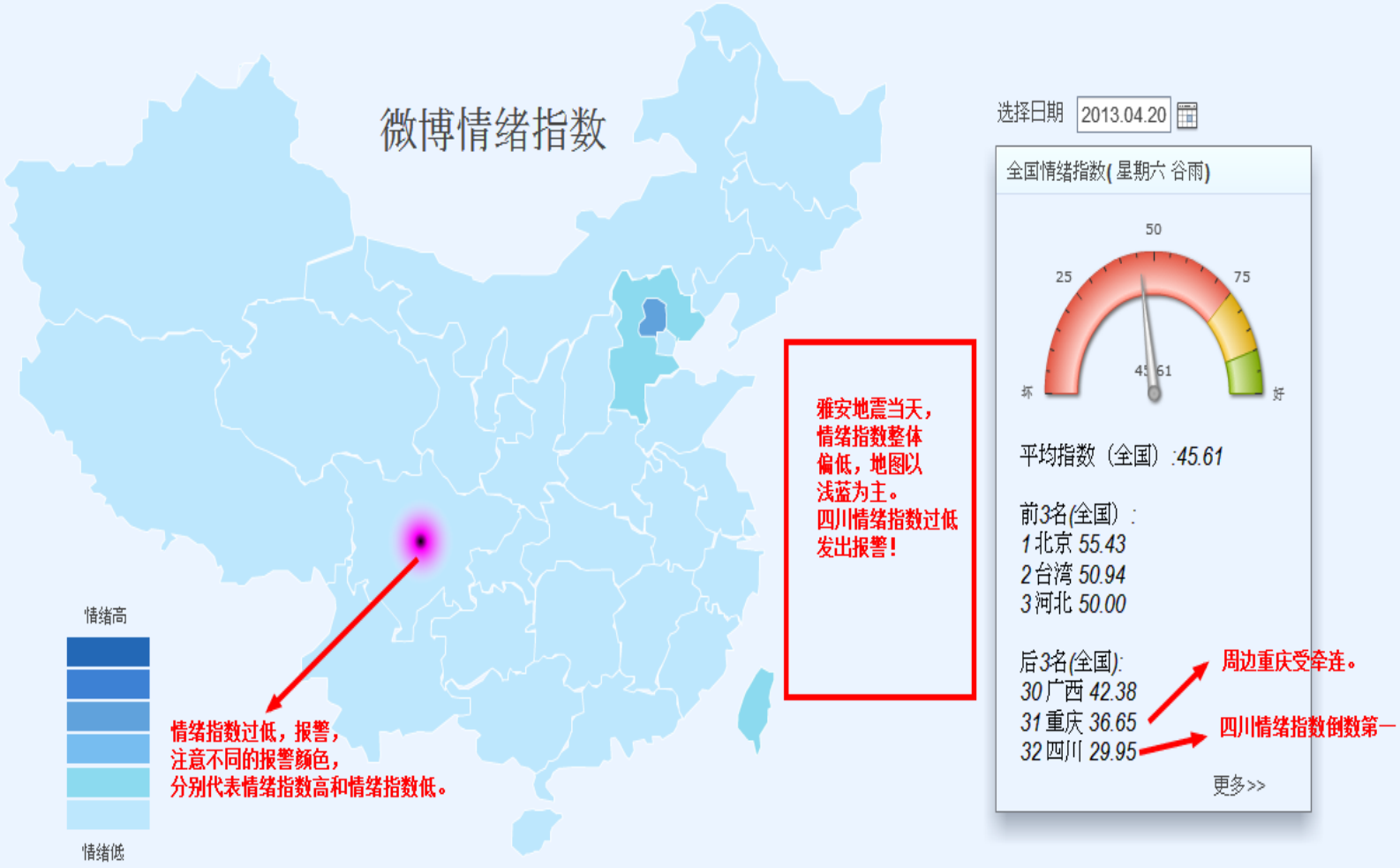
- 两种关系：上下位、同义
- 88万多词条



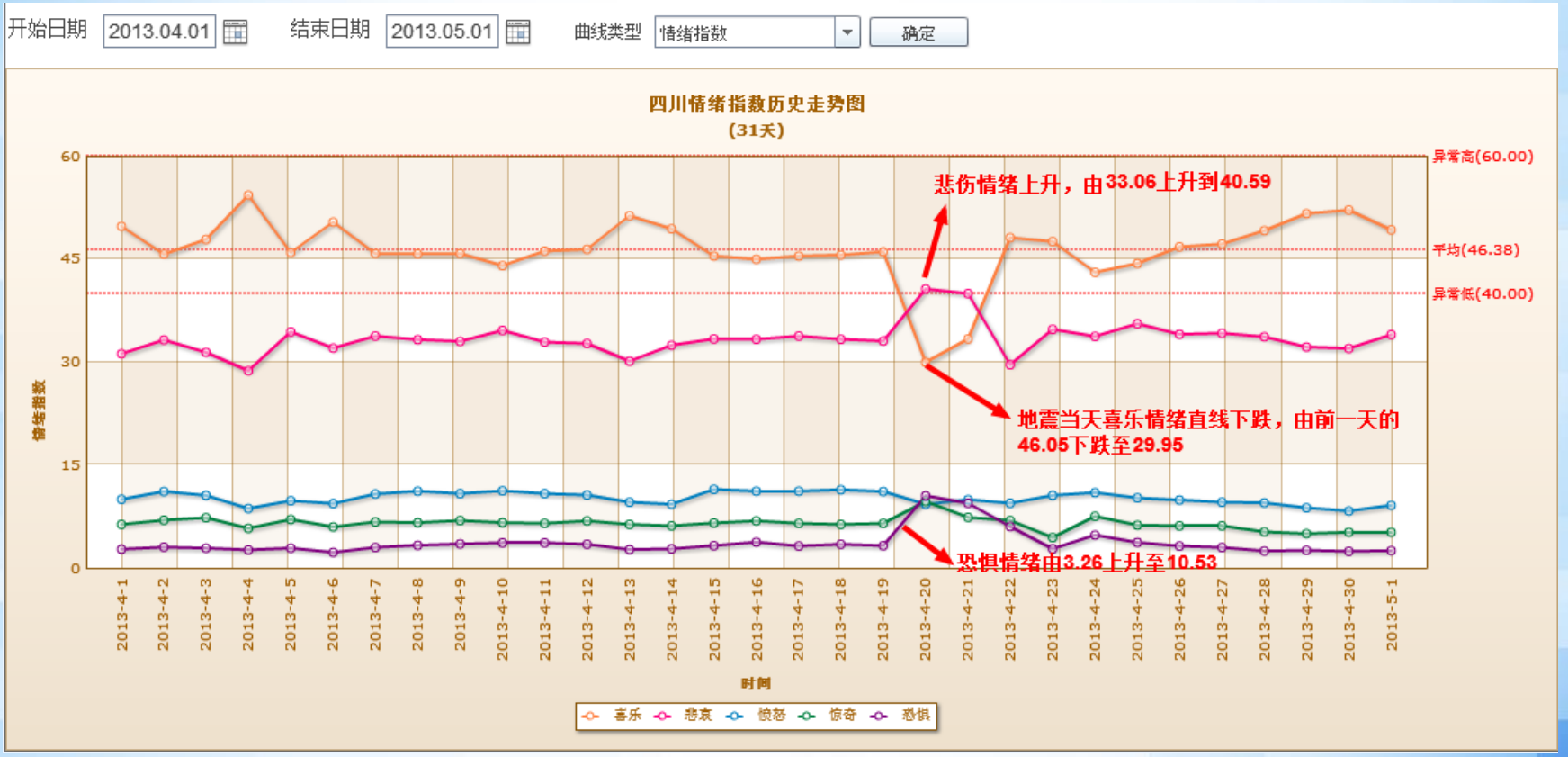



# 2013年4月20日四川雅安地震当天全国情绪分布

## 哈工大情绪地图： http://qx.8wss.com



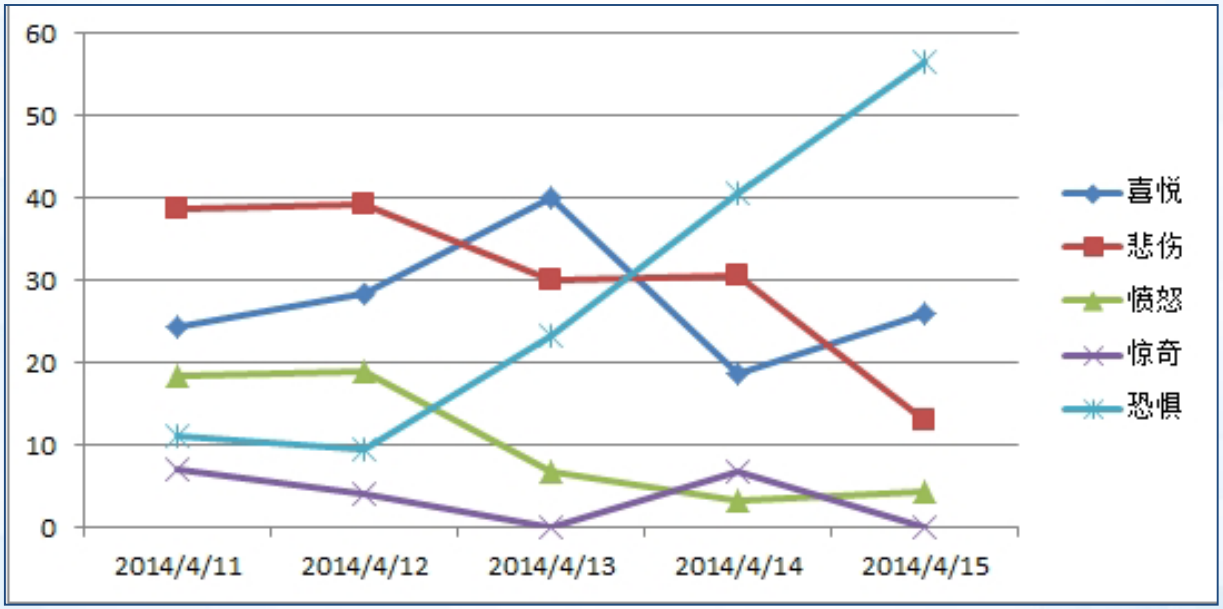
# 雅安地震前后四川省微博用户情绪的变化




五种情绪：喜、怒、悲、恐、惊

# 围绕热点事件的公众情绪变化曲线

- “兰州自来水苯超标”事件



# “社会媒体与语言处理研讨会”2012.12.8


- 徐
- 张华平，北京

软亚洲研究院，数据库

所，复杂系统

# “社会媒体处理

## 大会赞助商

### 钻石赞助商

北京拓尔思信息技术股份有限公司

### 金牌赞助商

海量信息技术有限公司

360互联网安全中心

### 银牌赞助商

北京微众文化传媒有限公司

数据堂（北京）科技有限公司

北京宏博知微科技有限公司

### 实物赞助商

新浪微博

## 参会单位

**成果名称：微博博主特征与行为大数据挖掘分析**

参展机构：北京理工大学
参展形式：口头报告+展台展示
联系人：张华平（kevinzhang@bit.edu.cn）

**成果名称：天玑学术网**

参展机构：中国科学院计算技术研究所

参展形式：口头报告+展台展示
联系人：刘知远

敬请各位专家批评指正！