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

《大数据分析导论》

教师简介 - 文继荣

中国人民大学信息学院教授、院长、博士生导师，大数据管理与分析方法研究北京市重点实验室主任，国家“千人计划”特聘专家。主要研究方向是大数据管理和分析、信息检索、数据挖掘和机器学习。自1999年起在微软亚洲研究院工作14余年，获50多项美国专利，在国际著名会议和期刊上发表论文100余篇，被同行引用12000余次，担任多个国际会议和期刊的程序委员和编委。

教师简介 - 窦志成

http://www.playbigdata.com

玩转大数据

参与了大量数据挖掘与信息检索、文本分析的项目和研究，拥有丰富的相关科研和项目经验

课程助教

- 秦绪博（博士一年级）
- 朱余韬（硕士二年级）
- 李娟（硕士二年级）
- 卢淑琪（本科三年级）

- 帅照在哪里？

《大数据分析导论》概览

- 一门面向文科生的大数据课程
- 培养文科生的大数据思维
- 让文科生也能动手做简单的大数据分析课题

大数据时代

- 大数据将成为未来主流计算平台

数据 + 计算能力

大数据+人文社科

- 大数据分析技术逐步开始应用在
  - 政府治理
  - 市场营销
  - 智慧司法
  - 股市预测
  - 企业管理
  - …

大数据+人文社科

- 大数据计算平台和架构

大数据+人文社科

- 大数据将成为文科应用和研究的新模式
  - 人文社会科学本质上是关于人的科学。是研究人与社会的行为规律的。
  - 各种APP和互联网+应用的出现，让人的行为可以通过数据记录下来
  - 大数据使得“人”变得可以精确刻画、可以量化度量的对象

大数据+人文社科

- 关于你的数据正在被采集

大数据+人文社科

- 社会数据

中国裁判文书网
China Judgements Online

国家企业信用信息公示系统
National Enterprise Credit Information Publicity System

东方财富网 eastmoney.com | 数据中心

国务院发展研究中心 信息网

政府工作报告

大数据+人文社科

- 大数据将成为文科研究新模式
  - 大数据的运用，可突破常规的基于人工分析、人工调研、少量案例分析等方法的局限性
  - 为人文社科问题的研究提供新思路

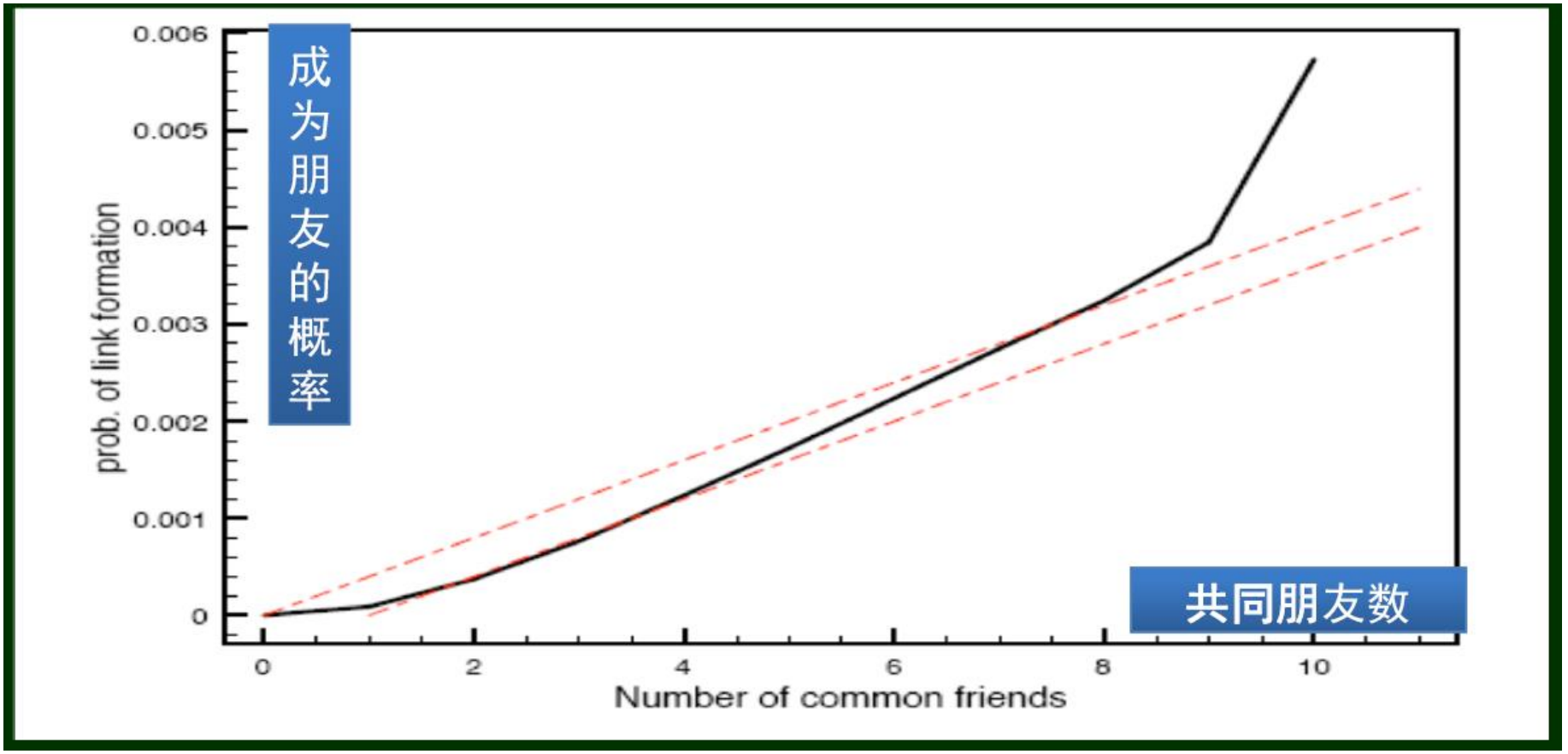
- 突破常规的思考：
  - 这个问题，能否从定性变成定量?
  - 这个问题，能否用大数据来验证?
  - 这个问题，是否能用大数据的方法进行分析?

计算社会学-三元闭包

- 三元闭包：如果两个互不认识的人有了一个共同的朋友，则他们两个将来成为朋友的可能性提高

- 如何用大数据来验证?
  - 定性-》定量：如果两个互不认识的人的共同朋友数越多，则他们两个在未来成为朋友的可能性越大
  - 选择合适的数据：某所大学的2万学生的通讯关系数据
  - 2006年， Science

计算社会学–三元闭包

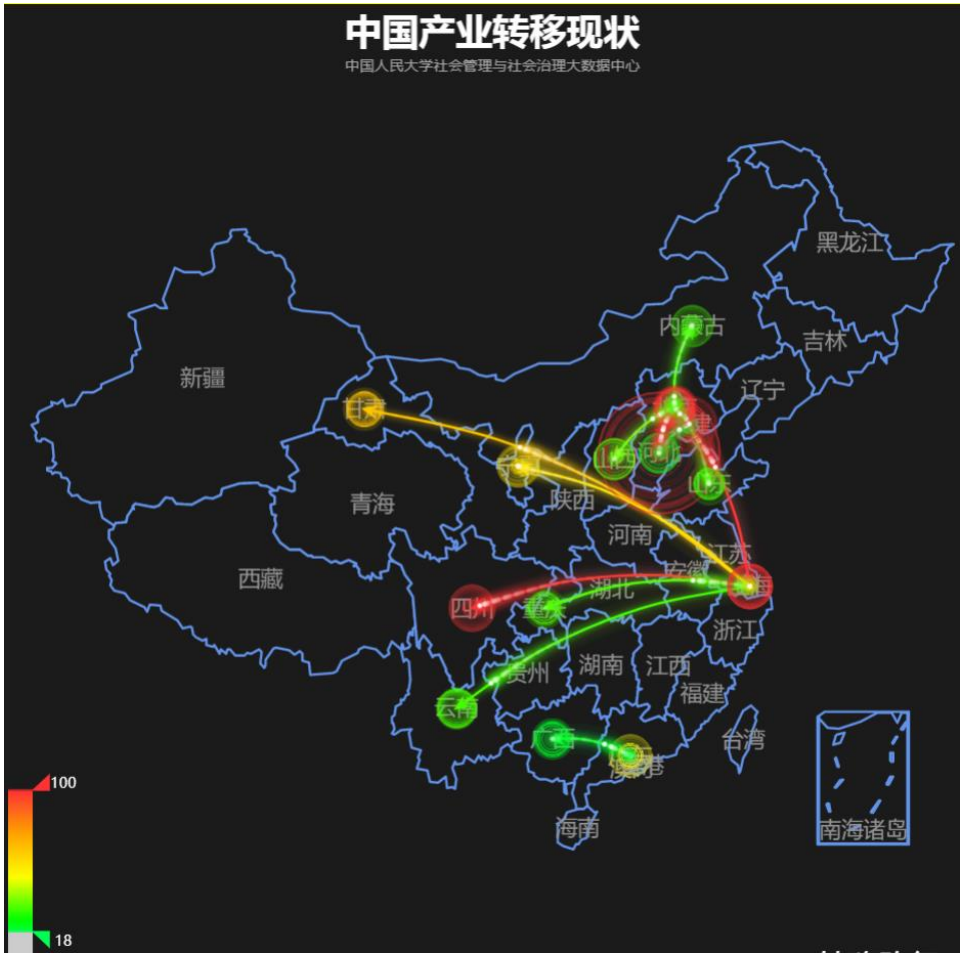

成为朋友的概率
prob. of link formation
0.006
0.005
0.004
0.003
0.002
0.001
0
0
2
4
6
8
10
Number of common friends
共同朋友数

中国产业转移现状

- 产业转移

目前北京的动物园批发市场已基本向天津西青、燕郊等地搬迁完毕，原有区域已成为科技、金融和公共服务的高地。

产业转移：
源地：北京
目的地：天津、河北

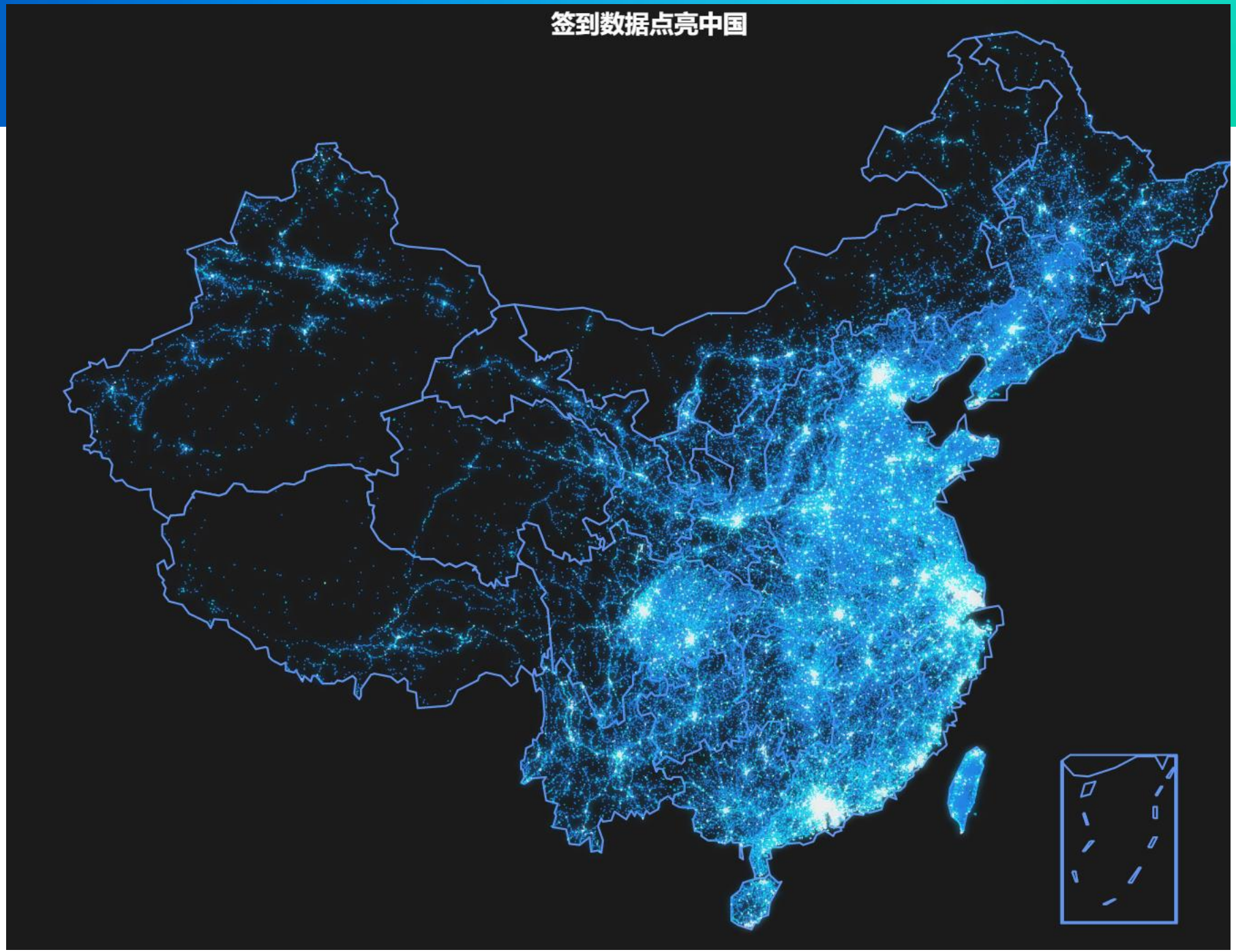


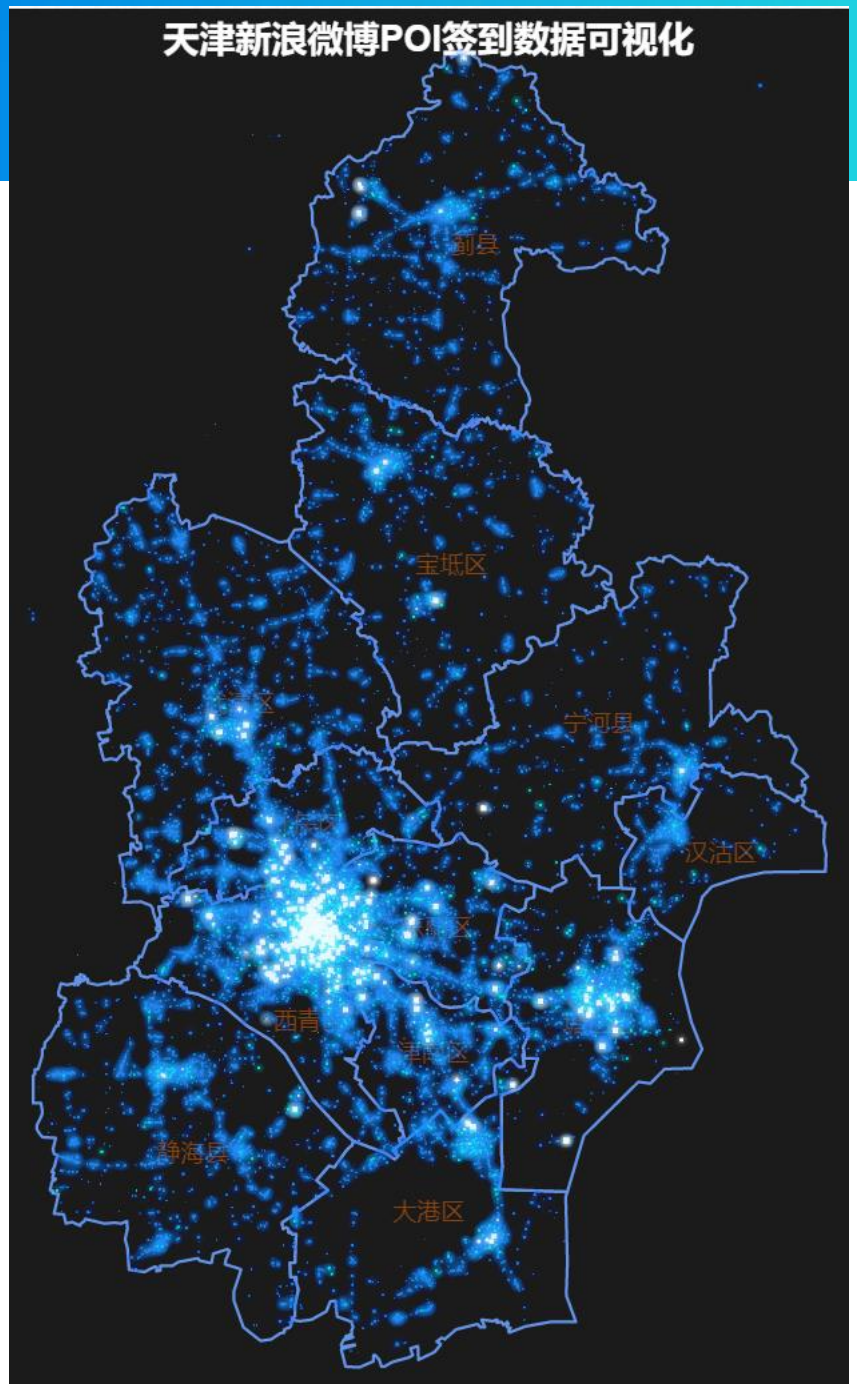
http://www.playbigdata.com/chanyeqianyi/

大数据+人文社科


全国网络扶贫行动大数据分析平台
国家级贫困县分布
显示贫困县
网络扶贫-推进力度
全国
东西部协作扶贫
贫困县通宽带情况
固定/移动宽带用户数
网络扶智
贫困县电商销售情况
农特产品上行销量(万元)
全国贫困县-优势资源
15496
0
南海诸岛
推进力度
网络覆盖
农村电商
网络扶智
信息服务
网络公益
832
贫困县
120亿
2016年 农特产品上行销量
1.20亿
2016年 移动宽带用户数
2,734万
2016年 固定宽带用户数
25
东西部协作省份
659
东西部协作区县
4574
共建共享联盟学校
223
双师学校

签到数据点亮中国




天津新浪微博POI签到数据可视化
蓟县
宝坻区
宁河县
汉沽区
西青
静海县
大港区

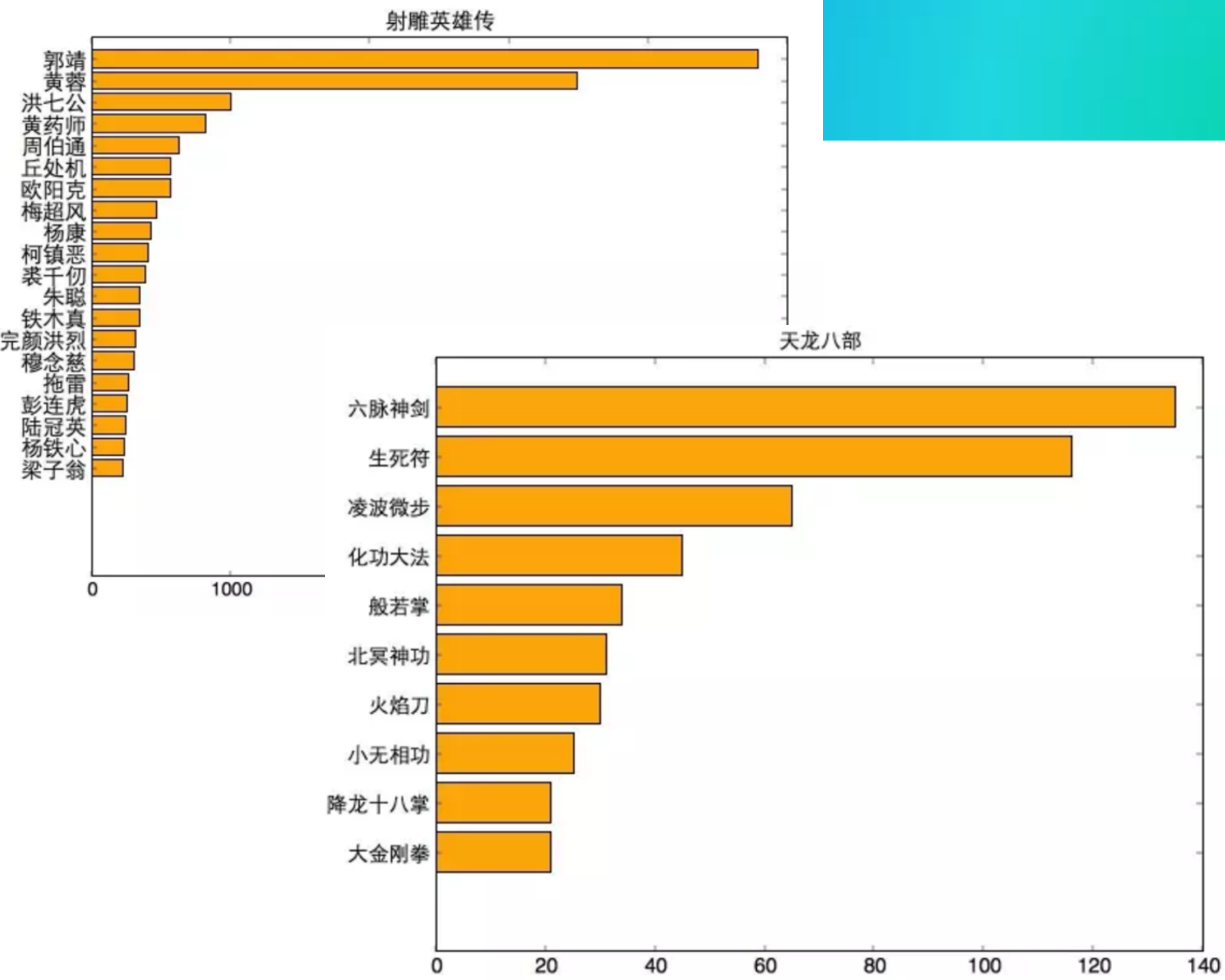
唐宋文学编年地图

- 中南民族大学教授王兆鹏主持
- http://sou-yun.com/poetlifemap.html

读武侠小说（来自微信）

- 机器学习：用Python读金庸武侠小说（1）侠之大者，为国为民之人物篇
- 机器学习：用Python读金庸武侠小说（2）一统江湖、称霸武林之门派、武功篇

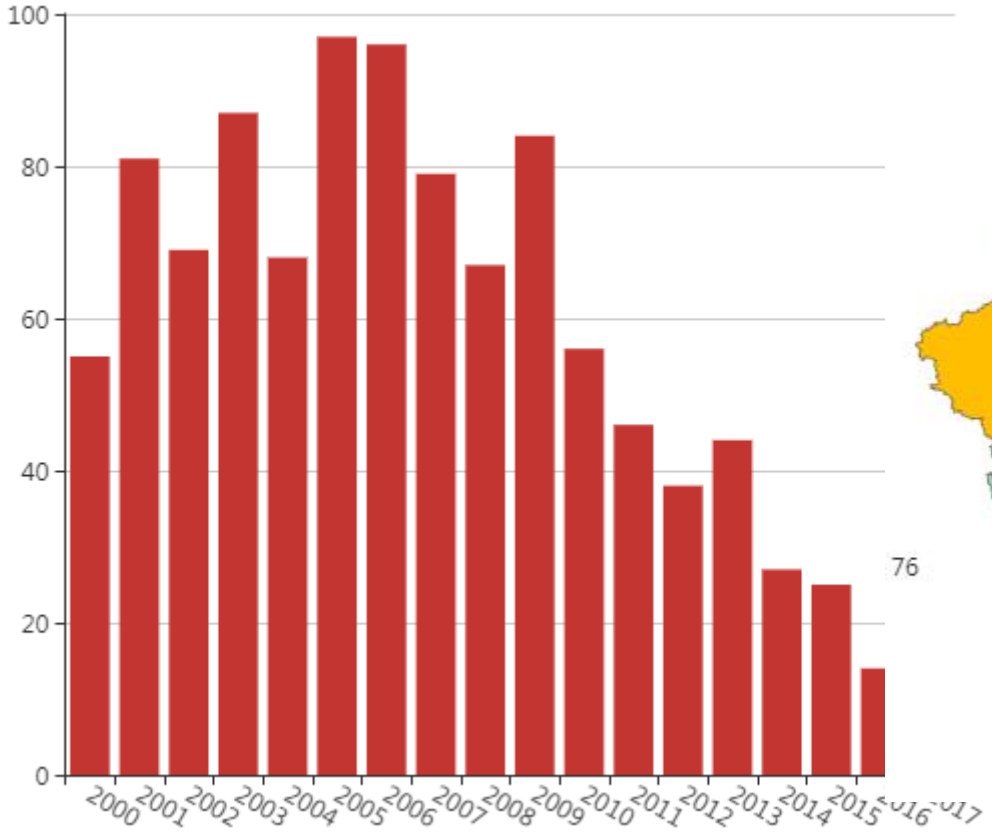
侠之大者，为国为民：角色出场次数排行榜。
少林武当、峨眉昆仑：门派出场次数排行榜。
亢龙有悔、神龙摆尾：武功出现次数排行榜。
人在江湖、身不由己：人物与门派的关系。
风花雪月、侠骨柔情：人物与人物的关系。
东邪西毒、南帝白丐：人物与武功的关系。
先诛少林、再灭武当：门派排行榜。
宝刀屠龙、号令天下：兵器排行榜。
天下武功、唯快不破：武功排行榜。
华山论剑、独孤求败：高手排行榜。

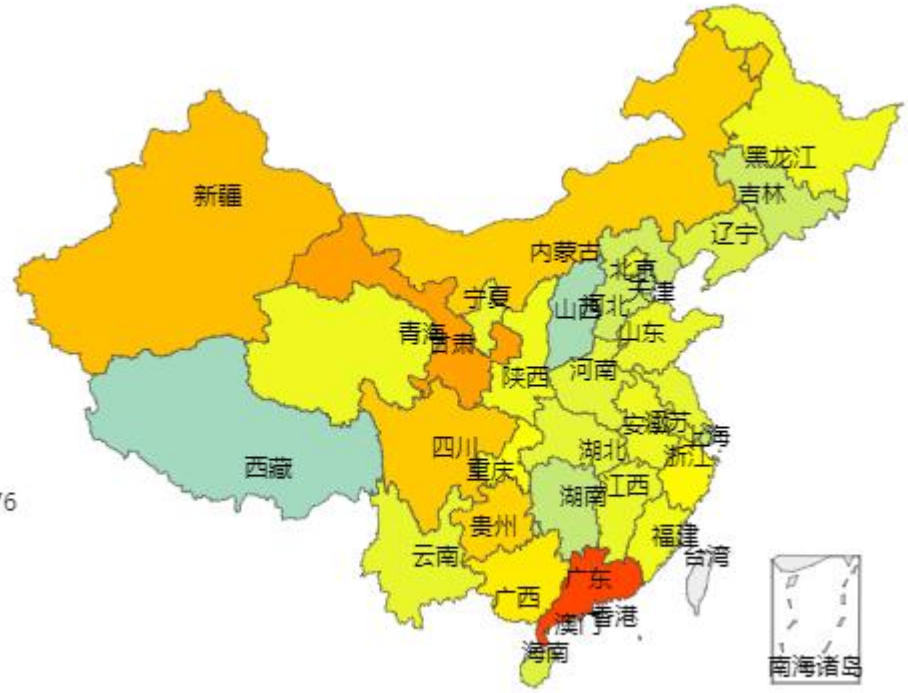

射雕英雄传
郭靖
黄蓉
洪七公
黄药师
周伯通
丘处机
欧阳克
梅超风
杨康
柯镇恶
裘千仞
朱聪
铁木真
完颜洪烈
穆念慈
拖雷
彭连虎
陆冠英
杨铁心
梁子翁
0
1000
天龙八部
六脉神剑
生死符
凌波微步
化功大法
般若掌
北冥神功
火焰刀
小无相功
降龙十八掌
大金刚拳
0
20
40
60
80
100
120
140

政府管理研究的几个问题

- 中国的计划生育政策
  - 中国政府对计划政策讨论最重视的是那几年?
  - 那个省最重视计划生育政策？哪个省最不重视?
- 群体性事件
  - 我国群体事件高发省份?
- 大数据
  - 我国从哪一年开始，大数据开始作为政府行为?

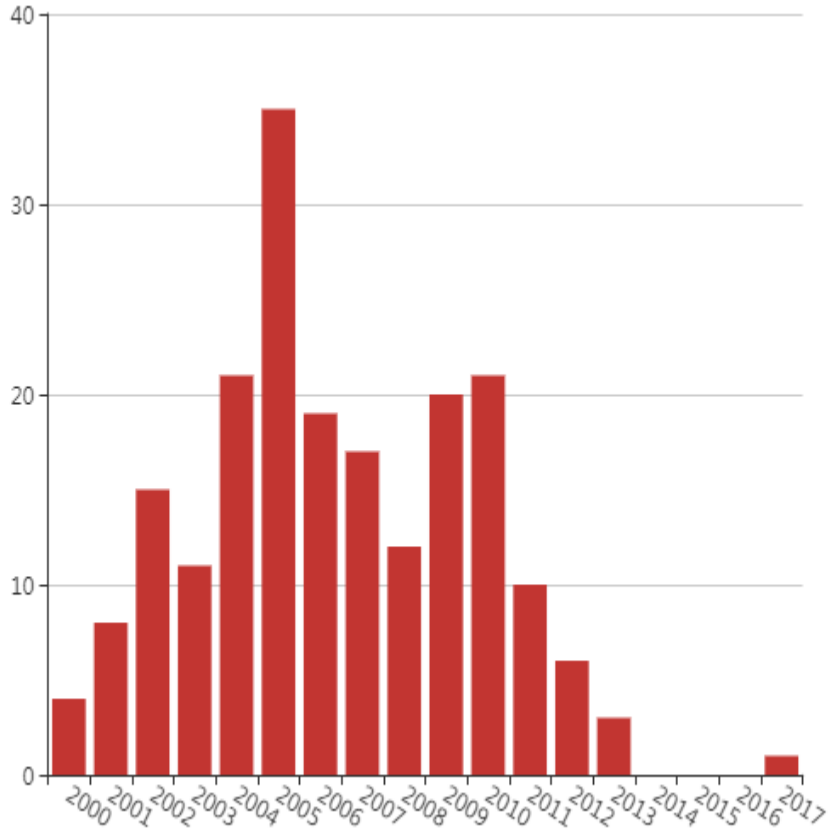
政府工作报告中隐藏的信息 - 计划生育

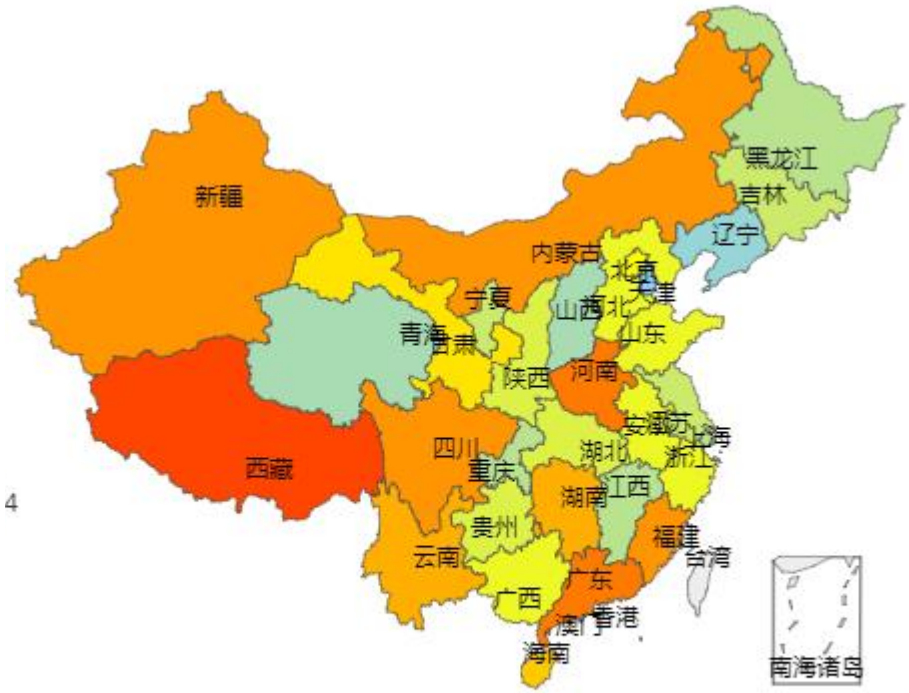




计划生育

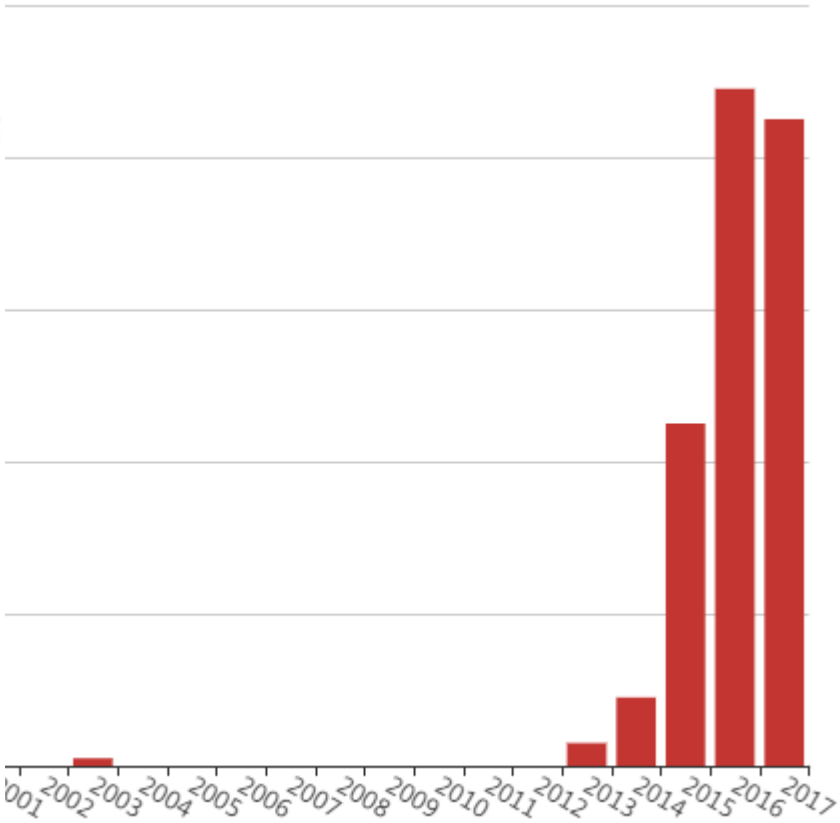
政府工作报告中隐藏的信息–群体性事件





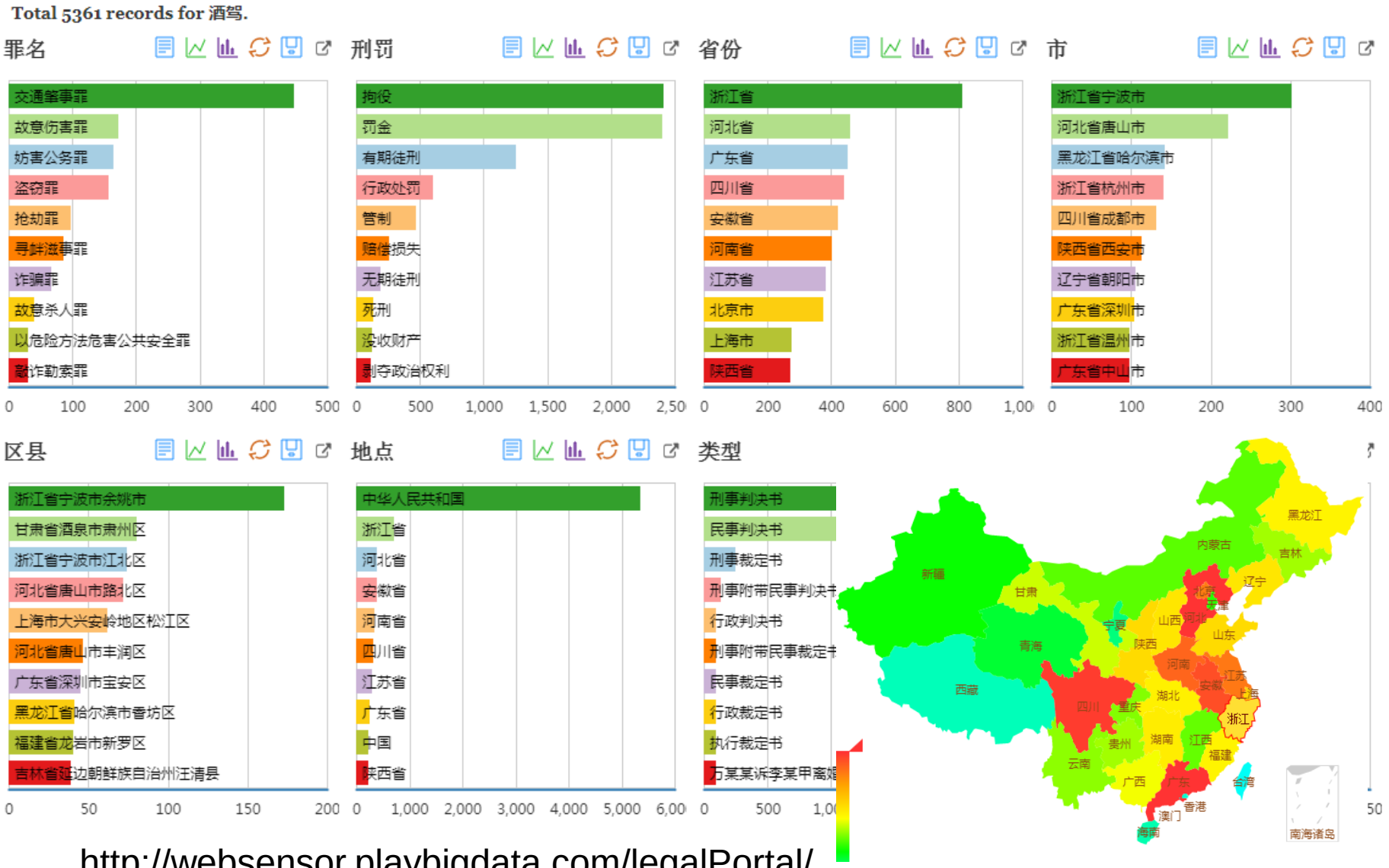
群体性事件

政府工作报告中隐藏的信息 – 大数据

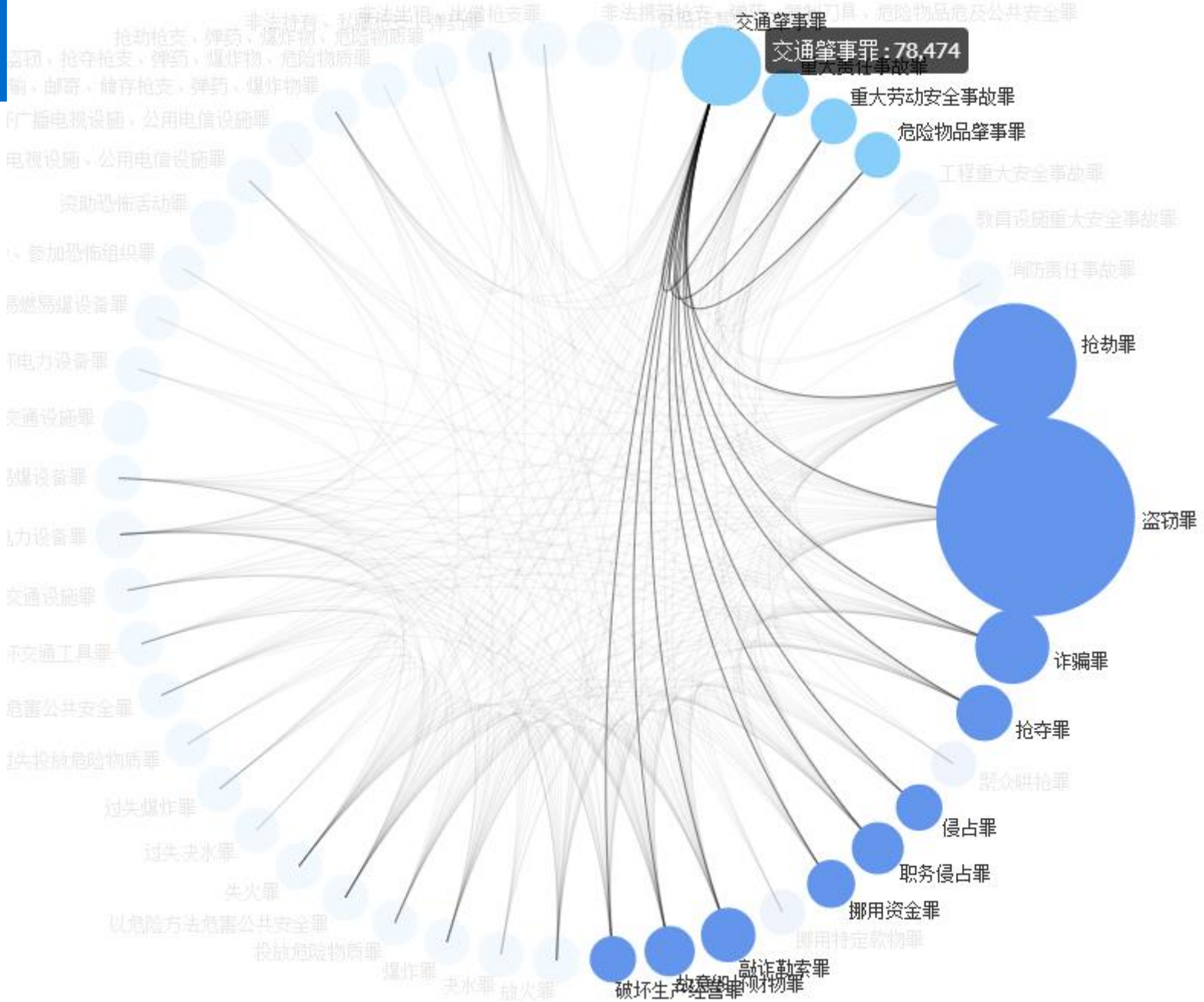


大数据

大数据+司法



http://websensor.playbigdata.com/legalPortal/


交通肇事罪
交通肇事罪：78,474
重大劳动安全事故罪
危险物品肇事罪
抢劫罪
盗窃罪
诈骗罪
抢夺罪
侵占罪
职务侵占罪
挪用资金罪
破坏生产经营罪

历史数据的价值


●●○○○ 中国移动 4G 14:10
返回 关闭 法狗狗法律机
交通肇事
重伤人数（人
1
2-3
死亡人数（人
事故责任
主要责任
全部责任
下一步
返回 关闭 法狗狗法律机器
选择地点
输入城市名称
北京-全省
所在省份
全国
北京
山东
江苏
上海
河北
陕西
湖南
福建
天津
四川
广西
海南
江西
开始查询
返回 关闭 法狗狗法律机
首页
交通肇事
基本预测
涉及法规
系统已学习250个案例，
重伤人数1人
死亡人数1人
全国
12-18个月
刑期预测
案件数
降低刑期 >
返回 关闭 法狗狗法律机器人
刑期（月）
降低刑期 >
71.6%
缓刑可能性
无缓刑
28.4%
有缓刑
71.6%
争取缓刑 >
查询结果仅供参考

自动定罪？

法律文书自动定罪

法律文书示例 | LSTM文本匹配 | CNN文本匹配 | 文本分类 | 多个二分类 | 方法介绍

胡某、陈某非法拘禁一审刑事判决书

淮南市田家庵区人民检察院指控：被告人胡某与被害人毕某因合伙经营饭店存在生意纠纷。2015年10月1日10时许，胡某约毕某在本区朝阳东路“上岛咖啡”店内见面协商相关债务事宜。二人见面后协商未果，胡某提出换个地方谈，毕某不同意，胡某便伙同被告人陈某、曹伟（另案处理）和陶某将毕某强行带上陈某驾驶的轿车。随后，四人将毕某带至大通区窑河大桥附近，下车后，胡某、陈某、曹伟、陶某对毕某实施殴打，当日13时许，胡某等人带着毕某来到本区淮河新城小区附近的&ldquo...

展开

RNN文本匹配 | CNN文本匹配 | 文本多分类 | 多个二分类 | 三种方法

丁俊杰诈骗罪刑事一审判决书

青冈县人民检察院指控：（一）、诈骗罪： 2014年5月初，被告人丁俊杰在其亲属生日宴中，遇到亲属大庆市居民李桂云，在吃饭过程中谎称自己承包工程需要融资并给予高额利息，骗取李桂云现金90000元。李桂云在当年6月中旬听亲属说可能被丁俊杰骗了，经多次索要丁俊杰偿还25000元，余款65000元丁俊杰拒不归还。 2015年3月被告人丁俊杰与同学葛岩取得联系后，得知葛岩正在销售化肥，于当年4月丁俊杰到葛岩家谎称有能力购买低价大庆尿素40吨，询问葛岩是否需要购买，葛岩同意购买后，为了骗取葛岩钱财，...

展开

法律文书示例 | LSTM文本匹配 | CNN文本匹配 | 文本分类 | 多个二分类 | 方法介绍

"淮南市田家庵区人民检察院指控：被告人胡某与被害人毕某因合伙经营饭店存在生意纠纷。2015年10月1日10时许，胡某约毕某在本区朝阳东路“上岛咖啡”店内见面协商相关债务事宜。二人见面后协商未果，胡某提出换个地方谈，毕某不同意，胡某便伙同被告人陈某、曹伟（另案处理）和陶某将毕某强行带上陈某驾驶的轿车。随后，四人将毕某带至大通区窑河大桥附近，下车后，胡某、陈某、曹伟、陶某对毕某实施殴打，当日13时许，胡某等人带着毕某来到本区淮河新城小区附近的“车速贷公司”准备让毕某写欠条，在进入公司时毕某趁胡某

预测

真实结果

非法拘禁罪

预测结果

非法拘禁罪	0.999904
聚众斗殴罪	0.203149

大数据和人文社科的衔接

- 人文社科向大数据思维转化：
  - 计算思维，定量研究
  - 数据思维：用数据说话
  - 挖掘和利用现有数据

常见桥段

A：听说你们是做大数据的，我们这有一堆数据，你帮我们分析一下呗。

B：好呀。你想分析什么？

A：我也不知道到。要不你先分析着看。做完了让我们看一眼。

B：。。。

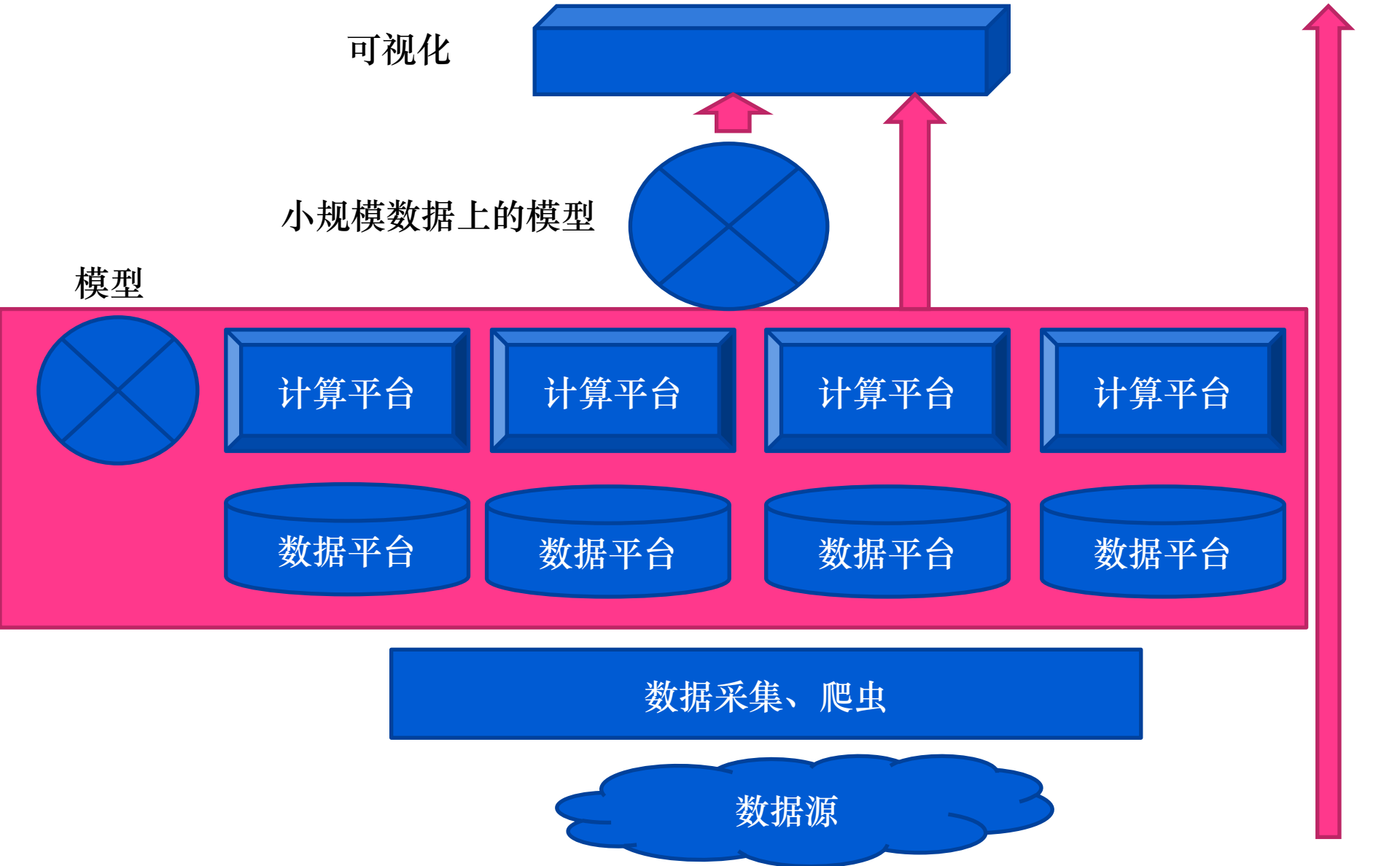
大数据+人文社科

- 大数据在人文社科领域的应用仍存在一定障碍
  - 人文社科的工作人员遇到大数据手足无措
  - 不了解大数据能够帮助他们解决哪些问题
  - 对大数据有过高或者过低的期望
  - 跨学科合作非常困难，沟通交流工作量巨大

课程目标

- 围绕大数据分析的生命周期和流程，通过结合文科专业背景的实验和案例，帮助人文社科专业的学生了解大数据思维，掌握大数据分析的基本原理和方法
- 力求让人文社科专业的学生能够理解大数据在解决相关问题时的基本思路，具备基本的大数据分析方法和技能，能够将大数据分析作为日后工作和科研上的工具
- 让文科生不仅了解大数据是什么，还能亲自进行大数据分析实战

通用大数据分析流程


可视化
小规模数据上的模型
模型
计算平台
计算平台
计算平台
计算平台
数据平台
数据平台
数据平台
数据平台
数据采集、爬虫
数据源

内容设计

- 大数据分析
  - 大数据思维及应用案例
  - 非结构化的文本分析
  - 数据挖掘与数据分析
  - 大数据平台
  - 数据可视化

- 编程语言：Python
  - 简单易学
  - 强大的数据挖掘和探索式数据分析功能
  - 良好的开源工具支持

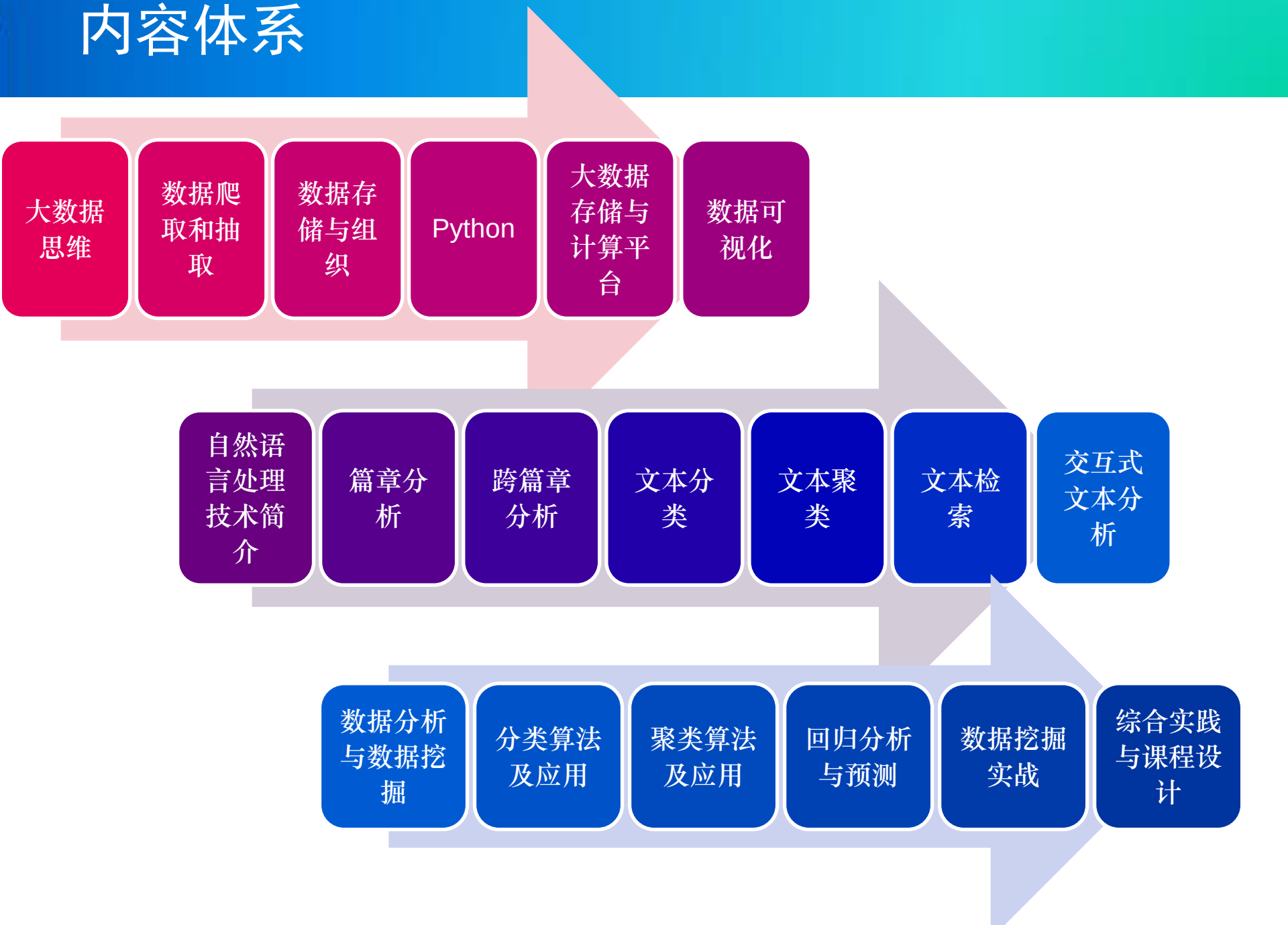
内容设计

- 围绕着一个或多个人文社科大数据应用案例
- 以非结构化文本大数据分析（分词、关键词抽取、命名实体抽取、情感分类、文本分类、主题分析等）为主线，以结构化大数据挖掘（分类、聚类、回归预测等）为辅
- 重点讲授在整个大数据分析的生命周期中的各个环节（数据采集、存储、预处理、表示、分析、挖掘、可视化） 的原理和工具
- 通过上机实验，力求让每个学生都能做一个大数据分析方面的小课题

内容设计

- 由浅入深
- 以案例教学为核心：一个大案例
- 培养大数据思维为主，动手实战为辅助
- 运用工具为主，动手编程为辅
- 基础算法为主，应用为辅
- 简单介绍大数据平台原理和思想
- 讲授与上机实习结合，多位助教协助

内容体系

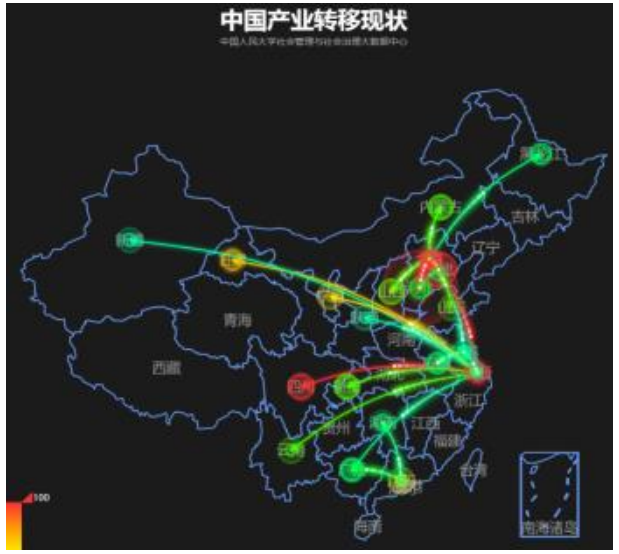

大数据思维
数据爬取和抽取
数据存储与组织
Python
大数据存储与计算平台
数据可视化
自然语言处理技术简介
篇章分析
跨篇章分析
文本分类
文本聚类
文本检索
交互式文本分析
数据分析与数据挖掘
分类算法及应用
聚类算法及应用
回归分析与预测
数据挖掘实战
综合实践与课程设计

大案例设计

- 政府工作报告分析
  - 具备时间+空间维度
  - 典型的文本数据（网页）
  - 数据量适中（很小）
  - 内容广泛（社会、经济、教育、司法等都有涉及）

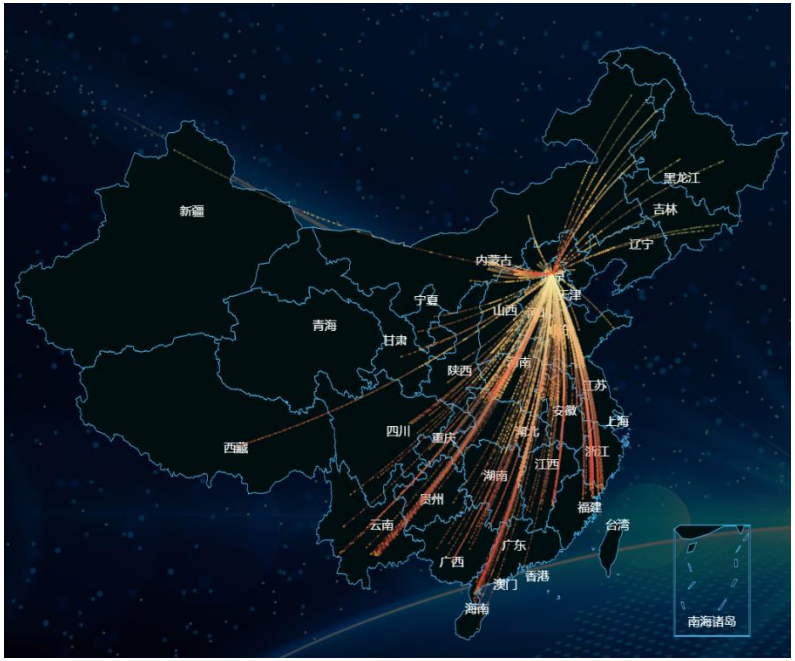
内容安排-1

- 大数据简介+大数据思维
  - 大数据思维：什么是大数据，大数据的重要性
  - 大数据在各个人文社科领域的应用示例等
  - 数据分析产品设计理念和基本形式



内容安排-2

- 数据可视化
  - 可视化的重要性以及典型案例
  - 常用可视化工具介绍（Excel+EChart）
  - 可视化图表制作实战（柱状图、曲线图、地图、词云等）；



内容安排-3

- 数据采集
  - 网页的基本原理
  - 数据爬取与信息抽取
  - 数据爬取工具
  - 爬取实战（单一页面、列表页面、循环爬取）

内容安排-4

- 文本分析
  - 什么是文本分析，文本分析常见任务
  - 基础自然语言处理算法：分词、词性识别、词性识别、实体抽取、关键词抽取等

内容安排–5

- 数据分析编程语言：Python
  - 基础Python编程内容 –简介，安装，变量，语句，函数，模块，列表，字符串，正则表达式等
  - Python下开源工具包使用
  - 文件处理，Json数据存储和转化, Excel文件生成
  - Scrapy爬虫框架使用
  - 自然语言处理工具调用

内容安排-6

- 简单数据分析
  - 简单数据统计分析
  - 数据序列化与存储
  - 数据分组聚合
  - 趋势图、timeline等制作
  - Excel + 数据透视表
  - Python交互式数据分析和可视化（matplotlib + seaborn 等）
  - 示例：词频统计等

内容安排-7

- 文本分类
  - 文本分类概念与应用
  - 文本分类算法简介
  - 话题模型详解
  - 实战（训练+预测）

内容安排-8

- 文本聚类
  - 聚类基本原理及应用场景
  - 常用文本聚类算法和工具
  - 聚类实战

内容安排-9

- 文本搜索与文本交互式分析
  - 文本检索系统原理与结构
  - 基于Solr构建检索与分析引擎
  - 数据导入导出，可视化对接

内容安排-10

- 数据挖掘
  - 数据挖掘概述，数据挖掘基本问题与基本流程
  - 数据分类、聚类和回归分析简介
  - Python数据挖掘基础，探索式数据分析
  - numpy, pandas使用

内容安排-11

- 大数据平台
  - 大数据分析平台概述
  - 数据存储和计算平台

内容安排-12

- 学生课程设计
  - 基于所学专业选定问题进行分析展示
  - 例如：政府工作报告分析展示
  - 也可自选其他题目
    - 司法文书分析：“食品安全问题”的研究
    - “中国人民大学”舆情分析
    - “雄安新区”、“一带一路”问题研究
    - “雾霾”问题
    - 上市公司，股票
    - …
  - 若干次课堂讨论

学生上机

- 上机
  - 信息学院提供机房
  - 课堂案例演练
  - 课程设计内容
  - 时间待定

考核方式

- 平时 50%
  - 若干作业
  - 课堂表现
- 课程设计 50%
  - 完成某数据分析产品并撰写设计报告或论文

课程网站

- 微人大课程中心
  - 课件下载
  - 提交作业
  - 两周后启用

教材和参考书

- 教材
  - 还没有，打算写一本
- 参考书
  - 《大数据时代》《大数据可视化》《数据之美》
  - 《Python基础教程》 Magnus Lie Hetland, 人民邮电出版社
  - 《Python数据分析与挖掘实战》 机械工业出版社
  - 《统计自然语言处理》
  - 《Python与数据挖掘》
  - 《Python大战机器学习：数据科学家的第一个小目标》
  - 《Python数据可视化》
  - 《数据科学实战手册R+Python》
  - 《Python机器学习实践指南》