



国家杰出青年科学基金1994-2004

组长: ZF1721402 杨子建

ZF1721138 巩孝刚

ZF1721413 袁 飞

ZF1721334 吴鹏飞

ZF1721424 张 磊

ZF1721439 赵礼鹏

项目贡献列表



| | 数据采集准备 | 数据采集 | 数据清洗 | 数据分析 | 数据可视化 | 贡献值 |
|-----|--------|------|------|------|-------|-------|
| 杨子建 | 100% | 0 | 0 | 0 | 100% | 16.6% |
| 袁飞 | 0 | 20% | 20% | 20% | 0 | 16.6% |
| 张磊 | 0 | 20% | 20% | 20% | 0 | 16.6% |
| 巩孝刚 | 0 | 20% | 20% | 20% | 0 | 16.6% |
| 赵礼鹏 | 0 | 20% | 20% | 20% | 0 | 16.6% |
| 吴鹏飞 | 0 | 20% | 20% | 20% | 0 | 16.6% |

数据采集准备



一.确认权威数据源：

我们组收集的目标数据为1994年至2004年国家杰出青年科学基金，首先寻找国家权威数据确定数据采集名单。

The screenshot shows the NSFC website with the following elements:












- Logo: 国家自然科学基金委员会 (National Natural Science Foundation of China)
- Navigation: 首页 (Home)
- Section: 杰出青年名录 (Outstanding Young Scientists List)
- Breadcrumb: 杰出青年专栏首页 > 杰出青年名录 > 附录
- Title: 附录：1994~2013年度国家杰出青年科学基金获资助者名单 (按年度、项目批准号顺序排列，不包括因故撤销者及未启动执行者)
- List of years (1994-2008):
 - 1994年, 1995年, 1996年, 1997年, 1998年
 - 1999年, 2000年, 2001年, 2002年, 2003年
 - 2004年, 2005年, 2006年, 2007年, 2008年

数据采集准备



二.收集互联网数据源:

确认了名单，先去网上寻找现成可编辑的数据，如百度文库、知网等，最好是excel格式的数据，word和pdf格式也接受，下载文本并与官网上的名单核对，保证数据准确。

-  [muchong.com]杰出青年科学基金资助名单1994-2011.xls
-  3组-ZF1721402杨子建-1994年到2004年 国家自然科学基金委员会杰出青年.xlsx
-  1994-1999.xlsx
-  1994年度“国家杰出青年科学基金”评审结果揭晓.pdf
-  1995年度国家杰出青年科学基金获得者名单(共计81名).pdf
-  1996年度国家杰出青年科学基金84位获选者名单（全）.pdf
-  1997年度国家杰出青年科学基金_省略_结果揭晓112名青年学者获得资助_汪平忠.pdf
-  1998年度国家杰出青年科学基金评审结果揭晓.pdf
-  1999年度国家杰出青年科学基金获资助者名单.pdf
-  2000-2004.xlsx
-  高性能计算杰出青年统计表表样.xlsx

数据采集准备



三.数据采集格式确认：

收集数据需要各组员统一标准，按作业要求编写收集维度的表头。

| A | B | C | D | E | F | G | H | I |
|----|----|------|----|----|------|------|----|------|
| 时间 | 姓名 | 出生日期 | 性别 | 籍贯 | 籍贯经度 | 籍贯纬度 | 单位 | 单位地址 |

| J | K | L | M | N | O | P |
|------|------|------|----|------|--------|--------|
| 单位经度 | 单位纬度 | 毕业学校 | 专业 | 专业代码 | 学士学位学校 | 学士学位时间 |

| Q | R | S | T | U |
|--------|--------|--------|--------|--------|
| 硕士学位学校 | 硕士学位时间 | 博士学位学校 | 博士学位时间 | 获得院士时间 |

数据采集准备



四.整理互联网数据源：

确认收集数据的维度，下面要做的就是整理下载的文档，例如pdf格式需要转换成可编辑的文档，word想办法将信息插入到excel中，最终收集的格式为excel，如图所示：

- [muchong.com]杰出青年科学基金资助名单1994-2011.xls
- 3组-ZF1721402杨子建-1994年到2004年 国家自然科学基金委员会杰出青年.xlsx
- 1994-1999.xlsx
- 1994年度“国家杰出青年科学基金”评审结果揭晓.pdf
- 1995年度国家杰出青年科学基金获得者名单(共计81名).pdf
- 1996年度国家杰出青年科学基金84位获选者名单（全）.pdf
- 1997年度国家杰出青年科学基金_省略_结果揭晓112名青年学者获得资助_汪平忠.pdf
- 1998年度国家杰出青年科学基金评审结果揭晓.pdf
- 1999年度国家杰出青年科学基金获资助者名单.pdf
- 2000-2004.xlsx
- 高性能计算杰出青年统计表表样.xlsx

数据采集准备



五.文档格式转换:

word文档转换为excel复制粘贴即可，pdf转换为excel比较麻烦，存在很多问题，例如有的pdf内容是扫描图片，有的是影印，如果直接用工具直接转化会失败。推荐两款工具：Adobe Acrobat 2017和Readiris Corporate 15

- [muchong.com]杰出青年科学基金资助名单1994-2011.xls
- 3组-ZF1721402杨子建-1994年到2004年 国家自然科学基金委员会杰出青年.xlsx
- 1994-1999.xlsx
- 1994年度“国家杰出青年科学基金”评审结果揭晓.pdf
- 1995年度国家杰出青年科学基金获得者名单(共计81名).pdf
- 1996年度国家杰出青年科学基金84位获选者名单（全）.pdf
- 1997年度国家杰出青年科学基金_省略_结果揭晓112名青年学者获得资助_汪平忠.pdf
- 1998年度国家杰出青年科学基金评审结果揭晓.pdf
- 1999年度国家杰出青年科学基金获资助者名单.pdf
- 2000-2004.xlsx
- 高性能计算杰出青年统计表表样.xlsx



五.文档格式转换：

Adobe Acrobat 2017 可对pdf进行编辑，如图所示，可以把不需要的信息编辑掉，包括一些图片噪音，修改完成后可将pdf导出为高清图片。

专业评审组会议，评选出 65 名候选人向评审委员会推荐。评委会在此基础上，经过认真、负责的终审，评定出 1994 年度“国家杰出青年科学基金”获得者陈永川等 49 名。

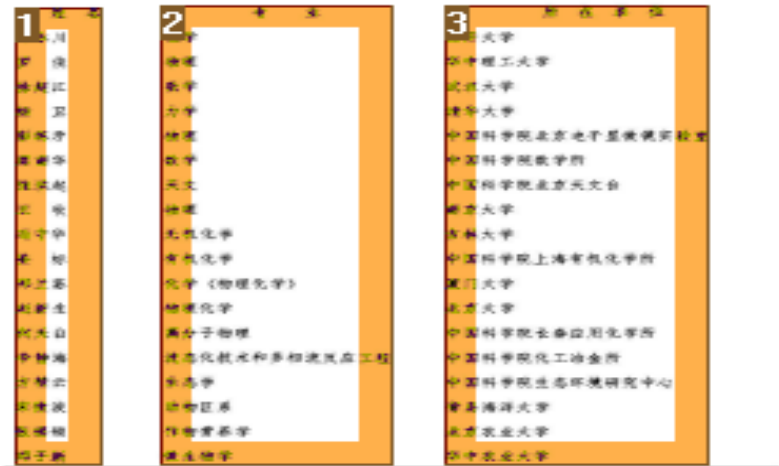
| 1994 年度“国家杰出青年科学基金”获得者名单 | | | |
|--------------------------|----------|-----------------|--|
| 姓 名 | 专 业 | 所 在 单 位 | |
| 陈永川 | 数学 | 南开大学 | |
| 罗 俊 | 物理 | 华中理工大学 | |
| 徐超江 | 数学 | 武汉大学 | |
| 杨 卫 | 力学 | 清华大学 | |
| 彭练矛 | 物理 | 中国科学院北京电子显微镜实验室 | |
| 席南华 | 数学 | 中国科学院数学所 | |
| 张洪起 | 天文 | 中国科学院北京天文台 | |
| 王 牧 | 物理 | 南京大学 | |
| 冯守华 | 无机化学 | 吉林大学 | |
| 姜 标 | 有机化学 | 中国科学院上海有机化学所 | |
| 郑兰荪 | 化学（物理化学） | 厦门大学 | |

数据采集准备



五.文档格式转换:

得到的高清图片用Readiris Corporate15进行文字识别, Readiris Corporate 15支持选择图片区域并识别区域内的文字:



数据采集准备



六.爬虫爬取数据:

转换成word, 复制数据
粘贴到excel中, 这样查
找到所有的名单及部分信
息后, 利用爬虫爬取百科
的数据, 进一步完善杰青
信息:

```
# 通过百度百科获取数据
def baidubaike(url, name):
    status, content = open_url(url)
    if status == 200:
        bs = BeautifulSoup(content, "html.parser")
        meta = bs.findAll('meta', {'name': 'description'})
        birthday = '?'
        native_place = '?'
        for tag in meta:
            desc = tag.attrs['content']
            if desc.find(name) > -1:
                obj = re.search('[0-9]*年.*生于[^\, \. \; ]*[市|县|区|道|乡|镇|村|县]
                birthday, native_place = birthday_place(obj)
        obj = re.search('[0-9]*年[^\, \. \; <]*毕业于[^\, \. \; <]*', content)
        time, school = school_time(obj)
        sex = '?'
        if content.find('男\n') > -1:
            sex = '男'
        elif content.find('女\n') > -1:
            sex = '女'
        return url, birthday, sex, native_place, time, school
    return url, '', '', '', '', ''

# 获取百度百科地址, key是单位用于多条时做区分
def get_baidubaike_url(name, key):
    sign = key
    if len(key) > 6:
        sign = key[0:6]
    url = 'https://baike.baidu.com/search?' + urllib.parse.urlencode({'word': na
    status, data = open_url(url)
    if status == 200:
        bs = BeautifulSoup(data, "html.parser")
        divs = bs.findAll('div', {'class': 'searchResult'}, True)
        for div in divs:
            for child in div.descendants:
                if child.name == 'dd':
                    href = ''
                    for ele in child.children:
```

数据采集准备



六.爬虫爬取数据:

爬虫需要匹配网页结构, 有的信息写在个人简介里, 有的写在正文中, 这就需要利用不同的爬虫爬取数据:

| | | | |
|--------------------|------------------|---------------------|--------|
| ... | ... | ... | ... |
| .idea | 2018/5/30 19:13 | 文件夹 | |
| baike | 2018/4/10 14:14 | 文件夹 | |
| common | 2018/4/10 14:12 | 文件夹 | |
| Include | 2017/12/11 20:31 | 文件夹 | |
| Lib | 2018/4/9 19:21 | 文件夹 | |
| Scripts | 2018/4/21 15:34 | 文件夹 | |
| tcl | 2018/4/9 19:13 | 文件夹 | |
| aspider.py | 2018/4/10 20:13 | Python File | 2 KB |
| bspider.py | 2018/4/10 22:59 | Python File | 3 KB |
| copyData.py | 2018/4/22 19:37 | Python File | 1 KB |
| datatemp.txt | 2018/4/14 14:33 | TXT 文件 | 3 KB |
| dict.py | 2018/4/22 9:26 | Python File | 1 KB |
| odem.xlsx | 2018/4/22 18:19 | Microsoft Excel ... | 83 KB |
| pip-selfcheck.json | 2018/4/21 15:32 | JSON File | 1 KB |
| RQL.py | 2018/5/14 21:21 | Python File | 1 KB |
| spiderTest.py | 2018/4/14 14:47 | Python File | 2 KB |
| temp.txt | 2018/4/14 14:37 | TXT 文件 | 5 KB |
| test.py | 2018/4/22 18:15 | Python File | 1 KB |
| text.xlsx | 2018/4/22 18:06 | Microsoft Excel ... | 168 KB |
| text_new.xlsx | 2018/4/22 18:20 | Microsoft Excel ... | 149 KB |

数据采集准备



七.人工初步数据校对:

爬虫的意义在于减轻工作量，但是扒下来的数据不全是准确的，这就需要人工校对，在初期只对生日毕业时间等校对，例如1994年获得杰青称号，1998年才出生的诡异数据，更详细的数据有利于后期组员查找时有更多匹配项，提高准确率减少工作时间。

| 时间 | 姓名 | 出生日期 | 性别 | 籍贯 | 籍贯经度 | 籍贯纬度 | 单位 |
|------|-----|------------|----|-------------|----------|----------|--------|
| 1994 | 陈永川 | 1964年3月 | ? | 四川南充 | 106.083 | 30.79528 | 南开大学 |
| 1994 | 罗俊 | 1956年11月 | | 湖北省仙桃 | 113.454 | 30.36495 | 华中理工大学 |
| 1994 | 徐超江 | 1956年 | ? | 湖北潜江 | 112.8969 | 30.42122 | 武汉大学 |
| 1994 | 杨卫 | 1954年2月 | ? | 北京 | 116.4273 | 39.90498 | 清华大学 |
| 1994 | 彭练矛 | 1962年9月 | ? | Noneaddress | | | 中国科学院 |
| 1994 | 席南华 | 1963年3月 | ? | 衡阳市祁东 | 112.0904 | 26.7999 | 中国科学院 |
| 1994 | 张洪起 | Noneinfo | | Noneaddress | | | 中国科学院 |
| 1994 | 王牧 | Noneinfo | ? | 南京 | 118.7966 | 32.05935 | 南京大学 |
| 1994 | 冯守华 | 1956年3月 | ? | 吉林省磐石 | 126.0604 | 42.94629 | 吉林大学 |
| 1994 | 姜标 | 1962年11月 | ? | 福建福安 | 119.6478 | 27.08805 | 中国科学院 |
| 1994 | 郑兰荪 | 1954年10月 | ? | 江苏省吴江 | 120.6379 | 31.15612 | 厦门大学 |
| 1994 | 赵新生 | 1957年8月 | ? | 安徽省祁门 | 117.7174 | 29.85406 | 北京大学 |
| 1994 | 何天白 | Noneinfo | ? | Noneaddress | | | 中国科学院 |
| 1994 | 李静海 | 1956年10月 | ? | 山西静乐 | 111.9394 | 38.35904 | 中国科学院 |
| 1994 | 方精云 | 1959年7月 | ? | 安徽怀宁 | 117.2239 | 31.85453 | 中国科学院 |
| 1994 | 宋微波 | 1958年 | | 山东微山县 | 117.1288 | 34.80655 | 青岛海洋大学 |
| 1994 | 张福锁 | Noneinfo | ? | 陕西凤翔 | 107.4102 | 34.51874 | 北京农业大学 |
| 1994 | 邓子新 | 1957年 | 男 | 湖北房县 | 110.7267 | 32.04009 | 华中农业大学 |
| 1994 | 孙方臻 | 1962年9月16日 | ? | 山东 | 120.3037 | 35.99004 | 中国科学院 |
| 1994 | 王志新 | 1953年8月10日 | ? | 北京 | 116.4273 | 39.90498 | 中国科学院 |
| 1994 | 陈建国 | 1945年7月 | ? | 山东荣成 | 122.4967 | 27.16516 | 北京大学 |

数据采集



成员分担查询任务

爬虫无法完成的工作智能通过人工完成，1994年至2004年获得杰青称号的有1395人，按人数分配任务，每人完成25%的工作量。

| A | B | C | D | E | F | G | H |
|------|-----|------------|----|-------------|----------|----------|--------|
| 时间 | 姓名 | 出生日期 | 性别 | 籍贯 | 籍贯经度 | 籍贯纬度 | 单位 |
| 1994 | 陈永川 | 1964年3月 | ? | 四川南充 | 106.083 | 30.79528 | 南开大学 |
| 1994 | 罗俊 | 1956年11月 | | 湖北省仙桃 | 113.454 | 30.36495 | 华中理工大学 |
| 1994 | 徐超江 | 1956年 | ? | 湖北潜江 | 112.8969 | 30.42122 | 武汉大学 |
| 1994 | 杨卫 | 1954年2月 | ? | 北京 | 116.4273 | 39.90498 | 清华大学 |
| 1994 | 彭练矛 | 1962年9月 | ? | Noneaddress | | | 中国科学院 |
| 1994 | 席南华 | 1963年3月 | ? | 衡阳市祁东 | 112.0904 | 26.7999 | 中国科学院 |
| 1994 | 张洪起 | Noneinfo | | Noneaddress | | | 中国科学院 |
| 1994 | 王牧 | Noneinfo | ? | 南京 | 118.7966 | 32.05935 | 南京大学 |
| 1994 | 冯守华 | 1956年3月 | ? | 吉林省磐石 | 126.0604 | 42.94629 | 吉林大学 |
| 1994 | 姜标 | 1962年11月 | ? | 福建福安 | 119.6478 | 27.08805 | 中国科学院 |
| 1994 | 郑兰荪 | 1954年10月 | ? | 江苏省吴江 | 120.6379 | 31.15612 | 厦门大学 |
| 1994 | 赵新生 | 1957年8月 | ? | 安徽省祁门 | 117.7174 | 29.85406 | 北京大学 |
| 1994 | 何天白 | Noneinfo | ? | Noneaddress | | | 中国科学院 |
| 1994 | 李静海 | 1956年10月 | ? | 山西静乐 | 111.9394 | 38.35904 | 中国科学院 |
| 1994 | 方精云 | 1959年7月 | ? | 安徽怀宁 | 117.2239 | 31.85453 | 中国科学院 |
| 1994 | 宋微波 | 1958年 | | 山东微山县 | 117.1288 | 34.80655 | 青岛海洋大学 |
| 1994 | 张福锁 | Noneinfo | ? | 陕西凤翔 | 107.4102 | 34.51874 | 北京农业大学 |
| 1994 | 邓子新 | 1957年 | 男 | 湖北房县 | 110.7267 | 32.04009 | 华中农业大学 |
| 1994 | 孙方臻 | 1962年9月16日 | ? | 山东 | 120.3037 | 35.99004 | 中国科学院 |
| 1994 | 王志新 | 1953年8月10日 | ? | 北京 | 116.4273 | 39.90498 | 中国科学院 |
| 1994 | 陈建国 | 1945年7月 | ? | 山东荣成 | 122.4867 | 37.16516 | 北京大学 |
| 1994 | 赵峰明 | 1952年10月 | ? | 浙江温州 | 120.0671 | 30.11600 | 中国医学科 |

数据清洗



| 时间 | 姓名 | 出生日期 | 性别 | 籍贯 | 籍贯经度 | 籍贯纬度 | 单位 | 单位地址 |
|------|-----|-----------|----|-------|------------|----------|--------------|----------|
| 1994 | 陈永川 | 1964/3/1 | 男 | 四川南充 | 106.082974 | 30.79528 | 南开大学 | 天津市南开区卫津 |
| 1994 | 罗俊 | 1956/11/1 | 男 | 湖北省仙桃 | 113.453974 | 30.36495 | 华中理工大学 | 湖北省武汉市洪山 |
| 1994 | 徐超江 | 1956年 | 男 | 湖北潜江 | 112.896866 | 30.42122 | 武汉大学 | 湖北省武汉市武昌 |
| 1994 | 杨卫 | 1954/2/1 | 男 | 北京 | 116.427287 | 39.90498 | 清华大学 | 北京市海淀区双清 |
| 1994 | 彭练矛 | 1962/9/1 | 男 | | | | 中国科学院北京电子显微镜 | 北京市西城区三里 |
| 1994 | 席南华 | 1963/3/1 | 男 | 衡阳市祁东 | 112.090357 | 26.7999 | 中国科学院数学所 | 北京市海淀区恒兴 |
| 1994 | 张洪起 | | 男 | | | | 中国科学院北京天文台 | 北京市海淀区中关 |
| 1994 | 王牧 | | 男 | 南京 | 118.796623 | 32.05935 | 南京大学 | 江苏省南京市鼓楼 |
| 1994 | 冯守华 | 1956/3/14 | 男 | 吉林省磐石 | 126.060427 | 42.94629 | 吉林大学 | 吉林省长春市朝阳 |
| 1994 | 姜标 | 1962/11/1 | 男 | 福建福安 | 119.64777 | 27.08805 | 中国科学院上海有机化学所 | 北京市西城区三里 |
| 1994 | 郑兰荪 | 1954/10/1 | 男 | 江苏省吴江 | 120.637851 | 31.15612 | 厦门大学 | 福建省厦门市思明 |
| 1994 | 赵新生 | 1957/8/1 | 男 | 安徽省祁门 | 117.717396 | 29.85406 | 北京大学 | 北京市海淀区颐和 |
| 1994 | 何天白 | | 男 | | | | 中国科学院长春应用化学所 | 北京市西城区三里 |
| 1994 | 李静海 | 1956/10/1 | 男 | 山西静乐 | 111.93944 | 38.35904 | 中国科学院化工冶金所 | 北京市西城区三里 |
| 1994 | 方精云 | 1959/7/1 | 男 | 安徽怀宁 | 117.223892 | 31.85453 | 中国科学院生态环境研究中 | 北京市海淀区双清 |
| 1994 | 宋微波 | 1958年 | 男 | 山东微山县 | 117.128828 | 34.80655 | 青岛海洋大学 | 山东省青岛市市南 |
| 1994 | 张福锁 | 1960/10/1 | 男 | 陕西凤翔 | 107.410155 | 34.51874 | 北京农业大学 | 北京市海淀区清华 |
| 1994 | 邓子新 | 1957年 | 男 | 湖北房县 | 110.72667 | 32.04009 | 华中农业大学 | 湖北省武汉市洪山 |
| 1994 | 孙方臻 | 1962/9/16 | 男 | 山东 | 120.303694 | 35.99004 | 中国科学院发育生物学研究 | 中关村南一条3号 |
| 1994 | 王志新 | 1953/8/10 | 男 | 北京 | 116.427287 | 39.90498 | 中国科学院生物物理研究所 | 北京市朝阳区大屯 |
| 1994 | 陈建国 | 1945/7/1 | 男 | 山东荣成 | 122.486658 | 37.16516 | 北京大学 | 北京市海淀区颐和 |
| 1994 | 杨焕明 | 1952/10/1 | 男 | 浙江温州乐 | 120.967147 | 28.11608 | 中国医学科学院基础医学研 | 北京市东城区东华 |

数据分析



根据得到的数据我们提出了4个问题：

- 一、杰出青年的申报单位是怎么分布的？
- 二、杰出青年有母校支持，得到称号的几率是不是高些？
出国留学是否有利于获得杰青称号？
- 三、画个四象限图，以获得称号时的年龄与获得称号时
本科毕业年限为依据，看看他们是怎么分布的？
- 四、获得杰青的学者从家乡出发求学，他们的迁徙图是怎
样的？有没有有意思的现象发生？

杰出青年申报单位的分布



2) 统计结果

拥有杰青数量最多的
单位TOP25

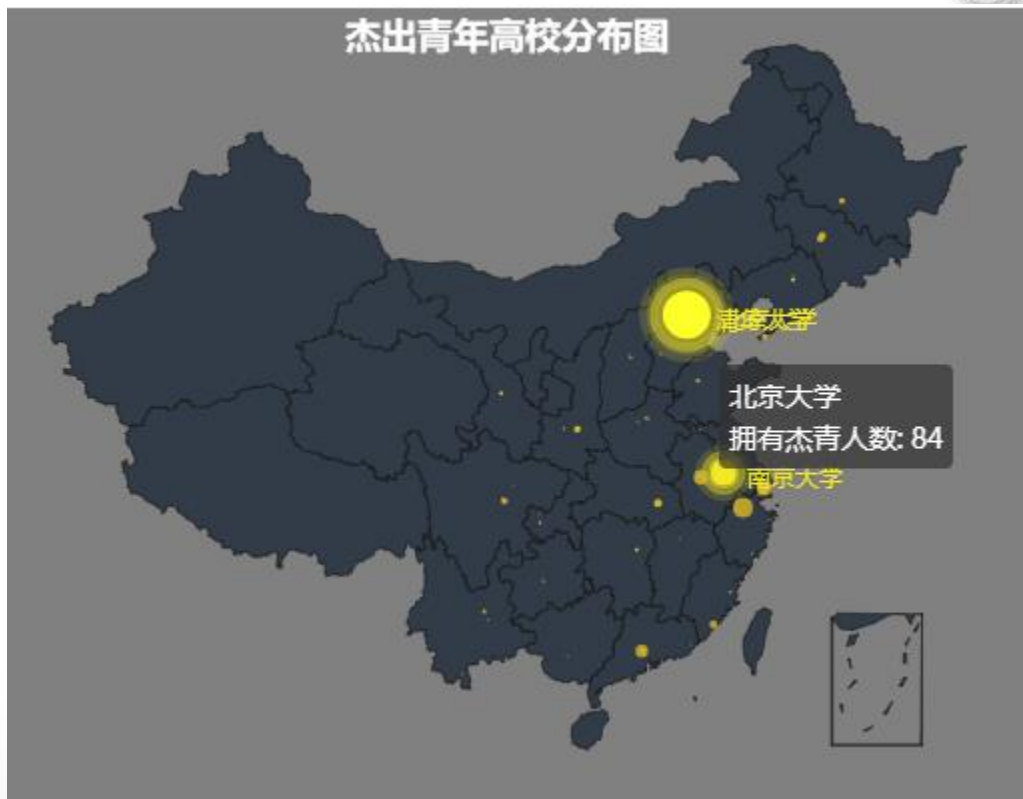
| | | | | |
|----|-----------------|------------|-----------|----|
| 1 | 北京大学 | 116.31088 | 39.99281 | 84 |
| 2 | 清华大学 | 116.326836 | 40.00366 | 72 |
| 3 | 南京大学 | 118.77943 | 32.055015 | 50 |
| 4 | 浙江大学 | 120.081699 | 30.302915 | 39 |
| 5 | 复旦大学 | 121.503584 | 31.296426 | 29 |
| 6 | 中国科学技术大学 | 117.269587 | 31.836842 | 29 |
| 7 | 中山大学 | 113.298379 | 23.096705 | 25 |
| 8 | 中国科学院物理研究所 | 116.331745 | 39.981833 | 24 |
| 9 | 中国科学院化学研究所 | 116.322509 | 39.990361 | 21 |
| 10 | 上海交通大学 | 121.433117 | 31.199008 | 20 |
| 11 | 中国农业大学 | 116.357122 | 40.005004 | 16 |
| 12 | 厦门大学 | 118.102555 | 24.436341 | 15 |
| 13 | 武汉大学 | 114.365248 | 30.53786 | 15 |
| 14 | 中国科学院数学与系统科学研究院 | 116.33253 | 39.982362 | 15 |
| 15 | 中国医学科学院 | 116.470664 | 39.915744 | 15 |
| 16 | 吉林大学 | 125.277062 | 43.823759 | 14 |
| 17 | 中国科学院长春应用化学研究所 | 125.406757 | 43.975508 | 14 |
| 18 | 东南大学 | 118.79484 | 32.054227 | 13 |
| 19 | 华中科技大学 | 114.414724 | 30.515977 | 13 |
| 20 | 中国科学院上海有机化学研究所 | 121.456924 | 31.192088 | 13 |
| 21 | 中国人民解放军第二军医大学 | 121.523917 | 31.30618 | 13 |
| 22 | 北京航空航天大学 | 116.347313 | 39.981771 | 12 |
| 23 | 北京师范大学 | 116.365798 | 39.961576 | 12 |
| 24 | 南开大学 | 117.167299 | 39.103693 | 12 |
| 25 | 四川大学 | 104.083748 | 30.630869 | 12 |

杰出青年申报单位的分布



3) 可视化结果

地图上黄色的点均为杰青申报单位，点越大说明该单位拥有的杰青越多。明显可以看出，北京地区和江浙地区盛产杰青学者，西部地区一个都没有，中部和中南部有零星分布，这反映了有能力支持获得杰青称号的单位集中在北京和江浙地区。



母校与留学



1) 数据清洗

该需求只需要分析杰青的申报单位以及学士硕士博士学校，所以需要清洗“单位名称”“学士学位”“硕士学位”“博士学位”三个属性：

| B | C | D | E | F | G |
|----------------|-------------|-------------|-------------|------|---------|
| 单位 | 本科 | 研究生 | 博士 | 是否出国 | 是否与单位重合 |
| 中科院 | 安徽大学 | | 复旦大学 | | 0 |
| 中科院 | 安徽农学院 | 日本信州大学 | 日本大阪市立大学 | 1 | 0 |
| 中科院 | 北京大学 | 北京大学 | 美国亚利桑那州 | 1 | 0 |
| 中科院 | 北京大学 | | | | 0 |
| 北京大学 | 北京大学 | 北京大学 | 美国加州大学伯克利分校 | 1 | 1 |
| 中科院 | 北京大学 | 中科院 | | | 1 |
| 中科院 | 北京大学 | | | | 0 |
| 北京医科大学 | 北京医科大学 | 北京医科大学 | 北京医科大学 | | 1 |
| 北京邮电大学 | 电子科技大学 | 北京邮电大学 | 北京邮电大学 | | 1 |
| 北京航空航天大学 | 阜新矿业学院 | 德国克劳斯塔尔工业大学 | 德国柏林工业大学 | 1 | 0 |
| 中科院 | 哈尔滨工业大学 | 哈尔滨工业大学 | 中科院 | | 1 |
| 中国医学科学院基础医学研究所 | 浙江大学 | 南京铁道医学院 | 丹麦哥本哈根大学 | 1 | 0 |
| 中科院 | 湖南怀化师范专科学校 | 华东师范大学 | 华东师范大学 | | 0 |
| 中科院 | 华北理工大学 | 北京科技大学 | 中科院 | | 1 |
| 华中科技大学 | 华中科技大学 | 中山大学 | 中科院 | | 1 |
| 华中农业大学 | 华中农业大学 | 华中农业大学 | 英国JohnInnes | 1 | 1 |
| 吉林大学 | 吉林大学 | | 吉林大学 | | 1 |
| 上海第二医科大学 | 江西省上饶地区卫生学校 | 上海第二医学院 | 法国巴黎第七大学 | 1 | 0 |
| 中科院 | 兰州大学 | | 中科院 | | 1 |

母校与留学



2) 统计结果

统计出杰青申报单位与杰青母校重合的数量，杰青求学期间是否留过学，我们的目的是了解申报单位是母校以及留学经历是否对获得称号有利：

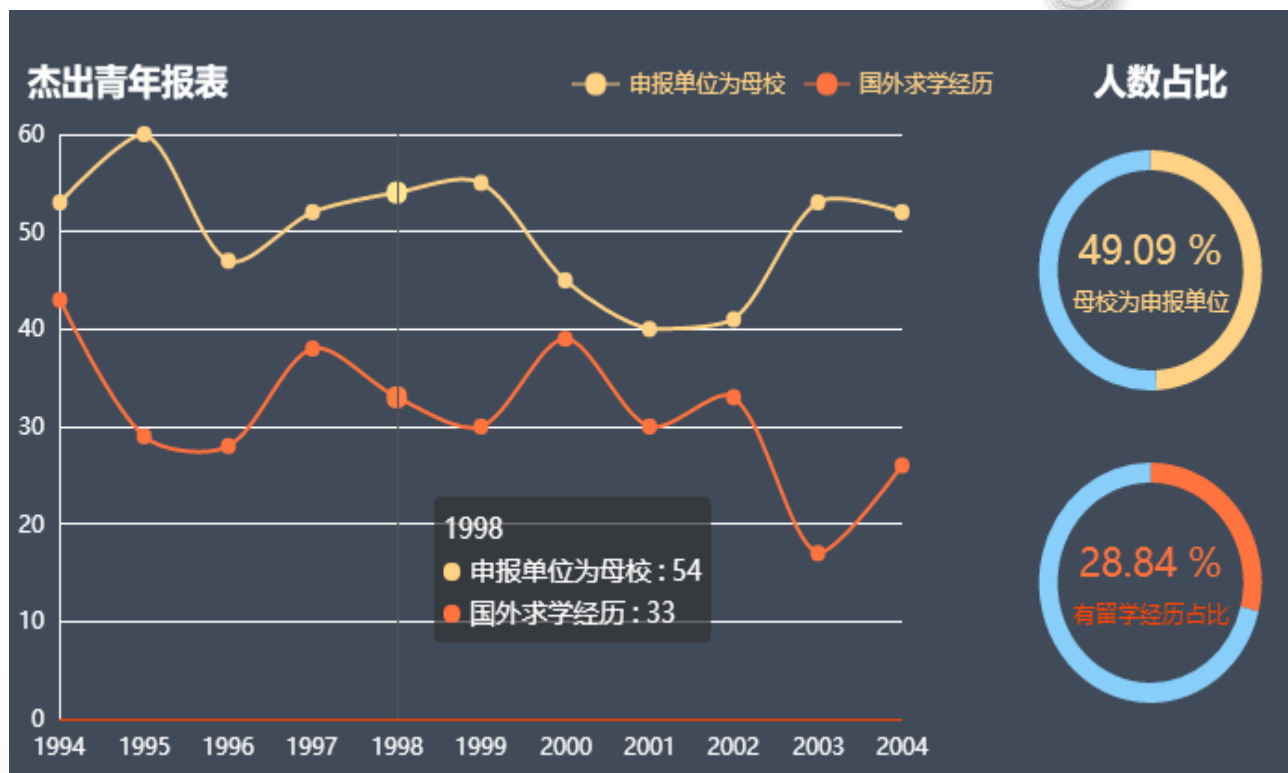
| 年份 | 每年的出国人数 | 每年的重合人数 | 每年的出国且重合人数 | | |
|------|---------|---------|------------|--|------|
| 1994 | 21 | 26 | 6 | | 49 |
| 1995 | 20 | 42 | 9 | | 70 |
| 1996 | 20 | 34 | 4 | | 72 |
| 1997 | 37 | 51 | 16 | | 98 |
| 1998 | 27 | 45 | 9 | | 83 |
| 1999 | 36 | 62 | 10 | | 119 |
| 2000 | 53 | 61 | 19 | | 135 |
| 2001 | 37 | 50 | 8 | | 125 |
| 2002 | 44 | 55 | 9 | | 135 |
| 2003 | 23 | 70 | 10 | | 132 |
| 2004 | 34 | 69 | 9 | | 133 |
| | | | | | 1151 |

母校与留学



3) 可视化展示

纵轴为百分比数据，
例如1998年母校与
单位重合的人数占当
年总人数的54%，1
994年到2004年总
占比为49%，说明
从母校申请杰青荣誉
称号确实成功率很大。

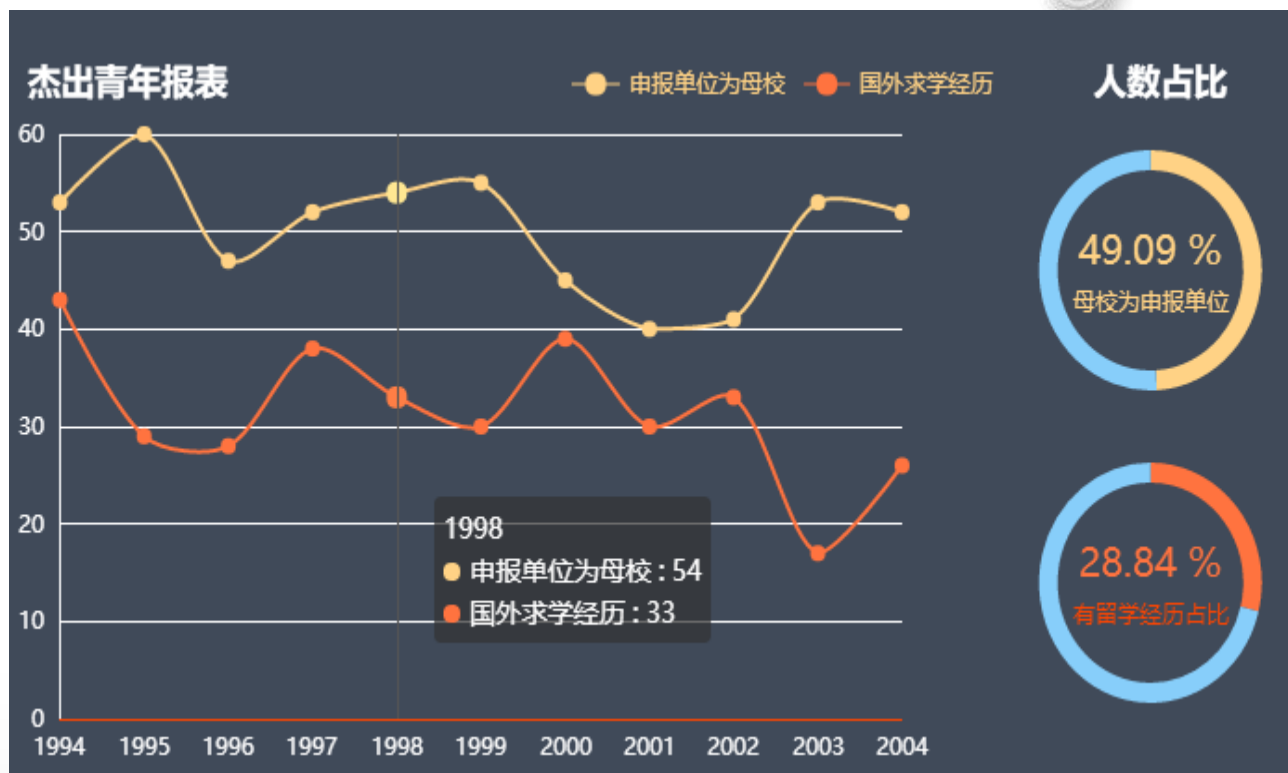


母校与留学



3) 可视化展示

纵轴为百分比数据，
例如1998年留学人数占当年总人数的33%，1994年到2004年总占比约为28%，也就是说大约每3个杰青就有一个有留学经历。

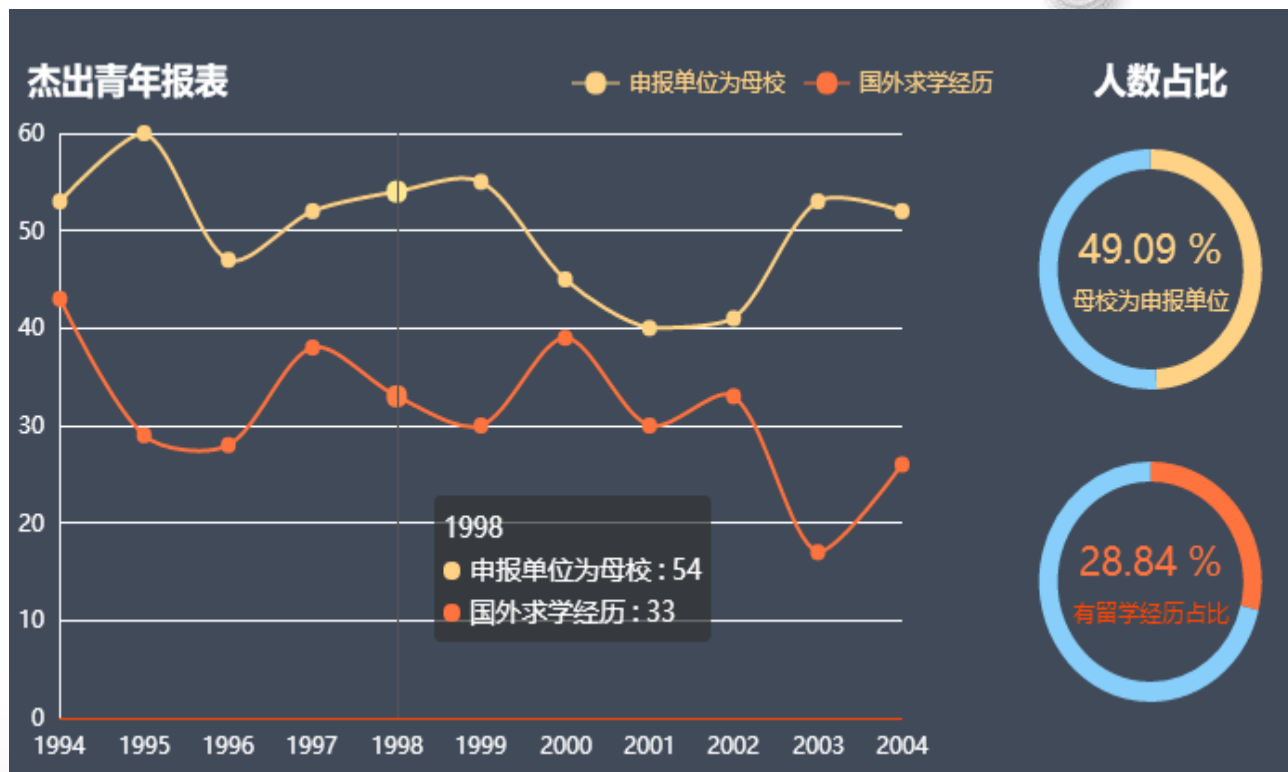


母校与留学



3) 可视化展示

观察黄色与橙色两条线，有一个有趣的现象：除了2000年到2002年，其他年份两条线的增减是相反的，母校为申报人数增多，获取称号的留学经历人数反而减少。



四象限图



1) 数据清洗

该需求只需要清洗杰青的年份、姓名、生日、本科毕业时间四个属性：

| | | | |
|------|-----|------|------|
| 1996 | 彭晓峰 | 1975 | 1983 |
| 1999 | 黎燕 | 1974 | 1996 |
| 1997 | 吴平 | 1970 | 1994 |
| 1997 | 马余刚 | 1968 | 1989 |
| 1994 | 陈永川 | 1964 | 1984 |
| 1995 | 麻生明 | 1965 | 1986 |
| 1999 | 李儒新 | 1969 | 1990 |
| 2002 | 封东来 | 1972 | 1994 |
| 2003 | 周志华 | 1973 | 1996 |
| 1994 | 魏炳波 | 1964 | 1983 |
| 1995 | 张亚平 | 1965 | 1986 |
| 1994 | 席南华 | 1963 | 1981 |
| 1995 | 张伟平 | 1964 | 1985 |
| 1996 | 卢柯 | 1965 | 1981 |

四象限图



2) 统计结果

根据清洗后的数据计算出杰青获得称号时的年龄以及本科毕业到获得杰青用了多少年，因为在数据收集过程中我们发现杰青年龄跨度很大，而且想他们本科毕业奋斗到获得国家认可需要多少年。

| | | | | | |
|------|-----|------|------|----|----|
| 1999 | 黎燕 | 1974 | 1996 | 25 | 3 |
| 1997 | 吴平 | 1970 | 1994 | 27 | 3 |
| 1997 | 马余刚 | 1968 | 1989 | 29 | 8 |
| 1994 | 陈永川 | 1964 | 1984 | 30 | 10 |
| 1995 | 麻生明 | 1965 | 1986 | 30 | 9 |
| 1999 | 李儒新 | 1969 | 1990 | 30 | 9 |
| 2002 | 封东来 | 1972 | 1994 | 30 | 8 |
| 2003 | 周志华 | 1973 | 1996 | 30 | 7 |
| 1994 | 魏炳波 | 1964 | 1983 | 30 | 11 |
| 1995 | 张亚平 | 1965 | 1986 | 30 | 9 |
| 1994 | 席南华 | 1963 | 1981 | 31 | 13 |
| 1995 | 张伟平 | 1964 | 1985 | 31 | 10 |
| 1996 | 卢柯 | 1965 | 1981 | 31 | 15 |
| 1999 | 李卫东 | 1968 | 1989 | 31 | 10 |
| 2003 | 刘嘉 | 1972 | 1995 | 31 | 8 |
| 2000 | 刘海燕 | 1969 | 1964 | 31 | 36 |
| 1995 | 龚旗煌 | 1964 | 1983 | 31 | 12 |
| 1998 | 谢毅 | 1967 | 1988 | 31 | 10 |
| 2000 | 张爱民 | 1969 | 1991 | 31 | 9 |
| 1994 | 彭练矛 | 1962 | 1982 | 32 | 12 |

迁徙图



1) 1994年可视化展示

可以看出杰青都
往北京地区集中，趋
势明显。



迁徙图



2) 1995年可视化展示

可以看出杰青迁徙集中在中部和东部，比94年更多样化。



迁徙图



3) 1996年可视化展示

可以看出杰青都往江浙地区集中，趋势明显。



迁徙图



4) 1997年可视化展示

可以看出杰青都往东南地区集中，趋势明显。



迁徙图



5) 1998年可视化展示

可以看出杰青都
往北京地区集中，趋
势明显。



迁徙图



6) 1999年可视化展示

可以看出杰青都往东部地区集中，趋势明显。



迁徙图



7) 2000年可视化展示

可以看出杰青都往中部东部地区集中，趋势明显。



迁徙图



8) 2001年可视化展示

可以看出杰青都
往北京地区集中，趋
势明显。



迁徙图



9) 2002年可视化展示

可以看出趋势呈现多样化，集中在五个地区。



迁徙图



10) 2003年可视化展示

可以看出杰青都
往北京地区集中，趋
势明显。



迁徙图



11) 2004年可视化展示

可以看出集中趋势多样化，总体还是从南往北走。



迁徙图



12)总结

杰青的教育迁徙趋势都是趋向教育和经济发达的地区，大致趋势都是从南往北走。直到2004年出现多种方向多种集中趋势，这是好事，说明更多的杰青投身于中部建设中，这样资源能更公平一些。



Thank You !

