高性能计算

——长江学者(2005-2011)

第四组

姓名	学号
易光启	ZF1721405
袁欢	ZF1721414
白玲	ZF1721101
党高峰	ZF1721118
相然	ZF1721340
董振	ZF1721124

目 录

1. 数据搜集
2. 数据整理
3. 数据分析
4. 课程总结

1.数据搜集

· 从2005年至2011年，一共有1269杰出青年，我们组每个人大概负责211个人左右信息的搜集
· 搜集的主要内容有：姓名、出生日期、性别、籍贯、单位及其经纬度、研究方向、获取称号时间、获取称号单位及其经纬度、本硕博毕业学校及其经纬度、本硕博专业、本硕博毕业时间。

2.数据整理

数据存在问题

(1)、信息不全
(2)、日期不统一
(3)、位置不统一

数据规范化

(1)、缺的属性统一补”NULL”

(2)、所有日期统一到年份

(3)、单位统一到一级单位

(4)、籍贯统一到市级

地名与单位转经纬度

(1)、申请百度地图API

(2)、python代码

(3)、转化结果

(1)、申请百度地图API


创建应用
回收站
每页显示30条
应用编号
应用名称
访问应用（AK）
应用类别
备注信息
(双击更改)
应用配置
11090767
map
服务端
设置 删除

(2)、python代码

```
def getlnglat(address):
    url = 'http://api.map.baidu.com/geocoder/v2/'
    output = 'json'
    ak = [redacted]
    add = quote(address) #由于本文城市变量为中文，为防止乱码，先用quote进行编码
    uri = url + '?' + 'address=' + add  + '&output=' + output + '&ak=' + ak
    req = urlopen(uri)
    res = req.read().decode() #将其他编码的字符串解码成unicode
    temp = json.loads(res) #对json数据进行解析
    return temp
```

(3)、转化结果

自动读取excel文档，并将转化的经纬度结果自动写入到excel文档对应的位置，实现批量自动转化，节省了人力，提高了效率

部分运行结果如下

获取称号单位		
单位名称	纬度X	经度Y
北京大学	39.999	116.315
北京应用	39.9622	116.381
复旦大学	31.3023	121.506
上海交通	31.2044	121.441
中科院数	30.6535	117.496
中科院数	30.6535	117.496
哈尔滨工	22.5434	113.961
湖南大学	28.1847	112.952
上海大学	31.282	121.465

3.数据分析

(1)、工具

(2)、分析内容

(3)、分析结果

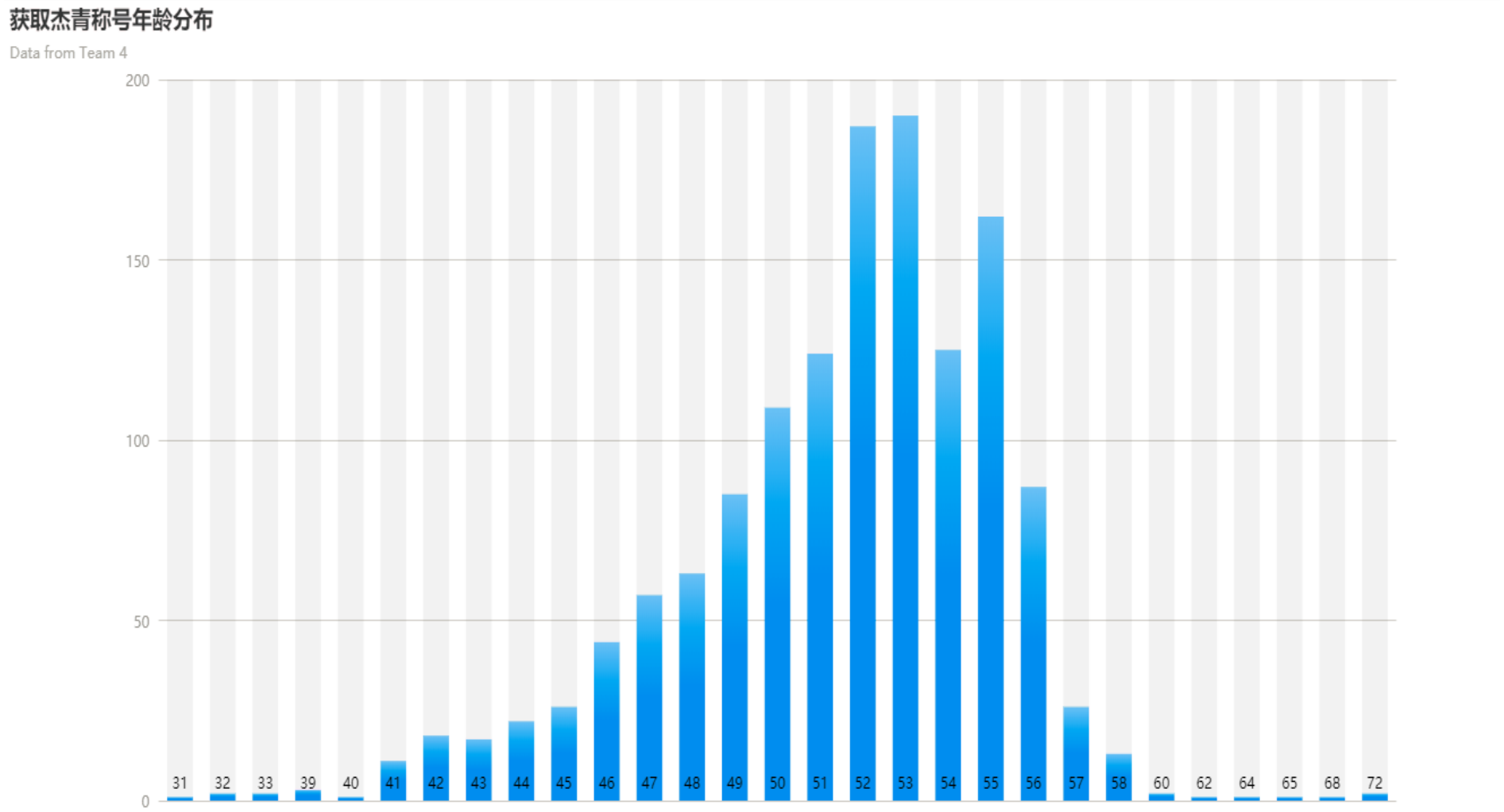
(1)、工具

百度Echarts+百度地图

(2)、内容

- 获称号与年龄的关系
- 籍贯与工作地的关系

获取称号与年龄的分布图



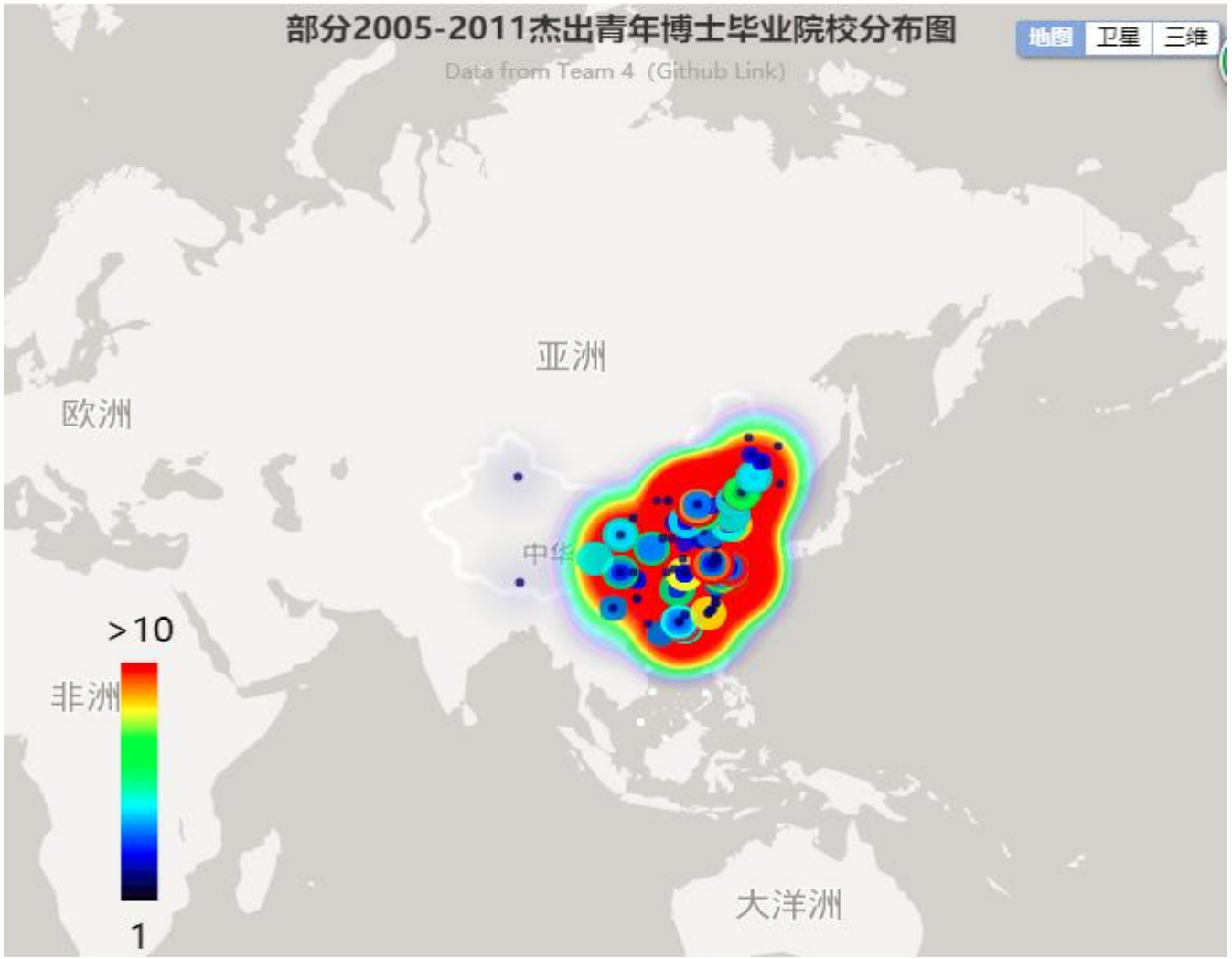
获奖与年龄的分析

本次研究内容为杰出青年现在年龄与获奖年龄分布情况。结合数据可视化的方法，我们做了部分杰出青年年龄与获奖年龄的分布图，横轴为年龄，纵轴为人数，得到杰出青年现在年龄与人数的柱状图。。

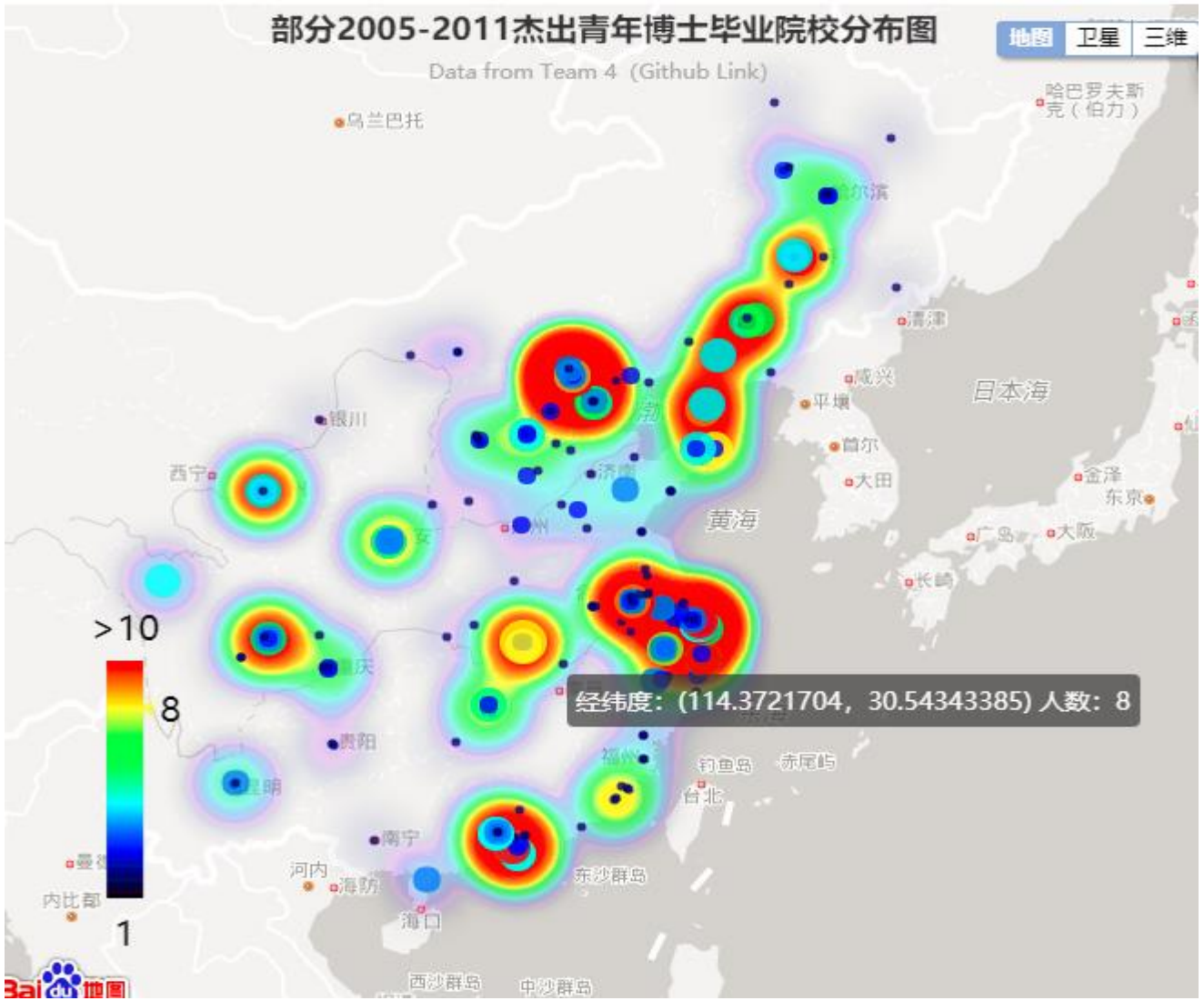
从分析结果来看，杰出青年获奖年龄段48-56岁所占比重最大。原因在于学习积淀的时间决定了大多数人50岁左右才会在学术上有所建树，不排除有些人十分勤奋、天资聪颖或者其他原因，在更低的年龄就在学术上有所建树。年龄比较大的人里面，60岁以上的，他的精力什么的有限，所以获得建树的人也比较少。

所以大家应该在现在努力积蓄力量，争取的50岁左右也获得很大的成就，毕竟按照大数据统计的情况来看是这样的，这个年龄获得成功的可能性还是很大的，但是在那个年龄之前，也要努力，万一成了那一小部分的人呢，当然在年龄小的时候没有太大成就不要气馁，毕竟很多人都是在50岁左右才获得。

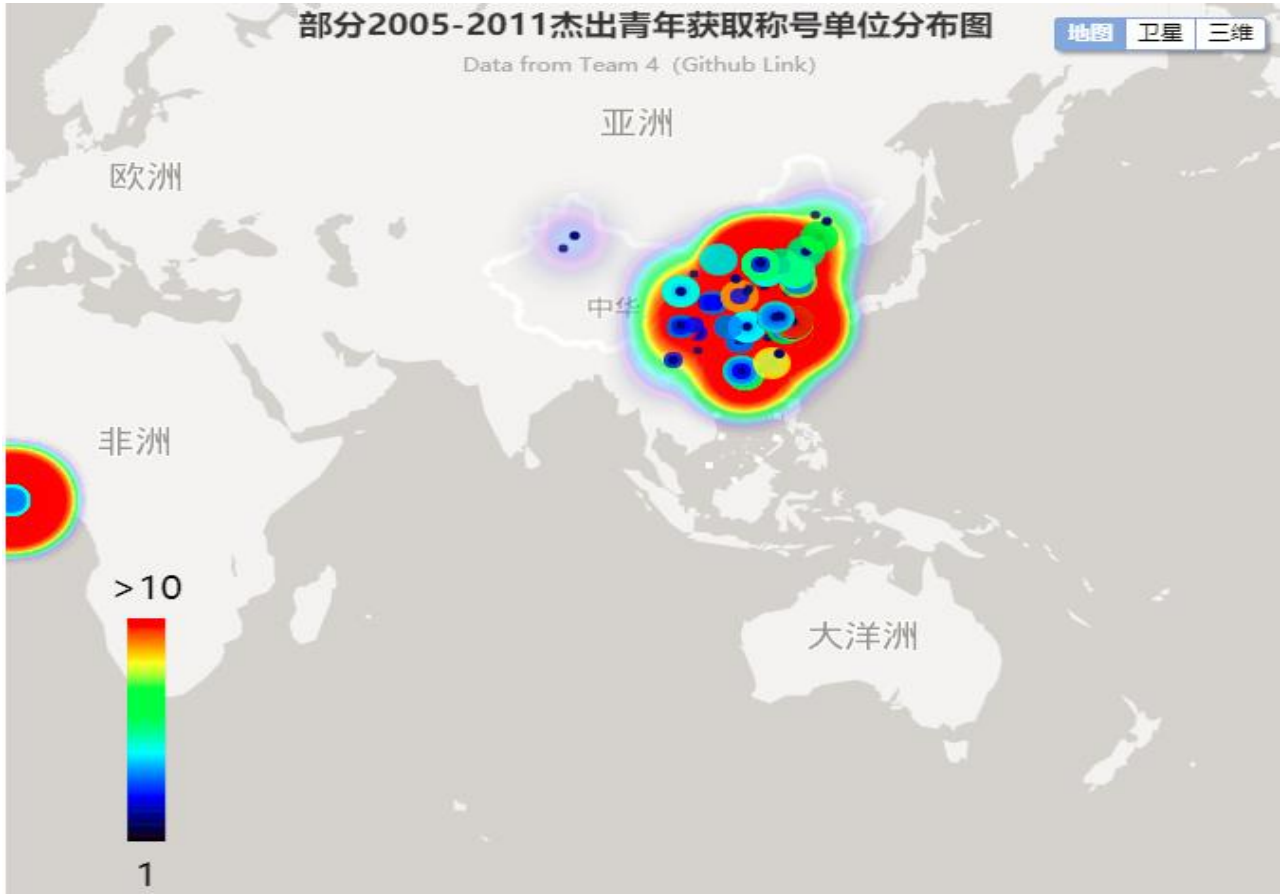
- # 博士毕业院校分布图


部分2005-2011杰出青年博士毕业院校分布图
Data from Team 4 (Github Link)
地图
卫星
三维
亚洲
欧洲
非洲
大洋洲
>10
1

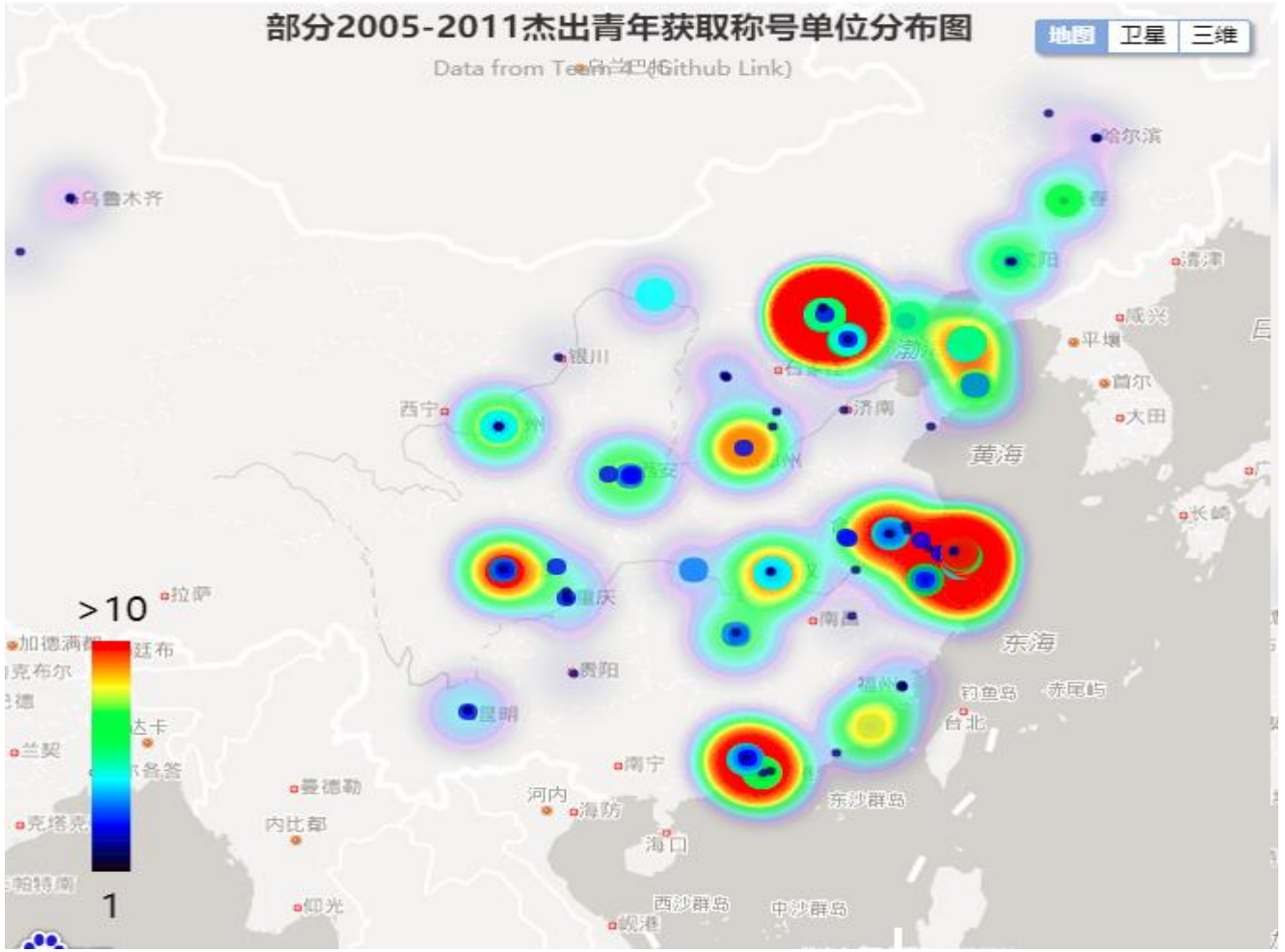
博士毕业院校分布图


部分2005-2011杰出青年博士毕业院校分布图
Data from Team 4 (Github Link)
地图
卫星
三维
日本海
黄海
>10
8
1
经纬度：(114.3721704，30.54343385) 人数：8

- # 工作单位分布图


部分2005-2011杰出青年获取称号单位分布图
Data from Team 4 (Github Link)
地图
卫星
三维
亚洲
欧洲
非洲
大洋洲
>10
1

工作单位分布图

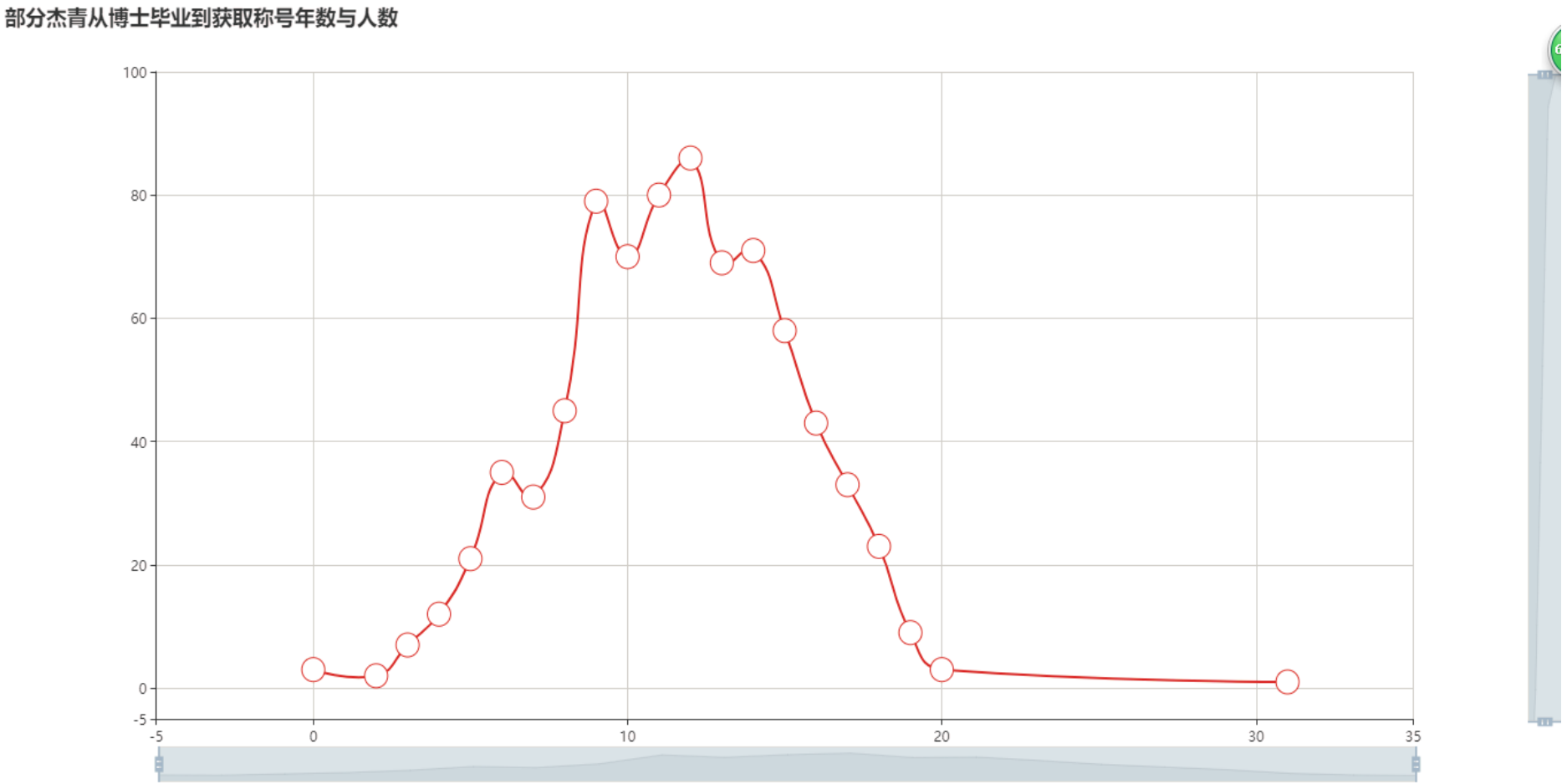

部分2005-2011杰出青年获取称号单位分布图
Data from Team (Github Link)
地图
卫星
三维
>10
1
哈尔滨
乌鲁木齐
银川
西宁
济南
黄海
东海
重庆
贵阳
昆明
南宁
河内
海防
海口
东沙群岛
西沙群岛
中沙群岛
台北
首尔
大田
平壤
清津
长崎
拉萨
内比都

地理分布的分析

本次研究内容为杰出青年博士毕业院校分布情况和现今所在地分布情况。结合数据可视化的方法，我们在地图上用不同的颜色标记出杰出青年们的博士毕业院校和现在工作学习的地点，颜色越暖（如红色）的地方代表人数越多，颜色越冷（如蓝色）的地方代表人数越少，能给人以很清晰的视觉认知。

从分析结果来看，北京、上海、广东、深圳等沿海地区的人数相对也比较多。而西北地区则比较少。其中的原因一方面大概是经济发达的地区往往更重视教育，另一方面经济发达的地区也有更多的钱力物力可以投资在教育上。而考虑杰出青年们现今分布的情况，则多处在高校或研究所分布较多的城市和省份，这个教育资源和当地的经济发展情况是基本上呈现正相关的。只有在人们的积极基础达到了一定的程度，才有可能去考虑更高的追求，去发展教育，人们的眼界和经济基础也是很相关的，所以说经济基础决定上层建筑是很正确的。

博士毕业到获取称号年数分布图



毕业到获得称号时间的分析

本次研究内容为杰出青年博士毕业到获得称号的时间的统计结果，用折线图表示了出来，
横坐标是毕业时间，纵坐标是在毕业多少年后获得称号的人数，曲线的高度越高，表明这个人数越多。

从分析结果来看，博士毕业十年左右获得这个称号的人是很多的，因为这个工作十年，积累的人脉，自己的能力、经验什么的都达到了一定的程度，而且这个时间也正是自己身体比较好的时候，身体是一切革命的本钱。但在这个时间之前，经验不足，能力不够，人脉积累不够，所以成功的可能性不大。但是工作很多年以后，这个时候人已经疲惫了，没有那么多的精力去追求这些了。

· 数据可视化分析结果（GitHub）

4.课程总结

通过这门课的学习，让自己学会了一些方法，比如从多个维度去分析总结，去发现数据之间的关联，横向、纵向、时间、空间等不同的维度，给自己的探索欲望提供了支持，以后可以通过这些方法去总结某些规律。

谢谢老师这学期的讲解，让我们开拓了视野，增长了见识。

- 最后

本次作业所有相关的内容已传到GitHub

github:https://github.com/AcKnight/CJSDataTeam4

谢谢大家