

# 网络科学第二讲

(应用案例--从《中国历代人物传记》中发现新知识)

2019版

罗铁坚

课程: <https://tjluo-ucas.github.io/ns/>

# 提纲

1. 文本实体抽取
2. 人物关系数据库
3. 人物关系学（Prosopography）
4. 时空数据分析（Spatial analysis）
5. 社会网络分析（Social network Analysis）

# 网络科学研究方法

## 基本假设：

- 1、社会、自然或技术“系统”是由不同部分有机组成。
- 2、这些关联的部分是存在某种“结构”的。
- 3、个体部分之间通过“结构”链接进行交互使系统得以运行。
- 4、系统内部的个体之间相互依存，并且任何个体的行为结果潜在地依赖其他个体的联合行为。
- 5、利用图论的知识来探讨系统的网络结构；研究个体节点的行为规律，采用博弈论的语言来建立基本模型。

## 研究方法：

- 1、确定研究问题（事件发生的原因或预测可能发展的趋势）
- 2、建模：构建网络图（节点和边的定义，需要领域知识）
- 3、分析：节点的行为模式或变化，全图的演化。
- 4、归纳：模式总类、发生的原因、调控手段
- 5、实验仿真和实际应用：收集什么数据、进行什么分析、给出何种决策。



# 中國歷代人物傳記資料庫

中央研究院歷史語言研究所、哈佛大學、北京大學中國古代史研究中心合作開發  
Developed through collaboration among Academia Sinica, Harvard University, and Peking University

何多士  
此公雅志  
高和始下  
惟美友  
然中  
福勝於  
於國神  
性厚  
生多  
於意醉  
北坡草堂

Lü Zuqian, whose style name was Bogong, was a grandson of the Right Assistant Director to the Imperial Secretary Haowen. His family had lived in Wuzhou since his grandfather's generation. The learning of Zuqian was based on family [tradition], and embodied the textual transmission from the Central Plain. When he grew up, Zuqian studied with Lin Zhiqi, Wang Yingchen, and Hu Xian respectively. Then he also befriended Zhang Shi and Zhu Xi, and his explication and inquiry became more sophisticated.

First he obtained official rank by way of the protection privilege. But later he obtained his Jinshi degree and also passed the special decree examination for "Erudite Learning and Exceptional Literary Composition." Then he was appointed to the School for the Imperial Clan in the Southern Outer Office. During the mourning period for his mother, when he stayed in Mt. Mingzhao (in Wuyi), literati from all directions raced there. He was appointed Erudite in the National University.

呂祖謙字伯恭，尚書右丞好問之孫也。自其祖始居婺州。祖謙之學本之家庭，有中原文獻之傳。長從林之奇、汪應辰、胡憲游，既又友張栻、朱熹，講索益精。

初，蔭補入官，後舉進士，復中博學宏詞科，調南外宗教。丁內艱，居明招山，四方之士爭趨之。除太學博士

“Factoids” in biographical texts

文本实体抽取

# 观察文本和问题，提出需要抽取的实体类型和值

1. Basic Biography: name, gender, dates (relationship to time)
2. Biography Addresses (relationship with places)
3. Alternate Names (relationship to names of different kinds)
4. Writings (relationship to learning)
5. Postings (relationship with government)
6. Mode of Entry into Government
7. Kinship (kin relationships with others)
8. Associations (non-kin relationships to others)
9. Social Status (relationship with modes of social distinctiveness)
10. Possessions (relationship to property as giver and receiver)
11. Events (participation in events)

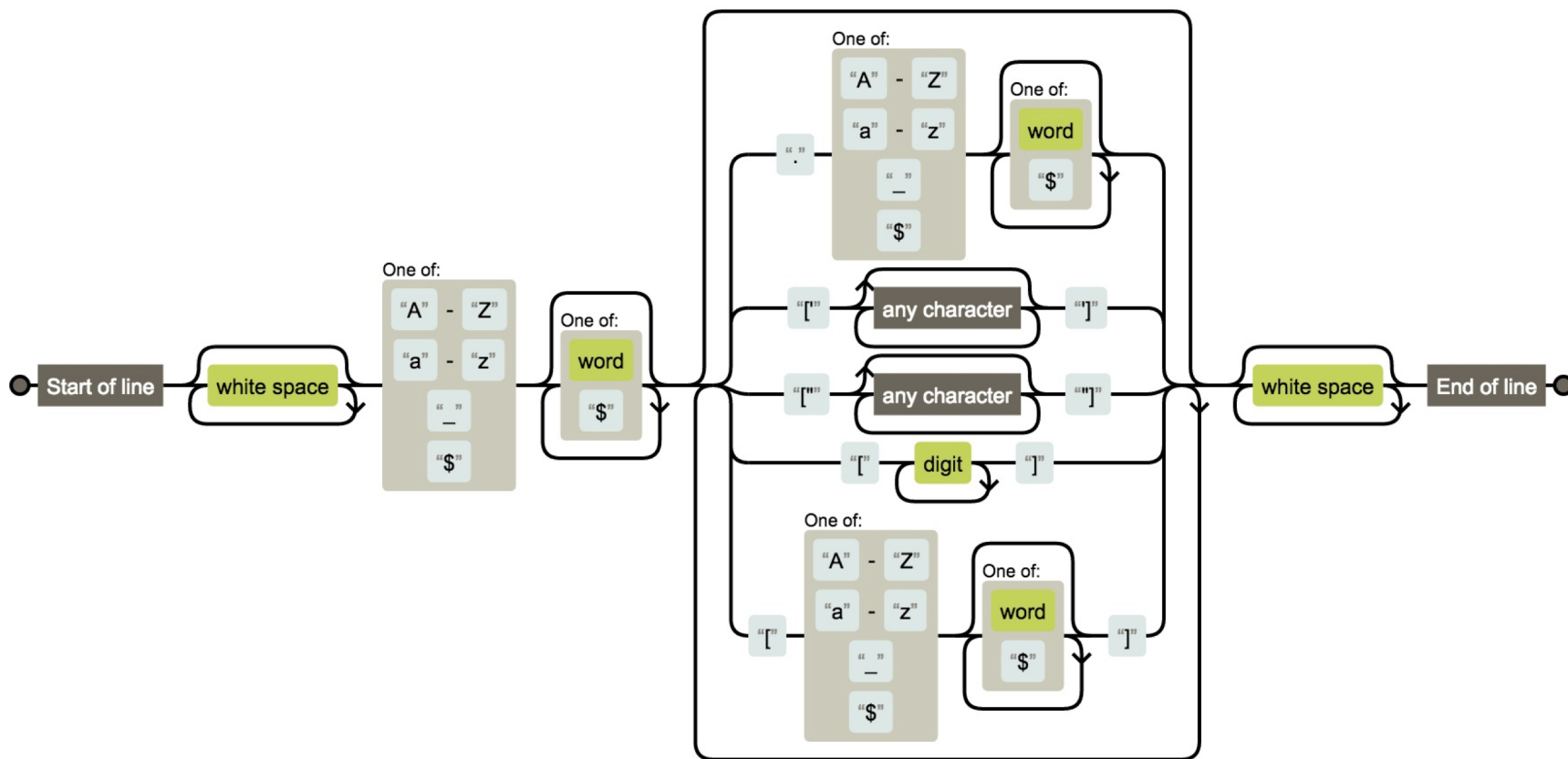
# 抽取实体的工具：正则表达式

- 正则表达式，又称规则表达式。（英语：Regular Expression，在代码中常简称为 regex、regexp 或 RE ），计算机科学的一个概念。**正则表达式通常被用来检索、替换那些符合某个模式(规则)的文本。**
- 正则表达式是对字符串（包括普通字符（例如，a 到 z 之间的字母）和特殊字符（称为“元字符”））操作的一种逻辑公式，就是用事先定义好的一些特定字符、及这些特定字符的组合，组成一个“规则字符串”，这个“规则字符串”用来表达对字符串的一种过滤逻辑。正则表达式是一种**文本模式**，模式描述在搜索文本时要匹配的一个或多个字符串。
- Python 从 **1.5** 版本开始增加了 **re** 模块，它提供 Perl 风格的正则表达式模式。re 模块使 Python 语言拥有全部的正则表达式功能。

为了证明正则表达式功能的强大之处，我们先用个小例子体现一下：

# 正则表达式的字符串模式匹配的原理

```
/^\\s*[A-Za-z_]([\\w$]*(?:\\. [A-Za-z_]([\\w$]*|\\['.*?'\\]|\\[".*?"\\]|\\[\\d+\\]|\\[[A-Za-z_]([\\w$]*\\)))*\\s*$
```





# 正则表达式实际应用例子

reign title number & year character death character

(元祐|元豐|...|紹興)[0-9]+年卒

## WDY

丁罕 (?~999), 潁州人。應募補衛士, 以戰功累遷指揮使。淳化中為澤州團練使, 知霸州, 河決, 以私錢募築, 民咸德之, 擢領靈環路都部署, 破李繼遷有殊功。後拜密州觀察使, 徙貝州。  
咸平二年卒, 子守德能世其家。

丁明 (1127~1211), 舊名騫, 字希閔, 後改名明, 字子公, 金壇人, 權子。閉門讀書二十年, 手編事類及史通考等書百餘卷。奉祠家居, 嘉定四年卒, 年八十五。鄉里私謚博雅先生。



# 用正则表达式找出命名实体 Named Entity Recognition (NER)

至元十八年为东阳县丞

年号	年号代码	任期	年份	地名	地名代码	官名	官名代码
至元	623	18	1281	东阳	18340	县丞	841

# 用正则表达式找出命名实体等的代码

```
1  #!/usr/bin/python2.5
2  import sys
3  import codecs
4  import getopt
5  import re from removeDups
6  import removeDups from copy
7  import copy# chinese string printing
8  def cprint(s):
9      print s.encode('utf-8')# name to string when ID unknown
10 def name2str(ids, name):
11     s = '<%s,' % name
12     for id in ids[name]:
13         s += '%s/' % id
14         s = s[:-1] s += '>'
15     return s # name to string when ID known
16 def name2strID(id, names, intervals):
17     name = names[id]
18     s = '<%s,%s,' % (name, id)
19     if intervals[id][0] and intervals[id][1]:
20         s += '%d~%d>' % (intervals[id][0], intervals[id][1])
21     else:
22         s += 'None~None>'
23     return s
```

# 从《中国历代人物传记》中发现新知识

## The China Biographical Database

– from anecdote to data to knowledge

文本实体抽取

人物关系数据库

人物关系学 (Prosopography)

时空数据分析 (Spatial analysis)

社会网络分析 (Social network Analysis)

# 中國歷代人物傳記資料庫

中央研究院歷史語言研究所、哈佛大學、北京大學中國古代史研究中心合作開發  
Developed through collaboration among Academia Sinica, Harvard University, and Peking University

## A Subset of the Data on Sima Guang: 有关司马光的部分数据：

Name 姓名	Dates 日期	Offices 任官	Associations 社会关系
Sima Guang 司馬光	1019- 1086	(1) 1059 度支勾 院 Budget Auditor; (2) 1085 门下侍郎 Executive of the Chancellery; (3) 1086 左仆射兼 门下侍郎 Left Executive, Dept of Ministries [.....]	(1) Yuanyou coalition member (元佑党); (2) An Dun 安惇 Desires opposed by; (3) Chao Buzhi 晁补之 Sacrificial prayer written by; (4) Chen Jian 陈荐 Sacrificial prayer written for; (5) Chen Min 陈敏 Honored by; (6) Cheng Yi 程頤 Recommended; (7) Ding Du 丁度 Sacrificial prayer written for; (8) Fan Chunli 范纯礼 Patron of; [.....]

# 中國歷代人物傳記資料庫

中央研究院歷史語言研究所、哈佛大學、北京大學中國古代史研究中心合作開發  
Developed through collaboration among Academia Sinica, Harvard University, and Peking University



关系型数据库的多个实体

People 人物

Offices 职官

Association Types 社会关系

Name 姓名	Dates 日期
Sima Guang 司馬光	1019-1086

Person 人物	Posting Date 任命日期	Office Title 官名
Sima Guang 司馬光	1059	度支勾院 Budget Auditor
Sima Guang 司馬光	1085	门下侍郎 Executive of the Chancellery
Sima Guang 司馬光	1086	左仆射兼门下侍郎 Left Executive, Dept of Ministries

Person 人物	Association Type 关系类型	Associate 社会关系人
Sima Guang 司馬光	Yuanyou member (元佑党)	(not applicable)
Sima Guang 司馬光	Desires opposed by	An Dun 安惇
Sima Guang 司馬光	Sacrificial prayer written by	Chao Buzhi 晁补之
Sima Guang 司馬光	Patron of	Fan Chunli 范纯礼
Sima Guang 司馬光	Sacrificial prayer written for	Ding Du 丁度

# 中國歷代人物傳記資料庫

中央研究院歷史語言研究所、哈佛大學、北京大學中國古代史研究中心合作開發  
Developed through collaboration among Academia Sinica, Harvard University, and Peking University

可以按照不同字段分別排序

Name 姓名	Dates 日期
Sima Guang 司馬光	1019-1086

Person 人物	Posting Date 任命日期	Office Title 官名
Sima Guang 司馬光	1059	度支勾院 Budget Auditor
Sima Guang 司馬光	1085	門下侍郎 Executive of the Chancellery
Sima Guang 司馬光	1086	左仆射兼門下侍郎 Left Executive, Dept of Ministries

Person 人物	Association Type 关系类型	Associate 社会关系人
Sima Guang 司馬光	Yuanyou member (元祐黨)	(not applicable)
Sima Guang 司馬光	Desires opposed by	An Dun 安惇
Sima Guang 司馬光	Sacrificial prayer written by	Chao Buzhi 晁補之
Sima Guang 司馬光	Patron of	Fan Chunli 范純禮
Sima Guang 司馬光	Sacrificial prayer written for	Ding Du 丁度



## (1) Biographical data are *coded* and stored in *tables*. 传记数据以代码的形式存储于关系数据表中。

- BIOG\_MAIN
- Biography Addresses
- Alternate Names
- Writings
- Postings
- Mode of Entry into Government
- Kinship
- Associations
- Social Status
- Possessions
- Events
- 基本数据
- 地址数据
- 别名数据
- 著述数据
- 任官数据
- 入仕途径
- 亲属数据
- 社会关系数据
- 社会区分数据
- 财产数据
- 事件数据



BIOG_ADDR 地址
<b>Person ID</b> Addr Type ID Place ID, etc

ALT NAMES 别名
<b>Person ID</b> Name Type ID Alt Name, etc

WRITINGS 著述
<b>Person ID</b> Text ID, etc

ENTRY 入仕
<b>Person ID</b> Entry ID Year, etc

Data tables are linked to each other via *Person IDs*.  
这些数据表通过人物代码关联起来。

POSTINGS 任官
<b>Person ID</b> <b>Postings ID</b> Office ID Start Date End Date, etc

BIOG_MAIN 基本数据
<b>Person ID</b> Name 姓名 Born Died Index Year Choronym ID Dynasty ID, Etc

ASSOCIATIONS 社会关系
<b>Person ID</b> Assoc Relation ID Associate ID, etc

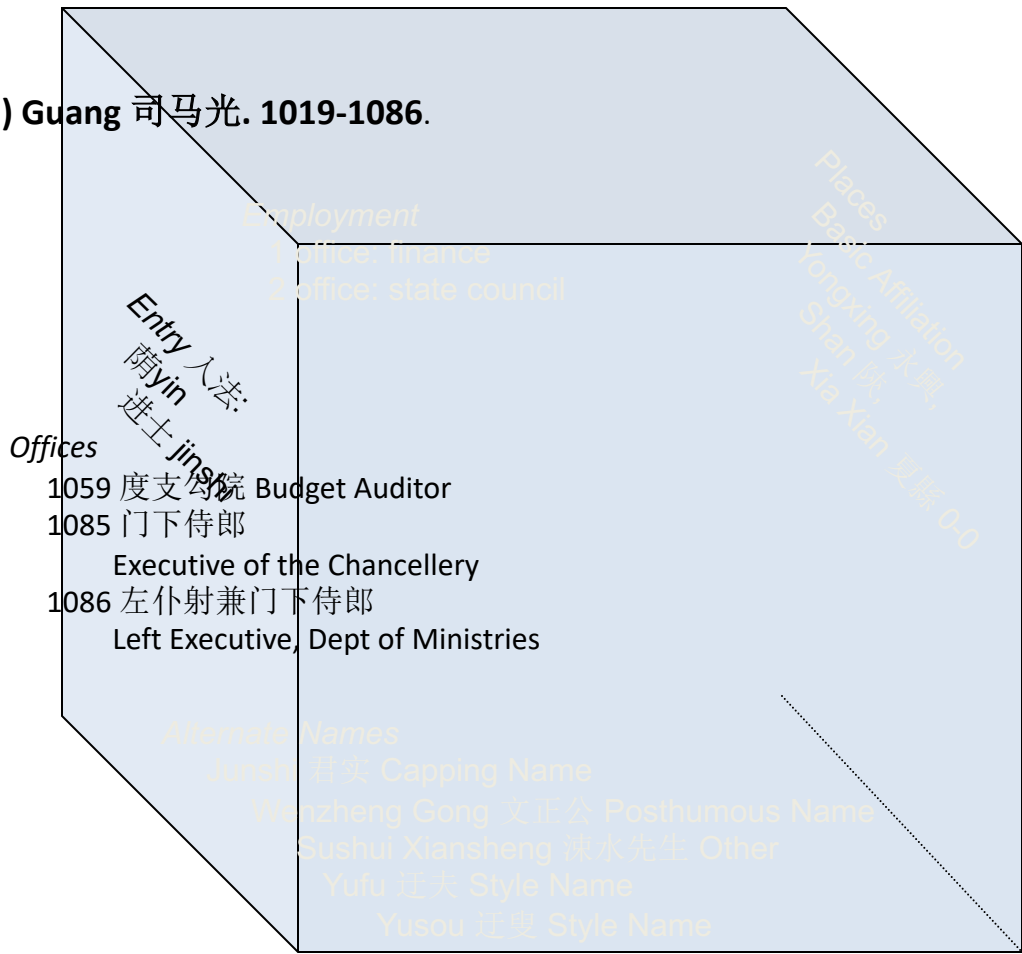
SOCIAL STATUS 社会区分
<b>Person ID</b> Status ID, etc

POST_ADDR 任官地
<b>Postings ID</b> Place ID, etc

KINSHIP 亲属
<b>Person ID</b> Kin Relation ID Kin ID, etc



Sima(1) Guang 司马光. 1019-1086.



人物的数据数据存在于各种实体（人物、地址等）中。

# 从《中国历代人物传记》中发现新知识

The China Biographical Database

– from anecdote to data to knowledge

文本实体抽取

人物关系数据库

人物关系学 (Prosopography)

时空数据分析 (Spatial analysis)

社会网络分析 (Social network Analysis)

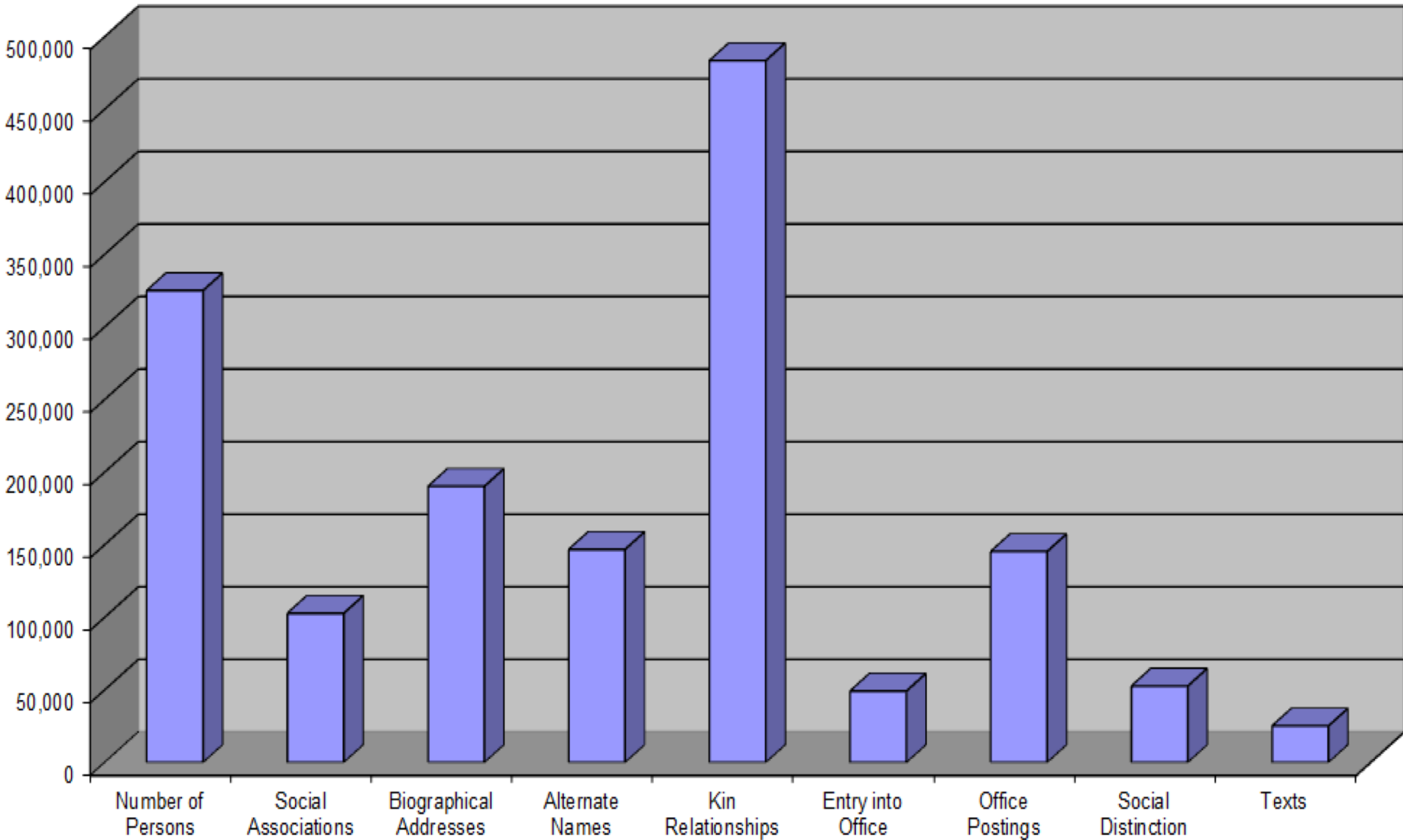
# Prosopography (人物关系学)

“人物关系学”探索研究一个特定群体的个体之间的关系。

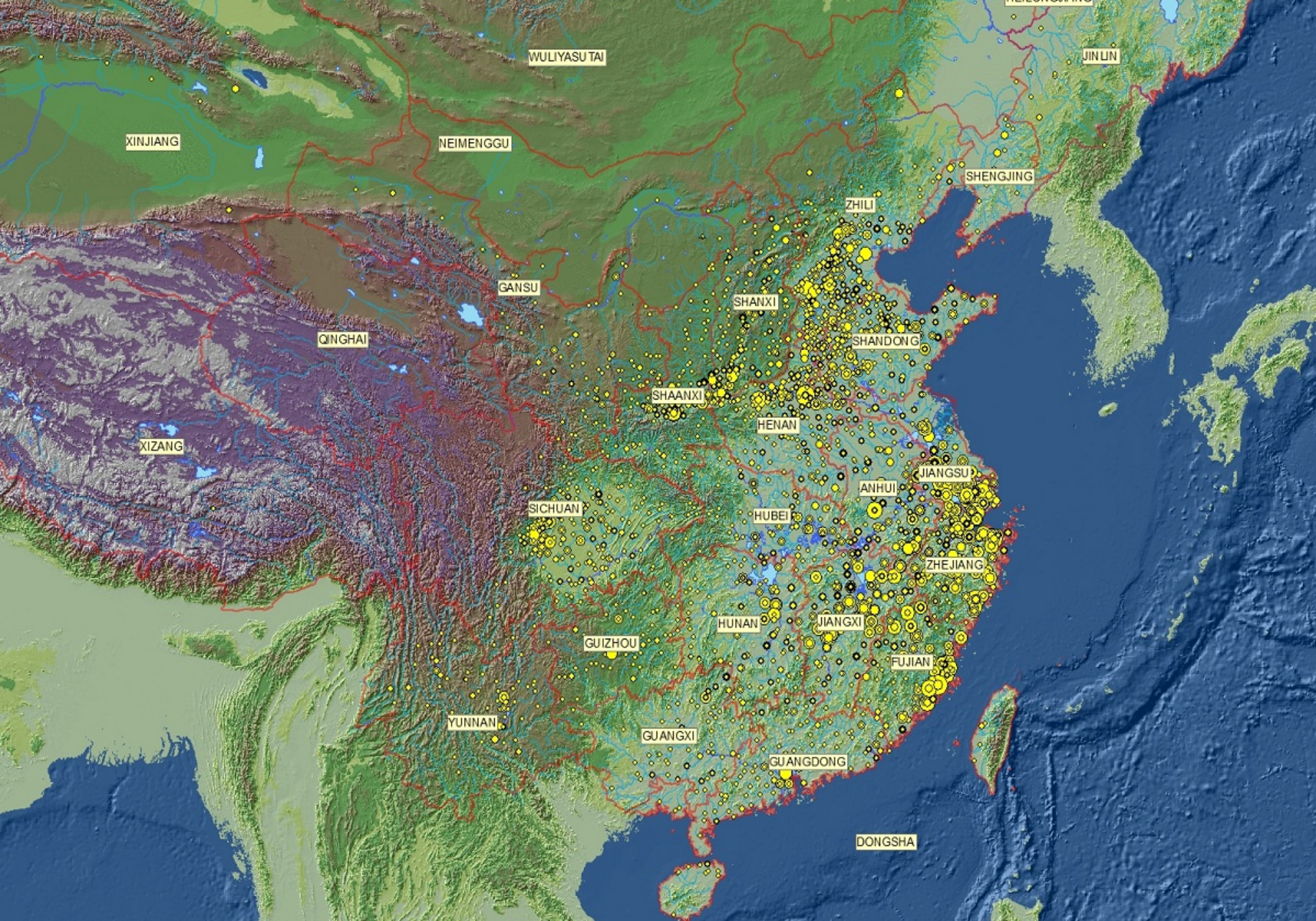
- 从历史文献中收集整理关于人物的出生和死亡、婚姻和家庭、社会出身和继承的经济地位、居住地、教育、个人财富的数额和来源,职业、宗教、任职经验等。
- 根据研究问题,对个体的各种类型的信息进行网络建模,推断或计算出个体间关系以及他们的影响程度等结论。



# CBDB Data December 2013

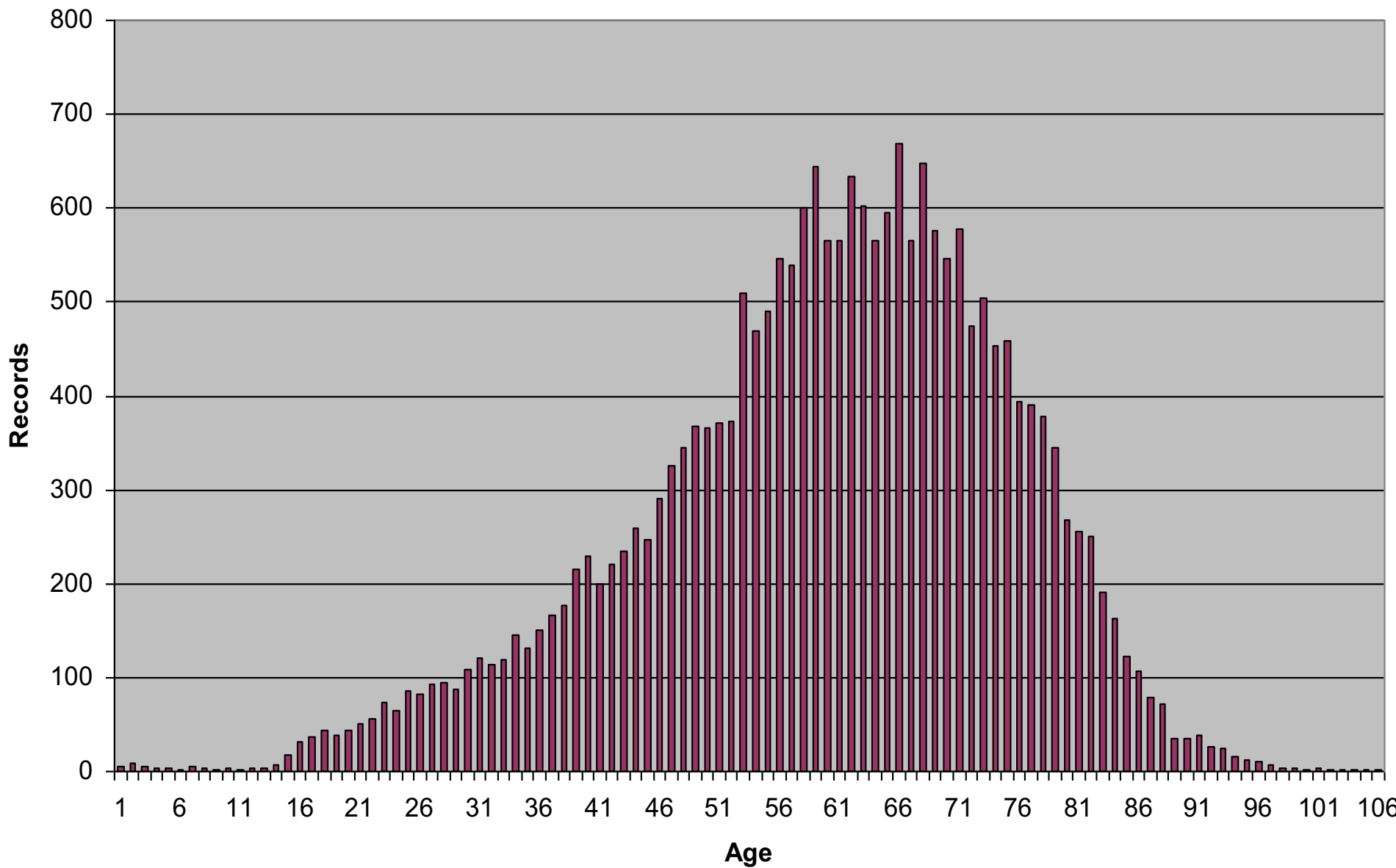


Period	Number
Tang	56432
5 Dynasties	1889
Song	47071
Liao	318
Jin	275
Yuan	19886
Ming	150585
Qing	37848
Minguo	3215
Other	20000



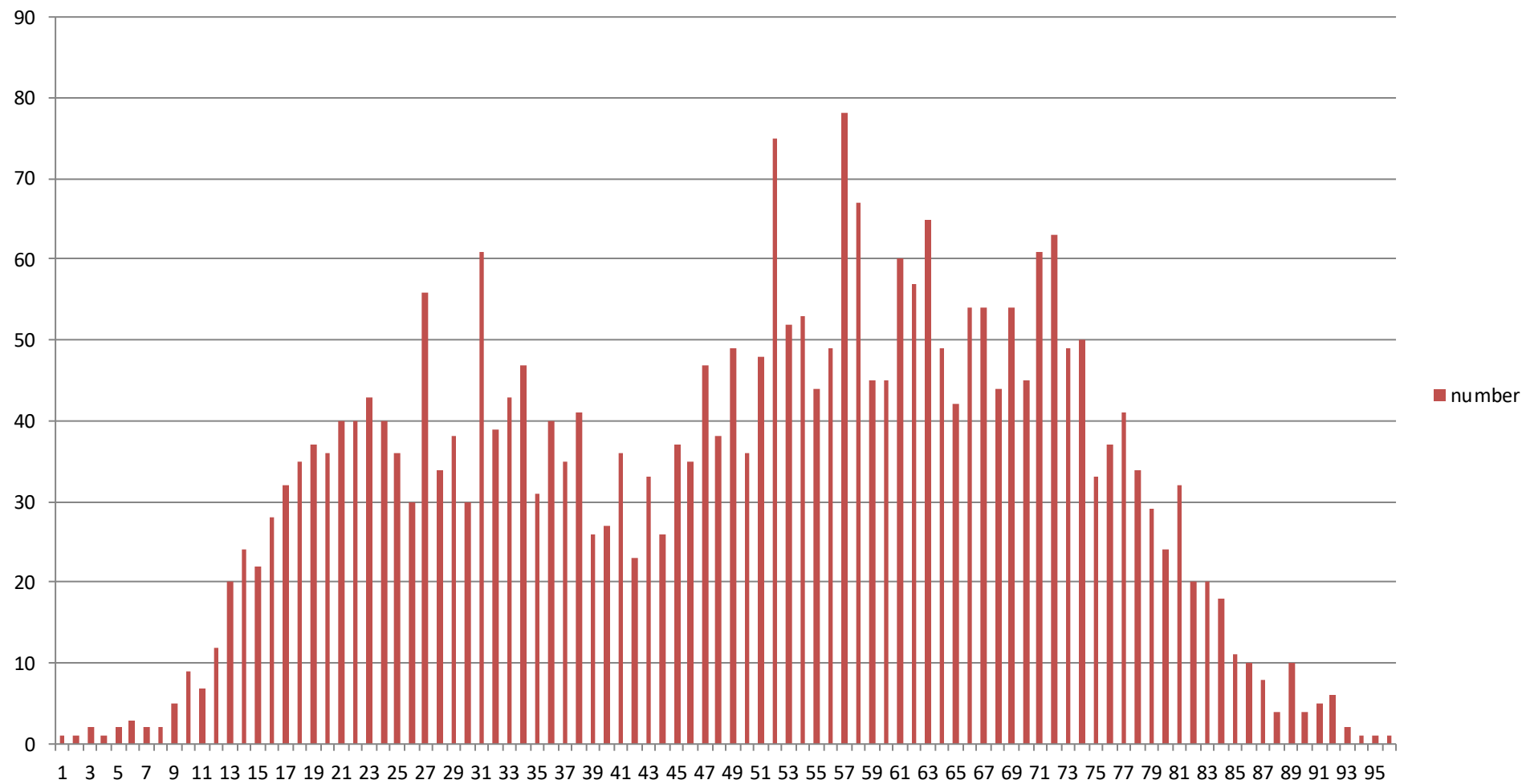

 中国地质大学  
 Cumulative Spatial Distribution of a Representative Sample of 67,000 CBDB Persons

# Age at Death-CBDB data Tang through Qing - 22270 persons





Age at Death of the 3072 Women in CBDB with Death Ages



# 从《中国历代人物传记》中发现新知识

The China Biographical Database

– from anecdote to data to knowledge

文本实体抽取

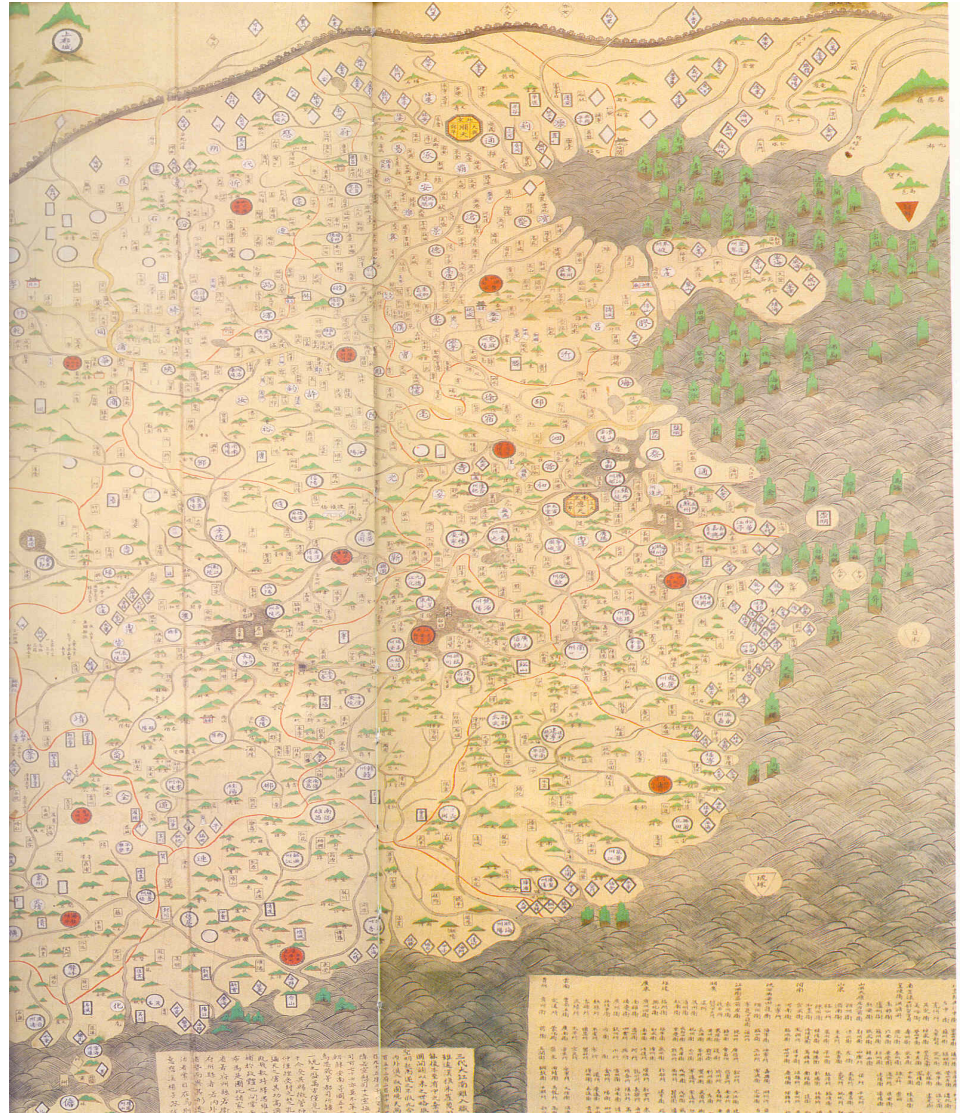
人物关系数据库

人物关系学 (Prosopography)

时空数据分析 (Spatial analysis)

社会网络分析 (Social network Analysis)

# 地图就是一个主张、命题和构想





# Chart of the Traces of Yu

每个方形的覆盖  
100 li

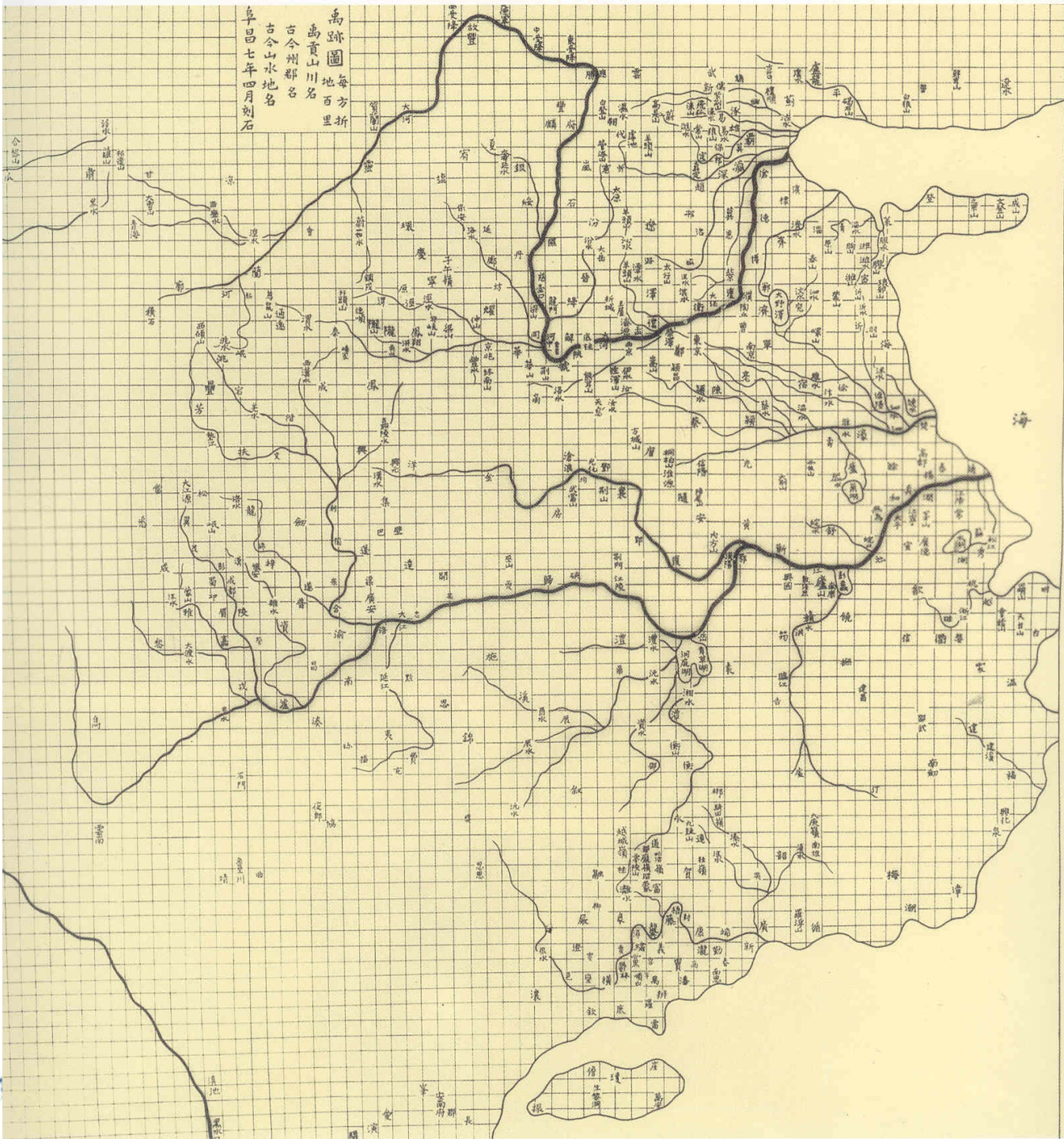
山和河的名字在  
“Tribute of Yu”

过去和现在都道府县  
的名称

过去和现在的山脉和  
河流的名称

Engraved in 4<sup>th</sup> month  
of the 7<sup>th</sup> year of  
Fuchang (=1136)

禹跡圖 每方折  
禹貢山川名  
古今州郡名  
阜昌七年四月刻石



# from the Army Map Service, ca. 1950/ CHGIS website

Harvard University Library Image Delivery

View full image in new window:



Click in thumbnail above to reposition detail segment.

**+** Zoom in



**-** Zoom out

3.22% of image shown

Screen size:

800 x 600

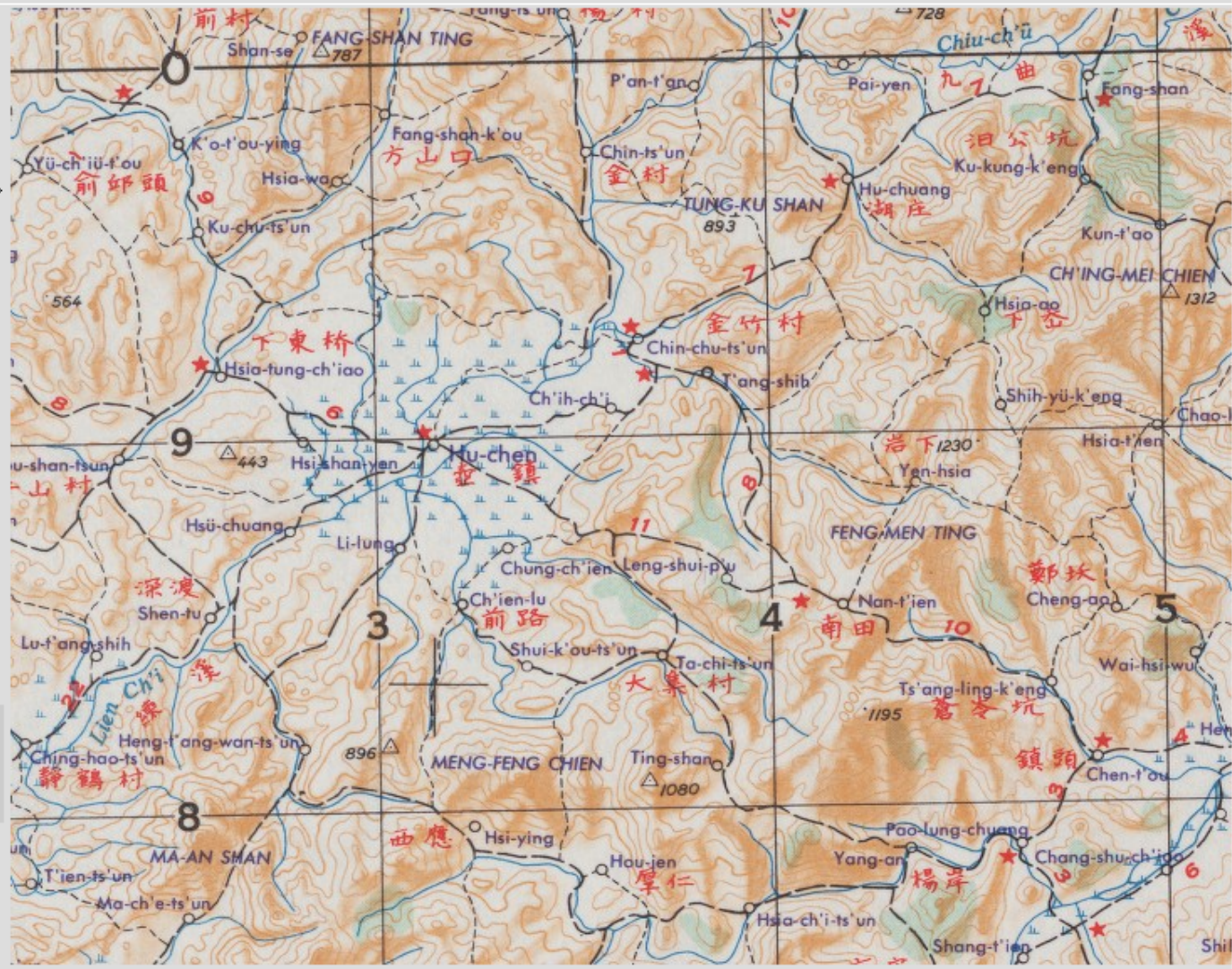
Clicking on image:

Zoom In

Rotate

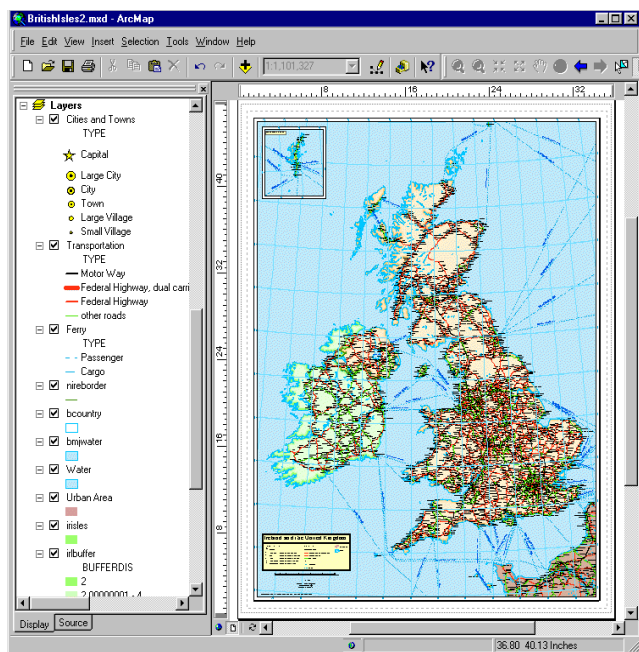
Help

[Copyright](#)

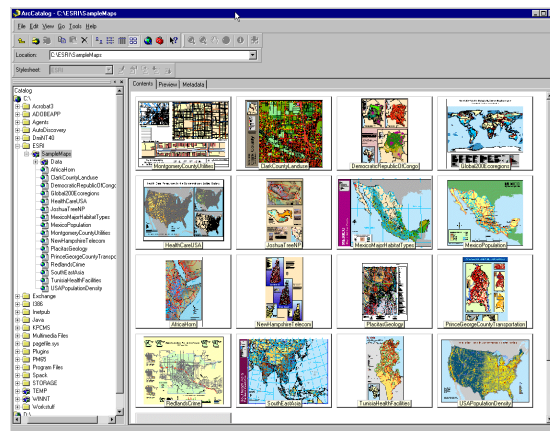
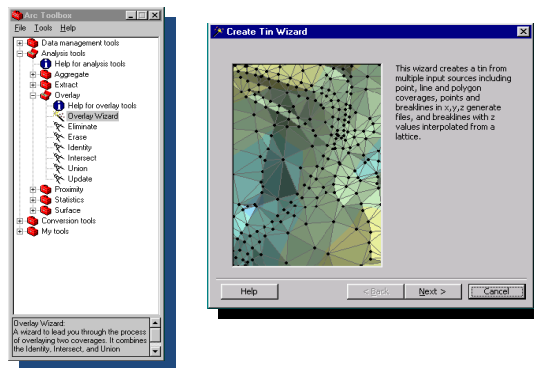


# GIS的三种视图

## 1. 智能交互式地图

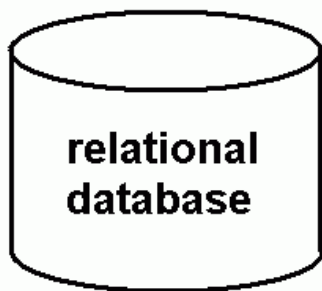


## 2. 一套工具和程序 – 执行任务的语言



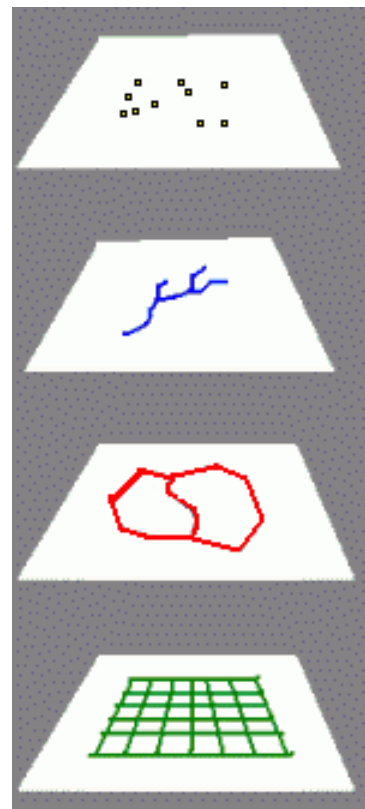
## 3. 管理信息系统

# 数字地名录



placename  
feature type  
coordinates  
valid date  
source note

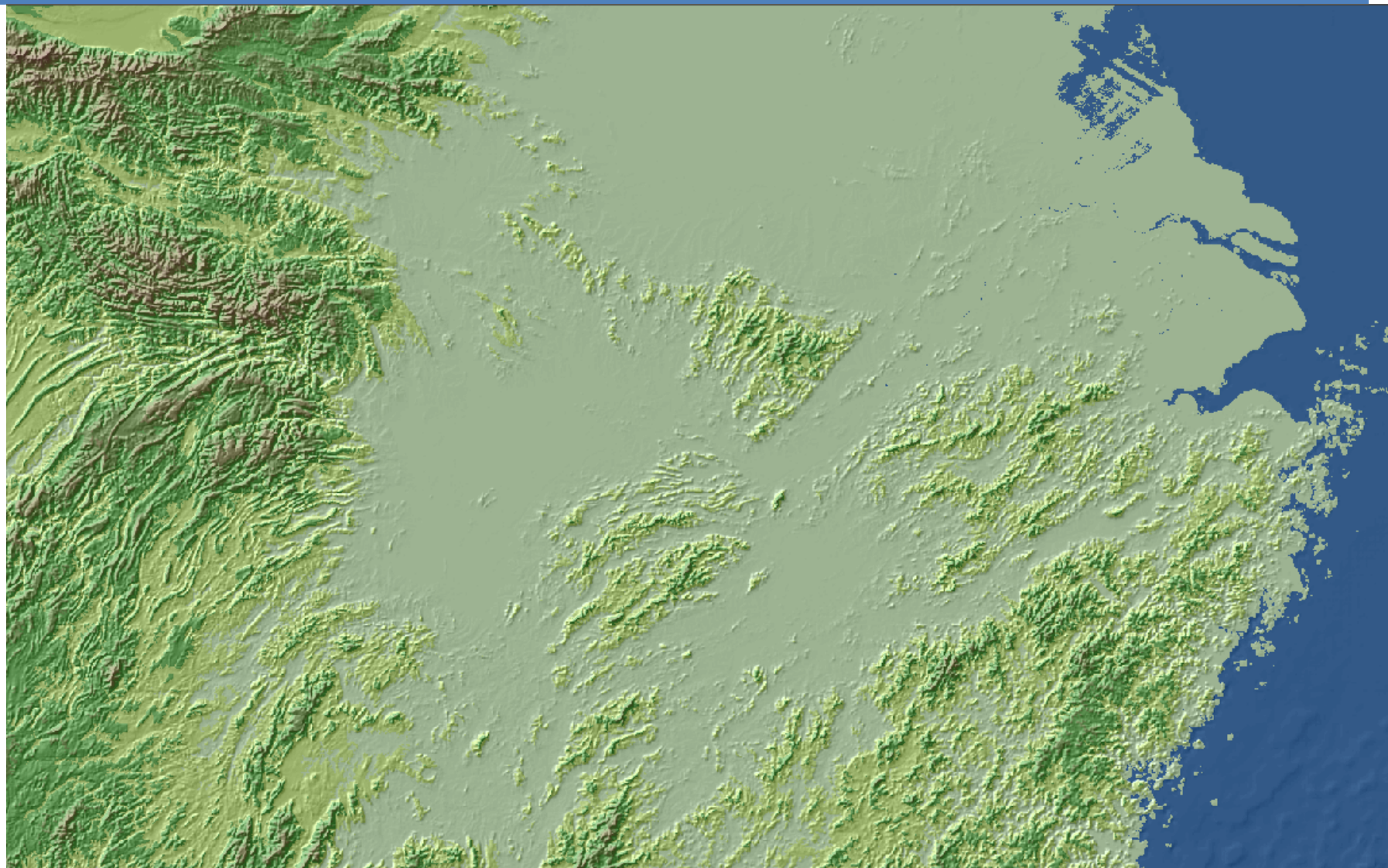
# GIS 层



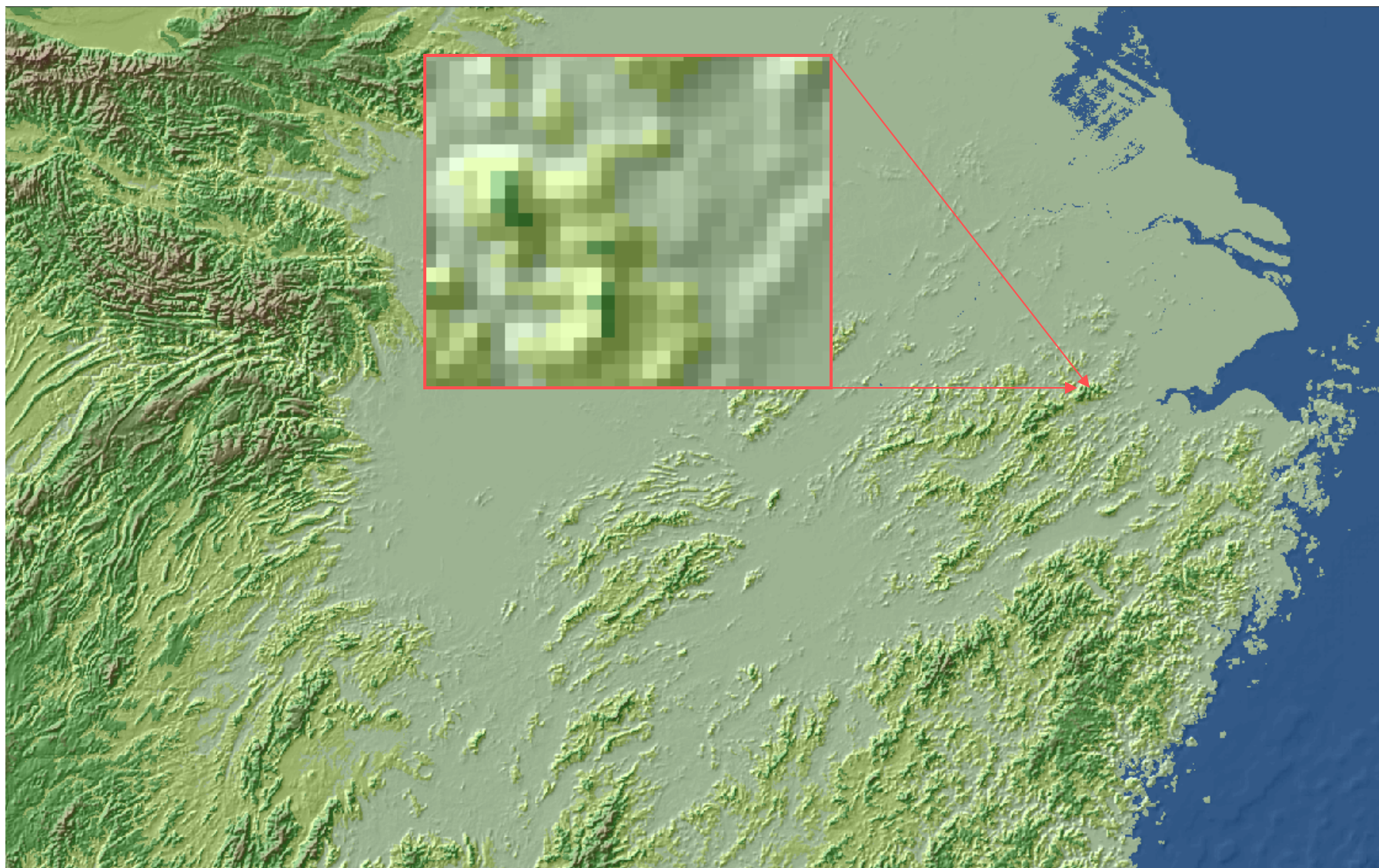
spatial  
objects



# 栅格数据: 数字高程模型 (DEM) or “网格数据”



# 栅格数据: 每个像素都有一个值



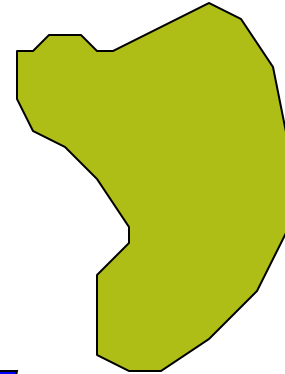
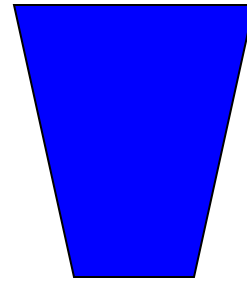
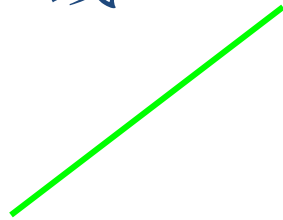
# 矢量数据

数字点、线和多边形, 用  $x$ 、 $y$  (和  $z$ ) 坐标绘制

点

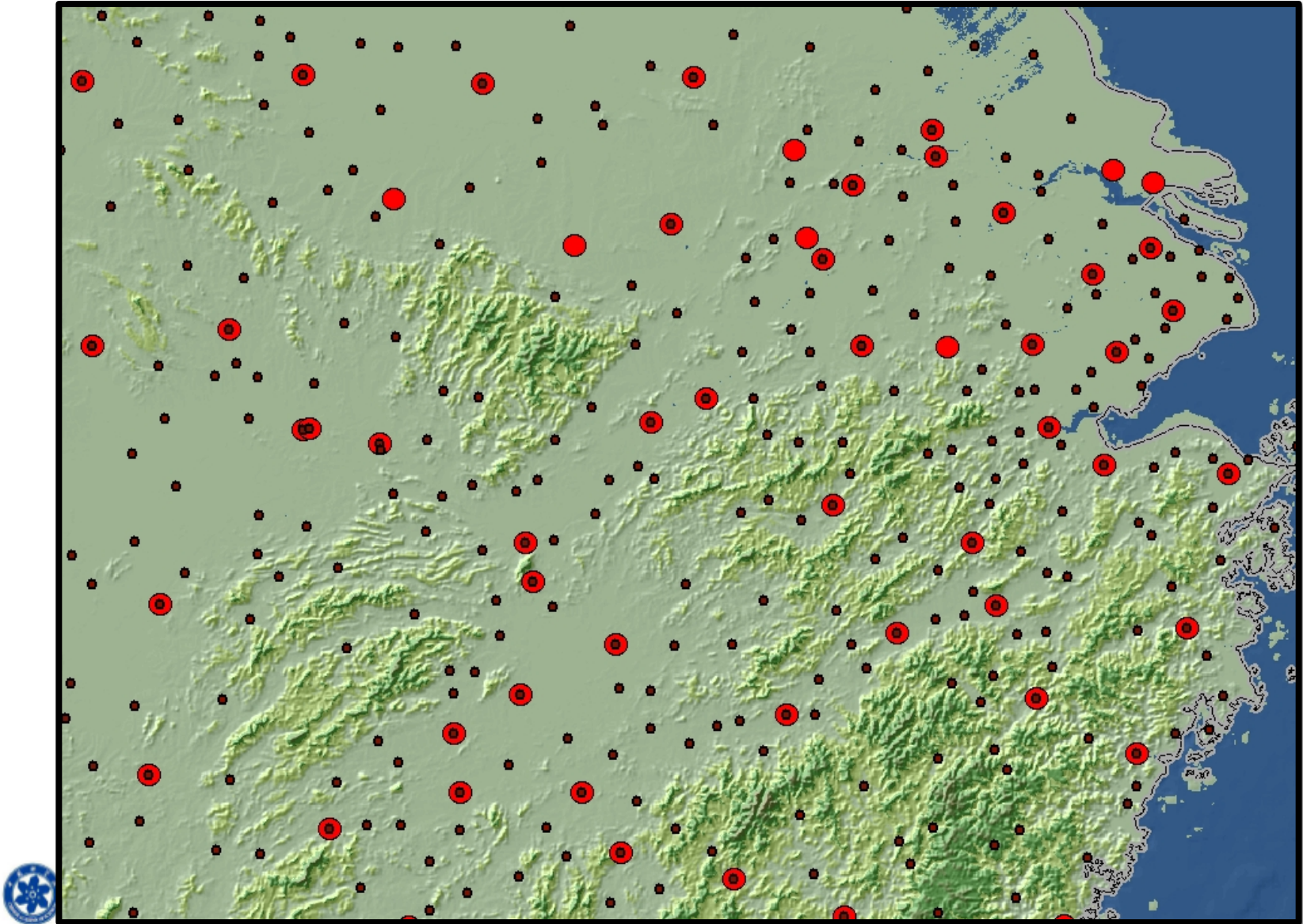


线



多边形

# 栅格和向量叠加, 州和县位置(1820)

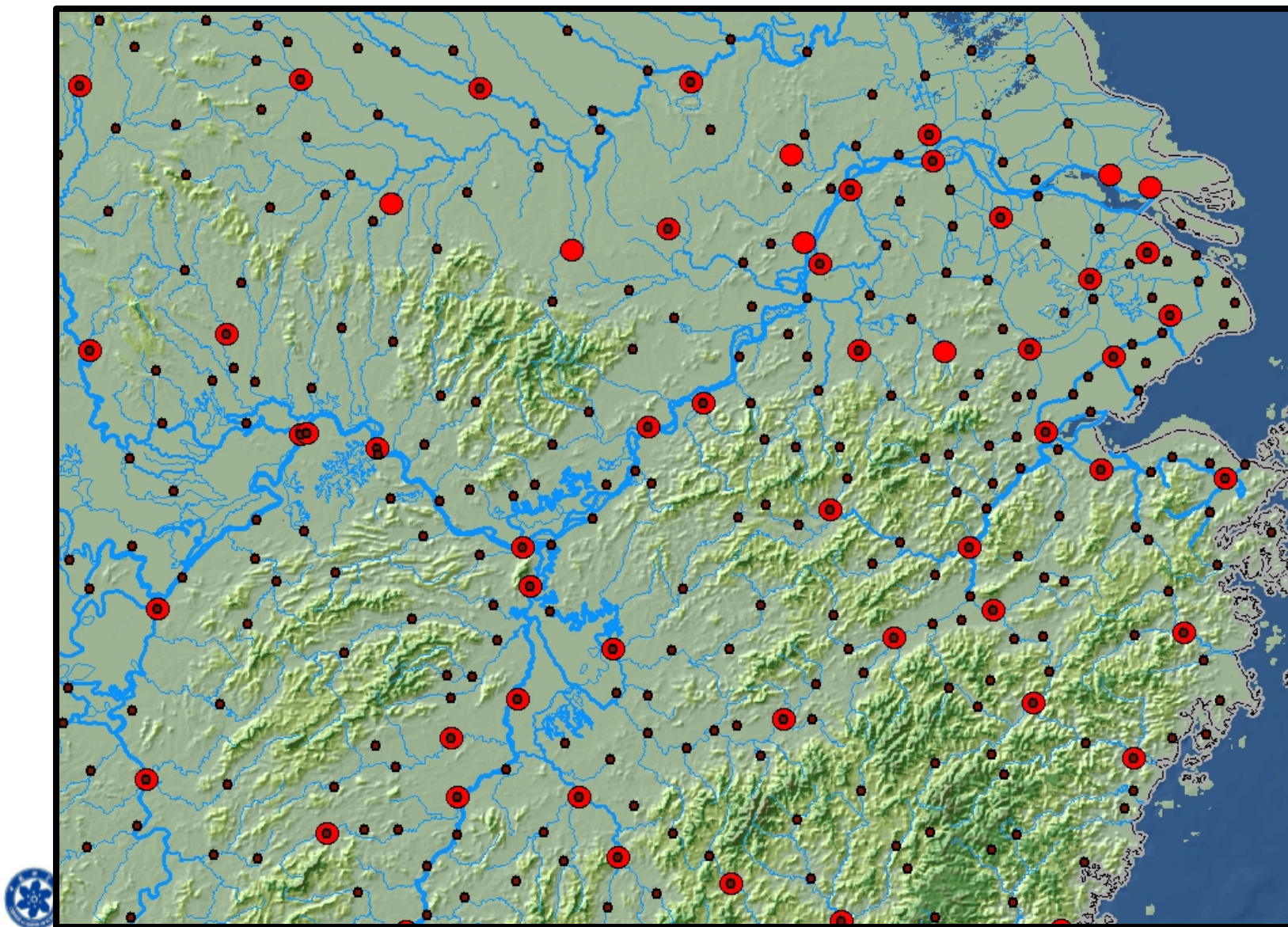


# 栅格和矢量叠加

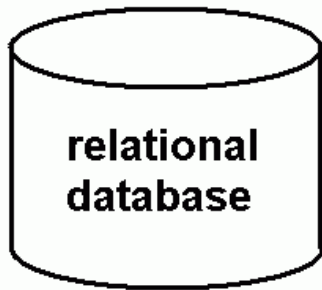
## 水道 (1820) 和 数字高程模型 (DEM)



# 州、县、水路和数据高程模型 (1820)

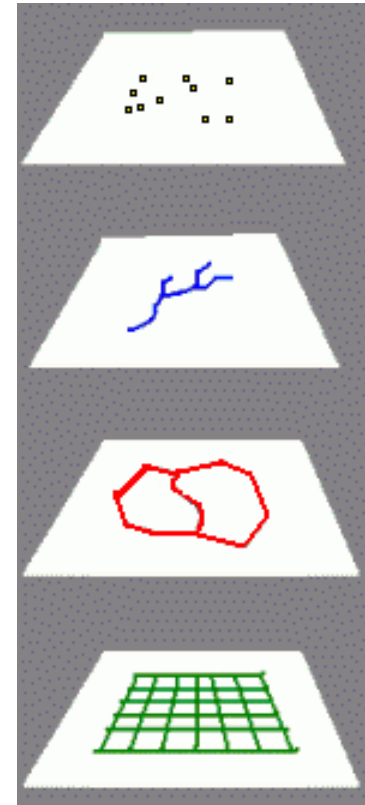


## 数字地名录



placename  
feature type  
coordinates  
valid date  
source note

## GIS 图层



spatial objects

CHGIS = gazetteer database + GIS layers

# database

linked to

# spatial data

PINYIN_NAM	FANTI_NA	PROV	SYS_ID
Xianyou Xian	仙游縣	福建	182
Putian Xian	莆田縣	福建	183
Anxi Xian	安溪縣	福建	184
Dehua Xian	德化縣	福建	185
Hui'an Xian	惠安縣	福建	186
Nan'an Xian	南安縣	福建	195
Tong'an Xian	同安縣	福建	196



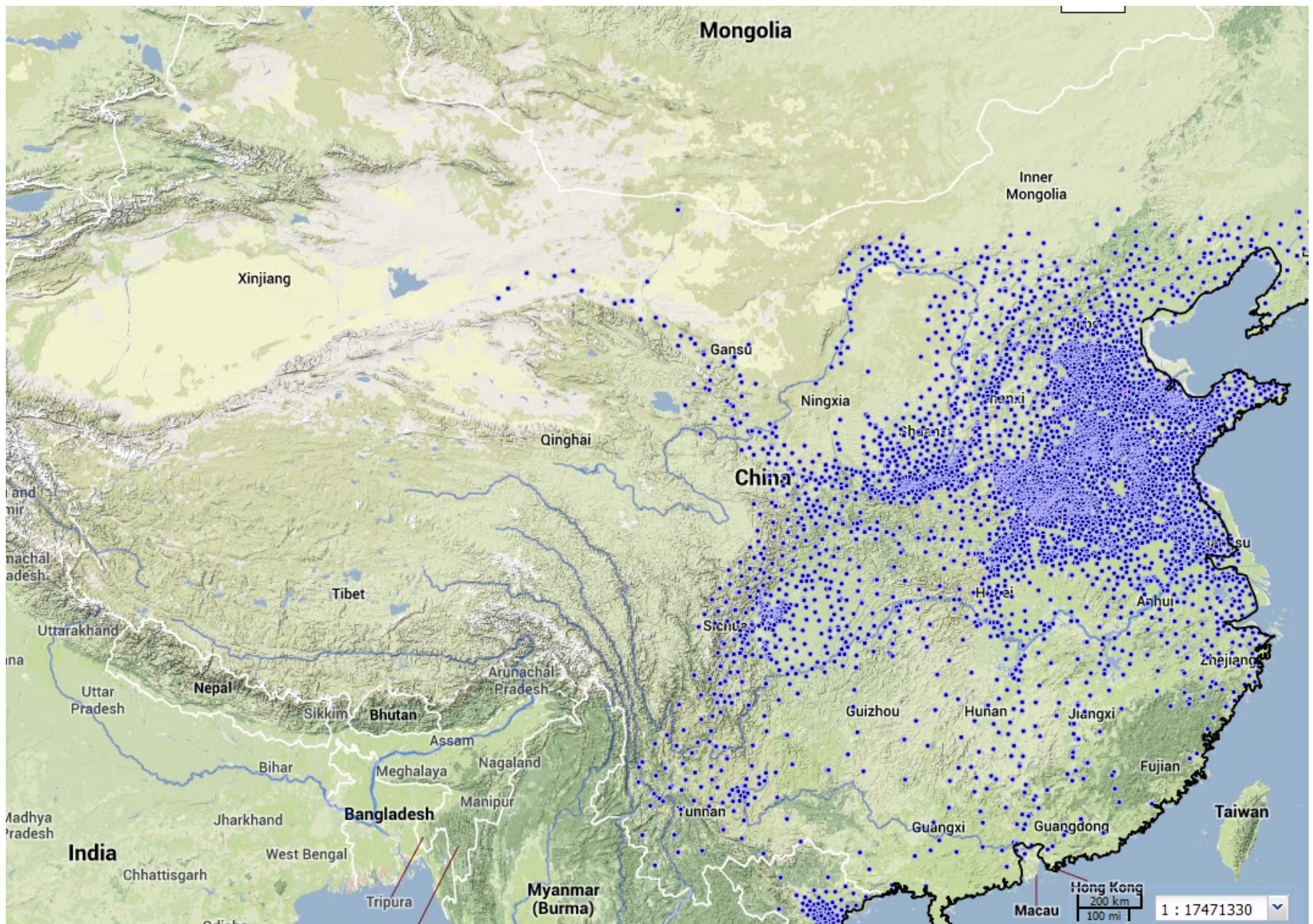
points

PINYIN_NA	HZ_FANTI	SYS_ID
Lai Xi	瀨溪	13462
Tao Xi	桃溪	14641
Lan Xi	藍溪	15043
Jin Jiang	晉江	15144



lines

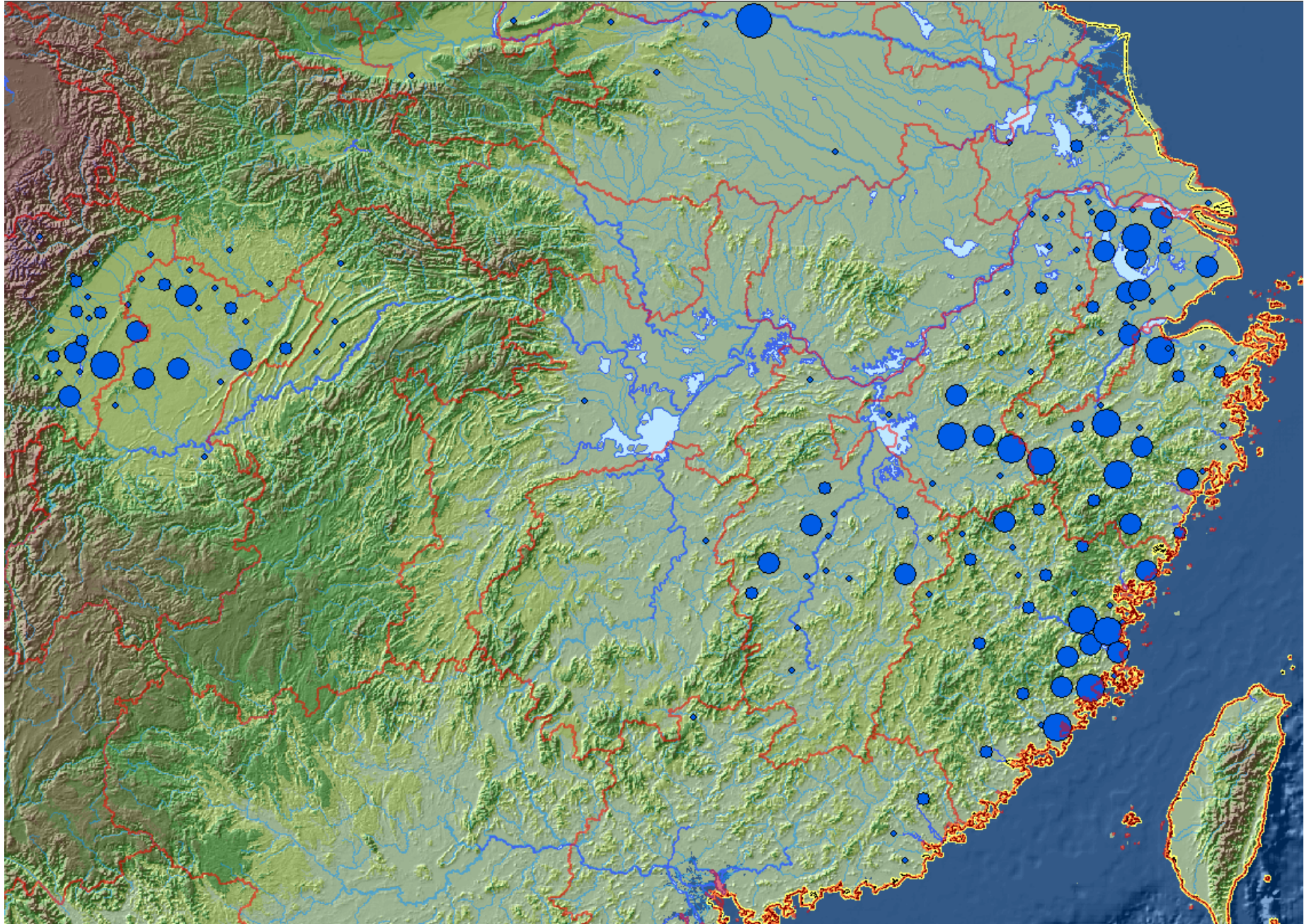


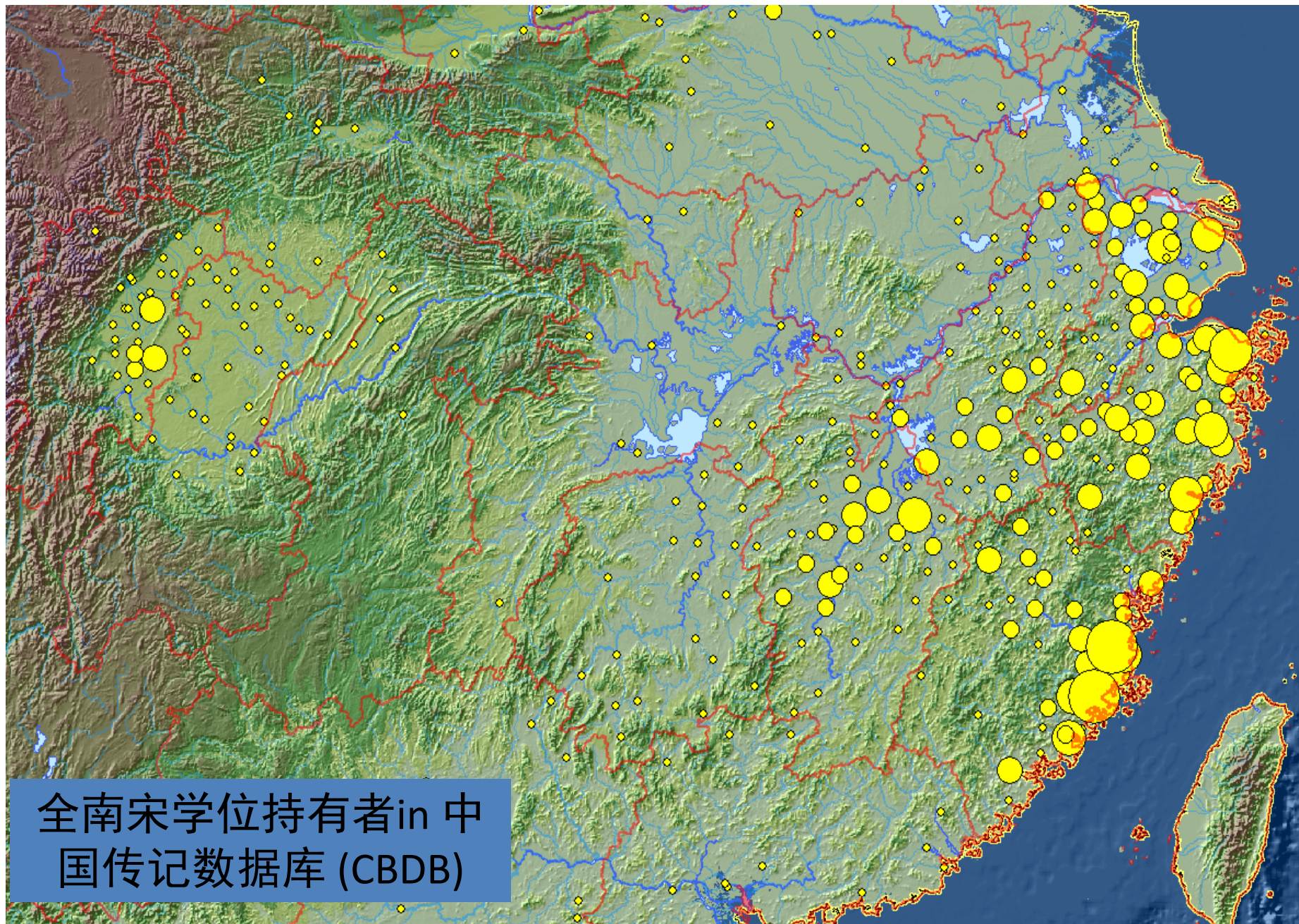



 公元2年登记的人口

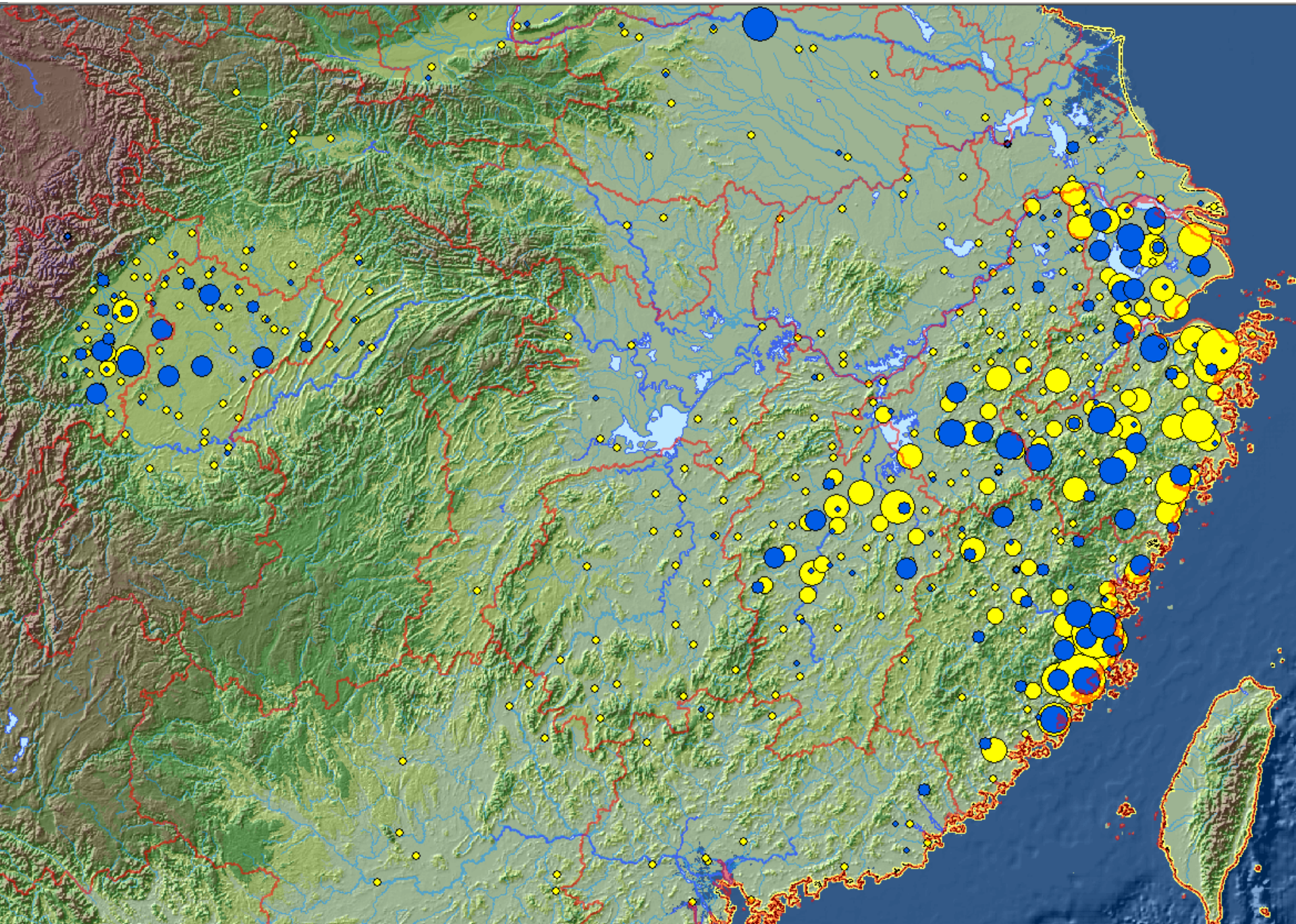


# 公元1148公务员考试考试地址名录

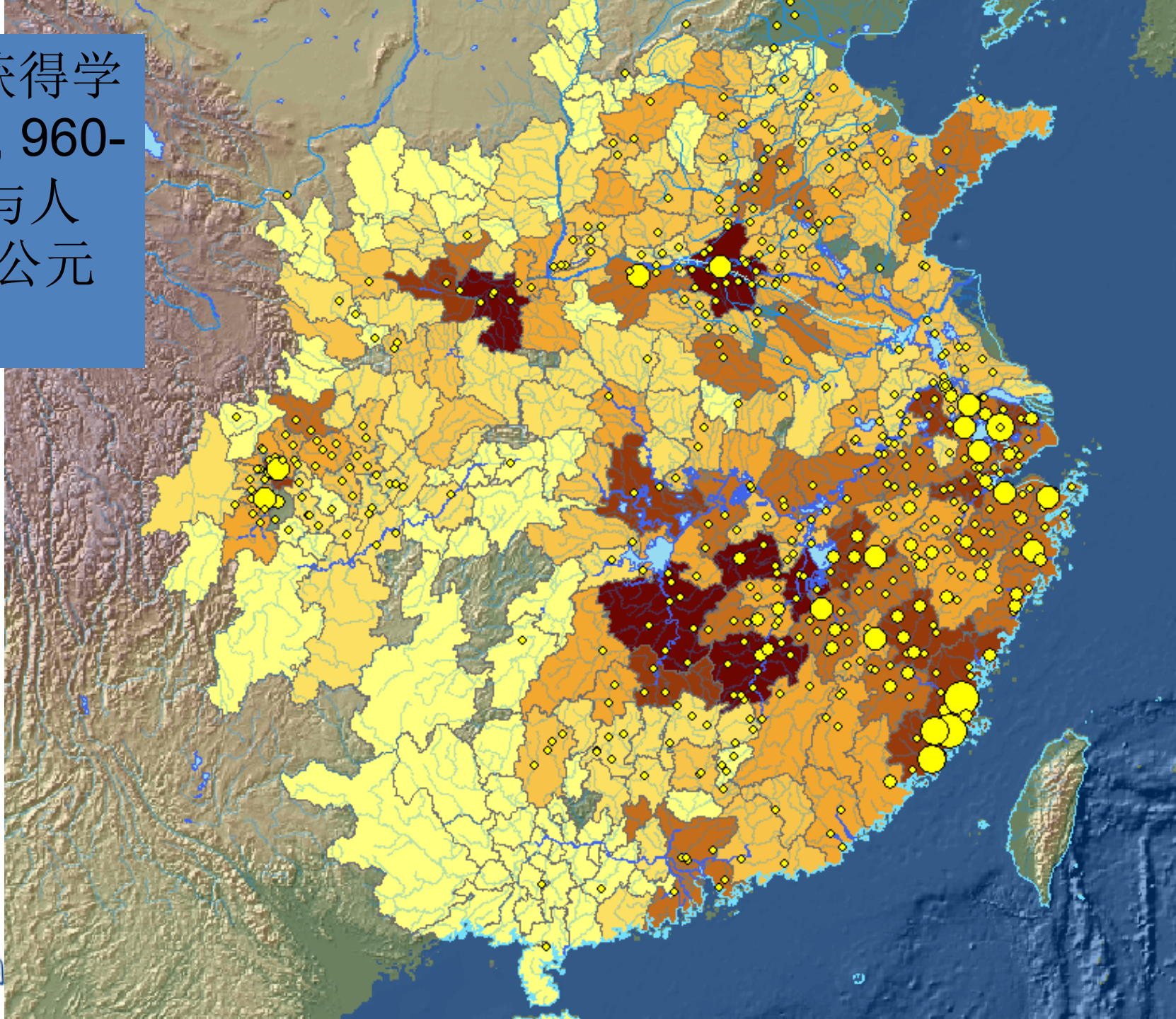




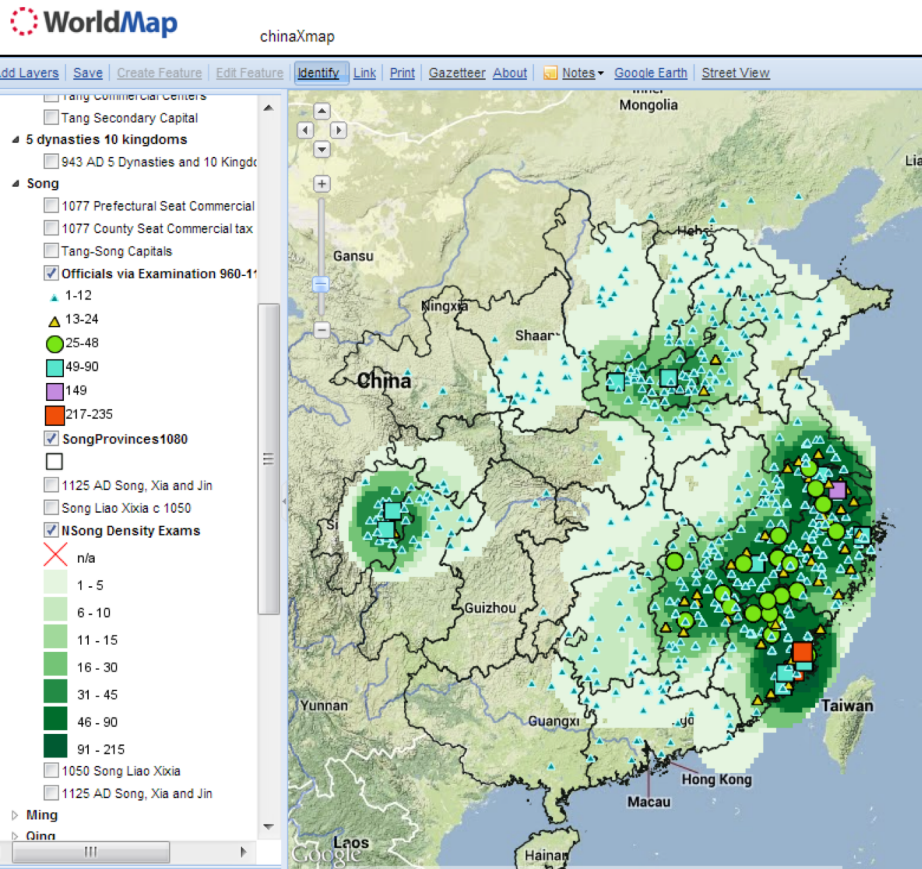
全南宋学位持有者in 中国传记数据库 (CBDB)



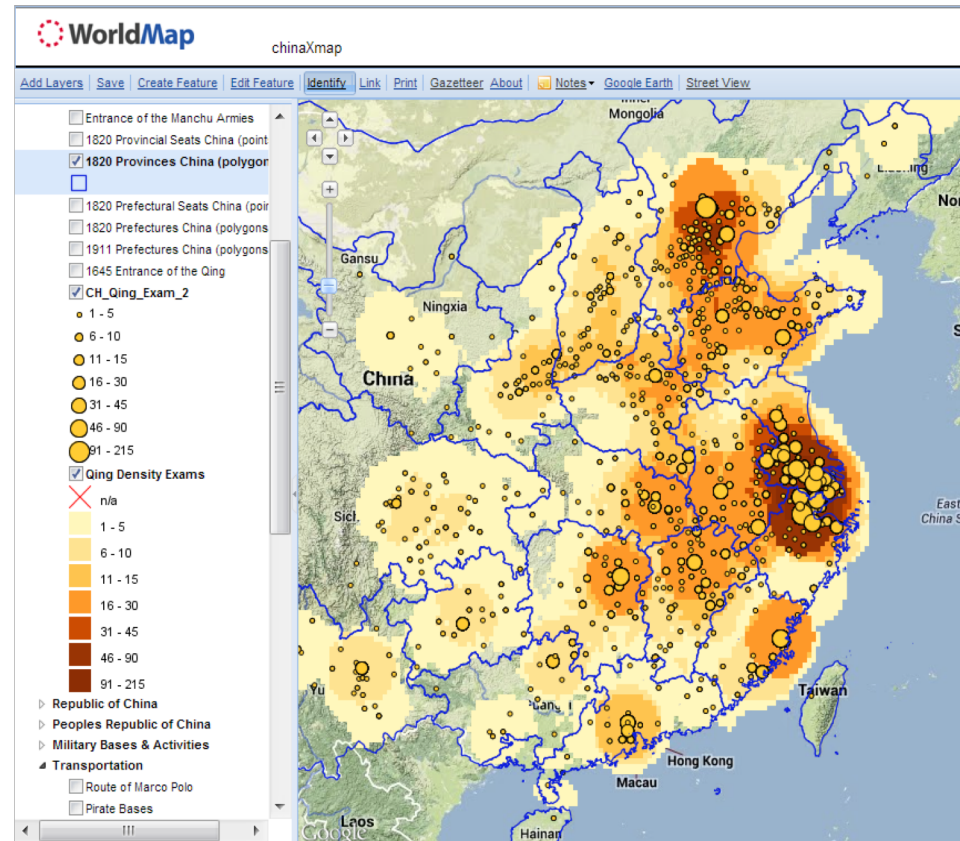
4730 获得学位的人, 960-1126, 与人口分布公元 1080



# 北宋的分布(960-1126) 进士 in CBDB



# 清代的分布(1644-1905)进士 in CBDB





19<sup>th</sup> c. “Busy” counties , Ming-Qing Postal Routes, 1990 Railroads





# China Historical GIS



Gazetteer Search Engine

List of Free Datasets

XML Web Service

Skinner Map Collection

中文版

home  
intro  
data  
members  
meetings  
docs

search tools

maps

with funding from the Henry Luce Foundation, the National Endowment for the Humanities

and support from

<http://www.fas.harvard.edu/~chgis>

Fairbank Center for Chinese Studies  
Harvard Asia Center  
Harvard Yenping Library  
Harvard Yenping Institute

© 2001-2010 - Harvard University and Fudan University

CHGIS  
Center for Geographic Analysis  
1737 Cambridge St, K021  
Cambridge, MA 02138  
tel: 617-495-2451  
new office location

China GIS



Japan GIS



Dynamic Maps



Map Scans



Featured EdSite



Sichuan Earthquake



# 从《中国历代人物传记》中发现新知识

The China Biographical Database

– from anecdote to data to knowledge

文本实体抽取

人物关系数据库

人物关系学 (Prosopography)

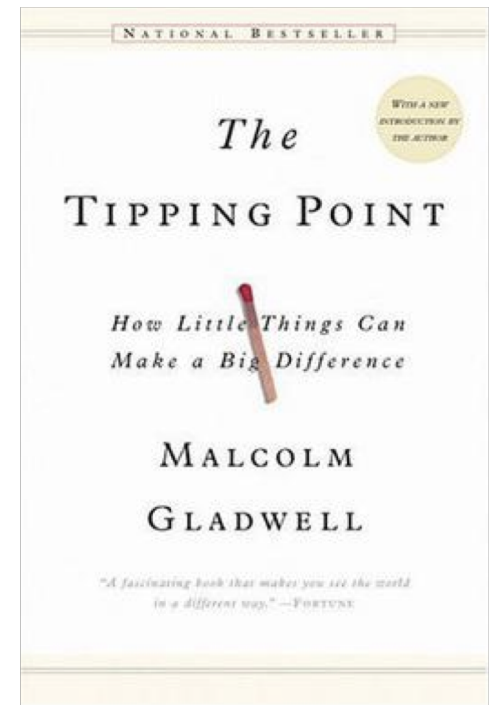
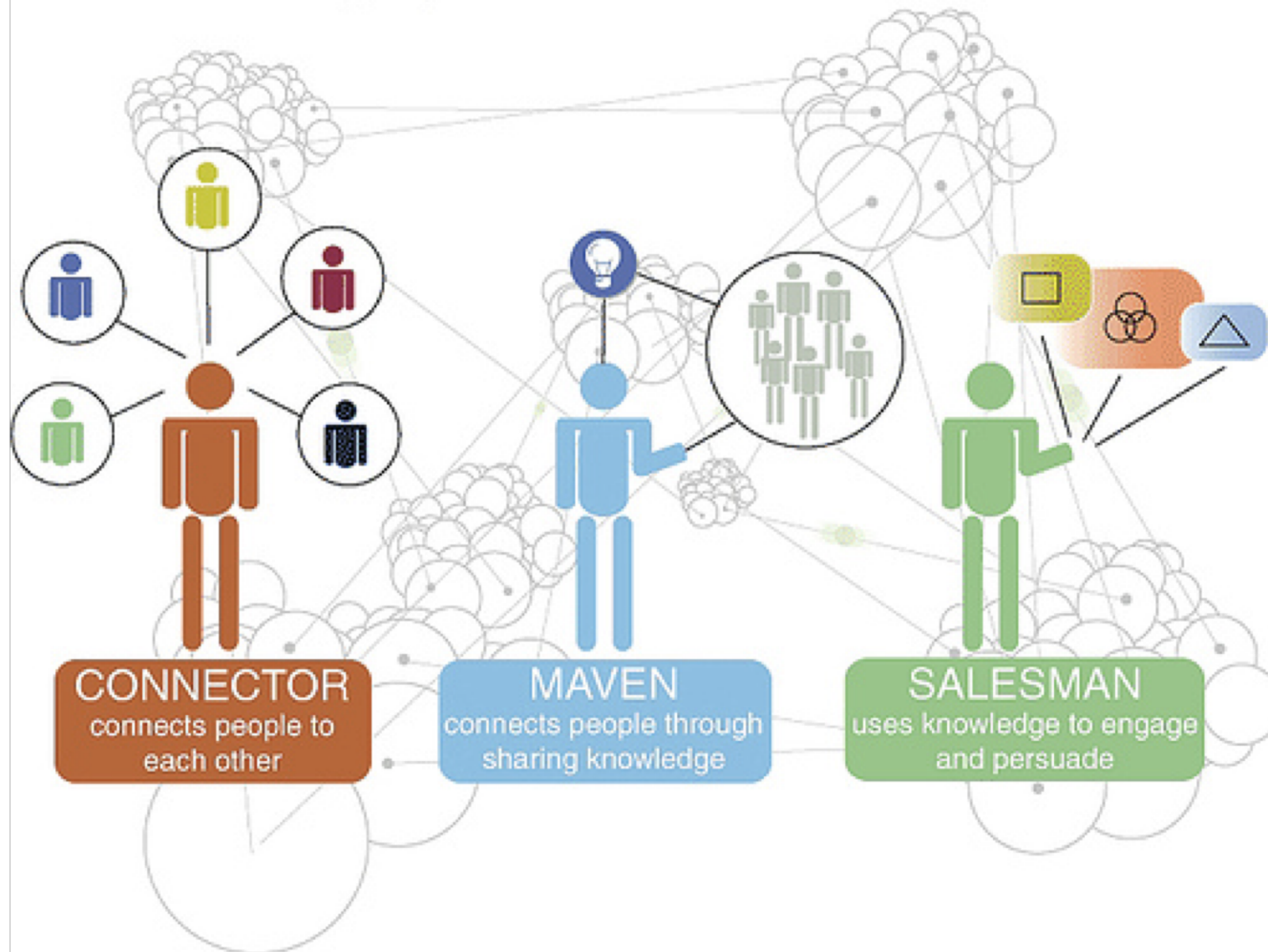
时空数据分析 (Spatial analysis)

社会网络分析 (Social network Analysis)

# 社交网络

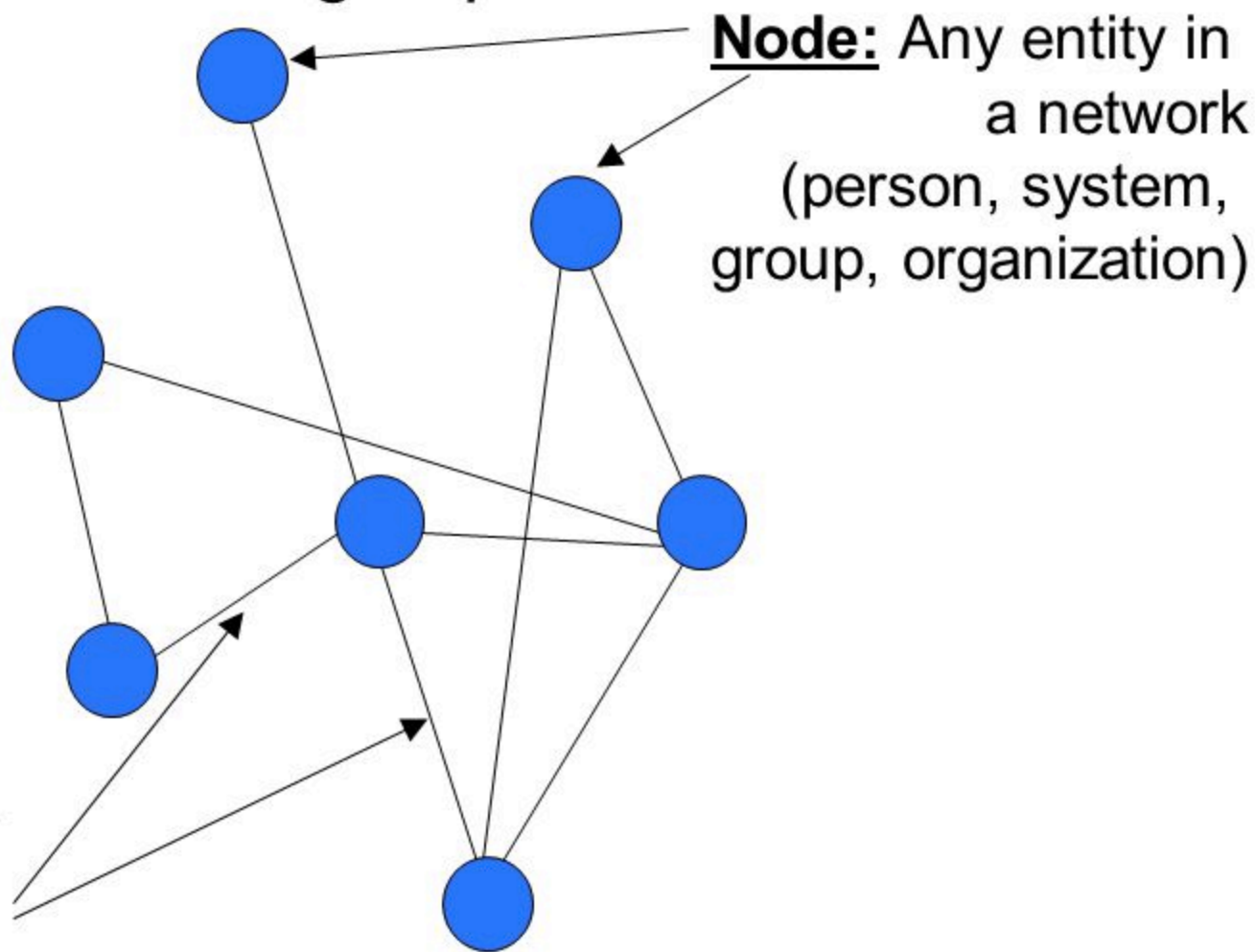
“社交网络分析是对人员、群组、机构、计算机、信息或知识等实体之间的关系和流向进行建模、度量和计算推理，以便找到个体行为如何影响群体看法或行动，以及群体如何影响个体的决策等原因。  
(valdis krebs, 2002年)。

The Tipping Point + The Social Media Network



# Social Network Analysis:

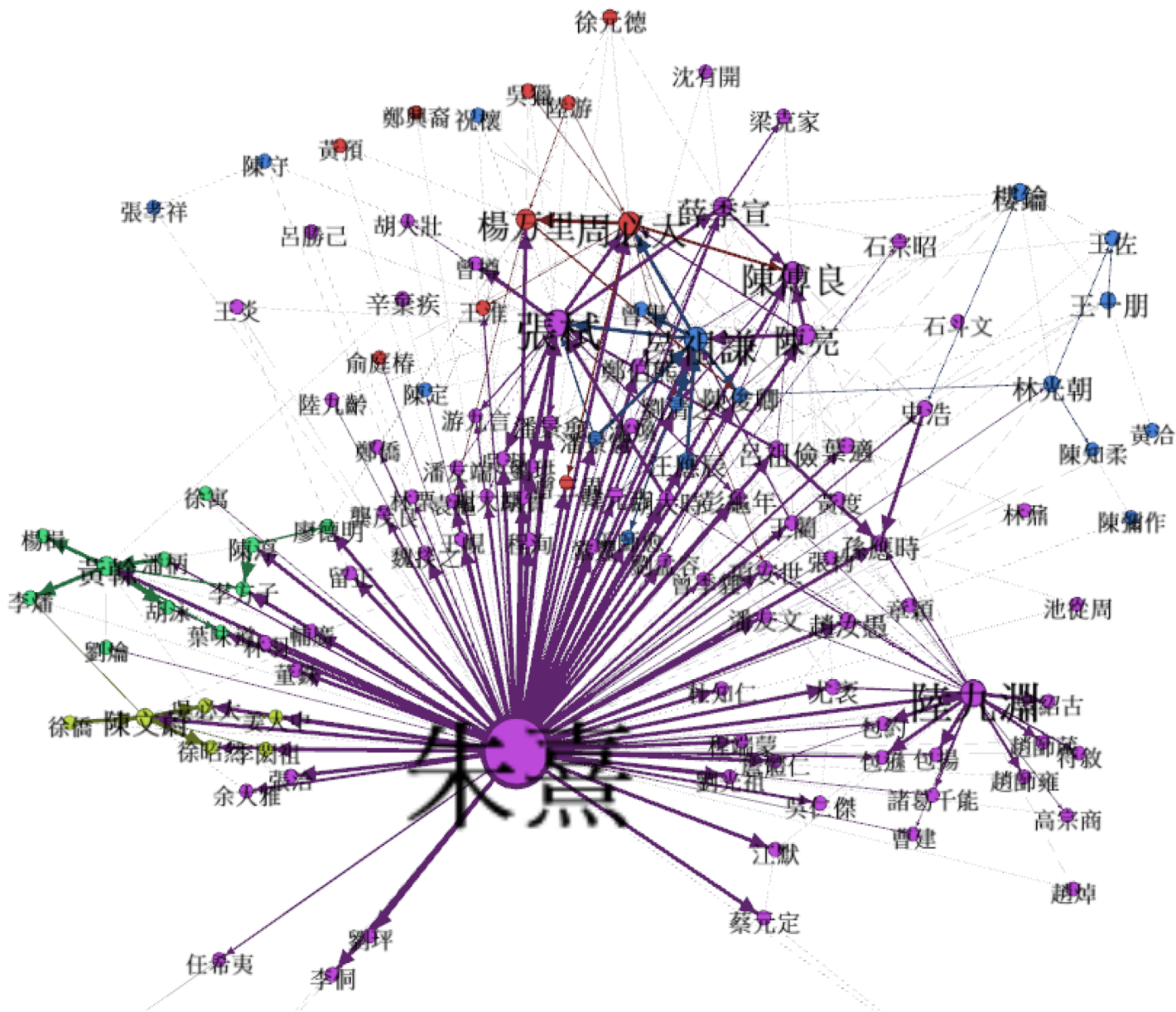
Focus on *interactions* between individuals/  
groups





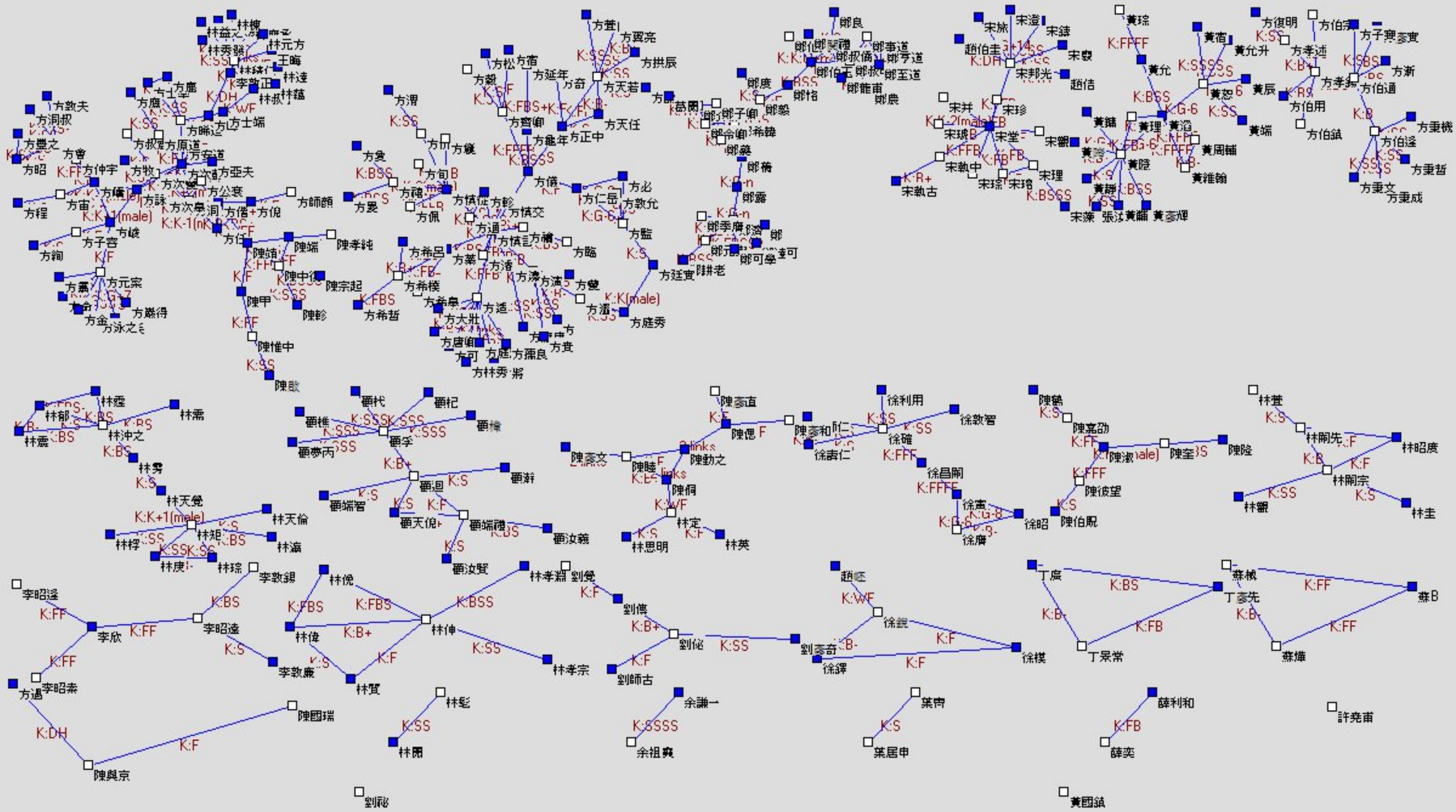


朱熹的信网. 显示至少两个字母的收件人, 并在它们之间进行交换。标签大小反映交换信件的数量; 颜色标识这些交换之间的子网

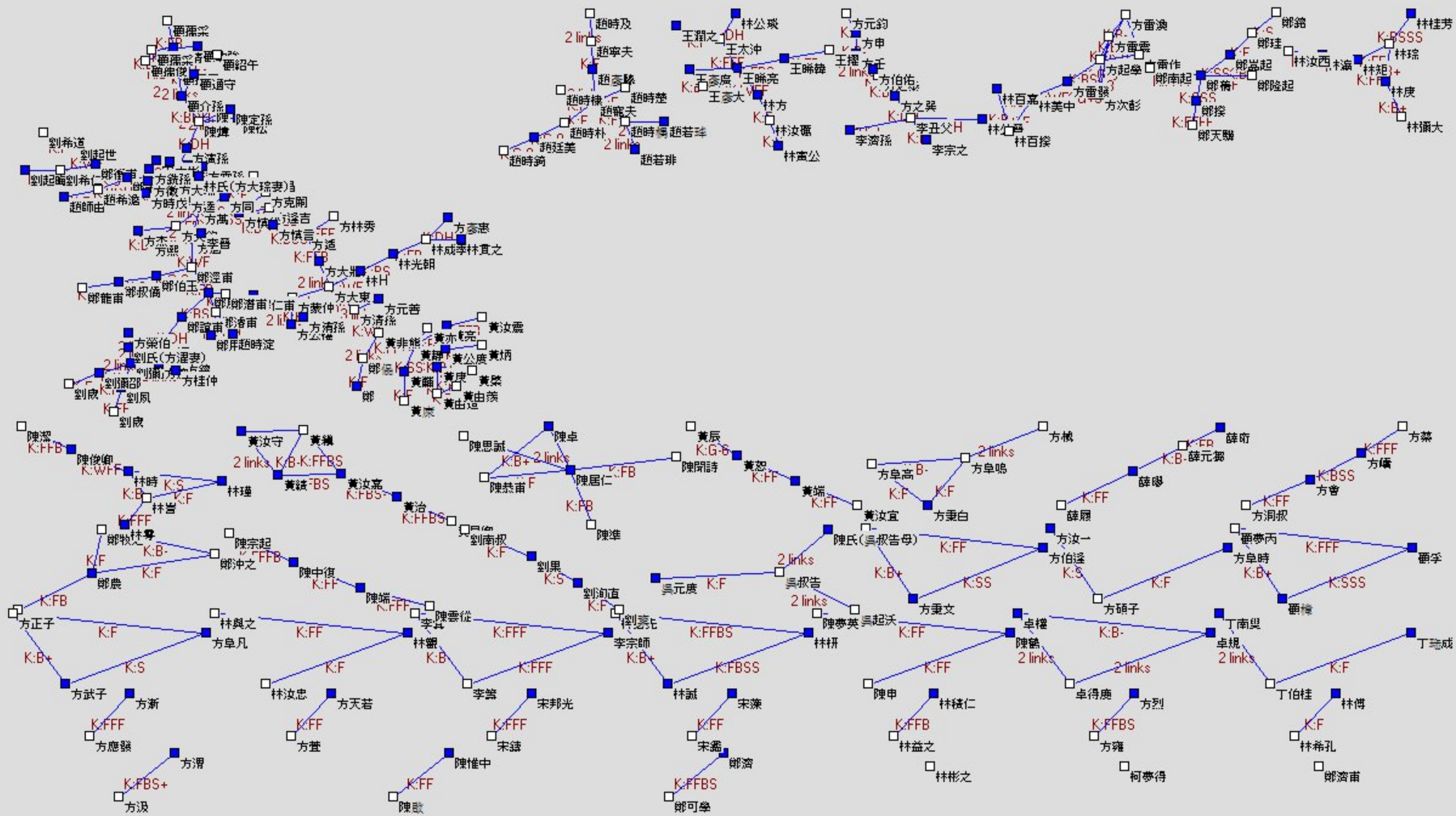






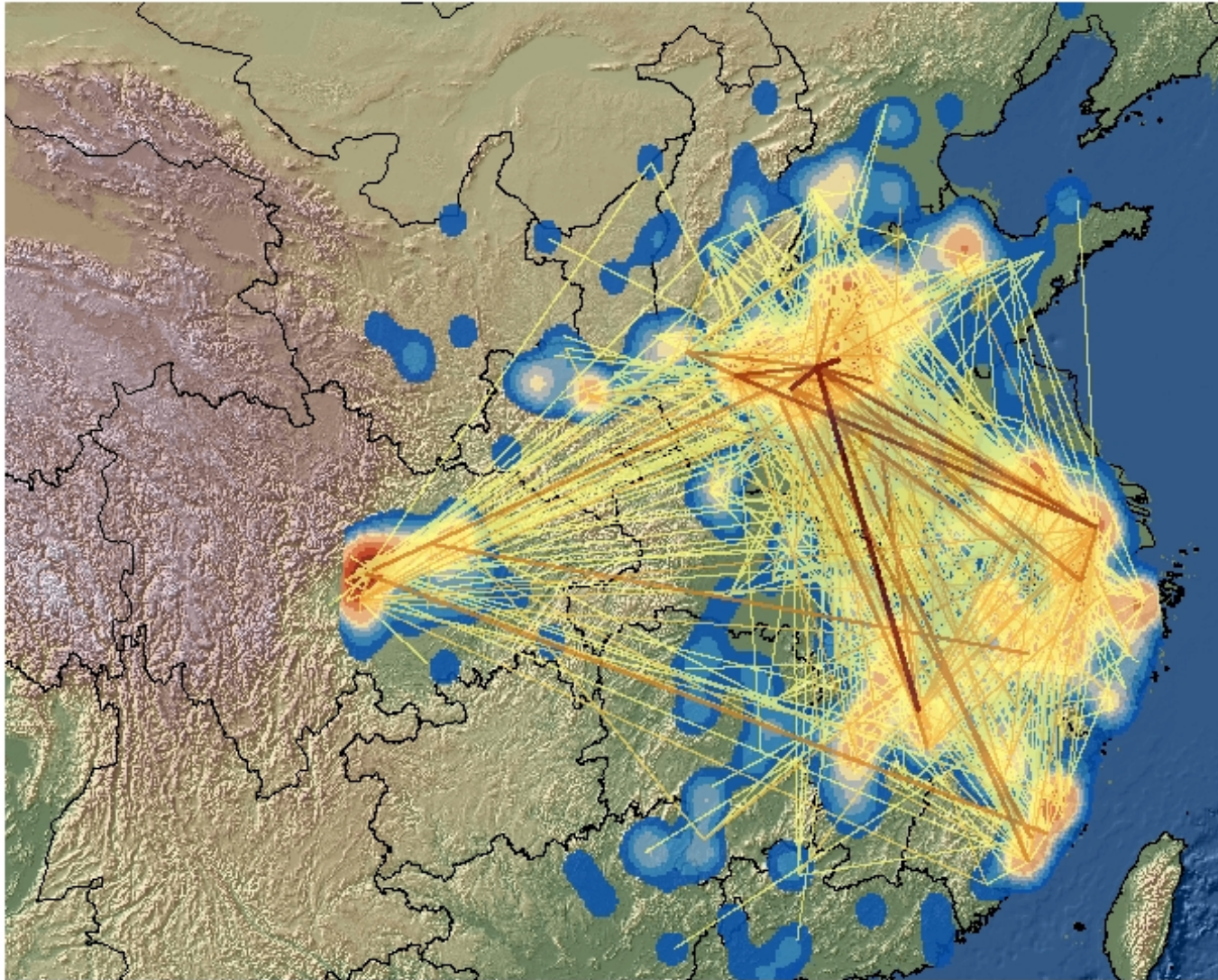


福建莆田直系亲属关系网络的构成, 对于获得了进士的人物之间的关系。 公元1050至 1100



福建莆田直系亲属关系网络的构成, 对于获得了进士的人物之间的关系。 公元1200至1250

# 中国传记数据库的文字交流空间密度, 在1060-1080 期间



2359  
persons out  
of 120,000  
currently in  
CBDB

Legend: Population Density



# 中國歷代人物傳記資料庫

中央研究院歷史語言研究所、哈佛大學、北京大學中國古代史研究中心合作開發  
Developed through collaboration among Academia Sinica, Harvard University, and Peking University



## 公开获取方式

# CHINA BIOGRAPHICAL DATABASE PROJECT (CBDB)

## 中國歷代人物傳記資料庫

Enter Biographical Data 輸入傳記資料

### Retrieve Data 數據查詢

Look Up Data on an Individual 按人查詢

Query by Methods of Entry into Government 按入仕途徑查詢

Query Office Holding 官職查詢

Query Kinship 查詢親屬關係

Query Associations 查詢社會關係

Query Social Networks 查詢社會關係網絡

Query Pair-wise Associations 查詢兩人之間社會關係

Exit 退出

Desktop Version

Last update: June 28 2011

單機版 2011年6月28日最後更新

# Historical Database Project (CBDB)

哈佛大学、北京大学中国古代史研究中心合作开发  
Developed in cooperation with Academia Sinica, Harvard University, and Peking University

[Search by Name](#) [Search by Address](#) [Search by Office](#) [Keyword Search](#) [Advanced Search](#) [Kinship and Social Associations](#) [Export Search Results](#) [Social Network Analysis](#) [User Guide](#) [Explanation of Terms](#)

## Search by Name

No Search Results.

Search



## Search by Name

### Individual Data

Name

Index Year  至

Dynastic Period

Address

Select

Search by Name allows you to identify an individual or retrieve a list of individuals that share basic biographical attributes. By default, the query includes everyone in the database, and the list is refined according to the manipulation of parameter values on the left column.

Read about "faceted search" in the CBDB Online Interface User Guide

### Examples

The most common usage of this interface is to isolate and identify a single individual by keyword search of name. If you search for **Wang Anshi** in pinyin, that query should return a single record (as of February 2011) corresponding to Wang Anshi 王安石 in the eleventh century.

If the individual is only known by a part of a given name, adjusting the temporal range helps narrow the list of search results. For example, 維 occurs commonly as part of a name (463 records as of February 2011), but delimiting the query to the Tang dynasty should narrow the search results to 5 records. Navigate to **Dynastic Period** and specify **Tang** from the dropdown list. Wang Wei 王维, whom we had in mind, appears in record #5.

Another possible use is to retrieve a list of individuals who share a surname. A search for the surname 獨孤 returns 54 records. This list can be refined so that it displayed those active only during the Northern Wei (5 records) or only during the Tang (32 records).

Click on the name of any person and the data in their CBDB file will be displayed.

Note: use "v" for "u"

### Technical description of the search parameters

#### Name

Retrieves records of individual(s) by keyword lookup of their name. The entry may be surname and given name, any of the alternate names, whole or partial, in Chinese or in pinyin.

#### Index Year

The year in which the person was (presumed to be) in his/her sixtieth year or the year of death if the subject died before the sixtieth year. By specifying a year range the results will be filtered for people whose index years are within the range. Index years are sometimes missing. They will be entered as new data justifies.

#### Dynastic Period

Automatically fills the range of index years by the dates of a major Chinese dynasty.

#### Address

Delimits the search results by the address associated with the individual(s). Click on **Select** to specify the location of the address.

Search by Address

Search Results 1067 Records

Search

ADDRESS

Address 0012784 務州 Wu Zhou

Select Clear Delete

Type of Address

Index Years

Dynastic Period



Basic Biographical Search Matches 1067 Records

1067 records, page 1 of 54, record 1 to 20, 20 records per page

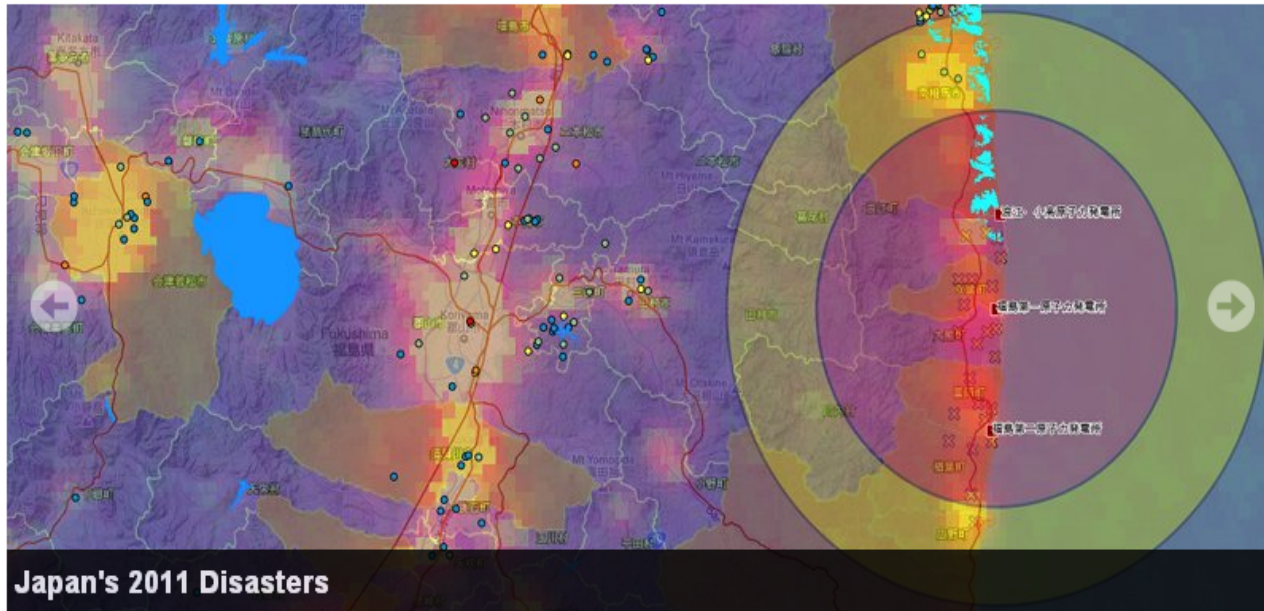
1 2 3 4 5 6 7 8 9 10

#	Name	Dynasty	Choronym	Alternate Names	外文名	Index Year
1	Qian Yu 錢通	宋Song	unknown宋詳			1109
2	Zhang Su 張肅	宋Song	unknown宋詳			1057
3	Zhang Fu 章服	宋Song	Yuzhang 豫章			1165
4	Zhao Buyou 趙不猷	宋Song	Song Taizong 宋太宗			1161
5	Jiang Yan 江衍	宋Song	unknown宋詳			1091
6	Fan E 范縉	宋Song	Gaoping 高平			1095
7	Hu Ze 胡則	宋Song	unknown宋詳			1022
8	Lv Bengzhong 呂弼中	宋Song	Dongping 東平			1139
9	Lv Yongzhong 呂用中	宋Song	Dongping 東平			1140
10	Mei Zhili 梅執禮	宋Song	unknown宋詳			1127
11	Pan Jinggui 潘景珪	宋Song	unknown宋詳			1192
12	Pan Lianggu 潘良貴	宋Song	unknown宋詳			1142
13	Su Zhou 蘇籀	宋Song	Zhaojun 越郡			1150
14	Su Ce 蘇策	宋Song	Zhaojun 越郡			1150
15	Tang Yaofeng 唐堯封	宋Song	unknown宋詳			1162
16	Wang Shixin 王師心	宋Song	Taiyuan 太原			1156
17	Ying Shunchen 應舜臣	宋Song	unknown宋詳			1076
18	Chen Liang 陳亮	宋Song	Wuxing 吳興			1194
19	Chen Shu 陳樞	宋Song	unknown宋詳			1151
20	Zhang Zhu 章著	宋Song	Yuzhang 豫章			1155

Create a Map

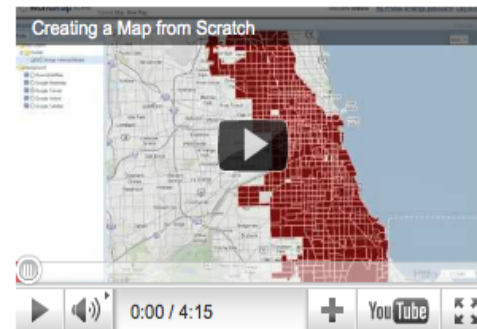
View a Map

About



Japan's 2011 Disasters

Build your own mapping portal and publish it to the world or to just a few collaborators. WorldMap is open source software.



[Watch the WorldMap Quick Start video](#)



<http://worldmap.harvard.edu/>