

南京大學
NANJING UNIVERSITY

量化历史数据库的构建与研究

南京大学

历史学院暨数字人文研究中心

梁晨

2020年9月

2022-1-21



第一讲

必要：微观历史数据库 的兴起与量化研究

■ 课程安排：

第一讲 必要：微观历史数据库的兴起

第二讲 可能：量化数据库与中国史研究

第三讲 如何：量化历史数据库的构建

第四讲 要点：从结构数据到追踪数据

第五讲 示例：中国现代知识阶层起源研究



史学界对数字人文的关注与热点



- 其一，大量历史文献数据库和专题数据库的建立与网络连接，加强了个人史料搜集和整理的能力，促进历史叙事的多样化；
- 其二，利用系统性史料构建适用于统计分析的量化数据库，通过对大数据的系统分析，“挖掘新事实、产生新认识”；
- 其三，反思数据库等技术在史学研究中的作用与影响，例如e-考据与传统考据并没有本质上的差别，仍需以深厚的史学功底为基础，强调传统史学训练的重要性。



史料数据化



- 1980年代的检索文献数据库→1990年代的学术出版物数据库→2000年代的量化历史数据库，互联网时代历史研究的材料组织和利用形式在不断变革。
- 这些变革使得史料拥有权的“唯一性”对史料获取的障碍大大降低，给研究带来了无限新可能。新的研究材料更适合运用全面、系统、精确的分析方法，更便于开展有说服力的比较研究。



1980年代可检索文献数据库



- 台湾中央研究院汉籍电子数据库（启动于1984年，包含二十五史、十三经等文献）
- Text Encoding Initiative 1987
- IRIS Intermedia project 1980s
- 80年代以后有越来越多的该类型数据库建设。如1993年开始的Rossetti Archive和美国国会图书馆、哈佛燕京图书馆、普林斯顿大学葛斯德图书馆以及台湾大学傅斯年图书馆“中文善本数位典藏计划”。
- 数据库之间还在相互连接。如1991年开始，目前由全世界30多家图书馆参加的中华古籍善本国际联合书目系统。



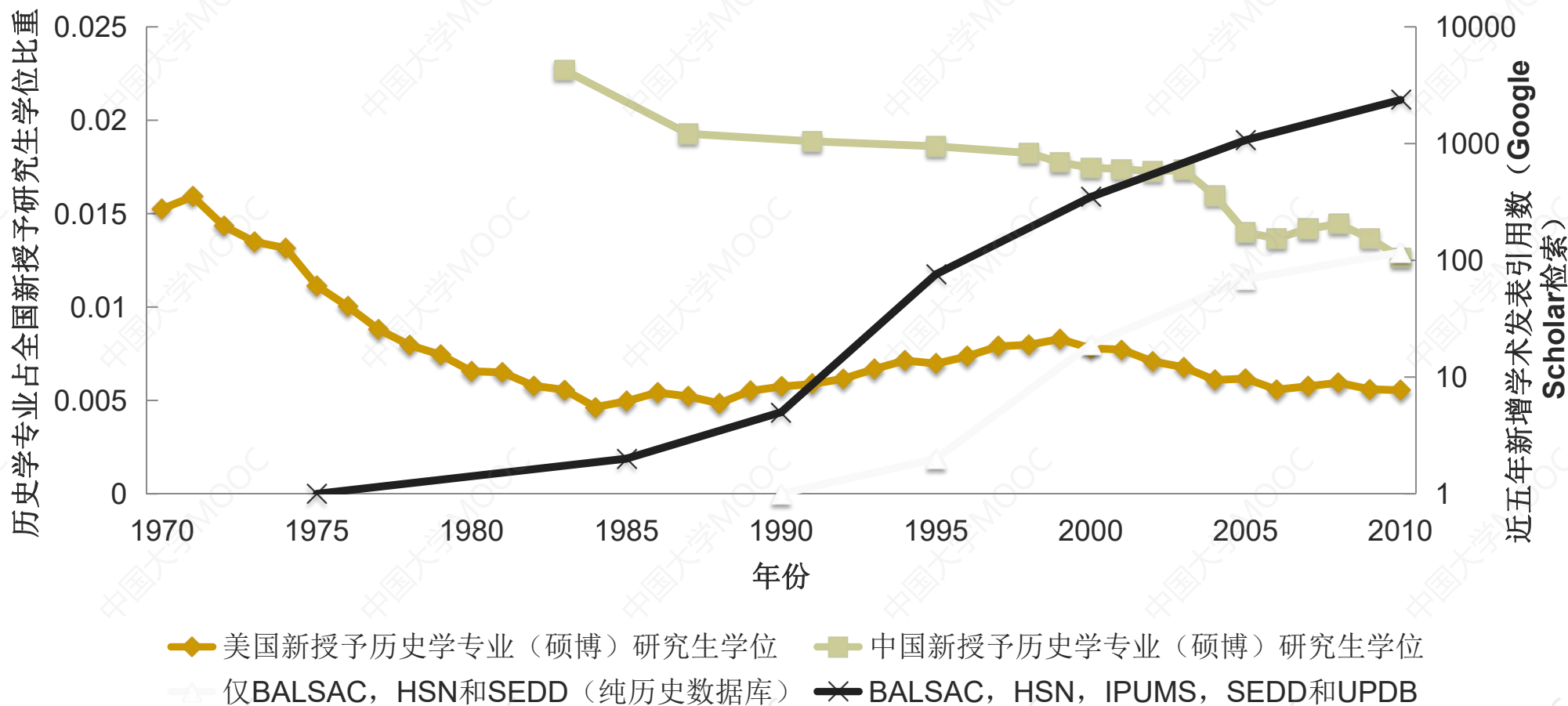
1990年代学术出版物数据库



- JSTOR（西文人文社会科学过刊数据库，1995年）
- 中国知网（集期刊杂志、博士论文、硕士论文、会议论文、报纸、工具书、年鉴、专利、标准、国学、海外文献资源为一体的网络出版平台，1999年始建）



2000年代大规模量化数据库





西方主要历史量化数据库



- Integrated Public Use Microdata Series (IPUMS) 整合公共围观数据系列
 - 750 million individual records, international, 1850-present
7亿5千万人，国际，1850年至今
 - Publicly available since 1993
1993年起公开
- Utah Population Database (UPDB) 犹他人口数据库
 - 7.4 million individuals in Utah, 1812-1992
7千4百万人，美国犹他州，1812-1992
- BALSAC Population Database 巴尔扎克人口数据库
 - 5 million individuals in Quebec, 1621-1992
5百万人，加拿大魁北克省，1621-1992
- Scanian Economic Demographic Database (SEDD) 斯堪尼亚经济人口数据库
 - 104000 individuals in Southern Sweden, 1813-2011
10.4万人，瑞典南部，1813-2001
 - Publicly available since 2013
2013年起公开
- Historical Sample of the Netherlands (HSN) 荷兰历史人口样本
 - 78000 individuals in the Netherlands, 1812-1992
7万8千人，荷兰，1812-1992
 - Publicly available since 1998
1998年起公开



量化数据库与历史研究



- 近年来，数字人文研究的发展，尤其是大规模历史量化数据库在收集、整理和构建方面的重大进展开始在这两种认识论分歧间架起桥梁。
- 中国研究中的这些大规模研究计划已经开始形成一种新的“求是型”学术。这不仅会重新定义中国近、现代以来的历史，而且会重新定义世界历史和社会理论。
- 实际上，这种研究范式在西方学界早已存在。但引人注目的是，大规模历史数据库对西方学界的重要影响是伴随着历史学科在全球范围内影响衰退同时出现的。



- 法国经济学家托马斯·皮盖蒂（Thomas Piketty）所著《二十一世纪资本论（Capital in the Twenty-First Century）》一书是依据多国二十世纪国民账户、收入、财产与税收等多种系统历史数据来研究资本主义社会中不平等的长期演化。该书曾高居各种英文畅销书排行榜首并引发热烈的学术讨论。
- 美国经济学协会主席克劳蒂亚·高丁（Claudia Goldin）与《经济学季刊（Quarterly Journal of Economics）》主编、前美国劳工部首席经济学家劳伦斯·凯兹（Lawrence Katz）合著的《教育与技术的竞赛（The Race between Education and Technology）》基于近一个世纪以来美国教育、职业和收入的个人层面微观数据讨论美国经济不平等的历史脉络和技术革新、教育进步对收入分配结构的影响。
- 这些成就在推动全球学术进步的同时，也彰显了大规模量化数据库在结合历史资料系统研究人类社会长期变化与延续等大问题上的作用。



“数字人文”研究需要构建的是量化数据库，而非文献数据库



- 真正的“数字人文”研究，其构建的数据材料应该能直接被电脑分析，特别是能被统计与运算，这也是“数字人文”在历史研究中最重要使用。
- 所谓量化数据库是指各种能够涵盖一定地域范围的、具有一定时间跨度的整体性大规模个人或微观层面信息的系统（一手）资料。这些资料按照一定的数据格式进行电子化，构建成为适用于统计分析软件的数据库并进行量化研究。
- 量化数据库研究多以“大数据”为基础，关注材料的系统性和可量化数据平台的构建，重视对长时段、大规模记录中的各种人口和社会行为进行统计描述及彼此间相互关联的分析，以此揭示隐藏在“大人口”（Big Population）中的历史过程与规律。



- 量化数据库的基础通常是系统化、结构化的资料，以便使用统计软件开展量化分析。
- 或是受1970年代国际计量史学研究中出现的问题的影响，当下的历史学家们对待量化方法并不都是欢迎或支持的。如有学者认为“所有的历史问题之所以重要就在于它难以被量化所回答”。
- 但从20世纪90年代以来至21世纪头10年，在国际史学界，计量史学的理论与方法非但没有被淘汰，反而已经走向了常态化。对有志研究人类社会的学者来说，无论他关心的是历史还是现实，都必须掌握数据分析方法而不能成为数据的奴隶。
- 依靠量化数据库，学者们对大量的系统数据材料进行了统计分析，发现了很多隐藏在海量史料中无法依靠传统阅读发现的新知识、新现象，西方学界也就产生了“依靠数据库发现知识” (Knowledge Discovery in Databases) 的说法。



量化数据库≠定量分析



- 定量分析是计算机技术与人文社会学科融合下产生的一种分析方法，可以针对各种文献进行某些主题的量化分析，属于“人文计算”或“数字人文”的范畴。例如有学者以正史文献为基础，统计魏晋隋唐时期士族所占的比例和影响；或利用一亿多字的文献，建立中国政治思想变迁的专业数据库，对观念和社会互动进行统计分析；或是由霍伊特·朗、苏真、戴安德等人发起的芝加哥大学文本实验室的跨语言、跨学科的ARTFL数字化项目，与上海图书馆合作，创建的大型期刊语料库平台（**Republic China Periodicals Corpus : 1918— 1949**），以此分析民国时期的思想观念和文化范式的发展特征。
- 需要注意的是，即使都属于数据库的统计分析，各个项目应用的计算机技术也是有很大差别的。
- 广义上讲，量化数据库和基于数据库的定量分析方法，属于数字人文的范畴，但是具体而言，从材料的选择和计算机技术的运用，都与历史学传统研究方法更为接近。量化数据库的构建多使用人口、土地、科举、缙绅录等单一性质的文献。



求是亦或解释？



- 十六世纪“科学革命”伊始，自然科学和人文社会科学的学术传统开始分化。



- 自然科学注重求是型学术



- 社会科学推崇解释型学术



- 十六世纪“科学革命”伊始，自然科学和人文社会科学的学术传统开始分化。
- 自然科学注重求是型学术。
- 社会科学推崇解释型学术。
 - 经典社会科学理论如托马斯·罗伯特·马尔萨斯人口学说即是这类解释型学术在社会科学领域的代表。（ Thomas Robert Malthus' *An Essay on the Principle of Population* , 1798/1803/1992）
 - 近些年来很多研究通过批判与揭示，挑战了我们过去的学术解释与理解，这可看成此学术形式在人文学科的表现。



大规模历史数据与“求是型”学术



- 数字人文研究中的大规模历史数据库项目最引人关注的特点是对大规模系统数据的重视以及由此形成的富有活力的分析和描述性的量化。
- 新历史数据让我们能够构建起新的研究目标。这些数据自身或隐藏其中的无声的“事实”，能够证明或挑战我们现有的学术理解。
- 基于量化数据库研究出的事实，含义往往较为一致，利于开展跨时段和地域文化的精确比较，能为我们掌握社会的历史和人类的行为提供一种更为全球化的理解，并进而构建起一种新的、自下而上、由繁入简的史观。



- 作为人文学科的历史学，其研究通常是对已知人类社会现象的理解或解释，是“问题驱动”的研究范式。
- 量化数据库研究不同，它从大规模、系统化史料的统计、运算出发，通过数据分析发现新的史实或现象，再以此为基础，重构我们的历史认知与理论，是一种“数据驱动”下侧重于发现而非解释的新研究范式。
- “数据驱动”是信息科学概念，运用到人文与社会科学界，更恰当的表述应为“量化数据驱动”（Quantitative data driven）。数据驱动下的历史研究，学者们要学会从文献叙述中发现问题转变为从数据分析中发现问题，提出问题并依靠数据计算确认现象并可能地解释现象。因此，“数字人文”尽管是依靠现代电脑技术发展起来的，但它绝不仅是一种技术¹⁶，而是一种方法论意义上的新研究范式。



- 由于“大数据”本身往往涵盖相对完整的“大人口”的多种“长时期”的具体信息，这不仅使得学者可以依托数据库来理解个体是如何被宏观社会环境所规范和影响的，也可以理解这些微观、个体行为又如何集合起来塑造和改变了宏观社会进程，甚至可以用来检验“中观”社会群体和因素，如家族、邻里社区等在宏观环境和微观行为互动下的特点和作用。
- 传统定量研究则倾向于将研究问题和具体操作抽象化，非常倚重数理模型和逻辑推导，强调复杂统计方法的运用，但往往不够重视实证材料或数据本身。
- 量化数据库研究是一种更为基本和宽泛的研究思路和方法，它既能够丰富、完善我们对微观人类历史和行为的认识，还能帮助我们构建更为可靠、踏实的宏大叙事，促进我们对人类社会长期发展规律的进一步认识。



开发与研究量化数据库的历史学价值



- 一：丰富了历史学研究的方法论与范式；
- 二：能有效利用大规模材料，实现不同材料的彼此链接，推动大规模、协作化研究形式的出现



第二讲

可能：量化数据库与中国史研究



- 对史学研究来说，构建量化数据库是兼及史料整理和问题研究的重要工具，代表历史学未来发展的一个新方向。其尊重材料、注重探求事实本源的方法导向，与历史学的学术传统是非常契合的。
- 中国历史学者的研究和教学在此方面仍停滞不前。一方面，大量掌握在历史学者手中的重要人类经验数据资料可能难以被利用，历史学发挥的自身基础学科功用的作为受到限制；另一方面，历史专业的学生对新研究方式方法缺乏足够的了解和学习，在数字时代的竞争力自然受到影响。
- 如何缩小差距、转变思路、在发挥自身优势的同时实施教育改革，仍是历史学科建设亟需思考和解决的问题



中国史的资料与量化数据库构建



■ 中国历史文献中类似材料非常丰富，传统官方文献中至少有三类非常适合数据库化和进行定量研究：

第一：历代户籍材料；

第二：与户籍材料相伴随同样历史悠久的土地及财产占有与分配登记材料

第三：至少自隋唐以来，考试（考核）就成为中国社会选拔精英人才的重要方式，历代皆有数量惊人的科考或官员铨选材料，这些材料历时长，系统化程度高，是不可多得的量化数据库素材。



中国史史料的数字化建设



- 中国史学界长期积累和正在进行的许多文献收集和整理项目，都具有构建大规模量化数据库的先天优势，预示着未来短期内大规模历史量化数据库及相关研究在中国蓬勃发展的广阔前景。
- 目前国内很多学术单位和图书档案管理单位都已经开始重视大规模、系统化历史材料的收集与整理，建立了一些文献保存或文献索引型数据库。
- 如国家与地方档案管的文献材料数码化工程和国家清史工程对清代文献的大力收集和整理；中国社会科学院近代史研究所对所藏约19万件珍贵档案的整理和电子扫描以及规模庞大的“抗日战争与近代中日关系文献数据平台”。



- 山西大学中国社会史研究中心多年来收集、整理了逾千万卷的山西地方社会历史档案
- 中山大学历史人类学研究中心长期以来致力于收集的以闽粤为中心的地方家族谱和民间契约文书；
- 上海交通大学历史系开展的以上海郊区、苏南、浙北地区为中心兼及中西部地区的“县级档案与契约文书的收集、整理与研究项目”已经汇集了8省50县约600万页的档案和30多万件的民间契约；
- 浙江大学地方文书与编纂中心则与浙江龙泉市档案馆合作，将该馆所藏晚清至民国时期总计17333个卷宗，88万余页的地方司法档案进行了整理、电子化和出版；
- 南京大学中华民国研究中心正与江苏省档案馆合作，准备将该馆所藏民国时期江苏高等法院逾12万卷档案整体电子化并整理出版，其体量将是龙泉司法档案的数倍。



- 这些大规模史料收集与文献数据库的构建对材料的永久保存意义重大，将很多分散在各地、各机构的重要史料进行了整合，一些数据库还支持检索功能。
- 这些电子化数据大多是向学界同行开放的，很多材料也可以实现远程查阅，这使得史料获取的难度大为降低。
- 虽然这类文献检索类数据库的保存格式和数据结构仍无法直接用于定量分析研究，但是已经为构建量化数据库奠定了坚实的材料基础，从而在较快实现向量化数据库的转变，进而促进系统、深入地利用这些历史材料进行分析研究分析。
- 可以预见，在未来短期内我国历史学在上述材料相关研究领域将会有长足的发展，不仅会在国家与社会、民间契约、宗族行为与司法诉讼等社会历史研究方向取得许多重大进展，更会为相关社会科学和国际比较研究提供全新动力。



数字化研究的教学与培训



■ 近年来，一些暑期集训课程在推动中国量化历史分析教育方面已经取得了很多进展，丰富的社会科学定量方法的暑期培训项目为学生和教师短期内提升分析方法的修养提供了可能：

- 1) 全面介绍量化史学分析方法和研究前沿的清华大学“量化历史讲习班”
- 2) 以具体数据库为基础注重训练数据分析和数据管理能力的上海交通大学的“中国多代人口数据库暑期学校”等



- 许多大学经管专业开设的发展经济学和经济史课程，也大量涉及量化历史研究的各种经典实例，为历史专业学生和学者深入学习相关理论和与分析应用提供了可能。
- 此外，Excel等大众化数据管理软件的普及与发展，对于数据库构建、管理数据和基本分析提供了便利。SPSS、SAS、STATA、R等专业统计软件在操作指令和交互界面方面也日趋人性化。
- 同时，全球社会科学的发展，为史学家和社会科学家的交流合作提供了可能。跨学科合作交流或团队的建立，可以帮助历史学家们突破技术壁垒。历史学家自身不可比拟的比较优势，更能够帮助他们有效地融入到新的研究潮流中来，对正确构建和分析历史量化数据库意义重大。



数字史学研究的中国意义



- 提倡依托于大规模量化数据库的学术研究，对推动整个人文社会科学的“求是型学术”、真正认识中国社会和历史发展的特征、平衡东西方学术见解均意义重大。
- 当前社会学科中较为普遍和“科学”的统计抽样调查等方式，尽管在数据收集的功效和人口层面的代表性等技术层面，统计抽样社会调查有其得天独厚的优势，但是由于西方理论范式先行，我们在中国实施抽样调查项目往往仍需要大量借助西方已有研究经验，通过研究直觉提出问题和设计问卷时难免会受到西方“已知”事实和理论的影响。许多重要研究成果需确认或修正了很多东西方“共性”的认识，但我们也在一定程度上丧失了认识中国“未知”特质、完善和平衡国际学界对不同社会差异的根本认识的可能。



- 构建大规模历史计量数据库，可以帮助中国历史学者们更“科学化”、“国际化”。
- 如果有越来越多的中国历史学者们通过构建大历史数据库，进行定量研究，中国历史学将更有条件系统展现出人类社会在某些方面的长时段与大转变，中国史学者的工作必然会更加引起世界的关注。
- 越来越多的计量数据库构建成功和研究的展开，还可为国内历史学专业学生的培养提供新素材和新方向，国内史学界应该对此给予必要的关注、重视。



- 从国际层面来说，一国学术研究的影响力与其国家实力和发展状况密切相关。中国的崛起与强大应该是本世纪最重要的国际变革。
- 因此，中国史学界应该有自信和觉悟，在扎实自身研究的同时，多参与国际合作和国际比较，向全世界展示更多的中国成就与中国经验。这种中国经验并不一定是几十年前我们常谈到的“唐宋文明”或“明清盛世”，而可能是20世纪以来的历史，比如革命历程与古老中国现代化的转变等。
- 正如我们在分析高等教育材料时发现的那样，百年来被认为落后的中国，在很多制度设计和社会效果上取得的成绩可能是西方社会无法比拟的。转换角度、扎实研究，中国史学对国际学术的贡献会大不一样。



李中清-康文林教授研究团队的数字化实践



- 李中清-康文林教授研究团队（以下简称李一康团队）一直倡导将大规模个人层面微观数据库构建与分析引入到历史研究中来，并积极进行实践。
- 通过多年的学术思考和积累，目前团队逐步围绕着四个规模类似，主题独立，但又彼此关联的个人层面微观历史数据库进行构建和研究工作。它们分别是近代以来的土地财产、大学生、职业技术人员和官员群体个人层面大规模历史数据库，这些数据库将帮助我们更深刻的理解社会经济史研究中非常关键的四个问题：谁获得财富，谁获得教育，谁获得职业，又是谁获得权力。



1、财富



- 中国多世代人口数据库——双城部分(China Multi-Generational Panel Dataset, Shuangcheng, CMGPD-SC)，提供了1866-1913年间生活在中国东北地区，今黑龙江省双城县的，基于人口统计学与社会经济特征的长时期的个体、家庭及其它相关方面的当地居民信息。
- C M G P D - S C 已经公开，可通过以下链接获取，网址：
<https://www.icpsr.umich.edu/icpsrweb/DSDR/studies/35292>，上网时间：
2019年3月6日。



- **CMGPD-SC**数据库所基于的原始历史材料主要来源于双城地区的八旗户口册以及土地登记册，由当地八旗管理机构编造和维护。
- 户口的登记首先以村庄为单位，村庄之下按户登记。每个条目首先记录户主的信息：其迁出地，民族，原属旗，职业，姓名，年龄，以及其他自上次登记后的重要人口信息变动情况。随后户口册记录了户主其他直系亲属（妻子，儿女，父母）的信息，包括与户主的关系，姓名，年龄，职业等，以及其他与户主共同居住的亲属的相关信息。户口册中对于所有家庭成员的索引标识都是根据他们与家庭主要成年男性的关系来确定的。
- 由于人口登记在土地分配中非常重要，在每年的第十一月，双城地区的政府每年都对户口册进行更新，并将更新过的户口册送至治理吉林的八旗将军处审阅。因此，与其他地区的八旗户口册三年一更新的制度，双城地区的旗人户口册是找到的唯一一个每年一次编制更新的户口册。双城八旗户口册在细节和完整度上都拥有较高的质量。



2、教育



- 李-康研究团队从**1980**年代中期起，便开始注意收集和整理中国教育精英的个人层面系统性历史档案资料，以期能深入、准确地掌握中国教育精英群体的长期变迁。
- 历经三十年的努力和多方合作，目前已初步掌握近**300**年来中国教育精英社会来源的主要信息（如家庭背景、地理分布等），并开展了量化数据库的构建和分析工作。
- 这些信息规模庞大、记录系统较为连续，几乎涵盖清代全部进士和官员以及大部分举人、贡生（**1644-1911**）、大部分中华民国时期的大学生（**1912-1949**）以及两所中华人民共和国时期国家与省级精英大学的全部大学生（**1950-2008**）。



民国学籍卡举例



40-3830 Li Cheng Tao

國立西南聯合大學 學生註冊片

學號：342011 姓名：李成濤 別號： 學籍：

籍貫：1. 楊 市 年齡：民國 34 年 19 歲 (出生於 15 年 11 月) 性別：男

入校年月：34 年 現在：2. 學院 第 二 年級 曾否入黨： 已婚未： 否

經過學校：(校名：1. 楊 12 大學 將來志願： 1. 183

(校址： 貴州 達縣

家長姓名：李成濤 關係：父子 職業： 家庭通訊處： 貴州 達縣 1. 183

保證人姓名： 關係： 職業： 兄姊、通訊處： 1. 183

學生通訊處： 校內： 校外：

備 註

卅四年度錄取新生 34 年度物 = 正式生 19

Parent's Name, Relation and Occupation

Guarantor's Name, Relation, Occupation

Place of Origins

Age

Sex

Marital Status

Family Address

Guarantor's Address

National Southwestern Associated University



3、职业



- 中国劳动力量化数据库（The China Workforce Database 或CWFD），是收集民国时期和建国后中国新兴职业群体信息的数据库，它是李-康研究团队这两年新启动的数据库构建和研究项目。
- 目前，该数据库初步拟定包含的职业群体类型，主要有大学教职员、医生、工程师、会计和律师教育和工作历史。建成后，该数据库将包括超过100,000名专业人士的个人微观记录。同时，该数据库具有和中国大学生量化数据库匹配连接的巨大潜力。



經濟部會計師登記證書稿底

姓名	牟敦輝
年 齡	二十九歲
籍 貫	山東日照
學 歷	國立北平大學（現改稱燕京大學）經濟系畢業
經 歷	曾任北平山東中學國文歷史教員農本局秘書 中國農民銀行總管理處會計處辦事員現任羅漢 中國農民銀行會計主任職務
會計師檢覈及格證書號數	會計檢字第一壹玖柒號
核准年月日	中華民國三十三年九月十八日
證書號數	捌拾柒號
部 長	秘書長 曹鈺斌
政務次長	司 長 王 珪
常務次長	科 長 王 珪
承辦人	王 珪

Source: Registration Card of Public Certified Accountant, Ministry of Economic Affairs



4.权力



- 清代保留至今的缙绅录文献规模浩大，提供了连续性的官员记录，是建立清代官员群体的大规模历史量化数据库的理想史料。
- 缙绅录是记录职官的职掌、姓名、出身、籍贯、字号等基本情况的专书，最早可以追溯到南宋的《班朝录》，在明朝已颇为流行。清代继承了出版官吏名册的传统，缙绅录成为记录全国职官信息最全面、系统的资料。
- 现存的清代缙绅录文献规模庞大，每种缙绅录几乎覆盖了当时所有的文职官员和部分八旗驻防官员，内容简洁、条目化、全覆盖，并且具有持续性的特征，非常适合建立长时段、大样本的量化数据库。缙绅录为学界提供了极为系统的个人层面微观数据，同时这些数据每季度出版，具有极好的历史连续性，而且至少几乎完全包含了**1760-1911**年间所有清政府官员的信息。



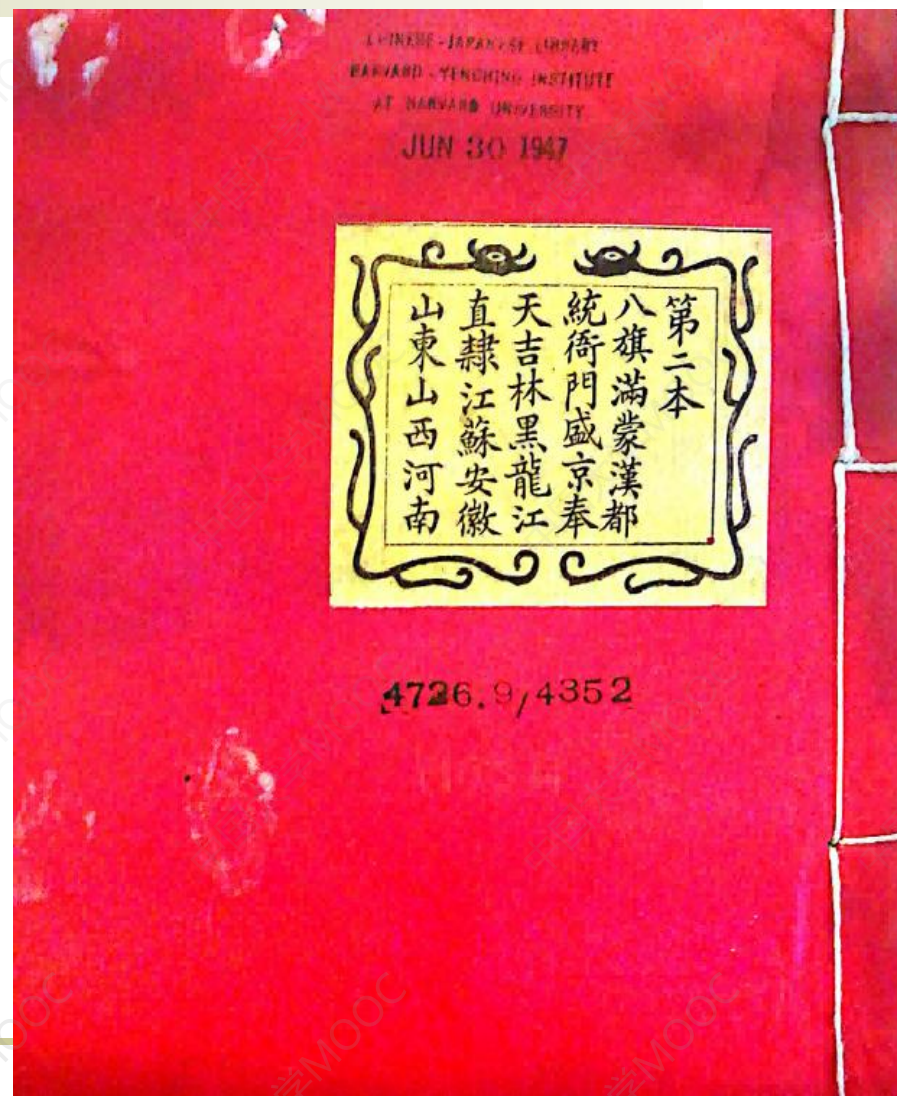
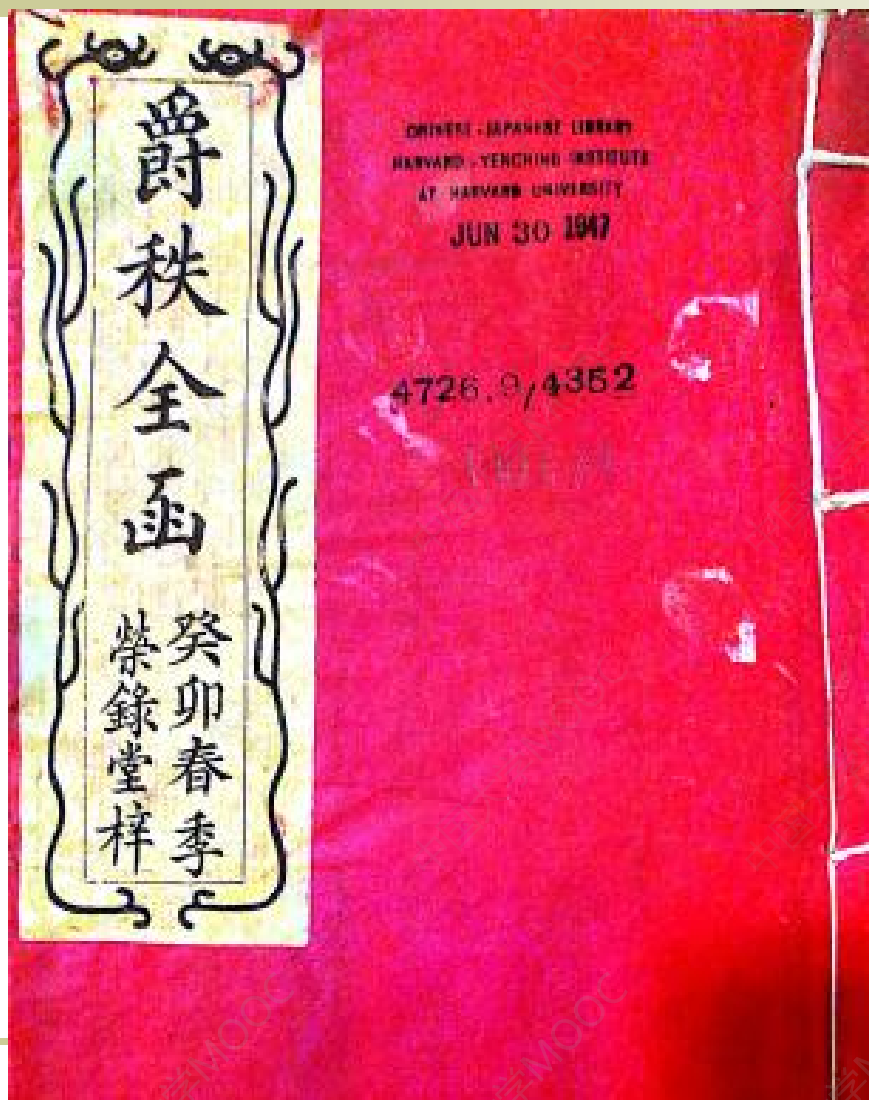
《清代缙绅录数据库》文献来源



- 以清代缙（搢）绅录文献为基础
- 清华大学图书馆现存230余种，其中209种已结集出版，即《清代缙绅录集成》（大象出版社2008年版）
- 国家图书馆、北京大学图书馆等十九个图书馆，另有120余种未出版
- 海外图书馆也藏有一定数量的缙绅录
- 目前录入的缙绅录数据以清华大学图书馆出版的《清代缙绅录集成》为基础



哈佛大学所藏缙绅录举例





《清代缙绅录》实例





量化数据库与历史学的未来走向



同时，本研究也展示了一种新的历史微观数据库的分析方法，利用该方法如何改变我们对历史和当代中国的理解。通过连接个人，他们的祖先和后代，以及他们不同的背景，不同的数据集和时间，我们可以更好地理解在不断变化着的政治价值、政府政策、人口演变、社会经济背景下的代际动态，从而了解：谁将获得权力，谁能获得财富，谁能获得教育，谁能获得工作。从而更好地衡量我们过去和现在的历史连续性和革命性变化。





Thank You From the Lee-Campbell Research Group



Web Access to Lee-Campbell Research Group:
<http://www.shss.ust.hk/lee-campbell-group/>



第三讲

如何：量化历史数据库的构建



■在数字人文浪潮的席卷之下，大数据的观念深入人心，但是如何定义大数据却存在分歧。有人提到数据库的容量可能要达到一个**PB**（即**1024TB**）才能称作大数据。有学者认为和今天计算机所能处理的数据相比，大部分历史数据库只能算是小数据。大数据、小数据如何区分，是否存在抽象的统一标准？计算机软件的处理能力是否能成为衡量数据大小的标准？

■有人指出大数据是“大小超出了传统数据库软件工具的抓取、存储、管理和分析能力的数据库”，但大数据的概念还是带有主观性的，并不能仅仅以超过多少**TB**为标准，不同的行业、领域，以及常用的通用软件，都可能影响到大数据的界定，可能没有统一的大数据标准。



- 脱离了具体的专业、数据结构和常用的处理数据的软件来讨论数据的大小，并没有实际意义，也会因为追求普遍标准而变得虚无化。
- 例如，量化数据库与检索型的历史数据库相比较，单纯按照数量单位（**TB**）来说，量化数据库要小得多，但因为两者数据结构不同，并没有可比性，更不要说与自然科学、信息产业等行业的数据库相比，所以脱离了具体的专业背景讨论数据的大小，并不是一个很好的选择。



构建量化数据库，首先是要选择合适的历史文献，选择的方式与检索型数据库完全不同。

检索型数据库要全面收集所有相关文献，追求大而全；构建量化数据库的目是为了进行定量研究，由于文献材料的复杂性，一般仅使用单一性质材料构建量化数据库。

在选择材料时，除了要遵循史学考据的所有学术规范，还要求史料具有系统性和延展性。



■ 考察历史文献是否适合构建量化数据库，主要从整体上评估，除了非伪造和可靠性，还包括从史料提取数据的数量和质量，可以概括为五个方面：

- (1) 从整体上考据文献的真实性；
- (2) 提取数据的难度；
- (3) 数据的时间和空间属性；
- (4) 数据的类别、完整度和数量等；
- (5) 数据的延展性。



(1) 量化数据库的构建与考据



- 量化数据库的考据从选择史料的时候就已经开始了。
- 和所有的描述性史学研究一样，首先要对文献进行考据，辨别真伪。量化数据库的建设需要巨大的时间周期和经费成本，考据工作的意义非常重大。
- 史学考据的学术规范，都是建立量化数据库必须遵循的规则。首先要考证史料的真伪，要谨慎的判断史料本身的立场和特征，所有史料都是来自于某些官方机构、私营书坊或个人，这些主体的意愿显然会影响史料的价值和质量。



- 需要特别强调的是，大规模数据的考据与个体数据的考据有一些差异。
- 用来创建量化数据库的历史文献，基本上都是系统的大规模的历史资料。所有史料都是对历史记忆的描述，存在一些错误在所难免的，在选择历史文献时，特别需要从整体的视角，全面的认识和考证史料的价值。
- 具体说来，对于描述性史学来说，如果某个数据或人物的记录存在错误，那么这个史料可能完全没有价值了，但对于系统性的大规模数据来说，存在少量的错误是在所难免的，需要从整体上考察史料的价值。



录入错误纠正和变量的考据、标准化



- 量化数据库在进行数据分析之前，必须进行大量的录入错误纠正和变量标准化工作。
- 研究组有长期稳定合作的专职录入人员，这在很大程度上保证了录入的准确率和效率，即便如此，录入错误也是不可避免的。
- 另外，历史文献中常常出现异体字或同音字表示同一专有名词的情况，缙绅录也不例外，地名、科名、旗分、铨选方法和职官等名称经常用不同的异体字和同音字表示。这在含义表述上没有问题，但如果使用程序识别，往往会识别为不同概念，直接影响结论的准确性，在进行分析之前，必须进行数据清理和标准化，这个过程也是考据的过程。



- 纠正录入错误和变量的标准化是一个非常艰难和漫长的过程。首先需要对相关领域的制度史深入了解，在此基础上制定纠错方案。与此同时，各种异体字和同音字的统一，也需要充分了解专有名词和习惯性的称谓，制定统一的标准化方案。这个过程就是历史学考据的过程，要求证据严谨。
- 其次，由于数据往往达到几百万条，量化数据库的考据也和以往的史学考据方法不同。在以往的考据思路指导下，要结合统计软件或计算机程序强大的批量处理功能，才可能完成量化数据库的考据和标准化——将传统的史学考据方法和计算机程序的批量处理能力相结合，演绎出新的考据形式。



■ 即使是在统计软件或计算机程序的支持下，可以批量的处理相同的错误或异体字、同音字等，也需要专业学者考订制度史和名词，制定修改方案，并进行人工核对，这是一个非常漫长、艰苦的过程。

■ 以《中国历史官员量化数据库——清代》为例，地名（籍贯和任职地）、旗分、科名和铨选方式等信息的标准化和纠错，不同的变量，方法各异，所有的错误、异体字、同音字往往也是有规律的，利用计算机程序批量修改和标准化处理。地名要考虑复杂的层级隶属关系。职官变量的数量最多，但由于清代官制的制约，所有的变化都遵循固定的程式化表达，在参考《清会典》、《大清会典事例》等关于清代官制的政书，以及《清代地方官制考》、《清代国家机关考略》等今人研究的基础上，对官职变量进行考证，确定官职变量和标准化名称之间的对应关系，将官职考据与计算机技术融合后，批量进行标准化处理，虽然思路简洁明确，但过程却是非常繁琐的。



(2) 提取数据的难度



- 是指将历史文献的主要信息提取出来的难度。量化数据库的建设过程，就是将主要内容以数据的形式提取出来，录入**Excel**等软件中，形成数据集合。
- 一般情况下，大规模的人口、户籍、土地、科举、缙绅录等资料是创建量化数据库的首选史料。这些史料虽然规模庞大，但条理清晰，格式固定，便于提取信息和录入。
- 因为量化数据库的录入仅是改变文献的格式，内容并不更改，即使没有专业的历史知识，只要经过培训，就可以快速而准确的掌握录入格式，核对起来也比较方便。
- 一般常用的统计学软件有**Excel**、**Stata**、**Spss**等，以及 **Python**、**R**等计算机语言



- 如果是描述性的史料，需要专业人士才能辨别信息，摘取变量，即使不考虑时间和经费成本，也很难找到足够多的专业人士进行多年连续的录入和核对。
- 另外，文本史料摘出的数据，信息要素可能参差不齐，影响数据的整体质量和完整性。



(3) 数据的时间和空间属性



- 历史文献的主要信息——可提取的每个数据，都要有明确的时间和空间属性，这一点非常重要，如此才能对每个数据进行时空定位，也是量化数据库进行精确的定量分析和微观研究的前提。
- 说到时间，至少具体到年份。如果材料时间跨度较长，百年甚至千年，并且均匀的连续分布，是比较理想的，不仅可以呈现历史变迁，还可以避免短期数据带来的以偏概全，影响结论的准确性。
- 每个信息还必须有明确的空间定位，具体到省、府、县（州），甚至可以定位到自然村。



- 为了方便理解（2）（3）两点，即提取数据和时空信息的重要性，这里以已经公开的《中国辽宁多代人口数据库（China Multi-Generational Panel Dataset - Liaoning，1749—1909）》为例进行讨论。这是李中清—康文林研究组历时20余年建立的中国清代人口量化数据库。
- 数据库人口的原始资料，来源于辽宁档案馆收藏的盛京内务府户口册，已经录入人口册28种，共计673册，总人口约达26万，分布在1350余个自然村中，是1749年至1909年的连续人口数据，共有150余万条记录。每条记录包括人口的时间和空间、姓名、身份地位、婚姻情况、出生死亡以及各种家庭状况等信息。
- 数据库的每一条记录的时间和空间分别可以具体到年份和自然村，时空定位明确，这样才能发挥定量分析方法的优势，可以进行微观的、精确的分析。



(4) 数据类别、完整度和数量等因素



- 理想数据的类别要足够丰富并相互关联，可以构建足够复杂的数据库，这样才能发挥定量分析方法的优势。
- 以王业键先生的“清代粮价数据库”为例，是自乾隆元年（1736）开始各体会向皇帝奏报的各省、府（直隶州厅）的主要粮食价格数据库，在研究清代经济史中发挥了巨大的作用，但变量的数量就显然有些单薄，没有足够的复杂性。
- 清代粮价数据库是经济史研究过程中很重要的材料，但如果作为独立的研究对象，数据类别仅有粮食的价格和相关的时间和空间信息，就显得略少。



参考文献



- 任玉雪：《量化数据库与史学考据方法》，《南京大学学报》2020年第5期；
- 梁晨、董浩、李中清：《量化数据库与历史研究》，《历史研究》，2015年第2期；
- 陈必佳：《再论〈缙绅录〉记载的准确性及其史料价值》，《清史研究》，2019年第4期。



第四讲

要点：从结构数据到追踪数据



- “互联”、“互动”是互联网的重要特点，历史数据库的构建与研究，一样应重视数据库自身信息的“内连接”和数据库之间信息的“外连接”，进而尽可能地实现“数据追踪”，由此方能体现出电子数据的优势，推动新技术环境下史学研究地深化与发展。
- “内连接”和“外连接”本是信息学科针对关系型数据库（**relational databases**）数据查询、分析时采取的方法，此处将这此概念移用于历史数据库的构建与研究中来。



- “内连接”主要包含两层含义，一是指某一数据库在建设时力争利用多来源史料中能相互匹配的信息进行连接以扩充数据库内容；二是指利用某一数据库进行研究时，应注意其中信息的相互连接，以拉长研究的时段、代际或空间范围等。
- “外连接”主要是指不同主题但内容、关系等具有连接可能的数据库，在开放的前提下，实现相应数据信息的连接，从而突破单一主题数据库信息范围与广度上的局限，从学术研究层面真正实现数据库建设与研究的“共享”与“共赢”。



- 历史数据库信息“内连接”与“外连接”的最终目的，是希望将存储于各主题数据库中的相关信息逐步从静态推向动态，实现“数据追踪”。
- 如利用多种材料或数据库的互相连接，可以连缀起某个个人和某些群体不同时间节点的信息，这就形成一种“追踪数据”。
- “追踪数据”的形成，使得某些历史上个人的经历能够丰富和连续得呈现出，对于理解、研究这些个人或群体的行为特征具有极高的学术价值。



“内连接”：不同史料与信息的连接



- 即便是单一主题的历史数据库，其包含的具体史料也是规模庞大，类型繁多的。而这些史料中的某些信息很可能可以被连接起来，形成数据链。
- 比如以金观涛为首构建的“中国近现代思想史专业数据库（1830-1930）”收录了多种晚清民国的报刊和个人著述，总字数达到了一起两千万。很多学者通过检索和统计工具的帮助，将这些海量的史料连接起来，分析近现代史上某些重要的概念，如“主义”、“青年”、“民族”等的形成演变，或“验证”，或“突破”了既有研究的结论，实现了学术创新。



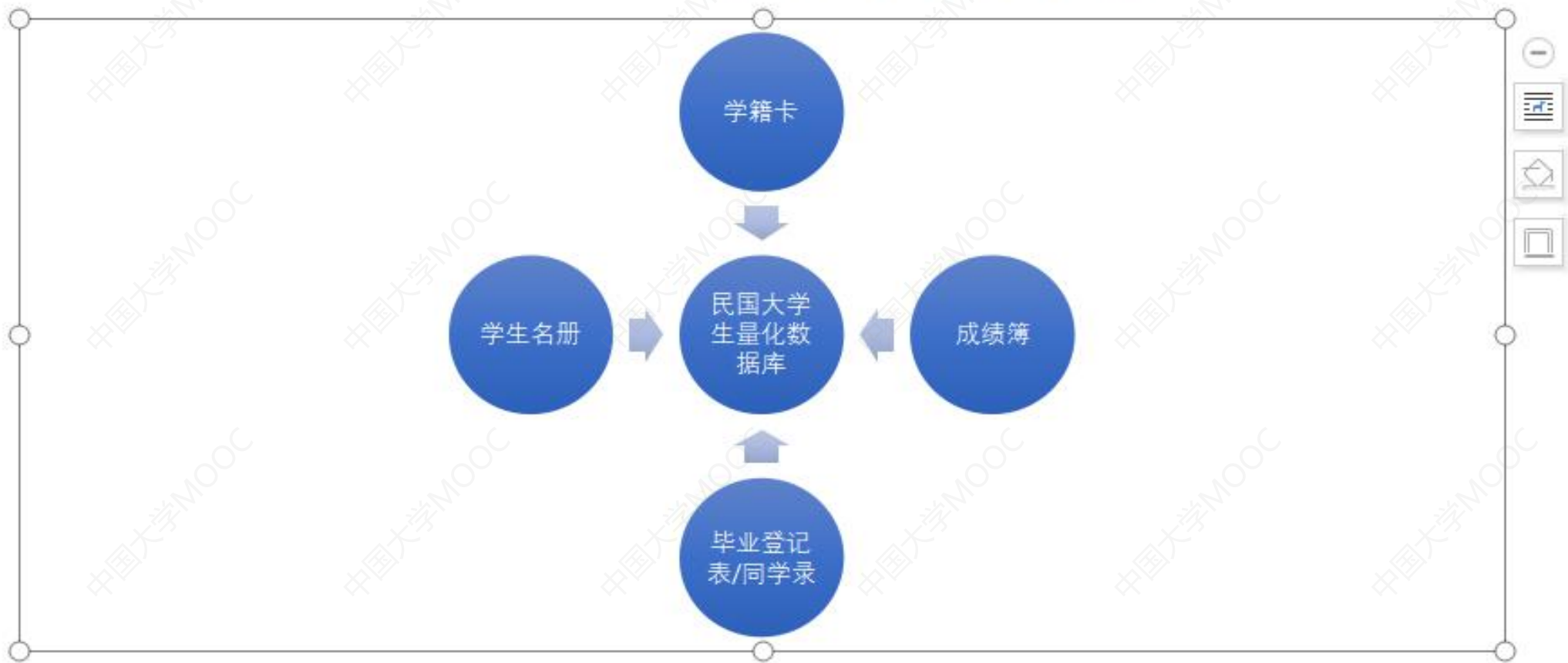
- 又如人物主题数据库是较为常见的一类历史数据库，通过内部连接，往往能够形成以人物为主干的信息链条。
- “民国大学生信息数据库”，即是以个人基本信息为基础的数据库。数据库中每一名学生的信息既有个人自身的，也有家长和保证人信息。通过数据库内部连接计算发现，很多家长本身也是大学生，能够连接形成追踪数据，进而获得三代人的基本信息，使得近现代史上多代际教育获得与社会流动研究具有了坚实的数据基础。而以亲友为主的保证人通常是学生最主要的社会关系，将不同学生的保证人信息进行连接，可以发现很多学生具有同一家长、共同亲友，这就使得学生的兄弟姐妹群体等重要社会关系网络被挖掘出来。



- 民国时期一方面并不存在全国性标准化的统一学生登记表格，因此民国大学各校所涵盖学生信息不相同；另一方面档案中，很多学校还保存了众多的非标准的登记表格，如同学录、调查表等，无论对有学籍登记表的学校，还是完全没有学籍登记表的信息，这些材料都可以很好的补充、扩充数据库，很好地帮助完成研究。
- 因此，在民国大学生数据库中，我们有**70%**左右的数据来自于学生入学时填写的信息，另有**30%**比例则来自其他材料，譬如成绩册、毕业生登记表、同学录等。



图 3-3 民国大学生数据库数据来源





“外连接”：不同主题数据库的连接



- 中国历代人物传记资料库（China Biographical Database Project, 简称CBDBB）的领导者，哈佛大学包弼德教授指出，随着“数字人文”的发展，历史数据库的建设大为增加，数字数据（史料）迅速膨胀，数据生产者之间的协作也就越来越重要。
- 通过数据库开放和协作，形成不同主题数据库的信息连接，是其中最能发挥数字数据优势，推动学术研究的一层。



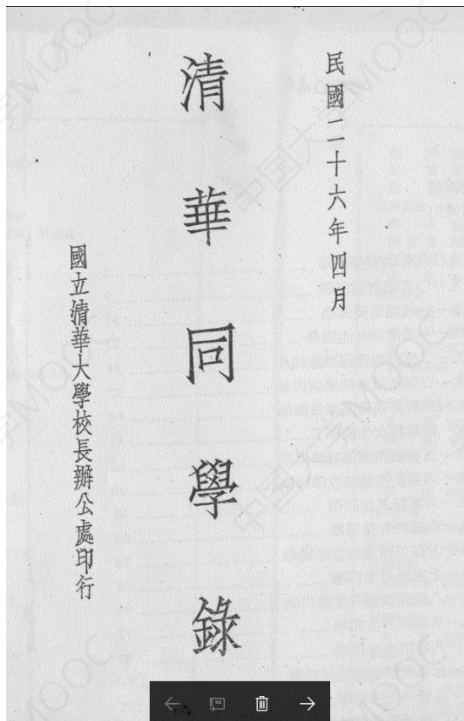
- “民国大学生信息数据库”收录的虽然主要是民国时期教育精英的基本信息，但数据库中早期学生的家长，往往已经上溯到了晚清时期。
- 香港科技大学康文林教授领导建设的清代《缙绅录》数据库，包含了整个清代官员信息，其中有一些宣统年间的数据，经过计算发现能够与“民国大学生信息数据库”中的学生家长信息进行匹配，这对理解民国大学生的家庭来源以及近代教育转型对家庭的影响都颇有助益。
- 《缙绅录》数据库本身又能够与包弼德教授领导的CBDB数据库中的明清人物、进士题名录等材料连接起来，这些官员人物的个人信息也由此大为丰富起来。



- “外连接”还可以应用于主题数据库与大型文献电子数据库的连接。越来越多的报纸、期刊乃至各位史料文献都逐步电子化，这类文献数据库中蕴含有大量的尚未结构化的历史信息，虽然多数数据库只提供开放检索而不能进行后台数据的直接抓取，但依靠手动检索和核对，还是能够与一些主题数据库实现系统性数据连接的。
- 比如在“民国大学生信息数据库”中有民国清华大学的学生资料，其中也包含1929年前清华全部庚款留美生。这批留美生是民国时期最为成功的系统留美者，梳理他们的职业获得轨迹，对理解民国社会发展和近代化等大问题意义重大。



材料：纸质材料与数据库的结合



游美畢業回國學生一覽表

第一批 宣統元年八月

姓名	籍貫	學位	現任職務	通信地址
張南良	江蘇無錫	森林碩士	長沙師範大學中學部主任兼	全上
張延金	江蘇金壇	電機工程師	南京東南大學	全上
張準	湖北枝江	科學士	南京東南大學	全上
張亮	廣東番禺	科學士	天津京奉鐵路局	全上
張慶	浙江鎮海	科學碩士	上海南洋中學	全上
張慶	廣東增城	科學碩士	廣東廣州高等師範	全上
程鴻法	江蘇無錫	機械工程師	上海北蘇州路三十三號	全上
程鴻法	江蘇吳縣	機械工程師	上海英大馬路一號白拔格洋行	全上
程鴻法	浙江吳興	科學碩士	哈爾濱中華總商會行會計	全上

清華同學錄

清華同學會歸國同學錄

清華同學會，是清華一九一三（癸丑）和一九一三（壬辰）兩班，在一九一三年夏天發起的，這會的宗旨，是聯絡校友，聯絡留學，當時也曾通過章程，選舉職員，等到一九一四（壬午）夏，又發出許多入會券，聯絡以前直接間接留學的同學入會，到一九一四（壬午）夏，一九一三和一九一四兩班，中等科學生十人，女生九人，同時啟行，他們在船上，重經章報，補充補給，就舉了周自齊為總理，兩先生為名譽會長，周自齊先生為名譽會員，趕到這一百多同學到了美國，這會總算有了三三兩兩畢業生為會員，他們對同學會甚是熱心，從起中文和英文的季報，後來日見發達，同學會通外國，所以將同學會改稱，為美國東部，和中西兩部支會，會員對會甚為踴躍，已經很好，對於國內清華同學，更是熱心的幫忙，從一九一三年成立，到一九二〇年，這幾年工夫，回國的會員漸次多了，所以又將國內支會，按本會的章程，無論何處地方，有了十個清華官費生，便可成一支部，一時成立的，有上海支會，北京支會，不久因有散居天津和山東的同學加入，所以北京支會，就改為華北支會。

附錄

八十五

留美回國之清華同學一覽表

第一批

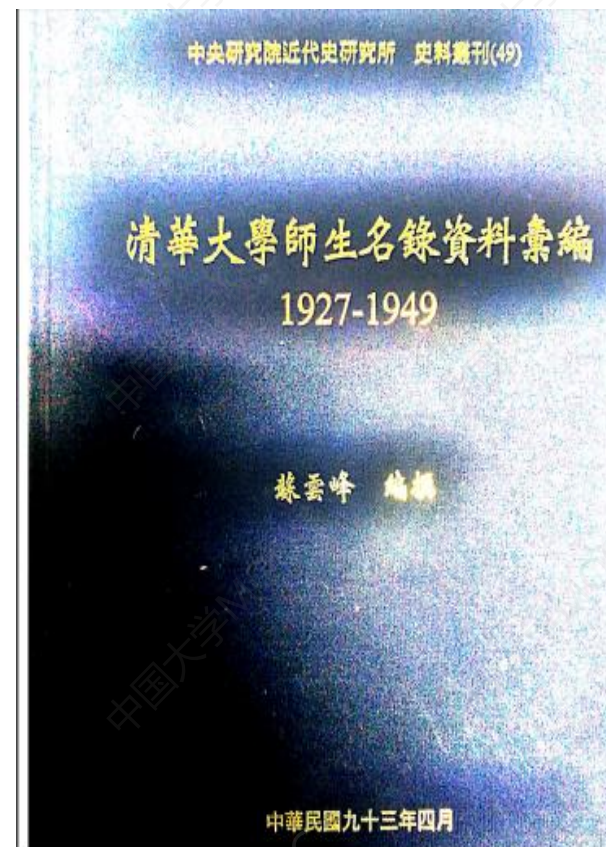
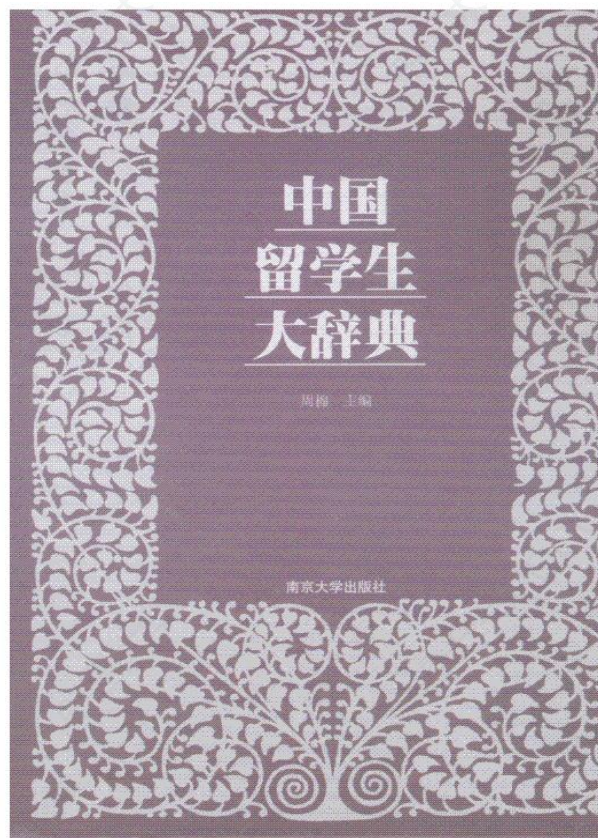
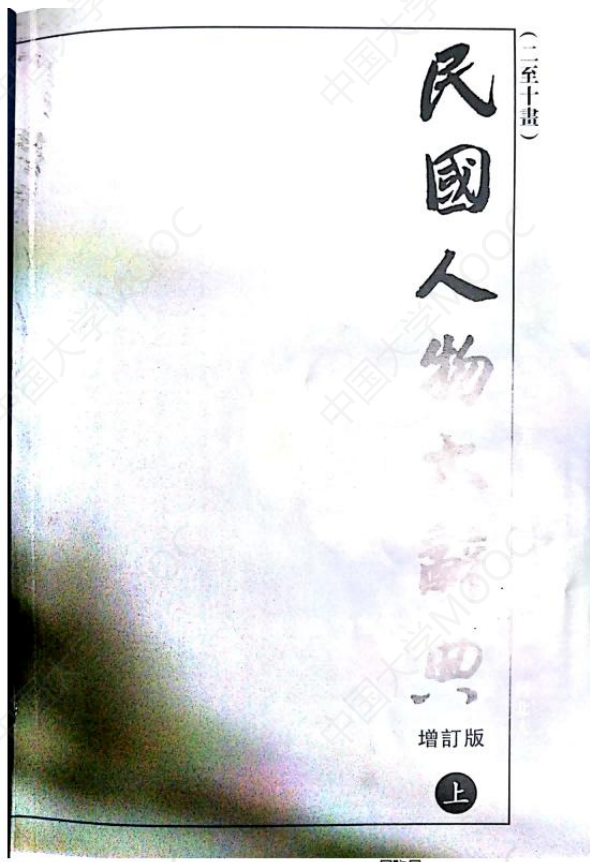
鄧培誦	北京東安門大街國海樓
盧景泰	廣東省城三甫街二十四號
戴濟	廣東省城三甫街二十四號
王長平	南京大學
金海	南京大學
吳玉麟	南京大學
李連發	南京大學
王璉	南京大學
張廷金	上海南京路
嚴家驊	上海南京路

留美回國之清華同學一覽表

羅惠倫	上海南京路
魏文彬	上海南京路
鄧照章	上海南京路
張昌運	上海南京路
賀裕慶	上海南京路
程義法	上海南京路
程義法	上海南京路
梅貽琦	上海南京路
張華	上海南京路
張福良	上海南京路
陳慶堯	上海南京路

■ 1937年 1920年 1922年 1925年

■ 依靠“内连接”，利用清华校方1937年出版过全校性的《同学录》，我们能获得60%留美生的唯一职业信息。





可检索数据库



[收書一覽](#) [分類檢索](#) [條目檢索](#) [全文檢索](#) [閱讀記憶](#) [使用幫助](#)

南京大學

檢索字詞	出版時間	期號
次檢索詞		
邏輯選項		
☐ 与		
☐ 或		
☐ 非		
範圍選項		
期 號 *請輸入數字		
出版時間 不限		
出版地點 不區分		
☒ 輸入簡化字請在口中打✓ 用繁體字檢索結果更精確		

[開始](#) [停止](#)

114.212.102.86[南京大學]歡迎您! [用戶中心](#) [退出](#)

全國報刊索引

見證近200年歷史變遷
Witnessing the Last 200 Years of Chinese History

☒ 所有資源 ☐ 近代期刊 ☐ 現代期刊 ☐ 中文報紙 ☐ 外文報紙 ☐ 行名錄

報道數據量超過5000萬條。揭示報刊數量達5萬余種的特大型二次文獻數據庫，目前每年更新500萬條。

☒ 全部類型 ☐ 正文 ☐ 圖片 ☐ 廣告

[高級檢索](#) [專業檢索](#) [文獻導航](#)



姓名	年份	性质	籍贯	专业1	专业2	专业3	学士1	时间	学士2	时间	硕士1	时间	硕士	时间	博士	时间	其他	时间	毕业学校1	毕业学校2	毕业学校3	职业	通讯处	职业2
王士杰	1909	考选留美同学	浙江奉化	文学	哲学			1909			1912								哈佛大学			浙江省立第四中学教员	浙江奉化大桥咸亨号转	
王仁辅	1909	考选留美同学	江苏昆山	算学			理学士	1913											哈佛大学			北平师范大学教授		
王长平	1909	考选留美同学	山东泰安	教育心理										博士	1914			密歇根大学						
王健	1909	考选留美同学	北平	化学			文学士	1914			理科硕	1916							威斯康辛大学	哥伦比亚大学		天津华北制革公司经理	天津河北大王庙街华北制革公司	
王进	1909	考选留美同学	浙江黄岩	化学工程			化学工程师												理海大学			美国明尼苏达大学化学学院		
方仁裕	1909	考选留美同学	江苏青浦	已故																				
朱复	1909	考选留美同学	江苏嘉定	土木工程			土木工	1913											理海大学			浙江省区救济院总干事	杭州吴山路求是里2号	
何杰	1909	考选留美同学	广东番禺	采矿			采矿工	1913			理科硕	1914							科罗拉多大学	理海大学		国立中山大学地质系教授，两广地质调查所所长		
李进霞	1909	考选留美同学	湖南湘乡	冶金			冶金工	1916											哥伦比亚大学			梧州三角嘴广西大学教授		
李鸣麟	1909	考选留美同学	江苏江宁	化学工程			化学工	1913											威斯康辛大学			南京实业部矿业司科长		
吴玉麟	1909	考选留美同学	江苏吴县	电机							理科硕士								温特沃斯技术学院			建设委员会威整堰电厂厂长	江苏苏州娄门外唯亭镇吴问心堂	
吴清度	1909	考选留美同学	江苏镇江	电机工程			文、理	1913			理科硕	1914							伊利诺伊大学	科罗拉多大学		铁道部技术标准委员会专任技术员		
刑契莘	1909	考选留美同学	浙江嵊县	造船工程	军舰制造		理学士	1914			理科硕	1916							麻省理工学院			青岛市政府工务局局长		
金邦正	1909	考选留美同学	安徽黟县	森林			理学士	1914			林业硕	1914							康奈尔大学	Ibid		北平上海商业储蓄银行经理		北平上)
金涛	1909	考选留美同学	浙江绍兴	土木工程			土木工	1912											康奈尔大学			平绥铁路工务处长		平绥铁I
邱培涵	1909	考选留美同学	浙江吴兴	农	商		理学士	1914			理科硕	1915							康奈尔大学	威斯康辛大学		上海永明人寿保险公司华经理		上海永F
秉志	1909	考选留美同学	河南开封	农			农学士	1911							博士	1913			康奈尔大学	Ibid		南京科学社生物研究所主任，北平静生生物调查所所长		南京科
胡刚复	1909	考选留美同学	江苏无锡	数理			学士	1913			硕士	1914			博士	1918			哈佛大学			上海大同大学校及上海交通大学教授		
范永增	1909	考选留美同学	江苏上海	卫生工程							理科硕	1913							麻省理工学院			上海七地局服务		
徐承宗	1909	考选留美同学	浙江慈溪	普通文科			文学士												哈佛大学			苏州疯人医院		
徐佩璋	1909	考选留美同学	江苏吴江	化学工程			理学士	1914											麻省理工学院			上海市公用局局长		上海市
高纶瑾	1909	考选留美同学	江苏江宁	铁道管理			学士	1912			硕士	1915							密歇根大学	宾夕法尼亚大学		衡州粤汉路交		胶济路)
唐悦良	1909	考选留美同学	广东中山	教育	政治		文学士	1913			文学硕	1915							耶鲁大学	普林斯顿大学		北京大学教授		北平东I
袁钟铨	1909	考选留美同学	江苏江宁	电机															M. I. T					
张廷金	1909	考选留美同学	江苏无锡	电机			机械工	1913			电气工	1914							俄亥俄大学	哈佛大学		交通大学电机工程学院院长兼教授及中华三极锐电公司经理		
张福良	1909	考选留美同学	江苏无锡	森林			学士	1913			林业硕	1915							耶鲁大学	Ibid	哈佛大学	全国经委会会计正		
张淮	1909	考选留美同学	湖北枝江	化学			S. B.	1915											麻省理工学院			清华大学化学系教授兼主任		
陈兆贞	1909	考选留美同学	广东番禺	铁道管理			B. S.	1915											麻省理工学院			上海交通大学服务		
陈煨	1909	考选留美同学	广东增城	化学工程			B. S.	1913			M. S.	1915							伊利诺伊大学	麻省理工学院		中山大学理工学院院长		
陈庆尧	1909	考选留美同学	浙江镇海	化学			B. S.				S. M.	1915							伊利诺伊大学	哥伦比亚大学		上海吴淞镇京沪铁路局机厂化验室服务		
陆宝淦	1909	考选留美同学	江苏常熟	农	化学		B. S.	1913			S. M.	1914							伊利诺伊大学	哥伦比亚大学				
梅貽琦	1909	考选留美同学	天津	电机			B. S.	1914											沃尔切斯特理工学院			国立清华大学校长		
程义法	1909	考选留美同学	江苏吴县	采矿							E. M.	1914							科罗拉多矿业学院			南京实业部矿业司司长		
陈义瀛	1909	考选留美同学	江苏吴县	机械工程							M. E.	1914							康奈尔大学			拔伯克锅炉公司工程师		
曾昭权	1909	考选留美同学	湖南湘乡	电机			S. B.	1915											麻省理工学院					
杨永言	1909	考选留美同学	江苏嘉定																威斯康辛大学			已故		
裘昌运	1909	考选留美同学	江苏无锡	农	经济														Univ. of Wis.	普渡大学和哥伦比亚大学		已故		
贺懋庆	1909	考选留美同学	江苏丹阳	造船工程			S. B.	1914											麻省理工学院			杭州市政府秘书长		
卢景泰	1909	考选留美同学	广东顺德	道路工程							M. A.	1915							哥伦比亚大学			北平平绥路路局工务处服务		
戴修鞠	1909	考选留美同学	湖南武陵	机械			B. S.	1915											麻省理工学院					
戴济	1909	考选留美同学	江苏吴江	工程化学			A. B.	1914											Maine Univ.			上海爱多亚路117号元丰公司服务		
谢兆基	1909	考选留美同学	浙江吴兴	化学工程							M. E.	1914							哥伦比亚大学					
魏文彬	1909	考选留美同学	河北密云	财政											Ph. D.	1915			哥伦比亚大学					
卞煦堃	1909	考选留美同学	广东番禺	文学	新闻学		文学学	1914	新闻	1916									普林斯顿大学	哥伦比亚大学		已故		
严家驹	1909	考选留美同学	福建闽侯	数理			B. S.	1913			M. S.	1914							伊利诺伊大学	哈佛大学		已故		





- 通过与目前已开放的各类近现代报刊数据库（如上海图书馆的“晚清、民国期刊全文数据库”）、现代出版物数据库（读秀、超星等）等，手动、逐一的检索每位留学生，能够连接上大量的职业信息。
- 到目前，留美生的平均职业属已经从0.6个上升到了7个左右，这为研究清华留美生的职业发展及其与中国近代化的关系等问题提供了可能。



■ 除此而外，我们还利用中国第二历史档案馆馆藏留学生信息、日华学会统计的留日学生名簿、兴亚院统计的《日本留学中华民国人名调》、中国留法比瑞同学会统计的《中国留法比瑞同学会同学录》，以及袁同礼组织统计的《中国留美同学博士论文目录》、《中国留英同学博士论文目录》和《中国留欧大陆各国同学博士论文目录》等包含欧美留学博士生信息等多种史料，构建了留学生信息库。

■ 该数据库目前共计**38729**条，变量包括姓名、性别、年龄、籍贯、国内学校、留学年份、留学国家、国外大学、所学科目、归国职业等。数据库中统计的学生留学年主要集中于**1872至1949**年间，留学国家主要包括日本、美国、英德法比瑞等西欧国家。



第五讲

示例：从微观数据到宏观历史的 中国知识阶层来源研究





民国大学生数据库和民国海外大学生数据库



○在2010年，梁晨首次提出准备做一个有关二十世纪上半叶的大学生来源研究，十年来，李-康团队创建了“民国大学生数据库”，包含了当今近一半的双一流大学，目前包含来自民国时期教育部所认可的29所大学的131,105个学生的160,253条记录，“民国海外大学生数据库”包含毕业于海外的40,000名留学生中的20,000名

Universities 大学

- 115731 名学生
- 55所大学中的19所

Institutes 学院

- 14482 名学生
- 75所学院中的9所

Specialized Schools 专科学校

- 892 名学生
- 1所专科学校



民国大学生数据库的变量



- 80%的中国大学生数据库基于学生登记（注册）卡建立，学生登记（注册）卡虽然会随着学校的不同而有所变化，但一般包含以下几点：

家庭和社会背景

- 父母和保证人的姓名、职业、居住地址及家庭成员

教育经历

- 过去及现在的学校及专业

个人信息

- 姓名、性别、籍贯、种族



民国大学生量化数据库学籍卡



40-3830

Li Cheng Tao

國立西南聯合大學學生註冊片

學號：342011 姓名：李成濤 別號： 學籍：

籍貫：上海 縣市 年齡：民國34年19歲(出生於15年11月)性別：男

入校年月：34年現在28學院第2年級曾否入黨： 已婚未：未

經過學校：校名：15412大學 將來志願 校址：上海

家長姓名：李望雲 關係：父子 職業： 家庭通訊處：上海

保證人姓名：李望雲 關係：父子 職業： 保證人通訊處：上海

學生通訊處：校內： 校外：

備註

卅四年度錄取名單新生 34年錄取名單=正式生 19



姓名(中文)李國鼎 (英文拼音) Lee Kwoon Ting 現年二十歲 生於光緒十九年三月三日 原籍安徽 縣人現居本

本縣永久通信地址 盱眙老子山丁乾豐號

父或保護人姓名 李相臣 職業 學界

及其通信處 同上

成績報告單應寄何處 盱眙老子山丁乾豐號

擔負經濟者何人 父 通信處何在 同上

曾否結婚 已婚 如曾結婚 現有子女幾人 無

信宗教否 知信 現信何宗教(孔釋道耶基督等) 知信基督教曾否受洗

受洗之日 年 月 日 關於何教會(中文) (英文)

教會地點安在 省 縣

曾肄業何處中學(校名) 江蘇第六師範 (地點) 清江浦

肄業時期自中華民國七年秋至中華民國九年秋 季止 共計若干學期 五學期

又曾在 學校肄業(地點)

時期自中華民國 年 季至中華民國 年 季止 共計若干學期

在何中學畢業(校名) 成美中學 (地址) 南京講堂街

畢業時期 中華民國十二年秋 季 該校為官立抑或私立或教會所設立 教會

中學畢業後曾否肄業其他專門或大學(校名) (地點)

肄業時期自中華民國 年 季至中華民國 年 季止 共計若干學期

又曾入 學校肄業(地點)

肄業時期自中華民國 年 季至中華民國 年 季止 共計若干學期

中學畢業後有無何種職業經驗 如有之係何種經驗

曾有幾年

考入本校後 志願肄業何科(文理農林商醫教育等) 文科

在本校大學畢業後從事何事 教育

保證書 (GUARANTOR'S CERTIFICATE)

保證人須以中文詳細填明以下諸條
The guarantor should fill in carefully in Chinese the spaces below

周寄高 現年二十六歲 安徽省無為縣人今願保學生李國鼎 肄業貴校
所有該生在校一切行為 概由保證人完全負責 此證即致金陵大學查照



民国高等教育的五个特点，1912-195



1. 学生和教员的选拔体系：选贤举能
2. International 15%学生45%教员有国际学位
3. Female 女性学生占20%学位
4. STEM dominant 50-75%学生选读与科技相关科目
5. New Social and Spatial Origins 新的社会与地理背景



I. 民国高等教育的政策目标与选贤举能的原则



- 科举制被废除后，民国新知识分子出现，这些知识分子继承了一个基本的信念和强大的目标
- 这个目标是要选拔和培养中国最优秀、最聪明的人，并以此恢复中国作为世界领先国家和社会的历史地位
- 这个基本的信念指的是，想要实现这样一个目标必须建立一个精英国家和社会
 1. Who were selected by objective measure of their cognitive ability, knowledge, and values 选贤举能
 2. Who were among the most capable and able in China at large 有教无类
- 大学的主要目的是教学，而不是做研究



I. 留学考试体系的建立



1909年，游美学务处通过考试录取了47名第一批直接留美生

民国初年，各级学校自设招生机构，命题、考试、阅卷、录取均由学校自行办理

1917年，教育部规定招生时，学生需要提供中学毕业证书，且在入学时接受选拔考试

1924年，教育部规定国立大学由入学考试成绩决定录取

1933年，南京国民政府教育部开始通过“比例招生法”介入大学招生

1935年，教育部进一步推行实际名额控制

1938年，教育部推行国立各院校统一考试，并成立统一招生委员会

1939年，统一招生委员会实现“制定统一命题”

1942年，因战事原因，恢复各大学自行招生

1946年，教育部公费和自费留学，均需填报考试报名书，参加统一留学考试，获教育部许可后，方可出国留学



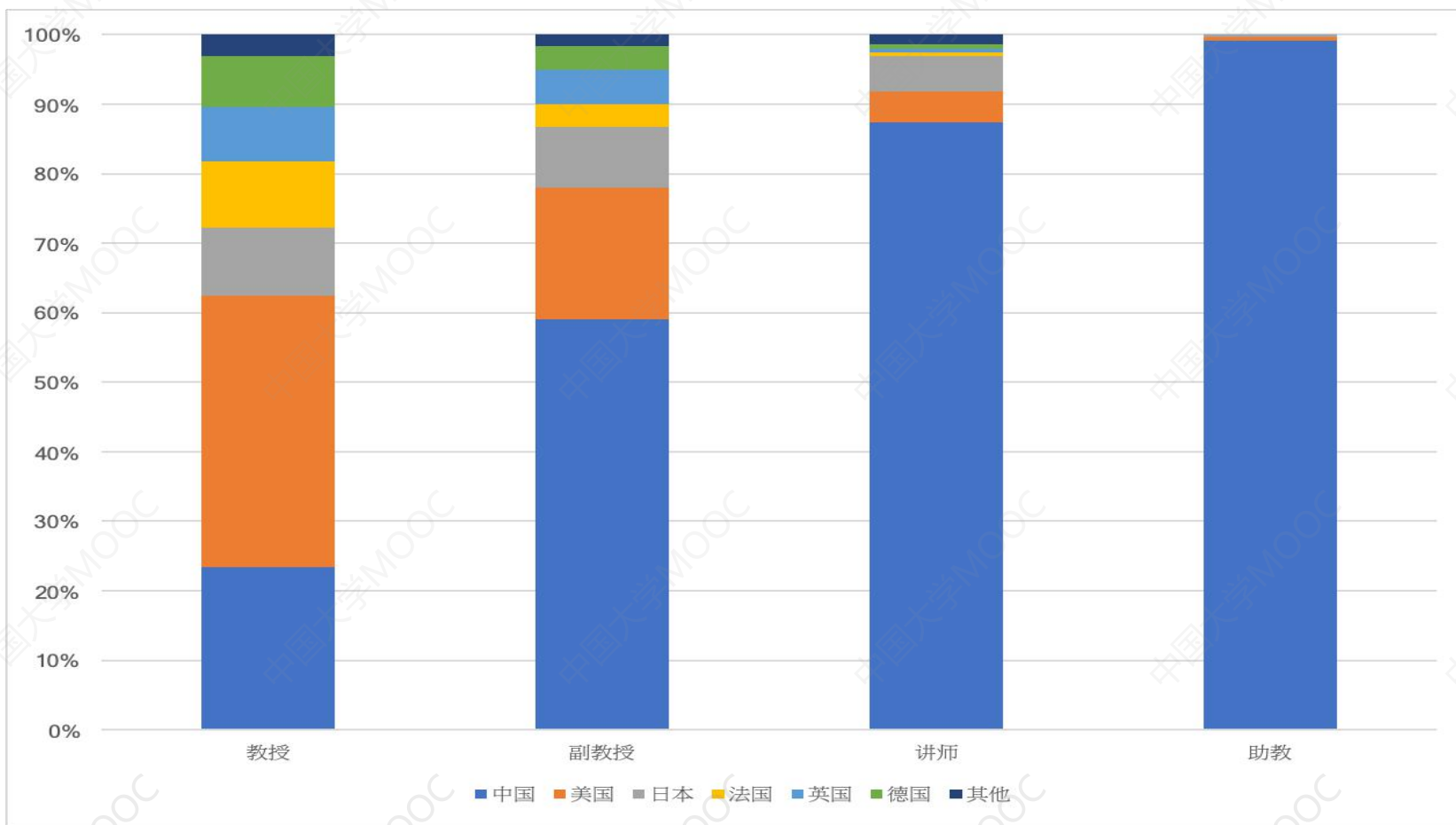
II. 留学教育，留学标准和领导力量



- 在20世纪上半叶，多达10%的中国大学生毕业于日本、欧洲和美国的国外大学，而不是中国的大学。
- 另有5%的大学生，尤其是从中国那些很有名望的大学里毕业的学生，他们继续在国外接受了进一步的硕博高等教育，其中5000人获得博士学位，超过10000的人获得硕士或专业硕士。还有更多的人出国留学，但没有参加正式的学位课程。
- 1/3至1/2的留学生通过考试选拔，或由中央政府出资，或由地方出资，或由外国政府或非政府组织资助（这一资助集中20世纪初期）
- 在民国大学所有教员中，这些国外大学的毕业生占比接近1/2，且正教授拥有海外求学经历的比例超过3/4。此外，根据CED-ROC：在总数为14997名工程师中有2276名（占比15%）是国外大学毕业生。

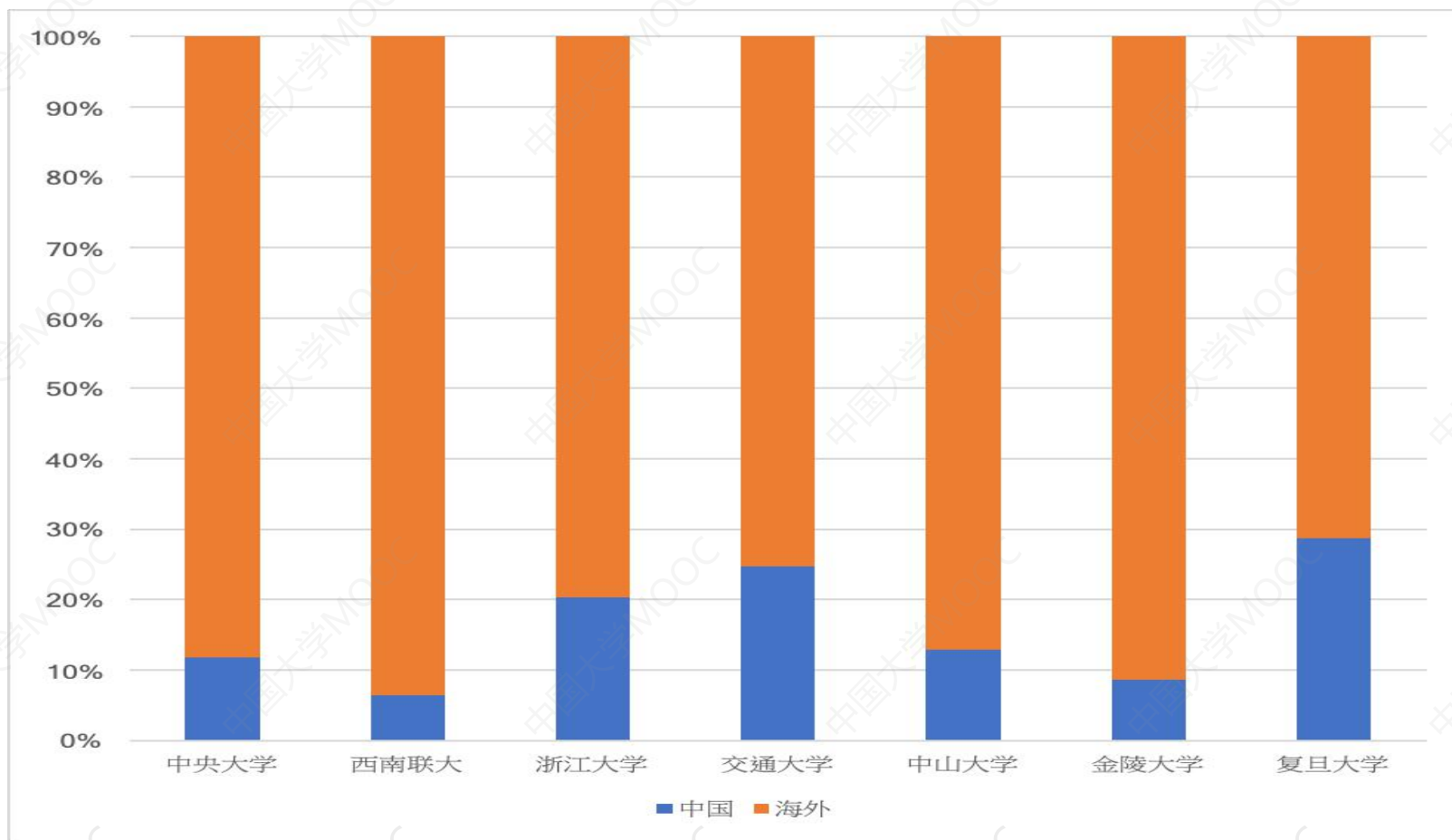


II. CUED-ROC: 受过国外教育的大学教员





II. 受过国外教育的国内顶尖大学正教授





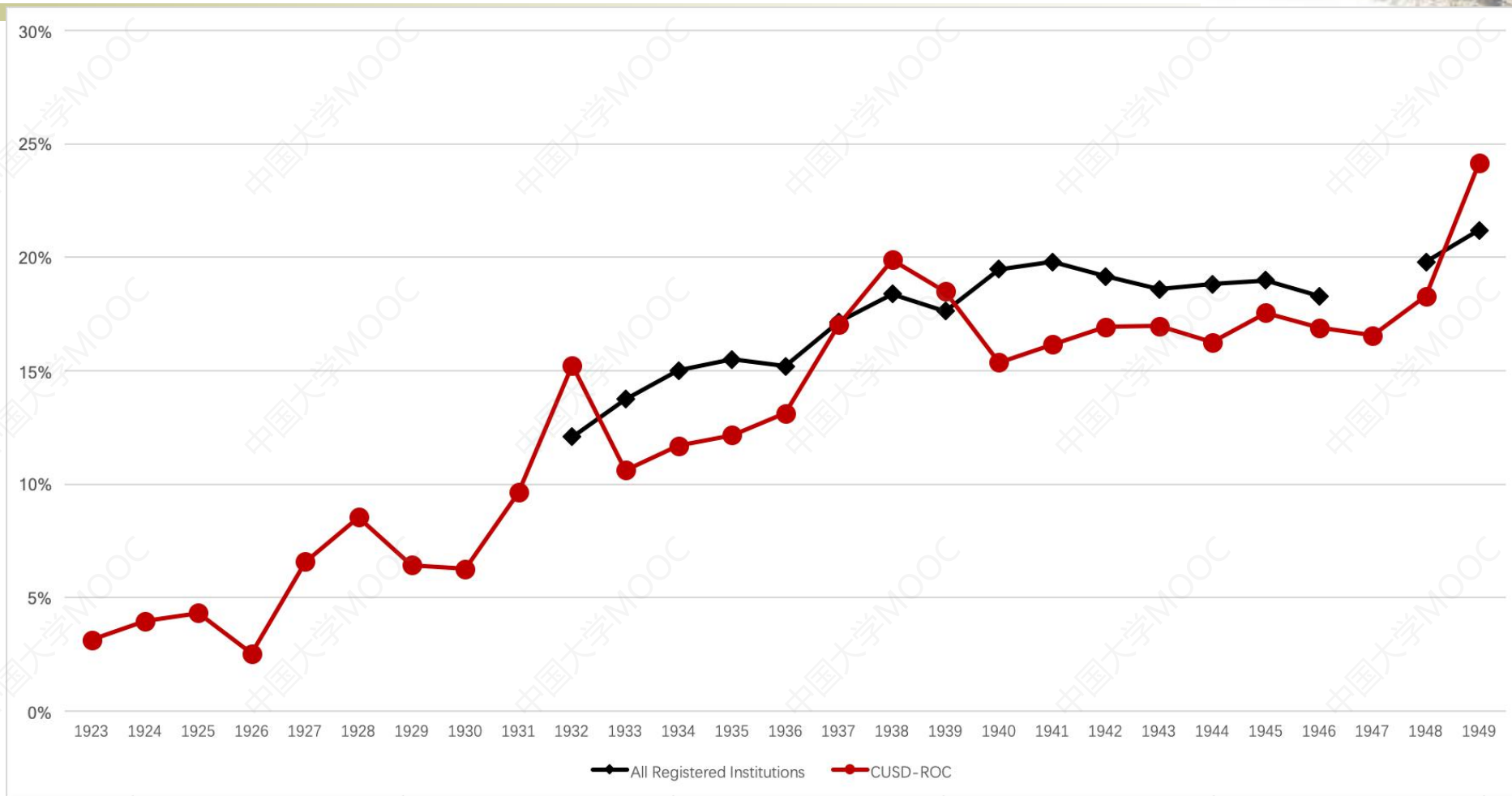
III. 高等教育中的女性与男女同校



- 尽管数千年来，中国一直有系统地排斥女性接受正规教育和带薪就业，但性别障碍与高等教育获得的关系始终变化无常。到20世纪初，性别障碍迅速且几乎完全地崩溃了。
- 大同大学于1916年、岭南大学于1918年、北京大学于1920年分别成为第一所接收女性学生的私立、教会、公立大学。
- 到1931年，中国41所注册大学中，只有3所是男女同校
- 与其他地方不同，这些男女同校的中国大学并没有单独的女性入学配额、单独的女子学院、单独的校园、单独的师资或专业。
- 女学生和男学生一样，是根据考试资格录取的，并接受与男学生相同的教育。



III.女大学生占比图



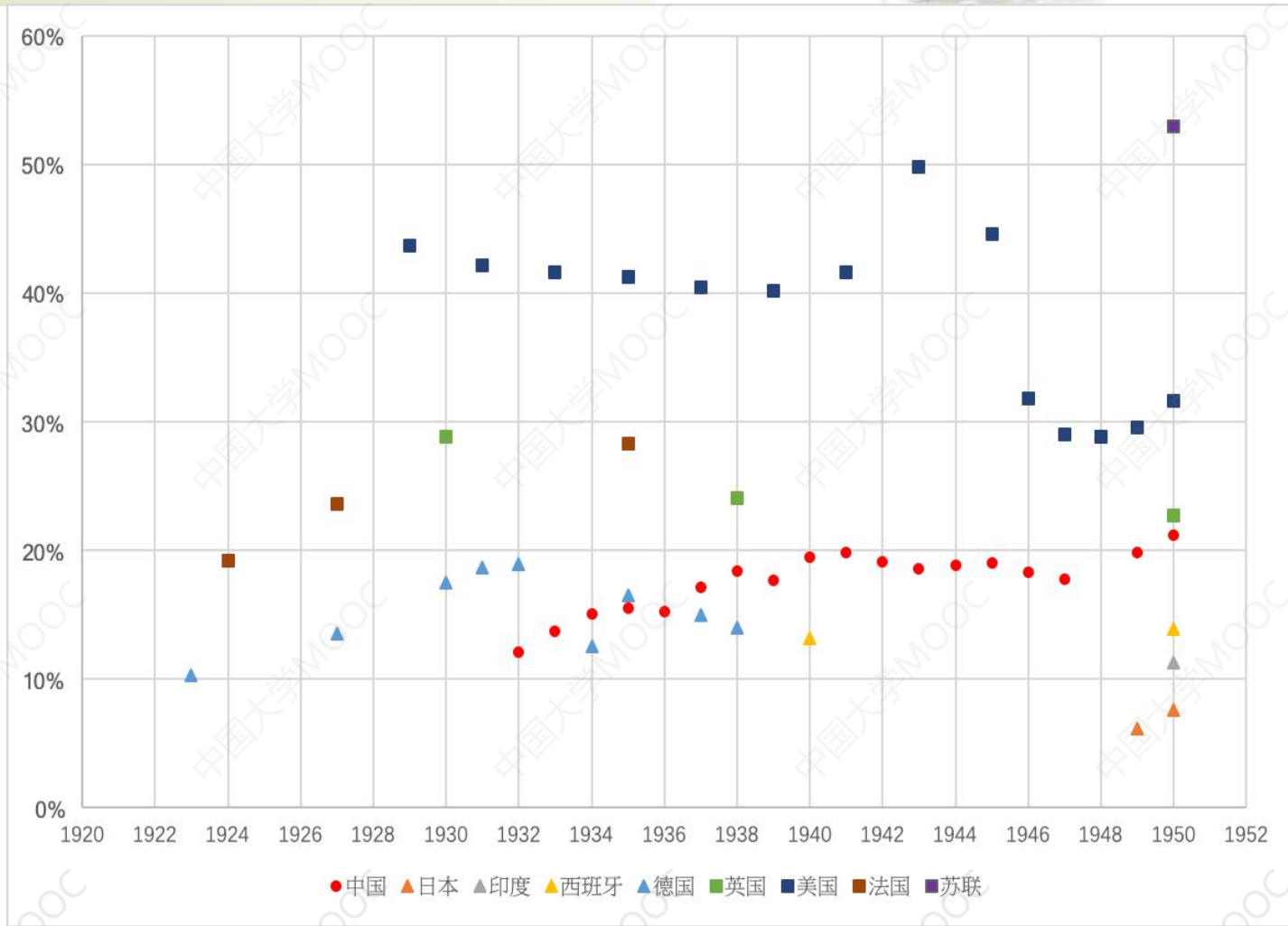
数据来源：全国1923-1946年数据来自《第二次中国教育年鉴》，1947年和1949年来自《中国教育成就：统计资料，1949-1983》。



III. 女大学生入学率的国际比较



- 总体而言，中华民国时期，女大学生占比约20%上下。
 - 然而，在人口众多的国家中，只有美国、英国和法国的女性大学入学率更高。
 - 中国女性的高等教育比例远远超过日本，与德国和西班牙等欧洲发达国家持平，甚至略高于后者。





IV. ROC：关注理、工、农、医

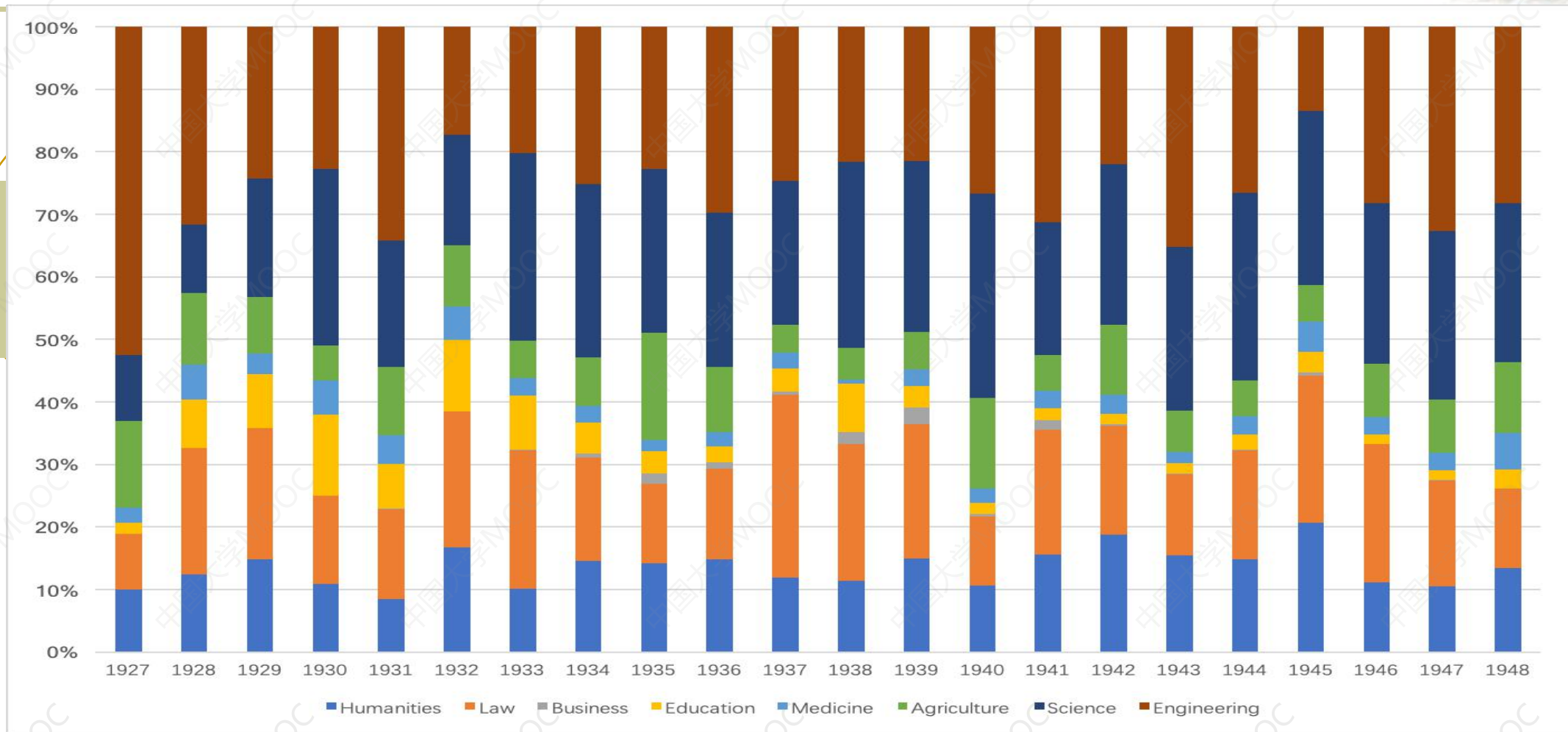


- 1931年，中华民国教育部正式将高等教育分为大学、研究所、专科学校三个层次，并设有人文、理学、法律、教育、农业、工程、商业、医学等八个学院。所有大学都必须包括至少三个学院，其中至少一个必须是理学、农学、医学或工程学。
- 1933年，中华民国教育部规定，任何一所高等教育研究所招收的人文科学学生不得超过理工科学生。
- 1935年，教育部进一步限制人文、法律、商科、教育等专业每年招收新生不得超过30人。
- 尽管这些规定对全国的影响尚不清楚，但根据中国最好的大学(1927年至1948年)的招生比例计算，理工农医占比几乎总是超过60%，在某些年份甚至增长到了75%。



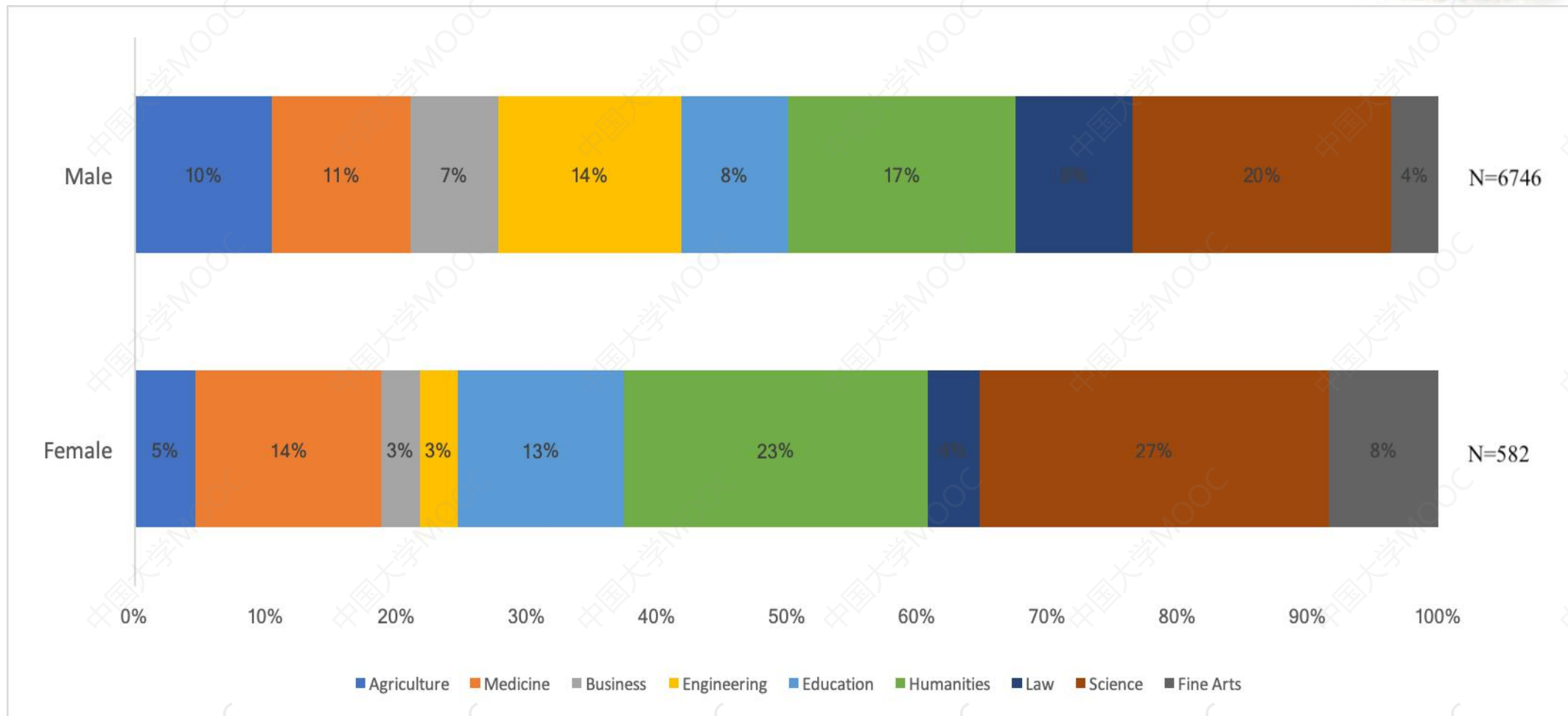
NANJING UNIVERSITY

IV. 1927-1948: 理工科在六所主要国立大学中占主导地位



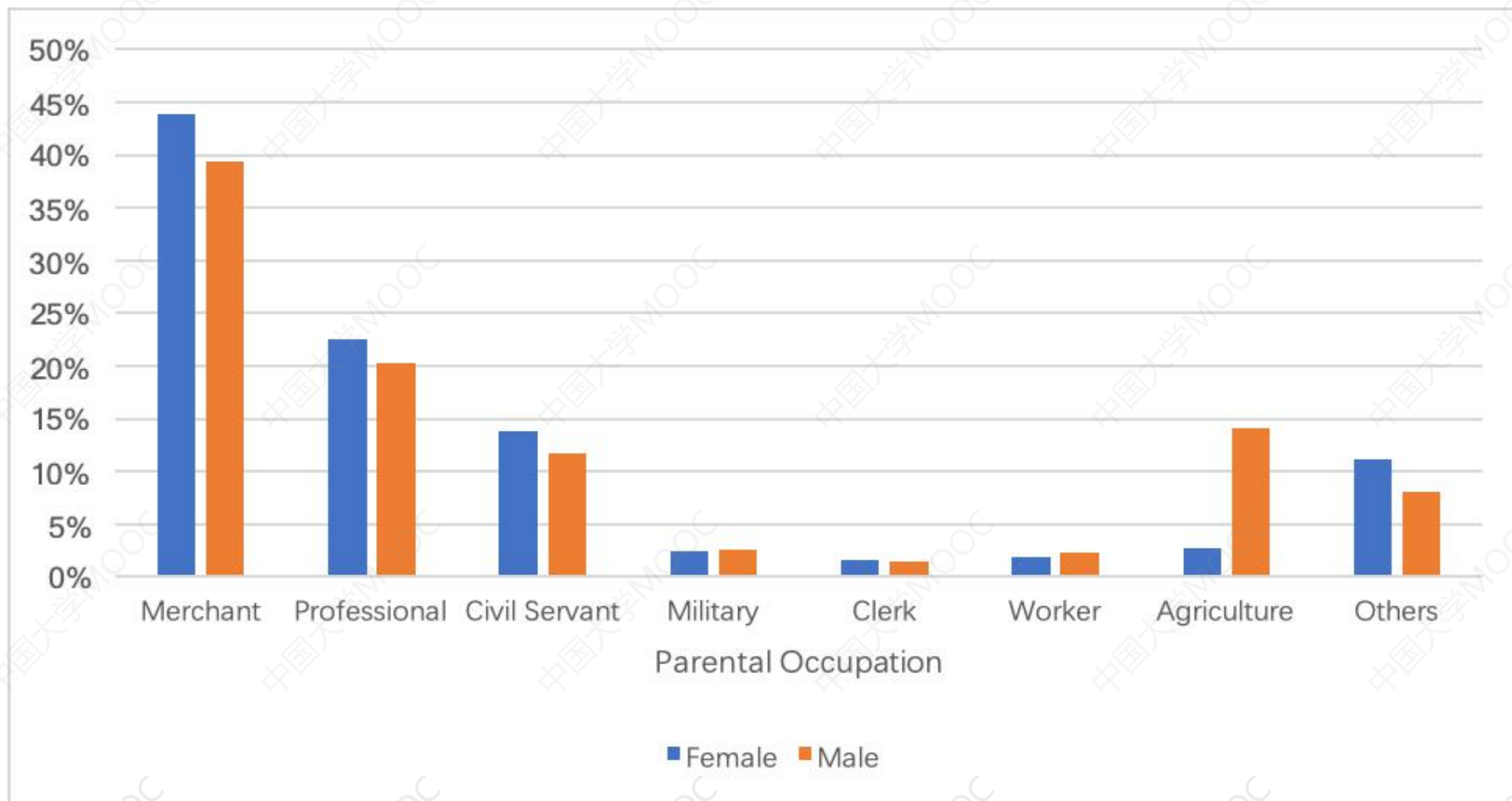


IV. 民国大学女教师理科出身比例高



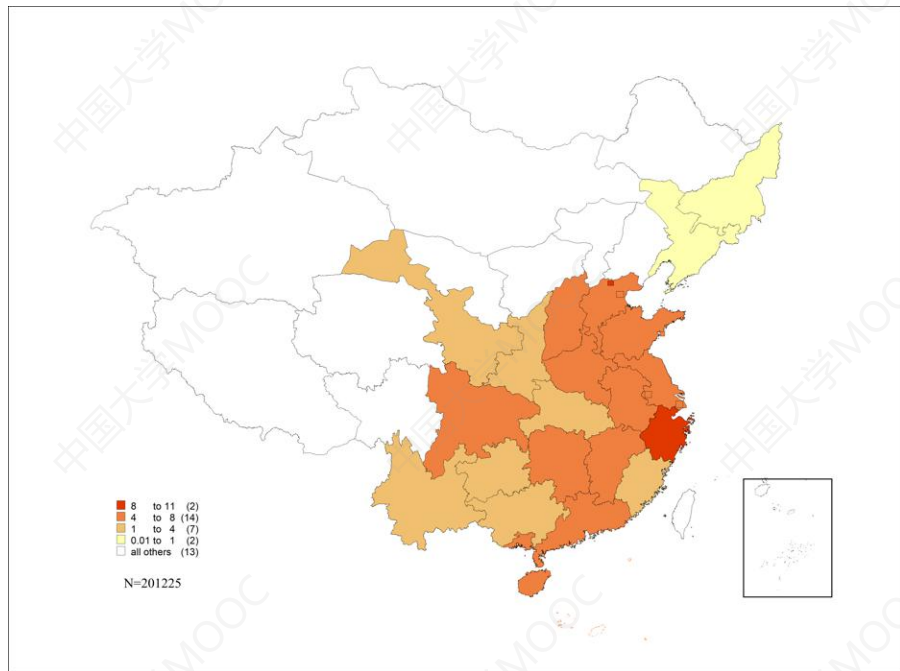


V. 学生多为商人及专业人士家庭出身：学生性别与家长职业

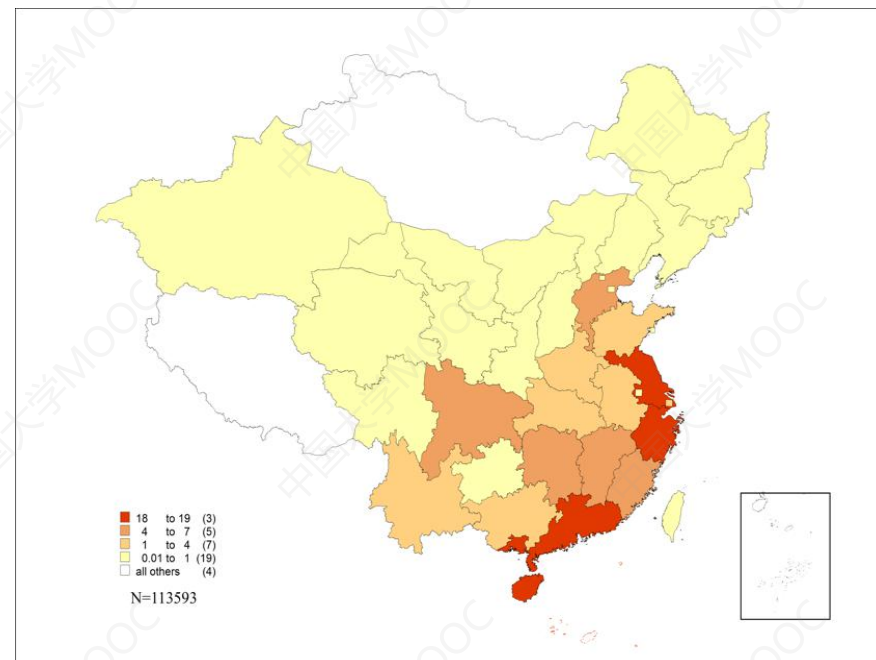




V. 清代科举精英与民国大学生的区别



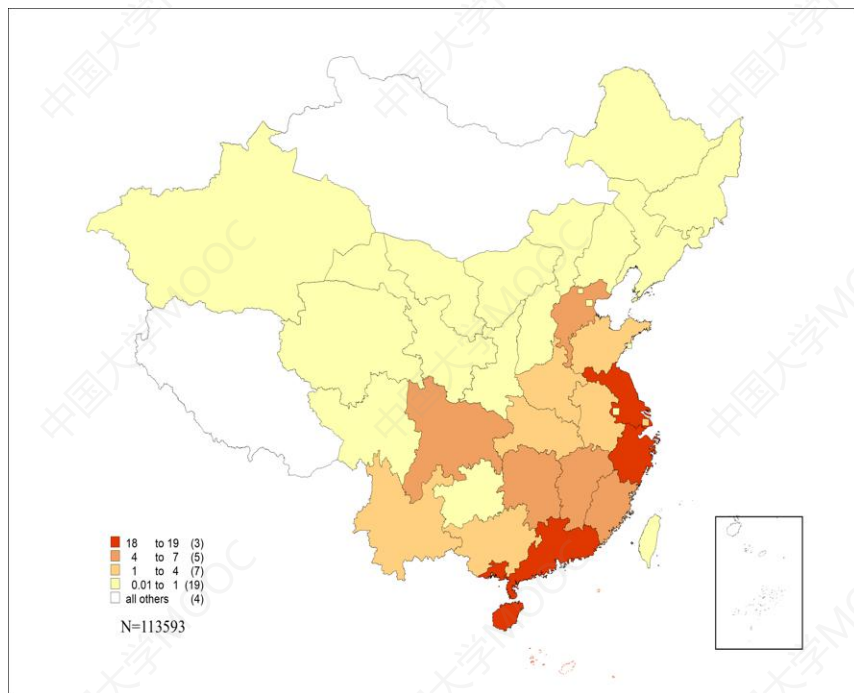
清代



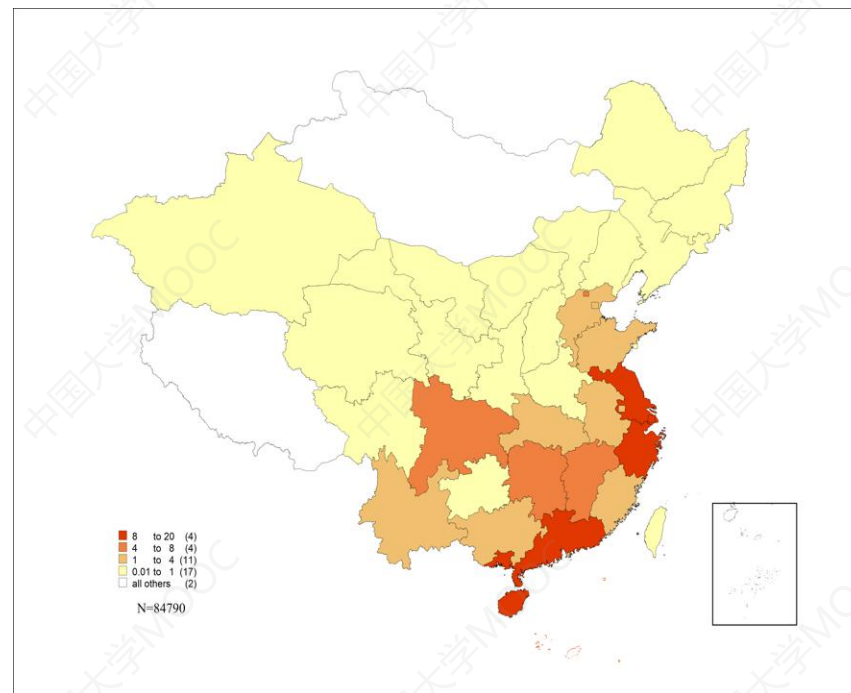
民国



V. 出生地及家庭住址:北京, 广州, 南京, 上海, 天津



5 Cities = 4 %



5 Cities = 36 %



中国的社会转型：1912-1952年



- 除了当代中国革命的胜利以外，民国时期另一个持久的成就是一个受过大学教育的中国社会的开始。
- 事实上，当代中国高等教育的许多显著特征，例如国际化、女性教育、理工科教育、城市化甚至地区性问题等，都可以追溯到这个时期。在辛亥革命之前，通过考试来选拔人才的做法备受关注；在中国共产党无产阶级革命之后，中国大学生的社会来源多样化以及中国研究型大学的最新发展更容易受人关注。
- 总之，通过客观考试进行选拔而产生的中国知识分子的可变性在20世纪绝大部分时间里“胜过”了政治及经济优势群体，且对我们理解中国当前的社会流动模式和社会分层有重要意义。