

用数据“写”历史——基于上海图书馆开放数据的数字人文阐释

李惠
南京农业大学
2026.7.7



目录

CONTENTS



南京農業大學
NANJING AGRICULTURAL UNIVERSITY

①

引言

②

研究对象

③

研究方法





典籍新生，数据里的文脉风华

上海图书馆开放数据竞赛自2016年起已成功举办10届，凝聚了一批数据创客团队，收获了大量优秀作品和创意。第十一届上海图书馆开放数据竞赛将继续为参赛者提供更加丰富和海量的历史人文数据。竞赛以“典籍新生，数据里的文脉风华”为主题，面向全球征集优秀应用、论文和AIGC作品。

➤ 创意论文赛道

- 作品形式：不少于8000字的学术性论文
- 鼓励**多学科交叉视角**，开展关于**数字人文**、公共服务及文化创新等领域的思辨与探索
- 典籍：国家重要文献，亦统称各种典册、书籍（大辞海）
- 新生：**重新获得生命**（大辞海）
- 文脉：文章的脉络，文章的布局和条理。亦指学术的流派或思维的线索（现代汉语词典）

➤ 从沉睡典籍到可感可触的文化现场

➤ 在时空经纬中探寻人物与历史

➤ 焕发非遗新生，演绎江南风韵、海派之美

➤ 让红色记忆焕发时代光芒

➤ 从公共行为洞见城市的文化脉动



利用**人工智能**、大数据等前沿技术，对海量**多模态**数据进行深度挖掘、关联与重构，实现文化资源的活化利用、创新表达与知识服务

The figure displays a collection of cultural heritage data points, organized into three main categories represented by colored boxes: Text (Text), Image, and Audio/Video.

Text Category:

- 家谱: 10.5万余种
- 盛宣怀档案: 15万余种
- 古籍循证数据: 130万余种
- 人名规范库: 135万余人
- 上海联编中心书目数据: 150万余种
- 红色文献书目数据: 8500余种
- 电影库数据: 4000余种
- 上海优秀历史建筑: 1086处
- 上海市不可移动文物: 3452处
- 红色旅游景点: 1199处
- 上海地名志: 10000余个
- 上海市历史文化事件: 1.5万余条
- 古籍刻工名录: 5800余人
- 上海年华机构名录: 2.1万余个
- 碑帖: 45件
- 其它基础知识库: 姓氏(600余个)、收藏机构名录(2000余个)、地理名词表(3800余个)、中国历史名城(2000余个)、中国名画(2000余幅)、中国古建筑(2000余座)、中国园林(2000余处)、中国石窟寺(2000余处)、中国古钱币(2000余种)、中国古陶瓷(2000余种)、中国古书画(2000余种)、中国古书籍(2000余种)、中国古地图(2000余幅)、中国古乐器(2000余种)、中国古服饰(2000余种)、中国古家具(2000余种)、中国古玩具(2000余种)、中国古文具(2000余种)、中国古兵器(2000余种)、中国古医书(2000余种)、中国古食谱(2000余种)、中国古药方(2000余种)、中国古秘方(2000余种)、中国古偏方(2000余种)、中国古验方(2000余种)、中国古秘方(2000余种)、中国古偏方(2000余种)、中国古验方(2000余种)
- 《近代上海城市文化》专题片: 12集
- 中文期刊论文数据: 98万余条
- 用户与借阅数据: 书目数据354万余条、读者数据10万余条、以上两者所产生的流通数据623万余条
- 东馆预约数据: 183万条
- 家谱原文: 14625种
- 古籍原文: 10802种
- 近代图书原文: 620种
- 借阅数据(2025): 69538条
- 戏单数据: 897条

Image Category:

Audio/Video Category:

➤ 家谱

- 收藏近4万种38万册家谱，涵盖567个姓氏，是全球收藏中国家谱原件最多的公藏机构
- 中国家谱知识服务平台、“君从何处来”动态迁徙展示、馆藏家风家训主题展、廉洁文化主题文献展、“阅人·阅城·阅世界”虚拟数字展览等
- 家族树（向帆、朱舜山）





李氏宗谱一百二十卷首二卷末一卷（江苏省常州市金坛市）📍 江苏省常州市

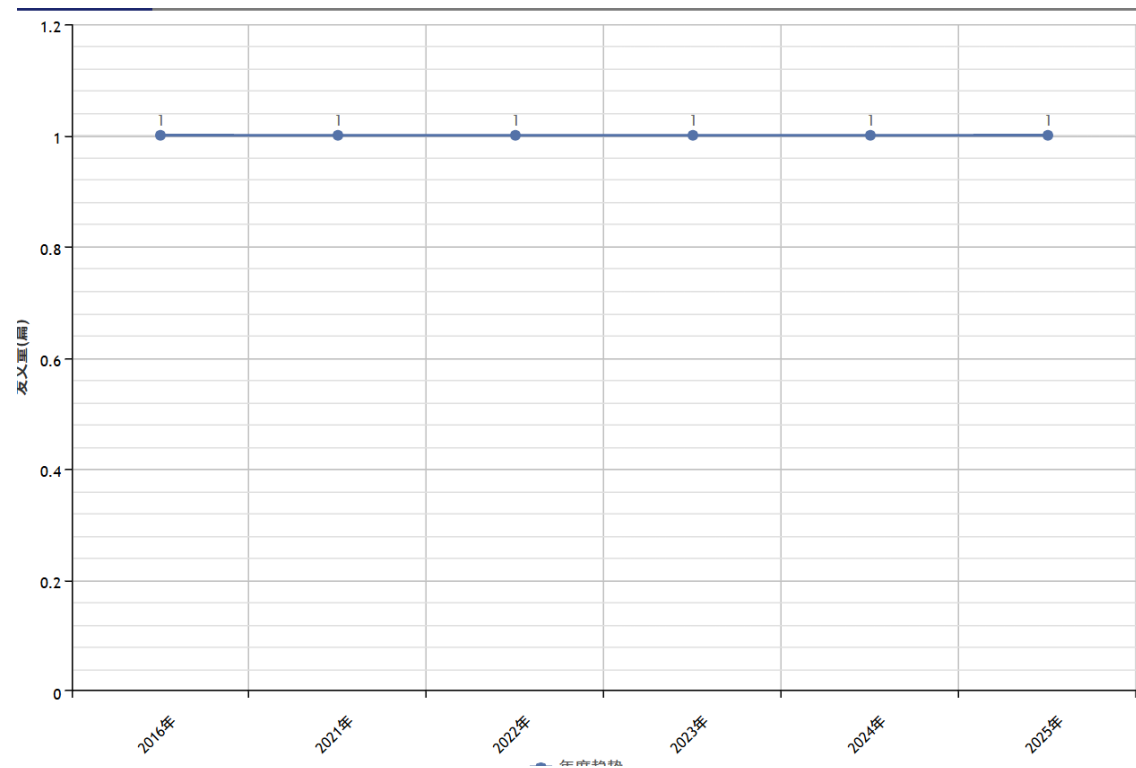
[查看](#)

[正 题 名]：李氏宗谱
[卷 数]：一百二十卷首二卷末一卷
[责 任 者]：李秉阳(纂修)
[堂 号]：见龙堂
[撰修时间]：清宣统元年(1909)
[装 订 项]：线装
[附 注 项]：书名据版心题。五修本。
[摘 要]：始祖宏义，字大能，唐季摄金坛，遂由晋陵溇湖西之湟里迁居金坛西冈。卷首上载序、凡例等，卷首下传赞，卷一世源图、总系，卷二派世系，卷一百一十五至一百二十二依次载诰敕、祠堂志、墓志、行实志、文翰志、履历志、分迁、古迹，卷末载颁谱字号、会谱定例等。谱收蔡元衡、王十朋等撰题序。书名据版心题。五修本。
[版 本]：木活字本
[数 量]：80册

馆藏地

馆藏机构(收藏者)	地址	全文
吉林大学图书馆	吉林省 长春市 南关区 吉林大学南岭校区校园中心	
上海图书馆	上海市 徐汇区 淮海中路1555号	DOI为STJP008958（索书号：923483-562） 

- 关键词检索



- 全文检索

全文 ▾

家谱 开发 挖掘 上海图书馆

×

Q

问AI

结果中检索

学术期刊

学位论文

图书

会议

报纸

年鉴

专利

标准

法律法规

检索范围: 学术期刊 全文: 家谱 开发 挖掘 上海图书馆 主题定制 检索历史 共找到 1,661 条结果

➤ 家谱

- 迁徙事件
- 社会流动性
- 命名文化
- 记载特征

A B C D E F G H J K L M N O P Q R S T W X Y Z

李 li

4730 种

先祖名人8704
始于1764

刘 liu

4340 种

先祖名人8498
始于1833

林 lin

1677 种

先祖名人2794
始于1644

罗 luo

985 种

先祖名人1888
始于1368

吕 lu

619 种

先祖名人1001
始于1368

梁 liang

521 种

先祖名人908
始于1368

廖 liao

499 种

先祖名人890
始于1368

卢 lu

474 种

先祖名人917
始于1368

陆 lu

438 种

先祖名人833
始于1239

赖 lai

252 种

先祖名人576

雷 lei

261 种

先祖名人493

楼 lou

225 种

先祖名人390

李

东汉《风俗通义》收。商朝末年有理征，一作李征，传说春秋时代哲学家老子姓李氏名耳。战国时有魏文侯师李悝（kuī），著《法经》。《急就章》列为汉代常见姓氏一。汉代有李广，成纪人，号飞将军。唐代有李白，字太白，大诗人，号诗仙。唐朝皇族姓李氏。《中国人名大辞典》收录李氏1806例，李氏为中国人数量最多、分布最广的三大姓氏（李王张）之一。宋《百家姓》列为第004。

查看李氏家谱

先祖名人

李开 李子基 李伯基 李荣诸 李章晟 换一批

家谱

醴陵罗田妙泉李氏四修族谱三十五卷（湖南省株洲市醴陵市）

先祖名人

李开 / 始迁祖

醴陵罗田妙泉李氏四修族谱三十五卷（湖南省株洲市醴陵市）

责任者：李晓岚

摘要：始迁祖开，字荣松，号光祖，明代人。存卷一至二十四、二十六、二十八至三十一、三十三至三十五。书名据版心、书签题。

堂号：惇裕堂

□ 需要参考多方资料，综合考证

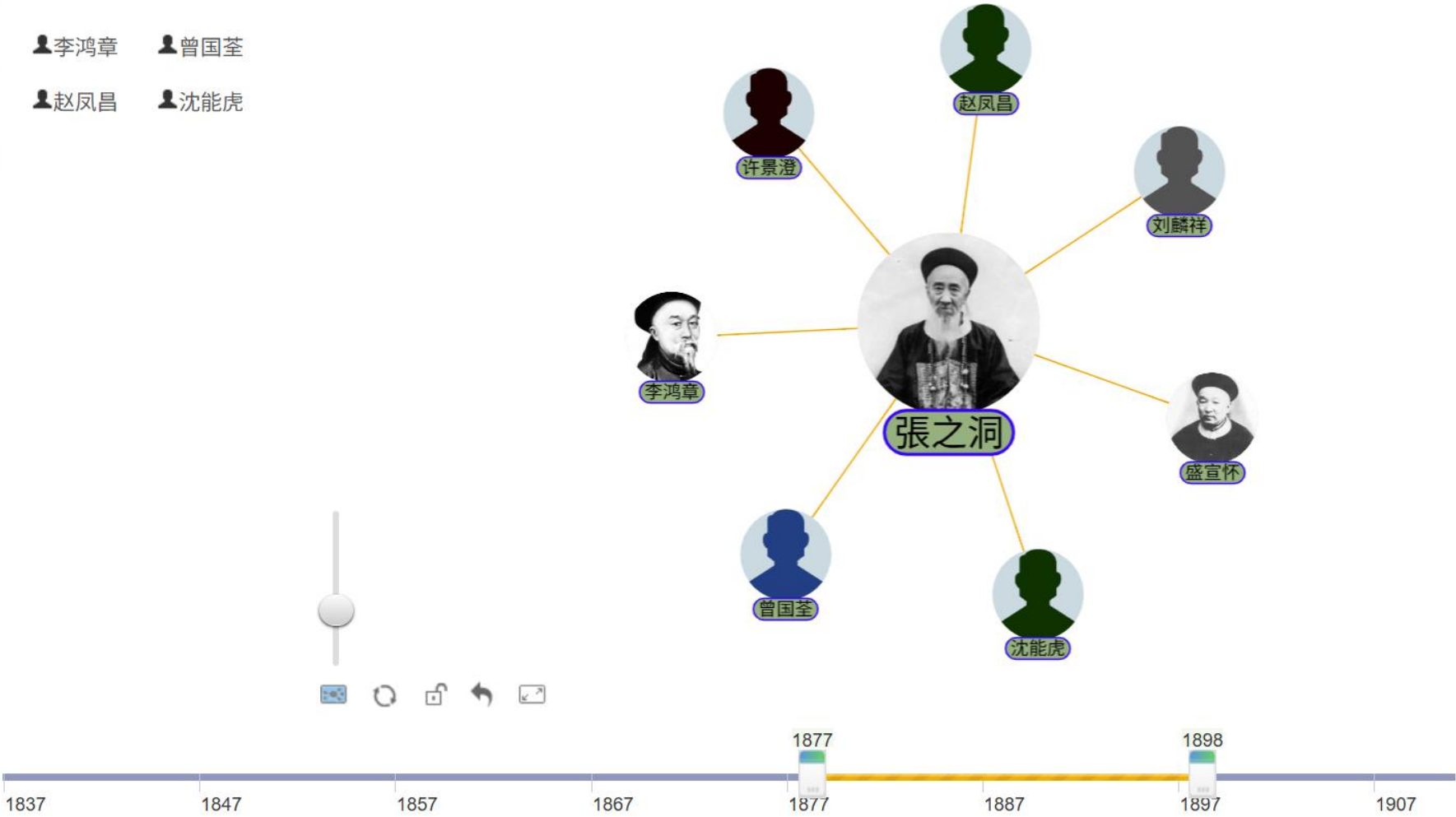
➤ 盛宣怀档案

- 15.7万件日记、文稿、信札、电报、账册、电文、合同、章程，是盛宣怀家族自1850至1936年间的记录，内容涉及政治、经济、社会、军事、外交、金融、贸易、教育各方面，被称为中国私人档案第一藏，是研究中国近代史的第一手史料宝库
- 上海图书馆盛宣怀档案知识库、众包平台

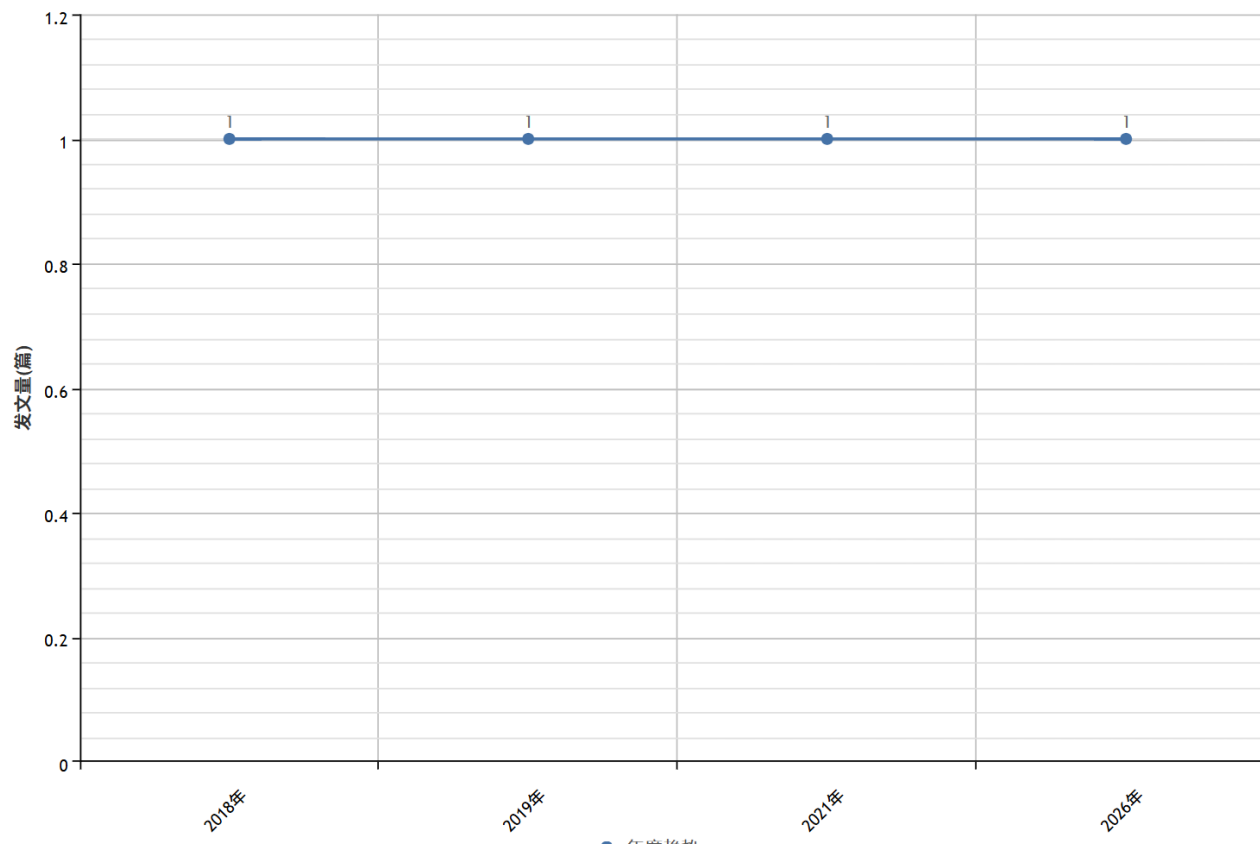


相关人物

- 盛宣怀
- 李鸿章
- 曾国荃
- 刘麟祥
- 赵凤昌
- 沈能虎
- 许景澄



- 关键词检索



- 全文检索

全文 | 盛宣怀 档案 挖掘

× Q 问AI 结果中检索

学术期刊 | 学位论文 | 图书 | 会议 | 报纸 | 年鉴 | 专利 | 标准 | 法律法规

检索范围: 学术期刊 全文: 盛宣怀 档案 挖掘 主题定制 检索历史

共找到 651 条结果

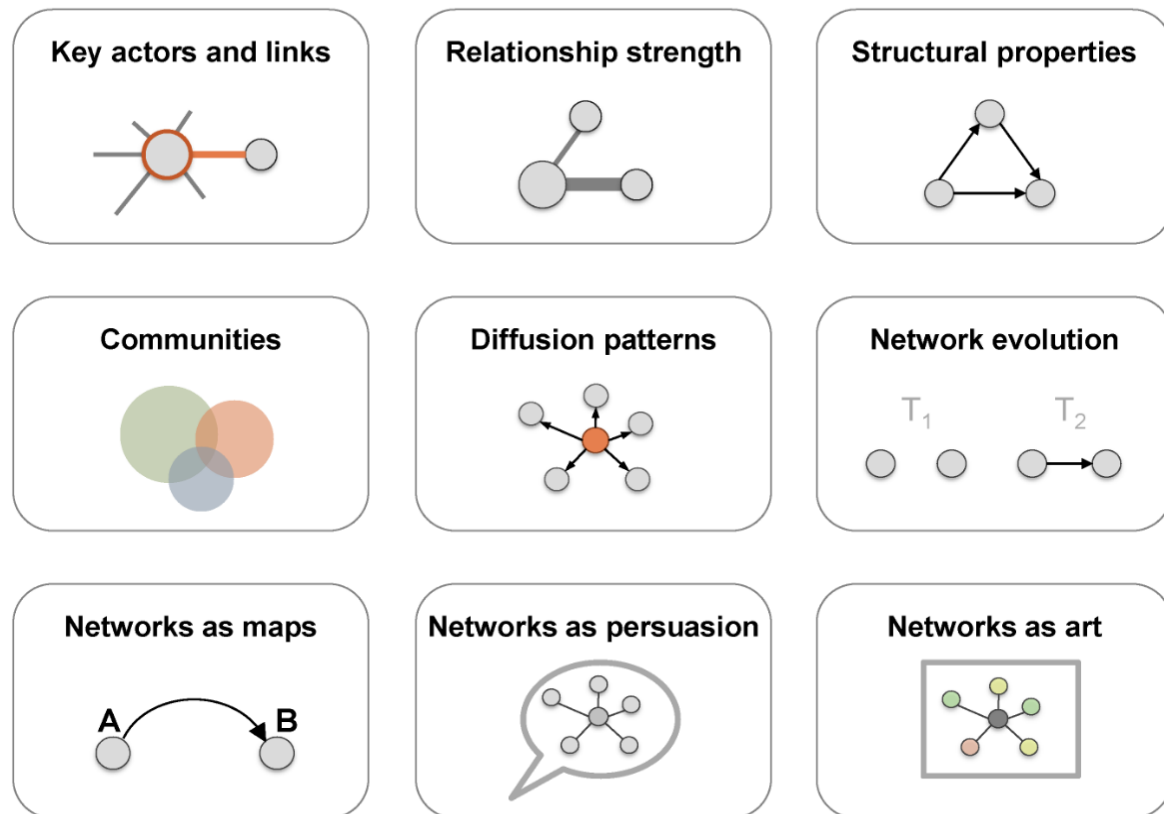
➤ 盛宣怀档案

- 朋友圈
- 历史事件
- 早期企业的雏形
- 记载特征

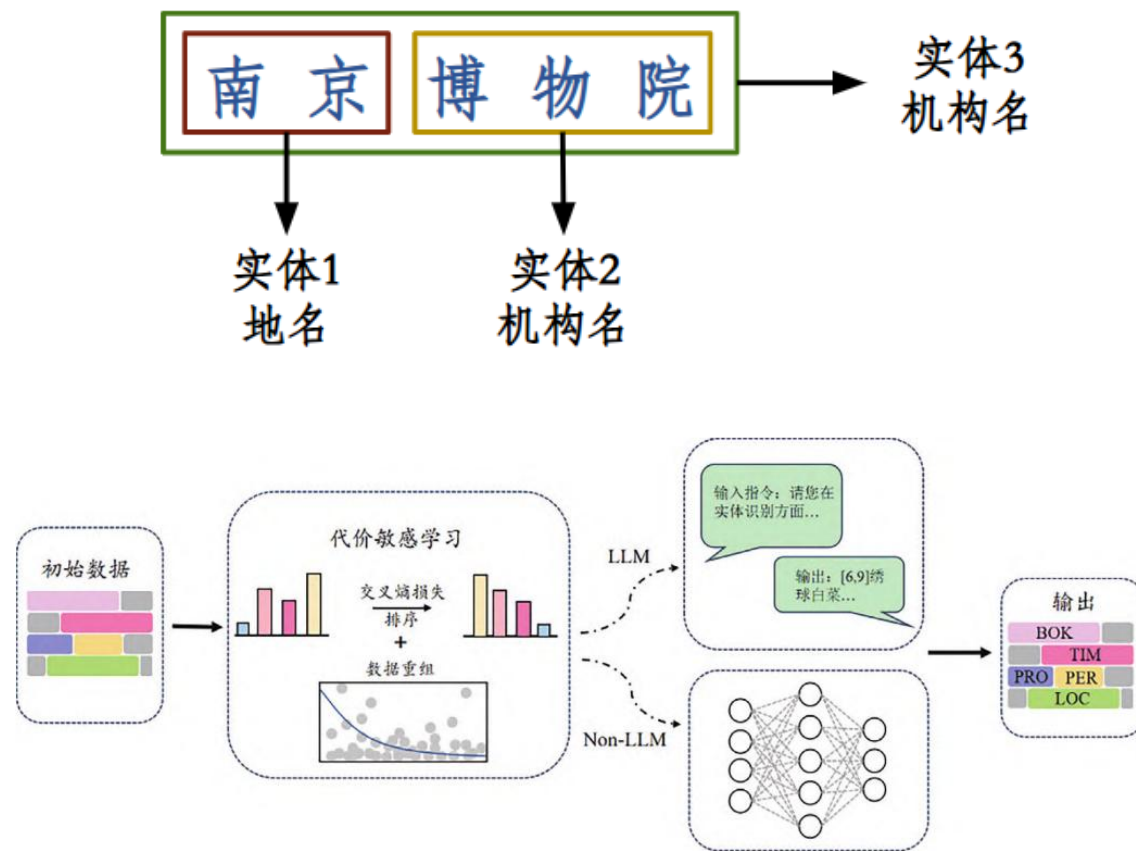
□ 需要参考多方资料，综合考证



- 不只是一种理论，也不只是一种方法 (Li, 2022)
- 使用网络和图的相关理论和方法，探索社会关系的具体实践，既关注**个体**行为，也关注**网络整体拓扑结构**反映的关系模式
- 对社会关系结构及其属性加以分析的一套规范和方法
- **网络度量方法**：网络中心性、连通分量、社群检测、传播路径、时空分析等



- 实体一般指存在且具有明确语义的对象或概念，如人物、地理位置、组织机构、时间等
- **实体识别**：自动判断文本中某一字符串，是否代表一个实体并确定其具体类别
- 参数微调、提示学习等 (Lin, 2026)
- 知识图谱、信息检索等任务的基础



- 传统的检索方法或问答系统，主要为基于原文的抽取式回答，由“检索+阅读理解”构成，生成能力较为欠缺，难以满足多实体、多关系的复杂知识问答需求（Li, 2026）
- **检索增强生成**（Retrieval-Augmented Generation, RAG）通过将外部知识源引入大语言模型的生成过程，在一定程度上缓解了大模型在事实性问答中的幻觉问题
- Naïve RAG, GraphRAG, Hybrid RAG, Agentic RAG等



- 用数字技术表现文献**背后的故事**
- “虽然知道阳光会影响植物的生长，但是我们从来没有确切地看过它是如何影响植物生长的（*We know that sunshine influences growth, but it's our first time truly understanding exactly how it influences*）（向帆专访，澎湃，2024）
- 与博物馆里的碑帖展览着重于展示文物之美不同，图书馆更侧重于向大众展示**历史文献的价值**，解读**历史文献背后的故事**，尝试“让书写在古籍里的文字活起来”，最终指向的还是**阅读文献本身**（黄薇，2023）



主要参考文献

- 向帆专访：把可视化作为方法. https://m.thepaper.cn/newsDetail_forward_29692712 [2026-07-01]
- “最古老”的历史文献插上“最科技”的翅膀，古籍变身虚拟世界、游戏空间. <https://www.jfdaily.com/wx/detail.do?id=665318> [2026-07-01]
- 上海图书馆开放数据竞赛. <https://opendata.library.sh.cn/> [2026-07-01]
- 珍档秘史—盛宣怀档案知识库. <https://sd.library.sh.cn/sd/home/index> [2026-07-01]
- 上海图书馆家谱馆. <https://www.library.sh.cn/special/GenealogyLibrary/> [2026-07-01]
- 大辞海. <http://www.dacihai.com.cn/> [2026-07-01]
- 现代汉语词典. <https://cd.hwxnet.com/> [2026-07-01]





Thank You

