



中国科学院成都文献情报中心

NATIONAL SCIENCE LIBRARY(CHENGDU), CHINESE ACADEMY OF SCIENCES



科技文献数据挖掘最新研究进展

汇报人：蔡祎然

导师：胡正银

单位：中国科学院成都文献情报中心

2024年9月20日

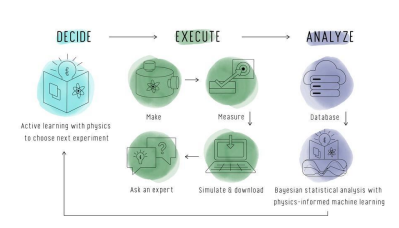
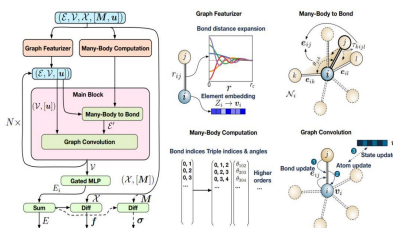
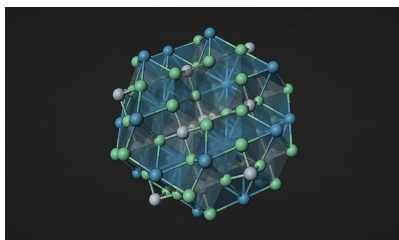
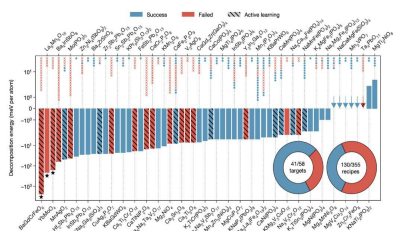


“AI+”时代，数据的重要性日益凸显



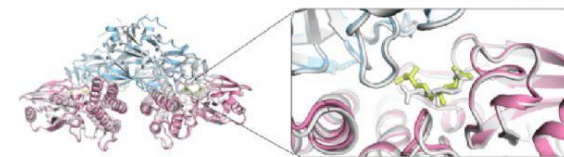
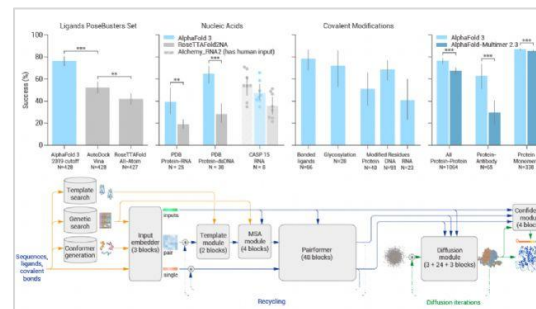
数智驱动（AI4S）创新成为国际前沿热点，科学数据是AI4S的基础

A-Lab (2023)



A-Lab利用Materials Project数据库、CSD剑桥晶体结构数据库中24000多种出版物的文献等的历史数据，生成合成配方，显著提升实验效率和成功率

AlphaFold 3 (2024)



Proteins with unique folds: The AziU3/U2 protein complex catalyses the formation of aziridine rings. at a novel catalytic site formed by a complex with a novel fold. AF3 correctly predicts this novel fold of the protein complex bound to the substrate (PDB ID 7WUX).

20万个已知结构



+

30万个未知结构



AlphaFold 3基于蛋白质结构数据库PDB和包含未知结构蛋白质序列的蛋白质知识库UniProt进行训练，能够高精度预测PDB内几乎所有分子类型复合物的模型

科技文献是一种国家的重要科技基础能力



科技文献领域数据来源广泛、共享和利用环节**多且复杂**、涉及主体多元



新范式下数据面临的问题及解决方案



- **数据/计算**资源分布在不同的学科/领域数据中心，数据处理流程**横跨不同的数据中心**
- 目标**数据体量一般较大**，迁移困难。由于安全及隐私方面考虑，数据不愿意公开。

多源异构数据整合难度大，缺乏统一标准规范



构建跨领域本体映射框架，制定数据整合标准

算法处理大规模数据效率低，精度和鲁棒性不足



运用大语言模型和先进的机器学习技术，优化算法结构

应用场景有限，可视化和解释性不足，难以发挥数据的应用价值



构建知识图谱，开发交互式可视化工具，优化应用策略





01

本体构建

02

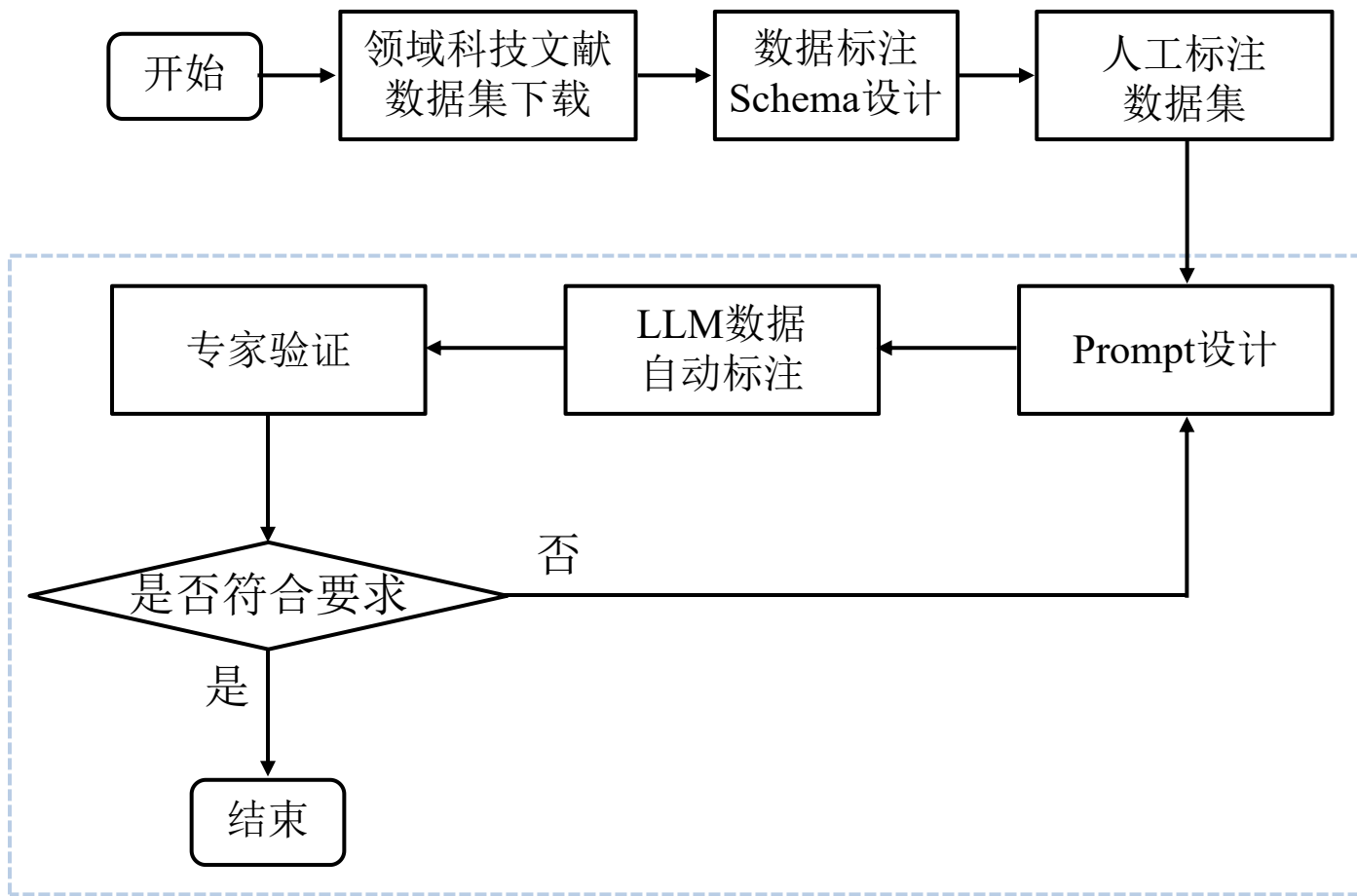
关键技术

03

应用场景

04

小结



LLM标注核心流程

■ 数据标注目标

以构建数据库为主要目标，以某一研究领域为主题，**从科技文献中摘取科学实体，进行标注、整理，建立数据库**，为科学研究提供帮助，协助知识推理和知识发现。

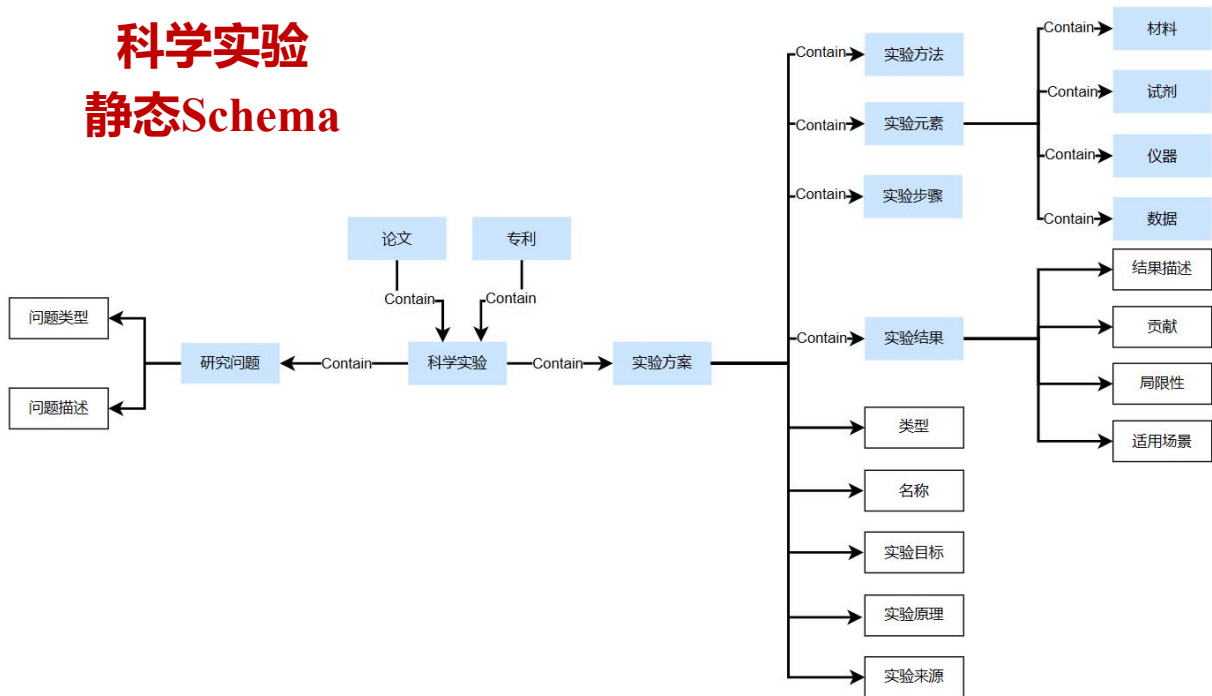
■ 科技文献数据源

科技文献中蕴含大量“可信、专业、规范”的知识与数据，科学实验是科技文献的核心内容，通过**对科学问题、科学原理、实验步骤等进行建模**，有效支撑数智驱动的知识发现与科学发现。

科技文献领域数据标注schema (以有机太阳能电池为例)

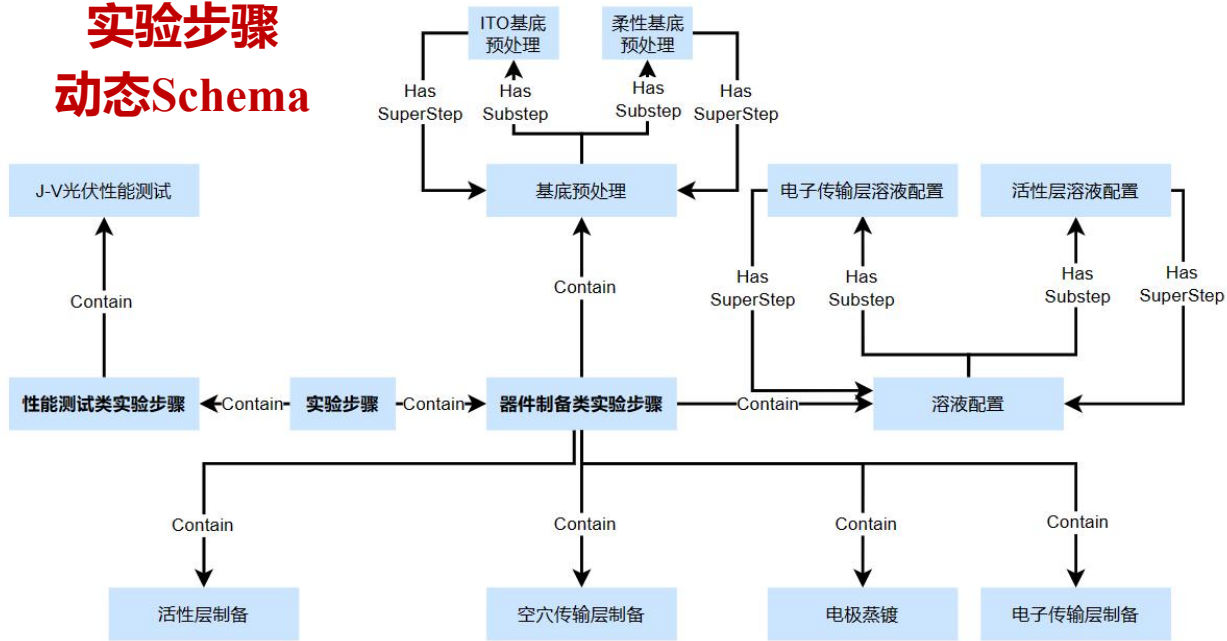


科学实验 静态Schema



- ① 将科学实验定义为**研究问题**和**实验方案**两大部分
- ② 将科学实验核心的**实验目标**、**实验原理**和**实验来源**等设置为实验方案对应的属性
- ③ 核心的实验方案下分**实验方法**、**实验元素**、**实验步骤**和**实验结果**共四个大类
- ④ 实验元素下分实验涉及的**材料**、**试剂**、**仪器**和**数据**

实验步骤 动态Schema



- ① 利用文本挖掘技术自底向上对静态本体进行迭代完善，来构建特定学科领域**实验步骤动态事件本体**
- ② 有机太阳能电池为例，将其实验步骤建模为**器件制备类**和**性能测试类**两种，前者共6步：溶液配置、基底预处理、电子传输层制备、活性层制备、空穴传输层制备电极制备；后者指J-V光伏性能测试

科技文献知识挖掘的思路

- 利用科技文献**大数据和深度学习技术**方法，从科技文献中**自动学习**获取科技文本挖掘的重要**知识**。
- **汇聚原始的科技文献与科学数据资源**，实现科技论文、科技图书、专利、科技报告、基金项目等多模态科技文献资源的集成、关联与融合。
- **实现科技文献全文内容的结构化解析**，包括章节段落的解析与定位、图表抽取与定位、图片的语义标签揭示，形成XML结构化文本。
- **科技文献中细粒度内容的挖掘和揭示**，包括文章中的文本结构标注、文章的逻辑结构标注、文章中的细粒度知识及知识关系标注(实体)。
- **构建高质量确信的科技文献本体知识库**，基于深层次知识及知识关联，建设科技文献本体和科技文献知识图谱。

第一步：领域科技文献资源汇聚整理



科技文献



专利文献



科技图书



基金项目



科技报告



实验数据



评论文章

第二步：科技文献全文结构化解析

Word解析 / PDF解析

科技文献解析与内容结构化

科技文献内容结构化，形成XML结构化文本

科技文献内容解析模型

章节段落解析与定位 / 图片表格抽取 / 图片语义化标签揭示

第三步：科技文献内容挖掘揭示



科技文献

知识对象
富集区定位

篇章级/句子级、
语义富集区
知识对象抽取

科学数据

知识对象

关键句子

第四步：科技文献本体知识库

科技文献本体建设

深层次知识关联

映射关联分析

数据质量控制

数据可靠性控制

科技文献知识图谱建设

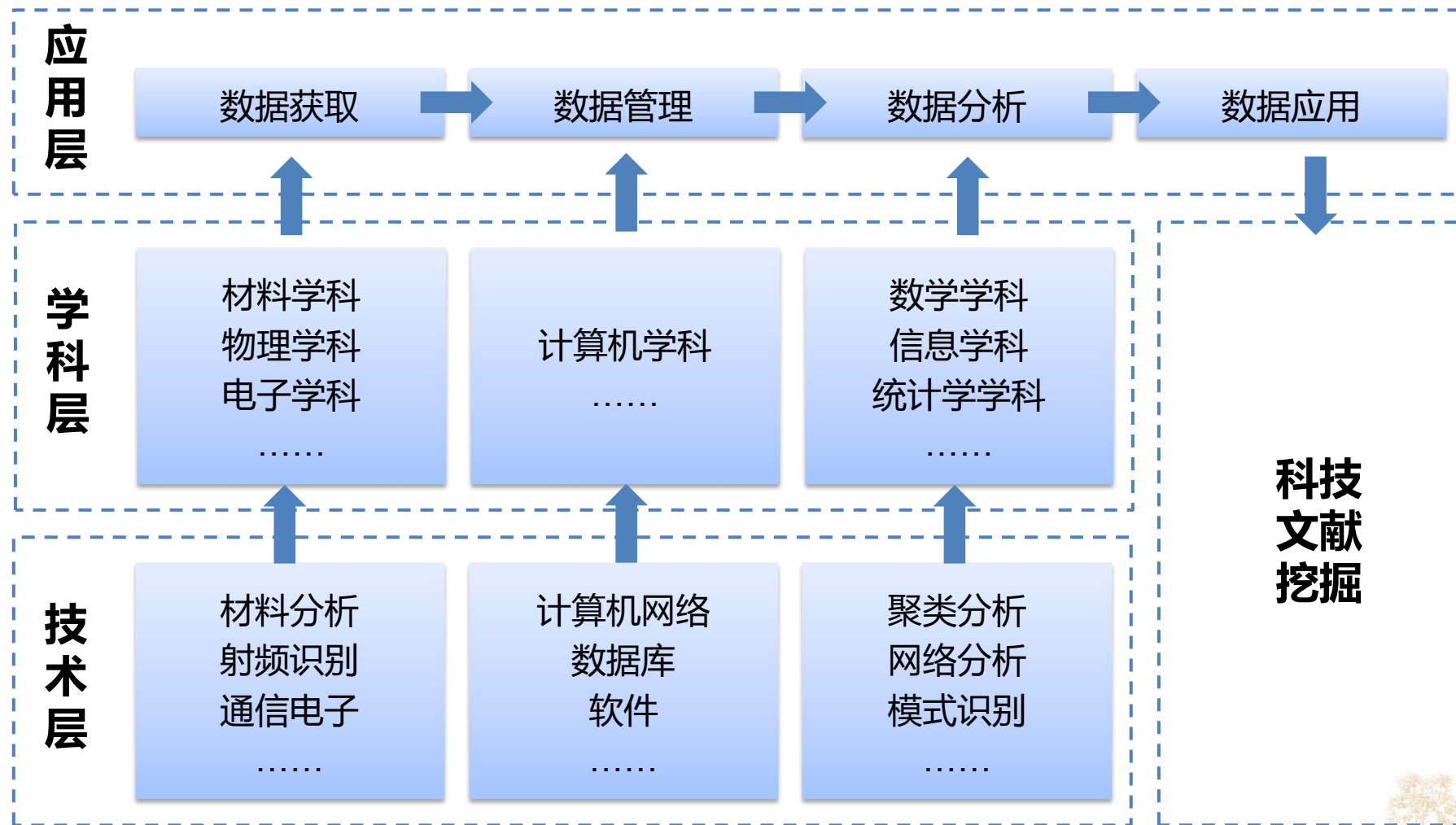
数据安全性鉴别

数据可信存储

数据深度聚类

数据智能管理

技术架构



科技文献领域数据挖掘关键技术比较



关键技术	优点	局限性	适用场景
数据挖掘	从大量数据中识别和提取潜在价值的信息 强调对知识单元和知识关联的识别与利用	对数据质量要求高 需要专业知识进行模型构建和解释	科技文献内容分析 研究趋势预测
文本挖掘	人工参与少 减少用户偏见与先验知识的限制	可能忽略上下文信息 对于专业术语的识别可能不准确	文献摘要自动提取关键词 研究趋势分析
知识图谱构建	构建科技领域实体和概念的关系网 支持复杂查询和推理	构建和维护成本高 对于动态数据更新有挑战	科技领域知识体系构建 科技文献关联分析
关联规则	发现文献中变量间的关联 识别研究领域的关联模式	可能产生大量无用规则 于稀疏数据效果不佳	研究领域关联分析 推荐相关文献
深度学习	自动学习特征，无需人工干预 提高文献检索的准确性	需要大量计算资源 模型训练时间长	文献内容的语义分析 精确检索文献
机器学习	处理不规则矩阵 对科技期刊文本进行分类、聚类、推荐等	需要大量标记数据 模型训练和调优耗时	文献影响力预测 适用于图形数据的分类
聚类分析	将相似的文献自动分组 无需预先定义类别	对于大规模数据集效率低 可能需要手动调整参数	文献相似性分析 文献自动分类
LDA主题建模	识别科技文献中的隐含主题 支持大规模文本分析	模型参数调整复杂 主题的解释性可能不直观	文献主题发现 研究趋势预测

通过集成大语言模型优化技术性能



结合**大语言模型**弥补方法技术的局限性，提升科技文献领域的数据处理和分析效率

科技文献领域模型

文献特点

科学语言的专业性和前沿性

特殊的文体结构

图片和表格多模态信息

引文网络错综复杂

学科特点

文献内容组织形式

学科研究方法和工具

数据呈现形式

专用知识组织体系

根据已有研究成果^[19]，在采用物理模型试验资料进行分析从而取得具体经验表达式时，可以采用递推方式求得经验关系，具体形式为：

$$\frac{R_s}{H_s} = f_1(m_d) f_2(m_u) f_3\left(\frac{H_s}{L_{op}}\right) f_4\left(\frac{d_w}{H_s}\right) f_5\left(\frac{B}{H_s}\right) \quad (1)$$

当断面上不带平台(折坡)时，经验关系式为：

$$R_s/H_s = f_1(m_d) f_2(m_u) f_3(H_s/L_{op}) \quad (2)$$

物理学文献通常包括复杂的数学方程和理论推导

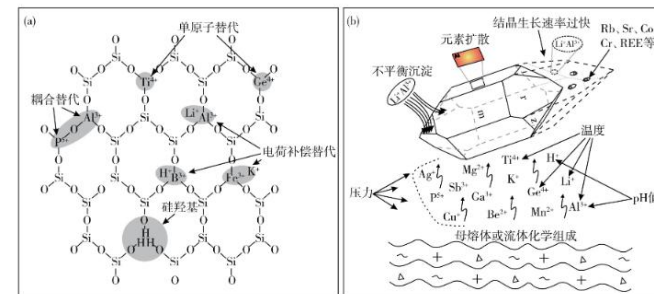


图1 石英晶体结构和常见微量元素赋存机制(a)及影响因素(b)示意图^[28]
Fig. 1 Schematic crystal structure showing the configuration of trace elements in quartz lattice (a) and the integrated influence factors of trace elements concentration in quartz crystal (b)^[28]

化学领域的文献则多涉及化合物的合成、结构分析和化学反应等，重视化学式和反应方程的使用

从科技文献中深度挖掘相关知识，并提供创新应用



科技文献

通用领域知识对象挖掘

专业领域知识对象挖掘

应用场景



篇章句子级
知识对象

- 研究背景句
- 研究目的句
- 研究方法句
- 研究结果句
- 研究结论句
- 研究问题句
- 概念定义句
-

短语级
知识对象

- 研究问题
- 方法模型
- 软件系统
- 度量指标
- 仪器设备
- 理论原理
- 概念定义词
- 数据资料
- 科学家
- 地点
-

物理
领域

现象
规律

数学
方法

理论
定理

实验
仪器

.....

化学
领域

有机
物

无机
物

化学
反应

化学
键能

.....

医学
领域

微生
物

疾病

靶标

药物

.....

计算机
领域

现象
规律

数学
方法

理论
定理

实验
仪器

.....

资源
环境
领域

双碳

水文

地质

矿物

.....

科技
政策
领域

科研
项目

规划
布局

保障
措施

科技
人才

.....

语义检索
精准推荐
智能监测
分析评估
领域画像
.....

案例1：科技文献知识AI引擎（SciAIEngine）



在SciAIEngine基础上开发了科技文献内容深度挖掘引擎，支撑科技文献内容中**多领域、多层次、细粒度知识对象**的自动标注。

功能

 语步识别 自动识别科技文献摘要的研究目的、研究方法、研究结果、研究结论等重要句子，显性揭示科技文献的重要内容。	 科技文献分类 自动对科技文献进行分类。AI引擎创新性地利用了多层分类模型以提升文献分类的准确率。	 关键词识别 自动识别科技文献中的关键词。以先进的关键词识别模型为基础。	 命名实体识别 自动识别科技文献中出现的命名实体，包括面向通用领域的人名、地名、机构名等以及面向专业领域的模型方法、仪器设备等。
 概念定义句识别 自动识别科技文献中的概念定义句子。	 文本标题生成 自动生成表达科技文献内容的文献标题。	 审稿人推荐 自动推荐适合评审某篇论文的审稿专家。	 投稿期刊推荐 自动推荐适合的投稿科技期刊。
 研究问题句识别 自动识别科技文献中表达研究问题的句子。	 引用句识别 自动识别科技文献中引用句并判断其引用情感和引用意图。	 深度聚类 自动聚类科技文献，生成类簇标签。	 结构化自动综述 自动综合科技文献，快速以结构化的形式自动梳理文献集的研究脉络。

语步识别

[英文摘要语步识别](#)

[中文摘要语步识别](#)

[基金项目语步识别](#)

科技文献分类

[中文科技文献分类](#)

[英文科技文献分类](#)

[中文政策文本分类](#)

[脑机接口领域分类](#)

[空间技术领域分类](#)

科技文献关键词识别

[中文科技文献关键词识别](#)

[英文科技文献关键词识别](#)

命名实体识别

[中文通用领域实体识别](#)

[英文通用领域实体识别](#)

[中文医学领域实体识别](#)

[中文科研实体识别](#)

[英文科研实体识别](#)

[中文科研实体关系识别](#)

[中文化合物实体识别](#)

[中文政策实体识别](#)

[英文物理学领域实体识别](#)

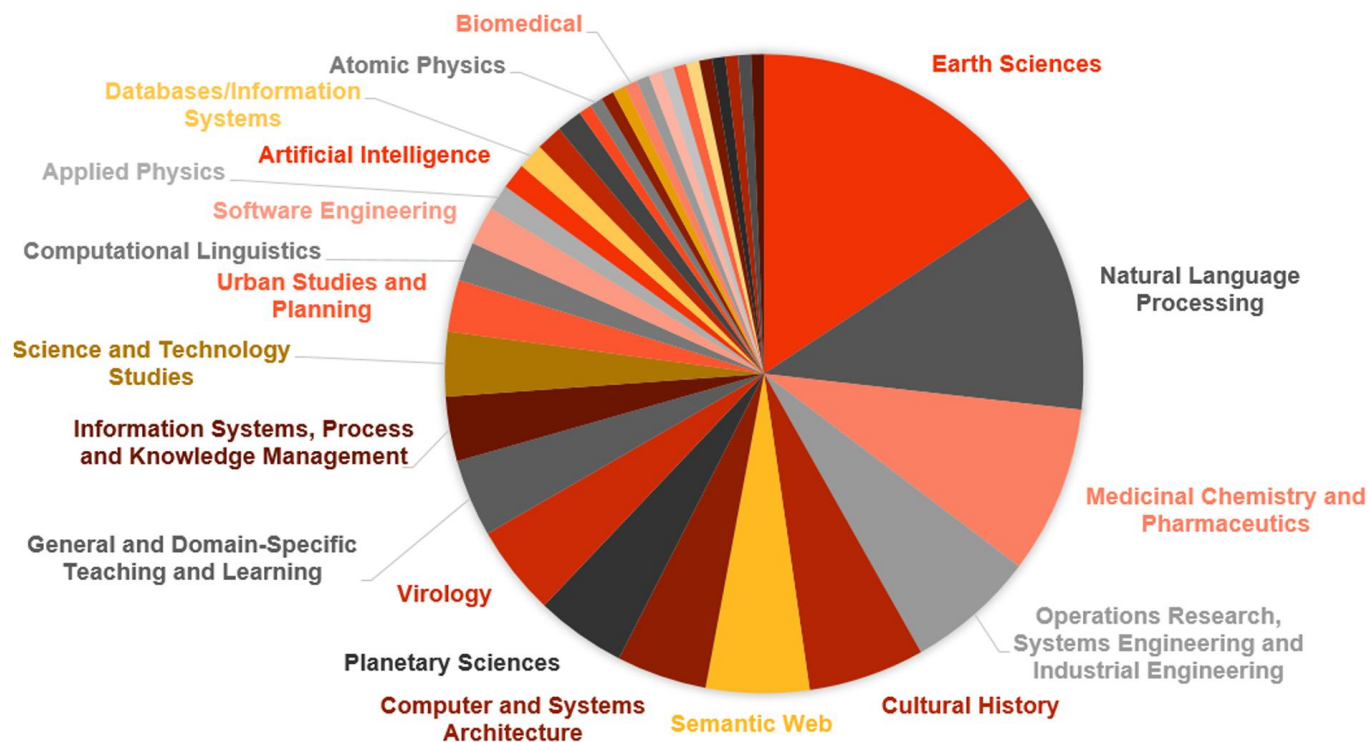
概念定义句识别

[英文科技文献定义句识别](#)

案例2：开放研究知识图谱 (ORKG)



ORKG 使科学知识对人类和机器都具有可操作性，从而实现了全新的机器辅助方式。这将帮助研究人员找到与其领域的相关贡献，并创建最先进的比较和评论。借助 ORKG，科学家可以以全新的方式探索知识，并在不同学科之间共享结果。



Nechakhin, Vladyslav et al. "Evaluating Large Language Models for Structured Science Summarization in the Open Research Knowledge Graph." ArXiv abs/2405.02105 (2024); n. pag.

案例3：知识服务平台



发现未知的已知——结合知识图谱与人工智能技术，融合多源数据、挖掘知识关联、促进知识发现，服务AI4S，为科研范式变革贡献文献情报力量。



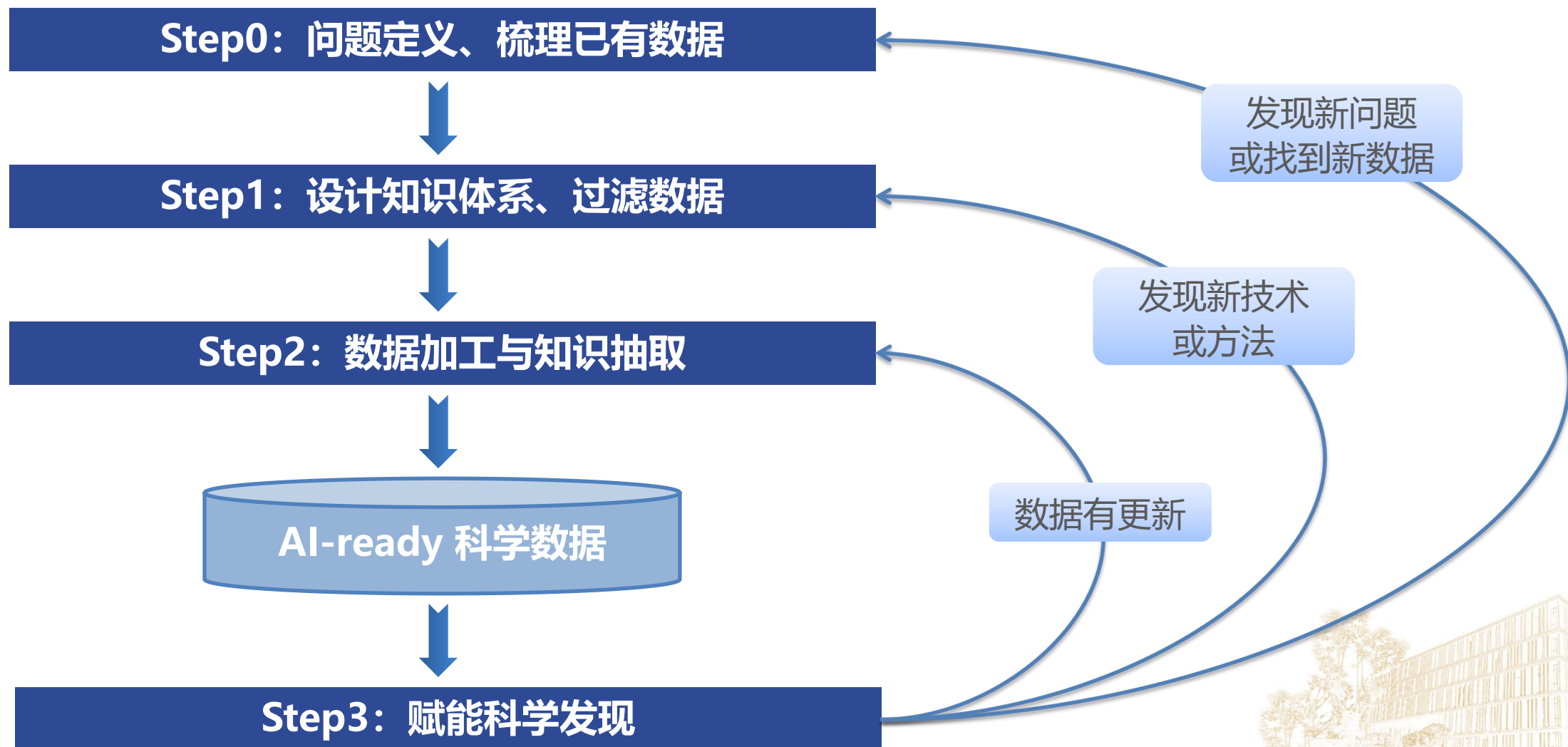
中国科学院知识服务平台



融合多源数据的干细胞知识发现平台

胡正银,刘蕾蕾,陈文杰,等. 面向科学知识发现的造血干细胞知识图谱构建研究[J]. 数据与计算发展前沿,2021,3(6):81-97. DOI:10.11871/jfdc.issn.2096-742X.2021.06.006.

科技文献数据挖掘流程：



现状：蓬勃发展的领域

技术进步：机器学习、自然语言处理技术不断进步。

应用广泛：在生物医学、工程、社会科学等多个领域得到应用。

数据量激增：科研数据量呈指数级增长，为研究提供丰富资源。

01

未来：持续创新与整合

多模态分析：结合文本、图像、声音等多种数据类型进行分析。

跨学科融合：促进不同学科间的知识和技术交流。

智能化工具：开发更智能的文献挖掘工具，提高研究效率。

02

挑战：需要克服的难题

数据质量：确保数据的准确性和可靠性，有效整合文献数据与科学数据。

隐私保护：在挖掘过程中保护个人和机构的隐私。

技术门槛：提高技术的可访问性，解决文献数据挖掘软件工具鲁棒性。

03

敬请批评指正!



NATIONAL SCIENCE LIBRARY(CHENGDU), CHINESE ACADEMY OF SCIENCES

中国科学院成都文献情报中心