

高能物理领域高质量科学数据体系建设与思考

胡皓，齐法制，张正德，熊东波，李庆梦，胡鹏等

中科院高能物理研究所

2025年8月26日

Outline

01 高能物理大数据和AI4HEP

02 高能物理/大装置科学数据体系

03 高质量科学数据集构建思考

04 总结与展望

Outline

01 高能物理大数据和AI4HEP

02 高能物理/大装置科学数据体系

03 高质量科学数据集构建思考

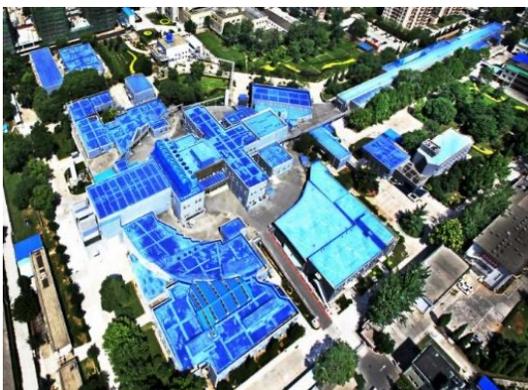
04 总结与展望

高能所



正式成立于1973年，科学院最大的基础研究机构，~ 1500职工，~ 1000研究生。

定位：国际领先的高能物理中心之一，世界先进水平的大型、综合性、多学科研究基地。



北京 玉泉路



广东 东莞



北京 怀柔



山东 济南



学科布局与主要研究方向



- 面向高能物理领域提供领域科研信息化能力，“服务 + 科研” 学科融合发展
- 人力资源
 - 55 正式职工，20 名项目聘用职工，35 学生/博士后
 - 计算机、粒子物理、光子和中子科学背景的交叉人才队伍
- 研究领域

Computing and Storage

- High Performance Computing
- High Throughput Computing
- Grid/Cloud computing
- Distributed storage

Network and Cyber security

- Datacenter and campus network
- Dedicated link for remote experiments
- International network collaboration

IT Services

- Database technology and application
- Conferencing Technology
- Institutional management tool

Scientific Software

- Open data and open science
- Scientific data management
- Scientific software framework

Innovation

- Big data
- AI for science
- Quantum computing

高能物理和大数据



高能物理

高能物理（粒子物理）是探索物质结构、宇宙起源及演化、生命起源等科学前沿问题的基础学科。

通过实现数据资源、软件工具、数据分析等资源能力的汇交和共享，国家高能物理科学数据中心已建设超过**90PB**的海量科学数据，主要包含5个领域：

粒子物理数据



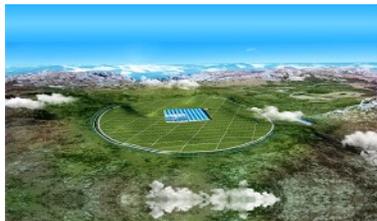
JUNO 江门中微子
BESIII 北京谱仪
DYB 大亚湾
CMS ATLAS LHCb
强子对撞
L3c L3宇宙线

中子科学数据



CSNS 散裂中子源

天体物理数据



LHAASO 高海拔宇宙线
YBJ-Asy 羊八井
ARGO-YBJ 羊八井
HXMT 硬X射线调制望远镜

光子科学数据



BSRF 北京同步辐射光源
HEPS 高能同步辐射光源

其他相关学科数据



HEPS 2025年建成，预计后
每年产生**300+PB**原始数据

中国/高能所牵头的高能物理大科学装置



国家高能物理科学数据中心



5大重要 应用场景



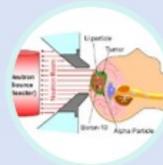
基础研究
粒子物理研究
中微子研究
宇宙线研究
空间科学研究



产业创新
特种材料研究
超导技术
微电子技术



人工智能
科学大模型
数据供给



生物医学
医学诊断
射线治疗



国际合作
国内外大科学装置与计划

6类数据加工与 服务方式

>24.3亿条
文件记录数

9个
典型数据库

62PB
数据资源

>10种
软件工具

>3.1亿人次
访问总量

>700PB
数据访问量/年

4种科学数据 汇交模式



科技项目汇交



文章关联数据汇交



大科学装置数据汇交



国际科学数据镜像

8类数据来源



粒子物理



天体物理



空间科学卫星



光子科学



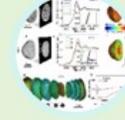
中子科学



正电子科学



国际合作实验



重大科研项目



文章数据汇交

5家数据资源 分中心



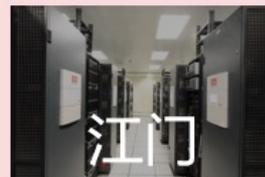
北京



东莞



稻城



江门



怀柔

跨地域分布式的数据和计算中心



分布式部署

- ◆ 北京总中心
- ◆ 大湾区分中心/广东省物质科学数据中心
- ◆ 怀柔分中心
- ◆ 稻城分中心
- ◆ 江门分中心

资源能力

- ◆ 数据存储：130PB 磁盘+100PB 磁带
- ◆ 科学计算：10 万 CPU 核+300 张 GPU 卡
- ◆ 网络能力：100Gbps国际通讯能力

1.6Tbps 内部高速互联

至怀柔分中心：100Gb/s

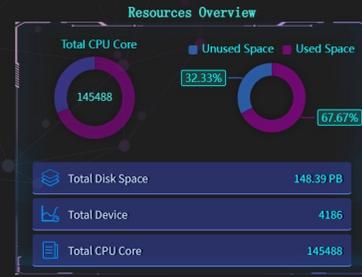
至大湾区分中心：20Gb/s

至稻城分中心：3Gb/s

至江门分中心：2Gb/s



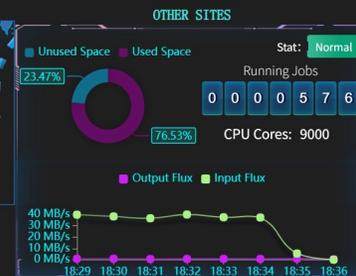
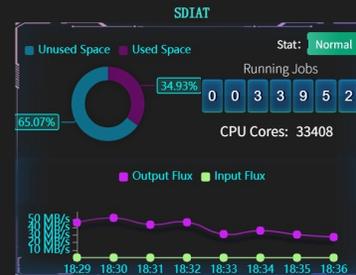
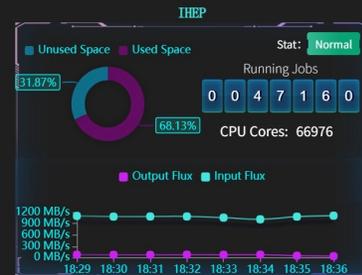
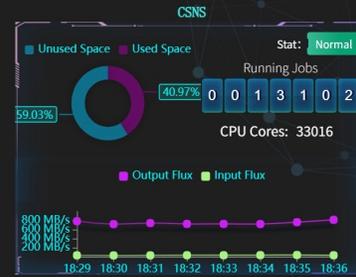
Computing Platform for High Energy Physics



Current Files Number
2,899,581,030

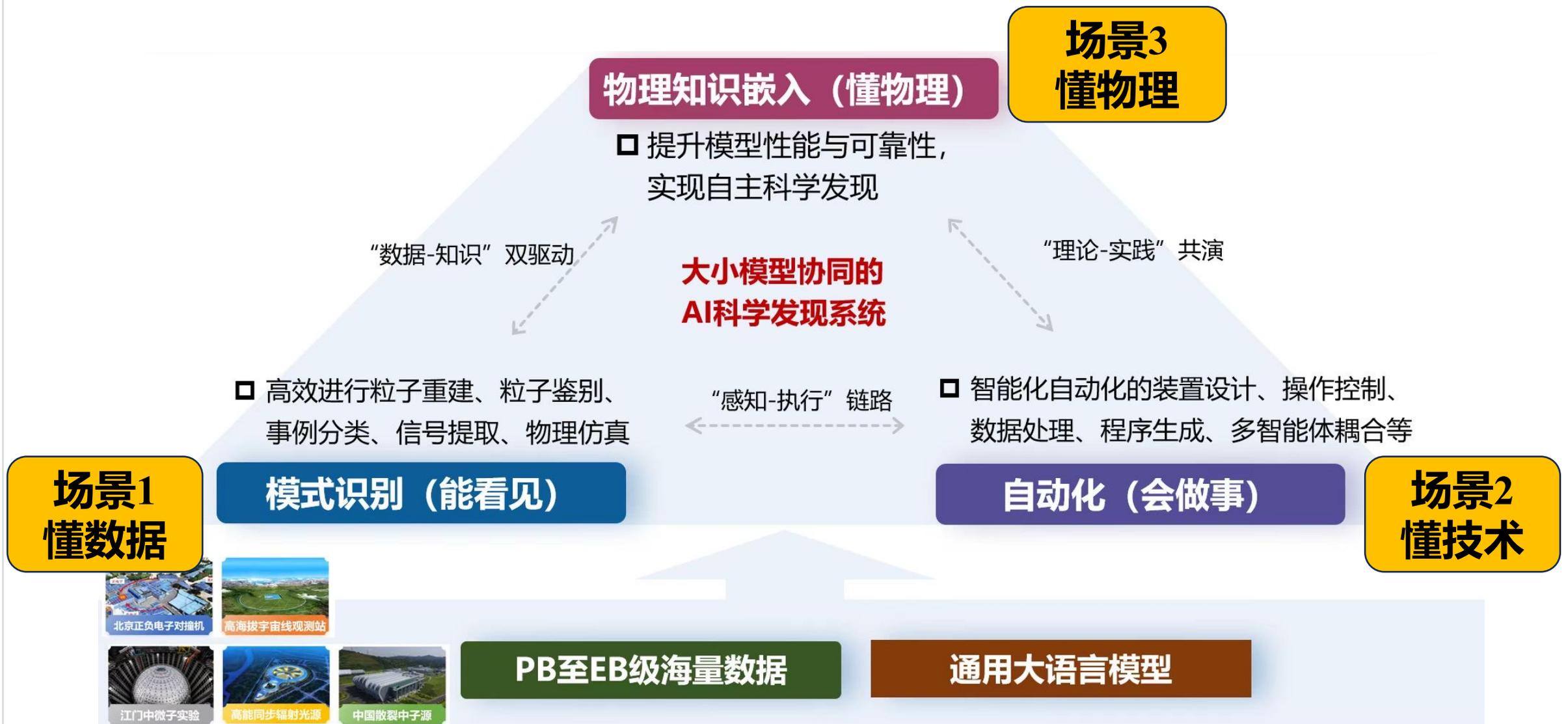
Current Running Jobs
97,660

Current Data Volume
100.4 PB



通过高能物理开放网络（High Energy Physics Open Connections）和开放数据中间件（Open Data Middleware, ODM），构建了集科学数据、异构算力和科学软件为一体的高能物理数据服务平台。汇聚百PB科学数据，十余万CPU核、三百张GPU卡的数据服务能力，数据访问量700PB/年。为粒子物理、天体物理、理论物理、光子科学、中子科学等多领域近万名科研人员和产业用户服务。

人工智能赋能高能物理科研活动的途径和场景



科学数据是AI4S/AI4HEP发展的基础



AI4HEP

提升科研效率

推动范式变革

促进科学发现



② 自动化

任务专用
AI算法研究

实时处理 ⑤

模拟 ①

重建 ③

⑨ 分析

通用AI
大模型探索

HEP
文本大模型

HEP
科学数据大模型

"赛博士"
科研智能体

HepAI 高能物理人工智能平台研发

国家高能物理科学数据中心
National HEP Data Center

AI-Ready科学数据集

数据集标准

数据标注工具

模型Hub

高性能I/O

Outline

01 高能物理大数据和AI4HEP

02 高能物理/大装置科学数据体系

03 高质量科学数据集构建思考

04 总结与展望

人工智能对数据体系提出了新的需求



□ 数据的模态：

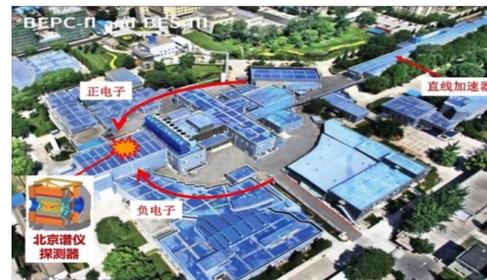
- 文本（文献、代码、逻辑过程）
- 图（Graph）、事例、团簇序列
- 图像、成像、谱学、衍散射

□ 数据需要AI-Ready

- FAIR（基础要求）
- 输入数据、**真值标签**、**AI任务**
- **高质量数据**

↑
**Garbage in,
Garbage out!**

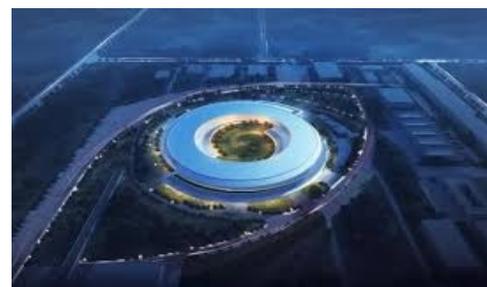
↑
额外需求



北京同步辐射装置



散裂中子源



合肥先进光源



高能光源



大装置数据体系是大科学装置中数据的采集、存储、处理、分析和使用的整体框架和机制。它涵盖数据从生成到最终使用的整个生命周期，并确保数据的质量、完整性、安全性和可用性。

数据策略与标准

□ 数据模型

- 表征数据
- 元数据

□ 数据标准

- 数据格式
- 元数据包

□ 数据策略

- 保存策略
- 使用机制

数据软件与工具

□ 采集与处理

- 实验设施运行控制
- 实验过程控制

□ 存储与管理

- 大数据存储
- 元数据管理

□ 分析与应用

- 科学数据分析软件
- 数据可视化

□ AI-Ready数据构建

- 数据标注与增强
- 多源数据融合

数据服务与供给

□ 面向领域科学家

- AI和数据接口

□ 面向AI研究人员

- 数据发现
- 数据解释

□ 面向AI机器人

- 数据管道
- 机器可用

数据体系1：策略与标准

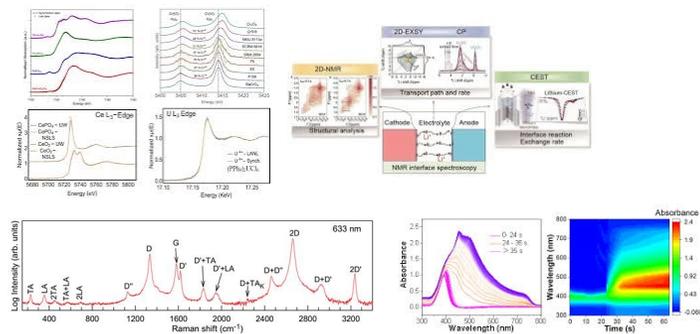


数据模型

表征数据



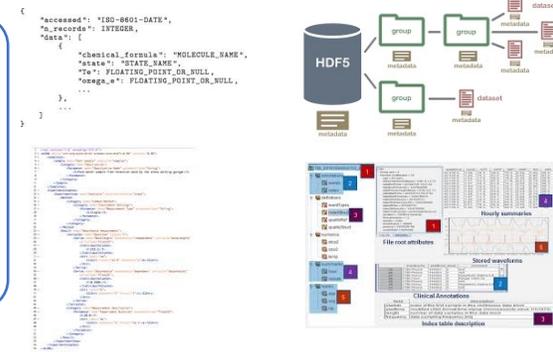
元数据



- 样品属性、实验环境、预处理方法等
- 可扩展、满足科研需求

数据标准

数据格式



数据策略



- 数据保存、利用机制

元数据标准



- 元数据实体信息由6个元数据包组成
- 每个包都有对应的实体或聚合实体

元数据实体集信息

数据集信息包

样品信息包

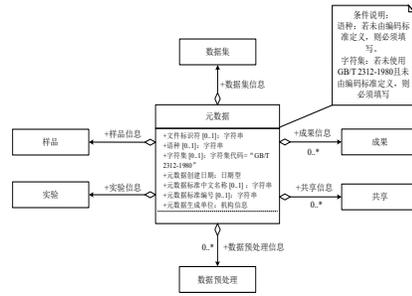
实验信息包

预处理信息包

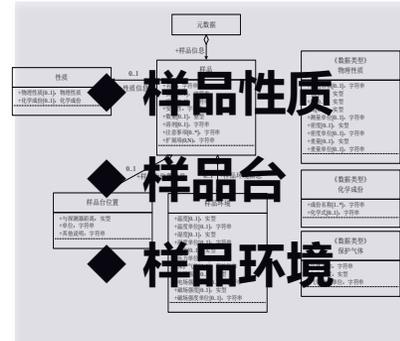
共享信息包

成果信息包

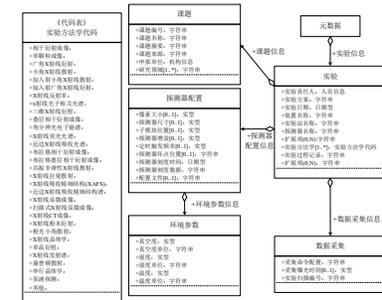
数据集信息包



样品信息包

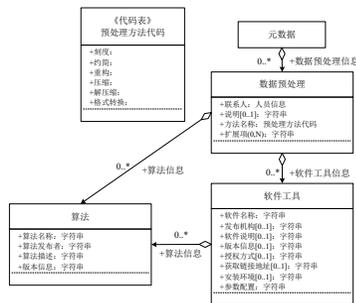


实验信息包

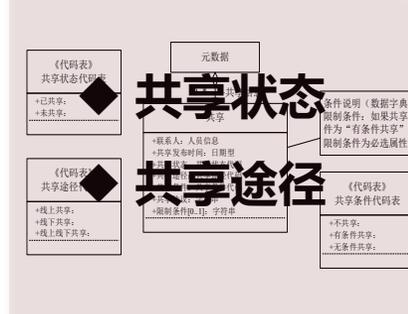


- 课题
- 探测器配置
- 环境参数
- 数据采集信息

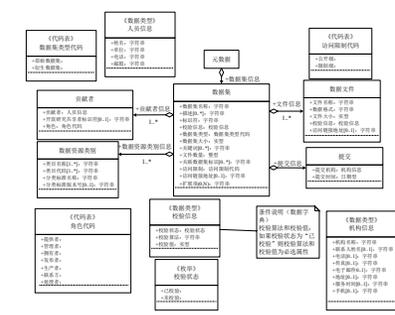
预处理信息包



共享信息包



成果信息包



- 数据资源类别
- 贡献者
- 数据文件
- 提交信息

数据体系1：策略与标准

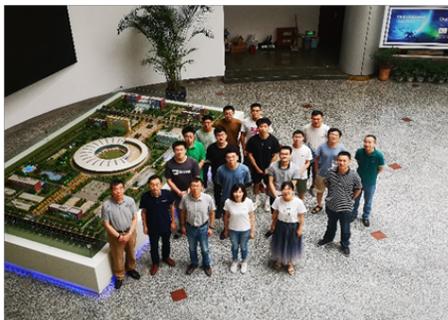


- 高能所、中科大、上科大、上海高研院等多个装置联合计算机网络信息中心，推动光源类大科学装置实验数据元数据标准，支撑光源科学数据管理、利用和共享。
- 推荐性国家标准《重大科技基础设施平台：光子与射线实验数据元数据》，2025年已国标立项。

论证意见：体系符合性好、必要性强、成熟度高，予以推荐立项



大科学装置数据策略



合肥 数据策略/标准研讨

SZAC 国家标准化业务管理平台

平台首页 制修订管理 外文版管理 标准样品管理 TC组织管理 TC工作平台

当前位置：国家标准计划项目信息

重大科技基础设施平台 光子与射线实验数据元数据 (计划号：20250950-T-306)

制修订：制定 标准性质：T 状态：**正常** 当前环节：**组织起草**

项目信息 文件&投票 业务流转 项目周期 协助办理记录 征求意见反馈 样书申领信息

项目信息 **编辑** 打印标准申报单 修改日志

基本信息			
计划号	20250950-T-306	项目编号	2024002968
原中文名称	重大科技基础设施平台 光子与射线实验数据元数据		
现中文名称	重大科技基础设施平台 光子与射线实验数据元数据		
英文名称	Major Scientific and Technological Infrastructure Platform: Metadata for Photon and Radiation Experiment Data		
标准性质	推荐性国家标准	制定/修订	制定
采标类型	无		
项目周期	18个月	标准类别	基础
ICS	35.240.01	CCS	

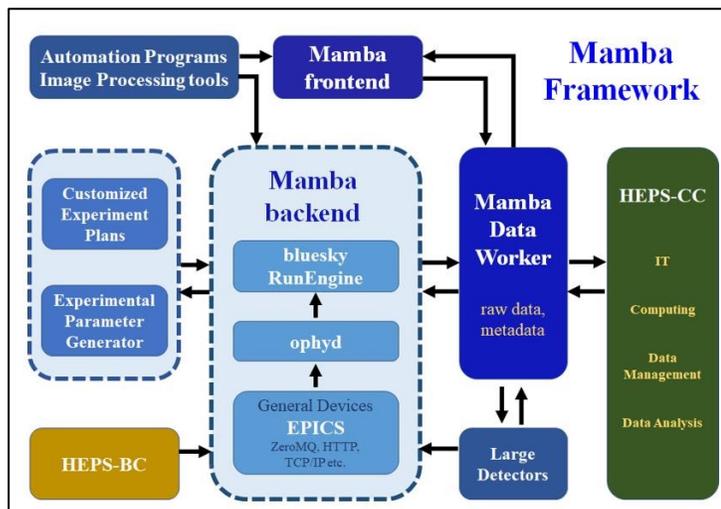
数据体系2：软件与工具



覆盖数据生产到使用的全生命周期的框架/工具/软件

数据采集和处理

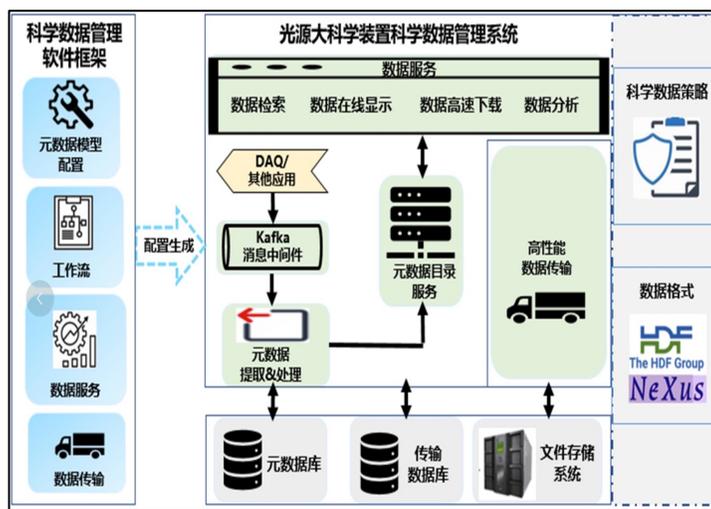
实验控制和数据采集框架Mamba



- ❑ 复杂实验过程控制
- ❑ EB级别大数据智能采集
- ❑ 高效在线解析

数据存储和管理

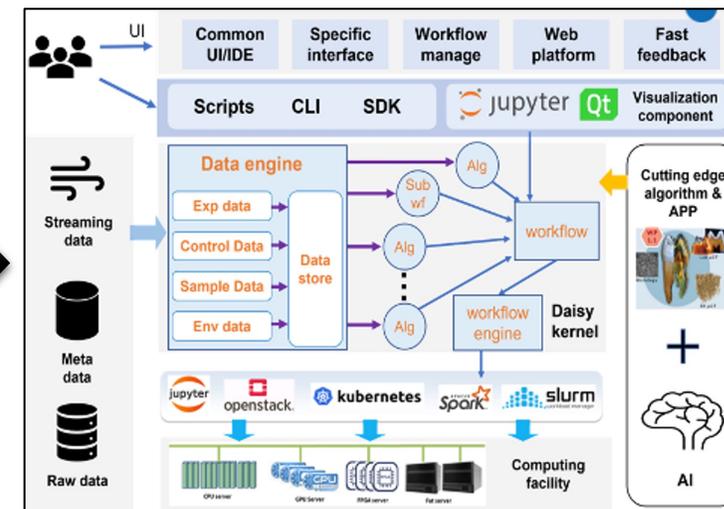
数据管理与服务框架Domas



- ❑ EB级别大数据存储
- ❑ 复杂的元数据管理
- ❑ 标准化数据格式

数据分析和应用

科学数据处理软件框架Daisy



- ❑ 多方法学及AI算法集成
- ❑ 高通量数据I/O
- ❑ 学科软件和工作流用户界面

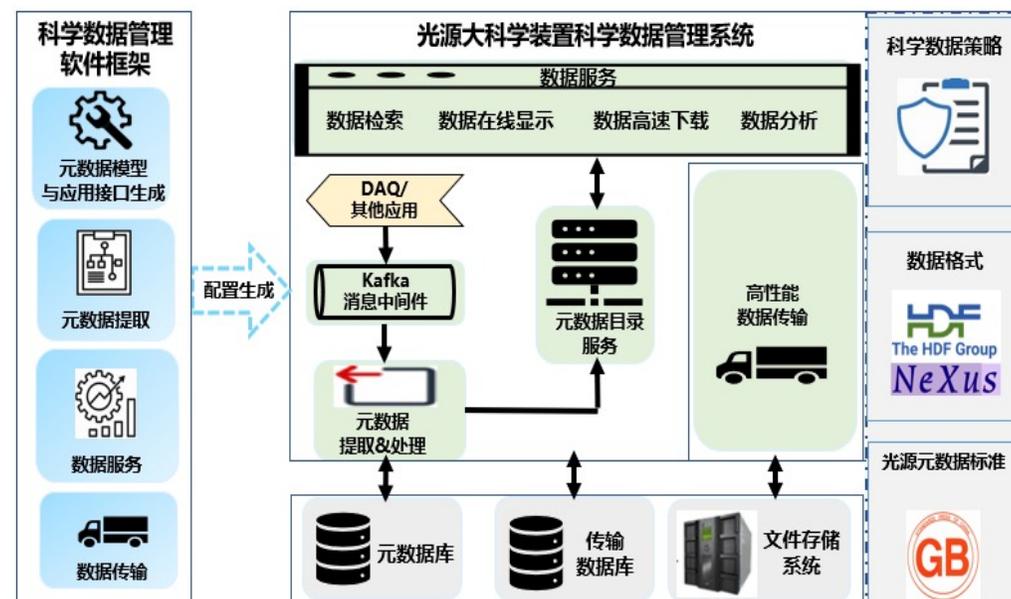
科学数据管理软件框架DOMAS



数据管理：对实验数据高效存储、组织和管理，提供数据全流程“一站式”服务，数据从**获取-->传输-->存储-->分析-->发布**等全生命周期管理，提升数据的利用、共享能力及效率

核心功能

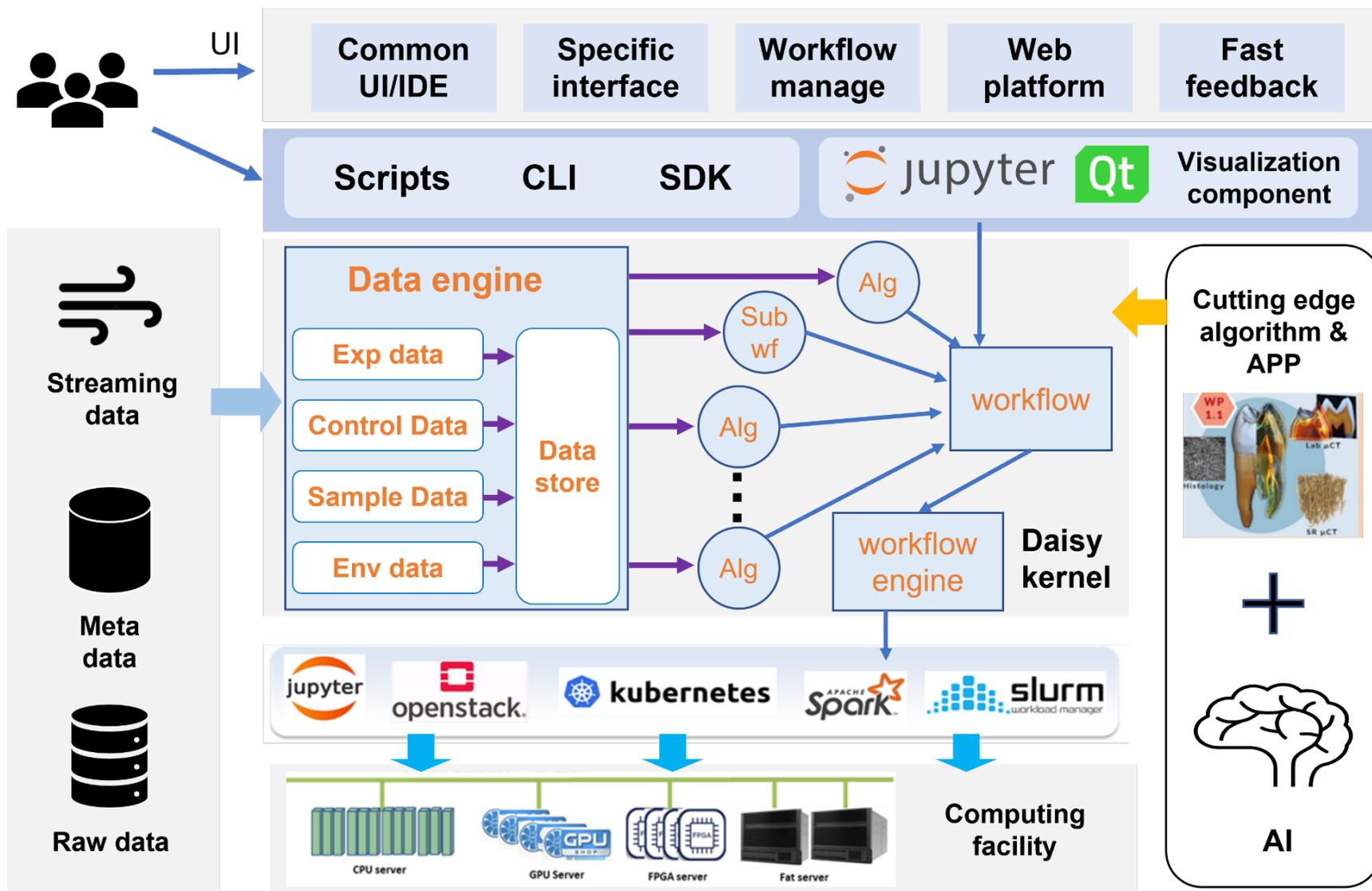
- ✓ 元数据模型：对元数据目录设计和存储，快速便捷地检索数据
- ✓ 元数据提取：从不同系统（DAQ、存储、用户服务、数据分析）获取元数据
- ✓ 数据存储：数据存储目录设计、权限设计
- ✓ 数据服务：提供数据访问、下载、分析与追踪服务
- ✓ 开放接口：标准接口，满足各系统/模块访问数据或元数据



软件框架的意义：

可应用于其他光源类装置或实验站的科学数据管理系统的快速开发，避免各装置进行独立开发而投入昂贵的人力和时间成本。

科学数据处理软件框架Daisy

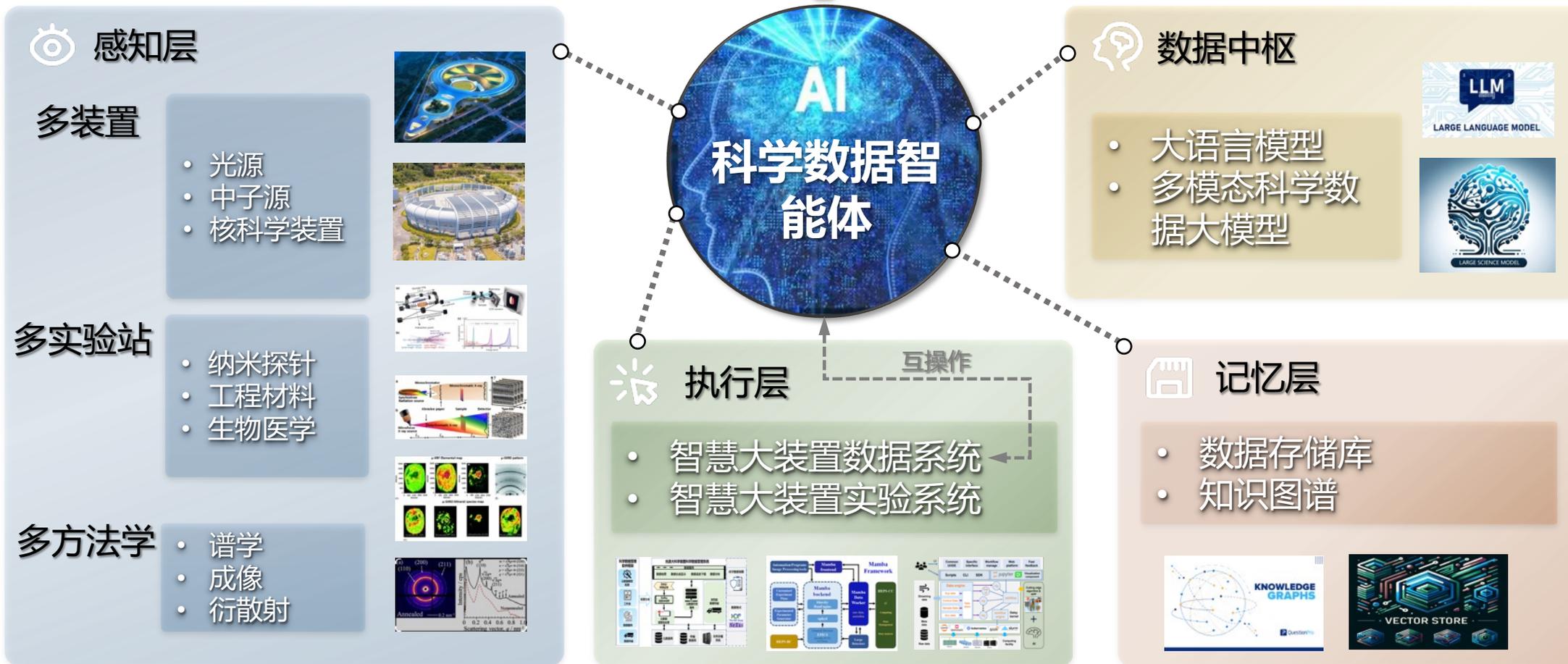


- 数据处理软件框架核心
- 满足新一代光源数据处理需求的衍生技术模块
 - 应对高通量数据I/O、多模态数据解析、多源数据接入的数据对象管理
 - 应对不同规模、不同通量、低延迟数据处理需求的弹性异构计算集群算力支持
 - 服务于学科方法学软件集成和发展的用户软件接口和软件开发环境支持
- 基于软件框架的学科专用应用软件以及针对灵活数据处理需求的通用工作流编排系统

数据智能体：解决高质量数据供给问题

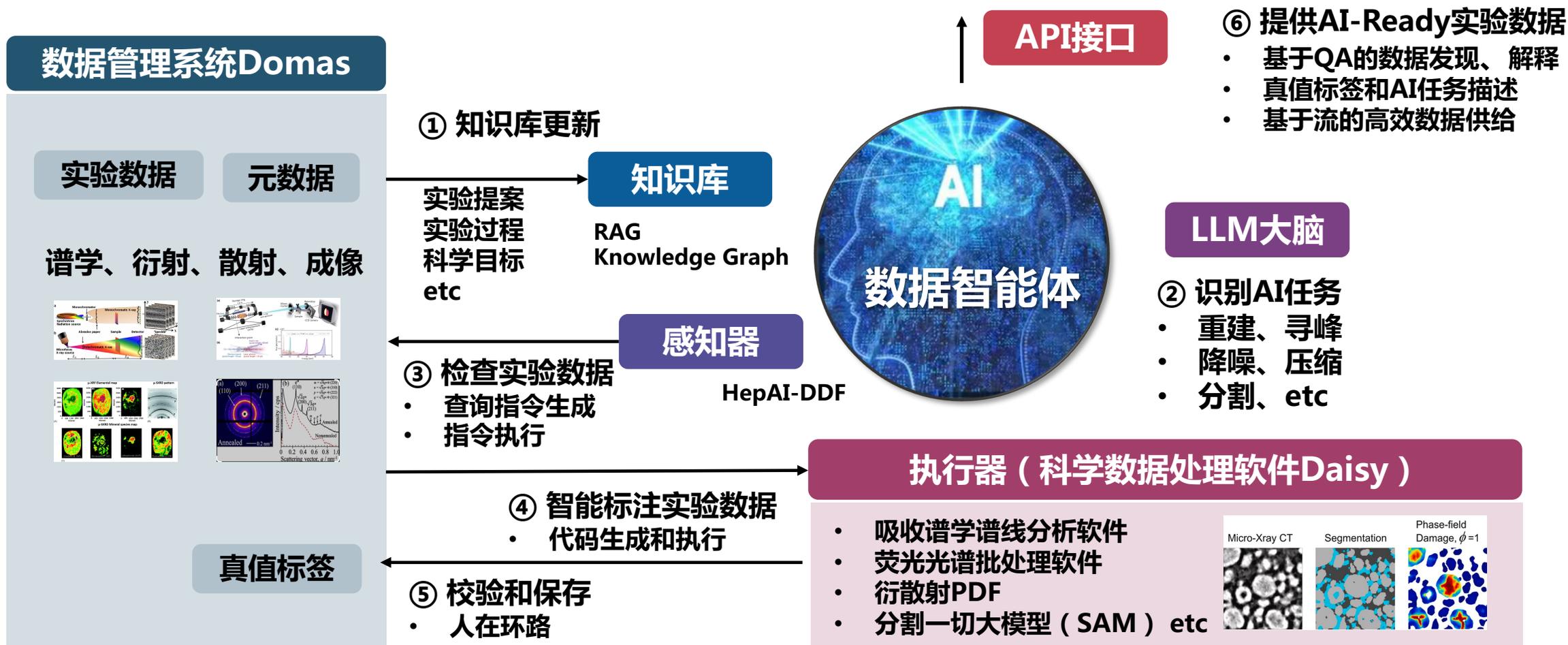


多模态科学数据的接入、解释、标注、融合对齐

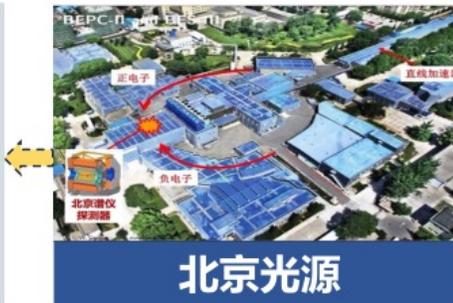




数据智能体



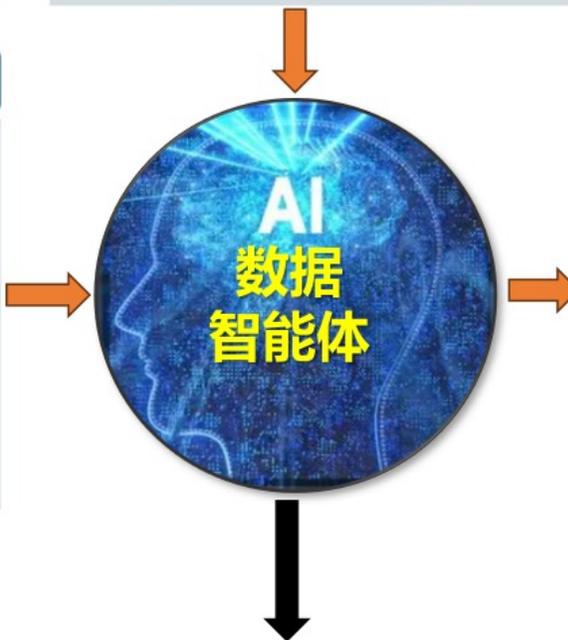
数据智能体：大装置科学数据的信息补全和处理



大装置数据清洗

BSRF

CSNS



智能体根据信息生成数据卡片：

装置信息：

- 名称：北京正负电子对撞机：4W1B-X射线荧光光谱分析实验站
- 装置方法学（实验方法）：基于同步辐射X射线的能量光栅和微区分析方法，采用超环镜和毫微秒级二次聚焦技术（样品处50微米聚焦光斑），结合多通道单色器对态密度固定能量（15.0keV）单色光，实现微区X射线荧光分析（μ-XRF）
- 基本用途：开展微区XRF分析和大型结构样品的元素二维扫描实验，涉及环境科学、生命科学、材料科学、地质科学、古气候、人文考古、医学、物理和国家安全特殊需求等领域

研究对象：

- 研究对象：纳米的物建理化平台-纳米的物安全性和代谢分析方法的研究
- 研究领域：生物学（历史痕迹显示主要研究领域为生物学）
- 研究目的：
- 研究问题：
- 实验方案/研究内容、实验方法：
- 实验方法学：

实验数据文件信息

数据文件储存大小及结构：

```

数据文件总大小: 222.64 GB
/ky-f4f4a-fc87-202408-Data-1884-2830813-8/
├── raw
│   ├── 8-5h-Q25-agg1-P_0_15 (3.00 GB)
│   ├── 8-5hwa_H5 (142.15 GB)
│   ├── 8-5hwa_H48 (142.43 GB)
│   ├── 8h-Q25-agg1-P_0_15 (4.00 GB)
│   ├── 8h-Q25-agg1-P_name_H5 (142.15 GB)
│   └── ... (27个同级别文件)
    
```

正在结合您的OCLIMAX中子散射模拟分析目的进行任务分解...

任务内容

源工具	任务名称	工具
源工具: OCLIMAX (v1.0.0)	分析Neutron的模拟分析中子散射数据方式	oclimax_introduce
源工具: OCLIMAX (v1.0.0)	使用AGU模拟的XRD文件	oclimax_convert
源工具: OCLIMAX (v1.0.0)	模拟中子散射模拟分析中子散射数据方式	oclimax_simulate
源工具: OCLIMAX (v1.0.0)	模拟中子散射模拟分析中子散射数据方式	oclimax_simulate
源工具: OCLIMAX (v1.0.0)	模拟中子散射模拟分析中子散射数据方式	oclimax_simulate

任务轮数 → >>>>> <<<<<<

当前正在执行的任务: [模拟中子散射模拟分析中子散射数据方式](#)

智能体根据处理要求进行进一步处理

我需要重建ID为673180d2c12a0000a30000a9的纳米-XCT imaging的数据,使用Daisy的tomopy reconstruction算法

AI DataAgent

正在提取需要处理的关键信息...

关键信息: 重建ID为673180d2c12a0000a30000a9的纳米-XCT imaging数据,使用Daisy的tomopy reconstruction算法.

数据处理的关键词为: ['重建ID', '673180d2c12a0000a30000a9', '纳米-XCT imaging', 'Daisy', 'tomopy reconstruction']

正在对数据: 673180d2c12a0000a30000a9 匹配合适的处理方式...

正在对数据: 673180d2c12a0000a30000a9 使用<tomography Reconstruction by Daisy>脚本进行处理...

- 多方法学数据清洗
- 多装置远程调用清洗

将缺失或者意义不明的实验数据根据实验背景和方法学信息进行补全，并推荐数据处理方法。

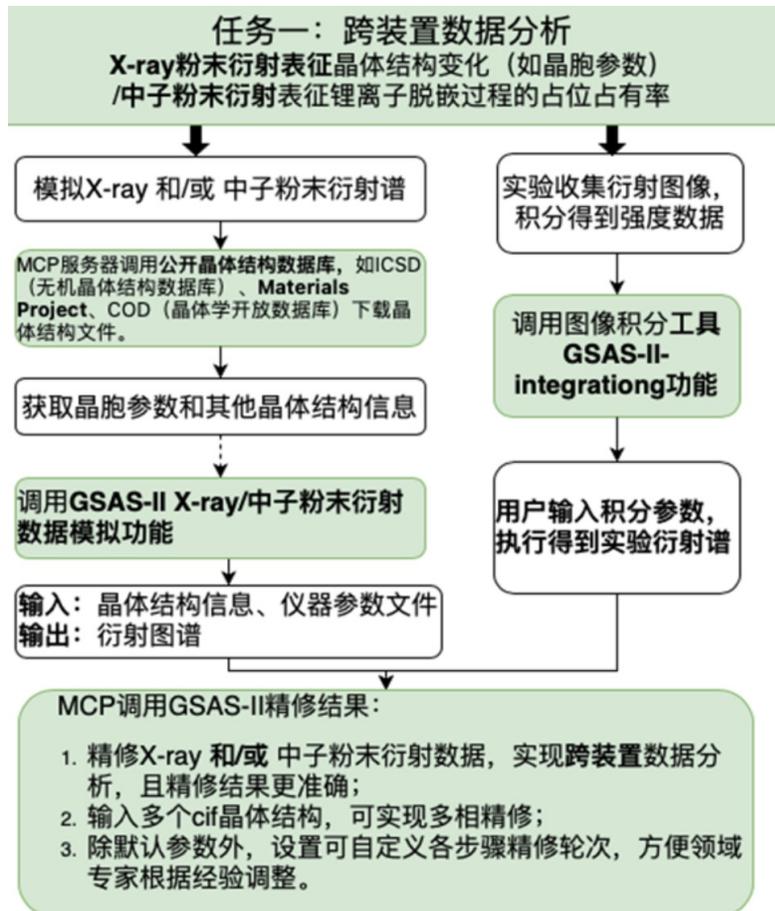
数据文件结构分析

1. 文件名规则解析
 - 时间/日期: 8-5hwa_H5 代表样品处理时间/日期 (0.5小时/小时)
 - 样品标识: CuS 表明研究对象为硫化亚铜材料 (符合“纳米的物建理化平台”课题)
 - 处理条件: agg1 可能指代的数据应用方案编号, P_0_15 可能表示不同处理状态 (如聚焦态和准态)
 - 数据格式: _P 为原始数据, _name 可能为样品/空测文件
2. 数据格式特性

数据智能体：构建跨装置高质量XRD实验和模拟融合数据集



- ◆ X射线衍射（XRD）是一种用于分析材料晶体结构的无损分析技术
- ◆ 初始结构→模拟得到模拟衍射谱→匹配实验衍射谱→精修，多次迭代，直至模拟与实验衍射谱匹配，得到晶体结构
- ◆ 数据智能体模拟人类分析过程，进行数据融合处理



- ◆ 基于Materials Project API的结构查询工具：
 - ◆ 带隙材料搜索接口
 - ◆ 其他接口
- ◆ 基于GSAS-II的XRD分析工具：
 - ◆ 基于CIF的XRD模拟
 - ◆ XRD tif图像积分
 - ◆ Xray和Neutron衍射数据精修
- ◆ 文件操作工具：
 - ◆ 文件查询
 - ◆ 文件操作
- ◆ Picgo图床工具





◆ 对晶体结构进行X射线衍射模拟结果：

输入晶体结构： 请直接对当前目录下的PbSO4的CIF文件进行XRD模拟

OI SCI-DataAgent Today at 14:44

SCI_DataAgent发言:

思考用时 19 秒

用户需要对PbSO4的CIF文件进行XRD模拟，需调用gsasii_xrd_process_agent的xrd_simulate_gsasii_post功能。已确认CIF文件路径为：/aifs/user/home/xiongdb/Vsproject/mutil_mcp_tools/files/thread_22578926-f5e3-48ef-873b-13a8fe7ca3e4/XRD_Neutron_power_diffraction/PbSO4-Wyckoff.cif。需要规划具体参数和执行流程。

由 planner_agent 进行接下来的操作，现在由 planner_agent 接手。

planner_agent发言:

思考用时 40 秒

任务计划如下:

| --- | --- | --- |

| 任务专家 | 任务内容 | 前提条件 |

| gsasii_xrd_process_agent | 调用GSAS-II的XRD模拟功能对PbSO4-Wyckoff.cif进行XRD图谱模拟 | |

gsasii_xrd_process_agent发言:

X-Ray Diffraction Simulation Results

| --- | --- |

| Item | Value |

| status | success |

输出模拟XRD谱

| project_file | /aifs/user/home/xiongdb/Vsproject/mutil_mcp_tools/files/thread_22578926-f5e3-48ef-873b-13a8fe7ca3e4/XRD_Neutron_power_diffraction/PbSO4_XRD_inst_d1a/PbSO4_XRD.gpx |

| data_file | /aifs/user/home/xiongdb/Vsproject/mutil_mcp_tools/files/thread_22578926-f5e3-48ef-873b-13a8fe7ca3e4/XRD_Neutron_power_diffraction/PbSO4_XRD_inst_d1a/PbSO4_XRD_data.csv |

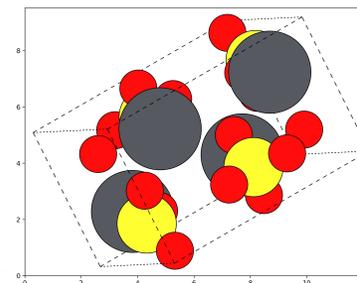
| reflections_file | /aifs/user/home/xiongdb/Vsproject/mutil_mcp_tools/files/thread_22578926-f5e3-48ef-873b-13a8fe7ca3e4/XRD_Neutron_power_diffraction/PbSO4_XRD_inst_d1a/PbSO4_XRD_refl.csv |

| plot_file | /aifs/user/home/xiongdb/Vsproject/mutil_mcp_tools/files/thread_22578926-f5e3-48ef-873b-13a8fe7ca3e4/XRD_Neutron_power_diffraction/PbSO4_XRD_inst_d1a/PbSO4_XRD_.png |

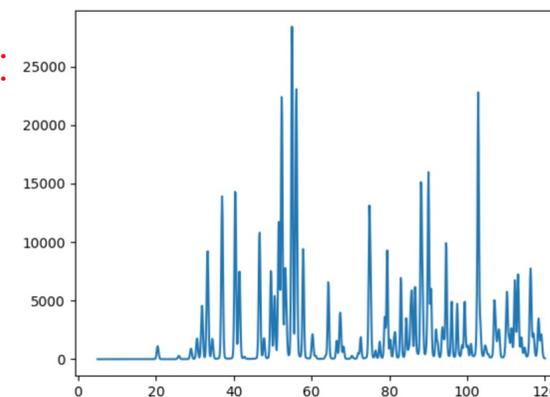
| error | None |

思考过程

输入展示：



输出展示：



Plan_Checker发言:

思考用时 10 秒

TERMINATE

□ AI-Ready数据：晶体结构与XRD谱模拟数据。

数据智能体:构建跨装置高质量XRD实验和模拟融合数据集



◆ X射线和中子粉末衍射实验数据精修实例：

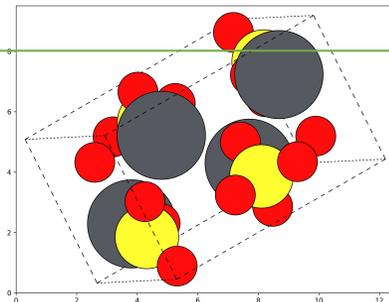
输入：X射线、中子粉末衍射实验数据和仪器参数、晶体结构

gsasii_xrd_process_agent发言：

已完成2次晶胞参数精修，结果如下：

精修质量指标：

1. X射线数据 (PWDR PBSO4.XRA Bank 1) :
 - Rwp值: 14.009 (较初始值降低约12%)
2. 中子数据 (PWDR PBSO4.CWN Bank 1) :
 - Rwp值: 4.715 (达到高精度收敛)



精修后晶胞参数 (Phase_1) :

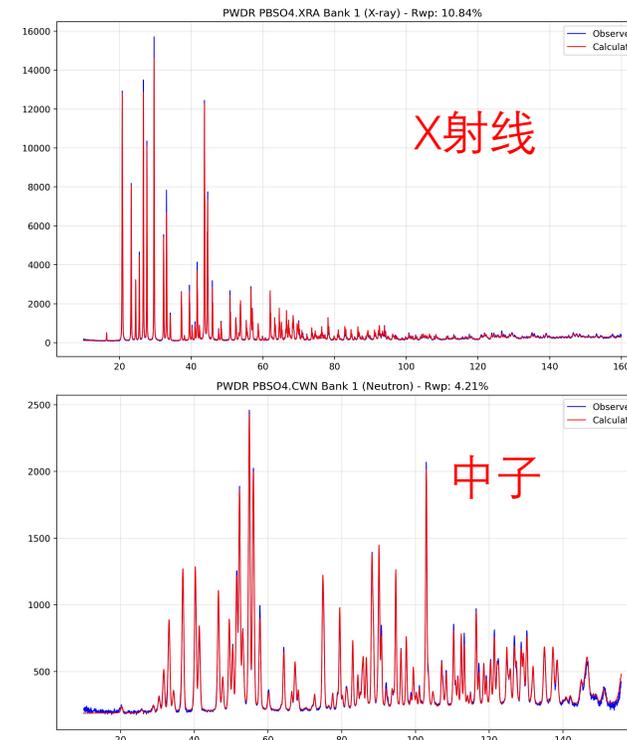
参数	长度 (Å)/角度 (°)	物理意义
a轴	8.4718	晶格沿x方向基本矢量
b轴	5.3932	晶格沿y方向基本矢量
c轴	6.9538	晶格沿z方向基本矢量
体积	317.72 Å ³	单胞空间占位量
晶系	正交晶系	$\alpha=\beta=\gamma=90^\circ$

需要我为您：

1. 生成GSAS-II精修报告 (PDF格式)
2. 绘制XRD实验-计算拟合对比图
3. 输出精修后CIF文件
4. 继续执行其他参数精修 (如原子位置/温度因子)

输出：X射线、中子粉末衍射模拟数据、精修后晶体结构

TERMINATE



□ 高质量融合数据：晶体结构、X射线/中子粉末衍射模拟数据、X射线/中子粉末衍射实验数据。



□ 根据不同AI任务，生成AI-Ready数据集

- 实验衍射谱+设备参数，预测晶体结构
- 晶体结构，预测模拟衍射谱
- . . .

Outline

01 高能物理大数据和AI4HEP

02 高能物理/大装置科学数据体系

03 高质量科学数据集构建思考

04 总结与展望

AI-Ready数据构建要素



□ AI-Ready数据集：已经过整理和预处理，适合直接用于训练和测试人工智能模型的数据集。

□ AI-Ready数据集的四要素



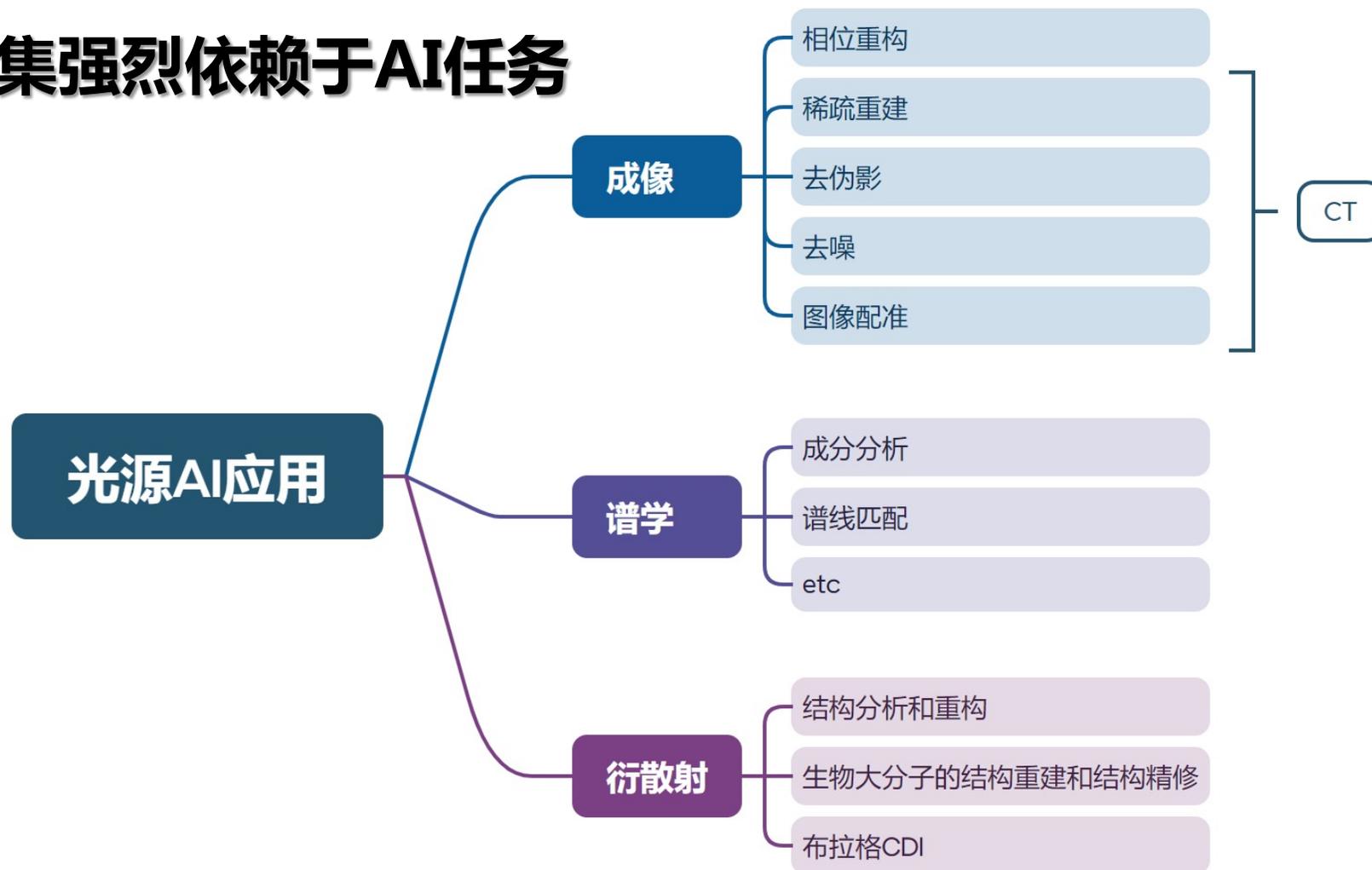
□ AI-Ready数据集强烈依赖于任务

AI-Ready数据构建



AI-Ready数据集强烈依赖于AI任务

- 一级任务：
- 二级任务：
- 三级任务？



AI任务需要输入什么、输出什么，从而反向思考数据怎么组织？

AI Ready数据四级成熟度模型

- 通过详细描述每个等级的特征，为数据集的AI-Readiness提供一个可快速评估的框架
- 主要从数据一致性、数据访问、元数据三个角度来设计考虑

级别：0 (Not AI-Ready)

数据一致性 未进行内部一致性的检查

数据访问 仅通过请求或订单系统对公众开放使用

元数据 没有数据字典可用，或数据字典不是机器可读格式（例如PDF）

级别：1 (Minimal)

数据一致性 手动一致性检查

数据访问 只有一个非程序化的访问选项，例如文件下载

元数据 数据字典是机器可读格式（例如CSV、XML、JSON）

级别：2 (Intermediate)

数据一致性 一致性检查部分自动化，部分结果有文档记录

数据访问 多种访问方式，且至少包括一个可编程访问方法（API）

元数据 数据字典使用机器可读的元数据标准

级别：3 (Optimal)

数据一致性 完全自动化的内部一致性检查和报告；且考虑了与社区数据集的外部一致性

数据访问 除多种交付选项（下载、API）外，还基于云和高性能计算提供的数据即服务（DaaS）

元数据 机器可读的元数据标准；参数与行业标准或其他机构数据集标准协调一致

数据智能体：加速数据底座构建



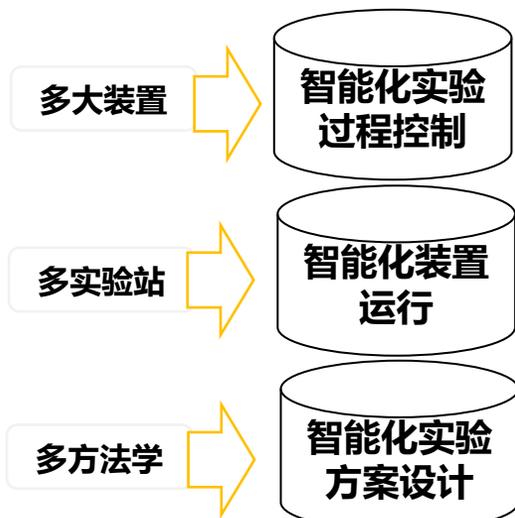
AI4Data

Data4AI

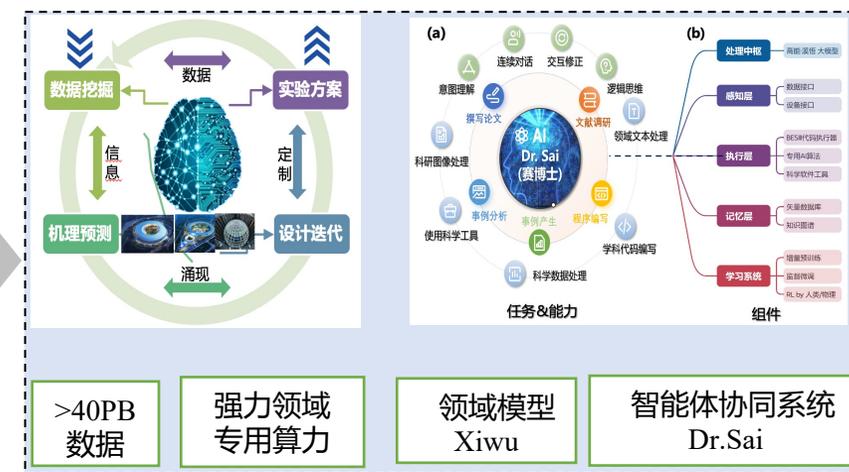
实验智能

数据智能

科研智能



数据软件框架及工具 科学数据融合和标准化



智慧型大科学装置

AI 驱动的科学发现系统

领域科研活动全生命周期





- AI加速科学研究范式的向**智能范式**变革
- 高能物理领域，目前充分利用大模型技术的两条思路
 - AI大模型 → 智能体 → 多智能体：学会复杂的物理分析过程，显著提升高级劳动的效率，促进新粒子发现
 - 探索科学数据大模型：从数据驱动的角度，理解数据、发现规律
- 新的模式对数据体系提出了新的要求
 - 数据策略和标准是基础，从数据采集的源头保障数据质量
 - 数据存储与管理解决数据海量和多元的问题，数据分析软件解决专业性强的问题
 - 数据智能体，通过多种工具构建AI-Ready数据，面向AI供给数据。



~~Thank you for listening~~



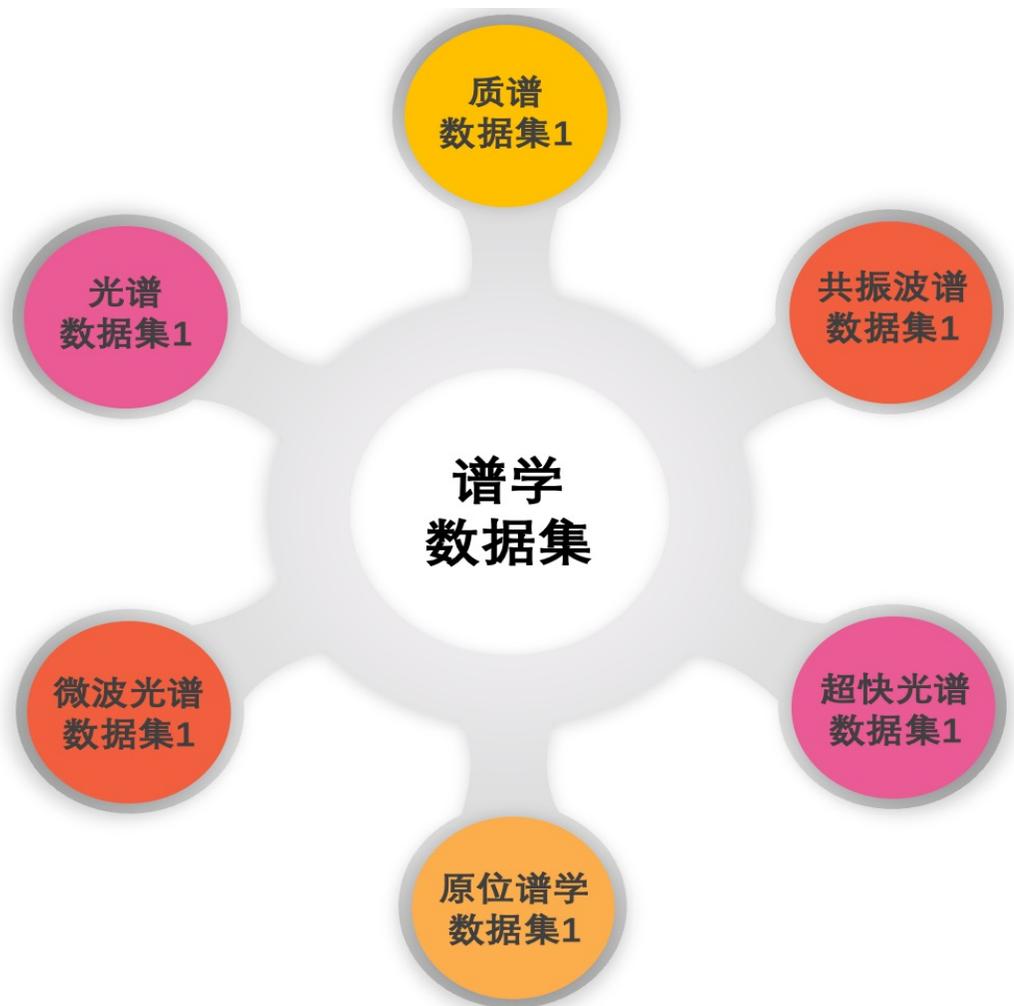
Backup

推动形成国家级（领域级）数据库/集



分布式数据库/数据集建设

逻辑集中统一、按规则访问和利用



基于智能体的AI-Ready数据构建方案



□ 数据融合智能体



大装置数据体系



大装置数据体系是大科学装置中数据的采集、存储、处理、分析和使用的整体框架和机制。它涵盖数据从生成到最终使用的整个生命周期，并确保数据的质量、完整性、安全性和可用性。

