AiDD AI+ 软件研发数字峰会
AI+ software Development Digital summit

AI驱动软件研发
全面进入数字化时代

中国·北京 08.18-19

AI+
software
Development
Digital
summit

人工智能工程化软件研发

龙明盛 清华大学

科技生态圈峰会 + 深度研习 ——1000 + 技术团队的共同选择

K+峰会

2023K+
全球软件研发行业创新峰会
上海站
会议时间 | 06.09-10

2023K+
全球软件研发行业创新峰会
北京站
会议时间 | 07.21-22

2024K+
全球软件研发行业创新峰会
深圳站
会议时间 | 05.17-18

K+峰会详情

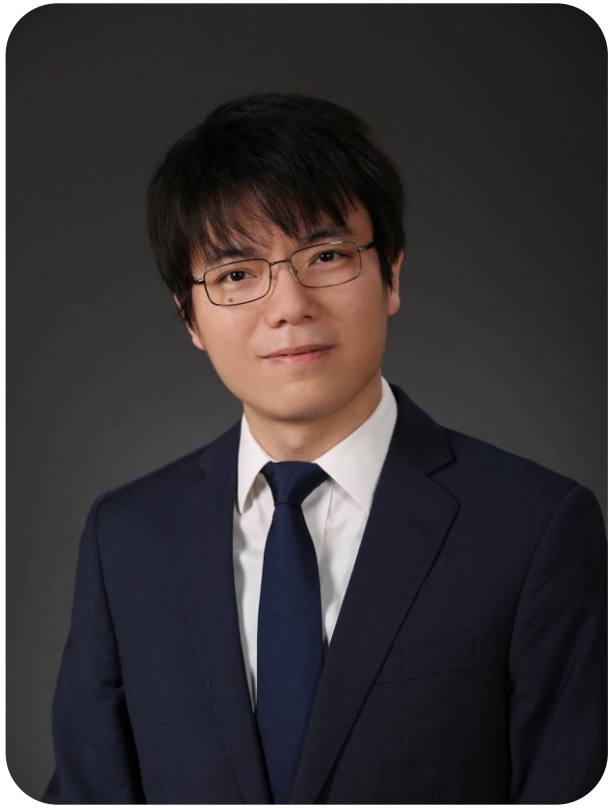
AiDD峰会

会议时间 | 08.18-19
AiDD AI+软件研发数字峰会
北京站

会议时间 | 11.17-18
AiDD AI+软件研发数字峰会
深圳站

AiDD峰会详情

演讲嘉宾



龙明盛

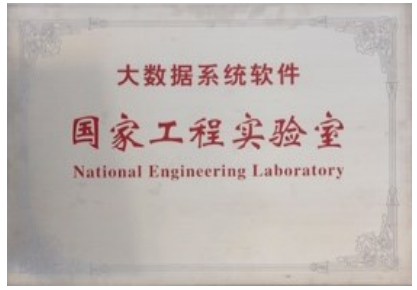
清华大学软件学院长聘副教授

清华大学长聘副教授、软件学院机器学习研究组负责人，国家优秀青年科学基金获得者，入选北京市科技新星和清华大学良师益友。主要研究领域为机器学习理论、算法与模型，专注于迁移学习、深度学习、科学学习及其在自然科学和软件工程中的应用。以第一或通讯作者发表Nature正刊/子刊和JMLR、TPAMI、ICML、NIPS、ICLR论文40余篇，谷歌引用2.6万次，三篇论文入选ICML和NIPS最具影响力论文。担任ICML、NIPS、ICLR、ICCV和CVPR（资深）领域主席，TPAMI和AIJ编委。获教育部技术发明一等奖、北京市科技进步一等奖、IJCAI-FTL时间检验奖，入选机器学习全球高影响力学者、爱思唯尔中国高被引学者、全球前2%科学家。

目录
CONTENTS

PART 01
团队介绍

团队介绍



团队带头人：
王建民教授

“清华数为”大数据系统软件团队是专注工业大数据系统软件的科研与工程团队。
团队先后研制了工业物联网时序数据库IoTDB、低代码开发工具DWF等产品，覆盖工业数据采集、管理、处理、分析与应用全生命周期。

工　业　数　据　软　件

APACHE IoTDB　　Anylearn　　FloK　　DWF

数据管理　　机器学习　　数据处理　　应用开发

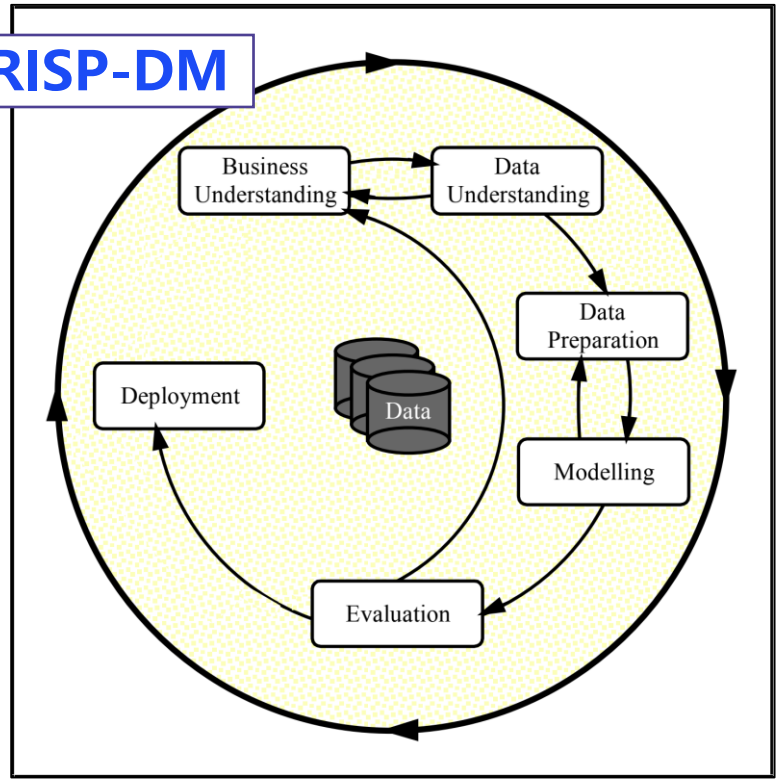
PART 01

引言：人工智能工程化方法论

人工智能工程化方法论

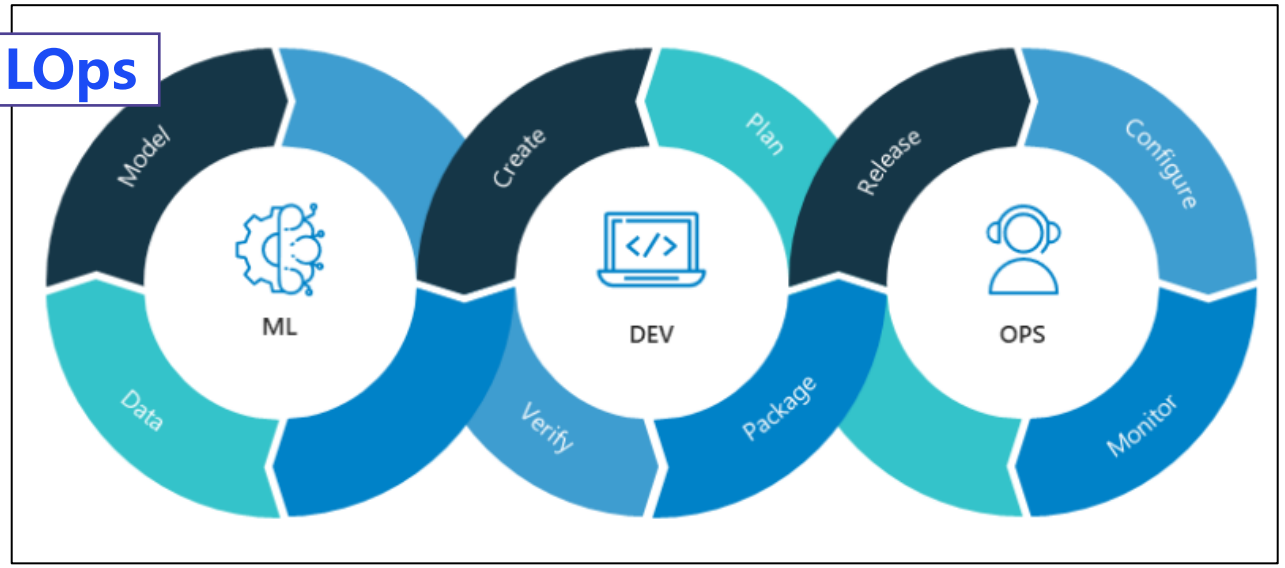
在数据、模型和服务层面上的**持续迭代**是智能软件研发的关键词。而人工智能工程化的核心在于**标准化**的研发流程和管理方法，通过体系化、规模化的分工协作和资产综合治理，提高研发**质量**和应用落地**效率**，推动人工智能技术为产业**持续赋能**。

CRISP-DM



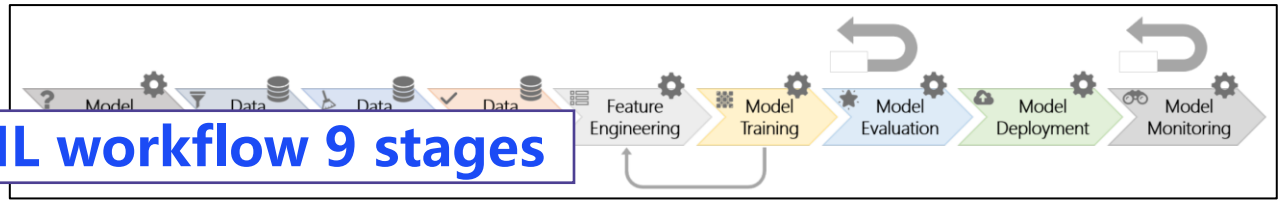
[5]

MLOps



[4]

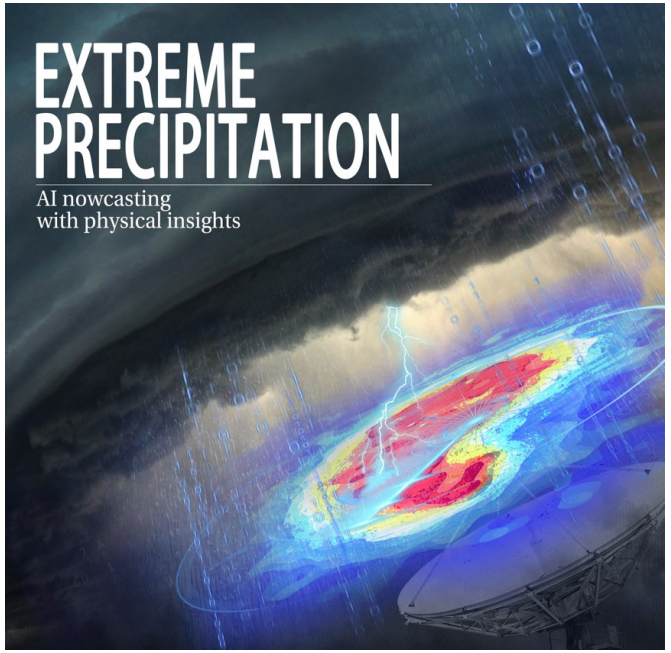
ML workflow 9 stages



[1]

PART 02
人工智能大模型研发案例

人工智能气象领域大模型研发成果



NEWS AND VIEWS | 05 July 2023

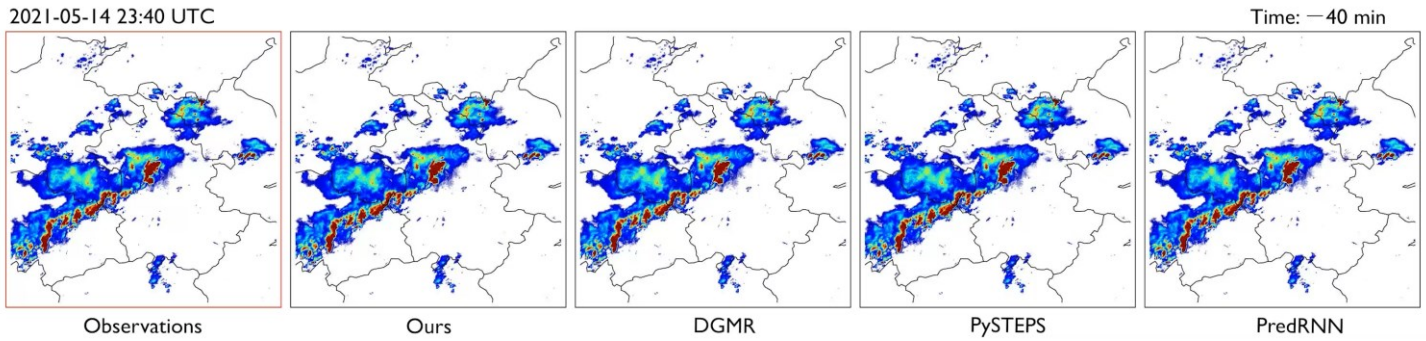
The outlook for AI weather prediction

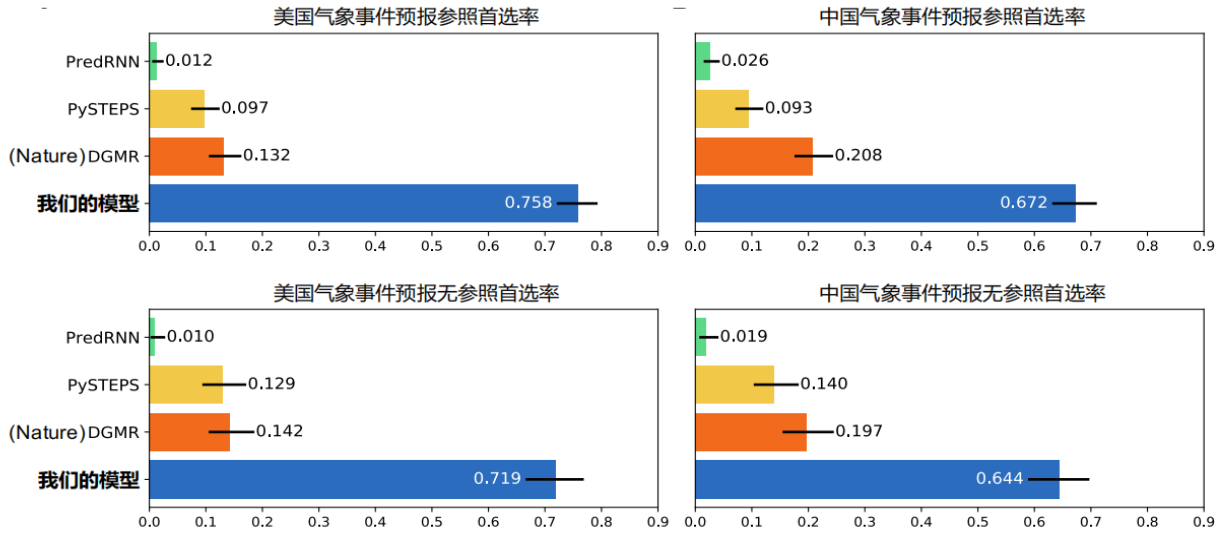
Two models demonstrate the enormous potential that artificial intelligence holds for weather prediction. But the risks involved demand that meteorologists learn to design, evaluate and interpret such systems.

NowcastNet

短临极端降水预测大模型

Nature正刊&专题报道[11]



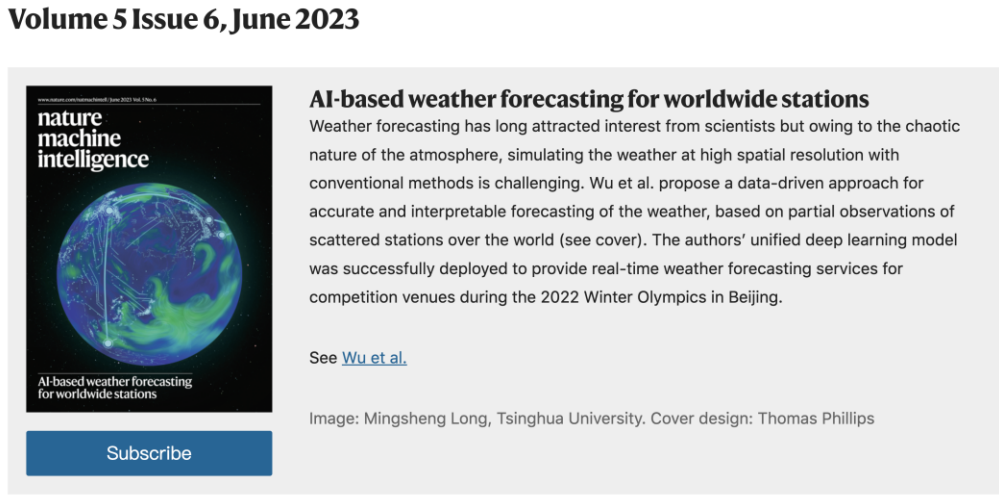


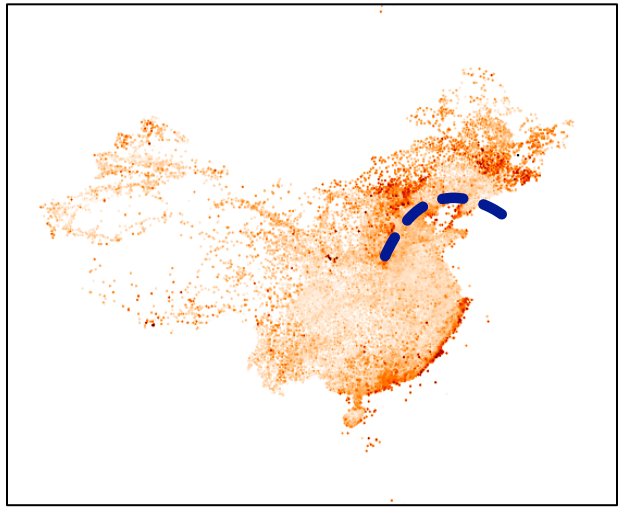
✓**通过全国62位一线预报员评测，性能大幅超过DeepMind**

✓**中国气象局业务系统（SWAN3.0）上线**

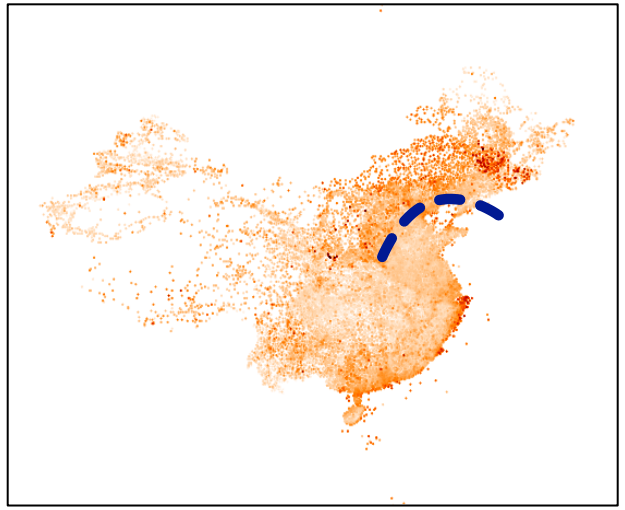
人工智能气象领域大模型研发成果

- ✓Corrformer短期气象预报大模型 [8]
- ✓首个**全球**自动站协同预报大模型
- ✓完成全球数万台自动气象站预报仅需**1秒**
- ✓入选Nature子刊（NMI）封面

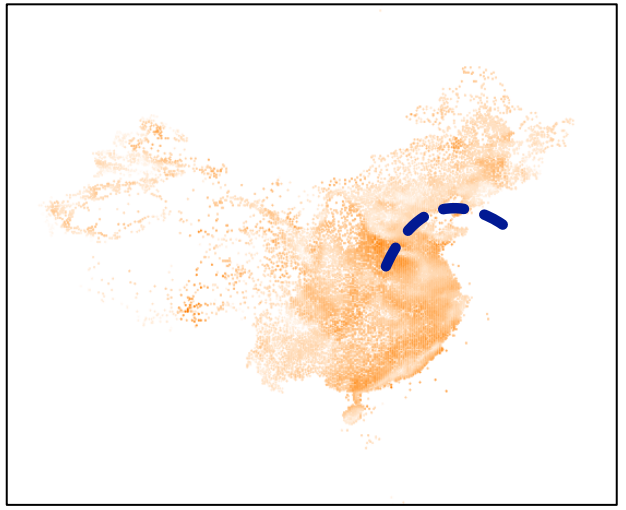




未来24小时真实观测



Corrformer预报结果



欧洲数值模式预报结果

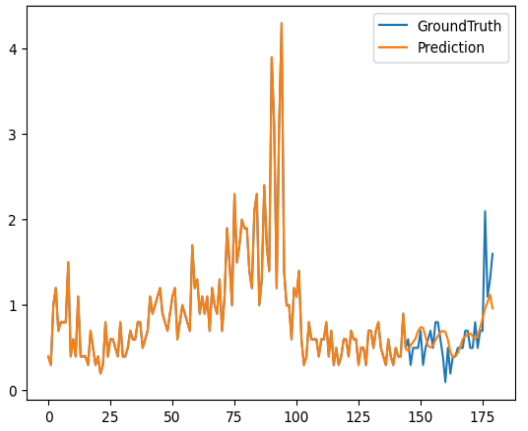
人工智能气象领域大模型研发成果

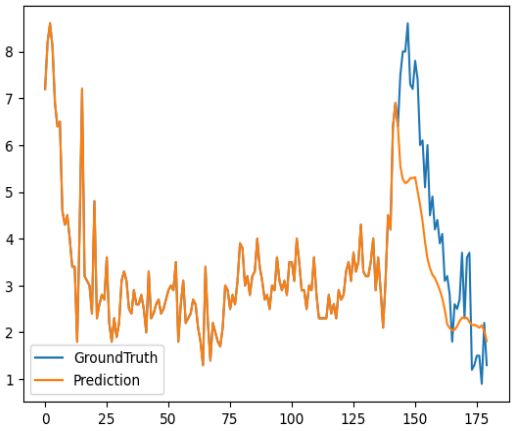
> “北京冬奥会是展现国家形象、促进国家发展、振奋民族精神的重要契机。”
>
> ——习近平

- ✓ Autoformer短期时序预测大模型[7]
- ✓ 唯一分钟级预报产品并在26站实时运行
- ✓ 误差比数值模式降低23%
- ✓ 获国家气象中心科技进步一等奖

为2022北京冬季奥运会提供场馆风速、温度预报，助力赛程规划、运动员备战，为北京冬奥会顺利开展发挥重要支撑作用。

实现基于实时气象观测的10分钟级风速、温度预报。在2022北京冬奥会场馆平均风速预报中，比主流数值预报误差下降23%，补齐了气象实时预报方面的短板。

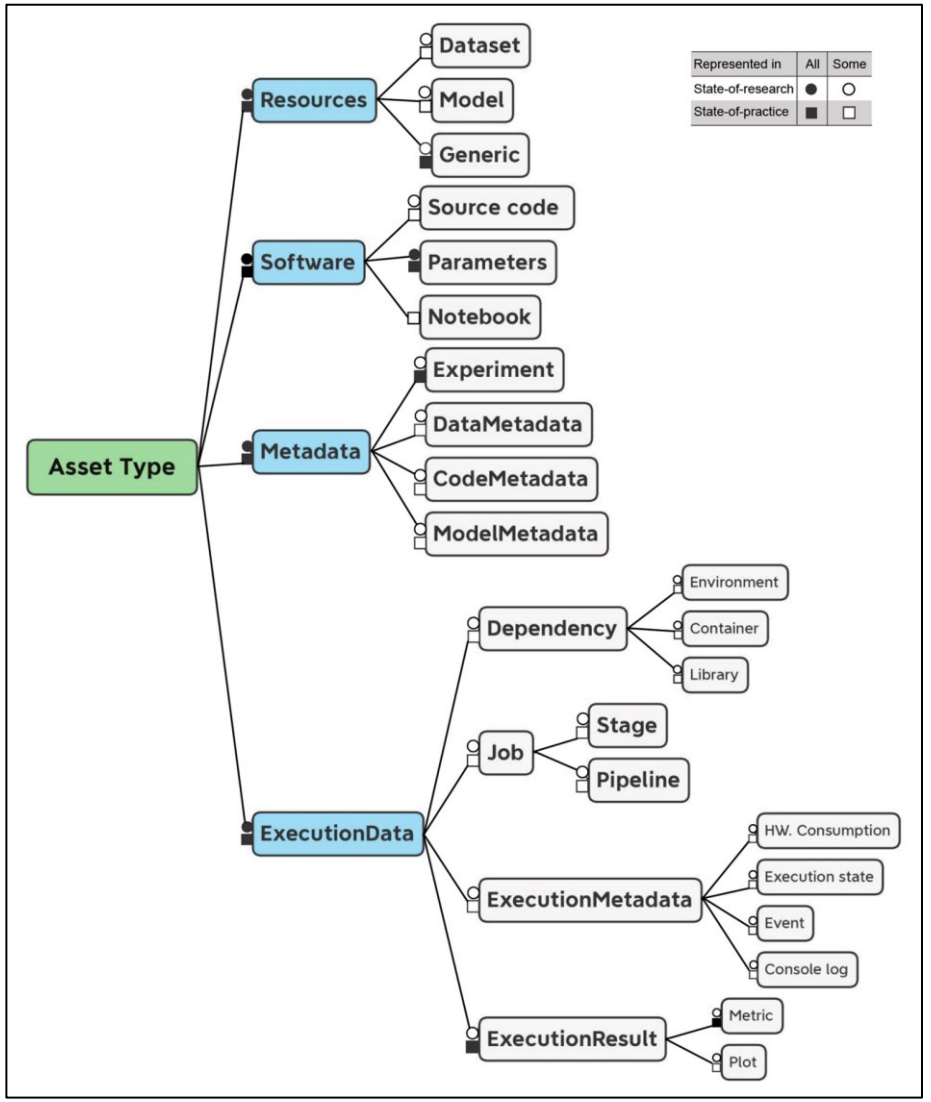




PART 03
人工智能大模型研发挑战

大模型时代人工智能工程化挑战

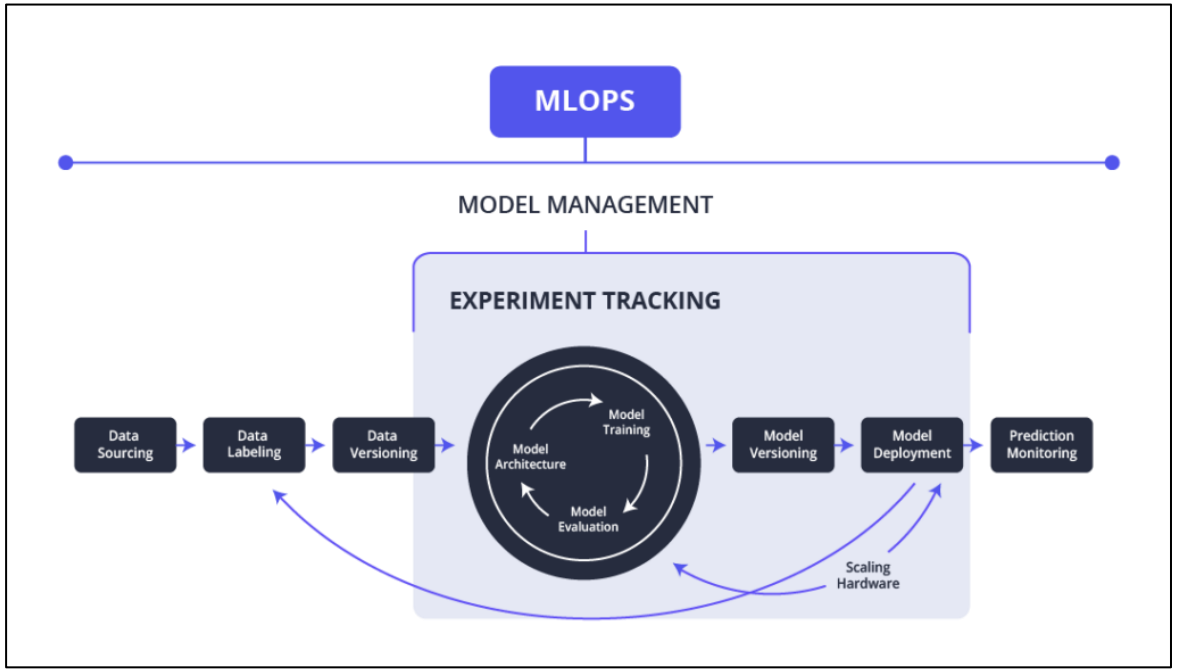
- **研发过程中的产物必然形成大量资产**
  - 如何管理与追溯?
  - 如何复用?
- **大模型的研发资产**
  - 近百TB处理后的数据集
  - 近千个模型参数文件
  - ……
- **缺乏集约式的存储和管理**
  - 数据集、预训练模型等资产碎片化
  - 资产间难以形成有机关联
  - 人员难以形成资产意识、重复造轮子



[2]

大模型时代人工智能工程化挑战

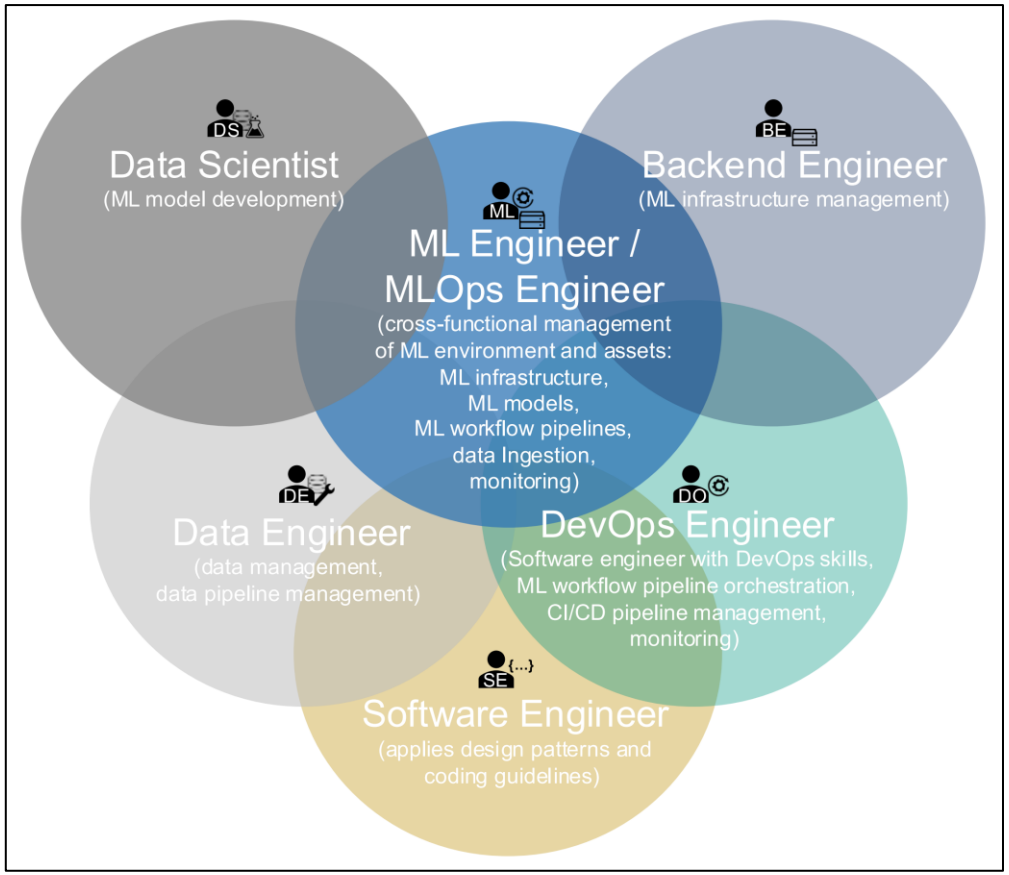
- **前人工作复现和新方案研发必然涉及大量迭代实验**
  - 如何记录与对比?
  - 如何分析与改进?
- **大模型的研发迭代**
  - 近万次大大小小的实验
  - 数千次算法代码变更
  - 几千份case study结果
  - ……
- **研发过程缺乏顶层设计**
  - 实验记录难以保证全面、客观
  - 经验和知识难以沉淀



[10]

大模型时代人工智能工程化挑战

- **研发工作必然由多人团队协作开展**
  - 如何组织与分工?
  - 如何沟通与汇报?
- **大模型的研发团队**
  - 项目管理、方案设计、前人工作复现
  - 数据收集、清洗、转换
  - ……
- **缺乏组织和共享机制**
  - 依赖“人传人”
  - 难以形成有效的沟通
  - 进度管理难以透明化



[3]

PART 04

清华数为Anylearn系统介绍

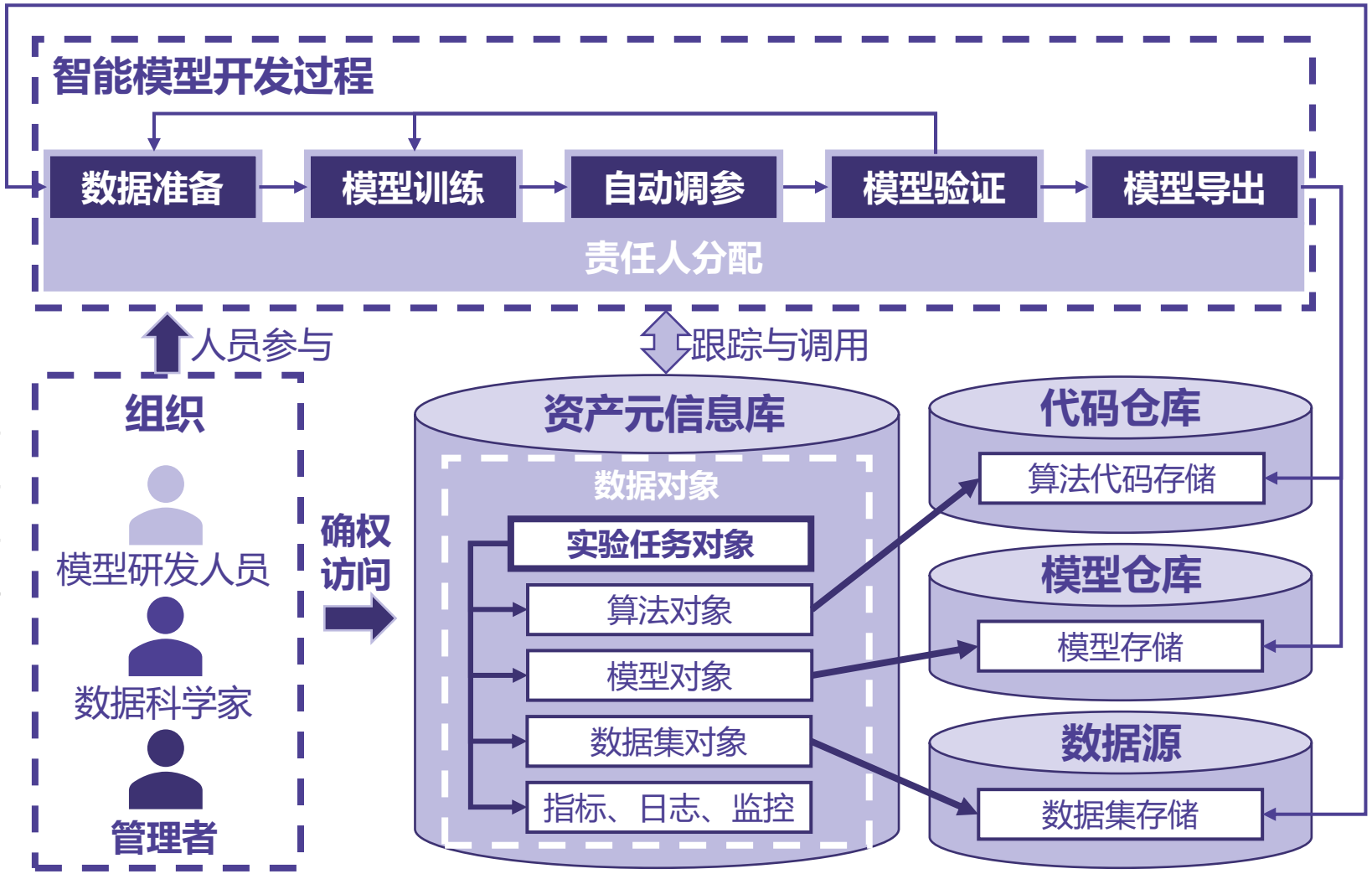
清华数为Anylearn

清华数为Anylearn是一款大数据机器学习研发管理系统。支持数据集、算法族、模型库等资产管理，支持机器学习研发过程管理、知识沉淀、模型迁移，满足资源统筹利用、团队高效协作等人工智能工程化需求。

Anylearn核心概念体系

- **标准化**机器学习资产管理与开发过程管理体系

高效率
可追溯
可复现

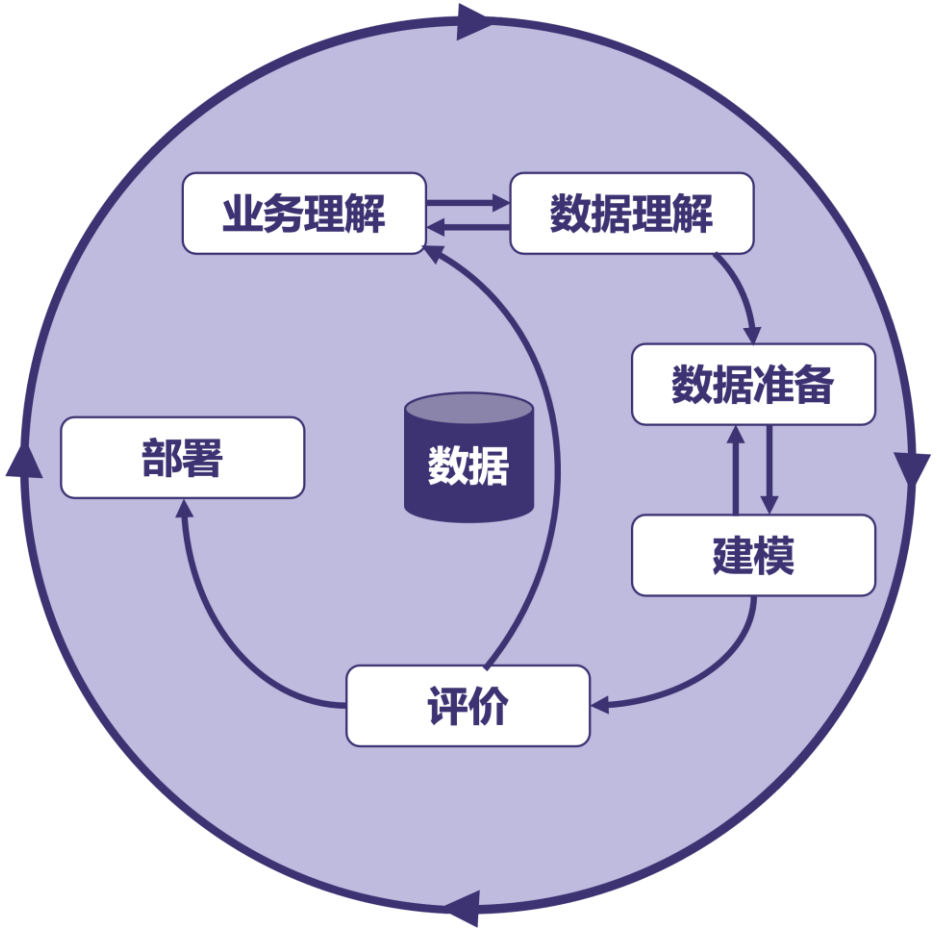


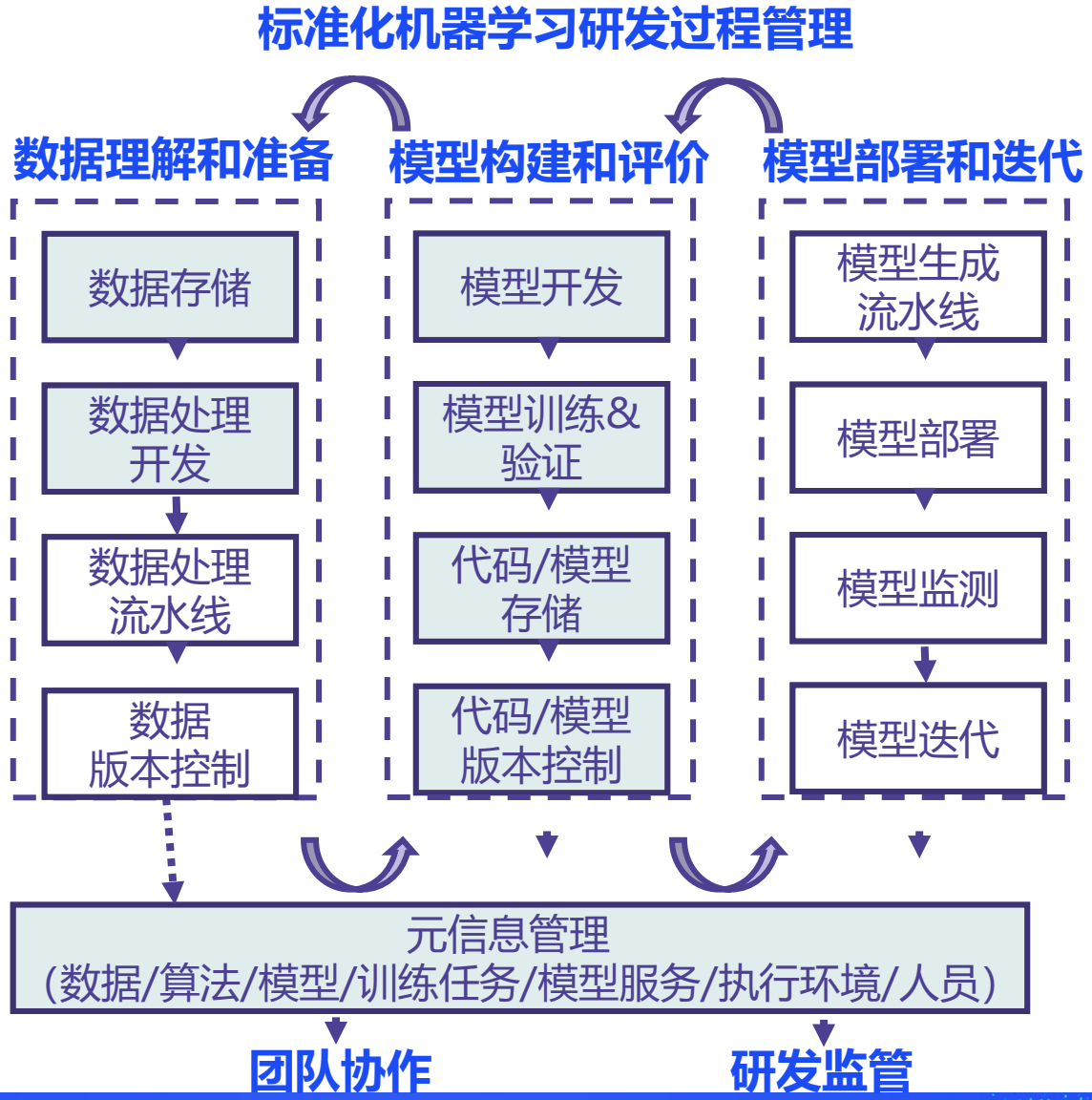
过程管理
- 组织关系管理
- 过程分工管理
- 团队项目管理
- 用户权限管理
- 可视化图表
- 研究报告

资产管理
- 资产关联查询
- 算法版本管理
- 模型版本管理
- 模型血统
- 数据版本管理
- 数据血统

Anylearn核心概念体系

- 方法论 → 实践 → 规范工具

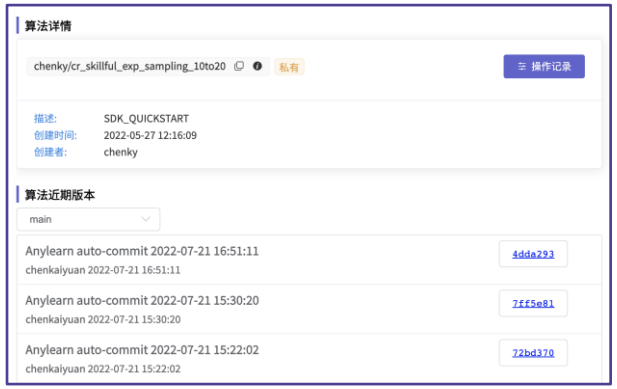




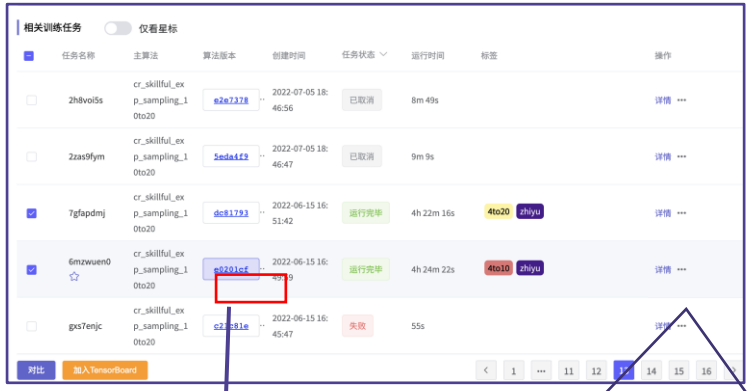
Anylearn机器学习研发管理

- **可追溯、可搜索、可对比、可复现**

算法

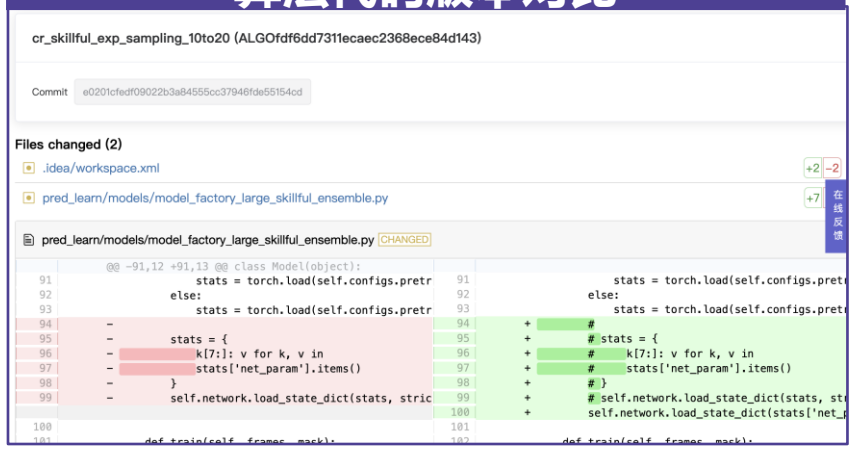


算法关联训练任务



训练输出模型

算法代码版本对比



训练元信息（超参数、执行环境）

Anylearn线上系统运行情况

建成高可用的GPU算力集群，部署Anylearn机器学习研发管理系统线上公开长活环境，共享池化多种类异构GPU，稳定支撑了多个人工智能项目的研发工作与多次教学任务（公网访问地址：https://anylearn.nelbds.cn/）。

- **主要用户**
  - 大数据系统软件国家工程研究中心
  - 清华大学校内师生
- **自2021年8月上线以来**
  - 累计增加共享数据集413个超100TB
  - 累计用户算法代码库7523个
  - 累计形成共享模型1029个
  - 累计训练任务数量5万余次
  - 累计执行训练时间超30万小时

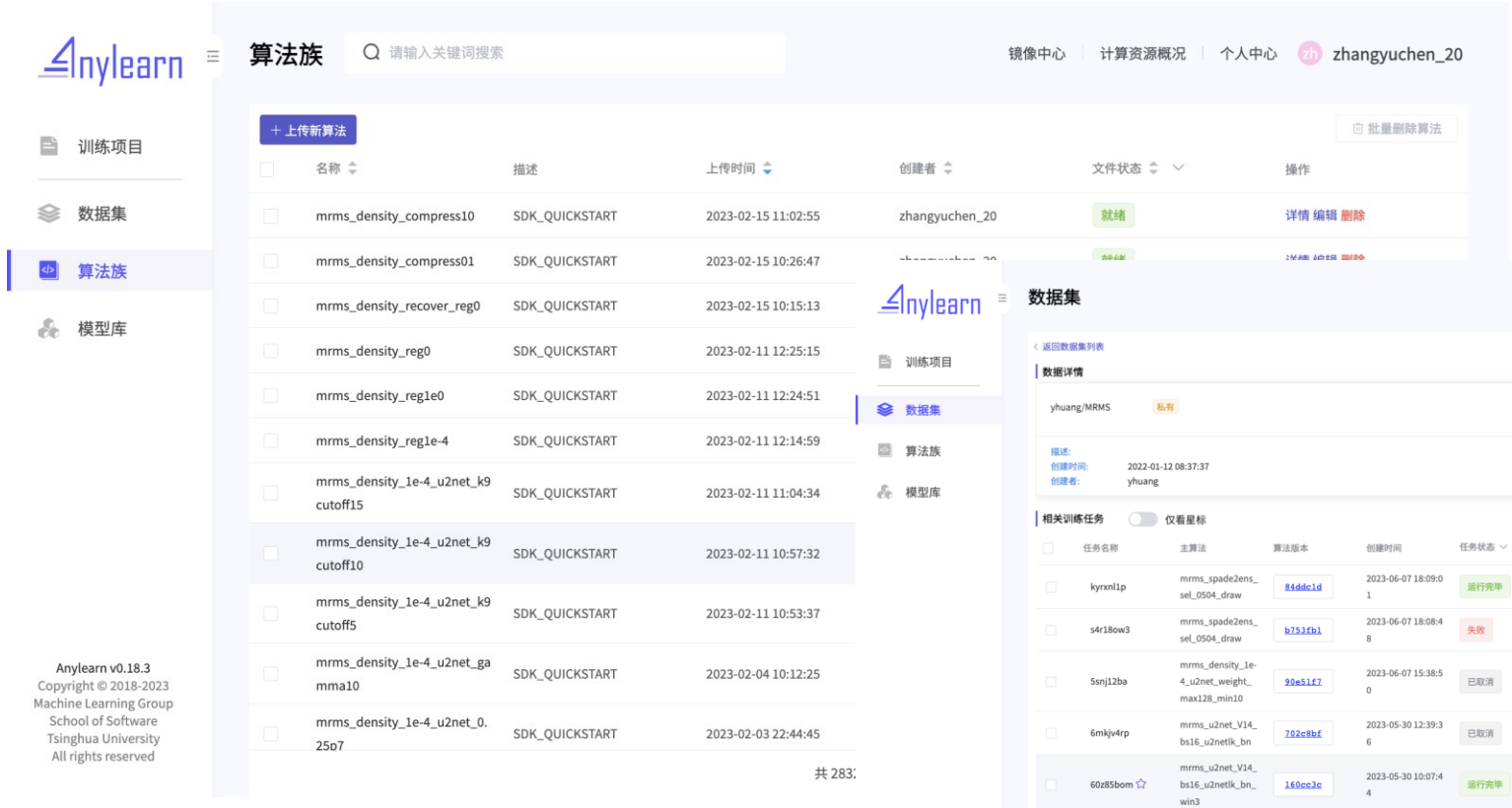
- **支撑人工智能研发工作**
  - 雷达回波外推基础模型研究
  - 冬奥风速预测模型研究
  - 制造任务智能调度方法研究
  - 新能源风速预测模型研究
- **支撑人工智能教学任务**
  - 《深度学习》作业平台2学期共74人
  - 《大数据基础》教学平台25人
  - 《软件工程实践与探索》训练平台10人

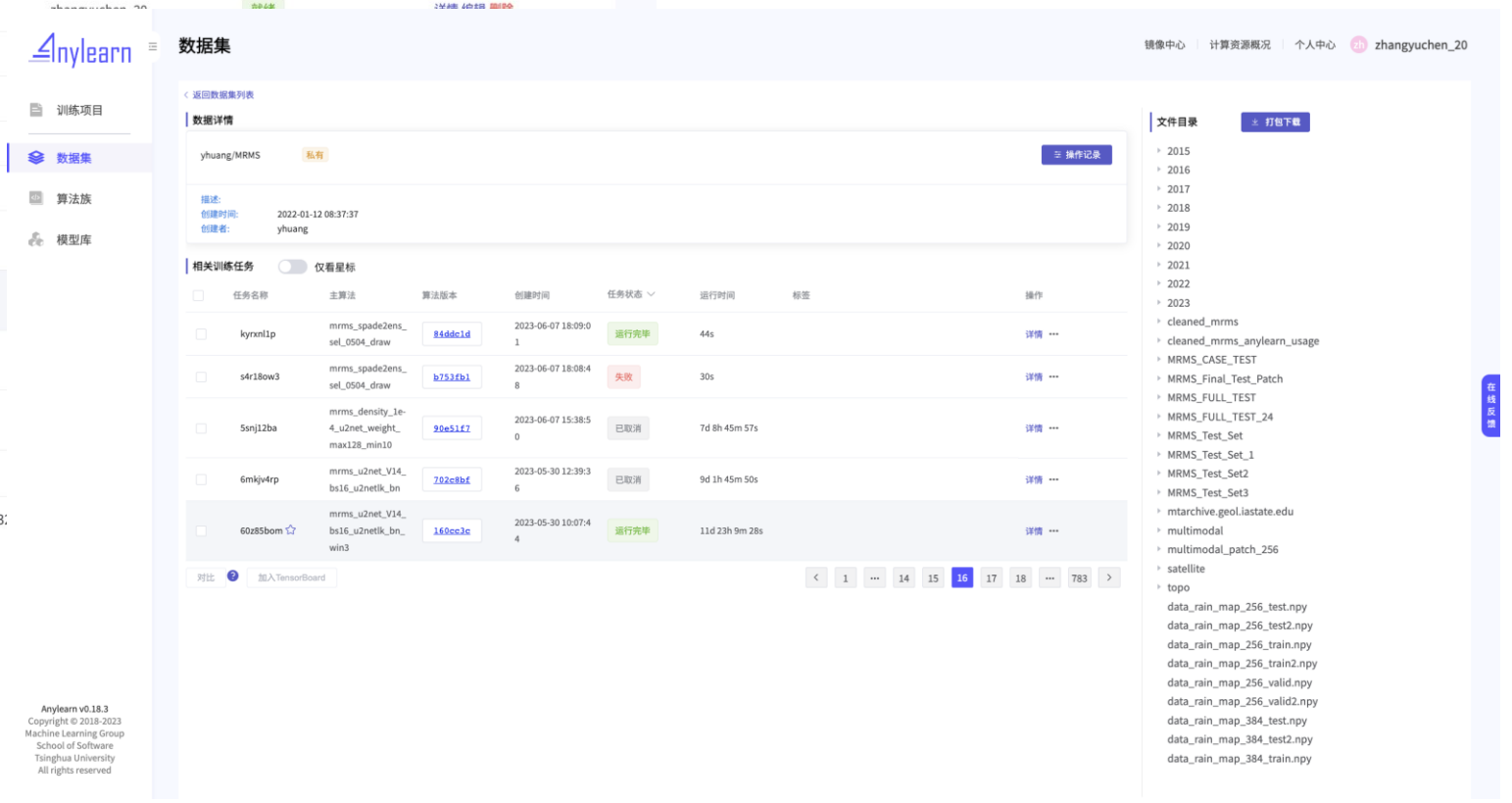
Anylearn线上系统运行情况

建成高可用的GPU算力集群，部署Anylearn机器学习研发管理系统线上公开长活环境，共享池化多种类异构GPU，稳定支撑了多个人工智能项目的研发工作与多次教学任务（公网访问地址：https://anylearn.nelbds.cn/）。

PART 05
清华数为Anylearn 对人工智能大模型研发的支撑

Anylearn助力领域大模型研发



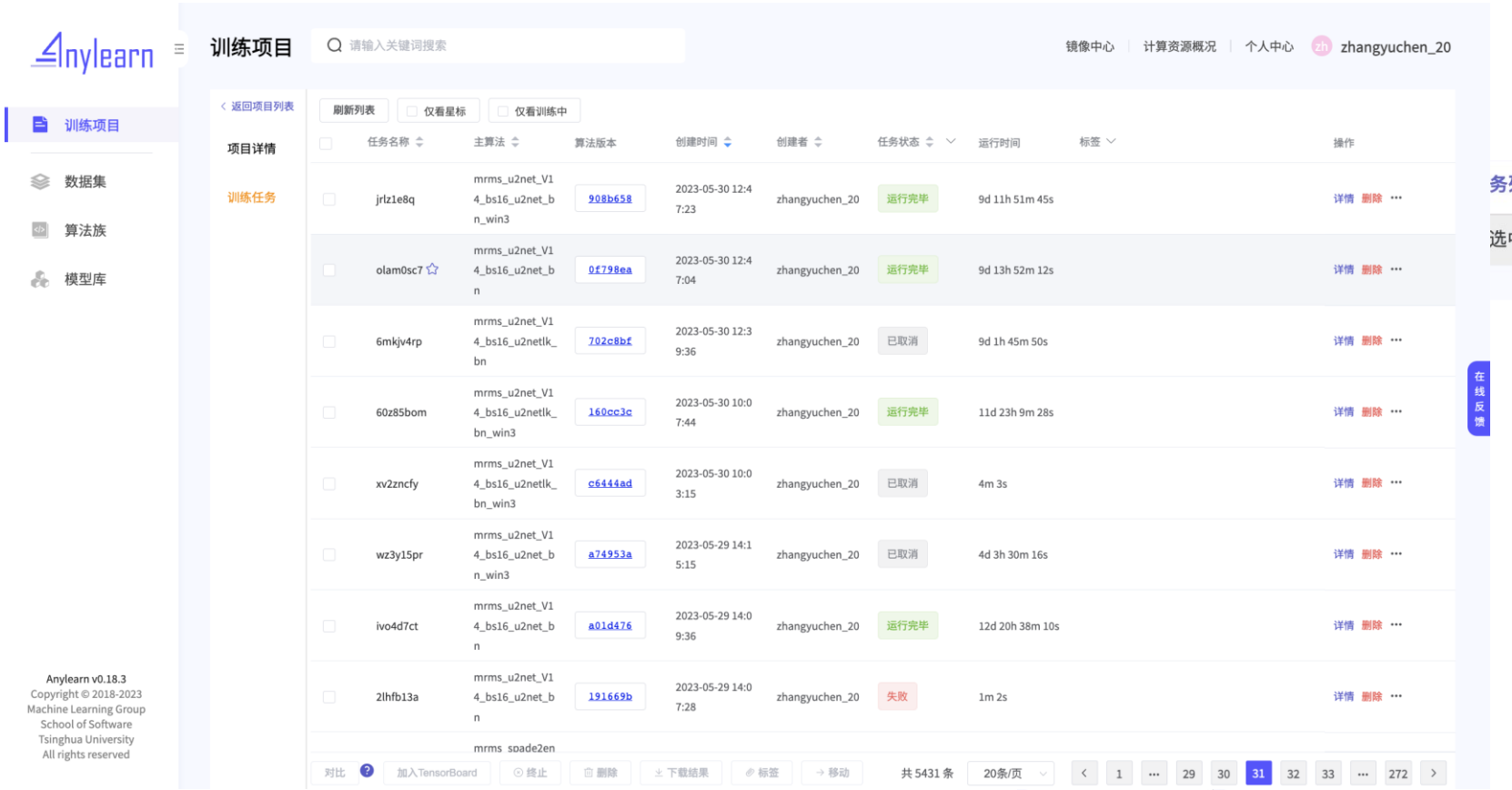


✓资产通过训练任务形成有机关联

✓几千次复现和方案

✓记录每一次代码变更

Anylearn助力领域大模型研发

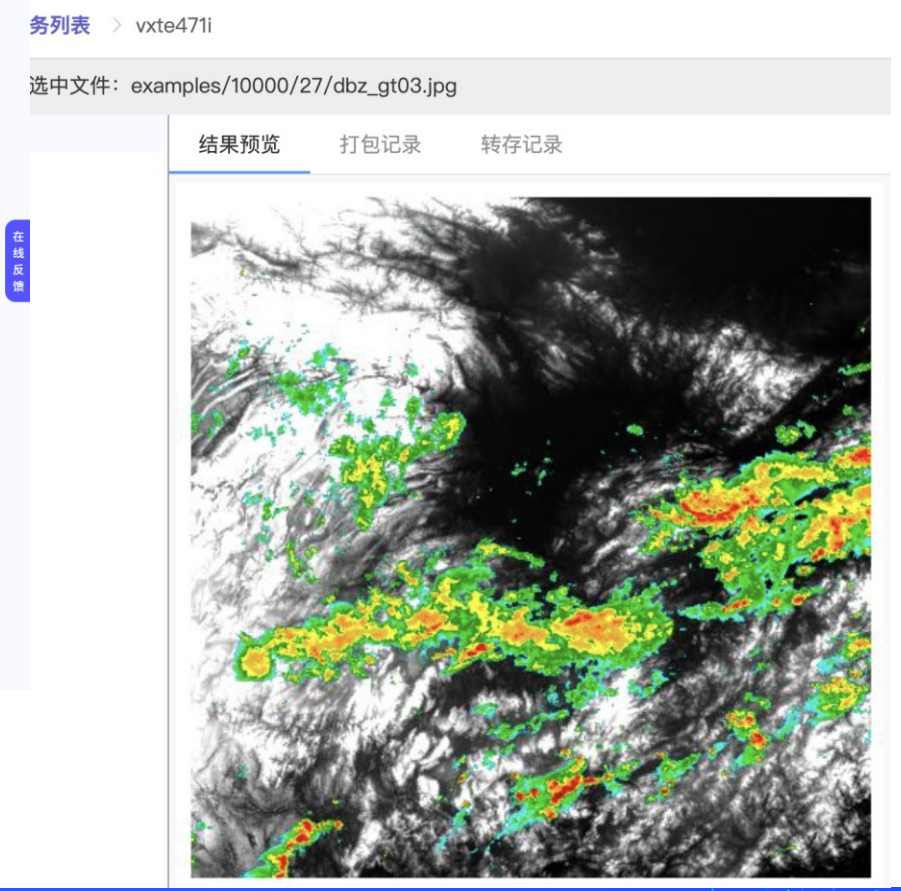


✓近万次实验记录

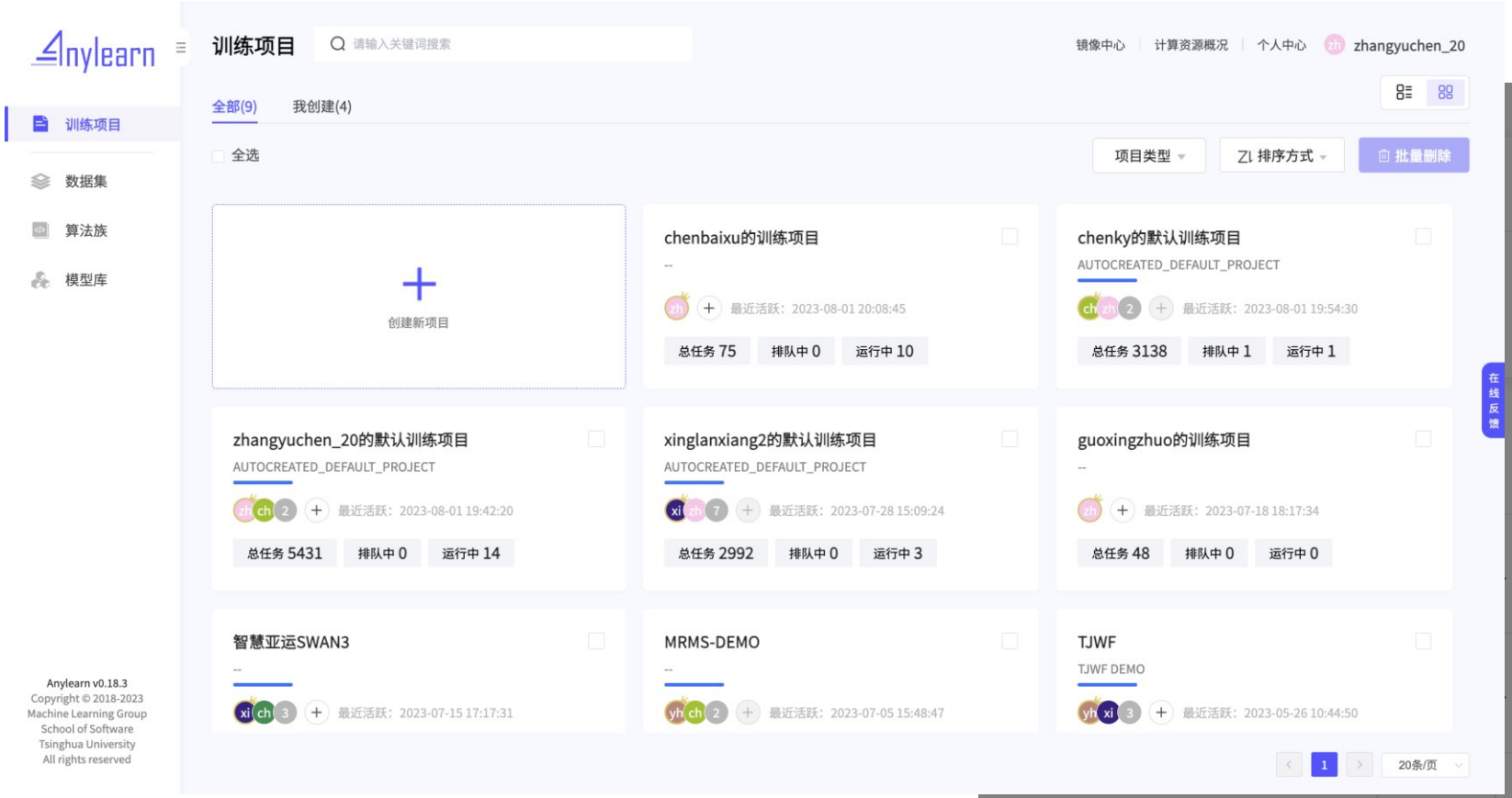
✓十万小时GPU训练时长

✓在线查看模型效果

✓Case study



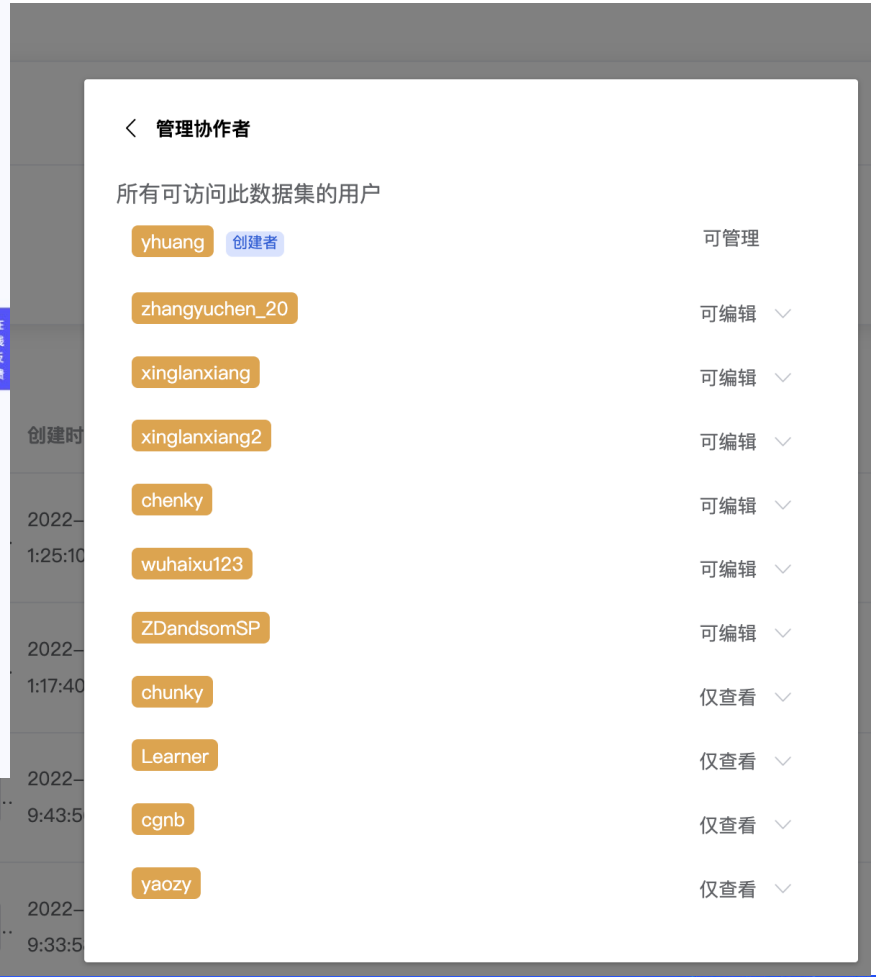
Anylearn助力领域大模型研发



✓多个项目并行推进

✓多人分工合作

✓研发团队共享资源

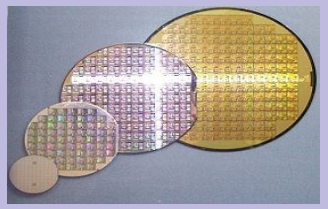


PART 06

总结与展望：人工智能软件研发支撑平台——工业大模型底座

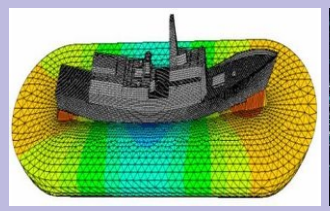
Anylearn工业大模型底座展望

晶圆检测 | 软件开发 | 农业 | 新能源 | 交通 | 大飞机装配 | 船舶 | 汽车

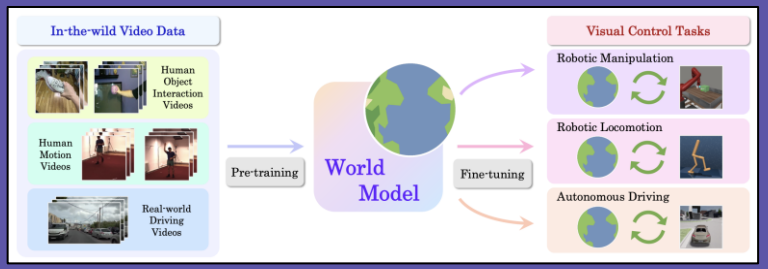






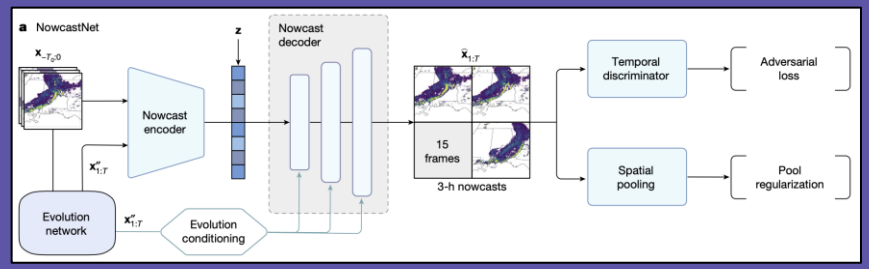



智能软件世界模型[9]



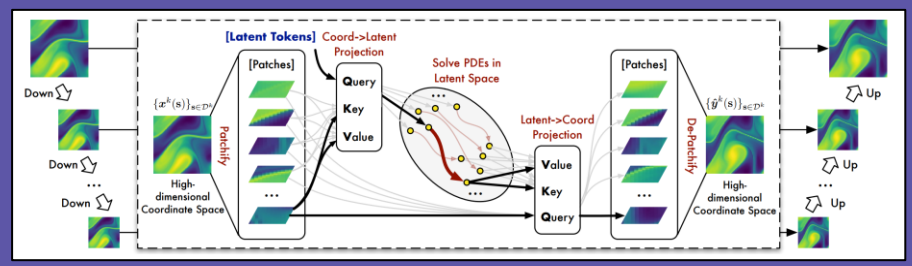
缺陷检测、智能编译、自动驾驶

地球气象基础模型[11]



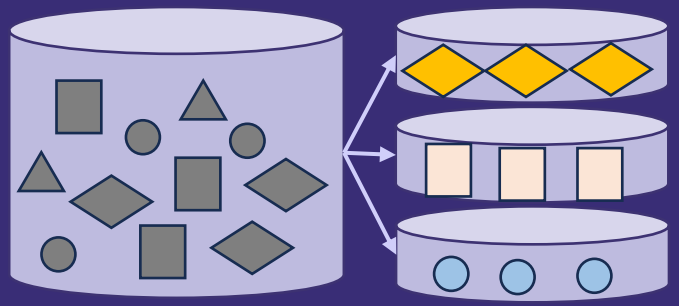
灾害天气预测、全球协同预报、气候推演

工业求解科学模型[6]

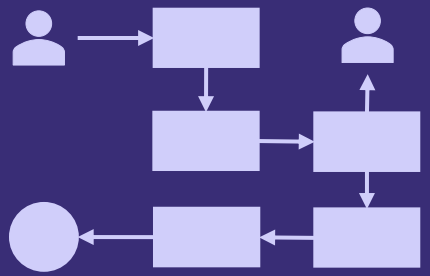


CAE前处理与求解、组合优化求解

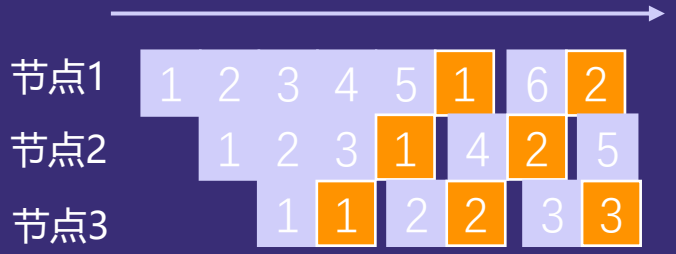
大模型研发资产管理



大模型研发过程管理



分布式高效数据读取



训、推、用一体

Anylearn工业大模型底座

参考文献

1. Amershi, S., Begel, A., Bird, C., DeLine, R., Gall, H., Kamar, E., Nagappan, N., Nushi, B., & Zimmermann, T. (2019). Software Engineering for Machine Learning: A Case Study. *International Conference on Software Engineering: Software Engineering in Practice (ICSE-SEIP)*, 291–300.

2. Idowu, S., Strüber, D., & Berger, T. (2022). Asset Management in Machine Learning: State-of-research and State-of-practice. *ACM Computing Surveys*, *55*(7), 1-35.

3. Kreuzberger, D., Kühl, N., & Hirschl, S. (2023). Machine Learning Operations (MLOps): Overview, Definition, and Architecture. *IEEE Access, 11*, 31866–31879.

4. Merritt, R. (2020). *What is MLOps?* NVIDIA Blog. https://blogs.nvidia.com/blog/2020/09/03/what-is-mlops/

5. Wirth, R., & Hipp, J. (2000). CRISP-DM: Towards a Standard Process Model for Data Mining. *International Conference on the Practical Applications of Knowledge Discovery and Data Mining*.

6. Wu, H., Hu, T., Luo, H., Wang, J., & Long, M. (2023). Solving High-Dimensional PDEs with Latent Spectral Models. *International Conference on Machine Learning*.

7. Wu, H., Xu, J., Wang, J., & Long, M. (2021). Autoformer: Decomposition Transformers with Auto-Correlation for Long-Term Series Forecasting. *Advances in Neural Information Processing Systems*.

8. Wu, H., Zhou, H., Long, M., & Wang, J. (2023). Interpretable weather forecasting for worldwide stations with a unified deep model. *Nature Machine Intelligence*, *5*(6), 602–611.

9. Wu, J., Ma, H., Deng, C., & Long, M. (2023). Pre-training Contextualized World Models with In-the-wild Videos for Reinforcement Learning. *(pre-print)*. https://arxiv.org/abs/2305.18499.

10. Your Ultimate Guide to ML Experiment Tracking. (n.d.). *Comet*. Retrieved August 10, 2023, from https://www.comet.com/site/lp/ultimate-guide-to-ml-experiment-tracking/

11. Zhang, Y., Long, M., Chen, K., Xing, L., Jin, R., Jordan, M. I., & Wang, J. (2023). Skilful nowcasting of extreme precipitation with NowcastNet. *Nature*, *619*(7970), 526–532.

AiDD AI+ 软件研发数字峰会
AI+ software Development Digital summit

感 谢 聆 听