

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+ 技术团队的选择



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



会议时间 | 08.18-19

AiDD AI+软件研发数字峰会
北京站



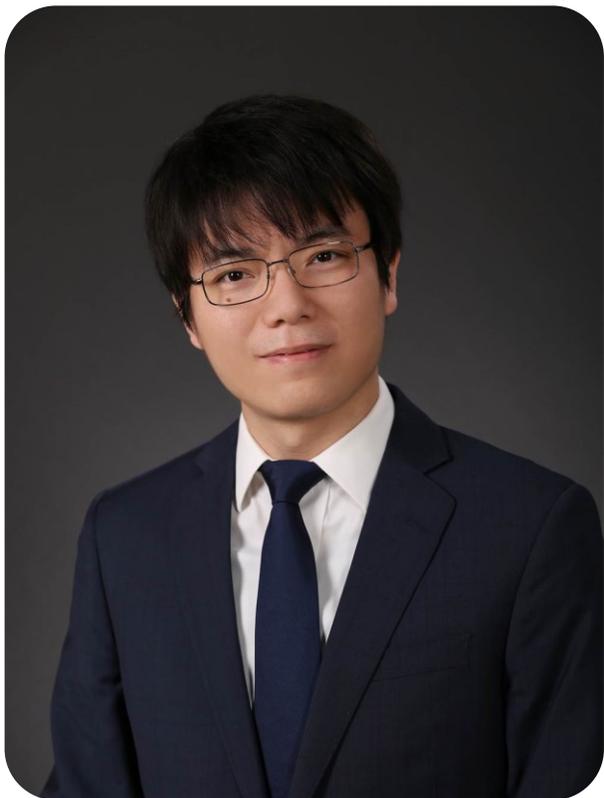
会议时间 | 11.17-18

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

清华大学软件学院院长聘副教授

清华学长聘副教授、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

1. 引言：人工智能工程化方法论
2. 人工智能大模型研发案例
3. 人工智能大模型研发挑战
4. 清华数为Anylearn系统介绍
5. Anylearn对大模型研发的支撑
6. 总结与展望：工业大模型底座

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



数据管理



机器学习



数据处理



应用开发

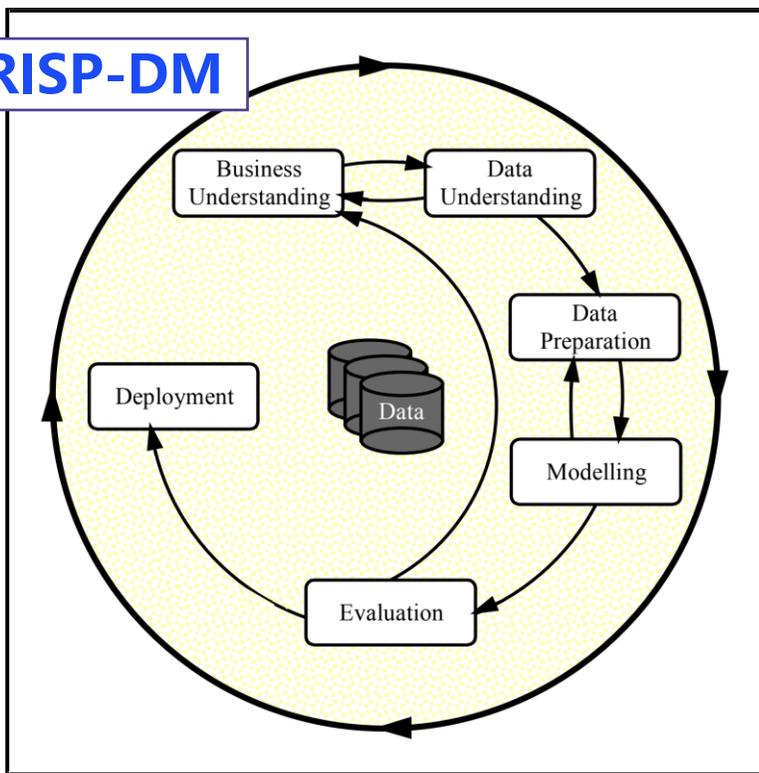
PART 01

引言：人工智能工程化方法论

人工智能工程化方法论

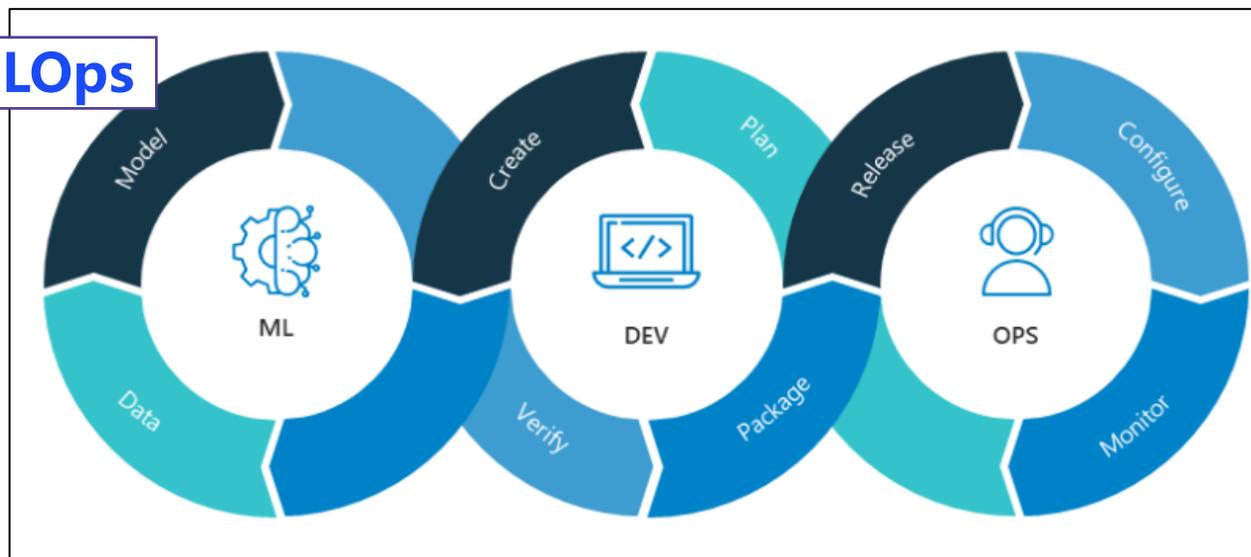
在数据、模型和服务层面上的**持续迭代**是智能软件研发的关键词。而人工智能工程化的核心在于**标准化**的研发流程和管理方法，通过体系化、规模化的分工协作和资产综合治理，提高研发**质量**和应用落地**效率**，推动人工智能技术为产业**持续赋能**。

CRISP-DM



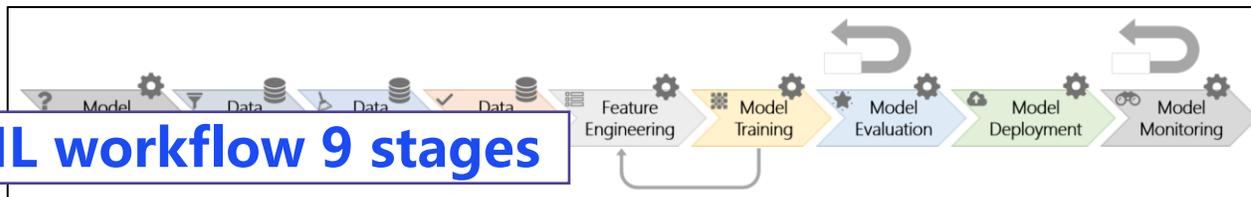
[5]

MLOps



[4]

ML workflow 9 stages



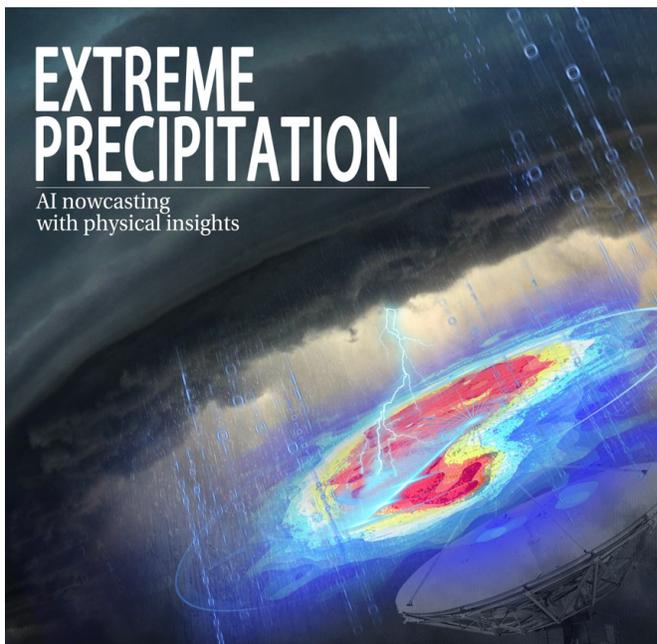
[1]

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

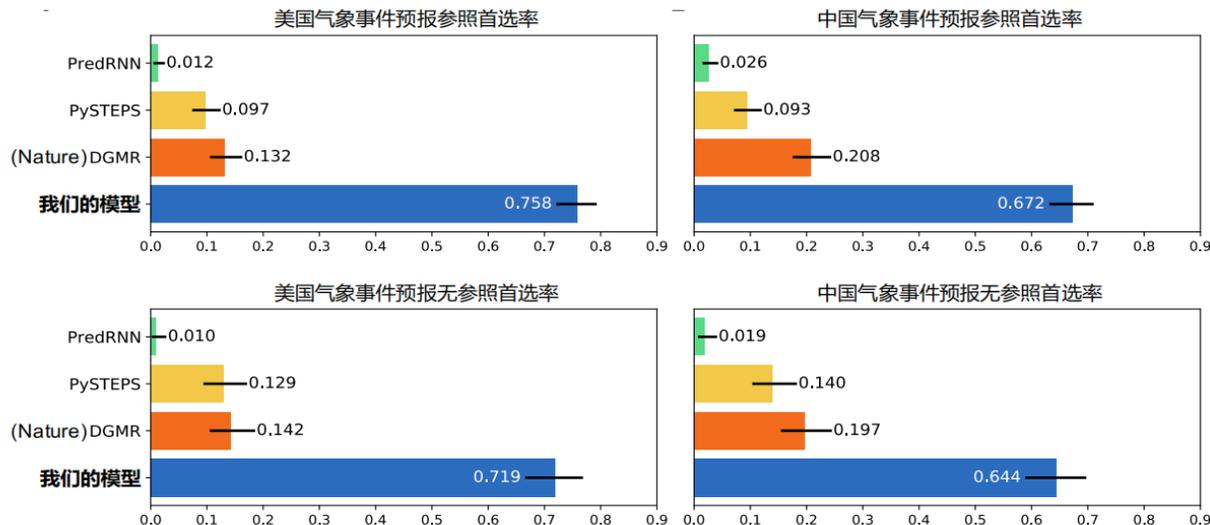
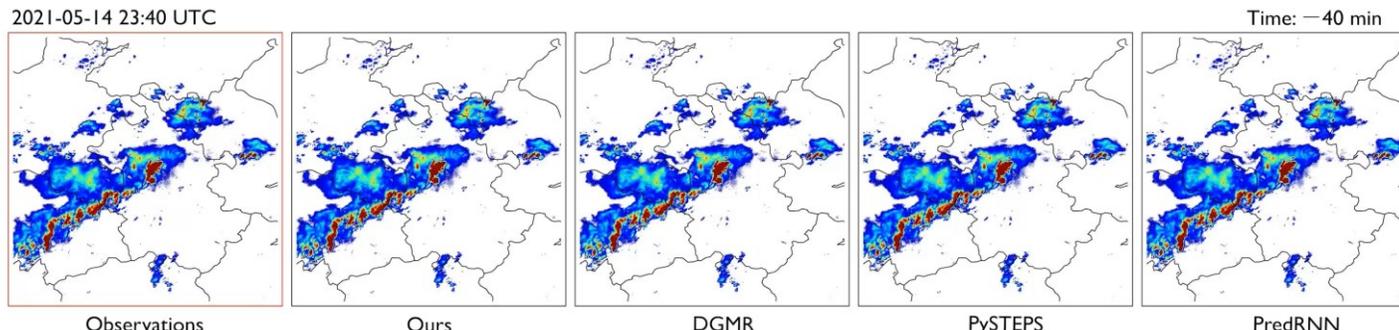
The outlook for AI weather prediction

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



✓ 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

AI驱动软件研发全面进入数字化时代

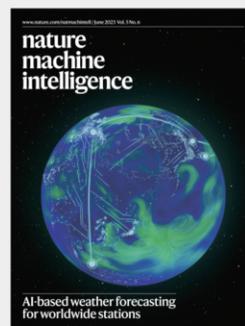
iDD AI+ 软件研发数字峰会
AI+ software Development Digital summit



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**

Volume 5 Issue 6, June 2023



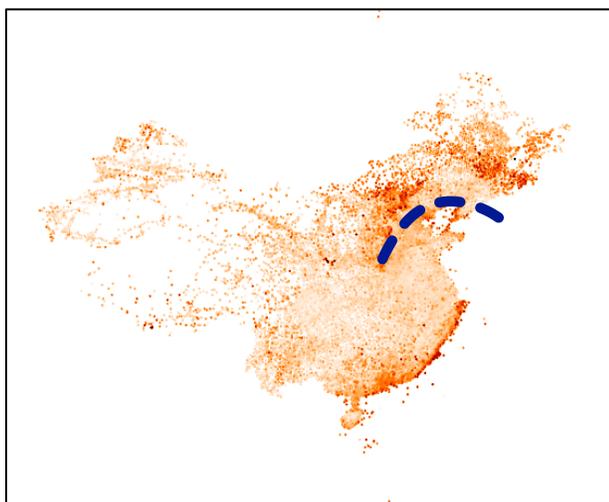
AI-based weather forecasting for worldwide stations

Weather forecasting has long attracted interest from scientists but owing to the chaotic nature of the atmosphere, simulating the weather at high spatial resolution with conventional methods is challenging. Wu et al. propose a data-driven approach for accurate and interpretable forecasting of the weather, based on partial observations of scattered stations over the world (see cover). The authors' unified deep learning model was successfully deployed to provide real-time weather forecasting services for competition venues during the 2022 Winter Olympics in Beijing.

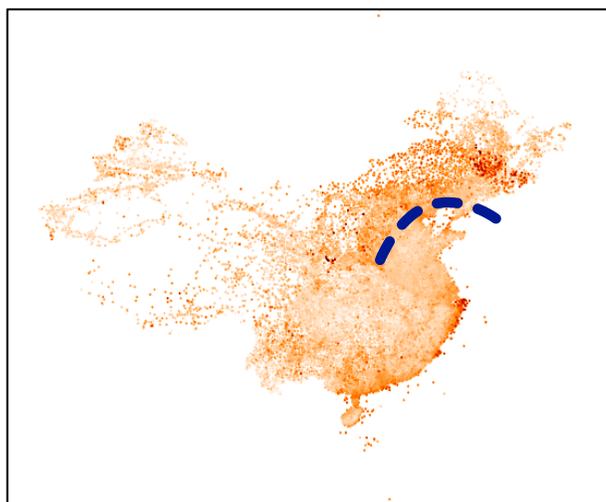
See [Wu et al.](#)

Image: Mingsheng Long, Tsinghua University. Cover design: Thomas Phillips

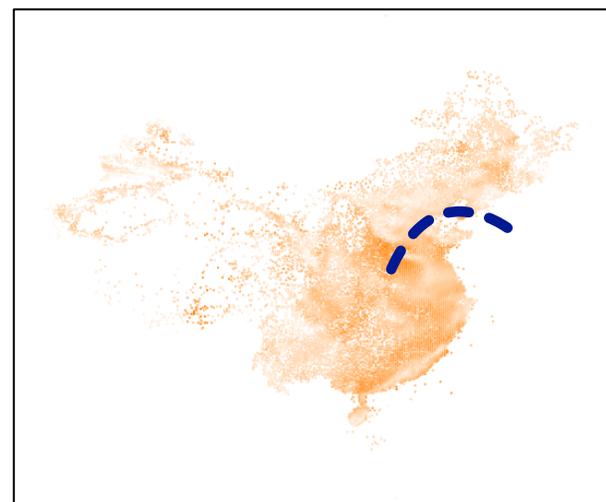
Subscribe



未来24小时真实观测



Corrformer预报结果



欧洲数值模式预报结果

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

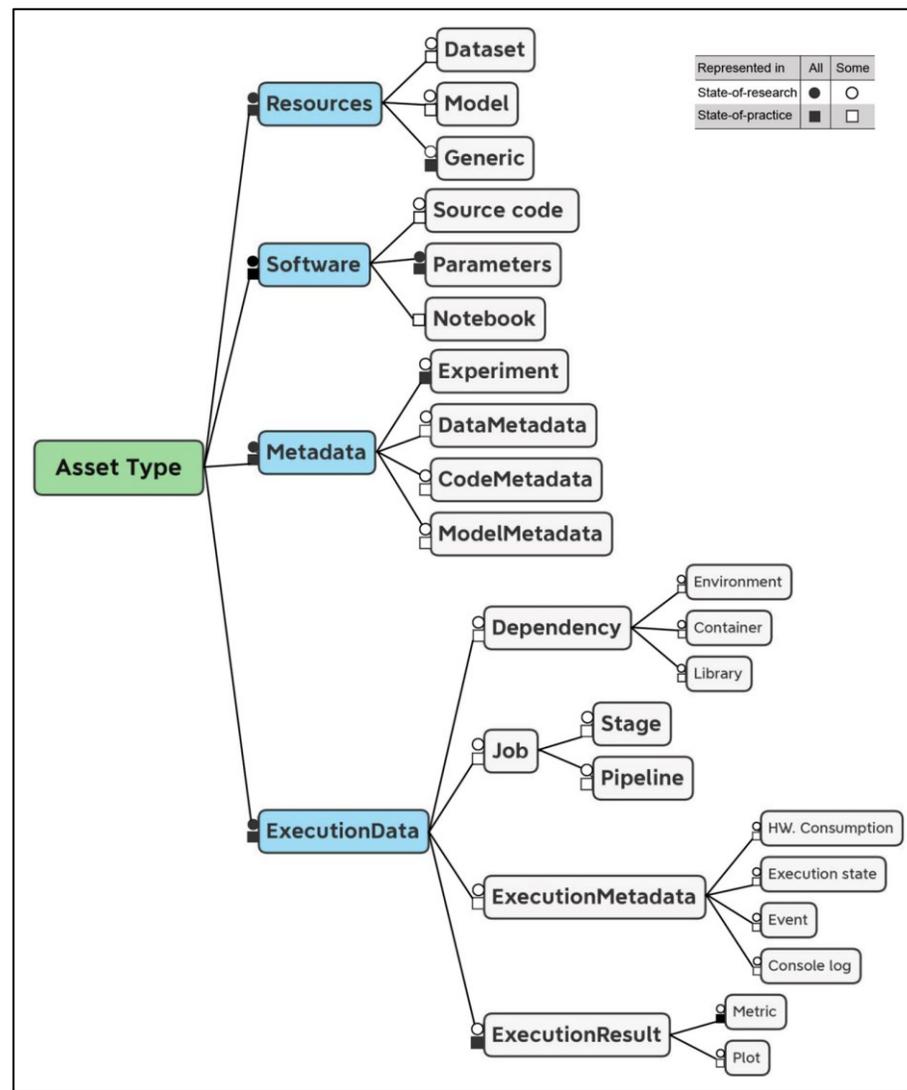
- 如何管理与追溯?
- 如何复用?

■ 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

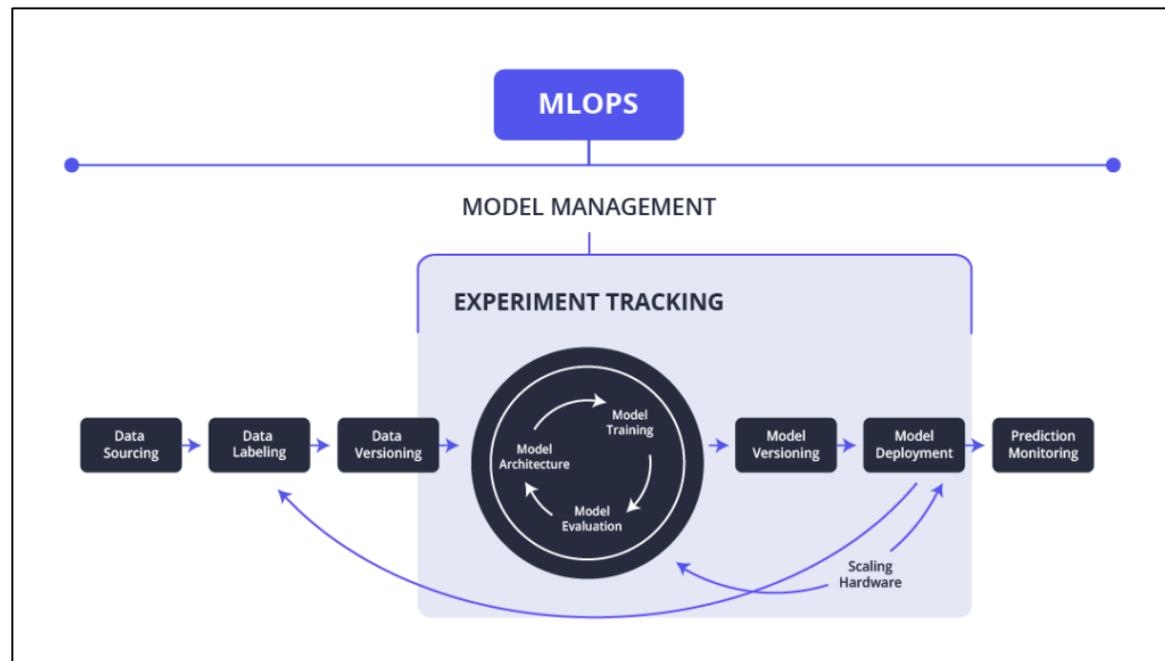
- 如何记录与对比?
- 如何分析与改进?

■ 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

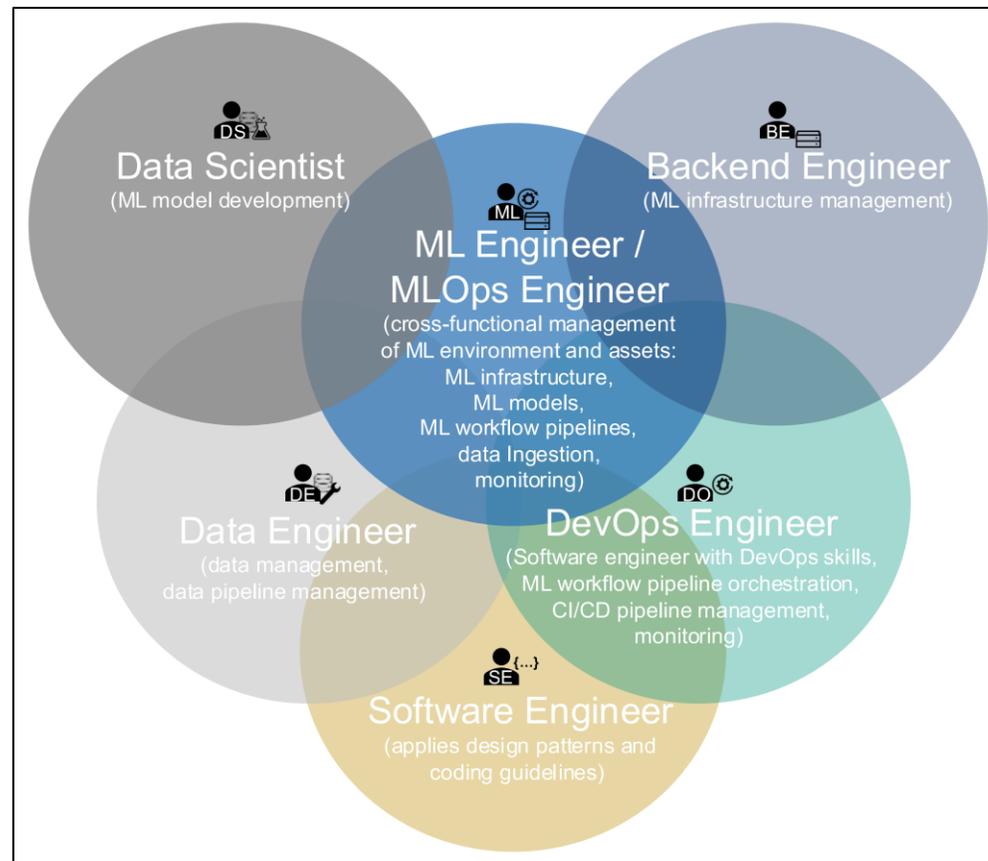
- 如何组织与分工?
- 如何沟通与汇报?

■ 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



PART 04

清华数为Anylearn系统介绍



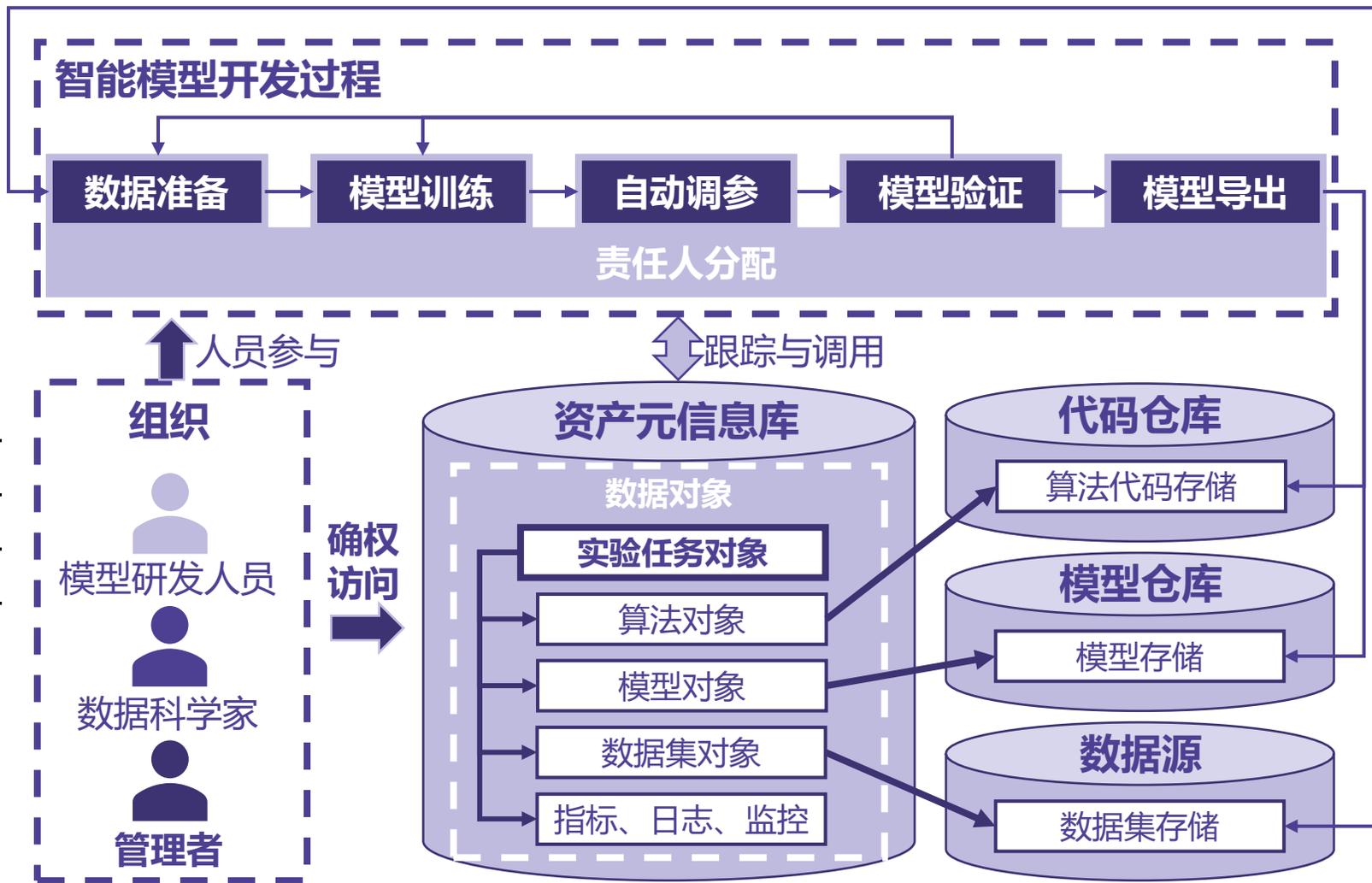
清华数为Anylearn

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

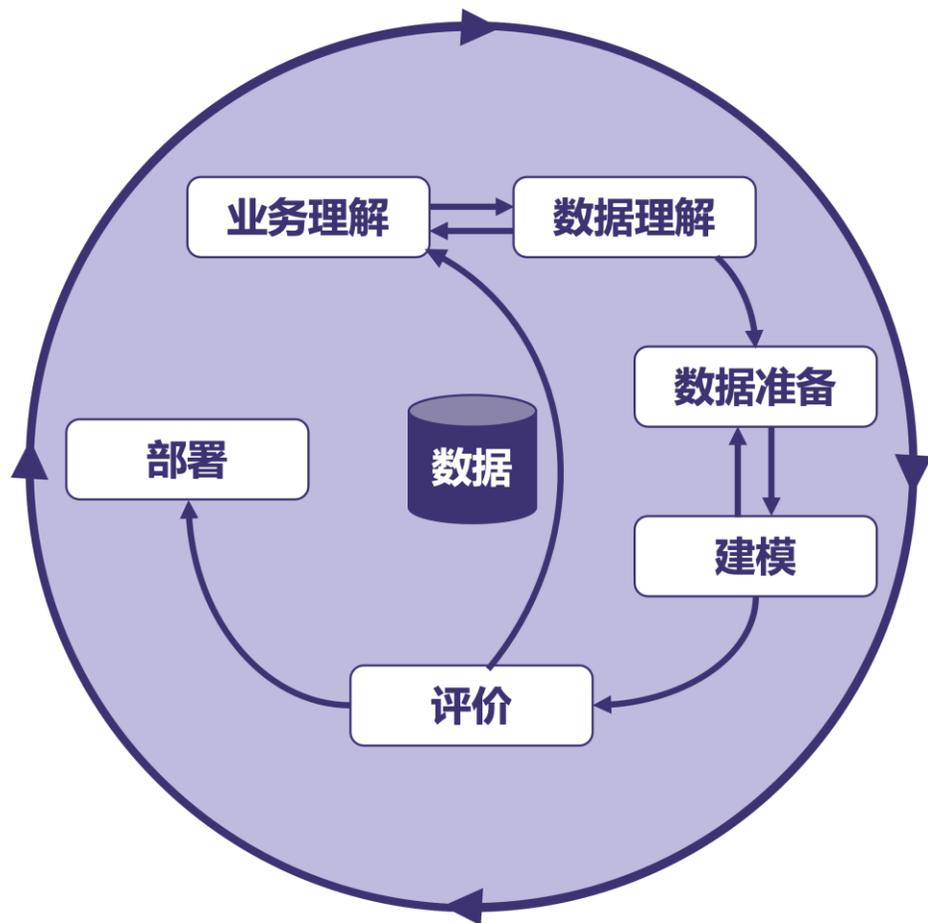
- 标准化机器学习资产管理与开发过程管理体系

高效率
可追溯
可复现

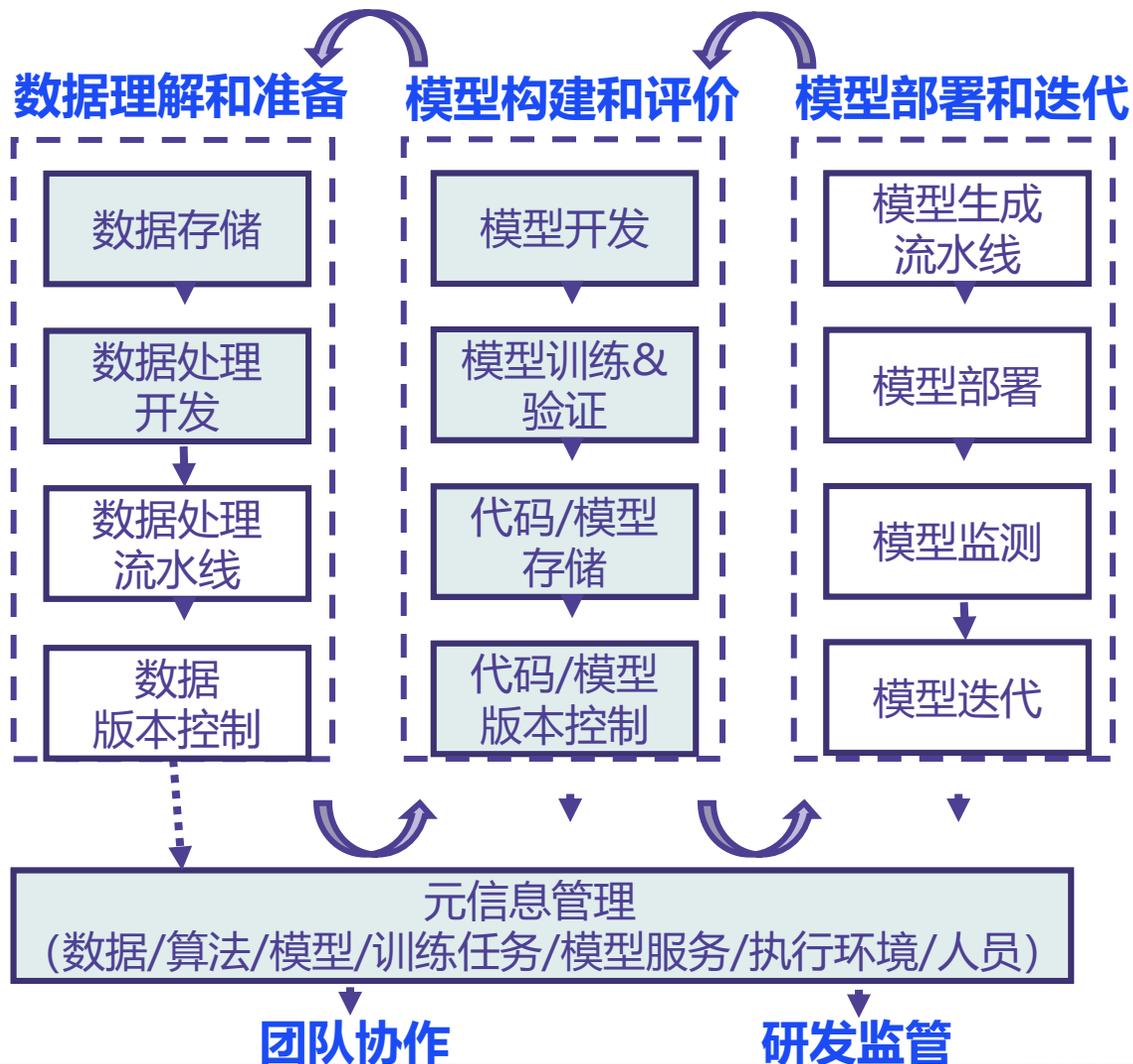


▶ Anylearn核心概念体系

- 方法论 → 实践 → 规范工具



标准化机器学习研发过程管理



Anylearn机器学习研发管理

- 可追溯、可搜索、可对比、可复现

算法

算法详情

chenky/cr_skillful_exp_sampling_10to20 私有 操作记录

描述: SDK_QUICKSTART
创建时间: 2022-05-27 12:16:09
创建者: chenky

算法近期版本

main

Anylearn auto-commit 2022-07-21 16:51:11 4dda293

chenkaiyuan 2022-07-21 16:51:11

Anylearn auto-commit 2022-07-21 15:30:20 7ffc581

chenkaiyuan 2022-07-21 15:30:20

Anylearn auto-commit 2022-07-21 15:22:02 72bd320

chenkaiyuan 2022-07-21 15:22:02

算法关联训练任务

相关训练任务 仅看星标

任务名称	主算法	算法版本	创建时间	任务状态	运行时间	标签	操作
2h8vois	cr_skillful_exp_sampling_10to20	62e1321	2022-07-05 18:46:56	已取消	8m 49s		详情
2zas9ym	cr_skillful_exp_sampling_10to20	5ed41f3	2022-07-05 18:46:47	已取消	9m 9s		详情
7fgapdmj	cr_skillful_exp_sampling_10to20	dc81731	2022-06-15 16:51:42	运行完毕	4h 22m 16s	4to10 zhiyu	详情
6mzwen0	cr_skillful_exp_sampling_10to20	e0201cf	2022-06-15 16:51:39	运行完毕	4h 24m 22s	4to10 zhiyu	详情
gx57enj	cr_skillful_exp_sampling_10to20	e21381e	2022-06-15 16:45:47	失败	55s		详情

对比 加入TensorBoard

训练输出模型

模型详情

chenky/mode-skillful-4to10-56w 私有

描述:

创建时间: 2022-06-15 15:45:44
创建者: chenky
来源算法: cr_skillful_exp_sampling_10to20
来源任务: ux12dzkc

算法代码版本对比

cr_skillful_exp_sampling_10to20 (ALGOdf6dd7311ecaec2368ecec84d143)

Commit e0201cfedf09022b3a84555cc37946fde65154cd

Files changed (2)

- .idea/workspace.xml (+2 -2)
- pred_learn/models/model_factory_large_skillful_ensemble.py (+7)

pred_learn/models/model_factory_large_skillful_ensemble.py (CHANGED)

```
@@ -91,12 +91,13 @@ class Model(object):
91     stats = torch.load(self.configs.pretr
92     else:
93     stats = torch.load(self.configs.pretr
94     #
95     # stats = {
96     #     k[7:]: v for k, v in
97     #     stats['net_param'].items()
98     # }
99     self.network.load_state_dict(stats, stric
100
101     def train(self, frame, mask):
102     stats = torch.load(self.configs.pretr
103     else:
104     stats = torch.load(self.configs.pretr
105     #
106     # stats = {
107     #     k[7:]: v for k, v in
108     #     stats['net_param'].items()
109     # }
110     self.network.load_state_dict(stats, stric
111     self.network.load_state_dict(stats['net_
112     def train(self, frame, mask):
```

训练元信息 (超参数、执行环境)

6mzwen0

任务状态: 运行完毕

任务ID: TRAI5d58ec8811ec9133d2ea4a6c6bc2

结果ID: FILE885cec8811ec9133d2ea4a6c6bc2

主算法: cr_skillful_exp_sampling_10to20

执行镜像: QUICKSTART_CARTOPY_PYTORCH1.9.0_CUDA11

资源配置: RADAR

RTX-3090-unique count 1
CPU count 16
Memory count 128
A-100-unique count 0

运行时长: 4h 24m 22s

创建时间: 2022-06-15 16:49:59

开始时间: 2022-06-15 16:50:05

完成时间: 2022-06-15 21:14:27

标签: 4to10 zhiyu

超参数配置

参数名	值
worker	1
is_training	0
small_data	false
device	cuda:0
cpu_worker	4
dataset_name	cr_png
save_dir	results/checkpoints
gen_frm_dir	results/examples
num_save_samples	6
model_name	skillful
reverse_input	0
img_height	512
img_width	512
train_width	256
train_height	256
img_ch	2
input_length	4

▶ 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

Anylearn

算法族

请输入关键词搜索

镜像中心 | 计算资源概况 | 个人中心 | zhangyuchen_20

训练项目

数据集

算法族

模型库

+ 上传新算法

批量删除算法

名称	描述	上传时间	创建者	文件状态	操作
mrms_density_compress10	SDK_QUICKSTART	2023-02-15 11:02:55	zhangyuchen_20	就绪	详情 编辑 删除
mrms_density_compress01	SDK_QUICKSTART	2023-02-15 10:26:47	zhangyuchen_20	就绪	详情 编辑 删除
mrms_density_recover_reg0	SDK_QUICKSTART	2023-02-15 10:15:13	zhangyuchen_20	就绪	详情 编辑 删除
mrms_density_reg0	SDK_QUICKSTART	2023-02-11 12:25:15	zhangyuchen_20	就绪	详情 编辑 删除
mrms_density_reg1e0	SDK_QUICKSTART	2023-02-11 12:24:51	zhangyuchen_20	就绪	详情 编辑 删除
mrms_density_reg1e-4	SDK_QUICKSTART	2023-02-11 12:14:59	zhangyuchen_20	就绪	详情 编辑 删除
mrms_density_1e-4_u2net_k9_cutoff15	SDK_QUICKSTART	2023-02-11 11:04:34	zhangyuchen_20	就绪	详情 编辑 删除
mrms_density_1e-4_u2net_k9_cutoff10	SDK_QUICKSTART	2023-02-11 10:57:32	zhangyuchen_20	就绪	详情 编辑 删除
mrms_density_1e-4_u2net_k9_cutoff5	SDK_QUICKSTART	2023-02-11 10:53:37	zhangyuchen_20	就绪	详情 编辑 删除
mrms_density_1e-4_u2net_ga_mma10	SDK_QUICKSTART	2023-02-04 10:12:25	zhangyuchen_20	就绪	详情 编辑 删除
mrms_density_1e-4_u2net_0_25p7	SDK_QUICKSTART	2023-02-03 22:44:45	zhangyuchen_20	就绪	详情 编辑 删除

共 283:

✓ 几千次复现和方案

✓ 记录每一次代码变更

Anylearn v0.18.3
Copyright © 2018-2023
Machine Learning Group
School of Software
Tsinghua University
All rights reserved

Anylearn v0.18.3
Copyright © 2018-2023
Machine Learning Group
School of Software
Tsinghua University
All rights reserved

✓ 资产通过训练任务形成有机关联

数据集

镜像中心 | 计算资源概况 | 个人中心 | zhangyuchen_20

数据集详情

任务名称	主算法	算法版本	创建时间	任务状态	运行时间	标签	操作
kyrn1lp	mrms_spade2ens_sel_0504_draw	856dc1d	2023-06-07 18:09:01	运行完毕	44s		详情 ...
s4r18ow3	mrms_spade2ens_sel_0504_draw	b753fb1	2023-06-07 18:08:48	失败	30s		详情 ...
5snj12ba	mrms_density_1e-4_u2net_weight_max128_min10	90e51f2	2023-06-07 15:38:50	已取消	7d 8h 45m 57s		详情 ...
6mkjv4rp	mrms_u2net_V14_bs16_u2nettk_bn	702e8bf	2023-05-30 12:39:36	已取消	9d 1h 45m 50s		详情 ...
60z85bom	mrms_u2net_V14_bs16_u2nettk_bn_win3	160cc3e	2023-05-30 10:07:44	运行完毕	11d 23h 9m 28s		详情 ...

对比 加入TensorBoard

文件目录

- 2015
- 2016
- 2017
- 2018
- 2019
- 2020
- 2021
- 2022
- 2023
- cleaned_mrms
- cleaned_mrms_anylearn_usage
- MRMS_CASE_TEST
- MRMS_Final_Test_Patch
- MRMS_FULL_TEST
- MRMS_FULL_TEST_24
- MRMS_Test_Set
- MRMS_Test_Set1
- MRMS_Test_Set2
- MRMS_Test_Set3
- mtarchive.geol.iastate.edu
- multimodal
- multimodal_patch_256
- satellite
- topo
- data_rain_map_256_test.npy
- data_rain_map_256_test2.npy
- data_rain_map_256_train.npy
- data_rain_map_256_train2.npy
- data_rain_map_256_valid.npy
- data_rain_map_256_valid2.npy
- data_rain_map_384_test.npy
- data_rain_map_384_test2.npy
- data_rain_map_384_train.npy

AI驱动软件研发全面进入数字化时代

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

Anylearn助力领域大模型研发

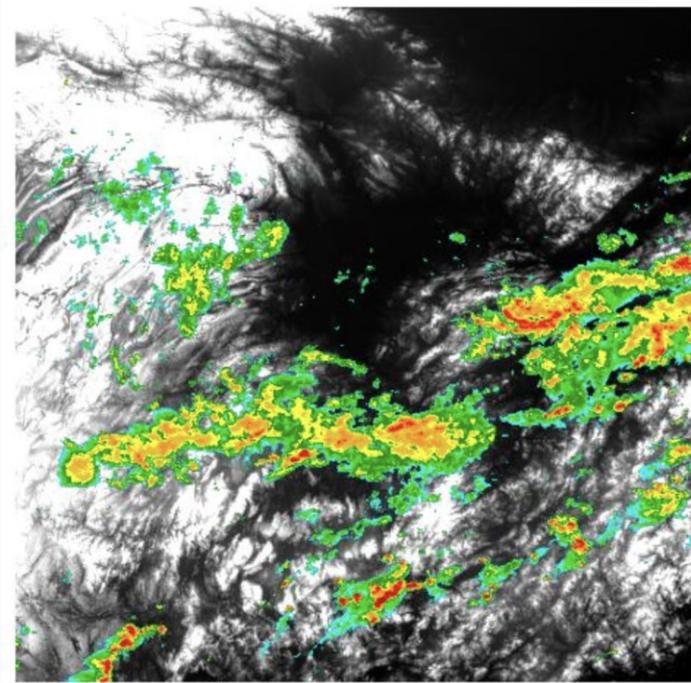
✓ 在线查看模型效果
✓ Case study

任务名称	主算法	算法版本	创建时间	创建者	任务状态	运行时间	操作
jrlz1e8q	mrms_u2net_V1_4_bs16_u2net_bn_win3	908b658	2023-05-30 12:47:23	zhangyuchen_20	运行完毕	9d 11h 51m 45s	详情 删除
olam0sc7	mrms_u2net_V1_4_bs16_u2net_bn	0f798ea	2023-05-30 12:47:04	zhangyuchen_20	运行完毕	9d 13h 52m 12s	详情 删除
6mkjv4rp	mrms_u2net_V1_4_bs16_u2netk_bn	702c8bf	2023-05-30 12:39:36	zhangyuchen_20	已取消	9d 1h 45m 50s	详情 删除
60z85bom	mrms_u2net_V1_4_bs16_u2netk_bn_win3	160cc3c	2023-05-30 10:07:44	zhangyuchen_20	运行完毕	11d 23h 9m 28s	详情 删除
xv2zncyf	mrms_u2net_V1_4_bs16_u2netk_bn_win3	c6444ad	2023-05-30 10:03:15	zhangyuchen_20	已取消	4m 3s	详情 删除
wz3y15pr	mrms_u2net_V1_4_bs16_u2net_bn_win3	a74953a	2023-05-29 14:15:15	zhangyuchen_20	已取消	4d 3h 30m 16s	详情 删除
ivo4d7ct	mrms_u2net_V1_4_bs16_u2net_bn	a01d476	2023-05-29 14:09:36	zhangyuchen_20	运行完毕	12d 20h 38m 10s	详情 删除
2lhb13a	mrms_u2net_V1_4_bs16_u2net_bn	191669b	2023-05-29 14:07:28	zhangyuchen_20	失败	1m 2s	详情 删除

任务列表 > vxte471i

选中文件: examples/10000/27/dbz_gt03.jpg

结果预览 打包记录 转存记录



✓ 近万次实验记录

✓ 十万小时GPU训练时长

AI驱动软件研发全面进入数字化时代

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

Anylearn助力领域大模型研发

训练项目 请输入关键词搜索

全部(9) 我创建(4)

项目类型 排序方式 批量删除

创建新项目

项目名称	最近活跃	总任务	排队中	运行中
chenbaixu的训练项目	2023-08-01 20:08:45	75	0	10
chenky的默认训练项目	2023-08-01 19:54:30	3138	1	1
zhangyuchen_20的默认训练项目	2023-08-01 19:42:20	5431	0	14
xinglanxiang2的默认训练项目	2023-07-28 15:09:24	2992	0	3
guoxingzhuo的训练项目	2023-07-18 18:17:34	48	0	0
智慧亚运SWAN3	2023-07-15 17:17:31			
MRMS-DEMO	2023-07-05 15:48:47			
TJWF	2023-05-26 10:44:50			

Anylearn v0.18.3
Copyright © 2018-2023
Machine Learning Group
School of Software
Tsinghua University
All rights reserved

✓研发团队共享资源

管理协作者

所有可访问此数据集的用户

用户名	权限
yhuang	可管理
zhangyuchen_20	可编辑
xinglanxiang	可编辑
xinglanxiang2	可编辑
chenky	可编辑
wuhaixu123	可编辑
ZDandsomSP	可编辑
chunky	仅查看
Learner	仅查看
cgnb	仅查看
yaozy	仅查看

✓多个项目并行推进

✓多人分工合作

AI驱动软件研发全面进入数字化时代

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

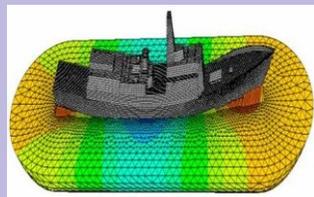
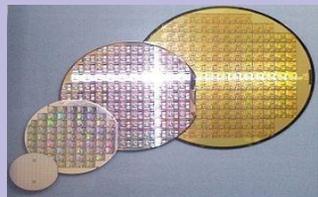
新能源

交通

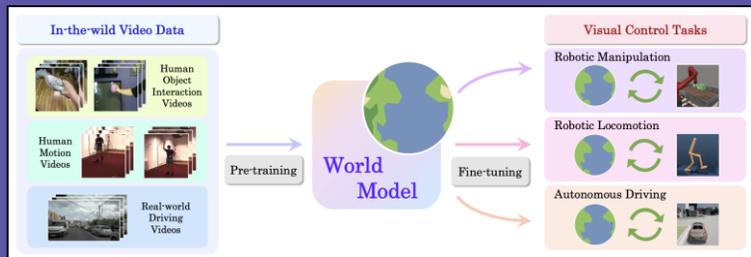
大飞机装配

船舶

汽车

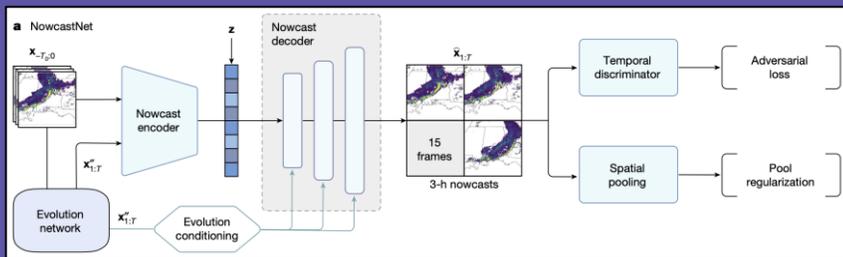


智能软件世界模型^[9]



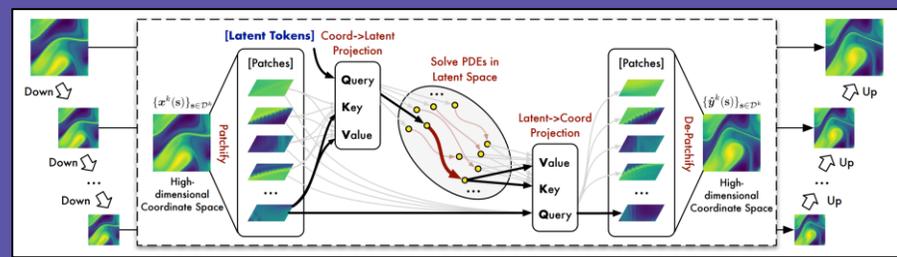
缺陷检测、智能编译、自动驾驶

地球气象基础模型^[11]



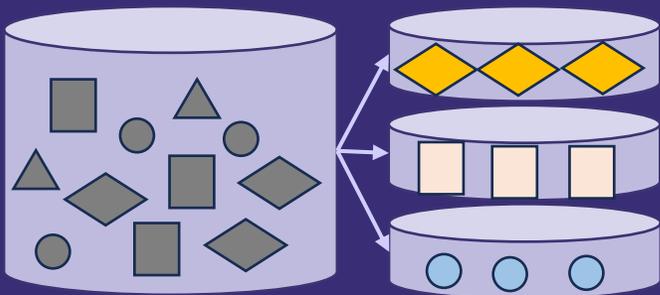
灾害天气预测、全球协同预报、气候推演

工业求解科学模型^[6]

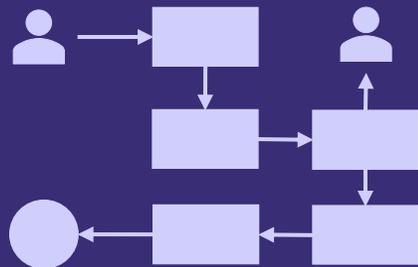


CAE前处理与求解、组合优化求解

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

节点1	1	2	3	4	5	1	6	2
节点2		1	2	3	1	4	2	5
节点3			1	1	2	2	3	3

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.

感谢聆听

