中国科学院国家天文台
NATIONAL ASTRONOMICAL OBSERVATORIES, CAS

大数据与大模型时代的天文研究

罗阿理　国家天文台

南开　2025.1.16

内容

- 天文大数据与分析工具的需求
- 天文图像数据和序列数据的深度学习：CNN
- 时间序列数据的深度学习：RNN（LSTM）
- Transformer架构与光谱数据
- 自监督的人工智能视觉工具ViT：MAE，BEIT
- 科学模型及国家天文台AI大模型规划与现状
- 基于Agent的AI科学家和天文AI体系建设

大样本观测是天文学新方向的数据基础

探索宇宙的边疆：天文学新方向

自然基金委《天文学十四五及中长期规划》景益鹏等

多信使天文学

使用引力波、中微子、宇宙线等非电磁手段来研究致密天体性质、丈量宇宙时空、追溯剧烈天体物理过程、检验基本物理规律

时域天文学

采用多波段、多时标方式研究动态宇宙，通过重复观测来揭示宇宙中各类天体的变化天体物理过程、检验基本物理规律

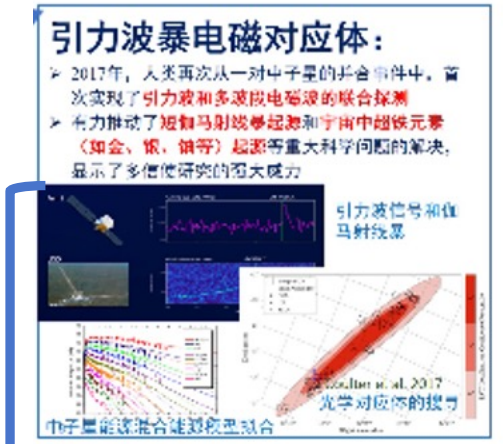
行星大科学

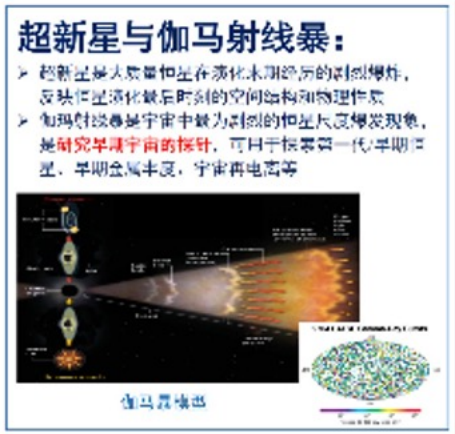
关乎生命起源的行星大科学，是集系外行星、太阳系行星、天体生物学、天体化学、地质学研究方法于一体的高度较差学科，旨在探索行星与生命的起源与演化

多波段、手段联合观测： 不同侧面、不同类型天体更加全面的信息

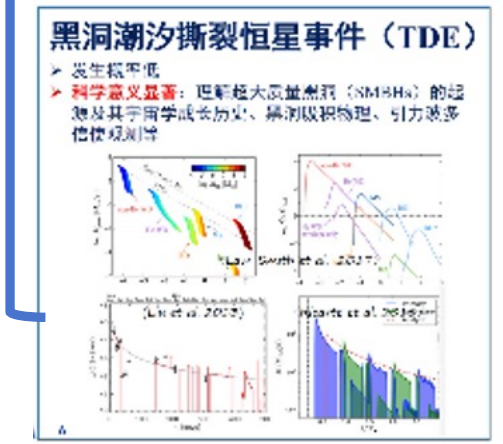
大天区面积深度巡天： 覆盖尽可能多的天体类型和数量

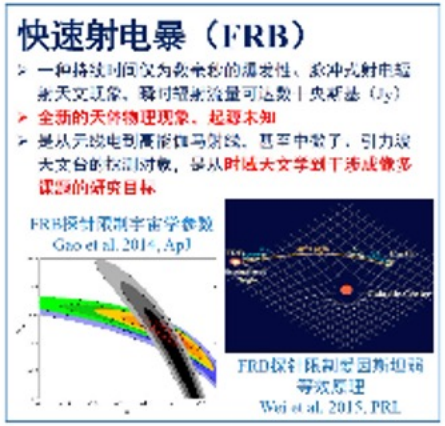
高频率采样、长期持续监测： 暂现源和变源的长期/短时标的变化特点





\+ + +



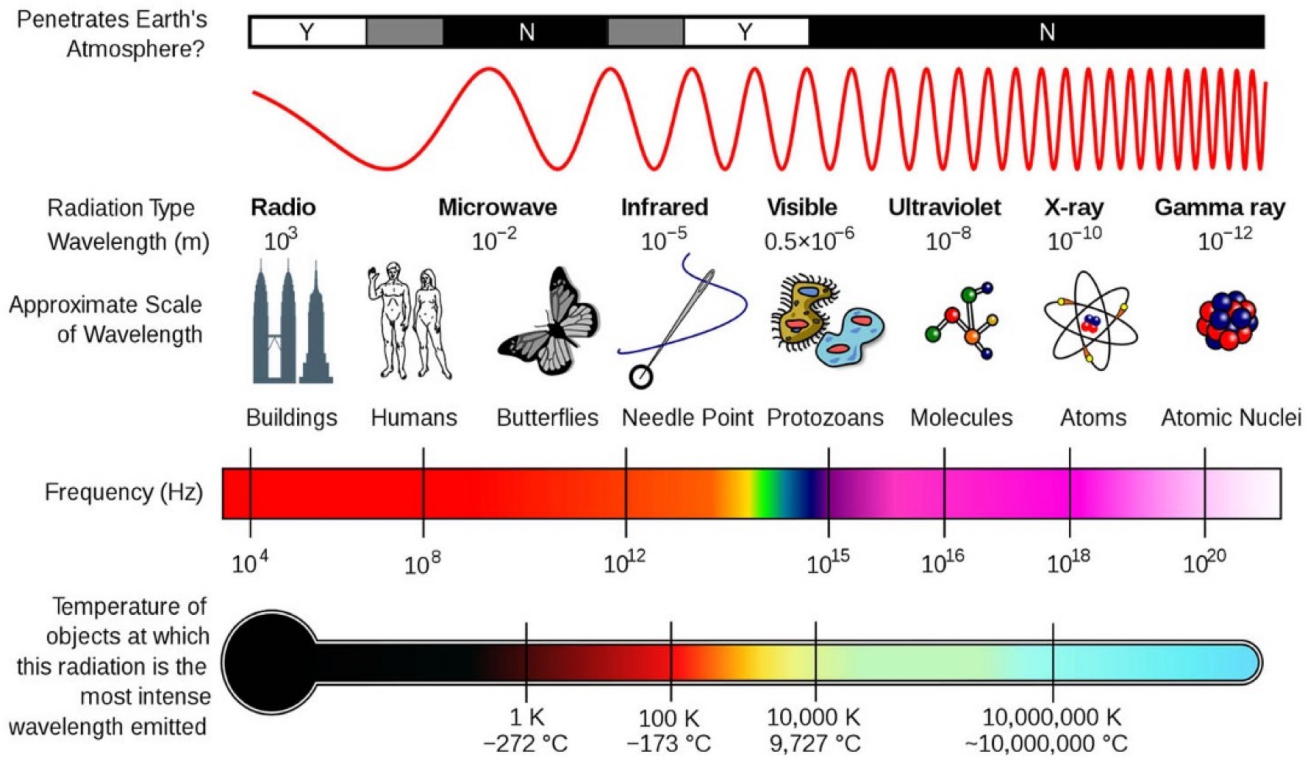


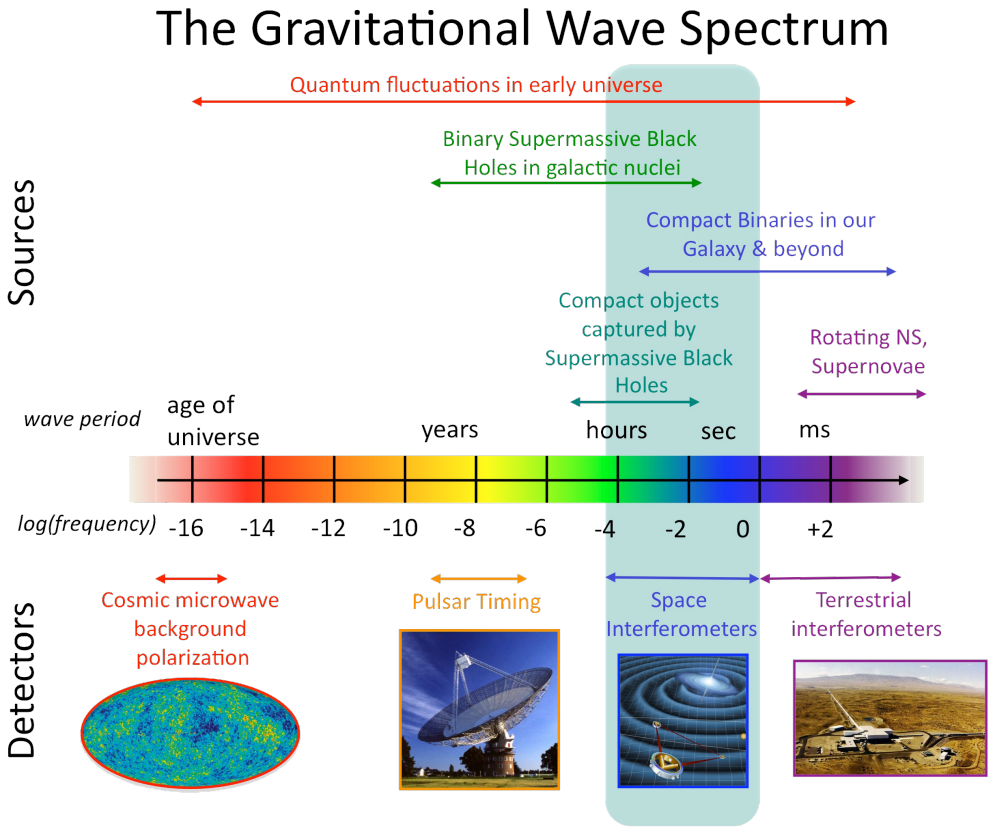
Credit to 刘继峰PPT

超大规模光谱巡天也是天文学新方向的基石

大规模光谱巡天成果多，领域广。为**暗能量和暗物质、黑洞、星系的组装、银河系结构和演化、恒星物理、行星科学、时域和多信使天文学**等重大科学前沿问题提供关键的数据基础。

电磁波是大样本天文学的最主要实测数据





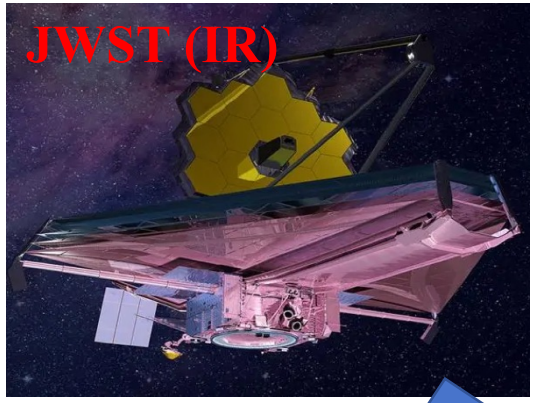
天文学是一门基于观测的交叉学科

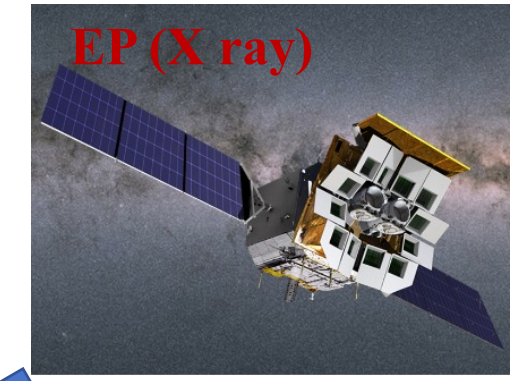
巡天观测望远镜

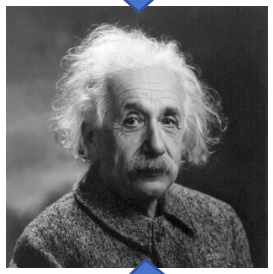
Optical /IR	Raido	UV	X-ray	Gamma ray
• SDSS • 2dF • RAVE • 2MASS • UKIDSS • WISE • Pan-STARRS • GAIA • ELT • LSST • LAMOST • DESI	• SKA • VLA • FAST	• FUSE • GALEX • CSST	• XMM-Newton • Rosat X • Chadra • HXT • EP	• Fermi • Swift • CGRO • INTEGRAL • HESS • Cos-B • SVOM

天文观测设施与大数据

空间
望远镜





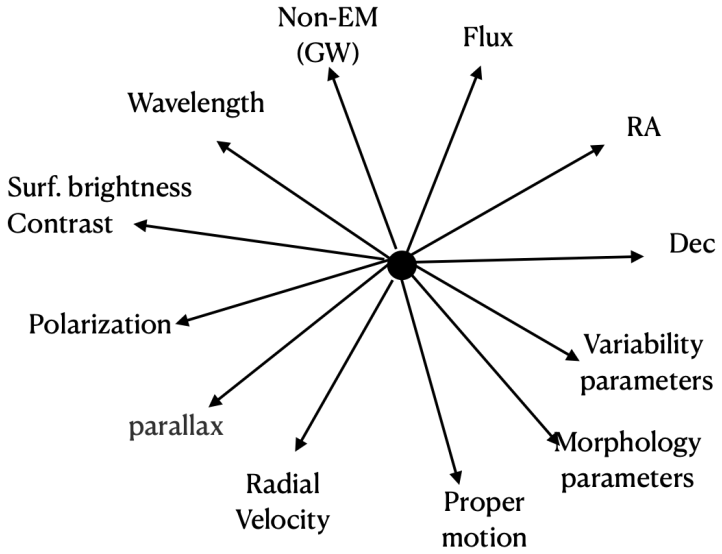


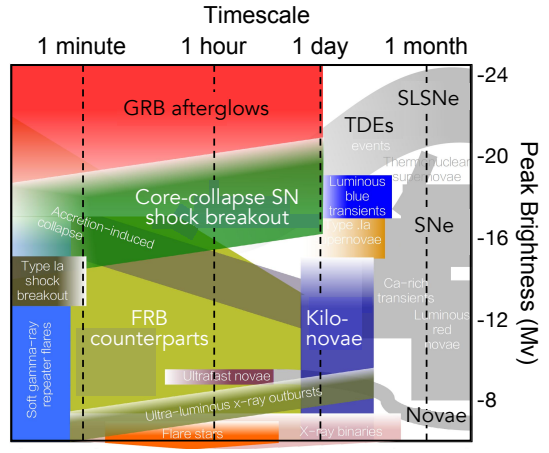
天体物理学家

地基
望远镜

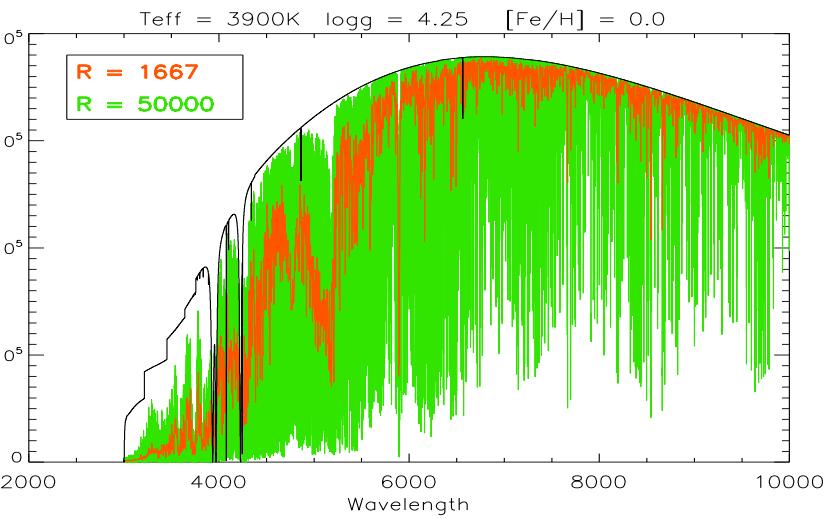
天文观测“大”数据的特点

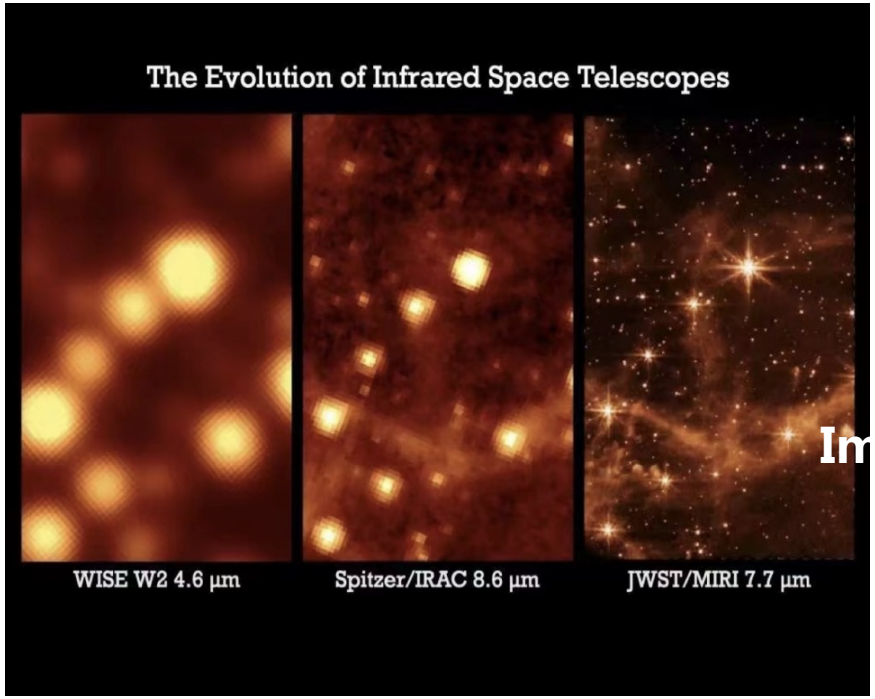
Sky Survey Projects	Data Volume
DPOSS (The Palomar Digital Sky Survey)	3 TB
2MASS (The Two Micron All-Sky Survey)	10 TB
GBT (Green Bank Telescope)	20 PB
GALEX (The Galaxy Evolution Explorer)	30 TB
SDSS (The Sloan Digital Sky Survey)	40 TB
SkyMapper Southern Sky Survey	500 TB
PanSTARRS (The Panoramic Survey Telescope and Rapid Response System)	~ 40 PB expected
LSST (The Large Synoptic Survey Telescope)	~ 200 PB expected
SKA (The Square Kilometer Array)	~ 4.6 EB expected
FAST	1 PB /week 1EB / 19years

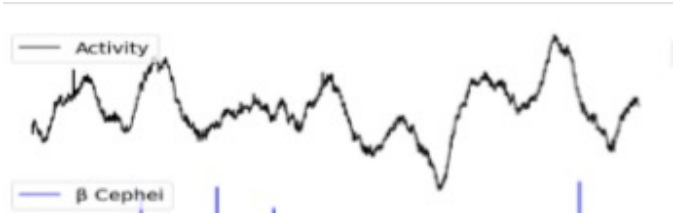


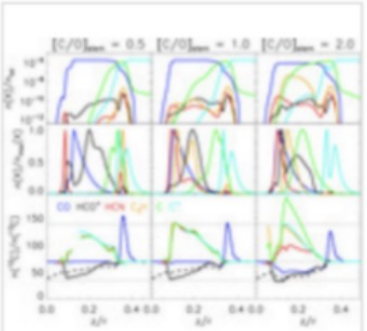


来自天文观测的大数据的主要类型

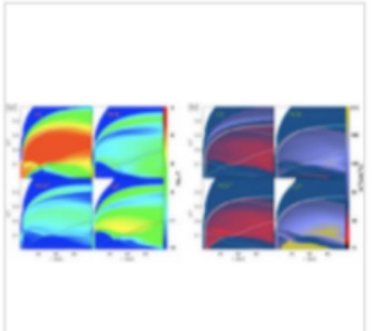




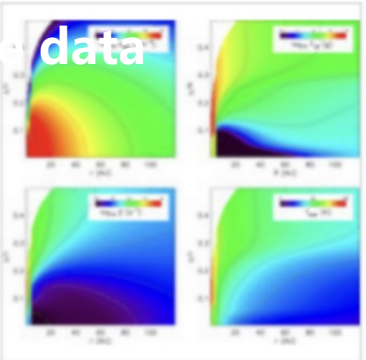




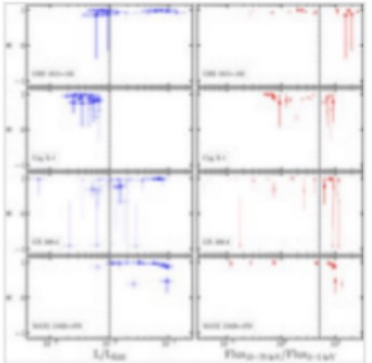
Vertical abundance (top) and isotope ratio (bot...



Abundance (a) and carbon isotope ratio (b) for ...



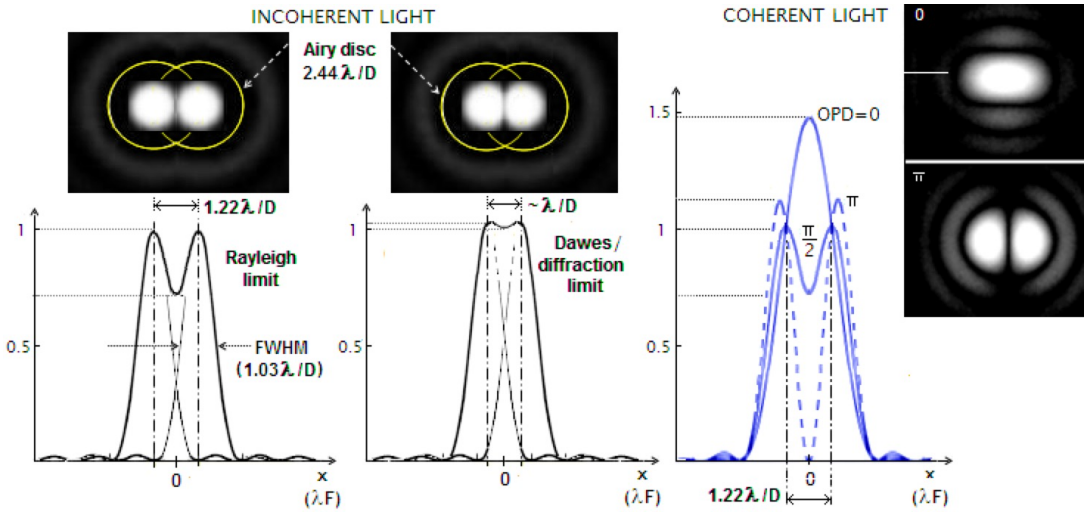
Physical parameters. gas number density (n ...



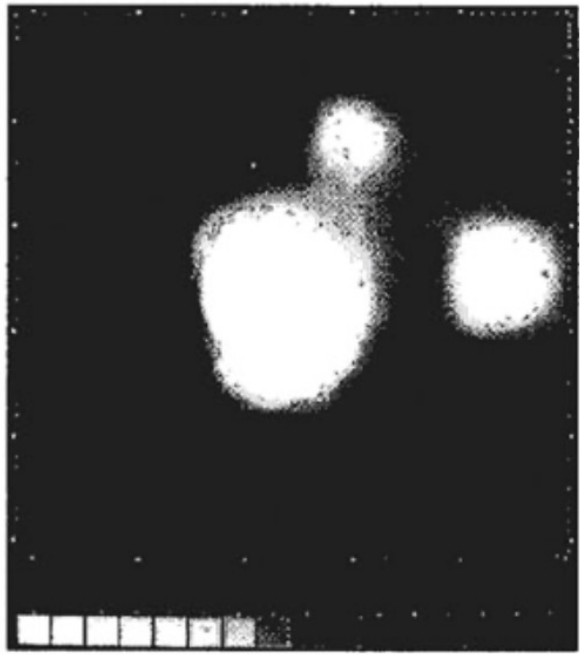
BH spin vs. Eddington fraction (left) and hardn...

巡天大数据的分辨率与信噪比

观测数据分辨率差异且普遍受到噪声干扰



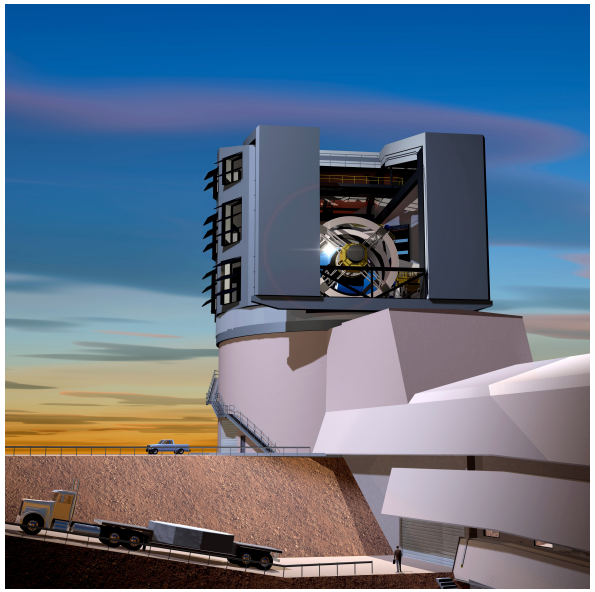
R～1.22λ/D λ～550nm, D～5m, R～0.028”



- ✓ 由3000张单次曝光的图像叠加而成的类星体图像(单次曝光 1/30s), 首先把每30幅进行叠加为1s曝光的图像，对齐后再叠加100张1s的图像。改善了信噪比，同时也改正了大气的抖动效应，因此也相应地提高了图像的分辨率。
- ✓ 这种技术分辨出4个不同成分，是同一个类星体的引力透镜效应。Subaru望远镜拍摄了它的红外图像。在红外图像中可以看到美丽的 Einstein环.

时域巡天数据的时间复杂度高

LSST



05-Mar-2024	COMP: Camera Pre-Ship Review at SLAC
30-May-2024	ComCam Reinstalled on TMA
01-Jul-2024	Dome Complete
04-Jul-2024	3-Mirror Optical System Ready for Testing
29-Aug-2024	Camera Ready for Full System AI&T
12-Dec-2024	LSSTCam Ready for On Sky (First Photon)
27-Jan-2025	System First Light with LSSTCam
23-May-2025	Test report: Final Pipelines Delivery
23-May-2025	COMP: Science Validation Surveys Complete
30-May-2025	Operation Readiness Review Complete

- LSST巡天的速度15秒曝光，视场10平方度，每晚大约20TB的原始数据
- 在十年的运行进行约20,000平方度，总曝光超过500万次，原始数据总量约为60 PB，并将产生一个20 PB的星表
- 处理后的总数据量将达到数百PB，接近1EB

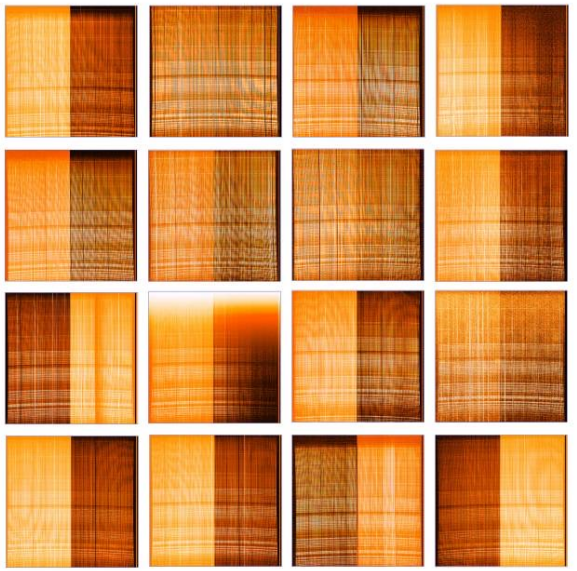
时域天文学的关键：速度+深度巡天

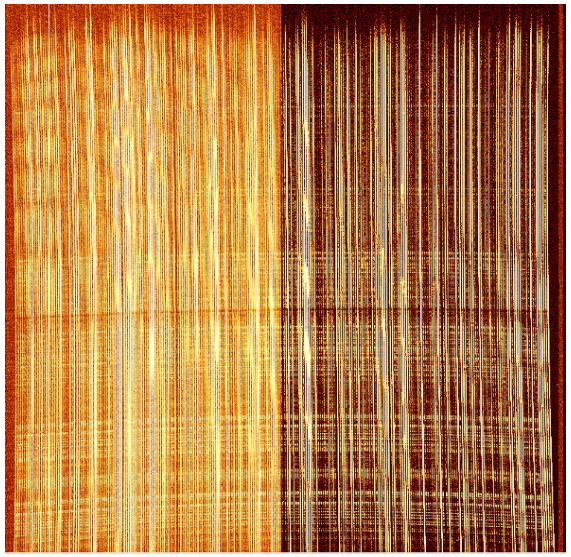
- 天文学的发展已经进入全面探索动态宇宙的新阶段，除引力波等热点问题外，也瞄准未知的新现象和新理论。
- 大天区和高频次的光学巡天是时域天文发展的重要方向。

项目	口径（米）	视场（平方度）	巡天面积（平方度）	Cadence
LSST	8.4	9.6	20000	3-4 天
ZTF	1.22	47	30000	~3天
WFST	2.5	6.5	20000	~4天
Mephisto	1.6	3.1	26000	~2周

- 技术接近瓶颈，单台望远镜已无法同时满足大天区、深星等、短时标时域巡天的需求。

大规模巡天光谱数据

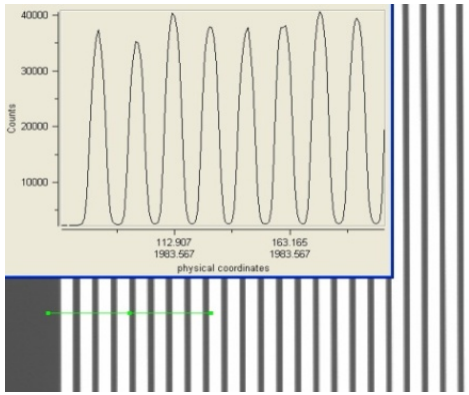




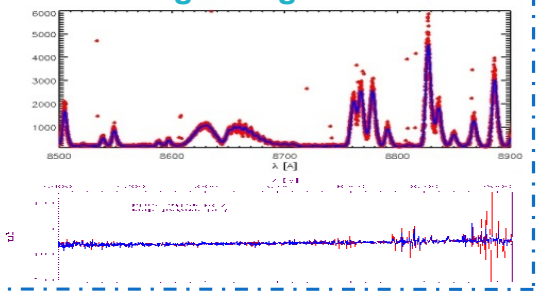
巡天光谱处理和定标的自动化

Reduction

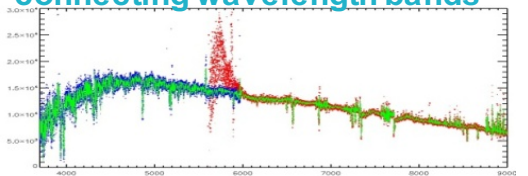
Tracing fibers & extracting spectra



Modeling the night sky & subtracting background

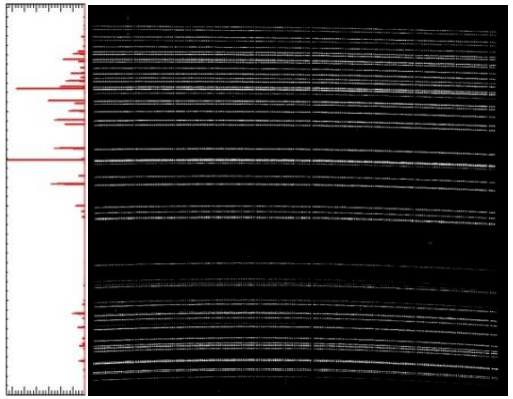


Co-adding multiple exposures & connecting wavelength bands

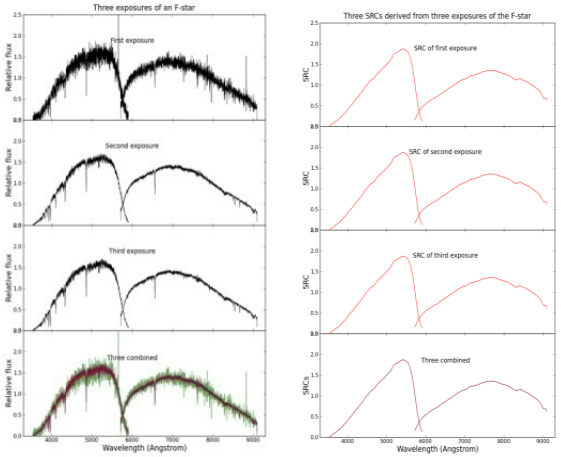


Calibration

Wavelength calibration with spectra of Arc lamps

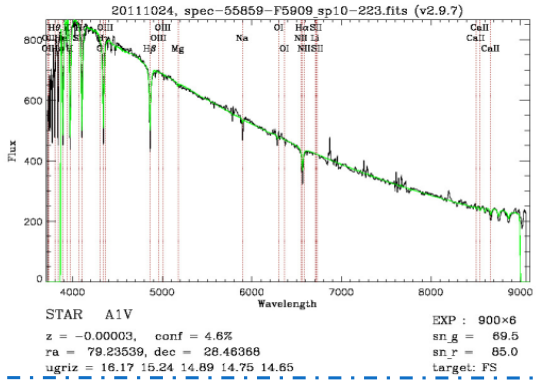


Flux calibration with standard stars (parameters are known)



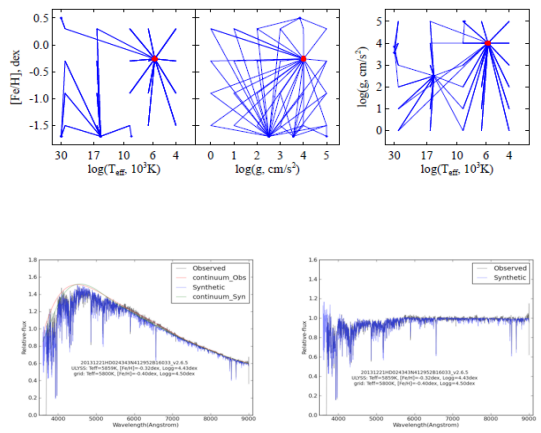
Spectral Analysis

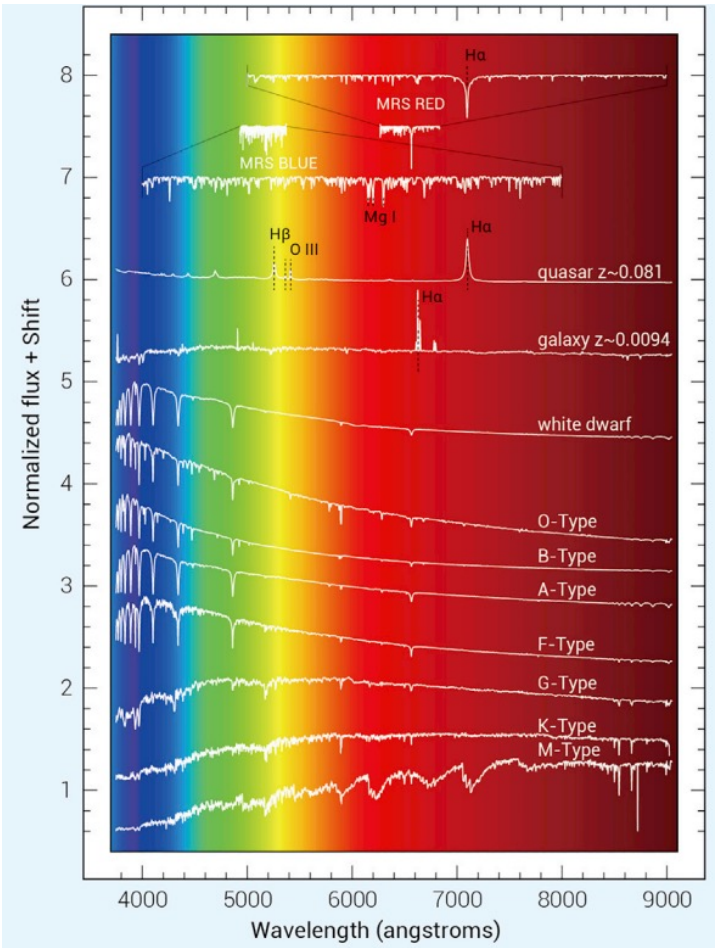
Looking for the best fit in a grid of template spectra



Stellar parameters using TGM interpolator

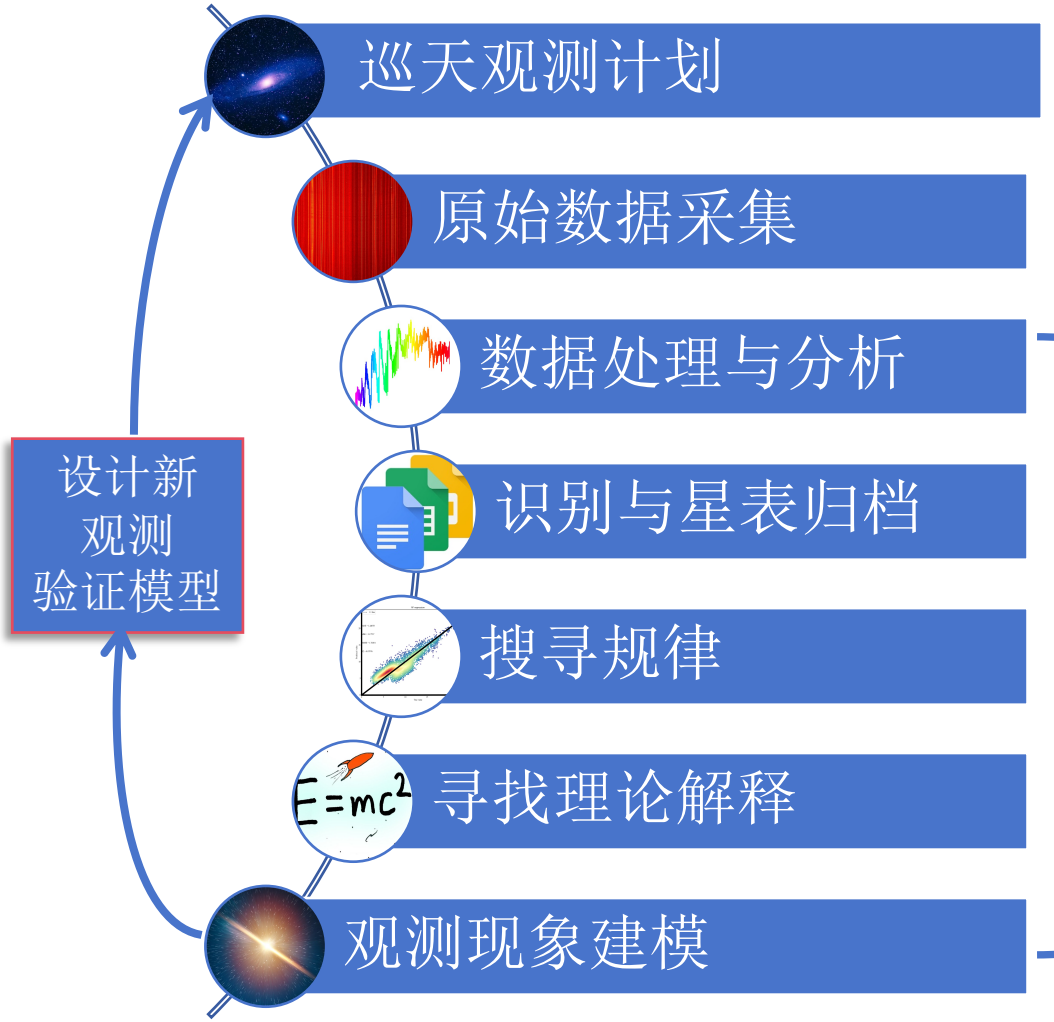
$$\mathrm{Obs}(\lambda) = \mathrm{P_n}(\lambda) \times [\,\mathrm{TGM}(\,\mathrm{T_{eff}}, \log\,\mathrm{g}, [\mathrm{Fe/H}], \lambda) \otimes \mathrm{G}(\,\mathrm{v_{sys}}, \sigma)]$$





针对天文研究链条的大数据分析技术

巡天观测的科学研究过程



对宇宙的更进一步理解

数据挖掘技术

- 信号处理
- 模式识别
- 机器学习
- 统计分析
- 可视化

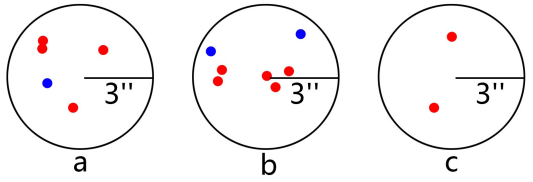
数据分析技术

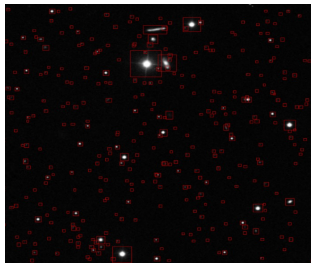
- 数据库技术
- 高性能计算

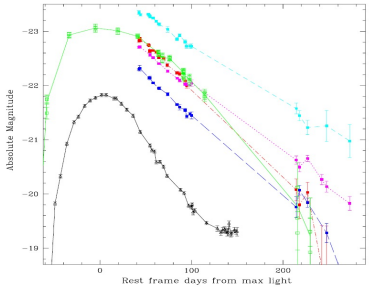
计算和数据管理技术

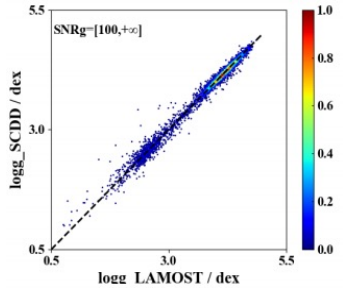
巡天大数据分析的一些共性问题

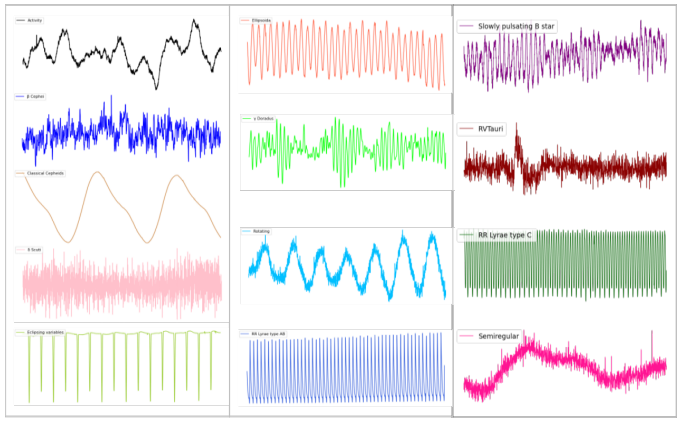
天体数据	类别	难题
数据形态	一维时变数据 二维空间数据 其它	同源证认：不同形态不同来源的数据证认为同一天体
天体类型	常见与稀有天体	数据挖掘
	变源	变源分类
	瞬变源	瞬变源预测
分辨率	低下	微小目标分类
	高清	物理参数回归











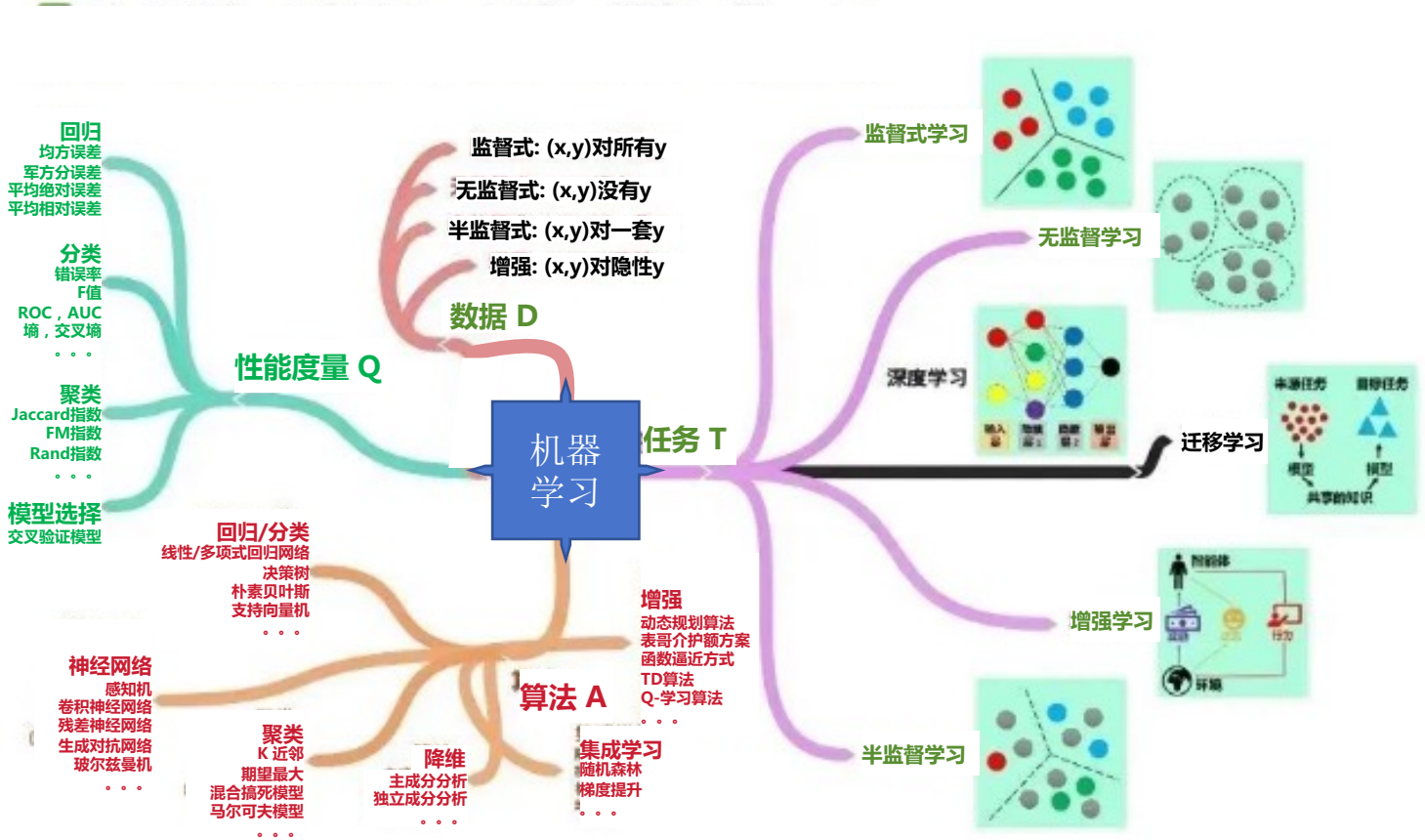
巡天数据的机器学习技术

常用的监督学习算法

Method	Accuracy	Interpretability	Simplicity	Speed
Naive Bayes classifier	L	H	H	H
Mixture Bayes classifier	M	H	H	M
Kernel discriminant analysis	H	H	H	M
Neural networks	H	L	L	M
Logistic regression	L	M	H	M
Support vector machines: linear	L	M	M	M
Support vector machines: kernelized	H	L	L	L
K-nearest-neighbor	H	H	H	M
Decision trees	M	H	H	M
Random forests	H	M	M	M
Boosting	H	L	L	L

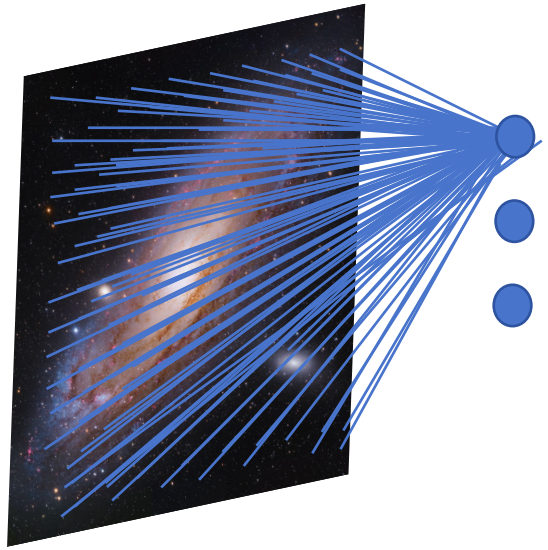
常用的非监督学习算法

Method	General Characteristics
Partitioning methods	– Find mutually exclusive clusters of spherical shape – Distance-based – May use mean or medoid (etc.) to represent cluster center – Effective for small- to medium-size data sets
Hierarchical methods	– Clustering is a hierarchical decomposition (i.e., multiple levels) – Cannot correct erroneous merges or splits – May incorporate other techniques like microclustering or consider object "linkages"
Density-based methods	– Can find arbitrarily shaped clusters – Clusters are dense regions of objects in space that are separated by low-density regions – Cluster density: Each point must have a minimum number of points within its "neighborhood" – May filter out outliers
Grid-based methods	– Use a multiresolution grid data structure – Fast processing time (typically independent of the number of data objects, yet dependent on grid size)



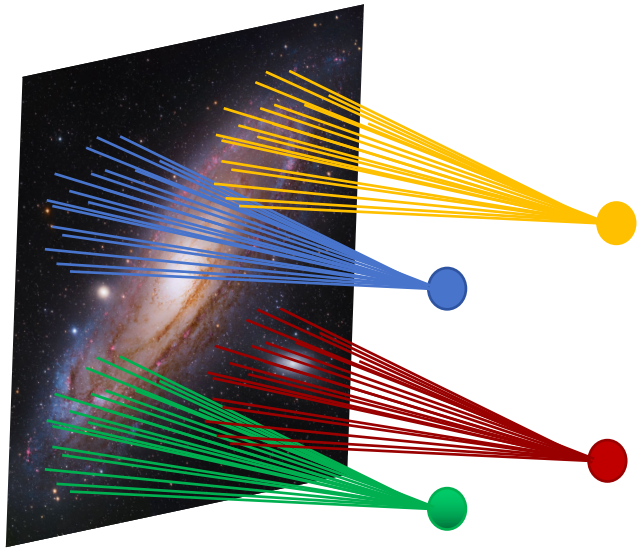
给定任务T和性能度量Q， 建立算法A从数据D中学习，
如果任务T上的性能Q随着数据D增多而改善

卷积神经网络（CNN）与图像和光谱数据

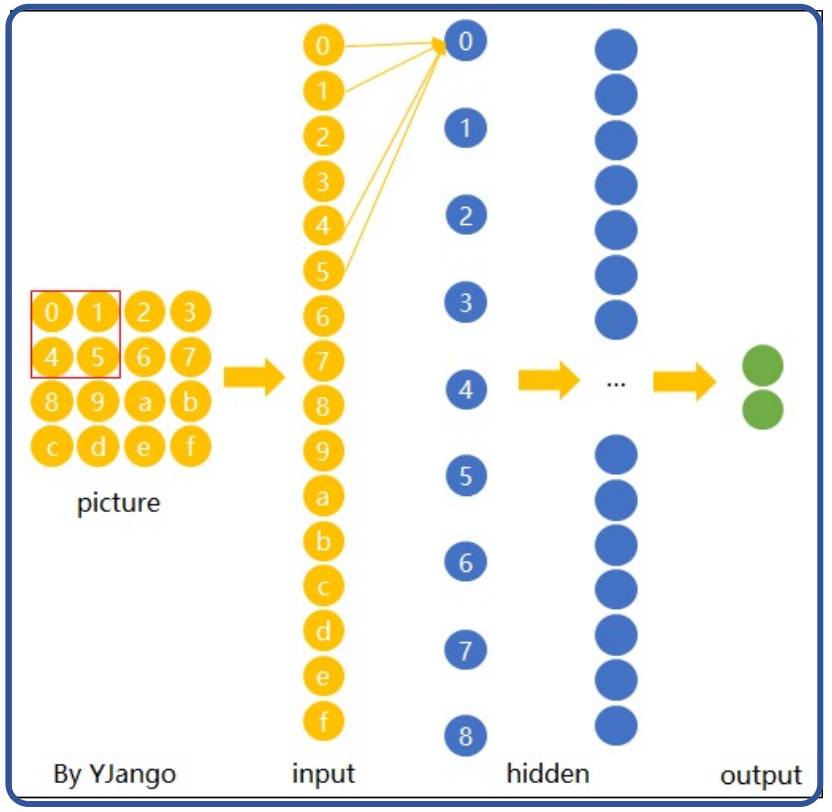


1k x 1k 图片 **全连接**
1M个隐层神经元=10^12权值

图像的空间联系是局部的，优于对全局的感受



1k x 1k 图片 **局部连接**
1M个隐层神经元，10 x10的局部感受野
100M个权值 **权值共享** **平移不变**



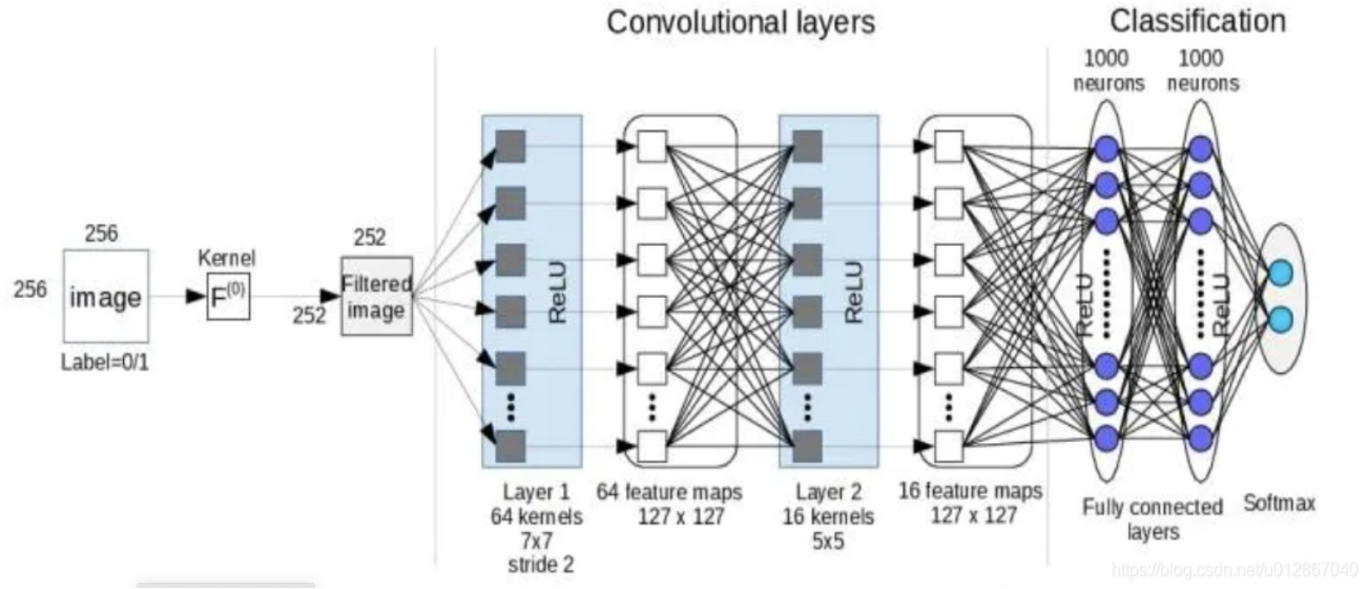
基于CNN的深度学习网络比较著名的有AlexNet，夺得ImageNet大赛2012冠军；随着深度神经网络的层数爆发式增长，带来了梯度消失和梯度爆炸等问题，使得模型训练的困难度增加。随之出现了**将神经网络某些层隔层短路**，跨层连接，被称为残差网络Residual Network(ResNet)。

CNN的“预训练+微调”概念

- 方式 I：使用具体任务的训练数据来训练深度网络学习图像表征，进而完成不同的任务（如分类、分割、目标检测等）。

- 方式 II：通常是**先用“海量”的数据让模型学到通用的图像表征，再进行下游具体任务数据的学习**，也就是**预训练+微调**的范式。

- 在CNN图像表征体系下，早期的预训练有另一个叫法是**迁移学习**，在BERT的预训练+微调范式流行之前就已经被广泛应用。迁移学习中，**传统CNN模型在做具体任务训练之前先在大量的图像任务数据集上进行预先训练**（如ImageNet分类任务数据集等）。**然后使用预训练的CNN权重初始化Backbone，并增加一些任务定制网络模块，完成在下游任务上的微调**（如Backbone+全连接层做分类任务）。

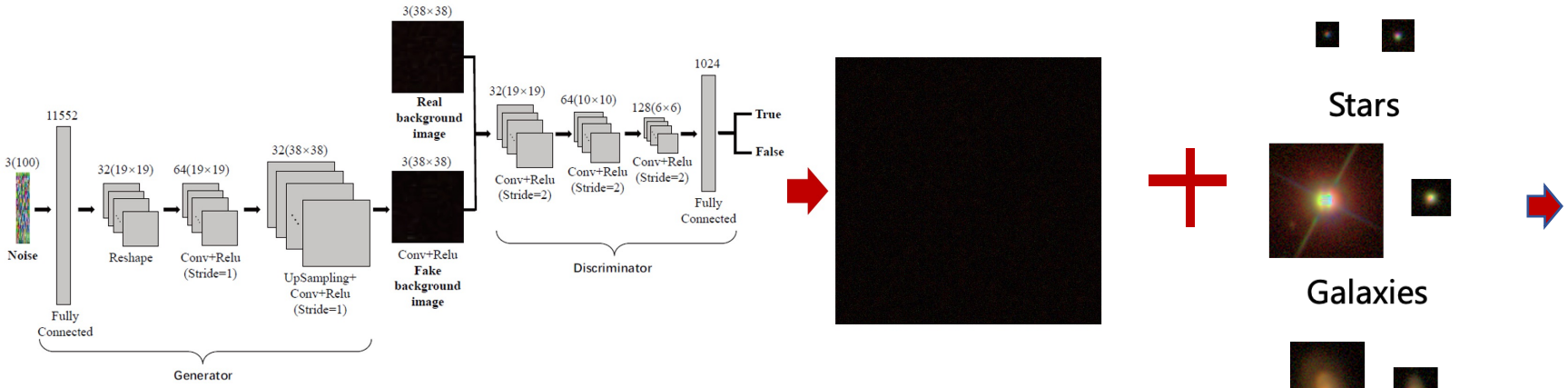


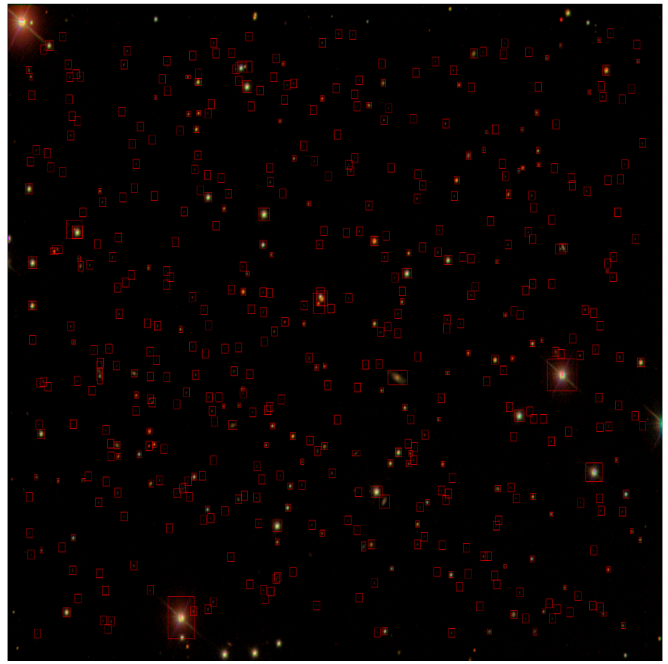
一个典型的CNN图像分析例子

目标：基于深度学习的测光星表构建

He et al. 2021 MNRAS

斯隆数字巡天（SDSS）已发布DR16版本的测光图像数据，共包含测光图像 1,630,817 x 5 张。为CSST做一个测光星表构建的AI版。



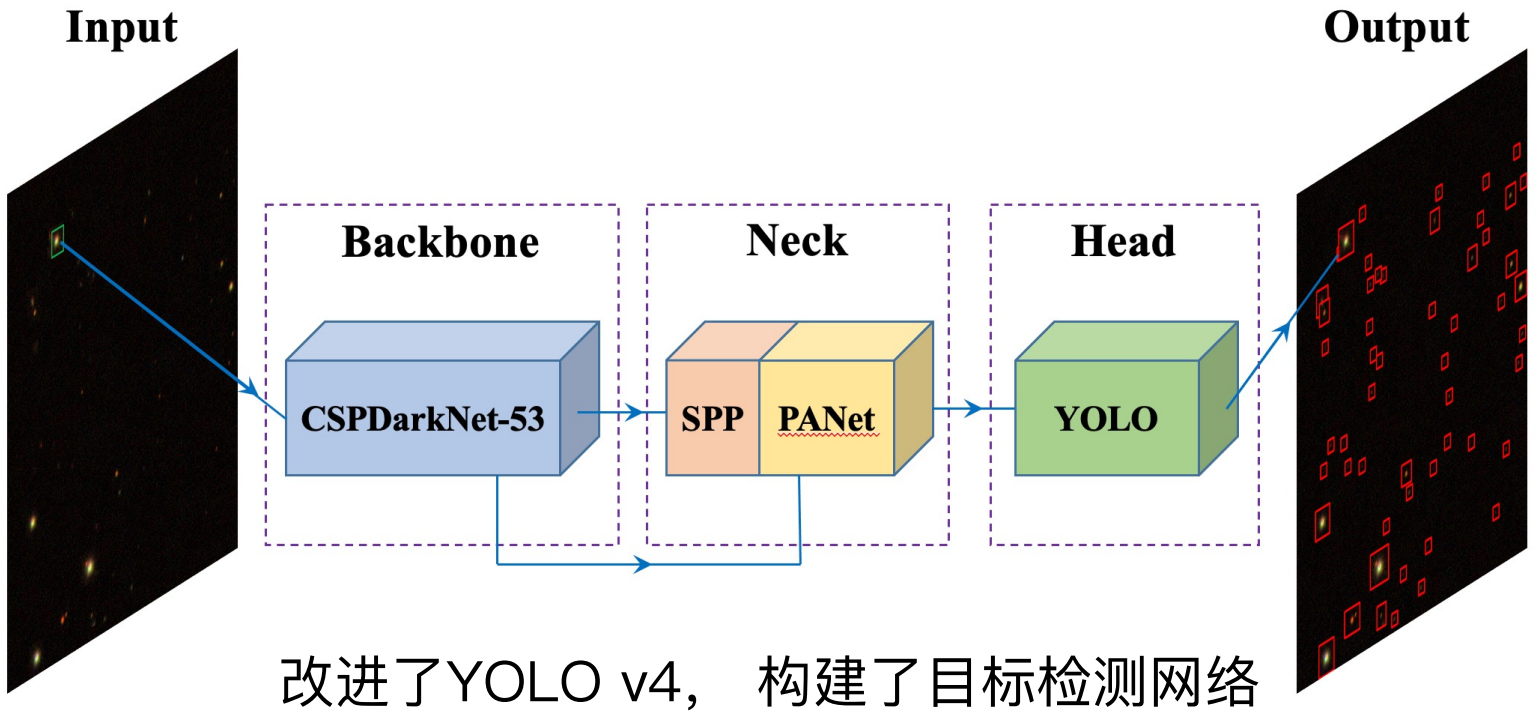


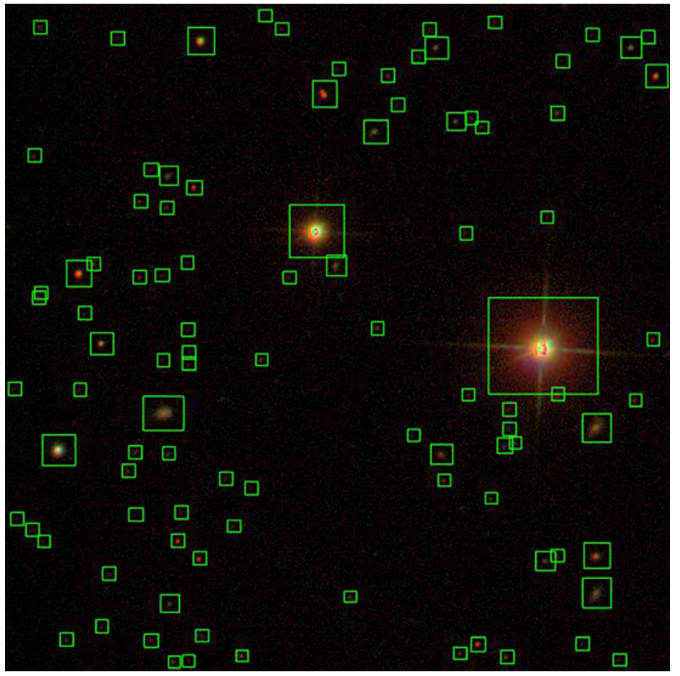
GAN网络生成训练图像样本10,000张

一个典型的CNN图像分析例子

目标：基于深度学习的测光星表构建

He et al. 2021 MNRAS





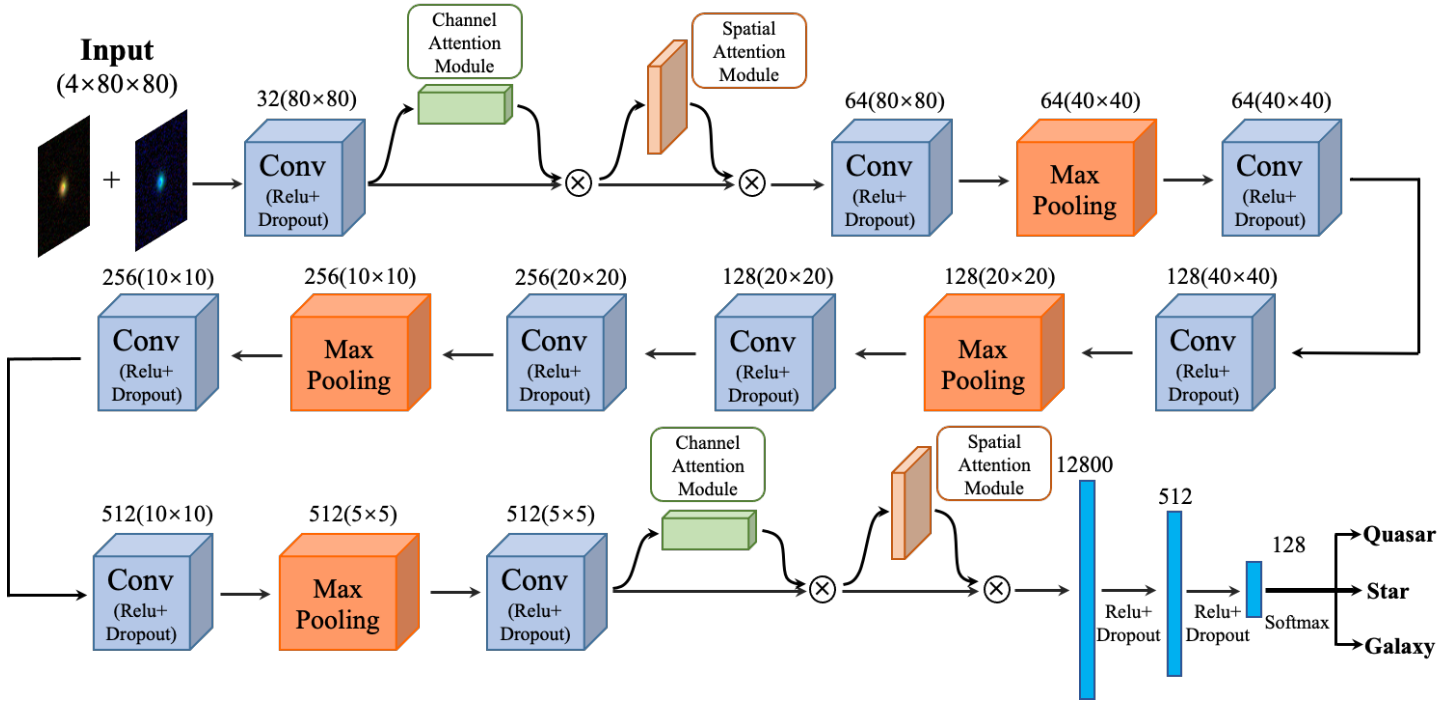
改进了YOLO v4， 构建了目标检测网络

随机选取SDSS图片编号3805，其中SDSS释放91659个目标，我们的网络得到92852个目标，位置误差约为0.87角秒。

一个典型的CNN图像分析例子

目标：基于深度学习的测光星表构建

He et al. 2021 MNRAS



		Prediction			
		Quasar	Star	Galaxy	Recall
Truth	Quasar	1894	86	20	96.6%
	Star	51	1947	2	95.7%
	Galaxy	15	1	1984	98.9%
	Precision	94.7%	97.4%	99.2%	

分类精度较高

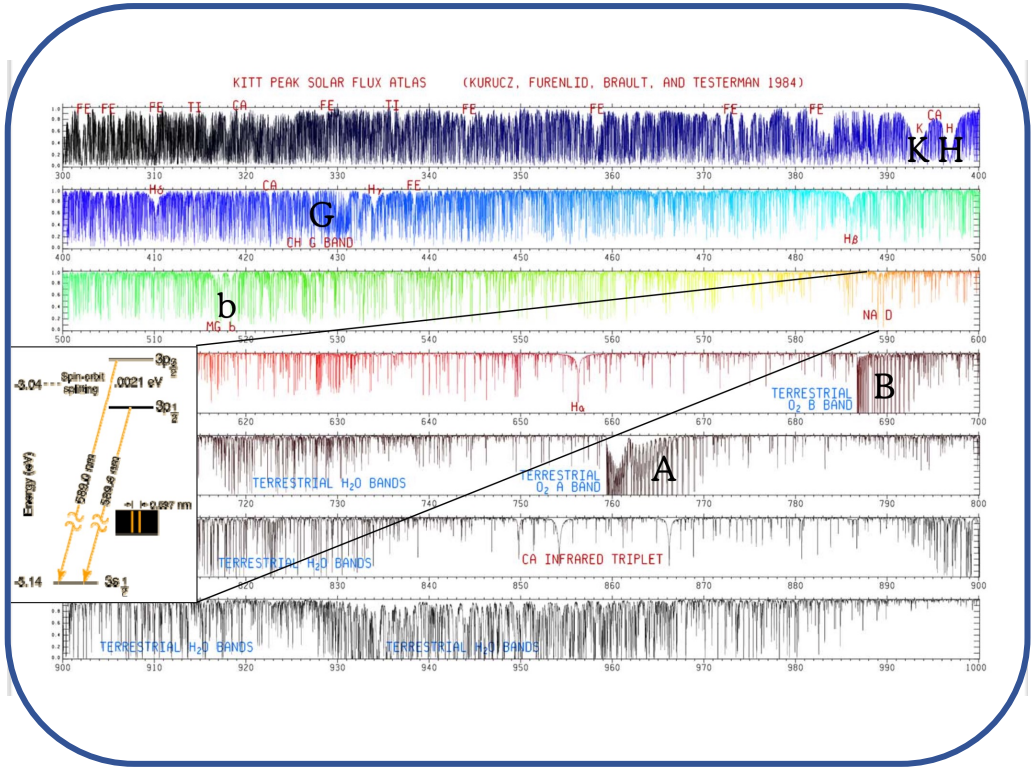
两种基于深度学习的光谱分析路线

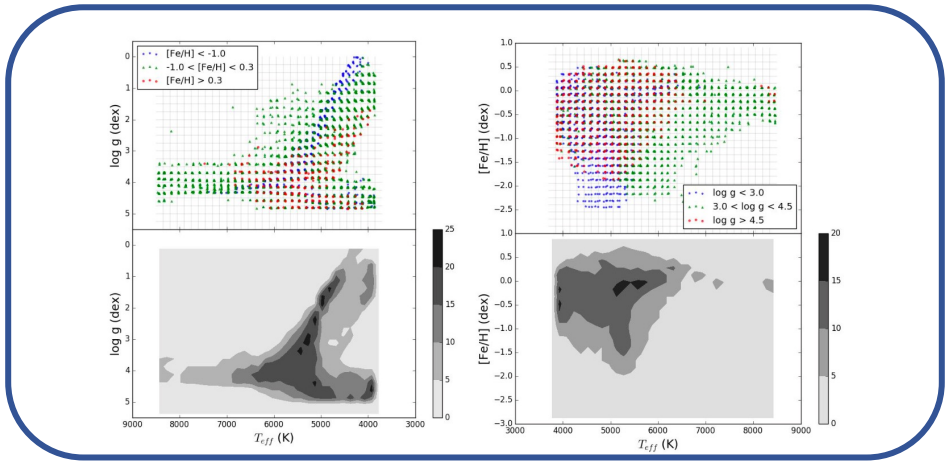
深度学习

从光谱学习物理参数

$$f(\vec{X}, \{a_1, a_2, ...\}) = \vec{y}$$

从物理参数学习光谱





Theoretical (Synthetic):

Kurucz MARCS PHOENIX

Empirical :

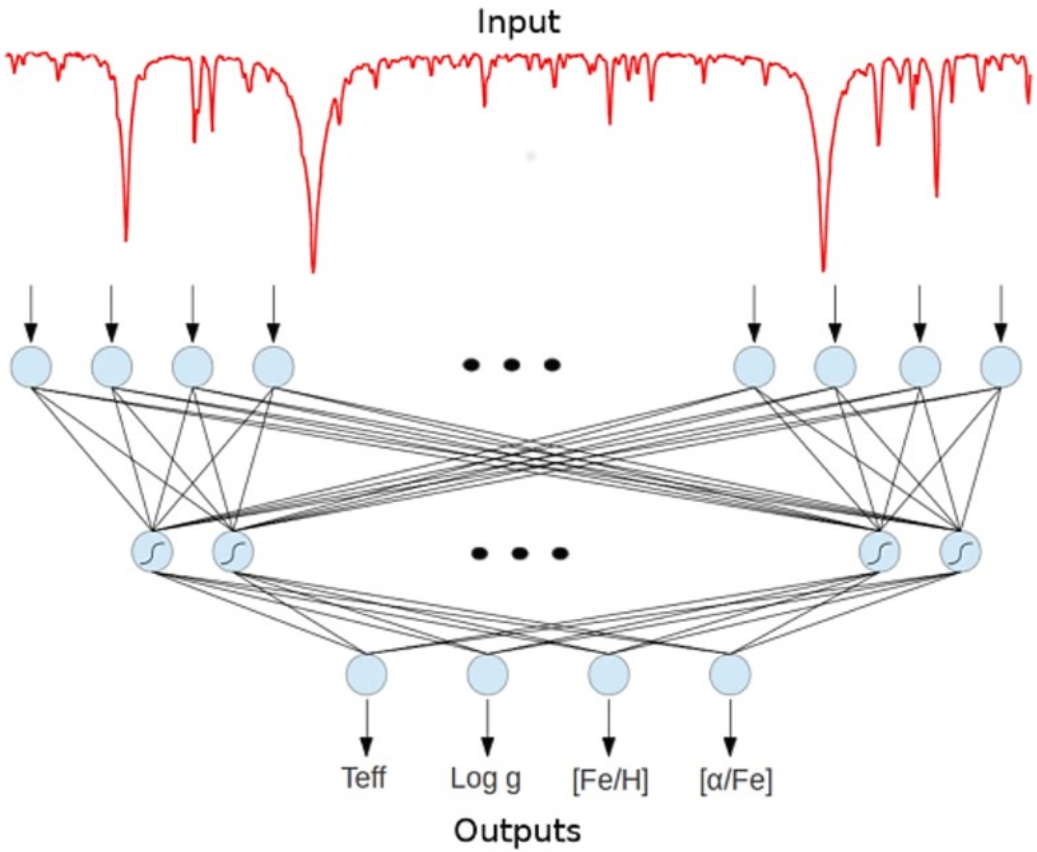
ELODIE MILES CFLIB STELIB

Apogee Mastar LAMOST

两种基于深度学习的光谱分析路线

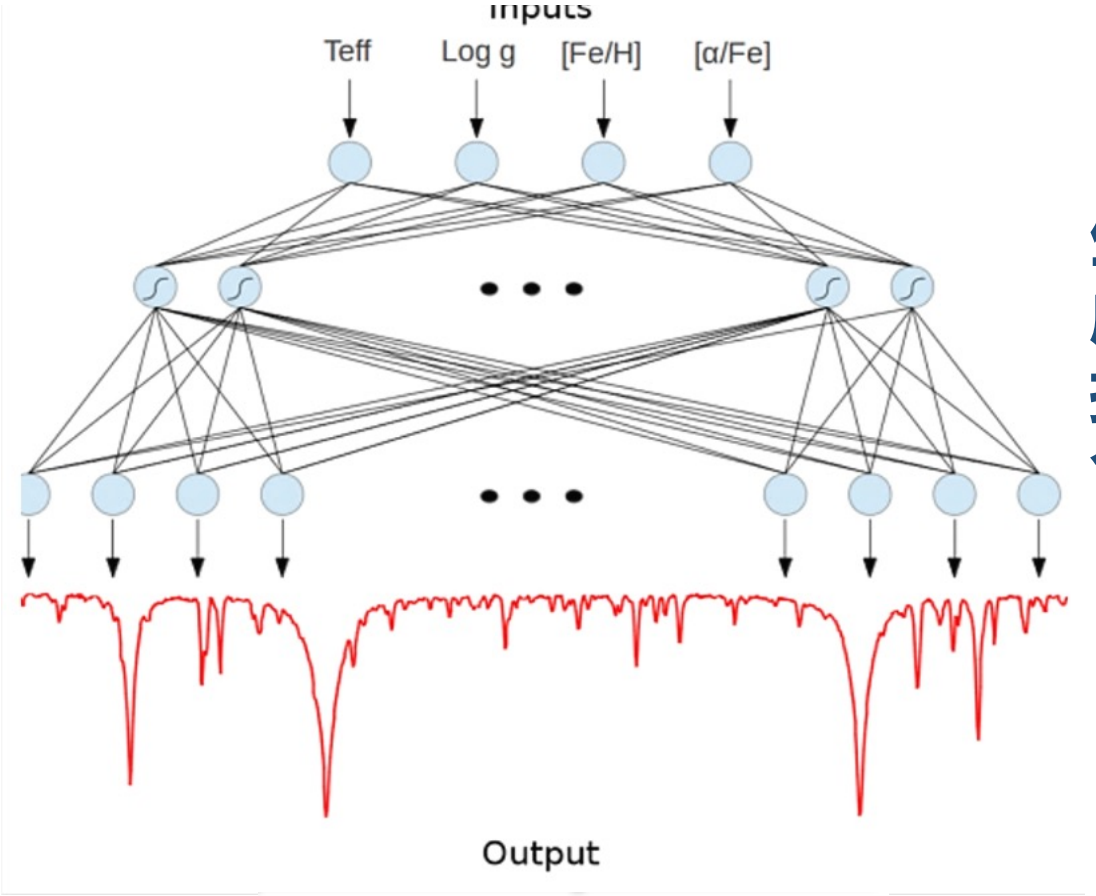
光谱生成拟合/端到端物理参数回归

直接回归



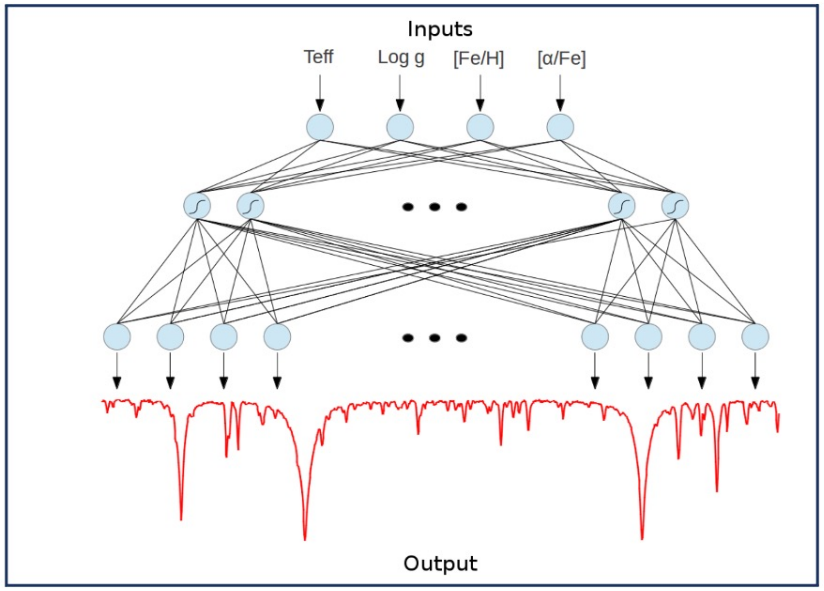
Bailer-Jones et al.1997
Manteiga et al. 2010
。 。 。

生成拟合

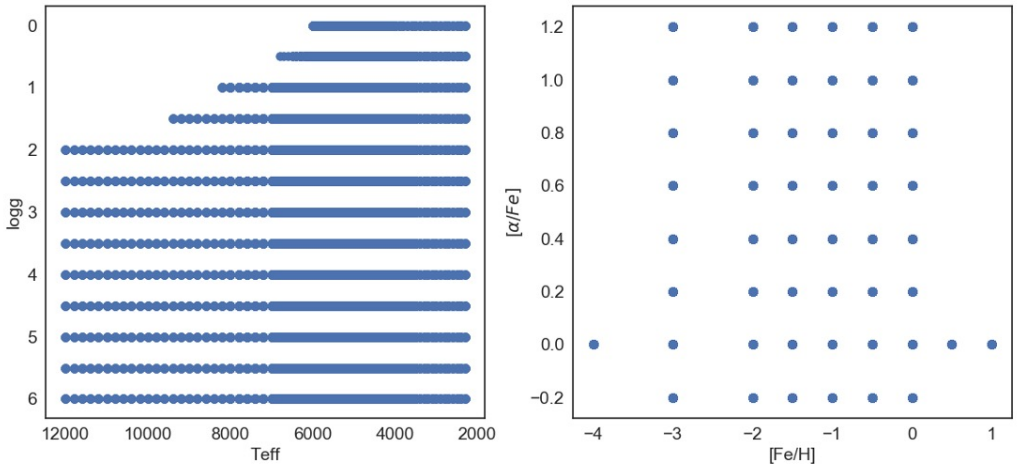


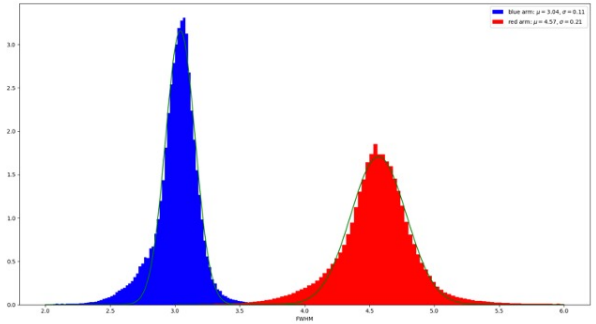
GANN: Dafonte et al. 2016
Payne: Ting et al. 2019
GSN: Wang et al.2019

光谱生成器（GSN）



模板库：PHOENIX model spectra（Husser et al. 2013）





- 蓝端FWHM: 3.04 A
- 红端FWHM: 4.75 A

Variable	Range	Step size
$T_{\rm eff}$ [K]	2300–7000	100
	7000–12 000	200
log g	0.0–+6.0	0.5
[Fe/H]	−4.0–−2.0	1.0
	−2.0–+1.0	0.5
[α/Fe]	−0.2–+1.2	0.2

利用光谱生成器进行MC采样生成

贝叶斯估计（Bayesian）

给定一条待测光谱，在参数空间通过MC采样5000次:

$T_{eff} \sim U(T_{eff\_LASP} - 3*err_{Teff\_LASP},\ T_{eff\_LASP} + 3*err_{Teff\_LASP})$,

$\log g \sim U(\log g_{\_LASP} - 3*err_{\log g\_LASP},\ \log g_{\_LASP} + 3*err_{\log g\_LASP})$,

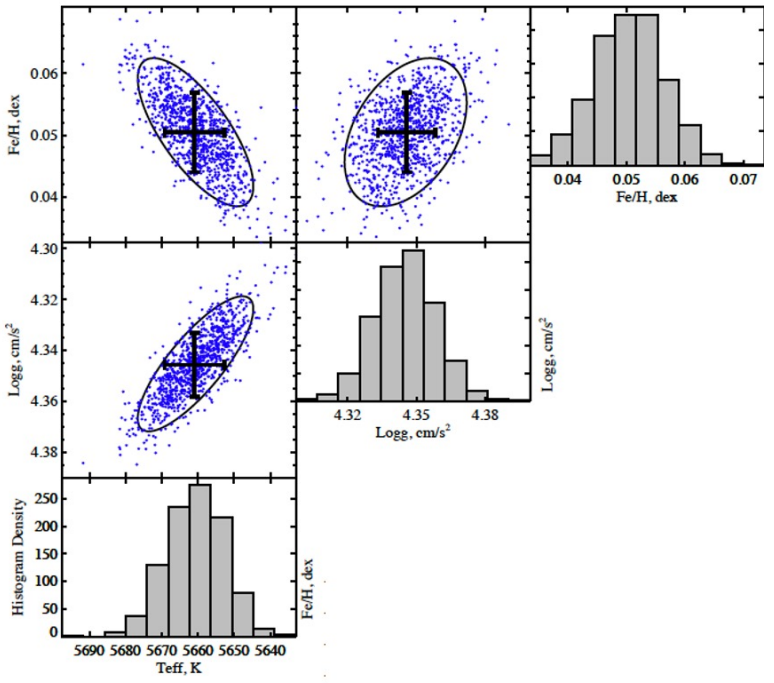
$[Fe/H] \sim U([Fe/H]_{LASP} - 3*err_{[Fe/H]\_LASP},\ [Fe/H]_{LASP} + 3*err_{[Fe/H]\_LASP})$,

$[\alpha/Fe] \sim U(\mu-0.3, \mu+0.3)$ 这里，$\mu = \begin{cases} 0, & 0 < [Fe/H] \\ 0.2, & -1 < [Fe/H] < 0 \\ 0.4, & [Fe/H] < -1 \end{cases}$

$$P(AP \mid S) = \frac{P(S \mid AP)P(AP)}{P(S)}$$

$$P(S \mid AP) = e^{-d/2}, \quad d = \sum_{\lambda_0}^{\lambda_n} \left(\frac{obs_\lambda - f_\lambda(AP) \times P_n(\lambda) - \mu_{pixerr}}{\sigma_i}\right)^2$$

$P(AP)$ 为恒星参数的先验分布



生成光谱的拟合结果

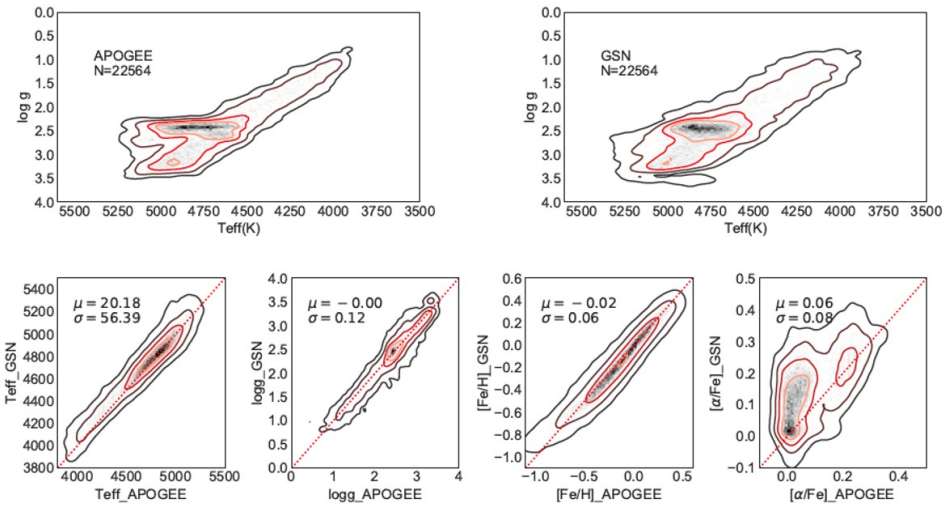
结果分析

LAMOST DR5 交叉 APOGEE DR14:

共计：81,131

符合以下条件： 22,564

- STARFLAG, ASPCAPFLAG
- APOGEE星表T_{eff}, log g, [Fe/H], [α/Fe]≠-9999
- S/N_g>=30

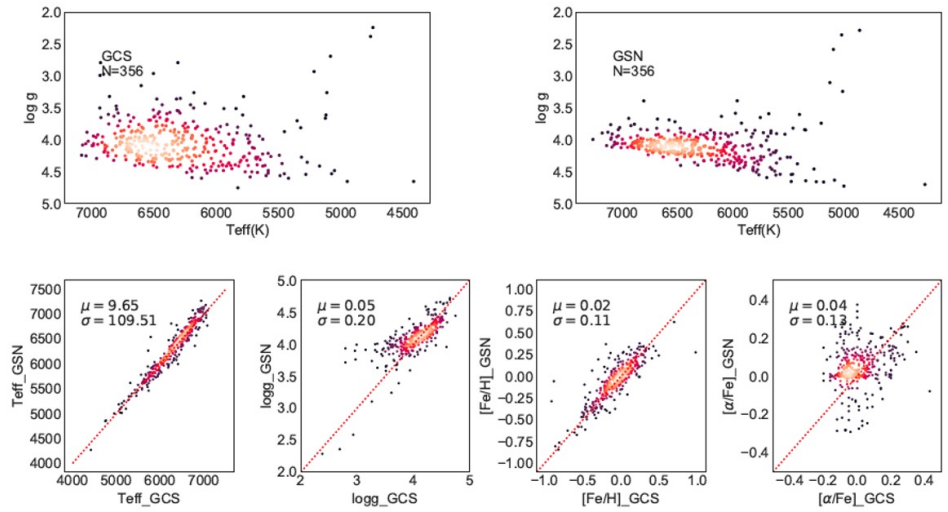


LAMOST DR5 交叉 GCS（Casagrande et al.2011）：

共计：553

符合以下条件： 356

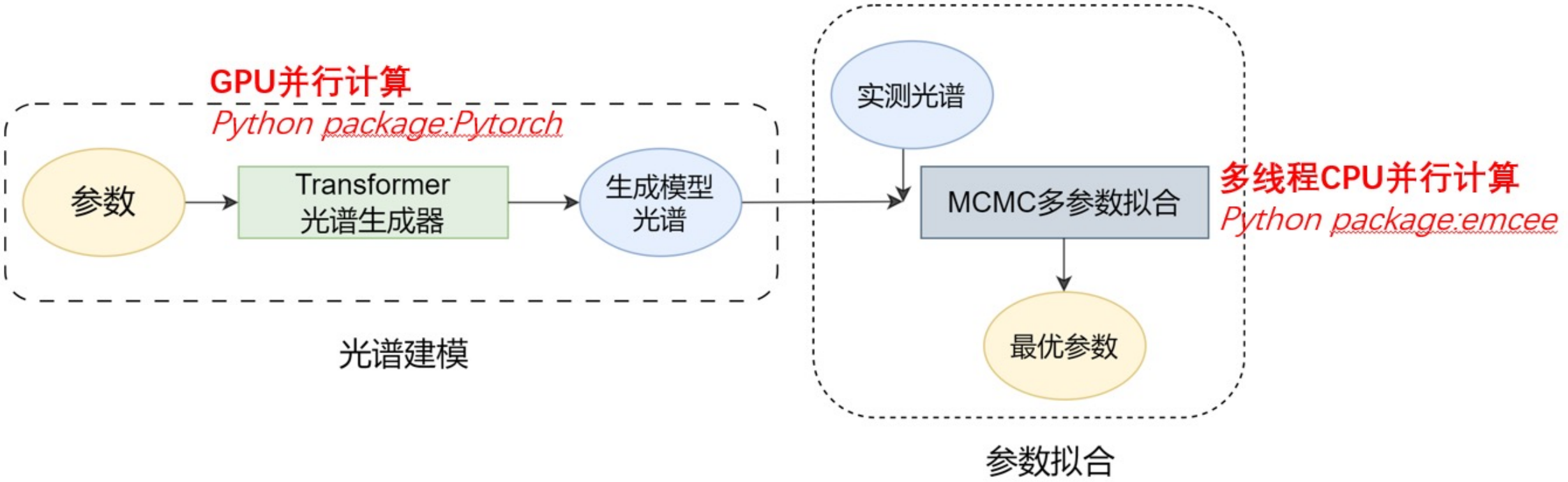
- GCS parameters (T_{eff}, log g, [Fe/H], [α/Fe]) 非空
- S/N>=30



基于 Transformer的光谱生成和MCMC拟合

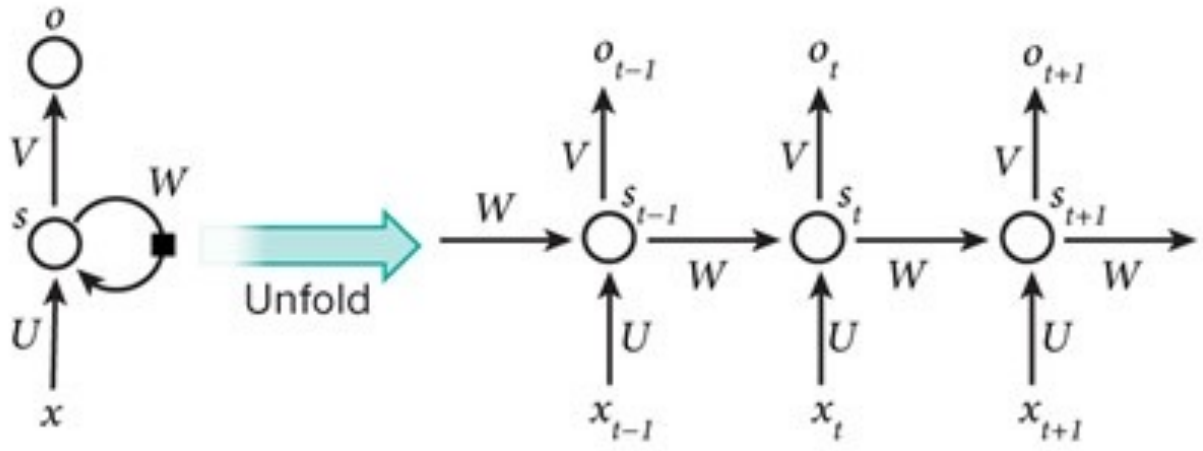
通过高维度参数空间光谱建模，从LAMOST、DESI光谱同时测定恒星大气参数和30多种化学元素丰度

1. 参数到光谱：
   利用Transformer算法+恒星大气物理模型进行高维度（~40维）光谱精确正向（生成式）建模
2. 光谱到参数：利用MCMC进行多参数（~40）拟合



循环神经网络RNN与时间序列

- 循环神经网络(RNN)是一种时间递归神经网络，适合序列数据。
- RNN的一个序列当前的输出与前面的输出也有关。RNN跟传统神经网络最大的区别在于每次都会将前一次的输出结果，带到下一次的隐藏层中，一起训练。

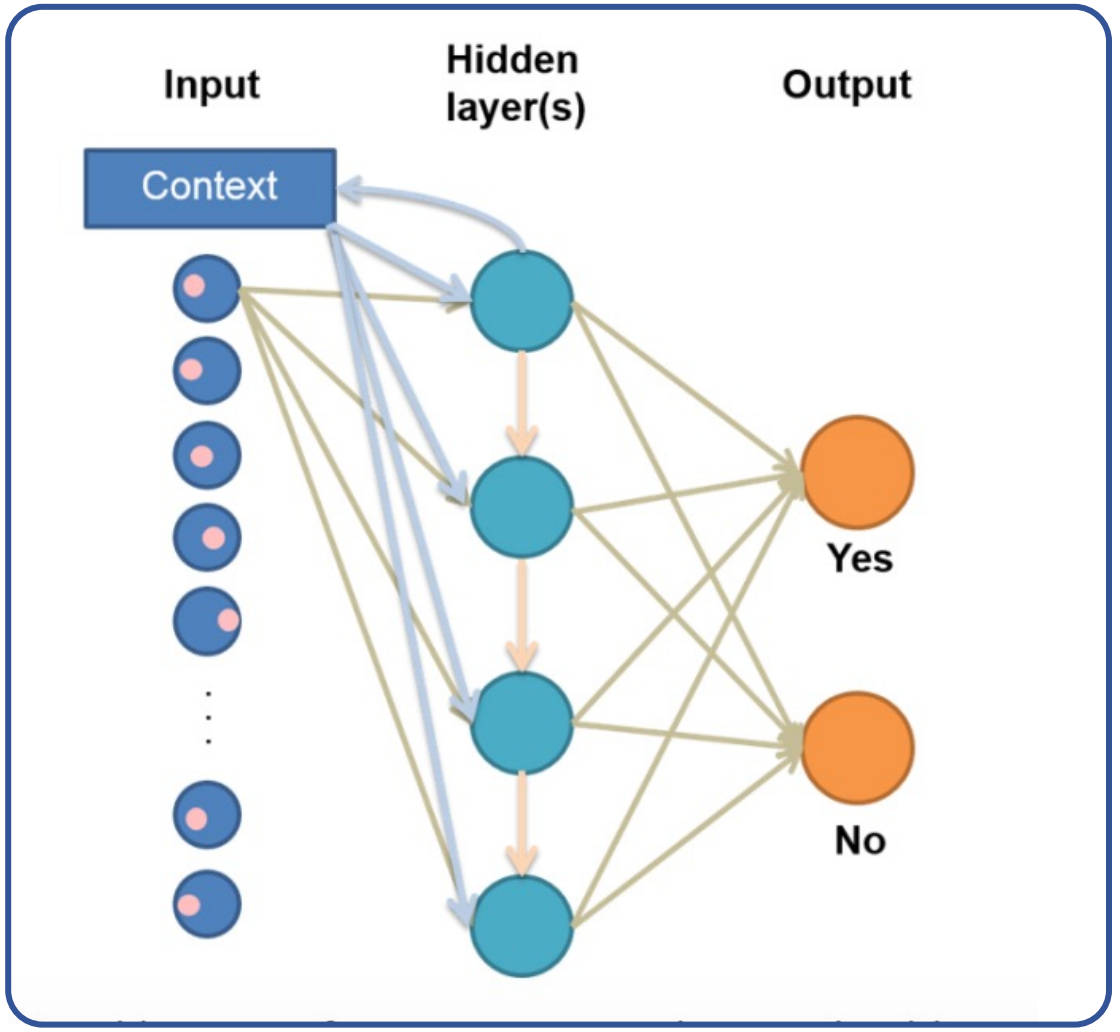


X 输入层的值
S 隐藏层的值，
U是输入层到隐藏层的权重矩阵，
0是输出层的值。
V是隐藏层到输出层的权重矩阵。

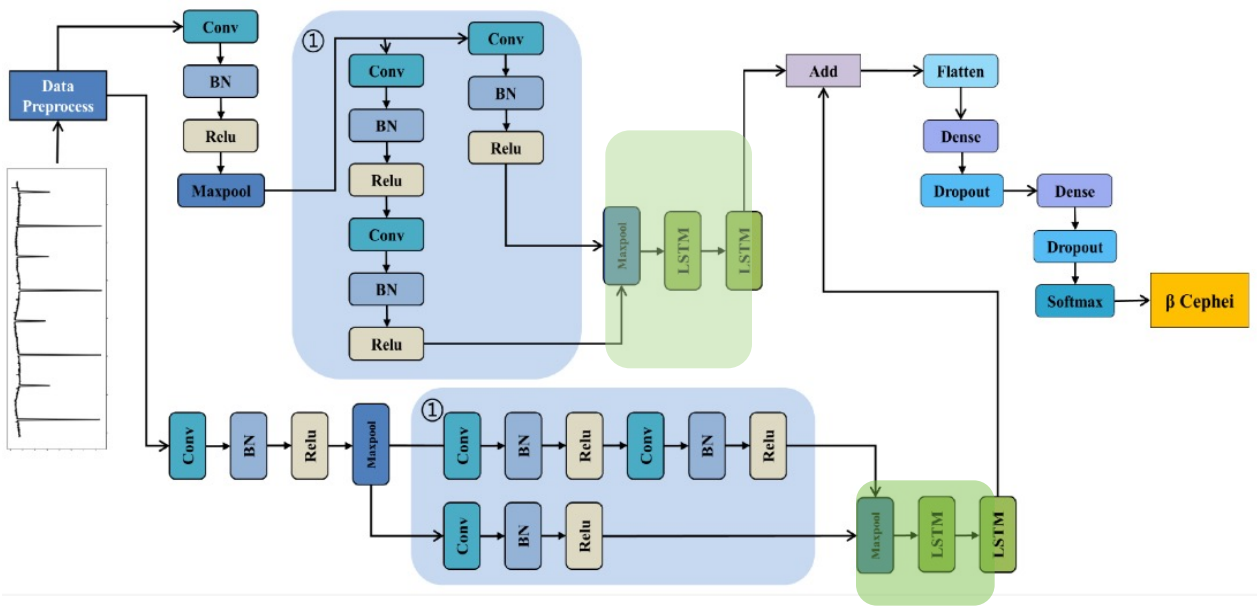
循环神经网络的隐藏层的值s不仅仅取决于当前这次的输入x，还取决于上一次隐藏层的值s。权重矩阵W就是隐藏层上一次的值作为这一次的输入的权重。

长短时记忆网络LSTM

- 为了解决RNN的短时记忆问题，提出了一种基于RNN的优化算法LSTM。
- LSTM可以避免梯度消失和爆炸。
- LSTM有三个门（遗忘门、输入门和输出门）来控制历史信息的作用，使整个网络能够更好地掌握序列信息之间的关系。
- LSTM选择记住重要信息，过滤掉噪声信息，减少了内存负担。



LSTM +CNN 的光变曲线分类案例

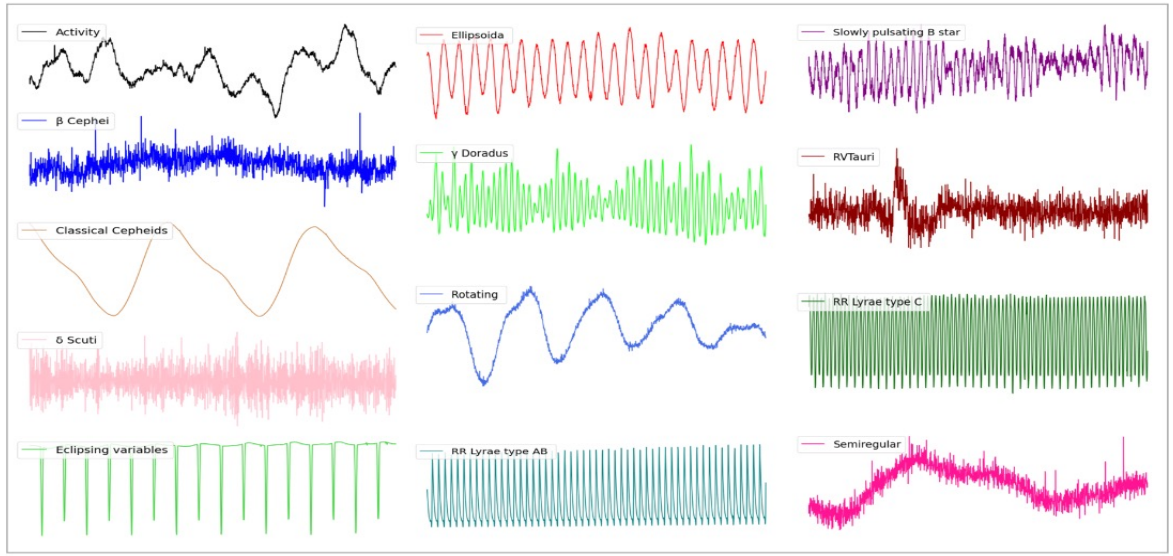


- 该方法包含两个相同的分支，每个分支由CNN结构和RNN结构组成。
- 每个分支内的CNN从输入数据中提取空间特征，从而检测局部模式和关系。
- 每个分支中的RNN捕获时间依赖性并学习连续数据中的长期模式并减少了内存负担。

光变曲线的分类：

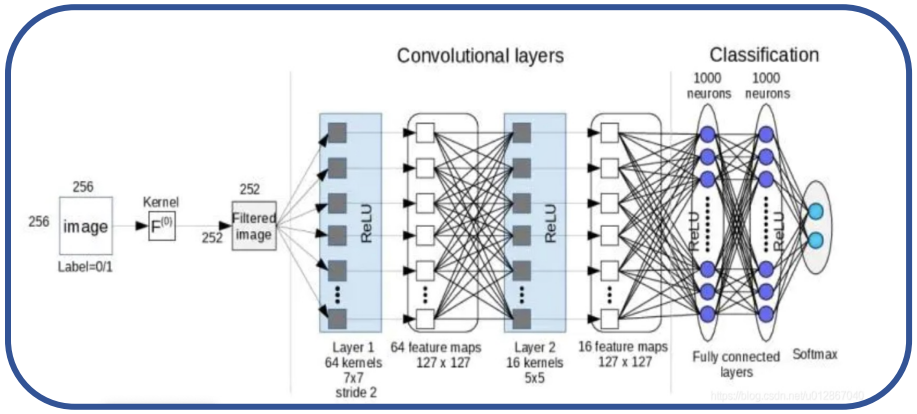
Activity, β Cephei, Classical Cepheids, ζ Scuti, Eclipsing variables, Ellipsoida, γ Doradus, Rotating, RR Lyrae type AB, RR Lyrae type C, RVTauri, Slowly pulsating B star, and Semi-regular classes.

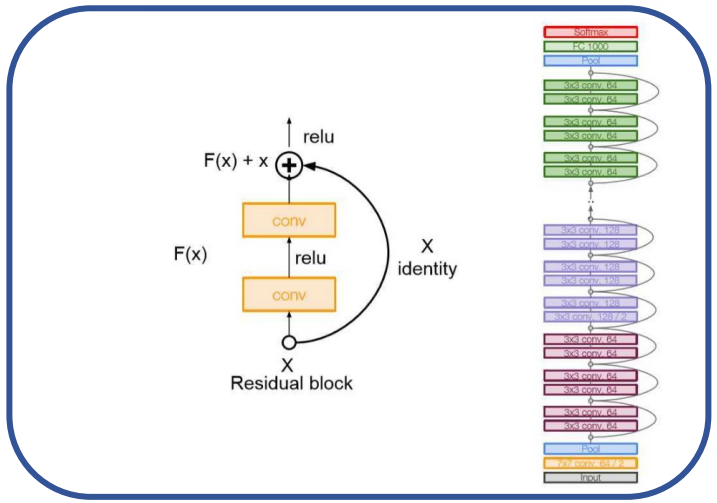
Yan J. et al 2023

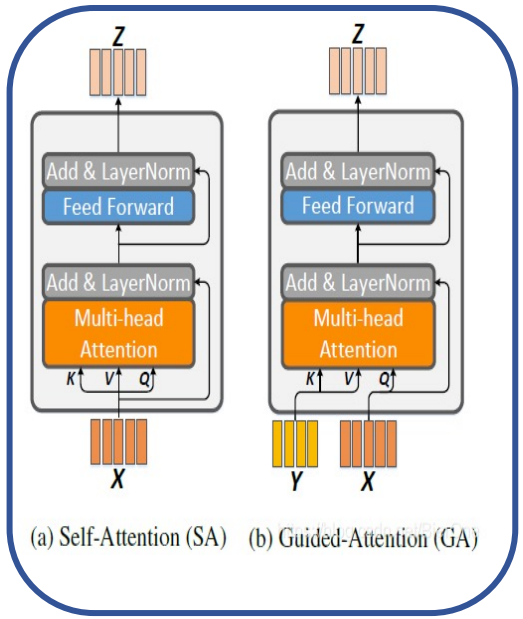


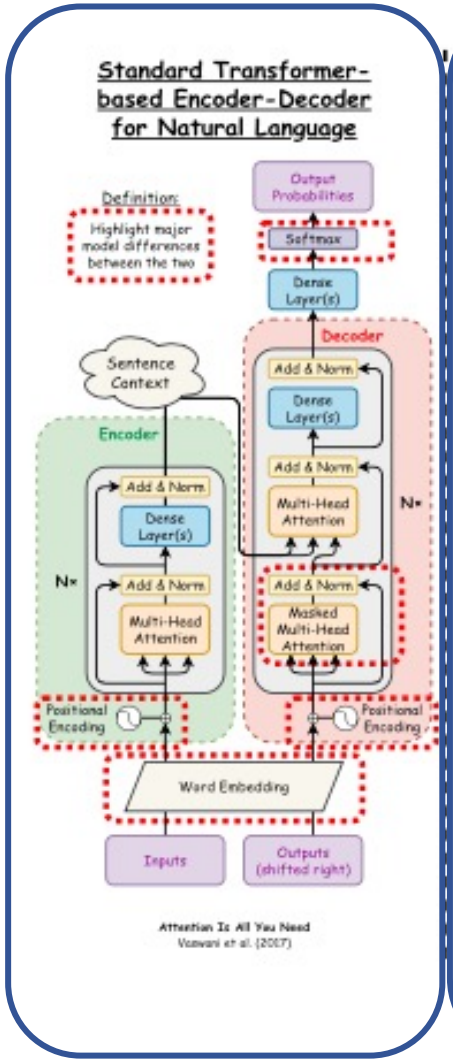
从判别式到生成式的人工智能

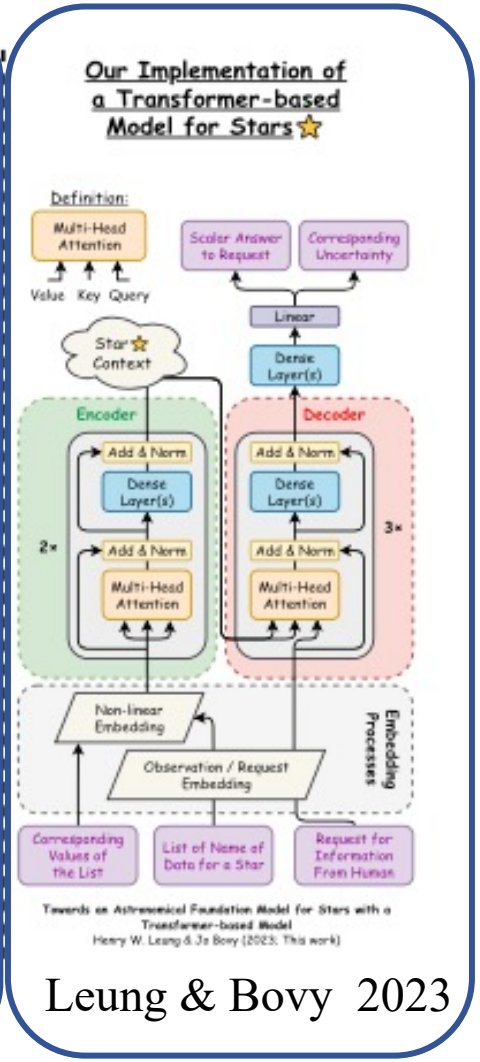
卷积神经网络的演化：对数据的压缩比大幅提高（复用性的提升）







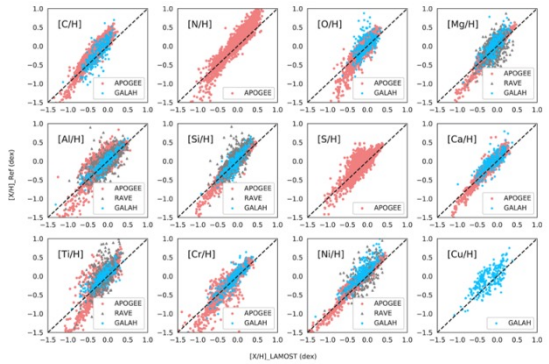




Leung & Bovy 2023

从深度学习CNN到大模型Transformer

基于卷积神经网络的深度学习网络 SPCANet



THE ASTROPHYSICAL JOURNAL

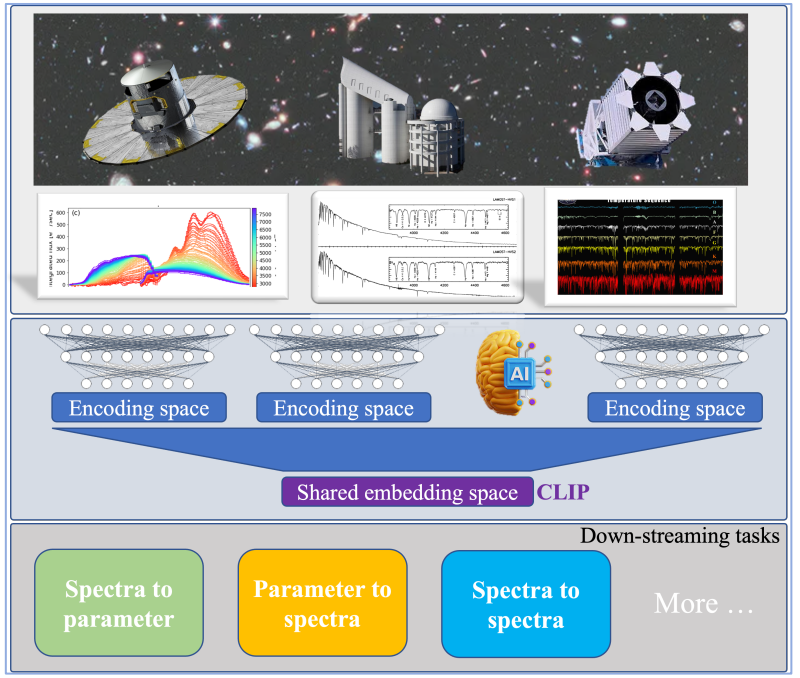
SPCANet: Stellar Parameters and Chemical Abundances Network for LAMOST-II Medium Resolution Survey

Rui Wang[1,2], A-Li Luo[1,2,3,4], Jian-Jun Chen[1], Wen Hou[1], Shuo Zhang[1,2], Yong-Heng Zhao[1,2], Xiang-Ru Li[5], Yong-Hui Hou[2,6], and LAMOST MRS Collaboration[1,4,7,8,9,10,11,12,13]

Published 2020 February 28 • © 2020. The American Astronomical Society. All rights reserved.
The Astrophysical Journal, Volume 891, Number 1
Citation Rui Wang *et al* 2020 *ApJ* **891** 23
DOI 10.3847/1538-4357/ab6dea

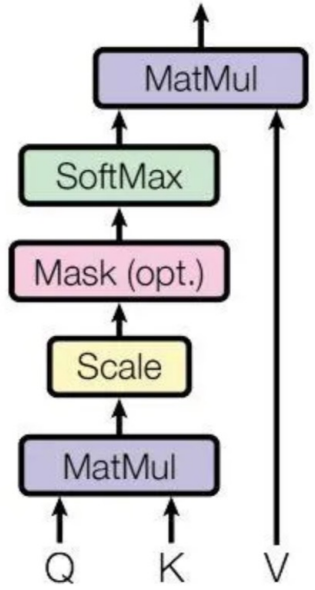
基于卷积神经网的SPCANet从LAMOST海量光谱获得精确元素丰度

基于transformer的光谱基座模型 AstroOne_Spec

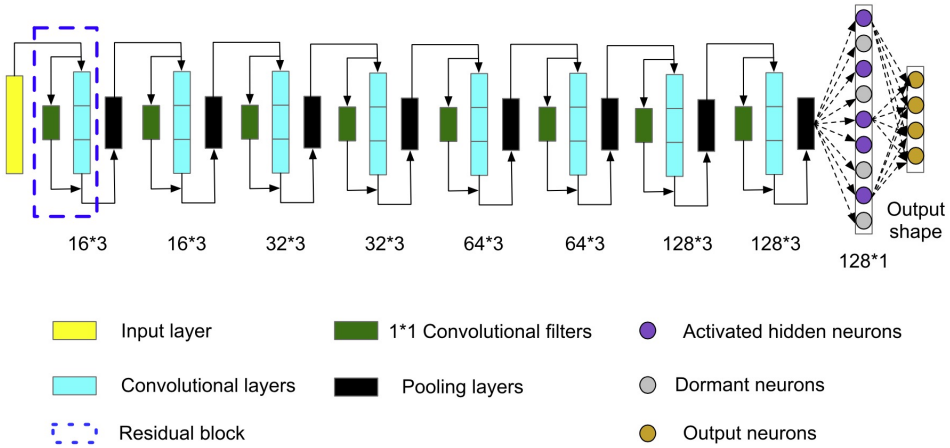


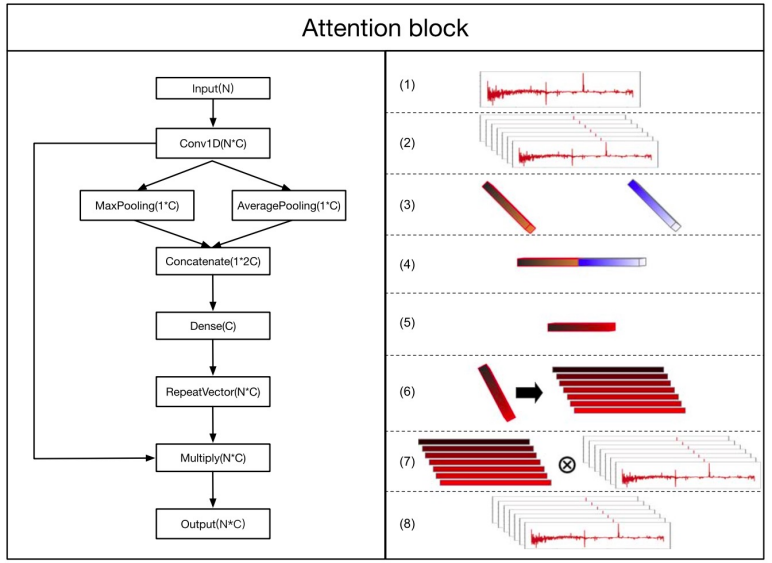
基于Transformer的SpecCLIP分析恒星的各种分辨率光谱，获得大气参数和元素丰度

Transformer架构的特点适合序列数据



Attention机制适合光谱数据　Zou et al. (PASP, 2020)



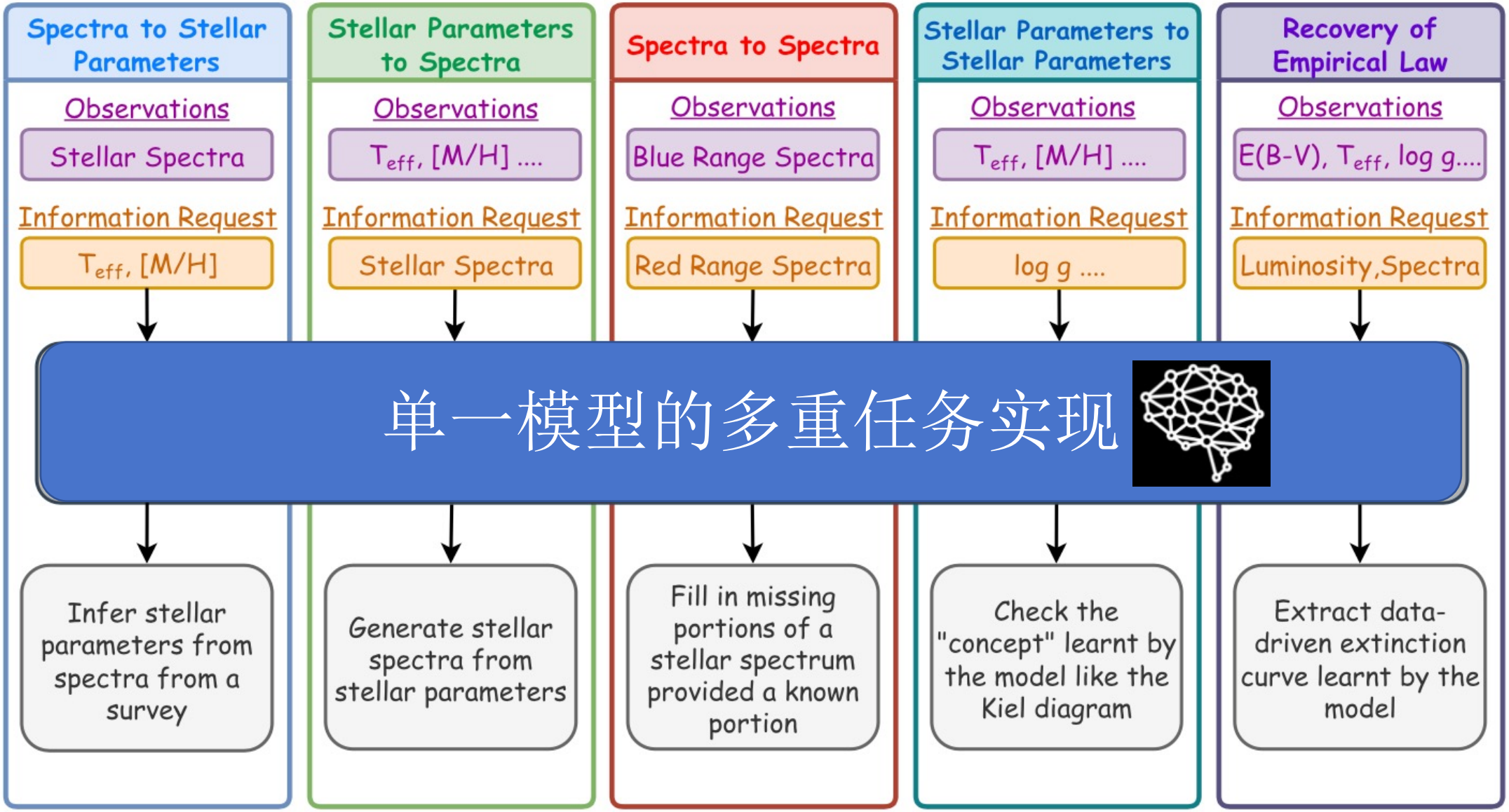


Encoder是由多头Attention和全连接神经网络构成：

- 编码器输入的句子首先会经过一个自注意力层。
- 解码器中的解码注意力层的作用是关注输入句子的相关部分，使用**多头注意力机制**。
- **并行计算**: 自注意力可以并行计算，可以有效地在硬件上进行加速。(因为在自注意力中可以并行的计算得分)
- **长距离依赖捕捉**:自注意力可以更好地处理长距离依赖关系，因为它不需要按顺序处理输入序列。

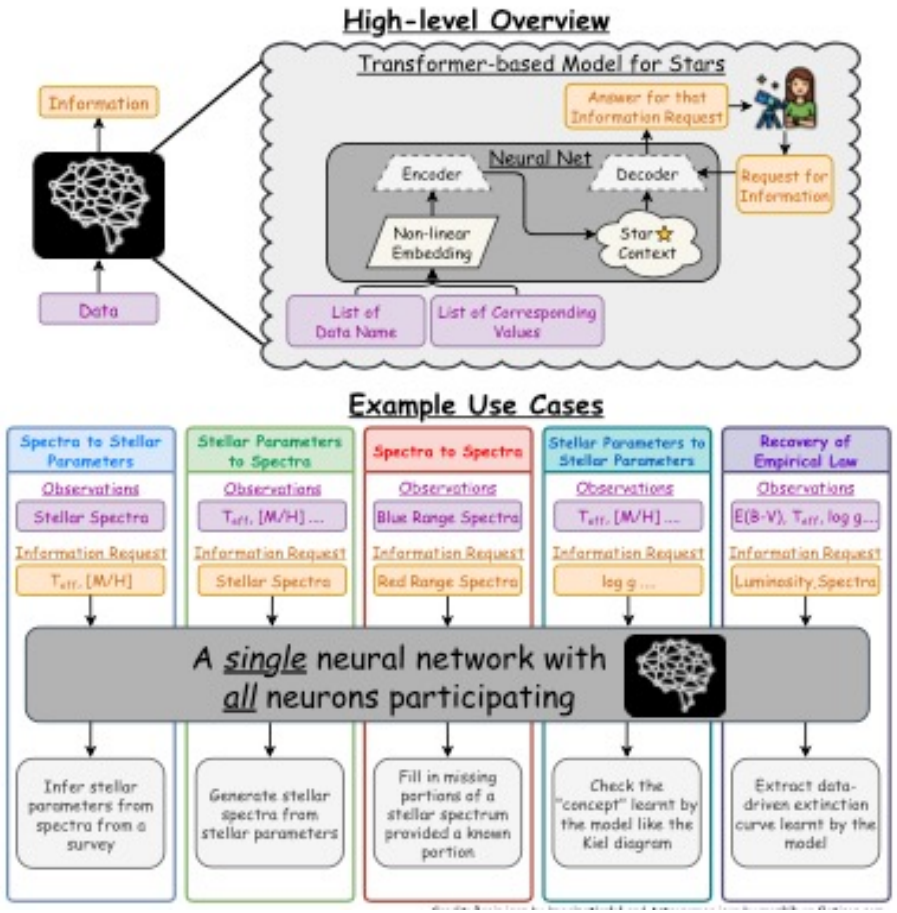
Model	Parameters	Data Set 1	Data Set 2	Data Set 3
1D SSCNN	5.18M	92.85%	98.45%	92.33%
C-Net	2.89M	88.75%	98.58%	91.97%
RC-Net	**0.388M**	93.33%	98.81%	93.07%
RAC-Net	0.623M	**93.52%**	**98.92%**	**93.25%**

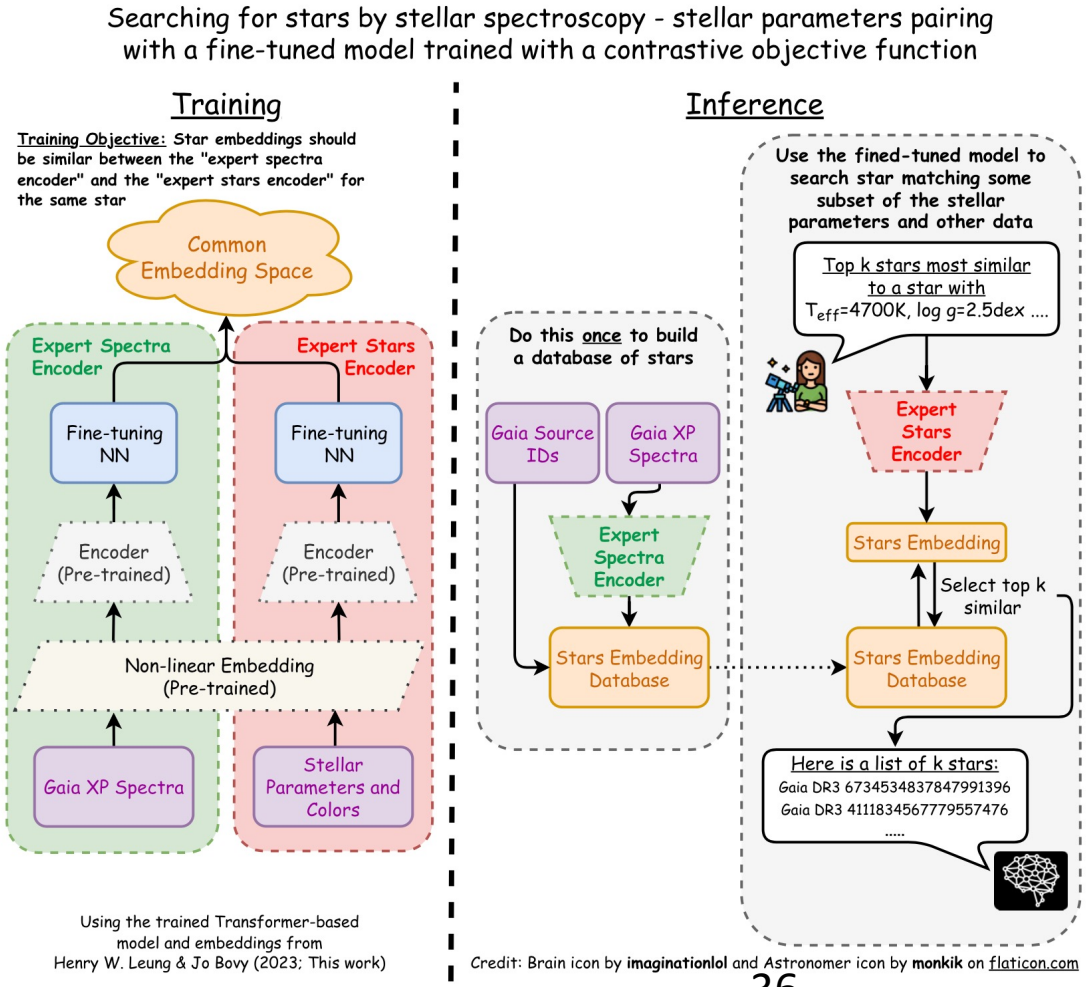
基于Transformer的恒星光谱基础模型



大样本恒星参数的神经网络回归

Transformer 大模型

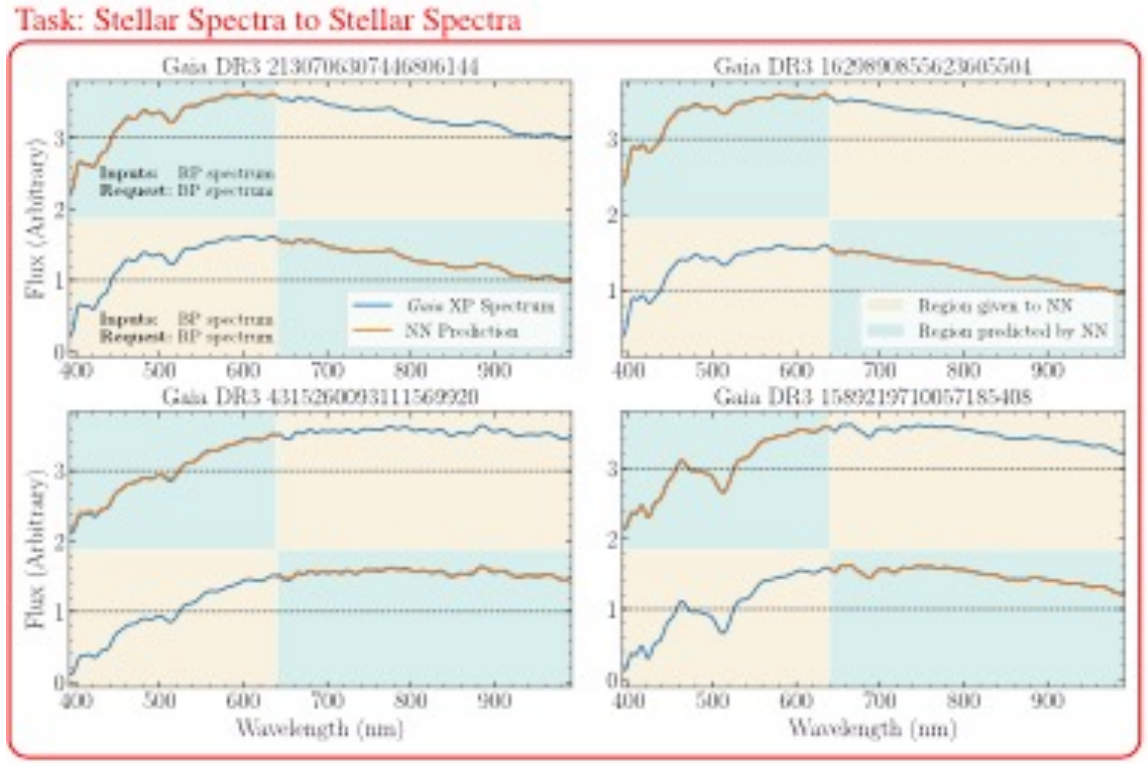


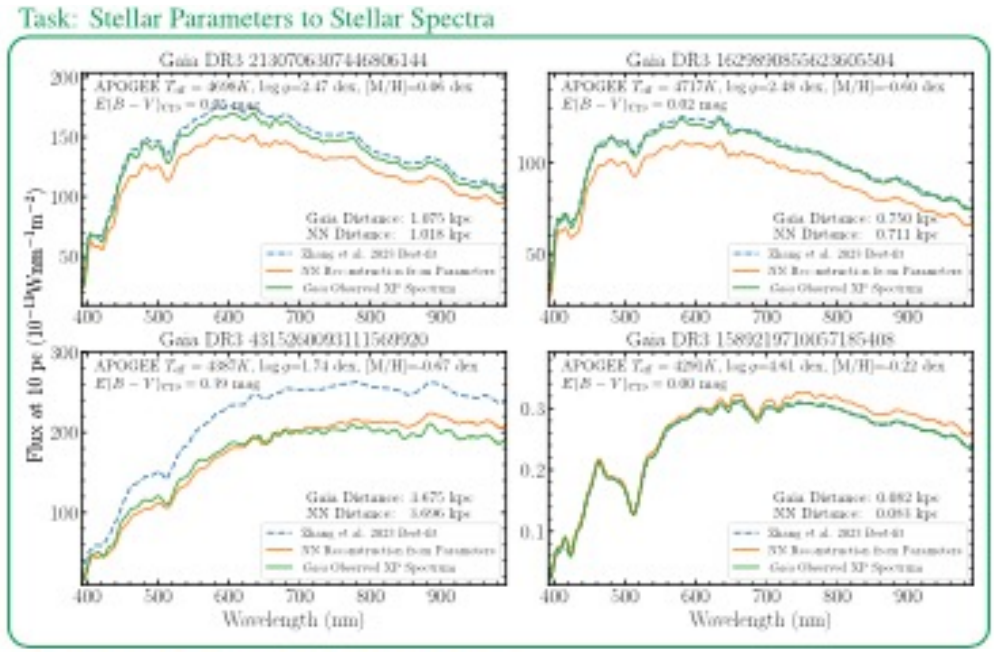


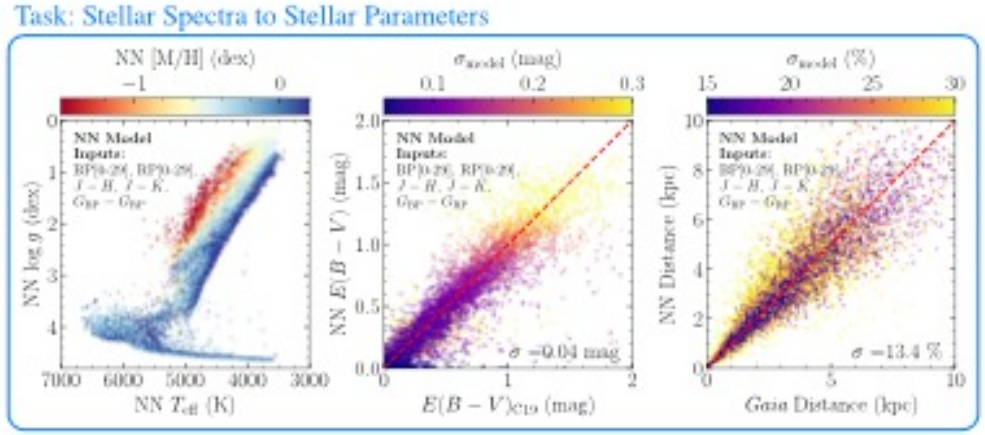
Leung H and Bovy J 2023

大样本恒星参数的神经网络回归

Transformer 大模型

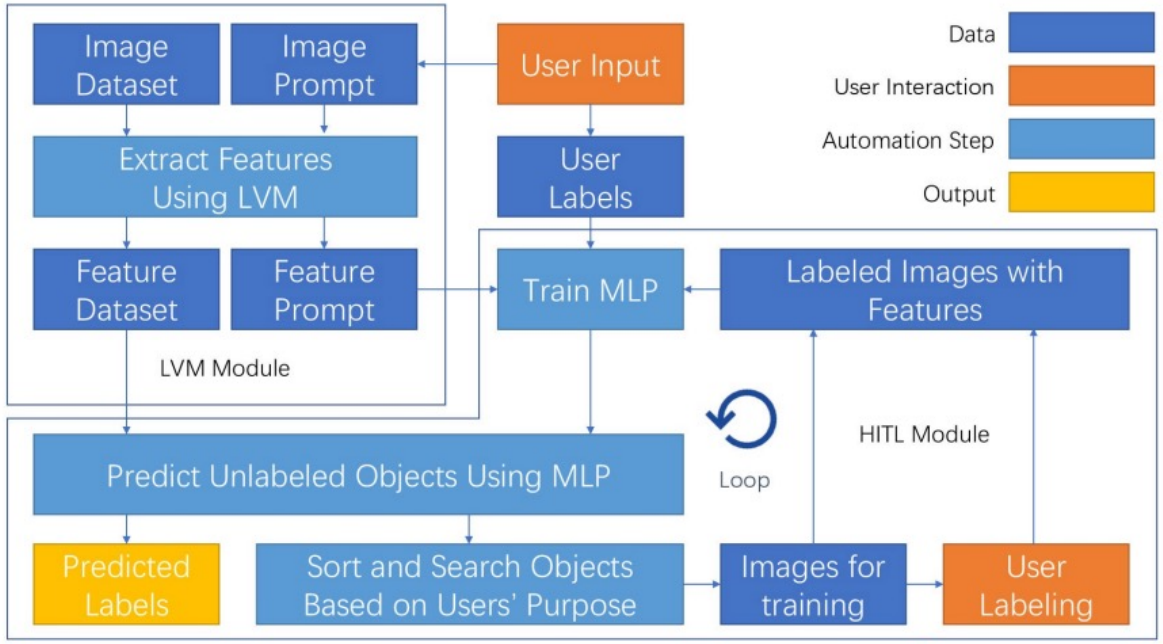


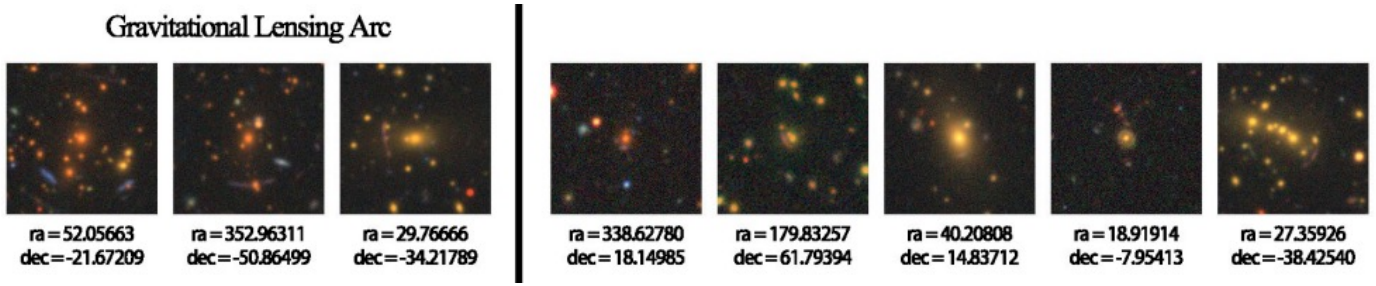




Leung H and Bovy J 2023

Transformer架构的特点也适合图像数据





只使用Transformer 的编码器

- **使用1亿张单个星系图像（从整个图像中剪切出来）来训练具有0.1B参数的变压器编码器。**

Chinese Physics C

PAPER

A versatile framework for analyzing galaxy image data by incorporating Human-in-the-loop in a large vision model*

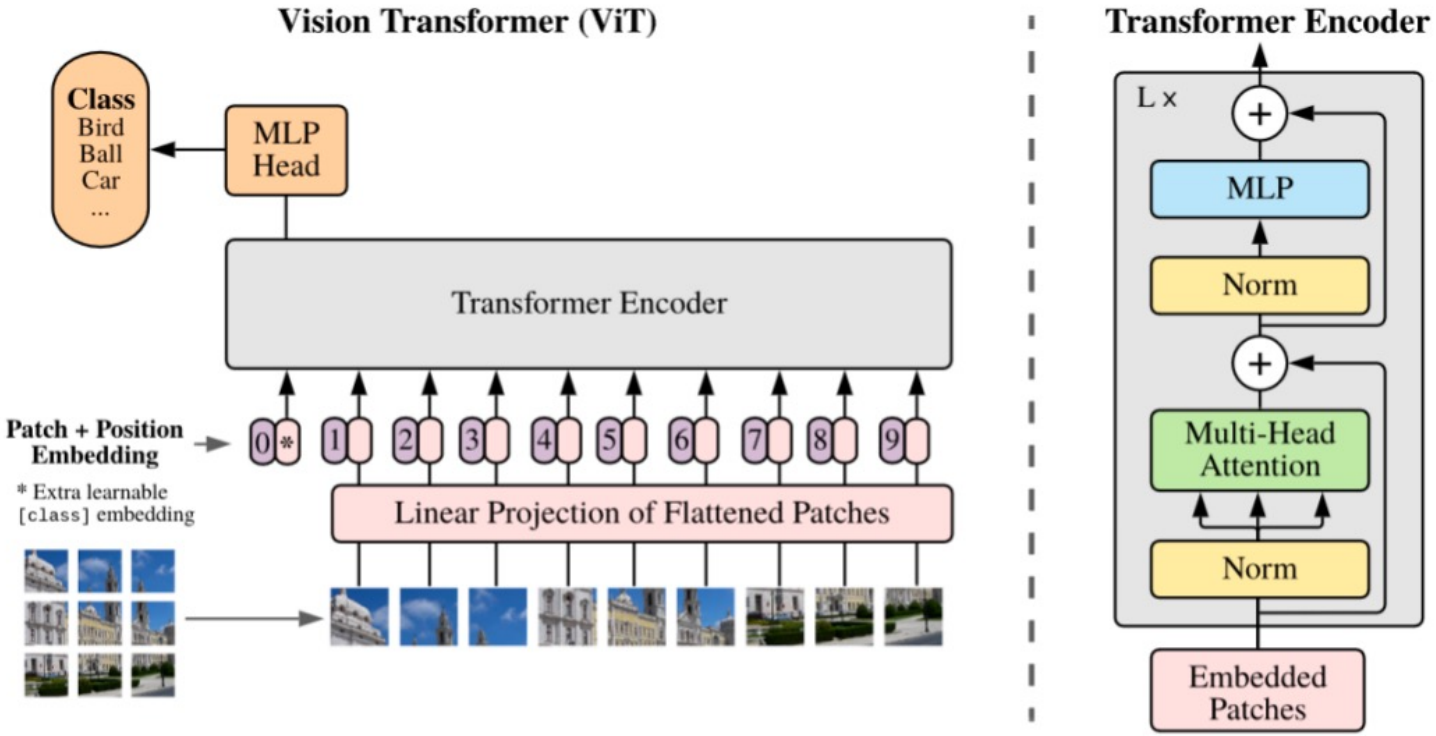
Ming-Xiang Fu (傅溟翔)[1,3,4], Yu Song (宋宇)[2], Jia-Meng Lv (吕佳蒙)[2], Liang Cao (曹亮)[2], Peng Jia (贾鹏)[2], Nan Li (李楠)[1,3,4], Xiang-Ru Li (李乡儒)[5], Ji-Feng Liu (刘继峰)[1,4], A-Li Luo (罗阿理)[3,6,7], Bo Qiu (邱波)[8] Show full author list

Chinese Physics C, Volume 48, Number 9

Citation Ming-Xiang Fu *et al* 2024 *Chinese Phys. C* **48** 095001

DOI 10.1088/1674-1137/ad50ab

基于Transformer的自监督工具ViT



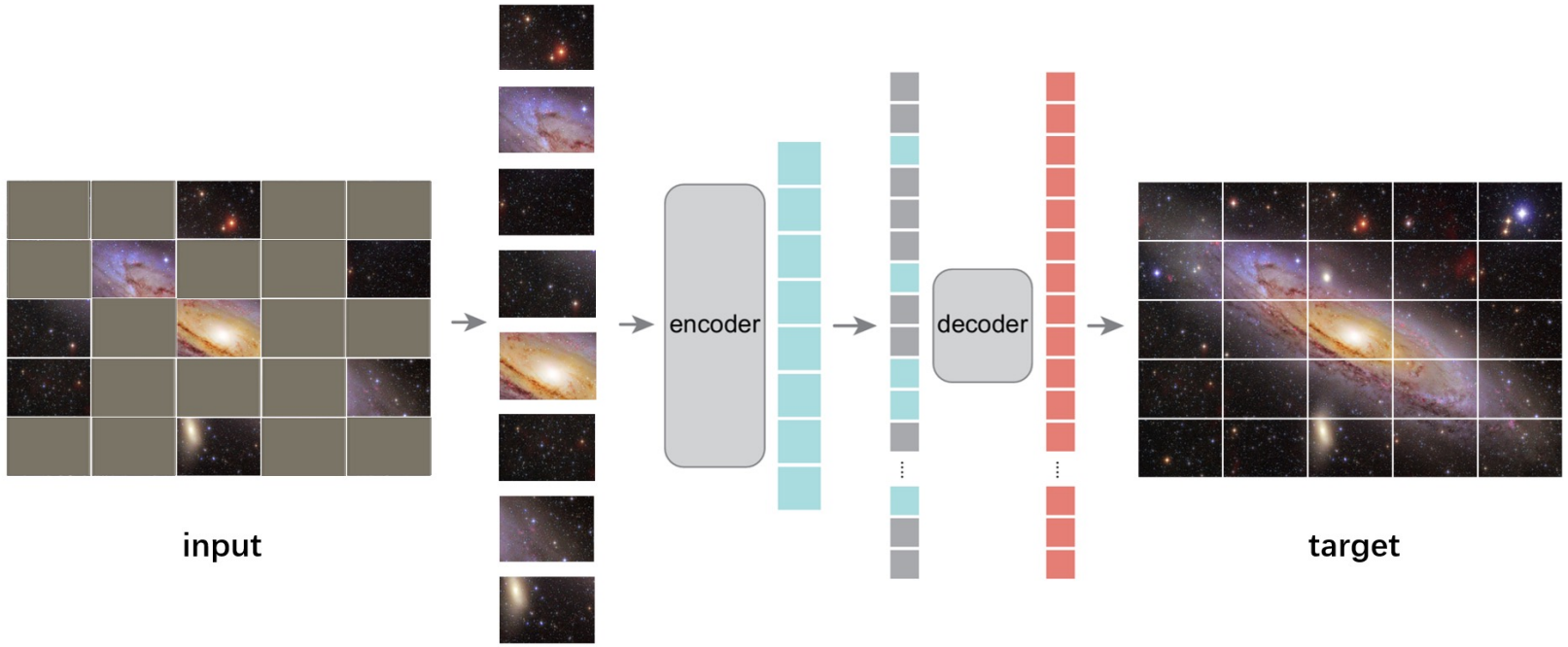
- ViT将输入图片平铺成2D的Patch序列（16x16），并通过线性投影层将Patch转化成固定长度的特征向量序列，对应自然语言处理中的词向量输入。
- 每个Patch可以有自己的位置序号，同样通过一个Embedding层对应到位置向量。
- Patch向量序列和图像位置向量相加作为Transfomer Encoder的模型输入。

ViT通过一个可训练的CLS token得到整个图片的表征，并接入全连接层进行下游分类。经过大量数据预训练，迁移到多个中等或小规模的图像库（ImageNet, CIFAR-100, VTAB 等）时，ViT取得了比CNN系的模型更好的结果，同时在训练时需要的计算资源大大减少。

自监督人工智能工具MAE

MAE： Masked Autoencoders

- 早期的ViT的预训练和CNN预训练一样，都是通过大规模的有监督分类任务数据集进行训练。
- 自监督预训练数据获取成本低、不需要标注、学习充分，比较有代表性的如Masked AutoEncoder（MAE）。
- MAE以ViT为基础模型，先对完整图片进行Patch掩码，接着使用一个Transformer Encoder对未Mask的Patch进行编码，然后通过相对小的Transformer Decoder模型还原被Masked Patch，从而实现模型的自监督预训练。



通过75%的高掩码率来对图像添加噪音，这样便很难通过周围的像素来对被掩码的像素进行重建，从而使编码器去学习图像中的语义信息。预训练之后，解码器被丢弃，编码器可以应用于未掩码的图像来进行识别任务。

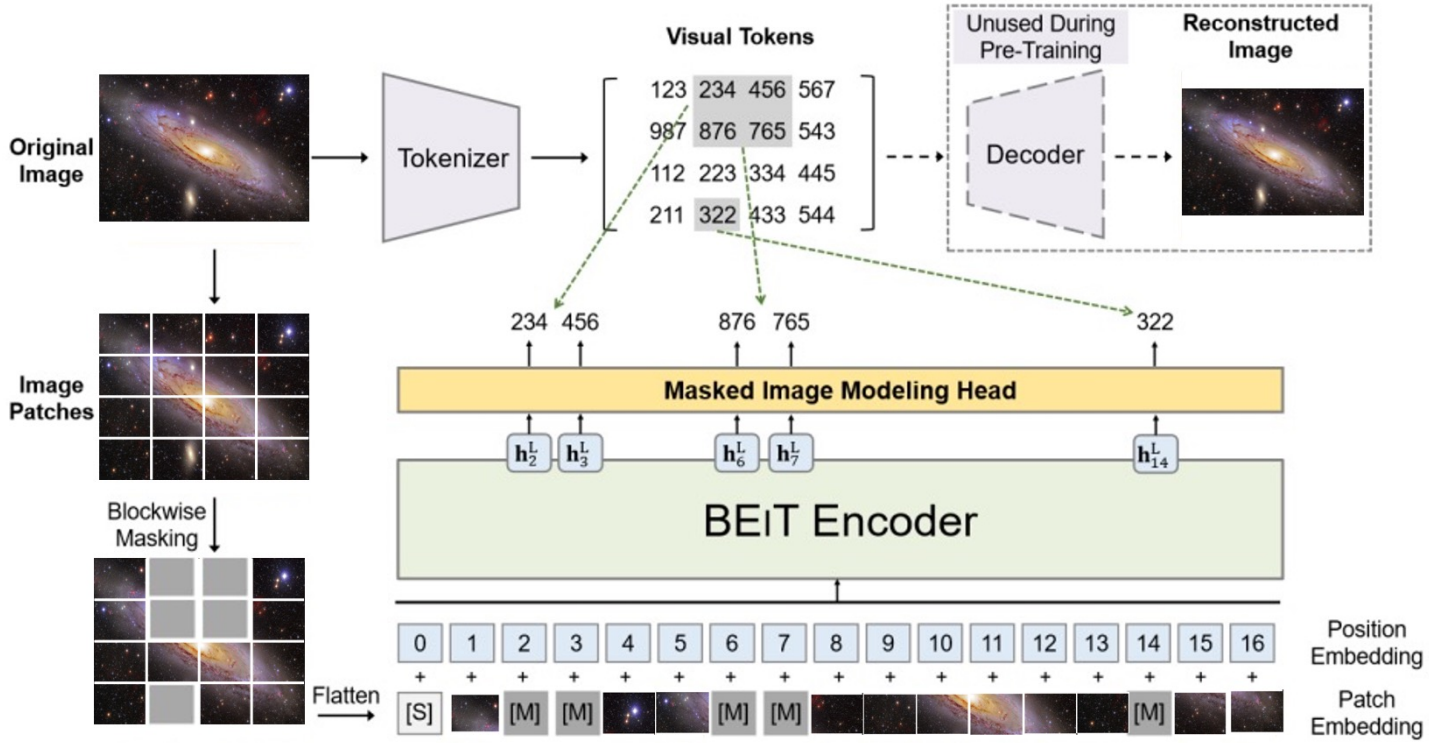
自然语言和图像特征的信息密度不同，文本数据是经过人类高度抽象之后的一种信号，信息是密集的，可以仅仅预测文本中的少量被掩码掉的单词就能很好的捕捉文本的语义特征。而图像数据是一个信息密度非常小的矩阵，包含着大量的冗余信息，恢复被掩码的像素并不需要太多的语义信息。

自监督人工智能工具BEIT

BERT Pre-Training of Image Transformers

与BERT的预训练框架对齐：

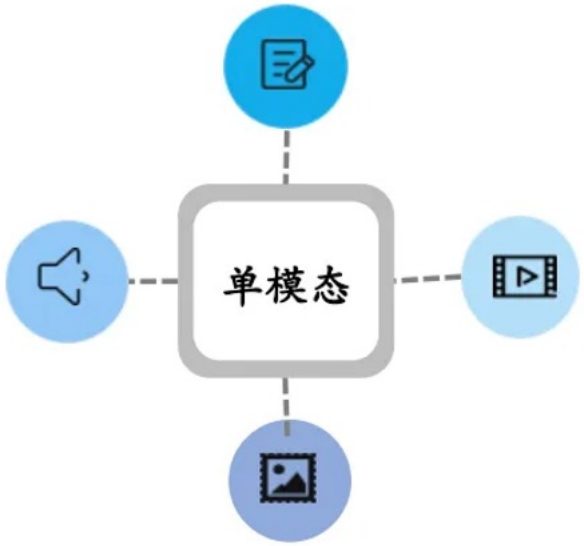
- **BEIT通过辅助网络模块先对视觉Patch进行向量化，得到整张图各部分的视觉Token ID。**
- **然后将视觉Patch视为自然语言中的单词进行掩码预测，**完成预训练流程。



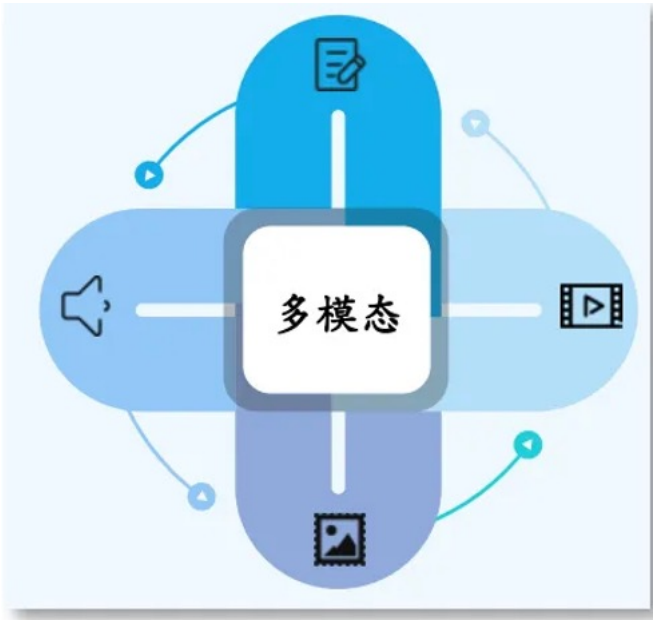
预训练的目标是基于被掩码的图像输入向量序列，预测源图像对应的视觉Token ID。

- 在预训练之前，BEIT先通过一个离散自回归编码器（dVAE）学习了一个“图像分词”器，可以将图像编码成离散的视觉Token集合。
- 在预训练阶段，输入的图片存在两个视角，一是图像Patch，另一个是视觉Token。BEIT随机对Patch进行掩码，并将掩码部分替换为特殊的Mask Embedding（[M]，图中的灰色部分），随后将掩码后的Patch序列输入到ViT结构的模型中。

从单模态学习到多模态学习

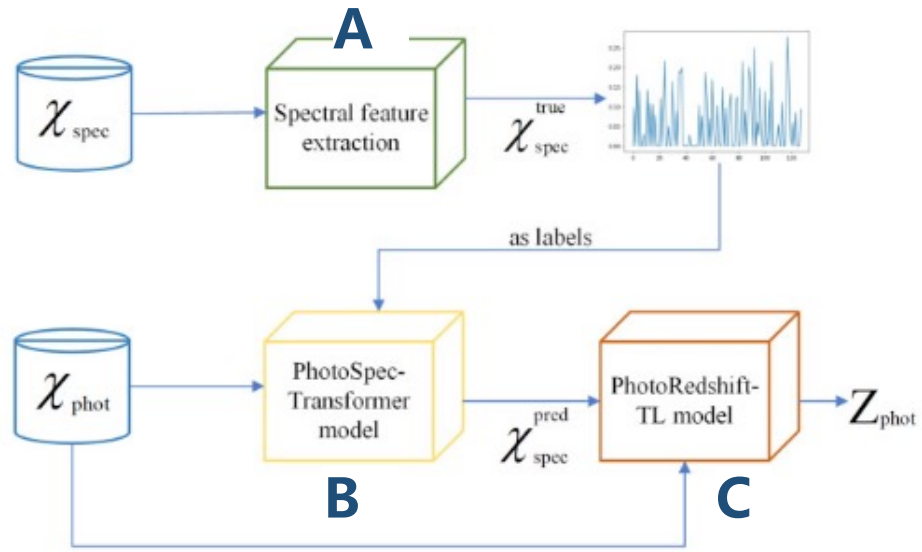


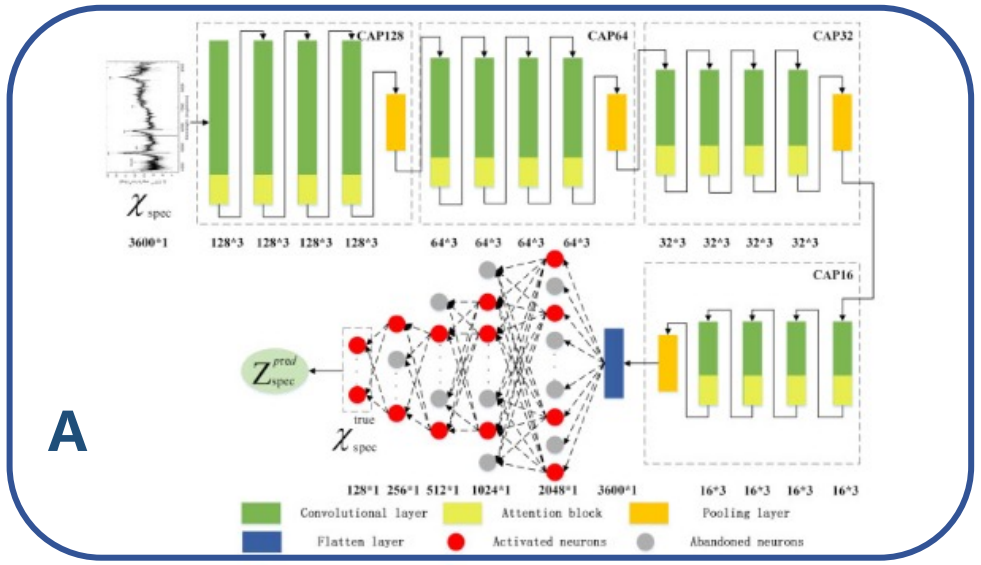
算法只能基于一种模态学习并
只能应用于该模态

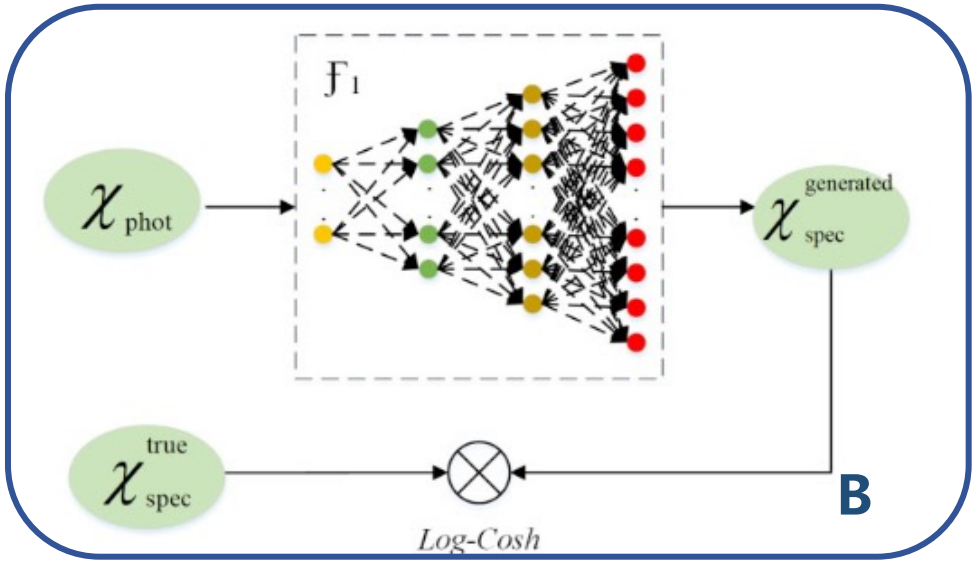


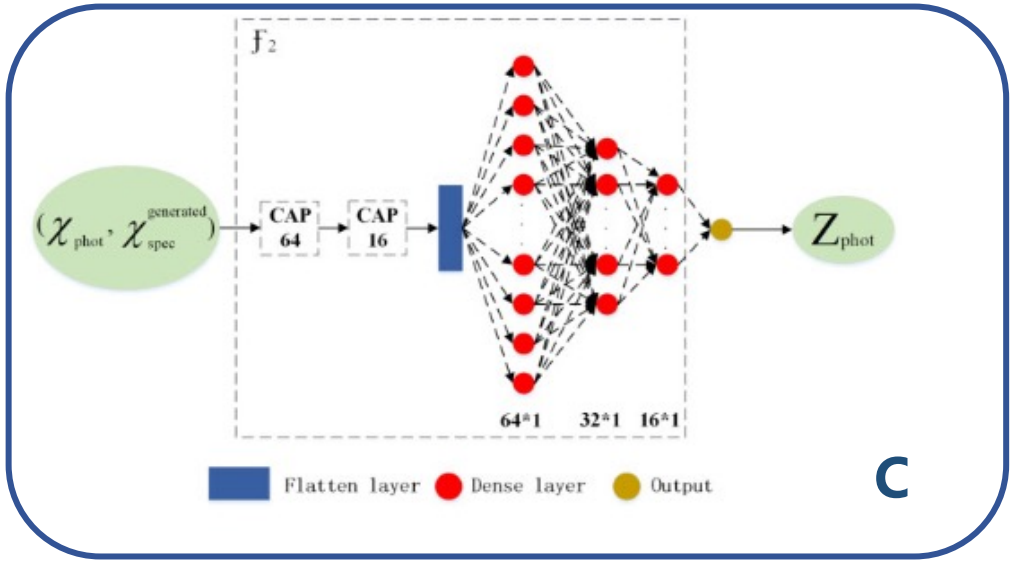
可以学习并应用于多种模态

多模态模型的尝试（类星体红移测量）









Hong et al 2023

多模态模型的尝试（类星体红移测量）

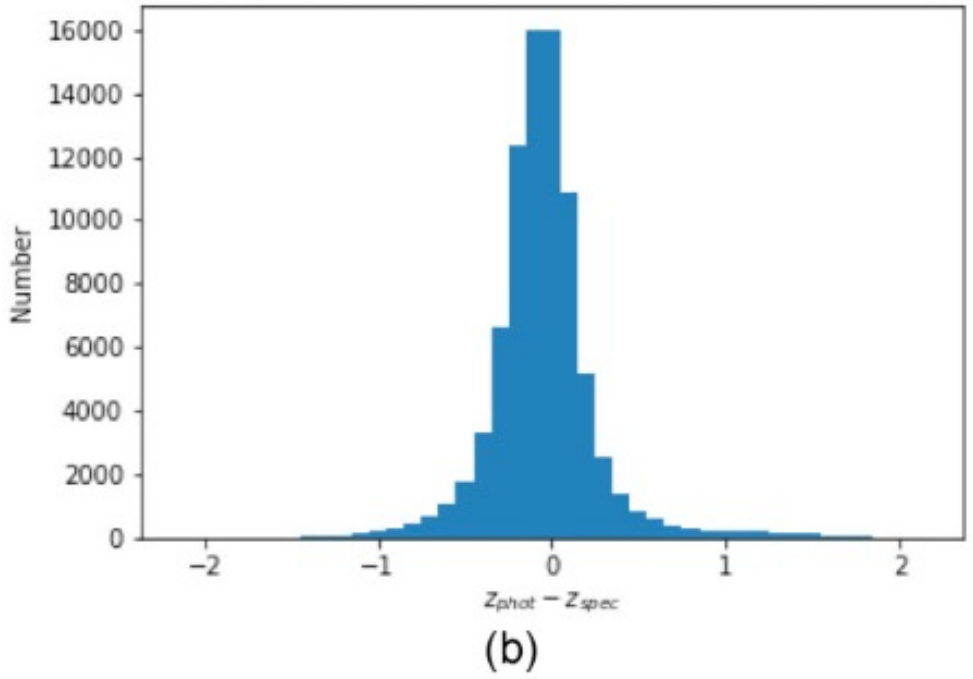


Figure 7. Prediction results of photometric redshift based on ANN.

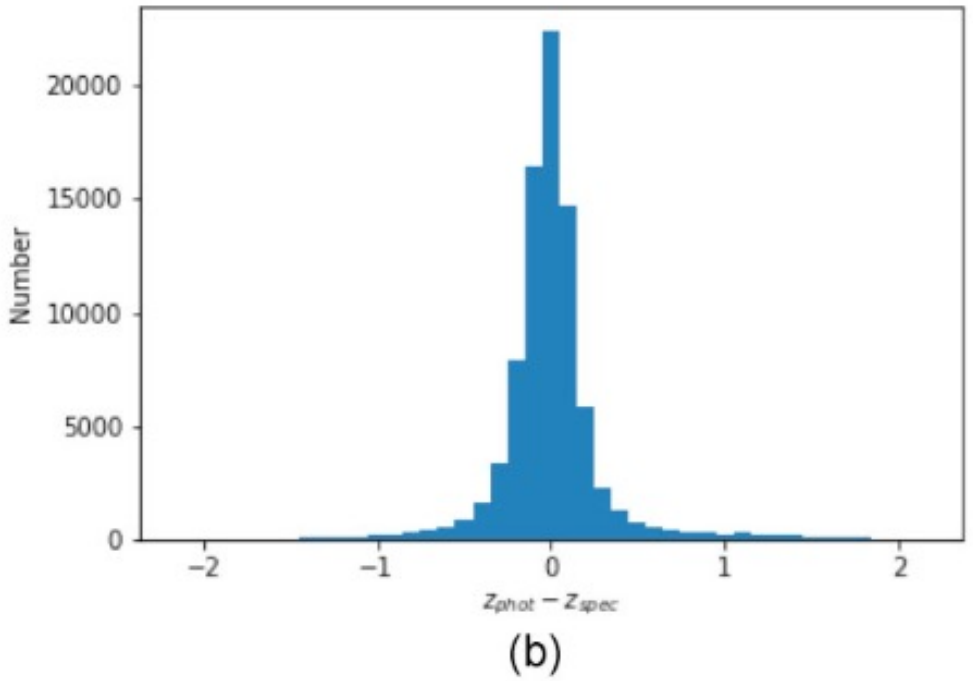
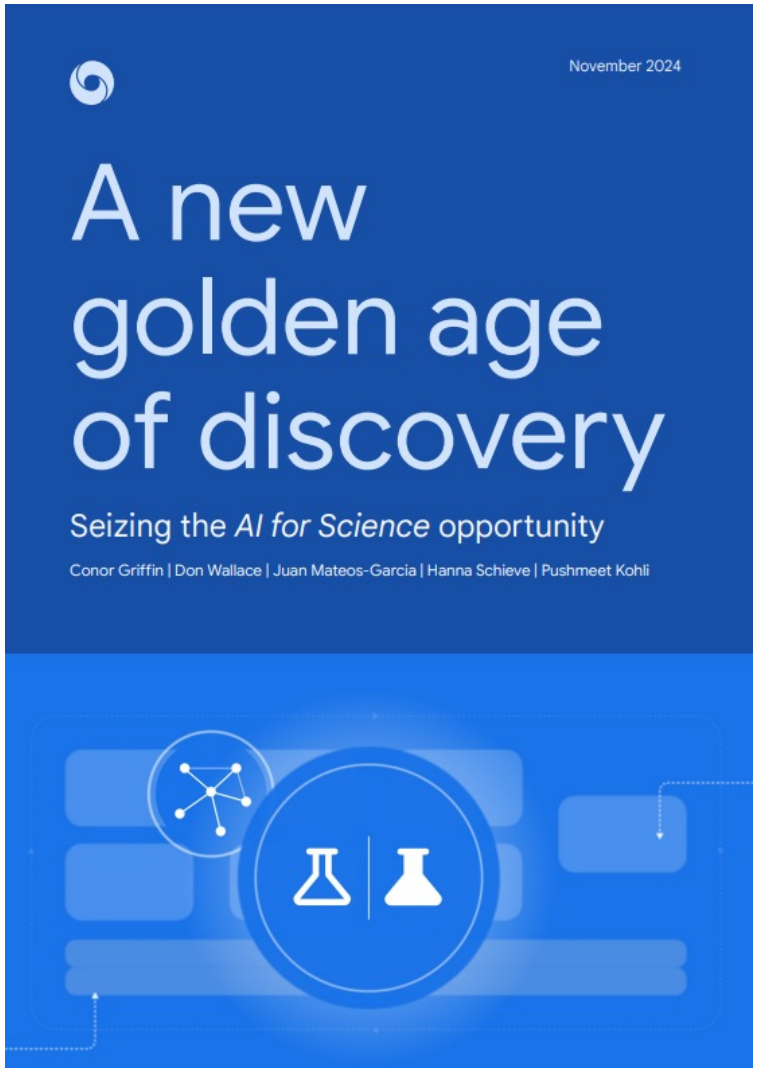


Figure 8. Prediction results of photometric redshift based on PhotoRedshift-MML.

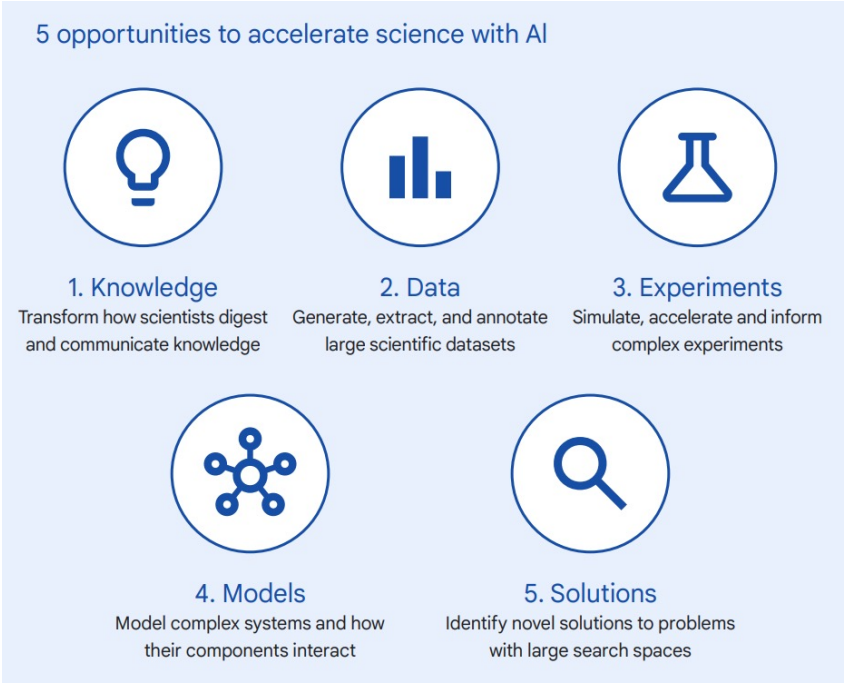
Hong et al 2023

人工智能大模型与科学



https://storage.googleapis.com/deepmind-media/DeepMind.com/Assets/Docs/a-new-golden-age-of-discovery_nov-2024.pdf

AI加速科学创新发现的黄金时代



- **知识**——改变科学家获取和传递知识的方式
- **数据**——生成、提取和标注大型科学数据集
- **实验**——模拟、加速并指导复杂实验
- **模型**——建模复杂系统及其组件之间的相互作用
- **解决方案**——为大规模搜索空间问题提出创新解决方案

AI for Science

“AI将颠覆基础科学研究，人类科学发展或迎来新的大航海时代。”

- 人工智能对科学研究的影响：AI 在数据处理和模式识别方面超越人类认知极限的能力，它将帮助研究人员探索新的科学维度和假设。
- 跨学科融合：AI 将打破学科边界，促进不同领域之间的融合，为解决复杂问题提供新的视角。

——微软首席科学家Eric Horvitz

“今天我们一定要做的一件事情是AI For Science。讲得稍微夸张一点，难以想象今天还有什么事情比AI For Science更重要。”

——沈向洋

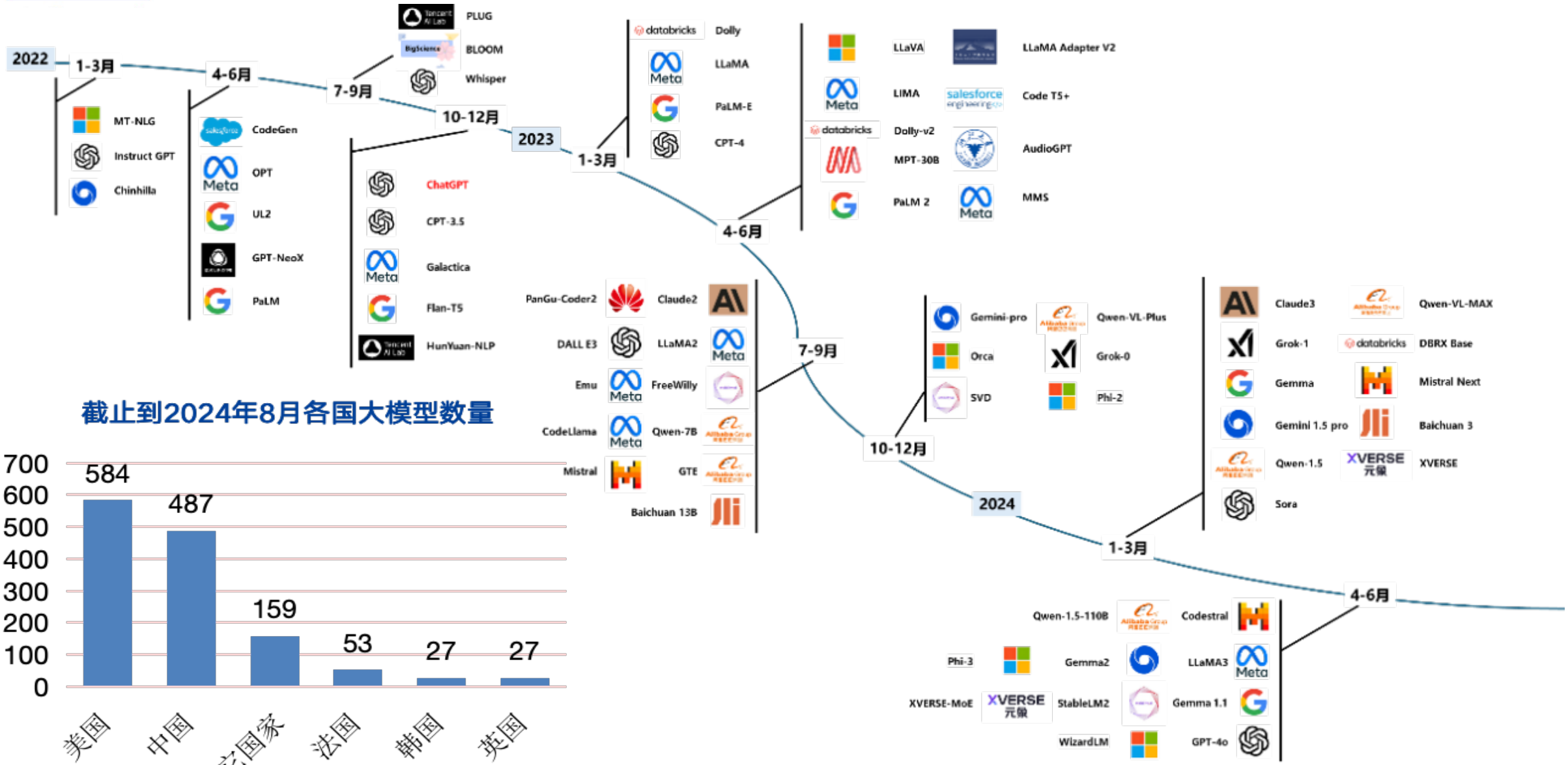
“未来五年，AI关注的重点将转向硬核领域——用AI加速科学和工程。因为这才是真正推动技术进步的引擎。”

——OpenAI科学家Jason Wei

“美国DOE建议重点打造多个大型基座模型”

——《美国人工智能驱动的科学研究新进展和部署动向》

主流的大模型产品大都是“文科生”模型



截止到2024年8月各国大模型数量

国家天文台与之江实验室的合作打造科学模型

选派40名科研骨干和学生赴之江实验室，共同训练人工智能大模型。从“种子班”起步，双方共同设计完成了系列大模型的首个——天文大语言模型AstroOne。为打造天文学领域型奠定了基础。

国家天文台人工智能的发展历程

90年代后期

神经网络热潮下**LAMOST**与中科院自动化所合作，开始了**海量光谱分析**方法研究。获863计划资助。

本世纪初开始

国台多个团组开展了机器（深度）的应用研究，在数据驱动的天文研究领域越来越多地获得**国家自然科学基金资助**

2010年开始

国台牵头与数所高校计算机院系合作举办“**海量天体数据挖掘研讨会**”年会，2023年更名为“AIDA年会”获国自然面上基金资助。

2011年

LAMOST巡天开始后，开展了大量的机器学习应用。自此，在天文顶刊发表大批机器学习为方法的**发现类**论文

2016年

国家天文台与**阿里云**战略合作挖掘天文大数据。**NADC**开启与阿里云的数据合作。同期，**FAST**开始了数据驱动的天文发现

2019年开始

CSST开始进行了AI在原始数据处理和海量巡天数据分析中的应用的探索

2020年开始

组织了“黑客松天文数据挖掘训练营”，2023年更名为“A^3训练营”。获联合重点基金资助。

2021年

司天工程成立人工智能小组，规划AI在司天巡天中的I应用

2022年底

在大语言模型背景下国台研发了“**星语**”模型

2023年

在大模型出现的背景下，成立人工智能工作组

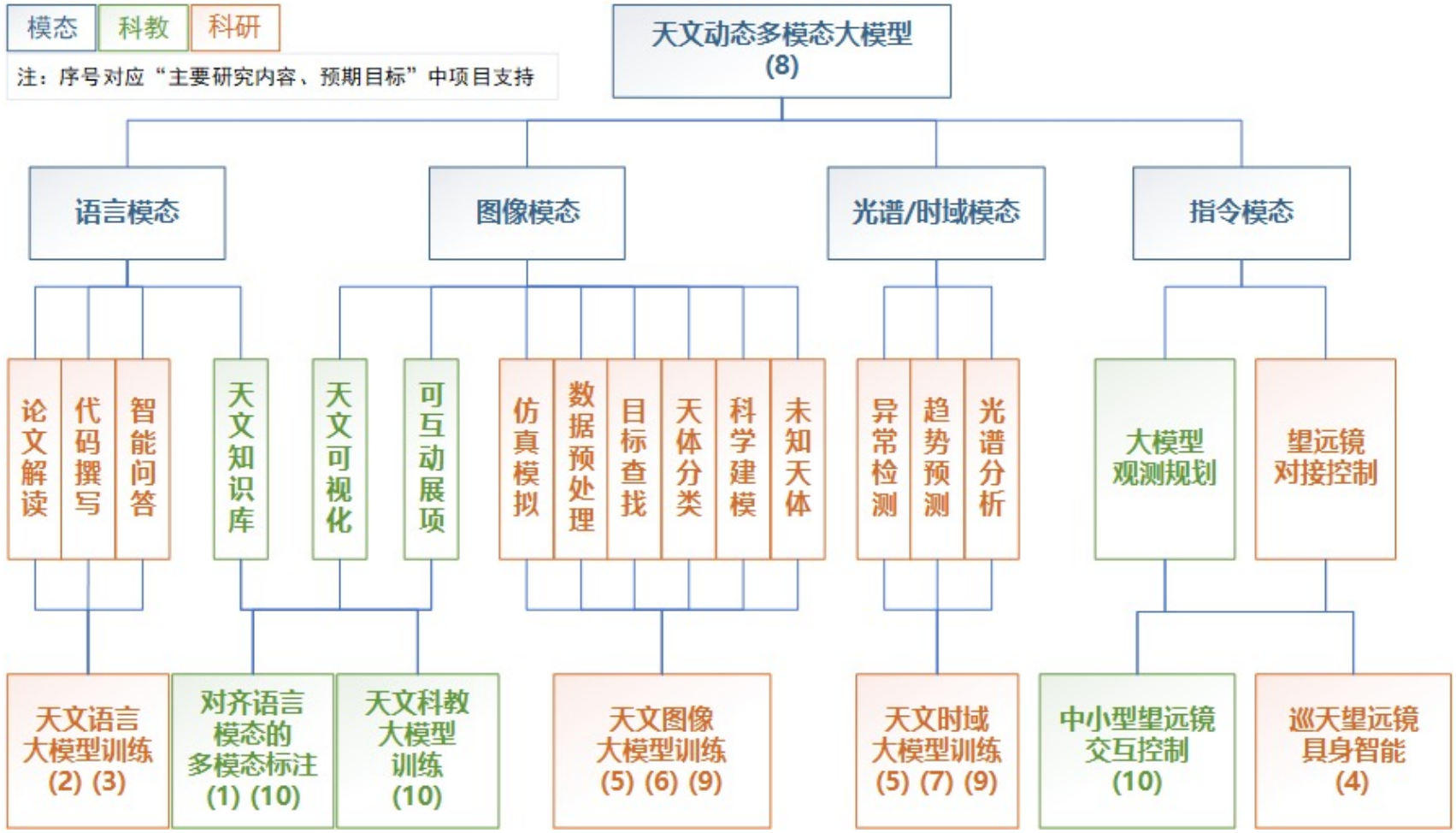
2024年

国台与**之江实验室**开展战略合作，训练天文垂直领域大模型，第一个模型是“天一”。

大数据+机器（深度）学习 → **多模态数据+大模型**

国家天文台AI工作组的目标

构建AI天文生态



主要任务

收集需求
发布指南
建立平台
制定标准
统筹资源

工作愿景

协调多方力量，建成天文人工智能生态，借力科研范式变革，实现中国天文对国际顶尖水平的加速追赶和超越!

打造科学模型的路径

1、科学语言模型

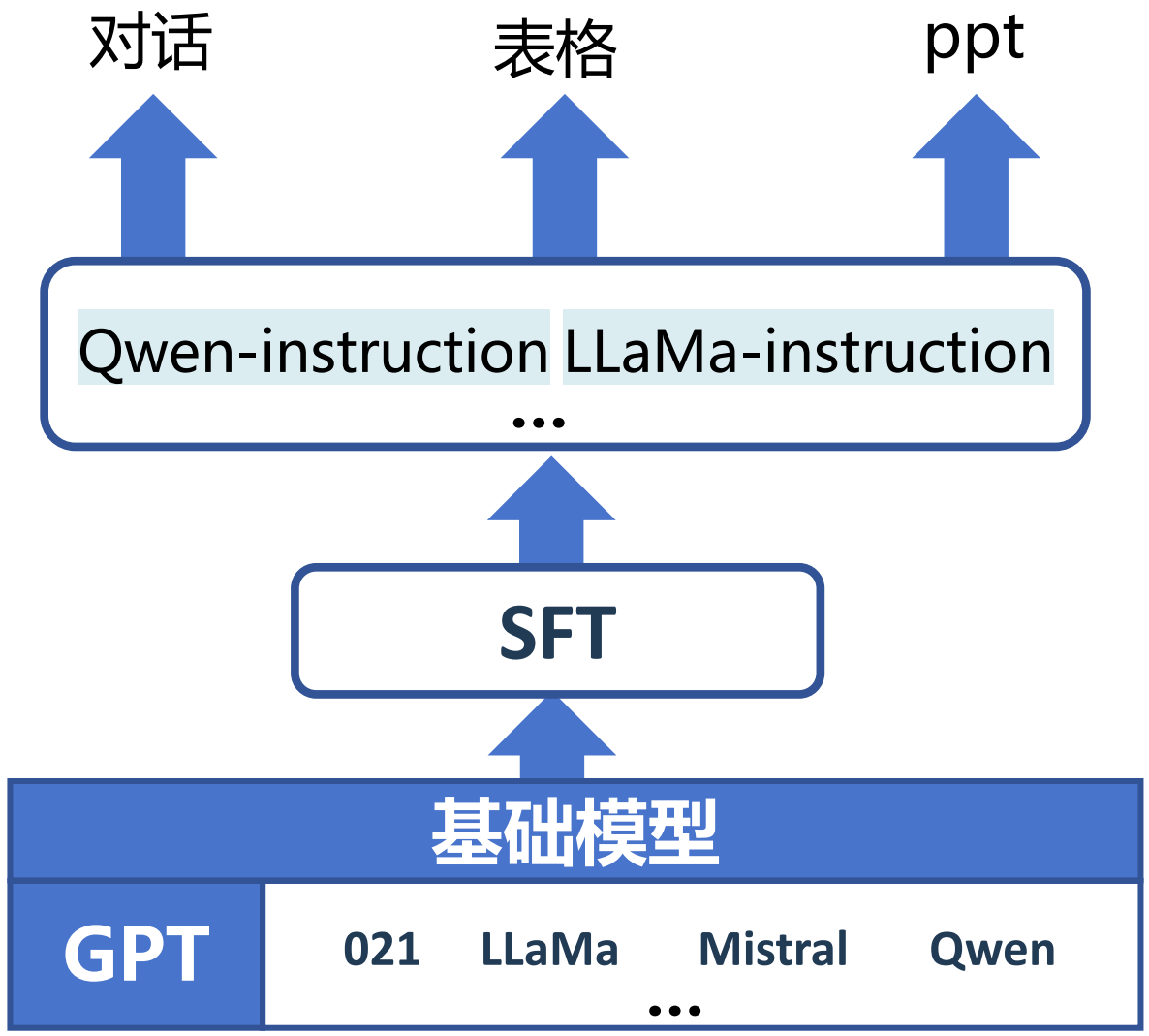
2、科学计算模型

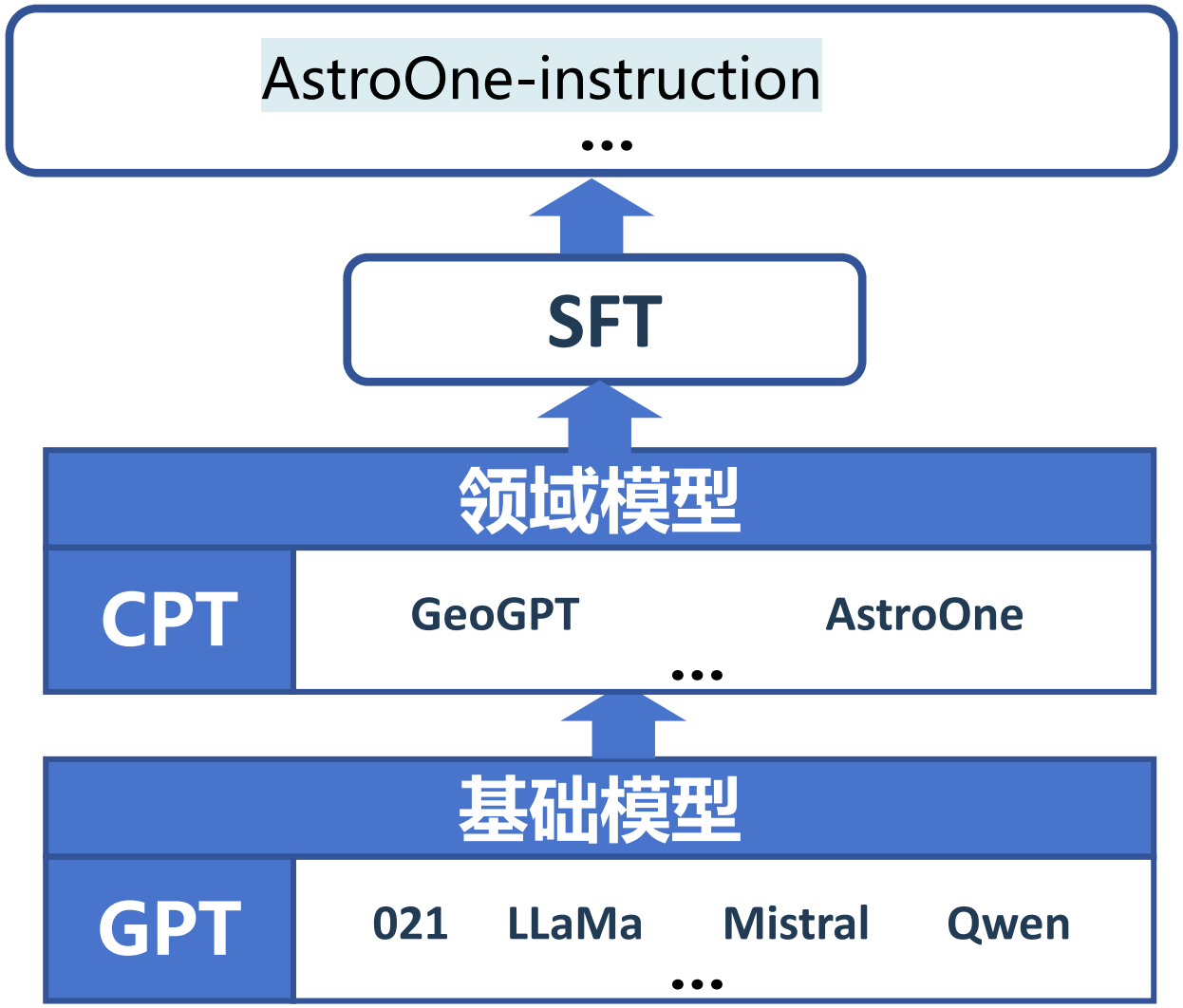
ChatGPT=GPT+Chat

Foundation model

Instruction Followling

科学语言模型

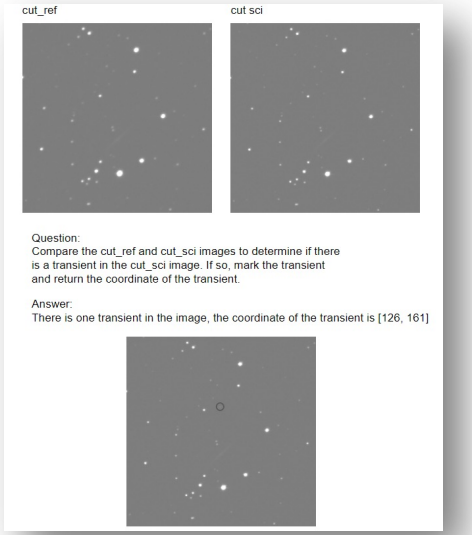

对话
表格
ppt
Qwen-instruction LLaMa-instruction
…
SFT
基础模型
GPT
021 LLaMa Mistral Qwen
…


AstroOne-instruction
…
SFT
领域模型
CPT
GeoGPT AstroOne
…
基础模型
GPT
021 LLaMa Mistral Qwen
…

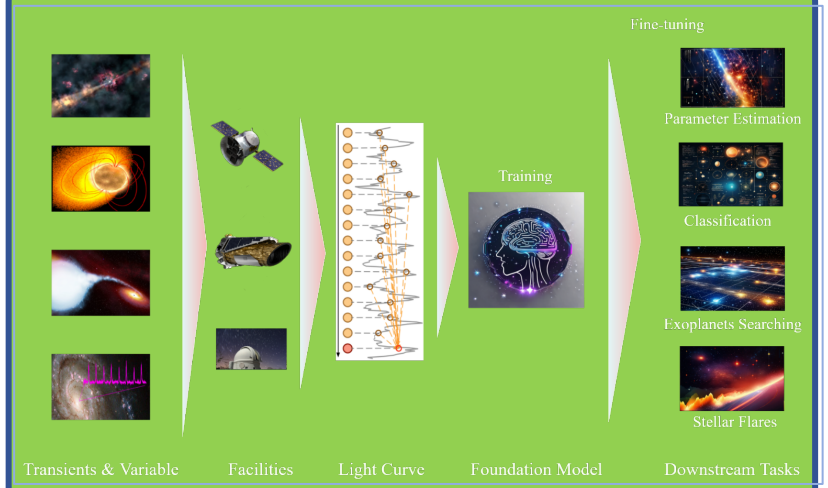
天文人工智能大模型的规划：1+1+X

天文学家常问问题收录；
科学思维链和提问范式；
快速检索已有研究成果，精确推荐；
天体复杂演化规律推理；
科学假设生成与启发框架；

知识和推理
语言模型



分析文献中数据
图文（多模态）模型



处理观测数据
X模型

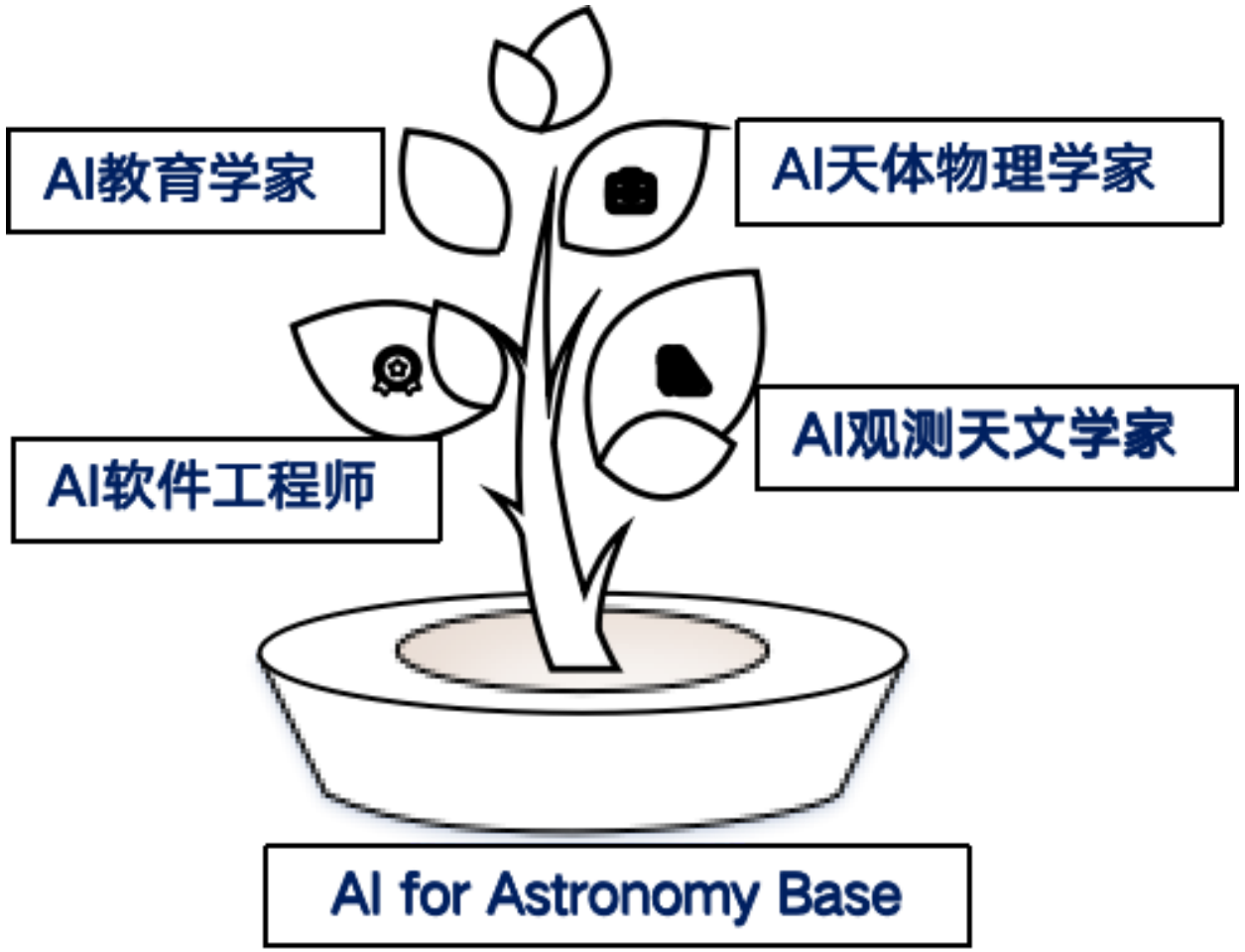
高可靠的
科学假设生成

强逻辑严谨推理的
科学分析

多模态多尺度的
科学规律揭示

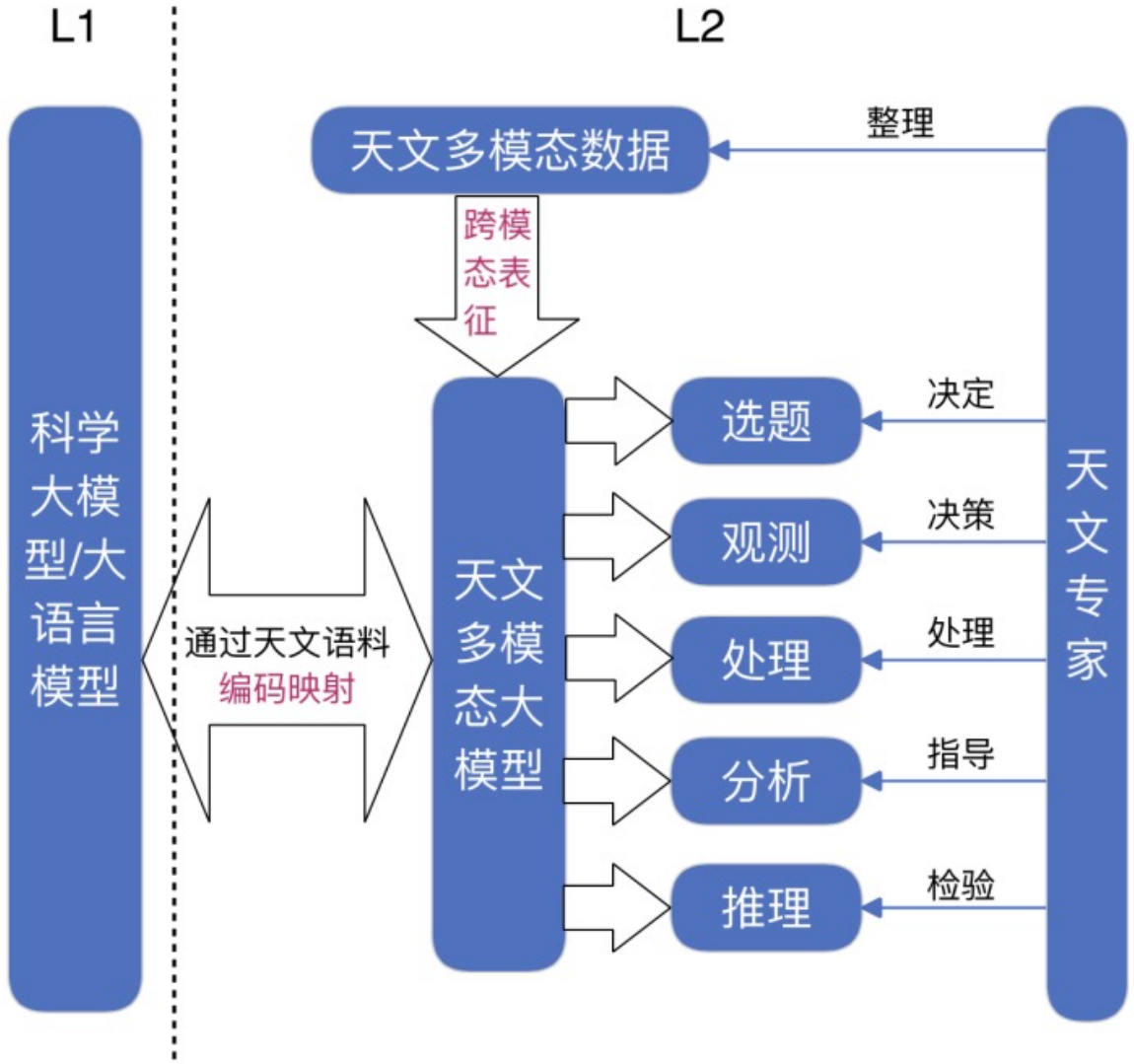
科学基础模型
评价体系

天文人工智能大模型的应用场景



- 大模型是针对**海量天文数据**的观测和处理科学研究的**人员严重不足、专用软件难以复用**等问题。
- 利用天文数据、知识库训练的**人工智能大模型**将天文学者从繁重的重复性劳动中解脱出来。
- 推动实现和部署基于大模型的、以智能体为内核的面向观测规划与协同、处理管线和代码生成、跨模态推理和科学假设生成的**全链条科研工作流。**

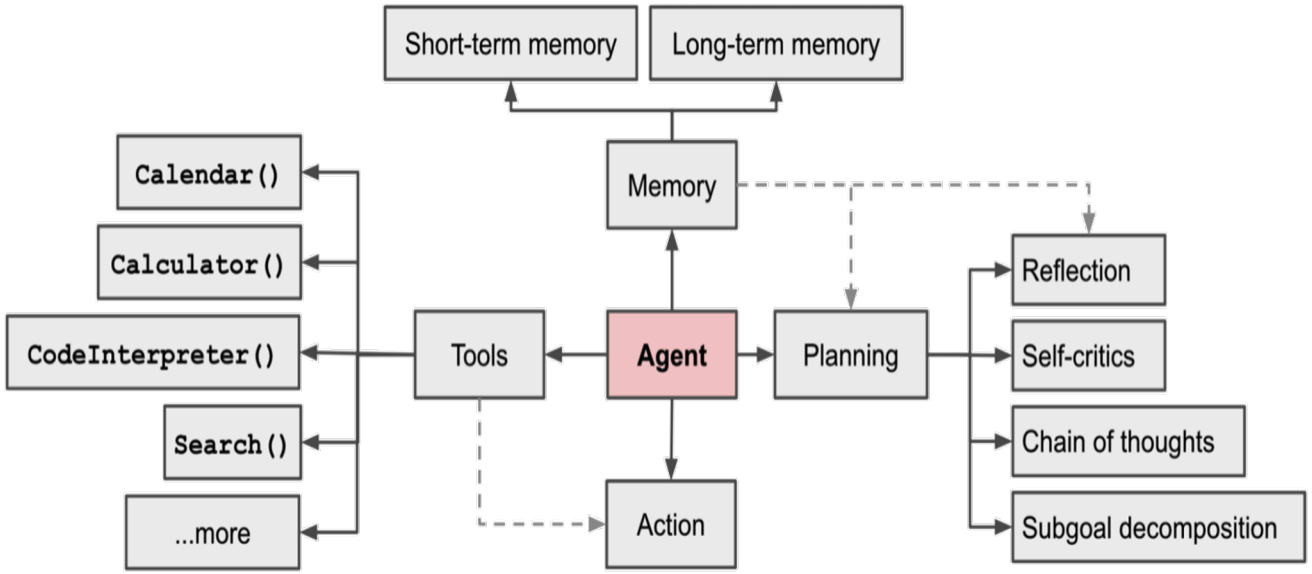
国家天文台的人工智能模型群



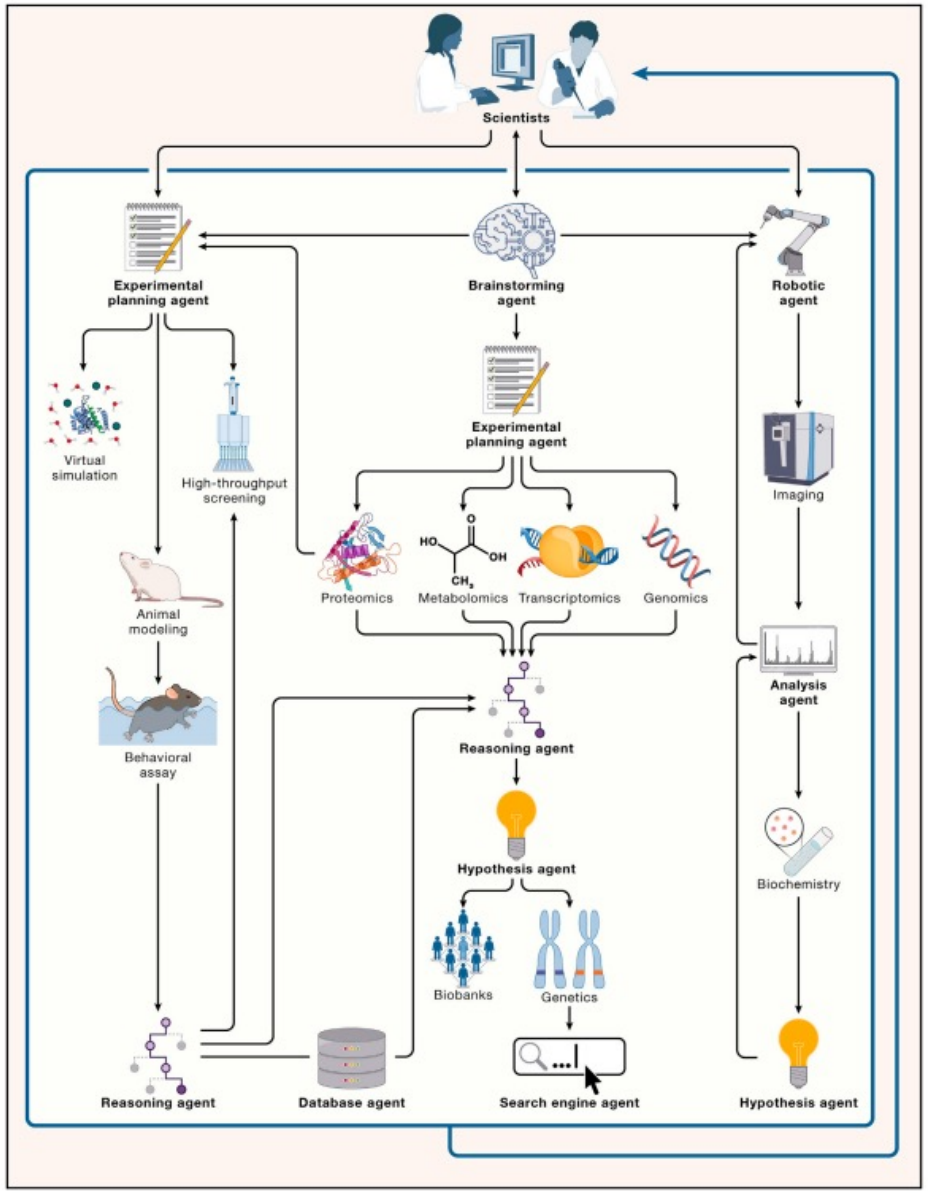
- 天文大语言模型：AstroOne
- 天文图文模型：AstroOne-VL
- 恒星光谱模型: AstroOne-SpecCLIP
- 光学时域模型： AstroOne-FALCO
- 测光模型：AstroOne-Photometry

依托1+1+X建设天文生态

基于Agent的AI科学家

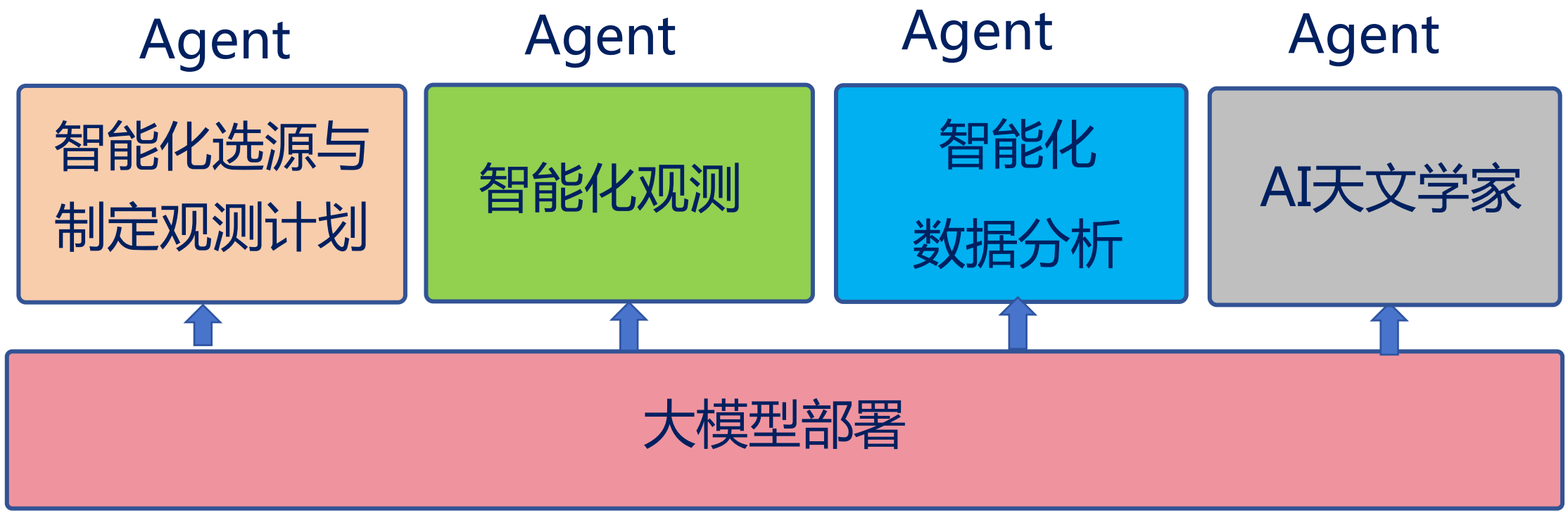


规划、记忆、工具、行动



Agent：从观测到数据分析到科学

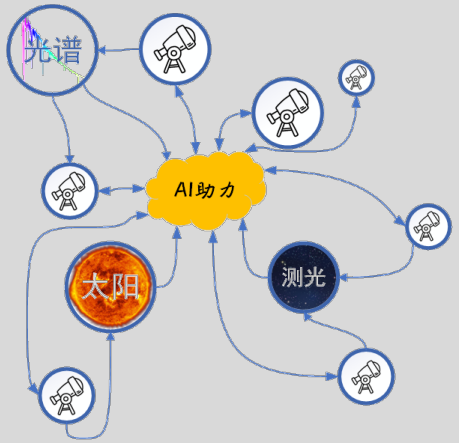
Agent自动化——帮助自主交付AI的技术，结合AI和自动化来动态感知环境、分析数据、引导不断变化的工作流程、自主作出决策。



Agent：多望远镜多数据源的全链条

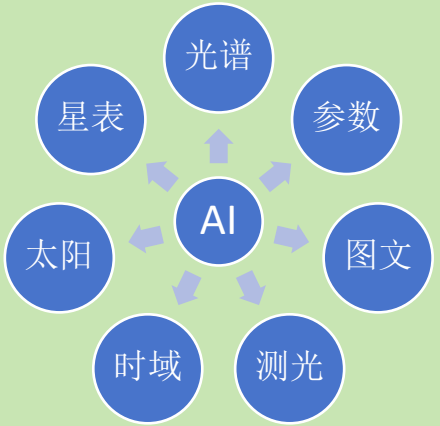
解决国家在天文领域的投入和产出之间的矛盾：
——人力匮乏、处理效率低、资源分布不均衡

智能化协同观测



➢**观测模式：**单望远镜观测难以同时获得高价值高时效性天文现象，亟需多观测设备的精准、及时联动获取数据，“老站”带“新站” ，东部带西部，**解决分布不均衡的问题**

智能化融合处理



多模态数据

➢**数据挖掘：**大装置获取的多波段单模态数据难以高效地提取天体目标的全面信息，成果产生效率较低，容易错失高价值天文现象，亟需跨模态数据集成和高效融合，**解决处理效率低下的问题**

➢**研究范式：**学科领域分工过细，存在知识壁垒，需要探索能够解决重大科学问题的跨领域研究路径，利用跨领域知识预测与重大科学问题相关联的天文现象，**解决我国天文人才匮乏的问题**

国台AI的发展愿景

□ 高水平持续增长语料库

- 天文高质量文本数据超500亿 tokens
- 超1000万对图文数据
- 天文科学数据超10PB

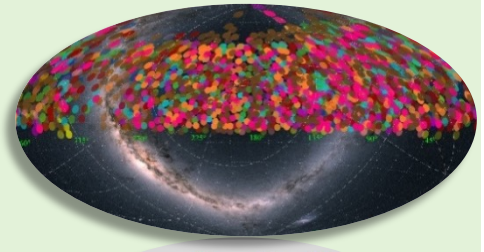
□ 通过对抗优化，促进基座模型发展

星语 VS 天一

HISPER

ASTRO ONE

□ 借助AI攻克天文技术难题，助力研究宇宙基本问题



➢ 天文事件产生的海量多模态数据及时处理，天文事件即时预警

➢ 打造具身智能望远镜，异地多设备联动，智能组网

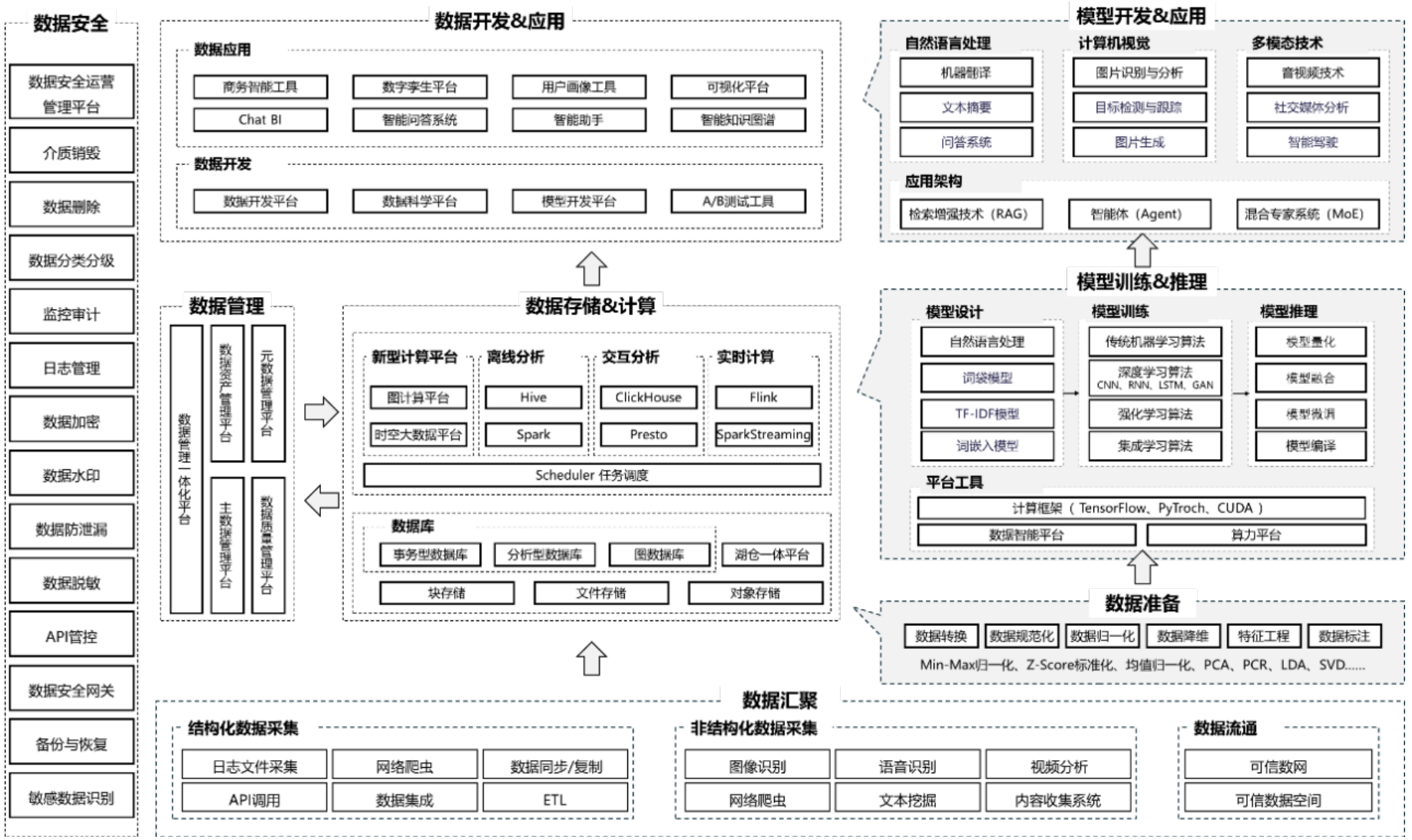
□ 持续发展天文AI模型群

具身智能望远镜

高度自动化的同时，通过AI进行观测决策，调度工具执行观测任务，减少大巡天阵列中人力消耗，提高观测效率。

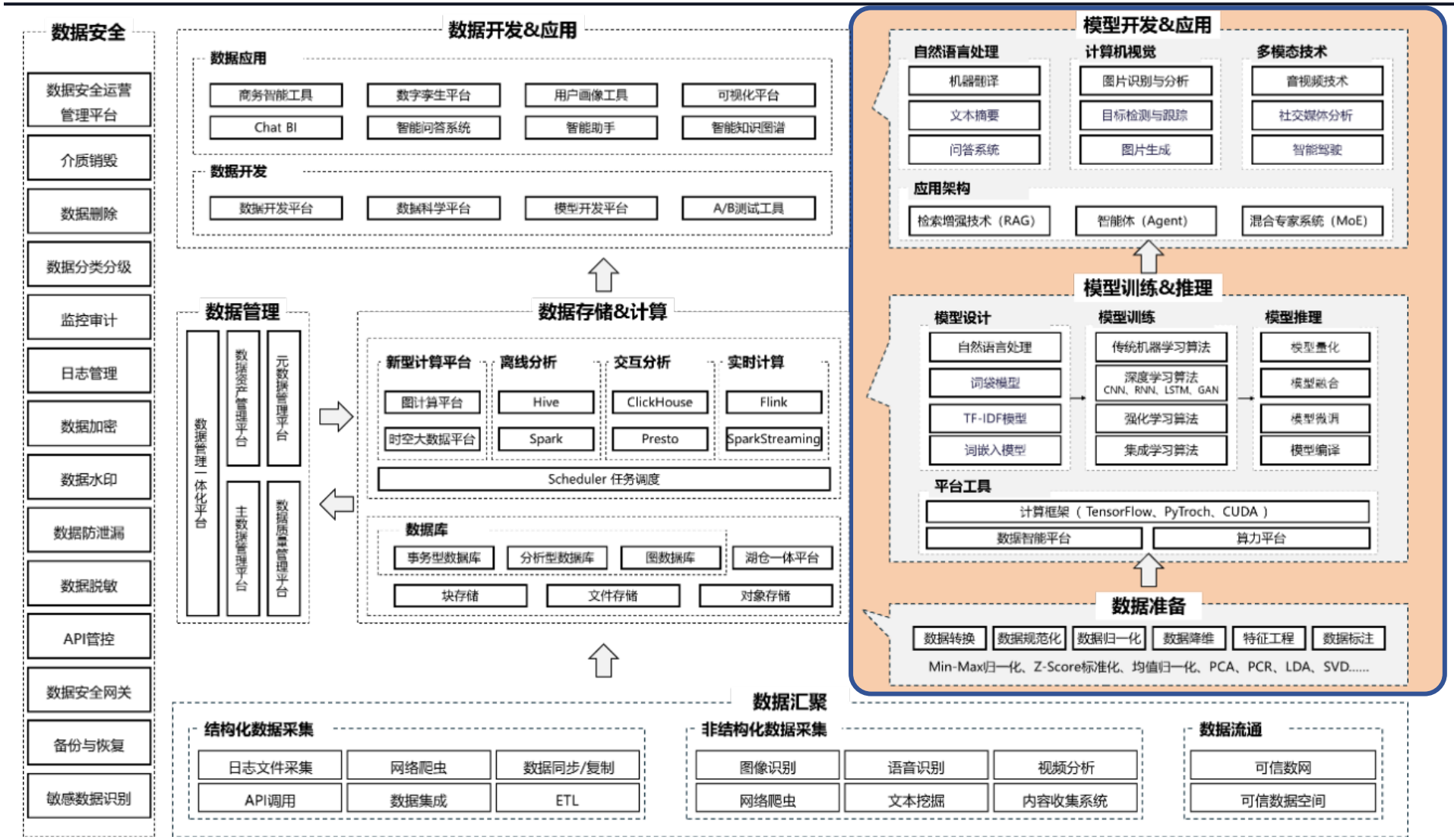
□ 推动人人用AI，促成天文研究新范式落地

天文数据+智能的标准体系



来自大数据技术标准推进委员会（中国通信标准化协会(CCSA)内设组织）

天文数据+智能的标准体系



来自大数据技术标准推进委员会（中国通信标准化协会(CCSA)内设组织）

总结

天文研究

从数据中**分析特性**，从特性中**探索规律**，从规律中凝练**理论和模型**，解释天文现象

为深度网络如何学习和计算提供**领域知识支撑**

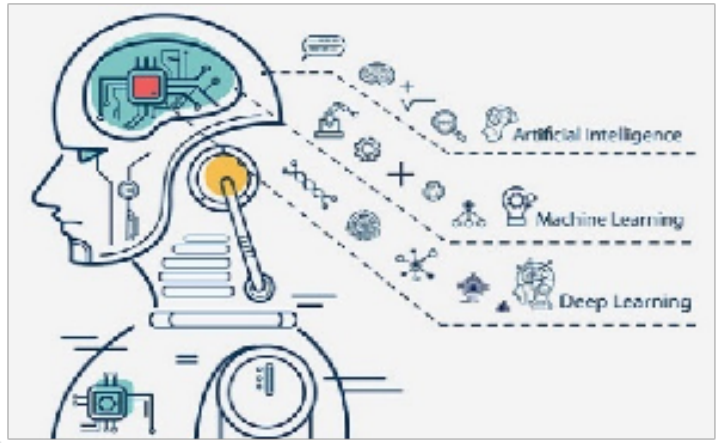
Science for AI

AI for Science

为天文学解决复杂问题提供**工具和方法**

人工智能

从数据中**提取特征**，从特征中学习**拟合**规律构建智能体模型，**预测和生成**目标任务内容



Q&A

欢迎批评指正

谢谢！