之江实验室 ZHEJIANG LAB
東南大學 SOUTHEAST UNIVERSITY
NAOC 中国科学院国家天文台 NATIONAL ASTRONOMICAL OBSERVATORIES, CAS

打造AI时代的天文数据知识关联搜索引擎

2025.11.02

汇报人：严笑然

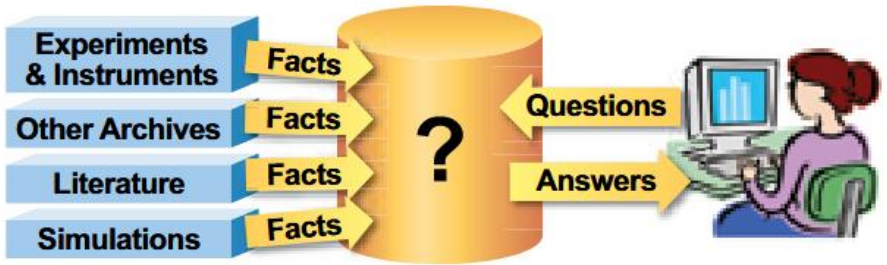
牵头单位：之江实验室

参与单位：东南大学、国家天文台

AI for Science 的前身eScience

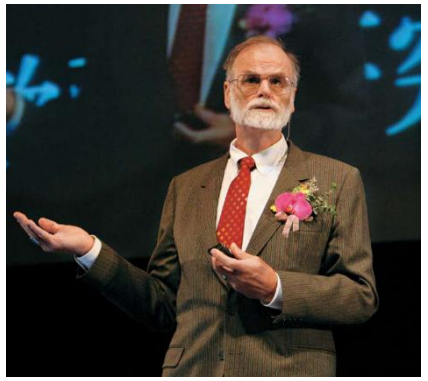
X-Info

- The evolution of X-Info and Comp-X for each discipline X
- How to codify and represent our knowledge



The Generic Problems

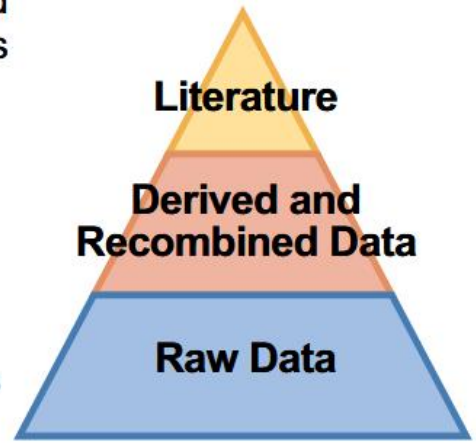
- Data ingest
- Managing a petabyte
- Common schema
- How to organize it
- How to reorganize it
- How to share it with others
- Query and Vis tools
- Building and executing models
- Integrating data and literature
- Documenting experiments
- Curation and long-term preservation



Jim Gray,
Turing awardee

All Scientific Data Online

- Many disciplines overlap and use data from other sciences
- Internet can unify all literature and data
- Go from literature to computation to data back to literature
- Information at your fingertips for everyone-everywhere



- Increase Scientific Information Velocity
- Huge increase in Science Productivity

从大数据到大模型，从云原生到AI原生

阿里云官方博客



目录

01 大数据时代的文献搜索

02 AI时代的天文数据知识关联搜索

03 Introducing 天问1.0Beta



Lessons from different fields

Data Phase	Astronomy	Twitter	Genomics
Acquisition	25 zetta-bytes/year	0.5–15 billion tweets/year	1 zetta-bases/year
Storage	1 EB/year	1–17 PB/year	2–40 EB/year
Analysis	In situ data reduction	Topic and sentiment mining	Heterogeneous data and analysis
	Real-time processing	Metadata analysis	Variant calling, ~2 trillion central processing unit (CPU) hours
	Massive volumes		All-pairs genome alignments, ~10,000 trillion CPU hours
Distribution	Dedicated lines from antennae to server (600 TB/s)	Small units of distribution	Many small (10 MB/s) and fewer massive (10 TB/s) data movement

doi:10.1371/journal.pbio.1002195.t001

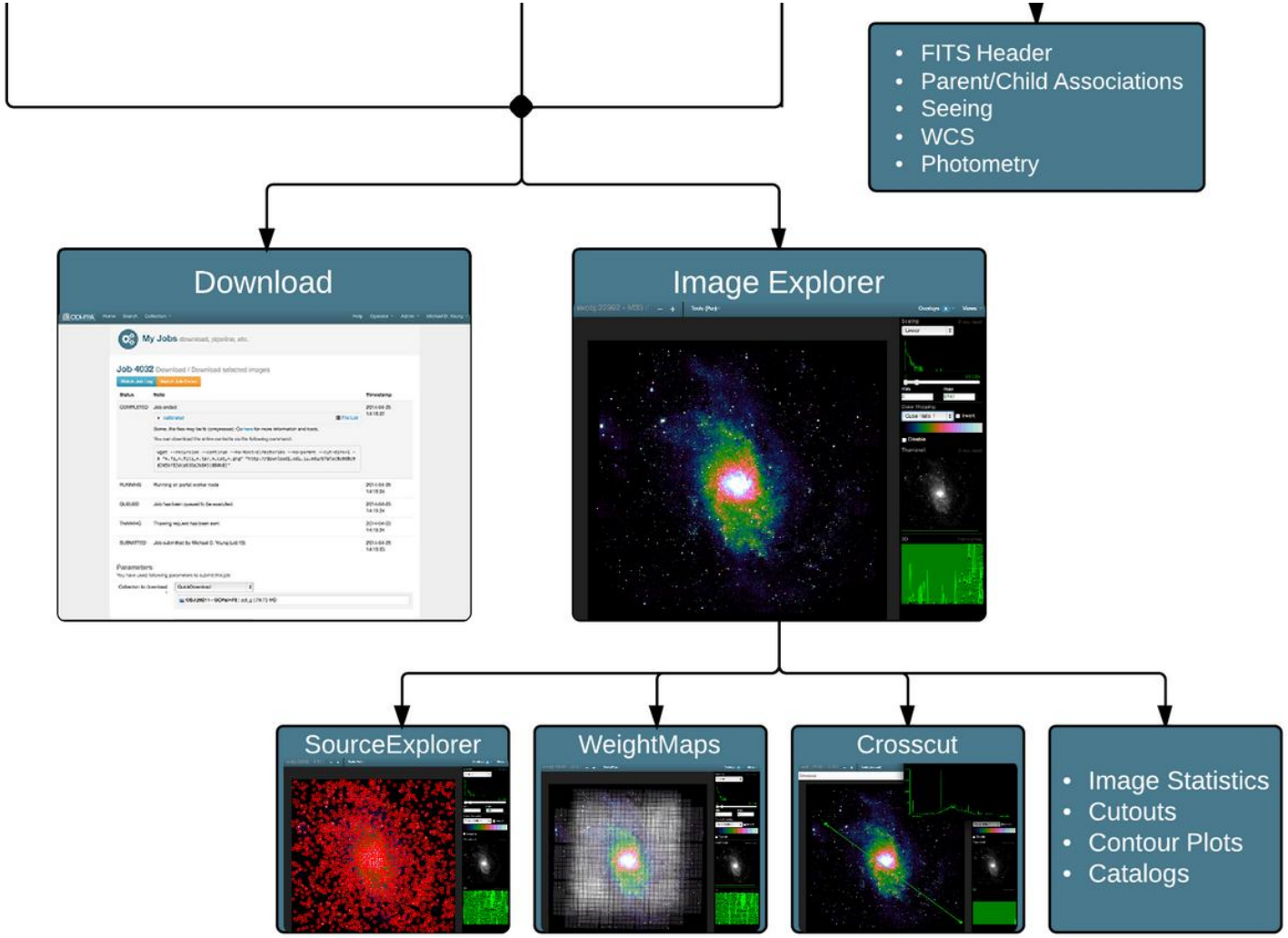
Volume **Velocity** **Variety**



1ZB = 1K EB = (1K)^2 PB

D Stephens et al. Big Data: Astronomical or Genomical? PLoS biology (2015)

Volume: data centric

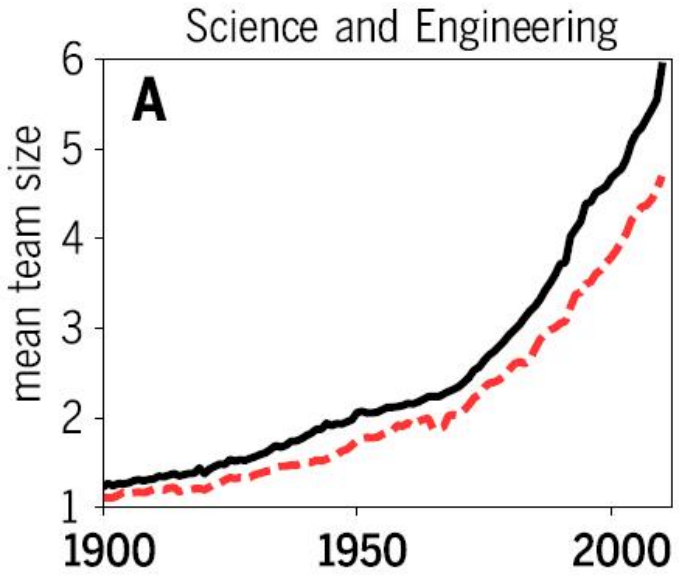


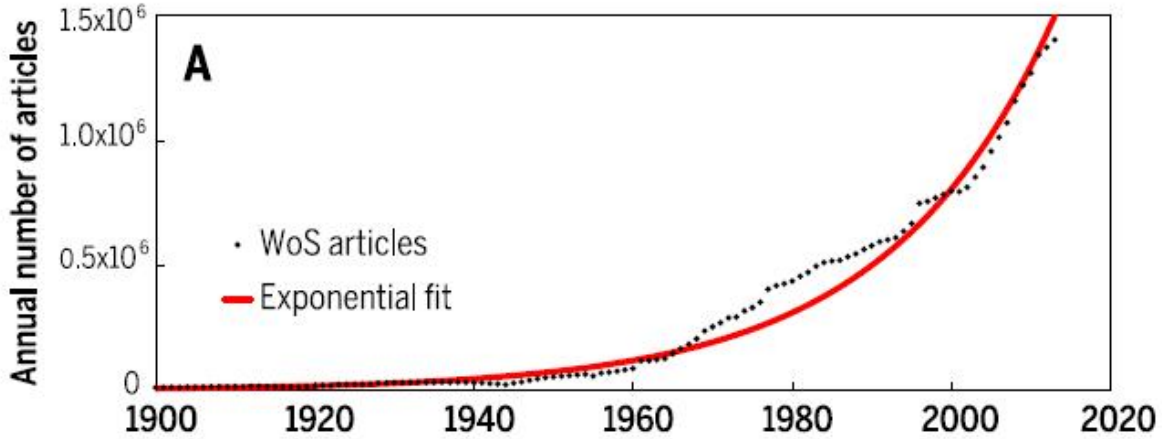
The ODI Pipeline, Portal, and Archive (ODI-PPA) is a web science gateway for a imager installed on the WIYN 3,5m telescope at Kitt Peak (Tucson, AZ)

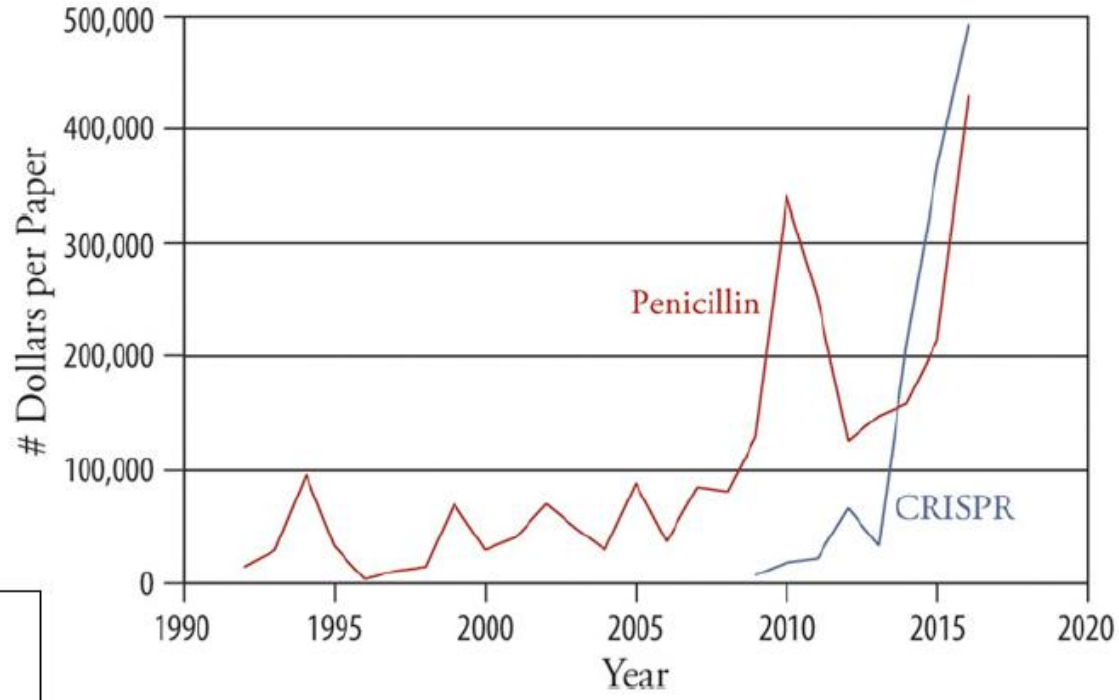
50 terabytes per year

The growth of science

Science of Science. Fortunato et al., Science Mar 2018

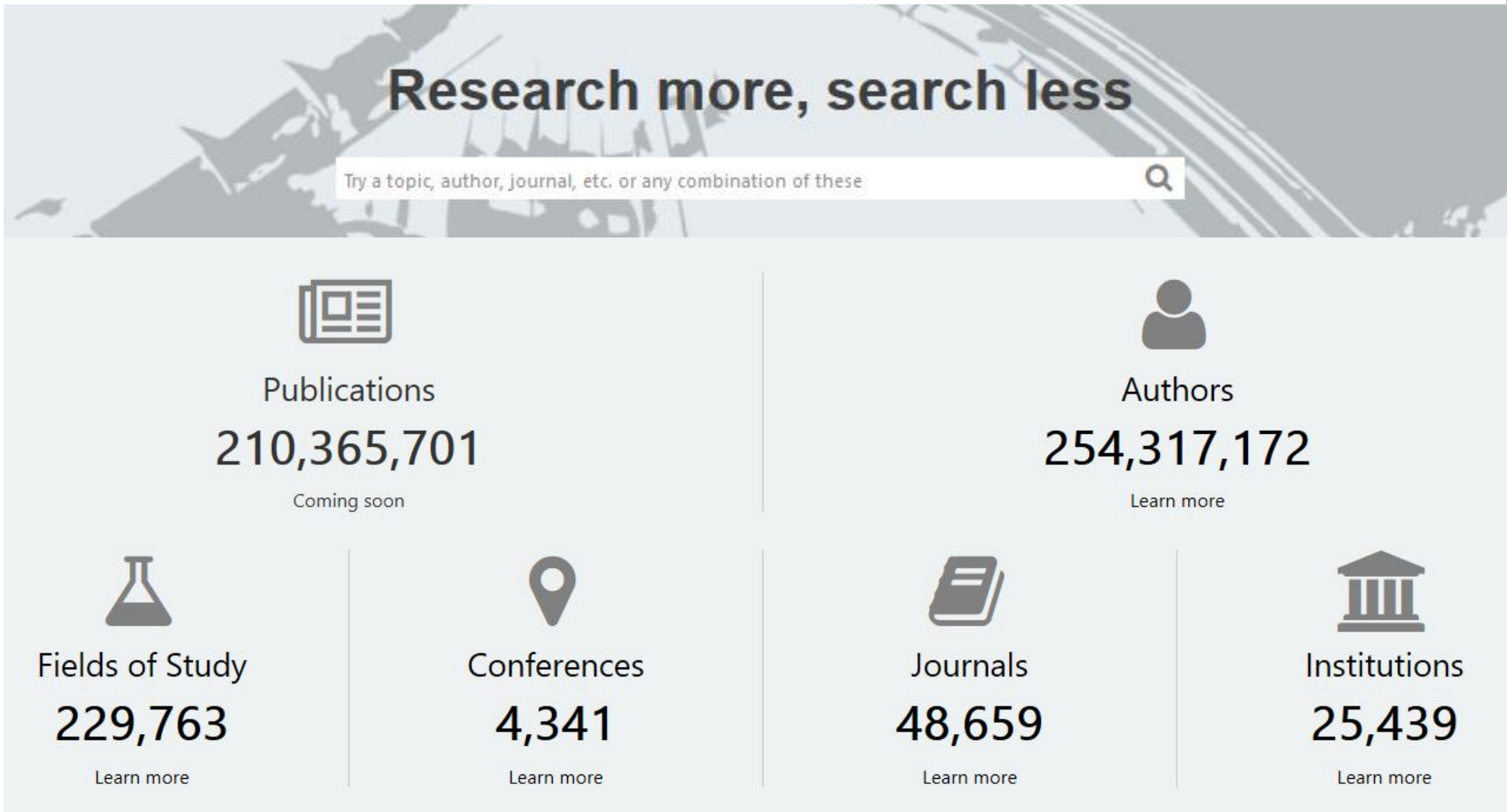






Shiffrin, Richard M., et al. "Scientific progress despite irreproducibility: A seeming paradox." PNAS (2018)

Microsoft Academic



The CADRE proposal



Big Data hosting is cumbersome, expensive, and requires technological expertise



Robert McDonald, former Associate Dean for Research and Technology at IU, suggested a shared hosting solution for all BTAA institutions



BTAA staff presented the solution to the Library Directors



Indiana University and Indiana University Network Science Institute completed the application for the Institute of Museum and Library Services (IMLS) grant

CADRE
Collaborative Archive & Data Research Environment

Overview of CADRE project

Xiaoran Yan
yan30@iu.edu

LIBRARIES

INDIANA UNIVERSITY
NETWORK SCIENCE INSTITUTE

B1G
ACADEMIC ALLIANCE

CC BY NC SA

Collaborative Archive & Data Research Environment (CADRE)



AUTHENTICATION

- Globus
- Shibboleth
- InCommon
- Custom

AUTHORIZATION

Granular Permissions

- Projects
- Teams
- Packages
- Archives
- Tools
- ...

CADRE DASHBOARD

Web-based Graphic Query Interface

JSON CSV XML

Query Results

Upload

Research Asset Commons (Marketplace)

Archives

Packages

Tools

Publish | Reuse | Reproduce

jupyter

Upload

JavaScript

custom scripts

APIs

- Relational Data API
- Graph Data API
- Text Analytics API
- Graph Analytics API
- Other APIs

COMPUTATION INFRASTRUCTURE

Apache Spark · neo4j · docker · PostgreSQL · JanusGraph

aws · Azure · Jetstream · Other computing resources

STORAGE INFRASTRUCTURE

Commercial Cloud Storage

Indexed Database (Relational & Graph)

- Web of Science (WoS)
- Microsoft Academic Graph (MAG)
- USPTO Database

Raw Data (Data Lake)

Institutional Storage

Local Storage

aws · Azure · Jetstream

The CADRE Search Engine

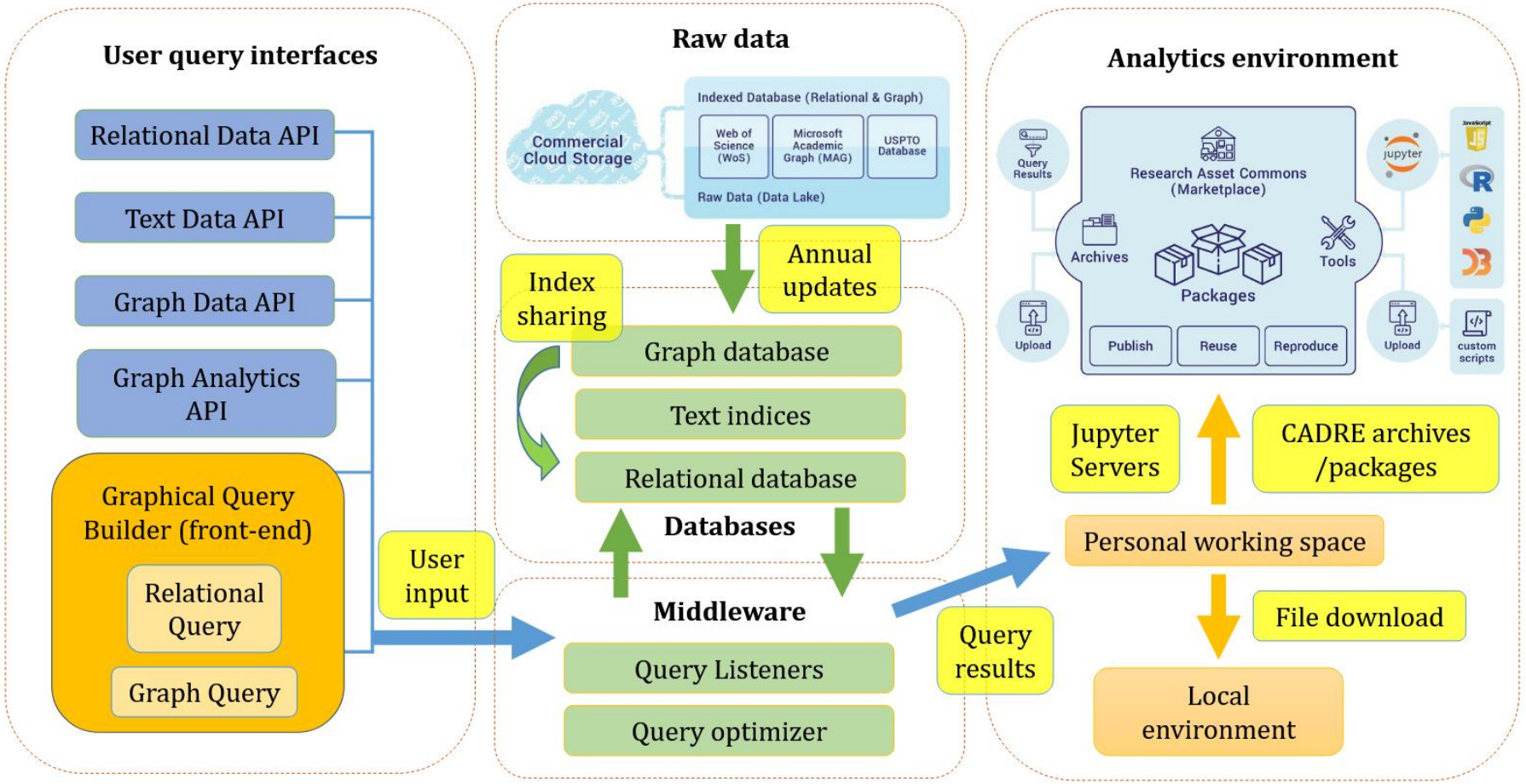
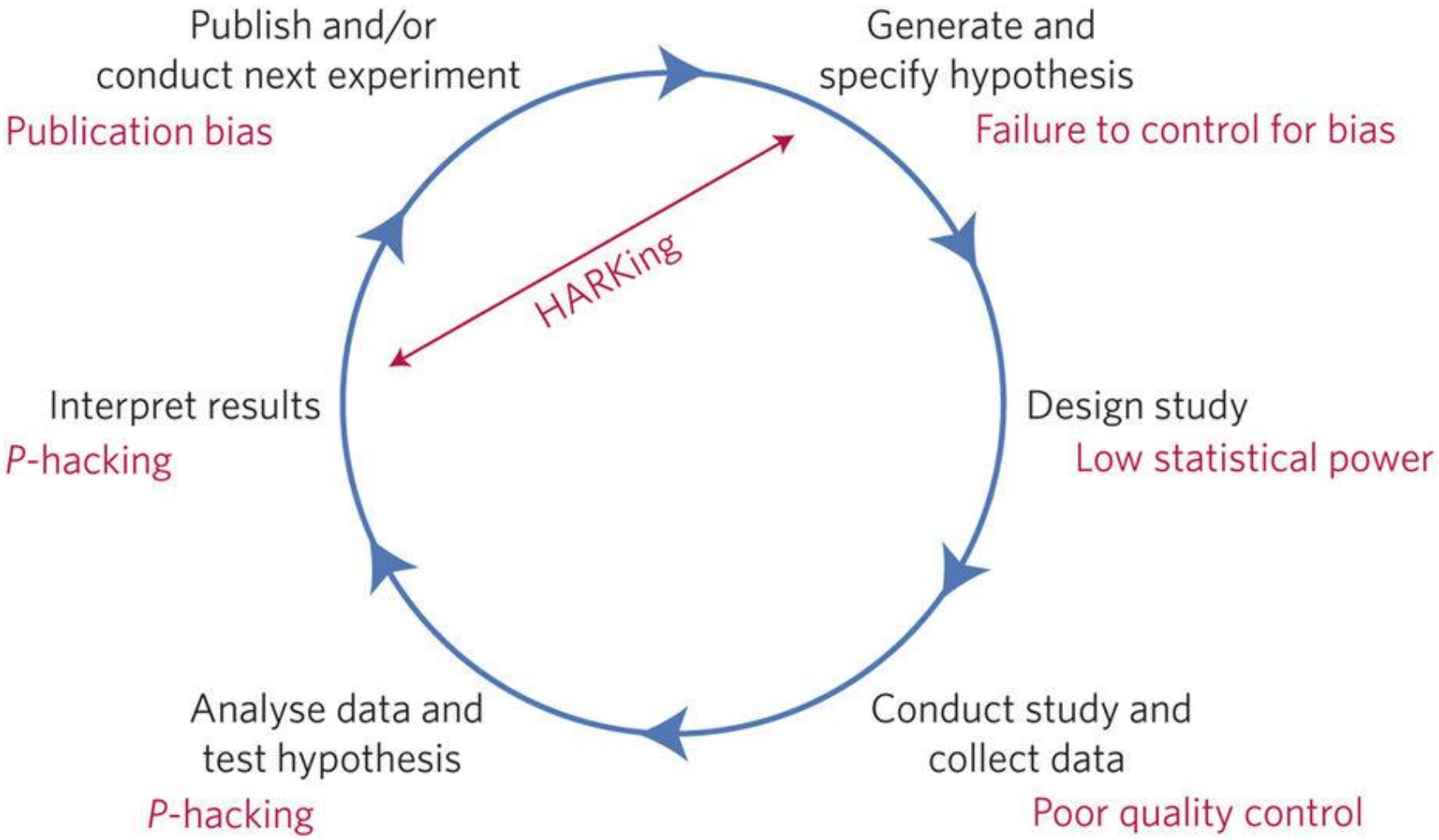


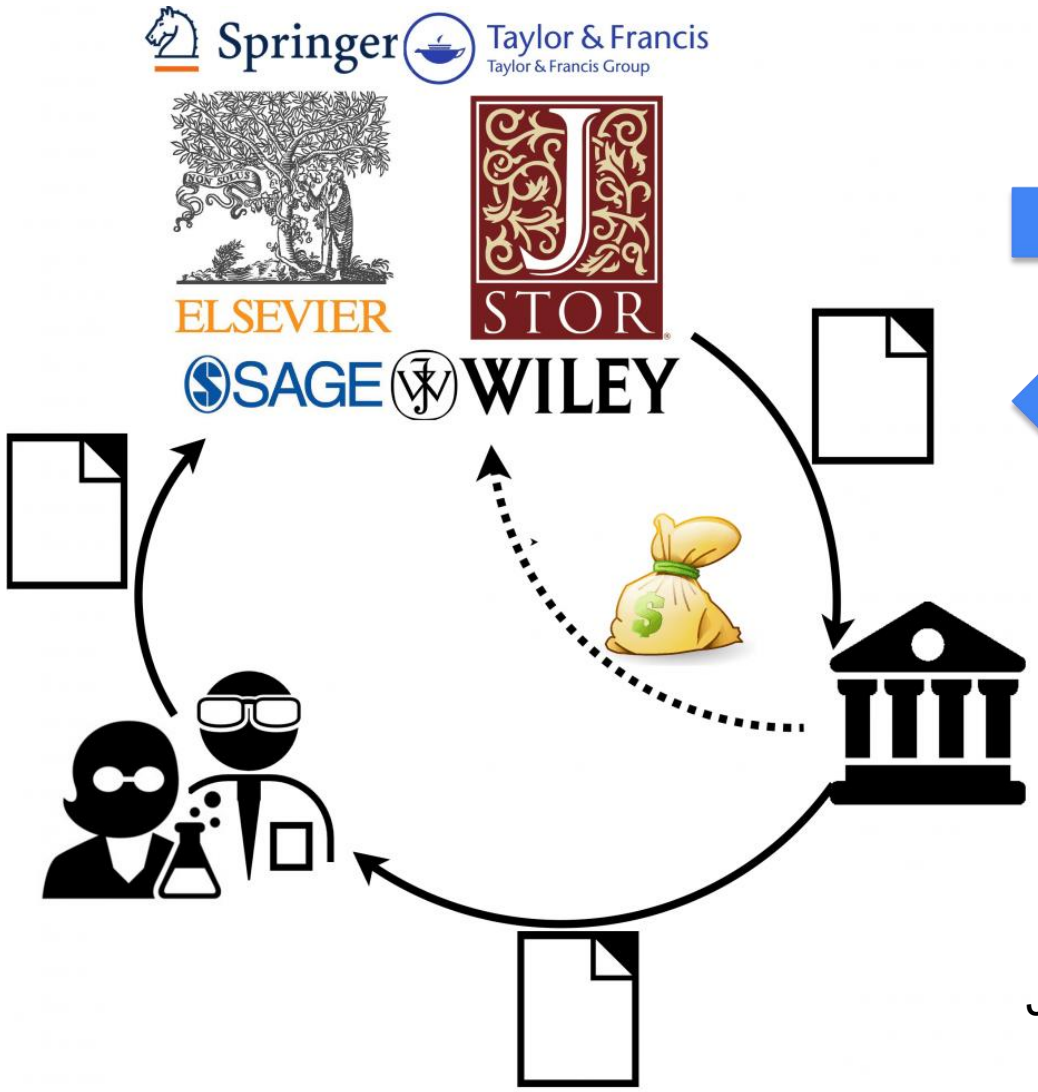
Figure 2: The flow of data on the CADRE platform

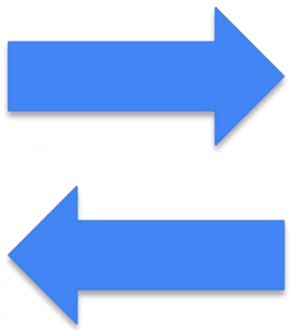
Veracity: the reproducibility “Crisis”



Marcus R. Munafò, et al. “A manifesto for reproducible science” (2017)

A collective effort





Microsoft

Clarivate Analytics

WEB OF SCIENCE™

Joshua Pearce, Potential Effects of Sci-Hub on Academic Publishing (2016)

Spectrum of Reproducibility





Computational → Statistical → Empirical

Stodden, Victoria. “Resolving Irreproducibility in Empirical and Computational Research” (2013)

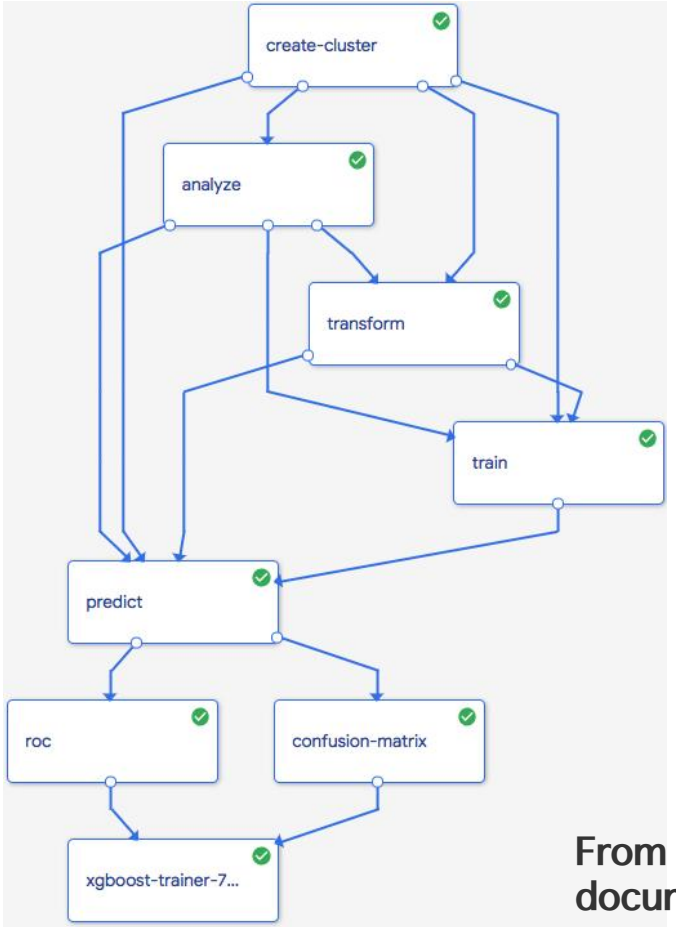
My personal story

Gephi plugin

- A new resolution
- The dream
- The nightmare

Researcher's real life

- 80% data cleaning?
- 40% is actually trying to get things to reproduce
- 20% is to deliver the result

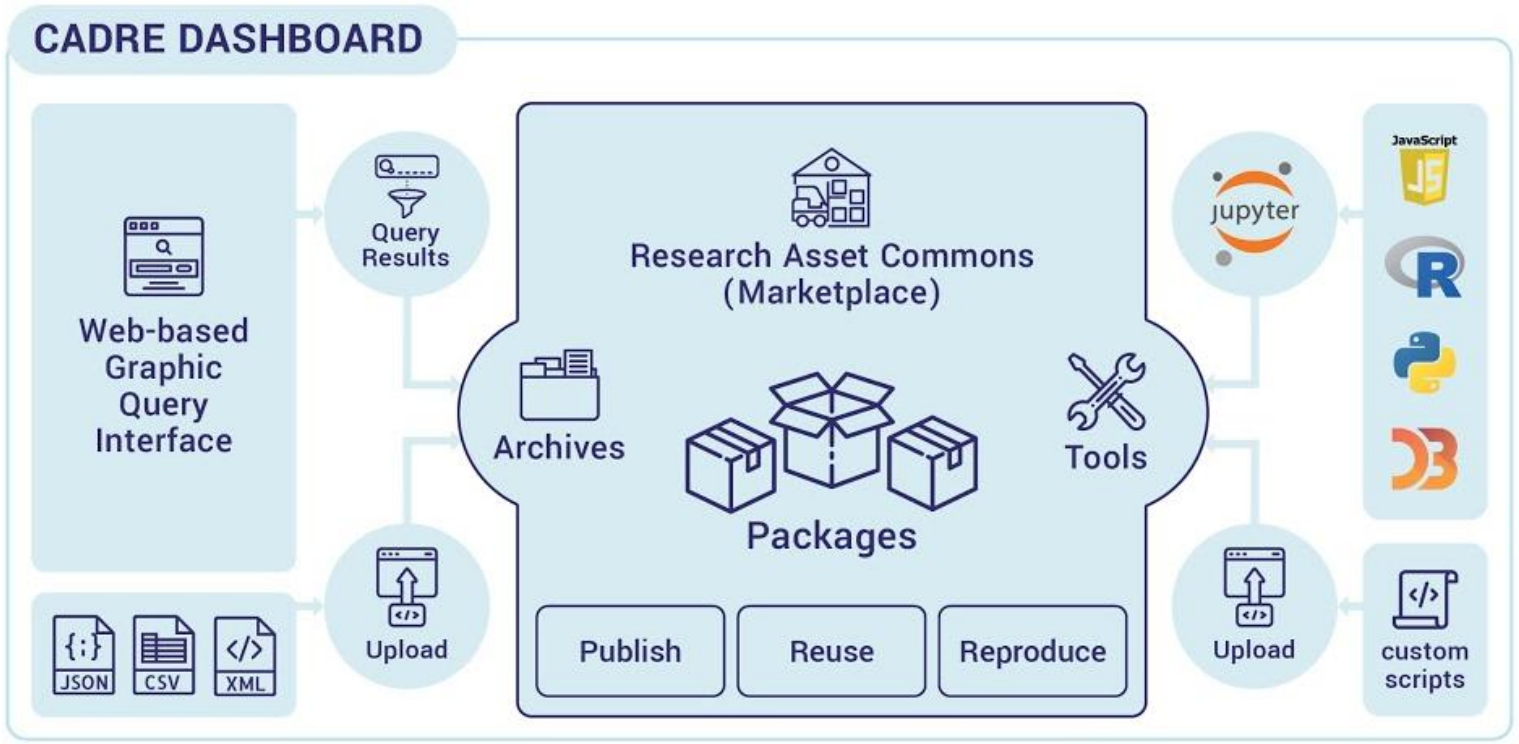


From Kubeflow documentation



The research Asset commons

Engagement and education in Big Data, reproducible

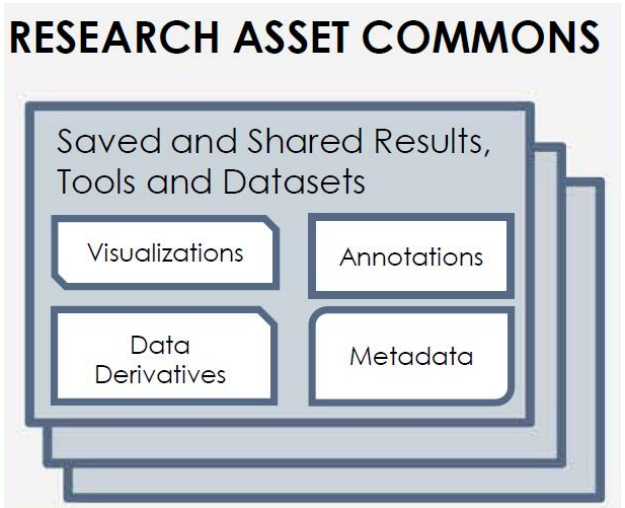


Sharing Data and Apps

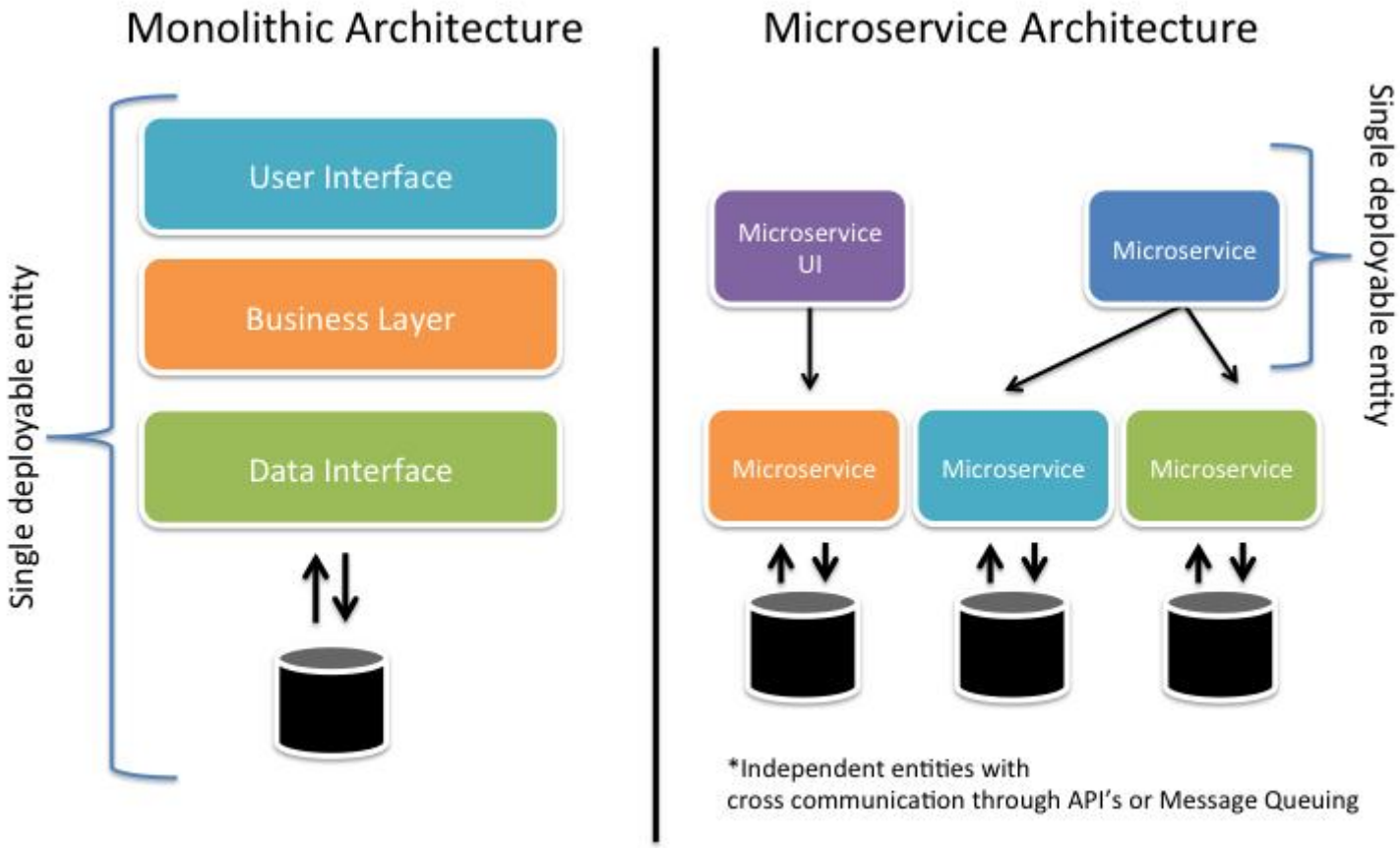
Upload data and share your Apps with other users from different insitutions and scientific domains.

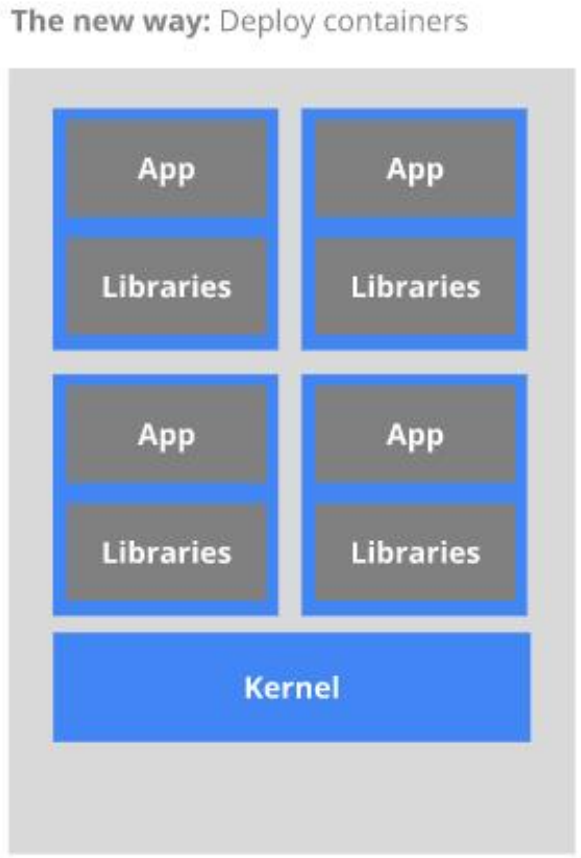
Publishing Results

Publish your research results with a single DOI with a full data provenance for a better reproducibility.



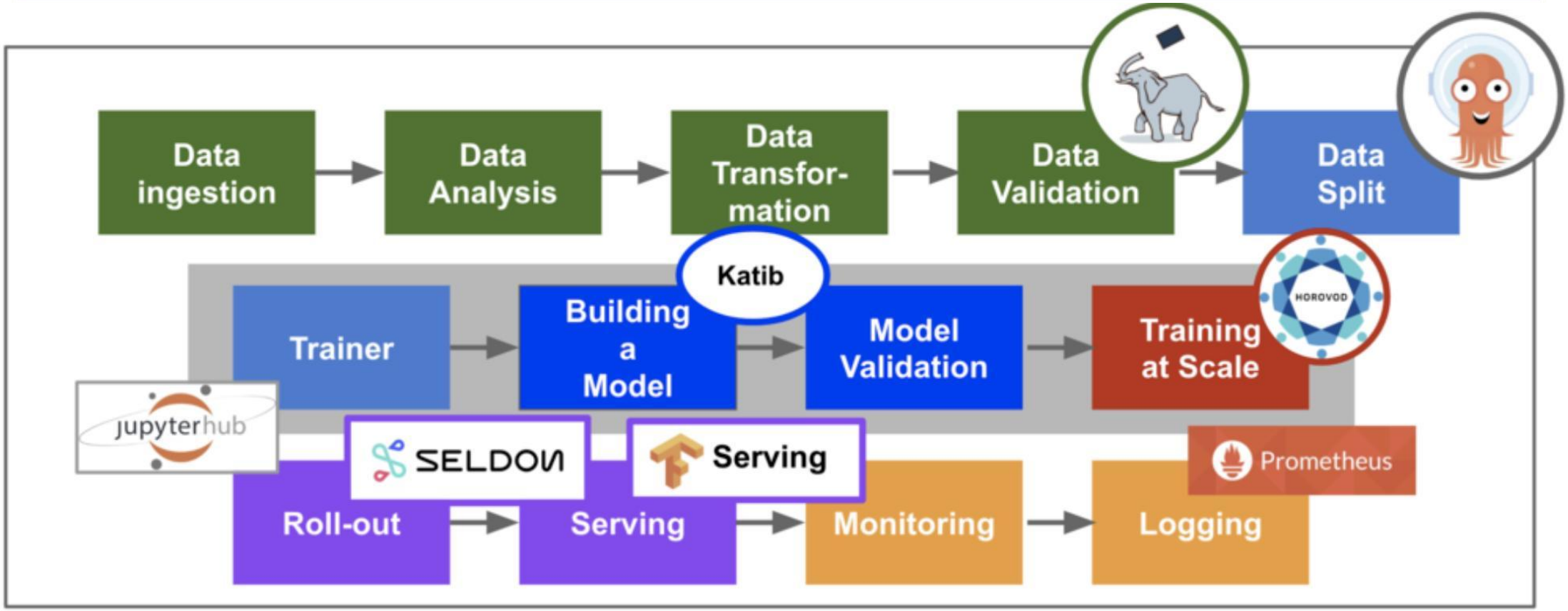
Lessons from the industry





Small and fast, portable
Uses OS-level virtualization

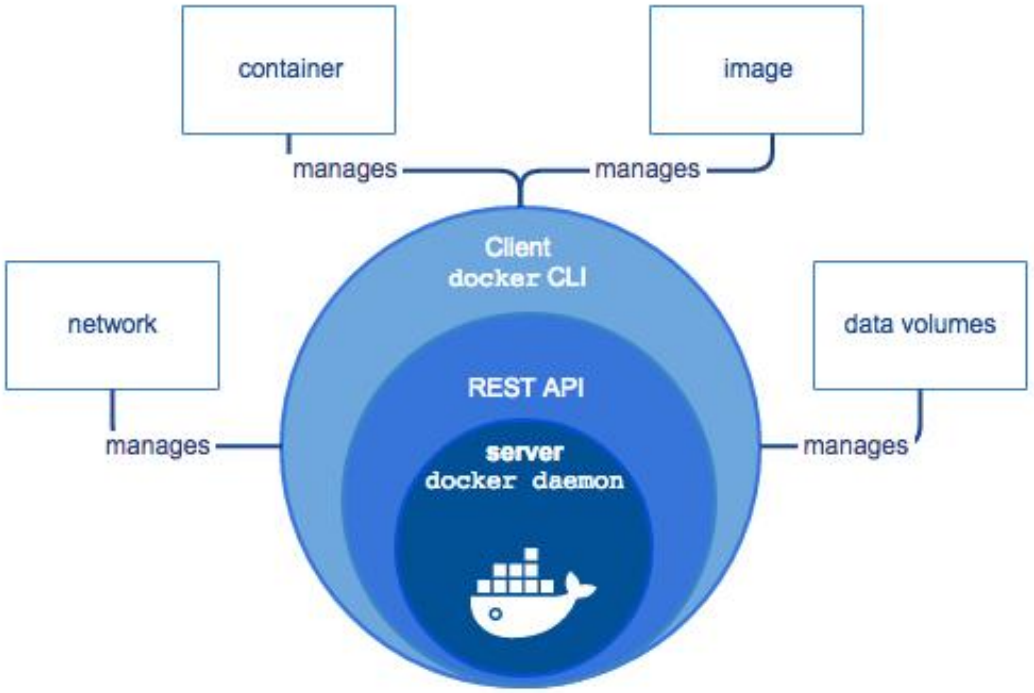
Lessons from the industry







Lessons from the industry



From https://kubernetes.io

Reproducibility

- Self-contained environment
- Code data separation
- Imaging and copying

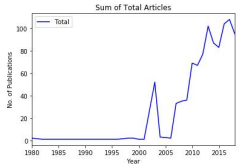
Scalability

- Kubernetes
- Parallelizable data and computing
- Hybrid cloud deployment



```
#Create plot:
inf_plot(x='Year', y=['Total'], color=['blue'], kind='line')
inf_plot = plt.legend(loc = 2)
inf_plot = plt.ylabel(ylabel='No. of Publications')
inf_plot = plt.title(label = 'Sum of Total Articles')

plt.show(inf_plot)
```




```
#Load libraries:
import pandas as pd
import matplotlib.pyplot as plt
```

```
#Load Data
corona_dat = pd.read_csv("query-results/corona_march_4e7ac01f-a8fc-4ee0-b679-7c0e861b63ee.csv", header = 0)

print(corona_dat)
```

```
#Aggregate Data
order = [4,0]
title_jour = corona_dat[[corona_dat.columns[i] for i in order]]

id_j = title_jour.groupby(['journal_display_name']).count()

id_j.reset_index(inplace=True)

id_j = id_j[1:-1]

id_j.columns=["Journal","Total

id_j_plt = id_j[id_j["Total"]
```

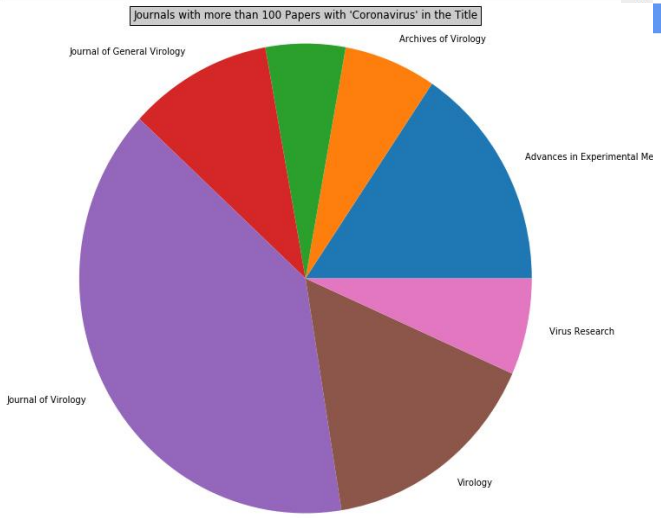
```
#Plot Data
plt.pie(id_j_plt["Total"], lab
plt.show()
```

```
id_j = id_j[1:-1]

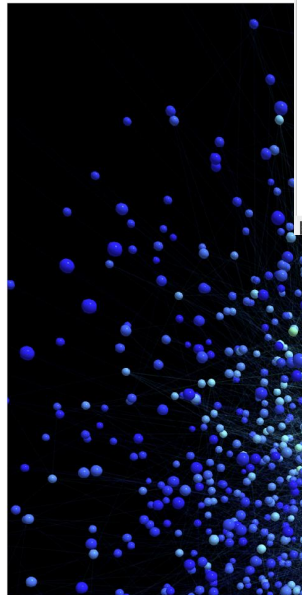
id_j.columns=["Journal","Total"]

id_j_plt = id_j[id_j["Total"]> 100]
```

```
#Plot Data
plt.pie(id_j_plt["Total"], labels=id_j_plt["Journal"], radius = 3)
plt.title("Journals with more than 100 Papers with 'Coronavirus' in the Title",bbox={'facecolor':'0
plt.show()
```



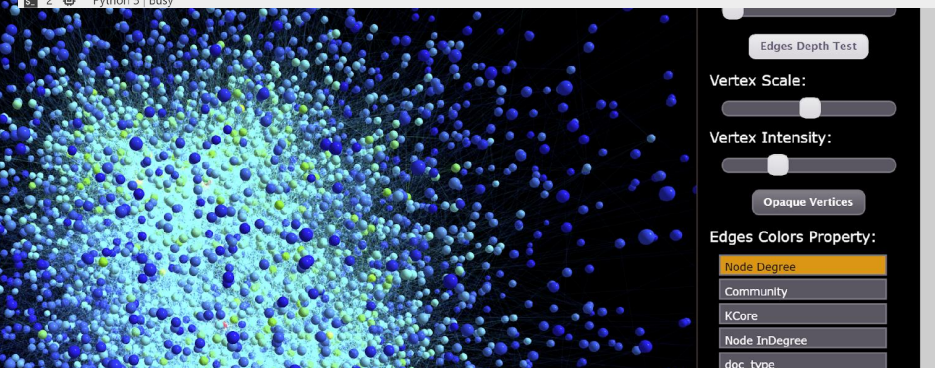

```
#Experimental
from IPython.display import IFrar
IFrame(src='https://filipinascime
```

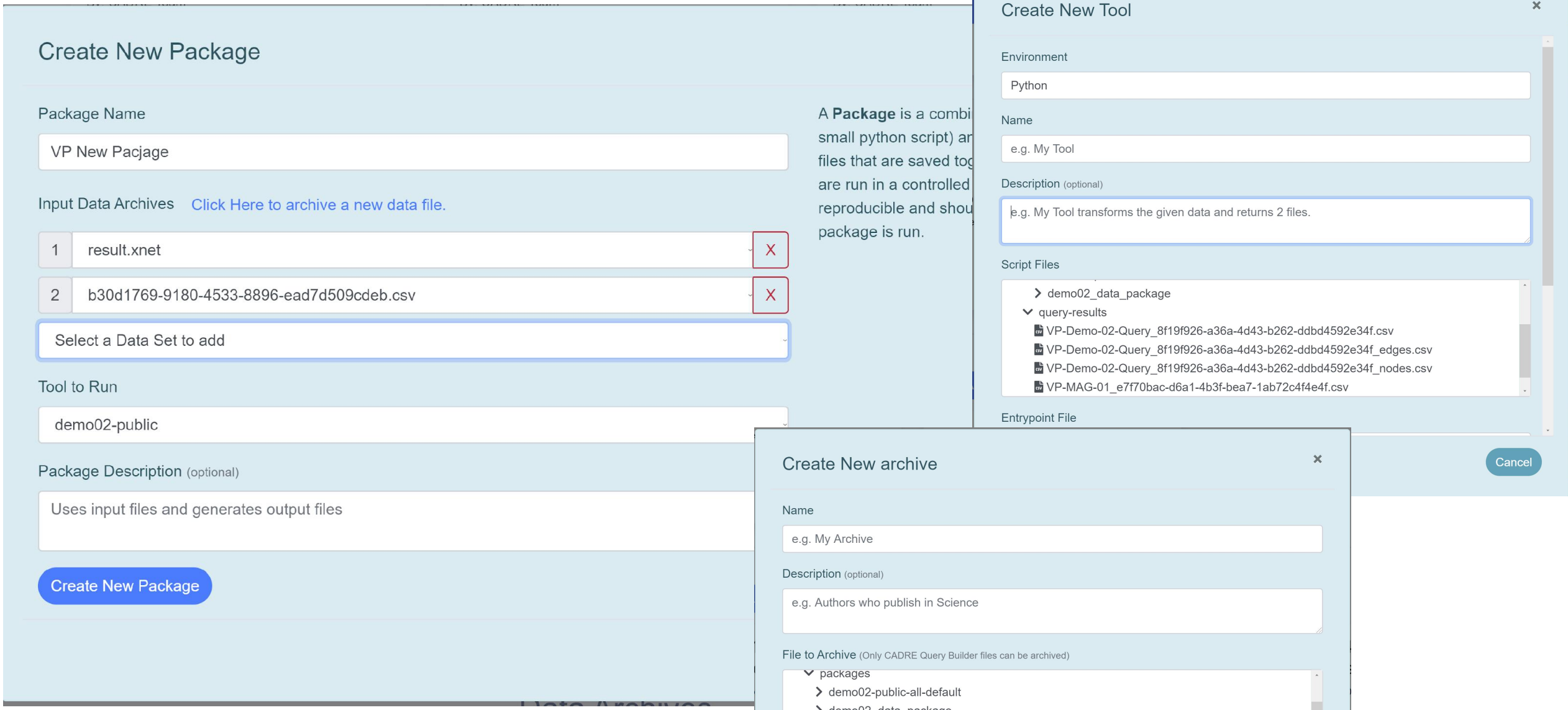


```
# Word Cloud for Abstracts
wc = WordCloud(background_color="white", max_words=2000, width=800,height=400, scale=1,contour_width=3, contour_color='white')
plt.figure(figsize=(20,10))
if("paper_abstract" not in graph.vertex_attributes()):
    textData = graph.vs["original_title"]
else:
    textData =  ["%s. %s"%(titleAndAbstract) for titleAndAbstract in zip(graph.vs["original_title"],graph.vs["paper_abstract"])]

wc.generate("\n".join(textData))
plt.imshow(wc, interpolation='bilinear')
plt.axis("off")
plt.show()
```





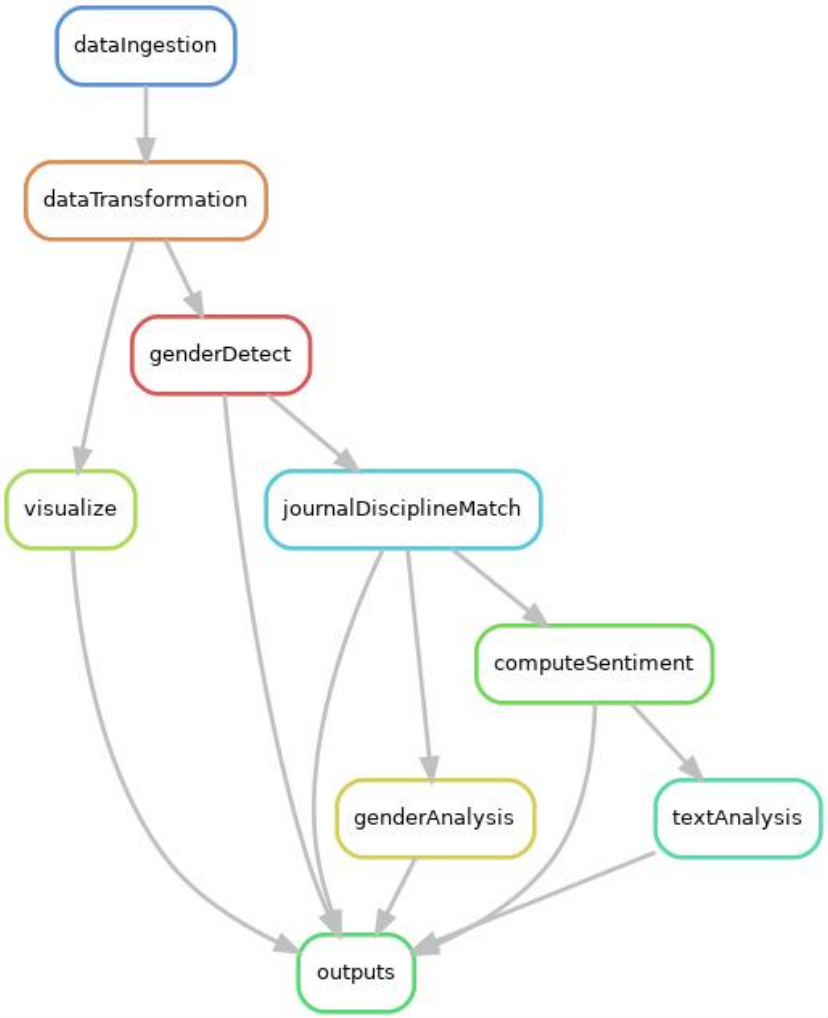

Create New Package
Package Name
VP New Pacjage
Input Data Archives
Click Here to archive a new data file.
1
result.xnet
X
2
b30d1769-9180-4533-8896-ead7d509cdeb.csv
X
Select a Data Set to add
Tool to Run
demo02-public
Package Description (optional)
Uses input files and generates output files
Create New Package
A Package is a combi
small python script) ar
files that are saved tog
are run in a controlled
reproducible and shou
package is run.
Create New Tool
Environment
Python
Name
e.g. My Tool
Description (optional)
e.g. My Tool transforms the given data and returns 2 files.
Script Files
demo02_data_package
query-results
VP-Demo-02-Query_8f19f926-a36a-4d43-b262-ddbd4592e34f.csv
VP-Demo-02-Query_8f19f926-a36a-4d43-b262-ddbd4592e34f_edges.csv
VP-Demo-02-Query_8f19f926-a36a-4d43-b262-ddbd4592e34f_nodes.csv
VP-MAG-01_e7f70bac-d6a1-4b3f-bea7-1ab72c4f4e4f.csv
Entrypoint File
Cancel
Create New archive
Name
e.g. My Archive
Description (optional)
e.g. Authors who publish in Science
File to Archive (Only CADRE Query Builder files can be archived)
packages
demo02-public-all-default
demo02_data_package
query-results
VP-Demo-02-Query_8f19f926-a36a-4d43-b262-ddbd4592e34f.csv
VP-Demo-02-Query_8f19f926-a36a-4d43-b262-ddbd4592e34f_edges.csv
Create New Archive

A live demo

Open science, communal culture, and women's participation in the movement to improve science

Mary C. Murphy, Amanda F. Mejia, Jorge Mejia, Xiaoran Yan, Sapna Cheryan, …
+ See all authors and affiliations

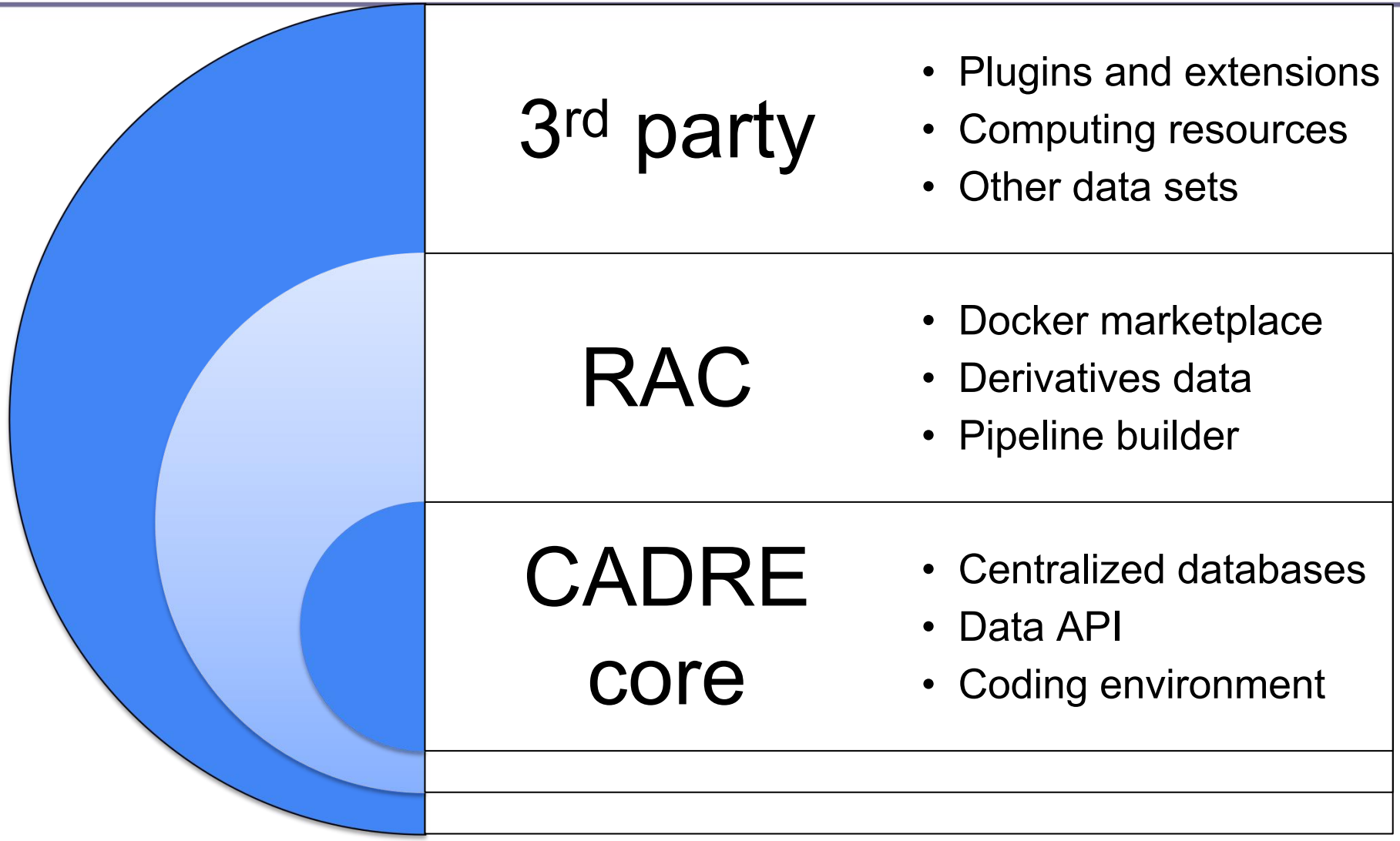
PNAS September 29, 2020 117 (39) 24154-24164; first published September 14, 2020; https://doi.org/10.1073/pnas.1921320117



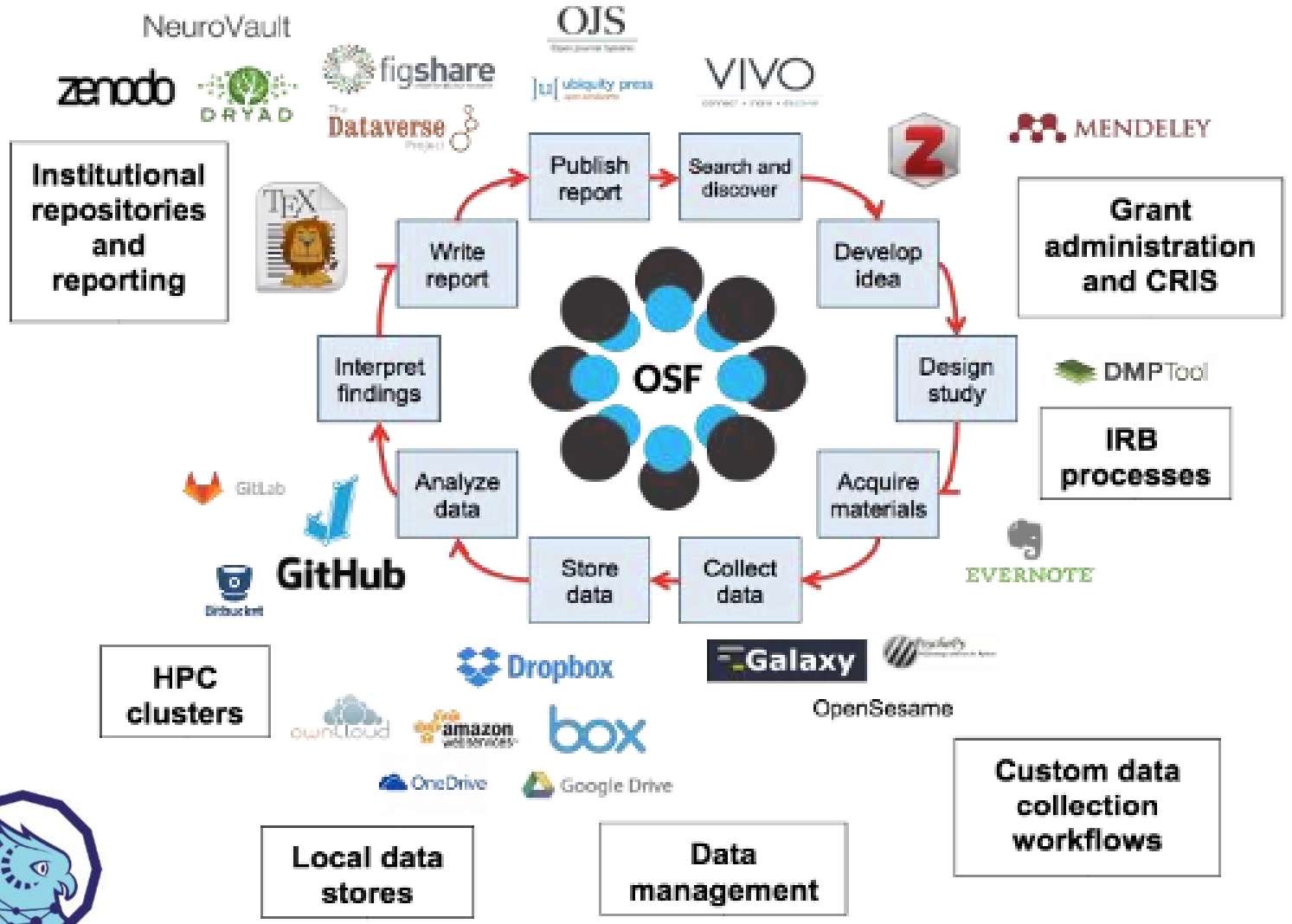
- **基于容器的鲁棒可重复流水线**
  - 云端数据自动导入清理（Spark代码）
  - 作者识别与性别识别（R代码）
  - 学术合作网络的社群结构分析（Java代码）
  - 合作者的情感识别+文本分析（Python代码）
  - 容器化的一键数据计算环境鲁棒复现

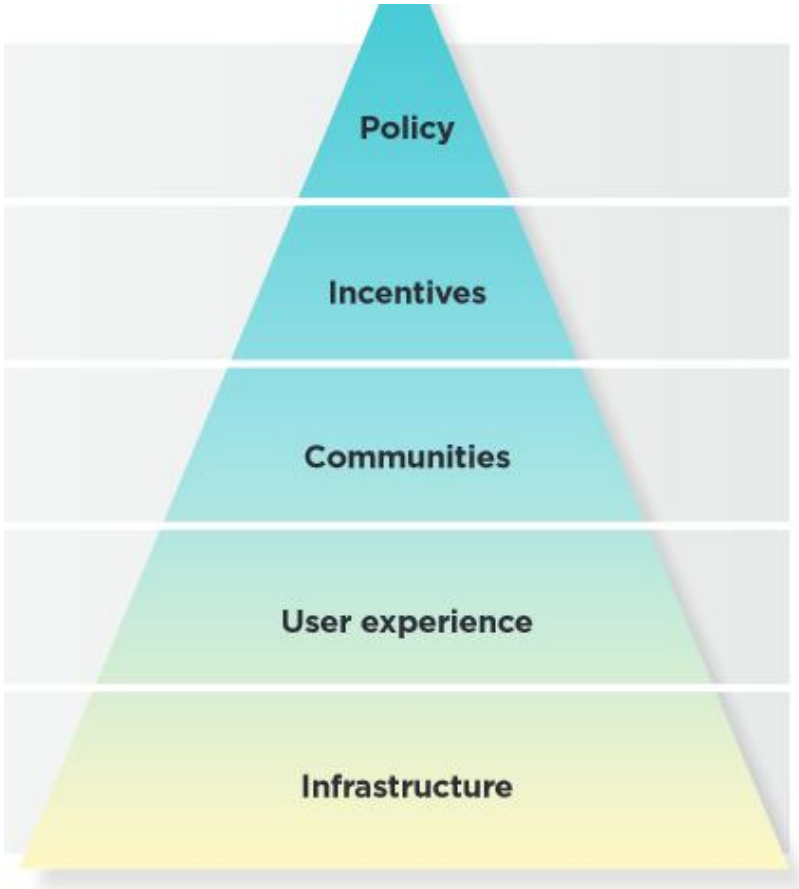


The CADRE ecosystem



Center for open science





目录

01 大数据时代的文献搜索

02 AI时代的天文数据知识关联搜索

03 Introducing 天问1.0Beta



总体目标和研发思路

总体目标：

打造能够智能**识别、关联、推荐**大规模多模态天文科学数据的自主可控软件系统，帮助专业科学工作者提高数据利用效率，推进跨学科合作和公众科普工作。

研发思路：

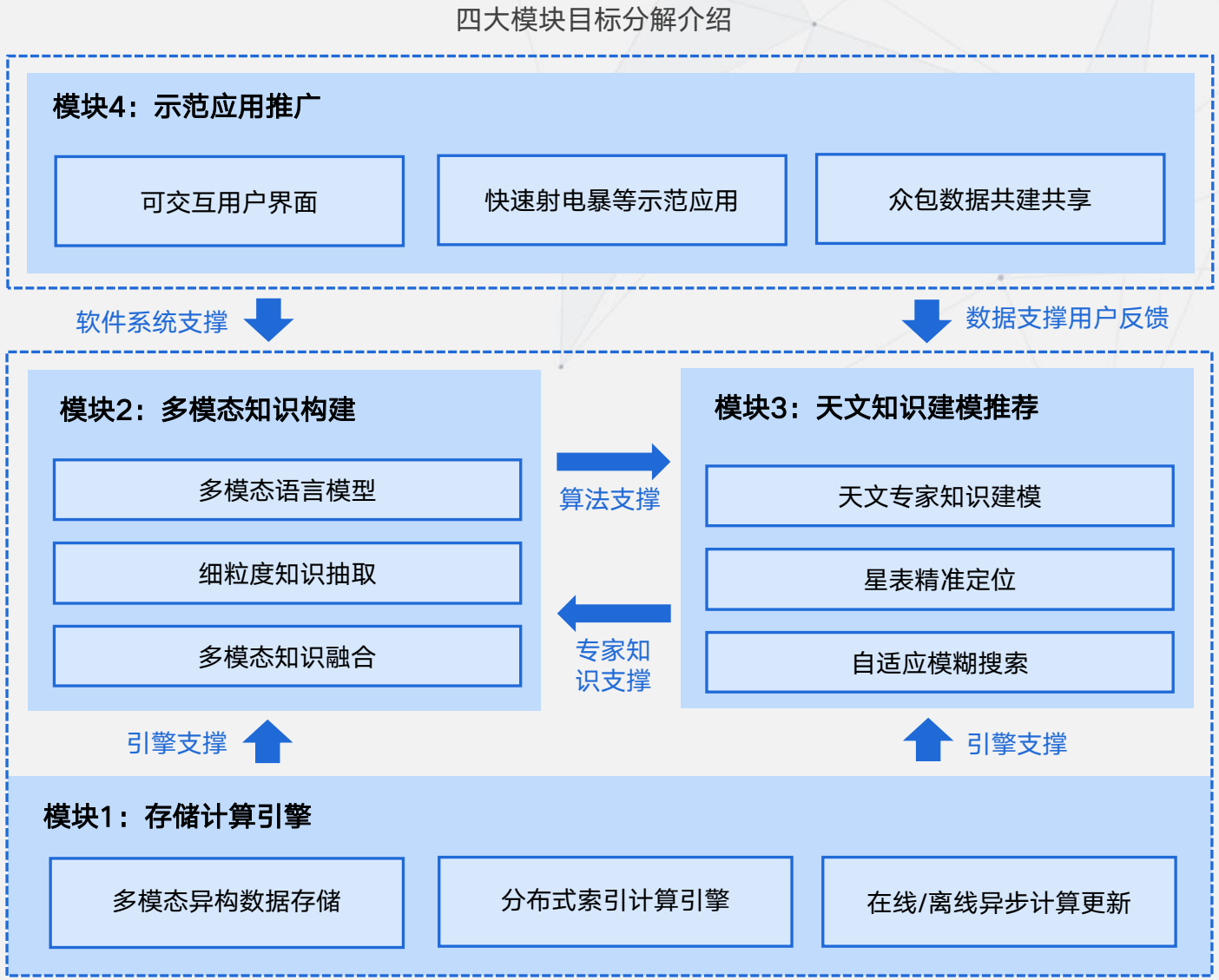
- 充分利用云原生**先进技术**的开源生态，自主研发海量多模态数据的存储计算引擎；
- 发挥团队在天文领域的**专业知识**和多模态知识图谱**智能算法**的优势，人机融合构建细粒度专业图谱；
- 利用团队的天文数据独家优势，基于**平台应用**推动开源数据集的众包共建，不断完善平台的用户体验和服务范围。



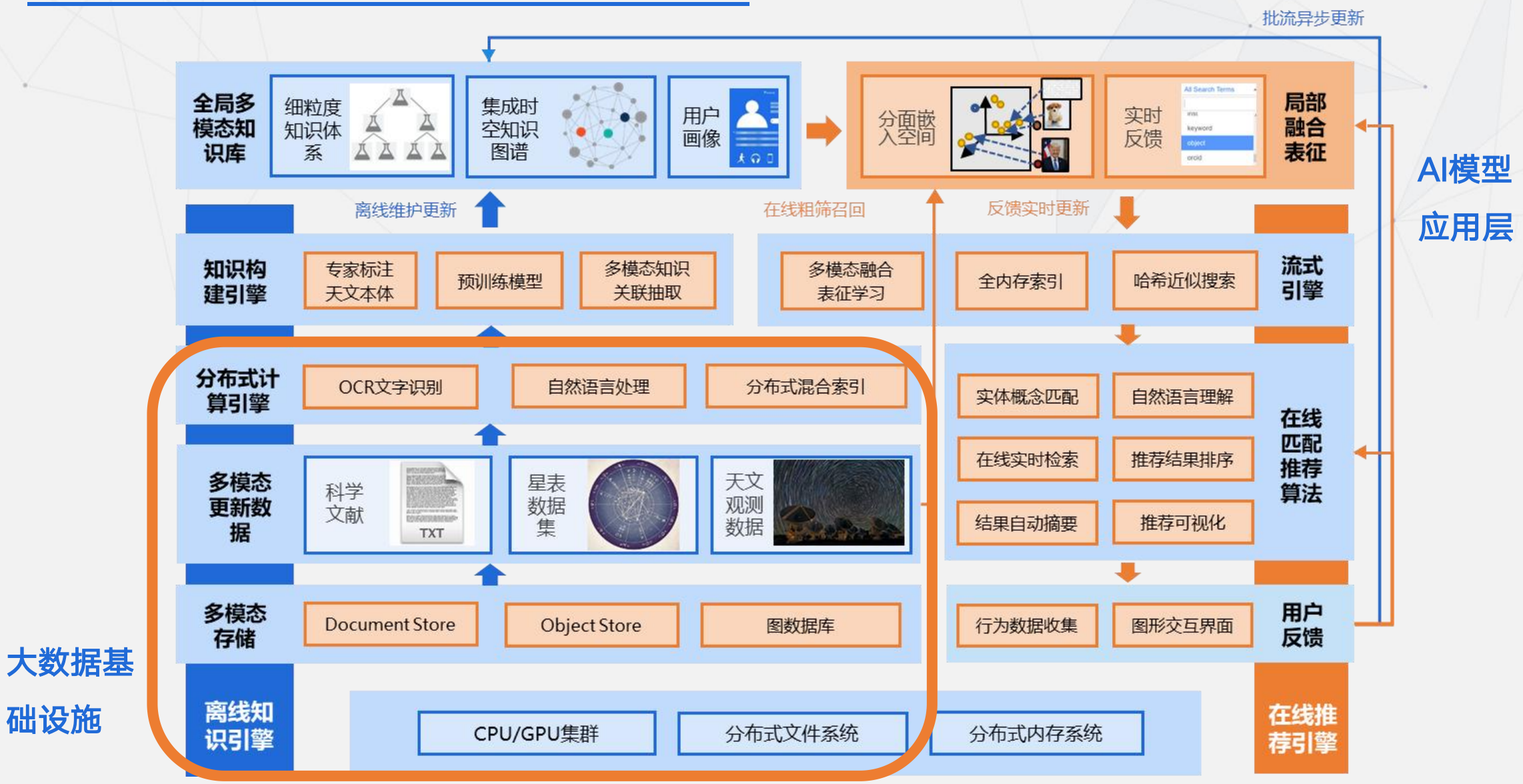
关键问题和任务分工

四大模块：

1. **大规模**多模态数据存储计算 （厉燕）
2. 跨模态细粒度知识**数据构建** （王萌）
3. 天文科学数据精准**自适应**推荐 （杨佳熹）
4. 天文推荐系统的应用部署和数据共建（陶一寒+冯毅）



多模态天文科学数据知识关联推荐引擎

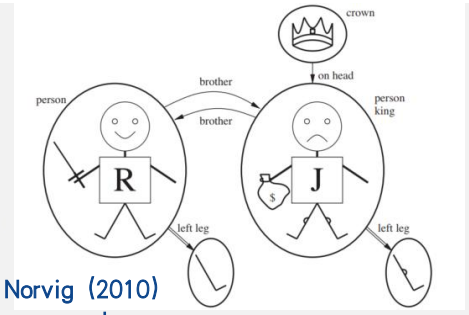


AI推理的两大流派

符号主义：逻辑推理

- 早期的小世界（逻辑驱动）
- 曾经辉煌的专家系统（知识库）
- 规则无法自主学习
- 推理能覆盖的范围极其有限
- 无法表征现实的不确定性

$$\neg Brother(LeftLeg(Richard), John)$$
$$Brother(Richard, John) \wedge Brother(John, Richard)$$
$$King(Richard) \vee King(John)$$
$$\neg King(Richard) \Rightarrow King(John).$$



Russell and Norvig (2010)
AI a modern approach

联结主义：统计推理

- 灵活的模型表征(神经网络)
- 高效的学习算法(反向传播)
- 超大规模的数据和模型
- 所得黑箱模型的不可理解
- 幻觉使得推理无法100%准确
- 对数据的依赖性

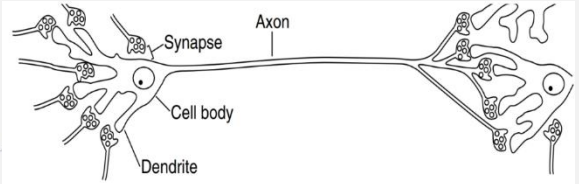


Figure 1.2 A simplified diagram of a neuron, from Crick and Asanuma (1986).

Luger (2009) Artificial Intelligence

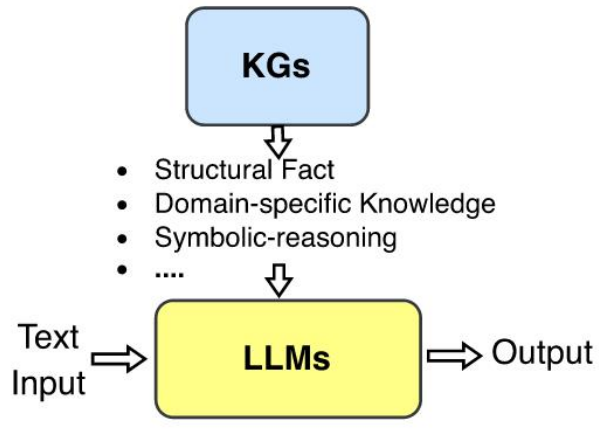
Rule Generation Prompt

```
Logical rules define the relationship between
two entities: X and Y. Each rule is written in
the form of a logical implication, which states
that if the conditions on the right-hand side
(rule body) are satisfied, then the statement on
the left-hand side (rule head) holds true.

Now we have the following rule samples:
husband(X,Y) <-- inv_wife(X,Y)
husband(X,Y) <-- father(X,Z_1) & daughter(Z_1,Y)
...
husband(X,Y) <-- father(X,Z_1) & sister(Z_1,Z_2)
& daughter(Z_2,Y)

Based on the above rules, please generate as
many of the most important rules for the rule
head: "husband(X,Y)"
```

?



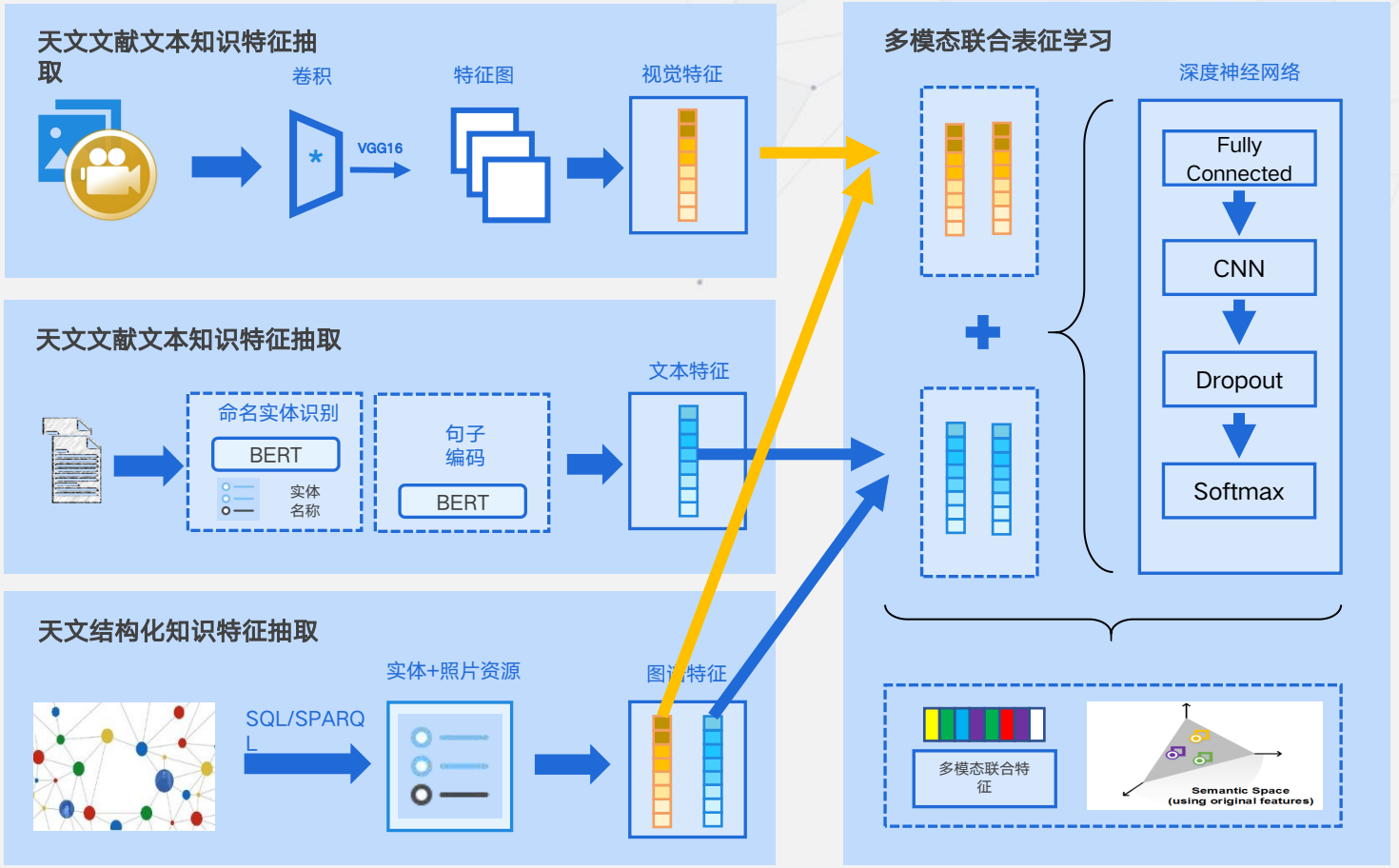
(a) KG-enhanced LLMs

Unifying Large Language Models and Knowledge Graphs: A Roadmap. *IEEE TKDE 2024*

多模态细粒度知识提取关联

技术方案：

- 针对天文学多模态知识领域专业性强、多模态天文数据底层异构的特点，研究基于自监督方式生成**多模态预训练语言模型**。
- 针对天文领域数据标注代价大、多模态天文知识关联挖掘的小样本问题，研究基于对比学习的跨模态**细粒度天文知识抽取**方法。
- 针对天文领域科学文献的实际下游搜索和推荐任务需求，利用跨模态冲突消解以及面向不同任务场景进行**天文知识的分面融合**。



任务1：大规模多模态数据存储计算

文档库

网页

论文

扫描图像

...

专利、论文、书籍

文档预处理

矫正

对齐

增强

多模态系统

视觉版面分析

语言模型

天文文献解析-识别-挖掘-关联流水线

文档内容解析

版面分析

表格解析

公式解析

文本识别

中文文档：PaddleOCR

英文文档：Nougat

其他工具：天壤平台等

可视化评测

Bad

Good

多模态语料库

数据结构化增强

数据清洗

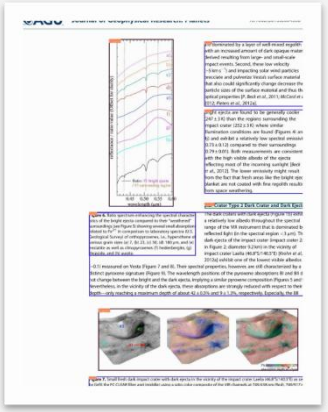
数据清洗

信息过滤

引用文献

……

文档结构化



天文领域知识建模与自适应推荐算法

2. 天文专业图谱的构建：

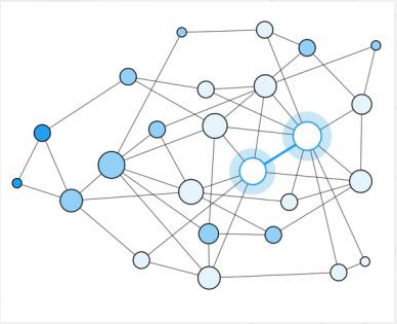
- 基于天文领域的论文中天体、数据集与观测点星表构建**细粒度天文时空知识图谱**，通过检索时空知识图谱观测点或天文现象坐标，研究周围的天体。
- 基于天文领域论文的作者与论文构建**天文领域科研社交网络图谱**。
- 基于天文领域权威分类体系与专家经验**设计关键词图谱的本体架构**，同时，基于关键词在文章中的共现关系，构建领域内关键词子图，用于挖掘领域中以及领域之间的热点研究方向。



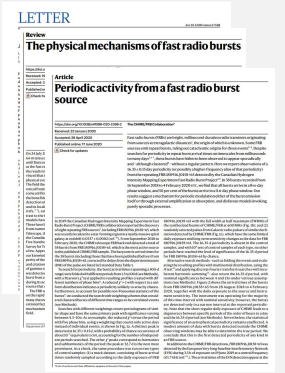
Show 100 entries Search:

N	Identifier	dist(asec)	Otype	ICRS (J2000) RA	ICRS (J2000) DEC	Mag U
1	NGC 6618 206	3.79	*	18 20 04.6531643256	-16 12 09.228889728	
2	CXOU J182004.9-161226	16.92	Y*?	18 20 04.959	-16 12 26.35	
3	[TBH2012] 631	24.32	cor	18 20 03.4	-16 12 31	
4	AGAL G014.979-00.609	29.70	Y*O	18 20 03	-16 12.6	
5	2MASS J18200426-1611292	42.10	Y*O	18 20 04.260	-16 11 29.21	
6	AGAL G014.982-00.624	51.15	Y*O	18 20 06.9	-16 12 48	
7	[ERG2015] 3097	53.35	smm	18 20 00.90	-16 11 55.1	
8	[BFT2007] 10	53.81	X	18 20 02.301	-16 12 55.45	
9	[SSS85] 85	58.32	MoC	18 20 03.77	-16 11 13.7	
10	AGAL G014.986-00.591	68.40	Y*O	18 20 00	-16 11.7	
11	[BFT2007] 23	68.84	X	18 20 06.945	-16 11 12.70	
12	2MASS J18200002-1612423	70.60	Y*O	18 20 00.0291717408	-16 12 42.315457392	
13	2MASS J18200066-1611120	80.30	Y*O	18 20 00.662	-16 11 12.07	
14	2MASS J18201037-1612069	85.78	Y*O	18 20 10.378	-16 12 06.90	
15	[CPA2006] N15	86.83	bub	18 20 05.4	-16 10 45	

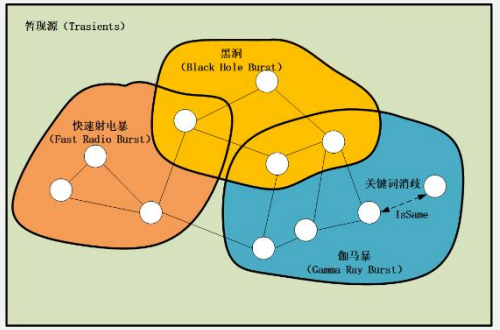
细粒度天文时空知识图谱



天文邻域科研社交网络图谱





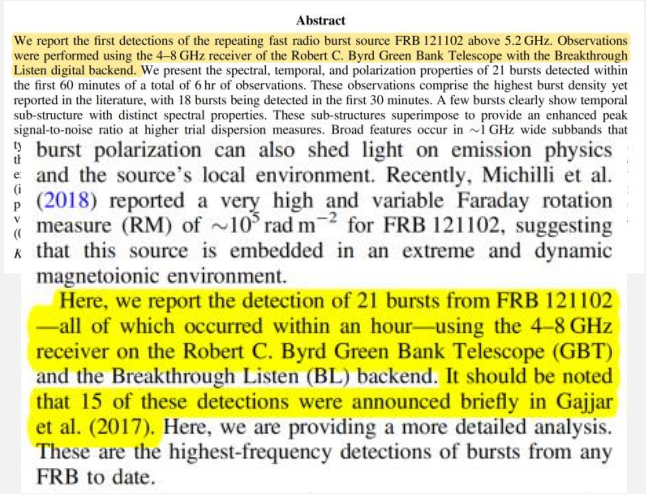


天文关键词图谱

天文领域知识建模与自适应推荐算法

1. 知识图谱本体设计：

- 通过对天文文献中摘要进行**细粒度知识抽取**，获得相关论文中天体类型、天体坐标、数据集与作者信息等具有关联关系的特征。
- 针对天文数据集进行**关键词搜索、特征匹配**，抽取数据集中的关联信息，与天文文献等其他模态的天文数据进行关联。
- 针对快速射电暴领域进行关键词归纳，对FRB文献进行专家标注，构建基于**天文知识的关键词体系**，并拓展到其他天文领域。



Abstract

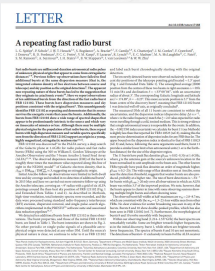
We report the first detections of the repeating fast radio burst source FRB 121102 above 5.2 GHz. Observations were performed using the 4–8 GHz receiver of the Robert C. Byrd Green Bank Telescope with the Breakthrough Listen digital backend. We present the spectral, temporal, and polarization properties of 21 bursts detected within the first 60 minutes of a total of 6 hr of observations. These observations comprise the highest burst density yet reported in the literature, with 18 bursts being detected in the first 30 minutes. A few bursts clearly show temporal sub-structure with distinct spectral properties. These sub-structures superimpose to provide an enhanced peak signal-to-noise ratio at higher trial dispersion measures. Broad features occur in ∼1 GHz wide subbands that burst polarization can also shed light on emission physics and the source's local environment. Recently, Michilli et al. (2018) reported a very high and variable Faraday rotation measure (RM) of $\sim 10^5$ rad m^{-2} for FRB 121102, suggesting that this source is embedded in an extreme and dynamic magnetoionic environment.

Here, we report the detection of 21 bursts from FRB 121102 —all of which occurred within an hour—using the 4–8 GHz receiver on the Robert C. Byrd Green Bank Telescope (GBT) and the Breakthrough Listen (BL) backend. It should be noted that 15 of these detections were announced briefly in Gajjar et al. (2017). Here, we are providing a more detailed analysis. These are the highest-frequency detections of bursts from any FRB to date.

天文文献细粒度知识抽取

Keywords 1/3

Keywords 1/2



LETTER

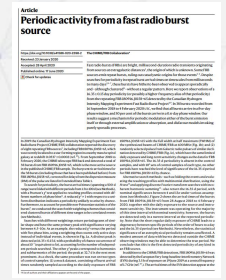
A repeating fast radio burst



Review

The physical mechanisms of fast radio bursts

Keywords 3/4

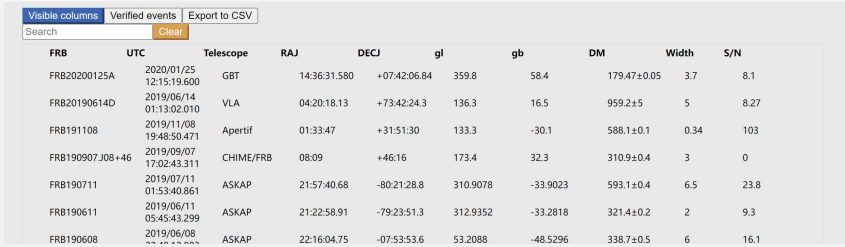


Article

Periodic activity from a fast radio burst source

Keywords 1/2/3/4

文献关键词提取与标注



Visible columns | Verified events | Export to CSV

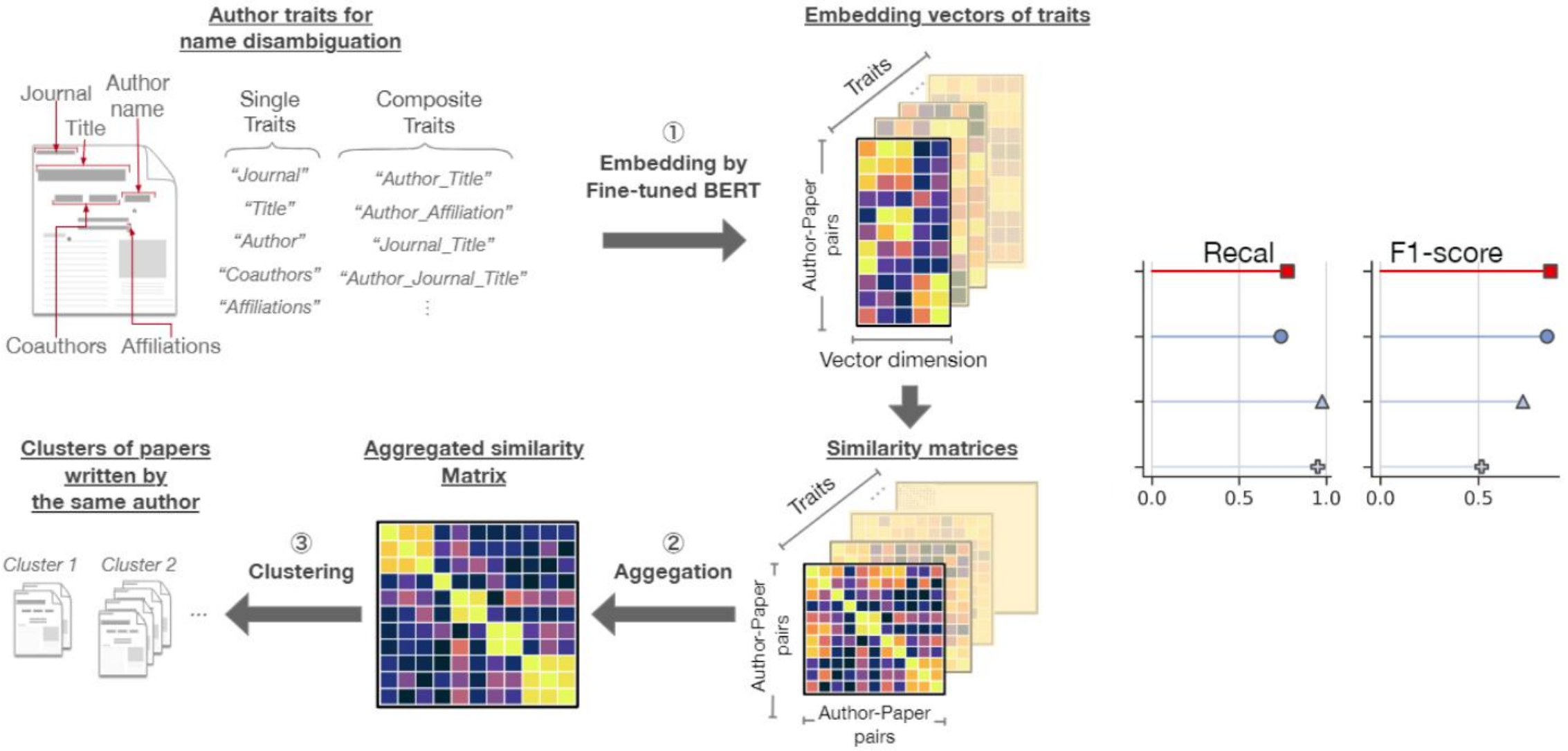
Search | Clear

FRB	UTC	Telescope	RAJ	DECJ	gl	gb	DM	Width	S/N
FRB20200125A	2020/01/25 12:15:19.600	GBT	14:36:31.580	+07:42:06.84	359.8	58.4	179.47±0.05	3.7	8.1
FRB20190614D	2019/06/14 01:13:02.010	VLA	04:20:18.13	+73:42:24.3	136.3	16.5	959.2±5	5	8.27
FRB191108	2019/11/08 19:48:50.471	Apertif	01:33:47	+31:51:30	133.3	-30.1	588.1±0.1	0.34	103
FRB190907.J08+46	2019/09/07 17:02:43.311	CHIME/FRB	08:09	+46:16	173.4	32.3	310.9±0.4	3	0
FRB190711	2019/07/11 01:53:40.861	ASKAP	21:57:40.68	-80:21:28.8	310.9078	-33.9023	593.1±0.4	6.5	23.8
FRB190611	2019/06/11 05:45:43.299	ASKAP	21:22:58.91	-79:23:51.3	312.9352	-33.2818	321.4±0.2	2	9.3
FRB190608	2019/06/08	ASKAP	22:16:04.75	-07:53:53.6	53.2088	-48.5296	338.7±0.5	6	16.1

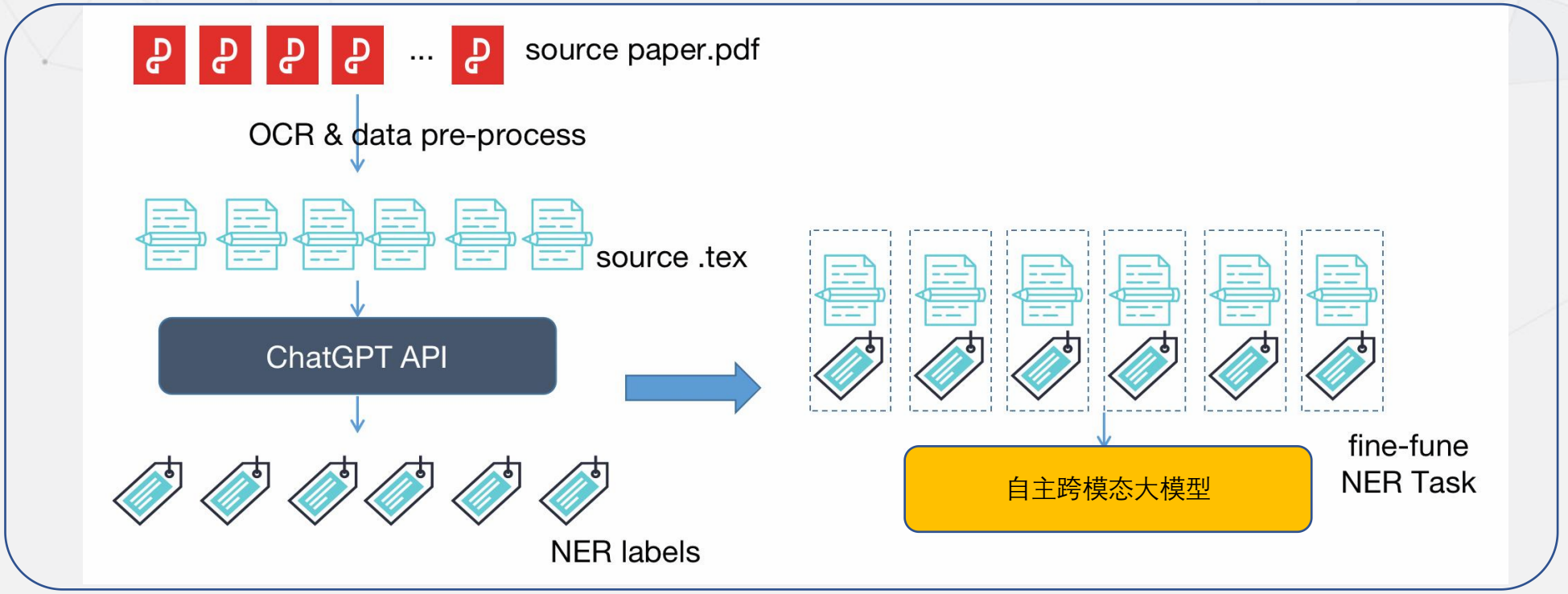
数据集关键信息提取

我们的工作

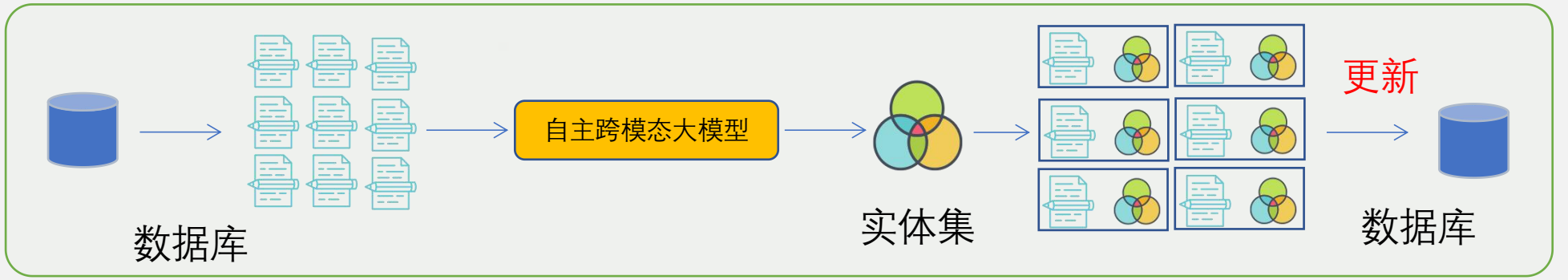
之江实验室 ZHEJIANG LAB



天文细粒度知识提取关联



标签制作/微调阶段



实体识别/数据库更新

整体框架：人机融合学科知识图谱持续迭代

文献同名作者消除歧义案例

- LLM抽取细粒度天文知识构造知识图谱
- 基于LLM挖掘生成不确定性的逻辑规则
- 软规则图谱推理

Rule Generation Prompt

```
Logical rules define the relationship between
two entities: X and Y. Each rule is written in
the form of a logical implication, which states
that if the conditions on the right-hand side
(rule body) are satisfied, then the statement on
the left-hand side (rule head) holds true.

Now we have the following rule samples:
husband(X,Y) <-- inv_wife(X,Y)
husband(X,Y) <-- father(X,Z_1) & daughter(Z_1,Y)
...
husband(X,Y) <-- father(X,Z_1) & sister(Z_1,Z_2)
& daughter(Z_2,Y)

Based on the above rules, please generate as
many of the most important rules for the rule
head: "husband(X,Y)"
```

Rule Samples

$$r_h \leftarrow r_1 \wedge r_2 \wedge r_5$$
$$r_h \leftarrow r_1 \wedge r_5 \wedge r_6$$
$$r_h \leftarrow r_1 \wedge r_3 \wedge r_4 \wedge r_5$$
$$\cdots$$

Prompt: Based on sampled rules, please generate important logical rules for relation r_h.

规则实例化采样

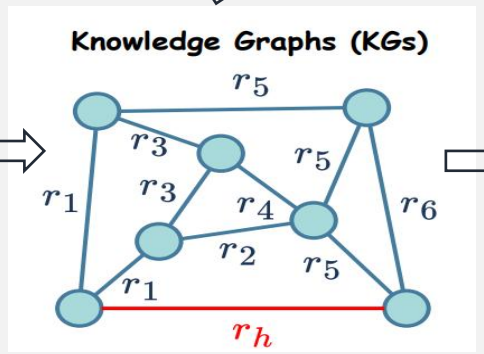
天文学科论文语料

+ Prompt

LLMs

知识抽取

关系、实体…



软逻辑推理

New Facts

(e_1, r_h, e_2)
(e_3, r_h, e_4)
…

生成候选规则

Output

```
husband(X,Y) <-- inv_wife(X,Y)
husband(X,Y) <-- father(X,Z_1) & daughter(Z_1, Y)
husband(X,Y) <-- nephew(X,Z_1) & aunt(Z_1,Y)
husband(X,Y) <-- husband(X,Z_1) & brother(Z_1, Y)
....
```

规则排序

$$r_h \leftarrow r_1 \wedge r_2 \wedge r_5 : 0.9$$
$$r_h \leftarrow r_1 \wedge r_5 \wedge r_6 : 0.7$$
$$\cdots$$

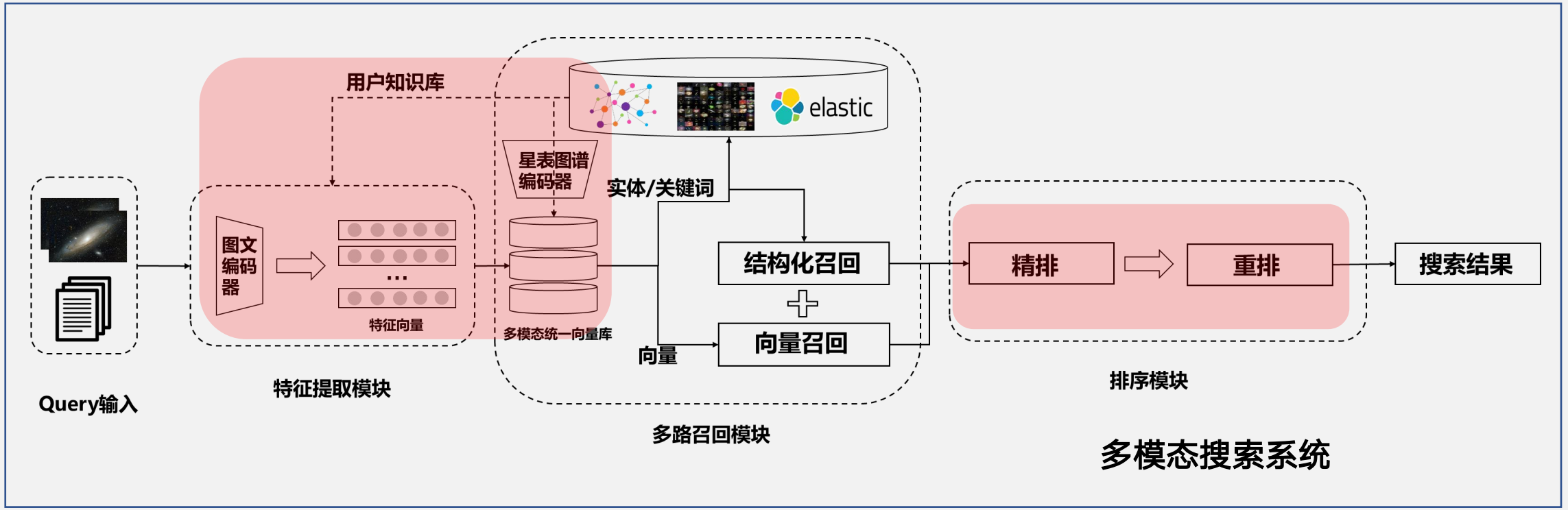
人在回路主动学习

基于混合专家多编码器架构的多路个性化检索

■ **目标：**一个**结合用户信息**（用户属性信息和历史交互信息）的**多模态**文献和数据资源**个性化搜索**系统。

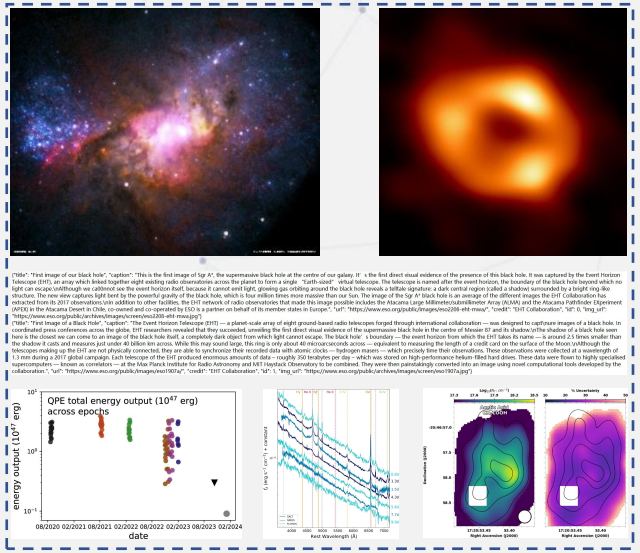
■ 工作特色（相对于ADS）：

1. 搜索引擎侧：新增支持多模态的文献检索能力；
2. 功能侧：增加基于用户知识库的个性化搜索能力以及增量数据处理、特征抽取、索引等能力。



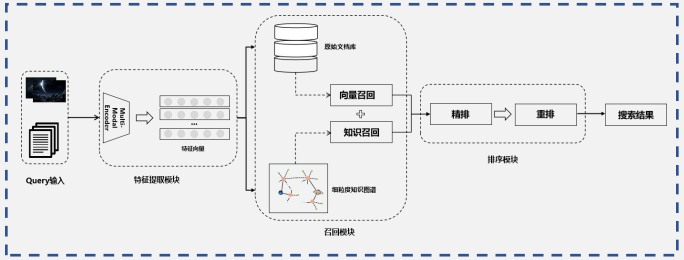
多模态数据表征融合对齐

数据相关



1. （已完成）天文文献统计可视化图文对数据 125+万对；
   a. Arxiv数据集 122万+对；
   b. ESO论文数据 3万+对。
2. （已完成）爬取和下载图文星体类数据 3个数据集 91+万张。

多模态搜索推荐系统



1、（已完成）多模态天文搜索引擎的设计。

2、（已完成）多模态天文搜索算法的验证和性能测试。

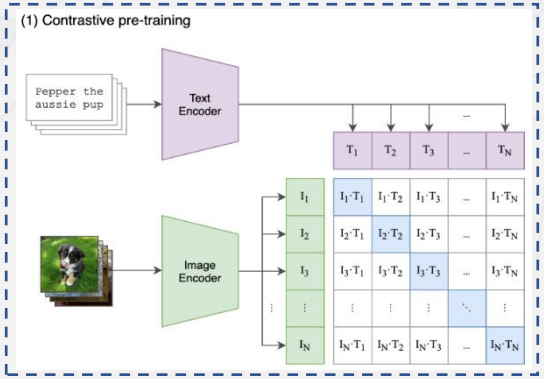
3、（已完成）多模态向量数据库的构建。

4、（进行中）基于天文领域数据集的模型微调和测评。

5、搜索引擎的集成、联调、测试以及部署。

6、国家天文台部署和测试。

天文领域多模态联合表征



1、（已完成）天文图文星体类数据图文表征模型基准测试。

2、（已完成）天文文献类数据图文表征模型基准测试。

3、（进行中）天文图文星体类数据图文表征模型SFT工作。

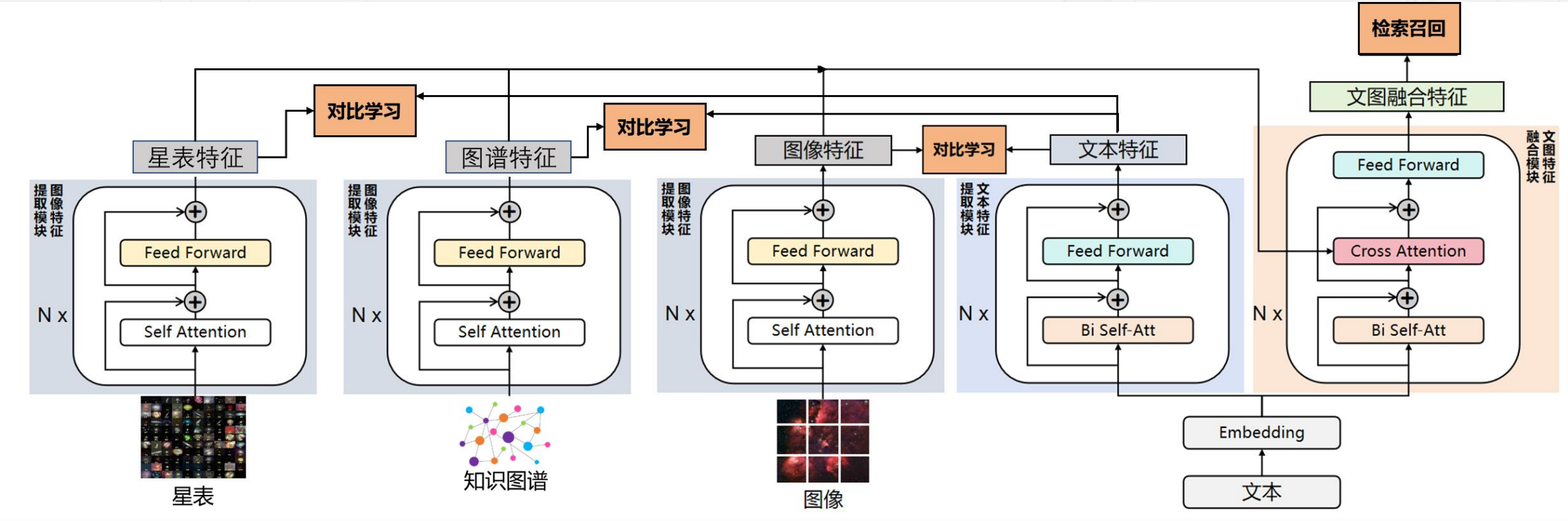
4、（进行中）天文文献类数据图文表征模型SFT工作。

5、研究针对天文领域数据的增强、自监督或者zero-shot等算法。

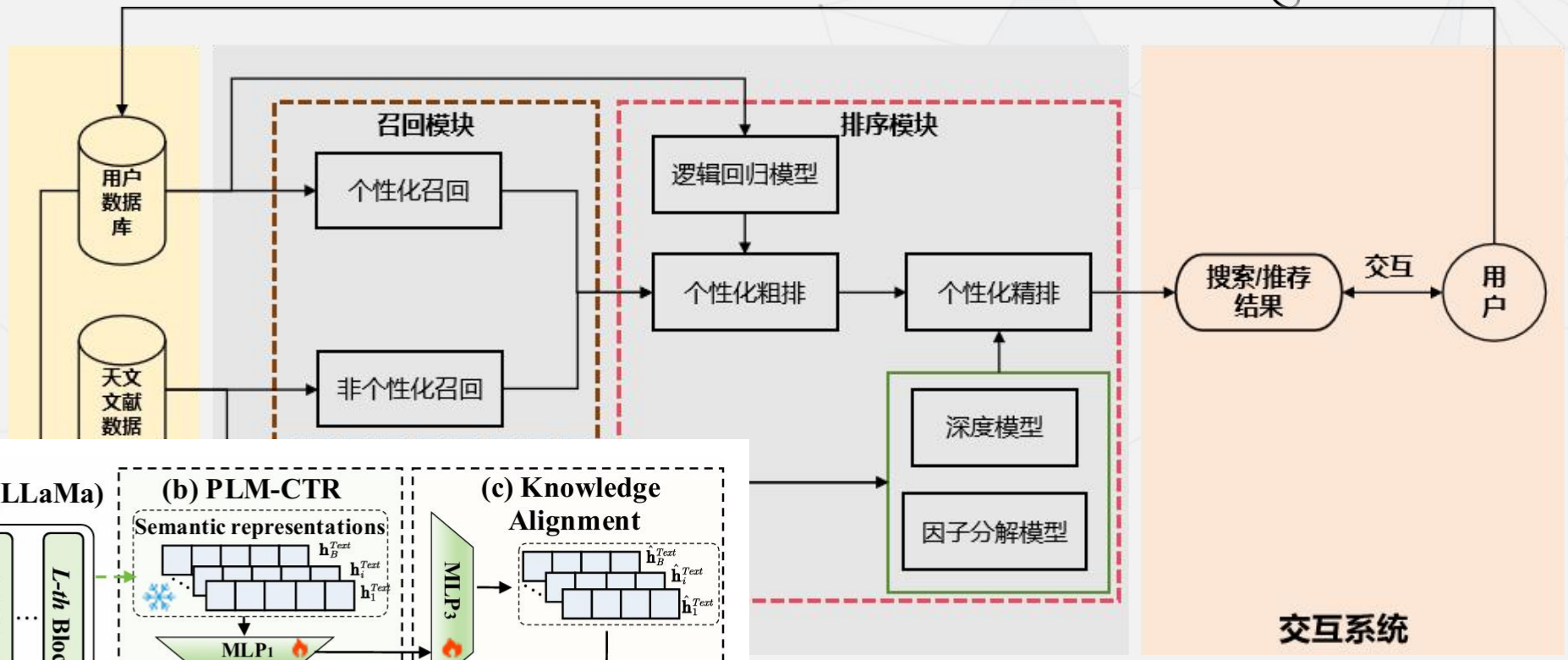
混合专家多模态融合召回

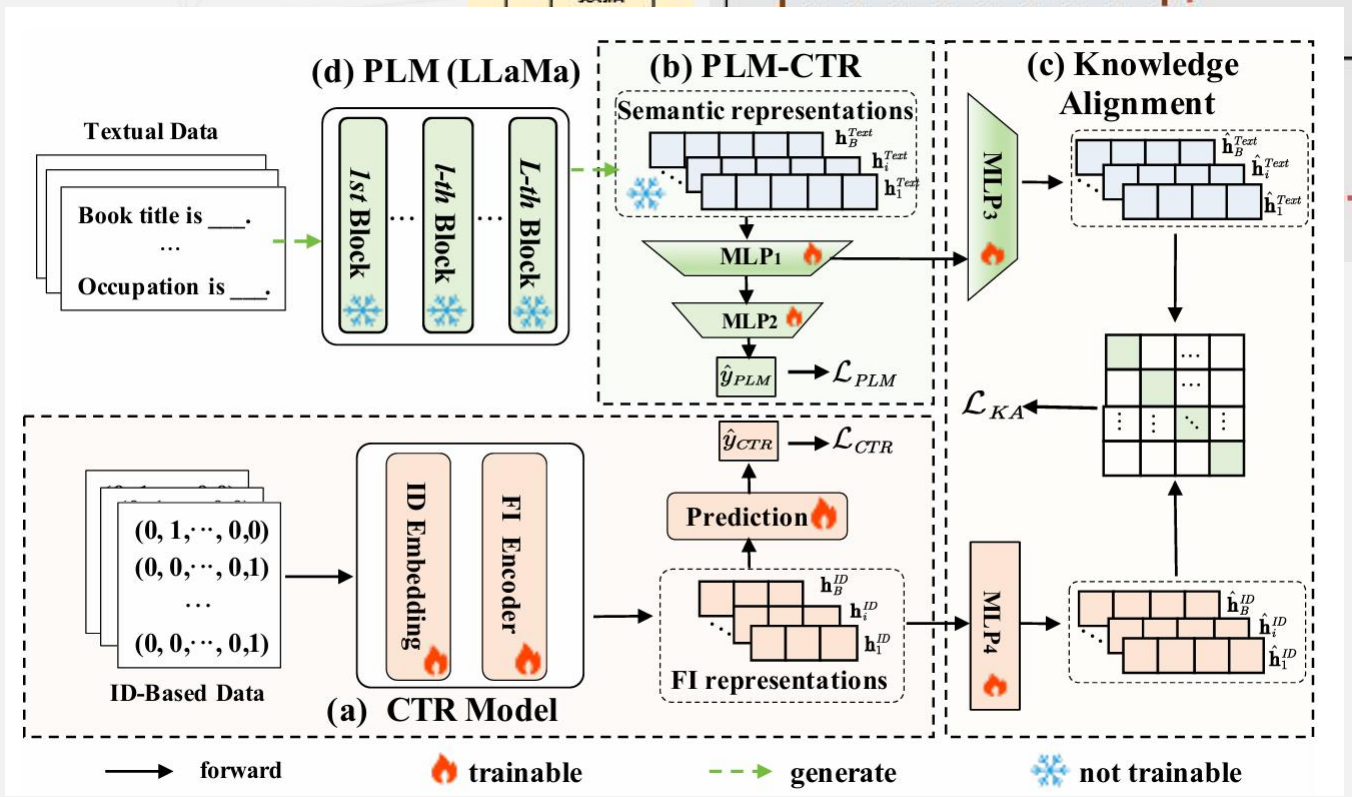
■ **目标：**

训练融合星表、知识图谱、图像、文本等多种模态的天文科学数据统一向量编码器，基于Bert类语言模型实现高效的多模态天文科学数据统一检索查询。



个性化重排推荐


用户数据库
天文文献数据
召回模块
个性化召回
非个性化召回
排序模块
逻辑回归模型
个性化粗排
个性化精排
深度模型
因子分解模型
搜索/推荐结果
交互
用户
交互系统


(d) PLM (LLaMa)
Textual Data
Book title is ___.
...
Occupation is ___.
1st Block
l-th Block
L-th Block
(b) PLM-CTR
Semantic representations
MLP1
MLP2
$\hat{y}_{PLM}$
$\mathcal{L}_{PLM}$
(c) Knowledge Alignment
MLP3
$\mathcal{L}_{KA}$
$\hat{y}_{CTR}$
$\mathcal{L}_{CTR}$
Prediction
(0, 1,⋯, 0,0)
(0, 0,⋯, 0,1)
...
(0, 0,⋯, 0,1)
ID-Based Data
ID Embedding
FI Encoder
FI representations
MLP4
(a) CTR Model
forward
trainable
generate
not trainable

目录

01 大数据时代的文献知识关联搜索

02 AI时代的天文数据知识关联搜索

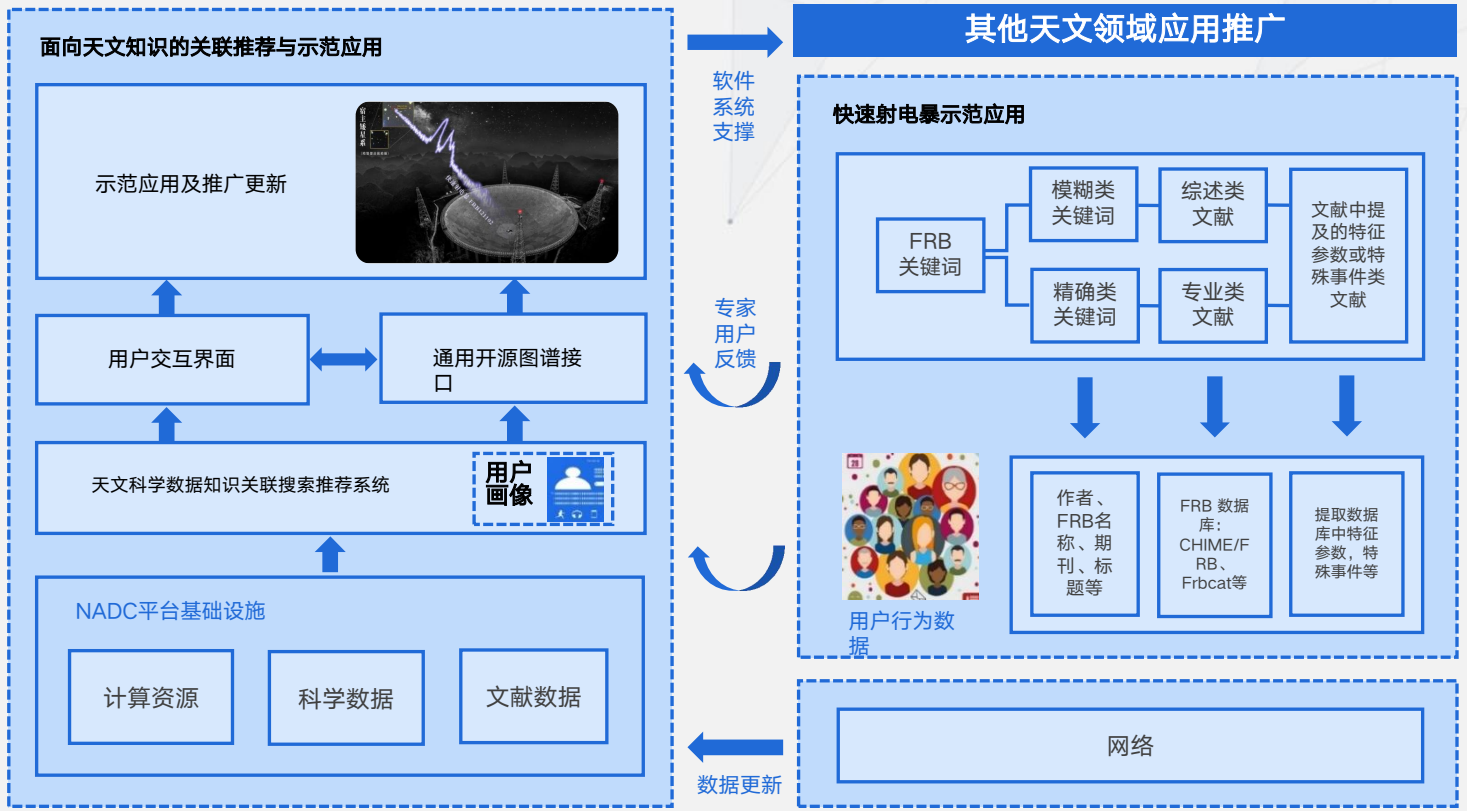
03 Introducing 天问1.0Beta



面向天文知识的关联推荐与示范应用

技术方案：

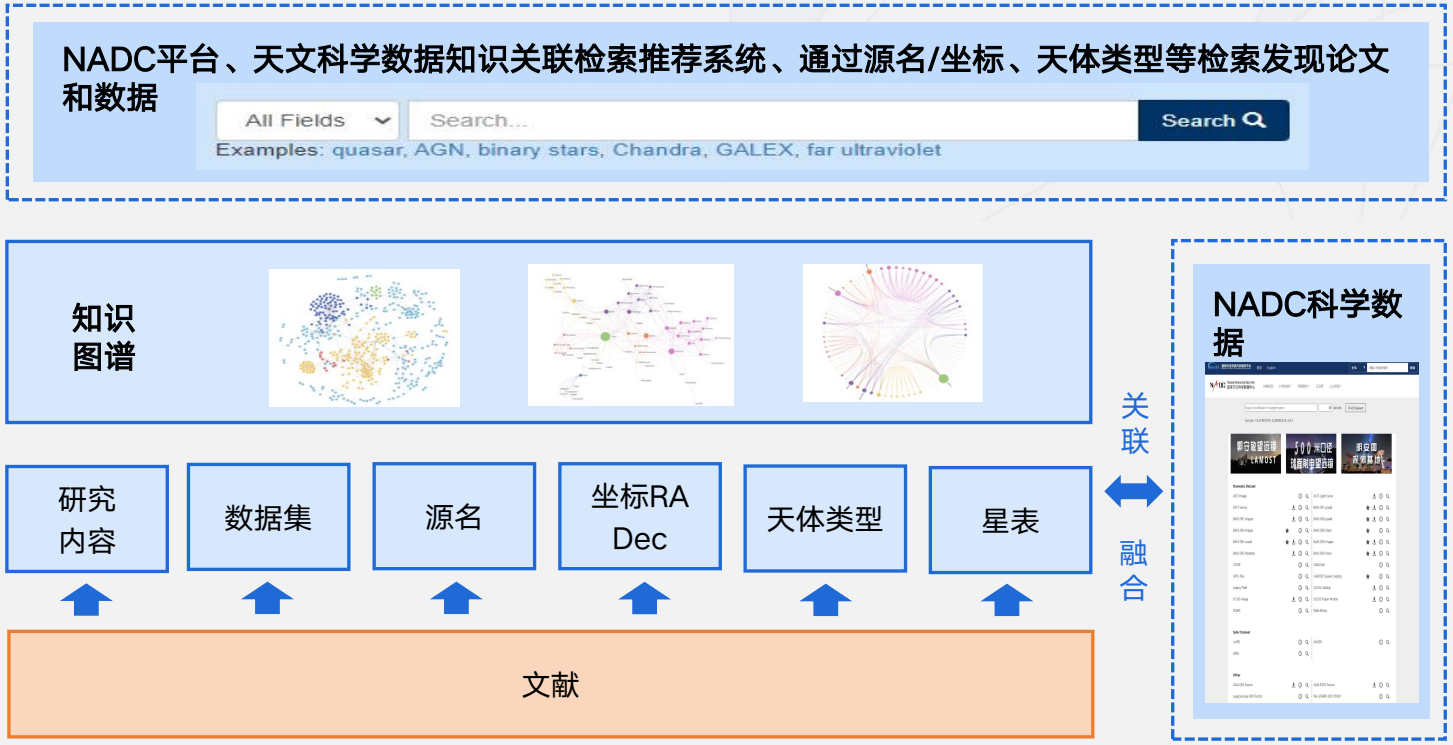
- 整合国家天文科学数据中心（NADC）的数据资源，将建立的天文知识关联融合推荐系统**部署在国家天文科学数据中心平台**，向天文领域的科研和公众用户开放测试。
- **以快速射电暴为系统示范案例**，进行知识建模、挖掘与推荐，推广并招揽合作专家参与，**探索反馈更新机制**，助力数据的标准化和开源众包共建。
- 基于快速射电暴示范应用，**总结设计通用接口和可交互的用户界面，**推广应用到新的研究领域，吸引相关领域专家参与，**实现数据众包收集和用户反馈机制**。



面向天文知识的关联推荐与示范应用

1. 系统在国家天文科学数据中心平台集成部署：

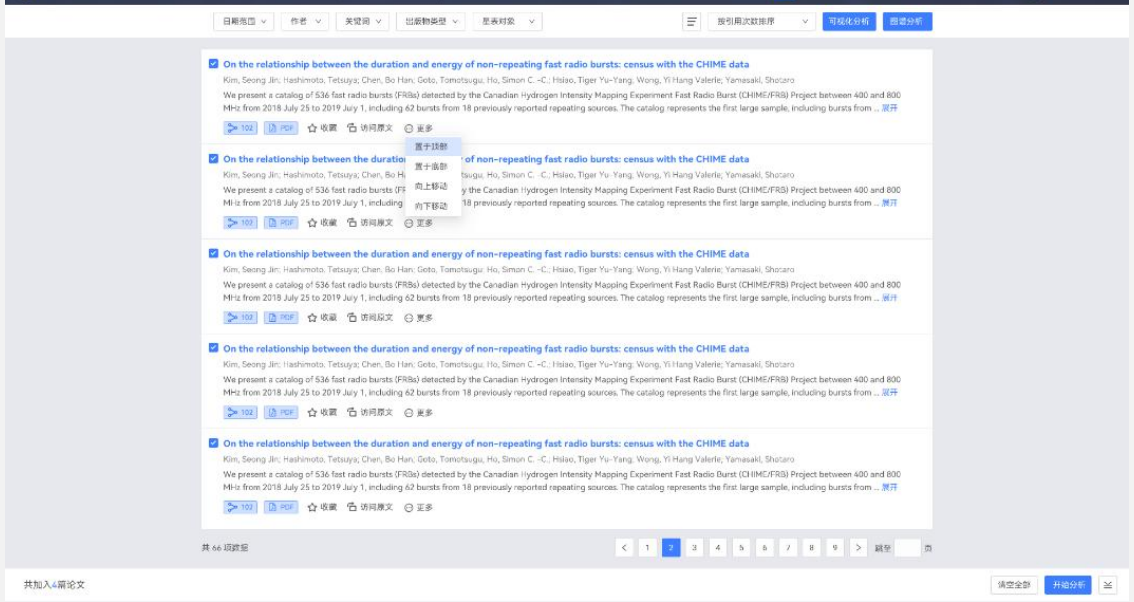
- 通过元数据以及坐标等信息，将NADC的数据资源与期刊文献中抽取构建的知识图谱进行关联融合，**建立文献和数据集、源名、坐标、天体类型、研究领域等之间的关联**
- 基于NADC平台部署检索推荐系统，能够通过源名/坐标、天体类型等检索获取多模态信息
- 围绕**快速射电暴领域**和**LAMOST数据集**开展示范应用



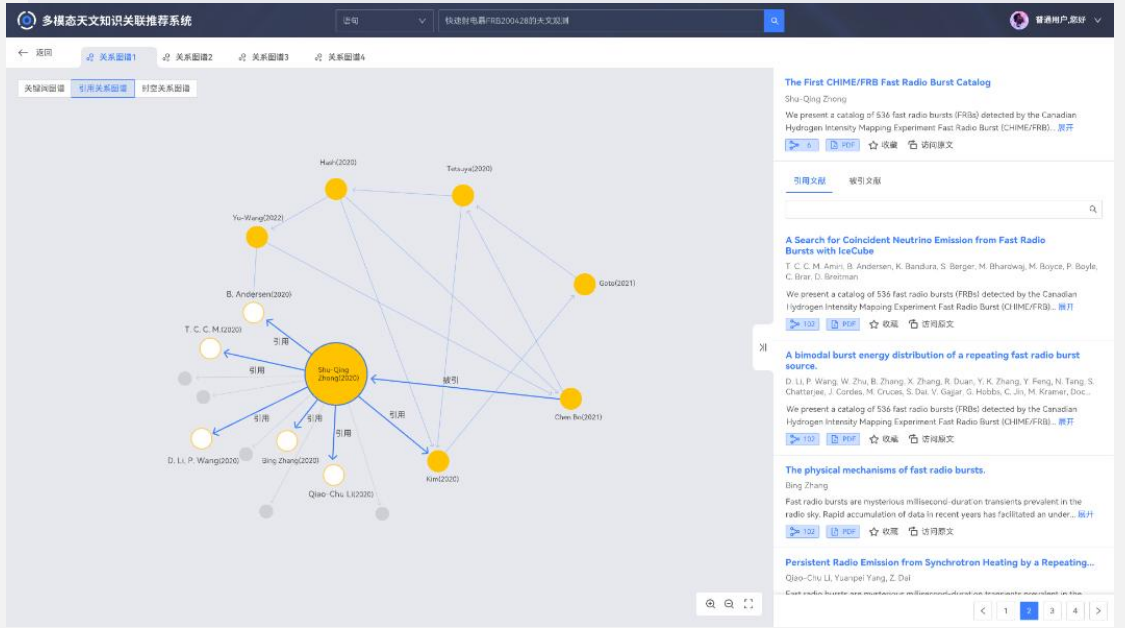
面向天文知识的关联推荐与示范应用

相关界面设计：

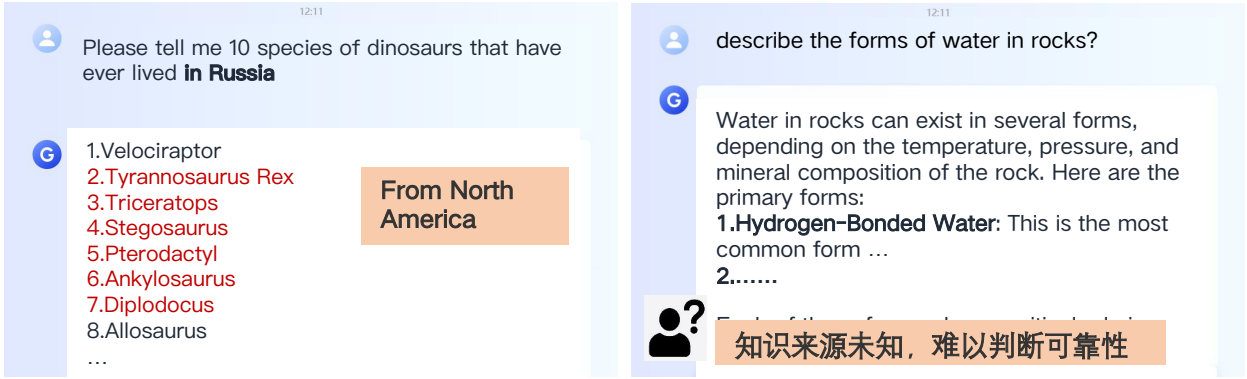






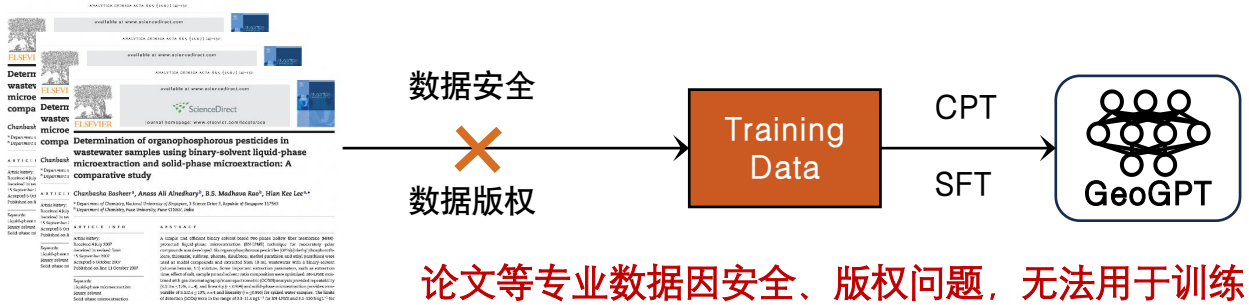


结构化数据与大模型结合：RAG



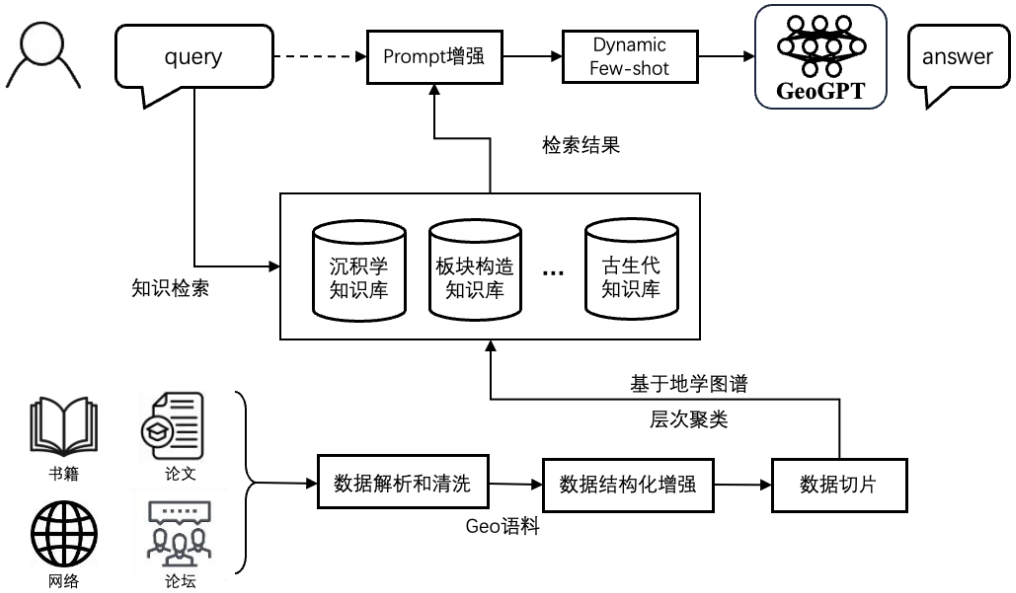
生成式模型存在幻觉，专业性不足

科研场景问答需要严谨的依据



论文等专业数据因安全、版权问题，无法用于训练

RAG技术方案



❑ 深时数字地球大科学计划（DDE）

RAG 问答增强评测

Base Model	0.531
GeoGPT	0.529
GeoGPT+RAG	0.666

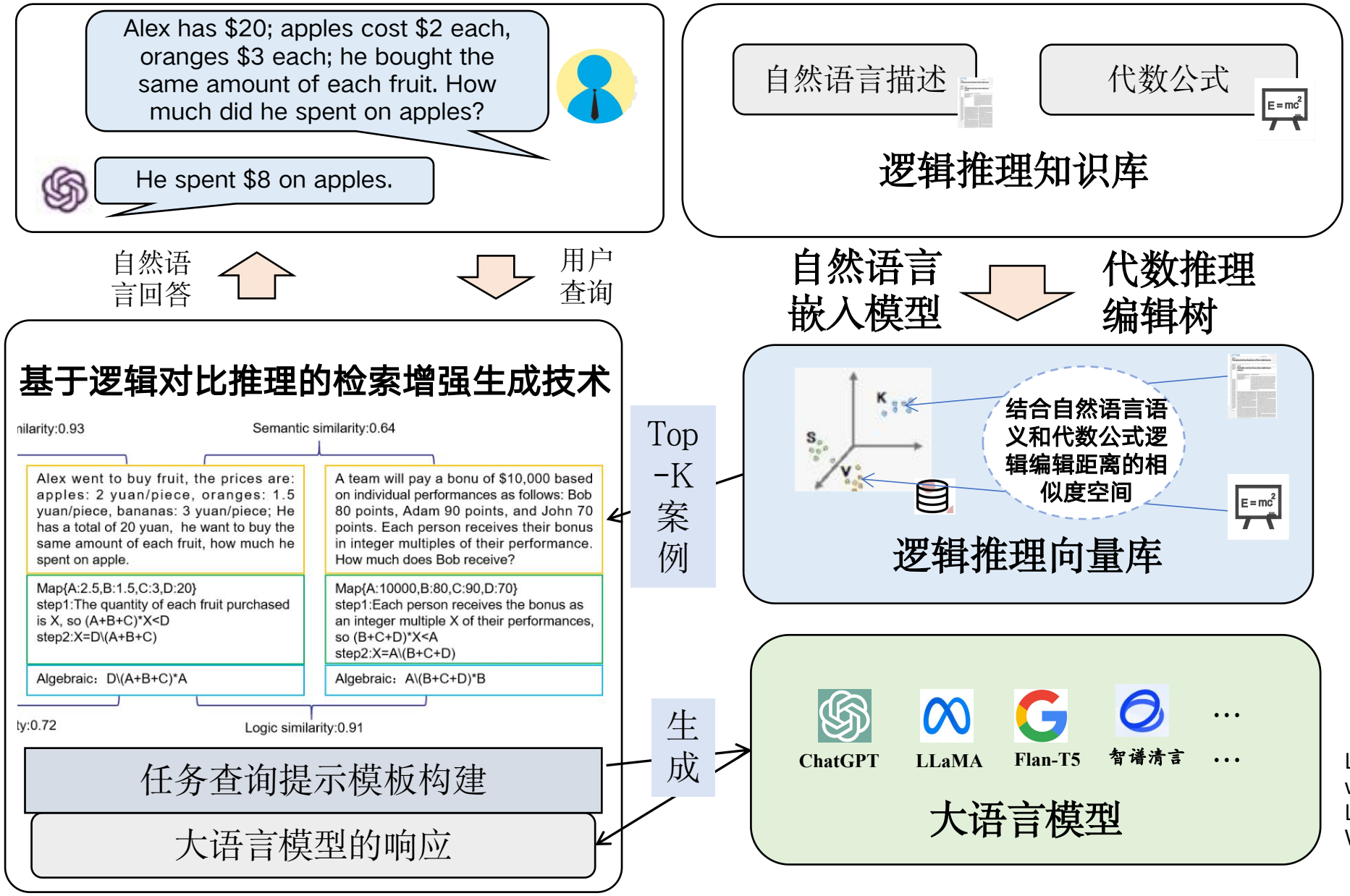
- 评测方法：基于公共库10W论文，利用大模型生成**938**题地学RAG测试集；
- 指标：回答召回率，即标准答案的论点在大模型回答中的比例；
- 结果：RAG 0.666 > No RAG 0.529，显著提升大模型回答的专业性。

RAG专家评测

	No RAG	公共库(10w)	个人库(100)
Bad Answer	7% (3)	0 (0)	0 (0)
General Answer	52% (22)	58% (7)	48% (20)
Good Answer	41% (17)	42% (5)	52% (22)

地质勘探领域专家评测，RAG的Good Answer的比例有提升，基本无Bad Answer

结构化知识与大模型结合：逻辑推理RAG



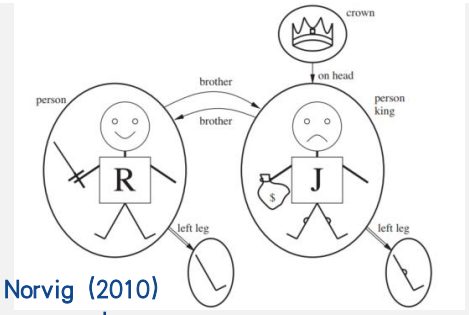
Logic Contrastive Reasoning with Lightweight Large Language Model for Math Word Problems, Manuscript

AI推理的两大流派

符号主义：逻辑推理

- 早期的小世界（逻辑驱动）
- 曾经辉煌的专家系统（知识库）
- 规则无法自主学习
- 推理能覆盖的范围极其有限
- 无法表征现实的不确定性

$$\neg Brother(LeftLeg(Richard), John)$$
$$Brother(Richard, John) \wedge Brother(John, Richard)$$
$$King(Richard) \vee King(John)$$
$$\neg King(Richard) \Rightarrow King(John).$$



Russell and Norvig (2010)
AI a modern approach

联结主义：统计推理

- 灵活的模型表征(神经网络)
- 高效的学习算法(反向传播)
- 超大规模的数据和模型
- 所得黑箱模型的不可理解
- 幻觉使得推理无法100%准确
- 对数据的依赖性

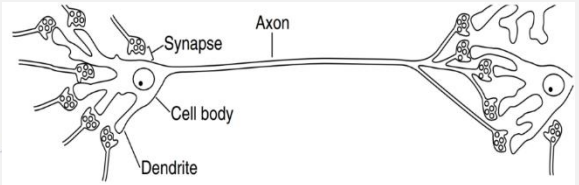
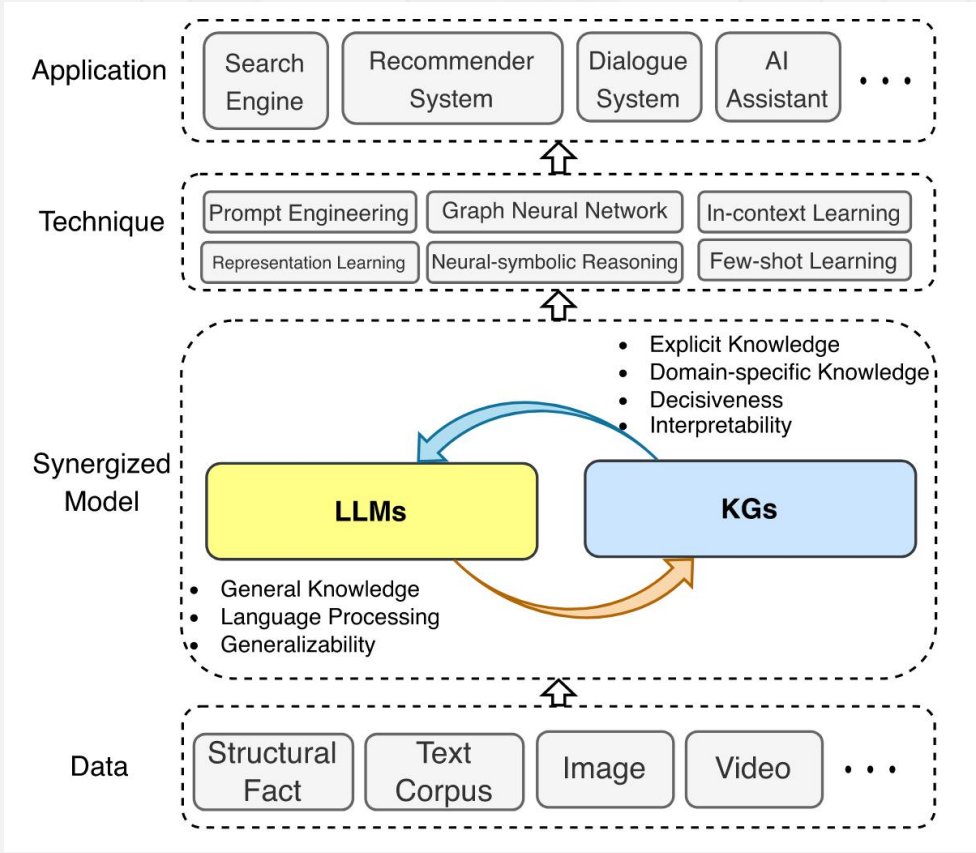


Figure 1.2 A simplified diagram of a neuron, from Crick and Asanuma (1986).

Luger (2009) Artificial Intelligence

?



Unifying Large Language Models and Knowledge Graphs: A Roadmap. *IEEE TKDE 2024*