

云计算、知识图谱、大模型 与天文科学数据搜索推荐

2023.03.24 >>>

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



《新一代人工智能发展规划》初步建立了群体智能的理论与技术体系，指明了群体智能在**人机物三元空间**感知、组织运行、协同演化等核心技术上的发展方向。围绕实验室**数据+知识双轮驱动**的布局，团队拟在群体层面重点突破三元空间数据汇聚交互、跨个体跨代际知识涌现和双轮驱动协同决策三大技术方向

感知

研究群智时空的分布式多模态感知、安全交互和结构组织原理，设计符合人类社会组织规律和法律规范的数据隐私、算法伦理的开放生态体系，实现群智时空的**高效组织、沉浸式交互和可信运行**；

表征

研究通用群体多模态知识表征，探索能够打通微观数据驱动到宏观知识抽象的跨尺度涌现机理，实现持续性回环评估与协同进化技术，从而实现**知识的跨个体共识与跨代际演进**；

决策

基于群体不断产生的大量数据和知识，研究多个智能体在复杂多变环境下的鲁棒计算以及协同博弈决策的问题，突破基于预训练模型、逻辑规则推理和群体共识的**双轮驱动融合推理决策**。

团队介绍



之江实验室
ZHEJIANG LAB

人工智能研究院
RESEARCH INSTITUTE OF
ARTIFICIAL INTELLIGENCE



严笑然博士, 之江实验室研究专家, 国家重点研发计划青年首席科学家



马萧博士, 之江实验室工程专家, 数智化康复装备浙江省工程研究中心专家委员



薛均晓博士, 之江实验室人工智能研究院研究专家, 擅长计算机图形学与多智能体仿真



姬朋立博士, 之江高级研究专员, 国青基金项目负责人, 擅长组合优化



周丽博士, 之江高级研究专员, 国青基金项目负责人, 擅长统计分析



陆亚飞博士, 浙江大学博士、之江实验室高级研究专员, 擅长复杂系统与控制



李清明博士, 之江高级研究专员, 从事联邦激励机制研究



张睿博士, 之江博士后, 擅长文本和结构化数据的表征学习



顾剑波硕士, 之江高级工程专员, 多年的业界算法工程落地经验



厉燕硕士, 之江高级工程专员, 多年数据分析和数据工程从业经验



张德文硕士, 之江高级工程专员, 多年界面交互&视觉设计经验



刘洋硕士, 之江高级工程专员, 多年大数据及分布式计算框架经验



陆矜菁硕士, 之江工程专员, 负责数据开发、部分后端开发



陈一家硕士, 之江工程专员, 擅长数据管理和数据分析工作



王刚硕士, 之江工程专员, 多年隐私计算与联邦学习研究经验



孙设, 硕士, 之江工程专员, 有联邦学习, 隐私计算领域的研究和工程实践经验



滕皓硕士, 之科后端工程师, 技术栈广泛, 算法开发经验丰富



侯炜华硕士, 之科工程师, 擅长大数据及后端服务开发



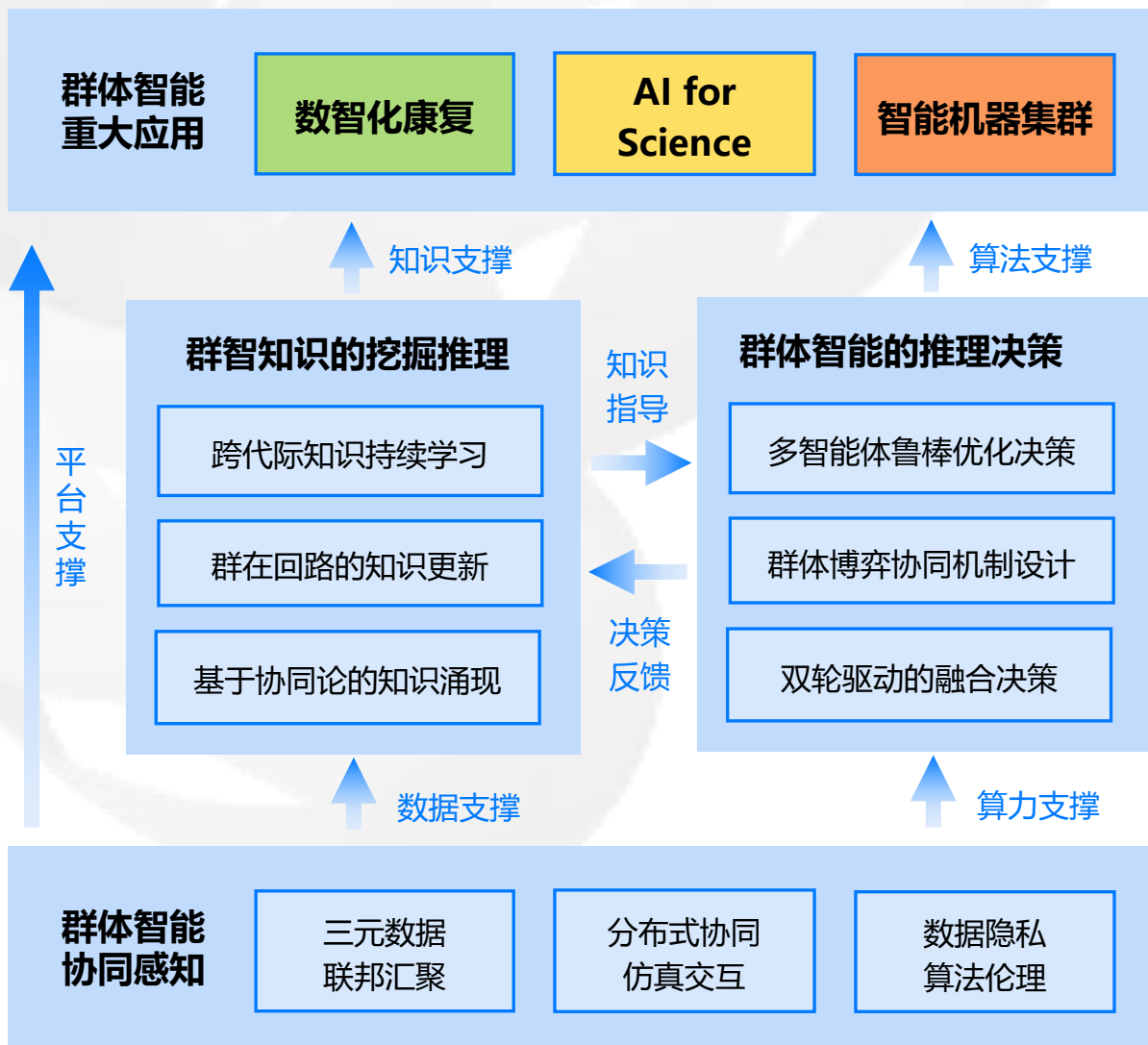
刁瑒, 之科工程师, 擅长数据可视化、OA系统开发

科研布局：已立项目和平台



之江实验室
ZHEJIANG LAB

人工智能研究院
RESEARCH INSTITUTE OF
ARTIFICIAL INTELLIGENCE



GJ实验室重大任务：跨媒体类人知识推理和持续学习

浦江实验室科研合同：类人智能的知识表达与计算

省级工程中心：数智化康复装备浙江省工程研究中心

之江实验室青年项目：基于多模态知识图谱构建和学习的研究

之江实验室青年项目：分布式图数据的价值评估与组合

之江实验室揭榜挂帅课题3：跨机构隐私保护机器学习建模平台开发和算法模型研究

之江实验室青年项目：基于概率软逻辑的主动学习方法研究

国家重点研发计划：多模态天文科学数据知识关联推荐系统

国家科学数据中心：国家天文科学数据中心之江实验室分中心

浙江省哲学社会科学实验室：智能社会治理实验室

之江实验室探索项目：交互式软规则动态知识图谱

部省联动重点研发子课题：机密计算微体系结构与可信执行环境

之江实验室装备项目：智能XX集群

请帮我们选出系统名称



1. 天问 (Astro Inquiry System)
2. 巡天 (AstroSearcher)
3. 洞天 (SkyInsight)
4. 天网 (Skynet)
5. 星河 (Galaxy)

目录

01 Big data and cloud computing

02 Cloud native computing and scientific reproducibility

03 Knowledge graphs and scientific search engine for Astrophysics

04 Large language models and the future of scientific search





CADRE

Collaborative Archive & Data Research Environment

Overview of CADRE project

Xiaoran Yan
yan30@iu.edu

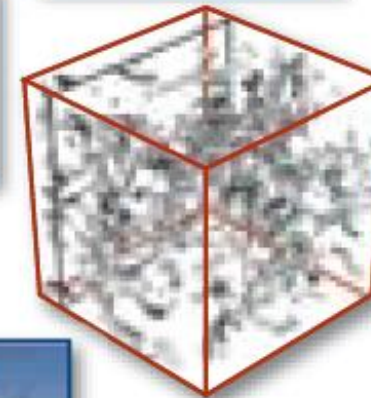


Science Paradigms

- Thousand years ago:
science was **empirical**
describing natural phenomena
- Last few hundred years:
theoretical branch
using models, generalizations
- Last few decades:
a **computational** branch
simulating complex phenomena
- Today: **data exploration** (eScience)
unify theory, experiment, and simulation
 - Data captured by instruments
or generated by simulator
 - Processed by software
 - Information/knowledge stored in computer
 - Scientist analyzes database/files
using data management and statistics



$$\left(\frac{\dot{a}}{a}\right)^2 = \frac{4\pi G\rho}{3} - K\frac{c^2}{a^2}$$

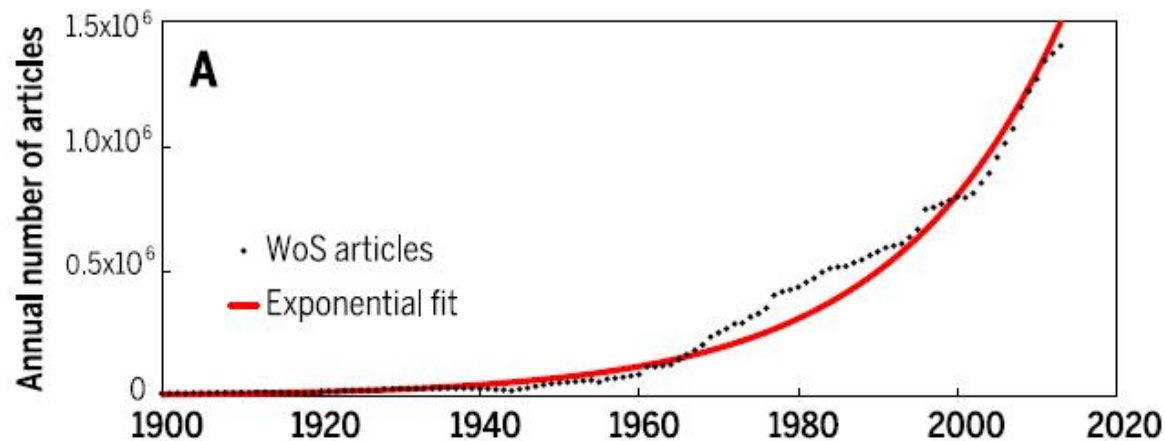
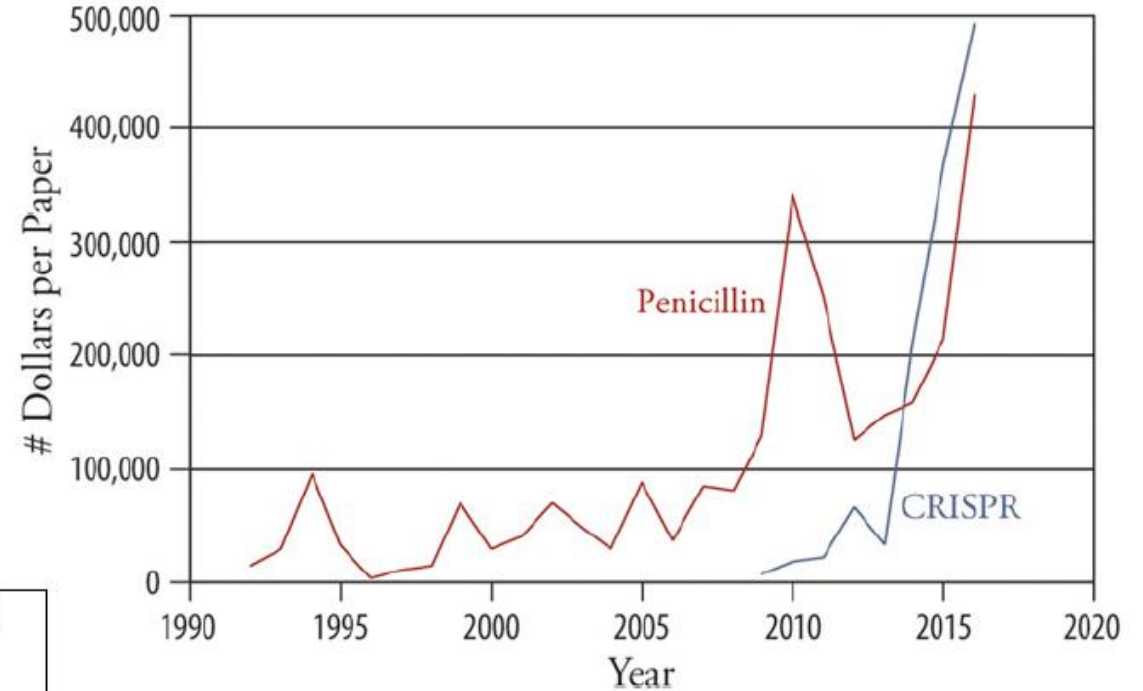
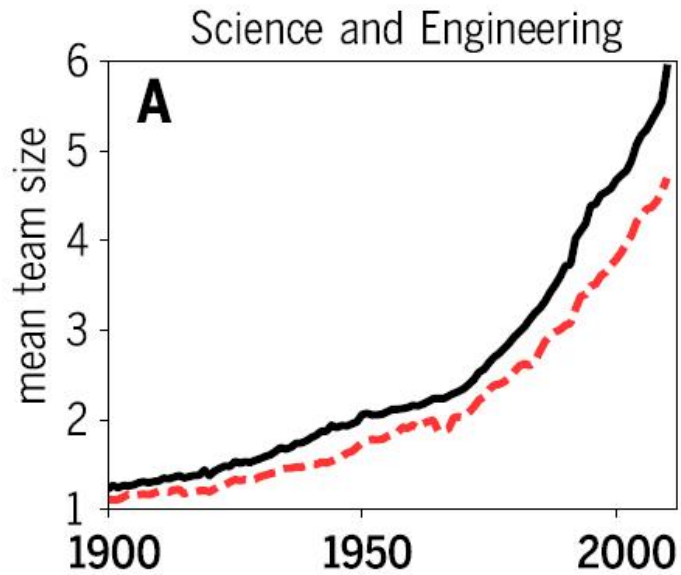


Tony Hey,
2009. *The
Fourth
Paradigm:
Data-
Intensive
Scientific
Discovery*



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

Research more, search less

Try a topic, author, journal, etc. or any combination of these



Publications

210,365,701

Coming soon



Authors

254,317,172

Learn more



Fields of Study

229,763

Learn more



Conferences

4,341

Learn more



Journals

48,659

Learn more



Institutions

25,439

Learn more



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



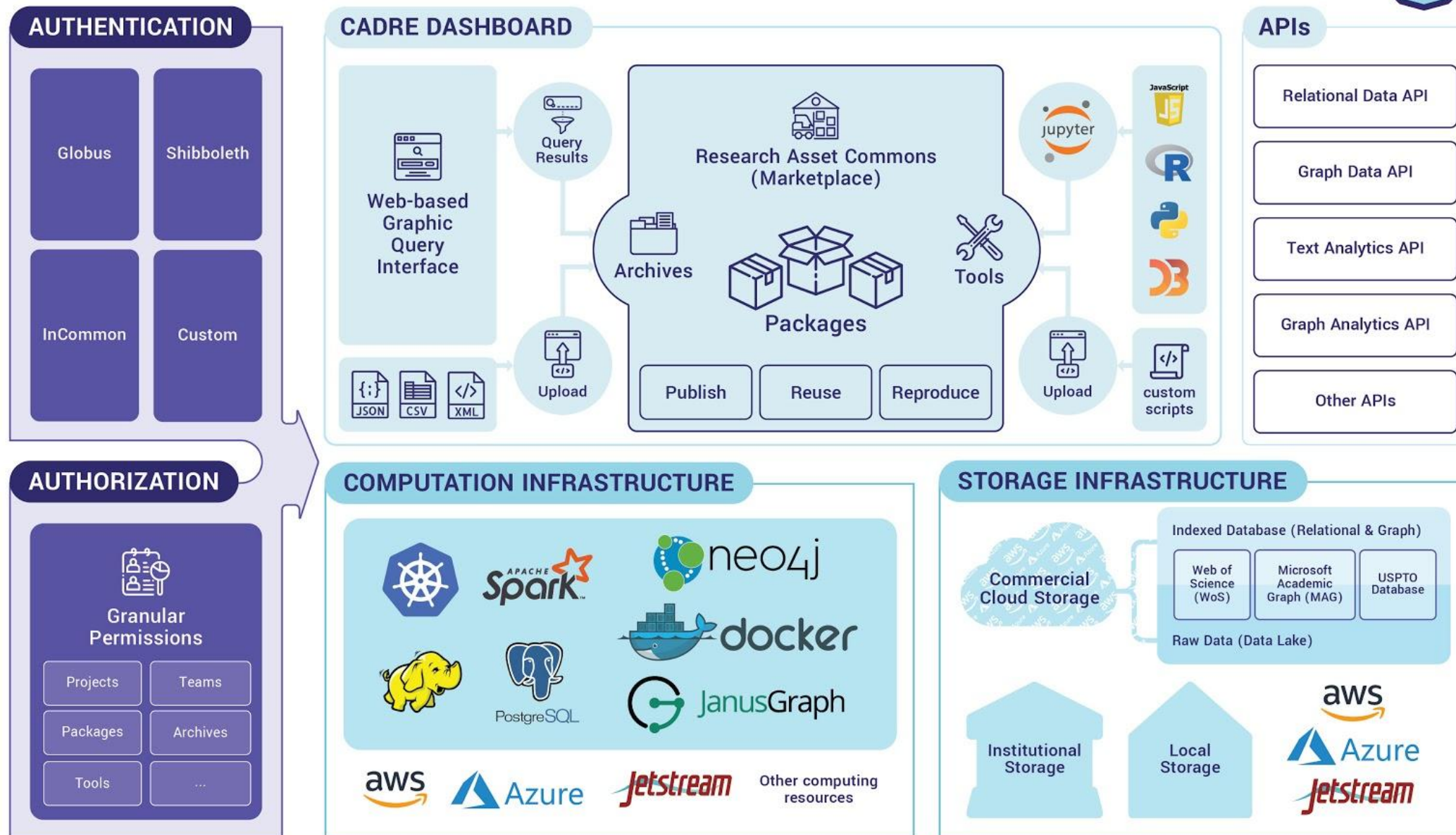
BTAA staff presented the solution to the Library Directors



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



Collaborative Archive & Data Research Environment (CADRE)





This project was made possible in part by the Institute of
Museum and Library Services LG-70-18-0202.

**New member: University
of Toronto Libraries!**

UNIVERSITY OF IOWA LIBRARIES
UNIVERSITY OF MICHIGAN LIBRARIES
MICHIGAN STATE UNIVERSITY LIBRARIES
UNIVERSITY OF MINNESOTA LIBRARIES
UNIVERSITY OF MARYLAND LIBRARIES
OHIO STATE UNIVERSITY LIBRARIES
PENN STATE UNIVERSITY LIBRARIES
PURDUE UNIVERSITY LIBRARIES
RUTGERS UNIVERSITY LIBRARIES
HEALTHPARTNERS INSTITUTE
PERVASIVE TECHNOLOGY INSTITUTE
MIDWEST BIG DATA HUB
SOUTH BIG DATA HUB
WEST BIG DATA HUB
MICROSOFT RESEARCH
WEB OF SCIENCE GROUP



CADRE Project Leadership



Jamie Wittenberg
Proj. Director
University of
Colorado, Boulder,
Library

Patricia Mabry
Co-Proj.
Director
HealthPartners
Institute

Valentin Pentchev
Co-Proj. Director
Indiana University
Network Science
Institute (IUNI)

Xiaoran Yan
Co-Proj. Director
AI research
Institute,
Zhejiang Lab

**Robert Van
Rennes**
Co-Proj. Director
Big Ten Academic
Alliance (BTAA)



Lessons from other fields

<u>Data Phase</u>	<u>Astronomy</u>	<u>Twitter</u>	<u>Genomics</u>
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

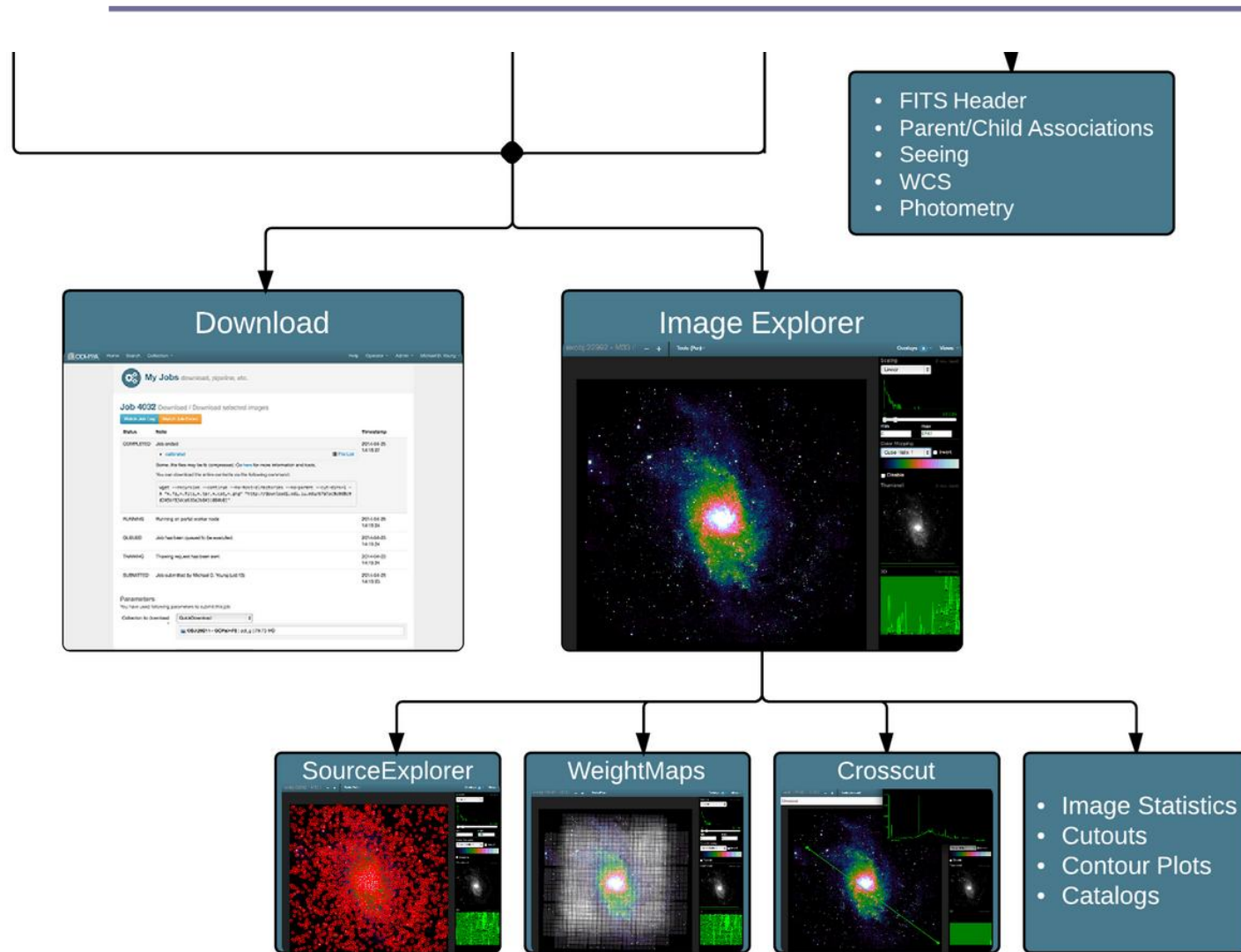
Variety



1ZB = 1K EB = (1K)² PB

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

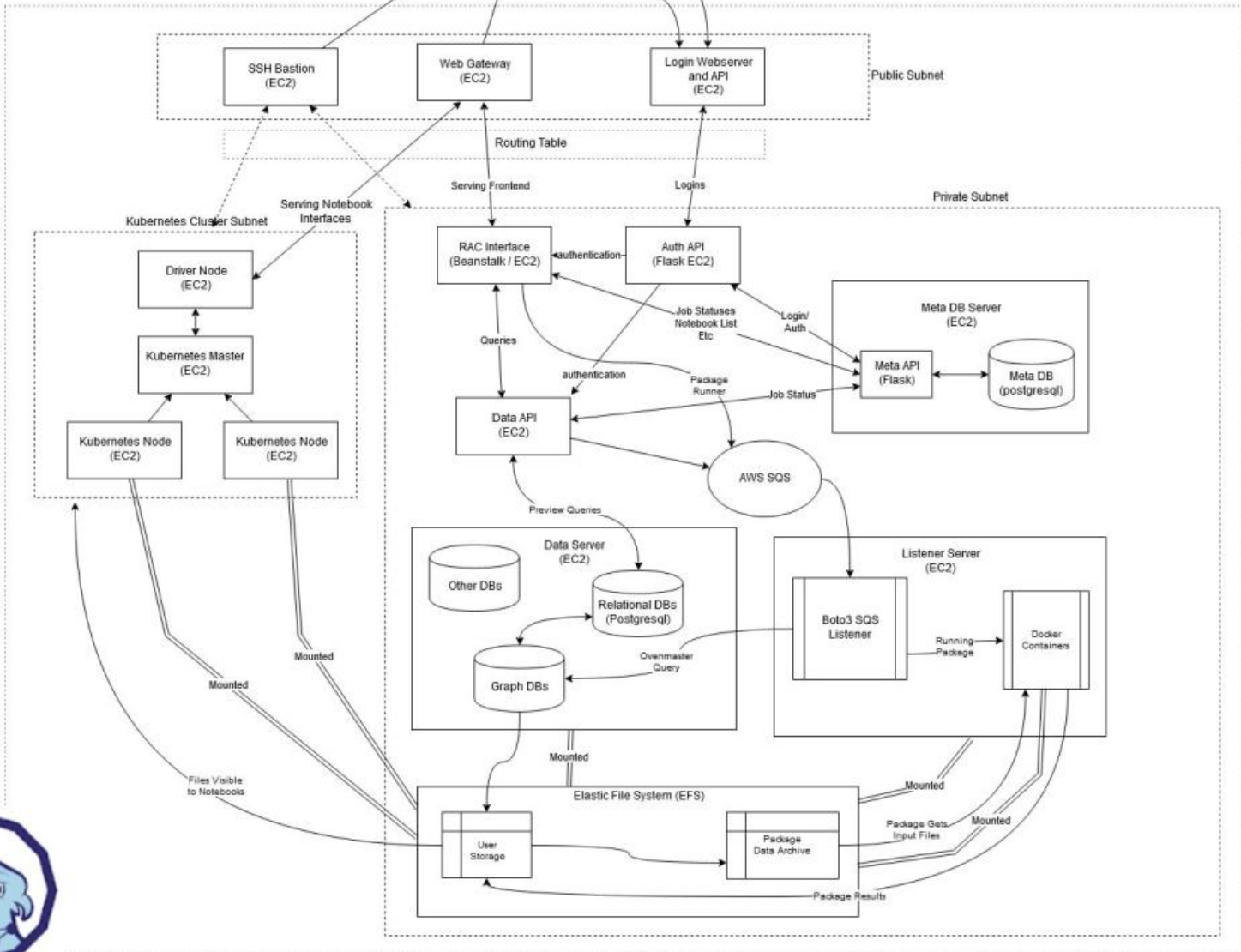
Volume: data centric



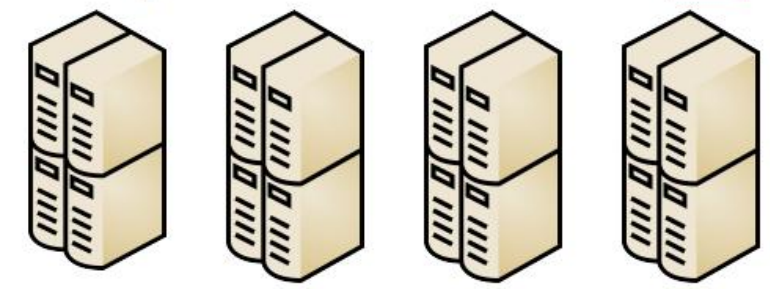
The [ODI Pipeline, Portal, and Archive \(ODI-PPA\)](#) is a web science gateway for an imager installed on the WIYN 3,5m telescope at Kitt Peak (Tucson, AZ)

50 terabytes per year





8 node Spark Cluster



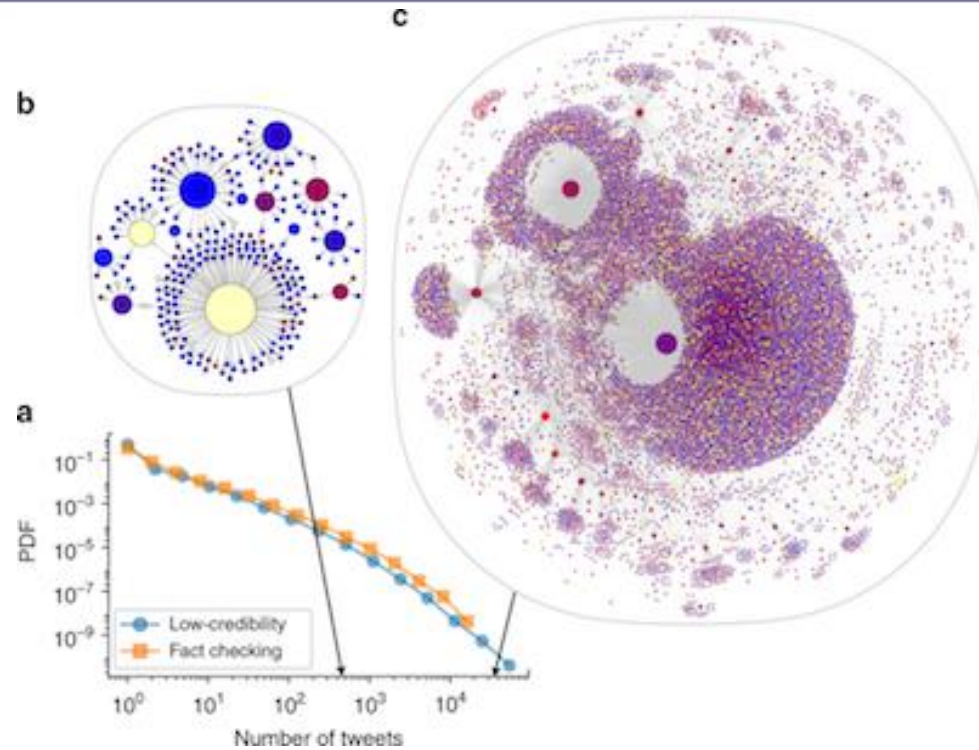
Large Memory compute node



Ability to expand to IU's High Performance and High Throughput systems



Velocity: live feeds and communities



Samples 10% of all tweets over 6 years
Active third party contributions

<http://osome.iuni.iu.edu/>

OSoMe Friends are unaffiliated, third party projects utilizing our APIs to do things that we think are useful.

- Hoaxy**: Visualize the spread of claims and fact checking.
- Botometer**: Check how bot-like a Twitter user behaves.
- Fakey**: Play this game to learn to recognize fake news on your social feed.
- Trends**: Compare when memes gain and lose popularity.
- Networks**: Explore who is discussing a meme and what memes are related.
- Maps**: Examine where people are talking about a meme over time.
- Movies**: Generate movies of how conversations about a meme evolve over time.
- EchoDemo**: Simulation demonstrating how two basic mechanisms of social media can lead to polarized social networks.
- Bot Electioneering Volume**: Visualizing the activity of likely bots on Twitter around the 2018 US midterm elections.
- API**: Query our data for your own analysis.
- Enhanced Data**: Enhanced data is available to all Indiana University students and faculty.
- Botson**: Chrome extension to detect and block twitter bots in your newsfeed.
- Probobot**: Twitter bot highlighting accounts with high bot scores.
- StattoBot**: Tweet this bot with another @account and it will reply with stats & bot score.



Variety: open collaborations

The image displays the Brain Life web interface, which is a platform for managing neuroimaging data and processes. The interface is divided into several sections:

- Header:** "Brain Life" logo, "Documentation", "New", and user profile "Soichi Hayashi".
- Navigation:** "Detail", "Datasets", "Processes", "Pipelines", "Publications".
- Process List:** A list of 1043 processes, with "Running (45)" highlighted. The first process is "rule:Data Normalize subject:153429" by Lindsey Kitchell, which is "FINISHED 2 hours ago" with the message "Service completed successfully".
- Process Details:** Shows "Staging Input" and "Output" for subject 153429, including a "Raw Output" section.
- Running Process:** A "RUNNING" process "Fitting the tensor model..." is shown with parameters: eddyCorrect: -1, rotateEvecsWithRx: false, rotateEvecsWithCanXform: false.
- 3D Visualization:** A 3D brain scan visualization showing fiber-like structures in yellow and green, with a control panel on the right for adjusting view and data.



目录

01 Big data and cloud computing

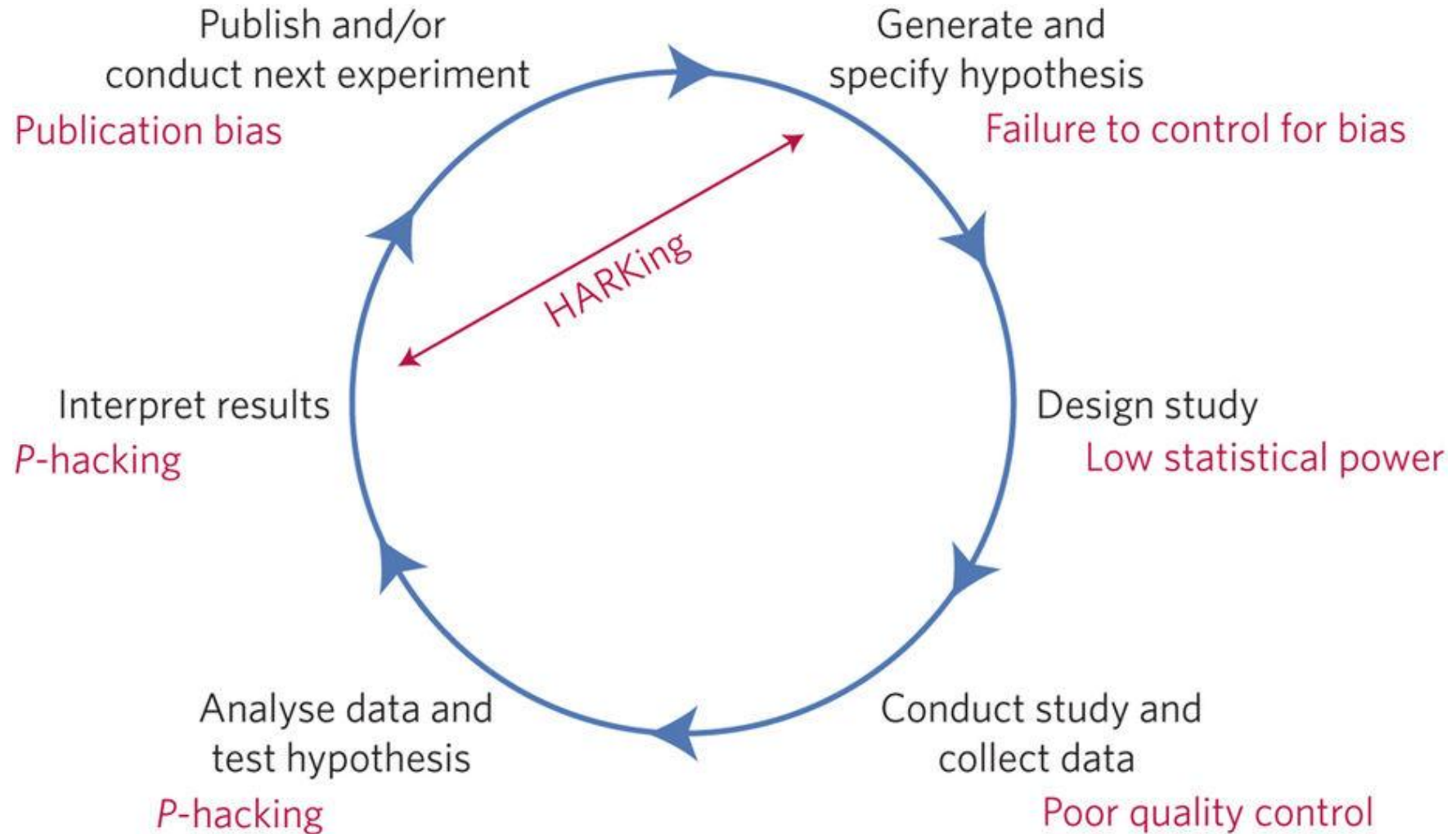
02 Cloud native computing and scientific reproducibility

03 Knowledge graphs and scientific search engine for Astrophysics

04 Large language models and the future of scientific search



Veracity: the reproducibility “Crisis”



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



Science Policy: Journals



“Data and materials availability: All data necessary to understand, assess, and extend the conclusions of the manuscript must be available to any reader of Science. After publication, all reasonable requests for materials must be fulfilled.”

There are still some exceptions-----

Any restrictions on the availability of data or materials, including fees and original data obtained from other sources (Materials Transfer Agreements), must be disclosed to the editors upon submission. “



Barbara R. Jasny
Deputy Editor,
Emeritus at
Science magazine



Science Policy: Funding agency



National Institutes of Health
Turning Discovery Into Health

 Search

[NIH Employee Intranet](#) | [Staff Directory](#) | [En Español](#)

[Health Information](#)

[Grants & Funding](#)

[News & Events](#)

[Research & Training](#)

[Institutes at NIH](#)

[About NIH](#)

[NIH Home](#) > [Research & Training](#)

RIGOR AND REPRODUCIBILITY

Rigor and Reproducibility

[Principles and Guidelines](#)

[Publications](#)

[Training](#)

[Meetings and Workshops](#)

[Expanded Guidelines](#)

[Application Instructions](#)

Rigor and Reproducibility

Two of the cornerstones of science advancement are rigor in designing and performing scientific research and the ability to reproduce biomedical research findings. The application of rigor ensures robust and unbiased experimental design, methodology, analysis, interpretation, and reporting of results. When a result can be reproduced by multiple scientists, it validates the original results and readiness to progress to the next phase of research. This is especially important for clinical trials in humans, which are built on studies that have demonstrated a particular effect or outcome.



Johns Hopkins University students in a laboratory. (Johns Hopkins University Photo)

Email Updates

To sign up for updates please enter your e-mail address.

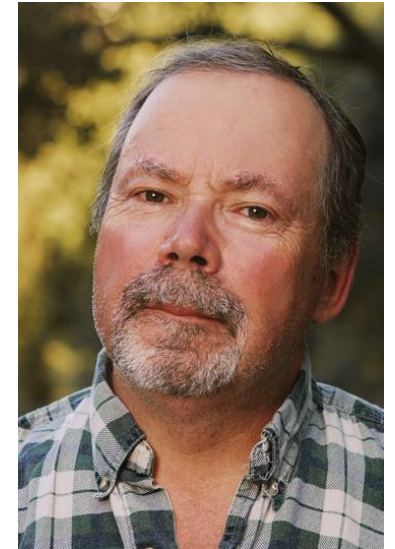
Related Links

[Letter from Dr. Stephen I. Katz: An Update on the NIH Initiative to Enhance Research Rigor and Reproducibility](#)

[Nature commentary on the Promise and Peril of Chemical Probes](#)

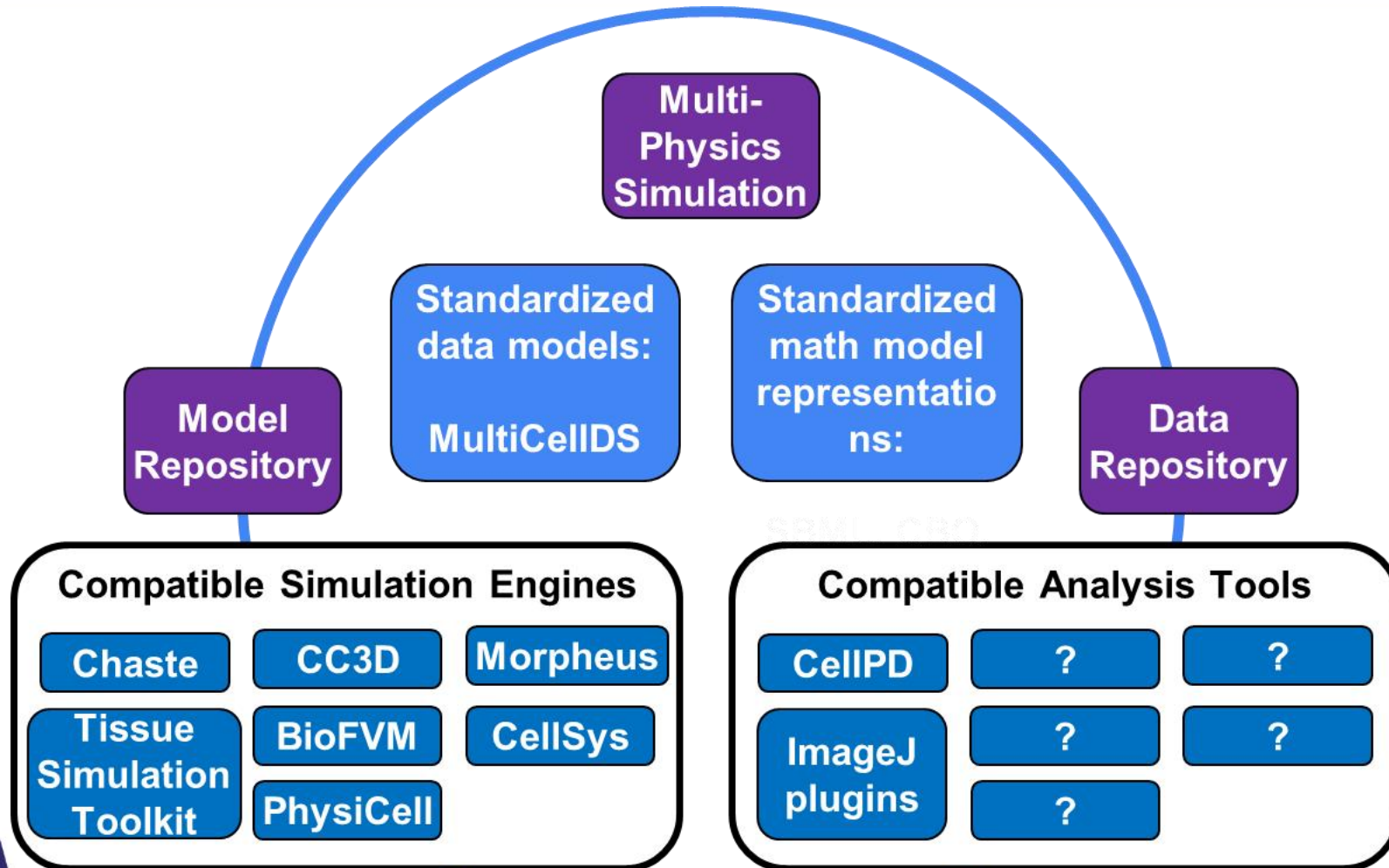
Contact Us

Please send email to NIHReprodEfforts@od.nih.gov.



Philip E. Bourne
Former NIH
Associate Director for
Data Science.
Led the Big Data to
Knowledge (BD2K)
initiative

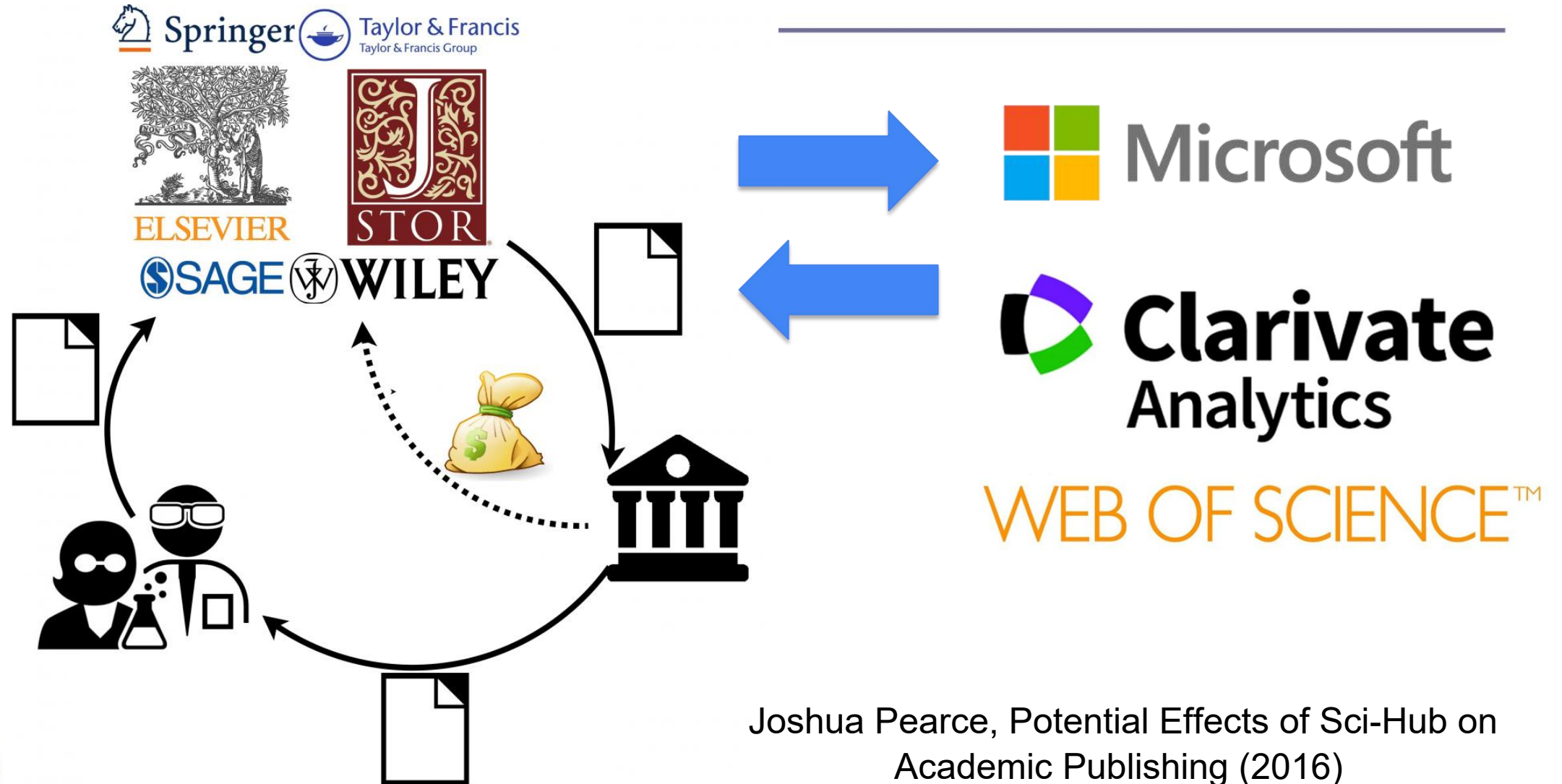
Science Policy: research communities



Paul Macklin
Associate
Professor of
Intelligent
Systems
Engineering at
Indiana
University



A collective effort



Spectrum of Reproducibility



Computational  Statistical  Empirical

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



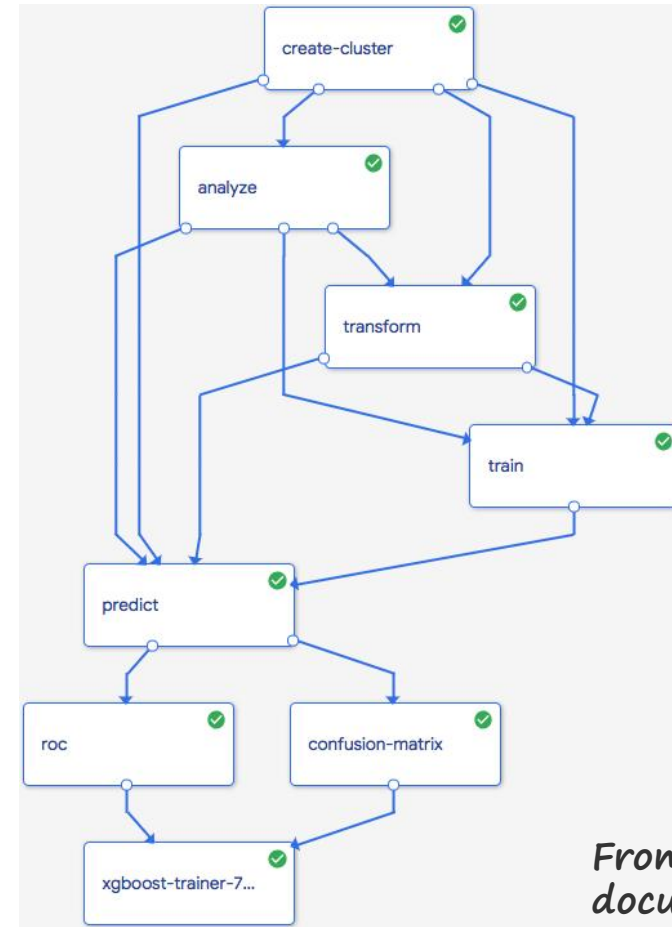
My personal story

Gephi plugin

- A new resolution
- The dream
- The nightmare

Researcher's real life

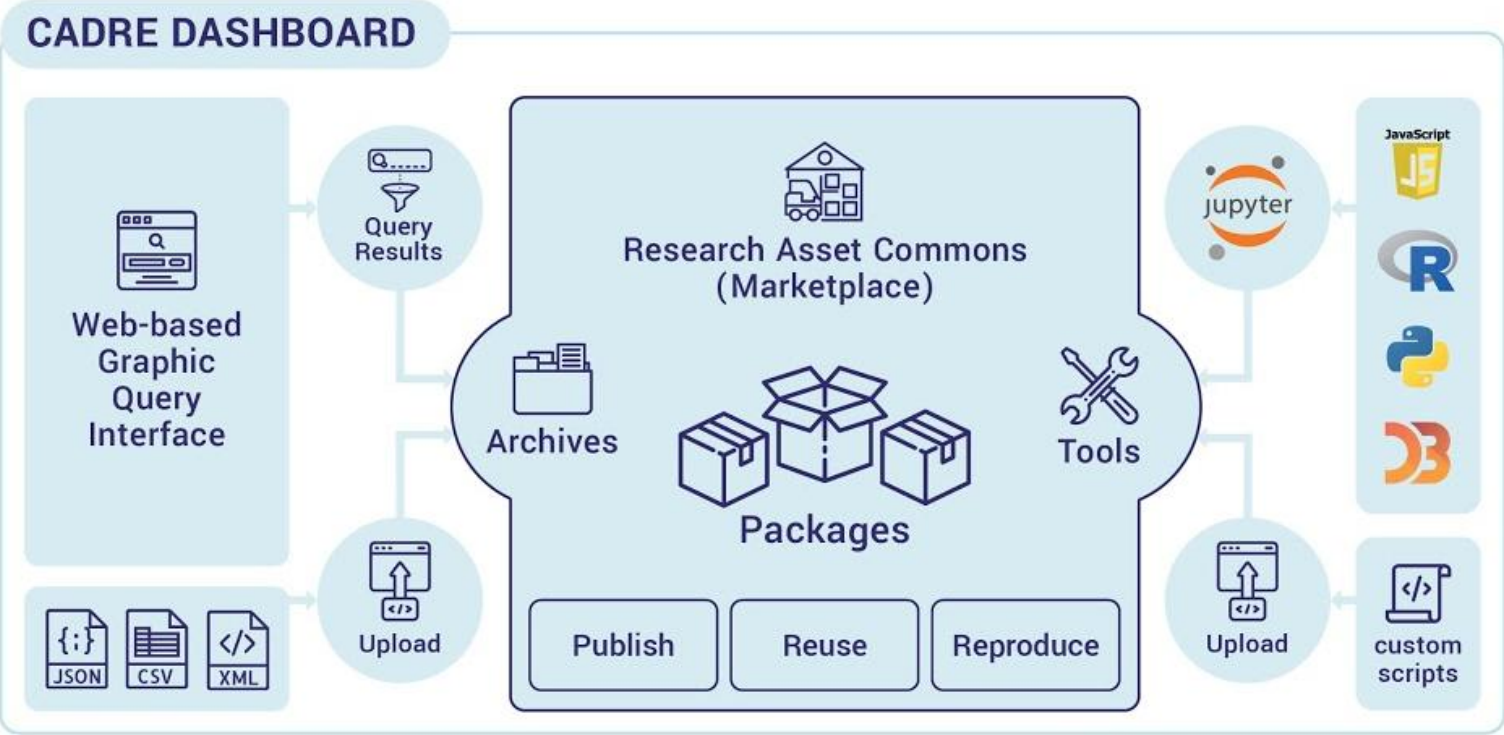
- 80% data cleaning?
- 40% is actually trying to get things to reproduce
- 20% is to deliver the result



From Kubeflow documentation

The research Asset commons

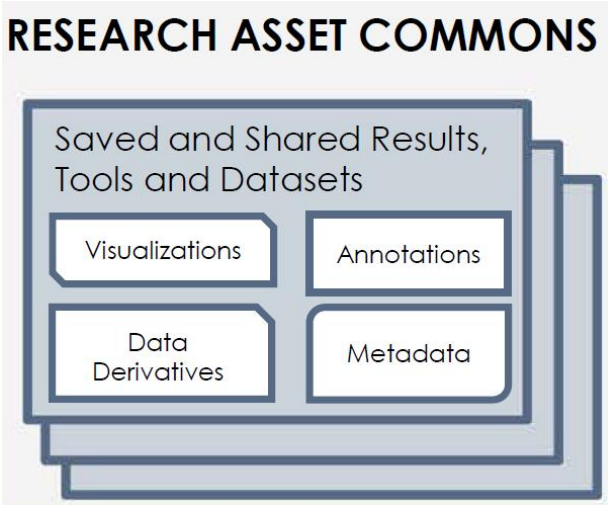
Engagement and education in Big Data, reproducible



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

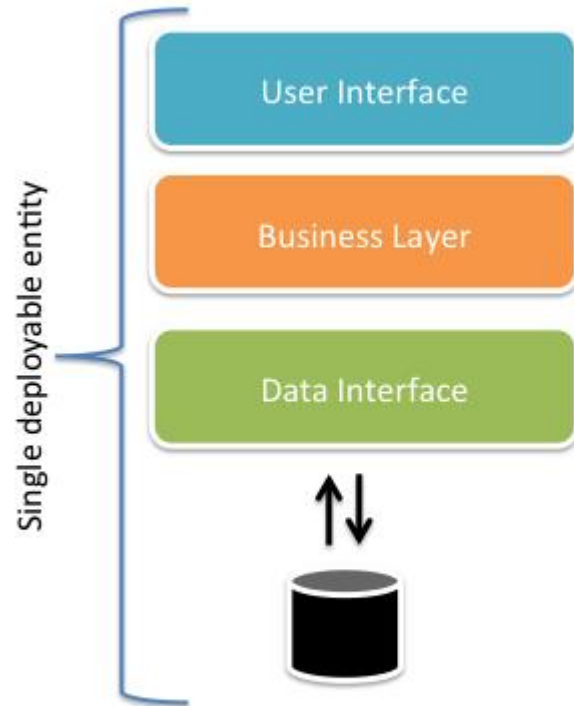


Publishing Results
Publish your research results with a single DOI with a full data provenance for a better reproducibility.

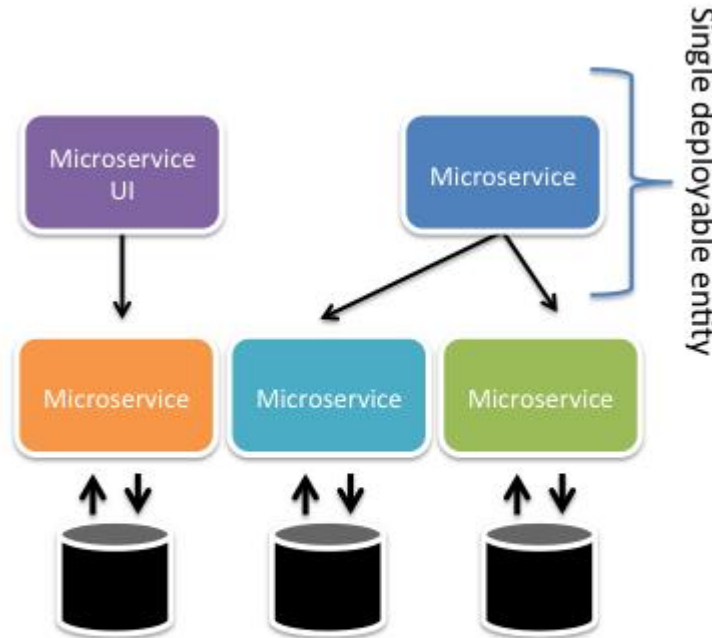


Lessons from the industry

Monolithic Architecture



Microservice Architecture



*Independent entities with cross communication through API's or Message Queuing

www.continuousautomation.com

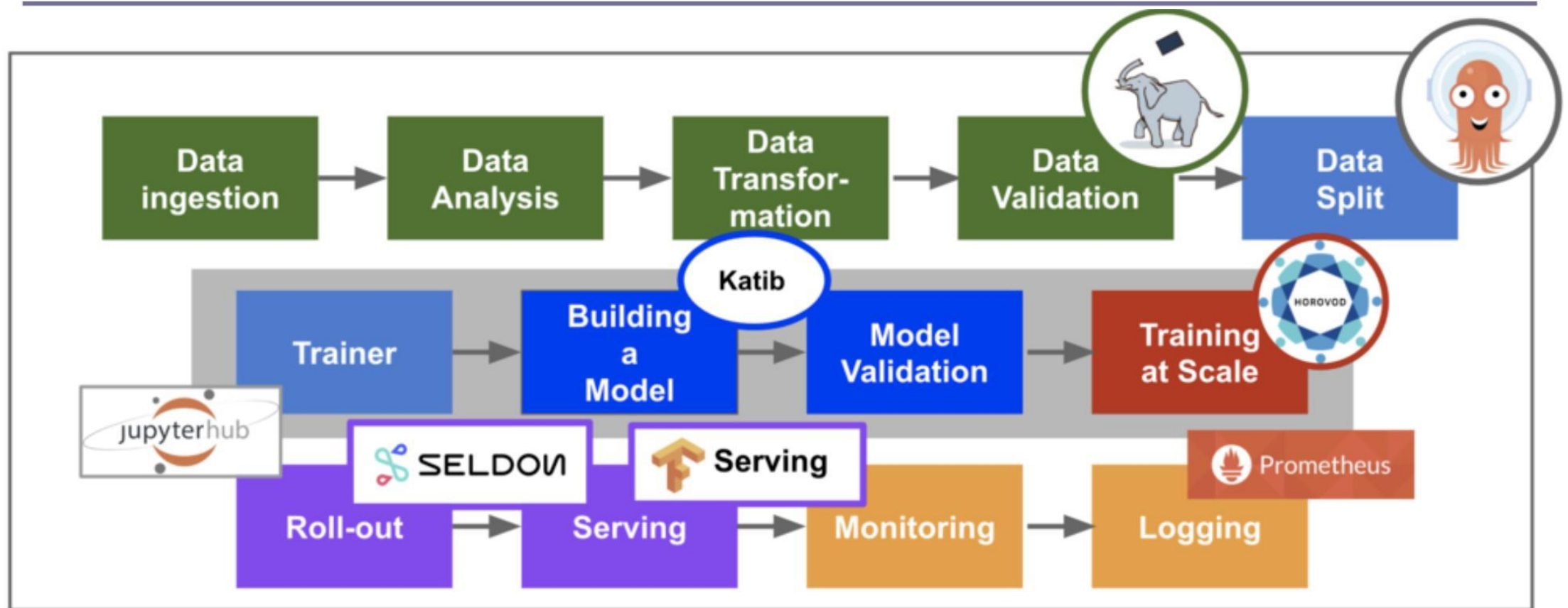
The new way: Deploy containers



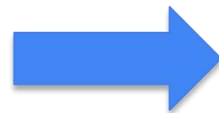
*Small and fast, portable
Uses OS-level virtualization*



Lessons from the industry



 + 
TensorFlow + kubernetes

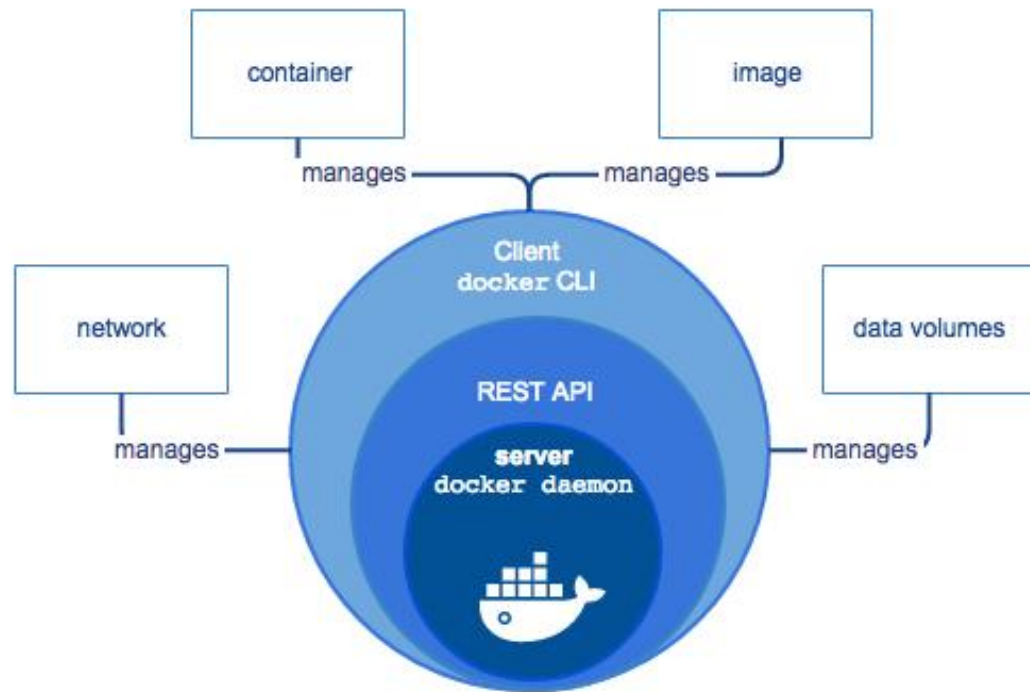



Kubeflow

www.programmersought.com



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



Kernel | Tabs | Settings | Help

Launcher | mag_info.csv | Feb_demo_01.py | Code

Last Modified	Count
6 1994	1
7 1996	1
8 1998	2
9 1999	2
10 2000	1
11 2001	1
12 2003	52
13 2004	3
14 2006	2
15 2007	33
16 2008	35
17 2009	36
18 2010	69
19 2011	67
20 2012	77
21 2013	102
22 2014	97
23 2015	83
24 2016	104
25 2017	108
26 2018	95

```

[5]: #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)

```

Corona.ipynb | Code

```

id_j = id_j[1:-1]

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

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

[22]: #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()

```

File | Edit | View | Run | Kernel | Tabs | Settings | Help

Corona.ipynb

```

[ ]: #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"]> 100]

[ ]: #Plot Data
plt.pie(id_j_plt["Total"], labels=id_j_plt["Journal"], radius = 3)
plt.show()

```

File | Edit | View | Run | Kernel | Tabs | Settings | Help

Part1.ipynb | Part2.ipynb

```

[5]: #Experimental
from IPython.display import IFrame
IFrame(src='https://filipinasci.com')

[5]:

```

cadre.iu.edu/jupyter/user/mvqwcz2dvornwq/lab

Corona.ipynb | Python 3

```

[ ]: #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))
plt.imshow(wc.generate(textData), interpolation='bilinear')
plt.axis("off")
plt.show()

```

Python 3 | Busy

Edges Depth Test

Vertex Scale: [Slider]

Vertex Intensity: [Slider]

Opaque Vertices

Edges Colors Property:

- Node Degree
- Community
- KCore
- Node InDegree
- doc_type

Create New Package

Package Name

VP New Pacjage

Input Data Archives [Click Here to archive a new data file.](#)

1 result.xnet



2 b30d1769-9180-4533-8896-ead7d509cdeb.csv



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 combination of small python script) and other files that are saved together and are run in a controlled environment. Each package is reproducible and should be able to be run when the 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

VP-Demo-02-Query_8f19f926-a36a-4d43-b262-ddbd4592e34f_nodes.csv

Create New Archive

A live demo

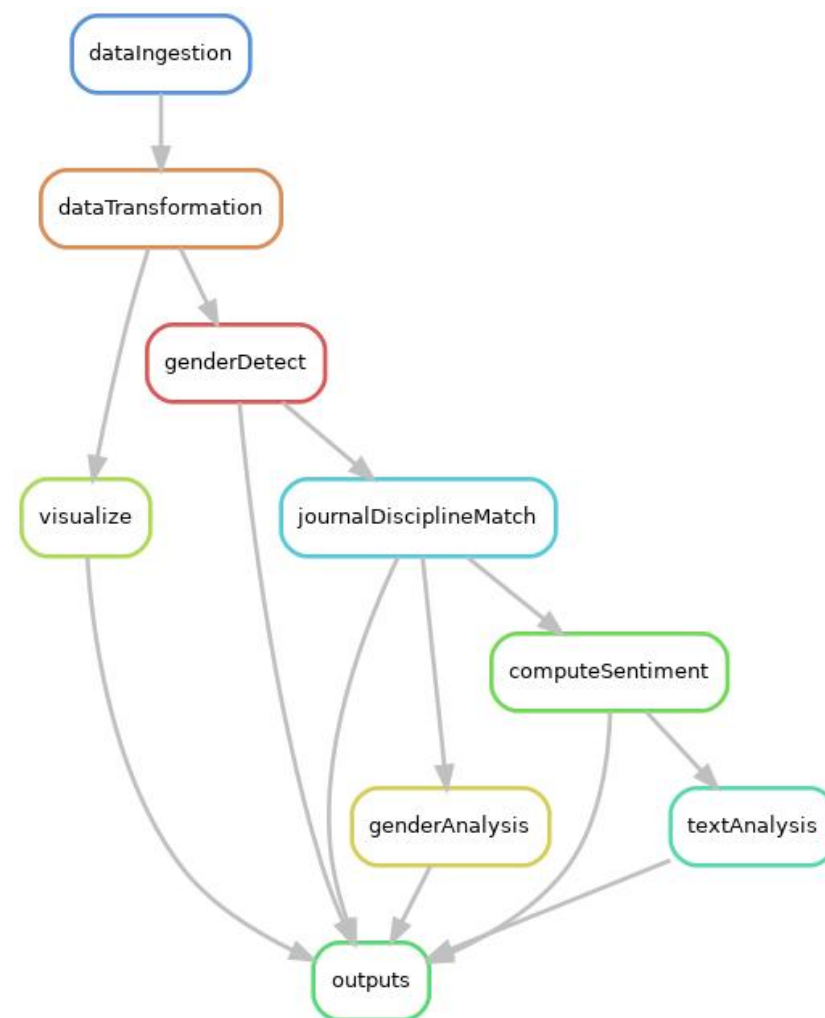
Open science, communal culture, and women's participation in the movement to improve science

 Mary C. Murphy,  Amanda F. Mejia,  Jorge Mejia,  Xiaoran Yan,  Sapna Cheryan, ...

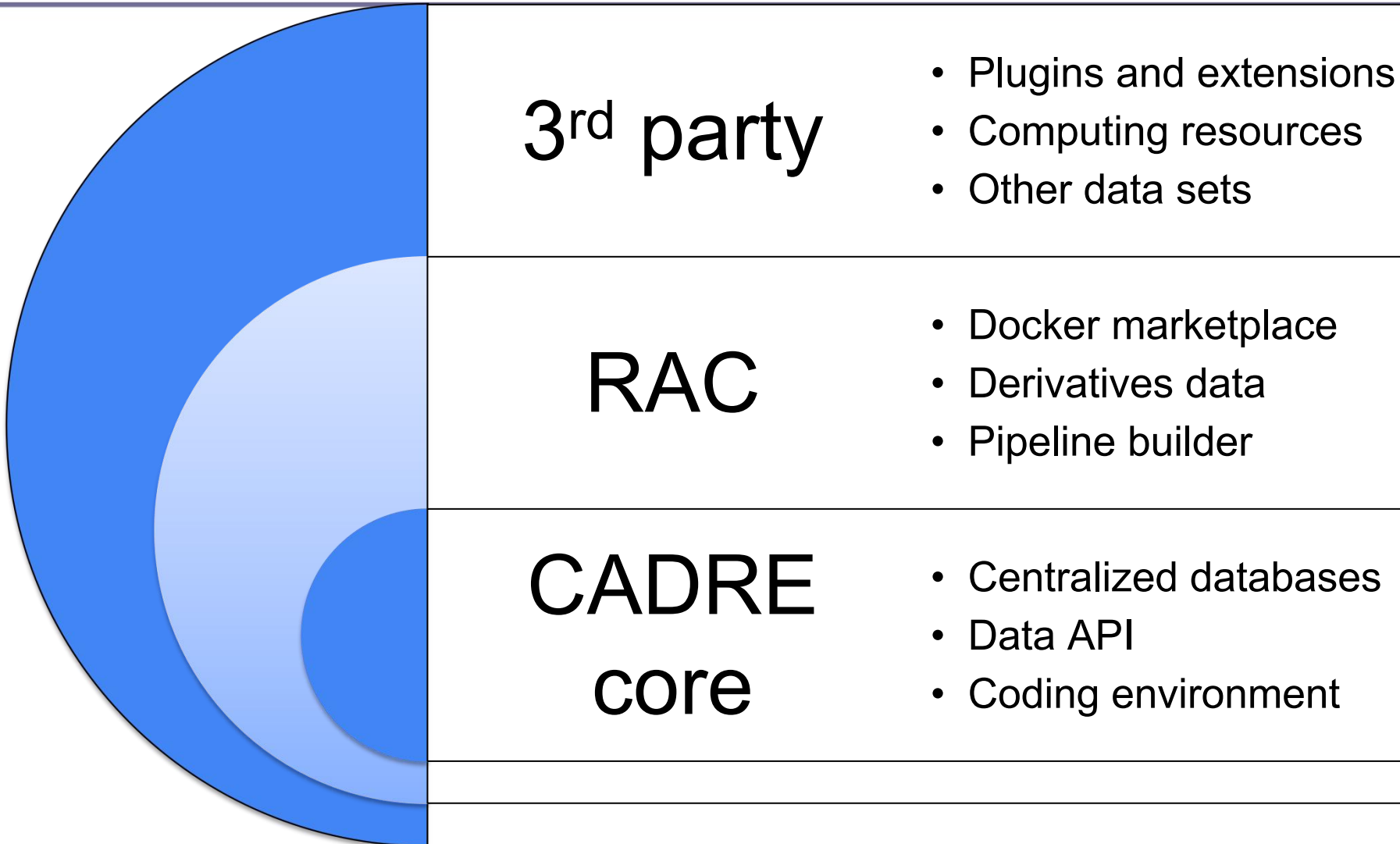
[+ See all authors and affiliations](#)

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

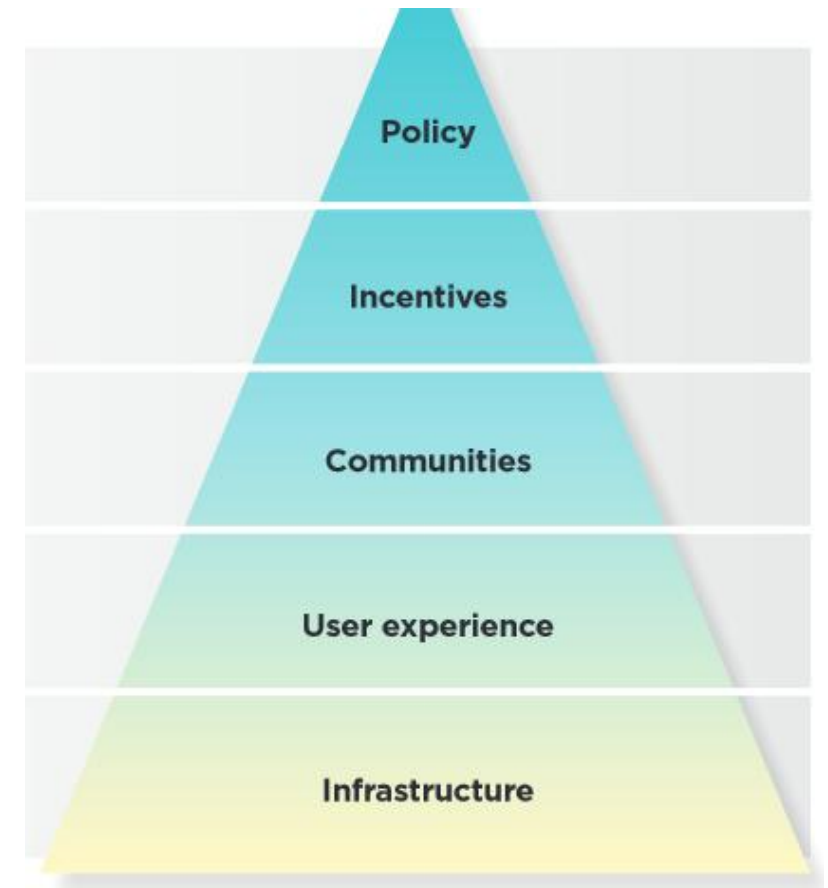
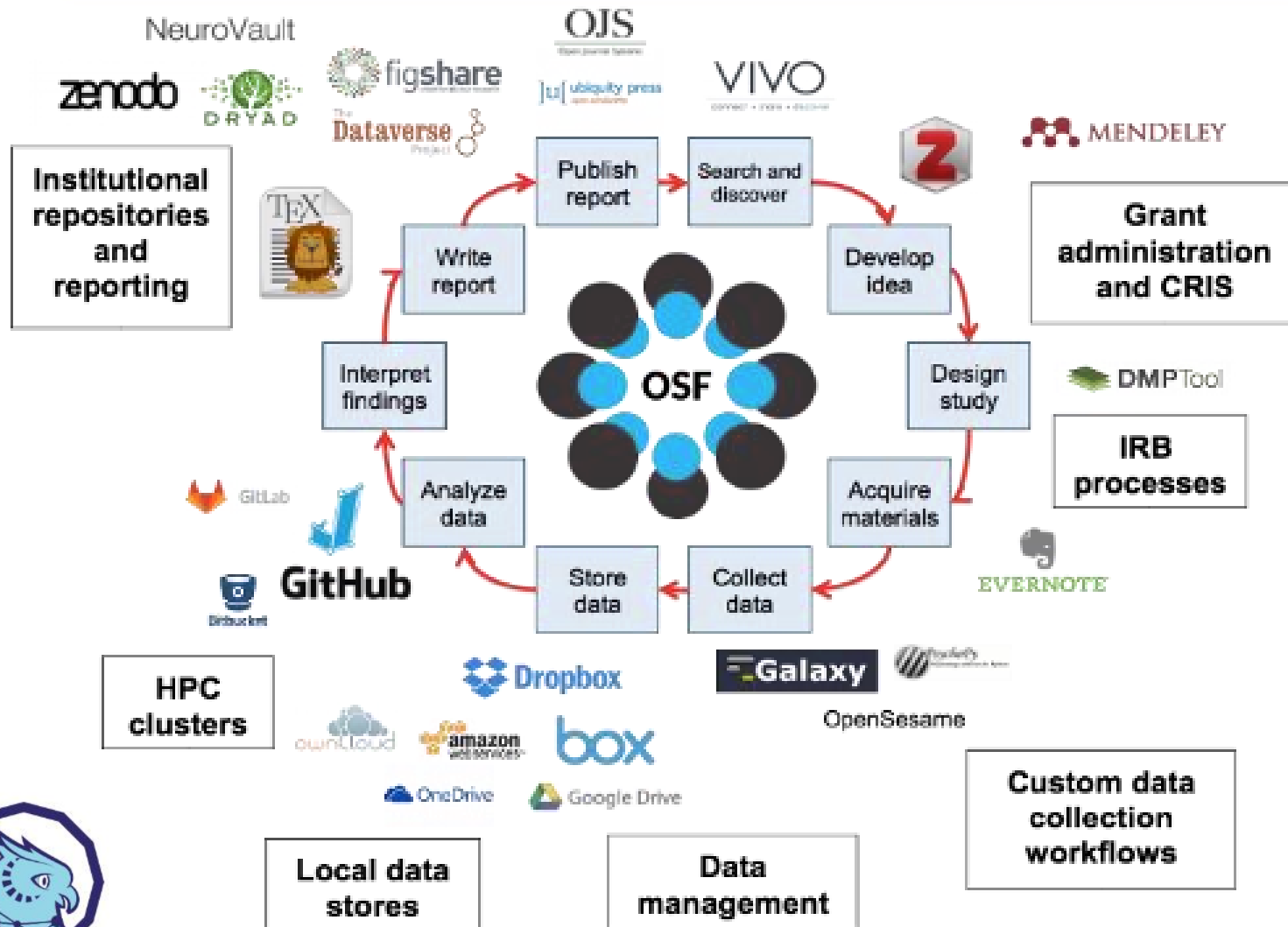
- ❑ **基于容器的鲁棒可重复流水线**
 - 云端数据自动导入清理 (Spark代码)
 - 作者识别与性别识别 (R代码)
 - 学术合作网络的社群结构分析 (Java代码)
 - 合作者的情感识别+文本分析 (Python代码)
 - 容器化的一键数据计算环境鲁棒复现



The CADRE ecosystem



Center for open science



目录

- 01 Big data and cloud computing
- 02 Cloud native computing and scientific reproducibility
- 03 Knowledge graphs and scientific search engine for Astrophysics
- 04 Large language models and the future of scientific search



本重点专项的总体目标是加强我国基础科研条件保障能力建设，着力提升**科学数据**等科研手段以及方法工具**自主研发与创新能力**；研制**可靠、耐用、好用、用户愿意用**的高端科学仪器。

应用

指南9.1：科学数据自主应用软件研发

构建**科学数据和天文知识体系**深度融合、面向提高天文科研工作效率和科普推广的**行业需求**的**文献深度分类**和个性化**自适应搜索推荐的自主系统**，并在**国家天文科学数据中心**展开示范应用。

数据

支持**多尺度、多模态、大规模科学数据**的收集、存储和**协同分析挖掘**。包括文献、图表、文本、实验数据等海量多模态数据的**开源自主**众包汇聚、**智能化追踪、关联**和更新。

知识

基于多模态预训练模型和知识图谱技术，研发**自主多模态领域知识**算法引擎，融合**天文学专家**的学科知识，实现**细粒度知识本体**构建，多模态实体的**智能识别、提取、关联和表达**。

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。



Astrophysics Data System

搜索功能：支持自然语言搜索

QUICK FIELD: Author First Author Abstract Year Fulltext All Search Terms

full:"super Earth"

Your search returned **6,897** results

object

Search for papers tagged with a specific astronomical object or at or near a set of coordinates

Syntax:
object:"object"

Example:
object:Andromeda

精准搜索：基于天文本体知识体系的（包括星表坐标）的查询搜索

AUTHORS

- Udry, S 270
- Seager, S 226
- Pepe, F 212
- Lovis, C 204
- Latham, D 200

COLLECTIONS

- astronomy 6.5k
- physics 1.1k
- general 234

REFEREED

INSTITUTIONS

KEYWORDS

PUBLICATIONS

BIB GROUPS

SIMBAD OBJECTS

- Star 2.3k
- Other 2.1k
- Galaxy 87
- Nebula 16
- X-ray 13

1 2022car..38215017P 2022/08
Random models for exploring planet compositions I: Uranus as an example
Podolak, Joshua I.; Malamud, Uri; Podolak, Morris

2 2022vaid.book..227R 2022/06 cited: 5
Origin and Dynamical Evolution of the Asteroid Belt
Raymond, Sean N.; Nesvorný, David

3 2022MNRAS.513.1544K 2022/06
The PEPSI exoplanet transit survey (PETS) I: investigating the presence of a silicate atmosphere on the super-earth 55 Cnc e
Keles, Engin; Mallonn, Matthias; Kitzmann, Daniel and 26 more

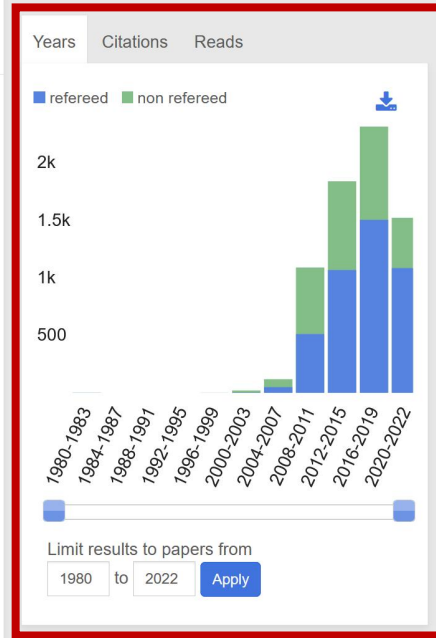
4 2022MNRAS.513..102C 2022/06
The TESS Triple-9 Catalog: 999 uniformly vetted exoplanet candidates
Cacciapuoti, Luca; Kostov, Veselin B.; Kuchner, Marc and 20 more

5 2022MNRAS.512.5552B 2022/06
A numerical inversion of $m \sin i$ exoplanet distribution: the sub-Saturn desert is more depleted than observed
Bertaux, Jean-Loup; Ivanov, Ivan

6 2022MNRAS.512.5228S 2022/06
The DRAKE mission: final results
Sarkar, Subhajit

7 2022MNRAS.512.5023F 2022/06
Sculpting the circumbinary planet size distribution through resonant interactions with companion planets
Fitzmaurice, Evan; Martin, David V.; Fabrycky, Daniel C.

推荐排序功能：基于分面搜索的关联排序和精细分类



相关论文、数据集等的可视分析

- Visualizations
- Citation Metrics
- Author Network
- Paper Network
- Concept Cloud
- Results Graph
- Operations ?
- Co-reads
- Reviews
- Useful
- Similar

Astrophysics Data System

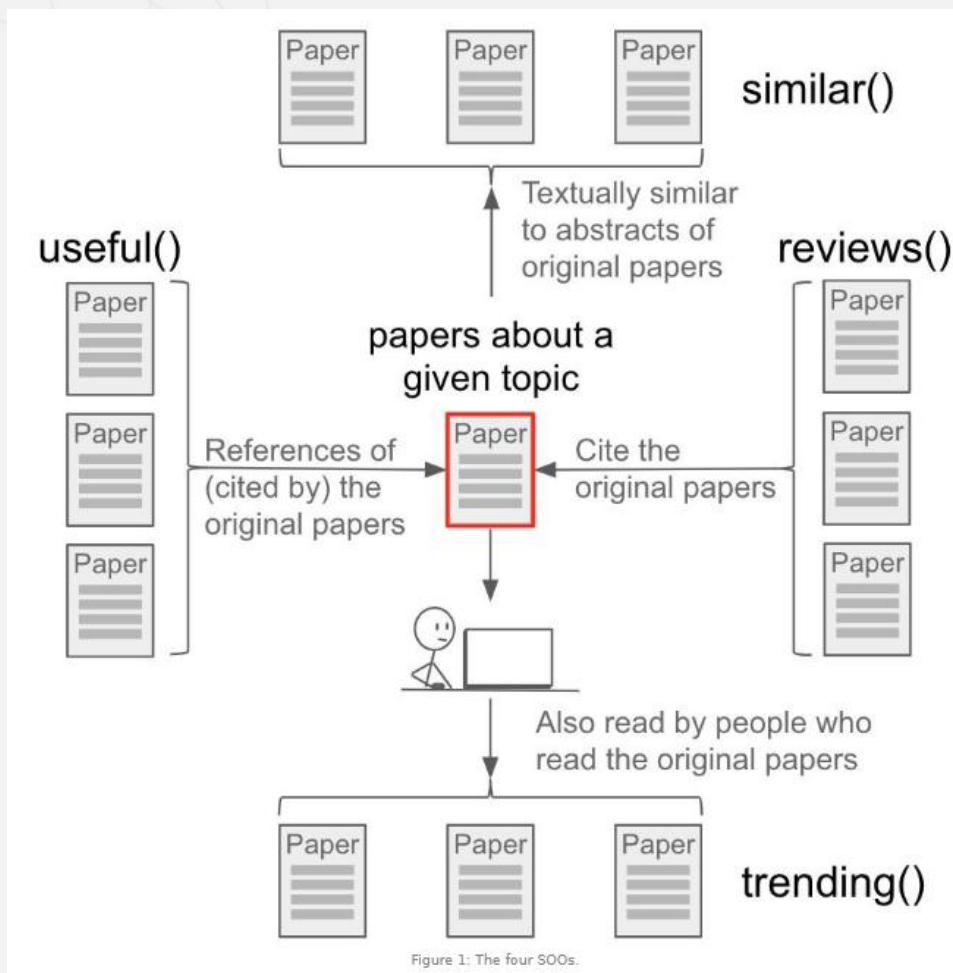
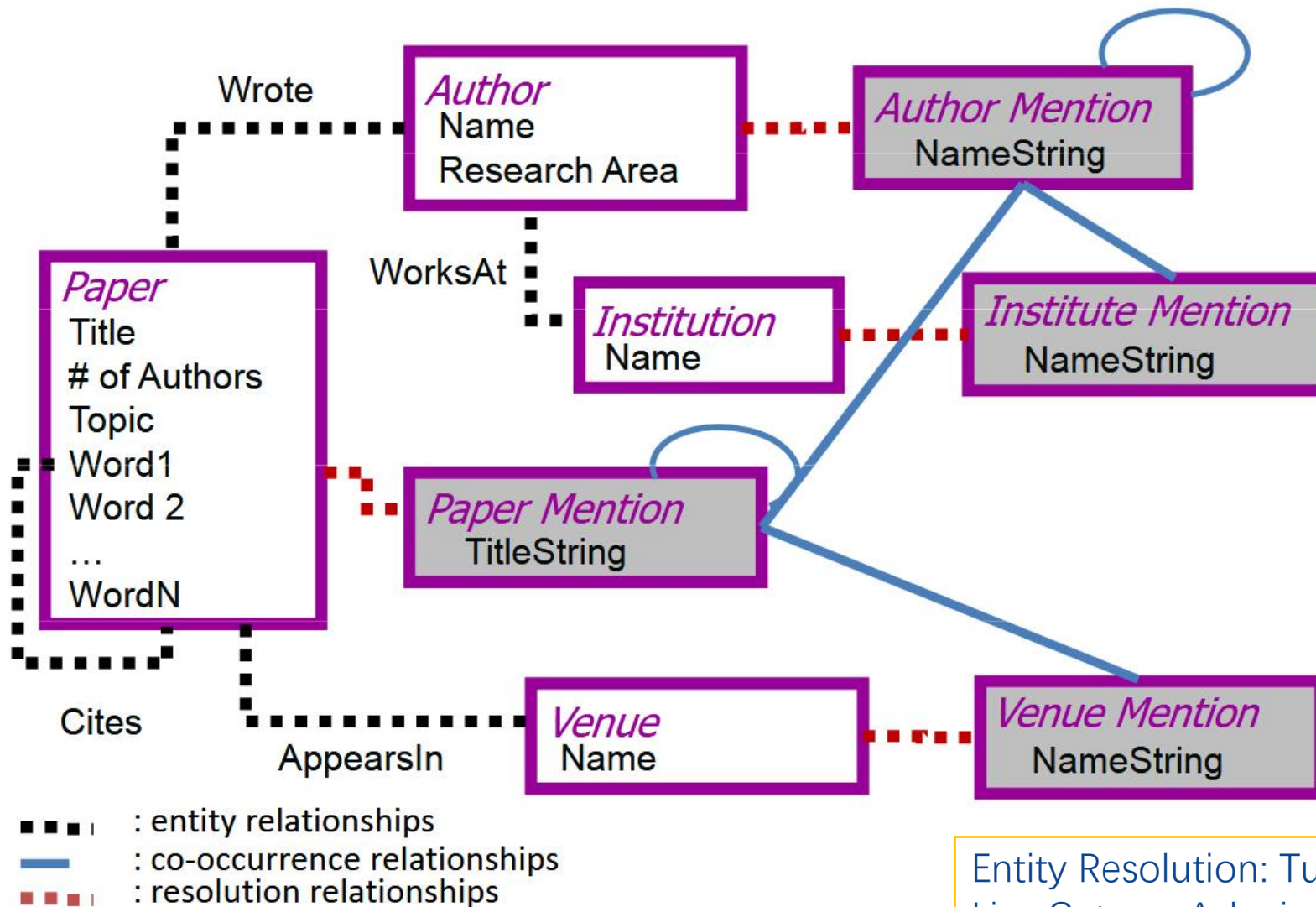
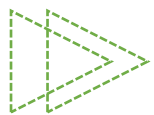


Figure 1: The four SOOs.

- ◆ 数据来源复杂，有多个数据协议与历史积累
- ◆ 早年就从关系型数据库转移到了XML+知识驱动文本索引
- ◆ 利用OCR技术提取摘要和引用
- ◆ 文章匹配消歧是技术上的一大难点
- ◆ DOI/ORCID的引入和群在回路信息收集/修正提出了OLTP+OLAP的双重需求
- ◆ 主要面对文字、XML搜索功能，缺乏图结构数据搜索引擎

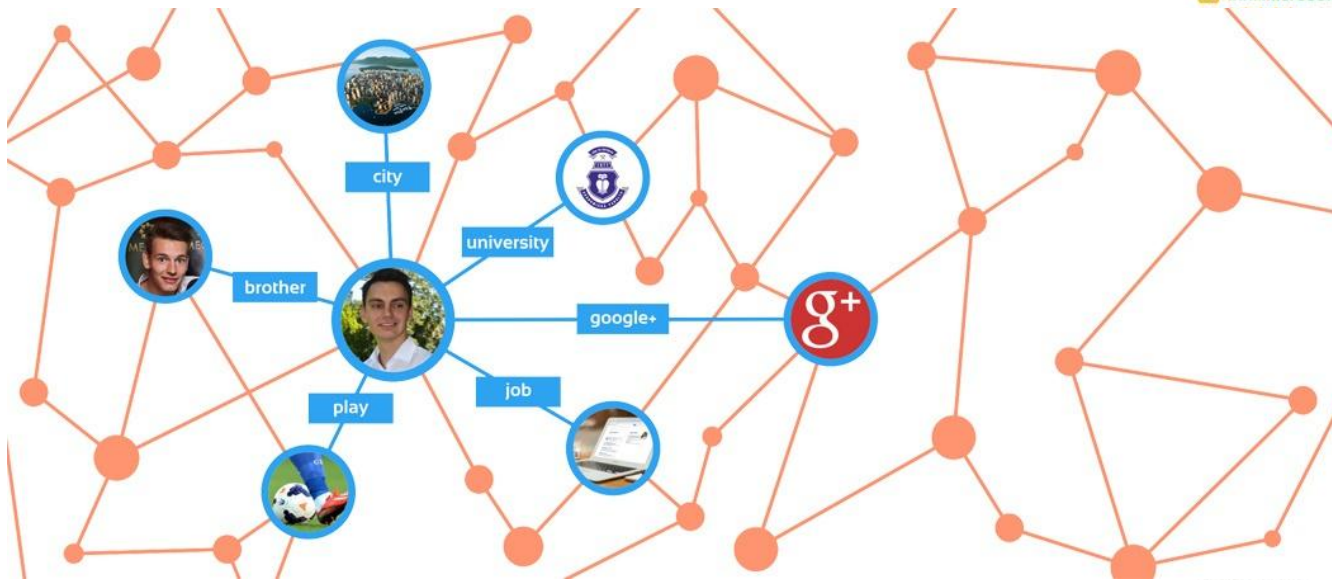
The ADS Team. "Second Order Operators in the NASA Astrophysics Data System." <https://doi.org/10.3847/25c2cfef.8d12c399>.

"ADS: The Next Generation Search Platform." arXiv, March 13, 2015. <http://arxiv.org/abs/1503.04194>.



Entity Resolution: Tutorial (VLDB2012)
Lise Getoor, Ashwin Machanavajhala

Knowledge graph and search engines



The screenshot shows a Google search for 'microsoft'. The search results include a snippet for 'Microsoft Store: Shop Now - MicrosoftStore.com', a 'Home Page' result, and a news article titled 'Mozilla blasts Microsoft for making it harder to switch to Firefox in Windows 10'. A large red arrow labeled 'Knowledge graph' points from the graph on the left to the search results. On the right, a detailed knowledge panel for Microsoft Corporation is displayed, including its stock price, CEO, and social media profiles.

Microsoft Corporation
Technology company

Microsoft Corporation /ˈmɪkrəsoʊˈft/ is an American multinational technology company headquartered in Redmond, Washington, that develops, manufactures, licenses, supports and sells computer software, ... Wikipedia

Customer service: 1 (800) 642-7676

Stock price: MSFT (NASDAQ) \$46.88 +0.59 (+1.27%)
Jul 30, 4:00 PM EDT - Disclaimer

CEO: Satya Nadella

Founded: April 4, 1975, Albuquerque, NM

Headquarters: Redmond, WA

Founders: Bill Gates, Paul Allen

Subsidiaries: Skype Technologies, Microsoft Store, Wintertals, More

Profiles

Twitter Facebook LinkedIn Google+ Instagram

People also search for View 15+ more

Nokia Sony Corporation Dell Intel Apple Inc.



Lessons learned and user stories

I want a unified query interface to both WoS and MAG at the same time.

Relational DB cannot support large scale citation queries.

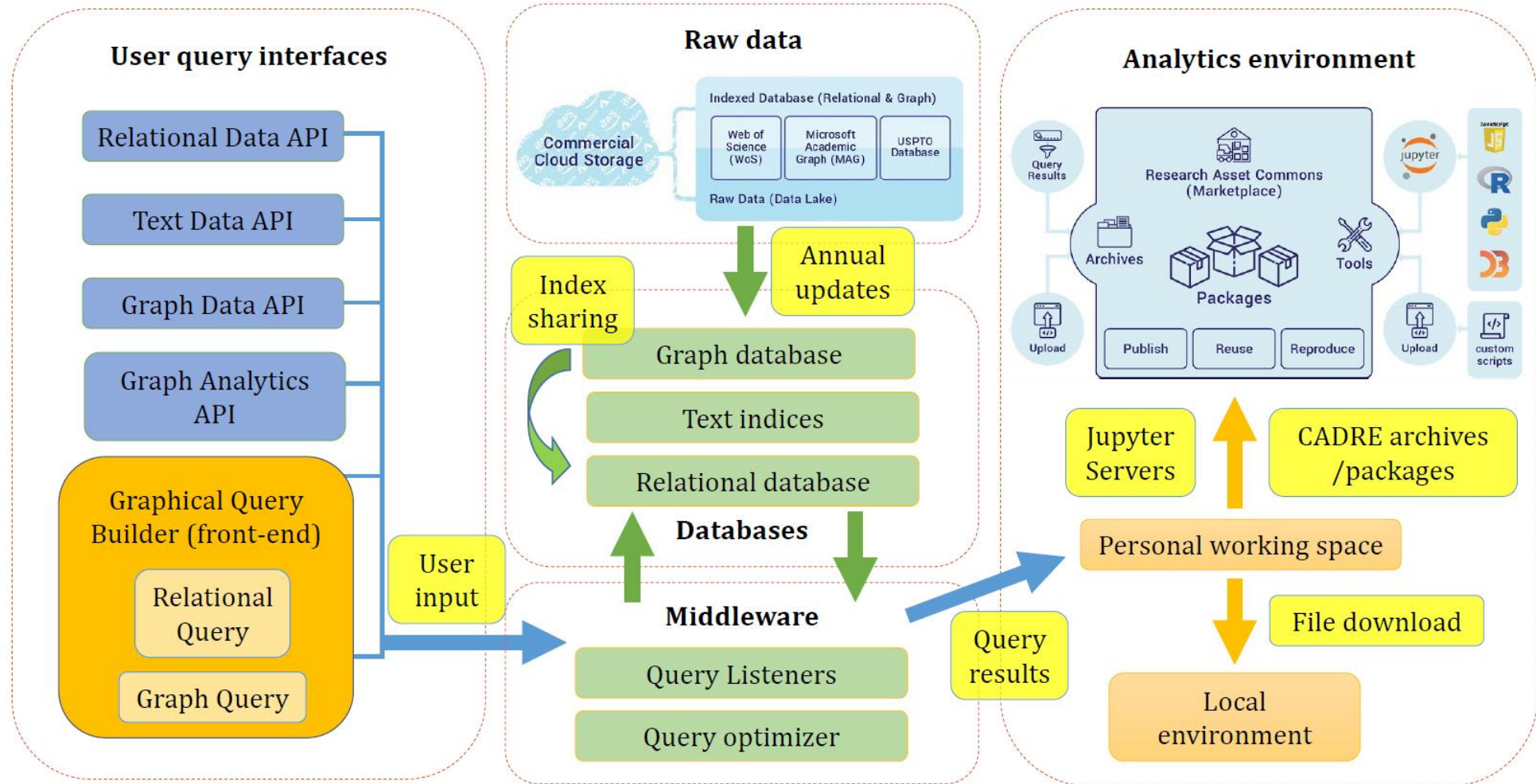
I need the ability to directly query the db from the Jupyter notebooks via code.

The work I want to do would greatly benefit from a graph DB.

Initial hybrid solution with PostgreSQL+Neo4j works but not scalable enough



Flow of data on CADRE



Metrics in Evaluating Graph Databases

Metric	Description	Priority
Scalability	The capability to host gigantic graph of millions of nodes and billions of edges	High
Text search	The capability to support both plain string and full text search against graph node and edge attributes	High
Query performance	The performance of the graph database on general-purpose and graph traversal queries	High
Cost effectiveness	Both the operational cost and software license cost to run the graph DBMS and the middleware built around it	High
Graph algorithm support	The availability of implementations of common graph metrics and algorithms	Medium
Ingestion	The time needed to populate each dataset in the graph database	Low



Benchmarking: Graph Databases

Database	Store Type	Model	Open Source	Language	Components
AgensGraph	Disk	Single	No	Cyber	Postgre SQL
JanusGraph	Disk	Cluster	Yes	Gremlin	Cassandra, Elasticsearch
MemoryGraph	Memory	Single	No	Gremlin	self-contained
Neo4j	Disk	Single	Yes	Gremlin	self-contained
RedisGraph	Memory	Single	No	Gremlin	Redis
TigerGraph	Disk	Single	No	GSQL	self-contained

Table: Benchmarking graph databases.



Benchmarking: Dataset

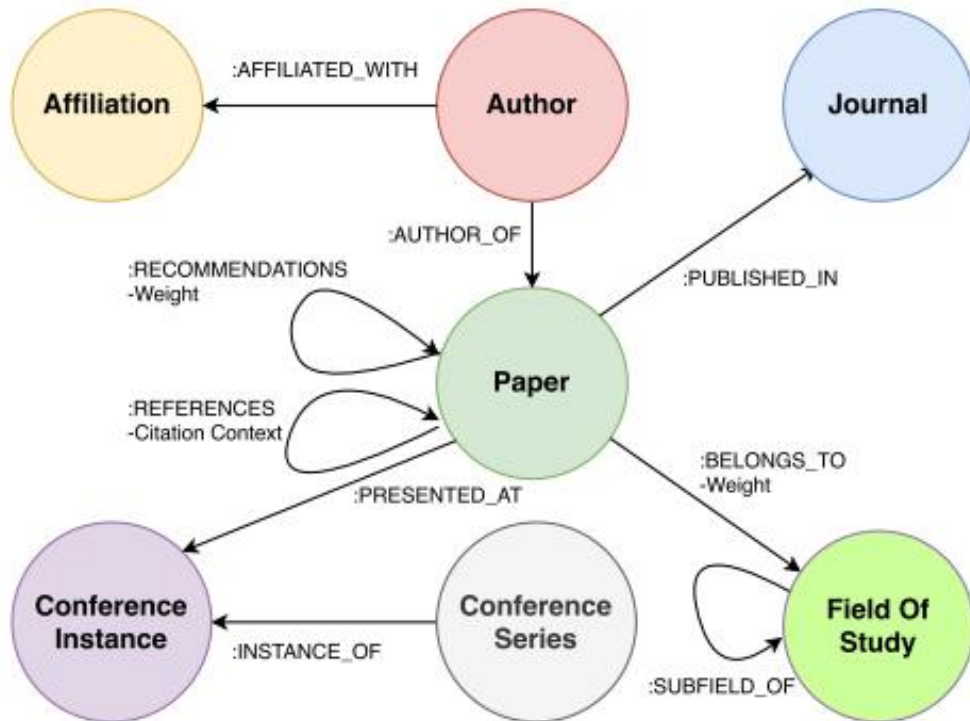


Figure: The Microsoft Academic Graph (MAG) schema

Dataset	Nodes (M)	Edges (M)	Size on disk (GB)
Full MAG	700	2,300	500
Reduced MAG	60	90	30

Table: The Microsoft Academic Graph (MAG) details



Benchmarking: Queries

Query	Type	Description
Q1	General-purpose	Fetches all attributes given a paper title.
Q2	General-purpose	Fetches all attributes for each author of a paper given its title
Q3	General-purpose	Fetches the fields of study a journal's publications belong to given a journal title
Q4	String Matching	Fetches all authors with papers matching the title criteria
Q5	Graph	Finds all citations for a given paper title.
Q6	Graph	Finds all references for a given paper title.
Q7	Graph	Fetches all 2-hop citations for a given paper title.
Q8	Graph	Fetches all 3-hop citations for a given paper title.
Q9	Graph	Fetches all 4-hop citations for a given paper title.
Q10	Graph	Fetches the citations for all papers in a given field of study name.

Table: Benchmarking queries.



Benchmarking: Query Performance

Query	JanusGraph	TigerGraph 1	TigerGraph 2	MemGraph	Neo4j	AgensGraph
Q1	1.82	880	480	2.26	82,370	1.86
Q2	3.66	17	14	1.88	6820	1850
Q3	69	190	28	29.89	120	100
Q4	varies	141670	32960	24950	15990	32310
Q5	2.09	17	13	3.44	6890	8160
Q6	1.4	30000	13	1.48	6860	8060
Q7	62.53	varies	varies	3.02	timeout	timeout
Q8	20900	varies	varies	8.05	timeout	timeout
Q9	timeout	varies	varies	5.32	timeout	timeout
Q10	varies	9620	100	3.04	138000	112910

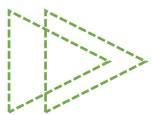
Table: Benchmarking query performance. Time is in milliseconds.



目录

- 01 Big scientific data and cloud computing
- 02 Cloud native computing and scientific reproducibility
- 03 Knowledge graphs and scientific search engine for Astrophysics
- 04 Large language models and the future of scientific search





国内外业界发展现状

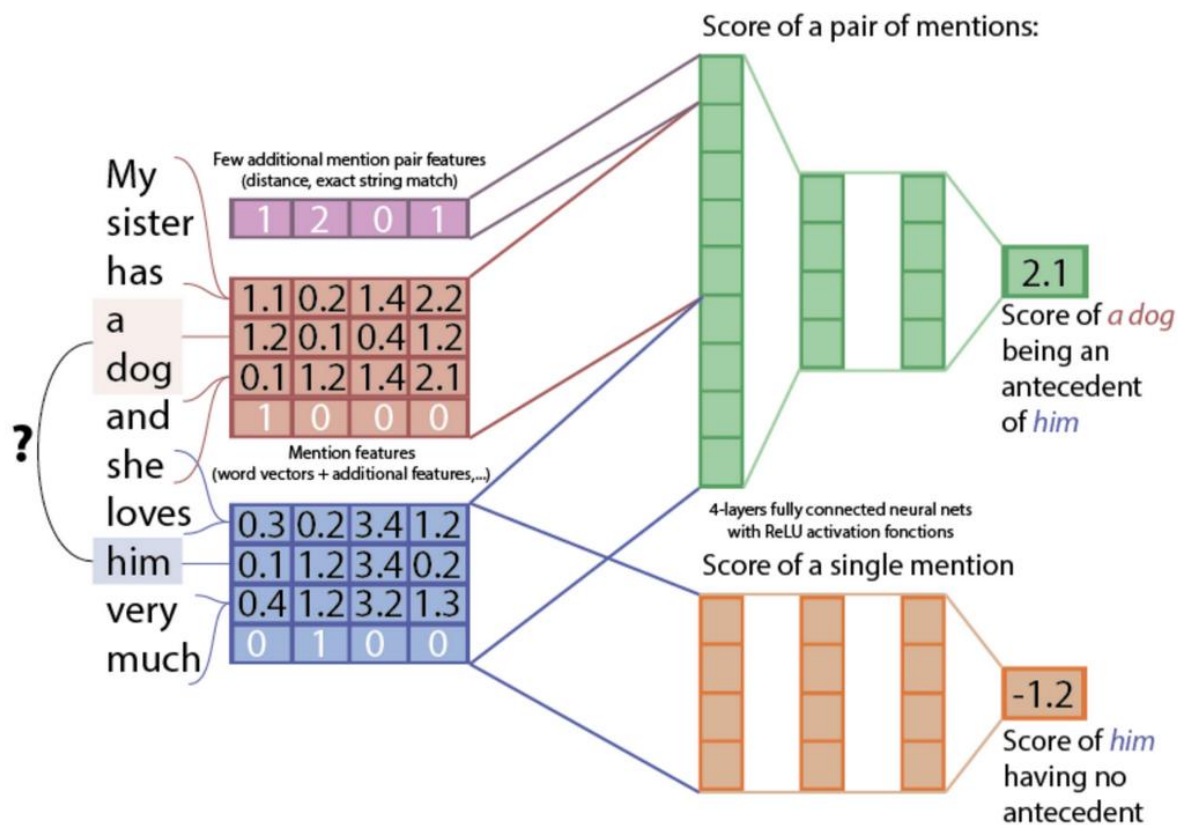


美团大脑：知识图谱的建模方法及其应用，2018



- 多为大型通用图谱
- Freebase & Wikipedia
- 海量数据+高度结构化的筛选
- 逐步过渡到property graph表征

- 高度依赖人工schema定义
- 需要大量资源长时间积累
- 缺乏交互式更新能力
- 专业小型图谱则多为人工构建



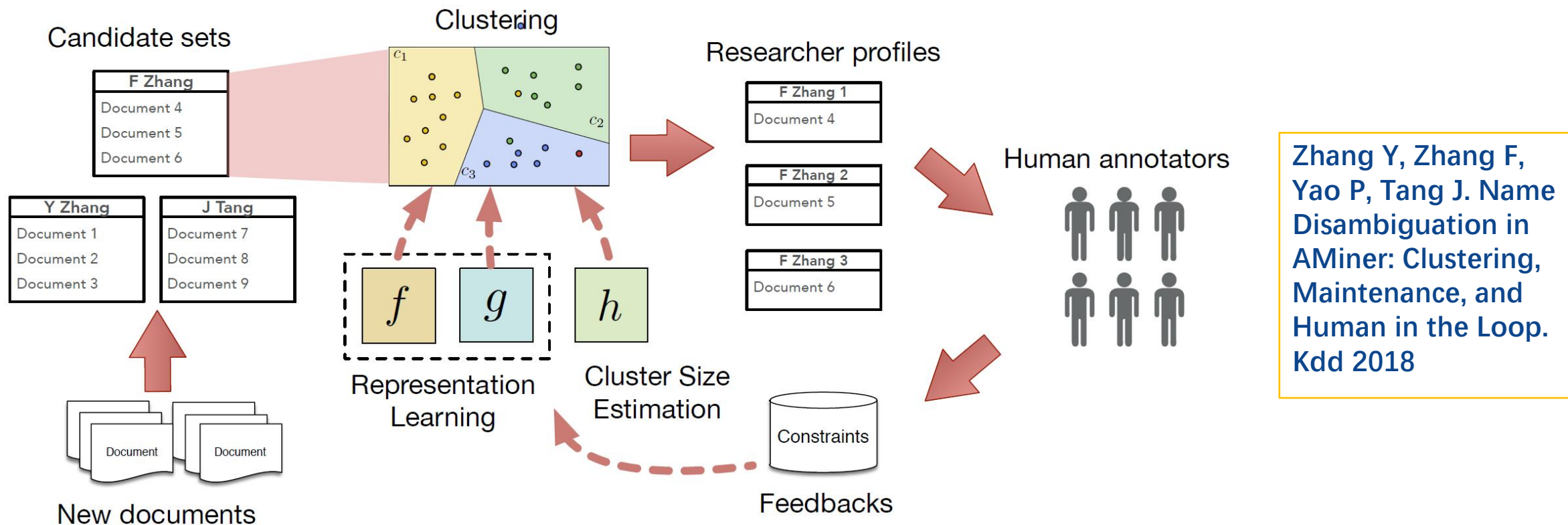
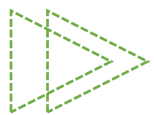
Rough sketch of the coreference scoring model.

深度模型算法

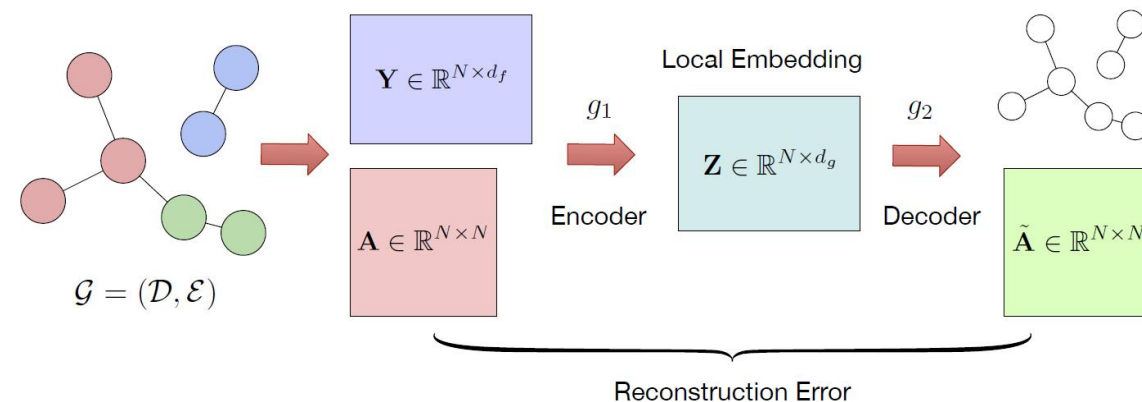
- 文本源数据的向量表征
- 深度语言模型的运用
- 知识图谱的嵌入算法
- 图卷积网络用于概率图模型的训练
- 极大地提高了算法可扩展性

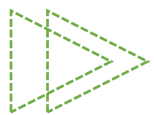
Four of the five winning teams in this Contest (UWA, BUPT-IBL, MIDAS-IIITD, and Lab1105) used NeuralCoref [43] for entity resolution.

<https://github.com/huggingface/neuralcoref>

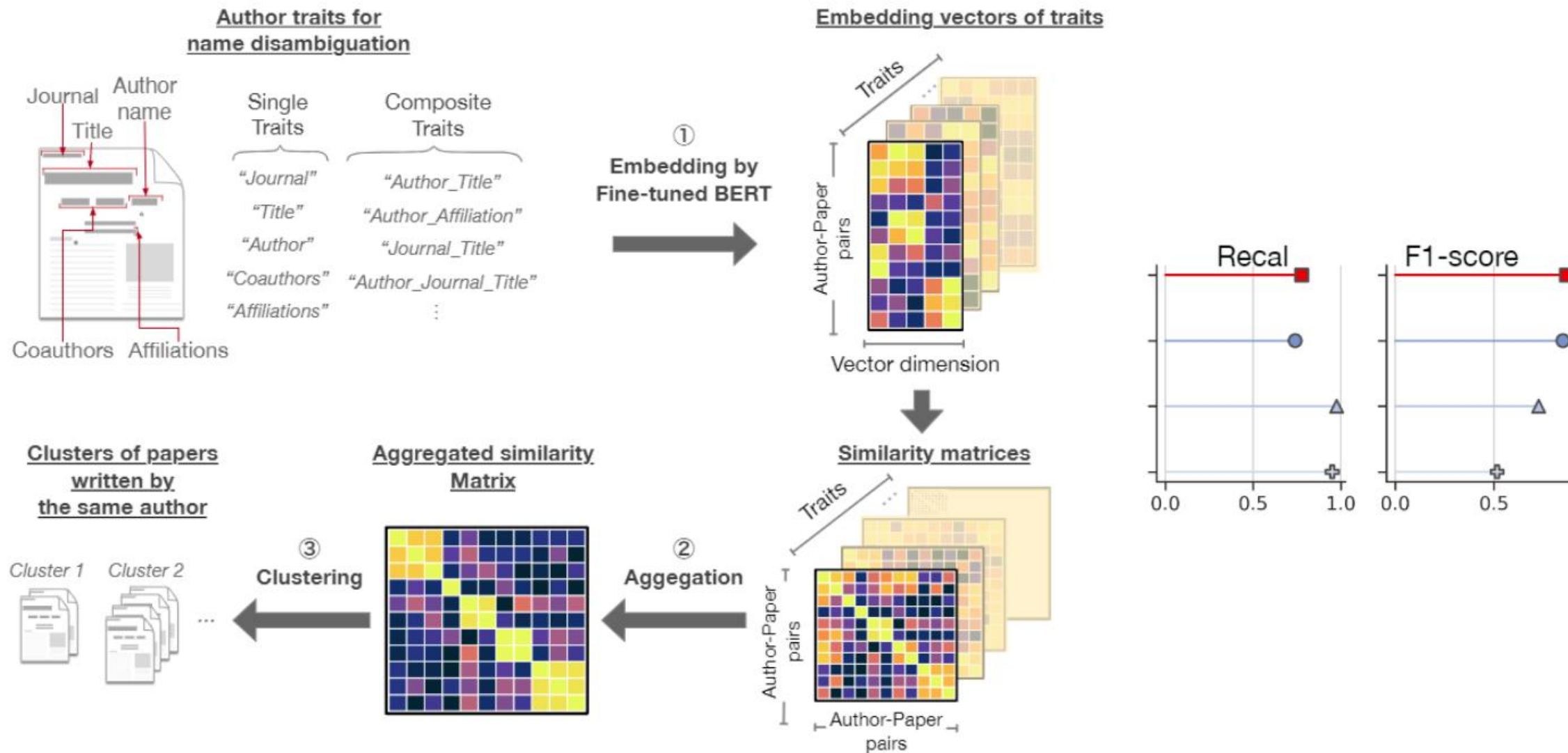


- 深度模型提高了数据吞吐能力（1.3亿作者，2亿文章，每月新增50万篇）
- 允许新数据以及人在回路的动态更新
- 放弃了图关联模型和其他实体的识别**





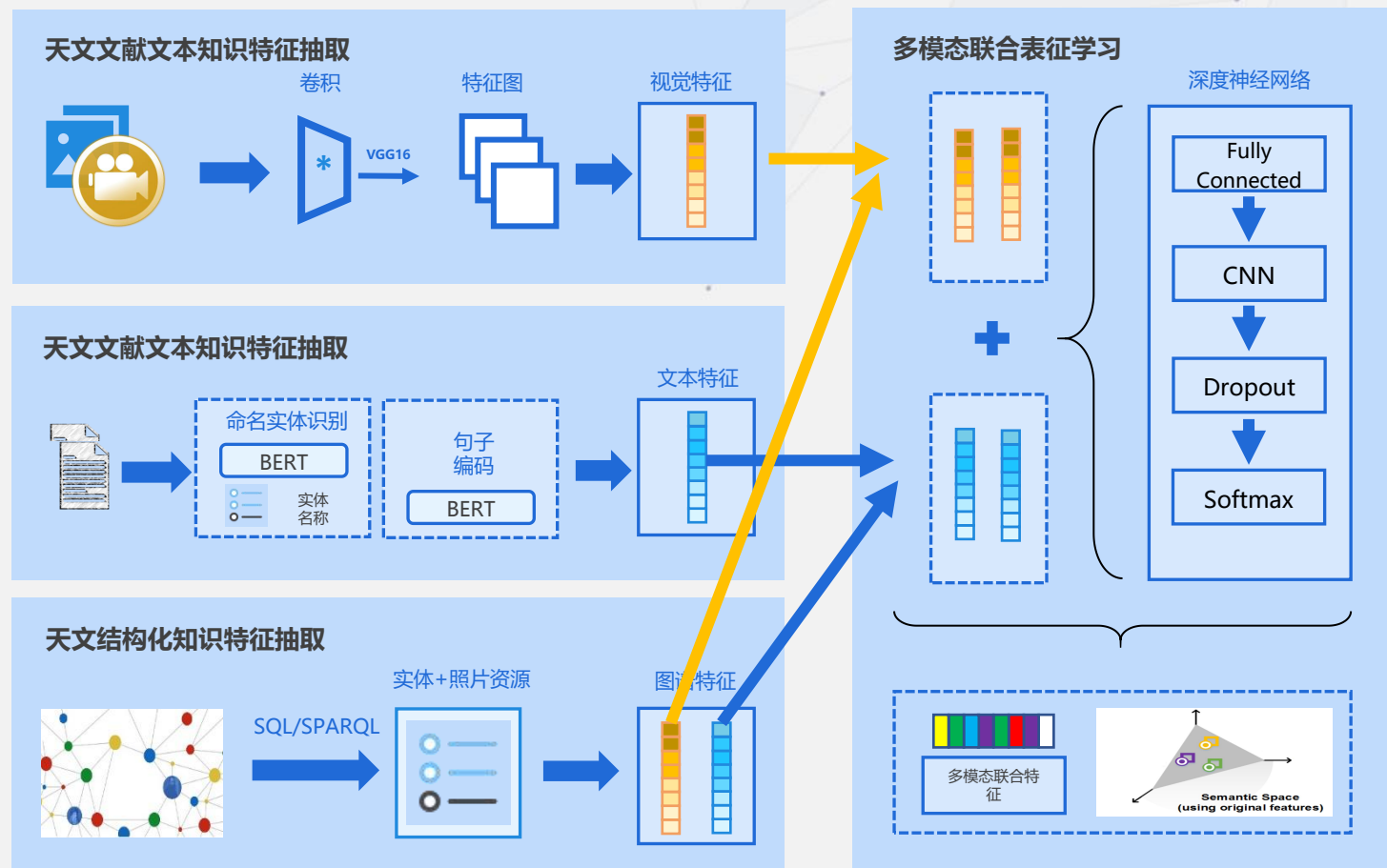
我们的工作



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**。

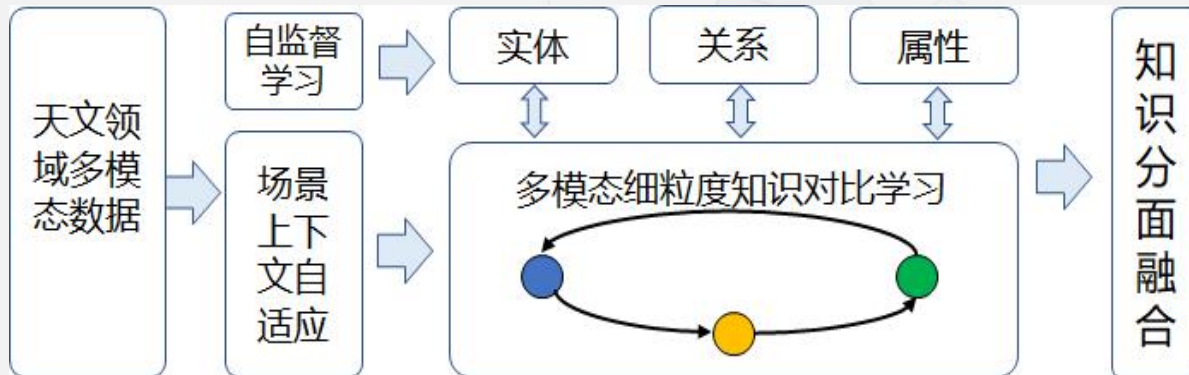








多模态细粒度知识提取关联

技术方案-多模态预训练语言模型

本课题设计“特征抽取—联合表征”两阶段的天文领域知识统一表征方法：

- 采用基于深度自监督学习模型进行单模态知识实例的特征抽取
- 采用基于多模态联合嵌入学习的跨模态表征方法，实现海量异构多模态知识特征的有效学习，实现跨模态异构知识联合表征

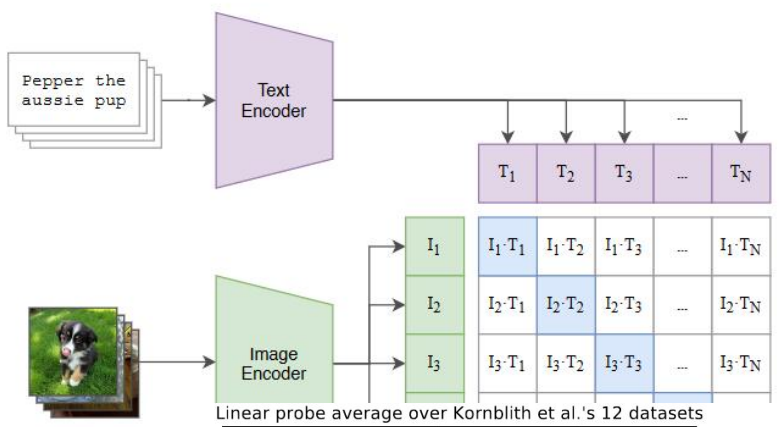


	Dataset Examples	ImageNet ResNet101	Zero-Shot CLIP	Δ Score
ImageNet		76.2	76.2	0%
ImageNetV2		64.3	70.1	+5.8%
ImageNet-R		37.7	88.9	+51.2%
ObjectNet		32.6	72.3	+39.7%
ImageNet Sketch		25.2	60.2	+35.0%
ImageNet-A		2.7	77.1	+74.4%

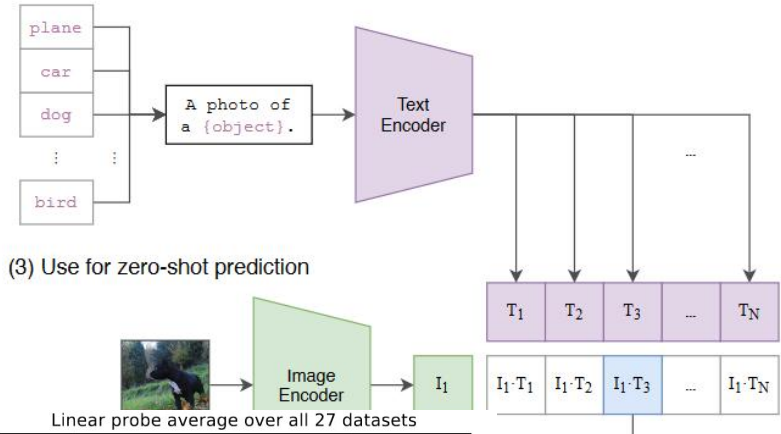
CLIP跨模态大模型

Learning Transferable Visual Models From Natural Language Supervision

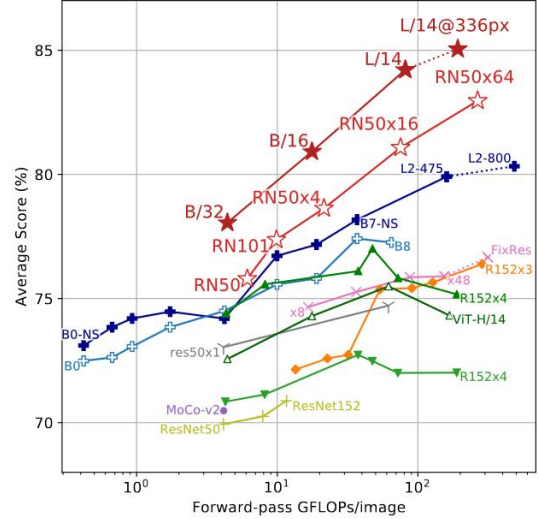
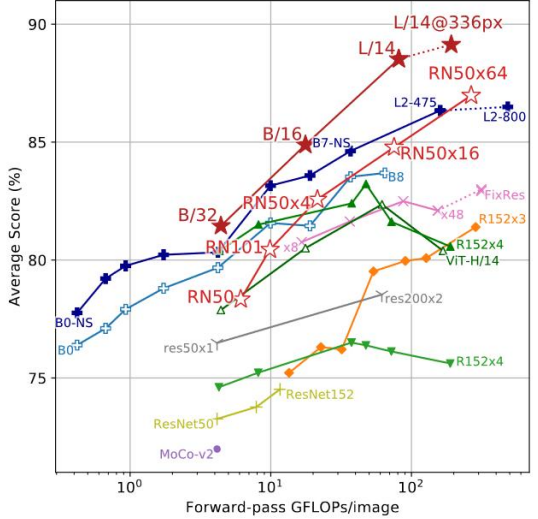
(1) Contrastive pre-training



(2) Create dataset classifier from label text



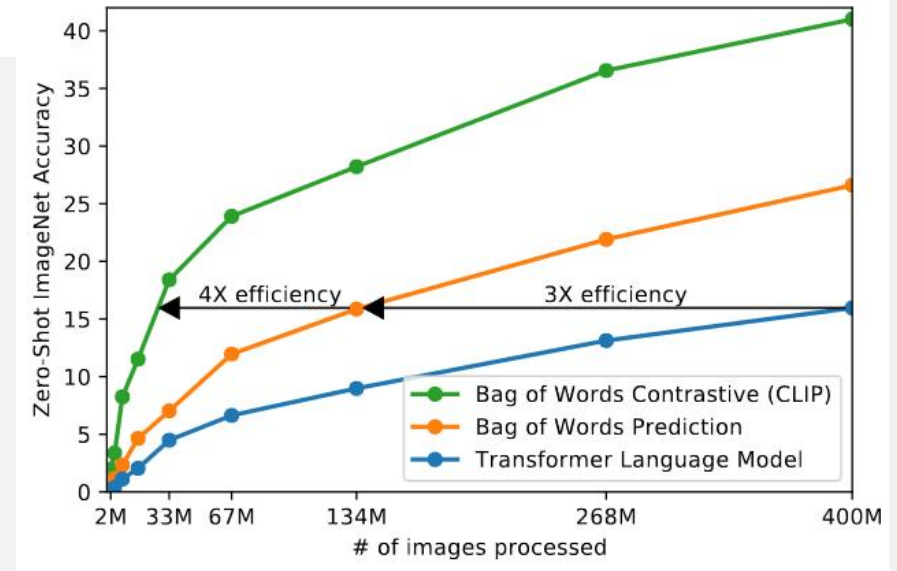
(3) Use for zero-shot prediction



- ★ CLIP-ViT
- ★ Instagram-pretrained
- ▲ ViT (ImageNet-21k)
- ★ CLIP-ResNet
- ◆ SimCLRv2
- ▲ BIT-M
- ◆ EfficientNet-NoisyStudent
- ▲ BYOL
- ▲ BIT-S
- ◆ EfficientNet
- ◆ MoCo
- + ResNet

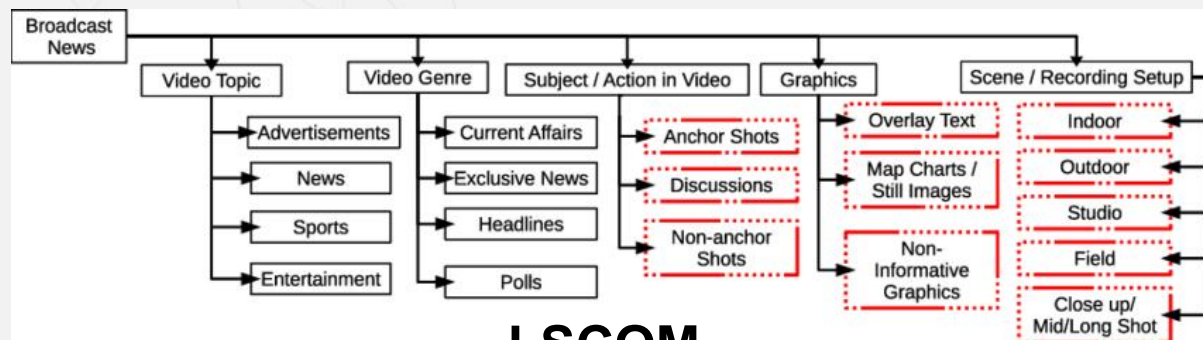
跨模态大规模预训练:

- 4M Text-image Data is all you need
- ViT版本模型训练需要6000 V100days
- Zero shot 迁移学习, 碾压监督学习
- 五花八门的改造应用



研究目标3：多模态科学数据的知识关联抽取融合与表征

(7)多模态领域本体和语料预训练模型：



LSCOM

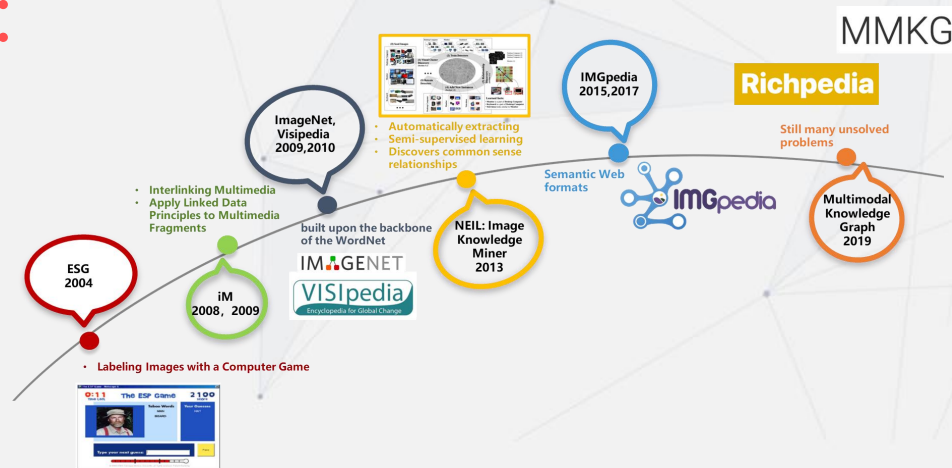
Large-Scale Concept Ontology for Multimedia, *IEEE Multimedia Magazine*, 13(3), 2006.

COMM

```

<?xml version="1.0" encoding="UTF-8" ?>
<Description xsi:type="ContentEntityType">
  <MultimediaContent xsi:type="ImageType">
    <Image id="IM01">
      <SpatialDecomposition>
        <StillRegion id="SR1">
          <Semantic>
            <Label><Name> Roosevelt </Name></Label>
          </Semantic>
        </StillRegion>
        <StillRegion id="SR2">
          <TextAnnotation>
            <!-- TextAnnotationType -->
            <KeywordAnnotation><Keyword> Churchill </Keyword></KeywordAnnotation>
          </TextAnnotation>
        </StillRegion>
        <StillRegion id="SR3">
          <Semantic>
            <Definition> <!-- Also TextAnnotationType -->
            <StructuredAnnotation><Who><Name> Stalin </Name></Who></StructuredAnnotation>
          </Definition>
        </StillRegion>
      </SpatialDecomposition>
    </Image>
  </MultimediaContent>
</Description>
      
```

COMM: A core ontology for multimedia annotation, *Handbook on Ontologies*, 2009

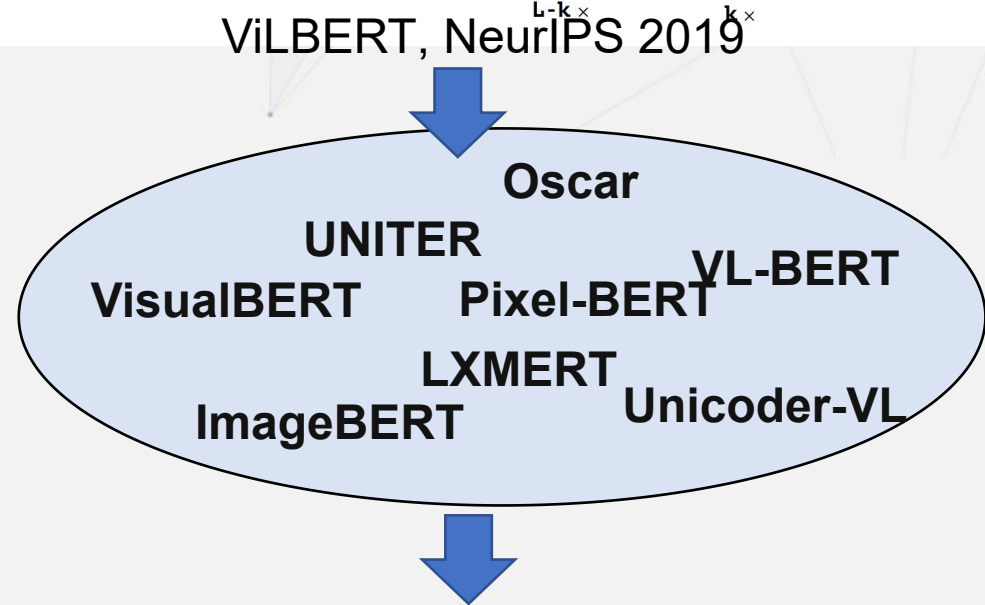
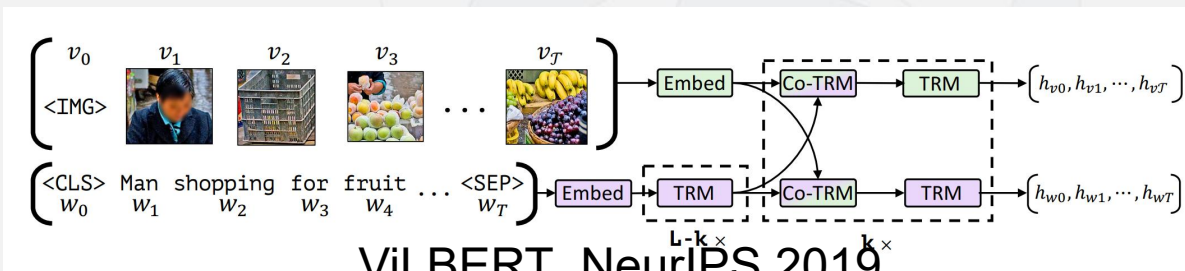
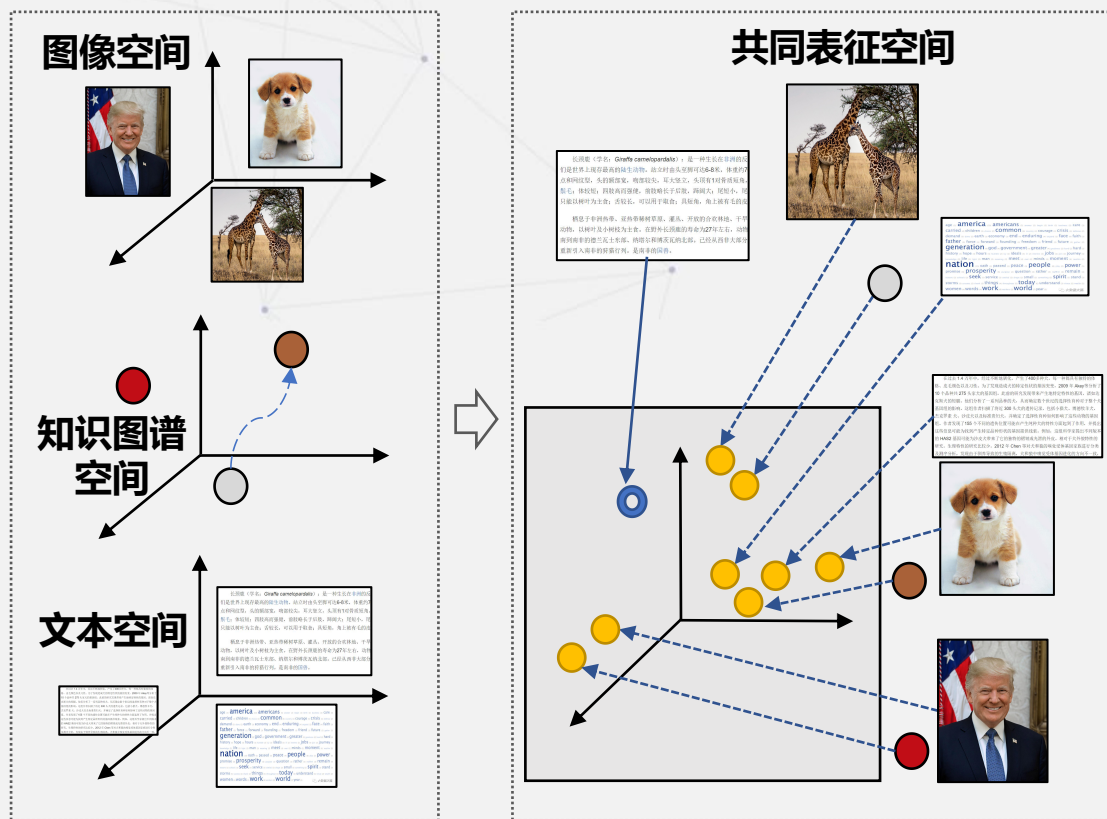


数据集	模态类型	跨模态语义关系	领域
DBpedia	文本、图像	不支持	开放域
Wikidata	文本、图像	支持	开放域
IMGpedia	文本、图像	支持	开放域
MMKG	文本、图像	支持	开放域
KgBench	文本、图像	支持	开放域
Richpedia	文本、图像	支持	开放域
知识森林	文本、图像、视频	支持	教育
百度知识图谱	文本、语音图像、视频	支持	开放域

跨模态语义关系建模是重点

研究目标3: 多模态科学数据的知识关联抽取融合与表征

(7)多模态领域本体和语料预训练模型:



Contrastive Learning

Efficiency **细粒度表征** Prompt
Few-shot Effect Analysis

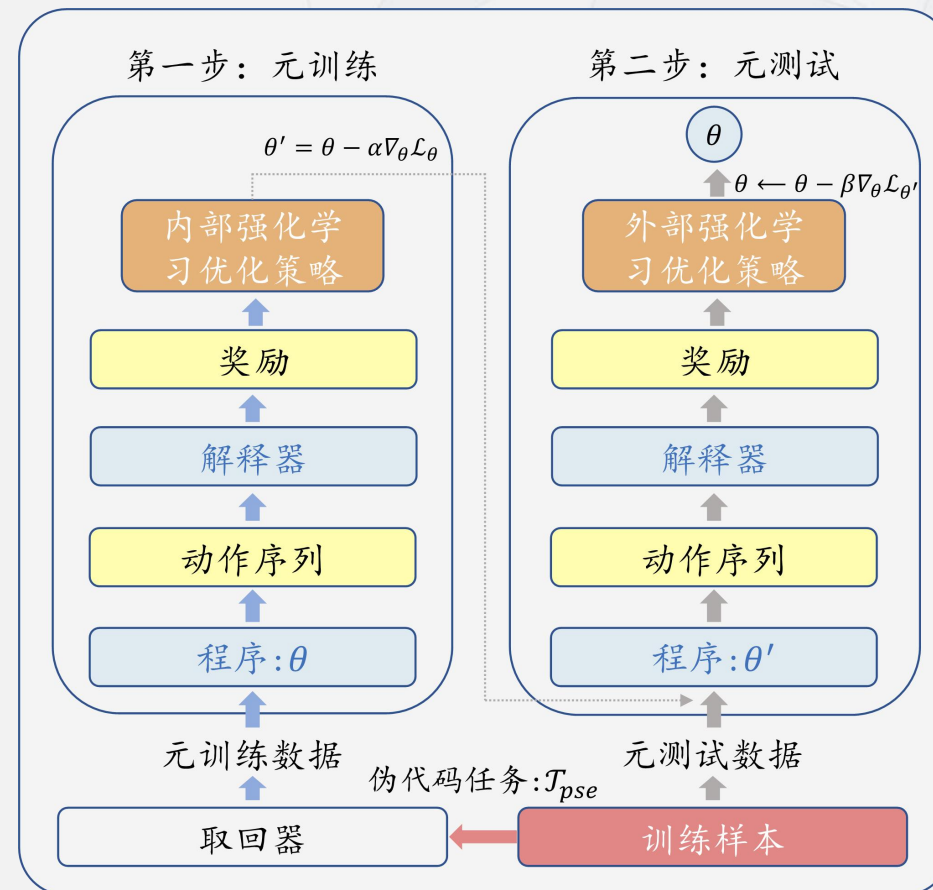
Multimodal machine learning: A survey and taxonomy. IEEE transactions on pattern analysis and machine intelligence 41.2 (2018): 423-443.

多模态细粒度知识提取关联

技术方案-细粒度天文知识抽取

在多模态数据统一表征建模的基础上，考虑不同任务场景中数据的表征需求和表达语义也存在差异性和动态性，设计一种基于对比学习与元框架的“学习-优化”多模态动态小样本知识抽取方法：

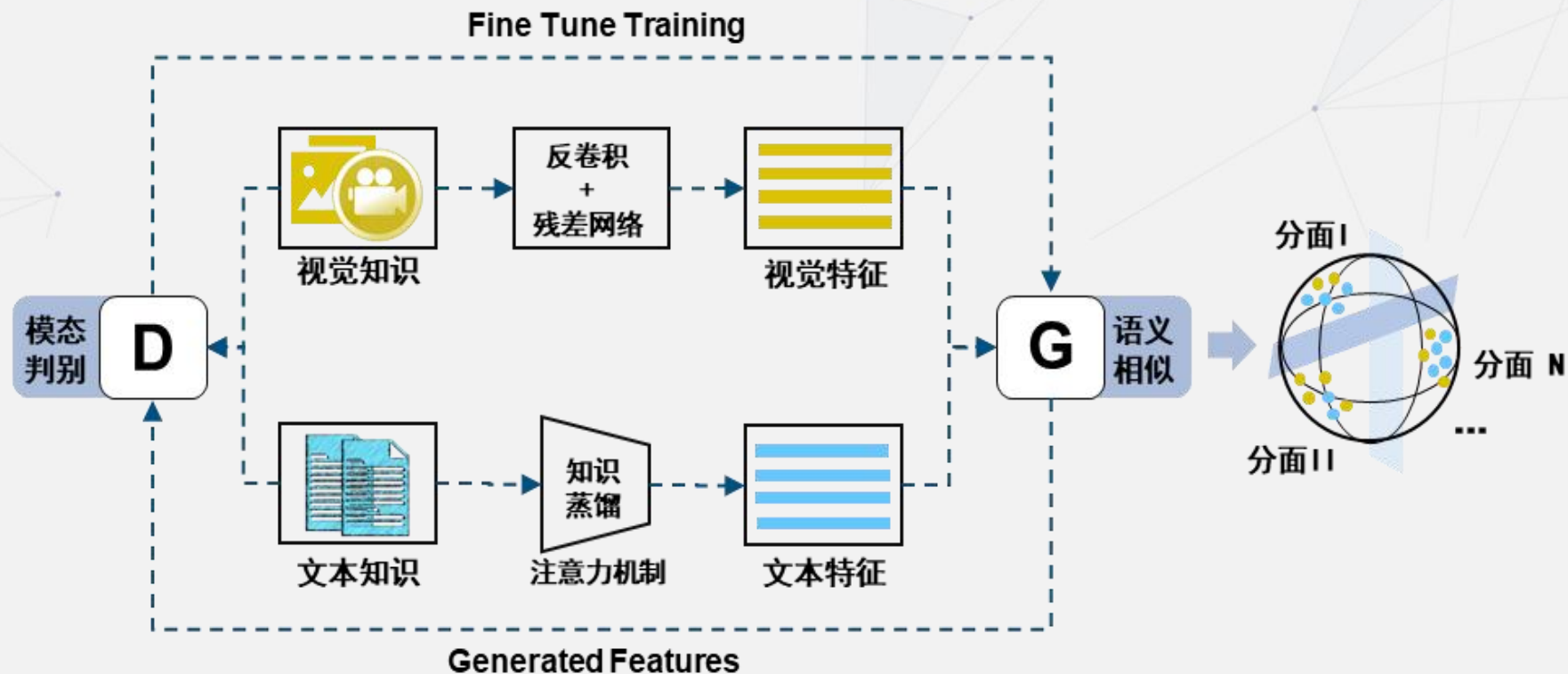
- 通过对比不同模态对于天文信息的表达，学习到知识在不同模态之间的共性表达
- 利用元抽取机制和预训练模型生成表征学习训练实例相似集，解决训练过程中的小样本或零样本问题。



多模态细粒度知识提取关联

技术方案-天文知识的分面融合

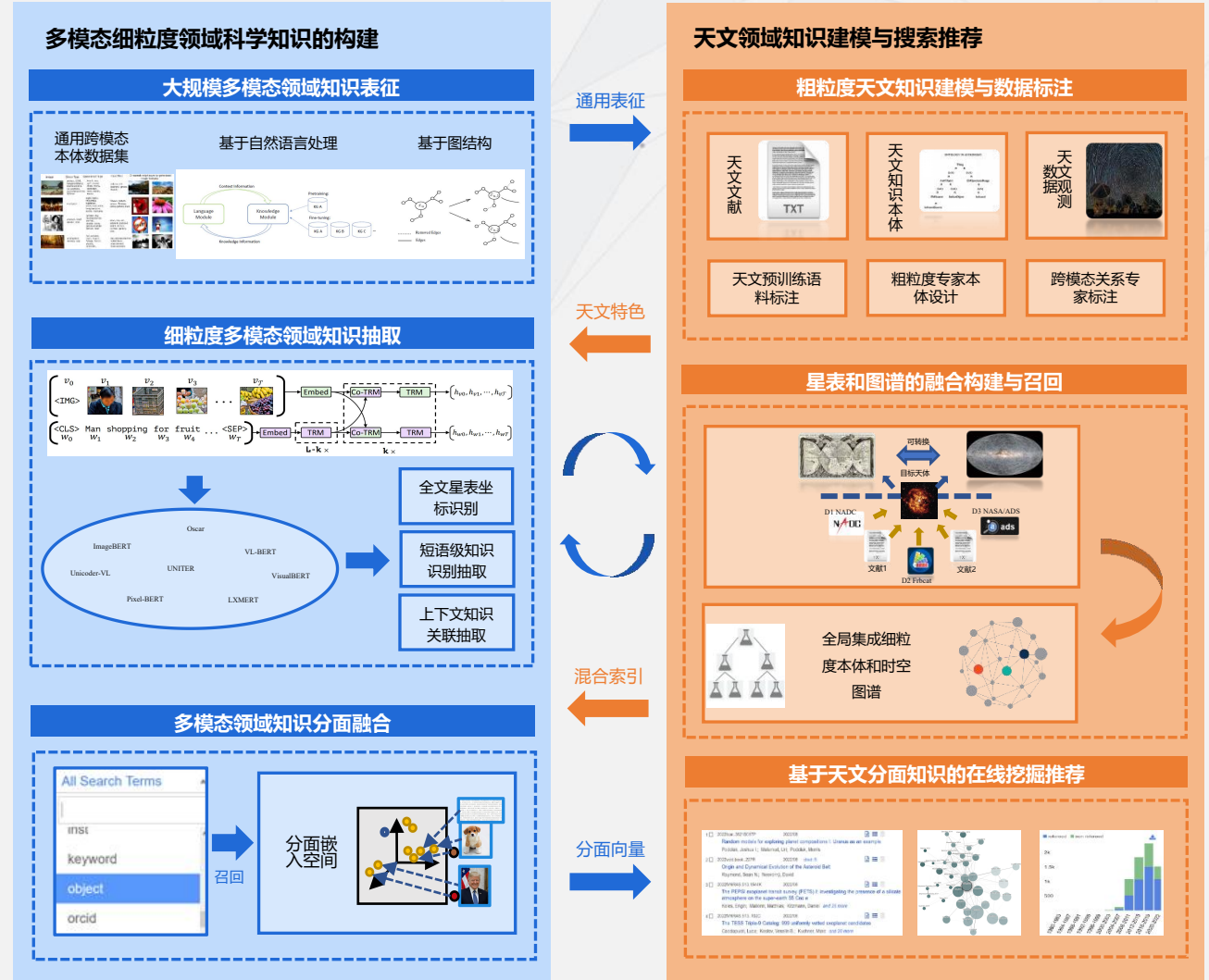
- 针对跨模态数据底层模态异构高层语义相关的特点,考虑当前天文多模态智能搜索、推荐场景,设计基于多视图特征迁移学习的跨模态知识融合算法。
- 针对属性、关系以及天文层级的知识融合,在分面层级采用基于“语义相似—模态判别”的对抗学习模式,实现面向多模态搜索和推荐完成所需的多模态异构知识分面融合。



天文领域知识建模与自适应推荐算法

技术方案:

- 基于天文专家的**数据标注和粗粒度知识本体设计**，通过多模态预训练模型进行知识抽取和关联，实现天文知识图谱的初步构建。
- 以天文领域的**数据特征**如天体分类为导引，运用多模态细粒度知识抽取技术，细化天文知识图谱。通过**星表坐标映射关联**技术，构建统一一定位的**细粒度天文时空知识图谱**。
- 基于分面融合知识的动态表征，设计基于分布式内存索引的排序算法，在分面层级实现**毫秒级的文字匹配、分面参数生成**和秒级的**多模态搜索推荐**。根据用户日志，优化时空图谱和相关模型参数。



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

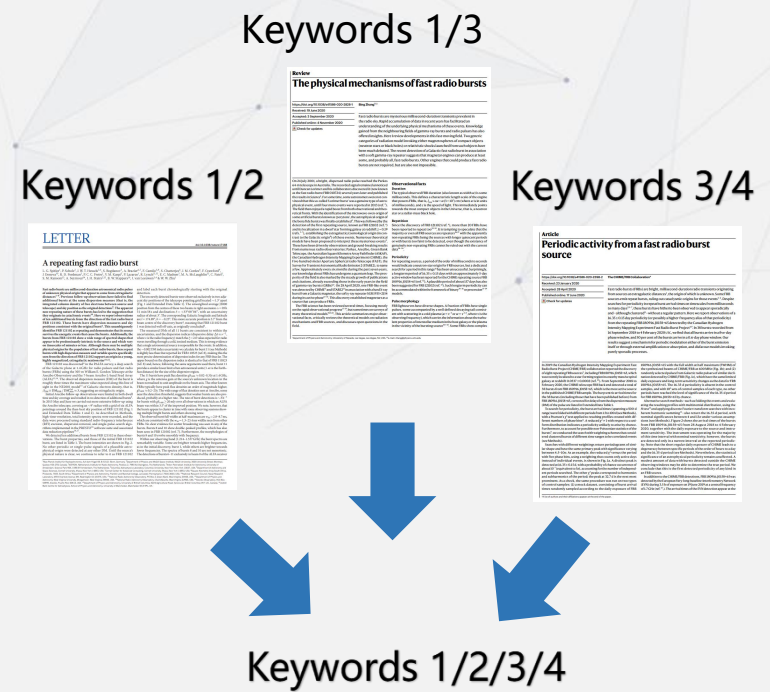
We report the first detections of the repeating fast radio burst source FRB 121102 above 5.2 GHz. Observations were performed using the 4–8 GHz receiver of the Robert C. Byrd Green Bank Telescope with the Breakthrough Listen digital backend. We present the spectral, temporal, and polarization properties 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

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 00:00:00.000	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



文献关键词提取与标注

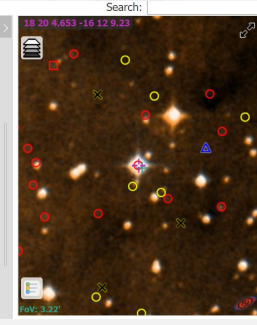
天文领域知识建模与自适应推荐算法

2. 图结构设计:

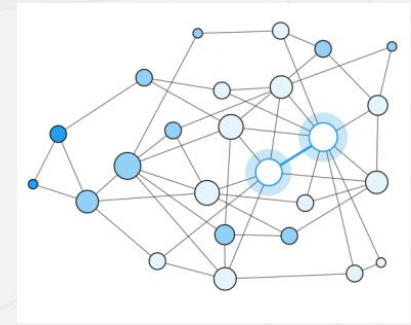
- 基于天文领域的论文中天体、数据集与观测点星表构建**细粒度天文时空知识图谱**，通过检索时空知识图谱观测点或天文现象坐标，研究周围的天体。
- 基于天文领域论文的作者与论文构建**天文领域科研社交网络图谱**。
- 基于天文领域权威分类体系与专家经验**设计关键词图谱的本体架构**，同时，基于关键词在文章中的共现关系，构建领域内关键词子图，用于挖掘领域中以及领域之间的热点研究方向。

Show 100 entries

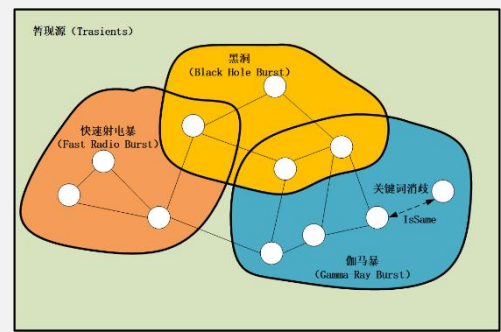
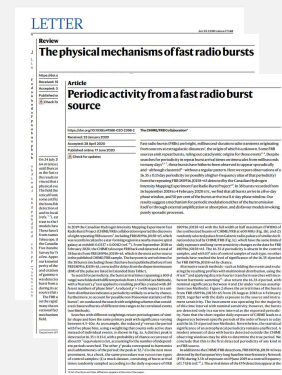
N	Identifier	dist(asec)	Otype	ICRS (J2000) RA	ICRS (J2000) DEC	Mag U
1	NGC 6618 206	3.79	*	18 20 04.6531643256	-16 12 09.228889728	
2	CXOU J182004.9-161226	16.92	Y??	18 20 04.959	-16 12 26.35	
3	[TBH2012] 631	24.32	cor	18 20 03.4	-16 12 31	
4	AGAL G014.979-00.609	29.70	Y*O	18 20 03	-16 12.6	
5	2MASS 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

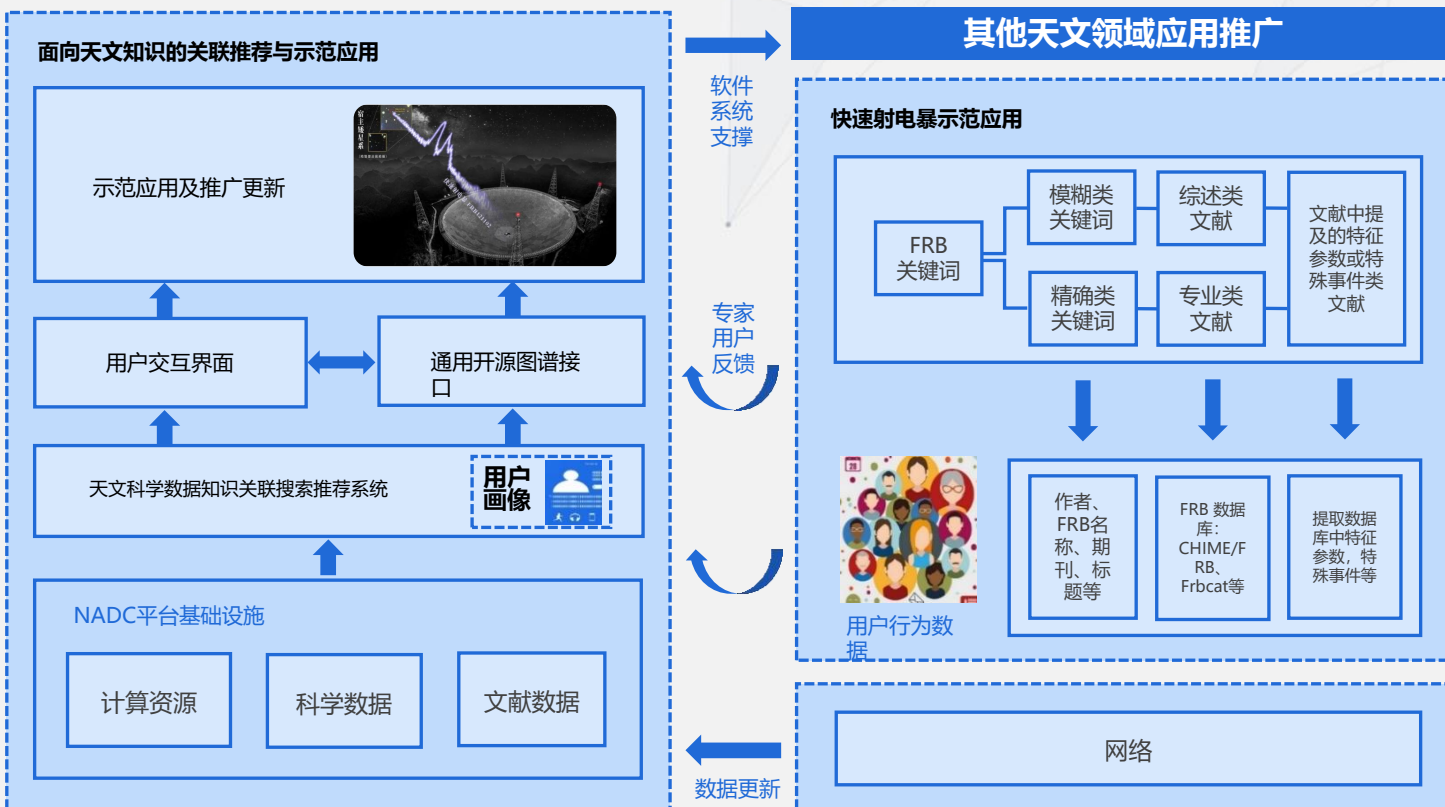


天文关键词图谱

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

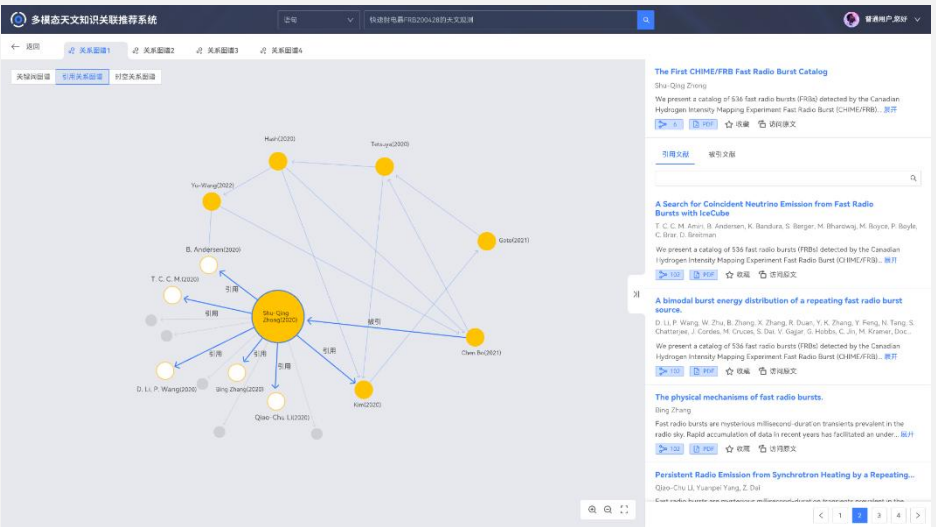
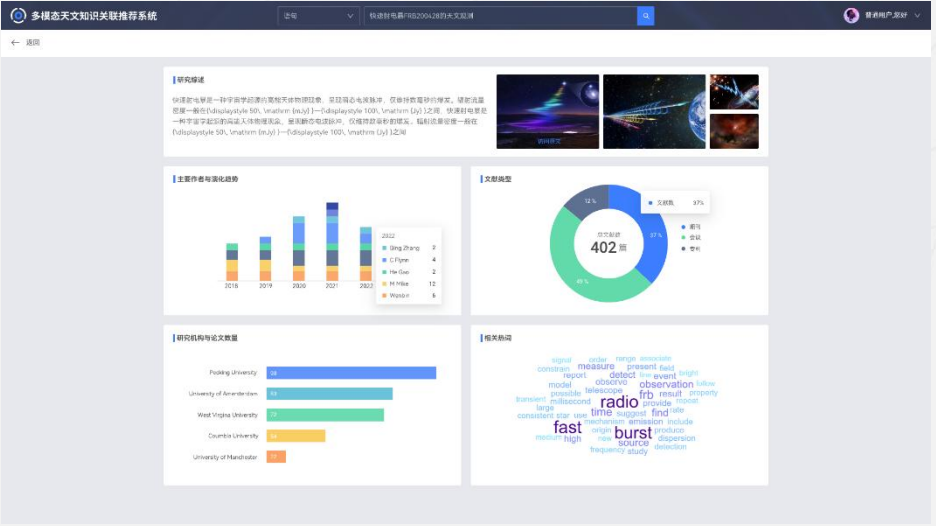
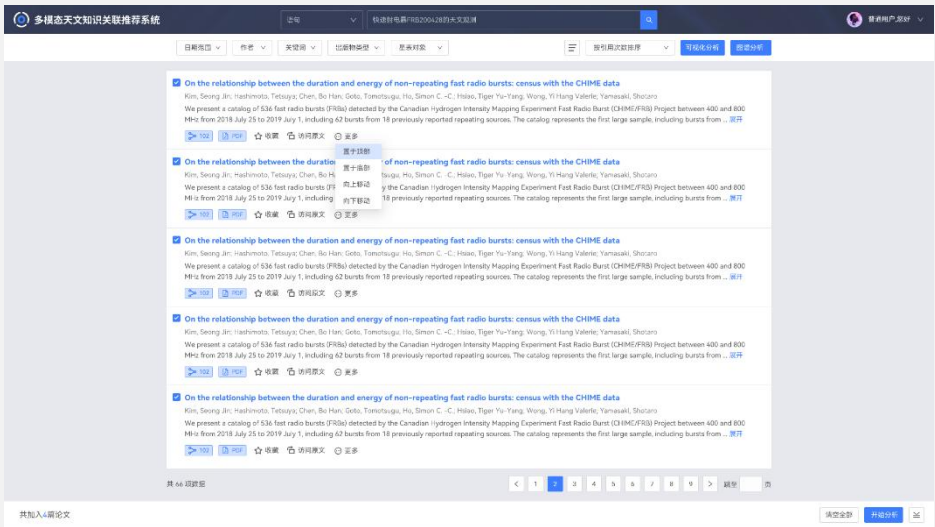
1. 系统在国家天文科学数据中心平台集成部署:

- 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:





请帮我们选出系统名称

1. 天问 (Astro Inquiry System)
2. 巡天 (AstroSearcher)
3. 洞天 (SkyInsight)
4. 天网 (Skynet)
5. 星河 (Galaxy)

谢谢参与我们的问卷调查!