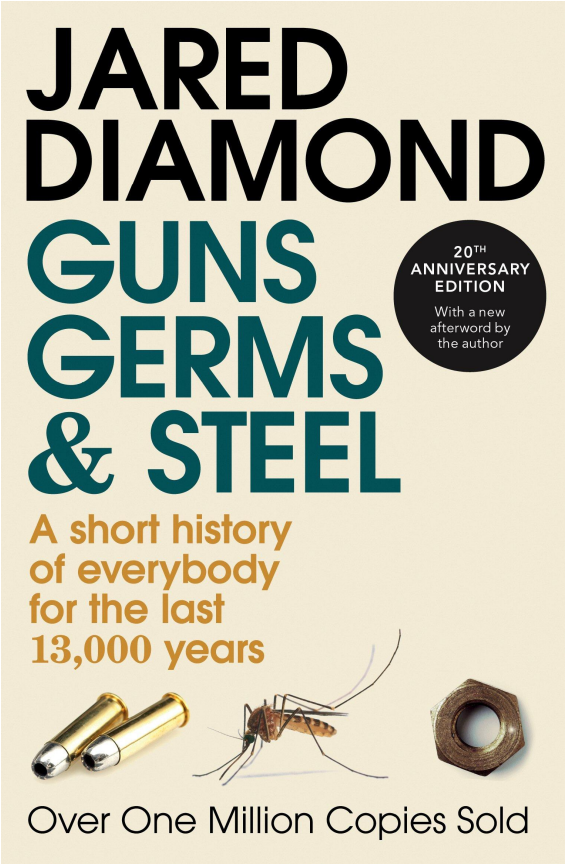
ΧΑΡΟΚΟΠΕΙΟ ΠΑΝΕΠΙΣΤΗΜΙΟ
HAROKOPIO UNIVERSITY

FORTH
Institute of Computer Science

Vasilis Efthymiou

# BITS STARS & KNOWLEDGE AI at the Frontiers of Astrophysics

SummerSoC
Service Oriented Computing


JARED
DIAMOND
GUNS
GERMS
& STEEL
20TH
ANNIVERSARY
EDITION
With a new
afterword by
the author
A short history
of everybody
for the last
13,000 years
Over One Million Copies Sold

Bits

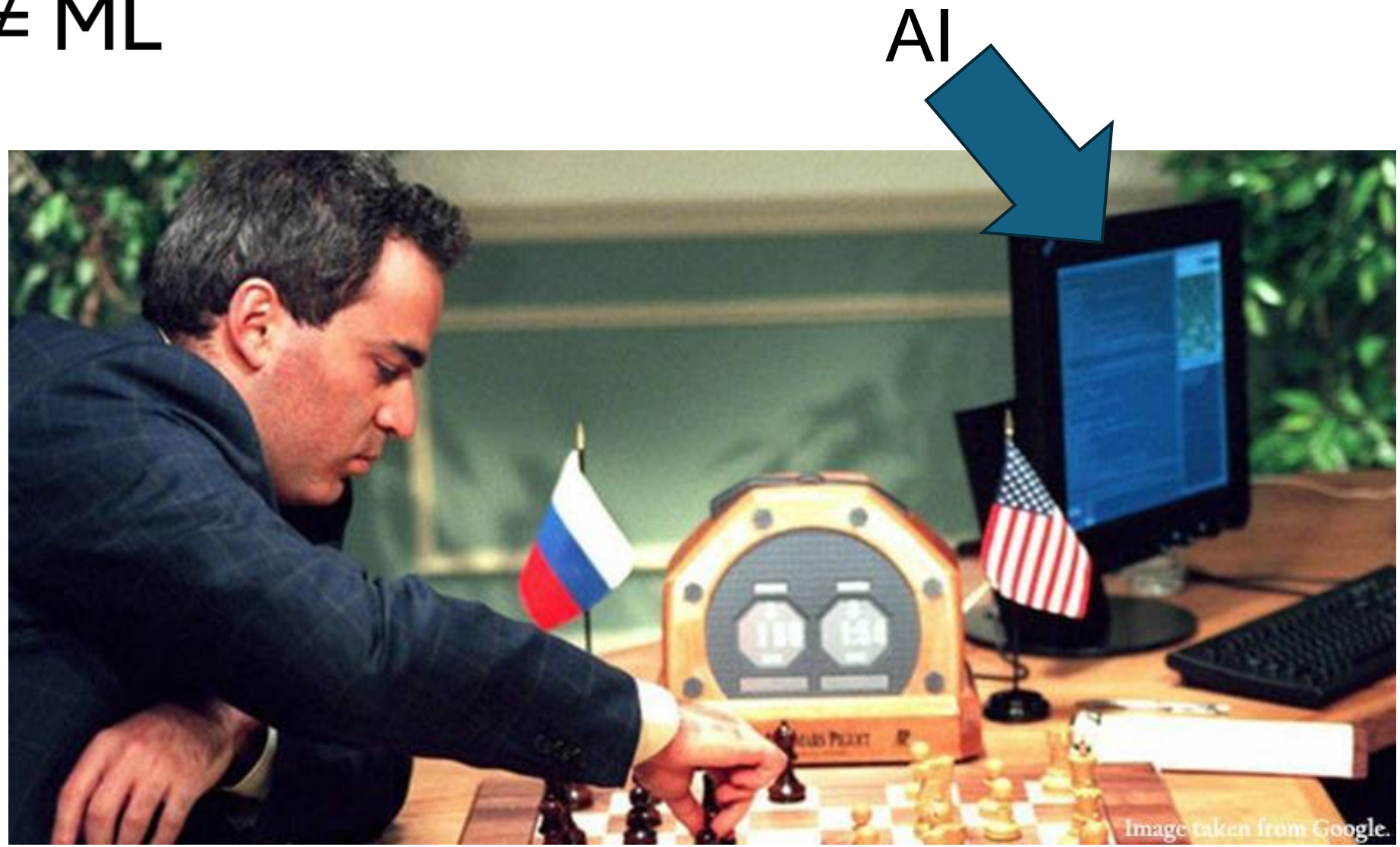
*"What would it take for a Foundation Model that understands (the lifecycle of) our Universe?"*

Knowledge

Stars

# Not there yet…

# AI ≠ ML



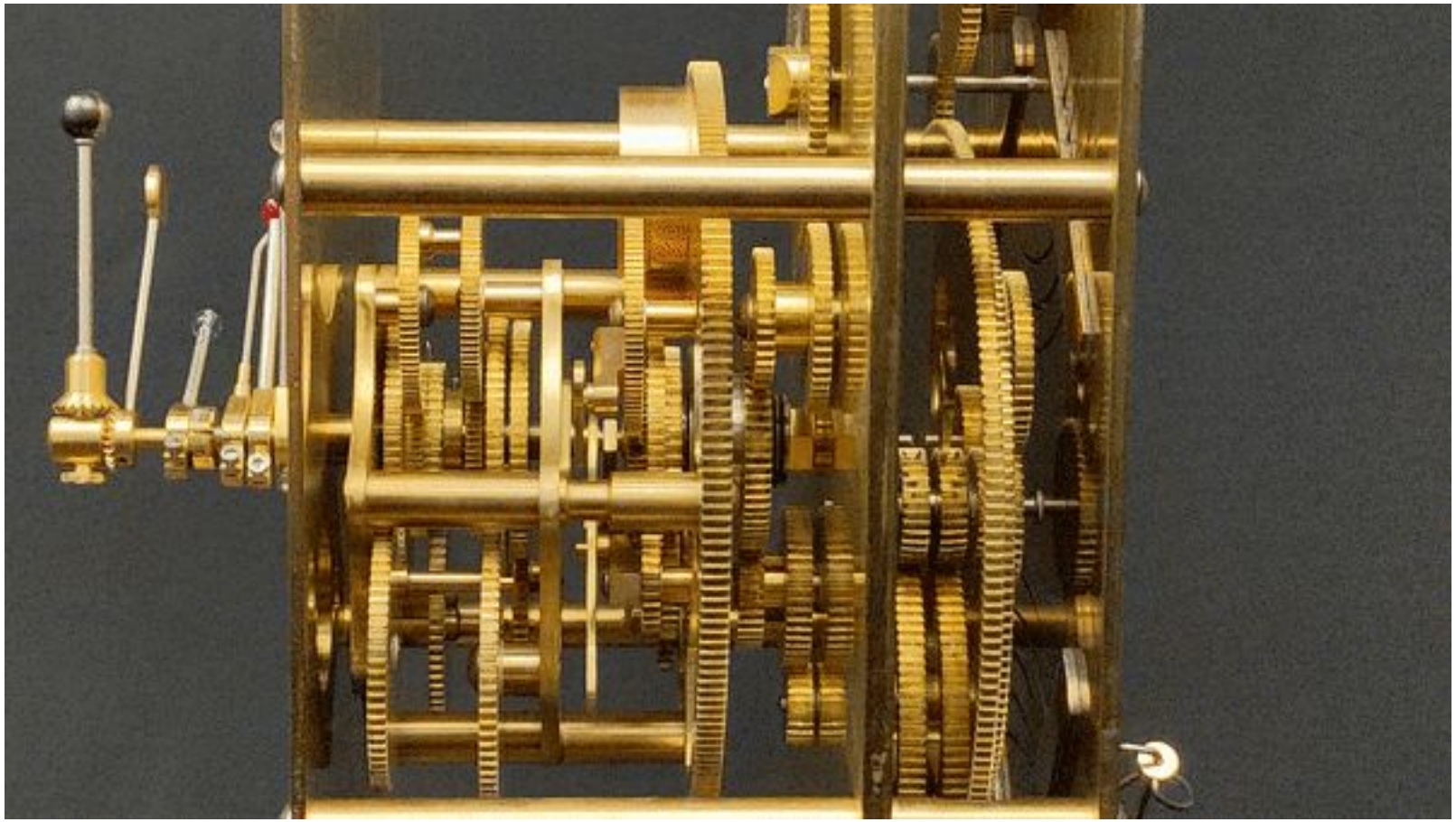
# Outline

- “Bits” (~ AI)
  - latest advances in deep learning
- “Stars” (~ astrophysics)
  - challenges in “traditional” astrophysics

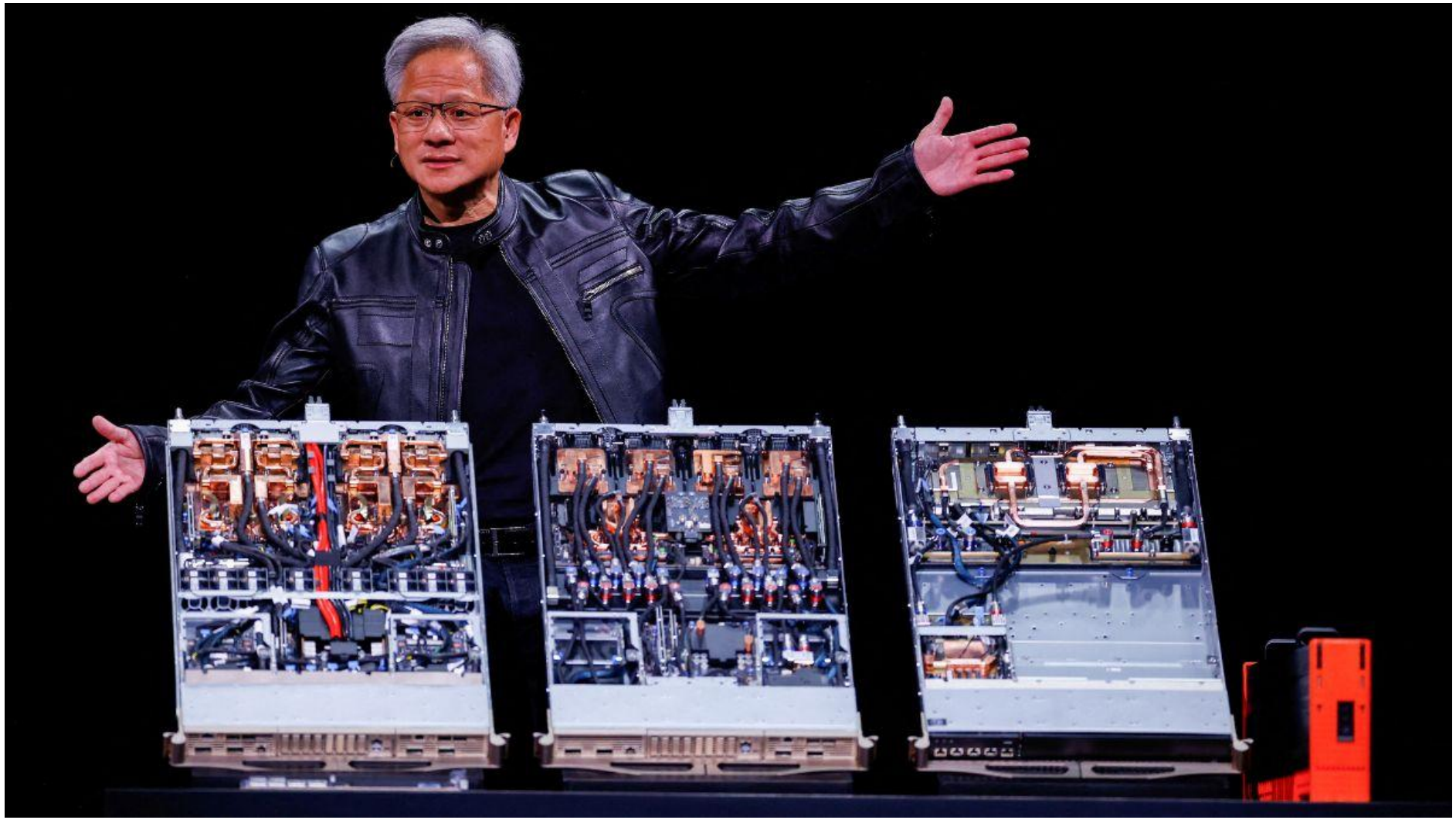
 UniversAI: Exploring the Universe with AI

- Knowledge…
  - … representation (knowledge graphs)
- The PARSEC project

# Bits



# … more bits



# Recent advances in AI

- Foundation models for science
  - SciBERT / Galactica / Claude
  - Retrieval-augmented generation (RAG) for data (e.g., catalog) querying
- Generative models for simulations
  - Diffusion models / GANs
- Graph Neural Networks / Graph Transformers
  - Detecting matches/missing links across one or more data sources
- Counterfactuals
  - Understanding how AI models (would have) work(ed)

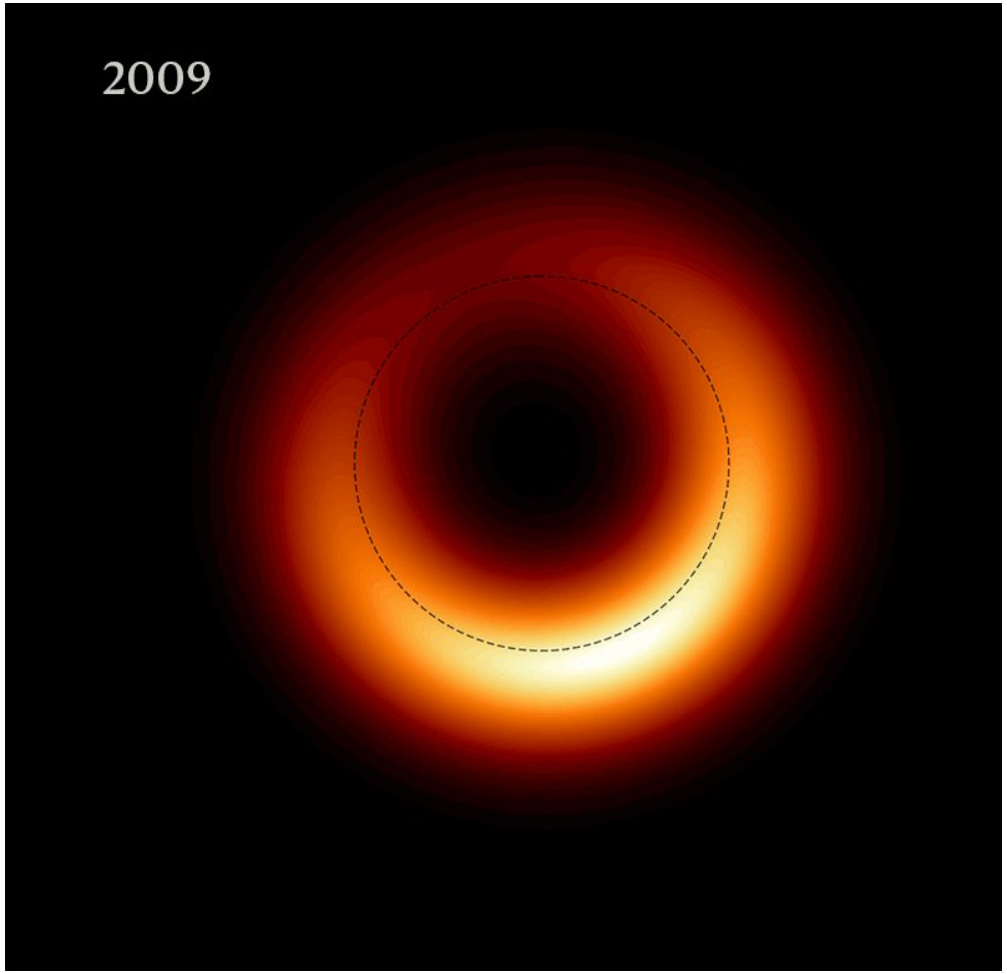
# Stars

# … more stars


will start in 2025
20 TB per night
60 PB in total

# AI & Astrophysics

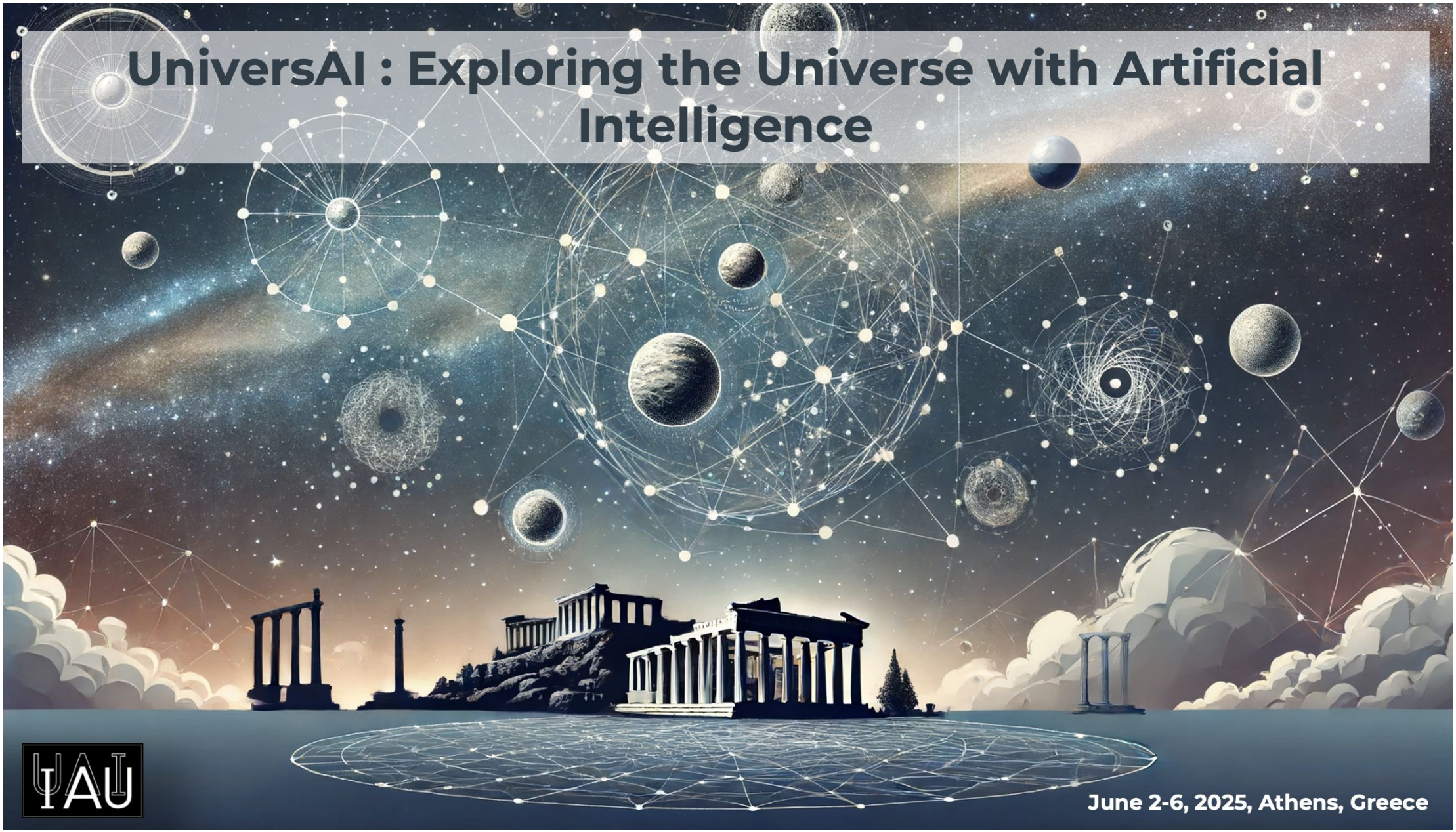
**The first-ever image of a black hole is now a movie…**




https://www.nature.com/articles/d41586-020-02717-3

# Recent advances in astronomy

- JWST & Multi-wavelength data fusion
  - Heterogeneous, high-res datasets across wavelengths and missions
  - Opportunity: how do we align, interpret and reason about them?
- Astroinformatics
  - Cosmological simulations now produce petabytes of synthetic universes
  - Opportunity: indexing, querying, similarity computations
- Citizen science
  - Galaxy zoo: large crowdsourced labels *(also: largely inconsistent)*
  - Opportunity: weakly/semi-supervised learning


UniversAI : Exploring the Universe with Artificial Intelligence
IAU
June 2-6, 2025, Athens, Greece

# What led to UniversAI?

- Cultural cross-talk
  - What is a “source”?
  - … “survey” vs “review”
  - … “referee” vs “reviewer”
  - … conference vs journal papers
- Falling behind the Big (AI) Bang / space missions / data repositories

- Idea: Let’s bring the best of the two communities together

# Invited talks

- **Jean-Luc Starck** *(CEA –Saclay)*
- **Christos Diou** *(HUA)*
- **Torsten Enßlin** *(MPA)*
- **Themis Palpanas** *(French University Institute)*
- **Tyson Littenberg** *(NASA)*
- **Angela Bonifati** *(Lyon 1 University)*
- **Meghyn Bienvenu** *(CNRS)*
- **Laurent Eyer** *(Geneva Astronomical Observatory)*
- **Federica Bianco** *(University of Delaware)*
- **Georgia Koutrika** *(Athena RC)*
- **Hendrik Müller** *(NRAO)*
- **Grigoris Tsagatakis** *(FORTH)*

# UniversAI – Highlights
# T. Littenberg – GW and AI

**The co-evolution of gravitational wave astrophysics and artificial intelligence**

Tyson B. Littenberg @ NASA Marshall Space Flight Center

Noise Reduction

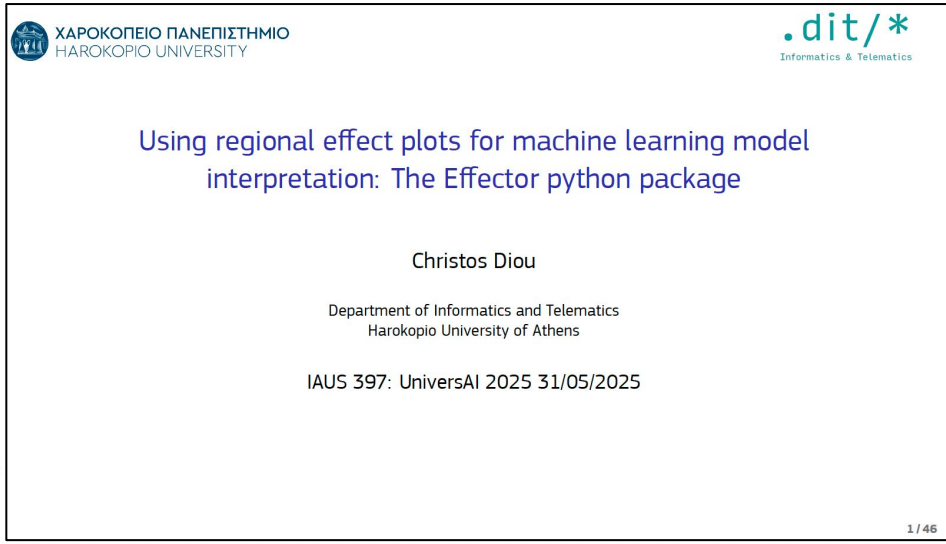
Computational Efficiency

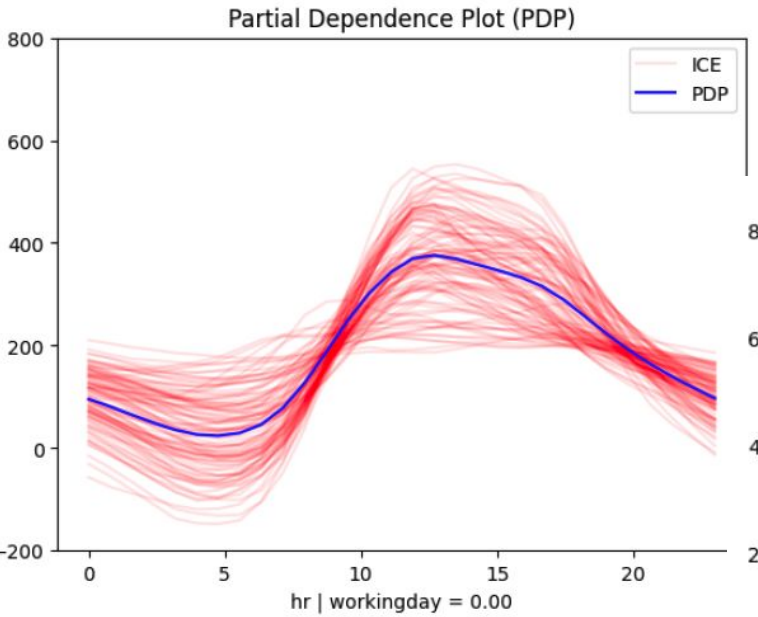
Event Classification

Source Extraction

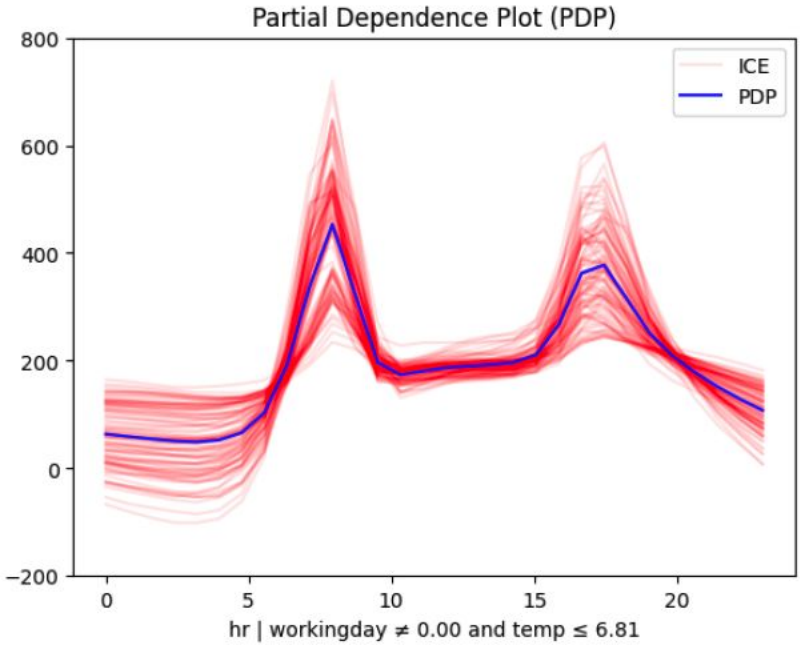
**For GW detection, interpretability is *critical*. Critical.**

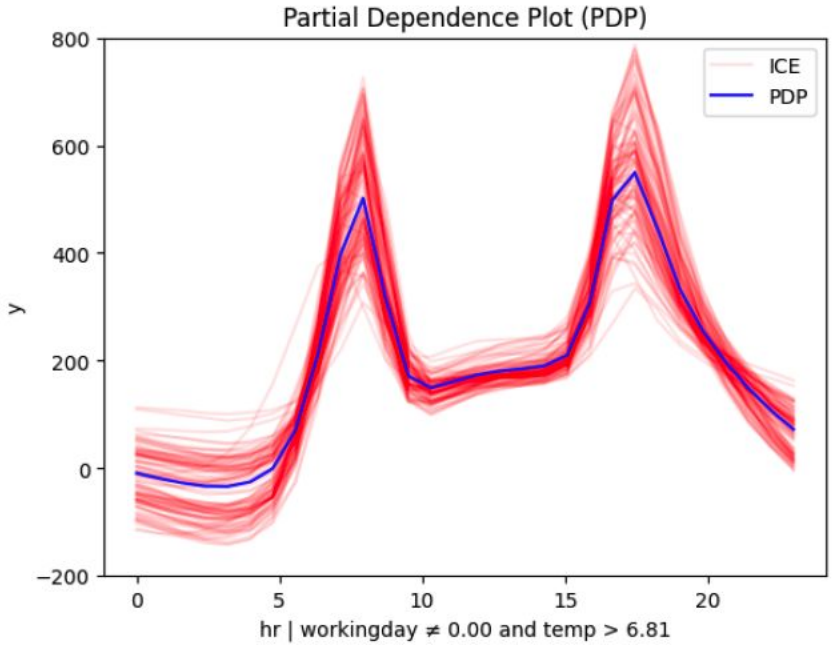
# UniversAI – Highlights
# C. Diou - XAI







# UniversAI – Highlights
# T. Palpanas – Anomaly Detect

**Machine Learning on Very Large Time Series Collections**

**Similarity Search and Subsequence Anomaly Detection**

Themis Palpanas

*Université Paris Cité*
*French University Institute*

diNo 381

## Subsequence Anomaly Detection

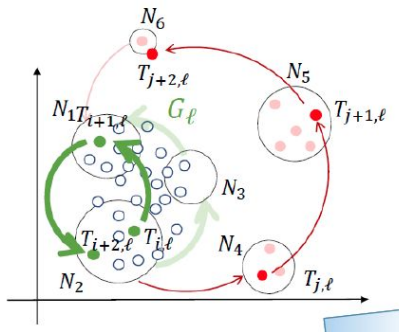
Publications

Boniol - PVLDB'20

### Proposed Approach: Series2Graph

Graph $G_\ell$ :

Given a time series $T$, and an input length $\ell$, we build a graph $G_\ell(\mathcal{N}, \mathcal{E})$, for which:




For instance:

The subsequence $T_{i,\ell+2}$, is a path in the graph $\mathrm{P_{th}}(T_{i,\ell+2}) = < \mathrm{T}_{i,\ell}, \mathrm{T}_{i+1,\ell}, \mathrm{T}_{i+2,\ell} > = < \mathrm{N}_4, \mathrm{N}_5, \mathrm{N}_6 >$ in $G_\ell$.

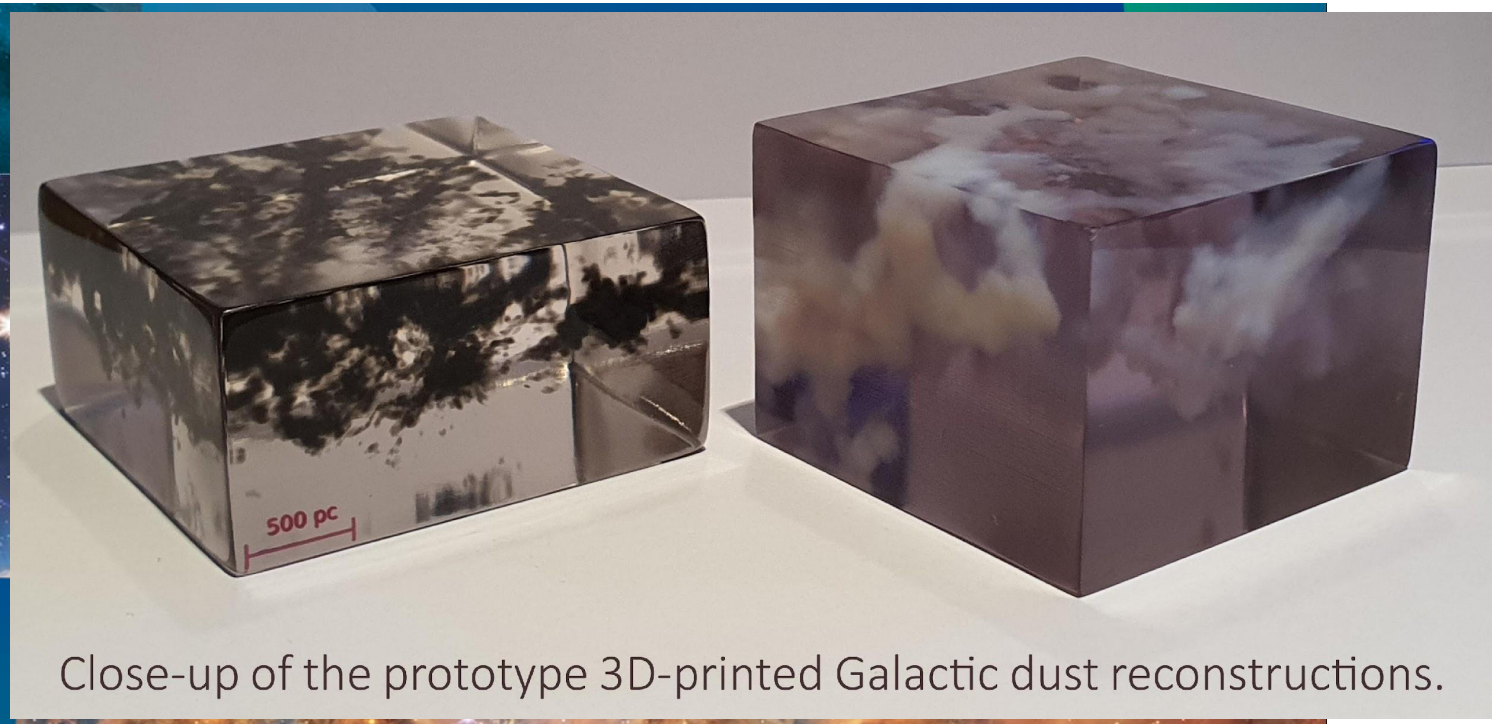
The subsequence $T_{j,\ell+2}$, is a path in the graph $\mathrm{P_{th}}(T_{j,\ell+2}) = < \mathrm{T}_{j,\ell}, \mathrm{T}_{j+1,\ell}, \mathrm{T}_{j+2,\ell} > = < \mathrm{N}_2, \mathrm{N}_1, \mathrm{N}_2 >$ in $G_\ell$.

$$\mathrm{Norm}\left(\mathrm{P_{th}}(T_{j,\ell+2})\right) \ll \mathrm{Norm}\left(\mathrm{P_{th}}(T_{i,\ell+2})\right)$$

unsupervised

Themis Palpanas - UniversAI - Jun 2025

# UniversAI – Highlights
# T. Enßlin – Milky Way Atlas


Close-up of the prototype 3D-printed Galactic dust reconstructions.

# UniversAI – Highlights
# … and many more


Diffusion Models for Emulating the Large-Scale Structure of the Universe in Modified Gravity Cosmologies

E. García-Farieta
Universidad de Córdoba
n: JK Riveros, P. Saavedra, H.J. Hortúa, I. Olier

Vision-Language Models for Radio Astronomy

Simone Riggi
simone.riggi@inaf.it
INAF
ISTITUTO NAZIONALE DI ASTROFISICA

Do Androids Dream of Exploding Stars and Receding Galaxies?

federica b. b
she/her
University of Delawa
Department of Physics and
Biden School of Public Policy and
Data Science Institu

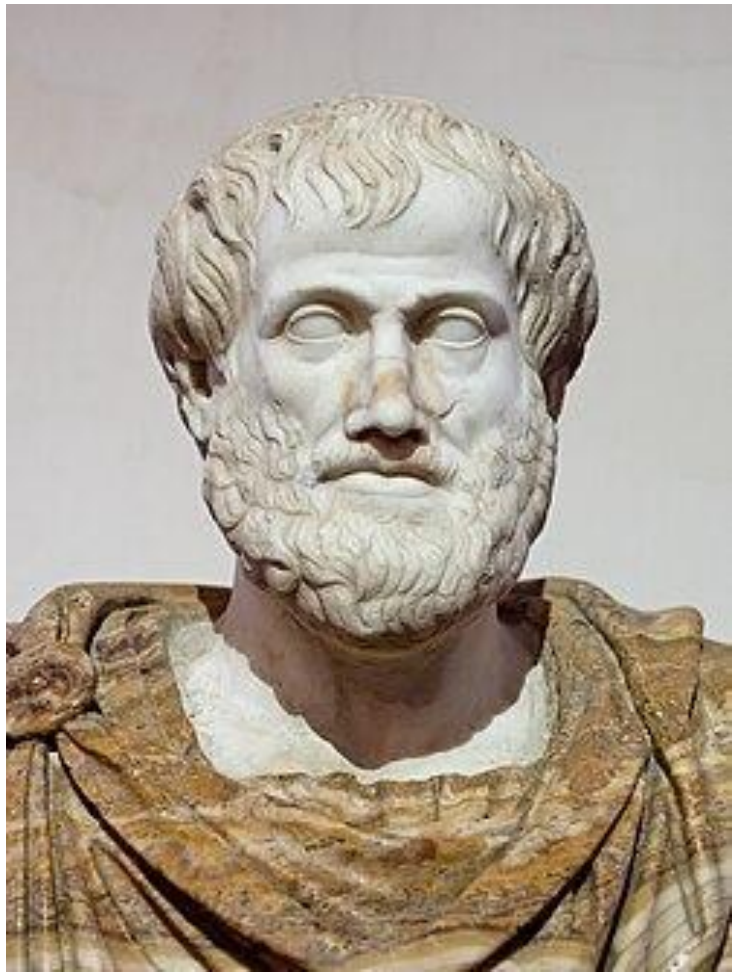
FALCO:
a Foundation model of Astronomical Light Curves for time dOmain astronomy

Yihan Tao
National Astronomical Observatories, Chinese Academy of Sciences
Xiaoxiong Zuo
LMU Munich, University Observatory / National Astronomical Observatories, Chinese Academy of Sciences

In collaboration with: Yang Huang, Zhixuan Kang, Huaxi Chen, Chenzhou Cui, Jashu Pan, Xiao Kong, Xiaoyu Tang, Henggeng Han, Haiyang Mu, Yuan-sen Ting, Yunfei Xu, Dongwei Fan, Guirong Xue, Ali Luo and Jifeng Liu

NADC National Astronomical Data Center 国家天文科学数据中心 NAOC 2025-06-05 · Athens · UniversAI2025 之江实验室 ZHEJIANG LAB

# Knowledge



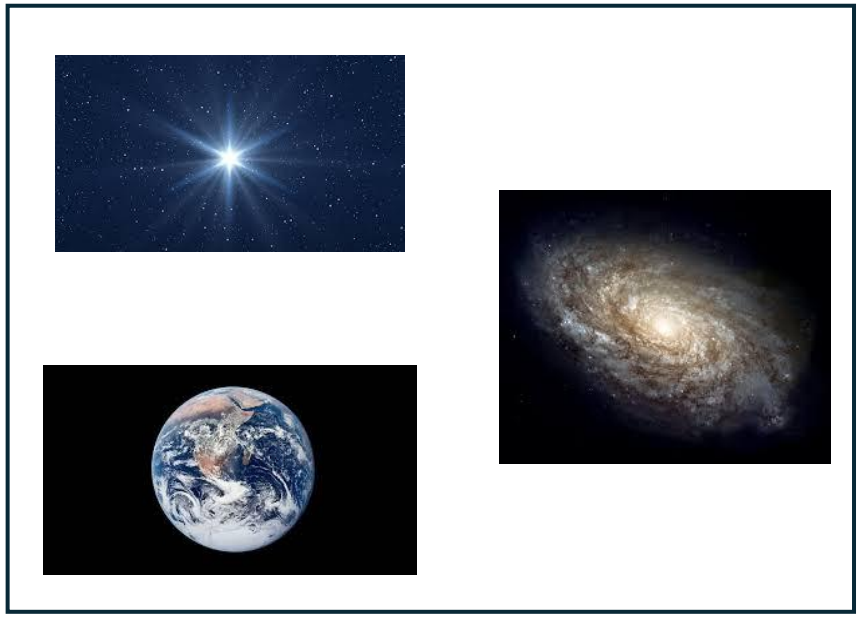
**Syllogism**:
All men are mortal.
Socrates is a man.
□
Socrates is a mortal.

# … more knowledge

# Knowledge Graphs (KGs)

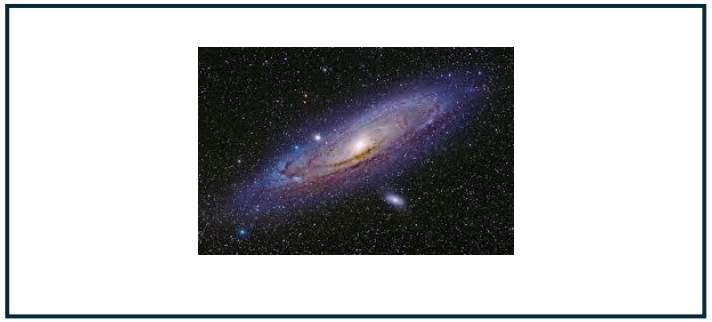
- **Class**: A group of entities that share common properties or characteristics. i.e., **Celestial bodies.**
- **Subclass**: Specific type of a broader class. i.e., **Galaxies.**
- **Individual/Instance**: Something or someone belonging to a class. i.e., **Milky Way.**

**Class**: **Celestial Bodies**



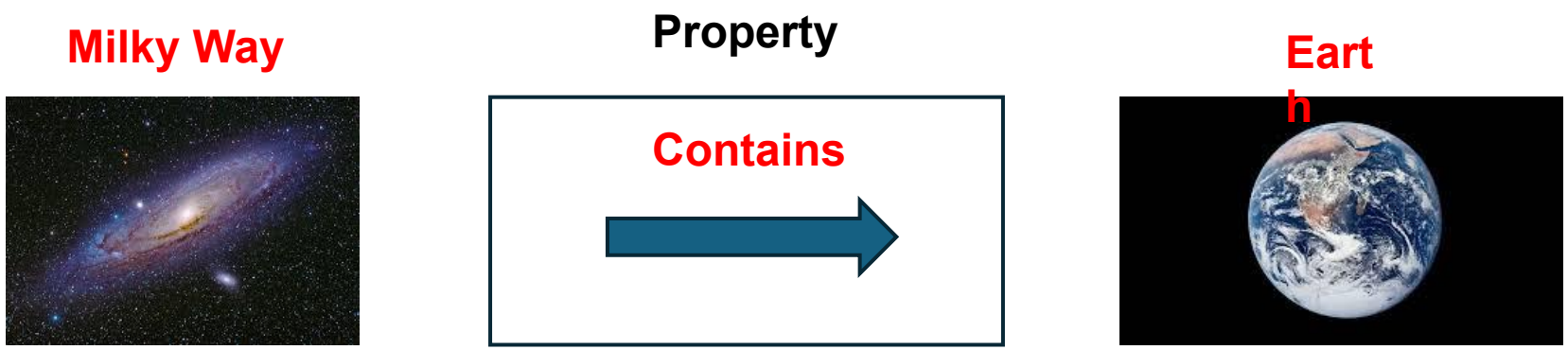
**Subclass**: **Galaxies**

**Individual**: **Milky Way**

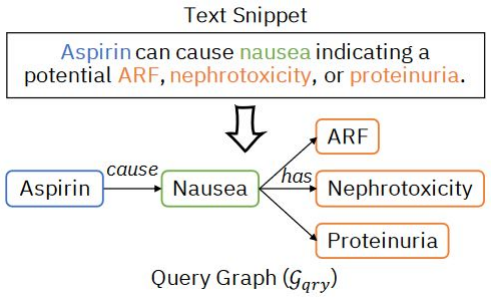


# Knowledge Graphs (KGs) (2)

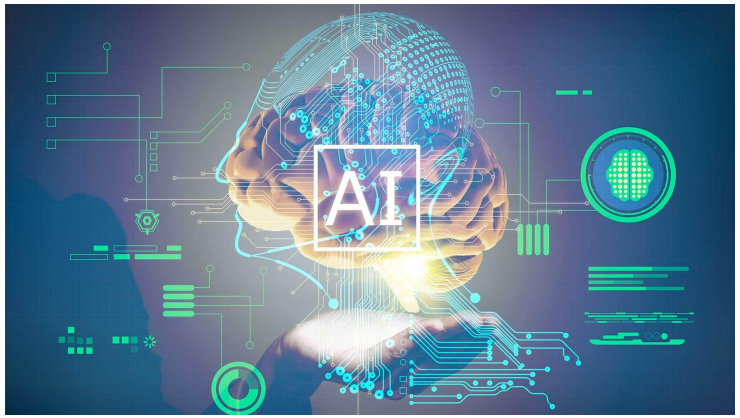
**Properties**: The attributes that connect individuals.

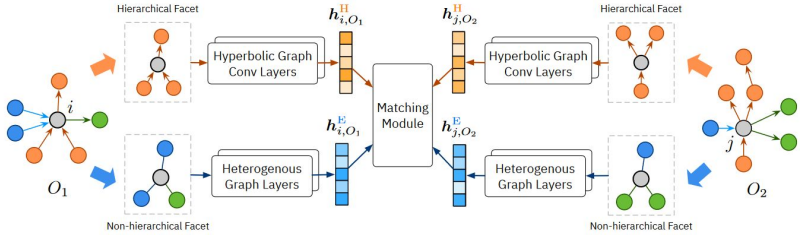


# Exploiting KGs for deep learning tasks

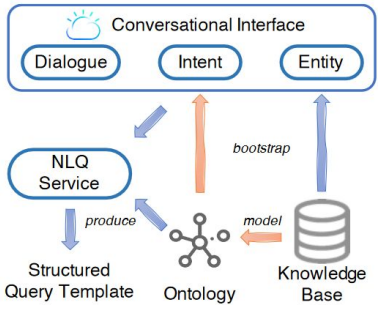



entity disambiguation




ontology matching




chatbots

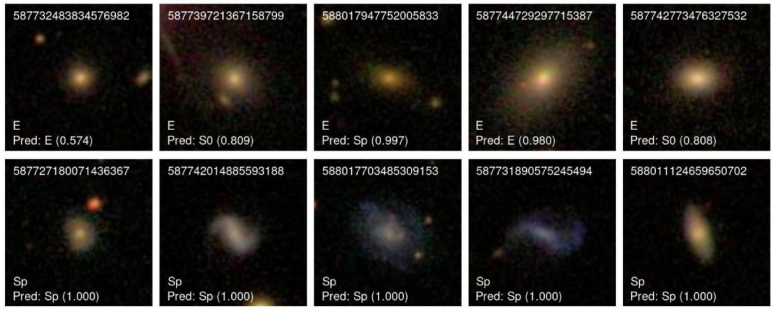
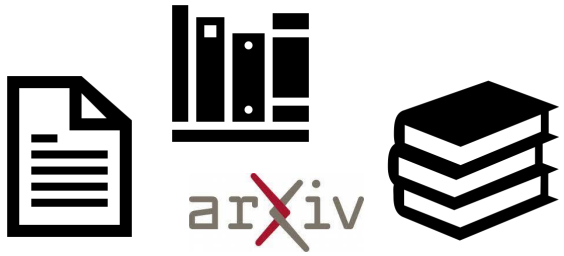



image classification

# Motivation




read papers

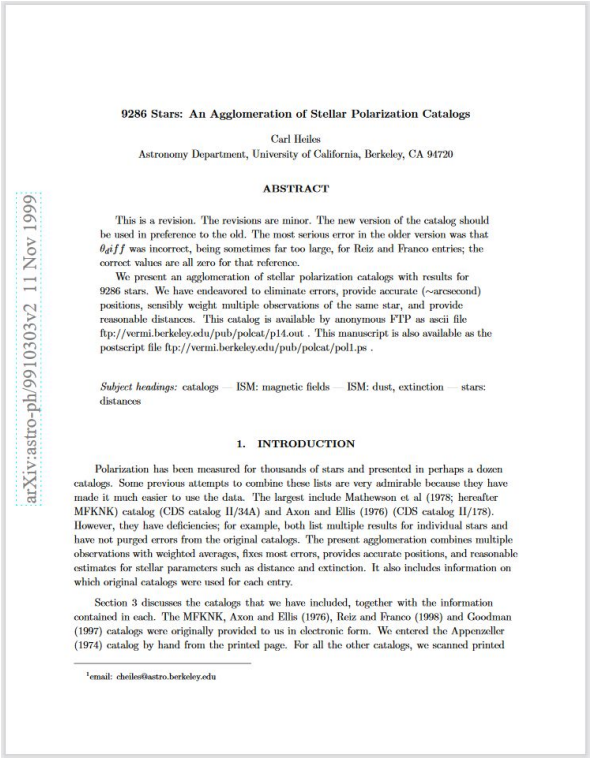
**Table B.1.** Spin and external pressure gradient estimates for blazars. Columns: (1) Name as given in H09, L17, (2) Alternative name, (3) Class (B is for BL Lacs, F for FSRQs), (4) Lorentz factor, (5) Spin, (6) External pressure gradient, (7) Reference for the Lorentz factor estimate.

| Name | Alt-name | Class | $\Gamma$ | $a$ | $s$ | Ref. |
|---|---|---|---|---|---|---|
| J0003-066 | NRAO 5 | B | 3.3 | 0.9 | 0.38 | H09 |
| J0016+731 | - | F | 6.8 | 0.59 | 0.64 | H09 |
| J0102+5824 | 0059+5808 | F | 12.0 | 0.6 | 0.83 | L17 |
| J0106+013 | OC 012 | F | 27.8 | 0.82 | 1.07 | H09 |
| J0136+4751 | 0133+476 | F | 9.5 | 0.49 | 0.76 | L17 |
| J0202+149 | 4C 15.05 | F | 9.9 | 0.61 | 0.76 | H09 |
| J0212+735 | - | F | 7.5 | 0.57 | 0.67 | H09 |
| J0217+0144 | PKS 0215+015 | F | 19.1 | 0.75 | 0.96 | L17 |

read tables



cross-check online astronomical catalogues and DBs



**9286 Stars: An Agglomeration of Stellar Polarization Catalogs**

Carl Heiles
Astronomy Department, University of California, Berkeley, CA 94720

**ABSTRACT**

This is a revision. The revisions are minor. The new version of the catalog should be used in preference to the old. The most serious error in the older version was that $\theta_{diff}$ was incorrect, being sometimes far too large, for Reiz and Franco entries; the correct values are all zero for that reference.

We present an agglomeration of stellar polarization catalogs with results for 9286 stars. We have endeavored to eliminate errors, provide accurate (~arcsecond) positions, sensibly weight multiple observations of the same star, and provide reasonable distances. This catalog is available by anonymous FTP as ascii file ftp://vermi.berkeley.edu/pub/polcat/p14.out . This manuscript is also available as the postscript file ftp://vermi.berkeley.edu/pub/polcat/pol1.ps .

*Subject headings:* catalogs — ISM: magnetic fields — ISM: dust, extinction — stars: distances

**1. INTRODUCTION**

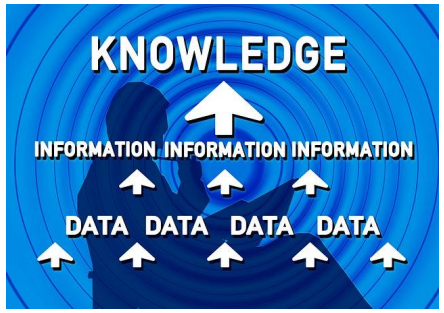
Polarization has been measured for thousands of stars and presented in perhaps a dozen catalogs. Some previous attempts to combine these lists are very admirable because they have made it much easier to use the data. The largest include Mathewson et al (1978; hereafter MFKNK) catalog (CDS catalog II/34A) and Axon and Ellis (1976) (CDS catalog II/178). However, they have deficiencies; for example, both list multiple results for individual stars and have not purged errors from the original catalogs. The present agglomeration combines multiple observations with weighted averages, fixes most errors, provides accurate positions, and reasonable estimates for stellar parameters such as distance and extinction. It also includes information on which original catalogs were used for each entry.

Section 3 discusses the catalogs that we have included, together with the information contained in each. The MFKNK, Axon and Ellis (1976), Reiz and Franco (1998) and Goodman (1997) catalogs were originally provided to us in electronic form. We entered the Appenzeller (1974) catalog by hand from the printed page. For all the other catalogs, we scanned printed

[1]email: cheiles@astro.berkeley.edu

arXiv:astro-ph/9910303v2 11 Nov 1999

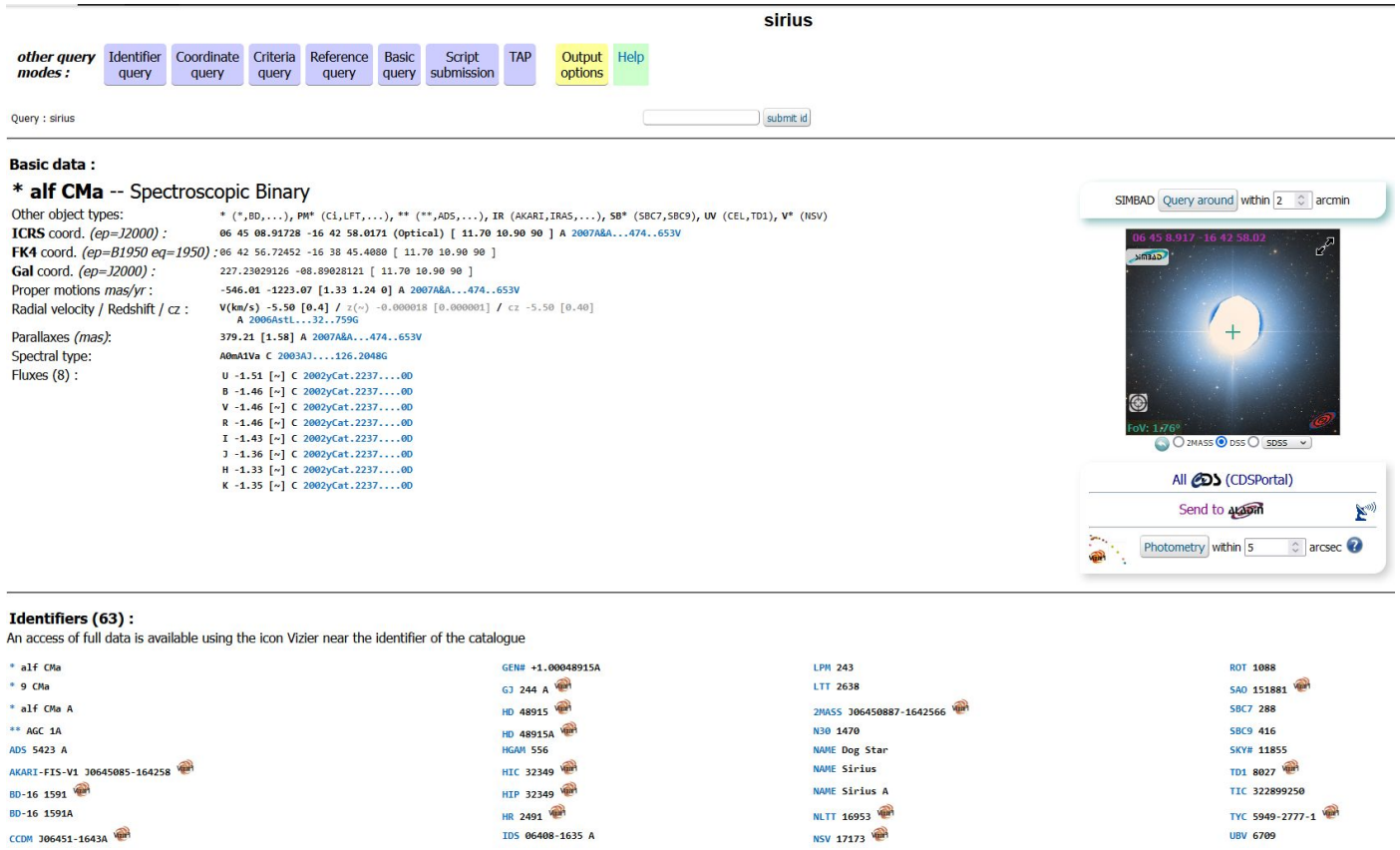
publish combined data and findings on them




combine data


*Kovlakas K., Zezas A., Andrews J. J., Basu-Zych A., Fragos T., Hornschemeier A., Kouroumpatzakis K., Lehmer B., Ptak. A (2021). "The Heraklion Extragalactic Catalogue (HECATE): a value added galaxy catalogue for multi-messenger astrophysics". MNRAS*

# The good news…



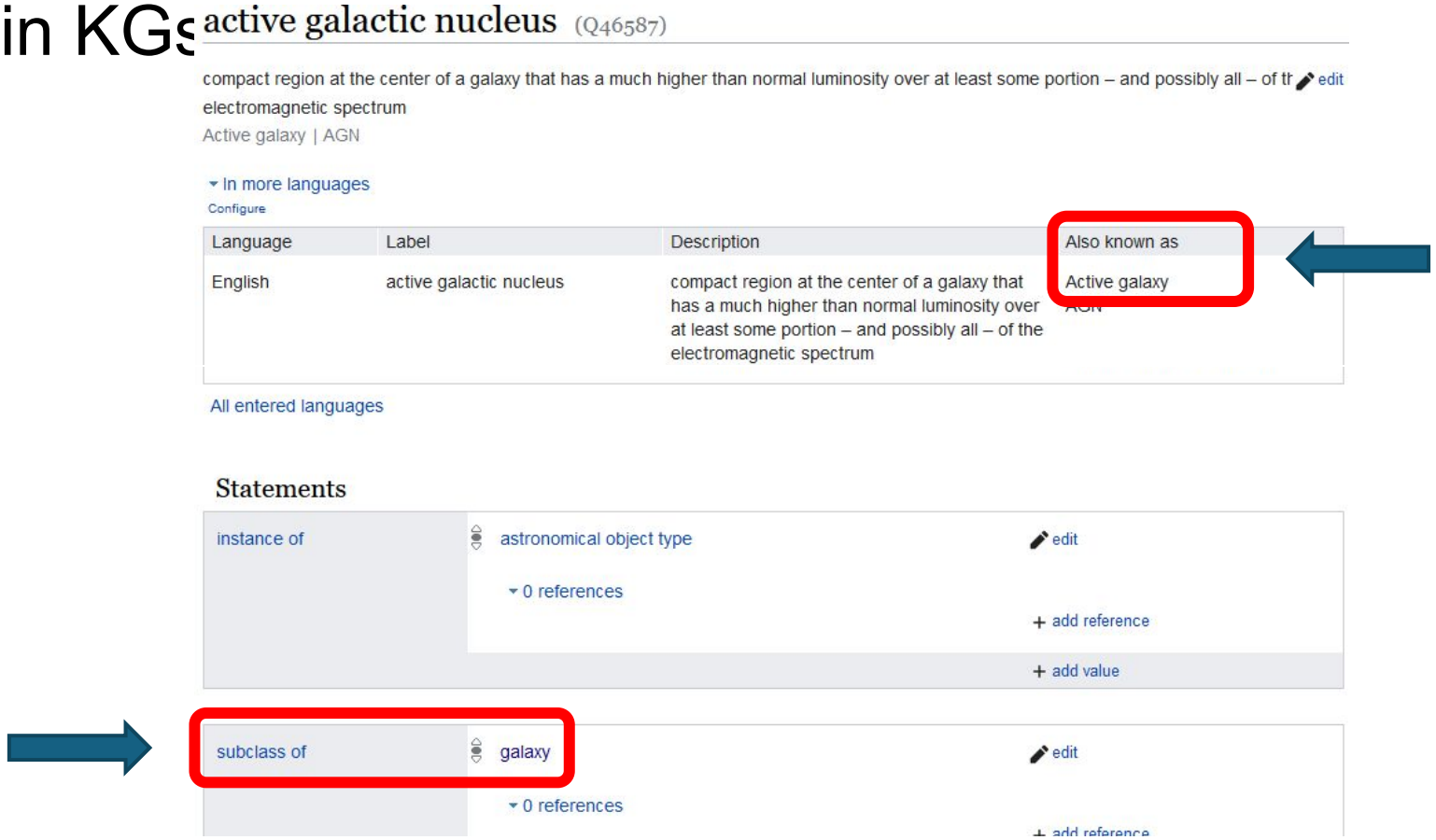

Sources:

- supplementary material optionally submitted by authors
- Huge manual work of curators
- User-provided feedback

http://simbad.u-strasbg.fr/Pages/guide/ch02.htx

# The bad news…

Existing cross-domain KGs



# The ugly news…

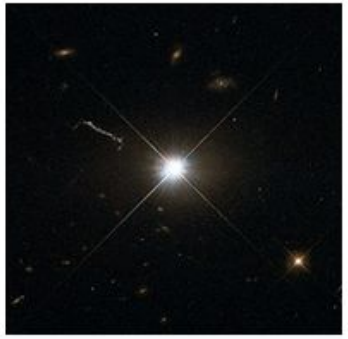
Existing cross-domain KGs:

DBpedia Browse using Formats Faceted Browser Sparql Endpoint

## About: Active galactic nucleus

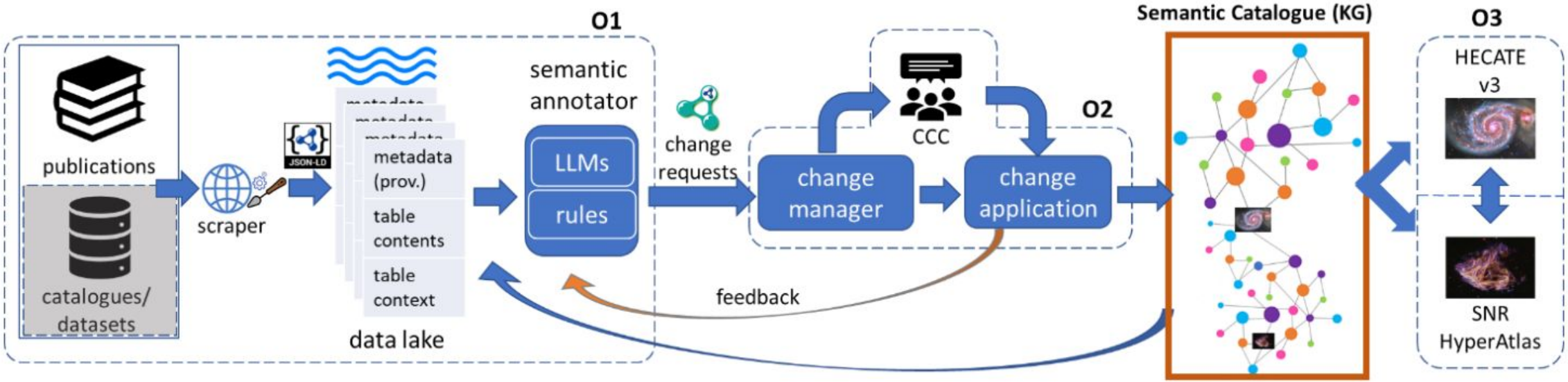
An Entity of Type: settlement, from Named Graph: http://dbpedia.org, within Data Space: dbpedia.org

An active galactic nucleus (AGN) is a compact region at the center of a galaxy that has a much-higher-than-normal luminosity over at least some portion of the electromagnetic spectrum with characteristics indicating that the luminosity is not produced by stars. Such excess non-stellar emission has been observed in the radio, microwave, infrared, optical, ultra-violet, X-ray and gamma ray wavebands. A galaxy hosting an AGN is called an "active galaxy". The non-stellar radiation from an AGN is theorized to result from the accretion of matter by a supermassive black hole at the center of its host galaxy.



https://isl.ics.forth.gr/PARSEC

# PARSEC



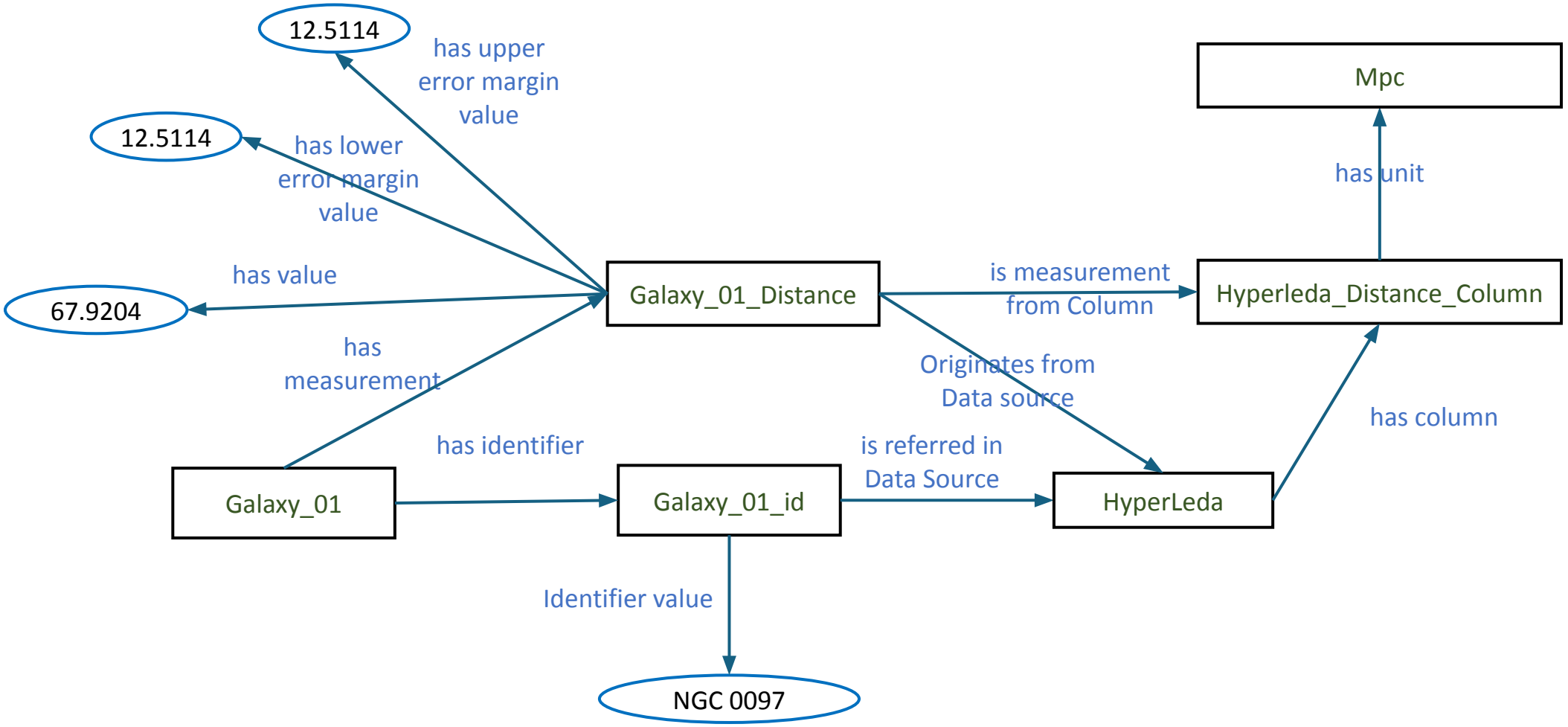
# DOCBO and HECAT3KG

- **DOCBO**: the ontology we created, containing 5 main classes:
  - Celestial Body, Identifier, Data Source, Measurement, Data Column
  - Public: https://zenodo.org/records/15388573

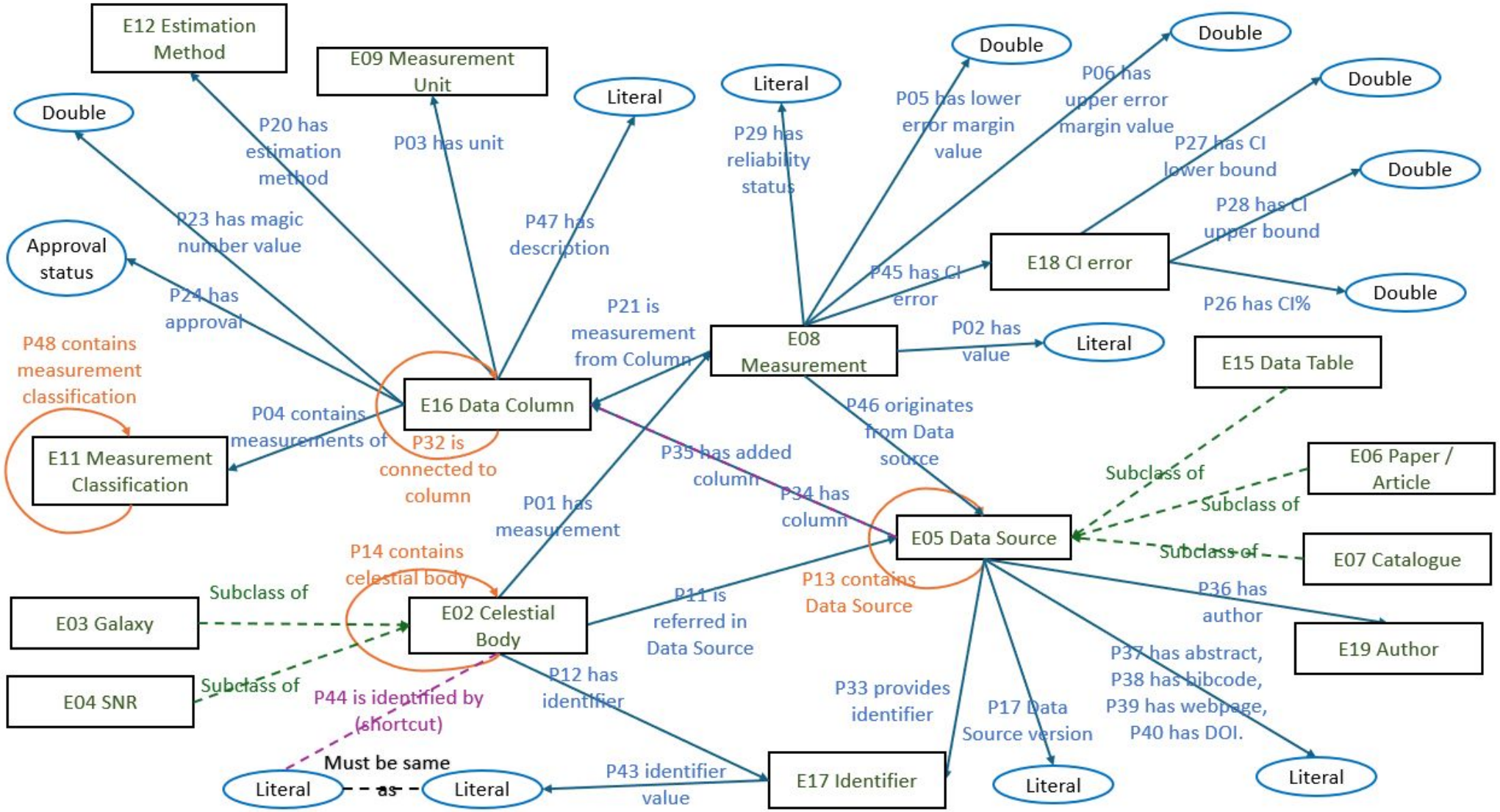
- **HECAT3KG**: A KG, following DOCBO, representing HECATE
  - following two mapping approaches (RML and X3ML)
  - 44 million RDF triples (~5GB)
  - Public: https://zenodo.org/records/15379419 (along with mappings)

| PGC | OBJNAME | ID_NED | ... | ID_2MASS | ... | RA | DEC | ... | D | E_D | ... |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 1442 | NGC0097 | NGC 0097 | | 00222998+2944433 | | 5.624916 | 29.745361 | | 67.9204 | 12.5114 | |

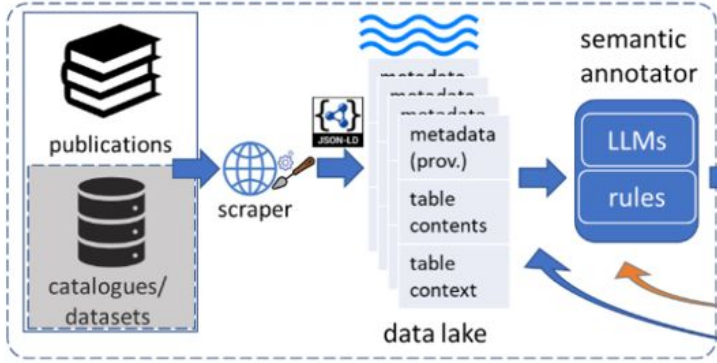
# HECAT3KG - Example



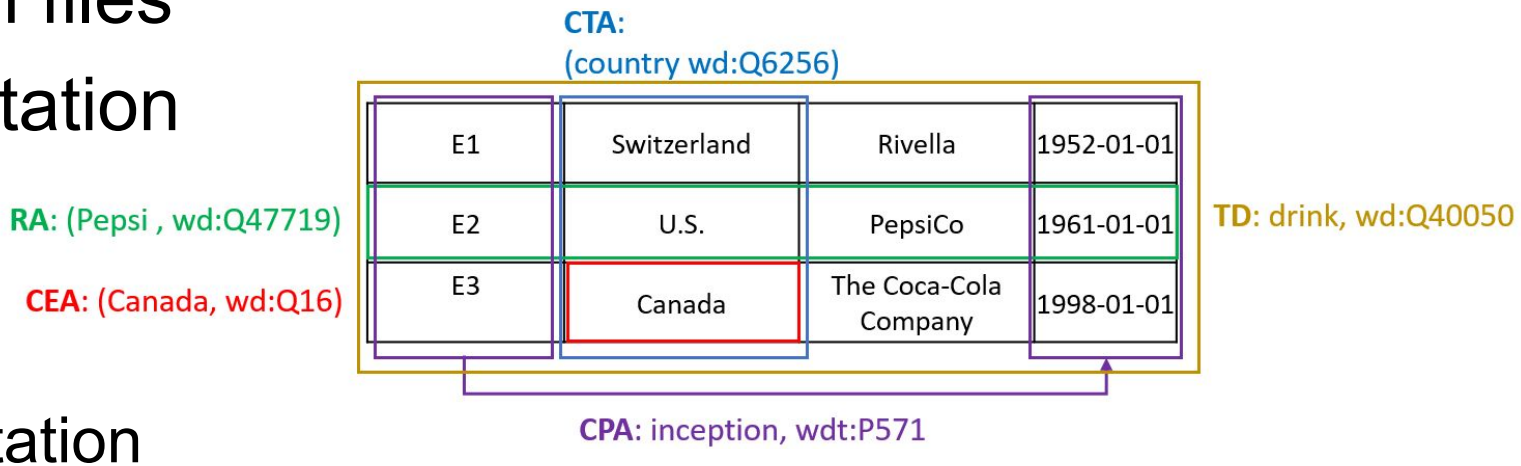
# DOCBO


E12 Estimation Method
E09 Measurement Unit
Double
Literal
Double
Double
Double
Double
Double
Double
Approval status
P20 has estimation method
P03 has unit
P29 has reliability status
P05 has lower error margin value
P06 has upper error margin value
P27 has CI lower bound
P28 has CI upper bound
P23 has magic number value
P47 has description
P24 has approval
P45 has CI error
P26 has CI%
E18 CI error
P21 is measurement from Column
E08 Measurement
P02 has value
Literal
E15 Data Table
P48 contains measurement classification
E16 Data Column
P04 contains measurements of
P32 is connected to column
P46 originates from Data source
E11 Measurement Classification
P35 has added column
P34 has column
Subclass of
E06 Paper / Article
Subclass of
P01 has measurement
E05 Data Source
Subclass of
E07 Catalogue
P14 contains celestial body
P13 contains Data Source
P36 has author
Subclass of
E03 Galaxy
E02 Celestial Body
P11 is referred in Data Source
E19 Author
P37 has abstract, P38 has bibcode, P39 has webpage, P40 has DOI.
Subclass of
E04 SNR
P44 is identified by (shortcut)
P12 has identifier
P33 provides identifier
P17 Data Source version
Must be same as
Literal
Literal
P43 identifier value
E17 Identifier
Literal
Literal

# Tabular Data Mining and Annota




- Read pdf/mrt publication files
- Semantic Table Interpretation
  - Table Topic Detection
  - Cell Entity Annotation
  - Column Type Annotation
  - Columns-Property Annotation
- Neuro-symbolic approach
  - Statistical models & LLMs to suggest annotations
  - Rules to reduce the search space



# Conclusions

- AI (Bits) is no longer just a tool; it's an integral part of astrophysics
- Astrophysics (Stars) is entering a golden era of data (e.g., LSST, SKA) that CS can uniquely address
- Ontologies/KGs (Knowledge) offer a bridge between raw data and a deep understanding (years of research founded in FOL)

- Interdisciplinary efforts like **UniversAI** show that collaboration isn't just possible—it's urgently needed.

# Thank you!