Knowledge Discovery Interest Group

Chair: Yihan Tao (NAOC & China-VO)

Vice-Chair: André Schaaff (CDS)

IVOA Interop

Görlitz, Germany

Nov 14, 2025

KD-IG

Knowledge Discovery is the task of **processing and analyzing astronomical datasets** with the aim of **extracting new knowledge**. This endeavor spans multiple disciplines including data access and exploration, visualization, artificial intelligent, statistical methods and workflow orchestration.

Roadmap

- Following the Malta KD-DSP joint session and its discussion regarding AI impact on VO standards, addressing how to **support AI workflows while enabling interoperability**
- **Reflecting the state of the scientific and technological art** in AI to the VO community, continue to interact with IVOA WGs/IGs
- Investigating standards and best practices for **making multimodal astronomical data ready for AI**

AI for Astronomy and VO

- Emerging paradigms of AI for Astronomy
  - **Foundation Models: Universal Representations**
    - From task-specific models to pre-trained general-purpose embeddings (e.g. AstroCLIP, AION-1, AstroM^{3},etc)
    - Enabling unified multi-modal understanding (images + spectra + light curves + text) and few-shot learning for rare objects
    - Requirements of massive, diverse, standardized pre-training datasets
  - **AI Agents: From Prediction to Autonomous Discovery**
    - Data centers are setting up agents and MCP servers for accessing data services and autonomous data analysis
    - Need for standardized interfaces for agents to interact with astronomical infrastructure and data resources
- Raising topics for VO as the essential infrastructure connecting AI and astronomical data
  - How to provide FAIR data for AI?
  - How could VO tools and protocols be applied and integrated with agent-based workflows?

KD activity in this Interop

joint session with DCP on "Using AI in Open Data Context"

Saturday November 15 @14:00 UTC+1: Room Wichernhaus "zenith"			
Speaker	**Title**	**Time**	**Abstract**
Gilles Landais	Trustworthy AI (and the Eu AI Act)	10+5	Quick report on the European Articial Interlligence act.
Thomas Aynaud (remote)	Software Heritage, project presentation and applications to AI	20+5	This talk will provide an overview of what Software Heritage is. We will highlight the data we collect and how we do that. It includes the source code and all the events that can be observed during the development process. We will then explore how to utilize and enrich this information to develop services relevant to AI and the actions we take to create a more open, responsible, and transparent AI.
Thomas Boch	First feedback on implementing MCP server to access CDS services	15+5	MCP (Model Context Protocol) is a standard aiming at connecting AI tools to external data sources. We will present the MCP server we developed at CDS, and demonstrate how it allows AI applications to access and consume data from Sesame, MocServer, HiPS2FITS, etc.
François Lanusse (remote)	Make astronomical data AI-ready across surveys: Lessons-learned from the Multimodal Universe Project	15+5	Deep Learning has seen a recent shift in paradigm, from training specialized models on dedicated datasets, to so-called Foundation Models, trained in a self-supervised manner on vast amounts of data and then adapted to solve specific tasks with state-of-the-art performance. This new paradigm has been exceptionally successful not only for large language models (LLMs) but in other domains such as vision models. However applications of this new approach in astrophysics are still very scarce, for reasons ranging from new architectures to the (surprising) lack of availability of suitable large scale datasets. In this talk, I will discuss our work on building the Multimodal Universe Dataset, the first attempt at building a framework for homogenizing data across many astronomical surveys specifically for AI training purposes. Besides discussing the technical solution we adopted, I will highlight some of the outstanding challenges to enable cross-matching across surveys and streaming data in a way that is compatible with large AI model training. Project page: https://github.com/MultimodalUniverse/MultimodalUniverse
All	Discussions	10	

CONTACTS

Subscribe the mailing list: http://mail.ivoa.net/mailman/listinfo/kdd

Chair： Yihan Tao (NAOC & China-VO)，y.tao@nao.cas.cn

Vice-Chair: André Schaaff (CDS), andre.schaaff@astro.unistra.fr