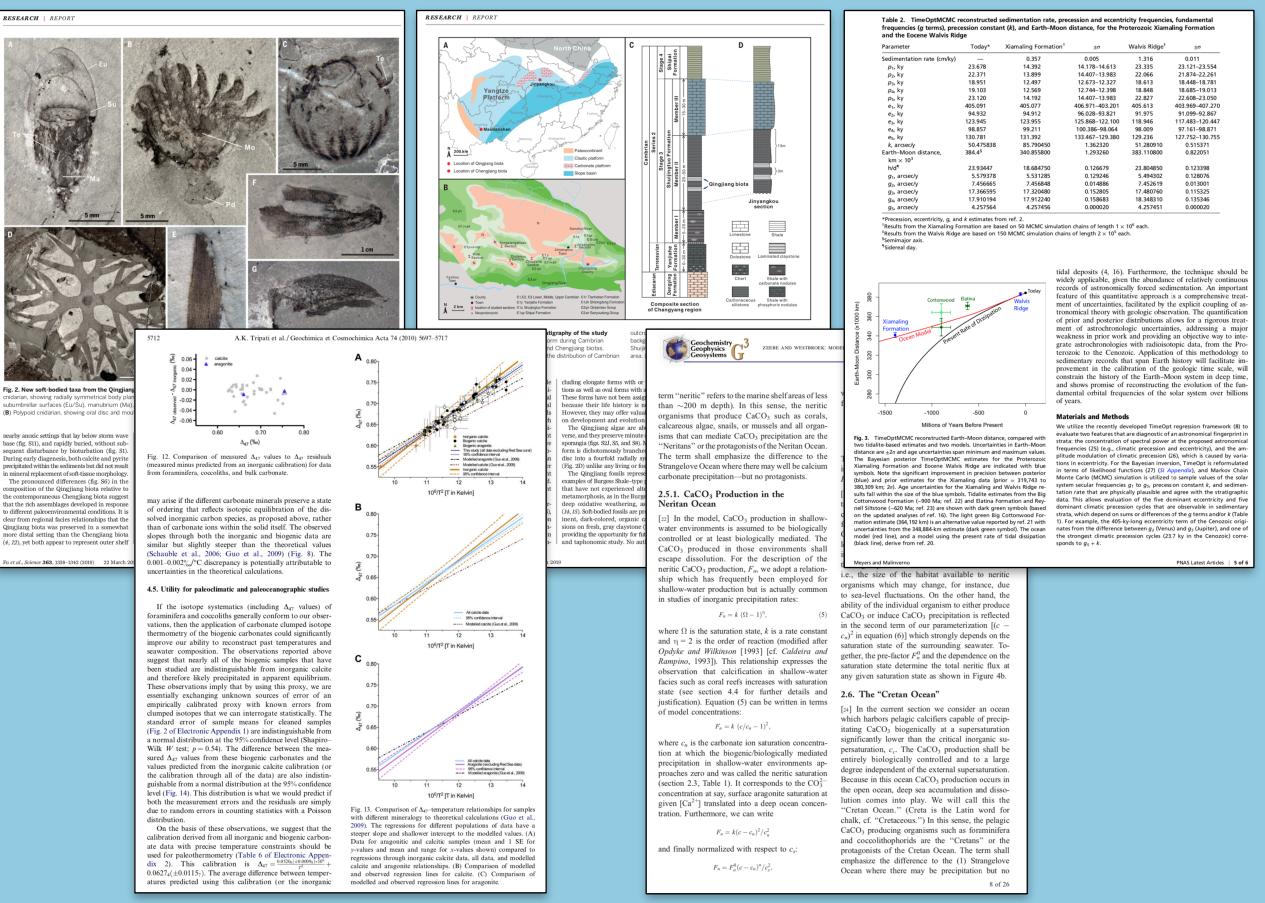


COSMOS: An AI platform for knowledge discovery and scientific model curation

Program Manager: Joshua Elliott, DARPA, ASKE

Goal: Develop an AI technical assistant capable of assimilating knowledge and data relevant to scientific models from across millions of publications.

Problem Overview



Scientific Literature

Scientific publications encapsulate technical knowledge and contain data and descriptions of phenomena necessary to parameterize and evaluate models. Using the published literature to construct and assess scientific models is challenging due to:

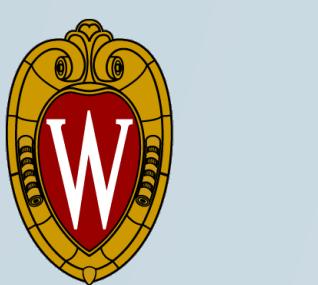
- Vast and rapidly growing numbers of publications distributed by multiple sources
- Heterogeneity of document formats and publishing conventions
- Disparate representations of information
- Uncertainty and conflicting information

Need an AI technical assistant to automate integration and analysis

Scientific Model Code

Much of the world's scientific knowledge is most explicitly encapsulated in model codes. Understanding models, re-calibrating them to observational or experimental data, and evaluating model output is often challenging due to:

- Large, complex, code bases with sparse, heterogeneous documentation
- Need to map model parameters and output to diverse representations in scientific publications
- Difficulty of identifying related phenomena not currently captured by model
- Uncertainty in model-data comparisons



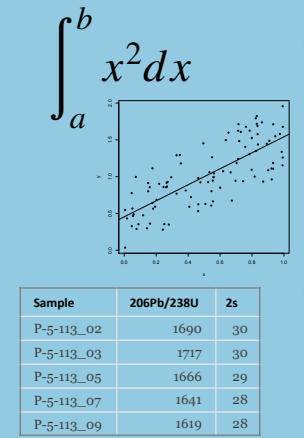
THE UNIVERSITY
of
WISCONSIN
MADISON

Theodoros Rekatsinas, Ankur Goswami, Ian Ross, Akshata Bhat, Miron Livny
Shanan Peters. More info: <https://github.com/UW-COSMOS>

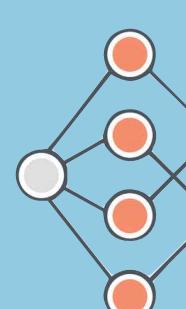
Solution Overview



- User-focused
- Answer questions, generate synthesis results
- Interactive evaluation and AI model training



- Extraction of sci. models, entity relations across tables, figures, equations
- Knowledge discovery across millions of papers



- Fine-tuned ML models over unified contextual representation to support diverse tasks
- Continual updating

Retrieve-and-Read Q&A

On-demand Knowledge Bases

Model Evaluation

Dataset Generation

COSMOS Microservices API

COSMOS Microservices

Automated container deployment, scaling, and management using Docker swarm; open source



COSMOS Universal Entity Representation

COSMOS Ingestion Layer

Deep Learning models for document analysis and recognition. Granular information extraction from tables, figures, text, and mathematical expressions.



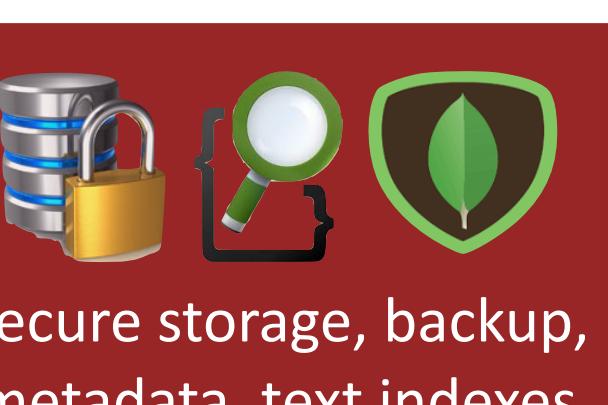
High Throughput Computing Layer



GPU, CPU nodes
DAGman job queue
encrypted binaries

Digital Library Layer (xDD)

Automated ingestion of full publications from multiple publishers, backed by institutional agreements: **10.6M documents, +10K daily**



www.darpa.mil

Distribution authorized to U.S. Government Agencies and their contractors

