GeoDeepDive: towards a TDM-ready digital library and machine reading system for the geosciences

Shanan E. Peters¹, Miron Livny², Ian Ross³, Chris Ré³
Depts. of Geoscience¹ and Computer Sciences², University of Wisconsin-Madison; Dept. of Computer Science, Stanford University³

Motivation: Leveraging “Dark Data” in the “Long Tail” of science

The scientific literature contains a large amount of valuable data distributed over millions of documents. Many questions require the synthesizing these data in order to generate novel aggregate datasets. This project aims to reduce the time and cost barriers to assembling synthetic literature-based datasets and to produce a new generation of probabilistic databases that remain tightly coupled to sources and that can improve and extend as new questions emerge and new capabilities are developed.

Knowledge Base Construction: DeepDive Pipeline

Leveraging a new TDM-ready library and supporting infrastructure

API for document and data discovery; can leverage the full range of TDM data products (basic functionality available now!)

Support for establishing dynamic linkages between existing databases and literature (see demo)

Push notifications when relevant new documents are acquired (coming soon)

Feedback system for improving/defining matched document sets (coming soon)

HTC capacity to rapidly redeploys software tool versions or new tools across all documents (basic functionality available now!)

Growing knowledge bases and large datasets to support advanced machine learning approaches

Literature Cited