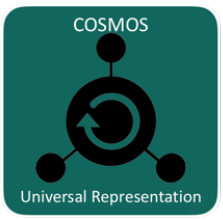
COSMOS: An AI platform for knowledge extraction from scientific publications

Theo Rekatsinas, Shanan Peters, Miron Livny

Outline

1. Challenge: Produce an AI technical assistant

2. COSMOS: Knowledge extraction as a service

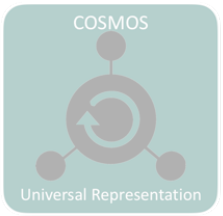


3. Demo: analyzing model code with COSMOS

Outline

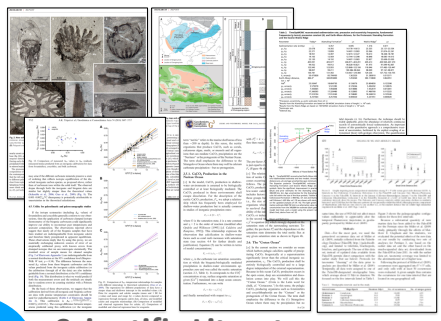
1. Challenge: Produce an AI technical assistant

2. COSMOS: Knowledge extraction as a service

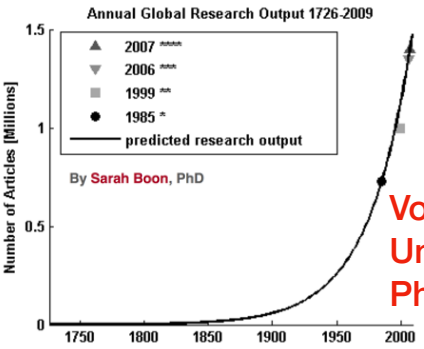


3. Demo: analyzing model code with COSMOS

Scientific workflows have a major bottleneck…

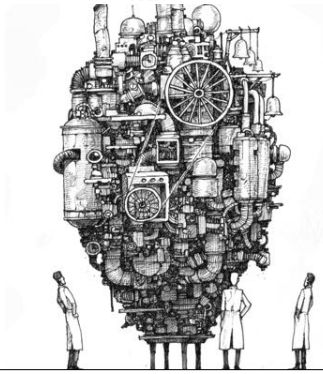


Scientific Literature



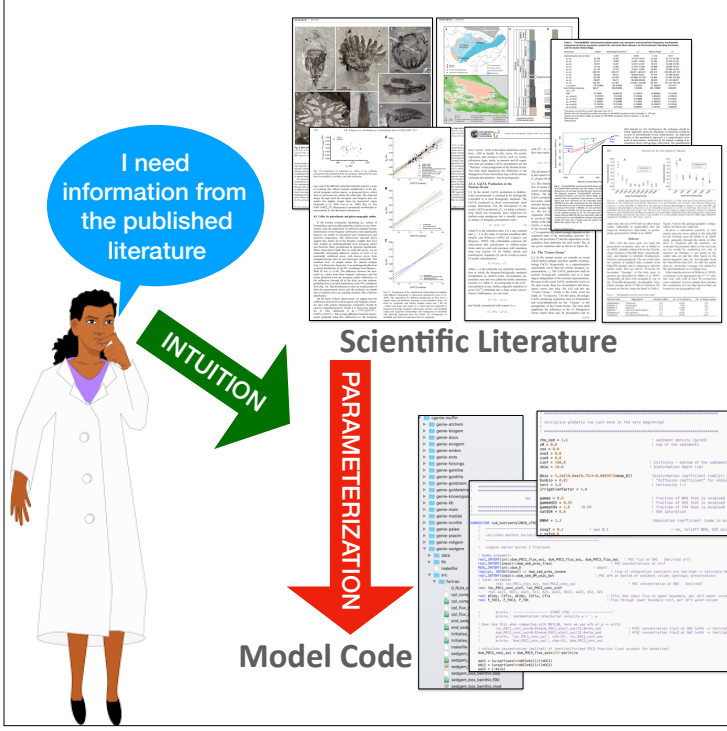
Voluminous
Unstructured
Physically scattered

Model Code



Monolithic
Complex
Difficult to assess

Scientific workflows have a major bottleneck…

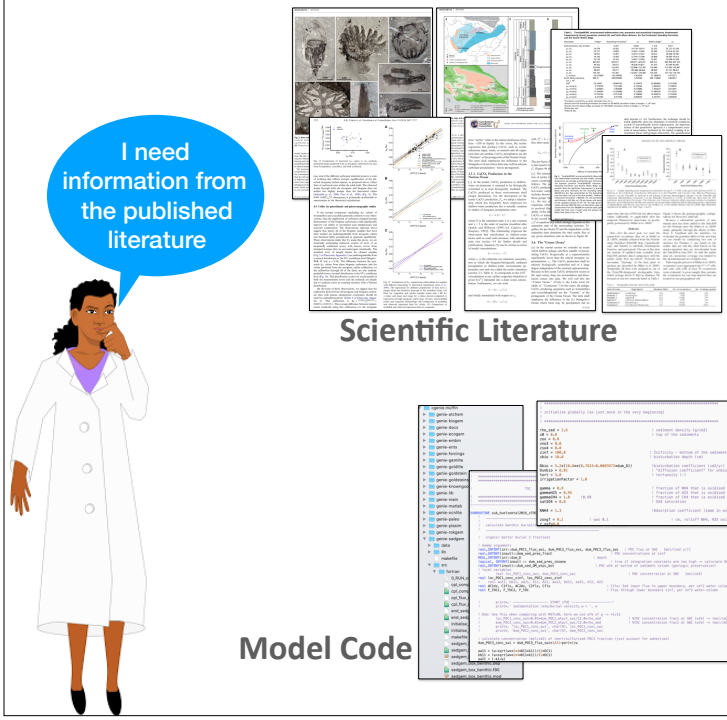


EMPIRICAL DATA

Model-data comparison a major pain-point!

PREDICTIONS

Scientific workflows have a major bottleneck…



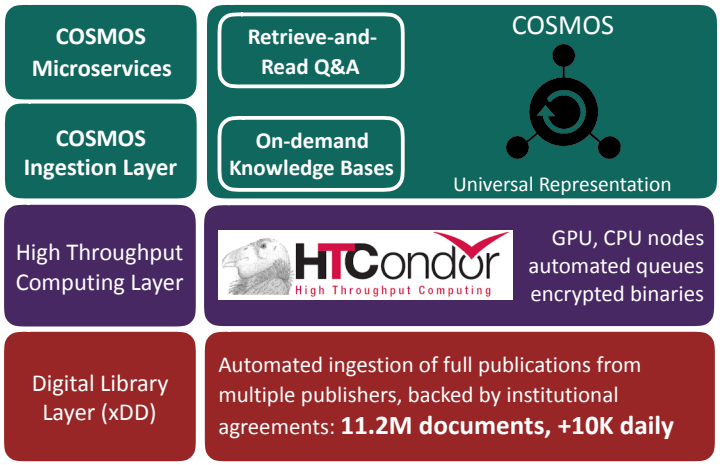
Deep-time Digital Earth aims to liberate data from collections such as the British Geological Survey's. BRITISH GEOLOGICAL SURVEY

Earth scientists plan to meld massive databases into a 'geological Google'

By Dennis Normile | Feb. 26, 2019 , 5:00 PM

In the past…volunteer earth scientists tried to do nearly everything themselves, including informatics and data management. [The Chinese] ***GBDB instead pays nonspecialists to input reams of data gleaned from earth science journals…***

xDD and COSMOS: an end-to-end stack for accelerating scientific discovery



- Ecosystem of lightweight, scalable services to locate, extract, and aggregate data and information from heterogeneous sources
- Supporting HTC infrastructure to parse and analyze documents, expose text via API
- Principled, automated access to new and archival publications spanning publishers

NSF USGS science for a changing world DARPA

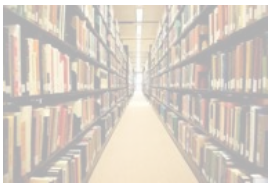
support 2014-2018 by NSF-ICER 1343760, DARPA ASKE
partial current support from USGS

xDD API:
https://geodeepdive.org/api
Code available at:
https://github.com/UW-COSMOS

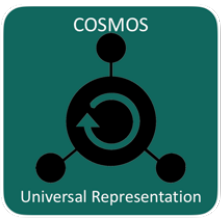
Correspondance:
Shanan Peters (peters@geology.wisc.edu)
Theodoros Rekatsinas (thodrek@cs.wisc.edu)
Miron Livny (miron@cs.wisc.edu)

Outline

1. Challenge: Produce an AI technical assistant

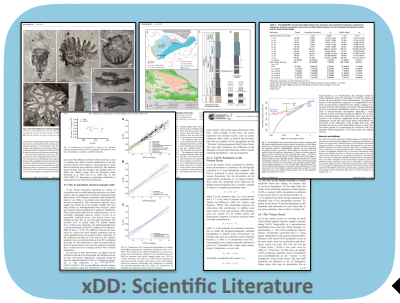


2. COSMOS: Knowledge extraction as a service

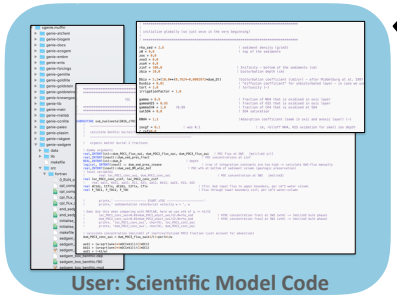


3. Demo: analyzing model code with COSMOS

Accelerating scientific discovery with COSMOS



xDD: Scientific Literature



User: Scientific Model Code

COSMOS

Universal Representation

- Retrieve-and-Read Q&A
- On-demand Knowledge Bases
- Model Evaluation
- Dataset Generation

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

- Entity relations across tables, figures, equations
- Discover knowledge in millions of papers

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

Goal: Knowledge extraction from high-variety input

Equation

$$\delta^{13}C = (R_{sample}/R_{standard} - 1) \times 10^3$$

Equation

Body Text

where R is $^{13}C/^{12}C$. The standard is Pee Dee Belemnite limestone that has been assigned a value of 0.0‰. The precisions of $\delta^{13}C$ determination were less than 0.2‰. POC and PON concentrations were determined using a TCD detector attached to the elemental analyzer.

For Chl *a* and pheophytin concentrations, POM samples were extracted in the dark for 12 h by 90% acetone, and their concentrations were measured by the fluorometric method (Japan Meteorological Agency, 1970), using a calibrated Turner Designs TD700 fluorometer. In this study, chlorophyll (Chl) was determined as the total pigment including pheophytin. PO_4-P was extracted filtrate by the ascorbic acid–Mo blue method (Strickland and Parsons, 1965), using a Technicon Auto Analyzer.

Section Header

3. Results

Context

Section Header

3.1. Variations in river discharge and riverine POM composition

Body Text

River discharge of the Kiso Rivers changed considerably during the observation period (Fig. 3). Discharge was low (<500 $m^3 s^{-1}$) until 22 June, and suddenly increased on 24 June (the first flood, ~2000 $m^3 s^{-1}$), reaching a peak flood on 28 June (the second flood, ~3000 $m^3 s^{-1}$). After that, it

during normal discharge. However, the concentration in the Nagara River at high discharge was the same level as that at normal discharge. After discharge, POC concentrations decreased in all rivers. $\delta^{13}C$ of POM in the Kiso River and the Nagara River varied from −27.3‰ to −23.1‰ and from −29.7‰ to −25.9‰, respectively. On the other hand, $\delta^{13}C$ of POM in the Ibi River remained fairly constant (ca. −30‰). The C/N ratios varied from 7.8 to 22.3 and reached the highest values during high discharge in all rivers.

Table

Table 1

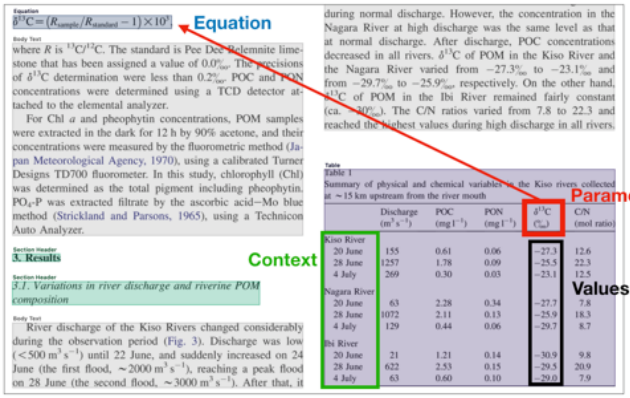
Summary of physical and chemical variables in the Kiso rivers collected at ~15 km upstream from the river mouth

Parameter

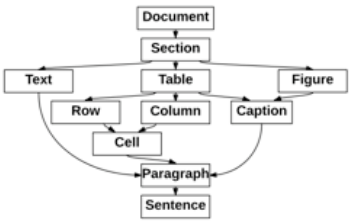
	Discharge ($m^3 s^{-1}$)	POC ($mg l^{-1}$)	PON ($mg l^{-1}$)	$\delta^{13}C$ (‰)	C/N (mol ratio)
Kiso River					
20 June	155	0.61	0.06	−27.3	12.6
28 June	1257	1.78	0.09	−25.5	22.3
4 July	269	0.30	0.03	−23.1	12.5
Nagara River					
20 June	63	2.28	0.34	−27.7	7.8
28 June	1072	2.11	0.13	−25.9	18.3
4 July	129	0.44	0.06	−29.7	8.7
Ibi River					
20 June	21	1.21	0.14	−30.9	9.8
28 June	622	2.53	0.15	−29.5	20.9
4 July	63	0.60	0.10	−29.0	7.9

Values

Core modules in COSMOS

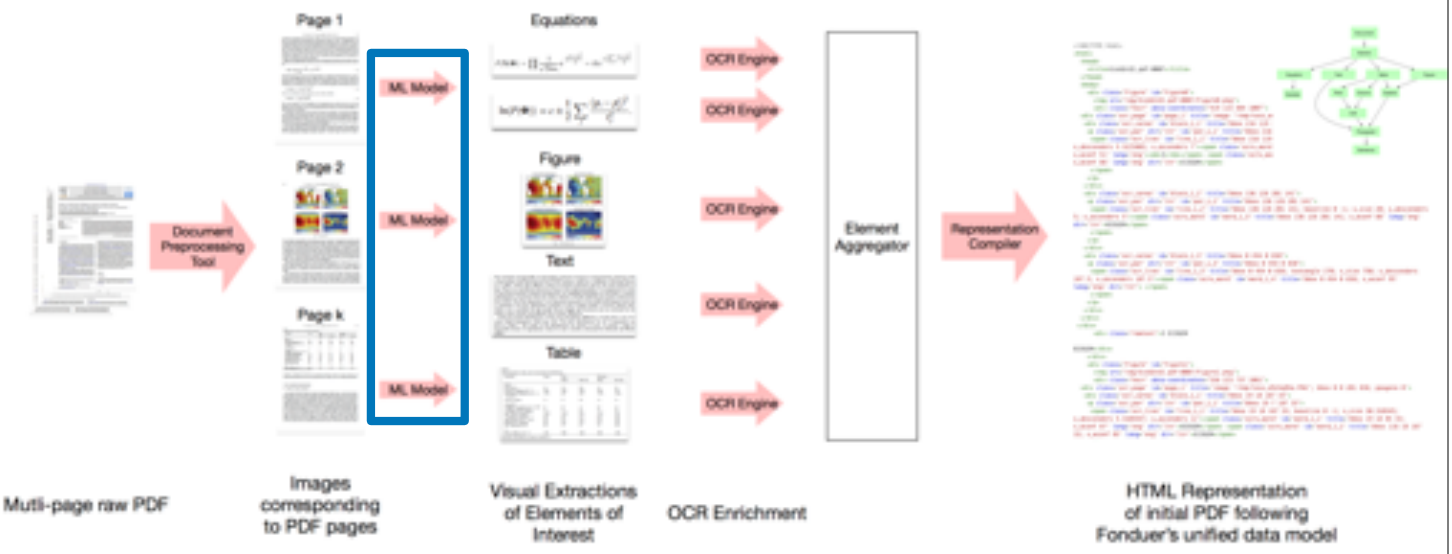


- **Module 1 [Ingest box]** : Identify semantically meaningful components of a PDF and store them in an easy to access to electronic format.



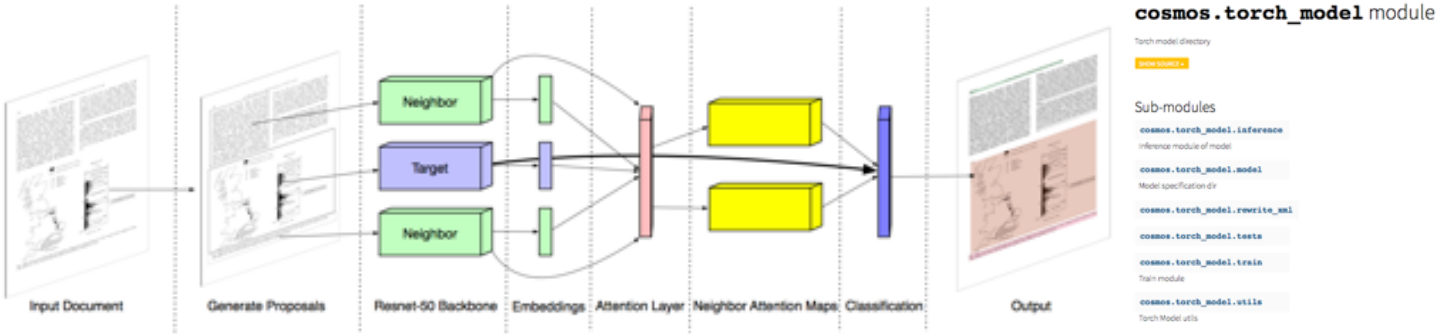
Output format: XML stored in MongoDB backend and queryable via ElasticSearch

Ingest Box: From PDF to XML

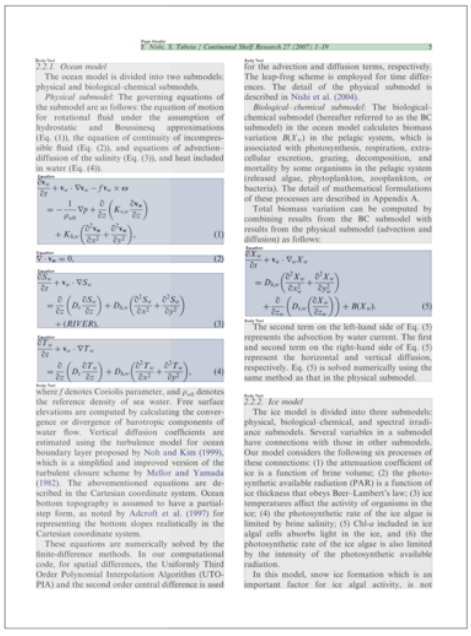


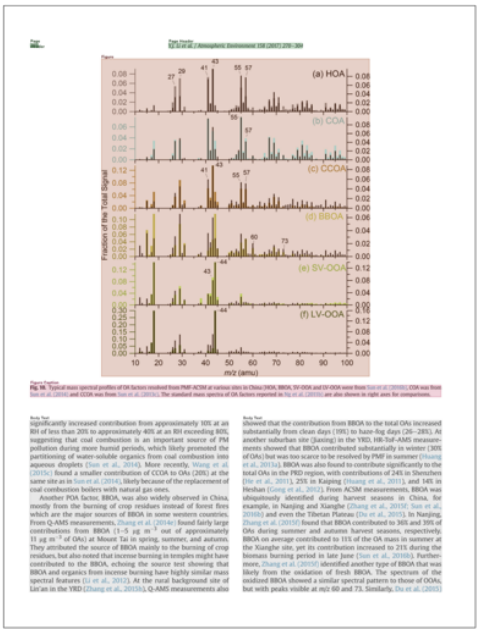
We need support for images to address the format heterogeneity across publications.
How do we get these elements from a scanned image?

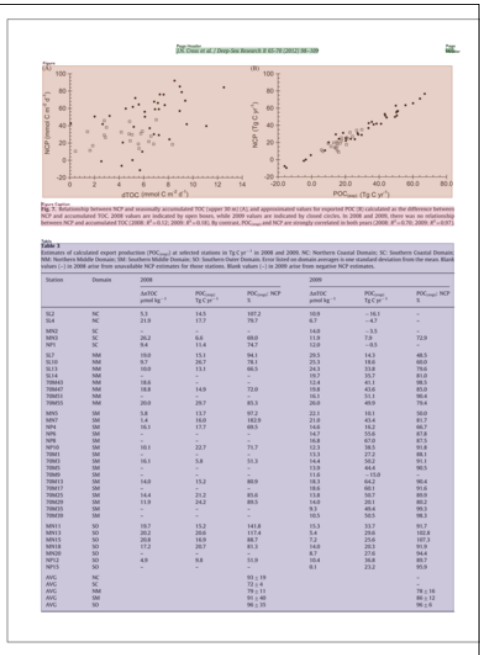
The COSMOS Attentive RCNN Model



New distributed representation (in the visual space) for each element in the page.

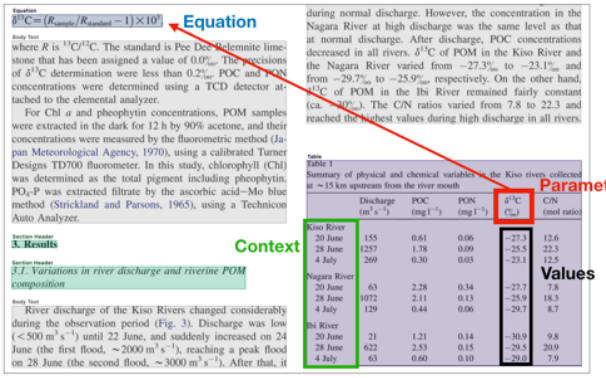






The output of COSMOS's object detection module:
tables, figures, equations, and associated text (captions, body text)

Core modules in COSMOS



- **Module 1 [Ingest box]** : Identify semantically meaningful components of a PDF and store them in an easy to access to electronic format.

- **Module 2 [Services]** : Identify semantically meaningful components of a PDF and store them in an easy to access to electronic format.

Ingestion augmentation services

LaTeX Specific OCR

Module for doing ocr for latex

Table and Figure Knowledge Base Construction

Module to construct tables and figures csvs

Equation/Body Text Knowledge Base Construction

Construct a knowledge base from equations and body text.

Branch: v0.3.0 Cosmos / services / ingestion / **src** /

code_ingestion.py

Knowledge extraction services

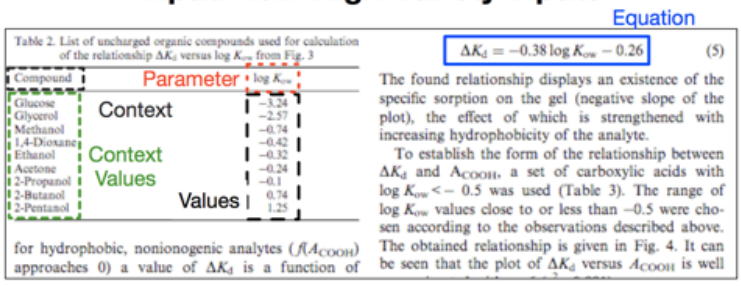
Branch: v0.3.0 Cosmos / services /

- question_answering_backend
- search_backend
- table_extractions
- word_embeddings

PyDoc for missing elements to be generated soon

Services: Knowledge extraction (equations/tables)

Input: Raw High-Variety Inputs



Output: Extracted Structured Knowledge

EID	Expression
E1	$\Delta K_d = -0.38 \log K_{ow} - 0.26$
E2	$G_O = \left[1 + (a + b(\mathrm{RH}*0.01) + c(\mathrm{RH}*0.01)^2)*\frac{\mathrm{RH}}{100 - \mathrm{RH}}\right]^{1/3}$

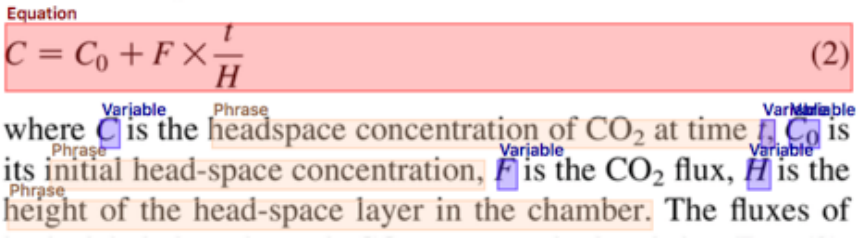
EID	Parameter	Context	Data
E1	logKd	Compound	DataFrame1
E2	*a*	Composition	DataFrame2
E2	*b*	Composition	DataFrame3
E2	*c*	Composition	DataFrame4

DataFrame1

Context	Values
Glucose	-3.24
Glycerol	-2.57
Methanol	0.74
1,4-Dioxane	-0.42
Ethanol	0.32
Acetone	-0.24
2-Propanol	-0.1
2-Butanol	0.74
2-Penthanol	1.25

COSMOS extracts knowledge from **multi-modal unstructured data** (text, tables, images, equations, diagrams)

Services: Model extraction



- The parse tree for extracted equation (red-colored bounding box) reveals symbols **C**, **t, Co, F,** and **H**
- The same symbols are recognized in the text below this equation (purple "Variable" tokens).
- Variable tokens are linked to descriptions using the output of Open-IE (using CoreNLP), linking the Variable tokens and the phrase tokens.
- This method can be further improved (e.g., F here not automatically associated with CO_2 flux).

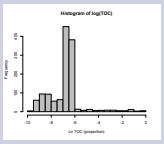
Services: Open-domain question answering

Model Factoid-Q&A: *What is "TOC" in my source code?*

Definition

Total Organic Carbon (TOC) is the amount of organic carbon in soil or a geological formation, particularly the source rock for a petroleum play.

Empirical Distribution



Entity-Aware Code Analysis

Fine-tuned for Q&A

Contextual Entity Embeddings

Model: trained with self-supervised learning

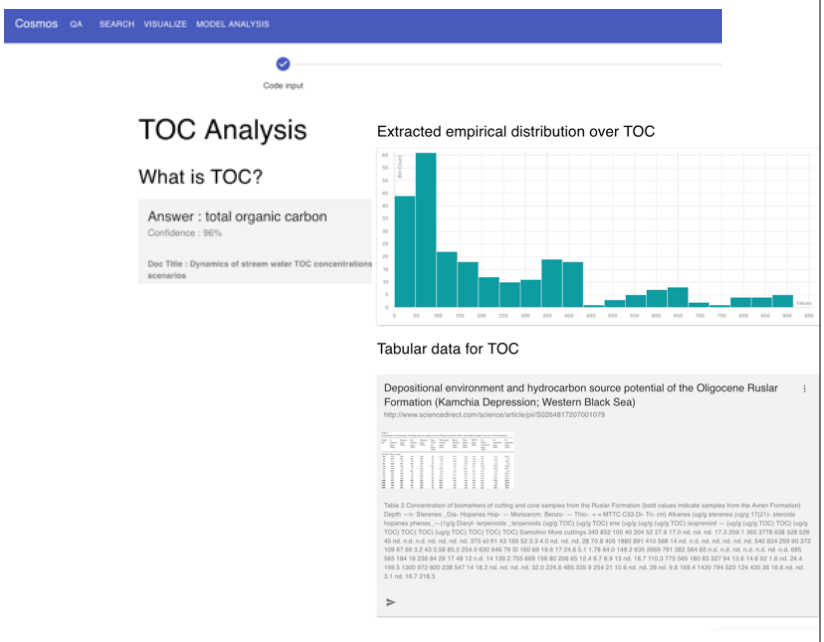
Multi-Headed Attention

Entity Embeddings

Text

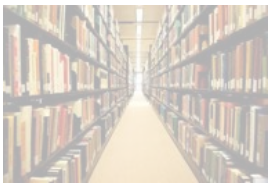
Graphs

Images

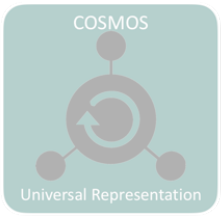


Outline

1. Challenge: Produce an AI technical assistant



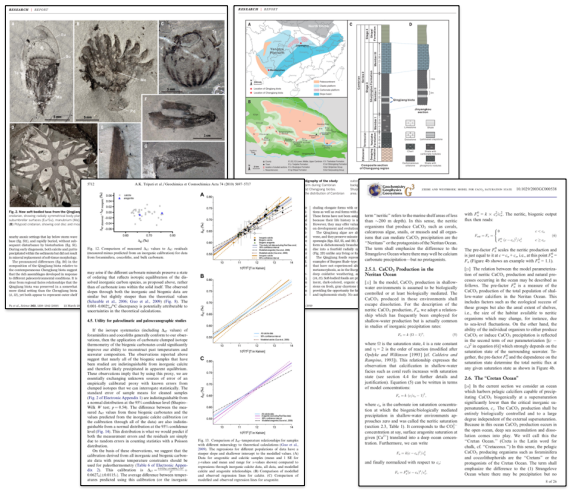
2. COSMOS: Knowledge extraction as a service



3. Demo: analyzing model code with COSMOS

Fundamental challenges

Direction 1: Retrieval for open-domain scientific discovery

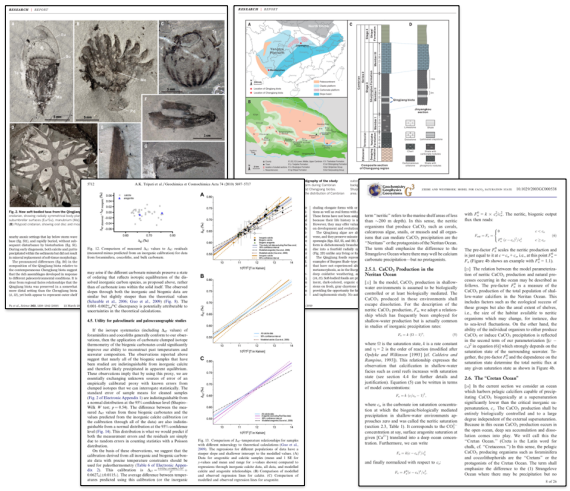


Question answering assumes strong supervision of the supporting evidence (i.e., retrieval is solved)

IR systems typically optimize different objectives (diversity) and can only personalize retrieval after collecting extensive evidence on its user

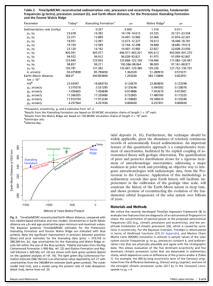
QA is fundamentally different from IR

Direction 1: Retrieval for open-domain scientific discovery



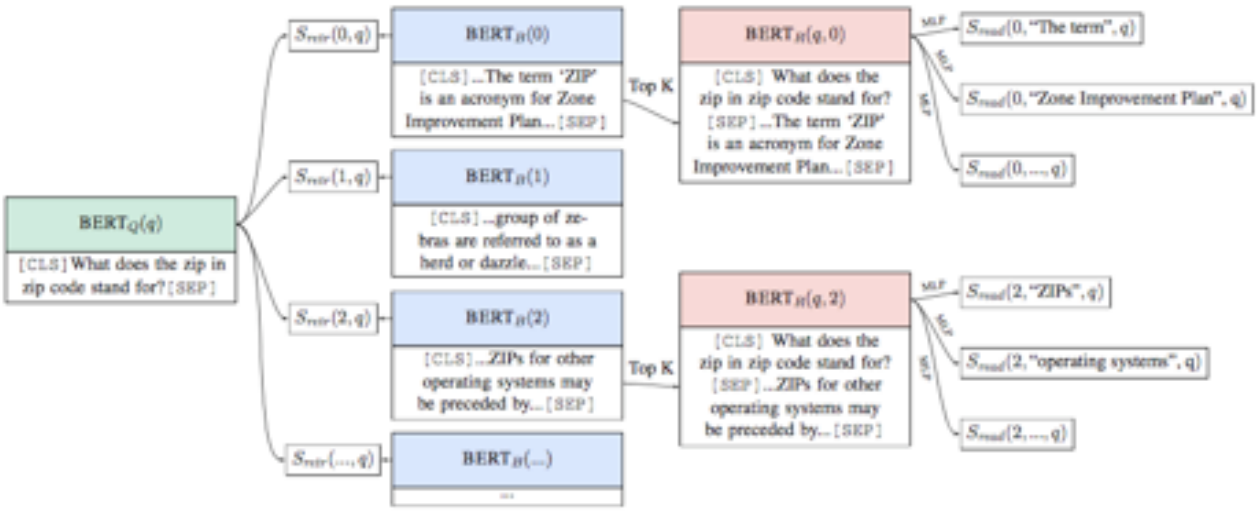
How can we leverage broader context for accurate retrieval?

Example Paper



Source Code

Direction 1: Retrieval for open-domain scientific discovery



Joint training of IR and QA (in an unsupervised manner)

Direction 2: Data management over multimodal data (indexing, querying, aggregation)

Screening Mammography

Think of SQL queries over numerical or categorical predicates combines with images or text descriptions (context)

Direction 2: Data management over multimodal data (indexing, querying, aggregation)

IMPRESSION :
No mammographic evidence of malignancy. Routine screening mammography is recomm

Mammogram BI-RADS: Overall: 1 - Negative.

As the teaching physician, I personally examined the radiologic study, reviewed the findi
interpretation.
EXAM:
MAMMOGRAM DIGITAL SCREENING BILATERAL W/TOMOSYNTHESIS

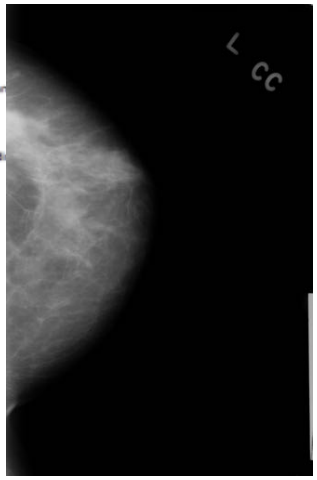
INDICATION:
Screening.

COMPARISON:
Compared to:

TECHNIQUE:
Current study was evaluated with Computer Aided Detection.

FINDINGS:
The breasts are heterogeneously dense, which may obscure small masses.

There are no significant masses, calcifications, or other findings.



Screening Mammography

Think of SQL queries over numerical or categorical predicates combines with images or text descriptions (context)

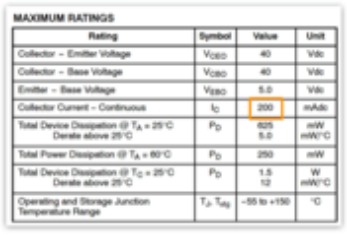
Conjunctive queries with similarity predicates (equality, inequality constraints over atomic data types merged with similarity queries)

We need a data model that merges relational algebra with linear algebra

Direction 3: On demand extraction of observational data

Input

Annotated examples on PDF



MAXIMUM RATINGS

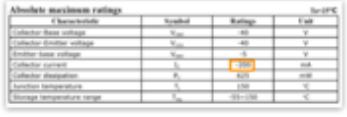
Rating	Symbol	Value	Unit
Collector – Emitter Voltage	V_{CEO}	40	Vdc
Collector – Base Voltage	V_{CBO}	40	Vdc
Emitter – Base Voltage	V_{EBO}	5.0	Vdc
Collector Current – Continuous	I_C	200	mAdc
Total Device Dissipation @ T_A = 25°C Derate above 25°C	P_D	625 5.0	mW mW/°C
Total Power Dissipation @ T_A = 60°C	P_D	250	mW
Total Device Dissipation @ T_C = 25°C Derate above 25°C	P_D	1.5 12	W mW/°C
Operating and Storage Junction Temperature Range	T_J, T_{stg}	−55 to +150	°C

2N3906-D.PDF

Absolute Maximum Ratings*

Symbol	Parameter	Value	Units
V_{CEO}	Collector-Emitter Voltage	40	V
V_{CBO}	Collector-Base Voltage	60	V
V_{EBO}	Emitter-Base Voltage	6.0	V
I_C	Collector Current - Continuous	200	mA
T_J, T_{stg}	Operating and Storage Junction Temperature Range	-55 to +150	°C

MMBT3904.PDF



Absolute maximum ratings

Parameter	Symbol	Ratings	Unit
Collector-base voltage	V_{CBO}	-60	V
Collector-Emitter voltage	V_{CEO}	-40	V
Emitter-base voltage	V_{EBO}	-5	V
Collector current	I_C	-200	mA
Collector dissipation	P_C	625	mW
Junction temperature	T_j	150	°C
Storage temperature range	T_{stg}	-55~150	°C

AUKCS04635-1.pdf

Task

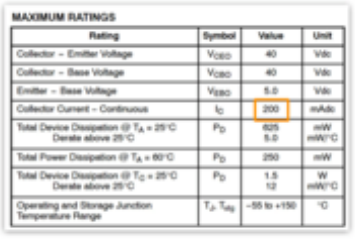
Extract the value of “Collector Current” for each model of transistor

How can a domain scientist extract the necessary observational data (from a collection of retrieved tables) without writing complex code?

Direction 3: On demand extraction of observational data

Input

Annotated examples on PDF

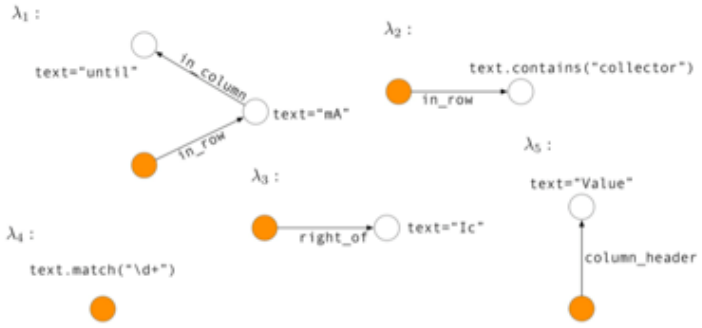


Task

Extract the value of “Collector Current” for each model of transistor

Output

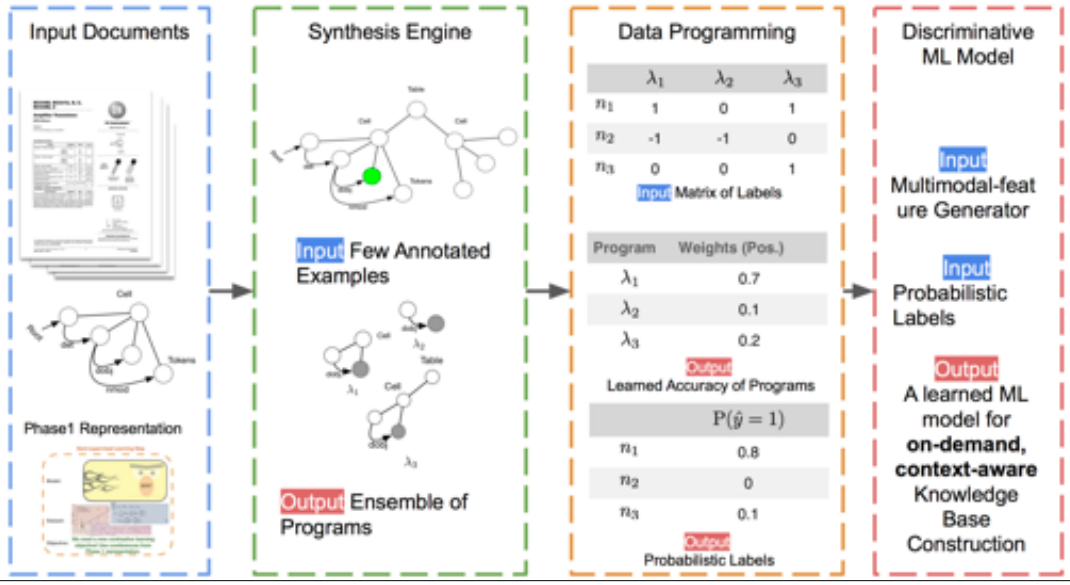
Ensemble of programs synthesized from the input examples



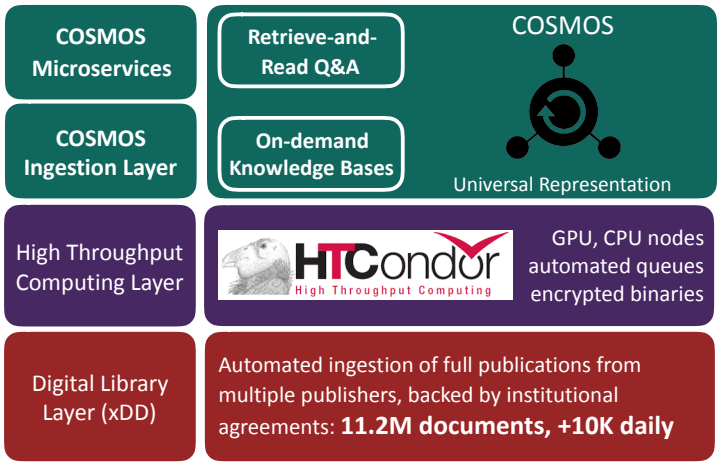
Weak Supervision powered by Program Synthesis

Direction 3: On demand extraction of observational data

The Hera Engine



xDD and COSMOS: an end-to-end stack for accelerating scientific discovery



- Ecosystem of lightweight, scalable services to locate, extract, and aggregate data and information from heterogeneous sources
- Supporting HTC infrastructure to parse and analyze documents, expose text via API
- Principled, automated access to new and archival publications spanning publishers

NSF USGS science for a changing world DARPA

support 2014-2018 by NSF-ICER 1343760, DARPA ASKE
partial current support from USGS

xDD API:
https://geodeepdive.org/api
Code available at:
https://github.com/UW-COSMOS

Correspondance:
Shanan Peters (peters@geology.wisc.edu)
Theodoros Rekatsinas (thodrek@cs.wisc.edu)
Miron Livny (miron@cs.wisc.edu)