

COSMOS: An AI platform for knowledge extraction from scientific publications

Theo Rekatsinas, Shanan Peters, Miron Livny

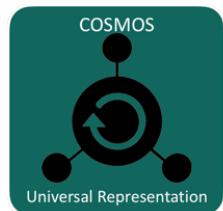


Outline

1. Challenge: Produce an AI technical assistant



2. COSMOS: Knowledge extraction as a service



3. Demo: analyzing model code with COSMOS

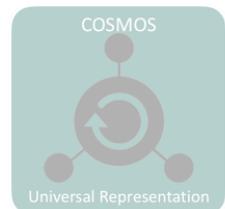


Outline

1. Challenge: Produce an AI technical assistant



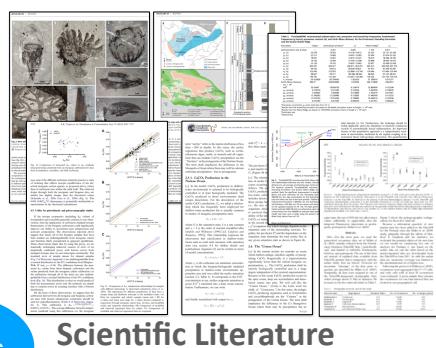
2. COSMOS: Knowledge extraction as a service



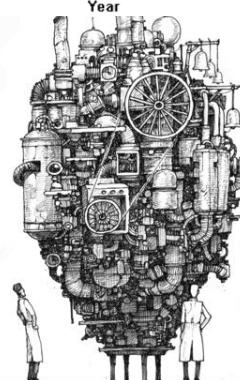
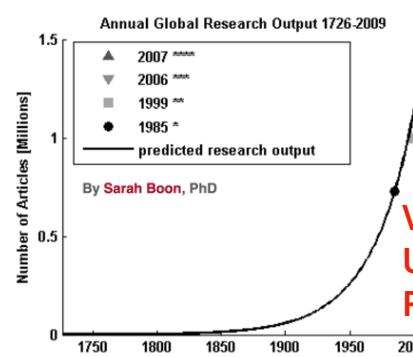
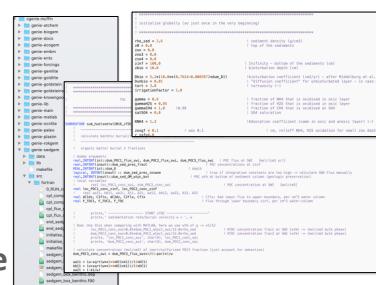
3. Demo: analyzing model code with COSMOS



Scientific workflows have a major bottleneck...



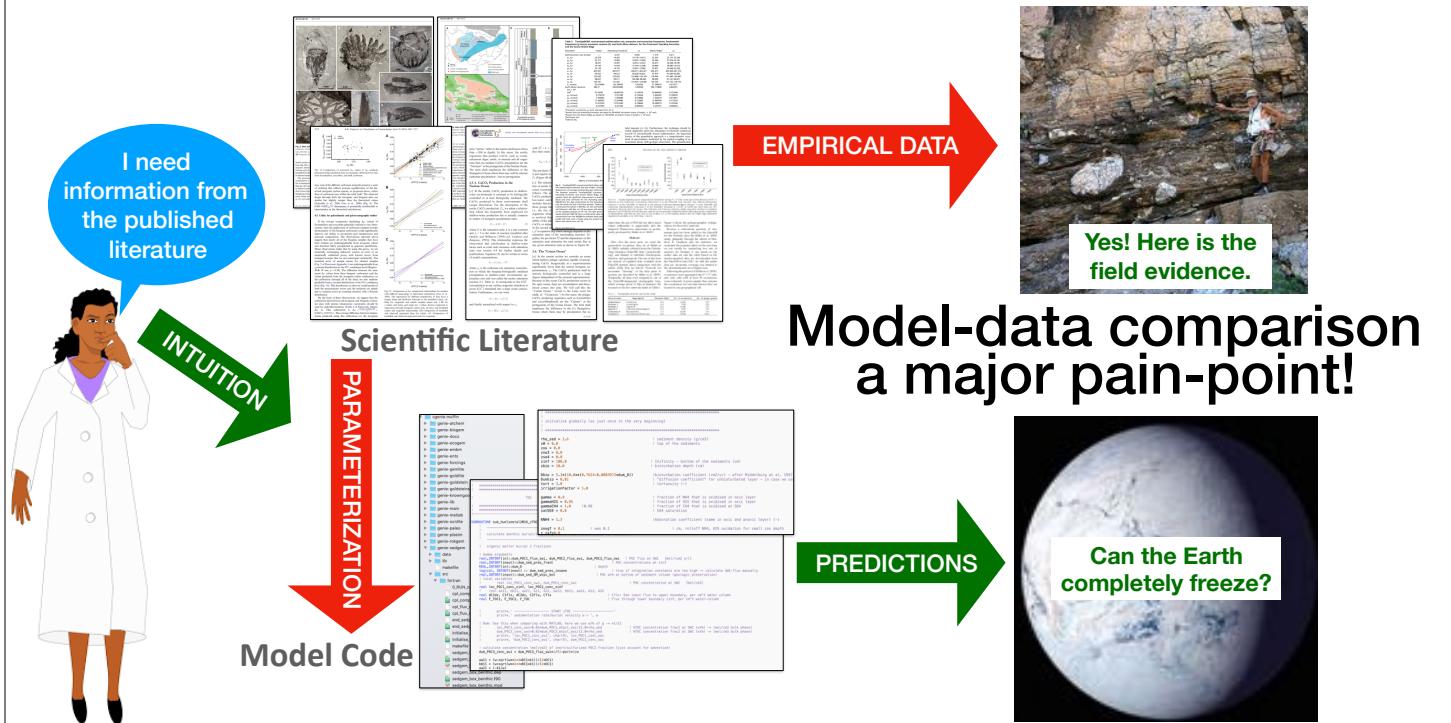
Model Code



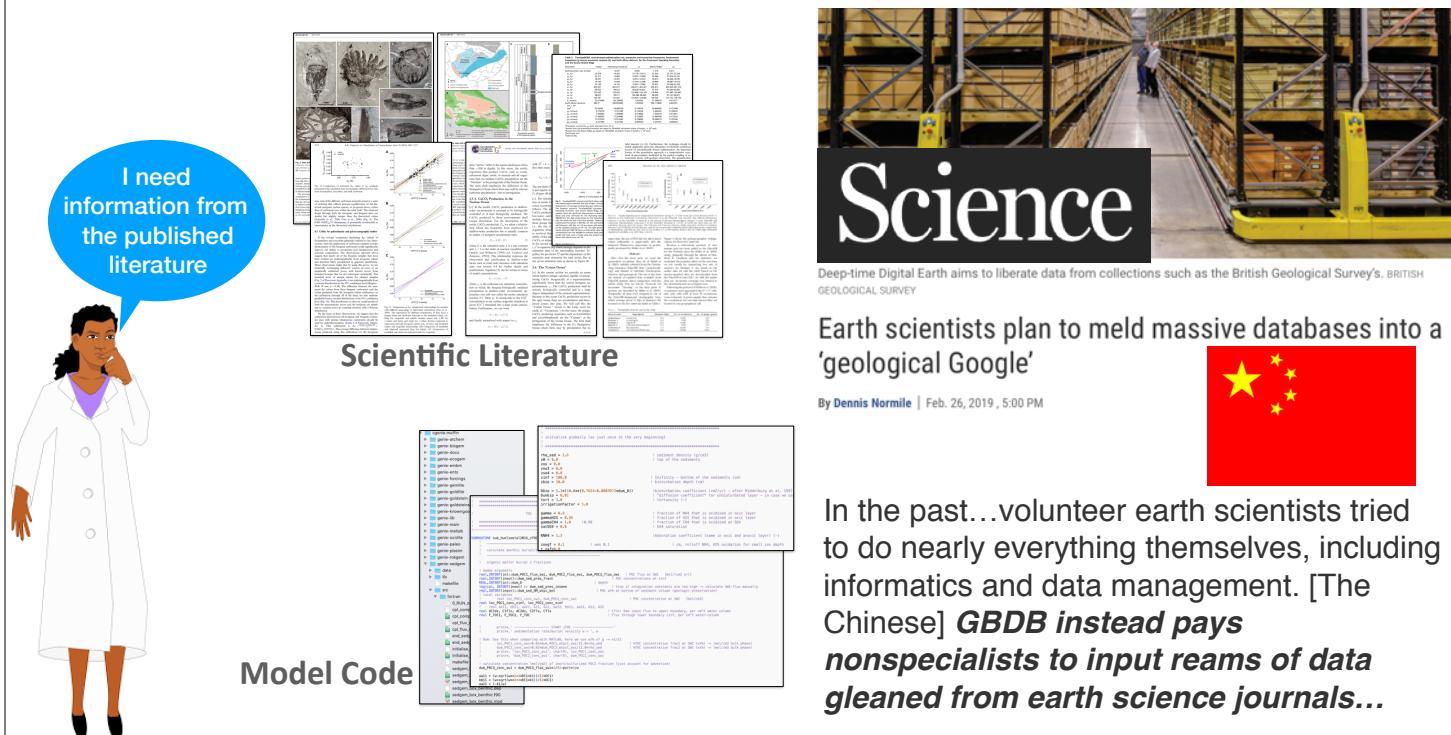
Monolithic
Complex
Difficult to assess

Voluminous
Unstructured
Physically scattered

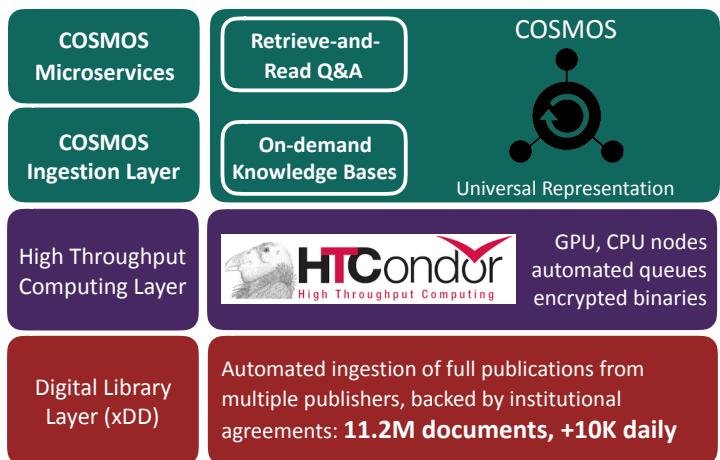
Scientific workflows have a major bottleneck...



Scientific workflows have a major bottleneck...



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



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

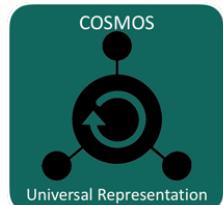
Correspondance:
Shanan Peters (peters@geology.wisc.edu)
Theodoros Rekatsinas (thodrek@cs.wisc.edu)
Miron Livny (miron@cs.wisc.edu)

Outline

1. Challenge: Produce an AI technical assistant



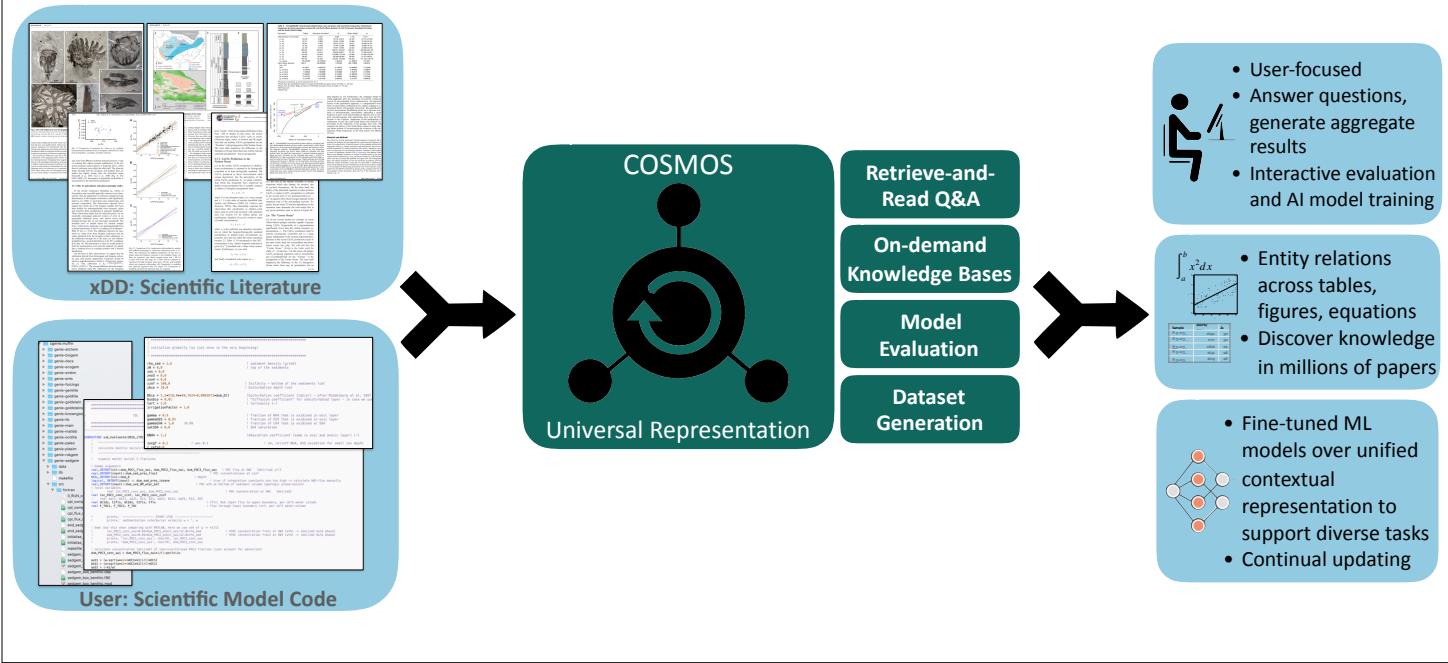
2. COSMOS: Knowledge extraction as a service



3. Demo: analyzing model code with COSMOS



Accelerating scientific discovery with COSMOS



Goal: Knowledge extraction from high-variety input

Equation

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

Equation

Context

Parameter

Values

Body Text

where R is $^{13}\text{C}/^{12}\text{C}$. The standard is Pee Dee Belemnite limestone that has been assigned a value of 0.0‰ . The precisions of $\delta^{13}\text{C}$ 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. PO₄-P was extracted filtrate by the ascorbic acid-Mo blue method (Strickland and Parsons, 1965), using a Technicon Auto Analyzer.

Section Header

3. Results

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 ($<500 \text{ m}^3 \text{ s}^{-1}$) until 22 June, and suddenly increased on 24 June (the first flood, $\sim 2000 \text{ m}^3 \text{ s}^{-1}$), reaching a peak flood on 28 June (the second flood, $\sim 3000 \text{ m}^3 \text{ s}^{-1}$). 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. $\delta^{13}\text{C}$ 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, $\delta^{13}\text{C}$ 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

	Discharge ($\text{m}^3 \text{ s}^{-1}$)	POC (mg l^{-1})	PON (mg l^{-1})	$\delta^{13}\text{C}$ (‰)	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

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.

Equation

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

Body Text

where R is $^{13}\text{C}/^{12}\text{C}$. The standard is Pee Dee Selemnite limestone that has been assigned a value of 0.07_{ppm} . The precision of $\delta^{13}\text{C}$ determination were less than 0.2_{ppm} . 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. PO₄-P was extracted filtrate by the ascorbic acid-Mn-blue method (Strickland and Parsons, 1965), using a Technicon Auto Analyzer.

Section Header
Results

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 ($<500 \text{ m}^3 \text{s}^{-1}$) until 22 June, and suddenly increased on 24 June (the first flood, $\sim 2000 \text{ m}^3 \text{s}^{-1}$), reaching a peak flood on 28 June (the second flood, $\sim 3000 \text{ m}^3 \text{s}^{-1}$). After that, it

Context

Table 1
Summary of physical and chemical variables in the Kiso rivers collected at $\sim 15 \text{ km}$ upstream from the river mouth

	Discharge ($\text{m}^3 \text{s}^{-1}$)	POC (mg l^{-1})	PON (mg l^{-1})	$\delta^{13}\text{C}$ (‰)	C/N (mol ratio)
Kiso River					Values
20 June	155	0.61	0.06	-27.7	12.6
28 June	1257	1.78	0.09	-25.5	22.3
4 July	269	0.30	0.05	-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
Be 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

Parameter

cosmos.ingestion module

Ingestion module for image03

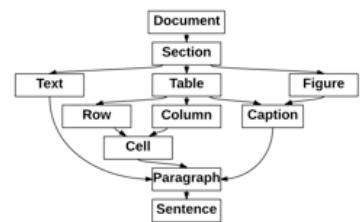
SHOW SOURCE ▾

Sub-modules

cosmos.ingestion.ingest_images

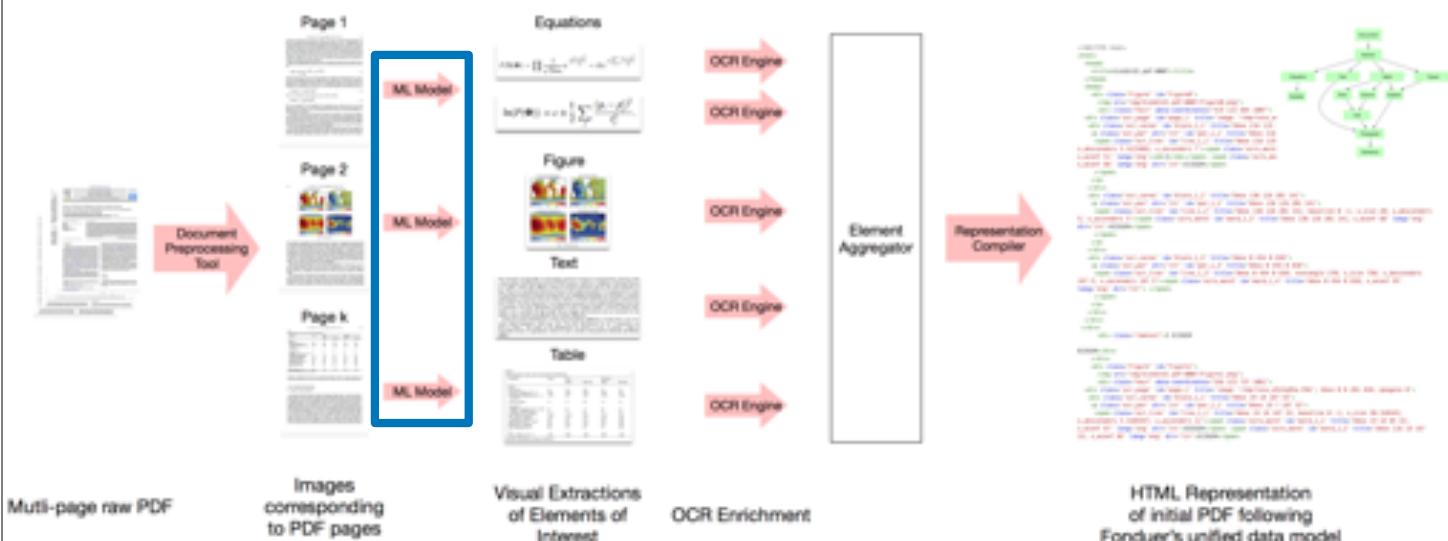
```
<div class="rawtext">A commercial sample (SSA-1) from Halliburton, Celle, Germany) containing (wt %) quartz 97.00, CaO 0.57, MgO 0.10, Al2O3 0.02(±0.01), and TiO2 0.00 (determined by X-ray diffraction). Microscopic analysis was used to estimate particle size. Silica flour was 1.657 cm2/g. Its average particle size (L158 value) was 32.7 nm. Specific density of the silica flour was found at 2.2 kg/l</div>
</div>
<div class="Equation" id="Equation8">

<div class="ocr_page" id="page_1" title="image_64071116_2719" data-score="0.5462715051">
<div class="ocr_carea" id="block_1" title="bbox_3 3 373 25">
<div class="ocr_scarea" id="line_1_1" title="bbox_3 3 373 25">
<div class="ocr_line" id="line_1_1" title="bbox_3 3 373 25" baseline="0" style="border: 1px solid black; position: relative; width: 370px; height: 25px; font-size: 1em; padding: 0 5px;">= div id="word_1_1" lang="eng" title="bbox_3 3 279 25">x_wcnf 63>ceil lang="eng" title="bbox_3 3 279 25">x_wcnf 82>ocid=</span> <span class="ocid" id="ocid_1_1" style="position: absolute; left: 0; top: 0; width: 100%; height: 100%;">?>retarder</span>
```



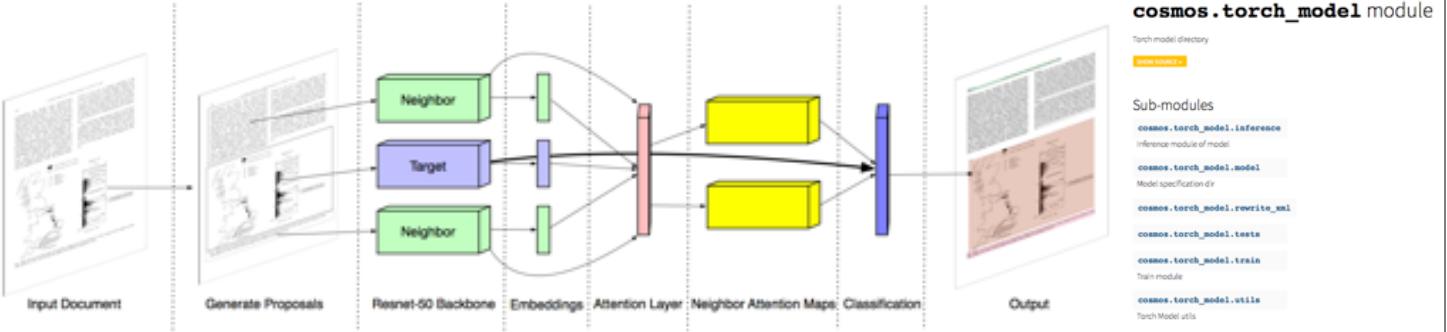
Output format: XML stored in MongoDB backend and queryable via ElasticSearch

Ingest Box: From PDF to XML

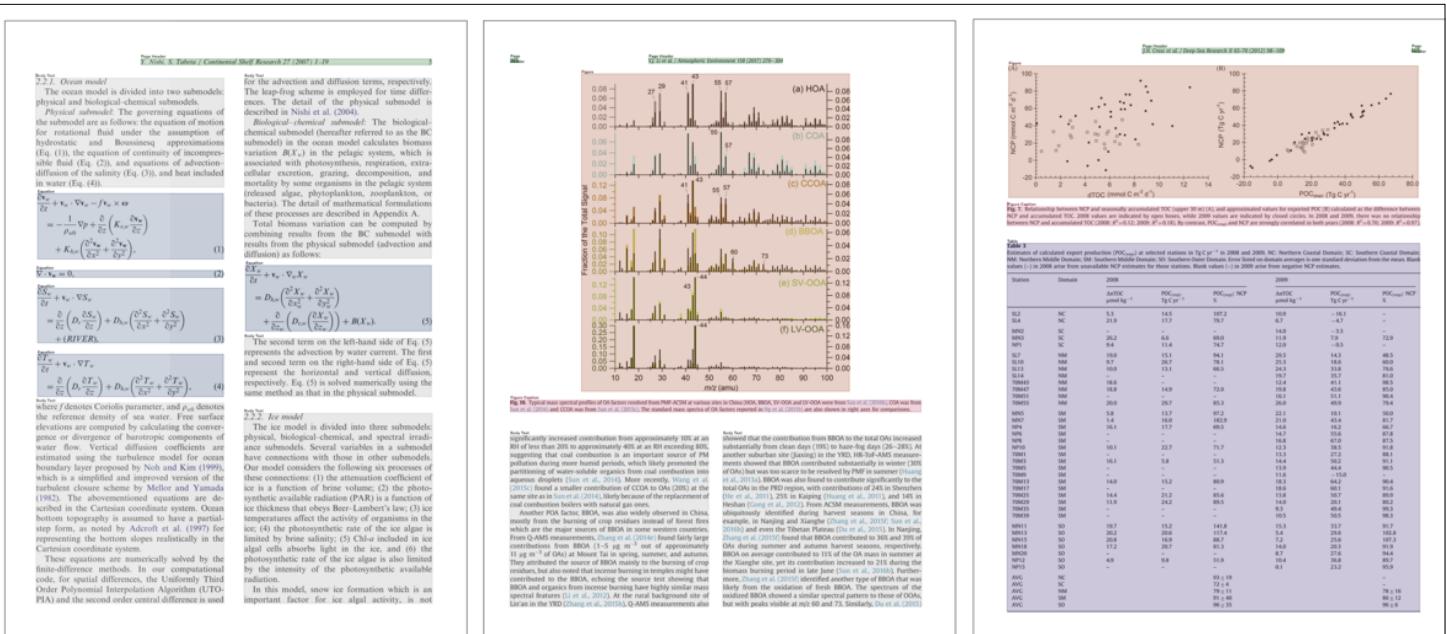


We need support for images to address the format heterogeneity across publications.
How do we get these elements from a scanned image?

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.

Equation

$$\text{Equation} = \frac{\text{POC}}{\text{POM}} = \frac{C_{\text{POC}}}{C_{\text{POM}}} = \frac{(R_{\text{sample}}/R_{\text{standard}}) - 1}{10} \times 10^3$$

where $R = 13^{\circ}\text{C}/1^{\circ}\text{C}$. The standard is Pure Deep-Bediente lime stone that has been assigned a value of 0.07°C . The precision of the POC/POM ratio was less than 2.7% . POC and PON concentrations were determined using a TCD detector attached to the elemental analyzer.

For Chl *a* and phytoplankton concentrations, TPO 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 *a* (Chl *a*) was determined as the total pigment including pheophytin. PO₄-P was extracted filtrate by the ascorbic acid-Mo blue method (Strickland and Parsons, 1965), using a Technicon Auto Analyzer.

Section Header
3. Results

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 ($< 500 \text{ m}^3 \text{s}^{-1}$) until 22 June, and suddenly increased on 24 June (the first flood, $\sim 2000 \text{ m}^3 \text{s}^{-1}$), reaching a peak flood on 28 June (the second flood, $\sim 3000 \text{ m}^3 \text{s}^{-1}$). 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. $\delta^{13}\text{C}$ 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, $\delta^{15}\text{N}$ of POM in the Ibi River remained fairly constant (ca. -2.6‰). The C/N ratios varied from 7.8 to 22.3 and reached the highest values during high discharge in all rivers.

River	Date	Discharge ($\text{m}^3 \text{s}^{-1}$)	POC (mg l^{-1})	PON (mg l^{-1})	$\delta^{13}\text{C}$ (‰)	C/N (mol ratio)	
Kiso River	20 June	155	0.61	0.06	-27.3	7.8	
	28 June	1257	1.78	0.09	-25.5	22.3	
	4 July	269	0.30	—	-23.1	12.5	
	Nagara River	20 June	63	2.28	0.34	-27.3	7.8
Ibi River	28 June	1072	2.11	0.13	-25.9	8.7	
	4 July	129	0.44	—	-26.0	8.7	
	Bri River	21 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	

Equation

Parameter

Values

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

Ingestion augmentation services	Knowledge extraction services
<p>X Specific OCR le for doing ocr for latex</p>	<p>Branch: v0.3.0 → Cosmos / services /</p>
<p>Table and Figure Knowledge Base Construction le to construct tables and figures csvs</p>	<ul style="list-style-type: none"><input checked="" type="checkbox"/> question_answering_backend<input checked="" type="checkbox"/> search_backend<input checked="" type="checkbox"/> table_extractions<input checked="" type="checkbox"/> word_embeddings
<p>Equation/Body Text Knowledge Base Construction uct a knowledge base from equations and body text.</p>	

Services: Knowledge extraction (equations/tables)

Input: Raw High-Variety Inputs

Compound	Parameter	$\log K_{ow}$
Glycine	Context	-3.24
Glycerol	Context	-2.57
Methanol	Context	-0.74
1,4-Dioxane	Context	-0.42
Ethanol	Context	-0.32
Acetone	Context	-0.24
2-Propanol	Context	-0.1
2-Butanol	Values	0.74
2-Pentanol	Values	1.25

for hydrophobic, nonionogenic analytes ($f(A_{COOH})$ approaches 0) a value of ΔK_d is a function of

$$\Delta K_d = -0.38 \log K_{ow} - 0.26 \quad (5)$$

The found relationship displays an existence of the specific sorption on the gel (negative slope of the plot), the effect of which is strengthened with increasing hydrophobicity of the analyte.

To establish the form of the relationship between ΔK_d and A_{COOH} , a set of carboxylic acids with

$\log K_{\text{ow}} < -0.5$ was used (Table 3). The range of $\log K_{\text{ow}}$ values close to or less than -0.5 were chosen according to the observations described above. The obtained relationship is given in Fig. 4. It can be seen that the plot of ΔK_d versus A_{COOH} is well

Output: Extracted Structured Knowledge

		DataFrame1	
EID	Expression	Context	Values
E1	$\Delta K_d = -0.38 \log K_{ow} - 0.26$	Glucose	-3.24
E2	$G_0 = \left[1 + (a + b(\text{RH}*0.01) + c(\text{RH}*0.01)^2) * \frac{\text{RH}}{100 - \text{RH}} \right]^{1/3}$	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-Pentanol	1.25

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

Services: Model extraction

Equation

$$C = C_0 + F \times \frac{t}{H} \quad (2)$$

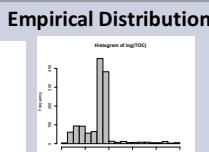
where C is the headspace concentration of CO_2 at time t , C_0 is its initial head-space concentration, F is the CO_2 flux, H is the height of the head-space layer in the chamber. The fluxes of

- The parse tree for extracted equation (red-colored bounding box) reveals symbols C , t , C_0 , 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).

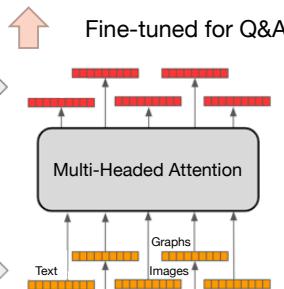
Services: Open-domain question answering

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.



Entity-Aware Code Analysis



Model: trained with self-supervised learning

TOC Analysis

What is TOC?

Answer : total organic carbon
Confidence : 98%

Doc Title : Dynamics of stream water TOC concentrations scenarios

Extracted empirical distribution over TOC

Tabular data for TOC

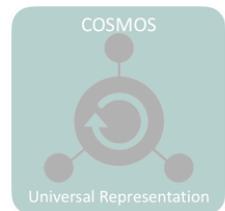
Depth	TOC
0	0.0
1	0.0
2	0.0
3	0.0
4	0.0
5	0.0
6	0.0
7	0.0
8	0.0
9	0.0
10	0.0
11	0.0
12	0.0
13	0.0
14	0.0
15	0.0
16	0.0
17	0.0
18	0.0
19	0.0
20	0.0
21	0.0
22	0.0
23	0.0
24	0.0
25	0.0
26	0.0
27	0.0
28	0.0
29	0.0
30	0.0
31	0.0
32	0.0
33	0.0
34	0.0
35	0.0
36	0.0
37	0.0
38	0.0
39	0.0
40	0.0
41	0.0
42	0.0
43	0.0
44	0.0
45	0.0
46	0.0
47	0.0
48	0.0
49	0.0
50	0.0
51	0.0
52	0.0
53	0.0
54	0.0
55	0.0
56	0.0
57	0.0
58	0.0
59	0.0
60	0.0
61	0.0
62	0.0
63	0.0
64	0.0
65	0.0
66	0.0
67	0.0
68	0.0
69	0.0
70	0.0
71	0.0
72	0.0
73	0.0
74	0.0
75	0.0
76	0.0
77	0.0
78	0.0
79	0.0
80	0.0
81	0.0
82	0.0
83	0.0
84	0.0
85	0.0
86	0.0
87	0.0
88	0.0
89	0.0
90	0.0
91	0.0
92	0.0
93	0.0
94	0.0
95	0.0
96	0.0
97	0.0
98	0.0
99	0.0
100	0.0
101	0.0
102	0.0
103	0.0
104	0.0
105	0.0
106	0.0
107	0.0
108	0.0
109	0.0
110	0.0
111	0.0
112	0.0
113	0.0
114	0.0
115	0.0
116	0.0
117	0.0
118	0.0
119	0.0
120	0.0
121	0.0
122	0.0
123	0.0
124	0.0
125	0.0
126	0.0
127	0.0
128	0.0
129	0.0
130	0.0
131	0.0
132	0.0
133	0.0
134	0.0
135	0.0
136	0.0
137	0.0
138	0.0
139	0.0
140	0.0
141	0.0
142	0.0
143	0.0
144	0.0
145	0.0
146	0.0
147	0.0
148	0.0
149	0.0
150	0.0
151	0.0
152	0.0
153	0.0
154	0.0
155	0.0
156	0.0
157	0.0
158	0.0
159	0.0
160	0.0
161	0.0
162	0.0
163	0.0
164	0.0
165	0.0
166	0.0
167	0.0
168	0.0
169	0.0
170	0.0
171	0.0
172	0.0
173	0.0
174	0.0
175	0.0
176	0.0
177	0.0
178	0.0
179	0.0
180	0.0
181	0.0
182	0.0
183	0.0
184	0.0
185	0.0
186	0.0
187	0.0
188	0.0
189	0.0
190	0.0
191	0.0
192	0.0
193	0.0
194	0.0
195	0.0
196	0.0
197	0.0
198	0.0
199	0.0
200	0.0
201	0.0
202	0.0
203	0.0
204	0.0
205	0.0
206	0.0
207	0.0
208	0.0
209	0.0
210	0.0
211	0.0
212	0.0
213	0.0
214	0.0
215	0.0
216	0.0
217	0.0
218	0.0
219	0.0
220	0.0
221	0.0
222	0.0
223	0.0
224	0.0
225	0.0
226	0.0
227	0.0
228	0.0
229	0.0
230	0.0
231	0.0
232	0.0
233	0.0
234	0.0
235	0.0
236	0.0
237	0.0
238	0.0
239	0.0
240	0.0
241	0.0
242	0.0
243	0.0
244	0.0
245	0.0
246	0.0
247	0.0
248	0.0
249	0.0
250	0.0
251	0.0
252	0.0
253	0.0
254	0.0
255	0.0
256	0.0
257	0.0
258	0.0
259	0.0
260	0.0
261	0.0
262	0.0
263	0.0
264	0.0
265	0.0
266	0.0
267	0.0
268	0.0
269	0.0
270	0.0
271	0.0
272	0.0
273	0.0
274	0.0
275	0.0
276	0.0
277	0.0
278	0.0
279	0.0
280	0.0
281	0.0
282	0.0
283	0.0
284	0.0
285	0.0
286	0.0
287	0.0
288	0.0
289	0.0
290	0.0
291	0.0
292	0.0
293	0.0
294	0.0
295	0.0
296	0.0
297	0.0
298	0.0
299	0.0
300	0.0
301	0.0
302	0.0
303	0.0
304	0.0
305	0.0
306	0.0
307	0.0
308	0.0
309	0.0
310	0.0
311	0.0
312	0.0
313	0.0
314	0.0
315	0.0
316	0.0
317	0.0
318	0.0
319	0.0
320	0.0
321	0.0
322	0.0
323	0.0
324	0.0
325	0.0
326	0.0
327	0.0
328	0.0
329	0.0
330	0.0
331	0.0
332	0.0
333	0.0
334	0.0
335	0.0
336	0.0
337	0.0
338	0.0
339	0.0
340	0.0
341	0.0
342	0.0
343	0.0
344	0.0
345	0.0
346	0.0
347	0.0
348	0.0
349	0.0
350	0.0
351	0.0
352	0.0
353	0.0
354	0.0
355	0.0
356	0.0
357	0.0
358	0.0
359	0.0
360	0.0
361	0.0
362	0.0
363	0.0
364	0.0
365	0.0
366	0.0
367	0.0
368	0.0
369	0.0
370	0.0
371	0.0
372	0.0
373	0.0
374	0.0
375	0.0
376	0.0
377	0.0
378	0.0
379	0.0
380	0.0
381	0.0
382	0.0
383	0.0
384	0.0
385	0.0
386	0.0
387	0.0
388	0.0
389	0.0
390	0.0
391	0.0
392	0.0
393	0.0
394	0.0
395	0.0
396	0.0
397	0.0
398	0.0
399	0.0
400	0.0
401	0.0
402	0.0
403	0.0
404	0.0
405	0.0
406	0.0
407	0.0
408	0.0
409	0.0
410	0.0
411	0.0
412	0.0
413	0.0
414	0.0
415	0.0
416	0.0
417	0.0
418	0.0
419	0.0
420	0.0
421	0.0
422	0.0
423	0.0
424	0.0
425	0.0
426	0.0
427	0.0
428	0.0
429	0.0
430	0.0
431	0.0
432	0.0
433	0.0
434	0.0
435	0.0
436	0.0
437	0.0
438	0.0
439	0.0
440	0.0
441	0.0
442	0.0
443	0.0
444	0.0
445	0.0
446	0.0
447	0.0
448	0.0
449	0.0
450	0.0
451	0.0
452	0.0
453	0.0
454	0.0
455	0.0
456	0.0
457	0.0
458	0.0
459	0.0
460	0.0
461	0.0
462	0.0
463	0.0
464	0.0
465	0.0
466	0.0
467	0.0
468	0.0
469	0.0
470	0.0
471	0.0
472	0.0
473	0.0
474	0.0
475	0.0
476	0.0
477	0.0
478	0.0
479	0.0
480	0.0
481	0.0
482	0.0
483	0.0
484	0.0
485	0.0
486	0.0
487	0.0
488	0.0
489	0.0
490	0.0
491	0.0
492	0.0
493	0.0
494	0.0
495	0.0
496	0.0
497	0.0
498	0.0
499	0.0
500	0.0
501	0.0
502	0.0
503	0.0
504	0.0
505	0.0
506	0.0
507	0.0
508	0.0
509	0.0
510	0.0
511	0.0
512	0.0
513	0.0
514	0.0
515	0.0
516	0.0
517	0.0
518	0.0
519	0.0
520	0.0
521	0.0
522	0

Outline

1. Challenge: Produce an AI technical assistant



2. COSMOS: Knowledge extraction as a service

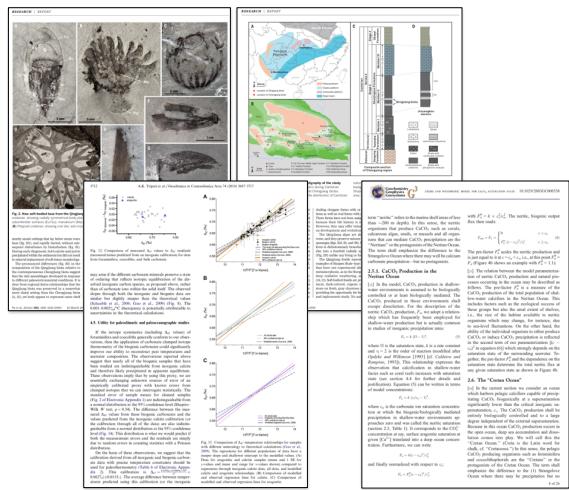


3. Demo: analyzing model code with COSMOS



Fundamental challenges

Direction 1: Retrieval for open-domain scientific discovery

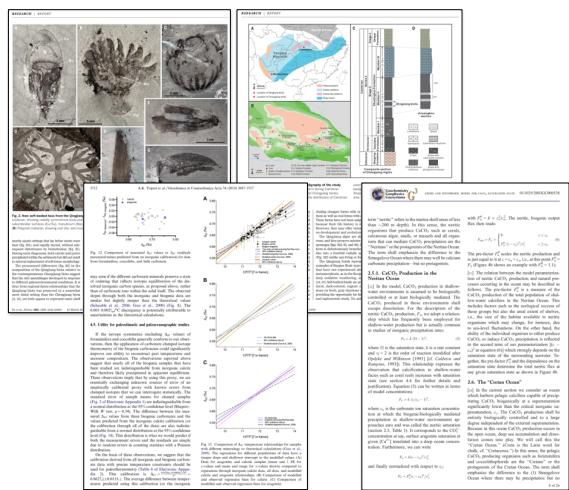


Question answering assumes strong supervision of the supporting evidence
(i.e., retrieval is solved)

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



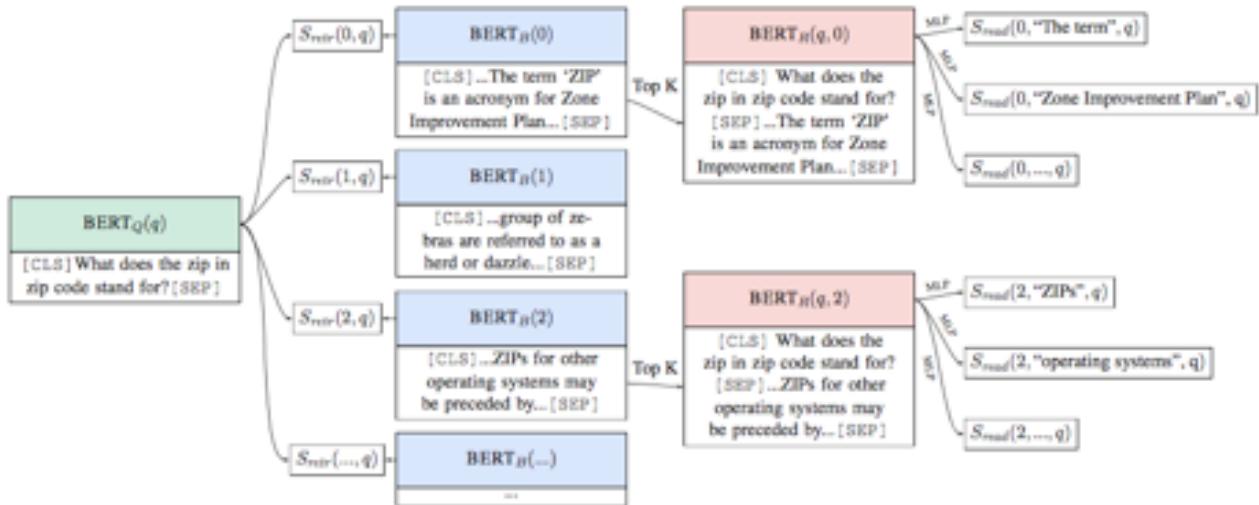
How can we leverage broader context for accurate retrieval?

Example Paper

Source Code

Sample	Sampling Location	Sampling Date	Sampling Time	Sampling Depth	Sampling Type	Sampling Method	Sampling Frequency	Sampling Interval	Sampling Duration	Sampling Notes
S1	North Sea	2023-01-01	08:00	0-10m	Surface	Net	1	1 day	24 hours	High productivity
S2	North Sea	2023-01-02	10:00	0-10m	Surface	Net	1	1 day	24 hours	Medium productivity
S3	North Sea	2023-01-03	12:00	0-10m	Surface	Net	1	1 day	24 hours	Low productivity
S4	North Sea	2023-01-04	14:00	0-10m	Surface	Net	1	1 day	24 hours	Very low productivity
S5	North Sea	2023-01-05	16:00	0-10m	Surface	Net	1	1 day	24 hours	Extremely low productivity
S6	North Sea	2023-01-06	18:00	0-10m	Surface	Net	1	1 day	24 hours	No productivity
S7	North Sea	2023-01-07	20:00	0-10m	Surface	Net	1	1 day	24 hours	No productivity
S8	North Sea	2023-01-08	22:00	0-10m	Surface	Net	1	1 day	24 hours	No productivity
S9	North Sea	2023-01-09	00:00	0-10m	Surface	Net	1	1 day	24 hours	No productivity
S10	North Sea	2023-01-10	02:00	0-10m	Surface	Net	1	1 day	24 hours	No productivity
S11	North Sea	2023-01-11	04:00	0-10m	Surface	Net	1	1 day	24 hours	No productivity
S12	North Sea	2023-01-12	06:00	0-10m	Surface	Net	1	1 day	24 hours	No productivity
S13	North Sea	2023-01-13	08:00	0-10m	Surface	Net	1	1 day	24 hours	No productivity
S14	North Sea	2023-01-14	10:00	0-10m	Surface	Net	1	1 day	24 hours	No productivity
S15	North Sea	2023-01-15	12:00	0-10m	Surface	Net	1	1 day	24 hours	No productivity
S16	North Sea	2023-01-16	14:00	0-10m	Surface	Net	1	1 day	24 hours	No productivity
S17	North Sea	2023-01-17	16:00	0-10m	Surface	Net	1	1 day	24 hours	No productivity
S18	North Sea	2023-01-18	18:00	0-10m	Surface	Net	1	1 day	24 hours	No productivity
S19	North Sea	2023-01-19	20:00	0-10m	Surface	Net	1	1 day	24 hours	No productivity
S20	North Sea	2023-01-20	22:00	0-10m	Surface	Net	1	1 day	24 hours	No productivity
S21	North Sea	2023-01-21	00:00	0-10m	Surface	Net	1	1 day	24 hours	No productivity
S22	North Sea	2023-01-22	02:00	0-10m	Surface	Net	1	1 day	24 hours	No productivity
S23	North Sea	2023-01-23	04:00	0-10m	Surface	Net	1	1 day	24 hours	No productivity
S24	North Sea	2023-01-24	06:00	0-10m	Surface	Net	1	1 day	24 hours	No productivity
S25	North Sea	2023-01-25	08:00	0-10m	Surface	Net	1	1 day	24 hours	No productivity
S26	North Sea	2023-01-26	10:00	0-10m	Surface	Net	1	1 day	24 hours	No productivity
S27	North Sea	2023-01-27	12:00	0-10m	Surface	Net	1	1 day	24 hours	No productivity
S28	North Sea	2023-01-28	14:00	0-10m	Surface	Net	1	1 day	24 hours	No productivity
S29	North Sea	2023-01-29	16:00	0-10m	Surface	Net	1	1 day	24 hours	No productivity
S30	North Sea	2023-01-30	18:00	0-10m	Surface	Net	1	1 day	24 hours	No productivity
S31	North Sea	2023-01-31	20:00	0-10m	Surface	Net	1	1 day	24 hours	No productivity
S32	North Sea	2023-02-01	22:00	0-10m	Surface	Net	1	1 day	24 hours	No productivity
S33	North Sea	2023-02-02	00:00	0-10m	Surface	Net	1	1 day	24 hours	No productivity
S34	North Sea	2023-02-03	02:00	0-10m	Surface	Net	1	1 day	24 hours	No productivity
S35	North Sea	2023-02-04	04:00	0-10m	Surface	Net	1	1 day	24 hours	No productivity
S36	North Sea	2023-02-05	06:00	0-10m	Surface	Net	1	1 day	24 hours	No productivity
S37	North Sea	2023-02-06	08:00	0-10m	Surface	Net	1	1 day	24 hours	No productivity
S38	North Sea	2023-02-07	10:00	0-10m	Surface	Net	1	1 day	24 hours	No productivity
S39	North Sea	2023-02-08	12:00	0-10m	Surface	Net	1	1 day	24 hours	No productivity
S40	North Sea	2023-02-09	14:00	0-10m	Surface	Net	1	1 day	24 hours	No productivity
S41	North Sea	2023-02-10	16:00	0-10m	Surface	Net	1	1 day	24 hours	No productivity
S42	North Sea	2023-02-11	18:00	0-10m	Surface	Net	1	1 day	24 hours	No productivity
S43	North Sea	2023-02-12	20:00	0-10m	Surface	Net	1	1 day	24 hours	No productivity
S44	North Sea	2023-02-13	22:00	0-10m	Surface	Net	1	1 day	24 hours	No productivity
S45	North Sea	2023-02-14	00:00	0-10m	Surface	Net	1	1 day	24 hours	No productivity
S46	North Sea	2023-02-15	02:00	0-10m	Surface	Net	1	1 day	24 hours	No productivity
S47	North Sea	2023-02-16	04:00	0-10m	Surface	Net	1	1 day	24 hours	No productivity
S48	North Sea	2023-02-17	06:00	0-10m	Surface	Net	1	1 day	24 hours	No productivity
S49	North Sea	2023-02-18	08:00	0-10m	Surface	Net	1	1 day	24 hours	No productivity
S50	North Sea	2023-02-19	10:00	0-10m	Surface	Net	1	1 day	24 hours	No productivity
S51	North Sea	2023-02-20	12:00	0-10m	Surface	Net	1	1 day	24 hours	No productivity
S52	North Sea	2023-02-21	14:00	0-10m	Surface	Net	1	1 day	24 hours	No productivity
S53	North Sea	2023-02-22	16:00	0-10m	Surface	Net	1	1 day	24 hours	No productivity
S54	North Sea	2023-02-23	18:00	0-10m	Surface	Net	1	1 day	24 hours	No productivity
S55	North Sea	2023-02-24	20:00	0-10m	Surface	Net	1	1 day	24 hours	No productivity
S56	North Sea	2023-02-25	22:00	0-10m	Surface	Net	1	1 day	24 hours	No productivity
S57	North Sea	2023-02-26	00:00	0-10m	Surface	Net	1	1 day	24 hours	No productivity
S58	North Sea	2023-02-27	02:00	0-10m	Surface	Net	1	1 day	24 hours	No productivity
S59	North Sea	2023-02-28	04:00	0-10m	Surface	Net	1	1 day	24 hours	No productivity
S60	North Sea	2023-02-29	06:00	0-10m	Surface	Net	1	1 day	24 hours	No productivity
S61	North Sea	2023-03-01	08:00	0-10m	Surface	Net	1	1 day	24 hours	No productivity
S62	North Sea	2023-03-02	10:00	0-10m	Surface	Net	1	1 day	24 hours	No productivity
S63	North Sea	2023-03-03	12:00	0-10m	Surface	Net	1	1 day	24 hours	No productivity
S64	North Sea	2023-03-04	14:00	0-10m	Surface	Net	1	1 day	24 hours	No productivity
S65	North Sea	2023-03-05	16:00	0-10m	Surface	Net	1	1 day	24 hours	No productivity
S66	North Sea	2023-03-06	18:00	0-10m	Surface	Net	1	1 day	24 hours	No productivity
S67	North Sea	2023-03-07	20:00	0-10m	Surface	Net	1	1 day	24 hours	No productivity
S68	North Sea	2023-03-08	22:00	0-10m	Surface	Net	1	1 day	24 hours	No productivity
S69	North Sea	2023-03-09	00:00	0-10m	Surface	Net	1	1 day	24 hours	No productivity
S70	North Sea	2023-03-10	02:00	0-10m	Surface	Net	1	1 day	24 hours	No productivity
S71	North Sea	2023-03-11	04:00	0-10m	Surface	Net	1	1 day	24 hours	No productivity
S72	North Sea	2023-03-12	06:00	0-10m	Surface	Net	1	1 day	24 hours	No productivity
S73	North Sea	2023-03-13	08:00	0-10m	Surface	Net	1	1 day	24 hours	No productivity
S74	North Sea	2023-03-14	10:00	0-10m	Surface	Net	1	1 day	24 hours	No productivity
S75	North Sea	2023-03-15	12:00	0-10m	Surface	Net	1	1 day	24 hours	No productivity
S76	North Sea	2023-03-16	14:00	0-10m	Surface	Net	1	1 day	24 hours	No productivity
S77	North Sea	2023-03-17	16:00	0-10m	Surface	Net	1	1 day	24 hours	No productivity
S78	North Sea	2023-03-18	18:00	0-10m	Surface	Net	1	1 day	24 hours	No productivity
S79	North Sea	2023-03-19	20:00	0-10m	Surface	Net	1	1 day	24 hours	No productivity
S80	North Sea	2023-03-20	22:00	0-10m	Surface	Net	1	1 day	24 hours	No productivity
S81	North Sea	2023-03-21	00:00	0-10m	Surface	Net	1	1 day	24 hours	No productivity
S82	North Sea	2023-03-22	02:00	0-10m	Surface	Net	1	1 day	24 hours	No productivity
S83	North Sea	2023-03-23	04:00	0-10m	Surface	Net	1	1 day	24 hours	No productivity
S84	North Sea	2023-03-24	06:00	0-10m	Surface	Net	1	1 day	24 hours	No productivity
S85	North Sea	2023-03-25	08:00	0-10m	Surface	Net	1	1 day	24 hours	No productivity
S86	North Sea	2023-03-26	10:00	0-10m	Surface	Net	1	1 day	24 hours	No productivity
S87	North Sea	2023-03-27	12:00	0-10m	Surface	Net	1	1 day	24 hours	No productivity
S88	North Sea	2023-03-28	14:00	0-10m	Surface	Net	1	1 day	24 hours	No productivity
S89	North Sea	2023-03-29	16:00	0-10m	Surface	Net	1	1 day	24 hours	No productivity
S90	North Sea	2023-03-30	18:00	0-10m	Surface	Net	1	1 day	24 hours	No productivity
S91	North Sea	2023-03-31	20:00	0-10m	Surface	Net	1	1 day	24 hours	No productivity
S92	North Sea	2023-04-01	22:00	0-10m	Surface	Net	1	1 day	24 hours	No productivity
S93	North Sea	2023-04-02	00:00	0-10m	Surface	Net	1	1 day	24 hours	No productivity
S94	North Sea	2023-04-03	02:00	0-10m	Surface	Net	1	1 day	24 hours	No productivity
S95	North Sea	2023-04-04	04:00	0-10m	Surface	Net	1	1 day	24 hours	No productivity
S96	North Sea	2023-04-05	06:00	0-10m	Surface	Net	1	1 day	24 hours	No productivity
S97	North Sea	2023-04-06	08:00	0-10m	Surface	Net	1	1 day	24 hours	No productivity
S98	North Sea	2023-04-07	10:00	0-10m	Surface	Net	1	1 day	24 hours	No productivity
S99	North Sea	2023-04-08	12:00	0-10m	Surface	Net	1	1 day	24 hours	No productivity
S100	North Sea	2023-04-09	14:00	0-10m	Surface	Net	1	1 day	24 hours	No productivity
S101	North Sea	2023-04-10	16:00	0-10m	Surface	Net	1	1 day	24 hours	No productivity
S102	North Sea	2023-04-11	18:00	0-10m	Surface	Net	1	1 day	24 hours	No productivity
S103	North Sea	2023-04-12	20:00	0-10m	Surface	Net	1	1 day	24 hours	No productivity
S104	North Sea	2023-04-13	22:00	0-10m	Surface	Net	1	1 day	24 hours	No productivity
S105	North Sea	2023-04-14	00:00	0-10m	Surface	Net	1	1 day	24 hours	No productivity
S106	North Sea	2023-04-15	02:00	0-10m	Surface	Net	1	1 day	24 hours	No productivity
S107	North Sea	2023-04-16	04:00	0-10m	Surface	Net	1	1 day	24 hours	No productivity
S108	North Sea	2023-04-17	06:00	0-10m	Surface	Net	1	1 day	24 hours	No productivity
S109	North Sea	2023-04-18	08:00	0-10m	Surface	Net	1	1 day	24 hours	No productivity
S110	North Sea	2023-04-19	10:00	0-10m	Surface	Net	1	1 day	24 hours	No productivity
S111	North Sea	2023-04-20	12:00	0-10m	Surface	Net	1	1 day	24 hours	No productivity
S112	North Sea	2023-04-21	14:00	0-10m	Surface					

Direction 1: Retrieval for open-domain scientific discovery



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

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

Mammogram BI-RADS: Overall: 1 - Negative.

As the teaching physician, I personally examined the radiologic study, reviewed the findings, and interpreted.

EXAM:
MAMMOGRAM DIGITAL SCREENING BI-LATERAL W/TOMOSYNTHESIS

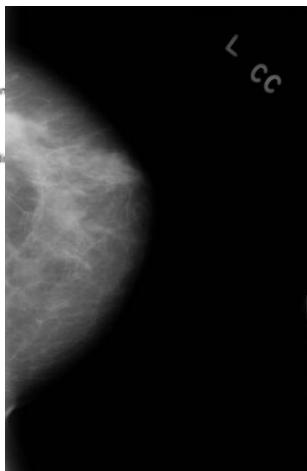
INDICATION:
Screening.

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)

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

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

Mammogram BI-RADS: Overall: 1 - Negative.

As the teaching physician, I personally examined the radiologic study, reviewed the findings and interpretation.

EXAM:
MAMMOGRAM DIGITAL SCREENING BILATERAL W/TOMOSYNTHESIS

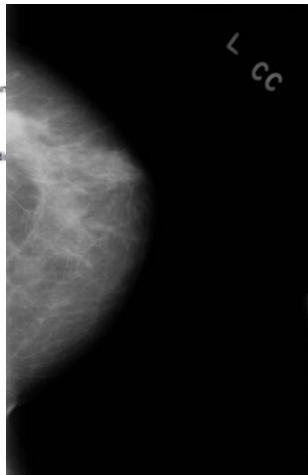
INDICATION:
Screening.

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 combined 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_{CB}	40	Vdc
Emitter – Base Voltage	V_{EBC}	5.0	Vdc
Collector Current – Continuous	I_C	200	mAdc
Total Device Dissipation ($\oplus T_A = 25^\circ\text{C}$)	P_D	420	mW
Derate above 25°C		5.0	mW/ $^\circ\text{C}$
Total Power Dissipation ($\oplus T_A = 25^\circ\text{C}$)	P_D	250	mW
Total Device Dissipation ($\oplus T_A = 25^\circ\text{C}$)	P_D	1.5	W
Derate above 25°C		12	mW/ $^\circ\text{C}$
Operating and Storage Junction Temperature Range	T_J, T_{SJ}	-55 to +150	$^\circ\text{C}$

2N3906-D.PDF

Absolute Maximum Ratings*			
Symbol	Parameter	Value	Units
	Collector-Emitter Voltage	50	V
	Collector-Base Voltage	50	V
	Emitter-Base Voltage	5.0	V
	Collector Current (Continuous)	200	mA
	T_J, T_{SJ} Operating and Storage Junction Temperature Range	-55 to +150	$^\circ\text{C}$

MMBT3904.PDF

Absolute maximum ratings			
Characteristic	Symbol	Rating	Unit
Collector-Emitter voltage	V_{CE}	400	V
Collector-Base voltage	V_{CB}	400	V
Emitter-Base voltage	V_{EB}	-5	V
Collector current	I_C	200	mA
Collector dissipation	P_D	600	mW
Collector-emitter reverse	V_{CEO}	50	V
Storage temperature range	T_S	-55/150	$^\circ\text{C}$

AUHKCS04635-1.pdf

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

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	mA/dc
Total Device Dissipation @ T _A = 25°C Derate above 25°C	P _D	625	mW mW/°C
Total Power Dissipation @ T _A = 60°C	P _D	250	mW
Total Device Dissipation @ T _A = 25°C Derate above 25°C	P _D	1.5	W
Operating and Storage Junction Temperature Range	T _J , T _{SJ}	-55 to +150	°C

2N3906-D.PDF

Absolute Maximum Ratings*			
Symbol	Parameter	Value	Units
V _{CE}	Collector-Emitter Voltage	50	V
V _{CB}	Collector-Base Voltage	50	V
V _{EB}	Emitter-Base Voltage	5.0	V
I _C	Collector Current - Continuous	200	mA
T _J , T _{SJ}	Operating and Storage Junction Temperature Range	-55 to +150	°C

MMBT3904.PDF

Absolute maximum ratings			
Symbol	Symbol	Rating	Unit
Collector-emitter voltage	V _{CE}	40	V
Collector-base voltage	V _{CB}	50	V
Emitter-base voltage	V _{EB}	5.0	V
Collector current	I _C	200	mA
Collector dissipation	P _D	625	mW
Operating temperature	T _J	-55 to +150	°C
Storage temperature range	T _{SJ}	-55 to +150	°C

AUKC04635-1.pdf

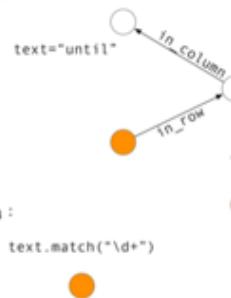
Task

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

Output

Ensemble of programs synthesized from the input examples

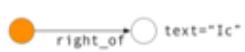
$\lambda_1 :$



$\lambda_2 :$



$\lambda_3 :$



$\lambda_4 :$



Weak Supervision powered by Program Synthesis

Direction 3: On demand extraction of observational data

The Hera Engine

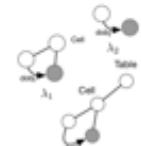
Input Documents



Phase1 Representation

Synthesis Engine

Input Few Annotated Examples



Output Ensemble of Programs

Data Programming

Program $\lambda_1 \quad \lambda_2 \quad \lambda_3$

Weights (Pos.)

$n_1 \quad 1 \quad 0 \quad 1$

$n_2 \quad -1 \quad -1 \quad 0$

$n_3 \quad 0 \quad 0 \quad 1$

Input Matrix of Labels

Program $\lambda_1 \quad \lambda_2 \quad \lambda_3$

$n_1 \quad 0.7 \quad 0.1 \quad 0.2$

$n_2 \quad 0.1 \quad 0.9 \quad 0.7$

$n_3 \quad 0.2 \quad 0.7 \quad 0.1$

Output Learned Accuracy of Programs

$P(\hat{y} = 1)$

$n_1 \quad 0.8 \quad 0.9 \quad 0.7$

$n_2 \quad 0.9 \quad 0.8 \quad 0.9$

$n_3 \quad 0.7 \quad 0.9 \quad 0.8$

Output Probabilistic Labels

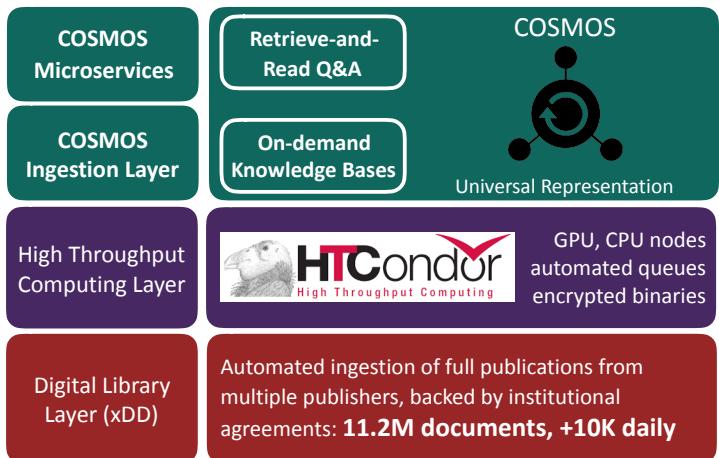
Discriminative ML Model

Input Multimodal-feature Generator

Input Probabilistic Labels

Output A learned ML model for on-demand, context-aware Knowledge Base Construction

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



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)