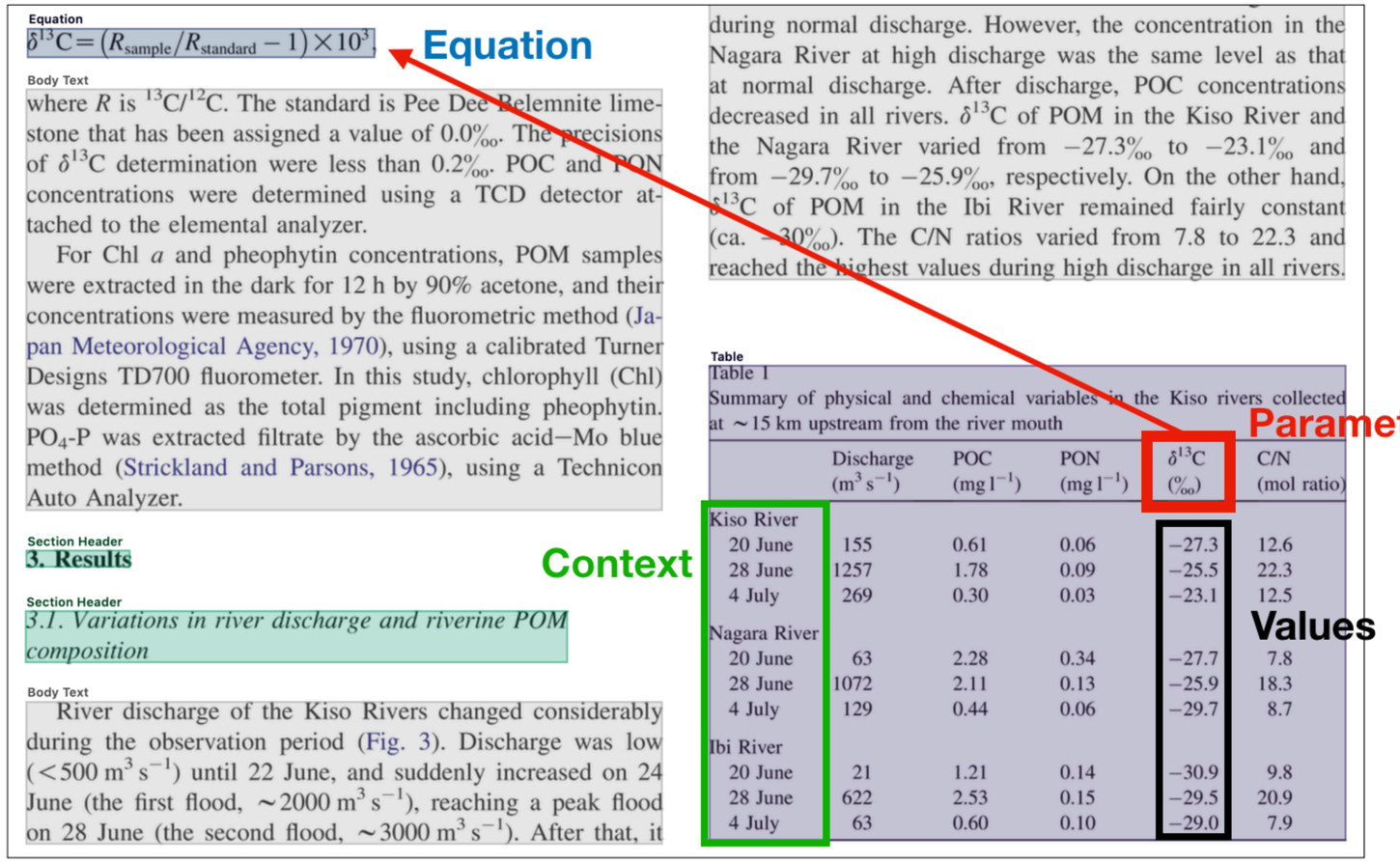
Extracting semantic knowledge and observational data for scientific models

(as a service)

Theo Rekatsinas, Shanan Peters, Miron Livny

Semantic knowledge extraction

- **Goal:** create repositories with ***machine-readable data*** about scientific models across different scientific domains
- **Semantic knowledge:** concepts, facts, relationships, entities, and observational data related to scientific phenomena and models
- **Our focus:** identify, extract, and collect semantic knowledge for scientific models from the published literature
- Foundational for model understanding, evaluating, and enhancing scientific models



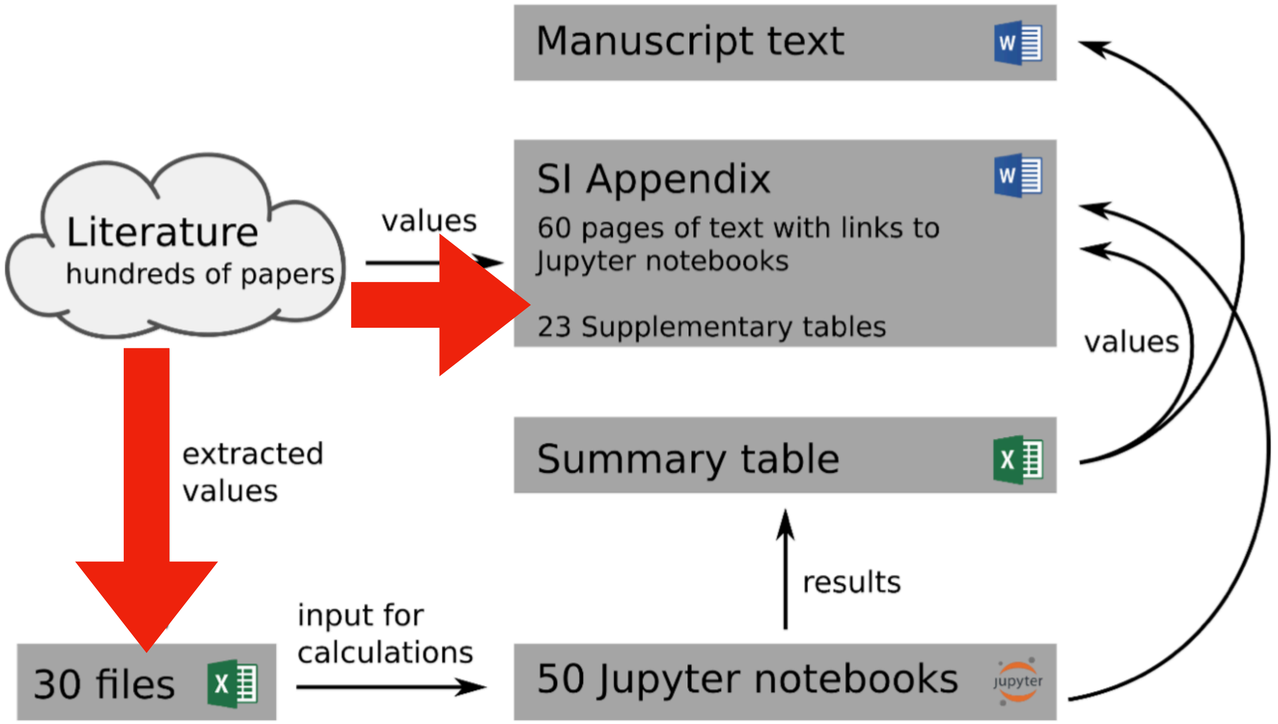
Challenge: reasoning about semantic context

Example: Understanding the biomass distribution on Earth

PNAS Proceedings of the National Academy of Sciences of the United States of America

The biomass distribution on Earth

Yinon M. Bar-On[a], Rob Phillips[b,c], and Ron Milo[a,1]



Analyzing the biomass on Earth

- The distribution and amount of biomass on Earth is fundamental to many models of carbon cycling and biological response to environmental changes
- Example: 2018 study depended on extracting published measurements of biomass and their context/classification
- Getting more such measurements is critical to building better models of biomass distribution

Automating data collection with xDD and COSMOS

Example use case

- We used xDD to find all document for a list of arthropod taxonomic names; **Result: 8,231 documents**
- Used COSMOS to extract page-level entities and tables; **Result: 30,701 table objects recognized**
- Used COSMOS to extract the content of tables relevant to the context of our analysis; **Result: 10,157 tables extracted**

Table

Independent variables	Insect variables: No. chewer species	Specialist ratio	Abundance	Biomass
1) Taxonomic isolation	−0.268	−0.346	−0.183	−0.174
2) Age	0.311	0.317	0.369	−0.198
3) Geographical distribution	0.364	−0.023	0.267	−0.041
4) Altitudinal range	0.697	0.440	0.463	0.531
5) Leaf flush	0.556	0.555	0.283	0.169
6) Young leaves	0.891	0.704	0.678	0.575
7) Leaf expansion	−0.098	−0.355	−0.183	0.230
8) Tree height	0.250	0.121	−0.035	−0.526
9) Leaf mass	0.088	−0.092	−0.251	0.106
10) Abundance	0.524	0.471	0.475	0.745
11) Leaf water	−0.477	−0.290	−0.135	−0.269
12) Leaf nitrogen	−0.279	−0.102	0.112	−0.076
13) Leaf palatability	0.225	−0.003	−0.093	−0.348
14) Log of ant abundance	0.588	0.809	0.645	0.315
15) Sampling effort	−0.211	−0.264	−0.165	0.325

and Basset Yves, *Local Communities of Arboreal Herbivores in Papua New Guinea: Predictors of Insect Variables*, **Ecology**, **77**(6), 1996, DOI: 10.2307/2265794

Preview table | Download pickled pandas dataframe | Download OCRed text | See full stored object

COSMOS can accelerate data collection while being robust and domain agnostic

Ongoing exploration: continental groundwater storage

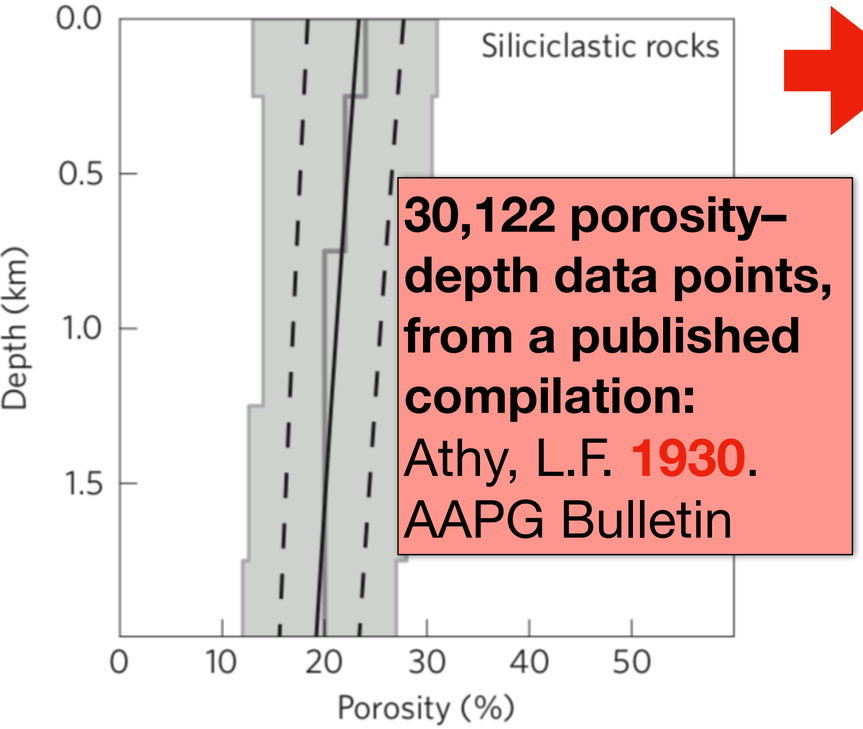
nature geoscience

Gleeson et al.

ARTICLES

PUBLISHED ONLINE: 16 NOVEMBER 2015 | DOI: 10.1038/NGEO2590

The global volume and distribution of modern groundwater



30,122 porosity–depth data points, from a published compilation:
Athy, L.F. **1930**. AAPG Bulletin

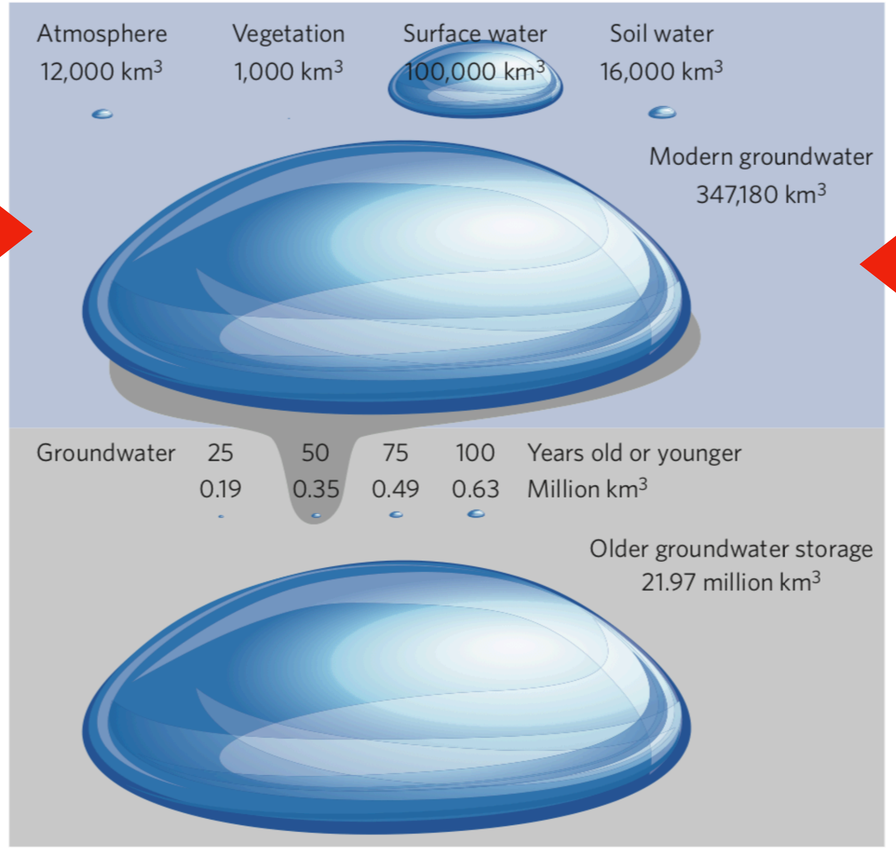
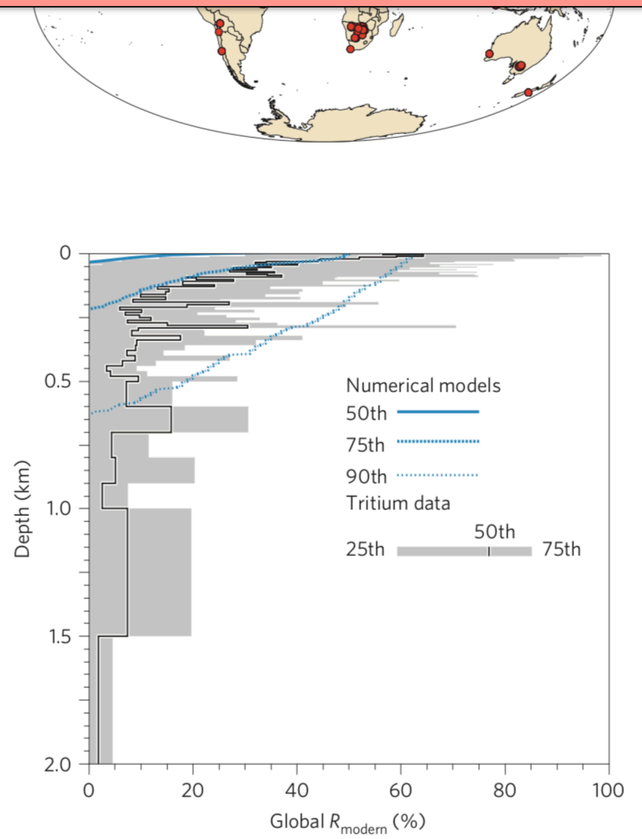


Figure 5 | The different volumes of water stored in the global water cycle.

3,769 measurements of ^{3}H in groundwater compiled from 160 published data sets.



Ongoing exploration: continental groundwater storage

COSMOS INPUT: NGDS vocabulary for specific measurement types and example context, xDD document acquisition/processing pipeline

NGDS
National Geothermal Data System

Automate location and aggregation of contextualized sample-based measurements and models relevant to groundwater, hydrothermal E

Improve groundwater models and inventories, national hydrothermal assessment

Table 1. Porosity is the helium porosity, permeabilit
***R* is measured resistivity in ohm·m, *F* is the normali**
(*n*) of 2, and *Rw* = 0.17 ohm.m, Error (%) is the er

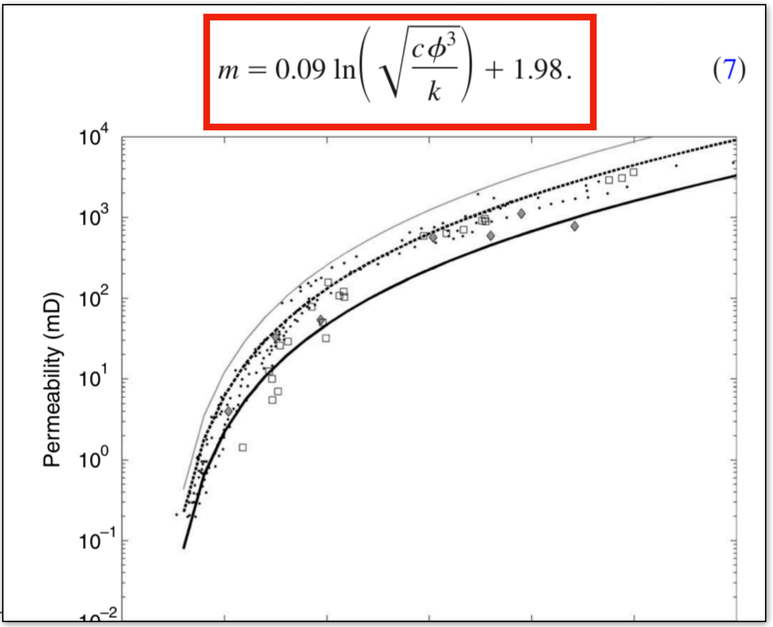
Sample	Porosity	Perm (mD)
A11	0.07	10
A16	0.07	6
A33	0.07	12
A82	0.08	7
A87	0.10	50
A89	0.08	26
A117	0.11	103
B31	0.11	107
B86	0.09	78
B101	0.11	121
B102	0.10	157
B108	0.08	29
F510	0.15	592
GT3	0.17	704
GW18	0.16	637
GW19	0.18	912
GW23	0.18	965
GW28	0.18	896
H27	0.25	3630
H42	0.24	2894
H74	0.24	3079
F410	0.06	1
F570	0.10	32

ABSTRACT

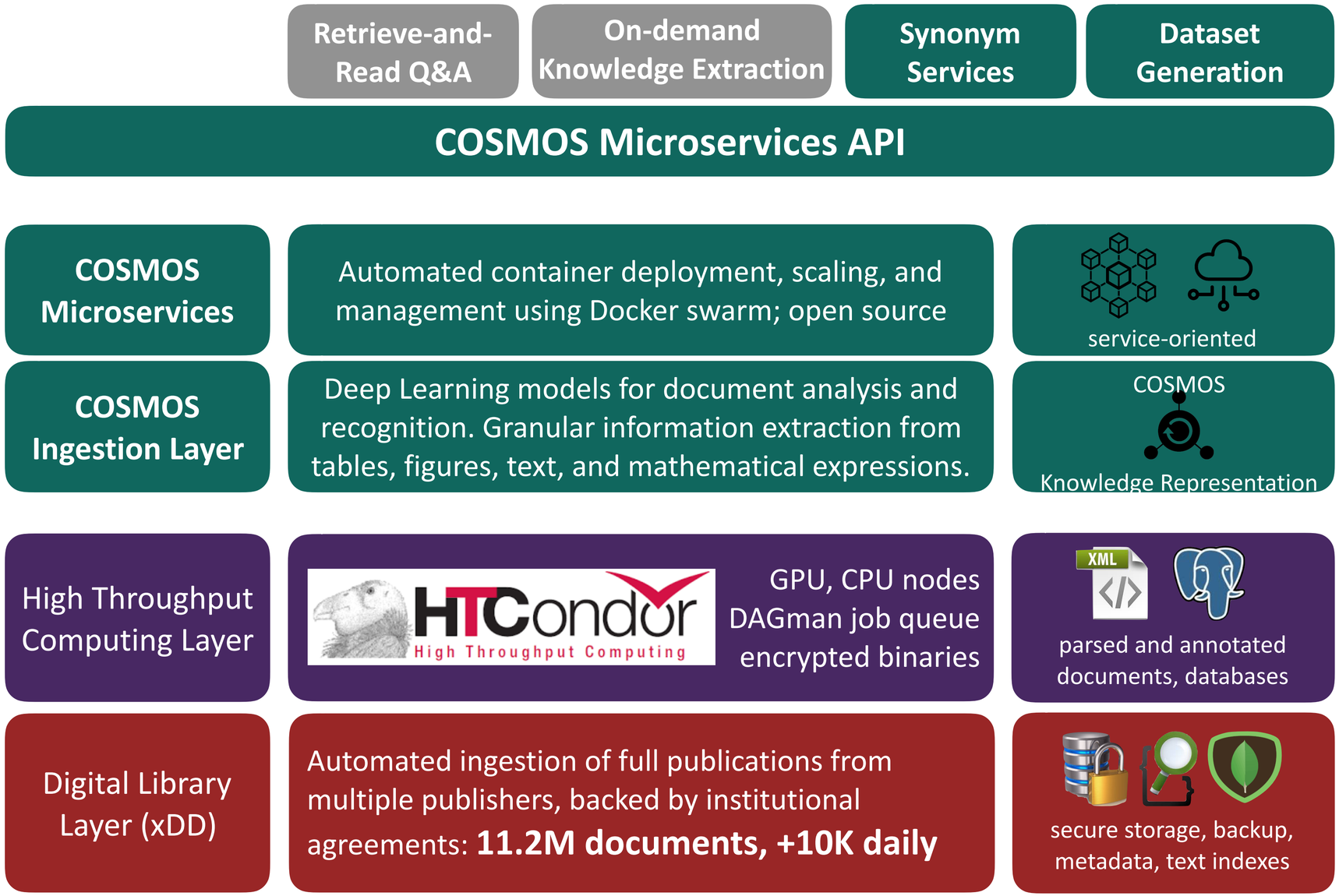
The relations among the resistivity, elastic-wave velocity, porosity, and permeability in Fontainebleau sandstone samples from the Ile de France region, around Paris, France were experimentally revisited. These samples followed a permeability-porosity relation given by Kozeny-Carman's equation. For the resistivity measurements, the samples were partially saturated with brine. Archie's equation was used to estimate resistivity at 100% water saturation, assuming a saturation exponent, $n = 2$. Using

0.65	88.29	216.49	9
0.76	19.34	65.67	6

$$m = 0.09 \ln\left(\sqrt{\frac{c\phi^3}{k}}\right) + 1.98. \quad (7)$$



Current status of COSMOS



Completed components of COSMOS

- a domain-agnostic ingest engine for unstructured data
- fine-grained entity and data tagging
- services for training natural language comprehension models over user-defined verticals
- multi-modal dataset generation
- analysis and tagging of model code (limited to Fortran)

On-going extensions and research

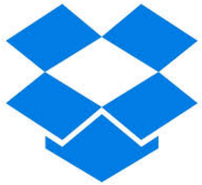
- On-demand knowledge and data extraction
- Retrieve-and-read Q&A
- Scaling of global-context embeddings
- Powering scientific applications (macroscopic analysis of phenomena)

: currently working on it

: alpha version completed

Our vision: democratize knowledge extraction

Agent



Applications

Services

Infrastructure

- An intelligent virtual assistant that can consider and reason about semantics across domains when linking/synthesizing artifacts from publications
- APIs that enable users to build custom applications over that assistant
- Services that bring fine-grained knowledge extraction within reach of every scientist—without requiring programming expertise (interactions based on natural language and point-and-click interfaces)
- Compute infrastructure and data to support diverse domains and disciplines

Knowledge extraction solutions are vertical-specific

Current view on Knowledge Extraction:

- **Closed-world extraction:** align to existing entities and attributes, e.g., (ID_Obama, place_of_birth, ID_USA)
- **Closed Information Extraction (IE):** align to existing attributes but extract new entities, e.g., (Theo, place_of_birth, Greece)
- **Open IE:** not limited by existing attributes (or entities)

: production ready

: limited successes for popular verticals

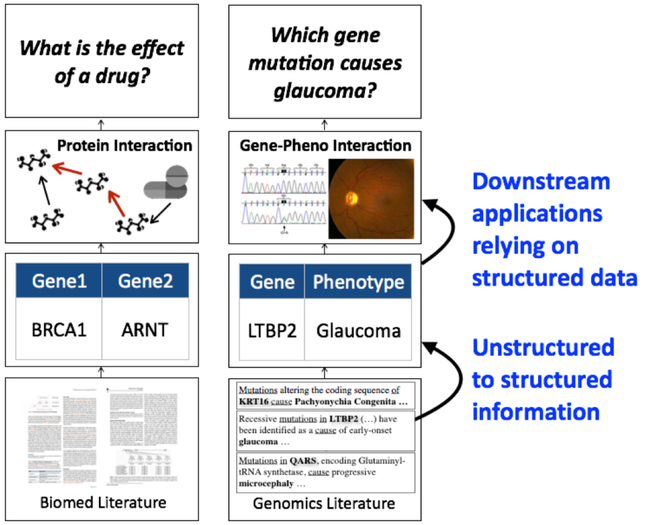
Example systems (from prior DARPA programs):

Memex Human Trafficking

DeepDive

- URL where the advertisement was found
- Phone number of the person in the advertisement
- Name of the person in the advertisement
- Location where the person offers services
- Rates for services offered

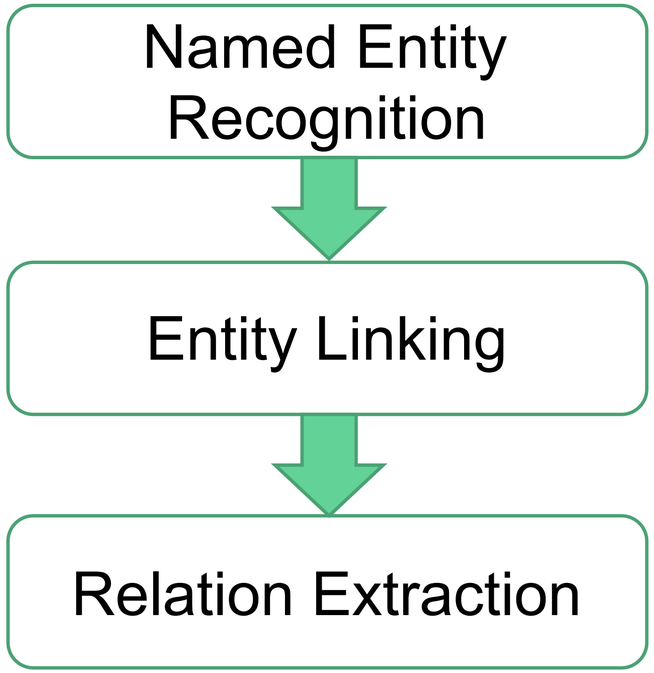
Simplifying Complexity in Scientific Discovery (SIMPLEX)



The bottleneck in Knowledge Extraction is human supervision

State-of-the-art Knowledge Extraction requires supervised machine learning

- **NER:** identify text-spans that are of interest; requires ***matchers or ML models trained*** on corpora of the target vertical
- **EL:** align text-spans to existing canonical entities; requires ***supervised ML models for matching***
- **RE:** align text-spans and identified entities to existing relations; requires ***specialized ML models for a domain schema*** (focus on specific relations in target domain)

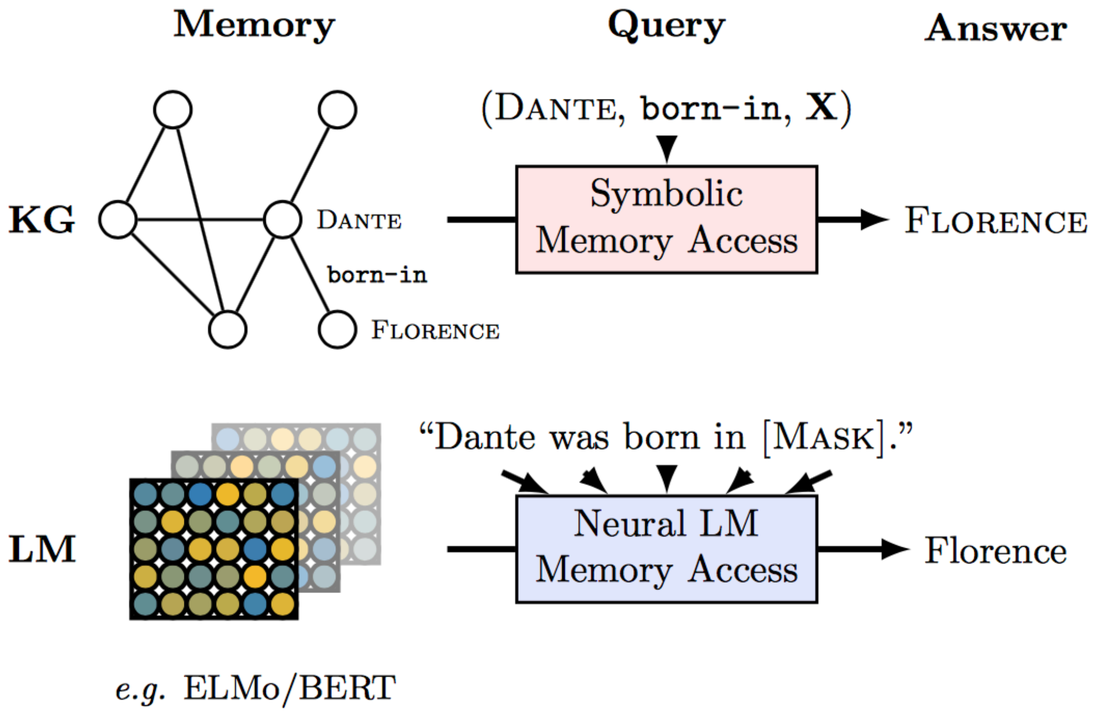


Steps in SOTA Knowledge Extraction

The case of unsupervised knowledge extraction with modern text-comprehension models

Language Models as Knowledge Bases?

Fabio Petroni[1] **Tim Rocktäschel**[1,2] **Patrick Lewis**[1,2] **Anton Bakhtin**[1]
Yuxiang Wu[1,2] **Alexander H. Miller**[1] **Sebastian Riedel**[1,2]
[1]Facebook AI Research
[2]University College London
{fabiopetroni, rockt, plewis, yolo, yuxiangwu, ahm, sriedel}@fb.com

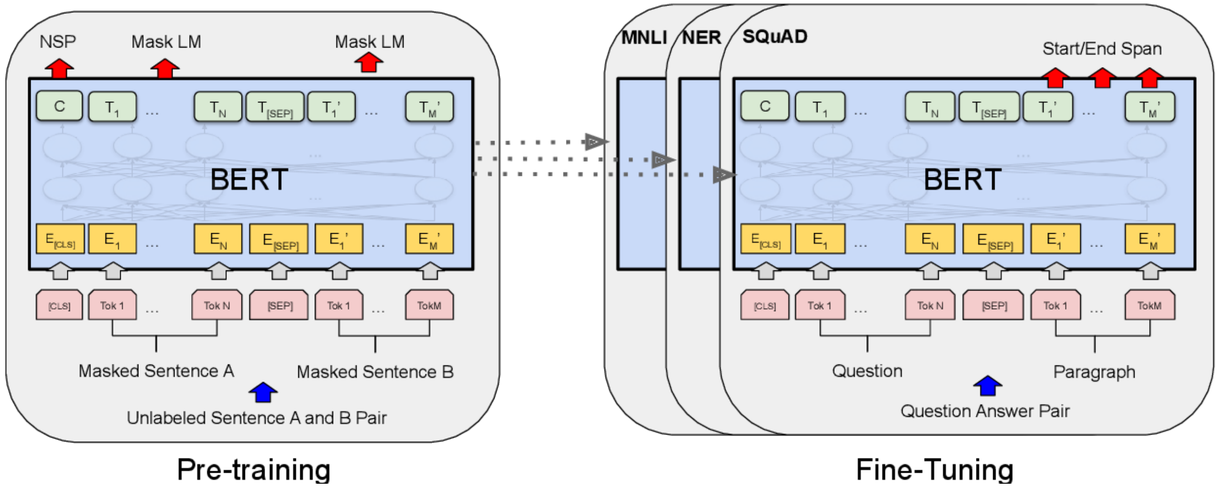


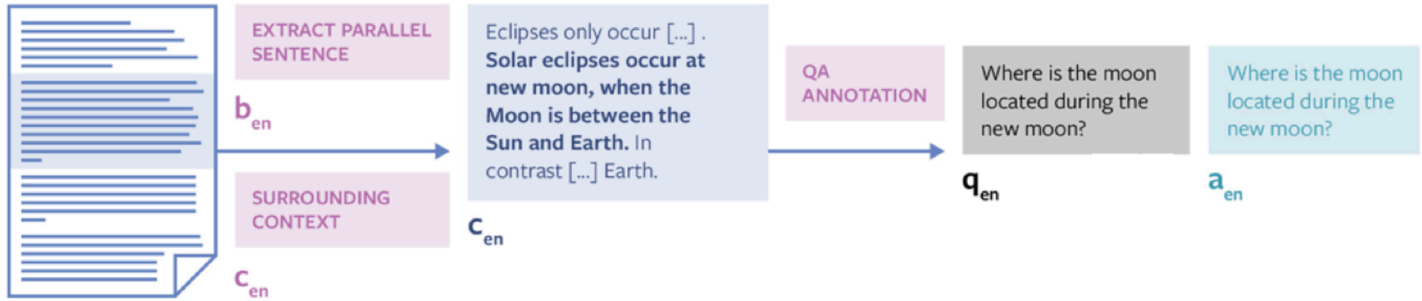
Contextual AI models for text comprehension
(unsupervised models) capture structured knowledge

The case of unsupervised relation extraction with modern text-comprehension models

Limitations:

- Models overfit on heavy-hitters (tend to predict most popular entities); **Cannot be applied to long-tail domains**
- Models need to be fine-tuned to downstream tasks; **Fine-tuning requires training a question-answering head for each domain**
- No end-to-end solution; **QA requires that one has already identified the context that contains a relevant extraction**





Re-Flex: Unsupervised relation extraction with contextual models

Unsupervised relation extraction (Overview)

- ***Step 1:*** The user provides a dictionary of tokens of interest (subjects); Ex.: set of rock formations, set of people, set of genes
- ***Step 2:*** The user describes a relation in Natural Language that these tokens participate in; Ex.: located in, plays for, is associated with disease
- ***Step 3:*** We complete a knowledge base of (subject, relation, object) tuples where the subjects and relations are specified as above. We only use unsupervised, contextual text comprehension models

Input: Components, parts of documents

Henrik Lundqvist grew up with his identical twin brother Joel in Åre, Jämtland, an area where alpine skiing is the v thoug Joel ice ho more p sports.

Austin Watson was born January 13, 1992, in Ann Arbor, Michigan, where he grew up with his er, ıry

NBA player	Position	Height (inches)	Weight (pounds)
Carmelo Anthony	Forward	80	230
Aron Baynes	Center	82	260
Matt Bonner	Forward/Center	82	240
Patrick Christopher	Guard	77	209
Norris Cole	Guard	74	170
Glen Davis	Forward/Center	81	289
Gorgui Dieng	Center	83	245
Cleanthony Early	Forward	80	219
Jrue Holiday	Guard	76	205
Jonas Jerebko	Forward	82	231
Andrew Nicholson	Forward	81	250
David West	Forward	81	240

Input: Text completion queries by combining subjects with relation tokens

Travis Hamonic plays for [MASK]
Morten Madsen plays for [MASK]
Henrik Lundqvist plays for [MASK]

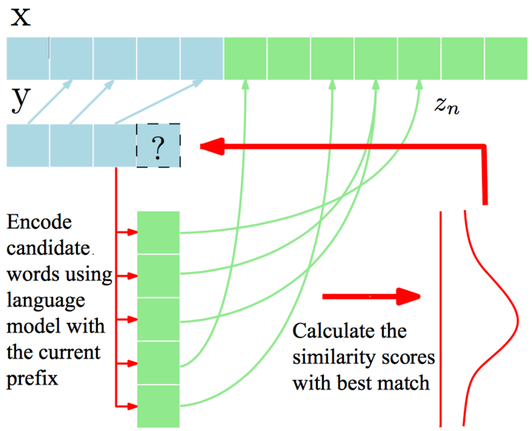
Re-Flex
Step 1: Align queries to relevant contexts
Step 2: Find most probable anchor answer in context (inference fine-tuning)
Step 3: Answer decoding

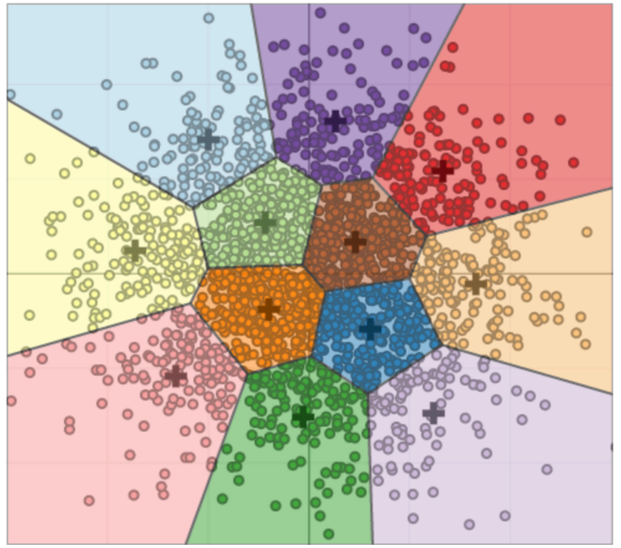
Output: Completed knowledge base

Travis Hamonic plays for **Calgary flames**
Morten Madsen plays for **Modo Hockey**
Henrik Lundqvist plays for **New York Rangers**

Inference fine-tuning by contextual matching

Condition on context tokens and re-weight their probabilities by computing their contextual similarity to probability distribution over the raw prediction of the model





We use Voronoi clustering to re-assign the probability mass (center correspond to tokens in the relevant context)

We perform inference-level fine-tuning by matching context to raw model predictions

Re-Flex: Unsupervised relation extraction with contextual models

RE-Flex achieves comparable or better performance than **supervised models**

RE-Flex yields **~40 F1 points improvement** against SOTA unsupervised methods

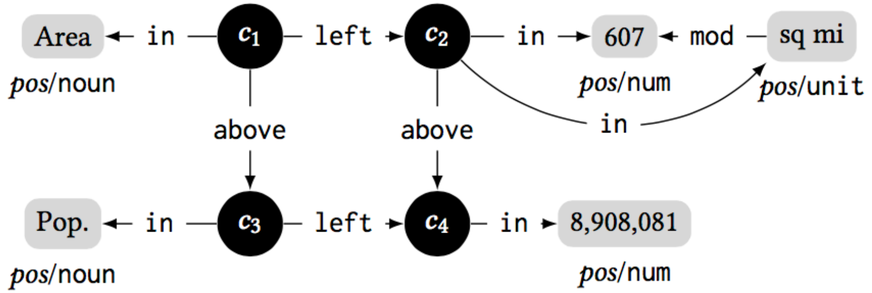
Dataset	Model	EM	F1
Google-RE	RE-Flex	91.1	91.1
	(Lewis et al., 2019)	54.3	68.2
	(Dhingra et al., 2018)	25.9	34.6
	BERT-Large+Squad2.0	53.6	77.0
	BERT-Large+ZSRE	83.2	88.2
T-REx	RE-Flex	67.1	67.1
	(Lewis et al., 2019)	13.6	18.5
	(Dhingra et al., 2018)	20.8	27.7
	BERT-Large+Squad2.0	41.2	51.0
	BERT-Large+ZSRE	46.8	52.2
ZSRE	RE-Flex	37.8	41.4
	(Lewis et al., 2019)	9.6	14.1
	(Dhingra et al., 2018)	12.3	17.4
	BERT-Large+Squad2.0	39.4	48.2
	BERT-Large+ZSRE	54.7	56.7

Ensemble Synthesis for Robust Data Extraction

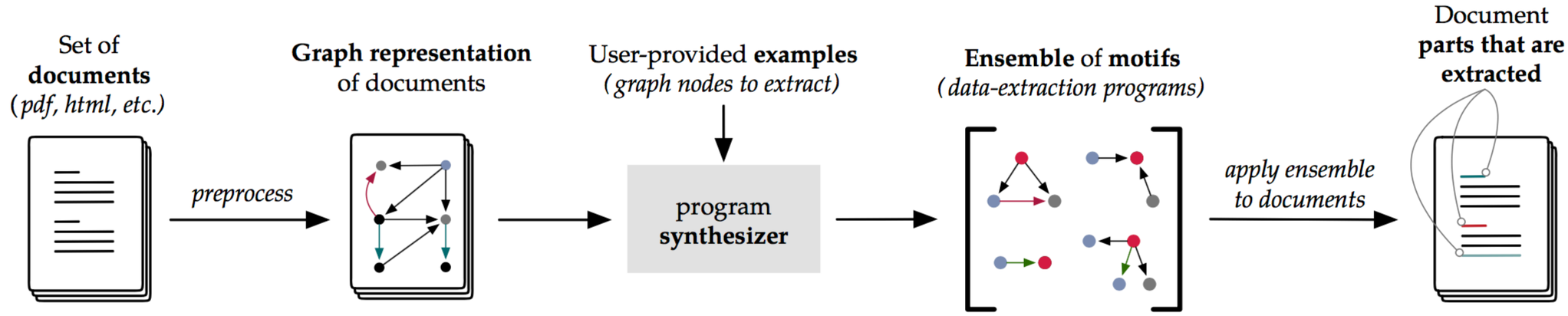
Interactive extraction over COSMOS generated datasets

- ***Step 1:*** Collect related document elements (paragraphs, tables, figures)
- ***Step 2:*** The user annotates few spans of interest
- ***Step 3:*** Exploit COSMOS's graph representation and apply program synthesis to generate graph-based programs that extract relevant data from the collection of Step 1
- ***Step 4:*** Iterate in an active learning loop to improve quality

Area	607 sq mi
Population	8,908,081

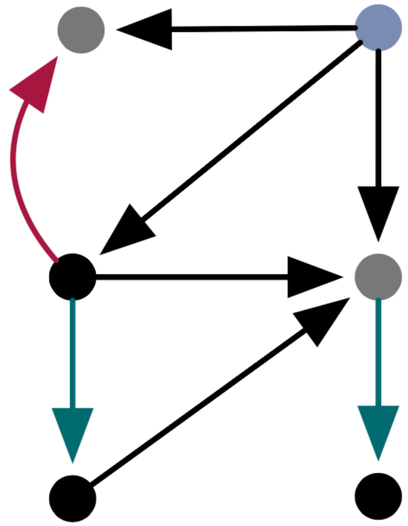


Example Graph Representation of a Table



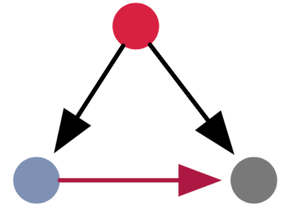
Program synthesis with Bayesian Ensembles

Input: Document graphs
(*unlabeled data*)
limited annotations by
expert (less than 10)

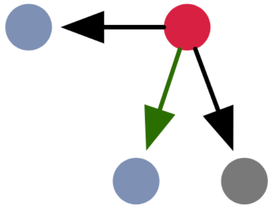


Program Synthesizer

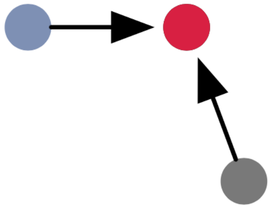
A new program synthesis approach that generate multiple programs that satisfy the input specification



Program 1

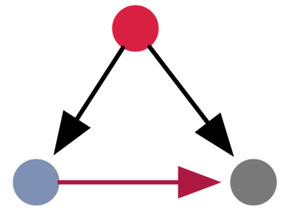


Program 2

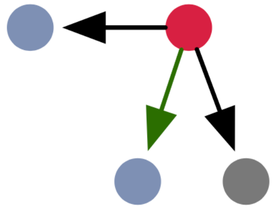


Program 3

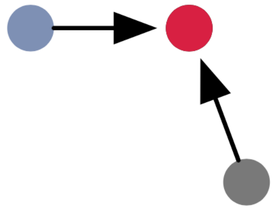
Program synthesis with Bayesian Ensembles



Program 1

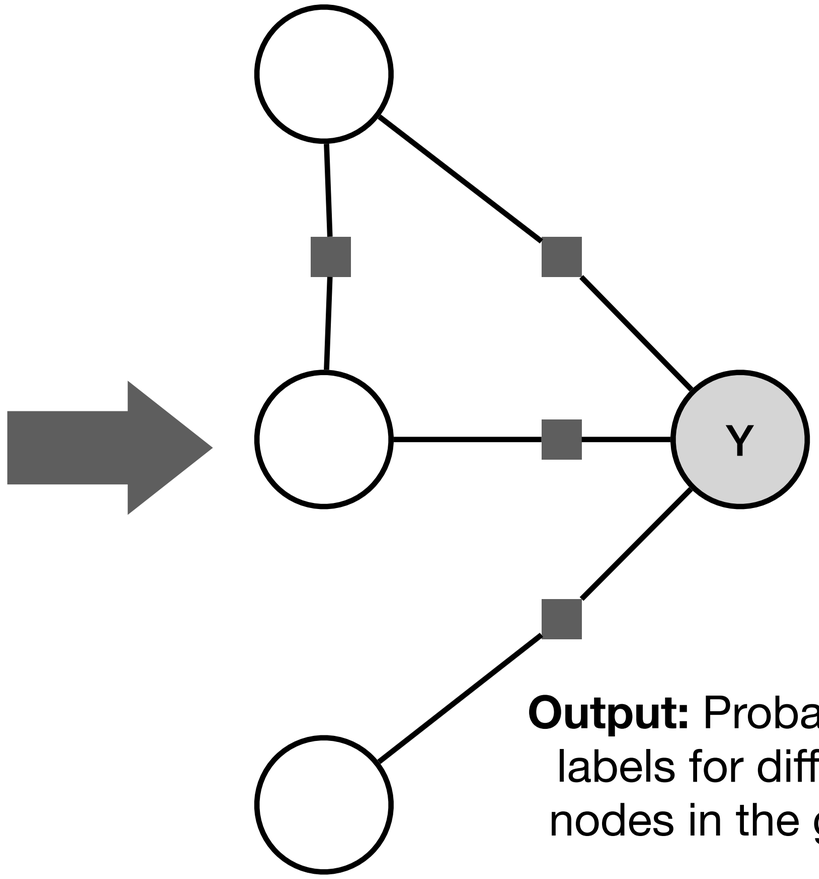


Program 2



Program 3

Generative Model



Output: Probabilistic labels for different nodes in the graph

$$Pr[y = 1 \mid \mathcal{M}(v)] \propto \exp\left(\sum_{M \in \mathcal{M}} \ln(\mathcal{A}(M)) \cdot \mathbb{1}[M(v) = 1]\right)$$

Active learning loop

initialize $FPR_0(M)$ for each motif $M \in \mathcal{M}$
initialize learning rate η

for i in $0, \ldots, l$ (l is an upper bound)
  update w_M for each motif M using $FPR_i(M)$
  query user for label y of v with the **highest entropy** (see text)

  if $y = 1$, then **for** $M \in \mathcal{M}$ s.t. $M(v) = 1$ (*true positives*)
    $FPR_{i+1} \leftarrow FPR_i(M) \cdot \exp(-\eta)$

  if $y = 0$, then **for** $M \in \mathcal{M}$ s.t. $M(v) = 1$ (*false positives*)
    $FPR_{i+1} \leftarrow FPR_i(M) \cdot \exp(\eta)$

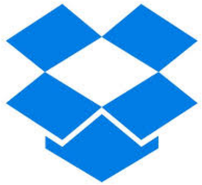
Active learning can help improve the accuracy estimate of each program

Dataset name	F1-score	# Ex.
Political (Cell)	1.0	1
Chemical (Span)	0.703	5
President (Cell)	0.66	1
Transistor (Cell)	0.65	1
Transistor (Span)	0.5	3

Competing or SOTA performance with less than five examples

Our vision: democratize knowledge extraction

Agent



Applications

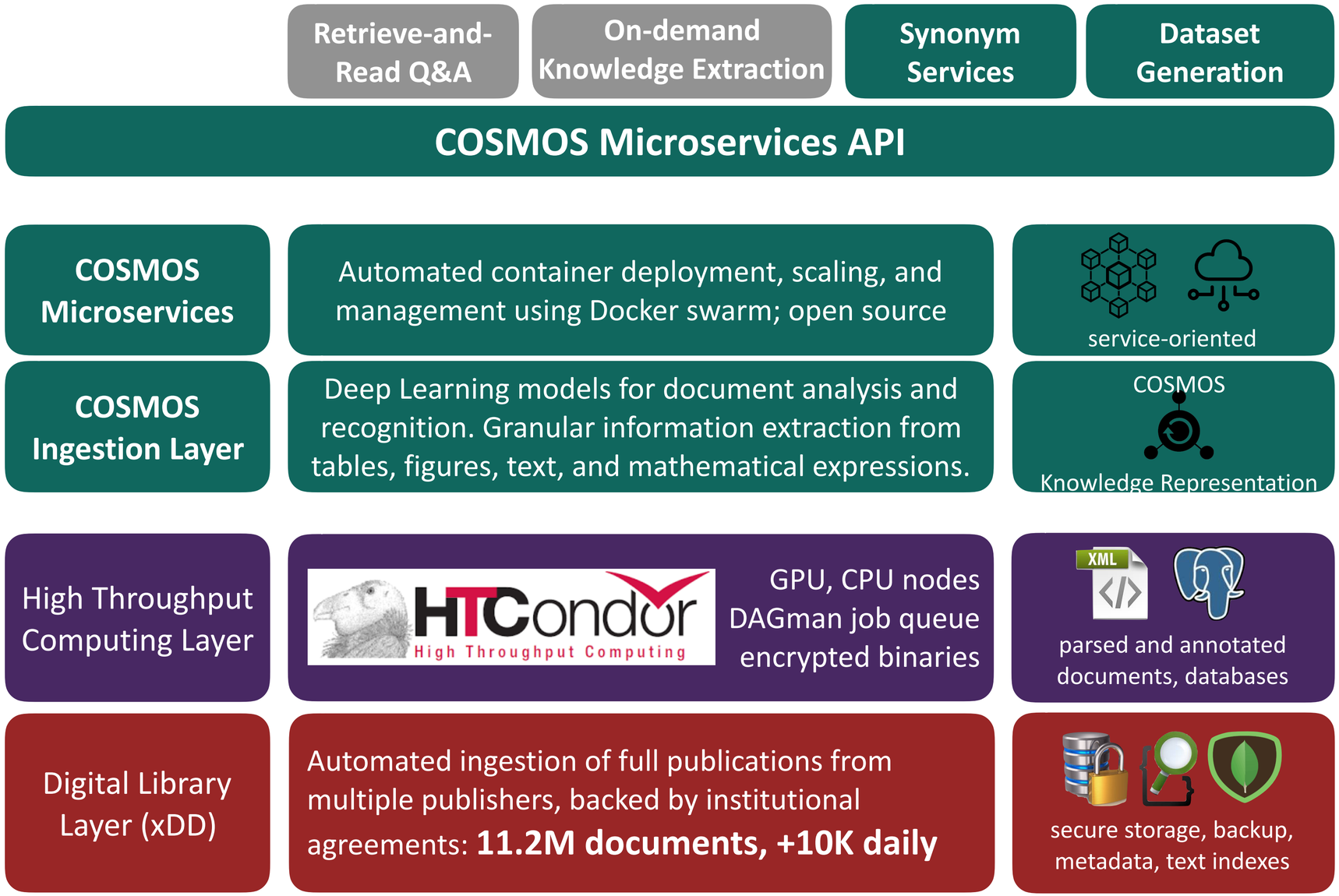
Services

Infrastructure

- An intelligent virtual assistant that can consider and reason about semantics across domains when linking/synthesizing artifacts from publications
- APIs that enable users to build custom applications over that assistant
- Services that bring fine-grained knowledge extraction within reach of every scientist—without requiring programming expertise (interactions based on natural language and point-and-click interfaces)
- Compute infrastructure and data to support diverse domains and disciplines

Supplemental Slides

Current status of COSMOS



Completed components of COSMOS

- a domain-agnostic ingest engine for unstructured data
- fine-grained entity and data tagging
- services for training natural language comprehension models over user-defined verticals
- multi-modal dataset generation
- analysis and tagging of model code (limited to Fortran)

On-going extensions and research

- On-demand knowledge and data extraction
- Retrieve-and-read Q&A
- Scaling of global-context embeddings
- Powering scientific applications (macroscopic analysis of phenomena)

: currently working on it

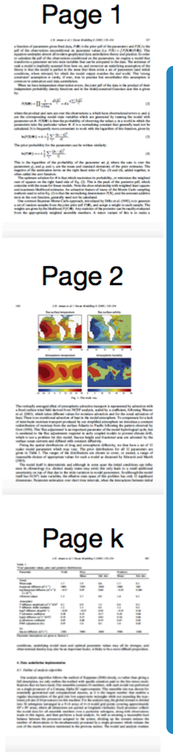
: alpha version completed

From PDFs to XML representations

COSMOS Ingestion Layer

Mutli-page raw PDF

Document Preprocessing Tool

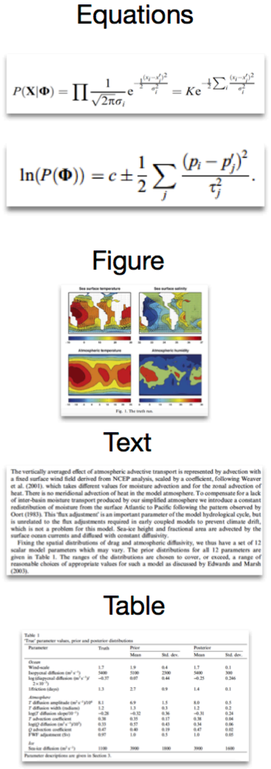


Images corresponding to PDF pages

ML Model

ML Model

ML Model



Visual Extractions of Elements of Interest

OCR Engine

OCR Engine

OCR Engine

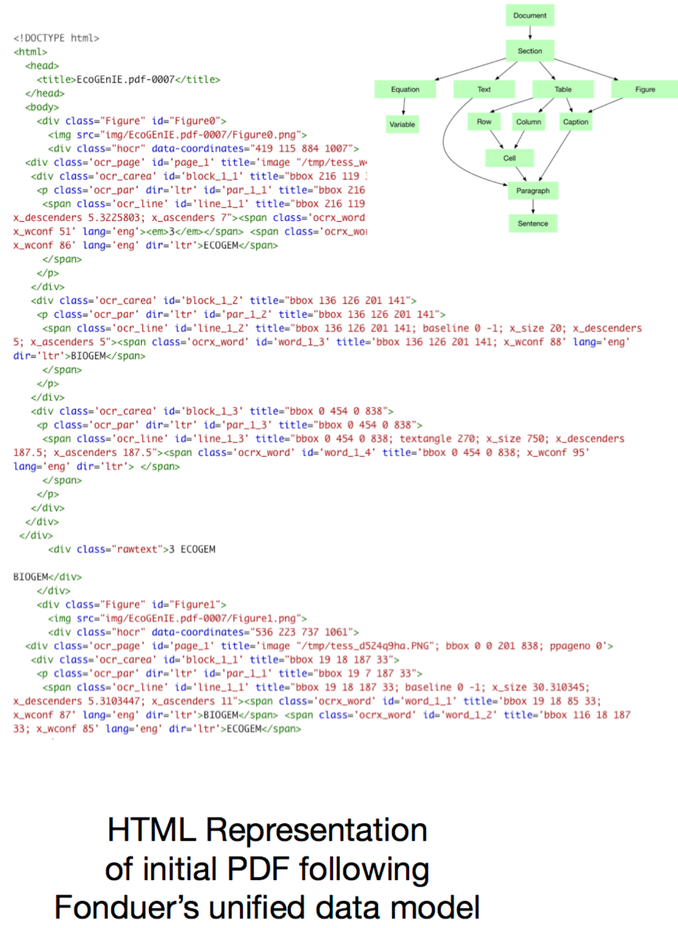
OCR Engine

OCR Engine

OCR Enrichment

Element Aggregator

Representation Compiler



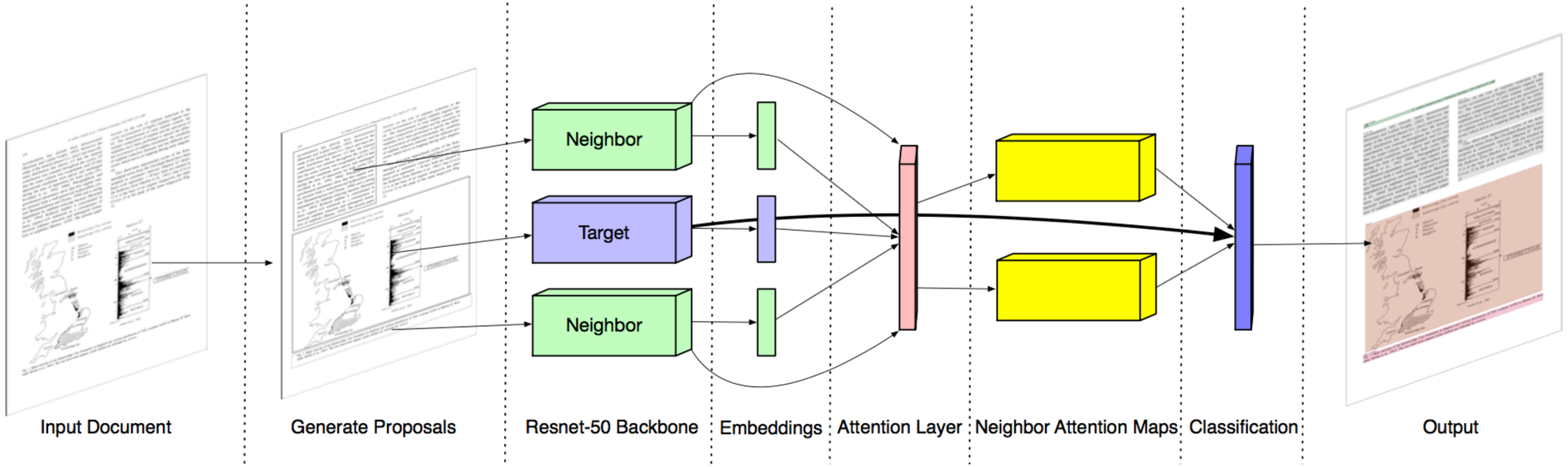
```
<!DOCTYPE html>
<html>
  <head>
    <title>EcoGEnIE.pdf-0007</title>
  </head>
  <body>
    <div class="Figure" id="Figure0">
      <img src='img/EcoGEnIE.pdf-0007/Figure0.png'>
      <div class="hocr" data-coordinates="419 115 884 1007">
  <div class='ocr_page' id='page_1' title='image "/tmp/tess_w
   <div class='ocr_carea' id='block_1_1' title="bbox 216 119
    <p class='ocr_par' dir='ltr' id='par_1_1' title="bbox 216
     <span class='ocr_line' id='line_1_1' title="bbox 216 119
x_descenders 5.3225803; x_ascenders 7"><span class='ocrx_word
x_wconf 51' lang='eng'><em>3</em></span> <span class='ocrx_wo
x_wconf 86' lang='eng' dir='ltr'>ECOGEM</span>
      </span>
    </p>
   </div>
   <div class='ocr_carea' id='block_1_2' title="bbox 136 126 201 141">
    <p class='ocr_par' dir='ltr' id='par_1_2' title="bbox 136 126 201 141">
     <span class='ocr_line' id='line_1_2' title="bbox 136 126 201 141; baseline 0 -1; x_size 20; x_descenders
5; x_ascenders 5"><span class='ocrx_word' id='word_1_3' title='bbox 136 126 201 141; x_wconf 88' lang='eng'
dir='ltr'>BIOGEM</span>
      </span>
    </p>
   </div>
   <div class='ocr_carea' id='block_1_3' title="bbox 0 454 0 838">
    <p class='ocr_par' dir='ltr' id='par_1_3' title="bbox 0 454 0 838">
     <span class='ocr_line' id='line_1_3' title="bbox 0 454 0 838; textangle 270; x_size 750; x_descenders
187.5; x_ascenders 187.5"><span class='ocrx_word' id='word_1_4' title='bbox 0 454 0 838; x_wconf 95'
lang='eng' dir='ltr'> </span>
      </span>
    </p>
   </div>
  </div>
 </div>
      <div class="rawtext">3 ECOGEM

BIOGEM</div>
    </div>
    <div class="Figure" id="Figure1">
      <img src='img/EcoGEnIE.pdf-0007/Figure1.png'>
      <div class="hocr" data-coordinates="536 223 737 1061">
  <div class='ocr_page' id='page_1' title='image "/tmp/tess_d524q9ha.PNG"; bbox 0 0 201 838; ppageno 0'>
   <div class='ocr_carea' id='block_1_1' title="bbox 19 18 187 33">
    <p class='ocr_par' dir='ltr' id='par_1_1' title="bbox 19 7 187 33">
     <span class='ocr_line' id='line_1_1' title="bbox 19 18 187 33; baseline 0 -1; x_size 30.310345;
x_descenders 5.3103447; x_ascenders 11"><span class='ocrx_word' id='word_1_1' title='bbox 19 18 85 33;
x_wconf 87' lang='eng' dir='ltr'>BIOGEM</span> <span class='ocrx_word' id='word_1_2' title='bbox 116 18 187
33; x_wconf 85' lang='eng' dir='ltr'>ECOGEM</span>
```

HTML Representation of initial PDF following Fonduer's unified data model

COSMOS' document segmentation model offers robustness against format heterogeneity across publications

The COSMOS Attentive RCNN Model

COSMOS Ingestion Layer



cosmos.torch_model module

Torch model directory

SHOW SOURCE »

Sub-modules

cosmos.torch_model.inference
Inference module of model

cosmos.torch_model.model
Model specification dir

cosmos.torch_model.rewrite_xml

cosmos.torch_model.tests

cosmos.torch_model.train
Train module

cosmos.torch_model.utils
Torch Model utils

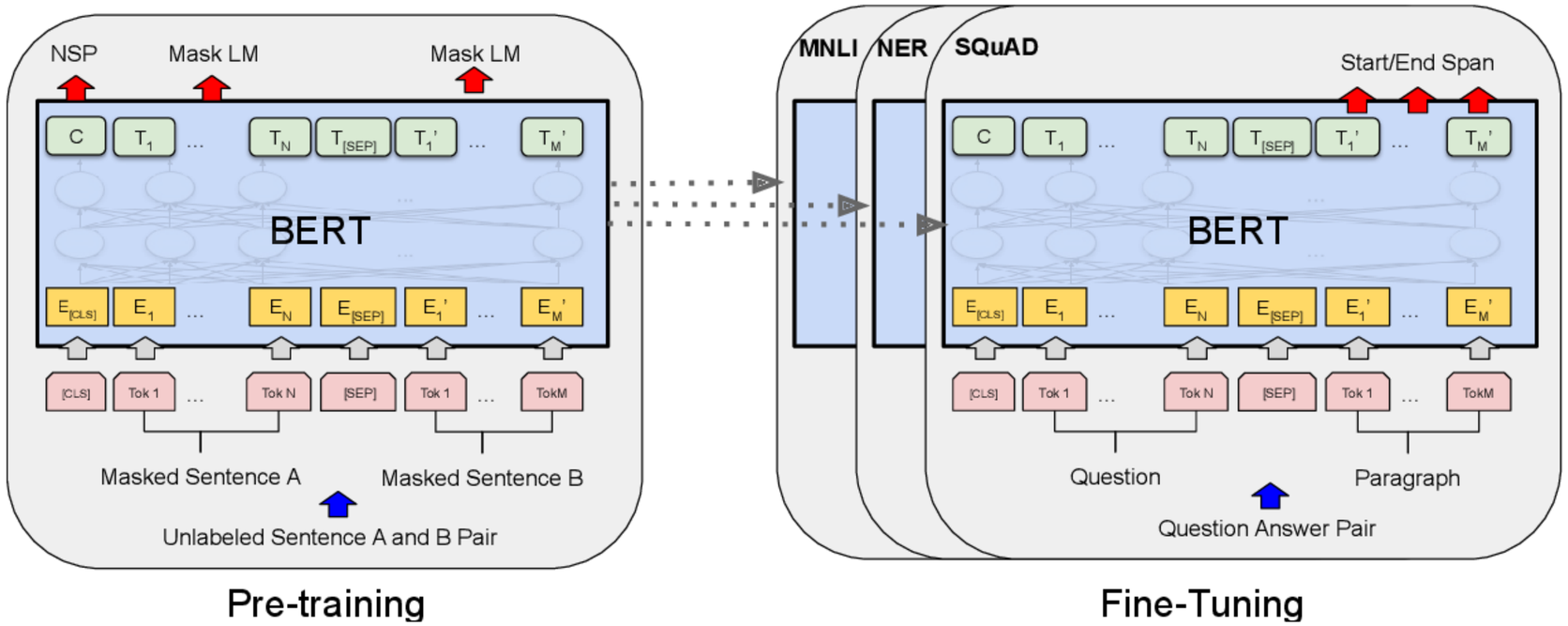
New distributed representation (in the visual space) for each element in the page.

AI2 Allen Institute for AI

In discussions with AI2 to deploy this model in Semantic Scholar.

Data representations in COSMOS

COSMOS
Knowledge Representation



Leaderboard

SQuAD2.0 tests the ability of a system to not only answer reading comprehension questions, but also abstain when presented with a question that cannot be answered based on the provided paragraph. How will your system compare to humans on this task?

Rank	Model	EM	F1
	Human Performance *Stanford University* (Rajpurkar & Jia et al. '18)	86.831	89.452
1 Mar 20, 2019	BERT + DAE + AoA (ensemble) *Joint Laboratory of HIT and iFLYTEK Research*	**87.147**	**89.474**
2 Mar 15, 2019	BERT + ConvLSTM + MTL + Verifier (ensemble) *Layer 6 AI*	86.730	89.286
3 Mar 05, 2019	BERT + N-Gram Masking + Synthetic Self-Training (ensemble) *Google AI Language* https://github.com/google-research/bert	86.673	89.147
4 Mar 16, 2019	BERT + DAE + AoA (single model) *Joint Laboratory of HIT and iFLYTEK Research*	85.884	88.621
5 Jan 15, 2019	BERT + MMFT + ADA (ensemble) *Microsoft Research Asia*	85.082	87.615

We adopt existing entity tagging based representations (e.g., fine-grained named entity tagging and indexing) and state-of-art context-aware representations based on deep learning models (transformer networks)

Use cases that COSMOS can contribute to

Use Case 1: continental groundwater storage

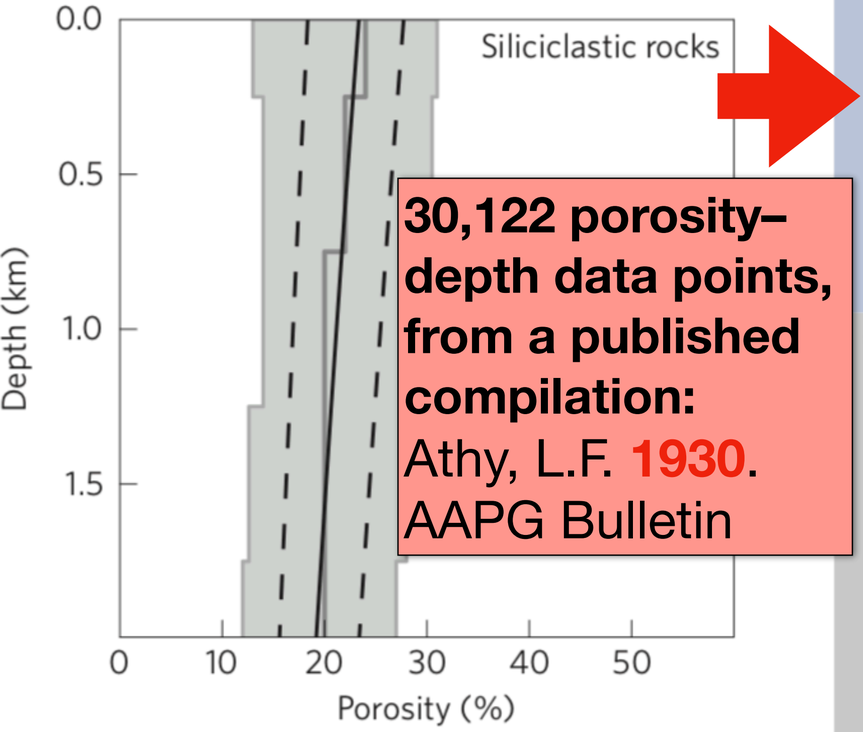
nature geoscience

Gleeson et al.

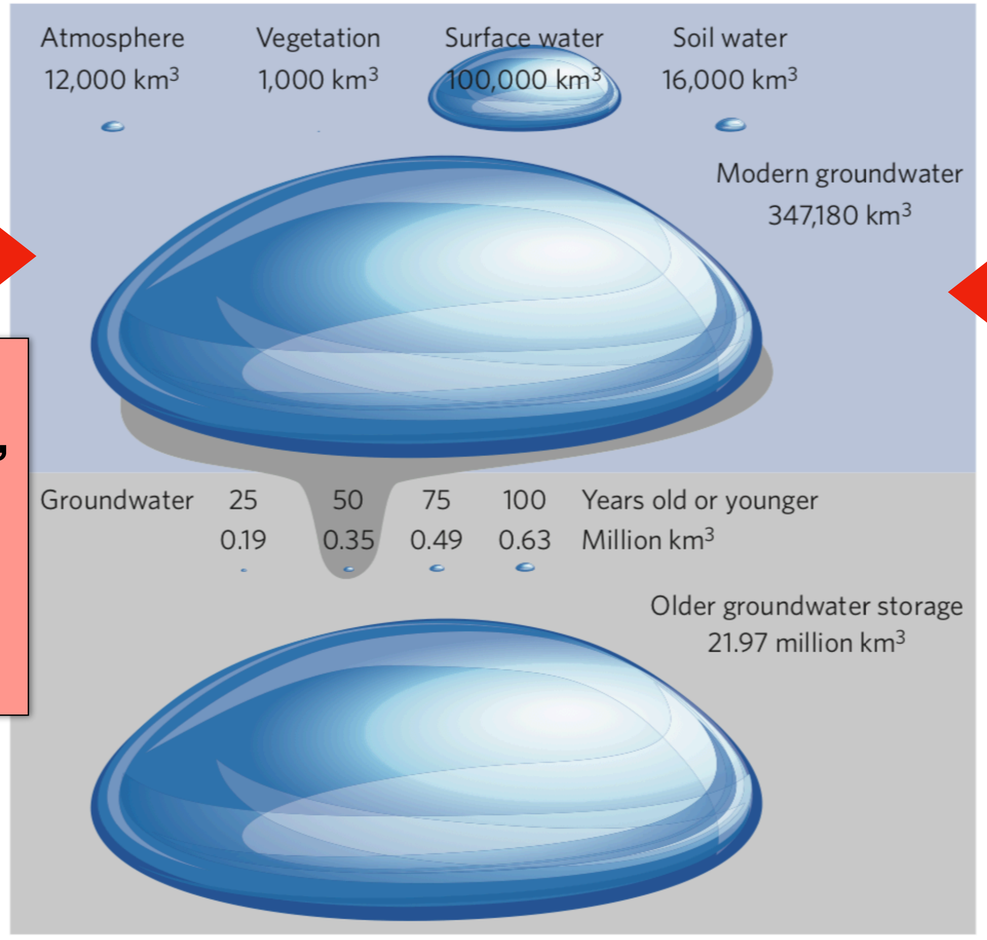
ARTICLES

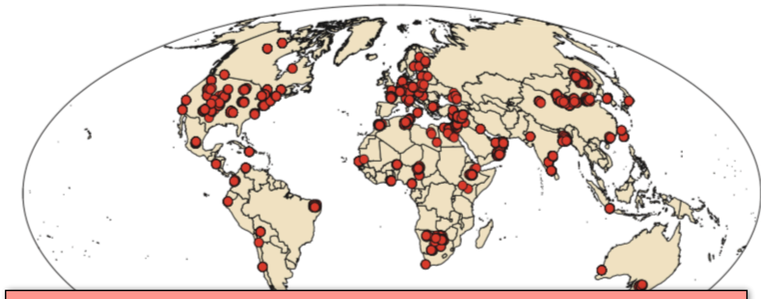
PUBLISHED ONLINE: 16 NOVEMBER 2015 | DOI: 10.1038/NGEO2590

The global volume and distribution of modern groundwater



30,122 porosity–depth data points, from a published compilation:
Athy, L.F. **1930**. AAPG Bulletin





3,769 measurements of ^{3}H in groundwater compiled from 160 published data sets.

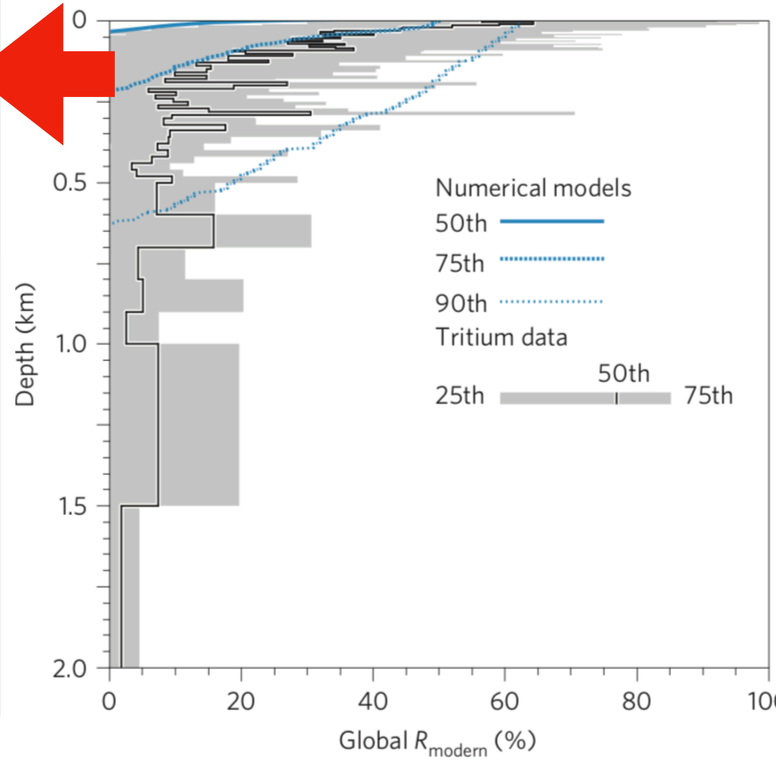


Figure 5 | The different volumes of water stored in the global water cycle.

Use Case 1: continental groundwater storage

COSMOS INPUT: NGDS vocabulary for specific measurement types and example context, xDD document acquisition/processing pipeline

NGDS
National Geothermal Data System

Automate location and aggregation of contextualized sample-based measurements and models relevant to groundwater, hydrothermal E

Improve groundwater models and inventories, national hydrothermal assessment

Table 1. Porosity is the helium porosity, permeabilit
***R* is measured resistivity in ohm·m, *F* is the normali**
(*n*) of 2, and *Rw* = 0.17 ohm.m, Error (%) is the er

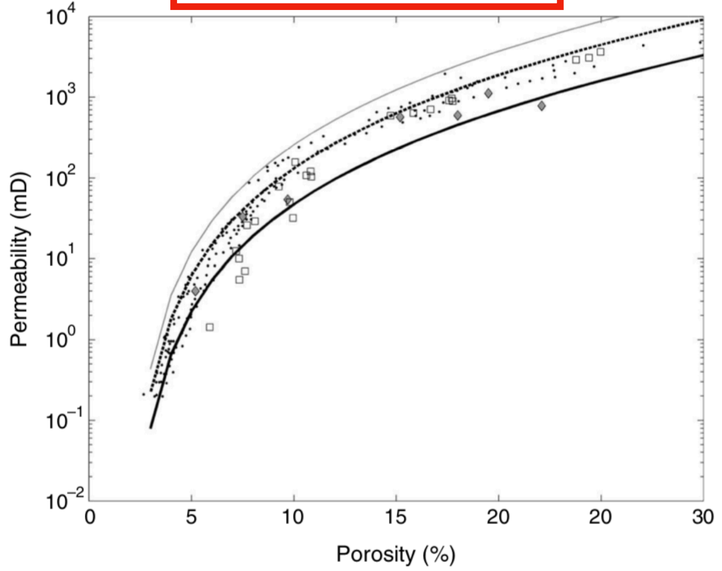
Sample	Porosity	Perm (mD)
A11	0.07	10
A16	0.07	6
A33	0.07	12
A82	0.08	7
A87	0.10	50
A89	0.08	26
A117	0.11	103
B31	0.11	107
B86	0.09	78
B101	0.11	121
B102	0.10	157
B108	0.08	29
F510	0.15	592
GT3	0.17	704
GW18	0.16	637
GW19	0.18	912
GW23	0.18	965
GW28	0.18	896
H27	0.25	3630
H42	0.24	2894
H74	0.24	3079
F410	0.06	1
F570	0.10	32

ABSTRACT

The relations among the resistivity, elastic-wave velocity, porosity, and permeability in Fontainebleau sandstone samples from the Ile de France region, around Paris, France were experimentally revisited. These samples followed a permeability-porosity relation given by Kozeny-Carman's equation. For the resistivity measurements, the samples were partially saturated with brine. Archie's equation was used to estimate resistivity at 100% water saturation, assuming a saturation exponent, $n = 2$. Using

0.65	88.29	216.49	9
0.76	19.34	65.67	6

$$m = 0.09 \ln\left(\sqrt{\frac{c\phi^3}{k}}\right) + 1.98. \quad (7)$$



Use Case 2: organic carbon content of Earth's crust

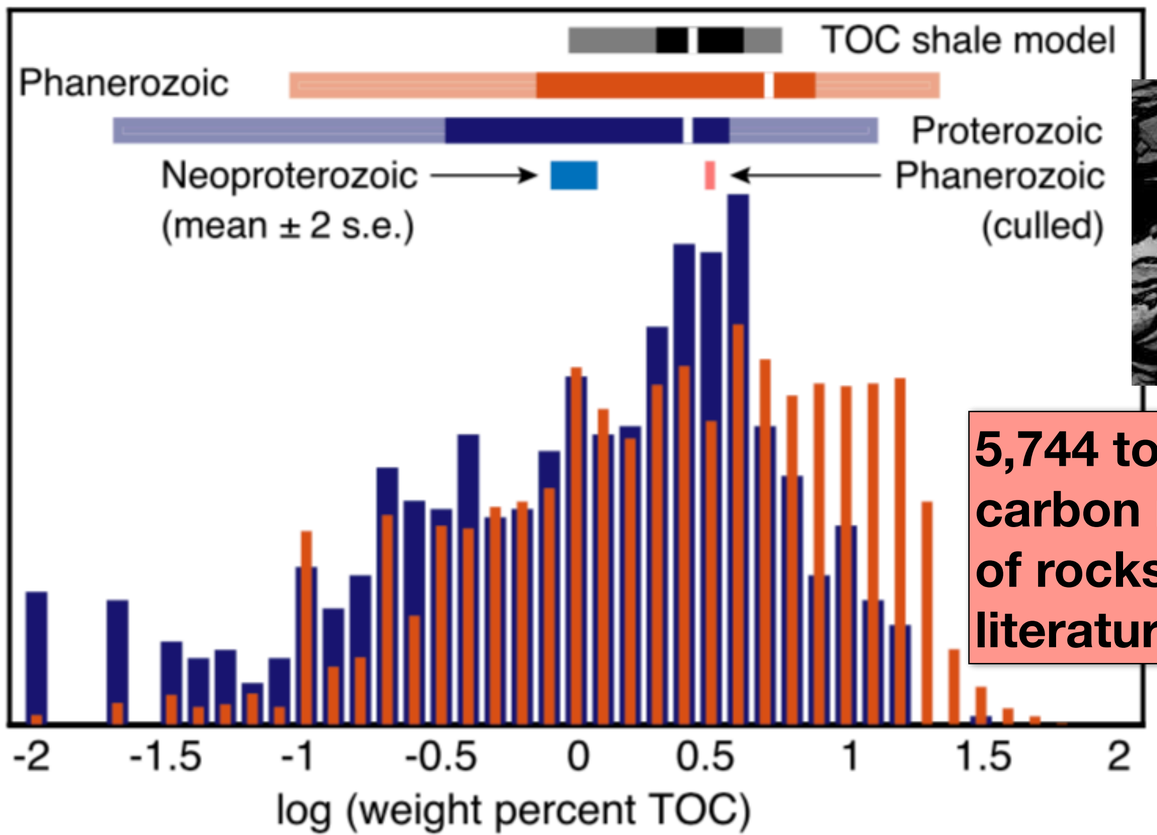
Contents lists available at ScienceDirect

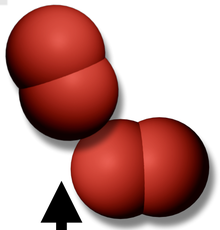
Earth and Planetary Science Letters

www.elsevier.com/locate/epsl

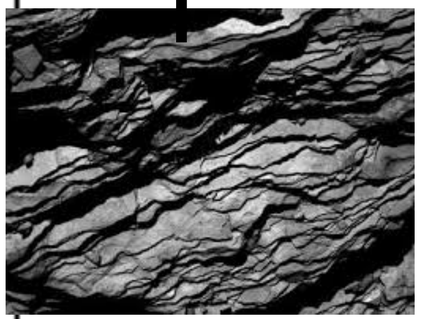
Atmospheric oxygenation driven by unsteady growth of the continental sedimentary reservoir

Jon M. Husson*, Shanan E. Peters

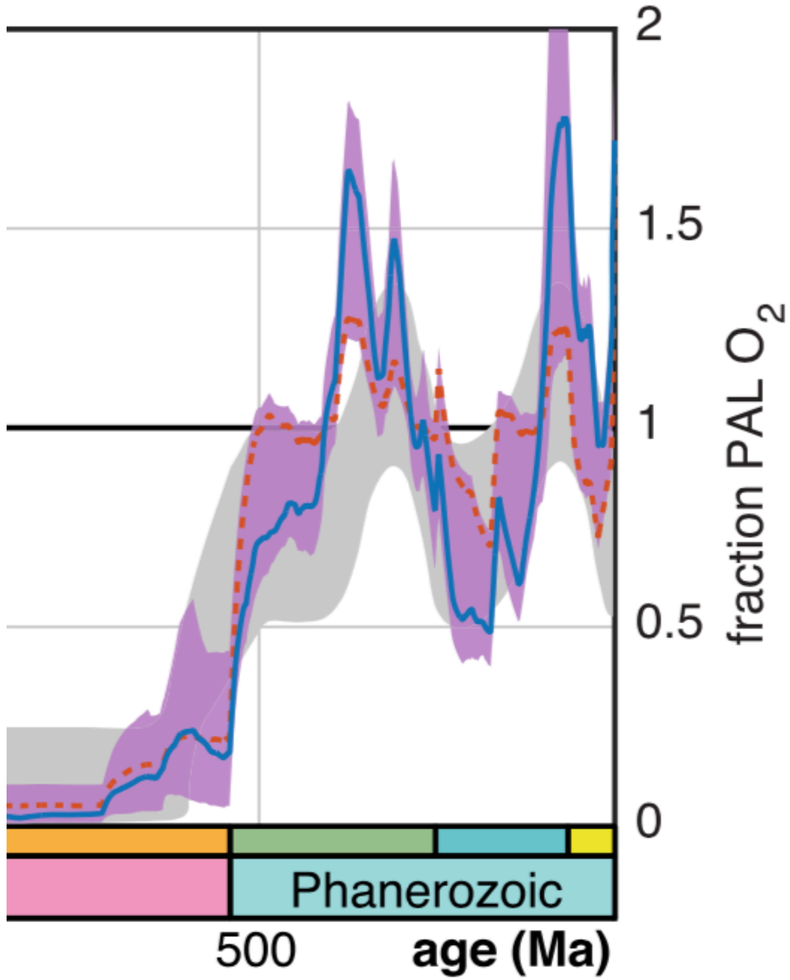




C gets buried
O_2 accumulates



5,744 total organic carbon measurements of rocks compiled from literature.



$$\frac{dM}{dt} = F_{org} - k_1 M - k_2 B_{org} \sqrt{M},$$

Use Case 2: organic carbon content of Earth's crust

COSMOS INPUT: vocabulary for specific measurement types and example context, xDD document acquisition/processing pipeline

Automate location and aggregation of contextualized sample-based measurements of organic carbon in rocks

USGS science for a changing world

Macrostrat

Improve inventories of unconventional fossil fuels, improve models of atmospheric oxygen

APPENDIX

Total Organic Carbon (TOC), Pyrolysis Data, and Organic Matter Types, Sites 535 (3450 m water depth) and 540 (2926 m water depth)

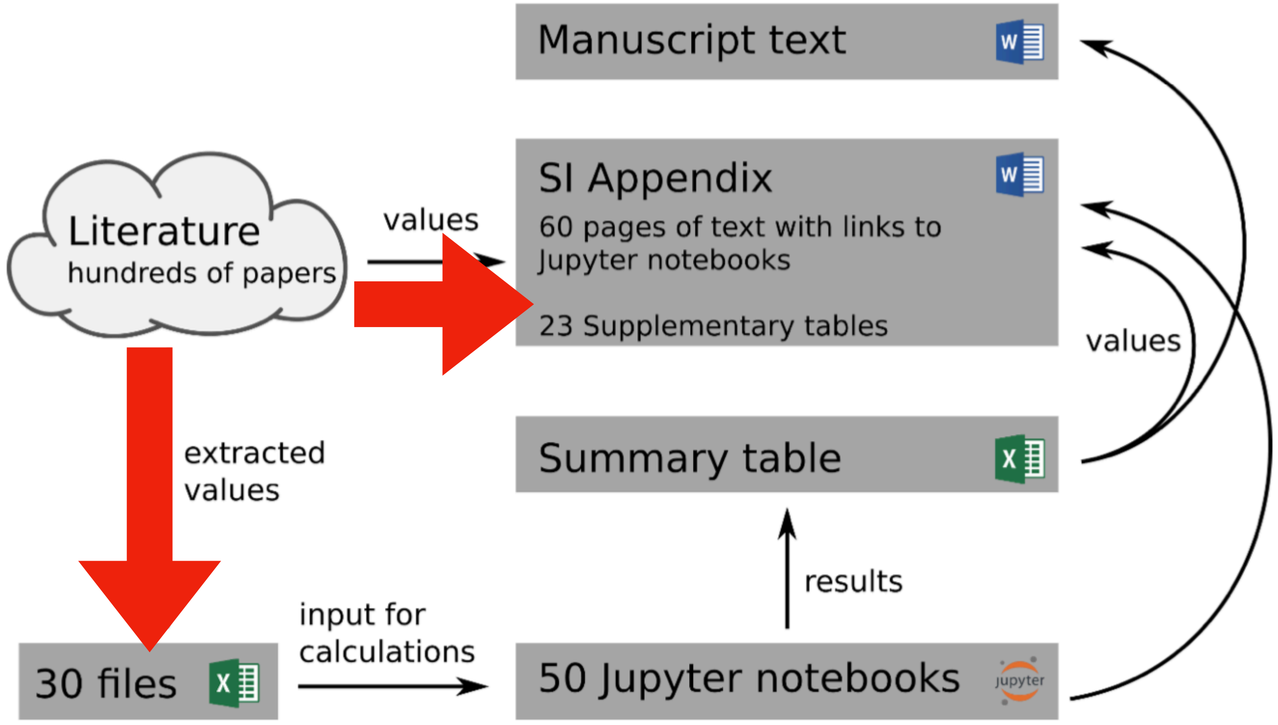
Lithologic Hole Unit	Hole	Core-section (interval in cm)	Sub-bottom depth (m)	Age[a]	Lithology	TOC (%)	Organic matter type[b] ST	PS	C	BT	AM	GA	SM	RB
I	535	2-2, 114–124	5	Pl.	Mud	1.15	40	5	25	10	15	5		
I		8-6, 120–130	68.7	Pl.	Clay	1.01	35	10	20	10	15	5	5	
I		15-4, 120–135	131.7	Pl.	Clay	0.89	25	10	20		30	10	5	
II		20-6, 139–150	182.5		Limestone	0.40			10		10	5	65	10
II		23-5, 137–150	209.4		Limestone	0.45								
II		27-2, 129–137	242.8		Limestone	0.38			5		10	10	65	10
II		31-5, 135–150	285.4	Ce.	Marly limestone	0.41								
II		35-4, 135–140	321.9		Limestone	0.32			5		25	60	5	5
II		38-3, 134–150	349.0		Limestone	0.23								
II		41-6, 120–133	381.7		Laminated limestone	1.15			5		40	45	5	5
III		48-3, 141–150	441.4	e.Ap.–l.Ha.	Laminated limestone	0.50			5		45	35	5	10
IV		52-3, 125–136	477.2		Laminated limestone	2.87	5		5		70m	5	5	10
IV		54-4, 120–131	492.7		Laminated marly limestone	1.01	5		5		70m	5	5	10
IV		57-5, 138–150	522.9	l.Ha.–e.Va.	Laminated marly limestone	0.27	5		10		45	30	5	5
IV		60-5, 120–129	550.2		Marly limestone	0.93	15	5	10		40	10	15	5
IV		63-4, 130–137	575.8		Marly limestone	0.22	10		15		45	15	10	5
IV		66-5, 137–150	604.4		Laminated marly limestone	1.06	5		10		55	20	5	5
V		69-4, 120–135	625.2	e.Va.–l.Be.	Limestone	0.17	10		20		30	15	20	5
V		72-3, 120–124	650.7	e.Va.–l.Be.	Limestone	0.18	20		10		30	10	25	5
I	540	1-2, 120–135	2.7	Pl.	Nannomarl	0.21	10		15		20	40	15	
II		4-4, 133–150	29.3	l.Mi.		0.19	10		15		20	40	15	
II		7-4, 120–133	57.7	l.Ol.		0.07								
II		10-3, 120–135	84.7	l.Ol.		0.07								
II		13-5, 135–150	116.4	l.Ol.		0.07								
II		16-5, 135–150	144.9	l.Ol.	Nannochalk	0.06								
II		19-4, 135–150	171.9	l.Ol.		0.06								
II		23-3, 134–150	208.3	e.Eo.		0.07								
II		26-3, 120–133	236.7	l.Eo.		0.05								
V		70-3, 135–150	654.9	l.Al.	Dark gray laminated	0.20	35		20		10	10	25	
V		73-2, 120–122	681.7	m.Al.	marly limestone	0.17	10		15		20	15	35	5
VI		78-3, 120–132	730.7	m.Al.	Banded limestone	0.14								

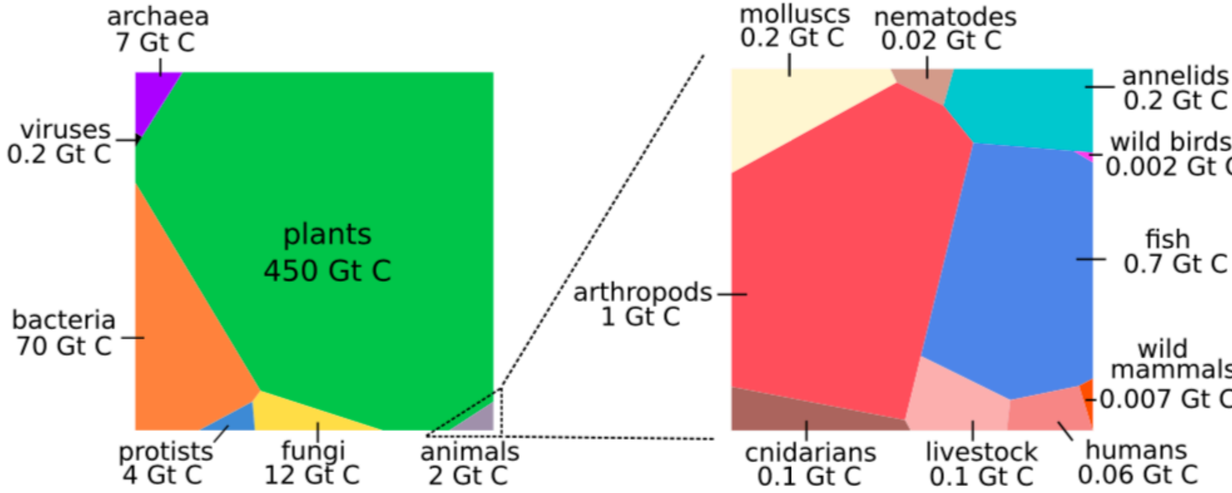
Use Case 3: biomass and its distribution on Earth

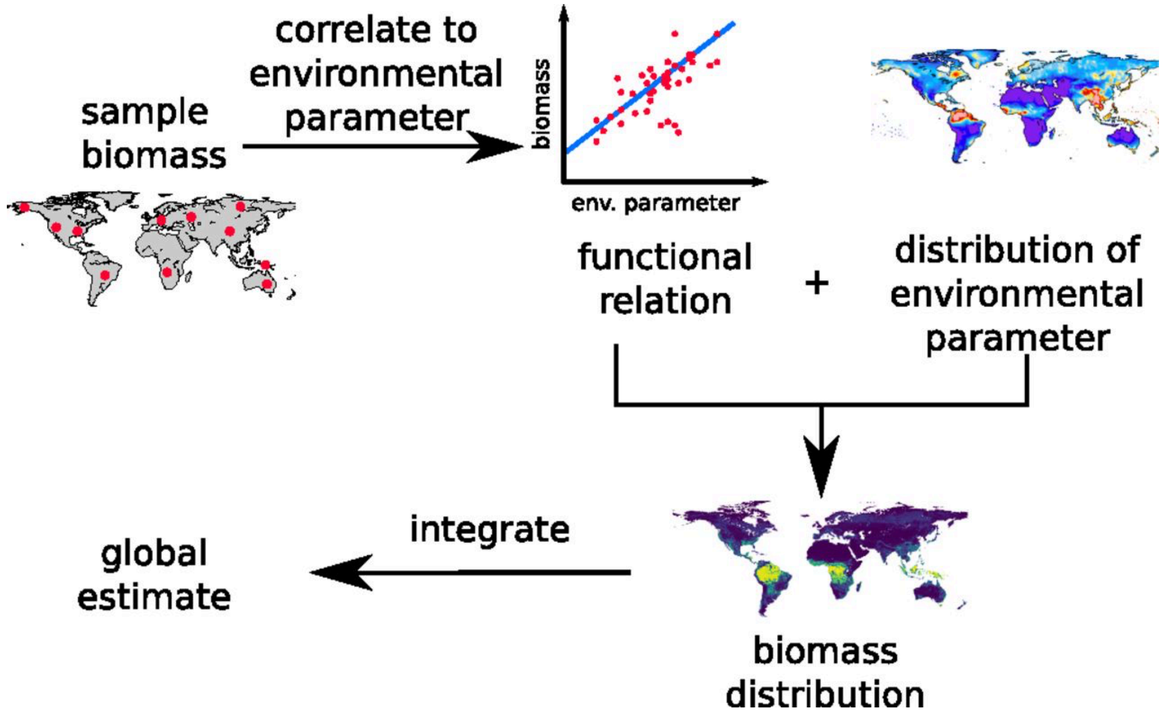
PNAS Proceedings of the National Academy of Sciences of the United States of America

The biomass distribution on Earth

Yinon M. Bar-On[a], Rob Phillips[b,c], and Ron Milo[a,1]







Use Case 3: biomass and its distribution on Earth

AMERICAN COMMITTEE FOR THE WEIZMANN INSTITUTE OF SCIENCE

SCIENCE FOR THE BENEFIT OF HUMANITY

Automate location and aggregation of contextualized sample-based measurements of biomass of specific organisms

Improve biomass inventories and build better models

COSMOS INPUT: vocabulary for specific measurement types, list of target taxonomic names, example context, xDD document acquisition/ processing pipeline

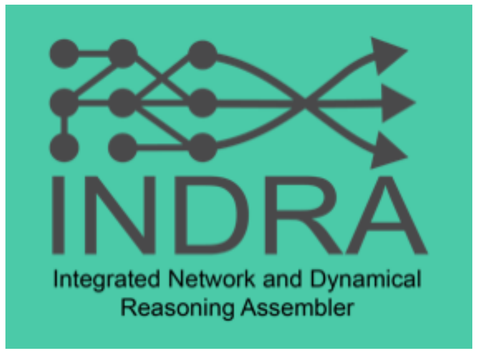
Table 1
Relative distribution (%) of soil macrofauna biomass in litter and soil layers, total density (ind m^{-2}) and total biomass (g m^{-2}) in a fores
(F), 4 years pasture (P4), and abandoned pasture (Pa) in Central Amazon

	F	P4	Pa 95	Pa 97
Relative distribution of macrofaunal biomass in litter (%)	3 (1)	0 (0)	0.5 (0)	0 (0)
Relative distribution of soil macrofaunal biomass in 0–5 cm (%)	13 (7)	6 (4)	27 (16)	4 (3)
Relative distribution of soil macrofaunal biomass in 5–10 cm (%)	3 (2)	19 (11)	44.5 (36.2)	12 (8)
Relative distribution of soil macrofaunal biomass in 10–25 cm (%)	81 (58)	75 (61)	28 (22)	84 (65)
Density of termites (ind m^{-2})	3608 (2113)	941 (689)	2117 (1455)	2266 (1652)
Density of ants (ind m^{-2})	2519 (2010)	322 (145)	2174 (1784)	402 (206)
Density of Oligachaeta (ind m^{-2})	136 (70)	284 (114)	0 (0)	0 (0)
Density of *P. corethrurus* (ind m^{-2})	0 (0)	0 (0)	390 (224)	136 (87)
Density of Isopoda (ind m^{-2})	35 (27)	5 (3)	38 (29)	110 (86)
Density of Diplopoda (ind m^{-2})	49 (28)	22 (17)	26 (19)	93 (78)
Density of Chilopoda (ind m^{-2})	54 (32)	0 (0)	158 (122)	10 (7)
Biomass of termites (g m^{-2})	5.29 (2.87)	0.61 (0.45)	3.43 (2.02)	4.28 (3.12
Biomass of ants (g m^{-2})	1.27 (0.98)	0.19 (0.14)	1.57 (1.21)	0.51 (0.30
Biomass of Oligachaeta (g m^{-2})	44.32 (26.24)	48.91 (29.21)	0 (0)	0 (0)
Biomass of *P. corethrurus* (g m^{-2})	0 (0)	0 (0)	45.12 (36.14)	7.76 (5.23
Biomass of Isopoda (g m^{-2})	0.02 (0.01)	0.21 (0.19)	0.05 (0.03)	1.28 (1.02
Biomass of Diplopoda (g m^{-2})	0.13 (0.06)	0.33 (0.25)	0.22 (0.13)	1.35 (1.11
Biomass of Chilopoda (g m^{-2})	0.96 (0.46)	0 (0)	0.55 (0.32)	0.07 (0.02

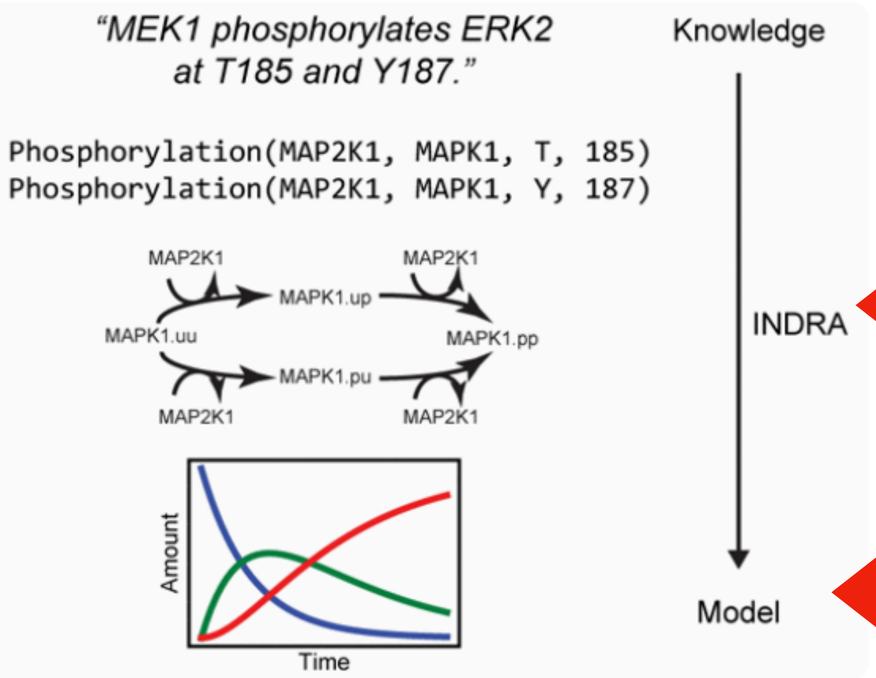
Means of 10 values, standard errors in parentheses.

Use Case 4: INDRA statements

COSMOS INPUT: large, general vocabulary, INDRA code, xDD document acquisition/processing pipeline

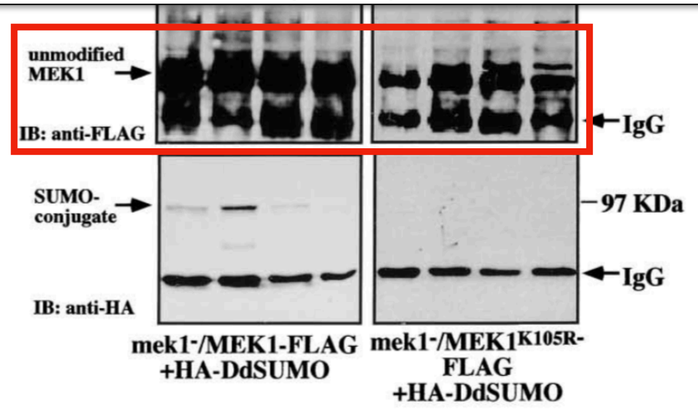


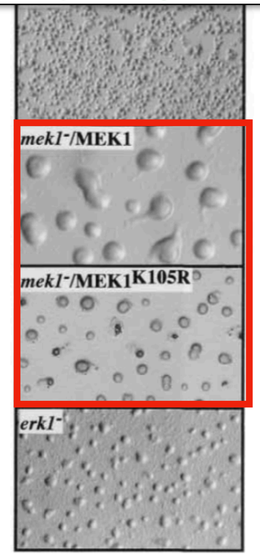

```
"title": "Regulated SUMOylation and Ubiquitination of DdMEK1 Is Required for Proper Chemotaxis",
"doi": "10.1016/S1534-5807(02)00186-7",
"coverDate": "June 2002",
"URL": "http://www.sciencedirect.com/science/article/pii/S1534580702001867",
"authors": "Sobko, Alex; Ma, Hui; Firtel, Richard A.",
"hits": 5,
"highlight": [
  ", San Diego 9500 Gilman Drive La Jolla, California 92093 Summary MEK1, which is required for aggregation",
  " to chemoattractant stimulation. SUMOylation is required for MEK1's function and its translocation from the",
  " to the cytosol and cortex, including the leading edge of chemotaxing cells. MEK1 in which the site",
  " of SUMOylation is mutated is retained in the nucleus and does not complement the mek1 null phenotype",
  ". Constitutively active MEK1 is cytosolic and is constitutively SUMOylated, whereas the corresponding"
```

API-based access to high-quality full-text mentions

Pipeline to generate INDRA statements





C.

Genotype (Strain)	Speed (μm/min)	Directional change (deg)	Roundness (%)	Directionality
Wild-type (KAx-3)	8.5±0.5	12.5±4.4	43.8±4.2	0.90±0.6
mek1-/MEK1 (ASF1)	8.1±0.4	12.4±0.8	41.3±4.8	0.89±0.05
mip1 null (HMF3)	5.7±0.3	41.1±4.5	51.9±0.7	0.59±0.06
mek1-/MEK1K105R	5.0±0.4	42.6±4.4	45.9±6.1	0.66±0.04