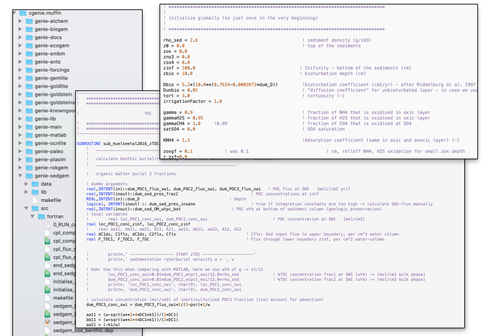
COSMOS: An AI platform for knowledge extraction from scientific publications

Theo Rekatsinas, Shanan Peters, Miron Livny

Publications are central to the scientific process

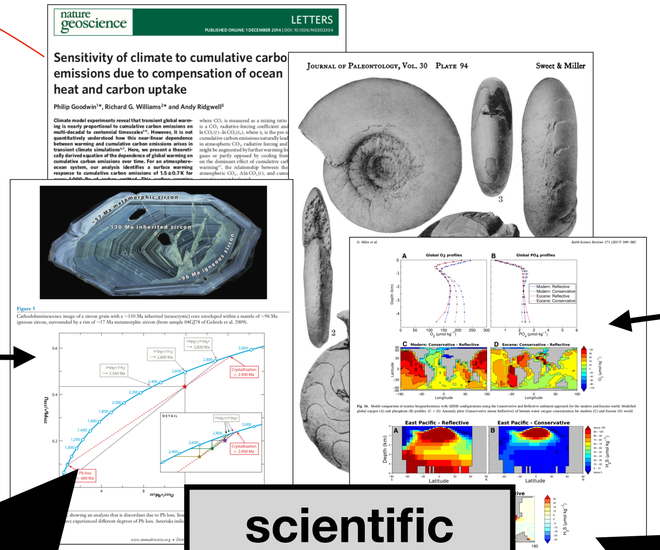
Structured Data

Scientific Models



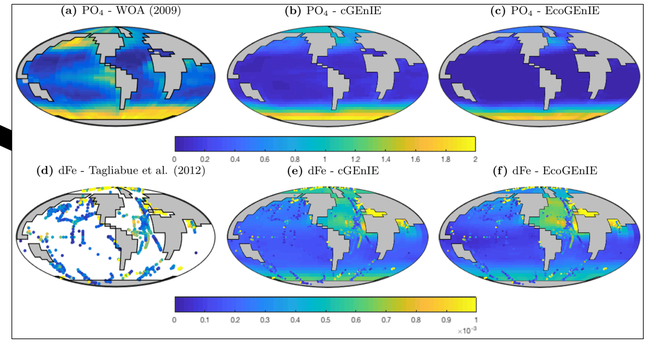
$$\frac{\partial c}{\partial t} = \nabla \cdot (D \nabla c) - \nabla \cdot (\mathbf{v} c) + R$$

deployed instruments



scientific publications, reports

“Dark Data”

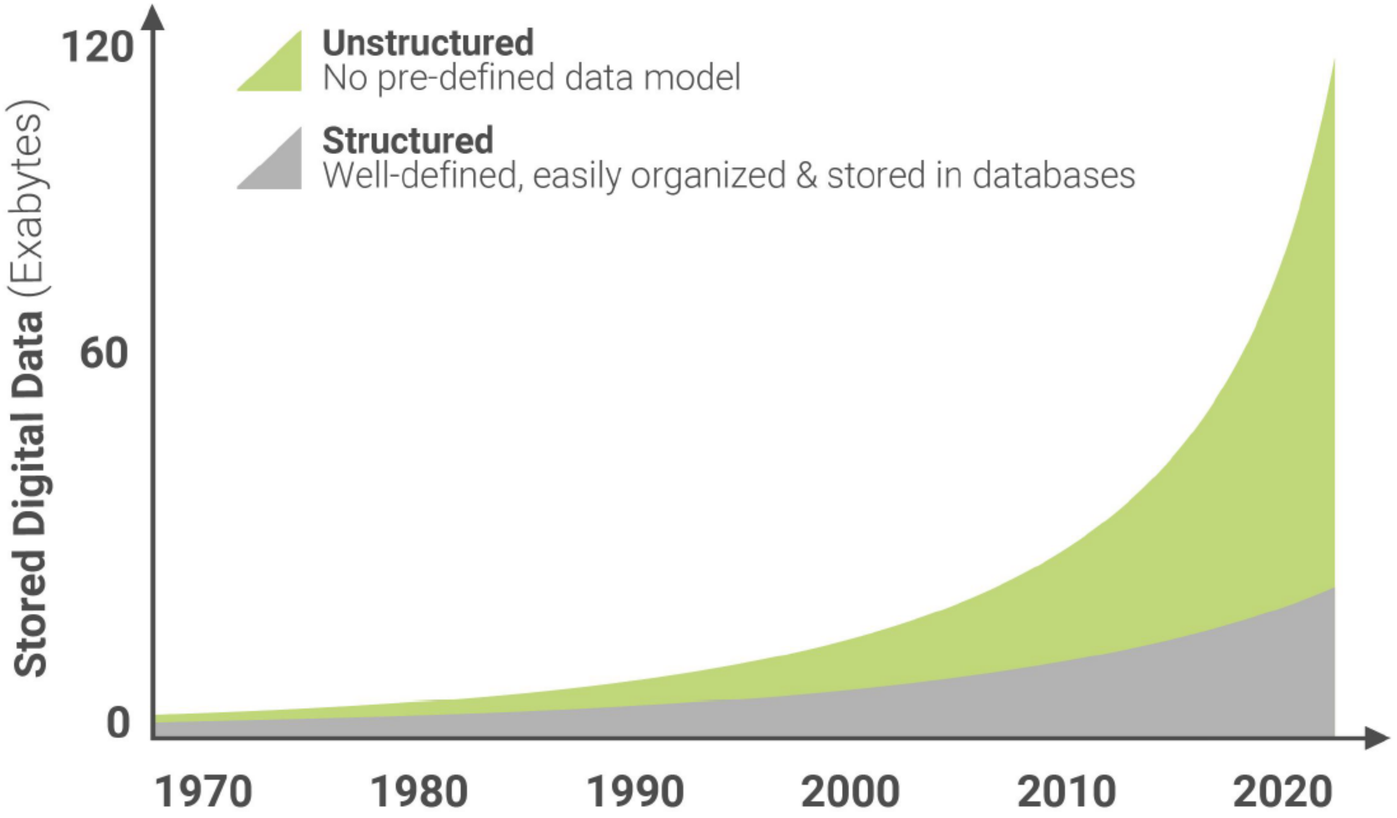


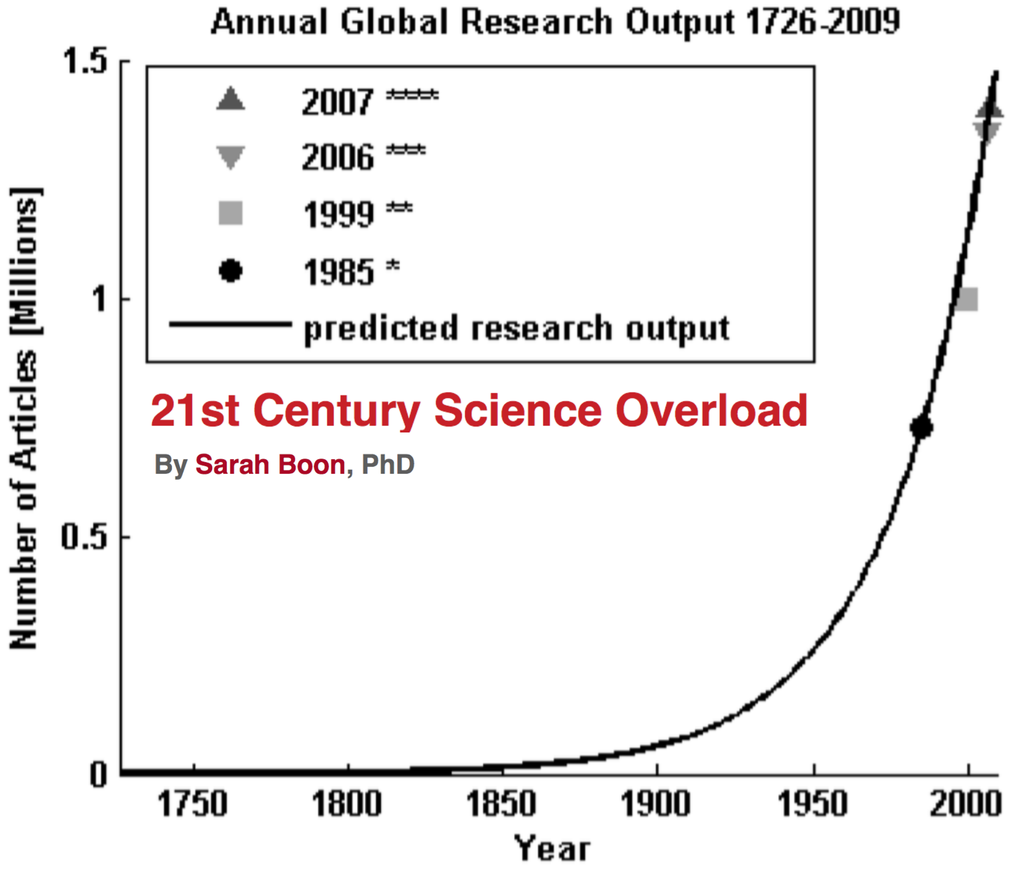
Model output

field observations

collections

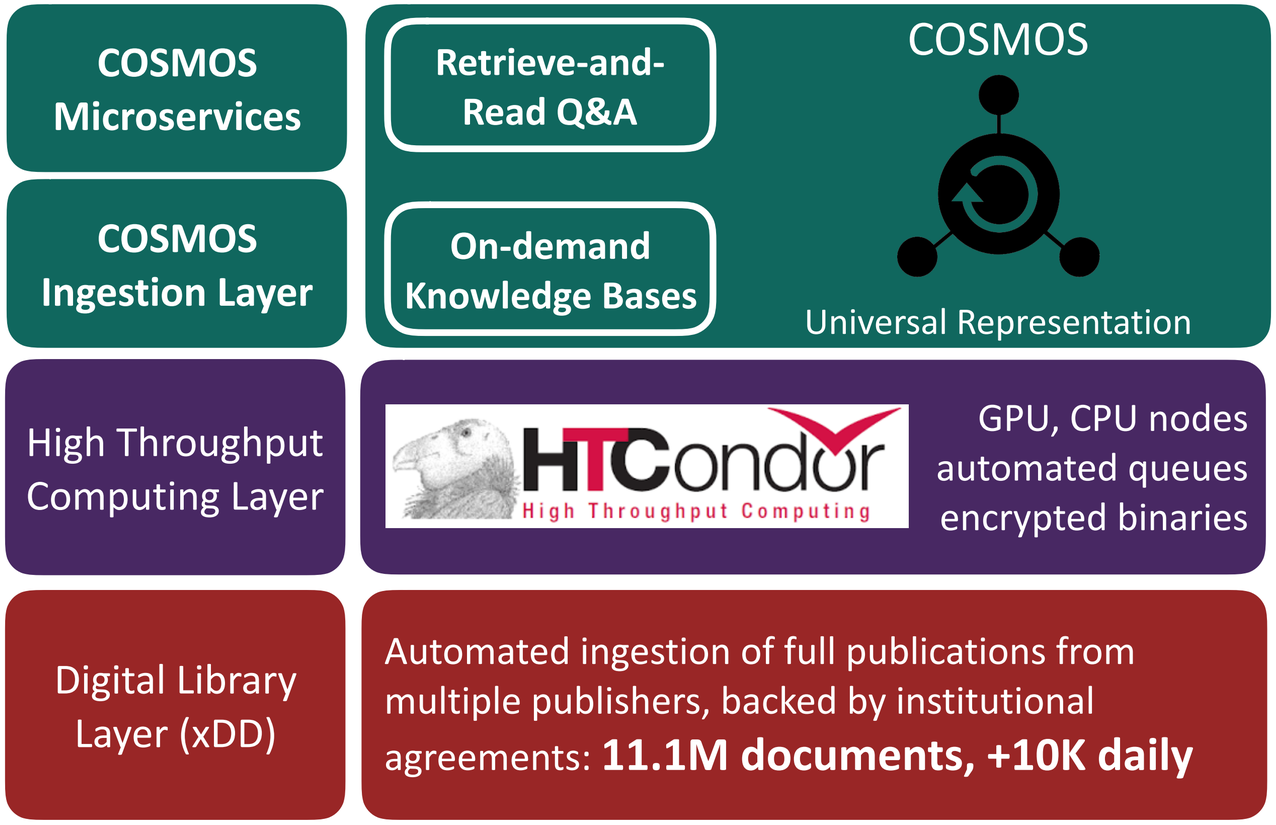
The gap between scientists and data is increasing





Automated **knowledge extraction** from unstructured data is required to close this gap and accelerate scientific progress.

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 data 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

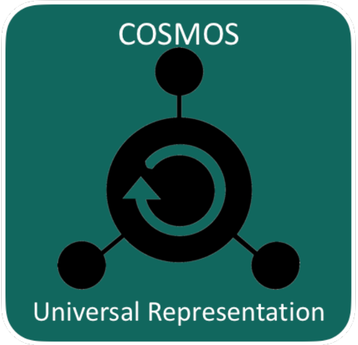
xDD API:
https://geodeepdive.org/api
Code available at:
https://github.com/UW-COSMOS

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

Outline

1. xDD: A portal to scientific publications and HTC

2. COSMOS: Knowledge extraction as a service

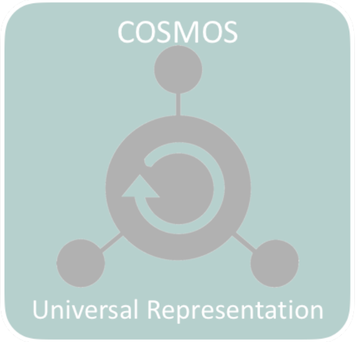


3. Demo: analyzing model code with COSMOS

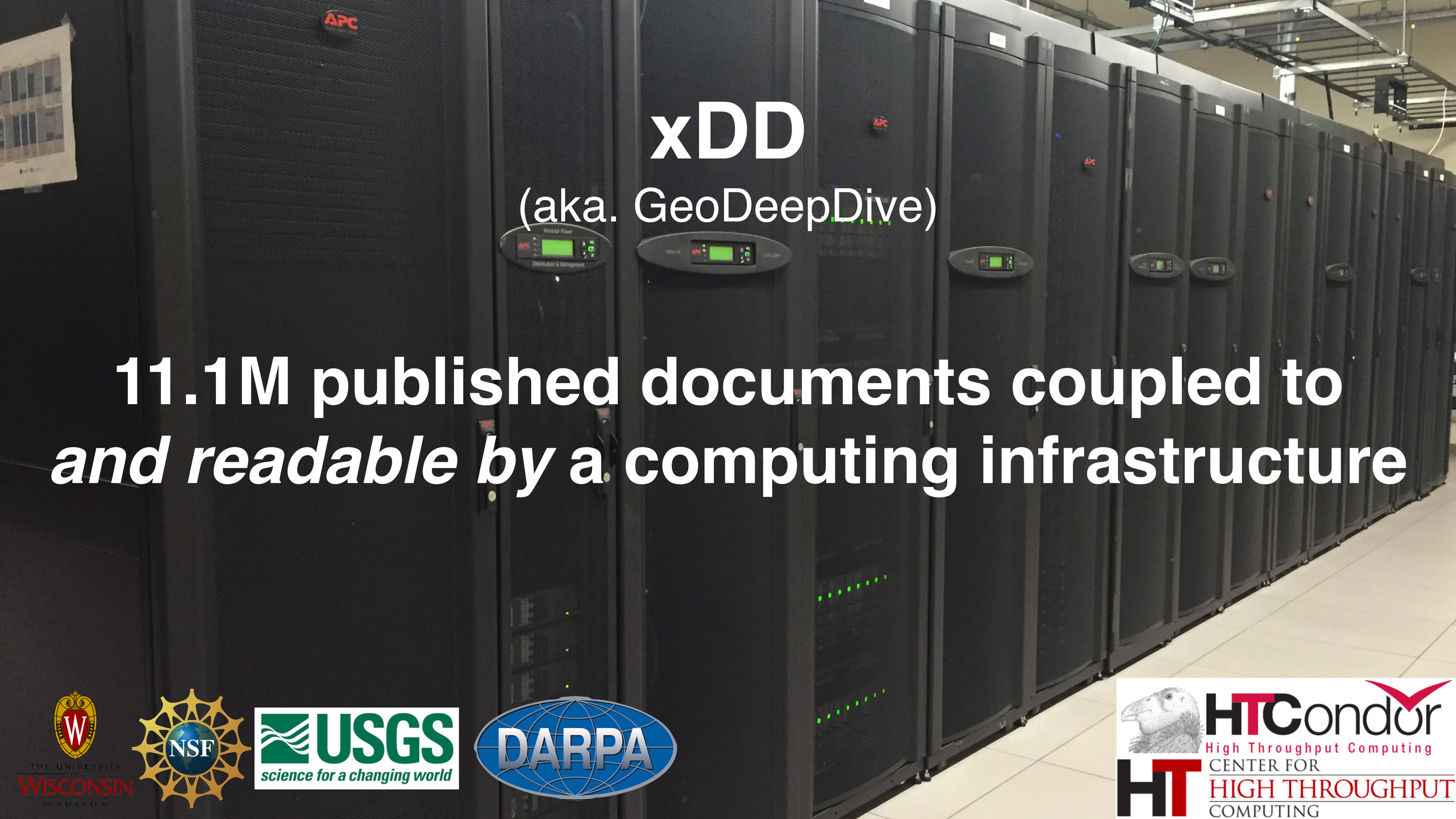
Outline

1. xDD: A portal to scientific publications and HTC

2. COSMOS: Knowledge extraction as a service



3. Demo: analyzing model code with COSMOS



xDD

(aka. GeoDeepDive)

11.1M published documents coupled to *and readable by* a computing infrastructure

THE UNIVERSITY of WISCONSIN MADISON

NSF

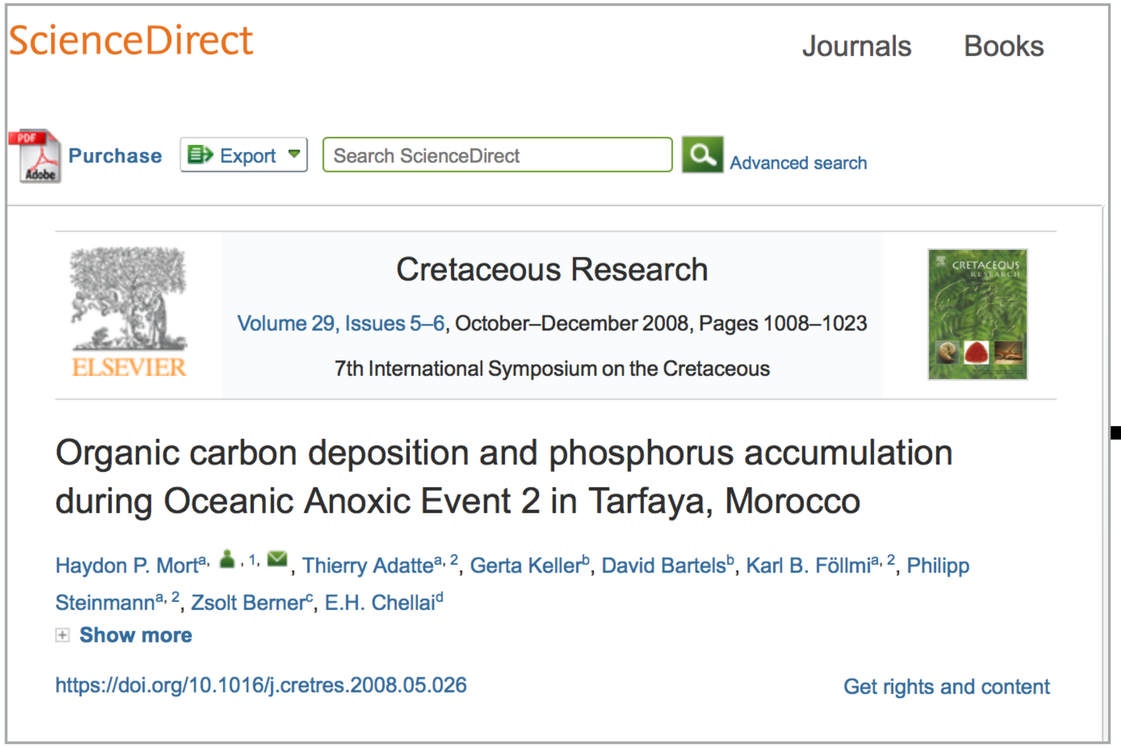
USGS science for a changing world

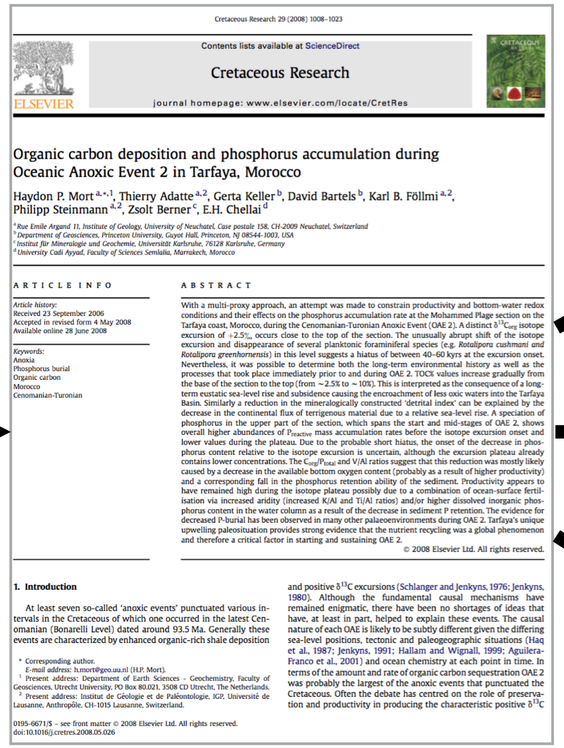
DARPA

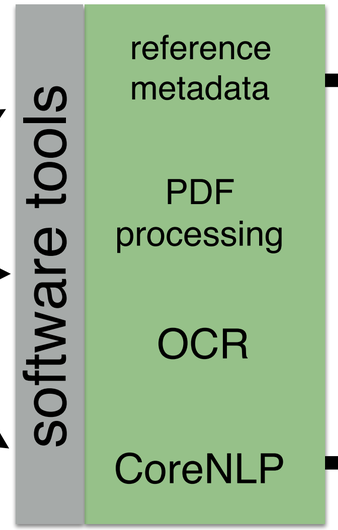
HTCondor High Throughput Computing

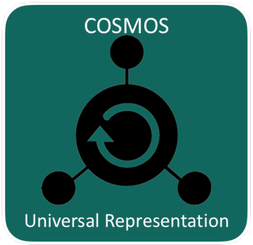
CENTER FOR HIGH THROUGHPUT COMPUTING

xDD Infrastructure Overview









publisher agreements

doc. fetching/storage

~10k/day

Vocabulary ingestion and labeling

geodeepdive.org/api/dictionaries

curated lists of terms
with hierarchy/context:

The Paleobiology Database
revealing the history of life

→ 350k
taxonomic names/
bio classification

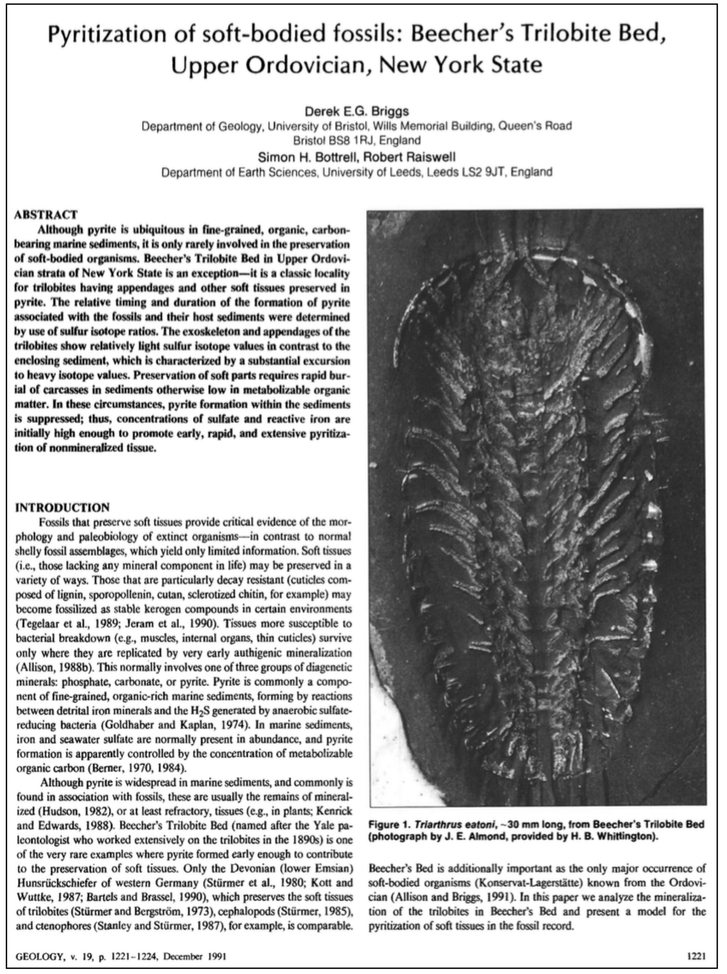
mindat.org

→ 6k
mineral names/
chemistry

Macrostrat

→ 45k
stratigraphic names/
hierarchy, rock types

GDD-supplied full text
and indexing capability



Pyritization of soft-bodied fossils: Beecher's Trilobite Bed, Upper Ordovician, New York State

Derek E.G. Briggs
Department of Geology, University of Bristol, Wills Memorial Building, Queen's Road
Bristol BS8 1RJ, England
Simon H. Bottrell, Robert Raiswell
Department of Earth Sciences, University of Leeds, Leeds LS2 9JT, England

ABSTRACT

Although pyrite is ubiquitous in fine-grained, organic, carbon-bearing marine sediments, it is only rarely involved in the preservation of soft-bodied organisms. Beecher's Trilobite Bed in Upper Ordovician strata of New York State is an exception—it is a classic locality for trilobites having appendages and other soft tissues preserved in pyrite. The relative timing and duration of the formation of pyrite associated with the fossils and their host sediments were determined by use of sulfur isotope ratios. The exoskeleton and appendages of the trilobites show relatively light sulfur isotope values in contrast to the enclosing sediment, which is characterized by a substantial excursion to heavy isotope values. Preservation of soft parts requires rapid burial of carcasses in sediments otherwise low in metabolizable organic matter. In these circumstances, pyrite formation within the sediments is suppressed; thus, concentrations of sulfate and reactive iron are initially high enough to promote early, rapid, and extensive pyritization of nonmineralized tissue.

INTRODUCTION

Fossils that preserve soft tissues provide critical evidence of the morphology and paleobiology of extinct organisms—in contrast to normal shelly fossil assemblages, which yield only limited information. Soft tissues (i.e., those lacking any mineral component in life) may be preserved in a variety of ways. Those that are particularly decay resistant (cuticles composed of lignin, sporopollenin, cutan, sclerotized chitin, for example) may become fossilized as stable kerogen compounds in certain environments (Tegelaar et al., 1989; Jeram et al., 1990). Tissues more susceptible to bacterial breakdown (e.g., muscles, internal organs, thin cuticles) survive only where they are replicated by very early authigenic mineralization (Allison, 1988b). This normally involves one of three groups of diagenetic minerals: phosphate, carbonate, or pyrite. Pyrite is commonly a component of fine-grained, organic-rich marine sediments, forming by reactions between detrital iron minerals and the H_2S generated by anaerobic sulfate-reducing bacteria (Goldhaber and Kaplan, 1974). In marine sediments, iron and seawater sulfate are normally present in abundance, and pyrite formation is apparently controlled by the concentration of metabolizable organic carbon (Berner, 1970, 1984).

Although pyrite is widespread in marine sediments, and commonly is found in association with fossils, these are usually the remains of mineralized (Hudson, 1982), or at least refractory, tissues (e.g., in plants; Kenrick and Edwards, 1988). Beecher's Trilobite Bed (named after the Yale paleontologist who worked extensively on the trilobites in the 1890s) is one of the very rare examples where pyrite formed early enough to contribute to the preservation of soft tissues. Only the Devonian (lower Emsian) Hunsrückschiefer of western Germany (Stürmer et al., 1980; Kott and Wuttke, 1987; Bartels and Brassel, 1990), which preserves the soft tissues of trilobites (Stürmer and Bergström, 1973), cephalopods (Stürmer, 1985), and ctenophores (Stanley and Stürmer, 1987), for example, is comparable.

Figure 1. *Triarthrus eatoni*, ~30 mm long, from Beecher's Trilobite Bed (photograph by J. E. Almond, provided by H. B. Whittington).

Beecher's Bed is additionally important as the only major occurrence of soft-bodied organisms (Konservat-Lagerstätte) known from the Ordovician (Allison and Briggs, 1991). In this paper we analyze the mineralization of the trilobites in Beecher's Bed and present a model for the pyritization of soft tissues in the fossil record.

labeled
entities, tuples

DOI:
10.1130/0091-7613(1991)019<1221:POSBFB>2.3.CO;2

Trilobita
Triarthrus
Climacograptus

pyrite

Frankfort Shale

exposed
via API

```
"term_hits": { ▼ 189470 properties, 6 MB
  "Navajosuchus novomexicanus": 7,
  "Ptilocolepidae": 2,
  "Geotrupidae": 652,
  "Macropoma lewesiensis": 11,
  "Simia morio": 7,
  "Anhanguera robustus": 4,
  "Montipora verrilli": 36,
  "Geffenina wangi": 6,
  "Shuvosaurus": 198,
  "Dryorhizopsidae": 1,
  "Ostrea antarctica": 12,
  "Pholidophorus dentatus": 10,
  "Oochorista": 6,
  "Sciurus arizonensis": 18,
  "Prosbole biexcisa": 3,
  "Toxopatagus": 271,
  "Sulcavitus": 169,
  "Brontops amplus": 2,
  "Melonella": 1087,
  "Bathrotomaria": 688,
  "Placochelyanus": 64,
  "Iliochione": 555,
  "Attenosaurus subulensis": 10,
  "Miccylotyrans": 4,
  "Deryeuma": 4,
  "Chaetabraeus": 2,
  "Cardinalis cardinalis": 1113,
  "Odontaster": 2532,
  "Coeloma": 384,
  "Megatrema": 36,
  "Xinjiangtitan": 4,
  "Tephrodytes brassicarvalis": 74,
  "Prolagus crusafonti": 42,
  "Streblascopora germana": 2,
  "Dichobunoidea": 11,
  "Cetotherium hupschi": 1,
  "Hauericeras (Gardeniceras) gardeni": 10,
  "Parevania": 15,
  "Histriobdellidae": 130,
  "Berosus (Berosus)": 13,
  "Dinornis elephantopus": 8,
  "Sternoxi": 11,
  "Bellimurina (Bellimurina)": 4,
  "Raphignathoidea": 76,
  "Cybelurus occidentalis": 4,
```


Define your own dictionaries:
https://github.com/UW-Deepdive-Infrastructure/dictionary_example

```
"term_hits": { ▼ 189470
  "Navajosuchus novomex
  "Ptilocolepidae": 2,
  "Geotrupidae": 652,
  "Macropoma lewesiensis
  "Simia morio": 7,
  "Anhanguera robustus":
  "Montipora verrilli":
  "Geffenina wangi": 6,
  "Shuvosaurus": 198,
  "Dryorhizopsidae": 1,
  "Ostrea antarctica": 1
  "Pholidophorus dentatu
  "Oochorista": 6,
  "Sciurus arizonensis":
  "Prosbole biexcisa": 3
  "Toxopatagus": 271,
  "Sulcavitus": 169,
  "Brontops amplus": 2,
  "Melonella": 1087,
  "Bathrotomaria": 688,
  "Placochelyanus": 64,
  "Iliochione": 555,
  "Attenosaurus subulensis": 10,
  "Miccylotyrans": 4,
    "pubname": "Earth-Science Reviews",
    "publisher": "Elsevier",
    "title": "The Cambrian palaeontological record of the Indian subcontinent",
    "coverDate": "Available online 11 June 2016",
    "URL": "http://www.sciencedirect.com/science/article/pii/S0012825216301179",
    "authors": "Hughes, Nigel C.",
    "hits": 4,
    "highlight": [ ▼ 4 items, 411 bytes
      " represented is the hyolithimorph Sulcavitus wynnei (Waagen, 1885) collected
      from several horizons within",
      ") Hy Sulcavitus  wynnei (Waagen), dorsum, Khussak Formation, Salt Range, SH,
      GSI 4118  (CMCIP 71490",
      ",) Kruse and Hughes in press, fig. 5C, scale bar: 2.5 mm; T) Hy  Sulcavitus
      wynnei (Waagen), venter",
      " 71490), Kruse and Hughes in press, fig. 5A, scale bar: 2.5 mm; U) Hy
      Sulcavitus wynnei (Waagen"
    ]
  },
  { ▼ 8 properties, 431 bytes
    "pubname": "Alcheringa: An Australasian Journal of Palaeontology",
    "publisher": "Taylor and Francis",
    "title": "Biostratigraphic potential of Middle Cambrian hyoliths from the eastern
    Georgina Basin",
    "coverDate": "2002 01",
    "URL": "http://www.tandfonline.com/doi/abs/10.1080/03115510208619263",
    "authors": "Kruse, Peter D.",
    "hits": 1,
    "highlight": [ ▼ 1 item, 94 bytes
      ",it is noteworthy that Sulcavitus possesses a distinctive deep median sulcus on
      the dorsum"
    ]
```


Define your own dictionaries:
https://github.com/UW-Deepdive-Infrastructure/dictionary_example

Pithecanthropus erectus

Mammalia - Primates - Hominidae

GeoDeepDive matched this taxon in 229 documents from 73 journals/publications:

- Soriano, M.. *The fluoric origin of the bone lesion in thePithecanthropus erectus femur*. **American Journal of Physical Anthropology January 1970**.
  ...The Fluoric Orig- in of the Bone Lesion in the *Pithecanthropus erectus* Femur Facultad de Medicina...

  ... that the *Pithecanthropus erectus* suffered a bone fluorosis of the t m e of the Periostitis deformans...

  ..., discovered by Dr. Eugene Dubois in 1891-1892 and attributed to the *Pithecanthropus erectus*, have...

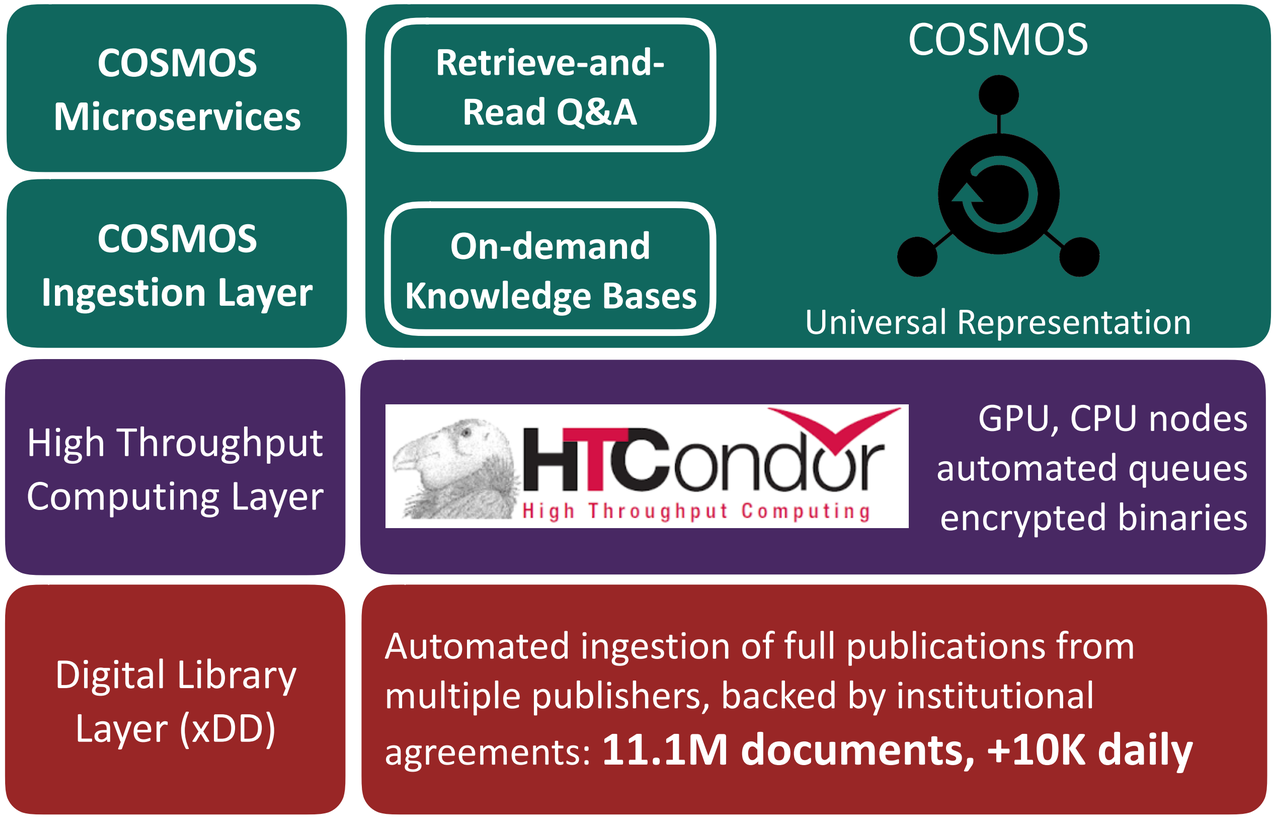
  ... a bone lesion very similar to that of the *Pithecanthropus erectus*. Both bones present several...

  .... PHYS.ANTHROP.,32: 49-58. 49 t Fig. 1 Femur of the *Pithecanthropus erectus*. (Taken from Julien, '69...

  Hide recognized terms from this document

Taxonomic names	Homo Homo erectus Pithecanthropus Pithecanthropus erectus
Lithologies	ash volcanic

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

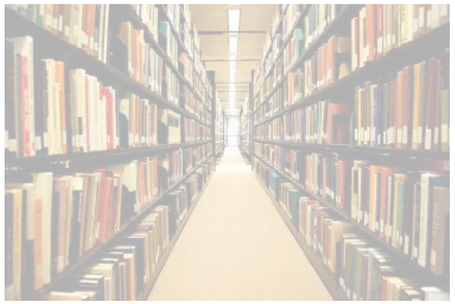
support 2014-2018 by NSF-ICER 1343760, DARPA ASKE
partial current support from USGS

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

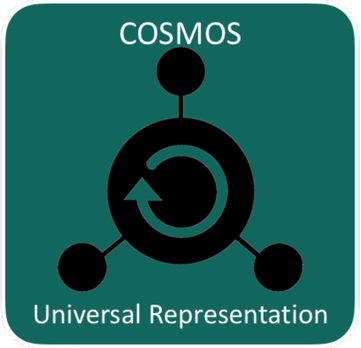
Correspondance:
Shanan Peters (peters@geology.wisc.edu)
Theodoros Rekatsinas (thodrek@cs.wisc.edu)
Miron Livny (miron@cs.wisc.edu)

Outline

1. xDD: A portal to scientific publications and HTC

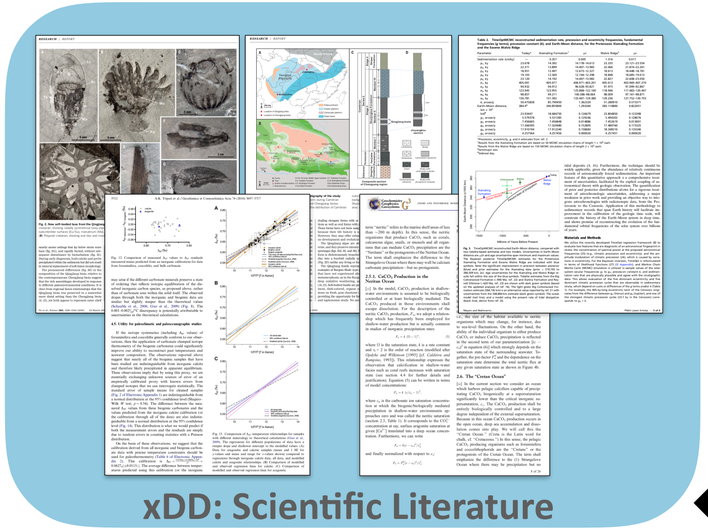


2. COSMOS: Knowledge extraction as a service

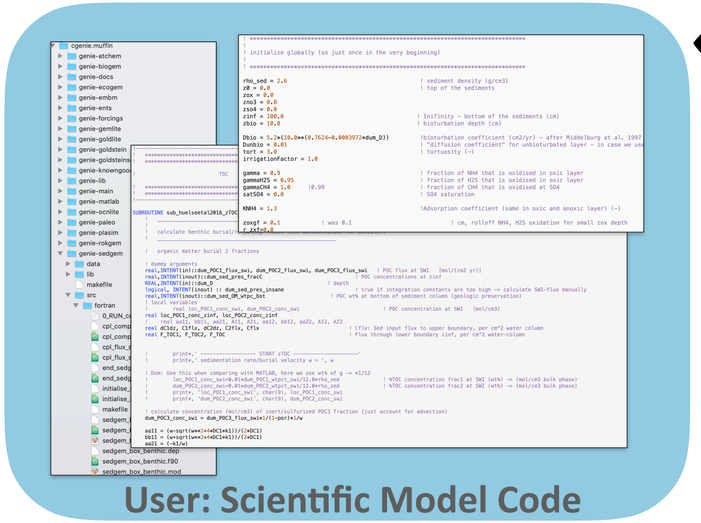


3. Demo: analyzing model code with COSMOS

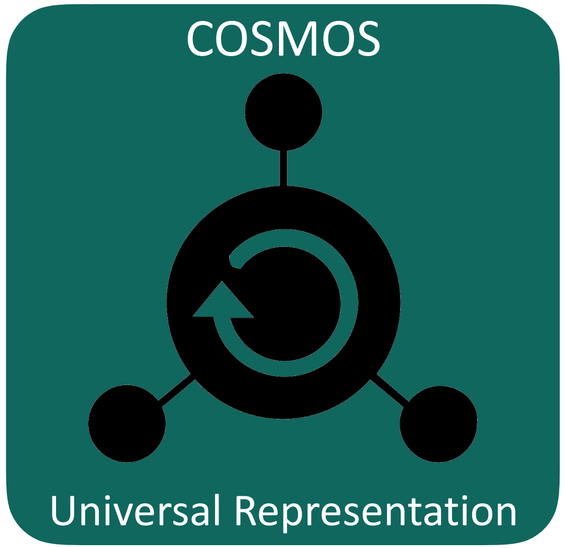
Accelerating scientific discovery with COSMOS



xDD: Scientific Literature

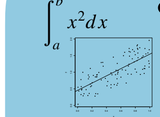


User: Scientific Model Code



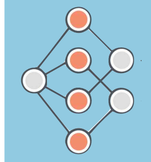
- Retrieve-and-Read Q&A
- On-demand Knowledge Bases
- Model Evaluation
- Dataset Generation

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



Sample	206Pb/	2s
P-S-113_	1690	30
P-S-113_	1717	30
P-S-113_	1666	29
P-S-113_	1641	28
P-S-113_	1619	28

- 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

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, $\sim 2000\ m^3\ s^{-1}$), reaching a peak flood on 28 June (the second flood, $\sim 3000\ m^3\ s^{-1}$). 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

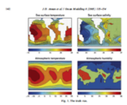
Values

From PDF to XML

Mutli-page raw PDF

Document Preprocessing Tool

Page 1



Page 2

Page k

Images corresponding to PDF pages

ML Model

ML Model

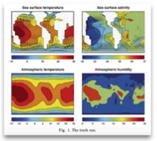
ML Model

Equations

$$P(\mathbf{X}|\mathbf{\Phi}) = \prod \frac{1}{\sqrt{2\pi}\sigma_i} e^{-\frac{(x_i - \mu_i)^2}{\sigma_i^2}} = Ke^{-\frac{1}{2}\sum_i \frac{(x_i - \mu_i)^2}{\sigma_i^2}}$$

$$\ln(P(\mathbf{\Phi})) = c \pm \frac{1}{2} \sum_j \frac{(p_i - p'_j)^2}{\tau_j^2}.$$

Figure

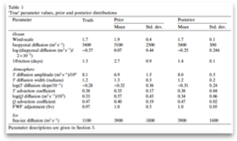


Text

The vertically averaged effect of atmospheric advective transport is represented by advection with a fixed surface wind field derived from NCEP analysis, scaled by a coefficient, following Weaver et al. (2001), which takes different values for moisture advection and for the zonal advection of heat. There is no meridional advection of heat in the model atmosphere. To compensate for a lack of inter-basin moisture transport produced by our simplified atmosphere we introduce a constant redistribution of moisture from the surface Atlantic to Pacific following the pattern observed by Oort (1983). This 'flux adjustment' is an important parameter of the model hydrological cycle, but is unrelated to the flux adjustments required in early coupled models to prevent climate drift, which is not a problem for this model. Sea-ice height and fractional area are advected by the surface ocean currents and diffused with constant diffusivity.

Fixing the spatial distributions of drag and atmospheric diffusivity, we thus have a set of 12 scalar model parameters which may vary. The prior distributions for all 12 parameters are given in Table 1. The ranges of the distributions are chosen to cover, or exceed, a range of reasonable choices of appropriate values for such a model as discussed by Edwards and Marsh (2003).

Table



Visual Extractions of Elements of Interest

OCR Engine

OCR Engine

OCR Engine

OCR Engine

OCR Engine

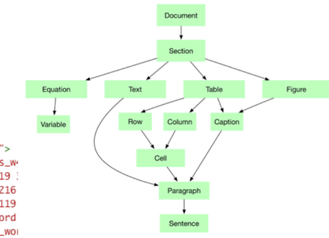
OCR Enrichment

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>
```

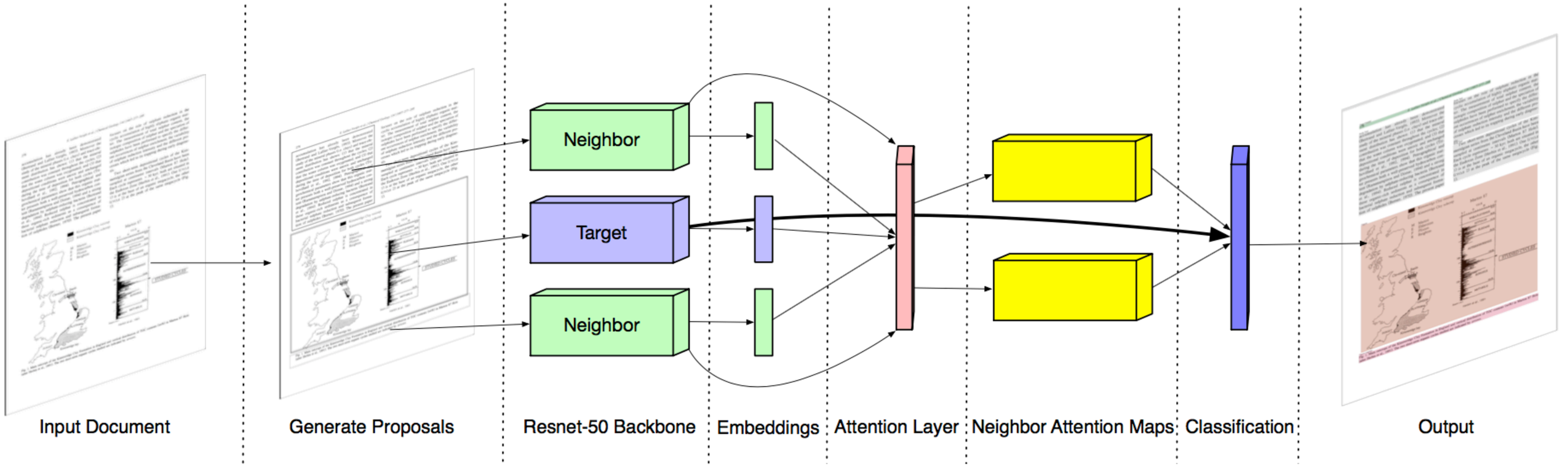


HTML Representation of initial PDF following Fonduer's unified data model

How do we get these elements from a scanned image?

We need support for images to address the format heterogeneity across publications.

The COSMOS Attentive RCNN Model



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

2.2.1. Ocean model

The ocean model is divided into two submodels: physical and biological–chemical submodels.

Physical submodel: The governing equations of the submodel are as follows: the equation of motion for rotational fluid under the assumption of hydrostatic and Boussinesq approximations (Eq. (1)), the equation of continuity of incompressible fluid (Eq. (2)), and equations of advection–diffusion of the salinity (Eq. (3)), and heat included in water (Eq. (4)).

$$\frac{\partial \mathbf{v}_w}{\partial t} + \mathbf{v}_w \cdot \nabla \mathbf{v}_w - f \mathbf{v}_w \times \boldsymbol{\omega} = -\frac{1}{\rho_{w0}} \nabla p + \frac{\partial}{\partial z}\left(K_{v,w} \frac{\partial \mathbf{v}_w}{\partial z}\right) + K_{h,w}\left(\frac{\partial^2 \mathbf{v}_w}{\partial x^2} + \frac{\partial^2 \mathbf{v}_w}{\partial y^2}\right), \quad (1)$$

$$\nabla \cdot \mathbf{v}_w = 0, \quad (2)$$

$$\frac{\partial S_w}{\partial t} + \mathbf{v}_w \cdot \nabla S_w = \frac{\partial}{\partial z}\left(D_v \frac{\partial S_w}{\partial z}\right) + D_{h,w}\left(\frac{\partial^2 S_w}{\partial x^2} + \frac{\partial^2 S_w}{\partial y^2}\right) + (RIVER), \quad (3)$$

$$\frac{\partial T_w}{\partial t} + \mathbf{v}_w \cdot \nabla T_w = \frac{\partial}{\partial z}\left(D_v \frac{\partial T_w}{\partial z}\right) + D_{h,w}\left(\frac{\partial^2 T_w}{\partial x^2} + \frac{\partial^2 T_w}{\partial y^2}\right), \quad (4)$$

where f denotes Coriolis parameter, and ρ_{w0} denotes the reference density of sea water. Free surface elevations are computed by calculating the convergence or divergence of barotropic components of water flow. Vertical diffusion coefficients are estimated using the turbulence model for ocean boundary layer proposed by Noh and Kim (1999), which is a simplified and improved version of the turbulent closure scheme by Mellor and Yamada (1982). The abovementioned equations are described in the Cartesian coordinate system. Ocean bottom topography is assumed to have a partial-step form, as noted by Adcroft et al. (1997) for representing the bottom slopes realistically in the Cartesian coordinate system.

These equations are numerically solved by the finite-difference methods. In our computational code, for spatial differences, the Uniformly Third Order Polynomial Interpolation Algorithm (UTOPIA) and the second order central difference is used for the advection and diffusion terms, respectively. The leap-frog scheme is employed for time differences. The detail of the physical submodel is described in Nishi et al. (2004).

Biological–chemical submodel: The biological–chemical submodel (hereafter referred to as the BC submodel) in the ocean model calculates biomass variation $B(X_w)$ in the pelagic system, which is associated with photosynthesis, respiration, extracellular excretion, grazing, decomposition, and mortality by some organisms in the pelagic system (released algae, phytoplankton, zooplankton, or bacteria). The detail of mathematical formulations of these processes are described in Appendix A.

Total biomass variation can be computed by combining results from the BC submodel with results from the physical submodel (advection and diffusion) as follows:

$$\frac{\partial X_w}{\partial t} + \mathbf{v}_w \cdot \nabla_w X_w = D_{h,w}\left(\frac{\partial^2 X_w}{\partial x_w^2} + \frac{\partial^2 X_w}{\partial y_w^2}\right) + \frac{\partial}{\partial z_w}\left(D_{v,w}\left(\frac{\partial X_w}{\partial z_w}\right)\right) + B(X_w). \quad (5)$$

The second term on the left-hand side of Eq. (5) represents the advection by water current. The first and second term on the right-hand side of Eq. (5) represent the horizontal and vertical diffusion, respectively. Eq. (5) is solved numerically using the same method as that in the physical submodel.

2.2.2. Ice model

The ice model is divided into three submodels: physical, biological–chemical, and spectral irradiance submodels. Several variables in a submodel have connections with those in other submodels. Our model considers the following six processes of these connections: (1) the attenuation coefficient of ice is a function of brine volume; (2) the photosynthetic available radiation (PAR) is a function of ice thickness that obeys Beer–Lambert's law; (3) ice temperatures affect the activity of organisms in the ice; (4) the photosynthetic rate of the ice algae is limited by brine salinity; (5) Chl-*a* included in ice algal cells absorbs light in the ice, and (6) the photosynthetic rate of the ice algae is also limited by the intensity of the photosynthetic available radiation.

In this model, snow ice formation which is an important factor for ice algal activity, is not

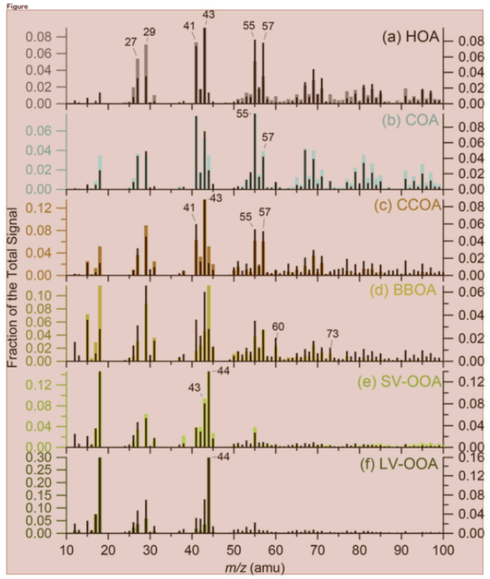


Fig. 10. Typical mass spectral profiles of OA factors resolved from PMF-ACSM at various sites in China (HOA, BBOA, SV-OOA and LV-OOA were from Sun et al. (2016b), COA was from Sun et al. (2014) and CCOA was from Sun et al. (2013c). The standard mass spectra of OA factors reported in Ng et al. (2011b) are also shown in right axes for comparisons.

significantly increased contribution from approximately 10% at an RH of less than 20% to approximately 40% at an RH exceeding 80%, suggesting that coal combustion is an important source of PM pollution during more humid periods, which likely promoted the partitioning of water-soluble organics from coal combustion into aqueous droplets (Sun et al., 2014). More recently, Wang et al. (2015c) found a smaller contribution of CCOA to OAs (20%) at the same site as in Sun et al. (2014), likely because of the replacement of coal combustion boilers with natural gas ones.

Another POA factor, BBOA, was also widely observed in China, mostly from the burning of crop residues instead of forest fires which are the major sources of BBOA in some western countries. From Q-AMS measurements, Zhang et al. (2014e) found fairly large contributions from BBOA (1–5 μg m^{-3} out of approximately 11 μg m^{-3} of OAs) at Mount Tai in spring, summer, and autumn. They attributed the source of BBOA mainly to the burning of crop residues, but also noted that incense burning in temples might have contributed to the BBOA, echoing the source test showing that BBOA and organics from incense burning have highly similar mass spectral features (Li et al., 2012). At the rural background site of Lin'an in the YRD (Zhang et al., 2015h), Q-AMS measurements also showed that the contribution from BBOA to the total OAs increased substantially from clean days (19%) to haze-fog days (26–28%). At another suburban site (Jiaxing) in the YRD, HR-ToF-AMS measurements showed that BBOA contributed substantially in winter (30% of OAs) but was too scarce to be resolved by PMF in summer (Huang et al., 2013a). BBOA was also found to contribute significantly to the total OAs in the PRD region, with contributions of 24% in Shenzhen (He et al., 2011), 25% in Kaiping (Huang et al., 2011), and 14% in Heshan (Gong et al., 2012). From ACSM measurements, BBOA was ubiquitously identified during harvest seasons in China, for example, in Nanjing and Xianghe (Zhang et al., 2015f; Sun et al., 2016b) and even the Tibetan Plateau (Du et al., 2015). In Nanjing, Zhang et al. (2015f) found that BBOA contributed to 36% and 39% of OAs during summer and autumn harvest seasons, respectively. BBOA on average contributed to 11% of the OA mass in summer at the Xianghe site, yet its contribution increased to 21% during the biomass burning period in late June (Sun et al., 2016b). Furthermore, Zhang et al. (2015f) identified another type of BBOA that was likely from the oxidation of fresh BBOA. The spectrum of the oxidized BBOA showed a similar spectral pattern to those of OOAs, but with peaks visible at *m/z* 60 and 73. Similarly, Du et al. (2015)

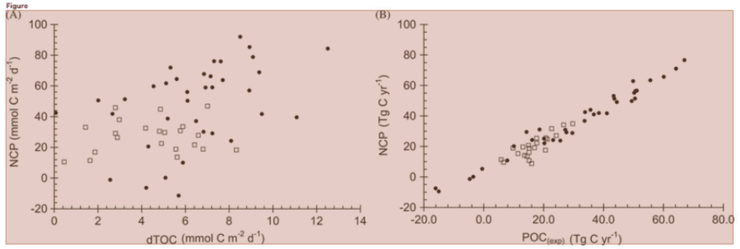


Fig. 7. Relationship between NCP and seasonally accumulated TOC (upper 30 m) (A), and approximated values for exported POC (B) calculated as the difference between NCP and accumulated TOC. 2008 values are indicated by open boxes, while 2009 values are indicated by closed circles. In 2008 and 2009, there was no relationship between NCP and accumulated TOC (2008: $R^2 = 0.12$; 2009: $R^2 = 0.18$). By contrast, $POC_{(exp)}$ and NCP are strongly correlated in both years (2008: $R^2 = 0.70$; 2009: $R^2 = 0.97$).

Table 3
Estimates of calculated export production ($POC_{(exp)}$) at selected stations in Tg C yr^{-1} in 2008 and 2009. NC: Northern Coastal Domain; SC: Southern Coastal Domain; NM: Northern Middle Domain; SM: Southern Middle Domain; SO: Southern Outer Domain. Error listed on domain averages is one standard deviation from the mean. Blank values (–) in 2008 arise from unavailable NCP estimates for those stations. Blank values (–) in 2009 arise from negative NCP estimates.

Station	Domain	2008			2009		
		ΔnTOC μmol kg^{-1}	$POC_{(exp)}$ Tg C yr^{-1}	$POC_{(exp)}$: NCP %	ΔnTOC μmol kg^{-1}	$POC_{(exp)}$ Tg C yr^{-1}	$POC_{(exp)}$: NCP %
SL2	NC	5.3	14.5	107.2	10.9	−16.1	–
SL4	NC	21.9	17.7	79.7	6.7	−4.7	–
MN2	SC	–	–	–	14.0	−3.5	–
MN3	SC	26.2	6.6	69.0	11.9	7.9	72.9
NP1	SC	9.4	11.4	74.7	12.0	−0.5	–
SL7	NM	19.0	15.1	94.1	29.5	14.3	48.5
SL10	NM	9.7	26.7	78.1	25.3	18.6	60.0
SL13	NM	10.0	13.1	66.5	24.3	33.8	79.6
SL14	NM	–	–	–	19.7	35.7	81.0
70M43	NM	18.6	–	–	12.4	41.1	98.5
70M47	NM	18.8	14.9	72.0	19.8	43.6	85.0
70M51	NM	–	–	–	16.1	51.1	90.4
70M55	NM	20.0	29.7	85.3	26.0	49.9	79.4
MN5	SM	5.8	13.7	97.2	22.1	10.1	50.0
MN7	SM	1.4	16.0	182.9	21.0	43.4	81.7
NP4	SM	16.1	17.7	69.5	14.6	16.2	66.7
NP6	SM	–	–	–	14.7	55.6	87.8
NP8	SM	–	–	–	16.8	67.0	87.5
NP10	SM	10.1	22.7	71.7	12.3	38.5	91.8
70M1	SM	–	–	–	13.3	27.2	88.1
70M3	SM	16.1	5.8	51.3	14.4	50.2	91.1
70M5	SM	–	–	–	13.9	44.4	90.5
70M9	SM	–	–	–	11.6	−15.0	–
70M13	SM	14.0	15.2	80.9	18.3	64.2	90.4
70M17	SM	–	–	–	18.6	60.1	91.6
70M25	SM	14.4	21.2	85.6	13.8	50.7	89.9
70M29	SM	11.9	24.2	89.5	14.0	20.1	80.2
70M35	SM	–	–	–	9.3	49.4	99.3
70M39	SM	–	–	–	10.5	50.5	98.3
MN11	SO	19.7	15.2	141.8	15.3	33.7	91.7
MN13	SO	20.2	20.6	117.4	5.4	29.6	102.8
MN15	SO	20.8	16.9	88.7	7.2	25.6	107.3
MN18	SO	17.2	20.7	81.3	14.0	20.3	91.9
MN20	SO	–	–	–	8.7	27.6	94.4
NP12	SO	4.9	9.8	51.9	10.4	36.8	89.7
NP15	SO	–	–	–	0.1	23.2	95.9
AVG	NC			93 ± 19			–
AVG	SC			72 ± 4			–
AVG	NM			79 ± 11			78 ± 16
AVG	SM			91 ± 40			86 ± 12
AVG	SO			96 ± 35			96 ± 6

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

Knowledge extraction from high-variety input

Input: Raw High-Variety Inputs

Table 2. List of uncharged organic compounds used for calculation of the relationship ΔK_d versus log K_{ow} from Fig. 3

Compound	log K_{ow}
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-Pentanol	1.25

Parameter

Context

Context Values

Values

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

Equation

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

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 $K_{ow} < -0.5$ was used (Table 3). The range of log K_{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

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

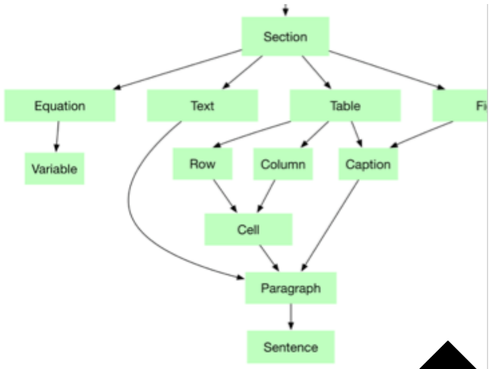
COSMOS extracts knowledge from **multi-modal unstructured data**
(text, tables, images, equations, diagrams)

Knowledge as a service in COSMOS

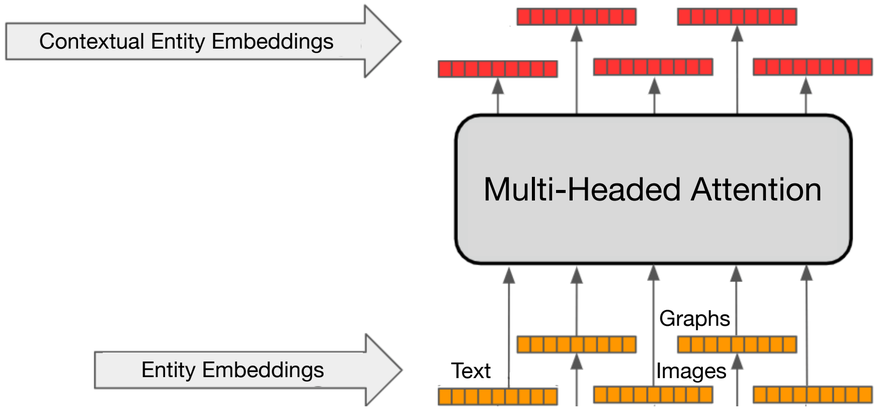
From the intermediate XML representation to knowledge bases and open-domain extraction

Semi-structured data representation

```
<!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>
```

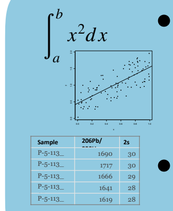


Contextual data representation

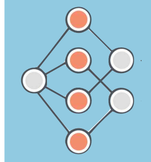


- **Retrieve-and-Read Q&A**
- **On-demand Knowledge Bases**
- **Model Evaluation**
- **Dataset Generation**

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

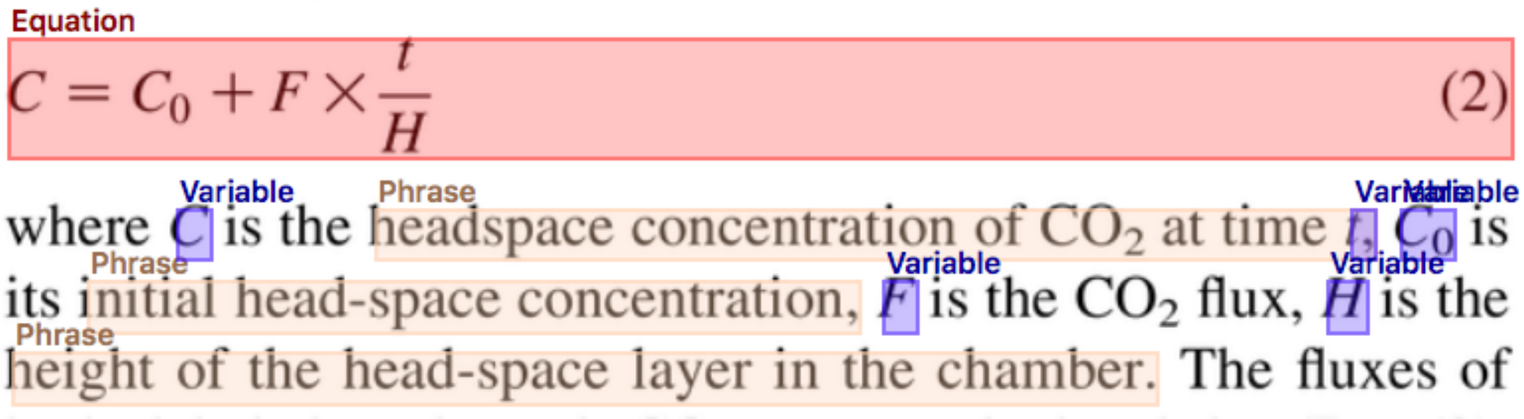


- Entity relations across tables, figures, equations
- Discover knowledge in millions of papers



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

From Semi-structured data to Equation - Variable Models



- 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).

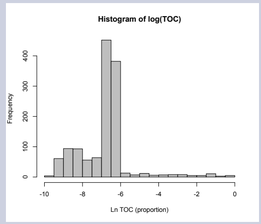
Open-domain retrieve and read interaction

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.

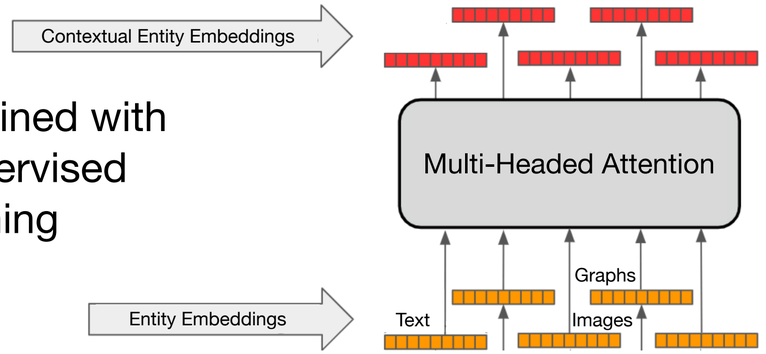
Empirical Distribution



Entity-Aware Code Analysis

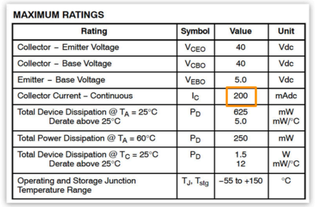
Fine-tuned for Q&A

Model: trained with self-supervised learning



Input

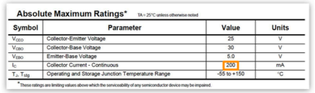
Annotated examples on PDF



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

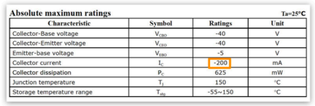


Absolute Maximum Ratings*

Symbol	Parameter	Value	Units
V_{CEO}	Collector-Emitter Voltage	45	V
V_{CBO}	Collector-Base Voltage	30	V
V_{EBO}	Emitter-Base Voltage	5.0	V
I_C	Collector Current - Continuous	200	mA
T_J, T_{stg}	Operating and Storage Junction Temperature Range	-55 to +150	°C

* These ratings are limiting values above which the serviceability of any semiconductor device may be impaired.

MMBT3904.PDF



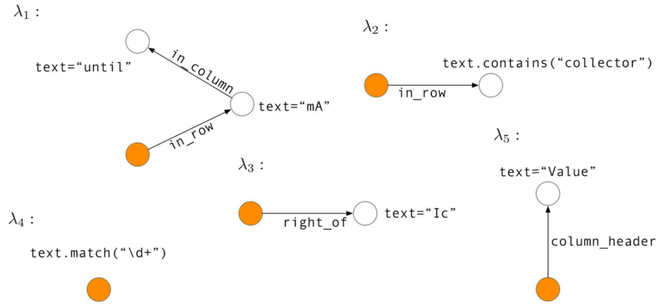
Absolute maximum ratings Ta=25°C

Characteristic	Symbol	Ratings	Unit
Collector-Base voltage	V_{CBO}	-40	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

Output

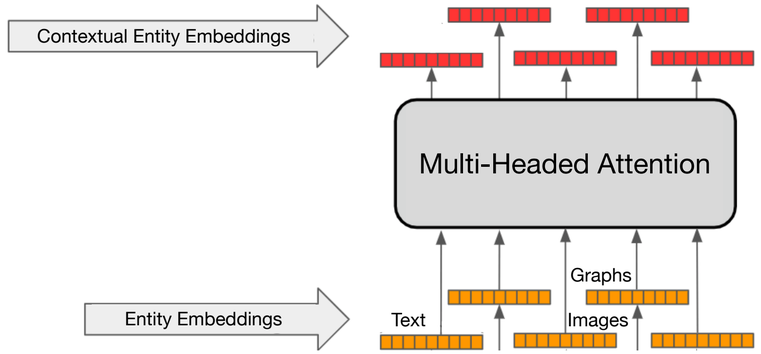
Ensemble of programs synthesized from the input examples



Task

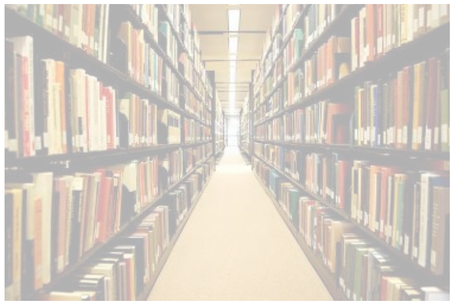
Extract the value of "Collector Current" for each model of transistor

Fine-tuned for knowledge base construction

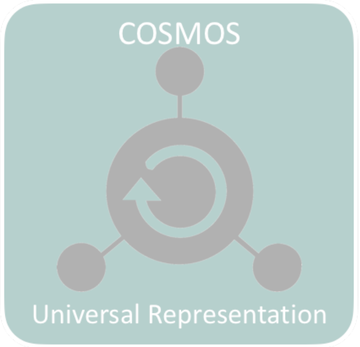


Outline

1. xDD: A portal to scientific publications and HTC



2. COSMOS: Knowledge extraction as a service



3. Demo: analyzing model code with COSMOS