COSMOS:
a data and AI platform for cross-disciplinary scientific discovery and model curation

Theodoros Rekatsinas, Shanan Peters, Miron Livny

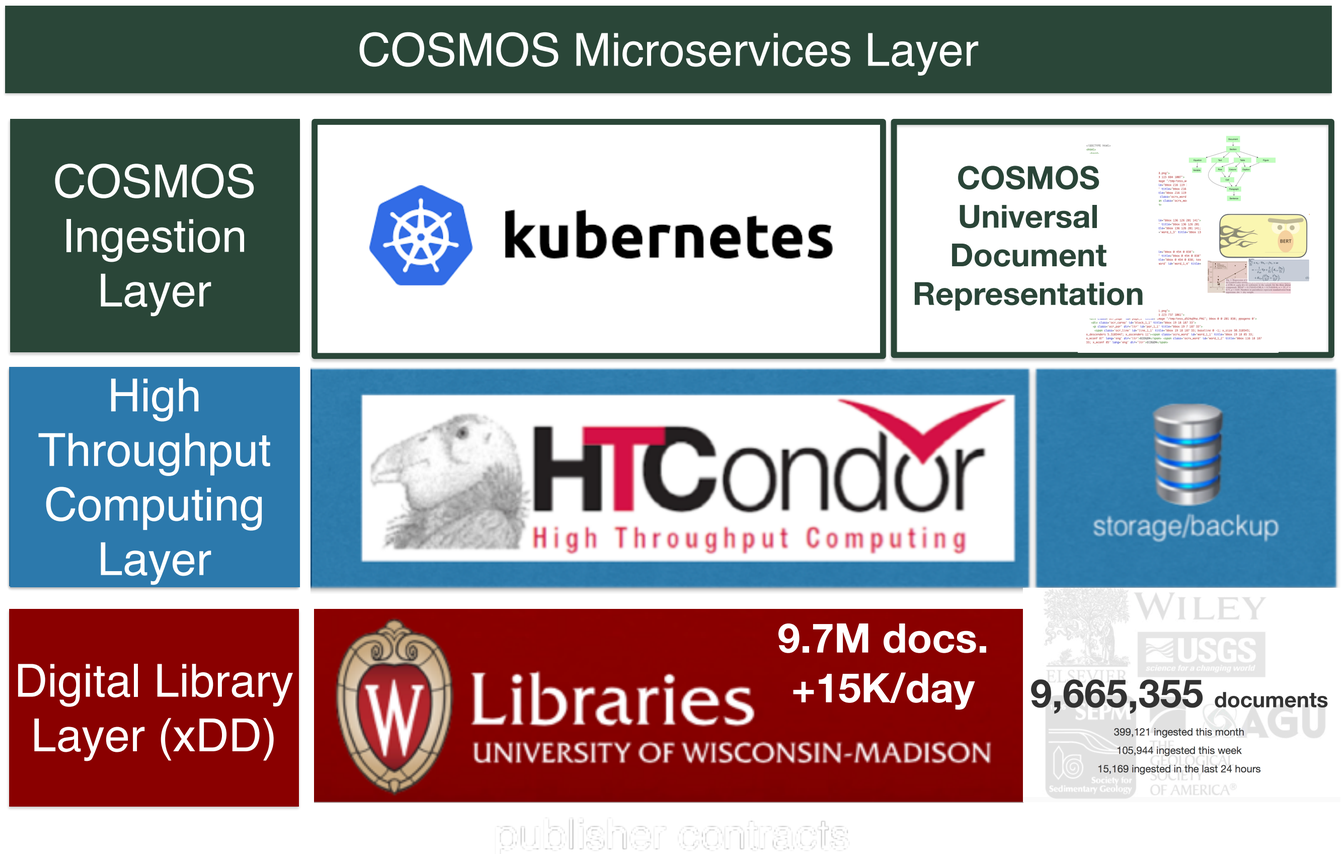
WISCONSIN
UNIVERSITY OF WISCONSIN–MADISON

HTCondor
High Throughput Computing

COSMOS is an end-to-end data, computing, and AI platform for leveraging scientific publications

COSMOS Microservices

Retrieve-and-read Q&A | **On-demand Knowledge Bases** | **Model Evaluation** | **Dataset Generation** | **Your Service**



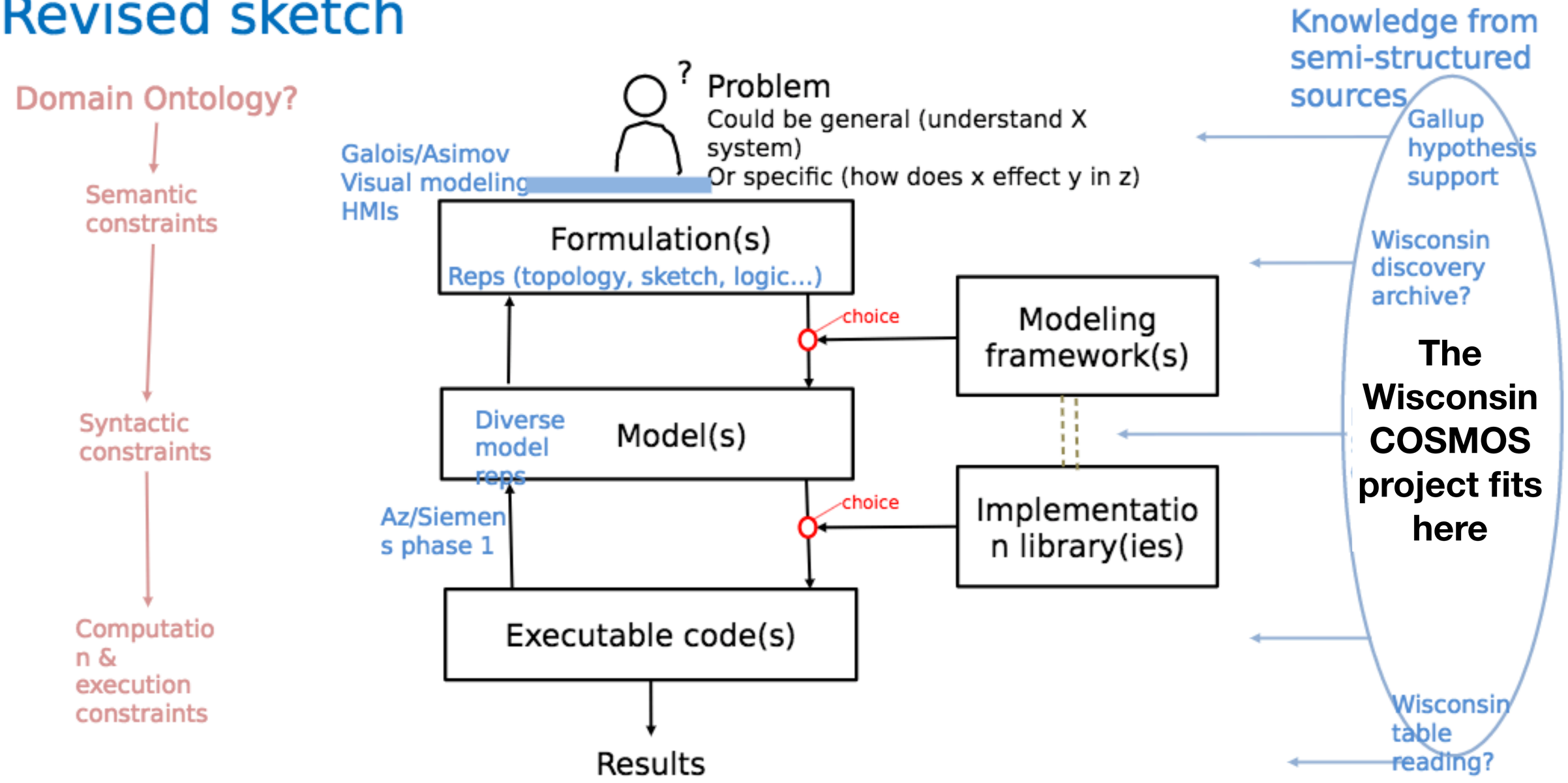
Key Requirements:

- Principled access to new and archival publications that span scientific domains and publishers
- Supporting HTC infrastructure
- Universal representation of published knowledge
- Ecosystem of lightweight, scalable services for scientific discovery

Without first three, the scope of scientific impact is limited!

Where does COSMOS belong?

Revised sketch



Short-term Goal for COSMOS:
Build a unified platform for Knowledge Extraction and Integration (1)

Input: documents that contain unstructured and structured data

Data formats:

- PDFs containing text, tables, equations, figures [priority: now supported]
- HTML documents (e.g., Wikipedia Pages) [coming soon:]
- Source Code [to be added during Phase 2]

Initial Output: an intermediate XML representation

- This is done for standardization purposes (we need a common representation for PDFs and HTML)
- We represent each document in XML as a sequence of text-boxes, equations, figures, figure captions, tables, table captions, etc.; the goal is an XML sequence that captures the natural layout of the input document
- Each XML node has meta-data to reconstruct the original document: bounding box, class type (i.e., text, equation, table, figure, etc), a cropped scanned image (if PDF), OCR extractions from the cropped image, or raw input for HTML
- This intermediate representation is stored in MongoDB and is our starting point for downstream knowledge extraction tasks.

Example XML representation

```
<div class="rawtext">A commercial sample (SSA-1 from Halliburton,
Celle, Germany) containing (wt %) quartz 97.60, CaO
0,57, MgO 0.18, A1203 0&#8216;17, and TiOz 0.06 (deter-
mined by X-ray &#64258;uorescence analysis) was used
Its specific surface area (Blaine method) was 1857
cmZ/g. Its average particle size (L150 value) was 32.7
um. Specific density of the silica flour was found at
2.65 kg/Ll</div>
</div>
<div class="Equation" id="Equation8">
<img src="img/59c42882cf58f1b82a0f37e0.pdf_2/Equation8.png"/>
<div class="hocr" data-coordinates="746 1697 1116 1719" data-score="0.5406793355941772">
<div class="ocr_page" id="page_1" title="image &quot;/tmp/tess_yjh4672i.PNG&quot;; bbox 0 0 374 26; ppageno 0">
<div class="ocr_carea" id="block_1_1" title="bbox 3 3 373 25">
<p class="ocr_par" dir="ltr" id="par_1_1" title="bbox 3 3 373 25">
<span class="ocr_line" id="line_1_1" title="bbox 3 3 373 25; baseline 0 0; x_size 23.279999; x_descenders 5.2799997; x_ascenders 6"><span class="ocrx_word
" dir="ltr" id="word_1_1" lang="eng" title="bbox 3 3 228 25; x_wconf 61">CaAMPS&#174;-co-itaconic</span> <span class="ocrx_word" dir="ltr" id="word_1_2"
lang="eng" title="bbox 237 7 279 25; x_wconf 82">acid</span> <span class="ocrx_word" dir="ltr" id="word_1_3" lang="eng" title="bbox 289 7 373 25; x_wconf
77">retarder</span>
```

The snippet shows two elements from a raw PDF document: (1) a raw text part after OCR is performed and (2) an equation element associated with the corresponding cropped image as well as elements associated with the equation.

Short-term Goal for COSMOS:
Build a unified platform for Knowledge Extraction and Integration (2)

The biggest challenge phased during Phase 1

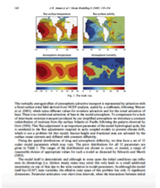
How to convert PDFs (with text, tables, equations, figures) to our intermediate XML representation?

Mutli-page raw PDF

Document Preprocessing Tool

Page 1

Page 2



Page k

Images corresponding to PDF pages

ML Model

ML Model

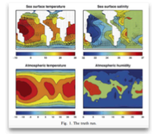
ML Model

Equations

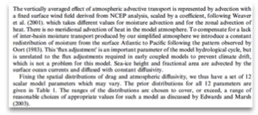
$$P(\mathbf{X}|\mathbf{\Phi}) = \prod \frac{1}{\sqrt{2\pi\sigma_i}} e^{-\frac{1}{2}\frac{(x_i - \hat{x}_i)^2}{\sigma_i^2}} = Ke^{-\frac{1}{2}\sum_i \frac{(x_i - \hat{x}_i)^2}{\sigma_i^2}}$$

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

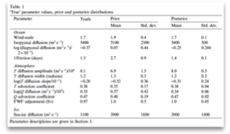
Figure



Text



Table



Visual Extractions of Elements of Interest

OCR Engine

OCR Engine

OCR Engine

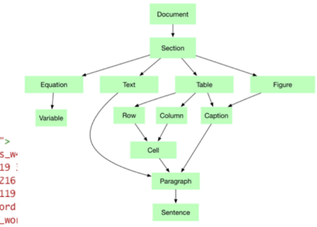
OCR Engine

OCR Engine

OCR Enrichment

Element Aggregator

Representation Compiler



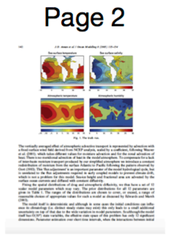

```
<!DOCTYPE html>
<html>
  <head>
    <title>EcoGEnIE.pdf-0007</title>
  </head>
  <body>
    <div class="Figure" id="Figure0">
      <img src='img/EcoGEnIE.pdf-0007/Figure0.png'>
      <div class="hocr" data-coordinates="419 115 884 1007">
  <div class='ocr_page' id='page_1' title='image "/tmp/tess_w
   <div class='ocrx_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='ocrx_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='ocrx_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='ocrx_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

From PDF to XML

Mutli-page raw PDF

Document Preprocessing Tool

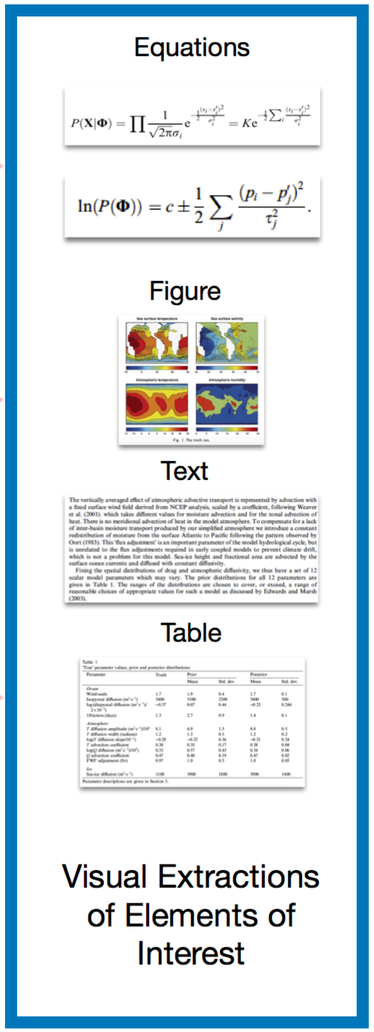


Images corresponding to PDF pages

ML Model

ML Model

ML Model



Visual Extractions of Elements of Interest

OCR Engine

OCR Engine

OCR Engine

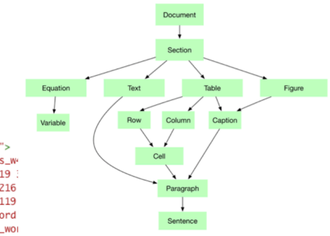
OCR Engine

OCR Engine

Element Aggregator

Representation Compiler

OCR Enrichment



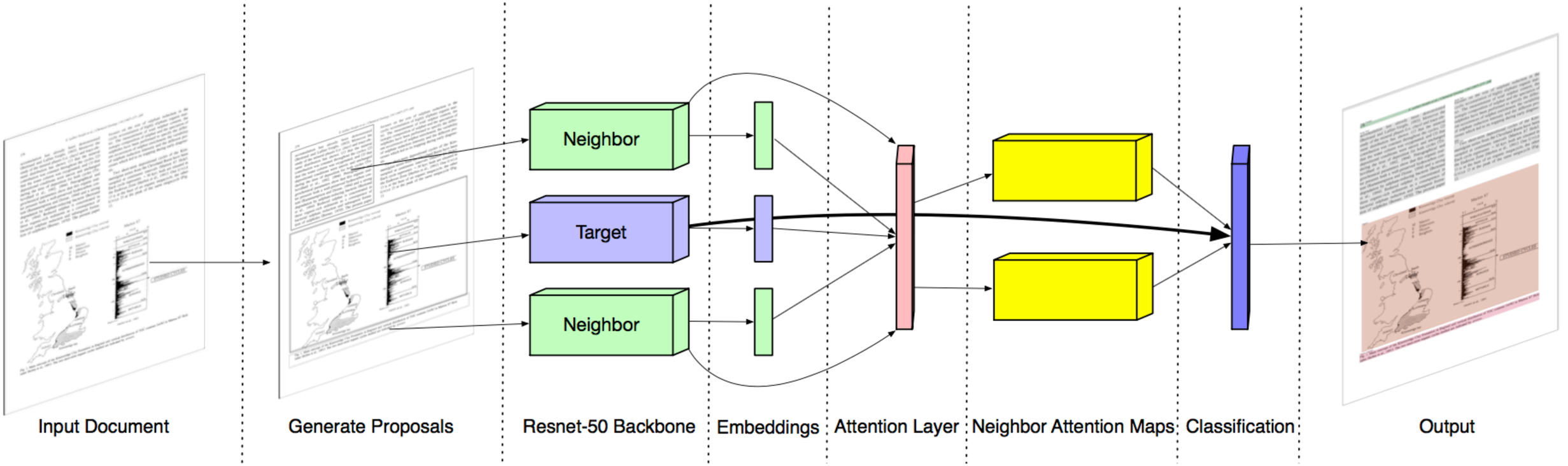

```
<!DOCTYPE html>
<html>
  <head>
    <title>EcoGEnIE.pdf-0007</title>
  </head>
  <body>
    <div class="Figure" id="Figure0">
      <img src='img/EcoGEnIE.pdf-0007/Figure0.png'>
      <div class="hocr" data-coordinates="419 115 884 1007">
  <div class='ocr_page' id='page_1' title='image "/tmp/tess_w
    <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

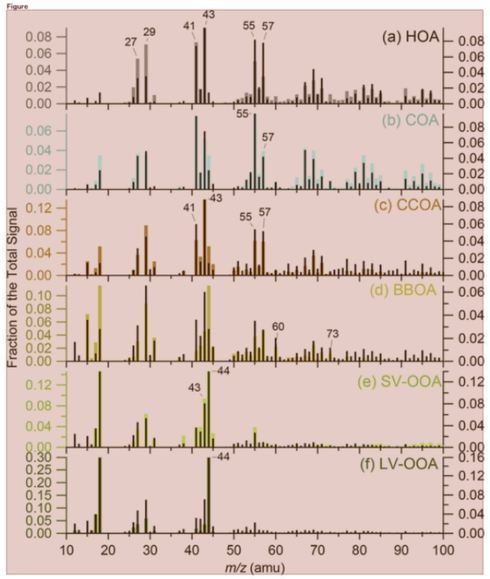


Fig. 10. Typical mass spectral profiles of OA factors resolved from PMF-ACSM at various sites in China (HOA, BBOA, SV-OOA and LV-OOA were from Sun et al. (2016b), COA was from Sun et al. (2014) and CCOA was from Sun et al. (2013c). The standard mass spectra of OA factors reported in Ng et al. (2011b)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

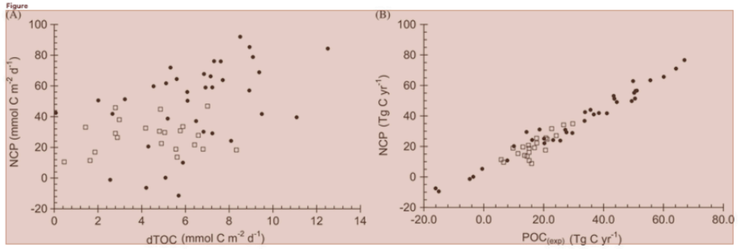


Fig. 7. Relationship between NCP and seasonally accumulated TOC (upper 30 m) (A), and approximated values for exported POC (B) calculated as the difference between NCP and accumulated TOC. 2008 values are indicated by open boxes, while 2009 values are indicated by closed circles. In 2008 and 2009, there 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)

Short-term Goal for COSMOS:
Build a unified platform for Knowledge Extraction and Integration (3)

Knowledge bases constructed from our XML representation [Phase 1 approach]

Step 1: We use the open-IE service from CoreNLP to identify all entities mentioned in different document elements

Step 2: We leverage co-reference of OCR extractions and proximity features (in the document space) to identify relations between elements (e.g., a table, an equation, a figure) and the entities identified in Step 1. We construct the following knowledge bases:

(1) Table to Entities (schema in database table <Table ID, Entity Name>)

This database relation associates a table element (for Phase 1 this is the entire extracted table) to an entity. The scope of relevant entities is currently limited to the caption [*During Phase 2 we will extend it to the entire document-We are designing a new document embedding that will allow us to answer such relatedness queries.*]

(2) Figure to Entities (schema in database table <Figure ID, Entity Name>)

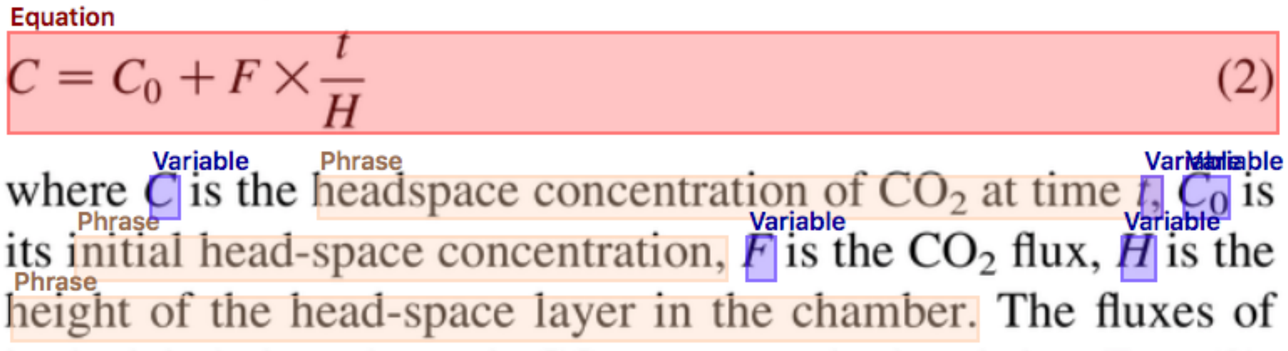
This database relation associates a figure (for Phase 1 this is the entire extracted figure) to an entity. The scope of relevant entities is currently limited to the caption [*During Phase 2 we will extend it to the entire document-We are designing a new document embedding that will allow us to answer such relatedness queries.*]

(3) Equation to Variables and Variable Definitions (schema in database tables <Equation ID, Variable Symbol, Natural Language Description>, <Equation ID, Parsed Equation Tree (representation to Latex)>)

This database relation associates each extracted equation with variables (symbols) that participate in the Equation. To this, we use the parsed equation tree (img2latex) and to identify all candidate symbols related to the equation. We then consider the text that surrounds the equation and match the symbols from the equation's parse tree to tokens in the surrounding text. Given the matched tokens we identify their natural language description by solving a relation extraction problem from the sentences that include them. The relations we extract associate a symbol token to a "description" text span (see example in next slide)

Short-term Goal for COSMOS:
Build a unified platform for Knowledge Extraction and Integration (4)

Example instance of the Equation - Variable - Description knowledge base

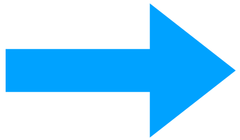


- Based on the parse tree of the extracted equation (red/pink-colored bounding box) we now that it contains symbols **C**, **t, Co, F,** and **H**
- These symbols are linked to the purple tokens in the text below this equation (these tokens are identified to be of type Variable). We have an Equation to Variable relationship here.
- These symbols are linked to descriptions that are formed by considered the output of Open-IE (using CoreNLP). The phrases correspond to definitions of these symbols. Here we form links between the Variable tokens and the phrase tokens.
- As shown our method can be improved (especially recall; e.g., F here not associated with CO2 flux).

The next short-term steps for COSMOS

Fine-grained Extraction of Table Elements

Parameter	Symbol	Value	Units
Nutrient quotas			
Minimum phosphate:carbon quota	$Q_{\mathrm{P}}^{\mathrm{min}}$	2.1×10^{-3}	mmol P (mmol C)$^{-1}$
Maximum phosphate:carbon quota	$Q_{\mathrm{P}}^{\mathrm{max}}$	1.1×10^{-2}	mmol P (mmol C)$^{-1}$
Minimum iron:carbon quota	$Q_{\mathrm{Fe}}^{\mathrm{min}}$	1.0×10^{-6}	mmol Fe (mmol C)$^{-1}$
Maximum iron:carbon quota	$Q_{\mathrm{Fe}}^{\mathrm{max}}$	4.0×10^{-6}	mmol Fe (mmol C)$^{-1}$
Temperature			
Reference temperature	T_{ref}	20	°C
Temperature dependence	A	0.05	-
Photosynthesis			
Maximum Chl-*a*-to-phosphorus ratio	$\theta_{\mathrm{N}}^{\mathrm{max}}$	48	mg Chl *a* (mmol P)$^{-1}$
Initial slope of P-I curve	α	3.83×10^{-7}	mmol C (mg Chl *a*)$^{-1}$(μEin m^{-2})$^{-1}$
Cost of biosynthesis	ξ	37.28	mmol C (mmol P)$^{-1}$
Grazing			
Optimum predator:prey length ratio	ϑ_{opt}	10	-
Geometric s.d. of ϑ	σ_{graz}	2.0	-
Total prey half-saturation	$k_{\mathrm{C}}^{\mathrm{prey}}$	5.0	mmol C m^{-3}
Maximum assimilation efficiency	λ^{max}	0.7	-
Grazing refuge parameter	Λ	-1	(mmol C m^{-3})$^{-1}$
Active switching parameter	s	2	-
Assimilation shape parameter	h	0.1	-
Other loss terms			
Plankton mortality	m	0.05	d^{-1}
Plankton respiration	$r_{i_{\mathrm{b}}=\mathrm{DIC}}$	0.05	d^{-1}
	$r_{i_{\mathrm{b}}\neq\mathrm{DIC}}$	0	d^{-1}
Partitioning of organic matter			
Fraction to DOM	β	0.66	-
Light attenuation			
Light attenuation by water	k_{w}	0.04	m^{-1}
Light attenuation by chlorophyll	k_{Chl}	0.03	m^{-1}(mg Chl)$^{-1}$



	0	1	2	3	4
0	b'Table4Size-independentmodelparame'	b'ters'	b''	b''	b''
1	b'Parameter'	b'Symbol'	b''	b'Value'	b'Units'
2	b'Nutrient quotas'	b''	b''	b''	b''
3	b'Minimum phosphate:carbon quota'	b'Q<s>min</s>'	b''	b'2.1\xc3\x9710\xe2\x88\x92<s>3</s>'	b'mmolP(mmolC)\xe2\x88\x92<s>1</s>'
4	b''	b'P'	b''	b''	b''
5	b'Maximum phosphate:carbon quota'	b'Q<s>max</s>'	b''	b'1.1\xc3\x9710\xe2\x88\x92<s>2</s>'	b'mmolP(mmolC)\xe2\x88\x92<s>1</s>'
6	b''	b'P'	b''	b''	b''
7	b'Minimum iron:carbon quota'	b'Q<s>min</s>'	b''	b'1.0\xc3\x9710\xe2\x88\x92<s>6</s>'	b'mmolFe(mmolC)\xe2\x88\x92<s>1</s>'
8	b''	b'Fe'	b''	b''	b''
9	b'Maximum iron:carbon quota'	b'Q<s>max</s>'	b''	b'4.0\xc3\x9710\xe2\x88\x92<s>6</s>'	b'mmolFe(mmolC)\xe2\x88\x92<s>1</s>'
10	b''	b'Fe'	b''	b''	b''
11	b'Temperature'	b''	b''	b''	b''
12	b'Reference temperature'	b'T<s>ref</s>'	b''	b'20'	b'\xe2\x97\xa6<s>C</s>'
13	b'Temperature dependence'	b'A'	b''	b'0.05'	b'-'
14	b'Photosynthesis'	b''	b''	b''	b''
15	b'MaximumChl-<s>a</s>-to-phosphorusratio'	b'\xce\xb8<s>max</s>'	b''	b'48'	b'mgChla(mmolP)\xe2\x88\x92<s>1</s>'
16	b''	b'N'	b''	b''	b''
17	b'Initial slope of P-I curve'	b'\xce\xb1'	b''	b'3.83\xc3\x9710\xe2\x88\x92<s>7</s>'	b'mmolC(mgChla)\xe2\x88\x92<s>1</s>(\xc2\xb5Ei...
18	b'Cost of biosynthesis'	b'\xce\xbe'	b''	b'37.28'	b'mmolC(mmolP)\xe2\x88\x92<s>1</s>'
19	b'Grazing'	b''	b''	b''	b''
20	b'Optimum predator:prey length ratio'	b'\xcf\x91<s>opt</s>'	b''	b'10'	b'-'
21	b'Geometrics.d.of<s>\xcf\x91</s>'	b'\xcf\x83<s>graz</s>'	b''	b'2.0'	b'-'
22	b'Total prey half-saturation'	b'k<s>prey</s>'	b''	b'5.0'	b'mmolCm\xe2\x88\x92<s>3</s>'
23	b''	b'C'	b''	b''	b''
24	b'Maximum assimilation ef\xef\xac\x81ciency'	b'\xce\xbb<s>max</s>'	b''	b'0.7'	b'-'
25	b'Grazing refuge parameter'	b'\xce\x9b'	b''	b'-1'	b'(mmolCm\xe2\x88\x92<s>3</s>)\xe2\x88\x92<s>1...
26	b'Active switching parameter'	b's'	b''	b'2'	b'-'
27	b'Assimilation shape parameter'	b'h'	b''	b'0.1'	b'-'
28	b'Other loss terms'	b''	b''	b''	b''
29	b'Plankton mortality'	b'm'	b''	b'0.05'	b'd\xe2\x88\x92<s>1</s>'
30	b'Plankton respiration'	b'ri<s>b</s>=DIC'	b''	b'0.05'	b'd\xe2\x88\x92<s>1</s>'
31	b''	b''	b''	b'0'	b'd\xe2\x88\x92<s>1</s>'
32	b''	b'ri<s>b</s>(cid:54)=DIC'	b''	b''	b''
33	b'Partitioning of organic matter'	b''	b''	b''	b''
34	b'Fraction to DOM'	b'\xce\xb2'	b''	b'0.66'	b'-'
35	b'Light attenuation'	b''	b''	b''	b''
36	b'Light attenuation by water'	b'k<s>w</s>'	b''	b'0.04'	b'm\xe2\x88\x92<s>1</s>'
37	b'Light attenuation by chlorophyll'	b'k<s>Chl</s>'	b''	b'0.03'	b'm\xe2\x88\x92<s>1</s>(mgChl)\xe2\x88\x92<s>1...
38	b''	b''	b'20'	b''	b''

- We already have a running prototype to convert tables to HTML (an example extraction is shown above).
- The output will be incorporated to our XML representation and our knowledge bases will be enhanced to capture fine-grained quantity-to-value relations.

Fine-grained Extraction of Figure Elements

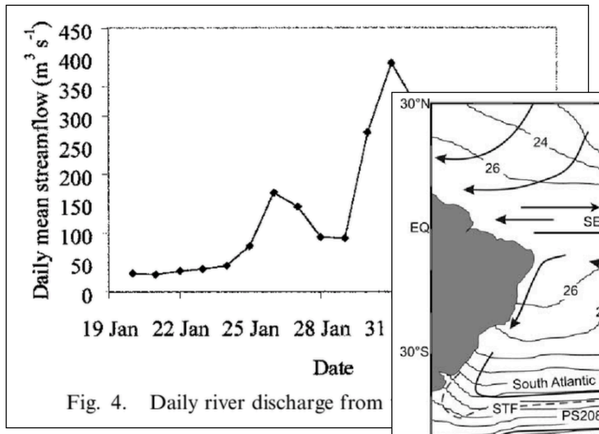
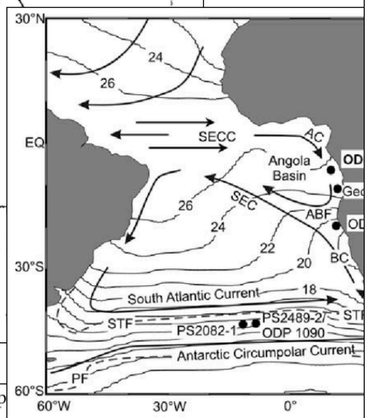


Fig. 4. Daily river discharge from



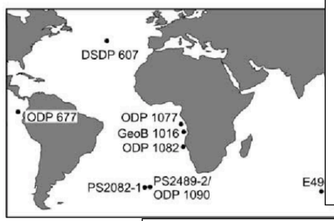


Figure F4. Photomicrograph of biogenic op
9H-4, 50 cm).

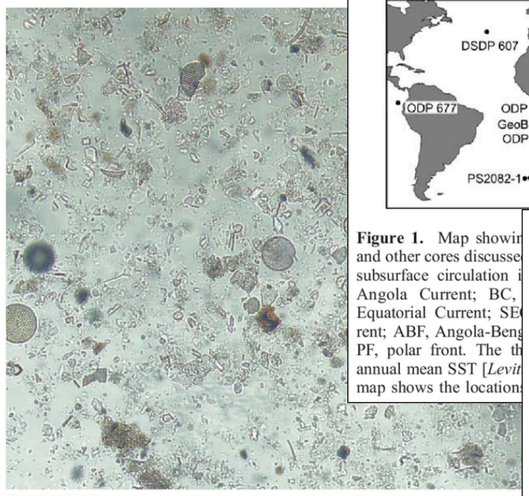


Figure 1. Map showir
and other cores discusse
subsurface circulation i
Angola Current; BC,
Equatorial Current; SE
rent; ABF, Angola-Beng
PF, polar front. The th
annual mean SST [*Levit*
map shows the location

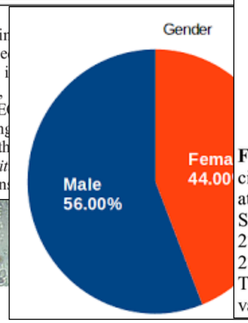


Figure F44. CWL acoustic logging data from Hole 1244E. The resistivity log is used as a reference indicator of the presence of gas hydrate. Low waveform amplitudes are also possible indicators of gas hydrate. SPh. foc. res. = spherically focused resistivity.

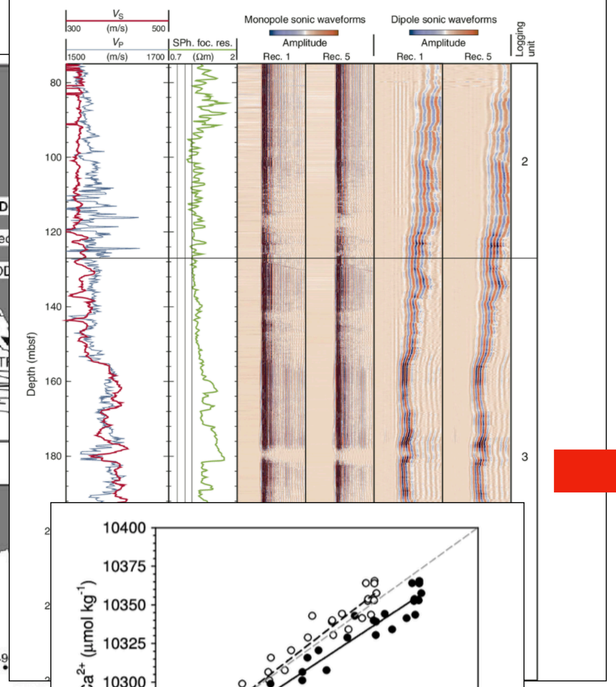
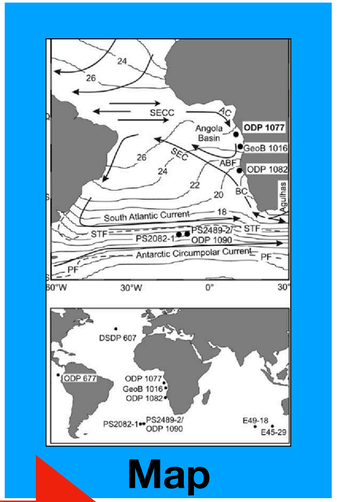
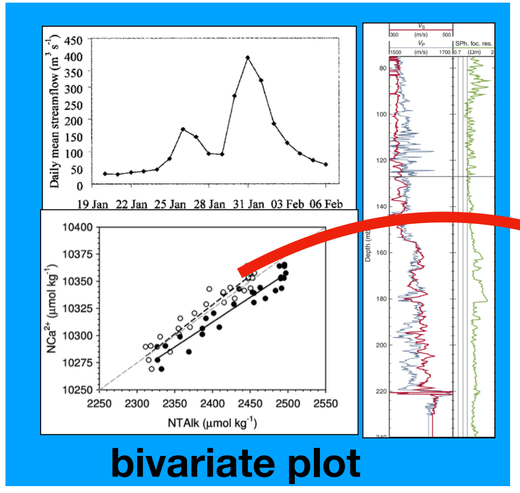


Figure 8. Ca^{2+} versus TAlk (open circles) or PTA (solid circles; PTA = TAlk + NO_3 + PO_4) in waters below 100 m at stations SEATS, K404, and BY01 in the South China Sea. Ca^{2+} = (0.54 ± 0.03)TAlk + (9038 ± 77), r = 0.96, n = 27; Ca^{2+} = (0.46 ± 0.03)PTA + (9214 ± 64), r = 0.96, n = 27. All of the data were normalized to a salinity of 35. The slope of the gray dashed line is equal to the theoretical value of 0.5.

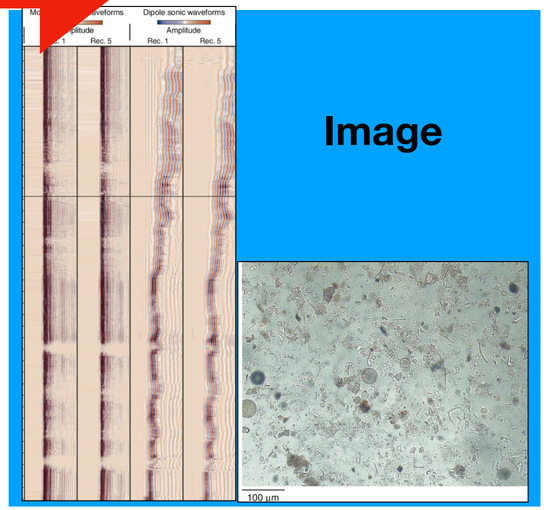


Map

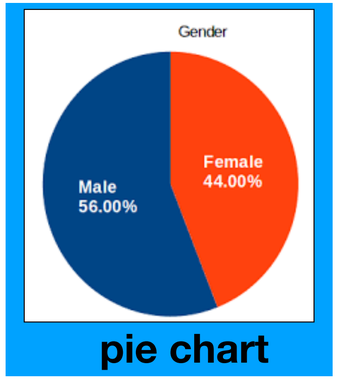


bivariate plot

x=2400
y=10300
units: $\frac{\mu mol}{kg}$



Image



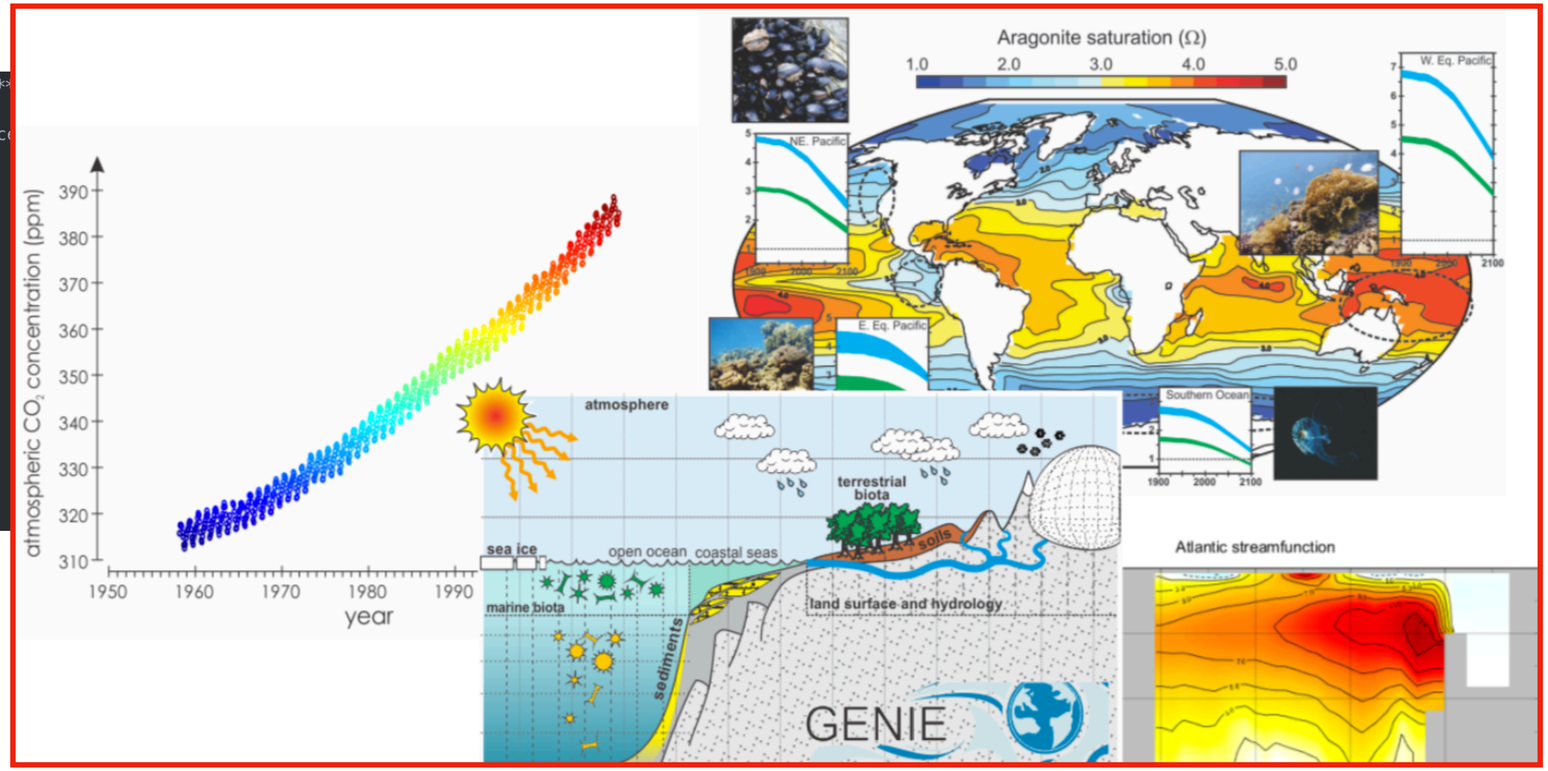
pie chart

- Classify extracted figure elements (i.e., id pie charts, histograms, images, etc).
- Train a CNN-based parser to extract measurements out of different figure classes
- Associate these measurements with different entities (e.g., units, variables, text entities)

Connecting scientific model code to metadata, data, and phenomena in scientific publications

```
! ******************************************************************************************
!
! CALCULATE BIOLOGICAL TRACER UPTAKE AT THE SURFACE OCEAN -- biologically induced (mass balanc
! NOTE: assume complete homogenization over mixed layer
!       => calculate uptake based on tracer concentrations in surface layer only
SUBROUTINE sub_calc_bio_uptake(dum_i,dum_j,dum_k1,dum_dt)
  ! dummy arguments
  INTEGER,INTENT(in)::dum_i,dum_j
  INTEGER,INTENT(in)::dum_k1
  real,intent(in)::dum_dt
  ! local variables
  INTEGER::k,l,io,is
  integer::loc_i,loc_tot_i
  real::loc_dPO4
  real::loc_dPO4_1,loc_dPO4_2
  real::loc_dPO4_sp,loc_dPO4_nsp
  real::loc_ohm,loc_co3                                                    !
  real::loc_frac_N2fix
```

```
! *** INITIALIZE VARIABLES ***
!
loc_dPO4 = 0.0
loc_delta_Corg = 0.0
loc_delta_CaCO3 = 0.0
!
loc_kPO4 = 0.0
loc_kPO4_sp = 0.0
loc_kPO4_nsp = 0.0
loc_kFe = 0.0
loc_kFe_sp  = 0.0
loc_kFe_nsp  = 0.0
loc_kSiO2 = 0.0
loc_kSiO2_sp = 0.0
```



- Parse cGENIE Fortran model code to identify equations, variables, constants, comments; locate and extract entities
- Parameterize model run for key times in Earth history (with UC Berkeley collaborators)
- Locate and extract relevant field data from scientific publications for model assessment and testing and revision

The next AI Explorations for COSMOS (going beyond ASKE)

Exploration 1: Self-supervised learning

1 - Semi-supervised training on large amounts of text (books, wikipedia..etc).

The model is trained on a certain task that enables it to grasp patterns in language. By the end of the training process, BERT has language-processing abilities capable of empowering many models we later need to build and train in a supervised way.

Semi-supervised Learning Step

Model:

Dataset:

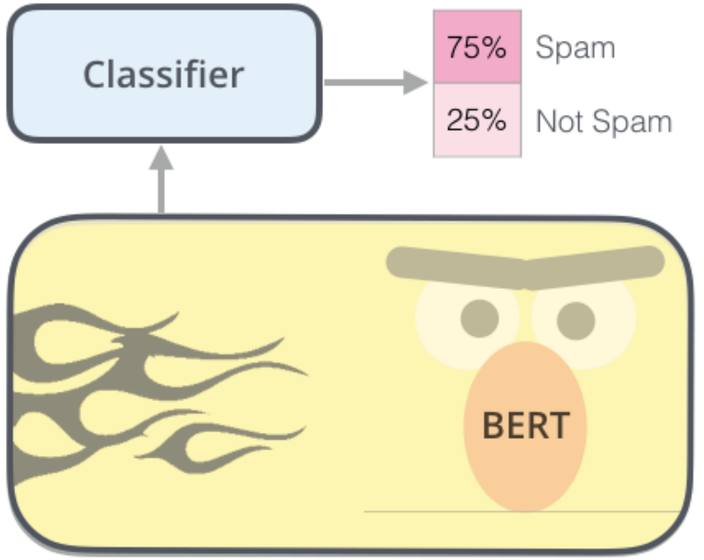
Objective:

Predict the masked word (langauge modeling)

2 - Supervised training on a specific task with a labeled dataset.

Supervised Learning Step

Model: (pre-trained in step #1)



Dataset:

Email message	Class
Buy these pills	Spam
Win cash prizes	Spam
Dear Mr. Atreides, please find attached…	Not Spam

Exploration 1: Self-supervised learning

1 - Semi-supervised training on large amounts of text (books, wikipedia..etc).

The model is trained on a certain task that enables it to grasp patterns in language. By the end of the training process, BERT has language-processing abilities capable of empowering many models we later need to build and train in a supervised way.

Semi-supervised Learning Step

Model:

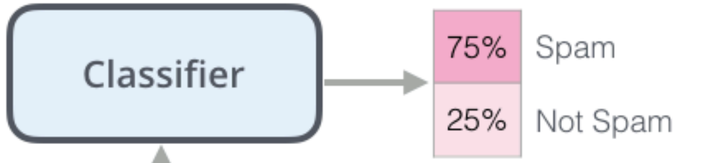
Dataset:

Objective: Predict the masked word (langauge modeling)

2 - Supervised training on a specific task with a labeled dataset.

Supervised Learning Step

Model:



Dataset:

Email message	Class
Buy these pills	Spam
Win cash prizes	Spam
Dear Mr. Atreides, please find attached…	Not Spam

BERT gives you a common self-supervised representation for diverse tasks

Exploration 1: Self-supervised learning

1 - Semi-supervised training on large amounts of text (books, wikipedia..etc).

The model is trained on a certain task that enables it to grasp patterns in language. By the end of the training process, BERT has language-processing abilities capable of empowering many models we later need to build and train in a supervised way.

Semi-supervised Learning Step

Model:

Dataset:

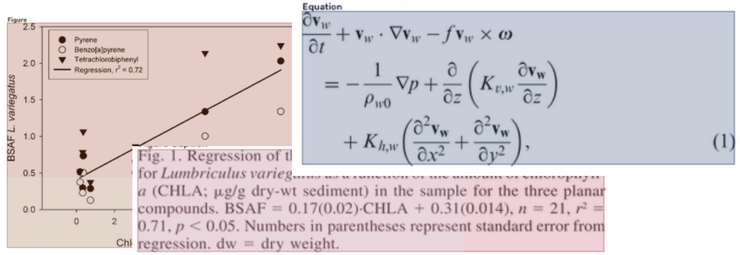


Fig. 1. Regression of t
for *Lumbriculus variegatus* as a function of the amount of chlorophyll *a* (CHLA; μg/g dry-wt sediment) in the sample for the three planar compounds. BSAF = 0.17(0.02)·CHLA + 0.31(0.014), $n = 21$, $r^2 = 0.71$, $p < 0.05$. Numbers in parentheses represent standard error from regression. dw = dry weight.

Equation

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

Objective:

We need a new contrastive learning objective! Use coreferences from Phase 1 representation.

2 - Supervised training on a specifi labeled dataset.

Model:
(pre-trained in step #1)

Q: Mercury , the planet nearest to the Sun , has **extreme surface temperatures** , ranging from 465 C in sunlight to -180 C in darkness . Why is there such a large range of temperatures on Mercury?

C1: The planet is *too small* to hold heat.

...

C4: The planet *lacks an atmosphere* to hold heat .

⇩ **Query1 = Q+C1**

Query1: Mercury , the planet nearest to the Sun , has extreme surface temperatures , ranging from 465 C in sunlight to -180 C in darkness . Why is there such a large range of temperatures on Mercury? The planet is too small to hold heat.

⇩ **Retrieving evidence**

S1: Other planets such as Mercury has **extreme** hot and cold **temperatures** .
S2: The planet Mercury is *too small* and has too little gravity to hold onto an atmosphere.
...

⇩ **Query4 = Essential-term(Q)+C4**

Query4: Mercury extreme surface temperatures. The planet lacks an atmosphere to hold heat .

⇩ **Retrieving evidence**

S1: The *lack of atmosphere* also contributes to the planet 's wild **temperature extremes** .
S2: Mercury is the closest planet to the sun and has a *thin atmosphere*, no air pressure and an **extremely** high **temperature**.
...

...

⇩ **Sending evidence to reader**

MRC

Open-domain retrieve-and-read interaction.
Essentially: question answering over complex scientific domains.

Exploration 1: Self-supervised learning

1 - Semi-supervised training on large amounts of text (books, wikipedia..etc).

The model is trained on a certain task that enables it to grasp patterns in language. By the end of the training process, BERT has language-processing abilities capable of empowering many models we later need to build and train in a supervised way.

2 - Supervised training on a specifi labeled dataset.

Model: (pre-trained in step #1)

Universal representation to enable a variety of downstream tasks Q&A (context-aware search engines), Knowledge Base Construction (Hera engine), Model Curation (hyper-parameter Evaluation), synthesis of new models

Dataset:

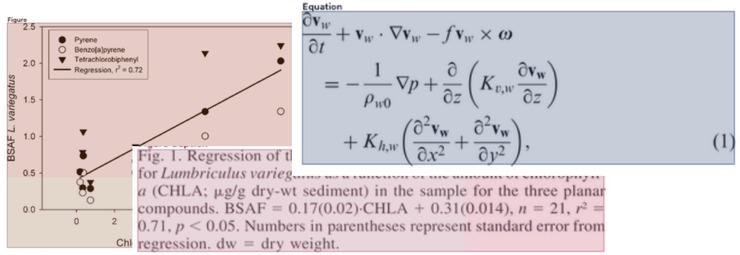


Fig. 1. Regression of t for *Lumbriculus variegatus* as a function of the amount of chlorophyll *a* (CHLA; µg/g dry-wt sediment) in the sample for the three planar compounds. BSAF = 0.17(0.02)·CHLA + 0.31(0.014), $n = 21$, $r^2 = 0.71$, $p < 0.05$. Numbers in parentheses represent standard error from regression. dw = dry weight.

Equation

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

Objective: We need a new contrastive learning objective! Use coreferences from Phase 1 representation.

Q: Mercury , the planet nearest to the Sun , has **extreme surface temperatures** , ranging from 465

C4

hot and cold **temperatures** .
S2: The planet Mercury is *too small* and has too little gravity to hold onto an atmosphere.
...

...

planet 's wild **temperature extremes** .
S2: Mercury is the closest planet to the sun and has a *thin atmosphere*, no air pressure and an **extremely** high **temperature**.
...

Sending evidence to reader

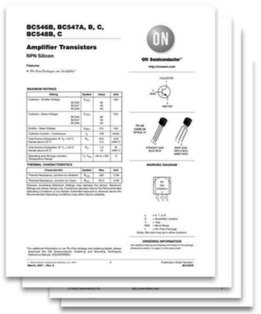
MRC

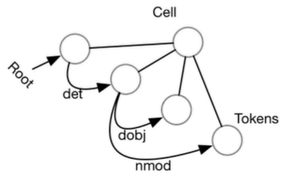
Open-domain retrieve-and-read interaction.
Essentially: question answering over complex scientific domains.

Exploration 2: Contextual AI for open-domain scientific discovery

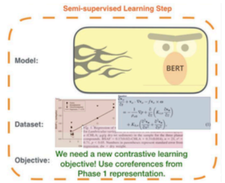
The Hera Engine

Input Documents

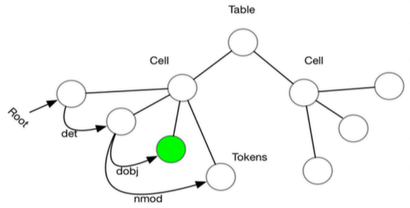




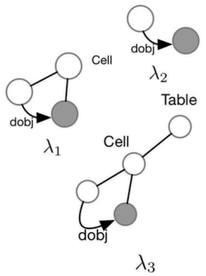
Phase1 Representation



Synthesis Engine



Input Few Annotated Examples



Output Ensemble of Programs

Data Programming

	λ_1	λ_2	λ_3
n_1	1	0	1
n_2	-1	-1	0
n_3	0	0	1

Input Matrix of Labels

Program	Weights (Pos.)
λ_1	0.7
λ_2	0.1
λ_3	0.2

Output
Learned Accuracy of Programs

	$\mathrm{P}(\hat{y}=1)$
n_1	0.8
n_2	0
n_3	0.1

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

Weak Supervision powered by Program Synthesis

Example: Extracting the Collector Current

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* TA = 25°C unless otherwise noted

Symbol	Parameter	Value	Units
V_{CEO}	Collector-Emitter Voltage	25	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

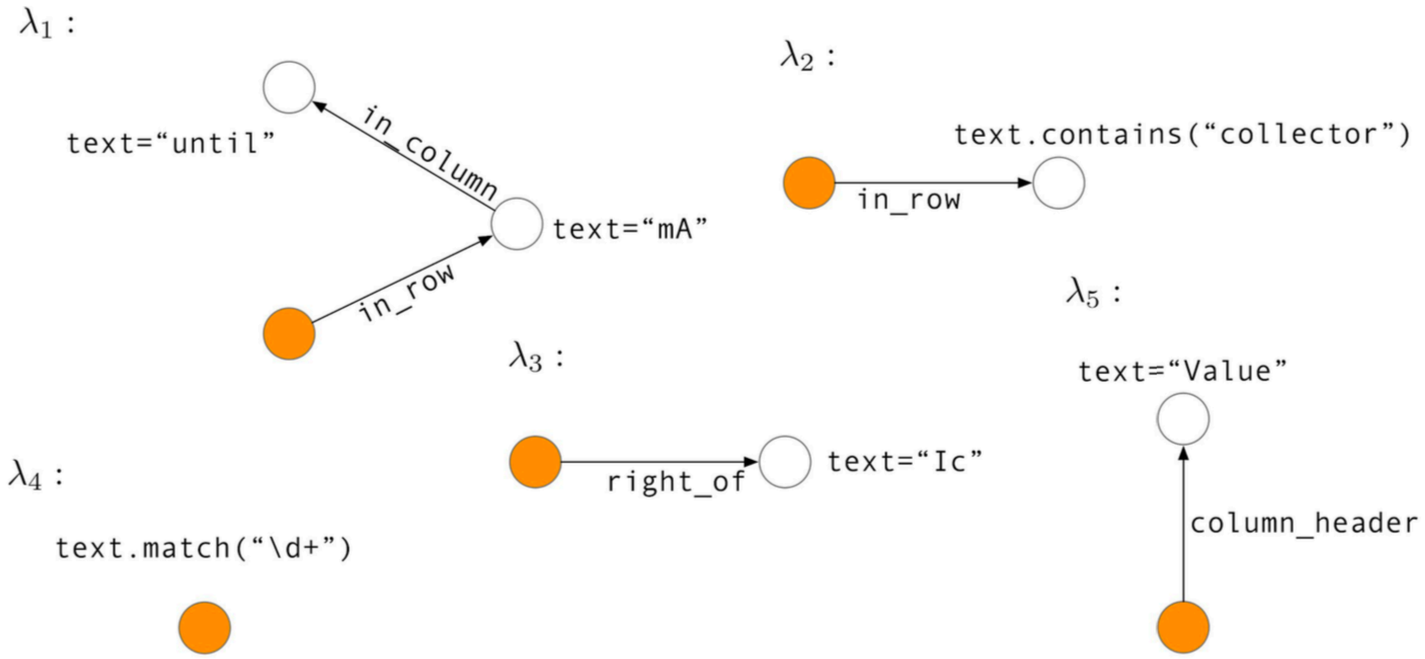
Absolute maximum ratings Ta=25°C

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

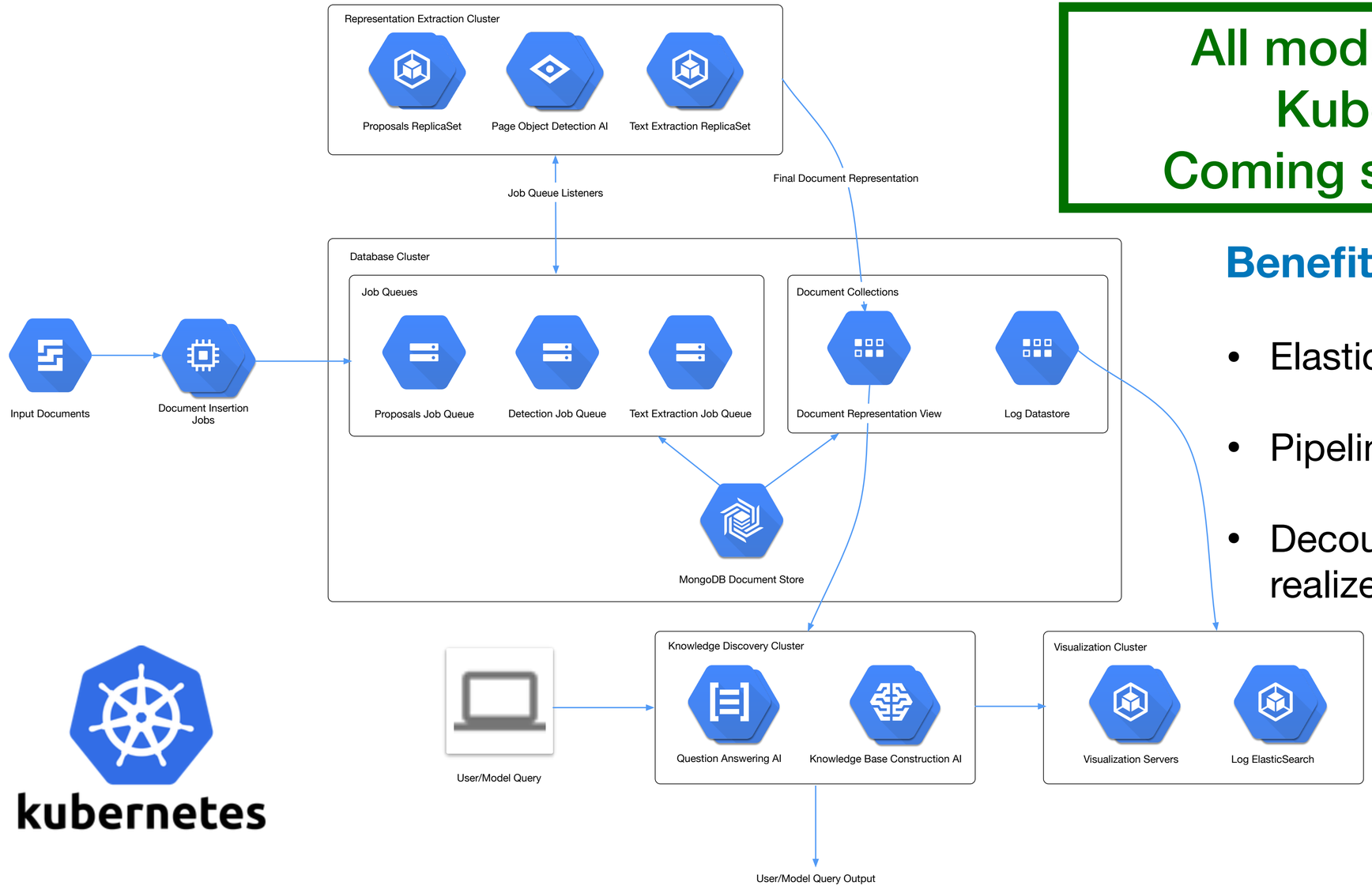
Ensemble of programs synthesized from the input examples



Task

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

Exploration 3: Self-supervised systems as a service



All modules of COSMOS are Kubernetes services. Coming soon in release v0.2.0!

Benefits

- Elasticity and easy scalability
- Pipelining of tasks
- Decoupled microservice architecture to realize an open AI platform

kubernetes