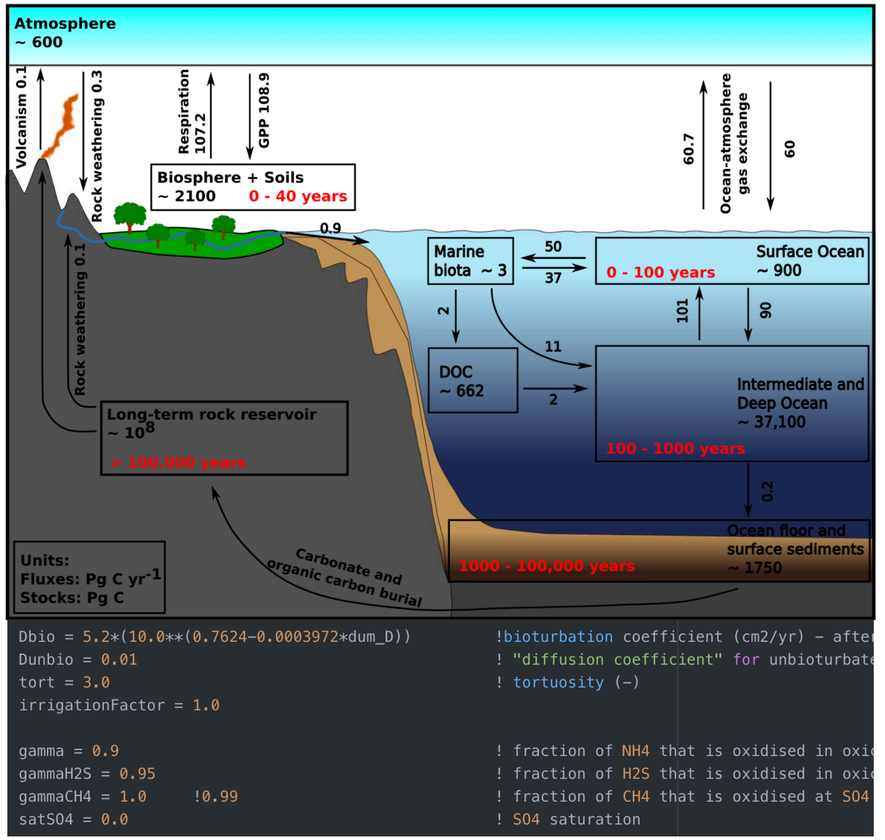
COSMOS: Curation Of Scientific MOdelS of the Earth system from publications

Theo Rekatsinas, Shanan Peters, Miron Livny
University of Wisconsin-Madison

$$F_{CaCO_3} = F_{CaCO_3,0}\left(1 + k_{Ca}\left(T - T_0\right)\right)\frac{R}{R_0}\frac{P}{P_0}$$

$$F_{CaSiO_3} = F_{CaSiO_3,0}e^{\frac{1000E_a}{RT_0^2}(T-T_0)}\left(\frac{R}{R_0}\right)^{\beta}\frac{P}{P_0}$$

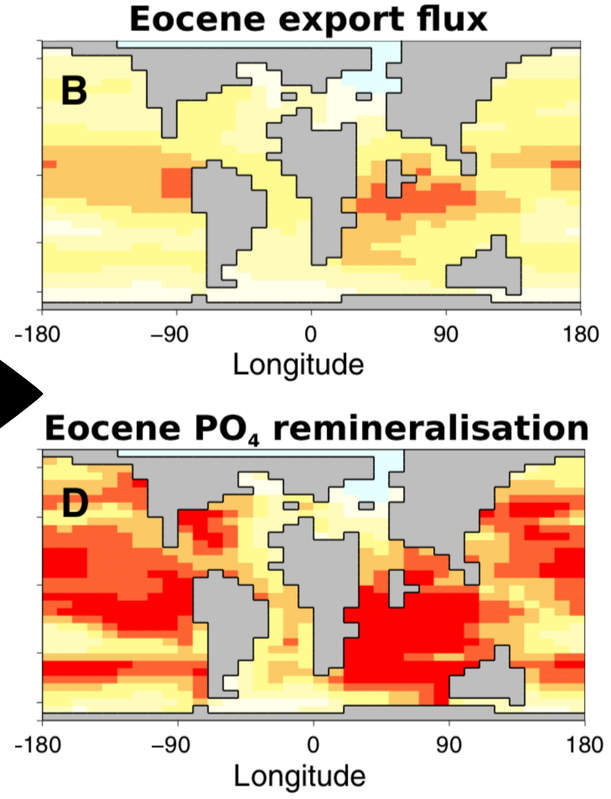
Variables	PICs of PMIP2 and CMIP5–PMIP3
Potential temperature (°C), N. Atlantic	2.9 to 6.4[a]
Potential temperature (°C), S. Atlantic	−1.6 to 2.0[c]
Salinity, N. Atlantic	34.8 to 35.5[a]
Salinity, S. Atlantic AMOC	34.6 to 35.0[c]
(1 Sv = 10^6 m^3 s^{-1})	12.64 to 23.02[e]

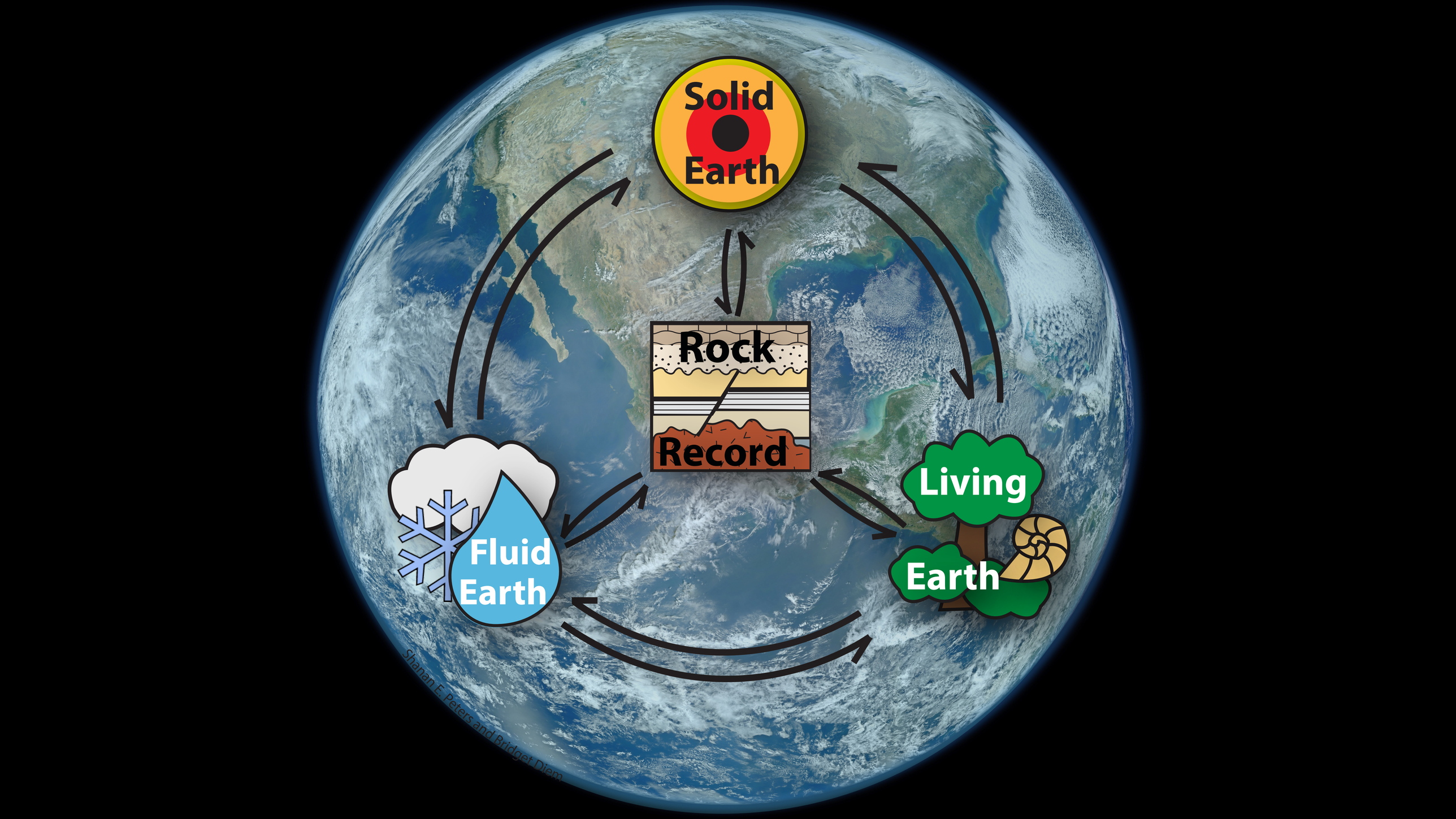



```
Dbio = 5.2*(10.0**(0.7624-0.0003972*dum_D))      !bioturbation coefficient (cm2/yr) - after
Dunbio = 0.01                                    ! "diffusion coefficient" for unbioturbate
tort = 3.0                                       ! tortuosity (-)
irrigationFactor = 1.0

gamma = 0.9                                      ! fraction of NH4 that is oxidised in oxi
gammaH2S = 0.95                                  ! fraction of H2S that is oxidised in oxi
gammaCH4 = 1.0      !0.99                        ! fraction of CH4 that is oxidised at SO4
satSO4 = 0.0                                     ! SO4 saturation
```

GENIE:
Grid ENabled Integrated Earth system model




Solid Earth
Rock Record
Living Earth
Fluid Earth
Shanan E. Peters and Bridget Diem

Structured Data

deployed instruments

field measurements

“Dark Data”

Structured Data

deployed instruments

field measurements

“Dark Data”

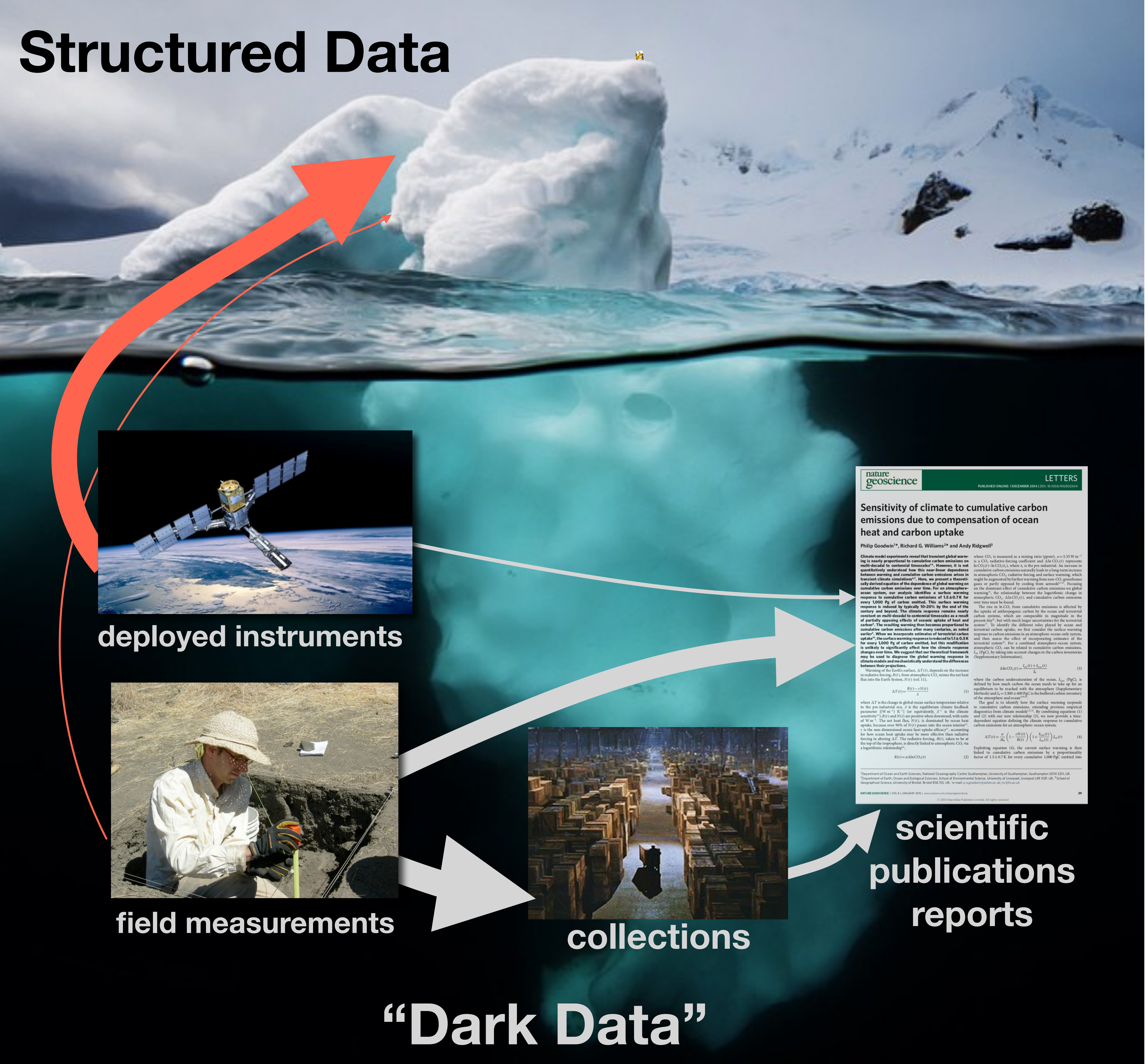
Structured Data

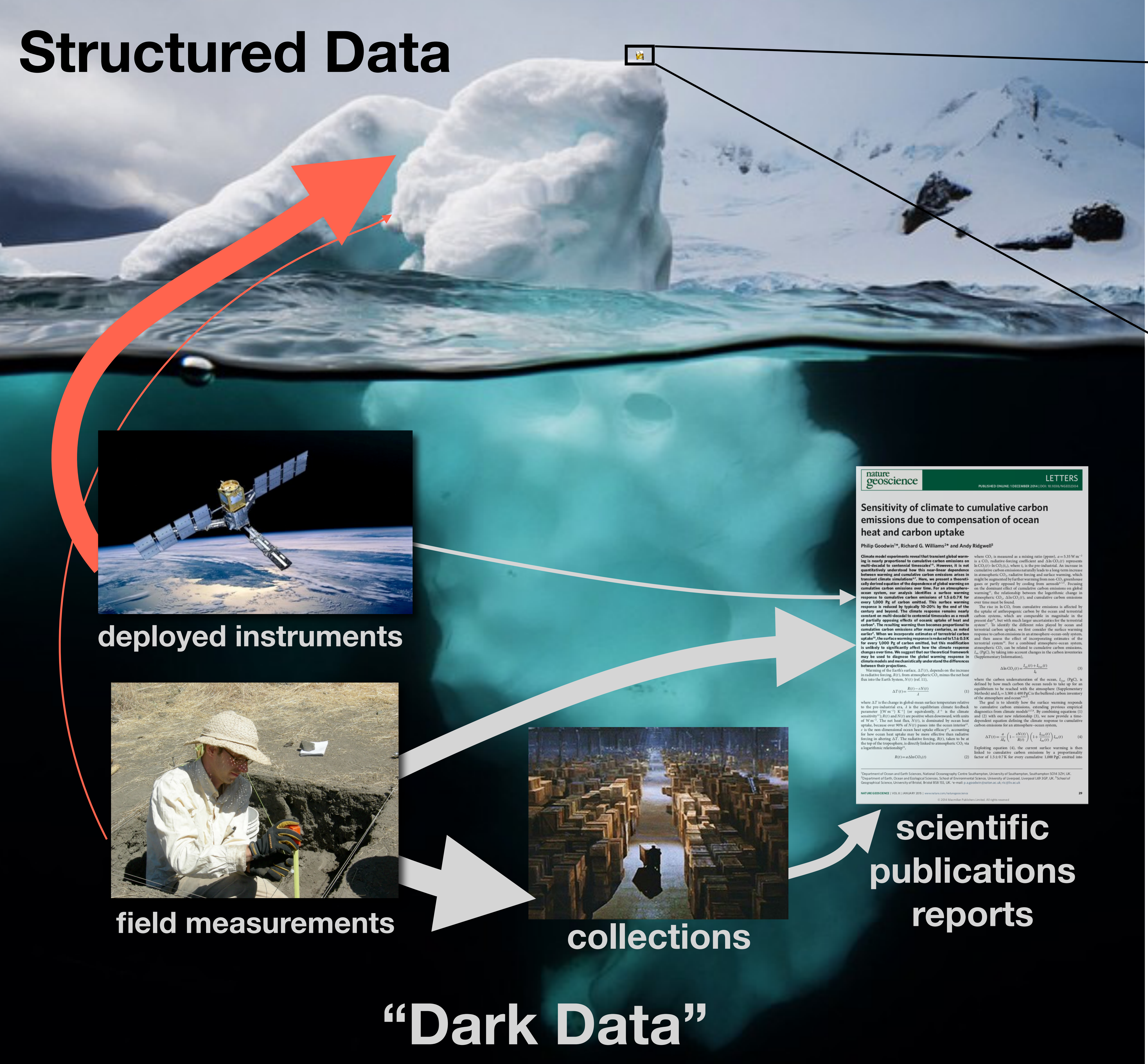
deployed instruments

field measurements

collections

“Dark Data”


Structured Data
deployed instruments
field measurements
collections
nature geoscience
LETTERS
Sensitivity of climate to cumulative carbon emissions due to compensation of ocean heat and carbon uptake
scientific publications reports
“Dark Data”



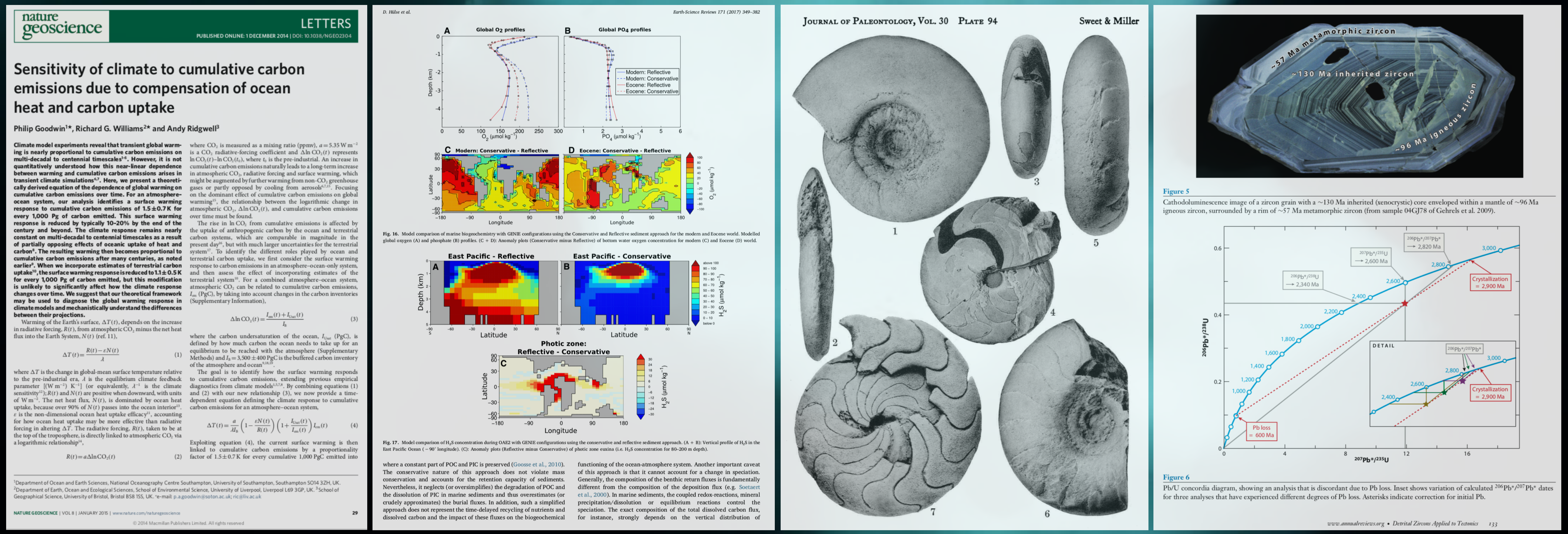
Scientific Models

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


Structured Data
Scientific Models
$$\frac{\partial c}{\partial t} = \nabla \cdot (D \nabla c) - \nabla \cdot (\mathbf{v} c) + R$$
deployed instruments
field measurements
collections
nature geoscience
LETTERS
Sensitivity of climate to cumulative carbon emissions due to compensation of ocean heat and carbon uptake
scientific
publications
reports
“Dark Data”


Structured Data
Scientific Models
$$\frac{\partial c}{\partial t} = \nabla \cdot (D \nabla c) - \nabla \cdot (\mathbf{v} c) + R$$
nature geoscience
LETTERS
Sensitivity of climate to cumulative carbon emissions due to compensation of ocean heat and carbon uptake
Philip Goodwin*, Richard G. Williams and Andy Ridgwell
deployed instruments
field measurements
collections
scientific publications reports
Model Predictions
“Dark Data”

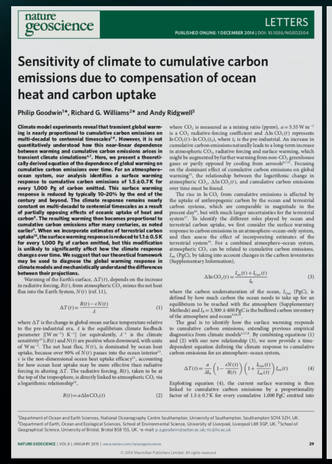
Structured Data



scientific publications are rich in
text, images, illustrations, tables, equations

“Dark Data”

Structured Data



nature geoscience

LETTERS

PUBLISHED ONLINE: 1 DECEMBER 2014 | DOI: 10.1038/NGEO2324

Sensitivity of climate to cumulative carbon emissions due to compensation of ocean heat and carbon uptake

Philip Goodwin[1]*, Richard G. Williams[2]* and Andy Ridgwell[3]

Climate model experiments reveal that transient global warming is nearly proportional to cumulative carbon emissions on multi-decadal to centennial timescales[1–4]. However, it is not quantitatively understood how this near-linear dependence between warming and cumulative carbon emissions arises in transient climate simulations[1–4]. Here, we present a theoretically derived equation of the dependence of global warming on cumulative carbon emissions. For a simple ocean–atmosphere system, our analysis identifies a surface warming response to cumulative carbon emissions of 1.5 ± 0.7 K for every 1,000 Pg of carbon emitted. This surface warming response is reduced by typically 50–20% by the end of the century and beyond. The climate response remains nearly constant on multi-decadal to centennial timescales as a result of the compensation of ocean heat uptake by ocean carbon[5–8]. The resulting warming then becomes proportional to cumulative carbon emissions after many centuries, as noted earlier[1]. When we incorporate estimates of terrestrial carbon uptake[9], the surface warming response is reduced to 1.1 ± 0.5 K for every 1,000 Pg of carbon emitted, but this modification is unlikely to significantly affect how the climate response changes over time. We suggest that our theoretical framework may be used to diagnose the global warming response in Earth system models and to statistically understand the differences between their projections.

Warming of the Earth's surface, $\Delta T(t)$, depends on the increase in radiative forcing, $R(t)$, from atmospheric CO_2 minus the net heat flux into the Earth System, $N(t)$ (ref. 11),

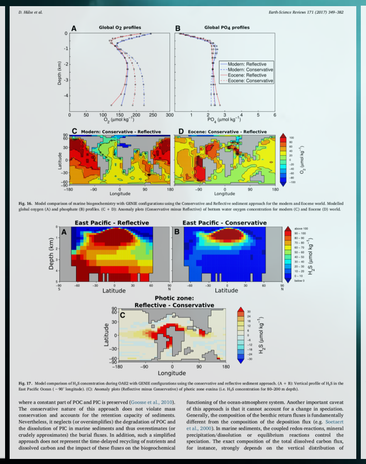
$$\Delta T(t) = \frac{R(t) - \lambda N(t)}{\lambda} \quad (1)$$

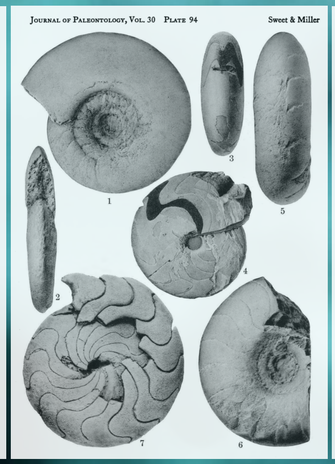
where ΔT is the change in global-mean surface temperature relative to the pre-industrial era, λ is the equilibrium climate feedback parameter (W m^{-2} K^{-1}) [or equivalently, λ^{-1} is the climate sensitivity[10]], $R(t)$ and $N(t)$ are positive when downward, with units of W m^{-2}. The net heat flux, $N(t)$, is dominated by ocean heat uptake, because over 90% of $N(t)$ passes into the ocean interior[5] (κ is the two-dimensional ocean heat uptake efficacy[6], accounting for the reduction in the rate of ocean heat uptake per unit radiative forcing in altering ΔT. The radiative forcing, $R(t)$, taken to be at the top of the troposphere, is directly linked to atmospheric CO_2 via a logarithmic relationship[12],

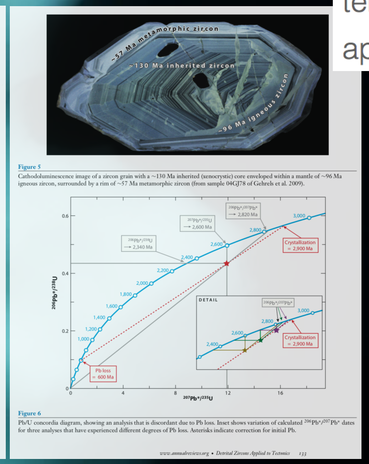
$$R(t) = \alpha \ln(CO_2(t) \quad (2)$$

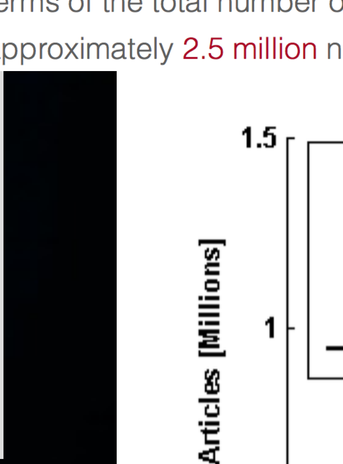
[1]Department of Ocean and Earth Sciences, National Oceanography Centre Southampton, University of Southampton, Southampton SO14 3ZH, UK. [2]Department of Earth, Ocean and Ecological Sciences, School of Environmental Science, University of Liverpool, Liverpool L69 3GP, UK. [3]School of Geographical Science, University of Bristol, Bristol BS8 1SS, UK. *e-mail: p.goodwin@soton.ac.uk; rlw@liv.ac.uk

© 2014 Macmillan Publishers Limited. All rights reserved









scientific publications are rich in
text, images, illustrations, tables, equations

“Dark Data”

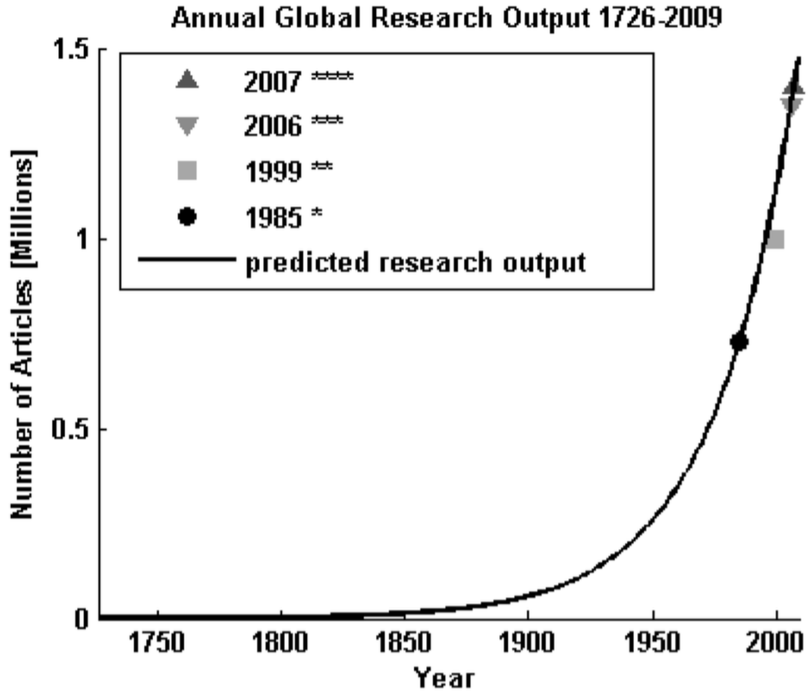
21st Century Science Overload

January 7, 2016

By Sarah Boon, PhD

Do you feel overwhelmed by the number of research papers in your field? Do you wonder if you're missing key ideas that could be critical for your research program? Does it feel like the deluge is only getting worse?

You're not imagining things. According to research from the University of Ottawa, in 2009 we passed the 50 million mark in terms of the total number of science papers published since 1665, and approximately 2.5 million new scientific papers are published each year.



COSMOS: overview of objectives

Biogeosciences, 4, 87–104, 2007
www.biogeosciences.net/4/87/2007/
© Author(s) 2007. This work is licensed
under a Creative Commons License.

Biogeosciences

Marine geochemical data assimilation in an efficient Earth System Model of global biogeochemical cycling

A. Ridgwell[1], J. C. Hargreaves[2], N. R. Edwards[3], J. D. Annan[2], T. M. Lenton[4], R. Marsh[5], A. Yool[5], and A. Watson[4]

[1]School of Geographical Sciences, University of Bristol, Bristol, UK

Table 1. EnKF calibrated biogeochemical parameters in the GENIE-1 model.

Name	Prior assumptions (mean and range[a])	Posterior mean[b]	Description
$u_0^{PO_4}$	1.65 μmol kg^{-1} yr^{-1} (0.3–3.0)	1.91 μmol kg^{-1} yr^{-1}	maximum PO_4 uptake (removal) rate (Eq. 3)
K^{PO_4}	0.2 μmol kg^{-1} (0.1–0.3)	0.21 μmol kg^{-1}	PO_4 Michaelis-Menton half-saturation concentration (Eq. 3)
r^{POC}	0.05 (0.02–0.08)	0.055	initial proportion of POC export as fraction #2 (Eq. 6)
l^{POC}	600 m (200–1000)	556 m	e-folding remineralization depth of POC fraction #1 (Eq. 6)
$r_0^{CaCO_3:POC}$	0.036 (0.015–0.088)[c]	0.022[d]	$CaCO_3$:POC: export rain ratio scalar (Eq. 8)
η	1.5 (1.0–2.0)	1.28	thermodynamic calcification rate power (Eq. 9)
r^{CaCO_3}	0.4 (0.2–0.6)	0.489	initial proportion of $CaCO_3$ export as fraction #2 (Eq. 11)
l^{CaCO_3}	600 m (200–1000)	1055	e-folding remineralization depth of $CaCO_3$ fraction #1 (Eq. 11)

[a] the range is quoted as 1 standard deviation either side of the mean

addition of calcium carbonate preservation in
ments and its role in regulating atmospheric C
elsewhere (Ridgwell and Hargreaves, 2007).

We have calibrated the model paramete
ocean carbon cycling in GENIE-1 by assimilat
vational datasets of phosphate and alkalinity u
ble Kalman filter method. The calibrated (me
dicts a global export production of particulate

particulate organic phosphorus ($F_{z=h_e}^{POP}$, in units of mol PO_4 m^{-2} yr^{-1}) is equated directly with PO_4 uptake (Eq. 1):

$$F_{z=h_e}^{POP} = \int_{h_e}^{0} \rho \cdot (1 - \nu) \cdot \Gamma \, dz \quad (4)$$

where ρ is the density of seawater and h_e the thickness of the euphotic zone (175 m in the 8-level version of this ocean model).

COSMOS: overview of objectives

1. Automate knowledge base construction (KBC) for equations, parameterizations, and descriptions of scientific models expressed in the published literature.

Biogeosciences, 4, 87–104, 2007
www.biogeosciences.net/4/87/2007/
© Author(s) 2007. This work is licensed
under a Creative Commons License.

Biogeosciences

Marine geochemical data assimilation in an efficient Earth System Model of global biogeochemical cycling

A. Ridgwell[1], J. C. Hargreaves[2], N. R. Edwards[3], J. D. Annan[2], T. M. Lenton[4], R. Marsh[5], A. Yool[5], and A. Watson[4]

[1]School of Geographical Sciences, University of Bristol, Bristol, UK

Table 1. EnKF calibrated biogeochemical parameters in the GENIE-1 model.

Name	Prior assumptions (mean and range[a])	Posterior mean[b]	Description
$u_0^{PO_4}$	1.65 μmol kg^{-1} yr^{-1} (0.3–3.0)	1.91 μmol kg^{-1} yr^{-1}	maximum PO_4 uptake (removal) rate (Eq. 3)
K^{PO_4}	0.2 μmol kg^{-1} (0.1–0.3)	0.21 μmol kg^{-1}	PO_4 Michaelis-Menton half-saturation concentration (Eq. 3)
r^{POC}	0.05 (0.02–0.08)	0.055	initial proportion of POC export as fraction #2 (Eq. 6)
l^{POC}	600 m (200–1000)	556 m	e-folding remineralization depth of POC fraction #1 (Eq. 6)
$r_0^{CaCO_3:POC}$	0.036 (0.015–0.088)[c]	0.022[d]	$CaCO_3$:POC: export rain ratio scalar (Eq. 8)
η	1.5 (1.0–2.0)	1.28	thermodynamic calcification rate power (Eq. 9)
r^{CaCO_3}	0.4 (0.2–0.6)	0.489	initial proportion of $CaCO_3$ export as fraction #2 (Eq. 11)
l^{CaCO_3}	600 m (200–1000)	1055	e-folding remineralization depth of $CaCO_3$ fraction #1 (Eq. 11)

[a] the range is quoted as 1 standard deviation either side of the mean

addition of calcium carbonate preservation in
ments and its role in regulating atmospheric C
elsewhere (Ridgwell and Hargreaves, 2007).

We have calibrated the model paramete
ocean carbon cycling in GENIE-1 by assimilat
vational datasets of phosphate and alkalinity u
ble Kalman filter method. The calibrated (me
dicts a global export production of particulate

particulate organic phosphorus ($F_{z=h_e}^{POP}$, in units of mol PO_4 m^{-2} yr^{-1}) is equated directly with PO_4 uptake (Eq. 1):

$$F_{z=h_e}^{POP} = \int_{h_e}^{0} \rho \cdot (1 - \nu) \cdot \Gamma \, dz \quad (4)$$

where ρ is the density of seawater and h_e the thickness of the euphotic zone (175 m in the 8-level version of this ocean model).

COSMOS: overview of objectives

1. Automate knowledge base construction (KBC) for equations, parameterizations, and descriptions of scientific models expressed in the published literature.

2. Automate KBC for empirical/experimental data and observations in publications that are semantically related to model inputs and predictions.

Biogeosciences, 4, 87–104, 2007
www.biogeosciences.net/4/87/2007/
© Author(s) 2007. This work is licensed
under a Creative Commons License.

Biogeosciences

Marine geochemical data assimilation in an efficient Earth System Model of global biogeochemical cycling

A. Ridgwell[1], J. C. Hargreaves[2], N. R. Edwards[3], J. D. Annan[2], T. M. Lenton[4], R. Marsh[5], A. Yool[5], and A. Watson[4]

[1]School of Geographical Sciences, University of Bristol, Bristol, UK

Table 1. EnKF calibrated biogeochemical parameters in the GENIE-1 model.

Name	Prior assumptions (mean and range[a])	Posterior mean[b]	Description
$u_0^{PO_4}$	1.65 μmol kg^{-1} yr^{-1} (0.3–3.0)	1.91 μmol kg^{-1} yr^{-1}	maximum PO_4 uptake (removal) rate (Eq. 3)
K^{PO_4}	0.2 μmol kg^{-1} (0.1–0.3)	0.21 μmol kg^{-1}	PO_4 Michaelis-Menton half-saturation concentration (Eq. 3)
r^{POC}	0.05 (0.02–0.08)	0.055	initial proportion of POC export as fraction #2 (Eq. 6)
l^{POC}	600 m (200–1000)	556 m	e-folding remineralization depth of POC fraction #1 (Eq. 6)
$r_0^{CaCO_3:POC}$	0.036 (0.015–0.088)[c]	0.022[d]	$CaCO_3$:POC: export rain ratio scalar (Eq. 8)
η	1.5 (1.0–2.0)	1.28	thermodynamic calcification rate power (Eq. 9)
r^{CaCO_3}	0.4 (0.2–0.6)	0.489	initial proportion of $CaCO_3$ export as fraction #2 (Eq. 11)
l^{CaCO_3}	600 m (200–1000)	1055	e-folding remineralization depth of $CaCO_3$ fraction #1 (Eq. 11)

[a] the range is quoted as 1 standard deviation either side of the mean

addition of calcium carbonate preservation in
ments and its role in regulating atmospheric C
elsewhere (Ridgwell and Hargreaves, 2007).

We have calibrated the model paramete
ocean carbon cycling in GENIE-1 by assimilat
vational datasets of phosphate and alkalinity u
ble Kalman filter method. The calibrated (me
dicts a global export production of particulate

particulate organic phosphorus ($F_{z=h_e}^{POP}$, in units of mol PO_4 m^{-2} yr^{-1}) is equated directly with PO_4 uptake (Eq. 1):

$$F_{z=h_e}^{POP} = \int_{h_e}^{0} \rho \cdot (1 - \nu) \cdot \Gamma \, dz \quad (4)$$

where ρ is the density of seawater and h_e the thickness of the euphotic zone (175 m in the 8-level version of this ocean model).

COSMOS: overview of objectives

1. Automate knowledge base construction (KBC) for equations, parameterizations, and descriptions of scientific models expressed in the published literature.

2. Automate KBC for empirical/experimental data and observations in publications that are semantically related to model inputs and predictions.

3. Remove major pain-point in model-data assimilation and improve pace and completeness of model assessment and improvement.

Biogeosciences, 4, 87–104, 2007
www.biogeosciences.net/4/87/2007/
© Author(s) 2007. This work is licensed
under a Creative Commons License.

Biogeosciences

Marine geochemical data assimilation in an efficient Earth System Model of global biogeochemical cycling

A. Ridgwell[1], J. C. Hargreaves[2], N. R. Edwards[3], J. D. Annan[2], T. M. Lenton[4], R. Marsh[5], A. Yool[5], and A. Watson[4]

[1]School of Geographical Sciences, University of Bristol, Bristol, UK

Table 1. EnKF calibrated biogeochemical parameters in the GENIE-1 model.

Name	Prior assumptions (mean and range[a])	Posterior mean[b]	Description
$u_0^{PO_4}$	1.65 μmol kg^{-1} yr^{-1} (0.3–3.0)	1.91 μmol kg^{-1} yr^{-1}	maximum PO_4 uptake (removal) rate (Eq. 3)
K^{PO_4}	0.2 μmol kg^{-1} (0.1–0.3)	0.21 μmol kg^{-1}	PO_4 Michaelis-Menton half-saturation concentration (Eq. 3)
r^{POC}	0.05 (0.02–0.08)	0.055	initial proportion of POC export as fraction #2 (Eq. 6)
l^{POC}	600 m (200–1000)	556 m	*e*-folding remineralization depth of POC fraction #1 (Eq. 6)
$r_0^{CaCO_3:POC}$	0.036 (0.015–0.088)[c]	0.022[d]	$CaCO_3$:POC: export rain ratio scalar (Eq. 8)
η	1.5 (1.0–2.0)	1.28	thermodynamic calcification rate power (Eq. 9)
r^{CaCO_3}	0.4 (0.2–0.6)	0.489	initial proportion of $CaCO_3$ export as fraction #2 (Eq. 11)
l^{CaCO_3}	600 m (200–1000)	1055	*e*-folding remineralization depth of $CaCO_3$ fraction #1 (Eq. 11)

[a] the range is quoted as 1 standard deviation either side of the mean

addition of calcium carbonate preservation in
ments and its role in regulating atmospheric C
elsewhere (Ridgwell and Hargreaves, 2007).

We have calibrated the model paramete
ocean carbon cycling in GENIE-1 by assimilat
vational datasets of phosphate and alkalinity u
ble Kalman filter method. The calibrated (me
dicts a global export production of particulate

particulate organic phosphorus ($F_{z=h_e}^{POP}$, in units of mol PO_4 m^{-2} yr^{-1}) is equated directly with PO_4 uptake (Eq. 1):

$$F_{z=h_e}^{POP} = \int_{h_e}^{0} \rho \cdot (1 - \nu) \cdot \Gamma \, dz \quad (4)$$

where ρ is the density of seawater and h_e the thickness of the euphotic zone (175 m in the 8-level version of this ocean model).

COSMOS: required components

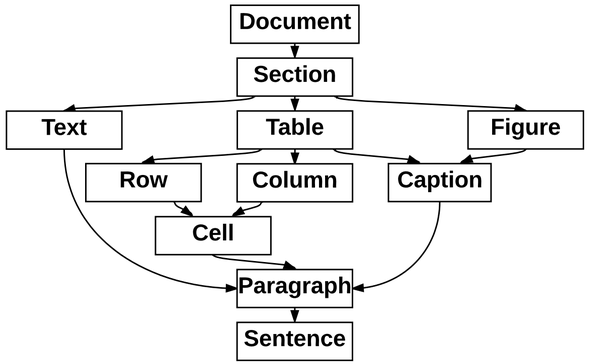
COSMOS: required components

1. Principled, automated access to scientific publications and the computing capacity and infrastructure required to repeatedly analyze them.

COSMOS: required components

1. Principled, automated access to scientific publications and the computing capacity and infrastructure required to repeatedly analyze them.

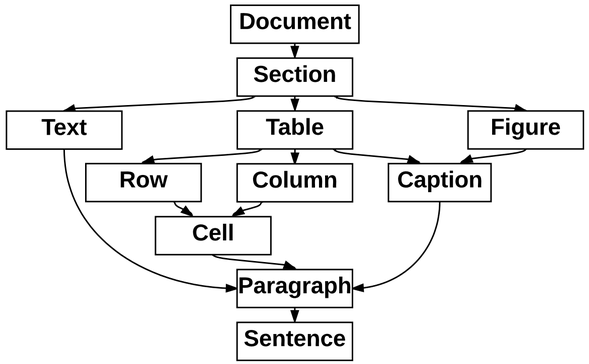
2. Models and techniques to represent and capture multi-modal data within publications.



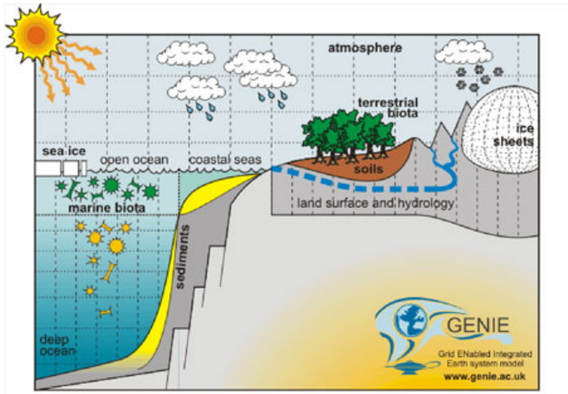
COSMOS: required components

1. Principled, automated access to scientific publications and the computing capacity and infrastructure required to repeatedly analyze them.

2. Models and techniques to represent and capture multi-modal data within publications.



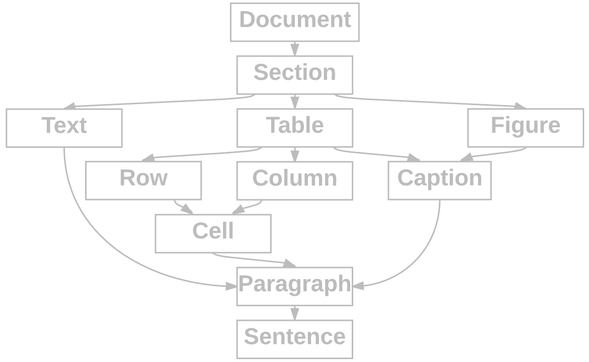
3. Earth system model with parameterizations and predictions that overlap with many different types of empirical data and observations in publications.



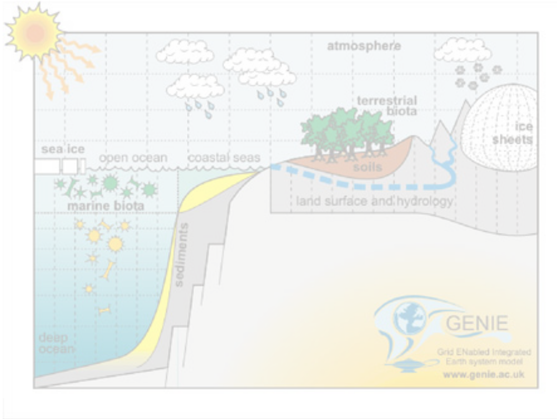
COSMOS: required components

1. Principled, automated access to scientific publications and the computing capacity and infrastructure required to repeatedly analyze them.

2. Models and techniques to represent and capture multi-modal data within publications.



3. Earth system model with parameterizations and predictions that overlap with many different types of empirical data and observations in publications.



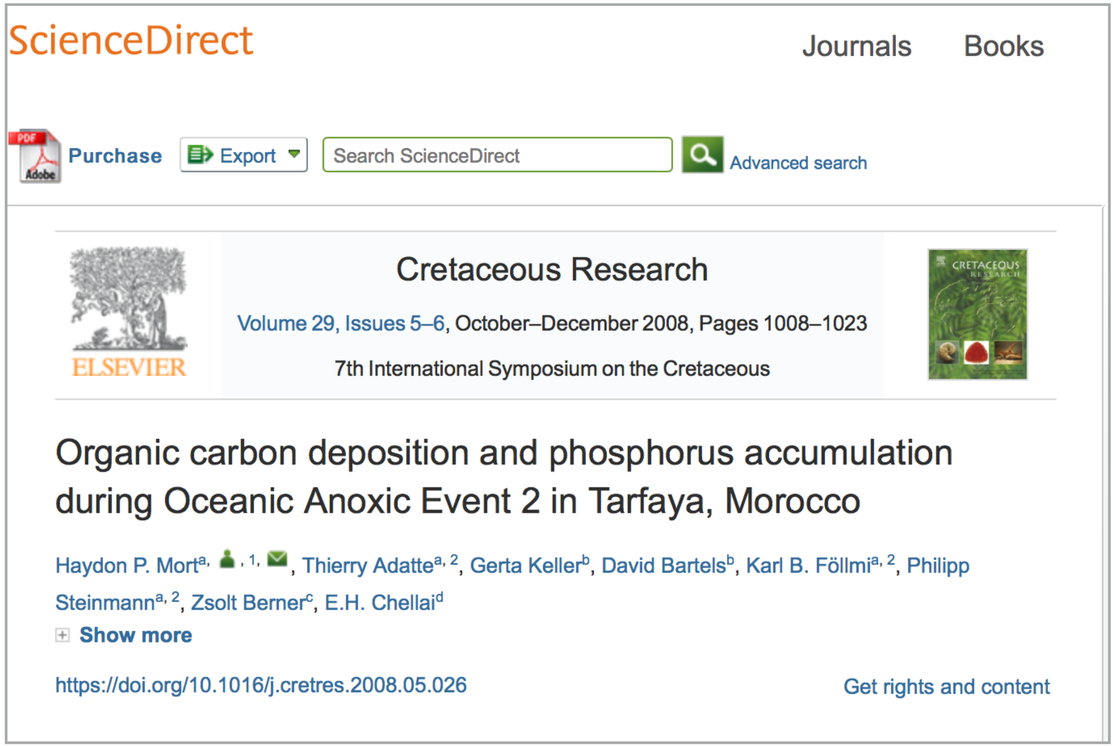
GeoDeepDive

(aka: xDD)

8.3M published documents coupled to *and readable by* a computing infrastructure

HTCondor
High Throughput Computing
CENTER FOR
HIGH THROUGHPUT
COMPUTING

GeoDeepDive Infrastructure Overview



publisher agreements

Libraries
UNIVERSITY OF WISCONSIN-MADISON

USGS
science for a changing world

THE GEOLOGICAL SOCIETY OF AMERICA®

PLOS ONE

ELSEVIER

WILEY
Publishers Since 1807

Canadian Science Publishing

Taylor & Francis
Taylor & Francis Group

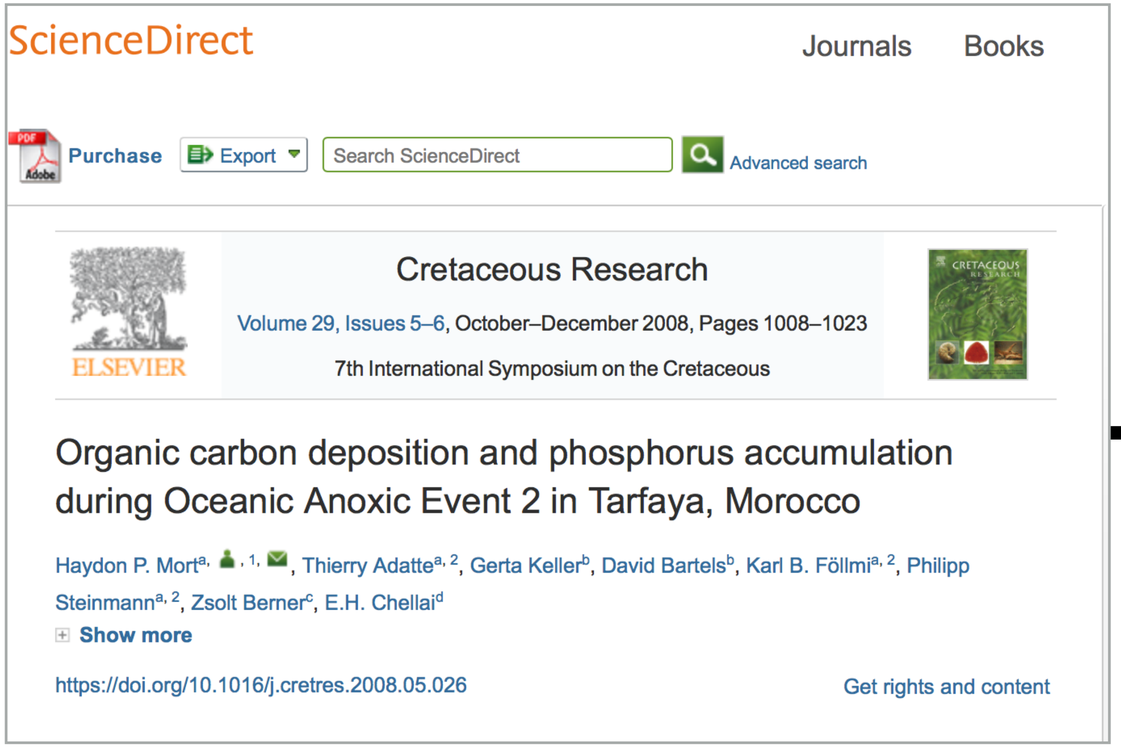
SEPM

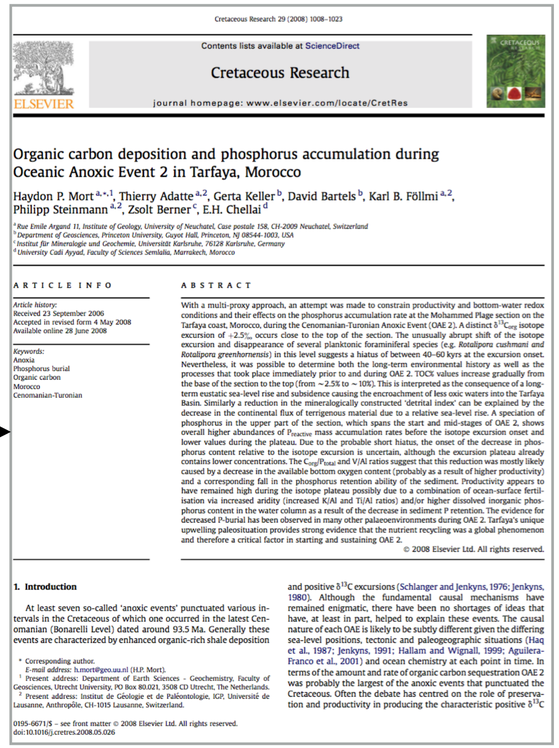
CAMBRIDGE UNIVERSITY PRESS

?

?

GeoDeepDive Infrastructure Overview





publisher agreements

Libraries UNIVERSITY OF WISCONSIN-MADISON

USGS science for a changing world

THE GEOLOGICAL SOCIETY OF AMERICA®

PLOS ONE

ELSEVIER

WILEY Publishers Since 1807

Canadian Science Publishing

Taylor & Francis Taylor & Francis Group

SEPM

CAMBRIDGE UNIVERSITY PRESS

?

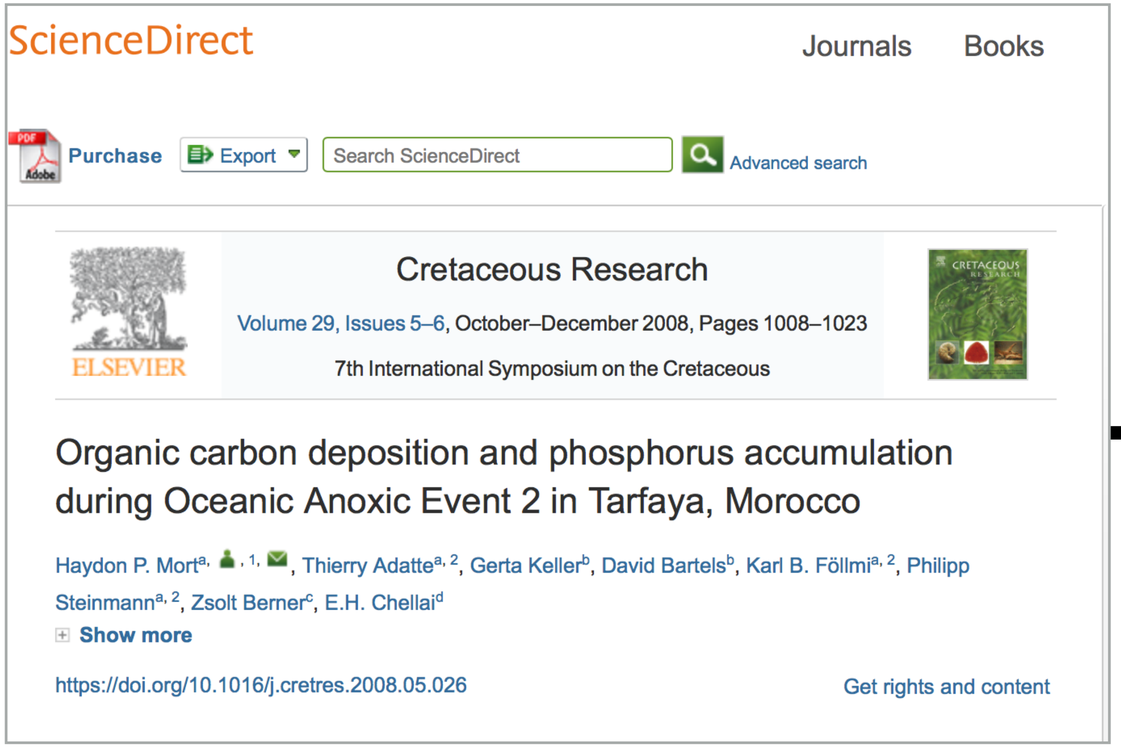
?

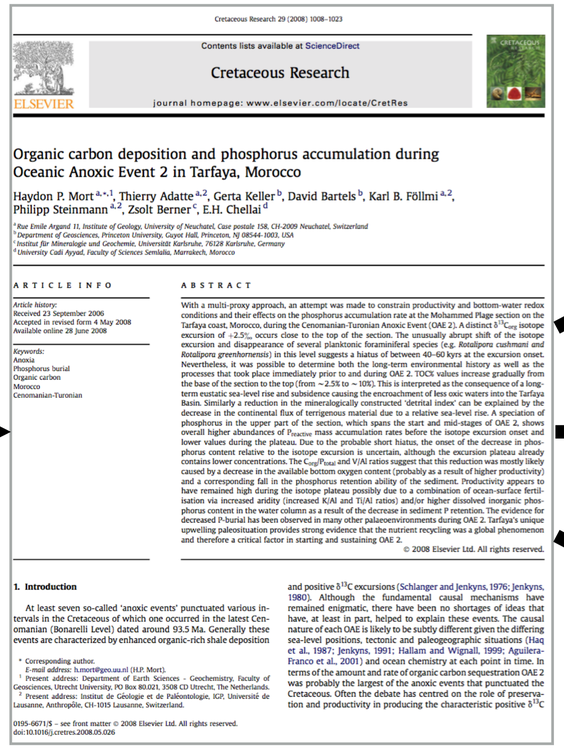
doc. fetching/storage

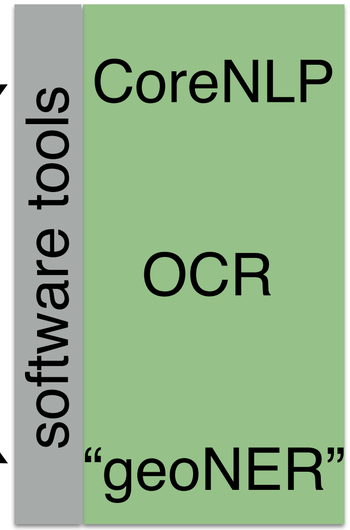
CENTER FOR HIGH THROUGHPUT COMPUTING

8-12k/day

GeoDeepDive Infrastructure Overview







publisher agreements

Libraries UNIVERSITY OF WISCONSIN-MADISON

USGS science for a changing world

THE GEOLOGICAL SOCIETY OF AMERICA®

PLOS ONE

ELSEVIER

WILEY Publishers Since 1807

Canadian Science Publishing

Taylor & Francis Taylor & Francis Group

SEPM Society for Sedimentary Geology

CAMBRIDGE UNIVERSITY PRESS

?

?

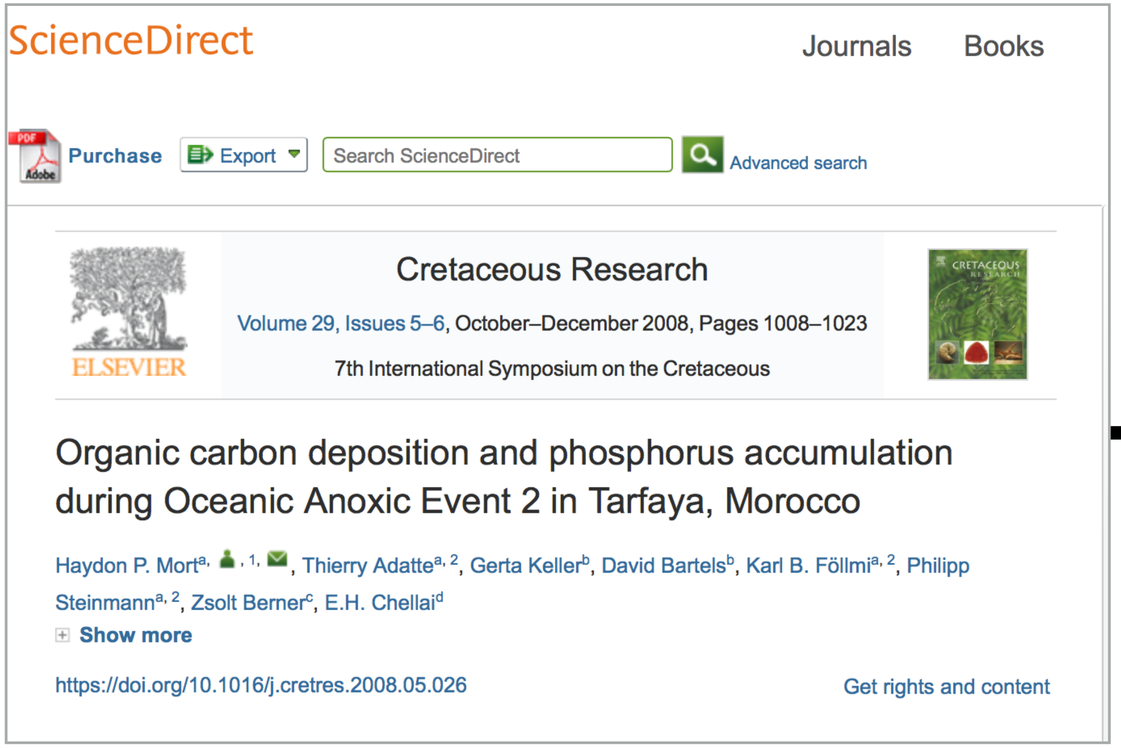
doc. fetching/storage

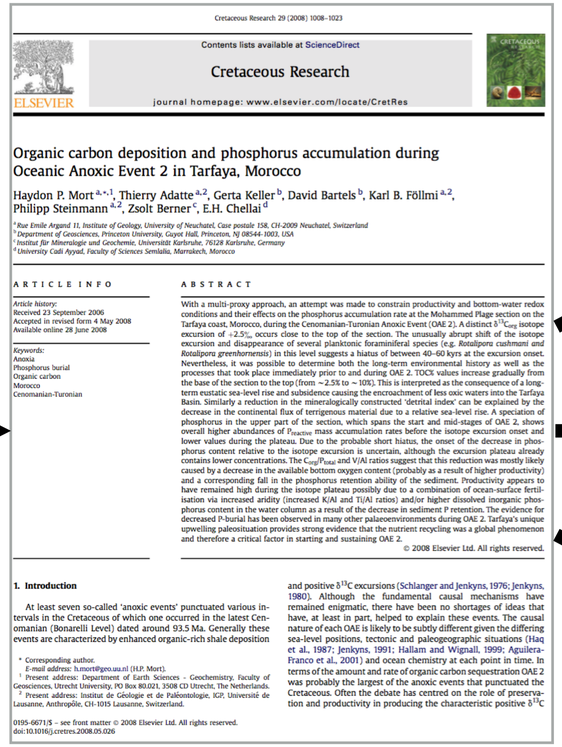
CENTER FOR HIGH THROUGHPUT COMPUTING

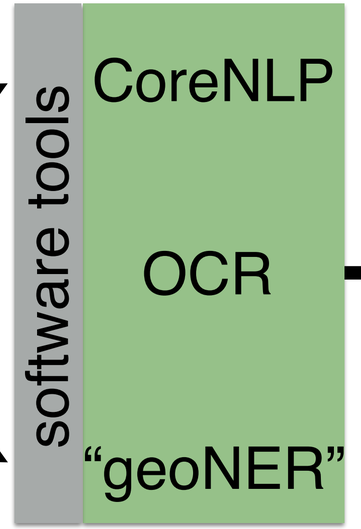
8-12k/day

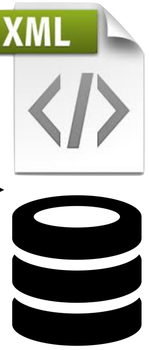
parsing, annotation, labeling

GeoDeepDive Infrastructure Overview









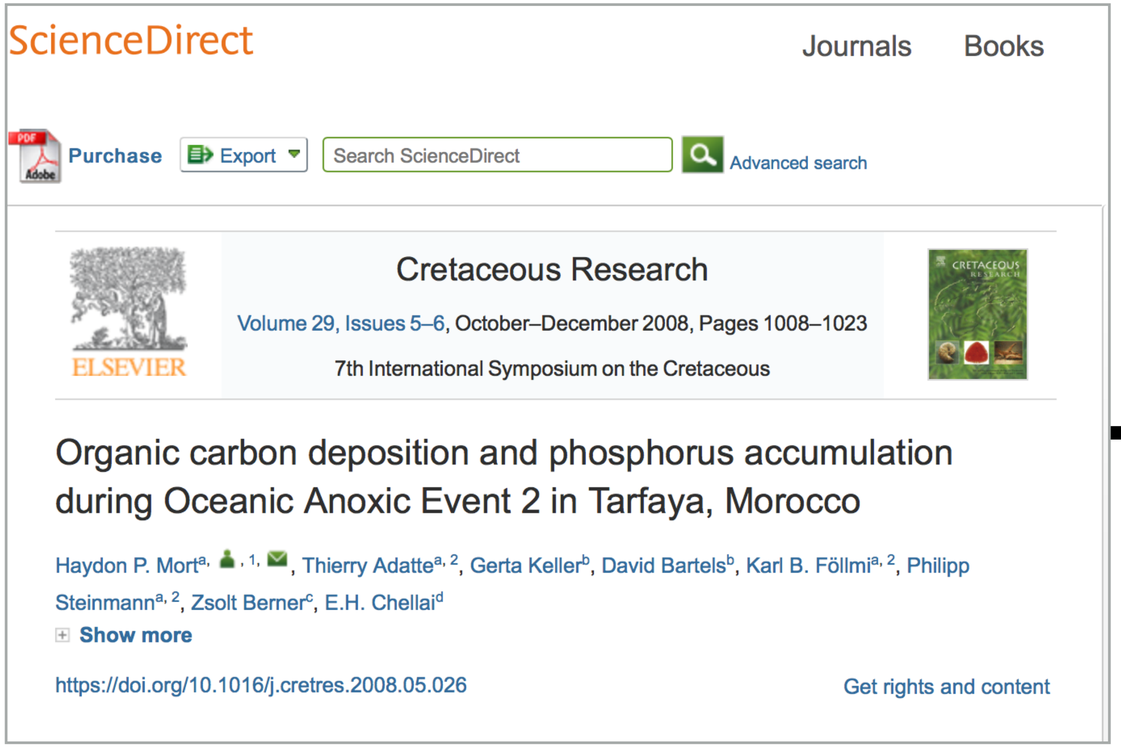
publisher agreements

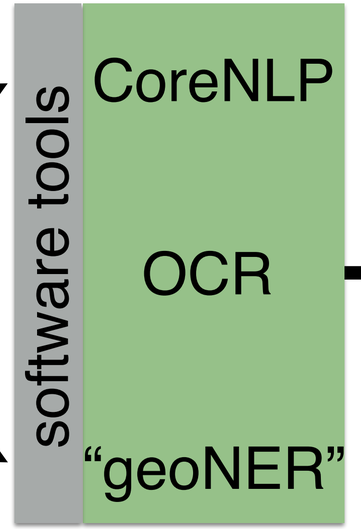
doc. fetching/storage

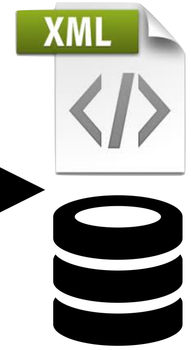
parsing, annotation, labeling

8-12k/day

GeoDeepDive Infrastructure Overview







publisher agreements

doc. fetching/storage

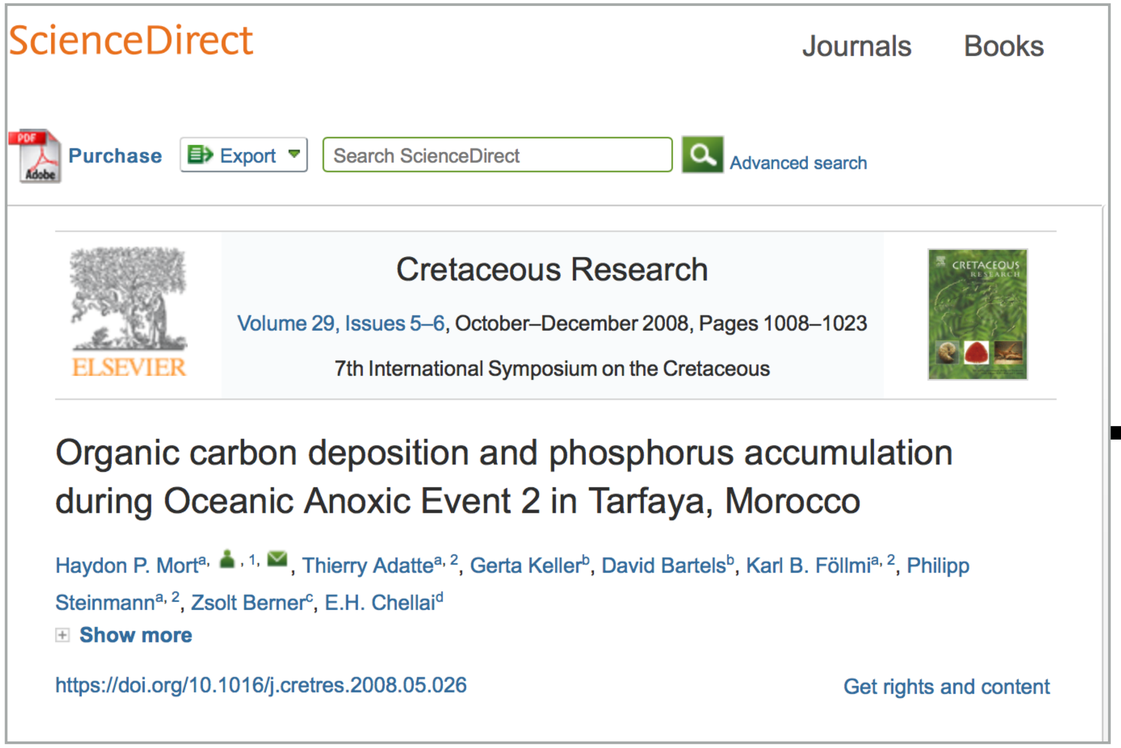
parsing, annotation, labeling

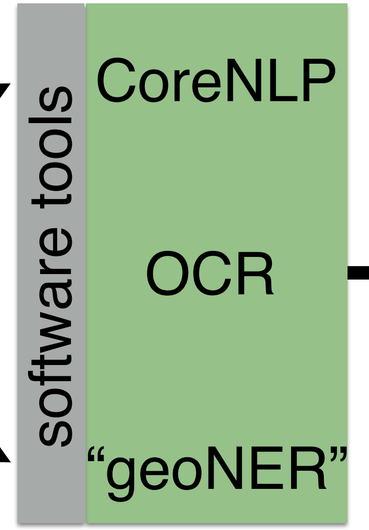
CENTER FOR HIGH THROUGHPUT COMPUTING

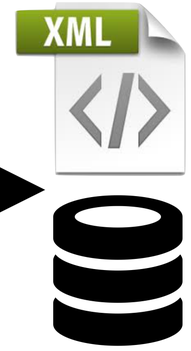
8-12k/day

COSMOS

GeoDeepDive Infrastructure Overview







publisher agreements

doc. fetching/storage

parsing, annotation, labeling

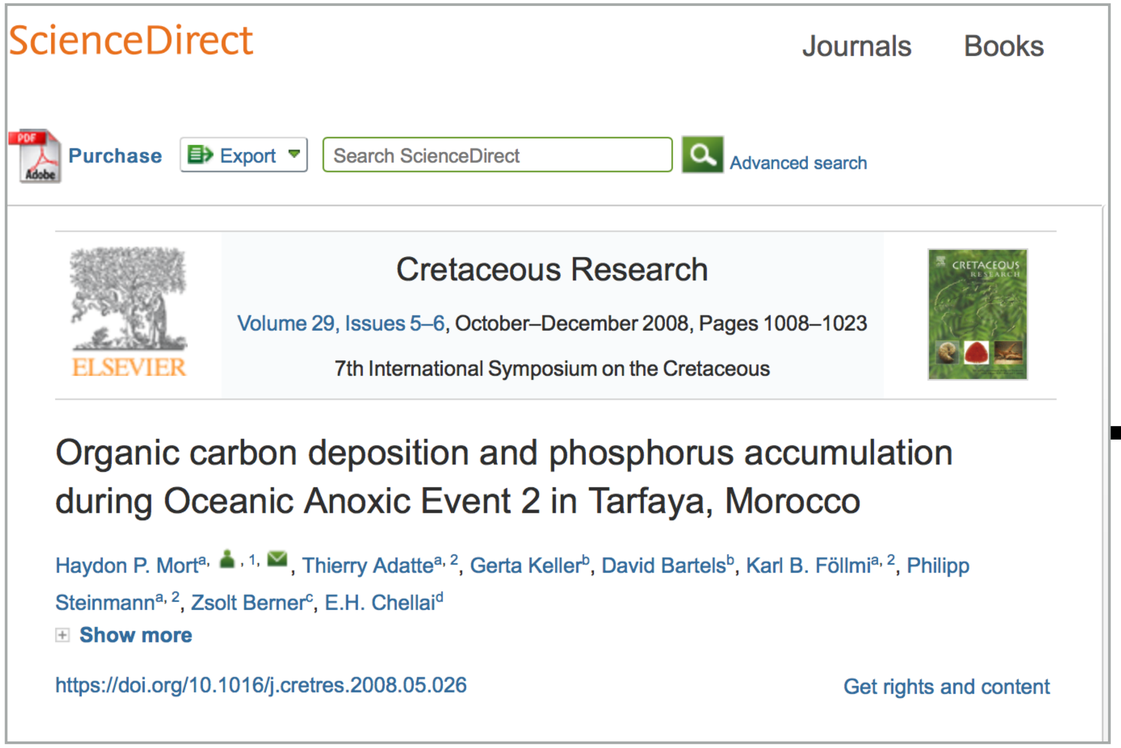
CENTER FOR HIGH THROUGHPUT COMPUTING

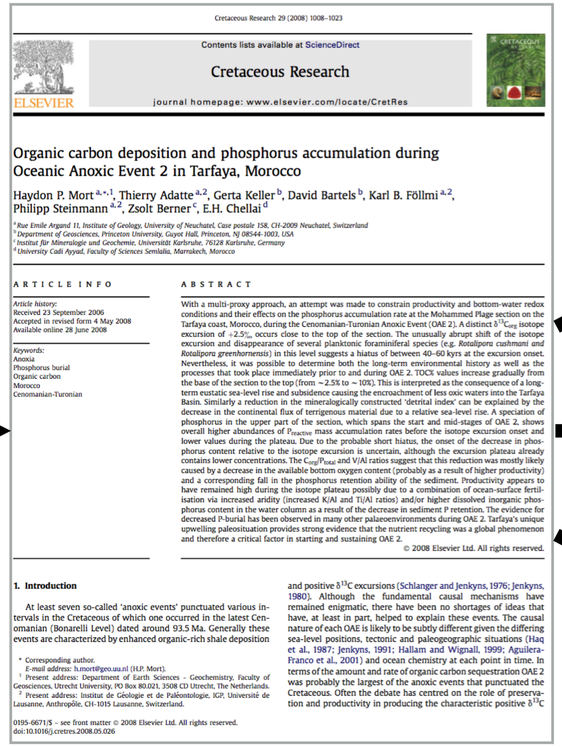
8-12k/day

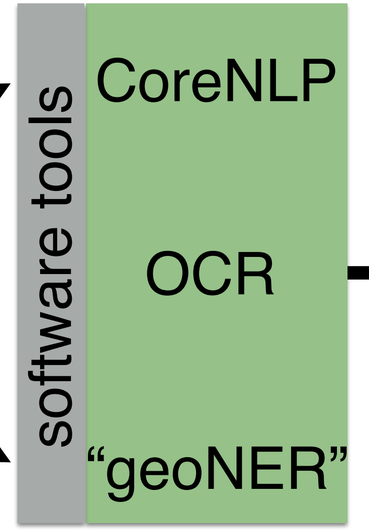
COSMOS

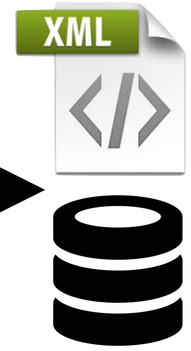
Science!

GeoDeepDive Infrastructure Overview









publisher agreements

doc. fetching/storage

8-12k/day

parsing, annotation, labeling

COSMOS

Science!

Original document data and metadata curation


```
<doi_record>
<journal_metadata language="en">
<full_title>Annals of the New York Academy of
Sciences</full_title>
</journal_metadata>
<journal_issue>
<publication_date media_type="print">
                <month>03</month>
                <year>1956</year>
</publication_date>
<journal_volume>
<volume>63</volume>
<doi_data>
<doi>10.1111/nyas.1956.63.issue-6</doi>
</doi_data>
<titles>
<title>TOPOLOGICAL CYTOCHEMISTRY</title>
</titles>
<contributors>
<person_name contributor_role="author"
sequence="first">
<given_name>M. J.</given_name>
<surname>Kopac</surname>
</person_name>
</contributors>
<first_page>1219</first_page>
<last_page>1235</last_page>
</pages>
```

🔒Original PDFs 🔒


```
{u'dc:title': u'Editorial Board',
 u'prism:coverDate': [{u'$':
u'2011-01-01', u'@_fa':
u'true'}],
 u'prism:coverDisplayDate':
u'January\u2013June 2011',
 u'prism:doi': u'10.1016/
S0753-3969(11)00046-2',
 u'prism:publicationName':
u'Annales de Pal\xe9ontologie',
 u'prism:startingPage': u'i',
u'prism:volume': u'97'}
```

Original document data and metadata curation


```
<doi_record>
<journal_metadata language="en">
<full_title>Annals of the New York Academy of
Sciences</full_title>
</journal_metadata>
<journal_issue>
<publication_date media_type="print">
                <month>03</month>
                <year>1956</year>
</publication_date>
<journal_volume>
<volume>63</volume>
<doi_data>
<doi>10.1111/nyas.1956.63.issue-6</doi>
</doi_data>
<titles>
<title>TOPOLOGICAL CYTOCHEMISTRY</title>
</titles>
<contributors>
<person_name contributor_role="author"
sequence="first">
<given_name>M. J.</given_name>
<surname>Kopac</surname>
</person_name>
</contributors>
<first_page>1219</first_page>
<last_page>1235</last_page>
</pages>
```

🔒Original PDFs 🔒

Fetched by GDD


```
{u'dc:title': u'Editorial Board',
 u'prism:coverDate': [{u'$':
u'2011-01-01', u'@_fa':
u'true'}],
 u'prism:coverDisplayDate':
u'January\u2013June 2011',
 u'prism:doi': u'10.1016/
S0753-3969(11)00046-2',
 u'prism:publicationName':
u'Annales de Pal\xe9ontologie',
 u'prism:startingPage': u'i',
u'prism:volume': u'97'}
```

Harmonized
Document Metadata (bibJSON files + mongoDB)

```
{"authors":"Kopac, M. J.",
"coverDate":"March 1956",
"publication_date":{"month":3,"year":1956},
"publisher":"Wiley-Blackwell",
"pubname":"Annals of the New York Academy of Sciences",
"sha1":"01b2fcc8a94a2bfbe785812aab28e45fcbe02d3f",
"startingPage":"1219",
"time":["2016-03-17T08:39:22.532-05:00"]
"title":"TOPOLOGICAL CYTOCHEMISTRY",
"vol":"63",
"doi":"10.1111/j.1749-6632.1956.tb32132.x",
"url":"http://doi.wiley.com/10.1111/j.
1749-6632.1956.tb32132.x"
…}

{
"coverDate":"January–June 2011",
"publication_date":{"month":6,"year":2011},
"publisher":"Elsevier",
"pubname":"Annales de Paléontologie",
"sha1":"6ee25a57571021e45faa1b545a5c59b9748ada7b",
"startingPage":"i",
"time":"2015-07-09T15:00:17.875-05:00",
"title":"Editorial Board",
"vol":"97",
"doi": "10.1016/S0753-3969(11)00046-2",
"url":"http://www.sciencedirect.com/science/article/pii/
S0753396911000462"
…
}
```

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

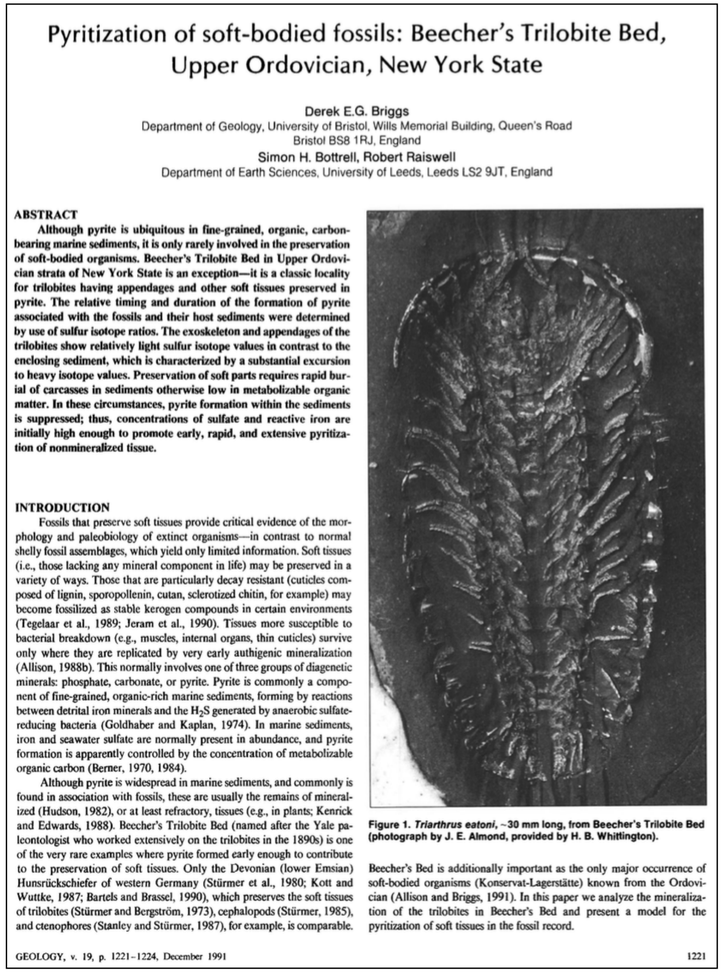
mindat.org

→ 6k
mineral names/
chemistry

Macrostrat

→ 45k
stratigraphic names/
hierarchy, rock types

GDD-supplied full text
and citation info.



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.

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

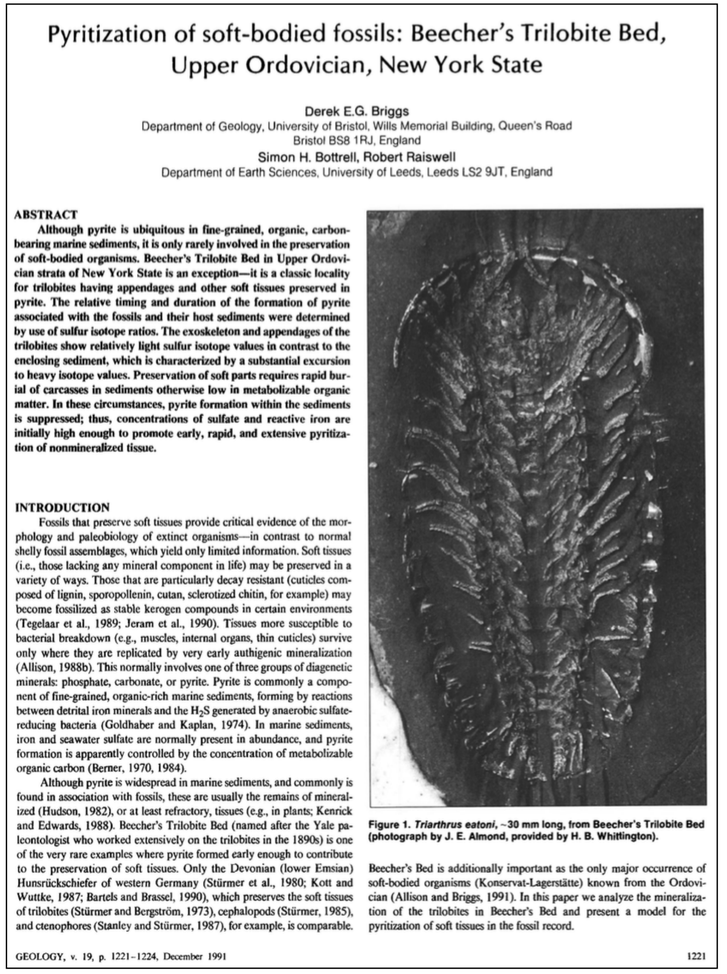
mindat.org

→ 6k
mineral names/
chemistry

Macrostrat

→ 45k
stratigraphic names/
hierarchy, rock types

GDD-supplied full text
and citation info.



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

Trilobita
Triarthrus
Climacograptus

pyrite

Frankfort Shale

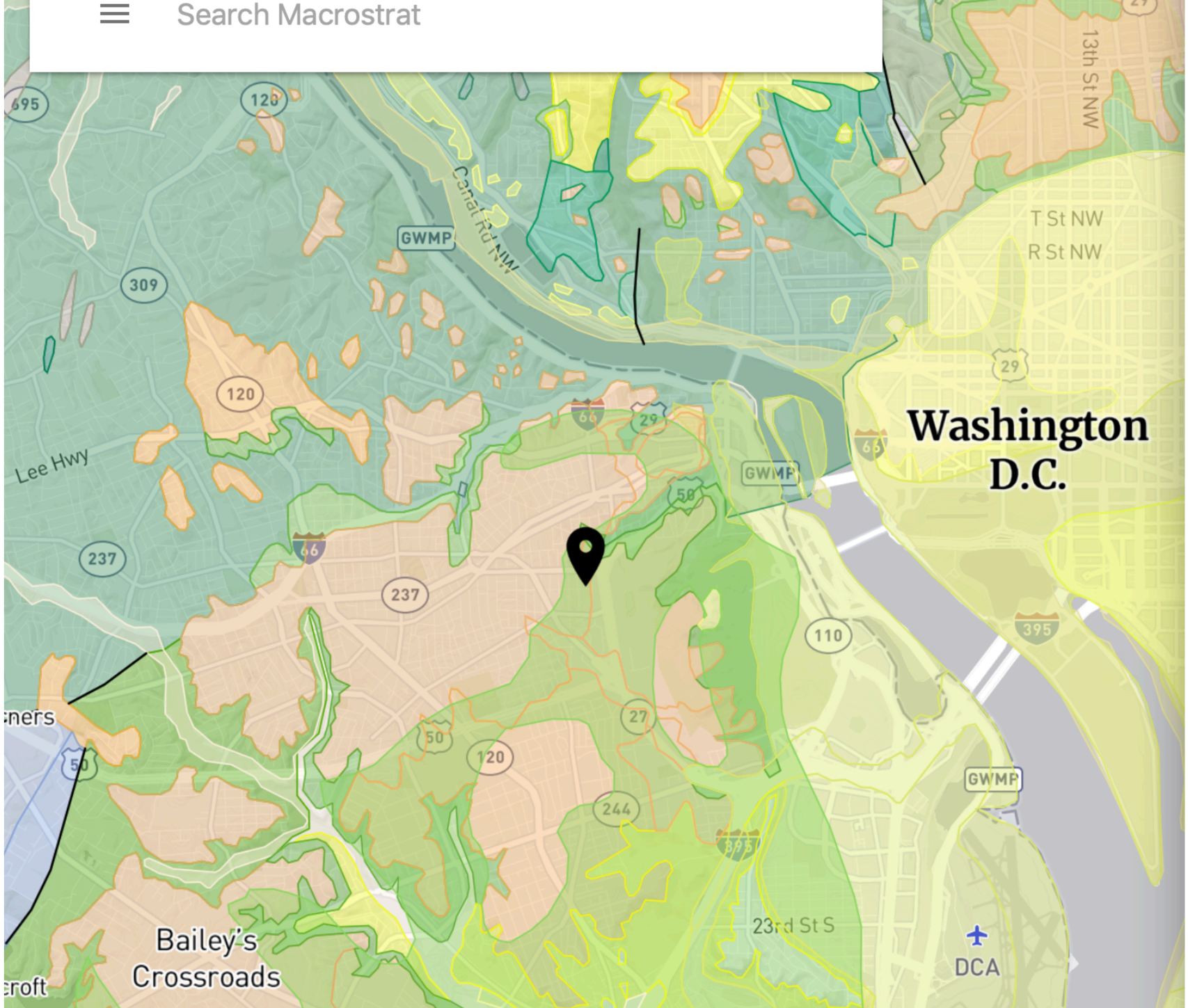
exposed
via API

```
"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



Primary Literature via GeoDeepDive

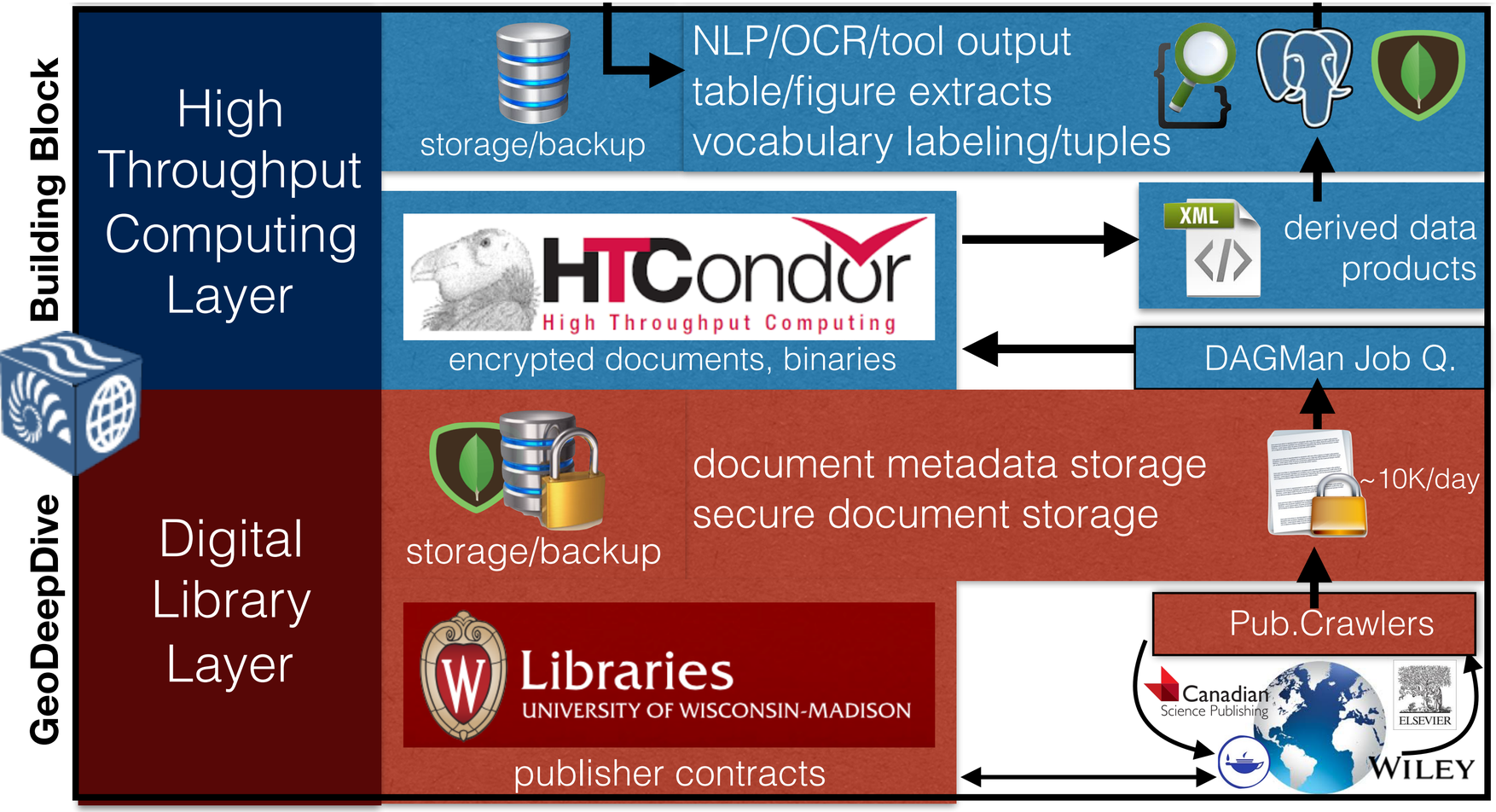
Palaeogeography, Palaeoclimatology, Palaeoecology Elsevier

Wolfe, Jack A., Upchurch, Garland R., 1987. **North American nonmarine climates and vegetation during the Late Cretaceous.**

...First , toothed angiosperm leaves appear by the Aptian (e.g. , Quercophyllum in Zone I of the **Potomac Group** ; Hickey and Doyle , 1977) , and virtually all other major physiognomic types of dicotyledonous leaves appeared by the early Cenomanian (Zone III of the **Potomac Group** ; Upchurch and Wolfe , 1987) ....

Uličný, David, Kvaček, Jiří, Svobodová, Marcela, Špičáková, Lenka, 1997. **High-frequency sea-level fluctuations and plant habitats in Cenomanian fluvial**

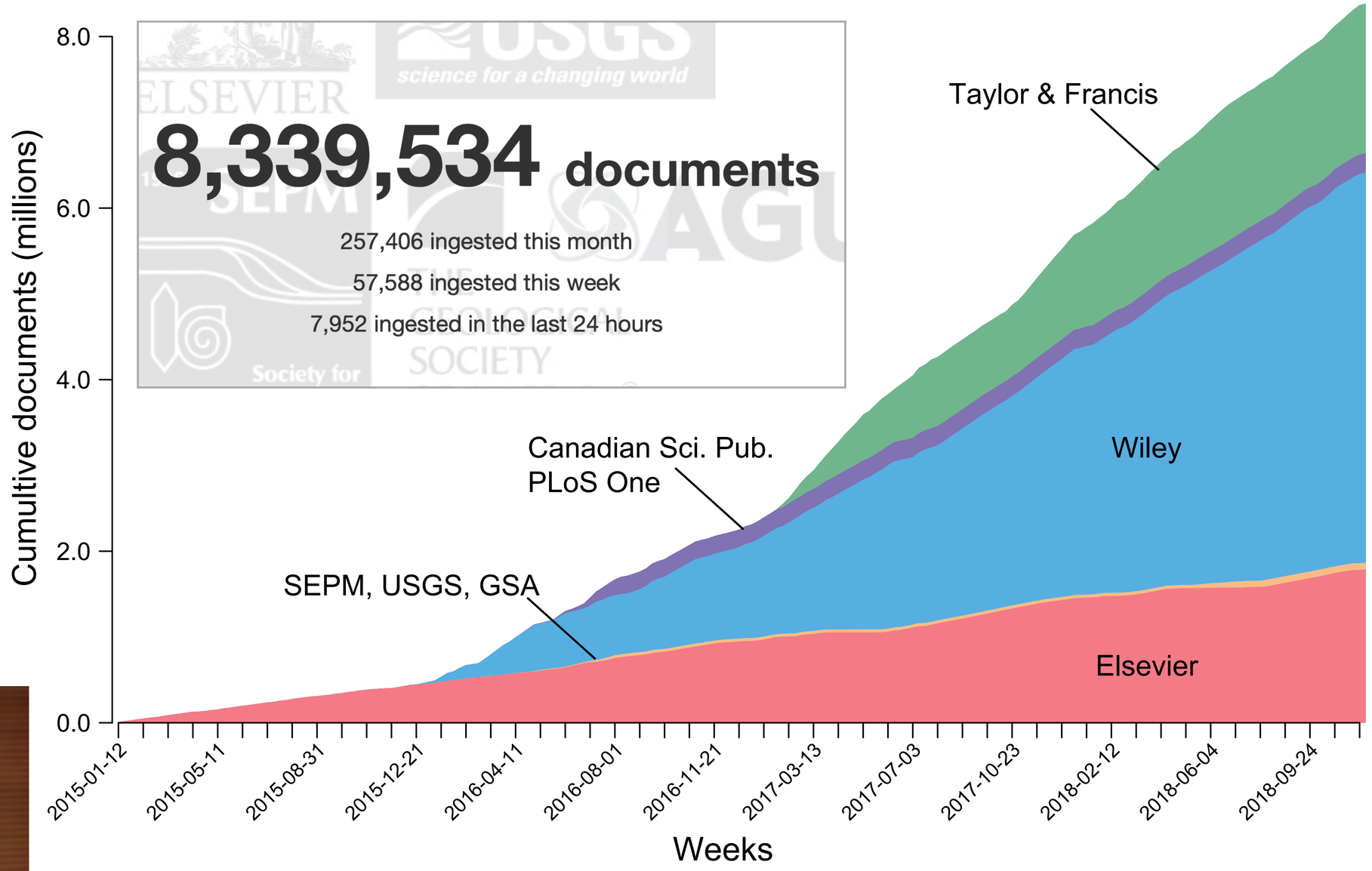
GeoDeepDive: key components



- Permissive, custom agreements with content owners that enable local storage, collaborative use.
- Automated system for acquiring documents and metadata from multiple providers.
- ElasticSearch for simple full-text word/phrase matching, cached results for large vocabularies.
- Harmonized document metadata for citation & linking to original content.
- Computing resources to rapidly re-analyze all documents.

NSF
USGS science for a changing world
support 2014-2018 by NSF-ICER 1343760
partial current support from USGS

✓ Principled access to scientific literature

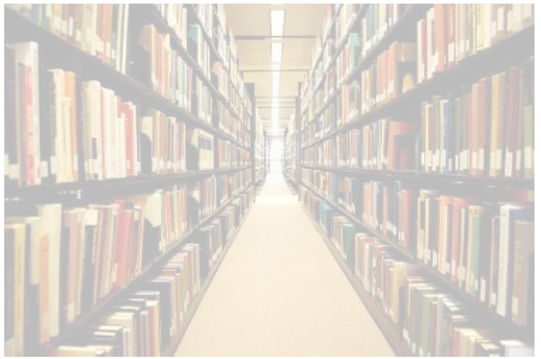


Miron Livny
Computer Sciences Dept.

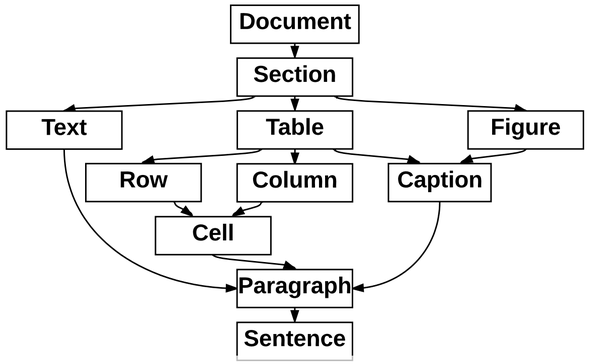
Ian Ross
Computer Sciences Dept.

COSMOS: required components

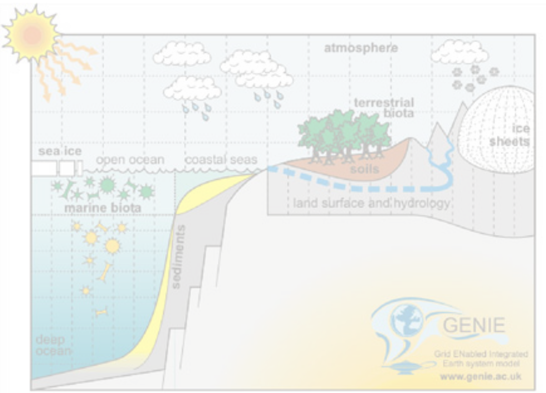
1. Principled, automated access to scientific publications and the computing capacity and infrastructure required to repeatedly analyze them.



2. Models and techniques to represent and capture multi-modal data within publications.



3. Earth system model with parameterizations and predictions that overlap with many different types of empirical data and observations in publications.



Fonduer: Information Extraction over Richly Formatted Data

Unstructured Information

Knowledge Base Construction

Structured Knowledge Base

TextRunner/ ReVerb

NELL

And many more…

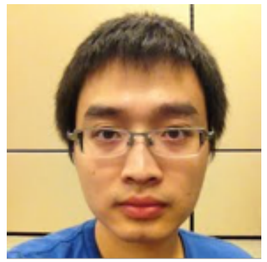
But, troves of "richly formatted" information remains untapped

Fonduer: Information Extraction over Richly Formatted Data

Our collaborators on Fonduer


Sen Wu


Luke Hsiao


Xiao Cheng


Braden Hancock


Philip Levis


Christopher Ré

Fonduer Users

FONDUER

Richly formatted data

Richly formatted data: information is expressed via textual, structural, tabular, and visual cues.

Transistor Datasheet (PDF)

SMBT3904...MMBT3904

NPN Silicon Switching Transistors

- High DC current gain: 0.1 mA to 100 mA
- Low collector-emitter saturation voltage

Maximum Ratings

Parameter	Symbol	Value	Unit
Collector-emitter voltage	V_{CEO}	40	V
Collector-base voltage	V_{CBO}	60	
Emitter-base voltage	V_{EBO}	6	
Collector current	I_C	200	mA
Total power dissipation	P_{tot}		mV
$T_S \leq 71°C$		330	
$T_S \leq 115°C$		250	
Junction temperature	T_j	150	°C
Storage temperature	T_{stg}	-65 ... 150	

Knowledge base construction from richly formatted data

Goal: extract maximum collector current from transistor datasheets

Transistor Datasheet

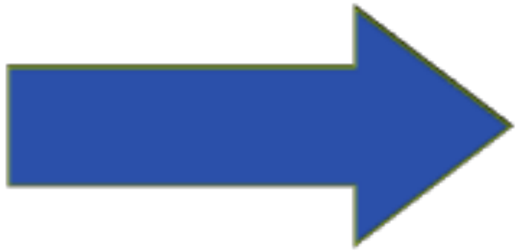
SMBT3904 .. MMBT3904

NPN Silicon Switching Transistors

- High DC current gain: 0.1 mA to 100 mA
- Low collector-emitter saturation voltage

Maximum Ratings

Parameter	Symbol	Value	Unit
Collector-emitter voltage	V_{CEO}	40	V
Collector-base voltage	V_{CBO}	60	
Emitter-base voltage	V_{EBO}	6	
Collector current	I_C	200	mA
Total power dissipation $T_S \leq 71°C$ $T_S \leq 115°C$	P_{tot}	 330 250	mV
Junction temperature	T_j	150	°C
Storage temperature	T_{stg}	-65 ... 150	

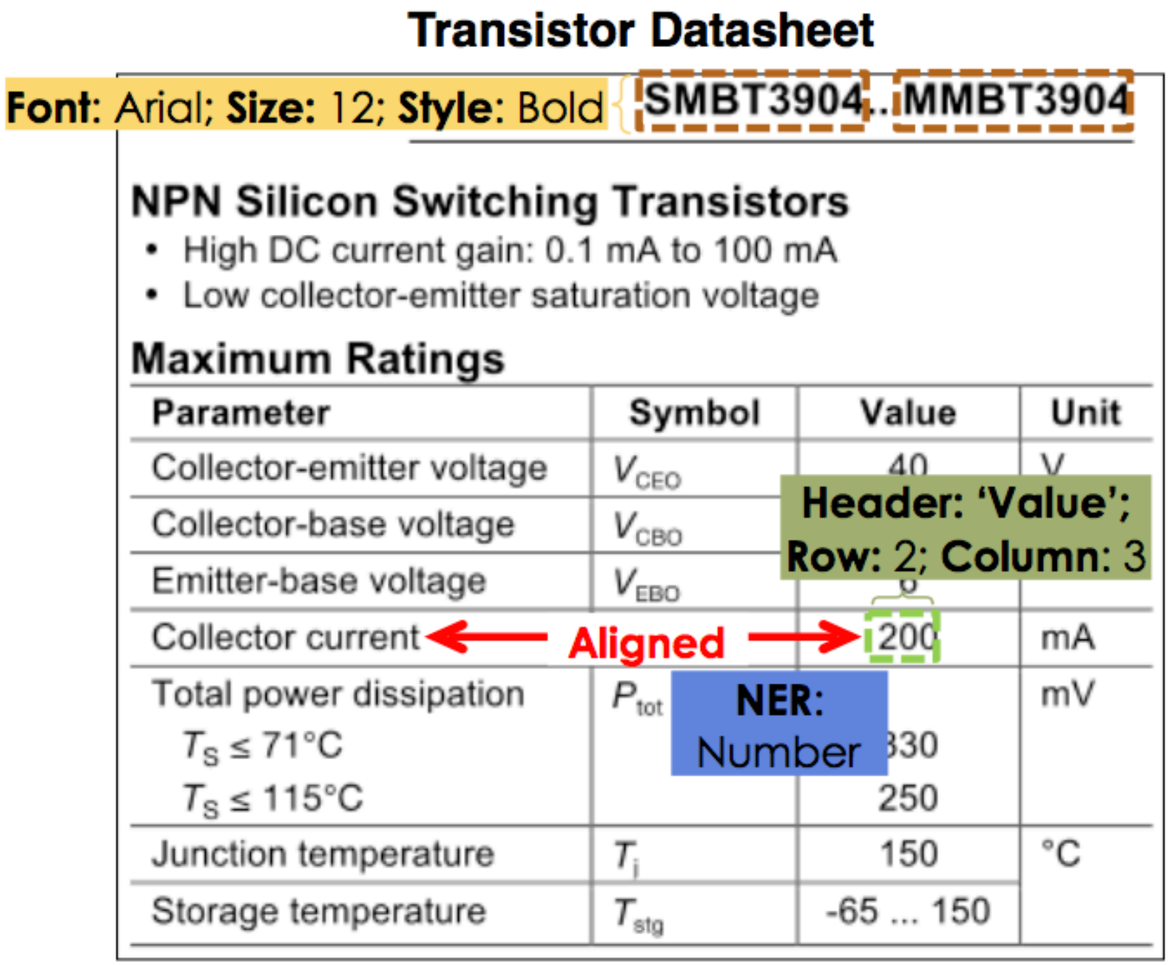


HasCollectorCurrent

Transistor Part	Current
SMBT3904	200mA
MMBT3904	200mA

Knowledge Base

Knowledge base construction from richly formatted data



In richly formatted data, semantics are expressed in textual, structural, tabular, and visual modalities throughout a document

Knowledge base construction from richly formatted data

Transistor Datasheet

SMBT3904...MMBT3904
NPN Silicon Switching Transistors
High DC current gain: 0.1 mA to 100 mA
Low collector-emitter saturation voltage
Maximum Ratings
Parameter Symbol Value Unit
Collector-emitter voltage VCEO 40 V
Collector-base voltage VCBO 60
Emitter-base voltage VEBO 6
Collector current IC 200 mA
Total power dissipation Ptot mV
TS ≤ 71°C 330
TS ≤ 115°C 250
Junction temperature Tj 150 °C
Storage temperature Tstg -65 ... 150

In richly formatted data, semantics are expressed in textual, structural, tabular, and visual modalities throughout a document

Conventional approach 1: Filter out other modalities besides unstructured text

Knowledge base construction from richly formatted data

Transistor Datasheet

Header SMBT3904.. MMBT3904

NPN Silicon Switching Transistors

- High DC current gain: 0.1 mA to 100 mA
- Low collector-emitter saturation voltage

Maximum Ratings **Table**

Parameter	Symbol	Value	Unit
Collector-emitter voltage	V_{CEO}	40	V
Collector-base voltage	V_{CBO}	60	
Emitter-base voltage	V_{EBO}	6	
Collector current	I_C	200	mA
Total power dissipation	P_{tot}		mV
$T_S \leq 71°C$		330	
$T_S \leq 115°C$		250	
Junction temperature	T_j	150	°C
Storage temperature	T_{stg}	-65 ... 150	

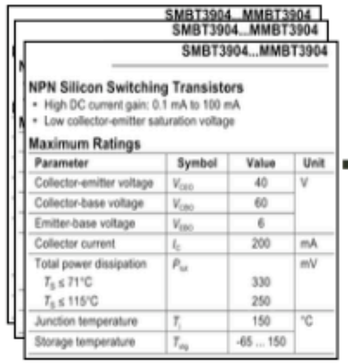
In richly formatted data, semantics are expressed in textual, structural, tabular, and visual modalities throughout a document

Conventional approach 1: Filter out other modalities besides unstructured text

Conventional approach 2: Limit the context scope to sentences or tables.

Problem: Misses important relations if you neglect multimodal information

Fonduer's pipeline


SMBT3904...MMBT3904

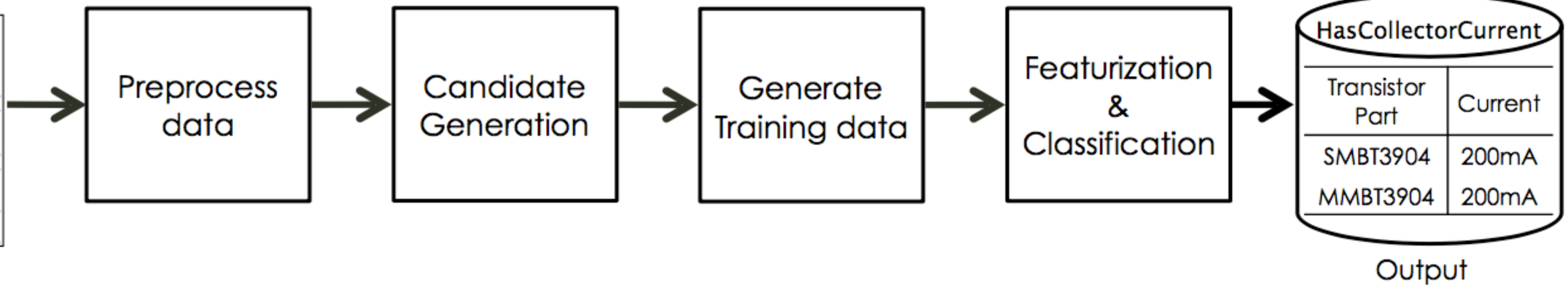
NPN Silicon Switching Transistors

- High DC current gain: 0.1 mA to 100 mA
- Low collector-emitter saturation voltage

Maximum Ratings

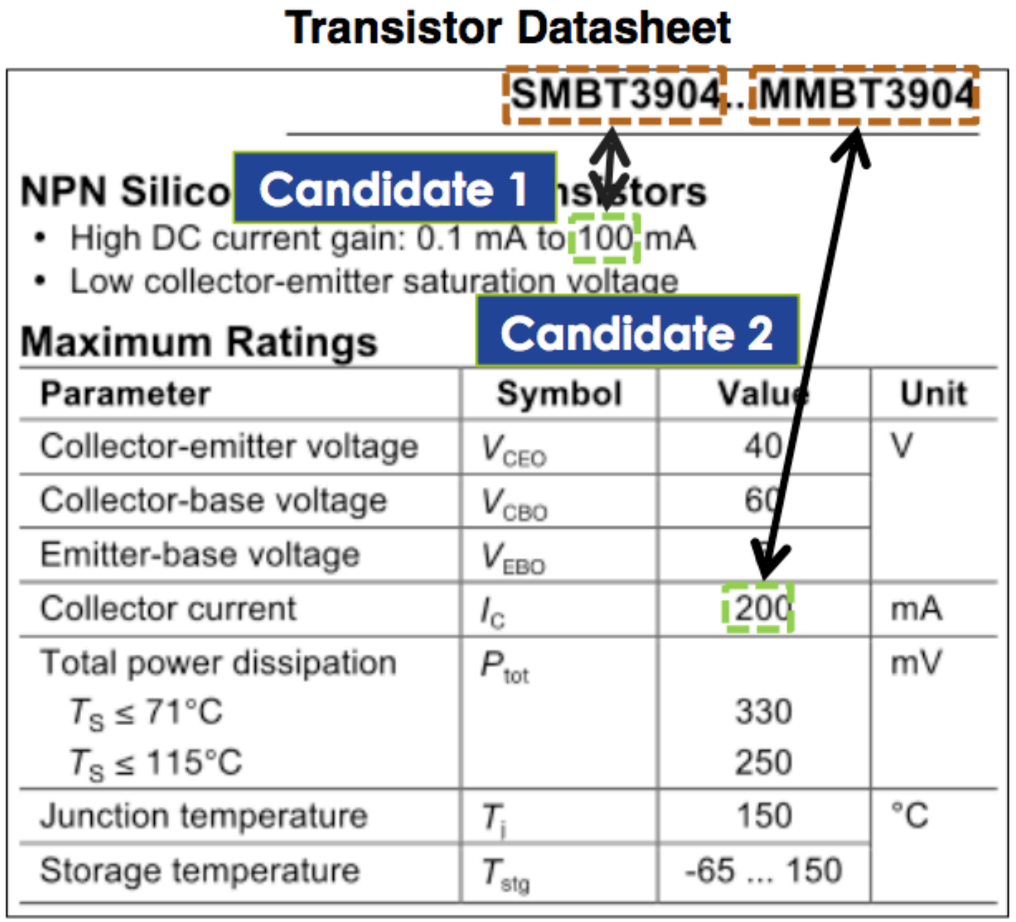
Parameter	Symbol	Value	Unit
Collector-emitter voltage	V_{CEO}	40	V
Collector-base voltage	V_{CBO}	60	
Emitter-base voltage	V_{EBO}	6	
Collector current	I_C	200	mA
Total power dissipation	P_{tot}		mV
$T_S \leq 71°C$		330	
$T_S \leq 115°C$		250	
Junction temperature	T_j	150	°C
Storage temperature	T_{stg}	-65 ... 150	

Data Input



Fonduer is a weakly supervised deep learning framework for knowledge base construction from richly formatted data

Multimodal weak supervision



Doc. level Candidates		Supervision	
		Manual	Labeling function
SMBT3904	100	✗	✗
MMBT3904	200	✓	✓

Weak supervision: express any supervision signal via labeling functions to generate training data

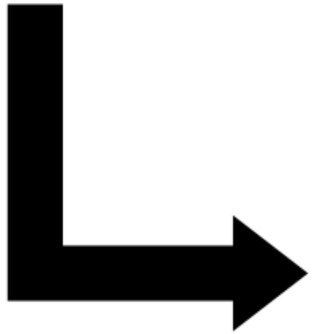
```
# Check if current is in the same row with keyword `collector`
def in_the_same_row_with(candidate):
  if 'collector' in
row_ngrams(candidate.current):
    return 1
  else: return -1
```

Multimodal weak supervision

Doc. level Candidates		Multimodal Supervision		
		Vertically aligned with 'Value'	Row ngrams contain 'mA'	'current' in sentence
SMBT3904	100	✗	∅	✓
SMBT3904	200	✓	✓	✗
SMBT3904	150	✓	✗	✗

∅=Abstain

Intuition: Use agreements / disagreements to learn the accuracy of LFs without ground truth



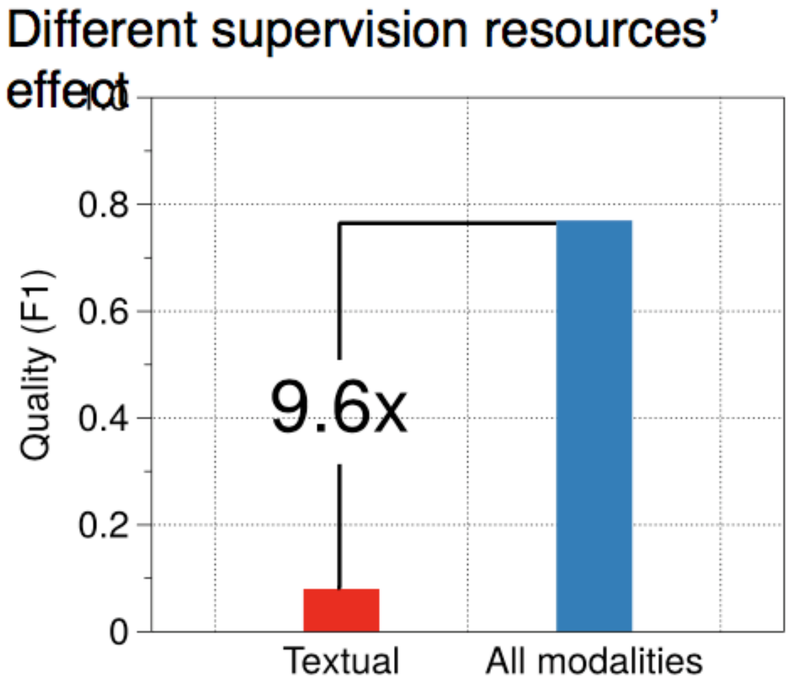
Data programming/MeTal

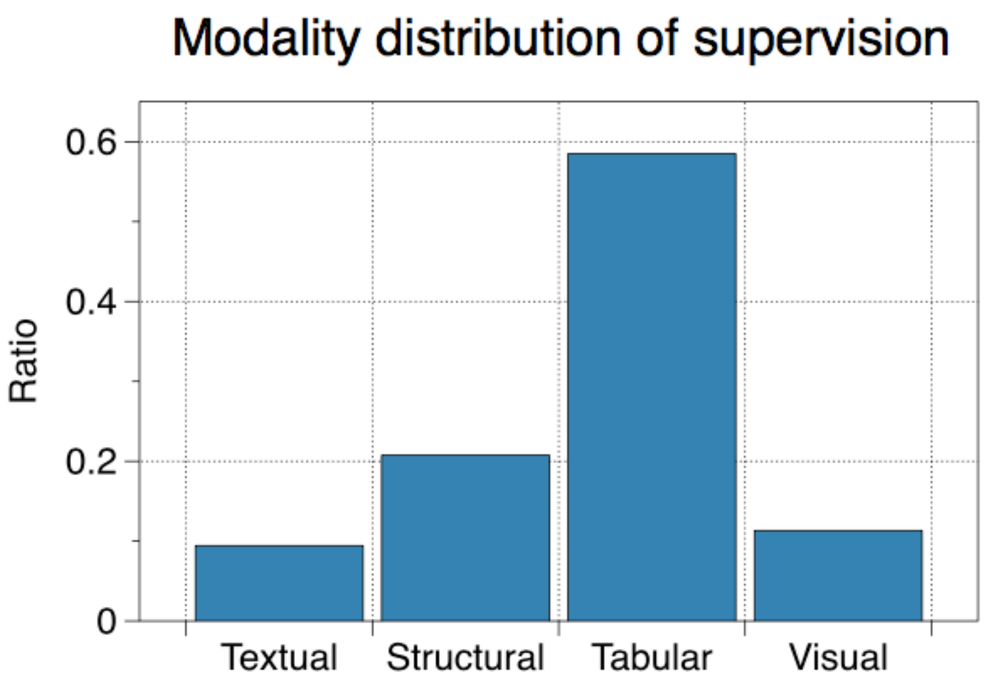
Output: Probabilistic Training Labels

SMBT3094	100	0.5
SMBT3094	200	0.85
SMBT3094	150	0.15

Multimodal supervision is key to quality

For transistor datasheets…





Users intuitively rely on multimodal information for supervision

Fonduer's data model

Richly formatted data

SMBT3904...MMBT3904

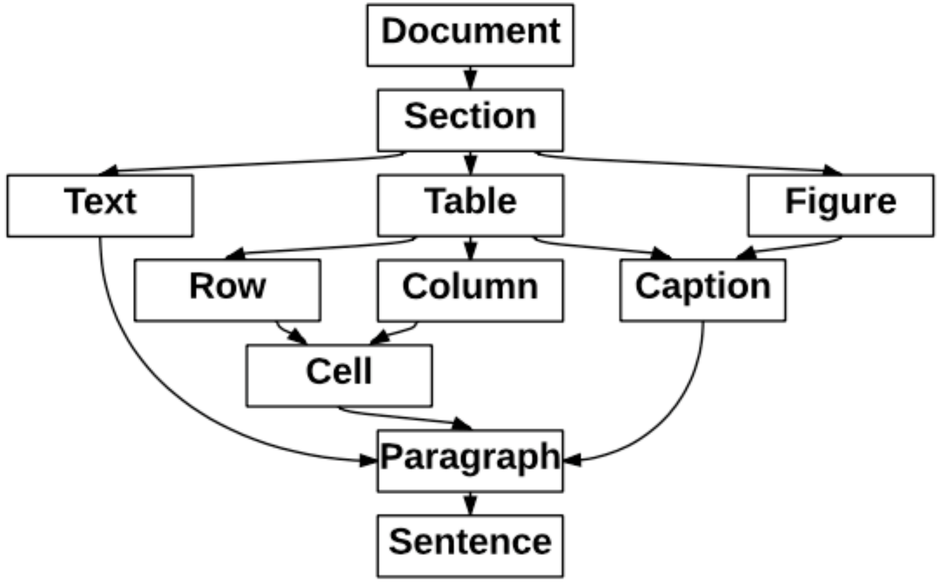
NPN Silicon Switching Transistors

- High DC current gain: 0.1 mA to 100 mA
- Low collector-emitter saturation voltage

Maximum Ratings

Parameter	Symbol	Value	Unit
Collector-emitter voltage	V_{CEO}	40	V
Collector-base voltage	V_{CBO}	60	
Emitter-base voltage	V_{EBO}	6	
Collector current	I_C	200	mA
Total power dissipation	P_{tot}		mV
$T_S \leq 71°C$		330	
$T_S \leq 115°C$		250	
Junction temperature	T_j	150	°C
Storage temperature	T_{stg}	-65 ... 150	

Data model



Fonduer automatically parses the richly formatted data into the data model that:

- ❑ Preserves structure/semantics across modalities
- ❑ Unifies a diverse variety of formats and styles
- ❑ Serves as the formal representation in KBC

Weakly Supervised KBC

- Fonduer helps build high-quality KBs from richly formatted data
- Allows users to leverage multimodal signals
- Augments LSTMs with features from each data modality to achieve high quality
- Fonduer is supporting real world applications

From Raw Documents to Fonduer's data model

Table 5
Proposal of ecological groups' (EGs) reassignments of AMBI index, based on Indicator Value (IndVal) coefficient and pollution condition of estuarine areas. Legend: Group 1 (estuaries with undisturbed or low disturbance conditions) and Group 2 (estuaries with medium to high disturbance conditions).

Indicator Value (IndVal) significant	Pollution condition (groups)	EG AZTI list	EG used for this study
40–60	Group 01	IV or V	III
		III	II
	Group 02	I or II	III
		III	IV
60–80	Group 01	IV or V	II
		III	V
	Group 02	I or II	IV
		III	I
80–100	Group 01	IV or V	I
	Group 02	I or II	V

correlation coefficients were calculated using the BIOESTAT v5.0 program (Ayres et al., 2007). Except for Spearman's rank correlation, the level of significance in all statistical analyzes was $\alpha = 5\%$.

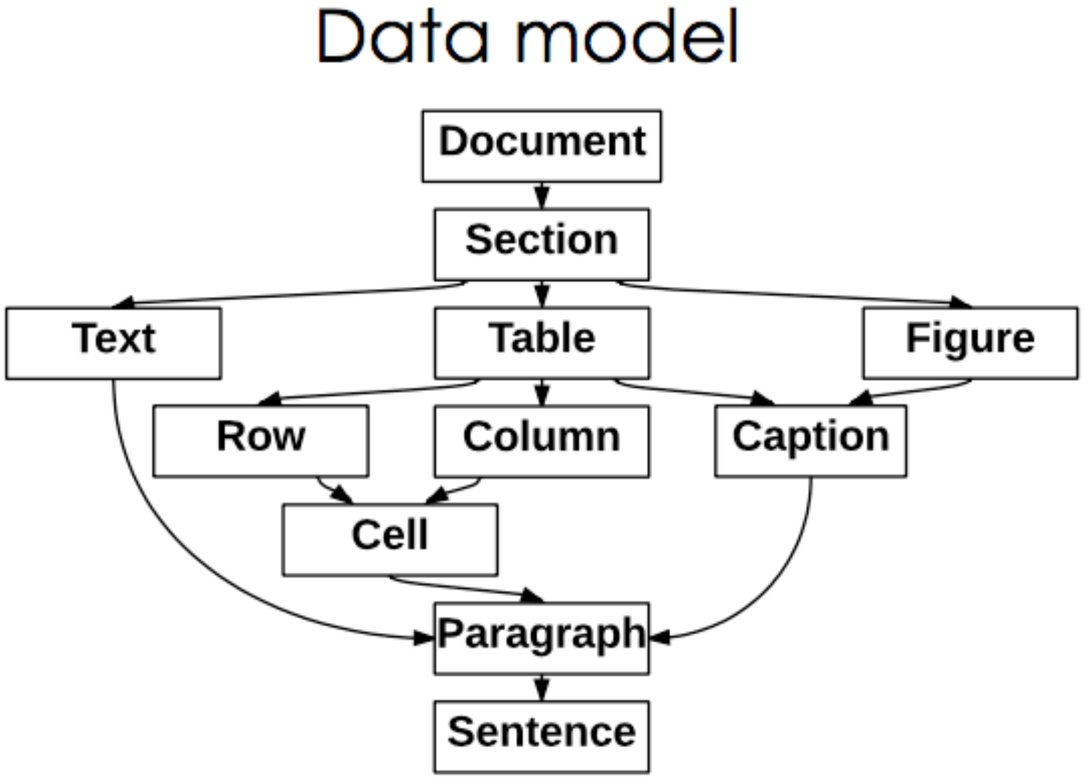
3. Results

3.1. Environmental data

Water parameters: Mean temperature ranged from 25.7 (±1.3) (in Mamucabas) to 30.5 (±1.9) °C (in Maracaípe), with little variation between seasons. Mean salinity values were similar to those obtained in the period of macrofaunal samplings (October 2007), with the majority of the sampling sites situated between polyhaline-euhaline zones. Salinity values ranged from a minimum of 11 (±5.5) (in Jaboatão) to a maximum of 33.5 (±3.7) psu (in Maracaípe) among seasons, excepted for Mamucabas and Pirapama, with low salinity values (oligohaline zones). At most sites, dissolved oxygen levels were found to be outside the normal limits for estuarine systems. On the other hand, only Pina Basin 1 (3.99 μmol l^{-1}), Jaboatão (6.31 μmol l^{-1}) and Paratibe (7.29 μmol l^{-1}) had higher ammonia-N

However, the majority (82%) was ascribed to an EG based on the classification for the same genus. For the following species, ecological groups were attributed according to the AZTI list for higher taxonomic levels (>family): *Anomalocardia brasiliana* (I), *Barantolla* sp. (V), *Capitellides* sp. (V), *Fabrisabella* sp. (I), *Megalomma* sp. (I), *Neomediomastus* sp. (V), *Pseudobranchiomma* sp. (I) and *Timarete* sp. (IV). Due to the lack of ecological information for tropical regions, twenty-one taxa remained without classification and were denominated as "not assigned".

Considering the definition of sites into groups using dissolved oxygen and disturbance levels, the IndVal coefficient revealed fifteen significant indicator species/taxa (Table 6). However, based on IndVal scale proposed, only four species had high indicator values (>40%): the polychaetes *Capitella* sp. and *Streblospio* sp., nematodes and the oligochaete *Tectidrilus* sp. In terms of ecological interpretation, *Capitella* sp. and *Tectidrilus* sp. were originally classified as EG_V on the AZTI list. However, the EG of this latter species was changed to EG_{II} based on its presence in unpolluted conditions. Both Nematoda and *Streblospio* sp. [originally considered tolerant (EG_{III})] were associated to sensitive and opportunistic groups by



Constructing Fonduer's input requires performing document segmentation.

Typical OCR focuses on text only!

Problem: Text-only segmentation obscures important table structure

Table 1. EnKF calibrated biogeochemical parameters in the GENIE-1 model.

Name	Prior assumptions (mean and range[a])	Posterior mean[b]	Description
$u_0^{PO_4}$	1.65 μmol kg^{-1} yr^{-1} (0.3–3.0)	1.91 μmol kg^{-1} yr^{-1}	maximum PO_4 uptake (removal) rate (Eq. 3)
K^{PO_4}	0.2 μmol kg^{-1} (0.1–0.3)	0.21 μmol kg^{-1}	PO_4 Michaelis-Menton half-saturation concentration (Eq. 3)
r^{POC}	0.05 (0.02–0.08)	0.055	initial proportion of POC export as fraction #2 (Eq. 6)
l^{POC}	600 m (200–1000)	556 m	e-folding remineralization depth of POC fraction #1 (Eq. 6)
$r_0^{CaCO_3:POC}$	0.036 (0.015–0.088)[c]	0.022[d]	$CaCO_3$:POC: export rain ratio scalar (Eq. 8)
η	1.5 (1.0–2.0)	1.28	thermodynamic calcification rate power (Eq. 9)
r^{CaCO_3}	0.4 (0.2–0.6)	0.489	initial proportion of $CaCO_3$ export as fraction #2 (Eq. 11)
l^{CaCO_3}	600 m (200–1000)	1055	e-folding remineralization depth of $CaCO_3$ fraction #1 (Eq. 11)

[a] the range is quoted as 1 standard deviation either side of the mean
[b] quoted as the mean of the entire EnKF ensemble
[c] assimilation was carried out on a $\log_{10}$ scale
[d] Note that the rain ratio scalar parameter is not the same as the $CaCO_3$:POC export rain ratio as measured at the base of the euphotic zone, because $r_0^{CaCO_3:POC}$ is further multiplied by $(\Omega-1)^{\eta}$ to calculate the rain ratio, where Ω is the surface ocean saturation state with respect to calcite (see Sect. 2.1). Pre-industrial mean ocean surface Ω is ~5.2 in the GENIE-1 model, so that the global $CaCO_3$:POC export rain ratio can be estimated using the 8-parameter assimilation as being equal to $(5.2-1)^{1.28} \times 0.022 = 0.14$.

(Table 1). Because we explicitly resolve the individual "components" (i.e., C, ^{13}C, P, ...) of organic matter, the GENIE-1 model can be used to quantify the effect of fractionation between the components of organic matter during remineral-

(e.g. Zhang et al., 2001, 2003) or by allowing the tracer transport of negative O_2 concentrations (e.g., Hotinski et al., 2001). We treat the remineralization of dissolved organic matter in an analogous manner if O_2 availability is insuffi-

Typical OCR focuses on text only!

Problem: Text-only segmentation obscures important table structure

Solution: Cast document segmentation as a vision problem

Table 1. EnKF calibrated biogeochemical parameters in the GENIE-1 model.

Name	Prior assumptions (mean and range[a])	Posterior mean[b]	Description
$u_0^{PO_4}$	1.65 μmol kg^{-1} yr^{-1} (0.3–3.0)	1.91 μmol kg^{-1} yr^{-1}	maximum PO_4 uptake (removal) rate (Eq. 3)
K^{PO_4}	0.2 μmol kg^{-1} (0.1–0.3)	0.21 μmol kg^{-1}	PO_4 Michaelis-Menton half-saturation concentration (Eq. 3)
r^{POC}	0.05 (0.02–0.08)	0.055	initial proportion of POC export as fraction #2 (Eq. 6)
l^{POC}	600 m (200–1000)	556 m	*e*-folding remineralization depth of POC fraction #1 (Eq. 6)
$r_0^{CaCO_3:POC}$	0.036 (0.015–0.088)[c]	0.022[d]	$CaCO_3$:POC: export rain ratio scalar (Eq. 8)
η	1.5 (1.0–2.0)	1.28	thermodynamic calcification rate power (Eq. 9)
r^{CaCO_3}	0.4 (0.2–0.6)	0.489	initial proportion of $CaCO_3$ export as fraction #2 (Eq. 11)
l^{CaCO_3}	600 m (200–1000)	1055	*e*-folding remineralization depth of $CaCO_3$ fraction #1 (Eq. 11)

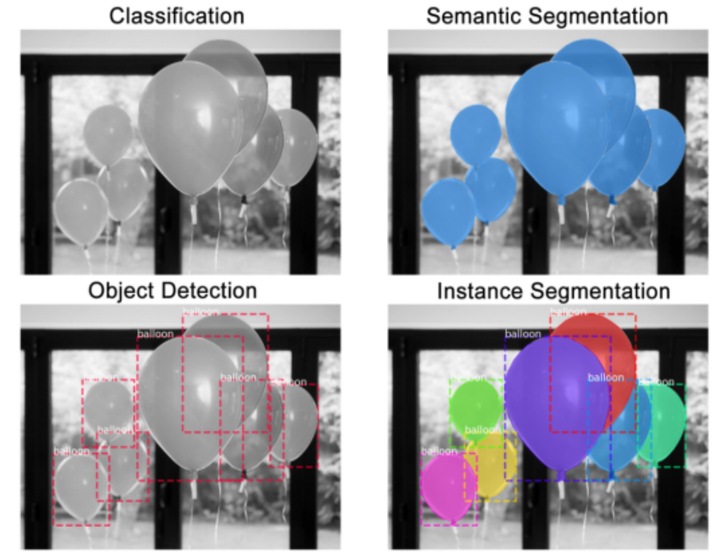
[a] the range is quoted as 1 standard deviation either side of the mean
[b] quoted as the mean of the entire EnKF ensemble
[c] assimilation was carried out on a $\log_{10}$ scale
[d] Note that the rain ratio scalar parameter is not the same as the $CaCO_3$:POC export rain ratio as measured at the base of the euphotic zone, because $r_0^{CaCO_3:POC}$ is further multiplied by $(\Omega-1)^{\eta}$ to calculate the rain ratio, where Ω is the surface ocean saturation state with respect to calcite (see Sect. 2.1). Pre-industrial mean ocean surface Ω is ~5.2 in the GENIE-1 model, so that the global $CaCO_3$:POC export rain ratio can be estimated using the 8-parameter assimilation as being equal to $(5.2-1)^{1.28} \times 0.022 = 0.14$.

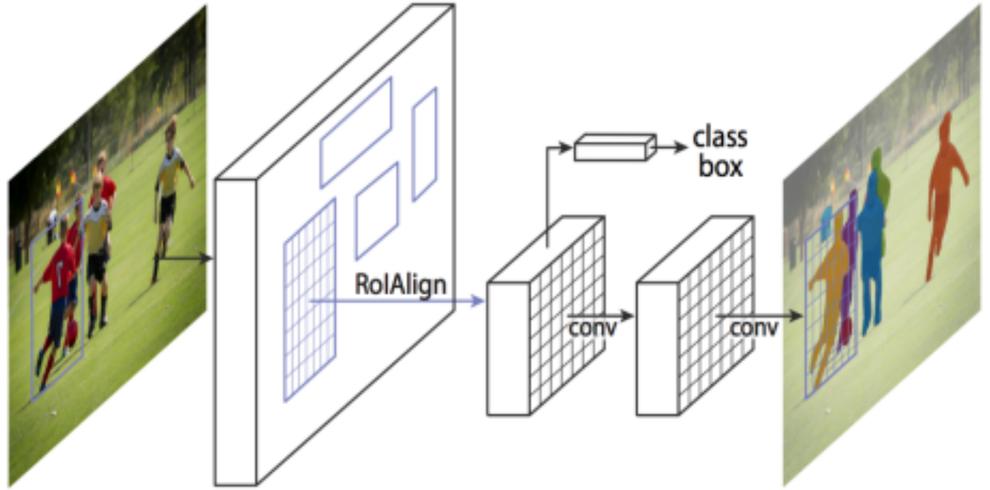
(Table 1). Because we explicitly resolve the individual "components" (i.e., C, ^{13}C, P, ...) of organic matter, the GENIE-1 model can be used to quantify the effect of fractionation between the components of organic matter during remineral-

(e.g. Zhang et al., 2001, 2003) or by allowing the tracer transport of negative O_2 concentrations (e.g., Hotinski et al., 2001). We treat the remineralization of dissolved organic matter in an analogous manner if O_2 availability is insuffi-

Mask-RCNN (He et al. 2018)

- State of the art image segmentation model
- Performs instance segmentation
- Identifies regions of interest (ROI) followed by object detection, classification, and mask detection in parallel
- We set mask to the bounding box with intention to refine bounding box





Mask R-CNN framework. Source: https://arxiv.org/abs/1703.06870

Data annotation: In-house labeler

Image tagger | *Click + drag to create item. Click existing item to adjust.* | Save | Clear changes | Next image

Table 1. Characteristics of 18 soil samples collected in FSSW-1 and FSSW-2

Parameter	Median	Range
Silt/clay content	36%	13–62%
"Sand" content	57%	34–69%
Organic matter content of "sand"	37%	10–75%
Total organic carbon	10%	1–34%
pH	5.8	5–6.1

tional Pesticides Monitoring Program of the U.S. Fish and Wildlife Service, involving the removal of feet, beaks, wingtips, and skin. Whole body concentrations of pesticides were then determined.

Surveys of birds and small mammals were performed by the Manomet Bird Observatory. Efforts were focused on identifying breeding pairs and nests of birds, live-trapping small mammals, and developing species inventories.

RESULTS

Site characteristics

The study area consisted of forested and shrubbed swamp/wetland whose soils are comprised of silty-sand, rich in organic matter (Table 1). The median value for total organic carbon (TOC) was 10% with a range of 1 to 34%; soils closer to the river tended to exhibit higher TOC values. The pH of the soils ranged from 5.0 to 6.1.

Concentrations of chemicals in soils

A 1985 investigation under Superfund [2] revealed elevated concentrations of various pesticides and selected metals such as arsenic in surface soils (Table 2). Highest concentrations occurred in the FSSW-1 area and the lowest in the on-site reference area, FSSW-3. 1,1,1-trichloro-2,2-bis-(*p*-chlorophenyl)ethane (DDT) and its metabolites (herein collectively referred to as DDTR) and chlordane were the chlorinated pesticides found in highest concentrations. The three composite and two hot spot soil samples (DB-3 and DB-12) collected in 1988 for the laboratory bioassays exhibited pesticide levels similar to those seen in 1985 (Fig. 2). Again, DDTR and chlordane were the predominant pesticides. A sample of soil collected in one of the more contaminated areas in FSSW-1 (near DB-3) in the 1989 sampling effort revealed elevated levels of arsenic (700 mg/kg) and total polynuclear aromatic hydrocarbons (63 mg/kg). Thirty-eight samples were analyzed for pesticides in 1989 as part

Table 2. Historical surface soil concentrations from 1985

Compound	Average concn. (mg/kg dry wt.) [Maximum concn.] FSSW-1[a]	FSSW-2[a]	FSSW-3[a]
Pesticides			
4,4′-DDD	70 [1,100]	4 [28]	1 [12]
4,4′-DDE	10 [47]	1 [5]	1 [2]
4,4′-DDT	61 [630]	2 [10]	1 [4]
Chlordane	143 [1,700]	17 [110]	2 [17]
Dieldrin	4 [32]	1 [1]	
PAHs			
Benzo[*a*]pyrene	1 [2]	1 [2]	
Fluoranthene	1 [5]		1 [1]
Total other PAHs	1 [22]		3 [7]
Metals			
Arsenic	80 [1,000]	17 [144]	20 [70]
Beryllium	1 [1]	1 [2]	1 [2]
Cadmium	1 [1]	1 [3]	1 [1]
Nickel	4 [25]	6 [21]	9 [32]
Lead	50 [721]	48 [128]	108 [215]
Silver	1 [4]		1 [1]
Zinc	50 [355]	32 [148]	71 [102]
Dioxin (2,3,7,8-TCDD)	0.003 [0.048]	0.00037 [0.001]	
Number of sampled stations	43	15	9

[a]Sample location.

Image tagger | *Click + drag to create item. Click existing item to adjust.* | Save | Clear changes | Next image

Page Header

Table Caption

Table 1. Characteristics of 18 soil samples collected in FSSW-1 and FSSW-2

Table

Parameter	Median	Range
Silt/clay content	36%	13–62%
"Sand" content	57%	34–69%
Organic matter content of "sand"	37%	10–75%
Total organic carbon	10%	1–34%
pH	5.8	5–6.1

Body Text

tional Pesticides Monitoring Program of the U.S. Fish and Wildlife Service, involving the removal of feet, beaks, wingtips, and skin. Whole body concentrations of pesticides were then determined.

Surveys of birds and small mammals were performed by the Manomet Bird Observatory. Efforts were focused on identifying breeding pairs and nests of birds, live-trapping small mammals, and developing species inventories.

Section Header

RESULTS

Section Header

Site characteristics

Body Text

The study area consisted of forested and shrubbed swamp/wetland whose soils are comprised of silty-sand, rich in organic matter (Table 1).

Body Text

The median value for total organic carbon (TOC) was 10% with a range of 1 to 34%; soils closer to the river tended to exhibit higher TOC values. The pH of the soils ranged from 5.0 to 6.1.

Section Header

Concentrations of chemicals in soils

Body Text

A 1985 investigation under Superfund [2] revealed elevated concentrations of various pesticides and selected metals such as arsenic in surface soils (Table 2). Highest concentrations occurred in the FSSW-1 area and the lowest in the on-site reference area, FSSW-3. 1,1,1-trichloro-2,2-bis-(*p*-chlorophenyl)ethane (DDT) and its metabolites (herein collectively referred to as DDTR) and chlordane were the chlorinated pesticides found in highest concentrations. The three composite and two hot spot soil samples (DB-3 and DB-12) collected in 1988 for the laboratory bioassays exhibited pesticide levels similar to those seen in 1985 (Fig. 2). Again, DDTR and chlordane were the predominant pesticides. A sample of soil collected in one of the more contaminated areas in FSSW-1 (near DB-3) in the 1989 sampling effort revealed elevated levels of arsenic (700 mg/kg) and total polynuclear aromatic hydrocarbons (63 mg/kg). Thirty-eight samples were analyzed for pesticides in 1989 as part

Table Caption

Table 2. Historical surface soil concentrations from 1985

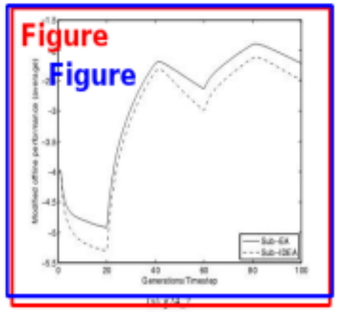
Table

Compound	Average concn. (mg/kg dry wt.) [Maximum concn.] FSSW-1[a]	FSSW-2[a]	FSSW-3[a]
Pesticides			
4,4′-DDD	70 [1,100]	4 [28]	1 [12]
4,4′-DDE	10 [47]	1 [5]	1 [2]
4,4′-DDT	61 [630]	2 [10]	1 [4]
Chlordane	143 [1,700]	17 [110]	2 [17]
Dieldrin	4 [32]	1 [1]	
PAHs			
Benzo[*a*]pyrene	1 [2]	1 [2]	
Fluoranthene	1 [5]		1 [1]
Total other PAHs	1 [22]		3 [7]
Metals			
Arsenic	80 [1,000]	17 [144]	20 [70]
Beryllium	1 [1]	1 [2]	1 [2]
Cadmium	1 [1]	1 [3]	1 [1]
Nickel	4 [25]	6 [21]	9 [32]
Lead	50 [721]	48 [128]	108 [215]
Silver	1 [4]		1 [1]
Zinc	50 [355]	32 [148]	71 [102]
Dioxin (2,3,7,8-TCDD)	0.003 [0.048]	0.00037 [0.001]	
Number of sampled stations	43	15	9

Table Note

[a]Sample location.

Qualitative Results



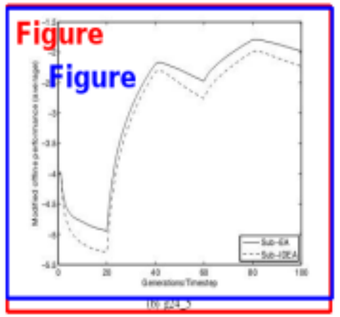


Fig. 5: Modified offline performance of Sub-EA and Sub-IDEA averaged over 30 runs

TABLE IV: Mean best-of-generation values (averaged over 30 runs)

	Sub-EA	Sub-IDEA
g24_2	-2.2130	-2.4890
g24_5	-2.2436	-2.2322

Table
Table

VI. Summary and Future Work

This paper highlights the benefits of Infeasibility Driven Evolutionary Algorithm (IDEA) for dynamic, constrained single objective optimization problems. The presence of infeasible solutions allow IDEA to approach the constrained optimum from the infeasible side as well as feasible side of the search space, thereby converging faster than conventional EAs which approach the optimum from feasible side only. The paper provides results of preliminary studies of the algorithm on two dynamic, constrained single objective optimization benchmarks. The results of using IDEA as a sub-evolve mechanism are certainly encouraging for the above problems. Its performance is currently being studied extensively for available constrained dynamic optimization problems.

The comparison of proposed algorithm has been made with a structurally similar algorithm in order to highlight the benefits of maintaining infeasible solutions for dynamic optimization problems. Currently studies are underway to compare the performance of IDEA with other existing algorithms.

Acknowledgment

The presented work was supported by grants from Defence and Security Applications Research Center (DSARC), Australian Defence Force Academy, University of New South Wales, Australia.

The work of Trung Thanh Nguyen is supported by the Overseas Research Scheme Award (ORS) and the School of Computer Science, University of Birmingham.

The work of Xin Yao is supported by an EPSRC grant (EP/E058884/1) on "Evolutionary Algorithms for Dynamic Optimisation Problems: Design, Analysis and Applications".

References

[1] H. K. Singh, A. Isaacs, T. Ray, and W. Smith, "Infeasibility Driven Evolutionary Algorithm (IDEA) for Engineering Design Optimization," in *Proceedings of 21st Australasian Joint Conference on Artificial Intelligence AI-08*, 2008, pp. 104–115.

[2] T. Ray, H. K. Singh, A. Isaacs, and W. Smith, "Infeasibility driven evolutionary algorithm for constrained optimization," in *Constraint Handling in Evolutionary Optimization*, ser. Studies in Computational Intelligence. Springer, in press.

[3] I. Hatzakis and D. Wallace, "Dynamic multi-objective optimization with evolutionary algorithms: a forward-looking approach," in *GECCO '06: Proceedings of the 8th annual conference on Genetic and evolutionary computation*. New York, NY, USA: ACM, 2006, pp. 1201–1208.

[4] T. Nguyen and X. Yao, "Benchmarking and solving dynamic constrained problems," in *IEEE Congress on Evolutionary Computation (CEC) 2009*, Accepted.

[5] J. Branke, "Memory enhanced evolutionary algorithms for changing optimization problems," in *Evolutionary Computation, 1999. CEC 99. Proceedings of the 1999 Congress on*, vol. 3, 1999, pp. 1874–1882.

[6] S. Baluja, "Population-based incremental learning: A method for integrating genetic search based function optimization and competitive learning," Carnegie Mellon University, Pittsburgh, PA, USA, Tech. Rep., 1994.

[7] S. Yang and X. Yao, "Experimental study of population-based incremental learning algorithms for dynamic optimization problems," *Soft Computing: A Fusion of Foundations, Methodologies and Applications*, vol. 9, no. 11, 2005.

[8] S. Yang, "Non-stationary problem optimization using the primal-dual genetic algorithm," *The 2003 Congress on Evolutionary Computation (CEC)*, vol. 3, pp. 2246–2253, December 2003.

[9] N. Mori, H. Kita, and Y. Nishikawa, "Adaption to a changing environment by means of the thermodynamical genetic algorithm," *Lecture Notes in Computer Science*, vol. 1141, pp. 513–522, 1996.

[10] P. Moscato, "On evolution, search, optimization, genetic algorithms and martial arts: Towards memetic algorithms," California Institute of Technology, Tech. Rep., 1989.

[11] T. Ray, A. Isaacs, and W. Smith, "A Memetic Algorithm for Dynamic Multiobjective Optimization," in *Multi-objective Memetic Algorithms*, ser. Studies in Computational Intelligence. Springer, 2008 (In Press).

[12] S. Yang and X. Yao, "Population-based incremental learning with associative memory for dynamic environments," *IEEE Transactions on Evolutionary Computation*, vol. 12, no. 5, pp. 542–561, Oct. 2008.

[13] K. Deb, A. Pratap, S. Agarwal, and T. Meyarivan, "A fast and elitist multiobjective genetic algorithm: NSGA-II," *Evolutionary Computation, IEEE Transactions on*, vol. 6, pp. 182–197, 2002.

[14] K. Deb and S. Agrawal, "Simulated binary crossover for continuous search space," *Complex Systems*, vol. 9, pp. 115–148, 1995.

name	# blocks	# variables # global	# variables avg./block	# generated function terms
Driver "A"	3903	1	1.0	2840
Driver "B"	3922	2	2.3	2951
Driver "C"	3925	2	2.0	2860
Driver "D"	4487	2	2.0	3124
Driver "E"	3933	4	4.0	2868
Driver "F"	4519	6	9.2	37365
Driver "G"	4521	5	13.4	4395
Driver "H"	6700	18	39.5	14612
Driver "I"	5429	3	4.3	9744
Driver "J"	8693	1	1.0	7250
Driver "K"	4569	7	20.7	29984

Table
Table

Fig. 3. Measurements from a preliminary investigation of the number of function terms generated for some small test programs. The richer boolean programs that these programs encode contain procedures with parameters and local state, which causes the blocks in the encoding to have varying numbers of variables on entry. The table shows both the number of global variables and the average number of variables per block. The boolean programs can contain unreachable blocks, which explains the fact that the number of function terms is sometimes smaller than the number of blocks.

the program's initial state, are given symbolically (k and m in the running example). If, instead of $\neg A(k, m)$, a complete set of explicit boolean values are used, as in:

$$\neg A(\text{false},\text{false}) \wedge \neg A(\text{false},\text{true}) \wedge \neg A(\text{true},\text{false}) \wedge \neg A(\text{true},\text{true})$$

Equation

then all function-term arguments will also be explicit boolean values, so there are only $N \cdot 2^K$ function terms, a *single* exponential.

There is a good reason we don't want to abandon the symbolic initial values in favor of the explicit ones: by using explicit values, we get *at least* 2^K function terms, because that's how many function terms we get for the start block alone. The numbers in Figure 3 show that the symbolic initial values can do better than that.

Interestingly enough, we can adjust the degree to which we use the two argument representations, by using the following simple equality: for any function b, expression e, and lists of expressions E_0 and E_1, we have:

$$b(E_0, e, E_1) = (\neg e \wedge b(E_0, \text{false}, E_1)) \wedge (e \wedge b(E_0, \text{true}, E_1)) \qquad (12)$$

Equation
Equation

Thus, if e is an expression other than an explicit boolean value, then the algorithm in Figure 2 can choose to [illegible] right-hand side of (12) instead of invoking the procedure *Instantiate*. The choice of which one to do would be determined heuristically. A possible heuristic is to use (12) whenever the argument e is "too complicated", as perhaps when the number of variables in e exceeds some threshold, or when e is anything but the symbolic initial value of the program variable corresponding to this function argument. By choosing the latter heuristic, for example, the number of different function terms is $N \cdot 3^K$, a single exponential as in the case of using only explicit boolean values; yet, by using this heuristic, the algorithm begins with just one negated start block function, not an exponential number of them as in the case of using only explicit boolean values.

I have yet to experiment with the symbolic-versus-explicit argument representations in my implementation.

Prediction
Ground Truth

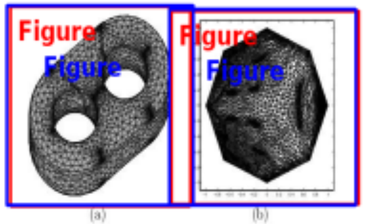


Figure 11: (a)Surface with genus 2 (b)Parameterization on $aba^{-1}b^{-1}cdc^{-1}d^{-1}$

disk $\mathbf{P}$ to $\mathcal{M}$ then all the 2D-corner vertices $w_1, w_2, ..., w_{2g-1}$ of $\mathbf{P}$ map to the same vertex Ω of the surface mesh $\mathcal{M}$. More precisely, the basepoint Ω is mapped $2g$ times if we deal with genus g. That means we have $2g$ different indices but those $2g$ points all have the same coordinates. Similarly, the points w_1 and w_2 which are portrayed in Fig. 4(b) maps to the same 3D points A of the surface mesh $\mathcal{M}$. For that reason, the point A has to be repeated twice. The vertices of the polygonal disk are chosen as

$$[\cos(s\pi/2g), \sin(s\pi/2g)] \quad \text{forall} s = 0, 1, ...2g-1. \qquad (20)$$

Equation
Equation

6 Constrained quadrangulation

In this section, we will describe a way to decompose a surface $\mathcal{M}$ into pieces of four-sided domains. In order to facilitate the presentation, we suppose that we have a parametrization $\mathcal{P}$ in disposition and that $\mathcal{M}$ is of genus zero and thus the parameter domain is a rectangle.

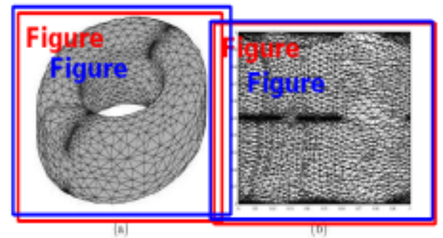
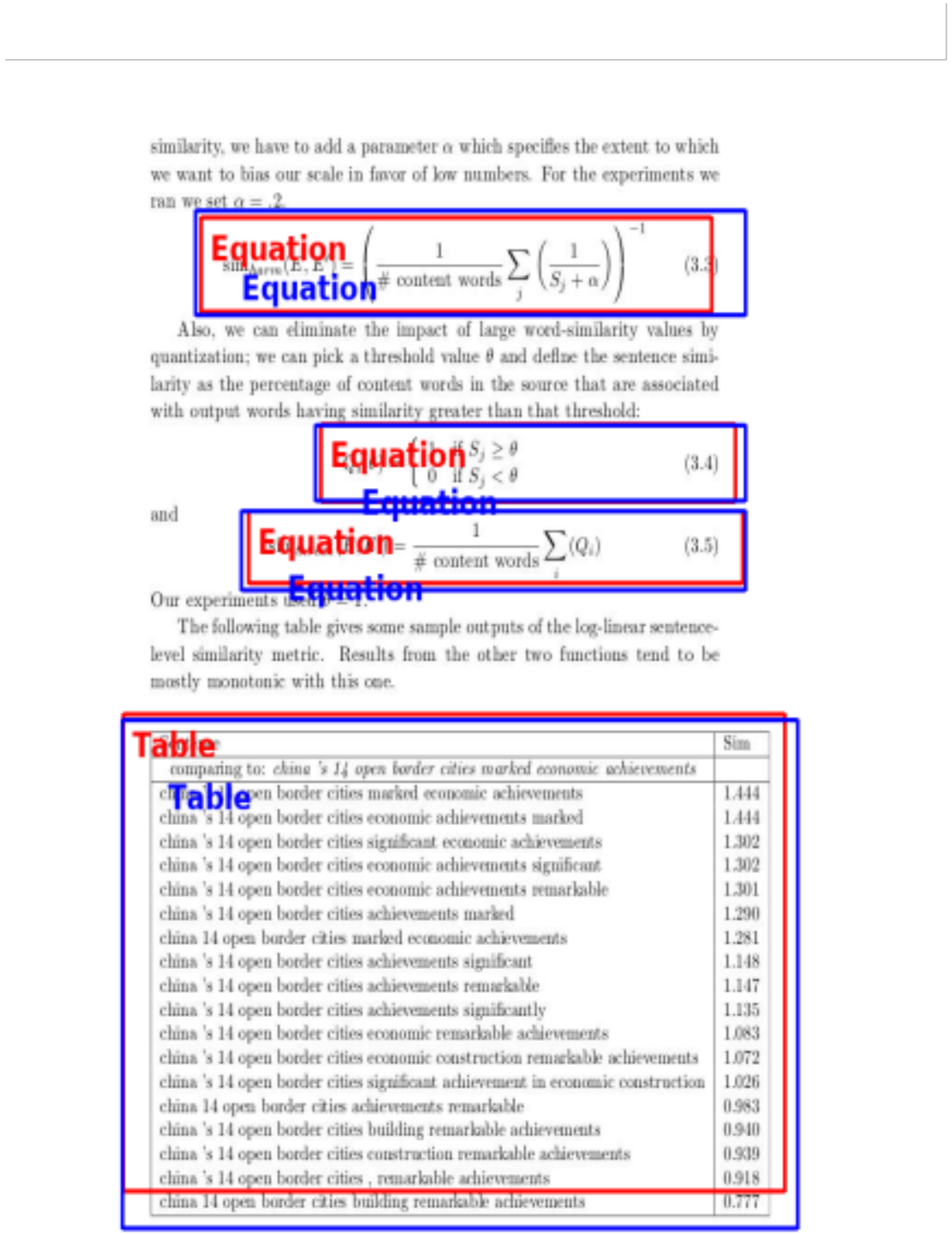


Figure 12: (a)Surface with genus 1 (b)Parameterization on $aba^{-1}b^{-1}$

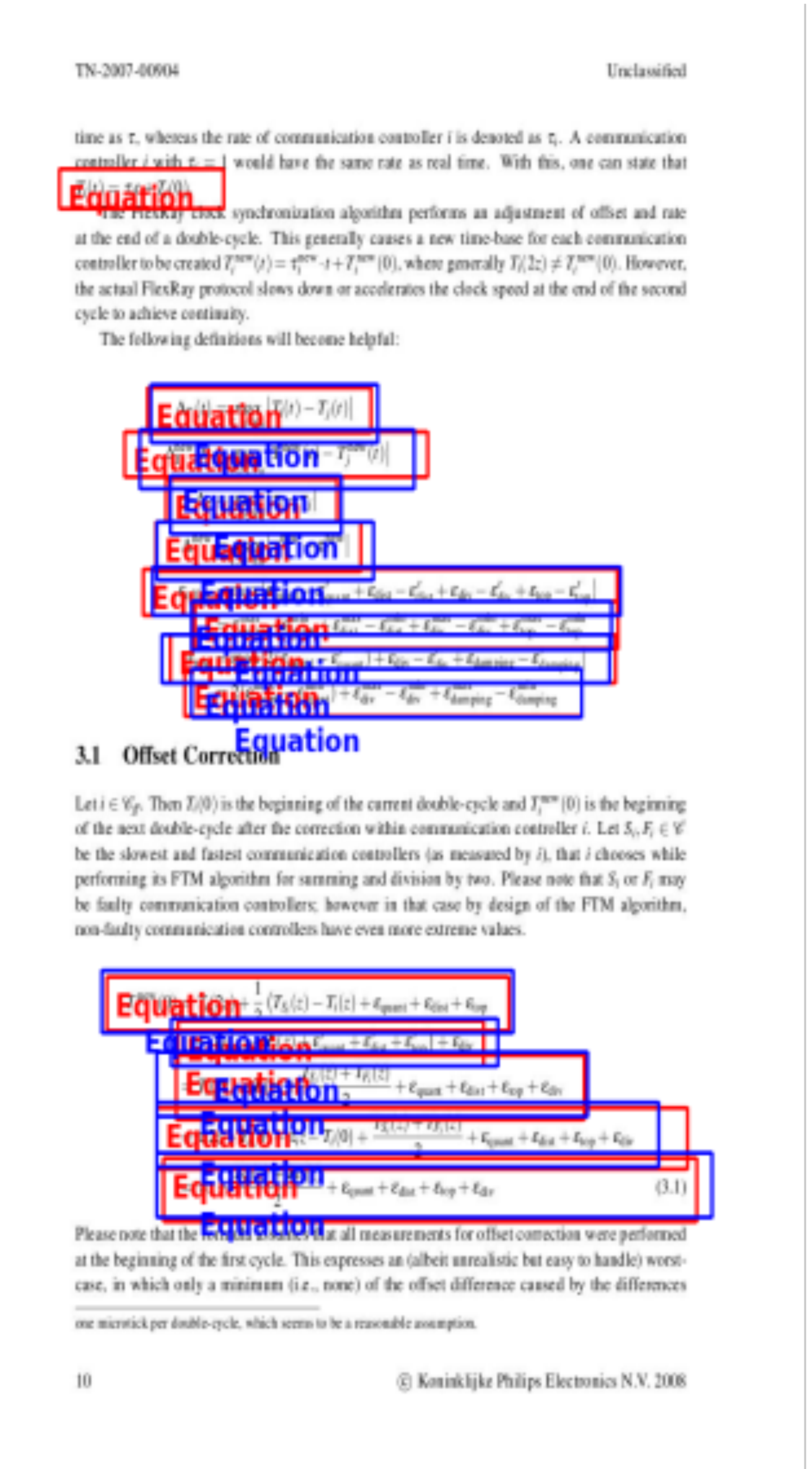
Qualitative Results

Prediction

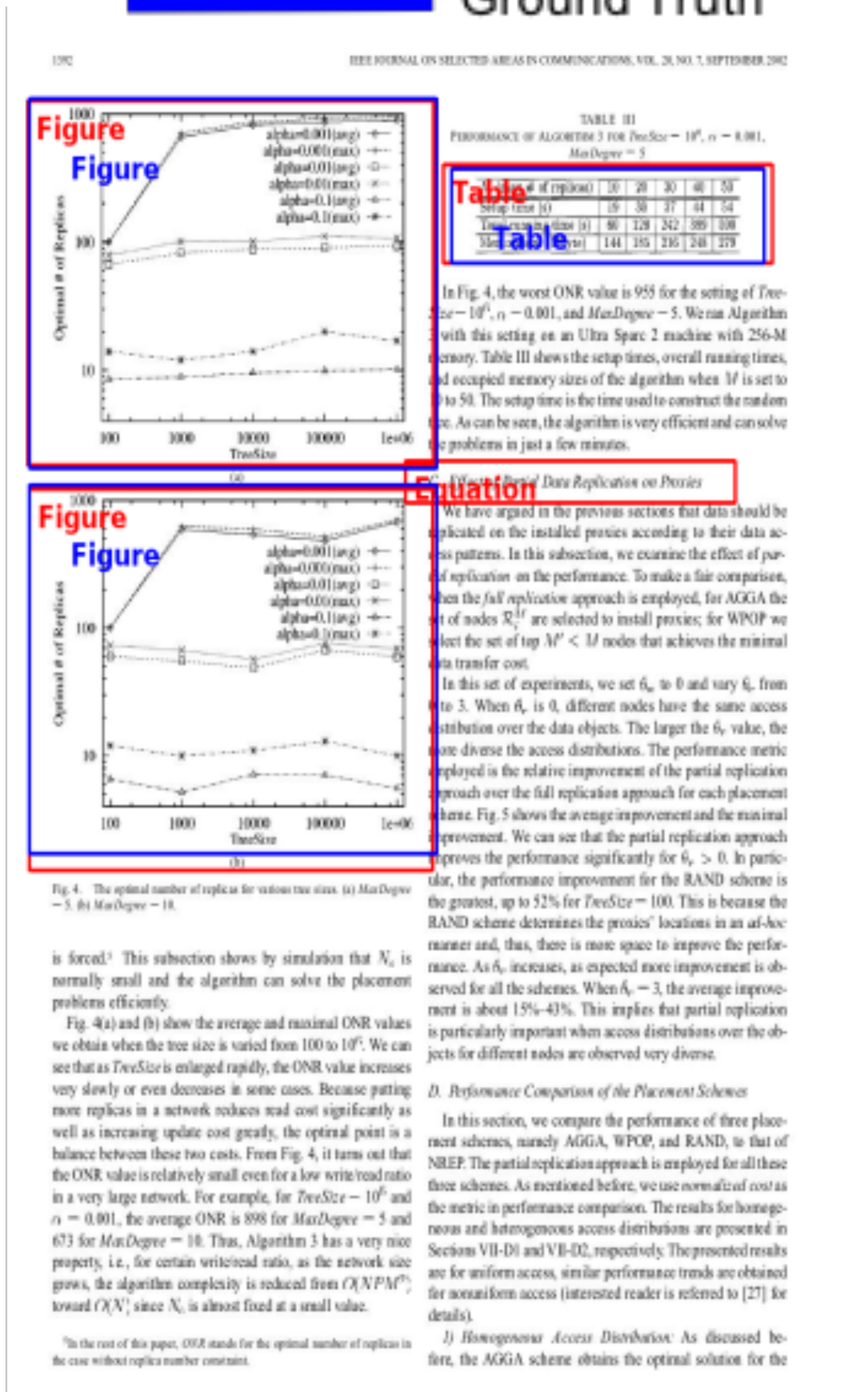
Ground Truth



Truncation



In-line equations

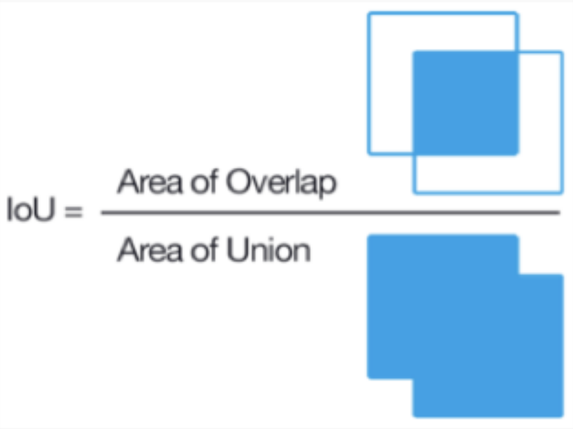


Inaccurate areas of interest

Quantitative Results

IoU = TP / (TP + FP + FN) (segmentation metric)

Average Precision (AP) -- TP / TP + FP
(classification metric)



Class Specific

Formula AP: 0.723
Formula IoU: 0.750

Figure AP: 0.734
Figure IoU: 0.768

Table AP: 0.888
Table IOU: 0.877

Next Steps: Improve Segmentation

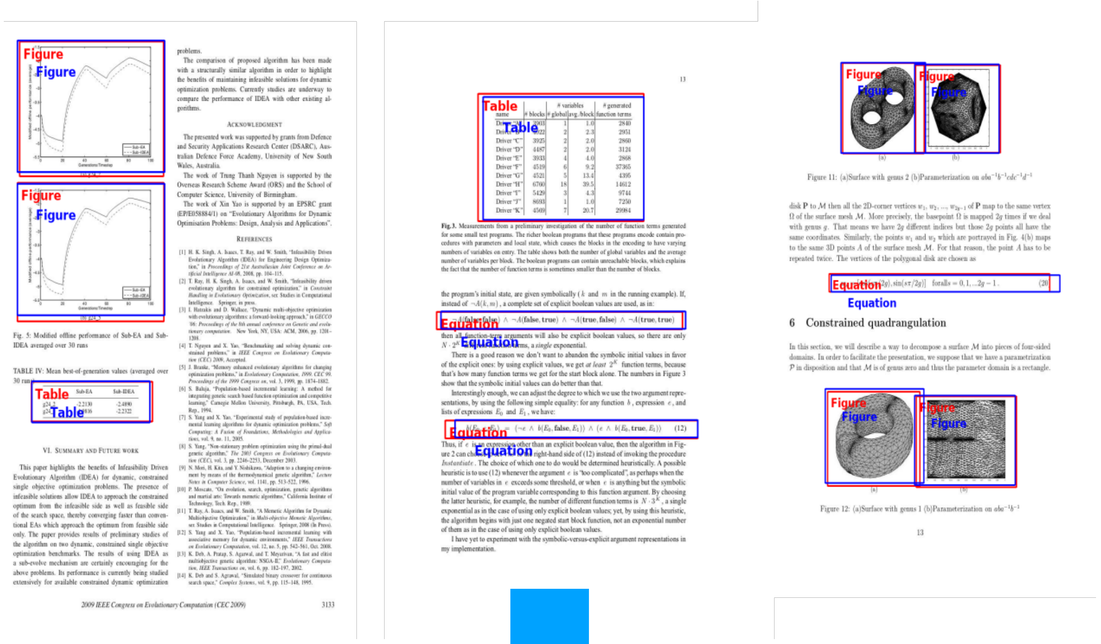
- Multi-modal segmentation
  - Core insight: we are not utilizing the actual text information in the document we're processing
    - IE: If a body text contains the word abstract, it's the abstract
  - Proposed solution:
    - When a region is proposed in the region proposal network, run OCR on the region and pass the text results to the next layer
- Better attention mechanism for document classification
  - Core insight: tradition segmentation is done in a rich pixel environment -- the background corresponds simply to non labelled objects (think a self driving car on the road)
  - Background in a document corresponds to null space. Partitioning null space is a challenge because there are many possible ways to validly partition it
  - Proposed solution:
    - Compensate for a lack of internal ROI info by providing additional context to an ROI
    - Attend to areas directly surrounding the ROI, then feed that output to the head.

Next Steps: OCR within individual segments

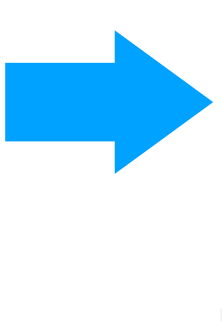
Name	Prior assumptions (mean and range[a])	Posterior mean[b]	Description
$u_0^{PO_4}$	1.65 μmol kg^{-1} yr^{-1} (0.3–3.0)	1.91 μmol kg^{-1} yr^{-1}	maximum PO_4 uptake (removal) rate (Eq. 3)
K^{PO_4}	0.2 μmol kg^{-1} (0.1–0.3)	0.21 μmol kg^{-1}	PO_4 Michaelis-Menton half-saturation concentration (Eq. 3)
r^{POC}	0.05 (0.02–0.08)	0.055	initial proportion of POC export as fraction #2 (Eq. 6)
l^{POC}	600 m (200–1000)	556 m	*e*-folding remineralization depth of POC fraction #1 (Eq. 6)
$r_0^{CaCO_3:POC}$	0.036 (0.015–0.088)[c]	0.022[d]	$CaCO_3$:POC: export rain ratio scalar (Eq. 8)
η	1.5 (1.0–2.0)	1.28	thermodynamic calcification rate power (Eq. 9)
r^{CaCO_3}	0.4 (0.2–0.6)	0.489	initial proportion of $CaCO_3$ export as fraction #2 (Eq. 11)
l^{CaCO_3}	600 m (200–1000)	1055	*e*-folding remineralization depth of $CaCO_3$ fraction #1 (Eq. 11)

Extract text modality to construct Fonduers data model

From Raw Documents to Fonduer's data model

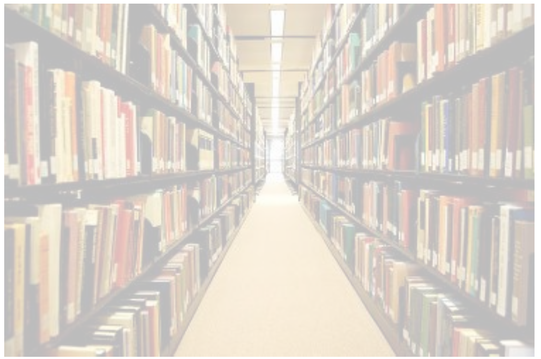


Name	Prior assumptions (mean and range[a])	Posterior mean[b]	Description
$u_0^{PO_4}$	1.65 μmol kg^{-1} yr^{-1} (0.3–3.0)	1.91 μmol kg^{-1} yr^{-1}	maximum PO_4 uptake (removal) rate (Eq. 3)
K^{PO_4}	0.2 μmol kg^{-1} (0.1–0.3)	0.21 μmol kg^{-1}	PO_4 Michaelis-Menton half-saturation concentration (Eq. 3)
r^{POC}	0.05 (0.02–0.08)	0.055	initial proportion of POC export as fraction #2 (Eq. 6)
l^{POC}	600 m (200–1000)	556 m	e-folding remineralization depth of POC fraction #1 (Eq. 6)
$r_0^{CaCO_3:POC}$	0.036 (0.015–0.088)[c]	0.022[d]	$CaCO_3$:POC: export rain ratio scalar (Eq. 8)
η	1.5 (1.0–2.0)	1.28	thermodynamic calcification rate power (Eq. 9)
r^{CaCO_3}	0.4 (0.2–0.6)	0.489	initial proportion of $CaCO_3$ export as fraction #2 (Eq. 11)
l^{CaCO_3}	600 m (200–1000)	1055	e-folding remineralization depth of $CaCO_3$ fraction #1 (Eq. 11)

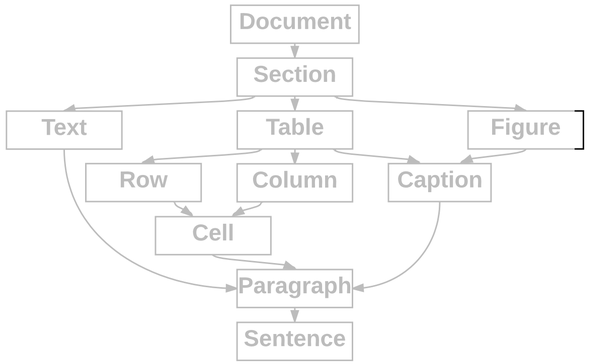


COSMOS: required components

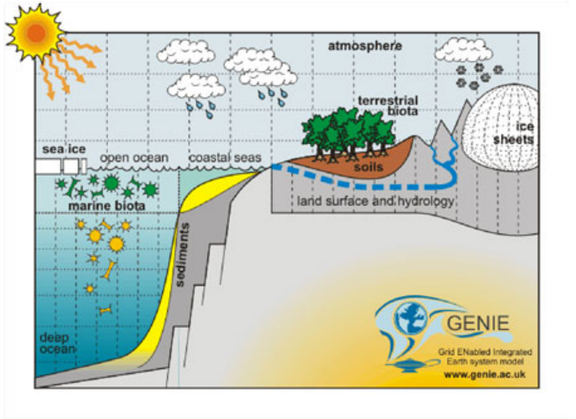
1. Principled, automated access to scientific publications and the computing capacity and infrastructure required to repeatedly analyze them.



2. Models and techniques to represent and capture multi-modal data within publications.



3. Earth system model with parameterizations and predictions that overlap with many different types of empirical data and observations in publications.



GENIE:
Grid ENabled Integrated Earth system model

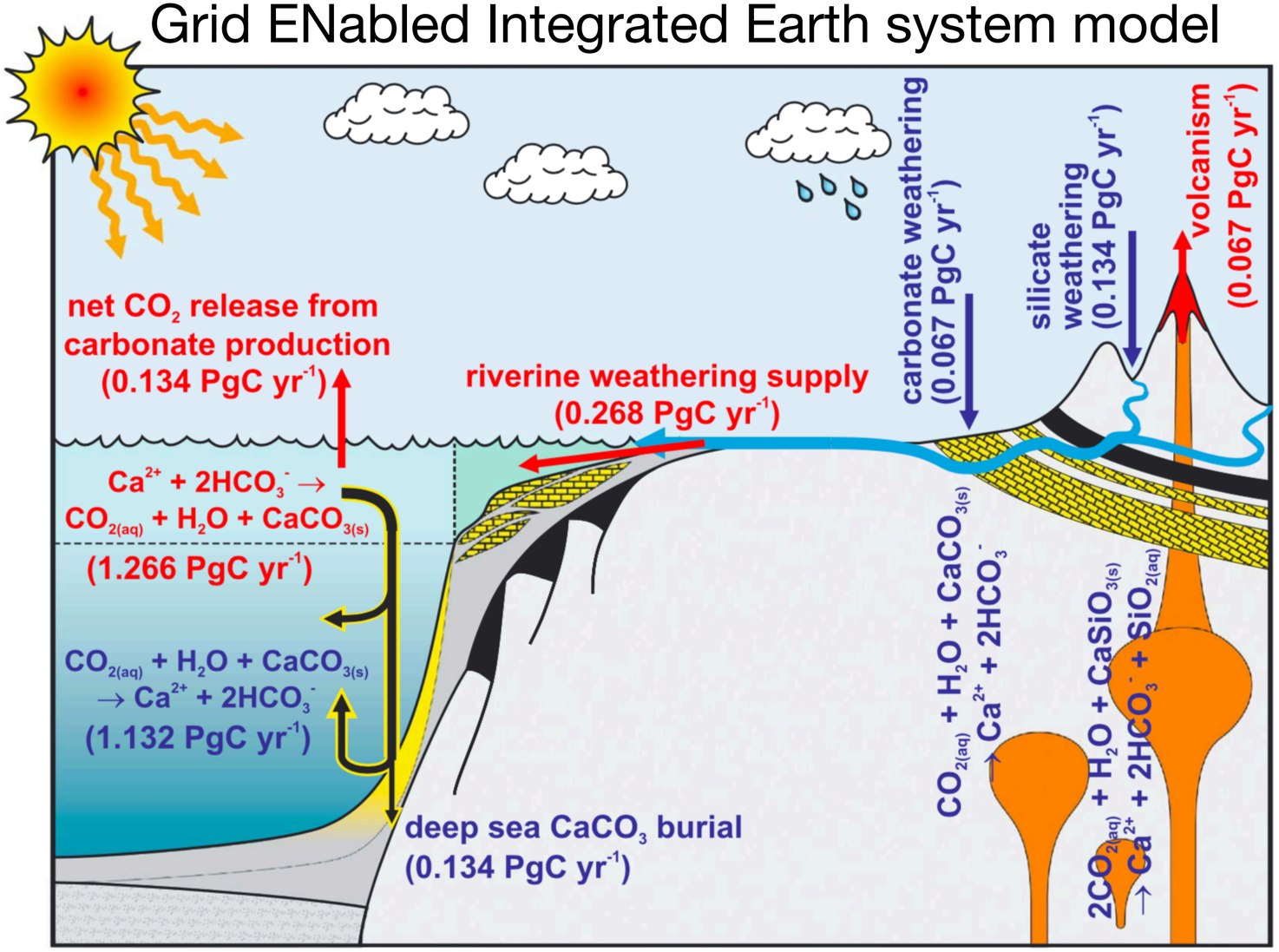
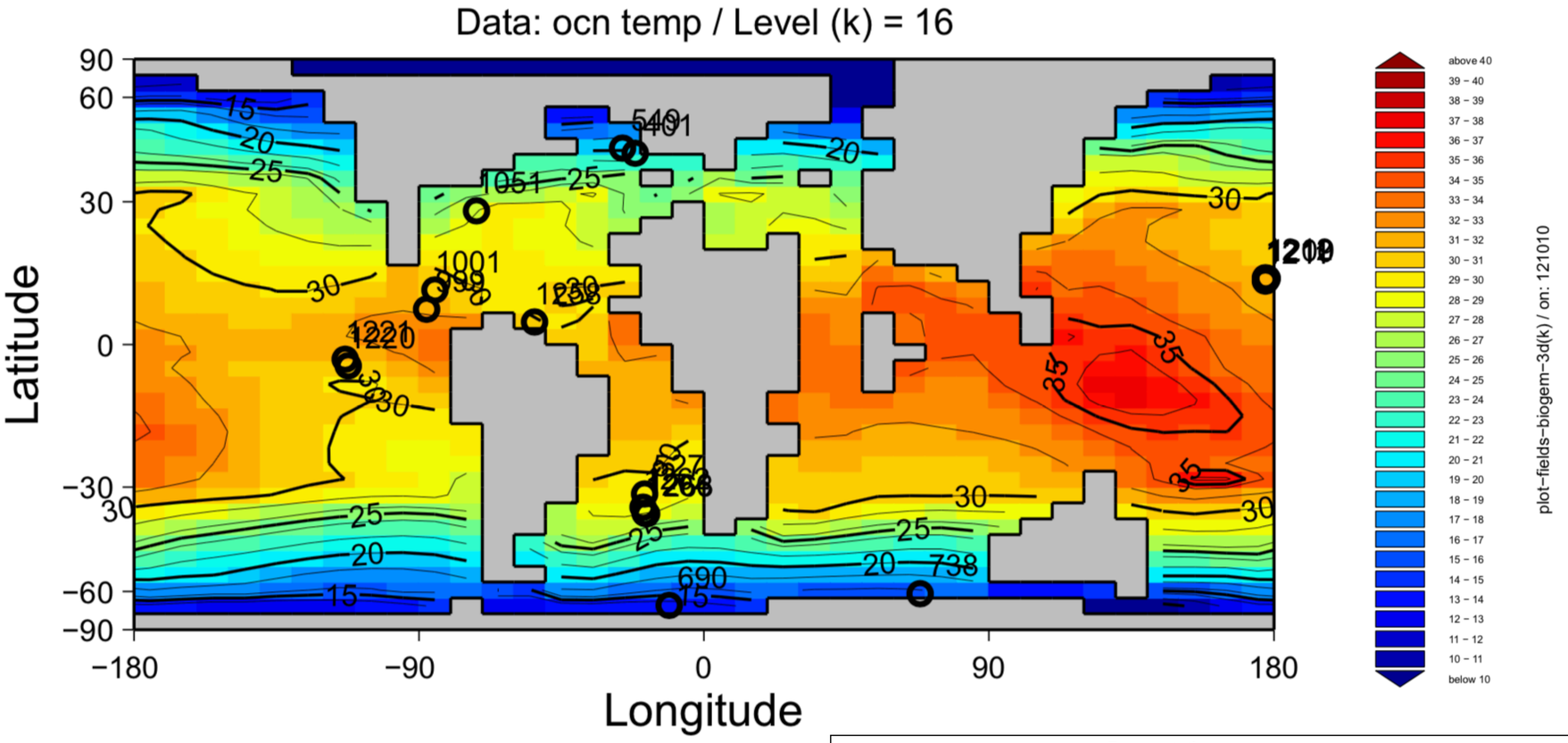


Figure 1. Illustration of the long-term (geological) carbon cycle fluxes. Shown are the long-term fluxes in the GENIE model at steady state. In red are sources of CO_2 to the atmosphere or ocean, and in dark blue are sinks of CO_2.

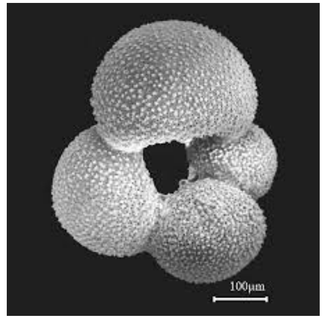
Colburn et al. (2015)

- Earth system model of intermediate complexity
  - highly parameterized; processes operating over small spatial and temporal scales aggregated into high-level parameterizations
  - more processes can be modeled and integrated over longer periods of model time, but increased uncertainty
- Includes both first-principle physics and empirical observations
- Makes predictions (e.g., stable C isotopic composition of limestone) that can be assessed with samples in field.

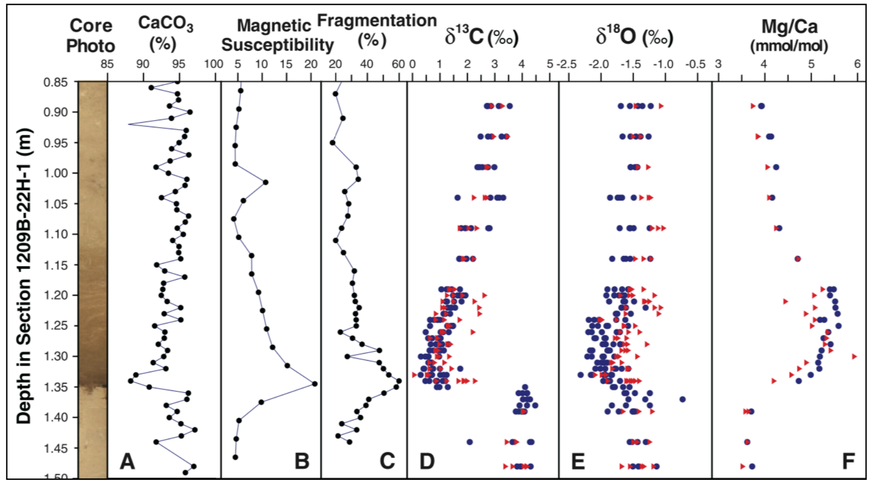
Example: cGENIE-predicted sea surface temperature 60 Myr ago

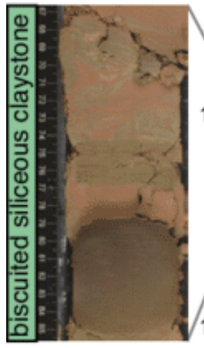


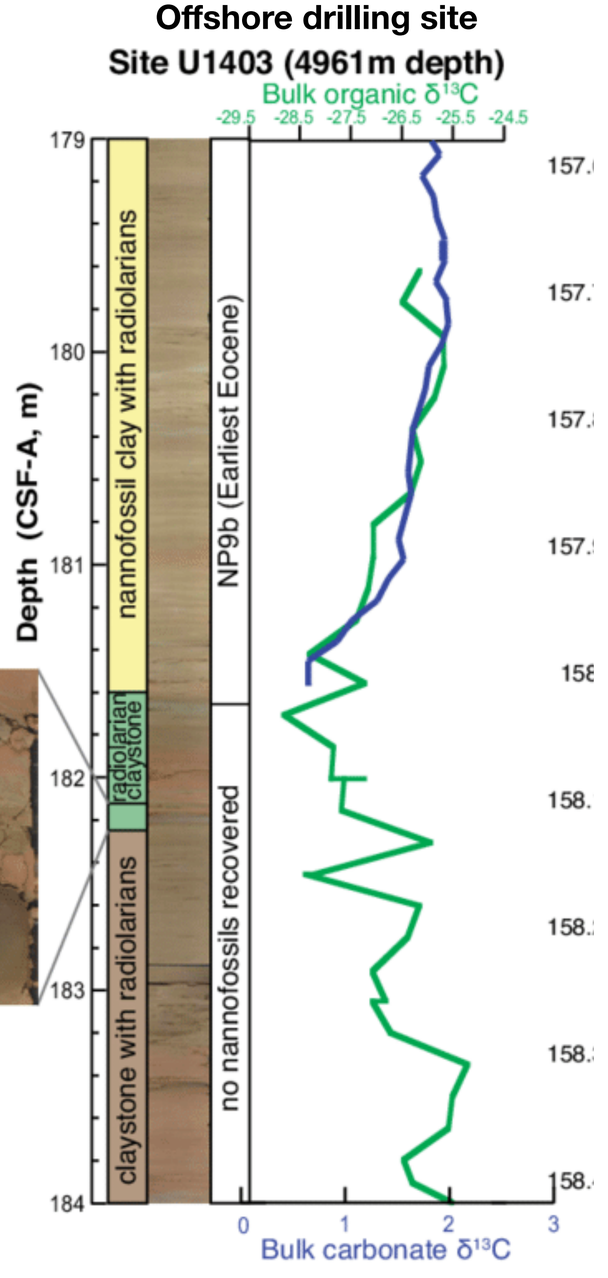
Ridgwell cGENIE user-manual (2017)



Paleotemperature proxy records (Saleska et al., 2003)







GENIE:

Grid ENabled Integrated Earth system model

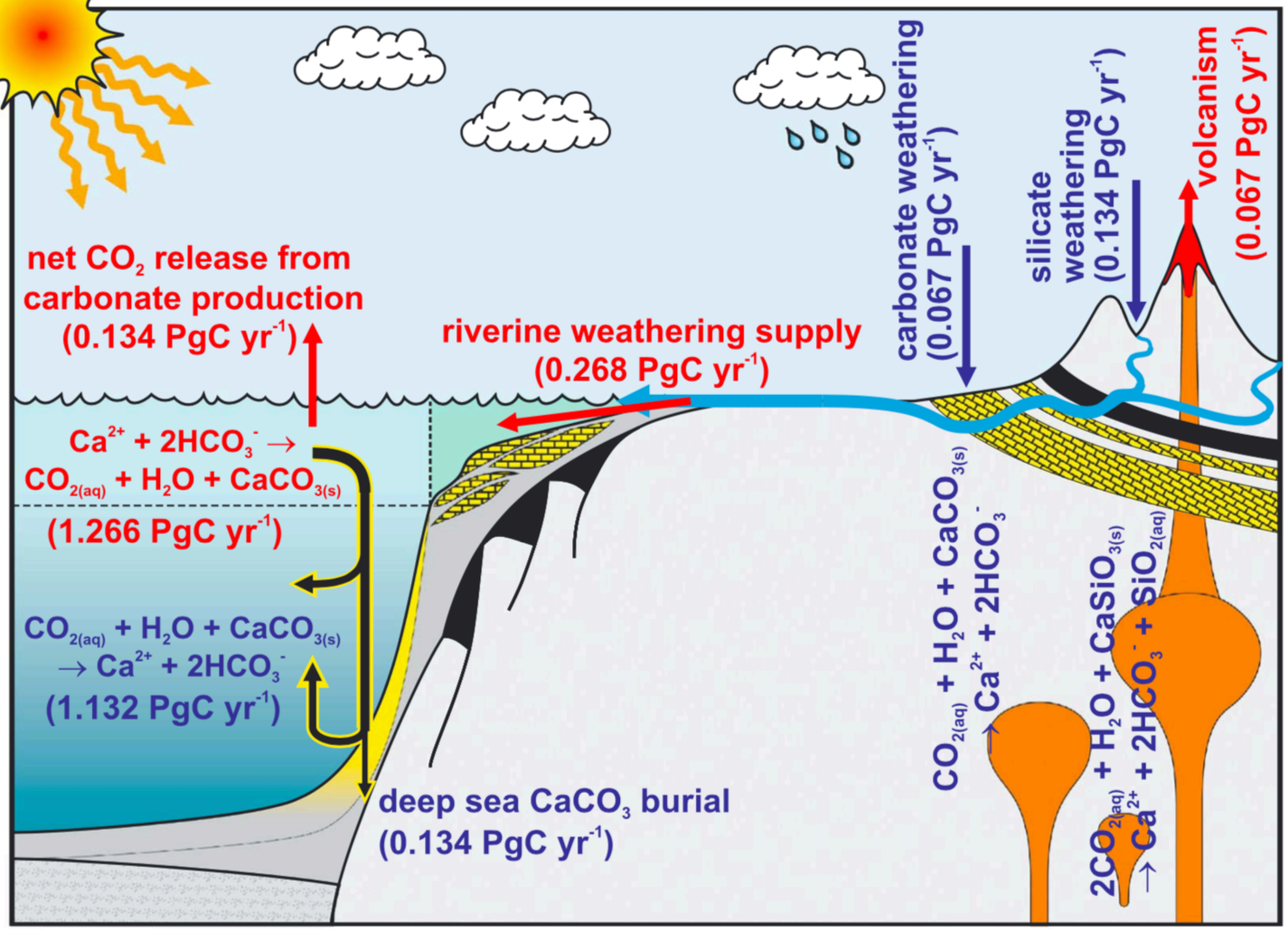
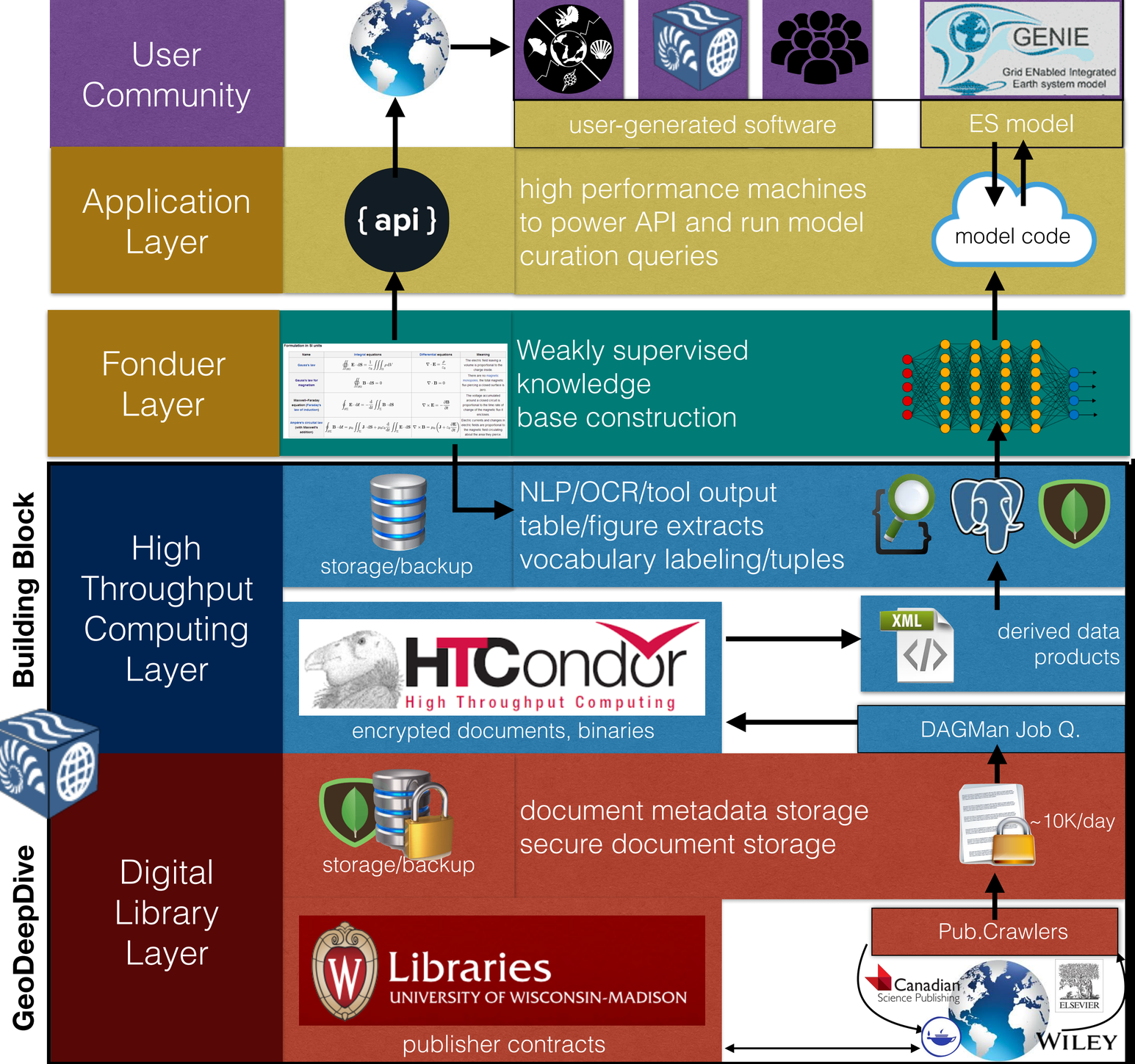


Figure 1. Illustration of the long-term (geological) carbon cycle fluxes. Shown are the long-term fluxes in the GENIE model at steady state. In red are sources of CO_2 to the atmosphere or ocean, and in dark blue are sinks of CO_2.

Colburn et al. (2015)

Phase II

Collaborator Seth Finnegan
UC Berkeley



ASKE COSMOS Layer

GeoDeepDive Layer