

Tailoring and Analyzing Topic-specific Crawls Using Semantics and Deep Learning

Ruth Duerr

Deep Learning and Semantics

The image is a composite of two parts. On the left is a white presentation slide with black text and logos. The title 'Deep Learning for Data-Driven AI' is at the top. Below it is the name 'Yoshua Bengio'. Underneath that is 'Research Data Alliance 10th Plenary Meeting'. At the bottom of the slide is the date '19 September 2017'. To the left of the text are the logos for MILA (blue hexagonal grid) and CIFAR (blue hexagonal grid). To the right is the logo for IVADO (colorful flower-like icon). A small image of a book cover for 'DEEP LEARNING' by Yoshua Bengio is shown, with the text 'PLUG: Deep Learning, MIT Press book is out, chapters will remain online'. On the right side of the image is a video frame of Yoshua Bengio, an older man with glasses and a light-colored shirt, speaking into a microphone. He is standing in front of a dark background. In the bottom right corner of the video frame, there is large blue and yellow text that reads 'RD10' with 'TH EME' written vertically next to the 'O'.

Deep Learning for Data-Driven AI

Yoshua Bengio

Research Data Alliance 10th Plenary Meeting

19 September 2017

CIFAR
CANADIAN
INSTITUTE
FOR
ADVANCED
RESEARCH

MILA

Université de Montréal

IVADO

PLUG: Deep Learning, MIT Press book is out,
chapters will remain online

DEEP LEARNING

RD10^{TH EME}

Tailoring and Analyzing Topic-specific Crawls Using Semantics and Deep Learning, presented by Ruth Duerr
at the 2018 ESIP Winter Meeting



Ronin Institute

Deep Learning and Semantics

The image shows a man with glasses and a mustache, wearing a light-colored shirt, standing behind a podium and speaking. He is holding a small device in his right hand. The podium has a large blue and yellow logo for 'RD10' on it. In the background, there is a screen displaying a slide with the title 'Deep Learning for Data-Driven AI' and the subtitle 'Still Far from Human-Level AI'. The slide lists several points about AI's limitations, including industrial successes based on supervised learning (illustrated with images of a dog and an ostrich) and learning superficial clues that can be easily fooled. The slide also notes that many more years of basic research are needed. The overall setting appears to be a conference or seminar.

Deep Learning for Data-Driven AI

Still Far from Human-Level AI

Re:

- Industrial successes mostly based on **supervised** learning



- Learning superficial clues, easy to fool trained networks
- Many more years of basic research needed

Université de Montréal

RD10
BETTER DATA BETTER DECISIONS
TRÈS BONNES DONNÉES BONNES DÉCISIONS
19-21 SEPTEMBER /SEPTEMBRE 2017 MONTREAL, CANADA
www.rd-alliance.org/rd

Deep Learning and Semantics

The image shows a man with glasses and a light-colored shirt speaking into a microphone at a podium. Behind him is a large screen displaying a presentation slide. The slide has a dark background with white text. At the top left, it says 'Deep Learning and Semantics' and 'Still Far from Human-Level AI'. Below that is a bulleted list: '• Industrial successes mostly based on **supervised** learning'. In the center of the slide is a graphic with a pink banner that reads 'data is the new oil' in large white letters, set against a background of various icons related to data and technology. To the left of the main slide, there are logos for 'MILA' and 'Université de Montréal'. On the right side of the slide, the text 'RD10' is visible.

Deep Learning and Semantics
Still Far from Human-Level AI

- Industrial successes mostly based on **supervised** learning

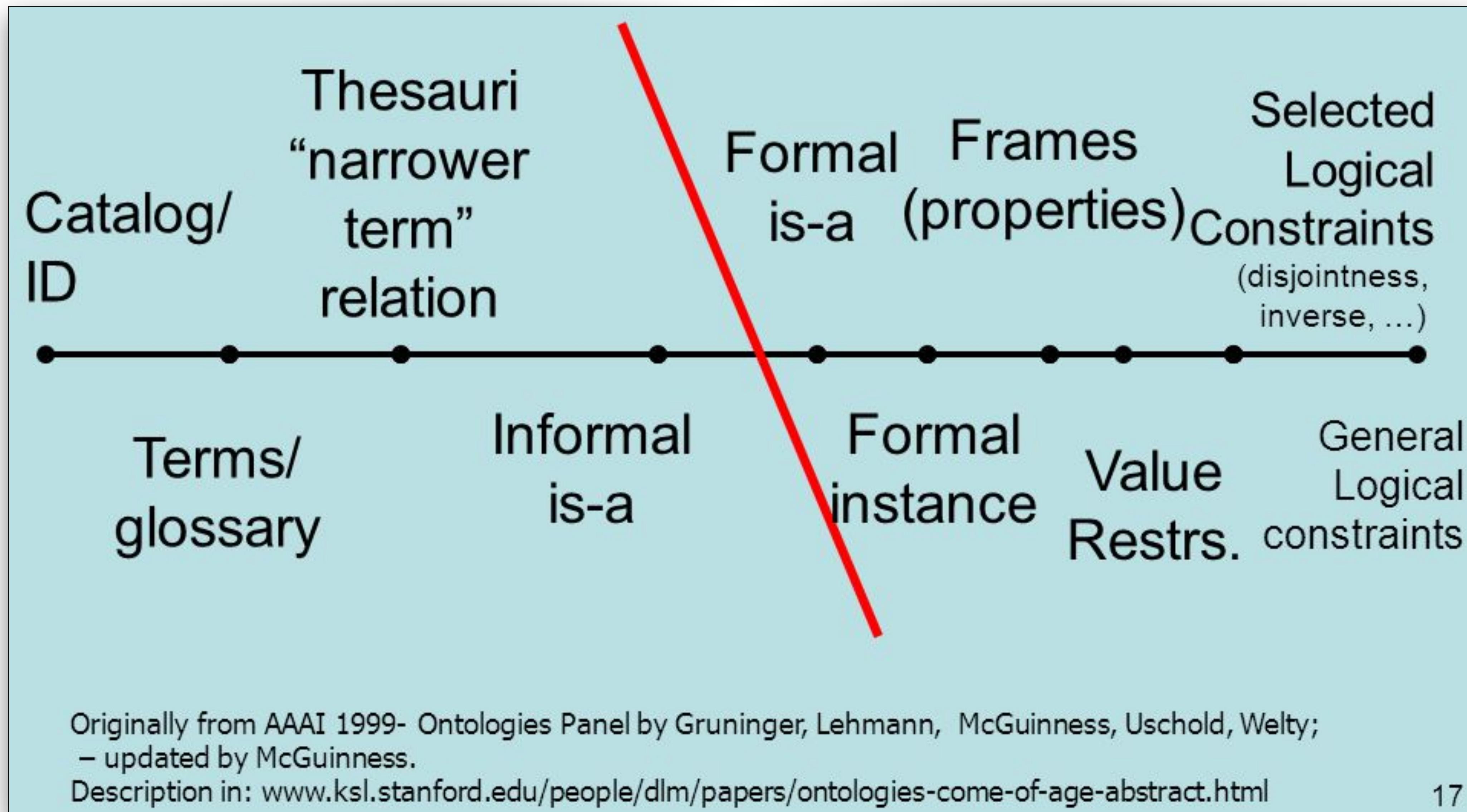
data is the new oil

- Learning systems based on neural networks
 - Because AI is based on ML, successful AI applications require DATA – lots of data
- Many more steps
 - The first step in any project:
 - what data is available and what data is needed, do we need to collect more, do we need to label it?

So, What is Labeled Data?

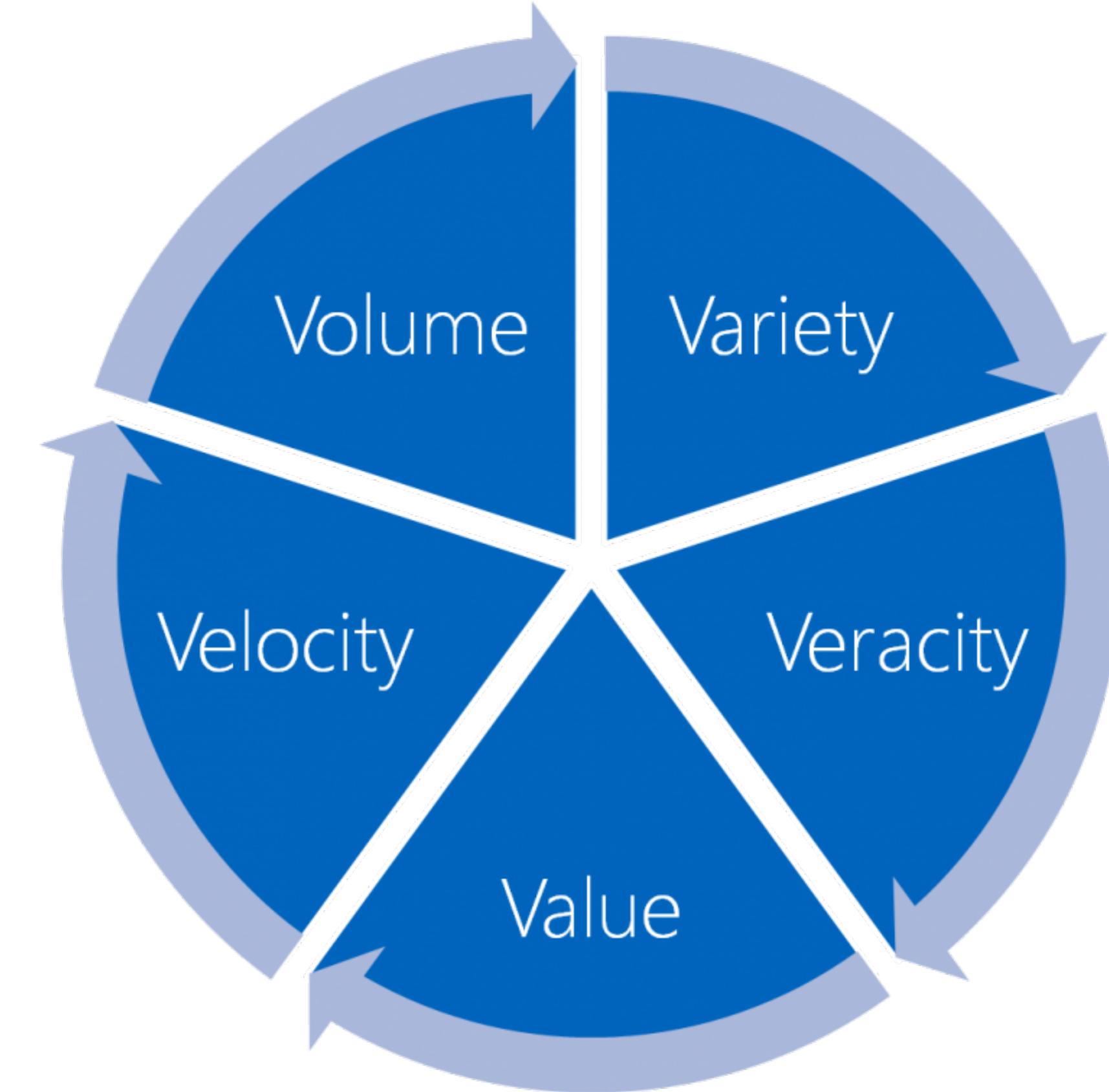
- There are infinitely many kinds
 - Nouns, verbs, adjectives, etc.
 - Entities (e.g., people, organizations, funders, etc.)
 - This image contains a hurricane (or person, or face, or animal, or building or whatever)

The Spectrum of Semantics - A Spectrum of Labeled Data?



So, What's the Problem?

- Domain data is highly distributed
- Domain data is extremely diverse
- Cataloging all of it is an impossible task
- What if we just leave everything where it is and find it, as needed, through *focused crawling*?



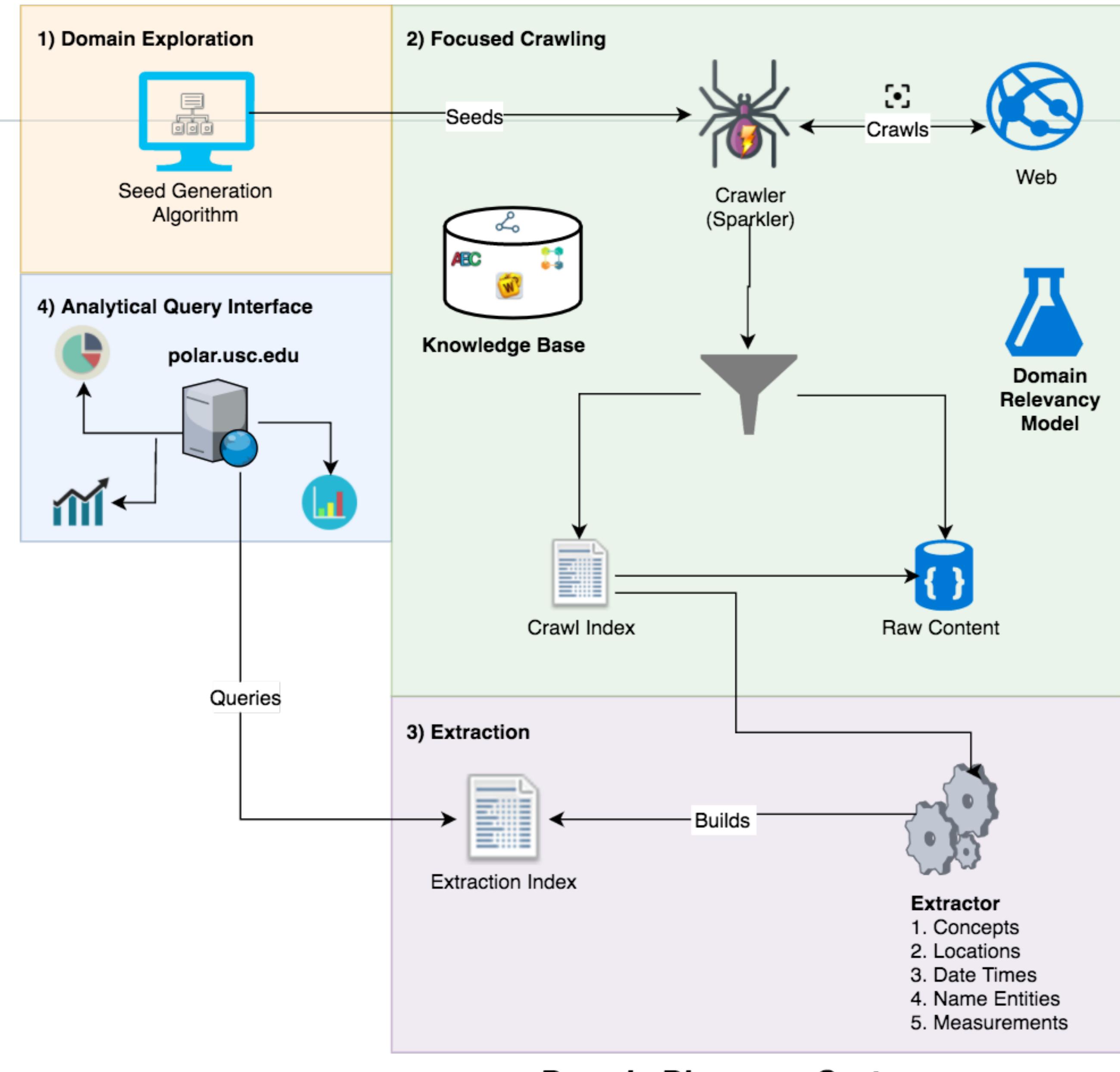
Applying ‘Big Data’ Technology to Domain (here Polar) Data

- Make it possible to query the body of accumulated knowledge about a domain, using natural language and deep learning
 - Find the applicable data and documents
 - Evaluate the structure and contents to effectively extract information
 - Store and index the information
 - Create interface to query the content (using NLP/ML)

Polar Deep Insights Architecture

Leverages prior work done under the DARPA MEMEX (<http://memex.jpl.nasa.gov/>), NSF Polar CyberInfrastructure activities, and community workshops

1. Domain Exploration - Create a URL seed list and domain relevancy model
2. Focused Crawling - Crawl the web using the seed list and model
3. Extraction - Use a number of extractors to extract content from the documents returned by the crawl
4. Analytical Query Interface - Use a variety of analytical tools to explore the extracted content



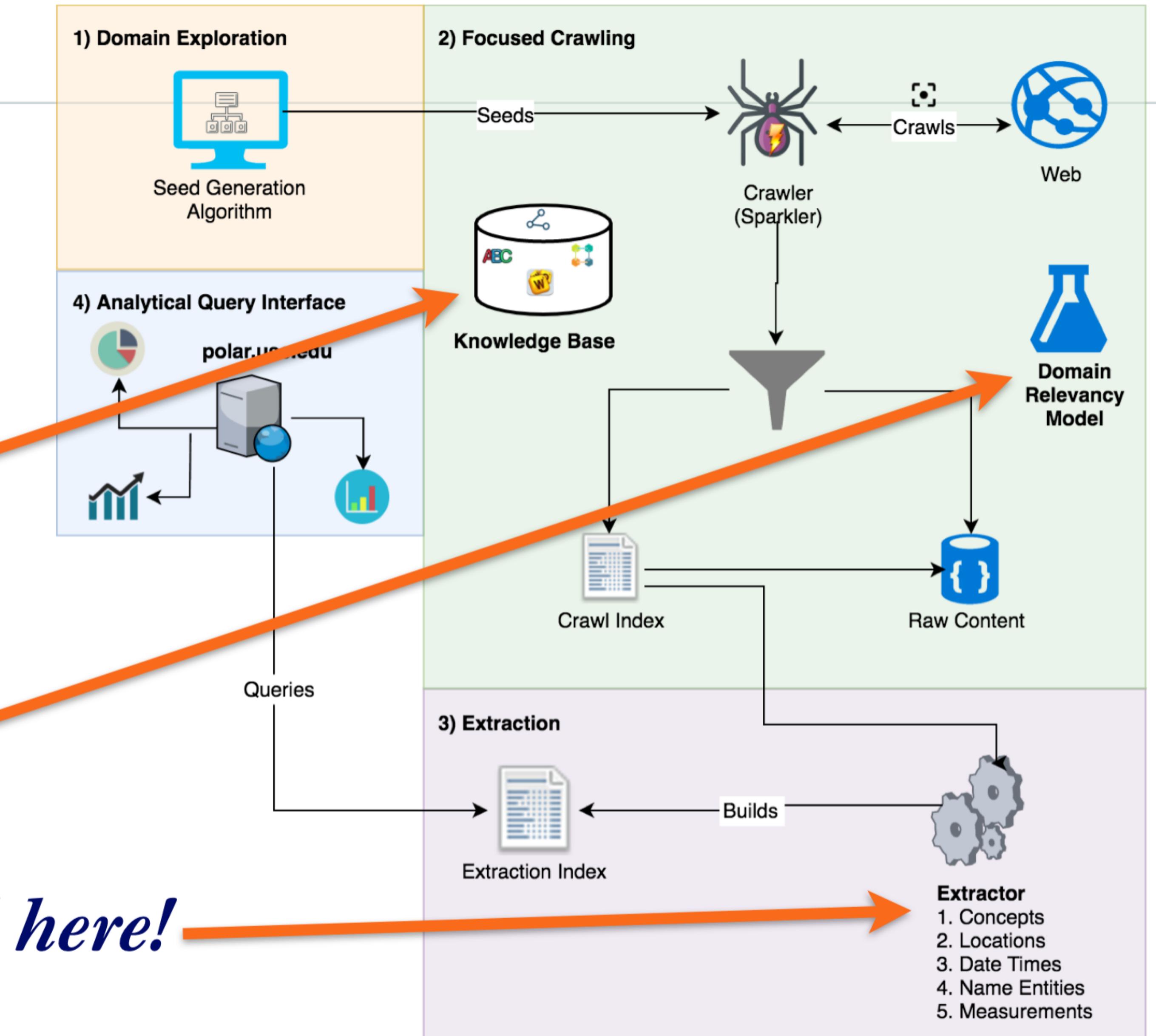
Domain Discovery System

Semantics is Everywhere!

Semantics is here!

And here!

And here!



Two Use Cases with Semantics in use with crawling and deep learning

- Sea ice use case - what insights can we get by simply applying domain expertise to a mostly existing set of tools
- Next steps with Ocean Observing Best Practices

Sea Ice Use Case - Crawler Inputs

Glossaries

http://pubs.usgs.gov/of/2004/1216/text.html	glacier terminology
http://unesdoc.unesco.org/images/0019/001925/192525E.pdf	glacier mass balance and related terms
http://www.aineva.it/previsori/Classificazione%20Internazionale/20080720_iacs_classif2008.pdf	snow terminology
https://nsidc.org/fgdc/glossary/	permafrost glossary (several pdf files required from website)
http://nsidc.org/cryosphere/glossary/all	general cryospheric glossary
http://www.jcomm.info/components/com_oe/oe.php?task=download&id=27226&version=March%202014&lang=1&format=1	sea ice glossary
http://globalcryospherewatch.org/reference/glossary.php	extensive compilation of cryospheric terms, most likely encompasses all of the above
http://www.spri.cam.ac.uk/resources/directory/organisations/	extensive list of polar organizations of various sorts

Sea Ice Use Case - Crawler Inputs

URL Seed List

- <http://arcticportal.org/>
- <http://ipa.arcticportal.org/>
- <http://nsidc.org/data/>
- <https://arcticdata.io/>
- <https://www.data.gov/climate/arctic-data/>
- <https://www.bas.ac.uk/data/>

Search Terms

- sea ice
- permafrost
- glacier
- ice sheet
- polar
- Arctic
- Antarctic
- snow
- sea ice thickness
- navigability
- calving
- ice berg
- snow water equivalent
- mass balance
- albedo

- After an initial crawl, a sample of the resulting documents were characterized as relevant, irrelevant, or possibly relevant and used to re-train the model

Sea Ice Use Case - Banana based query and analysis

The screenshot shows a web browser window for the URL polar.usc.edu/html/polar-deep-insights/#/config. The title bar says "Polar Deep Insights". The top navigation bar includes links for "Concept Editor", "Query Interface", and "Configure". The main content area is titled "Application configuration" with a status of "Complete". It lists several configuration parameters:

Field	Value
Elastic search endpoint	http://polar.usc.edu/elasticsearch
Elastic search extraction index	polar-deep-insights-complete
Elastic search extraction doc-type	docs
Elastic search measurements index	polar-measurements
Elastic search measurements doc-type	raw-measurements
Entity Count JSON path	http://polar.usc.edu/html/polar-deep-insights/entity-count.json
Sweet ontology path	http://polar.usc.edu/html/polar-deep-insights/sweet-ontology.ttl

On the right side, there is a note about a required concept ontology with download and creation links, and another note about document counts per-entity with a download link.

SAVE

[Github](#) . [Wiki](#)

 from IRDS.USC.EDU

Sea Ice Use Case - Banana based query and analysis

The screenshot shows a web browser window for the URL polar.usc.edu/html/polar-deep-insights/#/config. The title bar says "Polar Deep Insights". The top navigation bar includes "Concept Editor", "Query Interface", and "Configure". The main content area is titled "Application configuration Complete". It lists three tabs: "TREC-DD-PDF" (selected), "TREC-DD-SAMPLE", and "NSIDC-CRAWL". Below the tabs is a table with two rows:

Field	Value
Elastic search endpoint	http://polar.usc.edu/elasticsearch
Elastic search extraction index	polar-deep-insights-complete

A red box highlights a note and a "DOWNLOAD" button:

A concept ontology is required for this application to function. You can [create your own ontology](#) or download a predefined ontology from your elasticsearch index.

DOWNLOAD

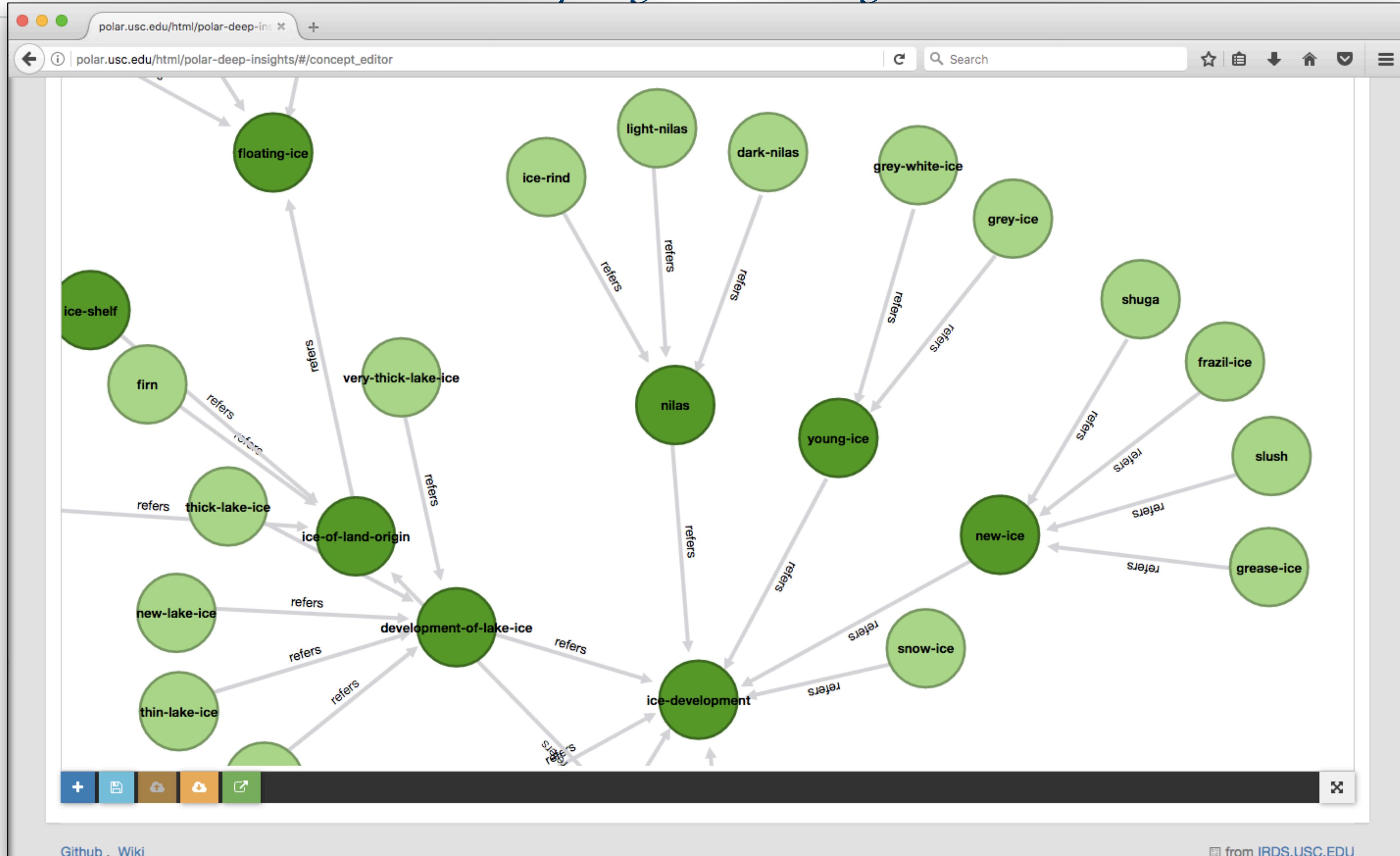
Below this, a large callout box contains the same text and a "DOWNLOAD" button:

A concept ontology is required for this application to function. You can [create your own ontology](#) or download a predefined ontology from your elasticsearch index.

DOWNLOAD

At the bottom left is a "SAVE" button. At the bottom right is a link to "Github . Wiki" and "from IRDS.USC.EDU".

Sea Ice Use Case - Banana based query and analysis



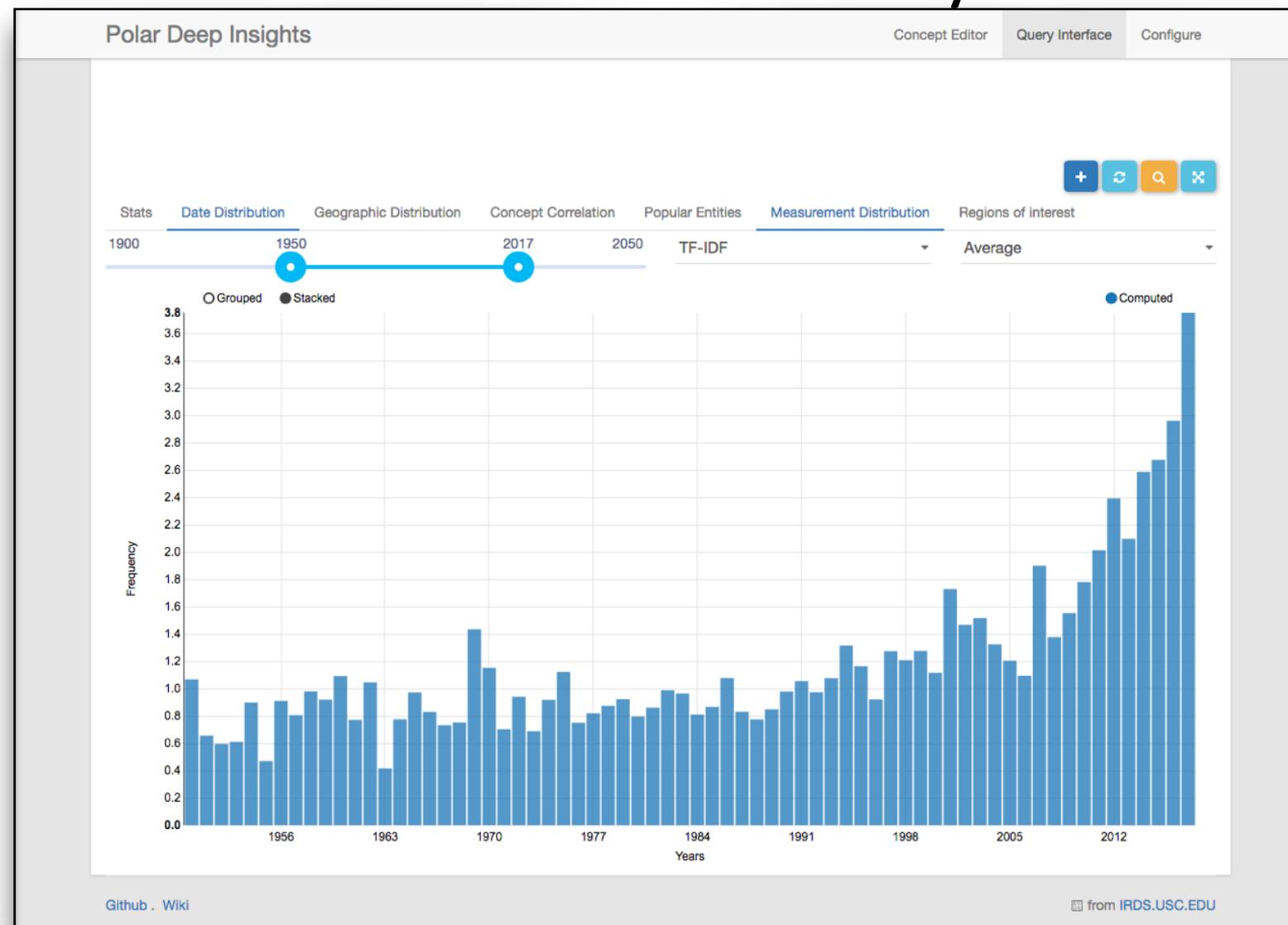
Github . Wiki

from IRDS.USC.EDU

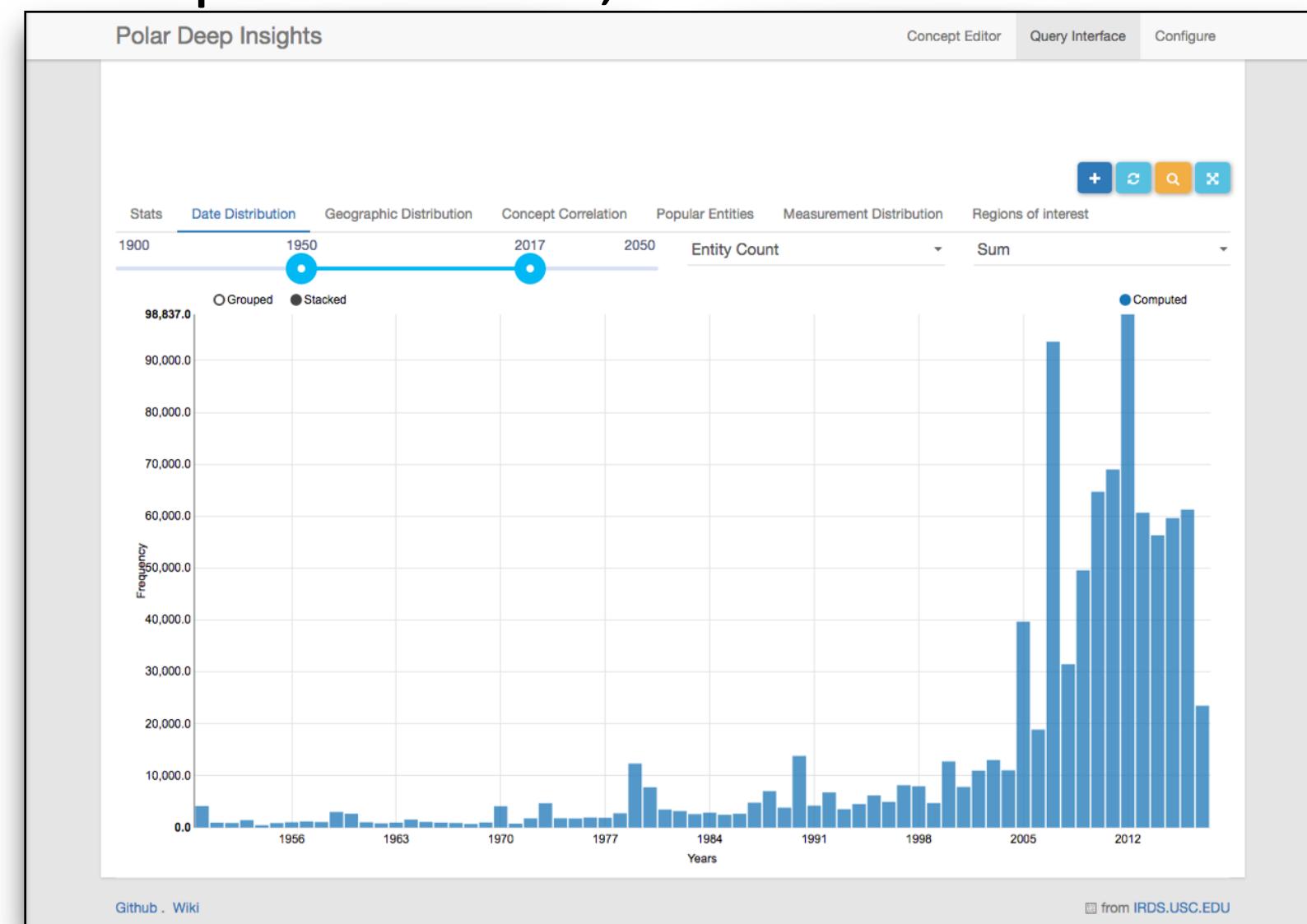
Sea Ice Use Case - Banana based query and analysis



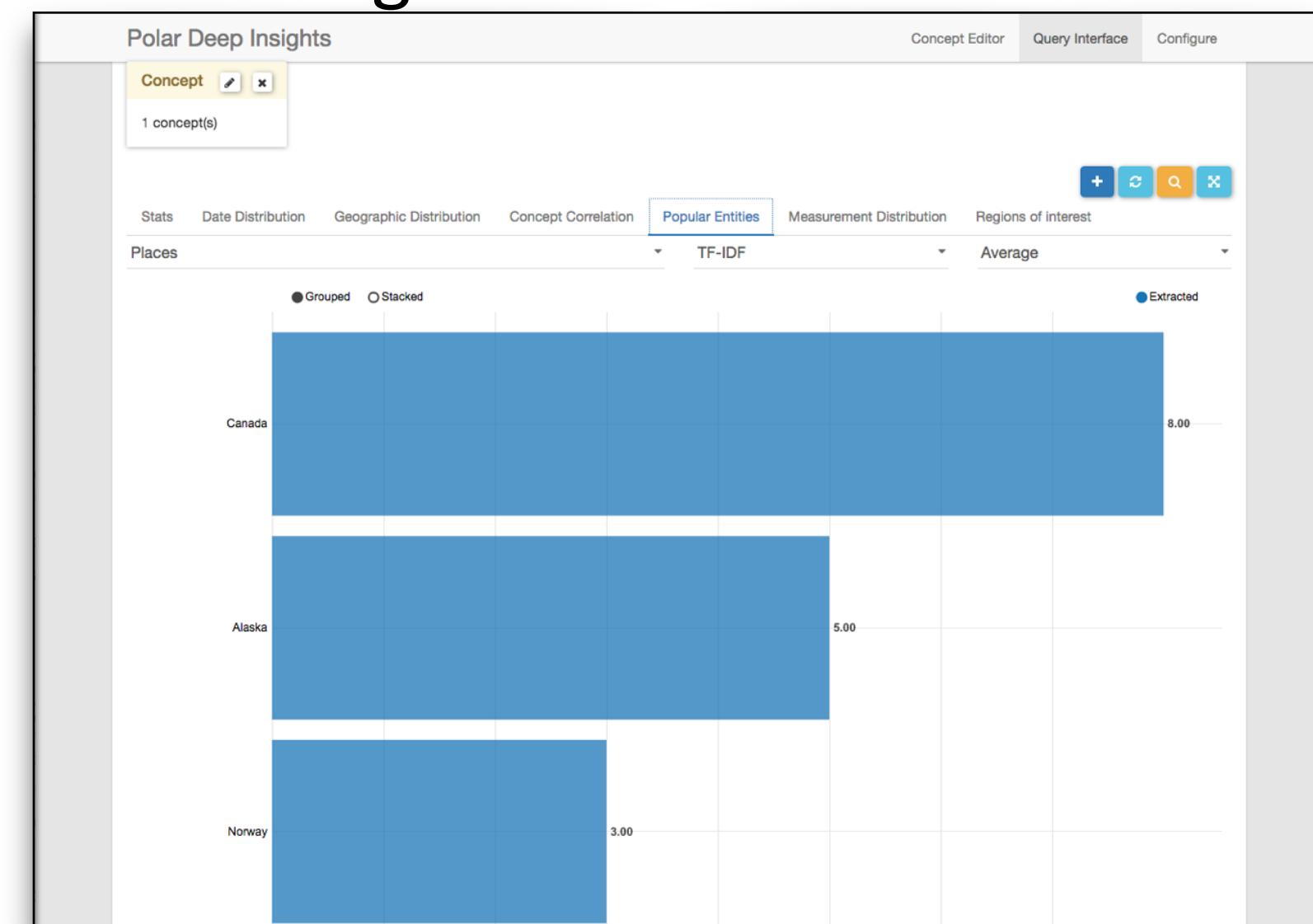
Distribution of documents by date



Distribution of documents that mention icebergs again by date - What's up with the spikes in 2005, 2007 and 2012?



Locations mentioned in documents that mentioned icebergs and that something was ice-bound.



Domain Exploration - Semi-automated Model Generation

Domain Discovery - Seed Generation

1 Generate a Model
ocean best practices 

Minimum 10 each
78  66  108 

More Options »

2 Create a Seed File


3 Start the Crawl

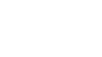
More Options »

4 Visit the Crawl Dashboard

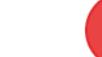
Title: Best Practices for Website Navigati
URL: https://ocean19.com/blog/best-pract

<iframe src="https://www.googletagmanager.com/ns.html?id=GTM-NG85BQL" height="0" width="0" style="display:none;visibility:hidden"></iframe>

THE ocean AGENCY

Title: Best Practices Teaser - Data.gov
URL: https://www.data.gov/ocean/best-pra

Title: OCADS - Guide to Best Practices for
URL: https://www.nodc.noaa.gov/ocads/oce

 NOAA NATIONAL CEI

OCADS Home
Access Data
Submit Data
About

Guide to Best Practices for Ocean

Title: Ocean - Best practices, tips and fu
URL: https://www.classy.org/blog/ocean/

Blocked by Content Security Policy

This page has a content security policy that prevents it from being loaded in this way.

Firefox prevented this page from loading in this way because the page has a content security policy that disallows it.

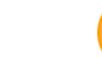
Title: Welcome to the Frontpage
URL: http://www.oceandatastandards.org/

OCEAN DATA STANDARDS

The Ocean Data Standards and Best Practice Project (ODSBP)

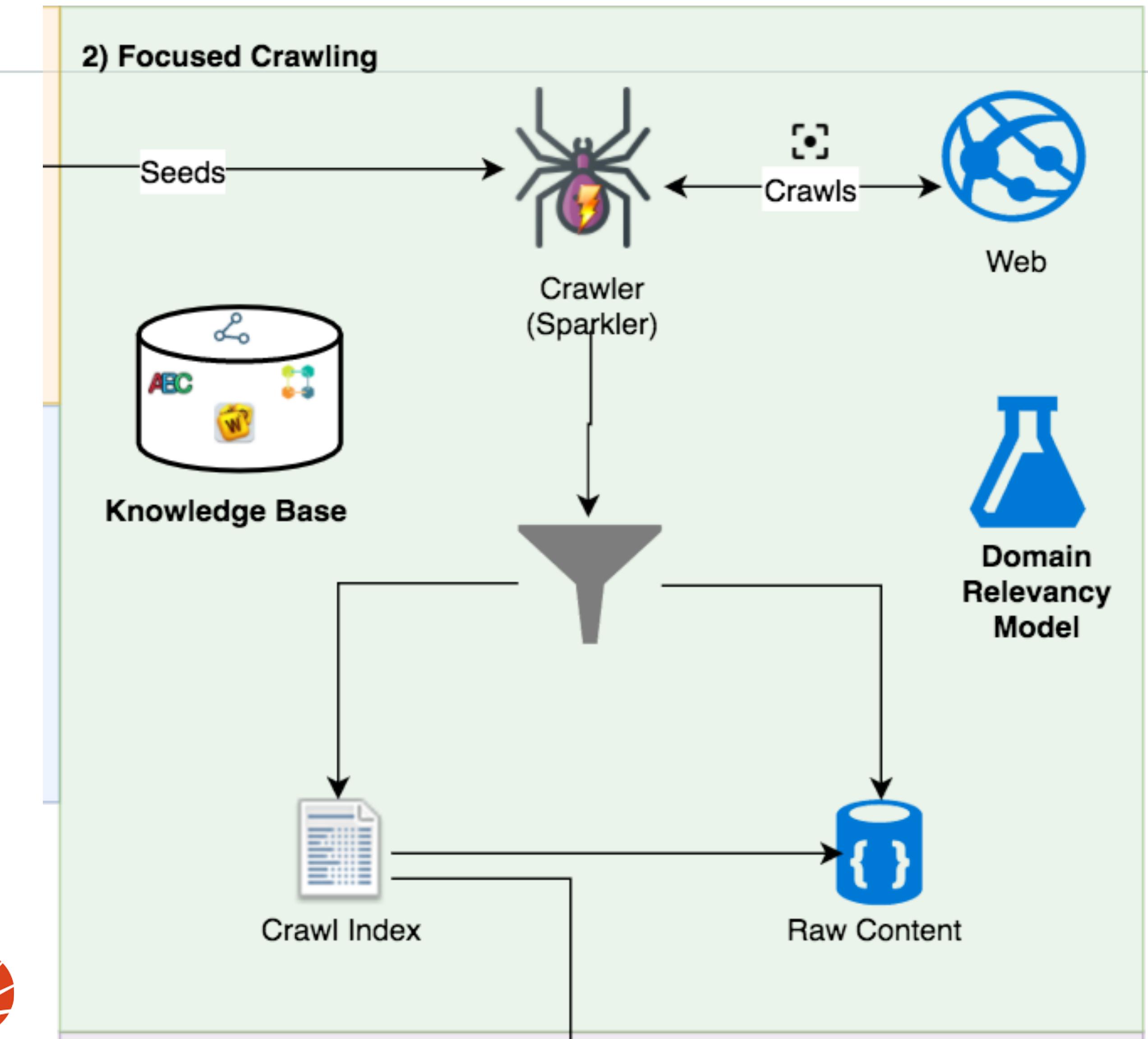
  

Title: Ocean - Best Practices - Data.gov
URL: https://www.data.gov/ocean/best-pra

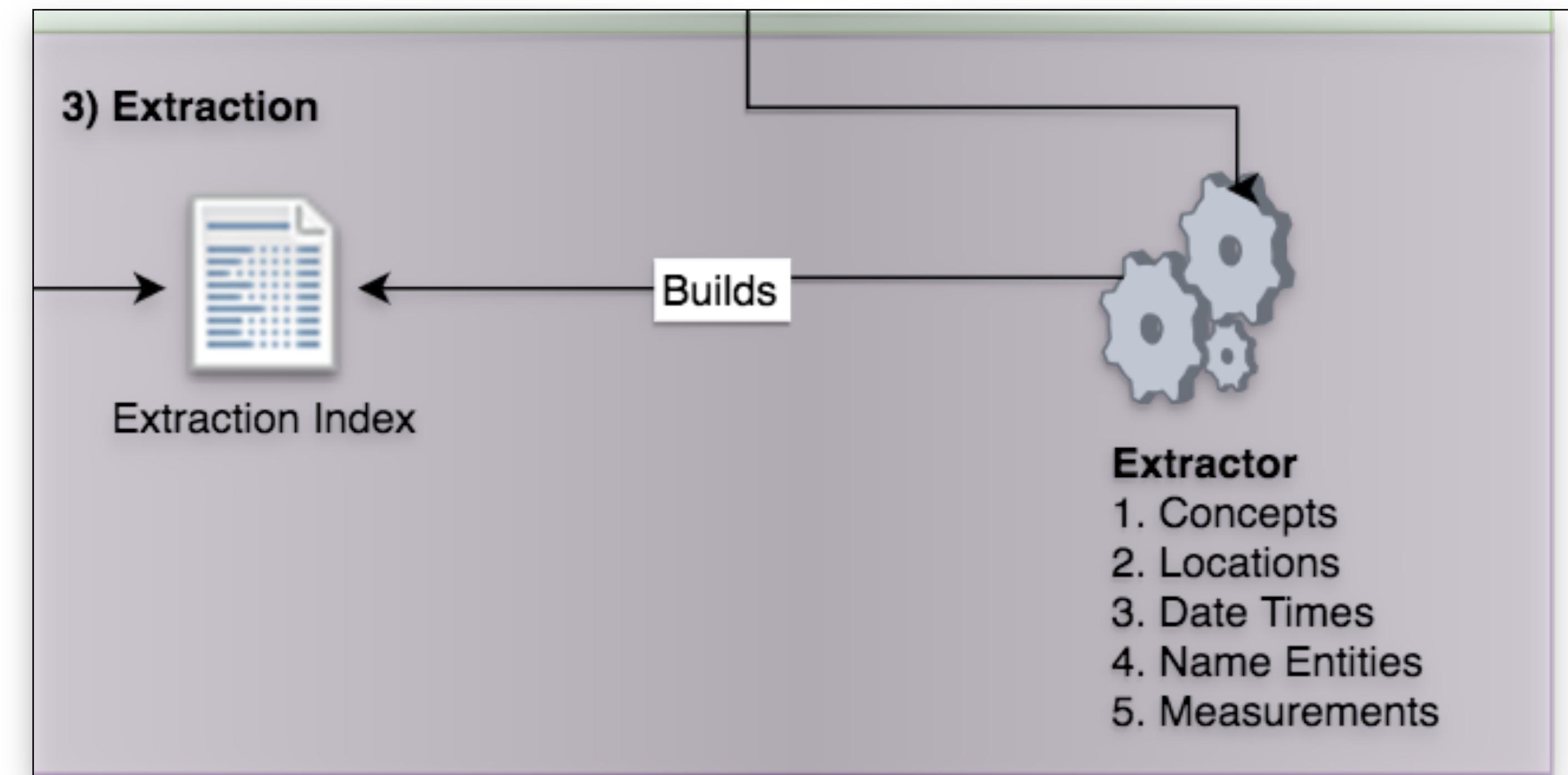
Focused Crawling

- Sparkler (<https://github.com/USCDatascience/sparkler>) is an extensible, highly scalable Web crawler that runs on top of Spark (vice Hadoop)
- Uses the domain relevancy model to find resources
- Avoids disrupting hosts being crawled
 - Partitions URLs by hostname and every node gets a different host to crawl
 - Inserts time delays between successive requests



Extraction

- Detects and extracts metadata, text, URLs
- Toolkit of parsers to extract
 - Concepts
 - Geographic locations
 - Dates and Times
 - Named Entities
 - Numerical measurements
- Creates an index for the extracted content



Polar Deep Insights - Facet-view based query and analysis

Query for documents mentioning the words “standards”, “ocean” and “sensors” using a keyword facet with terms “arctic” and “sea”

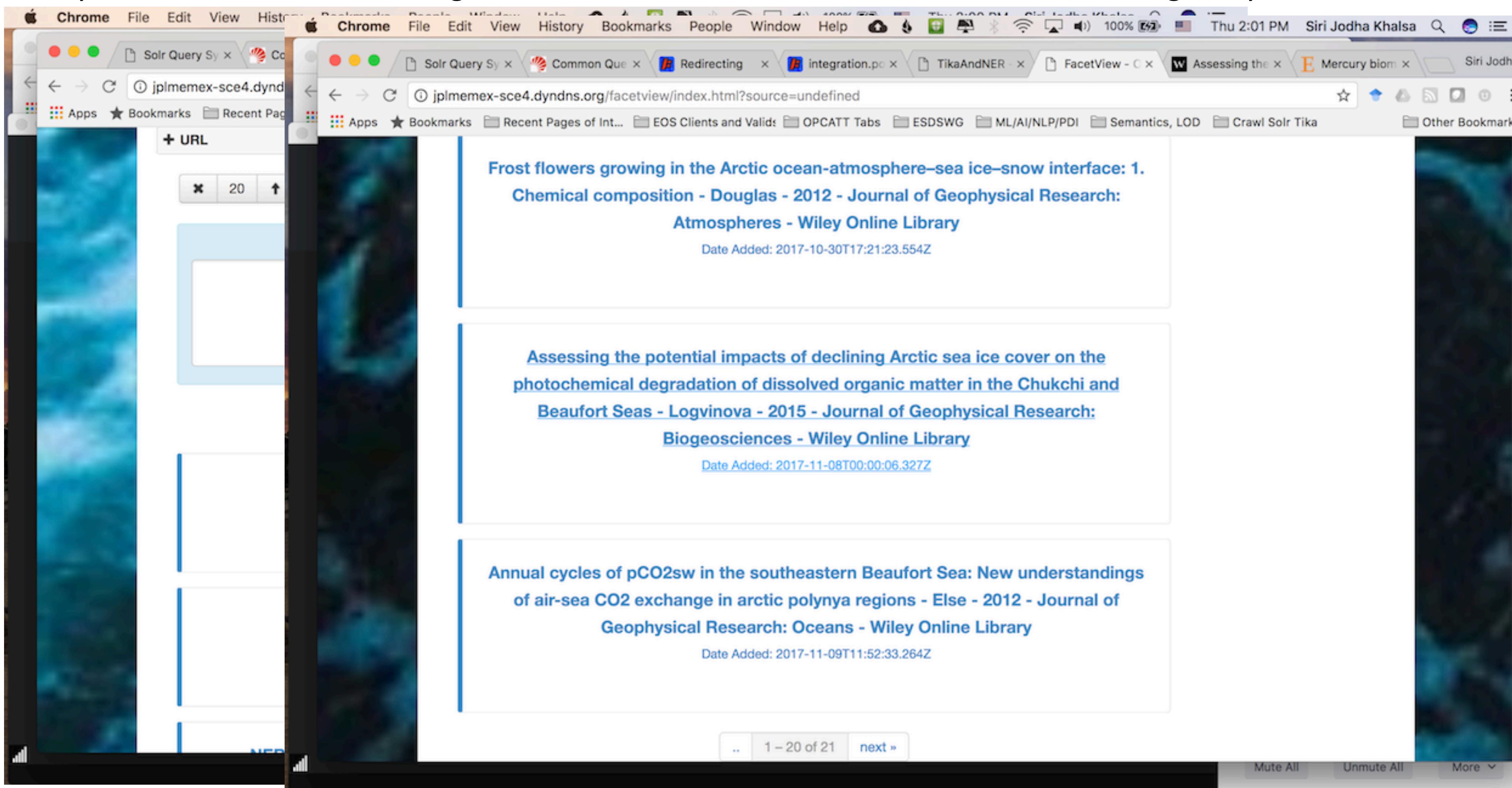
The screenshot shows a web browser window with the URL jpmemex-sce4.dyndns.org/facetview/index.html?source=undefined. The search query in the address bar is "standards ocean sensors". Below the address bar, there is a search interface with a "Date added" dropdown set to "20" and a "Share or save this search:" button. At the bottom of this interface are two blue buttons labeled "sea" and "arctic". The main content area displays three search results in cards:

- NERC - Wet UK summers linked to Arctic sea ice retreat**
Date Added: 2017-11-08T21:12:34.466Z
- NERC - CryoSat-2 mission reveals major Arctic sea-ice loss**
Date Added: 2017-11-08T21:06:01.658Z
- NERC - Ice thickness from CryoSat-2: Arctic sea ice thickness**
Date Added: 2017-11-08T21:06:01.658Z

At the bottom of the browser window, there are buttons for "Mute All", "Unmute All", and "More".

Polar Deep Insights - Facet-view based query and analysis

Query for documents mentioning the words “standards”, “ocean” and “sensors” using a keyword facet with terms “arctic” and “sea”



Polar Deep Insights - Facet-view based query and analysis

Query for documents mentioning the words “standards”, “ocean” and “sensors” using a keyword facet with terms “arctic” and “sea”

The screenshot shows a desktop interface with three separate Chrome browser windows open side-by-side. The central window is the most prominent, displaying a scientific article from onlinelibrary.wiley.com/doi/10.1002/2015JG003052/full. The article's abstract discusses the validation of PARAFAC models and the preparation of standards for DOC analyses. The word "standards" is highlighted in yellow throughout the text. To the right of the main content, a sidebar provides navigation links for the document's sections: Abstract, 1 Introduction, 2 Methods, 3 Results, 4 Discussion, 5 Implications and Conclusions, Acknowledgments, References, Related Content, and Citing Literature. The top of the screen features a Mac OS X-style menu bar with Apple, Chrome, and System preferences. The system tray at the bottom includes icons for battery level (100%), signal strength, and network connectivity.

Acknowledgements

This work would not have been possible without funding by NSF through ICER grant #1639675