



2



# *IN12C-02: (Polar) Domain Discovery with Sparkler*

Siri-Jodha S Khalsa<sup>1</sup>, Ruth Duerr<sup>2</sup>, Chris A Mattmann<sup>3</sup>,  
Nithin Krishna Ottilingam<sup>4</sup>, Karajeet Singh<sup>4</sup>, Luis Alberto Lopez<sup>1</sup>, Omid  
Davtalab<sup>4</sup>, Simin Karvigh<sup>4</sup>

4

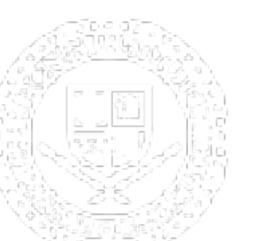


1



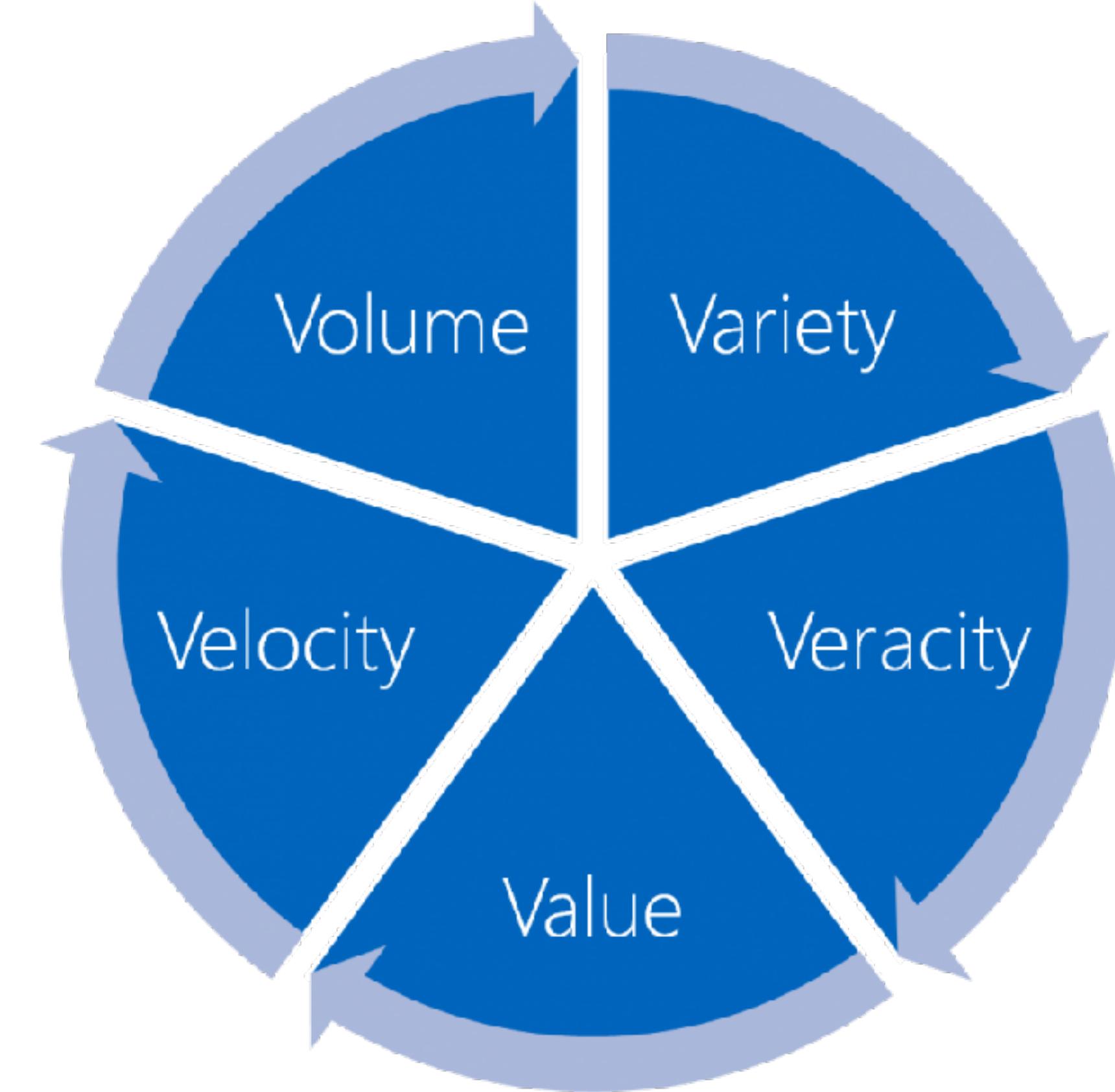
National Snow and Ice Data Center  
*Advancing knowledge of Earth's frozen regions*

3



# *So, What's the Problem?*

- Domain (polar) 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*?



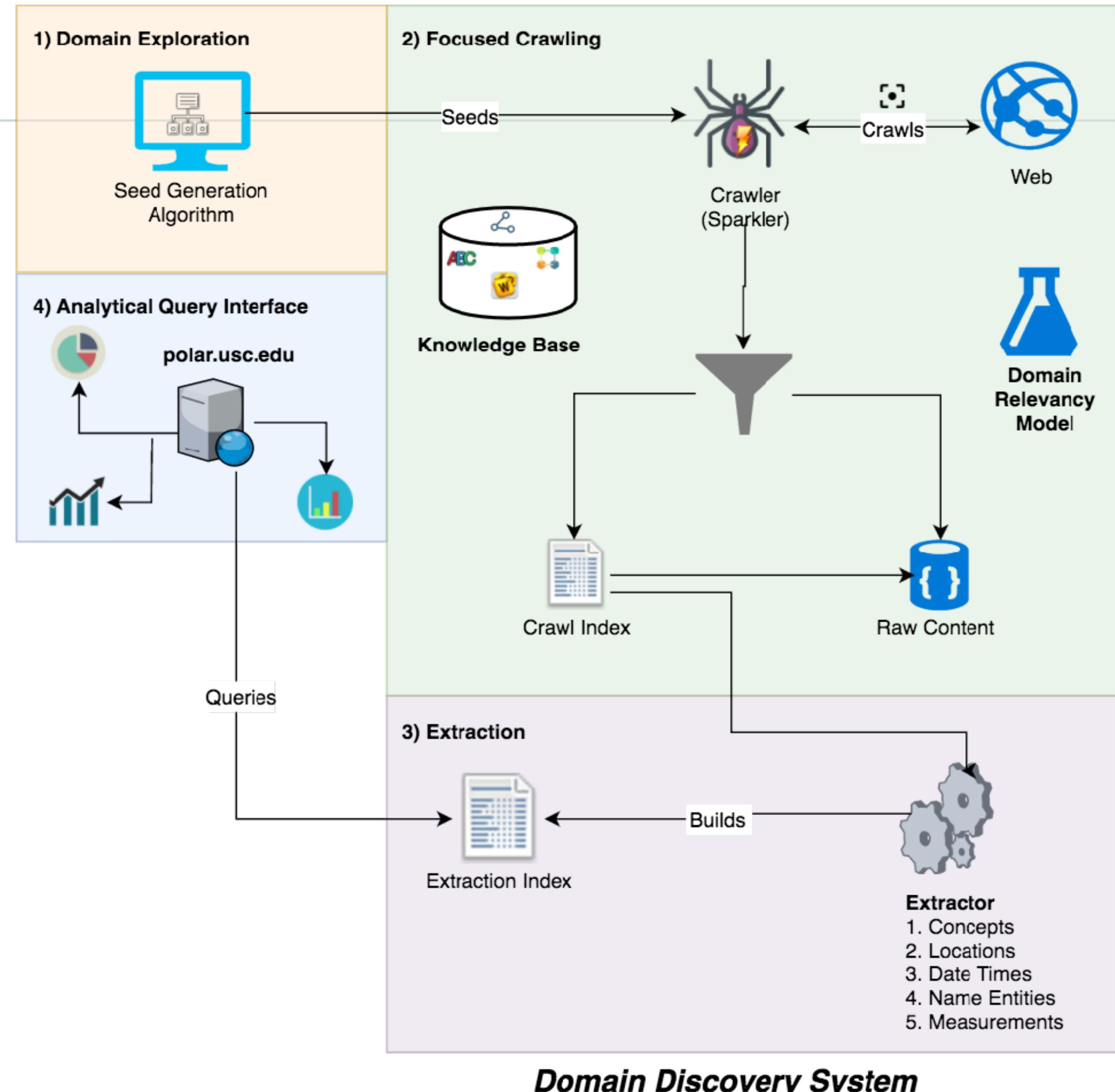
# *Applying ‘Big Data’ Technology to Domain (here Polar) Data*

- 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 Exploration - Create a URL seed list and domain relevancy model*

We are currently exploring two paths:

1. Subject Matter Expert model generation (from knowledge base, glossaries, search terms, and seed URL's)
2. Semi-automated model generation

Intent to compare efficiency and accuracy of each path (and their variations)

Two test cases so far:

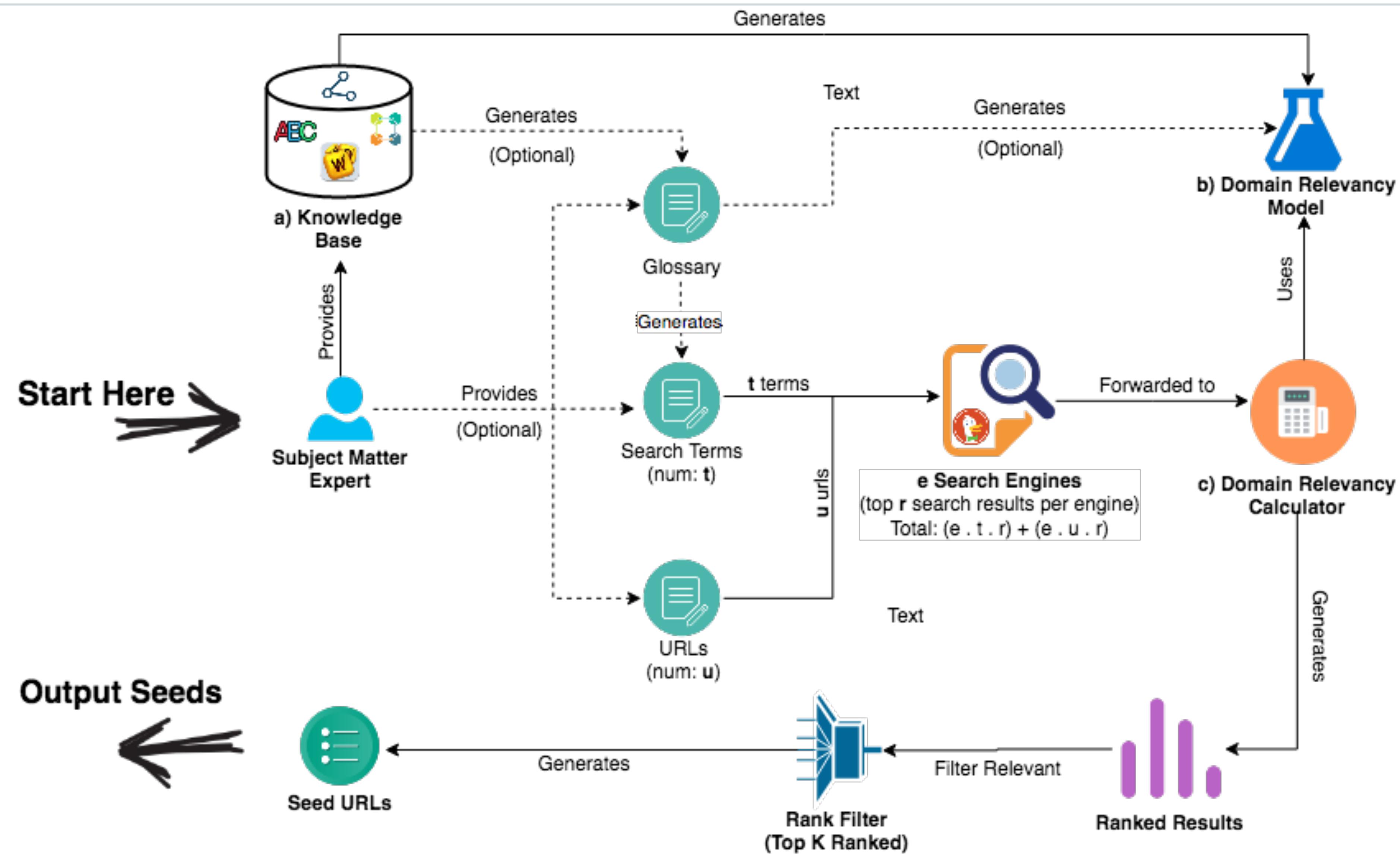
1. Sea ice based off the NSF-funded SSIII project's sea ice ontologies
2. Search terms, seed URL's, etc. provided by Jay Pearlman and Pier Luigi Buttigieg for their Integrated Oceanographic Data and Information Exchange Best Practices work

## **1) Domain Exploration**



**Seed Generation  
Algorithm**

# Domain Exploration - Subject Matter Expert Model Generation



# *Domain Exploration - Semi-automated Model Generation*

---

- Finding relevant URLs to crawl
  - Start with small set of seeds provided by domain experts
  - Feed these to general search engines and rank the additional found links according to text similarity and other measures
- Domain experts rate these URLs for relevancy
  - This annotated set of URLs then used to train a machine learning model to predict the ‘domain relevance’ of a given document

# 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**

**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/cce**

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

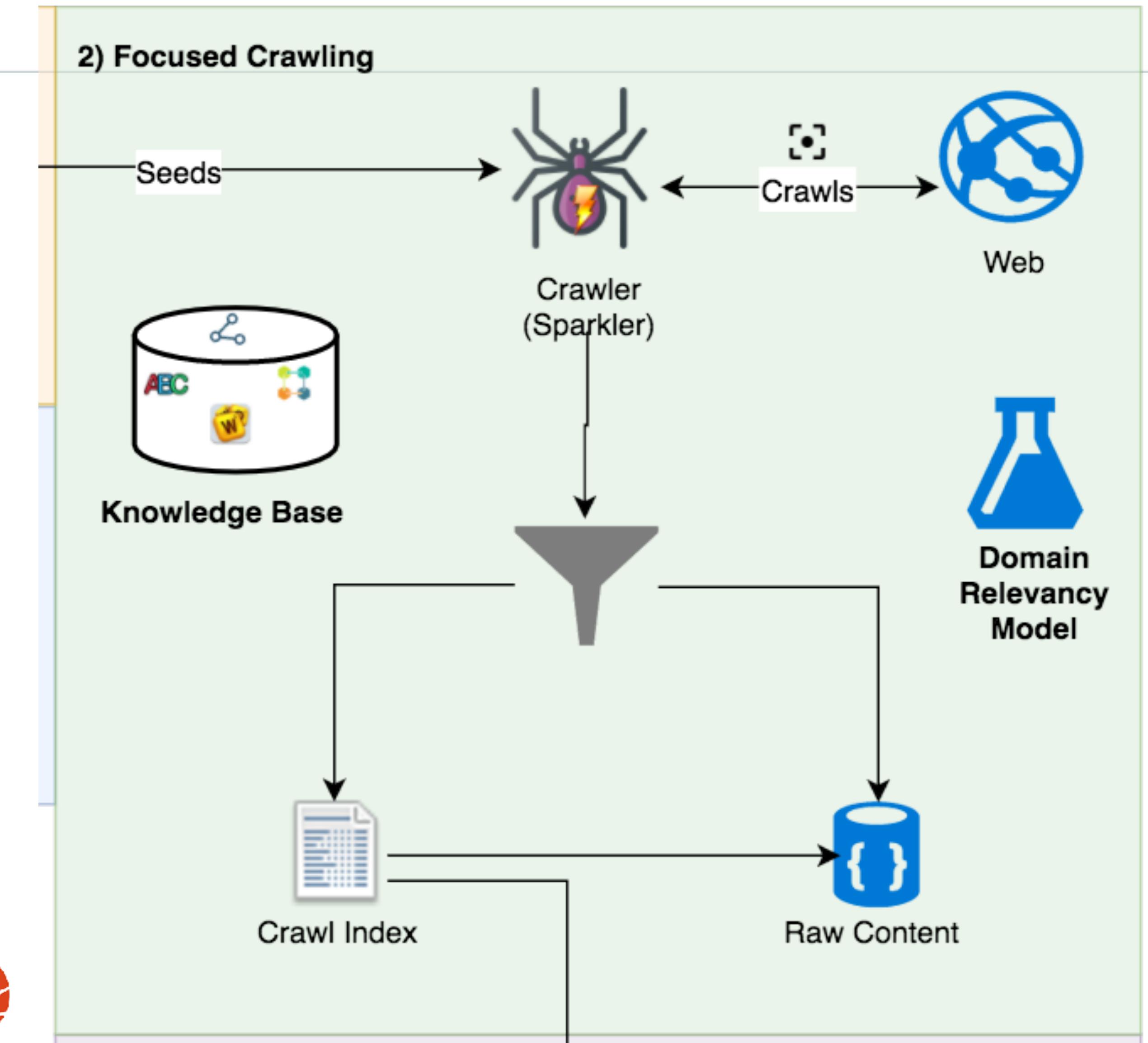
This page has a content security policy that prevents it from being loaded in this way.  
Firefox prevented this page from loading in this way because the page has a content security policy that disallows it.

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

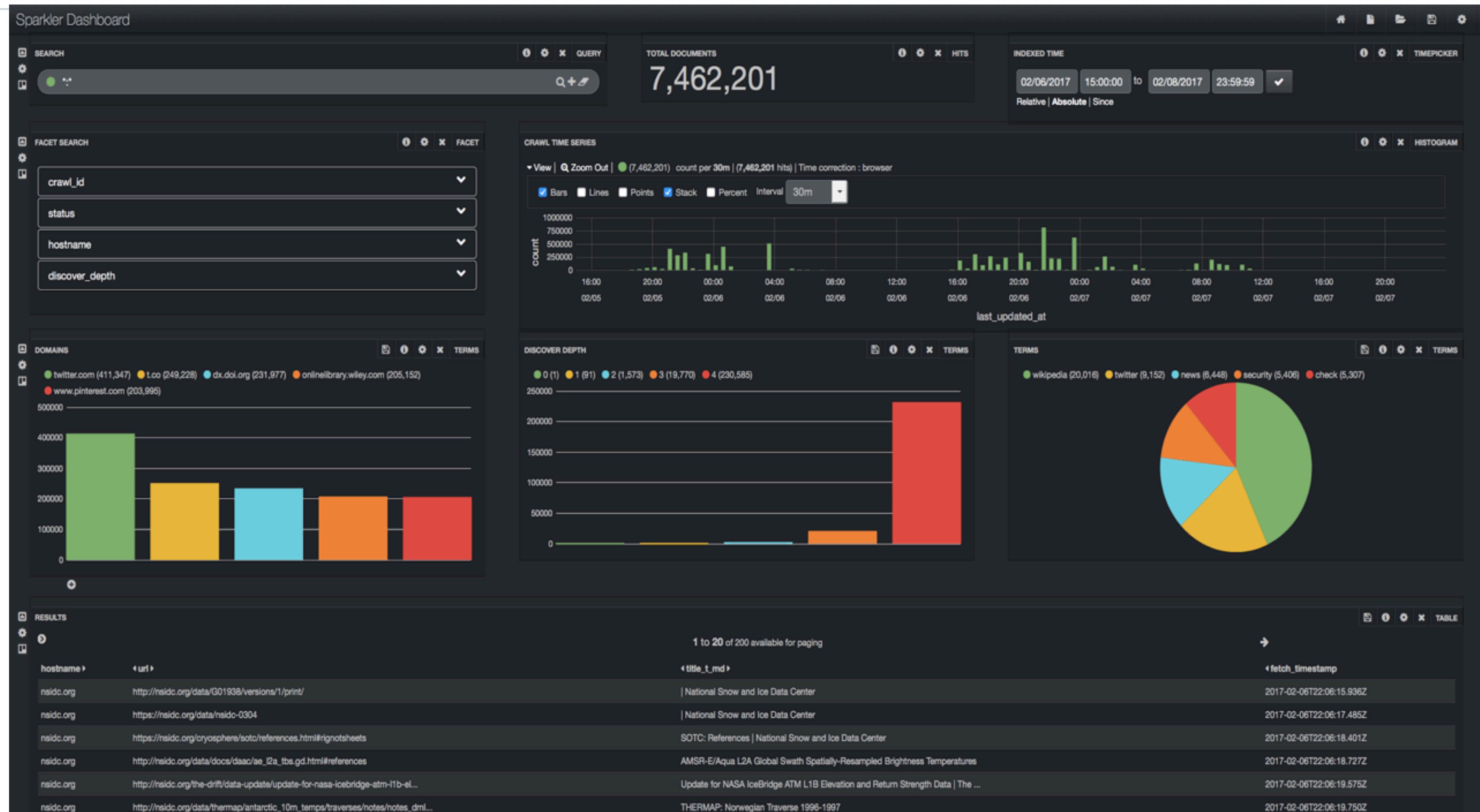
**Title: Ocean - Best Practices - Data.gov**  
**URL: https://www.data.gov/ocean/best-pra**

# 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

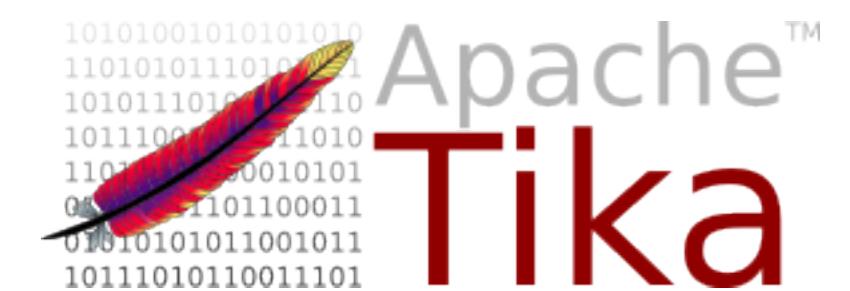
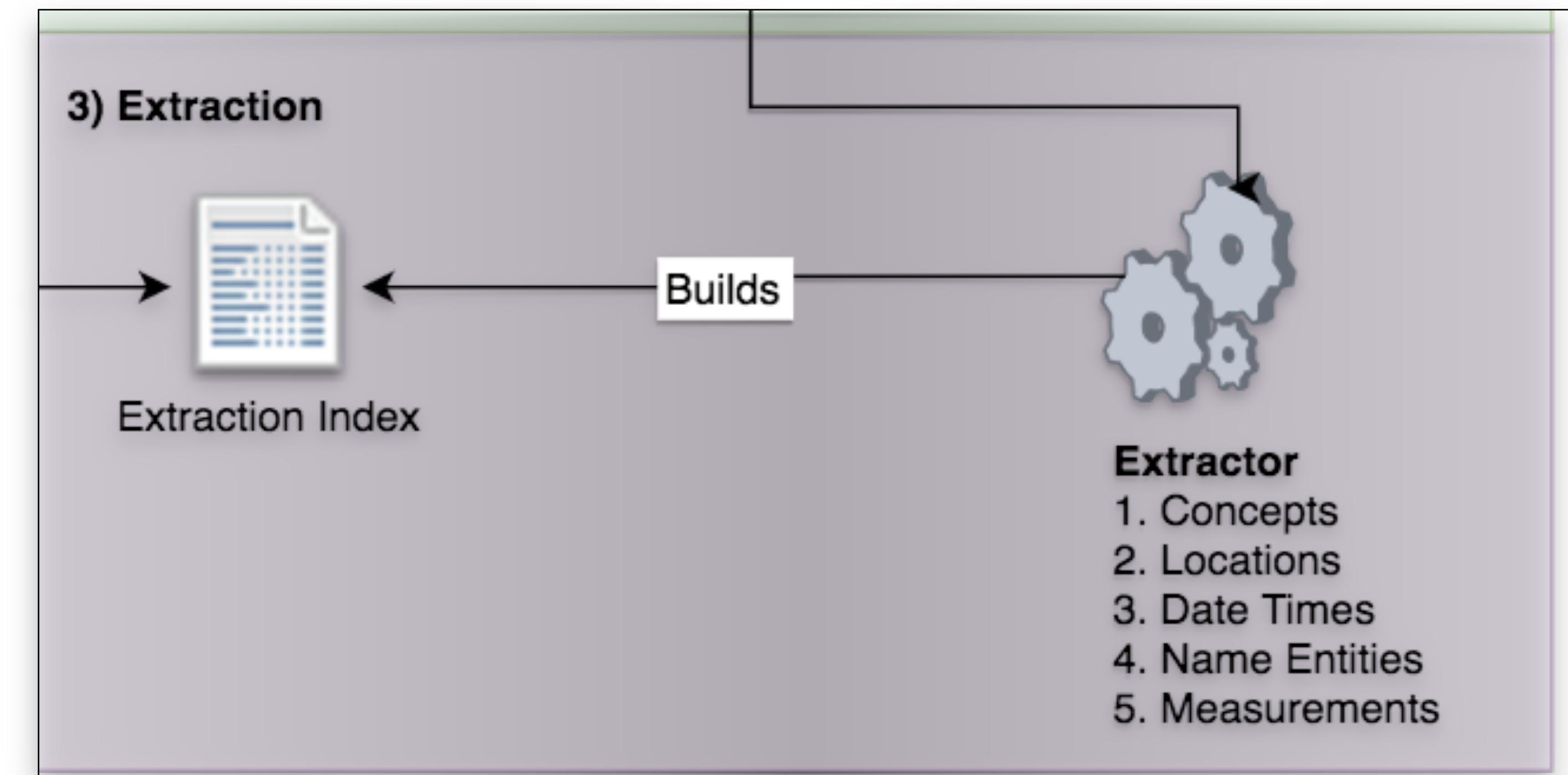


# Focused Crawling Dashboard



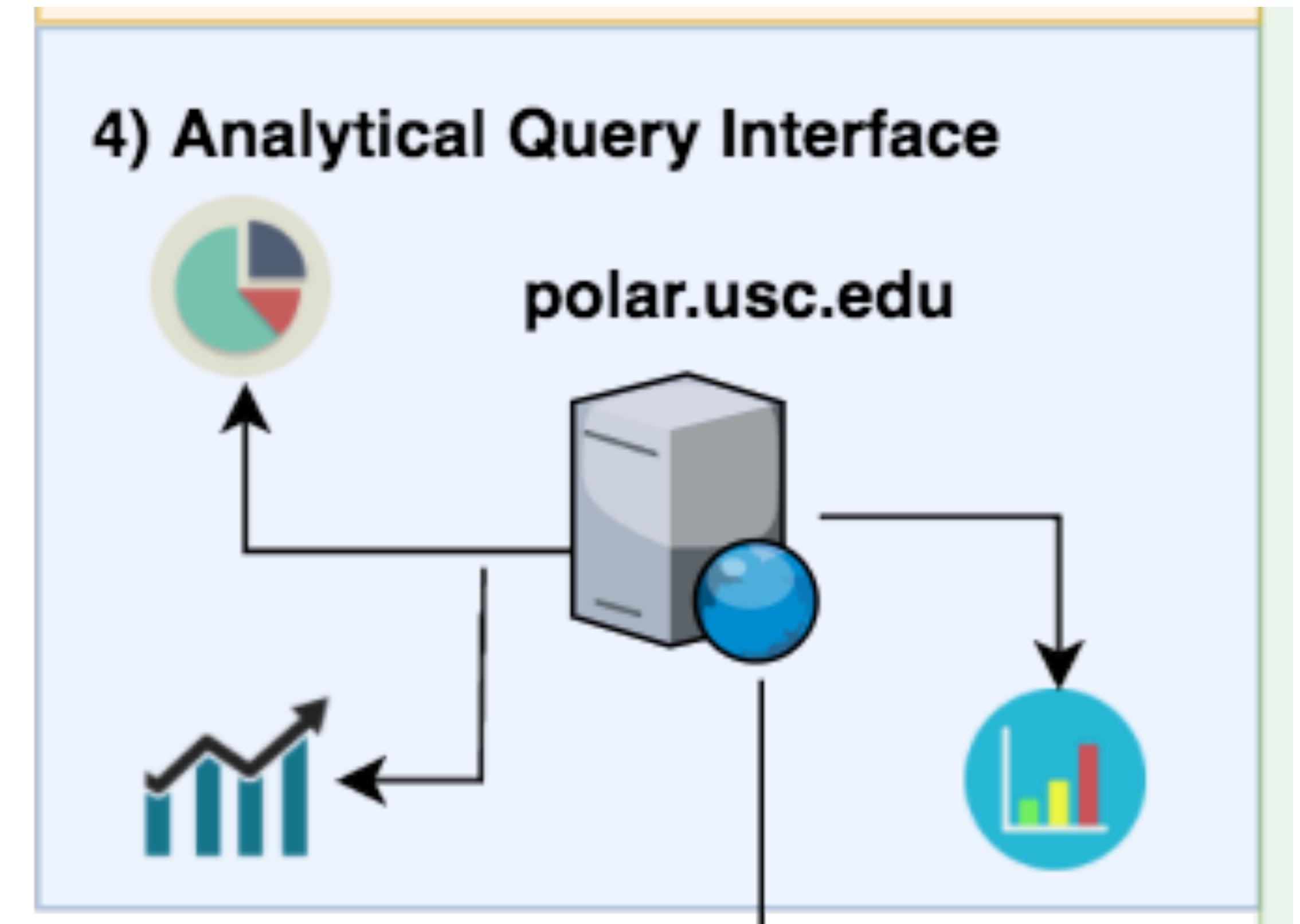
# Extraction

- 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

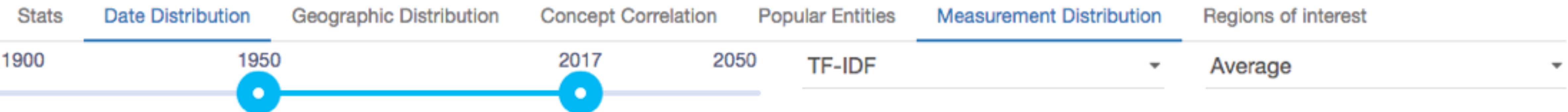


# Analytical Query Interfaces

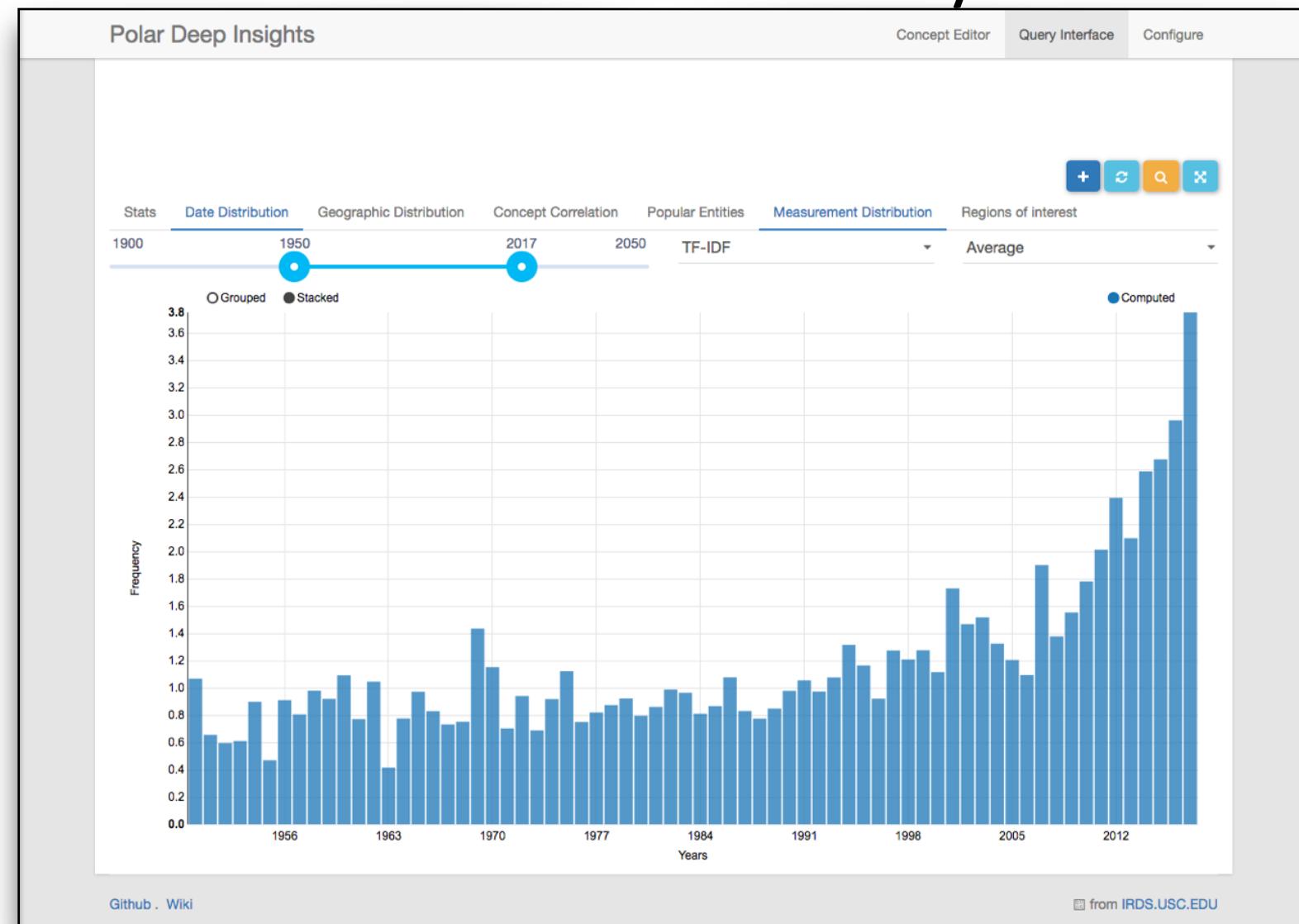
- This vast store of information is of little use without an efficient and intuitive means of querying it
- Polar Data Insights is experimenting with various tools that an user can interact with through different dashboards to query and visualize the data
  - Banana
  - FacetView



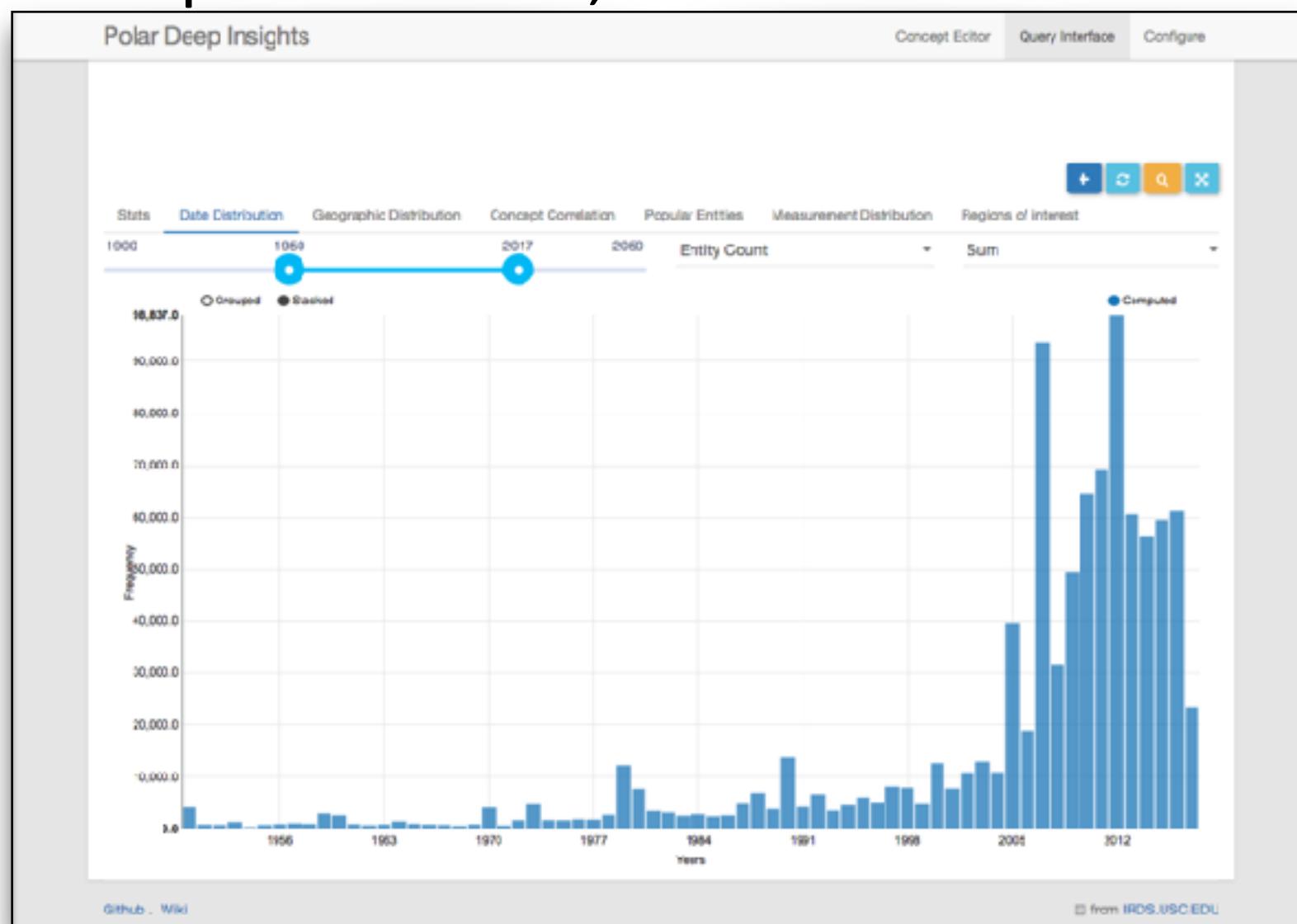
# Polar Deep Insights - Banana based query and analysis



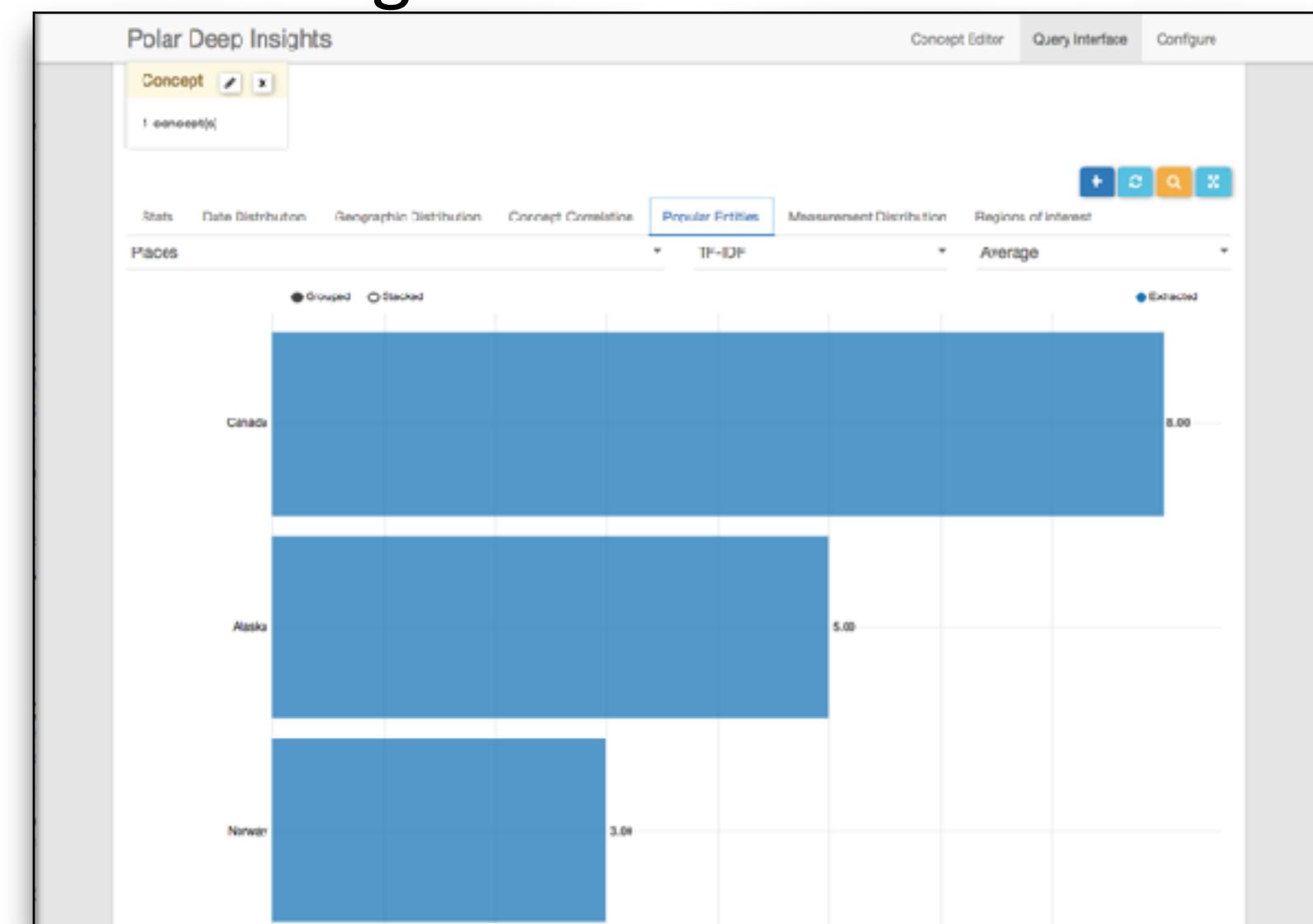
## Distribution of documents by date



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.



# 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 address bar containing `jplmemex-sce4.dyndns.org/facetview/index.html?source=undefined`. The search query in the search bar is `standards ocean sensors`. Below the search bar, there are buttons for `sea` and `arctic`. The main content area displays two search results:

- 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

# 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 multiple tabs open. The active tab displays a search results page for documents related to the Arctic and Sea. The results are presented in a faceted view, with three main items listed:

- Frost flowers growing in the Arctic ocean-atmosphere-sea ice-snow interface: 1.**  
Chemical composition - Douglas - 2012 - Journal of Geophysical Research: Atmospheres - Wiley Online Library  
Date Added: 2017-10-30T17:21:23.554Z
- Assessing the potential impacts of declining Arctic sea ice cover on the photochemical degradation of dissolved organic matter in the Chukchi and Beaufort Seas - Logvinova - 2015 - Journal of Geophysical Research: Biogeosciences - Wiley Online Library**  
Date Added: 2017-11-08T00:00:06.327Z
- Annual cycles of pCO<sub>2</sub>sw in the southeastern Beaufort Sea: New understandings of air-sea CO<sub>2</sub> exchange in arctic polynya regions - Else - 2012 - Journal of Geophysical Research: Oceans - Wiley Online Library**  
Date Added: 2017-11-09T11:52:33.264Z

At the bottom of the results page, there is a navigation bar with links for '1 – 20 of 21' and 'next »'. The browser interface includes a sidebar with a '+ URL' button and a menu bar with standard options like File, Edit, View, and History.

# 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 displays three separate Chrome browser windows side-by-side, all showing search results for the query "standards ocean sensors".

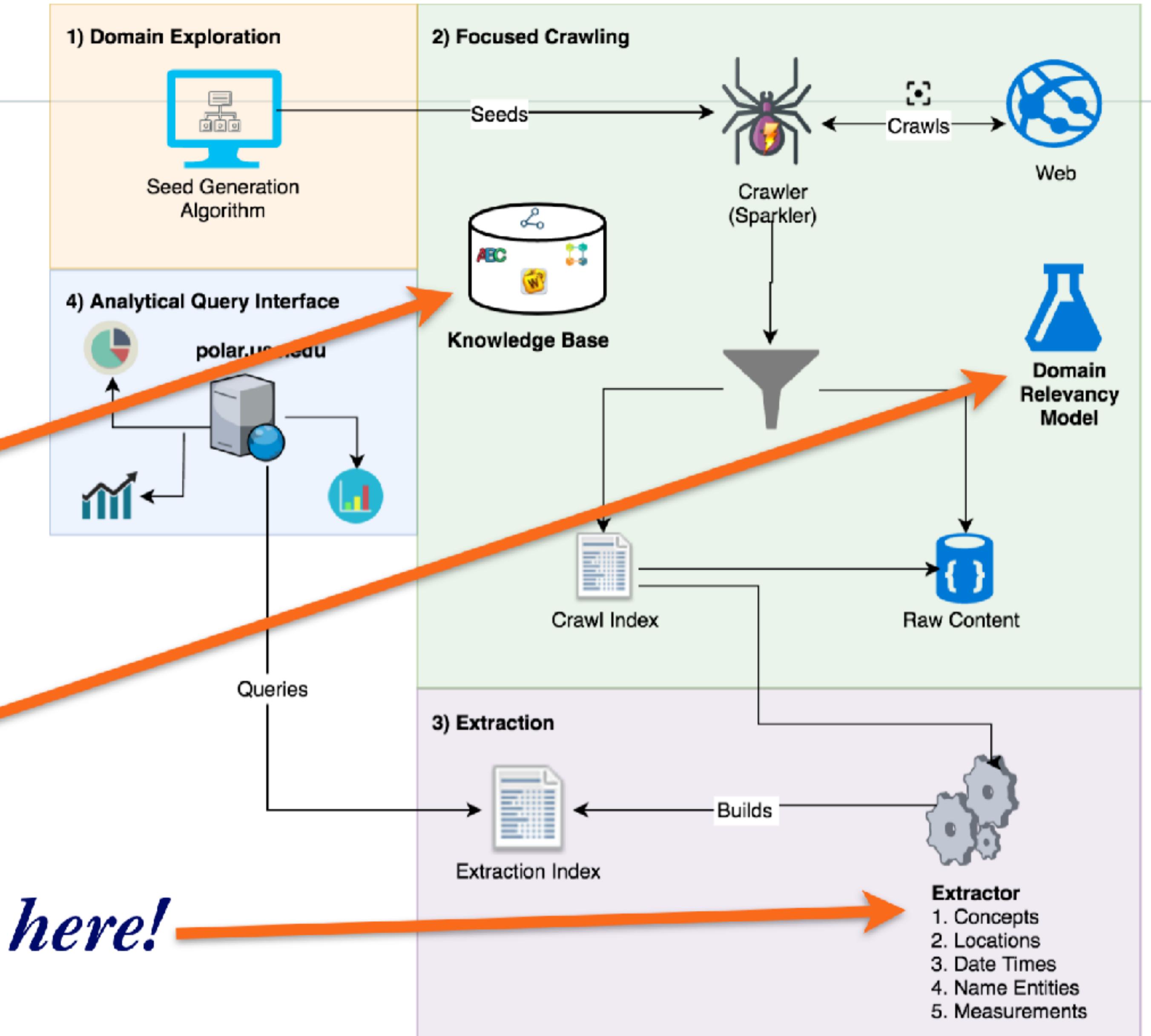
- Left Window:** Shows a search interface with a URL input field containing "jplmemex-sce4.dyndns.org". Below it are sections for "Assessing photochemistry", "Beaufort", and "Annual cycles of air-sea C".
- Middle Window:** Shows a search result titled "Frost flowers g Chemical co". It includes a sidebar with PDF, Info, References, and Figures options.
- Right Window:** Shows a search result from "onlinelibrary.wiley.com" with the URL "doi:10.1002/2015JG003052/full". The main content discusses PARAFAC models, model fitting, validation, and least squares results. A section titled "2.5 Dissolved Organic Carbon Analysis" is present, mentioning standards prepared by volumetric dilution of a stock solution containing 500 µM DOC. It also refers to a Consensus Reference Material (CRM) project and analyses of the CRM.

# Semantics is Everywhere!

*Semantics is here!*

*And here!*

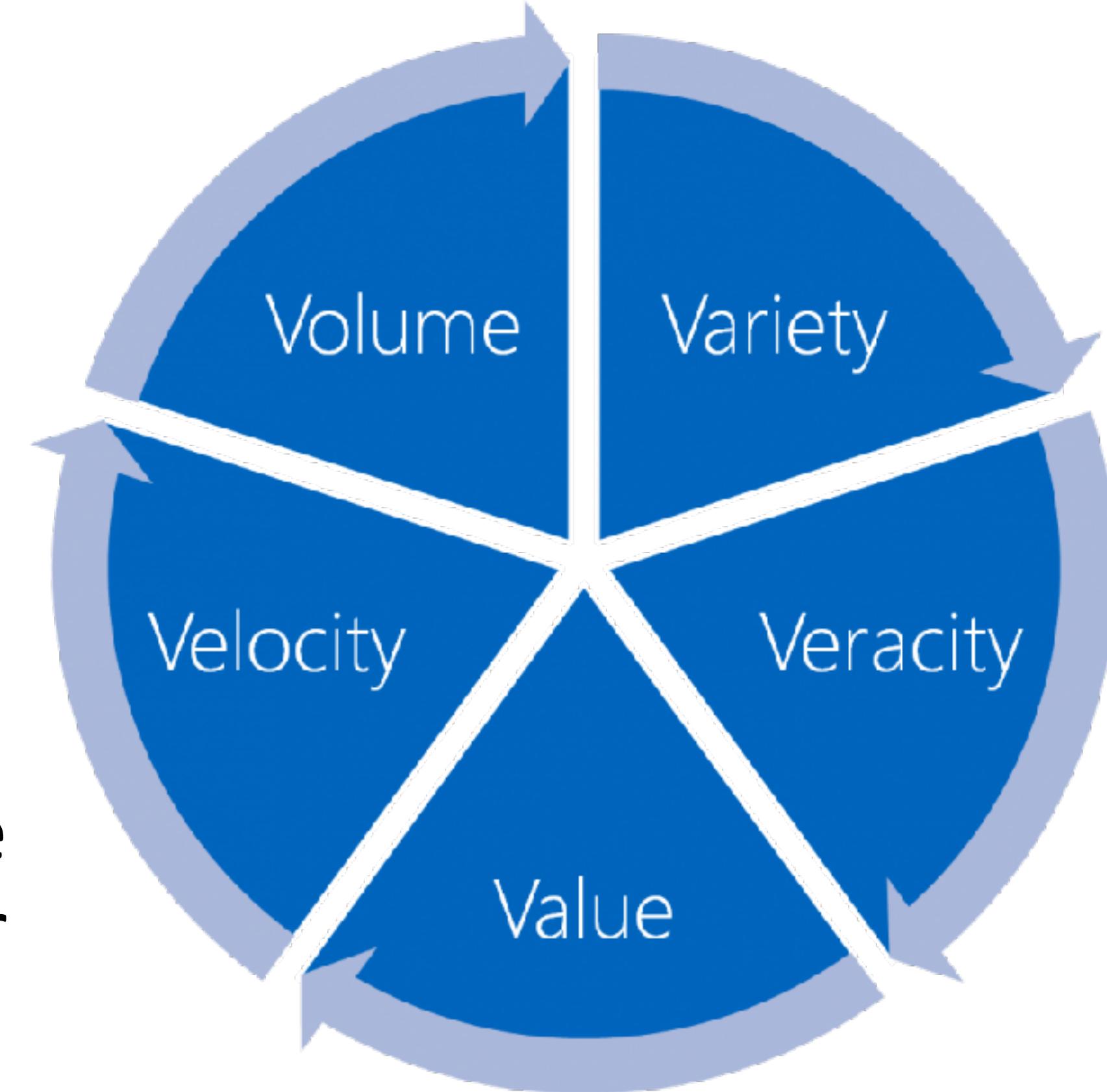
*And here!*



# *Summary and Conclusions*

---

- A new approach to data discovery and information extraction is required to make effective use of the wealth of textual and scientific data that is being generated
- An Open Source framework fosters community involvement in the development, and responsive evolution of the necessary tools
- These tools can provide the ability to address grand challenge questions concerning the state and trajectory of the Earth System and its Polar regions



## *Acknowledgements*

---

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