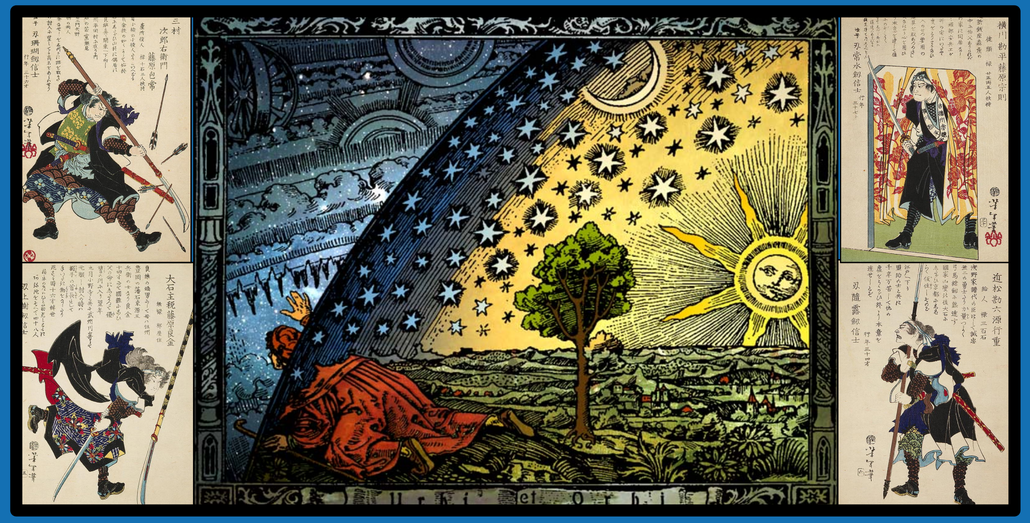
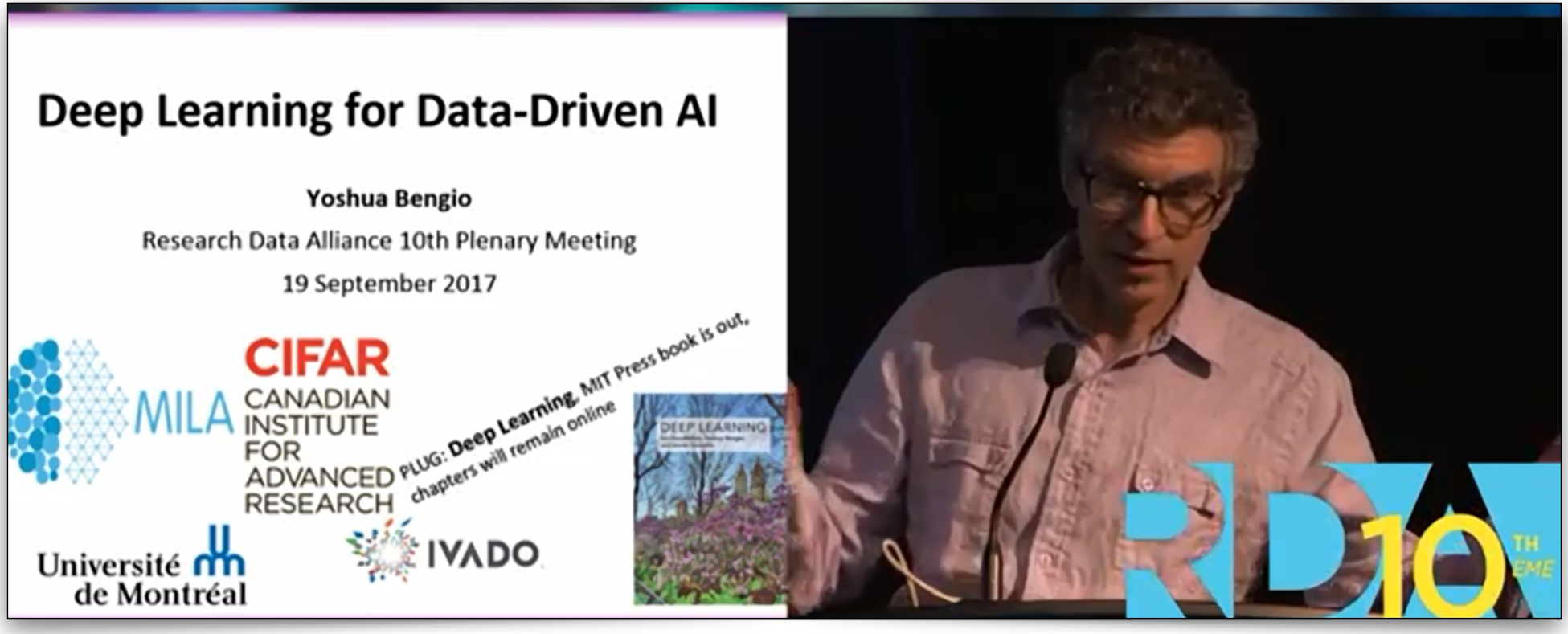


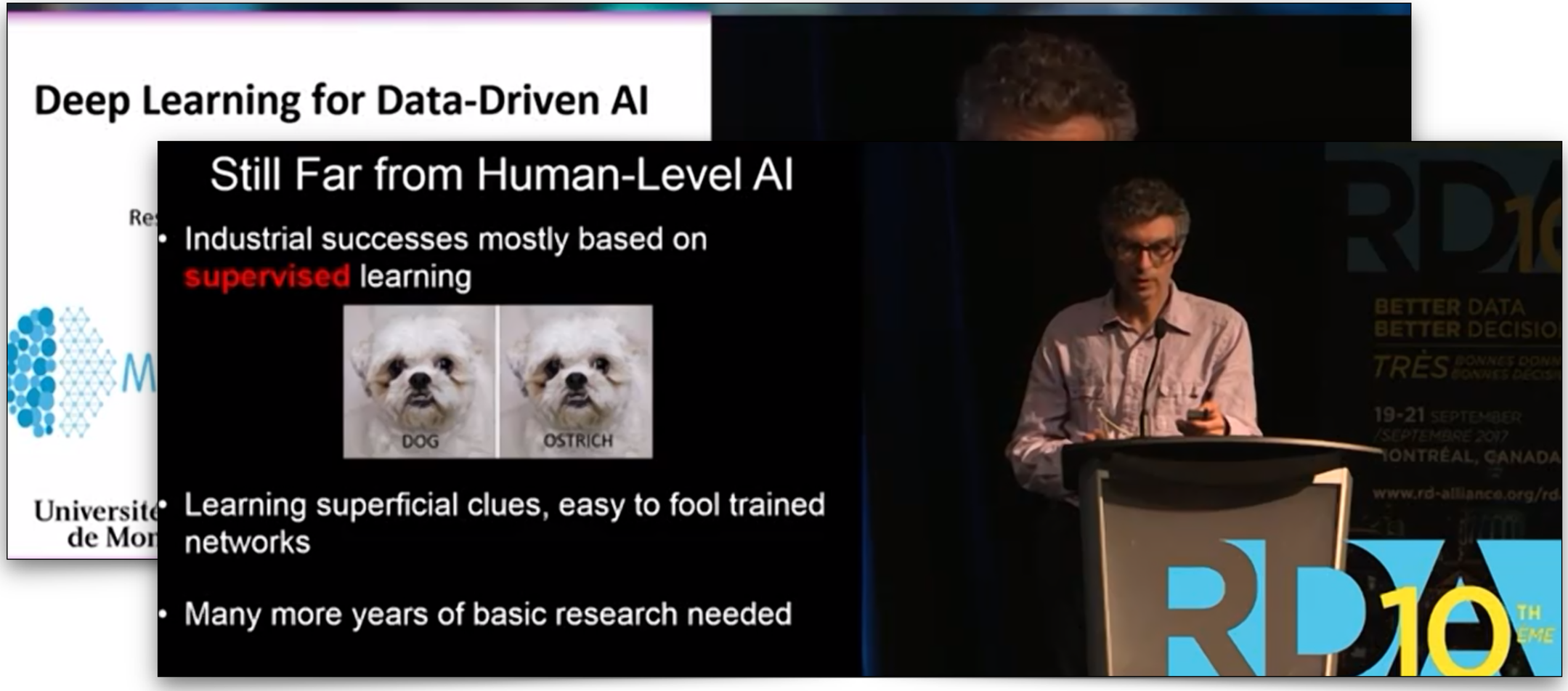
Tailoring and Analyzing Topic-specific Crawls Using Semantics and Deep Learning

Ruth Duerr

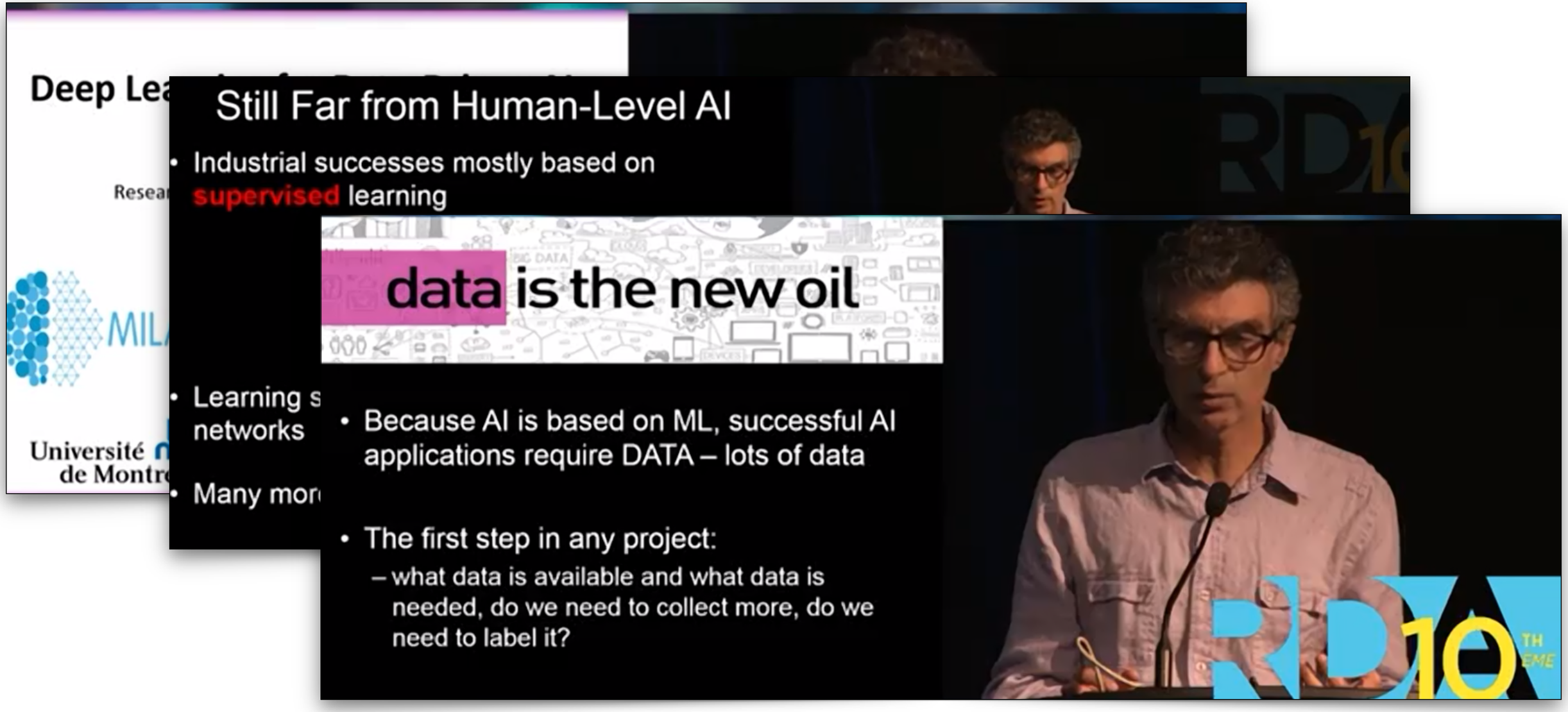
Deep Learning and Semantics



Deep Learning and Semantics



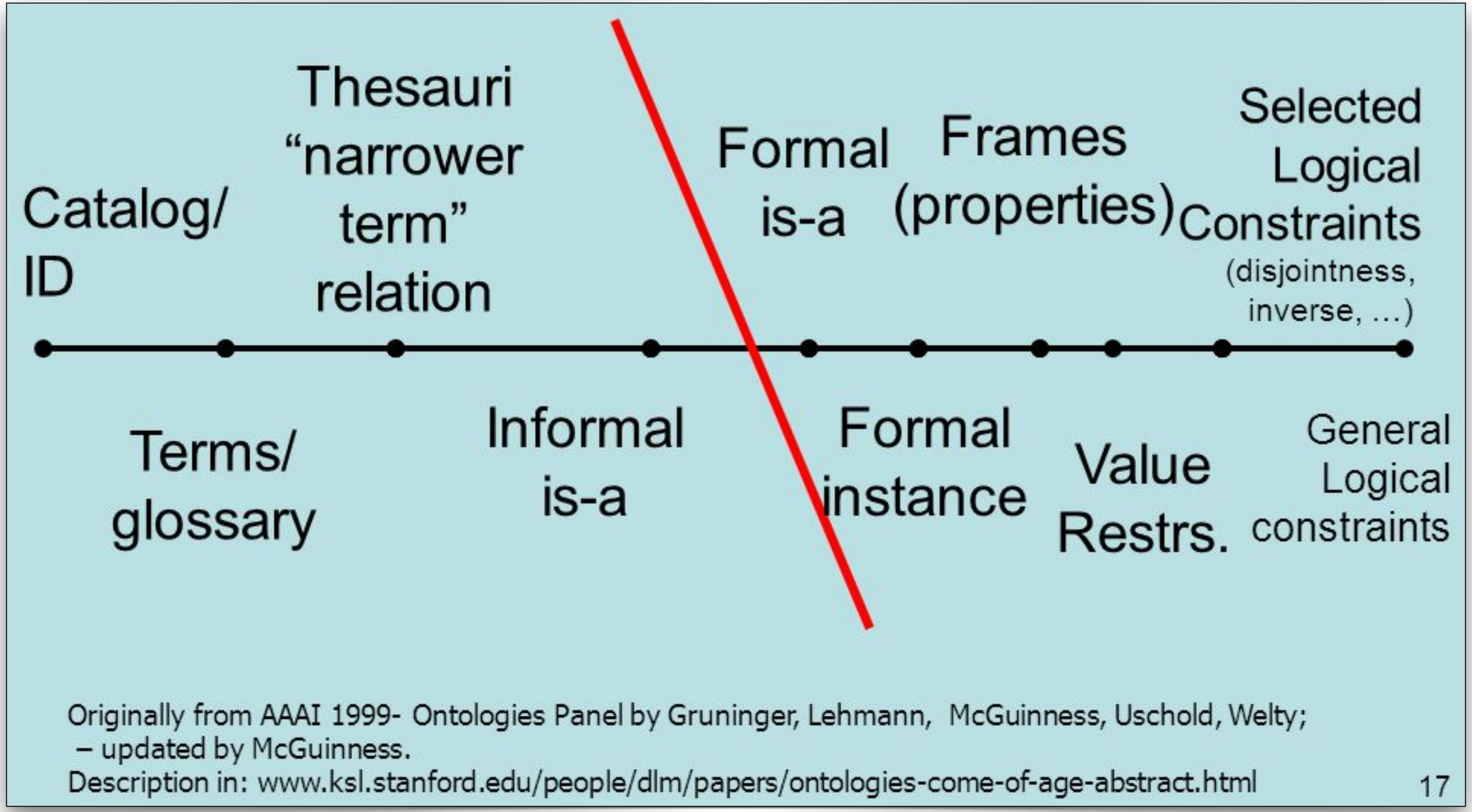
Deep Learning and Semantics



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?

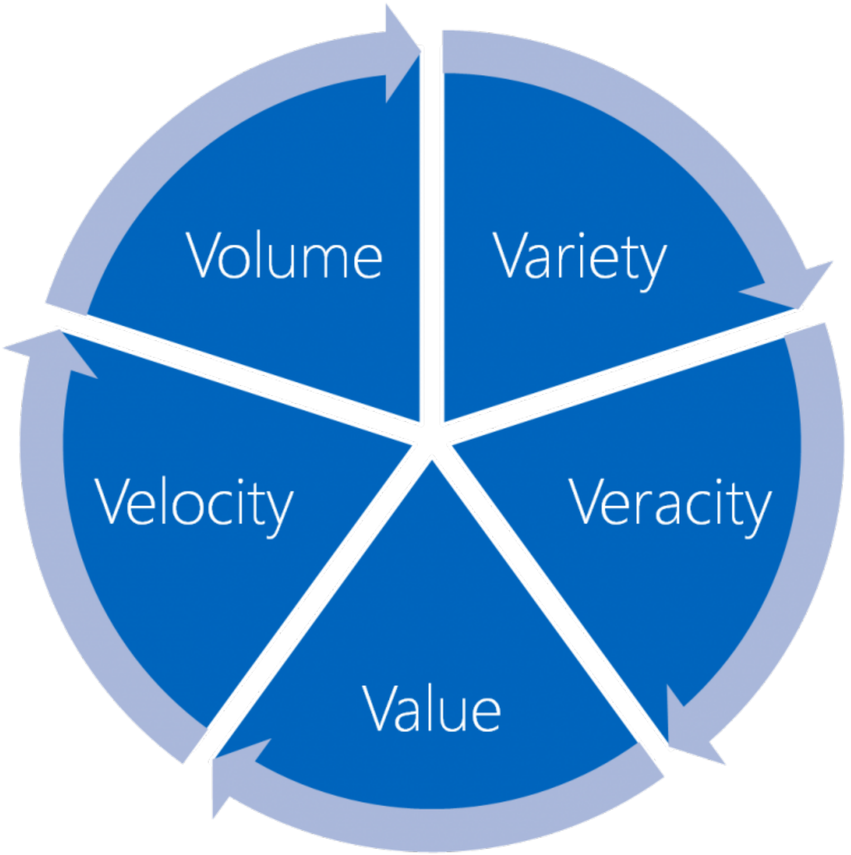


Originally from AAAI 1999- Ontologies Panel by Gruninger, Lehmann, McGuinness, Uschold, Welty; – updated by McGuinness.
Description in: www.ksl.stanford.edu/people/dlm/papers/ontologies-come-of-age-abstract.html

17

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*?



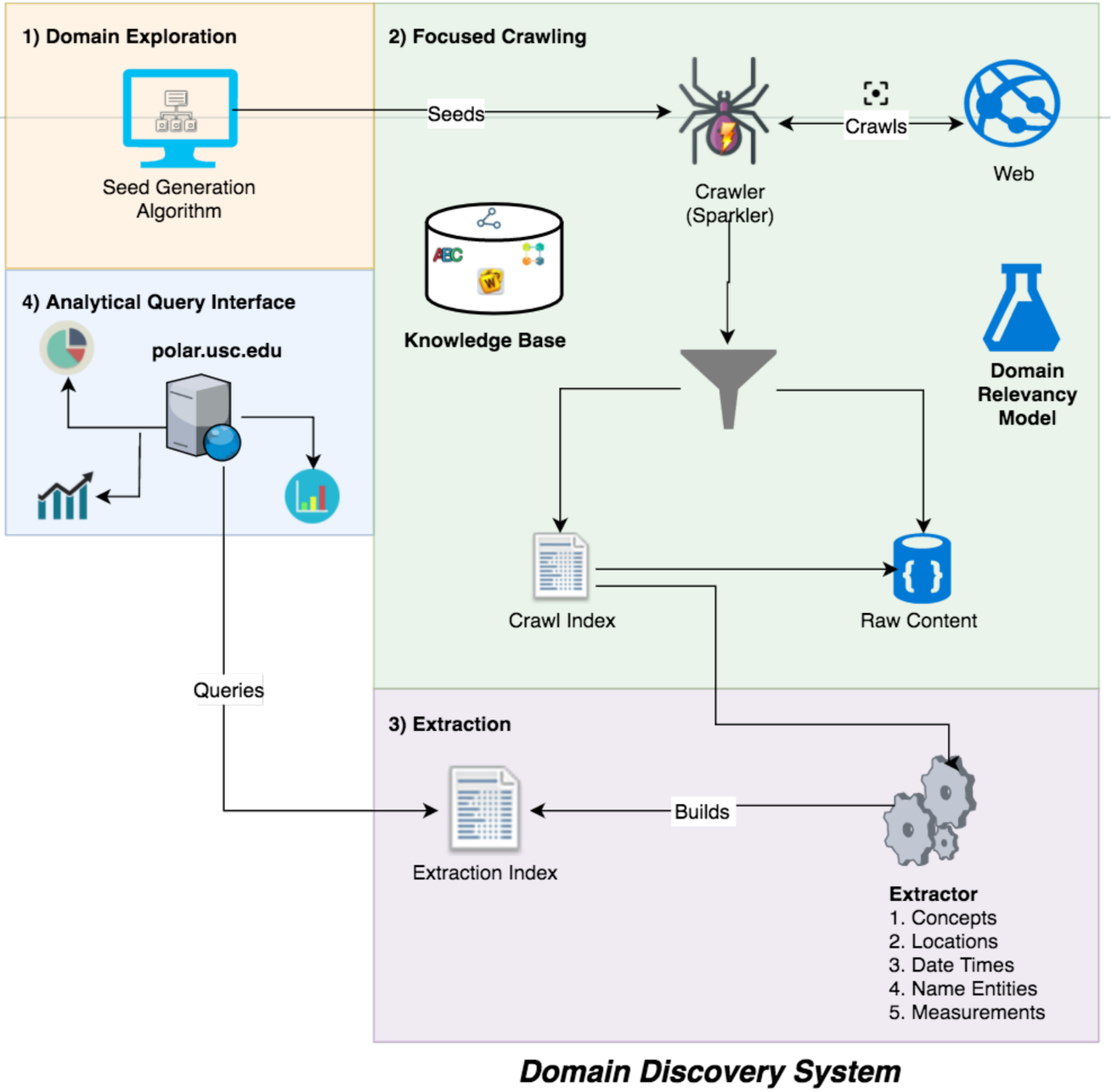
Applying "Big Data" Technology to Domain (here Polar) Data

- 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)

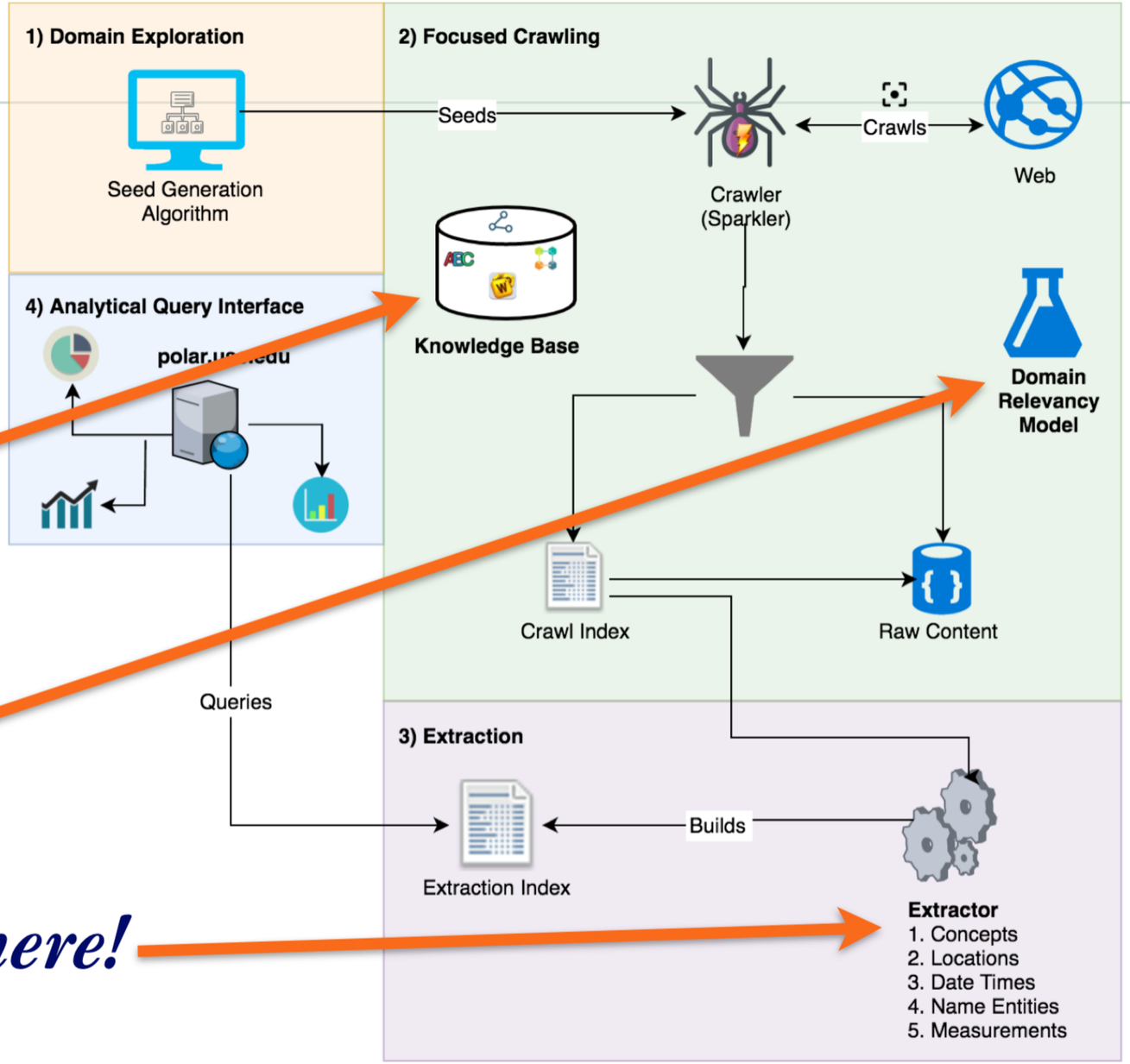
Polar Deep Insights Architecture

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



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

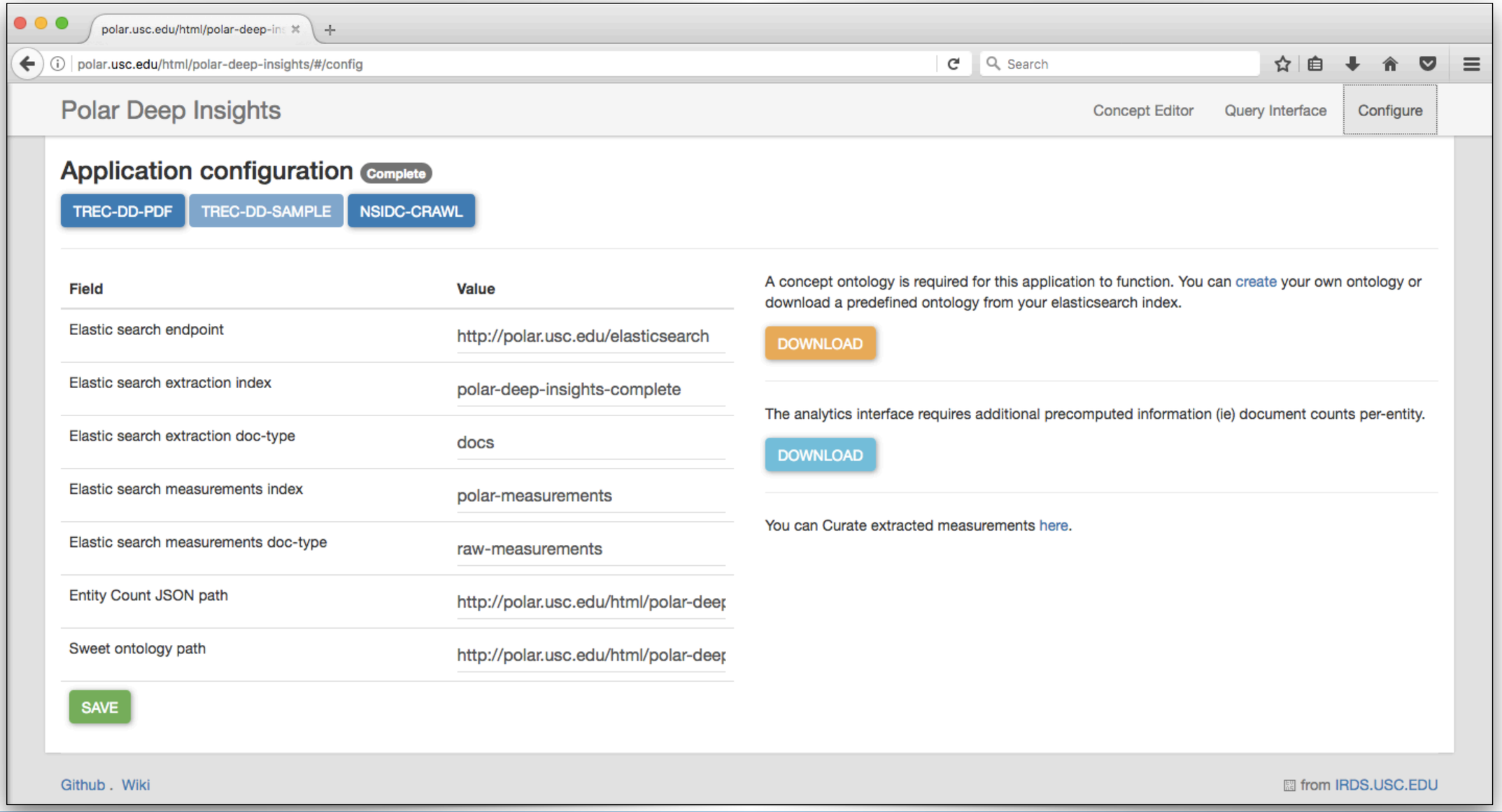
- http://arcticportal.org/
- http://ipa.arcticportal.org/
- http://nsidc.org/data/
- https://arcticdata.io/
- https://www.data.gov/climate/arctic-data/
- https://www.bas.ac.uk/data/

Search Terms

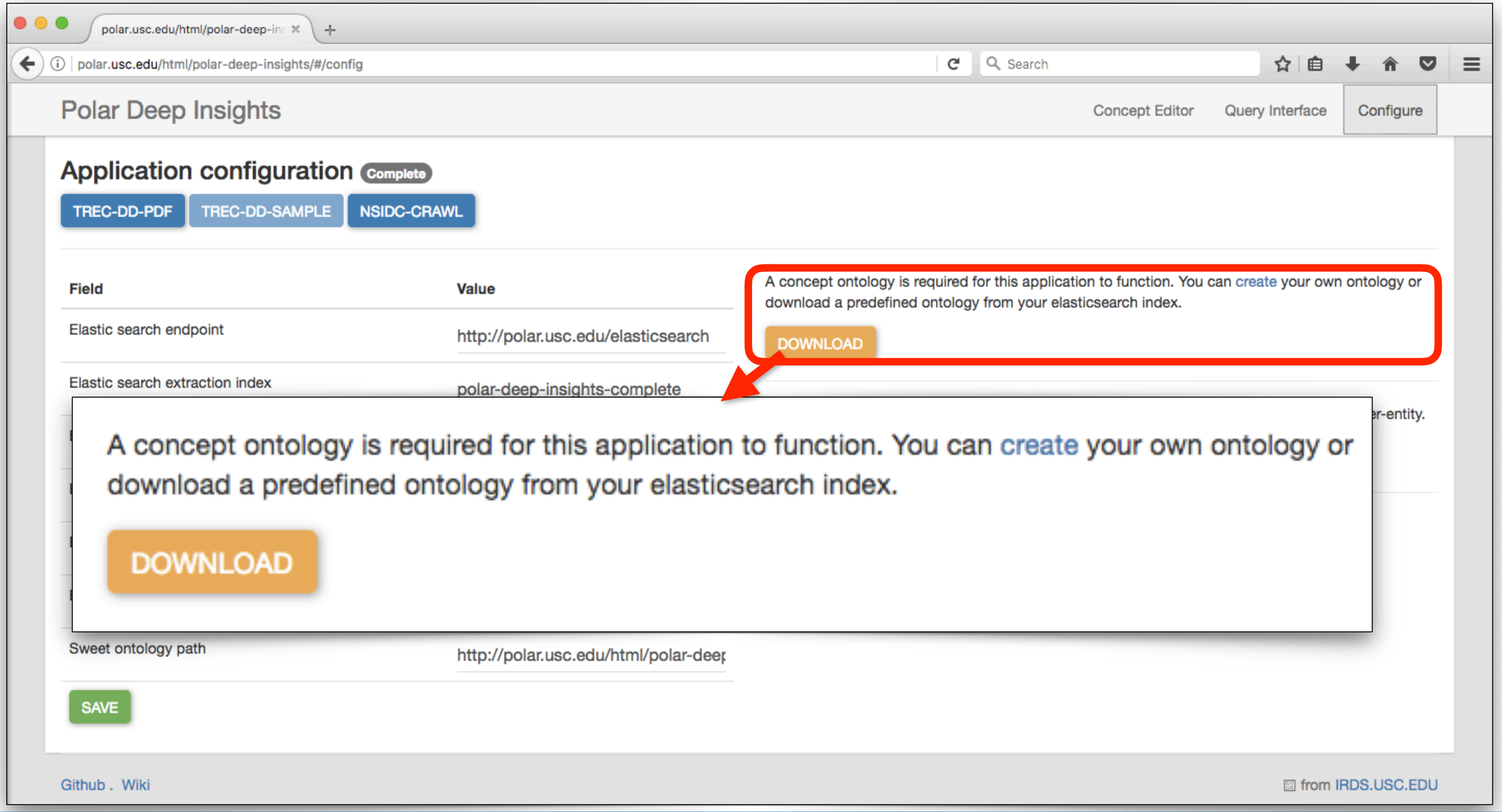
- sea ice
- permafrost
- glacier
- ice sheet
- polar
- Arctic
- Antarctic
- snow
- sea ice thickness
- navigability
- calving
- ice berg
- snow water equivalent
- mass balance
- albedo

- 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

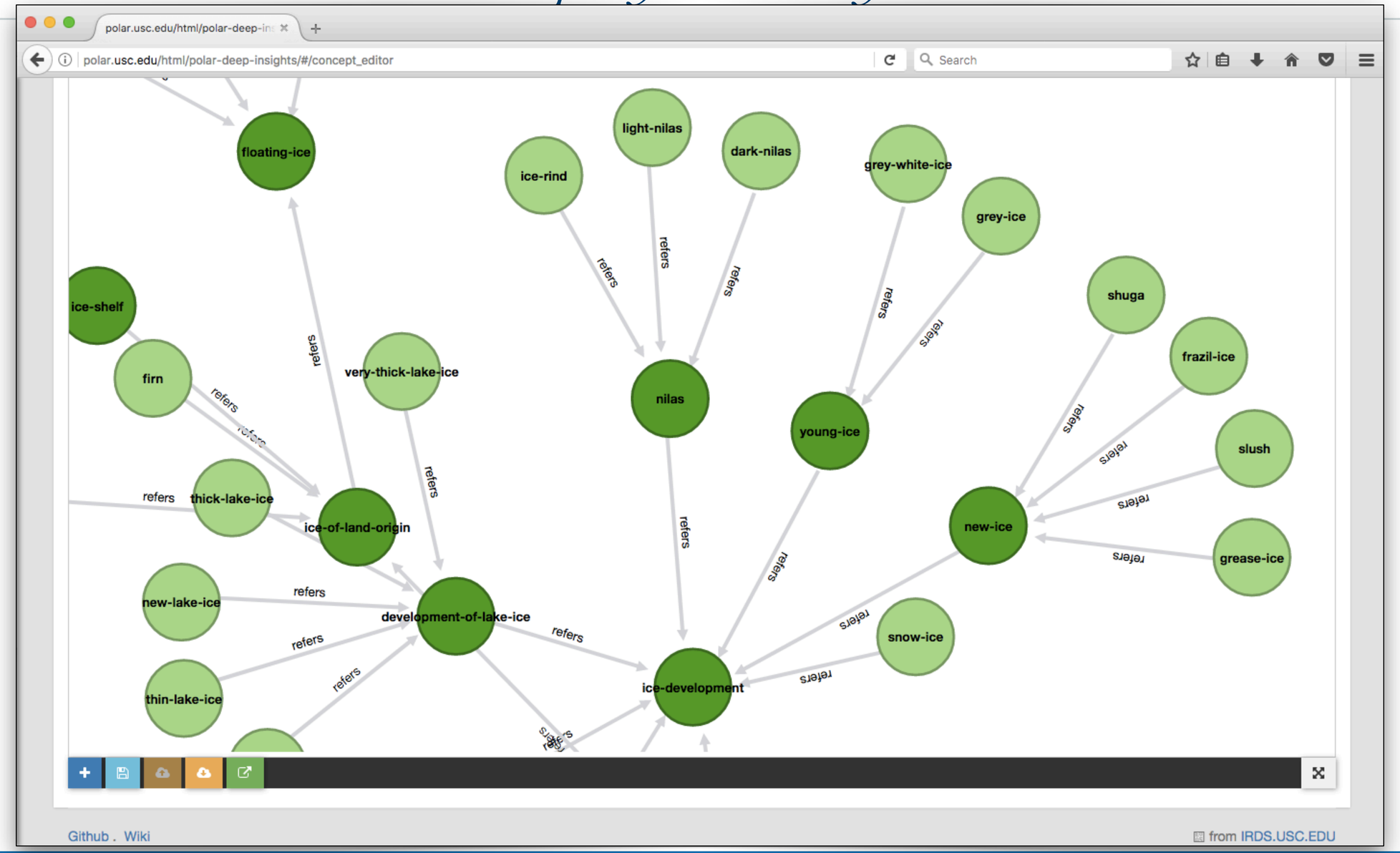
Sea Ice Use Case - Banana based query and analysis



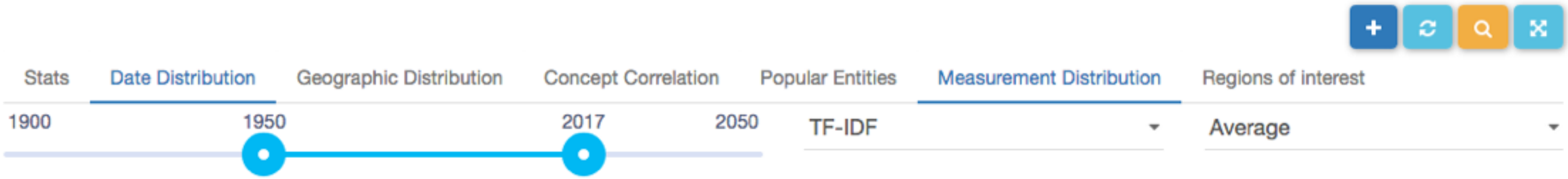
Sea Ice Use Case - Banana based query and analysis



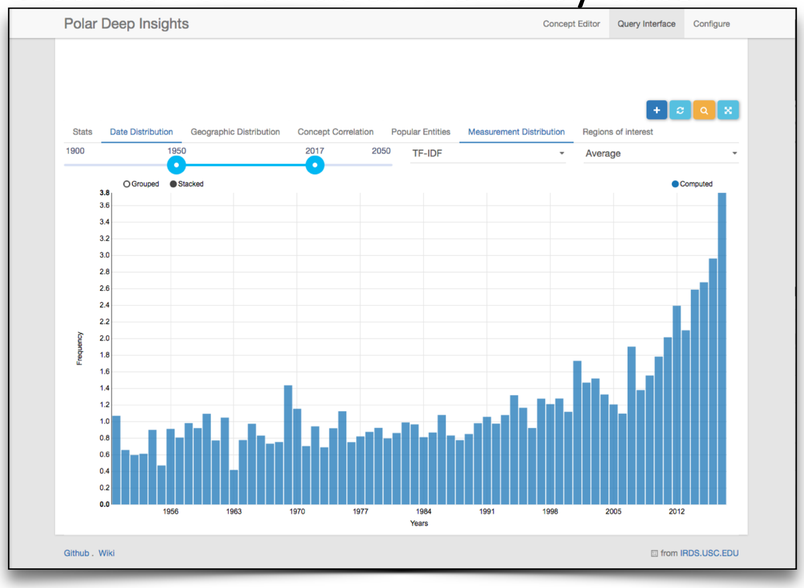
Sea Ice Use Case - Banana based query and analysis



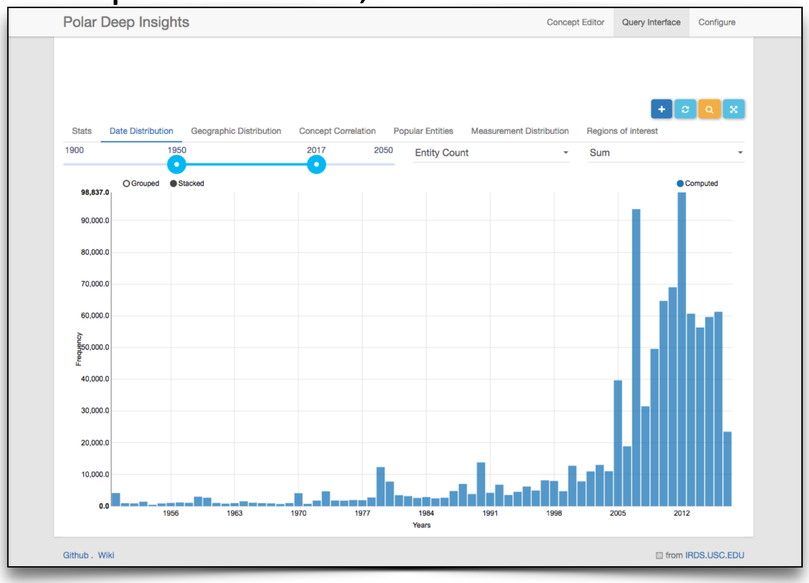
Sea Ice Use Case - Banana based query and analysis



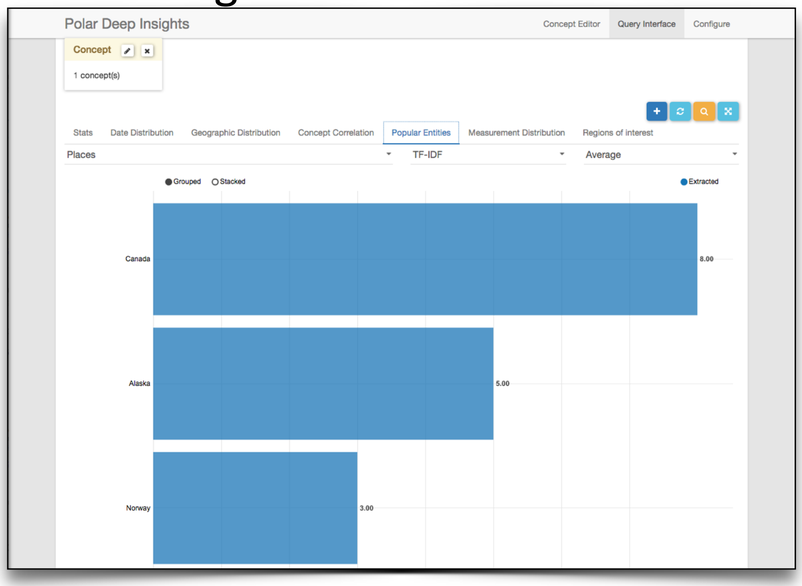
Distribution of documents by date



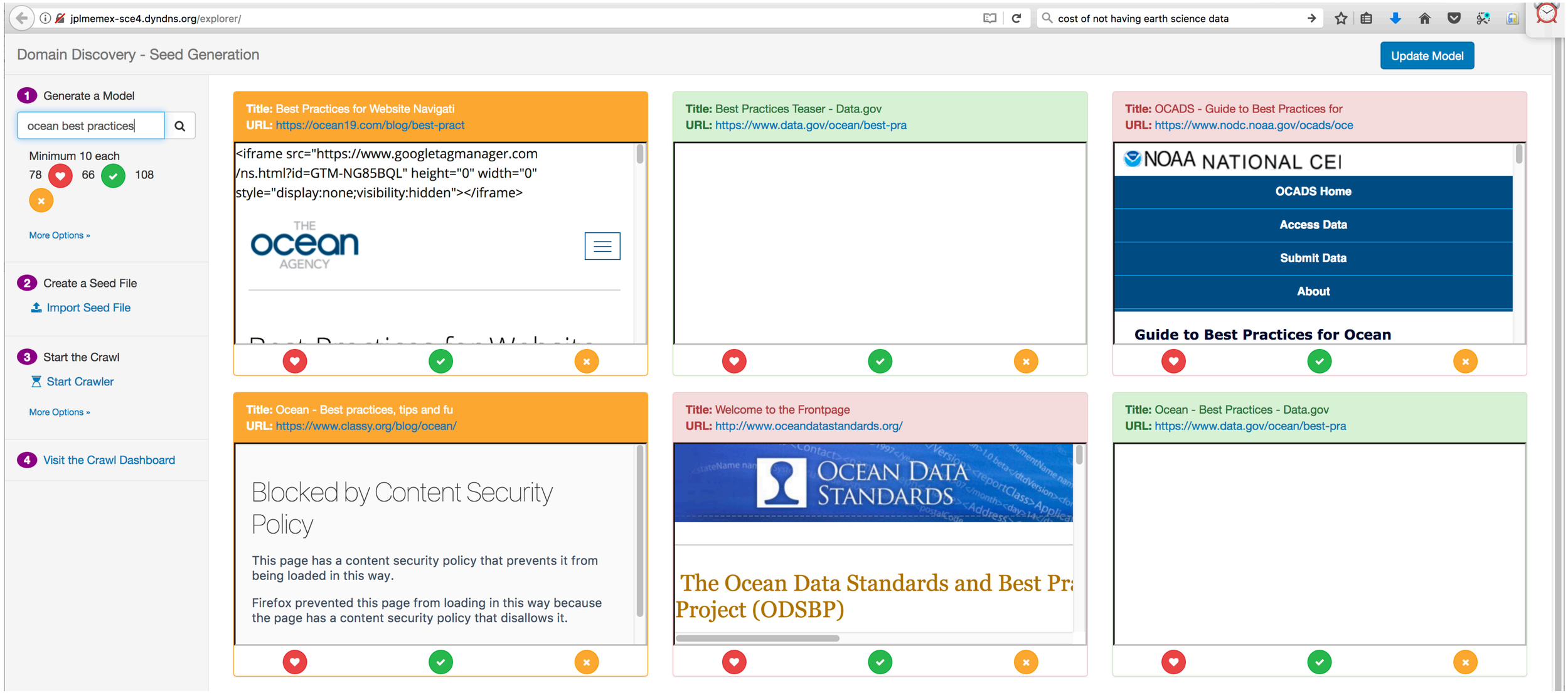
Distribution of documents that mention icebergs again by date - What's up with the spikes in 2005, 2007 and 2012?



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

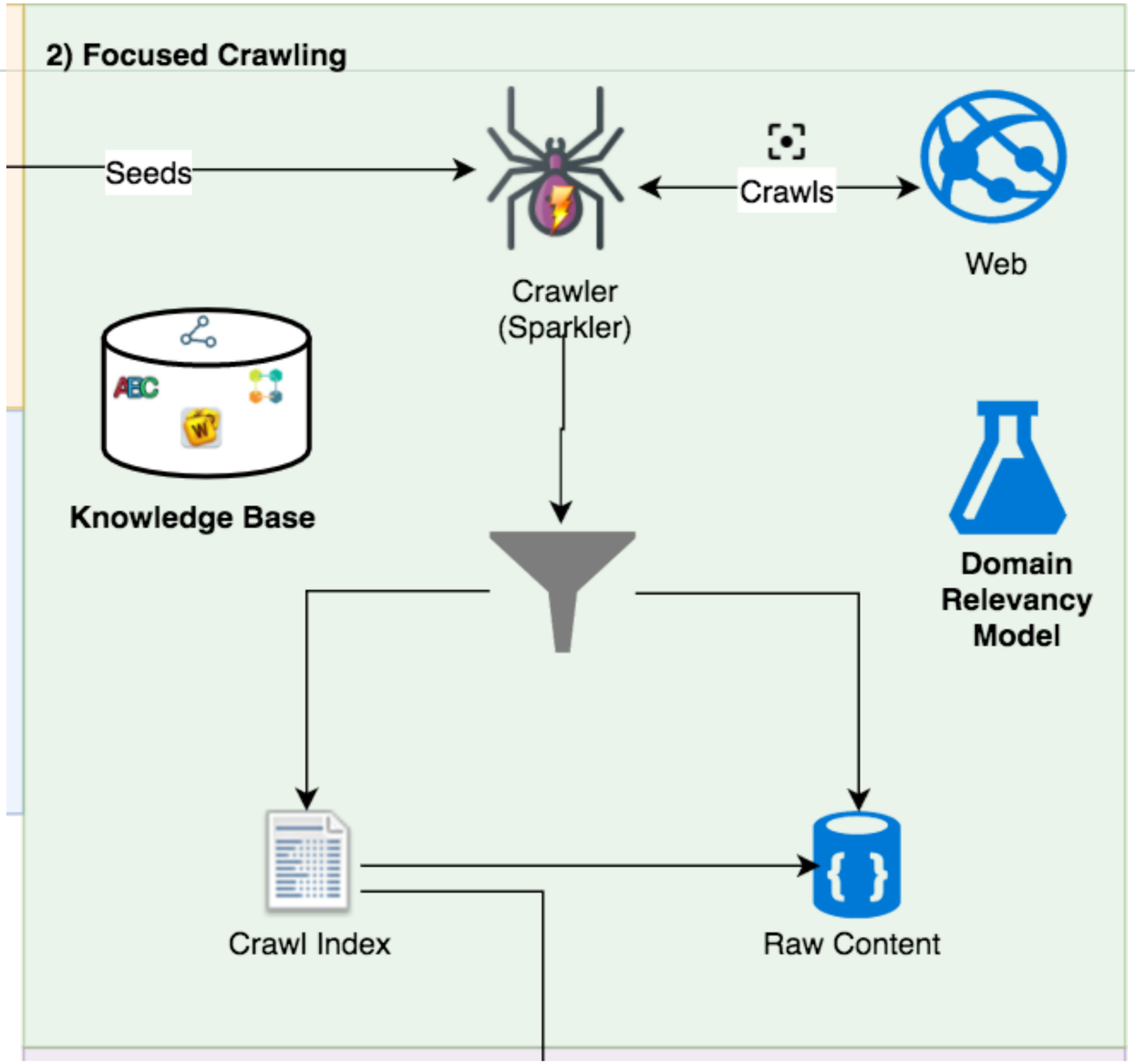


Domain Exploration - Semi-automated Model Generation



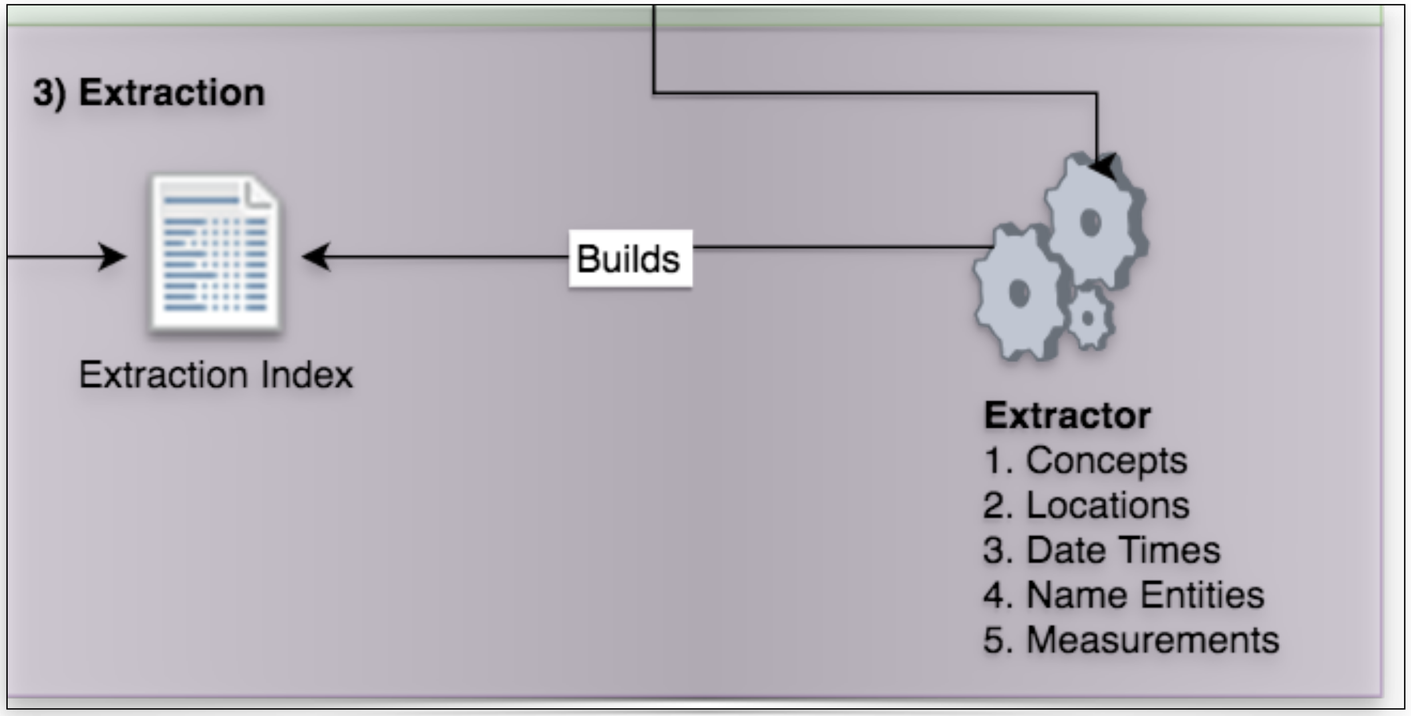
Focused Crawling

- 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

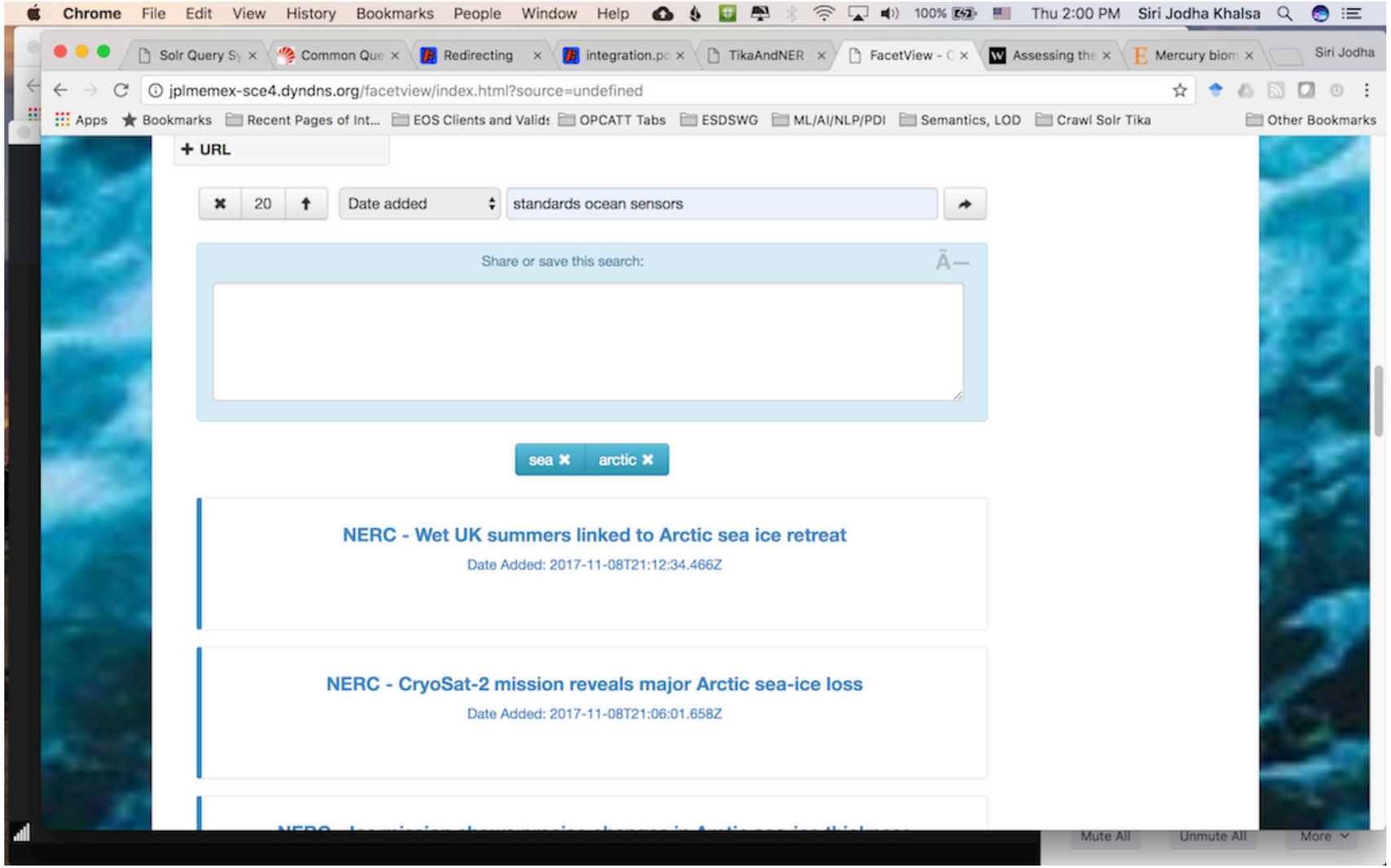
- Detects and extracts metadata, text, URLs
- Toolkit of parsers to extract
  - Concepts
  - Geographic locations
  - Dates and Times
  - Named Entities
  - Numerical measurements
- Creates an index for the extracted content



Apache™ Tika

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”

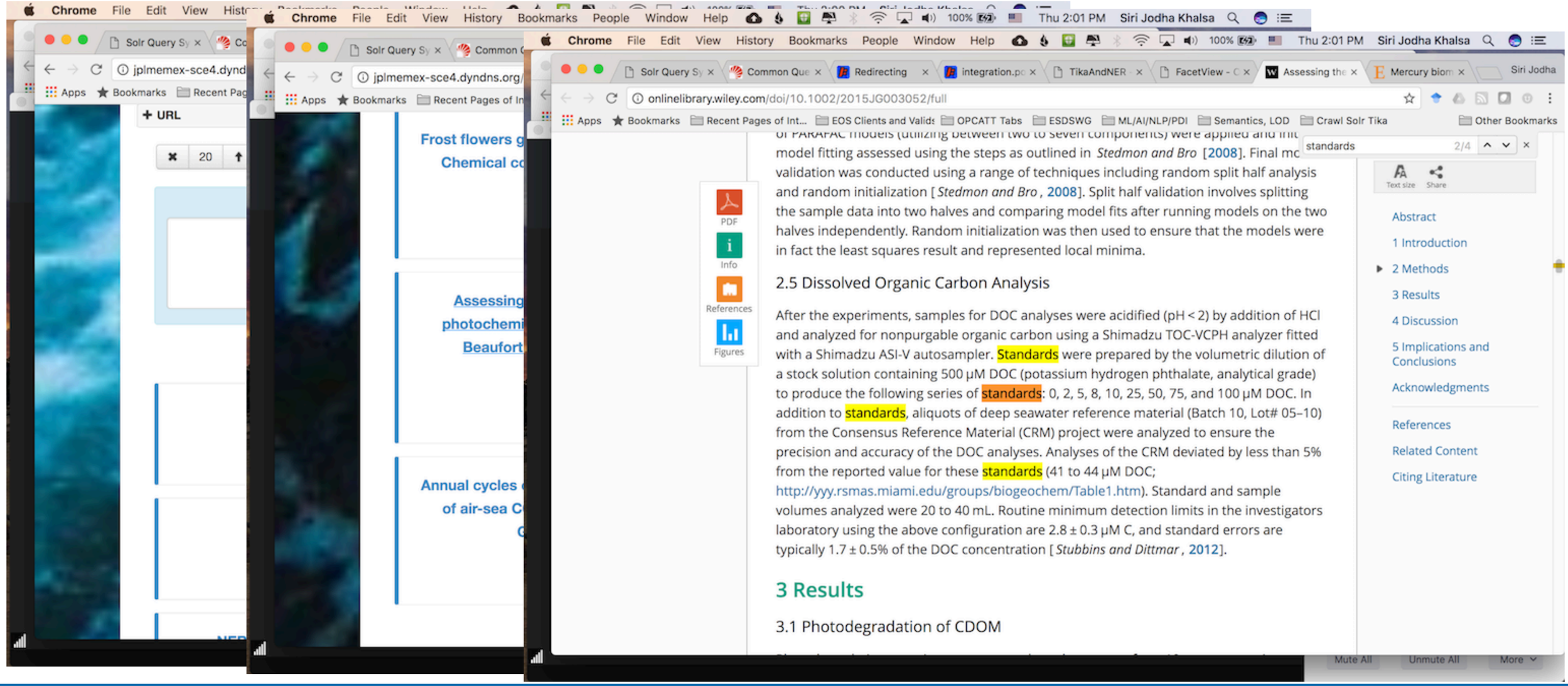


Polar Deep Insights - Facet-view based query and analysis

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

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”



Acknowledgements

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