Deep Web Crawling for Insights from Polar Data

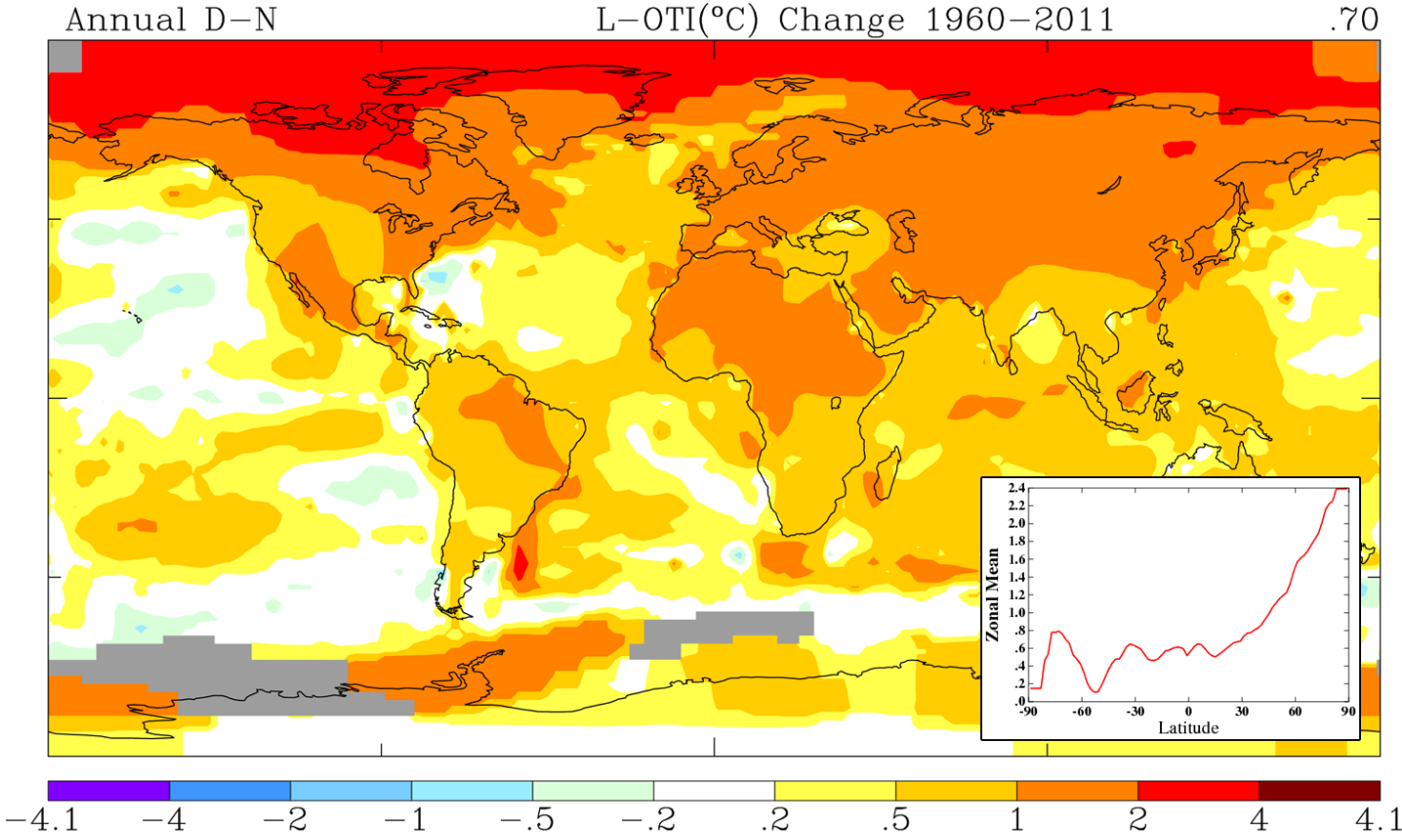
S. J. Khalsa[1], C. A. Mattman[2,4], R. E. Duerr[3]

[1]NSIDC/CIRES/U. Colorado, [2]NASA Jet Propulsion Laboratory,
[3]Ronin Institute, [4]University of Southern California

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

Making Sense of Polar Environmental Change

- The pace of environmental change at the Poles is accelerating, and the impacts of these changes are becoming more evident.
- The rate at which data and information about the global environment is being produced is also accelerating, but our ability to find and make sense of this information is lagging.
- Environmental information is no longer confined to major data repositories, but is distributed and heterogeneous, with much of it accessible on the Web.
- **New approaches to finding and making sense of this information are needed.**



Project Goals

- Funded by NSF, the **Polar Deep Insights** project is building a system to collect, analyze and make accessible interactively the wealth of numerical, textual and multimedia Polar information on the scientific Web.
- We are developing new information retrieval and analytics techniques that will link these varied form of information, bringing together unstructured and structured data to produce transformative insights.
- Major innovations include
  - A process for creating a Domain Relevance Model using Machine Learning
  - New methods of focused crawling using OS tools
  - Advanced methods for extracting content from web pages, documents and multimedia files.

Methods and Tools

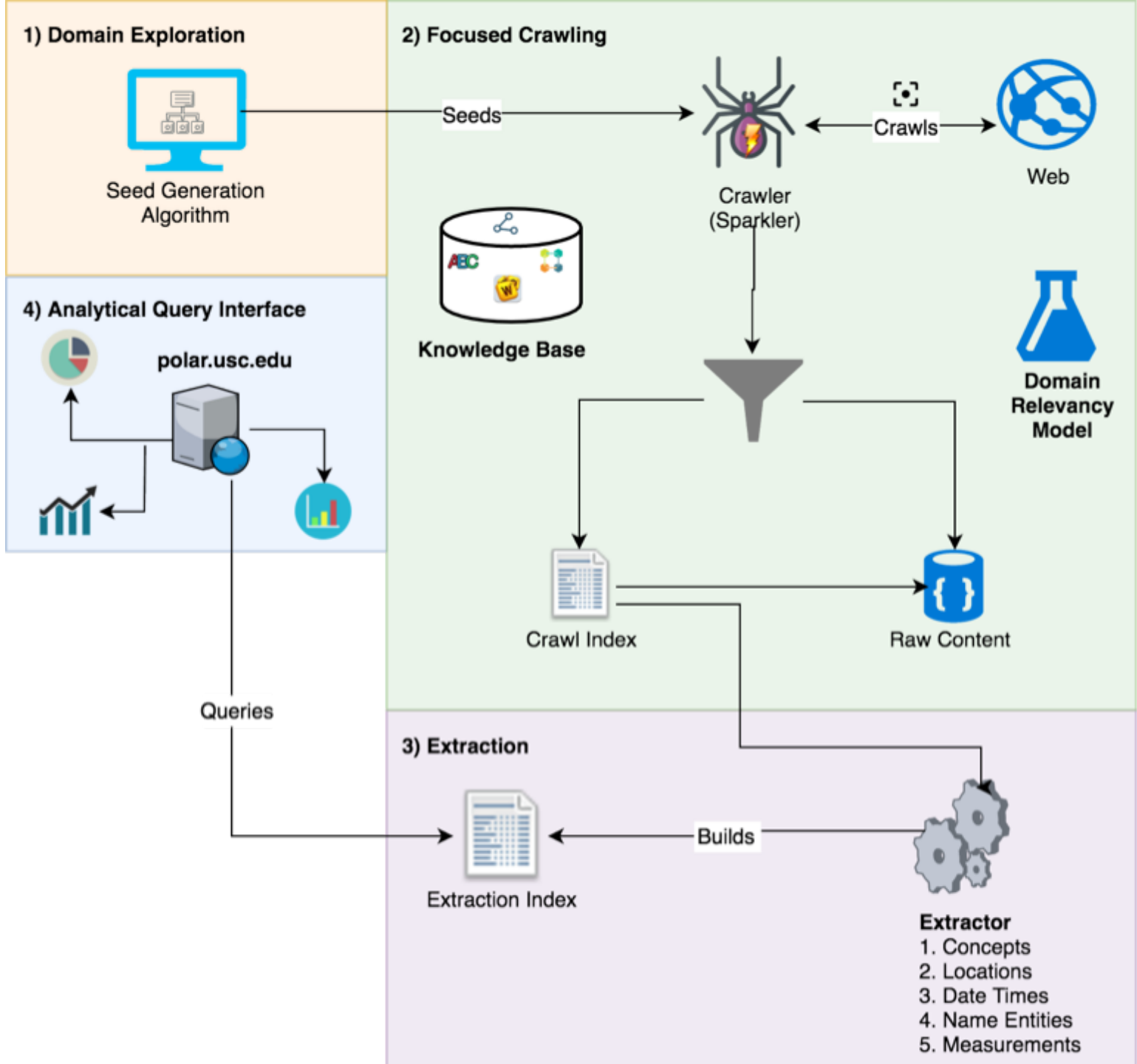
- A SVM-based machine learning algorithm to create a **Domain Relevance Model** for use in focused crawling
- An enhanced version of Apache Sparkler, an extensible, horizontally scalable **crawler** with high throughput capable of finding documents pertinent to the polar domain
- A **content enrichment** pipeline using Apache Tika
- A **user interface** using OS tools to display and interact with data.
  - Extracted named entities (Persons, Organizations, and Locations) are linked to relevant Polar topics and used to generate D3 based visualizations. FacetView connects to ElasticSearch and Solr, allowing faceted browsing, text-based search and display.

NSIDC

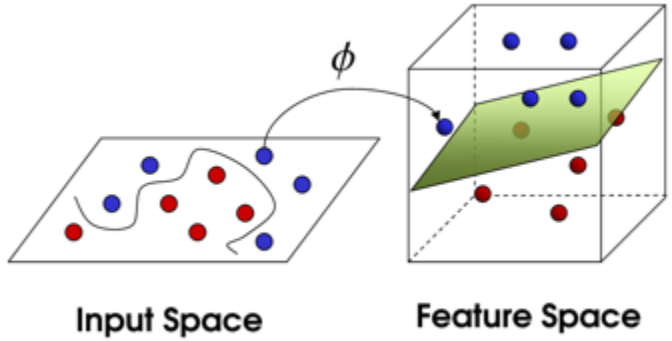
Polar Deep Insights Architecture

Leverages prior work done under the DARPA MEMEX, NSF Polar CyberInfrastructure activities, and community workshops

1) Domain experts provide URLs of web pages known to have relevant content; also domain vocabularies and ontologies
2) Crawler intelligently determines what part of the web it to crawl and what content to extract and index
3) Extractor goes through raw content pulling out named entities, spatial coordinates, measurements, units, dates and times, etc. and writes to index
4) Query Interface provides tools for exploring the extracted content. We are developing interfaces to respond to natural language queries



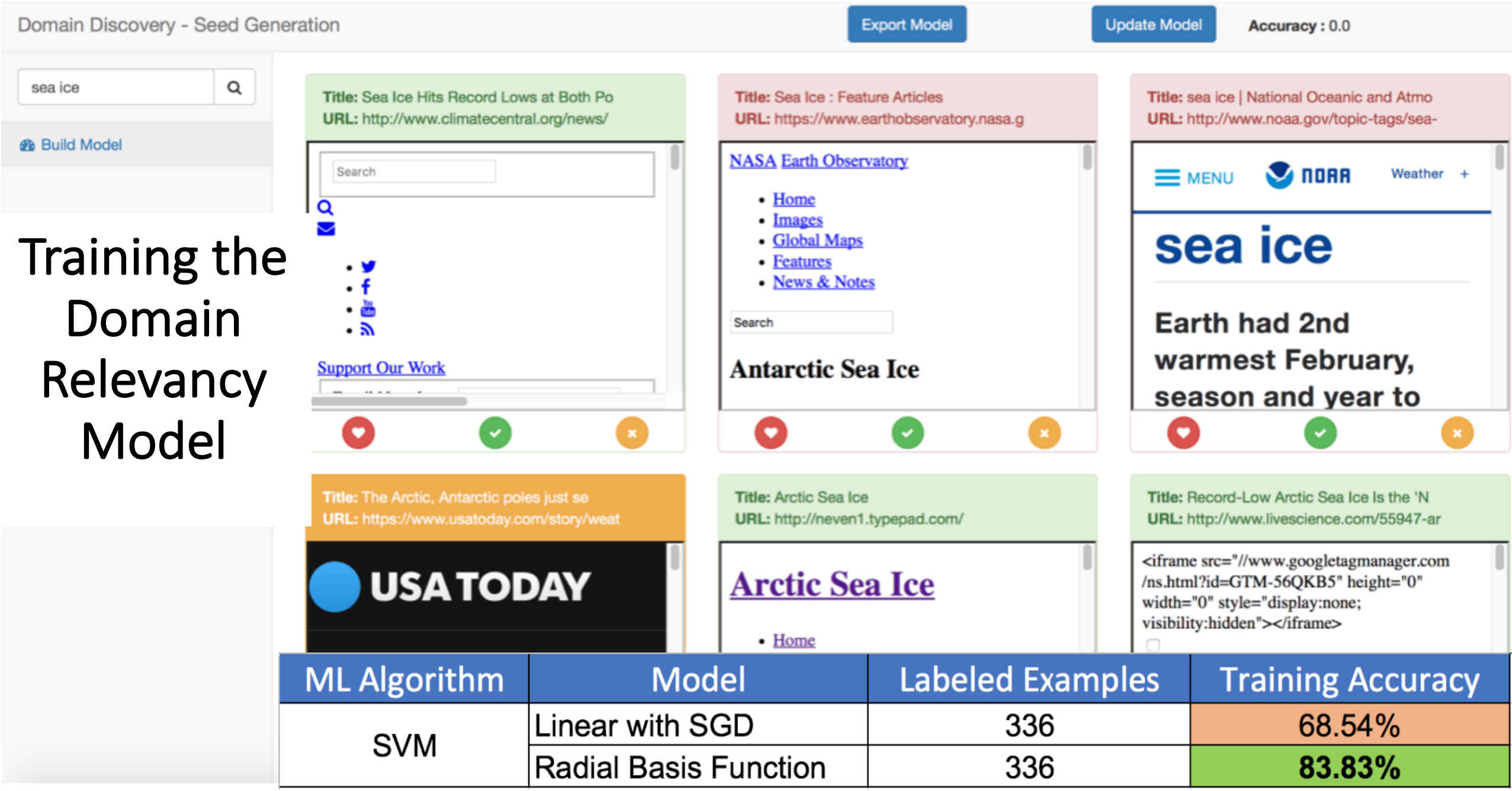
Domain Relevance Model



- Contextual and domain knowledge is supplied to the crawler by a machine learning model that predicts the relevance of a given document to the domain.
- We built a SVM (Support Vector Machine) classifier which works with domain ontologies and filtered text extracted from expert-selected web pages.
- Subject matter experts incrementally train the DRM by rating discovered documents.

NSIDC

Training the Domain Relevancy Model



ML Algorithm	Model	Labeled Examples	Training Accuracy
SVM	Linear with SGD	336	68.54%
	Radial Basis Function	336	**83.83%**

Finding Stuff in the World Wide (Wild) Web

Computers, and the Web changed the way we look for information

- But GSEs (general search engines) are not designed to look for data
  - Since they are “general” they are not tuned to the formats, methods, vocabulary of specific domains

NSIDC

Focused Crawling

- What do we look for?
  - Documents, datasets, images, video, data services specific to polar region
- How is this better than using Google?
  - GSEs lack domain knowledge and context; are agnostic to data content.
  - They lack basis for selectively crawling parts of a site that are specific to a particular domain of interest.
  - Are not specifically trained to interpret the contents of a document to predict whether it's of interest or not to a given domain

NSIDC

Web Crawling Basics

- A “crawler” is an agent or “bot” that systematically follows links (URLs)
  - Extract and index content for later searching
- Focused crawling means looking for specific content
- Requires “seed” URLs to get started
  - Also, supply keywords, glossaries or ontologies
- Crawlers consume resources on the systems they visit
  - Schedule, load, and "politeness" are important

NSIDC

Crawling Challenges

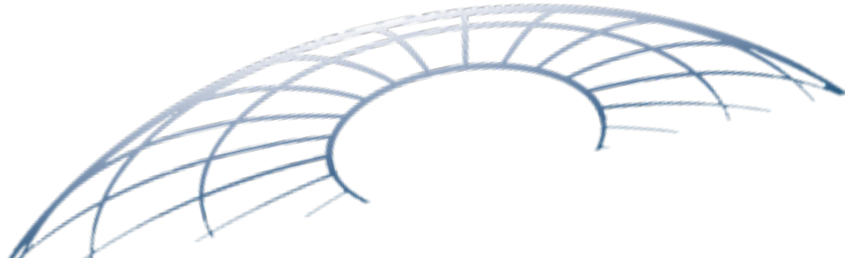
- Web pages may be ”dynamic”, presenting content in response to visitor actions
- Even focused crawling can return massive volumes of information
  - Further filtering and extraction of content necessary
- Need methods to work with the extracted content

Source: http://www.tfaforms.com/337137

Apache *Big Data* Technologies

- **Hadoop** - framework for distributed processing of massive unstructured data sets across computer clusters
  - Underpins an entire ecosystem of open-source applications for “Big Data”
- **Nutch** - an extensible, highly scalable Web crawler
- **Solr** - a means to index and query metadata through HTTP POST/GET requests
  - Provides faceted search, stemming, scoring, auto completion of query terms
- **Tika** - a content detection and analysis toolkit
  - To identify and extract information from various file types
- **Spark** - an alternative model for distributed computing
- **Sparkler** - a web crawler that runs on top of Spark
- **Kafka** – used to stream crawler output to a dashboard for real time monitoring

NSIDC



Extraction

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

NSIDC

Sparkler Dashboard

SEARCH QUERY

:

TOTAL DOCUMENTS HITS

7,462,201

INDEXED TIME TIMEPICKER

02/06/2017 15:00:00 to 02/08/2017 23:59:59

Relative | Absolute | Since

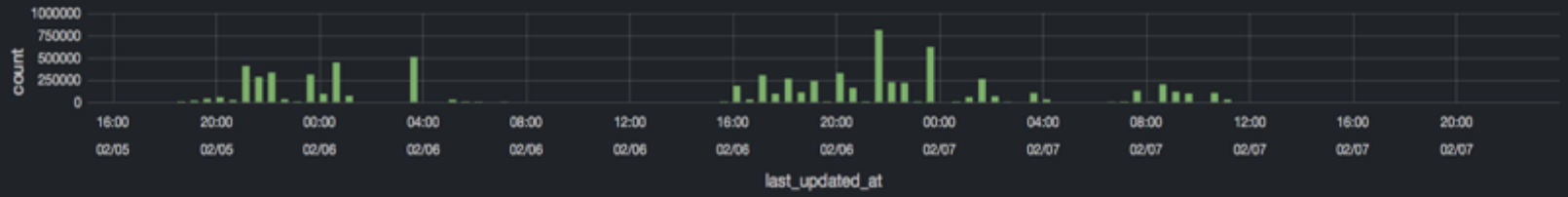
FACET SEARCH FACET

- crawl_id
- status
- hostname
- discover_depth

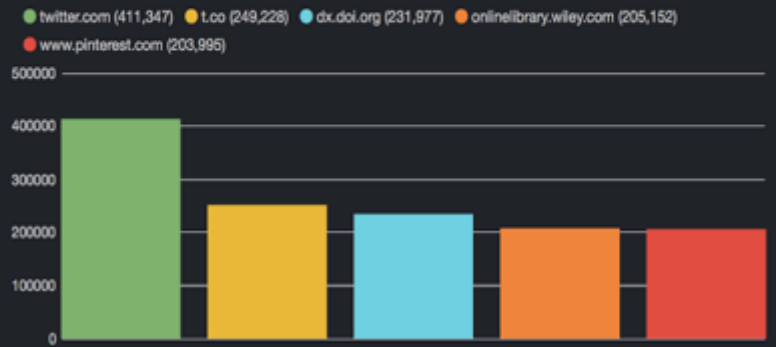
CRAWL TIME SERIES HISTOGRAM

View | Zoom Out | (7,462,201) count per 30m | (7,462,201 hits) | Time correction : browser

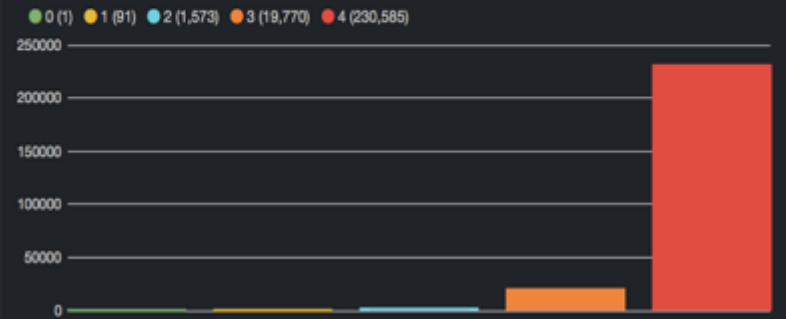
Bars Lines Points Stack Percent Interval 30m



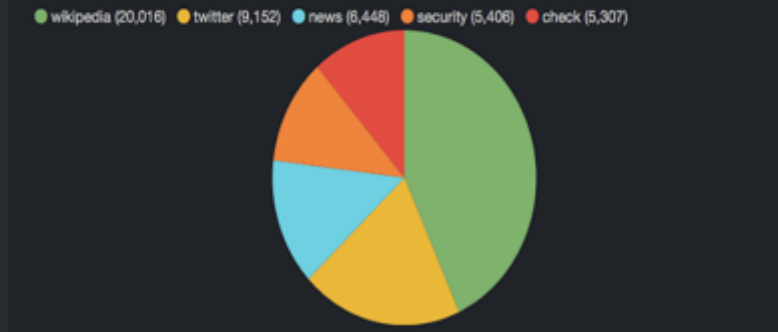
DOMAINS TERMS



DISCOVER DEPTH TERMS



TERMS TERMS



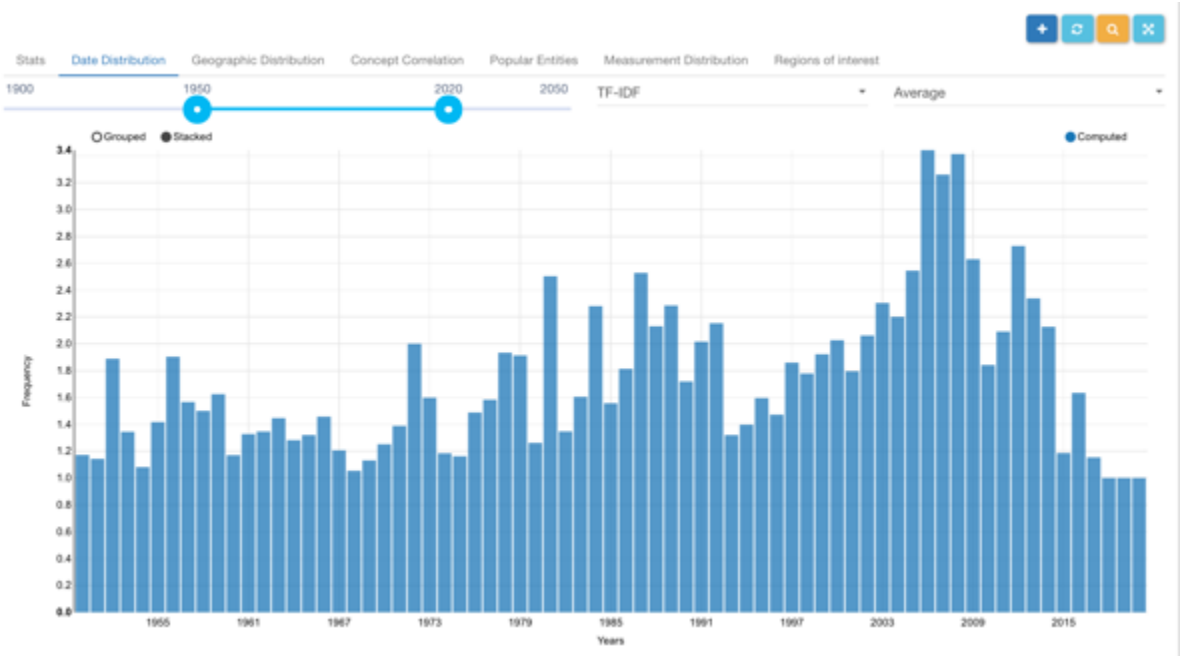
RESULTS TABLE

1 to 20 of 200 available for paging

hostname	url	title_t_md	fetch_timestamp
nsidc.org	http://nsidc.org/data/G01938/versions/1/print/	| National Snow and Ice Data Center	2017-02-06T22:06:15.936Z
nsidc.org	https://nsidc.org/data/nsidc-0304	| National Snow and Ice Data Center	2017-02-06T22:06:17.485Z
nsidc.org	https://nsidc.org/cryosphere/sotc/references.html#rignots/heets	SOTC: References | National Snow and Ice Data Center	2017-02-06T22:06:18.401Z
nsidc.org	http://nsidc.org/data/docs/daac/ae_l2a_tbs.gd.html#references	AMSR-E/Aqua L2A Global Swath Spatially-Resampled Brightness Temperatures	2017-02-06T22:06:18.727Z
nsidc.org	http://nsidc.org/the-drift/data-update/update-for-nasa-icebridge-atm-l1b-el...	Update for NASA IceBridge ATM L1B Elevation and Return Strength Data | The ...	2017-02-06T22:06:19.575Z
nsidc.org	http://nsidc.org/data/thermap/antarctic_10m_temps/traverses/notes/notes_dml...	THERMAP: Norwegian Traverse 1996-1997	2017-02-06T22:06:19.750Z

Query and Analysis

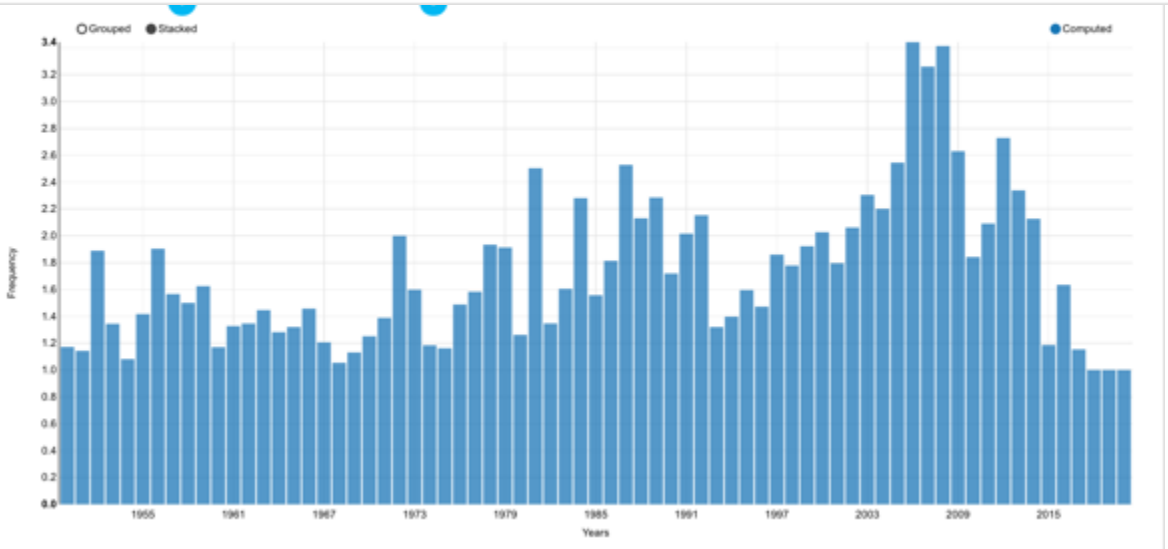
- This this vast store of information is of little use without an efficient and intuitive means of querying it
- Polar Data Insights is using various tools that an user can interact with through different dashboards to query and visualize the data



Query and Analysis

Stats | Date Distribution | Geographic Distribution | Concept Correlation | Popular Entities | Measurement Distribution | Regions of interest

- [illegible] intuitive means of querying it
- Polar Data Insights is using various tools that an user can interact with through different dashboards to query and visualize the data

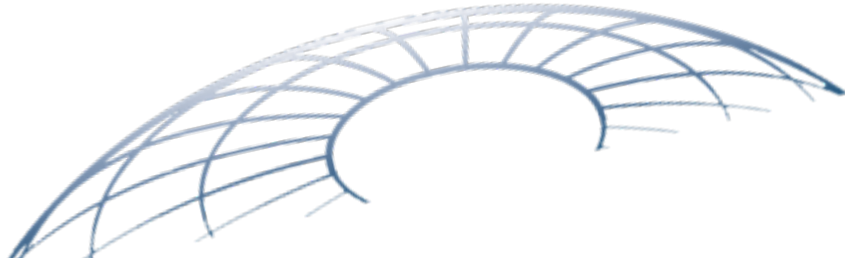


NSIDC

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 ensures responsive evolution of the tools
- These tools are providing the ability to address grand challenge questions concerning the state and trajectory of Polar regions

NSIDC



Project Personnel

- Chris Mattman (Principal Investigator)
- Siri Jodha Khalsa, Ruth Duerr (Domain Knowledge)
- Karanjeet Singh, Nithin Krishna Ottilingam (Developers)

Thanks for Your Attention

For More Information:

http://polar.usc.edu/

Polar Deep Insights Earth Cube:

http://bit.ly/2rEuWvn

National Snow and Ice Data Center

Advancing knowledge of Earth's frozen regions