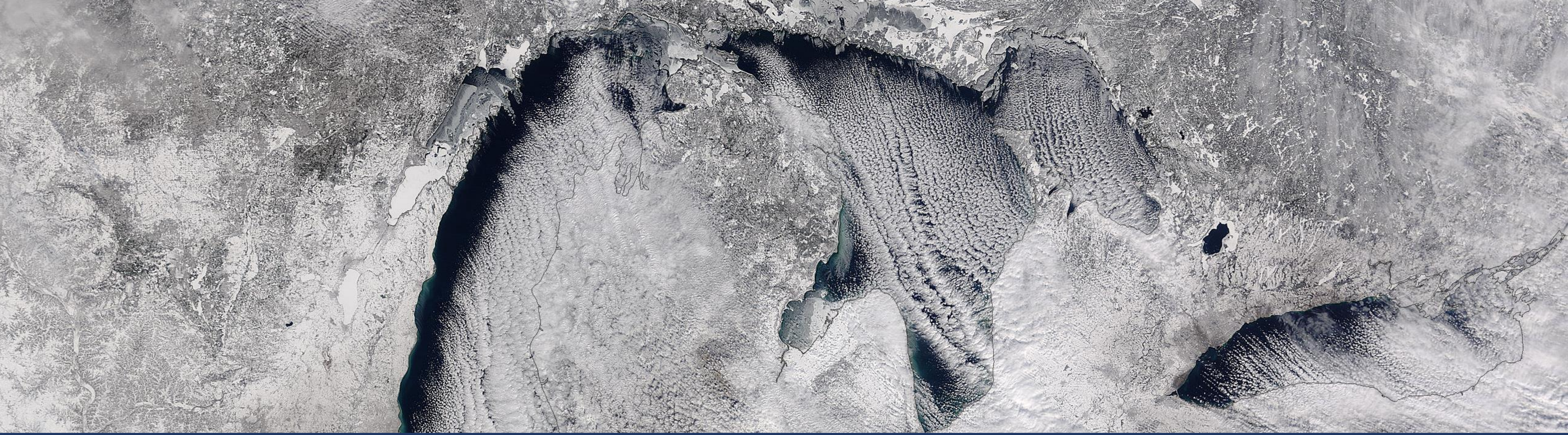


Developing a Knowledge Base for NASA Earth Science and Hydrologic Applications

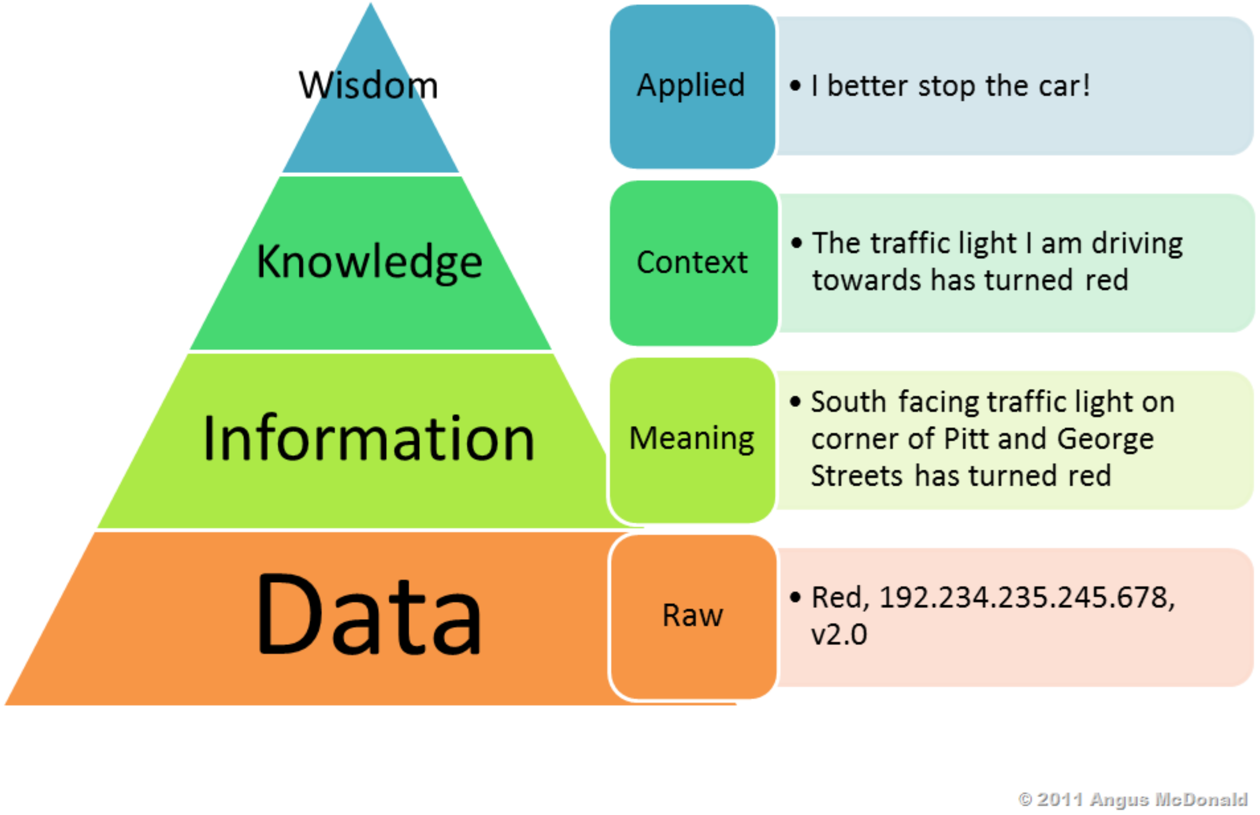
Amanda Weigel[1,2], Patrick Gatlin[1,3], Rahul Ramachandran[1,3], JJ Miller[1,2], Manil Maskey[1,3], Jia Zhang[1,3], Emily Berndt[3]

NASA/MSFC Data Science Informatics Group[1]
University of Alabama in Huntsville[2]
NASA Marshall Space Flight Center[3]
Carnegie Mellon University[4]

Data Use Challenges

What common challenges do Earth science and hydrologic data users encounter?

- Data **discovery**
- Data **use**
- Identifying **key resources** about the data.
  - Accessing introductory material (for unfamiliar users).
- Determining what **methods** to used – data processing, quality control and analysis.

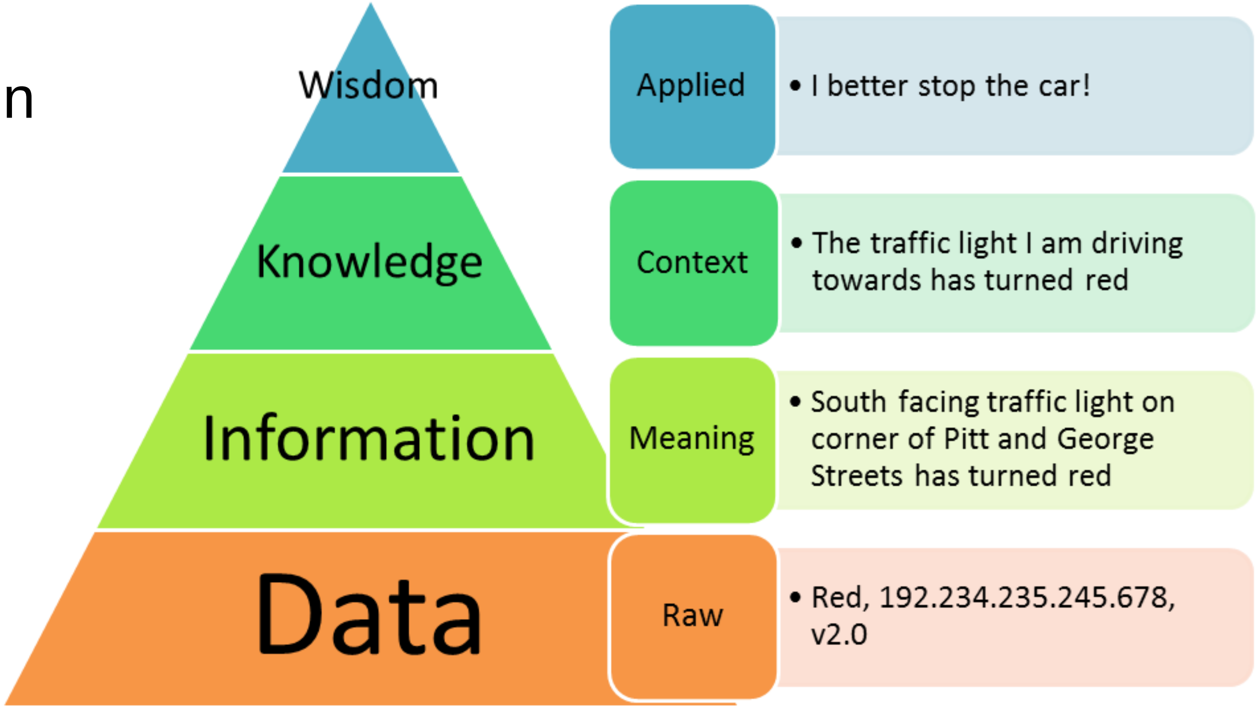


Data Use Challenges

To address these challenges, what difficulties are presented?

How can data and resources be linked in order to improve the data spin-up time?

How can we work to educate unfamiliar users?

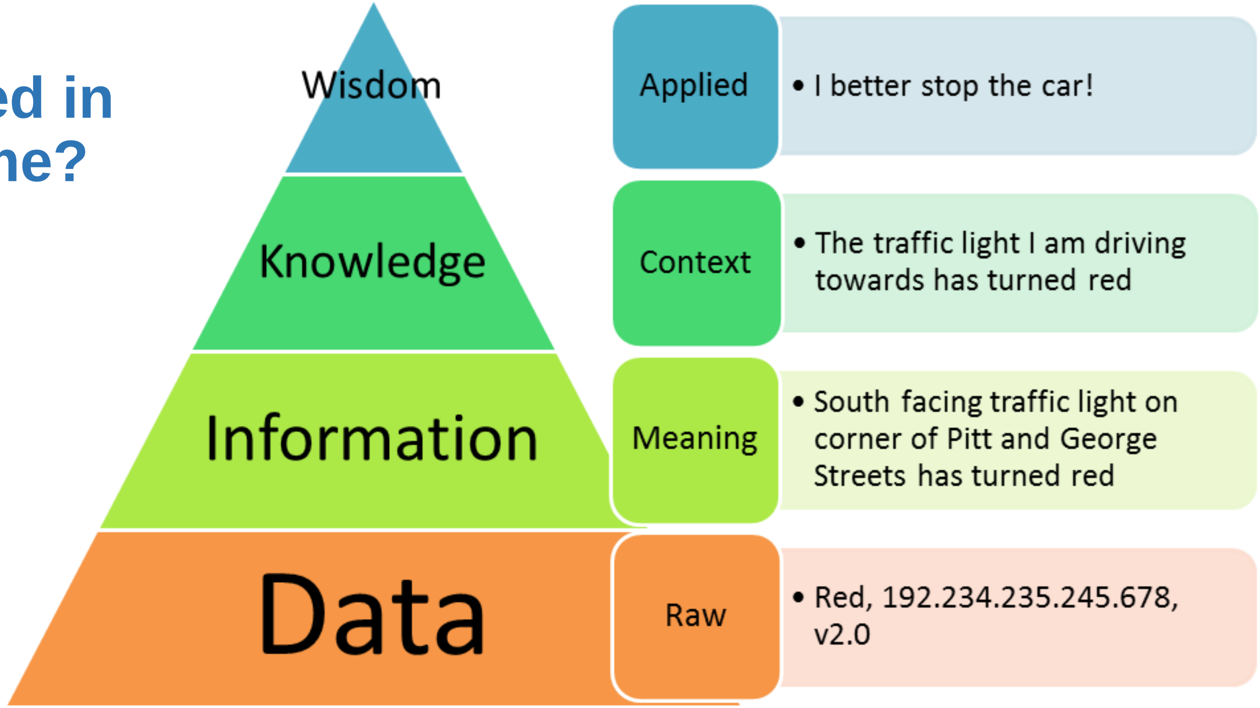


Data Use Challenges

To address these challenges, what difficulties are presented?

How can data and resources be linked in order to improve the data spin-up time?

How can we work to educate unfamiliar users?



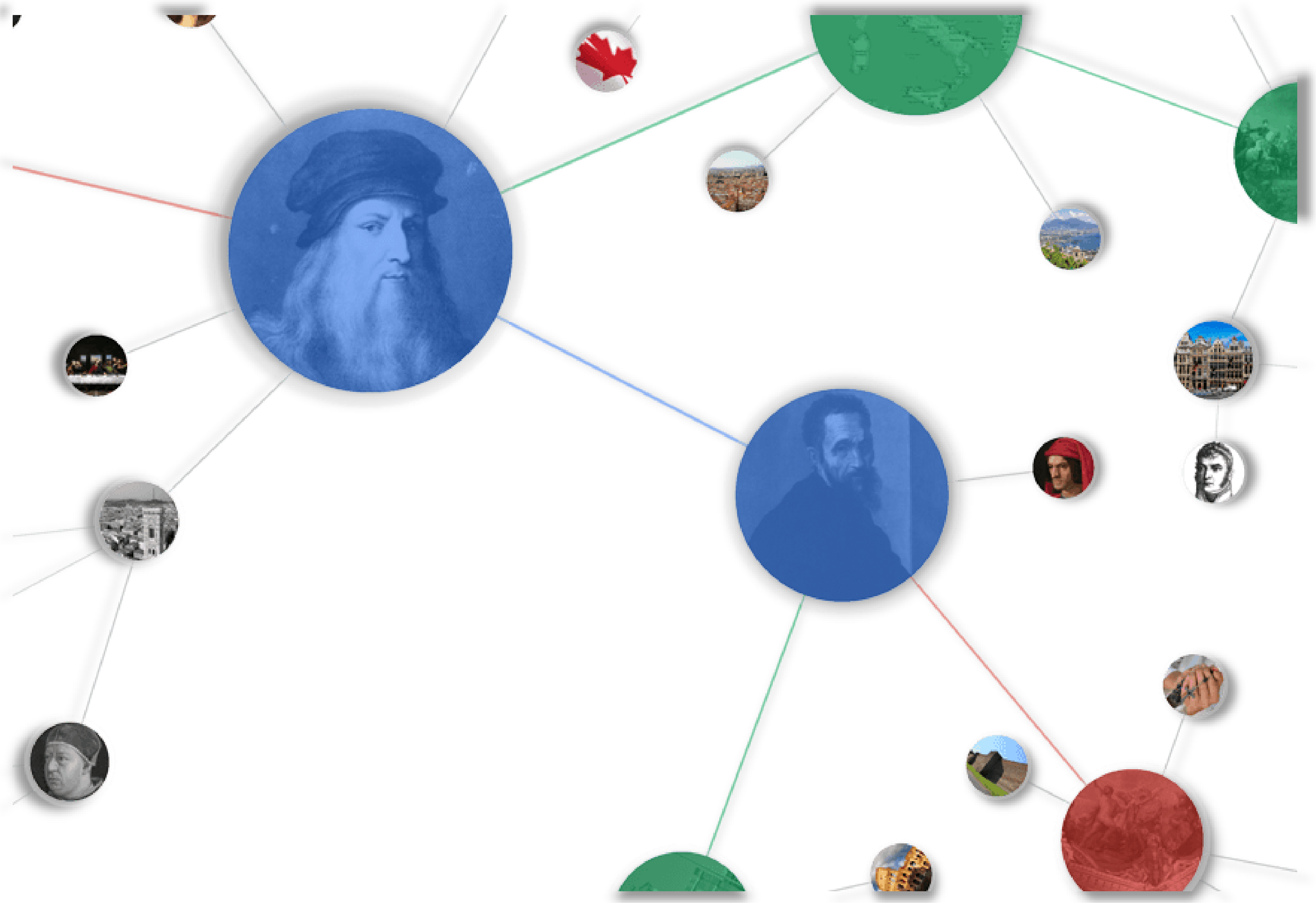
DSiG

What is a Knowledge Base?

- Think "**Google Search**".
- Developed by Google in 2012 to enhance the results of its search engine by systematically linking information.
- Aggregates structured and detailed information about a defined topic.
- Enables users to resolve their query without having to navigate and assemble information manually.
- Why not apply it to Earth science and hydrologic data and information?

Project Objectives

1. Identify key science information and develop an information model.
2. Extract key information from scientific literature (e.g. hypothesis, conclusions, methods, datasets, variables, etc.).
3. Link scientific knowledge to datasets, resources, services and scientists.
4. Develop a knowledge-based search capability for NASA Earth Science.



Google, The Knowledge Graph

Technical Approach

Terminology

What is an **entity**?

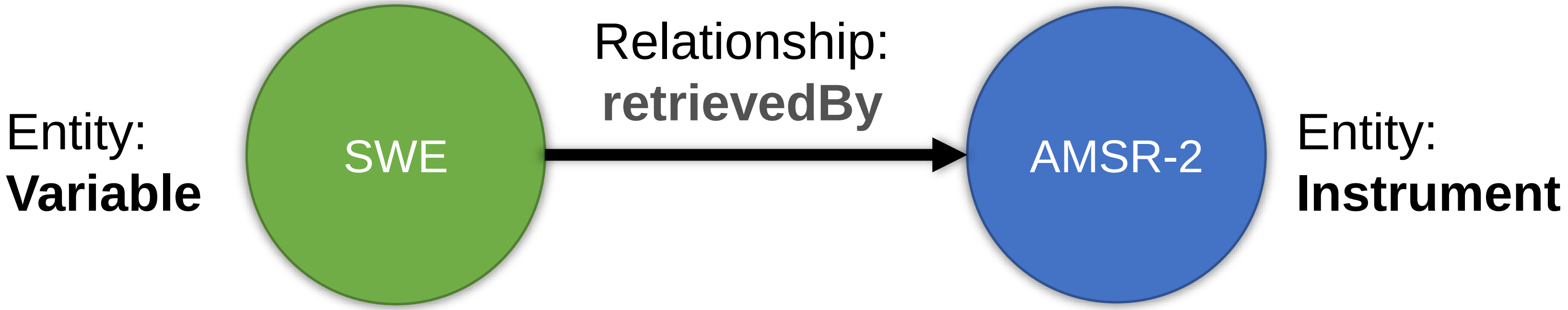
A thing with distinct and independent existence.

Examples: Variables, datasets, instruments, platforms etc.

What is a **relationship**?

The connection between two entities.

Example: “Snow water equivalent (SWE) is retrieved by AMSR-2”.

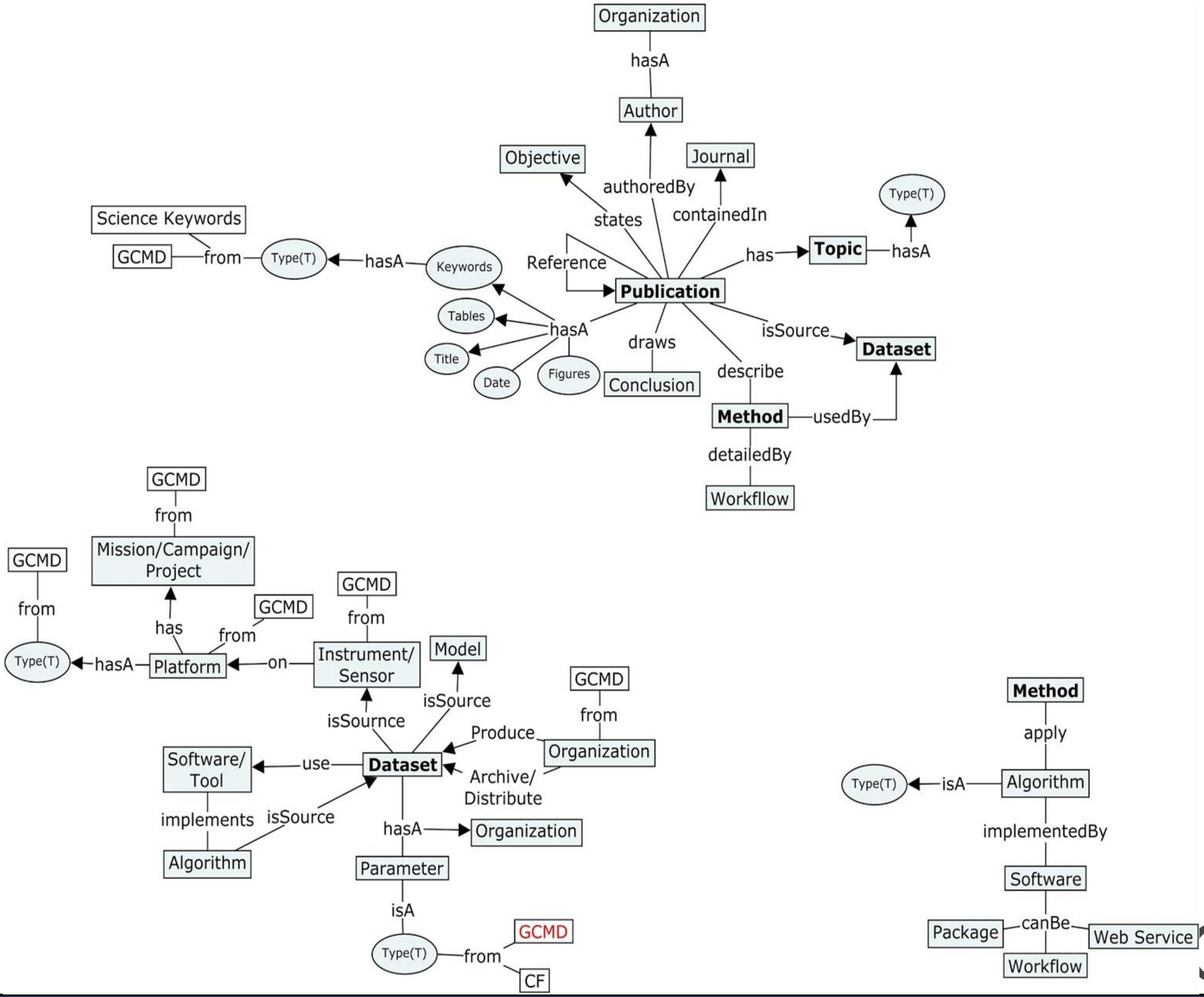


Technical Approach

Information Model

Development based on the Global Changing Information System (GCIS) information model

The information model defines entities and relationships pertinent to **NASA Earth science and hydrologic data, publications and resources**.



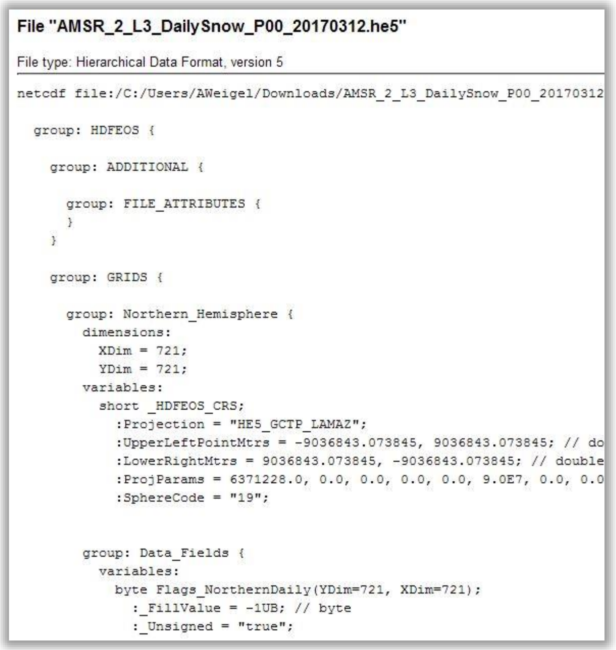
Technical Approach

Key Challenge

Knowledge base construction uses both structured and unstructured content (e.g., journal articles).

Structured Content

- Metadata, tables, controlled vocabularies



```
File "AMSR_2_L3_DailySnow_P00_20170312.he5"
File type: Hierarchical Data Format, version 5

netcdf file:/C:/Users/AWeigel/Downloads/AMSR_2_L3_DailySnow_P00_20170312
  group: HDFEOS {
    group: ADDITIONAL {
      group: FILE_ATTRIBUTES {
      }
    }
    group: GRIDS {
      group: Northern_Hemisphere {
        dimensions:
          XDim = 721;
          YDim = 721;
        variables:
          short _HDFEOS_CRS;
            :Projection = "HE5_GCTP_LAMAZ";
            :UpperLeftPointMtrs = -9036843.073845, 9036843.073845; // do
            :LowerRightMtrs = 9036843.073845, -9036843.073845; // double
            :ProjParams = 6371228.0, 0.0, 0.0, 0.0, 0.0, 9.0E7, 0.0, 0.0
            :SphereCode = "19";

        group: Data_Fields {
          variables:
            byte Flags_NorthernDaily(YDim=721, XDim=721);
              :_FillValue = -1UB; // byte
              :_Unsigned = "true";
```

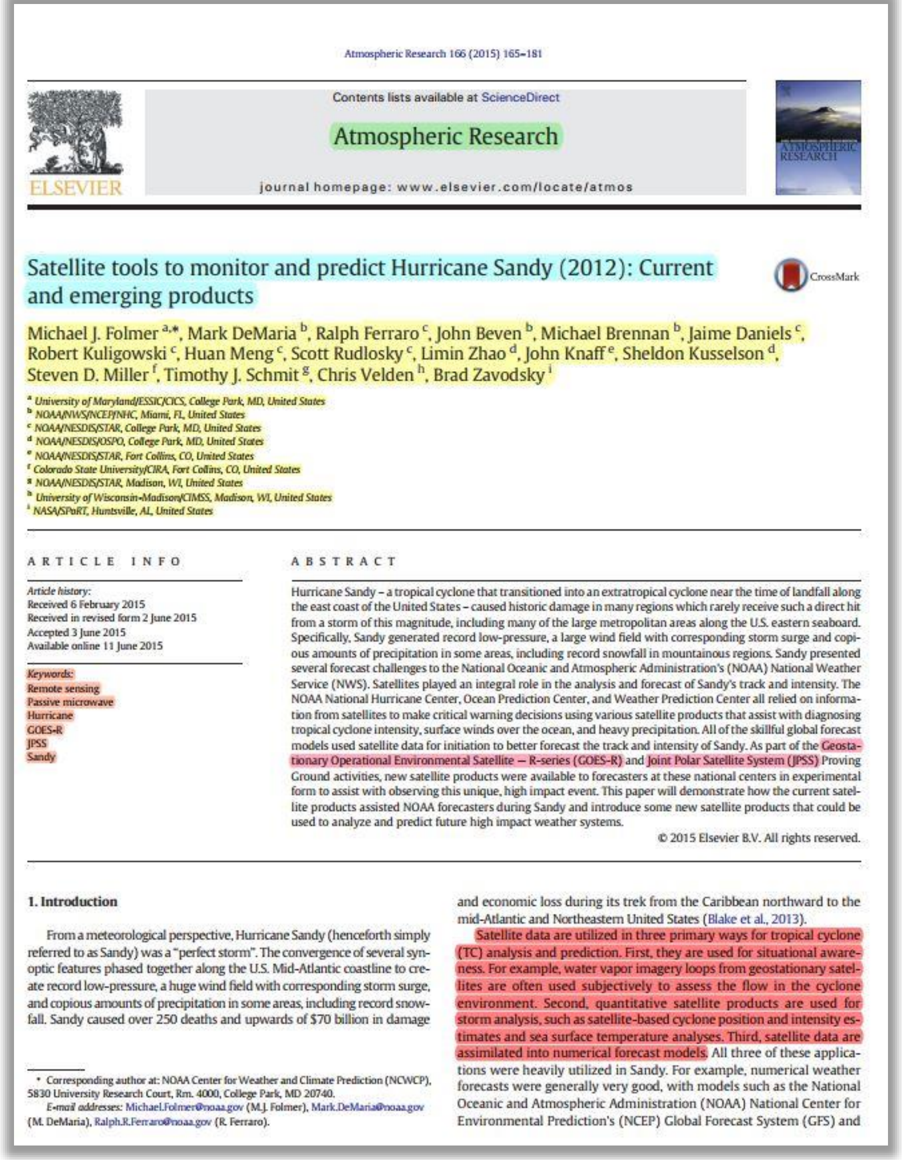
Unstructured Content

- Journal articles, ATB documents, user guides

Technical Approach

How do we extract entities from unstructured content?

- Natural language processing (**NLP**) – Computers analyzing, understanding and deriving meaning from human language.
- Semantic Entity Recognition (**SER**) – NLP technique used to identify entities in text.
- Use NLP (SER) techniques to identify entities within the unstructured text.
- Apply to journal publication text to extract and identify data, models, methods, people, and institutions (i.e., entities).
- Generate a truth set – Dictionary of known models, science keywords, CMR NASA Earth science data catalogue.



Atmospheric Research 166 (2015) 165–181

Contents lists available at ScienceDirect

Atmospheric Research

journal homepage: www.elsevier.com/locate/atmos

Satellite tools to monitor and predict Hurricane Sandy (2012): Current and emerging products

Michael J. Folmer [a,*], Mark DeMaria [b], Ralph Ferraro [c], John Beven [b], Michael Brennan [b], Jaime Daniels [c], Robert Kuligowski [c], Huan Meng [c], Scott Rudlosky [c], Limin Zhao [d], John Knaff [e], Sheldon Kusselson [d], Steven D. Miller [f], Timothy J. Schmit [g], Chris Velden [h], Brad Zavodsky [i]

[a] University of Maryland/ESSIC/CICS, College Park, MD, United States
[b] NOAA/NWS/NCEP/NHC, Miami, FL, United States
[c] NOAA/NESDIS/STAR, College Park, MD, United States
[d] NOAA/NESDIS/OSPO, College Park, MD, United States
[e] NOAA/NESDIS/STAR, Fort Collins, CO, United States
[f] Colorado State University/CIRA, Fort Collins, CO, United States
[g] NOAA/NESDIS/STAR, Madison, WI, United States
[h] University of Wisconsin-Madison/CIMSS, Madison, WI, United States
[i] NASA/SPoRT, Huntsville, AL, United States

ARTICLE INFO

Article history:
Received 6 February 2015
Received in revised form 2 June 2015
Accepted 3 June 2015
Available online 11 June 2015

Keywords:
Remote sensing
Passive microwave
Hurricane
GOES-R
JPSS
Sandy

ABSTRACT

Hurricane Sandy – a tropical cyclone that transitioned into an extratropical cyclone near the time of landfall along the east coast of the United States – caused historic damage in many regions which rarely receive such a direct hit from a storm of this magnitude, including many of the large metropolitan areas along the U.S. eastern seaboard. Specifically, Sandy generated record low-pressure, a large wind field with corresponding storm surge and copious amounts of precipitation in some areas, including record snowfall in mountainous regions. Sandy presented several forecast challenges to the National Oceanic and Atmospheric Administration's (NOAA) National Weather Service (NWS). Satellites played an integral role in the analysis and forecast of Sandy's track and intensity. The NOAA National Hurricane Center, Ocean Prediction Center, and Weather Prediction Center all relied on information from satellites to make critical warning decisions using various satellite products that assist with diagnosing tropical cyclone intensity, surface winds over the ocean, and heavy precipitation. All of the skillful global forecast models used satellite data for initiation to better forecast the track and intensity of Sandy. As part of the Geostationary Operational Environmental Satellite – R-series (GOES-R) and Joint Polar Satellite System (JPSS) Proving Ground activities, new satellite products were available to forecasters at these national centers in experimental form to assist with observing this unique, high impact event. This paper will demonstrate how the current satellite products assisted NOAA forecasters during Sandy and introduce some new satellite products that could be used to analyze and predict future high impact weather systems.

1. Introduction

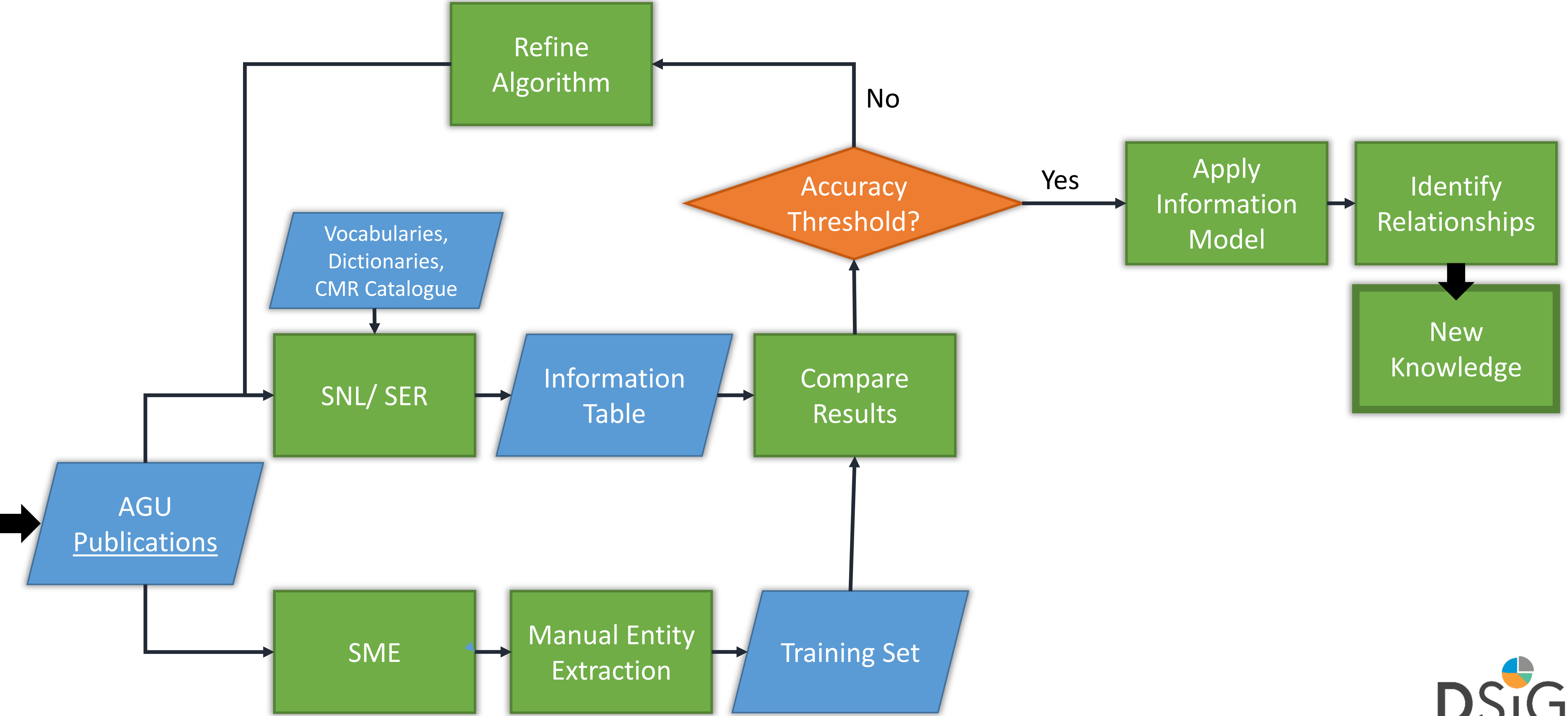
From a meteorological perspective, Hurricane Sandy (henceforth simply referred to as Sandy) was a "perfect storm". The convergence of several synoptic features phased together along the U.S. Mid-Atlantic coastline to create record low-pressure, a huge wind field with corresponding storm surge, and copious amounts of precipitation in some areas, including record snowfall. Sandy caused over 250 deaths and upwards of $70 billion in damage and economic loss during its trek from the Caribbean northward to the mid-Atlantic and Northeastern United States (Blake et al., 2013).

Satellite data are utilized in three primary ways for tropical cyclone (TC) analysis and prediction. First, they are used for situational awareness. For example, water vapor imagery loops from geostationary satellites are often used subjectively to assess the flow in the cyclone environment. Second, quantitative satellite products are used for storm analysis, such as satellite-based cyclone position and intensity estimates and sea surface temperature analyses. Third, satellite data are assimilated into numerical forecast models. All three of these applications were heavily utilized in Sandy. For example, numerical weather forecasts were generally very good, with models such as the National Oceanic and Atmospheric Administration (NOAA) National Center for Environmental Prediction's (NCEP) Global Forecast System (GFS) and

* Corresponding author at: NOAA Center for Weather and Climate Prediction (NCWCP), 5830 University Research Court, Rm. 4000, College Park, MD 20740.
E-mail addresses: Michael.Folmer@noaa.gov (M.J. Folmer), Mark.DeMaria@noaa.gov (M. DeMaria), Ralph.R.Ferraro@noaa.gov (R. Ferraro).

DSiG

Technical Approach



Preliminary Results

Example of extracted and populated information from unstructured sources

Carnegie Mellon University

Recommendation ▾ Resources ▾ Publication ▾ Analytics ▾ Collaboration ▾ Scientist ▾

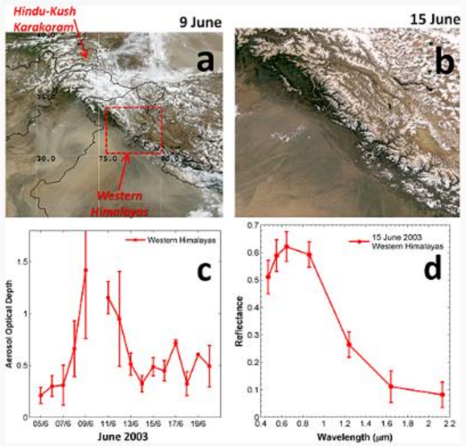
Satellite observations of desert dust-induced Himalayan snow darkening

MetaData | Figures and Captions | Analytics | Sections | Micro Article | References | Evaluation | Proofread

Paper Title	Satellite observations of desert dust-induced Himalayan snow darkening
Authors	Ritesh Gautam , Teppei J. Yasunari , Ritesh Gautam , N. Christina Hsu , William K.-M. Lau , Teppei J. Yasunari
Missing Acronym	BC (black carbon); HTP (Himalaya-Tibetan Plateau); WH (western Himalaya); WH (western Himalaya); SCF (snow cover fraction); IGP (Indo-Gangetic Plains); WH (western Himalaya); AOD (Aerosol Optical Depth); (NIR) (n); TOA (top-of-atmosphere); µm) ( ); RTM (radiative transfer model); SSA (single scattering albedo);
Id	55
Title	Satellite observations of desert dust-induced Himalayan snow darkening
Year	2013
Date	16-Mar 2013
Url	http://onlinelibrary.wiley.com/doi/10.1002/grl.50226/full
Document Id	55
Flag	0
Subject Category	Geology
Channel Name	GEOPHYSICAL RESEARCH LETTERS
Key Words	HYDROLOGICAL CYCLE; SPECTRAL ALBEDO; TIBETAN PLATEAU; MONSOON; AEROSOLS; PRODUCTS; GLACIERS; IMPACTS; CLIMATE; COVER
Author Keyword	Dust; Snow; Remote Sensing

Topic	climate(87.45%); dust(12.54%); aerosols(0.01%);
Topic Keywords	HYDROLOGICAL CYCLE; SPECTRAL ALBEDO; TIBETAN PLATEAU; MONSOON; AEROSOLS; PRODUCTS; GLACIERS; IMPACTS; CLIMATE; COVER
Instruments	MODIS(Moderate-Resolution Imaging Spectroradiometer)
Datasets	Transit (C1214558371-NOAA_NCEI) Terra 1 MODIS Imagery (C1214598068-SCIOPS) Aqua 1 MODIS Imagery (C1214598104-SCIOPS)
Variables	Toa cloud radiative effect, Solar zenith angle, Snow grain size, Normalized difference vegetation index, Surface snow area fraction, Tropopause instantaneous radiative forcing
Organizations	TOA, SCF, SSA, IGP, HTP
Sponsored Organizations/Projects	NASA

3 Extracted Figure/Table

Extracted Figure/Table	Caption
	(a) Satellite image of a major dust outbreak over south Asia, on 9 June 2003 from Terra/MODIS, indicating visibly dust-laden snow surface in the western Himalaya (WH); (b) zoom-in over WH on 15 June 2003; (c) daily AOD variations over the foothills, south of the WH snow cover; (d) MODIS spectral surface reflectance on 15 June indicating the VIS-NIR gradient for WH (30°N–34°N, 76°E–80°E), with error bars of ±1σ representing pixel-level variability.

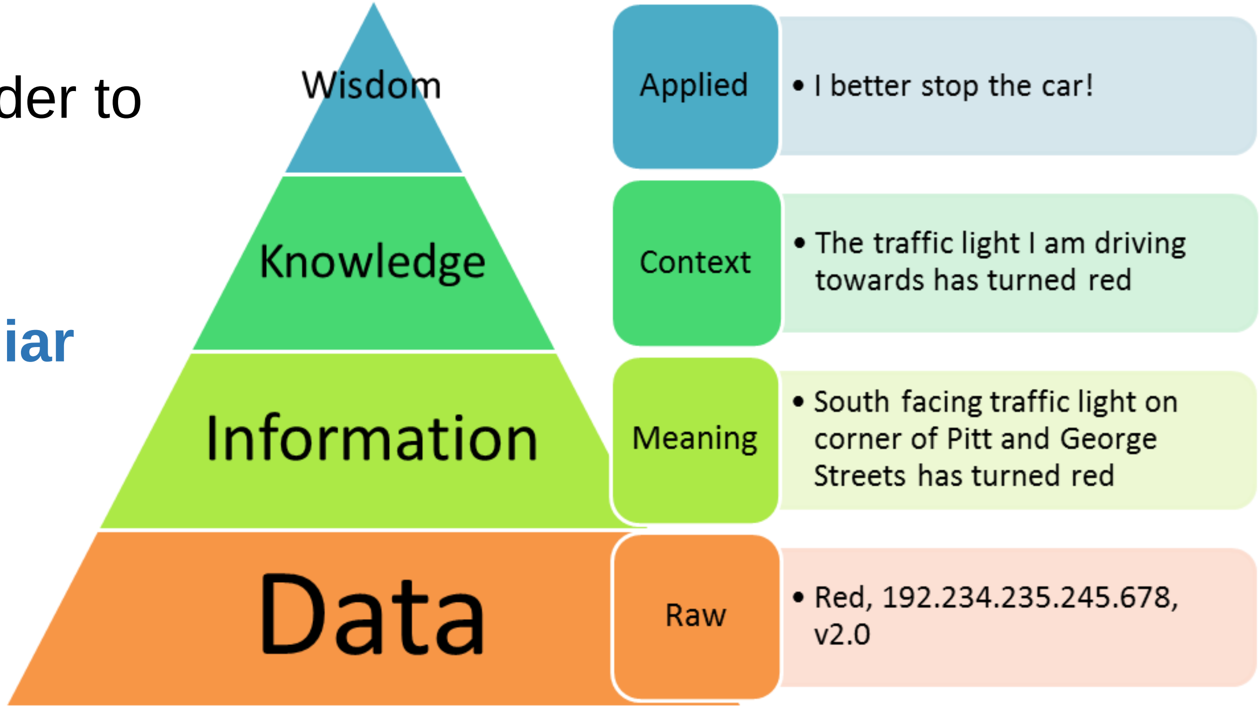
DSiG

Data Use Challenges

To address these challenges, what difficulties are presented?

How can data resources be linked in order to improve the data spin-up time?

How can we work to educate unfamiliar users?



DSiG

Other Resources

Challenge - Publications and technical documents often prove difficult for new and unfamiliar users to digest.

NASA Global Hydrology Resource Center (**GHRC**) Data Active Archive Center (**DAAC**)

NASA Short-term Prediction Research and Transition Center (**SPoRT**)

What resources are available to introduce data, methods and concepts?

- GHRC DAAC Data Recipes
- GHRC DAAC Micro Articles
- NASA SPoRT Quick Guides

NASA GHRC DAAC Data Recipes

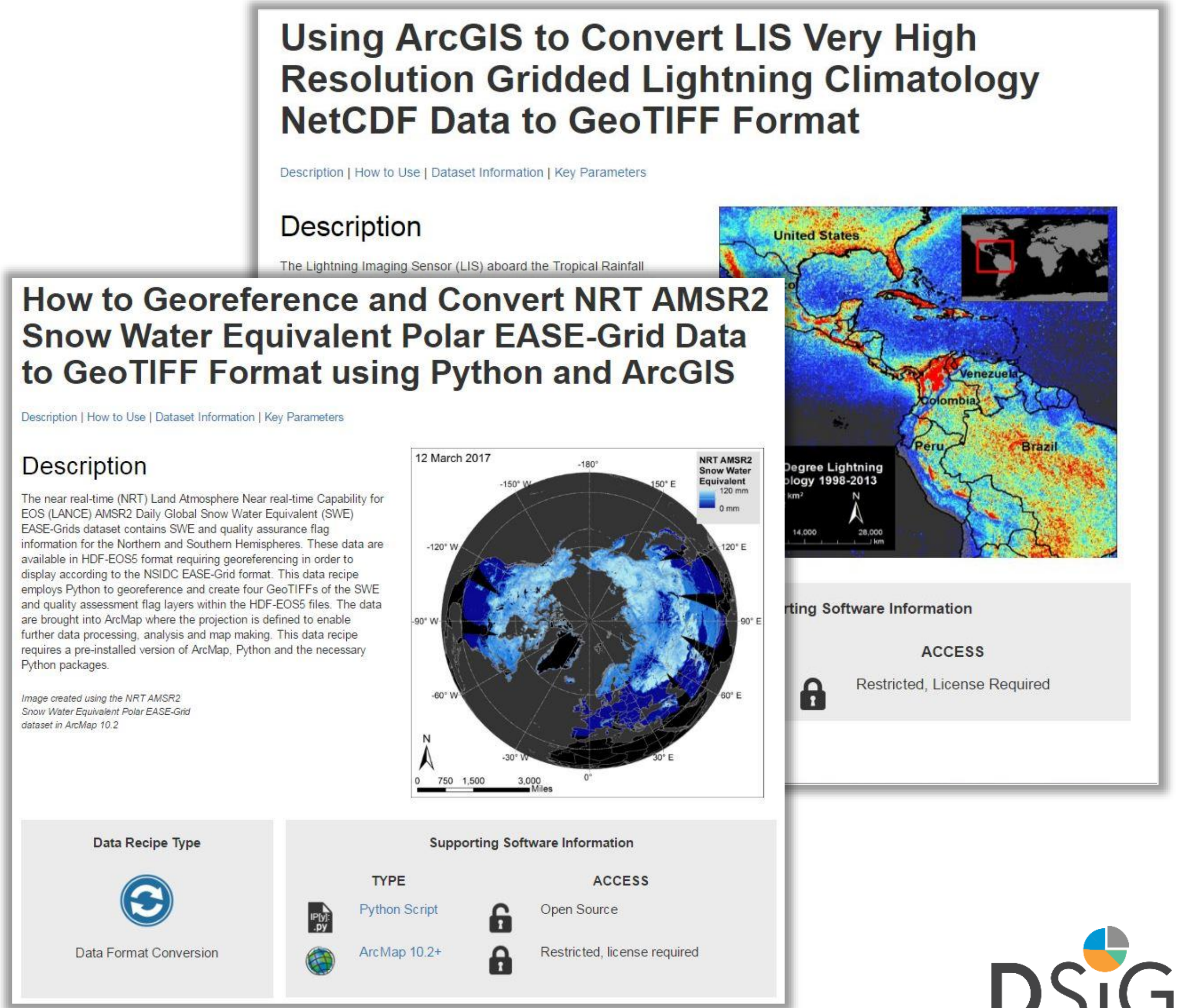
What is a Data Recipe?

Tutorials or step-by-step instructions to help users learn how to discover, visualize and use data, information, software and techniques.

Types of Data Recipes

- Using netCDF data in ArcGIS
- GHRC tool tutorials
- Python notebooks and scripts
- Data format conversions and georeferencing

[Discover GHRC Data Recipes Here]



NASA GHRC DAAC Micro Articles

What is a Micro Article?

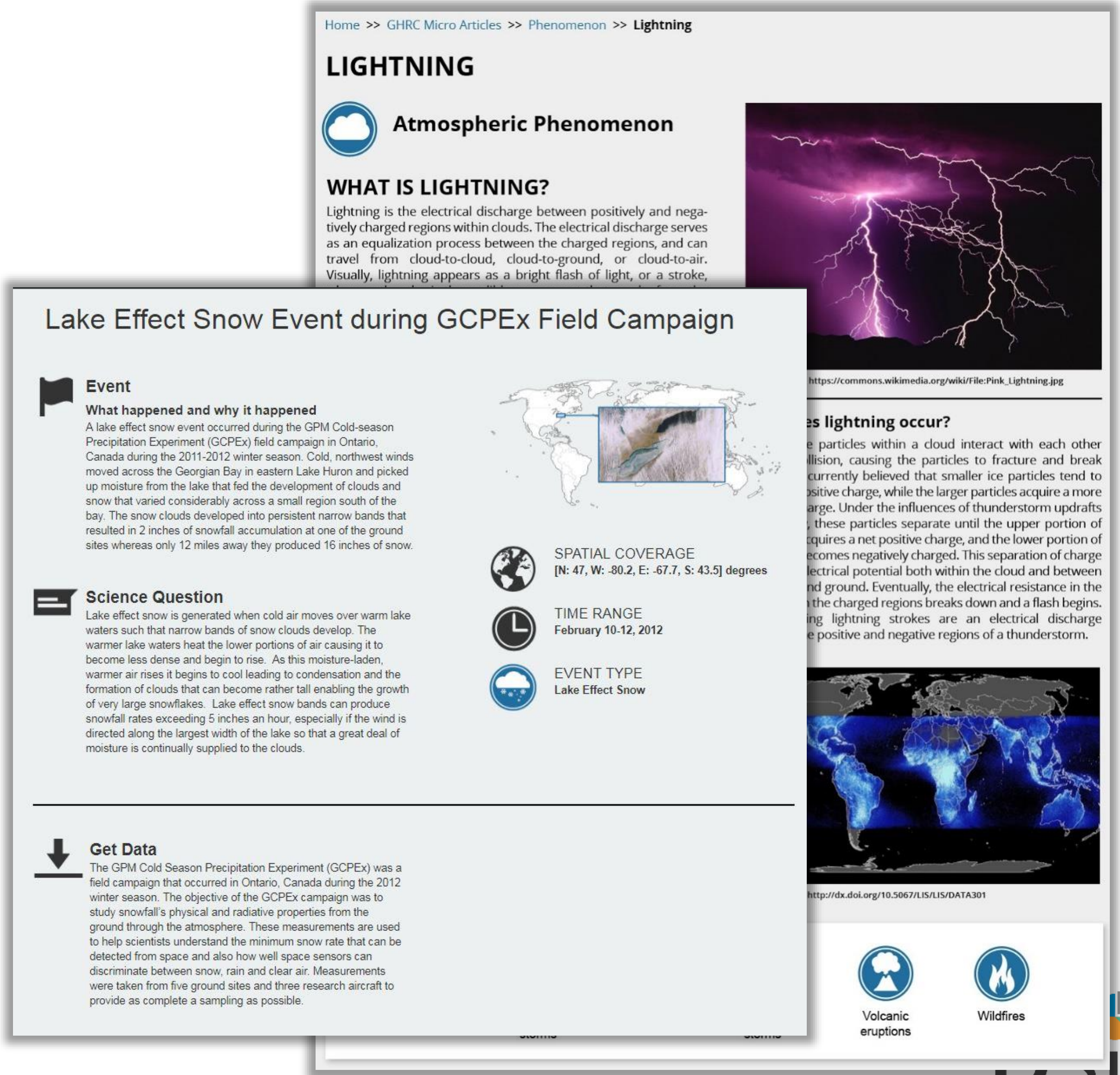
A short, interesting document that brings together data and key science concepts

Creates a knowledge base for users by curating around GHRC's data and science thematic areas

Types of Micro Articles

- Instruments
- Phenomena
- Events or Case Studies
- Publications

[Discover Micro Articles Here]



NASA SPoRT Quick Guides

What is a Quick Guide?

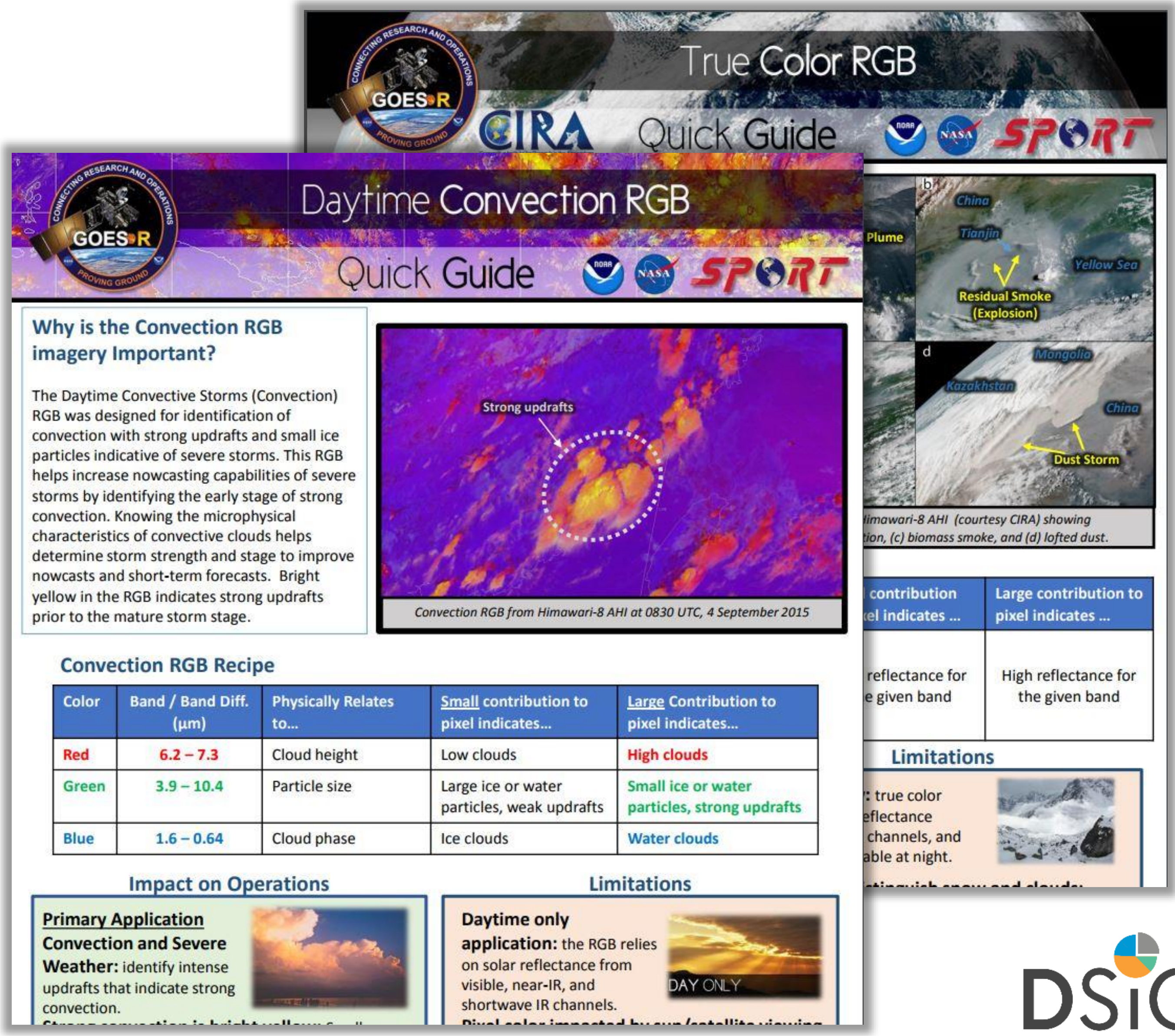
Short, easy to use resources that highlight key aspects of a data product or tool.

Intended to assists forecasters in quickly recalling information during times of operation.

Available Forms

- Download/print
- Interactive web browser
- Interactive through personal display system

[Discover SPoRT Quick Guides here]()


True Color RGB

Quick Guide

Daytime Convection RGB

Quick Guide

Why is the Convection RGB imagery Important?

The Daytime Convective Storms (Convection) RGB was designed for identification of convection with strong updrafts and small ice particles indicative of severe storms. This RGB helps increase nowcasting capabilities of severe storms by identifying the early stage of strong convection. Knowing the microphysical characteristics of convective clouds helps determine storm strength and stage to improve nowcasts and short-term forecasts. Bright yellow in the RGB indicates strong updrafts prior to the mature storm stage.

Convection RGB from Himawari-8 AHI at 0830 UTC, 4 September 2015

Convection RGB Recipe

Color	Band / Band Diff. (µm)	Physically Relates to...	Small contribution to pixel indicates...	Large Contribution to pixel indicates...
Red	6.2 – 7.3	Cloud height	Low clouds	High clouds
Green	3.9 – 10.4	Particle size	Large ice or water particles, weak updrafts	Small ice or water particles, strong updrafts
Blue	1.6 – 0.64	Cloud phase	Ice clouds	Water clouds

Impact on Operations

Primary Application
Convection and Severe Weather: identify intense updrafts that indicate strong convection.

Limitations

Daytime only application: the RGB relies on solar reflectance from visible, near-IR, and shortwave IR channels.

...imawari-8 AHI (courtesy CIRA) showing ...tion, (c) biomass smoke, and (d) lofted dust.

...contribution ...el indicates ...	Large contribution to pixel indicates ...
...reflectance for ...e given band	High reflectance for the given band

Limitations

...: true color ...eflectance ...channels, and ...able at night.

Other Resources

How do these benefit Earth science and hydrologic applications?

- These resources provide introductory information that is easy to read and understand without overwhelming users.
- Each point to additional documents for more detailed information.
- Each contain information on commonly used data, models, and software.
- They link directly to data, helping users understand a dataset and how to apply it towards research or applications.
- Populating this information within a knowledge graph allows users to search and discover information on data and methods for a broad user community.

DSIG

Next Steps

- Investigate generating easy to understand resources in a structured format to allow more seamless integration within the knowledge graph.
- Continue refining SER for Earth Science
- Continue building and evaluate a training set for SER (working with graduate students and SMEs)
- Scale efforts to to all Earth Science related journal titles in the Wiley Online Library
- Begin mining graphs to obtain new information
  - Prediction of relationship between entities (i.e., Network Link Prediction)
  - Automatic generation of new content (e.g., MicroArticles)

DSiG

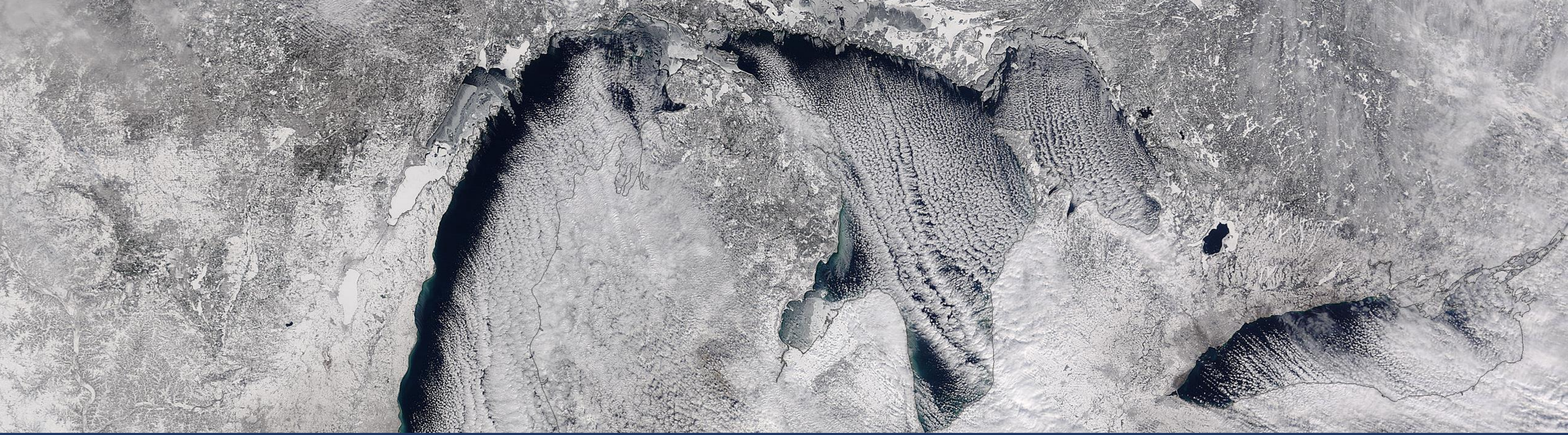
Benefits to NASA Earth Science and Hydrology

- Addresses the challenge in navigating the increasing volume of data and information.
- To provide an operational knowledge base to enhance NASA's Earth science research.

Beneficial Applications

- Hypothesis formulation and testing:
  - Automate the search for and compilation of background information.
  - Given a topic, what hypotheses have been tested?
  - What data/tools are being used to test a hypothesis?
  - Common paths to knowledge discovery.
- Mission development/review:
  - What kinds of instruments/parameters are needed to specify science objectives?
  - Impact of a mission by linking it with publications and dataset distribution.

DSIG



Thank you, questions?

amanda.m.weigel@nasa.gov

NASA/MSFC Data Science Informatics Group

DSiG Carnegie Mellon University THE UNIVERSITY OF ALABAMA IN HUNTSVILLE NASA