

Microsoft® Research

Faculty Summit 2010

Environmental Data Management

William Michener

Professor and Director of e-Science Initiatives

University of New Mexico, University Libraries System

Roadmap

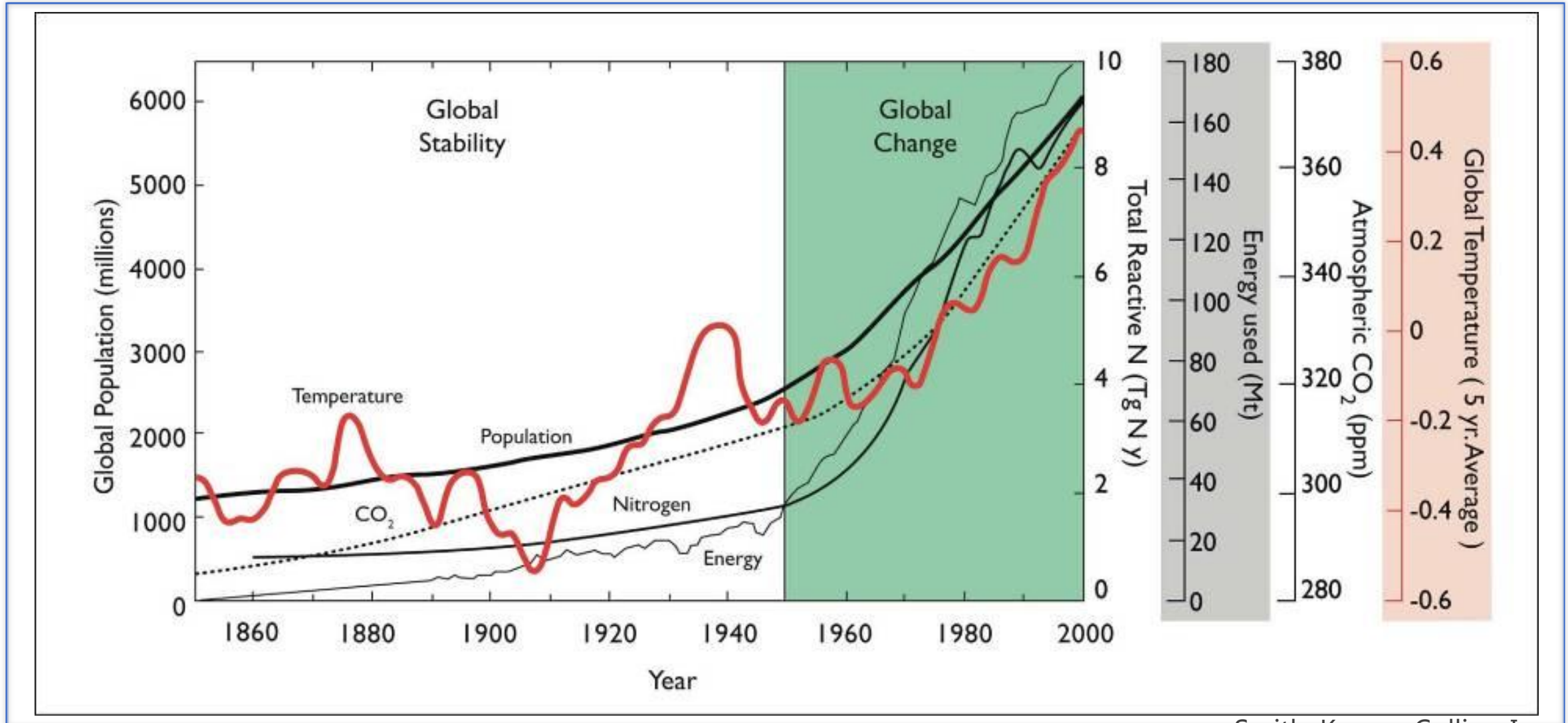
- Environmental Science Challenges
- Data Management Challenges
- DataONE: Part of the Solution
- Environmental Science 2020 – 3 case studies

Environmental Science Challenges

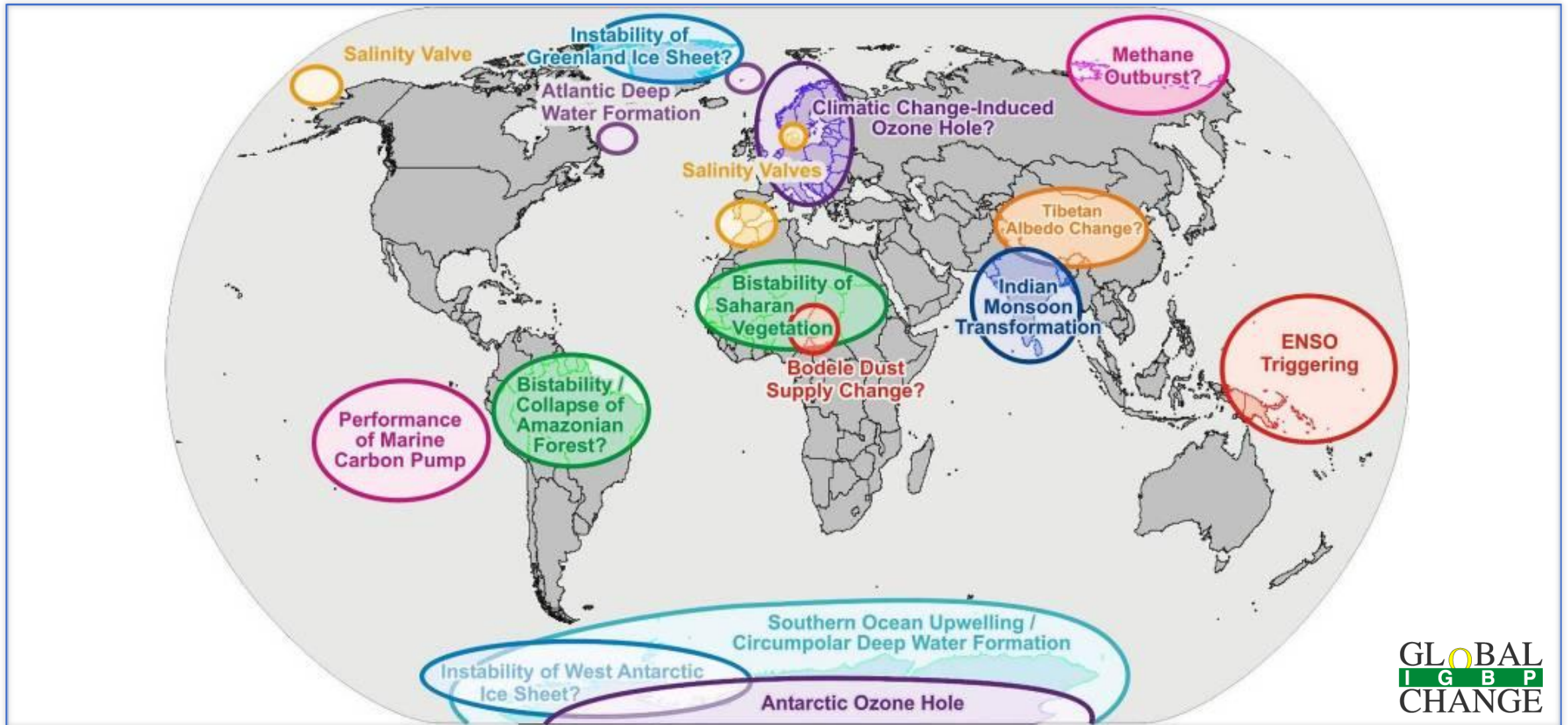
- Climate Change
- Biodiversity Loss
- Invasive Species
- Water Depletion
- Disease Spread
- Green Energy
- Habitat Loss
- ---



Environmental Science Challenges



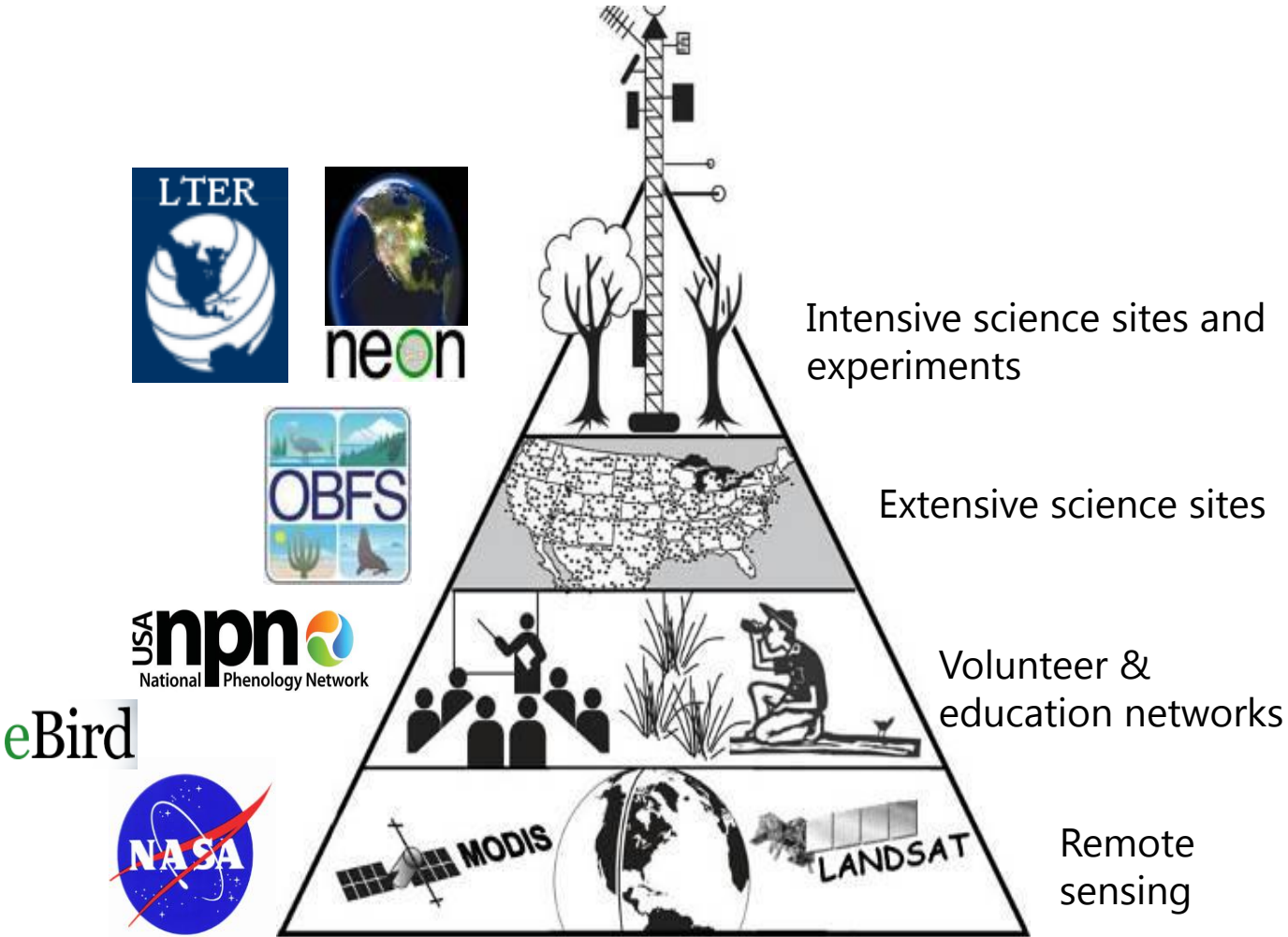
Environmental Science Challenges



Environmental Science Challenges

“Building the Knowledge Pyramid”

Decreasing Spatial Coverage
Increasing Process Knowledge



Adapted from CENR-OSTP

Data Management Challenges

1. Data Entropy
2. Data Discovery
3. Data Heterogeneity
4. Data Interpretation



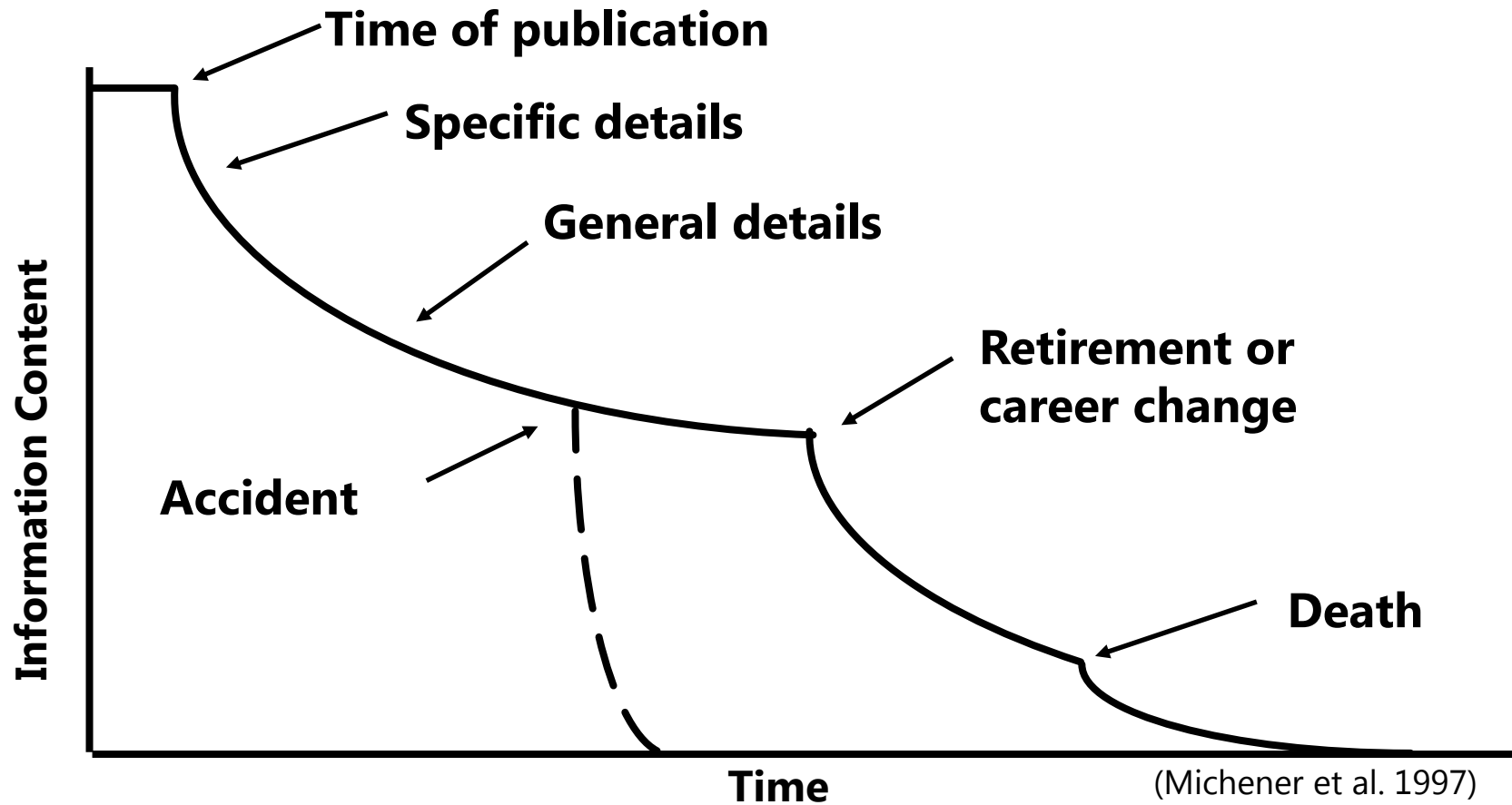
The
F O U R T H
P A R A D I G M

DATA-INTENSIVE SCIENTIFIC DISCOVERY

EDITED BY TONY HEY, STEWART TANSLEY, AND KRISTIN TOLLE

Data Management Challenges

1. Data Entropy



Data Management Challenges

2. Data Discovery



10s – 100s
Repositories



1,000s
Institutions



10,000s
Scientists



100,000s
Citizen
Scientists

Data Management Challenges

3. Data Heterogeneity

- Syntax
 - (format)
- Schema
 - (model)
- Semantics
 - (meaning)

Study A

METADATA (from EML)	
Study A:	White Mountains
Area col. units:	sq. meter
PIRU =	<i>Picea rubens</i>
BEPA =	<i>Betula papyifera</i>

date	site	species	area	count
10/1/1993	N654	PIRU	2	26
10/3/1994	N654	PIRU	2	29
10/1/1993	N654	BEPA	1	3

Study B

METADATA (from EML)	
Study B:	Green Mountains
Area sampled:	1 sq. meter
picrub =	<i>Picea rubens</i>
betpap =	<i>Betula papyifera</i>

date	site	picrub	betpap
31 Oct 1993	1	13.5	1.6
14 Nov 1994	1	8.4	1.8

Integrated Data

study	date	site	species	density
A	10/1/1993	N654	<i>Picea Rubens</i>	13.0
A	10/3/1994	N654	<i>Picea Rubens</i>	14.5
A	10/1/1993	N654	<i>Betula papyifera</i>	3.0
B	10/31/1993	1	<i>Picea Rubens</i>	13.5
B	10/31/1993	1	<i>Betula papyifera</i>	1.6
B	11/14/1994	1	<i>Picea Rubens</i>	8.4
B	11/14/1994	1	<i>Betula papyifera</i>	1.8

↑
metadata
'promoted'
to become
data

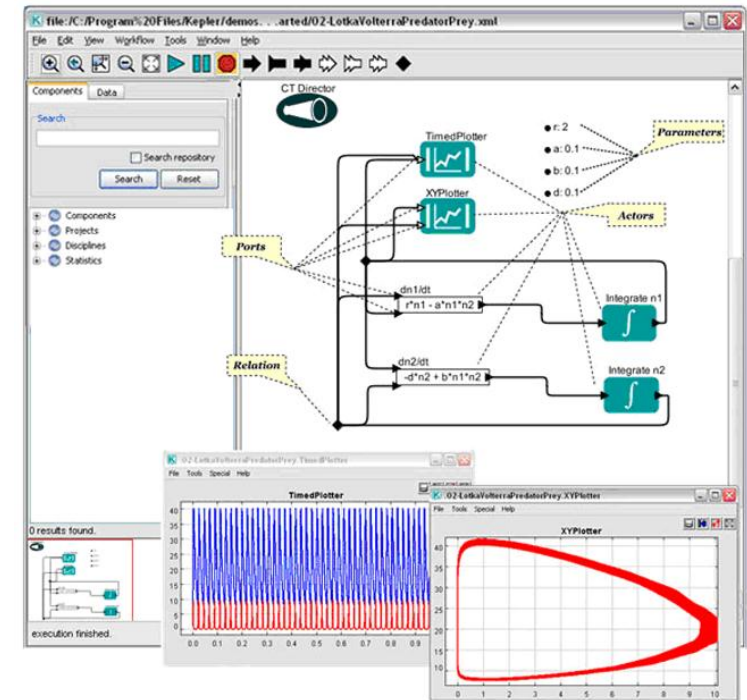
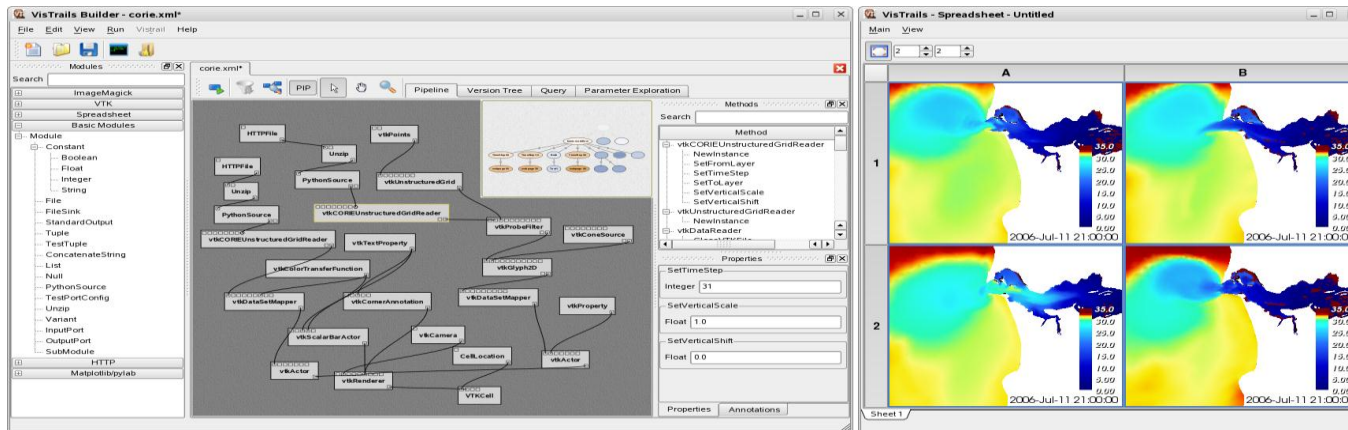
↑
format
normalized
using
metadata

↑
species metadata
from study B
is now data
(picrub/betpap
column headings)

↑
density
calculated
using
metadata

Data Management Challenges

4. Data Interpretation

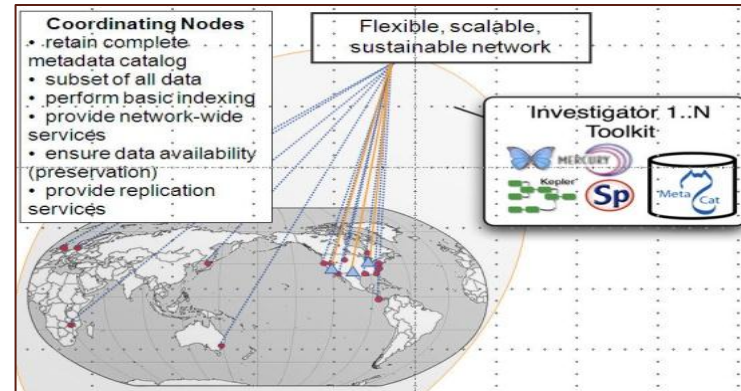


DataONE: Part of the Solution

Providing *universal access to data about life on earth and the environment that sustains it*

- engaging the scientist in the data curation process
- supporting the full data life cycle
- encouraging data stewardship and sharing
- promoting best practices
- engaging citizens
- developing domain-agnostic solutions

1. Build on existing cyberinfrastructure



2. Create new cyberinfrastructure



3. Support new communities of practice



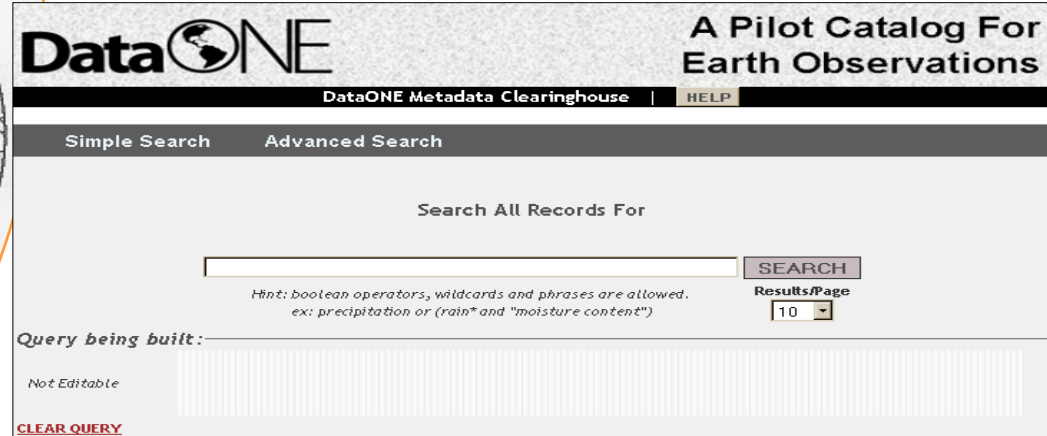
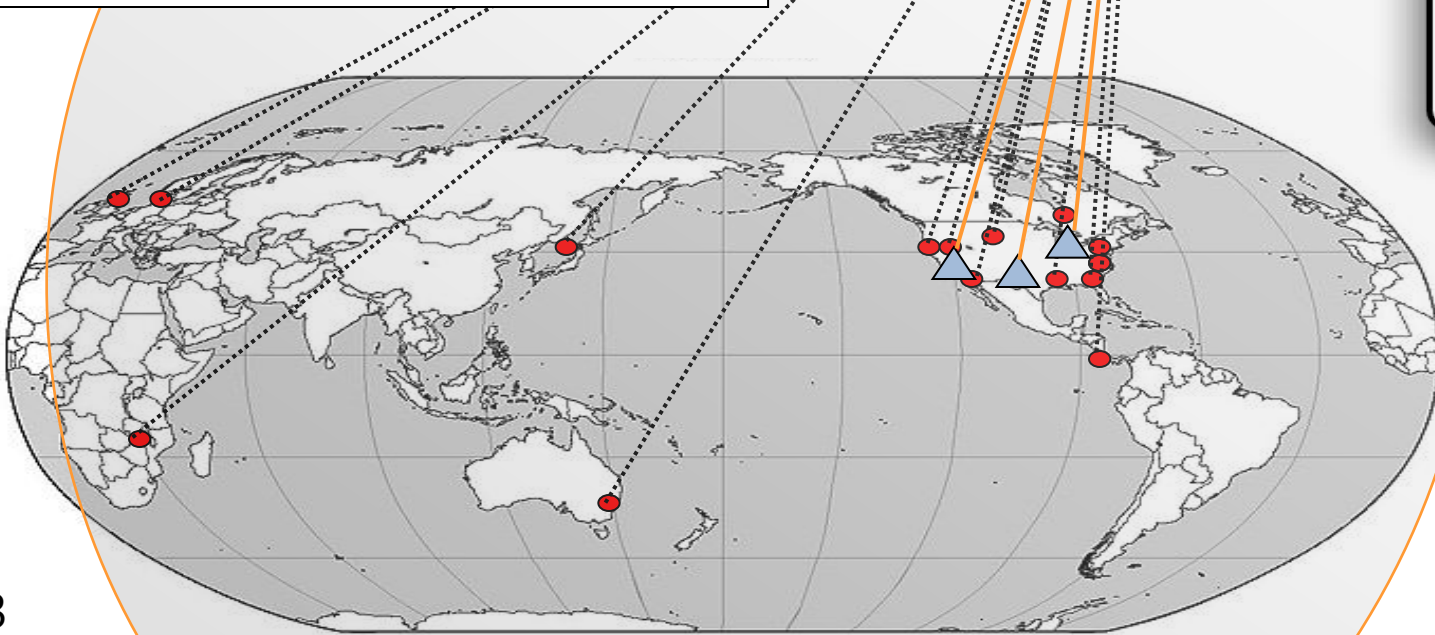
DataONE: Part of the Solution

Coordinating Nodes

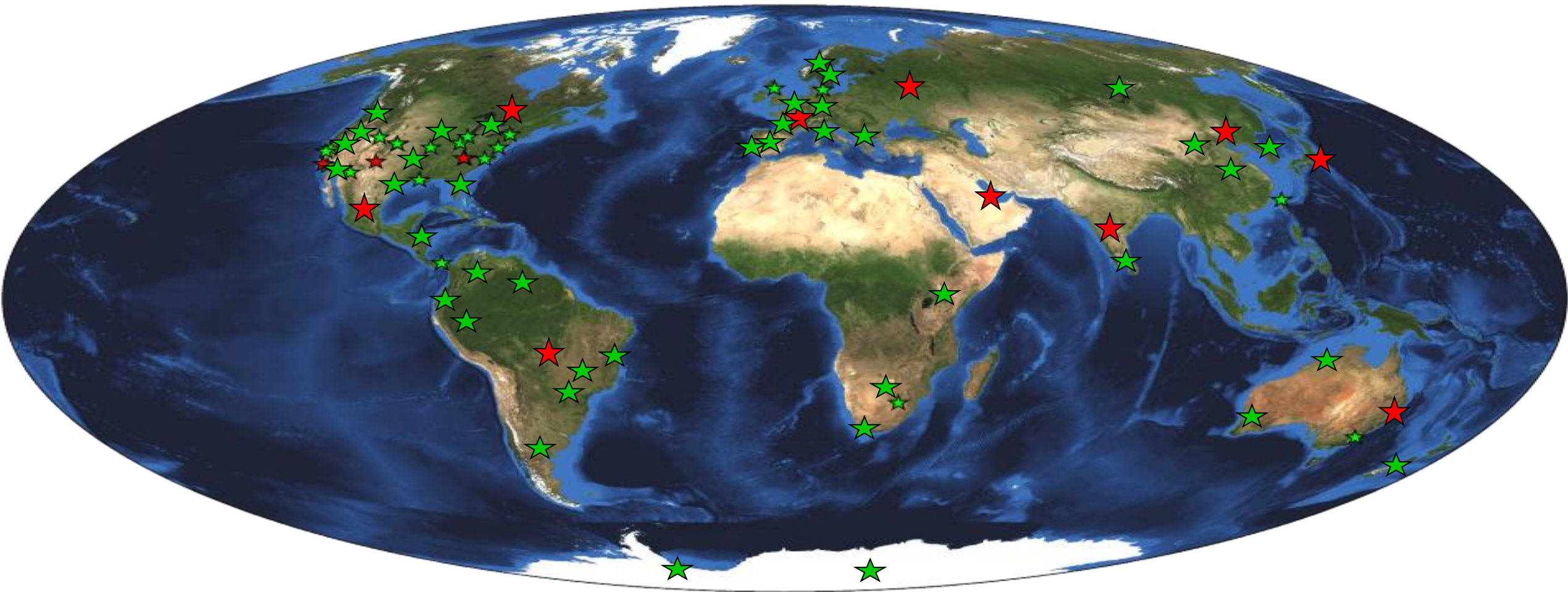
- Retain complete metadata catalog
- subset of all data
- Perform basic indexing
- Provide network-wide services
- Ensure data availability (preservation)
- Provide replication services

Flexible, scalable, sustainable network

Investigator 1..N Toolkit



DataONE: Part of the Solution



DataONE: Part of the Solution

1. Engage the community

- Assessments
- Usability studies
- DataONE Users Group



2. Build on existing CI (e.g., Investigator Toolkit)

- Many existing open source efforts exist
 - Data management: MATT, UDig, Specify
 - Analysis and modeling: R, Octave
 - Workflow systems: Kepler, Taverna, Triana, Pegasus
 - Grid systems: Condor, Globus, BOINC
 - Data and workflow portals: VegBank, myExperiment
- Commercial tools are extremely important too
 - Excel, MATLAB, SAS, ArcGIS
- DataONE: help communities build their own tools
 - Integrate, interoperate, stabilize
 - Create libraries to DataONE Service Interface

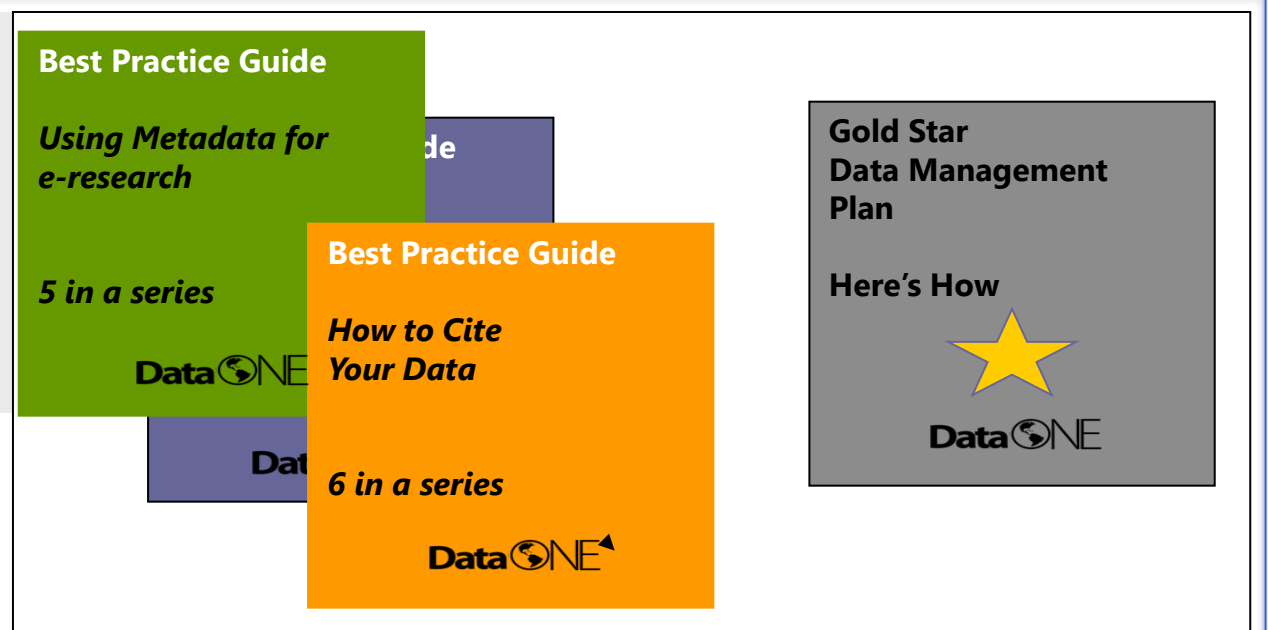


DataONE: Part of the Solution

3. Educate


Career Long Learning:

- best practice guides
- exemplary data management plans
- podcasts, web-casts
- workshops and seminars
- downloadable curricula



DataONE: Part of the Solution

4. Engage Citizens in Science



Project BudBurst
A National Phenology Network Field Campaign for Citizen Scientists

Learn why phenology is important

Participate!

Report your observations online

Does climate change affect budburst?

Download free materials

Map results from around the country

www.budburst.org

Logos for participating institutions: University of Wisconsin-Milwaukee, PCA, University of Montana, UCSB, Wisconsin, and others.



eBird



© 2007 Cornell Lab of Ornithology

Home

About

Project Gateway

References

Toolkit

Conference Proceedings

Discussion Forum

Citizen Science at the Cornell Laboratory of Ornithology

Welcome to Citizen Science Central!
A clearinghouse for ideas, news, and resources in support of citizen science—participate between volunteers and scientists that answer real-world questions.

IDEAS
About this Initiative
Discussion Forum

NEWS
News
Events

RESOURCES
Toolkit
References
Projects
Proceedings

169 Sapsucker Woods Road, Ithaca, NY 14850
1-800-643-BIRD | coms@cornell.edu

Open Notebook



USA **nphn** National Phenology Network

www.CitizenScience.org

5. Enable new science and demonstrate success



The screenshot shows the DataONE search interface. At the top left is the DataONE logo, and at the top right is the text "A Pilot Catalog For Earth Observations". Below this is a navigation bar with "DataONE Metadata Clearinghouse" and a "HELP" link. The main search area has tabs for "Simple Search" and "Advanced Search". The search prompt is "Search All Records For". There is a search input field, a "SEARCH" button, and a "Results/Page" dropdown menu set to "10". A hint below the input field reads: "Hint: boolean operators, wildcards and phrases are allowed. ex: precipitation or (rain* and 'moisture content')". Below the search area, it says "Query being built:" followed by a large, empty, non-editable text area. A "CLEAR QUERY" link is at the bottom left.

>70,000 Data Products

NBII Metadata Clearinghouse

Long Term Ecological Research (LTER) Network

ORNL Distributed Active Archive Center for Biogeochemical Data

Large Scale Biosphere-Atmosphere Experiment in Amazonia (LBA)

Organization of Biological Field Stations

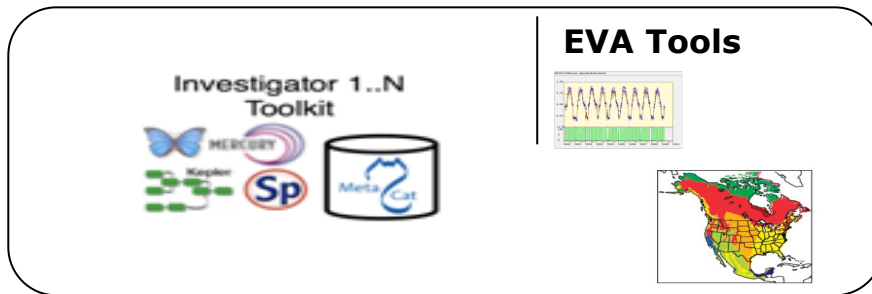
Inter-American Institute for Global Change Research (IAI)

MODIS and ASTER Products (LPDAAC)

National Phenology Network (USANPN)

DataONE: Part of the Solution

- Exploration, Visualization and Analysis (EVA) Working Group
 - Guide creation of scientific workflows and DataONE data exploration, visualization and analysis functionality.

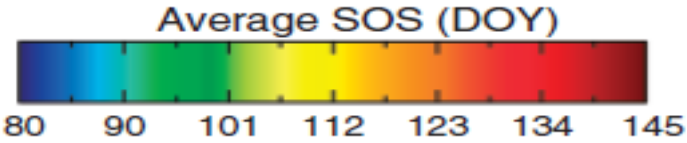
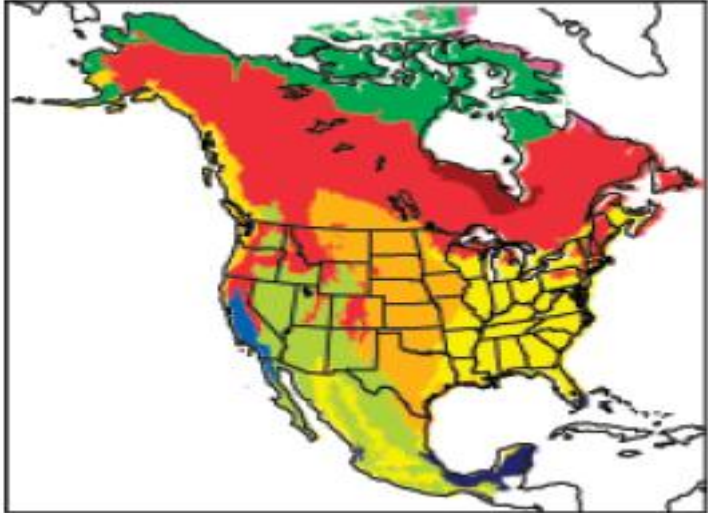


DataONE: Part of the Solution

1st EVA Example: Vegetation and Bird Phenology at 130,000 sites in North America



Phenology
(Start of Season)



(White et al. 2009)



Understand how bird migration patterns change through time

DataONE: Part of the Solution

The Process:

- Iterative data-intensive workflow – VisTrails
- Data synthesis/integration
- Spatio-Temporal Exploratory Model (STEM)
- Ornithologists/macroecologists



Occurrence of Indigo Bunting (2008)

Neotropical migrant that winters
in central and south America

Environmental Science 2020: 3 Case Studies

1. Conserving the World's Biodiversity
2. Acoustic Monitoring for Conservation
3. Assessing and Mitigating Environmental Risk

1. Conserving the World's Biodiversity

- Data are fundamental
 - Technology now greatly aids in digitization and character recognition
 - But, every specimen must still be “handled”
- Names are fundamental
 - Need for consistency both in organism taxonomic classifications and in mapping between classifications
 - Need extends to consistency in taxonomic representation of scientific names that are used in genes inserted in GM organisms
 - Huge implications for conservation, policy, and enforcement
- Possible solutions
 - Citizen science and crowd-sourcing
 - Emerging social web and linked-data technologies



2. Acoustic Monitoring for Conservation



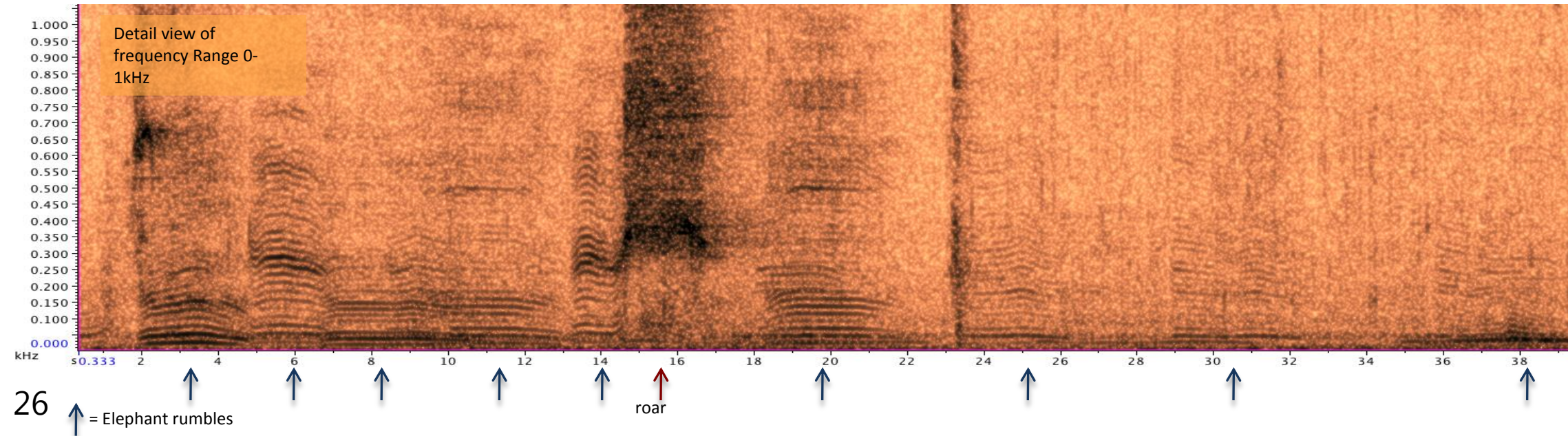
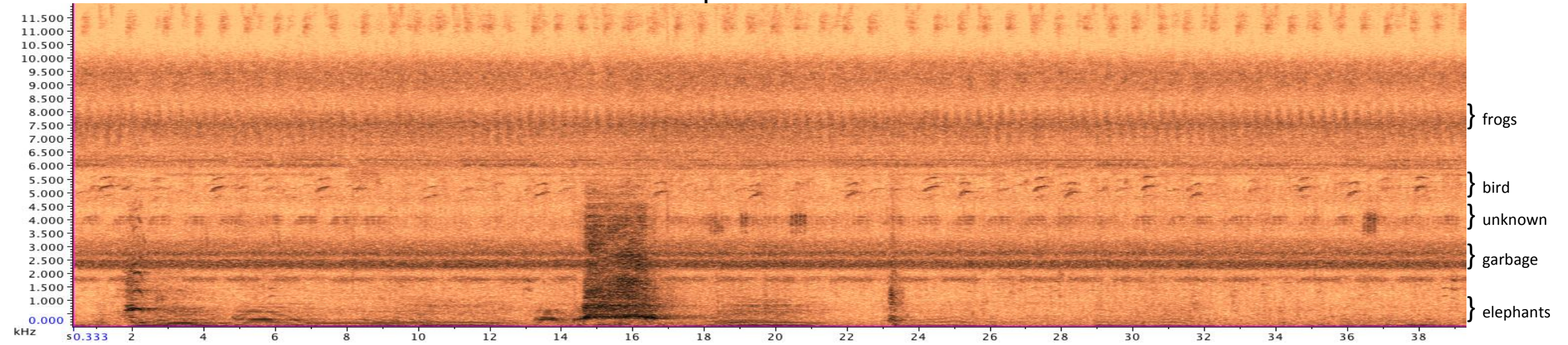
Forest Elephants live in the dense rainforests of Central Africa – hard to see – easier to hear

- The Elephant Listening Project uses autonomous acoustic recorders to capture the infrasonic calls of elephants 24/7
 - one unit 'listens' to > 3 km² of rainforest
 - no human interference, relatively low cost
- Unique findings from 2 years of acoustic monitoring:
 - discovered the most active hotspot for forest elephants in Gabon
 - documented behavior shifts in response to oil exploration – not detectable by standard methods
 - quantify actual hunting pressure by recording gunshots of hunters
 - determine elephant abundance, age/sex ratio based on different call signatures, quantify reproductive behavior
- The Potential for Conservation
 - richness of acoustic data: all species using acoustic communication, human disturbance indicators (vehicles, gunshots)
 - seasonal, daily, geographical, and historical activity patterns
 - abundance, presence/absence, comparative evaluations in space even without absolute estimate of density
 - continuous, autonomous acquisition

2. Acoustic Monitoring for Conservation



Complex Dense Data Matrix



2. Acoustic Monitoring for Conservation

- **Storing & Manipulating**
 - Ongoing project currently has 100,000 hrs of Congo Basin forest sounds
 - Current set of recorders gathering 216,000 hrs per year
 - Using optimal sampling rate = >31,000 GB of sound data/yr
- **Analysis Limits**
 - For elephants, current analysis requires visual examination of files
 - 0.54 hrs of analyst effort required per day of recording for minimum subsample of elephant calls
 - 2.15 hrs to completely analyze 24 hours of sound, just for elephants
- **Possible Solutions: Automated Detection & Manipulation Needed**
 - Need detectors that can process arbitrary length files for sounds of interest
 - Detector algorithms need user-set parameters for different sound targets
 - Methods to automatically cut out segments of files based on detector results to serve to analysts, maintaining time information
 - Methods to view context of clipped segments in original sound

3. Assessing and Mitigating Environmental Risk

eBird and
Gulf Coast
Oil Spill

The New York Times

Wednesday, June 2, 2010 Last Update: 3:19 PM ET

Estimates Suggest Spill Is Biggest in U.S. History



3. Assessing and Mitigating Environmental Risk

eBird and Gulf Coast Oil Spill

News

eBirders mobilize! Help survey Gulf Coast birds!

May 4, 2010

The ongoing oil spill disaster in the Gulf of Mexico will undoubtedly impact bird populations in the region for years to come. How can you help? eBirders can make a difference by surveying local beaches and marshes for birds. By getting out now and reporting the birds you find to eBird, your observations will provide a real-time snapshot of the region's birdlife, helping conservationists and researchers understand where, when, and how many of each species are currently occurring on local beaches and wetlands. If the oil does make landfall, we'll have recent data from all around the Gulf to help prioritize and focus conservation efforts. As time goes on, continued beach surveys will help conservationists assess the impacts of the spill. Act now! Survey birds tonight, tomorrow, and in the coming days on as many Gulf Coast beaches as possible. Read more to find out what to do.



**Brown Pelican, Sanibel Is., Florida.
Photograph by W.H. Majoros.**

3. Assessing and Mitigating Environmental Risk

eBird and Gulf Coast Oil Spill

[eBird - Gulf Coast Oil Spill Bird Tracker](#)

eBird | Gulf Coast Oil Spill Bird Tracker

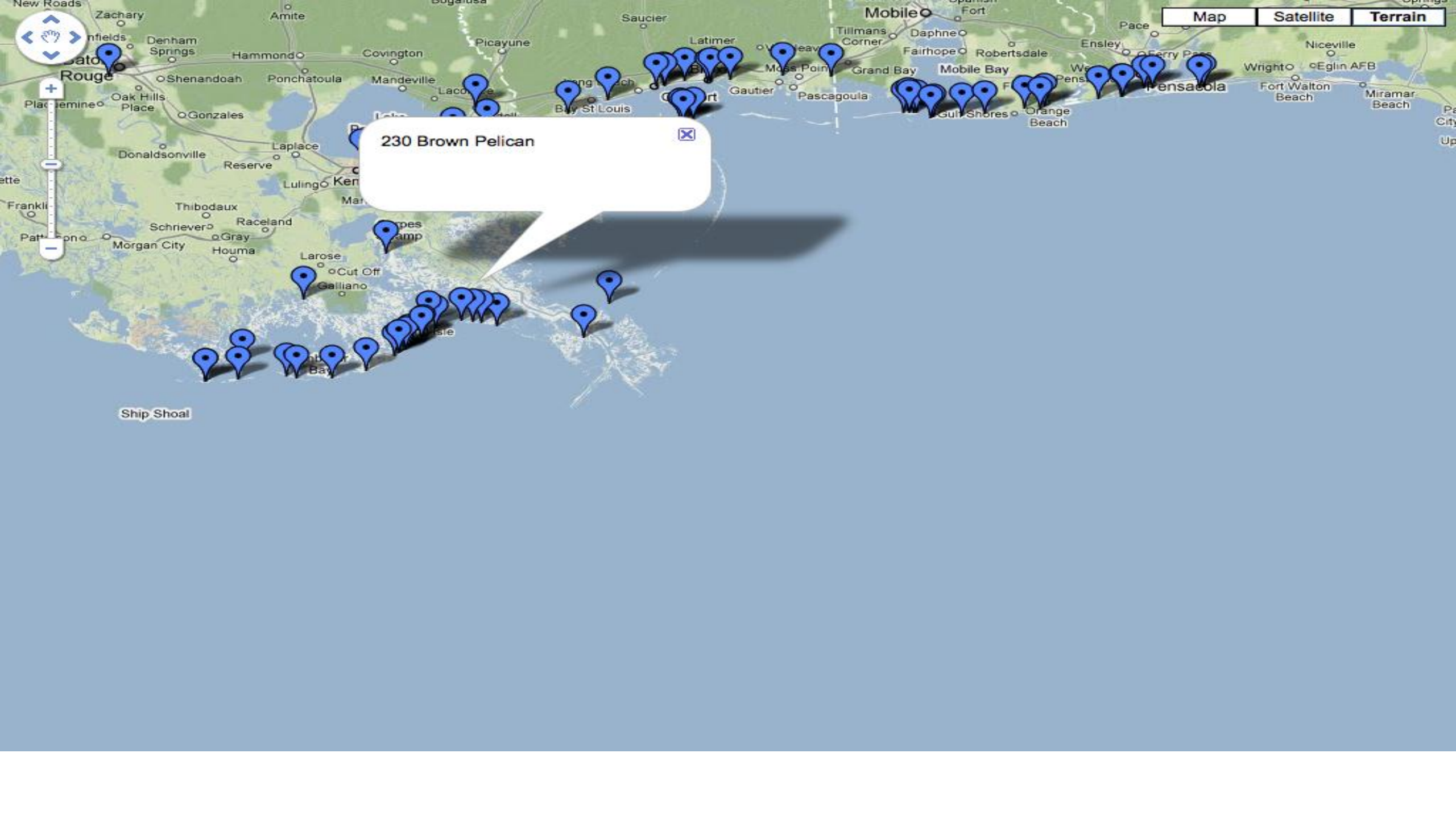
Hundreds of bird species could be impacted by the Gulf Coast oil spill, including species shown here.

Brown Pelican

Recently removed from the endangered species list, the rebounding population nests on coastal islands throughout the impact zone.

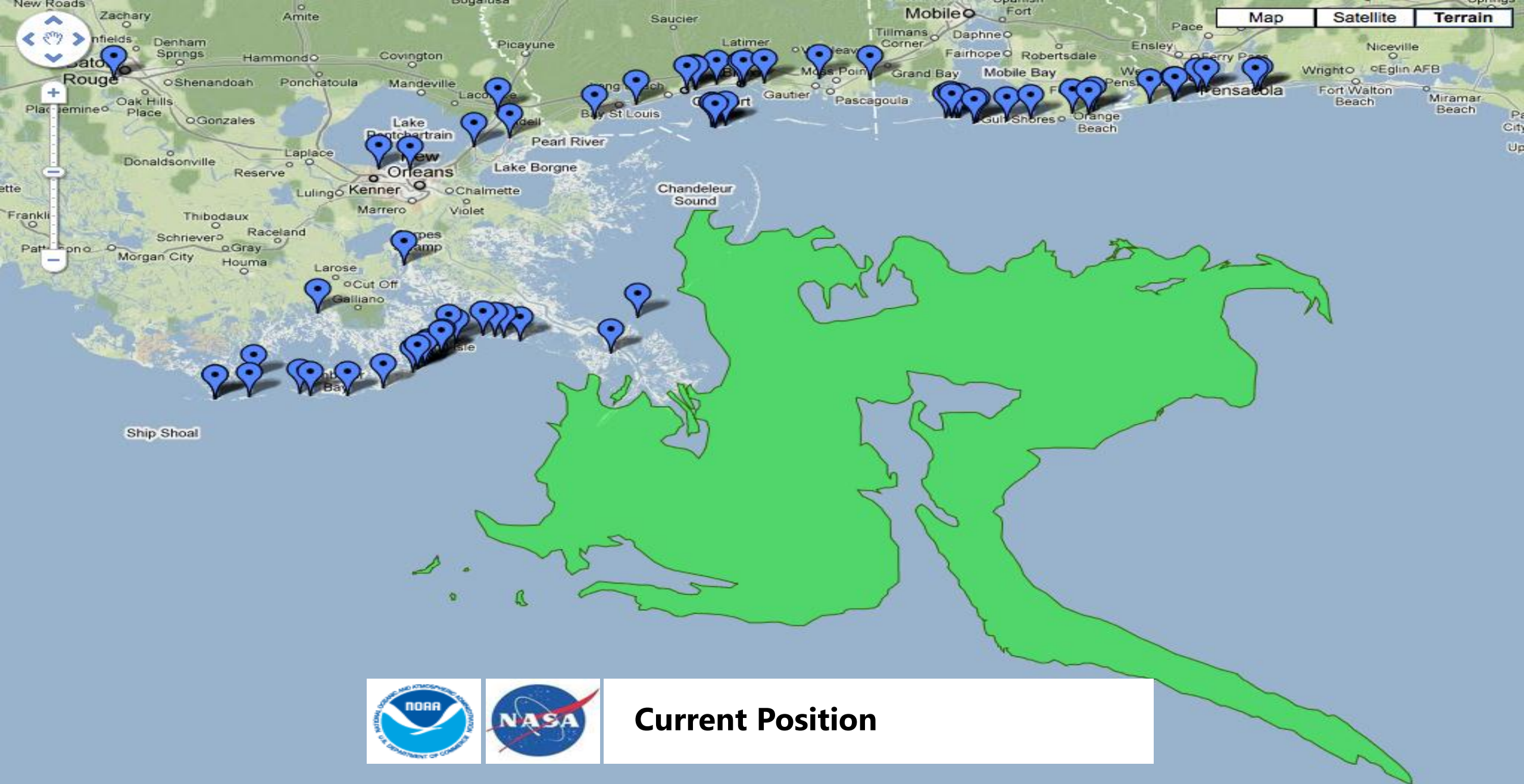


[Explore interactive sightings map](#)

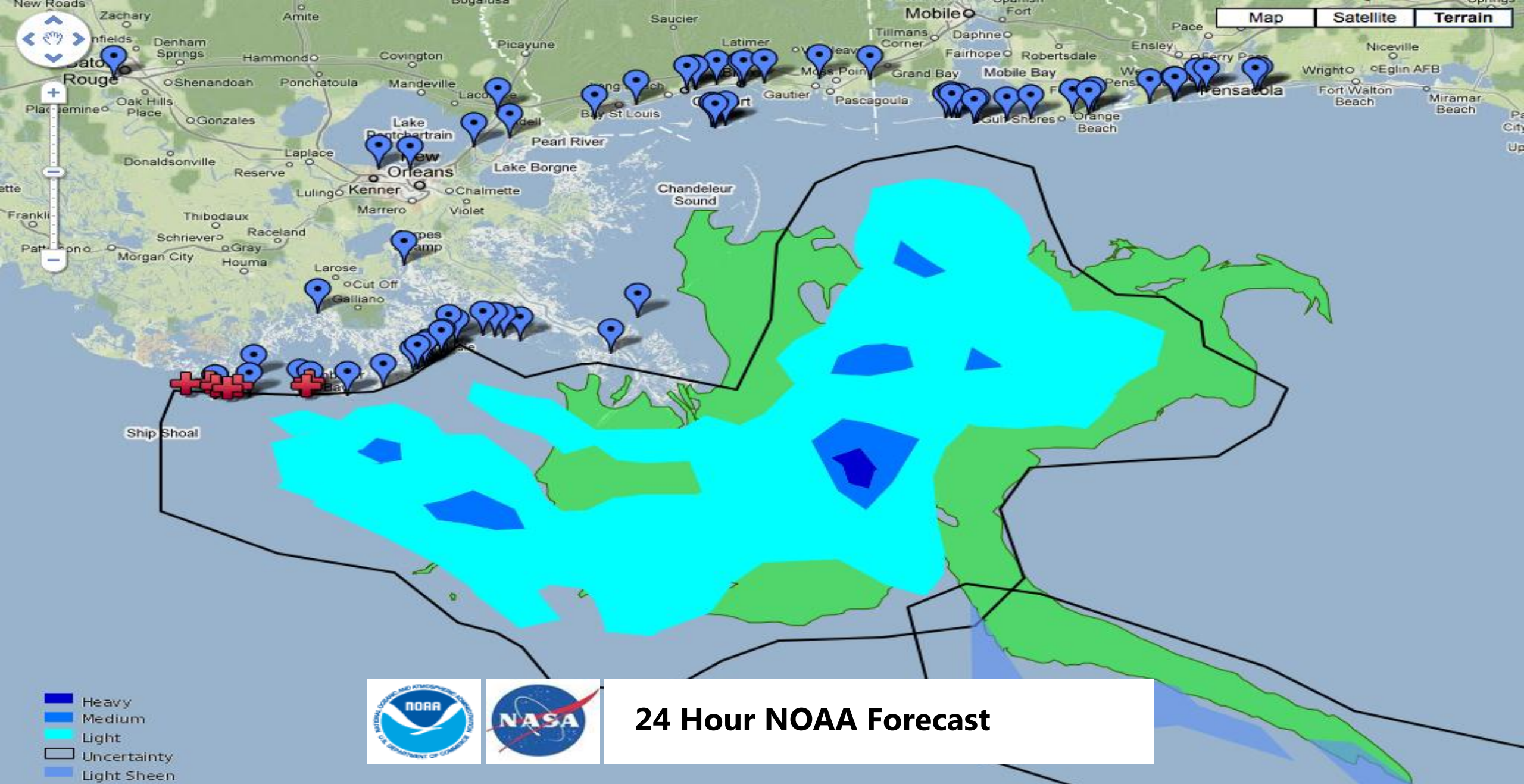


230 Brown Pelican

Ship Shoal



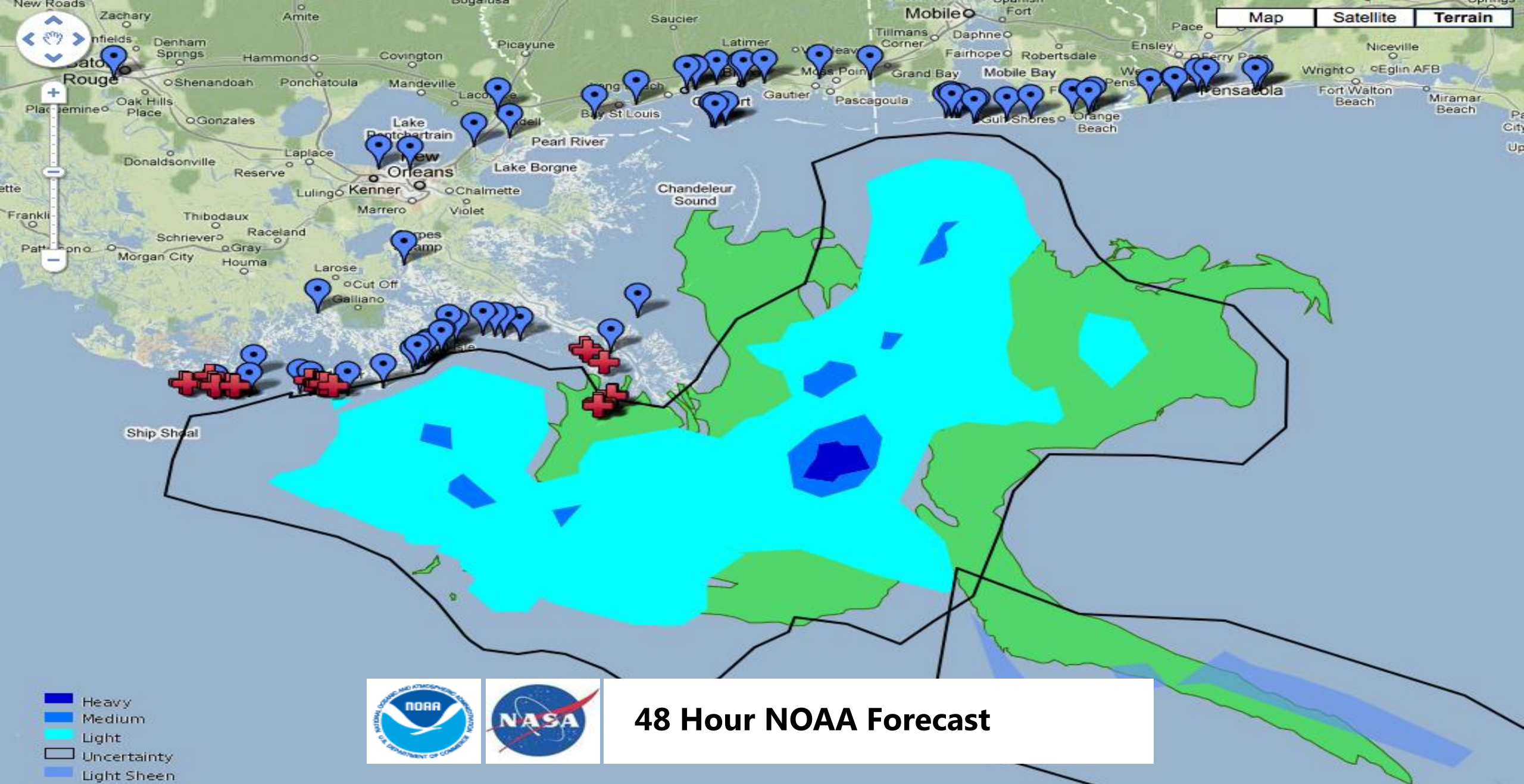
Current Position

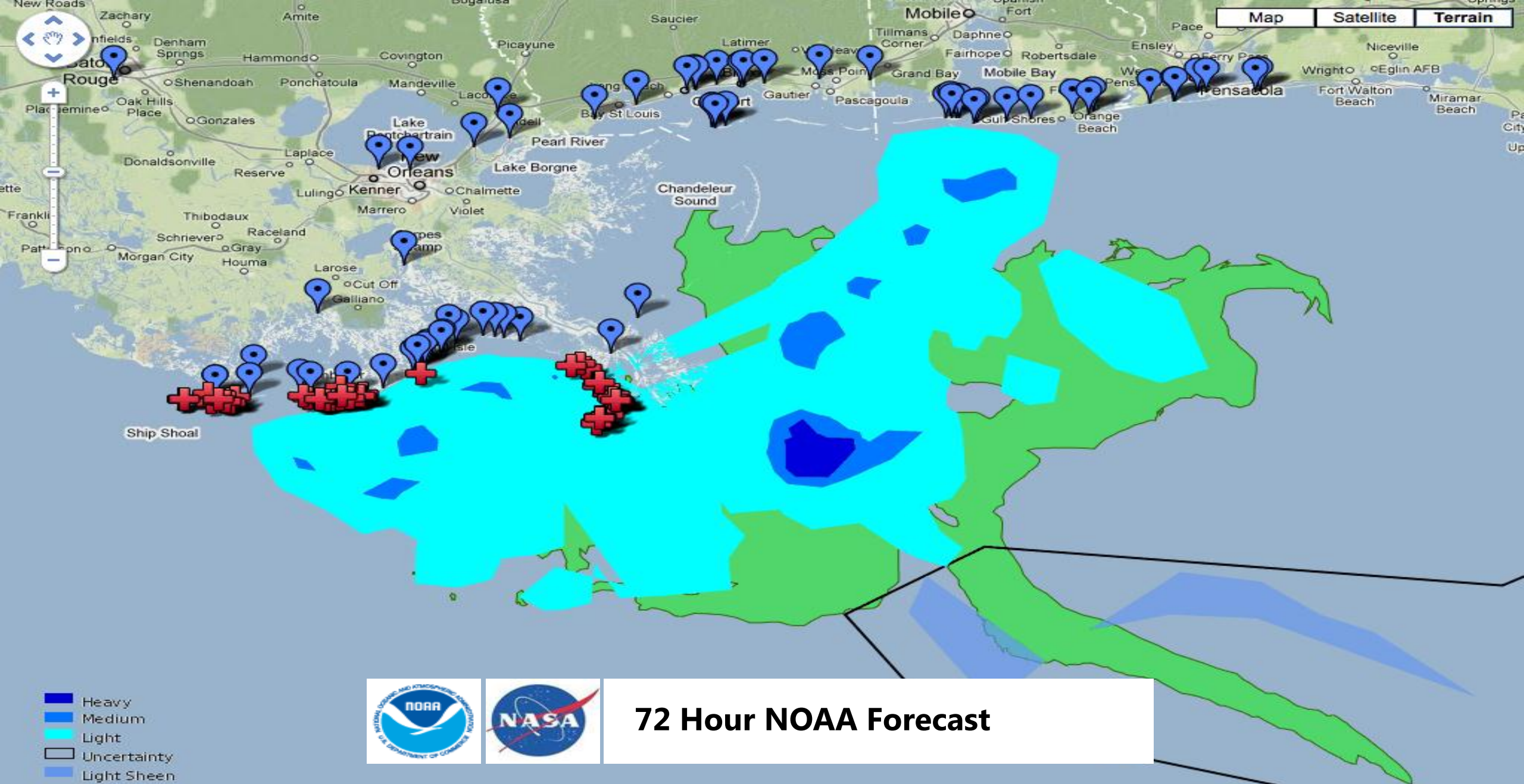


- Heavy
- Medium
- Light
- Uncertainty
- Light Sheen



24 Hour NOAA Forecast

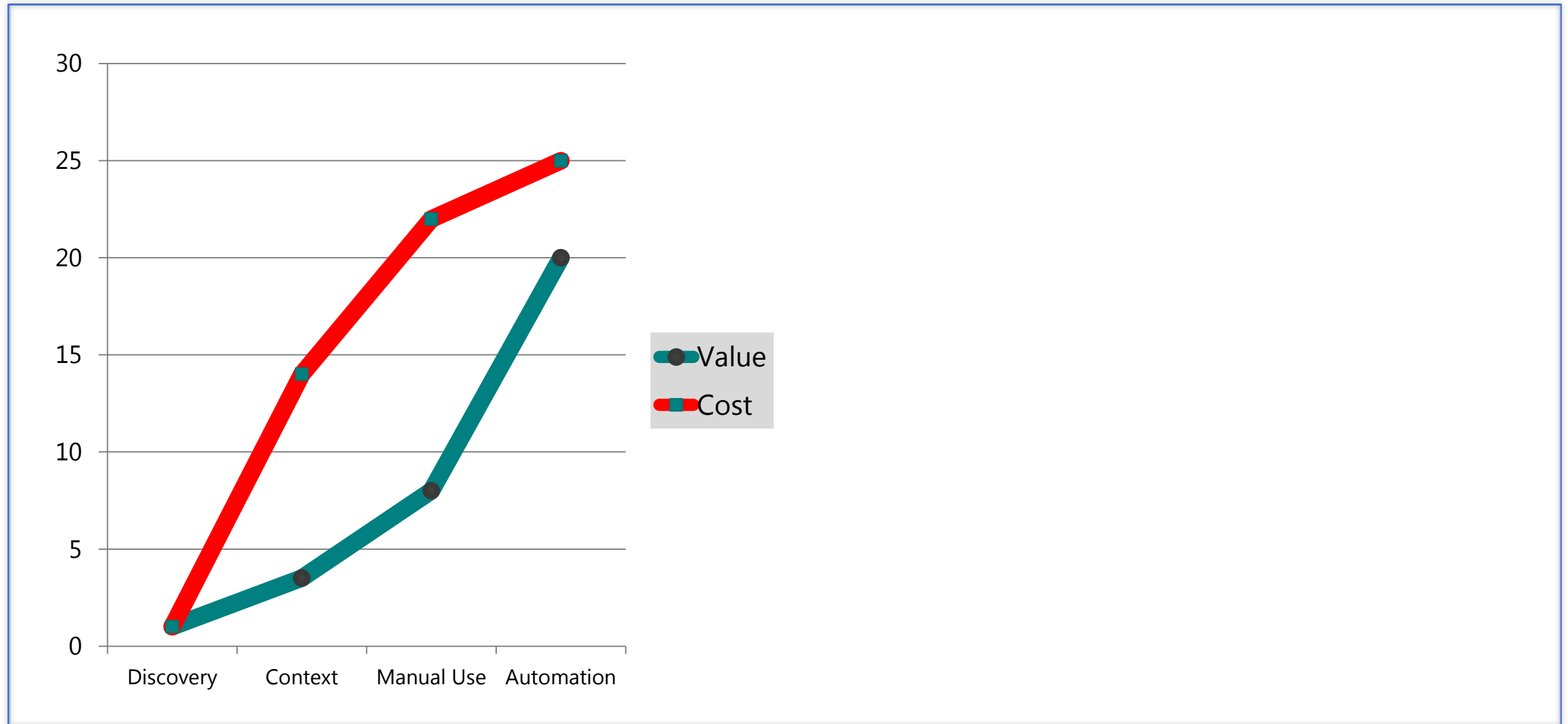




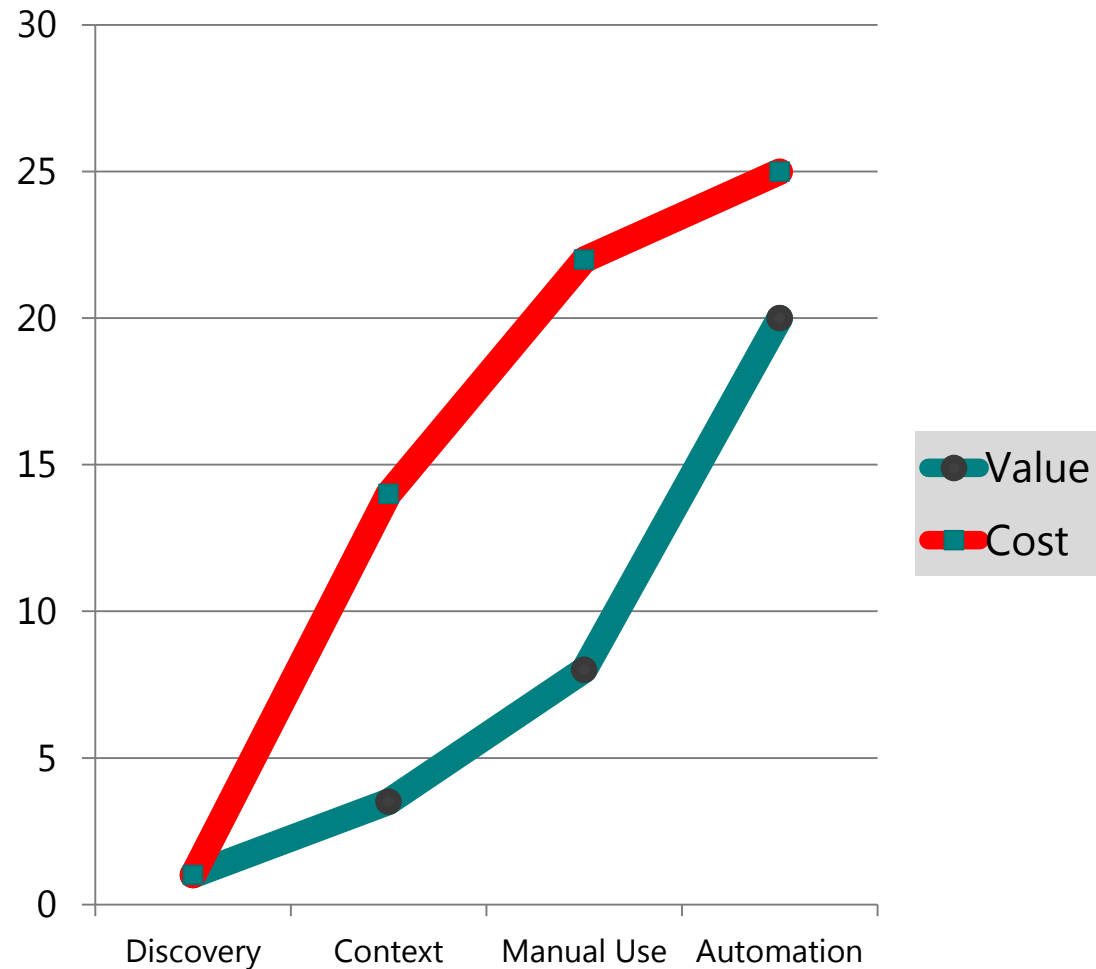
3. Assessing and Mitigating Environmental Risk

1. Forestall Data Entropy
 - Secure and replicated archives – e.g., DataONE
 - Comprehensive metadata
2. Enable Data Discovery
 - Expose data/metadata to registries, etc.
 - Provide discovery-level metadata
3. Mediate Data Heterogeneity
 - Adopt community standards
 - Support ontology development and semantic mediation tools (e.g., observation ontologies)
 - Invest in value-added (i.e., integrated) databases
4. Facilitate Data Interpretation
 - Comprehensive metadata
 - Expose workflows and data provenance
 - Advance the state of visualization tools

Metadata: The Grand Data Management Challenge:



Metadata: The Grand Data Management Challenge:



Technology Solutions:

- (Semi-) automation
- Metadata-capable sensors

Sociocultural Solutions:

- Education
 - good science
 - best practices

Technology and Socio-cultural:

- User-centered design
- Usability testing
- Tenure and promotion

Thanks!

- Cornell University – Steve Kelling, Peter Wrege
- Oak Ridge National Laboratory – Bob Cook
- University of Kansas – Dave Vieglais
- Funding from:
 - NSF DataNet, CISE Pathways Computational Sustainability and INTEROP Programs
 - Leon Levy Foundation
 - National Aeronautics and Space Administration

