

DataONE: Enabling Data-Intensive Biological and Environmental Research through Cyberinfrastructure

Presented by

John W. Cobb, Ph.D.

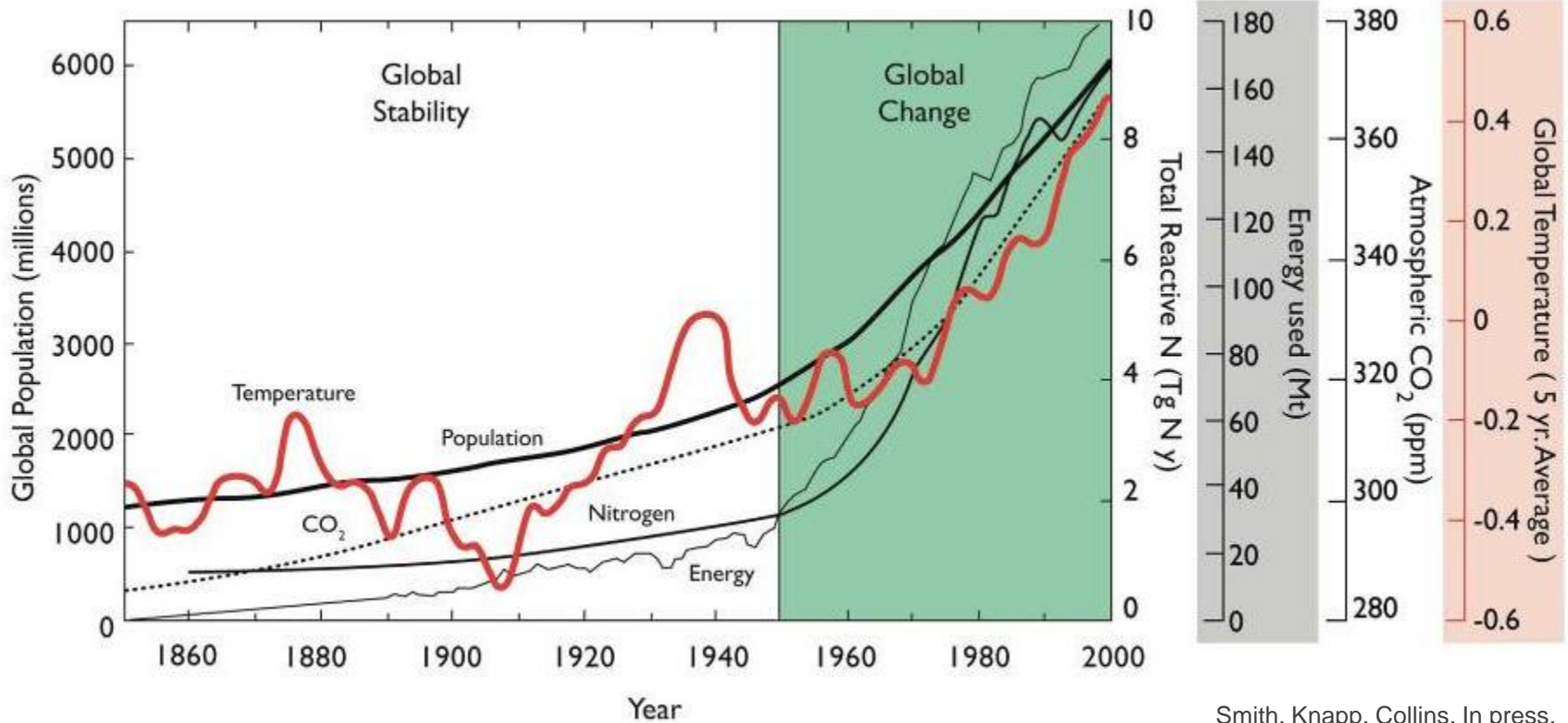
Computer Science and Mathematics
Office of Cyberinfrastructure
National Science Foundation

In collaboration with the DataONE team:

PI: Bill Michener, University of New Mexico, and ORNL, U. Tennessee, UC Santa Barbara, Cornell, U. Cal. Digital library, USGS, Duke U., U. N. Carolina, U. Ill.–Chicago, U. Kansas, Utah State U.

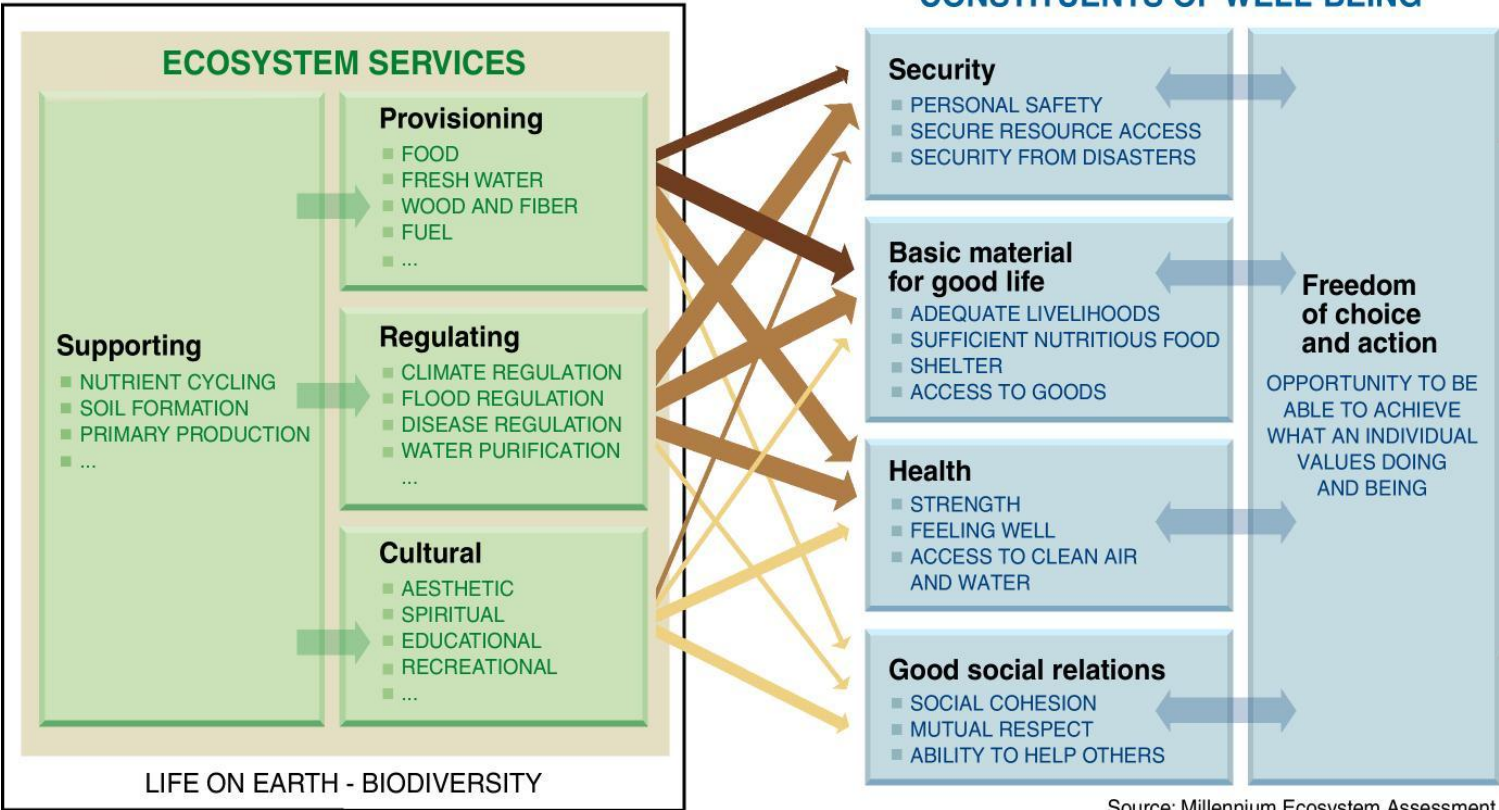


Global change research: Multiple data sources



Smith, Knapp, Collins. In press.

With coupled human and natural systems



Source: Millennium Ecosystem Assessment

ARROW'S COLOR
Potential for mediation by socioeconomic factors

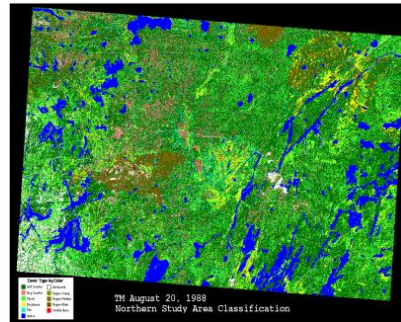
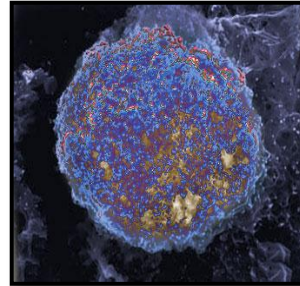
- Low
- Medium
- High

ARROW'S WIDTH
Intensity of linkages between ecosystem services and human well-being

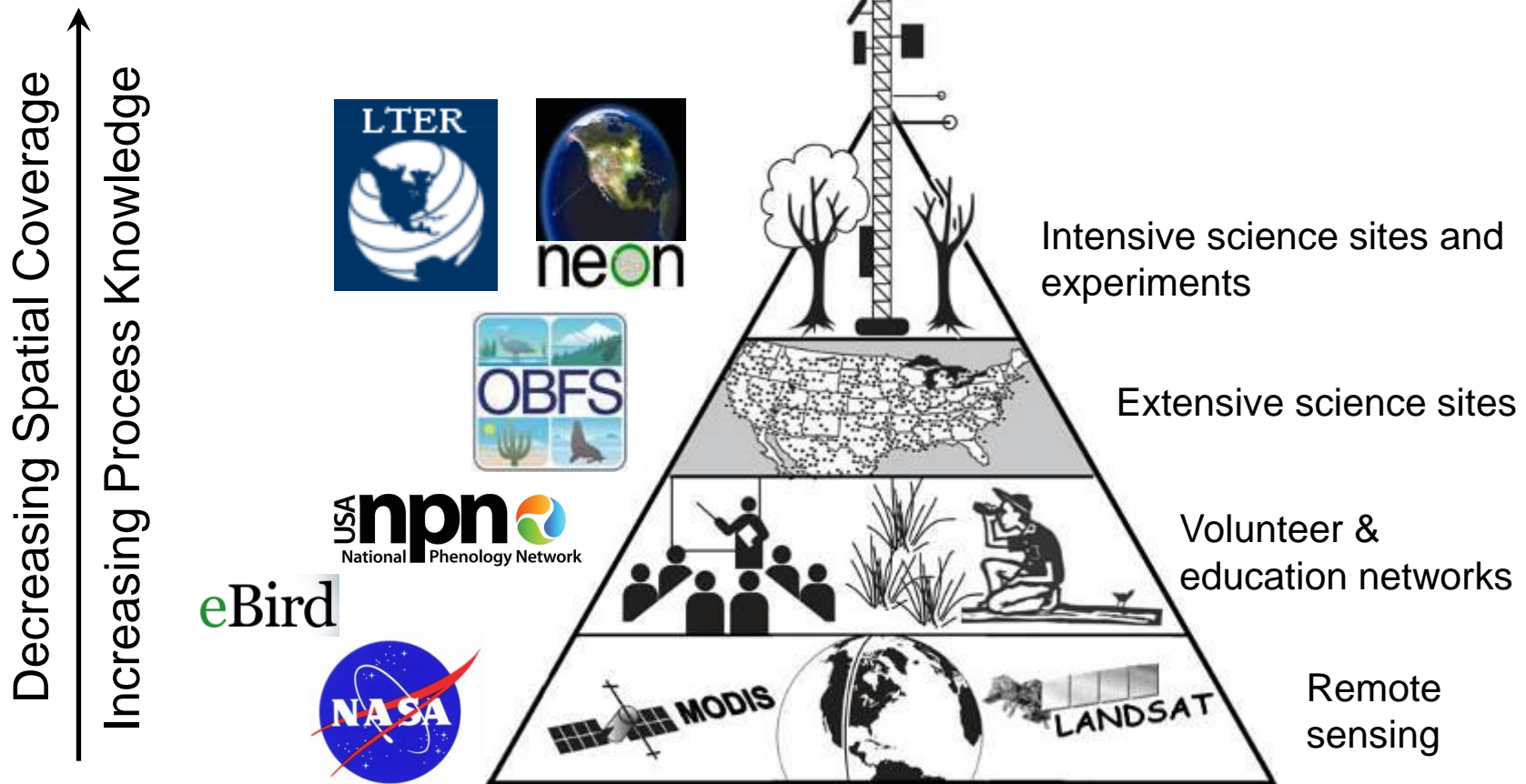
- Weak
- Medium
- Strong

DataONE data sources

- Research networks and environmental observatories
- Biological specimens
- Individual scientists
- Citizen scientists' data
- Natural resources and conservation data
- Observational data
- Global and continental land cover/land change and biogeochemical data



Heterogeneous data integration for scales small to large



Adapted from CENR-OSTP

Scattered data sources

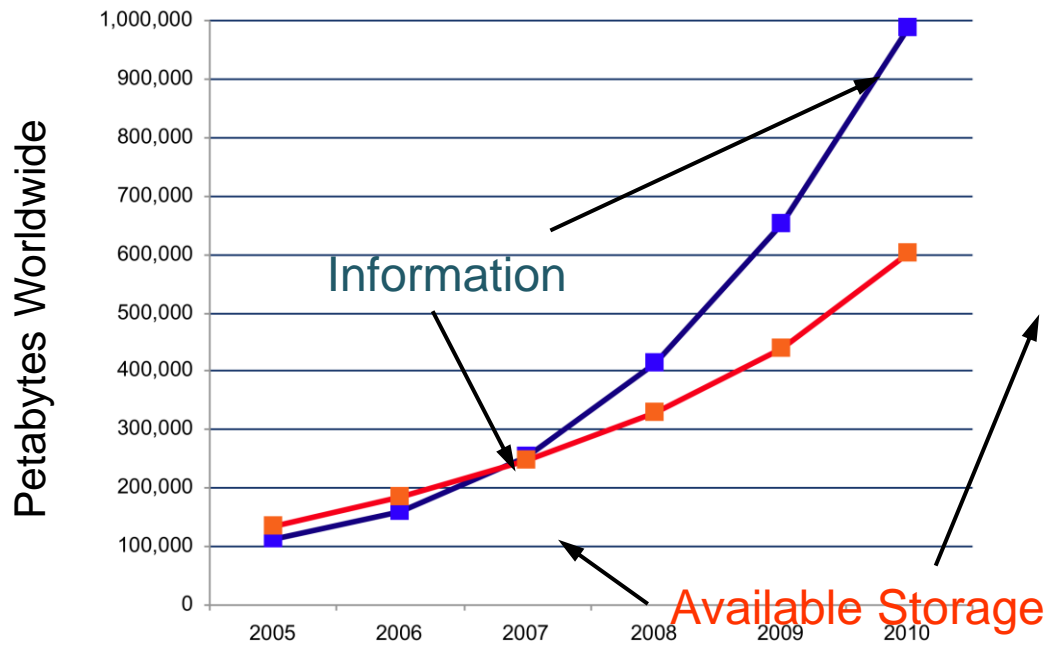
“finding the needle in the haystack”

Data are massively dispersed

- Ecological field stations and research centers (100s)
- Natural history museums and biocollection facilities (100s)
- Agency data collections (100s to 1000s)
- Individual scientists (1000s to 10,000s to 100,000s)

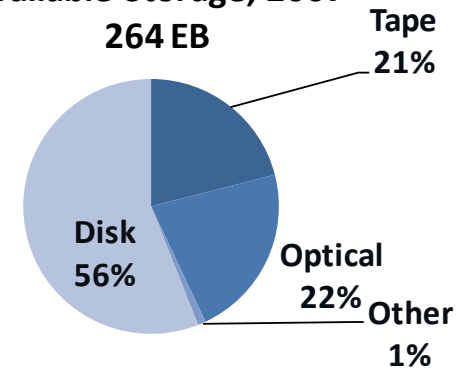


Data production exceeds storage



Transient information or unfilled demand for storage

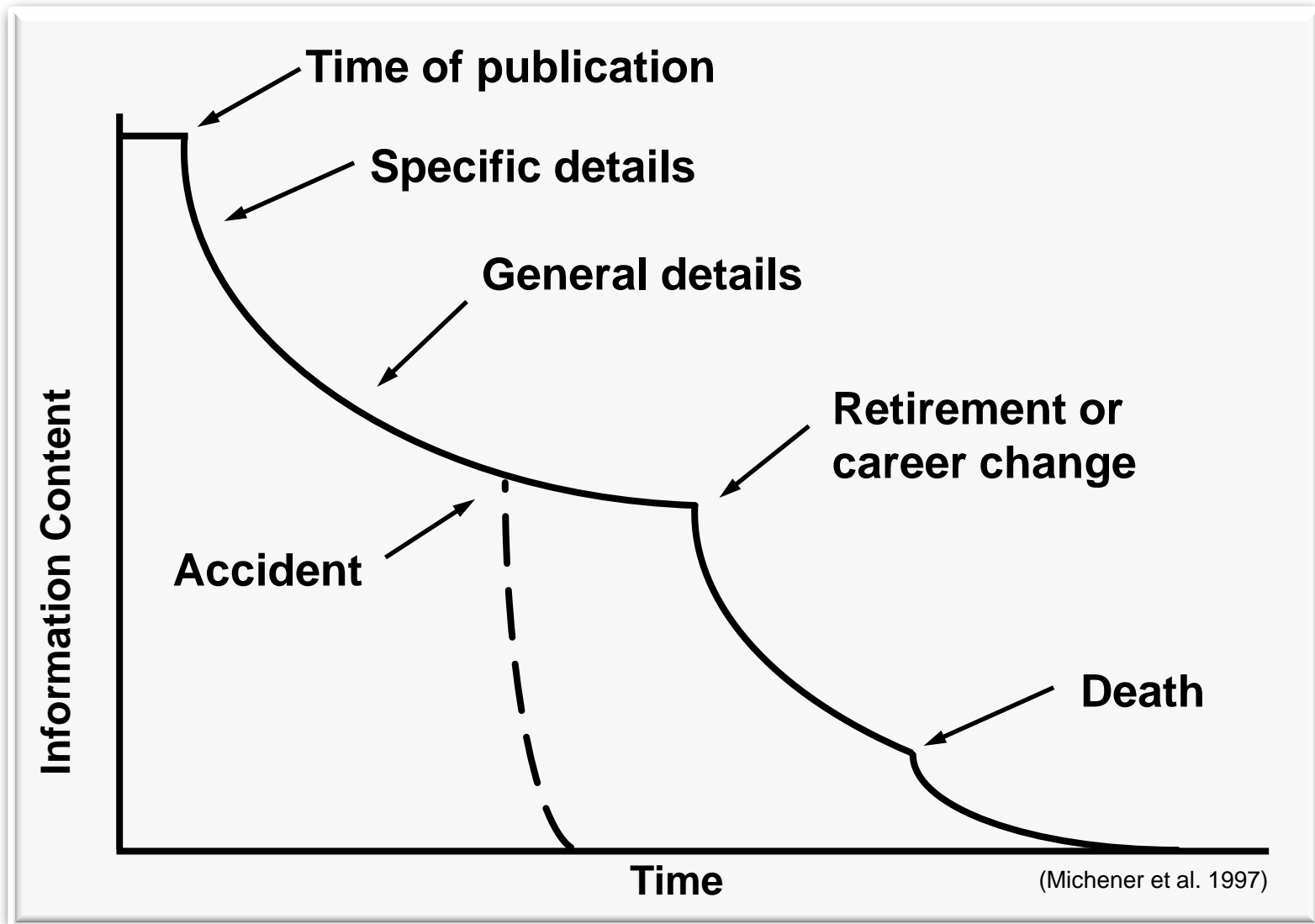
Available Storage, 2007
264 EB



Source: John Gantz, IDC Corporation: The Expanding Digital Universe

Poor data practice

“data entropy”

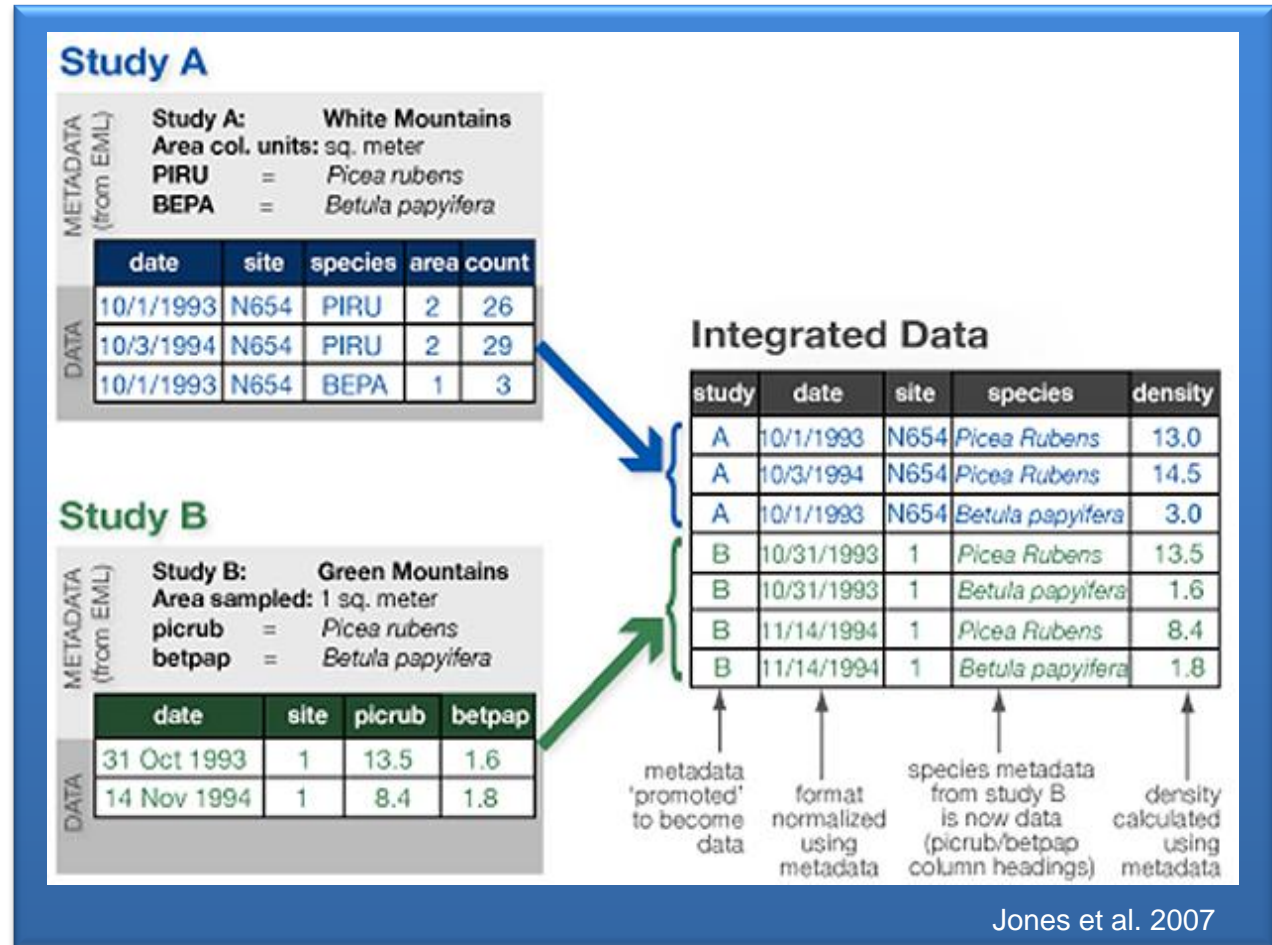


Data deluge

“the flood of increasingly heterogeneous data”

Data are heterogeneous

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



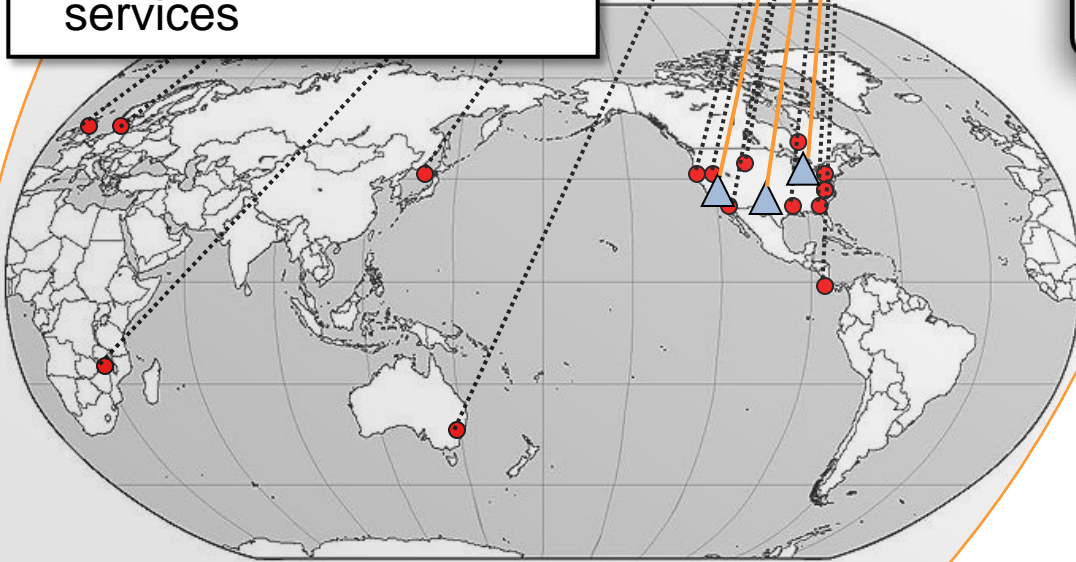
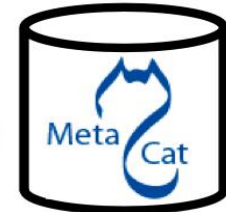
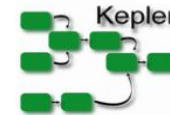
Distributed framework

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



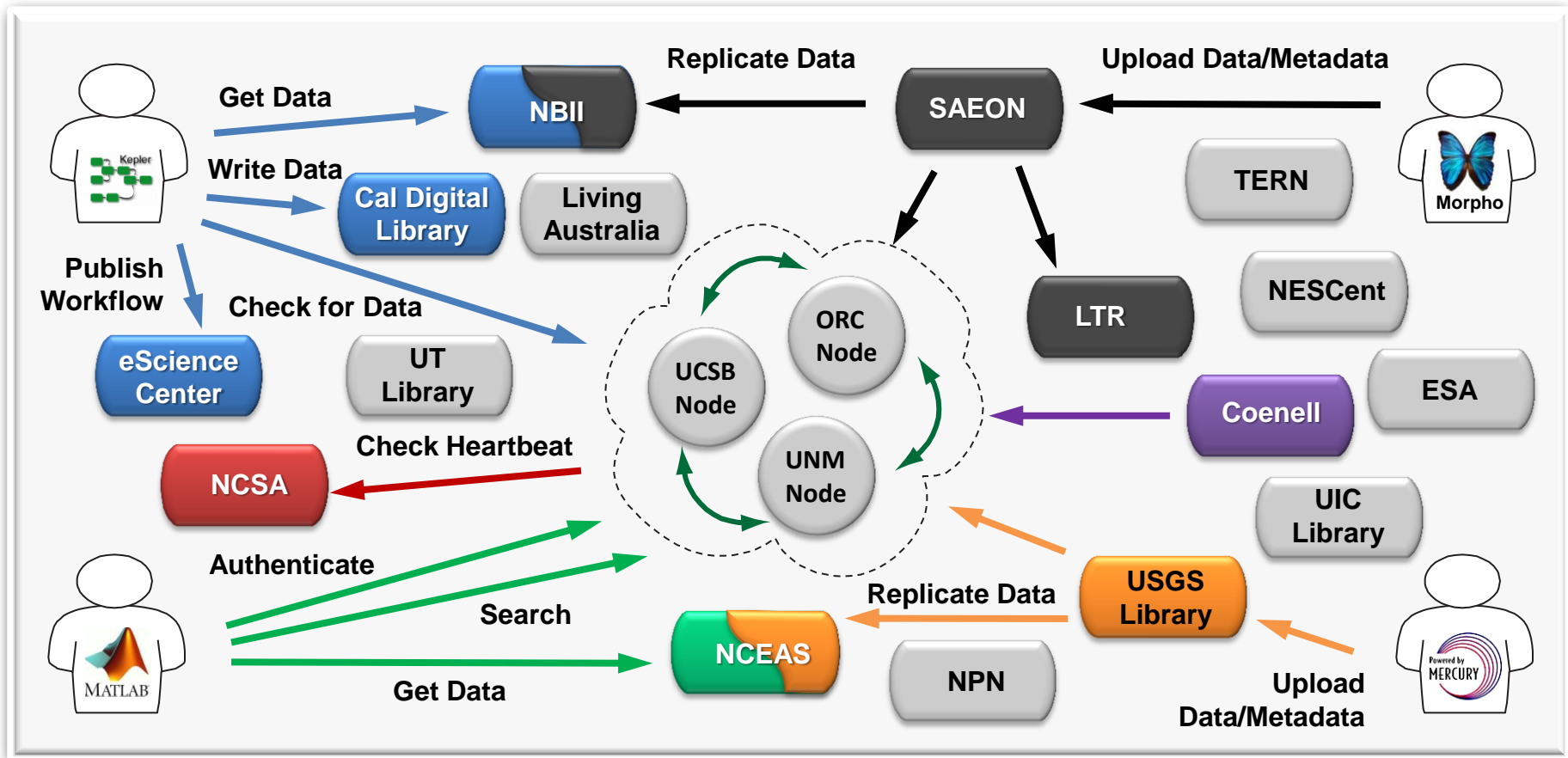
Examples of data holdings

Metadata interoperability across data holdings

Data Archive	Types of Data Managed	Metadata Standard(s)
	Biodiversity, taxonomic, ecological	BDP, DwC, DC, OGIS
	Biogeochemical dynamics, terrestrial ecological Earth observation imagery	DIF, BDP, ECHO
	Ecological, biodiversity, biophysical, social, genomics, and taxonomic	EML
	Avian populations and molecular biology	DwC
	Biological and taxonomic	DC subset
	Biophysical, biodiversity, disturbance, and Earth observation imagery	EML
	Biodiversity, biotic structure, function/process, biogeochemical, climate, and hydrologic	EML

BDP = Biological Data Profile	DIF = Directory Interchange Format
DC subset = Dublin Core subset	ECHO = EOS ClearingHouse
DwC = Darwin Core	EML = Ecological Metadata Language
DC = Dublin Core	OGIS = OpenGISZ

Supporting the data lifecycle



The data lifecycle

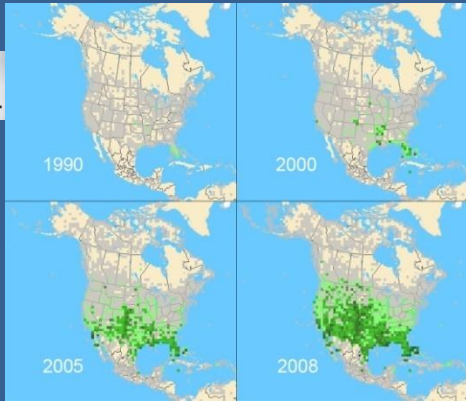
1. Deposition/acquisition/ingest
2. Curation and metadata management
3. Protection, including privacy
4. Discovery, access, use, and dissemination
5. Interoperability, standards, and integration
6. Evaluation, analysis, and visualization

Integrated data analysis

The Eurasian Collared Dove (*Streptopelia decaocto*) was introduced in the Bahamas in 1988. Since then it has spread across North America. There is concern that the Eurasian Collared Dove could compete with native dove species (i.e. Mourning Dove, *Zenaida macrura*, or White-winged Dove, *Zenaida asiatica*), both of which are economically beneficial. An analyst would like to predict how the invasive dove will spread over the next 20 years, and how it might impact the ranges of the other dove species.

First, the **analyst** searches DataONE **clearinghouse** to **discover** and **access** data on the distribution of the dove species. They find that a continent-wide network of citizen scientists has gathered information on the occurrence of Eurasian Collared Dove since it was introduced.

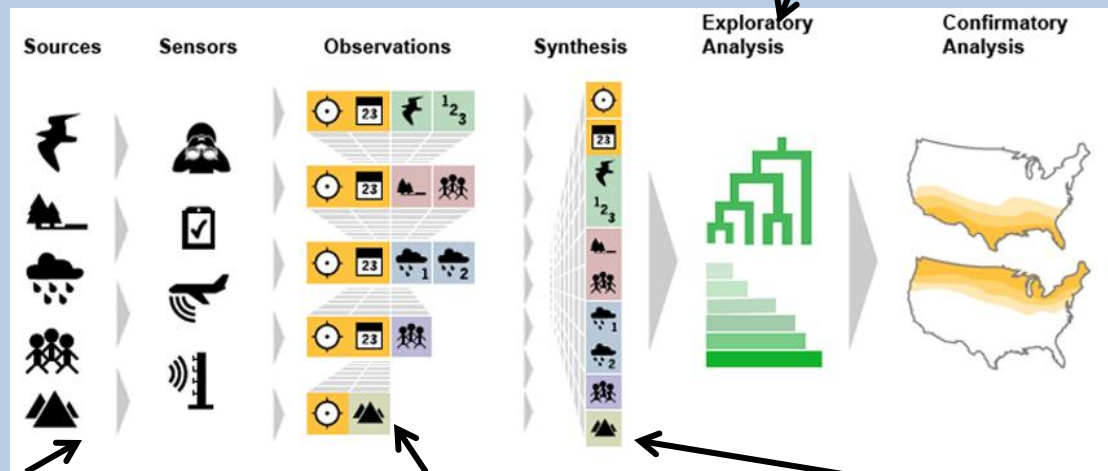
eBird



Range expansion of Eurasian Collared Dove

Second, analysis **workflows** are used to first explore and then predict patterns of species occurrence.

Exploratory analysis techniques that identify the factors that best predict species occurrence and drive hypotheses generation and predictive analysis.

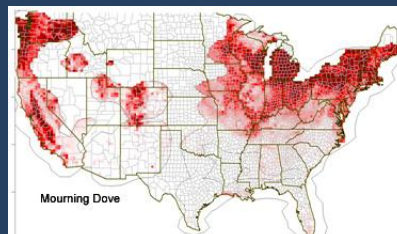
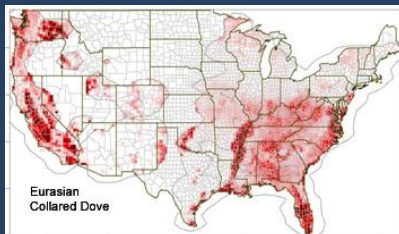


Sources of species observations are linked to landscape, climate, geographical and human factors.

Observations are available through the distributed network of DataONE data providers.

Data are organized via a core semantic model for observational data making **data synthesis** straightforward.

Third, accurate long-range forecasting **models** for each species are presented. **Predicted Ranges of 3 Dove species in 2025.**



Note: maps are examples of possible outcomes, and not actual representations of range.

Engaging 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: UWMILWAUKEE, pca, CHICAGO BOTANIC GARDEN, University of Montana, UCSB, UNIVERSITY OF WISCONSIN, UNIVERSITY OF TEXAS AT AUSTIN.

Images of various flowers and plants.



eBird

>Welcome to Citizen Science Central! - Citizen Science Project Toolkit Mozilla Firefox

<http://www.birds.cornell.edu/cscitoolkit>

CORNELL LAB of ORNITHOLOGY

Citizen Science Central

Welcome to Citizen Science Central!

A clearinghouse for ideas, news, and resources in support of citizen science—partnerships between volunteers and scientists that answer real-world questions.

- Home
- About
- Project Gateway
- References
- Toolkit
- Conference Proceedings
- Discussion Forum

IDEAS

- About this Initiative
- Discussion Forum

NEWS

- News
- Events

RESOURCES

- Toolkit
- References
- Projects
- Proceedings

© 2007 Cornell Lab of Ornithology Home | Login

CORNELL LAB of ORNITHOLOGY

169 Sapsucker Woods Road, Ithaca, NY 14850
1-800-663-8910 | cornellite@cornell.edu

Open Notebook

www.CitizenScience.org



USA npn
National Phenology Network

DataONE ... engaging diverse partners.

- Libraries and digital libraries
- Academic institutions
- Research networks
- NSF- and government-funded synthesis and supercomputer centers/networks
- Governmental organizations
- International organizations
- Data and metadata archives
- Professional societies
- NGOs
- Commercial sector



TeraGrid



THE UNIVERSITY of
NEW MEXICO



Contact

John W. Cobb, Ph.D

Computer Science and Mathematics

(865) 576-5439

cobbjw@ornl.gov

