Ecoinformatics and the Research Cycle

Deana Pennington
University of New Mexico
December 3, 2004

Grand Challenges in Ecology

- Alterations in biodiversity…exotic species, infectious disease
- Altered biogeochemical cycles at multiple spatial scales
- Climate change and variability, including ecosystem reponse to change
- Coupled human-natural ecosystems

Ecoinformatics

Tackling these question will require the use of all of the information available to us

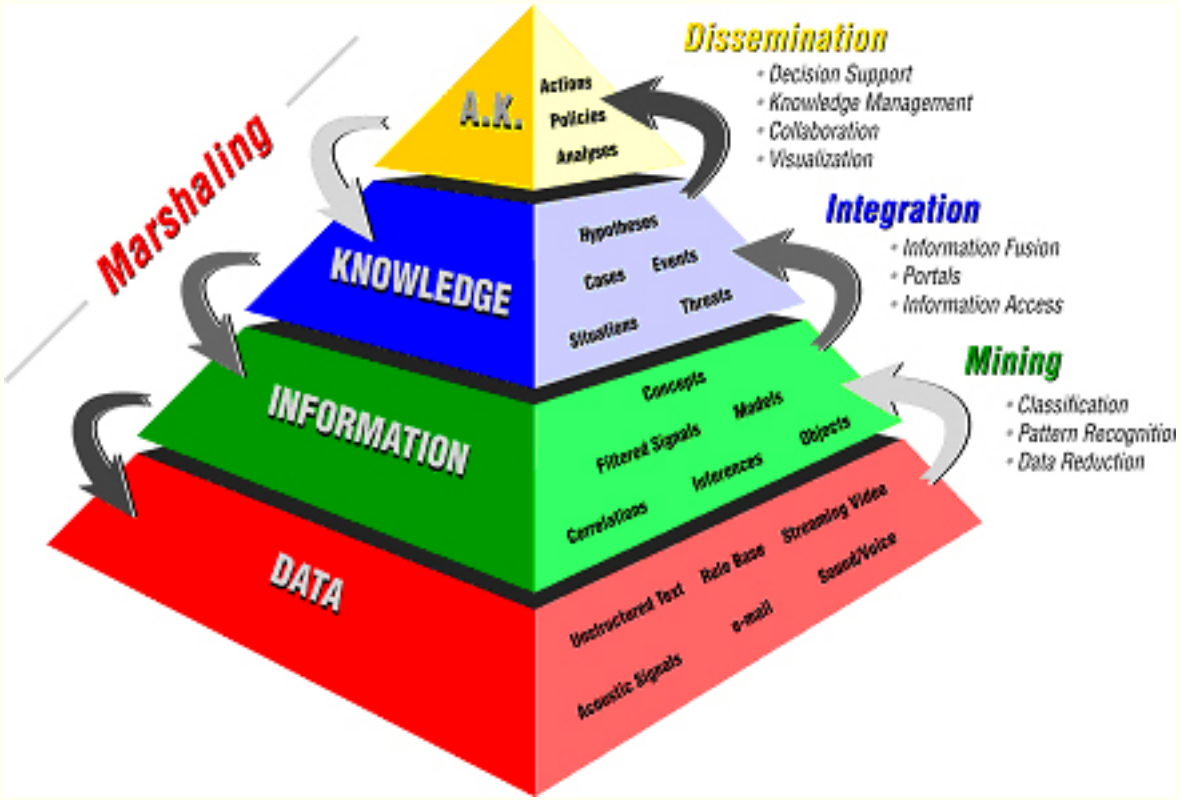
- 200+ years of data collection in US, 300+ globally
- Large and widely distributed data sets
- Data heterogeneity (text, Excel, GIS, DB, etc.)
- New data collection techniques: in situ sensor arrays
- Remotely-sensed imagery
- Scaling issues: space, time, levels (taxon)

Biodiversity and ecosystem informatics R&D has been identified as a critical national priority

–Computer-mediated collaboration

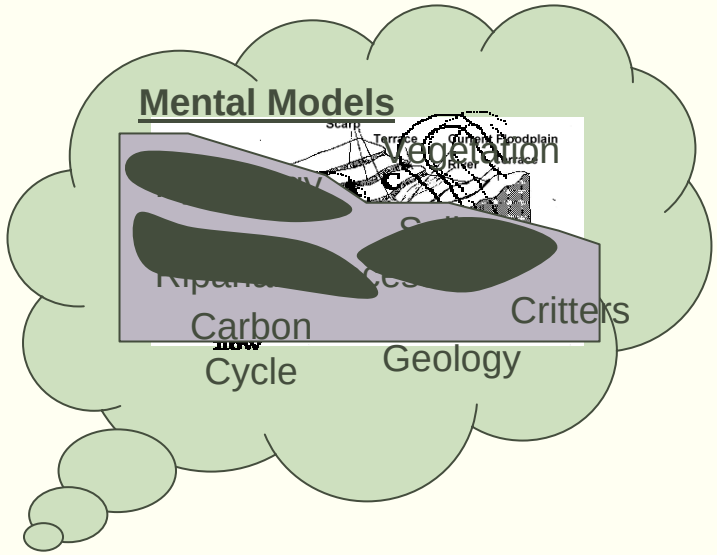
–New tools for synthetic understanding

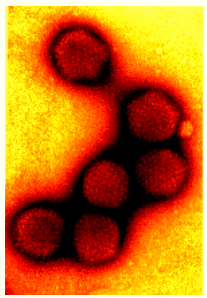
Knowledge Pyramid



From Los Alamos National Lab KDD Project

Technology-enabled science

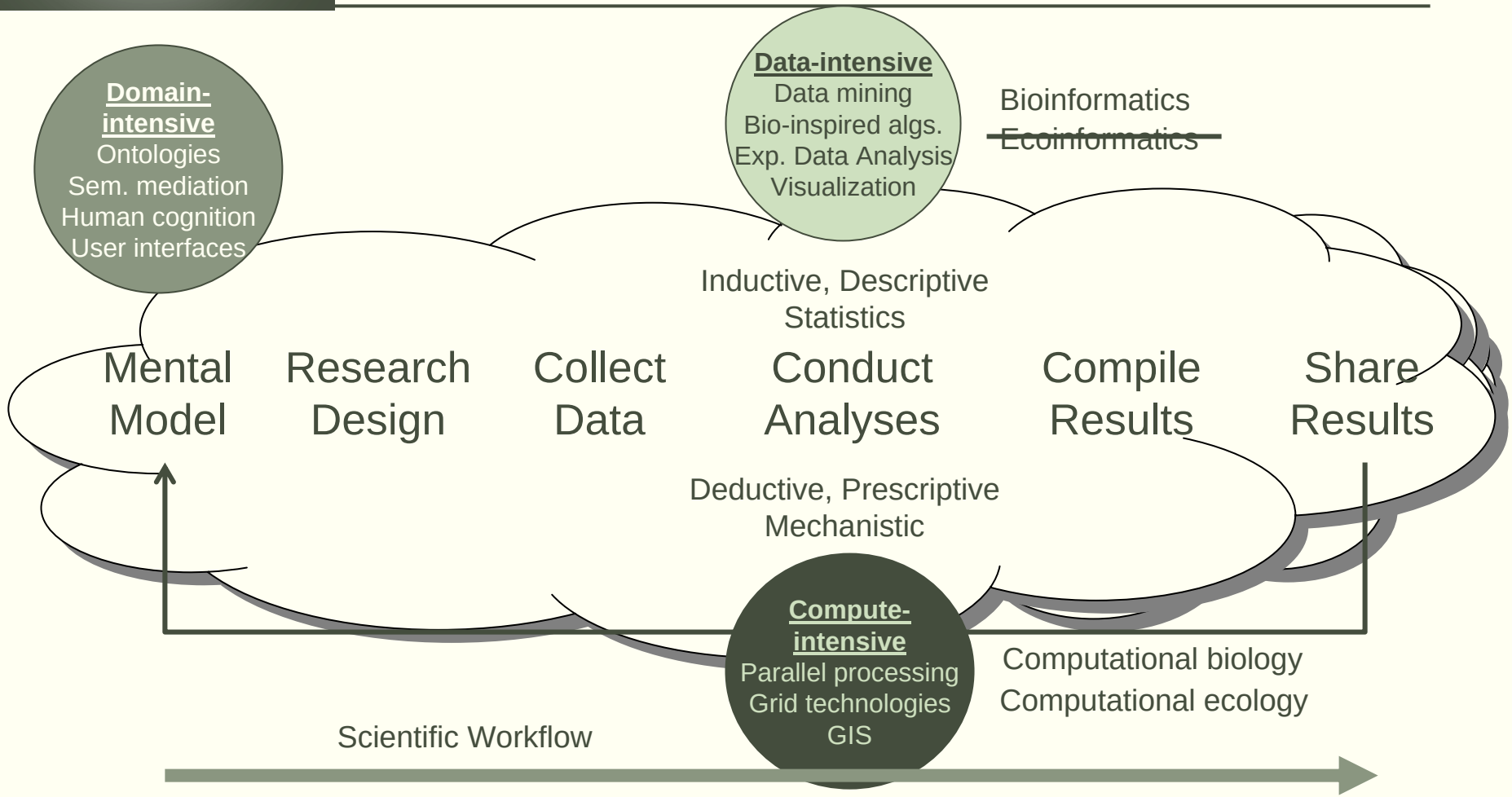




- Research design
- Data collection
- Analyses
- Publication venues

Data -> Actionable Knowledge requires context

Informatics and the Research Cycle



"Science Environment for (Ecological) Knowledge"

Collaborators

Partnership for Biodiversity Informatics (PBI)

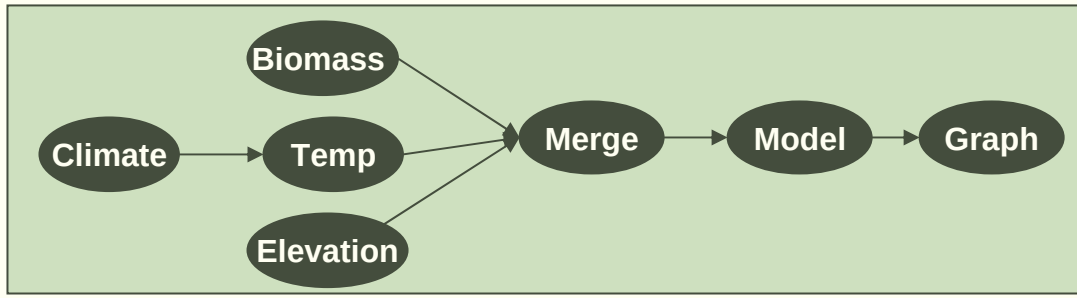
Productivity Example

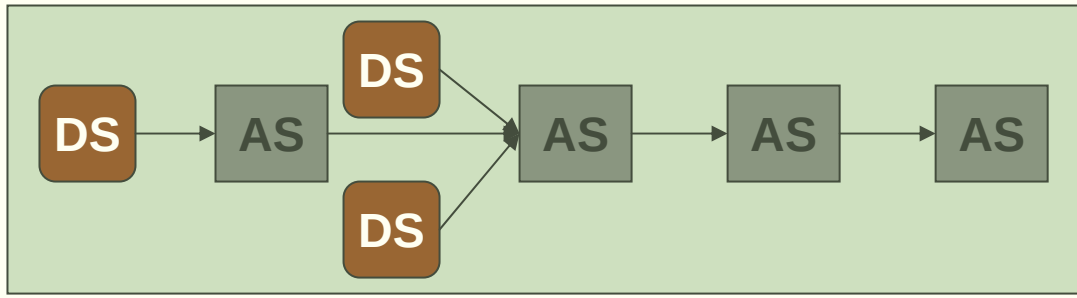
Mental Model

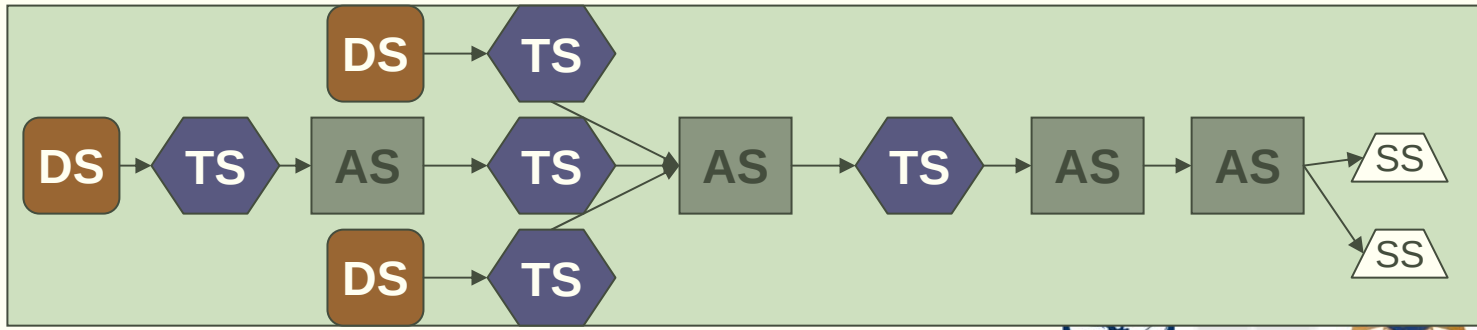
Conceptual Workflow

Abstract Workflow

Executable Workflow







C Concept

DS Data Step

AS Analysis Step

TS Transformation Step

SS System Step

SEEK

Semantic Mediation System
Kepler Workflow System

Technology-enabled

Mental Model

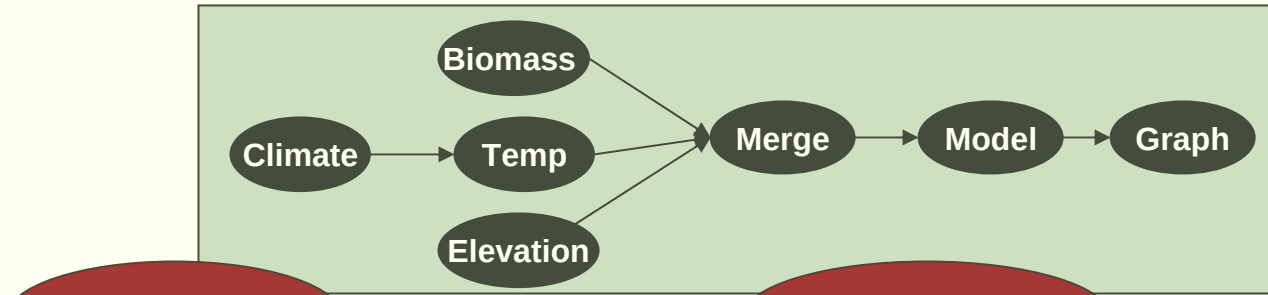
Ontologies

Biomass == *f* (Climate Elevation Et al.

Conceptual Workflow

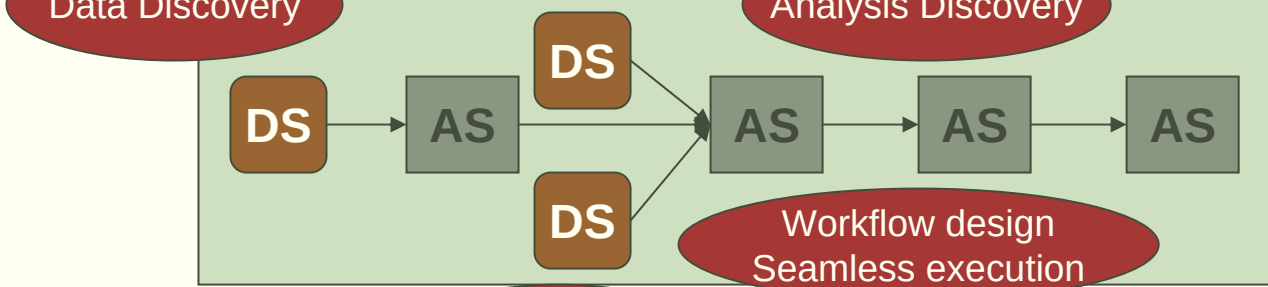
Abstract Workflow

Executable Workflow

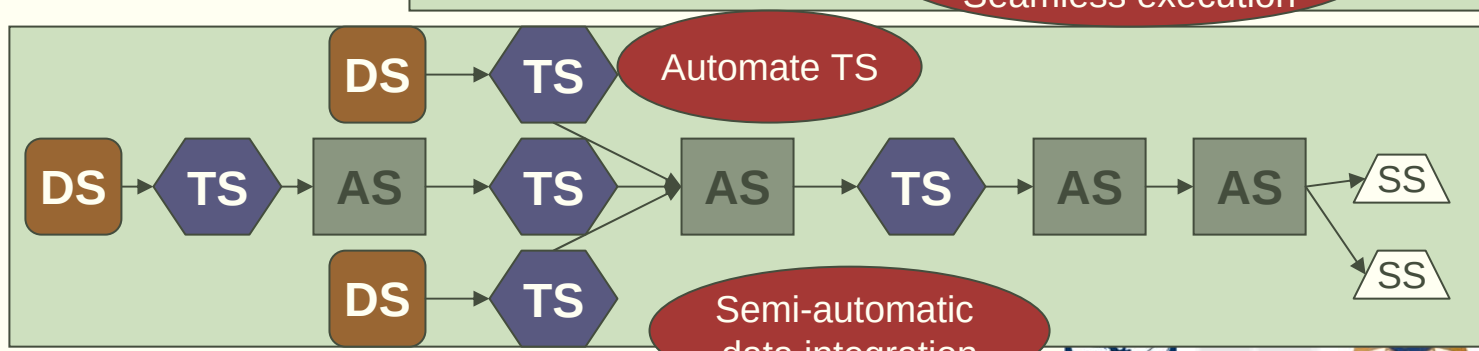


Data Discovery

Analysis Discovery



Workflow design
Seamless execution



Semi-automatic
data integration

C Concept

DS Data Step

AS Analysis Step

TS Transformation Step

SS System Step

KNB

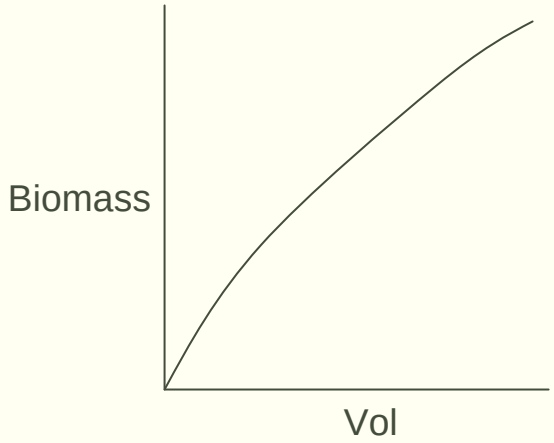
NPP Illustration

Protocol A
Non-destructive

Species A	30%	4 .0 cm
Species B	12%	1.2 cm
Species C	44%	6.3 cm

Protocol B
Destructive

323.1 g/cc



Seasonality: winter, spring, fall
Phenology: reproductive or vegetative
Scale

Data Integration

Location: Pine Mountain
Methods: Protocol A
Plot size: 5m x 5m

SPEC	COV99	HGT99	COV00	HGT00	COV01	HGT01
Veg1	30	4	35	4.2	35	4.4
veg2	12	1.2	11	0.9	10	0.3
veg3	44	6.3	46	6.6	51	6.9
…						

→ **Conversion Analysis** → YYMMDD,443.5,YYMMDD, 454.8,YYMMDD, 462.8, …

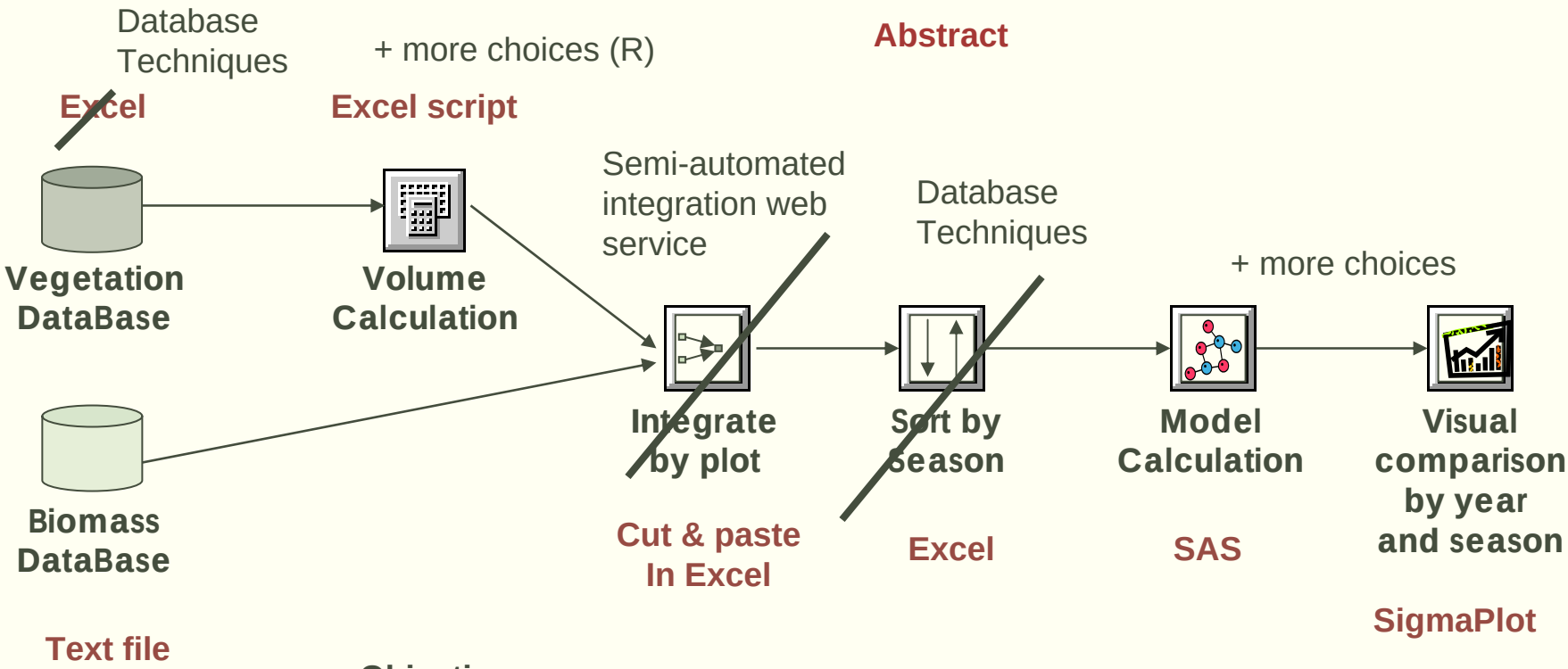
Location: White Mountain
Methods: Protocol B
Plot size: $1m^2$

TIME	BMASS
t1	323.1
t2	366.7
t3	383.2
…	

Data Integration:
Physical: file type
Logical: data organization
Semantic: space, time, methods

SITE	TIME	BMASS
PNM	t1	323.1
PNM	t2	366.7
WMT	t3	443.5
PNM	t4	383.2
WMT	t5	454.8
WMT	t6	462.8
…		

NPP Concept Workflow



Objective:
Many tasks
Many environments
Much cutting and pasting
➔ Single environment, automated integration

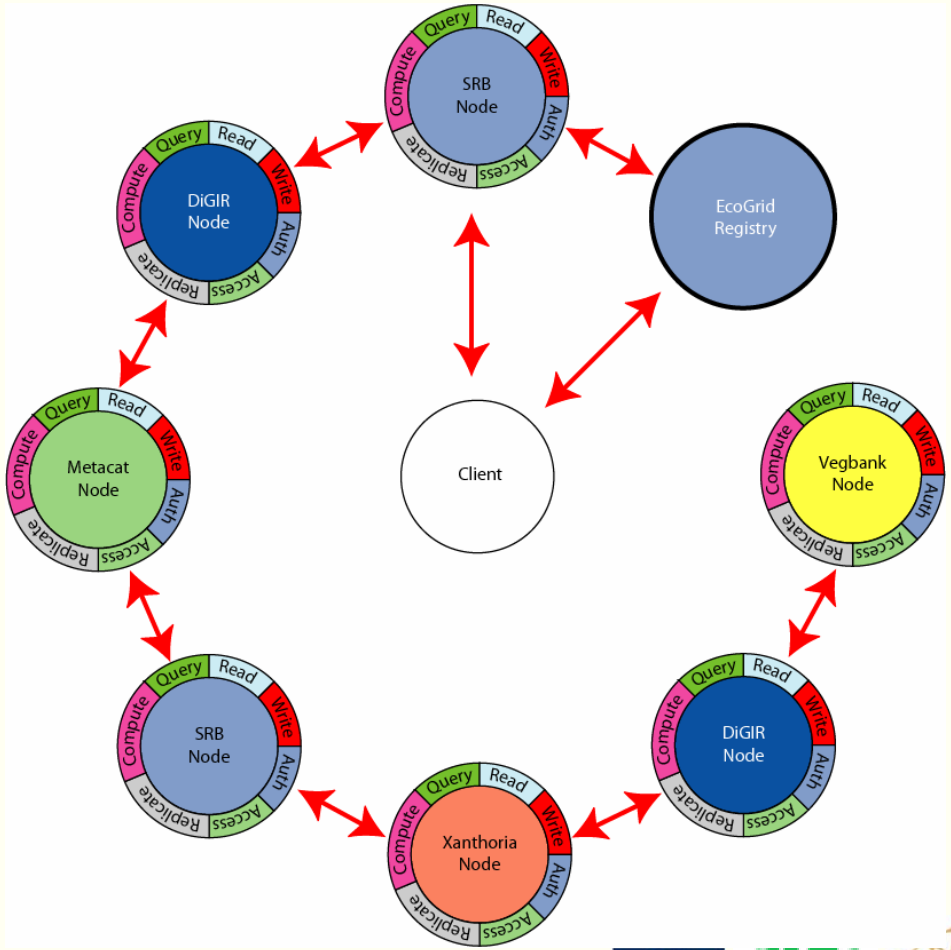
+ shared workflows

Automated Workflows

- Visual modeling
  - Single environment
  - Single platform
- Workflows:
  - Cross-platform
  - Cross-environment
  - Distributed data & analyses

Data nodes

Compute nodes

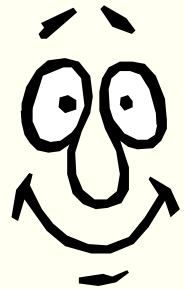


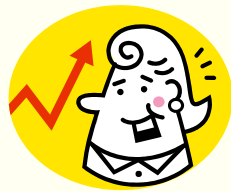
Metadata

- Decade-long effort to develop standards
- Ecological metadata language (EML)
- Database Registry/Morpho
- LTER minimum standards
- NSF requirements

Data Models, Database Design & QA/QC

Excellent database design
Excellent quality data
Excellent metadata



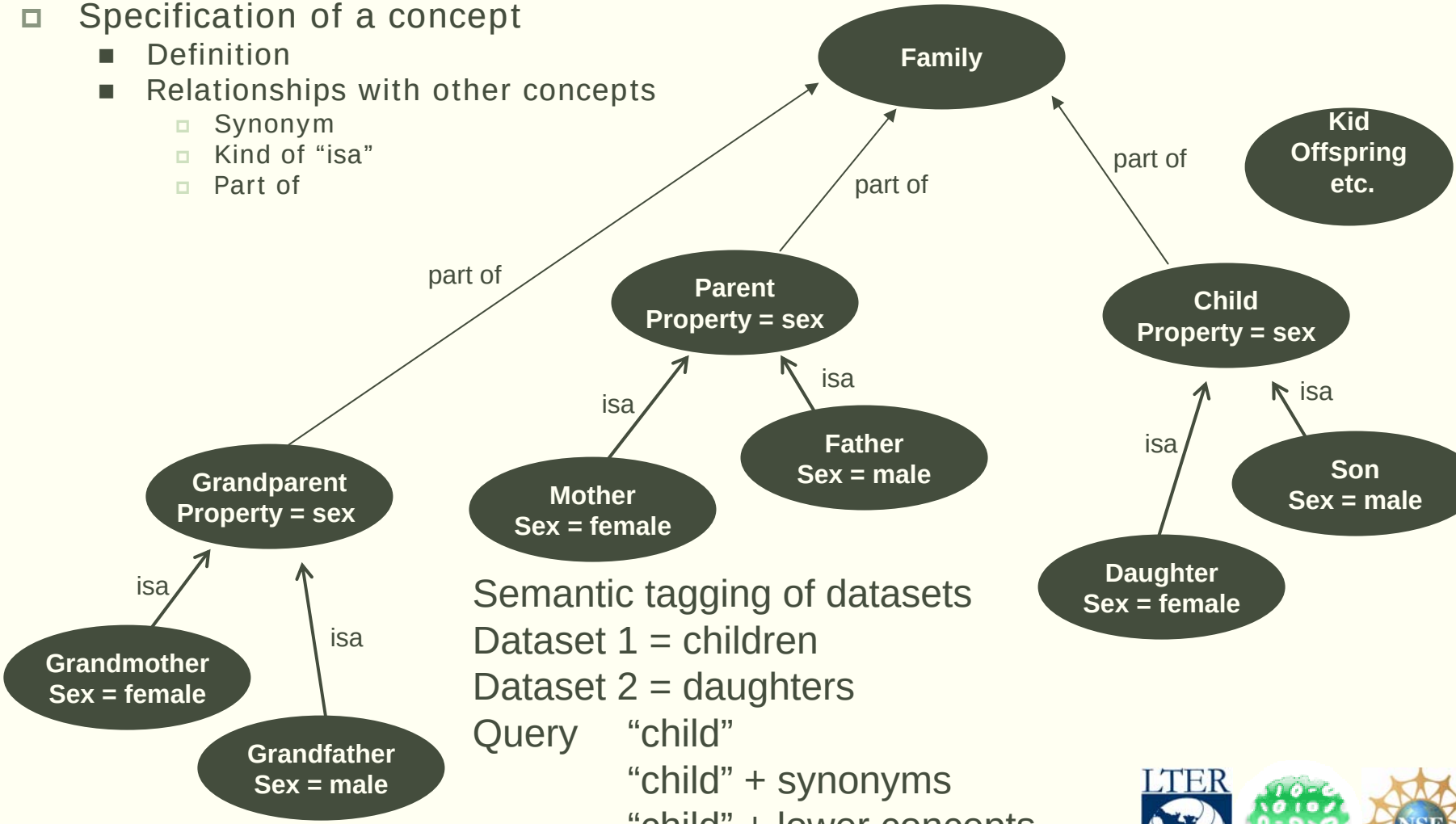


IT solutions

Poor database design
Poor quality data
Poor metadata

Ontologies

- Specification of a concept
  - Definition
  - Relationships with other concepts
    - Synonym
    - Kind of “isa”
    - Part of



Semantic tagging of datasets
Dataset 1 = children
Dataset 2 = daughters
Query “child”
“child” + synonyms
“child” + lower concepts

Ontologies: logical reasoning

- Specify all familial relationships (siblings, etc.)

Dataset 1

John is an instance of grandfather
Susan is his spouse
Jane is his daughter
Craig is his grandson

Dataset 2

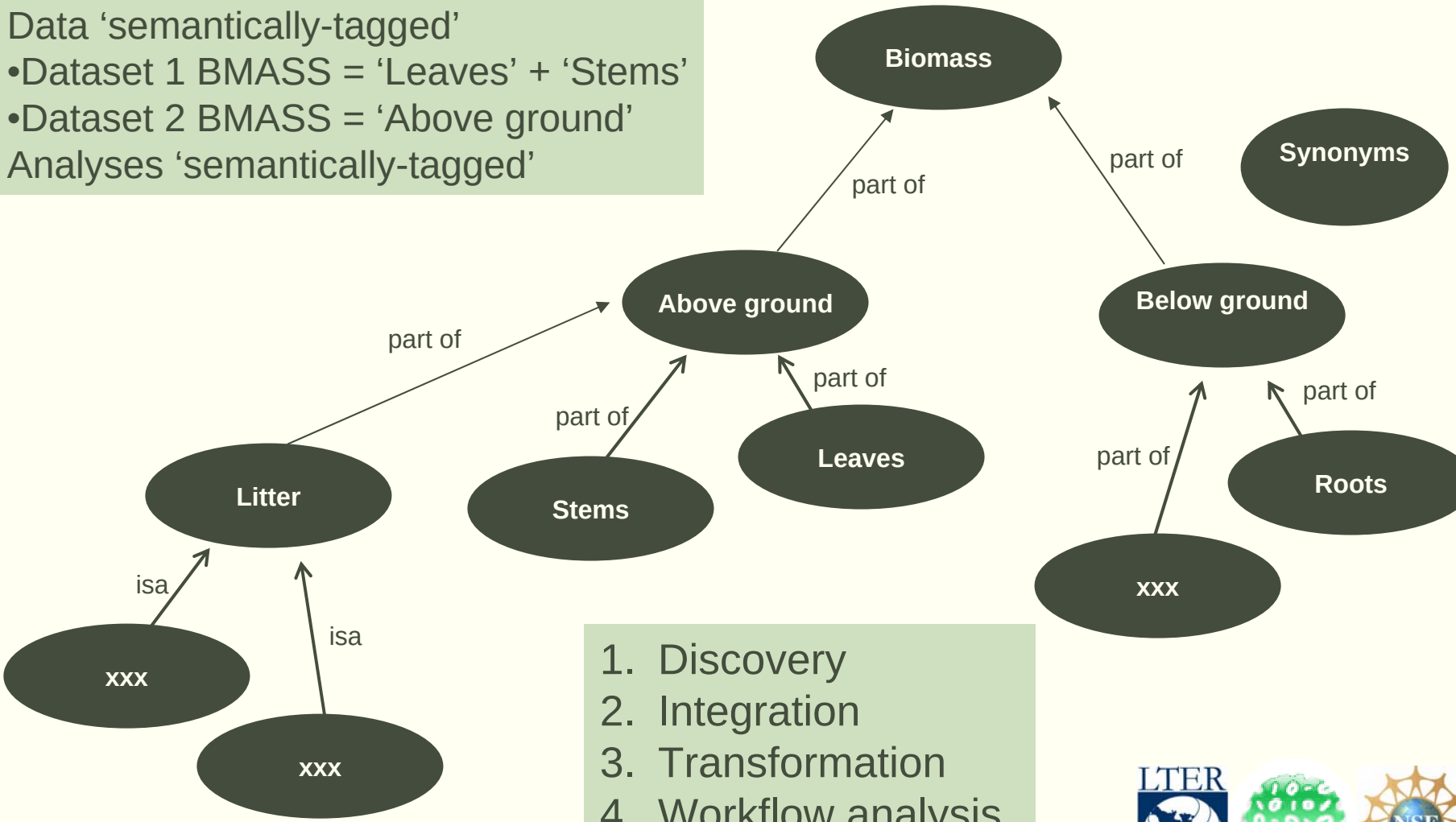
John is an instance of sibling
Laura is his sister
Connie is his parent
blah, blah, blah

Therefore, we can infer the relationship between Susan and Laura
… and everyone else in either dataset relative to John

Ontologies

Data ‘semantically-tagged’
- Dataset 1 BMASS = ‘Leaves’ + ‘Stems’
- Dataset 2 BMASS = ‘Above ground’

Analyses ‘semantically-tagged’



1. Discovery
2. Integration
3. Transformation
4. Workflow analysis

Ontologies

- Given these datasets, what models are on the system that will work?
- …are they *appropriate?*
- Given this model, search for data that will work, given these constraints

Ontology-Enabled Workflows

- Benefits
  - Complex queries for data and analyses
  - Sophisticated reasoning
  - Communication
  - Research design
- Shortcomings
  - Long-term domain challenge
  - Limits of use?

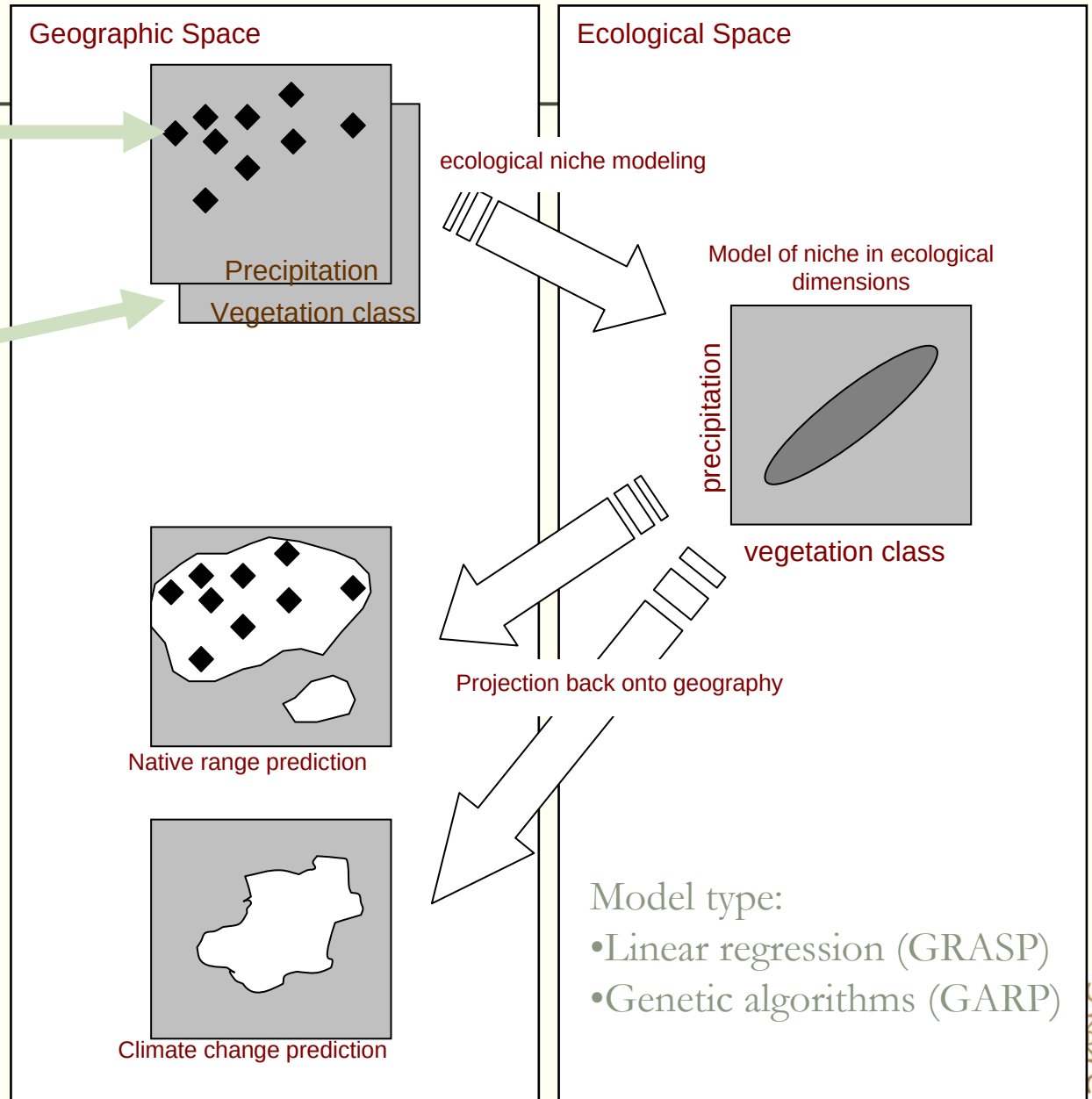
PARTNERSHIP for BIODIVERSITY INFORMATICS

Prototype project: Ecological Niche Modeling

Biodiversity information … e.g., data from museum specimens

Geospatial and remotely sensed data

Results used for integration with other data realms (e.g., human populations, public health, etc.)



Model type:
- Linear regression (GRASP)
- Genetic algorithms (GARP)

Modified from B. Michener

Mammal Project

Climate Change Analysis

2-3 dispersal scenarios

2 major evolutionary-computing algorithms (GA and NN)

21 GCM scenarios, including all IPCC scenarios

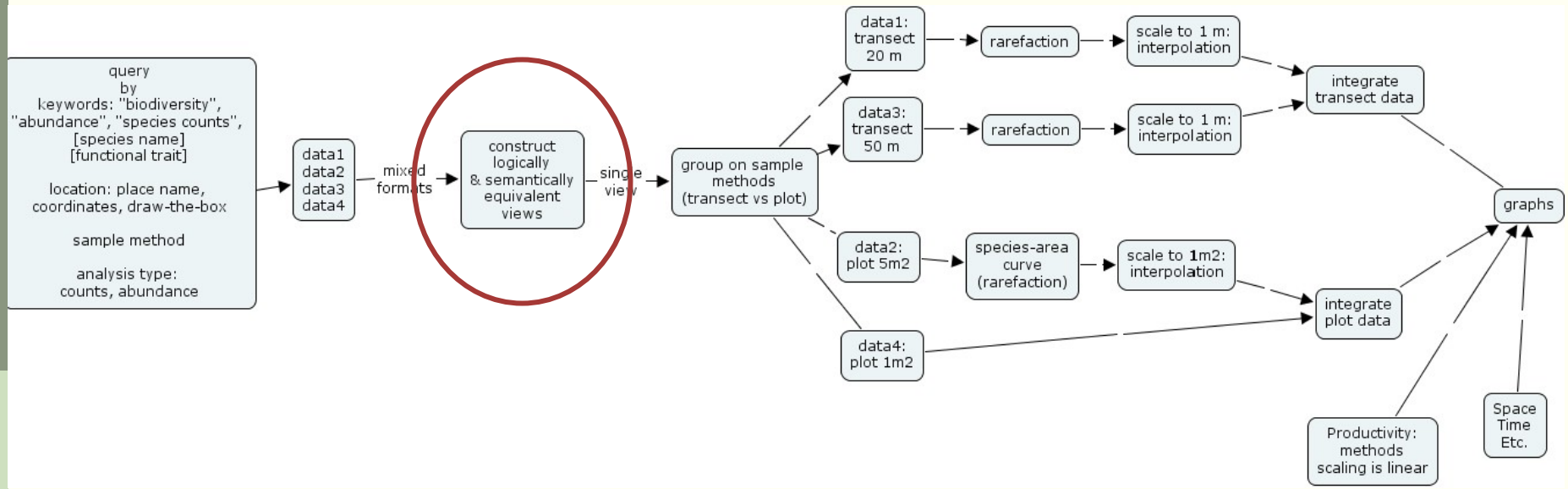
- 2000-3000 mammal species
- 100+ models/species/algorithm
- 2 algorithms
- 500,000 – 1,000,000 models
- Provide a hemisphere-wide view of mammal diversity
- Provide a massive comparison of CC implications

- Test large-scale implementation of Kepler/Grid functionality

Grassland biodiversity project

- Cross-site analysis of productivity, plant functional traits and plant species diversity
  - ARC, CDR, GCE, JOR, KBS, KNZ, NWT, SEV, SGS
- Biodiversity scaling issues
  - Former NCEAS working group

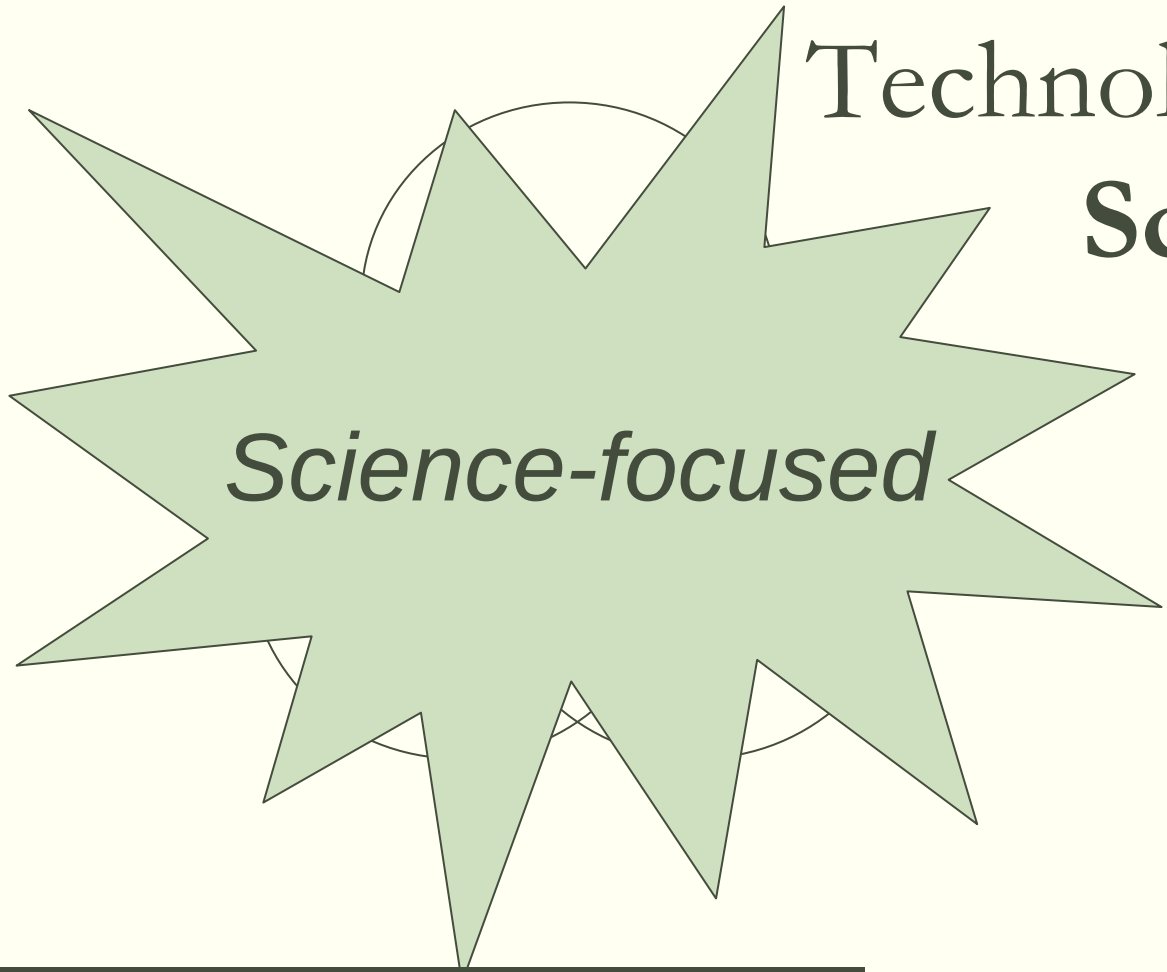
Biodiversity Integration Workflow



New, innovative analysis

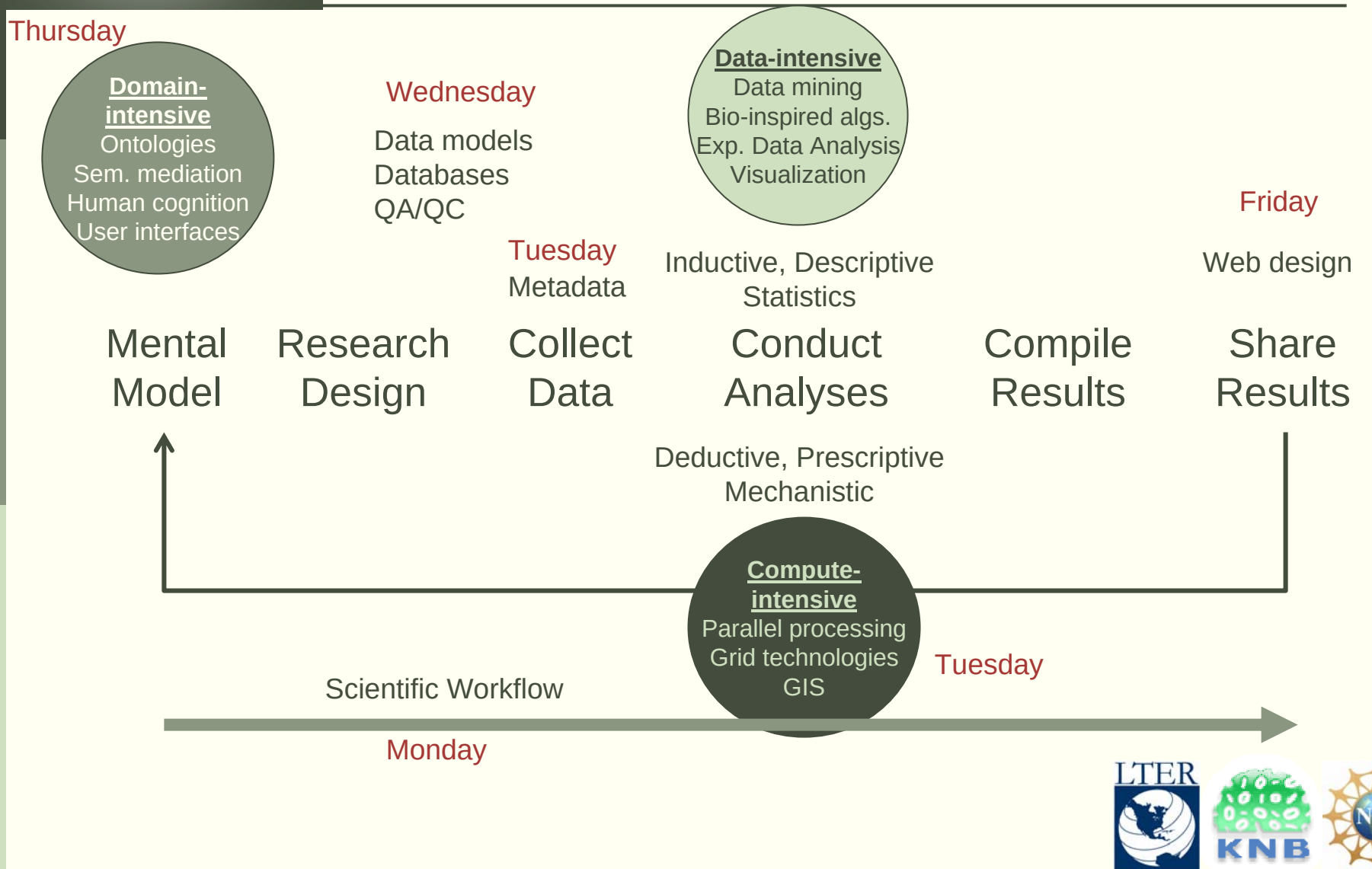
~~Technologic Systems for Scientists~~

Technology-enabled **Science**



A.K.A. “Take the red pill”

Informatics and the Research Cycle



Conclusion

"...Without advanced information processing, it would take decades to compile and analyze the incredible amounts of information that are produced by many of these instruments."

-Dr. Rita Colwell, Director NSF, 1998

Acknowledgements

This material is based upon work supported by the National Science Foundation under awards 0225676 for SEEK and 0225673 (AWSFL008-DS3) for GEON and by the Department of Energy under Contract No. DE-FC02-01ER25486 for SciDAC/SDM and by DARPA under Contract No. F33615-00-C-1703 for Ptolemy. Any opinions, findings and conclusions or recomendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation (NSF).

The National Center for Ecological Analysis and Synthesis, a Center funded by NSF (Grant Number 0072909), the University of California, and the UC Santa Barbara campus.

The Andrew W. Mellon Foundation.

PBI Collaborators: NCEAS, University of New Mexico (Long Term Ecological Research Network Office), San Diego Supercomputer Center, University of Kansas (Center for Biodiversity Research)

Kepler contributors: SEEK, Ptolemy II, SDM/SciDAC, GEON