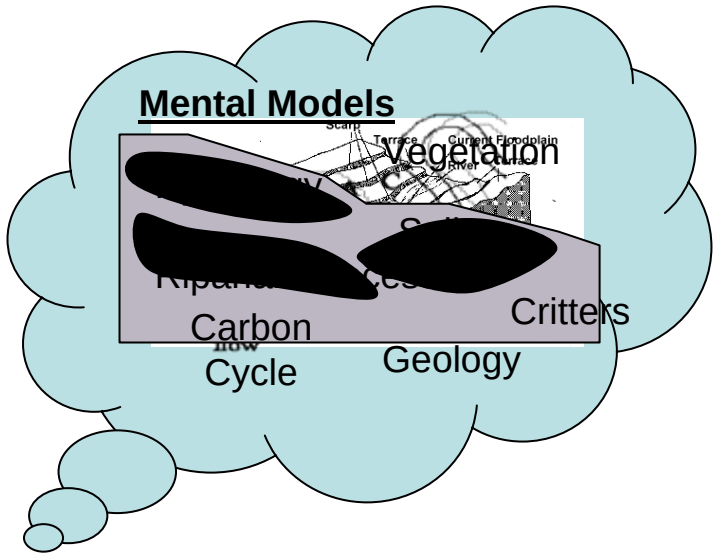
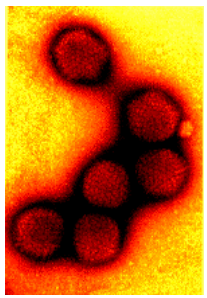
Research Design for Collaborative Computational Approaches and Scientific Workflows

Deana Pennington

January 9, 2006

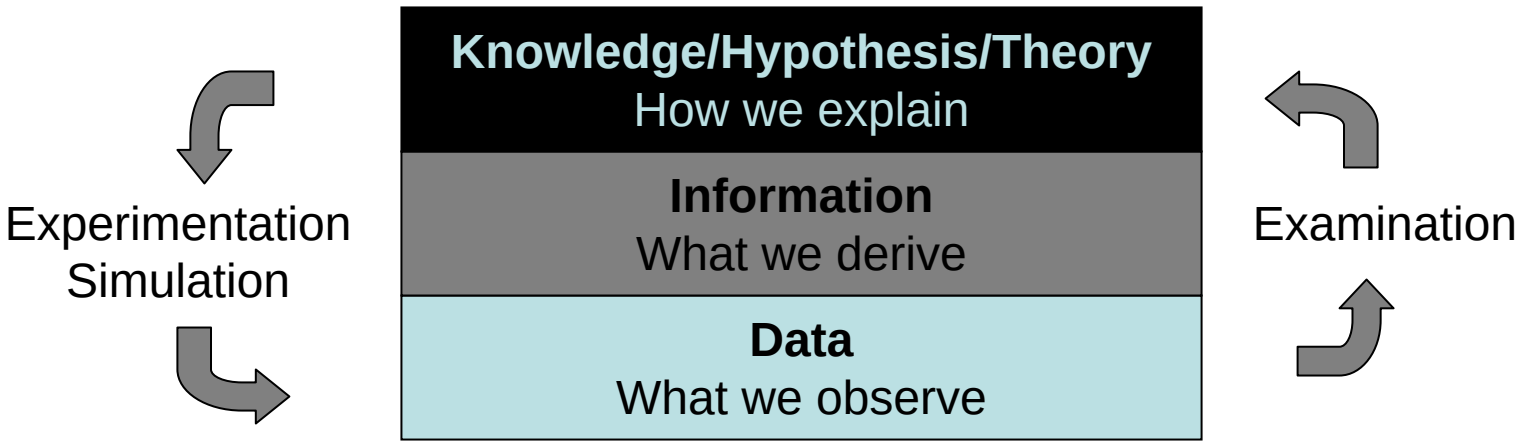
Conceptual Models





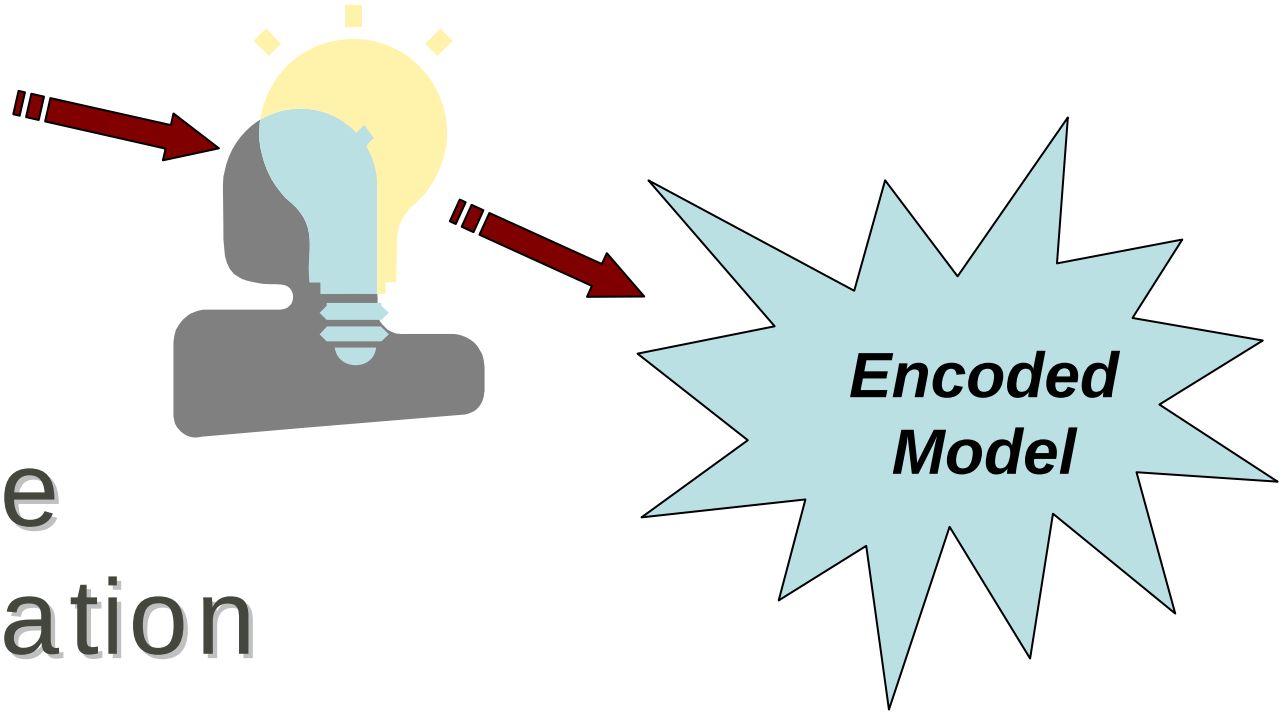
- Research design
- Data collection
- Analyses

Science in a nutshell



Data and information link to a conceptual model of reality.

SEEK
Andrews LTER



Knowledge Representation

- An approach for encoding a conceptual model
- For the purpose of automated, intelligent reasoning

NSF

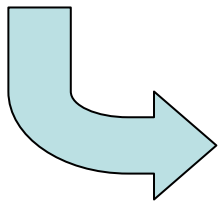
Types of Encoding

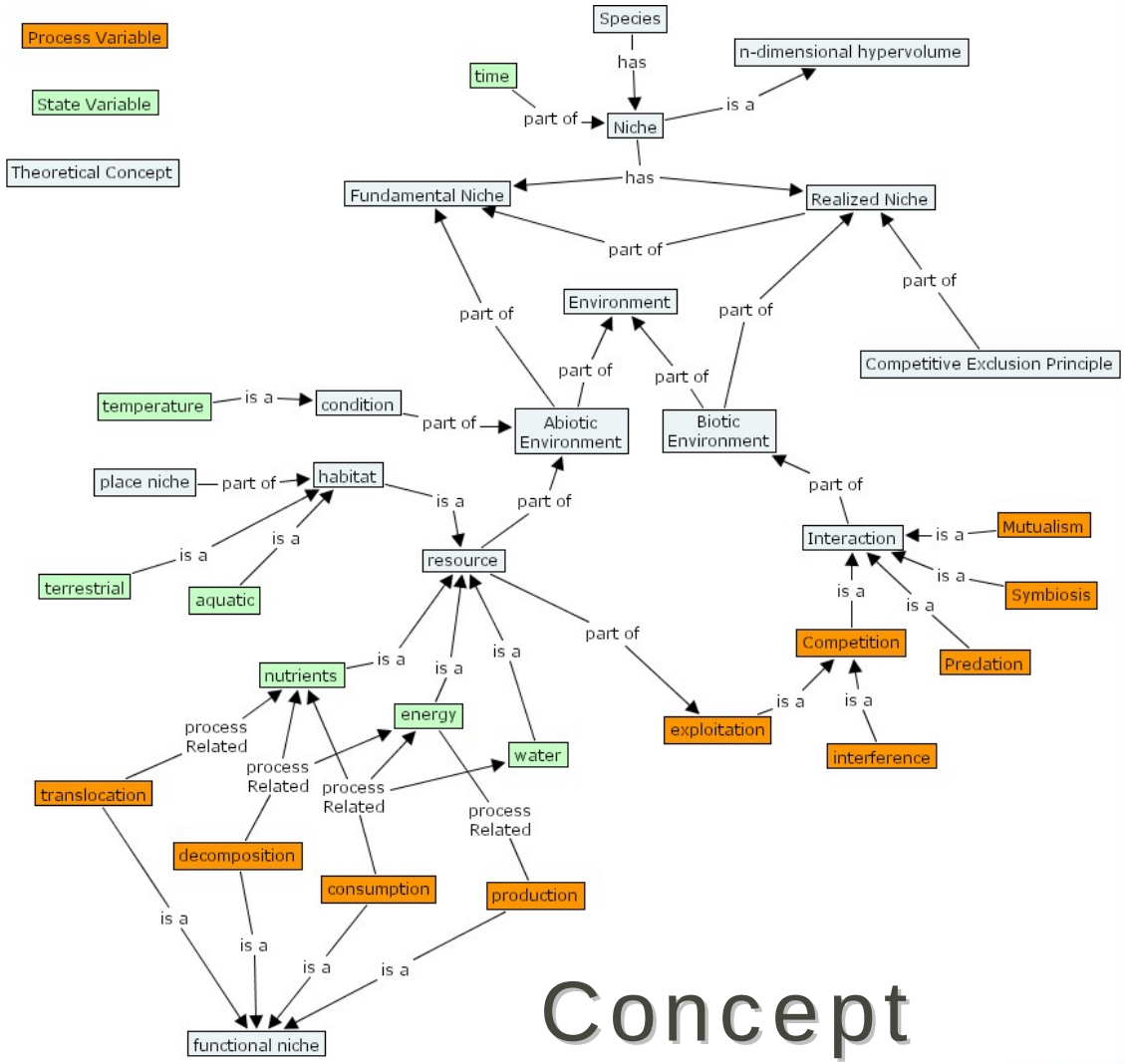
Expressiveness		Reasoning Capability
Natural language	**Semantic Networks**	**Ontologies**

All encodings are necessarily imprecise and inaccurate, because the only completely precise and accurate representation of reality is reality itself.

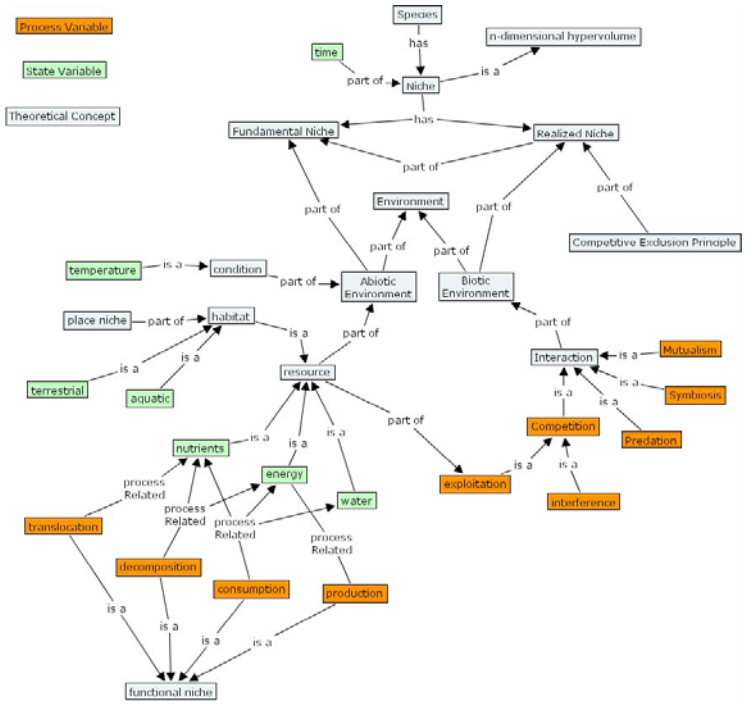
The “best” encoding depends on the objective of the user…

Semantic Networks (formal)





Concept Maps (informal)

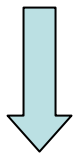


Concept models form the basis for other models in informatics

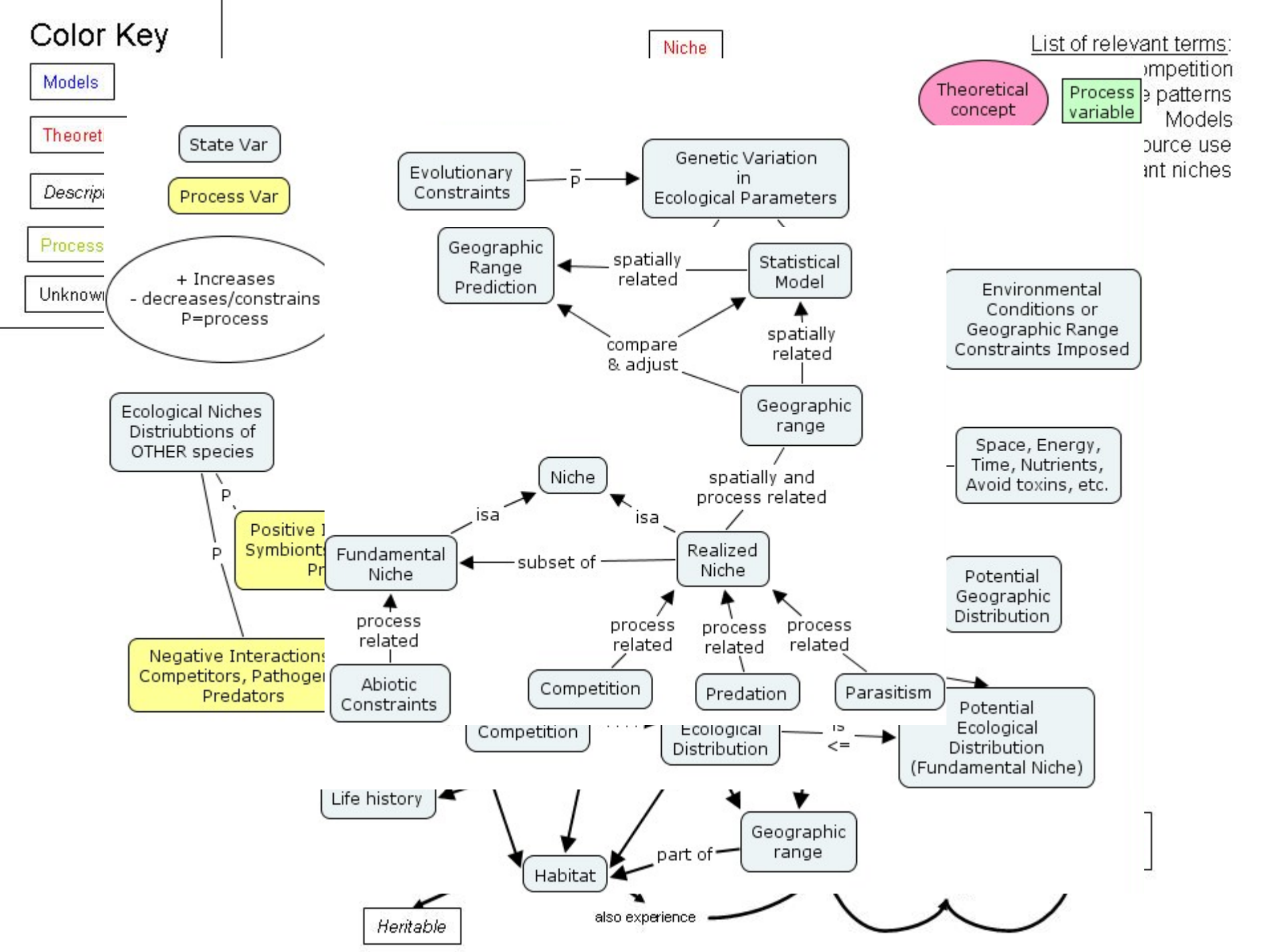
Data Model => Database design (Wednesday)

Ontology design (Friday)

Process Flow Diagrams



Conceptual Scientific Workflow


Color Key
Models
Theoret
Descripi
Process
Unknow
Niche
List of relevant terms:
mpetition
e patterns
Models
ource use
ant niches
Theoretical concept
Process variable
State Var
Process Var
+ Increases
- decreases/constrains
P=process
Evolutionary Constraints
P
Genetic Variation in Ecological Parameters
Geographic Range Prediction
spatially related
Statistical Model
compare & adjust
spatially related
Environmental Conditions or Geographic Range Constraints Imposed
Geographic range
Ecological Niches Distriubtions of OTHER species
Space, Energy, Time, Nutrients, Avoid toxins, etc.
Niche
spatially and process related
isa
isa
Positive I
Symbiont
Pr
Fundamental Niche
subset of
Realized Niche
Potential Geographic Distribution
process related
process related
process related
process related
Negative Interactions
Competitors, Pathoge
Predators
Abiotic Constraints
Competition
Predation
Parasitism
Competition
Ecological Distribution
<=
Potential Ecological Distribution (Fundamental Niche)
Life history
Geographic range
part of
Habitat
Heritable
also experience

Benefits

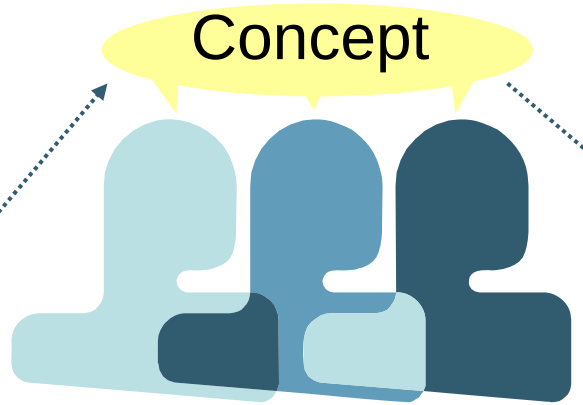
Learning:

- Making your own world view explicit
- Understanding others' perspectives
- Highlights common misconceptions
  - E.g. Is the realized niche necessarily more restricted than the fundamental niche?

Scientific discourse:

- Key decision points
  - E.g. Can a biotic property contribute to definition of the fundamental niche?
- Dimensionality
  - State space, functional interactions
  - Spatial interactions? Geography
  - Temporal interactions? Evolution

Ontological Commitment... = Scientific Agreement

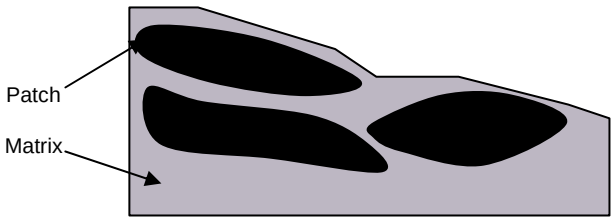


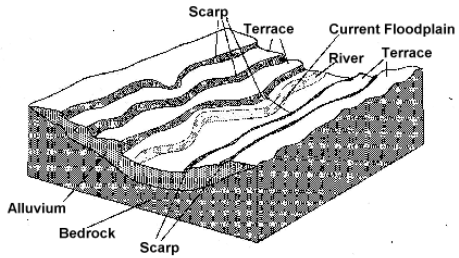
"*River*" *stands for*

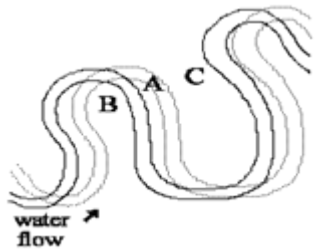
Ontologies:

- **Commit to a model (definition) of a domain**
- **Explicitly state assumptions concerning the model**
- **Have a wide scope (are general)**

Ontological Diversity… = Scientific Disciplinary Context

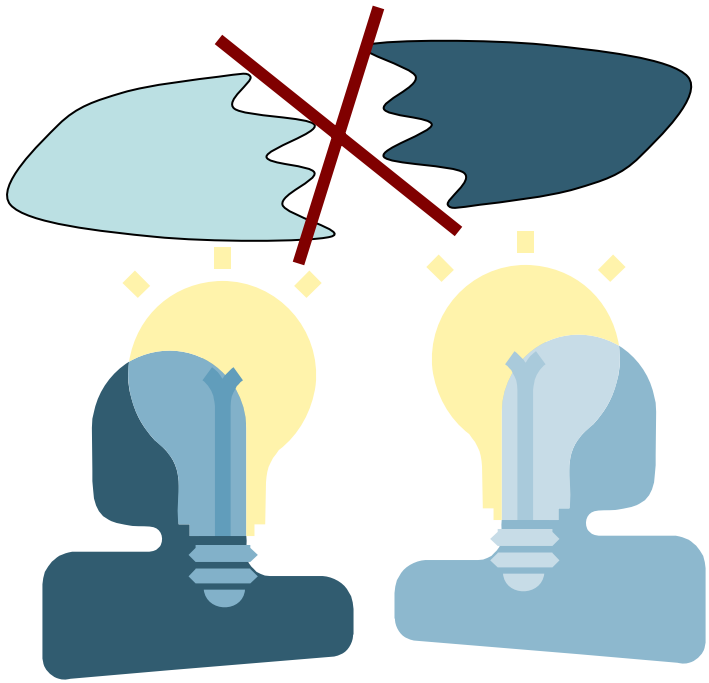






Ontologies make different perspectives explicit

Ontological Clash... = Scientific Debate!

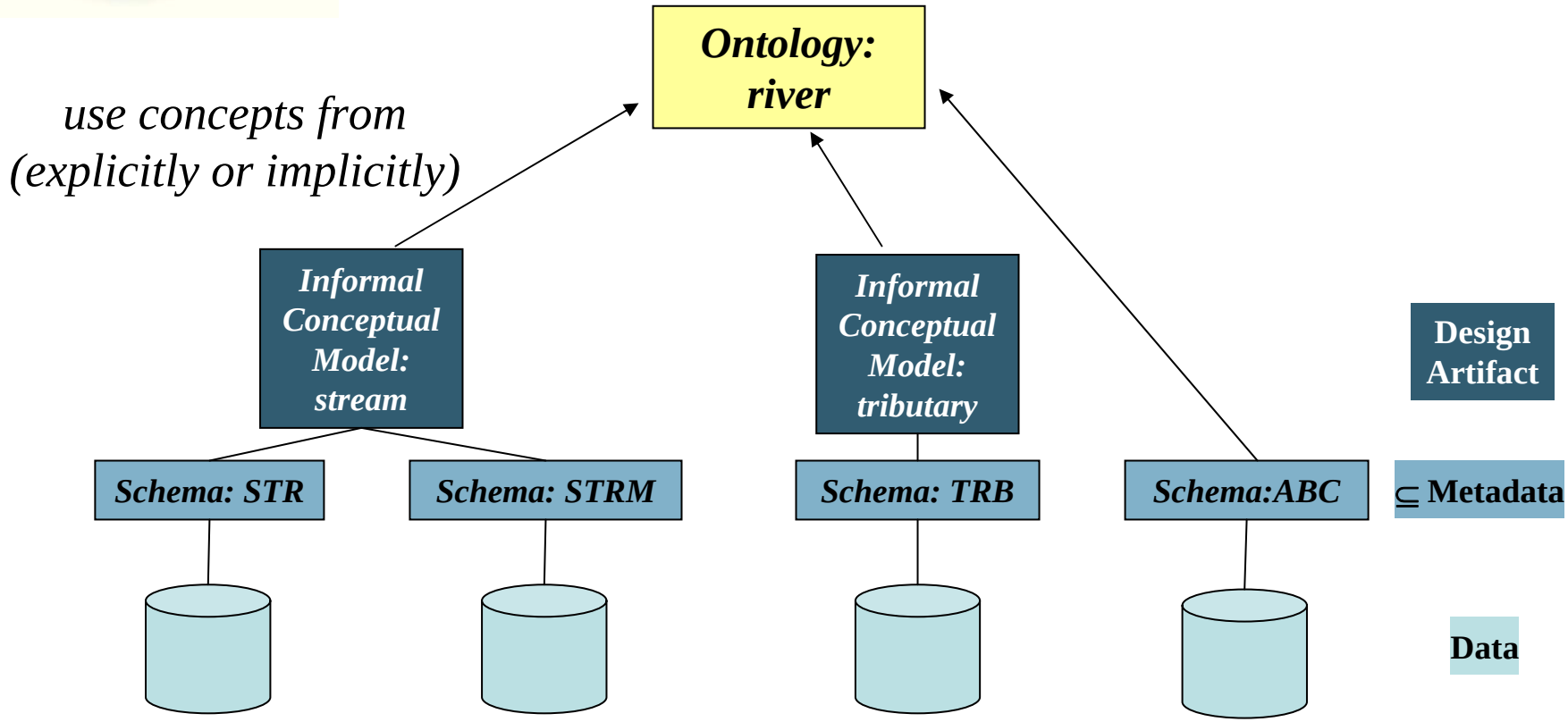


Ontologies can help clarify issues that go beyond terminological and perspective differences and make those issues explicit

In general

- The more explicit you can be about the conceptual framework that your data and analyses are linked to, the better
- Informal methods (publications, diagrams) work for exchange of information with a small community of colleagues
- Formal methods are required for automation
- The effort involved in constructing formal knowledge bases for automation have additional uses in education and dialogue

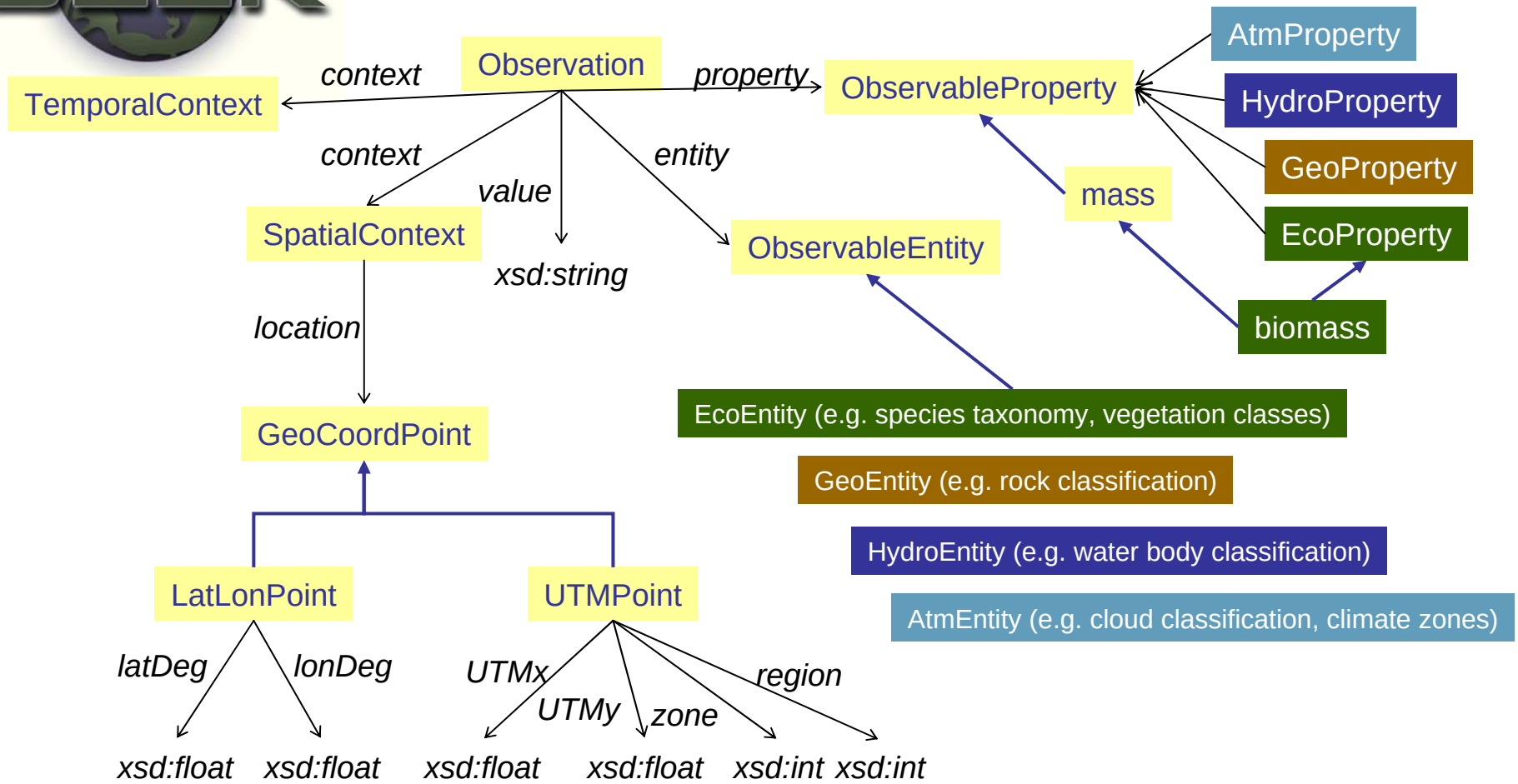
Ontological Commitment...



An ontology can then be used as a standard that supports exchange and integration of heterogeneous data sources and applications

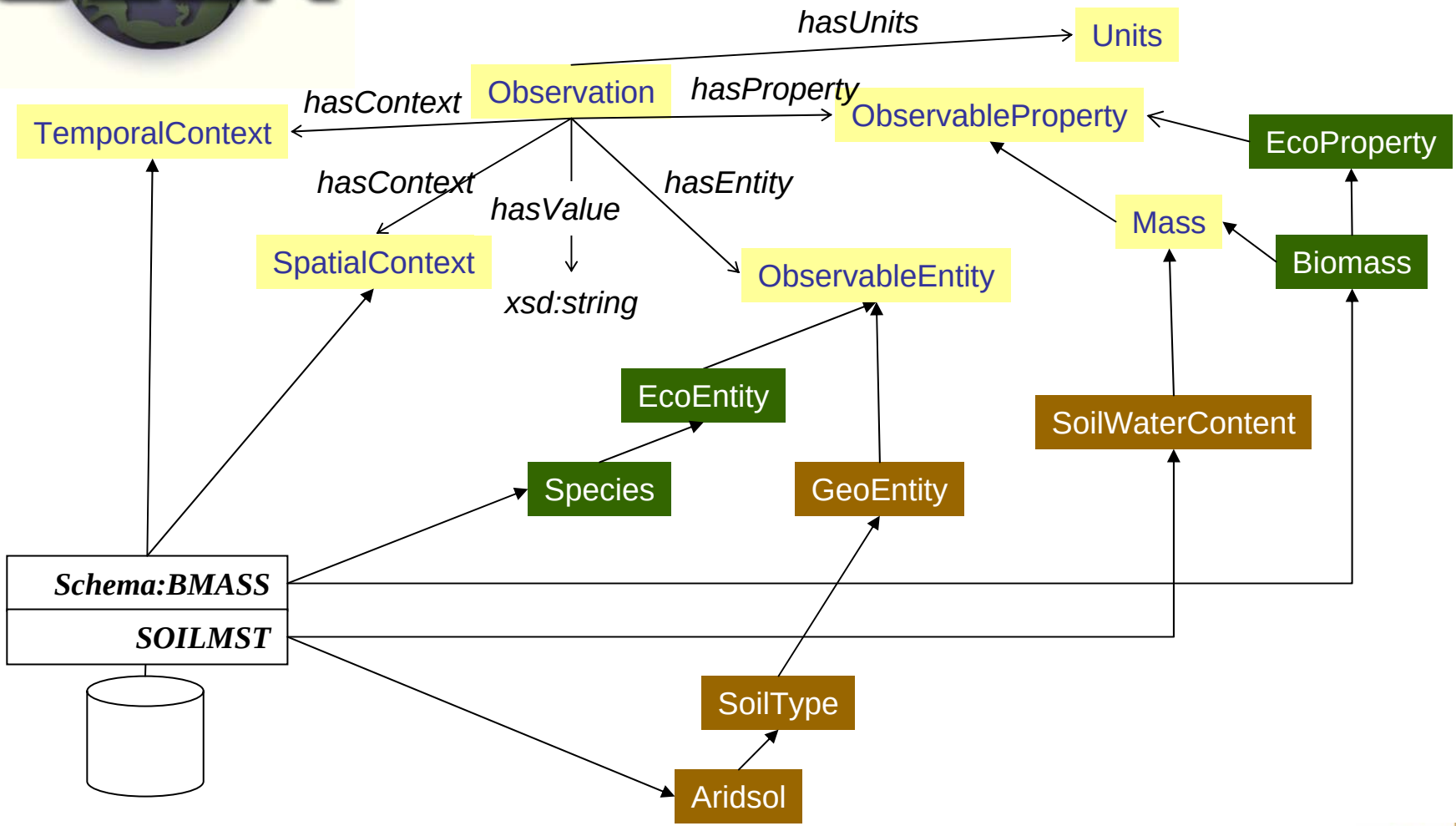
Shawn Bowers, 2005

Observation Ontologies



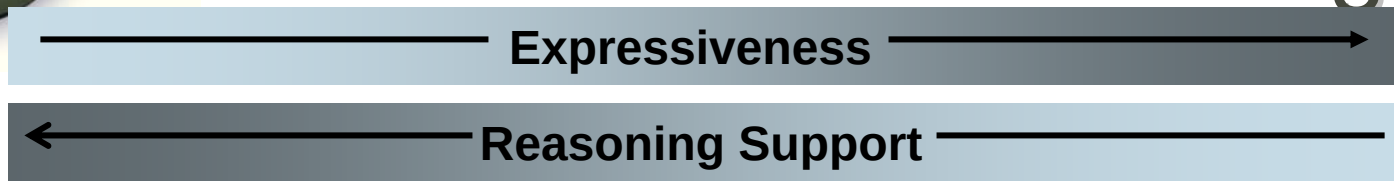
Some ontologies could be common to multiple disciplines; others will be discipline specific

Observation Ontologies



Data or resources developed within any given discipline should be able to be linked to ontologies from other disciplines, enabling cross-disciplinary sharing of resources

Language of Ontologies



OWL Full

OWL DL

OWL Lite

RDF Schema

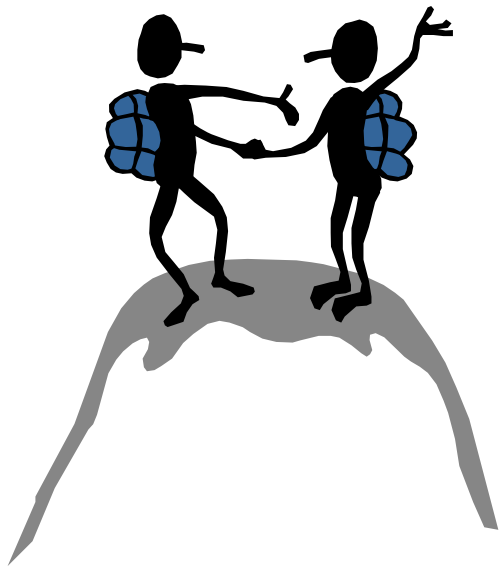
RDF

RDF = Resource Description Framework
OWL = Web Ontology Language

Standard languages are emerging that exploit Extensible Markup Language (XML). Different formal languages provide different benefits in terms of expressiveness and reasoning capabilities.

The “best” ontology language depends on the objectives of the user

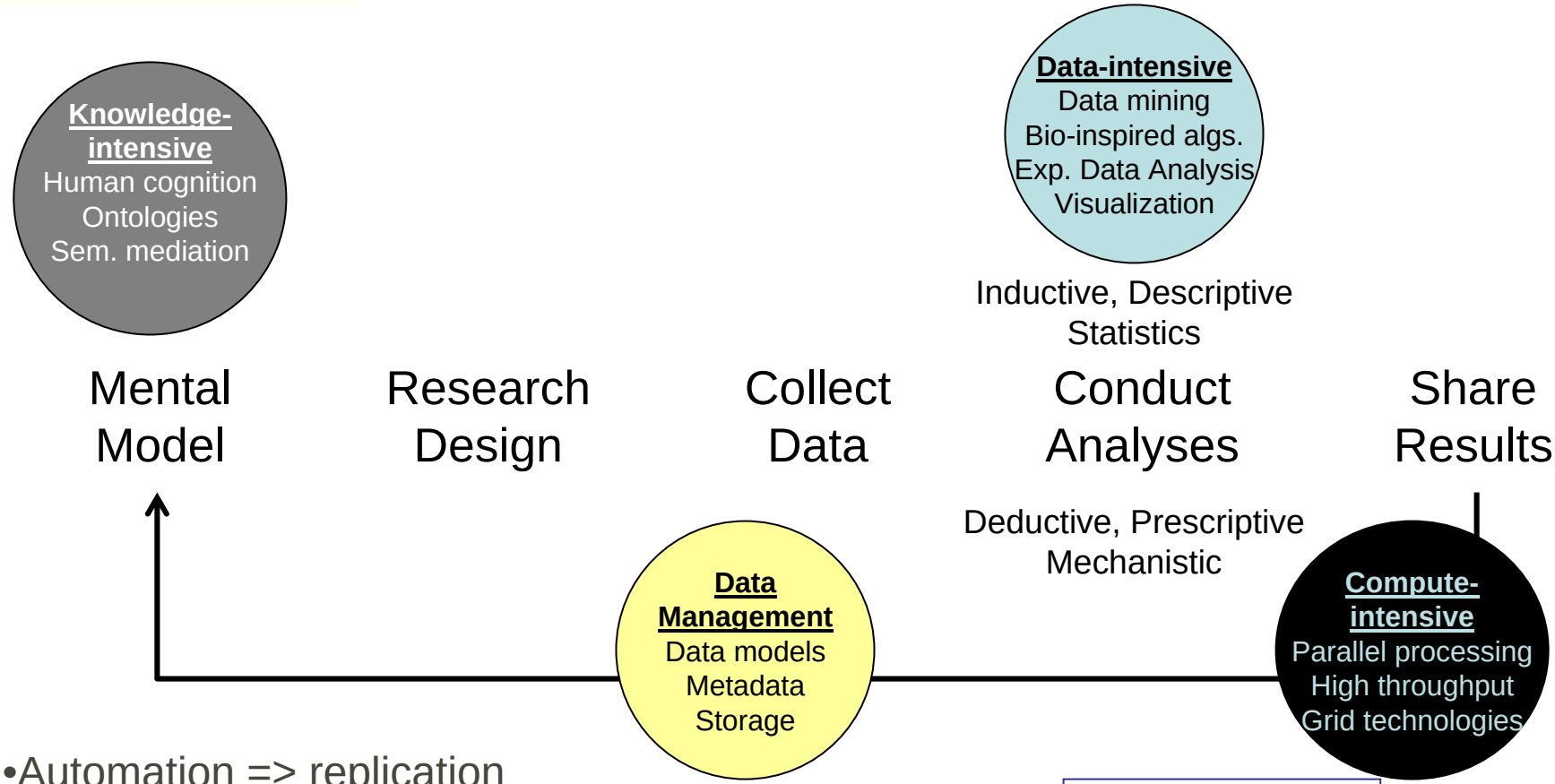
Ontology Development



Interdisciplinary effort between:

- Domain scientists who *have the knowledge*,
- Knowledge engineers who *formally express that knowledge*, and
- Computer scientists who *reason across the knowledge* from various systems

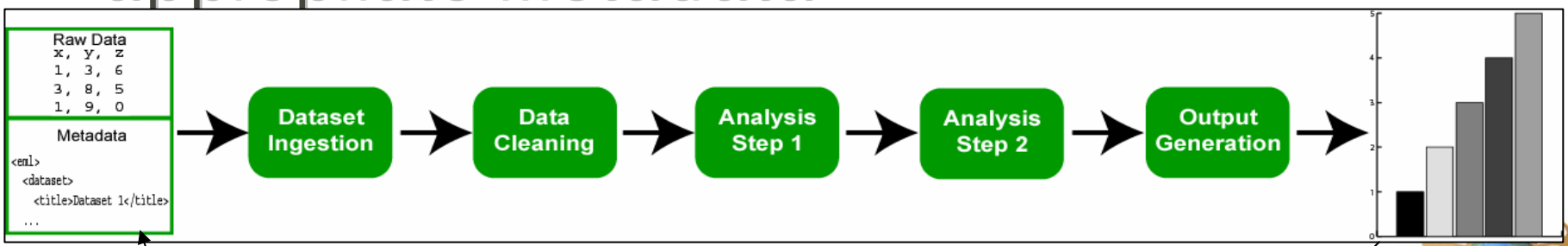
Informatics and the Research Cycle



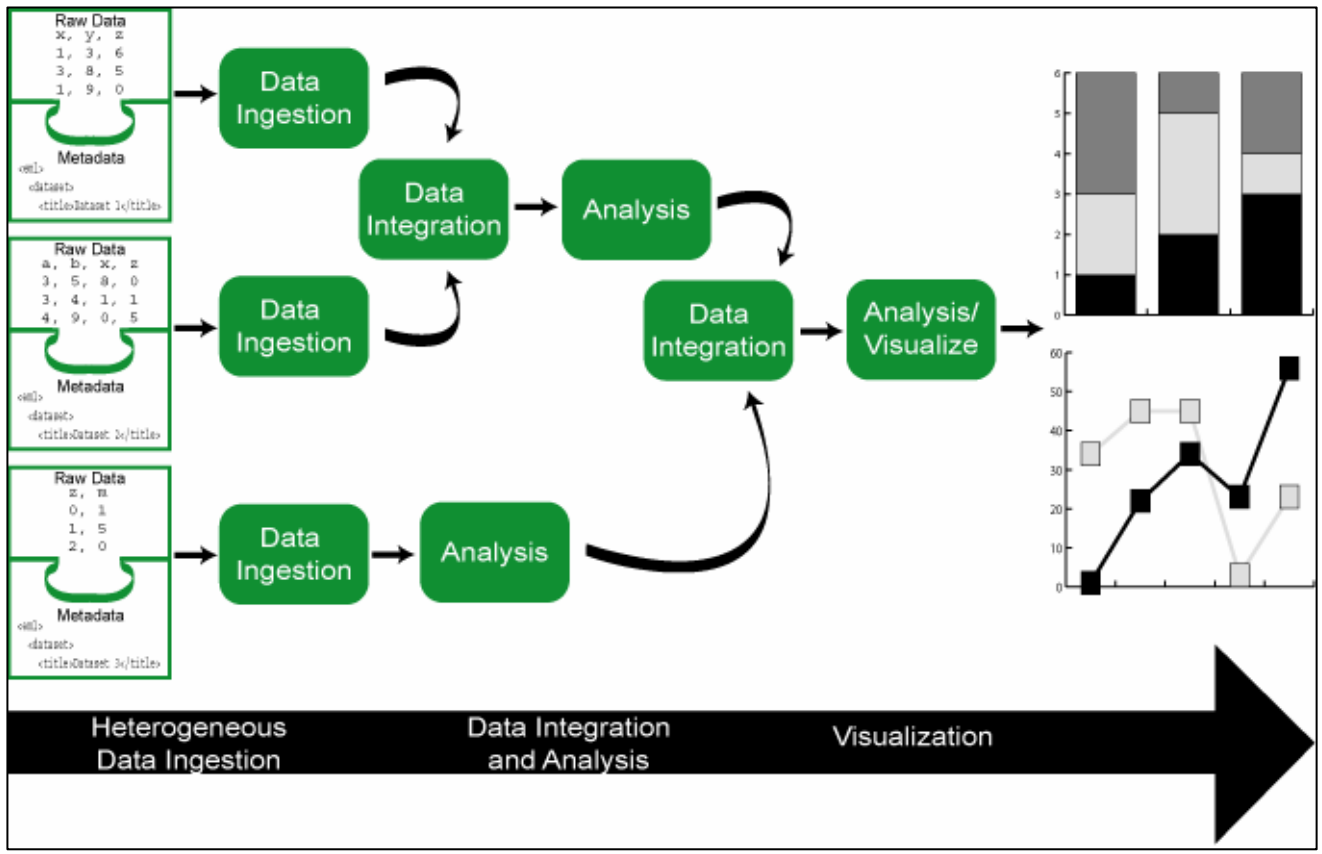
- Automation => replication
- Access to distributed resources
- Reusability & sharing
- ***Linked with knowledge-intensive approaches***

Scientific Workflows

- Model the way scientists work with their data now
  - Mentally coordinate export and import of data among software systems
- Workflows emphasize data flow
- Metadata-driven data ingestion
- Output generation includes creating appropriate metadata



- Not linear
- Involve multiple data sets
- Involve multiple analytical steps



Automated Workflows

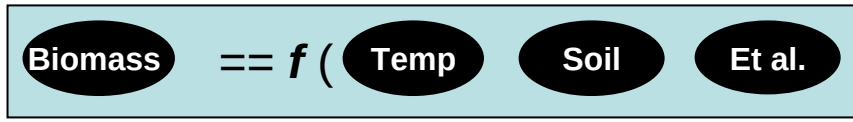
- Scripts
- Visual modeling environment

} Single platform, Single environment

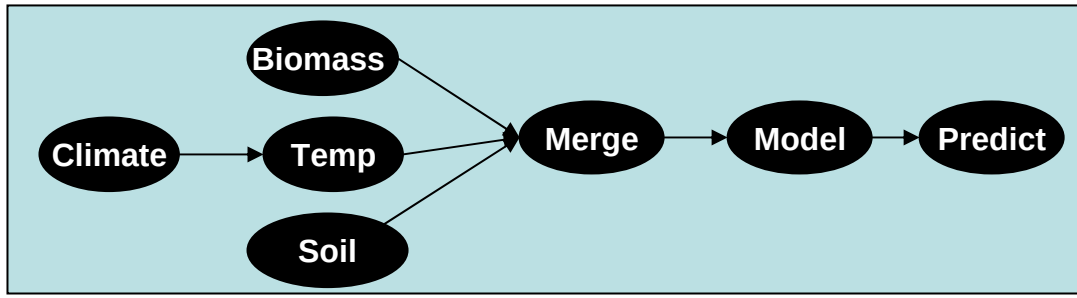
- Workflows:
  - Cross-platform
  - Cross-environment
  - Distributed data & analyses

Productivity Example

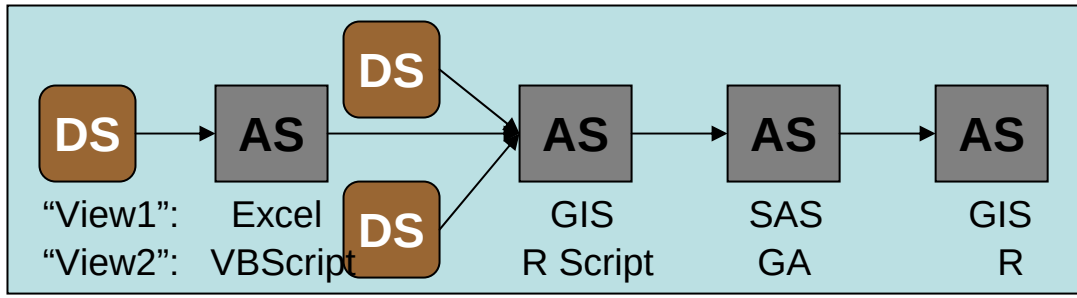
Mental Model



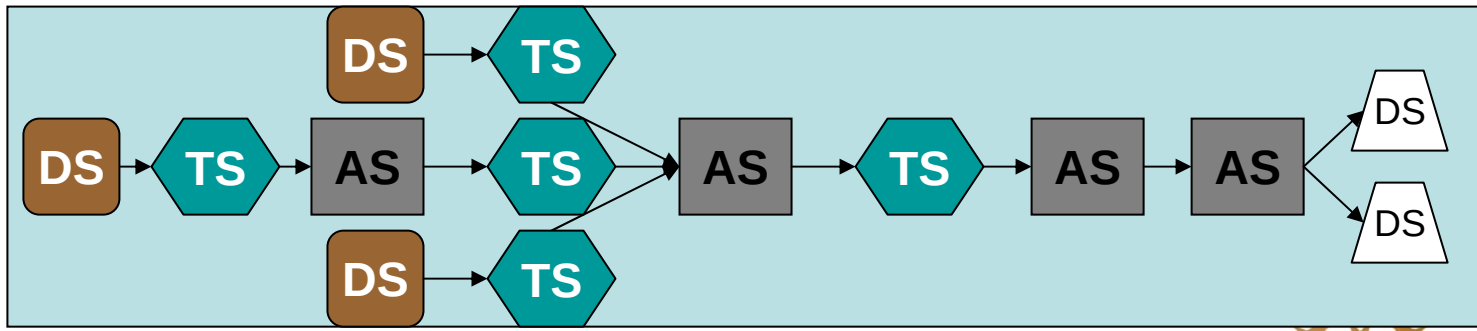
Conceptual Workflow



Abstract Workflow



Executable Workflow



Scientists design their research at the conceptual workflow level

- Often done on the fly over the period of time the research is being conducted

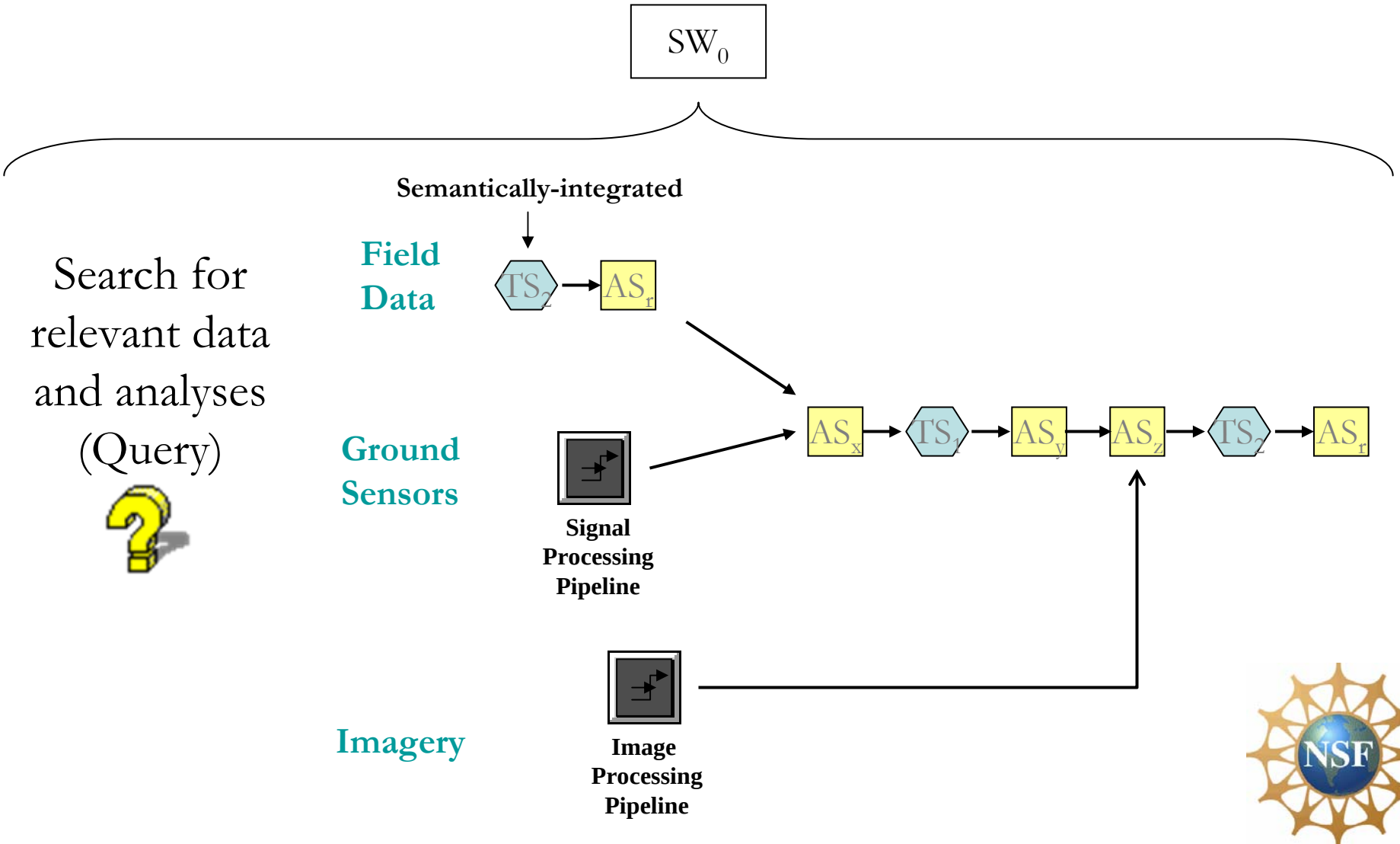
- For automated approaches, this must be well thought out from the beginning

- HOWEVER, because of the automation it is easy to modify the analysis and rerun it many times, so you are not locked into the original design

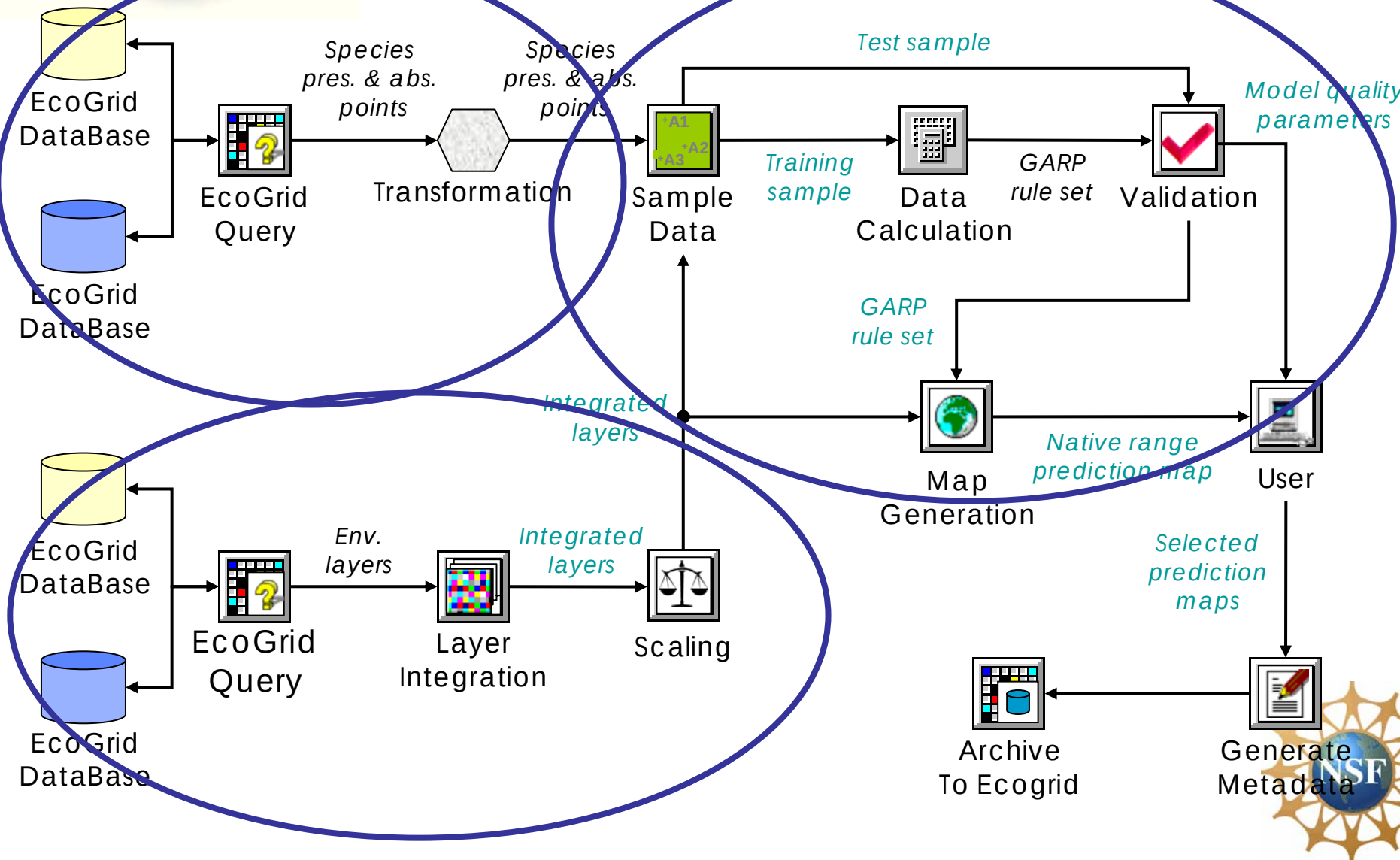
Benefits

- Reusable analysis steps, pipelines, and workflows
- Formal documentation of methods
  (output in report format)
- Reproducibility of methods
- Visual creation and communication of methods
- Versioning
- Automated data typing and transformation

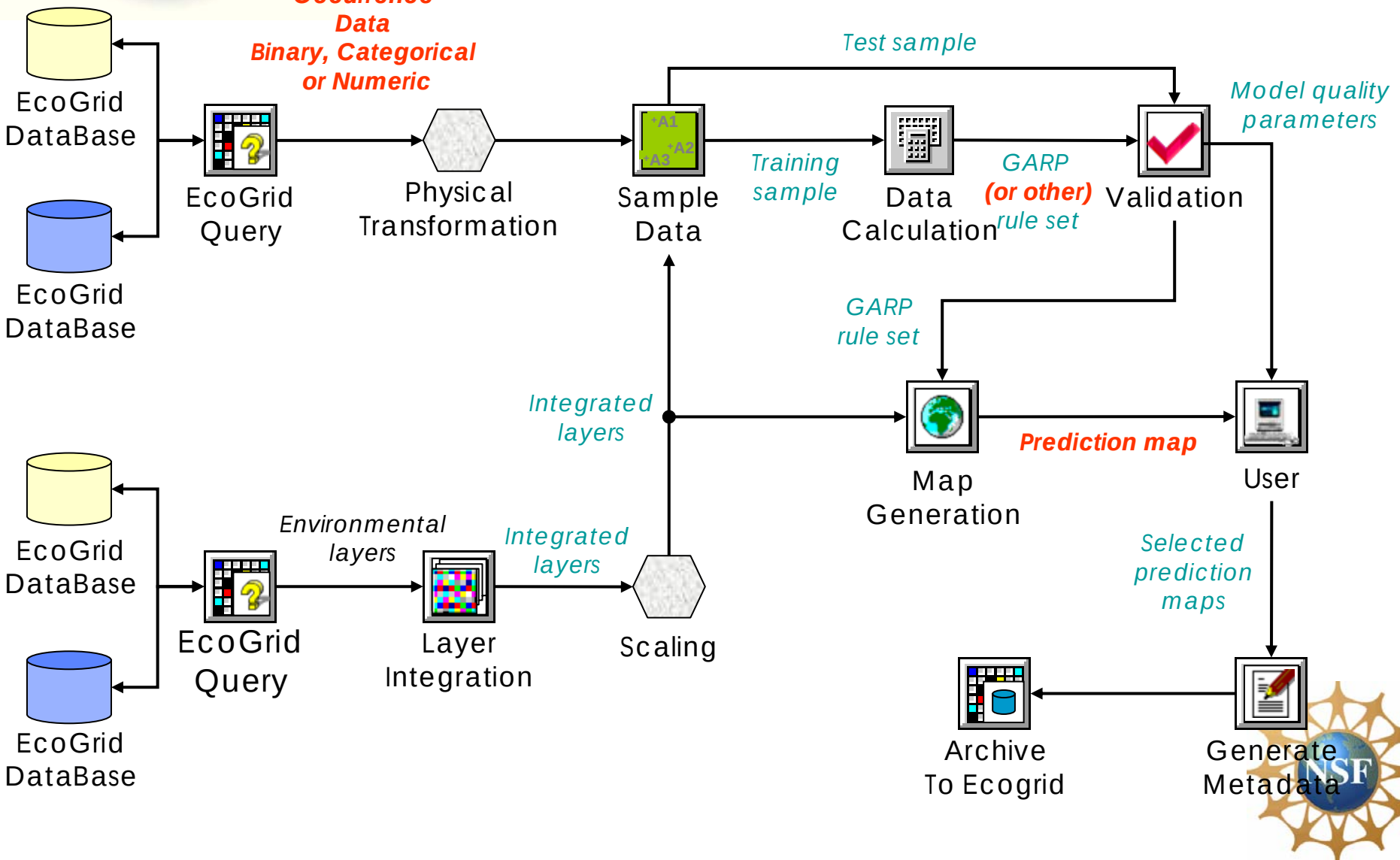
Nested workflows



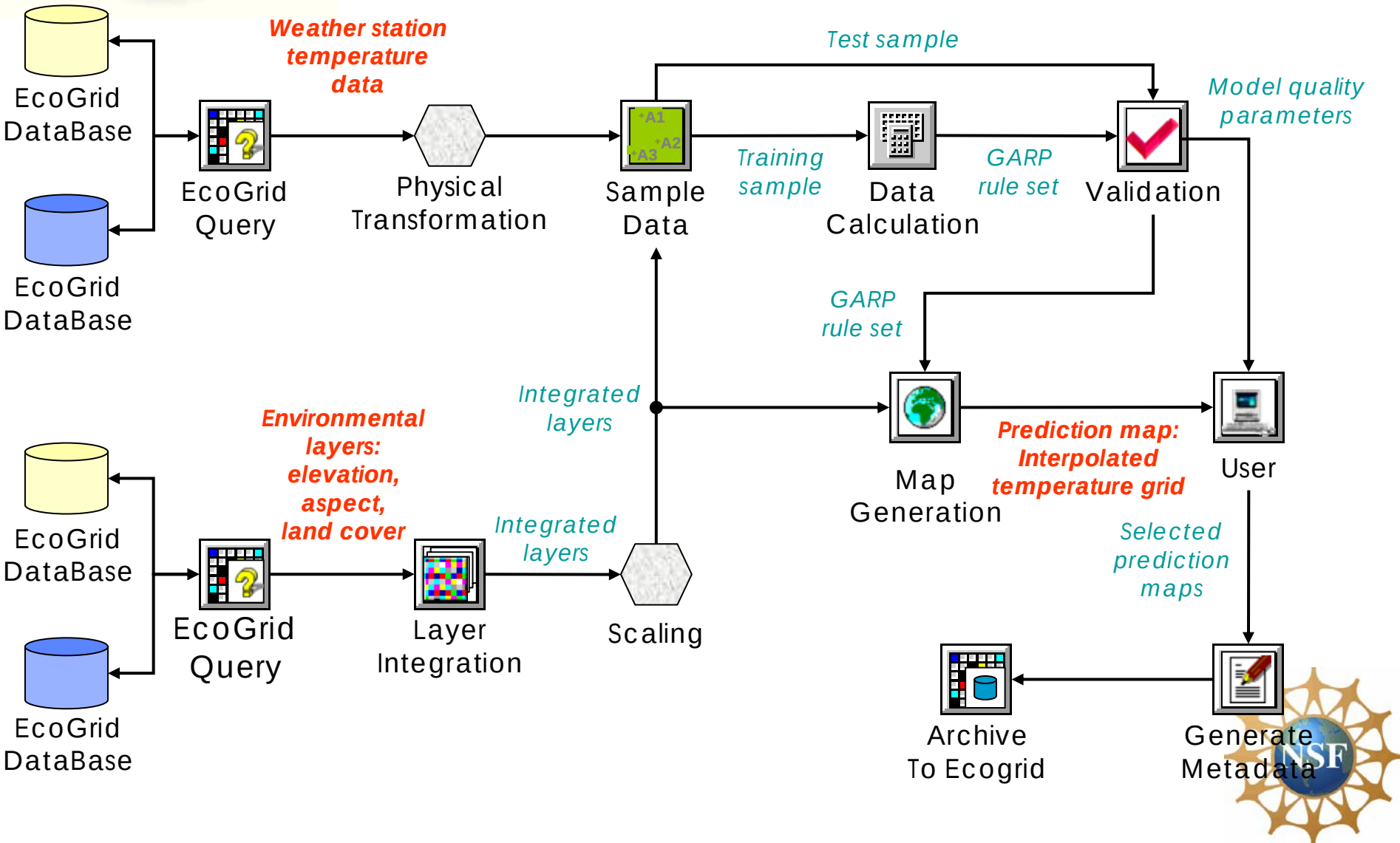
GARP Native-Species Pipeline (informal)



Generic Workflow



Temperature Interpolation Workflow



Exercise

1. Divide into groups of 4 (or so) with similar research interests
2. Pick a research topic to collaborate on
3. Brainstorm various mental models around that topic
4. Construct a conceptual workflow diagram for an analysis that could be conducted
5. Discuss how it could be reused for other related or unrelated analyses