Introduction to Scientific Workflows

Deana Pennington
University of New Mexico
January 3, 2005

What is a workflow?

How does climate affect productivity?
How could I test that?

Acquire data

What is a workflow?

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
...						

Sensor Dataset

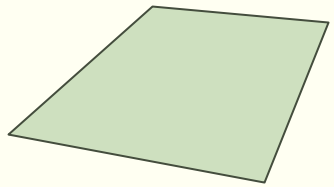


Image Dataset

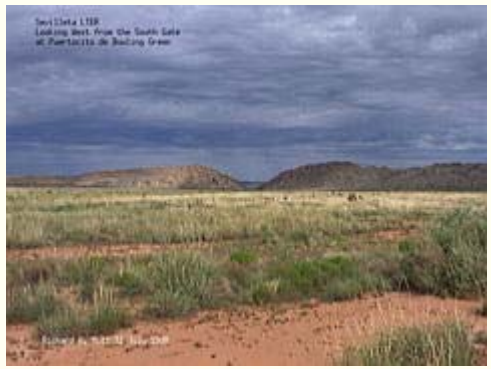
Weather Dataset

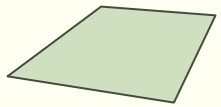
What is a workflow?

Search for compatible data

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

TIME	BMASS
t1	323.1
t2	366.7
t3	383.2
…	





What is a workflow?

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, …

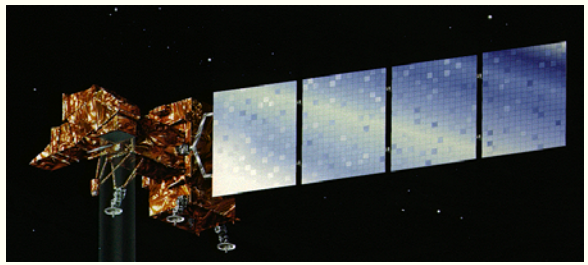
Location: White Mountain
Methods: Protocol B
Plot size: $1m^2$

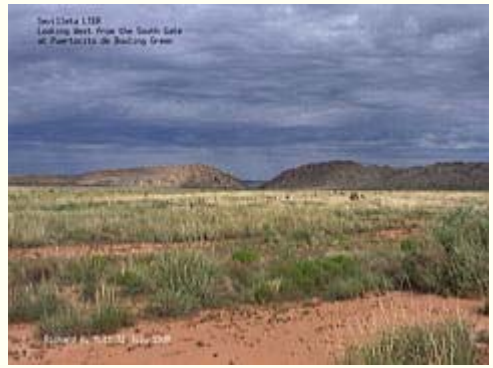
TIME	BMASS
t1	323.1
t2	366.7
t3	383.2
…	

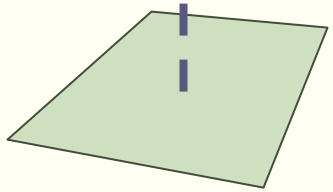
Data Integration:
Physical: file type
Logical: data organization
Semantic: space, time, methods

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
…		

What is a workflow?



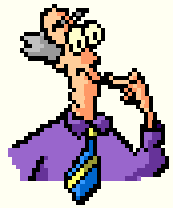




What is a workflow?

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
…		

GIS Expert



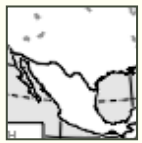
Sample 1, lat, long, presence

Sample 3, lat, long, absence

Sample 2, lat, long, presence

Biome

Elevation (m)



Mean annual temperature (C)

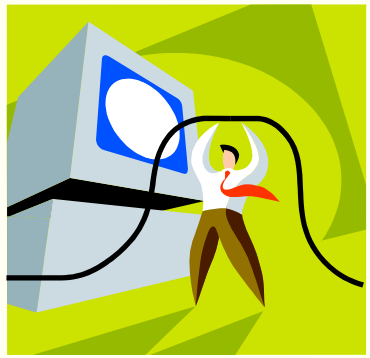
Integrated data:

323.1, biomeA, 2200m, 16C
366.7, biomeB, 2320m, 14C
443.5, biomeC, 1535m, 22C

What is a workflow?

Integrated data:

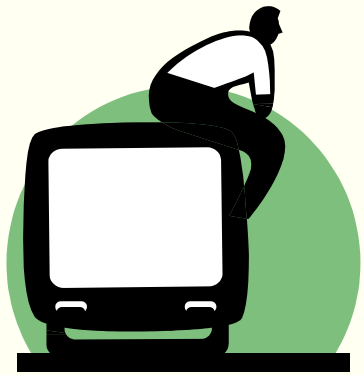
323.1, biomeA, 2200m, 16C
366.7, biomeB, 2320m, 14C
443.5, biomeC, 1535m, 22C



Big, ugly statistical program

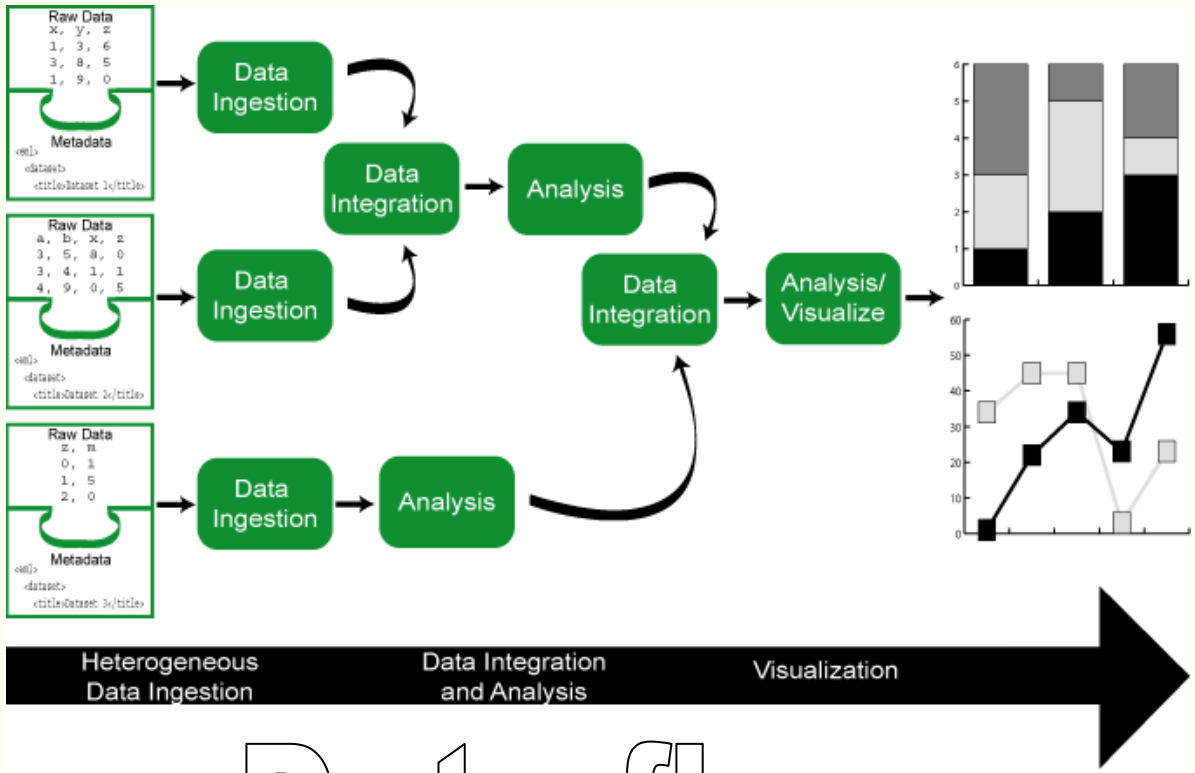
What is a workflow?

- What exactly did I do?
- Where did I put that piece of data?
- Who snatched the GIS expert away?
- Did I make that transformation before I…



What is a workflow?

Research Design



Reporting Sharing

Data flow

What is a workflow?

Data flow

Research Design

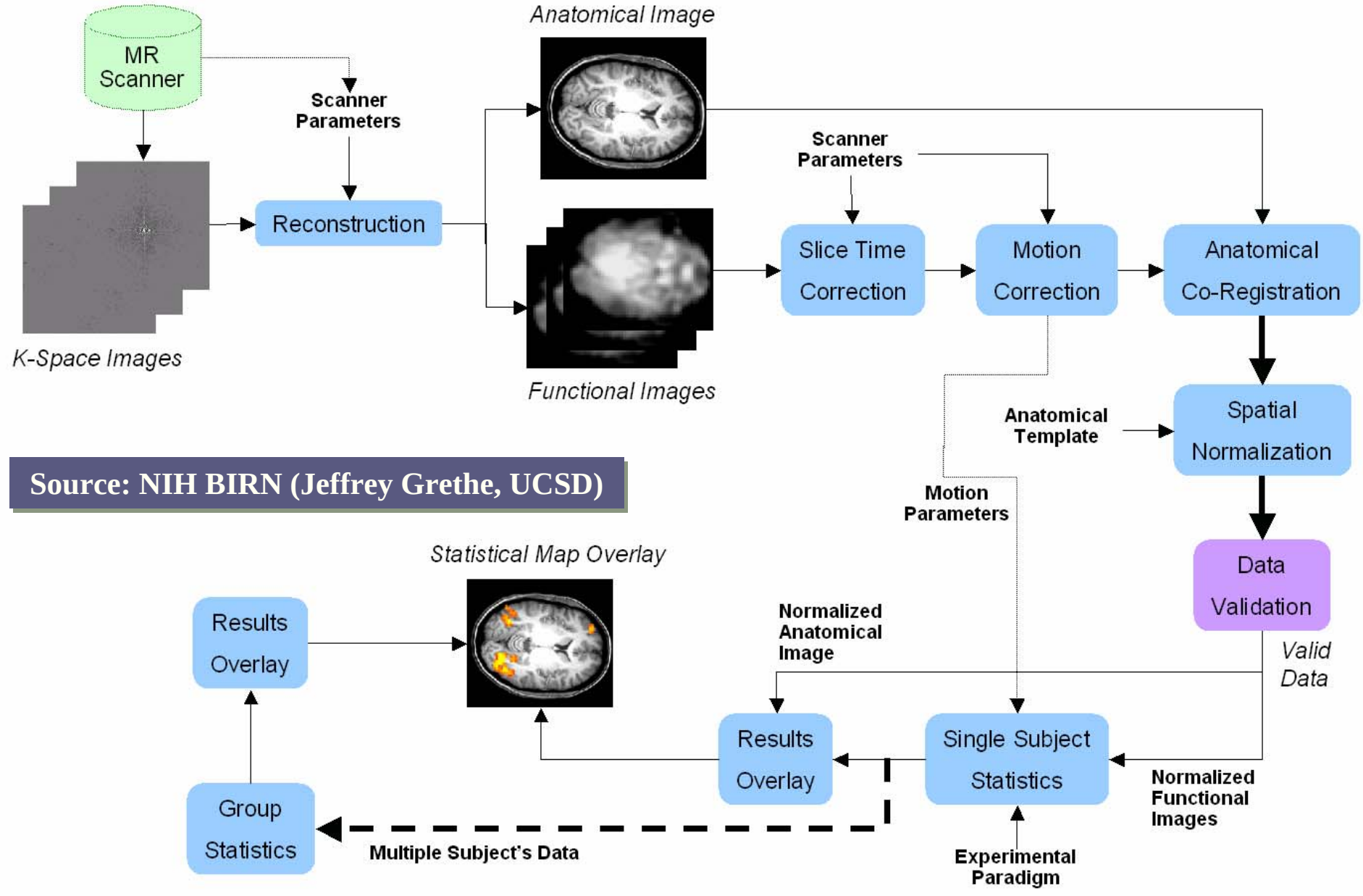
- Data discovery
- Analysis discovery
- Access to computational resources
- Automated data integration
- Automated transformation for analysis
- Workflow analysis

Reporting
Sharing

Workflow sharing

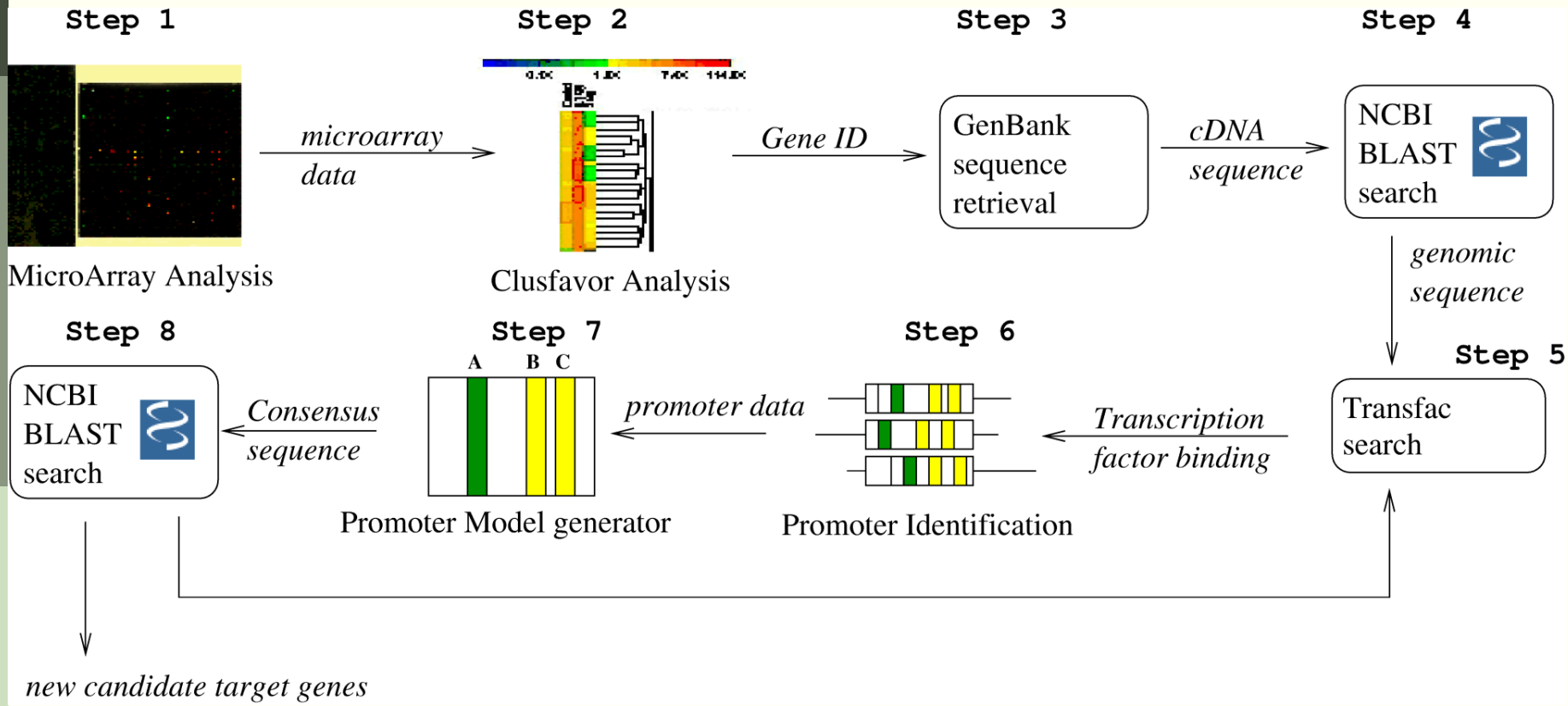
Workflow as methods metadata

Functional MRI Analysis Workflow



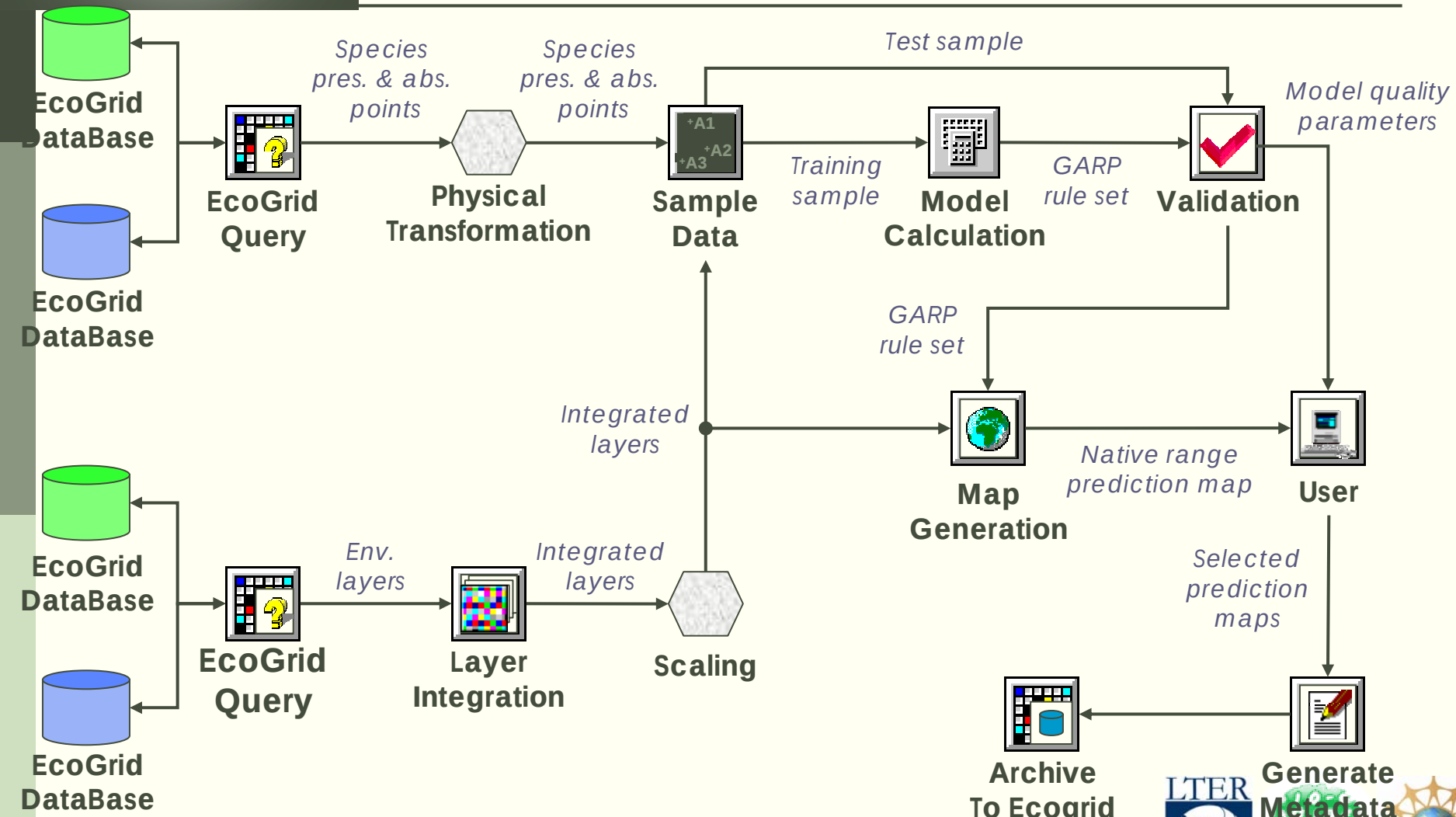
Source: NIH BIRN (Jeffrey Grethe, UCSD)

Promoter Identification Workflow (PIW)

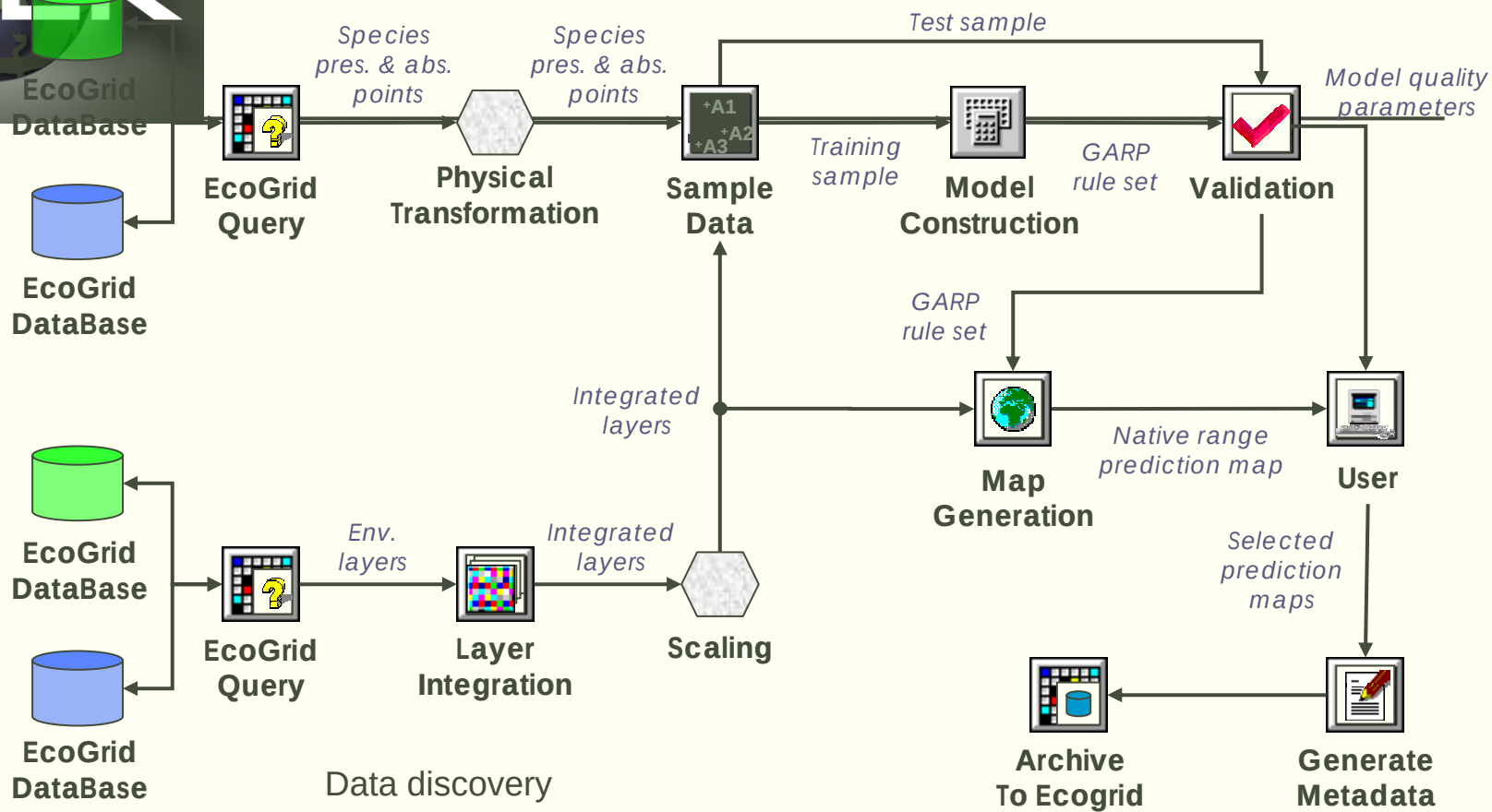


Source: Matt Coleman (LLNL)

Species Distribution Workflow

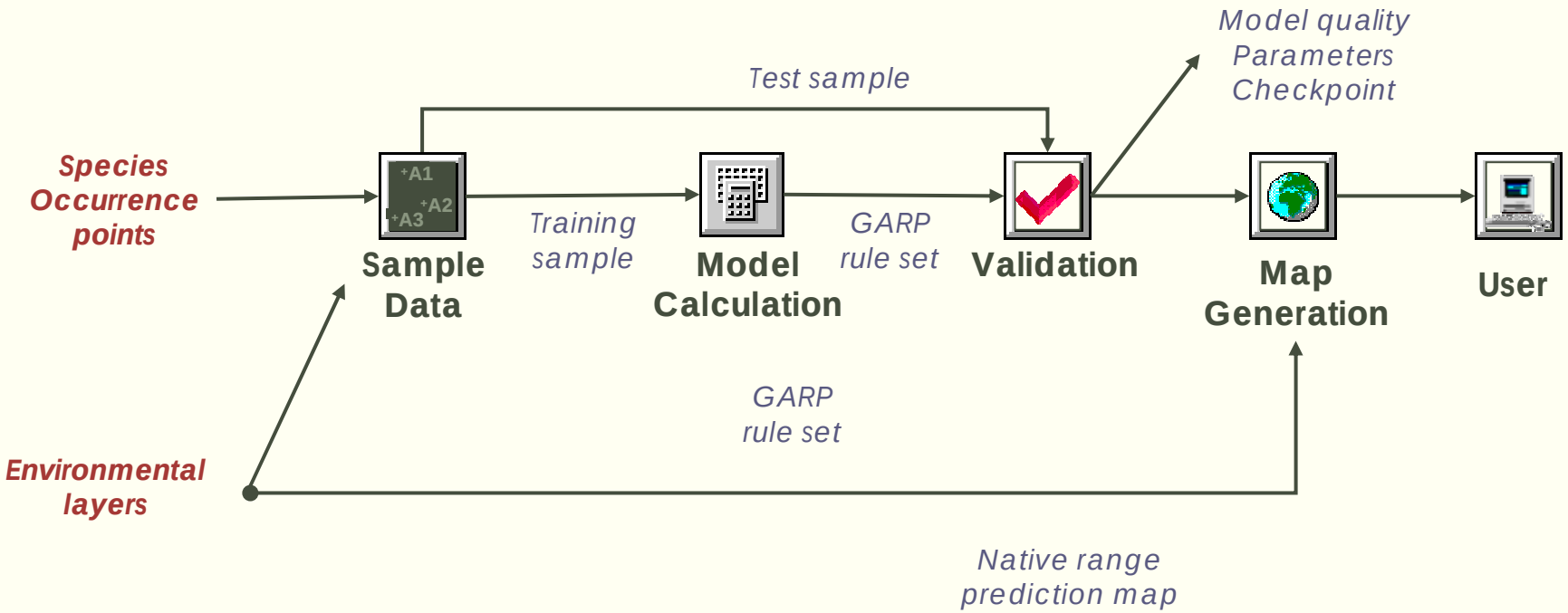


Species Distribution Workflow

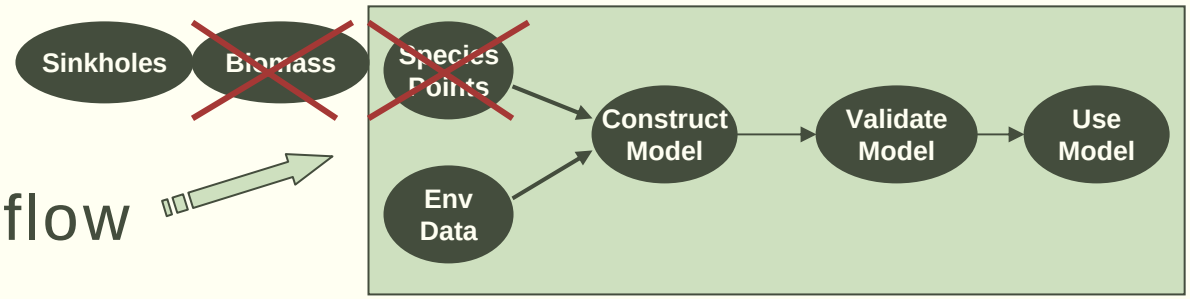


Data discovery
Analysis discovery
Access to computational resources
Workflow as methods metadata
Shared workflows
Automated data integration
Automated transformation for analysis
Workflow analysis

Abstract Workflow

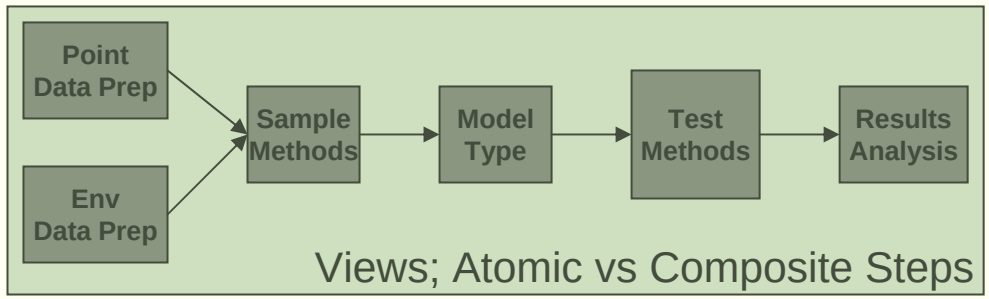


Hierarchical Workflows

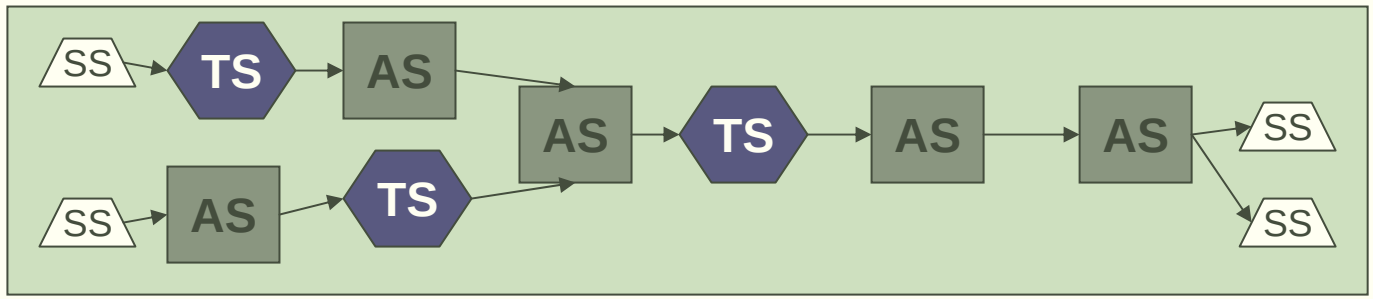


Conceptual Workflow

Abstract Workflow



Executable Workflow



- C — Concept
- AS — Analysis Step
- TS — Transformation Step
- SS — System Step

Workflow Reusability

- Reproduce/methods details
- Reuse within-discipline different data
- Reuse within-discipline different model/analyses
- Reuse between-discipline

Exercise: Conceptual and Abstract Workflows

1. Break into groups of 4
2. Pick a research topic
3. Construct an abstract workflow
4. Stop at 11:45
5. Group presentations

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