Introduction to Ecoinformatics

January 9, 2006

Deana Pennington
Research Assistant Professor
Co-director of SEEK
University of New Mexico

Goals

- Provide a context for understanding informatics, cyberinfrastructure, and how ecology fits into that context
- Explain what the SEEK project is and how it contributes to both ecology and informatics
- Put the topics that will be covered this week into this context
- Answer any general questions that you may have regarding any of the above!

What is Ecoinformatics?

Eco-Informatics?

Biodiversity and Ecosystem Informatics?

Relate to bioinformatics?

New: Cyberinfrastructure (CI)?

Old: Computational Science?

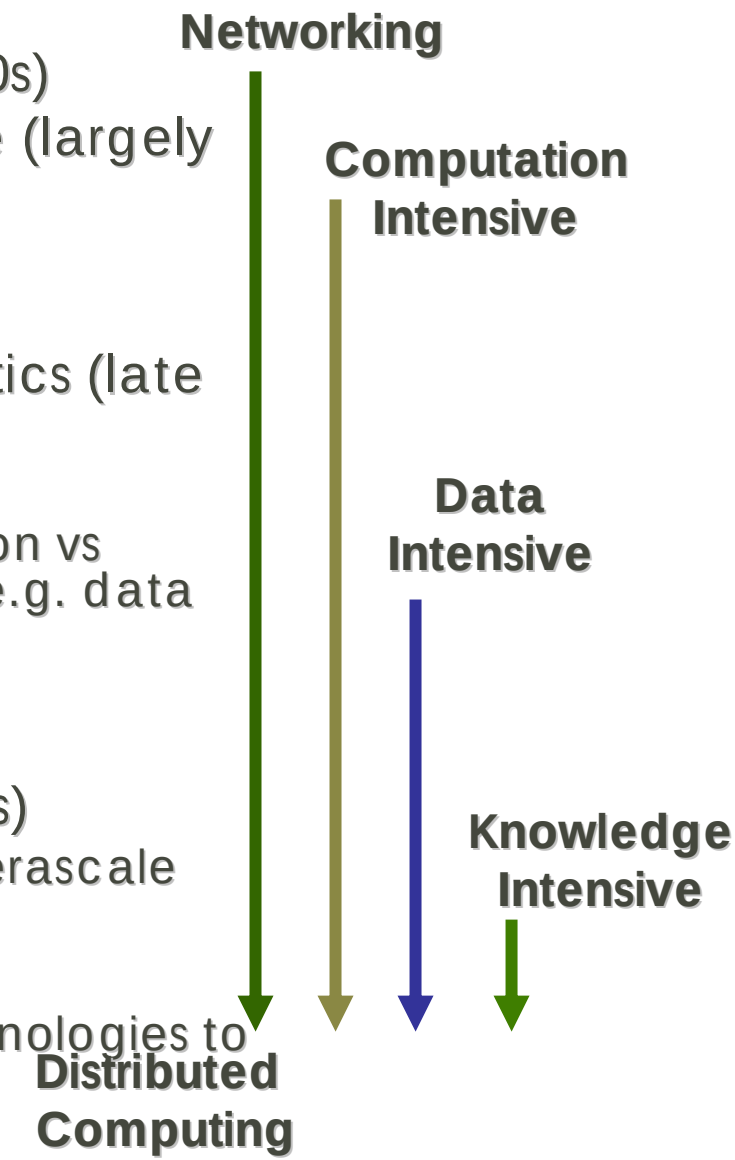
E-science (old: electronic science, new: enhanced science)

E-Ecology? E-Ek!

One letter shy of a Geek!

Some history

- First academic computing centers (1960s)
- Early activities in computational science (largely physics & chemistry) (1970s)
  - numerical analysis via computer
- NSFNet => Internet (1980s)
- Computational biology and bioinformatics (late 1980s - 1990s)
  - Molecular and genetic biology
  - Theoretic and mathematical computation vs Artificial intelligence & heuristic studies (e.g. data driven)
  - Computational science vs informatics
- NSF National Partnership for Advanced Computing Infrastructure (NPACI) (1990s)
  - Between supercomputing centers => Terascale computing
- CISE Information Technology Research
  - Large block of funding set aside for technologies to enable domain science (~ 2000)
- Cyberinfrastucture initiative



More recent history

- Atkins report (2003) "Revolutionizing Science and Engineering Through Cyberinfrastructure" - NSF Blue Ribbon Advisory Panel
  - CI = infrastructure based upon distributed computer, information and communication technology
  - Industrial economy => knowledge economy
- Presidential report: "Computational Science: Ensuring America's Competitiveness" (2005)
  - CS = the use of advanced computing capabilities to understand and solve complex problems
- Office of Cyberinfrastructure (2005)
  - CI-Team = workforce development
  - CI for Env Observatories = NEON (eco), OOI (ocean), CLEANER (env), COSMIC (climate)
  - High Performance Computing = petascale computing

Within ecology

- NRC report: "A Biologic Survey for the Nation" (1993)
  - Organismal biology & biodiversity
  - National Biological Information Infrastructure (NBII) => biologic information pipeline
- President's Committee of Advisers on Science and Technology Panel report: "Teaming with Life" (1998)
- Partnership for Biodiversity Informatics (PBI) (1999)
  - LTER Network Office, NCEAS, Nat'l History Museum at KU, San Diego Supercomputer Center
  - Knowledge Network for Biocomplexity (KNB) project initiated (1999)
- NSF, USGS (NBII), NASA joint workshop on Biodiversity and Ecosystem Informatics (BDEI) (2000, 2001, 2004)
- NBII – 10 nodes initiated (2001)
- NSF solicitation for BDEI (2001)
- PBI awarded $12M ITR research grant = SEEK project (2002)
  - Distributed computing, data-intensive and knowledge intensive approaches
  - Mathematical modeling and computation-intensive implicit
- NEON infrastructure & cyberinfrastructure

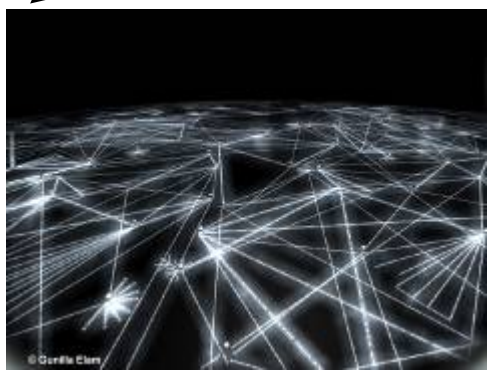
Why multi-scale distributed sensor-networking will transform ecology

Radioastronomy
Single Telescopes

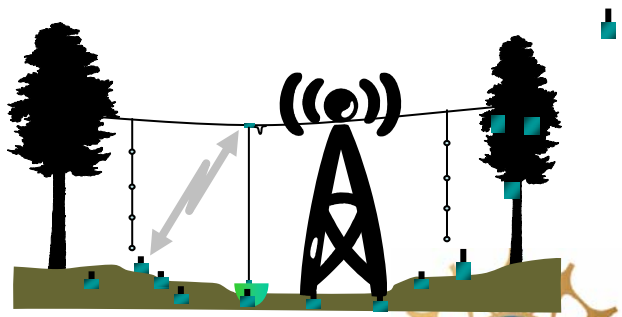
Computing
Supercomputers

Field ecology
Individual observations

… because it has done so over and over again



Internet



NEON

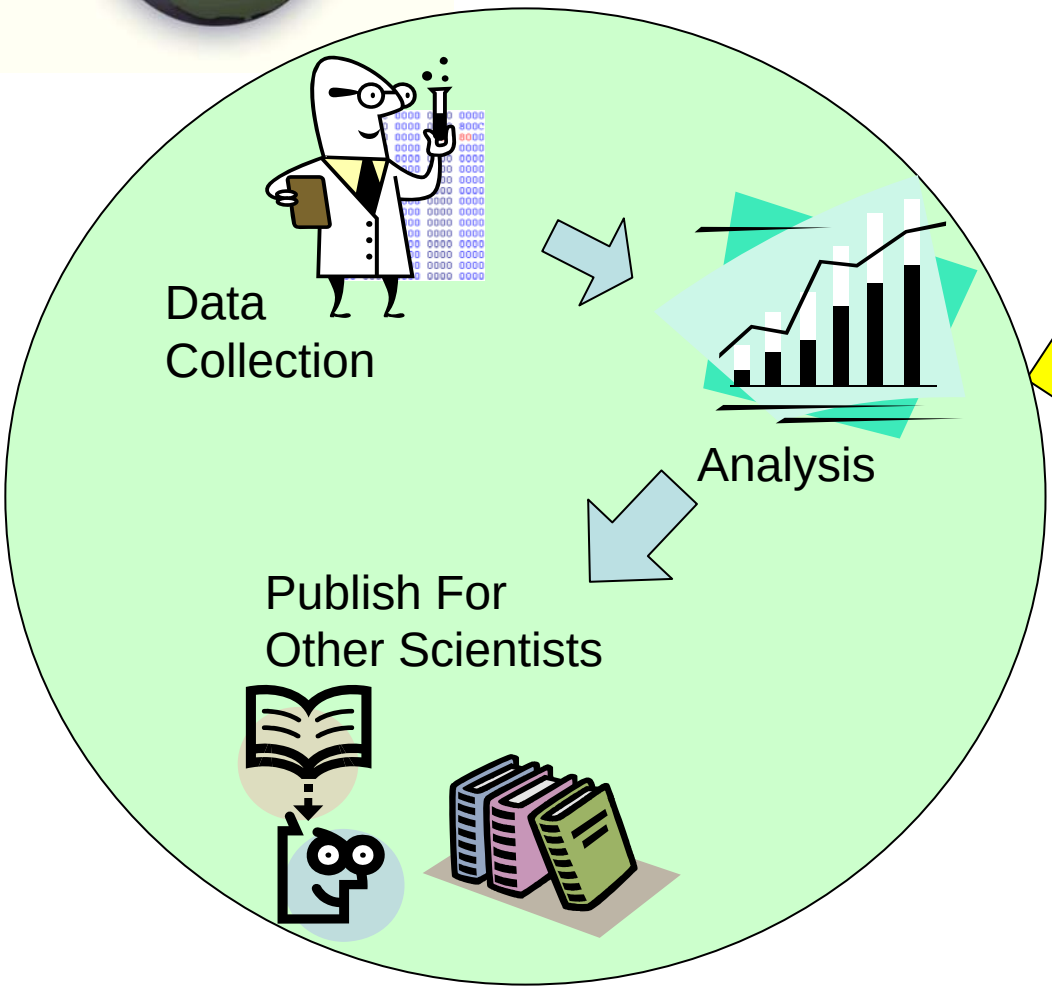
Michener, 2005

Which way ecologists?

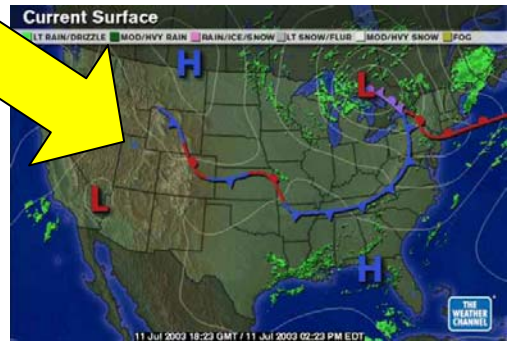
~~Business as usual~~

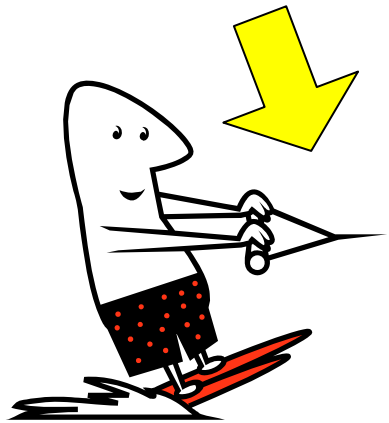
Get on the broad-scale, sensor-driven bandwagon

Move towards distributed, computationally-intensive, data-intensive, and knowledge-intensive approaches



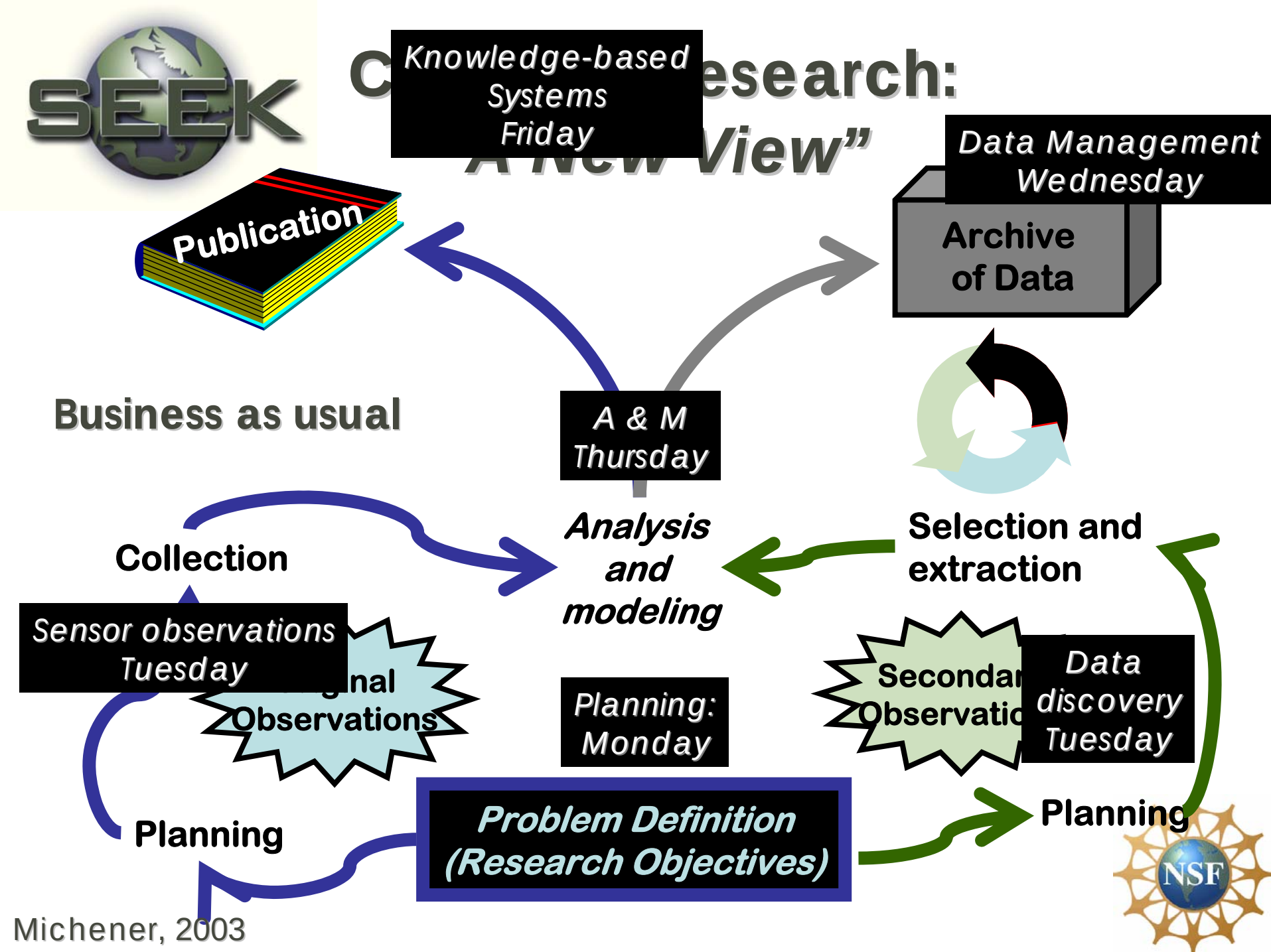
Translation to *broad scale*; data integration and/or modeling





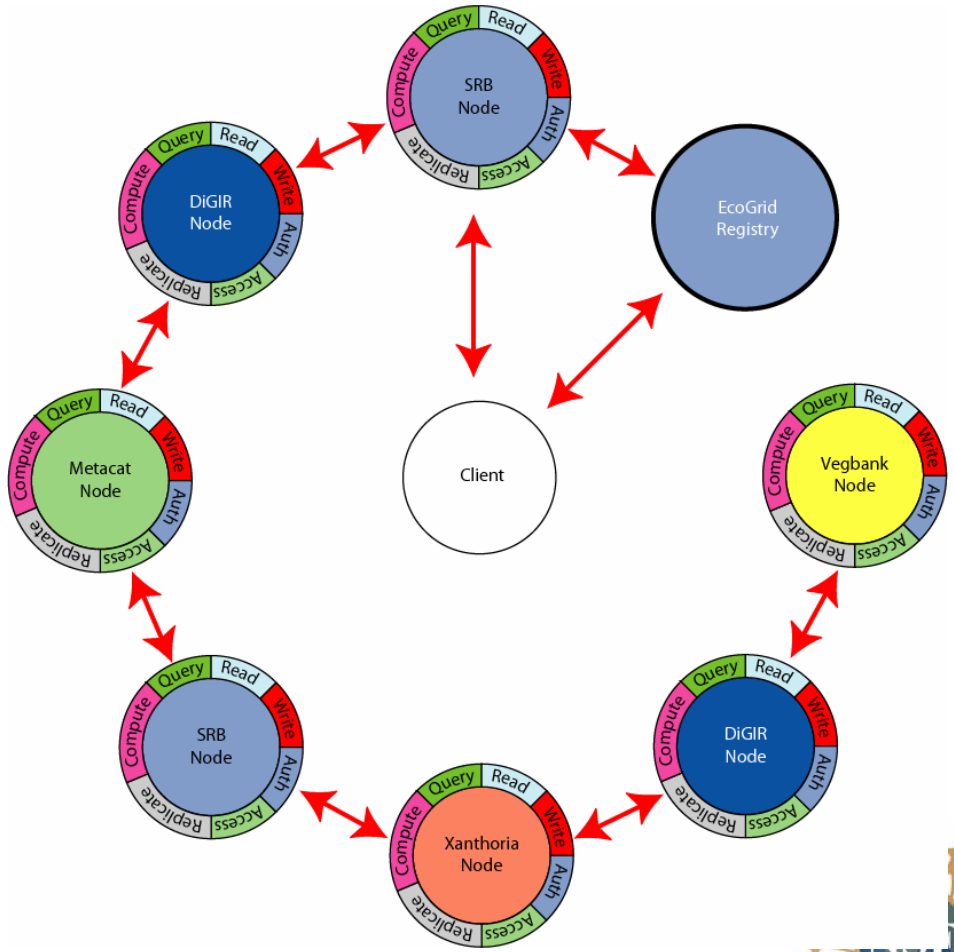
Business as usual

Use By Non-Scientists


SEEK
Knowledge-based Systems Friday
Publication
Data Management Wednesday
Archive of Data
Business as usual
A & M Thursday
Collection
Analysis and modeling
Selection and extraction
Sensor observations Tuesday
Observations
Planning: Monday
Data discovery Tuesday
Planning
Problem Definition (Research Objectives)
Planning
NSF
Michener, 2003

Internet vs Distributed Computing?

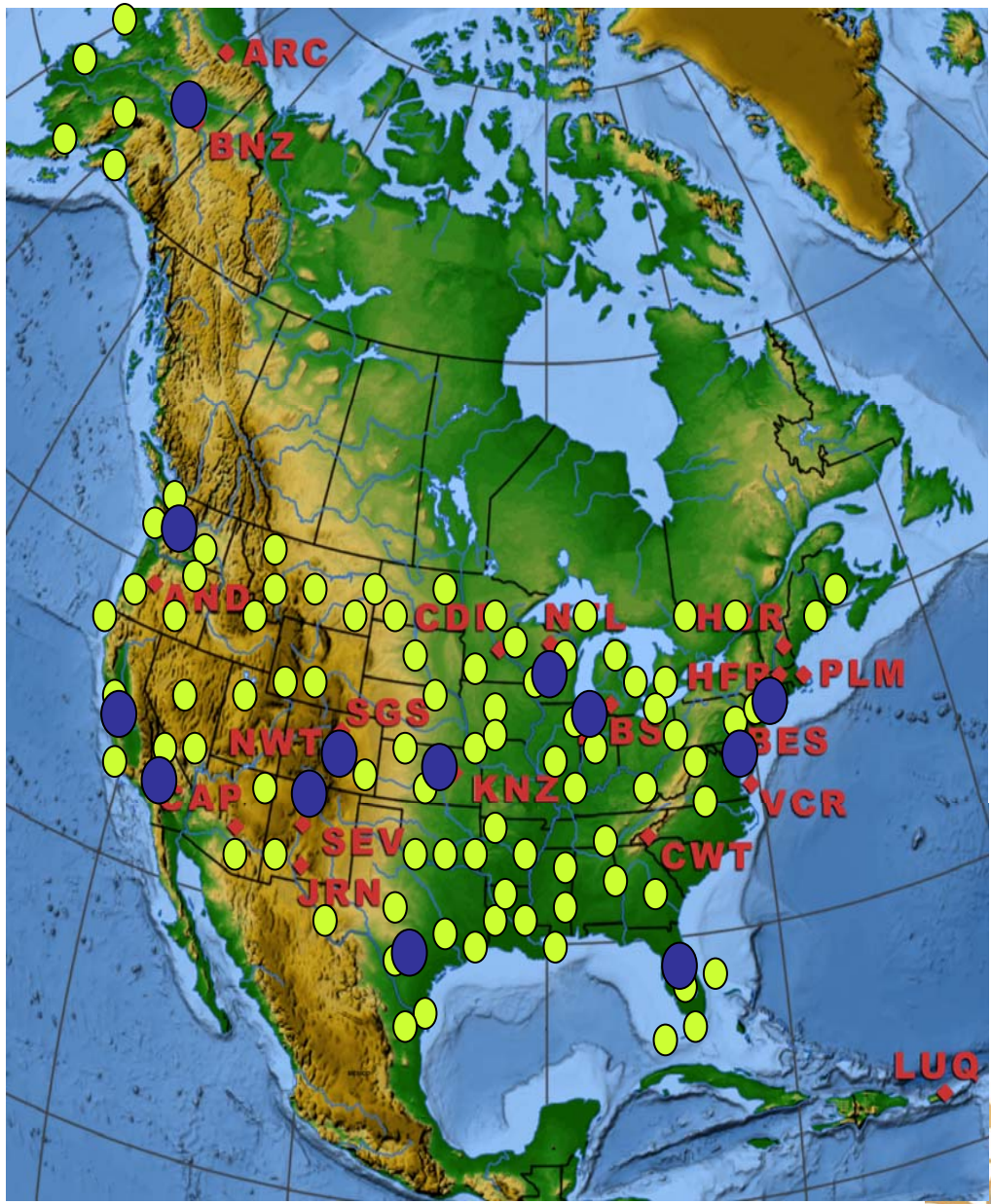
- Web => User goes to a URL and downloads what they need
- Web application => User goes to a URL and launches an application at a remote site, then downloads the results
- Distributed computing => Software at user's terminal finds the appropriate resource, goes to that URL, gets/does what is needed and returns it to the terminal for continued processing, seamlessly, hidden from user



Some U.S. Field Station Locations

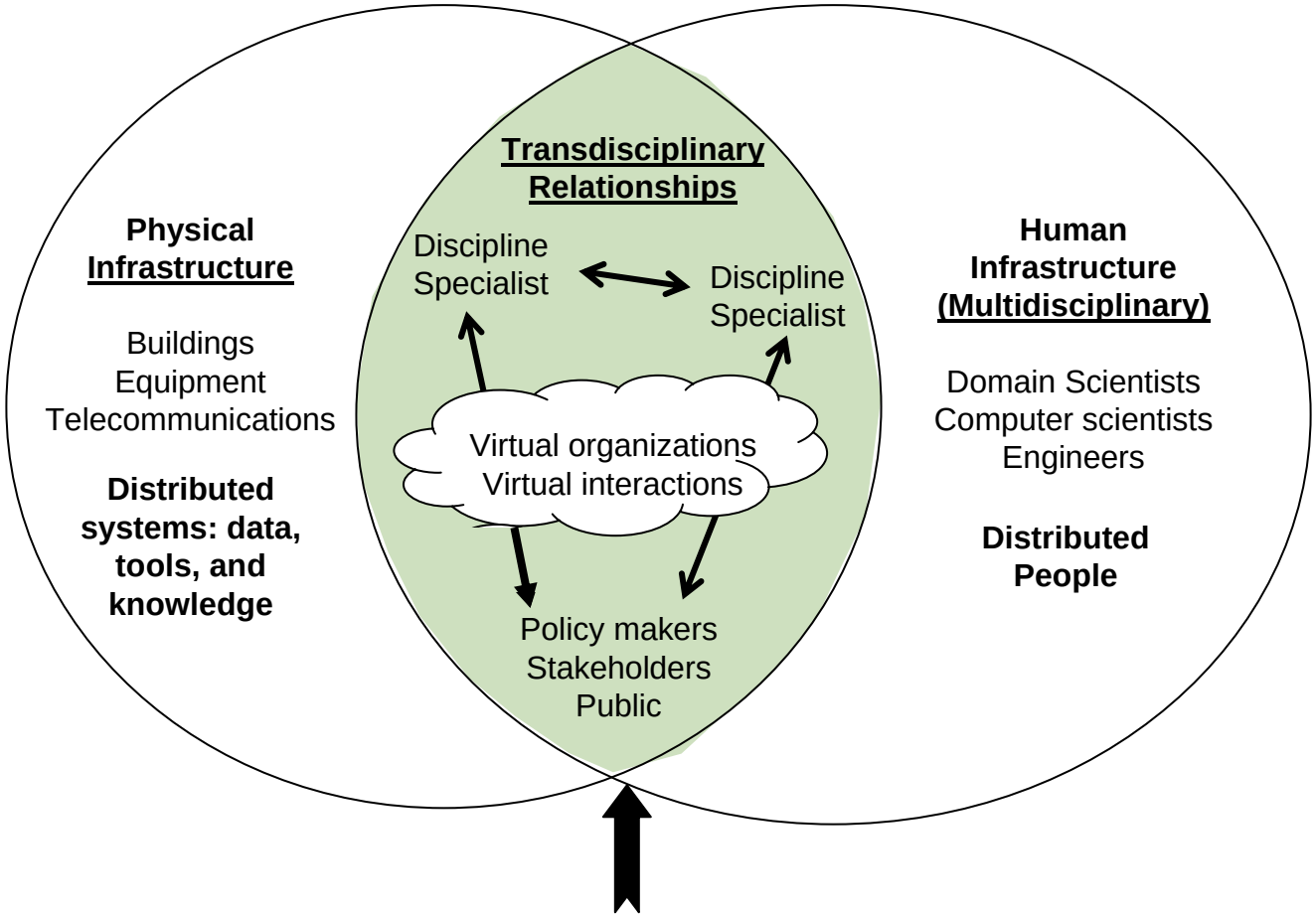
XXX LTER

OBFS



Cyberinfrastructure

CI = infrastructure based upon distributed computer, information and communication technology



Computational Science

CS = the use of advanced computing capabilities to understand and solve complex problems

Three pillars of science

1. Theory
2. Experimentation
3. Computational science: models of complex phenomena, analyze voluminous data

It is itself a discipline, also serves to advance all of science.

"The most scientifically important and economically promising research frontiers in the 21st century will be conquered by those most skilled with advanced computing technologies and computational science applications"

Presidential report: "Computational Science: Ensuring America's Competitiveness" (2005)

Computational Science

"…balanced investment in both computational science itself and *its applications across many domains*. Research in high-end architecture, systems software…differs from *research in the use of computational science to address challenging application problems*. Both kinds of research are important, but they require different expertise and generally are conducted by different people. It is a mistake to confound the two."

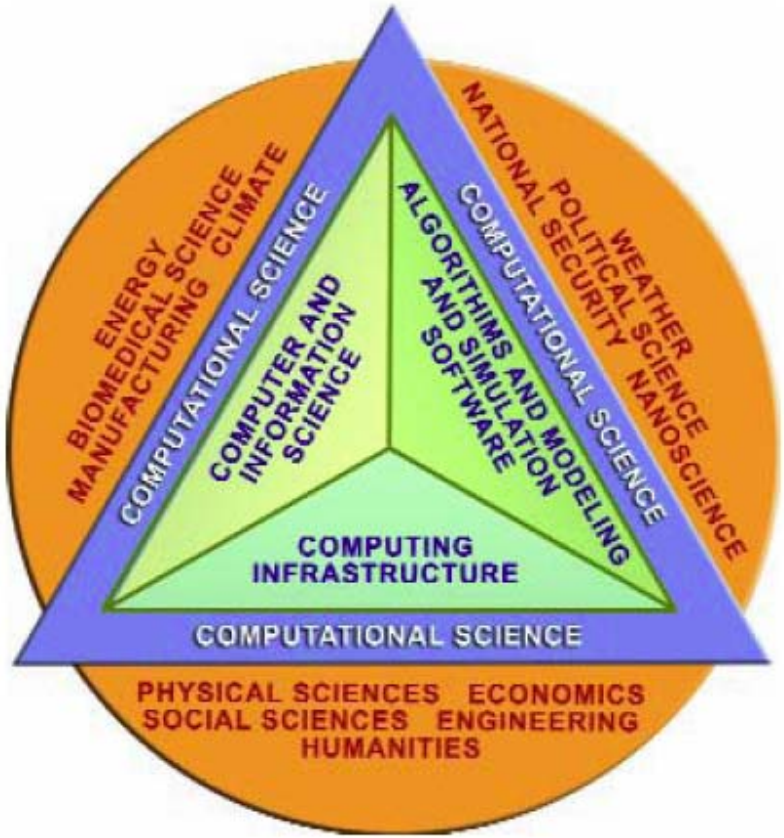


Figure 1

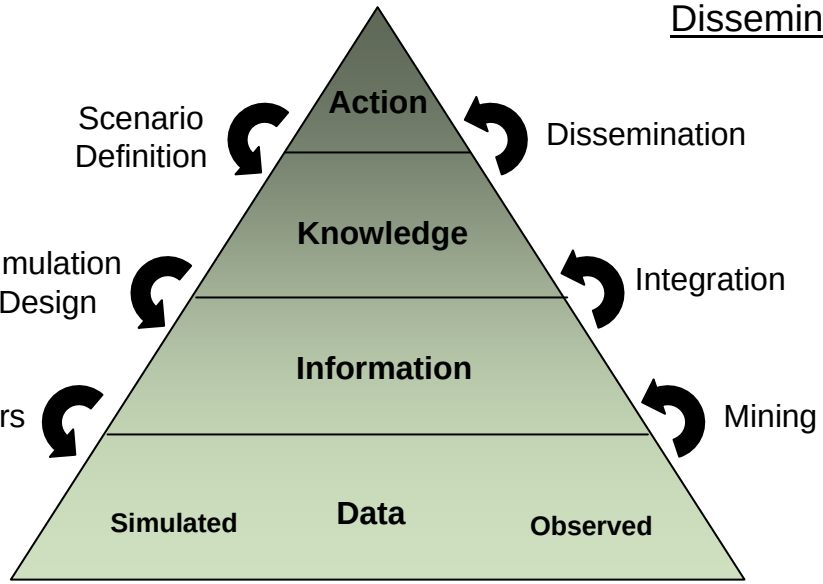
Presidential report: "Computational Science: Ensuring America's Competitiveness" (2005)

Computational Science

Knowledge Assessment Tools

Knowledge Discovery & Dissemination (KDD) Tools

Theoretical science, scenarios, forecasting, backcasting



Empirical science, data analysis & synthesis

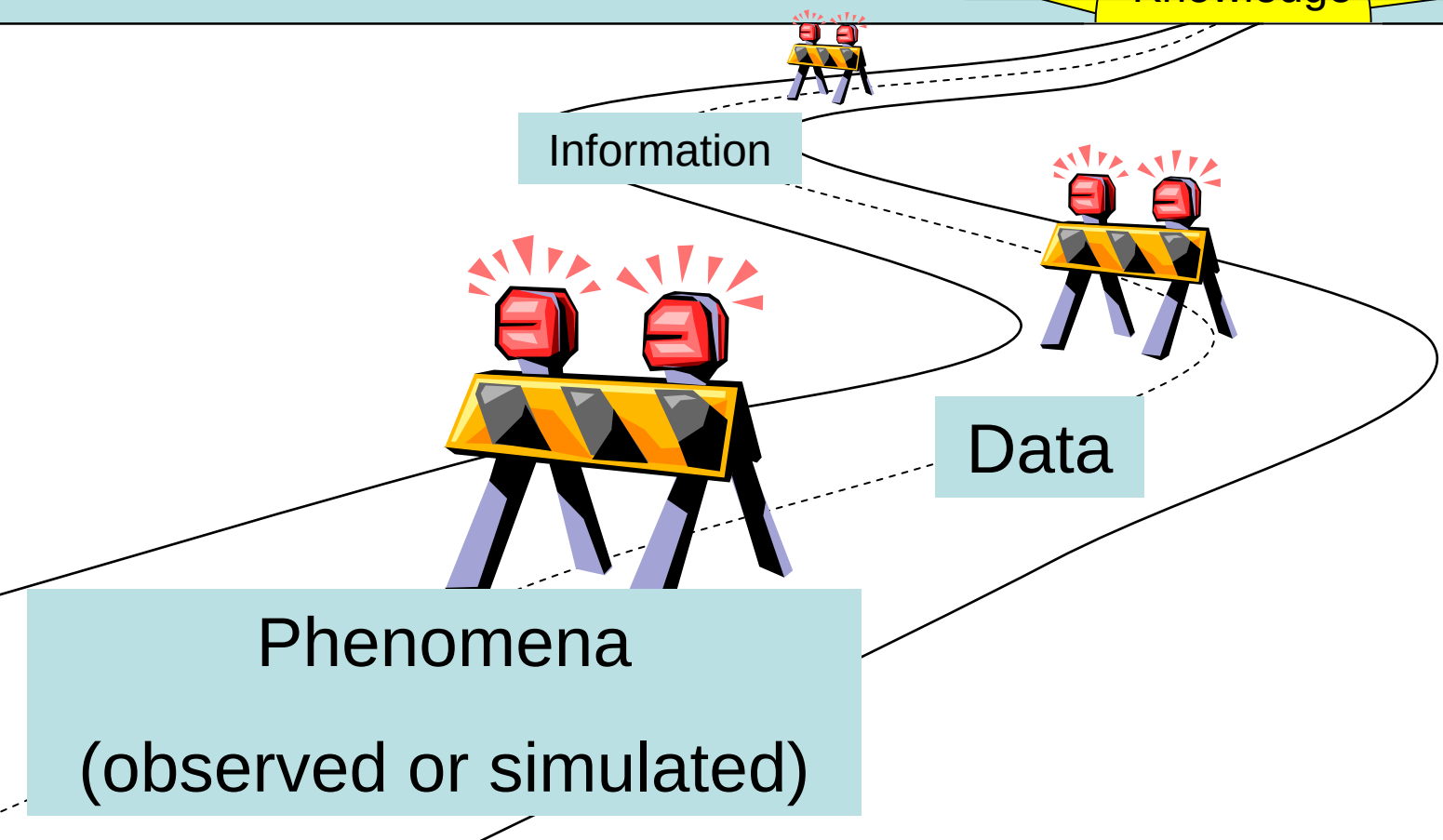
"Informatics"

Data Acquisition

Data Management & Discovery Tools

- **Challenges in ecology & environmental sciences**
  - **Distributed data and resources**
  - **Very small => very large databases**
  - **Heterogeneous data**
  - **Poor metadata**
  - **Implicit semantics**
  - **Georeferencing**
  - **Taxonomic freedom**

- Distributed computing
- Compute-intensive approaches
- Data-intensive approaches
- Knowledge-intensive approaches
  - SEEK project



Agenda

- What this week is NOT:
  - Software training (although we will do some of this)
- What this week is:
  - Preparation for engaging in computational science
  - Understanding the conceptual approaches of distributed computing, compute-intensive, data-intensive, and knowledge-intensive approaches
  - Explanation of the application of those approaches in SEEK
  - Training on tools being developed within SEEK
  - Exposure to some ongoing applications in ecology
- What we will cover:
  - Monday: Intro, research design in computational science
  - Tuesday: Data acquisition, especially GPS, embedded sensor networks, grid technologies & distributed computing
  - Wednesday: Data management
  - Thursday: Data analysis in Kepler
  - Friday: Data integration and synthesis (in Kepler)