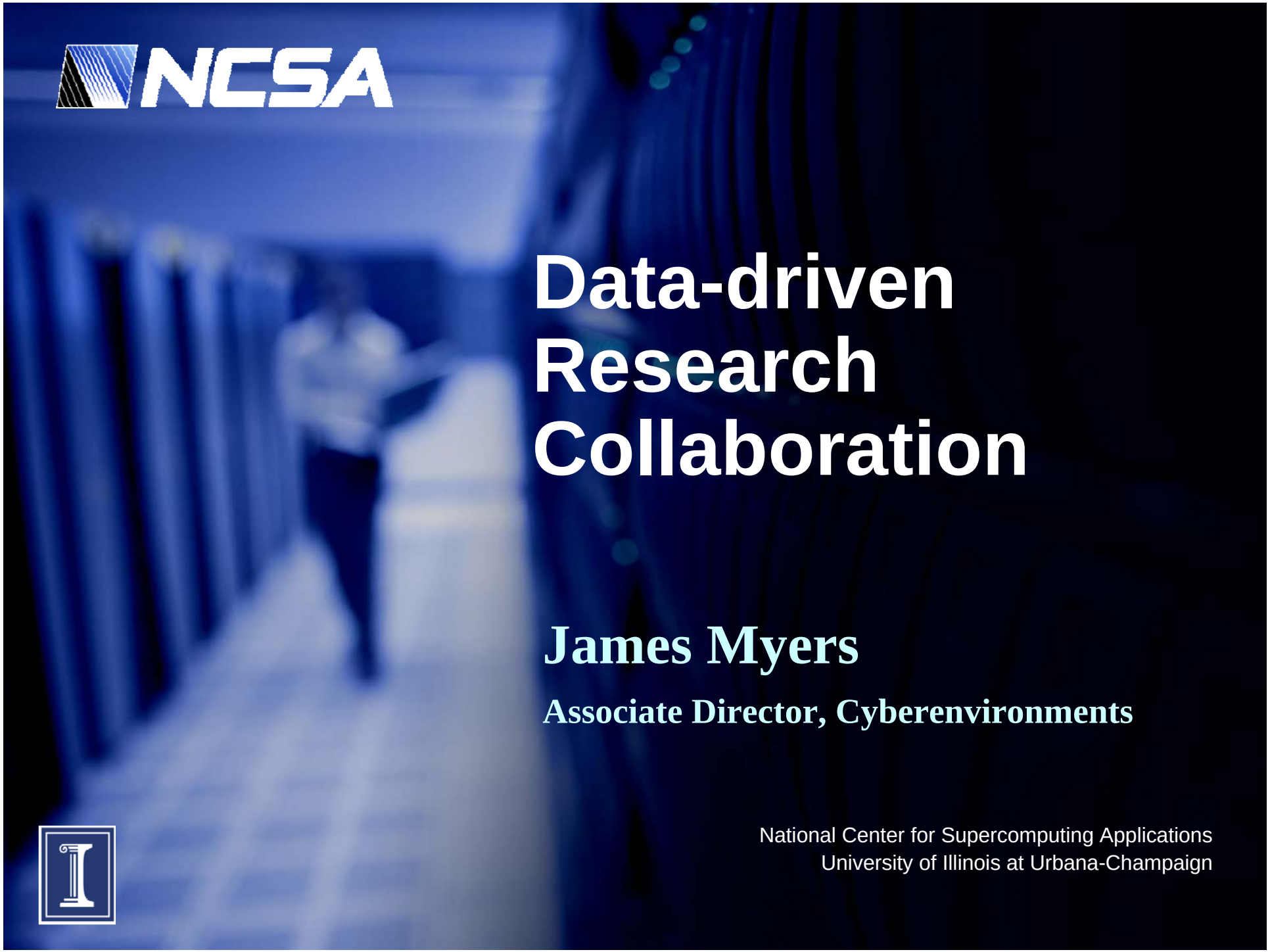


Data-driven Research Collaboration

James Myers

Associate Director, Cyberenvironments

National Center for Supercomputing Applications
University of Illinois at Urbana-Champaign

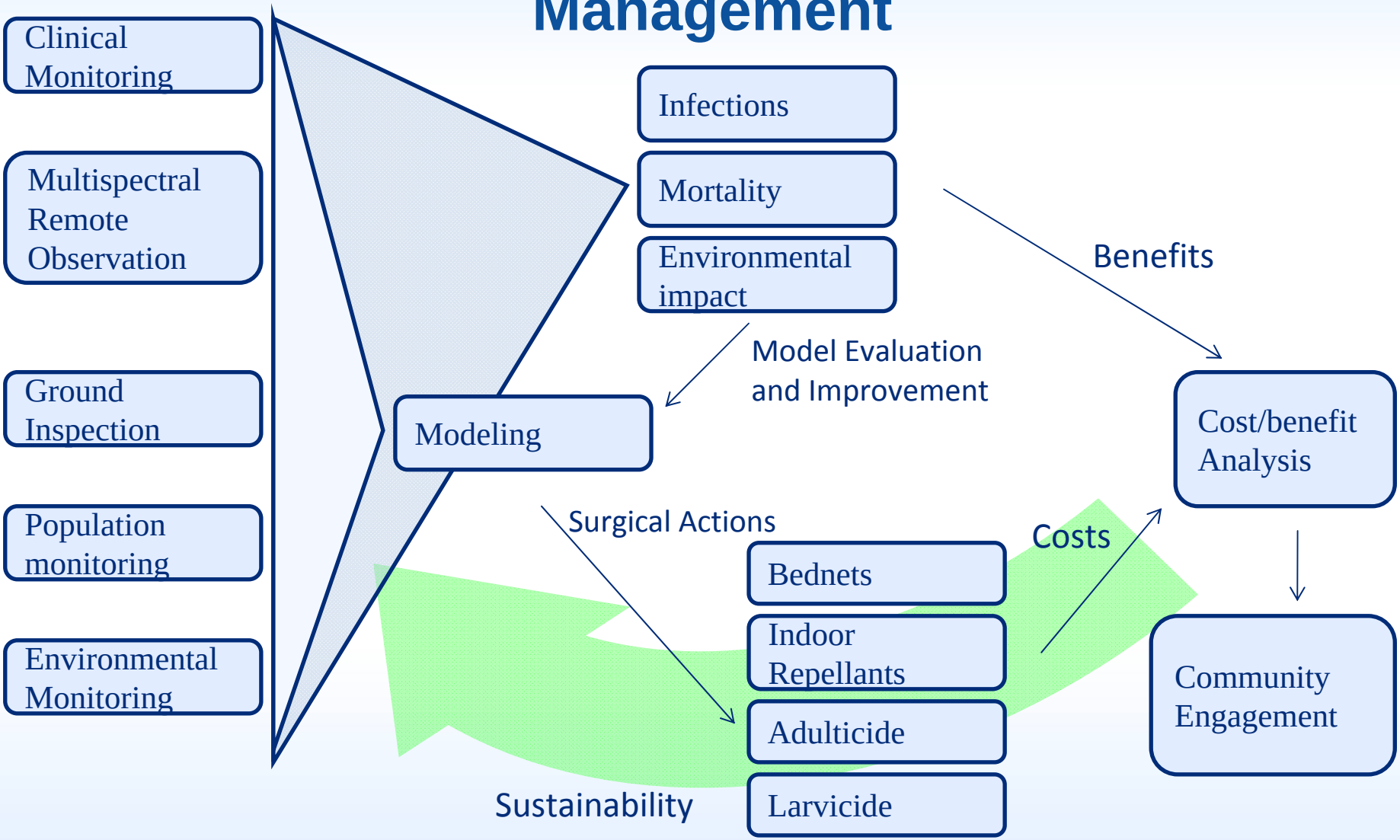
Outline

- **Project Areas**
  - Endemic Diseases
  - Seismic Safety
  - Eco-physiological modeling of Plant Growth
  - Data Center Collaborations
- **Requirements to support Community Research Collaboration**
- **Data Services vs. Serving Reference Data**

Integrated Malaria Management Consortium

Using advanced information systems to help control malaria

An Integrated Information System for Malaria Management



Countries

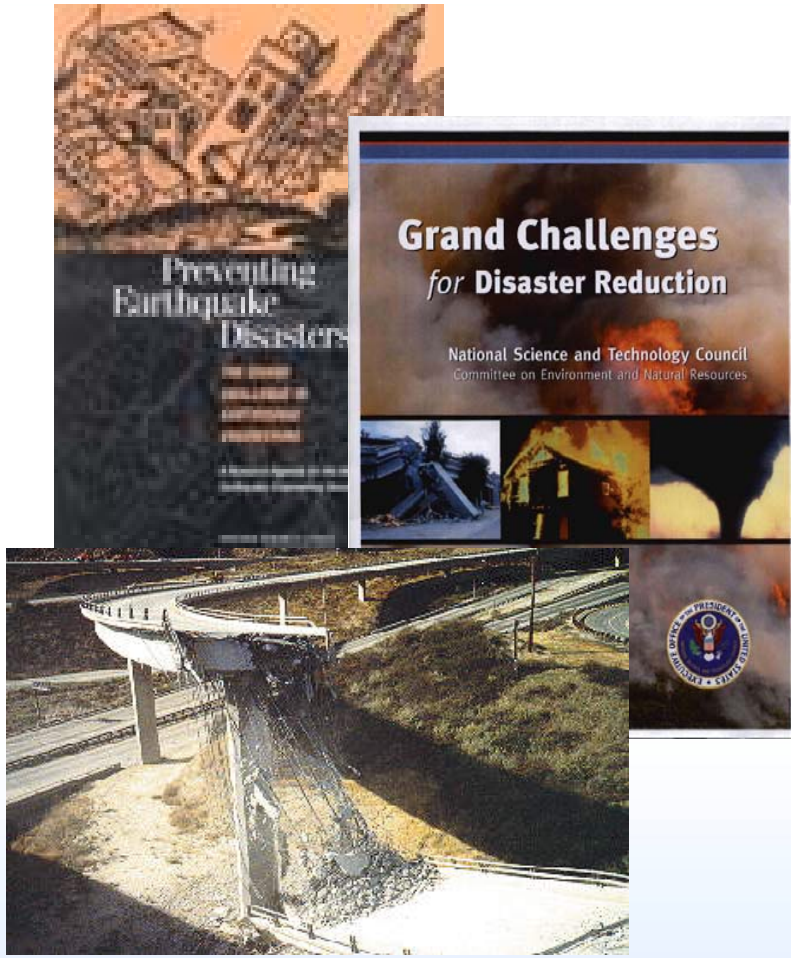
- **Formal Government support**
  - Costa Rica
  - Kenya
  - Zanzibar
- **Government level discussions**
  - Ecuador
  - Peru
  - Uganda
  - Zambia
- **Initial discussions**
  - Brazil
  - China
  - Liberia
  - New Guinea
  - Nicaragua
  - Panama
  - Rwanda
  - São Tomé and Príncipe
  - Singapore
  - South Africa

IMMC Participants

ADAPCO
Eastern and Southern Africa Centre for International Parasite Control
African Insect Science for Food and Health, ICIPE, Nairobi, Kenya
Kenya Medical Research Institute (KEMRI), Kisumu, Kenya
African Malaria Network, Tanzania
Medical Entomology, Epidemiology of Vector-borne Disease, University of the West Indies, Trinidad and Tobago
Arro-Gun Spray Systems LLC, Reno, NV
Awhere, Inc., CO
Millennium Institute, Arlington, VA
American Mosquito Control Association (AMCA), Mount Laurel, NJ
National Center for Supercomputing Applications (NCSA), University of Illinois, IL
College of Business Administration, University of Illinois at Chicago, IL
Fogarty Center, National Institutes of Health, Bethesda, MD
College of Public Health, University of Oklahoma, Oklahoma City, OK
National Institute for Medical Research (NIMR), London, England
College of Veterinary Sciences, Department of Pathobiology, University of Illinois, IL
Northwest Mosquito Abatement District (MAD), Northwest Cook County, IL
Congress of Racial Equality (CORE), New York, NY
The Office of Population Research, Princeton University, Princeton, NJ
Community Informatics Initiative, Graduate School of Library and Information Science, University of Illinois, IL
School of Economic, Political and Policy Sciences, University of Texas, Dallas, TX
Department of Entomology, University of California, Riverside, CA
School of Journalism, Mass Communication & Media Arts, Southern Illinois University, Carbondale, IL
Department of Epidemiology and Public Health, University of Miami, Miami, FL
St. Tammany MAD, Louisiana
Department of Geography, Arizona State University, Tempe, AZ
The Transparency and Accountability Network (Tr-Ac-Net), New York, NY
Division of Infectious Diseases, Gorgas Center for Geographic Medicine, University of Alabama-Birmingham, Birmingham, AL
The Whitney Laboratory for Marine Biology, University of Florida, St. Augustine, FL
Division of Vector Borne Diseases, Ministry of Health, Nairobi, Kenya
West Coast Aerial Applicators

US National Earthquake Hazard Reduction Program (NEHRP)

- Further developing performance-based seismic design
- Improving techniques for evaluating and rehabilitating existing buildings
- Developing earthquake-resistant lifeline components and systems
- Developing cost-effective strategies for reducing earthquake impacts on the built environment
- Improving the disaster resilience of communities
- Developing the nation's human resource base in the earthquake safety field



Post-Earthquake Information Management System (PIMS) 2008 Scoping Study

- **PIMS Scope**
  - Data collection, organization, and storage;
  - Data curation and quality assurance;
  - Information presentation, discovery, and retrieval;
  - Privacy and security;
  - Long-term data preservation;
  - Data standardization;
  - System evolution and change management;
  - Coordination with public, private, and governmental sources;
  - Best-practices for managing sparse data
  - Community adoption of PIMS.

FEMA

National Institute of BUILDING SCIENCES

PIMS Input Data

- **Perishable Data:**
  - Product of field investigations after hazard events
  - Field investigators
    - Engineers
    - Scientists
    - Government workers
    - Citizens
  - All data geo- and time-referenced
  - Examples of Primary Data
    - Form data on PDA
    - Notes on PDA
    - Photos
    - Videos
    - Recorded interviews

PIMS Input Data

- **Non-perishable data:**
  - From harvesting of existing information
  - Examples
    - Maps
    - Drawings
    - Structure inventories
  - Sources
    - National databases
    - Public and private organizations
    - Emergency management agencies, public safety departments, community building departments

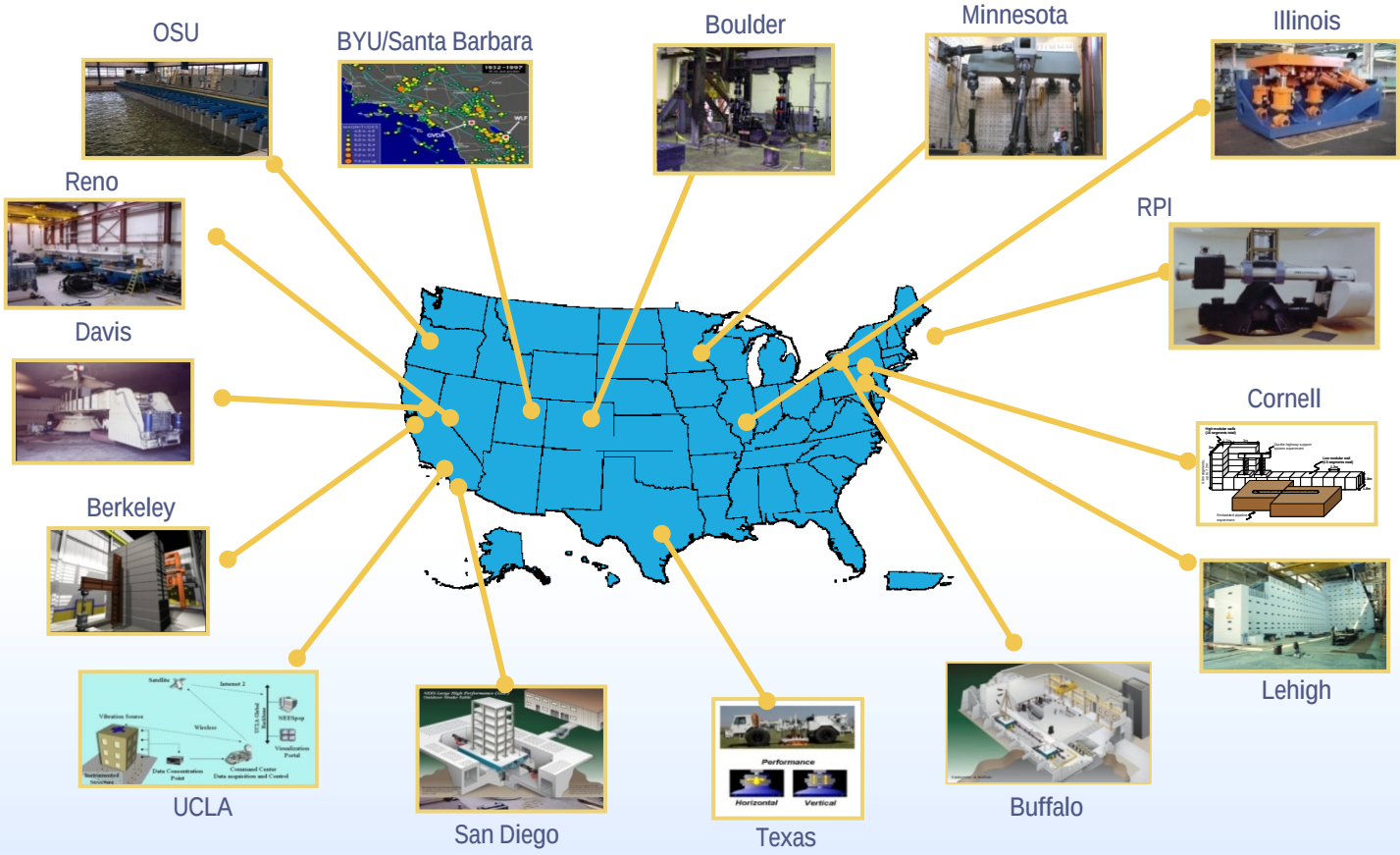
PIMS Envisioned End-Use

- **GIS-type interface**
- **Ability to extract lifelines/structures performance data based on:**
  - Location
  - Type of facility
  - Hazard level (ground shaking) experienced
  - Performance.

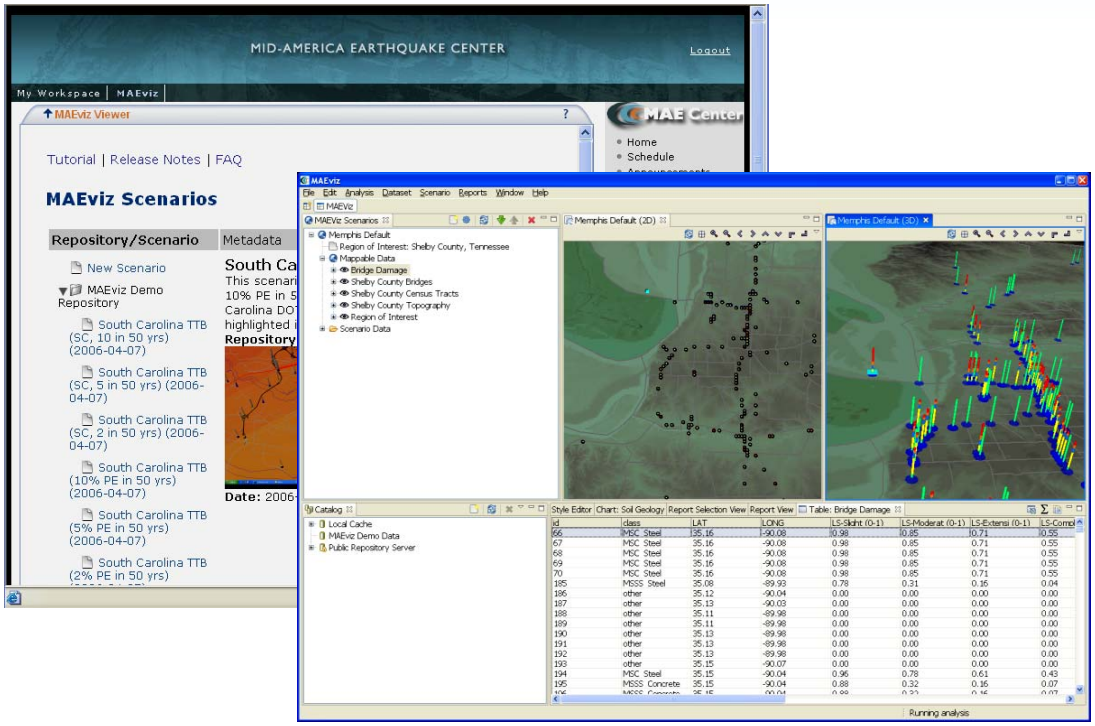
  For example -- What is the percentage of bridges with design feature X that have damage Y due to a hazard level of Z ?
- **Ability to output data in spreadsheets for statistical analysis**
- **Timeline for discussion: 50-100 years**

Network for Earthquake Engineering and Simulation (NEES)

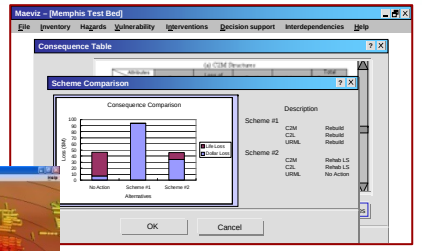
- *15 State-of-the-Art Shared Facilities*
- *NEES Central Data Repository*
- *Coordinated Cyberinfrastructure and Simulation Capabilities*

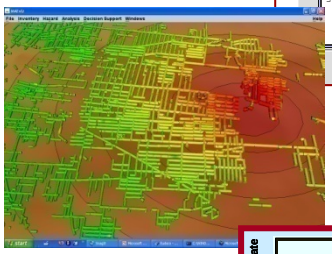


MAEViz: Consequence-Based Risk Management for Seismic Events



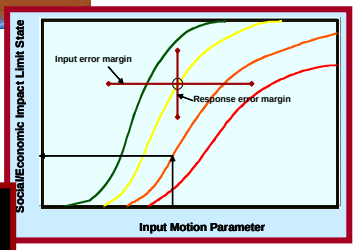
Decision Support

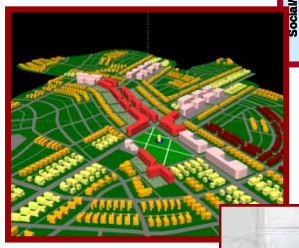




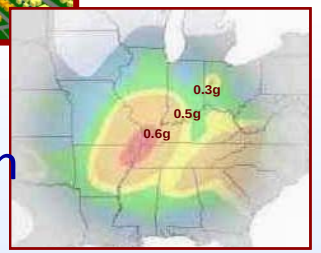
Damage Prediction

Fragility Models





Inventory Selection

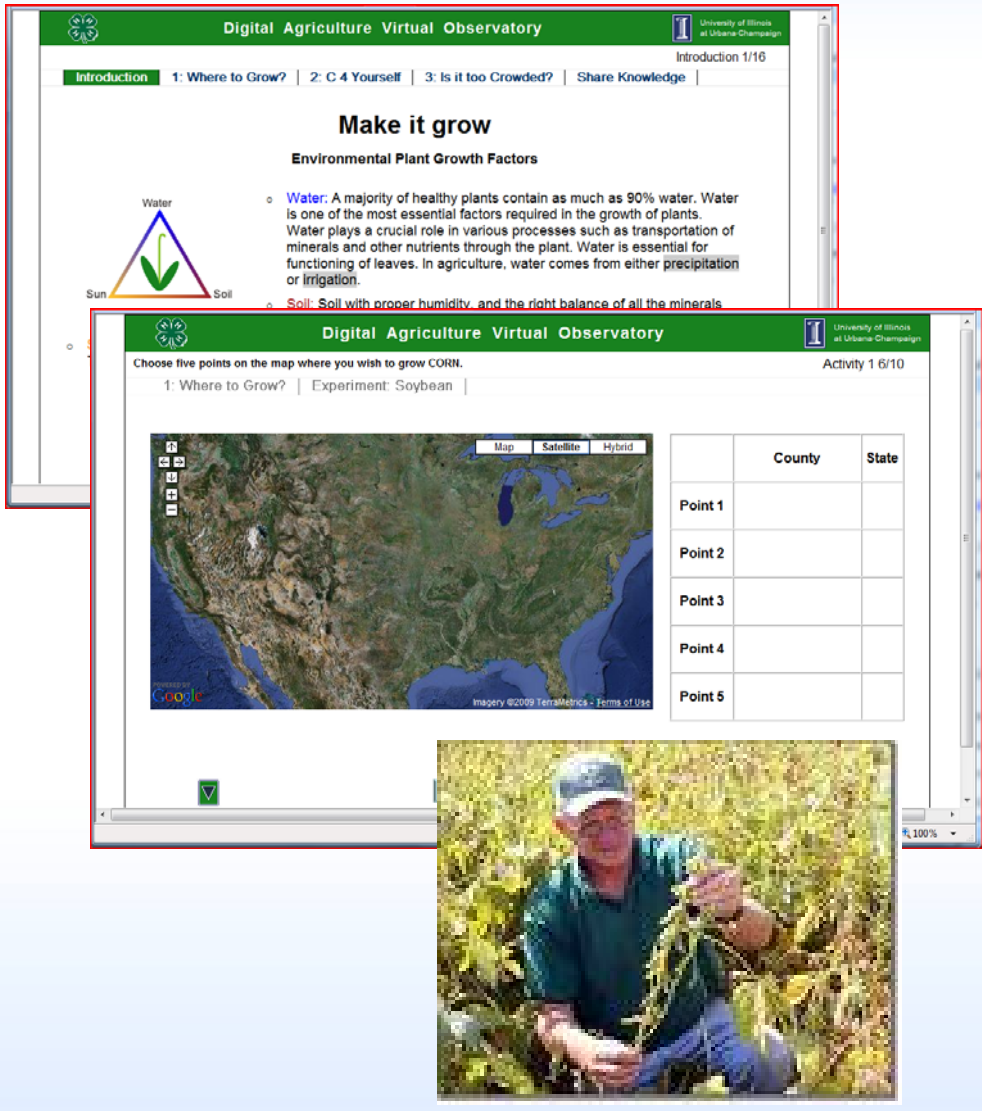


Hazard Definition

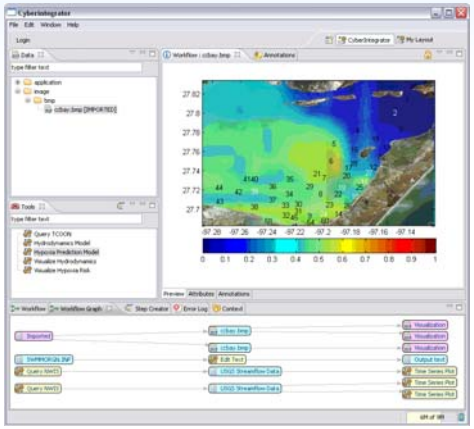
- Engineering View of MAE Center Research
- Physical through Socio-economic Analysis
- A "Cyberinfrastructure Aware" Application

Eco-physiological Modeling

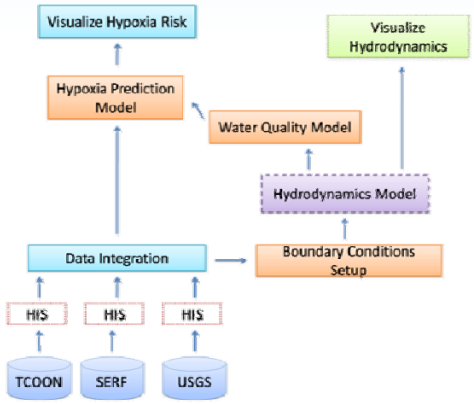
- Model from Xinguang Zhu and UI Institute for Genomic Biology
- Currently pursing educational use as a data-centric modeling service
- Potential connection to iPlantCollaborative.org

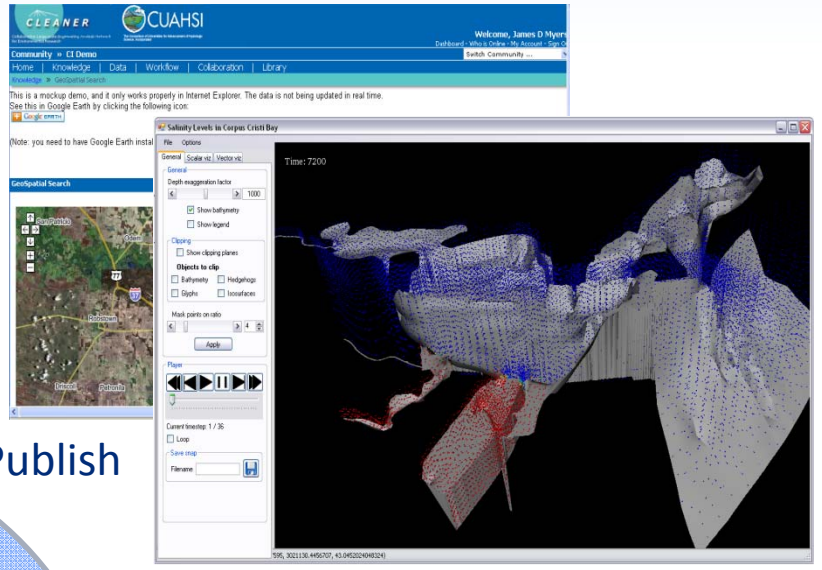


Digital Observatories

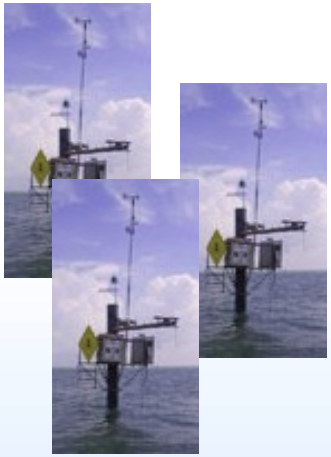


Model





Publish



Observe

From Basic Research to Societal Impact

Researchers

Policy Makers

Students

Citizens

Explore

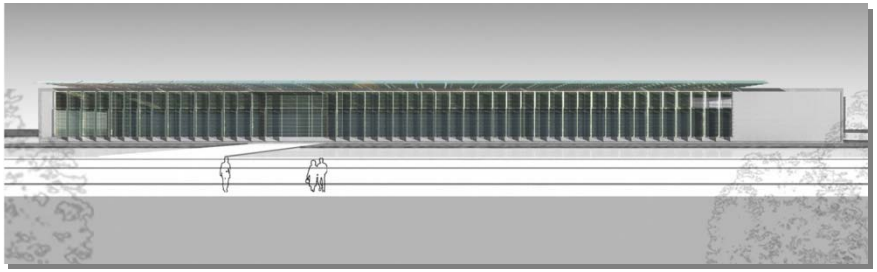
Understand

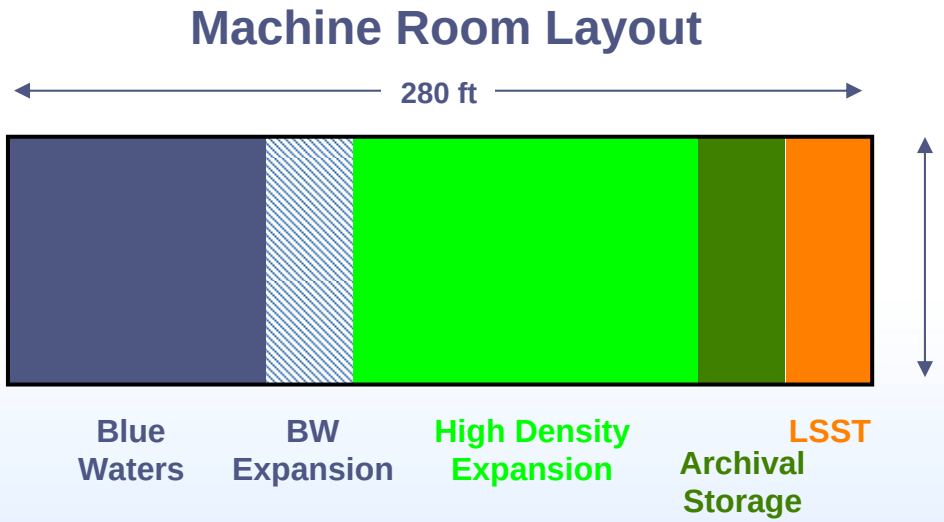
NCSA as a Data Center

- **College-level Unit of the University of Illinois**
  - Established in 1986 with funding from NSF and State of Illinois
  - One of two continuously funded NSF National Supercomputer Centers
  - Mission
    - Provide high-end computing resources to nation's scientists and engineers
    - Develop software needed to make full use of advanced computing systems
- **Staff**
  - 200+ Full time technical/professional staff
  - Students/postdocs/visiting scholars: varies
- **Computing Resources**
  - Six supercomputing systems: 144 TF
  - Archival storage system: 5 PB
  - Advanced visualization systems
  - Support > 2000 research groups

Blue Waters Petascale Computing System

- **Blue Waters**
  - NSF Flagship system
  - $208M
  - Multicore chips
  - >1 petaflop *sustained* performance
  - >200,000 cores
  - >800 terabytes of memory
  - >10 petabytes of user disk storage
  - On-line: July 2011





NCSA DIRECTORATES

PI	CET	CAC	ISL	AVL
PERSISTENT INFRASTRUCTURE	CYBER ENVIRONMENTS & TECHNOLOGIES	COMPUTING APPLICATIONS & COMMUNITIES	INNOVATIVE SYSTEMS LAB	ADVANCED VISUALIZATION LAB

← NCSA'S STRENGTHS ARE IN MULTI-DISCIPLINARY INTEGRATION →

← Astro, Bio, Geo, Enviro, Medical, Chemical, Humanites →

← PRIVATE & PUBLIC SECTORS →

← INTERNATIONAL & NATIONAL COMMUNITIES →

Common Characteristics Across Projects

- **Reference data required, but not sufficient**
- **Researchers**
  - Actively creating derived data products
  - Developing and publishing new methods
  - Interacting across disciplines to solve societal challenges

Data Sharing Implies Sharing

- **Community Data Curation - Contextualization**
- **Community Model Validation**
- **Community Resource (e.g. Data, Analysis Services) Publication**
- **Best-Practice Protocols**
- **Provenance Tracking and Reporting**
- **Event-Triggered Processing**

While recognizing

- **Data Heterogeneity**
- **Analysis Heterogeneity**
- **Resource (funding, infrastructure) Heterogeneity**
- **Social/Administrative Differences**
- **Temporal Mismatches**

- **Can we design realizing that these issues are inherent to research in science and engineering?**

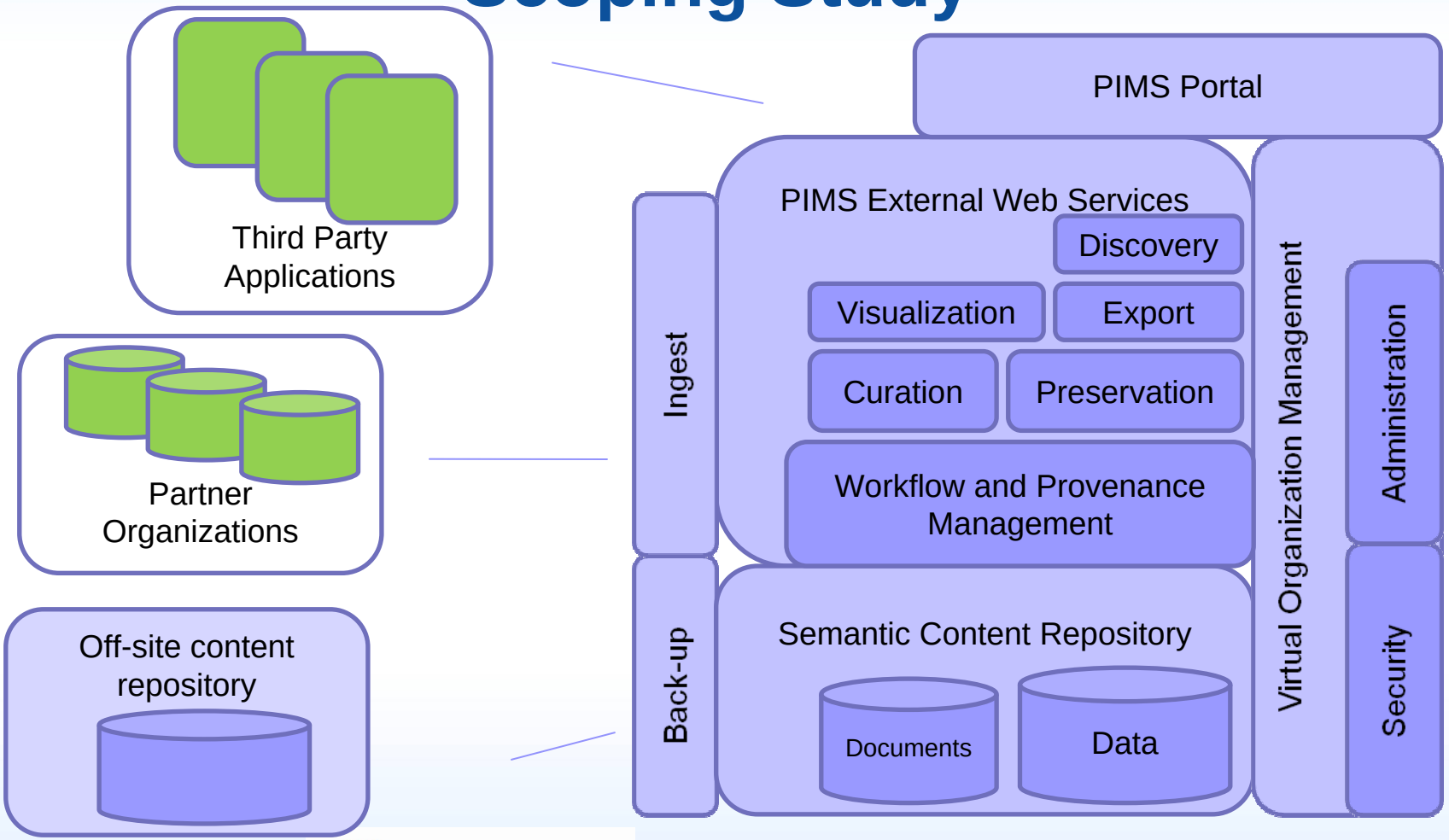
Key Design Concepts for Sharing Data Services

- **Explicit Separation of How from What:**
  - Content (type, global IDs, …) and Conceptual Context (metadata…)
  - Virtual Organizations/Social Networks (policies, resources, semantics, translation)
  - Process (workflow, provenance, …)
  - GUI Integration (portals, rich clients, …)
  - …
- **Ability to pass information through components that don't understand the details (everything is data)…**

…e-Science, Semantic Grid, Cyberenvironments, Web 2.0 …

…intelligence at the edges…

Post-Earthquake Information Management System (PIMS) 2008 Scoping Study



FEMA

National Institute of BUILDING SCIENCES

Conclusion

- Looking for opportunities driven by cross-disciplinary research efforts
- Designing to support end-to-end data management needs
- Supporting researchers in managing heterogeneous and evolving data and processes

- Opportunities in infrastructure development, community scale efforts, and data-intensive collaborative projects


Questions?