Cyberinfrastructure Framework for 21st Century Science and Engineering "CF21"

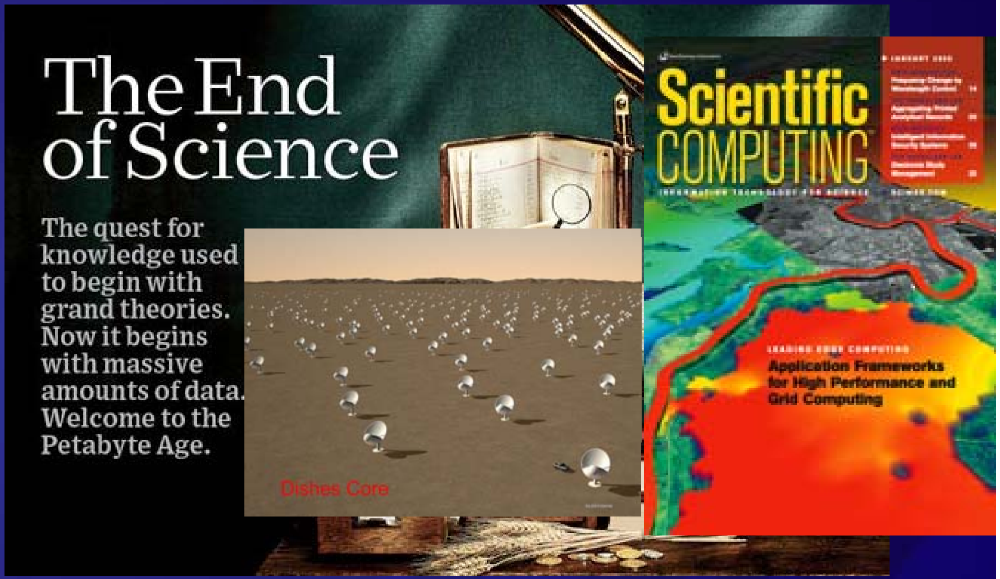
Dr. Jennifer M. Schopf
NSF, Office of CyberInfrastructure

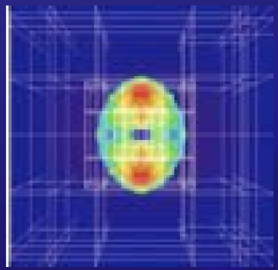
- This is Ed Seidel’s vision, with additional material compliments of
  - Alan Blatecky
  - José Muñoz
  - The rest of the NSF OCI office
  - Several ACCI members
  - Many conversations

Framing the Question

Science has been Revolutionized by CI

- Modern science
  - Data- and compute-intensive
  - Integrative
- Multiscale Collabs
  - Add'l complexity
  - Individuals, groups, teams, communities
- Must Transition NSF CI approach to address these issues





Along with *The Fourth Paradigm,* an emerging science of environmental applications

1. Thousand years ago — experimental science
   - Description of natural phenomena
2. Last few hundred years — theoretical science
   - Newton's Laws, Maxwell's Equations . . .
3. Last few decades — computational science
   - Simulation of complex phenomena
4. Today — data-intensive science

(from Tony Hey)

1. 1800s → ~1990 — discipline oriented
   - geology, atmospheric science, ecology, etc.
2. 1980s → present — Earth System Science
   - interacting elements of a single complex system (Bretherton)
   - large scales, data intensive
3. Emerging today — knowledge created to target practical decisions and actions
   - e.g. climate change
   - large scales, data intensive

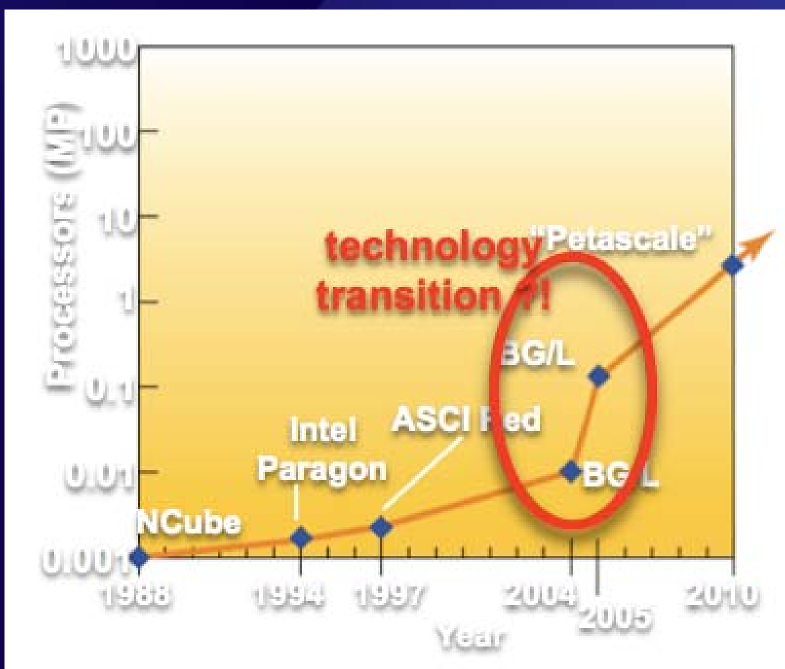
Jeff Dozier, University of California, Santa Barbara

Outline

- The Crises leading up to CF21
- A Cyberinfrastructure Ecosystem
  - Taskforces
- Upcoming Programs
  - Sustain
  - Advance
  - Experiment

Five Crises

- Computing Technology
  - Multicore: processor is new transistor
  - Programming model, fault tolerance, etc
  - New models: clouds, grids, GPUs,... where appropriate



Options on how to spend computing power

Model complexity

Temporal & spatial resolution

Scenarios, parameterizations, time period

Five Crises (cont)

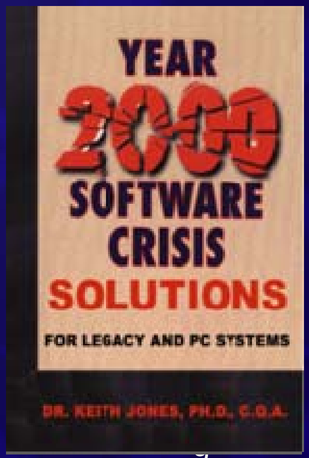
- Data, provenance, and viz
  - Generating more data than in all of human history: preserve, mine, share?
  - Archiving is hard – yet easy compared to curation
  - What about tracking data provenance?
  - How do we create "data scientists"?

Enormous, Irreplaceable Data Sets

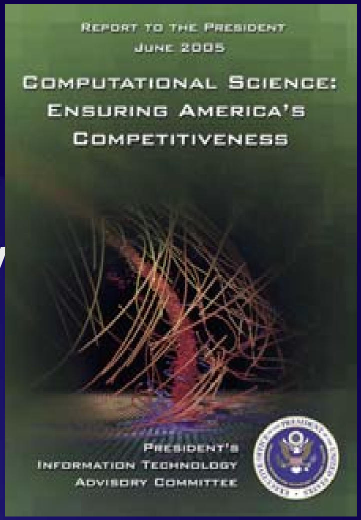
neon NATIONAL ECOLOGICAL OBSERVATORY NETWORK	~ 150 TB/Year
LSST Large Synoptic Survey Telescope	~ 30 TB/Night
Large Hadron Collider	~ 15 PB/Year
NASA E • O • S • D • I • S Earth Observing System Data & Information System	~ 64 TB/Year
earthscope	~ 40 TB/Year
Long tail of small science	?? TB/Year

Five Crises (cont)

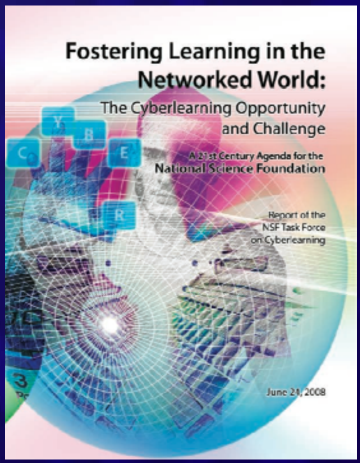
- Software
  - Complex applications on coupled compute-data-networked environments, tools needed
  - Modern apps: 10^6+ lines, many groups contribute, take decades
  - Science has become un-reproducible



Five Crises (cont)

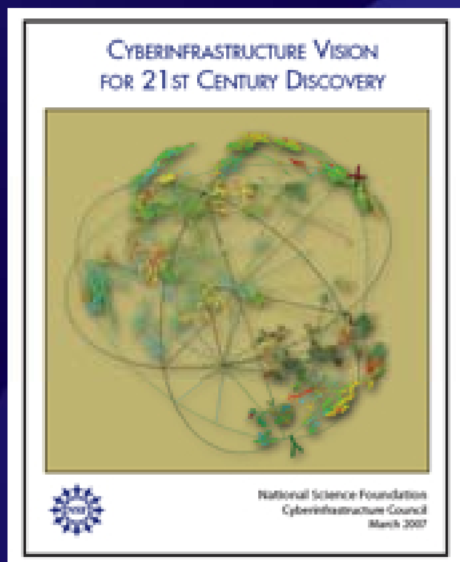


- Lack of Organizations for Multidisciplinary Computational Science
  - "*Universities must significantly change organizational* structures*: multidisciplinary & collaborative research are needed [for US] to remain competitive in global science*"
  - *"Itself a discipline, computational science advances all science…inadequate/outmoded structures within Federal government and the academy do not effectively support this critical multidisciplinary field"*
- Education- and the next generation
  - The CI environment is running away from us!
  - How do we develop a workforce to work effectively in this world?
  - How do we help universities transition?



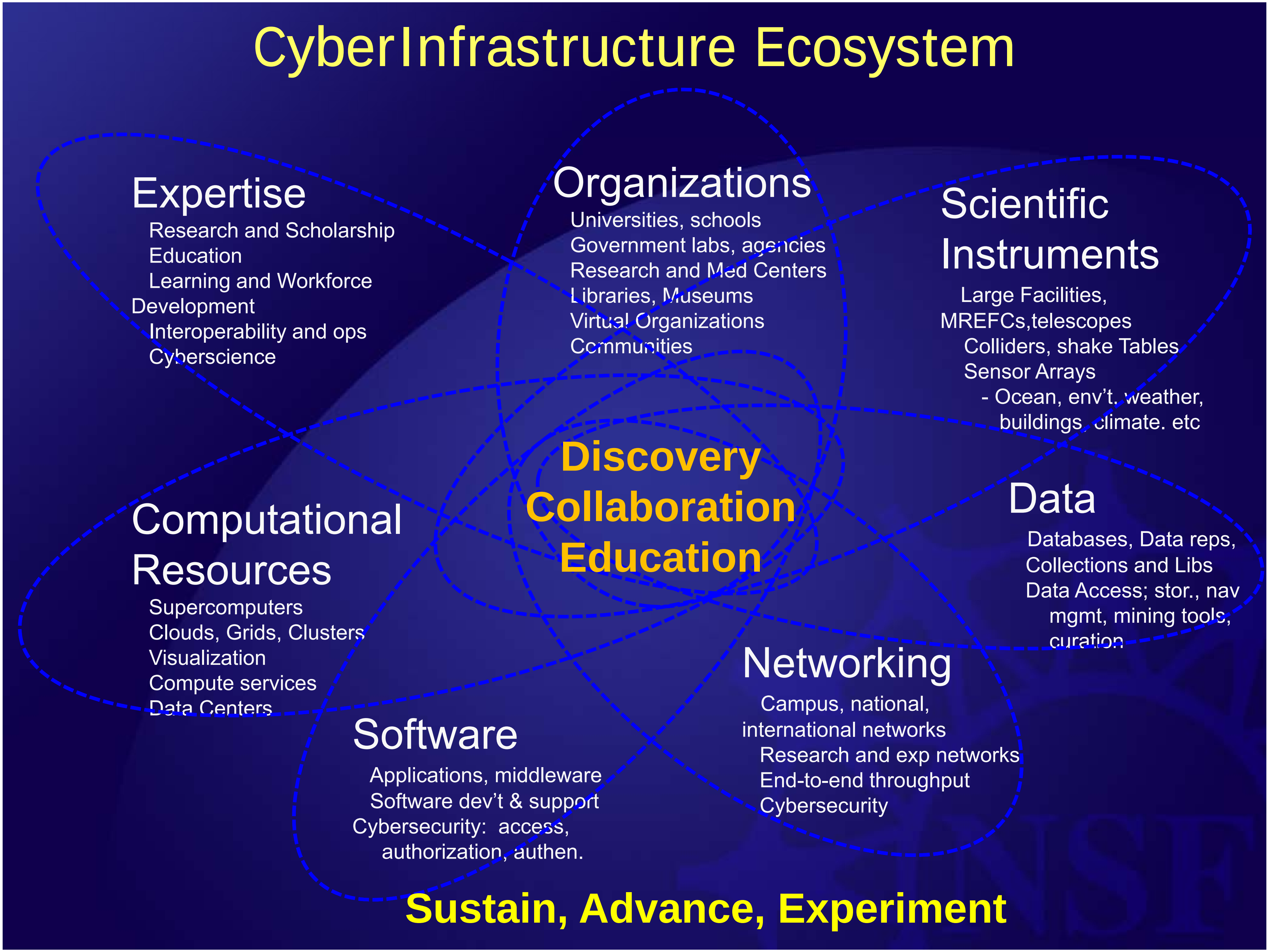
What is Needed?

An ecosystem, not components…



NSF-wide CI *Framework* for 21st Century Science & Engineering

People, Sustainability, Innovation, Integration


CyberInfrastructure Ecosystem
Expertise
Research and Scholarship
Education
Learning and Workforce Development
Interoperability and ops
Cyberscience
Organizations
Universities, schools
Government labs, agencies
Research and Med Centers
Libraries, Museums
Virtual Organizations
Communities
Scientific Instruments
Large Facilities, MREFCs,telescopes
Colliders, shake Tables
Sensor Arrays
- Ocean, env't, weather, buildings, climate. etc
Discovery
Collaboration
Education
Computational Resources
Supercomputers
Clouds, Grids, Clusters
Visualization
Compute services
Data Centers
Data
Databases, Data reps,
Collections and Libs
Data Access; stor., nav mgmt, mining tools, curation
Networking
Campus, national, international networks
Research and exp networks
End-to-end throughput
Cybersecurity
Software
Applications, middleware
Software dev't & support
Cybersecurity: access, authorization, authen.
Sustain, Advance, Experiment
NSF

“We seek solutions. We don't seek—dare I say this?—just scientific papers anymore”

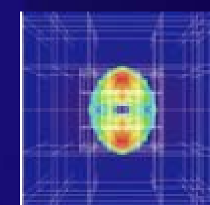
– Steven Chu
Nobel Laureate
US Secretary of Energy

Compliments Jeff Dozier, University of California, Santa Barbara

CF21: Cyberinfrastructure Framework...

- High-end computation, data, visualization for transformative science
  - Facilities/centers as *hubs of innovation*

- MREFCs and collaborations including large-scale NSF collaborative facilities, international partners



- Software, tools, science applications, and VOs critical to science, integrally connected to instruments

- Campuses fundamentally linked end-to-end; grids, clouds, loosely coupled campus services, policy to support

- People Comprehensive approach workforce development for 21st century science and engineering

Some observations

- Science and Scholarship are team sports
  - Competitiveness and success will come to those who can put together the best team, and can marshal the best resources and capabilities
- Collaboration/partnerships will change significantly
  - Growth of dynamic coalitions and virtual organizations
  - International collaboration will become even more important
- Ownership of data plus low cost fuels growth and number of data systems
  - Growth in both distributed systems and local systems
  - More people want to access more data
  - Federation and interoperability become more important

More observations

- More discoveries will arise from search approaches
  - Mining vast amounts of new and disparate data
  - Collaboration and sharing of information
- Mobility and personal control will continue to drive innovation and business
- Gaming, virtual worlds, social networks will continue to transform the way we do science, research, education and business
- The Internet has collapsed six degrees of separation and is creating a world with two or three degrees.

ACCI
Task Forces

Campus Bridging
Craig Stewart

Data (Viz)
Shenda Baker
Tony Hey

Software
David Keyes

Computing (Clouds Grids)
Thomas Zacharia

Education Workforce
Alex Ramerez

Grand Challenge VOs
Tinsley Oden

- Timelines: 12-18 months
- Advising NSF
- Workshop(s)
- Recommendations
- Input to NSF informs
  - CF21 programs
  - 2011-2 CI Vision Plan

Preliminary Task Force (TF) Results

- Computing TF Workshop Interim Report
  - Rec: Address sustainability, people, innovation
- Software TF Interim Report
  - Rec: Address sustainability, create long term, multi-directorate, multi-level software program
- GCC/VO TF Interim Report
  - Rec: Address sustainability, OCI to nurture computational science across NSF units
- Software Sustainability WS (Campus Bridging)
  - Rec: Open source, use sw eng practices, reproducibility

Innovation vs Sustainability

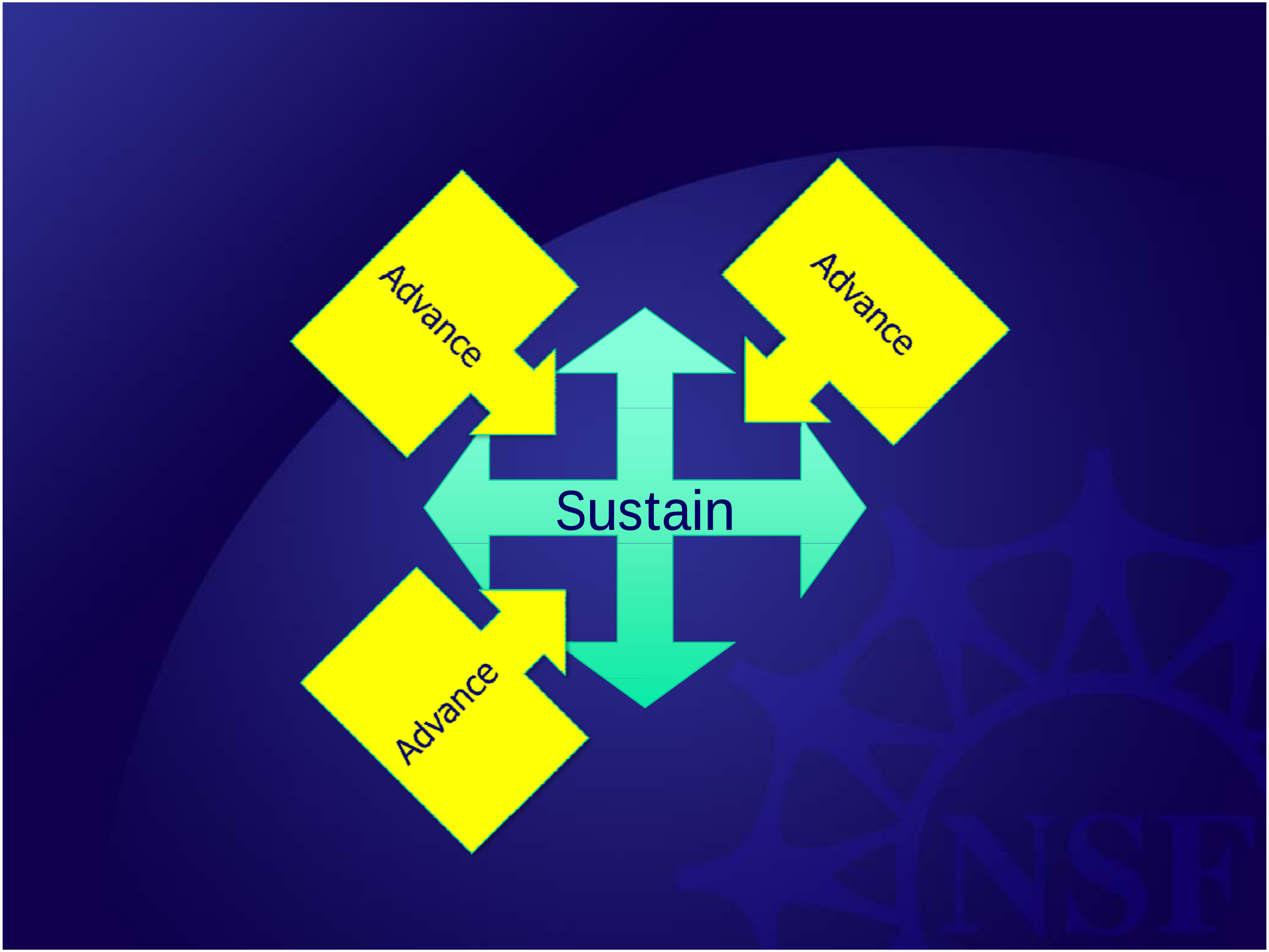
- Tension between:
  - Bleeding edge & tried and true
  - Novel and new & dependable
  - Might have a new way & method that always works

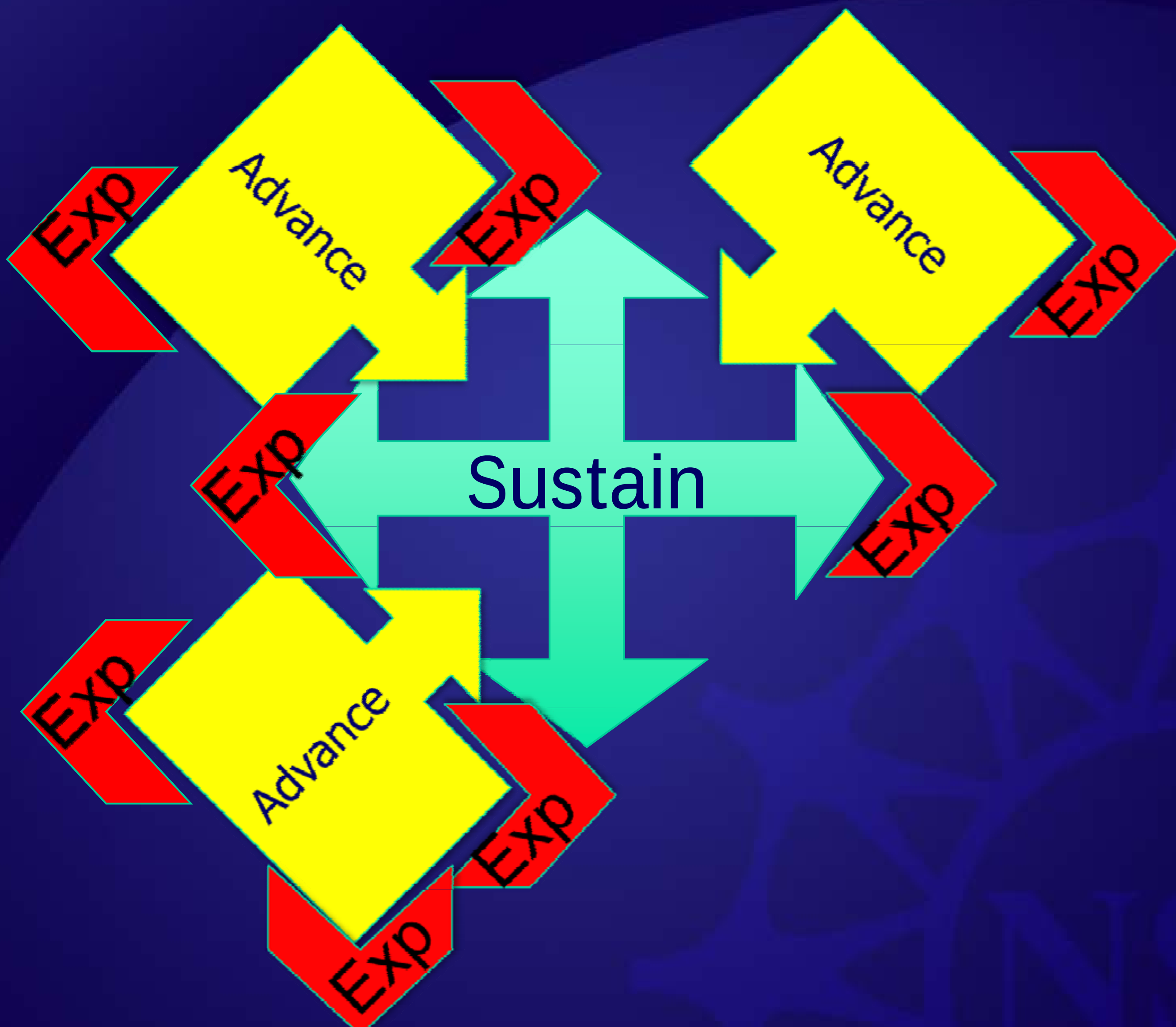
- We need a spectrum of approaches
  - Allow broad scale innovation
  - Continue to advance approaches
  - Yet sustain scientific disciplines

Over-arching Approach For Upcoming Programs

- Sustain
  - Large-scale “Institute”-style projects to promote long term approaches
  - Long term (5+ years), many PIs, and institutions
  - Highly multi-disciplinary, perhaps multi-agency
- Advance
  - Medium-scale collaborative teams to harden and expand successful experiments
  - Collaborative teams, multi-year (3-5)
- Experiment
  - Smaller scale, trials of new approaches


Sustain


Advance
Advance
Sustain
Advance
NSF


Advance
Exp
Exp
Advance
Exp
Exp
Sustain
Exp
Exp
Advance
Exp
Exp
Exp

CF21 Software Infrastructure for Sustained Innovation (SI2)

- Significant multiscale, long-term software program
  - Perhaps $200-300M over a decade
    - $10M identified in FY10 ($4M OCI/$6M Dirs)
    - $14M annual in OCI in future years
      - Catalyze significant funds from Dirs
- **Sustain**: Connected institutes, teams, investigators
  - Integrated into CF21 framework w/Dirs
  - 3-6 centers, 5+5 years, for critical mass, sustainability
- **Advance**: Numerous teams of scientists and computational and computer scientists with longer term grants
- **Experiment**: Many individuals w/short term grants, funded by OCI and directorates

Software, continued

- Ongoing discussions to build this program across NSF
  - Some of the institutes will be discipline specific
  - Some may be algorithm/tool themed (e.g., data, provenance, viz)
  - All should be fundamental to other programs (e.g., SEES)
  - Education, science applications, industrial partners linked deeply
- MREFC's, other large facilities need to participate
  - iPlant, NEON, LSST, etc...

What does Sustainability mean in the context of software?

- "Ability to maintain a certain process or state"
- In a biological context
  - Resources must be used at a rate at which they can be replenished
- In a software context
  - Creating software that can be used in broad contexts (reuse)
  - Funding models that encourage long-term support (beyond normal NSF grants)

Note: I'm defining software VERY broadly – everything in your environment, middleware, tools, numerical libraries, application codes, etc.)

One Future: Software As Infrastructure

- NSF should fund software sustainably the same way it does other infrastructure.
  - Same as telescopes, colliders, or shake tables
  - Line items in the directorate budgets
  - Constant or growing over time, reliably
  - Factor in “maintenance” and “replacement”
  - Eligible for programs like MRI and ARI
- Software is around even longer than hw
  - Hardware refresh ~3 years
  - Software can grow over decades
  - (what’s the right funding ratio of sw to hw in a large-scale CI project?)

However, if software is viewed as infrastructure by NSF...

- PIs must also treat it as such
  - Reliable, robust, reproducible, production-quality software
  - Reporting requirements (including uptime, usage statistics, and safety/security reporting)
  - Formal planning approach- including scheduling/estimation, requirements development, deployment plans, risk assessment, etc.
  - Teams with "professional engineering" backgrounds
- This program is a step in the right direction

Open Source

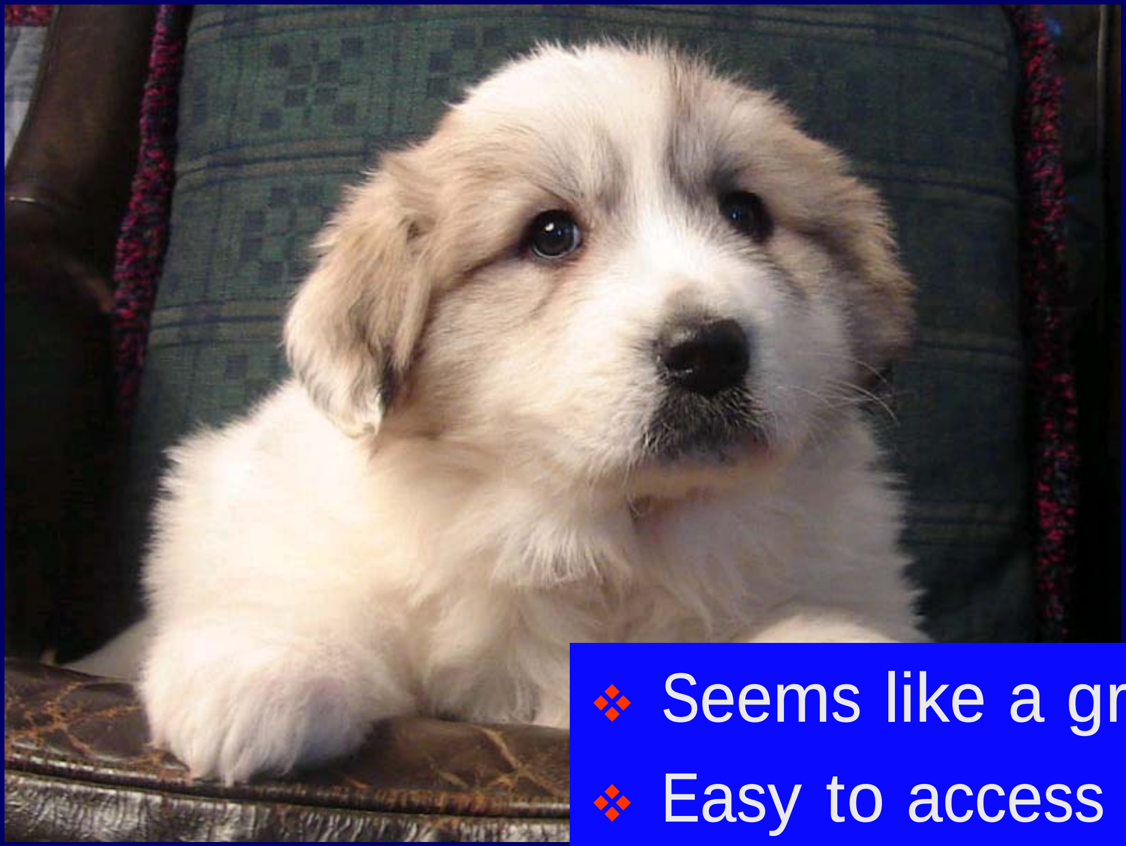
- Requirement for all current OCI programs
  - And many others across NSF
- Strongly encourages reuse
- Some people think simply open source is enough – it's not
- Necessary but not sufficient for sustainable software

Open Source software is free…

Free as in speech…

free as in beer, or…

Open Source Software is Like a Free Puppy

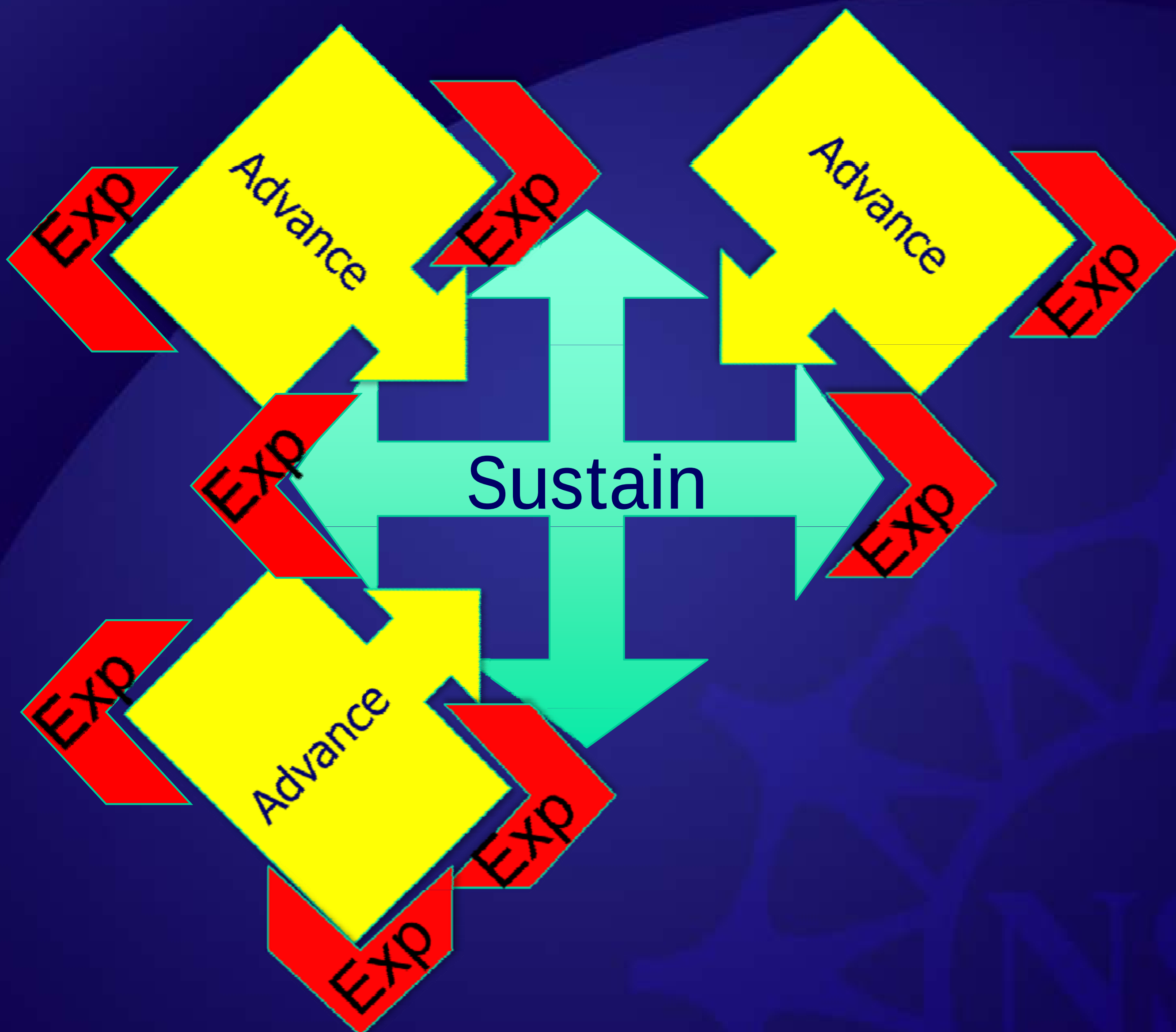


- Seems like a great bargain
- Easy to access
- Can catch you eye at a weak moment
- ...but sometimes more than you expected

- Long term costs
- Needs love and attention
- May lose charm after growing up
- Occasional clean-ups required
- Many left abandoned by their owners
- May not be quite what you think

For Sustainability to Work

- Fundamental change in culture for both development groups and funders
- Software institutes looking at an approach to sustain, advance and experiment is a first step


Advance
Advance
Advance
Exp
Exp
Exp
Exp
Exp
Exp
Exp
Exp
Exp
Sustain

Data Programs

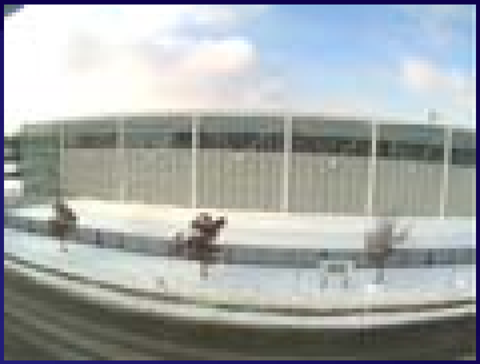
- DataNet: OCI Flagship Data Program
  - Focus on data-level interoperability and data preservation
- **Sustain**: 5 Centers, $20M, 5years (+5)
- **Advance**: eg. SDCI awards
  - ~3-4 year, $1-2M, support of data tools for broad set of applications and disciplines
- **Innovate**: eg. InterOp awards
  - Smaller scale, innovative use of data for new communities

2008 DataNet Awards

- DataNet Observation Network for Earth (PI: Michener)
  - Facilitates research on climate change and biodiversity, integrating earth observing networks
  - Emphasis on user community engagement, promote data deposition and re-use
  - Science question: What are the relationships among population density, atmospheric nitrogen, CO_2, energy consumption and global temps?
- Data Conservancy (PI: Choudhury)
  - Integrates observational data to enable scientists to identify causal and critical relationships in physical, biological, ecological, and social systems
  - User centered design paradigm, ethnographic studies
  - Science question: How do land and energy use in mega-cities impact the carbon cycle and climate change?

Planned CF21 HPC Program

- **Sustain**: Petascale-to-Exascale
  - 1-2 Sustainable facilities (~$200M+)
  - Likely NSF-DOE cooperation
  - 10 years (5+5)

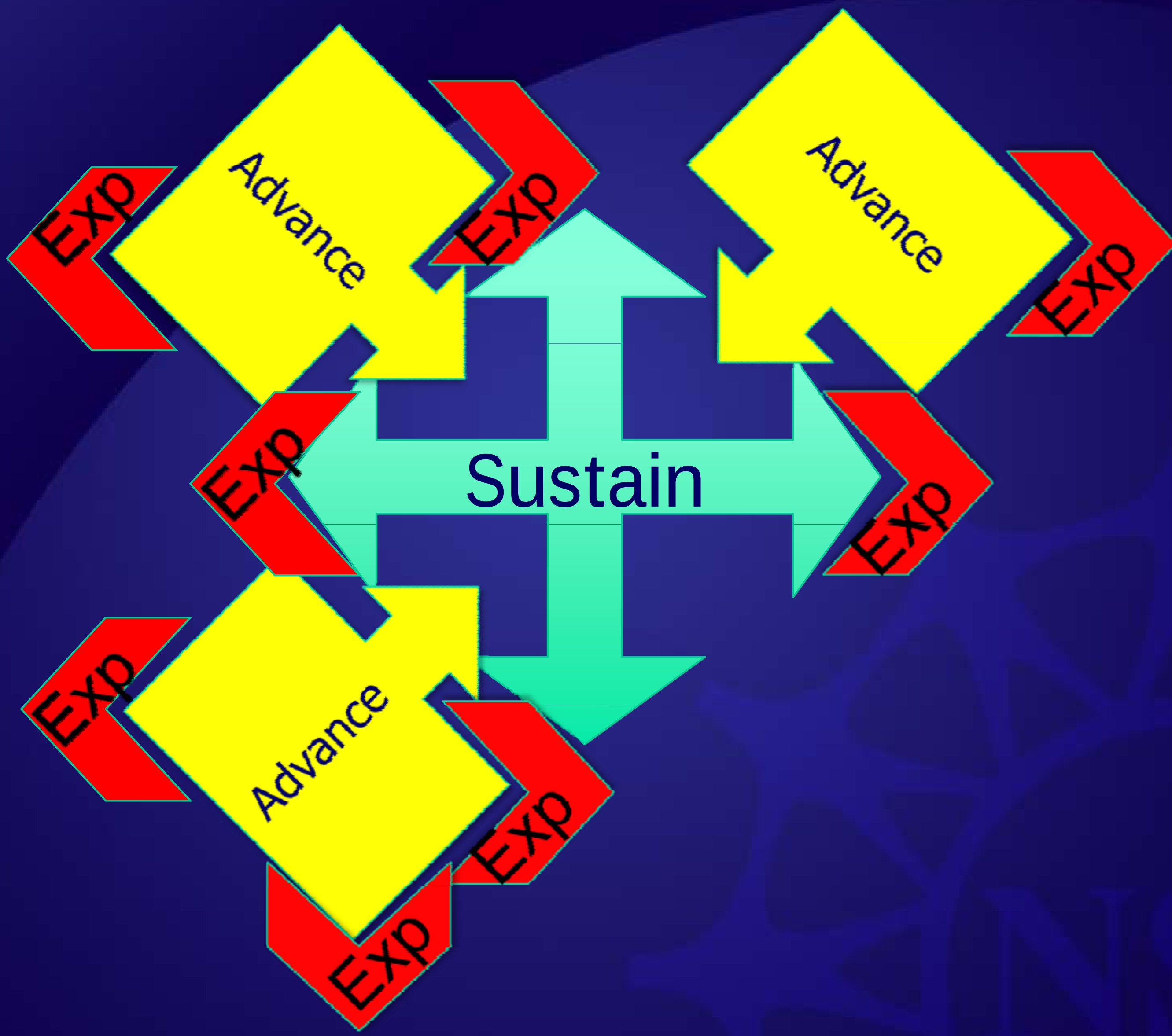


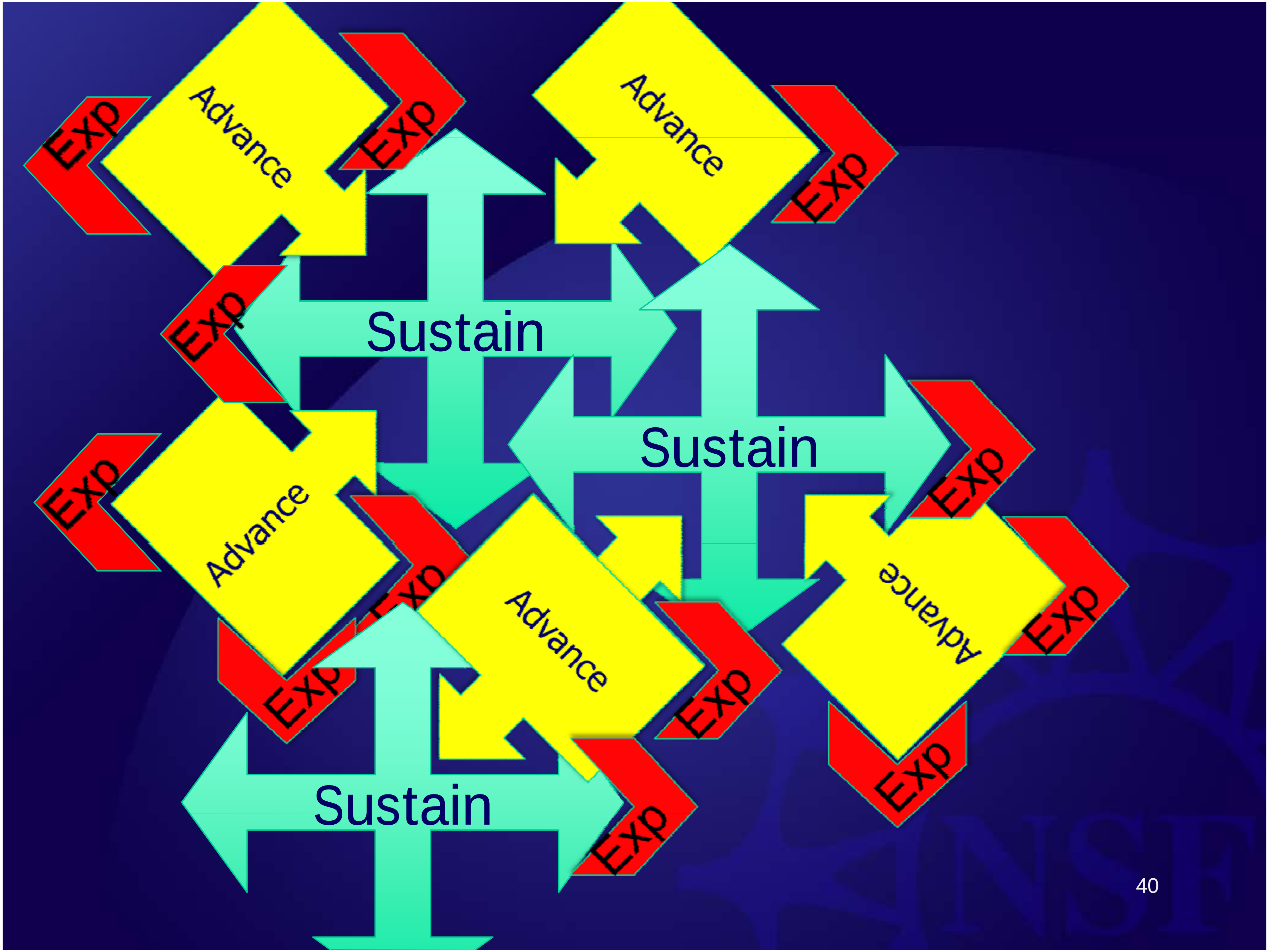
UIUC Petascale Facility:
$60M building!

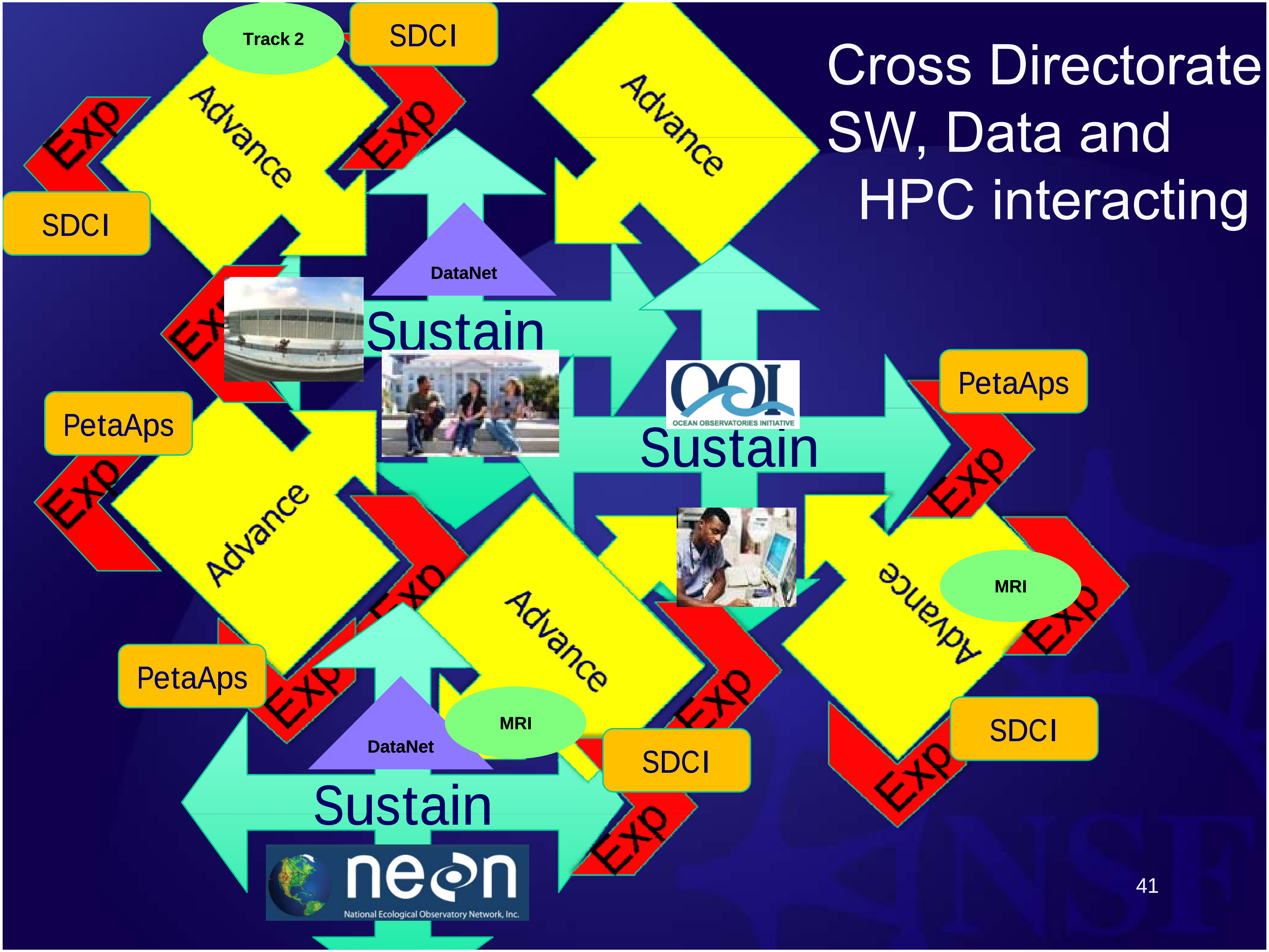
- **Advance**
  - 4-5 hubs of Excellence/Innovation, people, expertise
  - Mixture of data and compute-intensive centers, supporting broader array of services
- **Experiment**
  - Explore new architectures, couple with application/software dev

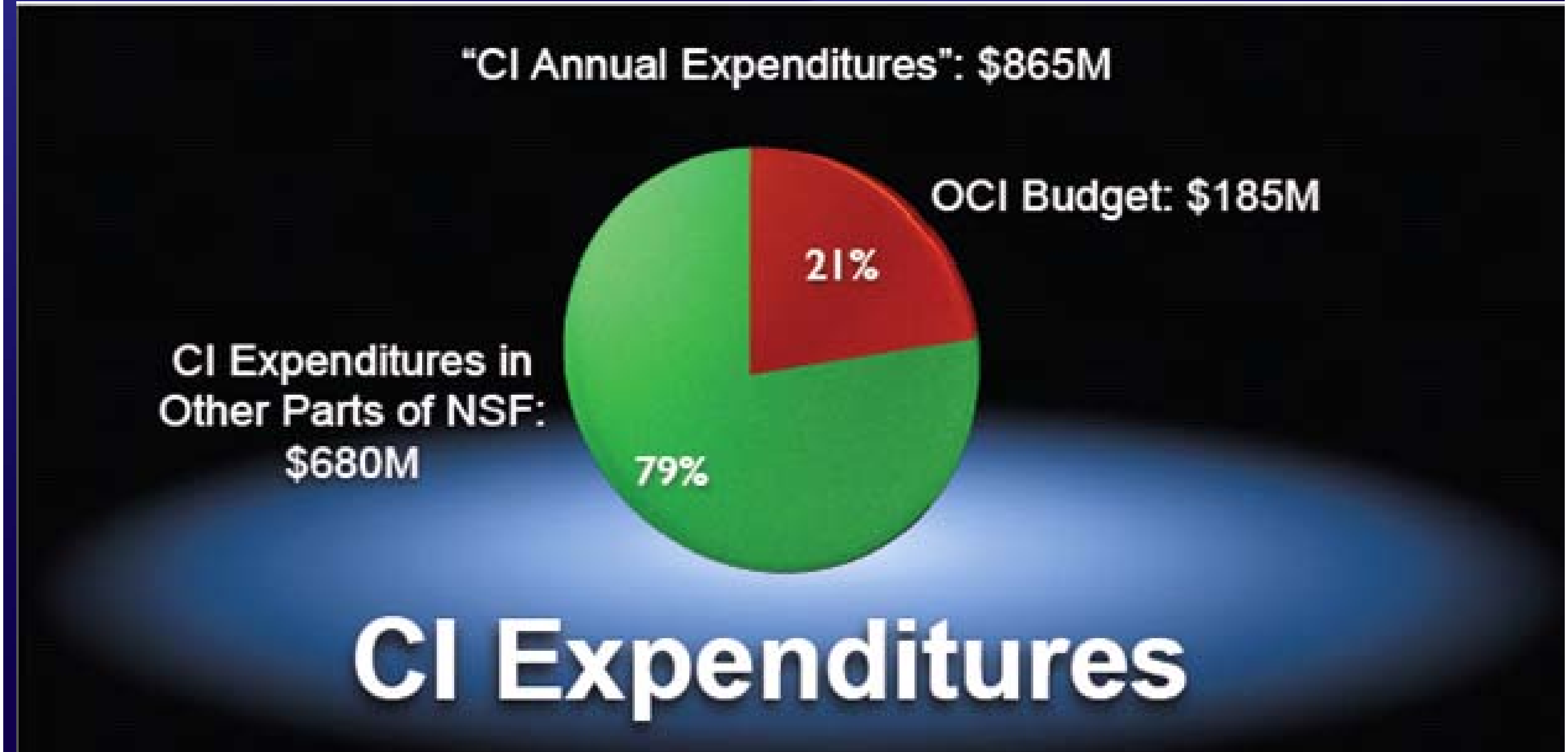
HPC Will Also Need

- Discipline specific connections
  - MRI, Divisional, Directorate programs can be aligned to connect in to this NSF-wide structure
    - Recommended common software, identity management, policy
    - Data, software sharing
- How does eXtreme Digital (XD), TeraGrid Phase 3 fit in?
  - Competition underway now
  - Foundation to build broader CF21 services in future


Advance
Exp
Exp
Advance
Exp
Sustain
Exp
Exp
Advance
Exp
Exp
Exp


Advance
Advance
Advance
Advance
Advance
Exp
Exp
Exp
Exp
Exp
Exp
Exp
Exp
Exp
Exp
Exp
Exp
Sustain
Sustain
Sustain

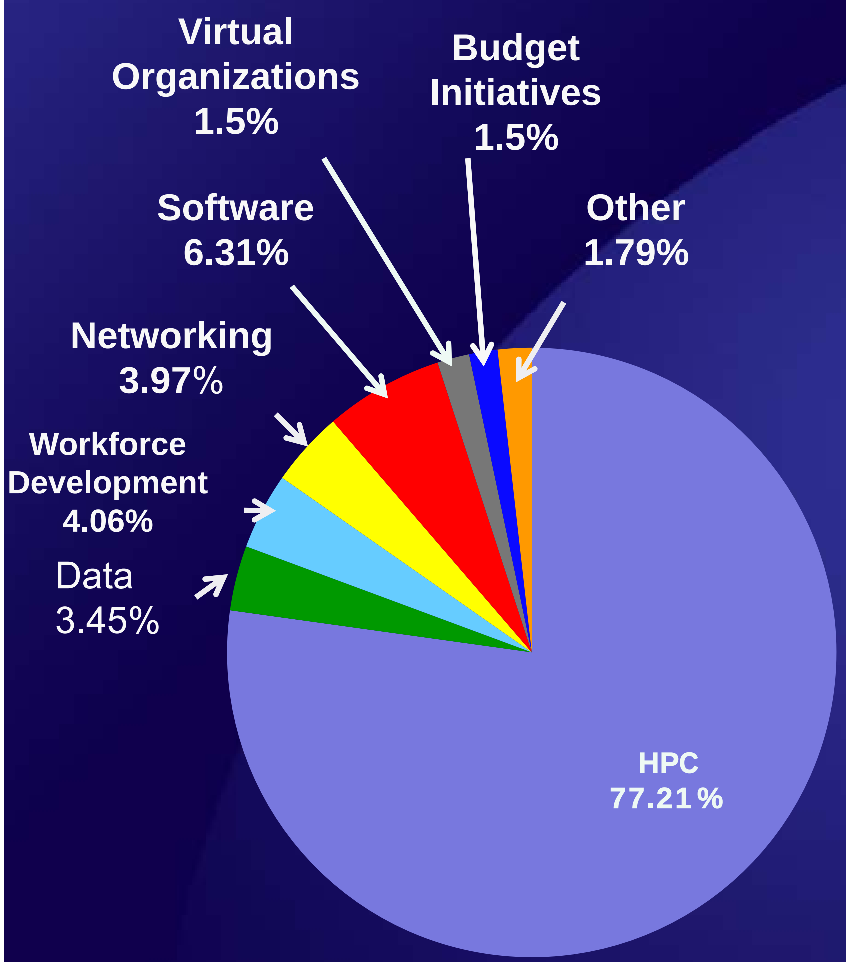

Cross Directorate SW, Data and HPC interacting
Track 2
SDCI
Advance
Exp
SDCI
DataNet
Sustain
OOI
OCEAN OBSERVATORIES INITIATIVE
PetaAps
MRI
DataNet
Sustain
neon
National Ecological Observatory Network, Inc.


"CI Annual Expenditures": $865M
OCI Budget: $185M
21%
CI Expenditures in Other Parts of NSF: $680M
79%
CI Expenditures

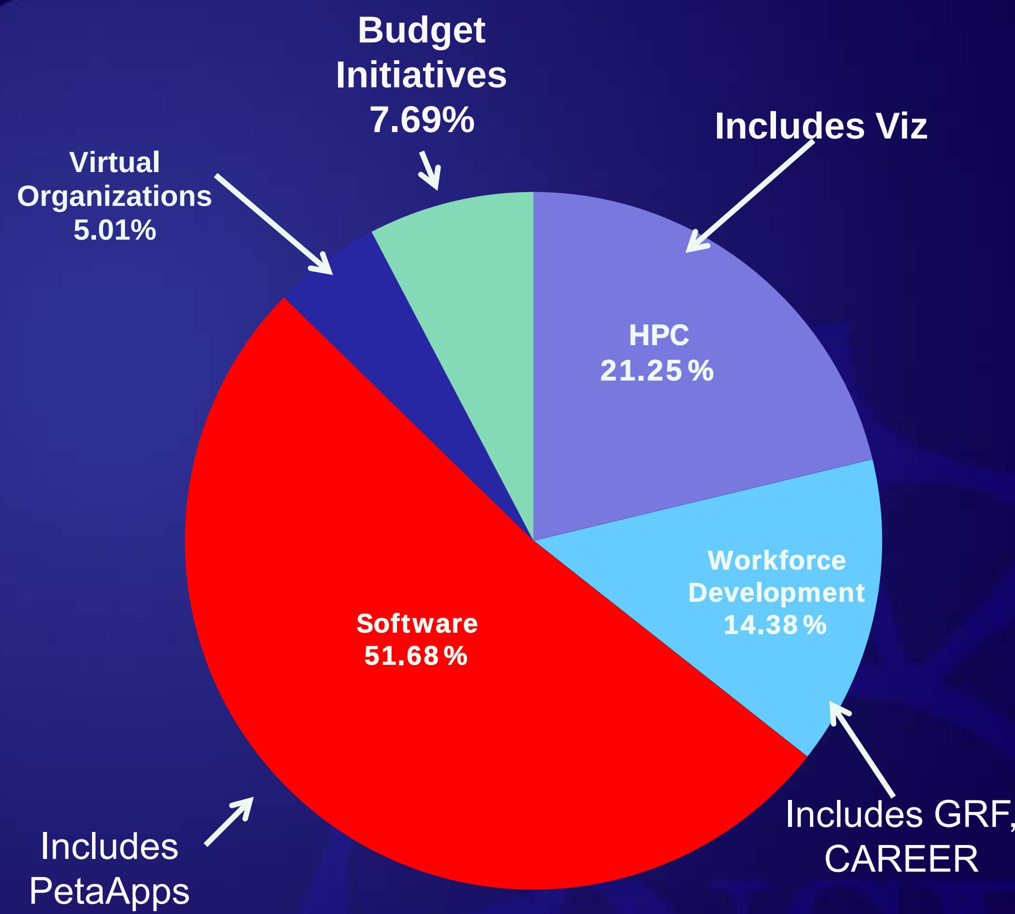
Outside of SW, Data, and HPC

- Postdoc program: CITracs
  - Emphasis on helping computational scientists learn about CI or vice versa
  - http://www.nsf.gov/pubs/2010/nsf10553/nsf10553.htm
- CI-TEAM: Training, Education, Advancement, and Mentoring for Our 21st Century WF
  - Prepare current and future generations of scientists, engineers, and educators
  - Design, develop, adopt and deploy cyber-based tools and environments for research and learning, both formal and informal
  - http://www.nsf.gov/pubs/2010/nsf10532/nsf10532.pdf

ARRA Catalyzed OCI Transition

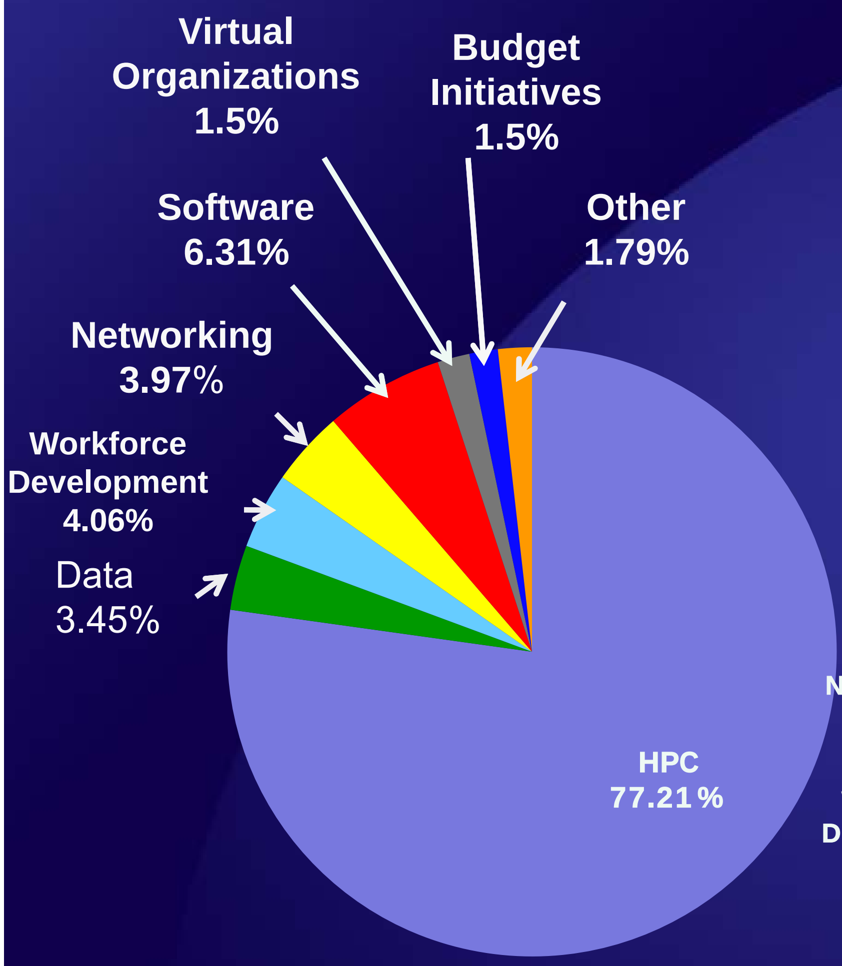


FY 09 Budget (Before ARRA)

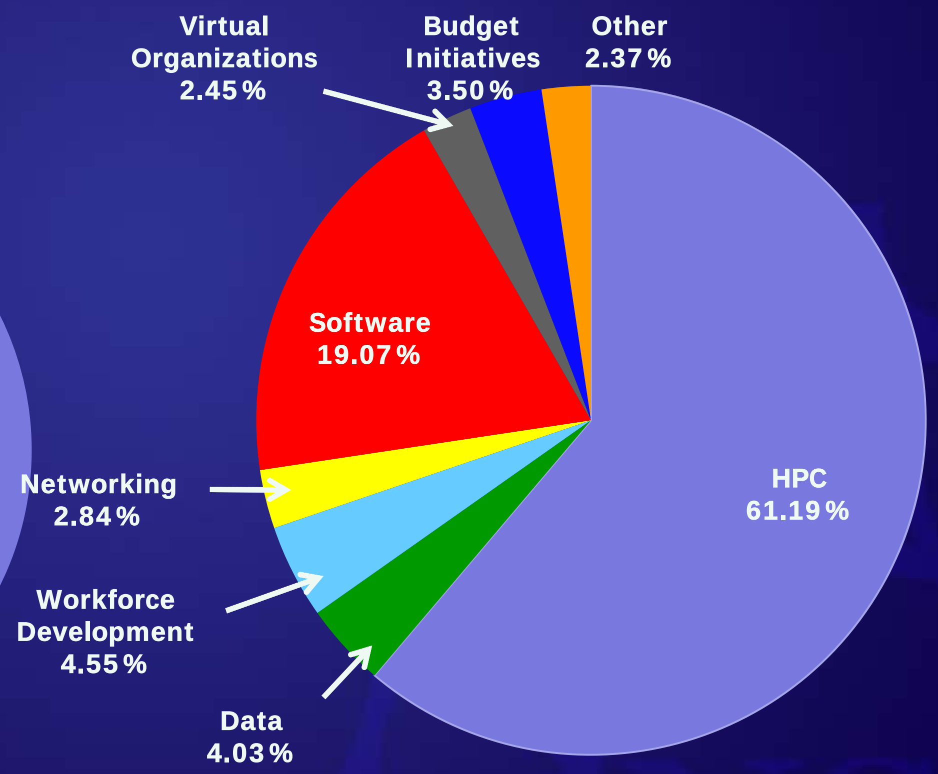


Recovery Act Funds

ARRA Catalyzed OCI Transition

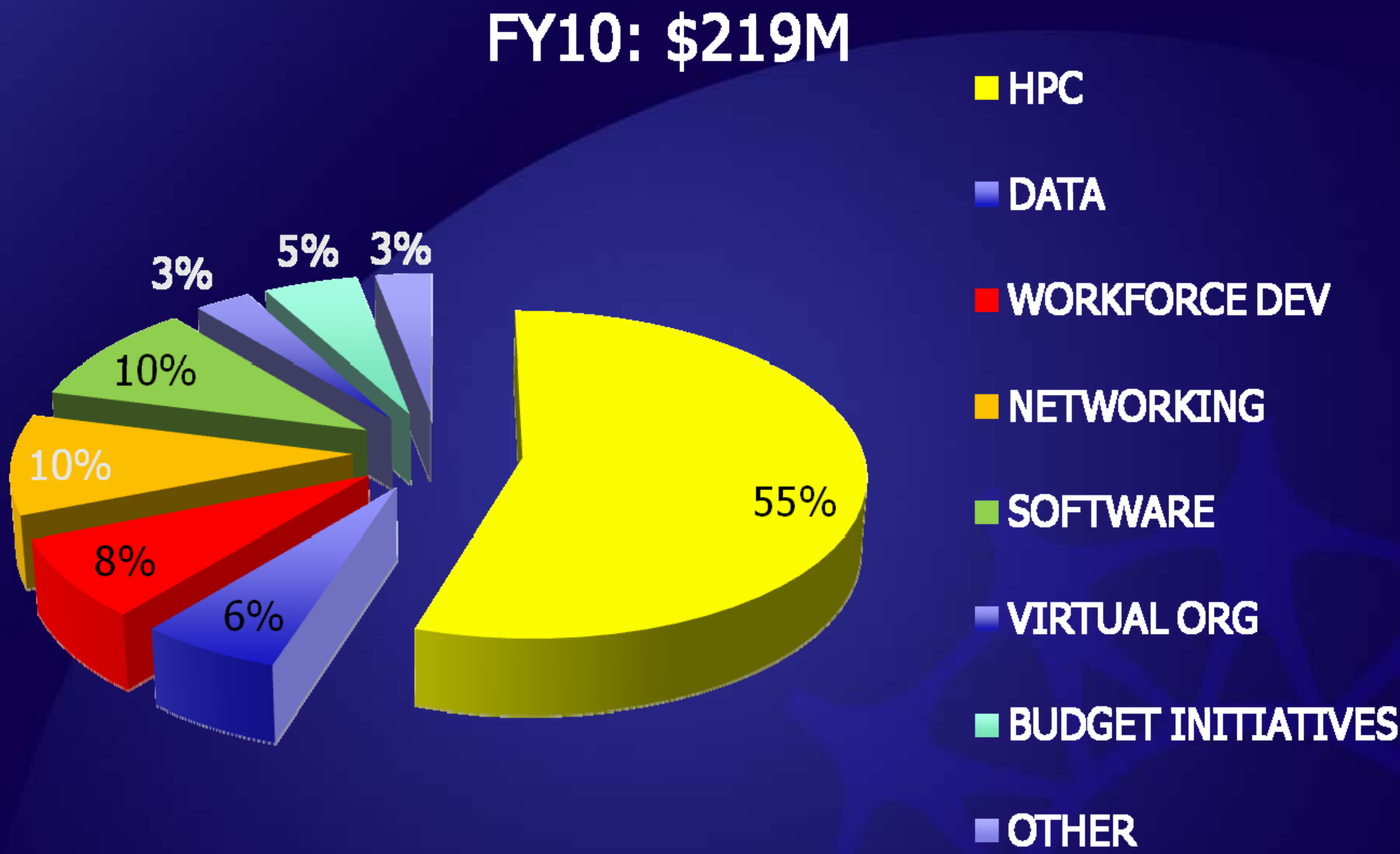


FY 09 Budget
(Before ARRA)



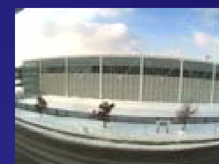
FY 09 Budget
(After ARRA)

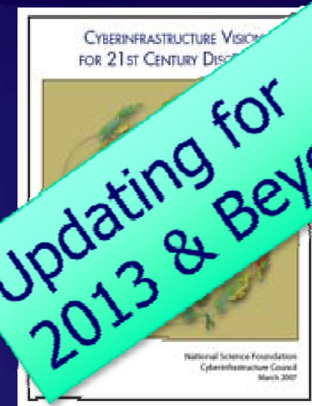
OCI BUDGET BREAKDOWN



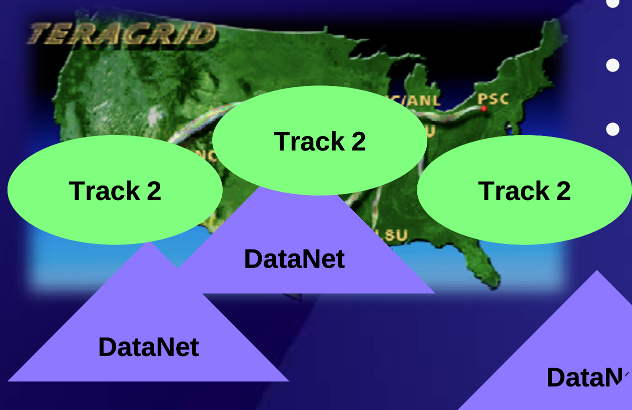
Roadmap and Timelines

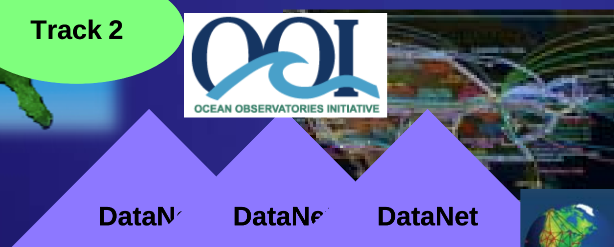
- National Petascale Facility
- CF21Computing program; hubs of innovation
- CF21 Software
- People, Data, VOs
- Better campus integration
- Major facilities CI planning





Updating for 2013 & Beyond





neon
National Ecological Observatory Network, Inc.

- NSF CF21 Strategic Plan 2012-2017
- Integration
- Stronger interagency interaction
- New science activities enabled

Task Force Reports and Workshops →

2010 2011 2012 2013

CF21 Strategy

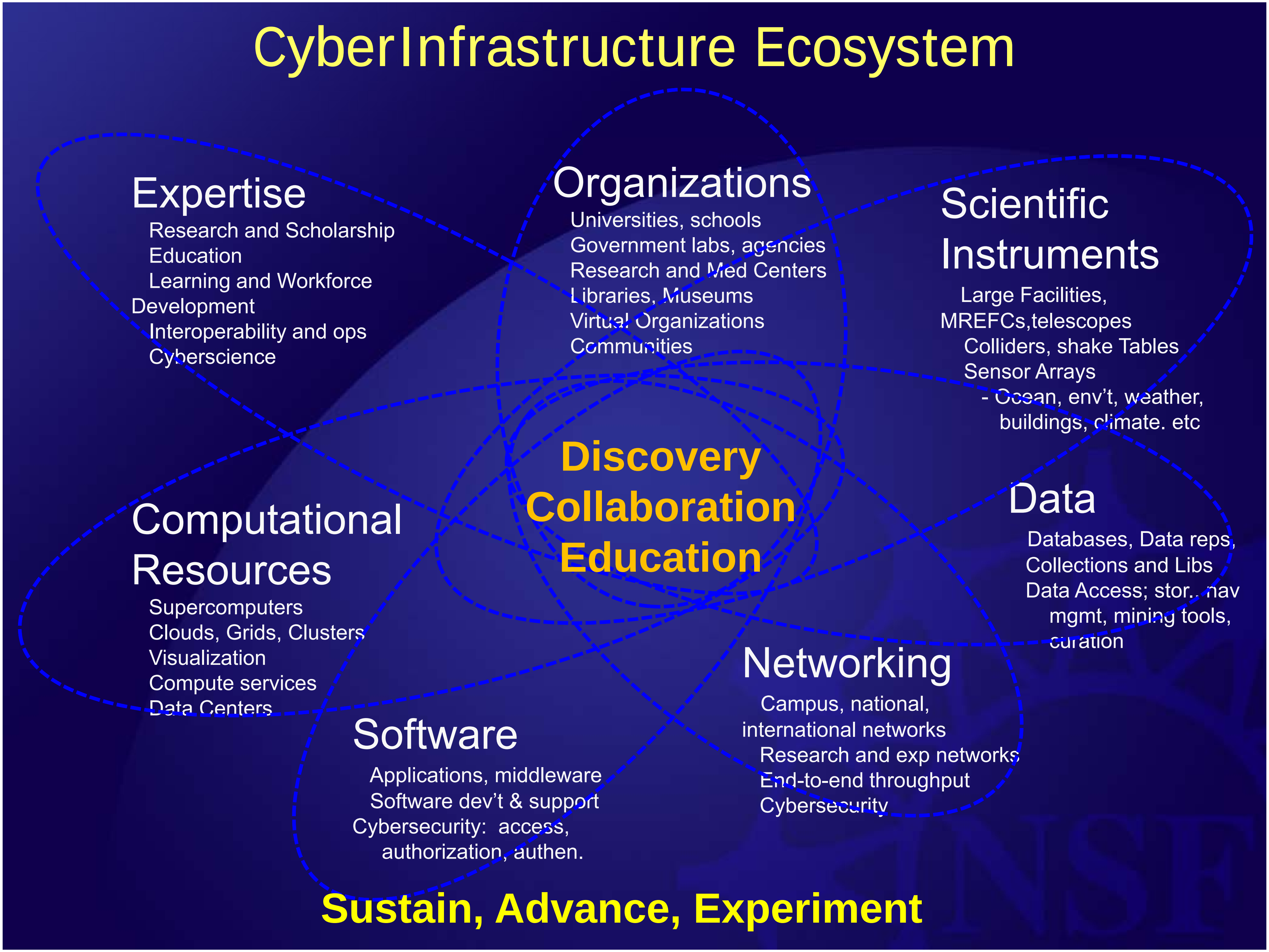
- Driven by science and engineering
- Intense coupling of data, sensors, satellites, computing, visualization, grids, software, VOs; entire CI ecosystem
- Better campus integration
- Major Facilities CI planning
- Task Forces and research community provides guidance and input
- All NSF Directorates involved

- Sustain, Advance, Experiment

Where does HubZero fit?

- Many Places!

- “PosterChild” for ReUse
- Platform for better communication
- Extension of access to resources, services, instruments


CyberInfrastructure Ecosystem
Expertise
Research and Scholarship
Education
Learning and Workforce Development
Interoperability and ops
Cyberscience
Organizations
Universities, schools
Government labs, agencies
Research and Med Centers
Libraries, Museums
Virtual Organizations
Communities
Scientific Instruments
Large Facilities, MREFCs,telescopes
Colliders, shake Tables
Sensor Arrays
- Ocean, env't, weather, buildings, climate. etc
Discovery
Collaboration
Education
Computational Resources
Supercomputers
Clouds, Grids, Clusters
Visualization
Compute services
Data Centers
Data
Databases, Data reps,
Collections and Libs
Data Access; stor., nav mgmt, mining tools, curation
Networking
Campus, national, international networks
Research and exp networks
End-to-end throughput
Cybersecurity
Software
Applications, middleware
Software dev't & support
Cybersecurity: access, authorization, authen.
Sustain, Advance, Experiment

More Information

- Jennifer M. Schopf
  - jschopf@nsf.gov
  - jms@nsf.gov
- Dear Colleague letter for CF21

http://www.nsf.gov/pubs/2010/nsf10015/nsf10015.jsp

- Software infrastructure for sustained innovation

http://www.nsf.gov/pubs/2010/nsf10551/nsf10551.pd

- CITeam

http://nsf.gov/pubs/2010/nsf10532/nsf10532.pdf

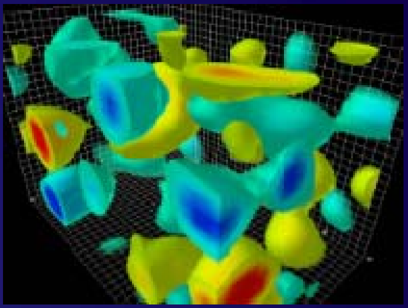
- CITraCS

http://nsf.gov/pubs/2010/nsf10553/nsf10553.pdf

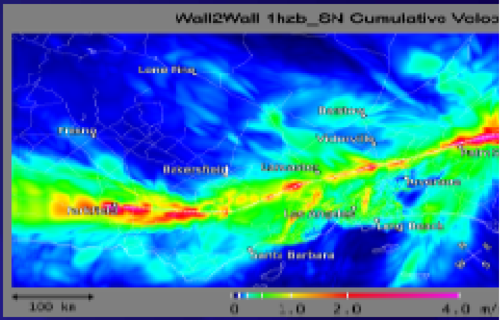
Backup

2009 PetaApps, CDI, CI-Reuse

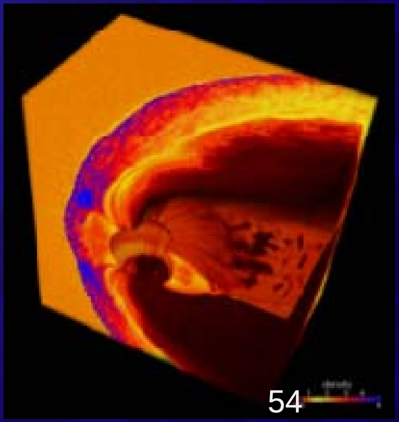
70% OCI ARRA: Innovations in software, apps, people



- PetaApps: OCI led, NSF-wide
  - Partners: MPS, CISE, ENG, GEO and SBE
  - 2009: $16M from OCI, matched for total of $35M!
  - 2007-9 Total: 42 awards, ~200 proposals, $60M
  - Equivalent to entire Track-2 award (including O&M)



- CDI: CISE led, NSF-wide
  - OCI a "Big 4" contributor in FY09! (CISE, ENG, OCI, MPS...), $63M total
  - OCI contributed to 22 awards, more than $10M



- CI Re-Use: Internal OCI-led NSF program
  - OCI venture fund of $4M to catalyze
  - CISE, GEO, OPP, BIO and MPS
  - 13 awards, > $20M investments catalyzed by OCI

The information value ladder

Data >>> Information >>> Insight

>>> Increasing value >>>

Forecasting

Reporting

Analysis

Integration

Done poorly

Distribution

Aggregation

Done poorly to moderately

Quality assurance

Collation

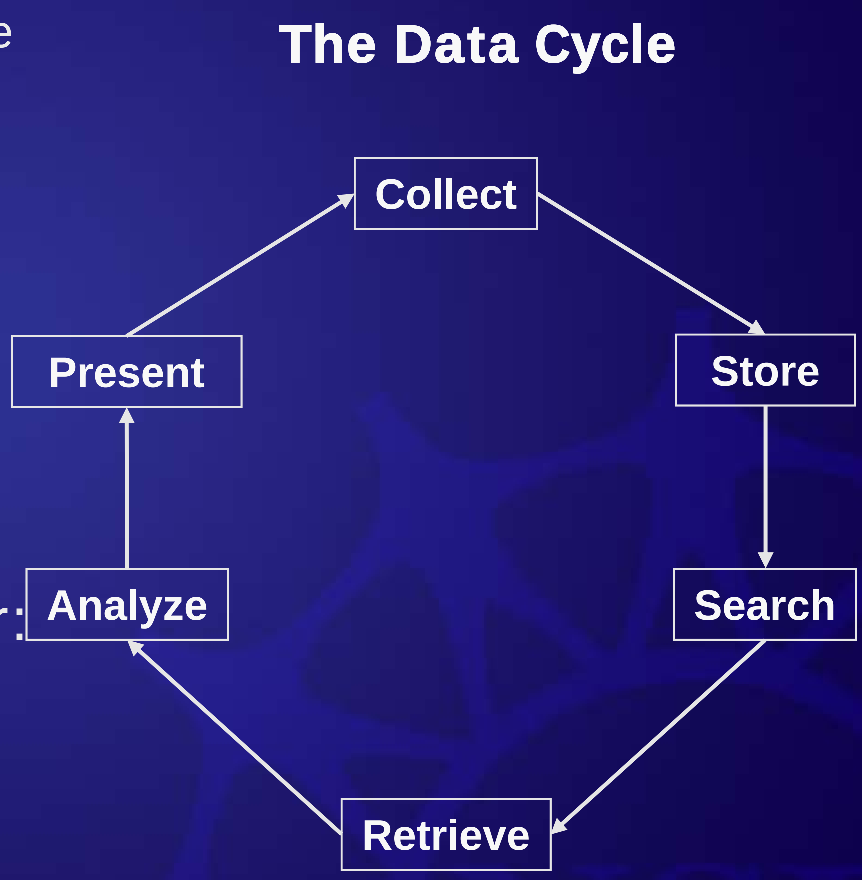
Monitoring

Sometimes done well, by many groups, but could be vastly improved

Slide Courtesy CSIRO, BOM, WMO

The data cycle perspective, from creation to curation

- The science information user:
  - I want reliable, timely, usable science information products
    - Accessibility
    - Accountability
- The funding agencies and the science community:
  - We want data from a network of authors
    - Scalability
- The science information author:
  - I want to help users (and build my citation index)
    - Transparency
    - Ability to easily customize and publish data products using research algorithms



Jeff Dozier, University of California, Santa Barbara

Organizing the data cycle

- Progressive "levels" of data
  - (Earth Observing System)
  - 0 Raw: responses directly from instruments, surveys
  - 1 Processed to minimal level of geophysical, engineering, social information for users
  - 2 Organized geospatially, corrected for artifacts and noise
  - 3 Interpolated across time and space
  - 4 Synthesized from several sources into new data products
- System for validation and peer review
  - To have confidence in information, users want a chain of validation
  - Keep track of *provenance* of information
  - Document theoretical or empirical basis of the algorithm that produces the information
- Availability
  - Each dataset, each version has a persistent, citable DOI (digital object identifier)

Abstract- 45 mins

- Today's science has been radically changed by advances in cyberinfrastructure (CI) – faster machines, better networking, more collaboration, shared data, and the ability to study vastly more complex problems than previously feasible. This talk will present the Office of CyberInfrastructure (OCI) vision for how NSF is addressing both the needs and opportunities raised by these advances in science and CI in terms of innovation, integration, sustainability and people. The CI ecosystem is growing and changing, and NSF is addressing this through extended community interactions through a series of task forces and new programs to be able to sustain, advance, and experiment with cyberinfrastructure, broadly construed.