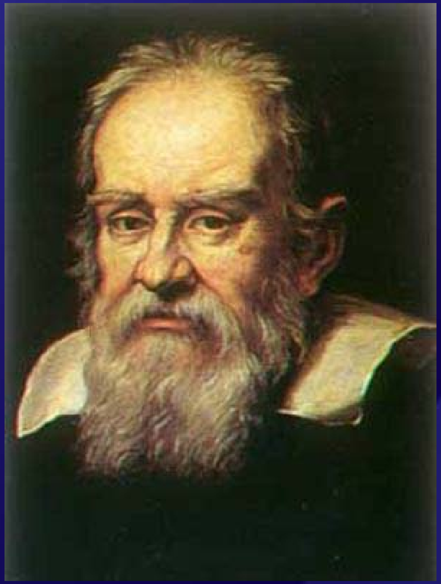
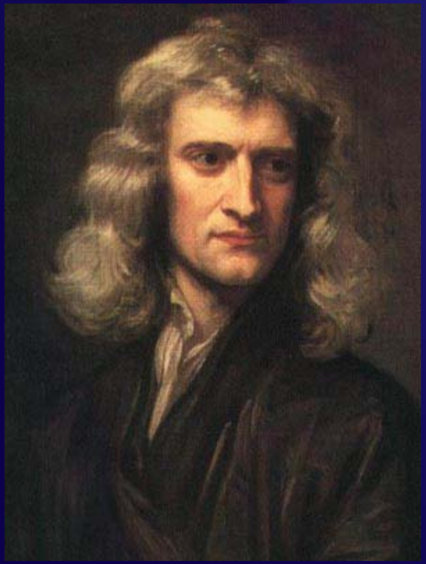
Cyberinfrastructure for Computation and Data-enabled Science & Engineering

Gabrielle Allen <gdallen@nsf.gov>
Program Director, OD/OCI
National Science Foundation

Transformation of Modern Science ...

Profound Transformation of Science
Gravitational Physics





- Galileo, Newton usher in birth of modern science: c. 1600
- Problem: single "particle" (apple) in gravitational field (General 2 body-problem already too hard)
- Methods
  - Data: notebooks (Kbytes)
  - Theory: driven by data
  - Computation: calculus by hand (1 Flop/s)
- Collaboration
  - 1 brilliant scientist, 1-2 student

Profound Transformation of Science

Collision of Two Black Holes

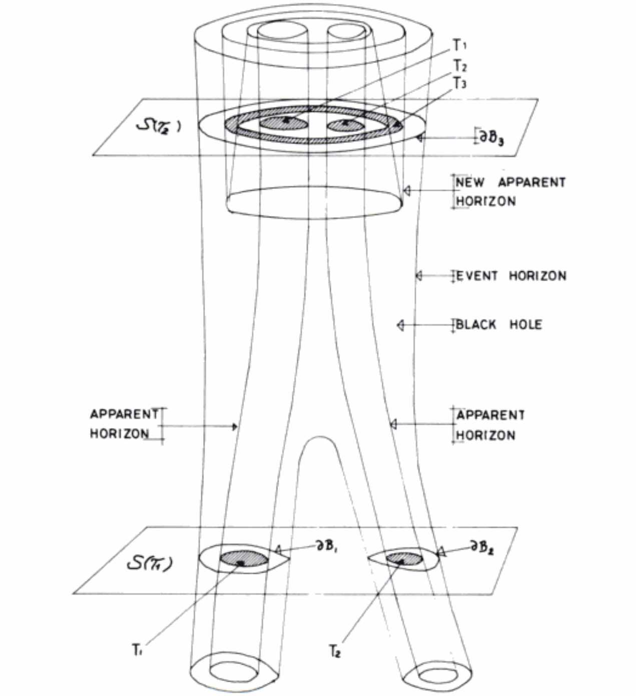
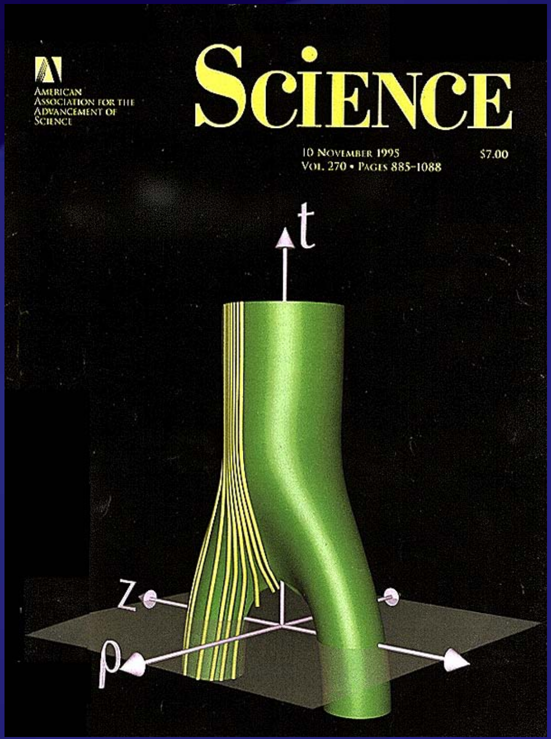
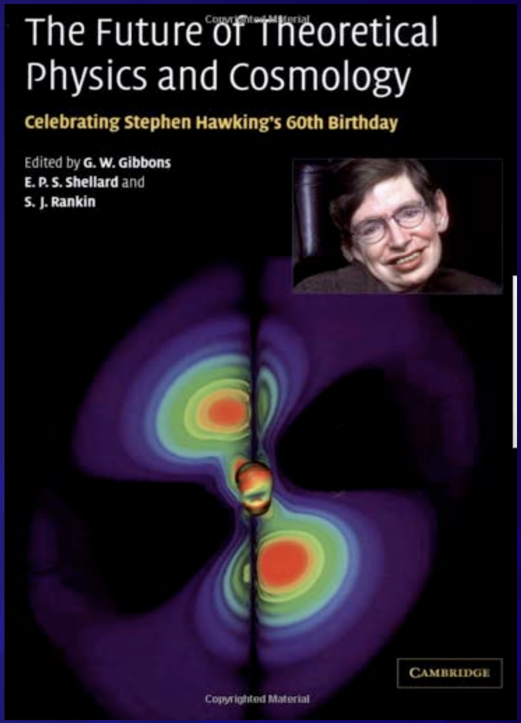


Figure 17. The collision of two Black Holes. The event horizons ∂B_1 and ∂B_2 merge to form the event horizon ∂B_3. The apparent horizons ∂T_2 do not merge but are enveloped by a new apparent horizon ∂T_3.





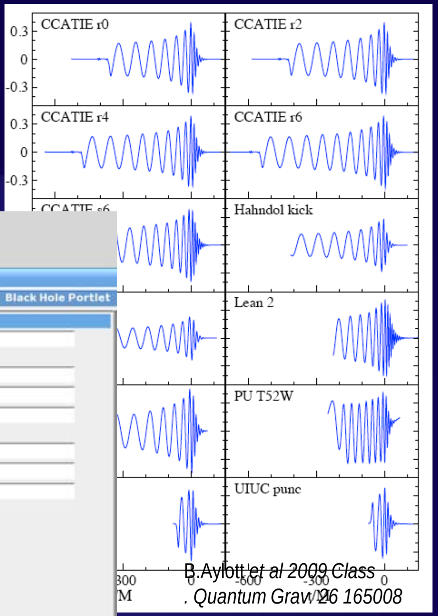
1972: Hawking. 1 person, no computer 50 KB

1995: 10 people, large computer, 50MB

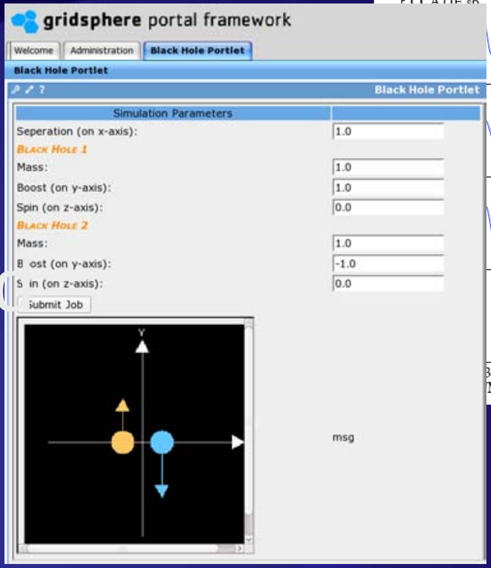
1998: 3D! 15 people, larger computer, 50GB

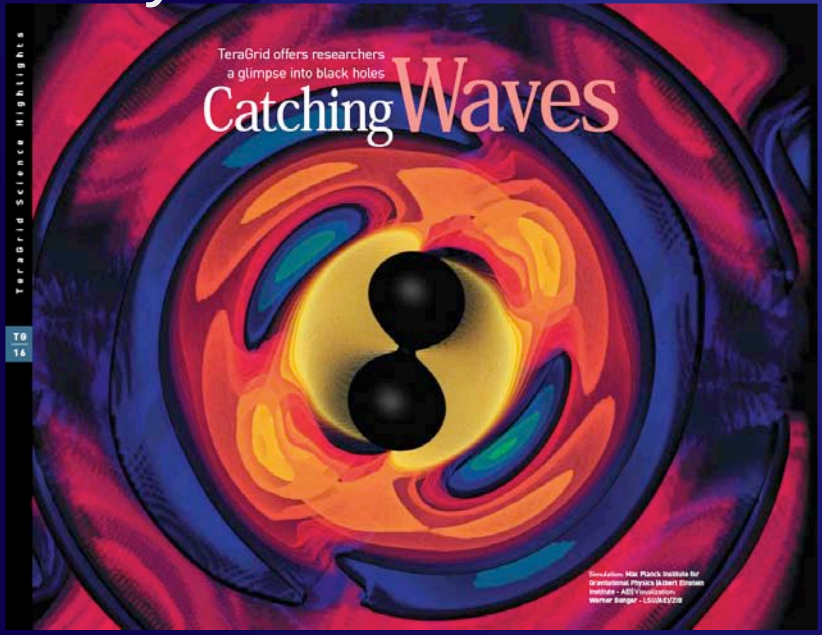
Black Holes: 2011

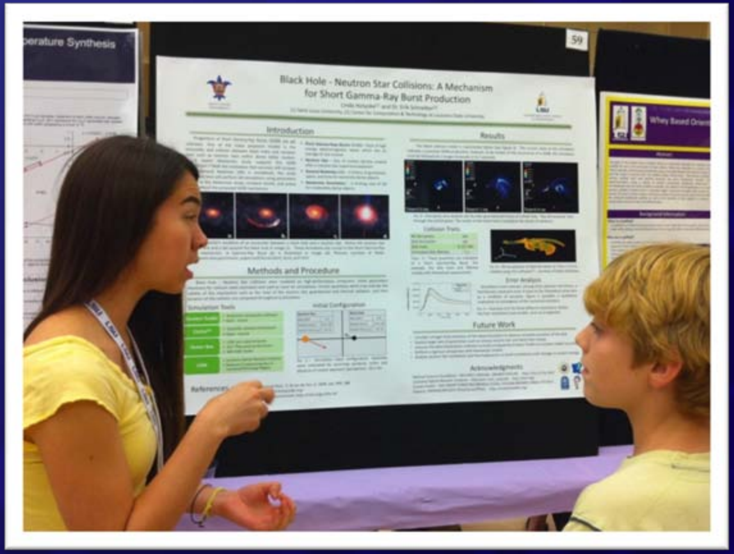
- 40+ year effort to model
- Now: Community codes, accurate waveforms
- Numerical waveforms used with GW data analysis, analytic GR



B.Aylott et al 2009 Class . Quantum Grav. 26 165008

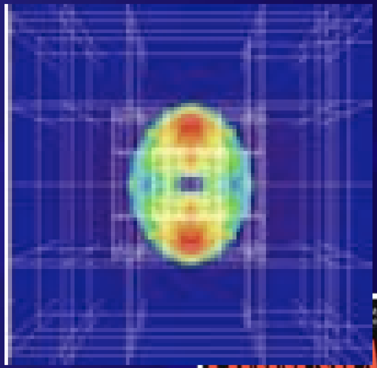






Just ahead: Complexity of Universe
LHC, Gamma-ray bursts!

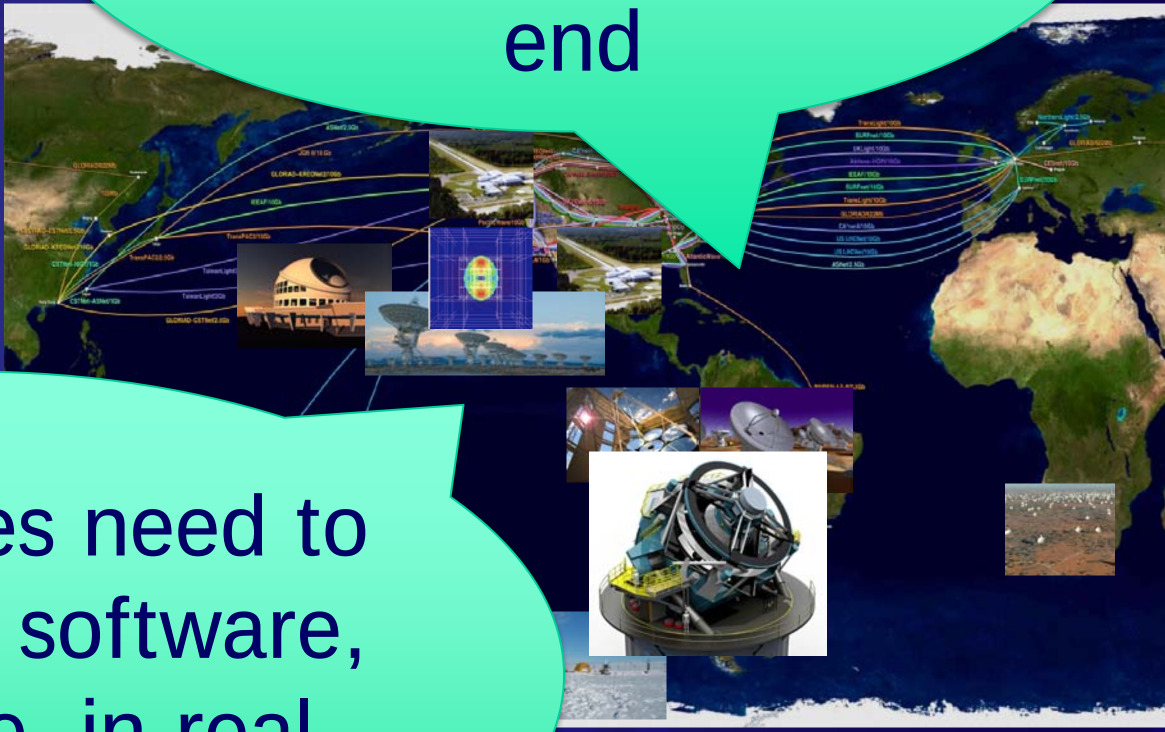
- Gamma-ray bursts!
  - Now: complex problems in relativistic astrophysics
  - Relativity, hydrodynamics, nuclear physics, radiation, neutrinos, magnetic fields: globally distributed collab!
  - Scalable algorithms, complex simulation codes, viz, PFlops*week, PB output!
- Gravity and general relativity are transformed
  - 4 centuries of small science, small data culture
  - 2-3 decades of radical change in both data (factors of 1000 per~5 years) and collaboration



Transient & Data-intensi

- New era: seeing events as they occur
  - (Almost) here now
    - ALMA, EVLA in radio
    - Ice Cube neutrinos
  - On horizon
    - 24-42m optical?
    - LIG
- Si
phys

Will require integration across disciplines, end-to-end

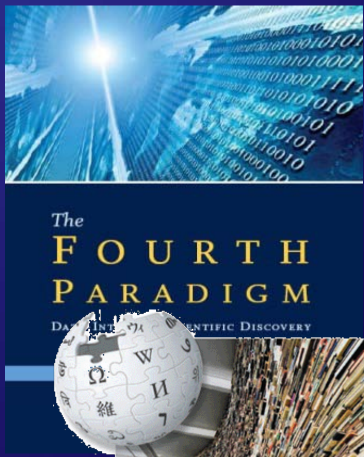


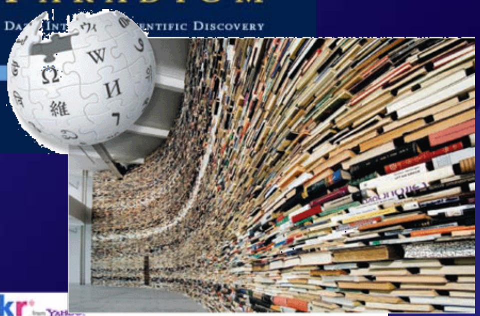
Communities need to share data, software, knowledge, in real time

Enabling Technologies/Paradigms: Rapidly Evolving

Windows Azure

amazon web services™



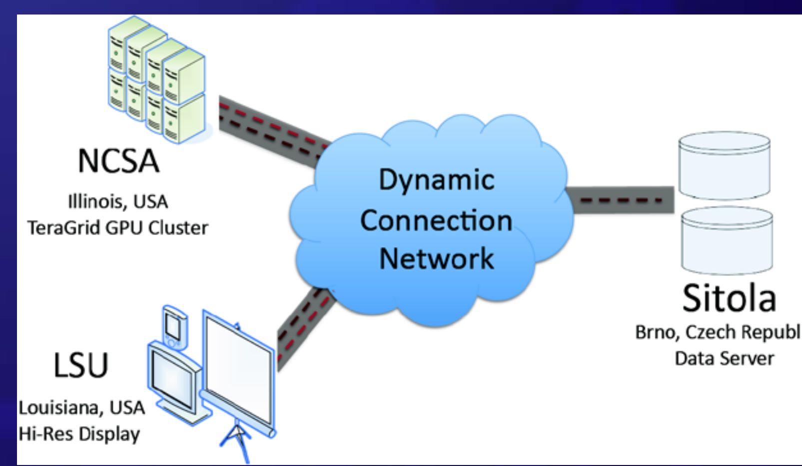


flickr

twitter

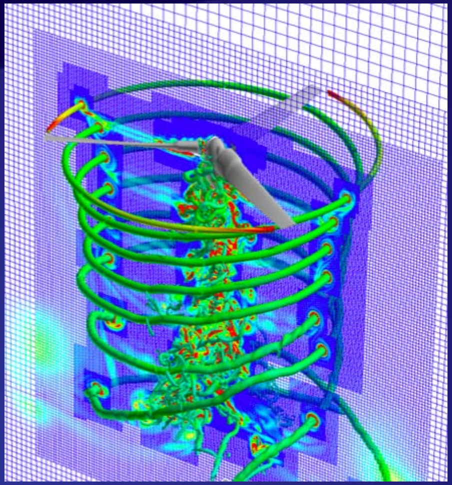
foursquare

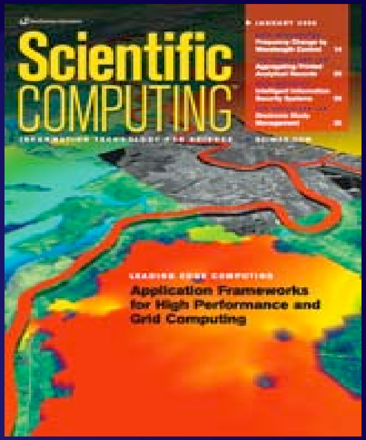
fluidinfo

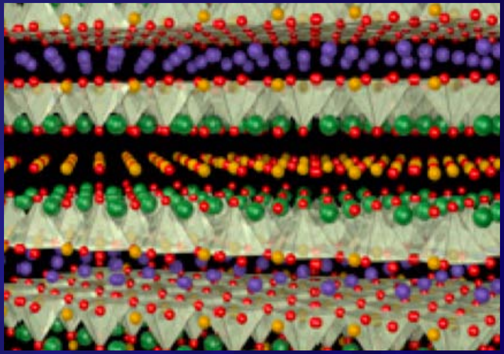


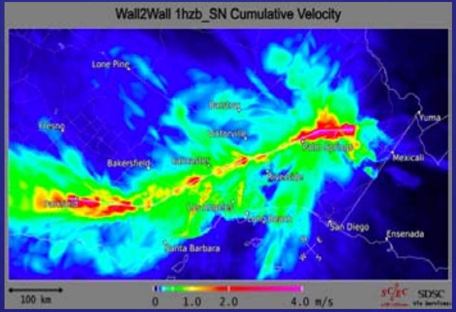
Grand Challenges, Simulations Combined with Digital Observations

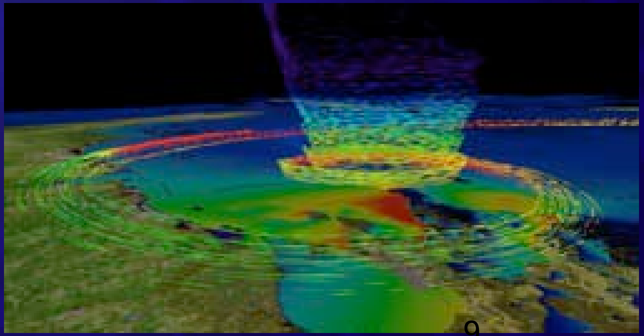
- Advanced New Materials
- Understanding Climate Change
- Quantum Chromodynamics and Condensed Matter Theory
- Semiconductor Design and Manufacturing
- Drug Design
- Energy through Fusion
- New Combustion Systems
- Astronomy and Cosmology
- Cardiovascular Engineering
- Water Sustainability
- Cancer Detection and Therapy
- CO_2 Sequestration











Computation and Data-enabled Science & Engineering (CDSE)

- Traditional computational science and engineering, dramatically enhanced by access to the full spectrum of CI-enabling-technologies: HPC, software, modern computational models and algorithms, data intensive computing, networking and storage, and visualization, as well as issues of education.
- A discipline in its own right at the intersection of applied and computational mathematics, computer science, and core science and engineering disciplines.

Community Recommendations for Change ....

NSF ACCI Task Force Reports

- Final recommendations presented to the NSF Advisory Committee on Cyberinfrastructure (Dec'10/Apr'11)
- Community input via over 25 workshops and BOFs, more than 1300 people involved
- Final reports on-line (April 2011)
  - http://www.nsf.gov/od/oci/taskforces/
- Already impacting ongoing and new programs in OCI

"Permanent programmatic activities in Computational and Data-Enabled Science & Engineering (CDS&E) should be established within NSF."

... Grand Challenges Task Force

"NSF should establish processes to collect community requirements and plan long-term software roadmaps"

... Software Task Force

"NSF should fund interdisciplinary research on the science of broadening participation"

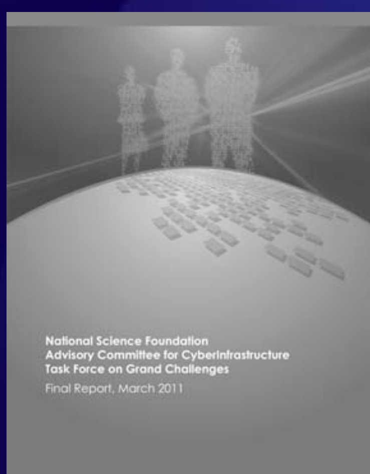
... Cyberlearning Task Force

ACCI Task Force Reports

Campus Bridging

CyberLearning

Data & Viz



Grand Challenges

HPC

Software

Recommendation of NSF Advisory Committee on Cyberinfrastructure (ACCI)

"The National Science Foundation should create a program in Computational and Data-Enabled Science and Engineering (CDS&E), based in and coordinated by the NSF Office of Cyberinfrastructure. The new program should be collaborative with relevant disciplinary programs in other NSF directorates and offices."

NSF can make a strong statement that will lead the Foundation, researchers it funds, and US universities and colleges generally, by recognizing Computational and Data-Enabled Science and Engineering as the distinct discipline it has clearly become.

Approved
Arden L. Bement, Jr.
Director
National Science Foundation

05/27/2010
Date

May 2010

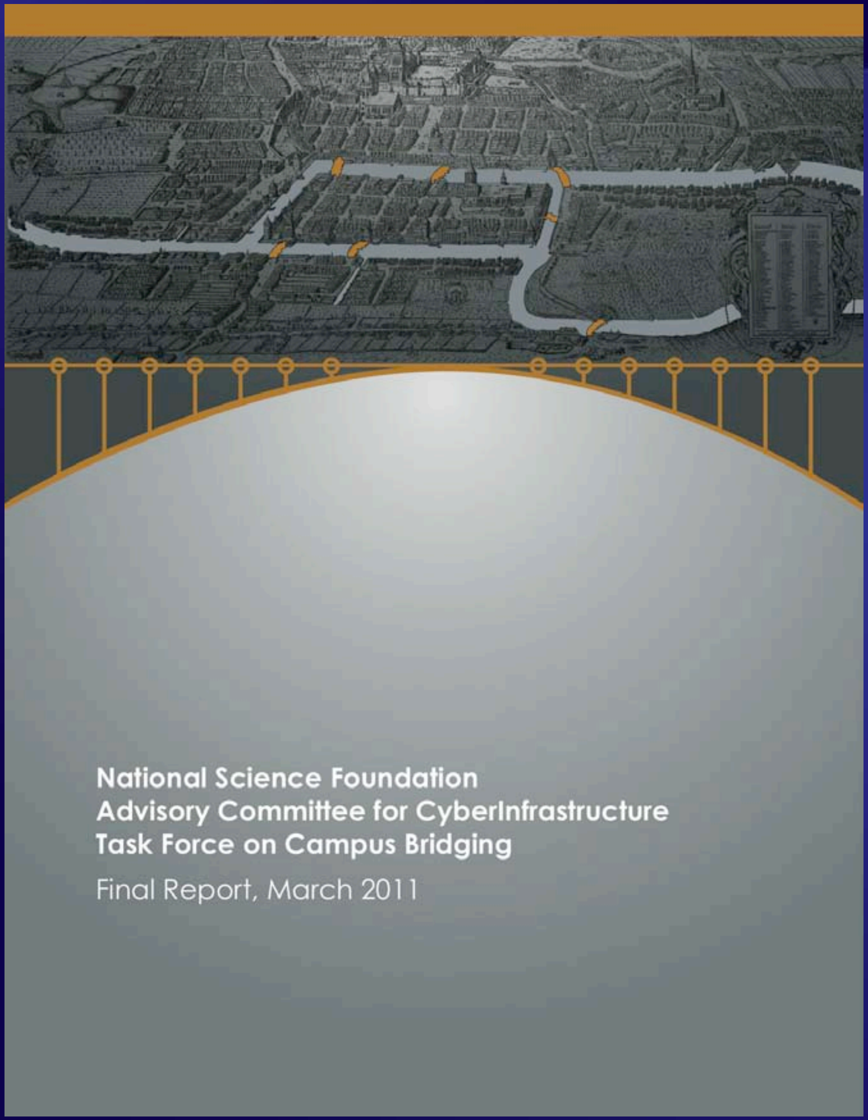
Grand Challenges Task Force

- Chairs:
  - Tinsley Oden
  - Omar Ghattas
  - John Leslie King

Grand Challenges Task Force Recommendations

- *Permanent, integrative* activities in CDS&E are critically needed at NSF to address current and emerging Grand Challenge Problems
- An interagency group in CDS&E should be established to address national goals and priorities and to ensure coordination of efforts
- Support of diverse HPC activities (hardware, methods, algorithms) should remain a high priority. University researchers need open access to these resources at all levels
- The development of robust, reliable and useable software at all levels needs to supported by NSF and recognized as an important component of the research portfolio of NSF
- Support CI for data and visualization
- Learn how to create grand challenge communities and VOs (and do it!)

Campus Bridging Task Force

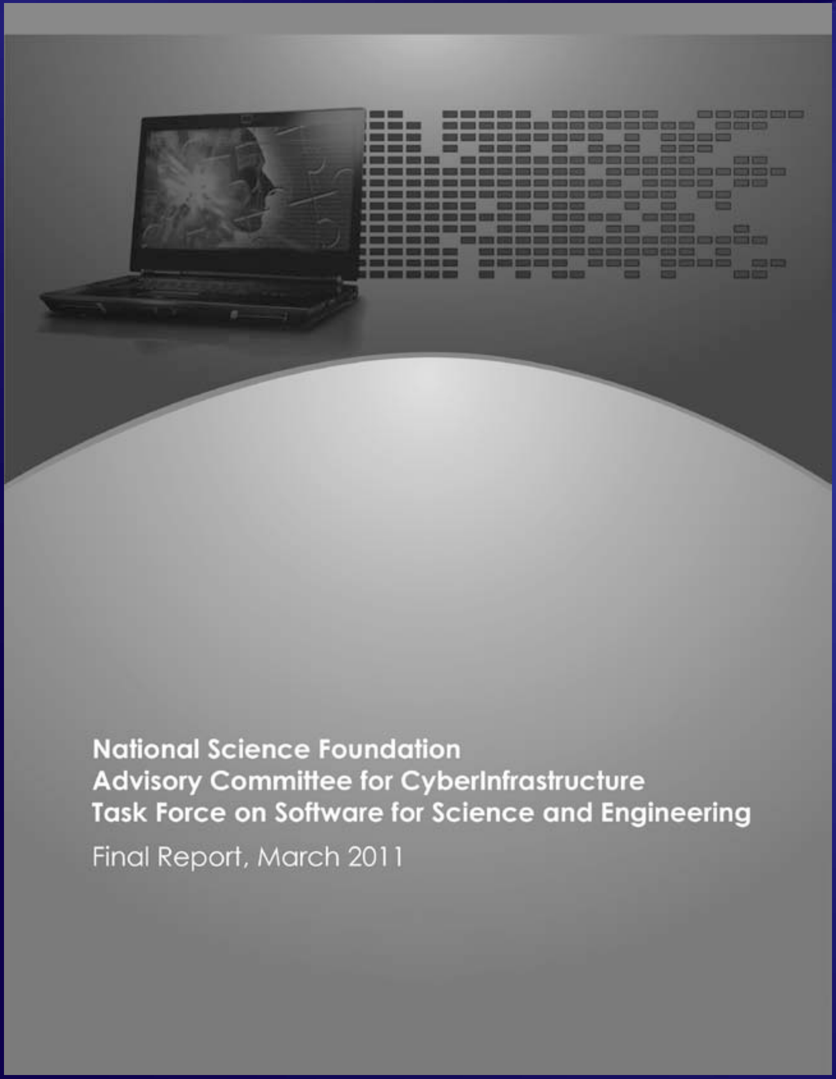


- Chairs:
  - Craig Stewart
  - Guy Almes
  - Jim Bottum

Campus Bridging Recommendations

- NSF should
  - Study successful campus CI implementations to document and disseminate the best practices for strategies, governance, financial models and deployment
  - Establish a blueprint and roadmap for national CI, including
    - Standard Authentication (InCommon)
    - MRI awards at campus level
    - National Data infrastructure, including national networking backbone
- Campuses should
  - Develop a Cyberinfrastructure master plan with the goal of identifying and planning for the changing research infrastructure needs of faculty and researchers
  - Work toward a goal of providing their educators and researchers access to a seamless Cyberinfrastructure which supports and accelerates research and education

Software for Science & Engineering Task Force



- Chairs:
  - David Keyes
  - Valerie Taylor

Software Task Force Recommendations

- Develop a multi-level (individual, team, institute), long-term program to support scientific software
- Promote verification, validation, sustainability, and reproducibility through software developed with federal support
- Develop a consistent policy on open source that promotes scientific discovery and encourages innovation
- Support software through collaborations among all of its divisions, related federal agencies, and private industry
- Utilize its Advisory Committees (including Directorate level) to obtain community input on software priorities

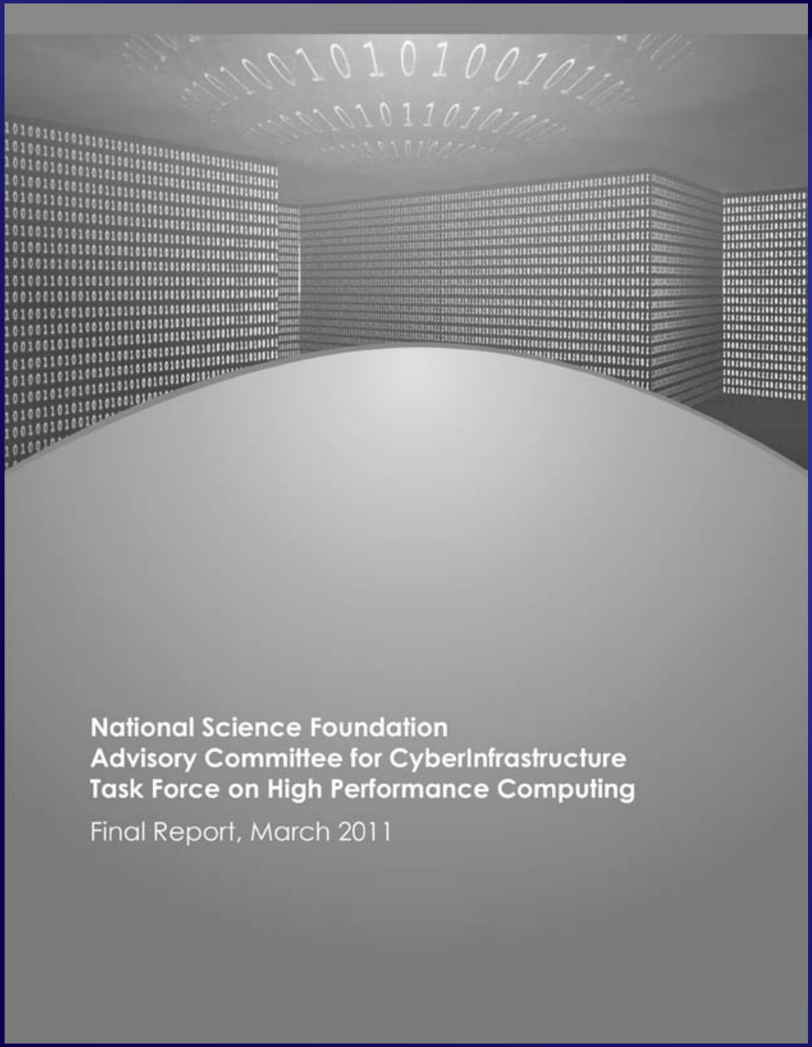
Data Task Force

- Chairs:
  - Tony Hey
  - Shenda Baker

Data Task Force Recommendations

- *Infrastructure*: NSF should recognize data infrastructure and services (including visualization) as essential research assets fundamental to today's science and as long-term investments in national prosperity
- *Culture Change*: NSF should reinforce expectations for data sharing; support the establishment of new citation models in which data and software tool providers and developers are credited with their contributions
- *Economic sustainability*: NSF should develop and publish realistic cost models to underpin institutional/national business plans for research repositories/data services
- *Data Management Guidelines*: NSF should identify and share best-practices for the critical areas of data management
- *Ethics and IP*: NSF should train researchers in privacy-preserving data access

HPC Task Force



- Chairs:
  - Thomas Zacharia
  - Jim Kinter

HPC Task Force Recommendations

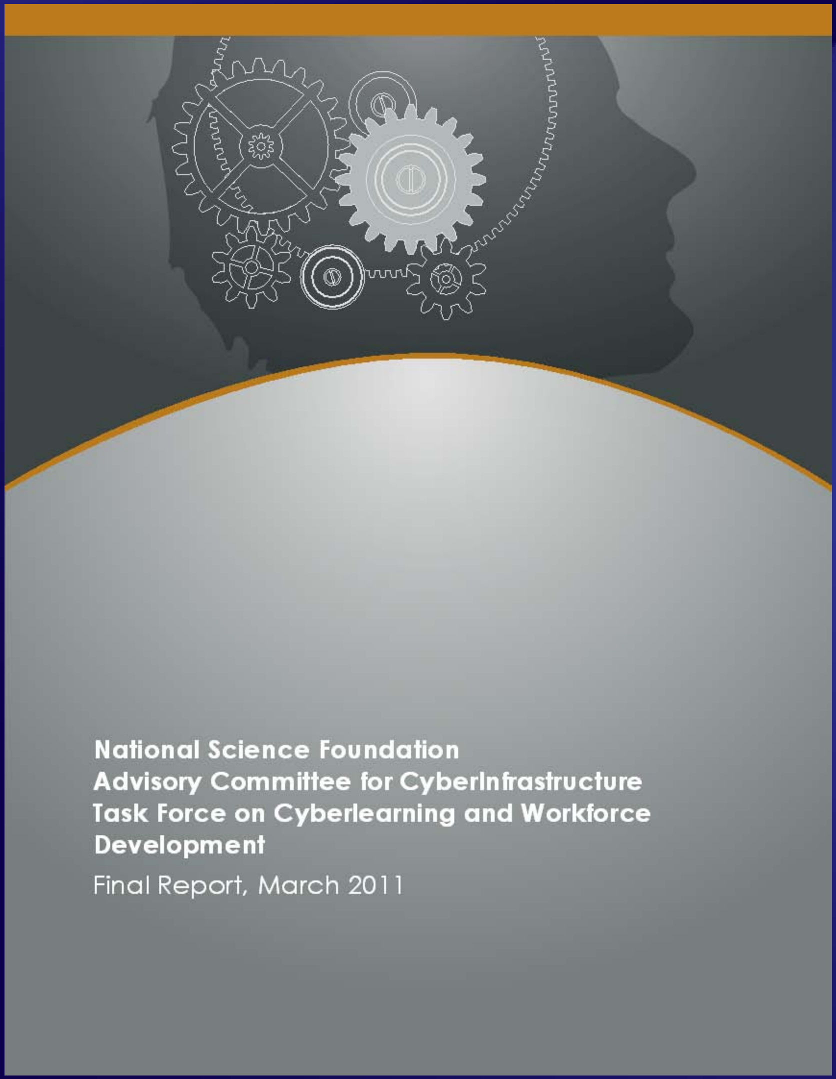
- Develop a *sustainable* model to provide the academic research community with access to a rich mix of HPC systems
  - 20-100 PF, integrated nationally, supported at campus levels
  - Invest now for exascale systems by 2018-2020

- Continue and grow a variety of education, outreach, and training programs to expand awareness and encourage the use of high-end modeling and simulation

- Broaden outreach to improve the preparation of researchers and to engage industry, decision-makers, and new user communities in the use of HPC as a valuable tool

- Provide funding for digital data framework to address the issues of knowledge discovery including co-location of archives and data resources with compute and visualization resources as appropriate

Cyberlearning and Workforce Development Task Force

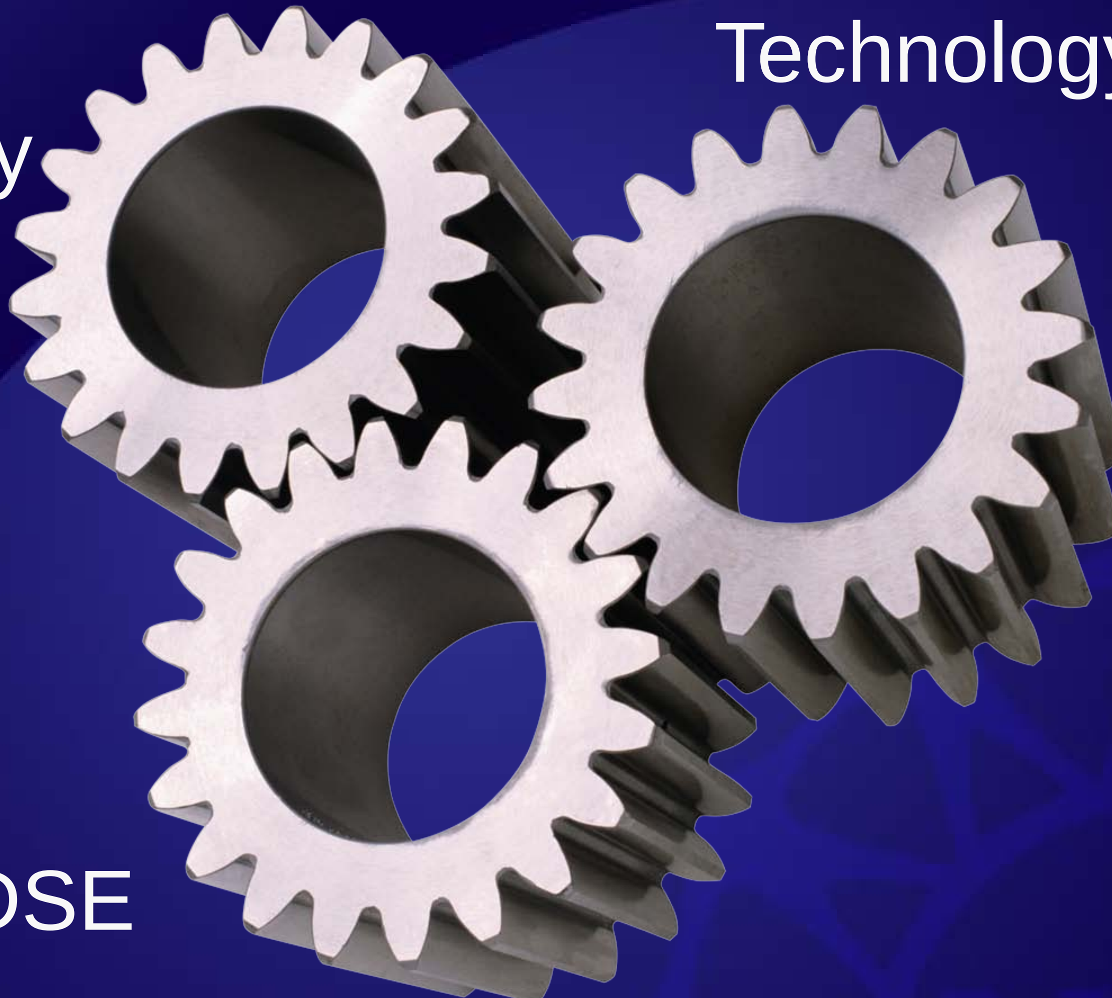


- Chairs:
  - Alex Ramirez
  - Geoffrey Fox

Cyberlearning and Workforce Development Task Force Recommendations

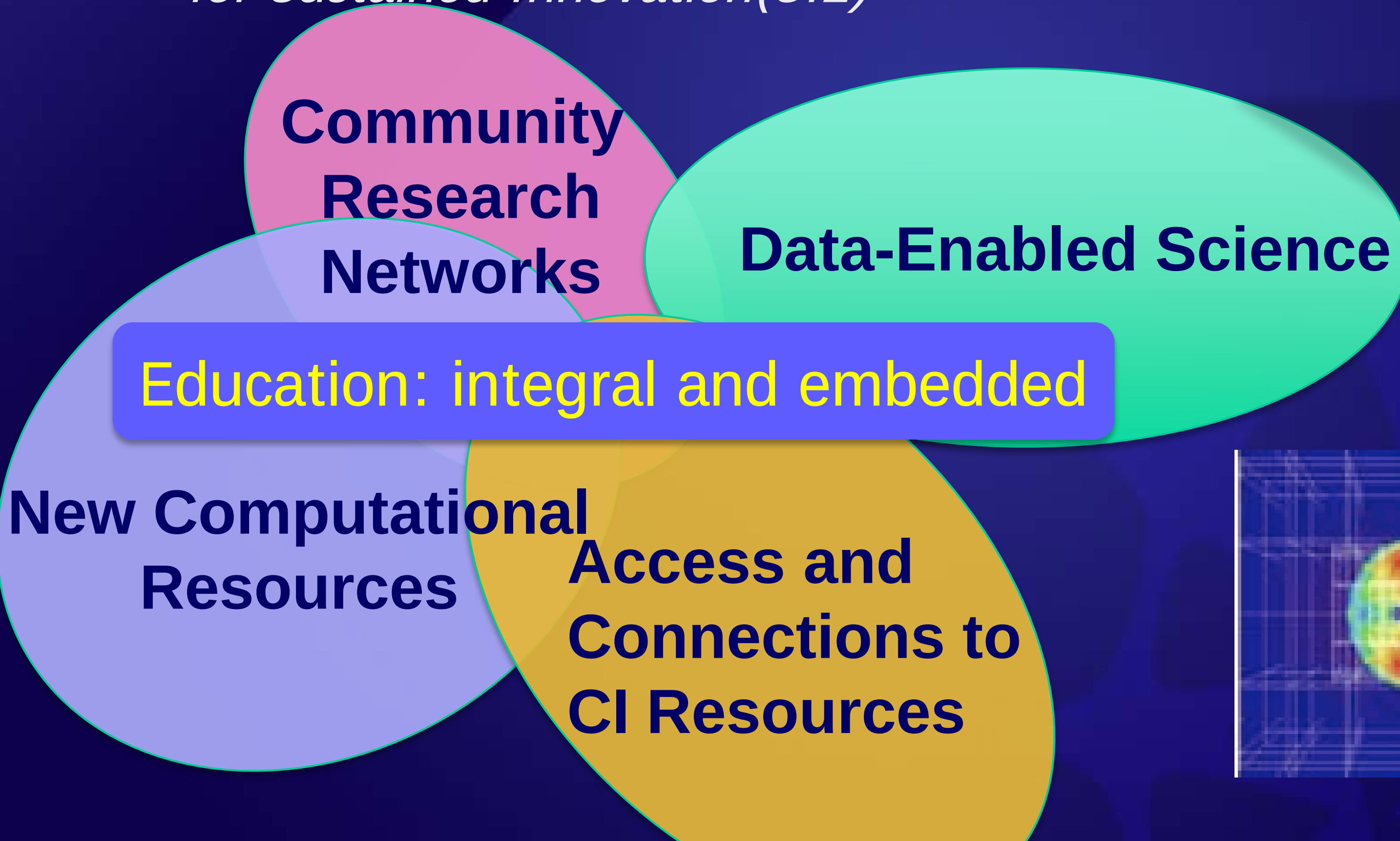
- *Overall*: *Continuous, Collaborative, Computation Cloud (C4)*
  - Pervasive/ubiquitous Internet-based, interacting devices, data sources, users to dominate research, education & all areas of human endeavor
- Promote cross-disciplinary, transformative research and education
  - Systemic change needed at all levels of education; university structures adjusted to train next generation scientists
- Invest in efforts to understand learning and research mechanisms and organizations in the new world of CI
  - Exploit and transform CI-enabled, STEM research advancements, tools, and resources for cyberlearning and workforce development purposes
- Focus on lifelong learning and professional development
- Strengthen leadership, fund research in broadening participation: elimination of underrepresentation of women, persons with disabilities, and minorities

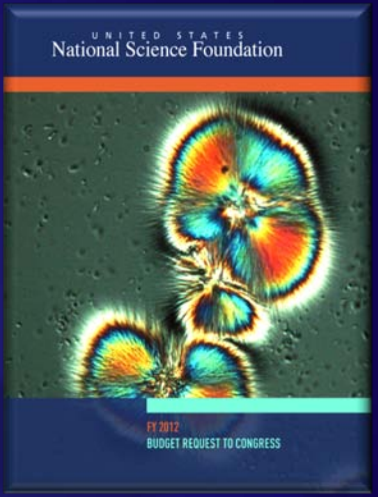
NSF Actions ...

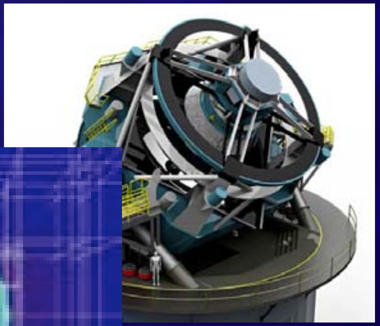

Scientific Discovery
Technology
CDSE

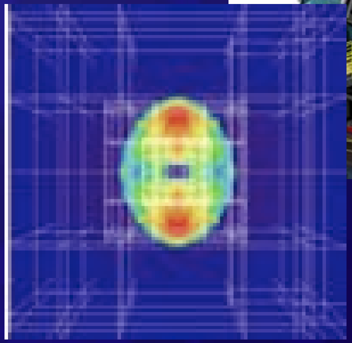
Cyberinfrastructure Framework for 21st Century Science and Engineering (CIF21)

- Coherent program *building on* other CI investments across NSF
  - *eXtreme Digital (XD), Software Infrastructure for Sustained Innovation(SI2)*









Data-Enabled Science
Thrust Area 1

- Data Services Program *(data)*
  - Provide reliable digital preservation, access, integration, and analysis capabilities for science and/or engineering data over a decades-long timeline
- Data Analysis and Tools Program *(information)*
  - Data mining, manipulation, modeling, visualization, decision-making systems
- Data-intensive Science Program *(knowledge)*
  - Intensive disciplinary efforts, multi-disciplinary discovery and innovation

THE CHRONICLE
of Higher Education

Dumped On by Data: Scientists Say a Deluge Is Drowning Research

Changes Coming at NSF for Data!

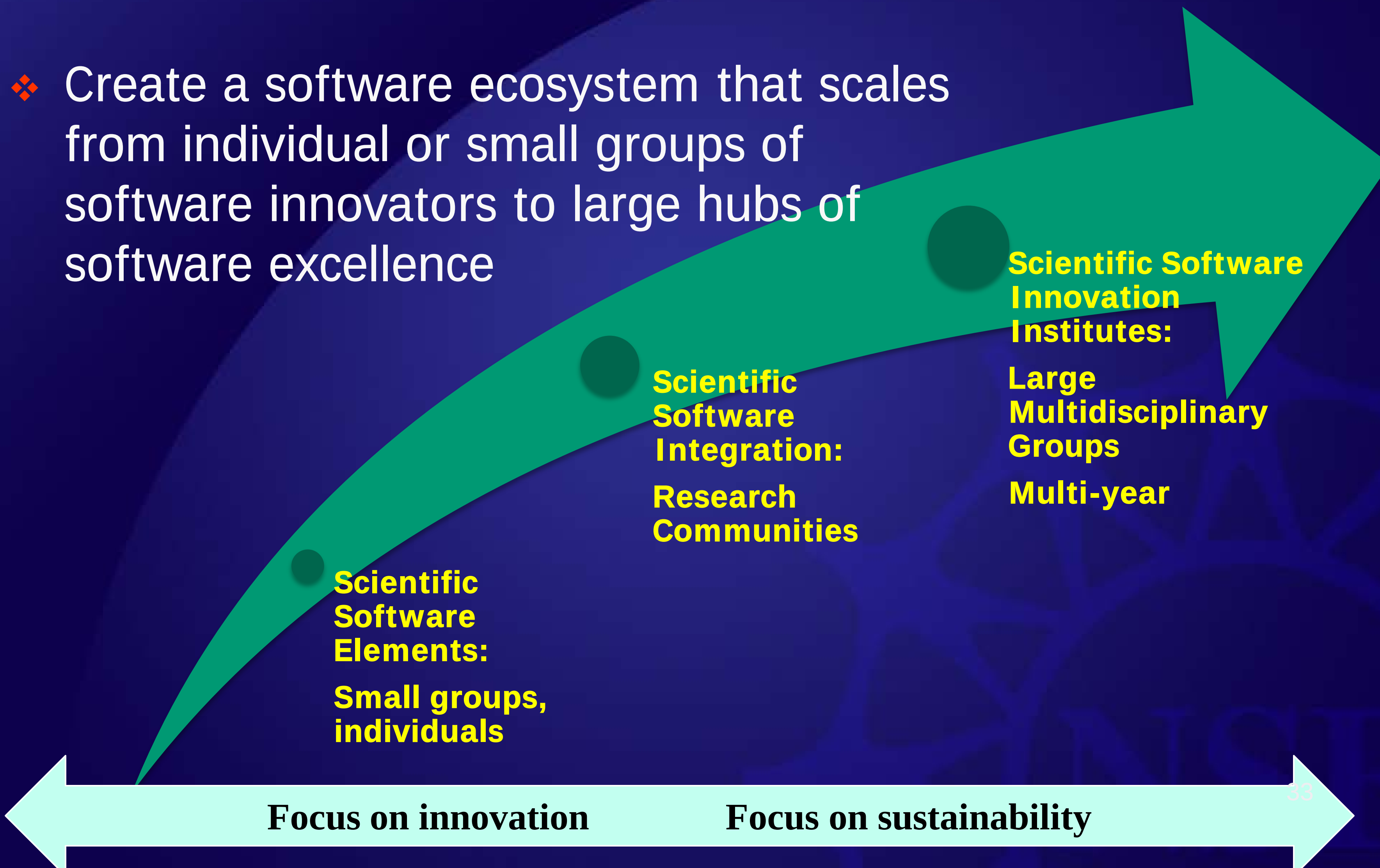
- Long-standing NSF Policy on Data
  - “*Investigators are expected to share with other researchers, at no more than incremental cost and within a reasonable time, the primary data… created or gathered in the course of work under NSF grants*”
- NSF now requires a Data Management Plan (DMP)
  - DMP will be 2-page supplement to the proposal
  - DMP subject to peer review; criterion for award
  - It will not be possible to submit proposals without DMP
  - Customization by discipline, program necessary
- Developing unifying data framework for science
  - Should connect globally; discussions underway with EU
- National Science Board beginning to examine policy for access and openness of data and publication

Sharing data, software will be needed for both interdisciplinary work and reproducibility

New Computational Infrastructure
Thrust Area 2

Creating Scalable Software Development Environments

- Create a software ecosystem that scales from individual or small groups of software innovators to large hubs of software excellence



Community Research Networks Thrust Area 3

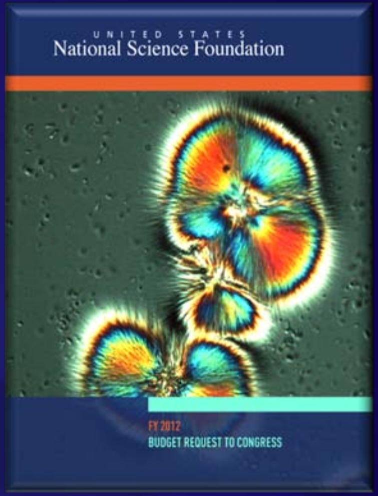
- New multidisciplinary research communities
  - Address challenges beyond individuals and disciplinary research communities
  - Support and optimize collaboration across small, mid-level and large community networks
  - Support SEES and new research communities
- Advanced research on community and social networks
  - Structures, leadership, fostering and sustainability
  - "virtuous cycle" providing feedback through formal evaluation and program iteration

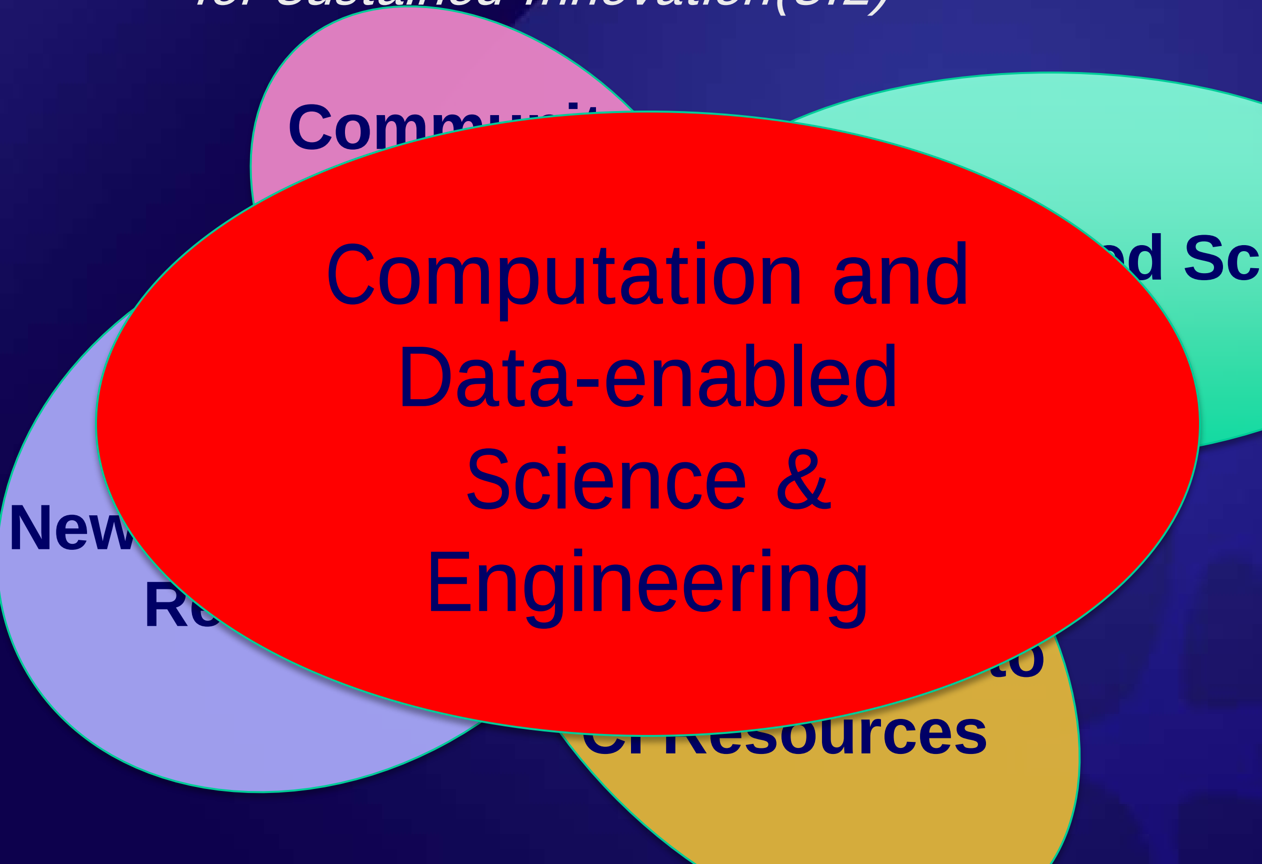
Access and Connectivity Thrust Area 4

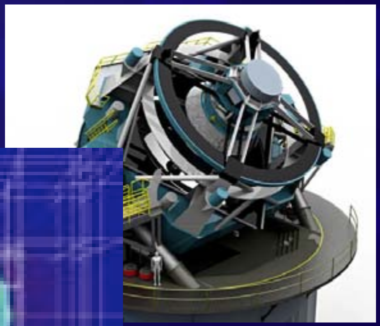
- Network connections and engineering program
  - Real-time access to facilities and instruments; Begins to tie in MREFC activities
  - Integration and end-to-end performance to provide seamless access from researcher to resource
- Cybersecurity – from innovation to practice
  - Deployment of identity management systems
  - Development of cybersecurity prototypes

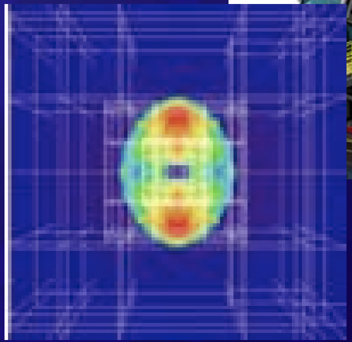
Cyberinfrastructure Framework for 21st Century Science and Engineering (CIF21)

- Coherent program *building on* other CI investments across NSF
  - *eXtreme Digital (XD), Software Infrastructure for Sustained Innovation(SI2)*



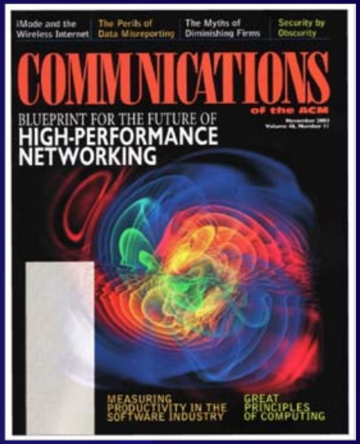






Take Home Lessons

- Science and society profoundly changing
- Comprehensive approach to CI needed to address complex problems of 21st century
  - All elements must be addressed, not just a few;
  - Many exponentials: data, compute, collaborate
- Data-intensive science increasingly dominant
  - Modern data-driven CI presents numerous crises, opportunities
- Academia and Agencies must address
  - NSF Responding through CIF21, changes in implementation of data policy, new programs



http://www.nsf.gov/od/oci/taskforces

(from ~ April 11th)