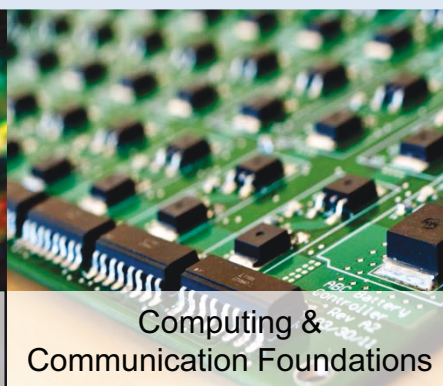# Harnessing the Data Revolution: A Perspective from the National Science Foundation

Computer & Network Systems

Advanced Cyberinfrastructure

Computing & Communication Foundations

Information & Intelligent Systems

Chaitan Baru

Senior Advisor for Data Science

Computer & Information Science & Engineering

National Science Foundation

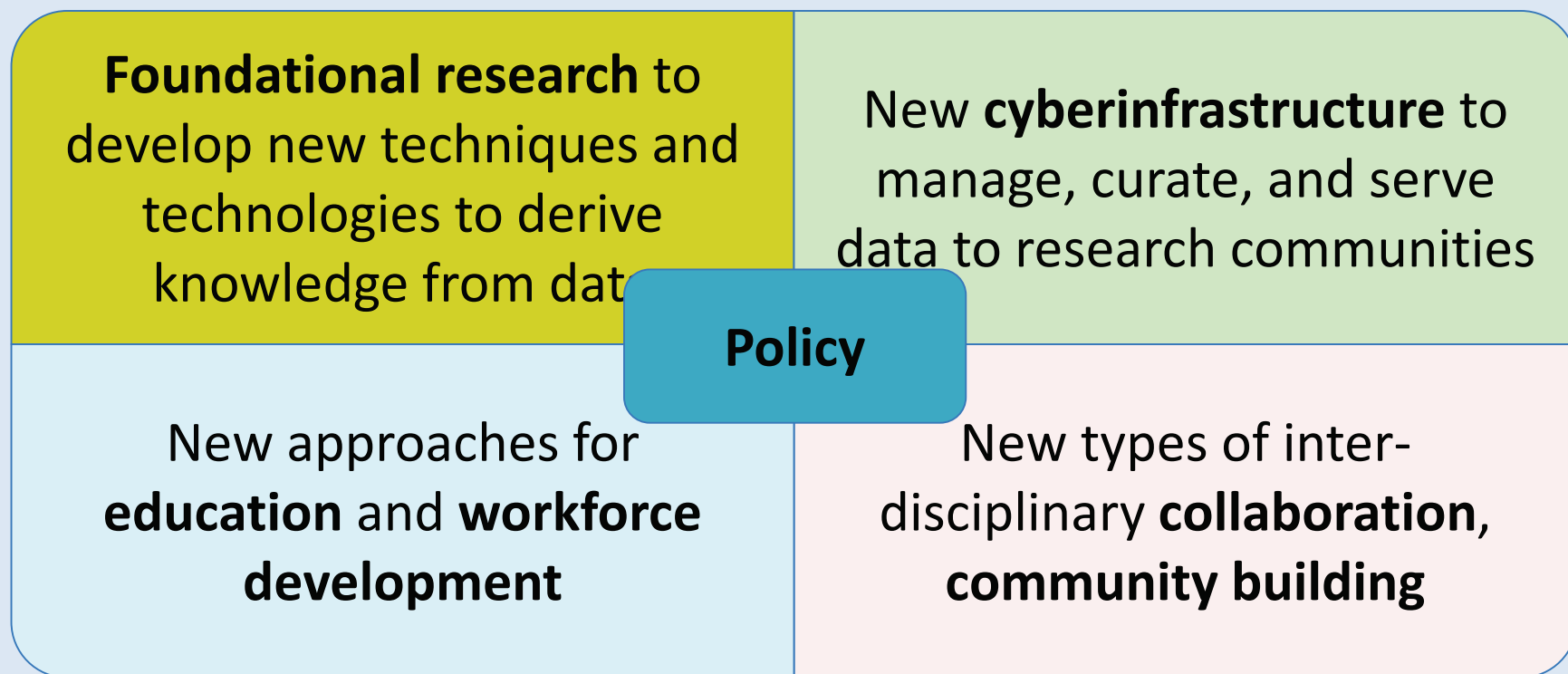**NDIC Meeting**

June 2016, Denver, CO

National Science Foundation's Mission

"To promote the progress of science; to advance the national health, prosperity, and welfare; to secure the national defense..."

NSF's Big Data / Data Science Programs
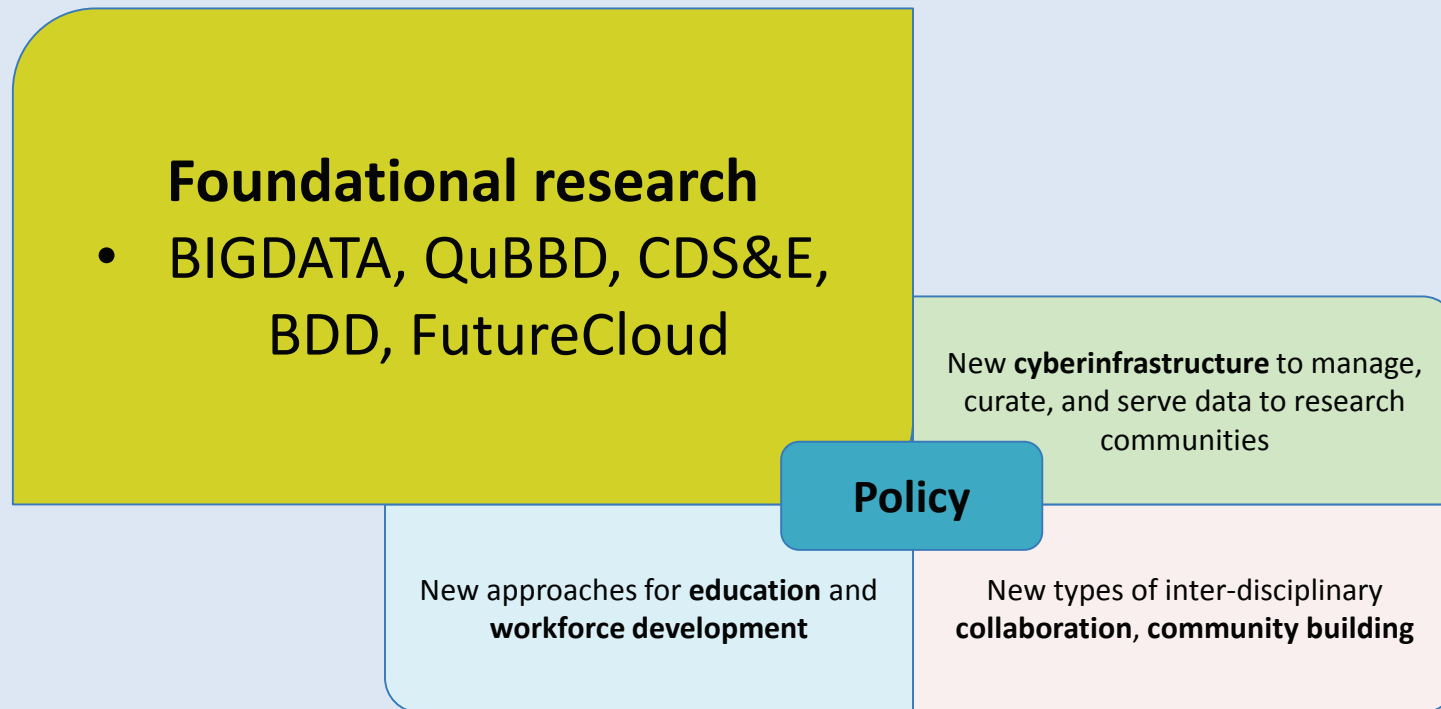
Foundational research to develop new techniques and technologies to derive knowledge from data

New **cyberinfrastructure** to manage, curate, and serve data to research communities

Policy

New approaches for **education** and **workforce development**

New types of inter-disciplinary **collaboration**, **community building**

Big Data / Data Science Research

Foundational research
- BIGDATA, QuBBD, CDS&E, BDD, FutureCloud

New **cyberinfrastructure** to manage, curate, and serve data to research communities

Policy

New approaches for **education** and **workforce development**

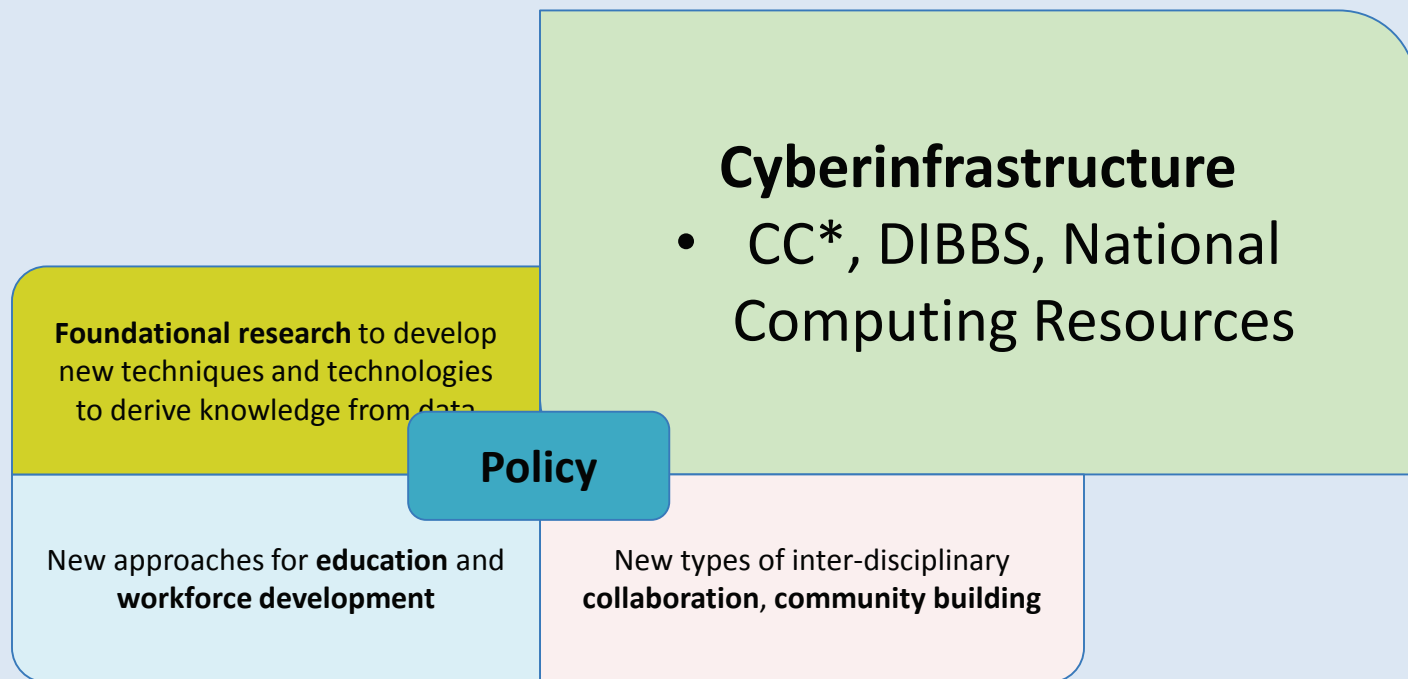New types of inter-disciplinary **collaboration**, **community building**

- BIGDATA (cross-Directorate)
  - Focus on Foundational Approaches and Innovative Applications
- QuBBD (MPS/DMS and NIH)
  - Quantitative Approaches in Biomedical Big Data

- CDS&E (cross-Directorate)
  - Computational and Data Science and Engineering
- BDD (CISE/CCF, IIS)
  - Big Data and Disaster Research
- FutureCloud (CISE/CNS)

Big Data / Data Science Infrastructure

Cyberinfrastructure
- CC*, DIBBS, National Computing Resources

Foundational research to develop new techniques and technologies to derive knowledge from data

Policy

New approaches for **education** and **workforce development**

New types of inter-disciplinary **collaboration**, **community building**

- CC*DNI (CISE/ACI)
  - Campus Cyberinfrastructure - Data, Networking, and Innovation Program
- DIBBS (CISE/ACI, with other Directorates)
  - Data Infrastructure Building Blocks
  - EarthCube (with GEO)

- RIDIR (SBE)
  - Resource Implementations for Data Intensive Research in the Social Behavioral and Economic Sciences
- BCC (EHR)
  - Building Community and Capacity in Data Intensive Research in Education

Big Data / Data Science Infrastructure

Foundational research to develop new techniques and technologies to derive knowledge from data

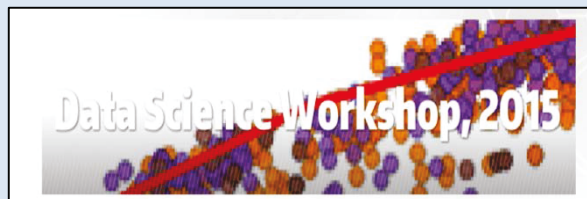New **cyberinfrastructure** to manage, curate, and serve data to research communities

Policy

Education and **workforce development**
- NSF Research Traineeship (NRT)

New types of inter-disciplinary **collaboration**, **community building**

- Data Science Education workshop, Aug 5-7, 2015, UW (CISE, EHR)
  - Brought together ~100 graduate students across different disciplines, engaged in data science research

Data Science Workshop, 2015

Big Data / Data Science Infrastructure

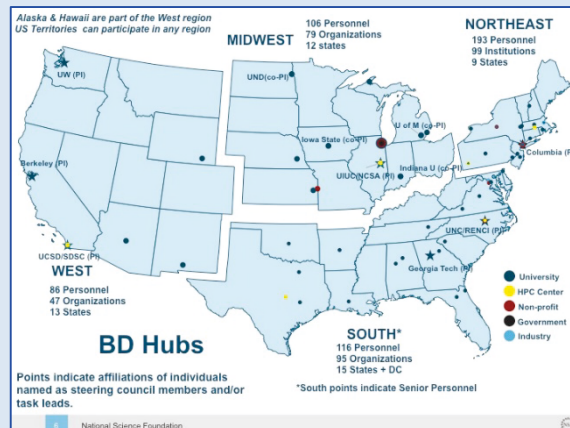**Foundational research** to develop new techniques and technologies to derive knowledge from data

New **cyberinfrastructure** to manage, curate, and serve data to research communities

Policy

New approaches for **education** and **workforce development**

Community building
- Big Data Regional Innovation Hubs and Spokes (BD Hubs/Spokes)

Alaska & Hawaii are part of the West region
US Territories can participate in any region

MIDWEST
106 Personnel
79 Organizations
12 states

NORTHEAST
193 Personnel
99 Institutions
9 States

UW (PI)
UND(co-PI)
U of M (co-PI)
Berkeley (PI)
Iowa State (co-PI)
Indiana U (co-PI)
Columbia (PI)
UIUC/NCSA (PI)
UCSD/SDSC (PI)
UNC RENCI (PI)
Georgia Tech (PI)

WEST
86 Personnel
47 Organizations
13 States

- University
- HPC Center
- Non-profit
- Government
- Industry

BD Hubs

SOUTH*
116 Personnel
95 Organizations
15 States + DC

Points indicate affiliations of individuals named as steering council members and/or task leads.

*South points indicate Senior Personnel

National Science Foundation

Also, Research Data Alliance, RDA (rd-alliance.org)

# 2017: Looking Ahead – New Initiatives

- Smart and Connected Communities (S&CC)
  - Improving quality of life, health, well-being, and learning in communities
- Smart and Autonomous Systems (S&AS)
  - Making possible intelligent physical systems that sense, perceive, and operate in dynamic, uncertain, and unanticipated environments
- Secure and Trustworthy Cyberspace (SaTC)
  - Securing our Nation's cyberspace
- Understanding the Brain (UtB)
  - Improving understanding of the brain
- National Strategic Computing Initiative (NSCI)
  - Maximizing benefits of HPC for scientific discovery and economic competitiveness
- Data for Scientific Discovery & Action (D4SDA)
  - Enabling 21st century science, engineering, and education to move toward effective use of digital data to advance discovery

NSCI and Data Science:
Two related national imperatives

- **High Performance Computing and Data Science**
  in support of science and engineering discovery and competitiveness
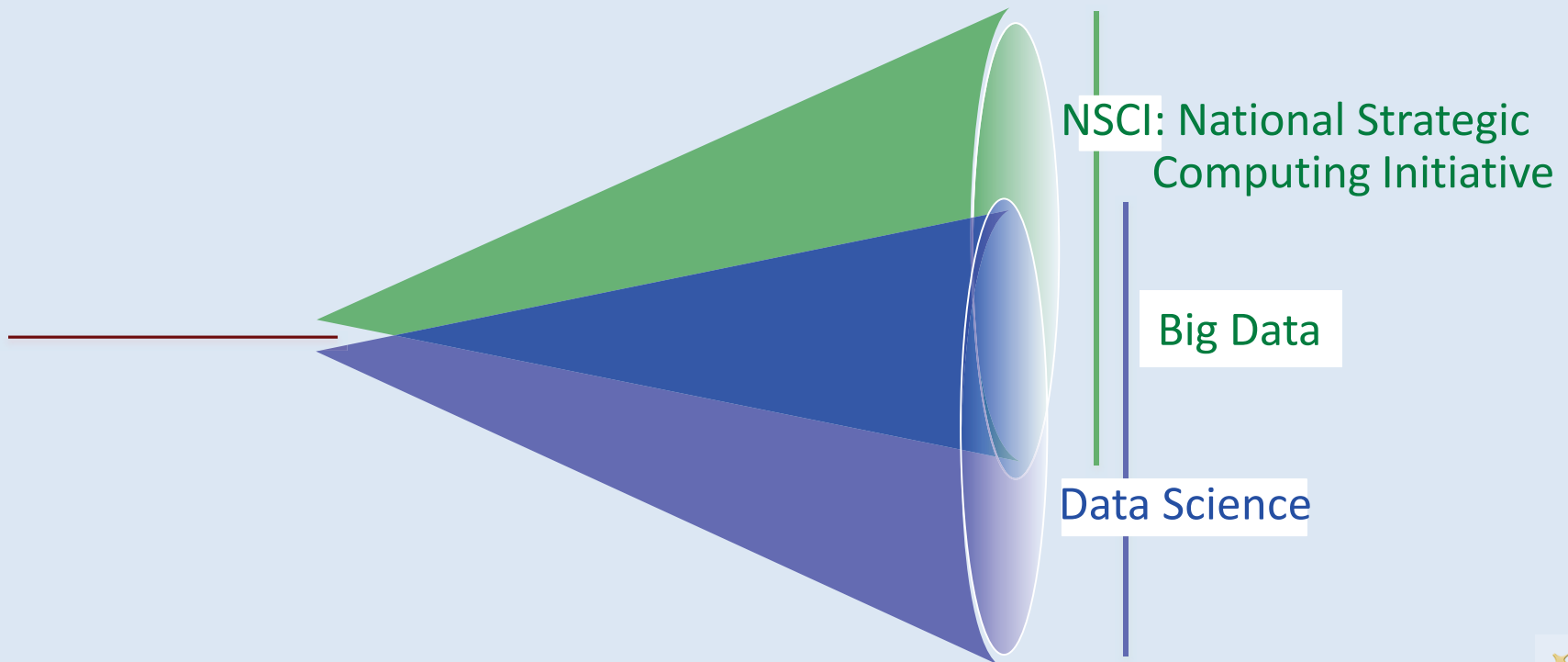
NSCI: National Strategic Computing Initiative

Big Data

Data Science

National Strategic Computing Initiative (NSCI)

Maximizing benefits of HPC for scientific discovery and economic competitiveness

CISE Investment: $19.70 M (2017)

- Partnership among CISE (co-lead), MPS (co-lead), ENG, and GEO

- Aligns with Executive Order 13702, *Creating a National Strategic Computing Initiative*

- NSF – with CISE leading – plays a central role in:

  - Scientific discovery advances
  - Broader HPC ecosystem for scientific discovery
  - Workforce development
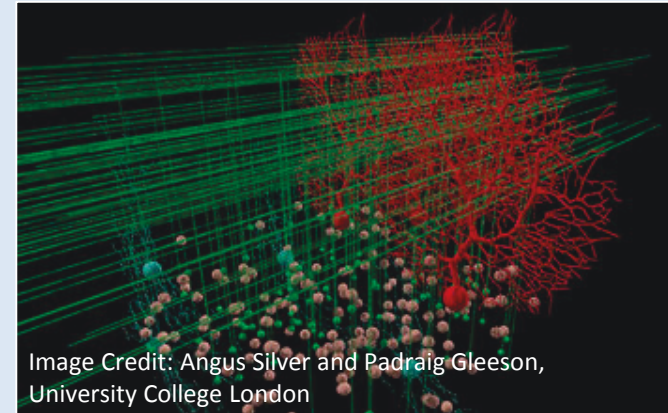
Image Credit: Angus Silver and Padraig Gleeson, University College London

Image Credit: NCSA/University of Illinois

Data for Scientific Discovery & Action (D4SDA)

Enabling 21st century science, engineering, and education to move toward effective use of digital data to advance discovery

CISE Investment: $19.60 M (2017)

- Aims to:

  - Promote foundational research in critical techniques, technologies

  - Provide innovative, reusable data and knowledge infrastructure to support data-intensive science

  - Enable/incent science community to address data governance, lifecycle issues

  - Educate data-savvy workforce of scientists, engineers, educators

NSF "Big Ideas"

Better understanding the changing Arctic is one item on a new list of big ideas that should shape the National Science Foundation's work.

NASA/Kathryn Hansen

NSF director unveils big ideas, with an eye on the next president and Congress

By **Jeffrey Mervis** | May. 10, 2016 , 3:30 PM

NSF "Big Ideas"

- Catalyze interest, investment in fundamental research

- Generate broad public appeal and attract partnerships with industry, private foundations, and academia

- Cutting-edge research agenda, appropriate for NSF

- Process: directorate discussions, white papers, AD retreat, subsequent collaborative refinement

NSF "Big Ideas"

RESEARCH
IDEAS

CISE will play a key role in many of these Big Ideas

Harnessing the Data Revolution

AD Retreat: CISE white paper
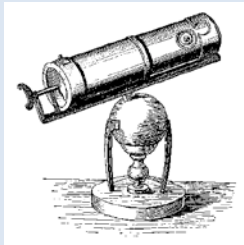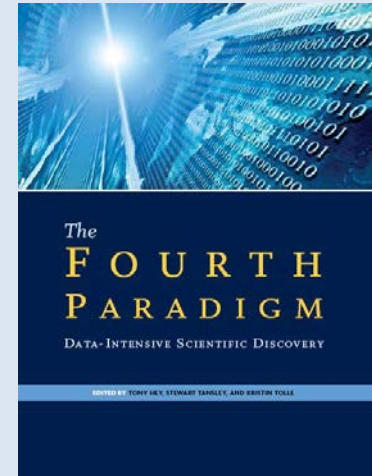
Experimental   Theoretical   Computational   Data

Vision: *move beyond isolated, standalone approaches for data science, services and infrastructure towards a cohesive, federated, national-scale approach that will* **harness** *the data revolution and* **transform** *U.S. science, engineering, and education over the next decade and beyond*

Data: critical across *all* of science
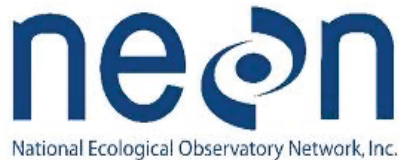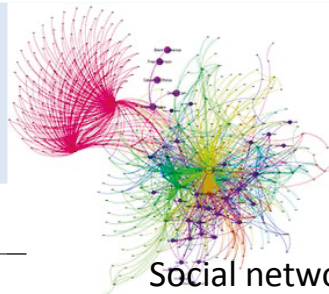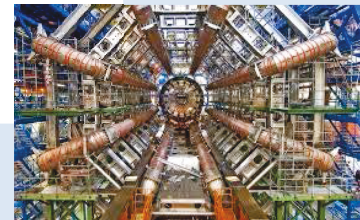
Harnessing the Data Revolution*

- *fundamental research in mathematics, statistics, computer and computational science*: data-driven discovery, decision-making

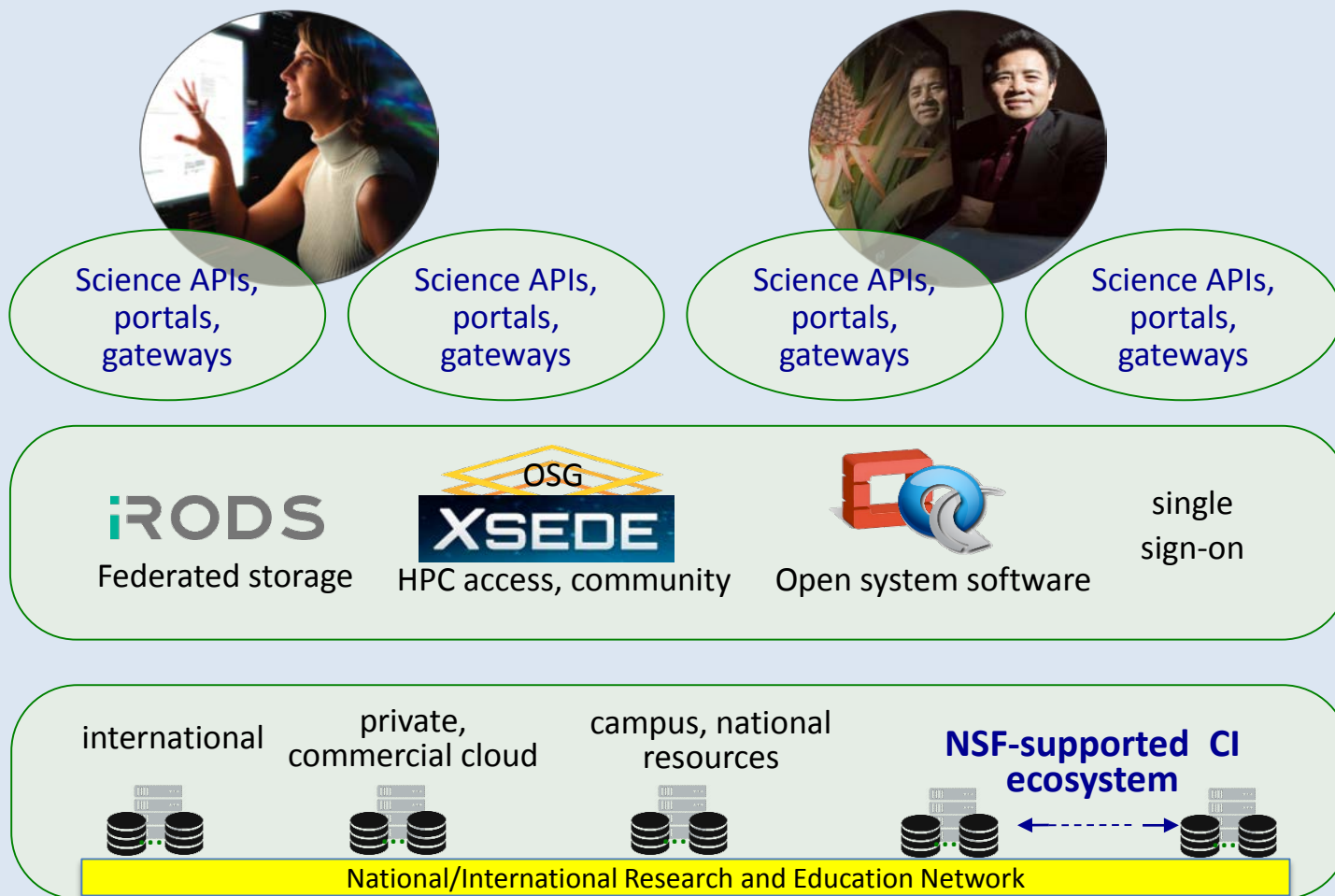- *fundamental research on data topics:* data discovery and integration, predictive analytics, data mining, machine learning; data semantics, open data-centric architectures and systems, reproducibility, privacy and protection, human-data interface

- *engagement of the research domains* supported across NSF: use advances in data science and CI to further research

- *embodiment of these innovations in a robust, comprehensive, open, science-driven, CI ecosystem*: accelerating a broad spectrum of data-intensive research, including large-scale, MREFC facilities

- *development and evaluation of innovative learning opportunities and educational pathways,* grounded in an education-research-based understanding of needs

A vision for research cyberinfrastructure

An open national research CI architecture

A vision for research cyberinfrastructure

Architecting an open national data infrastructure

Science APIs, portals, gateways

Science APIs, portals, gateways

Science APIs, portals, gateways

Science APIs, portals, gateways

New Data Services: Access, Discovery, Deep Analytics, Semantics

Existing CI services

international

private, commercial cloud

campus, national resources

NSF-supported CI ecosystem

National/International Research and Education Network
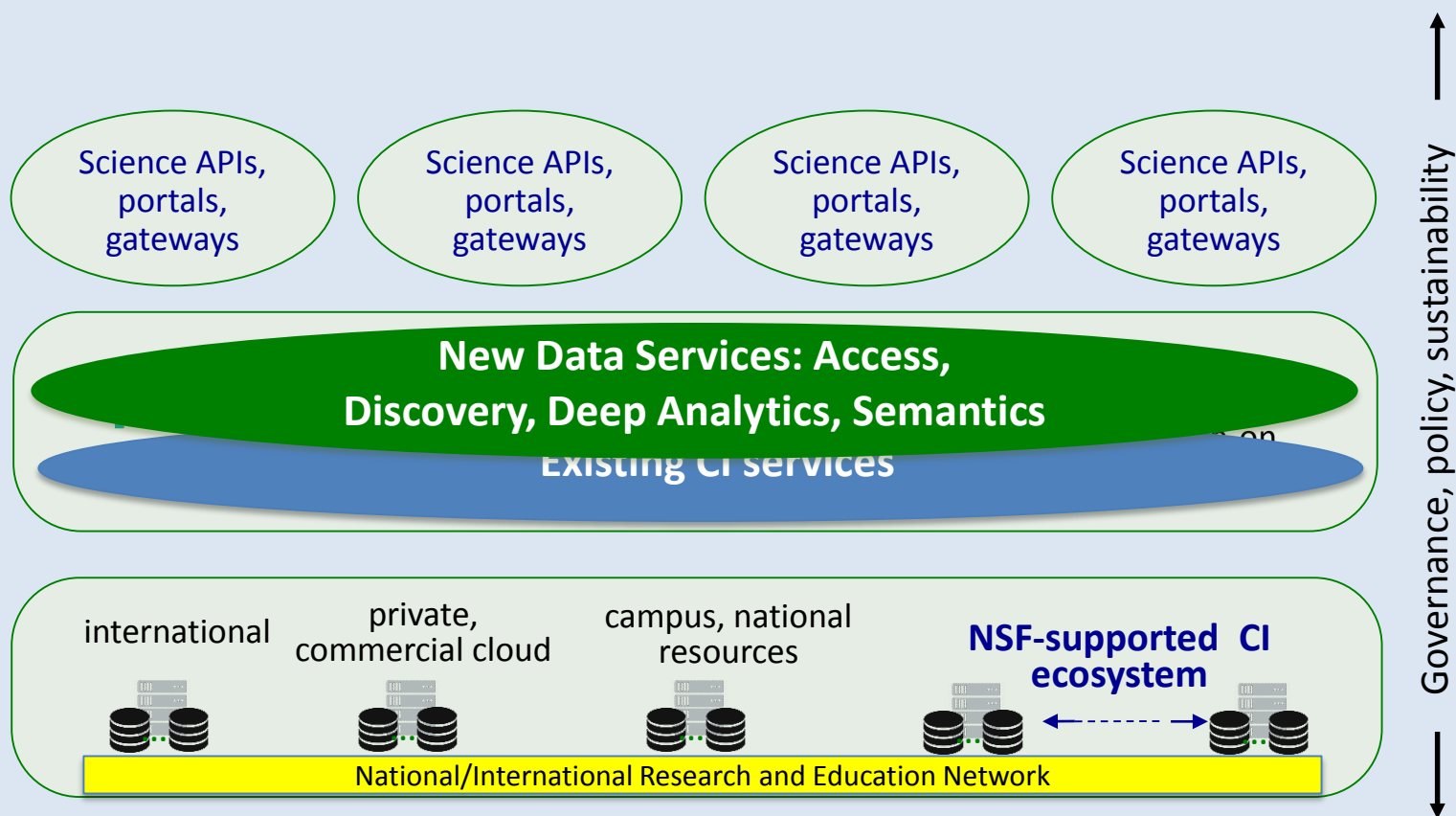
Governance, policy, sustainability

A vision for research cyberinfrastructure

Architecting an open national data infrastructure

Enabling and accelerating science drivers, including NSF initiatives & facilities

UtB, NBO

INFEWS

S&CC

Facilities, MREFC

New Data Services: Access, Discovery, Deep Analytics, Semantics

Existing CI services

international

private, commercial cloud

campus, national resources

NSF-supported CI ecosystem

National/International Research and Education Network

Governance, policy, sustainability
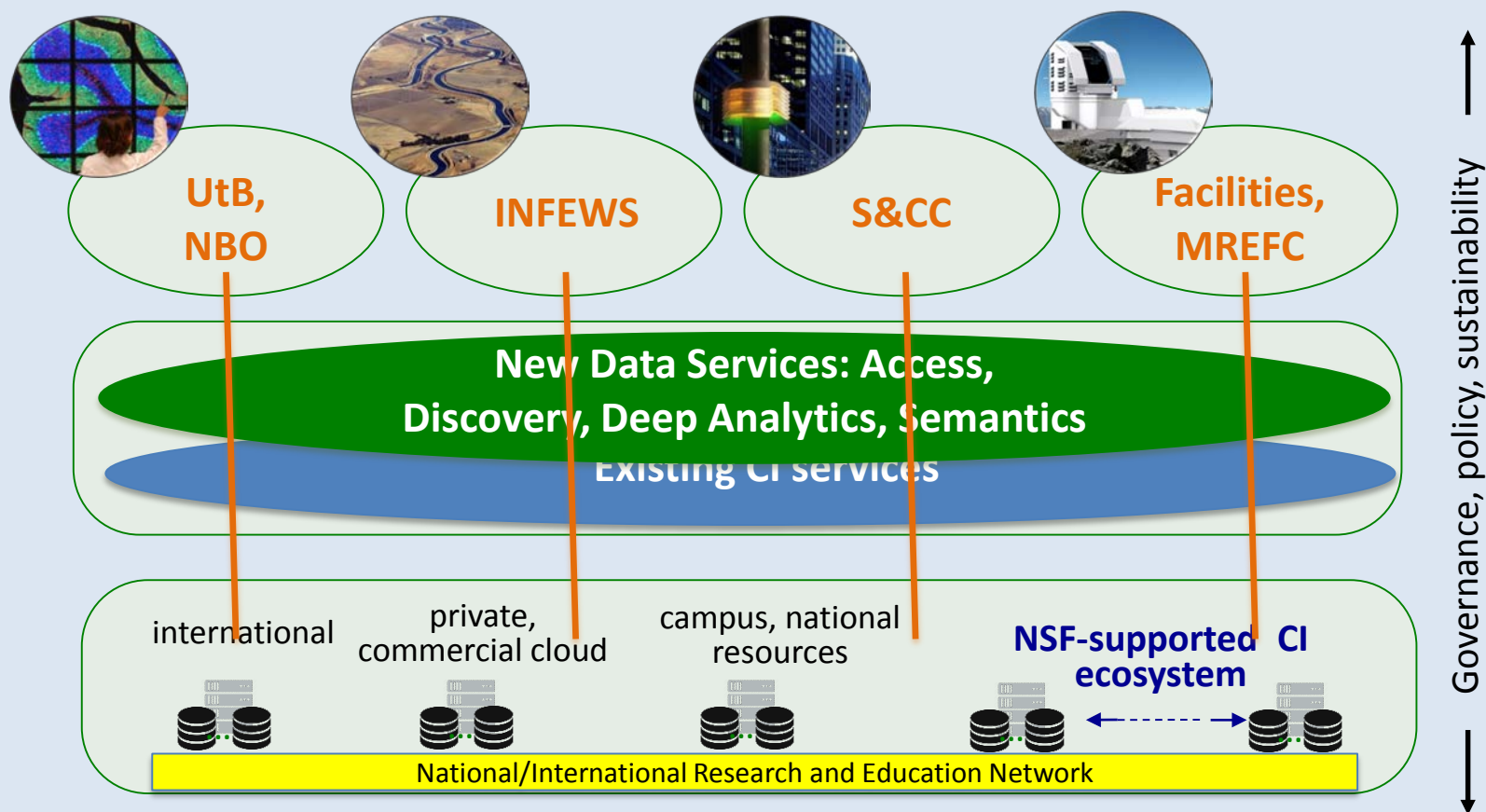
We are ready!

- NSF has help set the foundation:
  - CIF21 initiative
  - BIGDATA, ML investments
  - EarthCube
  - Research Data Alliance (RDA)
  - Data Hubs (and Spokes) investments
  - Resource Implementations for Data Intensive Research in the SBE (RIDIR)
  - DataONE
  - iDigBio
  - Public Access Plan
- Federal government partners
- Industry, state/local partners (e.g., Data Hubs)

Urgency: A national imperative

Research

Just as computational science transformed scientific research in the 20th century, so too can data science fundamentally transform scientific research in this century.

Research Infrastructure

Just as HPC and NSFNET enabled and accelerated new science and engineering research in the 20th century, so too can national data infrastructure enable and accelerate new science and engineering in this century.

If not us, who? If not now, when?

Some Related Events / Activities

- Recent meetings
    - **US-UK Health Data Science Workshop**, March 1-2, 2016, NIH Campus, Bethesda, MD. Hosted by Stanford University, in conjunction with the Research Councils of the United Kingdom (RCUK), NIH, NSF
    - NSF **BIGDATA PI Meeting**, April 20-21, Arlington, VA
    - NSF Workshop on **TFODS: Theoretical Foundations of Data Science**, April 28-30, 2016
- Upcoming events
    - CATS **Workshop on Causal Inference from Big Data**, Washington DC, June 2016
    - NAS Workshops on **Envisioning the Data Science Discipline: The Undergraduate Perspective**, Washington DC, 2016
    - NITRD Workshop on **Metrics for Assessing the Value of Digital Data Repositories,** Washington DC, 2016
    - **BD Hubs Annual Meeting,** Washington DC, 2016

An *amazing* time to be in CISE!
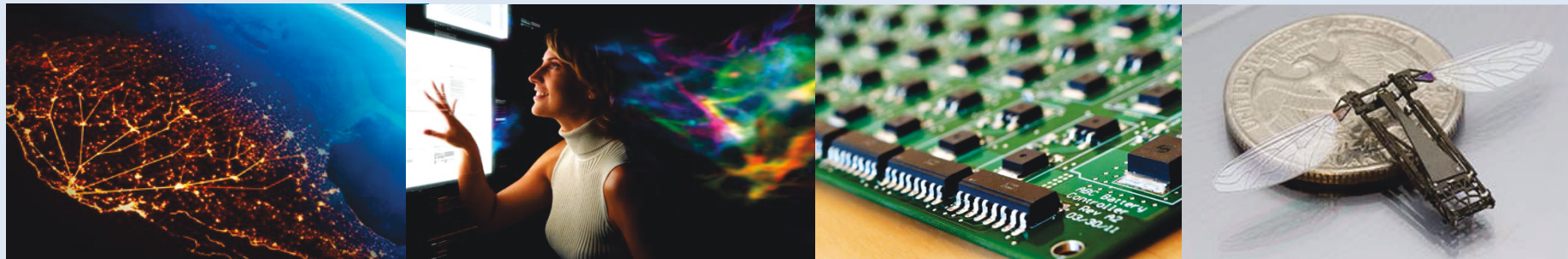
- **Ubiquity:** computing seems to be everywhere: science and engineering, workforce, societal
- **Engagement:** with many communities
- **Urgency:** rapidly expanding and evolving field - opportunity(!) in a time of fiscal uncertainty

Thanks!

- Chaitan Baru, cbaru@nsf.gov, (703) 292-4541

Smart & Connected Communities (S&CC)

Improving quality of life, health, well-being, and learning in communities

CISE Investment: $16.50 M

- Partnership among CISE, EHR, ENG, GEO, and SBE

- Network of regional research hubs to advance fundamental research

  - Advanced networking; physical sensors/devices; large-scale data management, analysis, and decision making, security, privacy

- Builds on previous investments in Urban Science and US Ignite

Image Credit: NSF

Image Credit: US Ignite
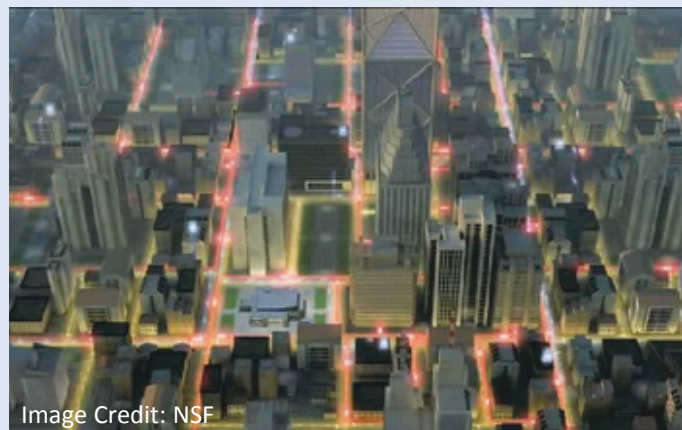
Smart & Autonomous Systems (S&AS)

Making possible intelligent physical systems that sense, perceive, and operate in dynamic, uncertain, and unanticipated environments

- Advances interstices of NRI and CPS
  - National Robotics Initiative (NRI): "co-robots" – robots that work alongside, or cooperatively with, people
  - Cyber-Physical Systems (CPS): deeply integrating computation, communication, control into physical systems, typically with provably correct behavior

Image Credit: NSF

- Exceeding today's capabilities in adaptability, autonomy, functionality, efficiency, reliability, safety, usability, recoverability, recyclability

Secure and Trustworthy Cyberspace (SaTC)

Securing our Nation's cyberspace

CISE Investment: $70.50 M

- Partnership among CISE, EHR, ENG, MPS, and SBE

- Aligns with 2016 *Federal Cybersecurity Research and Development Strategic Plan*

- CISE emphases:
  - Secure and Trustworthy Cyberspace program
  - NSF/Intel Partnership on Cyber-Physical Systems Security and Privacy (CPS-Security)

- Emerging areas:
  - Experimental testbeds
  - Science of privacy
  - Network and cloud security

Image Credit: ThinkStock

Image Credit: ThinkStock

Understanding the Brain (UtB)

Improving understanding of the brain

CISE Investment: $23.58 M

- Partnership among all NSF directorates
- Includes the BRAIN Initiative
- National Brain Observatory
  - Partnership among CISE, BIO, EHR, ENG, MPS, and SBE
  - Facilitate coordination of large-scale data collection efforts nationally and internationally to advance understanding of brain function
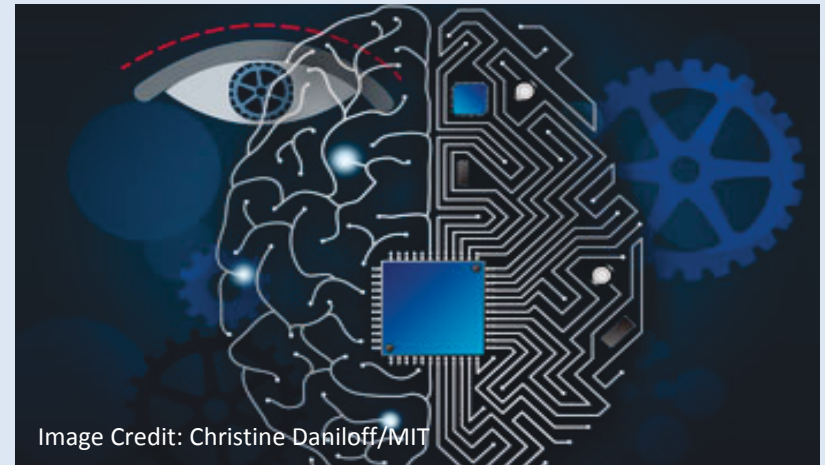

BRAIN INITIATIVE BRAIN RESEARCH THROUGH ADVANCING INNOVATIVE NEUROTECHNOLOGIES

Image Credit: Christine Daniloff/MIT