Sep 18th, 2:00 PM - 3:00 PM

Data Science R&D: Current Activities, Future Directions

Chaitan Baru
National Science Foundation

Data Science R&D: Current Activities, Future Directions

Chaitan Baru
Senior Advisor for Data Science, CISE
National Science Foundation

NSF's Perspective and Role

The NSF funds basic, curiosity driven research

To promote the progress of science;

to advance the national health, prosperity, and welfare;

to secure the national defense....

National Science Foundation
WHERE DISCOVERIES BEGIN

Big Data → Greater awareness of *all* aspects of data

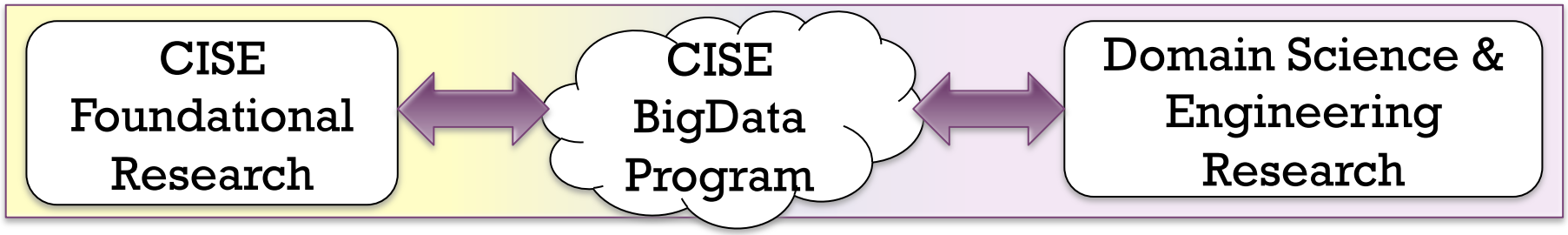
- It's not just about the cleaned up, structured data
- It's about
  - Data through it's entire lifecycle
  - All types of data
  - As much about the metadata as the data
  - And…effective, timely use of data in end applications
- Has rejuvenated and created a new research agenda around data

NSF: Research to Infrastructure Continuum

- NSF has programs across the spectrum
  - From foundational research programs to science facilities and cyberinfrastructure
  - ACI – Advanced CyberInfrastructure is a Division in CISE
- And in education and training

Big Data : Data Science :: Supercomputing : Computational Science

Big Data in the NSF CISE Context



- Big Data…

  → Data comes from the domains…

    → Big Data research needs to tie intimately to the domains

Examples of Big Science Data

- **LIGO** – Laser Interferometer Gravitational-Wave Observatory .. MPS/Physics
  - 1PB/year
- **LHC** – Large Hadron Collider
  - 4PB/year
- **LSST** – Large Synoptic Survey Telescope … MPS/Astronomy
  - 100 PB in 10 years
  - 10+ PB catalog database
- **NCAR** – National Center for Atmospheric Research … GEO/ Atmospheric Science
  - Multi-Petabytes of simulation data

Big Data in NSF Domains …

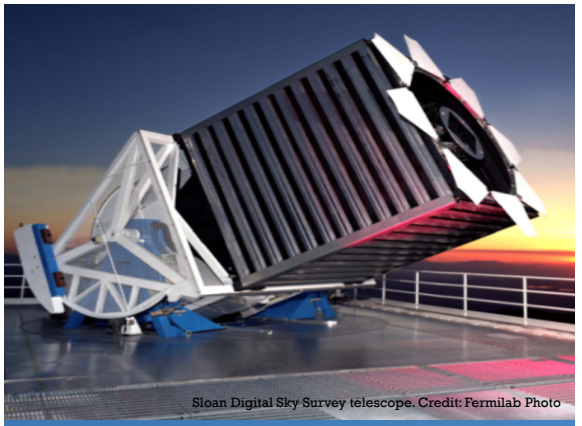
- **NCAR –** National Center for Atmospheric Research … GEO/Atmospheric Science
  - Multi-Petabytes of simulation data
- **EarthScope** … GEO/Earth Science
  - Seismic and Geodetic data archives at IRIS and UNAVCO
- **OOI** – Ocean Observing Initiative … GEO/Ocean Science
  - Just started. Data to be collected for 25 years

Big Data in NSF Domains … Persistent data

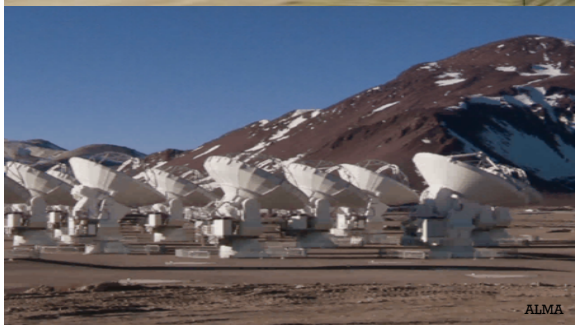
- **MGI** – Materials Genome Initiative… MPS/ Materials Science
  - Heterogeneous data collection. Novel initiative for this discipline.
- Most Are MREFC – New instrumentation projects ($100's M)
- Data is the new "instrument" !

LSST – Big Data in Astronomy

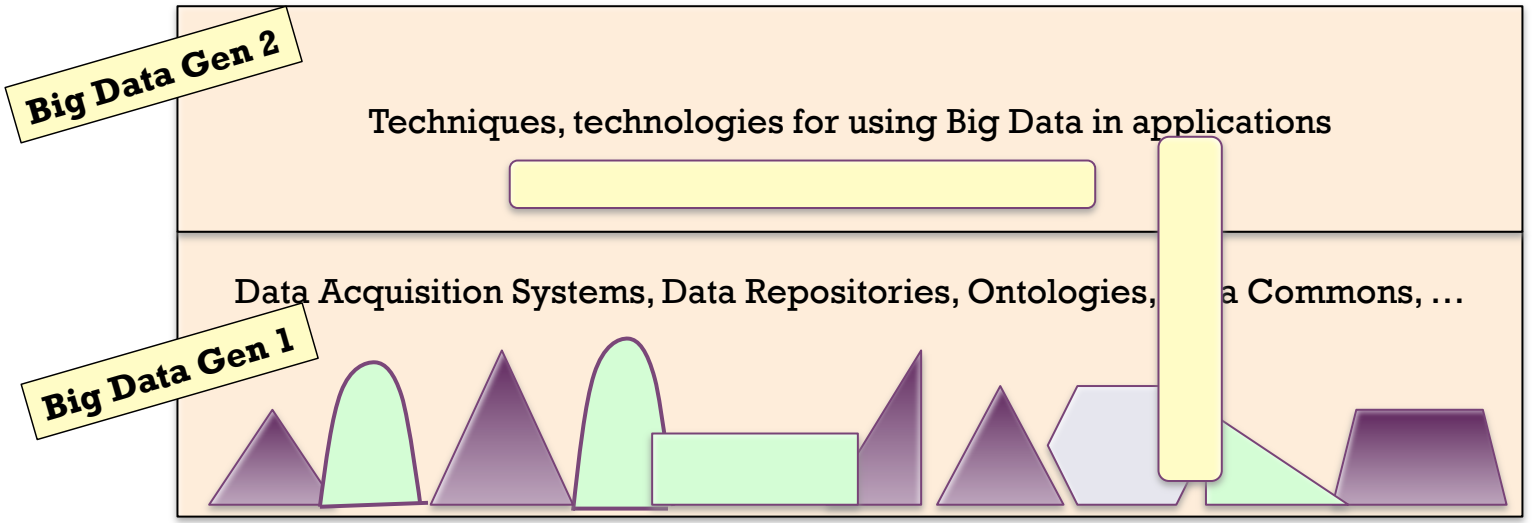
- The Sloan Digital Sky Survey in 2000, collected more data in its 1st few weeks than had been amassed in the entire history of astronomy
  - Within a decade, over 140 terabytes of information collected
- The Large Synoptic Survey Telescope due in Chile in 2016 will amass that quantity of data **in 10 days**


Sloan Digital Sky Survey telescope. Credit: Fermilab Photo


LSST


ALMA

From Infrastructure to Applications



Challenges

- Promoting, incentivizing inter-disciplinary research
- Providing researchers easy access to "interesting", "real" data
- Providing platforms for data science research
  - Hardware, Software, and Data
- Education and Development of a new pedagogy
- Sustainability of resources (data)
- Reproducibility of research
- Plugging the "brain drain"

NSF's Big Data / Data Science Investment Strategy

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

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 and Data Science at NSF / CISE

Foundational research

- BIGDATA (Foundations + Innovative Apps)
- Computational and Data Science and Engineering (CDS&E)
- Big Data for Disasters (BDD) FutureCloud

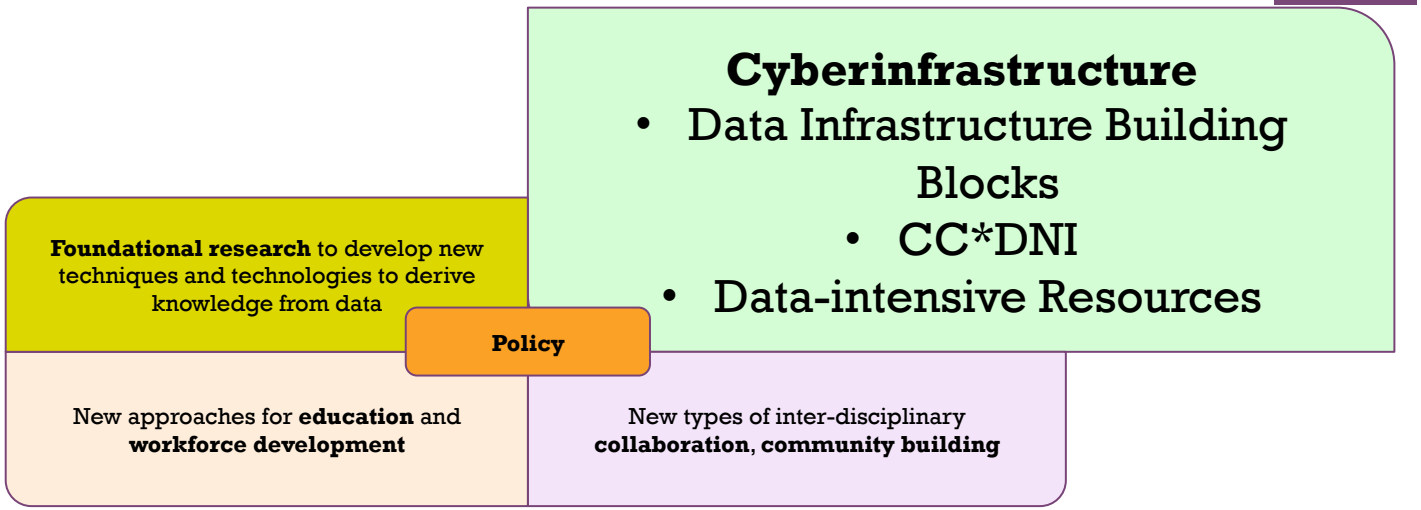
New **cyberinfrastructure** to manage, curate, and serve data to research communities

Policy

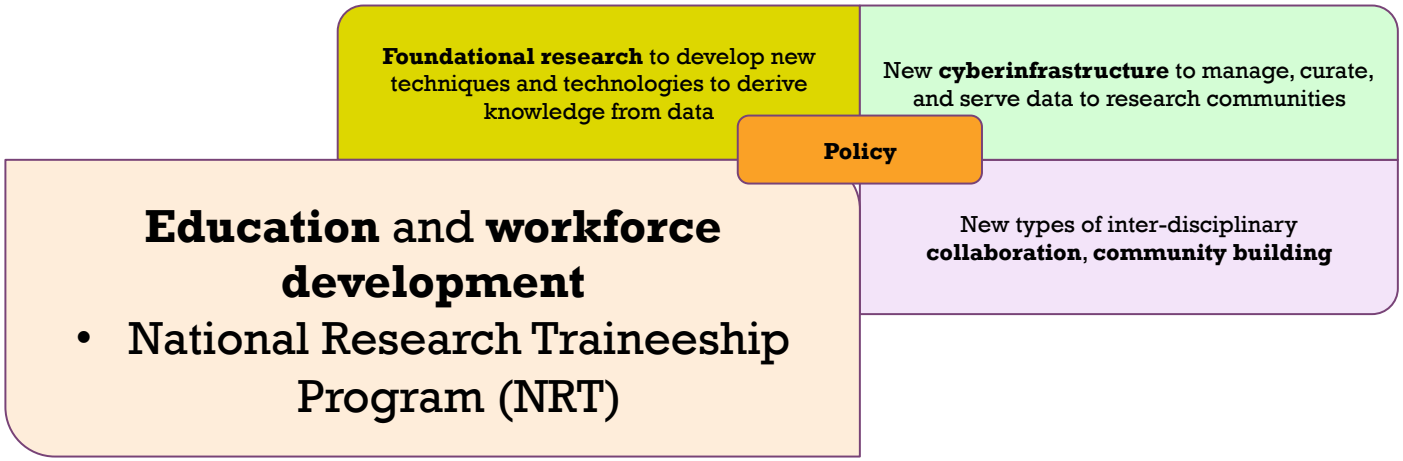
New approaches for **education** and **workforce development**

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

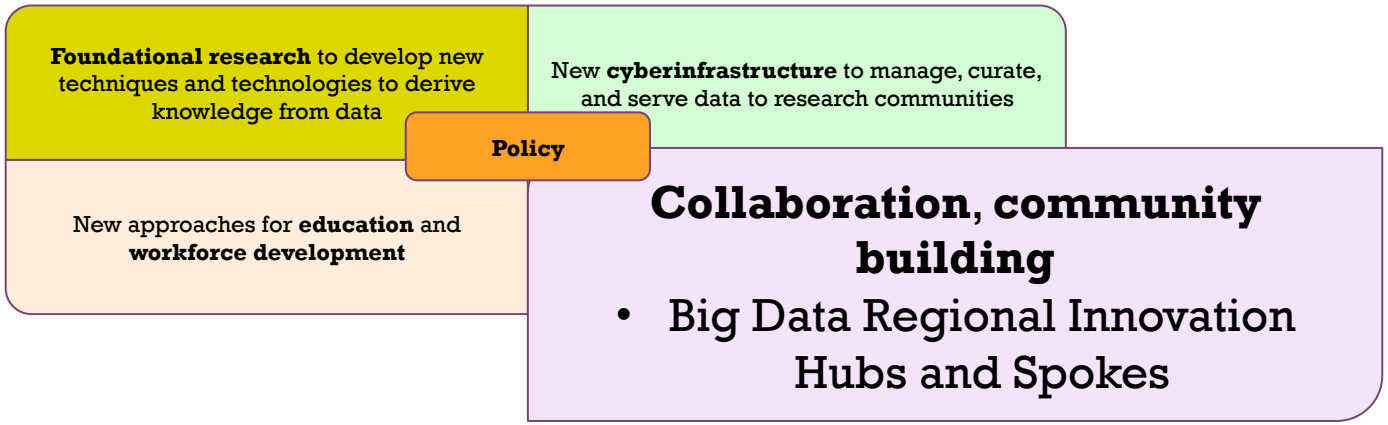
Big Data and Data Science at NSF / CISE



Big Data and Data Science at NSF / CISE



Big Data and Data Science at NSF / CISE



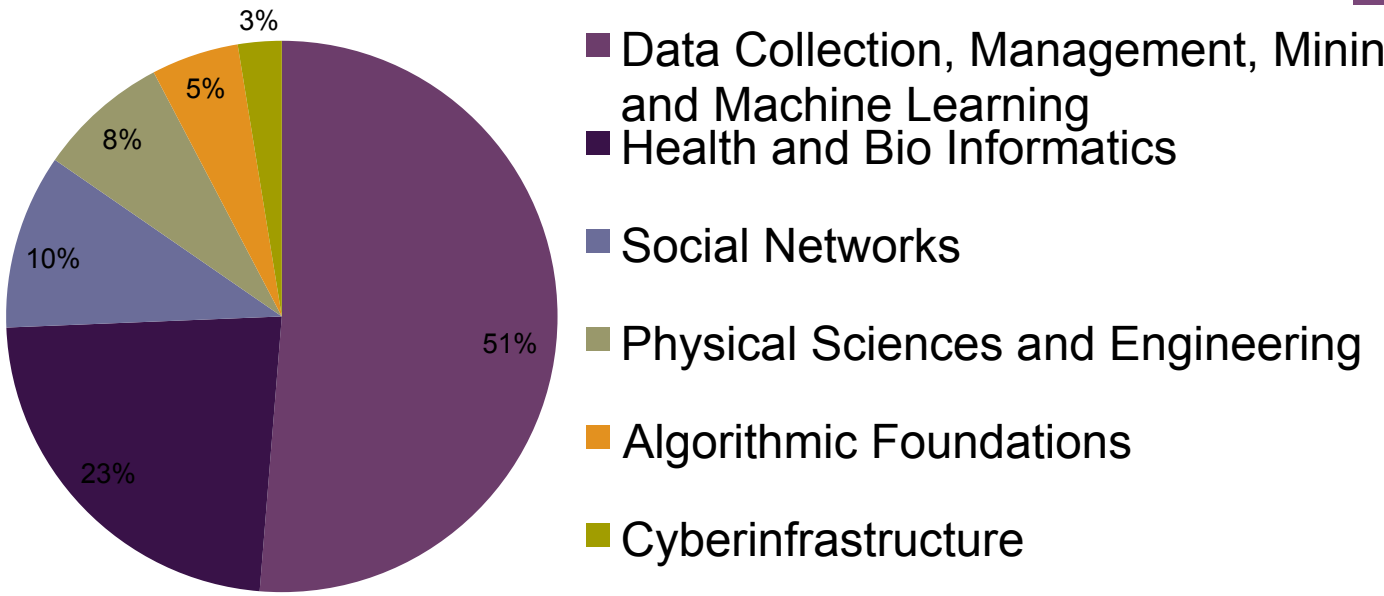
Big Data and Data Science at NSF / CISE

BIGDATA Research Program:
Critical Techniques and Technologies for Advancing Big Data Science and Engineering

- Foundational proposals: Fundamental, novel techniques, theories, methodologies and technologies of broad applicability
- Innovative Applications proposals: Novel techniques, theories, methodologies, and technologies of interest to at least one specific application
- Participation by all NSF Directorates, and Office of Financial Research (OFR), Treasury Dept

BIGDATA 2013: Solicitation

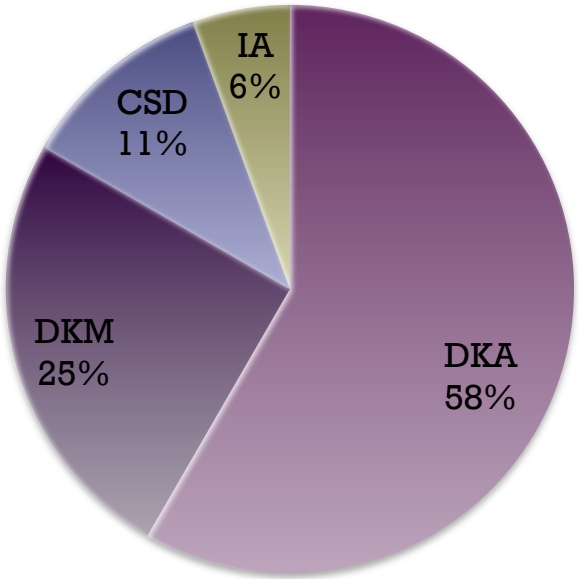
Percent by number of projects



In 2013, NSF and NIH awarded 45 projects

ranging from $250K/year for up to 3 years to $1M/year for up to 5 years.

BIGDATA 2014 Solicitation



(35 Projects)

DKA: Data and Knowledge Analytics; **DKM**: Data and Knowledge Management

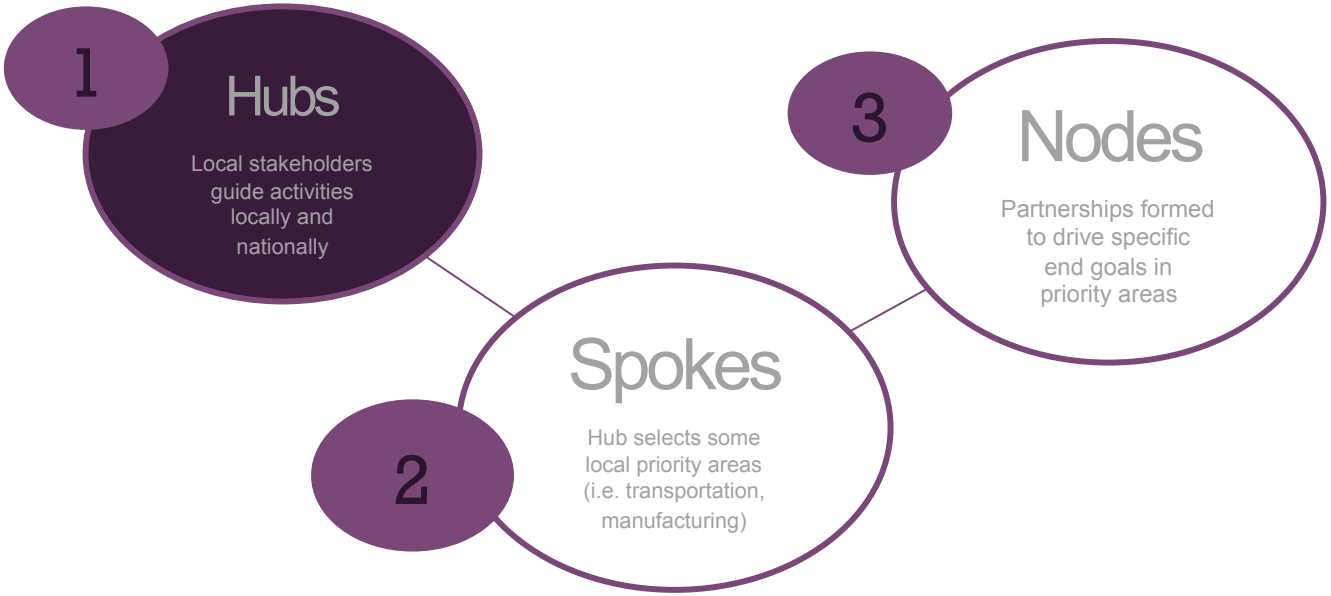
CSD: Computational Scientific Discovery; **IA**: Innovative Applications

BIGDATA 2015 Solicitation

- 30 funded projects
  - 42 proposals (12 multi-institutional collaborations)
  - 63% F; 20% IA; 17% Combined
- 6 EAGER grants

Big Data Regional Innovation Hubs

- "Hubs and Spokes"– A Nation-Wide Network to Foster Data Innovation



BDHubs: What are the Benefits?

INITIATE PARTNERSHIPS

Hubs will bring together academia, industry, non-profits, and government to initiate new partnerships.

By collectively ideating and bringing together resources from across sectors, partnerships can drive faster innovation and more novel ideas

COMMON RESOURCES

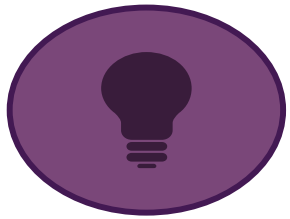
Participants can leverage the resources contributed by partners to Hub partnerships. Hubs can help develop "plug and play" infrastructure resources for partners.

Resource providers can find users that will develop novel applications for their infrastructure.

ACCESS TO TOP TALENT

In a world where demand for Big Data talent far exceeds supply, Hubs will connect partners with students in academia.

Projects with academia will train those students in projects of interest to partners before they even leave school.



SHARED BEST PRACTICES

Big Data practices, especially in a socio-technical context, are increasingly complex.

Partners can develop and share best practices in areas such as privacy, discrimination, and ethics to ensure adoption while minimizing unwanted consequences.

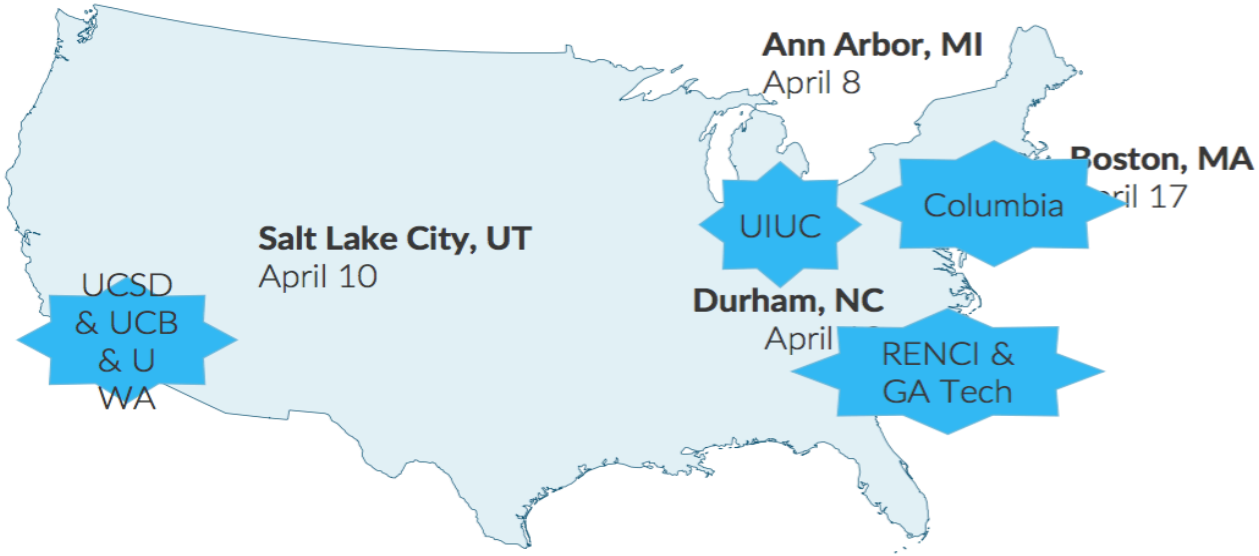
REDUCED COORDINATION COSTS

Partnerships always come with a logistical cost. With BDHubs, NSF will fund the staff and logistics support necessary for more complex collaborations, reducing overhead and maximizing benefits for participants.

Achieve collectively what is impossible individually

Yale Day of Data, September 18, 2015

www.usenix.org/bdhubs15



Throughout April 2015, NSF sponsored a series of

Regional Charrettes

usenix ASSOCIATION

knowinnovation

PRELIMINARY LISTED SECTORS OF INTEREST

Found in proposal drafts listed online

West	South	Northeast	Midwest
• Big Data Technology (Cloud, Spark)	• Big Data and Health Disparities	• Energy	• Food Water Energy
• Managing Natural Resources and Hazards	• Coastal Hazards	• Ethics	• Health Sciences, Life Sciences, Bioinformatics, Genomics
• Precision Medicine	• Industrial Big Data	• Privacy & Security	• Urban Sciences/Smart Cities
• Metro Data Science	• Materials and Manufacturing	• Finance	• Digital Agriculture (precision farming, sustainability,..)
• Data intensive discovery	• Habitat Planning	• Urbanization	• Advanced Manufacturing
		• Computational Science	• Social Network Science
		• Health	• Transportation
		• Education	• Business Analytics
		• Data Sharing	
		• Application of Data to Education	

Big Data: Recent / Upcoming Activities

- BD Hubs Charrette meeting (by invitation, for BD Hubs awardees)
  - November 3-5, 2015
- Data Science Meetup groups meeting (by invitation, for Meetup group coordinators)
- BIGDATA PI Meeting (for BIGDATA PI's), Feb 18-19, 2016
- Workshops on Big Data and IoT
  - First meeting organized by RENCI, UNC Chapel Hill, hosted by Cisco in San Jose, CA, on July 29-30, 2015.
  - Subsequent meetings planned, focusing on *Smart and Connected Communities* (with DHS), and *Personalized Medicine*

Recent / Upcoming Activities in Education

- Data Science Education workshop, Aug 5-7, Univ of Washington, Seattle, WA.
  - Sponsored by CISE and Education directorates
  - Bring together ~100 graduate students across different disciplines, engaged in data science research
- Workshop on *Envisioning the Data Science Discipline* (upcoming)
  - Joint among CISE, Education, DMS/Statistics, and Social Sciences programs at NSF; and NIH
  - To be organized by the National Academy of Science
- CISE Advisory Committee Subcommittee on Data Science Education

BIGDATA Inter-Agency Activities

- NITRD Big Data Senior Steering Group: 18 Federal R&D agencies interested in Big Data
  - Federal Big Data Strategic Plan under preparation
- NIST Public Working Group on Big Data
- XLDB.gov
  - A version of XLDB for government sector
- Collaborations with NIH
  - Quantitative Approaches to Biomedical Big Data (QuBBD)
  - Workshop on providing technical community access to "real" biomedical, medical data sets

Some Challenges…Need your inputs!

- How to "release" and provide access to interesting, "real" datasets
- Big Data Platforms at scale
  - What are the needs?
  - Is this like the supercomputer program in the "old days"?
- Privacy and ownership of data
  - Developing models to characterize, develop, nurture relationships between "data suppliers/holders" and "data collectors/ aggregators"
- Ethics and Data Science