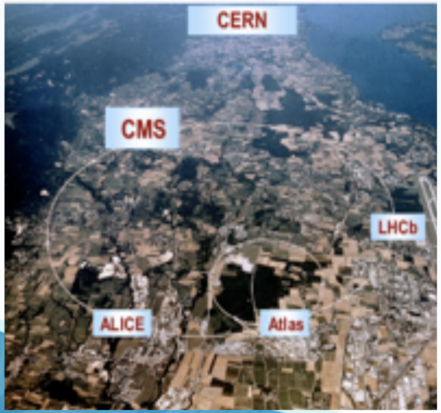
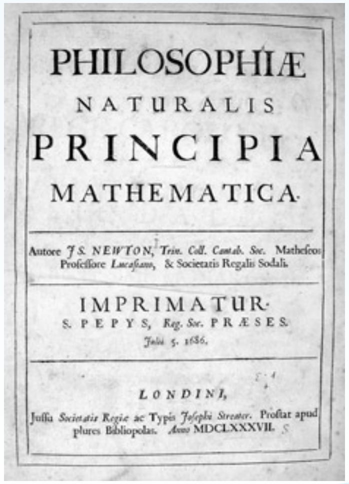
Midwest Big Data Hub

Edward Seidel
Director, NCSA
Founder Prof. of Physics, Prof of Astronomy

On behalf of the Midwest Big Data Hub



iMode and the Wireless Internet
The Perils of Data Misreporting
The Myths of Diminishing Firms
Security by Obscurity
COMMUNICATIONS
of the ACM
November 2003 Volume 46, Number 11
BLUEPRINT FOR THE FUTURE OF
HIGH-PERFORMANCE NETWORKING
MEASURING PRODUCTIVITY IN THE SOFTWARE INDUSTRY
GREAT PRINCIPLES OF COMPUTING



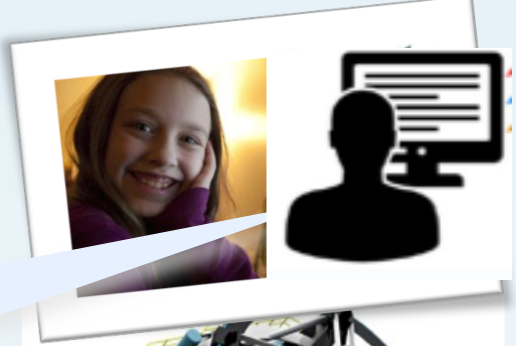
PHILOSOPHIÆ
NATURALIS
PRINCIPIA
MATHEMATICA.

Autore J S. NEWTON, *Trin. Coll. Cantab. Soc.* Matheſeos Profeſſore *Lucaſiano*, & Societatis Regalis Sodali.

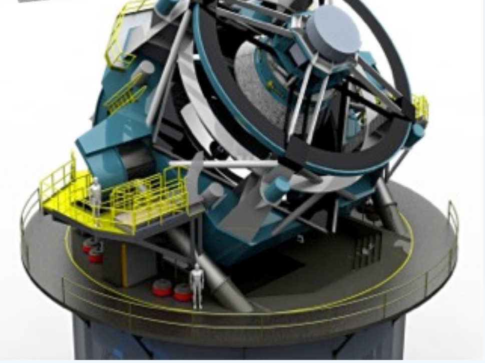
IMPRIMATUR.
S. PEPYS, *Reg. Soc.* PRÆSES.
Julii 5. 1686.

LONDINI,
Juſſu *Societatis Regiæ* ac Typis *Joſephi Streater*. Proſtat apud plures Bibliopolas. *Anno* MDCLXXXVII.

Data-enabled Transformation of Science



How can I publish, discover, verify data in this new world?



Astronomy 1500- 2000:

- Single scientist looks through telescope
- Record KB of data in notebook
- Require reproducibility

Sloan Digital Sky Survey 2000+

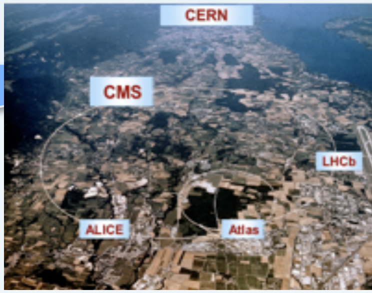
- Record data for decade (40TB)
- Serve to entire world
- Thousands of scientists work “together”

- DES (now)
  - 200GB/night
  - PB in decade
- LSST (6 years)
  - Record data for decade
  - SDSS/night!
  - 200 PB/decade

Big Data vs The Long Tail of Science

Fundamental Observation:
Science, industry, society…
communities communicate
by sharing data…

- Many “ al”
  - H
    pr d
- What abou other 99%)?
  - 1000s of biologists sequencing communities of organisms
  - Thousands of chemists and materials scientists developing a “materials genome”
    - Industry access can speed product development
  - Characteristics:
    - Heterogeneous, perhaps hand generated
    - Not curated, reused, served, etc…



MATERIALS GENOME

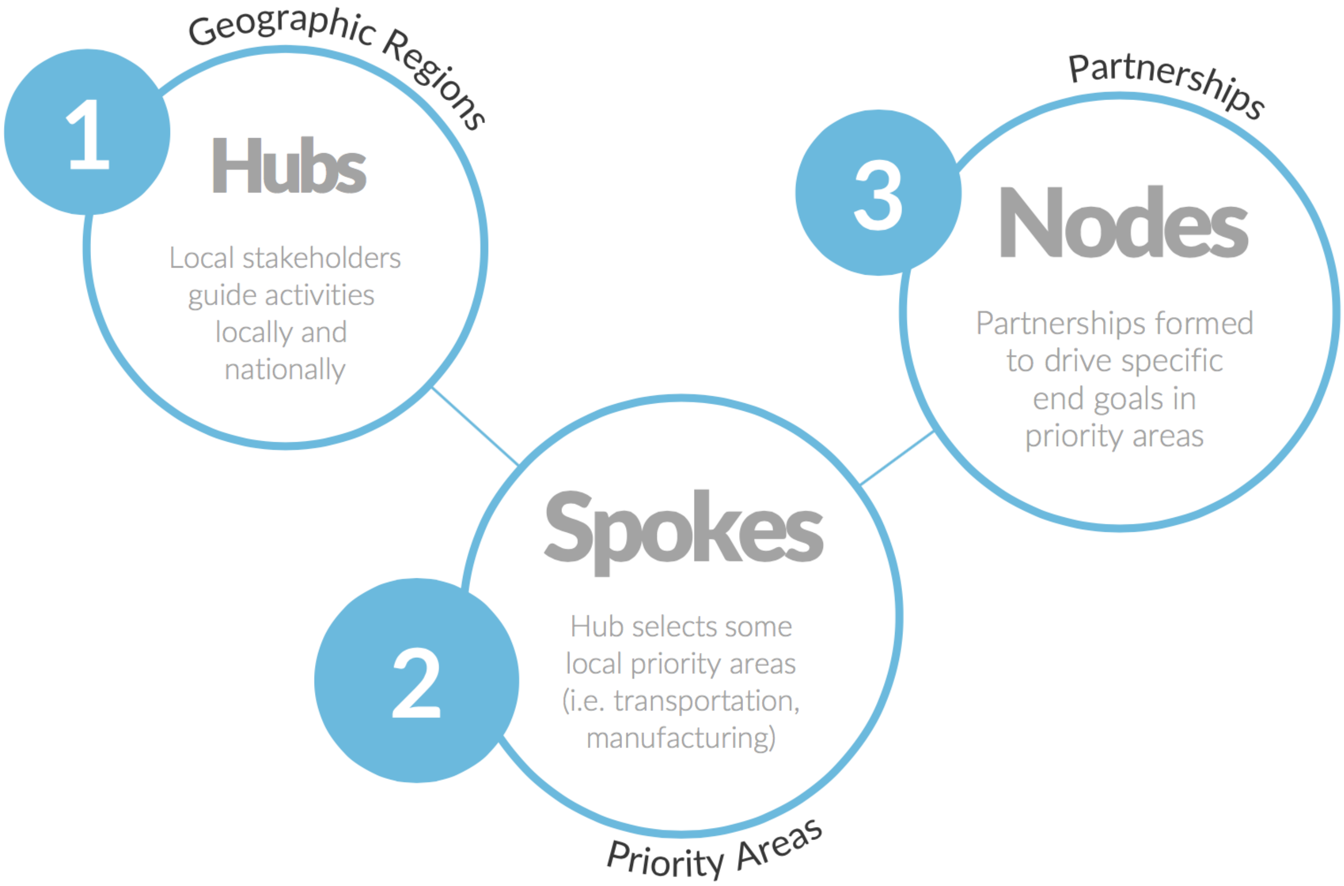
Open, Shareable Data: Critical for the future

- *Interdisciplinarity and complex problem solving*
  - Needed: ability to find, integrate results across communities
- *Public dissemination* of publicly funded research results
  - Needed: open, accessible results, searchable by public
- *Economic development*
  - Needed: availability of all the above to companies (MGI!)
- *Reproducibility of a scientific result*: heart of science
  - Needed: access to complete state of a result, including data, software, methods, (and the publication itself)
- *Accelerating discovery*: faster, deeper dissemination of results to other researchers; *Repurposing data* by others: extending in new ways
  - Needed: services to find, retrieve, analyze, describe data/results

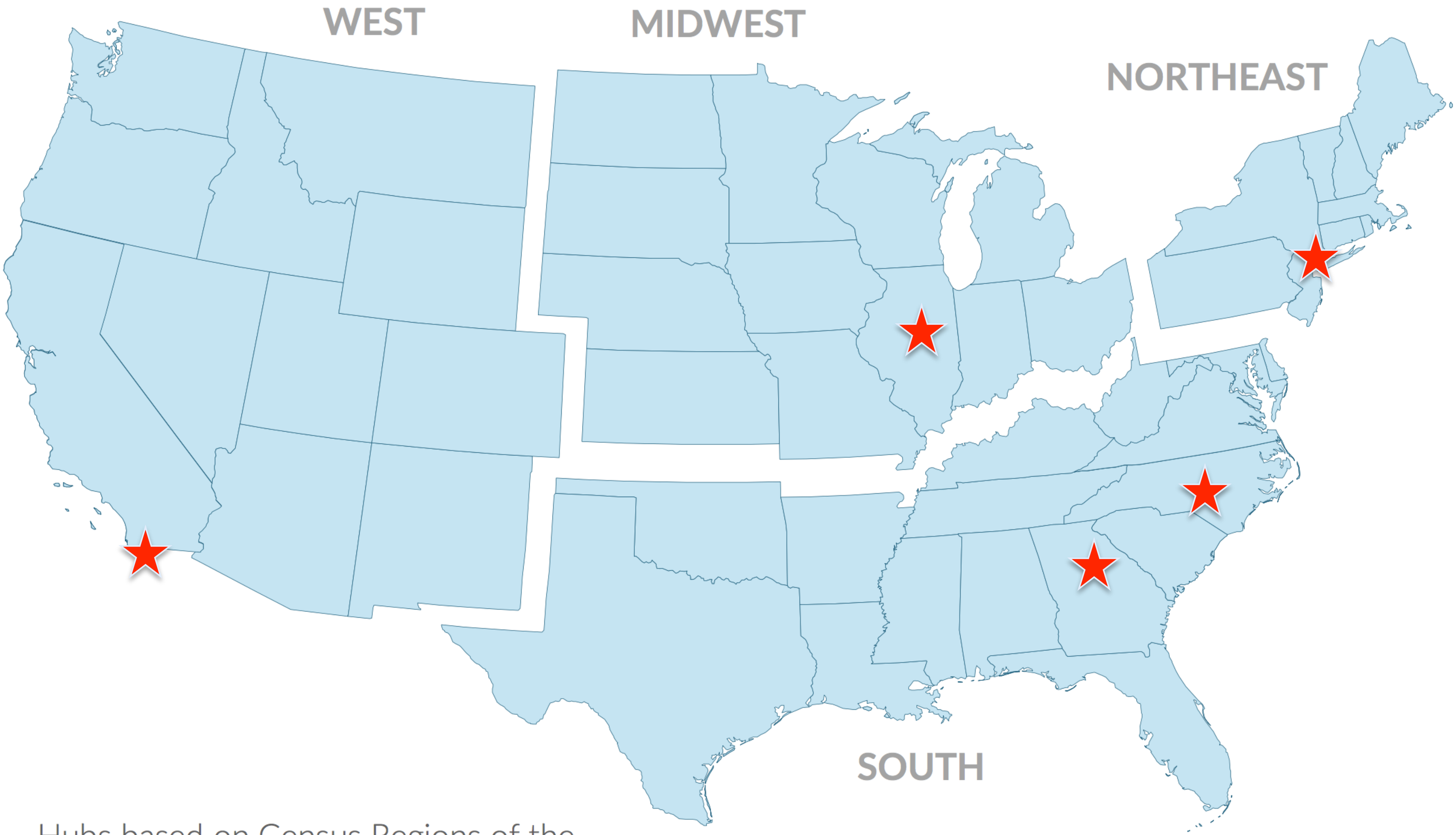
NSF Big Data Hubs

WHAT IS THE BDHUBS NETWORK?

"Hub and Spoke"– A Nation-Wide Network for Data Innovation



Geographic Regions
1
Hubs
Local stakeholders guide activities locally and nationally
Partnerships
3
Nodes
Partnerships formed to drive specific end goals in priority areas
Spokes
2
Hub selects some local priority areas (i.e. transportation, manufacturing)
Priority Areas



Hubs based on Census Regions of the

United States

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

SEEDCorn: Sustainable Enabling Environment for Data Collaboration (aka MBDH)

- A partnership of academia, government, industry, nonprofits
- Led by Illinois, Indiana, Iowa State, Michigan, North Dakota
- Over 100 partners already, e.g.…
  - Colleges, Universities, Medical Centers, of all types…
  - Industry, UI Labs, …
  - States, Cities of Chicago, Detroit, …
- Formal announcement Oct 28 (shhh!)

Why MBDH, and What does it do?

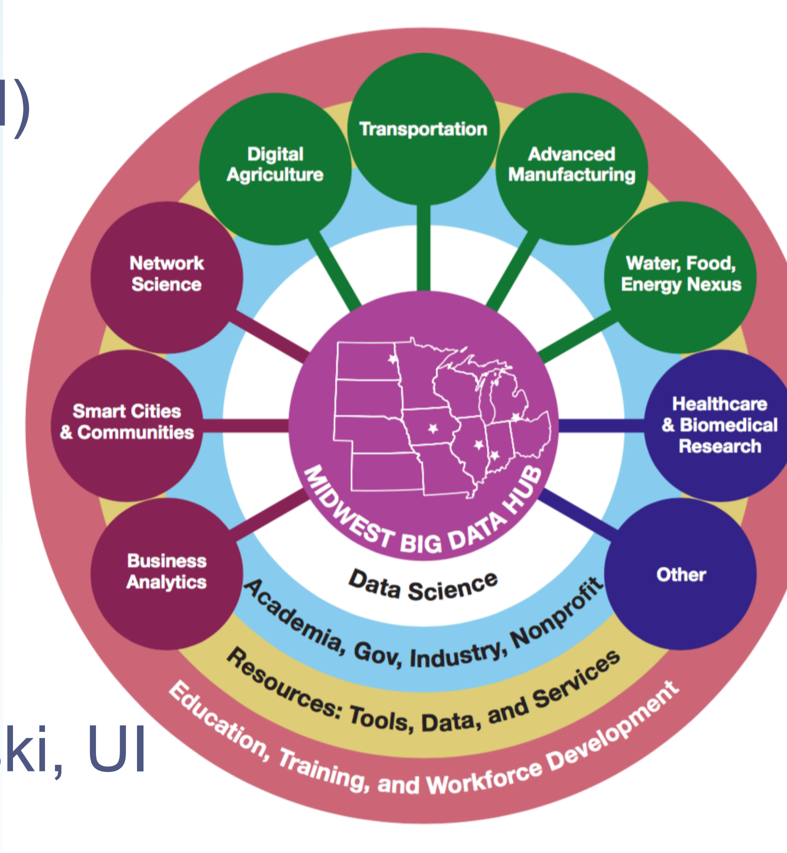
- Challenges in collecting, managing, serving, mining, and analyzing rapidly growing and increasingly complex data and information collections to create actionable knowledge and guide decision-making
- All sectors of society are profoundly impacted and need novel solutions that leverage the breadth of expertise in academia, industry, and government
- MBDH is a regional structure to bring together communities across region, nation
  - Working groups formed, interim steering committee, workshops planned
  - All sectors represented

Goals and Outcomes Expected

- Strengthening, creating and securing funding for 20-30 new public-private partnerships
  - Additional funding from agencies (NSF, NIH, DOE, NIST…), NGOs, governments, industry will be sought..
- Accelerating technology transfer projects
- Introducing new Big Data educational activities into universities, industry and government, including data policies, management, social impacts and best practices
- Starting pilots in a common data environment hosted by the National Data Service
- Developing and implementing new sustainability models

Spokes Supported by MBDH

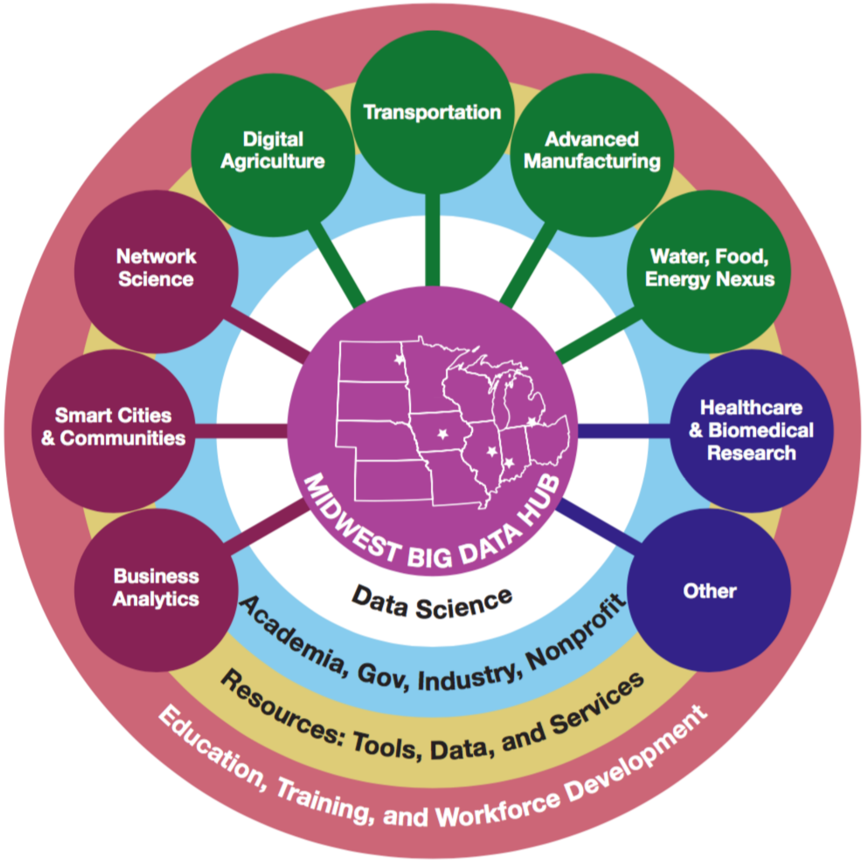
- Society
  - Network Science (Pescosolido, IN)
  - Urban Science (Catlett, UC)
  - Business Analytics (Chinnam, Wayne St)
- Natural & Built World
  - Digital Agriculture (Nusser, IA St)
  - Transportation (Jagadish, MI)
  - Advanced Manufacturing (Nowinski, UI Labs)
  - Water, Energy, Food (Nahrstedt, IL)
- Healthcare & Biomedical Research
  - Athey, MI



Transportation
Digital Agriculture
Advanced Manufacturing
Network Science
Water, Food, Energy Nexus
Smart Cities & Communities
Healthcare & Biomedical Research
Business Analytics
Other
MIDWEST BIG DATA HUB
Data Science
Academia, Gov, Industry, Nonprofit
Resources: Tools, Data, and Services
Education, Training, and Workforce Development

Crosscutting Rings Supported by MBDH

- Data Science
  - Plale, IN
- Education
  - Kliemann, IA St
- Data Tools and Services
  - Seidel, IL


Transportation
Digital Agriculture
Advanced Manufacturing
Network Science
Water, Food, Energy Nexus
Smart Cities & Communities
Healthcare & Biomedical Research
Business Analytics
Other
MIDWEST BIG DATA HUB
Data Science
Academia, Gov, Industry, Nonprofit
Resources: Tools, Data, and Services
Education, Training, and Workforce Development

Activities Planned

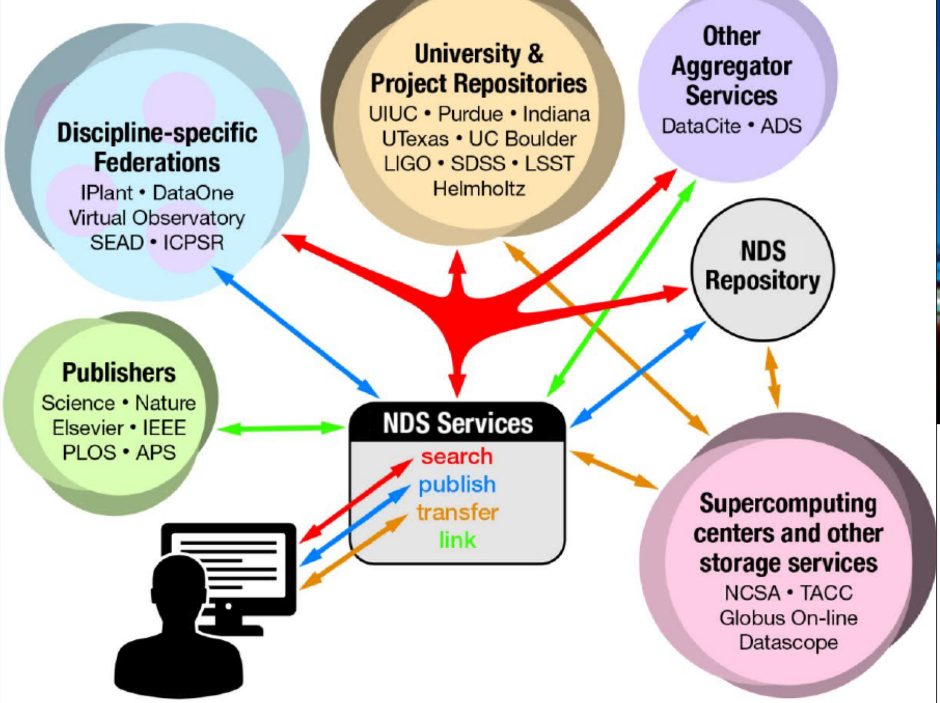
- NSF Charrette in DC Nov 3-5
  - All four hubs attending
- MBDH governance structure interim so we can function
  - Formal steering committee to be elected in spring
- Workshops
  - All Hands, various other workshops to organize spokes and rings
  - Throughout the region…see website for details (coming soon!)
- Pilot projects involving data services
- Proposals to fund projects will be developed
  - All agencies...NSF helps fund the hub that can support activities from anywhere...
  - Contact us if you want to join or have ideas

RELATED ORGANIZATIONS

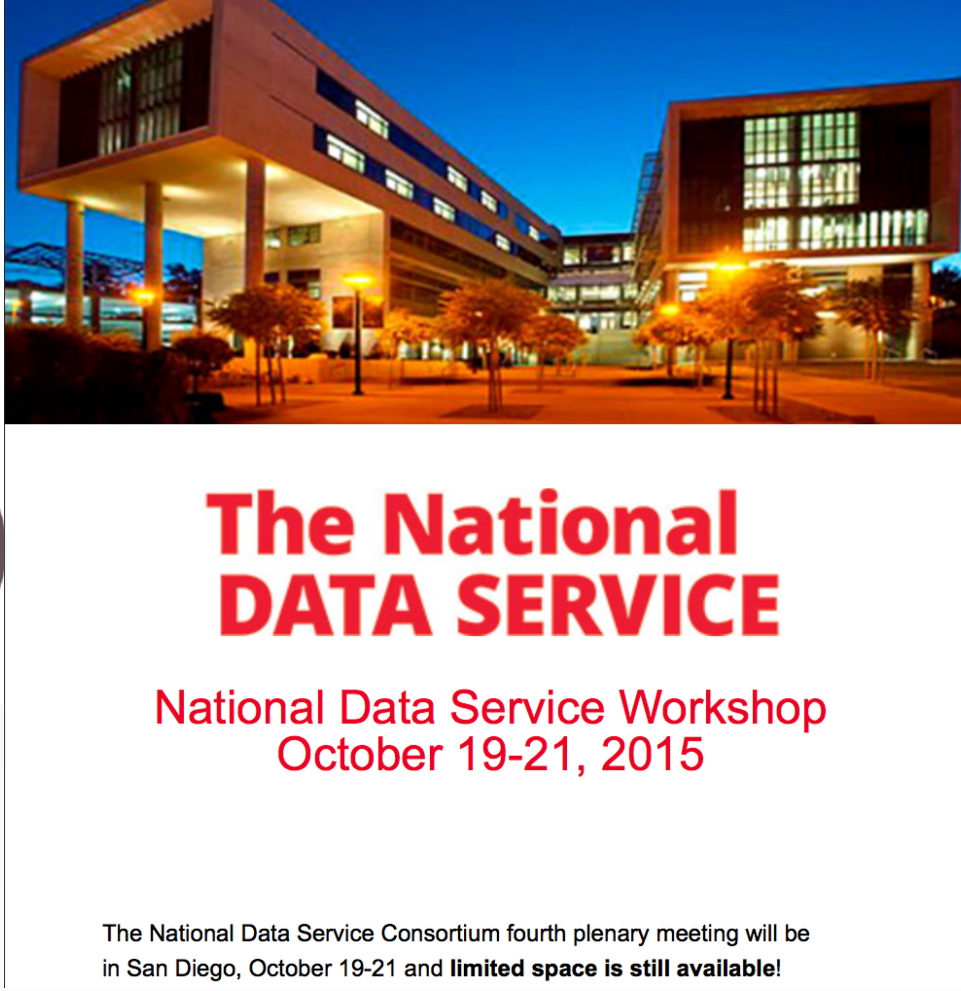
Research Data Alliance

- The Research Data Alliance (RDA) builds the social and technical bridges that enable open sharing of data.The RDA vision is researchers and innovators openly sharing data across technologies, disciplines, and countries to address the grand challenges of society.
- Over 3000 members, 104 countries
- Dozens of working and interest groups
  - E.g., Data Fabric, Agricultural Data, Life Sciences, etc…

Research Data Sharing without barriers



Providing a missing fabric…



The National DATA SERVICE

National Data Service Workshop
October 19-21, 2015

The National Data Service Consortium fourth plenary meeting will be in San Diego, October 19-21 and **limited space is still available**!

NATIONAL DATA SERVICE CONSORTIUM

NDS Lab and NDS Share

- NDS Labs
  - Target: friendly developers
  - A community support environment for developing, coordinating, deploying prototype service
  - Spinning disk, storage, virtual machines for and hosting services
  - Working with RDA to test/deploy WG outputs
- NDS Share
  - Target: friendly scientists
  - Experimental platform for sharing data
    - Enable anyone to create data collections, store data, get DOI
  - Include installations of community data sharing applications
- Numerous partners across USA (and elsewhere, e.g., Cardiff)
  - NDS meetings at NCAR, NIST, UT-Austin, San Diego

Summary

- Big Data Hubs are just starting (formal announcement Oct 28)
  - Broad public-private partnerships
- MBDH is a regional Hub working on behalf of entire Midwest Region
  - Executive Director Candidates sought asap
  - Plenty of time to join and help us launch!
  - See midwestbigdatahub.org for details as they develop
    - Working group white papers developed and will be there soon
- Other organizations have much to contribute
  - NSF DataNet, DIBBS projects, NIH BD2K, state, city, industry, more…can all contribute
  - NDS will support pilots requiring services; RDA has working/ interest groups of relevance
- Please join us!

Status: Numerous projects to help build out this vision

- John Towns, Interim NDS Director
- National Steering Committee Active
  - Technical advisory board for NDS Labs
- Next Meeting October 19-21 in San Diego
  - Can still register!! Nationaldataservice.org!
- Materials Data Facility
  - Funded by NIST
- RDA – NDS relationship developing
  - NDS as a place to build, test, deploy RDA outputs
  - Numerous projects under discussion
- NSF Big Data Hubs
  - BD Hub pilots will be supported

Basic Vision for Open Data and Publication Services

- Make it possible (easy) for anyone to:
    - Create a data collection and get an “identifier”…
    - Deposit it somewhere where it can be kept safe…
    - Provide services so others can find it, analyze it, repurpose it…
    - Link it to traditional (open, please!) publications…
        - OA aspects very important to this
- With these capabilities in place…and appropriate policies…
    - Many important things will happen…