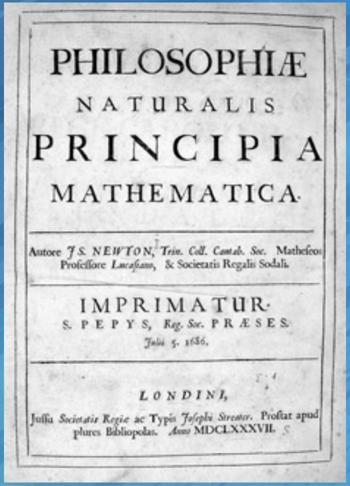
Data Intensive Science, Big Data Hubs and Services

Edward Seidel

Director, National Center for Supercomputing Applications

Founder Professor of Physics, Professor of Astronomy

University of Illinois at Urbana-Champaign

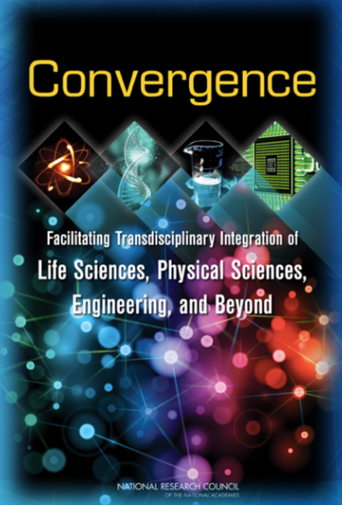


PHILOSOPHIÆ
NATURALIS
PRINCIPIA
MATHEMATICA.

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.



Convergence

Facilitating Transdisciplinary Integration of
Life Sciences, Physical Sciences,
Engineering, and Beyond

NATIONAL RESEARCH COUNCIL
OF THE NATIONAL ACADEMIES

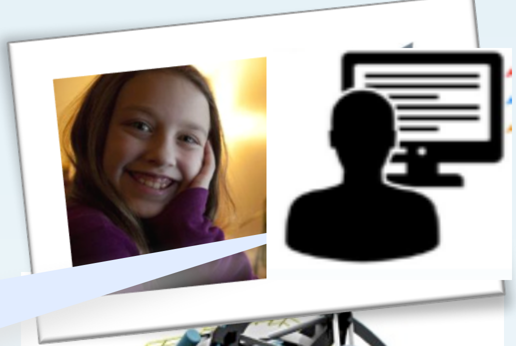
The End
of Science

The quest for
knowledge used
to begin with
grand theories.
Now it begins
with massive
amounts of data.
Welcome to the
Petabyte Age.

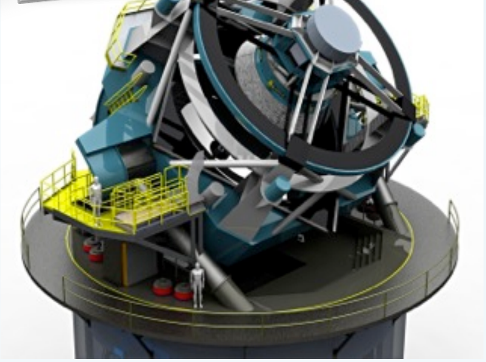
National Center for Supercomputing Applications
University of Illinois at Urbana–Champaign

VISION

Data-enabled Transformation of Science



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



Astronomy 1500- 2000:

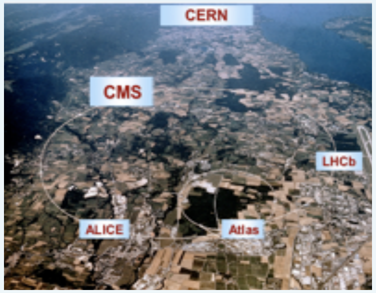
- Single scientist looks through telescope
- Record KB of data in notebook
- Require reproducibility

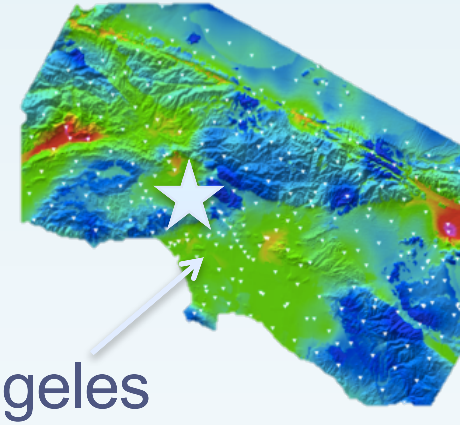
Sloan Digital Sky Survey 2000+

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

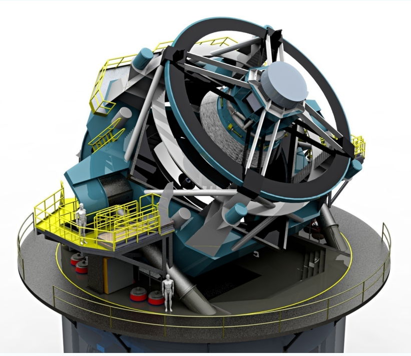
- DES (now)
  - 200GB/night
  - PB in decade
- LSST (6 years)
  - Record data for decade
  - SDSS/night!
  - 200 PB/decade

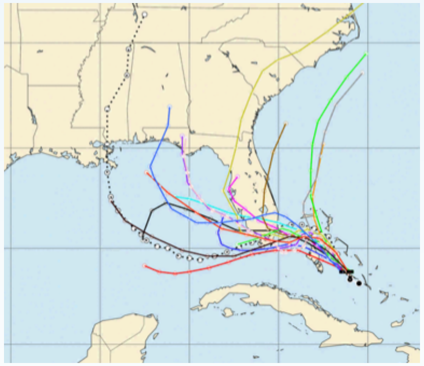
Scenarios like this in all fields

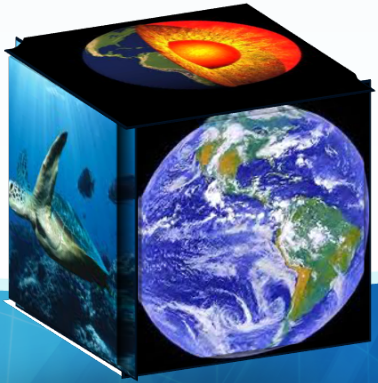


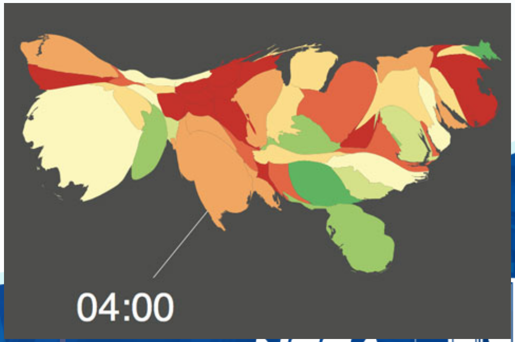


Los Angeles





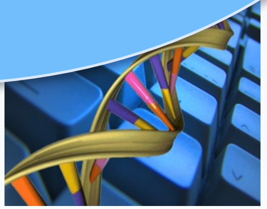




Big Data vs The Long Tail of Science

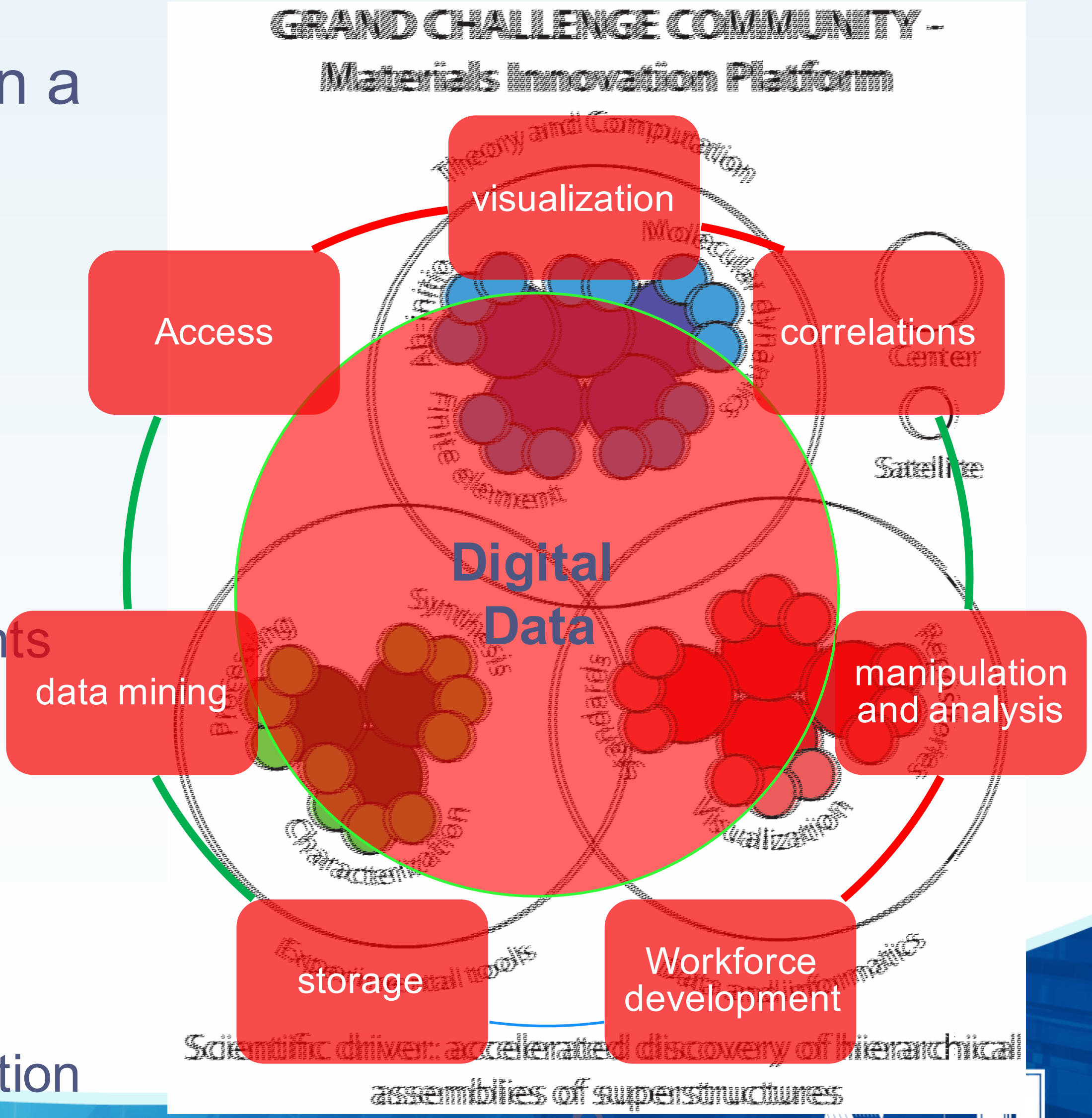
- Many "Big Data" projects are "speci
  - Tend to be highly organized, have s professionally curated, a lot attent
- What about the "Long Tail" (th
  - Thousands of biologists sequenc organisms
  - Thousands of chemist and materials "materials genome"
  - Millions of people "Tweeting"…
  - Characteristics:
    - Heterogeneous, perhaps hand generated
    - Not curated, reused, served, etc…

News Flash! NYT 6/3/13: Drug side effects discovered by mining web logs: paroxetine + pravastatin = high blood sugar!



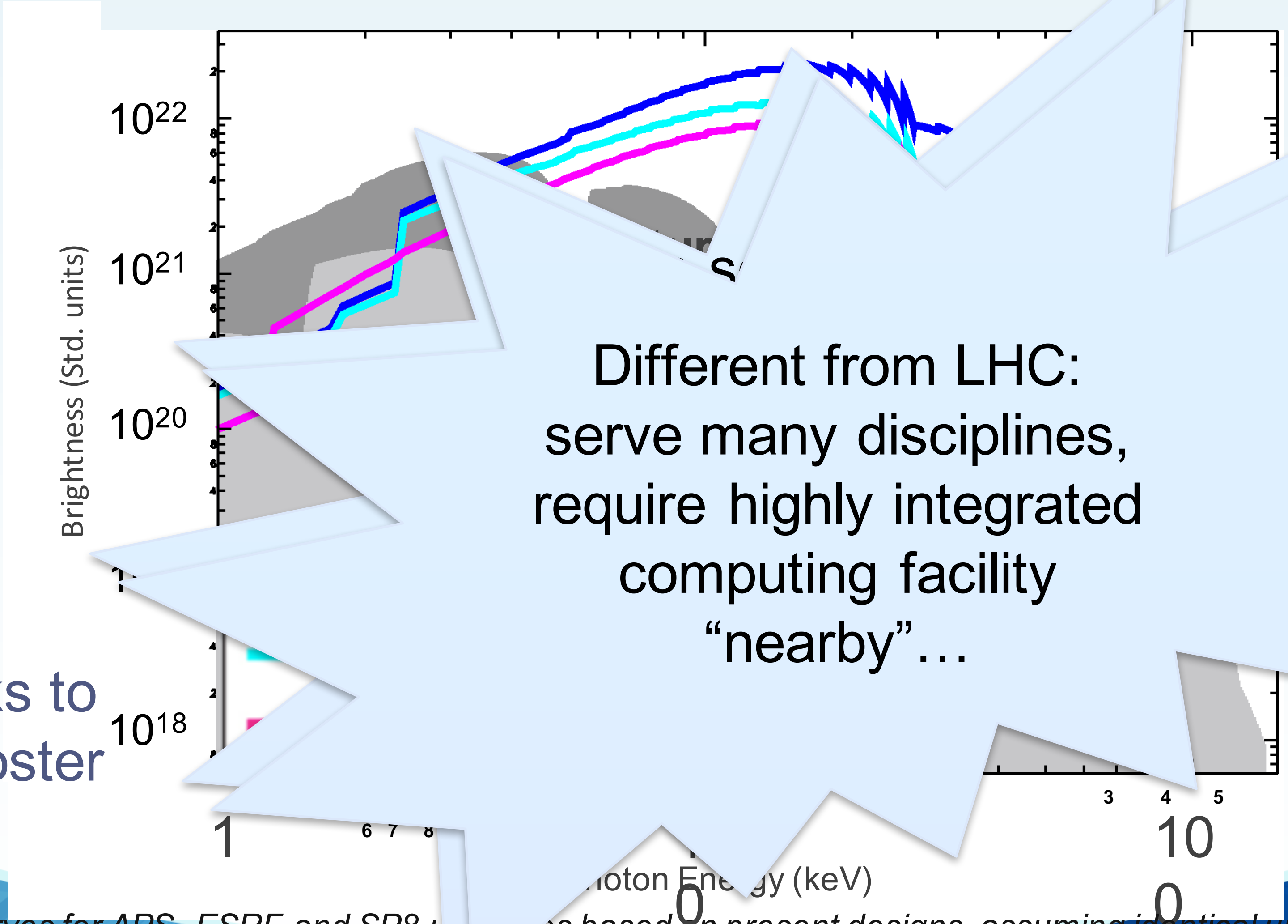
Materials Innovation
Will require Long Tail + Big Data services…

- Combining approaches in a digital world
  - Theory and computation
  - Instrumentation
  - Data and informatics
- Cyberinfrastructure
  - Software centers
  - Data services + Instruments
  - Computing
- Policy
  - Open data will accelerate discovery, enhance interdisciplinarity, speed innovation, commercialization



Advanced Photon Source Upgrade

Highly integrated computing/data services at ANL



Different from LHC: serve many disciplines, require highly integrated computing facility “nearby”…

Thanks to Ian Foster

Curves for APS, ESRF and SP8 u es based on present designs, assuming identical undulators

Basic Vision for Ope Services

- Make it po
  - Create a
  - Deposit it
  - Provide servic repurpose it…
  - Link it to traditiona ease!) publications…
    - OA aspects v portant to this
- With these capabilities in place
  - Many important things will happen…

“*We need to take steps to make scientific research data more liquid. The more we move towards open as the default for scientific research data, the more we will get out of the research enterprise. It is time to take deliberate steps to make that a reality*.” Mike Stebbins, White House OSTP

WHY

Open, Shareable Data: Critical for the future

- *Interdisciplinarity and complex problem solving*
  - Needed: ability to find, integrate results across communities
- *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
- *Economic development*
  - Needed: availability of all the above to companies (MGI!)
- *Public dissemination* of publicly funded research results
  - Needed: open, accessible results, searchable by public

BUILDING COMMUNITIES AND SERVICES TO ENABLE THEM

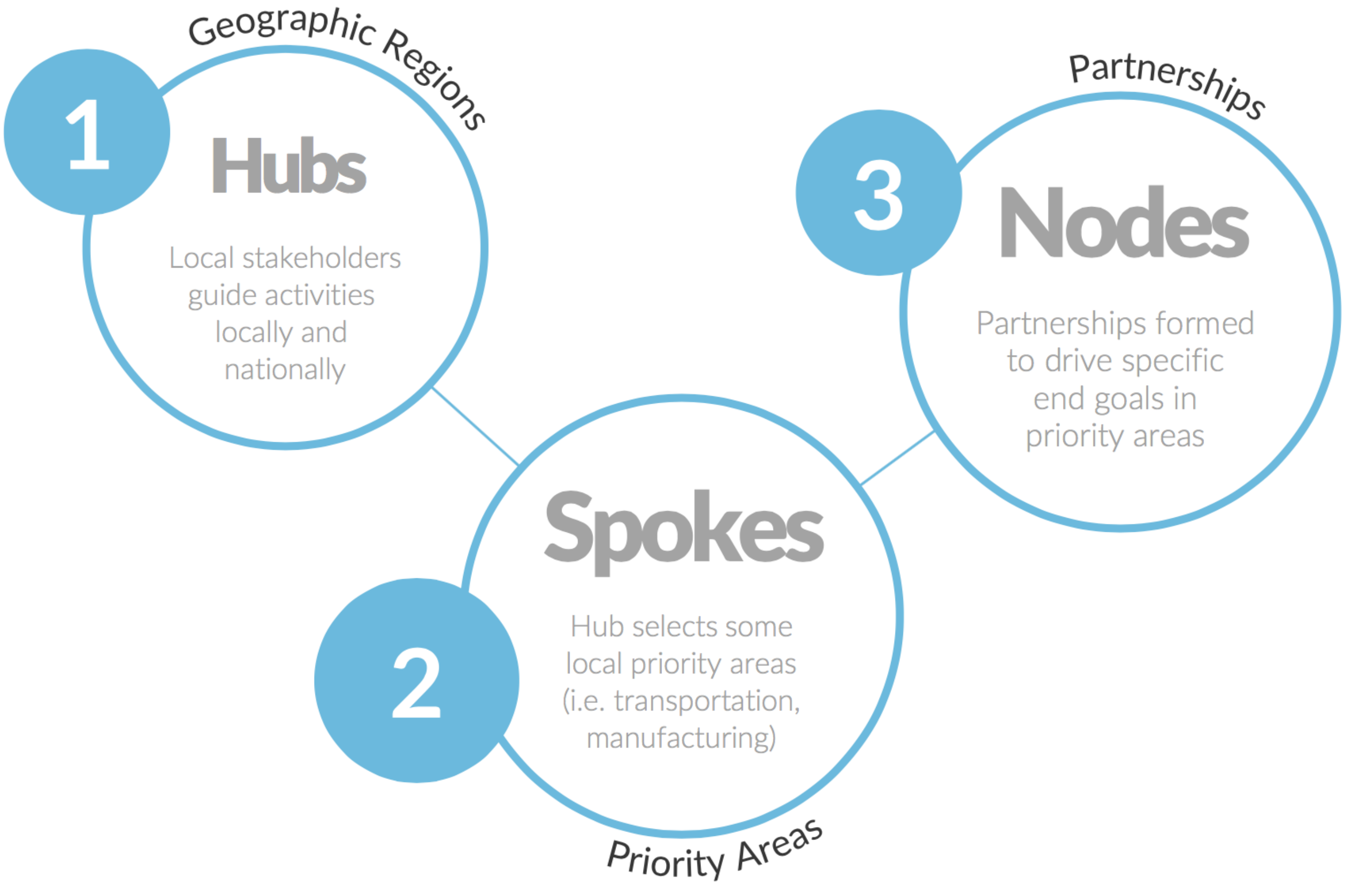
Building Grand Challenge Communities around Data

BIG DATA HUBS

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

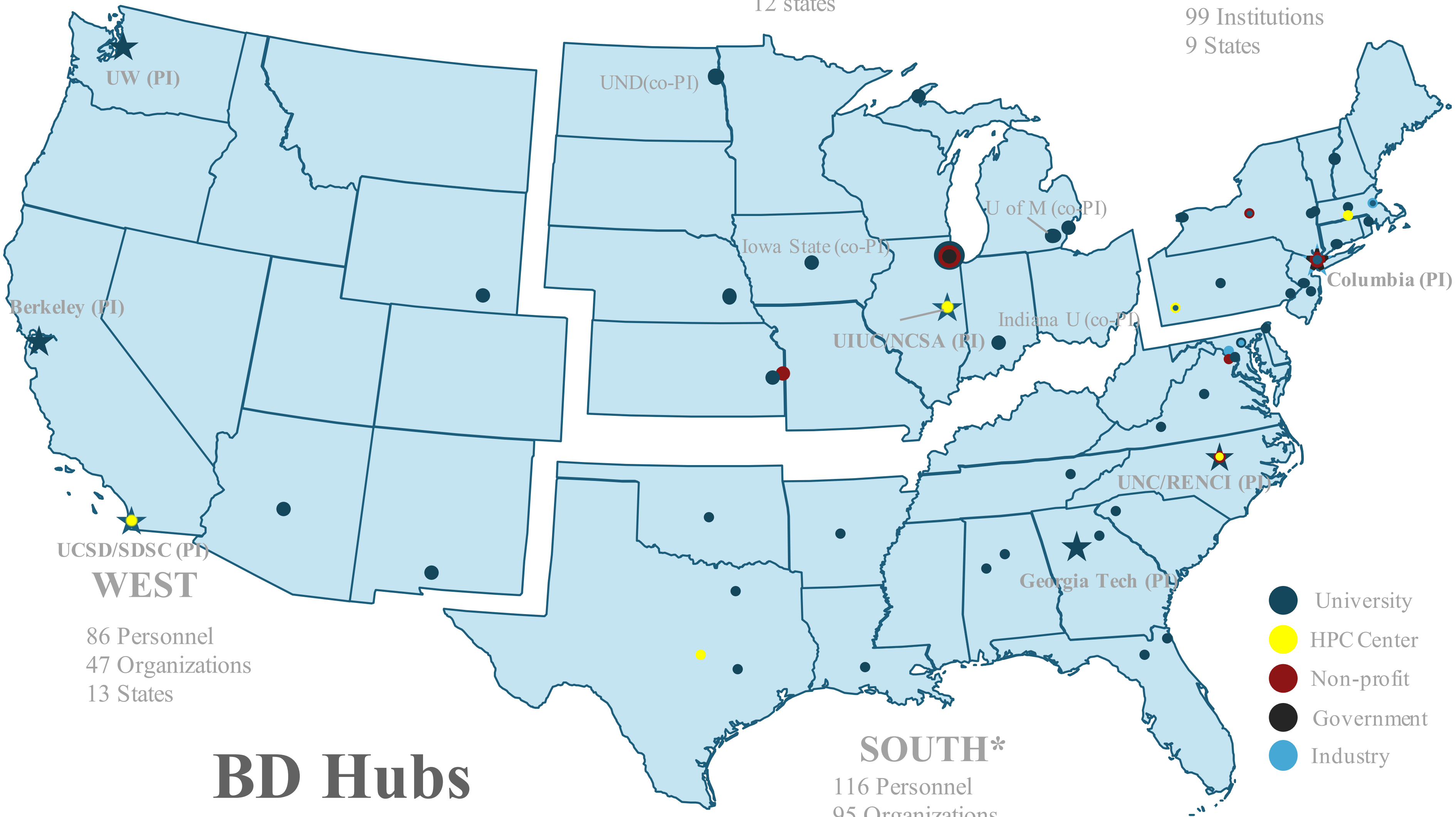
2

Hub selects some local priority areas (i.e. transportation, manufacturing)

Priority Areas

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

BD Hubs



MIDWEST

106 Personnel
79 Organizations
12 states

NORTHEAST

193 Personnel
99 Institutions
9 States

WEST

86 Personnel
47 Organizations
13 States

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

Midwest Big Data Hub (MBDH)

“Creating communities that effectively harness the growing power of data to solve societal and economic problems of relevance in the Midwest”

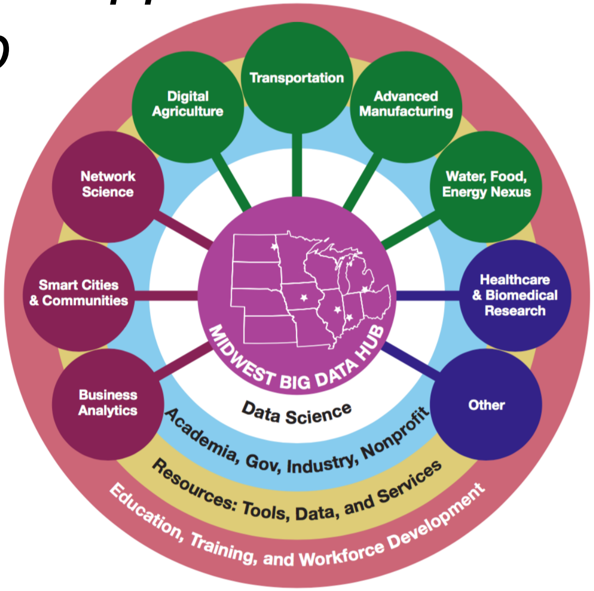
NSF Catalyzes MBDH with SEEDCorn: Sustainable Enabling Environment for Data Collaboration

- A partnership of academia, government, industry, nonprofits
- Over 100 partners already
  - Colleges, Universities, Medical Centers, of all types
  - Industry, Non-profits, NGOs
  - States, cities, communities

Spokes *Currently* Identified by MBDH

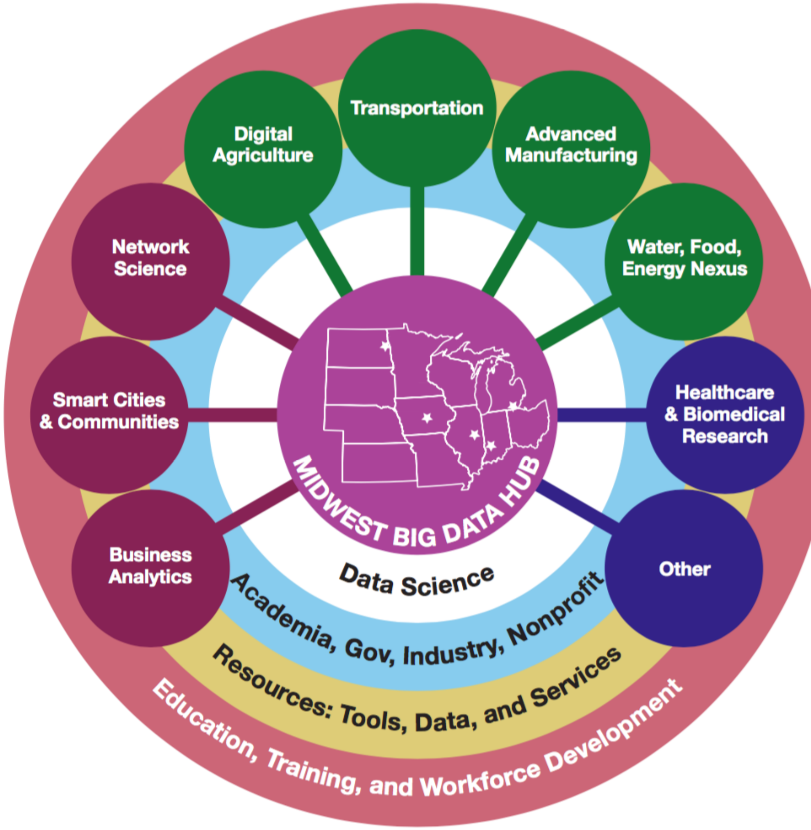
- **Network Science**
  - *Including Data Intensive Research in the Social, Behavioral, and Economic Sciences….*
- **Urban Science**
  - *Including Smart and Connected Communities…*
- **Business Analytics**
- **Digital Agriculture**
- **Transportation**
- **Advanced Manufacturing**
- **Food, Energy, Water**
- **Healthcare & Biomedical Research**
  - *Including neuroscience…*
- **Others as proposed…**
  - *Including Data Privacy*

Spokes are supported by the Hub


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**
  - *Including Data Intensive Research in the Social, Behavioral, and Economic Sciences…*
  - *Replicability and Reproducibility in Data Science*
- **Education**
  - *Including new approaches to STEM learning environments…*
- **Data Tools and Services**


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

Rings are cross-cutting, supporting all spokes

Goals and Outcomes/Impacts Expected

- Strengthening, creating and securing funding numerous new public-private partnerships
  - Additional funding from agencies (NSF, NIH, DOE, NIST, USDA…), NGOs, governments, industry will be sought
- Accelerating technology transfer projects
- Introducing new Big Data educational activities into universities, industry and government
  - Data policies, management, and best practices with real data for real impact
  - Will involve, train many young data scientists

Goals and Outcomes/Impacts Expected

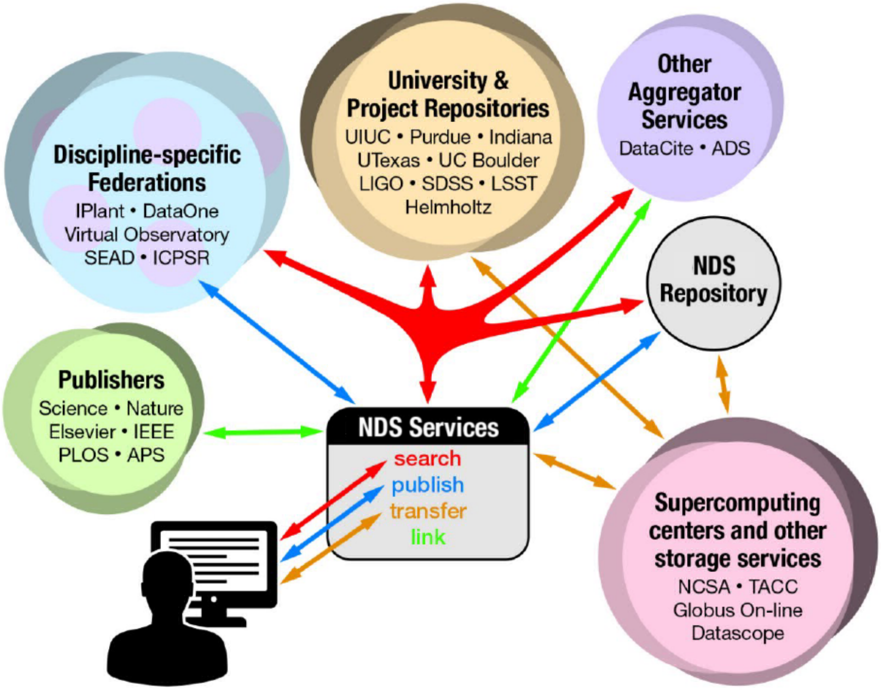
- Starting ***pilots*** in data environments (SEEDCorn!)
  - Collaborations will come together to develop and test new approaches to data sharing, policies, algorithms
  - Will work with various organizations to test pilots with real data
    - For example: helping farmers balance productivity and sustainability with detailed data on crop growth, soil conditions & environment
    - Research Data Alliance (RDA), National Data Service (NDS), other orgs. HPC centers supporting pilots
  - ¼ FTE funded to support communities like you!
- Developing and implementing new sustainability models
  - Models for long term data stewardship, private-public partnerships, educational practice
  - Different approaches will be needed!

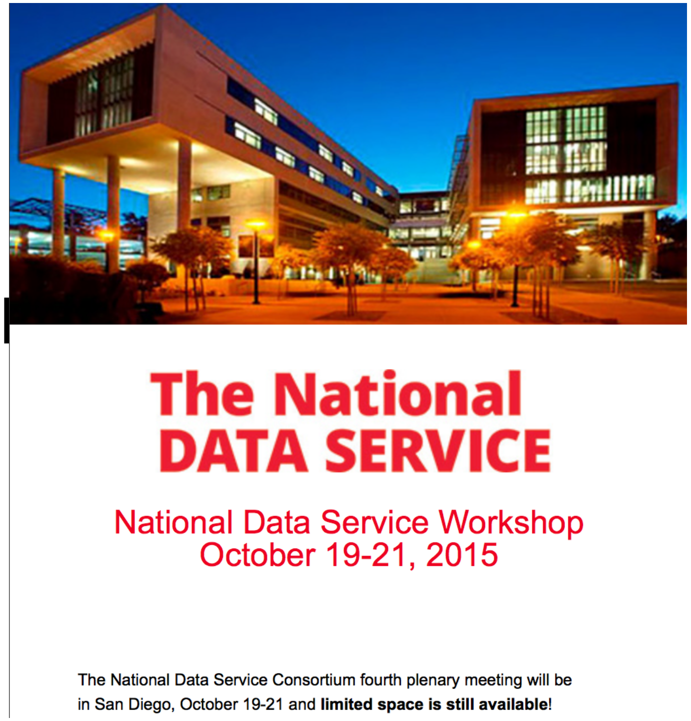
We are just getting started!

- We are bootstrappping our way to function
  - Executive Director sought!
  - You are invited to join!
- December: 45 LOIs for NSF Spoke Proposals!
  - Full proposals due in February…
- All Hands meeting in late March: TBD
- Check out our website at **midwestbigdatahub.org**
  - White papers, *interim* steering group leadership, and more...

Creating Federated Data Services to Support These Communities

NATIONAL DATA SERVICE CONSORTIUM





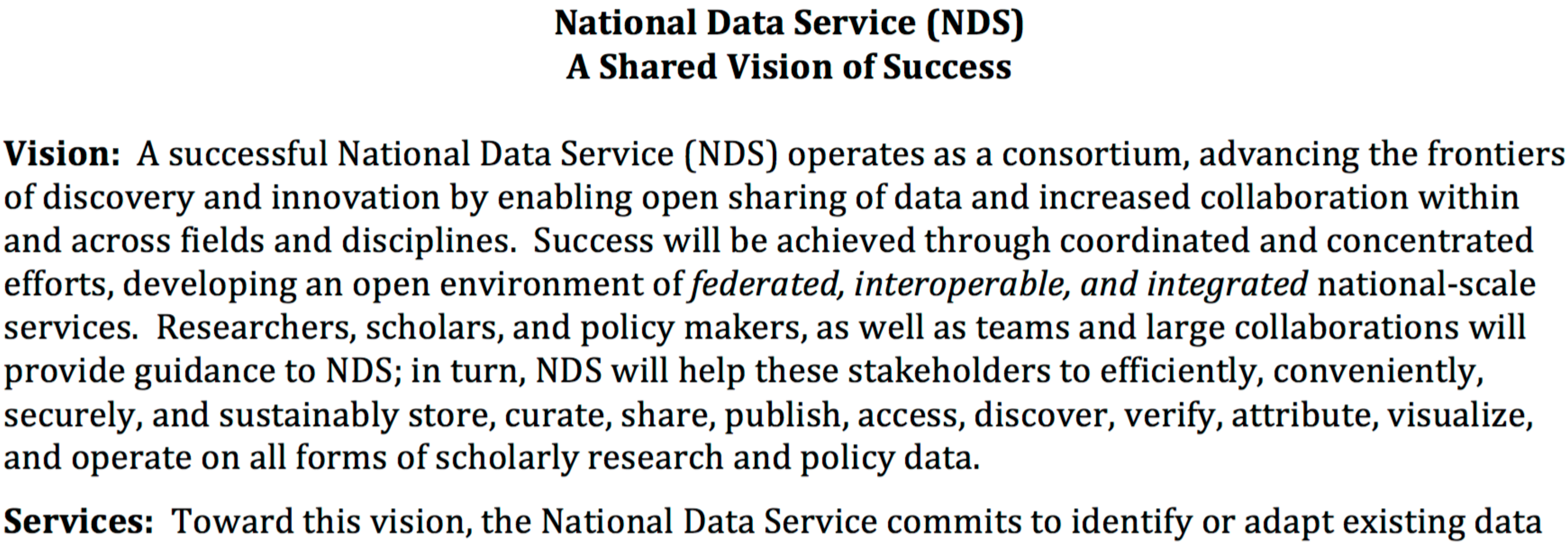
The National DATA SERVICE

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**!

Researchers, educators, students communicate by sharing data…this is central to enabling everything above! Services needed to make it work!

NDS Vision



National Data Service (NDS)
A Shared Vision of Success

Vision: A successful National Data Service (NDS) operates as a consortium, advancing the frontiers of discovery and innovation by enabling open sharing of data and increased collaboration within and across fields and disciplines. Success will be achieved through coordinated and concentrated efforts, developing an open environment of *federated, interoperable, and integrated* national-scale services. Researchers, scholars, and policy makers, as well as teams and large collaborations will provide guidance to NDS; in turn, NDS will help these stakeholders to efficiently, conveniently, securely, and sustainably store, curate, share, publish, access, discover, verify, attribute, visualize, and operate on all forms of scholarly research and policy data.

Services: Toward this vision, the National Data Service commits to identify or adapt existing data

  - NDS is a member of RDA and very active, e.g., workshops
- Extend/integrate efforts of individual projects
  - e.g., DataONE, SEAD, ICPSR, Dryad, publishers, etc

NDS Lab and NDS Share

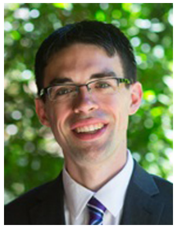
- NDS Labs
  - Target: friendly developers
  - A community support environment for d coordinating, deploying prototype s
  - Spinning disk, storage, virtual m and hosting services
  - Working with RDA to test/deplo
- NDS Share
  - Target: friendly scientists
  - Experimental platform for shari
    - Enable anyone to create da ctions, 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

RDA-NDS agreement to use NDS to test, deploy products of RDA working groups!

National Steering Committee

National Executive Committee

Technical Advisory Committee



Governance

The National
DATA SERVICE

Commitments

Resource Commitments

- NCSA, TACC, Globus, PSC, SDSC, Indiana, Notre all contributi
  - Fund
    - ND
    - Techn
    - Several resear just hired
- Resources
  - Federated OpenStack environments
    - Hundreds of cores, PBs of storage

NDS Director, other positions available!

First Funded Project: Materials Data Facility

- Ian Foster to describe current status
- We are very interested in developing this and getting you to help drive it
  - What services would help you?
  - What data sets would be of value?
  - How can we use the Chicago area collaborations to set the example for the nation?