Distributed Big Data and Analytics (DBDA)

Internet2 CINO Initiative Working Group Kickoff Meeting

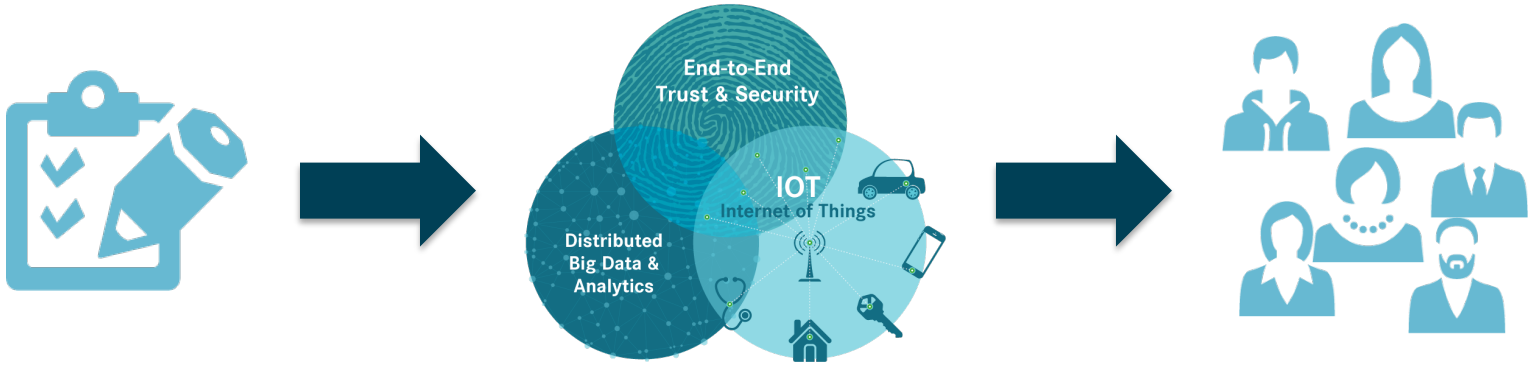
26 June 2015

Chairs
Alex Feltus, Clemson
Sam Gustman, USC
Marc Hoit, NC State

INTERNET2 POWERED BY COMMUNITY

Collaborative Innovation Program

Three new collaborative
Innovation Working Groups

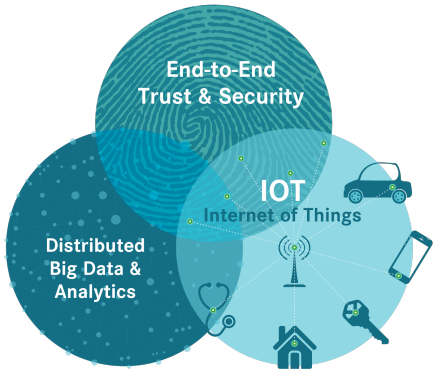


INTERNET 2

POWERED BY COMMUNITY

Established Three Collaborative Innovation Working Groups

- Each will formulate specific recommendations for a particular innovation initiative, including scope, value, milestones
  - Clarify value to members
  - Ensure the innovation is economically viable
  - Develop a scalable model to positively impact a significant segment of Internet2 membership
- Led by member representatives
  - May leverage member programs and facilities
  - Participation encouraged for all Internet2 members
- Working groups will operate within a broader Internet2 Collaborative Innovation Community with representatives from member organizations



Big Data = Big Issues!

Distributed Big Data & Analytics

UNIVERSITY RESEARCHERS INCREASINGLY NEED TO COLLABORATE *AND SHARE DATA AROUND THE WORLD.*

SCIENTISTS FROM **34 COUNTRIES** ARE USING DATA GENERATED BY THE LARGE HADRON COLLIDER TO STUDY THE SMALLEST KNOWN PARTICLES.

THE 1,000 GENOMES PROJECT COMPRISES OVER 2,500 WORLDWIDE DNA SAMPLES WITH A TOTAL DATASET SIZE OF **200 TERABYTES**

THE LARGE HADRON COLLIDER PRODUCES APPROXIMATELY **15 PETABYTES OF DATA ANNUALLY** OR ENOUGH TO FILL MORE THAN 1.7 MILLION DUEL-LAYER DVDs.

THE UNIVERSITY OF MARYLAND INSTITUTE FOR GENOME SCIENCE'S GENOMIC FILE SIZES ARE INCREASING FROM AN AVERAGE OF **5 GIGABYTES** TO **20 GIGABYTES.**[8]

TIME TO TRANSFER A GENOMIC FILE ACROSS THE GLOBE:

PUBLIC INTERNET = 26 HOURS

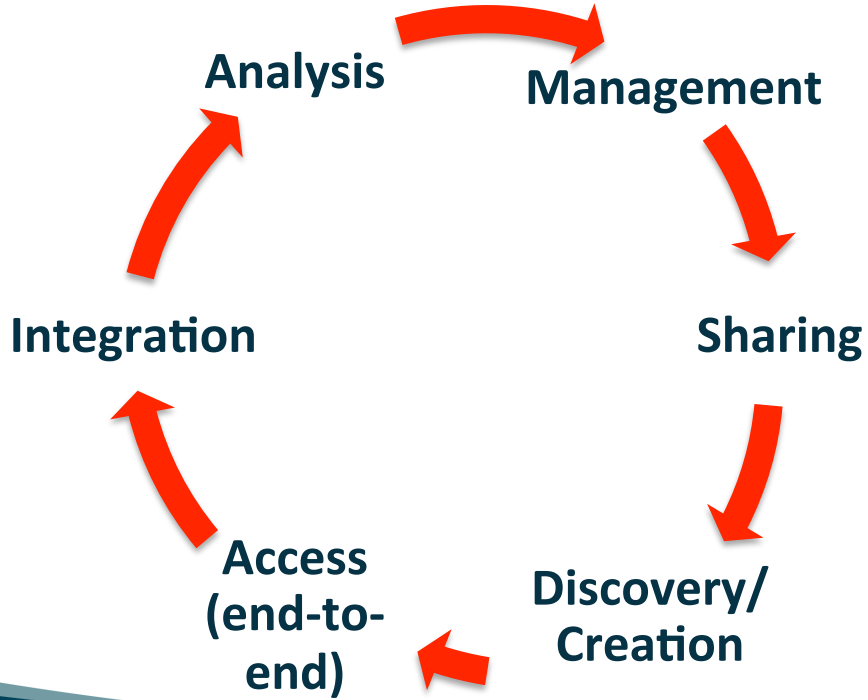
AIRPLANE = 17 HOURS

100 GBE INTERNET2 NETWORK= **30 SECONDS**

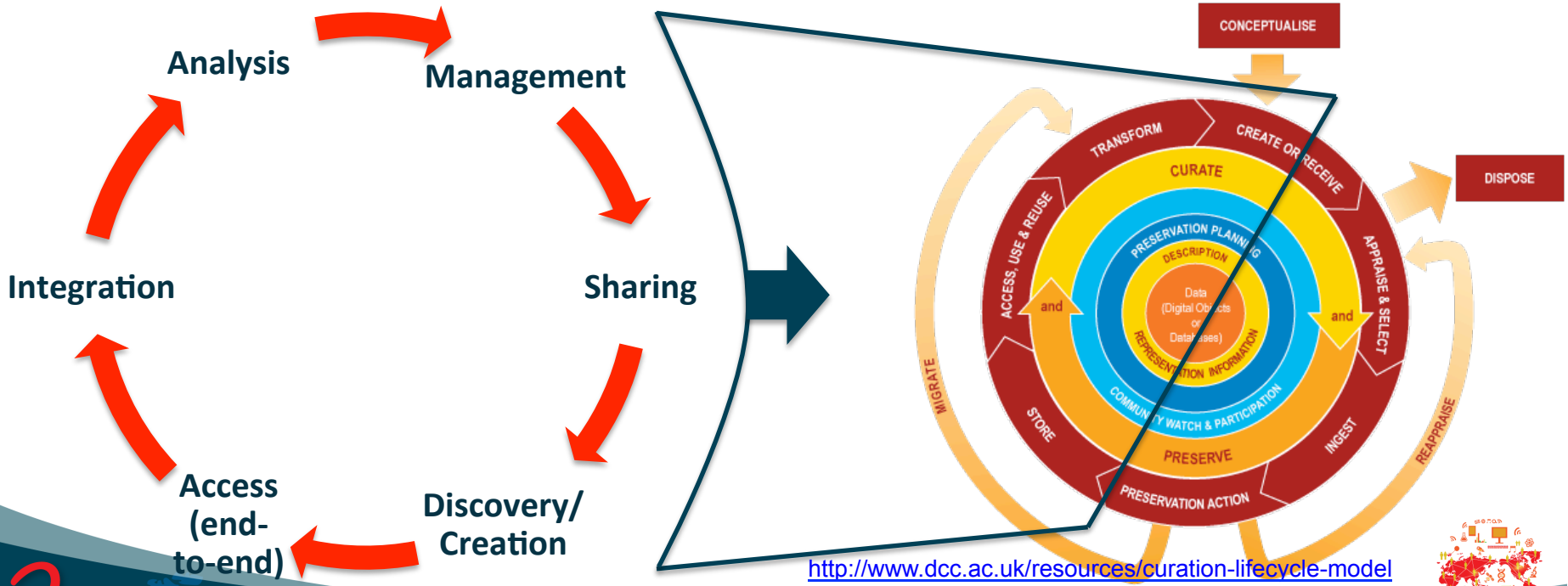
BEFORE UPGRADING ITS NETWORK AND STORAGE, UCLA'S LABORATORY OF NEURO IMAGING (LONI) **STRUGGLED TO ACCESS AND SHARE ITS REPOSITORY OF NEUROLOGICAL IMAGES** FOR MORE THAN 16,000 SUBJECTS – SOME IMAGES WERE AS LARGE AS **200 GIGABYTES**

THE 6,000-MEMBER NATIONAL NETWORK OF LIBRARIES OF MEDICINE DELIVERS TRILLIONS OF BYTES OF **DATA TO MILLIONS OF USERS EVERY DAY.**

A starting point: six focus areas



The six focus areas map to the DCC data life cycle model



http://www.dcc.ac.uk/resources/curation-lifecycle-model
DCC = Digital Curation Center

Challenges and opportunities

- Each of the focus areas has associated challenges for our members in:
  - Research
  - Research collaborations with campus, national and international scope
  - Teaching and learning
  - Enterprise operations

- These challenges, in turn, are opportunities for innovation that have broad impact across Internet2's membership
  - Shared risk and shared solutions tuned to the needs of higher education
  - Public-private and university-industry partnerships
  - Solutions that promote international collaborations in research and education

Management of data

- Challenges
  - Inclusion of librarians and those responsible for curation, archiving and preservation
  - Need cyber-infrastructure experts AND storage experts AND research scientists involved in system design
  - Need uniform shared processes for digital preservation of data

- Opportunities
  - Foster collaboration among Internet2 members and partners to identify and propagate best practices in data management
  - Propagate data management models at the campus-level that also support collaboration, discovery and sharing with the larger R&E community

Sharing of data

- Challenges
  - Storage bottlenecks can be a challenge, corollary to network bottlenecks
  - Distributed big data repositories and how to get the data through Internet2
  - Network trouble shooting is difficult
  - Need to serve the "missing middle" of researchers and scholars

- Opportunities
  - Identify and propagate campus-level infrastructure, policy and processes to promote data sharing
  - Identify, evaluate and communicate network, storage and metadata architectures that enable sharing

Discovery and creation of data

- Challenges
  - No uniformity in approach to engaging researchers early in a project to understand data storage and sharing requirements
  - “Long tail” researchers have difficulty finding, understanding and reusing data sets
  - Collaboration in building shared repositories regionally, within and across disciplines is difficult

- Opportunities
  - Compile and share best practices for implementing a life cycle approach to data at the campus level
  - Team with existing projects such as the RDA and NDS to implement services that meet members’ needs for discovery and re-use of data

Note: RDA = Research Data Alliance, NDS = National Data Service

End to end access of data

- Challenges
  - Hard to use the network – complicated sets of IT issues – need a cookbook for researchers – how to use the network, get access to services, use services
  - Researchers don't know what Internet2 is or how to access and leverage the network, including for their big data needs
  - End to End 100Gb connectivity internationally, or even domestically, is difficult to realize today
- Opportunities
  - Expand outreach of training programs in advanced network use to include researchers (e.g., perfSonar, Science DMZ, software defined networking)
  - Partner with NSF and NIH to develop network services related to accessing and using data sets (e.g., DANCES project, CloudLab, Chameleon Cloud)

Integration of data

- Challenges
  - Integration of existing data into new projects is difficult for many reasons: discovery, access, understanding data and metadata, ETL processes, …
  - Researchers generally do not have resources and expertise on their campuses to help with this
- Opportunities
  - Partner with the Data Carpentry group and Internet2 member representatives to identify areas of need in data integration and develop new or expand existing training programs
  - Work with members to propagate information about data science degree programs and to increase their impact in the Internet2 community
  - Partner with, e.g., the National Data Services and iPlant to develop data integration services of use to the Internet2 community

Analysis of Data

- Challenges
  - “impedance mismatch” between distributed data, analytics and network speeds; current research data sharing is "excel" scale, vs. needed Exascale
  - Workflow tools not generally understood or used in research
  - Tools for large-scale digital humanities are not represented in existing thinking about research cyberinfrasrtucture

- Opportunities
  - Team with XSEDE to evolve XSEDENet based on requirements from the Internet2 community of researchers and scholars
  - Team with the Open Science Grid to offer high throughput computing services
  - Work with the research workflow community to develop and propagate best practices to the Internet2 community

Outcomes and Deliverables

The WG will develop and communicate to Internet2 members:

- A list of projects including their owners, expectations, and timeline.
- Develop and publish white papers and recommendations for standards related to the DBDA use cases, research and implementations.
- Become a valuable resource to enable the development of DBDA research in our member organizations, including connections between experts and resources to support DBDA research
- Engage in formal and informal collaborations with Internet2 members, such as research, pilot implementations, test-beds, standards development, policy development, etc.
- Engage in formal and informal collaborations with outside organizations that are engaged with DBDA activities that may usefully complement the activities of Internet2 members.

Example Impact Areas

- Researcher/scholar and research IT staff “starter kits” for using advanced networking services in data-intensive work
- Identify the characteristics of a model reproducible architecture for campus research data repositories (from organization to infrastructure)
- Identify and promote the development of community-driven shared services for managing and analyzing large data sets (e.g., partner with the National Data Service, www.nationaldataservice.org, and Open Science Grid, www.opensciencegrid.org)
- Develop a “Big Data Days” campus awareness program and events, analogous to ongoing “Cyberinfrastructure Days” events (c.f. Google “cyberinfrastructure days”)

Timeline - Schedule

- **May - Initial Actions**
  - Group organization
- **June - Group Discussions and Drafts**
  - Identify initial challenges to be tackled, target opportunities and deliverables
  - Develop schedule, milestones, and deliverables
  - Establish monthly calls
- **July**
  - Establish any sub-groups for initial focus areas
  - Identify meetings and conferences to have presence
  - Engage with other Internet 2 Collaborative Innovation Community working groups on progress and overlap
- **August**
  - Plan for face-to-face meeting at TechEx
  - Draft presentations for TechEx
- **September**
  - Finalize presentations for TechEx
  - Plan for EDUCAUSE presence
- **October**
  - Present initial recommendations at TechEx
  - Post-TechEx meeting actions and revisions
  - Collaborations with other organizations
- **November/December**
  - Potentially publish white paper / report on DBDA Point-of-View, definition, best practices, needed practices, key technologies

Major Milestones

- June 11 - Initial co-leaders conference call
- June 26 - First conference call for the WG
- July 30 - Synopsis of sample member projects for DBDA
- August - Agree focus areas supporting member projects and needs for DBDA working group
- Oct 4-7 - Presentation and workshops at Internet2 TechEx, Cleveland, Ohio

Distributed Big Data & Analytics Project Template

- **Project/Research Title:**
- *Industry Sector:*
- *Science Sub-domain*:
- *Short Description of Project & Relation to Big Data*:
- *Potential Industrial Partners:*
- *Other Faculty/Organizations/Researchers Involved:*
- *Best Contact:*
- *Big Data Attributes (Image, text, geospatial, real-time, near-real time, sensor, distributed, other):*
- *Aggregate Data Size: Now _____, 2016 _____, 2017 _____, 2020 _____*

SAMPLE REAL:
Distributed Big Data & Analytics Project Template

- **Project/Research Title: Data Analytics of Campus-Scale Power System**
- *Industry Sector:* Electric Power Utility
- *Science Sub-domain*: Electrical Engineering
- *Short Description & Relation to Big Data:* The local electric power grid will be heavily instrumented on a campus containing a mix of residential sites, office spaces, industrial-scale electromechanical systems, and distributed energy sources. The instruments will be networked to a server that provides data for use in analytics focused on electric energy consumption, electric-service reliability, power quality, and local grid planning and design. The analytics will support research in local-grid technologies, distributed control of the electric grid, and power electronics, etc.
- *Potential Industrial Partners:* Duke Energy (Clemson's electric service provider), other electric utilities, power-industry instrumentation and electronics manufacturers, power-system monitoring and control vendors
- *Other Faculty Involved* All power faculty at Clemson, power research staff at Clemson's CURI site in Charleston, SoC faculty working in data analytics
- *Best Contact:* Dan xxxx, ECE Dept. Chair, <xxx@clemson.edu>
- *Big Data Attributes:* sensor, near-realtime, distributed, geospatial
- *Aggregate Data Size: Now 2 Tbyte, 2016: 4 Tbyte, 2017: 16 Tbyte, 2020: 1 Pbyte*

Communications Plan

- **Email:** The DBDA WG uses a mailing list for communication "cino-dbda@internet2.edu" and an archive of previous discussions is available on the web though the wiki.
- **Web:** The WG maintains a Wiki webpage hosted by Internet2 and available for the public at:https://spaces.internet2.edu/display/CWG/Distributed+Big+Data+and+Analytics
- **Conference Calls:** The DBDA Working Group meets over the phone on a monthly basis.
  - The DBDA WG will be invited to bi-monthly meetings of the Internet2 Collaborative Innovation Community for sharing of best practices, learnings and content with the other CIC working groups - End to End Trust & Security and Internet of Things.
- **Face-to-face meetings:** Occasionally the WG organizes face to face meetings. This could include meetings at the Internet2 TechEx and Internet2 Global Summit.
- **General public presentations:** The WG aims to establish presence in relevant conferences and meetings and present its progress to the public.

Next Steps

Call to Action for DBDA Working Group

Send a note to the DBDA Working Group leadership team **dbda-chairs@internet2.edu** with your interest areas related to this working group. e.g.,

- Use cases you are working on or are interested in
- Completed DBDA Project Template for your project(s)
- Data repositories of potential broad interest that may be generally accessible
- Reference architectures
- Research expertise on your campus / in your organization

Monthly team meeting – next one late July / Early August

Joint Collaborative Innovation Community call with all 3 working groups bi-monthly

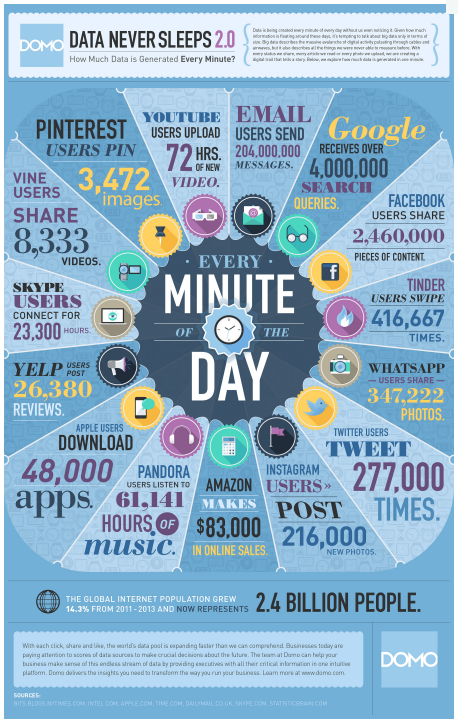
Questions, Discussion

Back-up

Potential sources of funding and information

- **Selected sources of research funding**
  - NSF - Critical Techniques and Technologies for Advancing Foundations and Applications of Big Data Science & Engineering (BIGDATA) - http://www.nsf.gov/funding/pgm_summ.jsp?pims_id=504767
  - DARPA – The Information Innovation Office (I2O) has several big data research programs with a significant budget increase for FY 2016 to $160M+
  - NSF Big Data Innovation hub (bdhub.info) as a collaborative funding mechanism between academia, NGOs, and industry.
- **Selected presentations given at the Global Summit**
  - http://meetings.internet2.edu/2015-global-summit/program
  - Internet2 Advanced Network Services Today
  - perfSONAR: Meeting the Community's Needs
  - Big Data Health Initiatives: PCORI, NIH Data Commons – Collaboration, Patient Networks, Data Clouds
  - Delivering Order-of-Magnitude Improvements to Research Data Throughput

The Growth of Big Data and Associated Analytics is a Global Phenomenon



- A billion hours ago, modern homo sapiens emerged.
- A billion minutes ago, Christianity began.
- A billion seconds ago, the IBM PC was released.
- A billion Google searches ago … was this morning.

Hal Varian, Google Chief Economist
"Beyond Big Data," 2013

About the Distributed Big Data and Analytics Working Group

Organize and apply Internet2 community capabilities to help members find solutions to problems in managing, finding, analyzing and visualizing data used in research and education

A few areas for innovation – many other possibilities

- Address challenges and opportunities big data management and analytics in distributed multi-institutional operations requiring complex compliance, security, identity, and privacy requirements
- Teaming within the community to leverage NSF, NIH and other federal funding opportunities related to big data
- Applications of software defined networking to big data operations

Some related initiatives

- Many individual member institution programs
- NSF programs: BIGDATA, CC-IIE/NIE/DNI, IIS, NIH BD2K, …
- EDUCAUSE ECAR Cyberinfrastructure Working Group
- National Academy Task Force on Cyberinfrastructure for 2020, …

Example application domains

- basic research
- digital humanities scholarship
- life sciences
- health care
- economics
- logistics
- network analysis
- security and privacy

Many other important areas

INTERNET2 POWERED BY COMMUNITY

Mission and Scope

The WG seeks to:

- Find ways to help enable the Internet2 community to address complex issues in big data management and analytics in both campus-level and geographically distributed multi-institutional operations
- Improve community awareness and understanding of DBDA and implications.
- Identify research gaps and needs to be addressed to be a catalyst and enabler for DBDA activities in our membership
- Foster collaboration among Internet2 members and partners to enable a global DBDA research and education plan
- Identify and support the investigation of DBDA use cases that may benefit Internet2 members (e.g., campus operations, health care, teaching and learning, etc.).
- Identify needs for and assist Internet2 in developing DBDA services
- Scope of participants includes all interested Internet2 members (universities, affiliates, NRENs, state/regional networks), industry partners, international partners

Vision and Value

- The DBDA WG at Internet2 focuses on the strengths of Internet2:
  - Its members
  - Advanced networking services for research and education
  - Trust, identity, security and cloud services

- The DBDA WG will leverage Internet2 resources and provide emphasis on the R&E community's needs and applications

Participation and Operations / Approach

- WG participation is open to all Internet2 members.
- The WG envisions a range of participants from individuals learning about DBDA to highly structured collaborations that may include teams of leading researchers and practitioners from multiple members.
- The DBDA WG will conduct open meetings for Internet2 members with open source results.

Possible Opportunities for Internet2 to Serve Members in DBDA Research and Operations

- Use of the very high performance Internet2 network and specialized services to move data and to link large data repositories and end users
- Stimulate scalable multi-institutional collaboration – both domestic and international
- Provide relevant services to the community through NET+ and other I2 Divisions
- Help communicate member activities through conferences
  - E.g., presentations and sessions at the Global Summit and Tech Exchange
  - Possible workshop series – include research topics and educational seminars, ensure sufficient depth in sharing critical information
  - Outreach through other community meetings (EDUCAUSE, SCXY, etc.)

Some Challenges for the Internet2 Community in Addressing DBDA -- 1

- Successful use of the Internet2 network requires a high degree of expertise – the "Wizard Gap" is still large for uses at high data rates and large volumes – only a few groups succeed with high end-end performance in making full use of the Internet2 100 Gbps backbone
  - From a campus perspective, data must traverse many different network administrative domains (a minimum of five for typical end-to-end use: campus, regional network, Internet2, a second regional, and a second campus)
  - Many groups working in DBDA research, education, and operations need more training and/or support in using Internet2 capabilities and services
  - Training and support include not only networking but also storage architectures and end-user systems

- Distributed collaboration, particularly internationally, requires inter-domain networking with diverse networking and end-user technologies and capabilities

- Many concepts and tools are available to improve infrastructure performance, but learning to apply them requires dedication and focus on "global" rather than campus network issues
  - perfSonar – wide area network performance monitoring and analysis
  - Science DMZ – removing institutional barriers to moving large amounts of data
  - Software Defined Networking – tools to create dedicated, purpose built networks for collaborations

Some Challenges for the Internet2 Community in Addressing DBDA -- 2

- Trade-offs between moving data to computing or moving computing to data
  - Trade-offs affected by ability to use the Internet2 network effectively and efficiently
- Uneven distribution of expertise in data curation, archiving, and preservation
- Lack of applicable standards, and exemplars, and lengthy development process
  - Vendor-driven rather than community-driven solutions
- The Internet2 community needs broader exposure to researchers
  - Many could find real value with more awareness of Internet2 and ability to access services

Strategies to address DBDA challenges

- Improved communications with broader segment of the R&E community
  - Help ensure researchers and educators are aware of and can access Internet2 services
- Establish communities of practice based on selected use cases
  - Use cases could include health care, Smart Grid, climate and environment, agriculture, security operations, …
  - Other types of use cases include access and sharing protocols, APIs for big data and big analytics operations, repository and hub management, …
  - Consider discussions with current members doing related work – e.g., teams who presented at the Global Summit
    - Exploration of common data sharing protocols and repository architectures
    - Campus-level partnership models including CIOs, CROs, and heads of libraries
- Facilitate research proposals to government and industry sponsors
  - NSF, NIH, DARPA, selected Internet2 corporate members