engineering laboratory

Global Research Data Infrastructure: Path Forward for Progress

NIST
National Institute of Standards and Technology
U.S. Department of Commerce

Dr. Chris Greer

Senior Executive for Cyber Physical Systems

NIST's Mission

- To promote U.S. innovation and industrial competitiveness by advancing **measurement science, standards**, and technology in ways that enhance economic security and improve our quality of life

NIST: Basic Stats and Facts

- Major assets
  - ~ 3,000 employees
  - ~ 2,800 associates and facilities users
  - ~ 1,300 field staff in partner organizations
  - Two main locations: Gaithersburg, Md., and Boulder, Colo.
  - Nobel Prize Winners: 1997, 2001, 2005, 2007, 2013

Internet of Things

If we had computers that knew everything there was to know about things—using data they gathered without any help from us—we would be able to track and count everything, and greatly reduce waste, loss and cost.

—Kevin Ashton, That 'Internet of Things' Thing, RFID Journal, July 22, 2009

Internet of Things

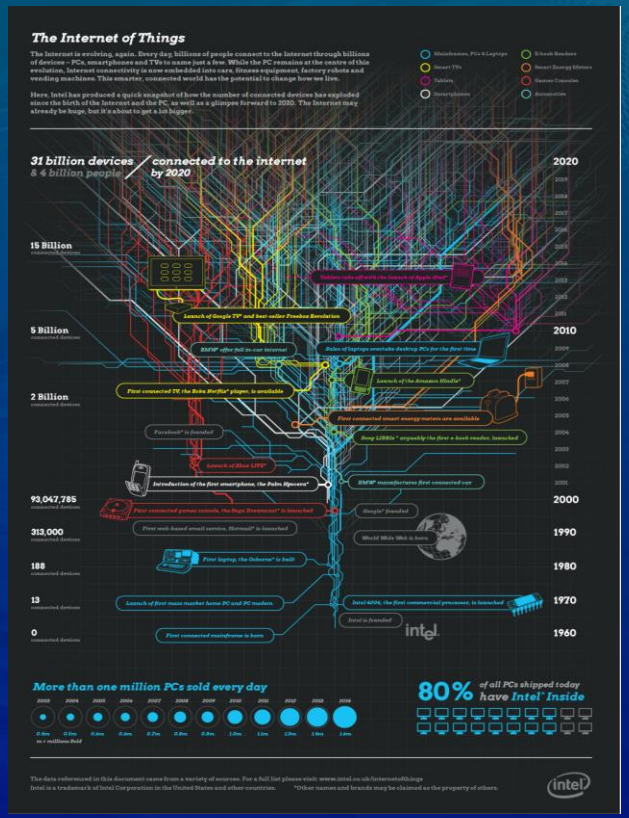
What are the defining characteristics of the “Internet of Things?”

- Scale
- Capability
- Reach

Internet of Things - Scale

Devices connected to the Web:

- 1970 = 13
- 1980 = 188
- 1990 = 313,000
- 2000 = 93,000,000
- 2010 = 5,000,000,000
- 2020 = 31,000,000,000



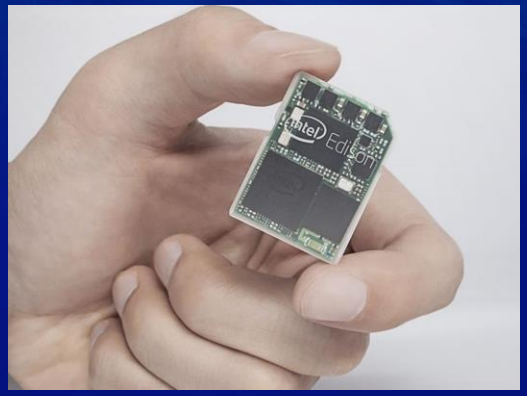
Source: Intel

Internet of Things - Capability

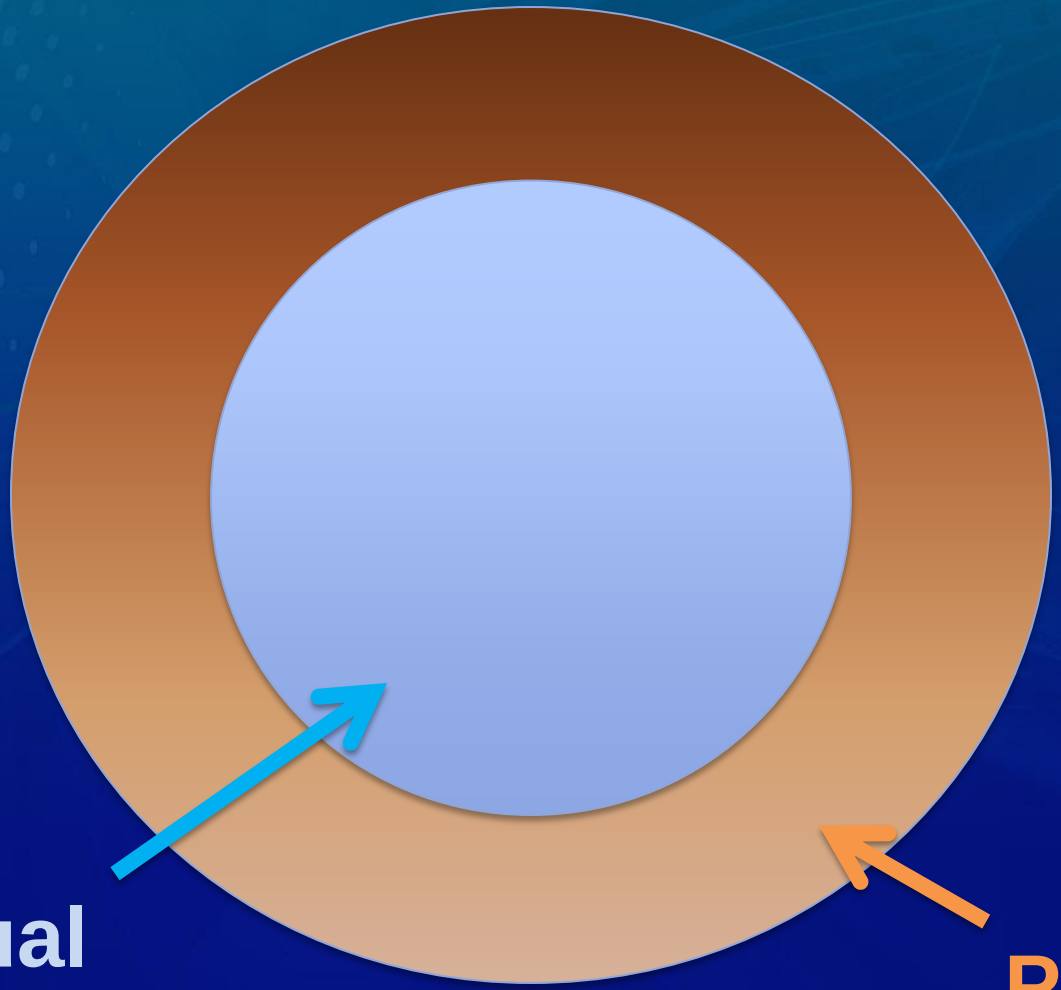
Intel Edison:

"It's a full Pentium-class PC in the form factor of an SD card,"

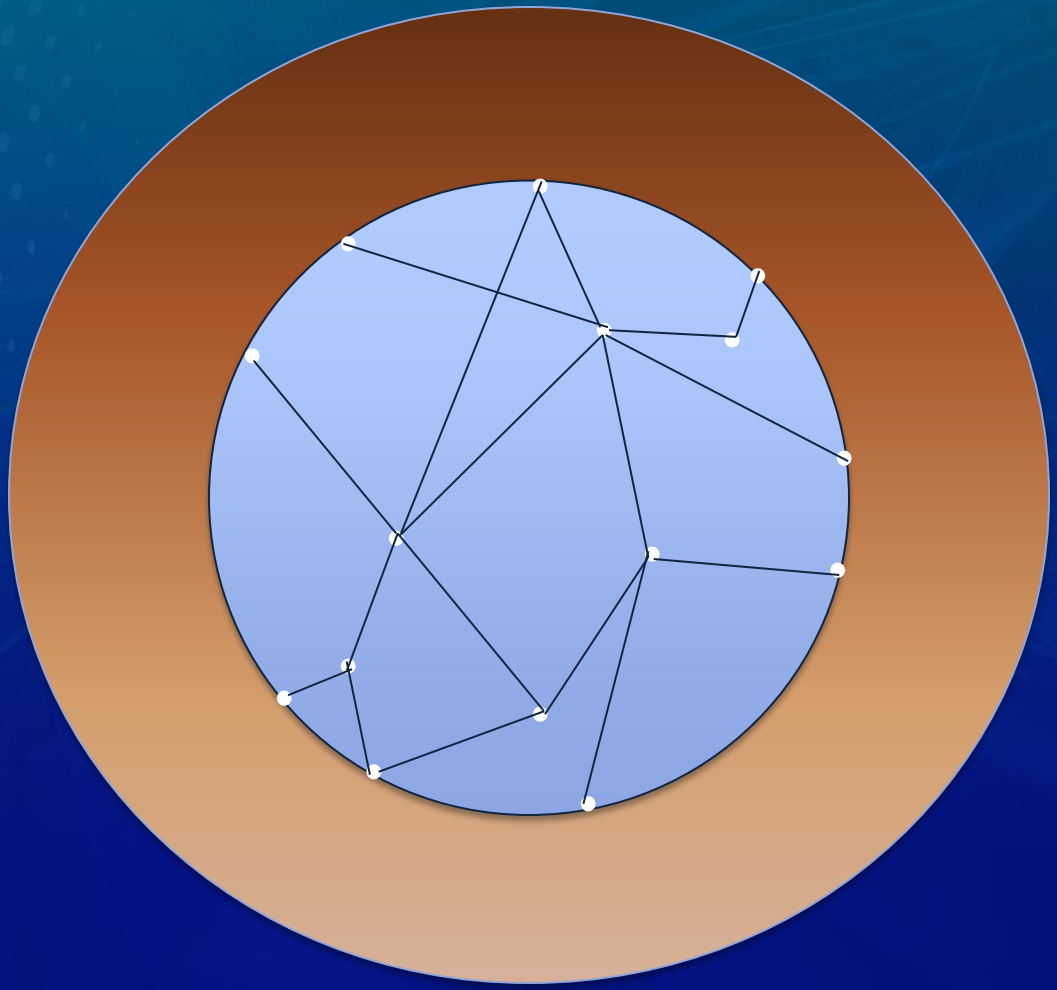
Intel CEO Brian Krzanich



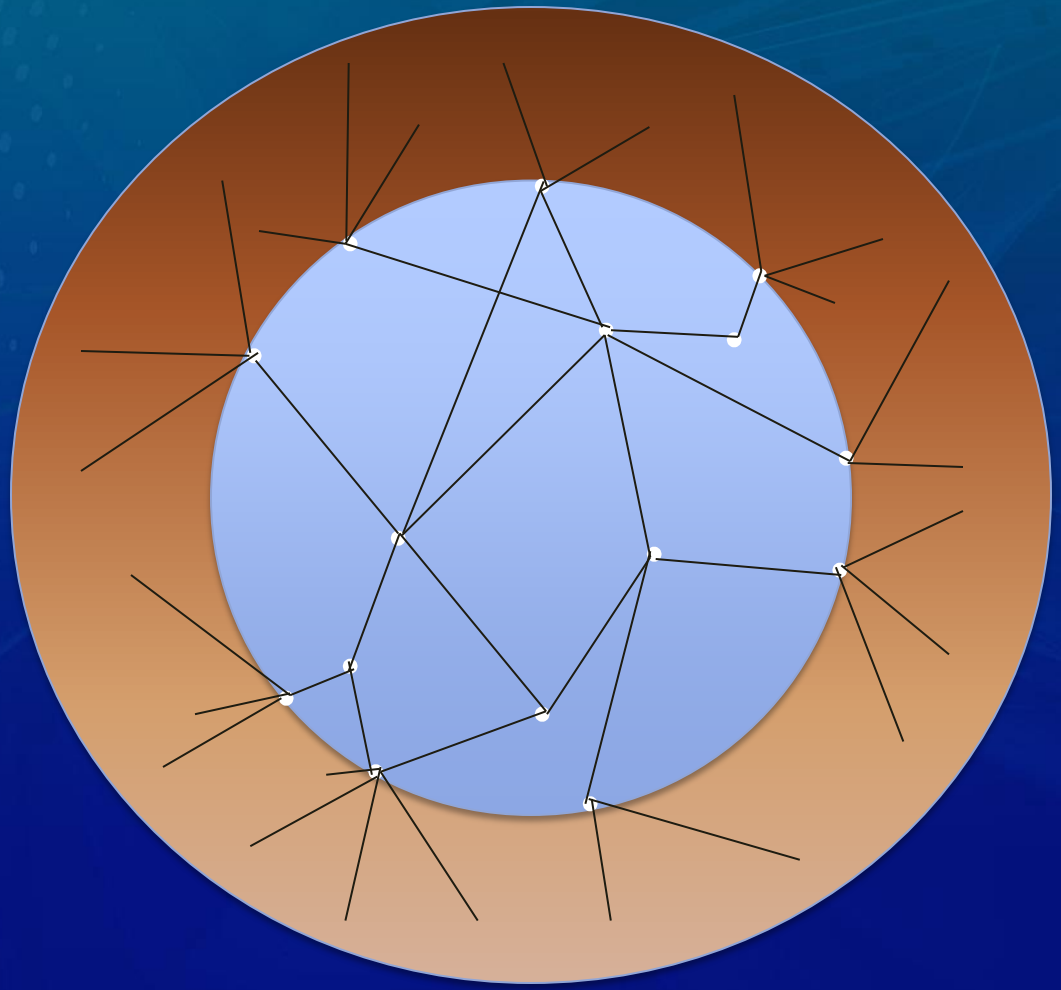
Internet of Things – Reach



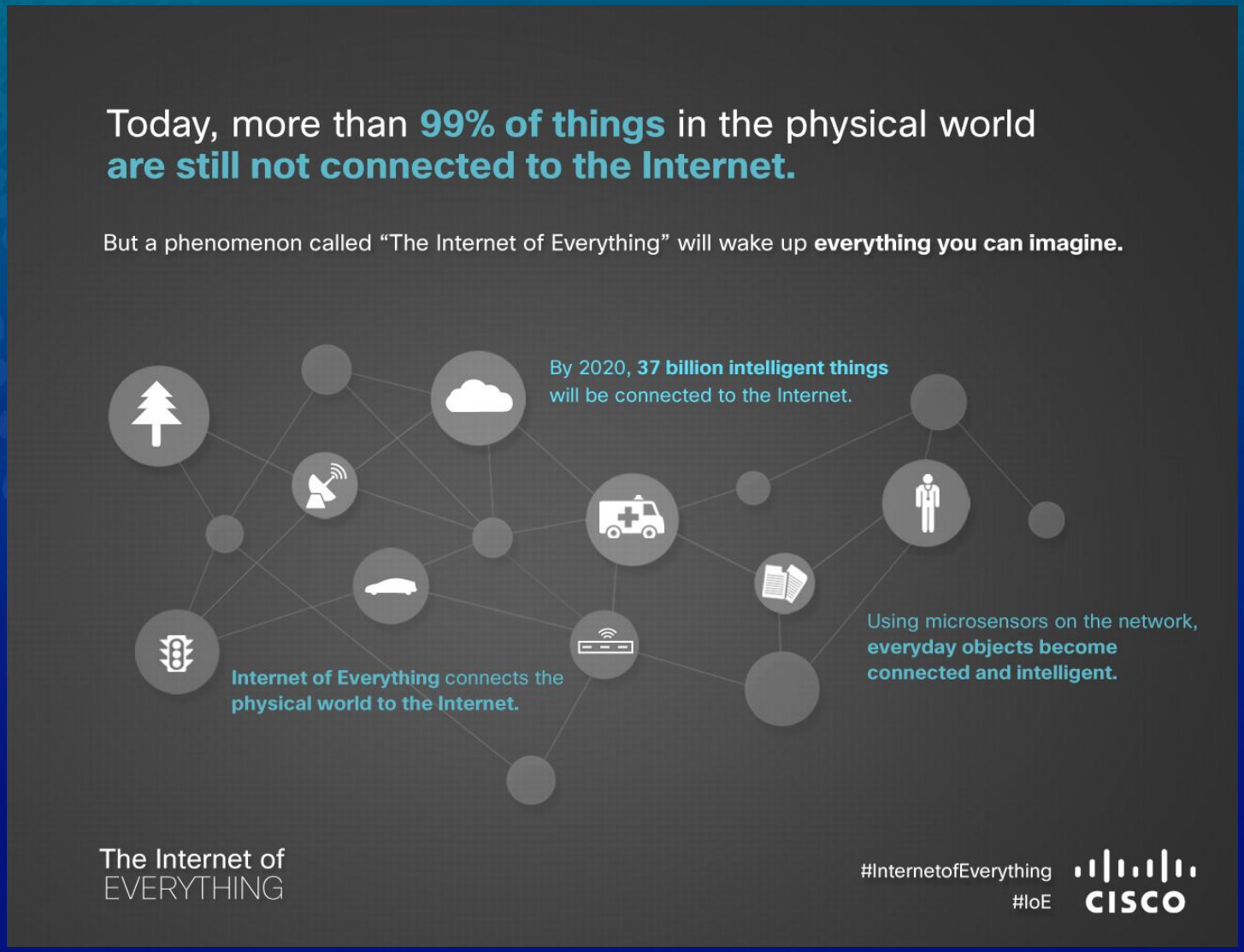
Internet of Things - Reach



Internet of Things - Reach



Cisco Internet of Everything



Big Data - Qualitative

Things You Don't Know

Questions You're Asking

Data Acquisition	BIG DATA
Conventional Data Analytics	Data-enabled Exploration

Questions You Haven't Thought Of

Things You Know

Credit: Jason Kolb, Applied Data Labs; Modified from the original at: www.applieddatalabs.com/content/new-reality-business-intelligence-and-big-data

Big Data - Quantitative

- Dimensions:
  - Volume
  - Velocity
  - Variety

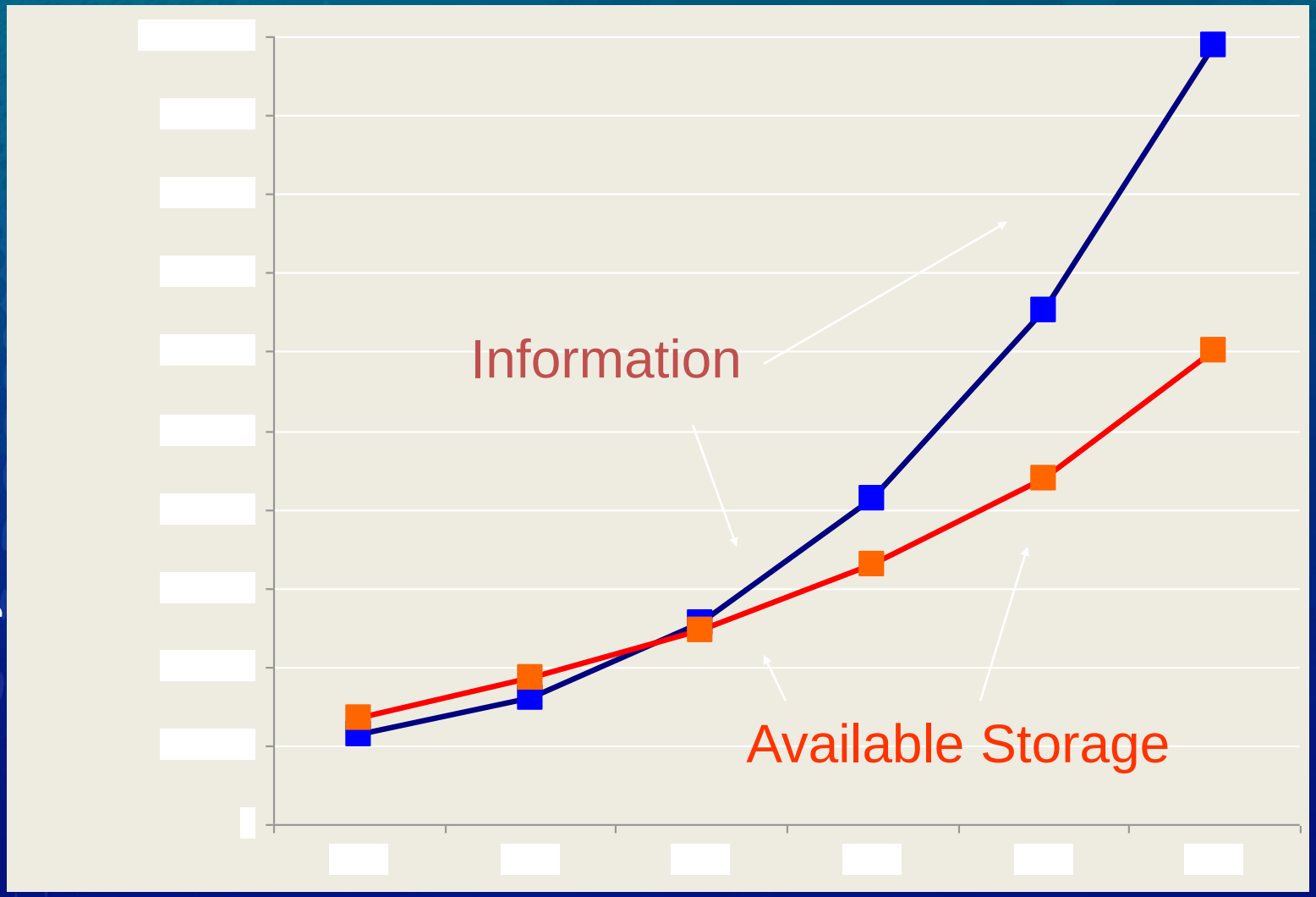
Big Data - Volume

IDC April 2014: The Digital Universe of Opportunities

- From 2013 to 2020, the digital universe will grow by a factor of 10 – from 4.4 trillion gigabytes to 44 trillion. It more than doubles every two years.
- In 2014, the digital universe will equal 1.7 megabytes a minute for every person on Earth.
- Data from embedded systems will grow from 2% of the digital universe in 2013 to 10% in 2020.
- In 2013, the available storage capacity could hold just 33% of the digital universe. By 2020, it will be able to store less than 15%.

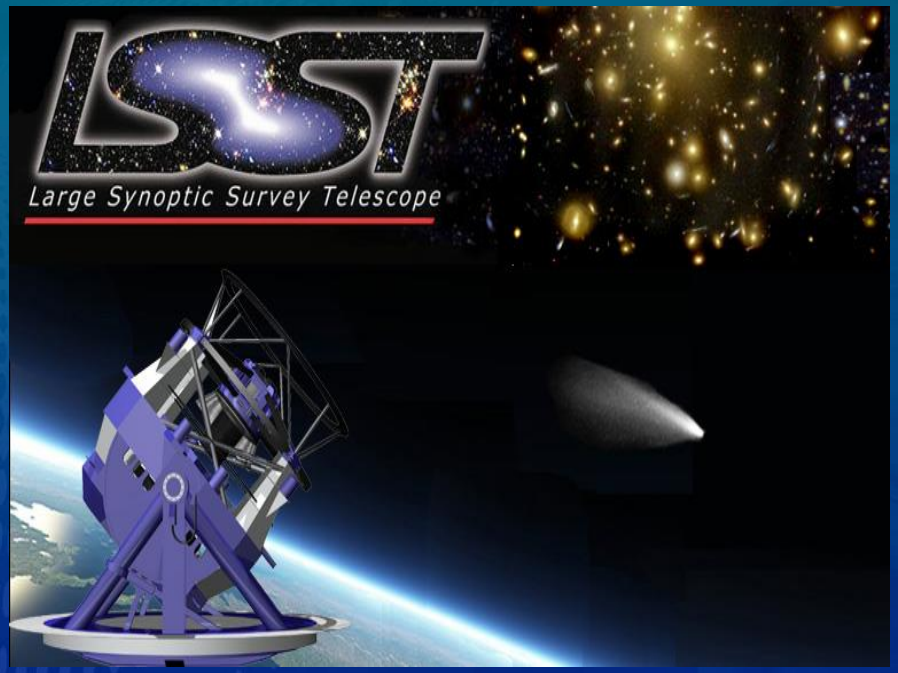
Source: IDC Corporation, http://idcdocserv.com/1678, sponsored by EMC

Big Data - Volume



Source: John Gantz, IDC Corporation, *The Expanding Digital Universe*

Big Data - Velocity



LSST:

"Suspended between its vast mirrors will be a three billion-pixel sensor array, which on a clear winter night will produce 30 terabytes of data. In less than a week this remarkable telescope will map the whole night sky …. And then the next week it will do the same again … building up a database of billions of objects and millions of billions of bytes."

Nature 440:383

- **Sloan Digital Sky Survey**
  - 140 Terabytes, year 2000 to present
- **LSST – Large Synoptic Survey Telescope**
  - Expect 140 Terabytes every 5 days
- **Square Kilometer Array**
  - Expect 140 Terabytes every 3 sec

Big Data - Variety

Combining Structured and Unstructured Data

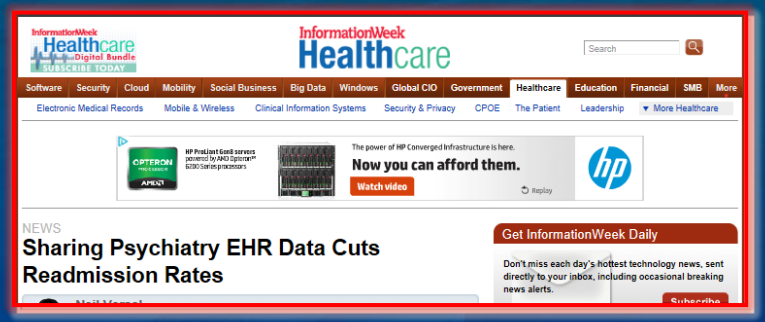
Source: IDC's Digital Universe Study, sponsored by EMC, June 2011

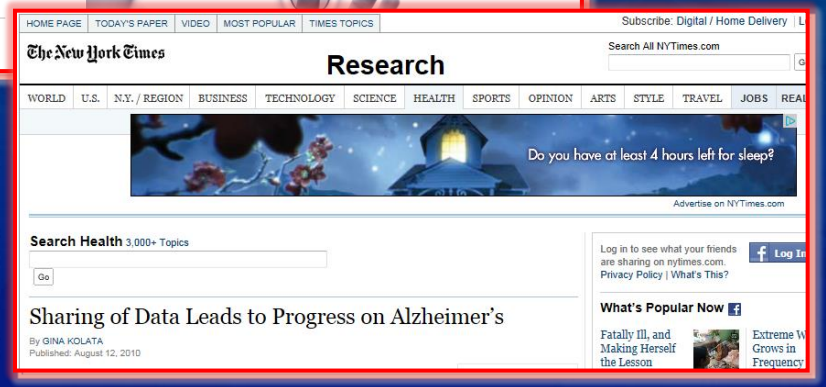
The time is right for progress:

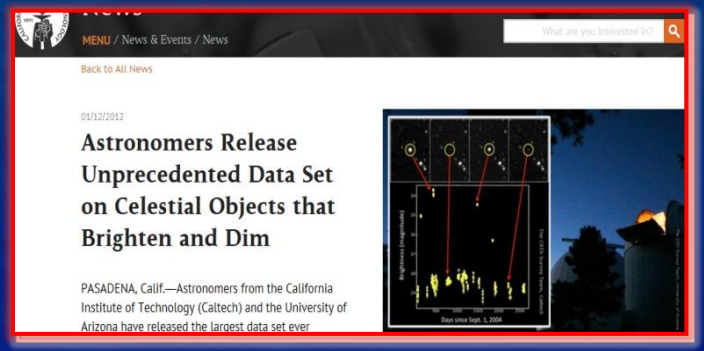
- The open exchange of research data has intrinsic value
- 21st Century science is global and digital
- Adequate technical capabilities are available today
- A global infrastructure can be operated on the basis of voluntary cooperation and consensus

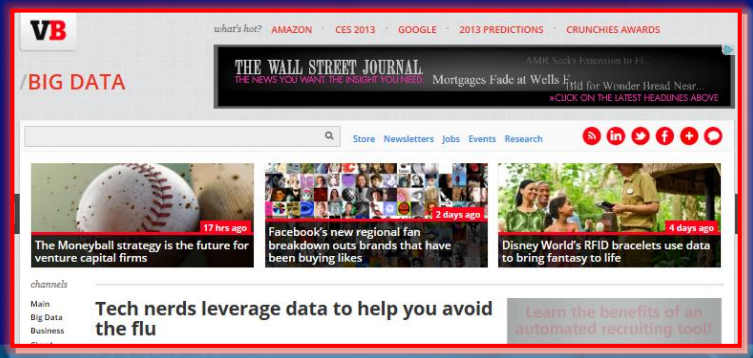
It's Not Just Your Data, It's Other People's Data

Data Sharing Fundamental to Data-Driven Innovation





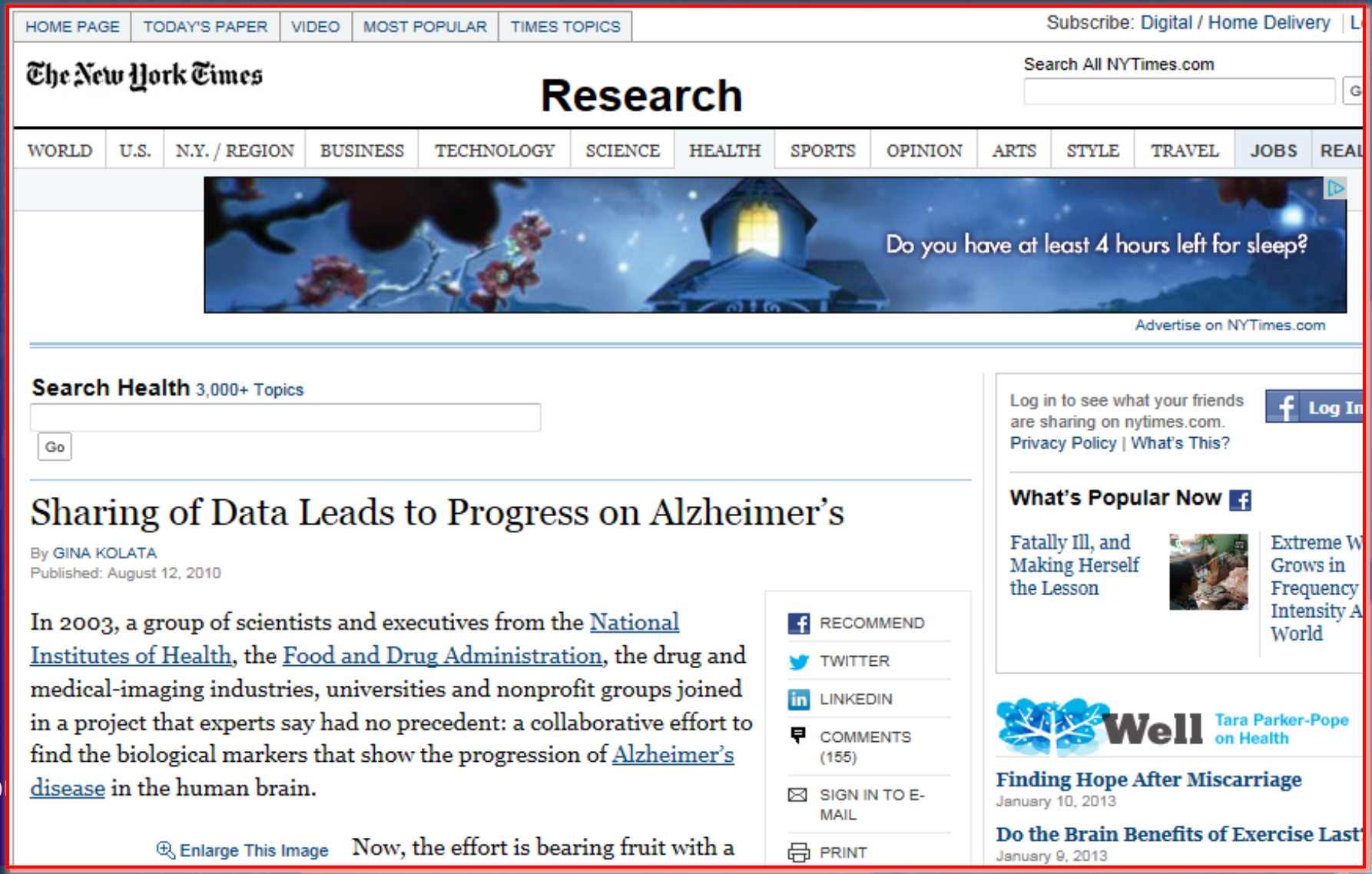




Source: Fran Berman
RPI

It's Not Just Your Data, It's Other People's Data

Data Sharing Fundamental to Data-Driven Innovation



So

It's Not Just Your Data, It's Other People's Data

Data Sharing Fundamental to Data-Driven Innovation

News

ADMISSIONS | VISIT | DIRECTORY | ACCESS

What are you interested in?

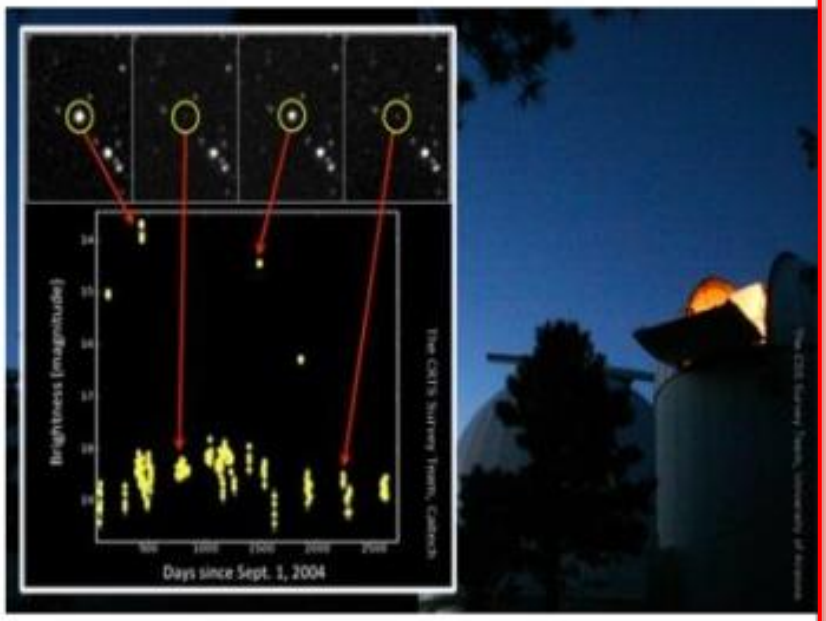
MENU / News & Events / News

Back to All News

01/12/2012

Astronomers Release Unprecedented Data Set on Celestial Objects that Brighten and Dim

PASADENA, Calif.—Astronomers from the California Institute of Technology (Caltech) and the University of Arizona have released the largest data set ever collected that documents the brightening and dimming of stars and other celestial objects—two



AN IMAGE OF A DWARF NOVA, WHICH IS A STAR SYSTEM WHERE MATERIAL FLOWS FROM A RED GIANT STAR TO A DENSE, COMPACT

So

disease in the human brain.

Enlarge This Image

Now, the effort is bearing fruit with a

SIGN IN TO E-MAIL

PRINT

Finding Hope After Miscarriage
January 10, 2013

Do the Brain Benefits of Exercise Last
January 9, 2013

What's been missing?

- Forum for reaching consensus and making decisions
- Basis for acting voluntarily on those decisions

Example of operating a global infrastructure on the basis of cooperation and consensus:

The Internet Society
and the
Internet Engineering Task Force (IETF)

Five Principles for an Open Data Infrastructure:

- Discoverable
- Accessible
- Understandable
- Manageable
- People

Characteristics of a coordinated global data effort:

- **Community-based** – Not a government organization, commercial entity or regulatory body
- **Open** – Membership is open, meetings are public, processes are transparent, and products are free
- **Balanced** – Organized on the principle of balanced representation for individual organizations and stakeholder communities

Characteristics of a coordinated global data effort:

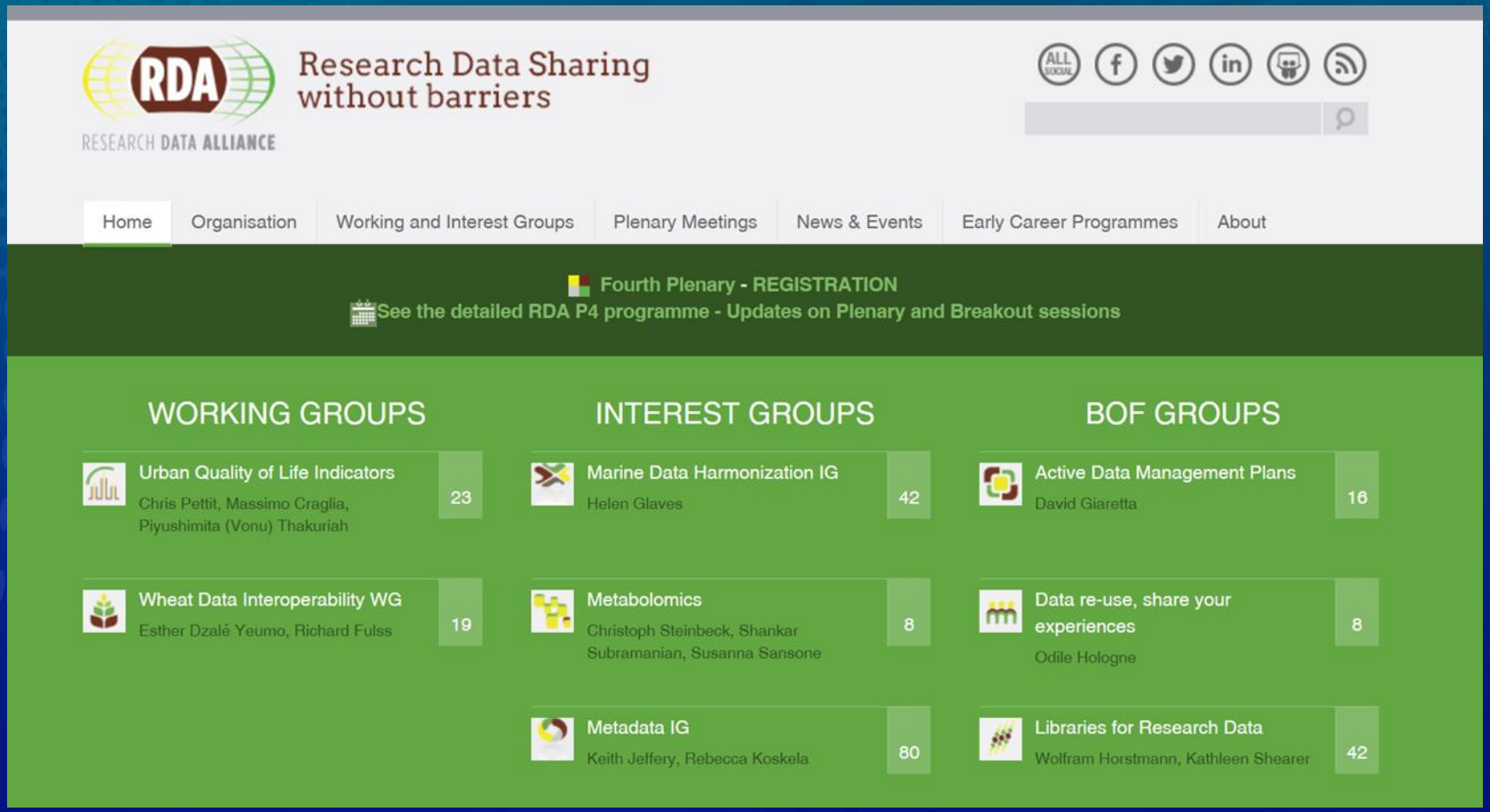
- **Consensus-driven** – Progress through rough consensus, voting to resolve disagreements as required
- **Harmonization-oriented** – Focused on harmonization and early deployment across standards, policies, technologies, tools, and other data infrastructure elements
- **Non-profit** - Does not design, promote, endorse, or sell commercial products or services

Research Data Alliance - RDA

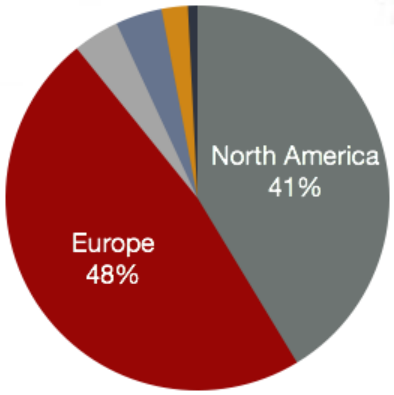
The Research Data Alliance (RDA) builds the social, organizatinal, and technical bridges that enable open sharing of data.

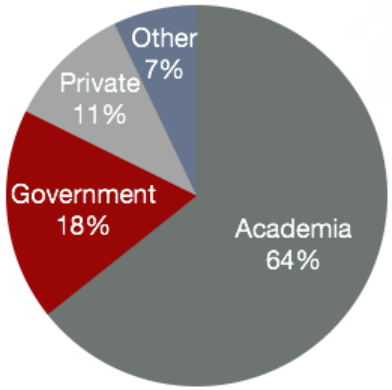
The RDA vision is researchers and innovators openly sharing data across technologies, disciplines, and countries to address the grand challenges of society.

Web Site: rd-alliance.org



The RDA Community Today: Over 1850 members from 80+ countries (as of 6/14)





traveltip.org

RDA and RDA/US are supported in part by the National Science Foundation.

Inside the RDA: Organizational Framework

Source: Fran Berman, RPI

RDA Council

- Fran Berman, Professor of Computer Science, Rensselaer Polytechnic Institute
- Patrick Cocquet, Chief Executive Officer, Cap Digital
- Tony Hey, Vice President, Microsoft Research Connections
- Kay Raseroka, Independent consultant and Trainer, IFLA Building Strong Library Associations
- Doris Wedlich, Chief Science Officer, Karlsruhe Institute of Technology (KIT)
- Ross Wilkinson, Executive Director, Australian National Data Service
- John Wood, Secretary General of the Association of Commonwealth Universities

RDA Approach:
CREATE → ADOPT → USE

RDA Members come together as

- **Working Groups** – 12-18 month efforts to build, adopt, and use specific pieces of infrastructure
- **Interest Groups** – longer-lived discussion forums that spawn Working Groups as specific pieces of needed infrastructure are identified.

Working Group efforts focus on the development and use of data sharing infrastructure

- **Code, policy, infrastructure, standards, or best practices that are adopted and used** by communities to enable data sharing
- **"Harvestable" efforts** for which 12-18 months of work can eliminate a roadblock
- **Efforts that have substantive applicability** to groups within the data community, but may not apply to everyone
- **Efforts for which working scientists and researchers can start today**

RDA and RDA/US are supported in part by the National

RDA Interest (IG) and Working Groups (WG) by Focus

Domain Science - focused

- Toxicogenomics Interoperability IG
- Structural Biology IG
- Biodiversity Data Integration IG
- Agricultural Data Interoperability IG
- Wheat Data Interoperability WG
- Digital Practices in History and Ethnography IG
- Defining Urban Data Exchange for Science IG
- Geospatial IG
- Marine Data Harmonization IG
- RDA/CODATA Materials Data Infrastructure and Interoperability IG
- Research Data Needs of the Photon and Neutron Science Community IG

Community Needs - focused

- Community Capability Model IG
- Engagement IG
- Development of Cloud Computing Capacity and Education in Developing World Research IG
- Ethics and Social Aspects of Data IG

Reference and Sharing - focused

- Data Citation WG
- Standardization of Data Categories and Codes WG
- RDA/CODATA Legal Interoperability IG
- Data Description Registry Interoperability Working Group

Data Stewardship - focused

- Research Data Provenance IG
- RDA/WDS Certification of Digital Repositories IG
- Preservation e-infrastructure IG
- Long-tail of Research Data IG
- RDA/WDS Publishing Data IG
- RDA/WDS Repository Audit and Certification Working Group
- Domain Repositories Interest Group

Base Infrastructure - focused

- Data Foundation and Terminology WG
- Metadata Standards Directory WG
- Practical Policy WG
- PID Information Types WG
- Data Type Registries WG
- Data in Context IG
- Big Data Analytics IG
- Data Brokering IG
- Federated Identity Management IG
- Metadata IG
- PID Interest Group
- Service Management IG

For additional information:

RDA Web Site:

rd-alliance.org

My Contact Info:

chris.greer@nist.gov