



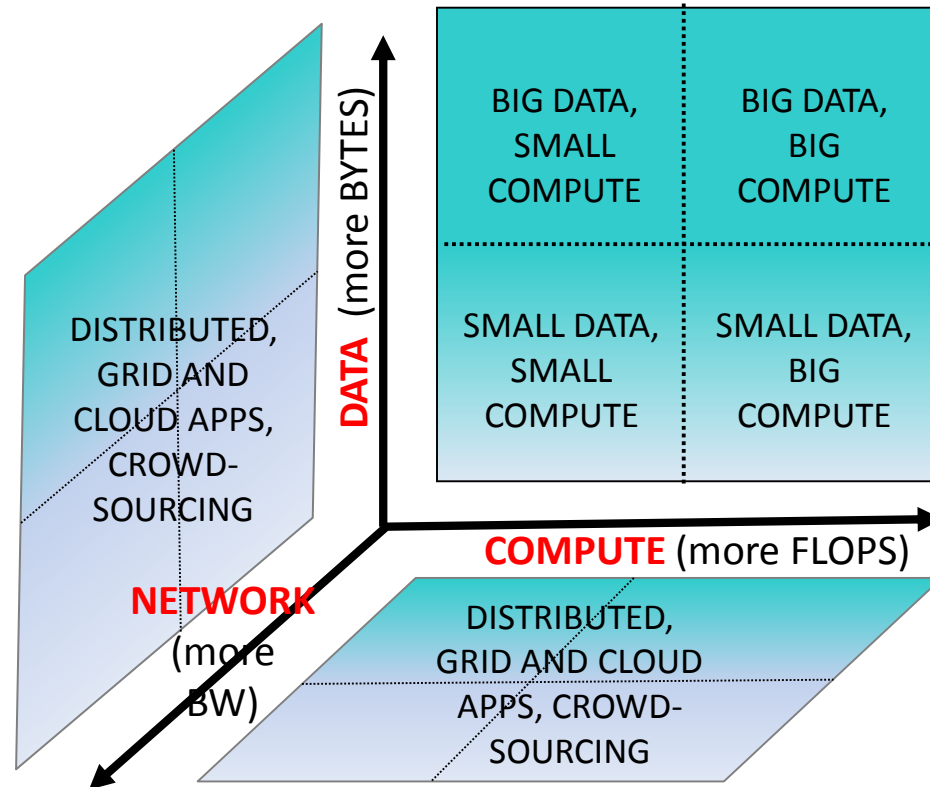
Building Data Sharing Infrastructure at a Global Scale – **the Research Data Alliance**

Dr. Francine Berman

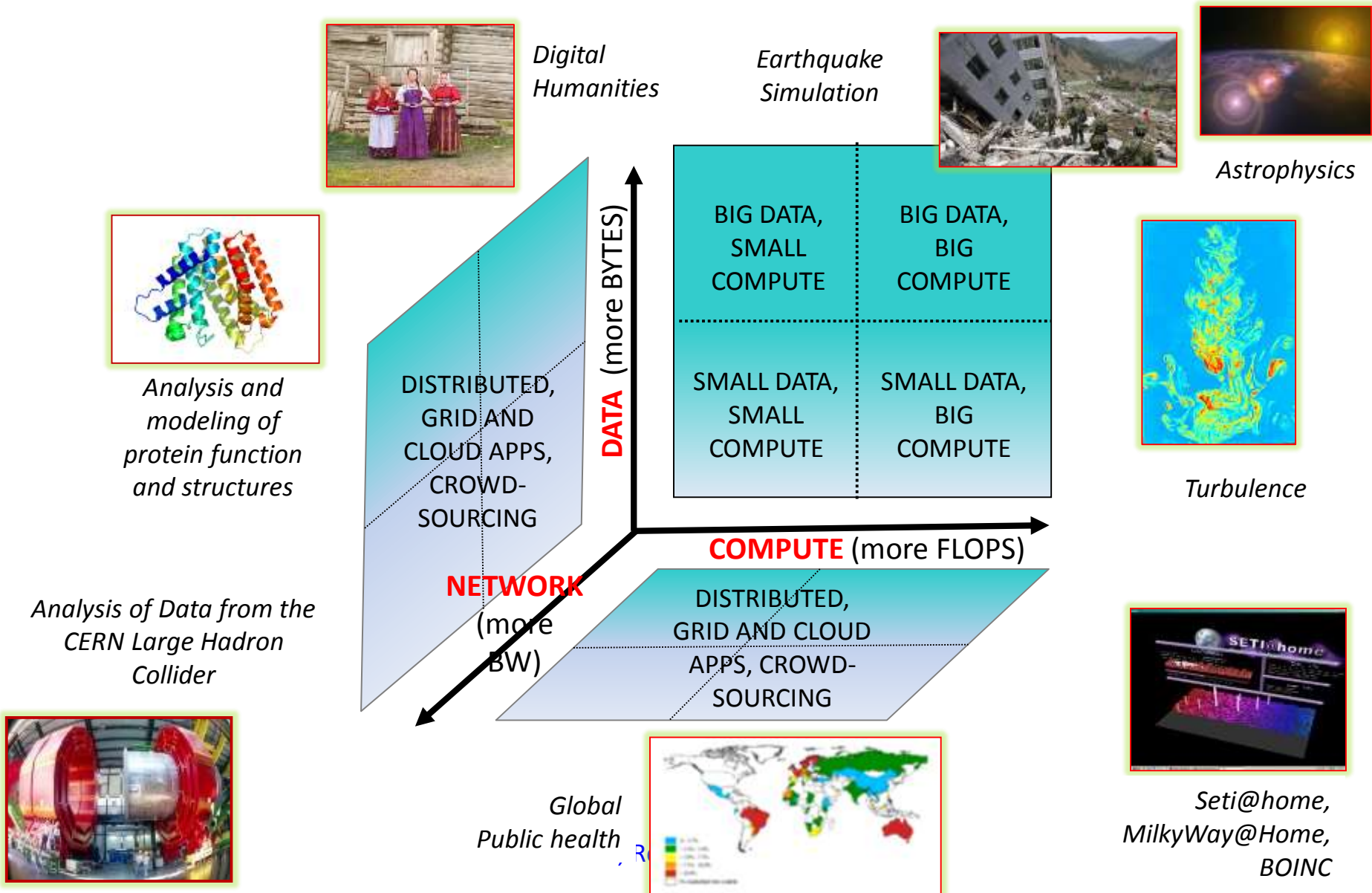
Chair, Research Data Alliance / US

Edward P. Hamilton Distinguished Professor of
Computer Science, Rensselaer Polytechnic
Institute

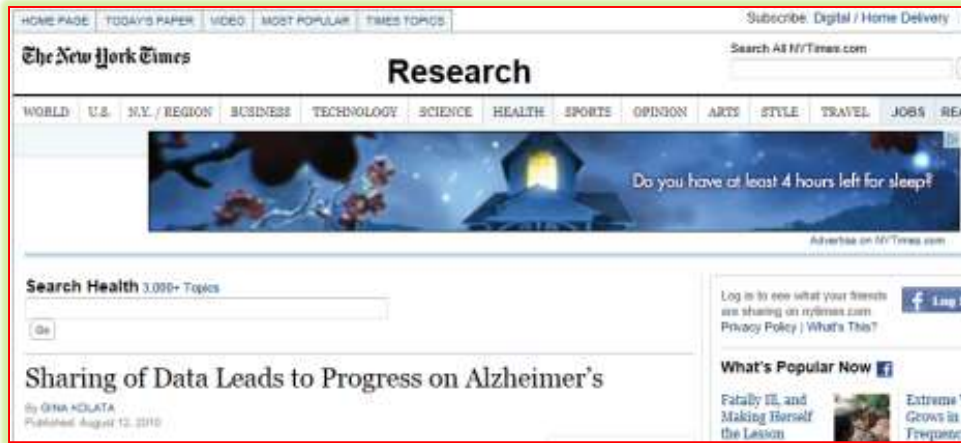
Today's Apps: Infrastructure Needs Span the Spectrum




Accessibility, Availability, Use and Re-Use of Data Accelerating Innovation Across the Board



Data Sharing Fundamental for Driving New Discovery and Advances



The screenshot shows the 'Research' section of The New York Times website. At the top, there are navigation links for 'HOME PAGE', 'TODAY'S PAPER', 'VIDEO', 'MOST POPULAR', and 'TIMES TOPICS'. The main header reads 'The New York Times Research' with a search bar. Below this is a horizontal menu with categories: 'WORLD', 'U.S.', 'N.Y. / REGION', 'BUSINESS', 'TECHNOLOGY', 'SCIENCE', 'HEALTH', 'SPORTS', 'OPINION', 'ARTS', 'STYLE', 'TRAVEL', 'JOBS', and 'REAL ESTATE'. A featured banner image shows a night sky with a house silhouette and the text 'Do you have at least 4 hours left for sleep?'. Below the banner is a search bar for 'Health' with 3,095+ topics. A featured article is titled 'Sharing of Data Leads to Progress on Alzheimer's' by Gina Kolata, published August 12, 2010. To the right, there are sections for 'Log in to see what your friends are sharing on nytimes.com' and 'What's Popular Now' featuring 'Fatally Ill, and Making Himself the Lesson' and 'Extreme V Grows in Frequency'.



The screenshot shows the 'Genetic Literacy Project' website with the tagline 'WHERE SCIENCE TRUMPS IDEOLOGY'. The site has a search bar and navigation links for 'HOME', 'SPECIAL SECTIONS', 'RESOURCES', and 'BROWSE'. On the right side, there are sections for 'Browse by' (Authors, Sources, or by our Advanced Search), 'More from this Source' (Guardian), and 'Public engagement in science'. A 'Popular Articles' section features a 'GMO' icon with a slash through it and the text '10 studies proving GMOs are harmful? Not if science matters'. Below this is an image of various fruits and the text 'From pink grapefruit to...'.

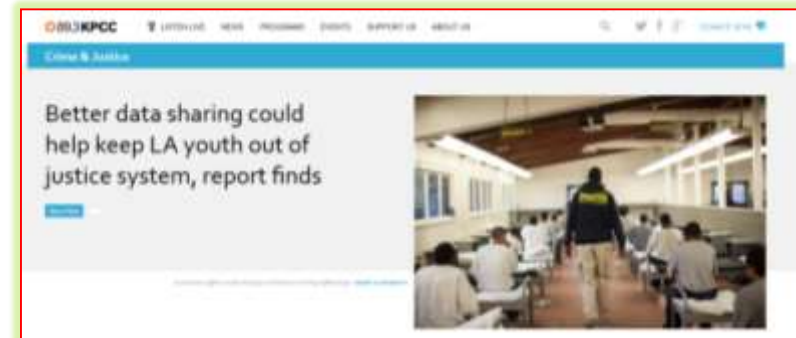
Sharing genomic data speeds cassava breeding for African farmers

Jean-Luc Jannink | April 3, 2015 | Guardian

When the time comes to harvest cassava, a subsistence farmer in Africa – often a woman – only to be told they lack sufficient starch content to fetch a price that the farmer's offered doesn't cover the cost of the fertiliser, let alone the cost of the labour.



The screenshot shows the 'InformationWeek Healthcare' website. The header includes 'InformationWeek Healthcare' and 'Digital Bundle'. Below the header is a navigation menu with categories: 'Software', 'Security', 'Cloud', 'Mobility', 'Social Business', 'Big Data', 'Windows', 'Global CIO', 'Government', 'Healthcare', 'Education', 'Financial', and 'SMB'. A sub-menu for 'Healthcare' includes 'Electronic Medical Records', 'Mobile & Wireless', 'Clinical Information Systems', 'Security & Policy', 'CPE', 'The Patient', 'Leadership', and 'More Healthcare'. A featured advertisement for HP is titled 'Now you can afford them.' with a 'Watch video' button. Below the ad, there is a 'NEWS' section with the headline 'Sharing Psychiatry EHR Data Cuts Readmission Rates' and a 'Get InformationWeek Daily' button. A small text box says 'Don't miss each day's hottest technology news, sent directly to your inbox, including occasional breaking news alerts.'



The screenshot shows the 'NPRCC' website with the headline 'Better data sharing could help keep LA youth out of justice system, report finds'. The page features a large image of a person standing in a courtroom, with other people seated at desks. The website has a navigation menu with links for 'HOME', 'ABOUT US', 'CONTACT US', 'SUPPORT US', and 'DONATE'.

Fran Berman, Research Data Alliance

Both Technical and Social Infrastructure Needed to support Data Sharing



Adopted Policy



Systems Interoperability



Common Types, Standards, Metadata



Sustainable Economics



Adopted Community Practice



Training, Education, Workforce

Getting the World Involved in Building / Coordinating Data Sharing Infrastructure: the Research Data Alliance

- **Research Data Alliance (RDA):** Global **community-driven organization** whose mission is to build the **social and technical bridges (infrastructure)** that enable data sharing.
- **Research Data Alliance Vision:** *Researchers and innovators **openly share data across technologies, disciplines, and countries** to address the grand challenges of society.*



RDA: Accelerate Data Sharing and Interoperability Across Cultures, Communities, Scales, Technologies

■ Technical parts of the data engine:

- Data type registries reference model
- Wheat data interoperability framework



Systems Interoperability



Common Types, Standards, Metadata

■ Rules of the road:

- Common agreement on data citation
- Common practice for data repositories



Policy and Practice

■ Better drivers

- Summer schools in data science and cloud computing in the developing world (with CODATA)
- Data management plan development and monitoring



Sustainable Economics



Training, Education, Workforce

RDA Approach: Solve Problems and Facilitate Progress

RDA Members come together as

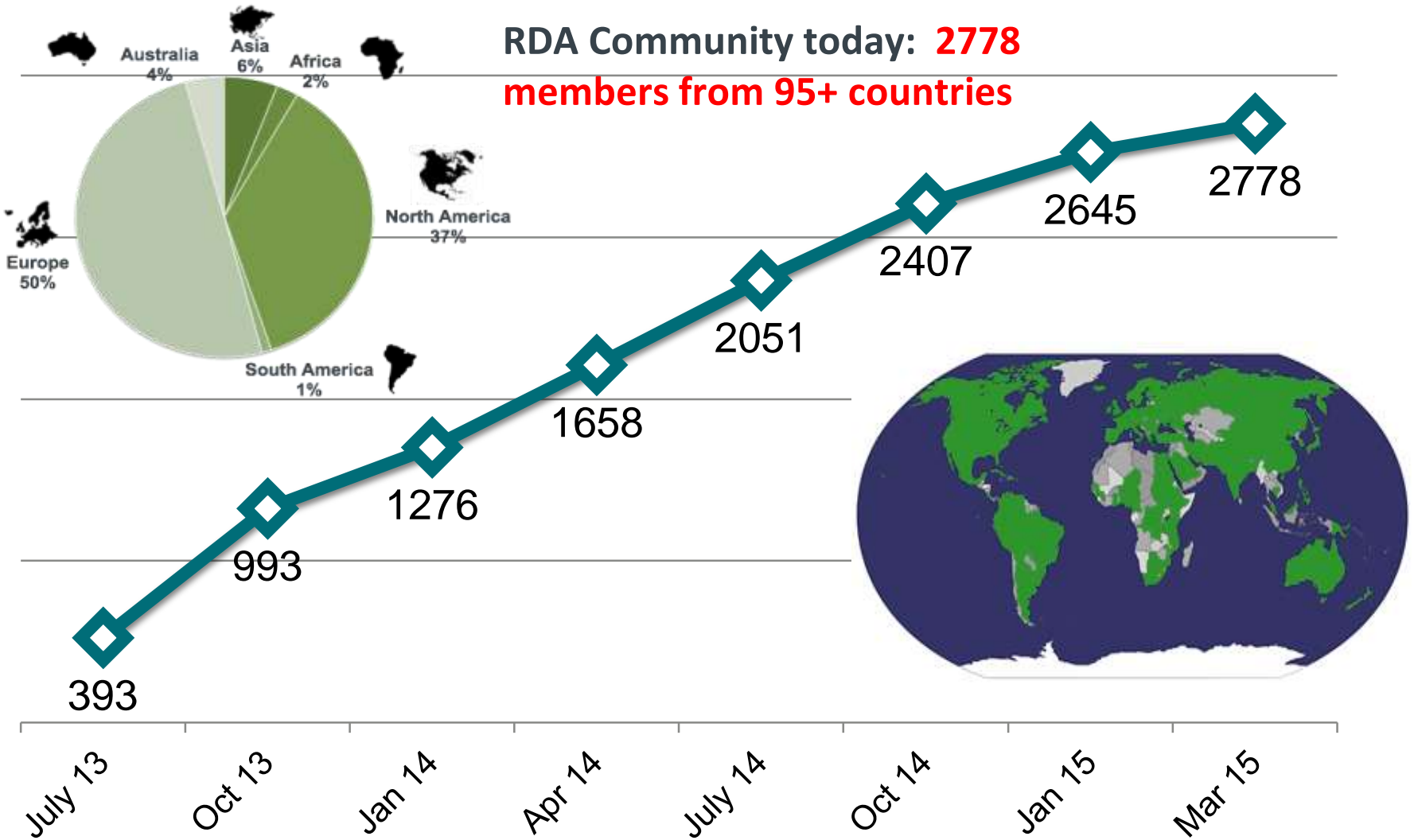
- **Working Groups** – 12-18 month efforts to **build, adopt, and use** specific pieces of infrastructure
- **Interest Groups** – longer-lived discussion forums that spawn Working Groups as specific pieces of needed infrastructure are identified.

RDA culture focuses on the pragmatic:

- **Working Groups must incorporate adopters** – no “build it and they will come”
- **Infrastructure must solve someone’s problem** but not necessarily everyone’s problems – not aiming for universal “esperanto” infrastructure
- **Amplify impact** when possible
 - community proactively enables **additional adopters** (communities, areas, organizations, projects that were not part of the original cohort) for whom RDA infrastructure work products are useful
 - RDA seeks to **collaborate with other organizations** to achieve their goals and strengthen the data community – RDA not looking for “world domination”

RDA Community @ 2: Precipitous Growth

RDA Community today: **2778**
members from 95+ countries



RDA Community at Work: Interest Groups as of March 2015



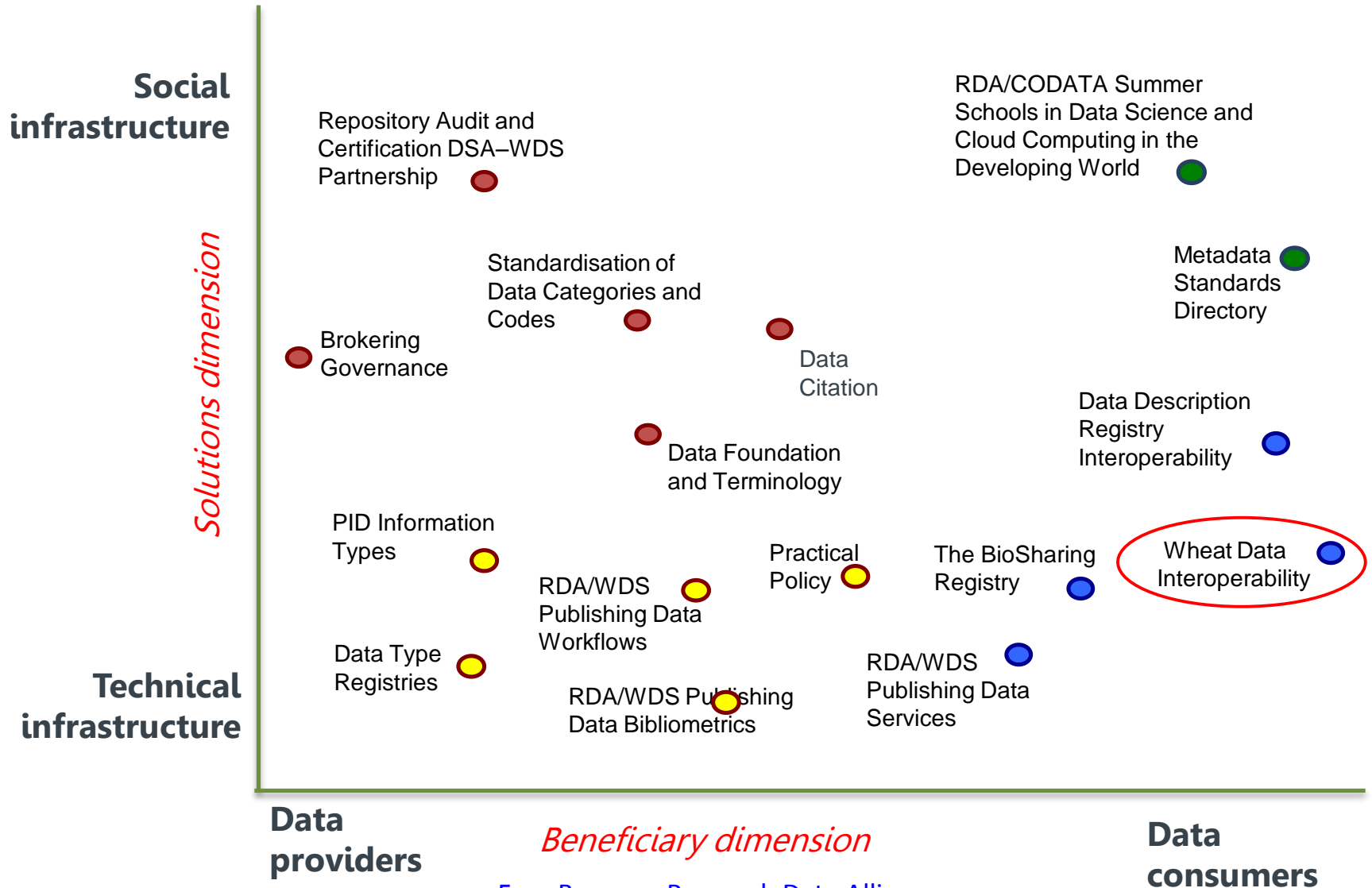
- | | | |
|---|---|---|
| 1. Agricultural Data Interoperability IG | 13. Domain Repositories Interest Group | Interest Group |
| 2. Active Data Management Plans* | 14. Education and Training on handling of research data | 27. RDA/CODATA Legal Interoperability IG |
| 3. Big Data Analytics IG | 15. ELIXIR Bridging Force IG | 28. RDA/CODATA Materials Data, Infrastructure & Interoperability IG |
| 4. Biodiversity Data Integration IG | 16. Engagement IG | 29. RDA/WDS Certification of Digital Repositories IG |
| 5. Brokering IG | 17. Federated Identity Management | 30. RDA/WDS Publishing Data Cost Recovery for Data Centres |
| 6. Community Capability Model IG | 18. Geospatial IG* | 31. RDA/WDS Publishing Data IG |
| 7. Data Fabric IG | 19. Libraries for Research Data | 32. Reproducibility IG |
| 8. Data for Development | 20. Long tail of research data IG | 33. Research data needs of the Photon and Neutron Science community |
| 9. Data Foundations and Terminology IG* | 21. Marine Data Harmonization IG | 34. Research Data Provenance |
| 10. Data in Context IG | 22. Metabolomics | 35. Service Management IG |
| 11. Development of cloud computing capacity and education in developing world research* | 23. Metadata IG | 36. Structural Biology IG |
| 12. Digital Practices in History and Ethnography IG | 24. PID Interest Group | 37. Toxicogenomics Interoperability IG |
- * in review

Domain Repositories Interest Group

(co-Chairs: George Alter/ICPSR, Peter Doorn/DANS, Ruth Duerr/NSIDC, Bob Hanisch/NIST + VAO)

- **Why:** Repositories critical for stewardship and preservation of research data. Common practice and policy can provide greater potential leverage and effectiveness. Exchange of ideas can improve user experience, economic sustainability.
- **What:** **RDA Domain Repositories Interest Group** brings together active data repositories serving many scientific disciplines. Discussions focus on sharing / creating good practice and collaborations around **data curation, dissemination, preservation** and **institutional sustainability**.
- **Value added:** RDA Domain Repositories Interest Group **working with other RDA groups** (data citation, metadata, certification of digital repositories) to adopt/amplify RDA infrastructure useful for their repositories
- **Impact:** Interest Group
 - helping build / strengthen individual repository organizations
 - creating community collaboration among repositories world-wide
 - developing a community that will improve stewardship options for domain researchers

RDA Working Groups Span a Broad Spectrum



Fran Berman, Research Data Alliance

Wheat Data Interoperability Working Group

co-Chairs: Esther Dzale Yeumo Kabore/French National Institute for Agricultural Research, Devika Madalli/Indian Statistical Institute, Johannes Keizer/Food and Agriculture Office of the UN

- **Why:** Wheat information systems needed to answer complex questions such as “*What genes and traits are relevant for understanding the impact of climate change on wheat plant productivity?*”. Diverse data on yield, market pricing, soil analysis, genomic and phenotypic information, etc. must be integrated / coordinated to address complex questions.
- **What:** **RDA Wheat Data Interoperability Working Group** developing a **common integration framework** for describing, representing, linking and publishing wheat data with respect to open standards to support wheat data sharing, use and re-use.
- **Work Products:** RDA Group will
 - Create common **standards and vocabularies** for wheat data management.
 - Create **framework** for Wheat Information System that integrates genomic annotations, phenotypes, genetic maps, physical maps, germplasm.
 - Facilitate access, discovery, use and re-use of Wheat Information System through **development / adoption of common metadata, vocabularies/ontologies/formats, good practice.**
- **Impact:** Working Group deliverables will be incorporated into the **Wheat Information System** of the Global Wheat Initiative and other international efforts, including the **Coherence in Information for Agricultural Research for Development (CIARD)** movement. Next steps: Framework will be adapted to other crops such as **Rice and Maize.**

Next Steps for RDA: Stay Pragmatic, Focus on Impact

More Infrastructure

Continuing pipeline of infrastructure deliverables adopted, used, coordinated and amplified to accelerate data sharing

More effective Community

Increasing coordination and collaboration between domains, sectors, organizations, communities. Effective advocacy for national and international data issues and communities.

Impact-focused Outreach

Stronger partnerships with industry, governments, domains, organizations.

Substantive engagement of students and early career professionals, greater spectrum of international cultures.

- **Next Plenaries** (Plenaries are both community and working meetings. Meetings held twice yearly around the world.):

- September, 2015: **Paris, France (P6)**
- March, 2015: **Tokyo, Japan (P7)**
- September, 2015: **Montreal, Canada or Washington, DC (P8)**
- March, 2016: **Barcelona, Spain (P9)**

Joining RDA:

Go to rd-alliance.org and register

- Must agree to RDA principles (openness, community-driven, etc.)
- Free for individuals



RDA/US: Collaborate Globally, Contribute Locally



■ = RDA/US Members
■ = No RDA/US Members, April '15

- **RDA/US = all U.S. members of RDA**
 - Currently ~1000 members of RDA in 46 states
- **RDA/US Mission:** To build RDA community in the U.S. and leverage RDA momentum to advance the U.S. data community

Current Activities:

Community Development

- **Student / Early Career** programs (supported by **NSF, Sloan Foundation**)
- **Targeted outreach** to data-enabled communities and organizations (supported by **NSF**)

Community support:

- RDA Deliverables **Adoption Amplification seed projects** (supported by **NSF**)
- **International Plenaries participation support** for RDA/US (**NSF**)
- **Coordination meetings** for RDA Working Groups (supported by **NIST, NSF**)

Organizational Support

- **U.S. RDA Plenaries** hosting (supported by **NSF** and sponsors)
- **RDA/US development** – leadership and community building (supported by **NSF**)

RDA/US team:

Steering Committee

- Fran Berman, RPI -- Chair
- Larry Lannom, CNRI – Vice-Chair
- Beth Plale, IU – Vice-Chair
- Kathy Fontaine, RPI – *Managing Director*

Student Resident: Candice Lanius, RPI

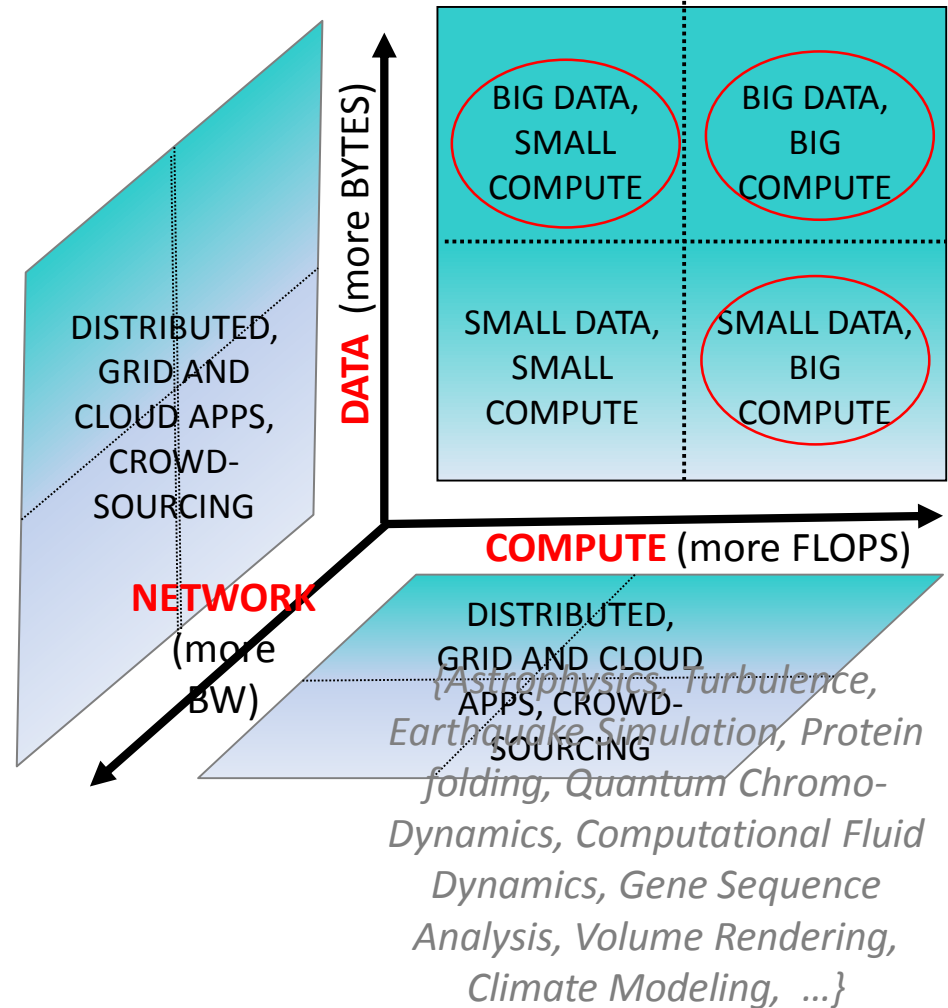
Communications Manager: Yolanda Meleco

Administrative Coordinator: Jamie Lupo-Petta

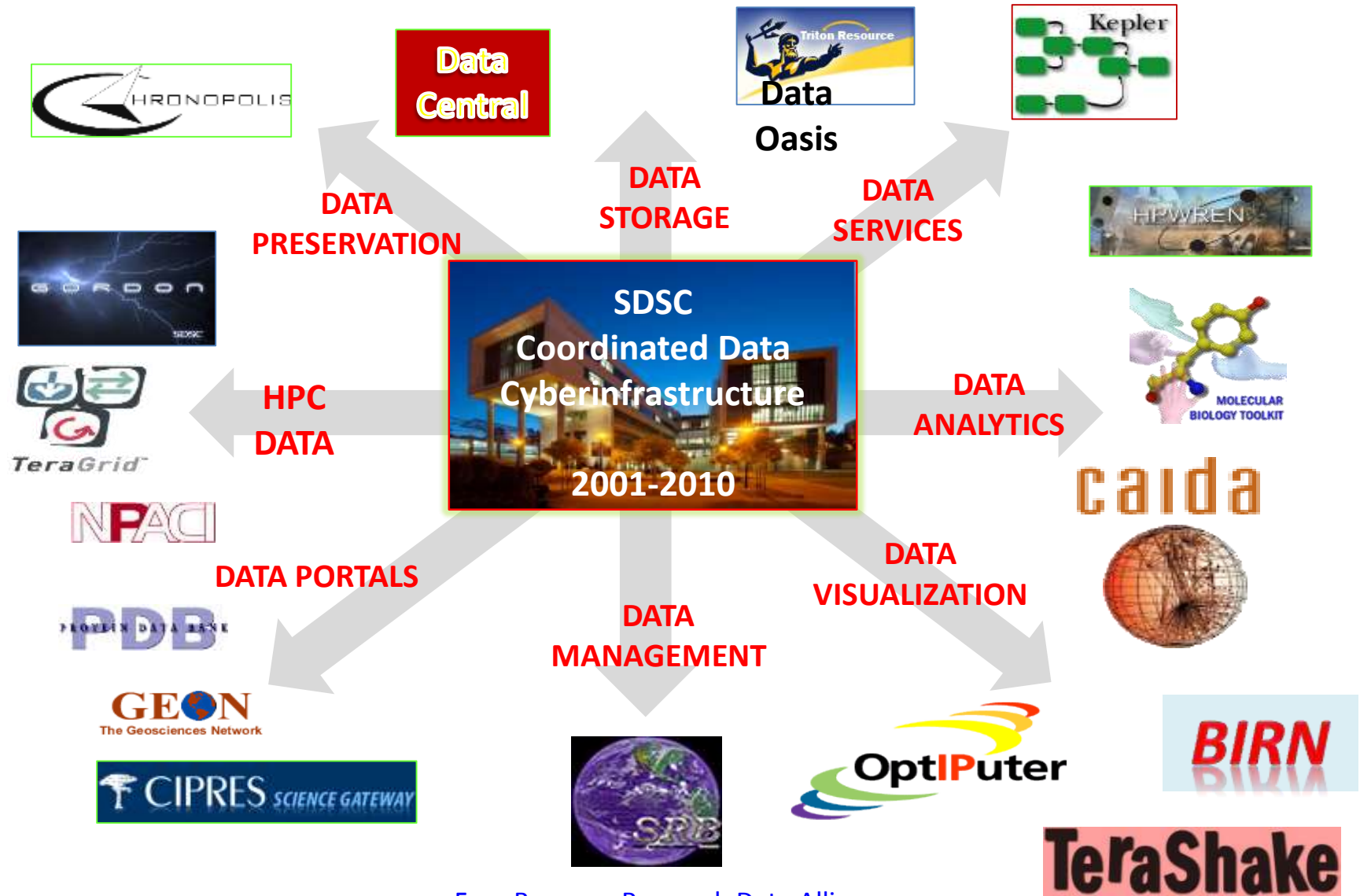
RDA and HPC: RDA an Opportunity to Advance HPC

- **Big compute and big data need extra infrastructure:**

- Data analysis services and tools
- Data visualization
- Gateways / portals / websites for community access
- Temporary data storage
- Options for long-term stewardship and preservation
- Targeted / strategic solutions to “move the compute or move the data” challenges



How can you address Big Data / Big Compute problems at an HPC Center? – San Diego Supercomputer Center 2001-2010



HPC Opportunity: RDA can be a vehicle to build / deploy / utilize and coordinate / promote infrastructure for Data- and Compute-intensive Applications

- Organizations that can support large-scale computations are in a unique position to build / deploy infrastructure as data transfer costs for some applications can be prohibitive.
- **RDA Interest Groups and Working Groups can be an important vehicle for building data infrastructure and developing useful policy / practice for HPC facilities**
 - Coordination of large-scale facilities benefits users
- Why not an RDA **Big Data** \cap **Big Compute** Interest / Working Group?



Opportunities to **Leverage Existing** and **Create New RDA** Groups to advance **Big Data / Big Compute Applications**

- | | | | |
|---|---|---|--|
| 1. Agricultural Data Interoperability IG | Repositories Interest Group | Interoperability IG | 1. Brokering Governance |
| 2. Active Data Management Plans* | 14. Education and Training on handling of research data | 29. RDA/WDS Certification of Digital Repositories IG | 2. Data Citation WG |
| 3. Big Data Analytics IG | 15. ELIXIR Bridging Force IG | 30. RDA/WDS Publishing Data Cost Recovery for Data Centres | 3. Data Description Registry Interoperability |
| 4. Biodiversity Data Integration IG | 16. Engagement IG | 31. RDA/WDS Publishing Data IG | 4. Data Foundation and Terminology WG |
| 5. Brokering IG | 17. Federated Identity Management | 32. Reproducibility IG | 5. Data Type Registries WG |
| 6. Community Capability Model IG | 18. Geospatial IG* | 33. Research data needs of the Photon and Neutron Science community | 6. Metadata Standards Directory Working Group |
| 7. Data Fabric IG | 19. Libraries for Research Data | 34. Research Data Provenance | 7. PID Information Types WG |
| 8. Data for Development | 20. Long tail of research data IG | 35. Service Management IG | 8. Practical Policy WG |
| 9. Data Foundations and Terminology IG* | 21. Marine Data Harmonization IG | 36. Structural Biology IG | 9. RDA/CODATA Summer Schools in Data Science and Cloud Computing in the Developing World |
| 10. Data in Context IG | 22. Metabolomics | 37. Toxicogenomics Interoperability IG | 10. RDA/WDS Publishing Data Bibliometrics WG |
| 11. Development of cloud computing capacity and education in developing world research* | 23. Metadata IG | 38. YOUR NEW INTEREST GROUP HERE | 11. RDA/WDS Publishing Data Services WG |
| 12. Digital Practices in History and Ethnography IG | 24. PID Interest Group | | 12. RDA/WDS Publishing Data Workflows WG |
| 13. Domain | 25. Preservation e-Infrastructure IG | | 13. Repository Audit and Certification DSA-WDS Partnership WG |
| | 26. Quality of Urban Life Interest Group | | 14. Repository Platforms for Research Data* |
| | 27. RDA/CODATA Legal Interoperability IG | | 15. The BioSharing Registry: connecting data policies, standards & databases in life sciences* |
| | 28. RDA/CODATA Materials Data, Infrastructure & | | 16. Wheat Data Interoperability WG |

17. YOUR NEW WORKING GROUP HERE

Rd-alliance.org: Research data sharing without barriers

Thank you



Fran Berman, Research Data Alliance