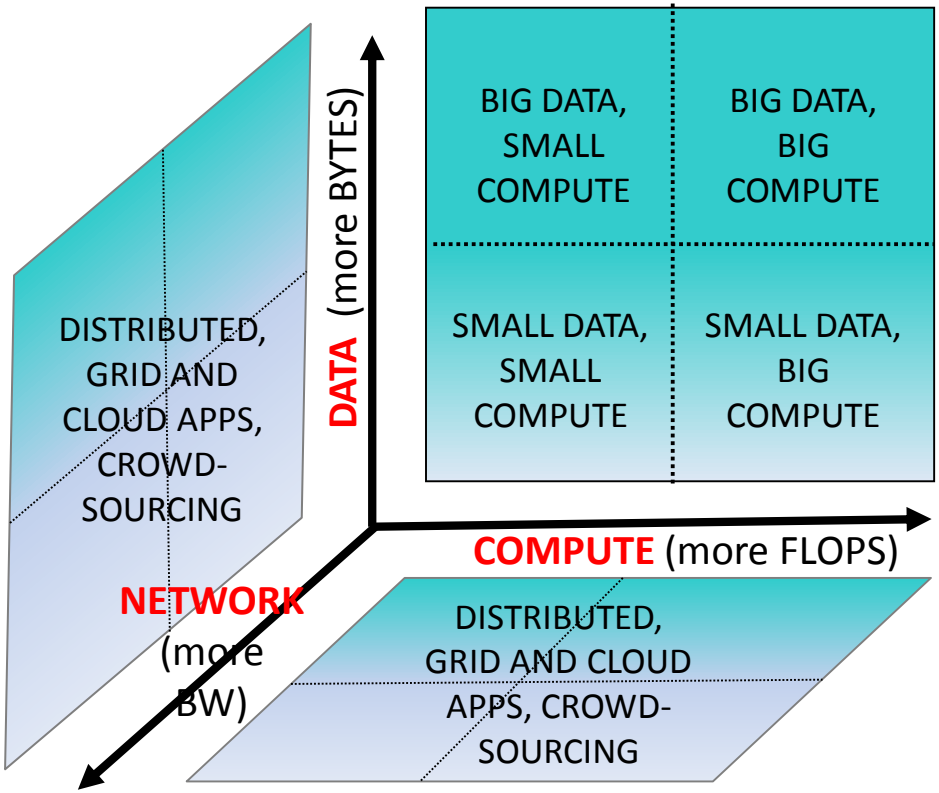
Building Data Sharing Infrastructure at a Global Scale – the Research Data Alliance

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



Fran Berman, Research Data Alliance

Accessibility, Availability, Use and Re-Use of Data

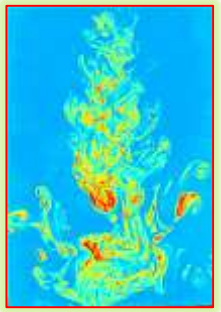
Accelerating Innovation Across the Board

Digital Humanities

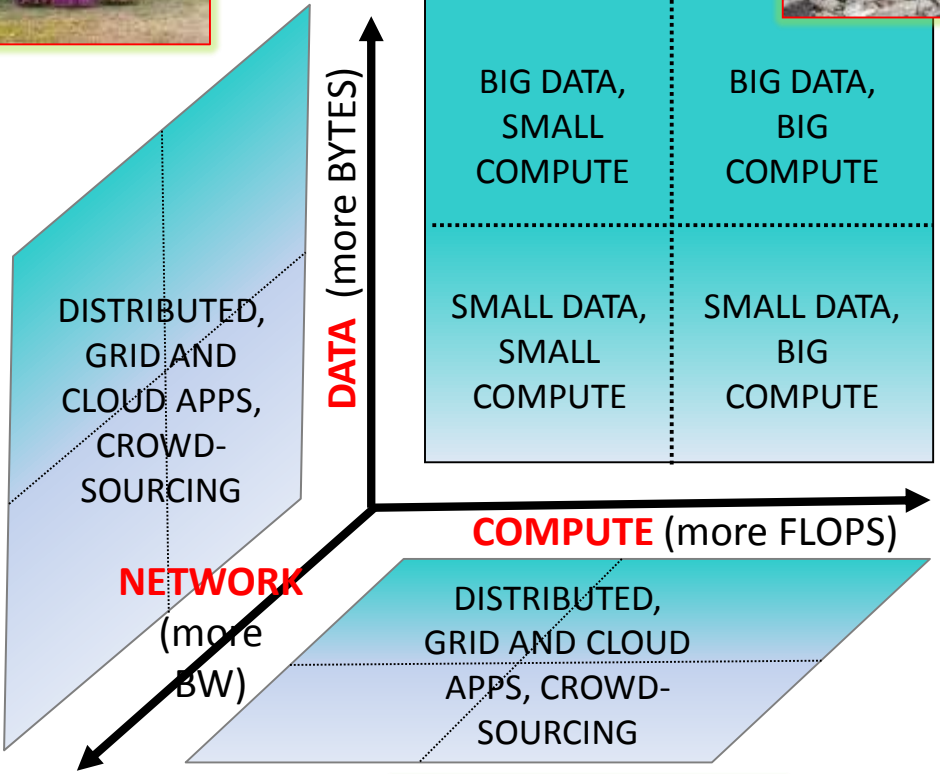
Earthquake Simulation

Astrophysics

Analysis and modeling of protein function and structures



Turbulence



Analysis of Data from the CERN Large Hadron Collider

Global Public health

Seti@home, MilkyWay@Home, BOINC

Data Sharing Fundamental for Driving New Discovery and Advances

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

Fran Berman, Research Data Alliance

Traffic Image: Mike Gonzalez

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
- **Rules of the road:**
  - Common agreement on data citation
  - Common practice for data repositories
- **Better drivers**
  - Summer schools in data science and cloud computing in the developing world (with CODATA)
  - Data management plan development and monitoring

Common Types, Standards, Metadata

Systems Interoperability

Policy and Practice

Sustainable Economics

Training, Education, Workforce

RDA Approach: Solve Problems and Facilitate Progress

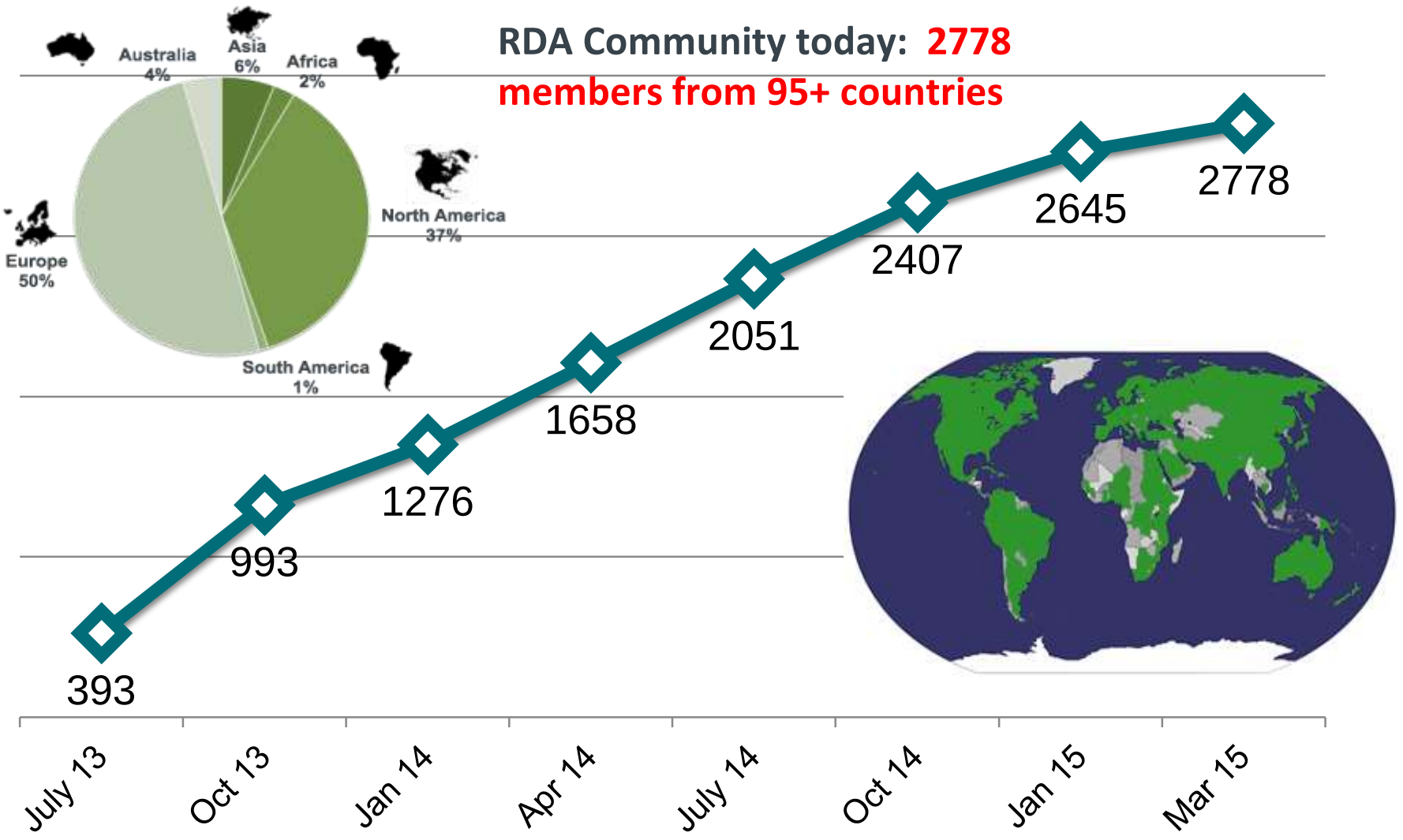
RDA Members come together as

- **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 at Work: Interest Groups as of March 2015

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

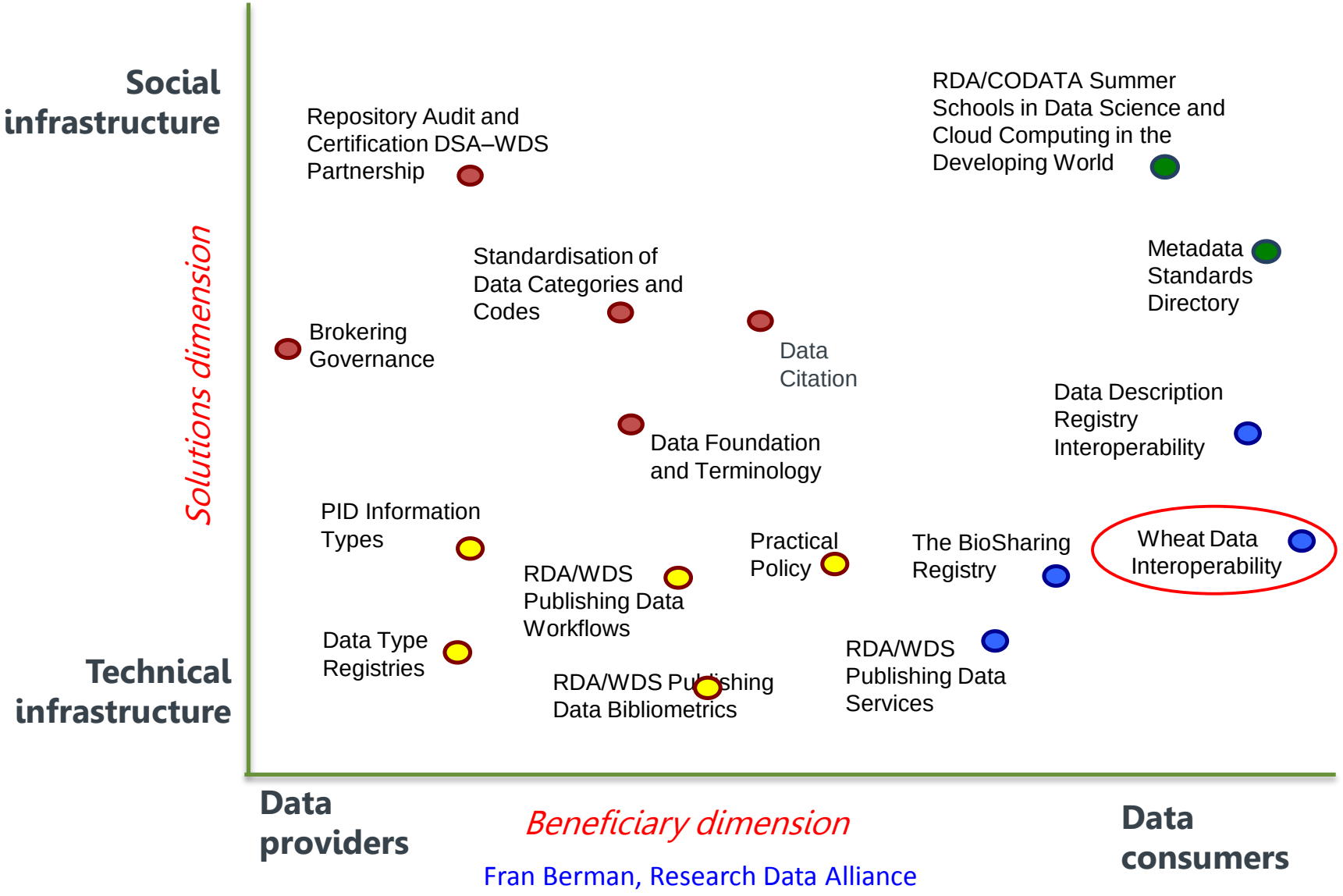
- **Why:** Repositories critical for stewardship and preservation of research data. Common practice and policy can provide greater potential leverage and effectiveness. Exchange of ideas can improve user experience, economic sustainability.

- **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

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

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

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

Fran Berman, Research Data Alliance

RDA/US: Collaborate Globally, Contribute Locally

- **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*

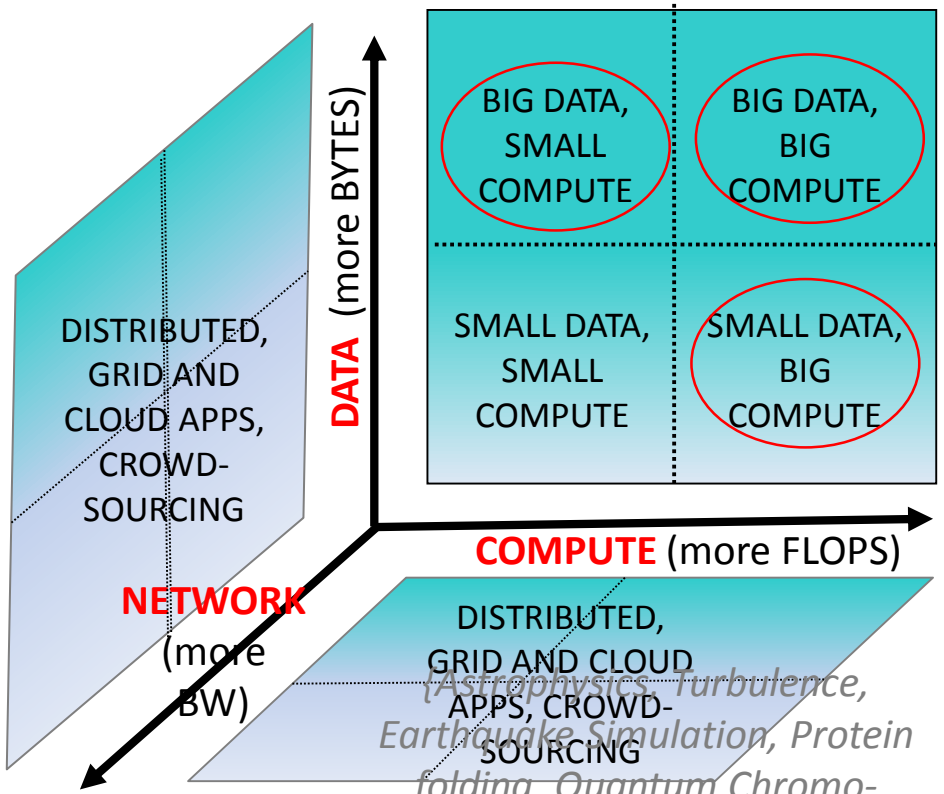
Student Resident: Candice Lanius, RPI

Communications Manager: Yolanda Meleco

Administrative Coordinator: Jamie Lupo-Petta

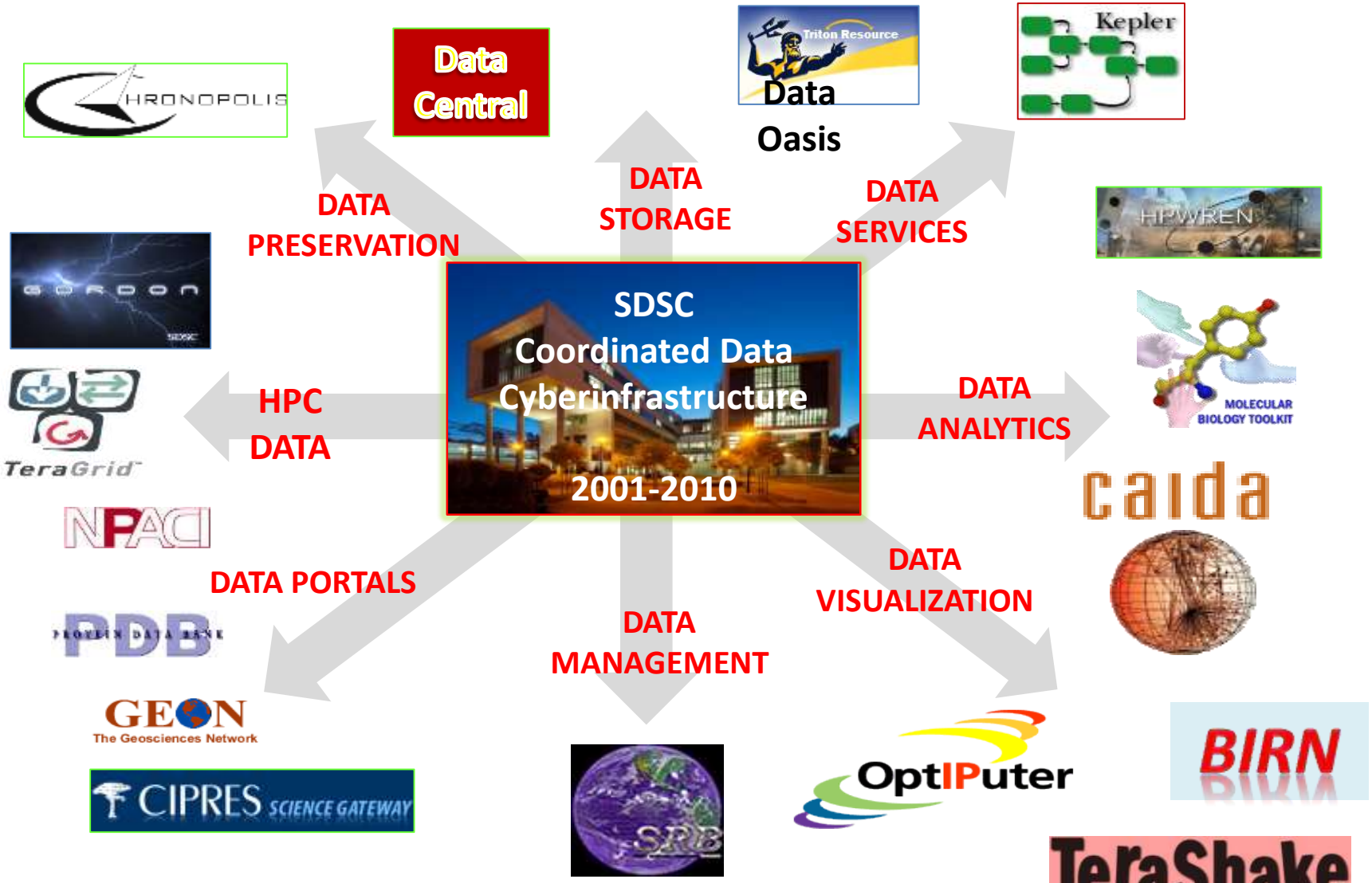
RDA and HPC: RDA an Opportunity to Advance HPC

- **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



{Astrophysics, Turbulence, Earthquake Simulation, Protein folding, Quantum Chromo-Dynamics, Computational Fluid Dynamics, Gene Sequence Analysis, Volume Rendering, Climate Modeling, …}

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

1. Brokering Governance
2. Data Citation WG
3. Data Description Registry Interoperability
4. Data Foundation and Terminology WG
5. Data Type Registries WG
6. Metadata Standards Directory Working Group
7. PID Information Types WG
8. Practical Policy WG
9. RDA/CODATA Summer Schools in Data Science and Cloud Computing in the Developing World
10. RDA/WDS Publishing Data Bibliometrics WG
11. RDA/WDS Publishing Data Services WG
12. RDA/WDS Publishing Data Workflows WG
13. Repository Audit and Certification DSA–WDS Partnership WG
14. Repository Platforms for Research Data*
15. The BioSharing Registry: connecting data policies, standards & databases in life sciences*
16. Wheat Data Interoperability WG
17. **YOUR NEW WORKING GROUP HERE**

Fran Berman, Research Data Alliance

Rd-alliance.org: Research data sharing without barriers

Thank you

Fran Berman, Research Data Alliance