

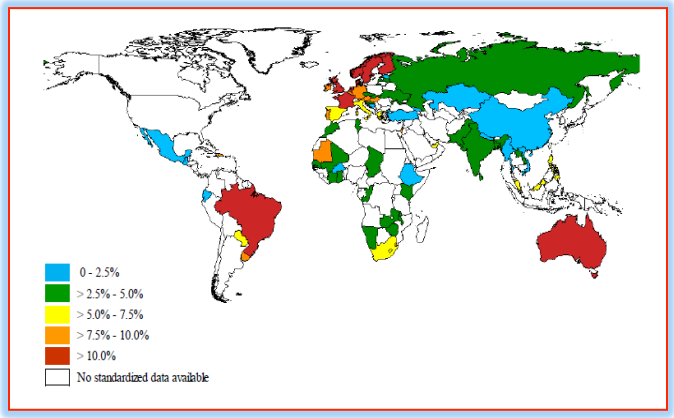
From Research Infrastructure to e-Innovation

Dr. Francine Berman

Chair, Research Data Alliance / US

Edward P. Hamilton Distinguished Professor in Computer Science, RPI

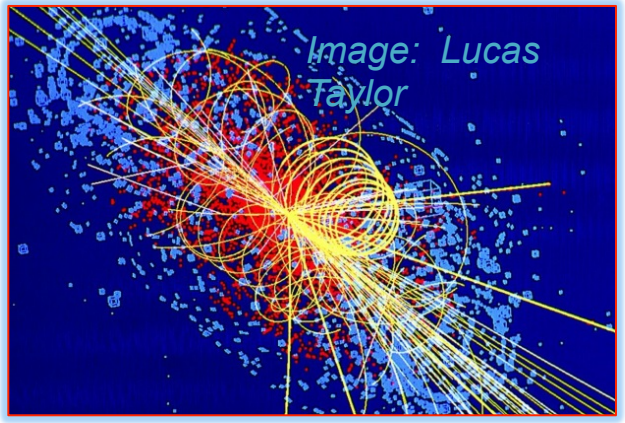
Research Data Needed to Drive Solutions to Complex Scientific and Societal Challenges



Who is most at risk to contract asthma?



How can we increase wheat yields?



How accurate is the Standard Model of Physics?

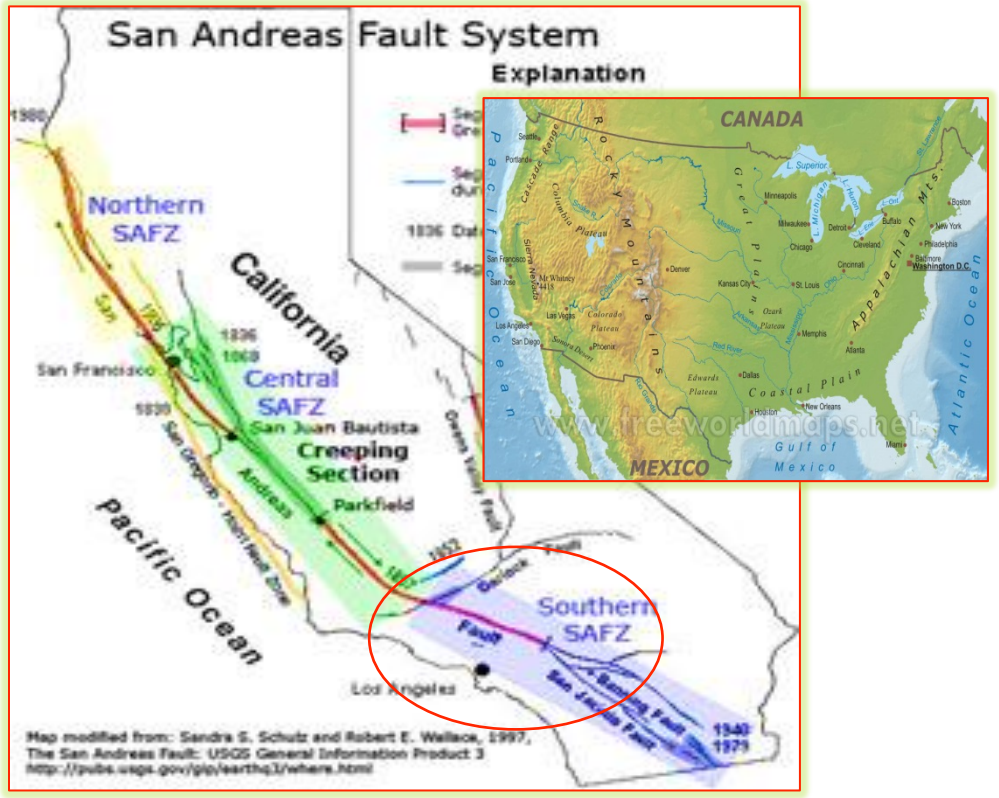


How can we best address energy needs and sustain the environment?

How Can We Respond to Large-Scale Earthquakes?

Earthquake simulations enable

- Enhanced **scientific understanding** of the physical world
- More strategic plans for bridge, building and other physical infrastructure reinforcements to increase **safety**
- Better **disaster response planning** for police, fire fighters, ER teams in high-risk areas to increase their effectiveness

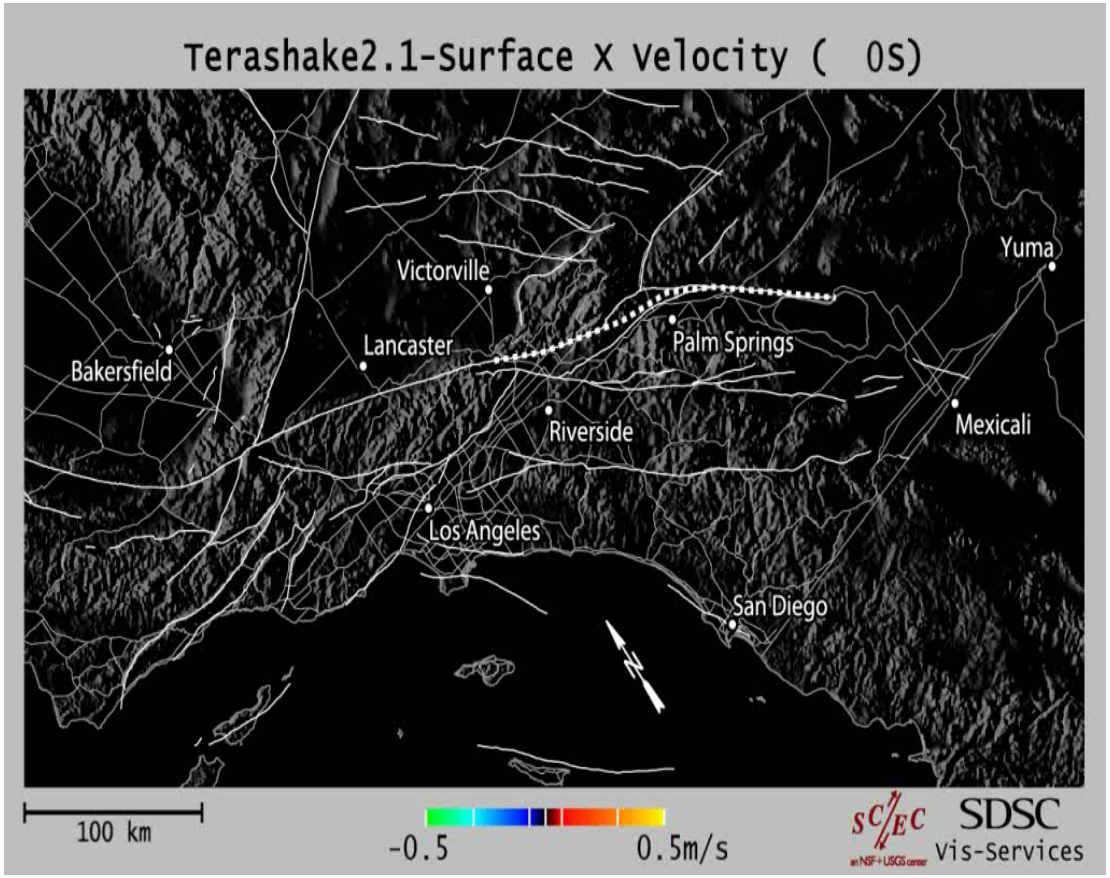


Simulation courtesy of Amit Chourasia, SDSC, Table information courtesy of Southern California Earthquake Center

TeraShake Simulation of 7.7 Earthquake on the Lower San Andreas Fault

Earthquake simulations enable

- Enhanced **scientific understanding** of the physical world
- More strategic plans for bridge, building and other physical infrastructure reinforcements to increase **safety**
- Better **disaster response planning** for police, fire fighters, ER teams in high-risk areas to increase their effectiveness



Simulation courtesy of Amit Chourasia, SDSC, Table information courtesy of Southern California Earthquake Center

"Bigger" Data Requires More Research Infrastructure

Earthquake simulations enable

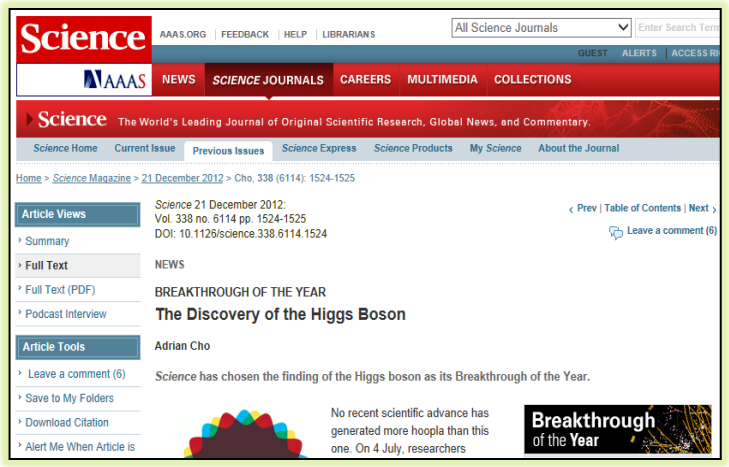
- Enhanced **scientific understanding** of the physical world
- More strategic plans for bridge, building and other physical infrastructure reinforcements to increase **safety**
- Better **disaster response planning** for police, fire fighters, ER teams in high-risk areas to increase their effectiveness

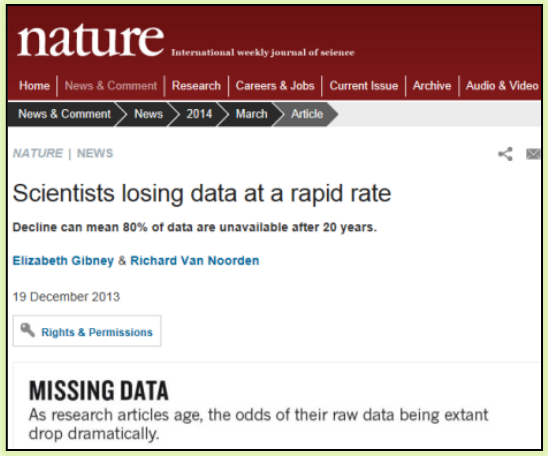
Estimated figures for simulated 240 second period, 100 hour run-time	**TERASCALE:** TeraShake domain (600x300x80 km^3)	**PETASCALE:** PetaShake domain (800x400x100 km^3)
Fault system interaction	NO	YES
Inner Scale	200m	25m
Resolution of terrain grid	1.8 billion mesh points	2.0 trillion mesh points
Magnitude of Earthquake	7.7	8.1
Time steps	20,000 (.012 sec/step)	160,000 (.0015 sec/step)
Surface data	**1.1 TB**	**1.2 PB**
Volume data	**43 TB**	**4.9 PB**

Simulation courtesy of Amit Chourasia, SDSC, Table information courtesy of Southern California Earthquake Center

Why Isn't Research Infrastructure a First-Class Priority?

	Research	Infrastructure
What's newsworthy?	New discovery, breakthroughs	Failure of systems
What is the value proposition?	Domain and national leadership and competitive-ness	Enabler of innovation across sectors
How is it funded?	One-time funding	Continuous, long-term support





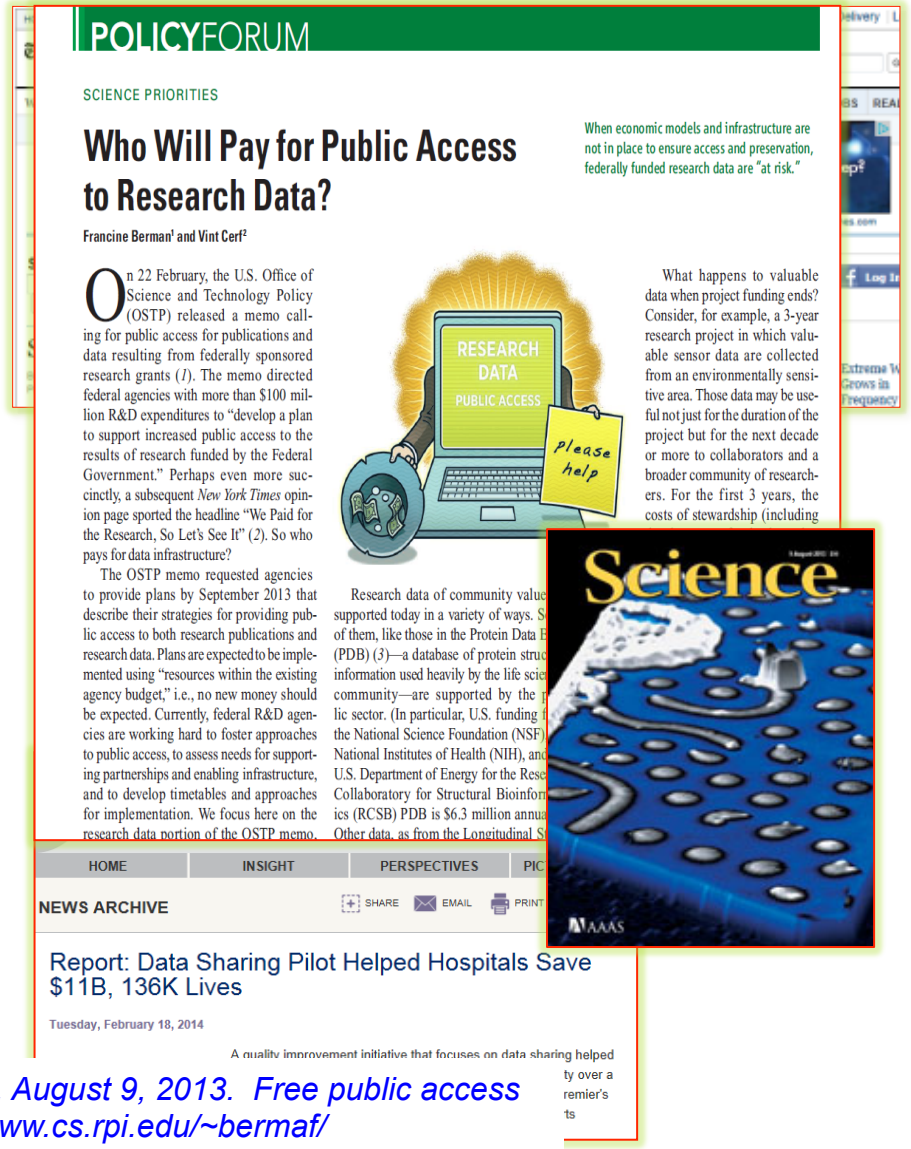




Rensselaer

Research Data is a Public Good

- National governments increasingly calling for public access to research data.
  - *Valuable to science and society as a driver of current and future innovation*
- **Research infrastructure is necessary** to support
  - *Use and re-use*
  - *Research reproducibility*
  - *Federally mandated data management plans*
  - *Public access to research data*



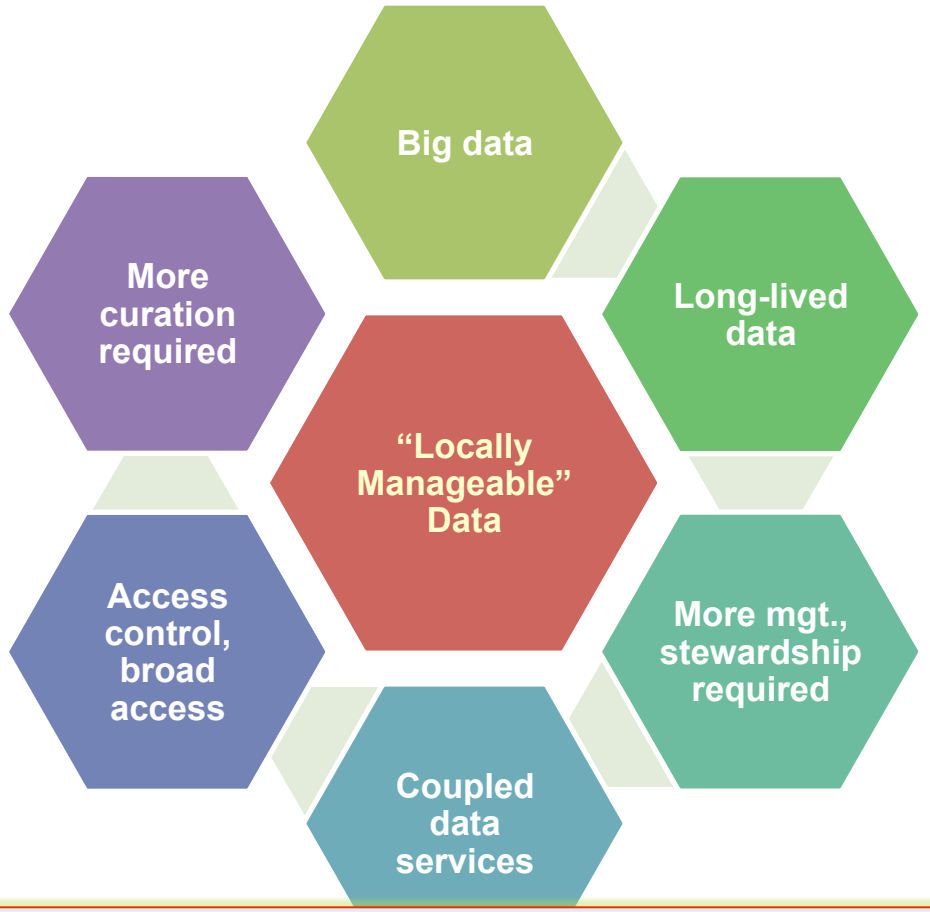
Article: Science Magazine, August 9, 2013. Free public access link at http:/www.cs.rpi.edu/~bermaf/

What Contributes to the Costs of Research Data Infrastructure?

Greater costs beyond the center (including "big" data)

The "data bill" includes

- *Technical infrastructure*
- *Social and organizational infrastructure*
- *Human infrastructure*



As a public good, public support is critical to ensure that research data is accessible and shared to maximize its potential for innovation

Community Driver: Research Data Alliance (RDA) Building Data Infrastructure World-wide to accelerate Data Sharing and Exchange

- RDA community efforts focus on building **social, organizational and technical infrastructure** to
  - *reduce barriers to data sharing and exchange*
  - *accelerate the development of coordinated global data infrastructure*



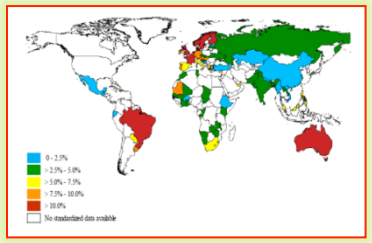
CREATE → ADOPT → USE

Relevant research data infrastructure: Tools, policy, community standards, and best practices **that are adopted and used** by communities to enable data sharing

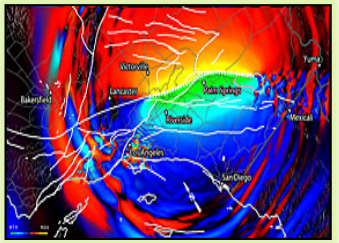
RDA
RESEARCH DATA ALLIANCE

Rensselaer

RDA Infrastructure Providing a Foundation for Access and Use of research data – Now and in the Future







Research Dissemination and Reproducibility

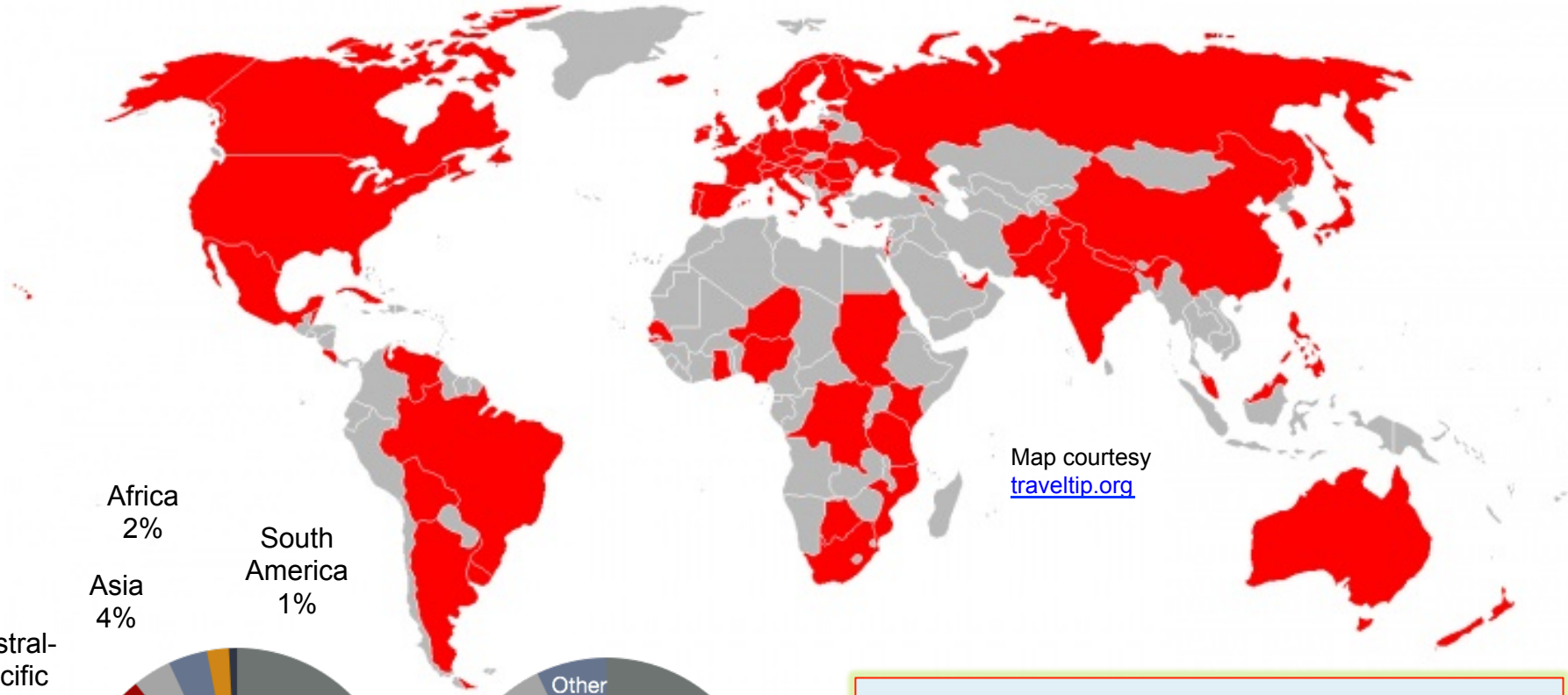
Data Use and Re-use

Data Access (now) and Preservation (later)

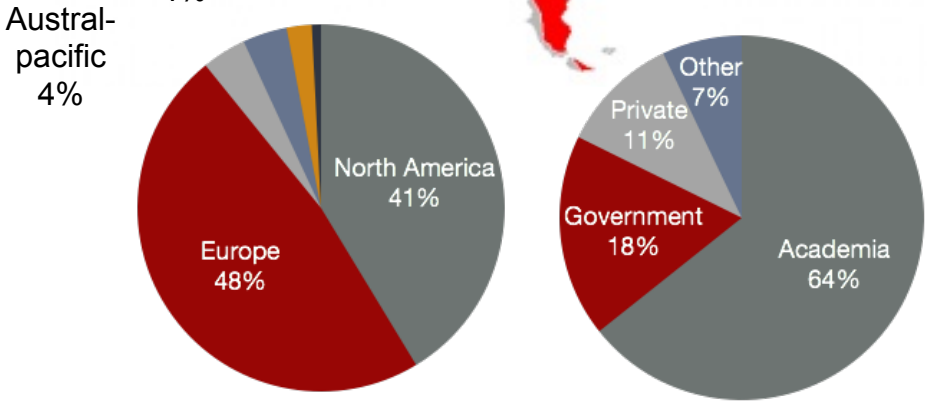
Data Discovery and Data Sharing

Technical, social, organizational, and human infrastructure

The RDA Community: Over 1500 members from 70+ countries (as of 15/3/14)



Map courtesy traveltip.org



Bootstrapping for Innovation:

Bottom-up community efforts and **top-down public sector interest** key to RDA's impact

Value Proposition: Research Infrastructure Needed to Drive e-Innovation

The Research landscape is changing

- Data driving innovation and discovery: Greater need for access, ease-of-use, interoperability of data
- Traditional modes of research recognition evolving: New approaches to collaboration / competition, publication, citation, analysis all involve digital data

The Educational landscape is changing

- University curricula becoming more data-driven
- Increasing integration of on-line / on-site options supported by data infrastructure
- More digital monitoring, tracking, accountability needed; more policy and regulation involving digital data

The Workforce is changing

- More data literacy required from everyone
- More data science embedded in everything
- Data scientists increasingly critical for competitiveness and leadership

Rensselaer

Thank You

