Sustaining the Data Ecosystem –

There is no free lunch but you still need to eat …

CCDSC 2016

Dr. Francine Berman

Chair, Research Data Alliance/US

Hamilton Distinguished Professor, RPI

Why does Sustainability Matter?

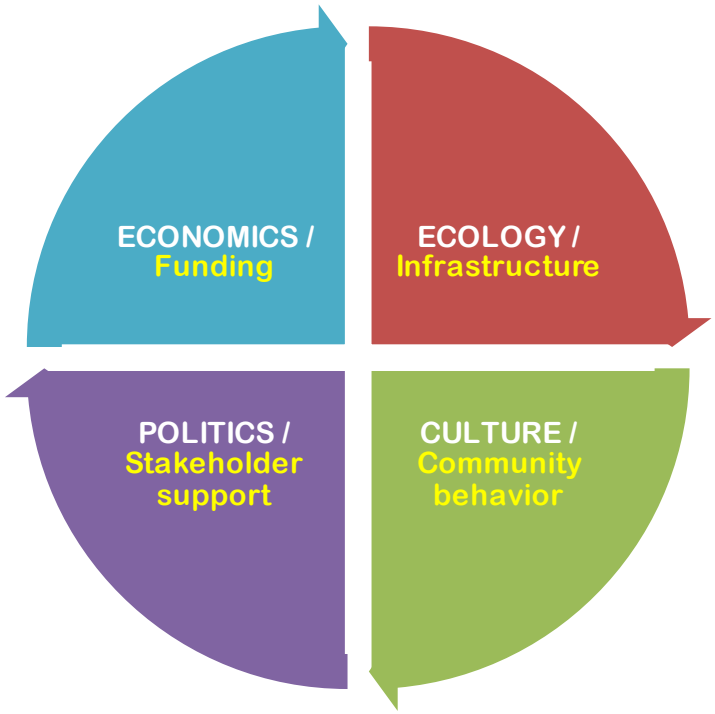
- Data drives discovery and innovation
- **Sustainable data ecosystem** necessary to support
  - Public access to research data
  - Use and re-use of data
  - Reproducibility of results
  - Data management plans
- Data stewardship and preservation fundamental: "Homeless" data ceases to exist

Social and Technical Approaches Both Needed for Sustainability

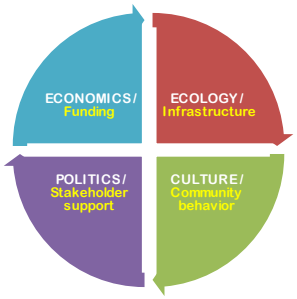
Sustainable development:

"development that meets the needs of the present without compromising the ability of future generations to meet their own needs."

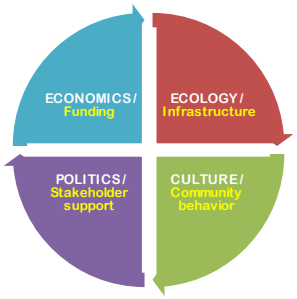
Our Common Future, U.N. Brundtland Commission



- Key components
  - Ecological sustainability
  - Cultural sustainability
  - Economic sustainability
  - Political sustainability



Ecology / Infrastructure -- Making data available isn't good enough



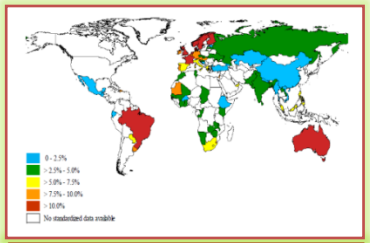
Ecology / Infrastructure -- Making data available isn't good enough

- Infrastructure needed to support data-driven research and innovation.
  - Data is not an asset if you don't know what it means.
  - Data is not useful if you can't find it.
  - Data needs to be in the right form for analysis.
  - Data needs to be preserved for results to be reproducible.

Rensselaer

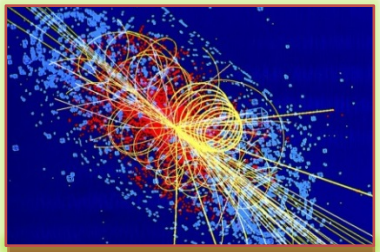
RDA
RESEARCH DATA ALLIANCE

Technical and Social Infrastructure Needed to Support Data-Driven Research

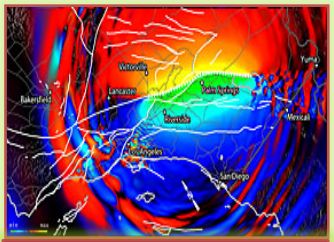


Who is at risk for asthma?

How do we increase agricultural productivity?



How accurate is the Standard Model of Physics?



What will happen in an earthquake?

Interoperability Frameworks

Curation Practice and Policy

Data Discovery Tools

Digital Object Identifiers

Common Metadata Standards

Sustainable Economics

Domain and Institutional Repositories

Data Analytics Algorithms

Data Access and Distribution Policy

Data Sharing Policy

Data Citation Standards

Auditing, Certification and Reporting Practice

Rensselaer

RDA RESEARCH DATA ALLIANCE

Fran Berman

Accelerating the building and coordinating better/more/useful data infrastructure – Research Data Alliance (RDA)

Research Data Alliance (RDA)
rd-alliance.org:

Global community-driven organization whose mission is to build and deploy **social and technical infrastructure** that enables data sharing.

Membership: **4300**+ from **110** countries, all sectors, and a broad spectrum of domains:

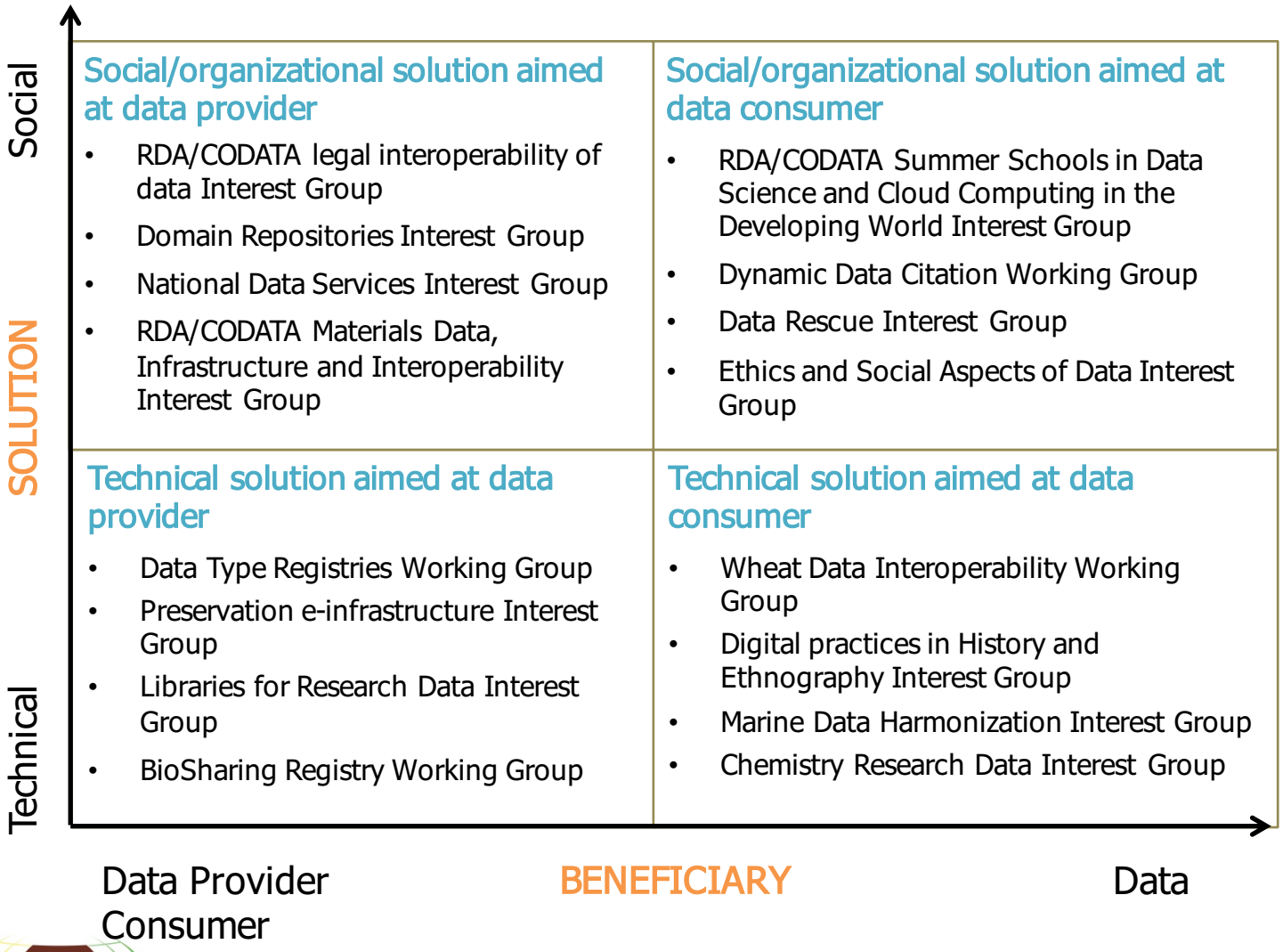
- Broad community spanning "**data consumers**" to "**data providers**" including domain scientists, data scientists, data professionals, information scientists, librarians, computer scientists, technologists, policy makers, educators, etc.

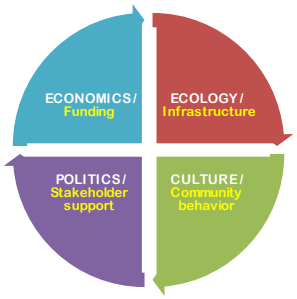
- **RDA Interest Groups** – *identify/explore data infrastructure needed to enable data-driven research*
  - Domain Repositories Interest Group
  - Chemistry Research Data Interest Group
  - Legal Interoperability Interest Group
  - Health Data Interest Group
  - <You initiate> Interest Group
- **RDA Working Groups** – *build and deploy infrastructure that addresses specific problems*
  - Dynamic Data Citation Working Group
  - Wheat Data Interoperability Working Group etc.
  - <You initiate> Working Group
- **Adopters** – *utilize RDA infrastructure to improve local environment for data sharing and data-driven research.*

RDA focus: (70+) RDA Working Groups and Interest Groups fostering better Curation, Management, Stewardship and Use

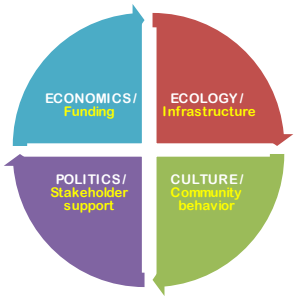
Policy, Good Practice, Community Standards, Education, Awareness, etc. …

Tools, frameworks, models, registries, portals, etc.





Economics / Funding – Who should pay the data bill and what do we need to support?



Economics / Funding – Who should pay the data bill and what do we need to support?

Data infrastructure costs increase with usage, stewardship and access requirements, perceived value

Greater costs at the extremes (including "big" data) …

Data Center Costs include

- Maintenance and upkeep
- Software tools and packages
- Utilities (power, cooling)
- Space
- Networking
- Security and failover systems
- People (expertise, help, infrastructure management, development)
- Training, documentation
- Monitoring, auditing
- Reporting costs
- Costs of compliance with regulation, etc.

Why are Infrastructure Investments such a hard sell?

- Quantifying opportunity cost a challenge
- Hard to "market" compared to more urgent/newsworthy/short-term competing priorities
- Business model must be sustainable and address infrastructure refresh and evolution

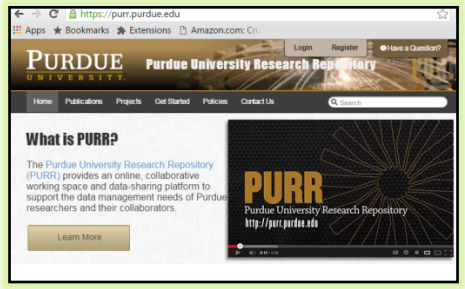
	Archival Storage Systems	Supercomputers
Metrics of Success	High reliability; Minimal data loss and damage	High Performance; good ranking on the Top500 list; application impact
Next Generation Systems	Smooth migration for data key: Preservation collections must migrate to new media without loss of data or disruption to users	Growth in capability/capacity key: Compatibility of systems not required although there should be application transition paths
Funding Model	No gaps. Funding must be available for continuous support of data collections	Serial "one time" funding for each new HPC resource possible

There's no free lunch but you still have to eat

How can we pay for/sustain research data and infrastructure?

Academic Sector

Create sustainable university library and domain repository stewardship options



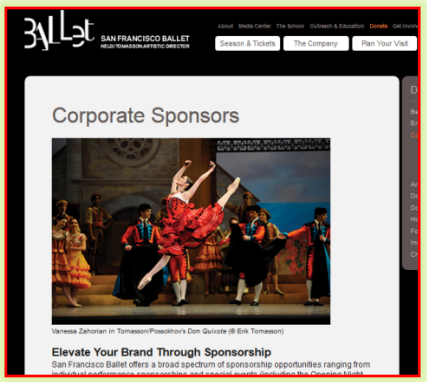
Not govt. supported → ?? → Govt. supported

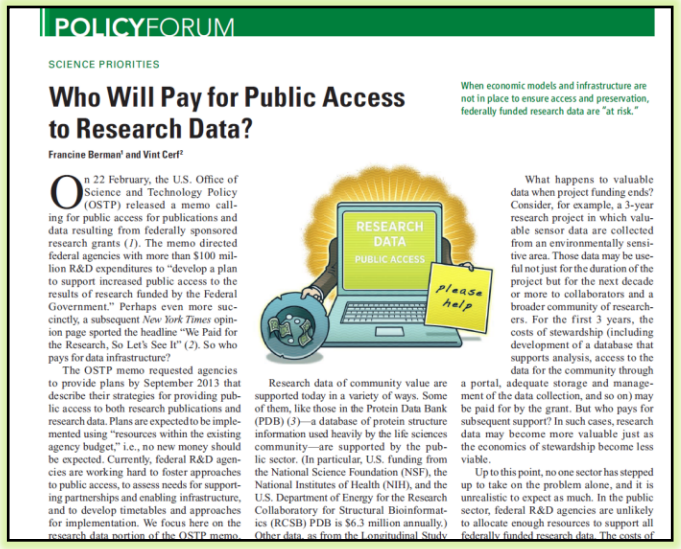
Public Sector

Clarify public sector stewardship commitments: articulate what data will / won't be supported

Private Sector

Facilitate private sector stewardship of public access research data as a public good




POLICYFORUM

SCIENCE PRIORITIES

Who Will Pay for Public Access to Research Data?

When economic models and infrastructure are not in place to ensure access and preservation, federally funded research data are "at risk."

Francine Berman[1] and Vint Cerf[2]

On 22 February, the U.S. Office of Science and Technology Policy (OSTP) released a memo calling for public access for publications and data resulting from federally sponsored research grants (*1*). The memo directed federal agencies with more than $100 million R&D expenditures to "develop a plan to support increased public access to the results of research funded by the Federal Government." Perhaps even more succinctly, a subsequent *New York Times* opinion page sported the headline "We Paid for the Research, So Let's See It" (*2*). So who pays for data infrastructure?

The OSTP memo requested agencies to provide plans by September 2013 that describe their strategies for providing public access to both research publications and research data. Plans are expected to be implemented using "resources within the existing agency budget," i.e., no new money should be expected. Currently, federal R&D agencies are working hard to foster approaches to public access, to assess needs for supporting partnerships and enabling infrastructure, and to develop timetables and approaches for implementation. We focus here on the research data portion of the OSTP memo.

Research data of community value are supported today in a variety of ways. Some of them, like those in the Protein Data Bank (PDB) (*3*)—a database of protein structure information used heavily by the life sciences community—are supported by the public sector. (In particular, U.S. funding from the National Science Foundation (NSF), the National Institutes of Health (NIH), and the U.S. Department of Energy for the Research Collaboratory for Structural Bioinformatics (RCSB) PDB is $6.3 million annually.) Other data, as from the Longitudinal Study

What happens to valuable data when project funding ends? Consider, for example, a 3-year research project in which valuable sensor data are collected from an environmentally sensitive area. Those data may be useful not just for the duration of the project but for the next decade or more to collaborators and a broader community of researchers. For the first 3 years, the costs of stewardship (including development of a database that supports analysis, access to the data for the community through a portal, adequate storage and management of the data collection, and so on) may be paid for by the grant. But who pays for subsequent support? In such cases, research data may become more valuable just as the economics of stewardship become less viable.

Up to this point, no one sector has stepped up to take on the problem alone, and it is unrealistic to expect as much. In the public sector, federal R&D agencies are unlikely to allocate enough resources to support all federally funded research data. The costs of

Public access version at http://www.cs.rpi.edu/~bermaf/

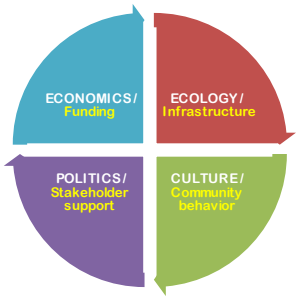
Individuals

Charge low-barrier-to-access fees for data / Advertise / Subscribe

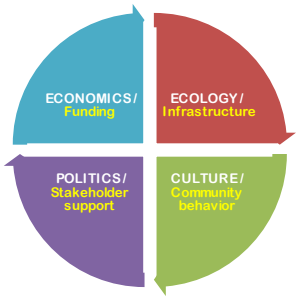
Evolve research culture to adapt what works in the private sector

Rensselaer

RDA RESEARCH DATA ALLIANCE

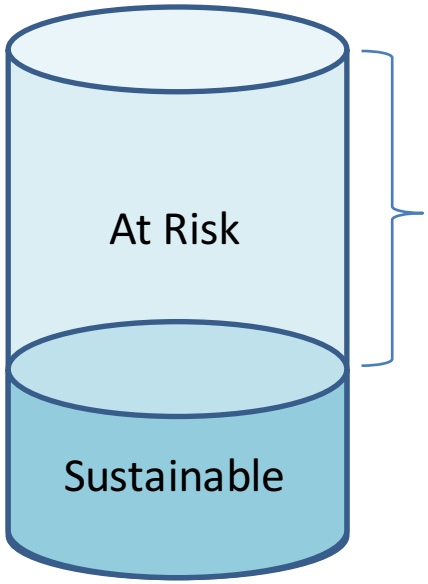


Culture / Community behavior – How can we minimize risk for valued open data?



Culture / Community behavior – How can we minimize risk for valued open data?

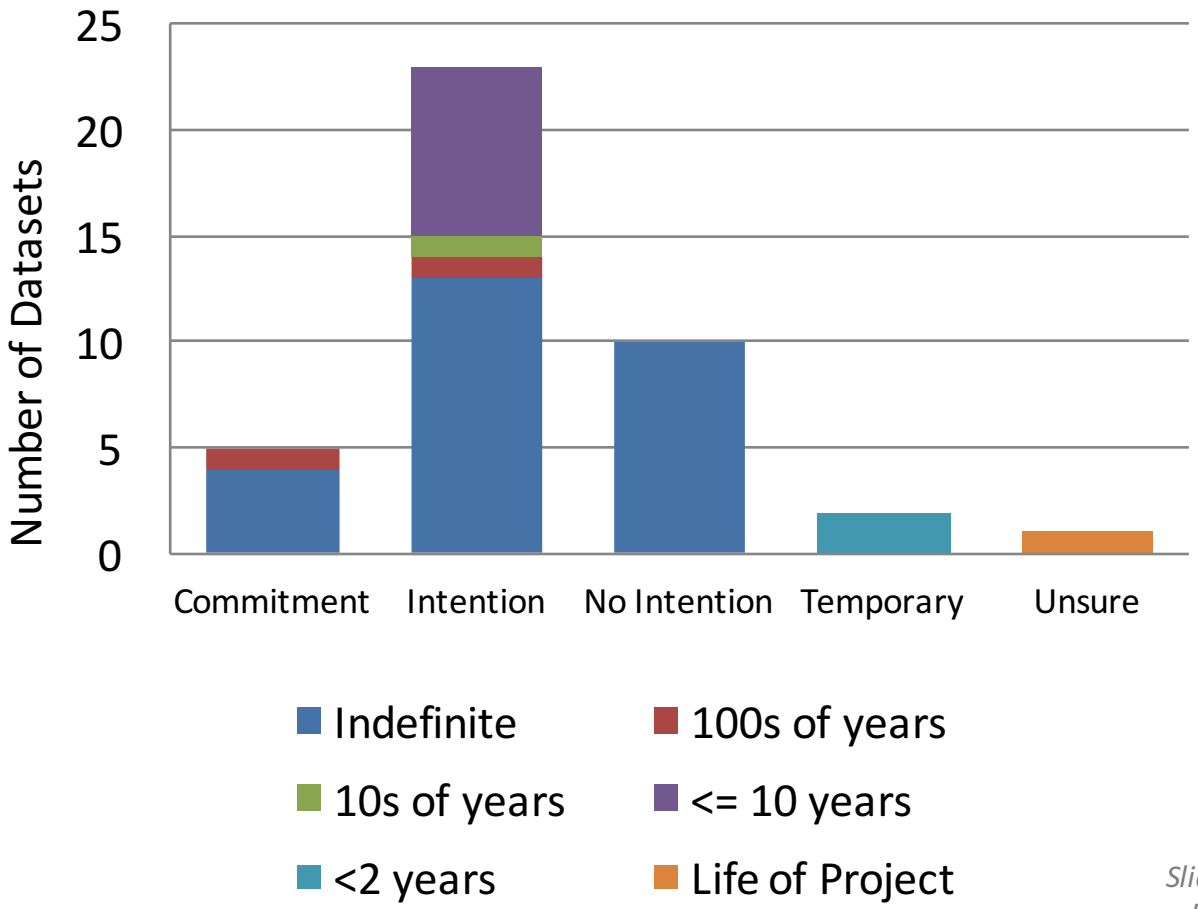
- How much public research data is at risk?
- U.S. National Institute of Health estimates for 2011 PubMed Central publications:
  - 12% of publication data sets deposited in recognized repositories, 88% of the data sets were invisible
  - Estimated approximately 200,000-235,000 invisible data sets generated NIH work published in 2011
  - 87% of the invisible data sets are new, 13% reflect data re-use
  - More than 50% of the datasets were derived from live human or animal subjects
- Community practice key to sustaining the data ecosystem



(Valued)
Sponsored
Research Data

Information from PLOS ONE
http://journals.plos.org/plosone/article?id=10.1371/journal.pone.0132735;
Graphic from http://www.colorado.edu/ibs/cupc/stewardship_gap/

Type of Commitment and Term of Value



Researchers believe their data have long-term value.

For datasets with >10 years of value:

- 2 out of 37 have a matching commitment
- ~1/4 have no explicit intention to preserve

Slide courtesy of Jeremy York from iPRES 16 Presentation. Stewardship Gap Project http://www.colorado.edu/ibs/cupc/stewardship_gap/

Many Stewardship gaps, many characterizations of "valued data"; Focused, strategic community practice can increase sustainability

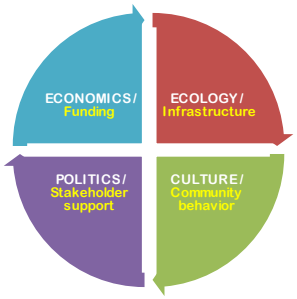
- **Resource Gaps:**
  - Insufficient funding
  - Insufficient staff
  - Insufficient information
  - Lack of facilities
- **Responsibility Gaps:**
  - Insufficient institutional and individual commitments
  - Differing expectations of researchers, stewards, and stakeholders
  - Insufficient stewardship and sustainability planning
  - Insufficient compliance with policy and regulation
- **Infrastructure gaps:**
  - Insufficient tools for management, use, discovery, preservation
  - Insufficient tools and frameworks for access and sharing

One approach does not fit all:

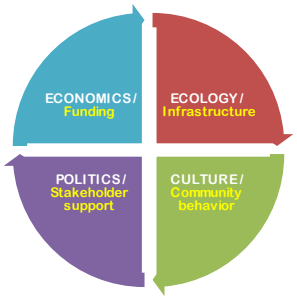
Differential policy, practice, resources, education/training, etc. can be used strategically to address various gaps

	Type of Gap			
Type of Value				

Information from the Stewardship Gap Project, http://www.colorado.edu/ibs/cupc/stewardship_gap/



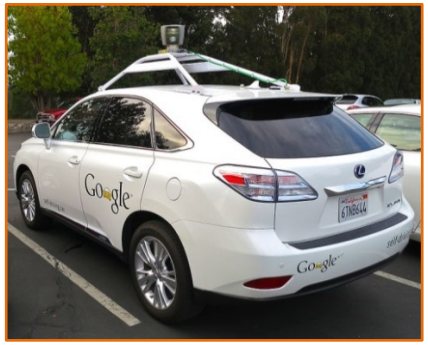
Politics / stakeholder support -- How to maximize benefits of data for the public good?



Politics / stakeholder support -- How to maximize benefits of data for the public good?

Internet of Things (IoT): Enabling environment or Lord of the Flies? How should the IoT be managed / organized?

- Who develops its "laws"?
- Who enforces them?
- Can you opt out?



Who is accountable when your self-driving car hits someone?

Which decisions should be made by technology?

When does your privacy matter more than the needs of others?

Does your computer know good from evil?

Rensselaer

RDA
RESEARCH DATA ALLIANCE

Wikimedia: Self-driving car Image courtesy of Steve Jurvetson, Mariordo; HAL9000 image from https://www.flickr.com/photos/zanotti/312159382; Robot image from http://hereandnow.wbur.org/2014/10/07/artificial-intelligence-strickland, iRobot, 20th Century Fox

What does Governance Mean for the IoT?

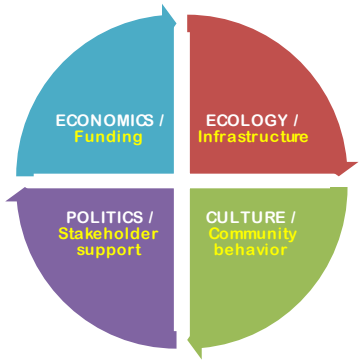
Adapting the World Governance Index (based on the UN Millennium Declaration), key governance themes span

Peace and Security	→	IoT Security, Trust, Safety, Crime
Democracy and Rule of Law	→	Legal framework for determining appropriate and inappropriate behavior, responsibility, accountability
Human Rights and Participation	→	IoT "Bill of Rights"? – Right to Privacy, Right to control information, Right to opt out, etc. Framework for promoting "equality" and penalizing "discrimination"
Sustainable development	→	Architectures, standards, policy, infrastructure, etc. to promote evolutionary and sustainable growth
Human development	→	Digital ethics, use of technology to advance / actualize its participants and contribute to well-being

IoT "Future Work" – Academic underpinnings and public development of governance, policy and social structures

- Is the IoT a society?
  - Who are its citizens? What are their rights? What is its ethnography? Will it be possible to live outside the IoT?
  - What should its ethical code be? What is the "common good"? Do we need "artificial ethics" in conjunction with artificial intelligence?
  - How do we implement and enforce social and governance structures for communities of devices, humans, systems, organizations, groups, hybrids? Does your toaster get a vote?

Social behavior begins with the individual: How you can help build a sustainable data ecosystem



- **ECOLOGY** / Infrastructure
  - Contribute to the development / adoption of data infrastructure for your problem/community and share it with others
  - Make your data accessible (as appropriate) by curating it and ingesting it into a publicly accessible repository
  - Create a data management plan that realistically describes what's needed throughout the entire data life cycle
- **ECONOMICS** / Funding
  - Budget realistically for the costs of data stewardship and preservation
  - Make data stewardship and preservation a fiscal priority for your project, institution or organization
- **CULTURE** / Community behavior
  - Contribute to or create a local / community culture of data sharing
  - Cite and publish your data when you write about your results.
  - Work with your professional societies and conferences to include "data sessions" and publications *(idea from Sibel Adali)*
- **POLITICS** / Stakeholder support
  - Make the case to stakeholders that data infrastructure is critical and a priority to ensure the accessibility of the data that drives innovation
  - Create / adopt / support policy and practice that enables the development and continued maintenance of sustainable stewardship, data sharing, and broad access