



Rensselaer



Data Policy for Open Science

Mark A. Parsons

0000-0002-7723-0950

Secretary General

Research Data Alliance

e-Infrastructure Reflection Group

Riga, Latvia

3 June 2015



Unless otherwise noted, the slides in this presentation are licensed by Mark A. Parsons under a Creative Commons Attribution-Share Alike 3.0 License

The first purpose of data policy should be to serve the objectives of the organization that sponsored the data collection.

Research Data Alliance



Vision

Researchers and innovators openly share data across technologies, disciplines, and countries to address the grand challenges of society.

Data policy should minimally address...



1. Data **sharing and access** requirements
2. Data **preservation and stewardship** requirements *and roles* (potentially including standards, documentation, protocols, etc.)
3. All the issues listed under data **security** (privacy, protection, legal issues, confidentiality, IPR, ownership), *as necessary for the nature and objectives of the data.*

... and encourage and reinforce appropriate norms of scientific behavior around data creation and use.

Preservation and Access

Two Peas in a Pod



- Scientific Data Stewardship:
 - “**preservation and responsive supply** of reliable and comprehensive data, products, and information for use in building new knowledge to...”
 - USGCRP, 1998
 - “the long-term preservation of the scientific integrity, monitoring and improving the quality, and the **extraction of further knowledge** from the data”
 - H. Diamond et al., NOAA/NESDIS, 2003
 - “Data stewardship encompasses all activities that preserve and **improve** the information content, **accessibility, and usability** of data and metadata.”
 - National Academy of Sciences, 2007

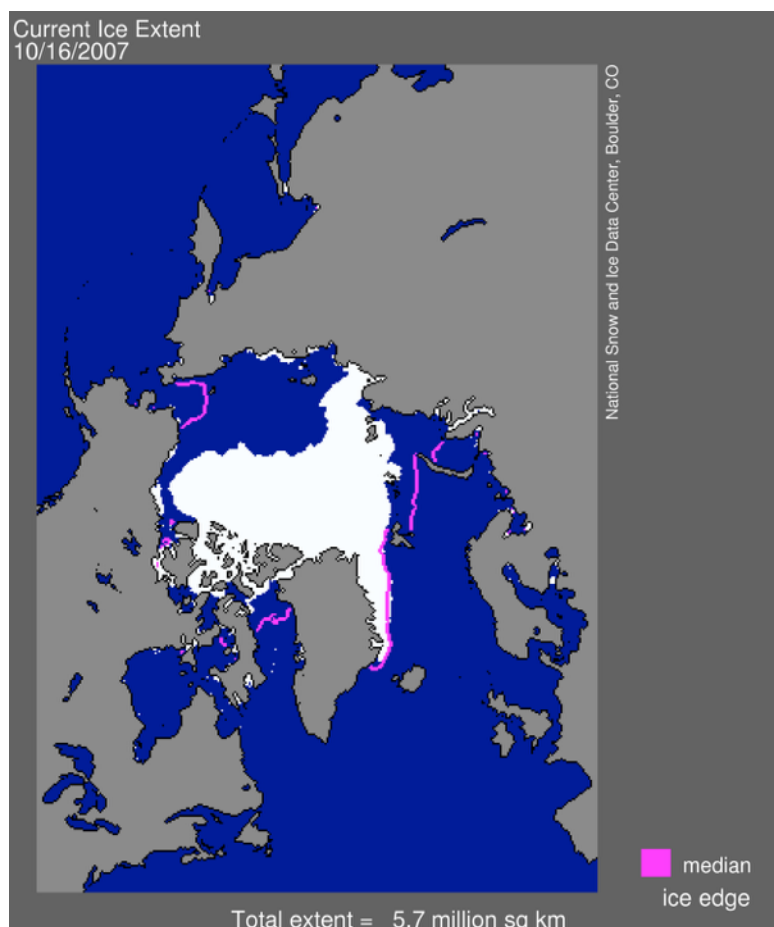


Access. What is it?

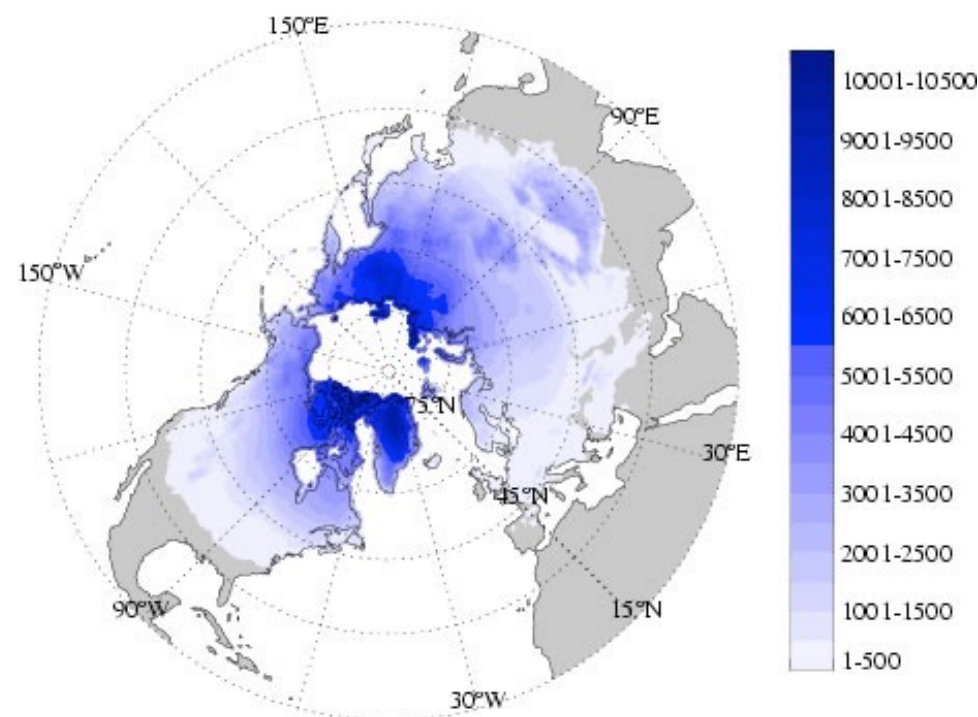
- Preservation requirements are well defined in the Open Archive Information System (OAIS) Reference Model, but
- No similar model for access requirements
- Not even a common definition of “access” and what restricts it
- Unique access requirements for
 - bio-medical, social science, humanities data
 - non-digital collections (physical samples, specimens, historical collections, etc.)
 - more...

What are the Data

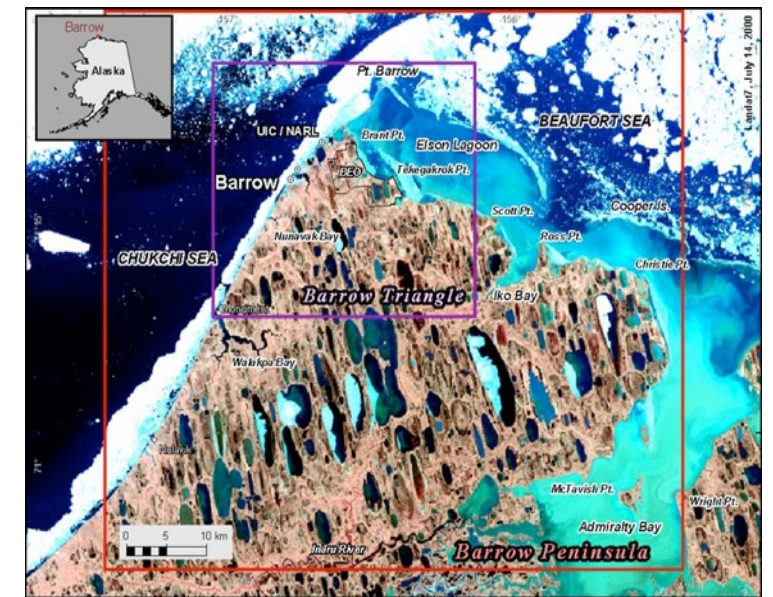
- National Science Board 2005:
 - Reference collections
 - Community or Resource collections
 - Research collections



Fetterer and Knowles. 2004. Sea Ice Index.
nsidc.org/data/seaice_index/



Zhang, T. et al. 2005. Northern Hemisphere EASE-Grid Annual Freezing and Thawing Indices, 1901 - 2002.
nsidc.org/data/ggd649.html



Manley, W. F. et al. 2005. Reduced-Resolution Radar Imagery, Digital Elevation Models, and Related GIS Layers for Barrow, Alaska, USA.
nsidc.org/data/arc3303.html

“A biologist would rather
share their toothbrush
than share their data”
—Carole Goble

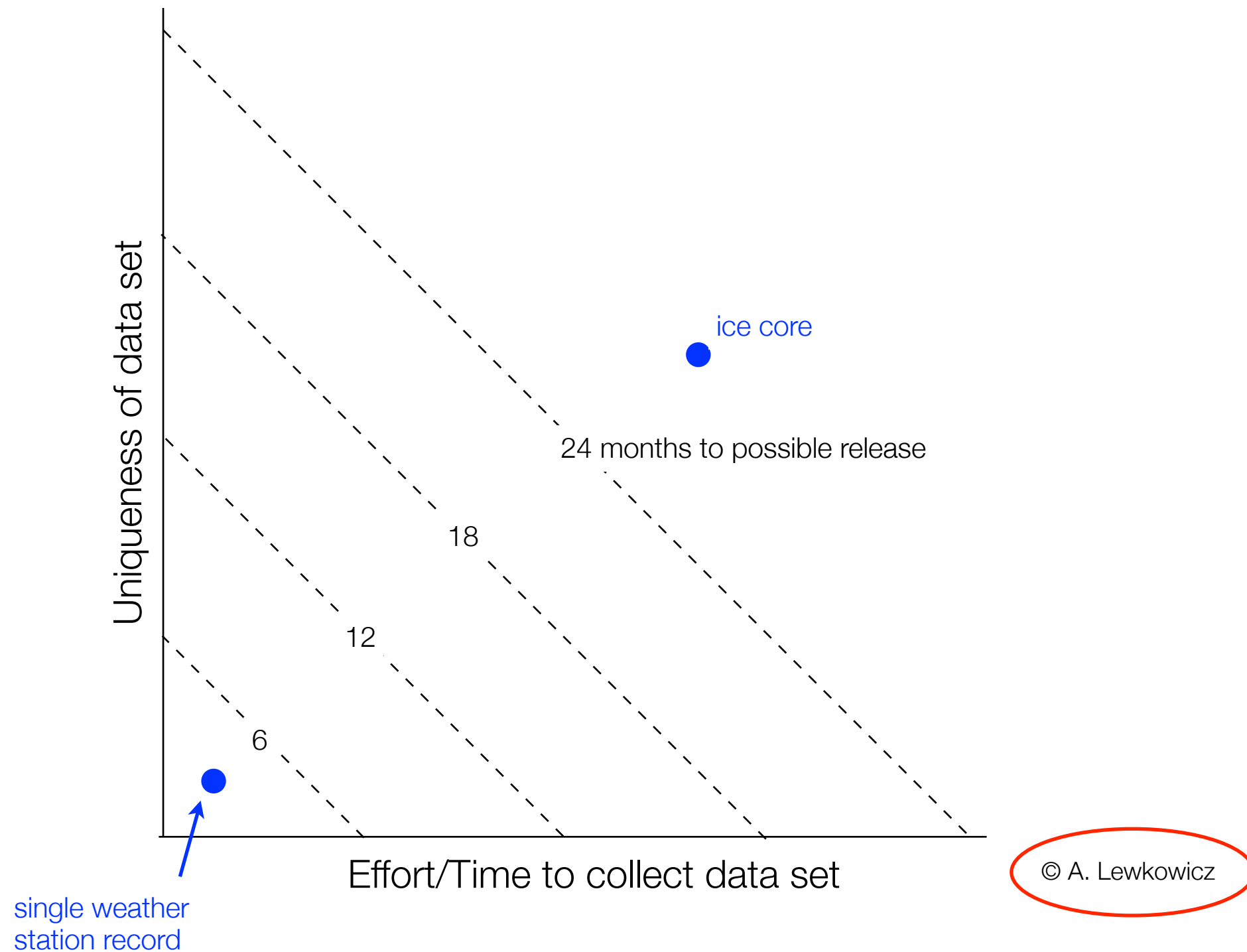




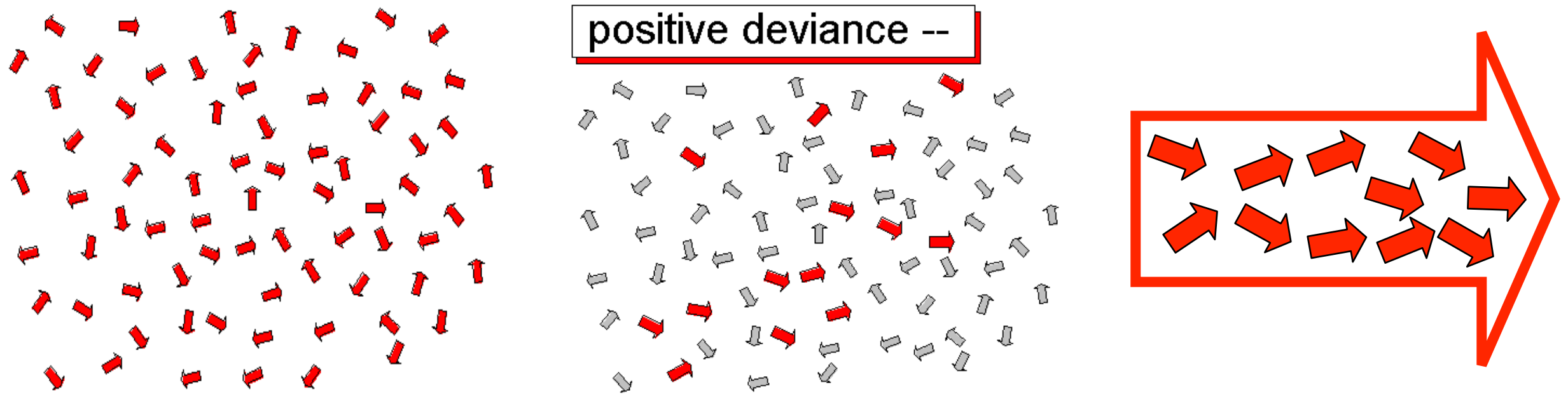
Identity—The “Real Polar Man”



A model for data release



Leadership Model: Positive Deviance



Positive deviance says that if you want to create change, you must scale it down to the lowest level of granularity and look for people within the social system who are already manifesting the desired future state. Take only the arrows that are already pointing toward the way you want to go, and ignore the others. Identify and differentiate those people who are headed in the right direction. Give them visibility and resources. Bring them together. Aggregate them. *Barbara Waugh*

What works



- Clear, open policy with timelines and backed up by active and engaged program managers.
- A ready, easy, and funded mechanism for data deposit.
- “Data Wranglers” —professional data managers responsible for identifying data for an archive and then encouraging and assisting data providers to publish their data.
- “Naming and shaming.” When there is public indication of who is sharing data and who is not, some people are quick to respond.
- Demonstrated value of data repositories. A clear, obvious value to the submitter for making their data available.
- Providing information to data providers about how their data are being used and by whom.
- Fair and formal credit and attribution to data providers. It is unclear how great an incentive attribution is, but if credit is not provided, it will dissuade many from contributing their data.

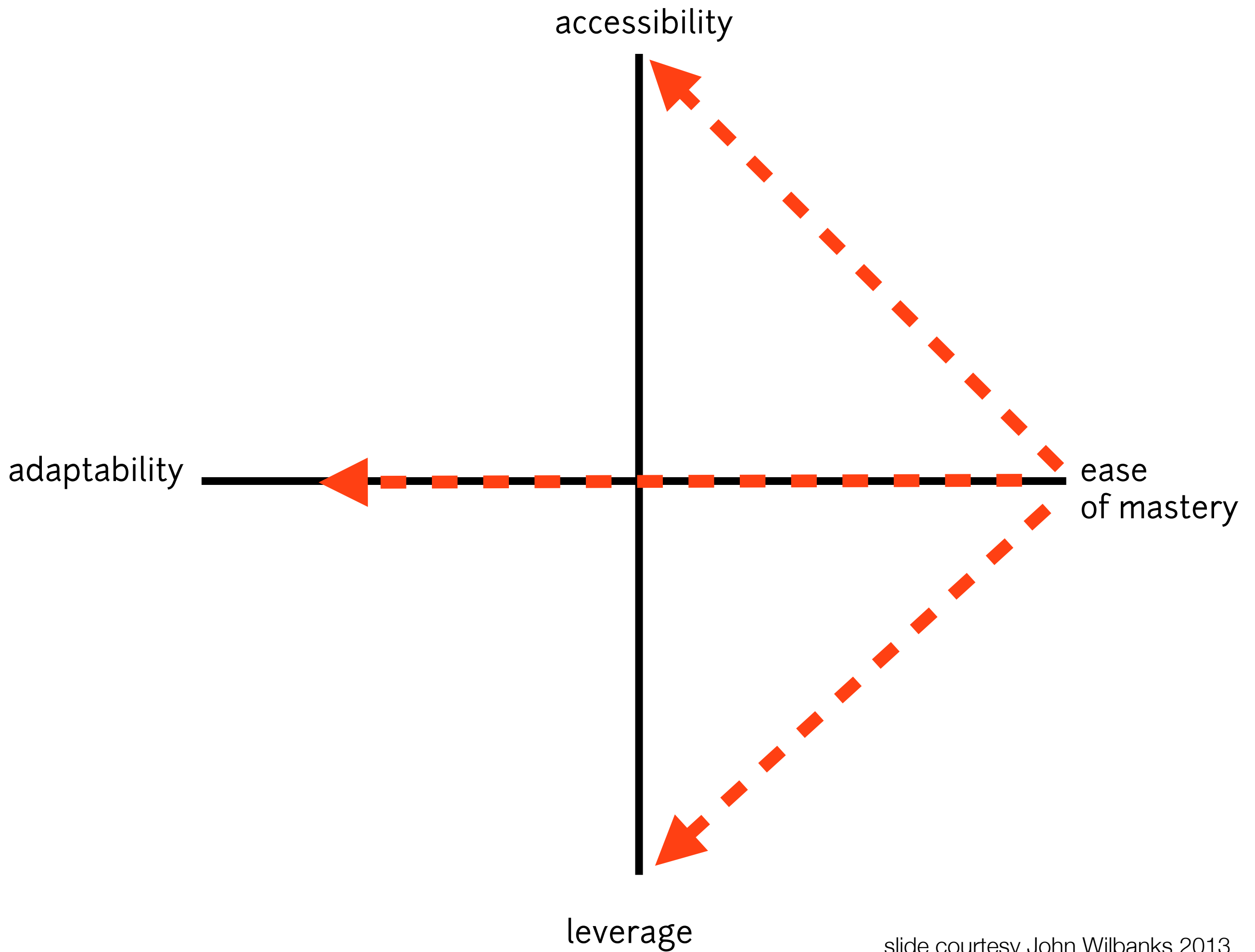
What needs work

- Harmonization of policy around the principles of open access and ethical use.
- Policy clarification of stewardship and “ownership” roles.
- Social science research on incentives for sharing and related issues of trust and scientific identity..
- Formal, recognized graduate and training programs in informatics and scientific data stewardship.
- Basic training on data management as a routine, required component of the graduate-level science curriculum.
- Identifying and recognizing appropriate credit mechanisms for appropriate work.
- Lowering the technical, social, and legal barriers to sharing

The generative value of data

- Generative value per Jonathan Zittrain (2008) as interpreted and extended to data by John Wilbanks:

“the capacity to produce **unanticipated change** through **unfiltered** contributions from **broad and varied audiences**.” —J. Zittrain
- Data become more generative by being more adaptable, more easily mastered, more accessible, and more connected and influential.
- Not net *present* value but net *potential* value.



Research Data Alliance



Vision

Researchers and innovators openly share data across technologies, disciplines, and countries to address the grand challenges of society.

Mission

RDA builds the **social and technical bridges** that enable open sharing of data.

“Create - Adopt - Use”



Adopted Policy



Systems
Interoperability



Common Types,
Standards, Metadata



Sustainable Economics



Adopted Community
Practice



Training, Education,
Workforce

RDA as a Policy Test Bed



- RDA is not a policy organisation, but it can help implement policy
 - Evolving data management plans to ongoing planning
 - Defining “register your data”
 - Answering the question of which metadata standard to use
 - Clarifying what is a certified repository
 - Sorting out the roles of institutional and domain repositories
 - Ensuring workflows are consistent across systems
 - Figuring out middleware/brokering governance models and interconnection between registries
 - Facilitating shared terminologies (e.g. biodiversity)
 - ...

Summary suggestions

- Mind your preservation and access—your stewardship
- Clarify and credit roles
- Promote and empower the champions—those who add generative value.
- Look for consensus and emergent norms from the data science community
- Iterate