(Realizing the dream of open data in the “long tail” of public agricultural research…)

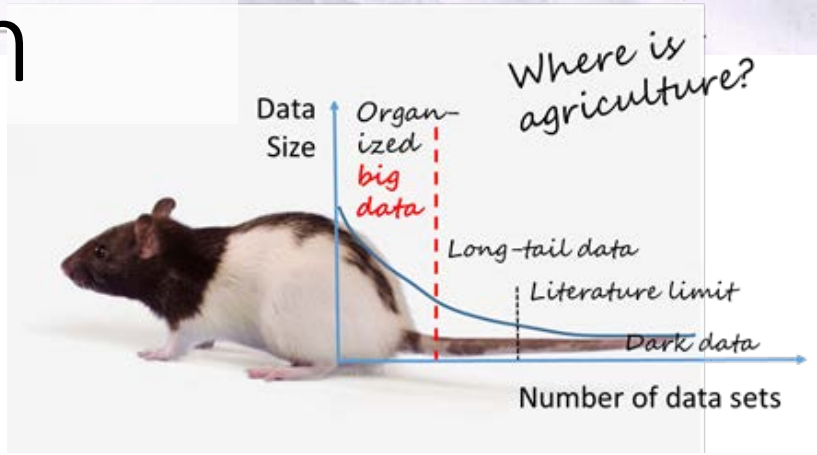
The response of the Contrarian Curmudgeon:

"You can have my data when you pry it out of my cold, dead hard drive"

Enabling Open-source Data Networks in Public Agricultural Research

Sylvie M. Brouder, Purdue Univ.

Interagency Working Group on Biological Data Sharing Workshop, Institute for Bioscience and Biotechnology Research, Rockville, MD, June 12, 2019



Enabling Open-source Data Networks in Public Agricultural Research

CAST Commentary QTA2019-1

Presented by

Sylvie Brouder, Ph.D.

Purdue University, Department of Agronomy

March 2019

#CASTreports2019

Task Force Members

Authors:

Sylvie Brouder (Chair)
Department of Agronomy, Purdue University
West Lafayette, Indiana

Alison Eagle
Sustainable Agriculture, Ecosystems Program,
Environmental Defense Fund, Raleigh, North Carolina

Naomi K. Fukagawa
Beltsville Human Nutrition Research Center, USDA–ARS
Beltsville, Maryland

John McNamara
Teaching Academy, Washington State University
Pullman (retired)

Seth Murray
Department of Soil and Crop Sciences
Texas A&M University, College Station

Cynthia Parr
National Agricultural Library, USDA–ARS
Beltsville, Maryland

Nicolas Tremblay
Agriculture and Agri-Food Canada
St-Jean-sur-Richelieu, Canada

Reviewers:

Marianne Stowell Bracke
Harriet Cheney Cowles Library, Whitworth University
Spokane, Washington

Paul Fixen
International Plant Nutrition Institute
Brookings, South Dakota (retired)

Jeffrey Volenec
Department of Agronomy, Purdue University,
West Lafayette, Indiana

CAST Liaison:

Drew Lyon
Extension and Research, Washington State University
Pullman

CAST®

What motivates this commentary?

2012 Drought, Purdue Ag. Comm.

Grand Challenge Questions

Next generation problem solving in agriculture:

- "Big" science
- Transdisciplinary research linking disciplines and **DATA**

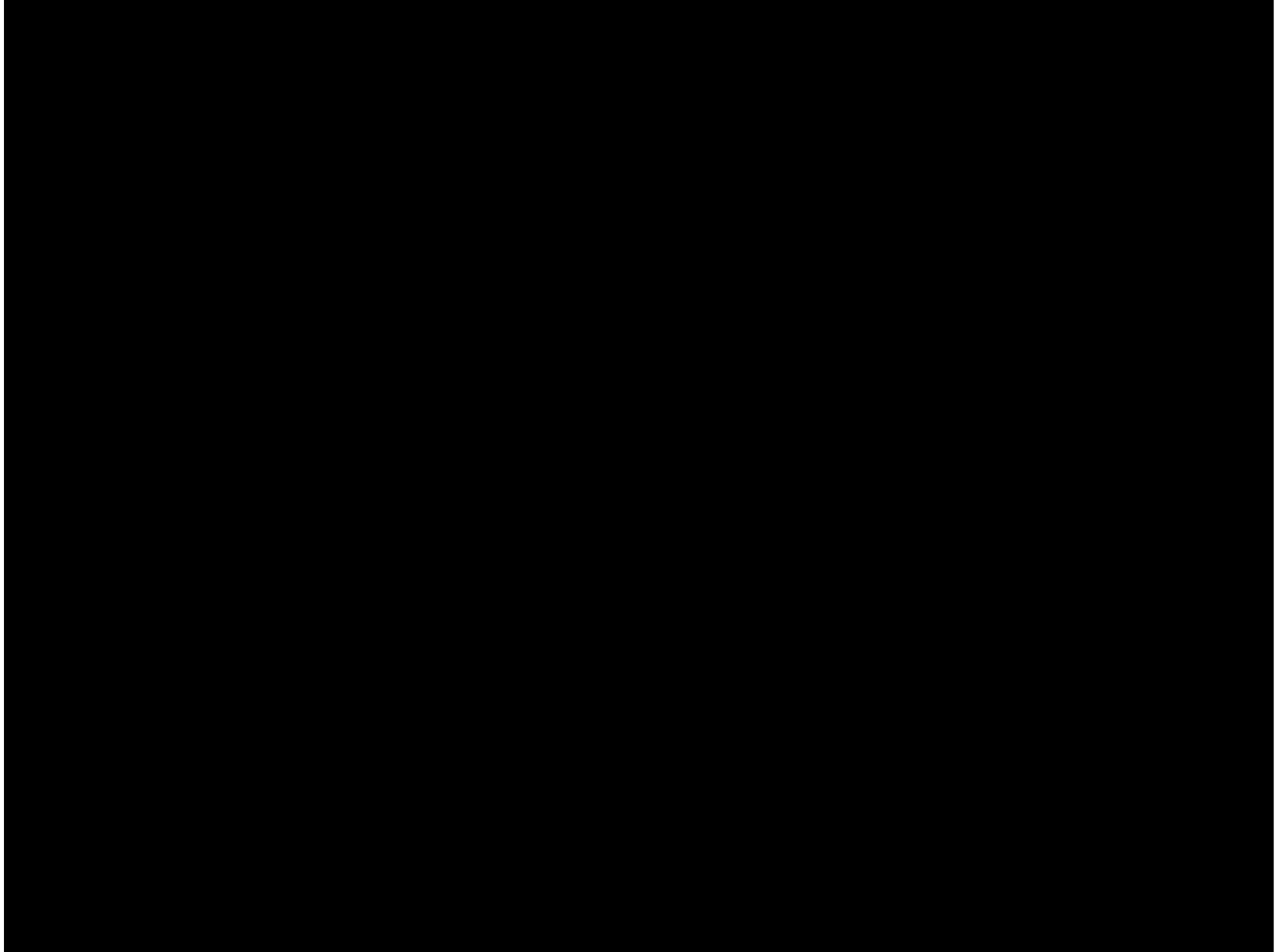
Major barrier to making better agricultural decisions:

- Lack of data sharing and data accessibility

Publically funded science belongs to the public and should be available for their use

Why data?: Professional “hats” → Director, Purdue Univ. Core Facility (1997 – present)

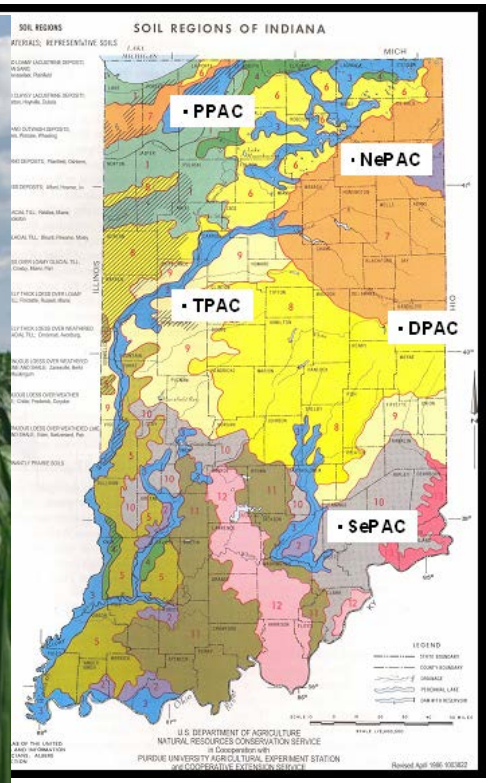
*Purdue’s Water Quality Field Station: A Core Facility for M x E research (**focus on N**) and bleak landscape without a DMP safety net…*



Why data? Other professional “hat”…

The Tri-State Recommendations

… and my 5-yr K experiment now entering its 19th year but still not enough data…



“Enabling Open-source Data Networks in Public Agricultural Research”

Goal

Advance the conversation among agricultural science partners to create a system conducive to data sharing and the team science to address grand-challenge questions in food systems

Purpose

Document the need for and anticipated benefits of developing data standards, incentivizing data sharing and building data-sharing infrastructure to meet the varied needs of agricultural researchers

Topics for today…

- Historical perspectives on data and the process of conducting agricultural research
  - Small science, the small-science environment, outcomes, and limitations
  - Why we currently don't share data
- Vision for the new data ecosystem
  - FAIR data, repositories, "knowledgebases"
- Four immediate imperatives
- Four critical strategies to facilitating data sharing
- Partnering for success
- The business model
  - Open data is not free ~ suggestions for who pays and how

What is "small" vs "big" science?

History Lesson: Why did we do “small” agricultural research?

Research tools and technologies limited data size and accessibility

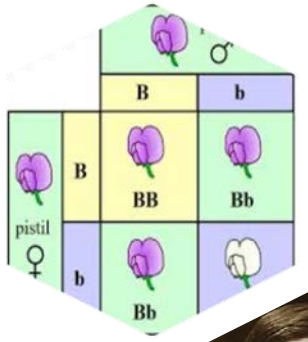
Rudimentary data collection and data-management tools

More limited capabilities of field and laboratory analytical tools

No or limited access to computational power

Data storage an individual decision and enterprise

Lots to be gained and was gained from “incremental” science



Gregor Mendel (1800s)

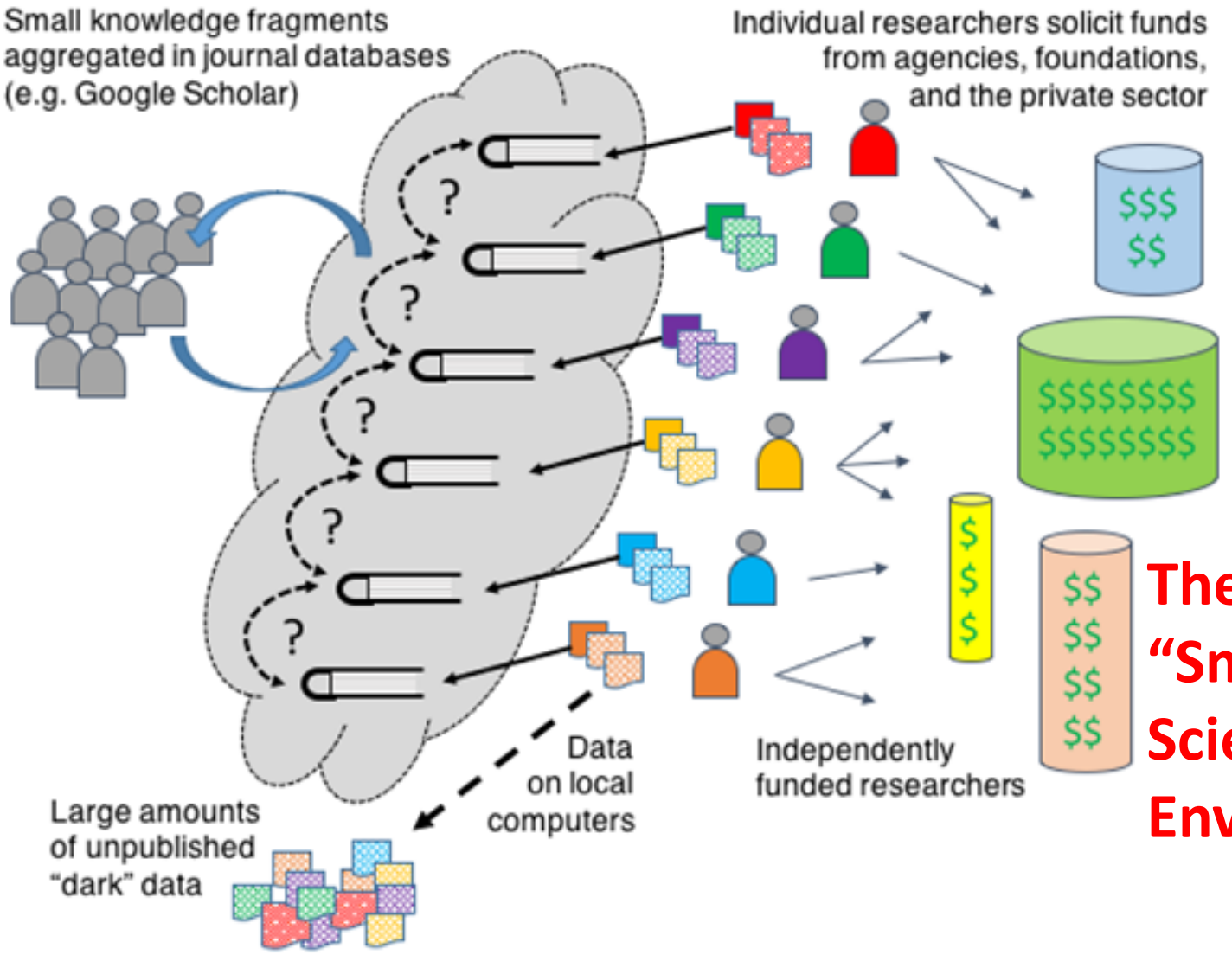
+ Ronald Fisher (1900s)

+ Gebisa Ejeta (now) …

How have we pursued agricultural research questions in the past?

Environmental Outcomes:

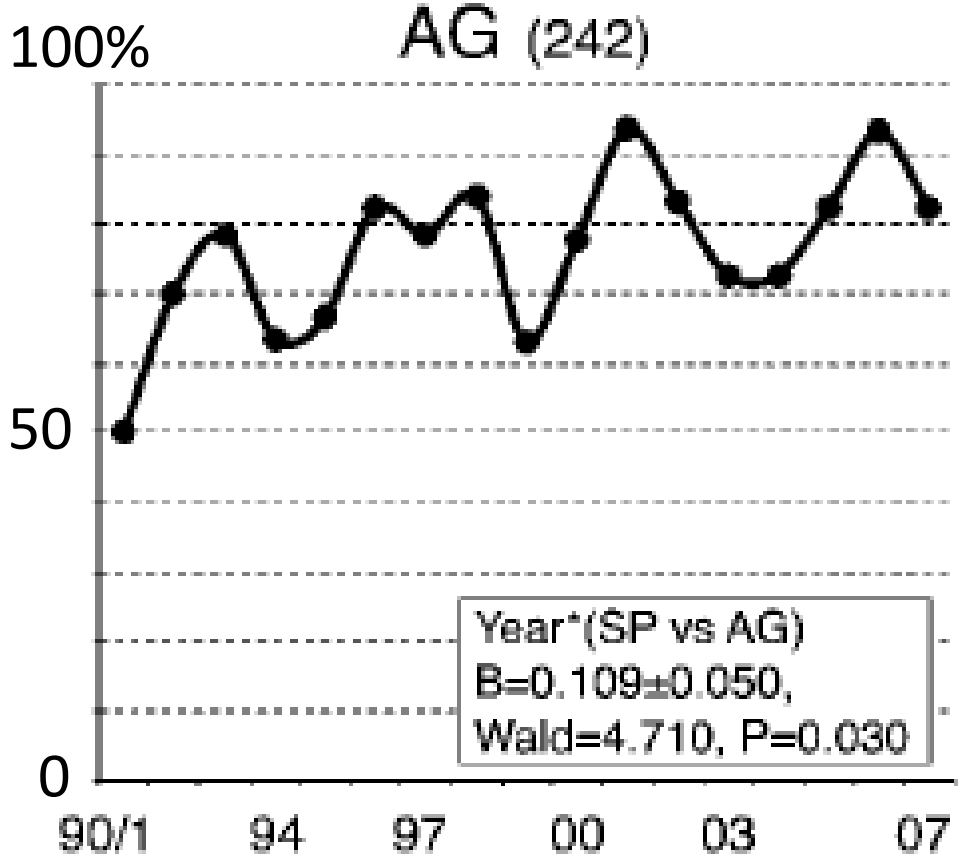
- Article data not "open" (subscription paywalls)
- No explicit linkages among knowledge fragments
- Research not transdisciplinary
- Lots of inaccessible and, often, lost data
- Data accessed via journals partial and potentially biased



The "Small" Science Environment

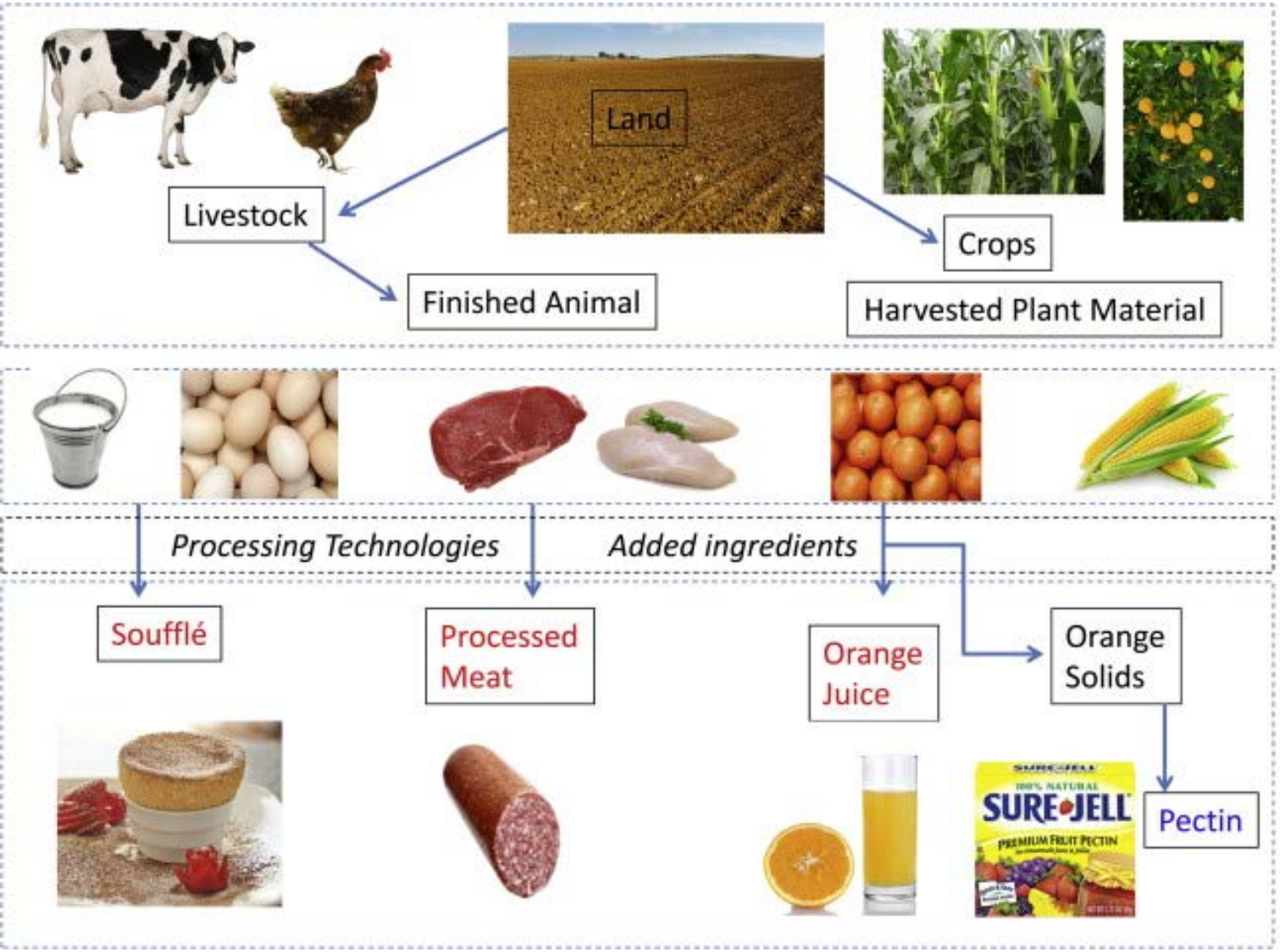
Policy and recommendations need to reflect all results; in journals, the null result is an endangered species!

- Review of 4,600 papers – all disciplines – 242 papers in agriculture
- Evaluated them for positive support of a "tested hypothesis"
  - Found 22% increase from 1990 to 2007
  - Asia > U.S. > Europe
- Hypothesize that:
  - Research becoming less pioneering and/or
  - Objectivity in research/publication is decreasing…



Negative results are disappearing from most disciplines and countries (Fanelli et al. 2012)

Achieving desired outcomes to complex problems with science: Single panacea vs. articulating strategies toward solutions…

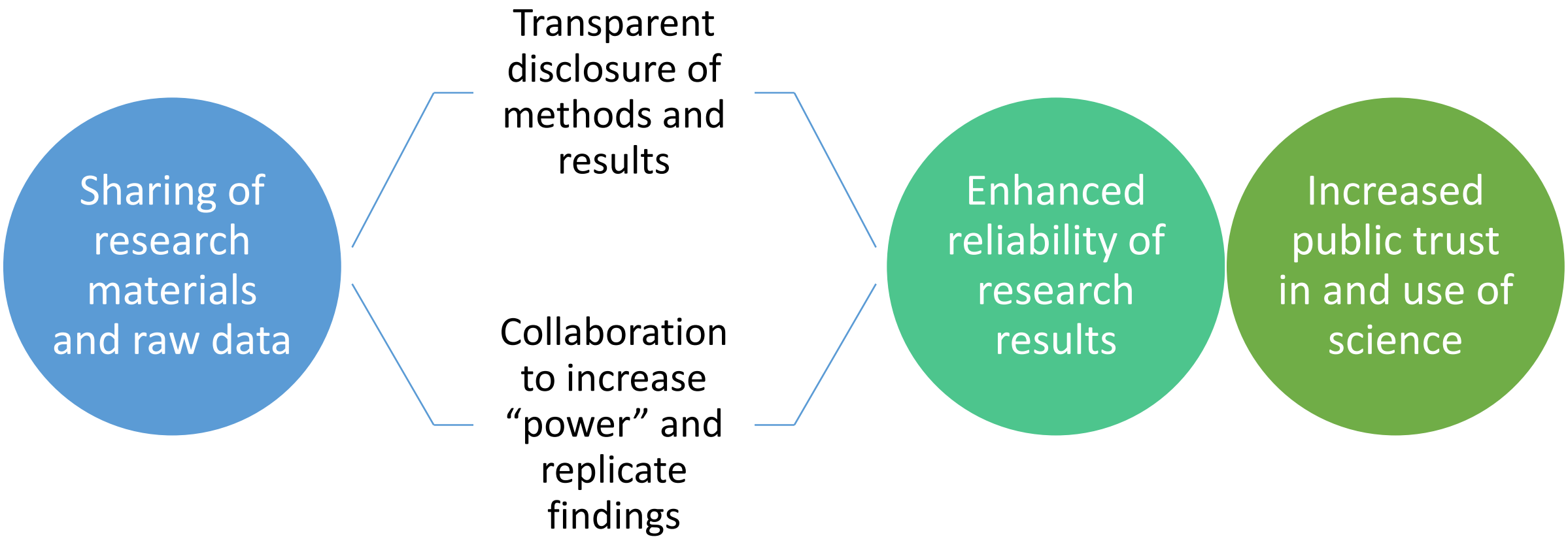


Small science can't address system tensions and trade-offs to weight a ***plurality of possible outcomes for multiple objectives***

"Abandon reductive approaches that imply one solution, and instead acknowledge the necessity of a non-reductionist approach that is integrative over different levels of detail."

A complex systems approach to address world challenges in food and agriculture (van Mil et al. 2014)

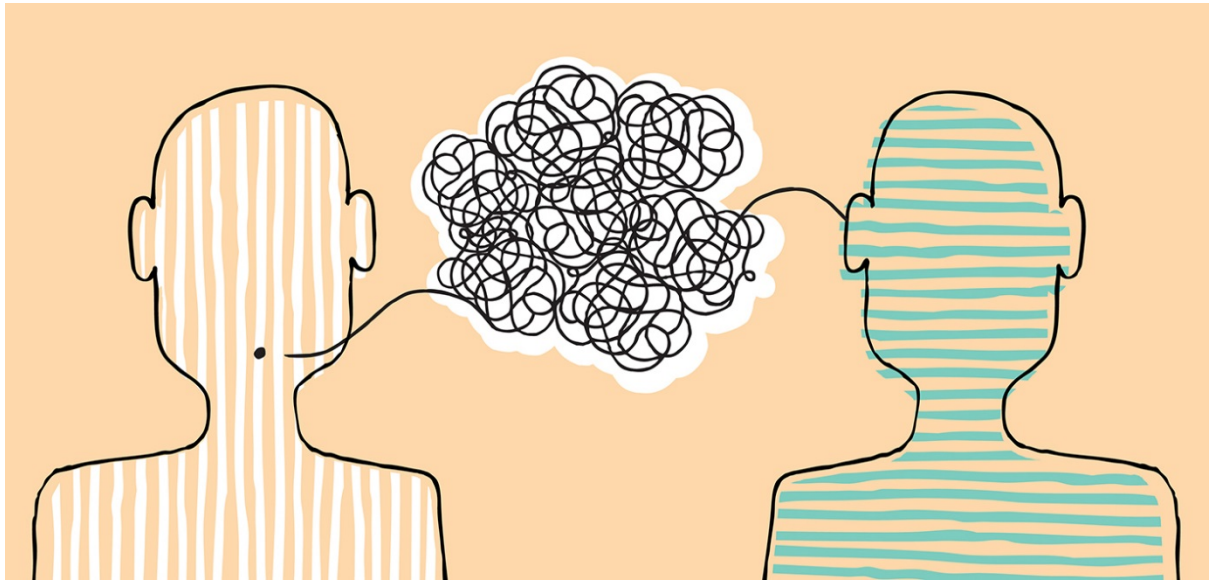
Other reasons to share: Increasing public trust in and use of science is founded on open data



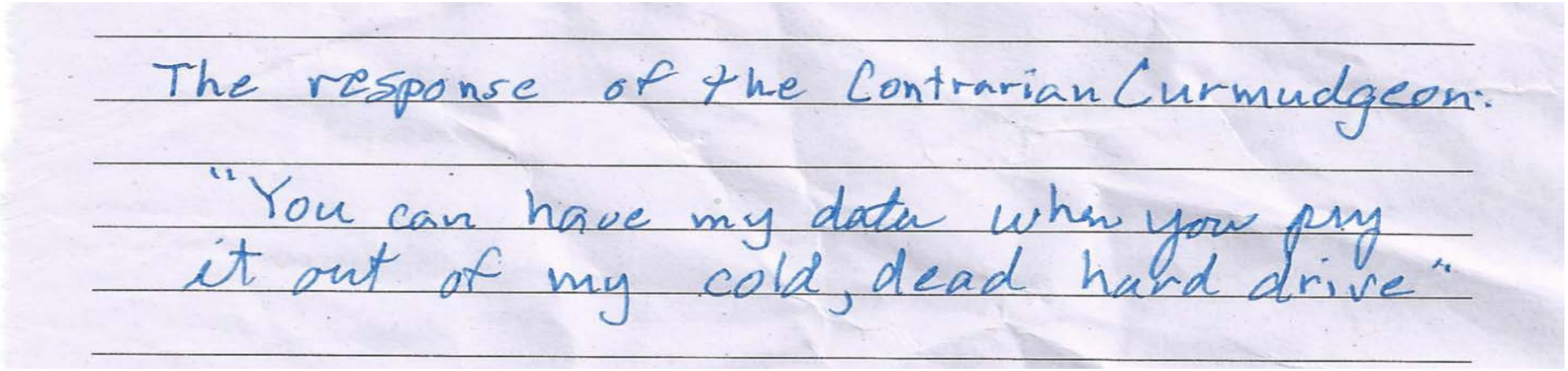
Make "big" data out of "small"

Why is so much data not shared and/or lost?

Multiple "vocabularies" among disciplines and domains

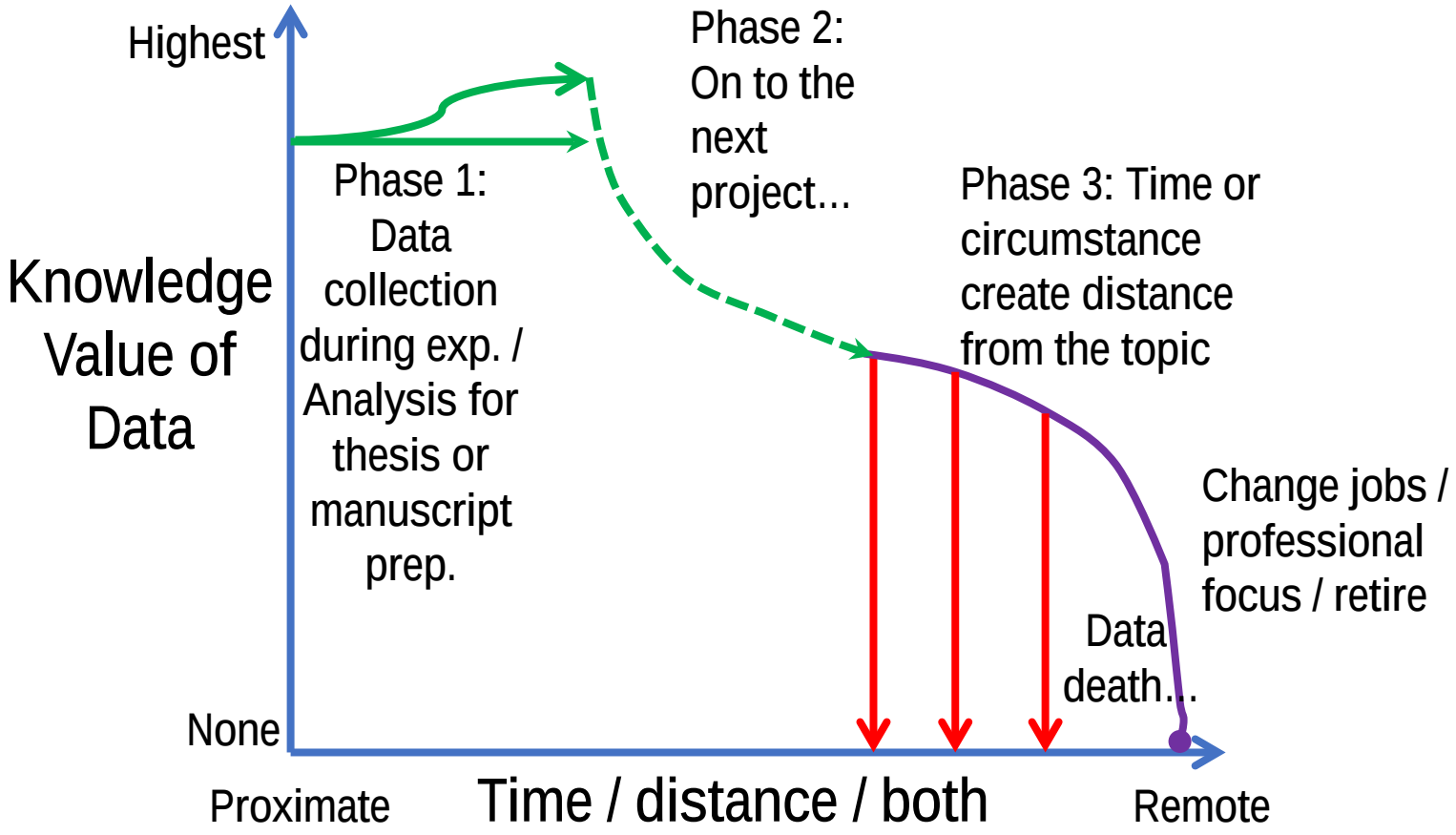


The "scoop" and other past perceptions of ownership…



Why is so much data not shared and/or lost?

Lack of proper, prior planning and low-barrier desktop tools ~ the effect of time on the "knowledge value" of data…



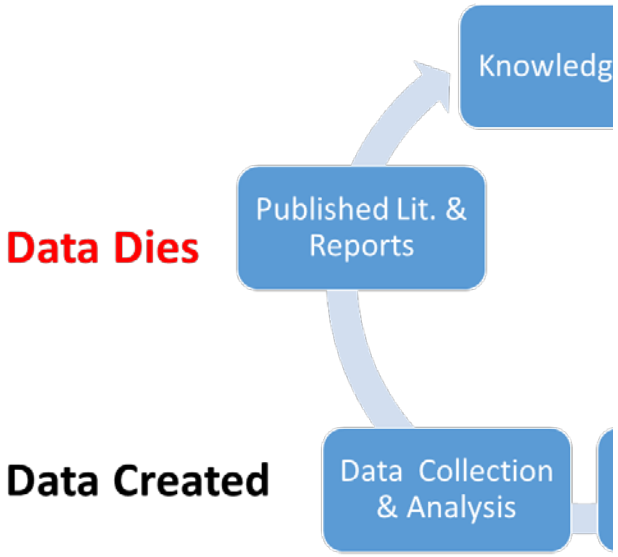
Modified from W. Michener, DataONE

Reality ~ data sharing is hard…

- It takes time and money
- Data literacy has not been part of undergraduate and/or graduate curricula
- Excel is low-barrier but it is ***inadequate*** to the task
- Once data are prepared for sharing, what do you do with it…?

FAIR data to facilitate data sharing and extend data lifecycles

Typical agricultural data



Agricultural data should be FAIR.
FAIR data are/can be...

Findable: described with a digital object identifier and rich metadata indexed in a searchable resource

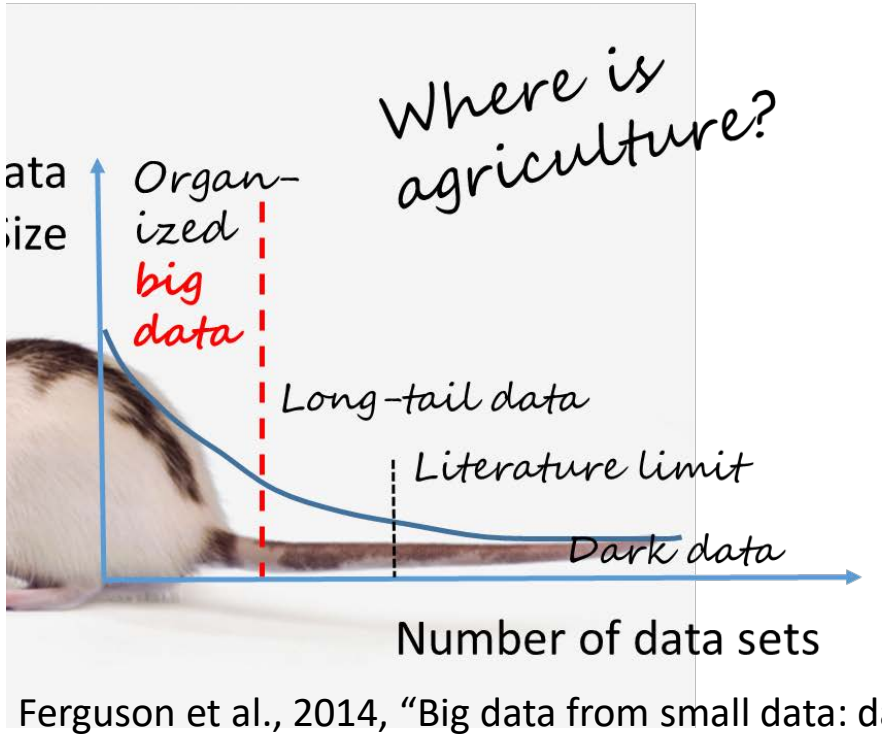
Accessible: retrieved using a standardized communication protocol (free, open, and universally implementable)

Interoperable: represented with a formal, shared, and broadly applicable language with FAIR vocabularies

Re-usable: richly described by a plurality of attributes (clear provenance, and usage license, and meets domain standards)

(Adapted from Wilkinson et al. 2016)

non-big data problem: Short
ong-tail data, and data lost to
nlightenment from medicine

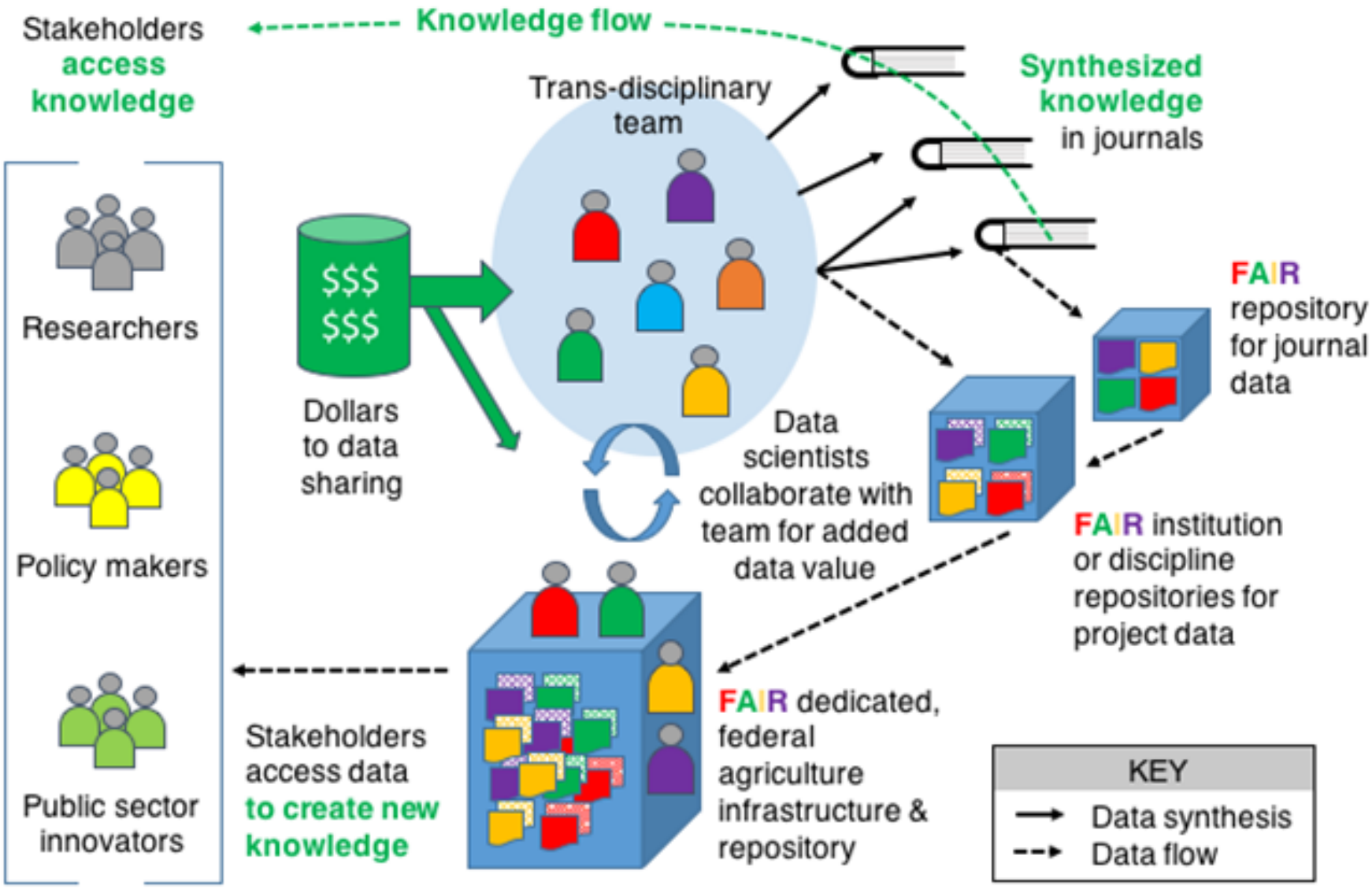


Ferguson et al., 2014, "Big data from small data: data
f neuroscience"

The data ecosystem for team science…

Key attributes:

- Data collection *a priori* anticipating reuse
- Article data is published; all project data are FAIR
- "Knowledgebases" with expert services enhance data collections and reuse
- Dedicated experts collaborate with teams and stakeholders for high-value data products ("fusions")
- Stakeholders access **knowledge** and data…



Examples of valuable but dark data in **Nutrient Mgmt. Research**

~ ***Recommendations must come from the "preponderance" of all evidence (not just the novel result that makes it to a journal…)***

Dark research data? (many IPNI proj.)

- **Orphaned data** ~ data collected but not used in experimental analysis (increasingly prevelant)
- **Null or failed studies** (reproving the null hypothesis) ~ no impact studies need to contribute to a "preponderance" of evidence
- **Confirmatory studies** ~ not novel so may not be publishable but still needed for preponderance of evidence

Dark non-research data (IPNI has long argued for their use in fert. recs.)

- Data from on-farm collaboratives and farmer-driven research efforts
- Data collected by farmers, CCAs, etc. in current management protocols (e.g. farm records)
- Monitoring data off equipment, etc.
- Other??

Most of these data are not well described thus "unavailable"…

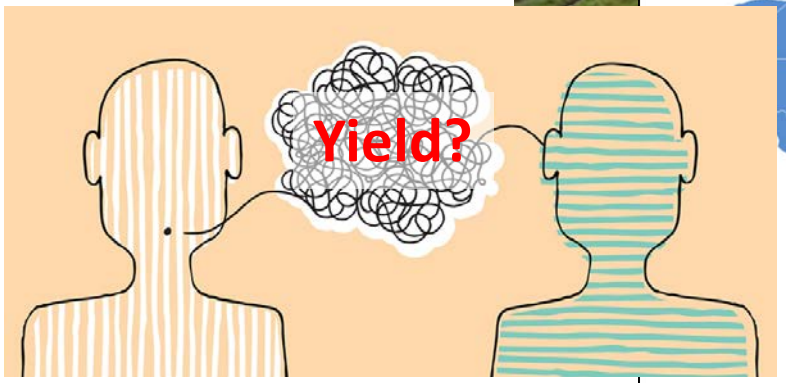
Facilitating data sharing to fully realize open access to public agricultural research

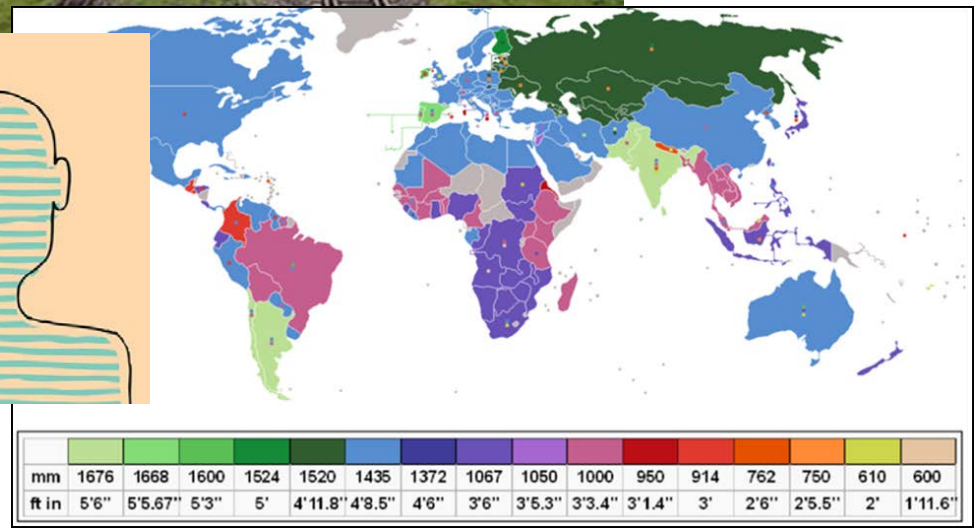
Imperatives

- **Development and implementation of "best practices" for data management in all federally funded projects**

Data and Metadata Standards

Necessary for getting from here to there…

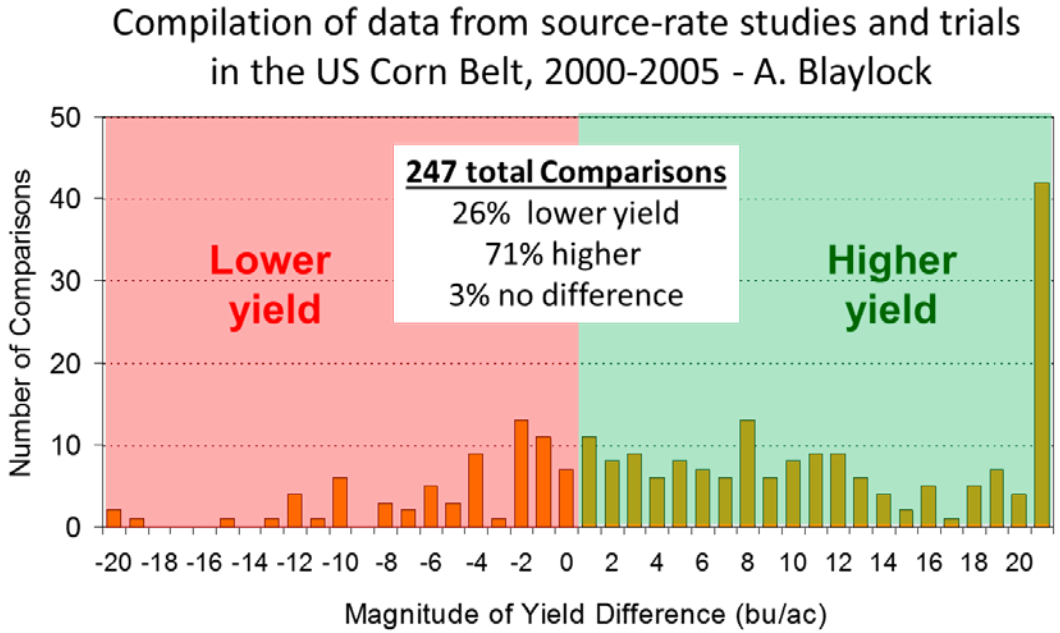




Facilitating data sharing to fully realize open access to public agricultural research

Imperatives

- Development and implementation of “best practices” in all federally funded projects
- **Incentives and mechanisms for making “grey and dark” data available**



Fertilizer Industry’s 4R Research Fund and Repository

4. Funding Schedule

The 4R Fund requires 10% of the funding be withheld until th
is submitted.

Date	Amount
year 1	$113,724
year 2	$106,903
year 3	$76,252
upon completion	$32,986
	(remaining 10%)

Facilitating data sharing to fully realize open access to public agricultural research

Imperatives

- Development and implementation of “best practices” in all federally funded projects
- Incentives and mechanisms for making “grey and dark” data available
- **Coordination among existing and emerging data initiatives, networks, and repositories**

Examples of initiatives, networks, repositories ~ illustrative not exhaustive!

Alliances, Coalitions, Networks		Repositories and Databases
The Research Data Alliance	DataONE	Ag Data Commons (ARS NAL)
The Ag Data Coalition	Cy Verse	Purdue Univ. Research Rep (PURR)
AgBioData	GODAN	Dryad
		Maize DB

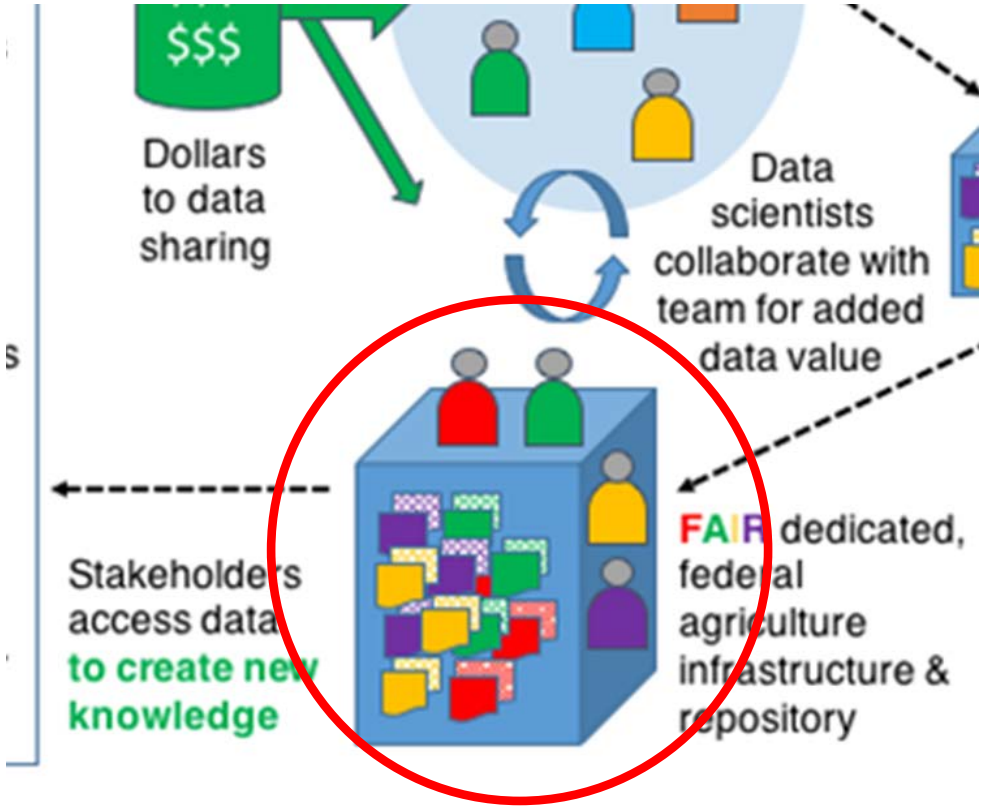
Sheer number poses a challenge to a coordinated landscape; each faces ***sustainability*** issues…

Facilitating data sharing to fully realize open access to public agricultural research

Imperatives

- Development and implementation of “best practices” in all federally funded projects
- Incentives and mechanisms for making “grey and dark” data available
- Coordination among existing and emerging data initiatives, networks, and repositories
- **Dedicated and sustainable infrastructure – hardware, software, and human resources – to curate, preserve, add value…**

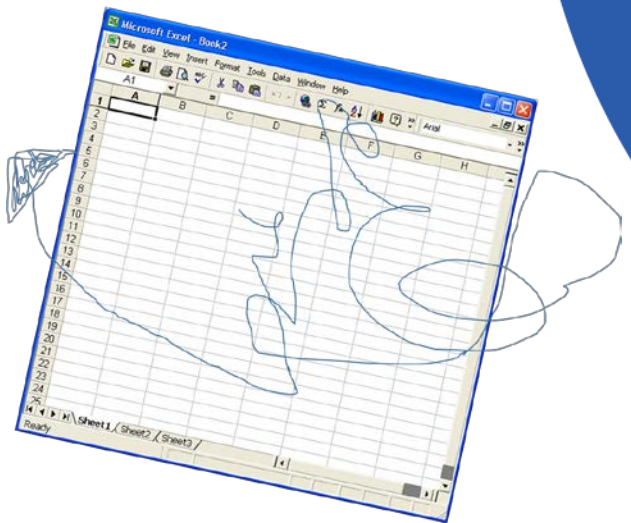
Critical need: A knowledgebase at the heart of the data ecosystem

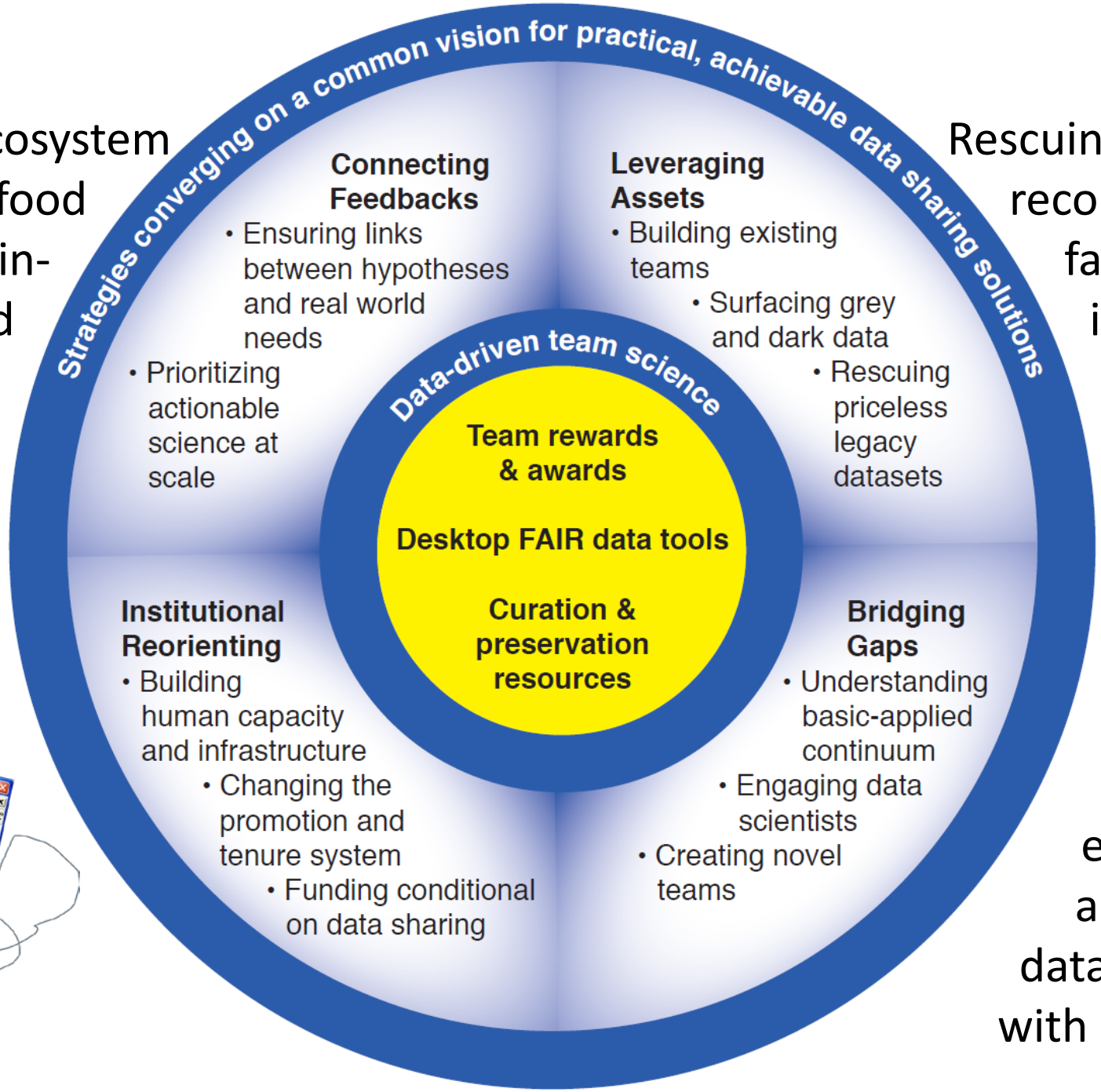


Strategies

Data to support ecosystem service valuation, food supply chain sustain-ability metrics, and user “Apps”

Curriculum for “digital natives”

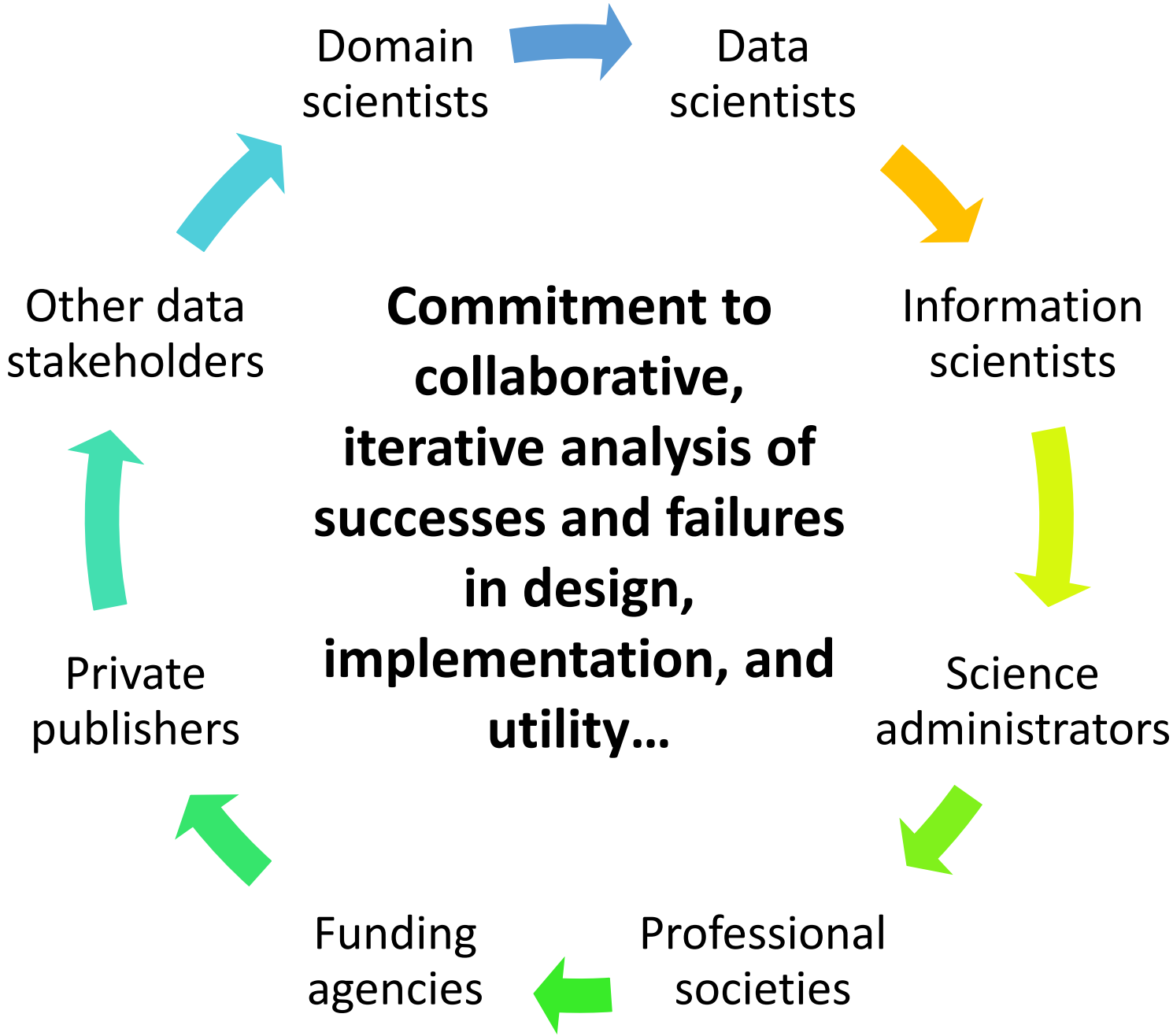




Rescuing long-term data records documenting farm management impacts on water quality (N, P)

Moving well beyond an agronomist simply asking an economist to play a cursory role or a data scientist to help with analytics or code

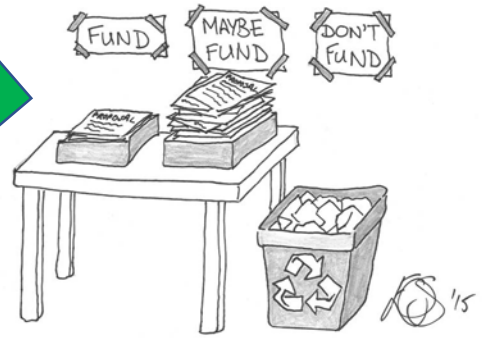
Partners


Domain scientists
Data scientists
Information scientists
Science administrators
Professional societies
Funding agencies
Private publishers
Other data stakeholders
Commitment to collaborative, iterative analysis of successes and failures in design, implementation, and utility…

Institutional reorienting & the eminence paradox: results of 7 million publicly-funded researchers behind paywalls…

JOE HOCKEY BRINGING HOME THE BACON

CONAN THE LIBRARIAN



Review panel categories.

WILEY-BLACKWELL

ELSEVIER

Taylor & Francis

SAGE Publishing

Springer

The Big 5 = > 50% of papers published in 2013 (Lariviere et al., 2015)

Open access ≠ cost free: Who pays for the ecosystem?

Open data infrastructure = public good

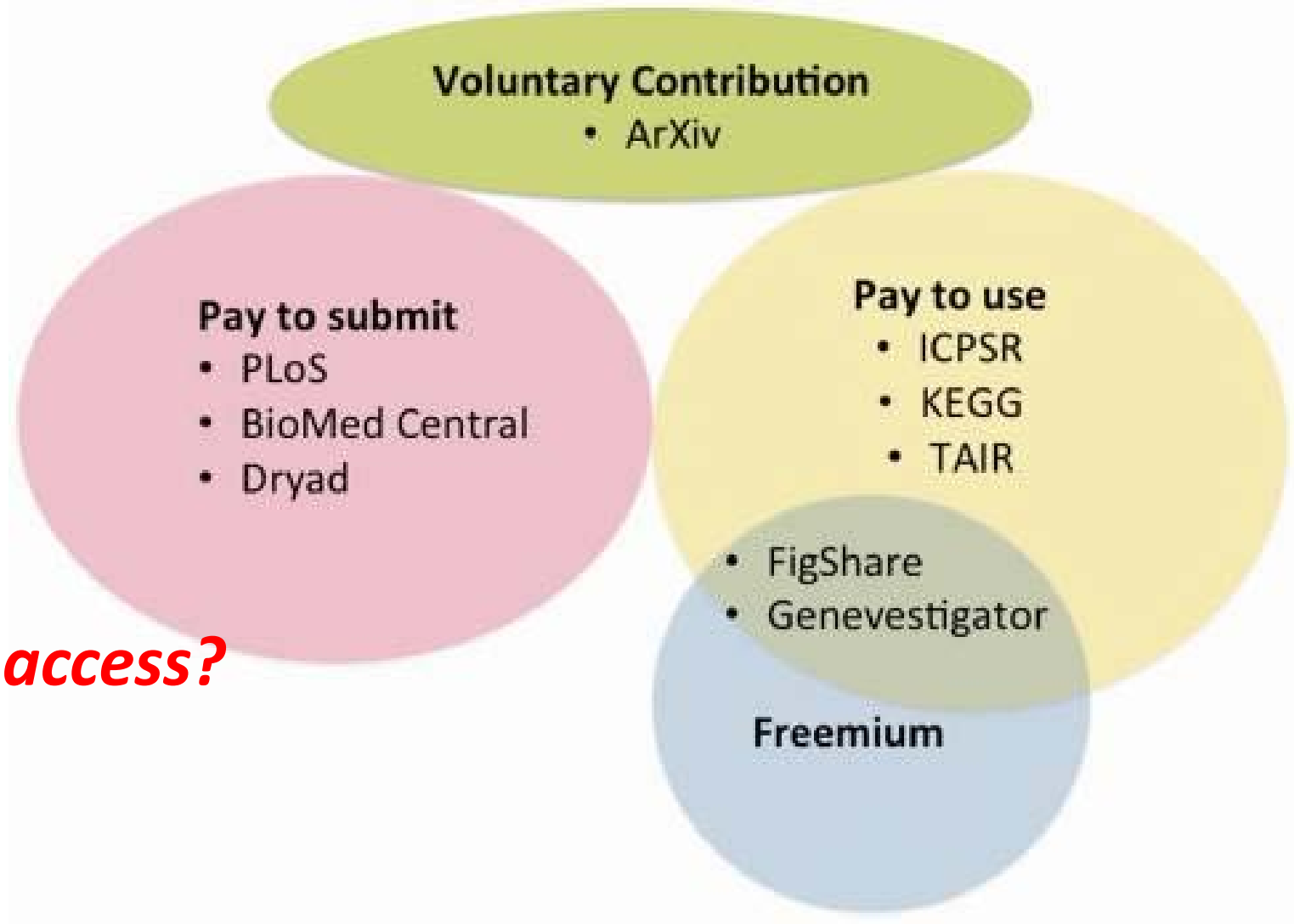
Core institutional support → 18-year survival / database support via short-duration research grants (current model) wasteful and inefficient

Category	N	Percent
Alive	53	16.3%
Alive - rebranded	23	7.0%
Archived	47	14.4%
Dead	203	62.3%
Total	326	100%

Longevity of biological databases (Attwood et al. 2015)

Who pays? User?

"who'd pay for YOUR data?!"



Subscription fee = open access?

Sustainable funding for biocuration: The Arabidopsis Information Resource (TAIR) as a case study of a subscription-based funding model (Reiser et al. 2016)

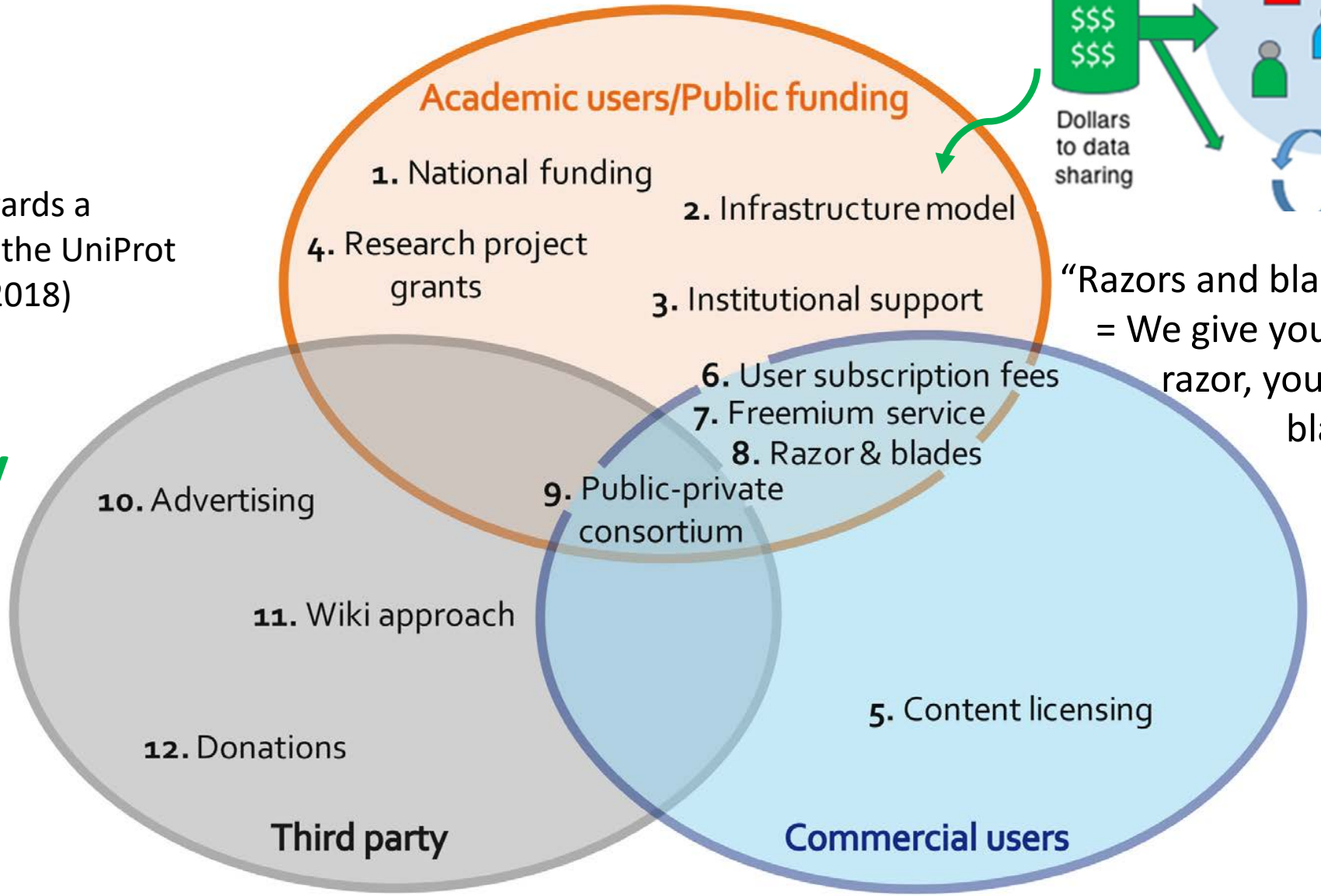
"Freemium" → data access is free but premium payments for additional services

12 funding models for data (knowledgebases) by origin of revenue

Funding knowledgebases: Towards a sustainable funding model for the UniProt use case (Gabella et al. 2017/2018)

Which are open access compliant / *equitable*? #1 – 4

Which are *stable*? #1, 2, maybe 3…



"Razors and blades" = We give you the razor, you buy blades

Final thoughts

- ✓Infrastructure: Most cost-effective, robust solutions may involve a mix of the proven with the innovative
- ✓Leadership and oversight: USDA Research and Economics Office
  - ➢Office of the Chief Scientist in partnership with the Office of the Chief Information Officer
  - ➢Why? Stewardship of public research data is a natural extension of their historic roles and responsibilities.
- ✓2018 Farm Bill opportunities?
  - ➢Creation of the Agriculture Advanced Research and Development Authority (AgARDA)
  - ➢Full appropriation of authorized funds will position AgARDA to address the data infrastructure needed for agricultural research to “make big data out of small.”
- ✓Getting started in D.C with AgARDA leading a partnership delivering benefits to all stakeholders in the agriculture data value chain?
  - ➢Convenings led by OCS and OCIO: Stakeholders in public listening sessions, USDA leadership in high-level meetings → create plan, operationalize, deploy