Title

From Data Creator to Data Reuser: Distance Matters

Authors
Borgman, Christine L.
Groth, Paul T.

Publication Date
2024-04-09

From Data Creator to Data Reuser: Distance Matters

Presentation by Christine L. Borgman, Distinguished Research Professor in Information Studies, University of California, Los Angeles, based on joint work with Paul T. Groth, University of Amsterdam

John P. Schneider Honorary Colloquium
iSchool, University of Texas, Austin
April 9, 2024
https://www.ischool.utexas.edu/events/572

Sharing research data is complex, labor-intensive, expensive, and requires infrastructure investments by multiple stakeholders. Open science policies focus on data release, yet reuse is also difficult and may never occur**.** Investments in data management could be made more wisely by considering who might reuse data, how, why, for what purposes, and when. Drawing upon empirical studies of data sharing and reuse, we develop the theoretical construct of *distance* between data creator and data reuser, identifying six distance dimensions that influence the ability to transfer knowledge effectively: domain, methods, collaboration, curation, purposes, and time and temporality. These dimensions are primarily social in character, with associated technical aspects that can decrease – or increase – distances between creators and reusers. We identify ways that data creators, data reusers, data archivists, and funding agencies can make data sharing and reuse more effective.

Christine L. Borgman conducts research in scientific data practices and information policy. Her publications in information studies, computer science, communication, and law include three award-winning books from MIT Press and more than 250 journal articles, conference papers, and other scholarly products. A Fellow of the American Association for the Advancement of Science and the Association for Computing Machinery, she has held visiting posts at Oxford, Harvard, and several European institutions. Professor Borgman is a member of the Library of Congress Scholars Council and the Board of Directors of the Electronic Privacy Information Center. Her honors and awards include the Paul Evan Peters Award from the Coalition for Networked Information, Association for Research Libraries, and EDUCAUSE; Award of Merit and the Research in Information Science Award, both from the Association for Information Science and Technology; and a Legacy Laureate of the University of Pittsburgh.

Preprint: Borgman, C. L., & Groth, P. T. (2024). *From Data Creator to Data Reuser: Distance Matters* (arXiv:2402.07926). arXiv. https://doi.org/10.48550/arXiv.2402.07926

From Data Creator to Data Reuser: Distance Matters

Christine L. Borgman

Distinguished Research Professor

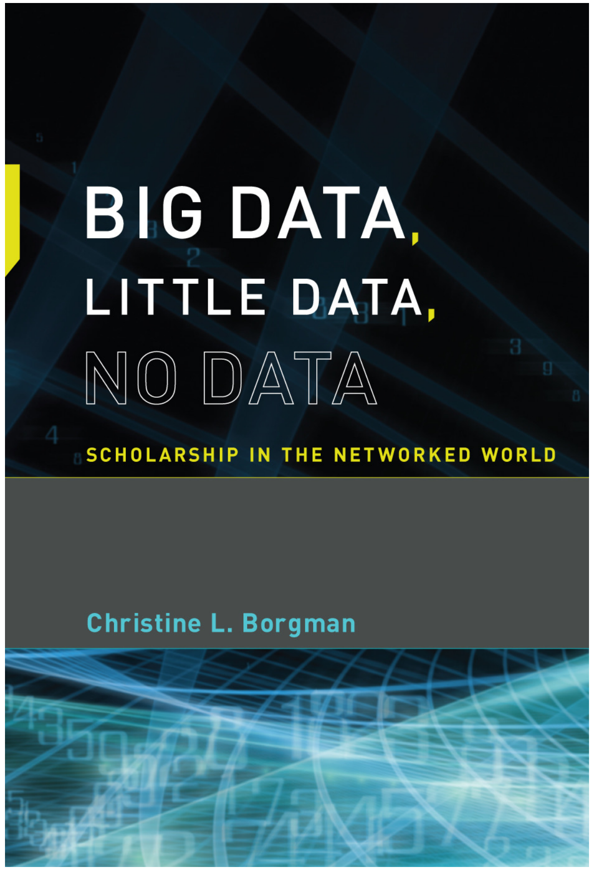
University of California, Los Angeles

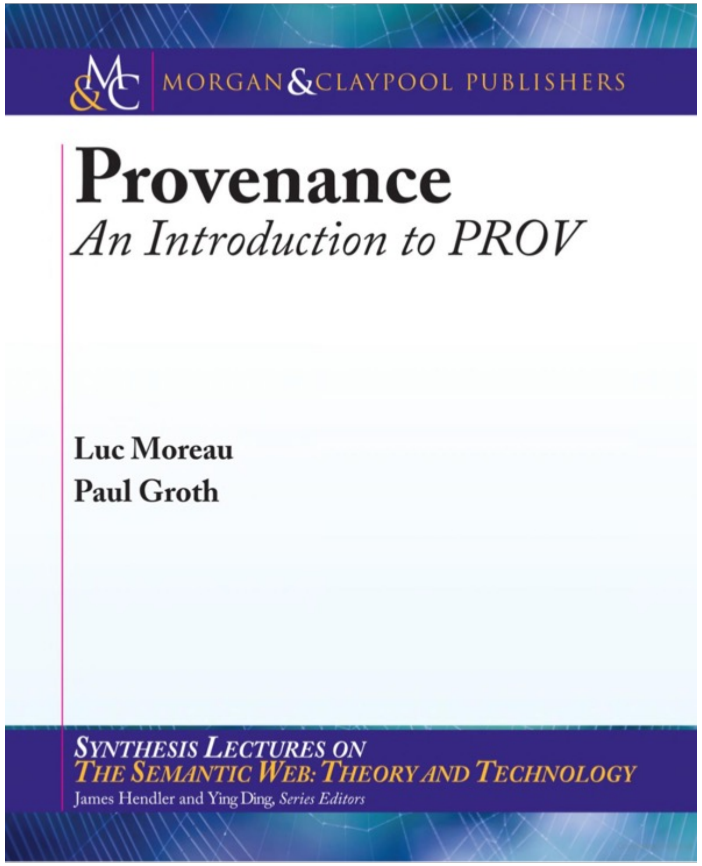
& Paul T. Groth

Professor of Algorithmic Data Science

University of Amsterdam

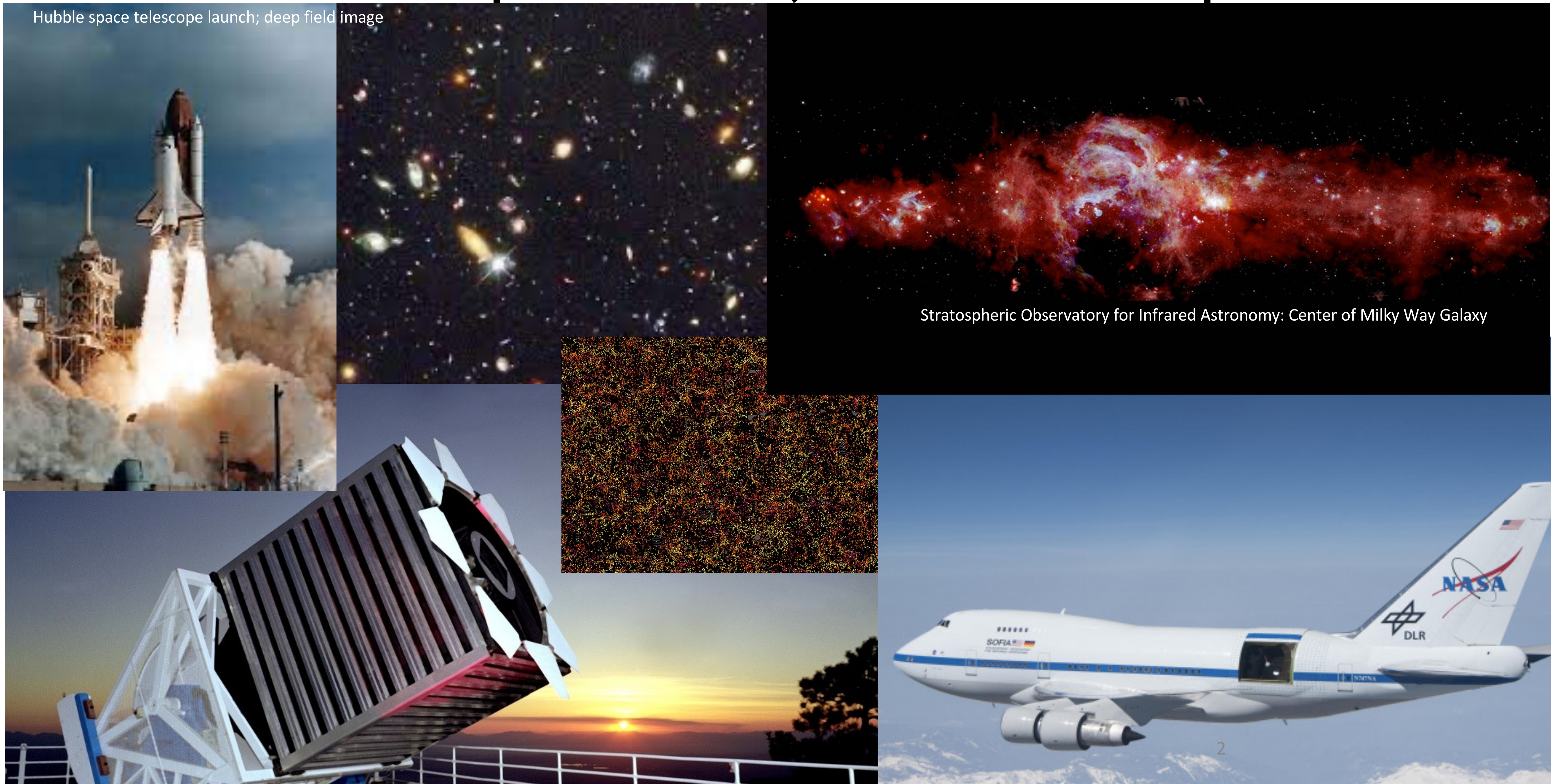
Distinguished Schneider Lecture
University of Texas, Austin
April 9, 2024





Borgman, C. L., & Groth, P. T. (2024). *From Data Creator to Data Reuser: Distance Matters* https://doi.org/10.48550/arXiv.2402.07926

Decades to acquire data, decades to preserve



Hubble space telescope launch; deep field image

Stratospheric Observatory for Infrared Astronomy: Center of Milky Way Galaxy

Research data for the public good

Why share research data?

- Reuse
- Reproduce
- Transparent
- Educate
- Required
  - Funding agencies
  - Journals

E·S·R·C ECONOMIC & SOCIAL RESEARCH COUNCIL

University of California CDL California Digital Library

National Science Foundation WHERE DISCOVERIES BEGIN

erc European Research Council Supporting top researchers from anywhere in the world

How to share research data?

- Deposit in data archive
- Publish data documentation
  - Research protocols
  - Codebooks
  - Software
  - Algorithms
- Link datasets to publication
- Cite data and software
- Develop instructional materials

NIH CDE Repository

Data sharing vs. Data reuse

- What are best investments in data sharing?
- What data are most likely to be reused?
- What factors influence data reuse?
  - Social
  - Technical
- How can data sharing be more effective ?
  - Data creators
  - Data reusers

Photo by Mathieu Stern on Unsplash

Borgman, C. L., & Bourne, P. E. (2022). Why It Takes a Village to Manage and Share Data. *Harvard Data Science Review*, *4*(3).
Borgman, C. L., & Brand, A. (2022). Data blind: Universities lag in capturing and exploiting data. *Science*, *378*(6626), 1278–1281.

How can data creators enable data reuse?

OPEN ACCESS Freely available online

PLOS | COMPUTATIONAL BIOLOGY

Editorial

Ten Simple Rules for the Care and Feeding of Scientific Data

Alyssa Goodman[1], Alberto Pepe[1]*, Alexander W. Blocker[1], Christine L. Borgman[2], Kyle Cranmer[3], Merce Crosas[1], Rosanne Di Stefano[1], Yolanda Gil[4], Paul Groth[5], Margaret Hedstrom[6], David W. Hogg[3], Vinay Kashyap[1], Ashish Mahabal[7], Aneta Siemiginowska[1], Aleksandra Slavkovic[8]

1 Harvard University, Cambridge, Massachusetts, United States of America, 2 University of California, Los Angeles, Los Angeles, California, United States of America, 3 New York University, New York, New York, United States of America, 4 University of Southern California, Los Angeles, California, United States of America, 5 Vrije Universiteit Amsterdam, Amsterdam, The Netherlands, 6 University of Michigan, Ann Arbor, Michigan, United States of America, 7 California Institute of Technology, Pasadena, California, United States of America, 8 Pennsylvania State University, State College, Pennsylvania, United States of America

Introduction

In the early 1600s, Galileo Galilei turned a telescope toward Jupiter. In his log book each night, he drew to-scale schematic diagrams of Jupiter and some oddly moving points of light near it. Galileo labeled each drawing with the date. Eventually he used his observations to conclude that the Earth orbits the Sun, just as the four Galilean moons orbit Jupiter. History shows Galileo to be much more than an astronomical hero, though. His clear and careful record keeping and publication style not only let Galileo understand the solar system, they continue to let *anyone* understand *how* Galileo did it. Galileo's notes directly integrated his **data** (drawings of Jupiter and its moons), key **metadata** (timing of each observation, weather, and telescope properties), and **text** (descriptions of methods, analysis, and conclusions). Critically, when Galileo included the information from those notes in *Sidereus Nuncius* [1], this integration of text, data, and metadata was preserved, as shown in Figure 1. Galileo's work advanced the "Scientific Revolution," and his approach to observation and analysis contributed significantly to the shaping of today's modern "scientific method" [2,3].

Today, most research projects are considered complete when a journal article based on the analysis has been written and published. The trouble is, unlike Galileo's report in *Sidereus Nuncius*, the amount of real data and data description in modern publications is almost never sufficient to repeat or even statistically verify a study being presented. Worse, researchers wishing to build upon and extend work presented in the literature often have trouble recovering data associated with an article after it has been published. More often than scientists would like to admit, they cannot even recover the data associated with their own published works.

Complicating the modern situation, the words "data" and "analysis" have a wider variety of definitions today than at the time of Galileo. Theoretical investigations can create large "data" sets through simulations (e.g., The Millennium Simulation Project: http://www.mpa-garching.mpg.de/galform/virgo/millennium/). Large-scale data collection often takes place as a community-wide effort (e.g., The Human Genome project: http://www.genome.gov/10001772), which leads to gigantic online "databases" (organized collections of data). Computers are so essential in simulations, and in the processing of experimental and observational data, that it is also often hard to draw a dividing line between "data" and "analysis" (or "code") when discussing the care and feeding of "data." Sometimes, a copy of the code used to create or process data is so essential to the use of those data that the code should almost be thought of as part of the "metadata" description of the data. Other times, the code used in a scientific study is more separable from the data, but even then, many preservation and sharing principles apply to code just as well as they do to data.

So how do we go about caring for and feeding data? Extra work, no doubt, is associated with nurturing your data, but care up front will save time and increase insight later. Even though a growing number of researchers, especially in large collaborations, know that conducting research with sharing and reuse in mind is essential, it still requires a paradigm shift. Most people are still motivated by piling up publications and by getting to the next one as soon as possible. But, the more we scientists find ourselves wishing we had access to extant but now unfindable data [4], the more we will realize why bad data management is bad for science. How can we improve?

This article offers a short guide to the steps scientists can take to ensure that their data and associated analyses continue to be of value and to be recognized. In just the past few years, hundreds of scholarly papers and reports have been written on questions of data sharing, data provenance, research reproducibility, licensing, attribution, privacy, and more—but our goal here is *not* to review that literature. Instead, we present a short guide intended for researchers who want to know why it is important to "care for and feed" data, with some practical advice on how to do that. The final section at the close of this work (Links to Useful Resources) offers links to the types of services referred to throughout the text. **Boldface lettering** below highlights actions one can take to follow the suggested rules.

Rule 1. Love Your Data, and Help Others Love It, Too

Data management is a repeat-play game. If you take care to make your data

Citation: Goodman A, Pepe A, Blocker AW, Borgman CL, Cranmer K, et al. (2014) Ten Simple Rules for the Care and Feeding of Scientific Data. PLoS Comput Biol 10(4): e1003542. doi:10.1371/journal.pcbi.1003542

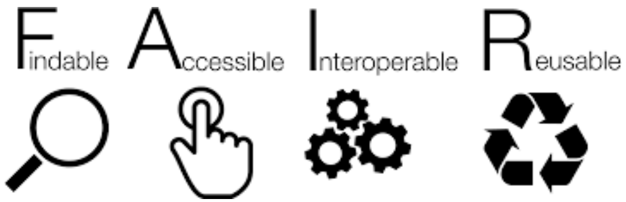
Editor: Philip E. Bourne, University of California San Diego, United States of America

Published April 24, 2014

Funding: The authors received no specific funding for writing this manuscript.

Competing Interests: The authors have declared that no competing interests exist.

* E-mail: alberto.pepe@gmail.com



www.nature.com/scientificdata

SCIENTIFIC DATA

Amended: Addendum

OPEN

SUBJECT CATEGORIES
» Research data
» Publication characteristics

Received: 10 December 2015
Accepted: 12 February 2016
Published: 15 March 2016

Comment: The FAIR Guiding Principles for scientific data management and stewardship

Mark D. Wilkinson *et al.*#

There is an urgent need to improve the infrastructure supporting the reuse of scholarly data. A diverse set of stakeholders—representing academia, industry, funding agencies, and scholarly publishers—have come together to design and jointly endorse a concise and measureable set of principles that we refer to as the FAIR Data Principles. The intent is that these may act as a guideline for those wishing to enhance the reusability of their data holdings. Distinct from peer initiatives that focus on the human scholar, the FAIR Principles put specific emphasis on enhancing the ability of machines to automatically find and use the data, in addition to supporting its reuse by individuals. This Comment is the first formal publication of the FAIR Principles, and includes the rationale behind them, and some exemplar implementations in the community.

Supporting discovery through good data management

Good data management is not a goal in itself, but rather is the key conduit leading to knowledge discovery and innovation, and to subsequent data and knowledge integration and reuse by the community after the data publication process. Unfortunately, the existing digital ecosystem surrounding scholarly data publication prevents us from extracting maximum benefit from our research investments (e.g., ref. 1). Partially in response to this, science funders, publishers and governmental agencies are beginning to require data management and stewardship plans for data generated in publicly funded experiments. Beyond proper collection, annotation, and archival, data stewardship includes the notion of 'long-term care' of valuable digital assets, with the goal that they should be discovered and re-used for downstream investigations, either alone, or in combination with newly generated data. The outcomes from good data management and stewardship, therefore, are high quality digital publications that facilitate and simplify this ongoing process of discovery, evaluation, and reuse in downstream studies. What constitutes 'good data management' is, however, largely undefined, and is generally left as a decision for the data or repository owner. Therefore, bringing some clarity around the goals and desiderata of good data management and stewardship, and defining simple guideposts to inform those who publish and/or preserve scholarly data, would be of great utility.

This article describes four foundational principles—Findability, Accessibility, Interoperability, and Reusability—that serve to guide data producers and publishers as they navigate around these obstacles, thereby helping to maximize the added-value gained by contemporary, formal scholarly digital publishing. Importantly, it is our intent that the principles apply not only to 'data' in the conventional sense, but also to the algorithms, tools, and workflows that led to that data. All scholarly digital research objects[2]—from data to analytical pipelines—benefit from application of these principles, since all components of the research process must be available to ensure transparency, reproducibility, and reusability.

There are numerous and diverse stakeholders who stand to benefit from overcoming these obstacles: researchers wanting to share, get credit, and reuse each other's data and interpretations; professional data publishers offering their services; software and tool-builders providing data analysis and processing services such as reusable workflows; funding agencies (private and public) increasingly

Correspondence and requests for materials should be addressed to B.M. (email: barend.mons@dtls.nl).
#A full list of authors and their affiliations appears at the end of the paper.

Goodman, A., et al. (2014). Ten Simple Rules for the Care and Feeding of Scientific Data. *PLoS Computational Biology*, *10*(4), e1003542. https://doi.org/10.1371/journal.pcbi.1003542

Wilkinson, M. D., et al. (2016). The FAIR Guiding Principles for scientific data management and stewardship. *Scientific Data*, *3*, 160018. http://dx.doi.org/10.1038/sdata.2016.18

Dimensions of Distance from Data Creator to Reuser

Social and Technical Distances

1. Domain
2. Methods
3. Collaboration
4. Curation
5. Purposes
6. Time and temporality

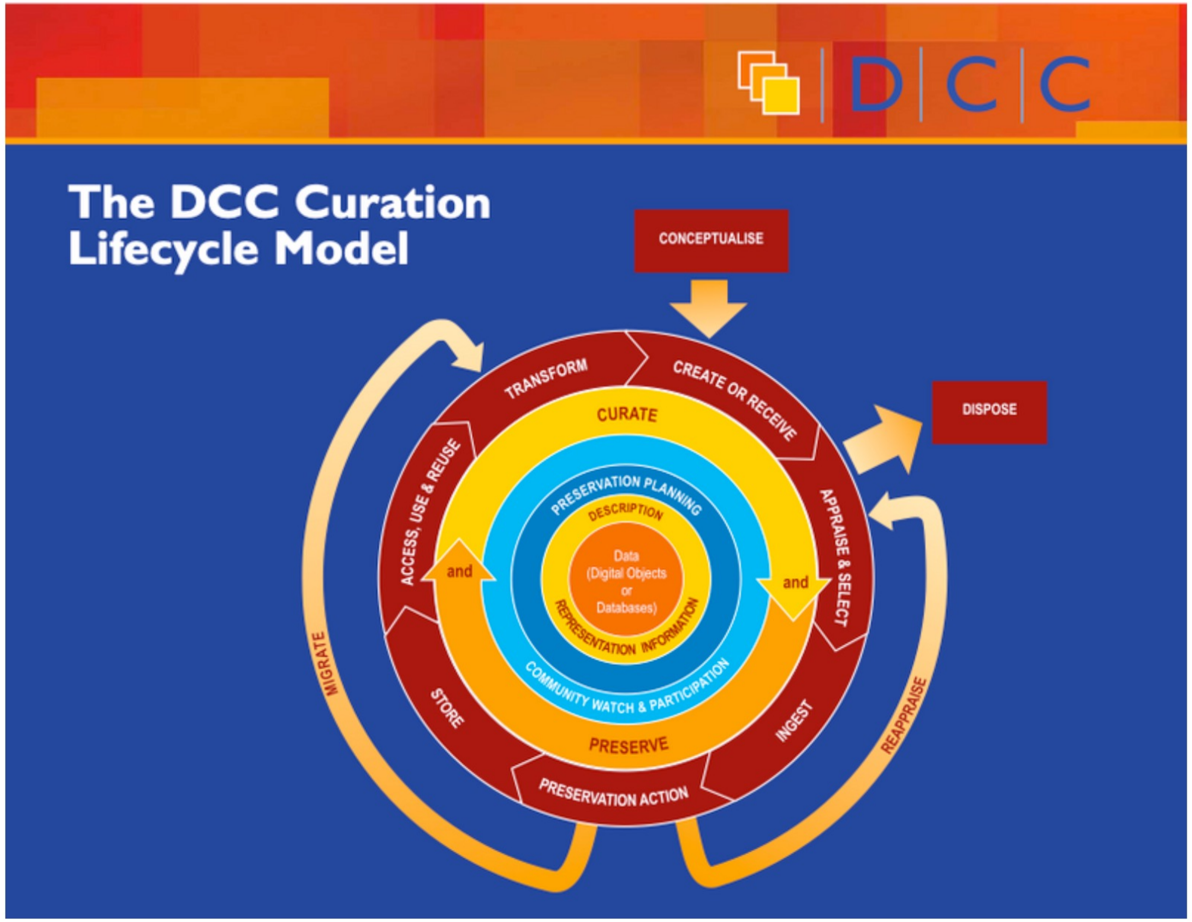
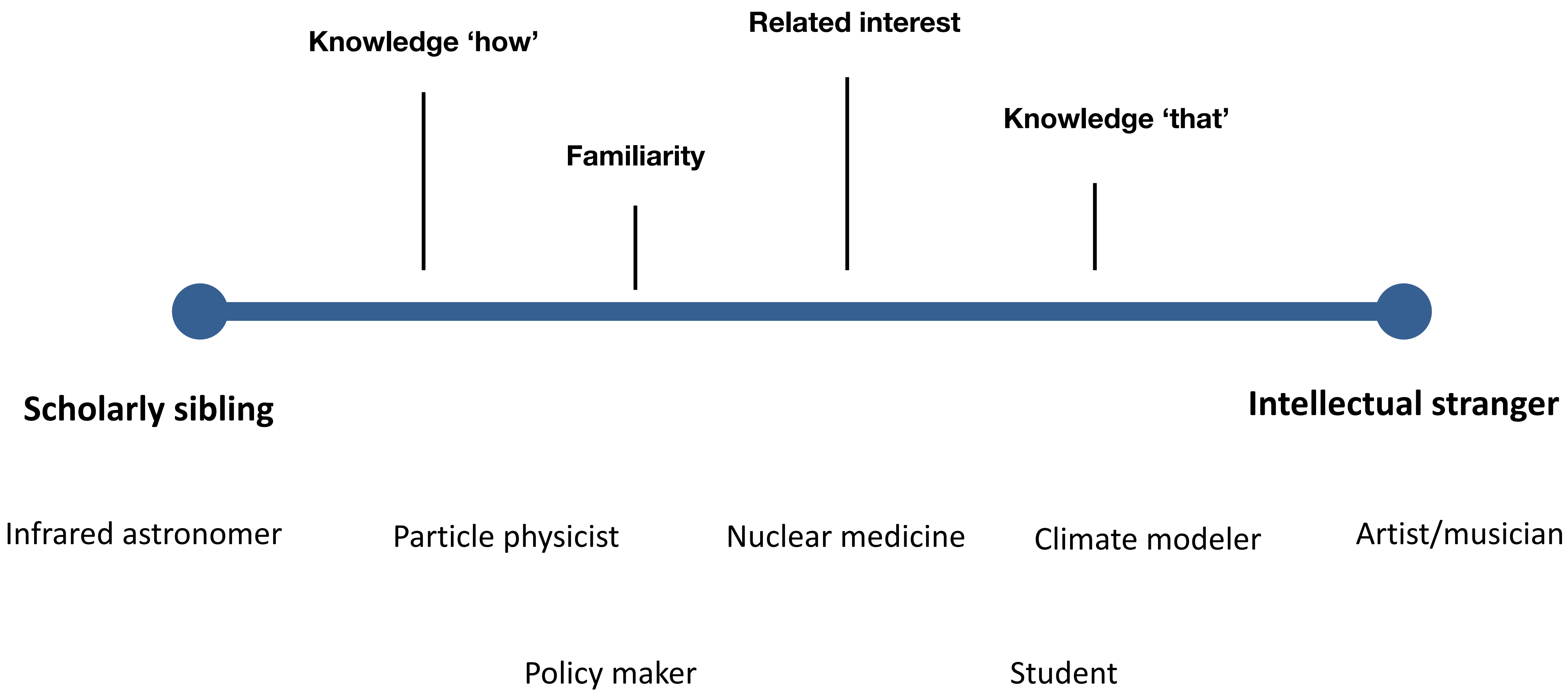


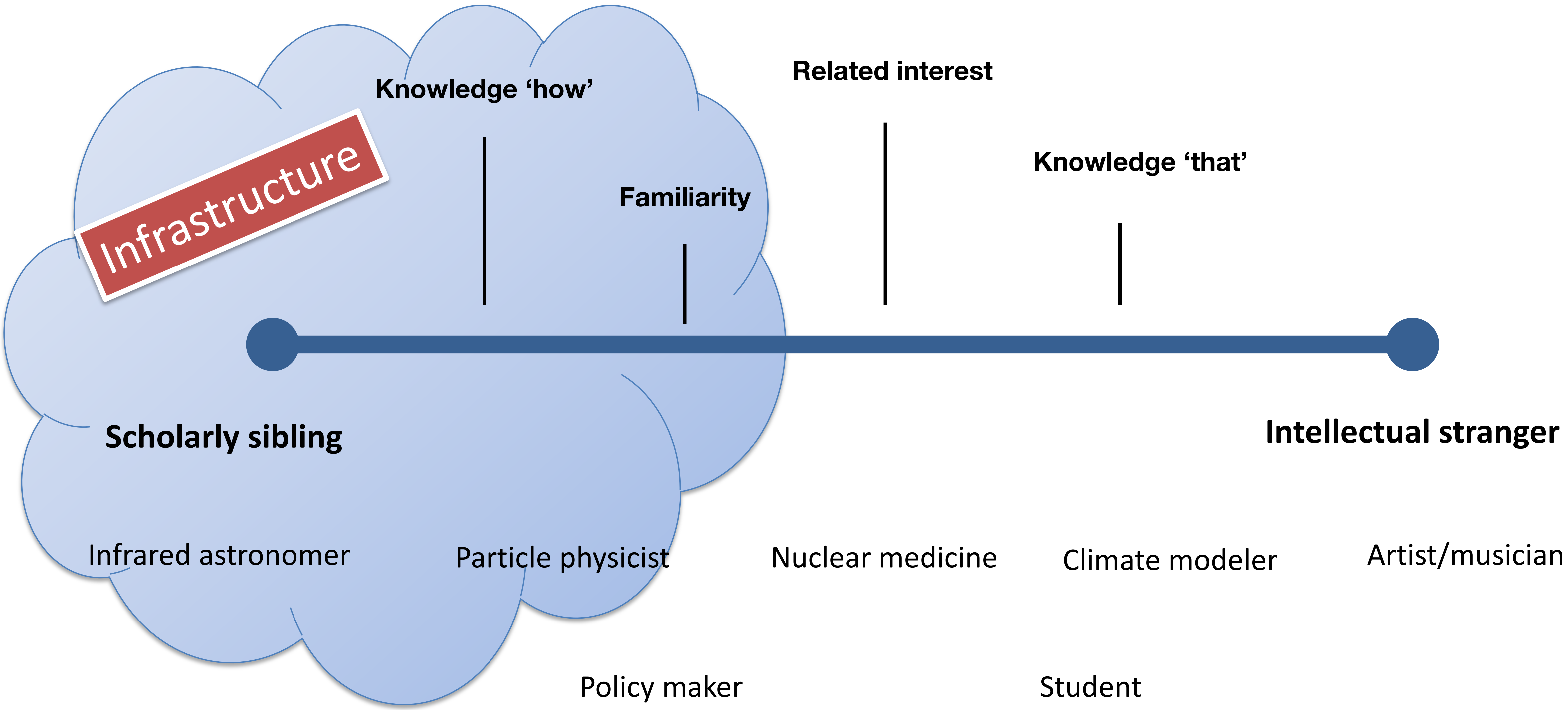
Figure 2. Digital Curation Center Curation Lifecycle Model (Higgins, 2008). Reprinted with permission of the Digital Curation Centre, U.K.

Borgman, C. L. (2019). The lives and after lives of data. *Harvard Data Science Review*, *1*(1). https://doi.org/10.1162/99608f92.9a36bdb6

Domain Distance



Domain Distance



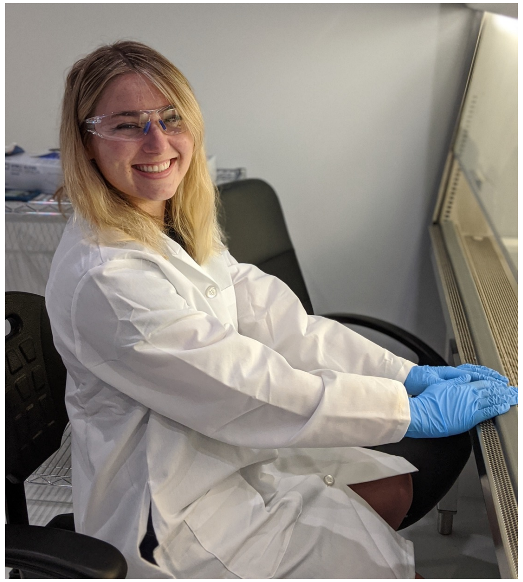
Data Creators’ Advantage

Comparative Data Reuse

- Ground truthing: calibrate, compare, confirm
- Instrument calibration
- Frequent, routine practice

Integrative Data Reuse

- Analysis: identify patterns, correlations, causal relationships
- Novel statistical analyses
- Rare, emergent practice

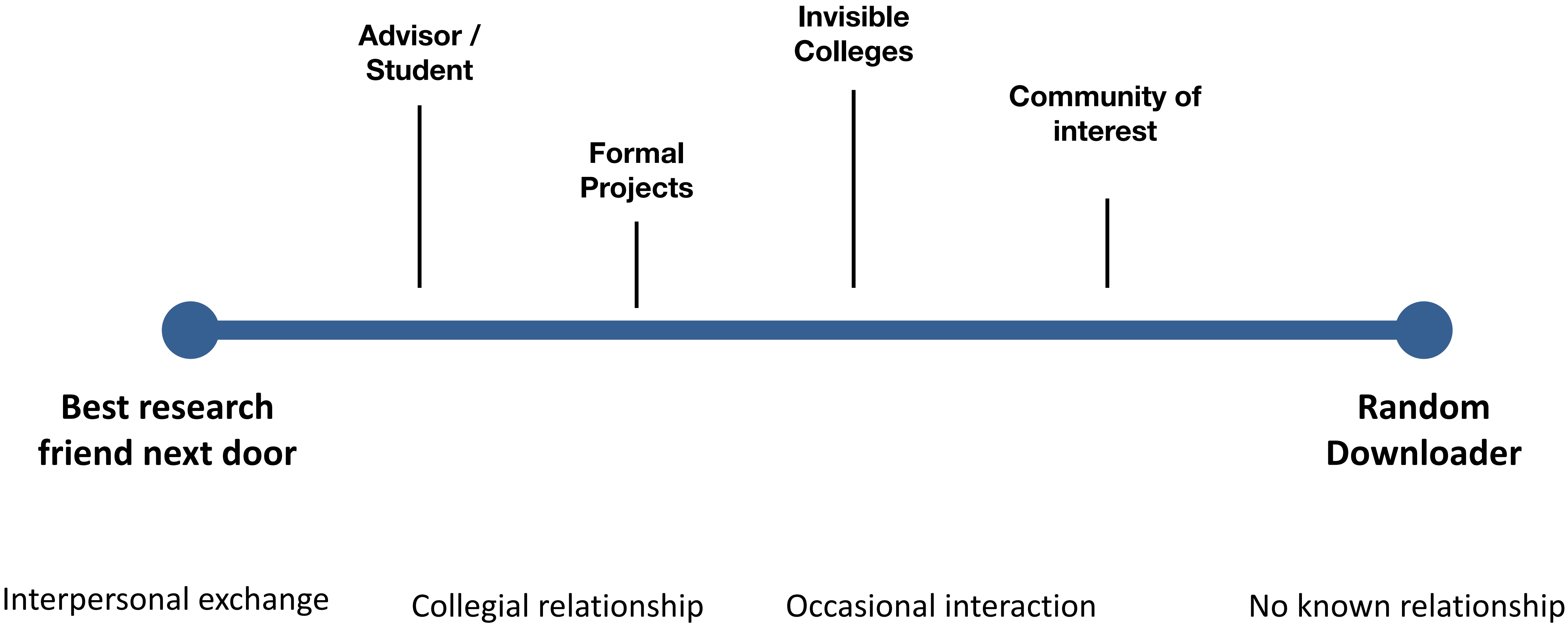


Bret Kavanaugh, Unsplash

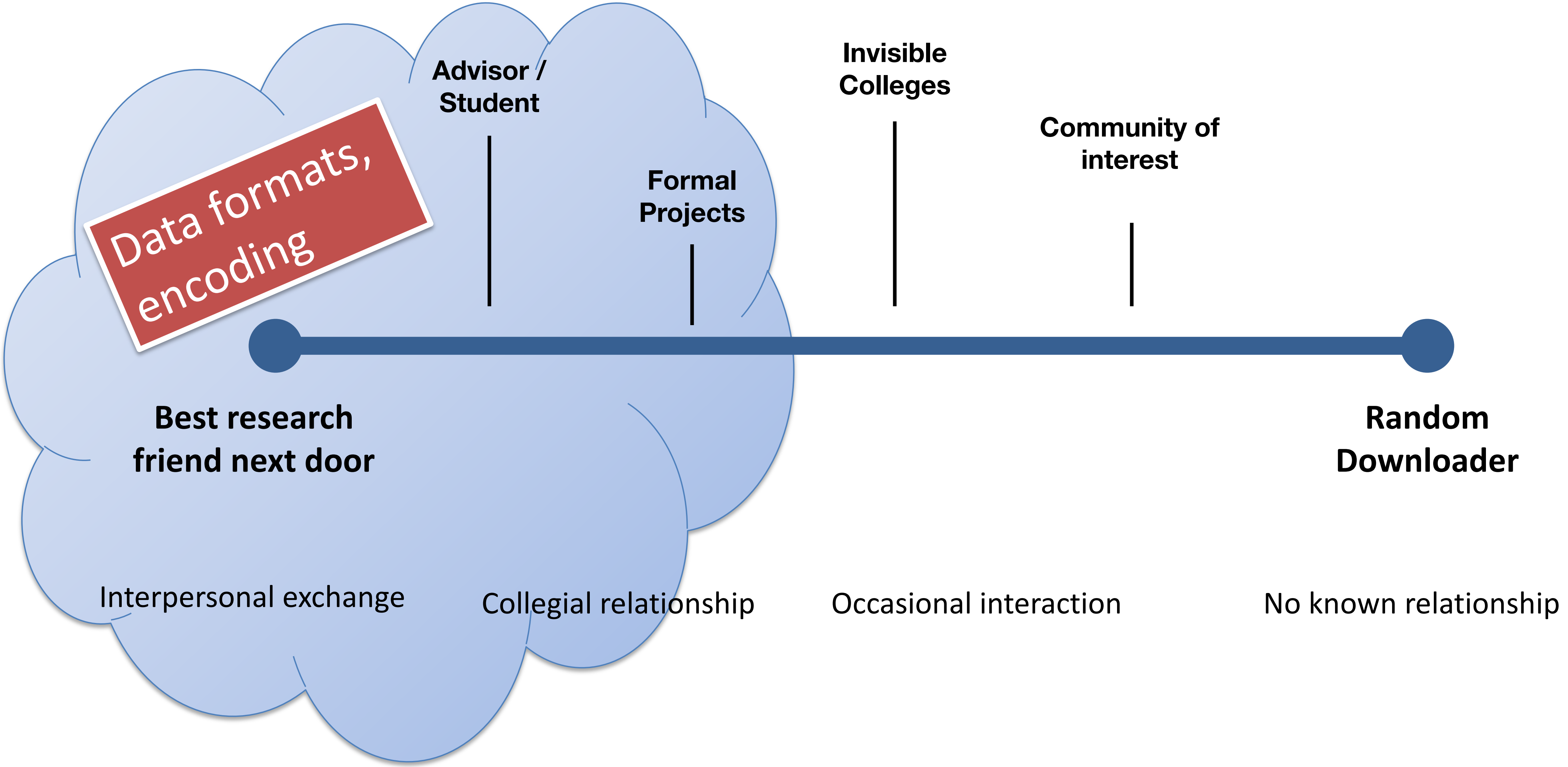
Pasquetto, I. V., Borgman, C. L., & Wofford, M. F. (2019). Uses and reuses of scientific data: The data creators’ advantage. *Harvard Data Science Review,* 1:2

National Cancer Institute

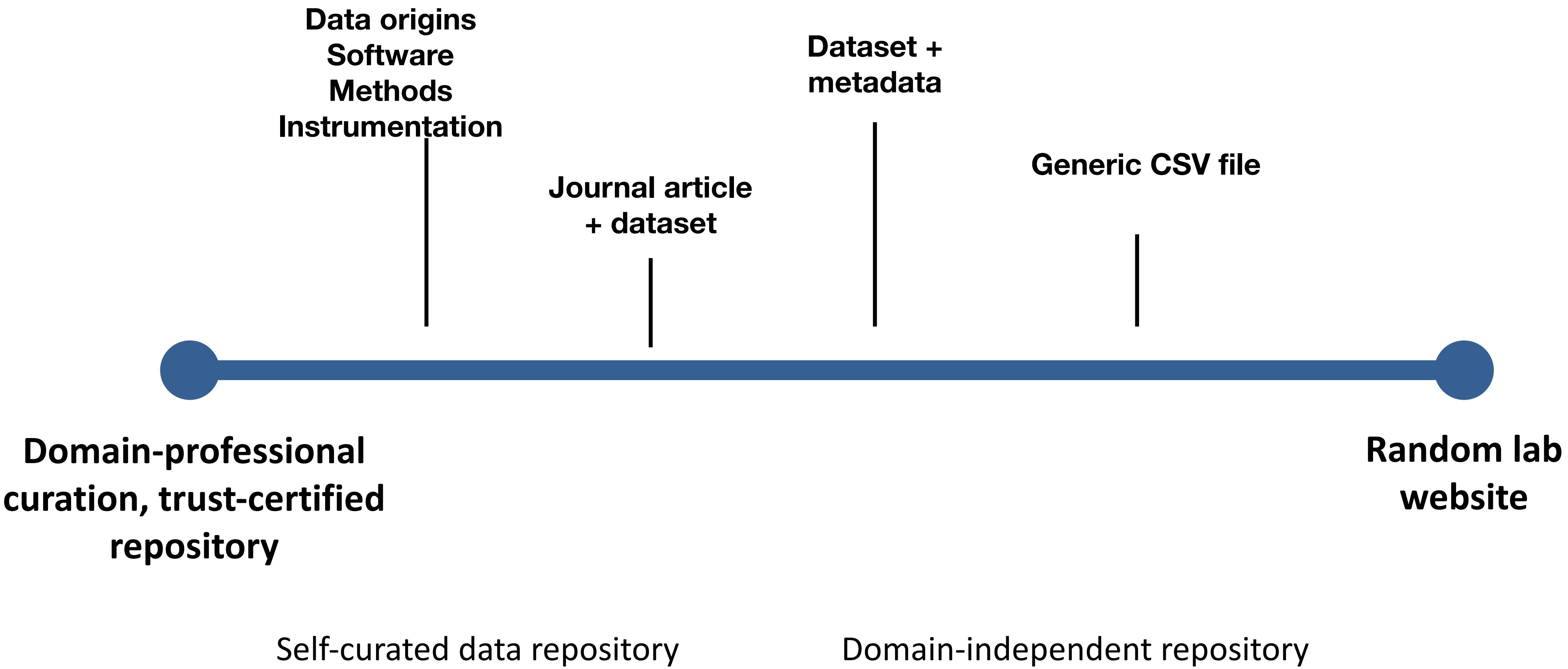
Collaboration Distance



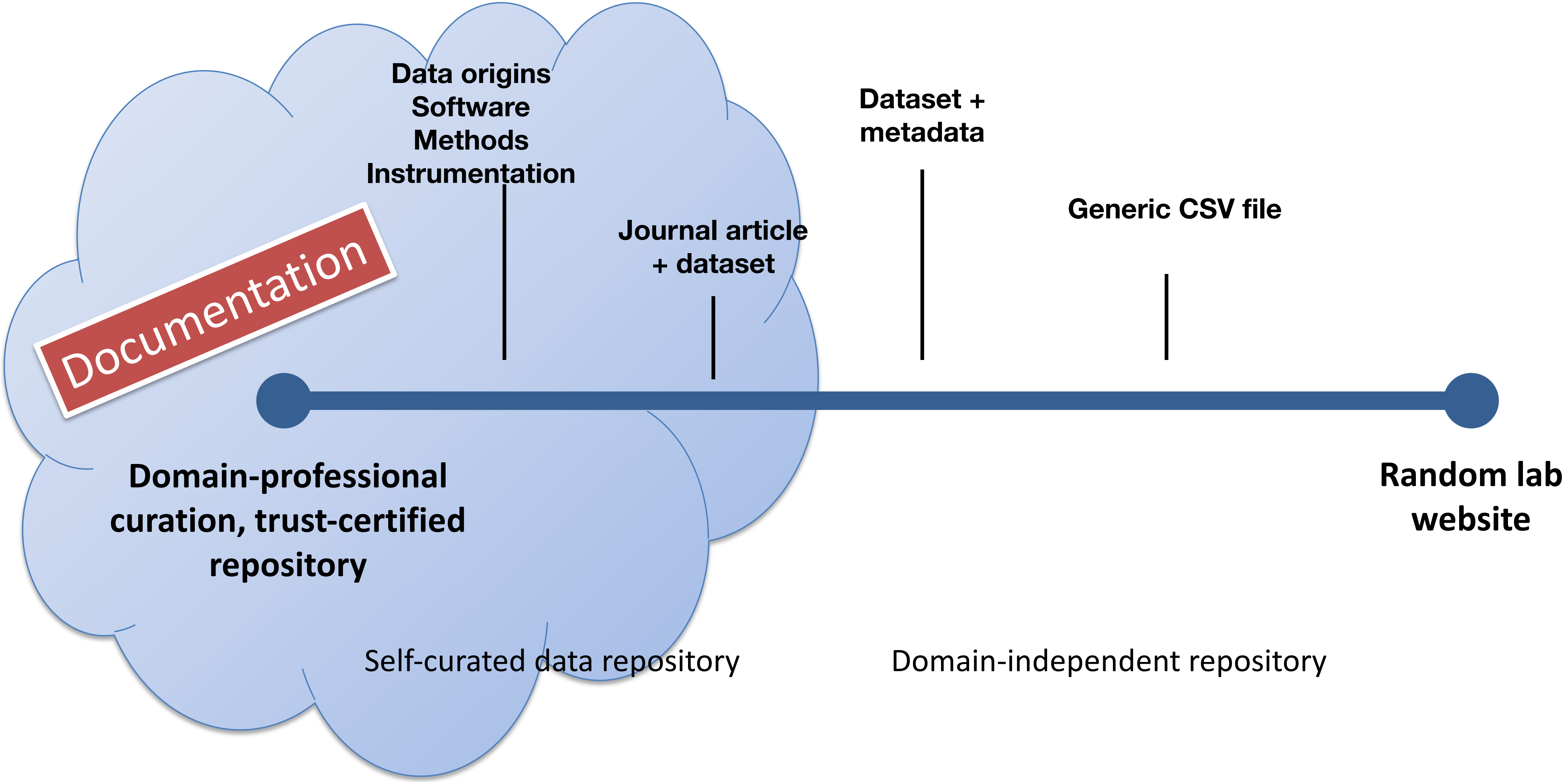
Collaboration Distance



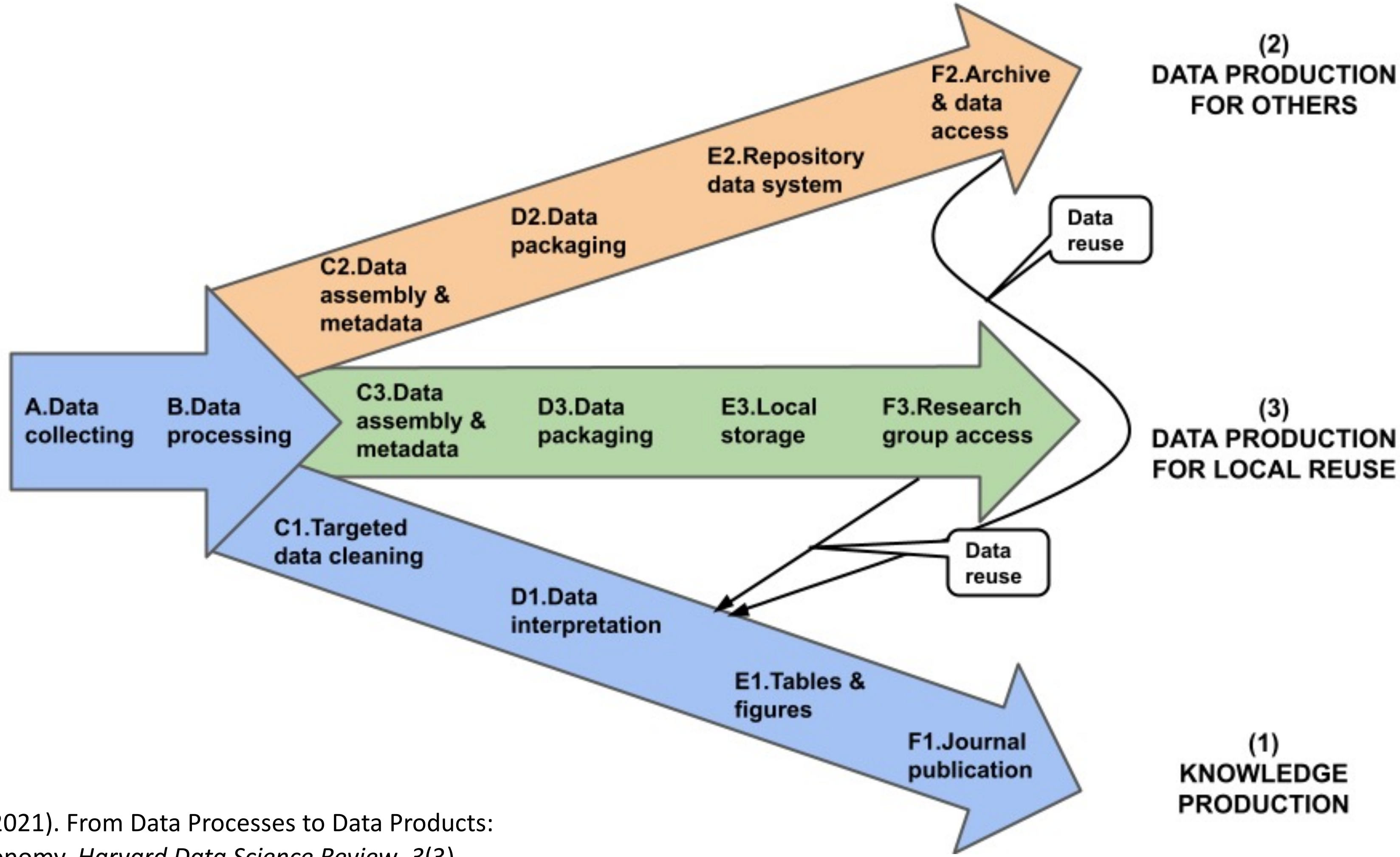
Curation Distance



Curation Distance



Data production, knowledge production, and reuse



- Borgman, C. L., & Wofford, M. F. (2021). From Data Processes to Data Products: Knowledge Infrastructures in Astronomy. *Harvard Data Science Review*, *3*(3).
- Baker, K. S., & Mayernik, M. S. (2020). Disentangling knowledge production and data production. *Ecosphere*, *11*(7), e03191.

Discussion and Implications

Image by Jean-Philippe Delberghe on Unsplash.com

Data creator ⇔ Data reuser

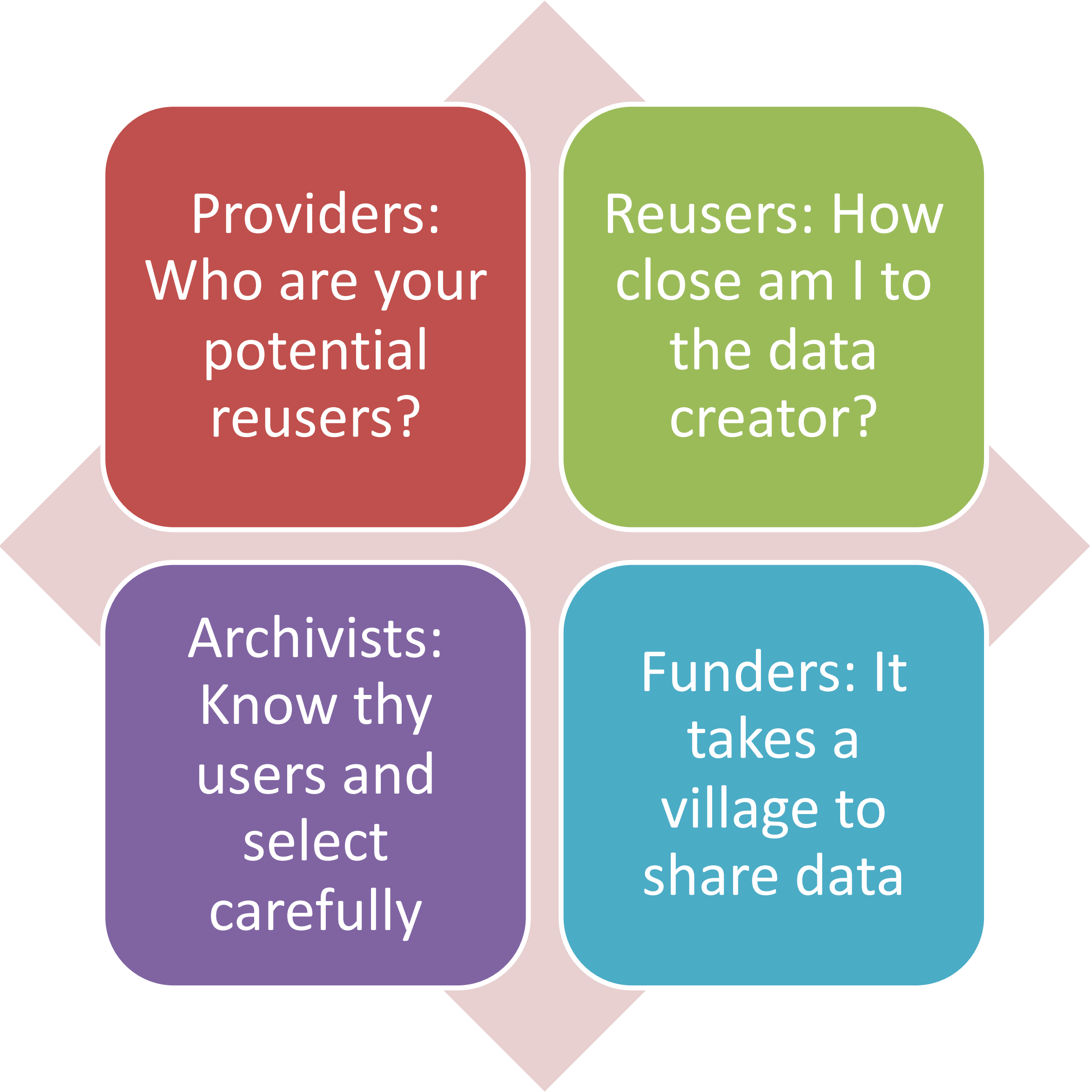
- Knowledge transfer
  - Direct, interpersonal
  - Intermediaries: archives
- Shorter distances
  - Easier, fuller transfer
  - Higher likelihood of occurrence
- Longer distances
  - Greater difficulty, expense
  - Lower likelihood of occurrence

Photo by Nadine Shaabana on Unsplash

Social and Technical Distances

Domain	⟷ Lubrication / friction ⟷	Infrastructure
Methods	⟷ Lubrication / friction ⟷	Standards
Collaboration	⟷ Lubrication / friction ⟷	Data formats
Curation	⟷ Lubrication / friction ⟷	Documentation
Purposes	⟷ Lubrication / friction ⟷	Interfaces
Time and temporality	⟷ Lubrication / friction ⟷	Provenance

Recommendations to Stakeholders



Recommendations to Data Creators

Identify the audiences for your data

Focus on reusers early in the research process

Greater distance => greater investment in data

- Infrastructure
- Standards
- Data formats
- Documentation
- Interfaces
- Provenance

Recommendations to Data Reusers

Identify your locus on each dimension

Assess investments needed to find reusable data

Assess investments needed to make data reusable

- Infrastructure
- Standards
- Data formats
- Documentation
- Interfaces
- Provenance

Recommendations to Data Archivists

Know where your users reside on each dimension

Greater distance => more investment in curation

Greater distance => more investment user services

- Infrastructure
- Standards
- Data formats
- Documentation
- Interfaces
- Provenance

Recommendations to Funding Agencies

Support investigators in targeted data release

Support more research on

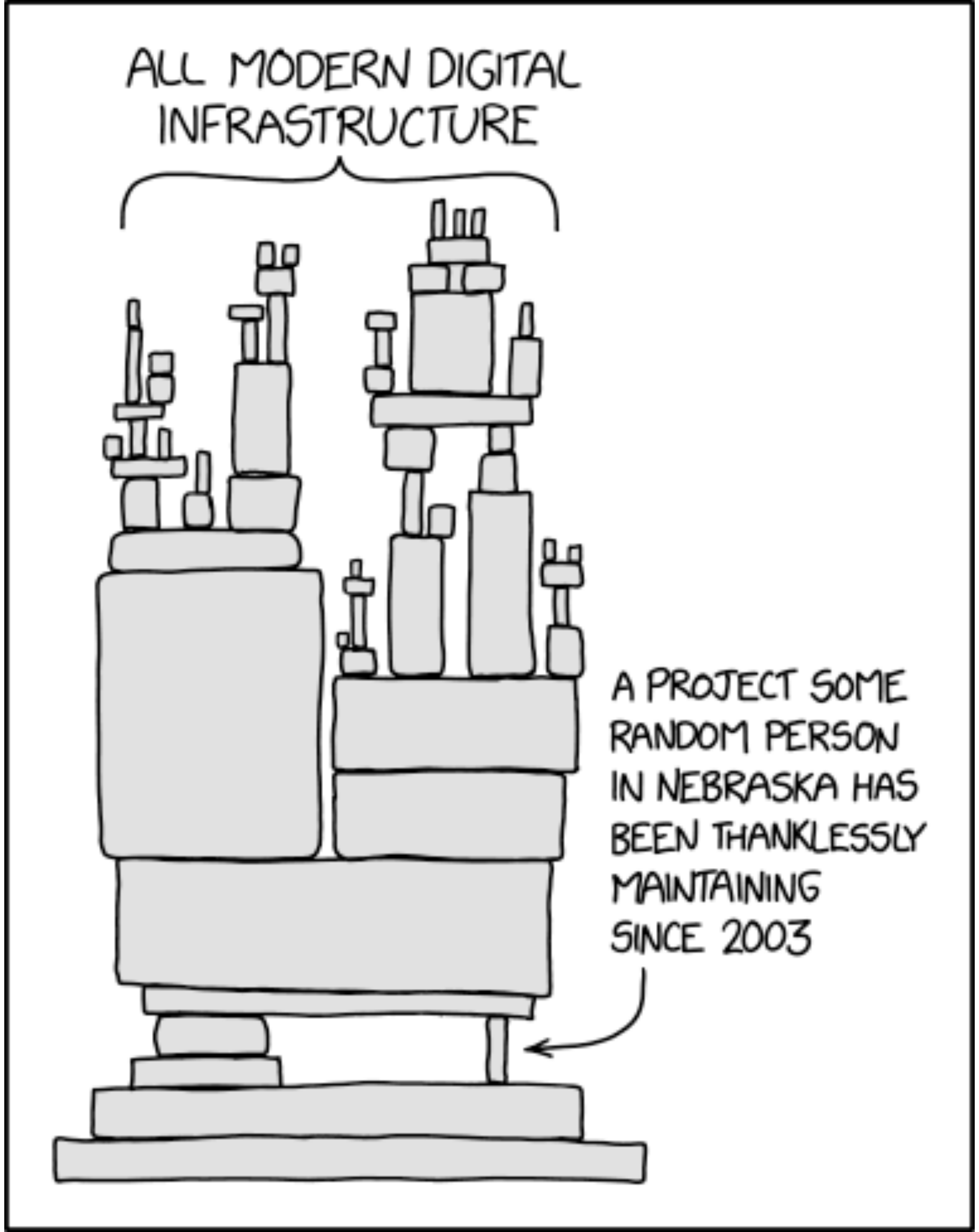
Infrastructure for data sharing and reuse

What data are reused, by whom, when, why, how

How to facilitate data reuse

How to develop ‘human infrastructure’ for research

How to improve knowledge exchange at distance



https://www.explainxkcd.com/wiki/index.php/2347:_Dependency

May all your problems be technical

Jim Gray, Turing Award Winner