Data & Materials Sharing[1]

Why, when, where, & how

Rick Gilmore Bruce Desmarais

Psychology/CSC

Political Science/SoDA

[1]NSF & NIH

Preliminaries

Agenda

- Why share data & materials
- When to share...
- Where to share...
- How to share...

Why share data

Transparency

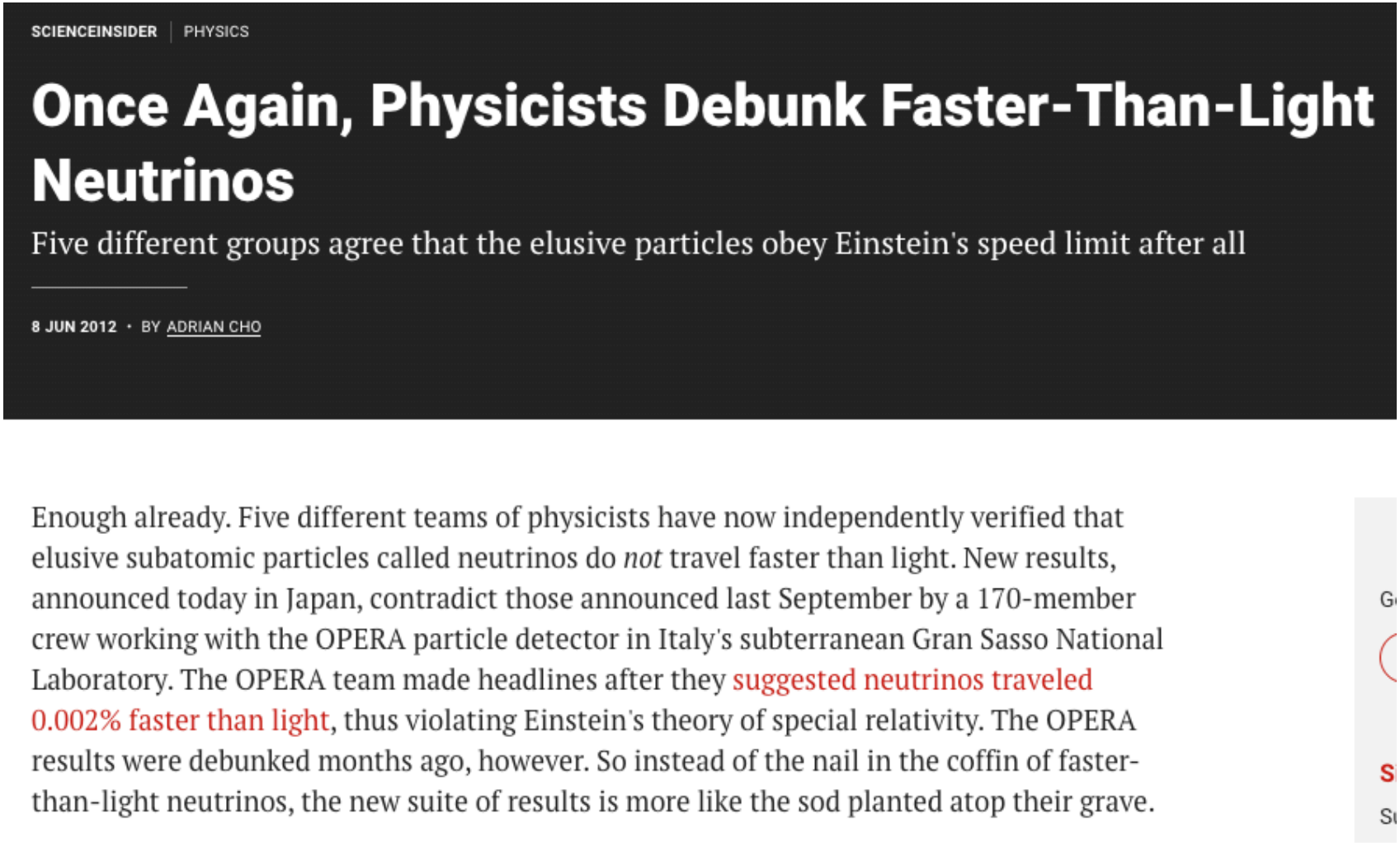
The Mind of a Con Man

Give this article 442

Accelerate discovery

Figure 2: (Pfizer 2020)

Improve rigor



SCIENCEINSIDER | PHYSICS

Once Again, Physicists Debunk Faster-Than-Light Neutrinos

Five different groups agree that the elusive particles obey Einstein's speed limit after all

8 JUN 2012 • BY ADRIAN CHO

Enough already. Five different teams of physicists have now independently verified that elusive subatomic particles called neutrinos do *not* travel faster than light. New results, announced today in Japan, contradict those announced last September by a 170-member crew working with the OPERA particle detector in Italy's subterranean Gran Sasso National Laboratory. The OPERA team made headlines after they suggested neutrinos traveled 0.002% faster than light, thus violating Einstein's theory of special relativity. The OPERA results were debunked months ago, however. So instead of the nail in the coffin of faster-than-light neutrinos, the new suite of results is more like the sod planted atop their grave.

Figure 3: (Cho 2012)

Internal reproducibility

- What's your project's "bus number"?
- Could someone else on your team reproduce analyses done by X?
- *Methods* reproducibility (Goodman, Fanelli, and Ioannidis 2016)

Requirement

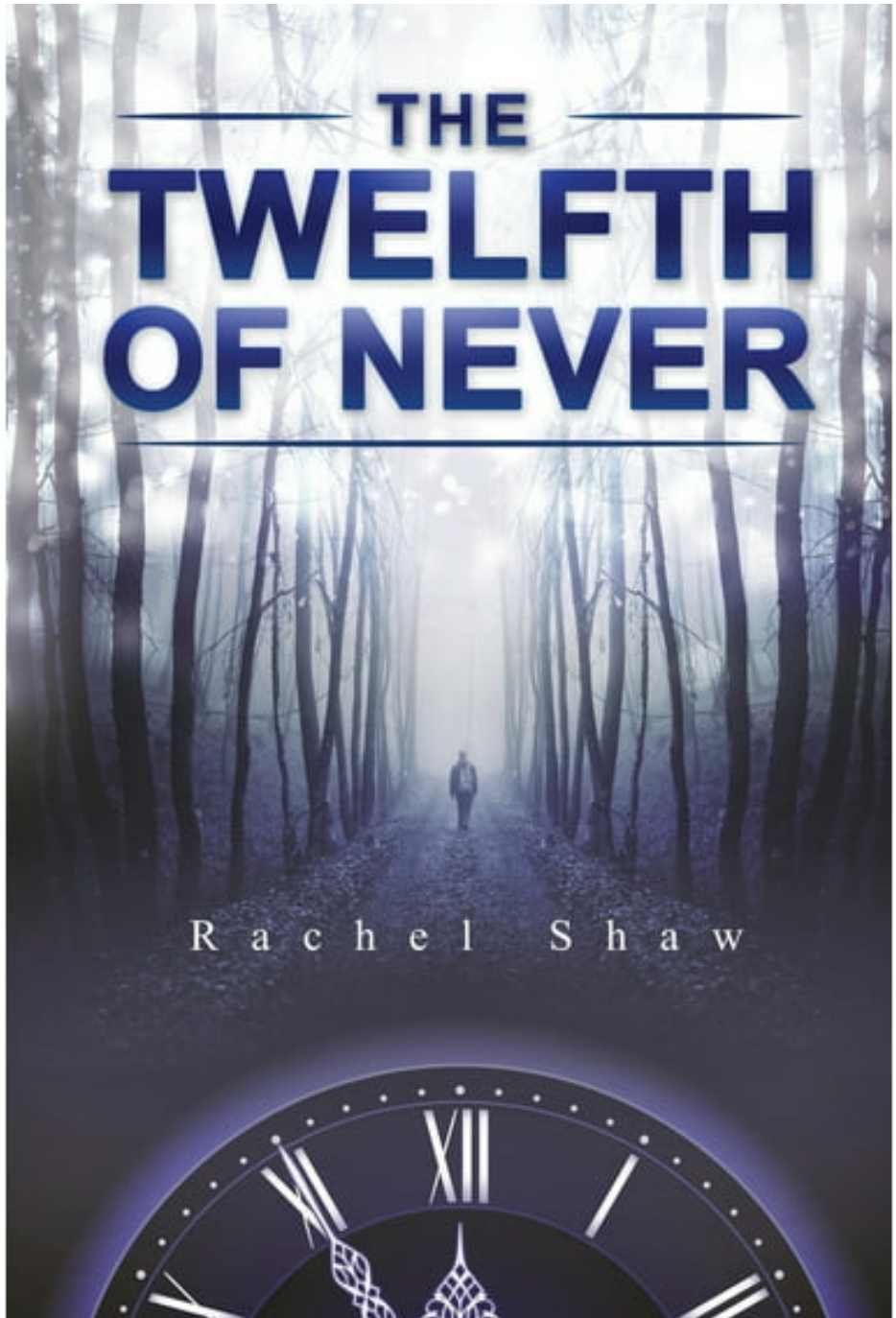
- Funder
- Journal
- Institutional

Ethical duty

Whereas regulators of human subjects research often view data sharing solely in terms of potential risks to subjects, we argue that the principles of human subject research require an analysis of both risks and benefits, and that such an analysis suggests that ***researchers may have a positive duty to share data*** *in order to maximize the contribution that individual participants have made. (Brakewood and Poldrack 2013)*

When to share

When NOT to share...



- When you've milked every last ounce of information out of a dataset, and all that remains is a lifeless, shriveled husk

On submission

- Some journals require that authors submit data for review. Ex, *Nature Scientific Data*

2 Select a repository for your data

When submitting a **Data Descriptor**, authors must deposit all relevant datasets in an appropriate public repository prior submission, and the completeness of these datasets will be considered during editorial evaluation and peer-review. Datasets must be made publicly available without restriction in the event that the Data Descriptor is accepted for publication (except reasonable controls related to human privacy issues or public safety).

- Reviewers have the option to review the data along with the manuscript.
- Bruce had experience of reviewer at NSD pointing out that numbers from the manuscript did not match data repository.

After acceptance

- Pre-publication verification (separate from review) increasingly common

GUIDELINES FOR DATA REPLICATION

The Journal of Politics operates a strict data replication policy effective January 1, 2021. It is the policy of the *JOP* to publish papers only if the data used in the analysis are clearly documented, available, and the empirical findings reproducible. Authors of accepted papers that contain empirical as well as simulation-based analyses are required to provide datasets, codes, and other relevant information necessary to facilitate replication. All manuscripts are accepted contingent on their replicability, which will initially be assessed by the *JOP* replication analyst(s) who will be assigned to your manuscript at the conditional acceptance stage. Manuscripts that are not replicable will be rejected for publication. This document provides guidelines on how to prepare data replication materials.

- Can take up to six months of back-and-forth with journal assistant.
- Leads to more careful organization pre-submission.

On publication

- Post and disseminate data and code with paper.
- Creates independent way to get cites.

SNAP judgments into the digital age: Reporting on food stamps varies significantly with time, publication type, a…
Benjamin W. Chrisinger, Eliza W. Kinsey, Ellie Pavlick, Chris Callison-Burch

Abstract
Introduction
Materials and methods
Results
Discussion

20. Bakshy E, Messing S, Adamic L. Replication Data for: Exposure to Ideologically Diverse News and Opinion on Facebook. Harvard Dataverse, V2; 2015.
View Article • Google Scholar

21. Roberts ME, Stewart BM, Tingley D, Lucas C, Leder-Luis J, Gadarian SK, et al. Structural Topic Models for Open-Ended Survey Responses. American Journal of Political Science. 2014;58: 1064–1082.
View Article • Google Scholar

End of grant period

- Major funders require data/material/publication sharing

NSF's data sharing policy

NSF-funded investigators are expected to share with other researchers, at no more than incremental cost and within a reasonable time, the primary data, samples, physical collections and other supporting materials created or gathered in the course of work under NSF awards.

Where to share data

Data repository

- Dataverse (linked to journal article)
- Databrary.org (esp. video/audio, identifiable data)
- ICPSR
- OSF.io
- PSU ScholarSphere
- PSU Data Commons
- NIH Data Archive
- Zenodo.org

Alternatives to repositories

- Supplemental materials
- Lab/project website

Data journal

List of data journals

Kindling, Maxi; Strecker, Dorothea

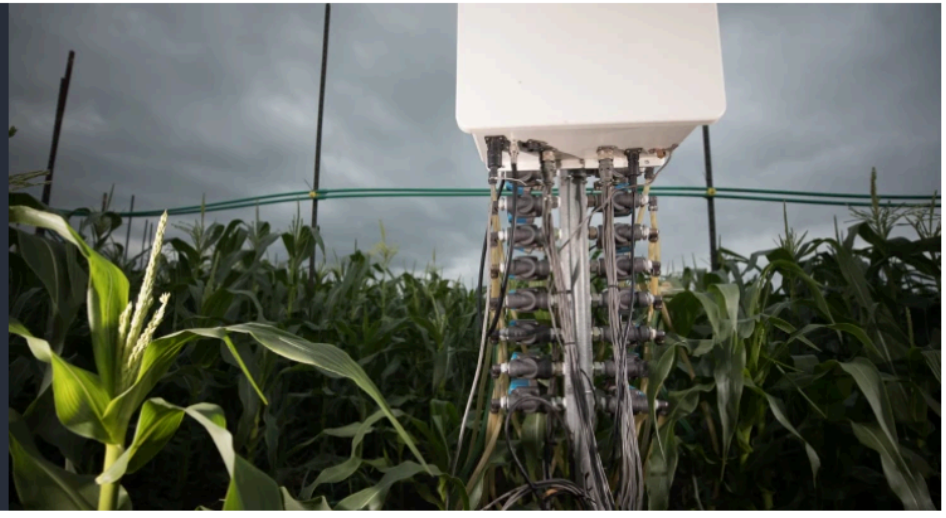
This document describes a dataset that aggregates information about 135 data journals.
Data journals focus on the publication of data papers -- a specialized publication type describing datasets, their collection and reuse potential that is peer-reviewed, citable and indexed.
This dataset includes a comprehensive list of data journals that was compiled by aggregating existing sources, as well as an overview of these sources.

The list is continually updated on GitHub, where additional information on data journals (URLs of data journal homepages) is provided: https://github.com/MaxiKi/data-journals

Preview

issn	journal_title	publisher	data_journal_type
2363-4952	RIDE - A review journal for digital editions and resources	Institut für Dokumentologie und Editorik	pure
2059-481X	Journal of Open Humanities Data	Ubiquity Press	pure
2296-7745	Frontiers in Marine Science	Frontiers	mixed
2054-345X	Human Genome Variation	Springer Nature	mixed
2032-6378	Journal of Astronomical Data	Vrije Universiteit Brussel	pure
2603-	Viticulture Data Journal	Pensoft Publishers	pure

Two decades of fumigation data from the Soybean Free Air Concentration Enrichment facility



Elise Kole Aspray, Timothy A. Mies ... Elizabeth A. Ainsworth

Data Descriptor | 20 April 2023

Announcements

Collection open for submissions

Scientific Data is open to submissions for this special collection: Meteorology and hydroclimate observations and models

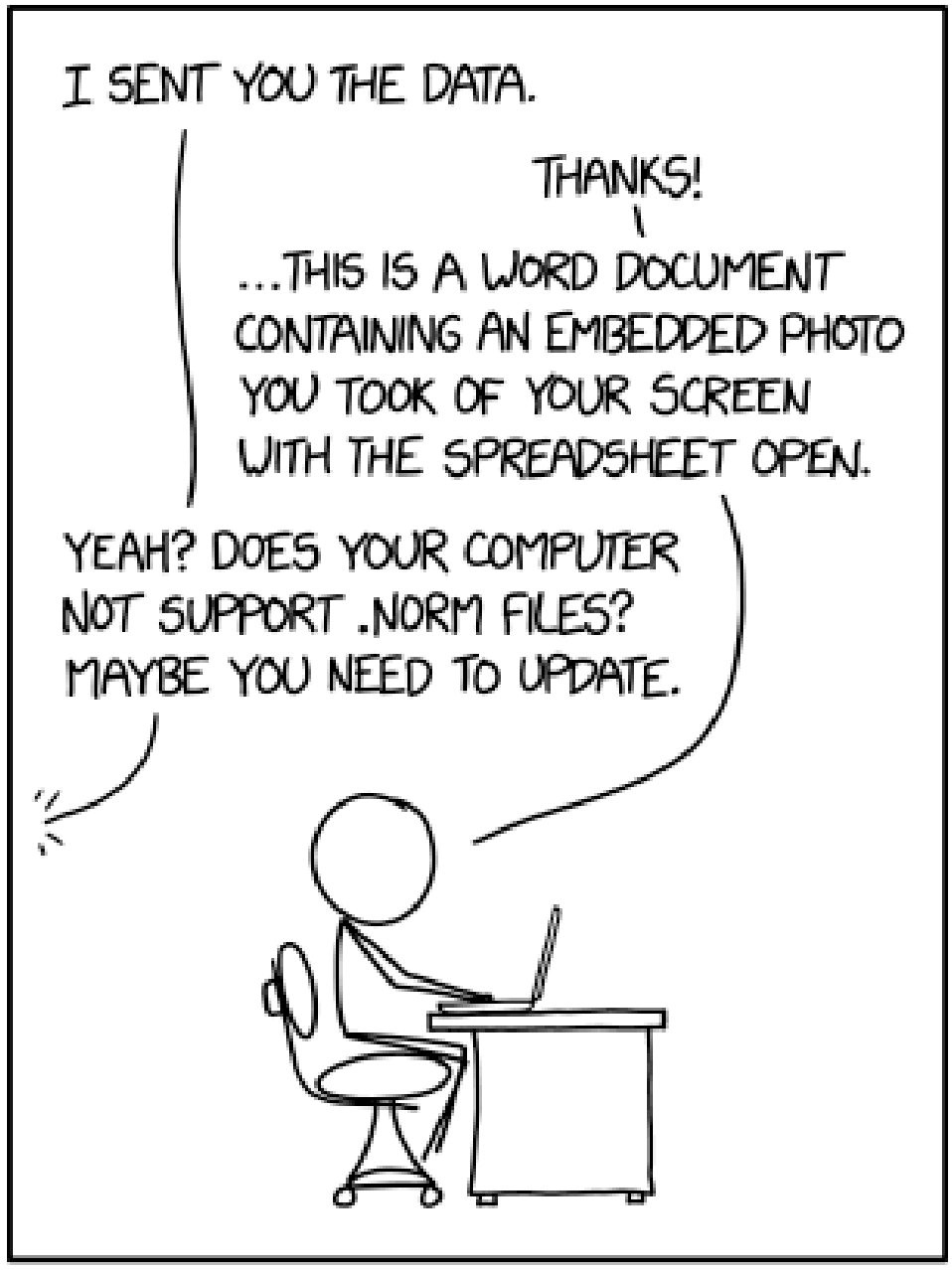
Open for submissions

www.nature.com/sdata/

n.b.– Gilmore is on editorial board.

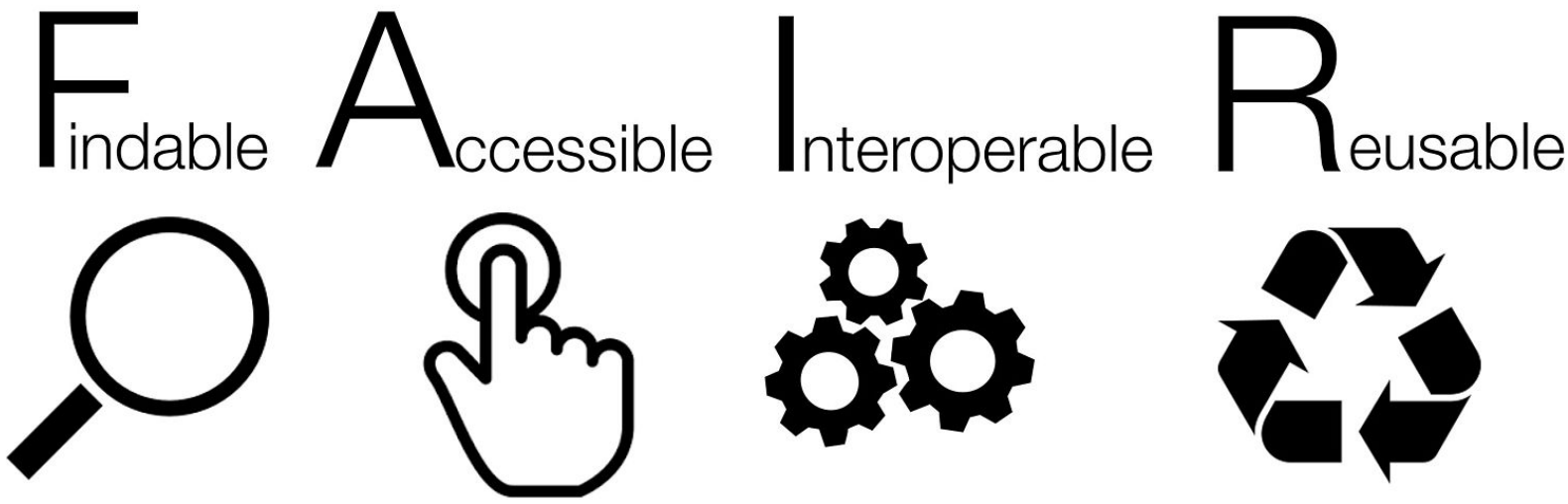
- Could be extra pub
- Especially good for early-career researchers

How to share data



SINCE EVERYONE SENDS STUFF THIS WAY ANYWAY, WE SHOULD JUST FORMALIZE IT AS A STANDARD.

FAIR principles



(Wilkinson et al. 2016)

FAIR in-practice

- **F**indable
  - Repository or journal with persistent identifier
- **Accessible**
  - Not behind paywall
- **I**nteroperable
  - {.txt, .csv} vs. {.docx, .xlsx}
- **R**eusable
  - Data definitions/dictionary
  - Code

CARE principles

Figure 5: (Carroll et al. 2021)

Seek permission to share

- Sensitive & identifiable data can (often) be shared with *restricted audiences* (i.e., researchers)
  - e.g., Databrary (databrary.org)
- Qualitative Data Repository (QDR; qdr.syracuse.edu)

Final thoughts

- Sharing $\neq$ public
- Plan for sharing as early as possible
- Data + code $\rightarrow$ reproducible analyses

Resources

About

This talk was produced using Quarto.

The source files are in R and R Markdown, then rendered to PDF slides in the beamer format. The PDF slides are hosted in a GitHub repo:

https://github.com/penn-state-open-science/bootcamp-2023-data-sharing

References

Bhattacharjee, Yudhijit. 2013. "The Mind of a Con Man." *The New York Times*, April. https://www.nytimes.com/2013/04/28/magazine/diederik-stapels-audacious-academic-fraud.html.

Brakewood, Beth, and Russell A Poldrack. 2013. "The Ethics of Secondary Data Analysis: Considering the Application of Belmont Principles to the Sharing of Neuroimaging Data." *NeuroImage* 82 (November): 671–76. https://doi.org/10.1016/j.neuroimage.2013.02.040.

Carroll, Stephanie Russo, Edit Herczog, Maui Hudson, Keith Russell, and Shelley Stall. 2021. "Operationalizing the CARE and FAIR Principles for Indigenous Data Futures." *Scientific Data* 8 (1): 108. https://doi.org/10.1038/s41597-021-00892-0.

Cho, Adrian. 2012. "Once Again, Physicists Debunk Faster-Than-Light Neutrinos." *Science*. https://doi.org/10.1126/article.27262.

Goodman, Steven N, Daniele Fanelli, and John P A Ioannidis.