Data Practices

Data Practices

An empirical view of what people creating, analyzing, and managing data *actually do*.
(or would do)
so that we can improve efficiency and reliability

V1. Data Practices	(how do we know what works?)
V2: What's going on in the lab?	(brace yourself; it ain't pretty)
V3: Data sharing	(no, no, no, no, no. It's *mine*!)
V4: Data Reuse	(if you didn't make it, it is hard to use it)

V2: What's going on in the lab?

Empirical extraction of vocabulary and processes

Empirical identification of bad behavior:

- metadata?? (for retrieval? use? interpretation? preservation? credit? reproducibility?)
- code documentation?
- code testing?
- workflow documentation?
- provenance availability?

and so on

Incentives to do better?

The problem (we're human, all too human)

Quasi-empirical studies

Much of the global analysis of research processes, data lifecycles, and data curation is

basically empirical,

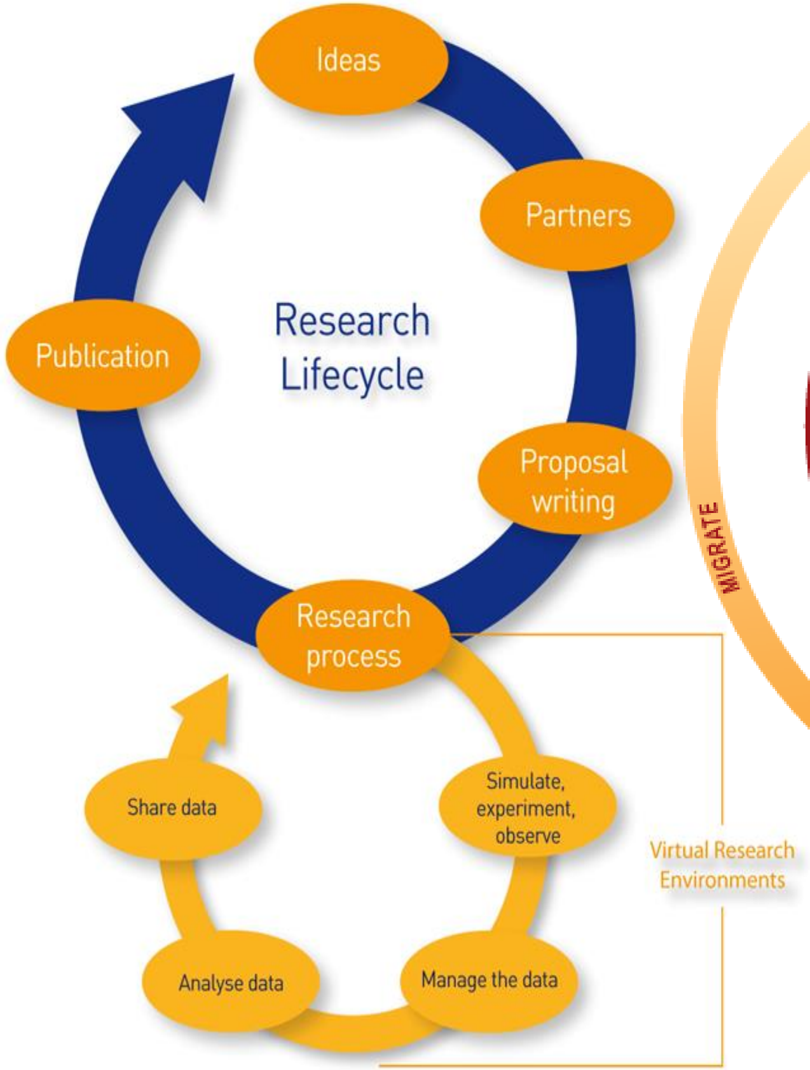
but at the same time casual, not rigorous

That does mean it is wrong;

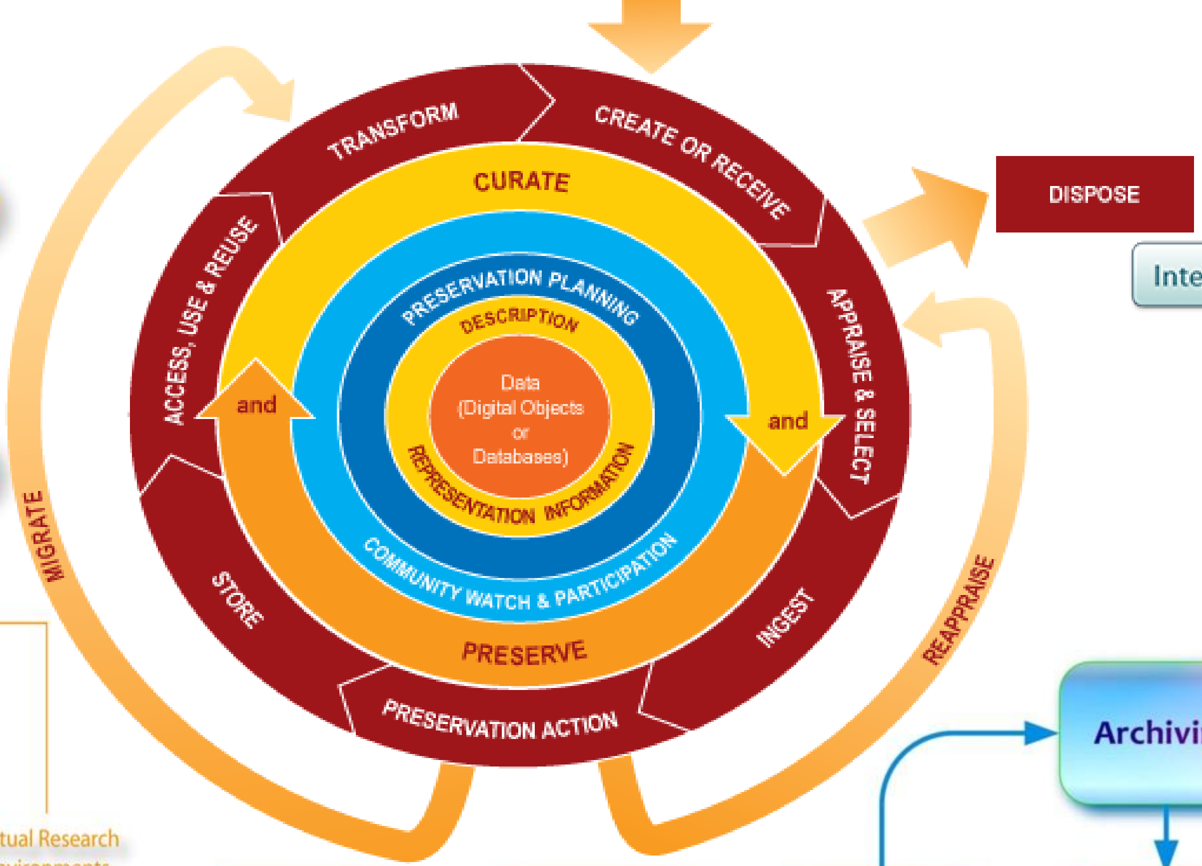
on the contrary: we don't always need a rigorous designed study

For instance:

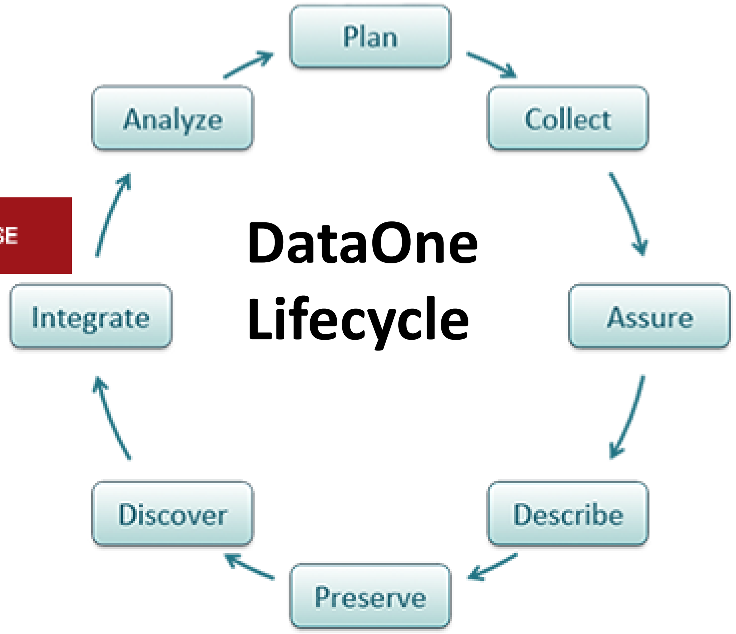
JISC/VRE Lifecycle



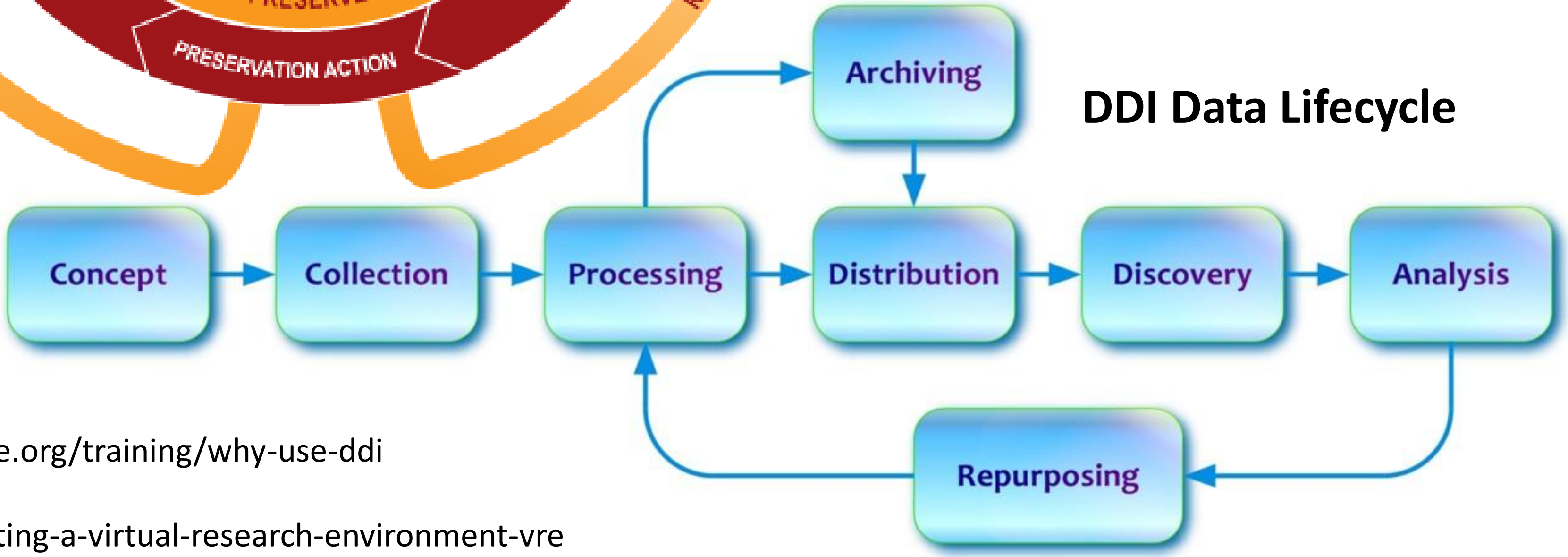
DCC Lifecycle



DataOne Lifecycle



DDI Data Lifecycle



Data Documentation Initiative. http://www.ddialliance.org/training/why-use-ddi
DataONE. https://www.dataone.org/data-life-cycle
JISC/VRE: https://www.jisc.ac.uk/full-guide/implementing-a-virtual-research-environment-vre

An empirically derived typology of research data practices

Designing research

Managing data

Generating and collecting

Processing

Analyzing, interpreting, and abstracting

Representing data

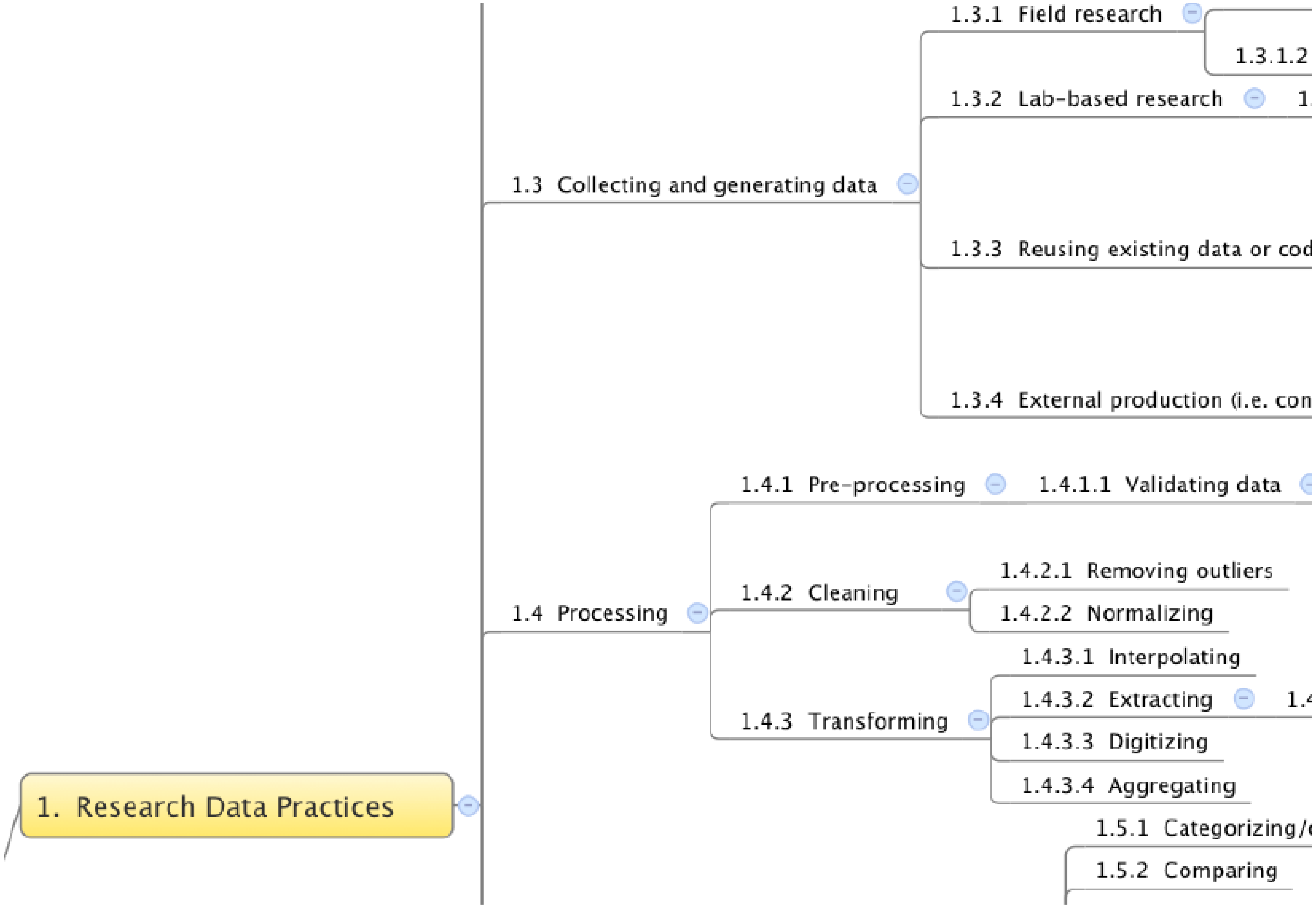
Sharing data and products

Attributing and citing data

Publishing data

Chao, T. C., Cragin, M. H., & Palmer, C. L. (2014). Data Practices and Curation Vocabulary (DPCVocab): An empirically derived framework of scientific data practices and curatorial processes. JASIST.

A fragment of an empirically derived “data practices and curation vocabulary” (DPCVocab)

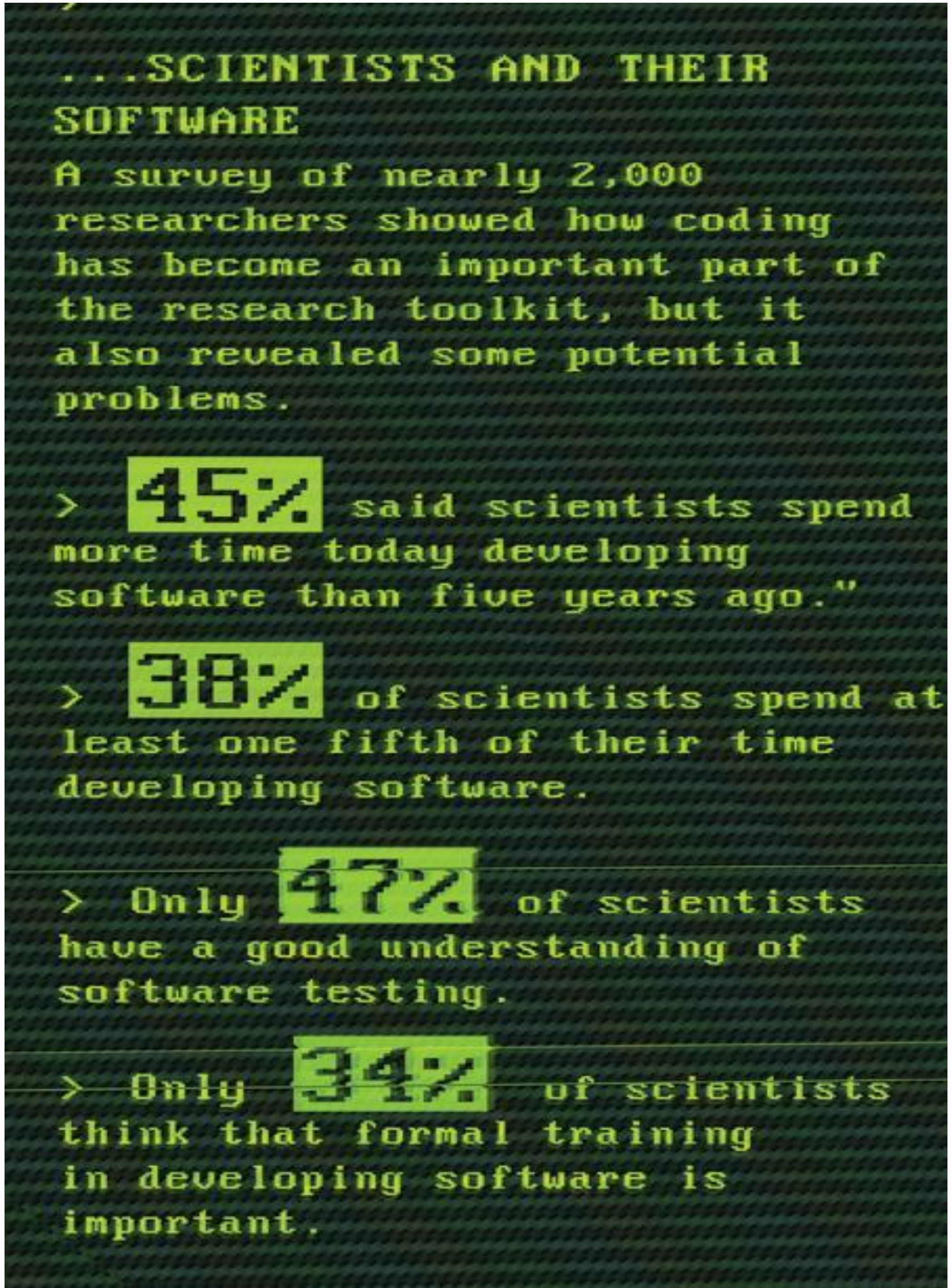


Chao, T. C., Cragin, M. H., & Palmer, C. L. (2014). Data Practices and Curation Vocabulary (DPCVocab): An empirically derived framework of scientific data practices and curatorial processes. JASIST. DOI: 10.1002/asi.

Discovering (ok, confirming) bad behavior

"As a general rule, researchers do not test or document their programs rigorously, and they rarely release their code, making it almost impossible to reproduce and verify published results generated by scientific software, say computer scientists. ... scientists often lack these communication and documentation skills"

— Zeya Merali "Computational science: ...Error . Why scientific programming does not compute"
Nature news feature (2010);
http://www.nature.com/news/2010/101013/full/467775a.html



Metadata failures

What metadata do you currently use to describe your data, if any?

Standards	2014 Responses
DC (Dublin Core)	7.1%
DwC (Darwin Core)	2.0%
DIF (Directory Interchange Format)	1.7%
EML (Ecological Metadata Language)	9.3%
FGDC (Federal Geographic Data Committee)	8.5%
ISO 19115 (Geographic Information-Metadata)	10.2%
OGIS (Open GIS)	7.2%
Standard within my lab	16.7%
Other	8.6%

None: 47.9%

(Tenopir et al., 2014)

Data storage

How much of your data do you currently store in the following locations?

	Most or all of my data
External hard disk/drive storage	83.3%
On my personal computer	65.3%
Dropbox/Google/Figshare/Cloud	57.2%
On my institution's server	37.7%
On the PI's server	28.4%
On a departmental server	23.1%
On paper in my office	13.7%
In my institution's repository	11.3%
In a domain repository	9.5%
Other data repository or archive	9.3%
In a publisher repository	2.4%

(Tenopir et ala., 2014)

Why is it so hard to be good?

We don’t need a behavioral economist
to tell us that we have have a hard time giving up short-term benefits for long-term benefits,
even when the long-term benefits are greater.

And that’s when the benefits accrue to *ourselves* (our future selves).

How much harder it is when much of the benefit accrues to *others*

This is why we don’t document code, test our code, add metadata to datasets, use standards, backup our files, avoid transformations at the command line, etc. etc.

Often the benefits seems indirect and elusive, and we can convince ourselves it is unnecessary

[“No time to document this, but no need either: how it works is self-evident.
And no need to test it: we were careful.
And no need to back up an earlier version; this one is better,
and I don’t think we used that earlier version for anything important . . . (or did we?)”]

Incentives for good data practices …?

Scientific value
- Better analysis and research outcomes

Credit
- Credit for data producers (metadata)
- Data sharing = increased citations (Pinowar, 2007)

Infrastructure
- Interoperable applications, systems, and data
- Reliability and reproducibility
- Efficiency
- Easier collaboration

Tenure and promotion assessment
- Measure of being a good data steward

???

References (General)

Ball, A. (2010). Data lifecycles. In Review of the State of the Art of the Digital Curation of Research Data. Project Report. Bath, UK: University of Bath.

Borgman, C.L. (2012). The conundrum of sharing research data. *Journal of the American Society for Information Science & Technology,* 63(6): 1059-1078.

Babeu, A. (2011). "Rome Wasn't Digitized in a Day": Building a Cyberinfrastructure for Digital Classics. Washington DC.

Chao, T. C., Cragin, M. H., & Palmer, C. L. (2014). Data Practices and Curation Vocabulary (DPCVocab): An empirically derived framework of scientific data practices and curatorial processes. JASIST.

Hanson, K., Surkis, A., & Yacobucci, K. (2012). Data Sharing and Management Snafu in 3 Short Acts [video].

Hey, A. J., Tansley, S., & Tolle, K. M. (2009). The fourth paradigm: data-intensive scientific discovery.

Cragin, M.H., Palmer, C.L., Carlson, J.R., & Witt, M. (2010). Data sharing, small science, and institutional repositories. *Philosophical Transactions of the Royal Society A*, *368*(1926), 4023-4038.

Pepe, A., Goodman, A., Muench, A., Crosas, M. & Erdmann, C. (2014). How Do Astronomers Share Data? Reliability and Persistence of Datasets Linked in AAS Publications and a Qualitative Study of Data Practices among US Astronomers. *PLoS ONE,* 9(8): e104798.

Research Information Network. (2008). To Share or Not to Share: Publication and Quality Assurance of Research Data Outputs.

Tenopir, C., Allard, S., Douglass, K., Aydinoglu, A. U., Wu, L., Read, E., Manoff, M., et al. (2011). Data sharing by scientists: Practices and perceptions. PloS ONE, 6(6), e21101.

Tenopir C, Dalton ED, Allard S, Frame M, Pjesivac I, Birch B, et al. (2015) Changes in Data Sharing and Data Reuse Practices and Perceptions among Scientists Worldwide.

References (Lifecycle models)

Ball, A. (2010). Data lifecycles. In Review of the State of the Art of the Digital Curation of Research Data. Project Report. Bath, UK: University of Bath.

DataONE. Data Lifecycle Model. https://www.dataone.org/data-life-cycle

Digital Curation Centre. DCC Curation Lifecycle Model. http://www.dcc.ac.uk/resources/curation-lifecycle-model

ICPSR. (2012). Guide to Social Science Data Preparation and Archiving. 5th Edition. https://www.icpsr.umich.edu/icpsrweb/content/deposit/guide/

UCF Libraries. Resarch Lifecycle. http://library.ucf.edu/about/departments/scholarly-communication/research-lifecycle/

UK Data Archive. Create and Manage Data – Research Data Lifecycle. http://data-archive.ac.uk/create-manage/life-cycle

USGS. The Data Lifecycle. https://www2.usgs.gov/datamanagement/why-dm/lifecycleoverview.php