

# Research Data Management - Introduction

Karl Benedict  
Director of Research Data Services & Associate  
Professor, College of University Libraries & Learning  
Sciences  
University of New Mexico

[kbene@unm.edu](mailto:kbene@unm.edu)

# Outline

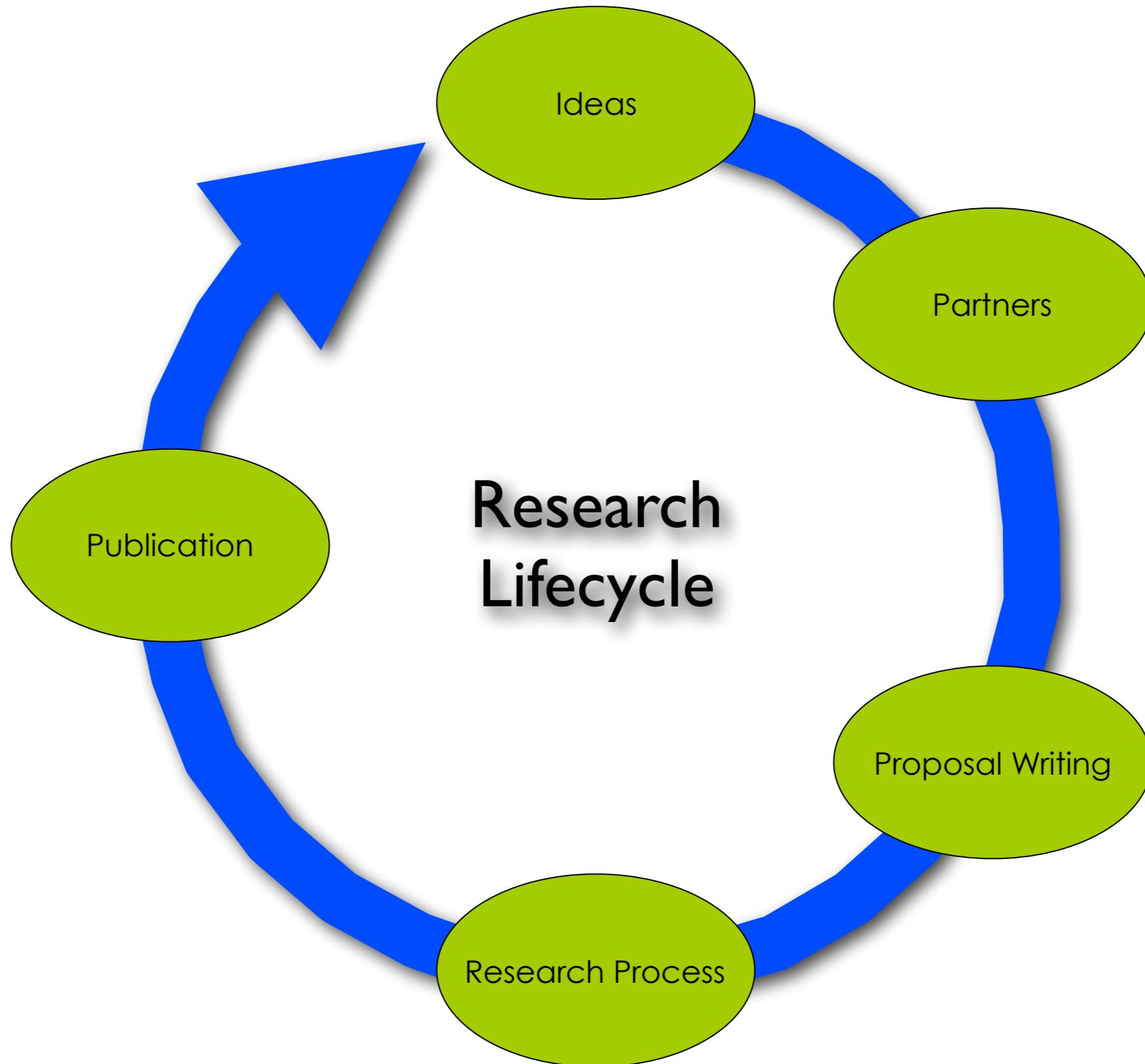
- Context
- Building
- “Data” and “Documentation”
- Some Definitions
- Some Recommendations
- An Analysis
- A Process

# Context

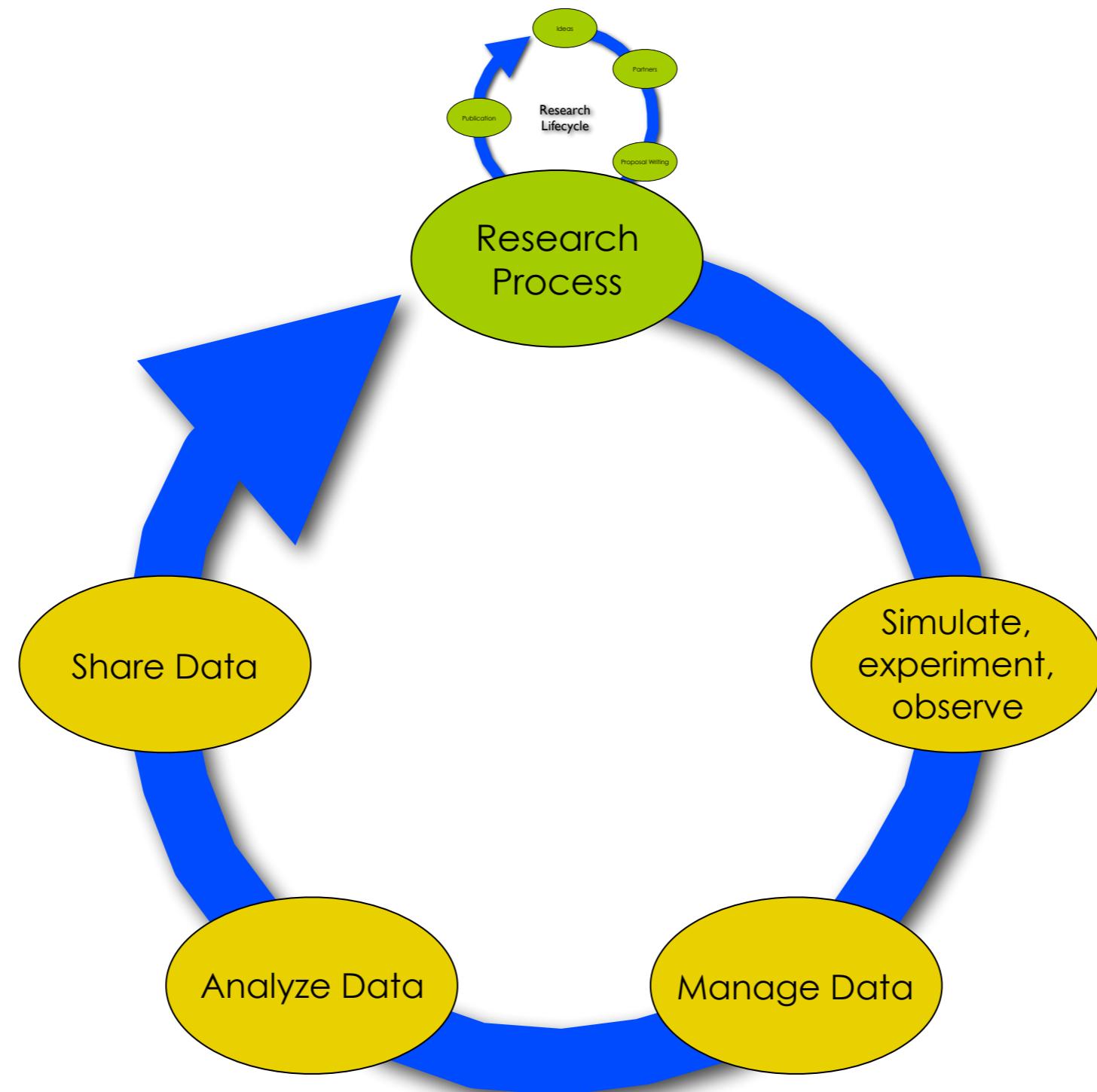
# Data Management Requirements

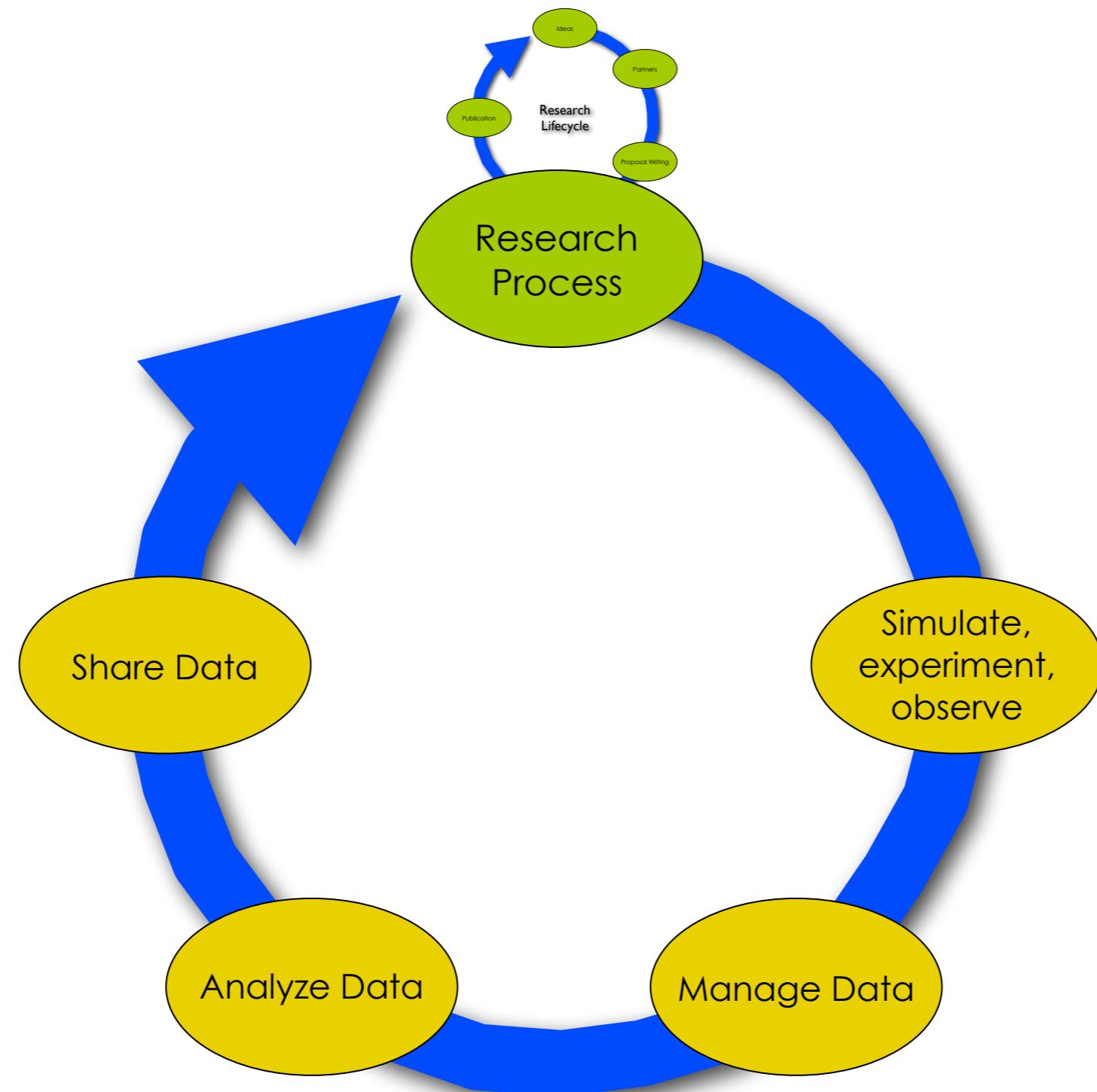
- Data Management Plans required by funding agencies
- Publications requiring links to supporting data
- Increasing collaborative research where data must be shared

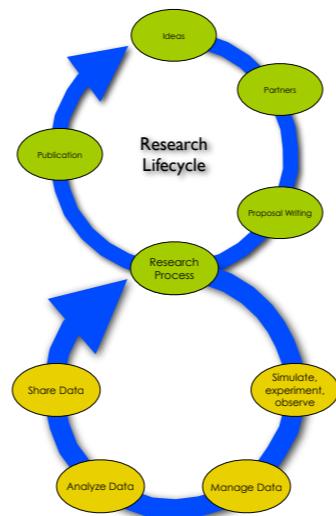
# A Cross-walk between the Research and Data Lifecycles

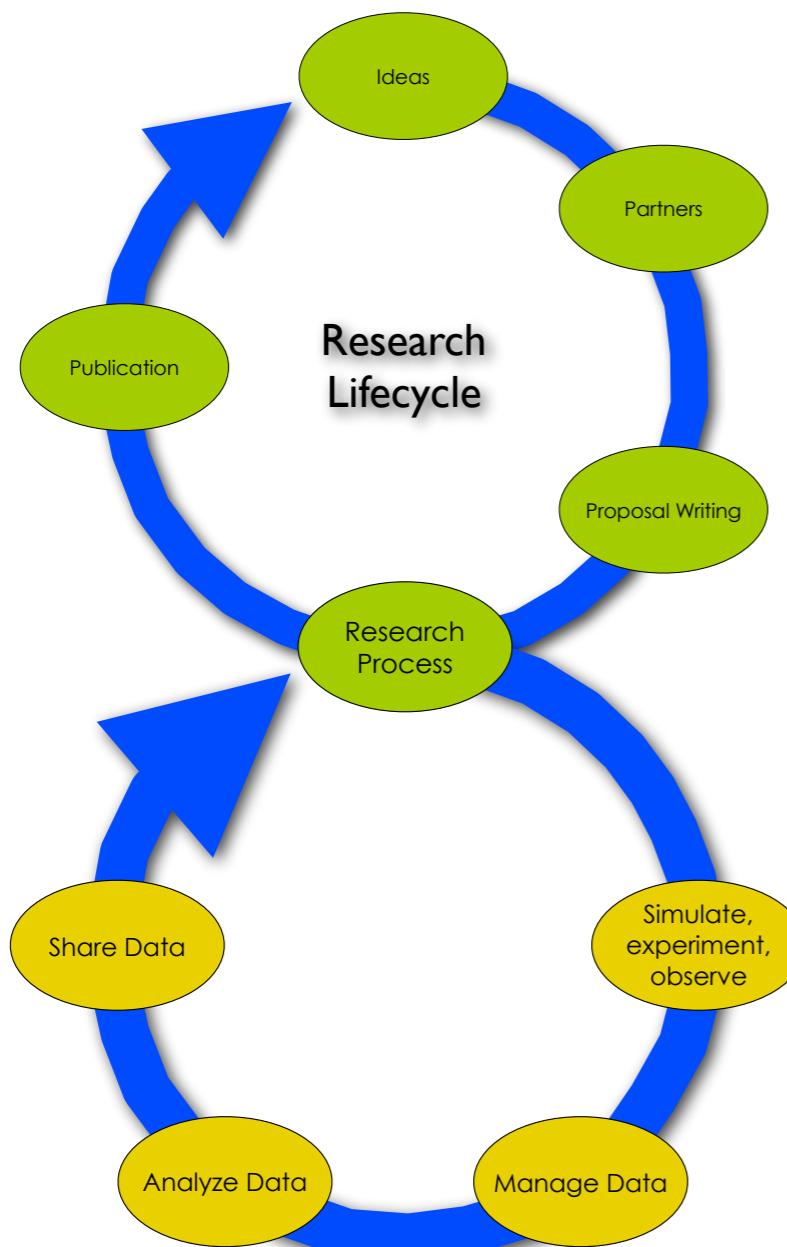






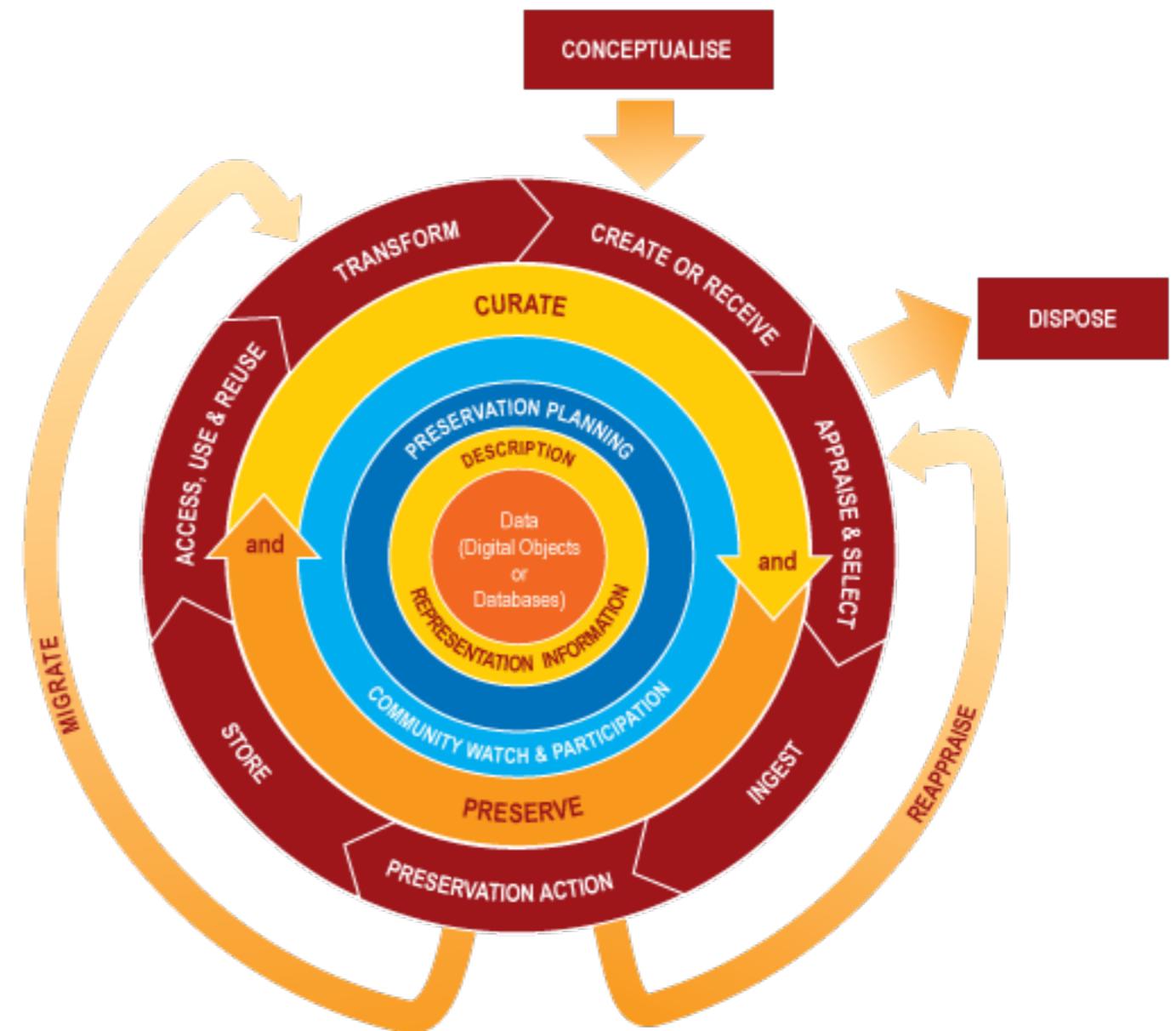






## JISC Research Lifecycle

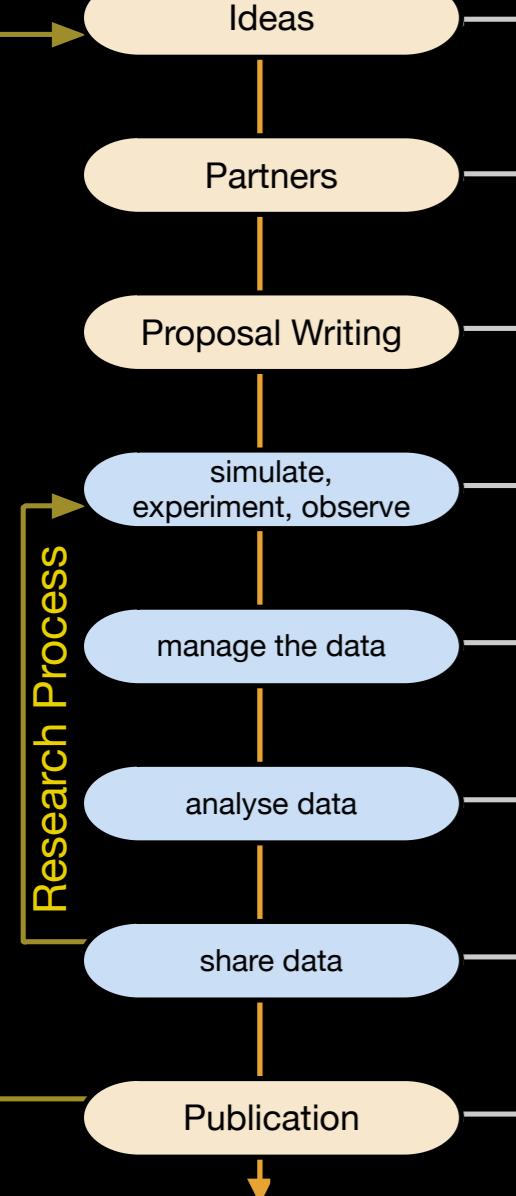
<http://www.jisc.ac.uk/whatwedo/campaigns/res3/jischelp.aspx>



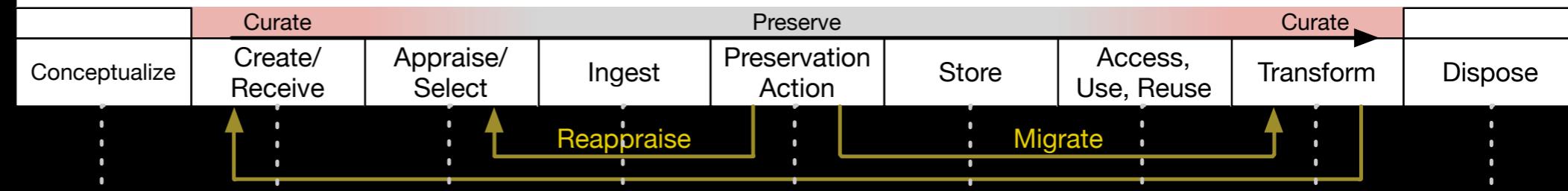
## DCC Data Curation Lifecycle

<http://www.dcc.ac.uk/resources/curation-lifecycle-model>

## Research Lifecycle



## Data Lifecycle



# Building



Photo by:  
Patrick Q

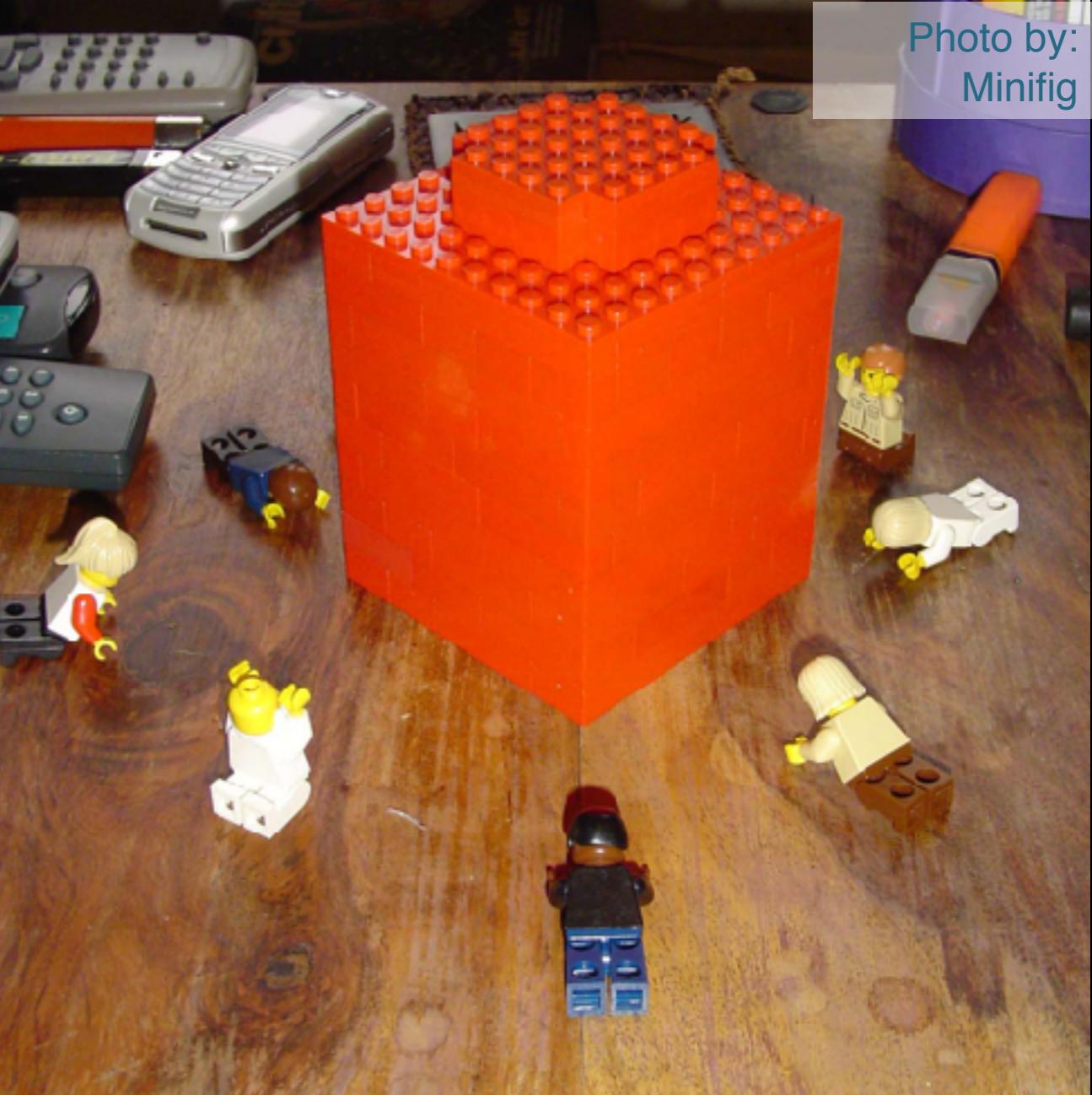


Photo by:  
Minifig

10 minutes  
Choose 15 pieces  
Place a being at the top  
Build for height (*impact*)  
Build for stability (*support*)

10:00



*Who “Won”?*

*Why?*

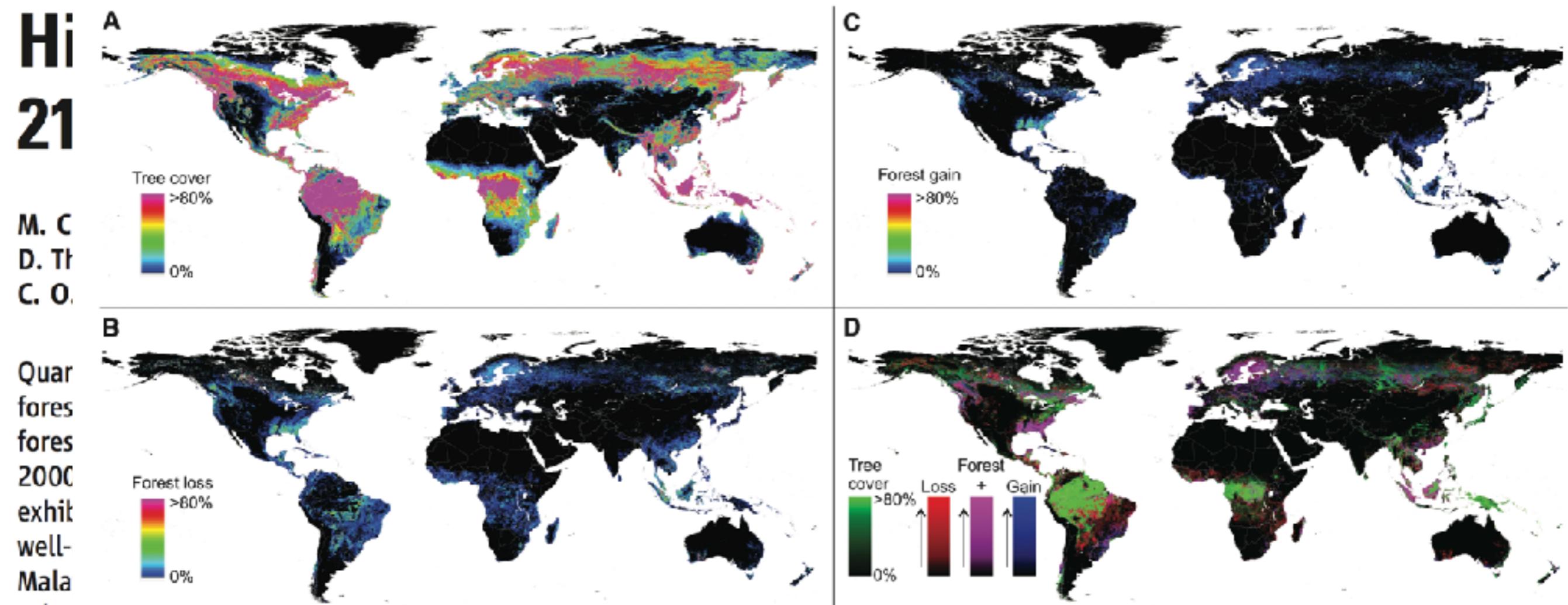
“Data” and  
“Documentation”



<http://www.youtube.com/watch?v=N2zK3sAtr-4&feature=share&list=FLdHSa0dF8hj5B-x4OaIBPWQ>

You are starting a new study, and you find a publication that is based on data key to your analysis ...

# Scenario 1



**Fig. 1. (A) Tree cover, (B) forest loss, and (C) forest gain.** A color composite of tree cover in green, forest loss in red, forest gain in blue, and forest loss and gain in magenta is shown in (D), with loss and gain en-

hanced for improved visualization. All map layers have been resampled for display purposes from the 30-m observation scale to a  $0.05^\circ$  geographic grid.

## High-Resolution Global Maps of 21st-Century Forest Cover Change

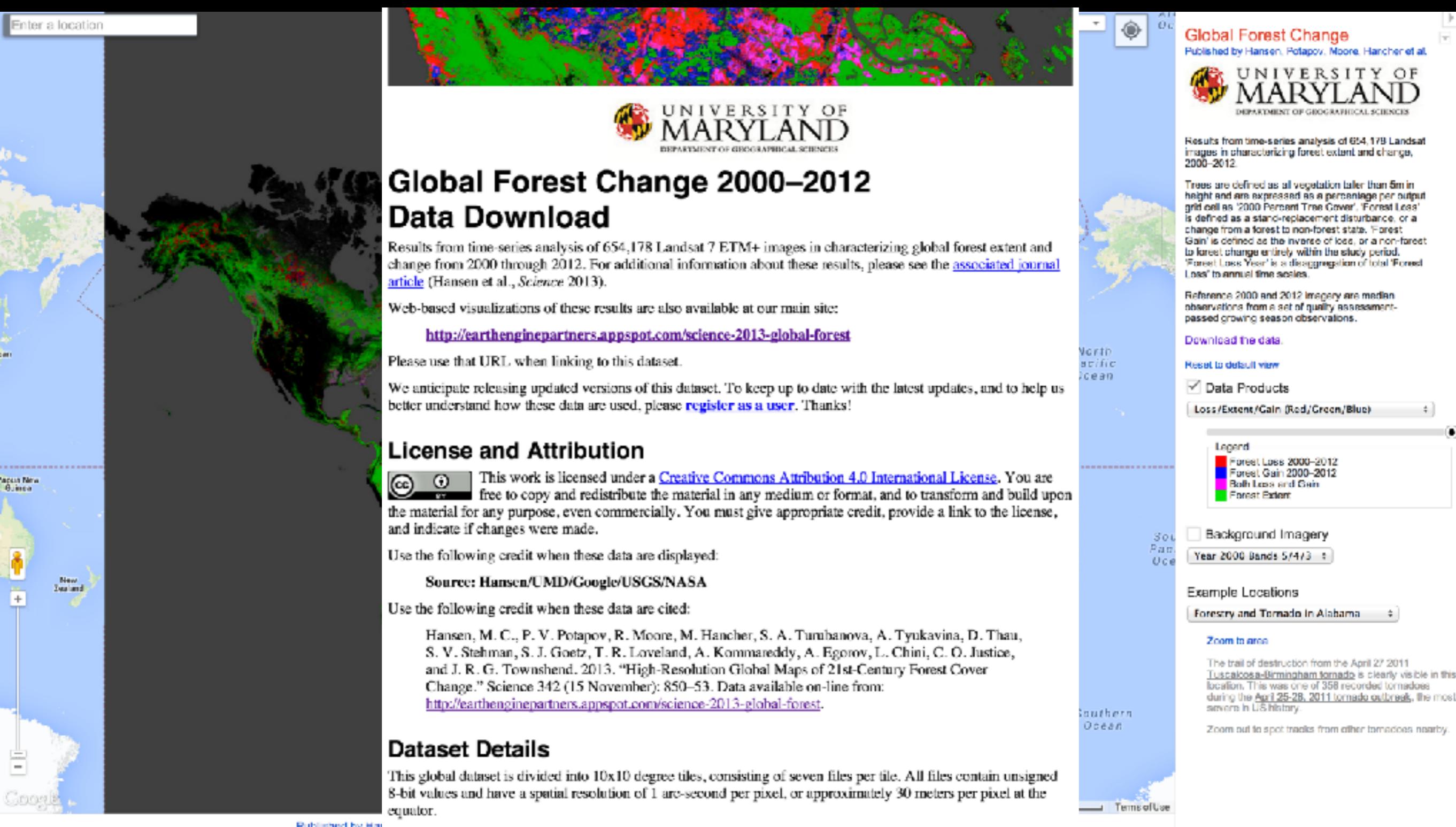
M. C. Hansen *et al.*  
Science 342, 850 (2013);  
DOI: 10.1126/science.1244693

# Questions . . .

Rate from 1 (impossible) to 5 (easy) the following

	1 (impossible)	2	3	4	5 (easy)
Data Discovery					
Access					
Understanding					
Use					

# Scenario 2



# Questions . . .

Rate from 1 (impossible) to 5 (easy) the following

	1 (impossible)	2	3	4	5 (easy)
Data Discovery					
Access					
Understanding					
Use					

discussion

# Some Definitions

# Data

*Data.* For the purposes of this document, data are any and all complex data entities from observations, experiments, simulations, models, and higher order assemblies, along with the associated documentation needed to describe and interpret the data.

- National Science Foundation (2007). *Cyberinfrastructure Vision for 21st Century Discovery*. National Science Foundation, Cyberinfrastructure Council. Washington, DC.  
<http://www.nsf.gov/pubs/2007/nsf0728/nsf0728.pdf>. pg. 22

# Documentation (AKA Metadata)

*Metadata.* Metadata are a subset of data, and are data about data. Metadata summarize data content, context, structure, interrelationships, and provenance (information on history and origins). They add relevance and purpose to data, and enable the identification of similar data in different data collections.

- National Science Foundation (2007). *Cyberinfrastructure Vision for 21st Century Discovery*. National Science Foundation, Cyberinfrastructure Council. Washington, DC. <http://www.nsf.gov/pubs/2007/nsf0728/nsf0728.pdf>. pg. 22

# Embargo

*Embargo.* A period during which access to research data is not allowed to certain types of users. This is either to protect the revenue of the publisher or (more generally) to protect the interests of other parties (for example, partner research organizations).

# License

*A licence in this context is a legal instrument for a rights holder to permit a second party to do things that would otherwise infringe on the rights held. The first thing to note is that only the rights holder (or someone with a right or licence to act on their behalf) can grant a licence; it is therefore imperative that the intellectual property rights (IPR) pertaining to the data are established before any licensing takes place.*

*How to License Research Data.* Digital Curation Centre.

<http://www.dcc.ac.uk/resources/how-guides/license-research-data#x1-20002>

# Some Recommendations

I'M EXHAUSTED FROM  
ALL OF THE BASIC  
RESEARCH I'M DOING.



IT'S TOO BAD THAT  
THE VALUE OF MY WORK  
WON'T BE QUANTIFIABLE  
FOR ANOTHER TEN  
YEARS.



I'D  
LIKE  
TO SEE  
YOUR  
LAB  
REPORT.



SO... THE  
NEW RULE  
IS THAT WE  
WRITE DOWN  
STUFF?

Dilbert.com DilbertCartoonist@gmail.com

© 2010 Scott Adams, Inc./Dist. by UFS, Inc.

What you need  
to know ...

*Who?*  
*What?*  
*Where?*  
*When?*  
*Why?*  
*How?*  
*Access?*



- **Who?** *Credit (researchers, sponsors), Questions, Responsibility, Role*
- **What?** *What was measured, Units, Aggregation*
- **Where?** *Geographic Location (define datum, Coordinate System, Method)*
- **When?** *Date, Time - structured, Consistent, Time Zone, Standards-based*
- **Why?** *Purpose for Data collection, Suggested Use, Known Limitations*
- **How?** *Instruments, Sensors, Algorithms, Models, Software*
- **Access?** *Licensing Terms, Embargo, Redistribution, modification*

# Organization

Define folder and file names  
and structure - and use it

Use meaningful names that  
include basic information  
(e.g. date, measurement,  
collection, etc.)

Unique

Avoid Spaces

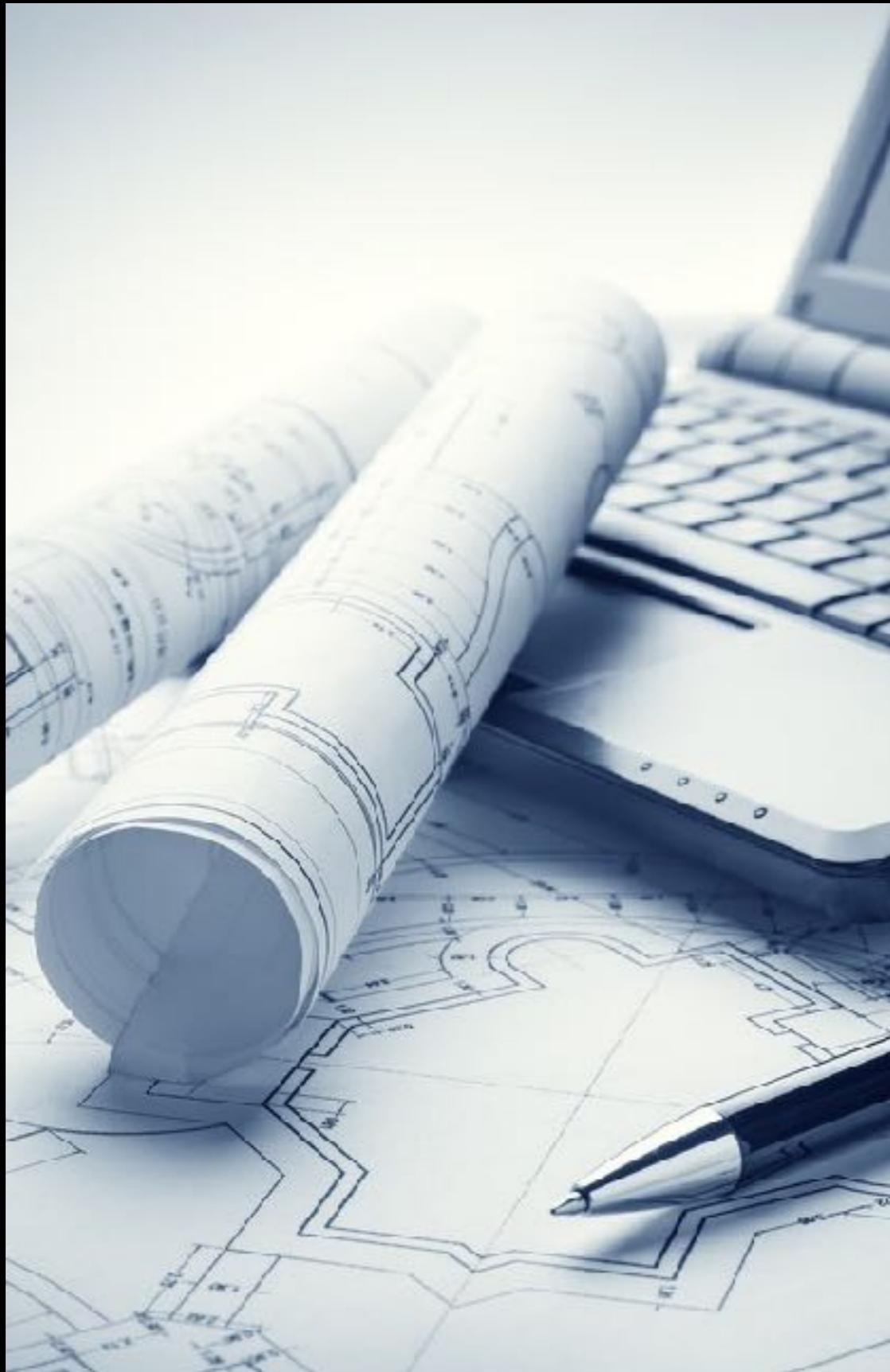
ASCII Characters only



A STORY TOLD IN FILE NAMES:

Filename	Date Modified	Size	Type
data_2010.05.28_test.dat	3:37 PM 5/28/2010	420 KB	DAT file
data_2010.05.28_re-test.dat	4:29 PM 5/28/2010	421 KB	DAT file
data_2010.05.28_re-re-test.dat	5:43 PM 5/28/2010	420 KB	DAT file
data_2010.05.28_calibrate.dat	7:17 PM 5/28/2010	1,256 KB	DAT file
data_2010.05.28_huh??.dat	7:20 PM 5/28/2010	30 KB	DAT file
data_2010.05.28_WTF.dat	9:58 PM 5/28/2010	30 KB	DAT file
data_2010.05.29_aaarrgh.dat	12:37 AM 5/29/2010	30 KB	DAT file
data_2010.05.29_#\$@*!&!.dat	2:40 AM 5/29/2010	0 KB	DAT file
data_2010.05.29_crap.dat	3:22 AM 5/29/2010	437 KB	DAT file
data_2010.05.29_notbad.dat	4:16 AM 5/29/2010	670 KB	DAT file
data_2010.05.29_woohoo!.dat	4:47 AM 5/29/2010	1,349 KB	DAT file
data_2010.05.29_USETHISONE.dat	5:08 AM 5/29/2010	2,894 KB	DAT file
analysis_graphs.xls	7:13 AM 5/29/2010	455 KB	XLS file
ThesisOutline!.doc	7:26 AM 5/29/2010	38 KB	DOC file
Notes_Meeting_with_ProfSmith.txt	11:38 AM 5/29/2010	1,673 KB	TXT file
JUNK...	2:45 PM 5/29/2010		Folder
data_2010.05.30_startingover.dat	8:37 AM 5/30/2010	420 KB	DAT file

Type: Ph.D Thesis Modified: too many times Copyright: Jorge Cham www.phdcomics.com



# Structure/Content

Consistent content

Separate data from analysis

Keep raw data separate

Focus on tabular structure for  
tabular data

Explicitly encode missing data,  
and document that encoding

Use meaningful column  
headings - while keeping short  
without spaces

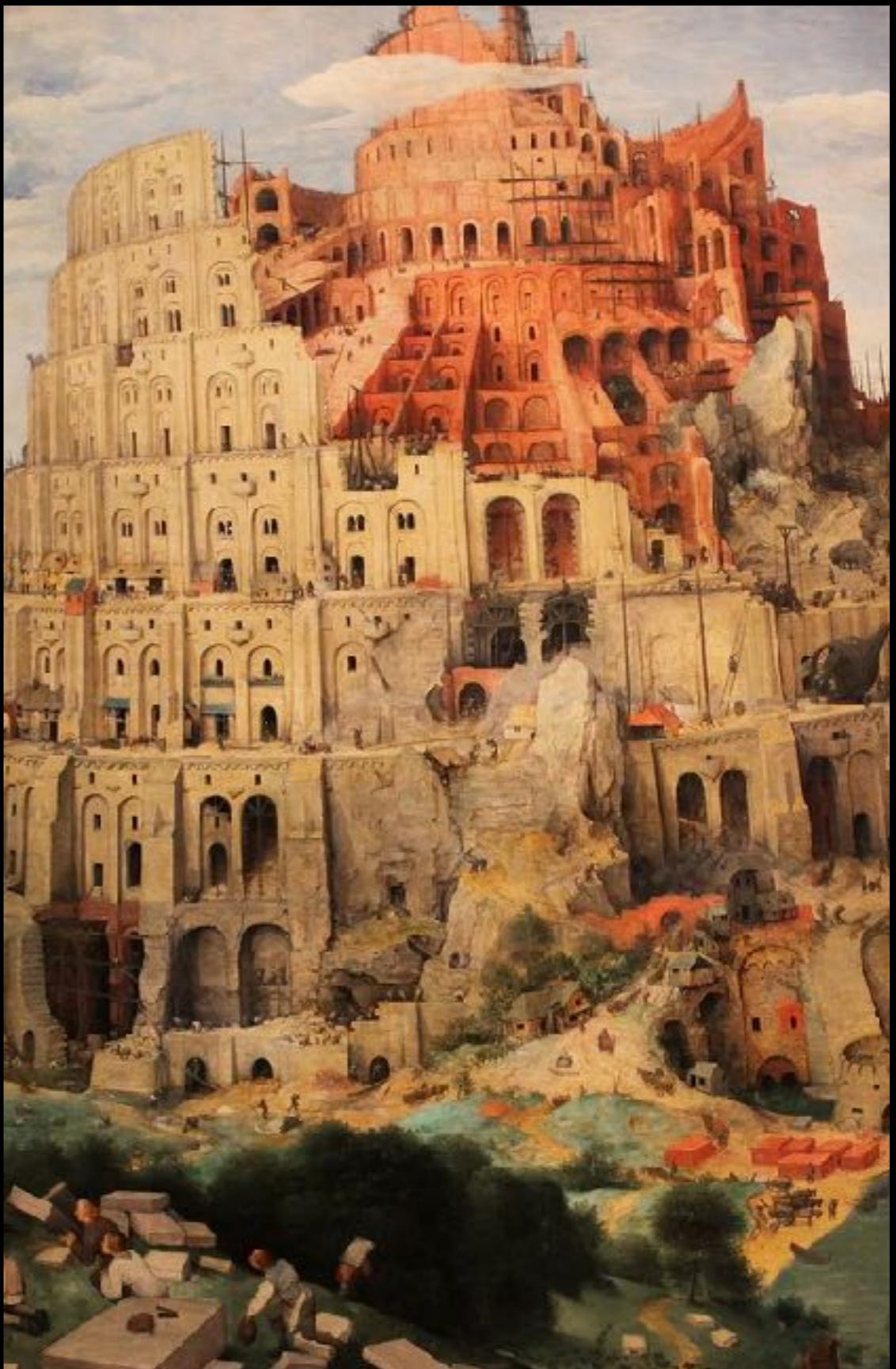
Include units

Data dictionary

# Formats

Plan for integration into an archive

Open Standards  
    >  
Proprietary ASCII  
    >  
Proprietary Binary - Documented  
    >  
Proprietary Binary



# Documentation

Many documentation standards

Machine and human readable

Commonly based on Extensible Markup Language (XML)

Wide variety of strategies/ methods/tools for creating documentation

Enables *Discovery, Use,* and *Understanding*

Who?

What?

Where?

When?

Why?

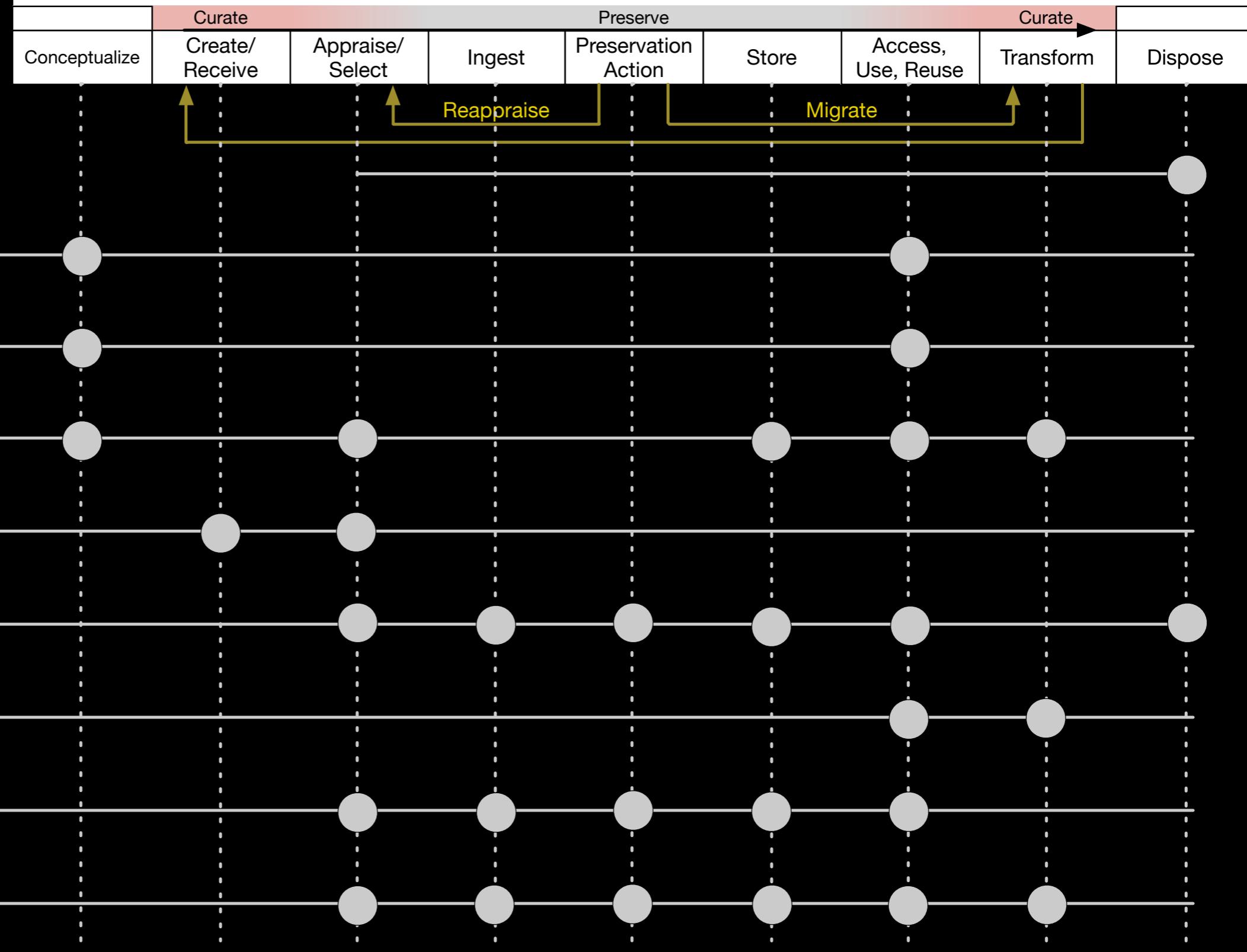
How?

Access?

# A Process

## Research Lifecycle

## Data Lifecycle



# Resources

- UNM Libraries - Research Data Services:  
Karl Benedict - [kbene@unm.edu](mailto:kbene@unm.edu)  
Jon Wheeler - [jwheel01@unm.edu](mailto:jwheel01@unm.edu)  
<http://libguides.unm.edu/data>
- Library of Congress: Sustainability of Digital Formats:  
<http://www.digitalpreservation.gov/formats/index.shtml>
- Digital Curation Centre: Disciplinary Metadata:  
<http://www.dcc.ac.uk/resources/metadata-standards>
- Library Hosted Infrastructure:  
Digital Repository: <http://digitalrepository.unm.edu>  
Lobo Drive: <http://lobodrive.unm.edu>  
LoboGit: <http://lobogit.unm.edu>
- Creative Commons  
<https://creativecommons.org/share-your-work/>

# Acknowledgements

- Unless otherwise noted, all images are from [shutterstock.com](http://shutterstock.com)
- The products and platforms upon which this work is based have been funded by
  - NSF EPSCoR Program (Track 1 [Awards: 0447691, 0814449, 1301346] and Track 2 awards [0918635, 1329470])
  - New Mexico Resource Geographic Information System
  - NASA ACCESS Program
  - University of New Mexico - University Libraries

# Questions?



Sign-In & Feedback: <https://www.surveymonkey.com/r/rds-signin>