Introduction to Research Data Management

Amanda L. Whitmire
Steven Van Tuyl
OSU Libraries

GOAL:

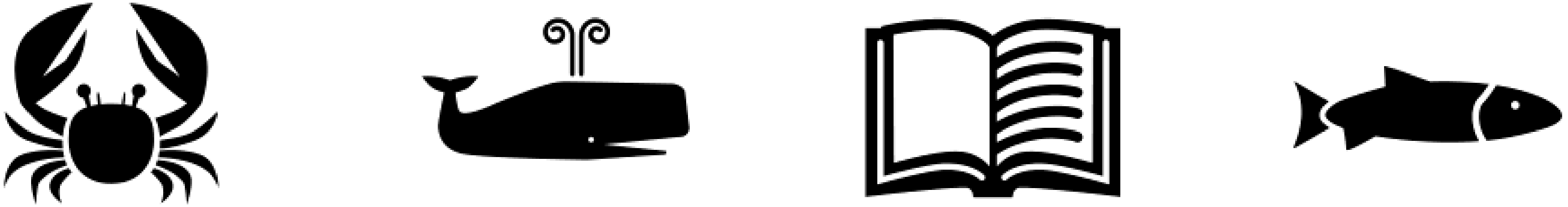
Achievable habits for implementing data management best practices into your workflow

What are data?

Research data is:

"...the recorded factual material commonly accepted in the scientific community as necessary to **validate** research findings."

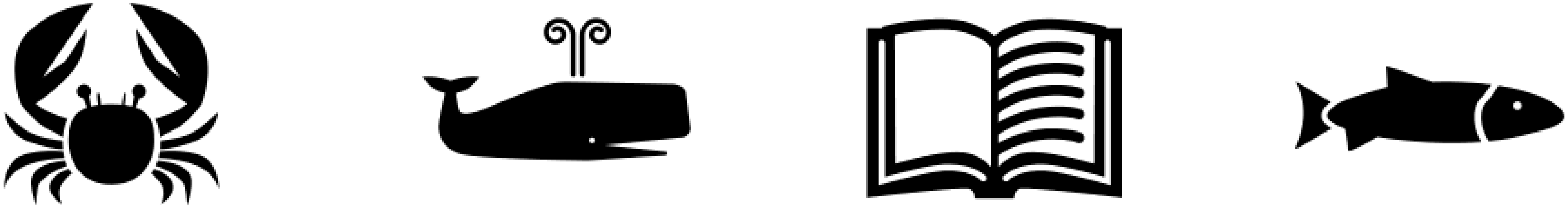
U.S. Office of Management and Budget, Circular A-110



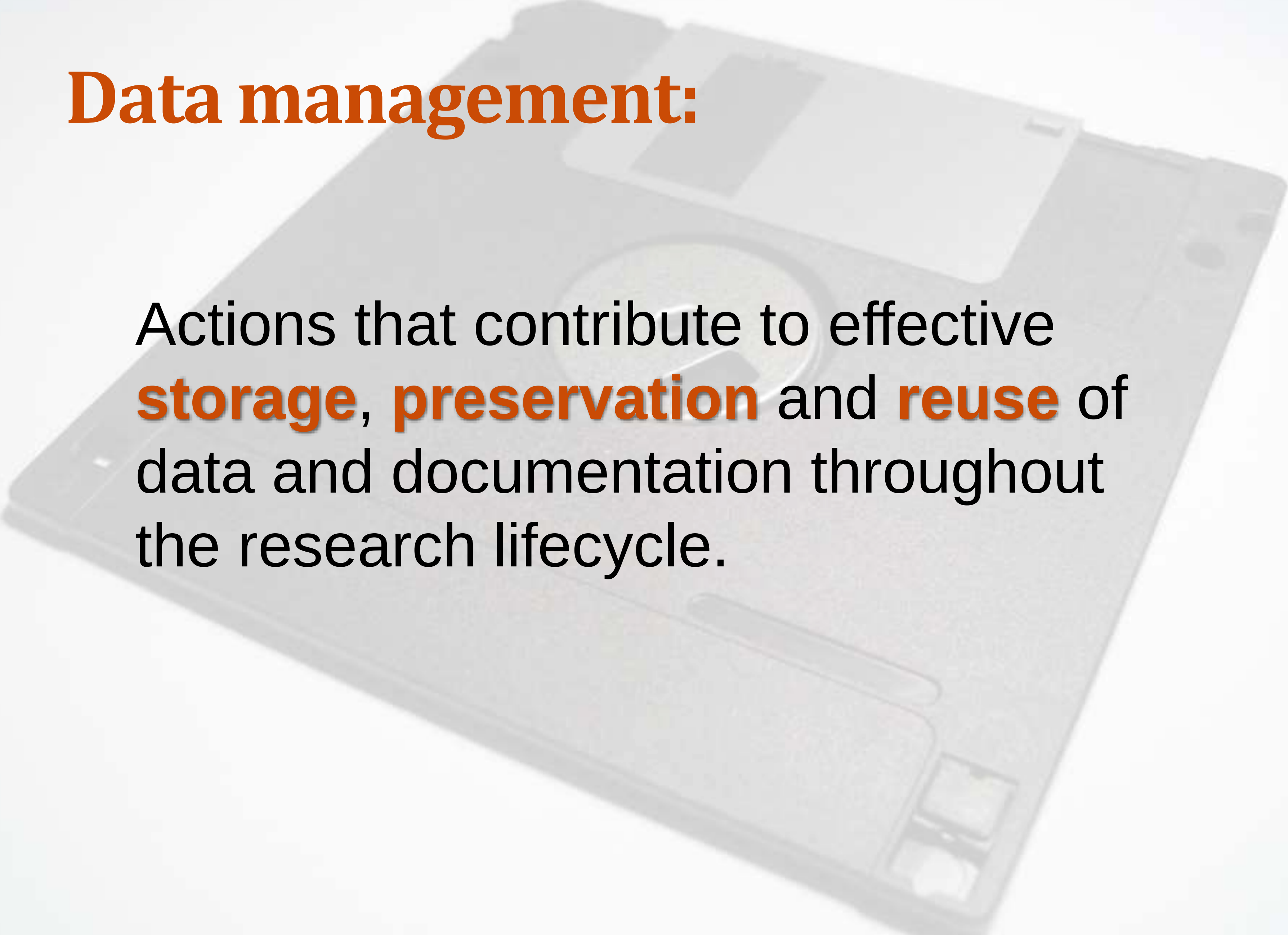
Research data is:

"Unlike other types of information, research data are collected, observed, or created, for the **purposes of analysis to produce and validate original research results**."

University of Edinburgh
MANTRA Research Data Management Training,
'Research Data Explained'

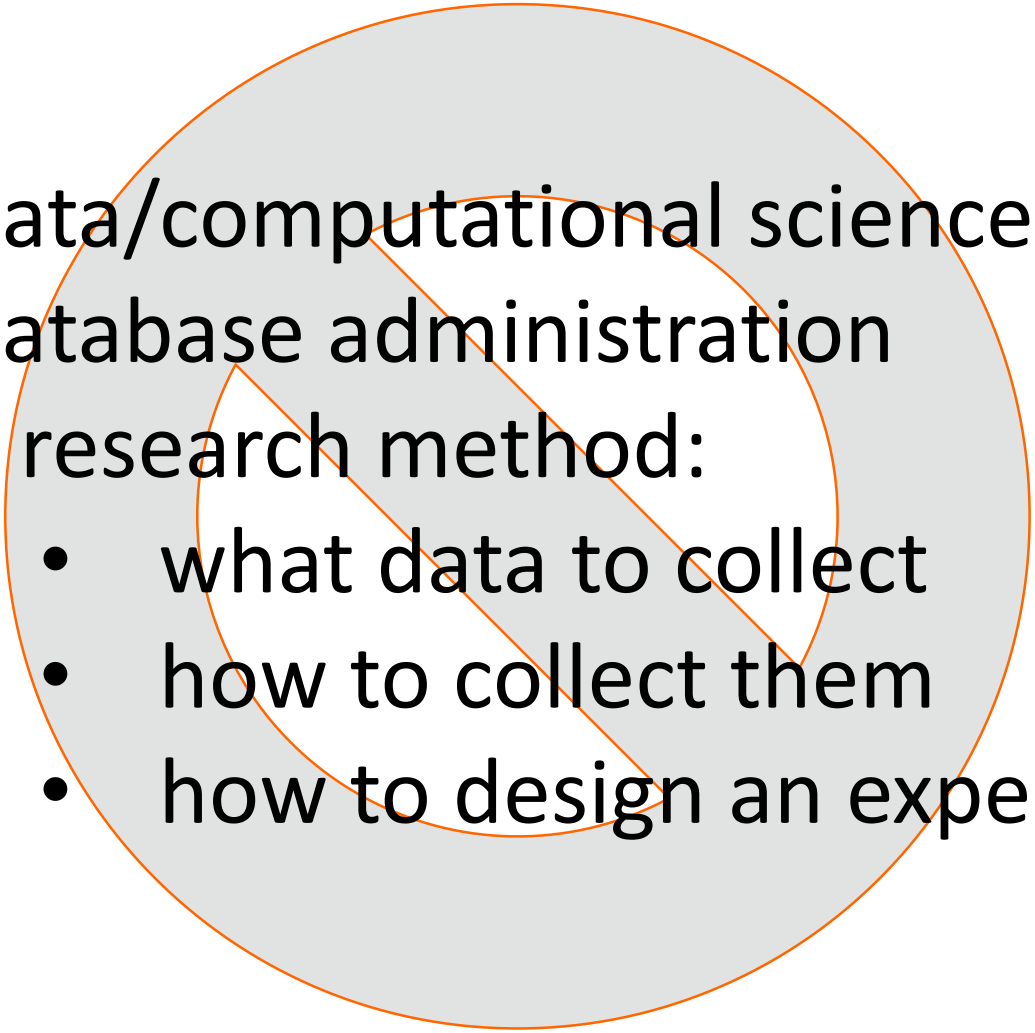


Data management:



Actions that contribute to effective **storage**, **preservation** and **reuse** of data and documentation throughout the research lifecycle.

What data management is not:



Data/computational science

Database administration

A research method:

- what data to collect
- how to collect them
- how to design an experiment

Why data management?

Further your field

Increase visibility & impact

Saves time

Protects investment

Increases research efficiency

Preservation

Funding agency requirements

image credit: http://www.flickr.com/photos/karenchiang/5041048588/

Further your field

Atul Butte
@atulbutte

New markers for preeclampsia, found using #opendata from @NCBI GEO: ncbi.nlm.nih.gov/pubmed/24195779 #ISMB @Carment

RETWEETS 11
FAVORITE 1

12:42 PM - 14 Jul 2014

Display Settings: Abstract

BMC Med. 2013 Nov 6;11:236. doi: 10.1186/1741-7015-11-236.

Integrating multiple 'omics' analyses identifies serological protein biomarkers for preeclampsia.

Liu LY, Yang T, Ji J, Wen Q, Morgan AA, Jin B, Chen G, Lyell DJ, Stevenson DK, Ling XB[1], Butte AJ.

Author information

Abstract

BACKGROUND: Preeclampsia (PE) is a pregnancy-related vascular disorder which is the leading cause of maternal morbidity and m
serological protein markers to diagnose PE with a multi-'omics' based discovery approach.

METHODS: Seven previous placental expression studies were combined for a multiplex analysis, and in parallel, two-dimensional gel
compare serum proteomes in PE and control subjects. The combined biomarker candidates were validated with available ELISA assa
(n=32) and control (n=32) samples. With the validated biomarkers, a genetic algorithm was then used to construct and optimize bioma

RESULTS: In addition to the previously identified biomarkers, the angiogenic and antiangiogenic factors (soluble fms-like tyrosine kina
factor (PlGF)), we found 3 up-regulated and 6 down-regulated biomakers in PE sera. Two optimal biomarker panels were developed f
assessment, respectively.

CONCLUSIONS: Both early and late onset PE diagnostic panels, constructed with our PE biomarkers, were superior over sFlt-1/PIGF
functional significance of these PE biomarkers and their associated pathways were analyzed which may provide new insights into the

PMID: 24195779 [PubMed - indexed for MEDLINE] Free full text

Increase visibility & impact

OPEN ACCESS PEER-REVIEWED

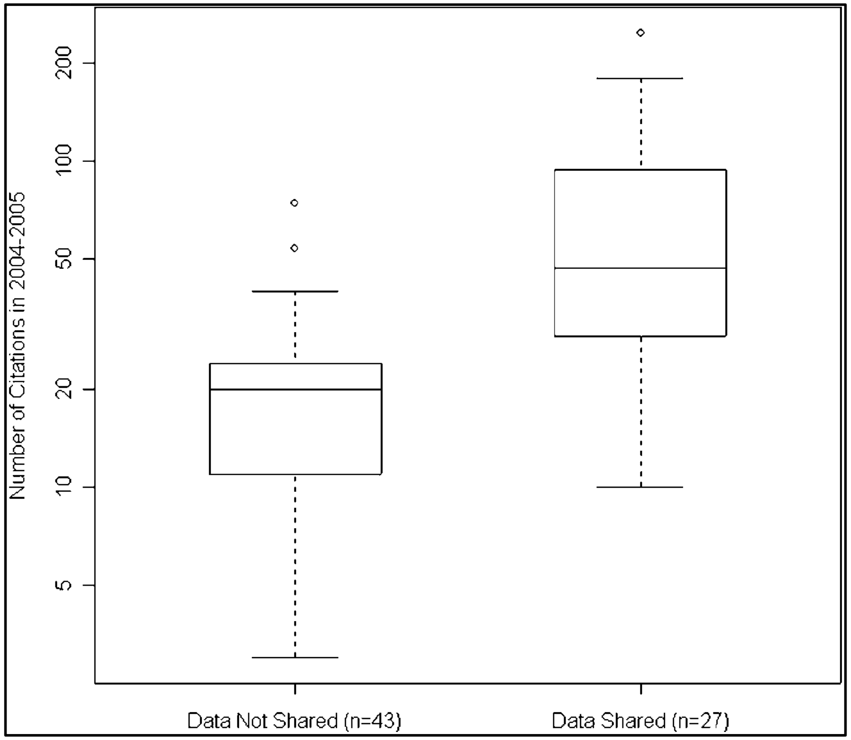
RESEARCH ARTICLE

34,276	97	359
VIEWS	CITATIONS	SAVES

Sharing Detailed Research Data Is Associated with Increased Citation Rate

Heather A. Piwowar, Roger S. Day, Douglas B. Fridsma

Published: Mar 21, 2007 • DOI: 10.1371/journal.pone.0000308 • Featured in PLOS Collections



85 cancer microarray clinical trial publications

69% increase in citations for articles w/publicly available data

Funder mandates

"…directed Federal agencies with more than $100M in R&D expenditures to develop plans to make the published results of federally funded research freely available to the public within one year of publication and requiring researchers to better account for and manage the digital data resulting from federally funded scientific research."

Posted by Michael Stebbins on February 22, 2013 at 12:04 PM EST

The Obama Administration is committed to the proposition that citizens deserve easy access to the results of scientific research their tax dollars have paid for. That's why, in a policy memorandum released today, OSTP Director John Holdren has directed Federal agencies with more than $100M in R&D expenditures to develop plans to make the published results of federally funded research freely available to the public within one year of publication and requiring researchers to better account for and manage the digital data resulting from federally funded scientific research. OSTP has been looking into this issue for some time, soliciting broad public input on multiple occasions and convening an interagency working group to develop a policy. The final policy reflects substantial inputs from scientists and scientific organizations, publishers, members of Congress, and other members of the public—over 65 thousand of whom recently signed a *We the People* petition asking for expanded public access to the results of taxpayer-funded research.

Which agencies are affected?

NASA

$54M

NSF

Department of Defense · United States of America ·

$35M

USDA
United States
Department of Agriculture

Department of Health & Human Services · USA

$23M

USGS
science for a changing world

Department of Energy · United States of America ·

National Oceanic and Atmospheric Administration
NOAA
U.S. Department of Commerce

United States · Environmental Protection Agency ·

Aspects of data management

DMPs/Planning

Storage & backup

File organization & naming

Documentation & metadata

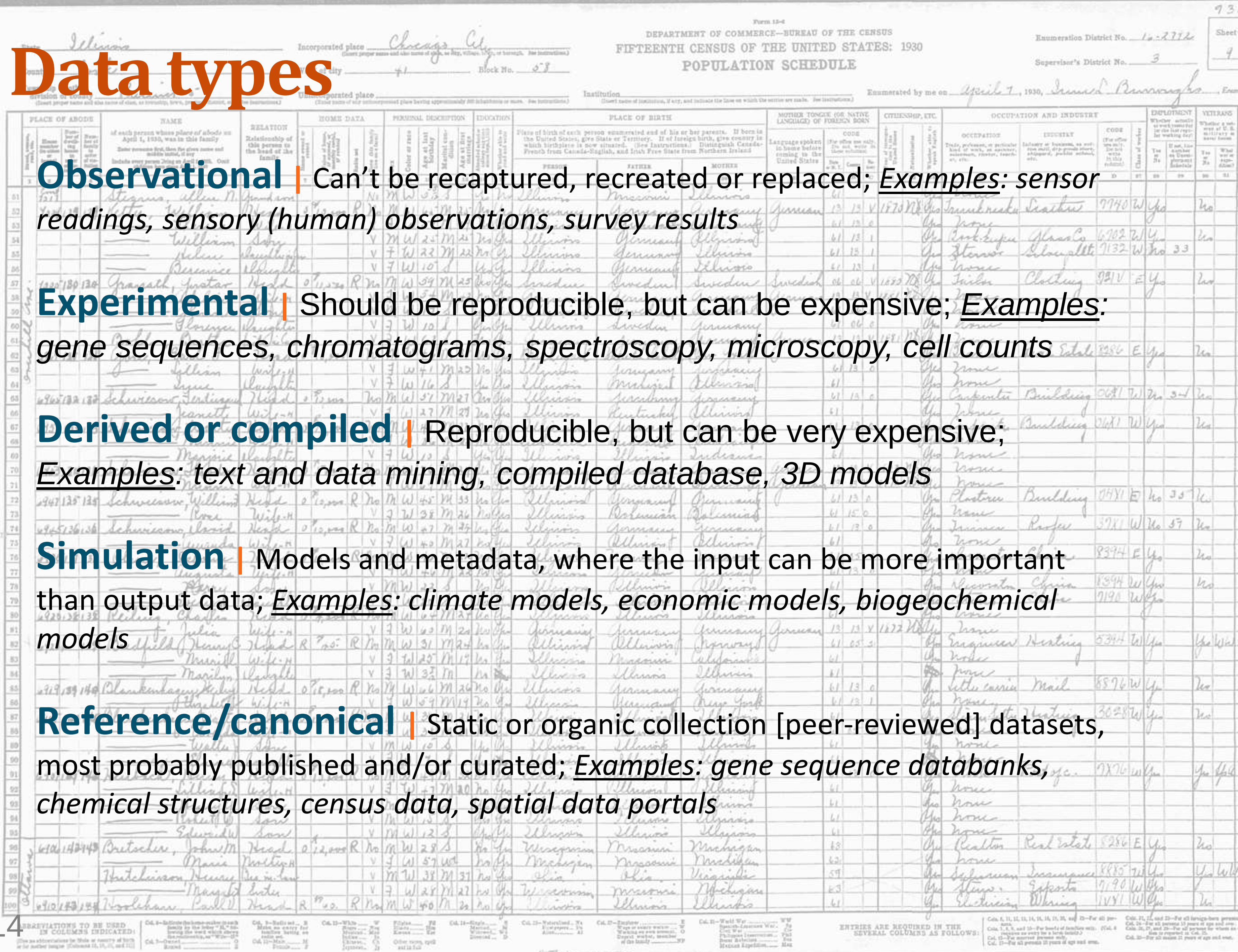
Legal/ethical considerations

Sharing & reuse

Archiving & preservation

Data types & formats

Data types



Observational | Can't be recaptured, recreated or replaced; _Examples_: *sensor readings, sensory (human) observations, survey results*

Experimental | Should be reproducible, but can be expensive; _Examples_: *gene sequences, chromatograms, spectroscopy, microscopy, cell counts*

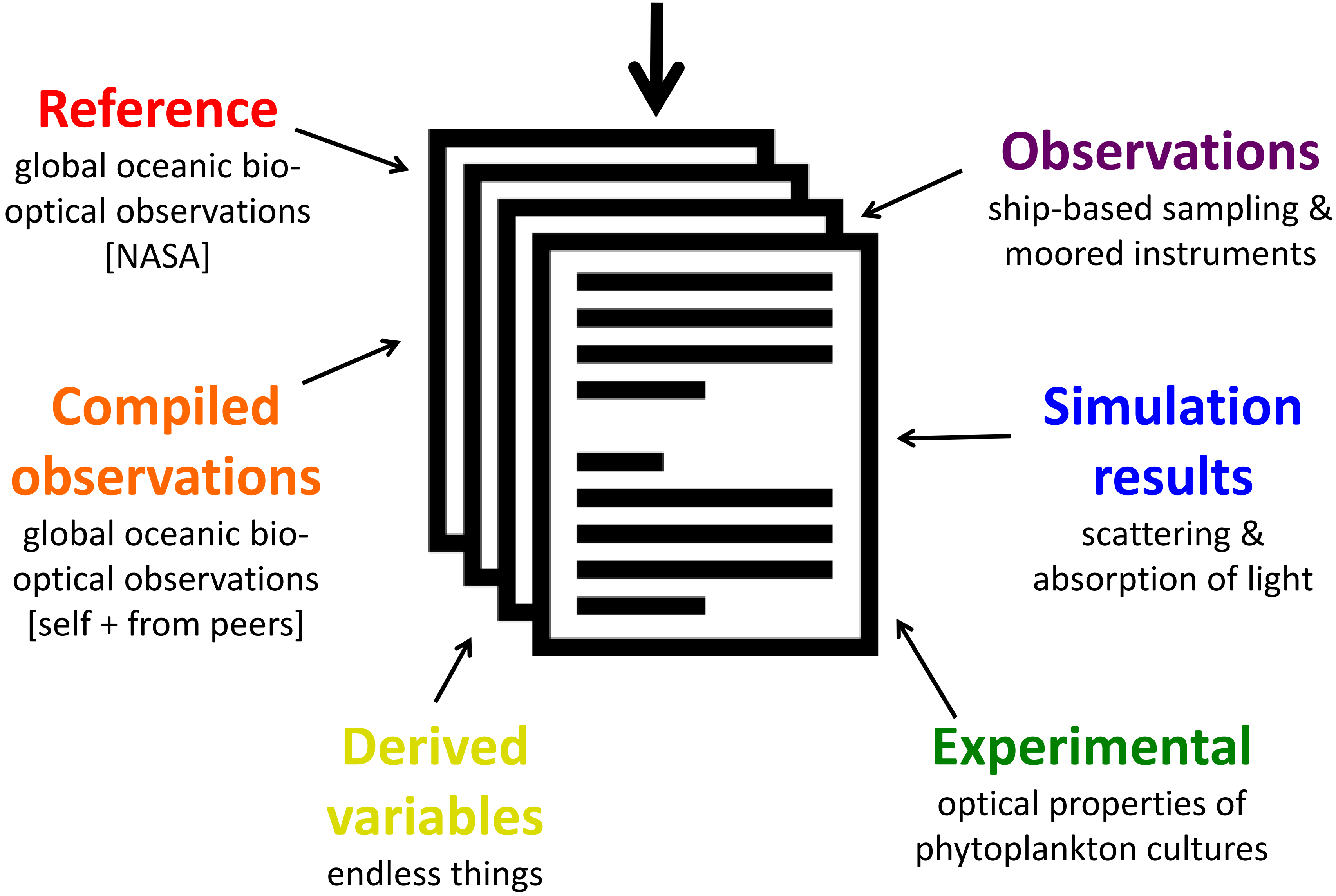
Derived or compiled | Reproducible, but can be very expensive; _Examples_: *text and data mining, compiled database, 3D models*

Simulation | Models and metadata, where the input can be more important than output data; _Examples_: *climate models, economic models, biogeochemical models*

Reference/canonical | Static or organic collection [peer-reviewed] datasets, most probably published and/or curated; _Examples_: *gene sequence databanks, chemical structures, census data, spatial data portals*

Amanda’s dissertation

The spectral backscattering properties of marine particles



Reference

global oceanic bio-
optical observations
[NASA]

Observations

ship-based sampling &
moored instruments

Compiled observations

global oceanic bio-
optical observations
[self + from peers]

Simulation results

scattering &
absorption of light

Derived variables

endless things

Experimental

optical properties of
phytoplankton cultures

Data types: another classification

Qualitative data “is a categorical measurement expressed not in terms of numbers, but rather by means of a natural language description. In statistics, it is often used interchangeably with "categorical" data.” See also: **nominal, ordinal**

Quantitative data “is a numerical measurement expressed not by means of a natural language description, but rather in terms of numbers. However, not all numbers are continuous and measurable.”

“My favorite color is blue-green.” vs. “My favorite color is 510 nm.”

Source: Wikibooks: Statistics

More common data types

Geospatial data has a geographical or geospatial aspect. Spatial location is critically tied to other variables.

Digital image, audio & video data

Documentation & scripts Sometimes, software code IS data; likewise with documentation (laboratory notebooks, written observations, etc.)

File Formats

"A **file format** is a standard way that information is encoded for storage in a computer file. It specifies how bits are used to encode information in a digital storage medium."   - Wikipedia

Data type

Qualitative, tabular experimental data

Possible formats

Excel spreadsheet (.xlsx)
Comma-delimited text (.csv)
Access database (.mdb/,accdb)
Google Spreadsheet
SPSS portable file (.por)
XML file

Reflective Writing: 60 seconds

What types & formats of data will you be generating and/or using?

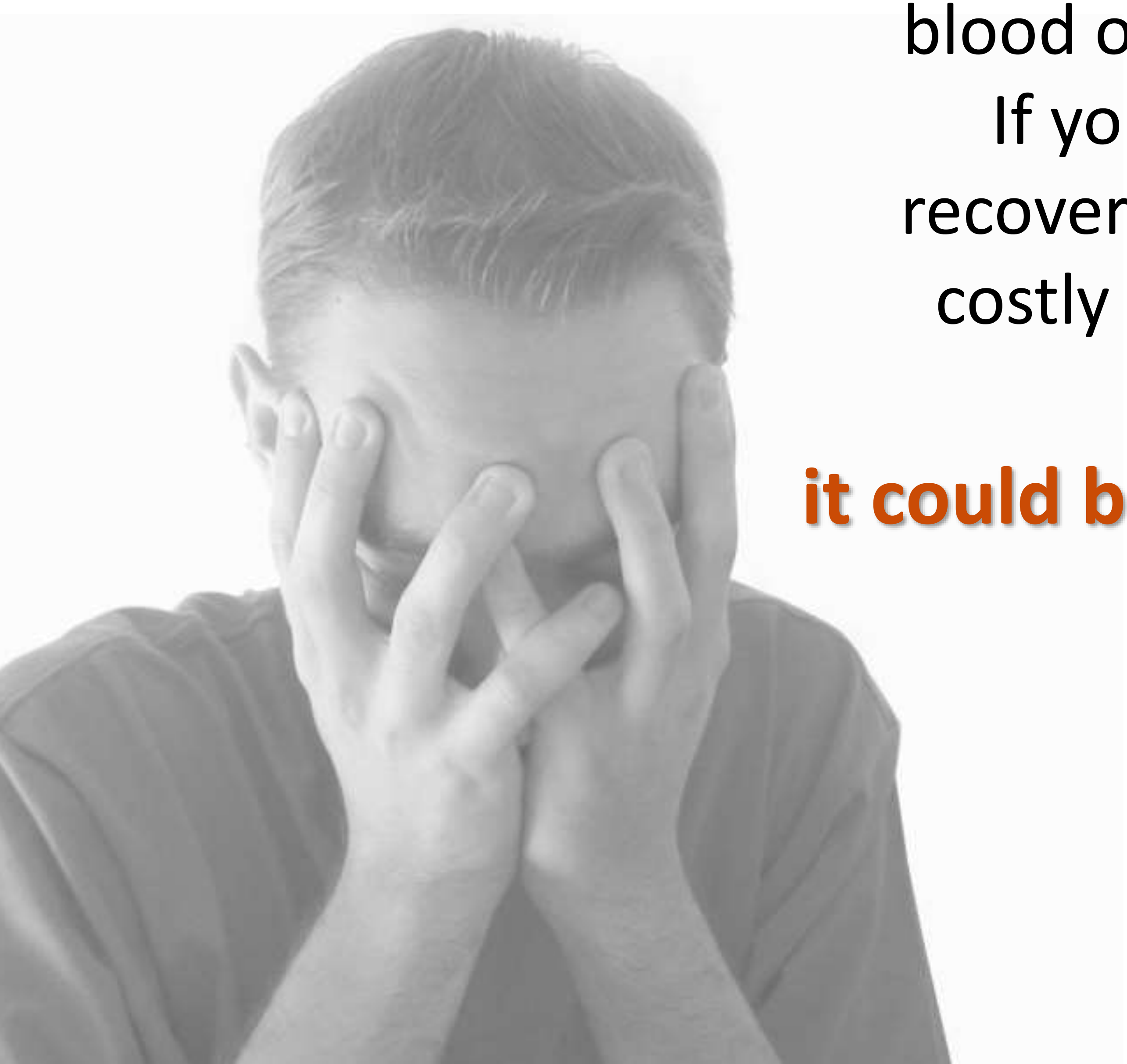
Observational | Experimental | Derived | Compiled |
Simulation | Reference/Canonical
Qualitative | Quantitative | Geospatial
Image/audio/video | Scripts/codes

The “big picture”

Where do you start?

Make a plan

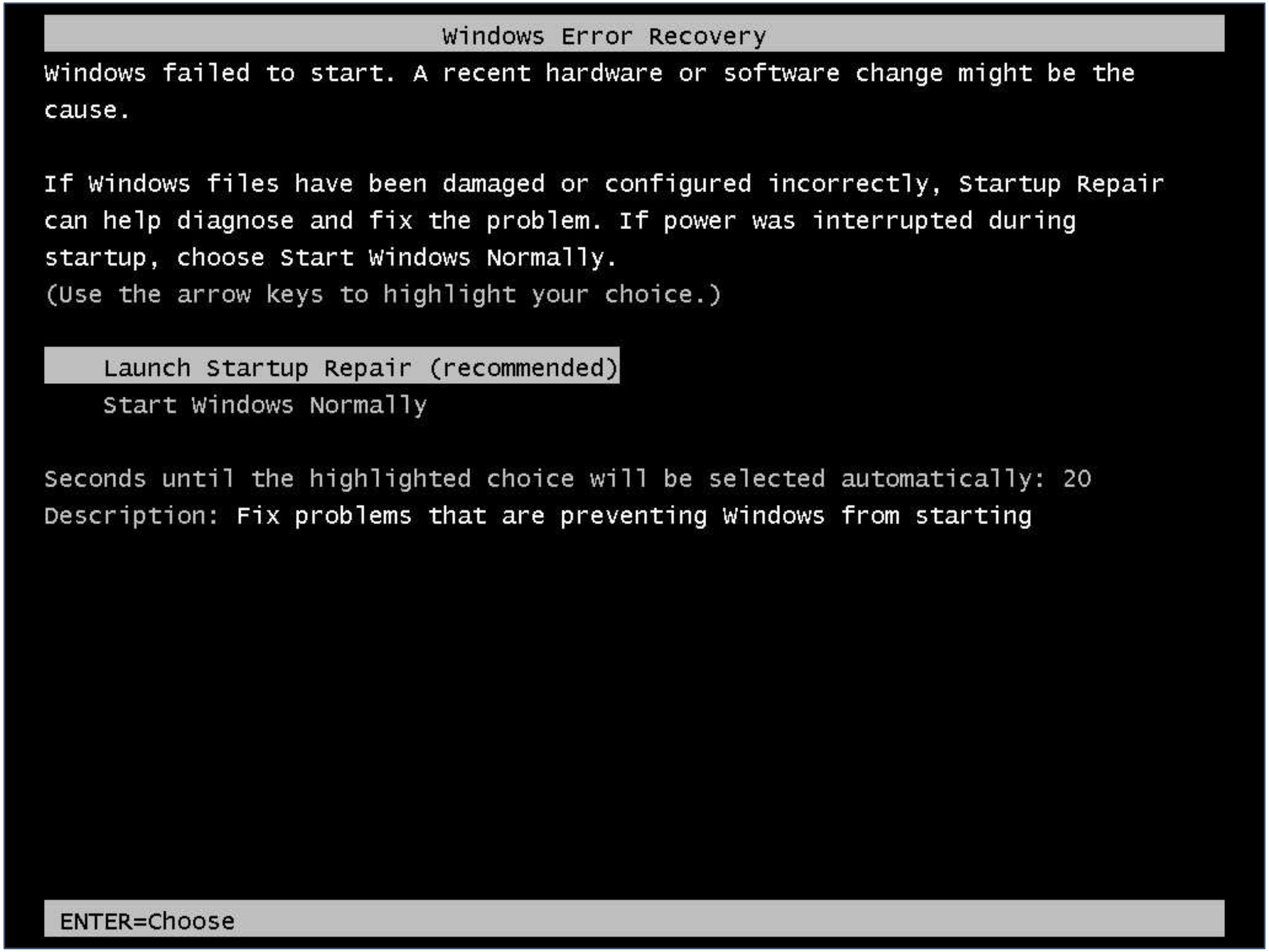
Data storage & backup



"Your data are the life blood of your research. If you lose your data recovery could be slow, costly or even worse…

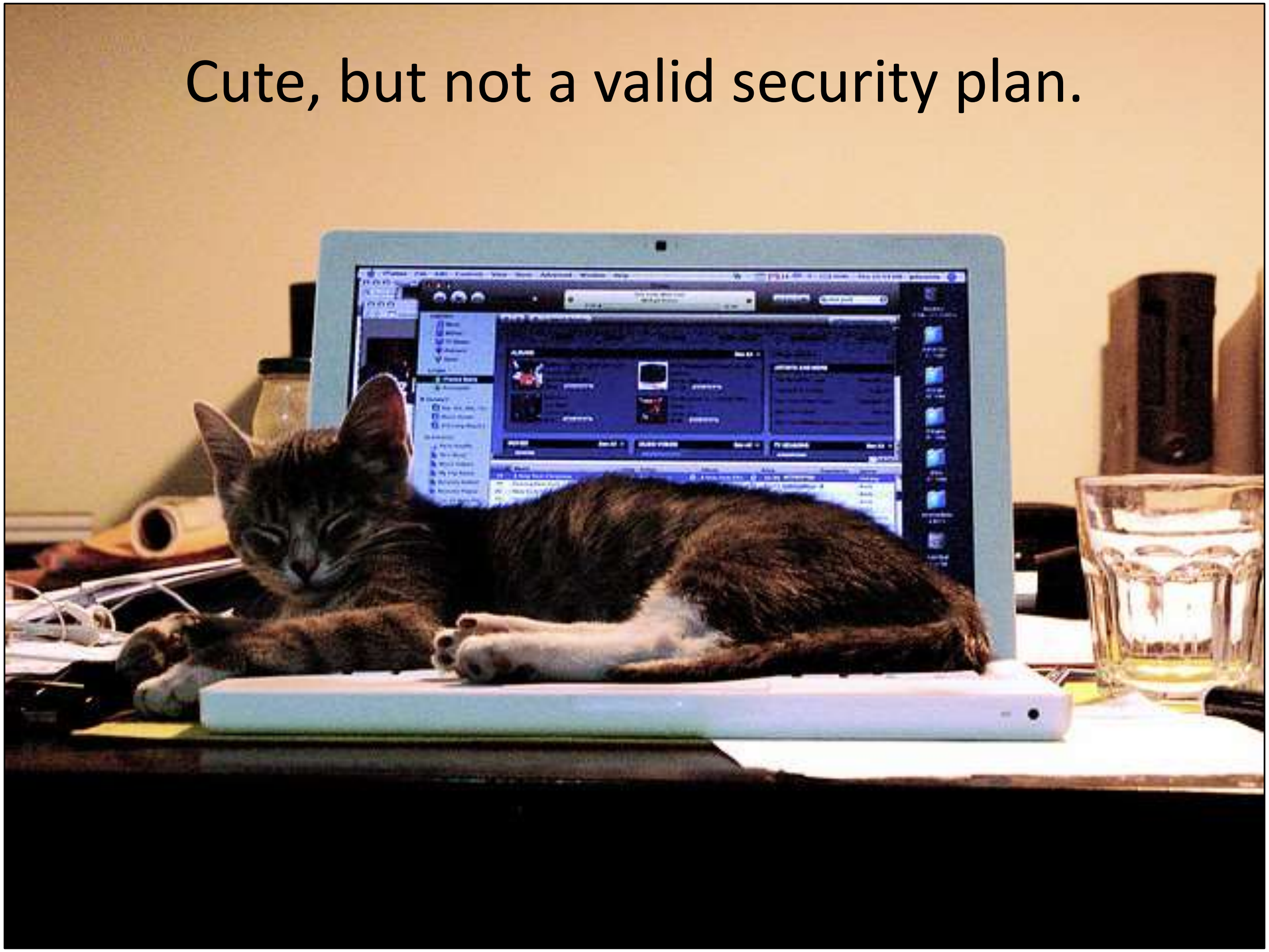
it could be impossible."

Most common loss scenario: drive failure



```
                      Windows Error Recovery
Windows failed to start. A recent hardware or software change might be the
cause.

If Windows files have been damaged or configured incorrectly, Startup Repair
can help diagnose and fix the problem. If power was interrupted during
startup, choose Start Windows Normally.
(Use the arrow keys to highlight your choice.)

    Launch Startup Repair (recommended)
    Start Windows Normally

Seconds until the highlighted choice will be selected automatically: 20
Description: Fix problems that are preventing Windows from starting


ENTER=Choose
```

This happens a lot:
physical theft & unintentional damage



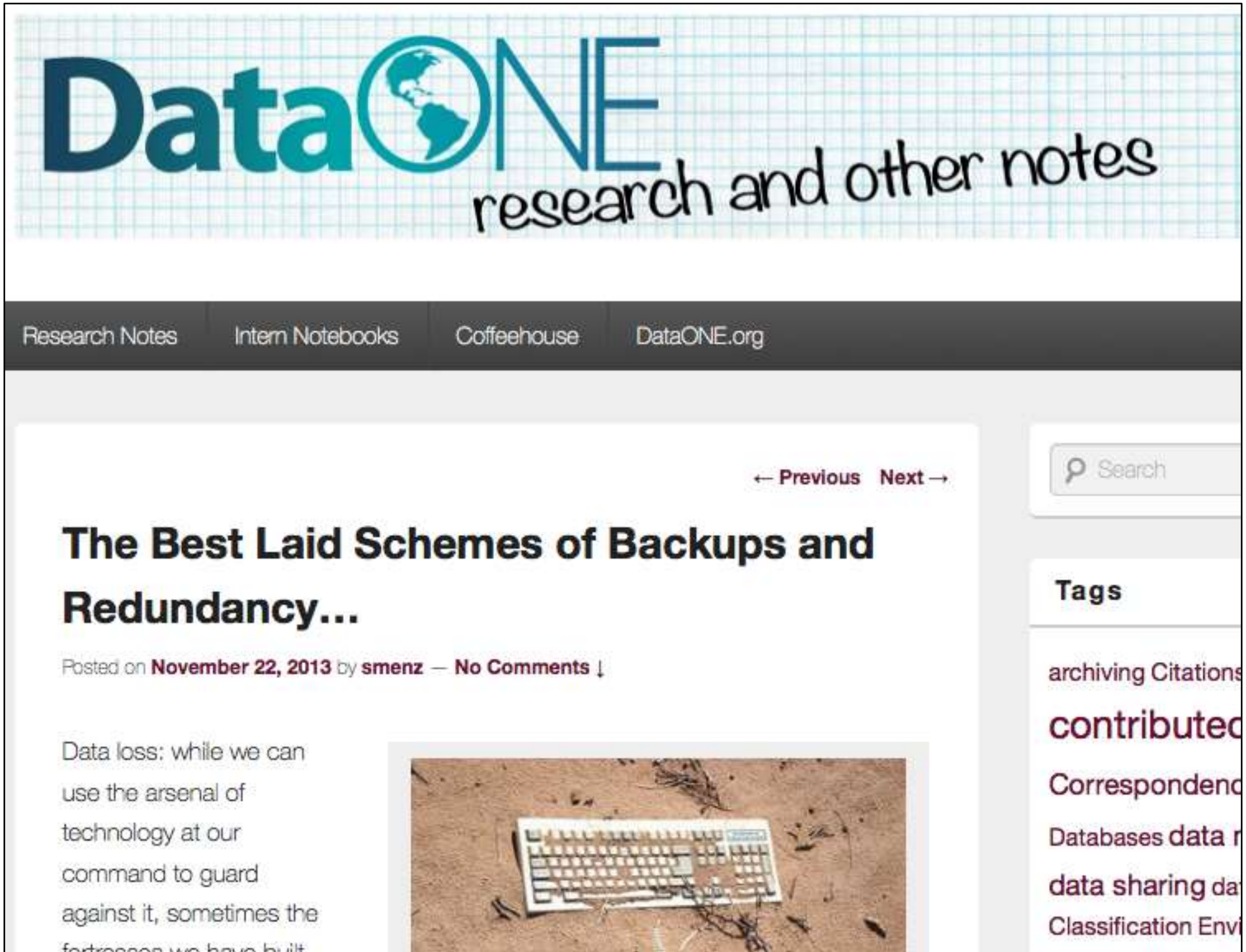
Cute, but not a valid security plan.

Rare, unexpected events happen

University of Southampton, School of Electronics and Computer Science, Southampton, UK, 2005

It CAN happen to you

Real-world lesson:
Audit your backups…



The Best Laid Schemes of Backups and Redundancy…

Posted on **November 22, 2013** by **smenz** — **No Comments ↓**

Data loss: while we can use the arsenal of technology at our command to guard against it, sometimes the fortresses we have built

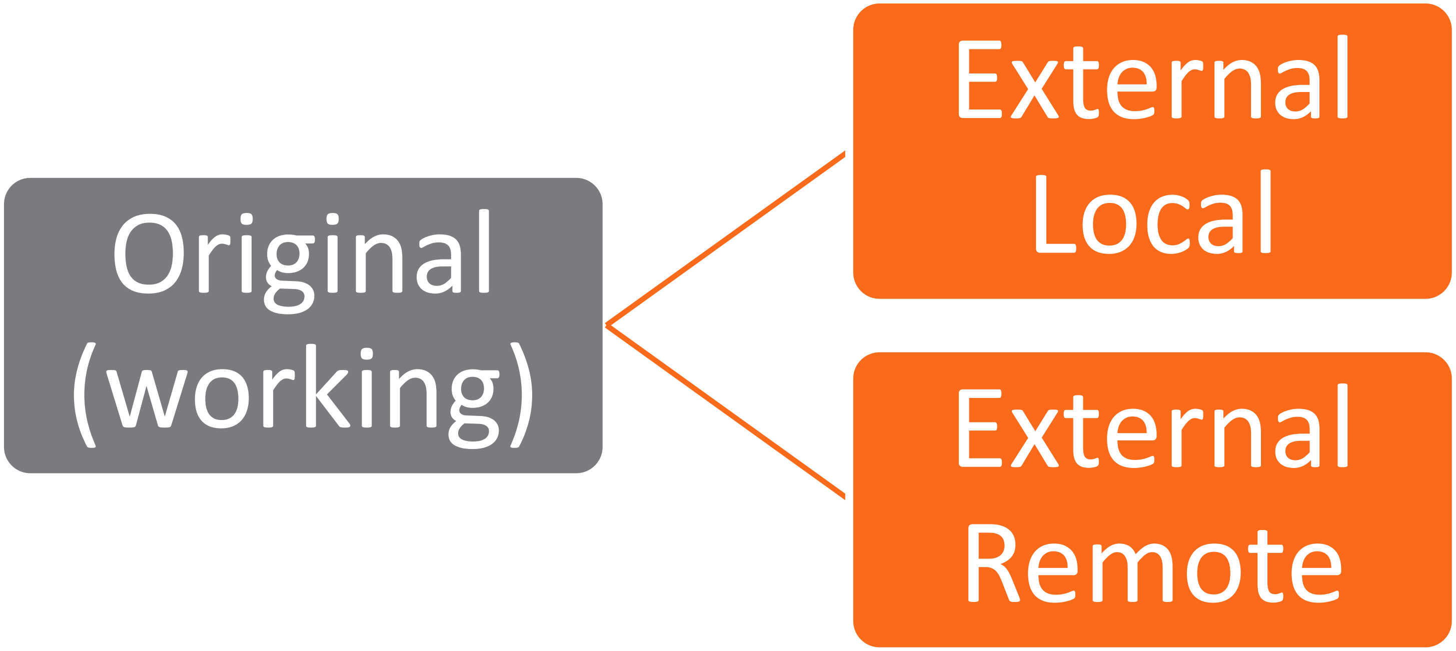
Data backup

> “Keeping backups is probably your most important data management task.”
>
> -Everyone

1. Some data backup is better than none.
2. Automated backups are better than manual.
3. Your data are only as safe as your last backup.

Data backup

Best Practice: 3 Copies of datasets



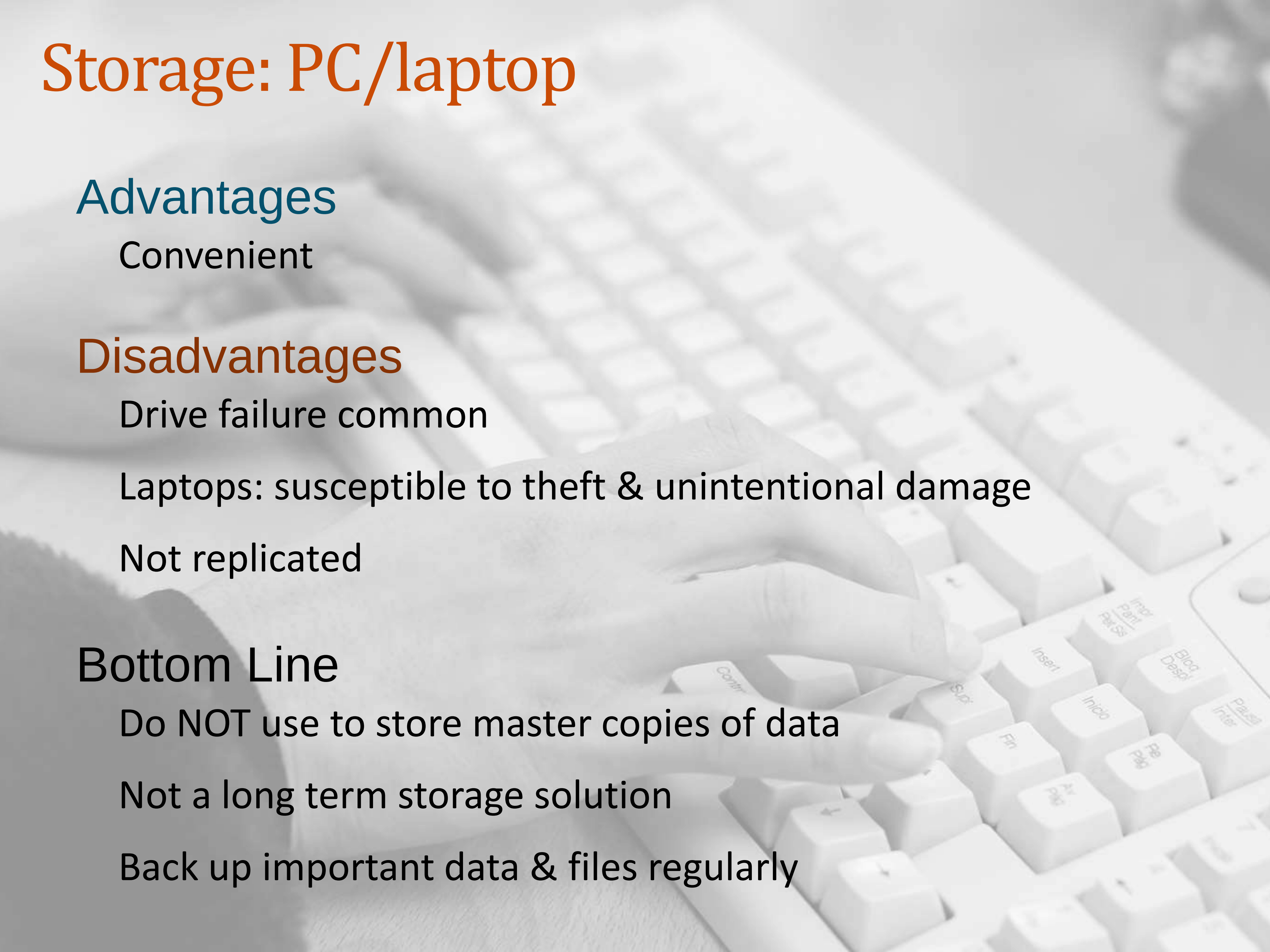
Original (working)

External Local

External Remote

Data storage options

1. Personal computers (PCs) & laptops
2. External storage devices
3. Networked Drives
4. Cloud servers



Storage: PC/laptop

Advantages

- Convenient

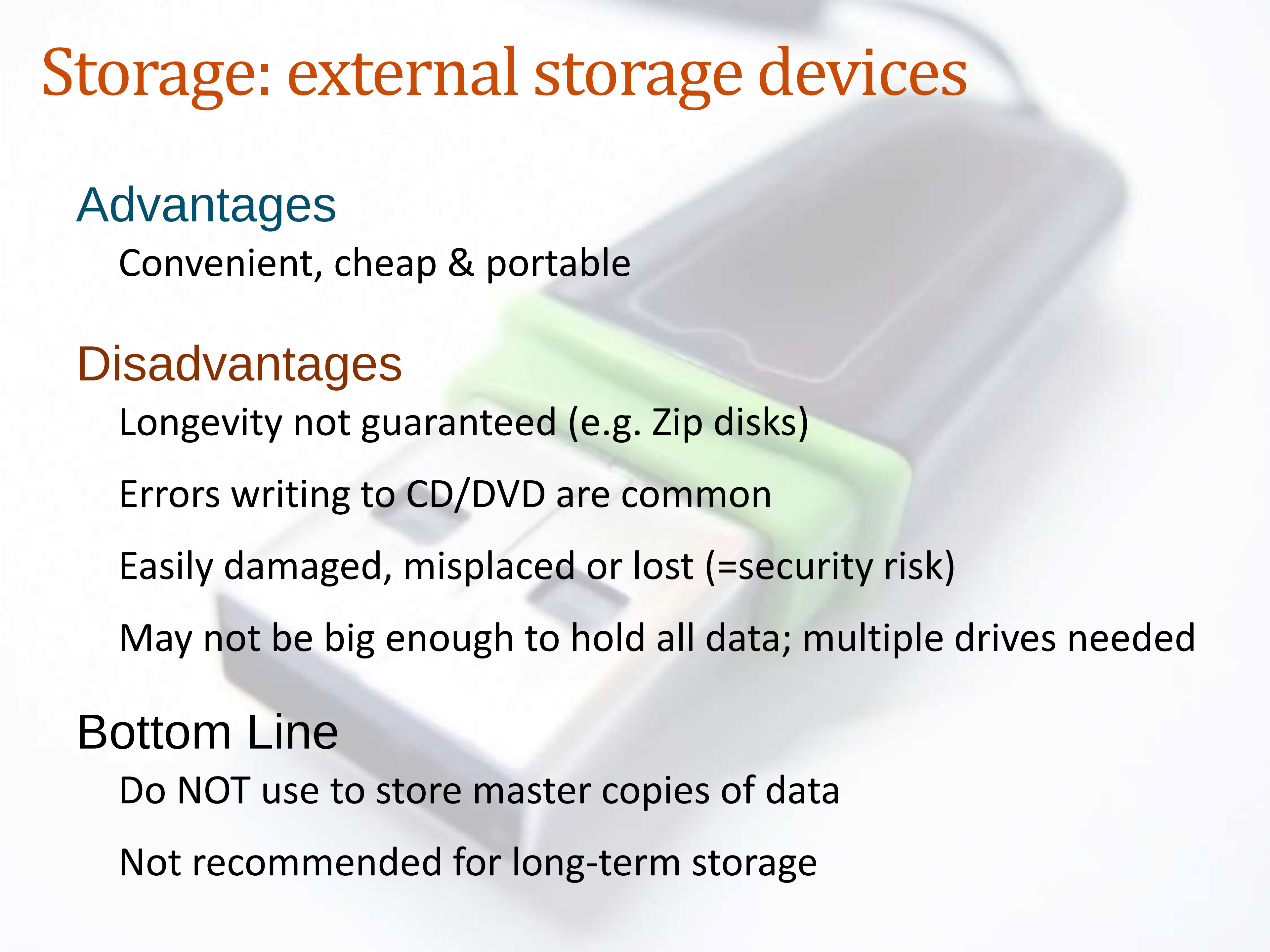
Disadvantages

- Drive failure common
- Laptops: susceptible to theft & unintentional damage
- Not replicated

Bottom Line

- Do NOT use to store master copies of data
- Not a long term storage solution
- Back up important data & files regularly

Storage: external storage devices



Advantages

- Convenient, cheap & portable

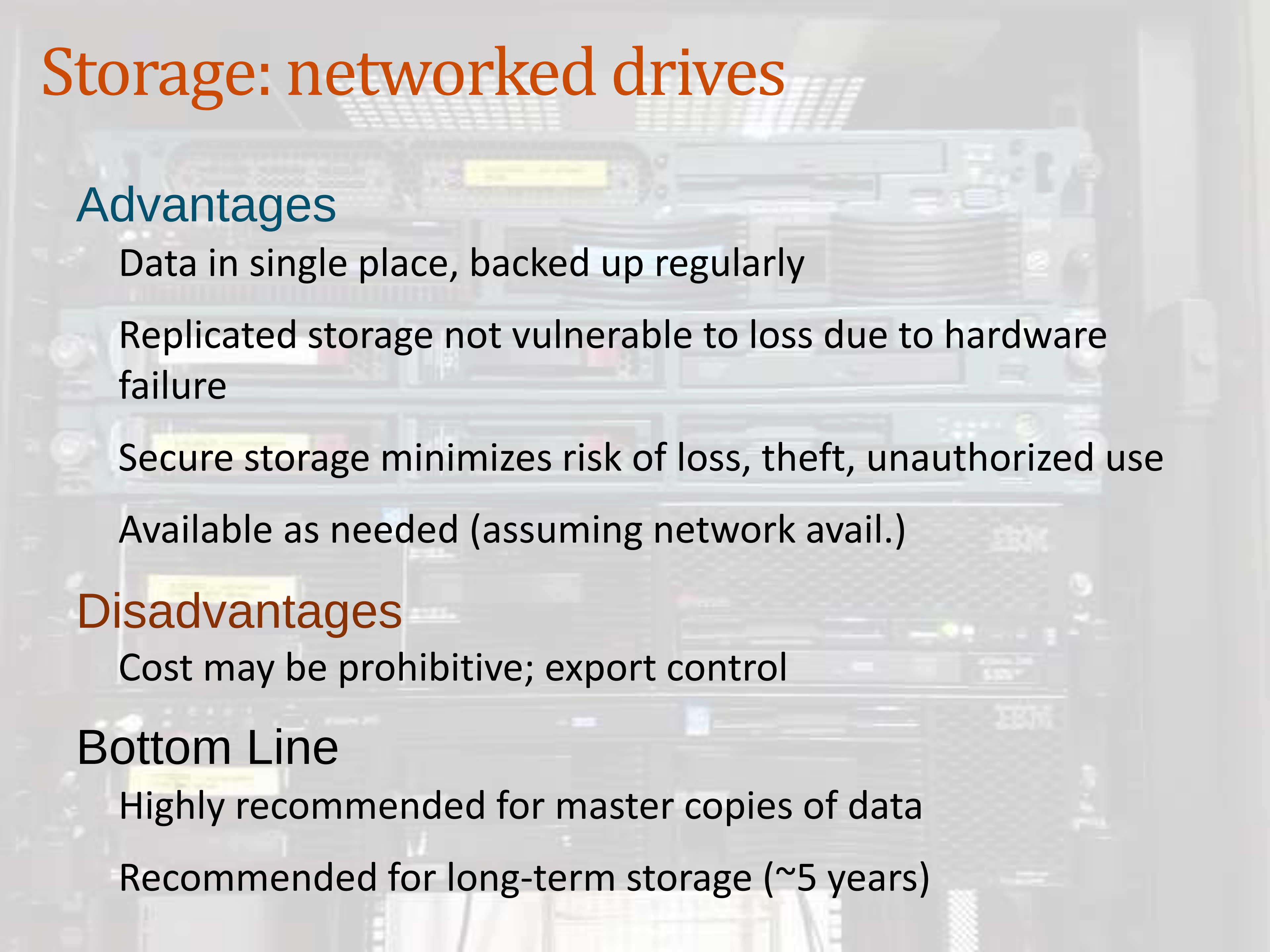
Disadvantages

- Longevity not guaranteed (e.g. Zip disks)
- Errors writing to CD/DVD are common
- Easily damaged, misplaced or lost (=security risk)
- May not be big enough to hold all data; multiple drives needed

Bottom Line

- Do NOT use to store master copies of data
- Not recommended for long-term storage

Storage: networked drives



Advantages

- Data in single place, backed up regularly
- Replicated storage not vulnerable to loss due to hardware failure
- Secure storage minimizes risk of loss, theft, unauthorized use
- Available as needed (assuming network avail.)

Disadvantages

- Cost may be prohibitive; export control

Bottom Line

- Highly recommended for master copies of data
- Recommended for long-term storage (~5 years)

Storage: cloud storage

Advantages

- Data in single place, backed up regularly
- Replicated storage not vulnerable to loss due to hardware failure
- Secure storage minimizes risk of loss, theft, unauthorized use

Disadvantages

- Cost may be prohibitive
- Upload/download bottleneck & fees
- Longevity?
- Export control

Bottom Line

- Possibly recommended for master copies of data
- Not recommended for in-process data, large files

Storage: Google Drive for OSU

Advantages

All same advantages of network & cloud storage

File sharing & collaboration w/variable access levels

Unlimited storage (GD), 30 GB non-GD

Automatic version control on GD

Disadvantages

30 GB may not be enough

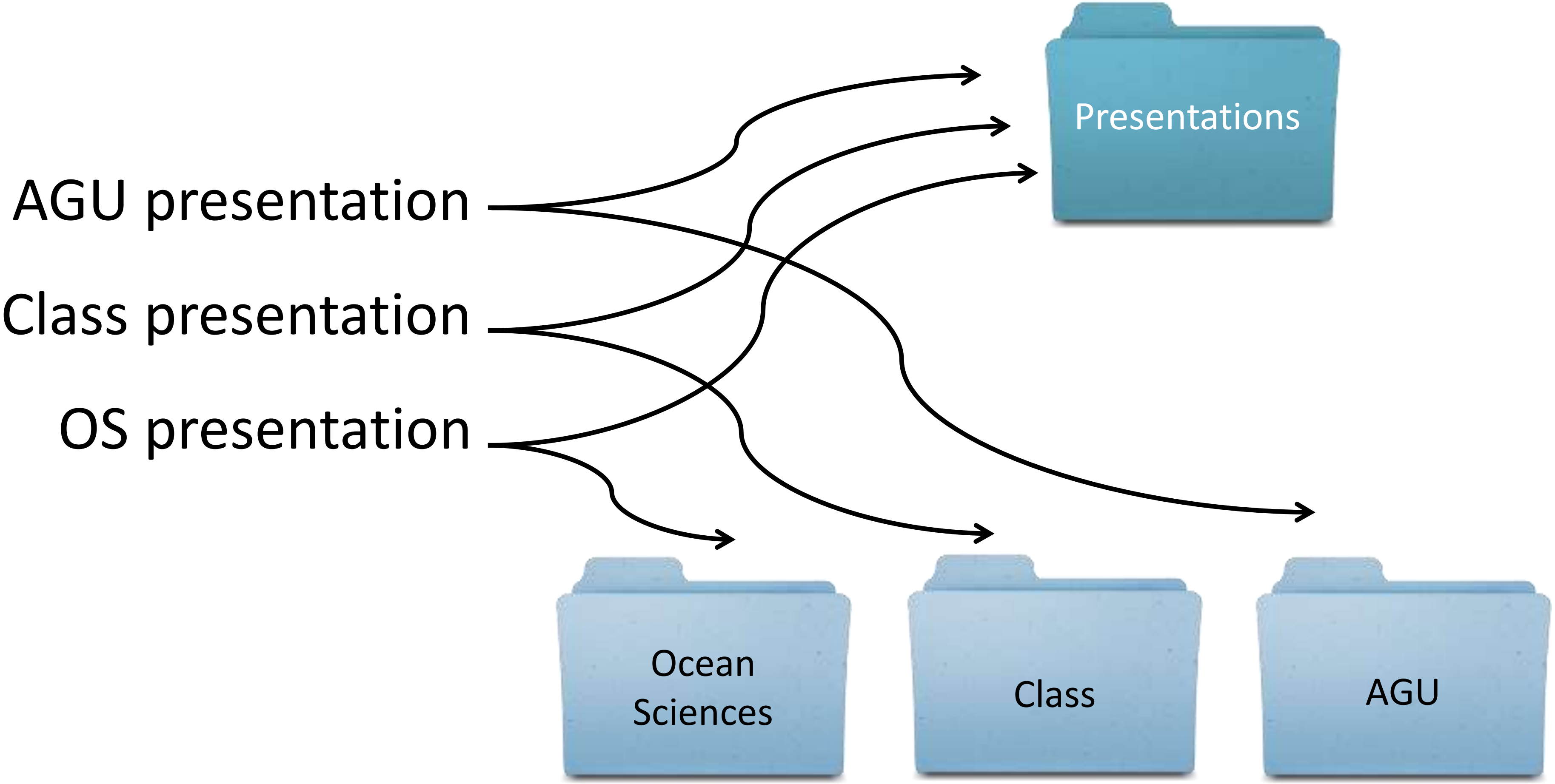
Upload/download bottleneck

Bottom Line

Possibly recommended for master copies of data

Possibly not recommended for in-process data, large files

Data organization



Data storage options

Local

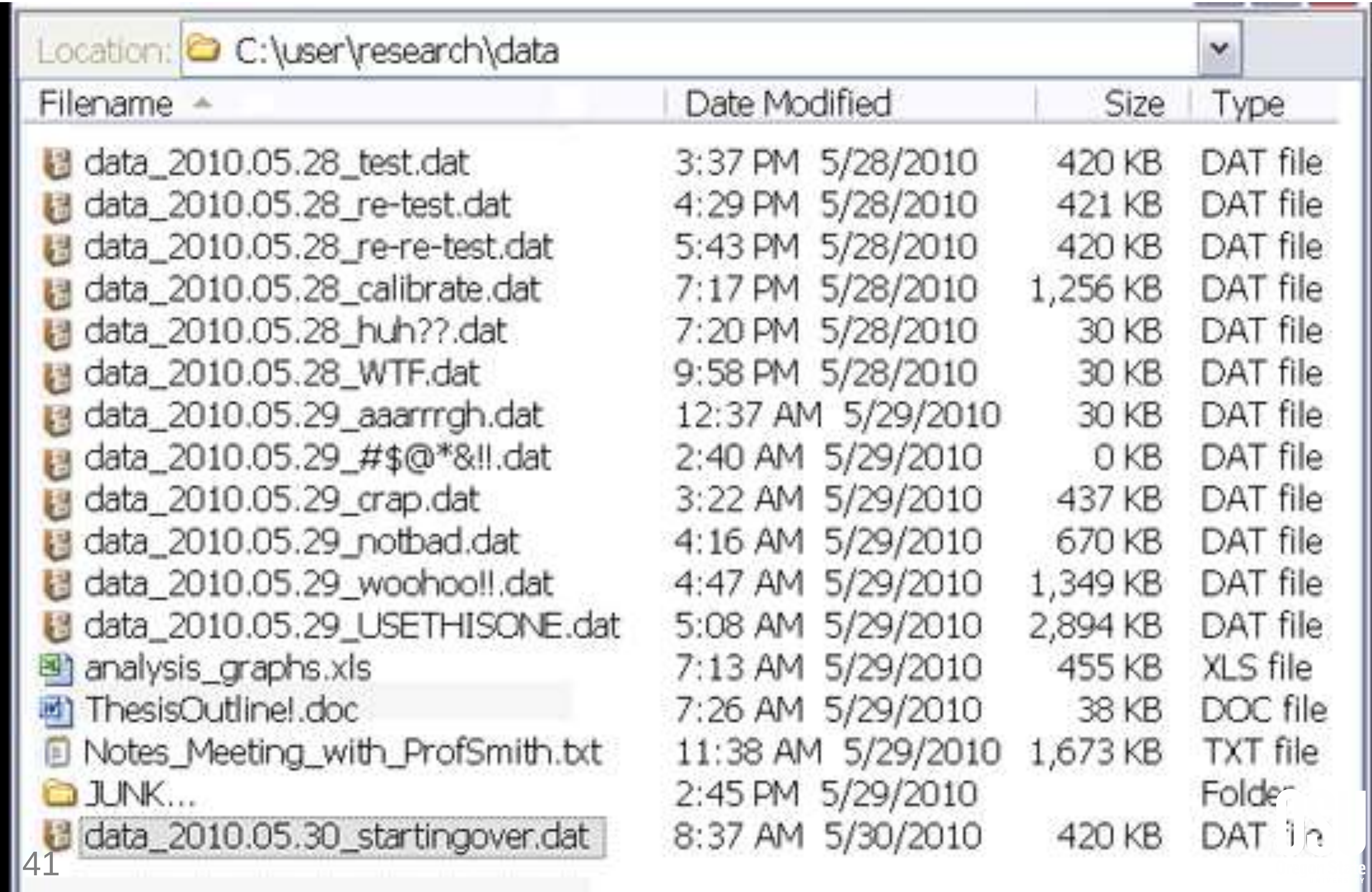
- Computer
- External storage
- Network server

Remote

- Network server
- Cloud storage
  - Google Apps
  - Box, Dropbox, etc.

File-naming conventions

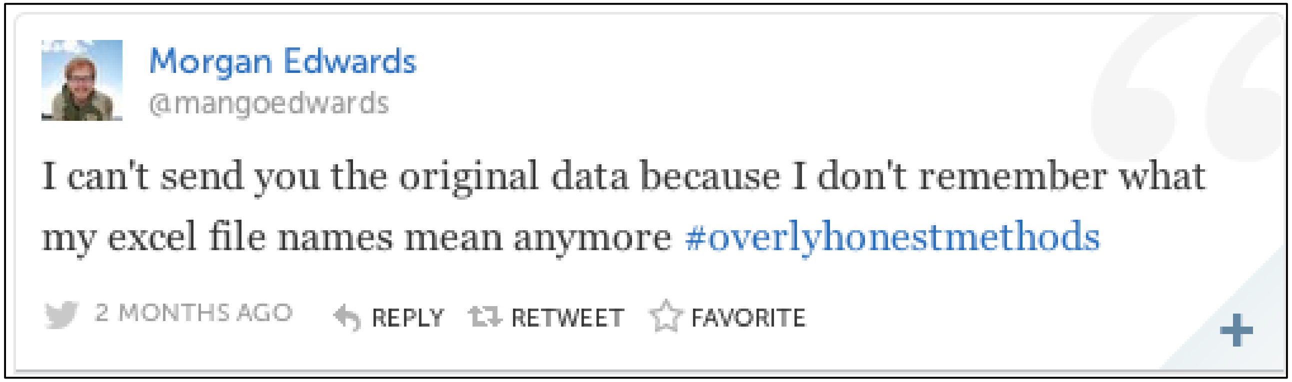
File naming strategy?



Location: C:\user\research\data

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_aaarrrgh.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		Folde
data_2010.05.30_startingover.dat	8:37 AM 5/30/2010	420 KB	DAT

#OverlyHonestMethods



Morgan Edwards
@mangoedwards

I can't send you the original data because I don't remember what my excel file names mean anymore #overlyhonestmethods

File naming conventions

s/n, variable

Retain order

Project_instrument_location_YYYYMMDDhhmmss_extra.ext

Index/grant

conditions

Leading zero!

File naming strategies

Order by date:

1955-04-12_notes_MassObs.docx
1955-04-12_questionnaire_MassObs.pdf
1963-12-15_notes_Gorer.docx
1963-12-15_questionnaire_Gorer.pdf

Order by type:

Notes_Gorer_1963-12-15.docx
Notes_MassObs_1955-04-12.docx
Questionnaire_Gorer_1963-12-15.pdf
Questionnaire_MassObs_1955-04-12.pdf

Order by subject:

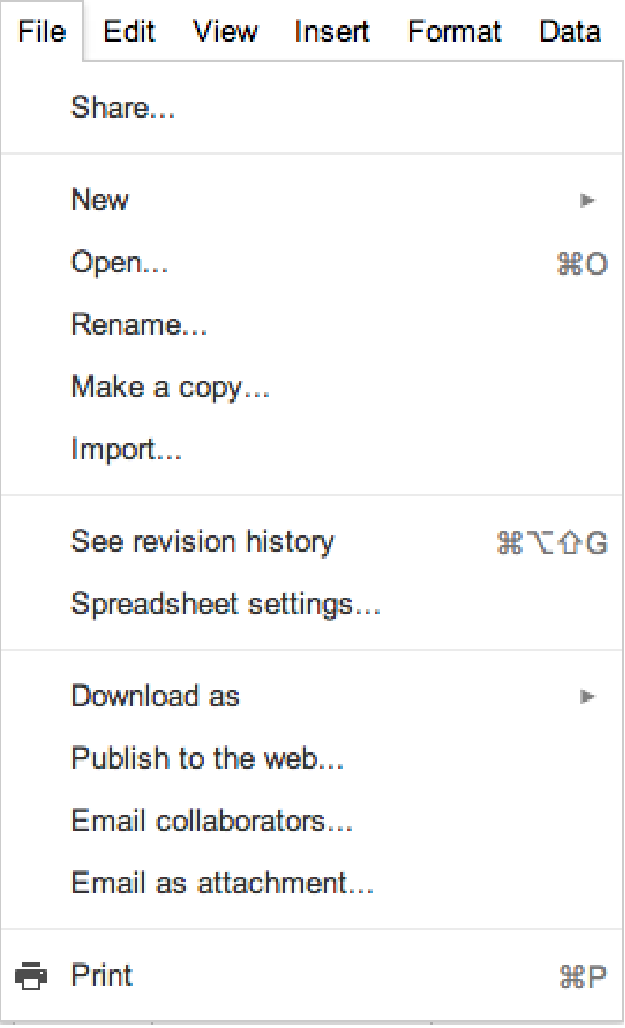
Gorer_notes_1963-12-15.docx
Gorer_questionnaire_1963-12-15.pdf
MassObs_notes_1955-04-12.docx
MassObs_questionnaire_1955-04-12.pdf

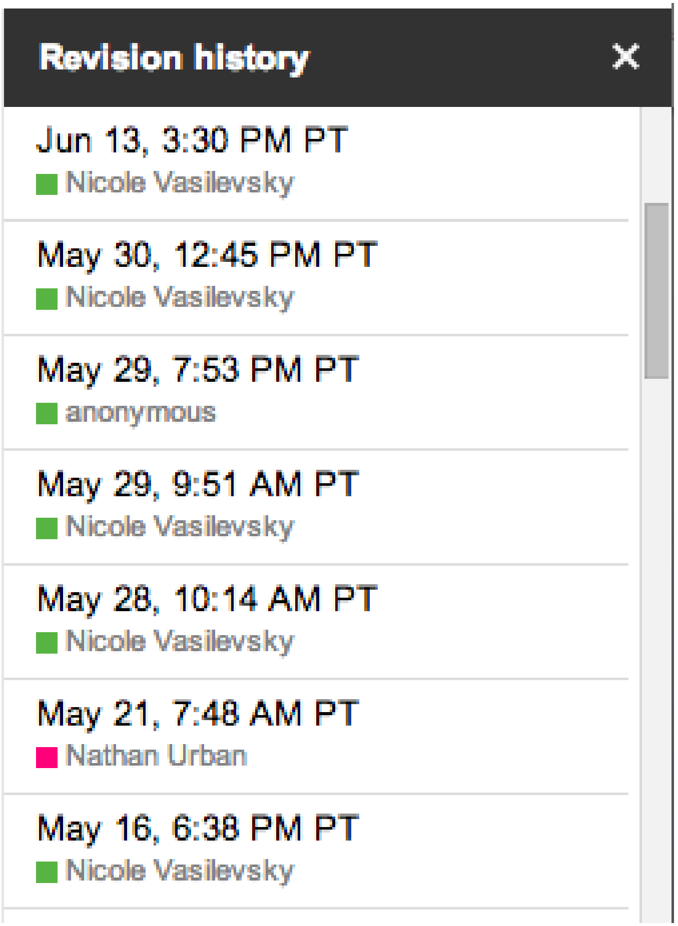
Forced order with numbering:

01_MassObs_questionnaire_1955-04-12.pdf
02_MassObs_notes_1955-04-12.docx
03_Gorer_questionnaire_1963-12-15.pdf
04_Gorer_notes_1963-12-15.docx

Version control

Google docs





Jun 13, 3:30 PM PT
🟩 Nicole Vasilevsky

May 30, 12:45 PM PT
🟩 Nicole Vasilevsky

May 29, 7:53 PM PT
🟩 anonymous

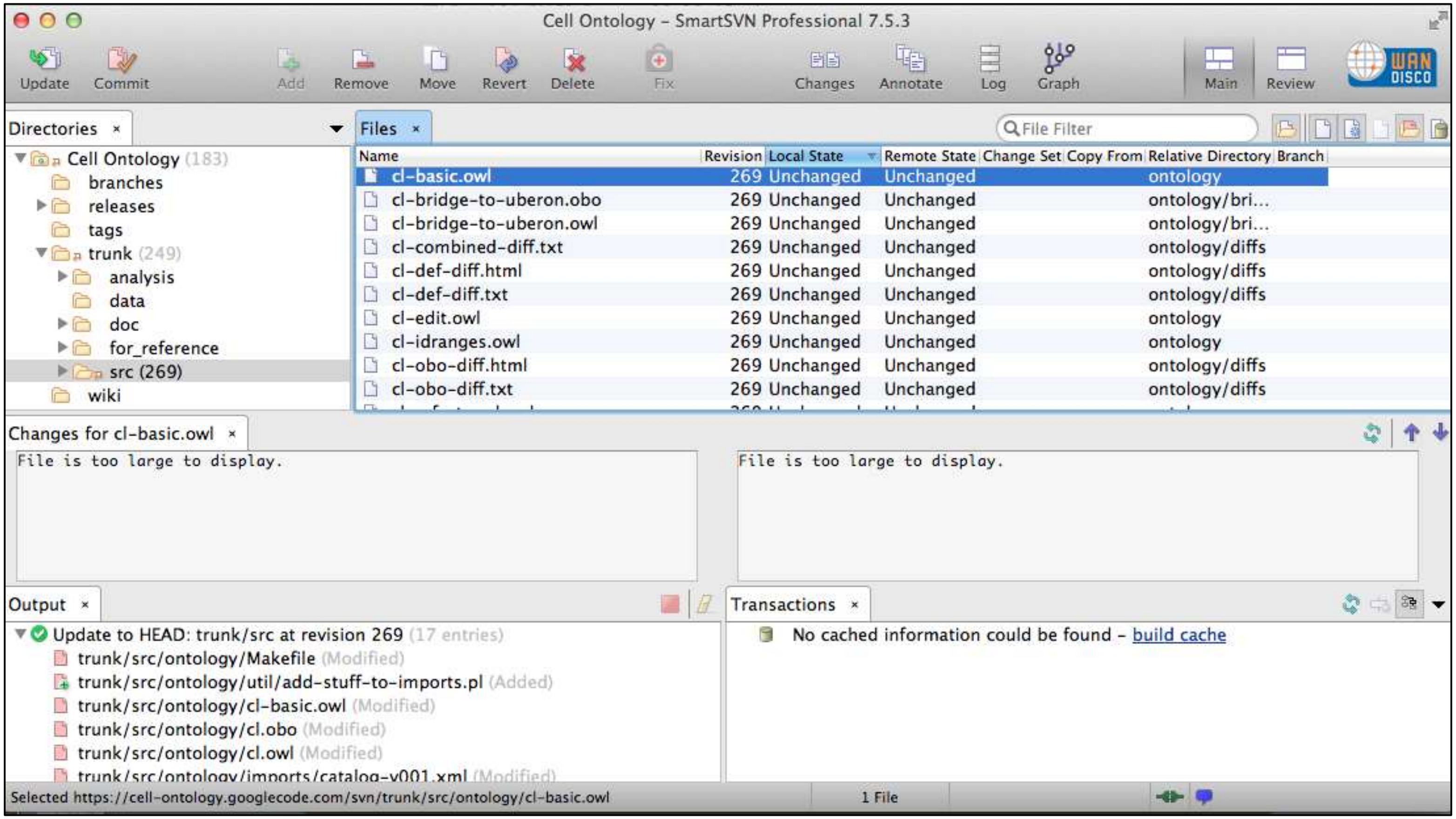
May 29, 9:51 AM PT
🟩 Nicole Vasilevsky

May 28, 10:14 AM PT
🟩 Nicole Vasilevsky

May 21, 7:48 AM PT
🟥 Nathan Urban

May 16, 6:38 PM PT
🟩 Nicole Vasilevsky

Version control



- Cell Ontology (183)
  - branches
  - releases
  - tags
  - trunk (249)
    - analysis
    - data
    - doc
    - for_reference
    - src (269)
  - wiki

Name	Revision	Local State	Remote State	Change Set	Copy From	Relative Directory	Branch
cl-basic.owl	269	Unchanged	Unchanged			ontology	
cl-bridge-to-uberon.obo	269	Unchanged	Unchanged			ontology/bri...	
cl-bridge-to-uberon.owl	269	Unchanged	Unchanged			ontology/bri...	
cl-combined-diff.txt	269	Unchanged	Unchanged			ontology/diffs	
cl-def-diff.html	269	Unchanged	Unchanged			ontology/diffs	
cl-def-diff.txt	269	Unchanged	Unchanged			ontology/diffs	
cl-edit.owl	269	Unchanged	Unchanged			ontology	
cl-idranges.owl	269	Unchanged	Unchanged			ontology	
cl-obo-diff.html	269	Unchanged	Unchanged			ontology/diffs	
cl-obo-diff.txt	269	Unchanged	Unchanged			ontology/diffs	

```
File is too large to display.
```

```
File is too large to display.
```

```
Update to HEAD: trunk/src at revision 269 (17 entries)
  trunk/src/ontology/Makefile (Modified)
  trunk/src/ontology/util/add-stuff-to-imports.pl (Added)
  trunk/src/ontology/cl-basic.owl (Modified)
  trunk/src/ontology/cl.obo (Modified)
  trunk/src/ontology/cl.owl (Modified)
  trunk/src/ontology/imports/catalog-v001.xml (Modified)
```

No cached information could be found – build cache

Selected https://cell-ontology.googlecode.com/svn/trunk/src/ontology/cl-basic.owl 1 File

The coolest Interface to (Sub)Version Control

TortoiseSVN

mercurial

git

Random suggestion

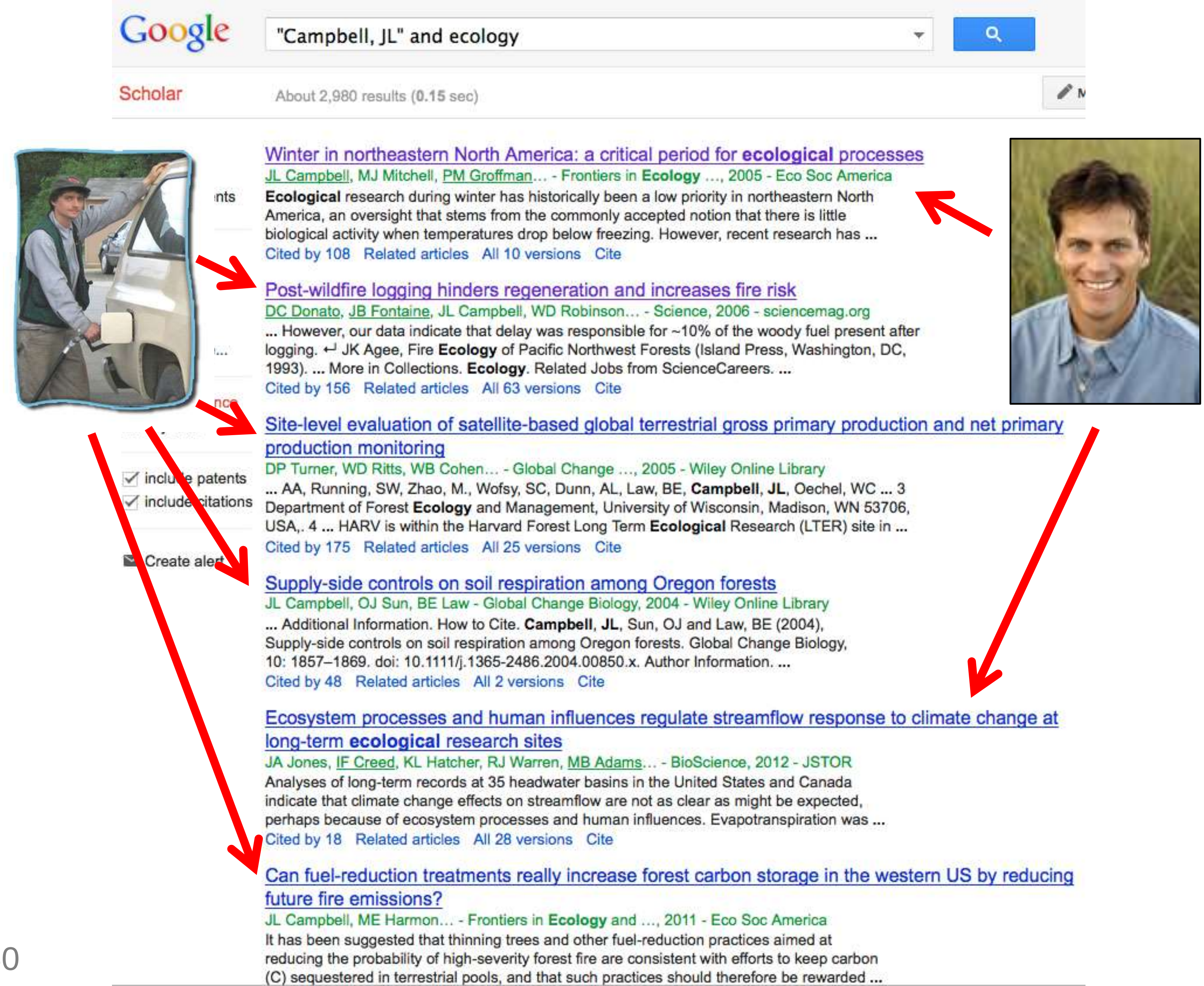
Disambiguate yourself

ORCID Connecting Research and Researchers

Open Researcher & Contributor ID

John L. Campbell
Forest Research Ecologist
Oregon State University, Corvallis, OR

John L. Campbell
Forest Research Ecologist
Center for Research on Ecosystem Change
US Forest Service, Durham, NC



"Campbell, JL" and ecology

About 2,980 results (**0.15** sec)

☑ include patents
☑ include citations

[Winter in northeastern North America: a critical period for **ecological** processes](#)
JL Campbell, MJ Mitchell, PM Groffman... - Frontiers in **Ecology** ..., 2005 - Eco Soc America
Ecological research during winter has historically been a low priority in northeastern North America, an oversight that stems from the commonly accepted notion that there is little biological activity when temperatures drop below freezing. However, recent research has ...
Cited by 108 Related articles All 10 versions Cite

[Post-wildfire logging hinders regeneration and increases fire risk](#)
DC Donato, JB Fontaine, JL Campbell, WD Robinson... - Science, 2006 - sciencemag.org
... However, our data indicate that delay was responsible for ~10% of the woody fuel present after logging. ↵ JK Agee, Fire **Ecology** of Pacific Northwest Forests (Island Press, Washington, DC, 1993). ... More in Collections. **Ecology**. Related Jobs from ScienceCareers. ...
Cited by 156 Related articles All 63 versions Cite

[Site-level evaluation of satellite-based global terrestrial gross primary production and net primary production monitoring](#)
DP Turner, WD Ritts, WB Cohen... - Global Change ..., 2005 - Wiley Online Library
... AA, Running, SW, Zhao, M., Wofsy, SC, Dunn, AL, Law, BE, **Campbell**, **JL**, Oechel, WC ... 3 Department of Forest **Ecology** and Management, University of Wisconsin, Madison, WN 53706, USA,. 4 ... HARV is within the Harvard Forest Long Term **Ecological** Research (LTER) site in ...
Cited by 175 Related articles All 25 versions Cite

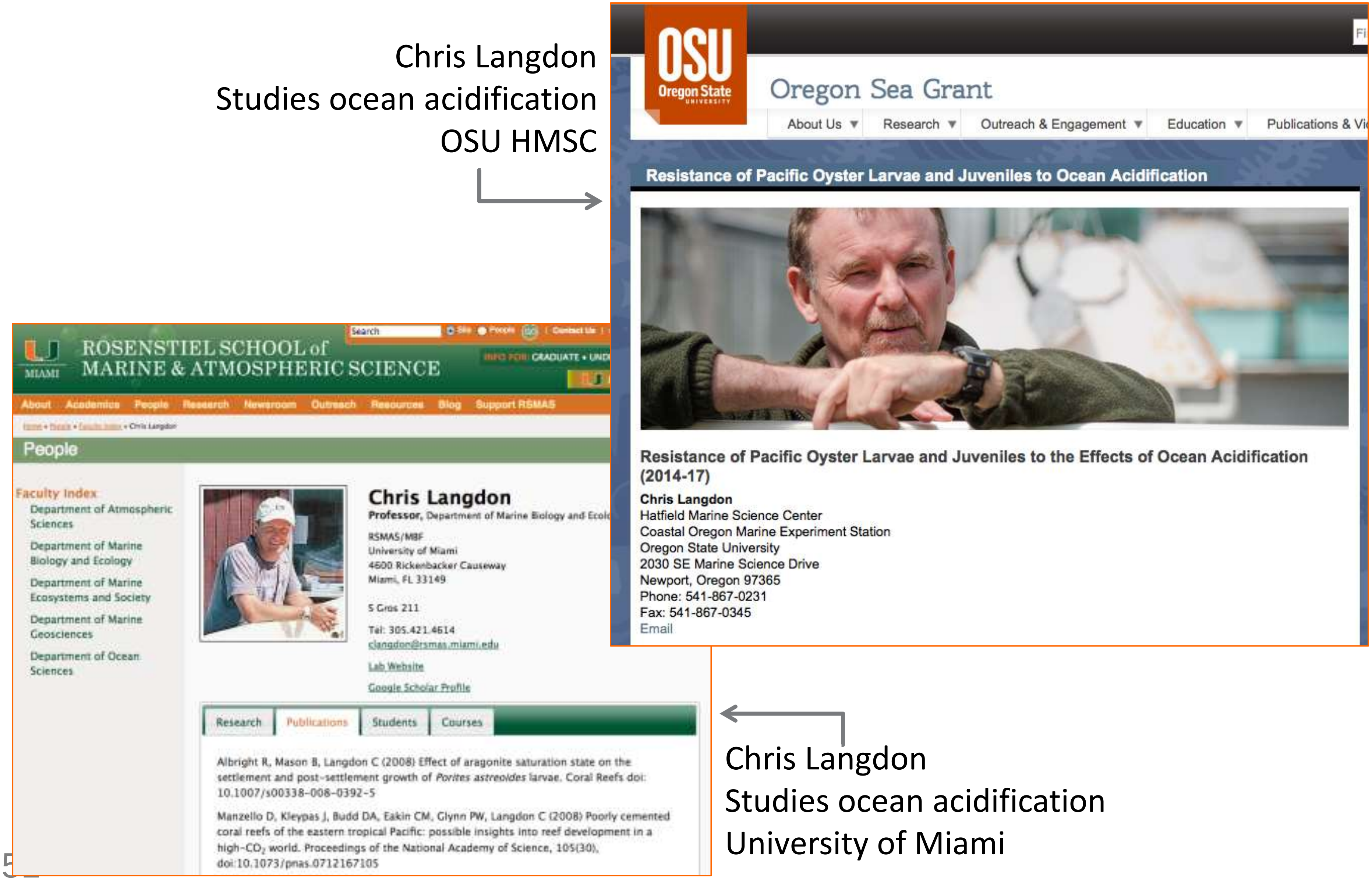
[Supply-side controls on soil respiration among Oregon forests](#)
JL Campbell, OJ Sun, BE Law - Global Change Biology, 2004 - Wiley Online Library
... Additional Information. How to Cite. **Campbell**, **JL**, Sun, OJ and Law, BE (2004), Supply-side controls on soil respiration among Oregon forests. Global Change Biology, 10: 1857–1869. doi: 10.1111/j.1365-2486.2004.00850.x. Author Information. ...
Cited by 48 Related articles All 2 versions Cite

[Ecosystem processes and human influences regulate streamflow response to climate change at long-term **ecological** research sites](#)
JA Jones, IF Creed, KL Hatcher, RJ Warren, MB Adams... - BioScience, 2012 - JSTOR
Analyses of long-term records at 35 headwater basins in the United States and Canada indicate that climate change effects on streamflow are not as clear as might be expected, perhaps because of ecosystem processes and human influences. Evapotranspiration was ...
Cited by 18 Related articles All 28 versions Cite

[Can fuel-reduction treatments really increase forest carbon storage in the western US by reducing future fire emissions?](#)
JL Campbell, ME Harmon... - Frontiers in **Ecology** and ..., 2011 - Eco Soc America
It has been suggested that thinning trees and other fuel-reduction practices aimed at reducing the probability of high-severity forest fire are consistent with efforts to keep carbon (C) sequestered in terrestrial pools, and that such practices should therefore be rewarded ...

Another Doppelgänger

Chris Langdon
Studies ocean acidification
OSU HMSC



Resistance of Pacific Oyster Larvae and Juveniles to Ocean Acidification

Resistance of Pacific Oyster Larvae and Juveniles to the Effects of Ocean Acidification (2014-17)

Chris Langdon
Hatfield Marine Science Center
Coastal Oregon Marine Experiment Station
Oregon State University
2030 SE Marine Science Drive
Newport, Oregon 97365
Phone: 541-867-0231
Fax: 541-867-0345
Email

Chris Langdon
Professor, Department of Marine Biology and Ecol

RSMAS/MBF
University of Miami
4600 Rickenbacker Causeway
Miami, FL 33149

S Gros 211

Tel: 305.421.4614
clangdon@rsmas.miami.edu

Lab Website

Google Scholar Profile

Albright R, Mason B, Langdon C (2008) Effect of aragonite saturation state on the settlement and post-settlement growth of *Porites astreoides* larvae. Coral Reefs doi: 10.1007/s00338-008-0392-5

Manzello D, Kleypas J, Budd DA, Eakin CM, Glynn PW, Langdon C (2008) Poorly cemented coral reefs of the eastern tropical Pacific: possible insights into reef development in a high-CO_2 world. Proceedings of the National Academy of Science, 105(30), doi:10.1073/pnas.0712167105

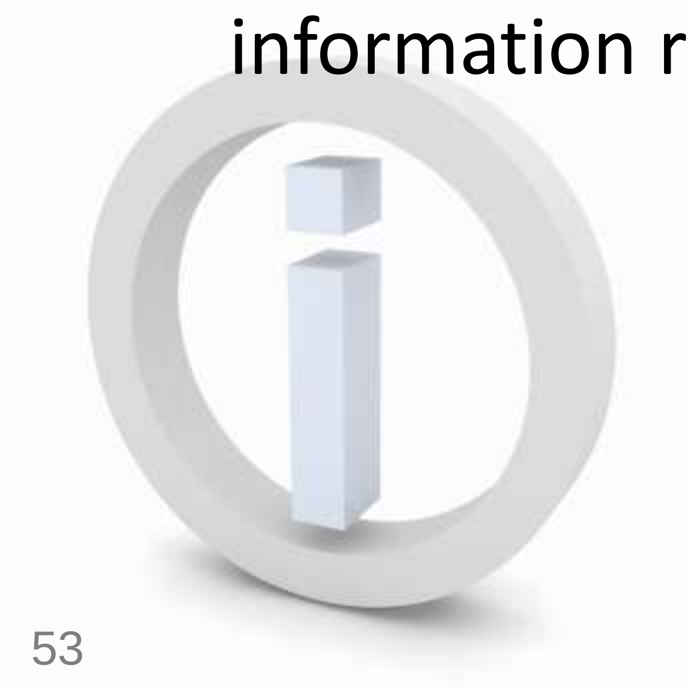
Chris Langdon
Studies ocean acidification
University of Miami

Metadata

What is metadata?

- Data about data
- Structured information that describes, explains, locates, or otherwise makes it easier to retrieve, use, or manage an information resource.

NISO, *Understanding Metadata*



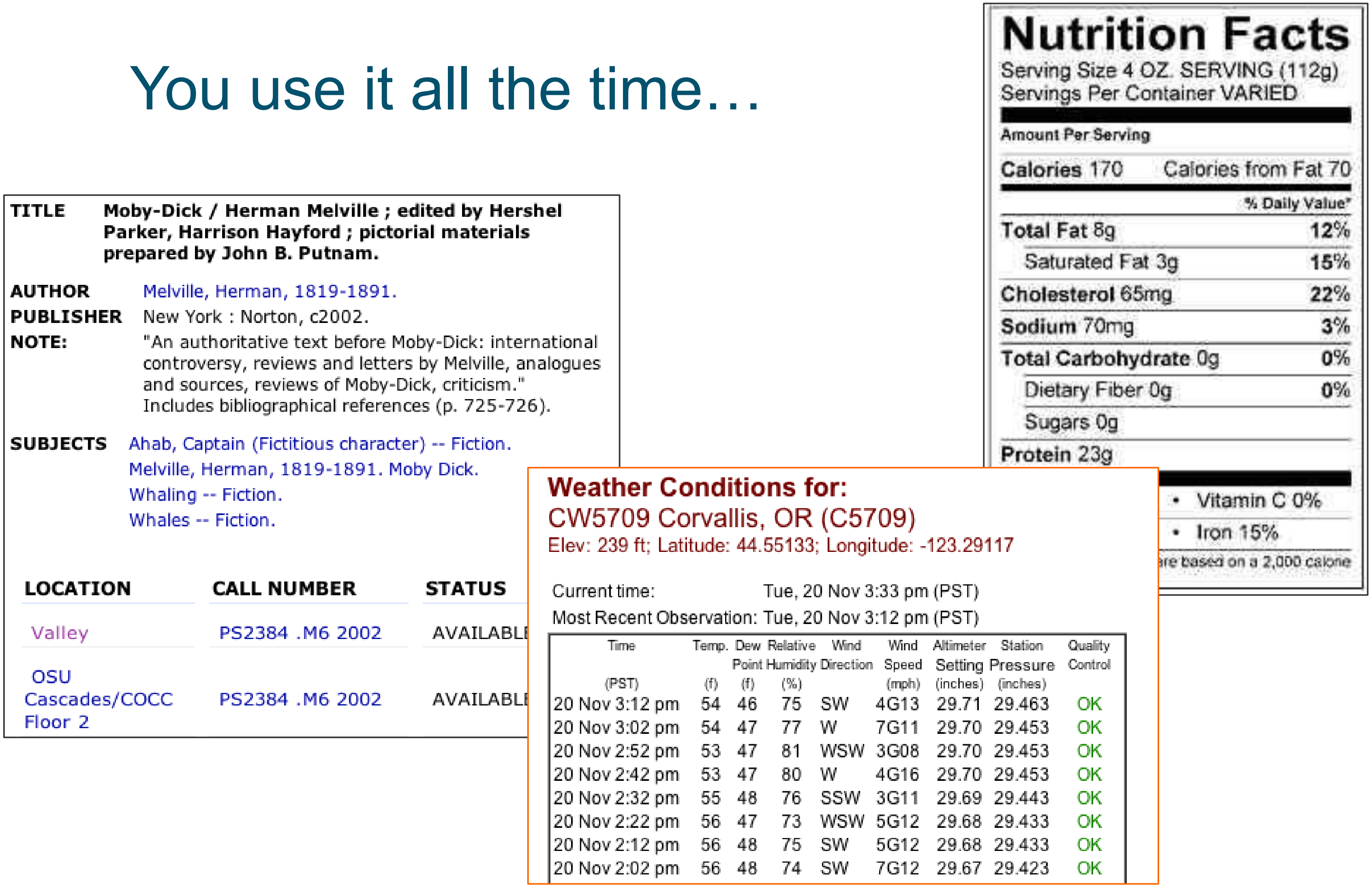
Metadata

“The metadata accompanying your data should be written for a user ***20 years into the future*** -- what does that person need to know to use your data properly? Prepare the metadata for a user who is unfamiliar with your project, methods, or observations.”

Oak Ridge National Laboratory Distributed Active
Archive Center for Biogeochemical Dynamics
(ORNL DAAC)

Metadata in real life

You use it all the time…



TITLE	**Moby-Dick / Herman Melville ; edited by Hershel Parker, Harrison Hayford ; pictorial materials prepared by John B. Putnam.**
AUTHOR	Melville, Herman, 1819-1891.
PUBLISHER	New York : Norton, c2002.
NOTE:	"An authoritative text before Moby-Dick: international controversy, reviews and letters by Melville, analogues and sources, reviews of Moby-Dick, criticism." Includes bibliographical references (p. 725-726).
SUBJECTS	Ahab, Captain (Fictitious character) -- Fiction. Melville, Herman, 1819-1891. Moby Dick. Whaling -- Fiction. Whales -- Fiction.

LOCATION	CALL NUMBER	STATUS
Valley	PS2384 .M6 2002	AVAILABL
OSU Cascades/COCC Floor 2	PS2384 .M6 2002	AVAILABL

Nutrition Facts

Serving Size 4 OZ. SERVING (112g)
Servings Per Container VARIED

Amount Per Serving

Calories 170	Calories from Fat 70
	% Daily Value*
Total Fat 8g	**12%**
Saturated Fat 3g	**15%**
Cholesterol 65mg	**22%**
Sodium 70mg	**3%**
Total Carbohydrate 0g	**0%**
Dietary Fiber 0g	**0%**
Sugars 0g	
Protein 23g	

- Vitamin C 0%
- Iron 15%

re based on a 2,000 calorie

Weather Conditions for:

CW5709 Corvallis, OR (C5709)

Elev: 239 ft; Latitude: 44.55133; Longitude: -123.29117

Current time: Tue, 20 Nov 3:33 pm (PST)
Most Recent Observation: Tue, 20 Nov 3:12 pm (PST)

Time (PST)	Temp. (f)	Dew Point (f)	Relative Humidity (%)	Wind Direction	Wind Speed (mph)	Altimeter Setting (inches)	Station Pressure (inches)	Quality Control
20 Nov 3:12 pm	54	46	75	SW	4G13	29.71	29.463	OK
20 Nov 3:02 pm	54	47	77	W	7G11	29.70	29.453	OK
20 Nov 2:52 pm	53	47	81	WSW	3G08	29.70	29.453	OK
20 Nov 2:42 pm	53	47	80	W	4G16	29.70	29.453	OK
20 Nov 2:32 pm	55	48	76	SSW	3G11	29.69	29.443	OK
20 Nov 2:22 pm	56	47	73	WSW	5G12	29.68	29.433	OK
20 Nov 2:12 pm	56	48	75	SW	5G12	29.68	29.433	OK
20 Nov 2:02 pm	56	48	74	SW	7G12	29.67	29.423	OK

Major metadata standards

Darwin Core | **biological diversity, taxonomy**

Dublin Core | **general**

DDI (Data Documentation Initiative) | **social & behavioral sciences**

DIF (Directory Interchange Format) | **environmental sciences**

EML (Ecological Metadata Language) | **ecology, biology**

ISO 19115 | **geographic data**

01101101 01100101 01110100 01100001
01100100 01100001 01110100 01100001

Metadata examples

Santa Barbara Coastal Long Term Ecological Research (LTER)

web link

Bureau of Labor Statistics, Consumer Price Index, 1913-1992

web link

Legal & ethical considerations

Data sharing & reuse

"…digitally formatted scientific data resulting from unclassified research supported wholly or in part by Federal funding should be stored and **publicly accessible** to search, retrieve, and analyze."

Office of Science and Technology Policy
The White House

How to preserve & share data

PANGAEA

RCSB PDB
PROTEIN DATA BANK

figshare
credit for **all** your research

ICPSR
A PARTNER IN SOCIAL SCIENCE RESEARCH
INTER-UNIVERSITY CONSORTIUM FOR POLITICAL AND SOCIAL RESEARCH

DRYAD

National Oceanographic Data Center

How to preserve & share data

ScholarsArchive@OSU

SA@OSU Home

Open Access at OSU

Increase the visibility and impact of your scholarship with ScholarsArchive@OSU.

About ScholarsArchive@OSU

ScholarsArchive@OSU is Oregon State University's digital service for gathering, indexing, making available and storing the scholarly work of the Oregon State University community. It also includes materials from outside the institution in support of the university's land, sun, sea and space grant missions and other research interests.

- FAQ
- Open Access

Search ScholarsArchive@OSU

Advanced Search

Browse

All of ScholarsArchive@OSU
    Communities & Collections
    Issue Date
    Authors
    Titles
    Subjects

My Account

Login
Register

Discover

Author
    Oregon State University. Extension Service (2848)
    Forest Products Laboratory (U.S.) (1912)
    Oregon State University. Agricultural

Data journals

Earth Syst. Sci. Data, 4, 47-73, 2012
www.earth-syst-sci-data.net/4/47/2012/
doi:10.5194/essd-4-47-2012
© Author(s) 2012. This work is distributed
under the Creative Commons Attribution 3.0 License.

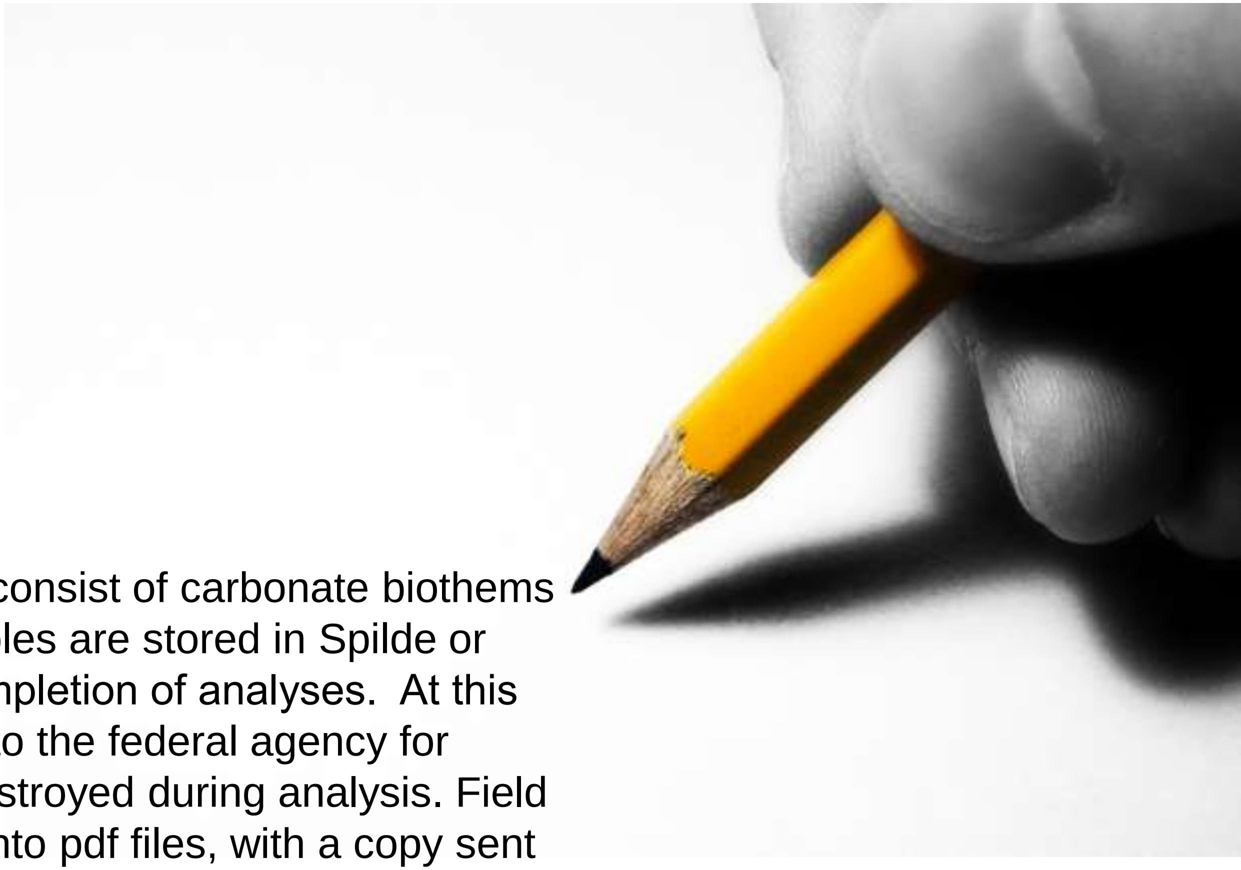
Database of diazotrophs in global ocean: abundance, biomass and nitrogen fixation rates

Y.-W. Luo[1], S. C. Doney[1], L. A. Anderson[2], M. Benavides[3], I. Berman-Frank[4], A. Bode[5], S. Bonnet[6], K. H. Boström[7], D. Böttjer[8], D. G. Capone[9], E. J. Carpenter[10], Y. L. Chen[11], M. J. Church[8], J. E. Dore[12], L. I. Falcón[13], A. Fernández[14], R. A. Foster[15], K. Furuya[16], F. Gómez[17], K. Gundersen[18], A. M. Hynes[19,*], D. M. Karl[8], S. Kitajima[16], R. J. Langlois[20], J. LaRoche[20], R. M. Letelier[21], E. Marañón[14], D. J. McGillicuddy Jr.[2], P. H. Moisander[22,], C. M. Moore[23], B. Mouriño-Carballido[14], M. R. Mulholland[24], J. A. Needoba[25], K. M. Orcutt[18], A. J. Poulton[26], E. Rahav[4], P. Raimbault[6], A. P. Rees[27], L. Riemann[28], T. Shiozaki[16], A. Subramaniam[29], T. Tyrrell[23], K. A. Turk-Kubo[22], M. Varela[5], T. A. Villareal[30], E. A. Webb[9], A. E. White[21], J. Wu[31], and J. P. Zehr[22]**

[1]Department of Marine Chemistry and Geochemistry, Woods Hole Oceanographic Institution, Woods Hole, Massachusetts 02543, USA
[2]Department of Applied Ocean Science and Engineering, Woods Hole Oceanographic Institution, Woods Hole, Massachusetts 02543, USA
[3]Instituto de Oceanografía y Cambio Global, Universidad de Las Palmas de Gran Canaria, 35017, Las Palmas de Gran Canaria, Spain
[4]The Mina and Everard Goodman Faculty of Life Sciences, Bar-Ilan University, Ramat Gan 52900, Israel
[5]Instituto Español de Oceanografía, Centro Oceanográfico de A Coruña, 15080 A Coruña, Spain
[6]IRD-INSU-CNRS, Laboratoire d'Océanographie Physique et Biogéochimique, UMR 6535, Centre d'Océanologie de Marseille, Aix Marseille Université, France
[7]Department of Natural Sciences, Linnaeus University, 39182 Kalmar, Sweden
[8]School of Ocean and Earth Science and Technology, University of Hawaii, Honolulu, Hawaii 96822, USA
[9]Department of Biological Sciences and Wrigley Institute for Environmental Studies, University of Southern California, Los Angeles, California 90089, USA
[10]Romberg Tiburon Center, San Francisco State University, Tiburon, California 94920, USA
[11]Department of Marine Biotechnology and Resources, National Sun Yat-sen University, Kaohsiung 80424, Taiwan
[12]Department of Land Resources and Environmental Sciences, Montana State University, Bozeman, MT 59717, USA
[13]Laboratorio de Ecología Bacteriana, Instituto de Ecología, Universidad Nacional Autónoma de México, Mexico

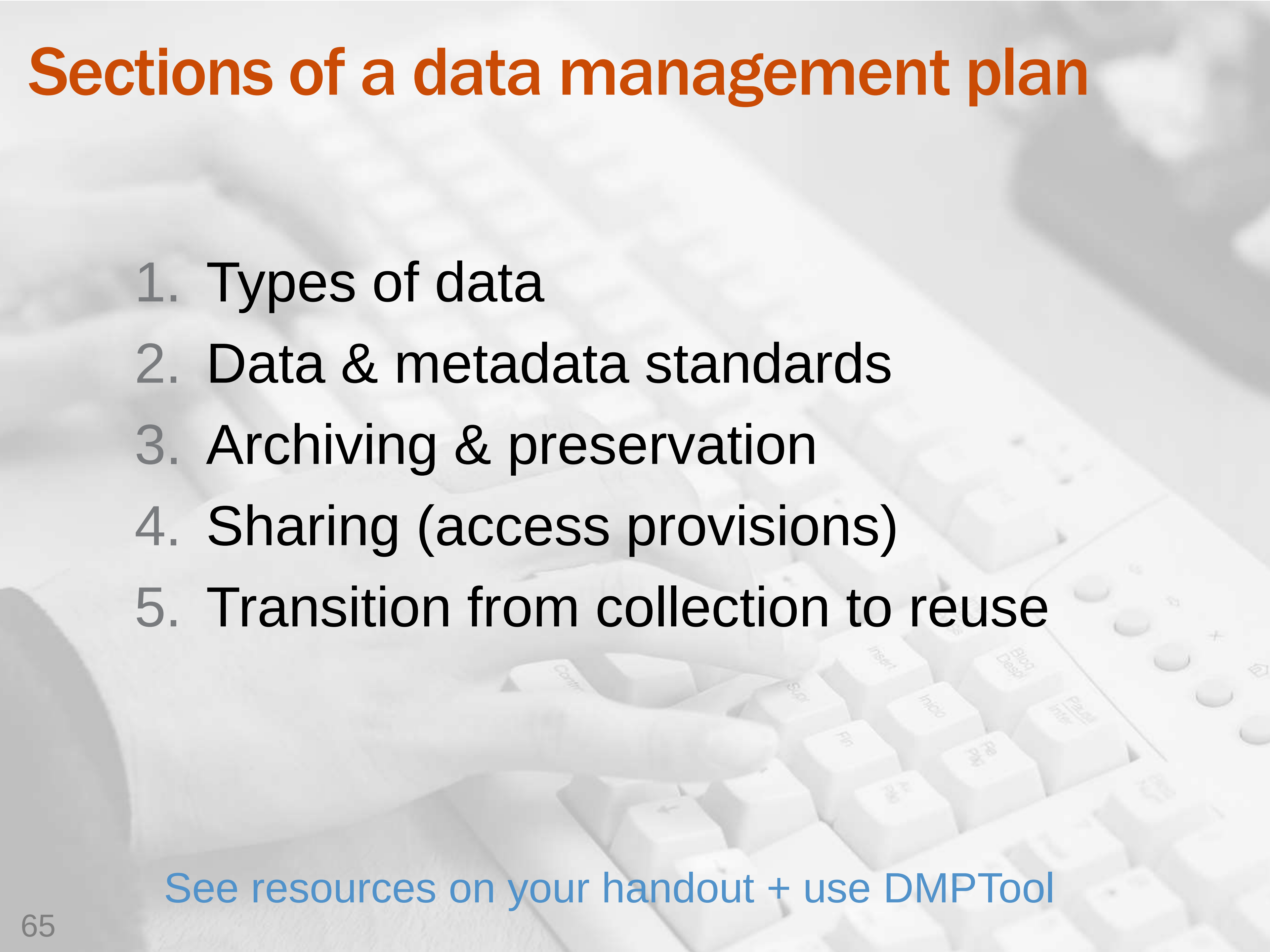
Data management plans

What is a data management plan?



…Physical specimens consist of carbonate biothems and filtered water samples are stored in Spilde or Northup's labs until completion of analyses. At this time they are returned to the federal agency for museum curation or destroyed during analysis. Field notes will be scanned into pdf files, with a copy sent to Carlsbad Caverns National Park or other federal cave manager. SEM images will be saved in the tif format. …

Sections of a data management plan

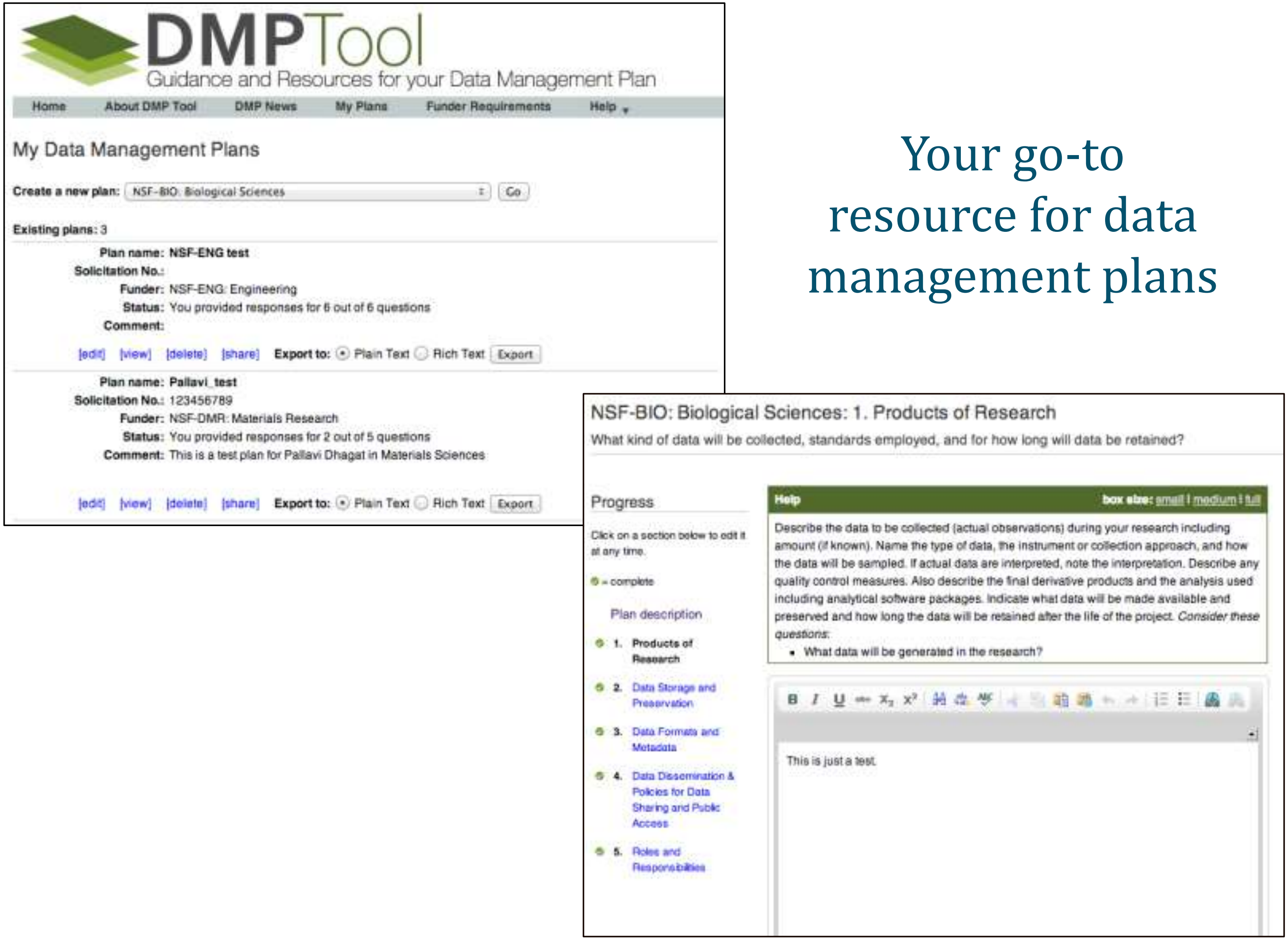


1. Types of data
2. Data & metadata standards
3. Archiving & preservation
4. Sharing (access provisions)
5. Transition from collection to reuse

See resources on your handout + use DMPTool

DMPTool
Build your Data Management Plan

| Visit DMPTool



My Data Management Plans

Create a new plan: NSF-BIO: Biological Sciences [Go]

Existing plans: 3

Plan name: NSF-ENG test
Solicitation No.:
Funder: NSF-ENG: Engineering
Status: You provided responses for 6 out of 6 questions
Comment:

[edit] [view] [delete] [share] **Export to:** ● Plain Text ○ Rich Text [Export]

Plan name: Pallavi_test
Solicitation No.: 123456789
Funder: NSF-DMR: Materials Research
Status: You provided responses for 2 out of 5 questions
Comment: This is a test plan for Pallavi Dhagat in Materials Sciences

[edit] [view] [delete] [share] **Export to:** ● Plain Text ○ Rich Text [Export]

NSF-BIO: Biological Sciences: 1. Products of Research

What kind of data will be collected, standards employed, and for how long will data be retained?

Progress

Click on a section below to edit it at any time.

● = complete

Plan description

- ● **1. Products of Research**
- ● 2. Data Storage and Preservation
- ● 3. Data Formats and Metadata
- ● 4. Data Dissemination & Policies for Data Sharing and Public Access
- ● 5. Roles and Responsibilities

Help — **box size:** small | medium | full

Describe the data to be collected (actual observations) during your research including amount (if known). Name the type of data, the instrument or collection approach, and how the data will be sampled. If actual data are interpreted, note the interpretation. Describe any quality control measures. Also describe the final derivative products and the analysis used including analytical software packages. Indicate what data will be made available and preserved and how long the data will be retained after the life of the project. *Consider these questions:*

- What data will be generated in the research?

This is just a test.

Your go-to resource for data management plans

Contact information

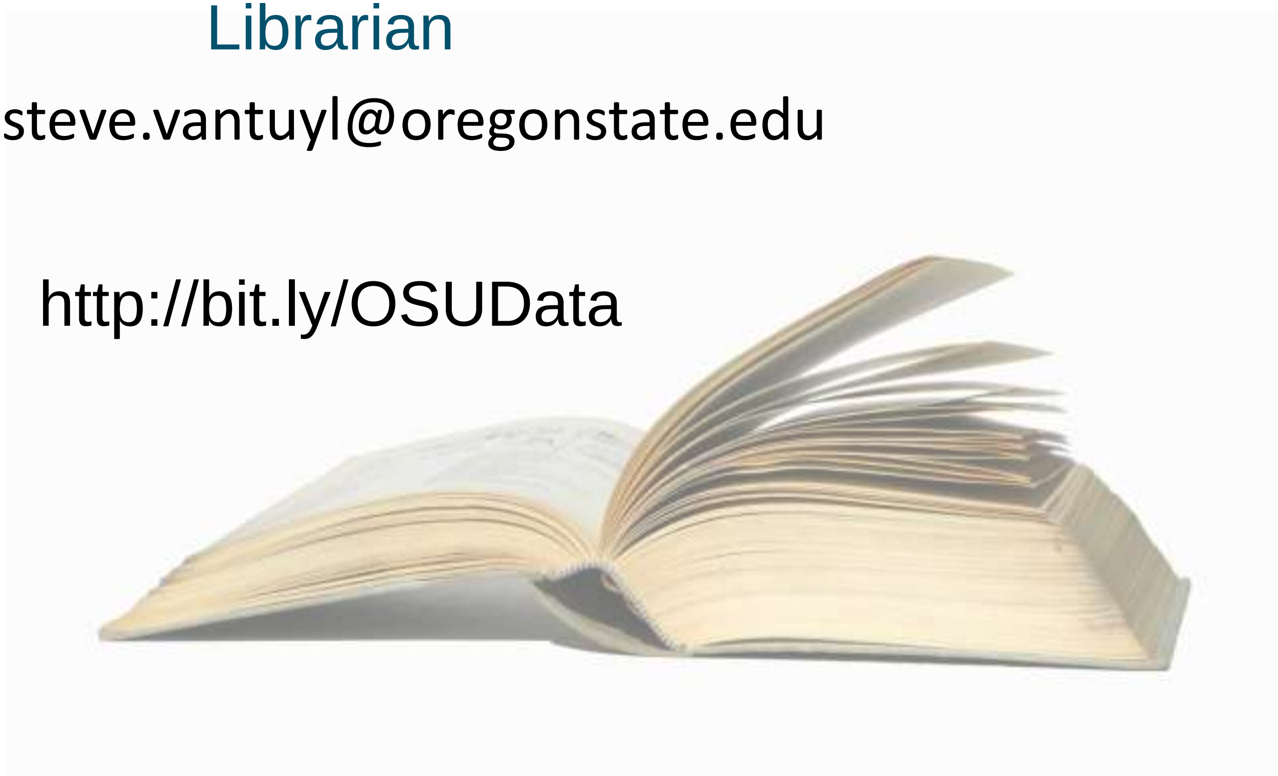
Amanda Whitmire | Data Management Specialist

amanda.whitmire@oregonstate.edu

Steve Van Tuyl| Data and Digital Repository Librarian

steve.vantuyl@oregonstate.edu

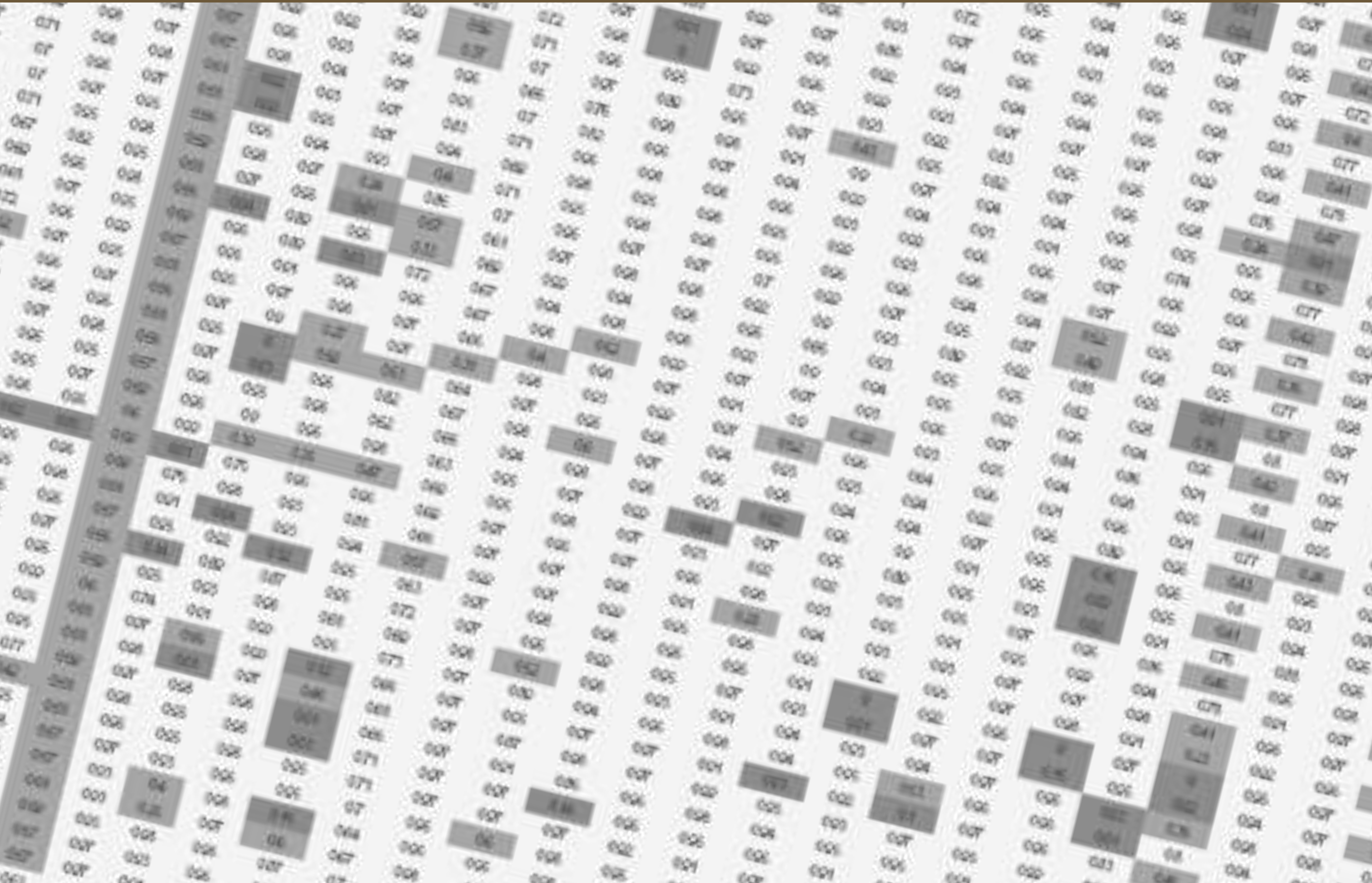
http://bit.ly/OSUData

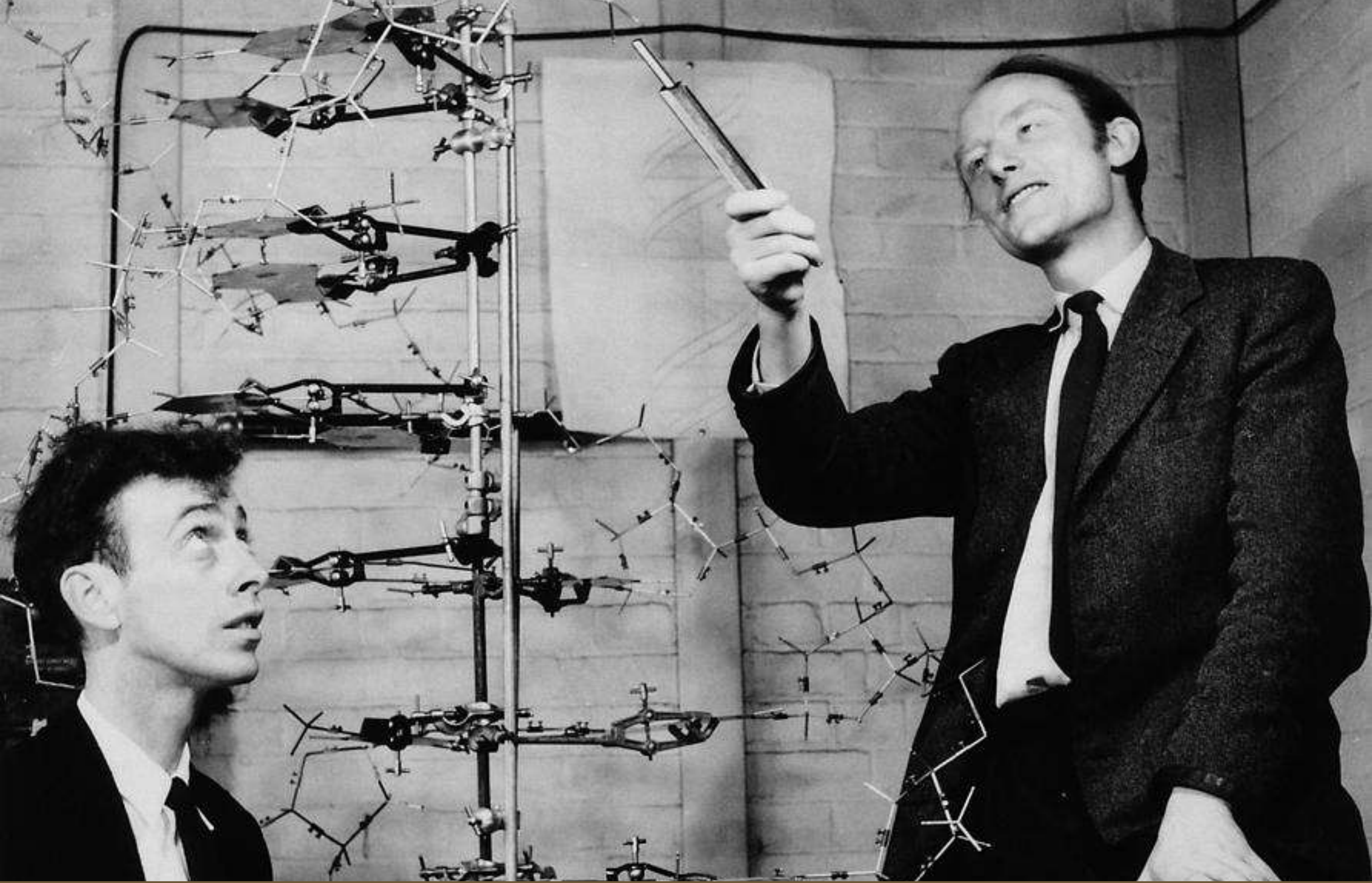


end

Extra / Outdated Slides

Data does not speak for itself...

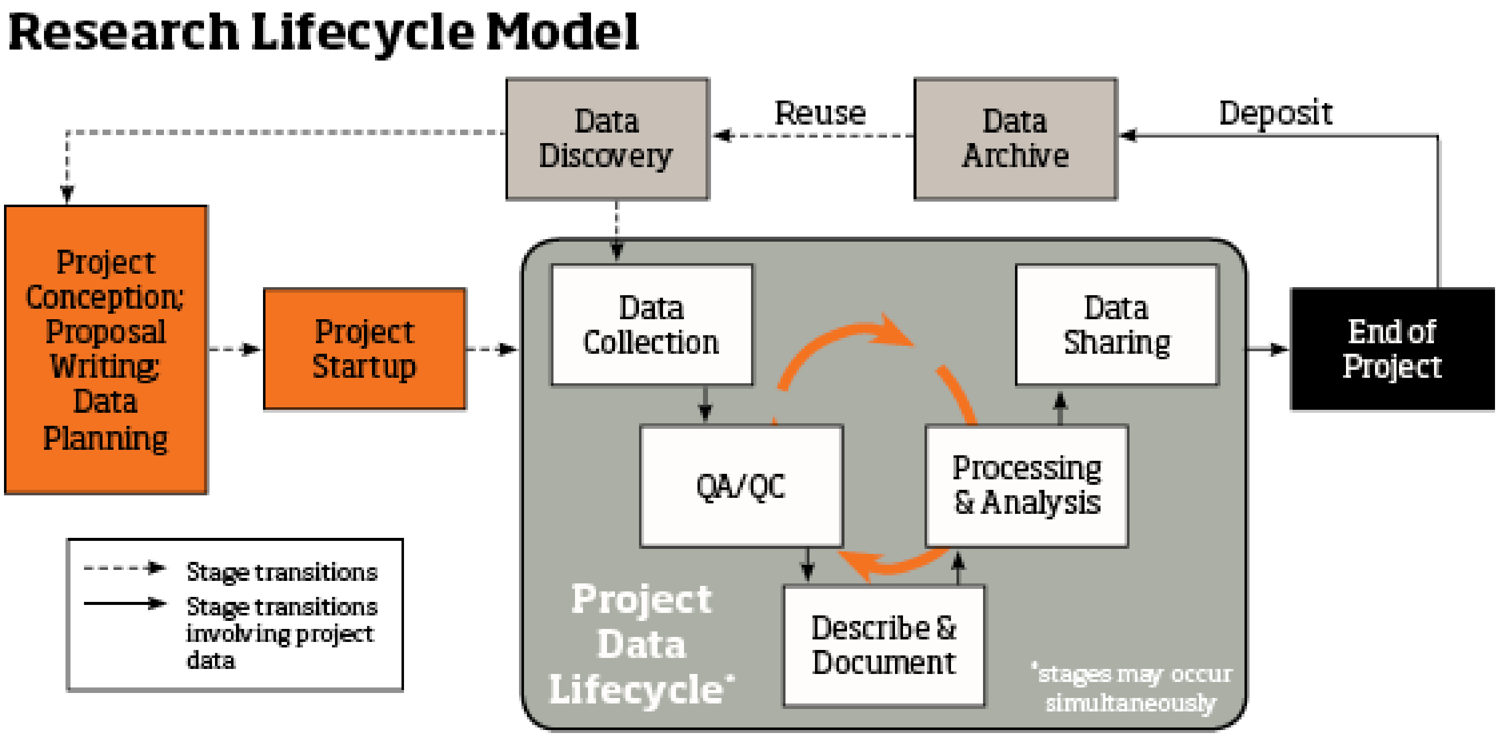




YOU speak for YOUR data

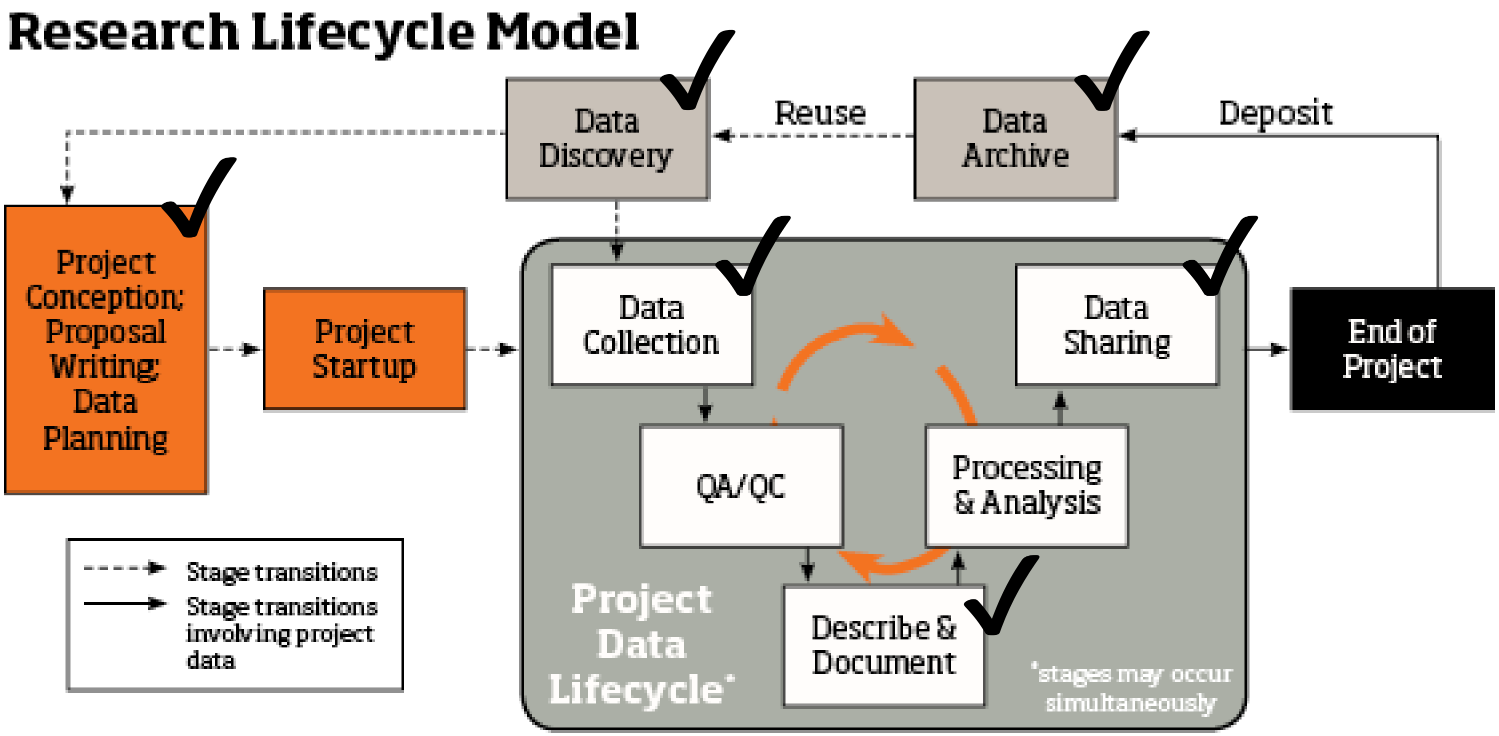
But first, you need to manage it

Data management in the lifecycle



Adapted from: http://dmconsult.library.virginia.edu/lifecycle
http://www.icpsr.umich.edu/icpsrweb/content/deposit/guide/#cycle

How can OSU Libraries help?



Adapted from: http://dmconsult.library.virginia.edu/lifecycle
http://www.icpsr.umich.edu/icpsrweb/content/deposit/guide/#cycle

Copyright – who owns the data?

Copyright applies to text, images, recordings, videos etc. in a digital form, in the same way that it would to analog versions of those works. The code of computer programs (both the human readable source code and the machine readable object code) is protected by copyright as a literary work. Data compilations such as datasets and databases can be protected by copyright in the literary works category, which includes ‘tables’ or ‘compilations’. A table or compilation, consisting of words, figures or symbols (or a combination of these) is protected if it is a literary work and has the required degree of originality.

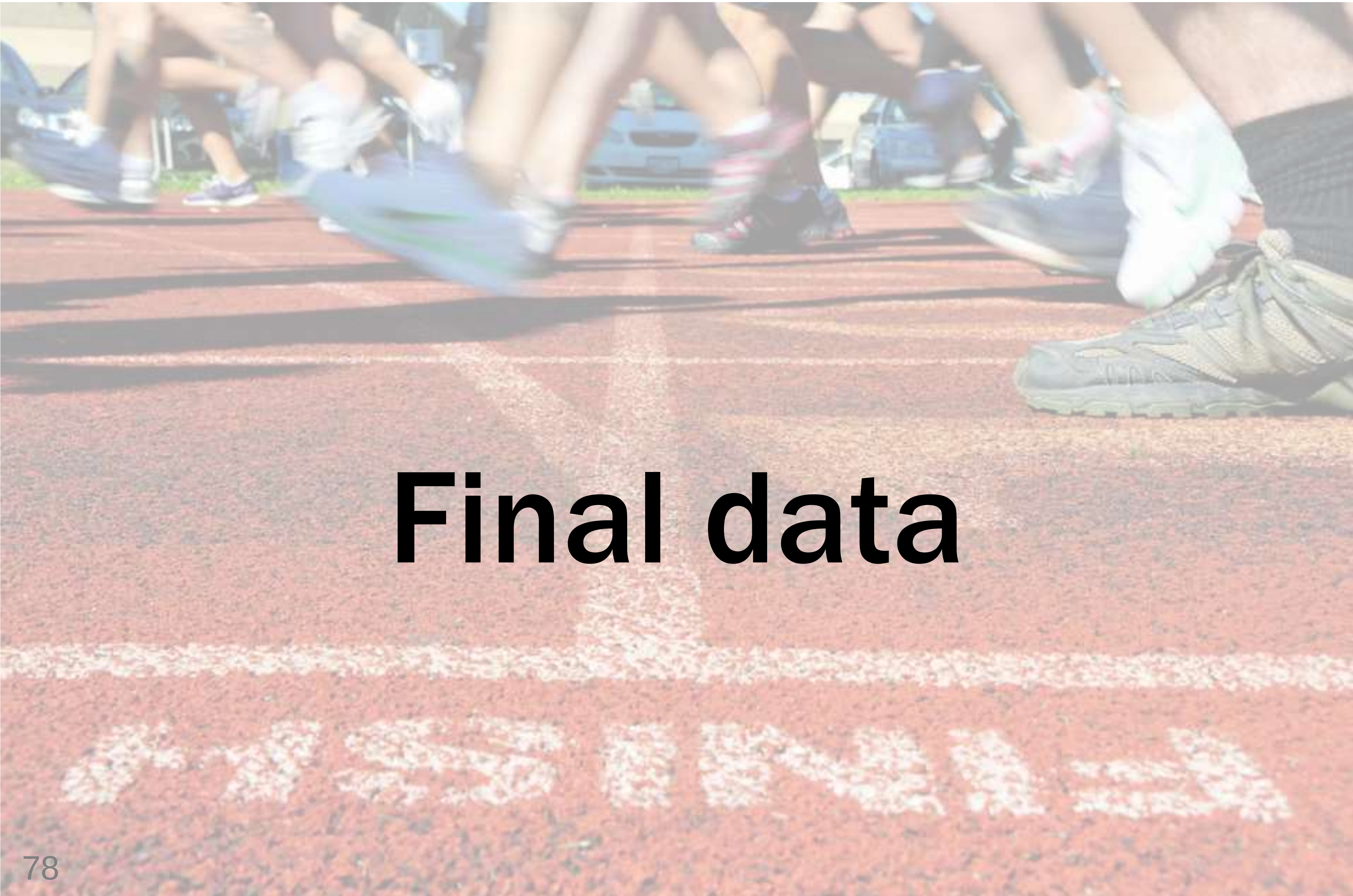
Types, formats & stages of data

Raw data

Types, formats & stages of data

Intermediate data

Types, formats & stages of data



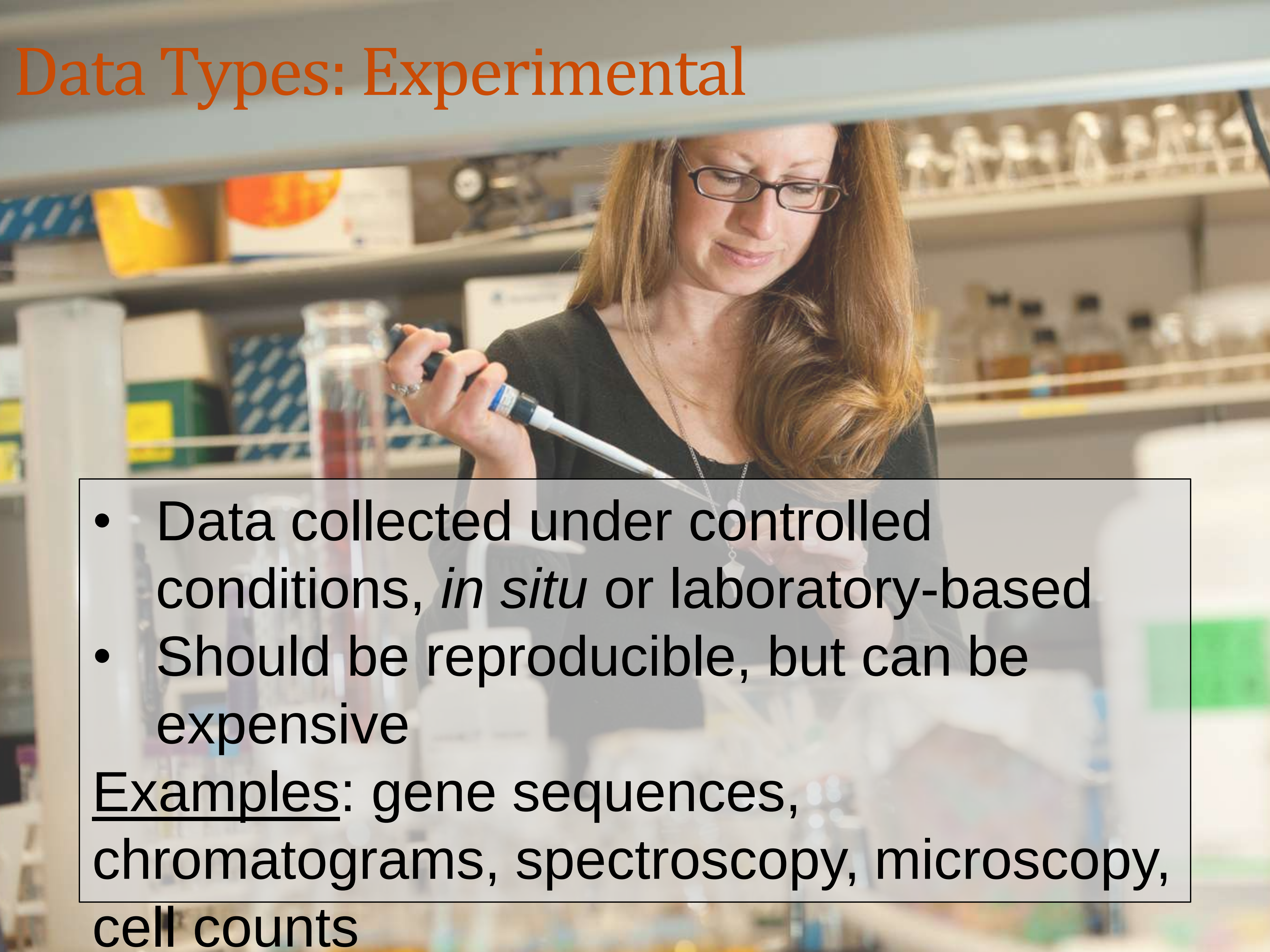
Final data

Data Types: Observational

- Captured *in situ*
- Can’t be recaptured, recreated or replaced

Examples: Sensor readings, sensory (human) observations, survey results

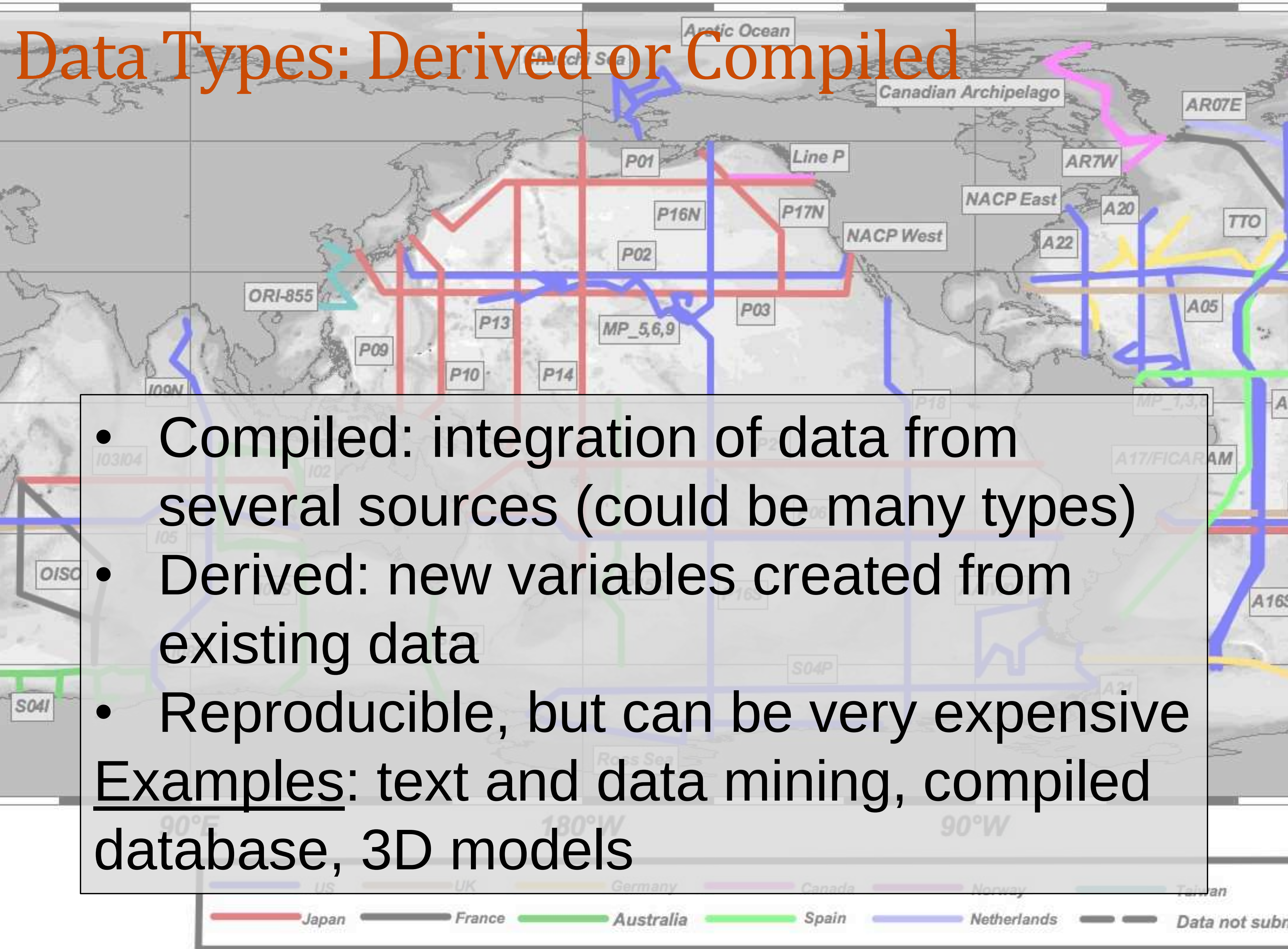
Data Types: Experimental



- Data collected under controlled conditions, *in situ* or laboratory-based
- Should be reproducible, but can be expensive

Examples: gene sequences, chromatograms, spectroscopy, microscopy, cell counts

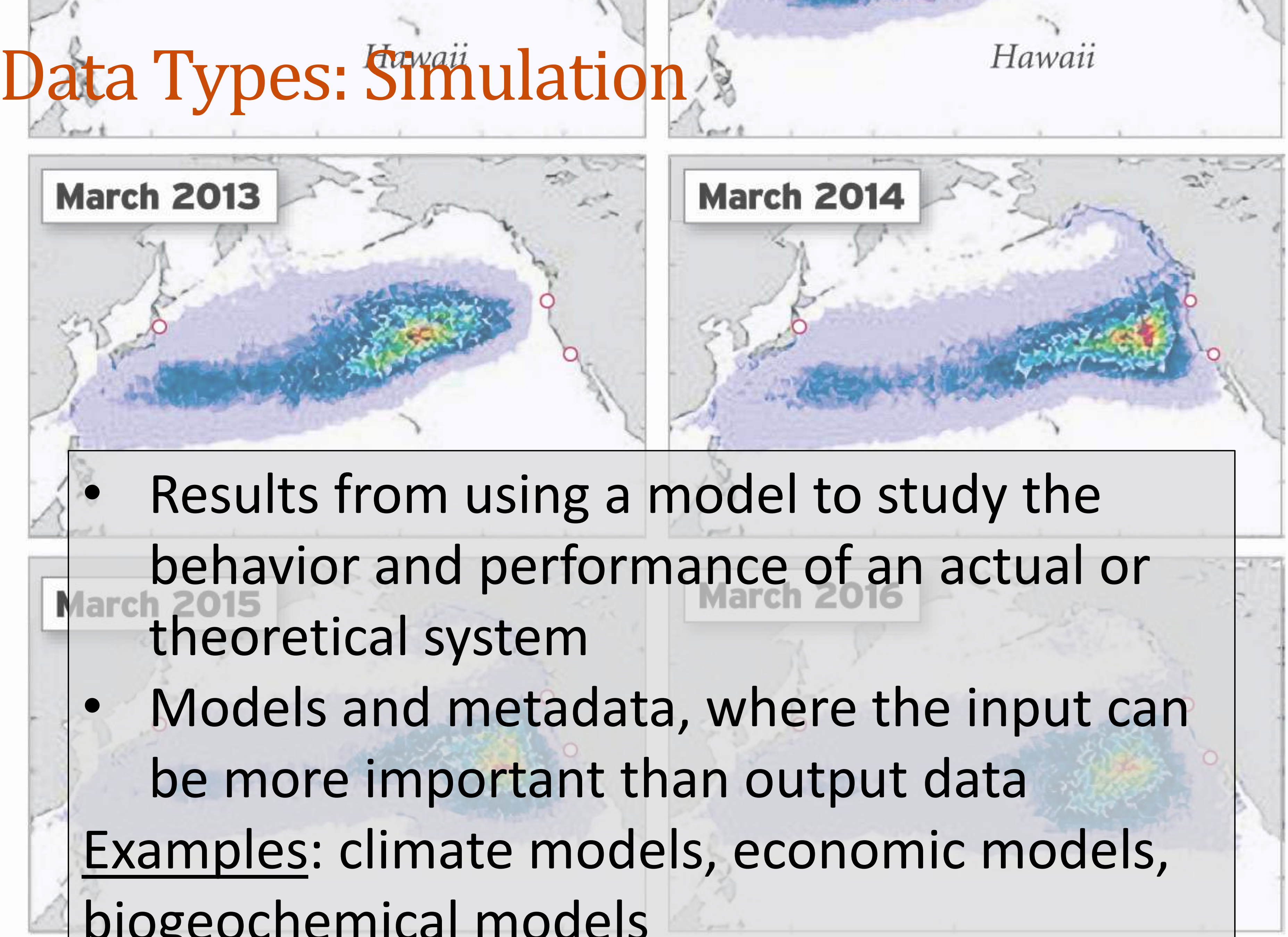
Data Types: Derived or Compiled



- Compiled: integration of data from several sources (could be many types)
- Derived: new variables created from existing data
- Reproducible, but can be very expensive

<u>Examples</u>: text and data mining, compiled database, 3D models

Data Types: Simulation



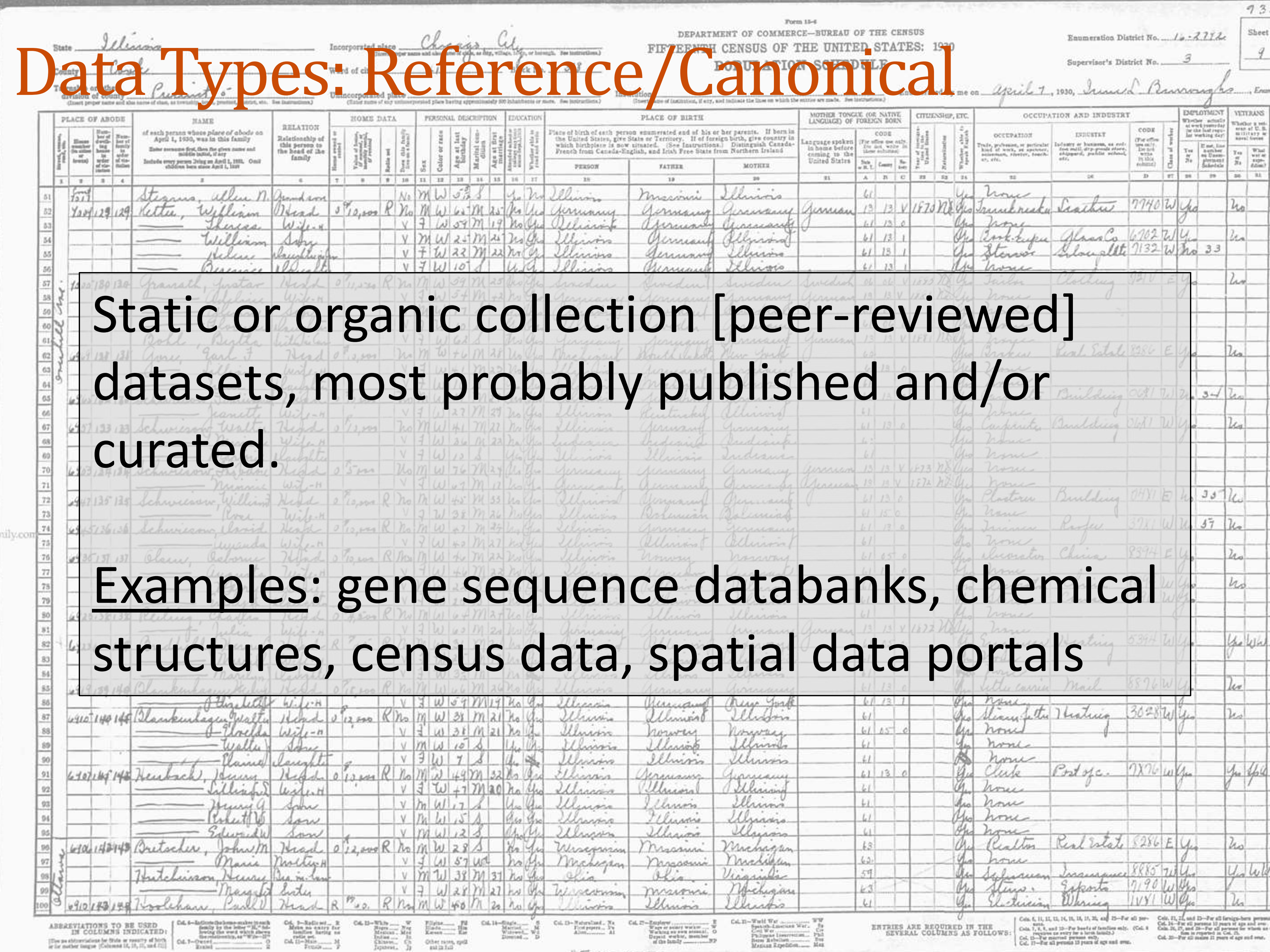
Source: Nikolai Maximenko, International Pacific Research Center

THE OREGONIAN

- Results from using a model to study the behavior and performance of an actual or theoretical system
- Models and metadata, where the input can be more important than output data

Examples: climate models, economic models, biogeochemical models

Data Types: Reference/Canonical


DEPARTMENT OF COMMERCE—BUREAU OF THE CENSUS
FIFTEENTH CENSUS OF THE UNITED STATES: 1930

Static or organic collection [peer-reviewed] datasets, most probably published and/or curated.

<u>Examples</u>: gene sequence databanks, chemical structures, census data, spatial data portals

Data storage & preservation

About backups...

Do them.

University of Southampton, School of Electronics and Computer Science
Southampton, UK, 2005

Plan for unexpected events

Metadata demonstration

Colectica