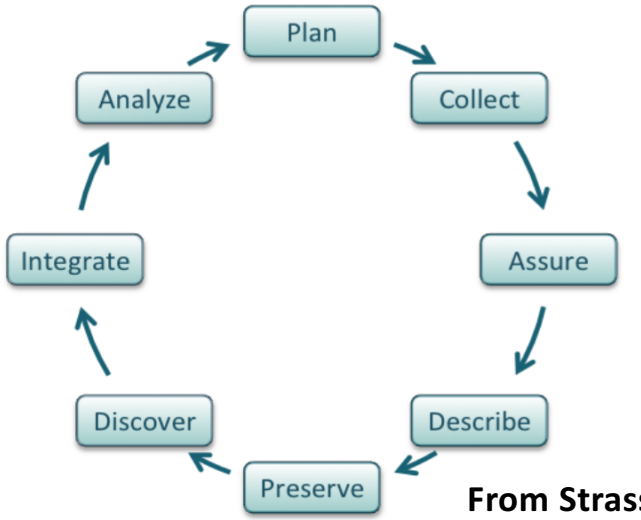
How to think About Your Data: Introduction to Data Science & Management

(*what they don't – but should - teach you about the scientific method*)

Brian J. Enquist

(borrowing heavily from presentations by C.A. Strasser and Ethan White)



From Strasser et al.

Some Simple Guidelines for Effective Data Management

- Most scientists spend much time thinking about the types of data they need to further their studies
- Relatively little effort is spent considering how to store, analyze, share their data
- it is increasingly important to store and document scientific data in ways that facilitate: (i) Open Science; and (ii) their effective retrieval and interpretation in the future.

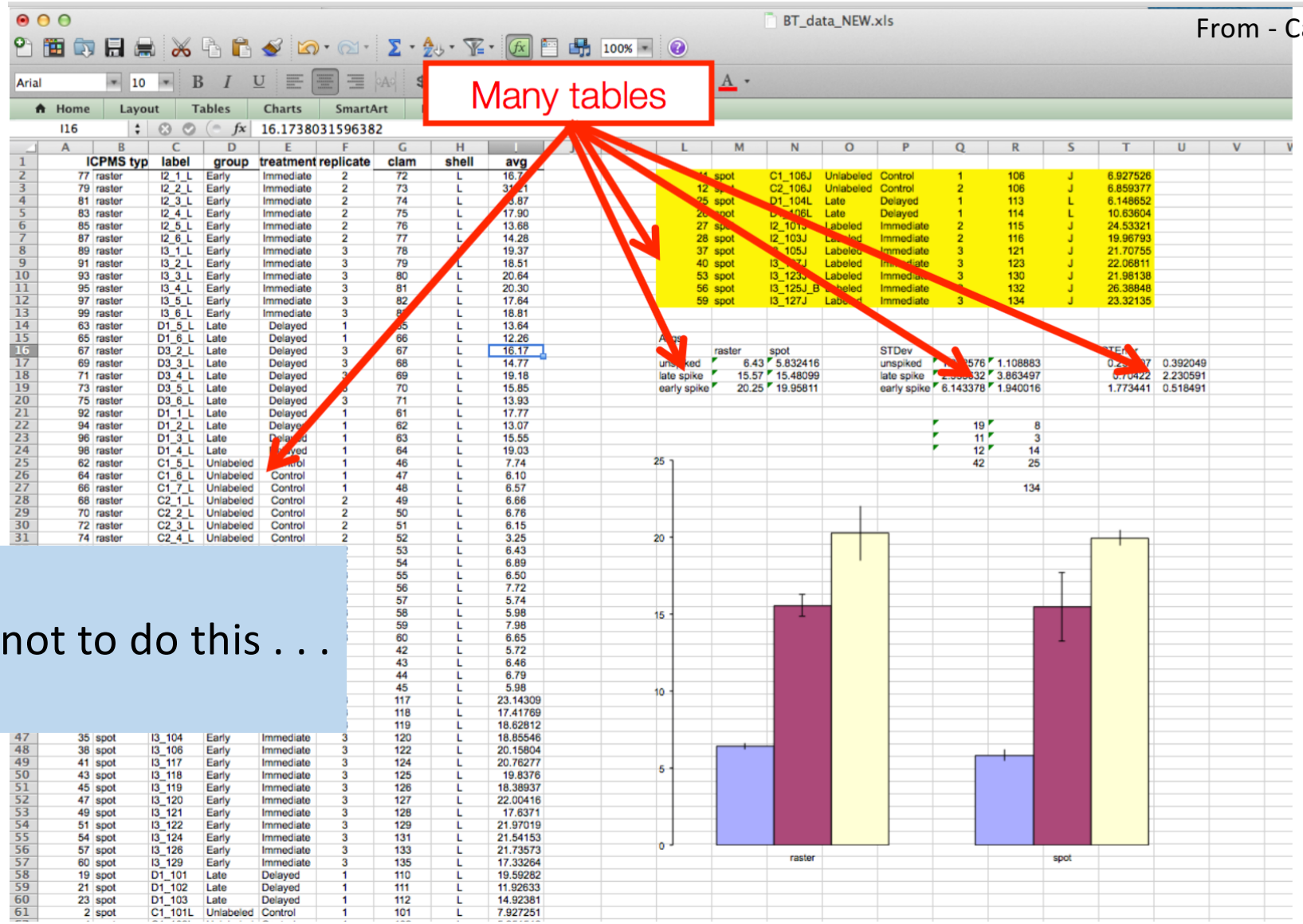
Borer et al. 2009 Bulletin of the Ecological Society of America

Scientists are bad at data management.

From - Carly Strasser

From Flickr by robertpaulyoung

From - Carly Strasser



Results

Differentiating power of trait sets

For the seven-PFT classification, the mean fraction of correctly predicted PFTs ($\overline{fcp}$) for each trait set was plotted against the number of traits included per trait set (Fig. 2). $\overline{fcp}$ increased with an increasing number of traits included and increased to 0.73 (with a set of five traits). From four traits onwards, trait sets did not significantly differ from each other ($P > 0.85$), whereas sets with three and six traits were not significantly differe other either ($P = 0.18$). For sets of six traits, the pre decreased in comparison to sets of five traits. Althou no clear relationship between $\overline{fcp}$ and the number

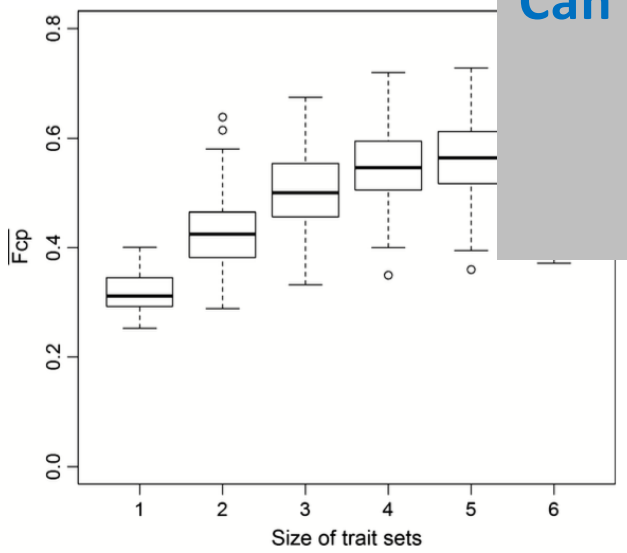


Fig. 2 Mean fraction correctly predicted ($\overline{fcp}$) of all tested trait sets against size of tested trait sets for the seven-plant functional type (PFT) classification. Box plots show the median (middle line), the 25 and 75% quartiles (hinges), the outer value within the 1.5 × interquartile range (whiskers) and outliers (open circles).

Predictability of PFTs in the best performing trait sets

The 10% trait sets (for 2–6 traits) with the highest $\overline{fcp}$ ('best performing trait sets') were selected for further analyses. For seven PFTs, it comprised trait sets with a $\overline{fcp}$ of 0.62–0.73 (κ range 0.45–0.70): this included one two-trait set, nine three-trait sets, 22 four-trait sets, 15 five-trait sets and two six-trait sets (total of 49). The proportion of correctly predicted species, averaged per PFT (Table 2), reflecting the extent to which a PFT could be dis-

Can you document exactly how you got from data collection to your graph??

fused with any other PFT as well. Better predictability of some PFTs was not necessarily caused by a more detailed PFT definition, as reflected in less variation in trait values (Fig. S1); shrubs were the third best performing PFT (Table 2), but encompassed a wide range of species and associated trait values.

Most selected traits in best performing trait sets

Because the number of sets in which a trait occurred varied, we did not compare the absolute number of occurrences in the best performing trait sets. Instead, we compared occurrences of a trait in these best performing trait sets relative to the occurrence of a trait in the full multitrait dataset (2–6 trait sets). The trait with the highest relative occurrence was specific root length (SRL)

From - Carly Strasser

1

From Flickr by ransomtech

Didn't share the data

Didn't document the data (metadata)

Didn't document provenance/workflow

From - Carly Strasser

2

From Flickr by ransomtech

NO Reproducibility
Transparency
Reuse

From - Carly Strasser

From Flickr by Mark Sardella

Plan before data collection

Six Rules of Thumb for Thinking About Data Collection -> Publication

1. Use a scripted program for analysis.

Python

- Analysis scripts are written records of the various steps involved in processing and analyzing data, and provide a form of analytical “metadata.”
- This scripted approach is in contrast to a “GUI-driven” analysis
- Such GUI-based analyses often seem convenient when working through your data and analysis, but rarely leave a clear accounting of exactly what you have done

From - Carly Strasser

2. Think about your naming conventions for your files!

Design file naming scheme

Planning

Use descriptive file names*

- Unique
- Reflect contents

Bad: Mydata.xls
2001_data.csv
best version.txt

Better: Eaffinis_nanaimo_2010_counts.xls

- Study organism
- Site name
- Year
- What was measured

*Not for everyone

3. Store data in nonproprietary software formats (e.g., comma delimited text file, .csv); proprietary software (e.g., Excel, Access) can become unavailable, whereas text files can always be read.

- If your data files are stored using proprietary software, when this software is no longer available, your data also will disappear

4. Always store an uncorrected data file with all its bumps and warts. Do not make any corrections to this file; make corrections within a scripted language.

- When you make corrections post facto to an original data file you could easily be changing something that you later discover was correct in its original form, or maybe you make a mistake while correcting
- **Using a scripted language, you can re-run analyses as well as transformations and corrections to your data,** by using your original data as input, but saving the changes to a separate data file

5. Atomize your data! record a single piece of data (unique measurement) only once; separate information collected at different scales into different tables.

- In other words, create a relational database.

Example of effective data management for the long term.

Site-info-VegBiodiv.csv

Site	Latitude	Longitude	AvgPrecipitation	DomPlantCommu

Abund-data-VegBiodiv_2007.csv

Date	Site	SpName	Abundance

Species-info-VegBiodiv.csv

SpName	Vert1-Invert0	Endo1-Ecto0	AvgMass	TrophicStatus

- When analyzing your data, you can then merge these separate data files together, using your scripted program to link data tables using these key fields.

- In relational databases, this is called a 'join." This method of joining tables together via keys is the basis for a relational database.

From - Carly Strasser

Atomize

During collection

One piece of information per cell

	A	B	C	D	E	F	G	H
1								
2								
3								
4								
5								
6	Address			Number	Street	City	State	Zip
7	415 20th St, Oakland CA 94612		VS	415	20th St	Oakland	CA	94612
8								
9								
10								

From - Carly Strasser

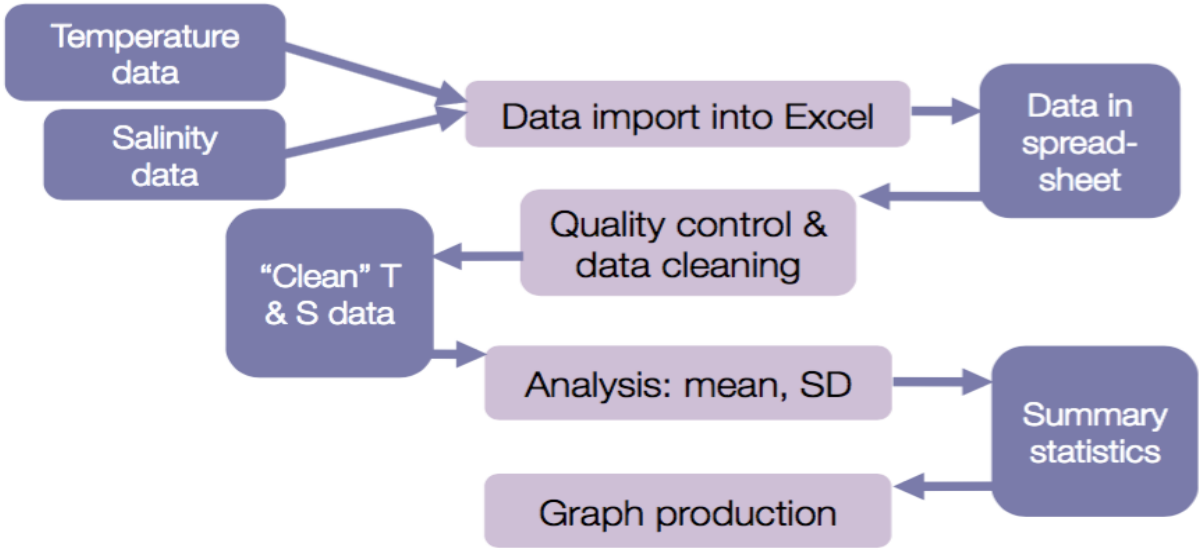
6. All science is a workflow – but we rarely document it and rarely can we reconstruct all parts of it!

Document your workflow

During collection

Workflow: how you get from the raw data to the final products of your research

Simple workflow: flow chart



From - Carly Strasser

Document your workflow

During collection

Workflow: how you get from the raw data to the final products of your research

Simple workflow: commented script

- R, SAS, MATLAB…
- Well-documented code is
  - Easier to review
  - Easier to share
  - Easier to use for repeat analysis

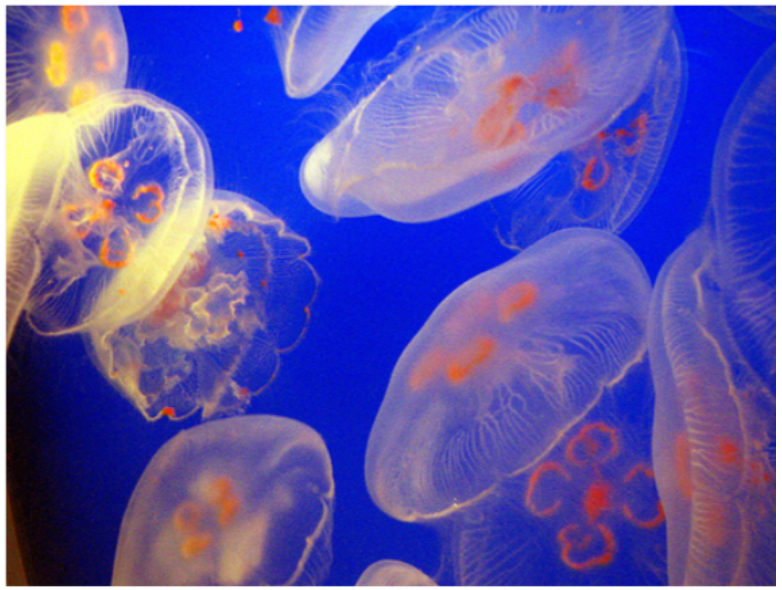
From - Carly Strasser

Document your workflow

During collection

Workflows enable

- Reproducibility
- Transparency
- Reuse



From Flickr by merlinprincesse

What are the most important practices that researchers could implement to make their data sets ready to share with other researchers?

Value of Data Sharing

Data sharing IS a good thing:

- Promotes open science and initiation of new research
- Promotes data longevity
- Promotes data reusability
- Greater exposure of your data to other potential projects & collaborators
- Increased citation of source papers (Piwowar, 2007)
- Confirmation of results from publications (Thornton et al., 2005)
- Generation of value added products
- Possibility for future research collaborations
- More value for the sponsor's research investment

Best Practices for Preparing Environmental Data Sets to Share and Archive[1]

Les A. Hook, Suresh K. Santhana Vannan, Tammy W. Beaty, Robert B. Cook, and Bruce E. Wilson
September 2010

The Seven Best Practices for Preparing Data Sets to Share:

1. Define the Contents of Your Data Files
    1. Units, parameter names, formats
2. Use Consistent Data Organization
3. Use Consistent File Structure & Stable File Formats For Tabular and Image Data
4. Assign Descriptive File Names
5. Perform Basic Quality Assurance
6. Assign Descriptive Data Set Titles
7. Provide Documentation

Best Practices for Preparing Environmental Data Sets to Share and Archive[1]

Les A. Hook, Suresh K. Santhana Vannan, Tammy W. Beaty, Robert B. Cook, and Bruce E. Wilson
September 2010

Four Best Practices for Organization and Documentation

• • • •

Every new data project should include

1. README.txt
2. Data Dictionaries
3. Metadata
4. The project data files (or database) used in all analyses

Readme.txt file(s)

What do include? Top-level project information

- Project Name
- Project Summary
- Any previous work on the project? and where its located
- Funding information
- Primary contact(s) information
- Your name and title (if you aren't the primary contact)
- Other people working on the project
- Location of data and supporting info
- Organization and naming conventions used for the data

Data Dictionaries

What do include? Top-level project information

- Units
- Format
- Min and max values
- Variable names
- Variable definitions
- How its measured

From - Carly Strasser

Metadata

Create metadata

During collection

Metadata: data reporting

WHO created the data?

WHAT is the content
of the data set?

WHEN was it created?

WHERE was it collected?

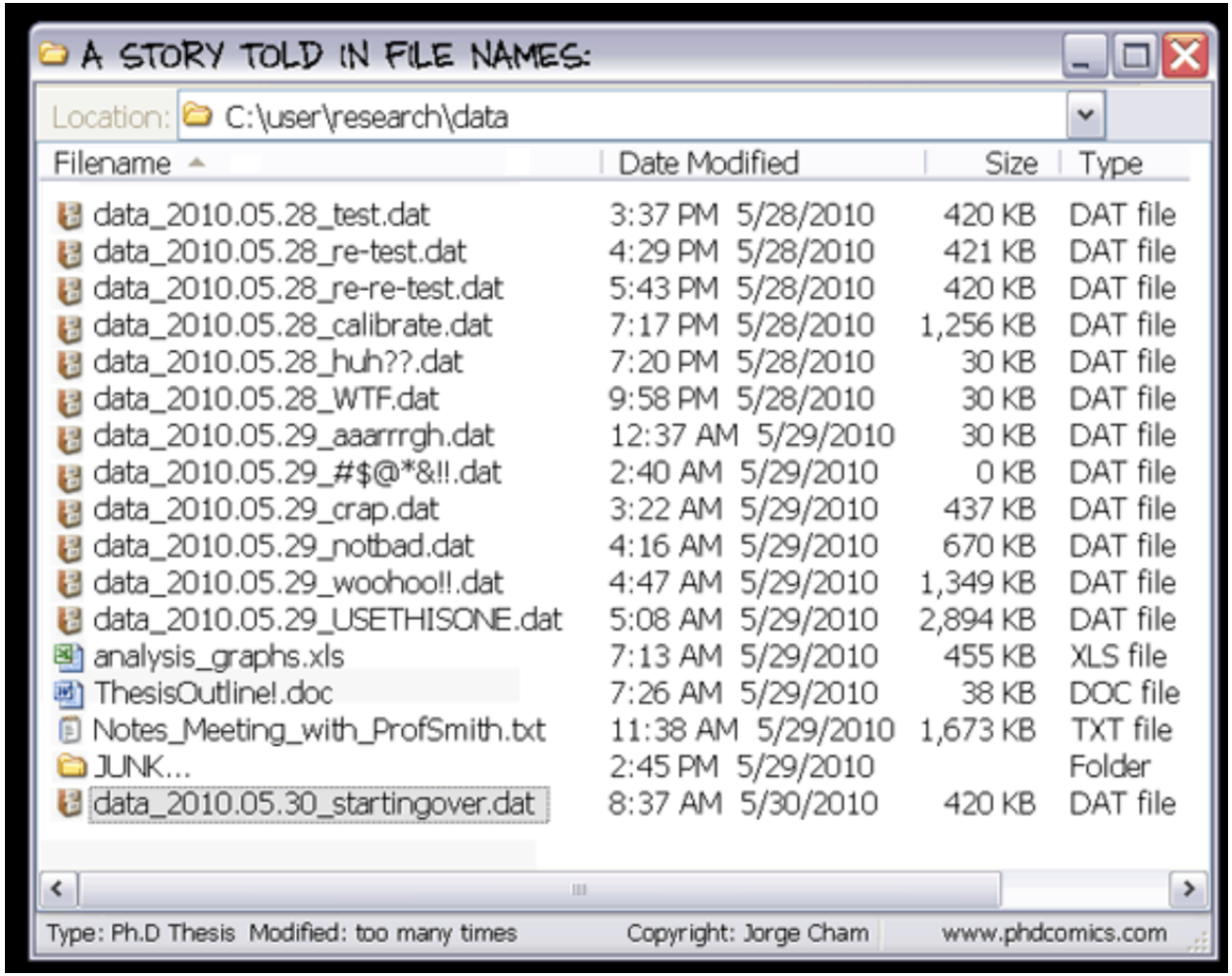
HOW was it developed?

WHY was it developed?

From Flickr by /Michael Patrick

Best practices for scientific computing

1. Use version control
2. Test your code
3. Automate repetition



From Ethan White

1. Use version control

Track changes on steroids.
Tracks every change every made. Stores the full state of the code.
Like a lab notebook for computing.

Use Github!

Never. Lose. Anything.
Easily unbreak your code. Experiment with impunity. Essential for collaborative coding.

http://github.com

From Ethan White

2. Test your code

Test individual pieces of code.
Does my code still do the same thing. Automate testing.

From Ethan White

3. Automate repetition

Let the computer do the work. Make fewer mistakes.

From Ethan White

Learn More

Borer, Elizabeth T., Eric W. Seabloom, Matthew B. Jones, and Mark Schildhauer. 2009. Some Simple Guidelines for Effective Data Management. Bulletin of the Ecological Society of America. 90:205—214. [doi:10.1890/0012-9623-90.2.205]

Barton, C., R. Smith and R. Weaver . 2010. Data Practices, Policy, and Rewards in the Information Era Demand a New Paradigm. Data Science Journal. IGY95-IGY99.

Michener, W K. 2006. Meta-information concepts for ecological data management. Ecological Informatics. 1:3-7.

Hook et al. Best Practices for Preparing Environmental Data Sets to Share and Archive https://www.dataone.org/best-practices

Streasser et al. DataOne - Primer on Data Management: What you always wanted to know*

Read anything from Ethan White on best practices for data and computation in Ecology – see his FigShare account

Learn More

Train scientists to:

Produce science faster

That is more correct

And more reproducible

http://software-carpentry.org, @swcarpentry

Best practices for scientific computing

Greg Wilson (@gvwilson), D. A. Aruliah, **C. Titus Brown (@ctitusbrown),** Neil P. Chue Hong, Matt Davis, Richard T. Guy, Steven H. D. Haddock, Katy Huff, Ian M. Mitchell, Mark Plumbley, Ben Waugh, Ethan P. White, Paul Wilson

http://arxiv.org/abs/1210.0530

From Ethan White