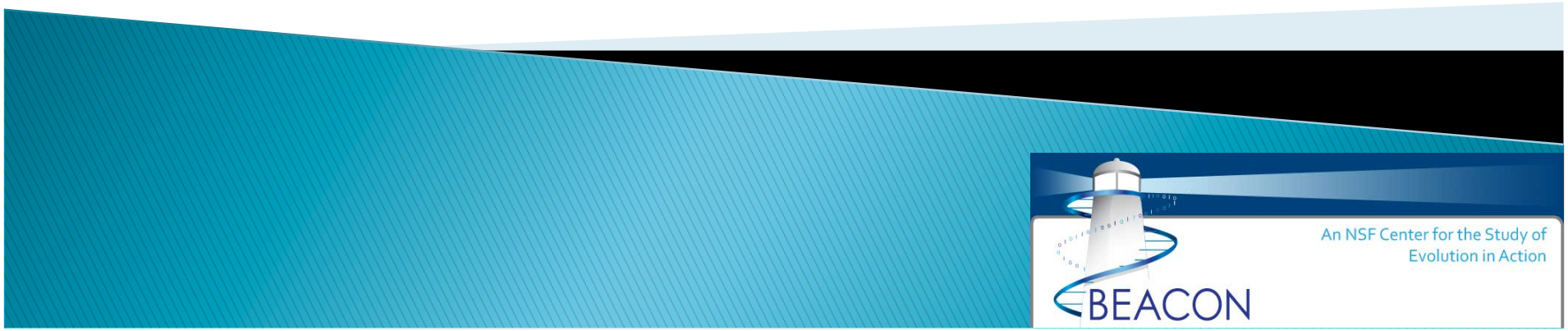
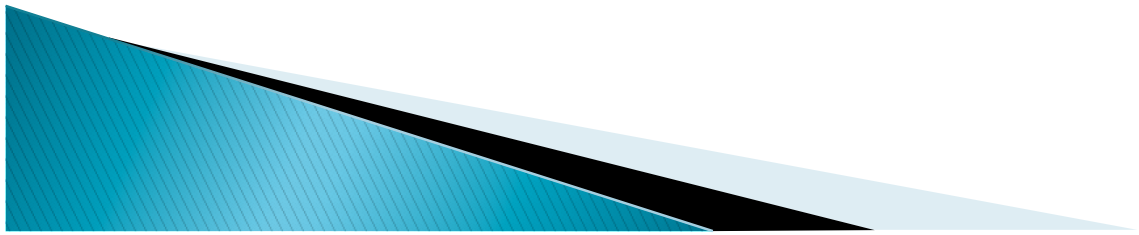
Data archiving, management and reproducible research for ecology and evolution.

Thursday August 15th 2013

Ian Dworkin

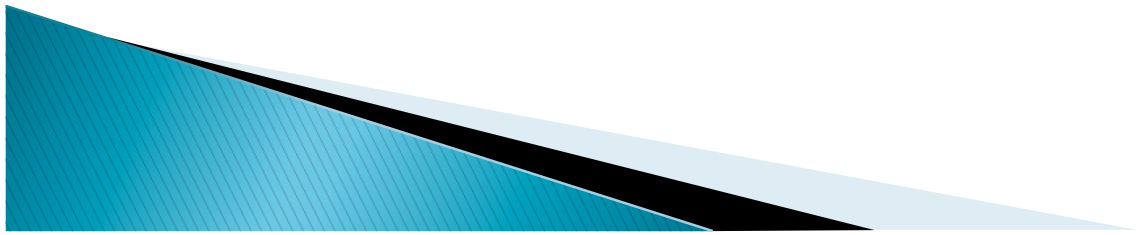


How do we make science a more open and reproducible endeavor?



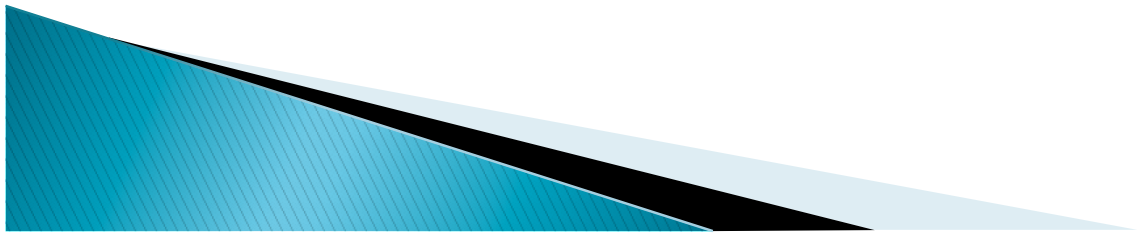
How do we make science a more open and reproducible endeavor?

Is publishing your paper enough?



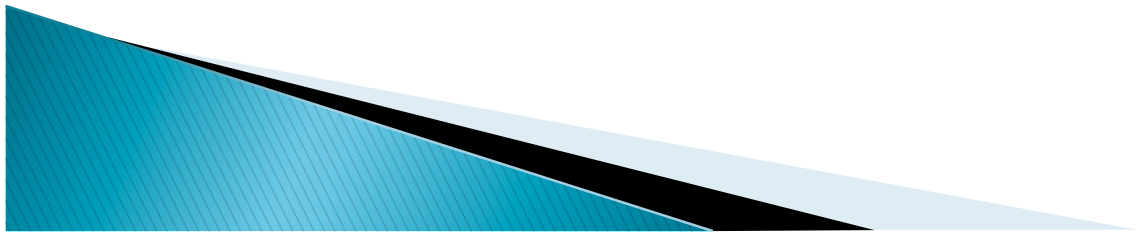
How do we make science a more open and reproducible endeavor?

Data? Meta-data? Scripts?

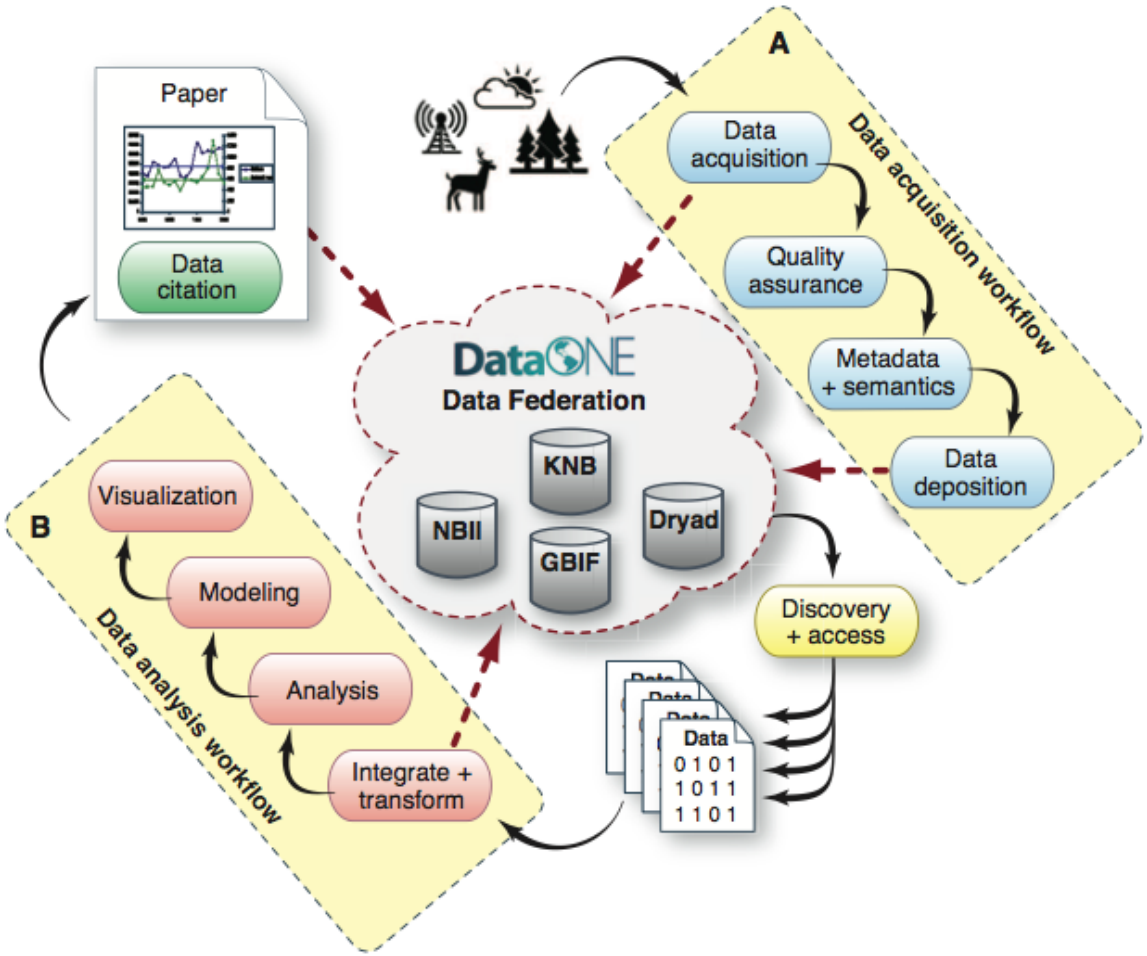


You want to make this process as easy as possible for yourself

- In the process of organizing yourself to share your data and analysis pipelines, you will make yourself a better scientist.



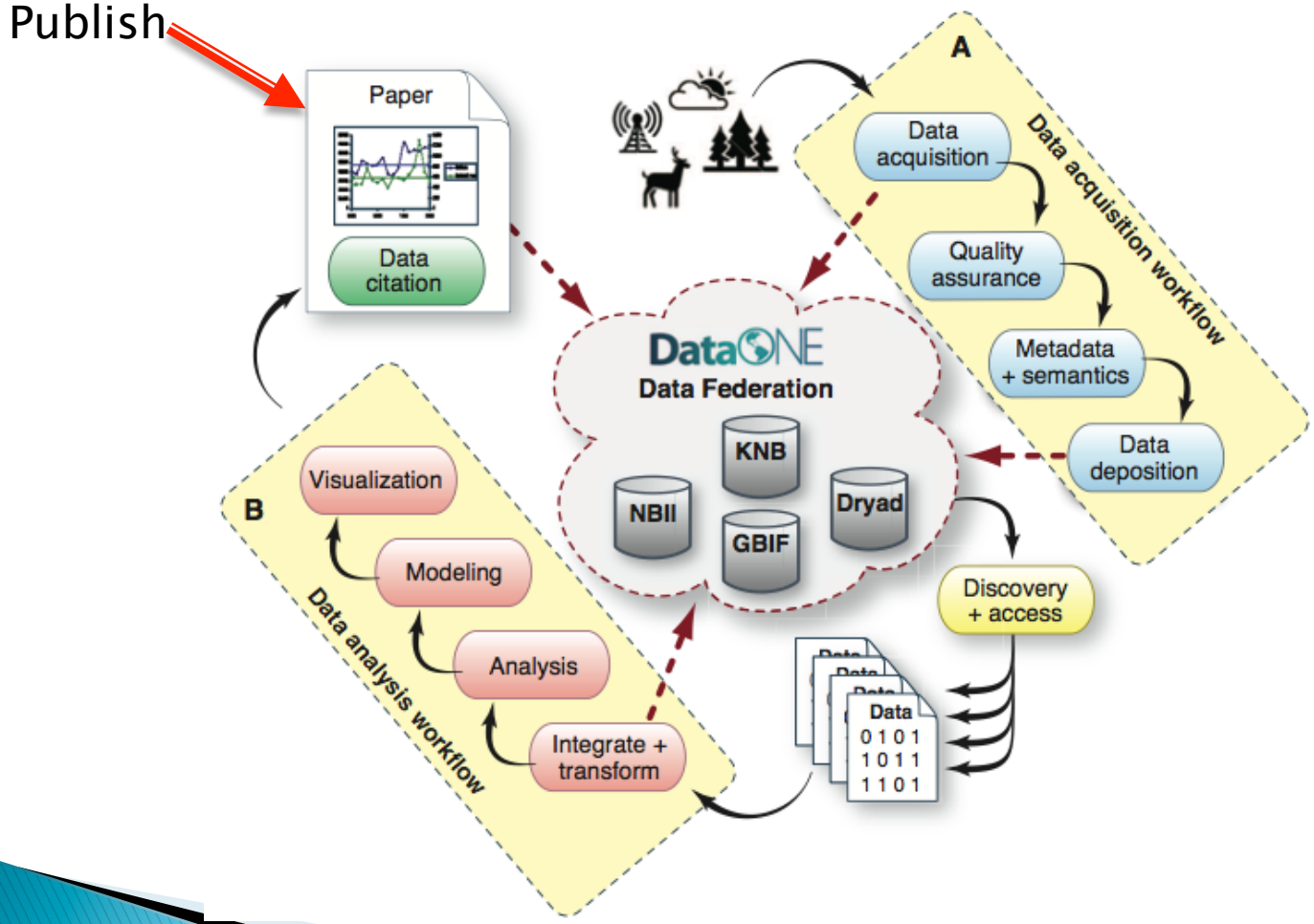
How do we make science a more open and reproducible endevour?



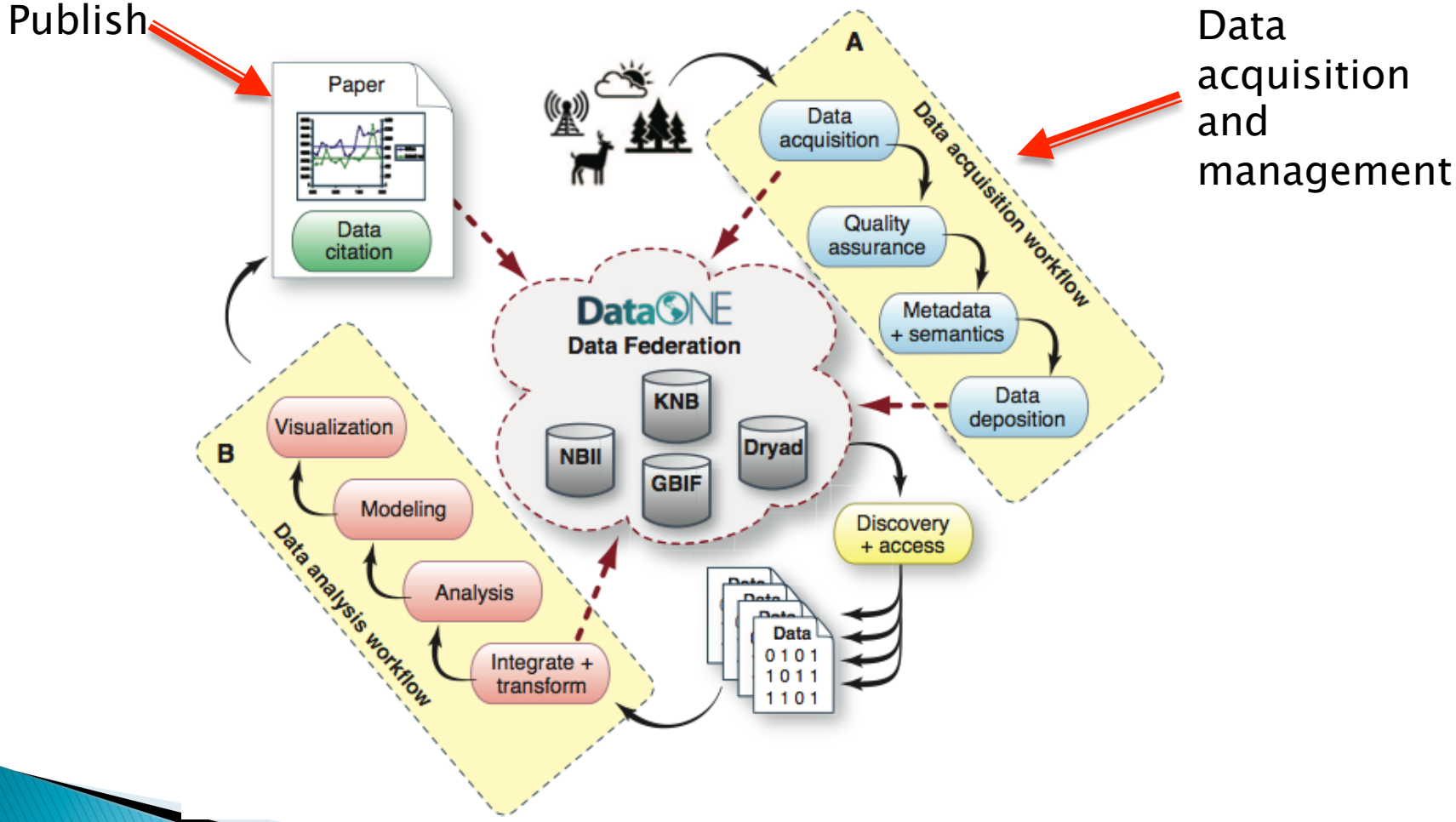
Reichman et al 2011 Science, 331:703-705

http://www.sciencemag.org/content/331/6018/703.full.html

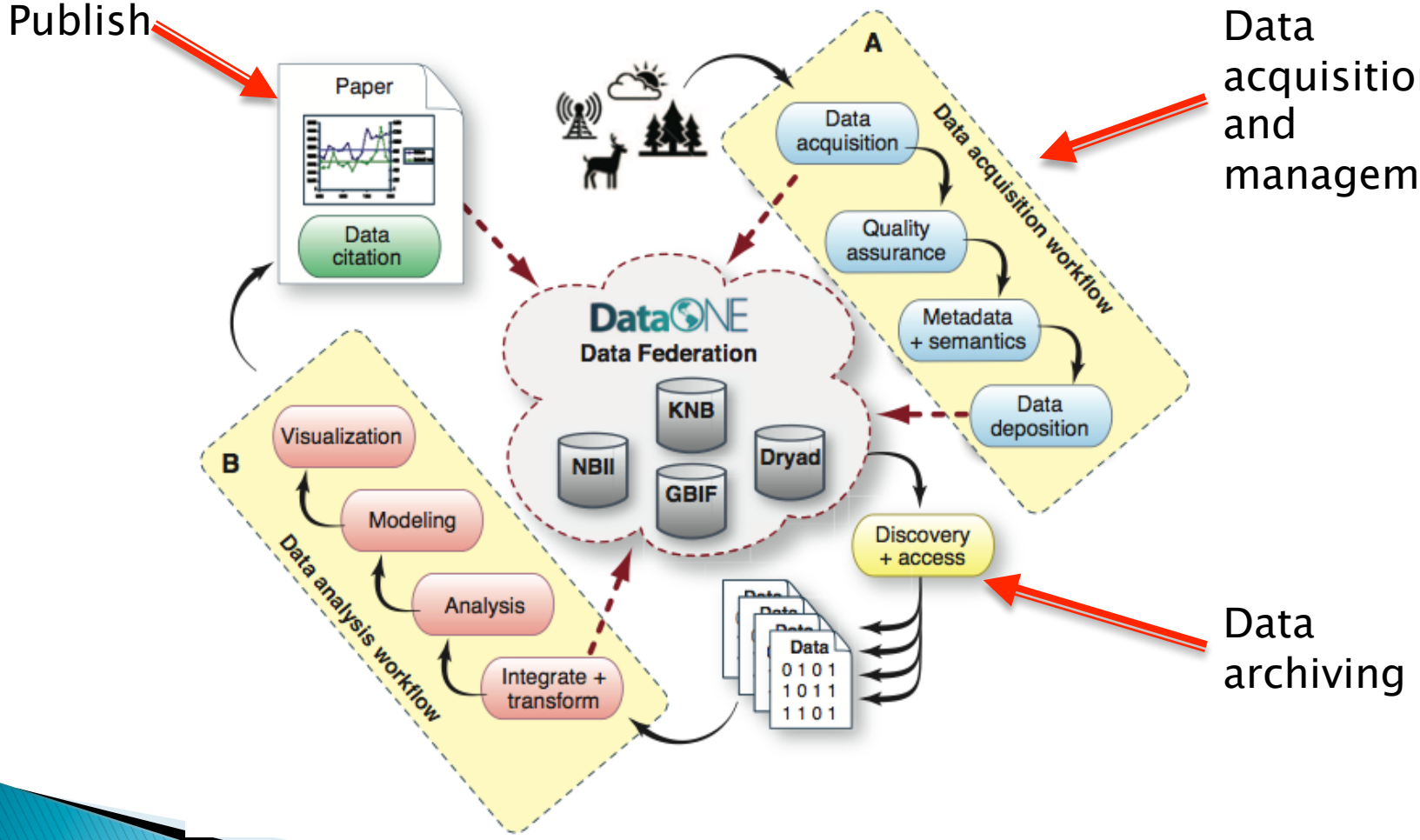
How do we make science a more open and reproducible endevour?



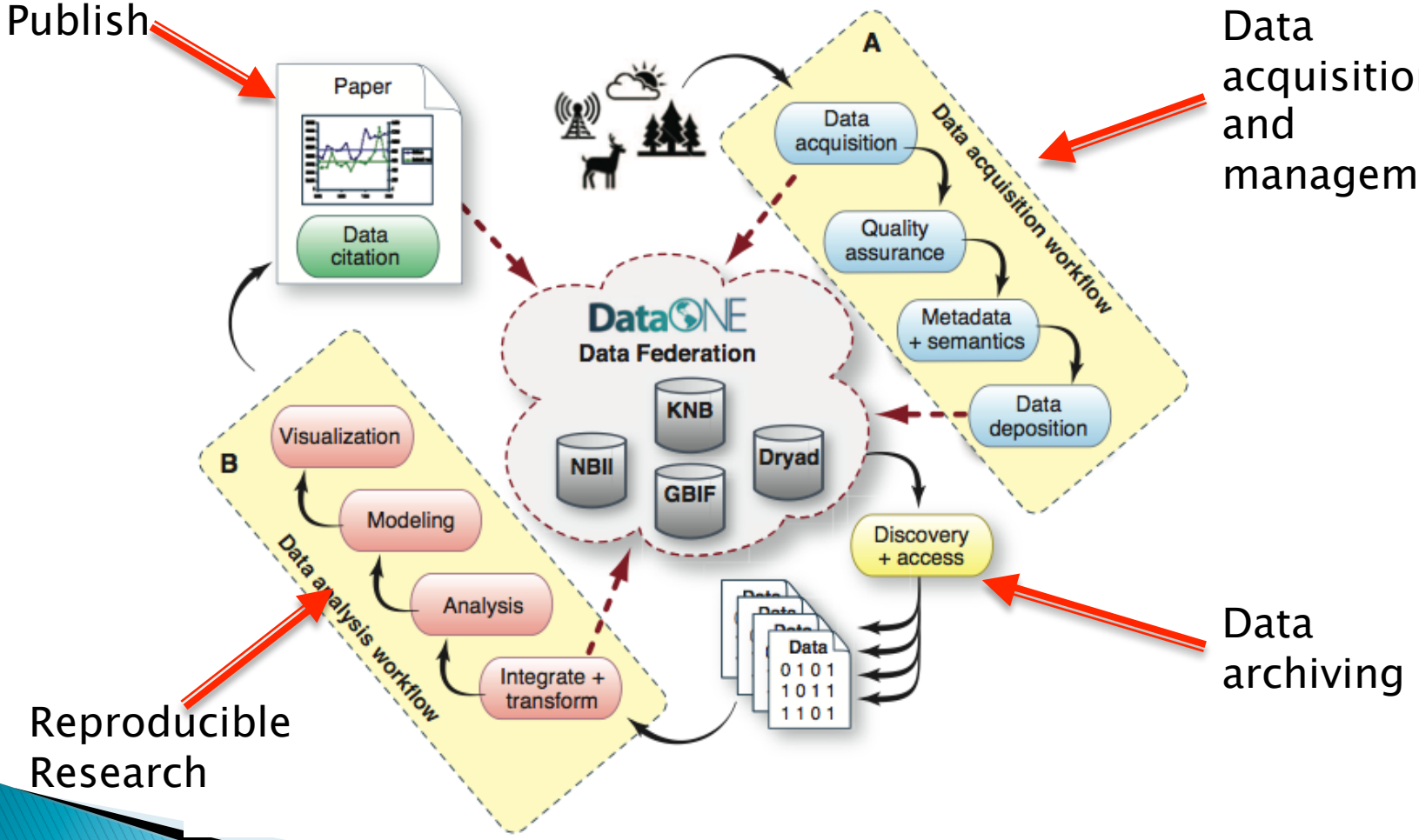
How do we make science a more open and reproducible endevour?



How do we make science a more open and reproducible endevour?



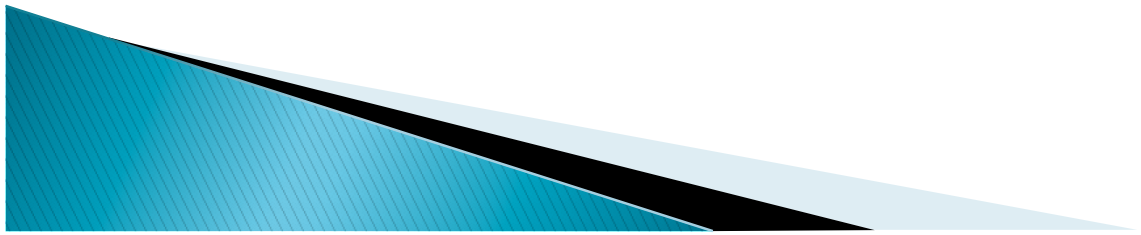
How do we make science a more open and reproducible endevour?



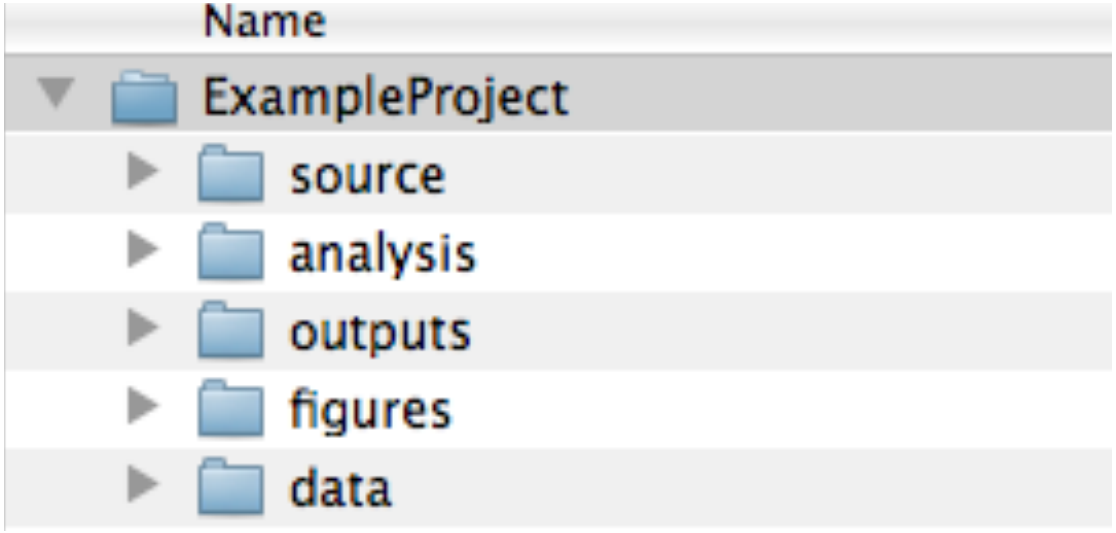
So how do you make this easy on yourself

- At the onset of your research, organize your workflow with the goal of making it as easy as possible to share.

- With a few simple “rules” you will save yourself an enormous amount of time while making your data available.



Rule number 1:Organize your project in a simple and clear manner.

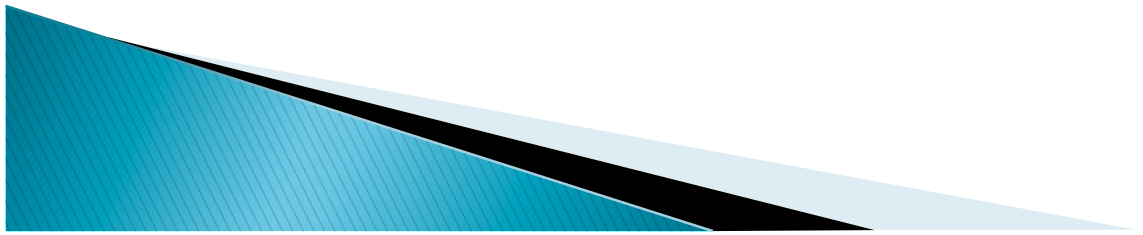


- http://nicercode.github.io/blog/2013-04-05-projects/
- http://nicercode.github.io/blog/2013-05-17-organising-my-project/

Rule number 2: Always strive to store raw data as plain text files

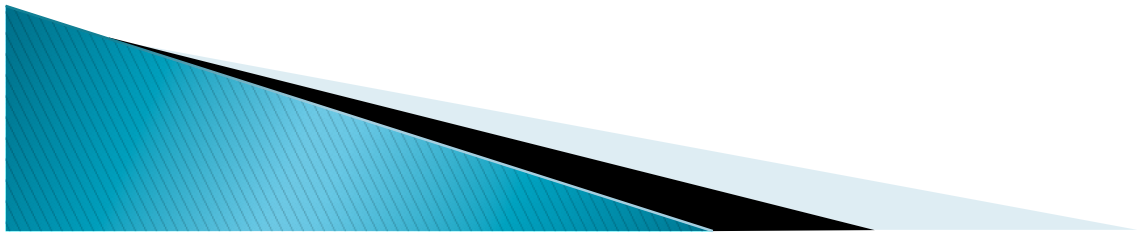
- File formats for many commercial programs (SAS, JMP, Excel,) can not be easily read by other programs.

- Worse still. As new versions of software emerge, no guarantee of being able to open older file formats (Excel).



Rule number 3: version control analysis (and if possible, the data).

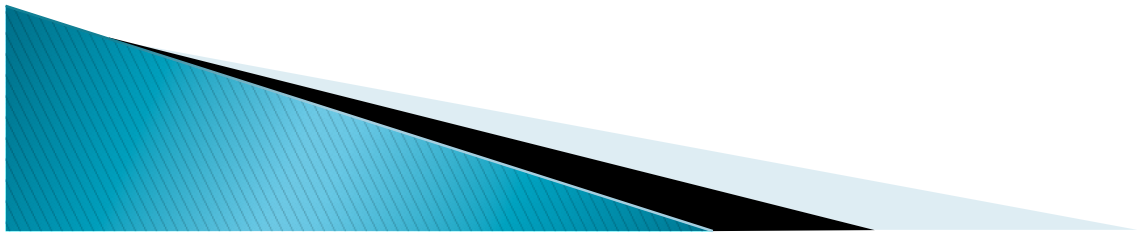
- Best not to alter your raw data in the original files. Construct scripts to do so.
- Learn to use version control (for R- Rstudio makes this really easy with git).
- Helps reduce the possibility of generating dozens of analysis scripts.



But what if I use a GUI based analysis software (i.e. JMP)

- Most statistical software packages with GUIs allow you to generate the underlying script for your analysis as a textfile.

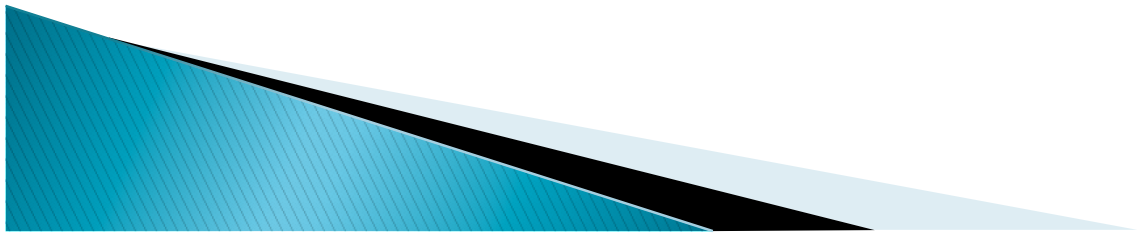
- This can be used to repeat your analysis and to share.



With version control your analysis pipeline gets its own "lab notebook"

- Using version control (like git) will allow you to always add comments about what changes were made to your files, which serves as a reminder of both when and why you changed them.

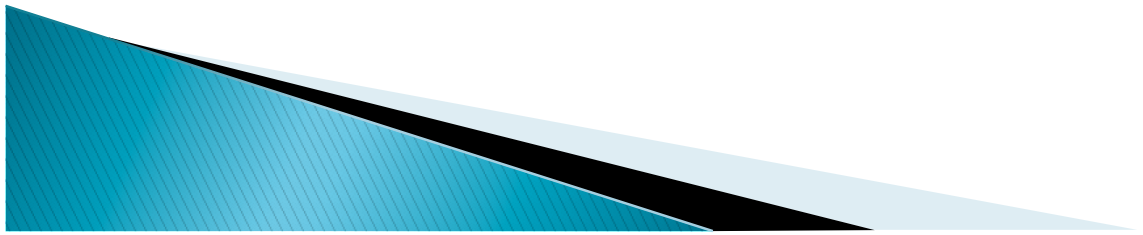
- It is your lab notebook for the analysis.



Rule number 4: Write a readme file right at the beginning of your analysis

- Don't assume that any collaborators or future users (or future you) will know what variables are in your data set.

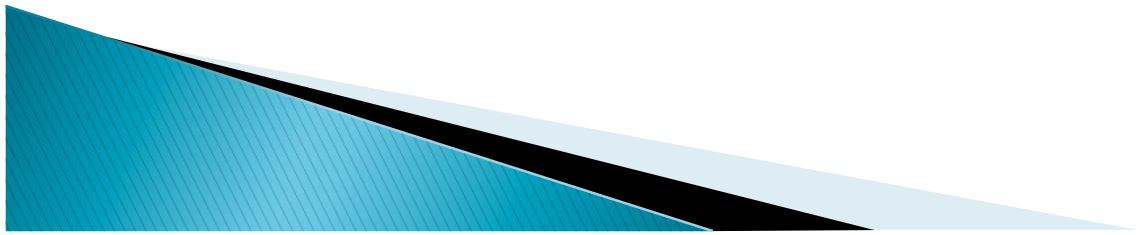
- Write a short readme file containing some basic meta-data for the project including explanations of the variables, how and where they were collected (and why).



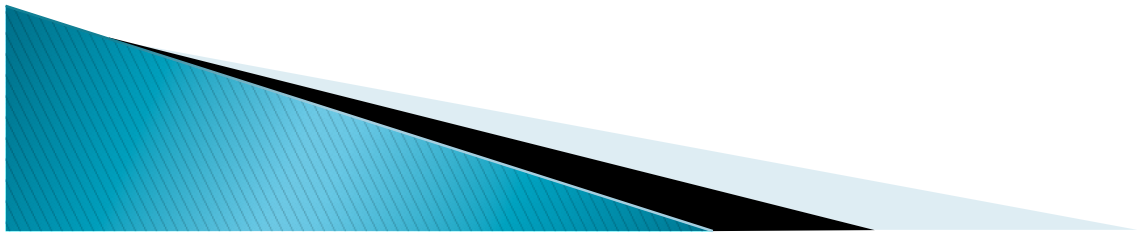
Sharing data

- In addition backing up data for yourself, getting it out there as soon as possible (in conjunction with manuscript submission or acceptance) is crucial.

- For Evolution, DRYAD does a great job of curating your data, and they tweet each new submissions (which seems to increase the readership of your papers).



Slides for previous presentations.

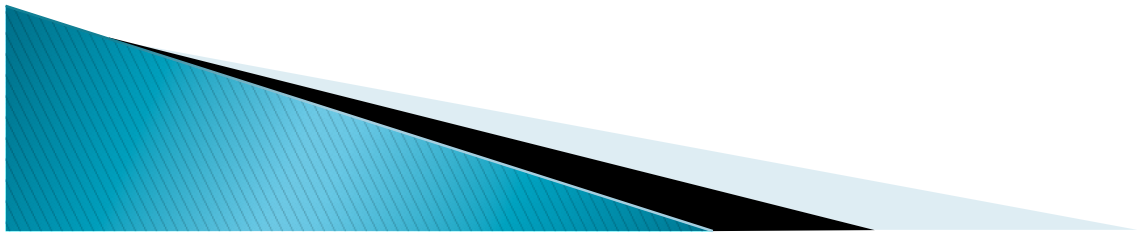


A Primer on Data management

- It is also essential to check out DataONE
- They have a tool for generating your NSF Data management plan.
- Links to software to manage data.
- Best Practices for managing data, including tutorials.

Simple Guidelines (adapted from Borer et al. 2009) for data management and archiving.

For a more detailed overview please see the best practices section at DataONE, as well as the tutorials.



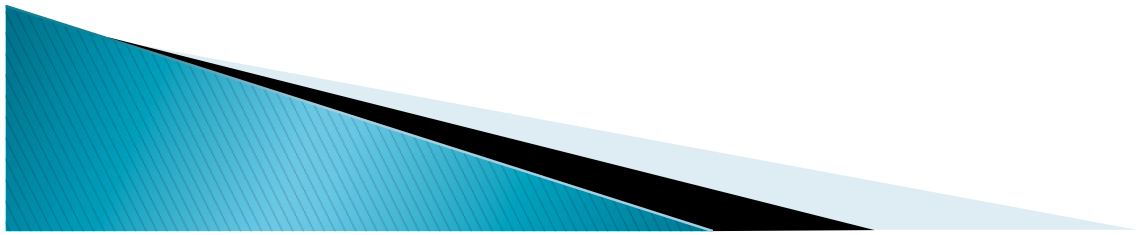
Simple Guidelines for data management/archiving:Think about your data and meta-data even before you start collecting it!

- It is best to start considering how to share the meta-data associated with the data (variable names, organisms, when it was collected, anything else needed to replicate results not included in the paper) **while doing the analysis and manuscript preparation.**
- Consider using a controlled vocabulary as well.

When should I share my data? What data?

- The general advice (and requirements) is to share the data used in a manuscript to coincide with the publication of that manuscript.

- Databases such as DRYAD allow for an embargo for up to a year, to provide you (the author) more opportunity in using that data for additional research, without concern of competition.

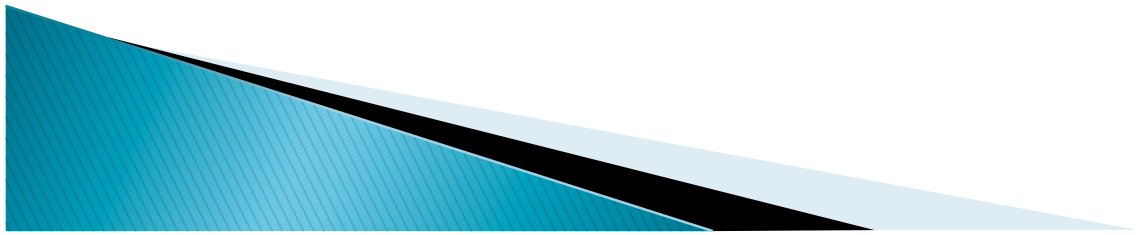


What data should I share.

- The general requirement is to share all of the raw data (warts and all) required to reproduce the analyses that were performed in your paper.
- There is no requirement to share unpublished data (even if it is related).
- However several databases allow for the publication of such data (at the authors discretion) in particular for long term data sets.
- It is advisable to share your analysis script/ pipeline especially if you remove outliers, perform transformations etc…

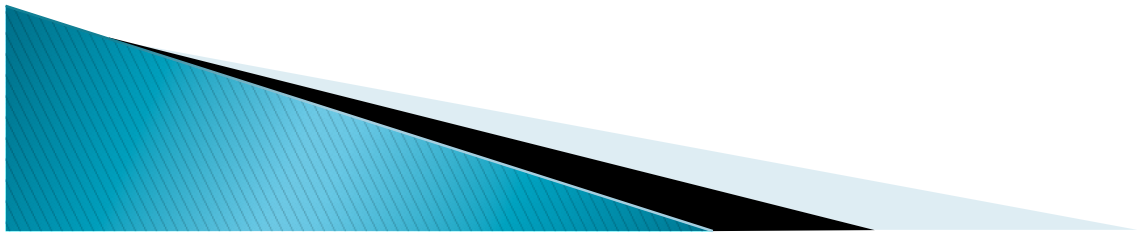
Outline of the questions for the workshop

1. Why should I share my data?
2. When should I share my data?
3. **When should I not share my data?**
4. How should I share my data?
5. Where should I share my data?
6. Final thoughts on data sharing, and the basics of reproducible research.



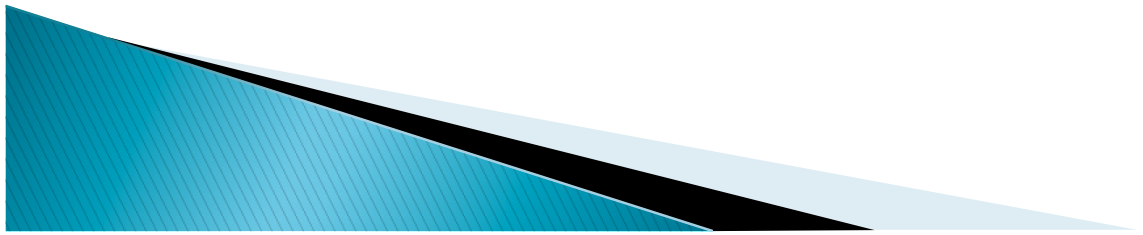
When should I not share my data?

- An often heard concern is ".. What if I did something wrong in the analysis…"
- This is not a valid reason for not sharing your data!!! Indeed this is a great reason to share it, so that your results can be verified, and mistakes corrected.
- Plus this can lead to additional citations!



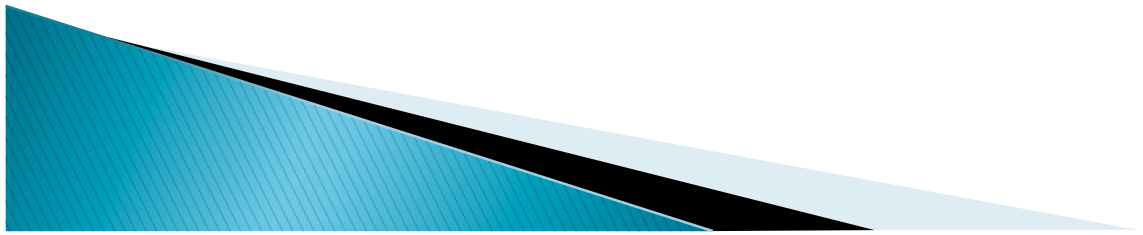
When should I not share my data?

- There are legitimate reasons not to share data in some situations.
- For example, if you have GIS data that may point to organisms that are of particular conservation concern. You may be able to remove the GIS data, and make the rest available.



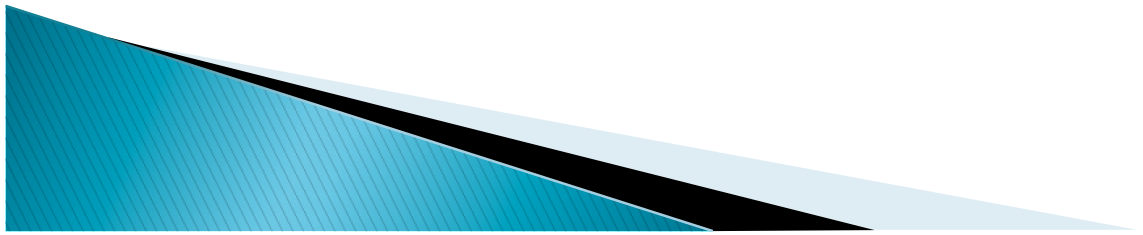
Outline of the questions for the workshop

1. Why should I share my data?
2. When should I share my data?
3. When should I not share my data?
4. How should I share my data?
5. **Where should I share my data?**
6. Final thoughts on data sharing, and the basics of reproducible research.



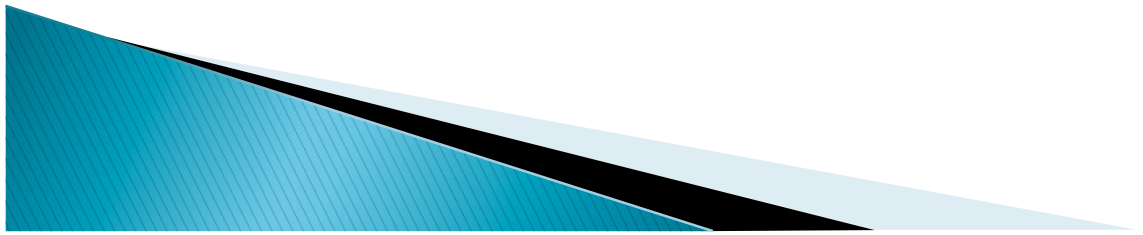
Where should I archive my data?

- There are a few options:
- Your website.
- University website.
- Journal Supplementary information.
- **Public or Community Data repositories.**



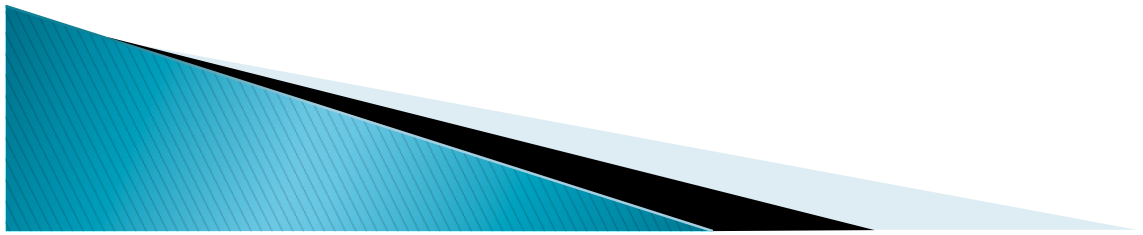
Where should I archive my data?

- Personal websites (even on a University server) are not ideal, as it may be difficult to permanently maintain access to the url, in particular after a researcher departs.
- Supplementary material sections at journals are problematic for TWO reasons.
  - Often there is a **paywall**, so users without subscriptions (personal or institutional) can not access the data.
  - Journals are **NOT** doing an adequate job of maintaining the supplementary materials.
    - See Anderson et al. 2006. BMC Bioinformatics 7:260



Where should I archive my data?

- There are now numerous general (i.e. dataone) and data specific repositories available.
- Most are designed for long term storage, with backups.
- Here are a few examples.



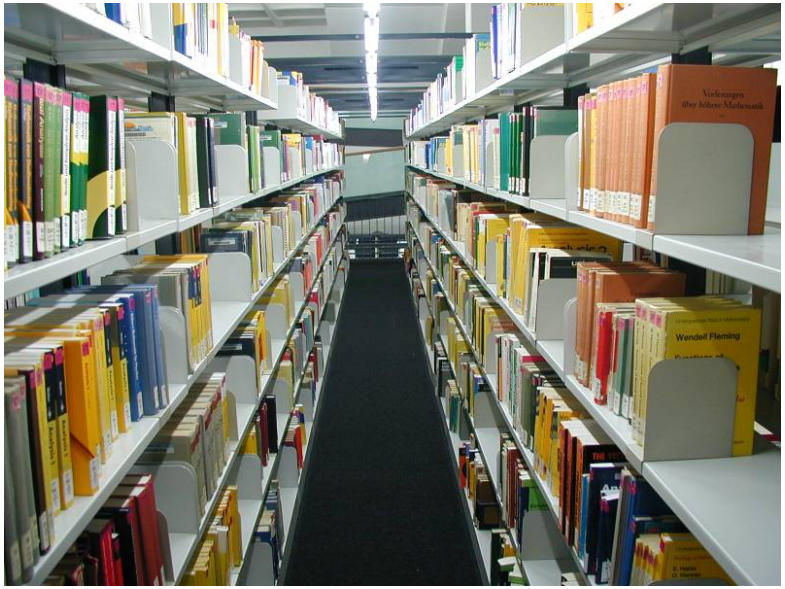
Can this vision be achieved by specialized databases?

- There are a growing number of specialized databases to which deposition is expected (Genbank, Treebase)
  - And others are emerging (Morphbank, PDB, etc)
- A world in which every datatype had its own required database, each with its own submission system
  - Would be a huge burden on authors
  - Would inevitably leave some data orphaned
  - Might never be financially possible

https://www.nescent.org/wg/dryad/images/4/4b/Dryad.bbl.feb08.ppt

What is the alternative?

- A catch-all digital library for data that are
  - Heterogeneous
  - Idiosyncratically structured



https://www.nescent.org/wg/dryad/images/4/4b/Dryad.bbl.feb08.ppt

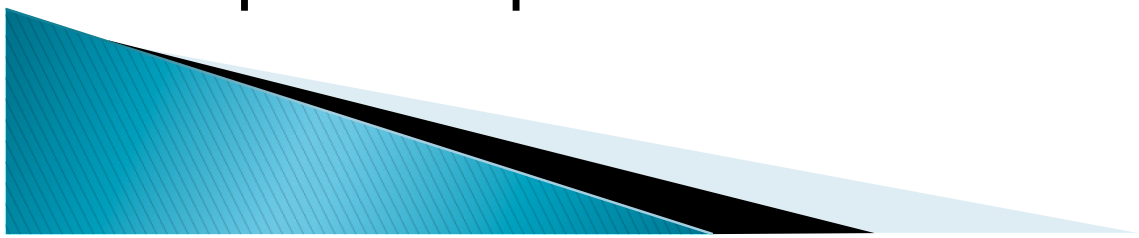
How should I share my data?

Clearly you need to choose an appropriate archive for your discipline and the data type.

Many particular data types have very specific requirements for submission (NCBI GEO).

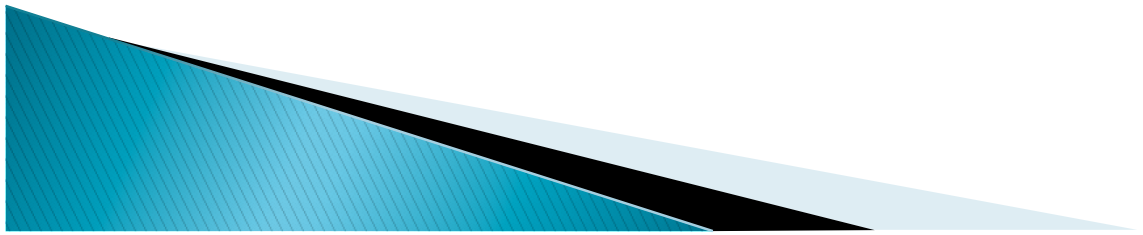
Others such as DRYAD are more flexible, but they are still manually curated before being accepted.

Generally a short clear “readme” file associated with the data will be enough to clear up most specific questions associated with the data.



Outline of the questions for the workshop

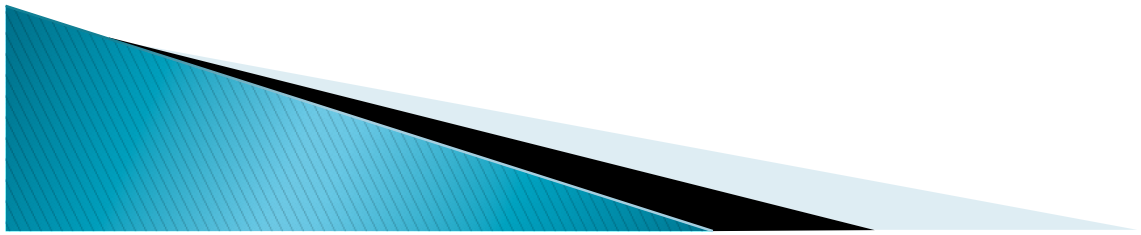
1. Why should I share my data?
2. When should I share my data?
3. When should I not share my data?
4. **How should I share my data?**
5. Where should I share my data?
6. Final thoughts on data sharing, and the basics of reproducible research.



How should I share my data? A Primer on Data management

- Read:
  - Some Simple Guidelines for Effective Data Management, Borer ET, Seabloom EW, Jones MB, Schildhauer M (2009). Bulletin of the Ecological Society of America 90(2), 205–214. doi:10.1890/0012-9623-90.2.205.
  - http://www.esajournals.org/doi/abs/10.1890/0012-9623-90.2.205
  - Whitlock, MC. 2010. Data archiving in ecology and Evolution: best practices. TREE 26(2):61–65

Also See:
http://researchdata.wisc.edu/share-your-data/data-access-2/
http://www.vprgs.msu.edu/files_vprgs/data_interview_HANDOUT.pdf

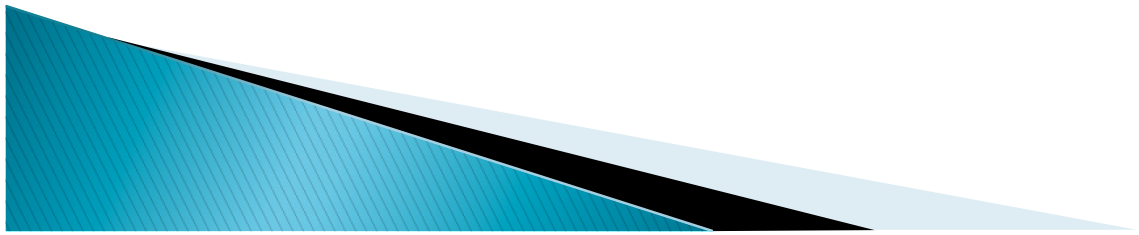


A Primer on Data management

- It is also essential to check out DataONE
- They have a tool for generating your NSF Data management plan.
- Links to software to manage data.
- Best Practices for managing data, including tutorials.

Simple Guidelines (adapted from Borer) for data management and archiving.

For a more detailed overview please see the best practices section at DataONE, as well as the tutorials.

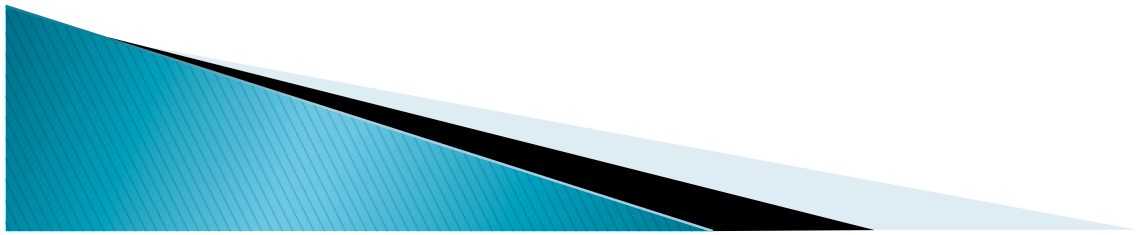


Simple Guidelines for data management/archiving:Think about your data and meta-data even before you start collecting it!

- It is best to start considering how to share the meta-data associated with the data (variable names, organisms, when it was collected, anything else needed to replicate results not included in the paper) **while doing the analysis and manuscript preparation.**
- Consider using a controlled vocabulary as well.

Simple Guidelines for data management/archiving: Think about storage (backups), administration etc..

- You do not want to lose your data. How do you organize, backup and administer the data?
- Backup policy? Data Storage? Data management?



Simple Guidelines for data management/archiving: Provide raw data.

- Make sure that you provide raw data (warts and all).
- If appropriate **also** provide transformed data (i.e. For complex transformations).

- Alternatively (and preferred), provide raw data with instructions/scripts on how to perform the transformations or analysis (See section on Reproducible Research).

Simple Guidelines for data management/archiving: Store data in non-proprietary formats, preferably in flat (ascii) text files.

- Not everyone has access to excel or access.
- Such file formats change (try opening an excel document from 1990).
- very simple “flat” formats such such as tab/ space (.txt), or comma seperated values (.csv).
- preferably in ascii over unicode.

http://repositories.lib.utexas.edu/recommended_file_formats

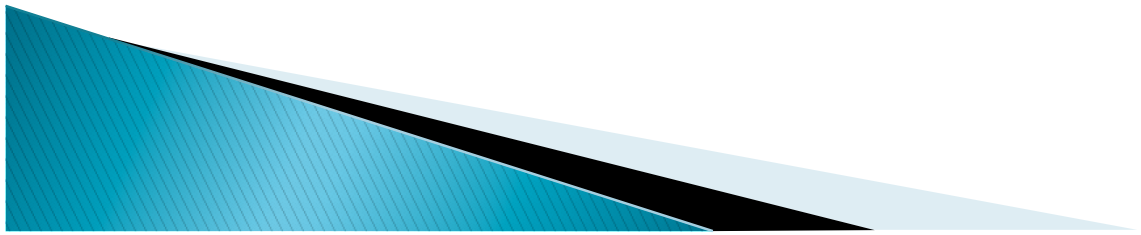
Simple Guidelines for data management/archiving: Use suitable field delimiters.

- Commas (preferred) or tabs
- semi-colons can be problematic for import into some software.

http://repositories.lib.utexas.edu/recommended_file_formats

Simple Guidelines for data management/archiving: Use clear and consistent codes for file names, and variables within the data file.

- Names should be clear enough so other people can understand.
- Consistent (including spelling and case, and variants on abbreviations).

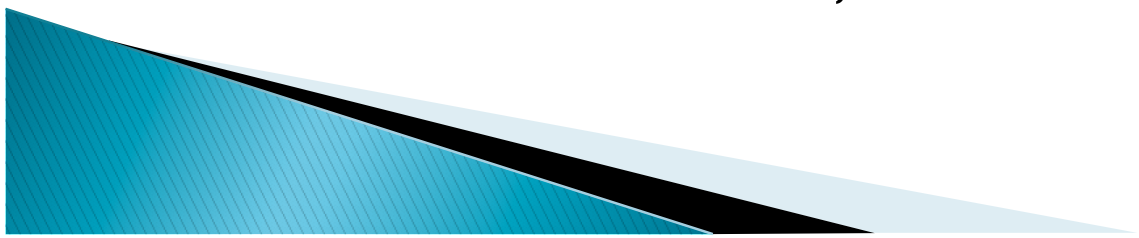


Simple Guidelines for data management/archiving: Do Not change the final data set

- GUI based spreadsheet programs seem useful, but it is very easy to change aspects of the data, without recording your changes in a log.
- If possible make a read-only copy of the data.
- Make all changes in the data with a script. If not possible, log any changes in an associated meta-data file.

Simple Guidelines for data management/archiving: Make a meta-data file.

- Write a meta-data file (describing the variables in the data, where it was collected, when it was collected, by whom, details of the experiment, and changes or outliers or issues about data quality).
- Much more to this, so see link above.

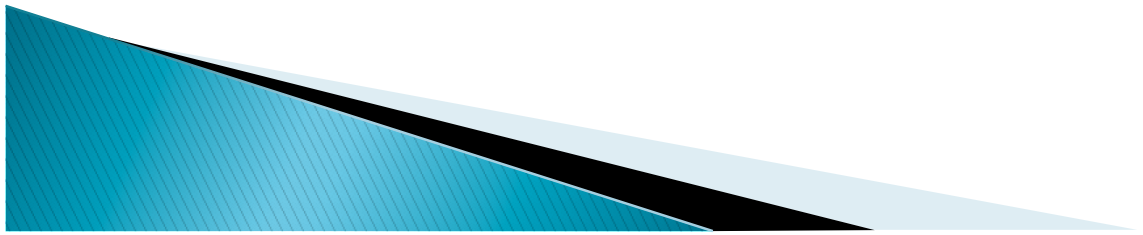


Simple Guidelines for data management/archiving: confirm a match between descriptions of data and the data themselves

- Write a meta-data file (describing the variables in the data, where it was collected, when it was collected, by whom, details of the experiment, and changes or outliers or issues about data quality).
- Have someone who did not write and collect the data examine the data and meta-data to make sure they match!

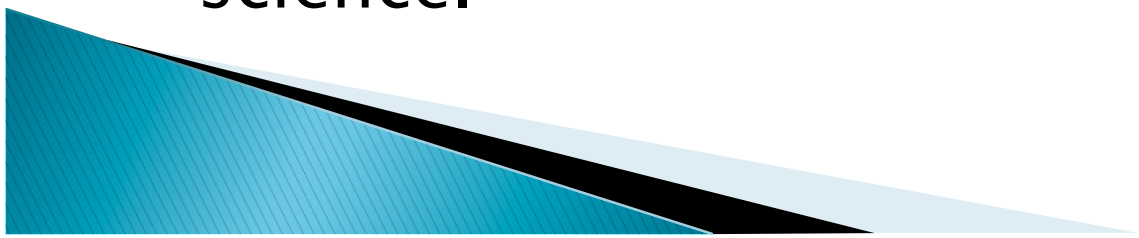
Simple Guidelines for data management/archiving: Archive your data

- Identify suitable repository for the data
- citation and document provenance for your dataset.
- Provide Identifier for the dataset.



Reproducible research

- The data, and meta-data is only one part of the equation. Most analyses have become increasingly complicated and computational.
- Thus having just the raw data may not be enough to reproduce your results.
- Point and click based analysis software, while convenient, hinders reproducibility of science.

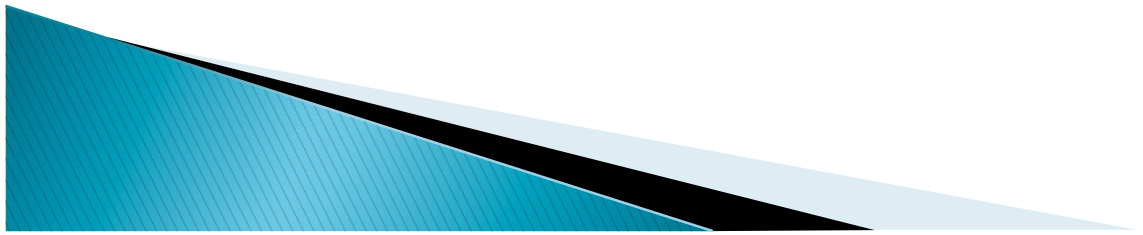


Reproducible research

- If you do perform a complicated analysis, but do not provide the scripts to run them, it is unlikely that others will be able to reproduce what you have done.
- Indeed, it may lead to problems if they assume you did one thing, and you actually did another.
- Thus you should provide the script(s) that can replicate your whole analysis in your published manuscript (including figures).
- This includes removal of outliers, data transformations.

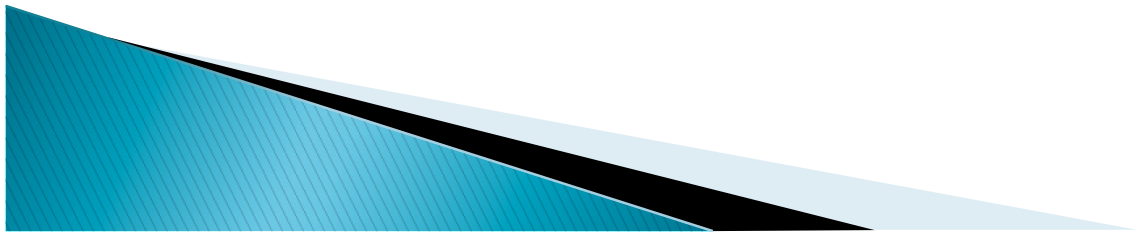
Reproducible research

- Re-run the analysis that you published with the data you are about to submit. Can you replicate your results? If you can not, no one else will be able to either!!!!!!***
- Also provide (either in the script or meta-data) the versions of the software (and libraries) you used. Even the OS can be reported.

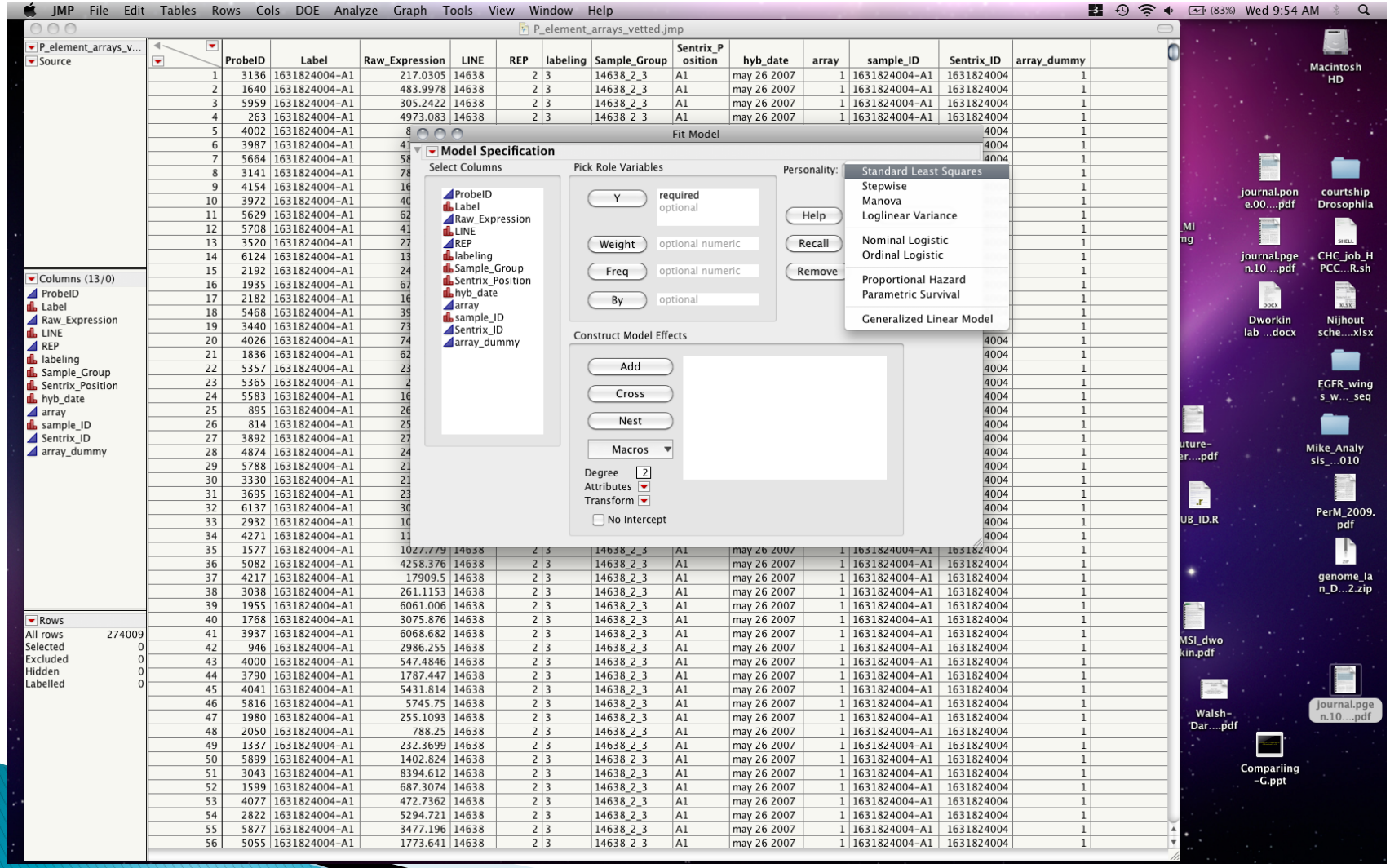


Reproducible research

- The very idea of reproducible research is to provide the "whole environment" of data and computational pipeline with the manuscript itself.
- There are many tools that enable reproducible research.
- Scripting languages like R (Sweave) and python (Pweave) have many tools that aid in this.

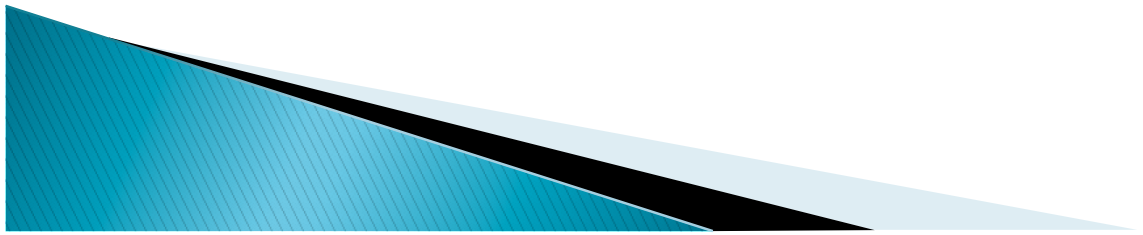


The problem with GUI analysis tools.



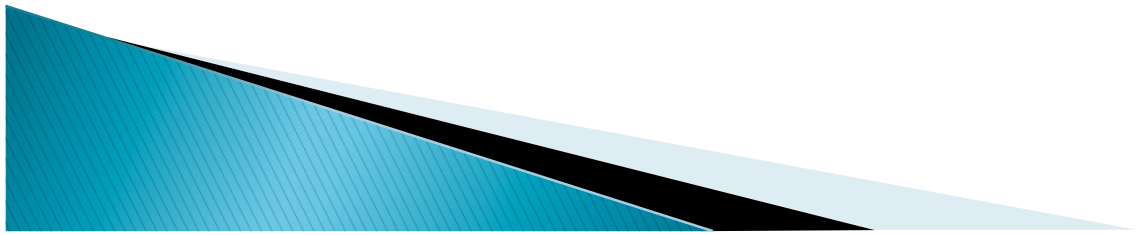
Reproducible research

- If at all possible use open analysis tools (such as R, python, etc..) or even proprietary languages (SAS, matlab, mathematica) but provide the scripts as “metadata” used for the analysis in the paper.
- Remember to run it before submitting it to the database, to make sure you can replicate the results of your paper. Include relevant information of software/library versions with the metadata.

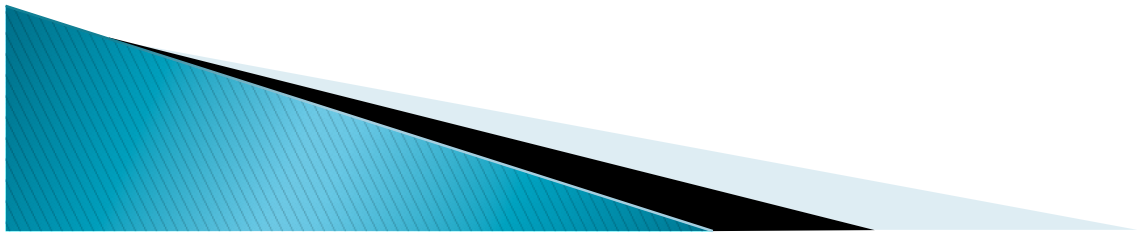


But what if I only have access to GUI based analysis tools for my work.

- Then keep a detailed “analysis” notebook of how you performed the analysis (buttons selected, options clicked, etc…). Provide this with the data.
- The software likely has a log or macro language that can provide the underlying script which can be archived.



Extra stuff (questions)



Unpublished data in Dryad

- Datasets of special historical, educational and scientific significance
  - Particularly long-term data collections that are greater than the sum of their parts.
  - NESCent's Distinguished Visiting Scholar program.
- Prepublished data from NESCent scientists.
- Otherwise, unpublished data is not accepted.

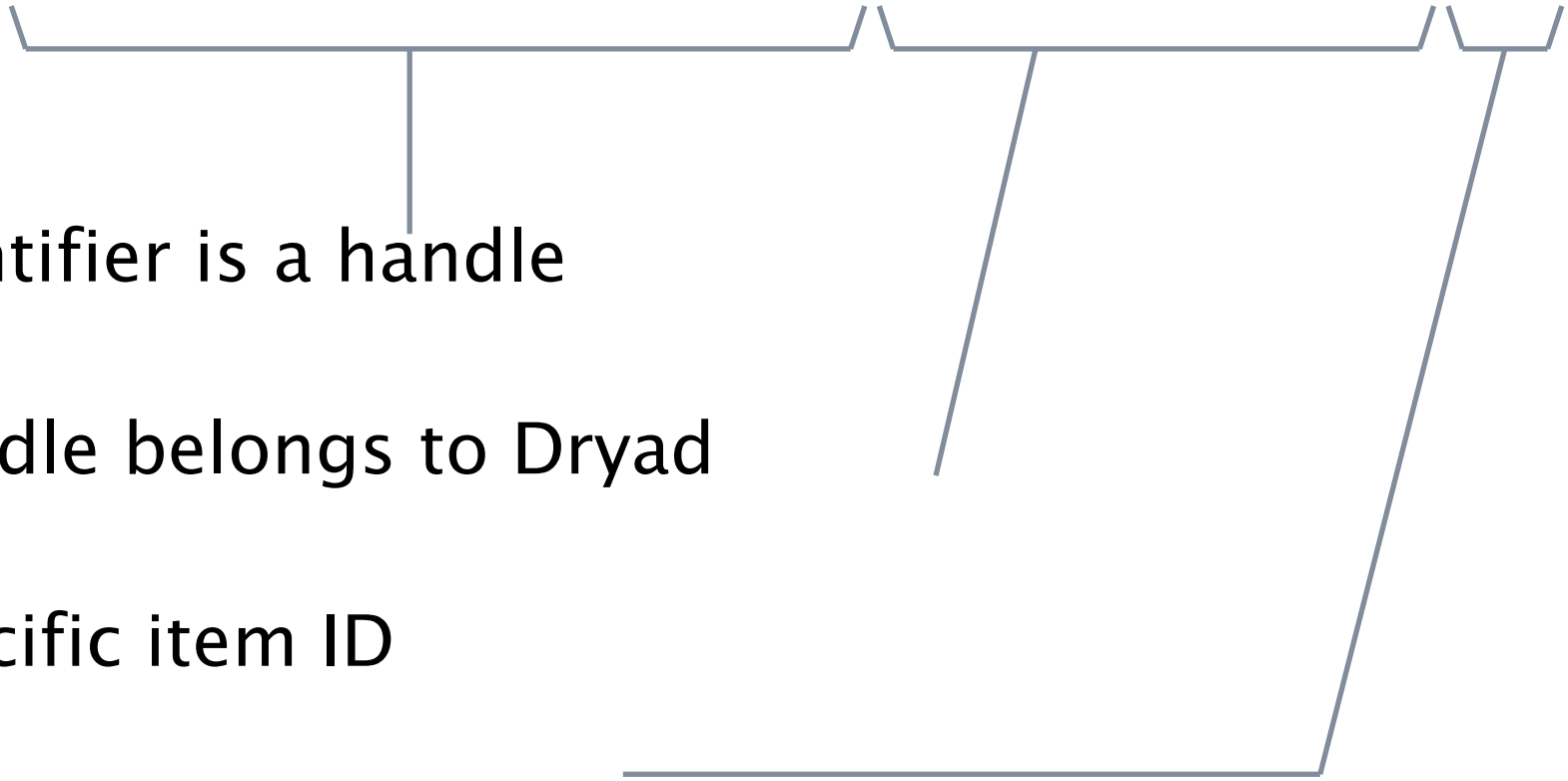
https://www.nescent.org/wg/dryad/images/4/4b/Dryad.bbl.feb08.ppt

Issues of intellectual property

- Some data can be copyrighted (e.g. images)
  - Most data cannot, although law varies by jurisdiction
- Most experts downplay the role of legal agreements in regulating data reuse within a "Science Commons"
- Instead, scientific norms (for reuse, and for attribution) are thought to be more appropriate
  - When immediate data sharing might be truly deleterious, an embargo is an option.

https://www.nescent.org/wg/dryad/images/4/4b/Dryad.bbl.feb08.ppt

http://hdl.handle.net/10255/dryad.23



Identifier is a handle

Handle belongs to Dryad

Specific item ID

https://www.nescent.org/wg/dryad/images/4/4b/Dryad.bbl.feb08.ppt

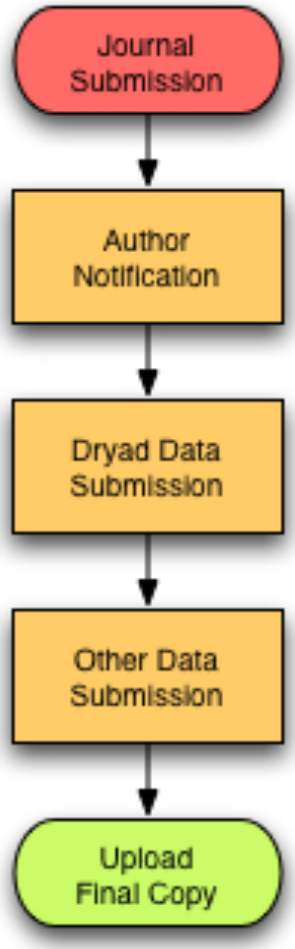
DRYAD

- Paper citation
  - Sidlauskas B (2007) Testing for Unequal Rates of Morphological Diversification in the Absence of a Detailed Phylogeny: A Case Study From Characiform Fishes. *Evolution* 61 (2), 299–316, doi:10.1111/j.1558-5646.2007.00022.x

- Data citation
  - Sidlauskas B (2007) Relative warps. hdl:10255/dryad.23

https://www.nescent.org/wg/dryad/images/4/4b/Dryad.bbl.feb08.ppt

DRYAD: Submission process



https://www.nescen[illegible]rg/wg/dryad/images/4/4b/Dryad.bbl.feb08.ppt

Panton Principles (http://pantonprinciples.org/)

Science is based on building on, reusing and openly criticising the published body of scientific knowledge.
For science to effectively function, and for society to reap the full benefits from scientific endeavours, it is crucial that science data be made open.

By open data in science we mean that it is freely available on the public internet permitting any user to download, copy, analyse, re-process, pass them to software or use them for any other purpose without financial, legal, or technical barriers other than those inseparable from gaining access to the internet itself. To this end data related to published science should be explicitly placed in the public domain.

http://pantonprinciples.org/

Formally, we recommend adopting and acting on the following principles:

1. Where data or collections of data are published it is critical that they be published with a clear and explicit statement of the wishes and expectations of the publishers with respect to re-use and re-purposing of individual data elements, the whole data collection, and subsets of the collection. This statement should be precise, irrevocable, and based on an appropriate and recognized legal statement in the form of a waiver or license. When publishing data make an explicit and robust statement of your wishes.

2. Many widely recognized licenses are not intended for, and are not appropriate for, data or collections of data. A variety of waivers and licenses that are designed for and appropriate for the treatment of data are described here. Creative Commons licenses (apart from CCZero), GFDL, GPL, BSD, etc are NOT appropriate for data and their use is STRONGLY discouraged. Use a recognized waiver or license that is appropriate for data.

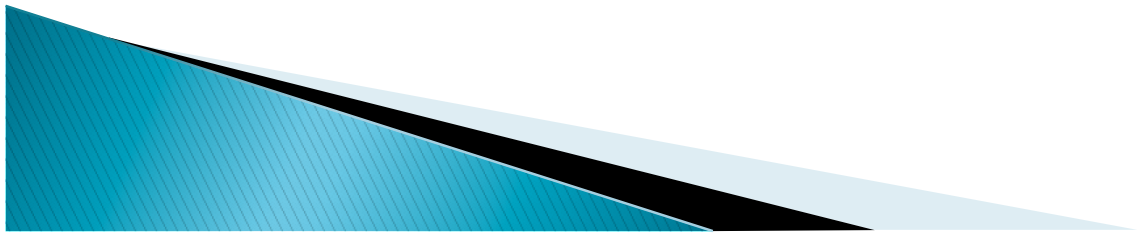
3. The use of licenses which limit commercial re-use or limit the production of derivative works by excluding use for particular purposes or by specific persons or organizations is STRONGLY discouraged. These licenses make it impossible to effectively integrate and re-purpose datasets and prevent commercial activities that could be used to support data preservation.
If you want your data to be effectively used and added to by others it should be open as defined by the Open Knowledge/Data Definition – in particular non-commercial and other restrictive clauses should not be used.

4. Furthermore, in science it is STRONGLY recommended that data, especially where publicly funded, be explicitly placed in the public domain via the use of the Public Domain Dedication and Licence or Creative Commons Zero Waiver. This is in keeping with the public funding of much scientific research and the general ethos of sharing and re-use within the scientific community.

Explicit dedication of data underlying published science into the public domain via PDDL or CCZero is strongly recommended and ensures compliance with both the Science Commons Protocol for Implementing Open Access Data and the Open Knowledge/Data Definition.

Comments during the workshop

- A number of participants raised the point of data “ownership”, which is jointly “owned” by the scientists (grad students, post-docs and PI), but does not belong to any one individual.
- The other issue was on whether the “university” owns the research, and whether there is copyright issues in submitting data. However, it does not seem that this is a concern.



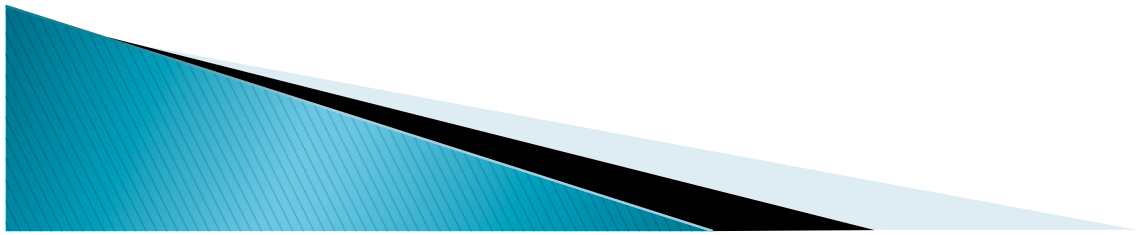
Data management and Ownership issues (at MSU).

- See the following links:

http://www.vprgs.msu.edu/dataguidelines

http://www.vprgs.msu.edu/node/1439

http://www.lib.msu.edu/about/diginfo/collect.jsp



For more information

- http://www.lib.msu.edu/about/diginfo/ldmp.jsp
- http://www.dataone.org/resources
- (for best practices, tools and how to make an effect dataplan)