Good Enough Practices: Data Management

Review Over Papers:

"Good enough practices in scientific computing" by Wilson et al, 2017

"Best Practices for Scientific Computing" by Wilson et al, 2014

Nicholas A. Del Grosso

“As with bench experiments, not everything must be done to the most exacting standards; however, scientists need to be aware of best practices both to improve their own approaches and for reviewing computational work by others.”

Wilson et al, 2017

Save the Raw Data

Never overwrite with “Cleaned Up” Data

Back up the raw data in multiple locations!

Protect Your Raw Data from Accidental Modification

Log the Raw Data’s Hash to Be Able to Check for Changes

Create the data you wish to see in the world.

Use Open File Formats

CSV, JSON, YAML, XML, HDF5

Use self-descriptive variable names and data codes

Don’t use artificial codes for missing data.

NA, not -99

Store essential metadata in a descriptive, regularly-patterned filename

Ruthlessly Eliminate Duplication.

Normalize Data.

Every piece of data must have a single authoritative representation in the system (4a).

Keep Raw Data Canonical

Use the same version of the raw data everywhere.

Keep Constants Constant.

Physical constants ought to be defined exactly once.

site	1999	2000
Whitehorse	745	2666
Yellowknife	37737	80488
Inuvik	212258	

Wide Table

site	year	cases
Whitehorse	1999	745
Whitehorse	2000	2666
Yellowknife	1999	37737
Yellowknife	2000	80488
Inuvik	1999	212258

“Tidy” Long Table

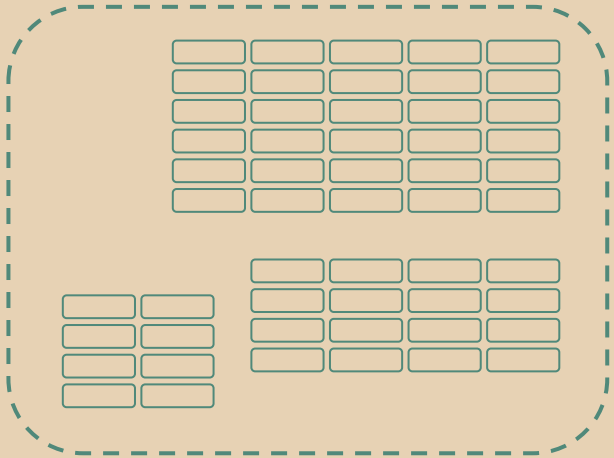
- Records of Observations
- Simpler to Analyze with Statistical Software
- Elegantly handles “Missing” Data.

Share Data in “Tidy” structures to ease statistical analysis

Anticipate the need to use multiple records

ID	site	year	cases
1	Whitehorse	1999	745
2	Whitehorse	2000	2666
3	Yellowknife	1999	37737
4	Yellowknife	2000	80488
5	Inuvik	1999	212258
6	Inuvik	2000	213766

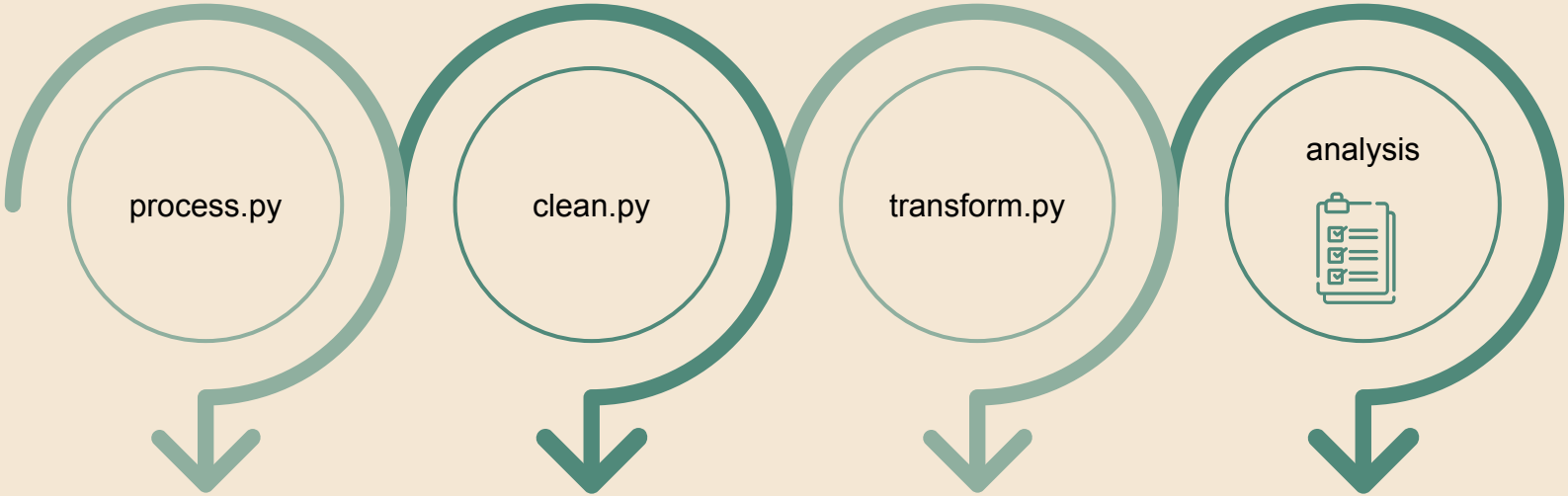
Add a UUID or PK/Dataset combination for every record



Use File Formats that Support Multiple Datasets

JSON, SQLITE, HDF5,
NetCDF, NWB

Record all the steps used to process data.



Submit data to a reputable DOI-issuing repository

Thanks!

DO YOU HAVE ANY QUESTIONS?

delgrosso.nick@gmail.com
Github: https://github.com/nickdelgrosso
linkedin: https://www.linkedin.com/in/nick-del-grosso/