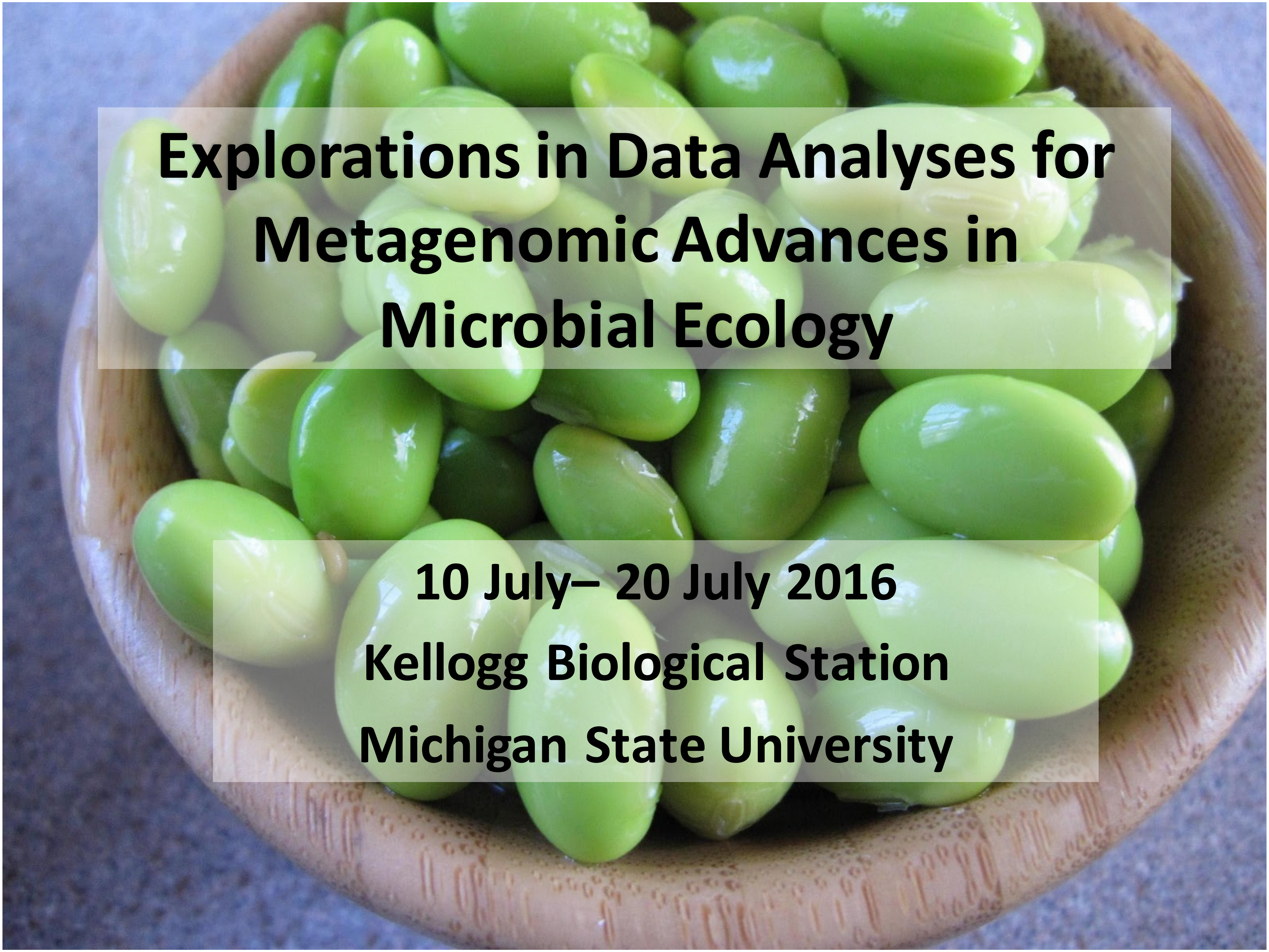


Explorations in Data Analyses for Metagenomic Advances in Microbial Ecology

10 July– 20 July 2016

Kellogg Biological Station

Michigan State University

What is a workflow?

- Exactly what you tell the computer to execute the analysis
- Each optimized step in a computing analysis
  - Verbatim scripts that were executed
  - Annotated:
    - Software versions used
    - Description of what the software is doing/goal of that step
    - Brief notes on deviations from default options
- Workflows can include different software (e.g., PANDAseq to QIIME to R), and should also include all “formatting steps” needed to move between tools – hopefully you don’t need to manually format too much; avoid if possible

Workflows should be mindlessly complete – the computer is a literal beast

The Peanut Butter and Jelly Robot

https://www.youtube.com/watch?v=leBEFaVHllE

https://www.youtube.com/watch?v=Y-UEdr1wofM

Table discussions - Etherpad

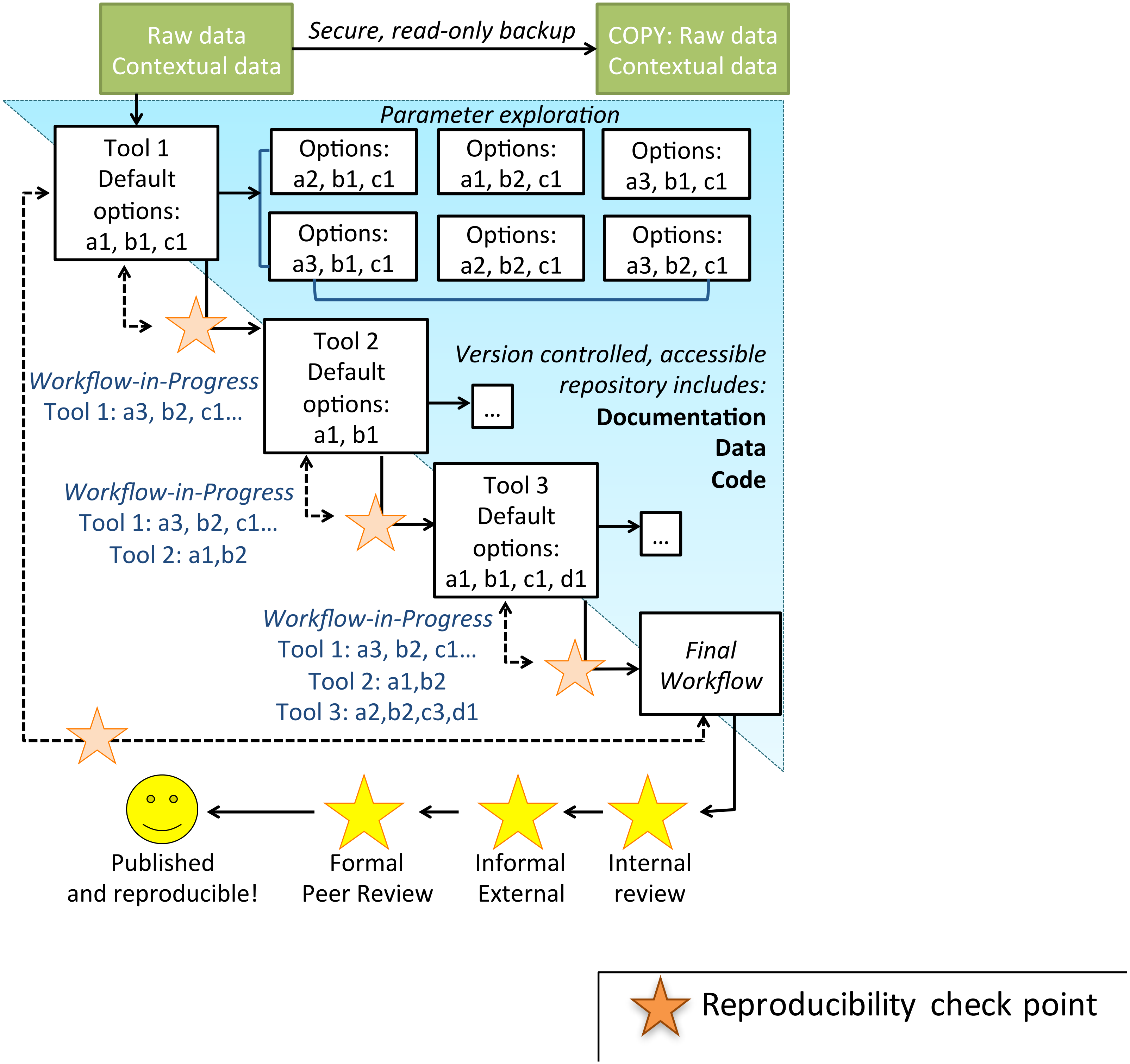
- What are some steps that you take when you start an analysis workflow?
- What is the very first step? The very last?
- What is the most important step?
- What is the best strategy/advice that you use for analysis?
- What is something in a computing workflow that you currently do, but you want to improve?

Computing Workflows for Biologists

- Papers of interest:
  - Wilson et al. 2014. Best practises for Computing. *PLoS Computational Biology*
  - Nobel 2009. Organizing Computational Biology Projects. *PLoS Computational Biology*
  - Sandve et al. 2013. Ten simple rules for reproducible computational research. *PLoS Computational Biology.*
    - » *All of these references are posted in our Mendeley group*

Our suggestions for an analysis approach

1. Adopt a systematic, iterative exploration of parameter space.
   - Include “sanity checks”
   - Focus on exploring the parameters that *matter* for your objective/hypothesis
   - Organize your output and input for *someone who isn’t you*
2. Work towards an optimized, seamless workflow.
3. Implement *reproducibility check-points.*
4. Maintain computing notes just as you would experimental notes
5. Do your part: cultivate a shared responsibility for reproducibility of results and data management


Raw data
Contextual data
Secure, read-only backup
COPY: Raw data
Contextual data
Parameter exploration
Tool 1
Default
options:
a1, b1, c1
Options:
a2, b1, c1
Options:
a1, b2, c1
Options:
a3, b1, c1
Options:
a3, b1, c1
Options:
a2, b2, c1
Options:
a3, b2, c1
Tool 2
Default
options:
a1, b1
…
Version controlled, accessible
repository includes:
Documentation
Data
Code
Workflow-in-Progress
Tool 1: a3, b2, c1…
Workflow-in-Progress
Tool 1: a3, b2, c1…
Tool 2: a1,b2
Tool 3
Default
options:
a1, b1, c1, d1
…
Workflow-in-Progress
Tool 1: a3, b2, c1…
Tool 2: a1,b2
Tool 3: a2,b2,c3,d1
Final
Workflow
Published
and reproducible!
Formal
Peer Review
Informal
External
Internal
review
Reproducibility check point

Naming Conventions

Exampl

20_A_T

int 1,

rep1)

Exampl

Ashley’

A

Ashley

Exampl

ALS1, A

Improv

ALS01, ALS02, ALS03…ALS10, ALS11

Our samples, e.g.
C01_05102014_R1_D01

C01 – Centralia core site 1
Date 05102014 – 05 Oct 2014
R1 – core 1 (there were sometimes multiple cores from the same site)
D01 – DNA extraction replicate 1 D01- DNA extraction rep 1

…
F – forward read; R = Reverse read

Subsampling – sometimes required for a large dataset to work efficiently through a workflow

- Check out our tutorial about subsampling:

- https://github.com/edamame-course/2015-tutorials/blob/master/final/2015-06-23-QIIME1.md#ampliconsubsampling