Pacific Northwest
NATIONAL LABORATORY

Beyond Box Checking: Best Practices for Managing Your Data and Code

Casey Burleyson and Chris Vernon

U.S. DEPARTMENT OF ENERGY BATTELLE

PNNL is operated by Battelle for the U.S. Department of Energy

Goal For This Talk

In this presentation we will present one approach to documenting the data and code associated with your publications.

Casey – Data Management Lead

Chris – Lead Software

Open Science Thesis

Adopting open science principles and methods will help you to get work done by reducing barriers to entry, gaining economies of scale, and avoiding duplication of effort.

Reproducibility: The ability to reproduce the results of others (and yourself) is a critical component of the overall scientific method and is fundamental to ensuring scientific integrity.

Extensibility: The ability to extend the work of others (and yourself) is critical to enabling the community to scaffold experiments across projects and over time and ultimately increase the pace of scientific discovery.

Pacific Northwest
NATIONAL LABORATORY

Open Science: An Analogy

Open Science: An Analogy

2003-2007

2010

2007-2010

2007

2021

Open Science: An Analogy

1) **Asking a compelling question**
   - *Experience; Knowledge*
2) **Experimental design & model formulation**
   - *Data; Code; Knowledge*
3) **Model forcing**
   - *Data; Code; Knowledge*
4) **Multi-model experiments**
   - *Data; Code; Knowledge*
5) **Model analysis**
   - *Data; Code; Knowledge*
6) **Publication**

Journals are now requiring that data be made available at publication. Of this workflow, what is typically archived?

1) **Asking a compelling question**
   - *Experience; Knowledge*
2) **Experimental design & model formulation**
   - *Data; Code; Knowledge*
3) **Model forcing**
   - *Data; Code; Knowledge*
4) **Multi-model experiments**
   - *Data; Code; Knowledge*
5) **Model analysis**
   - *Data; Code; Knowledge*
6) **Publication**

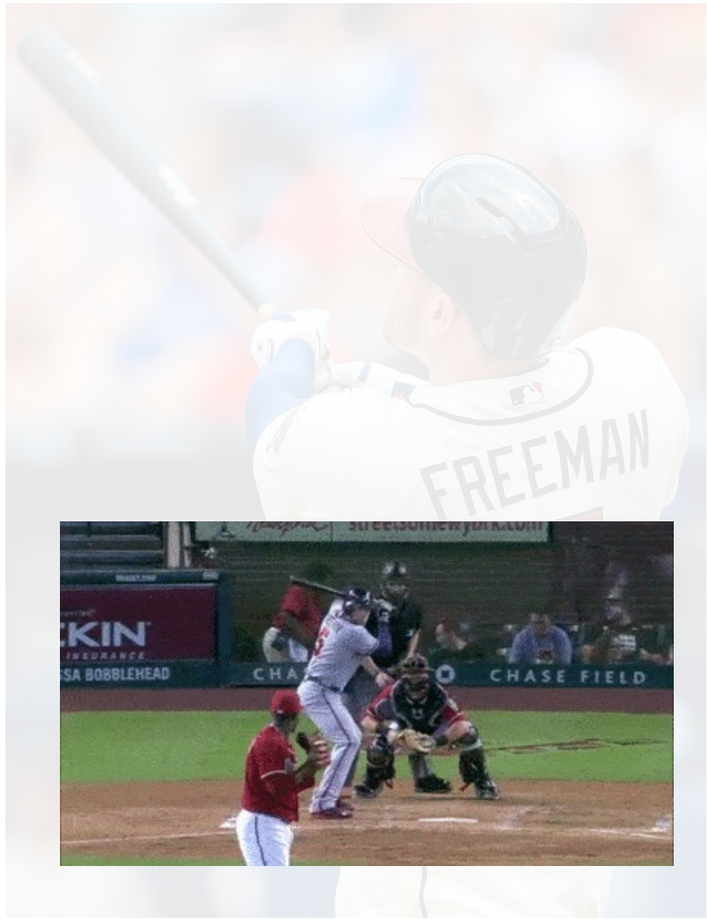
Journals are now requiring that data be made available at publication. Of this workflow, what is typically archived?

2003-2007

2010

2007-2010

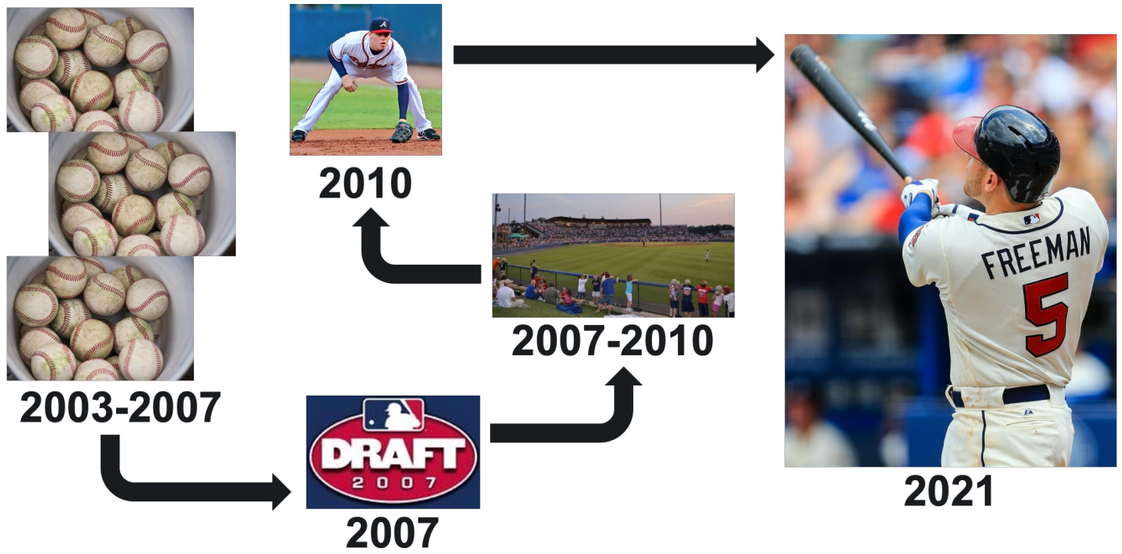
2007



2021

What is Stopping Us?

- Demands more intentional and explicit workflows for developing and managing data, models, experiments, and publications
- Lack of training
- Temporal mismatches between cost (now) and benefits (later)
- Easy to be overwhelmed by the plethora of tools and myriad of mandates

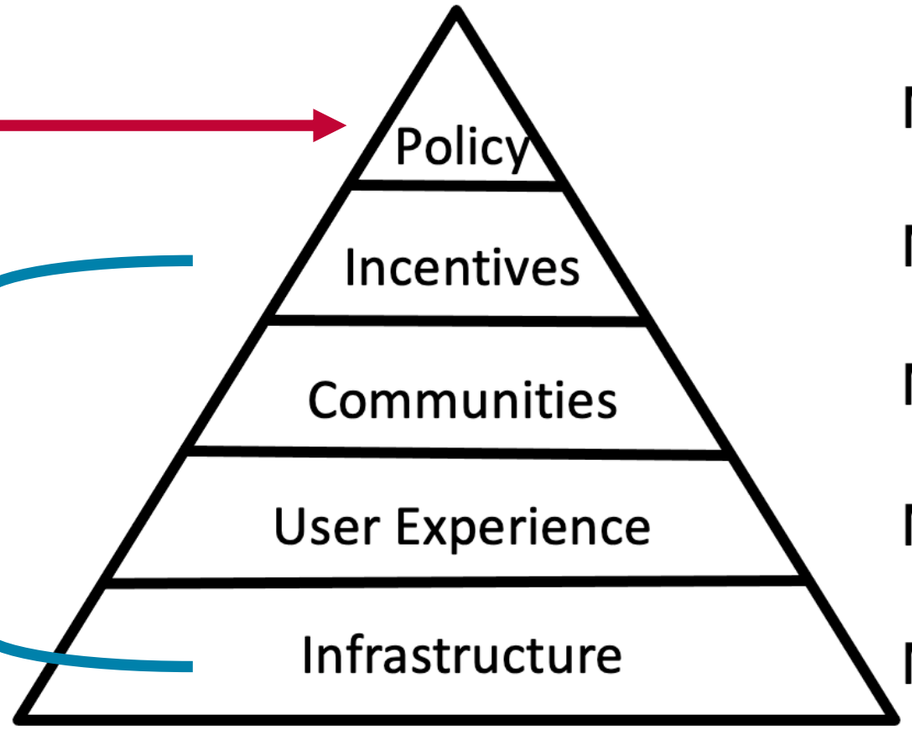


"Individual practitioners or institutions tend to have their own data creation and management practices, driven by their own mandates, historical protocols, and the organic evolution of data. This results in a wide diversity of types, practices, and models for characterizing and handling data, which significantly hinders reproducibility of science and reuse of data." – IHTM Workshop Report

Facilitating Open Science

Journals and funders largely skipped to this end of the pyramid…

The open science community is tackling these foundational elements of the pyramid…



Make it required

Make it rewarding

Make it normative

Make it easy

Make it possible

Conceptual diagram from Brian Nosek of the University of Virginia and the Center for Open Science

Open Science: An Analogy

1) **Asking a compelling question**
   - *Experience; Knowledge*
2) **Experimental design & model formulation**
   - *Data; Code; Knowledge*
3) **Model forcing**
   - *Data; Code; Knowledge*
4) **Multi-model experiments**
   - *Data; Code; Knowledge*
5) **Model analysis**
   - *Data; Code; Knowledge*
6) **Publication**