Data for Data Science

fredhutch.io
Fred Hutchinson Cancer Research Center

Sign in here: https://hackmd.io/@k8hertweck/introData

Course outline

- Class 1: Data entry and creating spreadsheets
- **Class 2: Organizing data and project files**
- Class 3: Documenting data with metadata
- Class 4: Data manipulation and reproducibility

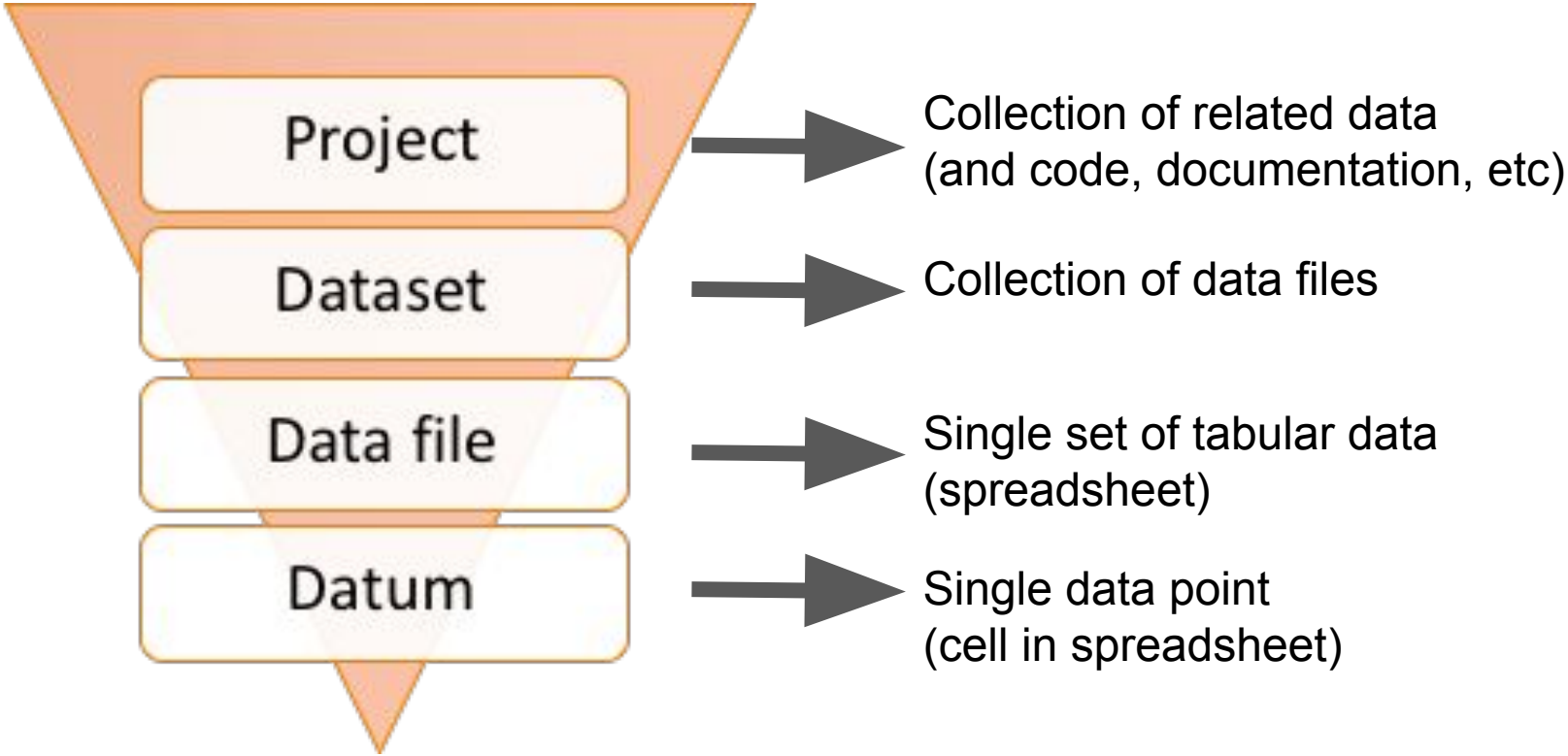
Today’s objectives

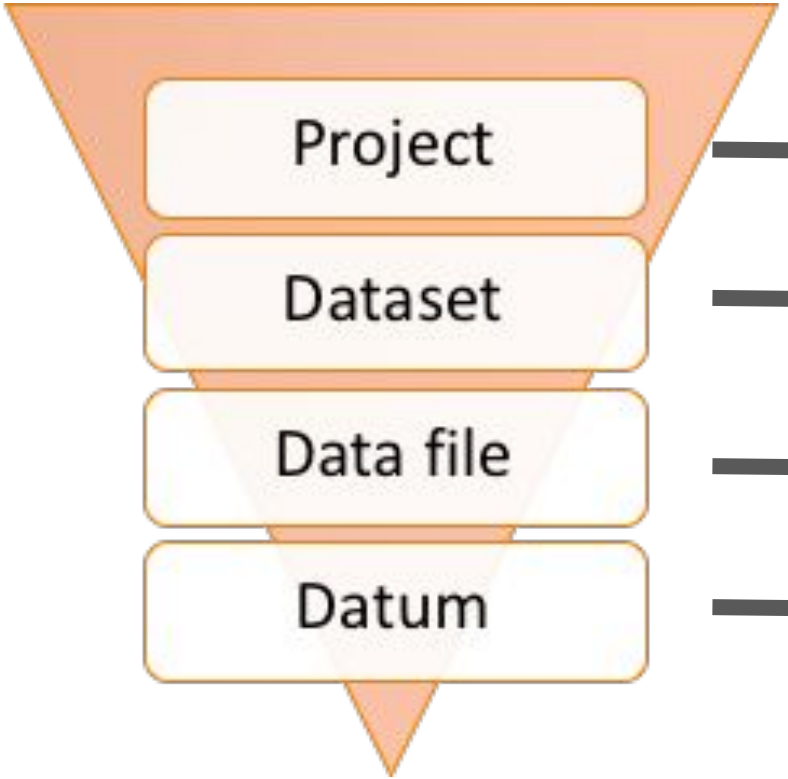
After today’s class, you should be able to:

- Describe different types of data and data files
- Implement best practices in organizing data (project) files
- Access resources available at Fred Hutch to manage data files

In the context of your work, what are “data”?

What are some examples of a “dataset”?


Project
Dataset
Data file
Datum
Collection of related data
(and code, documentation, etc)
Collection of data files
Single set of tabular data
(spreadsheet)
Single data point
(cell in spreadsheet)



How to organize collections of related files for a single project?

How to standardize data files?

How to organize data in a spreadsheet?

What goes in a single cell?

Good Enough Practices in Scientific Computing

1. **Data management**: saving both raw and intermediate forms, documenting all steps, creating tidy data amenable to analysis.
2. **Software**: writing, organizing, and sharing scripts and programs used in an analysis.
3. **Collaboration**: making it easy for existing and new collaborators to understand and contribute to a project.
4. **Project organization**: organizing the digital artifacts of a project to ease discovery and understanding.
5. **Tracking changes**: recording how various components of your project change over time.
6. **Manuscripts**: writing manuscripts in a way that leaves an audit trail and minimizes manual merging of conflicts.

Items in blue will be a focus of this class

Project Organization: Overview

Put each project in its own directory, which is named after the project.

Divide work into projects based on overlap of data and code.

The examples presented here are for a generic analysis that isn't based in code.

```
|-- CITATION.txt  (how to reference project)
|-- README.txt    (overview of project)
|-- LICENSE.txt   (how project/code can be used)
|-- data/
|-- doc/
|-- results/
|-- analysis/
```

For this presentation, files have a suffix, folders end in /

Modified from Good Enough Practices

Project Organization: Documentation

Put text documents associated with the project in the `doc` directory.

If you use other software or tools for these tasks, it would be useful to describe that here.

```
|-- doc/
|   |-- notebook.txt (electronic lab notebook)
|   |-- manuscript.doc (manuscript draft)
```

Modified from Good Enough Practices

Project Organization: Raw data

```
|-- data/
|   |-- original_data.xlsx  (raw data, before tidying)
|   |-- data.csv  (raw data, with tidy data principles applied)
|   |-- README.txt  (metadata, or information about data)
```

Put raw data and metadata in a `data` directory

(we'll spend the next class session talking more about metadata!)

Modified from Good Enough Practices

Project Organization: Results

Put files generated during cleanup and analysis in a `results` directory.

A separate `figures` directory may also be useful.

```
|-- results/
|   |-- filtered_data.csv
|   |-- summarized_results.csv
|-- figures/
|   |-- scatterplot.csv
```

Modified from Good Enough Practices

Project Organization: Results

For projects that have many intermediate steps, your results directory may be separated into additional folders

```
|-- results/
|   |-- QC_sequences/
|   |-- assembly/
|   |-- variant_analysis/
```

Modified from Good Enough Practices

Project Organization: Analysis

Put files used for data analysis and creating figures in `analysis`

This may include files created from analysis programs, like Tableau, Geneious, etc.

```
|-- analysis/
|   |-- Sept_analysis.xlsx
```

Modified from Good Enough Practices

Project Organization: Analysis

If you are writing code for analysis, it is more appropriate to separate files into:

- Source code in the `src` directory.
- External scripts or compiled programs in the `bin` directory.

```
|-- bin/
|   |-- labmates_script.R
|-- src/
|   |-- data_analysis.R
```

Modified from Good Enough Practices

Project Organization: Naming files and folders

Why do names matter? Name all files to reflect their content or function.

- Use appropriate file suffixes
- Avoid spaces and special characters
- Avoid sequential numbers
- Avoid relative position in manuscript (Fig1, Fig2, etc)

```
|-- raw_data/
|   |-- insulin_experiment_2017.csv
|-- figures/
|   |-- clinical_scatterplot.jpg
```

Modified from Good Enough Practices

Project Organization Summary

1. Put each project in its own directory, which is named after the project.
2. Include license, citation, and README documents for the project
3. Put text documents associated with the project in the `doc` directory.
4. Put raw data and metadata in a `data` directory and files generated during cleanup and analysis in a `results` directory.
5. Put files associated with analysis or creating figures in the `analysis` directory.
6. Name all files to reflect their content or function.

These guidelines may vary based on the lab, project, or organization!

Modified from Good Enough Practices

Project organization activity

Reorganize the files in sample_project such that:

- Files are named consistently with no special characters
- Files are separated into relevant folders
- A file called README describes the basic file structure

What are some challenges associated with organizing data files for a project?

What are “big data”?

What is a project?

- Challenges:
  - Data may be used across multiple projects with different collaborators
  - Long-term projects may have intermediate and longitudinal analysis
- Solutions:
  - Focus on specific deliverables (publications, reports, etc)
  - Develop workflows that allow standardization, not necessarily total synchronization

How to store/access data files? Challenges

- Big data:
  - Many data files: hundreds or thousands
  - Large data files: MB, GB, TB?
  - Files with different data types: genomic, clinical, simulations, statistical output
- Data security: balancing protection of sensitive information with allowing access to researchers
- Data transfer: maintaining data integrity while publishing, sharing, and archiving data

How to store/access data files? Solutions

- Solutions:
  - Organize data into projects and otherwise plan for long term, team-based approaches
  - Fred Hutch Data Science Wiki: Data Storage section describes available resources
  - Office hours available for various groups at Fred Hutch who can help provide recommendations

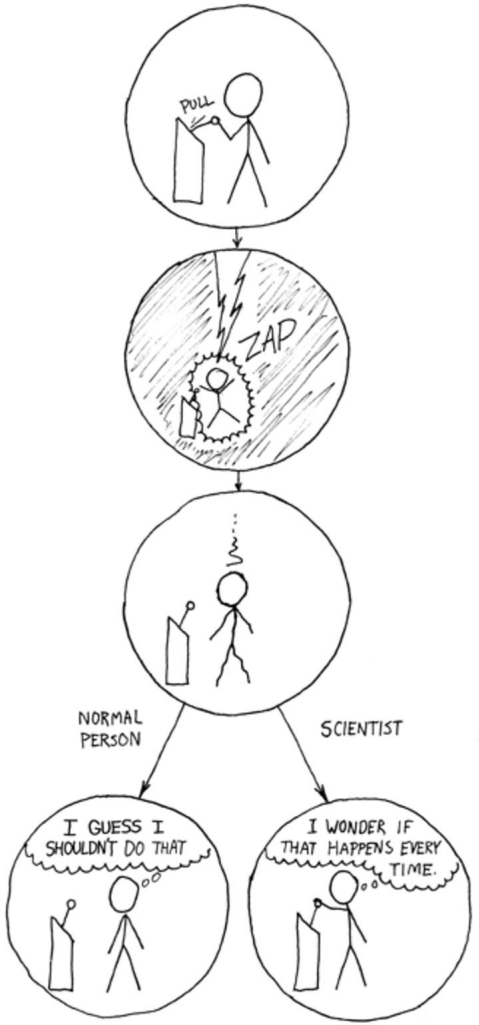
“Legacy” issues with data

- Challenges:
  - You may “inherit” data from coworkers/collaborators
  - Long-term studies may have suboptimal, unscalable strategies
- Solutions:
  - If a project is nearing completion, it may not be worth adjusting strategies and you should focus on what is necessary for publication/archival
  - Identify a natural point of transition (if possible), and document how management has changed

Reproducibility and transferability are big data management challenges!

- *Reproducibility*: obtaining the same results multiple times
  - Confirm previously published scientific results
  - Automate large-scale data analysis projects

- *Transferability*: using data multiple times
  - Among researchers
  - Among research questions



Challenges with the data life cycle

We tend to be most concerned about data collection and analysis.

Best practices in data management help facilitate other steps in this process as well!

Primer on Data Management

Summary

- Applying best practices for project management can help streamline the process of analyzing and reporting your data
- Plan ahead to help both yourself and other people who may need to use the data later
- Every project is different, and may require specific practices to suit its needs

Next time: documenting data with metadata

Resources

- A Quick Guide to Organizing Computational Biology Projects
- Good Enough Practices in Scientific Computing
- Fred Hutch Biomedical Data Science Wiki: Section on Data Storage in Scientific Computing