UCDAVIS DataLab
Data Science and Informatics

README, Write Me!

Digital Project Management

March 4th, 2021

Victoria Farrar, Dr. Pamela L. Reynolds

vsfarrar@ucdavis.edu
plreynolds@ucdavis.edu

Please keep videos off and stay muted to reduce bandwidth.

Please note: Today's workshop will be recorded and posted to DataLab's website.

UC DAVIS
DataLab
Data Science and Informatics

Research
Training
Community

»» Increase UC Davis' research impact via data-driven expertise.

»» Support the next generation of data-capable researchers and students.

»» Foster and coordinate data-enabled researchers and university units.

Supporting innovation, accelerating research for the entire community.

http://datalab.ucdavis.edu

Workshop Overview

- Principles and Levels of Metadata
- READMEs
- Data Dictionaries / Codebooks
- Workflow Diagrams
- Studio: Document your own data!

UCDAVIS
DataLab
Data Science and Informatics

Learning Objectives

By the end of this workshop, you should be able to:

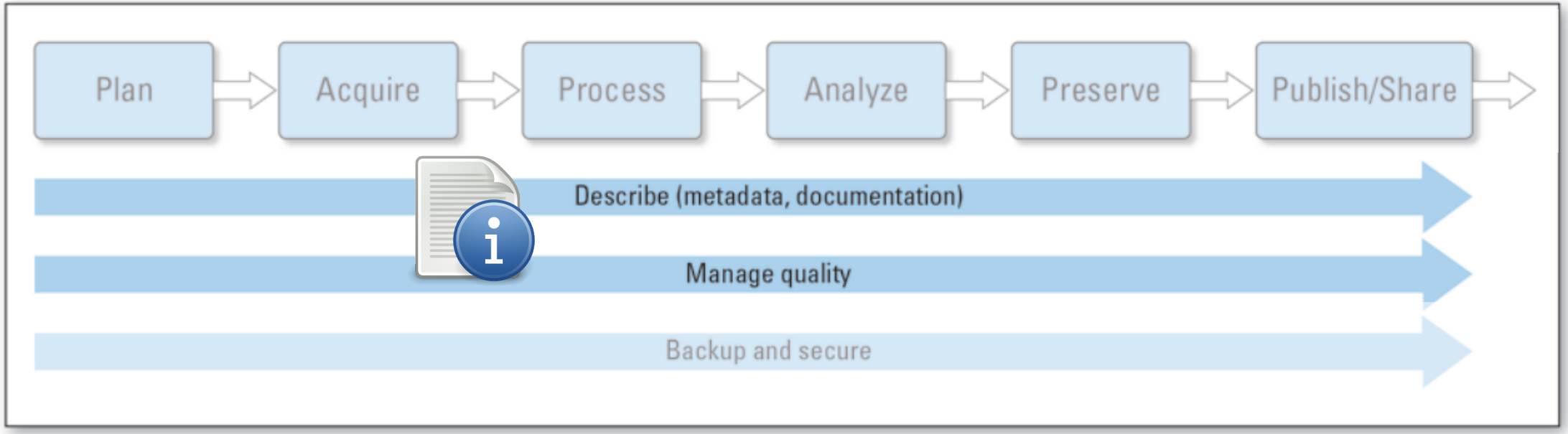
- Define metadata
- Understand the importance of collecting metadata / data documentation
- Recognize the different ways metadata can be recorded at the project and dataset level
- Locate the relevant metadata standard for your field
- Produce a data dictionary / codebook for a tabular dataset
- Document the organization of a data-driven project using a README
- Represent how their data is processed and analyzed using a workflow diagram

What is metadata?

"structured information that describes, explains, locates, or otherwise makes it easier to retrieve, use, or manage an information resource. Metadata is often called ***data about data*** *or* ***information about information****"*

- National Information Standards Organization

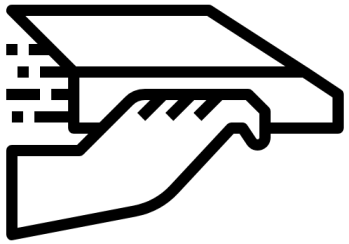
Metadata across the data lifecycle



Source: USGS

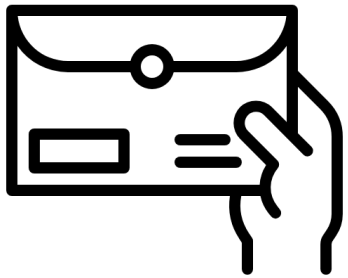
Why collect metadata?

As a data collector or creator,

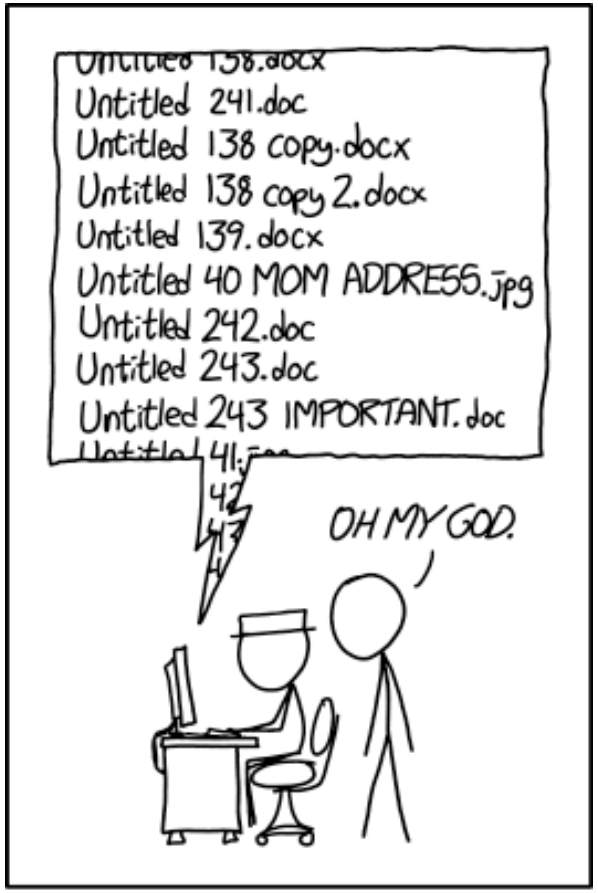


- Helps other researchers find, interpret and use your data
- Helps you use your own data in the future

As a data user,



- Enables you to find and interpret data from other researchers
- Helps you reproduce analyses or analyze data in new ways



PROTIP: NEVER LOOK IN SOMEONE ELSE'S DOCUMENTS FOLDER.

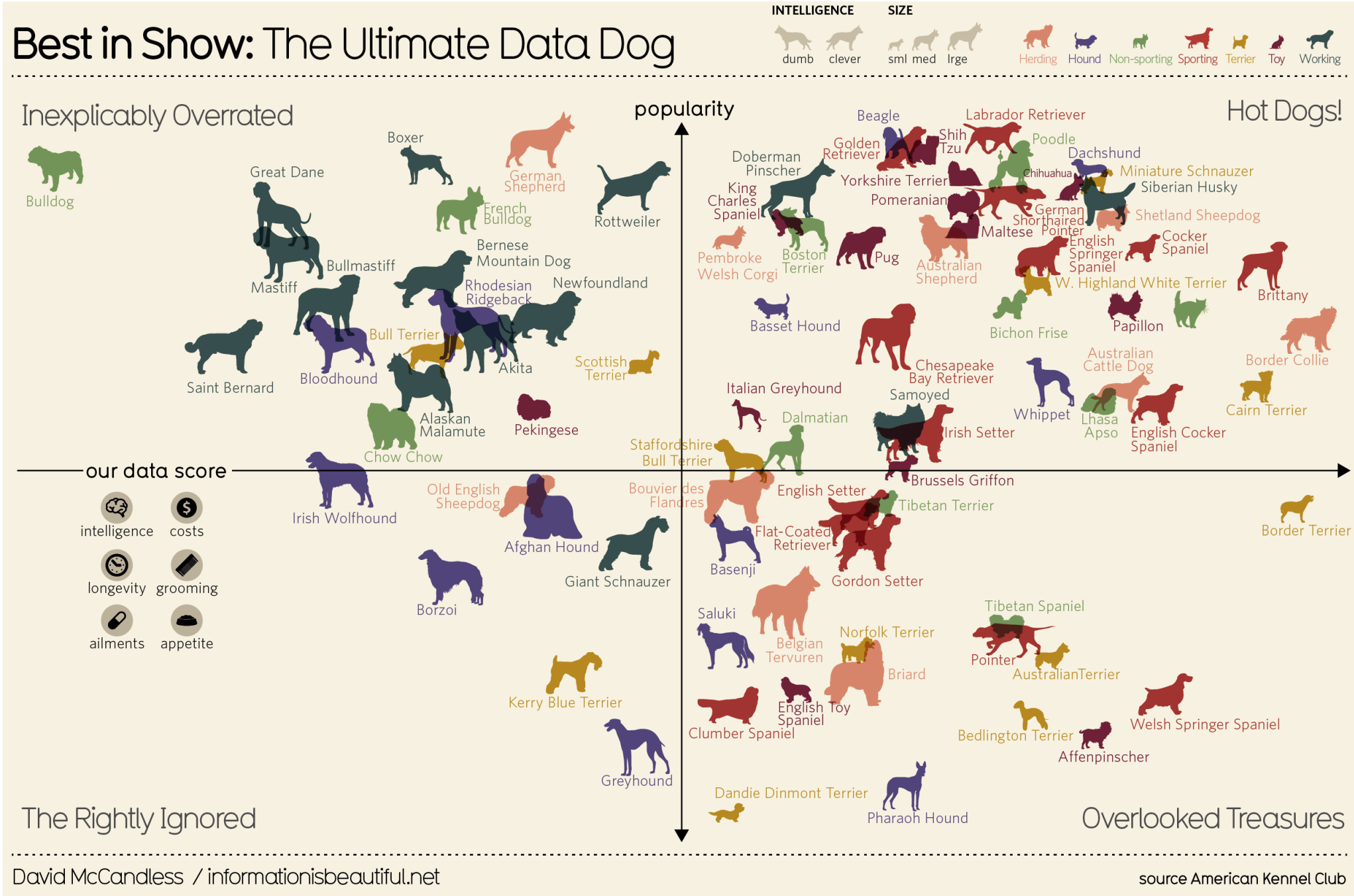
UCDAVIS DataLab
Data Science and Informatics

Why collect metadata?

Activity: What should we document?

What should be included in metadata or documentation for a dataset?

Let’s ask some questions about the following figure to see what we would want to know.



David McCandless / informationisbeautiful.net

Drop your response into the chat.

UCDAVIS DataLab
Data Science and Informatics

What is one question you could answer right now about YOUR data that would make it more accessible or useable?

Drop your responses in the chat.

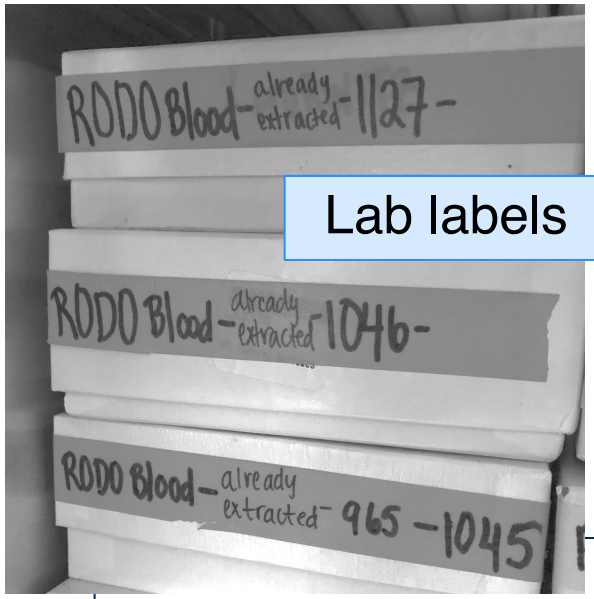
UCDAVIS DataLab
Data Science and Informatics

Metadata can cover…

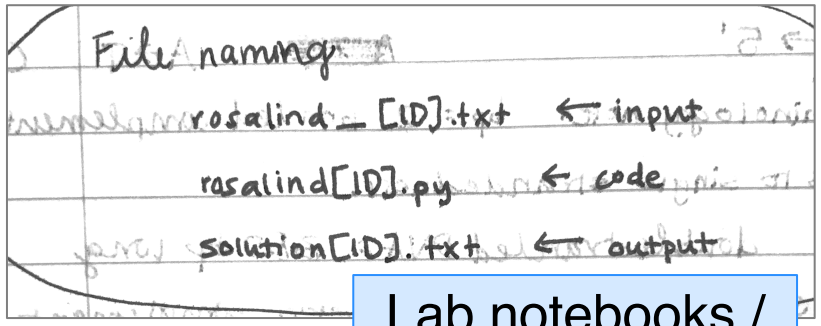
- What processes were used for creating or collecting the data?
- What do the values in the table mean?
- What software do I need in order to read the data?
- Can I give these data to someone else?
- Why were the data created?
- How is the data organized?
- What does the data mean?
- How should the data be cited?
- If I want to reproduce the analysis, how should I do so?

Metadata can take many different forms

… and you're probably already collecting it!



Lab labels



Lab notebooks / Notes to self

- scicomm_course_grades.csv
- scicomm_data_joined
- scicomm_data_joined.csv
- scicomm_final_survey_numeric.csv
- scicomm_final_survey.csv
- scicomm_group_assgn.csv
- scicomm_initial_survey.csv
- scicomm_key.xlsx

File names

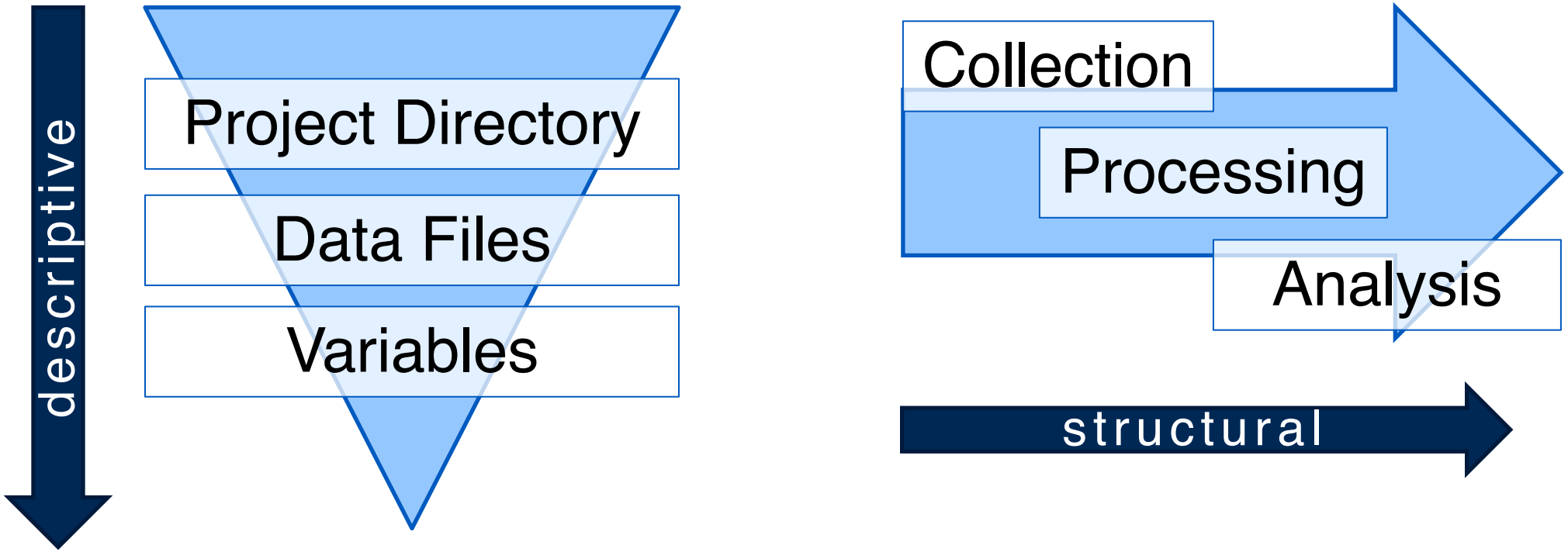
On Sun, Nov 29, 2020 at 4:48 PM Victoria Farrar <vsfarrar@ucdavis.edu> wrote:

I just uploaded the code (matches what is currently on Github) to the folde comes /code_files/2020-11-29_github.
The data this code references is the most recent file from Brad, 2020-09-16_ wrangled.csv.
Let me know if you also want a copy of the cleaned output from the data parsing file.

Emails / notes to collaborators

UCDAVIS DataLab
Data Science and Informatics

Levels of Metadata



Modified from Illinois Research Data Service

Data Management Plans (DMP)

Documentation is only one part of the data life cycle and only one part of data management.

As you begin a data-driven project, you may want to consider making a data management plan (DMP).

UCDAVIS
DataLab
Data Science and Informatics

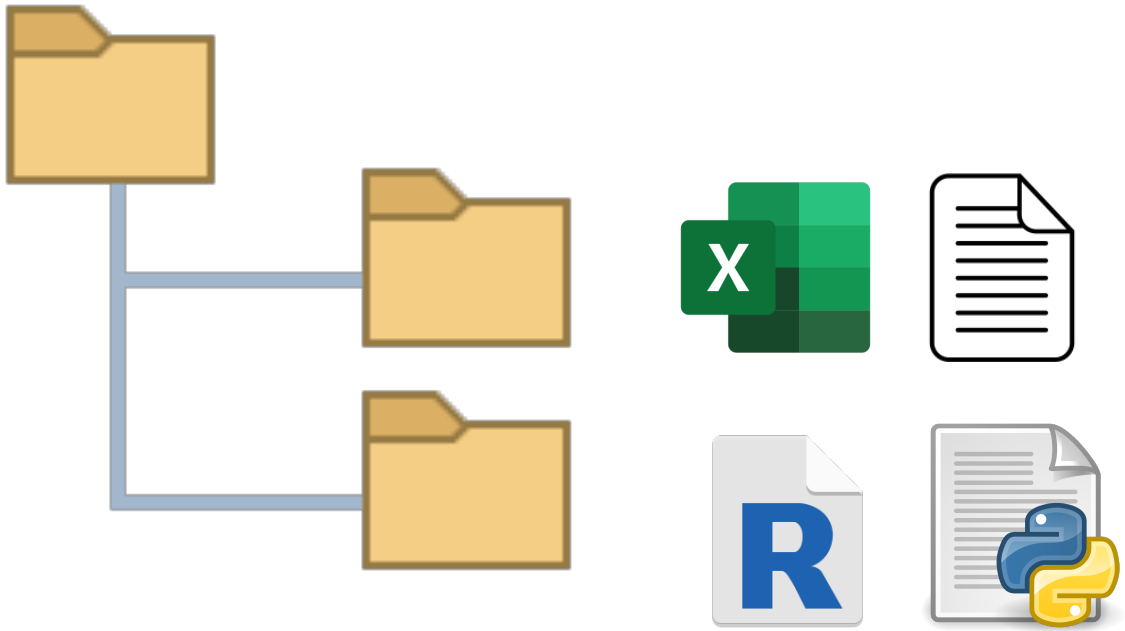
Project Level: File Directory

descriptive

Project Directory

Data Files

Metadata: README



Modified from Illinois Research Data Service

UCDAVIS DataLab
Data Science and Informatics

READMEs

- Text files that serve as a “map” of the organization, contents, and background of your data-driven project

- Examples:
  - README.txt example (University of Arizona)
  - README.md example (DataLab Hack for California)

READMEs should include …

- Project title, investigator, institution and dates of project
- Brief description of methods for **data collection** and **generation**
- Brief description of **data processing and analysis**
- Brief description of what each **folder and/or data file contains**
- Date each file was created
- Overview of the **file naming scheme**

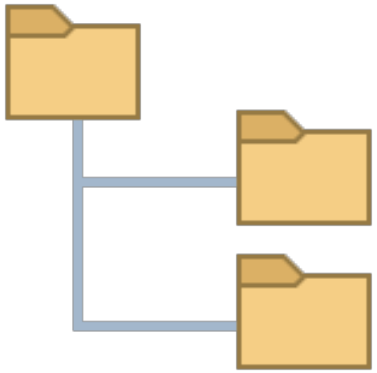
UCDAVIS DataLab
Data Science and Informatics

README Best Practices

- Be written in plain-text (.txt) or markdown (.md)
- Live in the main directory of the project folder or file they describe
- Be named so that it is clear what project / file it describes
- Be formatted similarly across different projects
- Use standardized formats, such as ISO 8601 for dates (YYYY-MM-DD)

File Directory Organization

- Keep all your project files in a single project folder
- Organize files into folders by workflow step, or by result or product



A) Organized by File type

Example.A
- Code
  - Step.1
  - Step.2
- Data
  - Processed
  - Raw
- Results
  - Figure.1
  - Figure.2
  - Models
- readme.txt

B) Organized by Analysis

Example.B
- Figure.1
  - Code
  - Data
  - Results
- Figure.2
  - Code
  - Data
  - Results
- readme.txt

File Naming Schemes

Example file naming conventions:

[investigator]_[method]_[subject]_[YYYYMMDD]_[version].[ext]

[project #] _[method]_[version]_[YYYYMMDD].[ext]

[YYYYMMDD] _[version]_[subject]_[datacollector].[ext]

[type of file]_[author]_[date].[ext]

File Naming Schemes

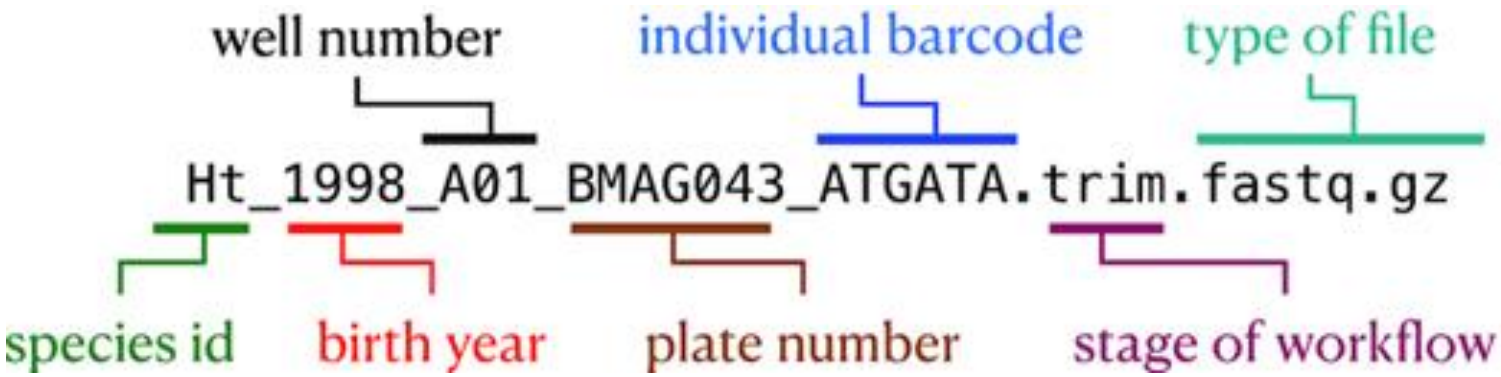
Example file naming conventions:

[investigator]_[method]_[subject]_[YYYYMMDD]_[version].[ext]

[project #] _[method]_[version]_[YYYYMMDD].[ext]

[YYYYMMDD] _[version]_[subject]_[datacollector].[ext]

[type of file]_[author]_[date].[ext]

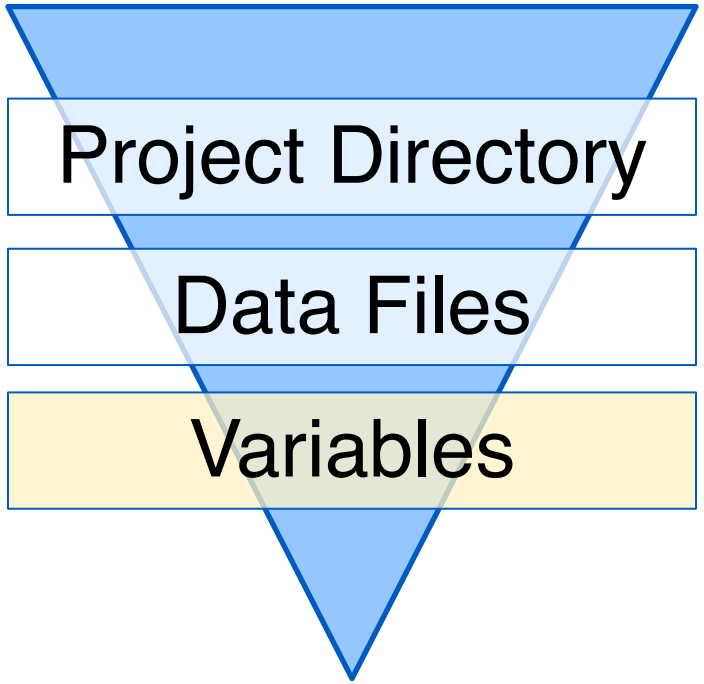


Reiter et al (2021)

UCDAVIS DataLab
Data Science and Informatics

Dataset Level: Variables

descriptive



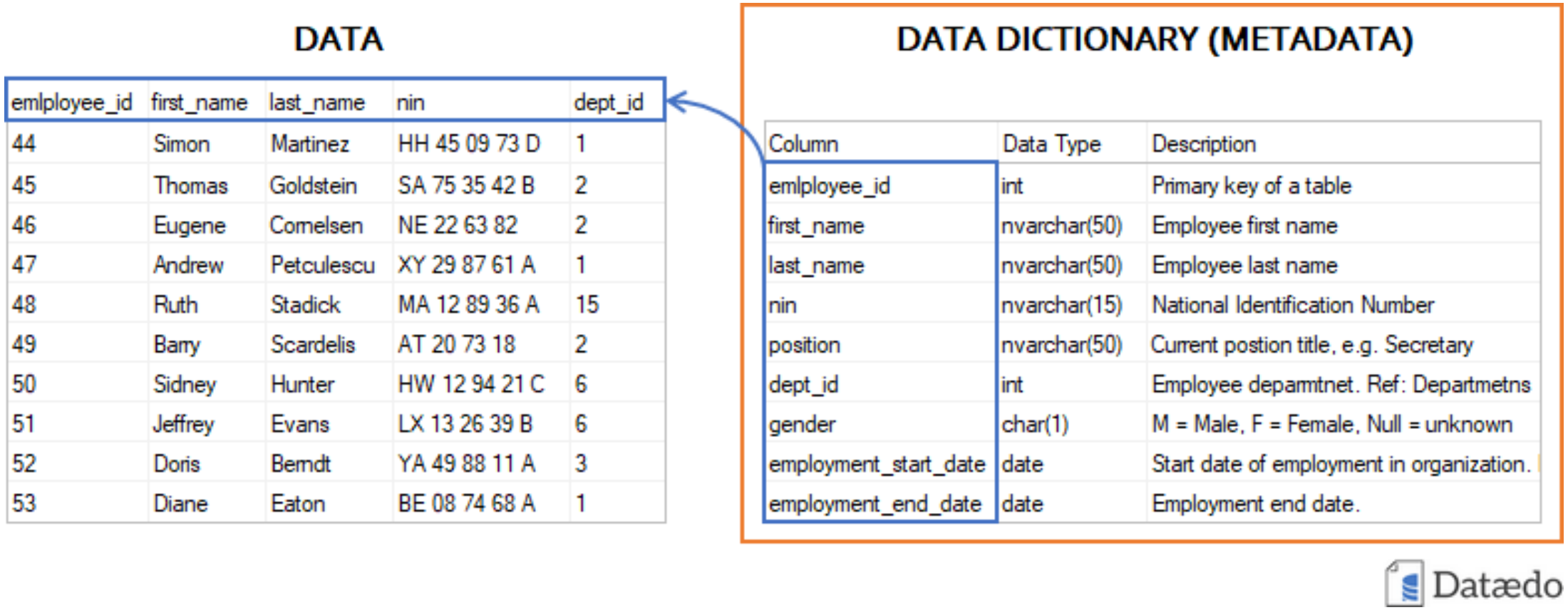
Metadata: Data Dictionary or Codebook

Id	Var1	Var2	Var3
1			
2			

Modified from Illinois Research Data Service

UCDAVIS DataLab
Data Science and Informatics

Data Dictionary Example



Datædo

UCDAVIS DataLab
Data Science and Informatics

Codebook Example

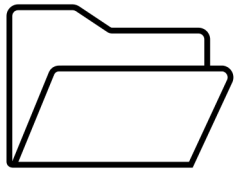
Column	Code	Label	Class
race	0		chear:White
race	1		chear:BlackOrAfricanAmerican
race	2		chear:OtherRace
edu	0	high school degree or less	chear:HighSchoolOrLess
edu	1	technical college or some college	chear:SomeCollegeorTechnicalSchool
edu	2	college graduate	chear:CollegeGraduate
smoke	0	no smoking in pregnancy	chear:NonSmoker
smoke	1	some smoking in pregnancy	chear:Smoker

Rashid et al (2020)

Data Dictionaries should include …

- List of variables / column names and a description of what they include
- The type and the units the variable should be entered in
- The acceptable range or list of values for a variable
- How missing data is coded across the dataset or in specific variables

Data Dictionary Best Practices

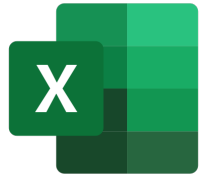


/project

README_project.txt

data1.csv

data2.csv



data3.csv

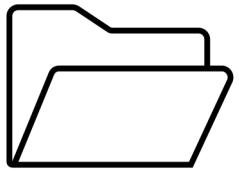
data1_dictionary.csv

data2_dictionary.csv

data3_dictionary.csv

UCDAVIS DataLab
Data Science and Informatics

Data Dictionary Best Practices



/project

README_project.txt

data1.csv

data2.csv

data1_dictionary.csv

data2_dictionary.csv

```
README for PROJECT


#####
Files and Organization


data1.csv : Measurement data from X
machine, collected on DATE. Includes
measurements for X number of units.

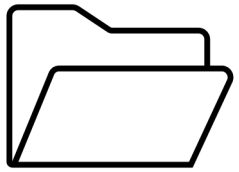

data1_dictionary.csv: Data dictionary for
data1.csv, describes variables, formats,
and units.


….
```

data3_dictionary.csv

UCDAVIS DataLab
Data Science and Informatics

Data Dictionary Best Practices



/project

README_project.txt

data.csv

```
README for PROJECT

…

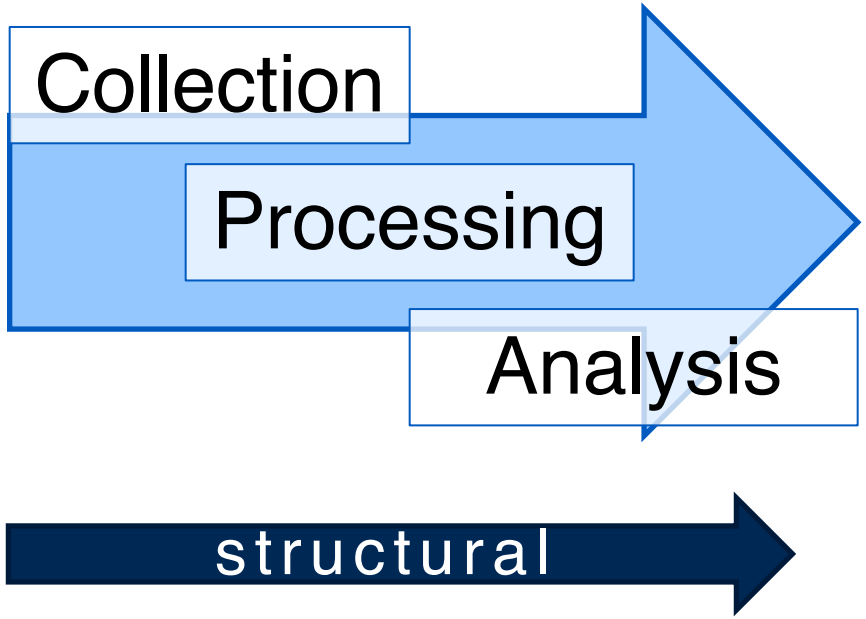
__
Data Dictionary for data1.csv

Var1: Variable 1, numeric, the
measurement of component X, measured
in cm.
Var2 : Variable 2, character, the name
of contributor Y, listed Last,First
format.

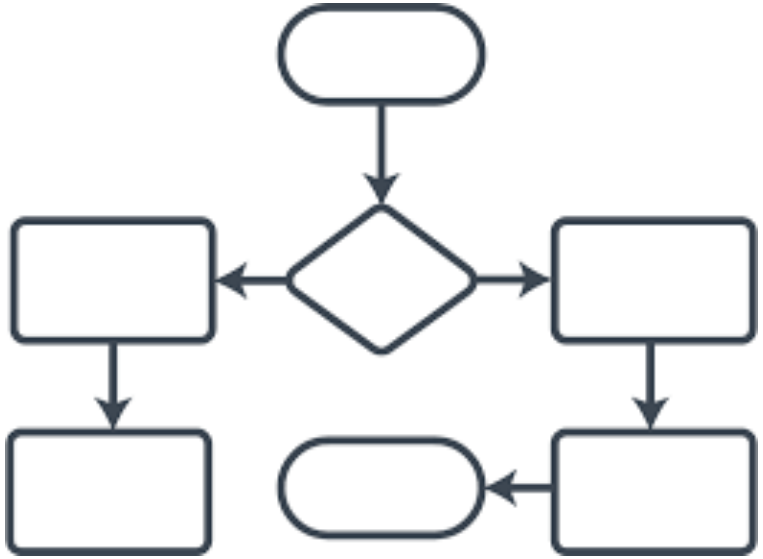
…
```

UCDAVIS DataLab
Data Science and Informatics

Project Level: Workflow Processes



Metadata: Workflow Diagram

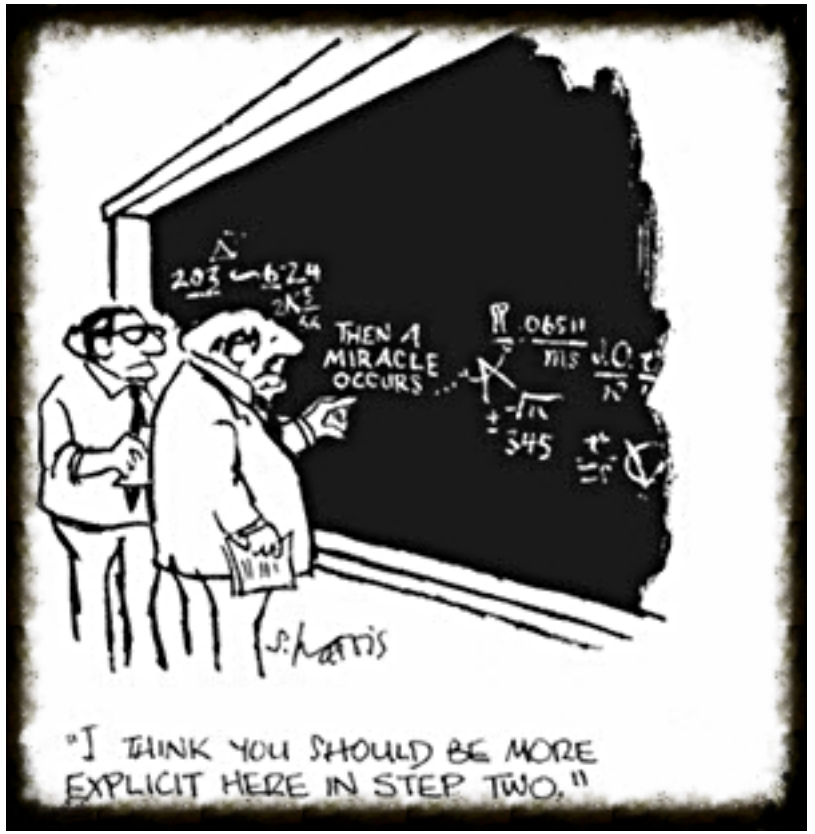


Modified from Illinois Research Data Service

UCDAVIS DataLab
Data Science and Informatics

Structural metadata can go in a README

- How was data processed or filtered from the raw data?
- What was your analysis workflow?
- What software or programming language are required?
- How were any results files, such as summary tables or figures, produced?



UCDAVIS
DataLab
Data Science and Informatics

```
Prepare for cleaning:
1. Download and install OpenRefine

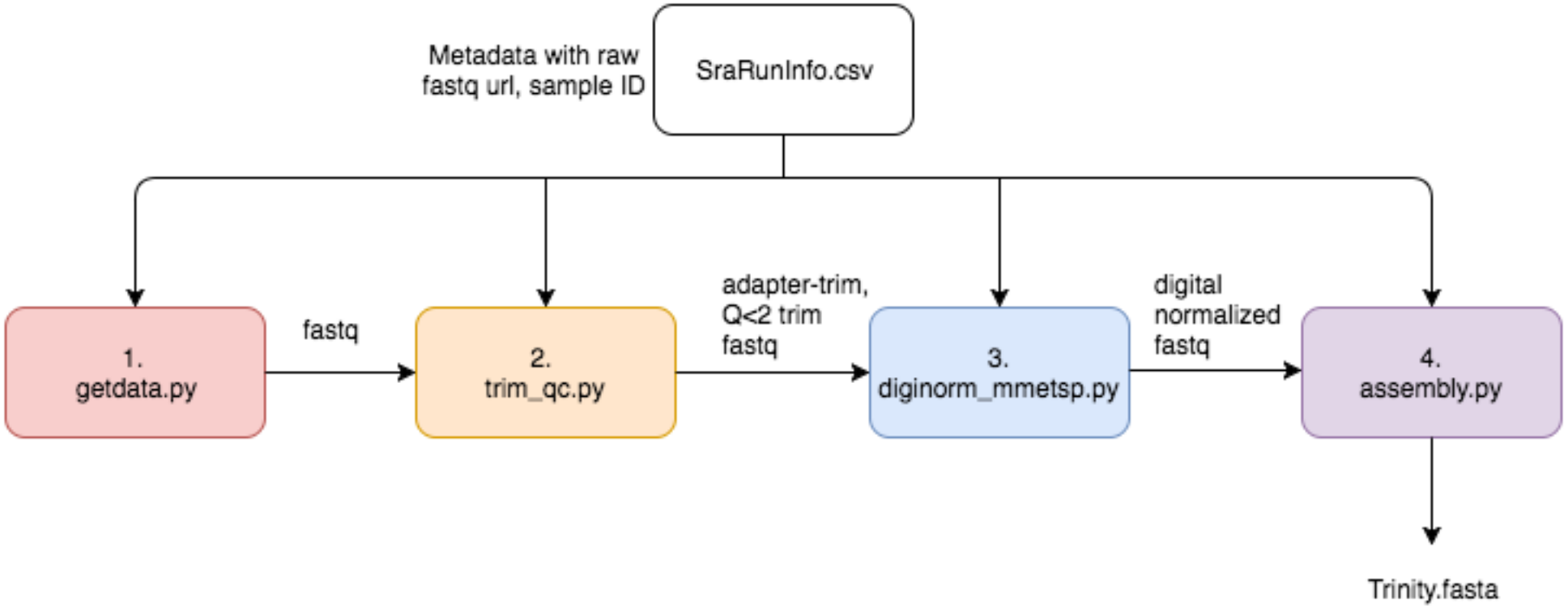
Prepare for running Python:
1. If not already installed, install virtualenv
   * ```pip install virtualenv```
1. Create a project folder:
   * ```virtualenv -p <location of Python3 install directory> <name of project>```
1. Activate the virtual environment:
   * ```source <name of project>/bin/activate ```
1. Install libraries:
   * ```pip install -r requirements.txt ```
```

```
### Run the analysis

1. Run ```survey_2014_analysis.py```:
    * ```python survey_2014_analysis.py```
1. This summarises the reseponses to the survey, by groups the answers to each question and counting how many times each one
occurs. It stores the results in a series csv files (one per question) in the ```output/summary_csvs/``` directory.
1. Run ```comparison_new_old_results.py```:
    * ```python comparison_new_old_results.py```
1. This takes the results of the summary files produced by the ```survey_2014_analysis.py``` and compares them against the
results from the original analysis. It stores the results of that analysis in a series of csv files (one per question) in the
```output/comparison_summary_csvs/``` directory.
```

From University of Arizona example README
```

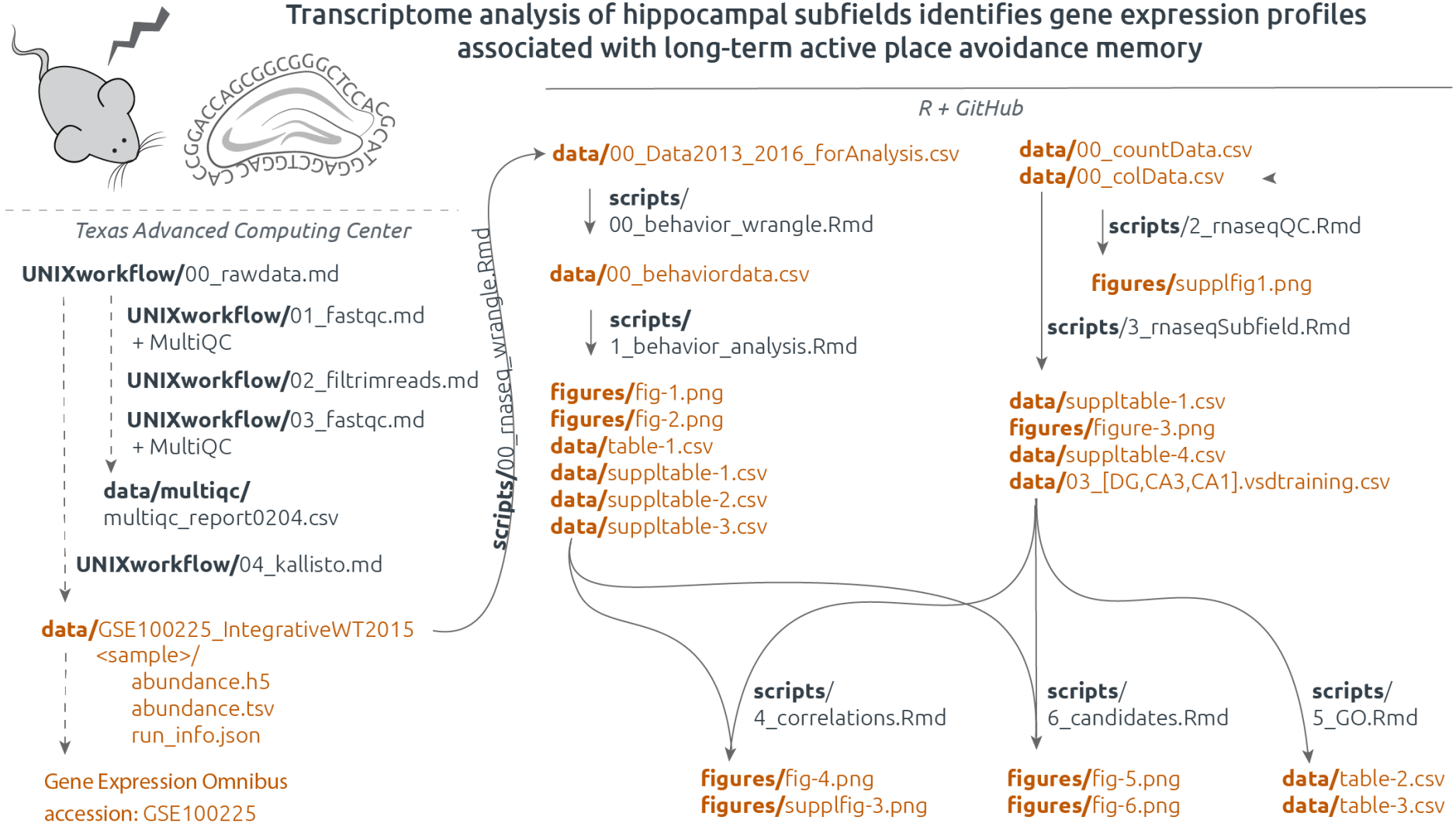
Workflow diagrams can add a visual element



Github: @dib-lab

UCDAVIS
DataLab
Data Science and Informatics

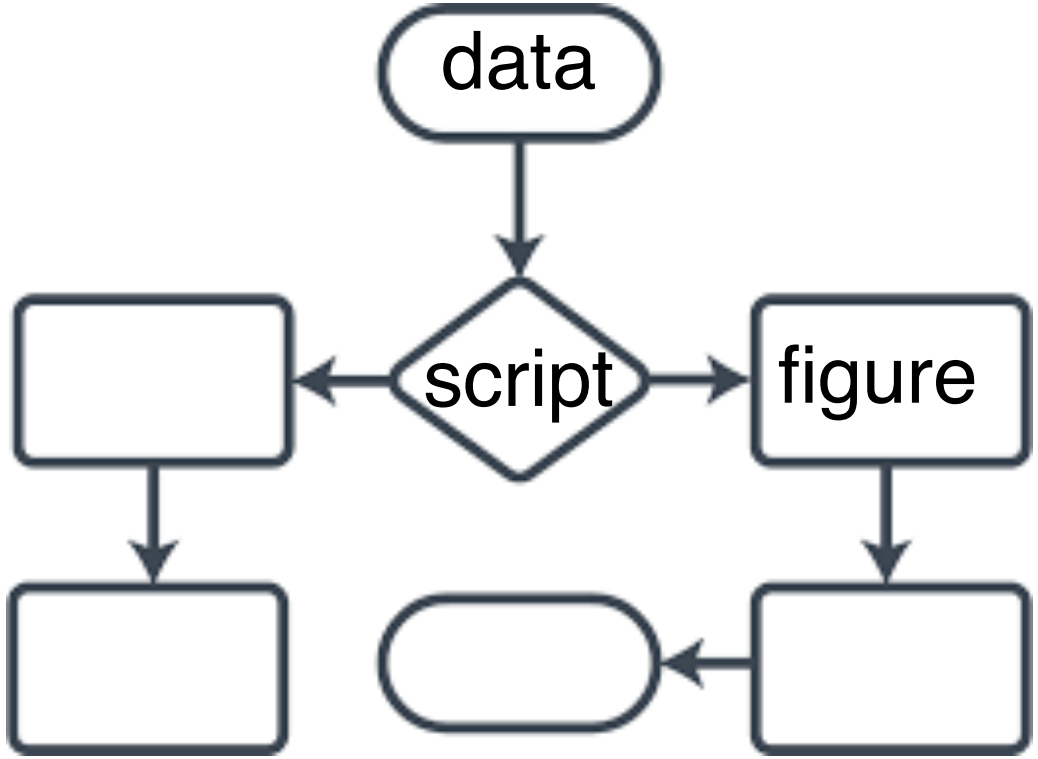
Workflow diagram examples



Github: @raynamharris

UCDAVIS DataLab
Data Science and Informatics

Workflow diagram tips



Metadata can be structured and standardized

- Funding agencies, data depositories, and databases require more structured metadata
- This **structured metadata** is usually produced in specific language formats, like XML or EML
- Fields have created specific **metadata standards** for their discipline

An example of XML:

```
<?xml version="1.0" encoding="UTF-8"?>
<xs:schema xmlns:xs="http://www.w3.org/2001/XMLSchema"
    xmlns="http://champ-project.org/stdMethod"
    xmlns:champ="http://champ-project.org/champ"
    xmlns:dcterms="http://purl.org/dc/terms/"
    elementFormDefault="qualified" attributeFormDefault="unqualified"
    targetNamespace="http://champ-project.org/stdMethod" version="1.0" xml:lang="en">

    <xs:import namespace="http://champ-project.org/champ" schemaLocation="http://champ-project.org/images/schema/champ.xsd"/>
    <xs:import namespace="http://purl.org/dc/terms/" schemaLocation="http://dublincore.org/schemas/xmls/qdc/2008/02/11/dcterms.xsd"/>
    <xs:element name="summary" substitutionGroup="dcterms:abstract"/>
    <xs:element name="performanceData" substitutionGroup="champ:exampleData"/>
    <xs:element name="citation" substitutionGroup="dcterms:bibliographicCitation"/>

    <xs:element name="stdMethod" type="methodType"/>

    <xs:complexType name="methodType">
        <xs:sequence>
            <xs:element ref="dcterms:title" minOccurs="1" maxOccurs="1"/>
            <xs:element ref="summary" minOccurs="1" maxOccurs="1"/>
            <xs:element ref="champ:analyte" minOccurs="1" maxOccurs="unbounded"/>
            <xs:element ref="champ:matrix" minOccurs="1" maxOccurs="unbounded"/>
            <xs:element ref="champ:analyticalFocus" minOccurs="0" maxOccurs="1"/>
            <xs:element ref="champ:applicationArea" minOccurs="0" maxOccurs="1"/>
            <xs:element ref="champ:analysisType" minOccurs="0" maxOccurs="1"/>
            <xs:element ref="champ:analysisFormat" minOccurs="0" maxOccurs="1"/>
            <xs:element ref="champ:analysisInterferent" minOccurs="0" maxOccurs="unbounded"/>
            <xs:element ref="champ:analysisLocale" minOccurs="0" maxOccurs="1"/>
            <xs:element ref="champ:instrument" minOccurs="1" maxOccurs="unbounded"/>
            <xs:element ref="champ:setting" minOccurs="0" maxOccurs="unbounded"/>
            <xs:element ref="champ:chemical" minOccurs="0" maxOccurs="unbounded"/>
            <xs:element ref="champ:solution" minOccurs="0" maxOccurs="unbounded"/>
            <xs:element ref="champ:samplingProtocol" minOccurs="0" maxOccurs="1"/>
            <xs:element ref="champ:storageCondition" minOccurs="0" maxOccurs="unbounded"/>
            <xs:element ref="champ:analysisTimeframe" minOccurs="0" maxOccurs="1"/>
            <xs:element ref="champ:procedure" minOccurs="1" maxOccurs="1"/>
            <xs:element ref="champ:quality" minOccurs="0" maxOccurs="1"/>
            <xs:element ref="performanceData" minOccurs="0" maxOccurs="1"/>
            <xs:element ref="citation" minOccurs="0" maxOccurs="unbounded"/>
        </xs:sequence>
        <xs:attribute name="id" type="xs:string"/>
    </xs:complexType>

</xs:schema>
```

UCDAVIS DataLab
Data Science and Informatics

Metadata can be structured and standardized

- Field-specific terminology and data codes can be found with these standards as well
- Ask the UC Davis librarians for information about more standardized metadata

Let’s get documenting!

Begin documenting your own data or project using templates

Join Breakout Room
(click “Join”)

Watch a demo building documentation for a dataset

Stay in Main Room
(click “Not Now”)

UCDAVIS
DataLab
Data Science and Informatics

UCDAVIS

DataLab

Data Science and Informatics