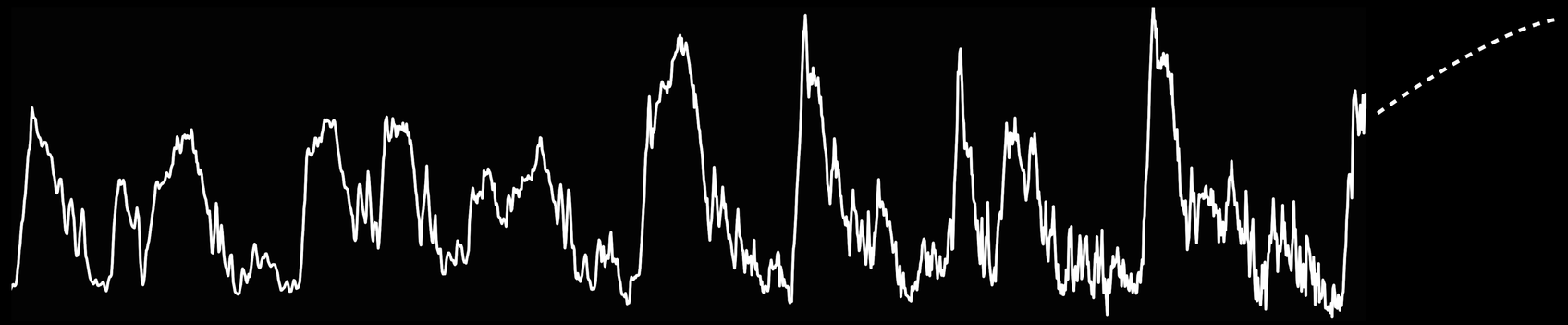
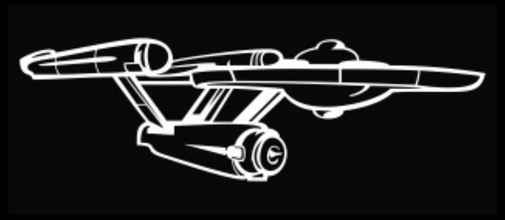


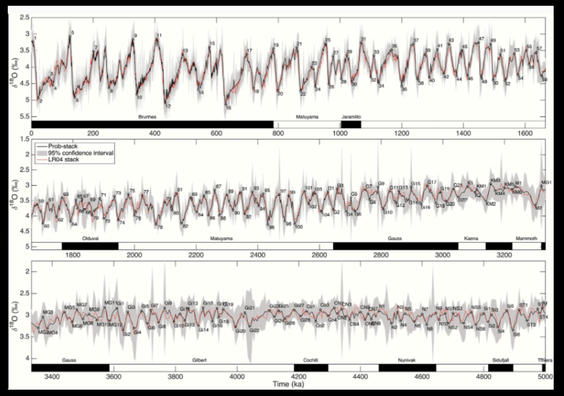


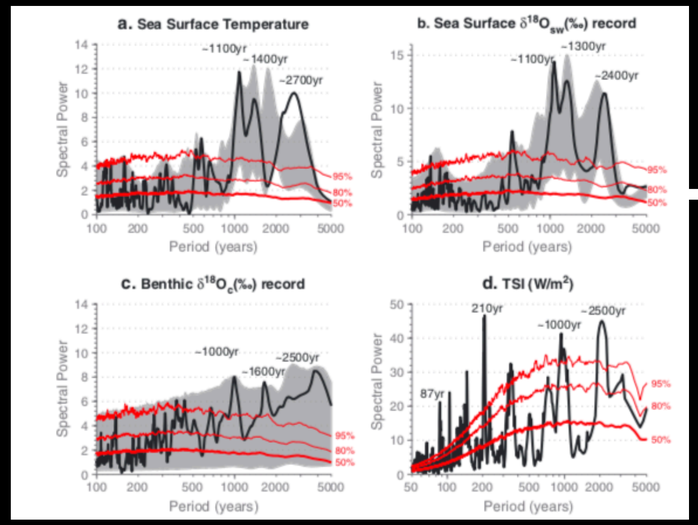
Research Data Management

Deborah Khider

USC Viterbi
School of Engineering
Information Sciences Institute

My Data Journey

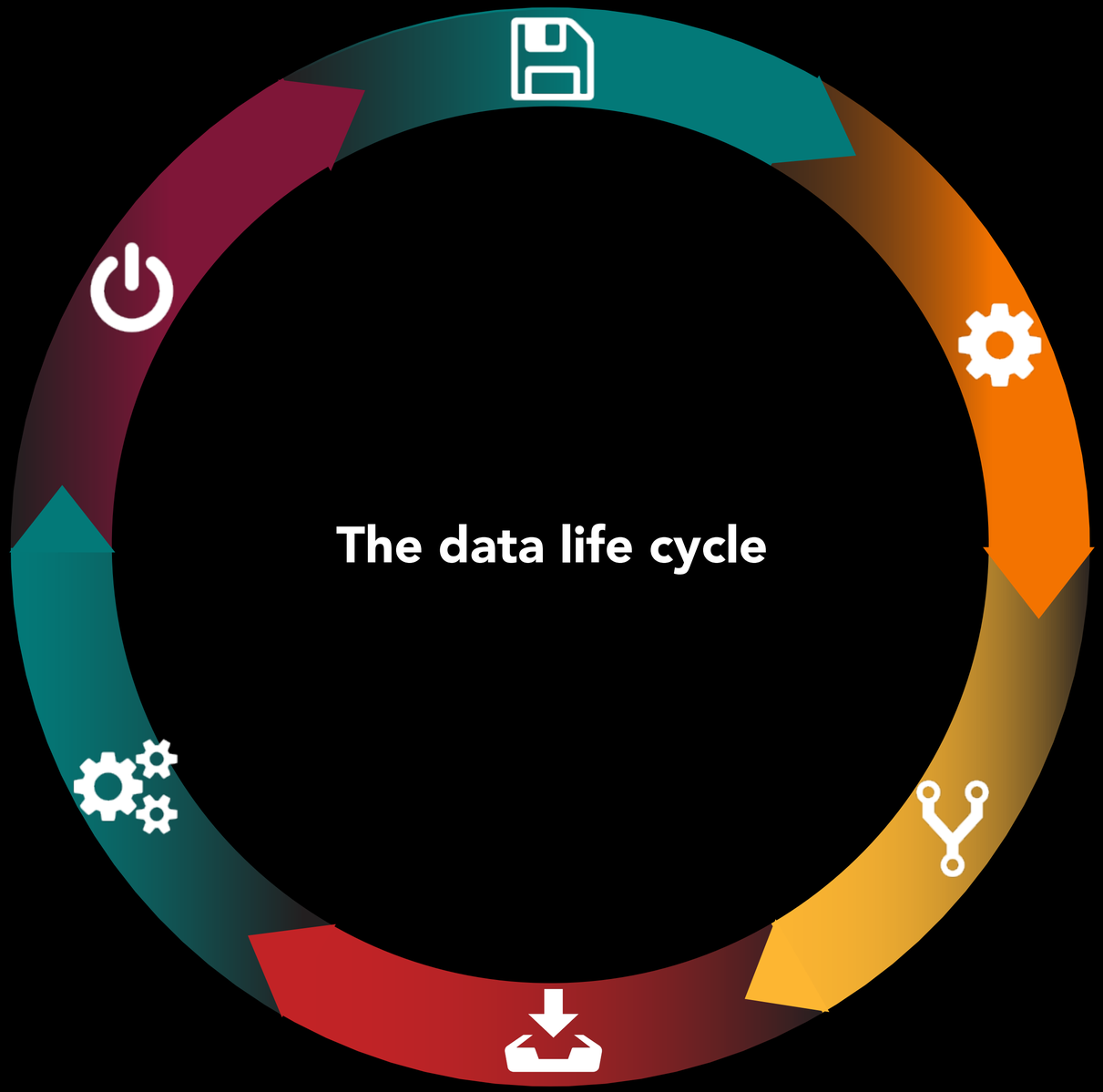


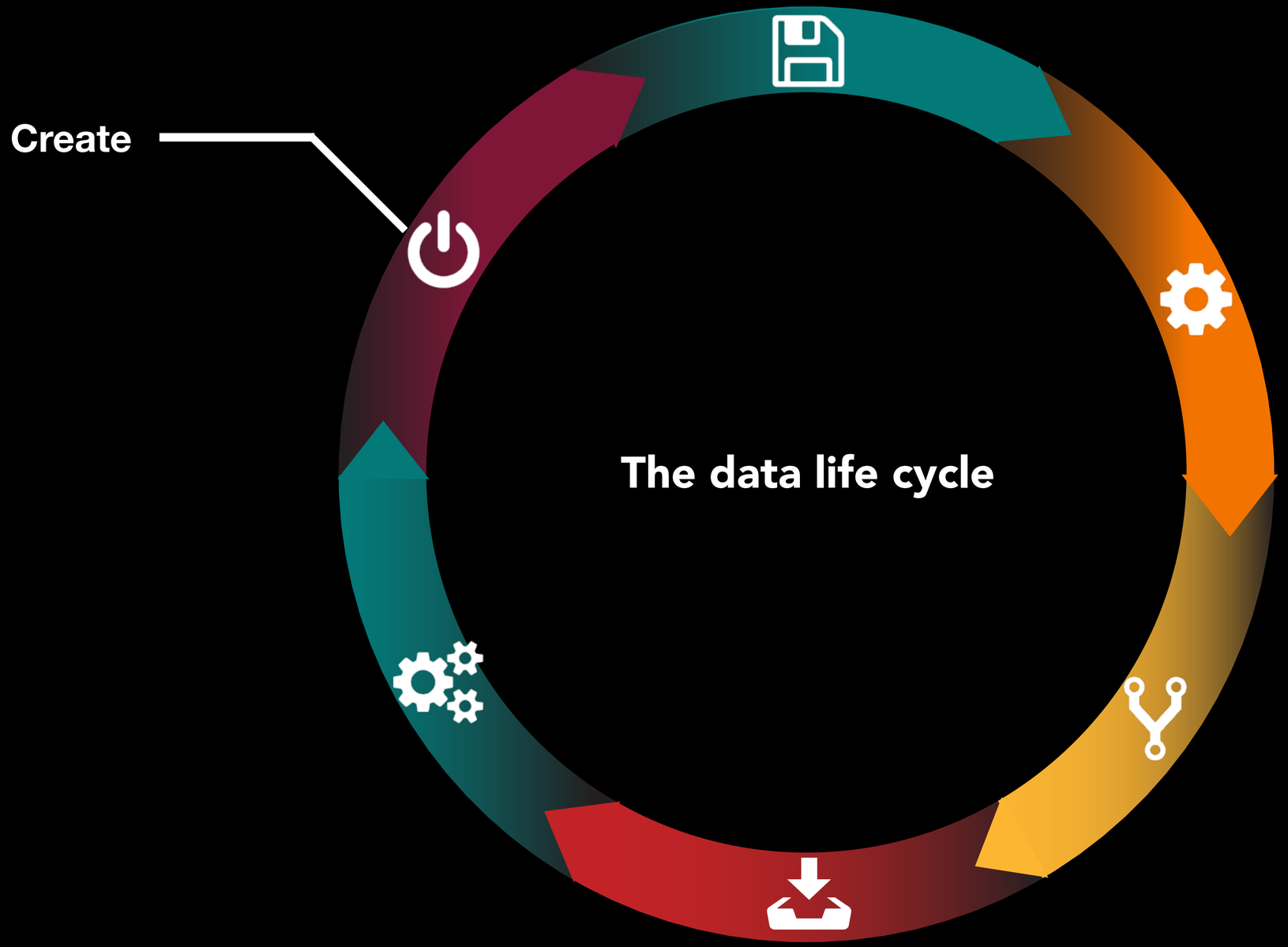


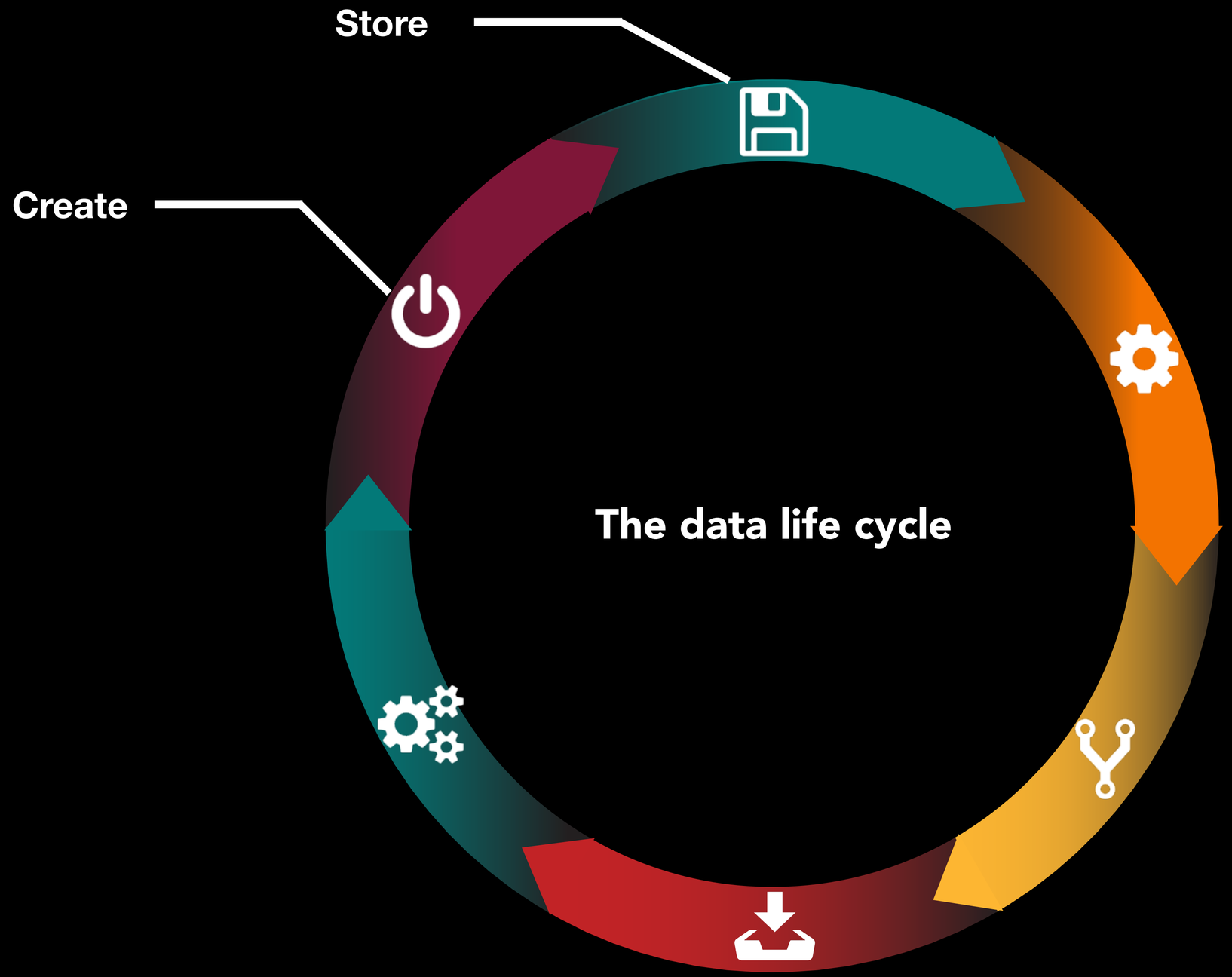
LinkedEarth

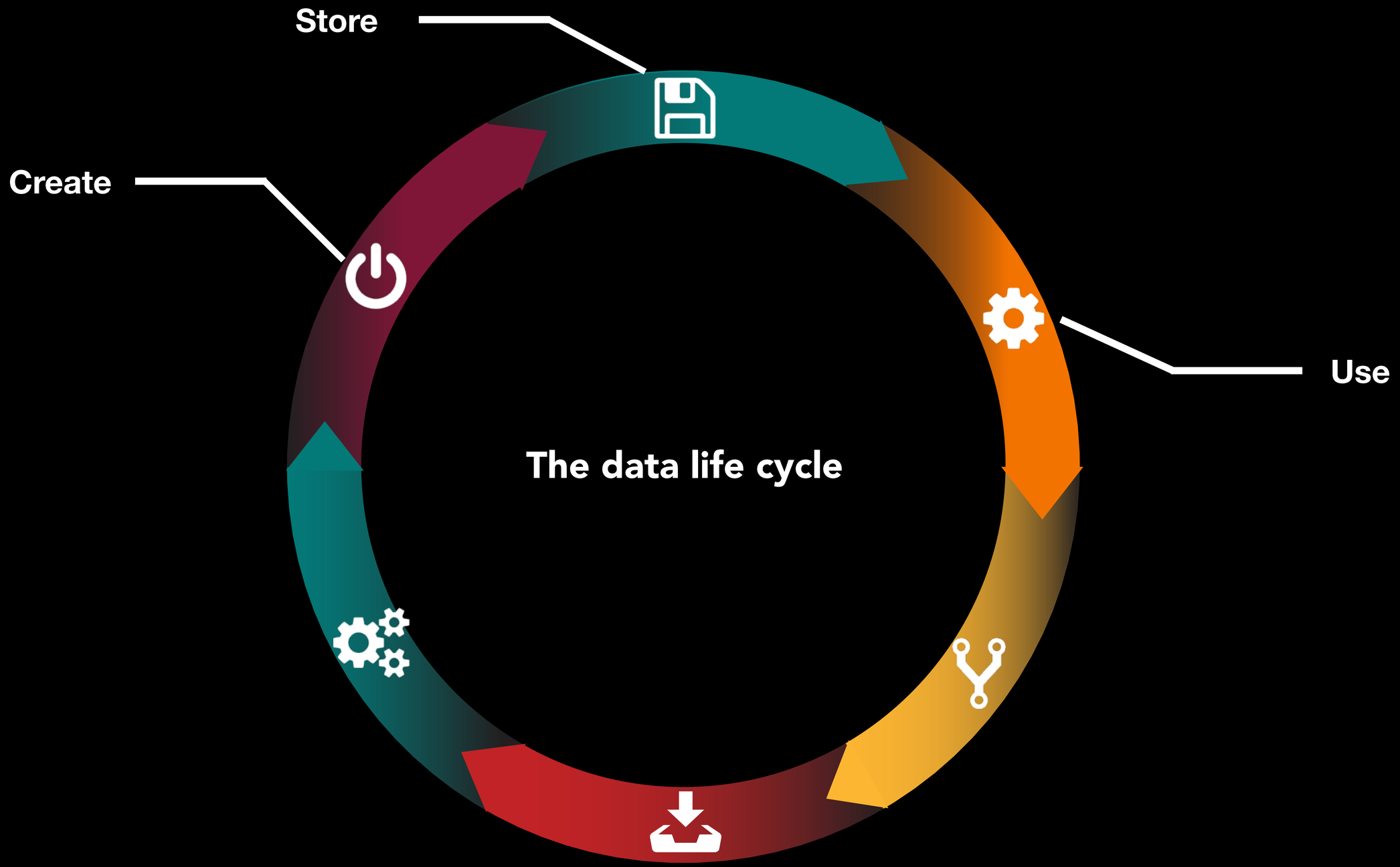
mint
model integration

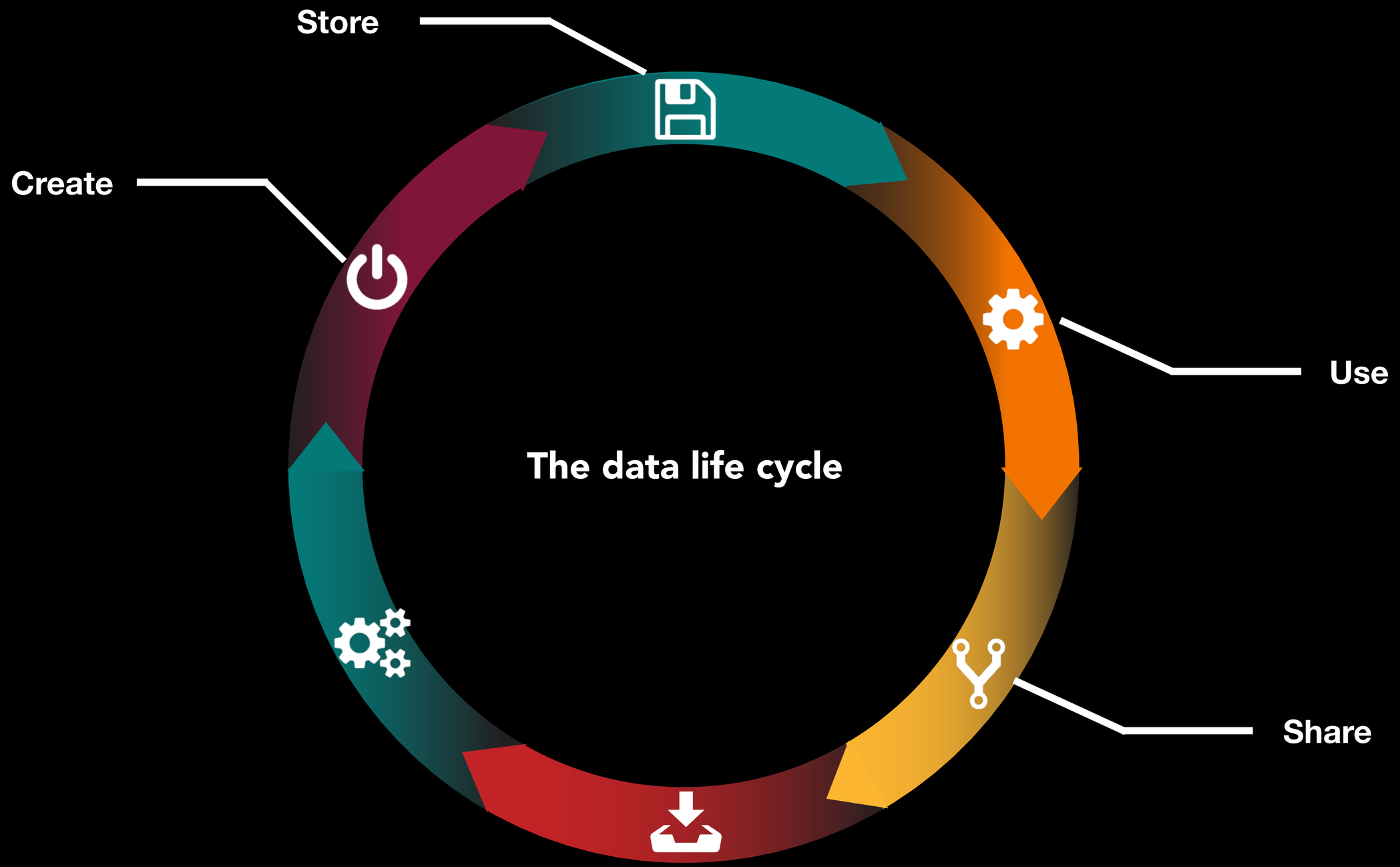
Model Integration through Knowledge-Rich Data and Process Composition

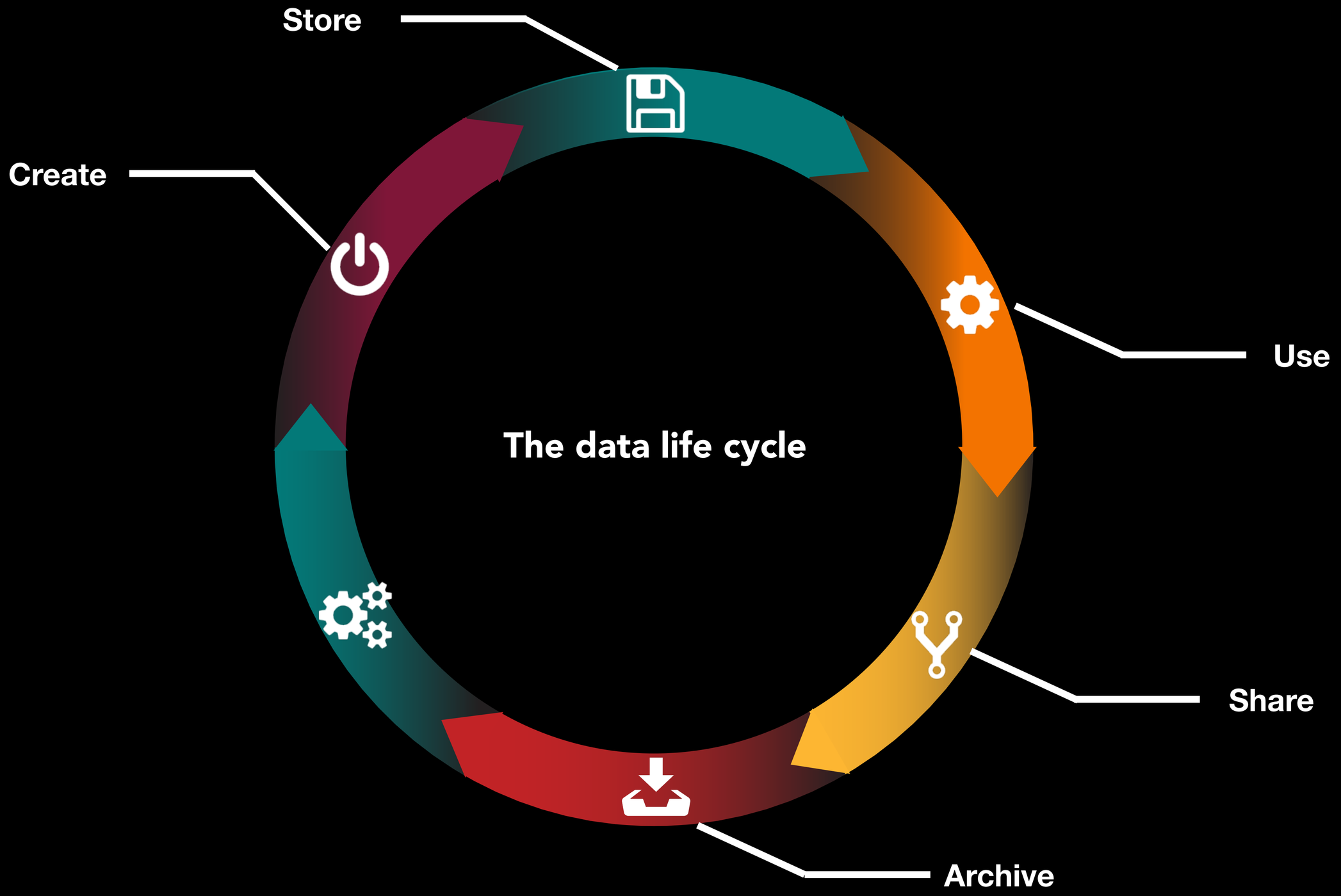

The data life cycle

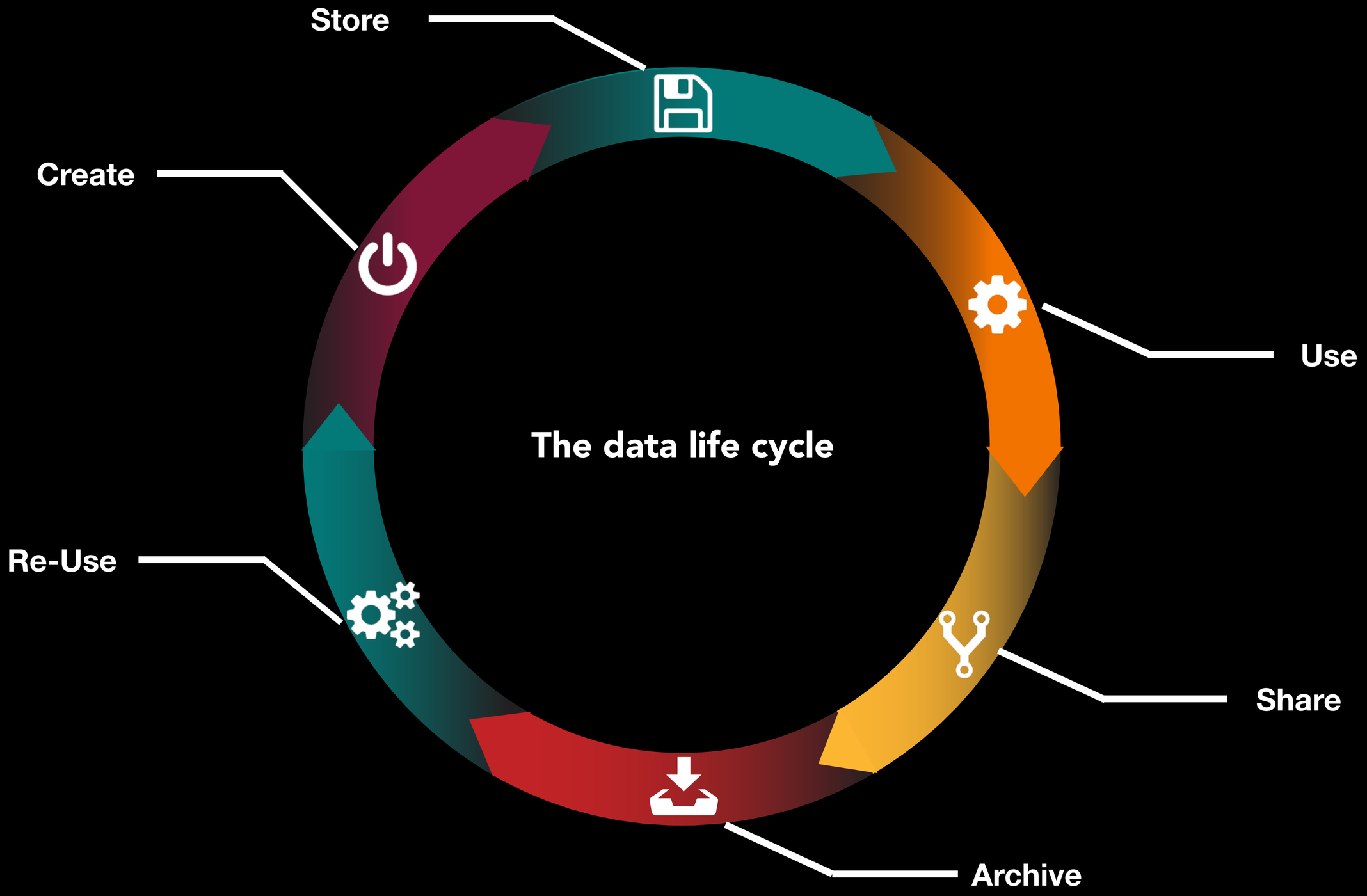

Create
The data life cycle


Store
Create
The data life cycle


Store
Create
Use
The data life cycle


Store
Create
Use
Share
The data life cycle


Store
Create
Use
The data life cycle
Share
Archive


Store
Create
Use
Re-Use
Share
Archive
The data life cycle

Create

Create

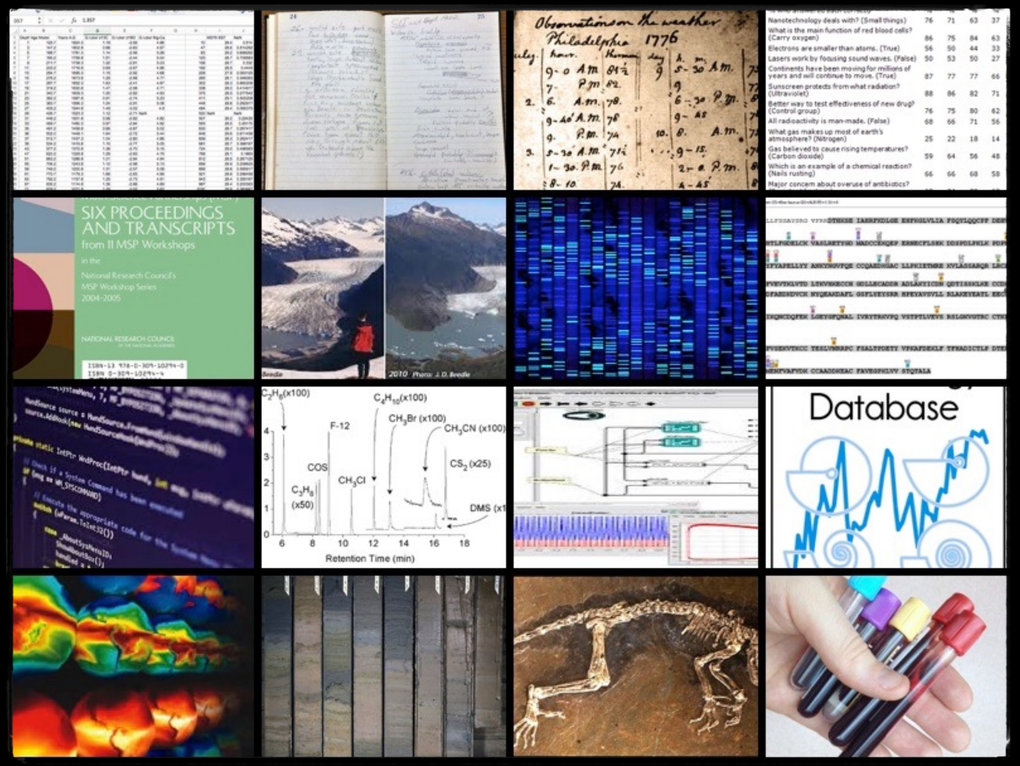
DATA:

"...THE RECORDED FACTUAL MATERIAL COMMONLY ACCEPTED IN THE SCIENTIFIC COMMUNITY AS NECESSARY TO VALIDATE RESEARCH FINDINGS."

Create

DATA:

"…THE RECORDED FACTUAL MATERIAL COMMONLY ACCEPTED IN THE SCIENTIFIC COMMUNITY AS NECESSARY TO **VALIDATE** RESEARCH FINDINGS."



Create

DATA:

"...THE RECORDED FACTUAL MATERIAL COMMONLY ACCEPTED IN THE SCIENTIFIC COMMUNITY AS NECESSARY TO **VALIDATE** RESEARCH FINDINGS."

METADATA:

METADATA IS INFORMATION ABOUT THE DATA THAT PROVIDES CONTEXT KEY TO **UNDERSTAND** WHAT THE DATA REPRESENTS.

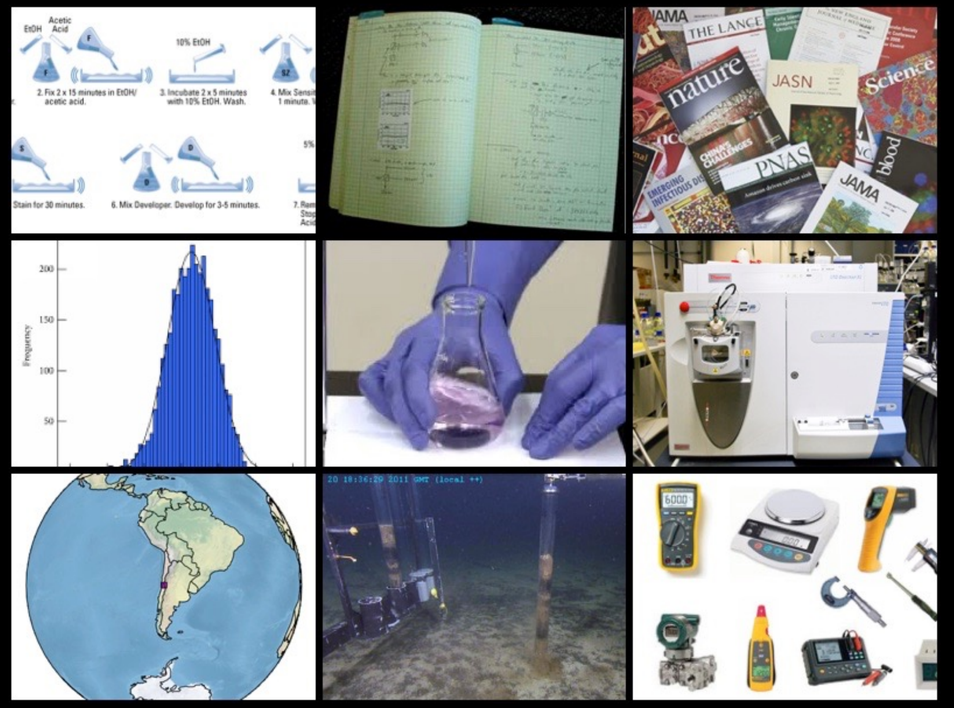
Create

DATA:

"...THE RECORDED FACTUAL MATERIAL COMMONLY ACCEPTED IN THE SCIENTIFIC COMMUNITY AS NECESSARY TO **VALIDATE** RESEARCH FINDINGS."

METADATA:

METADATA IS INFORMATION ABOUT THE DATA THAT PROVIDES CONTEXT KEY TO **UNDERSTAND** WHAT THE DATA REPRESENTS.



Create

DATA:

"...THE RECORDED FACTUAL MATERIAL COMMONLY ACCEPTED IN THE SCIENTIFIC COMMUNITY AS NECESSARY TO **VALIDATE** RESEARCH FINDINGS."

METADATA:

METADATA IS INFORMATION ABOUT THE DATA THAT PROVIDES CONTEXT KEY TO **UNDERSTAND** WHAT THE DATA REPRESENTS.

DATA MANAGEMENT:

ACTIONS THAT CONTRIBUTE TO EFFECTIVE **STORAGE**, **PRESERVATION**, AND **REUSE** OF DATA AND METADATA THROUGHOUT THE RESEARCH LIFECYCLE.

Why do you need to know about data management?

- Scientists are changing

Credit: http://www.scientificpaperofthefuture.org

Why do you need to know about data management?

- Scientists are changing

Open Data

Open Access

Open Publications

Open Source

Credit: http://www.scientificpaperofthefuture.org

Why do you need to know about data management?

- Scientists are changing
- Publishers are changing

AGU ADVANCING EARTH AND SPACE SCIENCE

JOIN GIVE LOGIN RENEW

WHAT IS NEEDED?

AGU requires that the underlying data needed to understand, evaluate, and build upon the reported research be available at the time of peer review and publication. Additionally, authors should make available software that has a significant impact on the research. This entails:

1. Depositing the data and software in a trusted repository, as appropriate, and preferably with a DOI
2. Including an Availability Statement as a separate paragraph in the Open Research section explaining to the reader where and how to access the data and software
3. And including citation(s) to the deposited data and software, in the Reference Section.

Click on the headings below for detailed information on:

- Models & Simulations
- Journal-Specific Data Guidance
- International Geo Sample Numbers

WHAT DATA NEEDS TO BE AVAILABLE?

Primary and processed data used for your research should be preserved and made available. Generally, the underlying data are considered to be the types of data usually preserved in domain repositories for each discipline. These may include raw data, but are usually the processed or refined data that support and lead to the described results and allow other readers to assess your conclusions and build off your work.

In your paper, cite these data, as well as any data you used from other sources, and include information about access to the data in the availability statement. For **model or simulation data**, follow journal specific guidance on prioritizing preserved output; in general, availability of software is most important.

Very large data (greater than 1 terabyte or TB) can be a challenge to preserve as there often fees and additional resources required. One option to consider, institutions often offer solutions for data preservation and compliance. Again, refer to the journal specific guidance for more information or email DataHelp@agu.org.

https://www.agu.org/Publish-with-AGU/Publish/Author-Resources/Data-and-Software-for-Authors

Why do you need to know about data management?

- Scientists are changing
- Publishers are changing

AGU ADVANCING EARTH AND SPACE SCIENCE

JOIN GIVE LOGIN RENEW

WHAT IS NEEDED?

AGU requires that the underlying data needed to understand, evaluate, and build upon the reported research be available at the time of peer review and publication. Additionally, authors should make available software that has a significant impact on the research. This entails:

1. Depositing the data and software in a trusted repository, as appropriate, and preferably with a DOI
2. Including an Availability Statement as a separate paragraph in the Open Research section explaining to the reader where and how to access the data and software
3. And including citation(s) to the deposited data and software, in the Reference Section.

Click on the headings below for detailed information on:

- Models & Simulations
- Journal-Specific Data Guidance
- International Geo Sample Numbers

WHAT DATA NEEDS TO BE AVAILABLE?

Primary and processed data used for your research should be preserved and made available. Generally, the underlying data are considered to be the types of data usually preserved in domain repositories for each discipline. These may include raw data, but are usually the processed or refined data that support and lead to the described results and allow other readers to assess your conclusions and build off your work.

In your paper, cite these data, as well as any data you used from other sources, and include information about access to the data in the availability statement. For **model or simulation data**, follow journal specific guidance on prioritizing preserved output; in general, availability of software is most important.

Very large data (greater than 1 terabyte or TB) can be a challenge to preserve as there often fees and additional resources required. One option to consider, institutions often offer solutions for data preservation and compliance. Again, refer to the journal specific guidance for more information or email DataHelp@agu.org.

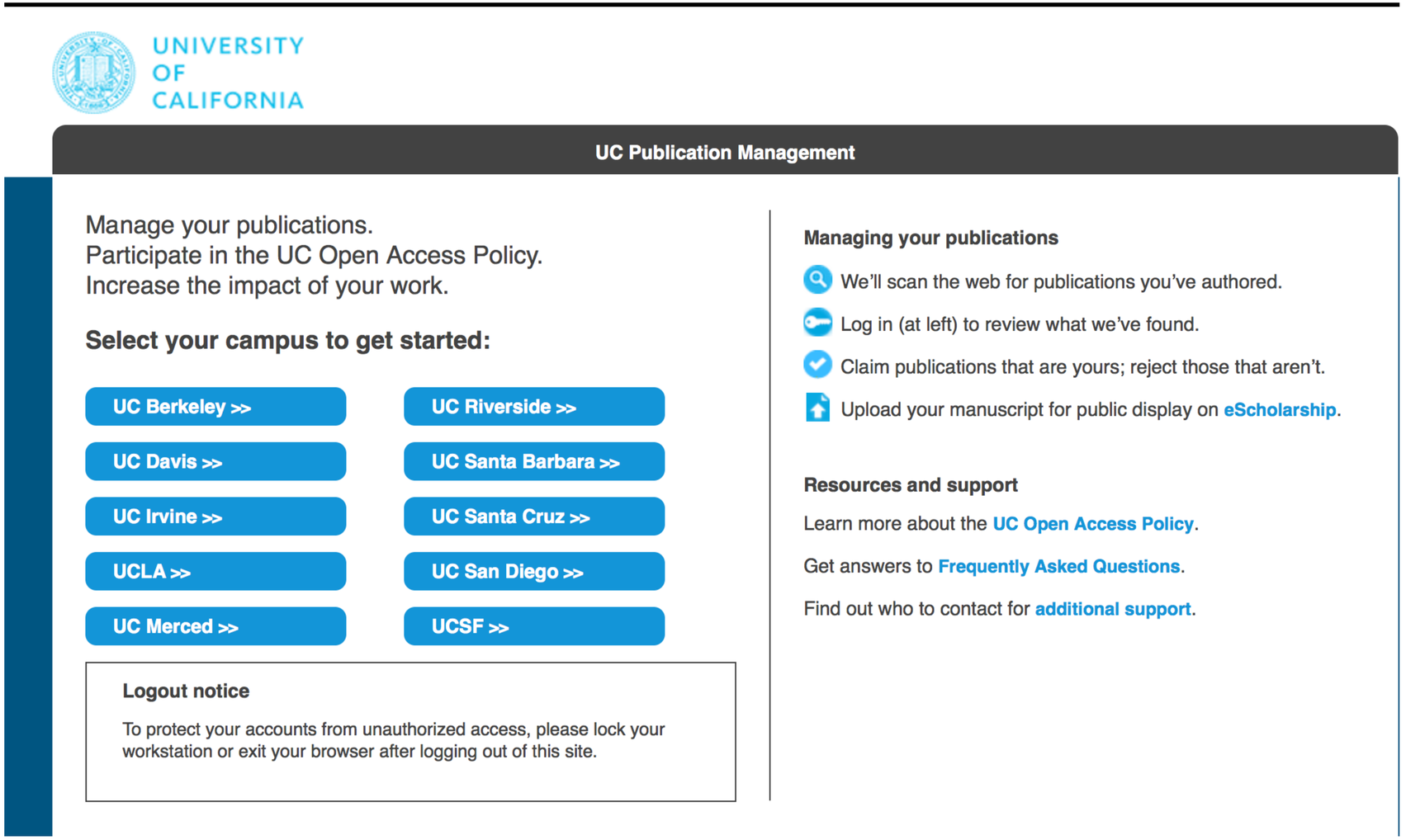
https://www.agu.org/Publish-with-AGU/Publish/Author-Resources/Data-and-Software-for-Authors

Why do you need to know about data management?

- Scientists are changing
- Publishers are changing
- Funders are changing

Why do you need to know about data management?

- Scientists are changing
- Publishers are changing
- Funders are changing
- Universities are changing



Why do you need to know about data management?

- Sharing your data is not only required but it also helps you!

Sharing Detailed Research Data Is Associated with Increased Citation Rate

Heather A. Piwowar, Roger S. Day, Douglas B. Fridsma

Published: March 21, 2007 • DOI: 10.1371/journal.pone.0000308 • Featured in PLOS Collections

Principal Findings

We examined the citation history of 85 cancer microarray clinical trial publications with respect to the availability of their data. The 48% of trials with publicly available microarray data received 85% of the aggregate citations. Publicly available data was significantly ($p = 0.006$) associated with a 69% increase in citations, independently of journal impact factor, date of publication, and author country of origin using linear regression.

Modern Scientific Articles

Traditional Published Articles

Text:

Narrative of method, the data is in tables, figures/plots, the software used is mentioned

Modern Published Articles

Text:

Narrative of method, the data is in tables, figures/plots, the software used is mentioned

Data:

Supplementary Materials, pointers to data repositories

Credit: http://www.scientificpaperofthefuture.org

Modern Scientific Articles

Traditional Published Articles

Text:

Narrative of method, the data is in tables, figures/plots, the software used is mentioned

Modern Published Articles

Text:

Narrative of method, the data is in tables, figures/plots, the software used is mentioned

Data:

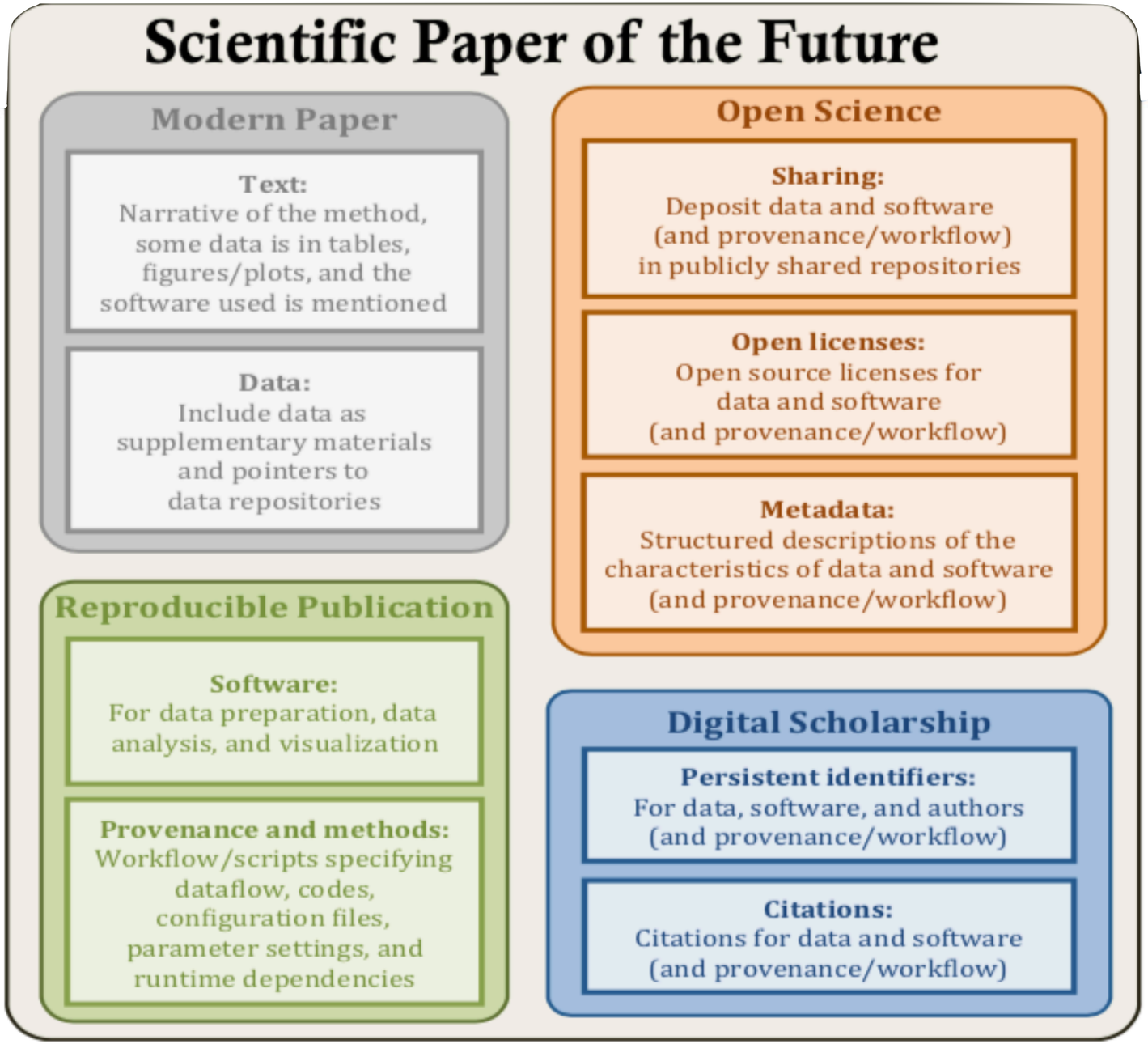
Supplementary Materials, pointers to data repositories

NOT Published, loosely recorded

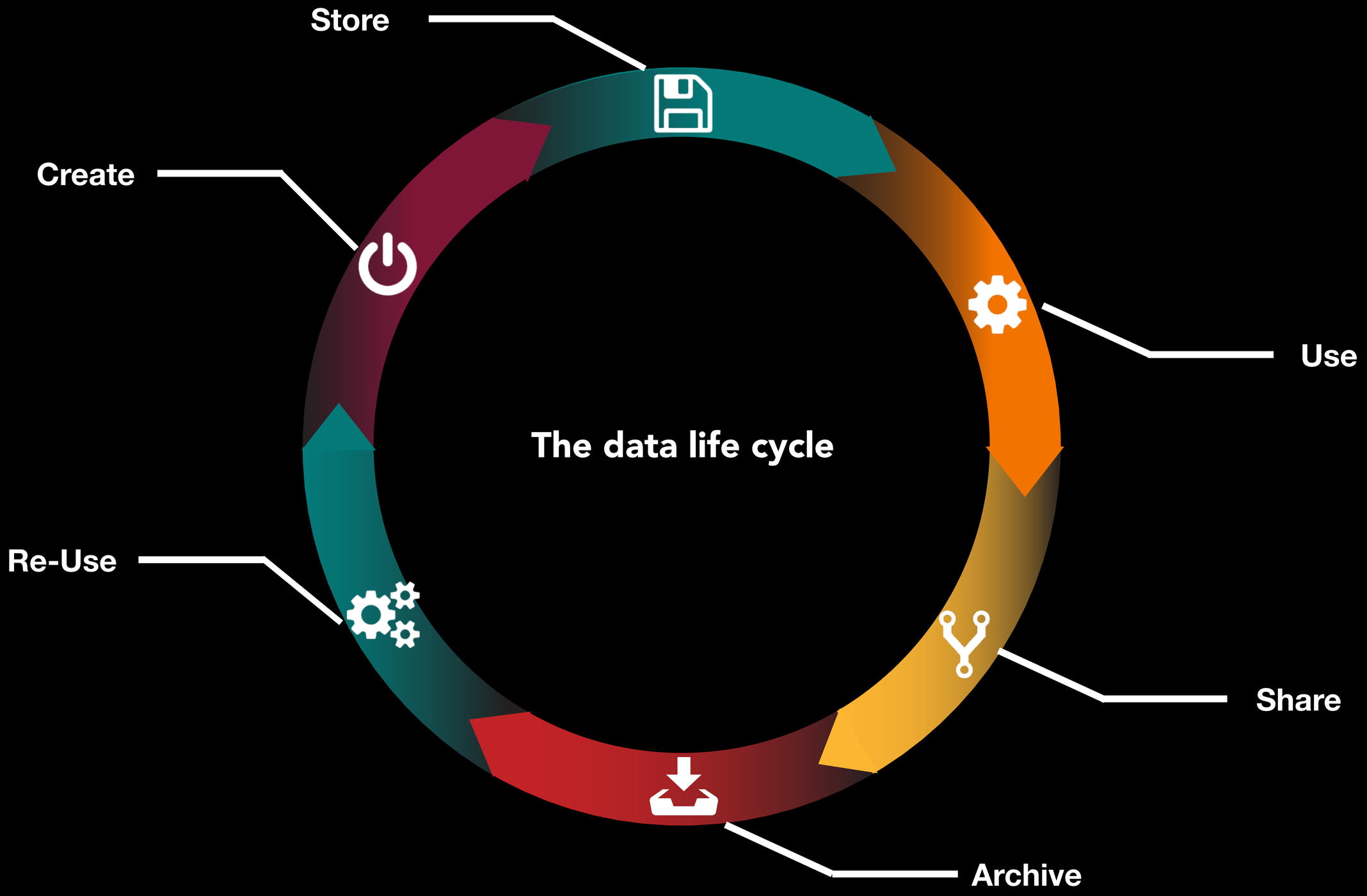
Software:

Scripted codes, documentation, manual steps

Credit: http://www.scientificpaperofthefuture.org



Credit: http://www.scientificpaperofthefuture.org


Store
Create
Use
Re-Use
Share
Archive
The data life cycle

Store

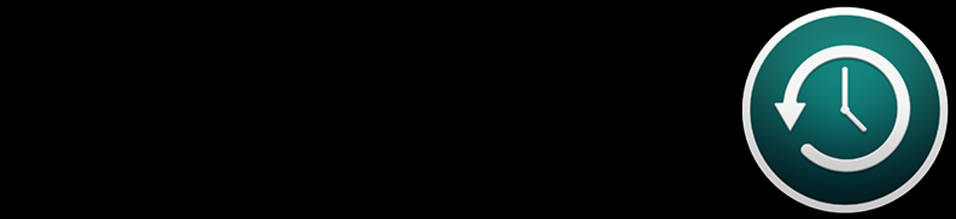
Store

AND

Store

AND

Store



AND

Store

AND

Store

AND

Google Drive

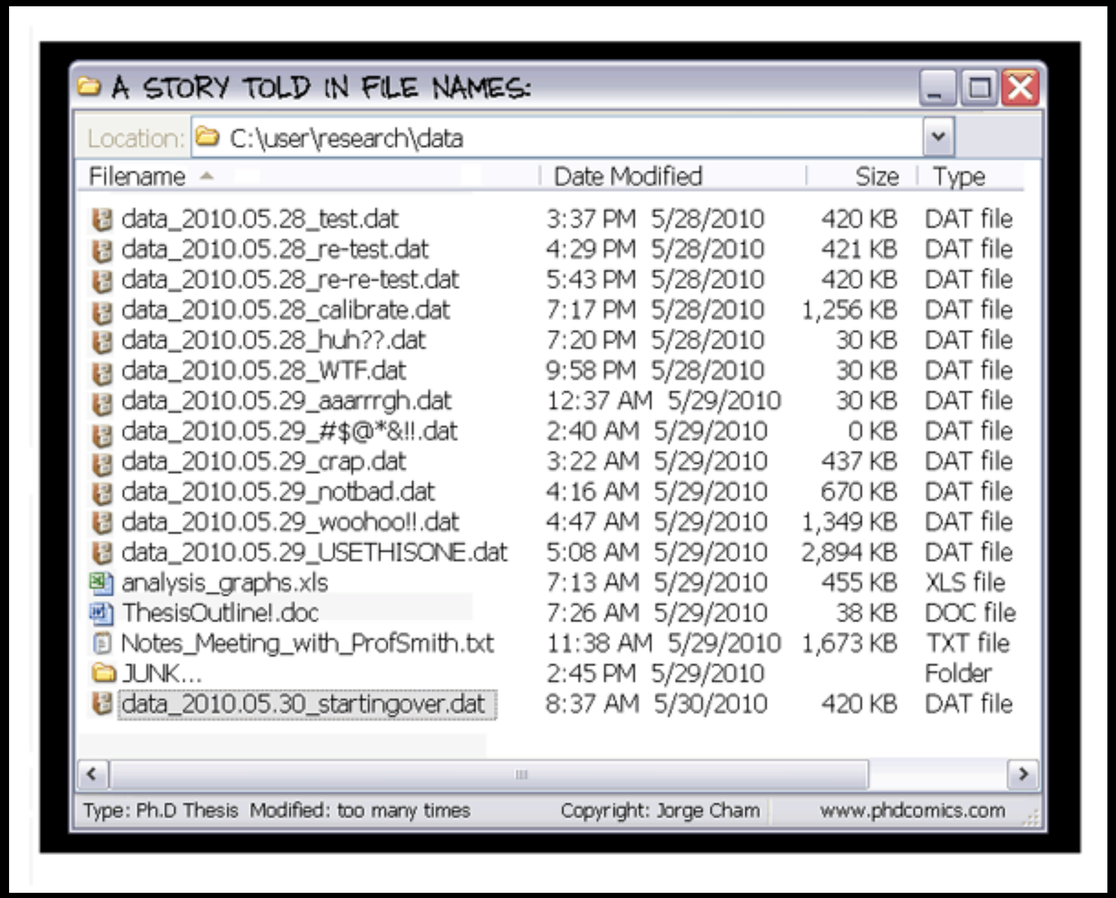
Dropbox

iCloud

Keep in mind:

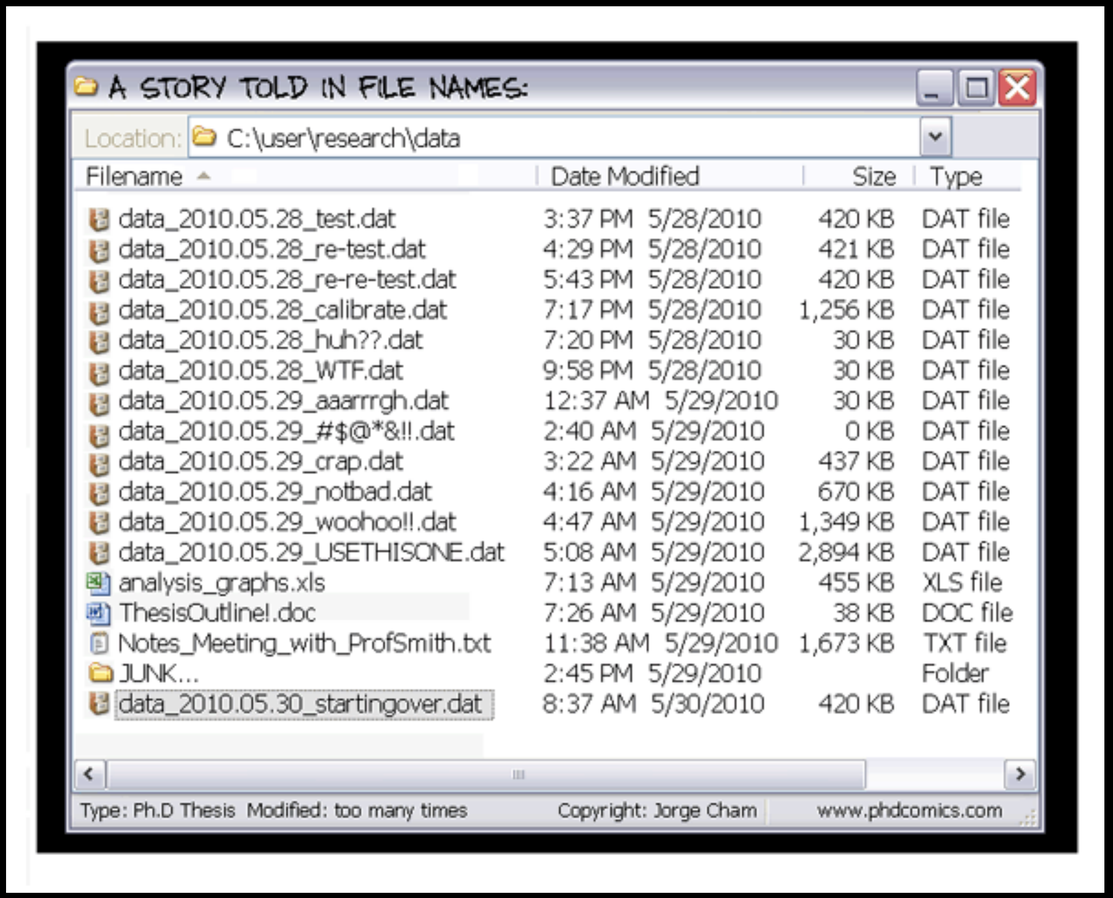
1. Some data backup is better than none
2. Automated backups are better than manual
3. Your data is only as safe as the last backup

File Name and Organization



PhDcomics

File Name and Organization



PhDcomics

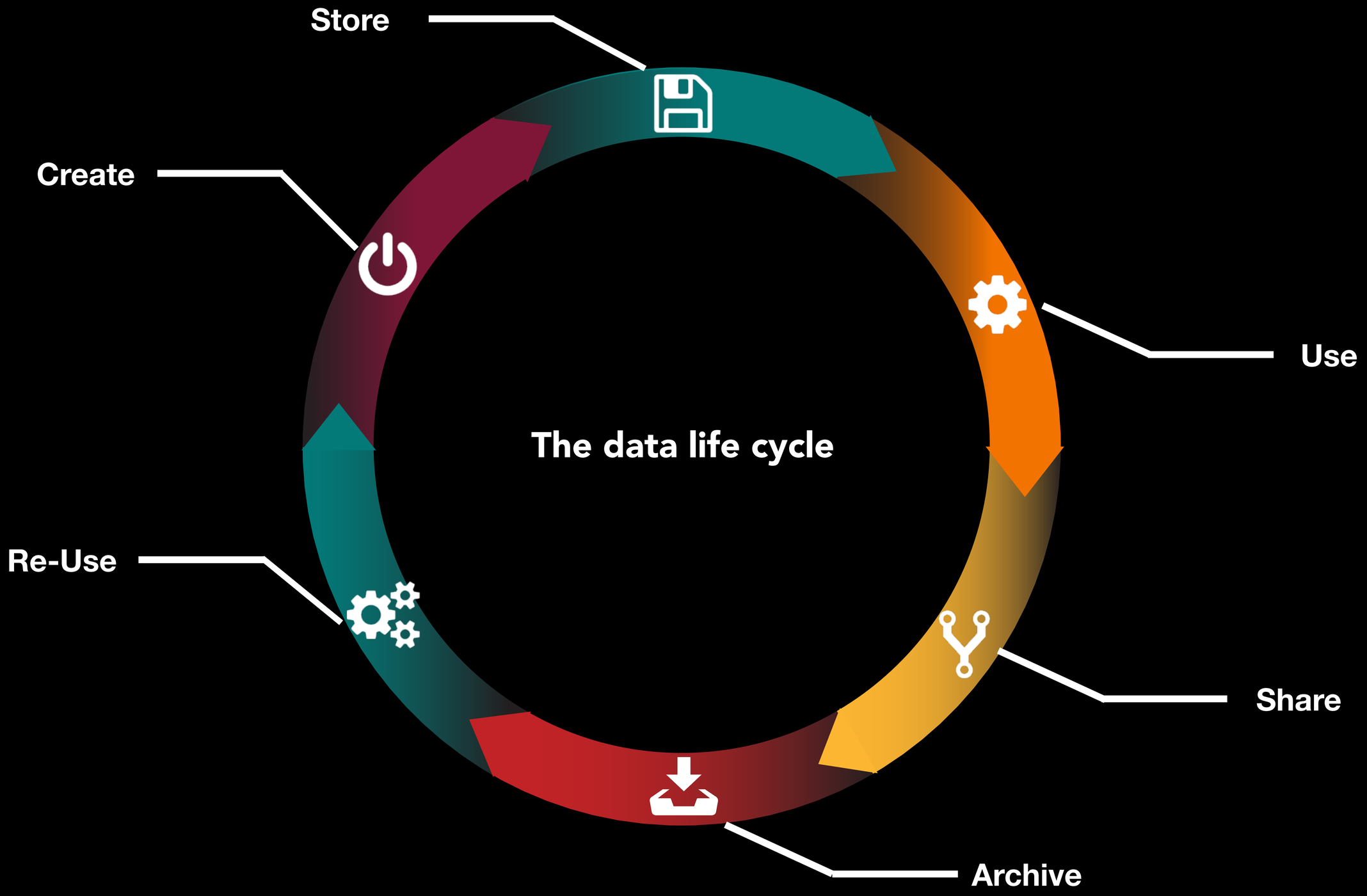
File Name and Organization

Project_Date_Description

File Name and Organization

KEEP A FILE ABOUT YOUR FILES

- **ReadMe**: Description of what the files/folders contain

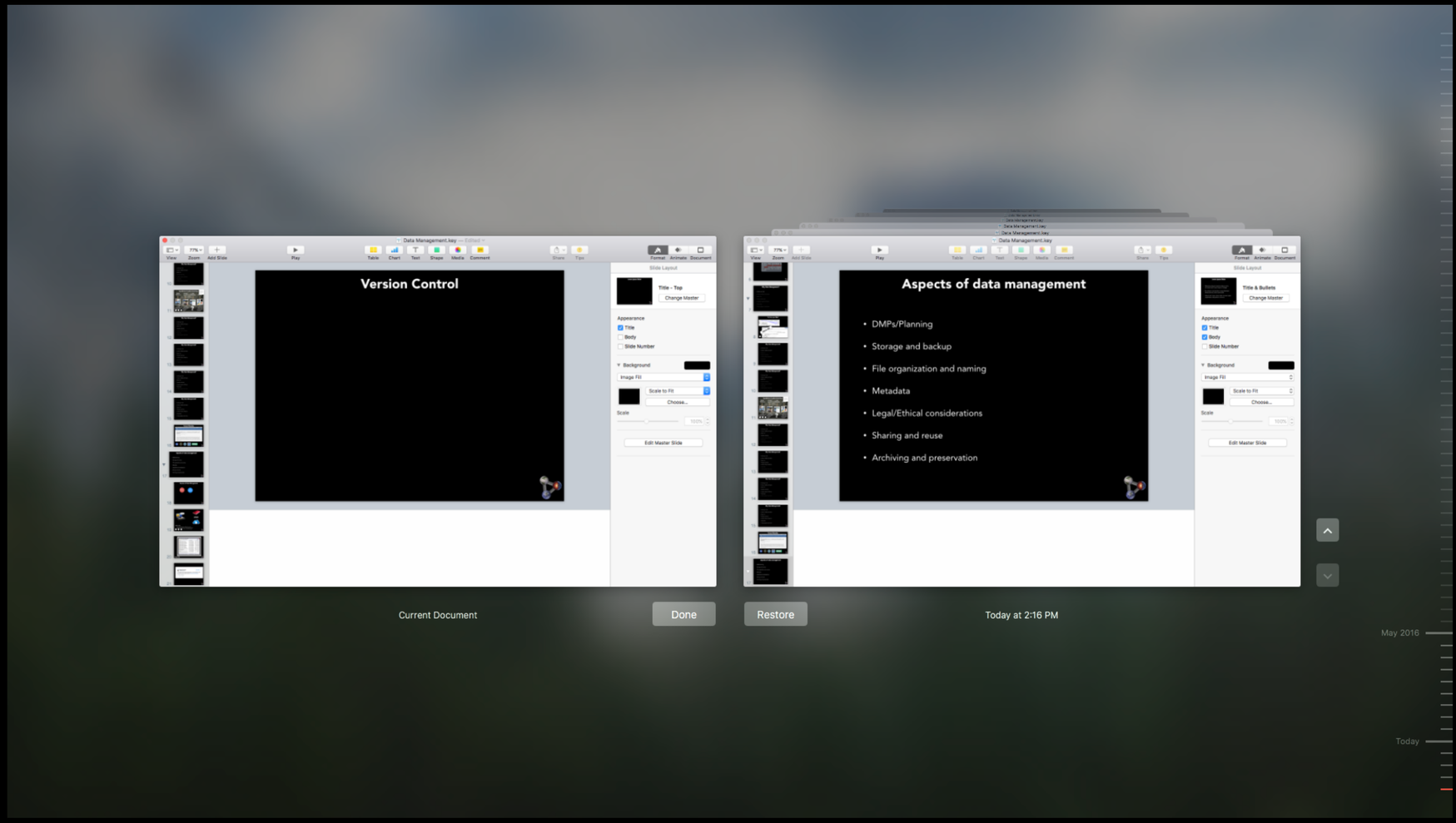

Store
Create
Use
Re-Use
Share
Archive
The data life cycle

KEEP YOUR RAW DATA RAW!

- Always keep the original data as raw as possible. Create new versions of dataset if you perform any data cleaning.
- Even more important when calibrating datasets or applying normalization, filters,…

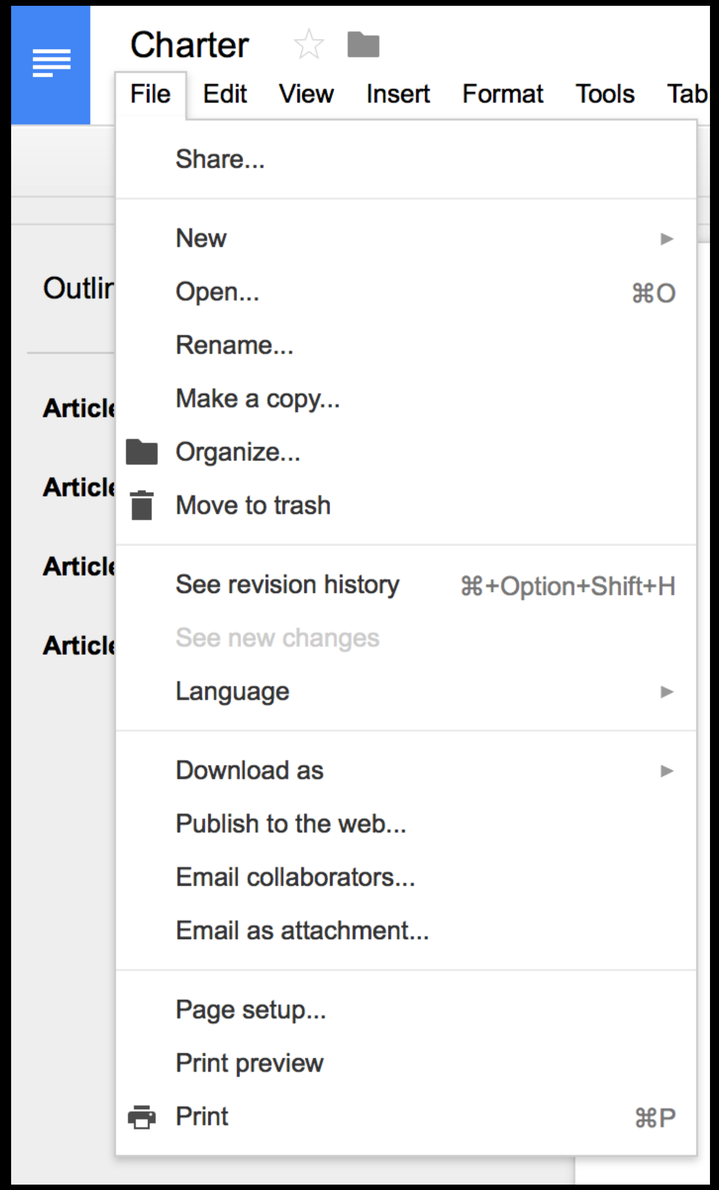
Version Control

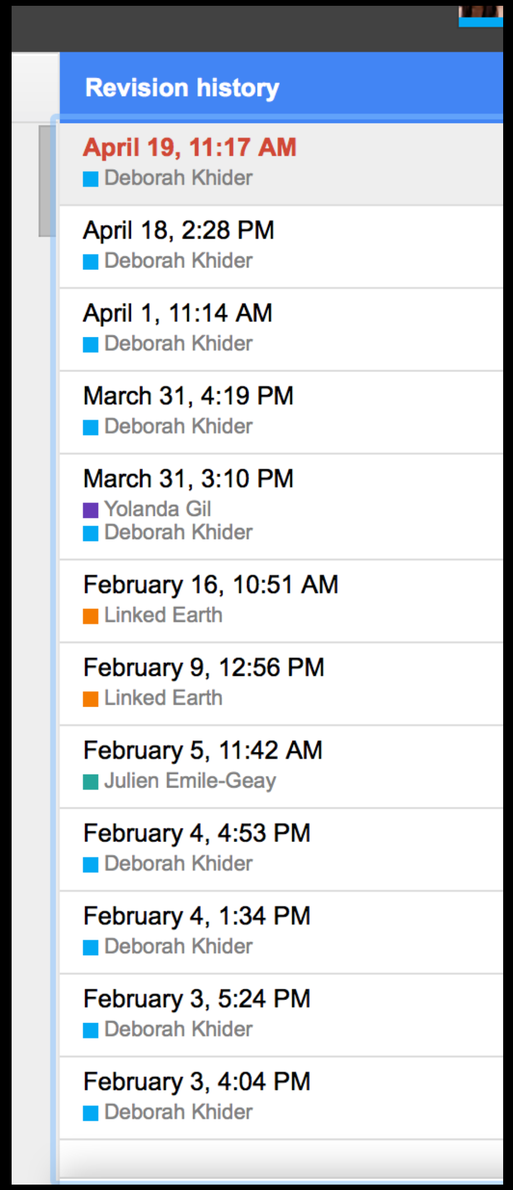
Version Control



Built into Pages, Numbers, Keynote on the Mac…

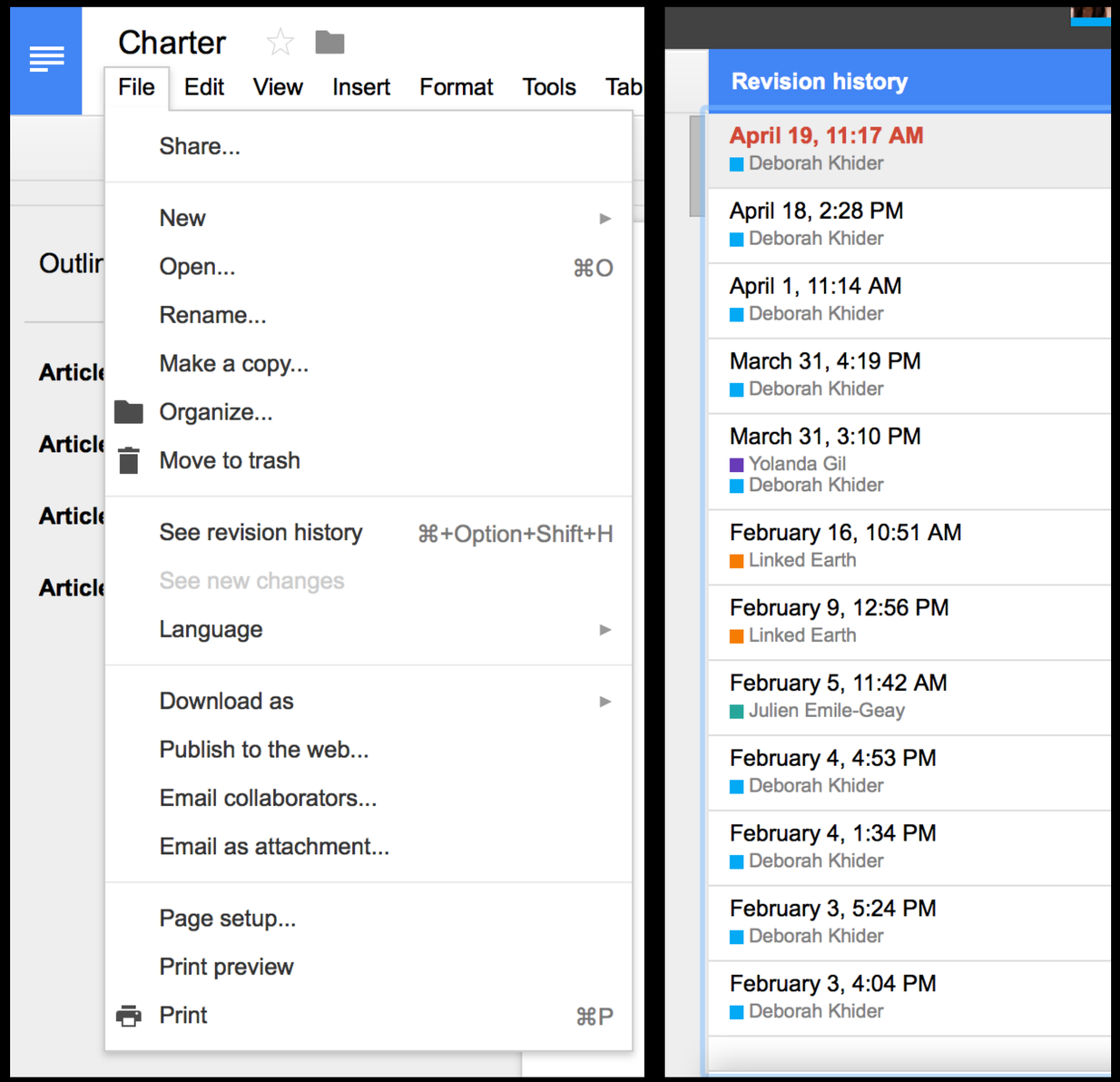
Version Control





... as well as Google apps

Version Control

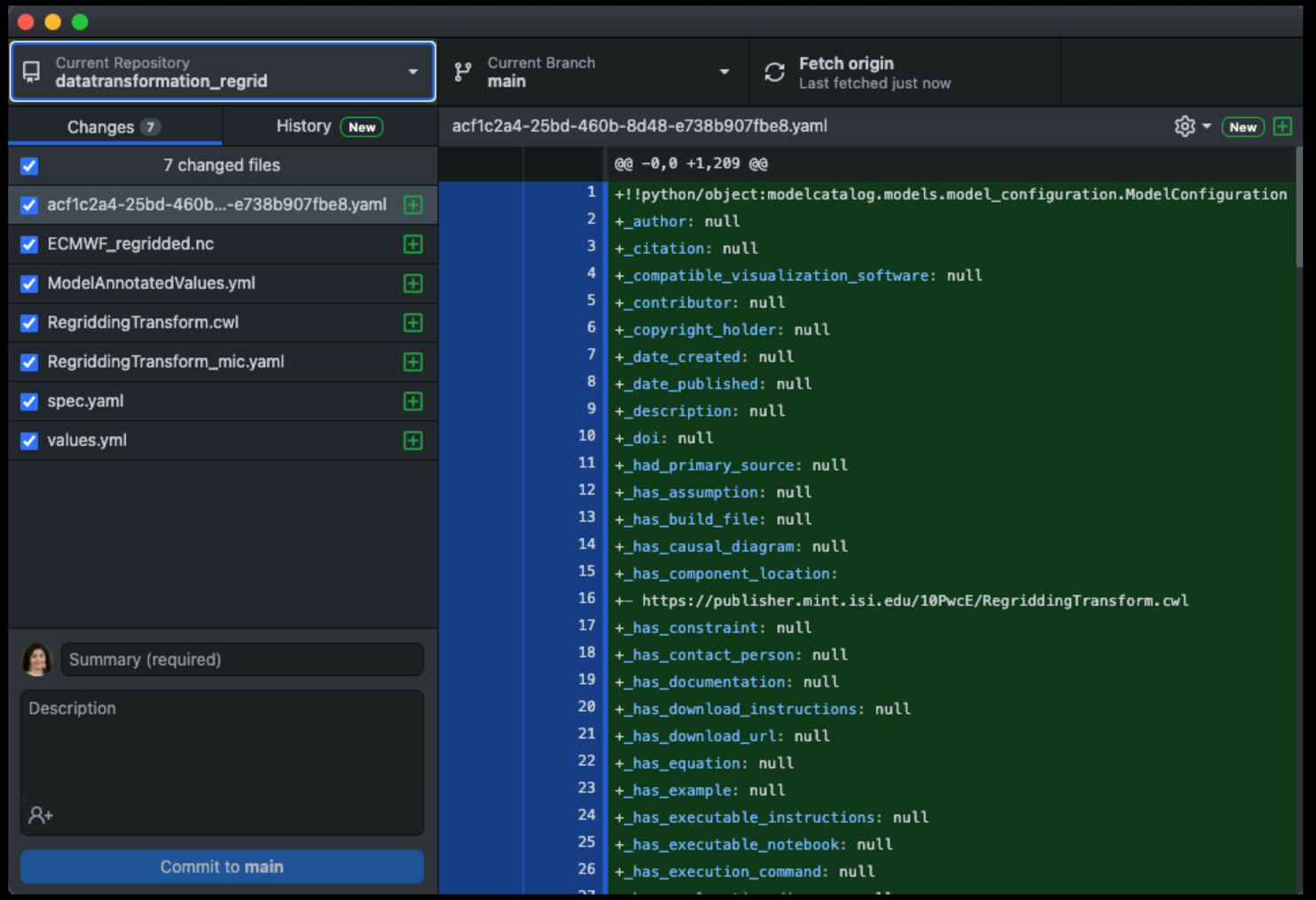


Overleaf

Authorea

... as well as Google apps

Version Control



Use GitHub (code)

Version Control

Revision history

Raw source Inline diff

Only show named versions

Feb 24, 2021 8:22 PM
Deborah Khider

Wed Feb 24 2021 20:22:11 GMT-0700 (MST)
Drought Period.ipynb

Text cell <QZ5Ef8bvPz9M>
#%% [markdown]

```
1
2
3
4 # Drought prediction for South Sudan and Ethiopia
5
6
7 This notebook trains a CNN to predict drought as expressd as SPI
8
9
```

Text cell <ilCArfsLS5fC>
#%% [markdown]

```
1 ## Software requirement
```

Code cell <2y_fuUzgRtK->
#%% [code]

```
1 !pip install netCDF4
```

Execution output from Nov 20, 2020 11:42 AM

Stream

```
Collecting netCDF4
[?25l  Downloading https://files.pythonhosted.org/package
[K     |████████████████████████████████| 4.3MB 5.1MB/s
[?25hRequirement already satisfied: numpy>=1.9 in /usr/lo
Collecting cftime
[?25l  Downloading https://files.pythonhosted.org/package
[K     |████████████████████████████████| 296kB 30.0MB/s
[?25hInstalling collected packages: cftime, netCDF4
Successfully installed cftime-1.3.0 netCDF4-1.5.4
```

Code cell <UWWai2L8QGMN>
#%% [code]

Wed Feb 24 2021 20:22:11 GMT-0700 (MST)
Drought Period.ipynb

Text cell <QZ5Ef8bvPz9M>
#%% [markdown]

```
1
2
3
4 # Drought prediction for South Sudan and Ethiopia
5
6
7 This notebook trains a CNN to predict drought as expressd as SPI
8
9
```

Text cell <ilCArfsLS5fC>
#%% [markdown]

```
1 ## Software requirement
```

Code cell <2y_fuUzgRtK->
#%% [code]

```
1 !pip install netCDF4
```

Execution output from Nov 20, 2020 11:42 AM

Stream

```
Collecting netCDF4
[?25l  Downloading https://files.pythonhosted.org/package
[K     |████████████████████████████████| 4.3MB 5.1MB/s
[?25hRequirement already satisfied: numpy>=1.9 in /usr/lo
Collecting cftime
[?25l  Downloading https://files.pythonhosted.org/package
[K     |████████████████████████████████| 296kB 30.0MB/s
[?25hInstalling collected packages: cftime, netCDF4
Successfully installed cftime-1.3.0 netCDF4-1.5.4
```

Code cell <UWWai2L8QGMN>
#%% [code]

Google Colab

Don’t Forget the Metadata!!!

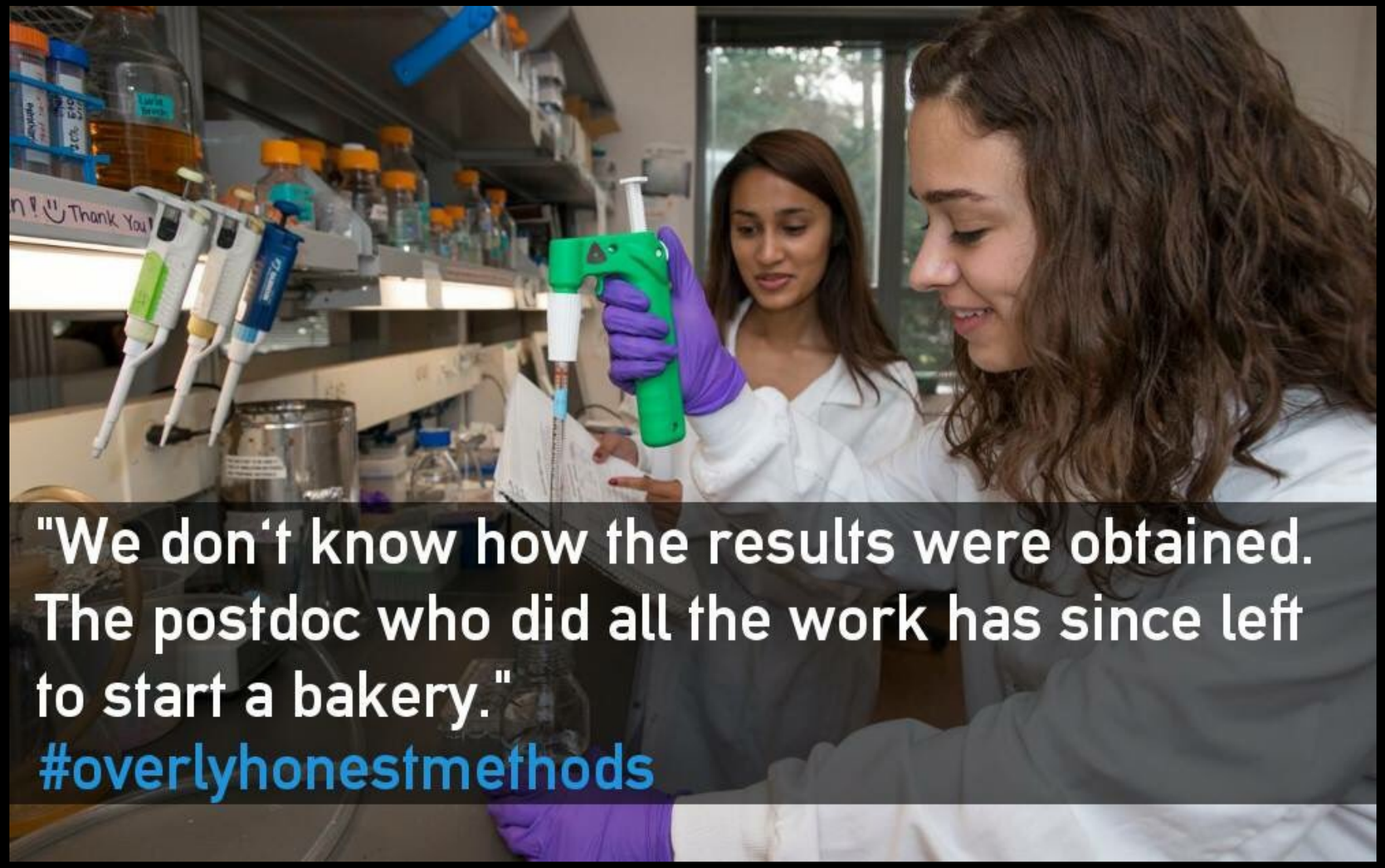
TYPES OF METADATA

- **Descriptive Metadata:** Location, collection frequency, object, etc…
- **Data Characteristics**: Size, statistical properties,…
- **Provenance metadata:** Instrument, Method/Software, Parameters…

Don’t Forget the Metadata!!!

USE OF METADATA

- **Facilitate** reuse by others



Don’t Forget the Metadata!!!

USE OF METADATA

- **Facilitate** reuse by others
- Support **queries** on data repository

In [4]:

```
import json
import requests

url = "http://wiki.linked.earth/store/ds/query"

query = """PREFIX core: <http://linked.earth/ontology#>
PREFIX wiki: <http://wiki.linked.earth/Special:URIResolver/>
PREFIX rdfs: <http://www.w3.org/2000/01/rdf-schema#>
SELECT distinct ?a
WHERE {
{
    ?dataset wiki:Property-3AArchiveType ?a.
}UNION
{
    ?w core:proxyArchiveType ?t.
    ?t rdfs:label ?a
}
}"""

response = requests.post(url, data = {'query': query})
res = json.loads(response.text)

print("The following archive types are available on the wiki:")
for item in res['results']['bindings']:
    print ("*" + item['a']['value'])
```

```
The following archive types are available on the wiki:
*marine sediment
*coral
*lake sediment
*glacier ice
*tree
*documents
*speleothem
*sclerosponge
*borehole
*hybrid
*bivalve
*Rock
```

Don’t Forget the Metadata!!!

USE OF METADATA

- **Facilitate** reuse by others
- Support **queries** on data repository
- **Explain a data analysis** by providing context for the data

Don’t Forget the Metadata!!!

USE OF METADATA

- **Facilitate** reuse by others
- Support **queries** on data repository
- **Explain a data analysis** by providing context for the data
- Enable **automated data integration**

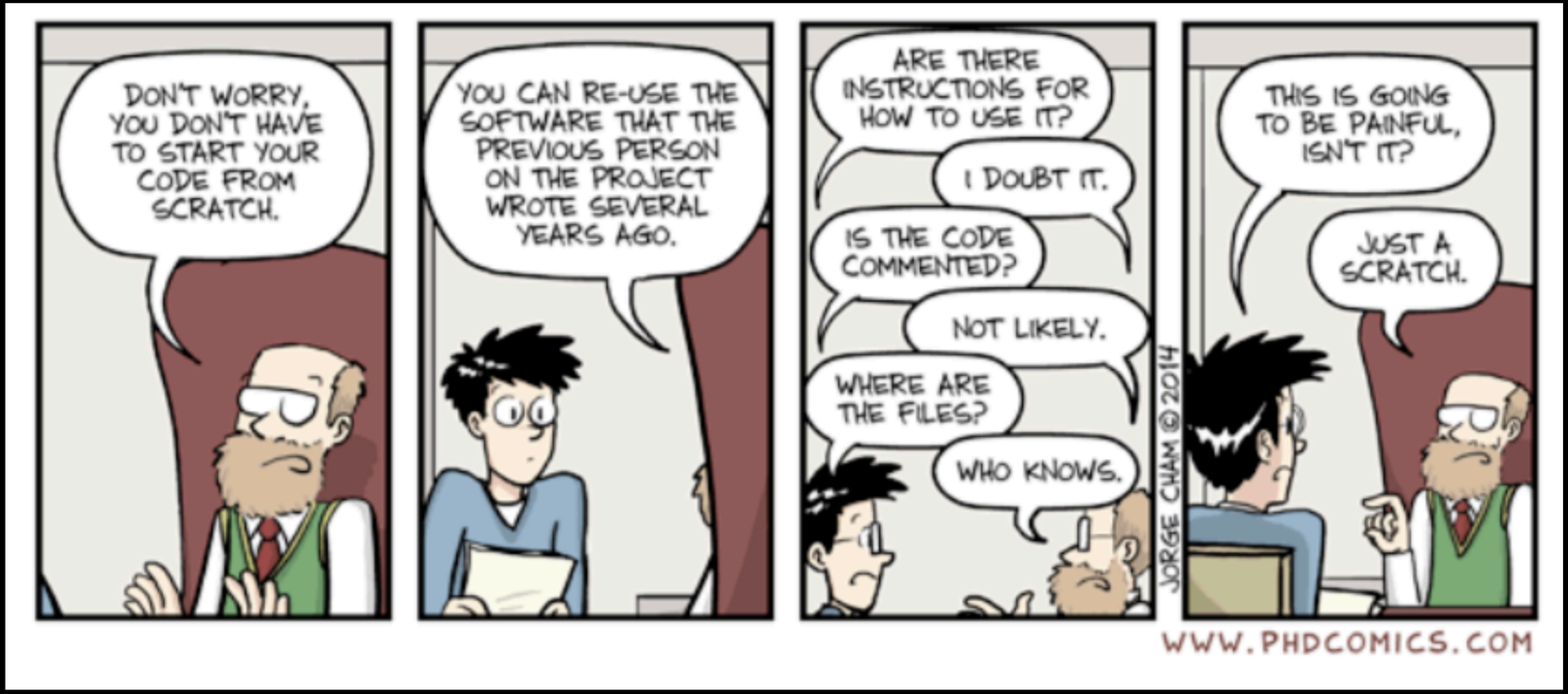
THE BIGGEST LIE I TELL MYSELF IS "I DON'T NEED TO WRITE THAT DOWN I'LL REMEMBER."

Everyone at some point in their life.

Use

WRITE IT DOWN!

- **Methods:** Laboratory, statistics, data cleaning…
- **Comment your code:** 1 line of code = 1 line of comment
- **Meeting notes**



Keep code and explanation together!

Notebook

jupyter spectrogram (autosaved)

File Edit View Insert Cell Kernel Help Python 3

Markdown CellToolbar

Simple spectral analysis

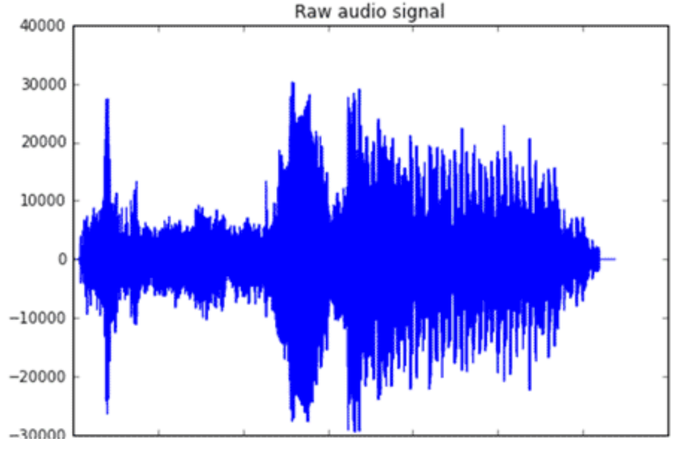
An illustration of the Discrete Fourier Transform

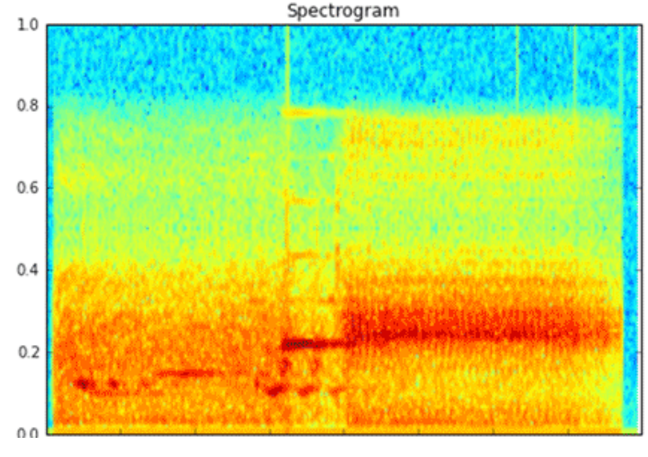
$$X_k = \sum_{n=0}^{N-1} x_n exp^{\frac{-2\pi i}{N} kn} \quad k = 0, \ldots, N-1$$

In [2]:
```
from scipy.io import wavfile
rate, x = wavfile.read('test_mono.wav')
```

And we can easily view it's spectral structure using matplotlib's builtin specgram routine:

In [5]:
```
fig, (ax1, ax2) = plt.subplots(1,2,figsize(16,5))
ax1.plot(x); ax1.set_title('Raw audio signal')
ax2.specgram(x); ax2.set_title('Spectrogram');
```





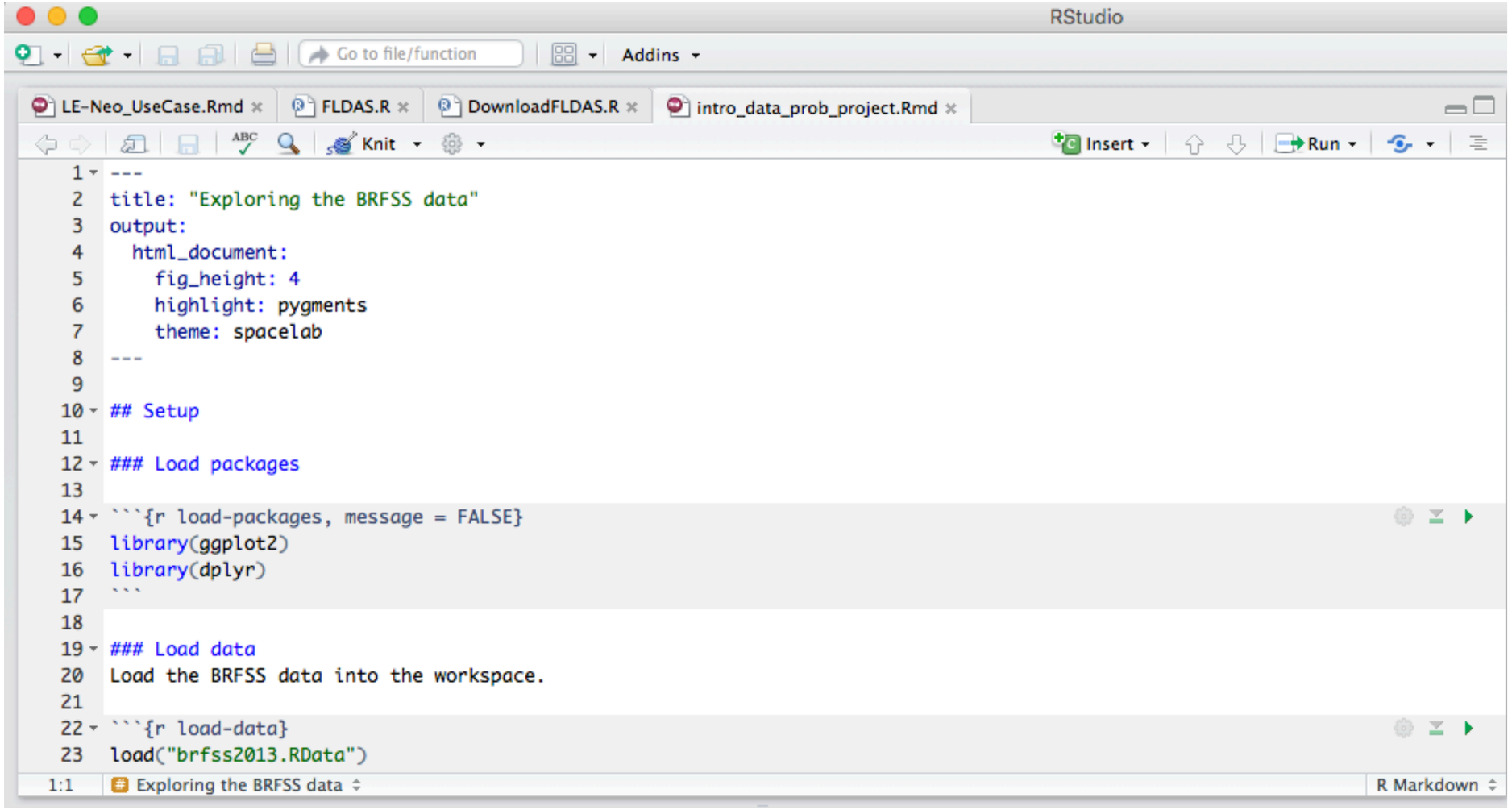
Keep code and explanation together!

jupyter

Notebook

R Markdown

from RStudio



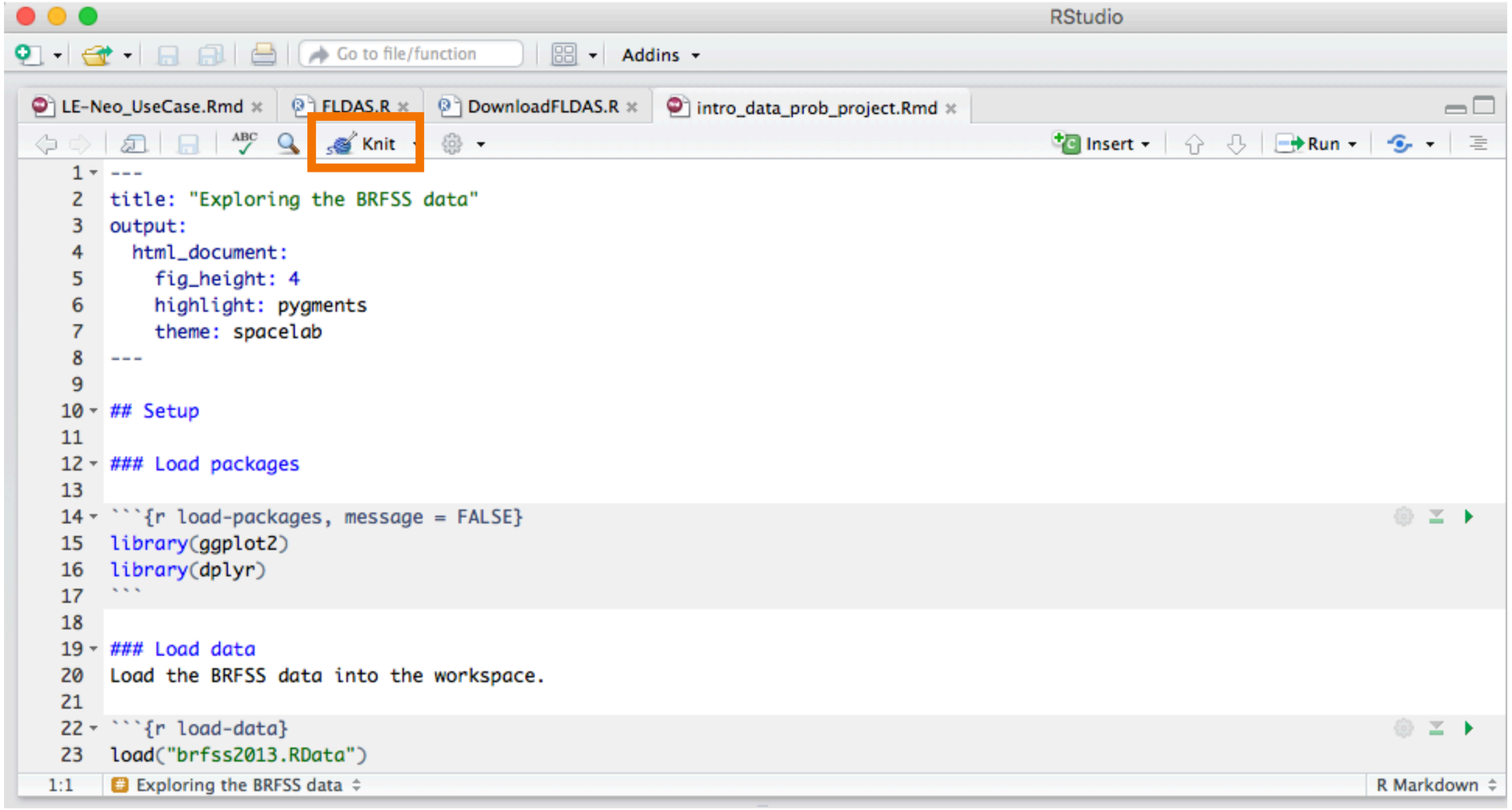
Keep code and explanation together!

jupyter

Notebook

R Markdown

from RStudio



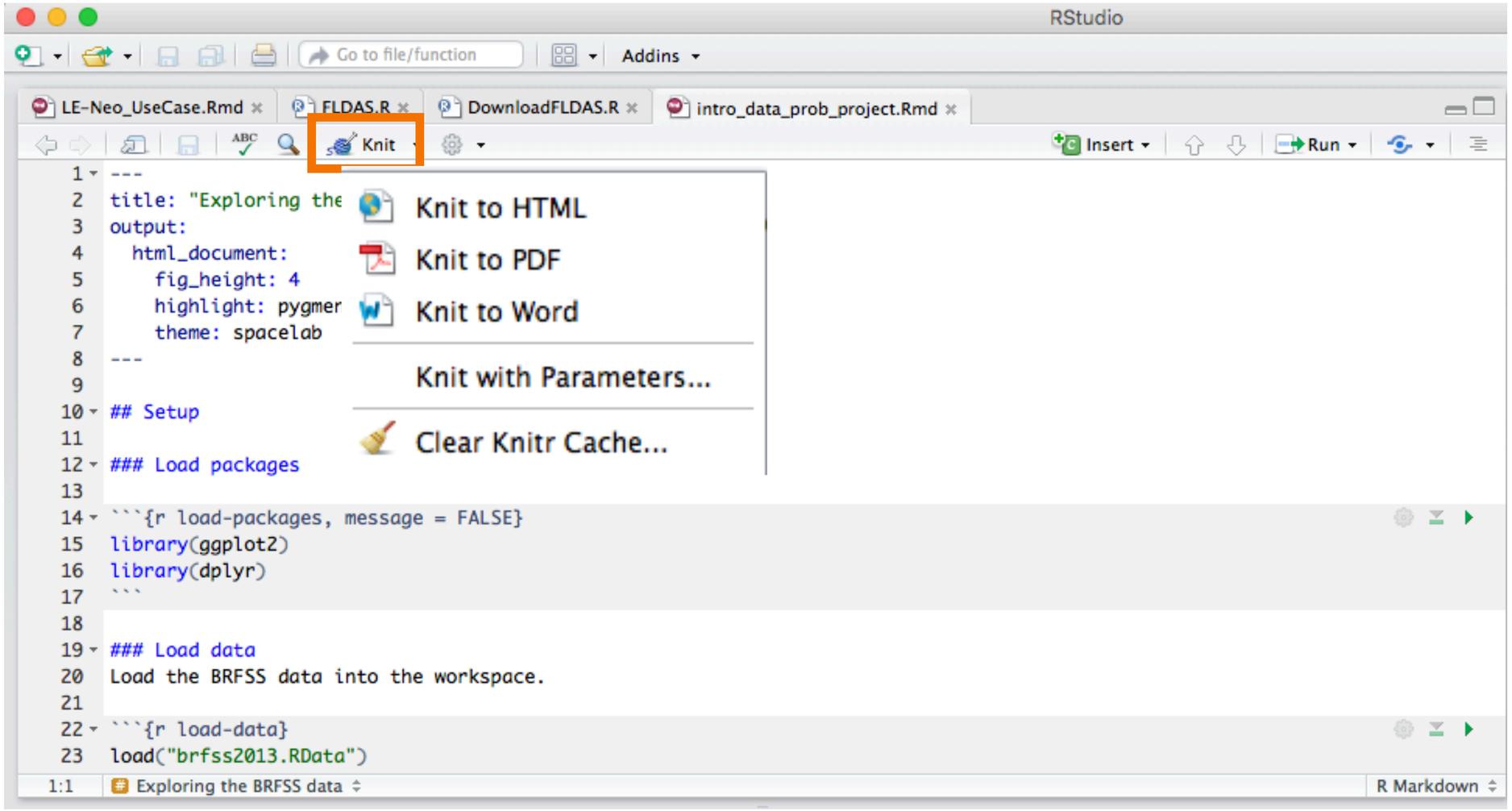
Keep code and explanation together!

jupyter

Notebook

R Markdown

from RStudio



Keep code and explanation together!

jupyter

Notebook

R Markdown

from RStudio

Part 3: Exploratory data analysis

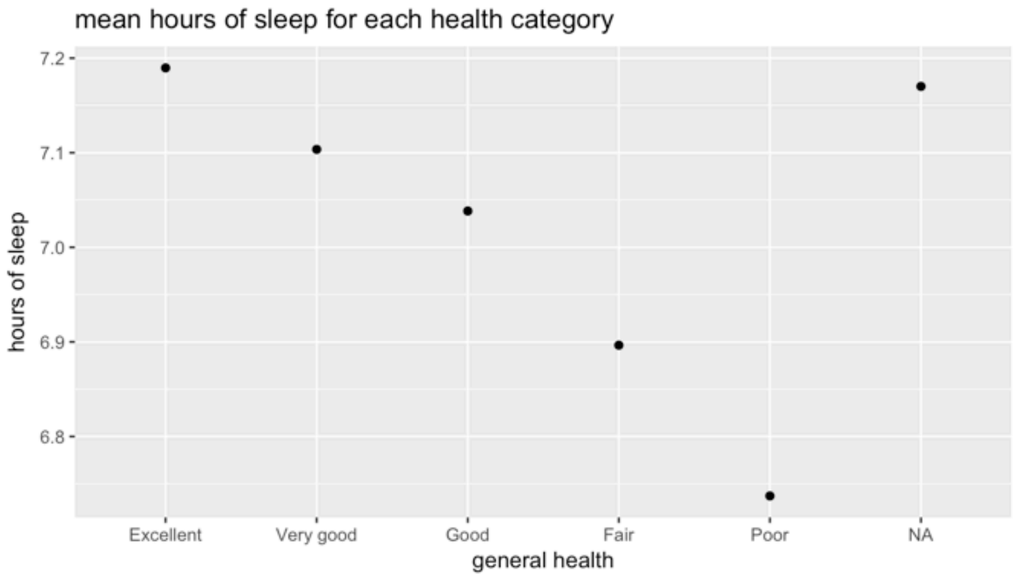
Research question 1: Relationship between sleep and general health

```
#Remove outliers (can't really sleep more than 24hours)
sleeptime <- filter(brfss2013, sleptim1<=24)
```

```
## Warning: package 'bindrcpp' was built under R version 3.2.5
```

```
# Summarize the data
healthSleep <- sleeptime %>%
    group_by(genhlth) %>%
    summarise(meanSleep = mean(sleptim1))

#Plot the general health/mean of sleep time
ggplot(healthSleep, aes(genhlth, meanSleep)) + geom_point(aes(genhlth, meanSleep)) +
    labs(title="mean hours of sleep for each health category", x = 'general health', y='hours of sleep')
```



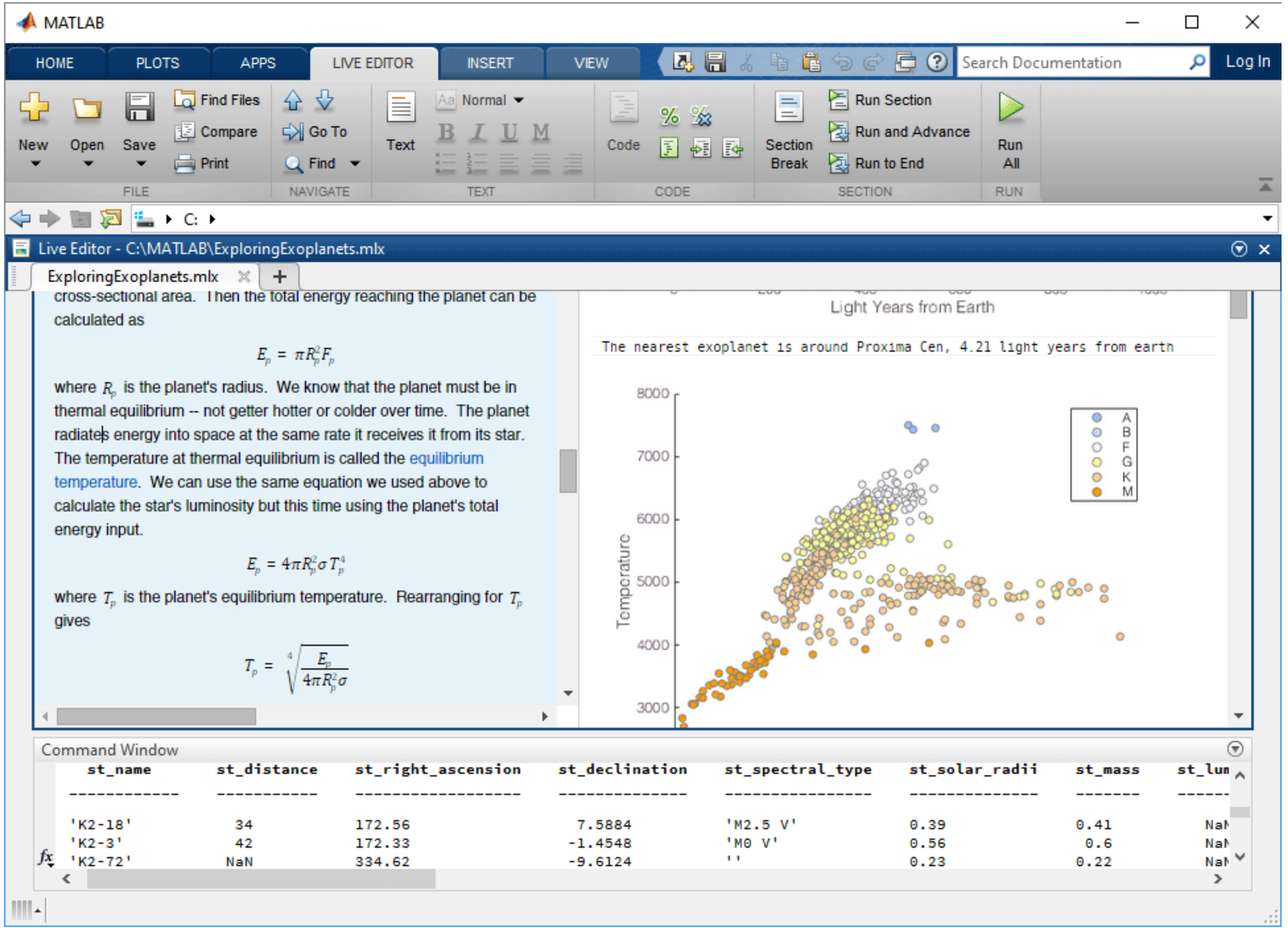
From the plot above, there appears to be a general correlation between health and the amount of sleep.

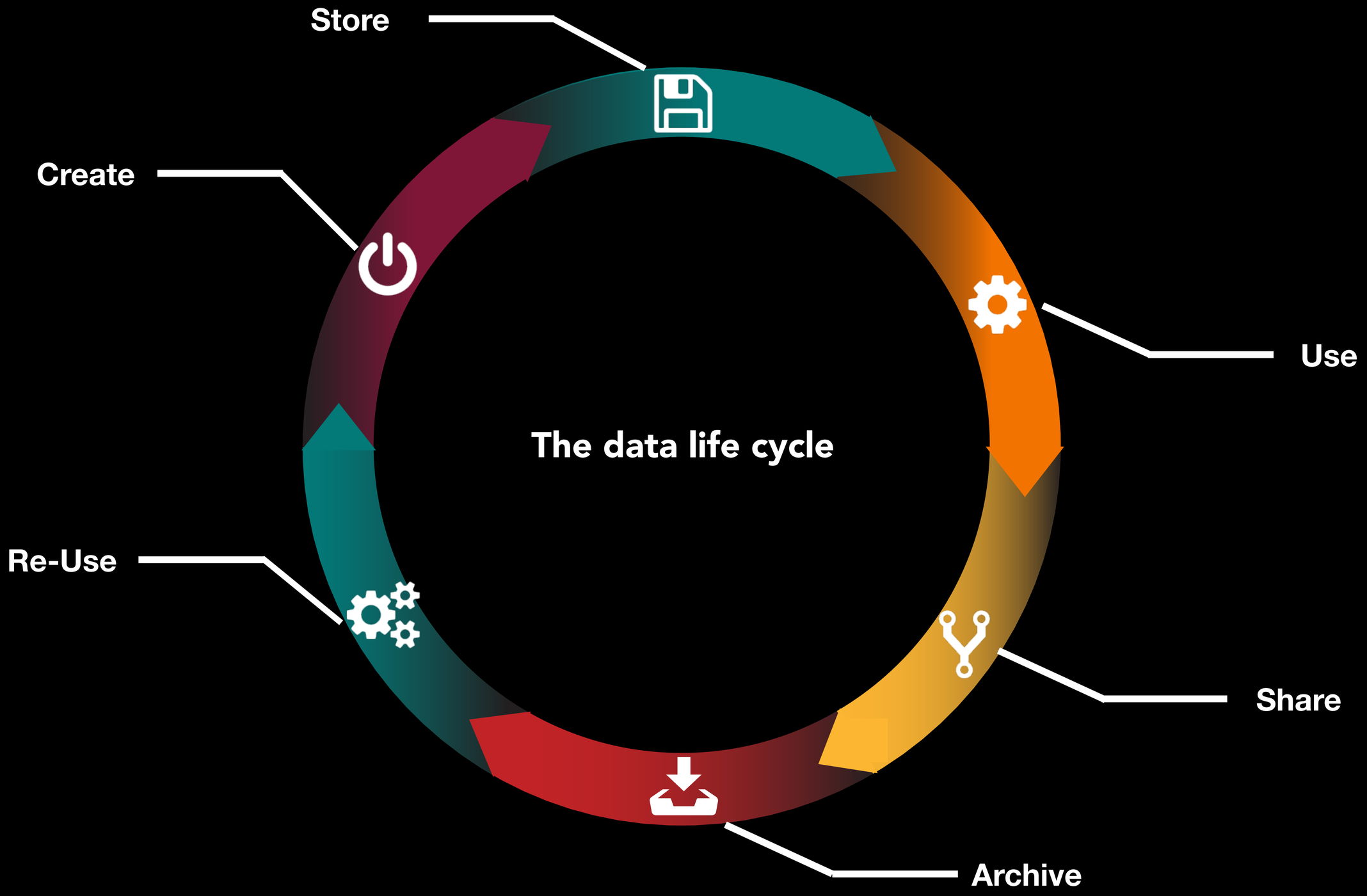
Keep code and explanation together!

jupyter Notebook

R Markdown from RStudio

MathWorks® MATLAB Live Editor




Store
Create
Use
Re-Use
Share
Archive
The data life cycle

Share: Physical Samples

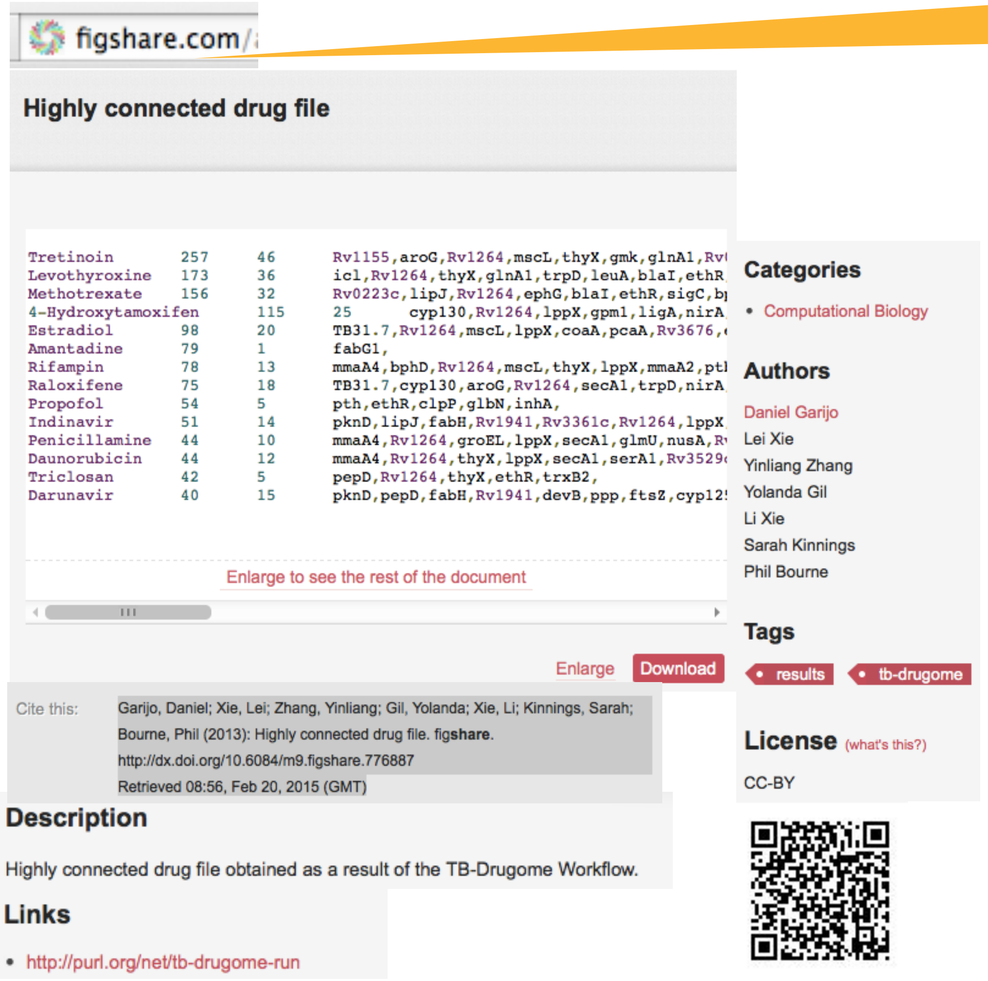
INTERNATIONAL GEO SAMPLE NUMBER: IGSN

- Globally unique and persistent identifier for physical samples in the Earth Sciences
- To obtain a number, go to http://www.geosamples.org/
- Record and register quality metadata for your samples
  - At a minimum: Location, contact, access restrictions, lithology
- Use IGSNs in your publications: text, data tables,…

Credit: http://www.scientificpaperofthefuture.org

Share: Data

1. Publication in a shared repository


figshare.com/

Highly connected drug file

Tretinoin	257	46	Rv1155,aroG,Rv1264,mscL,thyX,gmk,glnA1,Rv
Levothyroxine	173	36	icl,Rv1264,thyX,glnA1,trpD,leuA,blaI,ethR
Methotrexate	156	32	Rv0223c,lipJ,Rv1264,ephG,blaI,ethR,sigC,bp
4-Hydroxytamoxifen	115	25	cyp130,Rv1264,lppX,gpm1,ligA,nirA
Estradiol	98	20	TB31.7,Rv1264,mscL,lppX,coaA,pcaA,Rv3676,
Amantadine	79	1	fabG1,
Rifampin	78	13	mmaA4,bphD,Rv1264,mscL,thyX,lppX,mmaA2,ptl
Raloxifene	75	18	TB31.7,cyp130,aroG,Rv1264,secA1,trpD,nirA
Propofol	54	5	pth,ethR,clpP,glbN,inhA,
Indinavir	51	14	pknD,lipJ,fabH,Rv1941,Rv3361c,Rv1264,lppX
Penicillamine	44	10	mmaA4,Rv1264,groEL,lppX,secA1,glmU,nusA,Rv
Daunorubicin	44	12	mmaA4,Rv1264,thyX,lppX,secA1,serA1,Rv3529c
Triclosan	42	5	pepD,Rv1264,thyX,ethR,trxB2,
Darunavir	40	15	pknD,pepD,fabH,Rv1941,devB,ppp,ftsZ,cyp12

Enlarge to see the rest of the document

Enlarge Download

Cite this: Garijo, Daniel; Xie, Lei; Zhang, Yinliang; Gil, Yolanda; Xie, Li; Kinnings, Sarah; Bourne, Phil (2013): Highly connected drug file. figshare. http://dx.doi.org/10.6084/m9.figshare.776887 Retrieved 08:56, Feb 20, 2015 (GMT)

Description

Highly connected drug file obtained as a result of the TB-Drugome Workflow.

Links

- http://purl.org/net/tb-drugome-run

Categories

- Computational Biology

Authors

Daniel Garijo
Lei Xie
Yinliang Zhang
Yolanda Gil
Li Xie
Sarah Kinnings
Phil Bourne

Tags

results tb-drugome

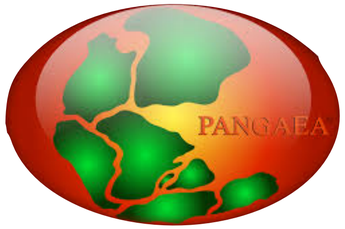
License (what's this?)

CC-BY

Share: Data

Open
Core Data

Formerly the National Climatic Data Center (NCDC)... more about NCEI »

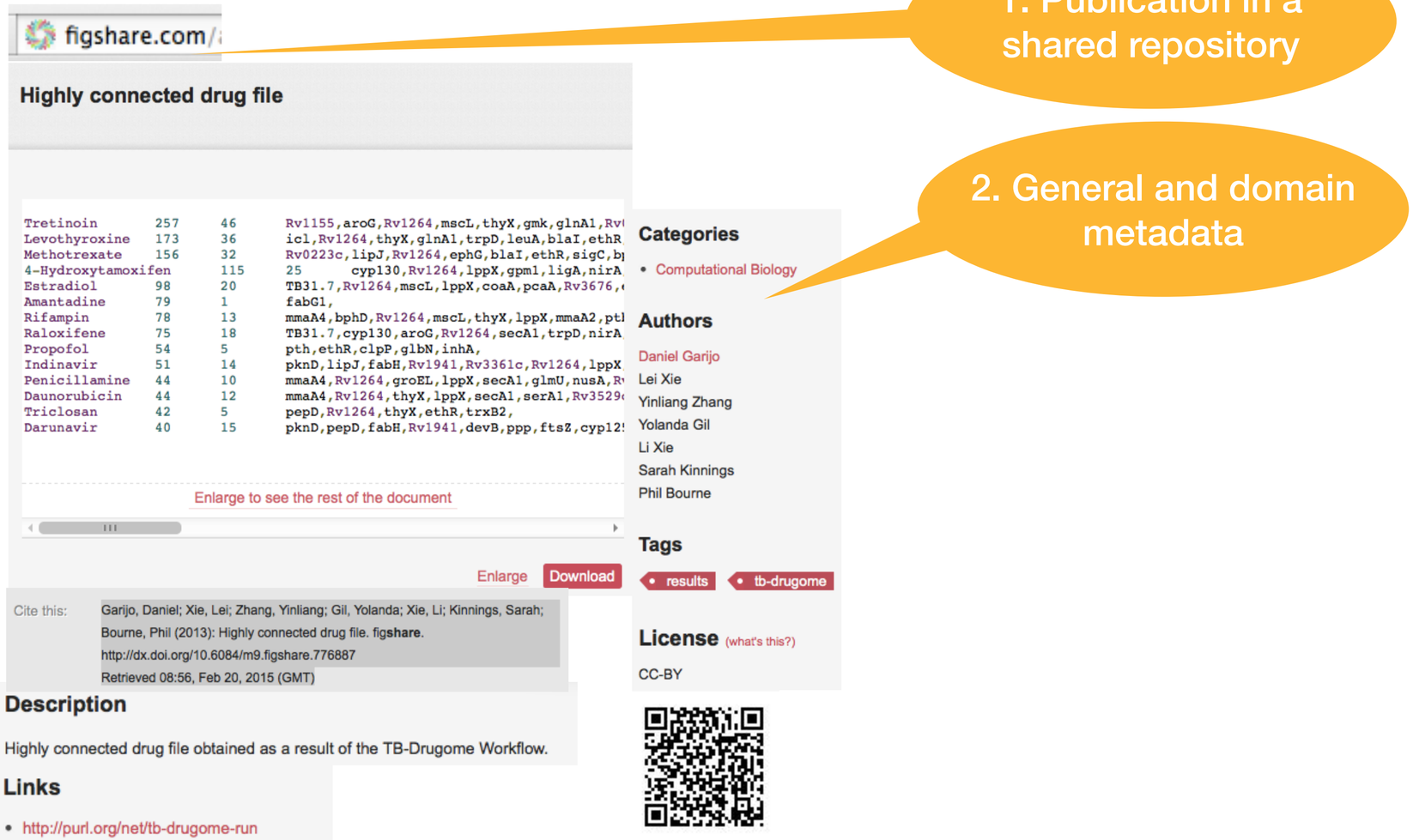


Neotoma

Share: Data

1. Publication in a shared repository

2. General and domain metadata



Share: Data

General

- Dataset name/title
- Description
- Creator(s)
- Publication date
- License
- Publisher/contact
- Version
- Resource type
- Location of the data

Domain Specific

- Categories
- Keywords/tags
- Related Links
- A data repository in a given discipline may request metadata using accepted standards

Share: Data

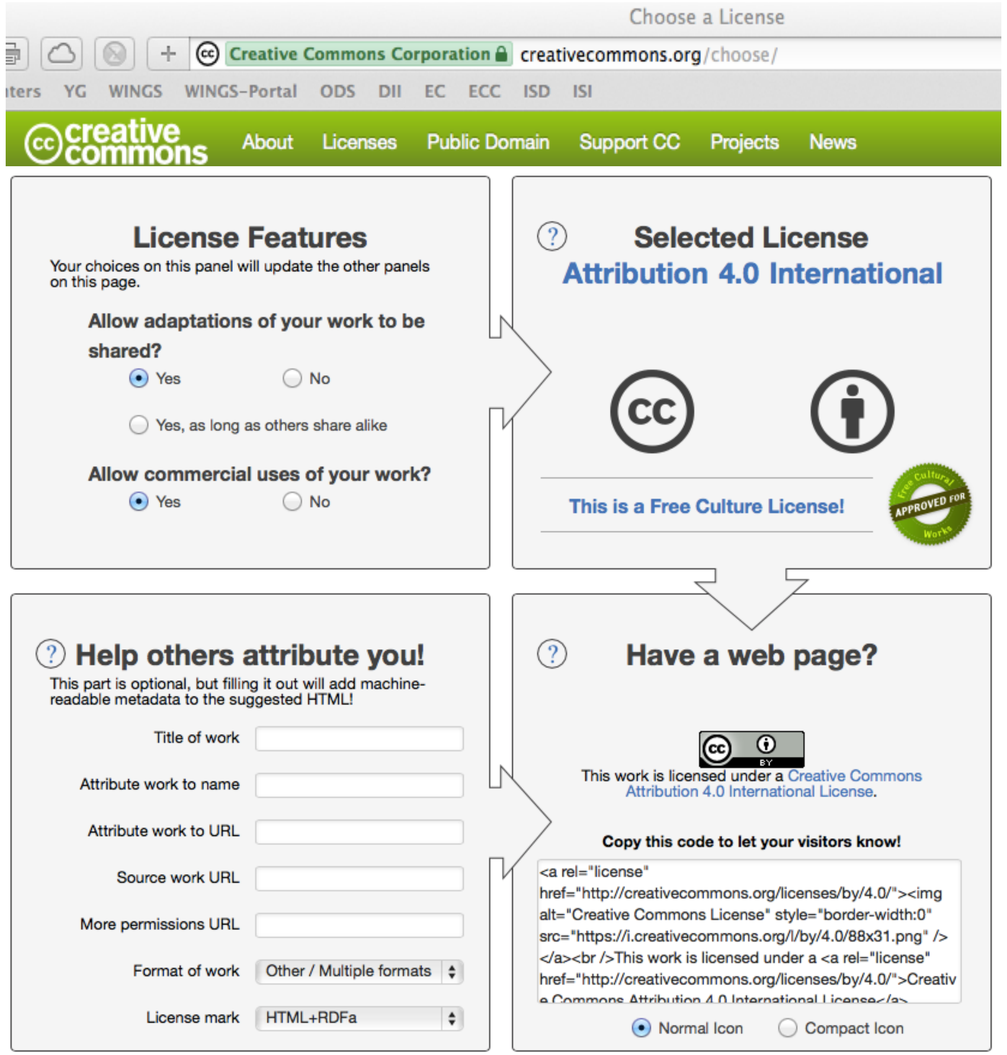
General

- Dataset name/title
- Description
- Creator(s)
- Publication date
- License
- Publisher/contact
- Version
- Resource type
- Location of the data

Domain Specific

- Categories
- Keywords/tags
- Related Links
- A data repository in a given discipline may request metadata using accepted standards

Share: Data



Recommended: CC-BY and CC0

Attribution
CC BY

This license lets others distribute, remix, tweak, and build upon your work, even commercially, as long as they credit you for the original creation. This is the most accommodating of licenses offered. Recommended for maximum dissemination and use of licensed materials.

CC0 (datasets)

CC0 can be particularly important for the sharing of data and databases, since it otherwise may be unclear whether highly factual data and databases are restricted by copyright or other rights. Databases may contain facts that, in and of themselves, are not protected by copyright law.

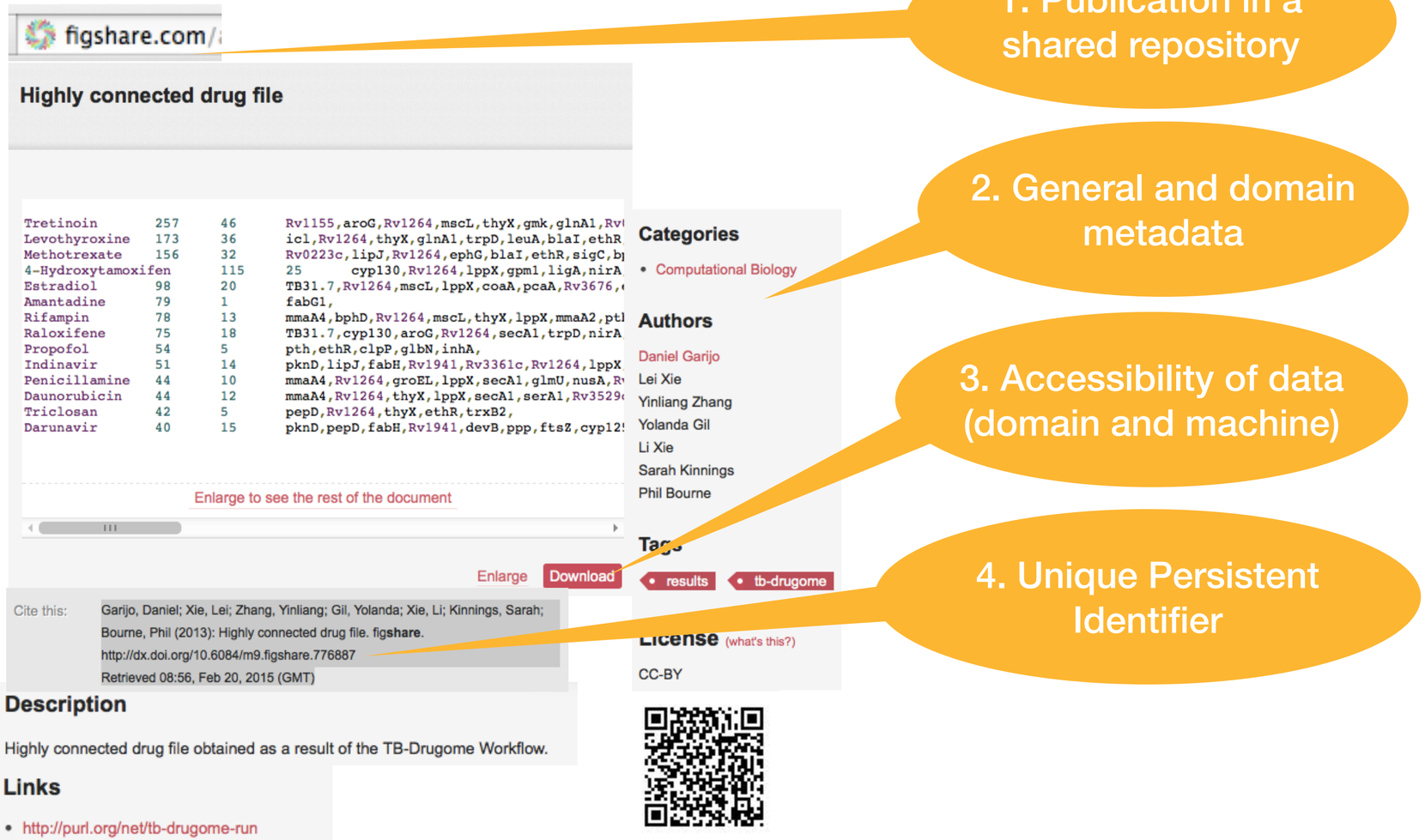
CC0 is recommended for data and databases and is used by hundreds of organizations. It is especially recommended for scientific data. Although CC0 doesn't legally require users of the data to cite the source, it does not take away the moral responsibility to give attribution, as is common in scientific research.

http://creativecommons.org/licenses/

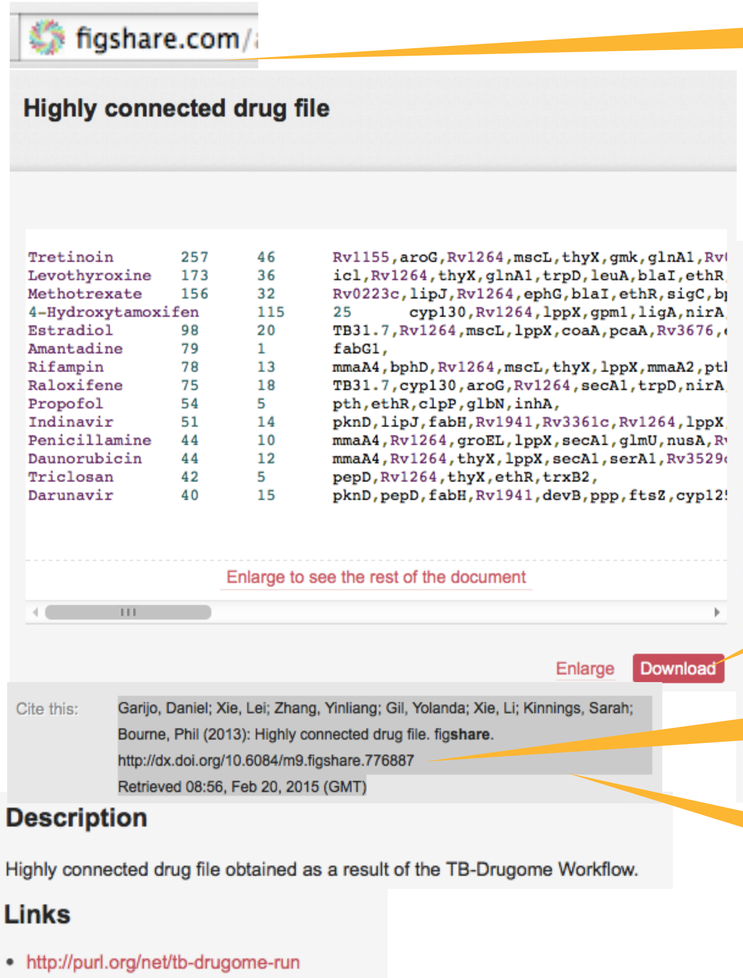
Share: Data

figshare.com/
Highly connected drug file
Tretinoin 257 46 Rv1155,aroG,Rv1264,mscL,thyX,gmk,glnA1,Rv
Levothyroxine 173 36 icl,Rv1264,thyX,glnA1,trpD,leuA,blaI,ethR
Methotrexate 156 32 Rv0223c,lipJ,Rv1264,ephG,blaI,ethR,sigC,b
4-Hydroxytamoxifen 115 25 cyp130,Rv1264,lppX,gpm1,ligA,nirA
Estradiol 98 20 TB31.7,Rv1264,mscL,lppX,coaA,pcaA,Rv3676,
Amantadine 79 1 fabG1,
Rifampin 78 13 mmaA4,bphD,Rv1264,mscL,thyX,lppX,mmaA2,pt
Raloxifene 75 18 TB31.7,cyp130,aroG,Rv1264,secA1,trpD,nirA
Propofol 54 5 pth,ethR,clpP,glbN,inhA,
Indinavir 51 14 pknD,lipJ,fabH,Rv1941,Rv3361c,Rv1264,lppX
Penicillamine 44 10 mmaA4,Rv1264,groEL,lppX,secA1,glmU,nusA,R
Daunorubicin 44 12 mmaA4,Rv1264,thyX,lppX,secA1,serA1,Rv3529
Triclosan 42 5 pepD,Rv1264,thyX,ethR,trxB2,
Darunavir 40 15 pknD,pepD,fabH,Rv1941,devB,ppp,ftsZ,cyp12
Enlarge to see the rest of the document
Enlarge
Download
Cite this:
Garijo, Daniel; Xie, Lei; Zhang, Yinliang; Gil, Yolanda; Xie, Li; Kinnings, Sarah; Bourne, Phil (2013): Highly connected drug file. figshare.
http://dx.doi.org/10.6084/m9.figshare.776887
Retrieved 08:56, Feb 20, 2015 (GMT)
Description
Highly connected drug file obtained as a result of the TB-Drugome Workflow.
Links
http://purl.org/net/tb-drugome-run
Categories
Computational Biology
Authors
Daniel Garijo
Lei Xie
Yinliang Zhang
Yolanda Gil
Li Xie
Sarah Kinnings
Phil Bourne
Tags
results
tb-drugome
License (what's this?)
CC-BY
1. Publication in a shared repository
2. General and domain metadata
3. Accessibility of data (domain and machine)

Share: Data


figshare.com/
Highly connected drug file
Tretinoin 257 46 Rv1155,aroG,Rv1264,mscL,thyX,gmk,glnA1,Rv
Levothyroxine 173 36 icl,Rv1264,thyX,glnA1,trpD,leuA,blaI,ethR
Methotrexate 156 32 Rv0223c,lipJ,Rv1264,ephG,blaI,ethR,sigC,b
4-Hydroxytamoxifen 115 25 cyp130,Rv1264,lppX,gpm1,ligA,nirA
Estradiol 98 20 TB31.7,Rv1264,mscL,lppX,coaA,pcaA,Rv3676,
Amantadine 79 1 fabG1,
Rifampin 78 13 mmaA4,bphD,Rv1264,mscL,thyX,lppX,mmaA2,pt
Raloxifene 75 18 TB31.7,cyp130,aroG,Rv1264,secA1,trpD,nirA
Propofol 54 5 pth,ethR,clpP,glbN,inhA,
Indinavir 51 14 pknD,lipJ,fabH,Rv1941,Rv3361c,Rv1264,lppX
Penicillamine 44 10 mmaA4,Rv1264,groEL,lppX,secA1,glmU,nusA,R
Daunorubicin 44 12 mmaA4,Rv1264,thyX,lppX,secA1,serA1,Rv3529
Triclosan 42 5 pepD,Rv1264,thyX,ethR,trxB2,
Darunavir 40 15 pknD,pepD,fabH,Rv1941,devB,ppp,ftsZ,cyp12
Enlarge to see the rest of the document
Enlarge
Download
Cite this:
Garijo, Daniel; Xie, Lei; Zhang, Yinliang; Gil, Yolanda; Xie, Li; Kinnings, Sarah; Bourne, Phil (2013): Highly connected drug file. figshare.
http://dx.doi.org/10.6084/m9.figshare.776887
Retrieved 08:56, Feb 20, 2015 (GMT)
Description
Highly connected drug file obtained as a result of the TB-Drugome Workflow.
Links
http://purl.org/net/tb-drugome-run
Categories
Computational Biology
Authors
Daniel Garijo
Lei Xie
Yinliang Zhang
Yolanda Gil
Li Xie
Sarah Kinnings
Phil Bourne
Tags
results
tb-drugome
License (what's this?)
CC-BY
1. Publication in a shared repository
2. General and domain metadata
3. Accessibility of data (domain and machine)
4. Unique Persistent Identifier

Share: Data



1. Publication in a shared repository

2. General and domain metadata

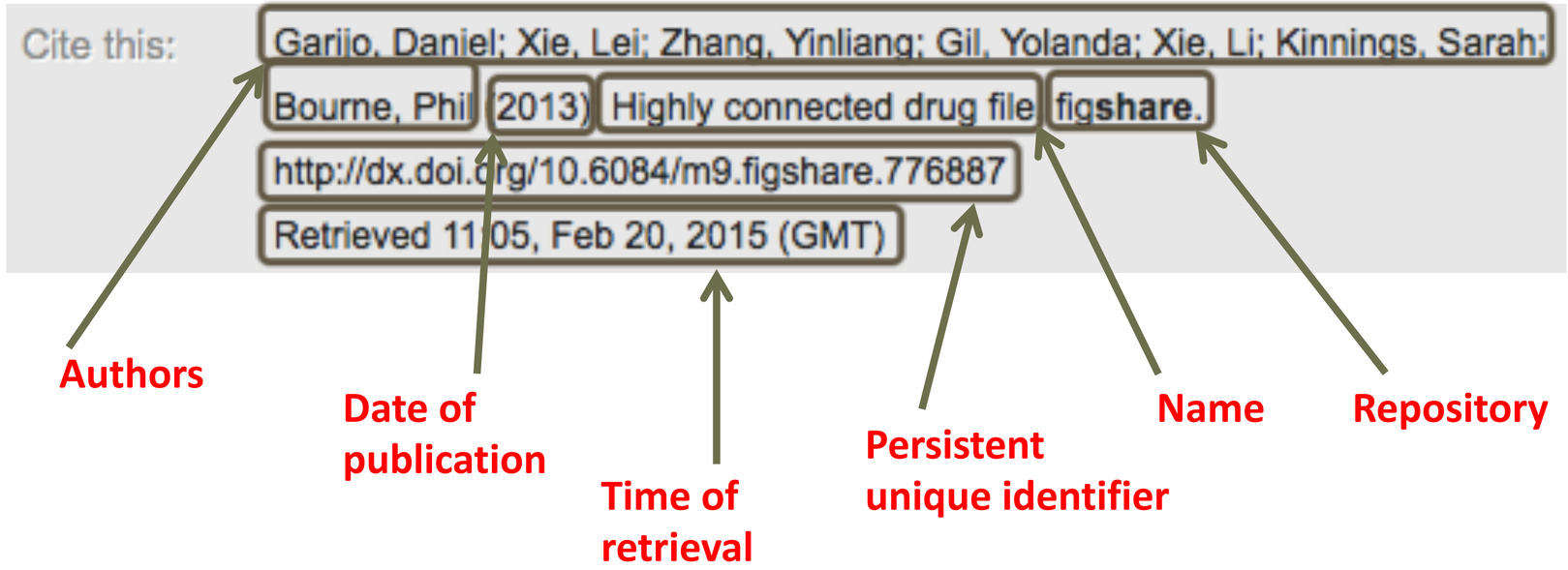
3. Accessibility of data (domain and machine)

4. Unique Persistent Identifier

4. Citation preference

Share: Data

Citing Data: Data repositories and journals often specify how to cite data



Share: Code

Why is Scientific Software not Shared?

- “No one would use my code if I shared it”
- “My code is really bad”
- “My code is not ready to be shared”
- “Sharing my software will take a lot of time”
- “I won’t get anything out of sharing my software”
- “I’ve shared software before, bad things happened”
- “I work for the government”
- “I want to commercialize my software”
- “I don’t want anyone to commercialize my software”
- “I don’t know where to start”

Credit: http://www.scientificpaperofthefuture.org

Share: Code

Why is Scientific Software not Shared?

- “No one would use my code if I shared it”
- “My code is really bad”
- “My code is not ready to be shared”
- “Sharing my software will take a lot of time”
- “I won’t get anything out of sharing my software”
- “I’ve shared software before, bad things happened”
- “I work for the government”
- “I want to commercialize my software”
- “I don’t want anyone to commercialize my software”
- “I don’t know where to start”

Credit: http://www.scientificpaperofthefuture.org

Share: Code

Best practices for making software available

- Accessible from a public location
  - Software repository

Share: Code

Best practices for making software available

- Accessible from a public location
  - Software repository

github SOCIAL CODE HOSTING

launchpad

GitLab

sourceforge

figshare

ATLASSIAN Bitbucket

Share: Code

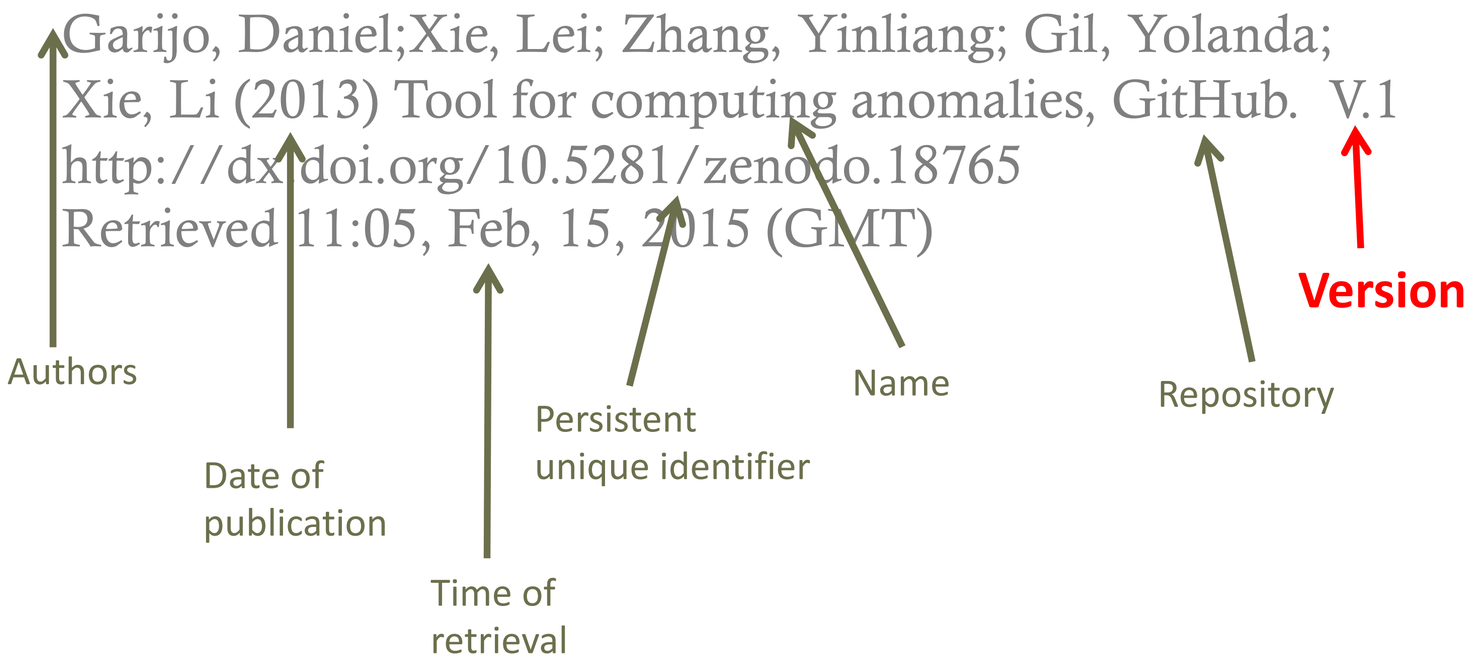
Best practices for making software available

- Accessible from a public location
- License
  - Open source license: reduce constraints and enable software developers to make their source code available to the public
  - “Copyleft”license (GNU General Public License)
  - “Permissive” license

Share: Code

Best practices for making software available

- Accessible from a public location
- License
- Citation



https://zenodo.org

Share: Code

Best practices for making software available

- Accessible from a public location
- License
- Citation
- Executable via a zero-install environment (in the cloud)

binder

Turn a Git repo into a collection of interactive notebooks

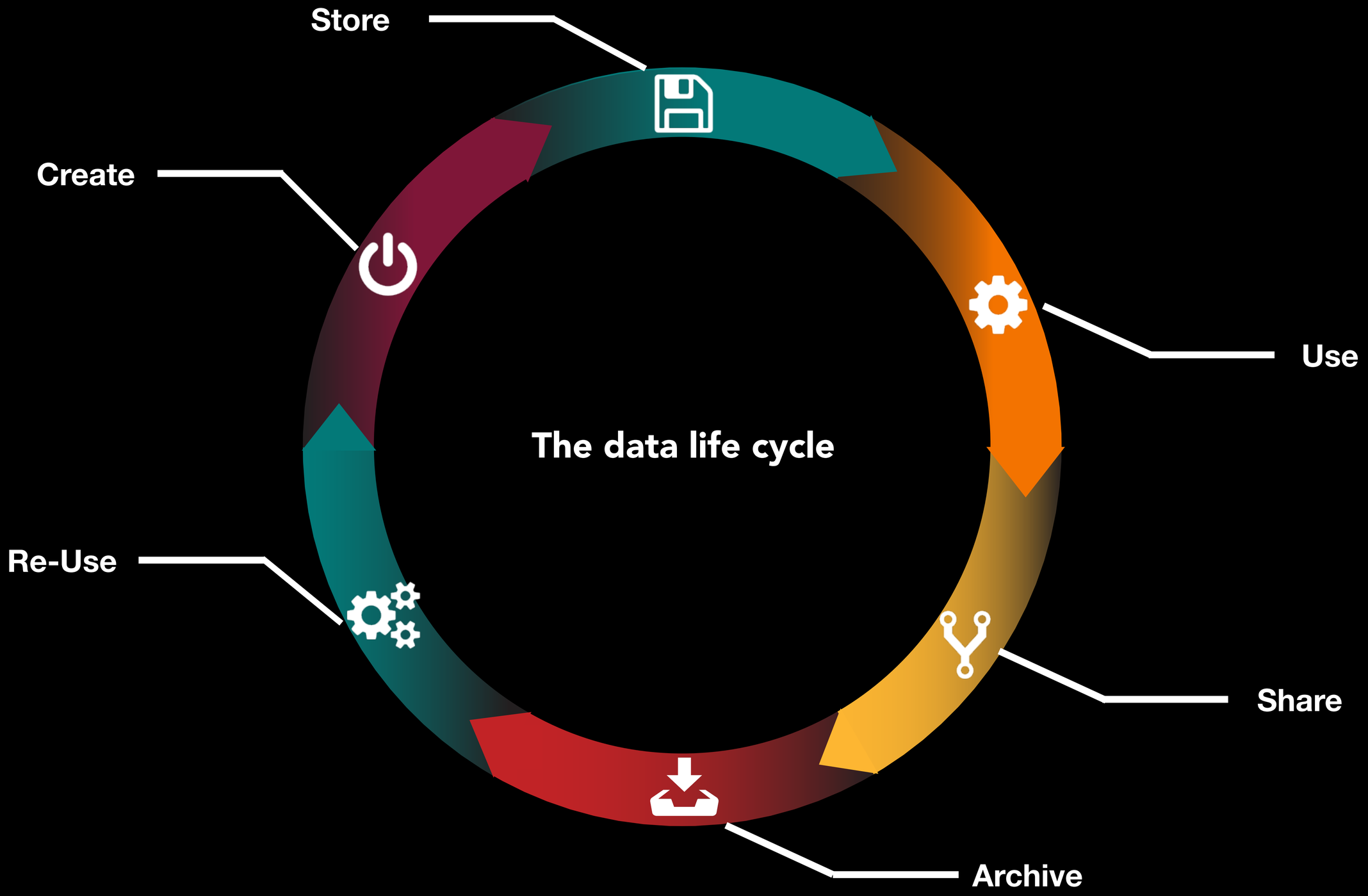
How it works

1. **Enter your repository information**
   Provide in the above form a URL or a GitHub repository that contains Jupyter notebooks, as well as a branch, tag, or commit hash. Launch will build your Binder repository. If you specify a path to a notebook file, the notebook will be opened in your browser after building.

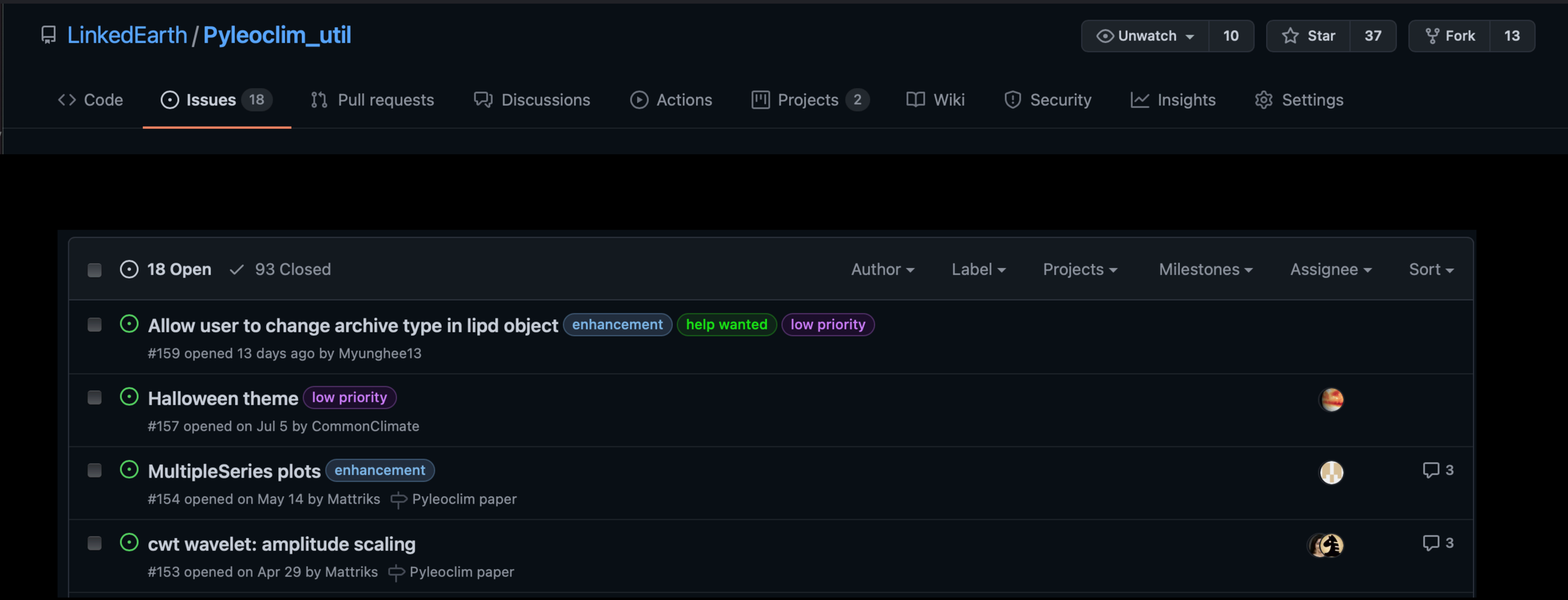
2. **We build a Docker image of your repository**
   Binder will search for a dependency file, such as requirements.txt or environment.yml, in the repository's root directory (more details on more complex dependencies in documentation). The dependency files will be used to build a Docker image. If an image has already been built for the given repository, it will not be rebuilt. If a new commit has been made, the image will automatically be rebuilt.

3. **Interact with your notebooks in a live environment!**
   A JupyterHub server will host your repository's contents. We offer you a reusable link and badge to your live repository that you can easily share with others.

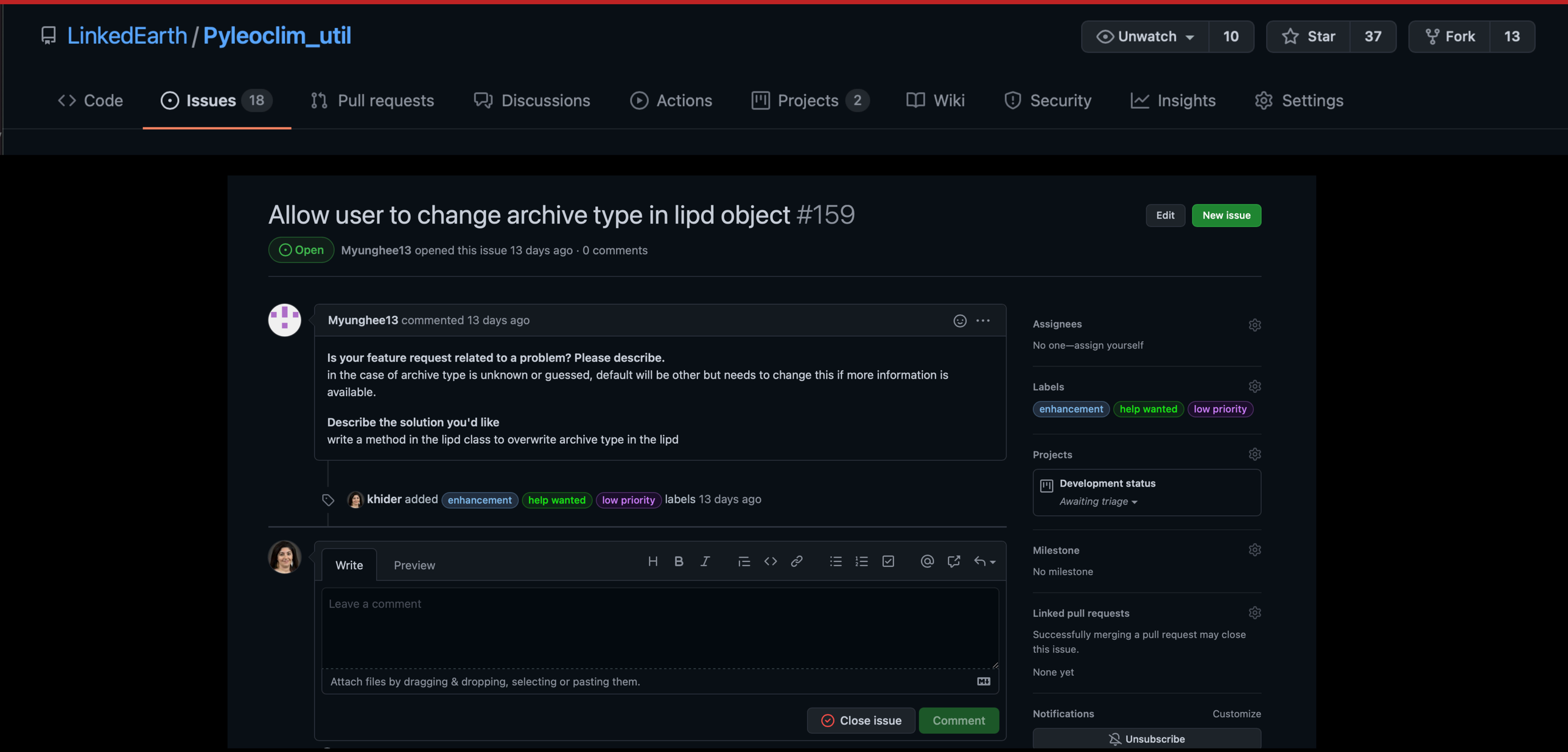
https://mybinder.org
https://github.com/LinkedEarth/paleoHackathon


Store
Create
Use
The data life cycle
Re-Use
Share
Archive

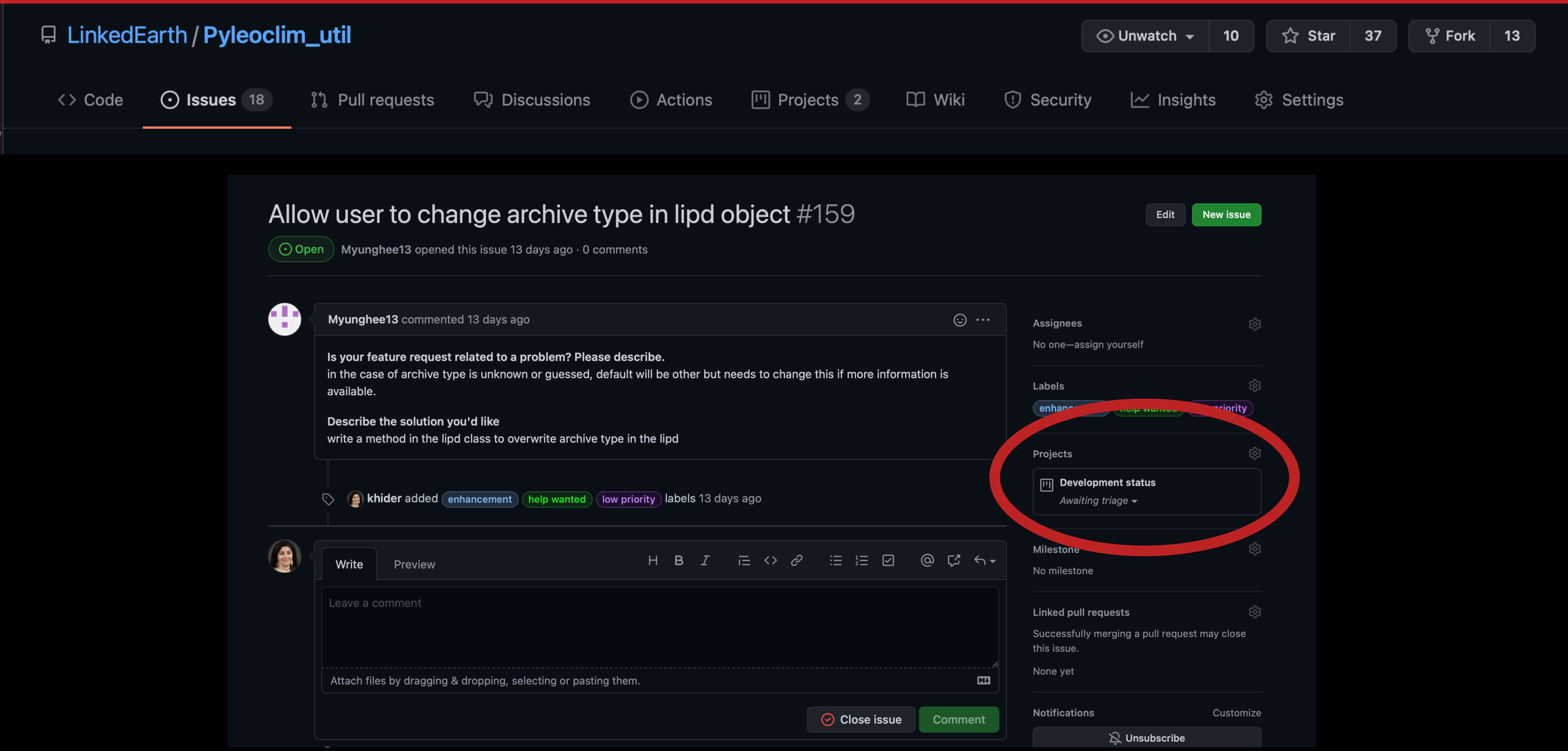
Project management with GitHub



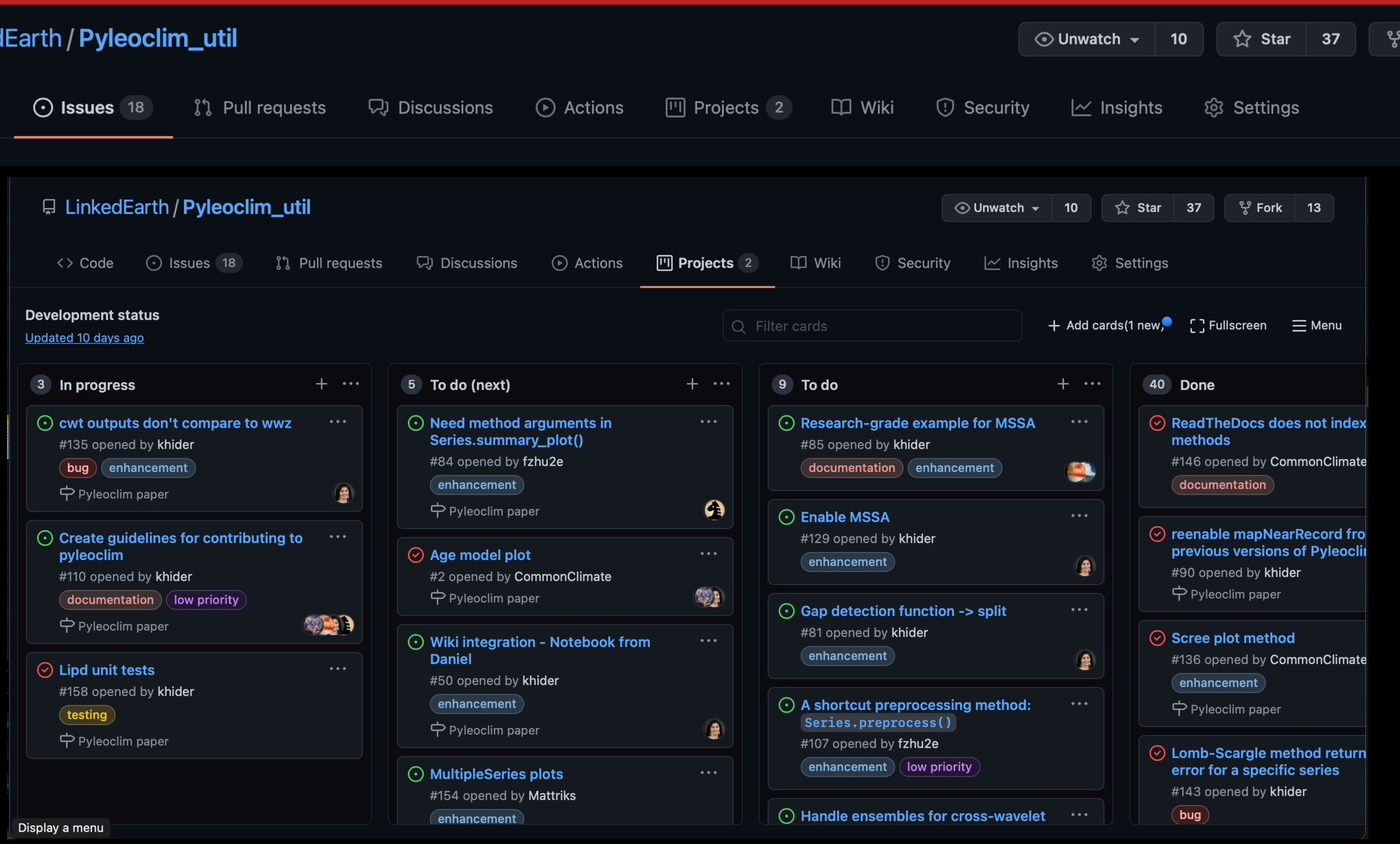
Project management with GitHub



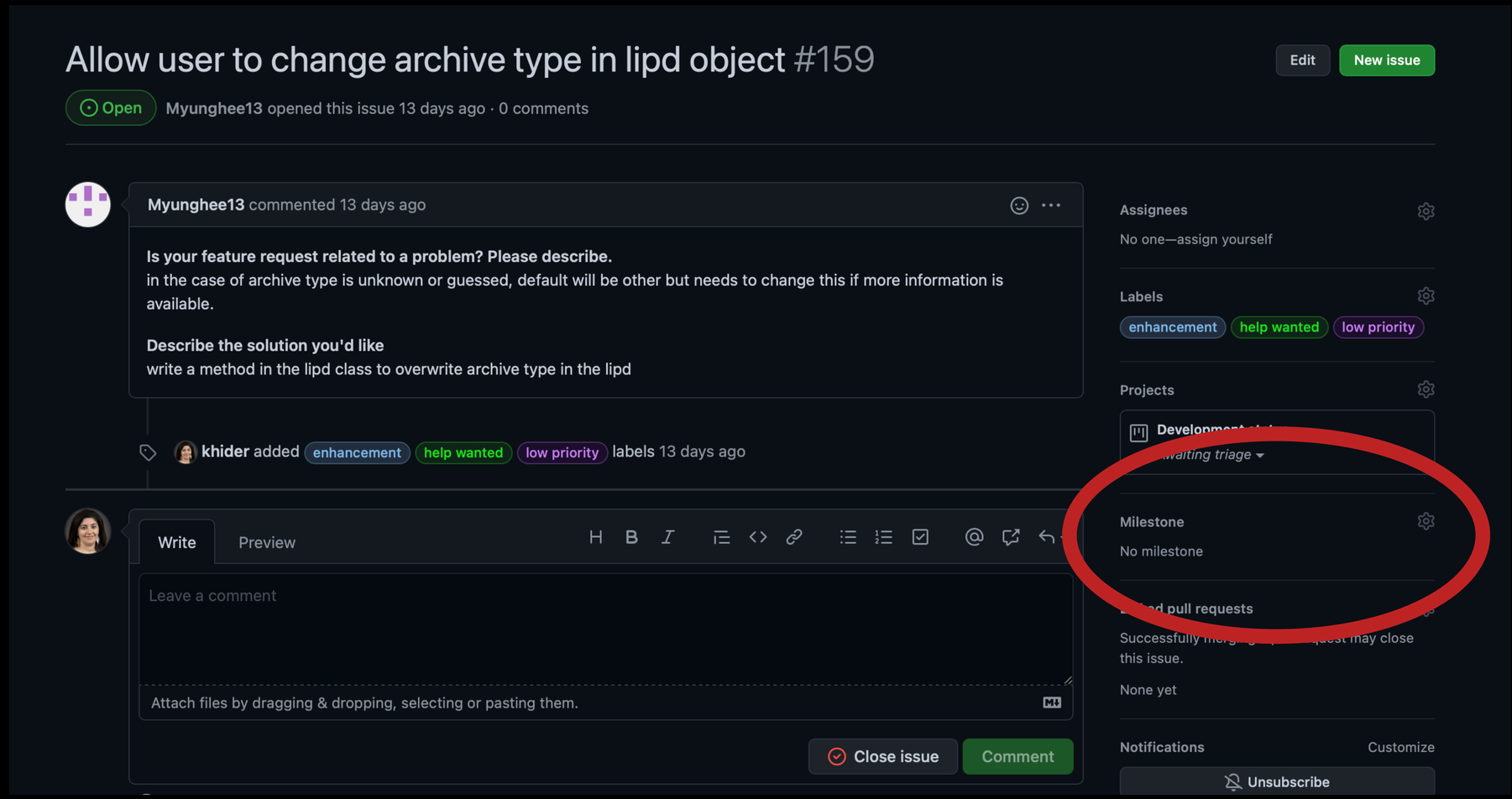
Project management with GitHub



Project management with GitHub



Project management with GitHub



Project management with GitHub

LinkedEarth / Pyleoclim_util

Unwatch 10 | Star 37 | Fork 13

Code | Issues 18 | Pull requests | Discussions | Actions | Projects 2 | Wiki | Security | Insights | Settings

Labels | Milestones

New milestone

2 Open | 4 Closed | Sort

Pyleoclim paper

Past due by 3 months | Last updated about 3 hours ago

Submission of the Pyleoclim package to PP.

62% complete 6 open 10 closed

Edit Close Delete

Support functionalities for Big Paleo Data

Due by August 31, 2021 | Last updated 5 months ago

Functions needed to be incorporated into Pyleoclim to support analysis tools needed for Big Paleo Data work.

0% complete 1 open 0 closed

Edit Close Delete

References

- Geoscience Paper of the Future:
  - Gil, Y., & . (Ed .) .. (2016, April 17). The Geoscience Paper of the Future: OntoSoft Training (Version 9). figshare. https://doi.org/10.6084/m9.figshare.1586773.v9

http://www.scientificpaperofthefuture.org/gpf/

Slides Availability

doi: 10.6084/m9.figshare.6510305

https://figshare.com/s/999787b6f9f6416266b1

License: CC BY 4.0