

Research Data Management

Simple Ways to Make your Research Life Easier

Tom Morrell

BE/Bi 103

November 22, 2017

Current Research Data Practices



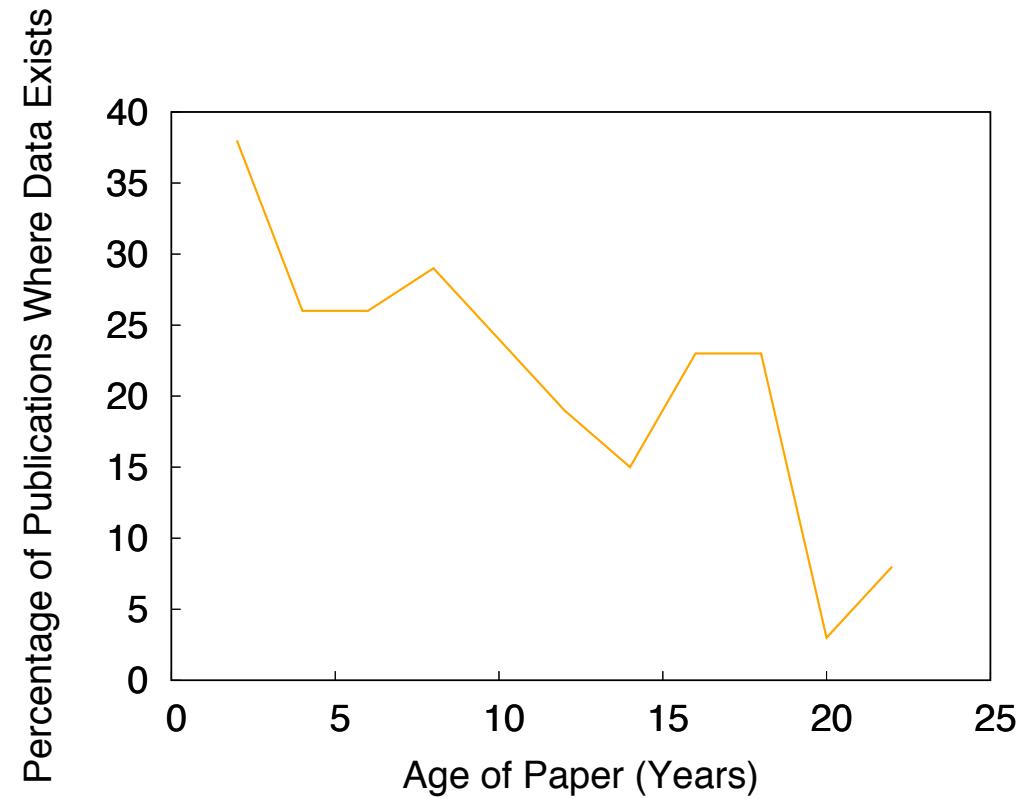
Most researchers store data on local computer hard drives

Researchers report that finding data is their biggest challenge

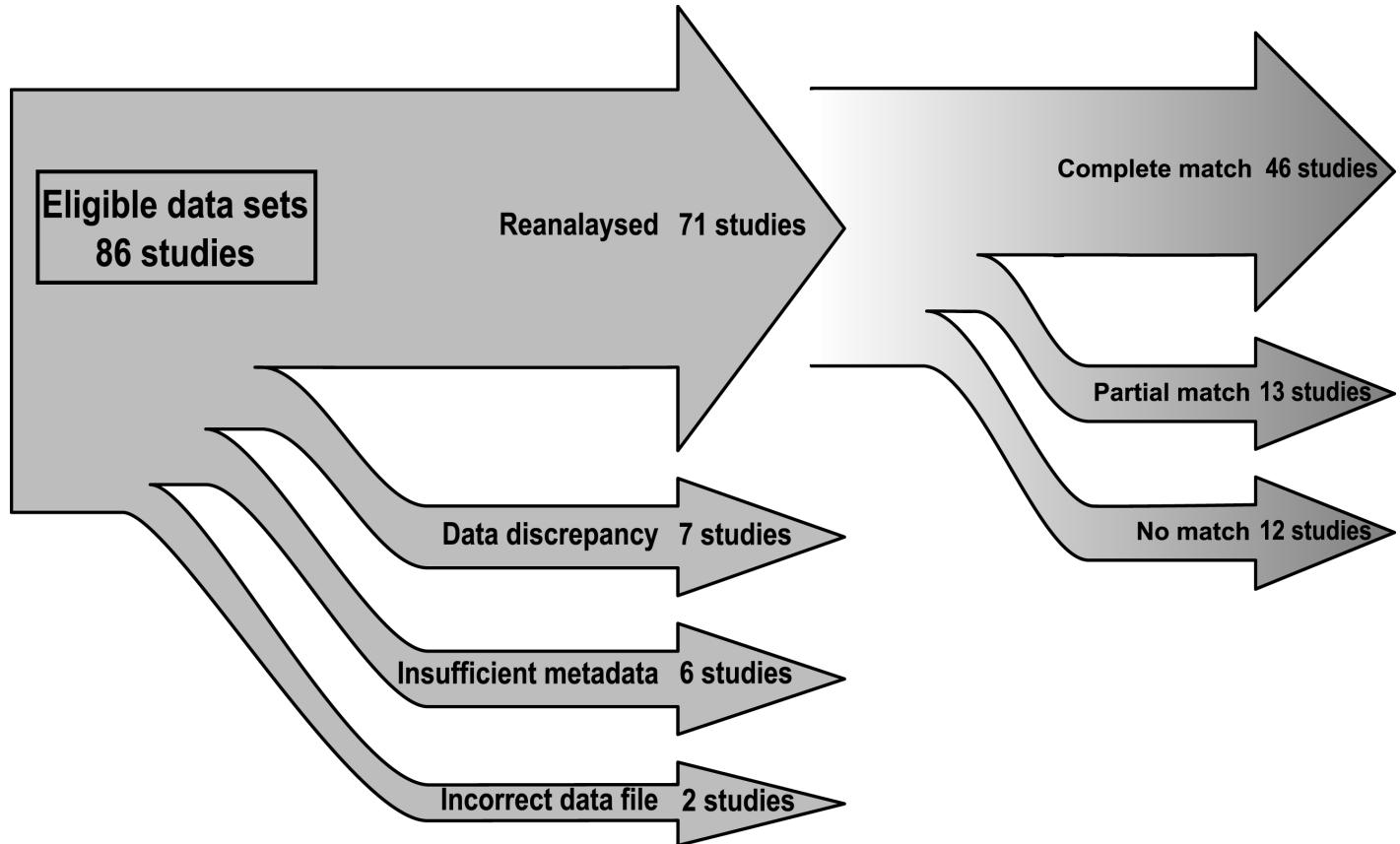
- Akers, K. G. & Doty, J. Disciplinary differences in faculty research data management practices and perspectives. *Int. J. Digit. Curation* **8**, 5–26 (2013). (Emory)
- Shen, Y. Strategic Planning for a Data-Driven, Shared-Access Research Enterprise: Virginia Tech Research Data Assessment and Landscape Study. *Coll. Res. Libr.* **77**, 500–519 (2016).

How Reusable is Research Data Today?

- Morphological characteristics of plants and animals
 - 516 publications using a specific analysis technique between 1991 and 2011
 - 25% of emails didn't work
 - 38% didn't respond to email
 - 13% didn't have data
 - 4% didn't want to share
 - Received 19% of data
 - Availability decreased with time

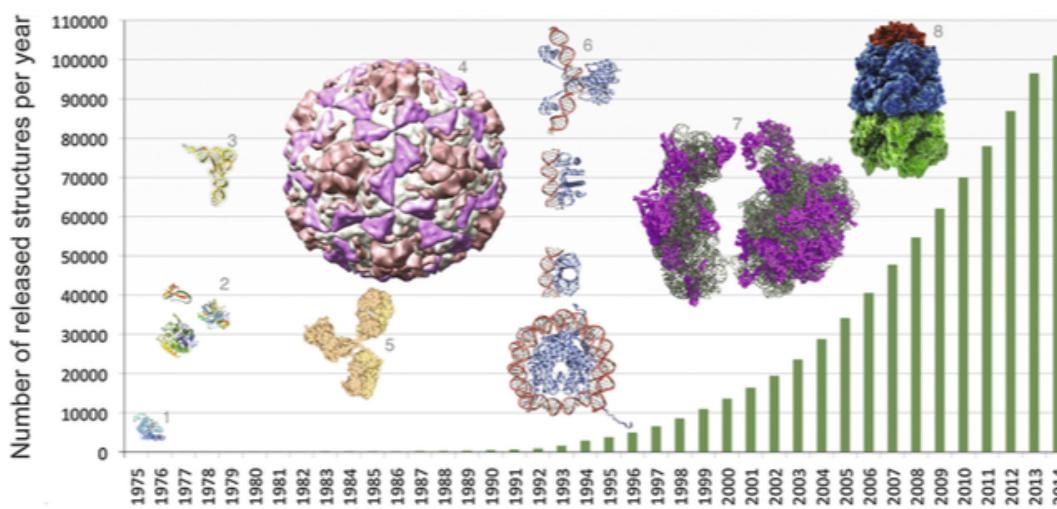


Data Quality

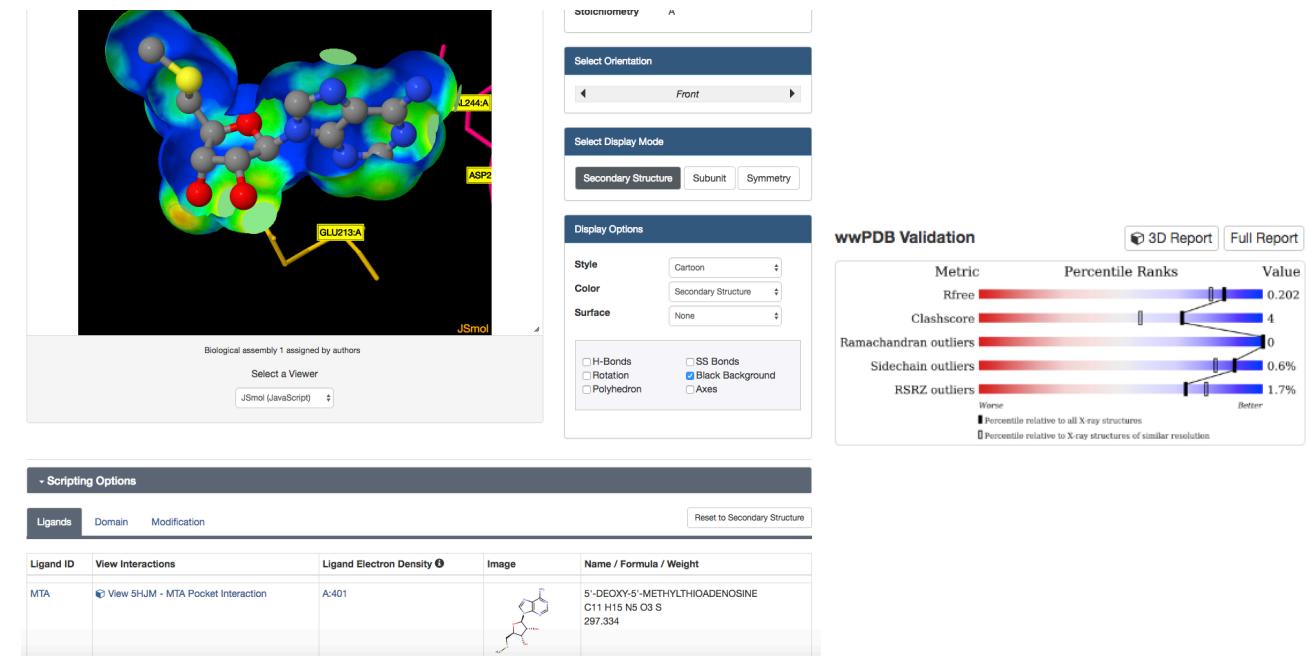


On Average, 13%
of Papers Have
Usable Data

Why is it better to have data available?



www.rcsb.org



Why is it better to have data available?

“Digitally formatted scientific data resulting from unclassified research supported wholly or in part by Federal funding should be stored and publicly accessible to search, retrieve, and analyze.”

2013 OSTP Memo

Data Management Plans

- Expected Data
- Data Formats and Metadata
- Access to Data
- Data Archiving

Why is it better to have data available?

Journals requiring complete data availability:



Journals requiring some data availability:

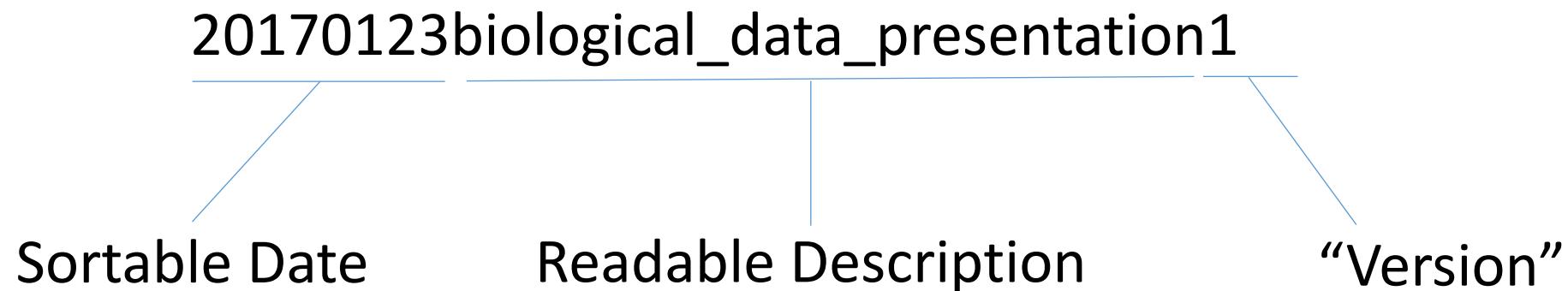


Simple Solutions

- Choose a file naming/organization scheme
- Save reasonable files
- Use reliable storage
- Think about sharing

Naming

- Trying to recreate your work months/years later is hard
- Choosing a consistent naming system makes things easier

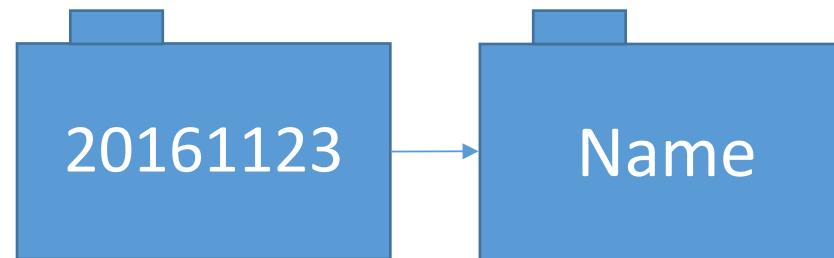


Data Architectures

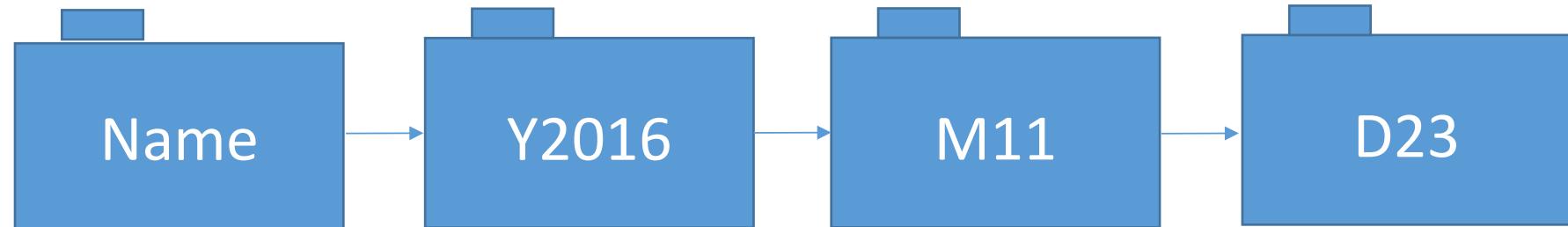
Simple



Date Based



Complex



Dataset

Automatically Manage Metadata/Documents

<https://github.com/caltechlibrary/dataset>

<https://doi.org/10.22002/D1.297>

Save Reasonable Files

- Human-readable text files are best (.txt, .csv)
- Non-proprietary files are better than proprietary
- Do analysis with scripts if possible
- Save both input and output files as space allows

Active Data Storage

- Small amounts of data (GB) are easy
- TB-scale data require planning
 - Need a system that will be reliable
 - Network-Attached Storage (Local RAID array)
 - Cloud Storage

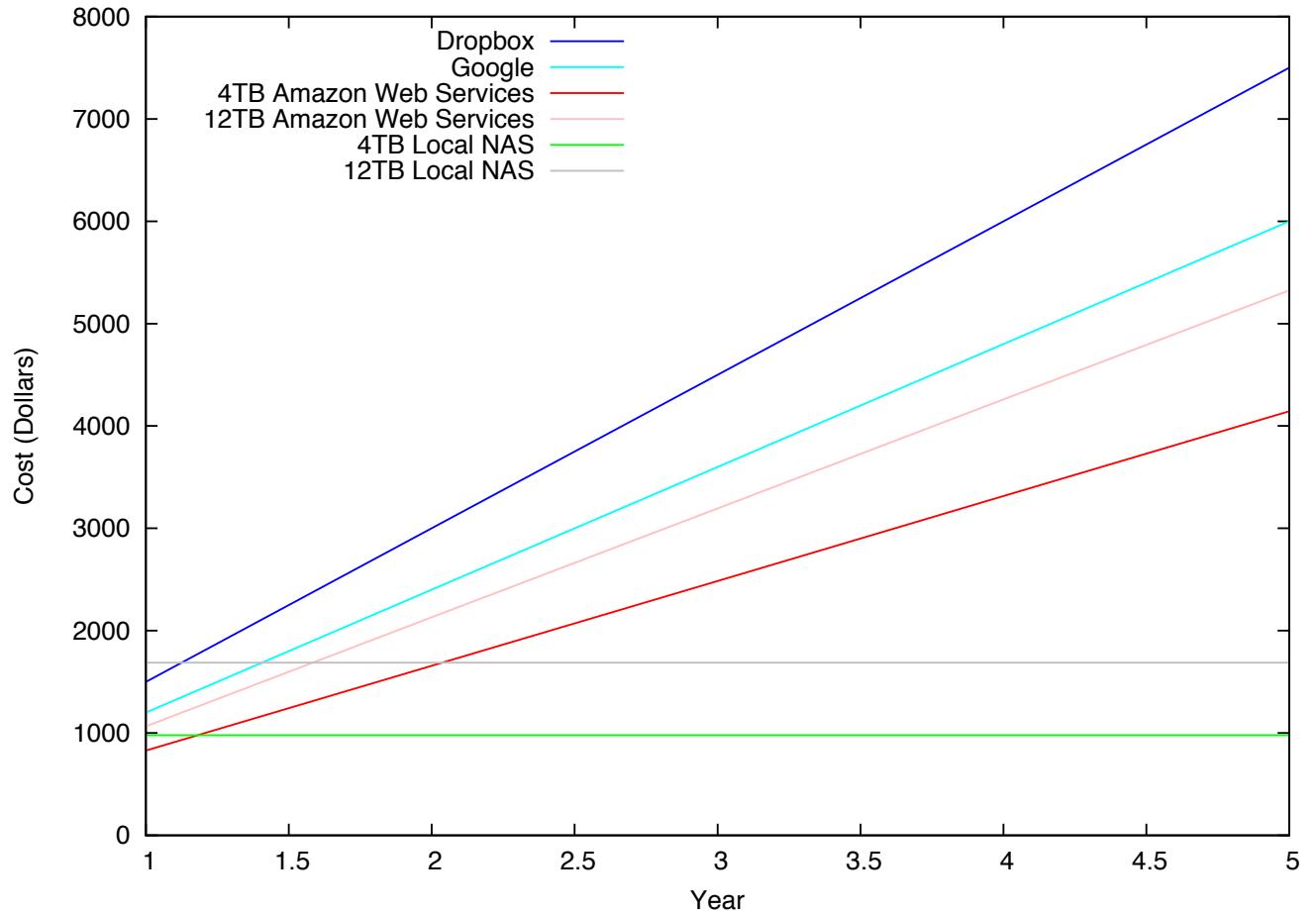
Network-Attached Storage

- Small computer with array of hard disks
- Consumer/Prosumer devices
- Low Cost (4 TB-\$425; 42 TB-\$3000)
- Need to plan space requirements
- Need to manage



Cloud Storage

- Defined or flexible storage
- Vendor Managed
- Continuous cost
- Limited by bandwidth
- Dependent on vendor



Disaster Recovery

- What Happens in a Disaster?
- Use 2 mirrored NAS units in 2 locations
- Mirror NAS to cloud storage
(Box.com - imss.caltech.edu/box)



Data Sharing

- FAIR (Findability, Accessibility, Interoperability, Reusability)
 - Subject Repositories
 - General Repositories
 - Institutional Repositories

Subject Repositories

- Protein Data Bank
- GenBank
- Wormbase
- Pangaea
- Long Term Ecological Research Data Portal
- Good listing: journals.plos.org/plosone/s/data-availability
- Thousands more: www.re3data.org



General Repositories



- Zenodo (CERN-Free)
- Dryad (NCState-\$120 per submission + Space)
- Figshare (20GB Max)
- Mendeley Data (Elsevier-Free)
- Dataverse (Harvard-Free)

CaltechDATA

- Available at data.caltech.edu
- Easy to describe and upload files
- All records get a DOI (permanent, registered link)
- Integration with Github
- API for accessing data
- Library takes care of preserving and maintaining access to files



California Institute of Technology
Research Data Repository



GitHub

Discoverability

- CaltechDATA site search
- DOIs appear in DataCite search
- and Search Engines

The screenshot shows a DataCite search results page. The search bar at the top contains 'VSWIR microimaging'. Below it, a search result is displayed for a dataset titled 'Identifying and Quantifying Mineral Abundance through VSWIR Microimaging Spectroscopy: A Comparison to XRD and SEM'. The result is attributed to Leask, Ellen K. and Ehlmann, Bethany L. The page includes navigation links for Works, People, Data Centers, and Members, along with a sign-in button. On the right side, there are filters for Resource Type (Dataset, Author), and detailed counts for each: 1 Dataset, 1 Author, 1 Ehlmann, Bethany L., and 1 Leask, Ellen K. Below the main search area, there's a sidebar for Resource Types, Publication Year (2017), and Data Centers (Caltech).

The screenshot shows a Google search results page for 'VSWIR microimaging'. The search bar at the top contains 'VSWIR microimaging'. Below it, several search results are listed, all related to VSWIR microimaging spectroscopy. One result is a PDF titled 'MICROIMAGING VSWIR SPECTROSCOPY INSTRUMENTS FOR...' by AA Fraeman, dated Sep 28, 2016. Another result is a PDF titled 'IDENTIFYING AND QUANTIFYING MINERAL ABUNDANCE...' by EK Leask, dated Mar 13, 2017. Both results are from the Caltech library.

Identifying and Quantifying Mineral Abundance through VSWIR ... - DOIs
<https://doi.org/10.22002/D1.222> ▾
Mar 13, 2017 - Identifying and Quantifying Mineral Abundance through VSWIR Microimaging Spectroscopy: A Comparison to XRD and SEM. Dataset.

Citations



California Institute of Technology
Research Data Repository

TCCON data from Caltech (US), Release GGG2014.R1

Dataset 2017-09-08 CaltechDATA

Download Edit

Details

Authors Wennberg, P. O.; Wunch, D.; Roehl, C. M.; Blavier, J.-F.; Toon, G. C.; Allen, N. T.

Contributors California Institute of Technology, Pasadena, CA (US)

Description The TCCON (Total Carbon Column Observing Network) is a network of ground-based Fourier Transform Spectrometers that record direct solar absorption spectra of the atmosphere in the near-infrared. From these spectra, accurate and precise column-averaged abundances of atmospheric constituents including CO₂, CH₄, N₂O, HF, CO, H₂O, and HDO, are retrieved. This data set contains observations from the TCCON station at the California Institute of Technology, Pasadena, USA.

Publication Date 2017-09-08

Subject(s) atmospheric trace gases, CO₂, CH₄, CO, N₂O, column-averaged dry-air mole fractions, remote sensing, FTIR spectroscopy, TCCON

DOI 10.14291/tccn.ggg2014.pasadena01.R1/1182415

Version GGG2014.R1

Format application/x-netcdf

<https://doi.org/10.14291/tccn.ggg2014.pasadena01.R1/1182415>

Related Identifier(s)

IsDocumentedBy (URL): https://tccn-wiki.caltech.edu/Network_Policy/Data_Use_Policy/Data_Description
IsDocumentedBy (URL): <https://tccn-wiki.caltech.edu/Sites>
IsPartOf (URL): <http://tccnadata.org>
IsDocumentedBy (DOI): 10.14291/tccn.ggg2014.documentation.R0/1221662
IsCitedBy (DOI): 10.5194/amt-9-683-2016
IsCitedBy (DOI): 10.5194/amt-9-227-2016
IsCitedBy (DOI): 10.5194/amt-9-3491-2016
IsCitedBy (DOI): 10.5194/amt-9-3527-2016
IsNewVersionOf (DOI): 10.14291/tccn.ggg2014.pasadena01.R0/1149162
IsPartOf (DOI): 10.14291/TCCON_GGG2014
IsCitedBy (DOI): 10.3390/rs8050414

Update Record

remote sensing

Title / Keyword Journal Remote Sensing ▾
Author / Affiliation Section all
Article Type all Special Issue all Advanced Search

Volume 8, Issue 5

Article Versions

- Abstract
- Full-Text PDF [2676 KB]
- Full-Text HTML
- Full-Text XML
- Full-Text Epub
- Article Versions Notes
- Supplementary material

Related Info

- Google Scholar

Remote Sens. 2016, 8(5), 414; doi:[10.3390/rs8050414](https://doi.org/10.3390/rs8050414) Open Access Article

Comparison of XH₂O Retrieved from GOSAT Short-Wavelength Infrared Spectra with Observations from the TCCON Network

Eric Dupuy ^{1,*}, Isamu Morino ¹, Nicholas M. Deutscher ^{2,3}, Yukio Yoshida ¹, Osamu Uchino ¹, Brian J. Connor ⁴, Martine De Mazière ⁵, David W. T. Griffith ², Frank Hase ⁶, Pauli Heikkinen ⁷, Patrick W. Hillyard ^{8,9}, Laura T. Iraci ⁸, Shuji

48. Wennberg, P.O.; Wunch, D.; Roehl, C.; Blavier, J.F.; Toon, G.C.; Allen, N. TCCON Data from California Institute of Technology, Pasadena, California, USA, Release GGG2014R1; Carbon Dioxide Information Analysis Center; Oak Ridge National Laboratory: Oak Ridge, TN, USA, 2014. [Google Scholar] [CrossRef]

<https://doi.org/10.3390/rs8050414>

Citation

Email Alert

California Institute of Technology
Research Data Repository

Dear Paul Wennberg,

Your CaltechDATA work "TCCON data from Caltech (US), Release GGG2014.R1" has been cited in:

1. Dupuy E, Morino I, Deutscher N, et al. Comparison of XH₂O Retrieved from GOSAT Short-Wavelength Infrared Spectra with Observations from the TCCON Network. *Remote Sensing*. 2016;8(5):414. doi:10.3390/rs8050414.

This link has been added to your CaltechDATA record at [10.14291/tccn.ggg2014.pasadena01.R1/1182415](https://doi.org/10.14291/tccn.ggg2014.pasadena01.R1/1182415).

Best,

CaltechDATA Alerting Service

Is this incorrect? Let us know at data@caltech.edu

This email was sent by the Caltech Library, 1200 East California Blvd., MC 1-43, Pasadena, CA 91125, USA

[unsubscribe](#)



California Institute of Technology
Research Data Repository

Demo

Use Cases

Thesis Preparation

<https://doi.org/10.7907/Z9NC5Z7H>



Engineered Viral Vectors and Developed Tissue Clearing Methods for Single-cell Phenotyping in Whole Organs

Citation

Chan, Ken Yee (2017) Engineered Viral Vectors and Developed Tissue Clearing Methods for Single-cell Phenotyping in Whole Organs. Dissertation (Ph.D.), California Institute of Technology. doi:10.7907/Z9NC5Z7H. http://resolver.caltech.edu/CaltechEScholarship/3030?utm_id=17222&utm_17

A central question in biology is how different cell types interact with each other and their native environment to form complex functional systems and networks. Although our ability to investigate this question has considerably expanded from the development of genetically encoded tools, some limitations still persist. For example, it is challenging to efficiently deliver transgenes to specific cell types in whole organs. Additionally, it is challenging to efficiently deliver transgenes into difficult-to-target areas through direct injection or injection via a catheter. These challenges are compounded by the complexity of the tissue system, which limits our ability to extensively study these areas. Therefore, tools and methods that overcome these limitations are highly sought after. In recent years, researchers have been developing tissue clearing technologies to render whole organs transparent for optical interrogation and characterizing viral cascades and engineering viral vectors for non-invasive gene delivery.

Tissue clearing techniques for three dimensional optical interrogation were invented over a century ago. However, these earlier methods used harsh organic chemicals and failed to retain the tissue's native fluorescence. In the early 1990s, researchers developed a technique called CLARITY that used newly generated transgenic mouse lines that allowed for cell type-specific expression of fluorescent transgenes or to express fluorescent proteins in all cells. This technique was later improved by the addition of a fixative agent, which addressed these limitations by further developing and standardizing a tissue clearing method that utilizes the vesicular to perfuse clearing reagents. This technique, called perfusion-assisted agent release *in situ* (PARI), enables the preservation of (i) clearing of soft tissue, (ii) preservation of native fluorescence, and (iii) preservation of epitopes compatible with IHC.

Data files uploaded
during writing process



California Institute of Technology
Research Data Repository

- <https://doi.org/10.22002/D1.234>
- <https://doi.org/10.22002/D1.235>
- <https://doi.org/10.22002/D1.236>
- <https://doi.org/10.22002/D1.237>

Software in CaltechDATA

<https://doi.org/10.22002/D1.218>



Tau -- a lightweight tool for specifying and verifying tiny automata models

Software
2017-02-27
CaltechDATA

Download

Edit

Details

Authors

Holzmann, Gerard Department of Computing & Mathematical Sciences, Caltech

Description

Abstract:
Tau is a small Tcl/Tk application that can be used to quickly specify and formally verify small automata models (the name 'tau' is short for 'tiny automata'). It is used as a teaching aid in CS118, a course on the formal verification of asynchronous software systems using logic model checking. Tau requires the availability of a standard C compiler (e.g., gcc) and a recent version of the Spin model checker (e.g., Version 6.4.3 or later) as background tools.

<https://doi.org/10.22002/D1.240>



tmorrell/caltechdata_plot: Update AWS Documentation

Software
2017-05-24
CaltechDATA

Download

Edit

Details

Authors

Tom Morrell Caltech

Description

Abstract:
Includes new documentation with improved AWS setup.
Other:
Demo interactive plotting tool that uses Bokeh server to produce an interactive plot by calling the caltechDATA (Invenio 3) API

Use Cases



New Results

An allosteric theory of transcription factor induction

Manuel Razo-Mejia, Stephanie L. Barnes, Nathan M. Belliveau, Griffin Chure, Tal Einav, Rob Phillips
doi: <https://doi.org/10.1101/111013>

This article is a preprint and has not been peer-reviewed [what does this mean?].

Abstract Info/History Metrics Supplementary material Preview PDF

Abstract

Allosteric molecules serve as regulators of cellular activity across all domains of life. We present a general theory of allosteric transcriptional regulation that permits quantitative predictions for how physiological responses are tuned to environmental stimuli. To test the model's predictive power, we apply it to the specific case of the ubiquitous simple repression motif in bacteria. We measure the fold-change in gene expression at different inducer concentrations in a collection of strains that span a range of repressor copy numbers and operator binding strengths. After inferring the inducer dissociation constants using data from one of these strains, we show the broad reach of the model by predicting the induction profiles of all other strains. Finally, we derive an expression for the free energy of allosteric transcription factors which enables us to collapse the data from all of our experiments onto a single master curve, capturing the diverse phenomenology of the induction profiles.

<https://doi.org/10.1101/111013>

HOME | ABO

Search



California Institute of Technology
Research Data Repository

<https://doi.org/10.22002/D1.224>
<https://doi.org/10.22002/D1.227>
<https://doi.org/10.22002/D1.228>
<https://doi.org/10.22002/D1.229>

Paper Website
on GitHub

The screenshot shows a light blue header with the Caltech logo and a navigation menu with links to "ABOUT", "ANALYSIS", "DATA", "PEOPLE", and "ACKNOWLEDGEMENTS". Below the menu, it says "Philips Lab · GitHub Repo". The main content area has a light gray background with a green oval containing handwritten mathematical notes. Above the oval, a blue arrow points upwards with the text "Data Files". The notes in the oval contain the following equation:

$$\text{Fold-Change} \approx \left(\frac{R}{R + K_{dR}} \frac{e^{-K_{dP} P}}{N_{RS}} \right)^{-1}$$

The text below the notes reads:

An Allosteric Theory of Transcription Factor Induction

This website serves as a record for the experimental and theoretical work described in the publication "An Allosteric Theory For Transcription Factor Induction" by Manuel Razo-Mejia*, Stephanie Barnes*, Nathan Belliveau, Griffin Chure, Tal Einav*, and Rob Phillips (*contributed equally).

The paper can be found on the [bioRxiv](#) and [arXiv](#). You can download PDFs of the current version and the supplementary information below.

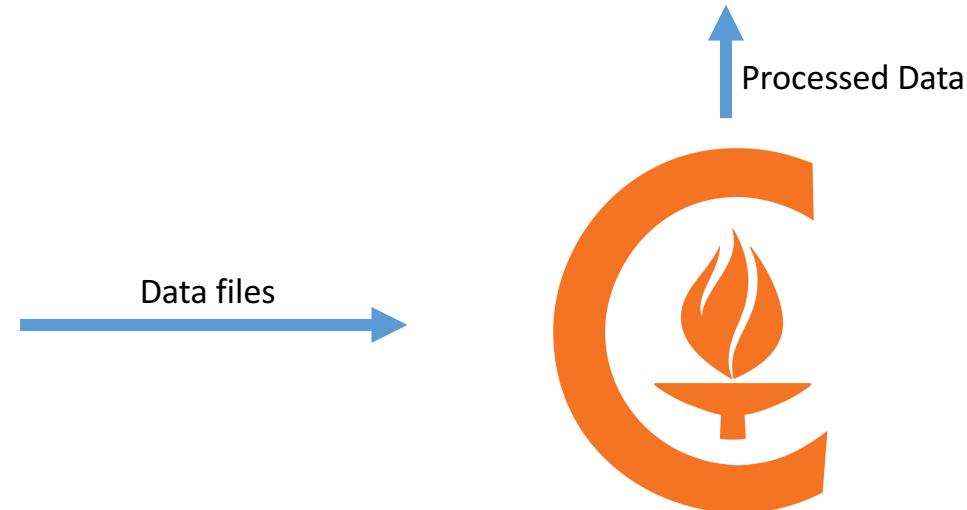
- Main Text
- Supplementary Information

https://rpgroup-pboc.github.io/mwc_induction
<https://doi.org/10.22002/D1.299>

Use Case - TCCON



Total Carbon Column Observing Network (TCCON)
29 Data Collection Sites Around the World



Use Case - TCCON



tccon.ornl.gov

TCCON Data Archive

HOME GGG2014 GGG2012 GGG2009

Total Carbon Column Observing Network (TCCON)

The TCCON Data Archive

TCCON is a network of ground-based Fourier Transform Spectrometers recording direct solar spectra in the near-infrared spectral region. From these spectra, accurate and precise column-averaged abundances of CO₂, CH₄, N₂O, HF, CO₂, H₂O and HDO are retrieved. The HF and HDO retrievals are uncalibrated and hence preliminary. Data are updated monthly on the first of the month. The data become publicly available no later than one year after the measurements are recorded, and many sites choose to release their data much sooner.

For the latest TCCON information, please visit the [TCCON Wiki](#). For citation information and our data policy, please see our [Data Use Policy](#). For site-specific information and data analysis descriptions, please read the [Data Description](#). Auxiliary data (column averaging kernels, a priori profiles) are included in the netCDF files provided below. Information on how to use our column averaging kernels and a priori profiles can be found on our [Auxiliary Data](#) page.

A technical report describing the GGG2014 TCCON data version can be found on the [documentation](#) page. Our telluric line list can be downloaded from the [atm](#) page. Our solar line list can be downloaded from the [solar](#) page. A program to generate our a priori profiles can be downloaded from the [a priori](#) page. Please note that the a priori profiles used in the TCCON retrievals are included in the data files below. If you need to produce TCCON a priori profiles for locations and times where there are no TCCON measurements, please use the program linked above.

The TCCON is closely affiliated with the Network for the Detection of Atmospheric Composition Change Infrared Working Group (NDACC-IRWG). In contrast with TCCON, which produces column-averaged dry-air mole fractions, the NDACC produces vertical profiles of the concentrations of many of the same gases and several others. The NDACC website and links to their database can be found at [www.ndacc.edu/irwg](#).

[Sign up to the TCCON Users email list to get email updates on TCCON data releases.](#)
Note that the website is self-signed; you can safely add an exception.

[Login for TCCON Partners](#)

Private data files

Sites

Ascension Island

- [@ae20120522_20120831.nc](#)
- [@ae20130317_20130618.nc](#)
- [@ae20130911_20131229.nc](#)
- [@ae20140108_20140716.nc](#)
- [@ae20140717_20141019.nc](#)
- [@ae20141021_20141231.nc](#)
- [@ae20150101_20150310.nc](#)
- [@ae20150311_20150409.nc](#)
- [@ae20150410_20150630.nc](#)
- [@ae20150701_20150926.nc](#)
- [@ae20151005_20151218.nc](#)

Public data files

Index of /2014Public/ascension01

Name	Size	Date Modified
[parent directory]		
README.txt	11.8 kB	10/20/14, 5:00:00 PM
ae20120522_20161221.public.nc	10.1 MB	5/31/17, 5:25:00 PM

Automatically released 1x/month

Departmental Server at Caltech

Login for TCCON Partners TCCON Data Archive GGG2014 GGG2012 GGG2009

Total Carbon Column Observing Network (TCCON)



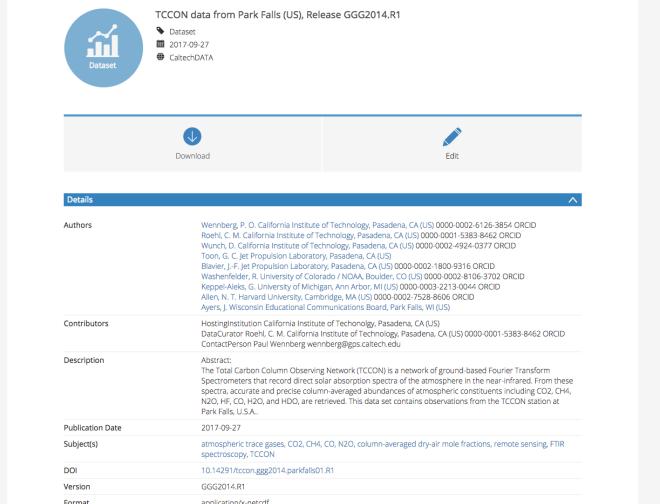
TCCON is a network of ground-based Fourier Transform Spectrometers recording direct solar spectra in the near-infrared spectral region. From these spectra, accurate and precise column-averaged abundances of CO₂, CH₄, N₂O, HF, CO₂, H₂O and HDO are retrieved and reported here. A technical report describing the retrievals is found [here](#) and telluric spectral line lists used in the retrievals are publicly available.

Data in netCDF format are publicly available no later than one year after the spectra are recorded; many sites release their data earlier. Citation and data use requirements are included in the license associated with each record. Column averaging kernels and a priori profiles are included in the files. Information on how to use these can be found [here](#). To produce TCCON a priori profiles for locations and times where there are no TCCON measurements, a stand-alone program can be [downloaded](#).

[Sign up to the TCCON Users email list to get email updates on TCCON data releases.](#)

[tccondata.org](#)

CaltechDATA



<https://doi.org/10.14291/tccn.ggg2014.parkfalls01.R1>

Demo – CaltechDATA + Interactive Plotting



Orange-brown jasper spectrum
Dataset
2017-02-16
CaltechDATA



Download



Edit



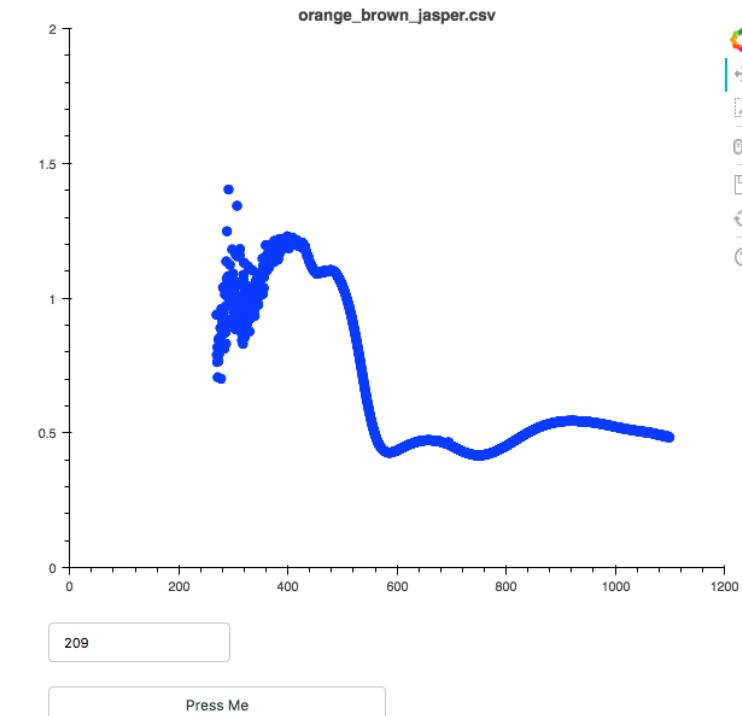
Red jasper spectrum
Dataset
2017-02-14
CaltechDATA



Download



Edit



plots.caltechlibrary.org

doi.org/10.22002/D1.240

Caltech Library Data Management Services

- Want to chat about data issues?
- Data management plan development
- Consultations on storage technologies or file organization

data@caltech.edu

tmorrell@caltech.edu

626-395-3827

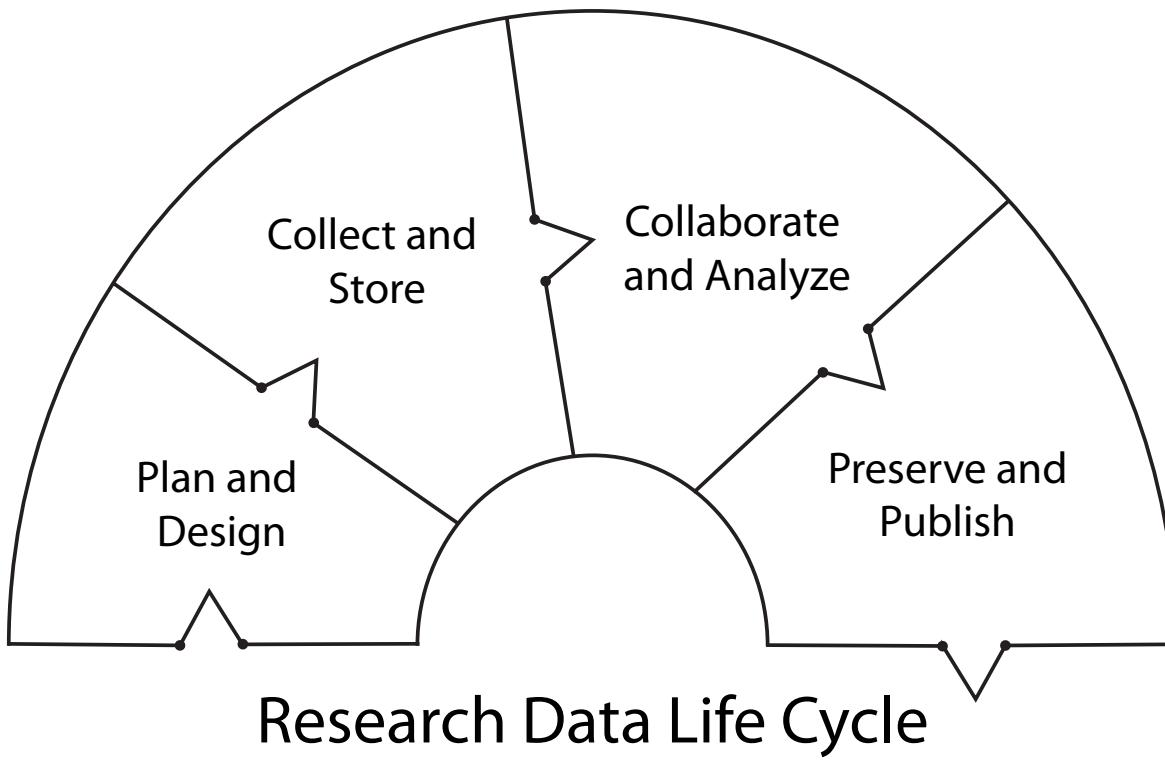
Research Software Workshop

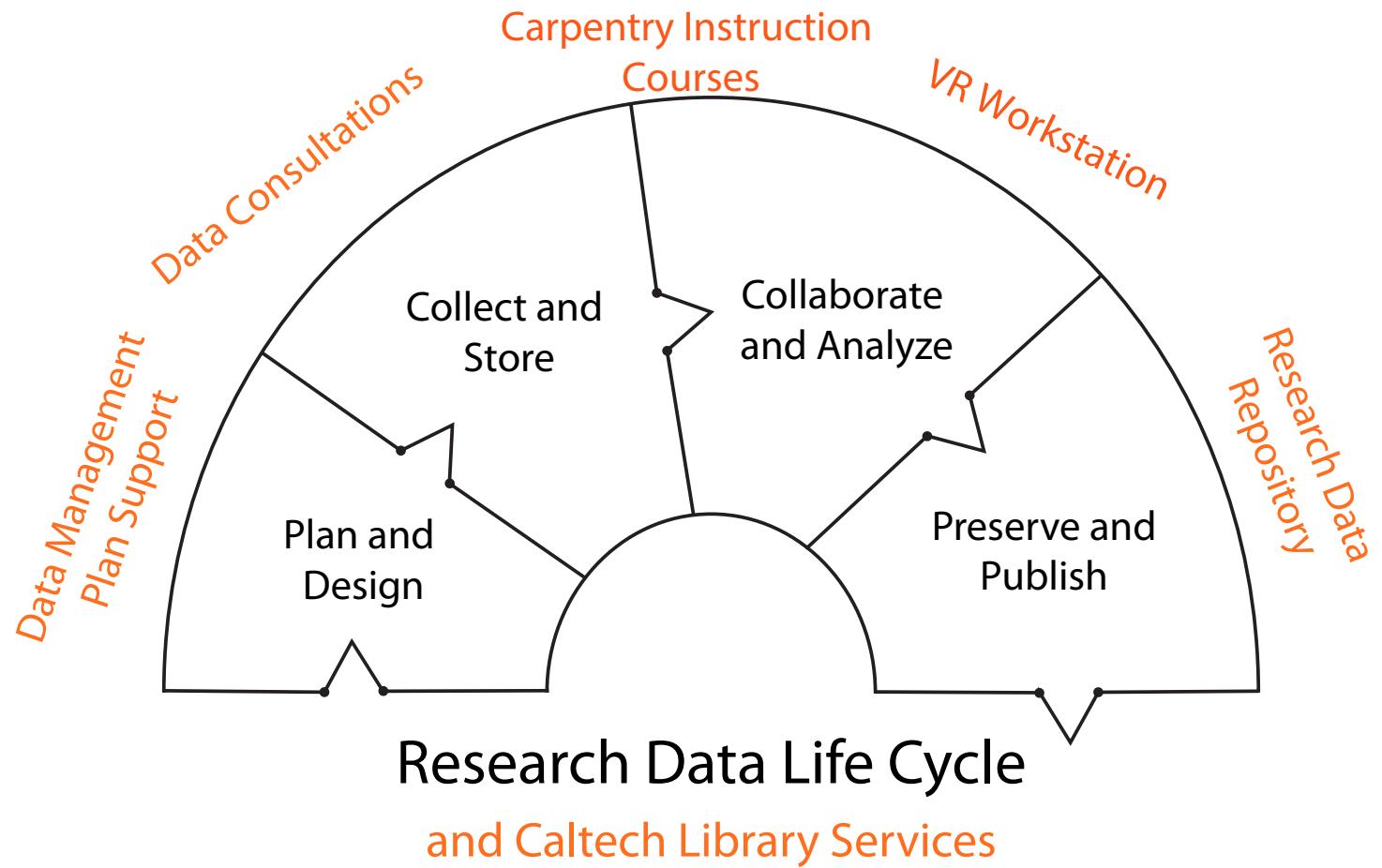
Dan Katz (NCSA)

January 31st 12-2

Lunch Provided

Registration Required





Things to Think About

- Choose a file naming/organization scheme
- Save reasonable files
- Use reliable storage
- Think about sharing