Practical Tools For Research in Economics: Lessons From R4R Program

> Yun-Ting Yeh (R4R 2023 Cohort) Department of Economics Nov 30, 2023

0

# Outline



RESEARCH, INNOVATION & IMPACT Data Science Institute



Arizona Institute for Resilience



### **Brief Introduction:**

- Roots for Resilience (R4R) program
- Foundational Open Science Skills (FOSS) workshop
  - Version Control: Git, GitHub
  - Documentation: GitHub Pages

### **Useful Tools for Research:**

- CyVerse
- Others (Data Visualization Tools, Al-powered Search)

## Roots for Resilience (R4R) Program

"The Roots for Resilience Program provides training and support to select graduate students on open, reproducible science and computational infrastructure to enhance research focused on resiliency in the environment."

- This program is led by Arizona Institute for Resilience (AIR), CyVerse, and the Data Science Institute (DSI)
- Cohort consists of 10 graduate students from various diciplines in UA
- Provides \$7,000 fellowship with requirements of weekly meetings (FOSS lectures, DSI workshops, discussions, etc.), learning journals, capstone project, and department presentation (5-10 hours/week)
- Next cohort is Fall 2024. Check out <a href="https://datascience.arizona.edu/r4r">https://datascience.arizona.edu/r4r</a> if interested!

## Foundational Open Science Skills (FOSS) Workshop

Week 0: The Unix Shell, Git and Github

Week 1: Introduction to Open Science

- Six Pillars of Open Science: Open Access Publications, Open Data, Open Educational Resources, Open Methodology, Open Peer Review, Open Source Software
- Open data does not necessarily mean publicly available
- FAIR & CARE principles

#### Week 2: Managing Data

- Data Management Plan
- Licenses

#### Week 3: Intro to Project Management

Governance Operation document

#### Week 4: Documentation & Communication

Build GitHub Pages website using MkDocs

#### Week 5: Version Control

• Git, GitHub

#### Week 6,7: Repeatibility

Containers, Docker, GitHub Codespaces

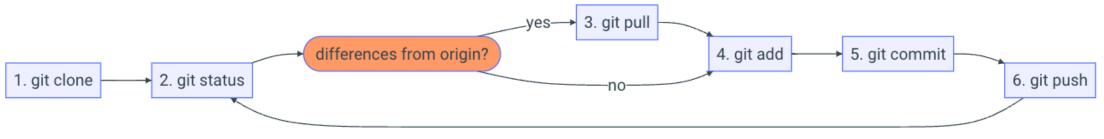


## Version Control

- Keeping track of changes you made to your files can be helpful
  - Efficient version control systems are valuable...
  - Even more valuable for collaborative work!
- Git: a command-line program for version control
  - Once you make a repository ('folders'), Git keeps track of changes you made to files in that repository and stores the changes in .git folder
  - Stores only the *differences* between different versions of files, instead of saving the whole files
  - Can be used on local machines alone
- **GitHub**: a cloud-based hosting service
  - Can remotely host your Git repositories
  - Can serve as a backup of the repositories, facilitate collaboration, ... etc.

## **Version Control**

- Basic commands (locally):
  - In Terminal, go to your folder: cd folder\_name
  - Add changes to keep track of: git add file\_name
  - Commit all added changes: git commit -m "message for this change"
  - Check status: git status
- Basic cycle (involving remote repositories):



Source: https://foss.cyverse.org/05\_version\_control/

• More info on Git and GitHub: <a href="https://foss.cyverse.org/05\_version\_control/">https://foss.cyverse.org/05\_version\_control/</a>

## Example: GitLab

♥ □ + ₩	yunting1121 > Payment Adoption > Commits			
	master v payment-adoption		Author ~ Search by message	
Q Search or go to	Feb 25, 2021			
Project P Payment Adoption	analyze relationship between ind payment adoption and merce yunting1121 authored 2 years ago	hant acceptance	936afb58 🛱	
🛠 Pinned 🗸	analyze individual payment first usage		b81be9e1 🛱 🗁	
lssues	yunting1121 authored 2 years ago			
Merge requests 0	Feb 24, 2021			
Ba Manage >	drop observations within a week of the first observation yunting1121 authored 2 years ago		yunting1121 > Payment Adoption > Commits > 383265cc	
🗇 Plan >	yunting1121 authored 2 years ago	D 11 C	Commit 383265cc $[^{e_1}_{\mathbb{C}}$ authored 2 years ago by $\bigoplus$ yunting1121	Browse files Options ~
Code ~ Merge requests 0	add sample definition; drop cashless subsample definition yunting1121 authored 2 years ago	Q Search or go to	add sample definition; drop cashless subsample defir	nition
Repository	create individual transaction level payment usage panel data yunting1121 authored 2 years ago	Toject		
Branches	yunting1121 authored 2 years ago	🖈 Pinned 🗸	parent 6a02ff87  Branches > Branches containing commit	
		Issues		
		Merge requests 0	No related merge requests found	
		යි Manage >		
		➡ Plan >	Changes 1	
		Code	Showing 1 changed file $\checkmark$ with 23 additions and 22 deletions	Show whitespace changes Inline Side-by-side
		Merge requests 0	✓ ➡ coding/coding_ind_pmt_panel.R C	+23 -22 💭 View file @ 383265cc
		Repository	00 -117,25 +117,26 00 rm(Data_dum, Data_est, period_list)	00 -117,25 +117,26 00 rm(Data_dum, Data_est, period_list)
		Branches	<pre>117 mutate(first_use3 = row_number() == 1) %&gt;% ungroup 118</pre>	<pre>117 mutate(first_use3 = row_number() == 1) %&gt;% ungroup 118</pre>
		Commits	119 120 - # 3. Extract individuals who use cashless after policy for the	119 120 + # 3. Propoerly define sample
		Tags Repository graph	first time 121 - # create after policy indicator	121 + # check data starting date for individuals
		Compare revisions	122 - df <- df %>% mutate(after = ifelse(date >= "2019-10-01", 1, 0))	122 + id_first_ob <- df %>% select(id, date) %>% arrange(id, date) %>%
		Snippets	123	123 + group_by(id) %>% filter(row_number() == 1) %>% ungroup 124
			124 - # summarise individual first usage before/after policy 125 - id_usage3 <- df %>% group_by(id, after, pmt3) %>%	125 + # latest date of entering sample 126 + # "2019-09-25"
		♥ Secure >	<pre>126 - summarize(use = sum(first_use3)) %&gt;% ungroup %&gt;% 127 - right_join(crossing(id = id_list, pmt3 = levels(df\$pmt3)))</pre>	127 + max(id_first_ob\$date)
		Deploy >	128 129 - # extract individuals who didn't use cashless before policy but	128 129 + # check month distribution
			127 - # extract individuals who didn't use cashtess before policy but	TEA . W CHOCK MOLEN GIGET IDOLION

### Documentation

- Project documentation can be valuable for others to understand our approaches
  - Even more valuable for future ourselves (and our collaborators)!
  - Do you think you can understand your codes and data one year from now? five years from now?
- Effective documentation must consider the followings:
  - Clarity, Completeness, Accuracy, Organization, Relevance
- Public documentation: GitHub Pages
  - Can render website using MkDocs
  - Can create website context simply using Markdown
  - Your website will be automatically updated every time you commit a change
  - Hands-on tutorial: <u>https://foss.cyverse.org/04\_documentation\_communication/#hands-on-building-a-github-pages-website-using-mkdocs</u>

## Example: GitHub Pages Website

#### • Website created for FOSS Capstone Project

🗳 🛛 Project Website		Q Se	arch	ি foss-capstone ৫০ ¥০		
Project Website Home Data Management Plan Governance Operations	Data Management Data Management Plan This document records data manage Firms: Water Pollution and Beverage dataset used in the project. This doc proceeds.	ement plan for my 2 <sup>nd</sup> year project, "C Manufacturers' Entry". It currently fo	cuses on the main nges as the project	On this page         Data Management Plan         1. Data Description         2. Data Collection         3. Data Storage and Backup         4. Data Organization and Documentation         5. Data Preservation and Long-Term Access         6. Data Sharing and Access         7. Data Storage and Preservation and Long-Term Access         6. Data Sharing and Access         7. Data Storage and Preservation and Long-Term Access         6. Data Sharing and Access	Q Type () to search Settings	
	<ol> <li>Data Description</li> <li>Historial Business Data is provided</li> <li>Data includes company information</li> </ol>	Files	foss-capstone / docs / Data_Management_Plan.md □			99c1ba0 · last month 🕥 <b>History</b>
	2021. The data is stored in csv files.	<ul> <li>Q Go to file (t)</li> <li>&gt;</li></ul>	1 ~ ## Data Managem 2  3 4 This document rr 5 6  7 8 #### 1. Data De	ent Plan ecords data management plan for my 2nd year scription		Raw [] 2 P v C

# CyVerse



Website: https://cyverse.org/

Learning Center: <a href="https://learning.cyverse.org/">https://learning.cyverse.org/</a>

Self-guided Course: <a href="https://cyverse-learning-materials.github.io/cyverse\_mooc/">https://cyverse-learning-materials.github.io/cyverse\_mooc/</a>

- CyVerse is a powerful computational infrastructure built to handle huge datasets and complex analyses
- Offer subscriptions with different tiers (including free tier)
- Main features
  - Data Store: cloud-based data storage
  - Discovery Environment: platform with simple web interface for managing and sharing data or running analyses
  - Classrooms...

## CyVerse: Advantages

- UA students have *free* Pro account
  - 3 TB of data storage
  - 20,000 compute units (core hours on CPU)/year
  - Access to GPU
- Easy data transfer
- Easy data sharing with other users
- Less wait on using computation resources
- Can keep the account even after graduation! (May need to pay for subscription if not using free account)

# CyVerse: Sign up

- Create account at <u>https://user.cyverse.org/signup</u> using UA email (to get free Pro account)
- Go to user portal

<image/>	CYVERSE USER PORTAL	PHELP
Welcome to the CyVerse User Portal The User Portal is the place to customize your CyVerse experience. Request acce services and explore learning materials and workshops.	Services	
Welcome to the CyVerse User Portal The User Portal is the place to customize your CyVerse experience. Request acce services and explore learning materials and workshops.		
CyVerse User Portal The User Portal is the place to customize your CyVerse experience. Request accel services and explore learning materials and workshops.		Welcome to the
services and explore learning materials and workshops.		CyVerse User Portal
		The User Portal is the place to customize your CyVerse experience. Request access to services and explore learning materials and workshops
LEARNING CENTER GETTING STARTED WEBINAR		
CLOSE AND DON'T SHOW AGAIN		LEARNING CENTER GETTING STARTED WEBINAR
My Services	My Services	
My Services		CLOSE AND DON'T SHOW AGAIN
	Discovery Environment Use hundreds of bioinformatics apps and manage data in the	CLOSE AND DON'T SHOW AGAIN

## CyVerse: Discovery Environment

Click 'launch' on Discovery Environment or go to <u>https://de.cyverse.org/</u>

≡	
<ul> <li>↑</li> <li>↑</li></ul>	Evolution CyVerse
	CyVerse is ready for its next Evolution. While we have made great strides in diversifying our funding, over the next two years, we will explore more possibilities with Phoenix Bioinformatics, a pioneer in helping research infrastructures define their futures. We look forward to sharing our progress, but in the meantime, we would value your thoughts via this survey.
∑ Solo	Tour the Discovery Environment
	Welcome ytyeh to CyVerse! Would you like a quick tour of the Discovery Environment? Dismiss Take a Tour Resource Usage
?	Your current subscription tier is Pro. Effective Period: 2023-09-18 23:00:41 - 2024-09-18 23:00:41
	Data Storage Using 0.98% as of 2023-11-29 18:55:34CPU Consumption Using 82:82% as of 2023-11-29 02:11:30 $_{768.0 \text{ GiB}}$ $_{1.5 \text{ TiB}}$ $_{2.3 \text{ TiB}}$ $_{10000 \text{ Core Hours}}$ $_{10000 \text{ Core Hours}}$ $_{10000 \text{ Core Hours}}$ $_{10000 \text{ Core Hours}}$ $_{20000 \text{ Core Hours}}$ $_{10000 \text{ Core Hours}}$ $_{10000 \text{ Core Hours}}$

## CyVerse: Data Store

• Click the second icon on the left panel

					Data 👻	■ 👌 🐥 Y
♠	n tytyeh ◄				C Refresh Folder	Tupload
•		Name 🛧	Last Modified	Size		Customize Columns
		analyses	2023-09-19 12:00:32	-		* * *
		smog-check	2023-09-26 14:27:13	-		
>_			$\langle \cdot \rangle$	1 > 100 -		
ର୍						
盦						
\$						
?						

## CyVerse: Data Store

 Can easily share files/folders with other users (You'll need to know their username) and give permissions (Read/Write/Own)

≡	<b>O</b>				Data 👻	8 🕁	99+	Y
♠	<b>f</b> ytye	eh ✔ ▶ 🖿 smog-check		C Refresh Details	I Folder 🗍 🚹 Upload 👻 📑 Add	i To Bag <	Share	More Actions
¢		Name 🛧	Last N	lodified	Size		Custor	mize Columns
		Coded-data	2023-	11-23 18:48:23	•			:
		Sharing					×	:
▶_		-						:
] ð		Search By CyVerse username, email, or group name						:
		Who Has Access		Resources				:
		User Permission		coded-data				
\$		mikeda@arizona.edu		/iplant/home/ytyeh/smog-check/coded-data				:
?		Username: mikeda University of Arizona	Write -			:		
								:
						Cancel	Done	:
		out_reg_fail_all_smoke.tex	2023-	10-07 18:24:04	1.9 KiB			:
		out_reg_fail_all_vis_int_temp.tex	2023-	10-16 12:39:38	1.9 KiB			:
		out_reg_fail_all_vis.tex	2023-	10-07 18:25:48	1.9 KiB			:
		out_reg_fe_cost_obd.tex	2023-	10-25 15:15:06	928 B			:
			×	1 > 100 -				

## CyVerse: Data Store

• Click 'Shared With Me' and find the user that shares data with you

≡				Data	<b>-</b> ?	🕹 🔑 Y	
A	Shared W	/ith Me ▾			C Refresh	1 Details	Add To Bag : More Actions
		Name 🛧	Last Modified	Size			Customize Columns
		Imcguire	2018-10-23 18:48:11	-			0 0 0
		lwang	2022-03-01 18:31:59				6 0 0
		maizecode	2015-10-13 22:03:19	-			0 0 0
		mian	2023-03-13 10:58:29	-			0 0 0
ର୍		mikeda	2023-09-22 17:54:38	-			:
Ê		moontree1985	2017-11-15 12:39:18	-			*
\$		mtruncatula	2014-08-18 08:55:36	-			:
?		mwoodhouse	2022-02-18 05:06:51	-			* * *
		nevp	2022-09-09 10:59:25	-			* * *
		nimkulip	2014-05-07 13:40:30	-			* • •
		nirav	2022-07-09 10:27:14	-			*
		nmiller	2022-01-24 17:37:19	-			* *
		nowlanf	2022-11-09 08:38:02	-			* • •
		okamuralab	2019-04-02 19:46:01	-			*
			< 1 2 > 100 -				

# CyVerse: Apps

- Various apps to choose from for analyses, including interactive ones
- Ex. Command line Interface, R, Python (probably not Stata...)

♠	Featured Apps	<b>~</b>	Filter	✓ Manage Tools ► Instant Launch	es + Create
		Name 🛧	Туре		
		Cloud Shell	VICE		* * *
<b>J</b>		JupyterLab Datascience	VICE		0 0 0
		JupyterLab Earthlab	VICE		0 0 0
		JupyterLab Geospatial	VICE		* *
<b>o</b>		JupyterLab QIIME2	VICE		0 0 0
		Rocker RStudio Geospatial	VICE		0 0 0
\$		Rocker RStudio Stan	VICE		0 0 0
?		Rocker RStudio Verse	VICE		0 0 0
		Ubuntu Desktop	VICE		0 0 0
		Ubuntu Desktop GIS	VICE		0 0 0
		VS Code	VICE		0 0 0
			•		

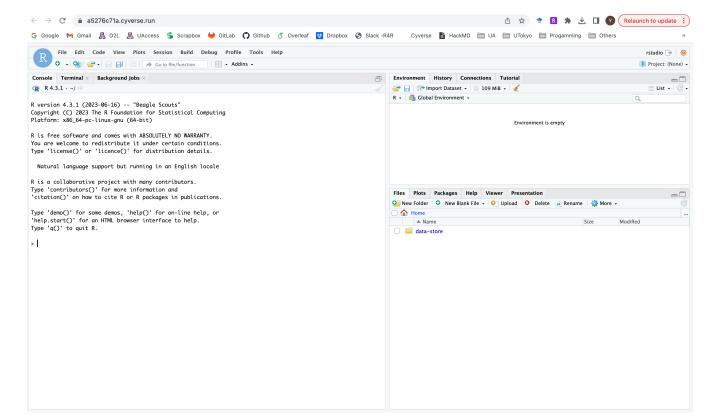
## CyVerse: Apps

• Click the app, set maximum cores you need, and click launch

				Analyses 👻	?	ŵ	
♠	Rocker_RStudio_Verse_analysis I Submitted - 2023-11-29 21:24:4 2	Go to Analysis	C Refresh Request Help : More Actions				
	Analysis ID	63f3a3	48-8f38-11ee-b26e-008cfa5ae621				
		Сору А	nalysis ID				
	Арр	Rocker	RStudio Verse				
≻_	Output folder	/iplant/	nome/ytyeh/analyses/Rocker_RStudio_Verse_analysis1-2023-11-30-04-24-42.7 📭				
	Start date	2023-11	-29 21:24:42				
â	End date	-					
	Owner	ytyeh					
\$	Status History						^
?	1: Interactive - Submitted						 ~
	Parameters						^
	Name $\downarrow$		Туре	Value			
	* This page will refresh automatically. * You c	an bookmark this page to	come back to it later.				

## CyVerse: Apps

- Launch an R Studio session on your browser!
- Can access Data Store to load/save your data and codes
- Will need to install packages every time you open a new session...



## CyVerse: Analyses

#### • Can check or terminate analyses here

≡	•					Analyses 👻	? 👉	99 <del>+</del> (	Y
♠	View -	nalyses 👻 Filter 👻			C Refresh S Terminate	🚺 Details	ld To Bag	Share	: More Actions
<b>P</b>		Name	Status	Start date $\psi$	Duration				
		Rocker_RStudio_Verse_analysis1	Completed	2023-11-29 21:24:42	2 minutes		•		
		Rocker_RStudio_Geospatial_analysis1	Completed	2023-11-29 02:06:19	5 minutes		▶ # 0		
		Rocker_RStudio_Verse_analysis1	Completed	2023-11-29 01:55:29	11 minutes		▶ # 0		
<u>&gt;_</u>		Rocker_RStudio_Geospatial_analysis1	Completed	2023-11-28 21:13:14	about 5 hours		▶ # 0		
୍ଦ୍	$\checkmark$	Rocker_RStudio_Verse_analysis1	Running	2023-11-28 12:12:03	1 day		⊗ = ⊠	8 🕕	
		Rocker_RStudio_Verse_analysis1	Completed	2023-11-28 12:01:16	about 9 hours		▶ # 0		
血		Rocker_RStudio_Verse_analysis1	Completed	2023-11-28 11:31:14	41 minutes		🖻 🎞 🕕		
\$		Rocker_RStudio_Geospatial_analysis1	Completed	2023-11-27 12:35:19	about 7 hours		🖻 🎞 🕕		
?		Rocker_RStudio_Verse_analysis1	Completed	2023-11-27 11:51:37	1 day		🖻 🎞 🕕		
		Rocker_RStudio_Verse_analysis1	Completed	2023-11-26 17:59:37	about 4 hours		🖻 🛱 🕕		
		Rocker_RStudio_Verse_analysis1	Completed	2023-11-24 15:10:13	about 21 hour	S	🖻 🎞 🕕		
		Rocker_RStudio_Verse_analysis1	Completed	2023-11-23 18:37:44	38 minutes		◘ ♯ 0		
		Rocker_RStudio_Verse_analysis1	Completed	2023-10-26 13:17:27	about 4 hours		🖻 🛱 🚯		
		Rocker_RStudio_Geospatial_analysis1	Completed	2023-10-25 12:50:36	1 day		◘ ♯ 0		
		Rocker_RStudio_Verse_analysis1	Completed	2023-10-25 12:21:02	1 day		0 🛱 🖸		
		Cloud Shell analysis1	Completed	2023-10-23 21:54:50	about 12 hour	S			
				< 1 > 100 ▼					

## CyVerse: Data Transfer

- Move data between Data Store and your local computers
  - Discovery Environment web interface (limit: <2GB upload)
  - Cyberduck desktop app (similar to Globus for UA HPC)
  - GoCommands
- GoCommands
  - Command line tool which is operated in terminal
  - For Windows, use PowerShell or CommandPrompt
  - Once installed with configurations set up, very handy to transfer data

## CyVerse: Data Transfer

- GoCommands
  - Follow <u>https://learning.cyverse.org/ds/gocommands/</u> for installation and configuration setup
  - Upload data using put subcommand:

./gocmd put --progress /local\_directory /iplant/home/cyverse\_username/destination\_folder

• Download data using get subcommand:

./gocmd get --progress /iplant/home/cyverse\_username/target\_file /local\_destination

• Can easily transfer files or folders!

Last login: Wed Nov 29 22:17:15 on ttys000				Deres		( <b>0</b>	- 01
<pre>(base) yyeh@Yun-TingdeMBP ~ % ./gocmd putprogre ck/Output/plot/ois/overall_failrate.png /iplant/heta</pre>						Smo	gcr
/Users/yyeh/Dropbox/SmogCheck/Output/plot/ois/ove	-		•				
/Users/yyeh/Dropbox/SmogCheck/Output/plot/ois/ove-		0.0%	[				
/Users/yyeh/Dropbox/SmogCheck/Output/plot/ois/ove		0.0%	[				
/Users/yyeh/Dropbox/SmogCheck/Output/plot/ois/ove		0.0%	[				
'Users/yyeh/Dropbox/SmogCheck/Output/plot/ois/ove <sup>,</sup>		0.0%	[				
/Users/yyeh/Dropbox/SmogCheck/Output/plot/ois/ove <sup>,</sup>	·	0.0%	[				
'Users/yyeh/Dropbox/SmogCheck/Output/plot/ois/ove <sup>,</sup>	·	0.0%	[				
/Users/yyeh/Dropbox/SmogCheck/Output/plot/ois/ove <sup>,</sup>		done!	[36.	86KB	in	623	ms]
(base) yyeh@Yun-TingdeMBP ~ %			2001	CORD		020	

## More Tools...

#### **Data Visualization Tools**

- Voyager: <u>https://vega.github.io/voyager/</u>
- RAWGraphs: <u>https://www.rawgraphs.io/</u>

#### **Generative AI in Google Search**

- Only need Chrome browser & Google account
  - Click 'Lab icon' on a new page and turn on SGE (Search Generative Experience)
  - Only available for users in the US
- For some search, you'll have an option to generate an Al-powered overview
  - Additional option to enter a new conversation model and ask follow up questions

## Example: Generative AI in Search

- Particularly helpful in *locating* policy documents...
- Comes with links to references!

Google	why did california bar introduce ois technology for smog inspection? $ imes$ 4 $ extsf{eq}$ 3 $ imes$ 4				(i)	蛊		Y
	Perspectives News Images Qui Videos Shopping Books Maps Flights		ŀ	All filters 👻 Tools			SafeSe	arch 👻
	Generative AI is experimental. Info quality may vary. Learn more							
	The Bureau of Automotive Repair (BAR) introduced the BAR On-Board Diagnostic Inspection System (BAR-OIS) <b>to improve vehicle</b> <b>emissions testing</b> . The BAR-OIS is required for the following vehicles: • Most gasoline vehicles from 2000 and newer	ARTMENT OF CONSUMER AFFA.	Geotab and the California BAR: What	Smog Check program legislation aims for clean air benefits				
	<ul> <li>Hybrid vehicles</li> <li>Alternative-fuel vehicles</li> </ul>	🚯 Califor	Nov 17, 2020 Geotab	Feb 23, 2010				
	Most diesel vehicles from 1998 and newer	Californi ;	Coolab :	A Californi :				
	The BAR-OIS aims to: 🔛							
	Improve air quality							
	Reduce the cost of Smog Checks							
	Reduce the time it takes to get a vehicle inspected							
	The BAR also introduced a Continuous Testing Program (CTP) Pilot program. This program uses telematics to reduce the need for in-person Smog Checks. 👽							

↔ What is California Bar OIS? ↔ What is the new smog law in California in 20	23?
Ask a follow up	

## Resources

- Prompt Engineering (GPTs): <u>https://ua-data7.github.io/introllms/chatgpt\_prompts/</u>
- Find Al Tools: <u>https://github.com/ua-data7/LearningResources/wiki/Al-Tools-Landscape</u>
- Data Service at Library: <a href="https://data.library.arizona.edu/data-management">https://data.library.arizona.edu/data-management</a>
- FOSS materials: <u>https://foss.cyverse.org/</u>
- DataLab workshops by DSI: <u>https://www.datascience.arizona.edu/calendar</u>
- Workshop recordings: <u>https://www.youtube.com/c/CyverseOrgProject/videos</u>
- Events for researchers: <a href="https://researchbazaar.arizona.edu/">https://researchbazaar.arizona.edu/</a>
  - Coffee and Code, Hacky Hour, Community Events, ... etc.

# Thank you!

Contact me at <u>yyeh@arizona.edu</u> if you have any questions ©

**Reference:** 

This material is based upon work supported by the National Science Foundation under Award Numbers DBI-0735191, DBI-1265383, and DBI-1743442. URL: www.cyverse.org

0