

Intro to Informatics for the Brain Resilience Study

Part 1: FAIR workflows and using Cedar

Session Outline

- Interactive
 - you'll need to be on a computer to follow along
- Make sure you have your login info for Digital Research Alliance (Compute Canada)
 - Should be the same credentials you use to log into <https://ccdb.alliancecan.ca/security/login>
 - Also have your MFA device ready
- Feel free to interrupt or leave a message in the chat if you have questions, need something repeated, or are having trouble with any steps
- This session will be recorded – but only for internal use, within the BRS

FAIR and Open Science Principles

The **FAIR** data principles



Findable

To identify data for both humans and computers by computerising metadata that facilitate searching for specific datasets.



Accessible

Data is stored properly -for long term- so that it can easily be accessed and/or downloaded with well-defined access conditions. These could be access to the metadata (only) or getting access to the actual data.



Interoperable

The ability to combine different datasets either by humans or by computers. Therefore multiple agreements have to be made with respect to the terminology used to prevent ambiguities of the meanings of these terms.



Reusable

Data should be ready to be used for future research and to be further processed using computational methods. This requires adequate information about how the data were obtained and processed (provenance), and an appropriate license.

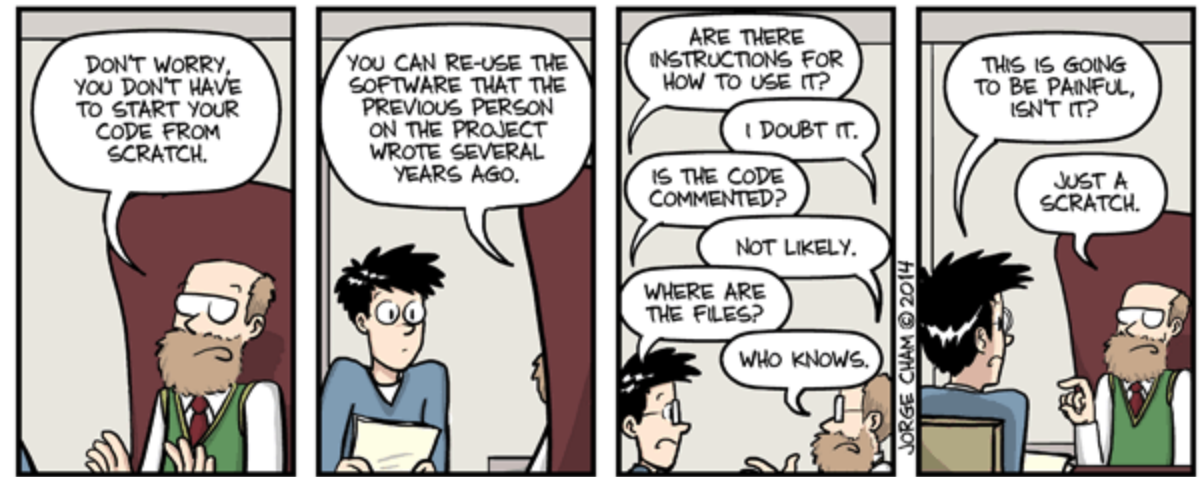
Wilkinson et al (2016) *Sci Data*
<https://doi.org/10.1038/sdata.2016.18>

Why do we care about FAIR?

- Reproducibility
 - **Makes work easier for you** and anyone who uses/shares your workflows/data
- Why FAIR? See:
 - [Fundamentals of Data Management for the Brain Resilience Study](#)
 - [Recording](#)
 - [Paper on FAIR](#)

Piled Higher and Deeper by Jorge Cham

www.phdcomics.com



title: "Scratch" - originally published 3/12/2014

How have we made the Brain Resilience Study FAIR, so far?

- Centralized storage
 - We've chosen a single location on Cedar to access the shared datasets from
- Easy and reproducible workflows
 - Machine readable data (BIDS, predictable file format)
 - TSV/CSV, tabular format
 - Standardized file name format and directory structure
 - Automated workflows
 - Processing & analysis is done with code > reproducible
 - Git/GitHub for tracking changes to scripts
- Metadata and Documentation
- Safeguarding data integrity
 - Three copies of data (Cedar, Cedar automated back-up, data acquisition laptops & paper copies)
 - Validation of transcribed data
 - Main shared dataset is write-protected to prevent alteration
 - Working collaboratively with generated data: Setting permissions
 - Data flows: main dataset -> scratch -> projects -> main dataset (if applicable), downloads and uploads

Cedar: Introduction

What is Cedar?

- Cedar is a *supercomputer that can be accessed via the internet* to accomplish high-performance computing tasks
- It is widely used to *store AND process large research data sets*.
- *Not just storage!*

Why Cedar?

- Goal: most work is done on Cedar instead of locally
 - Scale, reproducibility, longevity, dataset growth
 - Compute resources, standardized environments, debugging
 - Training and support will be provided
 - Data sharing

How do I get on Cedar?

- Through your computer's *Terminal app*



Local
(Your Computer)

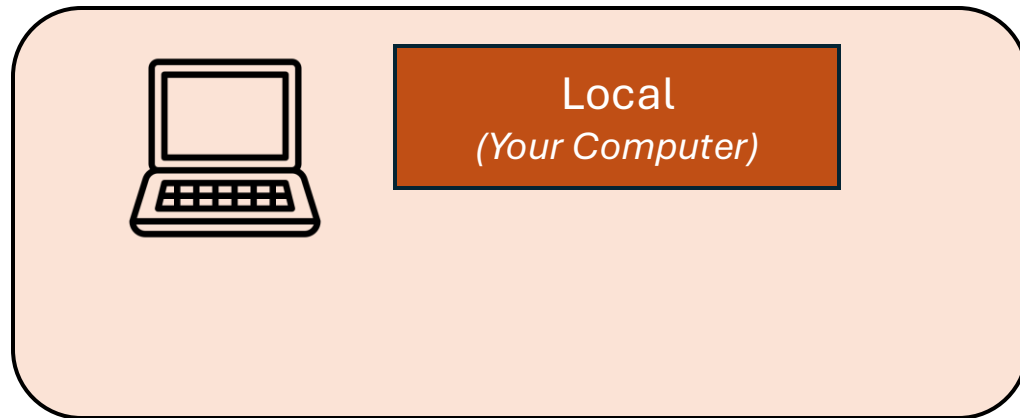
Terminal:

```
brs — -zsh
Last login: Tue Jan 28 12:59:50 on ttys039
brs@d207-023-164-249 ~ %
```

```
Windows PowerShell
Copyright (C) Microsoft Corporation.

Install the latest PowerShell for new features and improvements!

PS C:\Users\Justin>
```



Let's open our File Explorer/Finder with:

- **open .** (mac)
- **start .** (windows)

You can move around your local computer with

- **Powershell/Terminal** (command line)
- **File Explorer/Finder** (graphical user interface, or *GUI*)

You can look at files and directories (folders)

Powershell /Terminal:

```
PS C:\Users\Justin\my_folder> pwd
```

```
Path
```

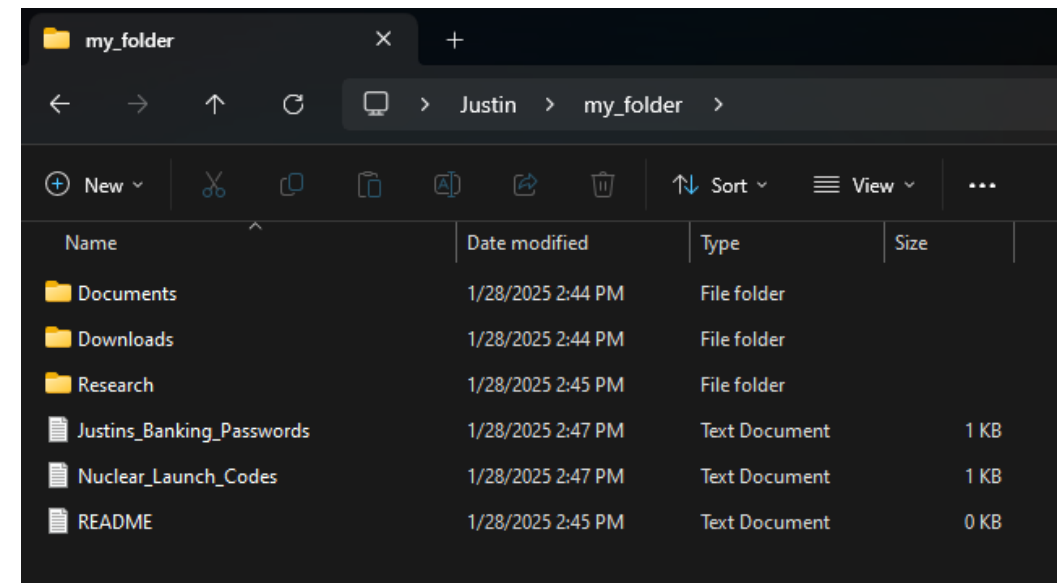
```
C:\Users\Justin\my_folder
```

```
PS C:\Users\Justin\my_folder> ls
```

```
Directory: C:\Users\Justin\my_folder
```

Mode	LastWriteTime	Length	Name
d----	1/28/2025 2:44 PM		Documents
d----	1/28/2025 2:44 PM		Downloads
d----	1/28/2025 2:45 PM		Research
-a----	1/28/2025 2:47 PM	36	Justins_Banking_Passwords.txt
-a----	1/28/2025 2:47 PM	445	Nuclear_Launch_Codes.txt
-a----	1/28/2025 2:45 PM	0	README.txt

File Explorer/Finder:

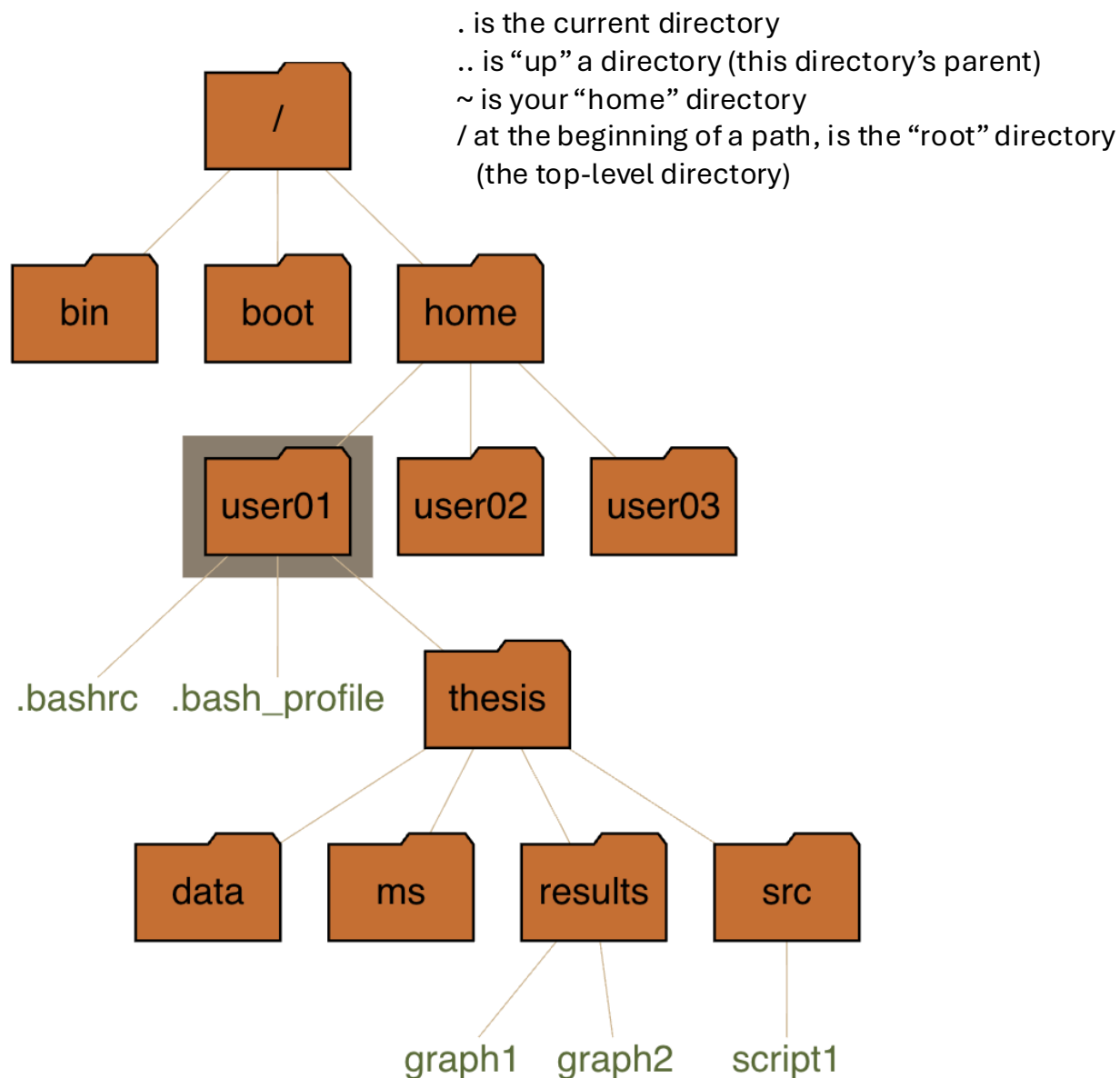




Local
(Your Computer)

Fundamental Navigation Commands

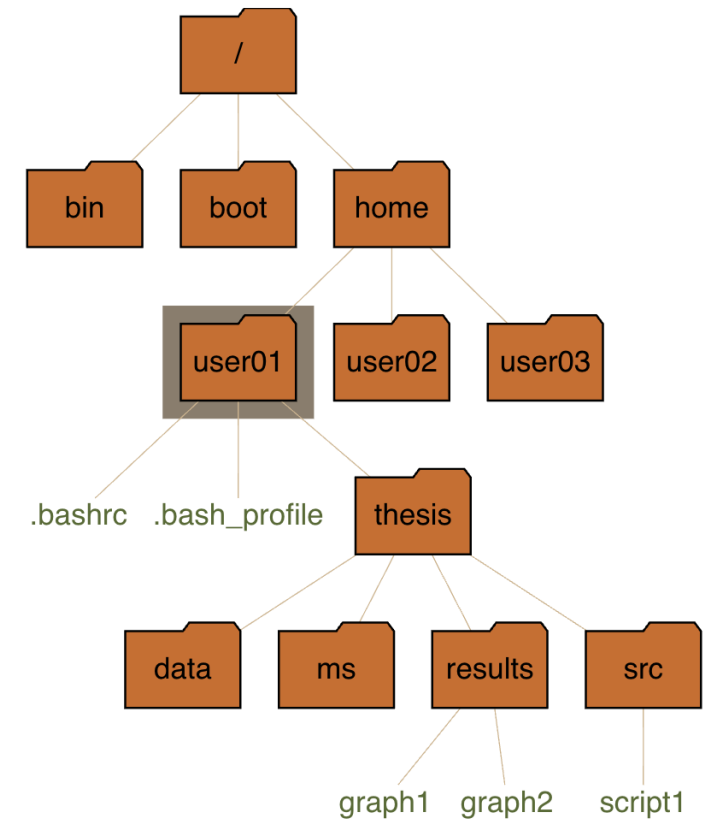
- ***pwd***: Print Working Directory
 - Shows you where you are
- ***cd***: Change Directory
 - Changes where you are
 - Absolute vs relative file paths
 - *cd /home/user02*
 - *cd ..*
 - *cd ../../boot*
 - *cd thesis/results*
- ***ls***: List Files / Directories in Working Directory
 - Shows what files and folders are inside the folder you're in
 - Can also show the contents of another folder
 - *ls /home/user03*
 - *ls thesis*
- ***cat***: Shows contents of a file



Directories

Using pwd, cd, & ls to Navigate Directories

- *Sample Problems:*
 - From the user01 directory, what are 2 ways to navigate to the results directory?
 - From the results directory, what are 2 ways to print the content of the src directory?





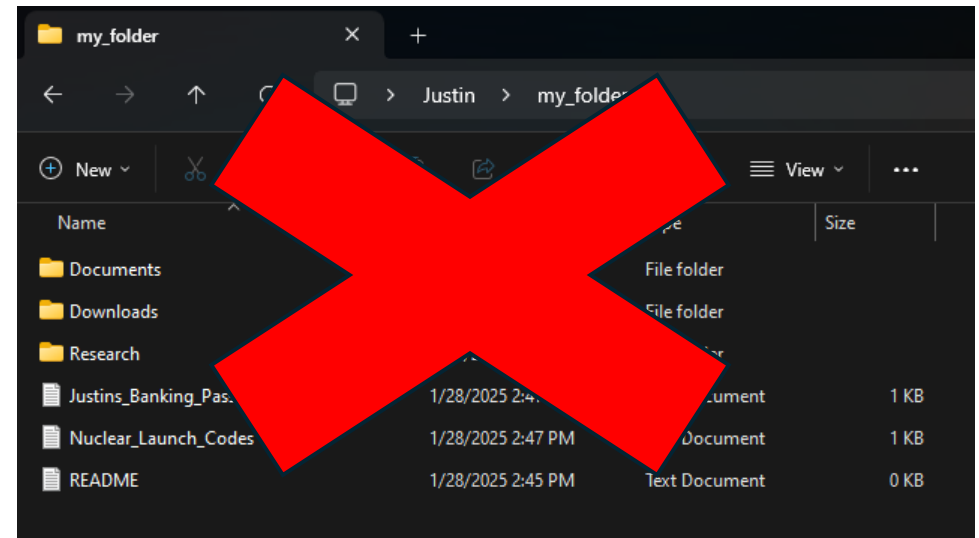
Cedar
(Digital Research Alliance Server)
(Cedar's Supercomputers)

- To access the BRS data, we need to be on the Cedar servers, not our local computers
- The commands we used to move around on your local computer are the same on cedar:
 - *ls, cd, pwd*
- There is NO GRAPHICAL USER INTERFACE (GUI) on Cedar
 - e.g. no Finder app or File Explorer on Cedar
- We need to access Cedar through **Terminal**

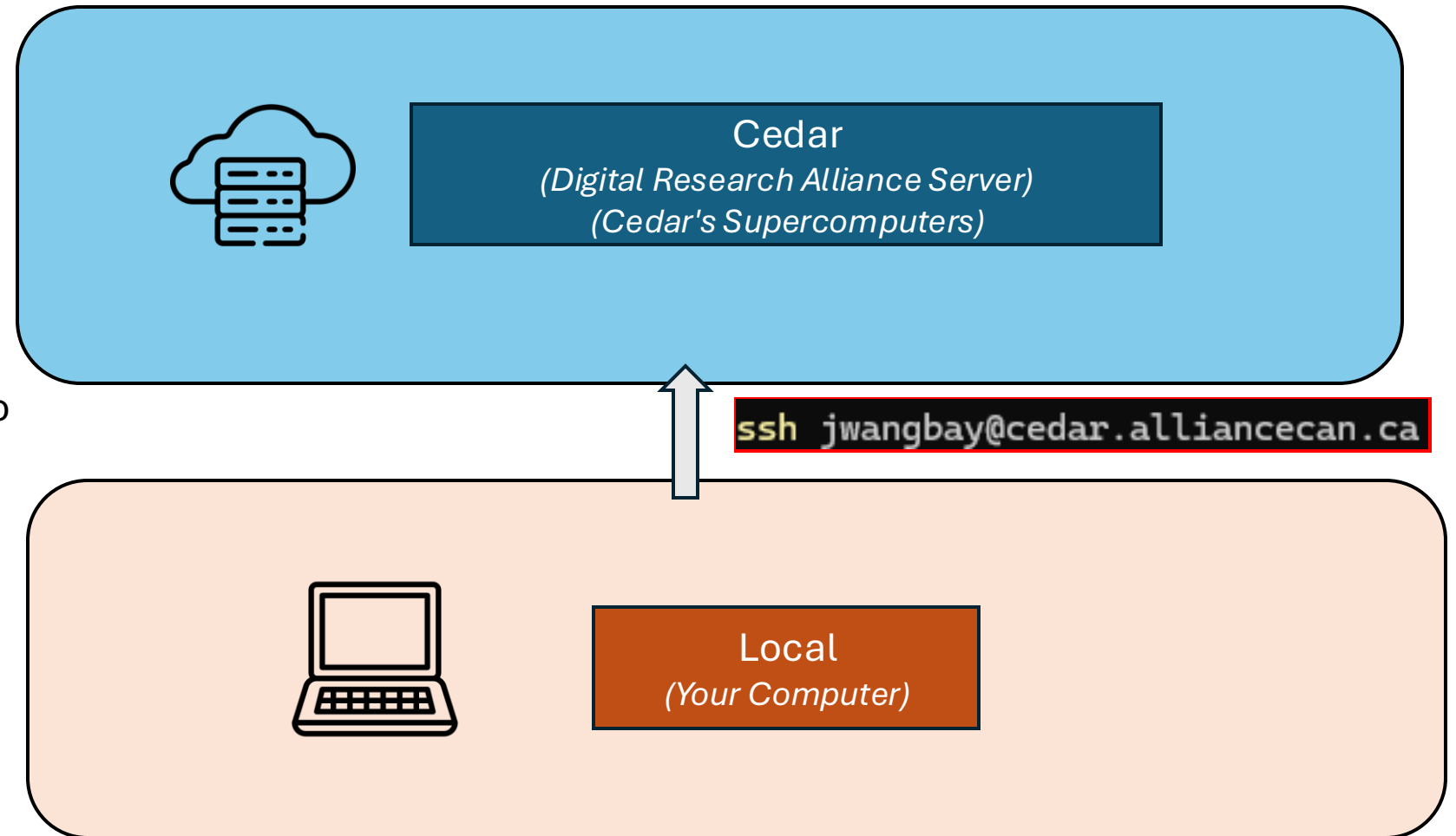
Powershell/Terminal:

```
jwangbay@cedar1:~/scratch  x  +  v  -  □  x
[jwangbay@cedar1 ~]$ pwd
/home/jwangbay
[jwangbay@cedar1 ~]$ cd scratch/
[jwangbay@cedar1 scratch]$ ls
AD TVB_Cam-CAN_PHD_Scripts.tar.gz
[jwangbay@cedar1 scratch]$ |
```

File Explorer/Finder:



Accessing Cedar



The `ssh` command allows you to enter Cedar's servers

What ssh looks like:

Logging in to Cedar

1. Open Terminal
2. Type *ssh <username>@cedar.alliancecan.ca* into the command line
3. Fill out your password and complete the MFA

```
PS C:\Users\Justin> ssh jwangbay@cedar.alliancecan.ca
The authenticity of host 'cedar.alliancecan.ca (206.12.124.2)' can't be established.
ED25519 key fingerprint is SHA256:a4n68wLDqJhjtePn04T698+7anVavd0gdpiECLByLAU.
This host key is known by the following other names/addresses:
  C:\Users\Justin/.ssh/known_hosts:1: cedar.computecanada.ca
Are you sure you want to continue connecting (yes/no/[fingerprint])? yes
Warning: Permanently added 'cedar.alliancecan.ca' (ED25519) to the list of known hosts.
(jwangbay@cedar.alliancecan.ca) Password:
(jwangbay@cedar.alliancecan.ca) Duo two-factor login for jwangbay

Enter a passcode or select one of the following options:

1. Duo Push to phone (iOS)
2. Duo Push to pad (iOS)

Passcode or option (1-2): 1
Success. Logging you in...
Success. Logging you in...
Last login: Mon Feb 10 22:41:48 2025 from 172.103.139.165

=====
Welcome to Cedar! / Bienvenue sur Cedar!

For information see: https://docs.alliancecan.ca/wiki/Cedar
Email support@tech.alliancecan.ca for assistance and/or to report problems.

Pour plus d'information lisez : https://docs.alliancecan.ca/wiki/Cedar
Écrivez à support@tech.alliancecan.ca pour obtenir de l'aide ou rapporter un
problème.
=====
```

How to tell if you're in Cedar



Cedar

*(Digital Research Alliance Server)
(Cedar's Supercomputers)*

 jwangbay@cedar1:~

[jwangbay@cedar1 ~]\$ |

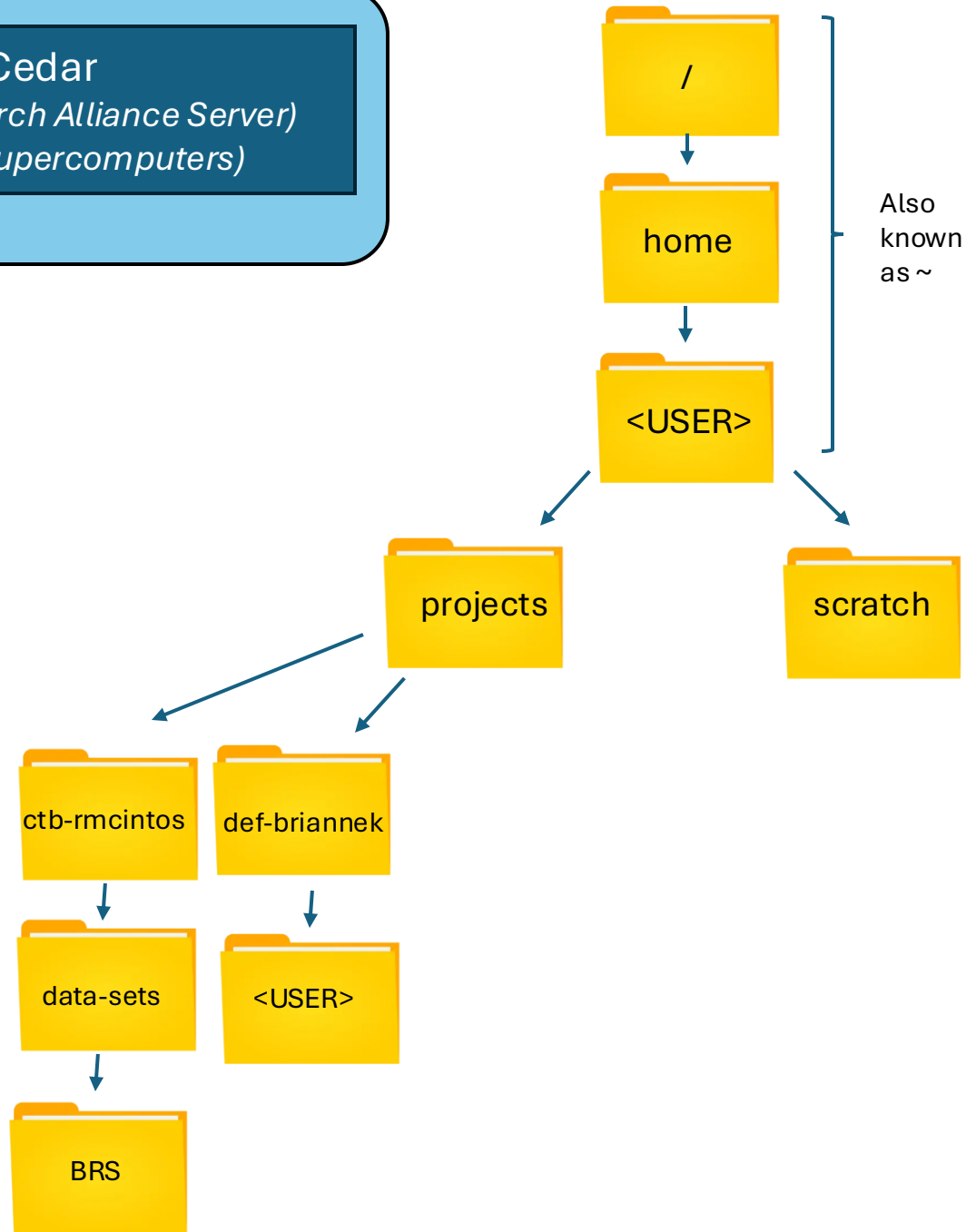
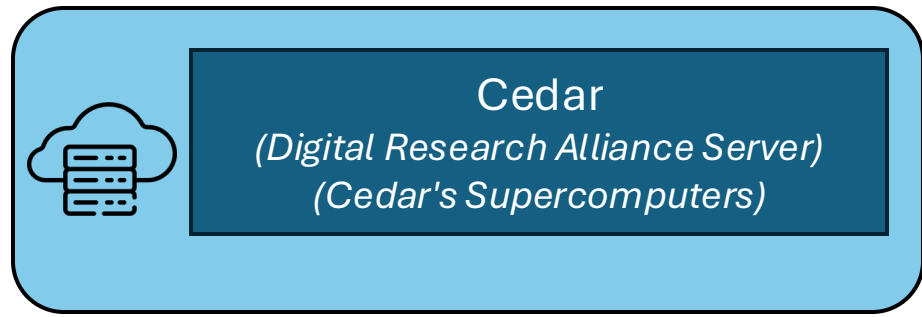
Cedar: Basic Navigation Using Bash


Fundamental Navigation Commands

1. *pwd*: Print Working Directory
 - a) *pwd*
2. *cd*: Change Directory
 - a) *cd <path>*
3. *ls*: List Files / Directories in Working Directory
 - a) *ls*
 - b) *ls <path>*
4. *cat*: Shows contents of a file
 - d) *cat <path_to_file>*
5. *mkdir*: creates a directory
6. *nano*: Creates/edits a file
 - a) *nano <path_to_file>*
 - b) *ctrl+x* closes the file, press *y* to save changes, enter to confirm and exit nano
7. *cp*: copies a file
 - a) *cp <source> <destination>*
8. *rm*: Deletes a file
 - a) *rm -i <path_to_file>*
 - b) *rm -ir <path_to_dir>*
 - c) be CAREFUL! There is no recycling bin
9. *ctrl+c*: cancels whatever you're doing

The same commands you use to navigate your local computer can be used to navigate Cedar when you're connected.

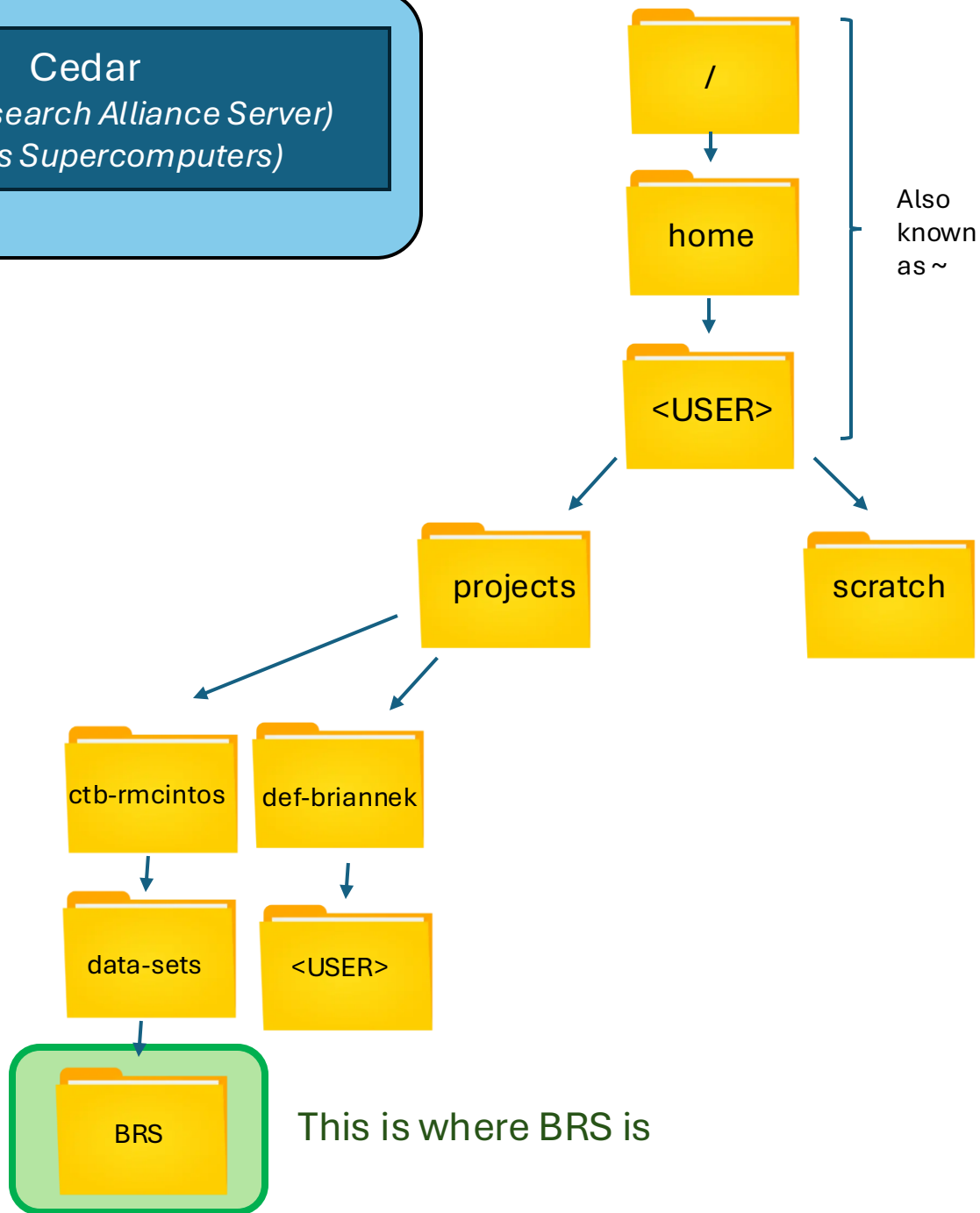
nano is especially important on Cedar because there is no GUI text editor

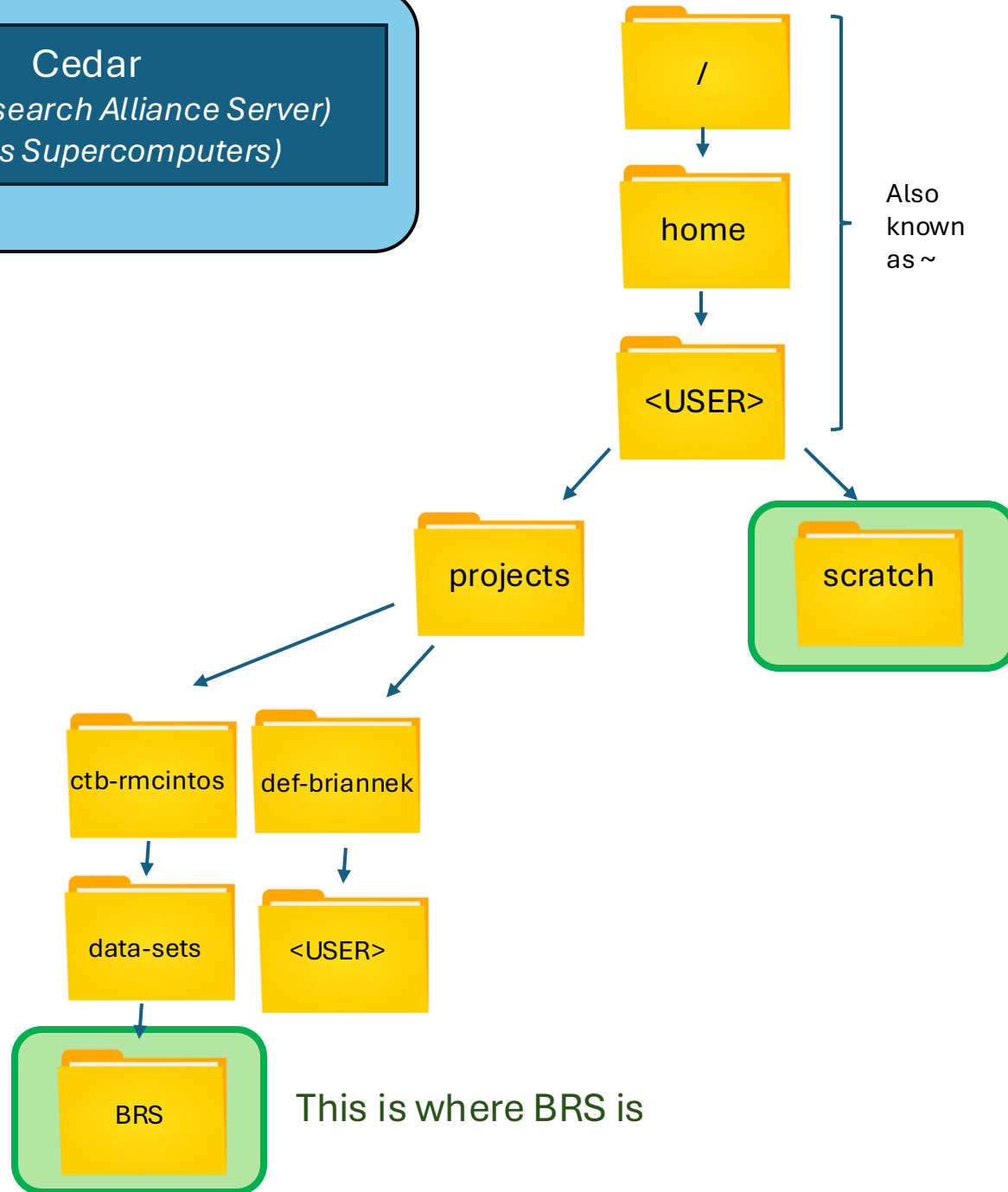
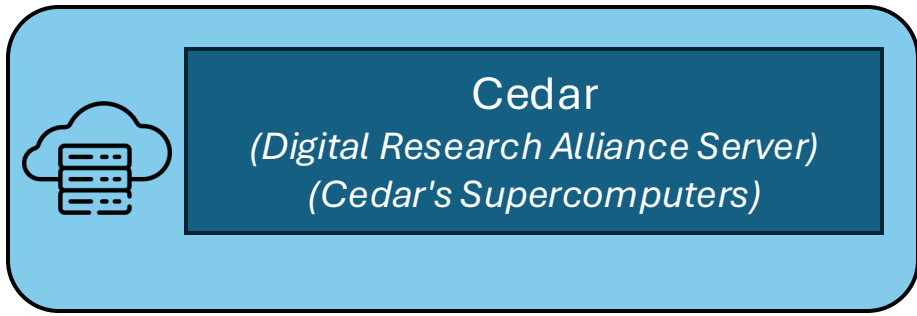




Cedar

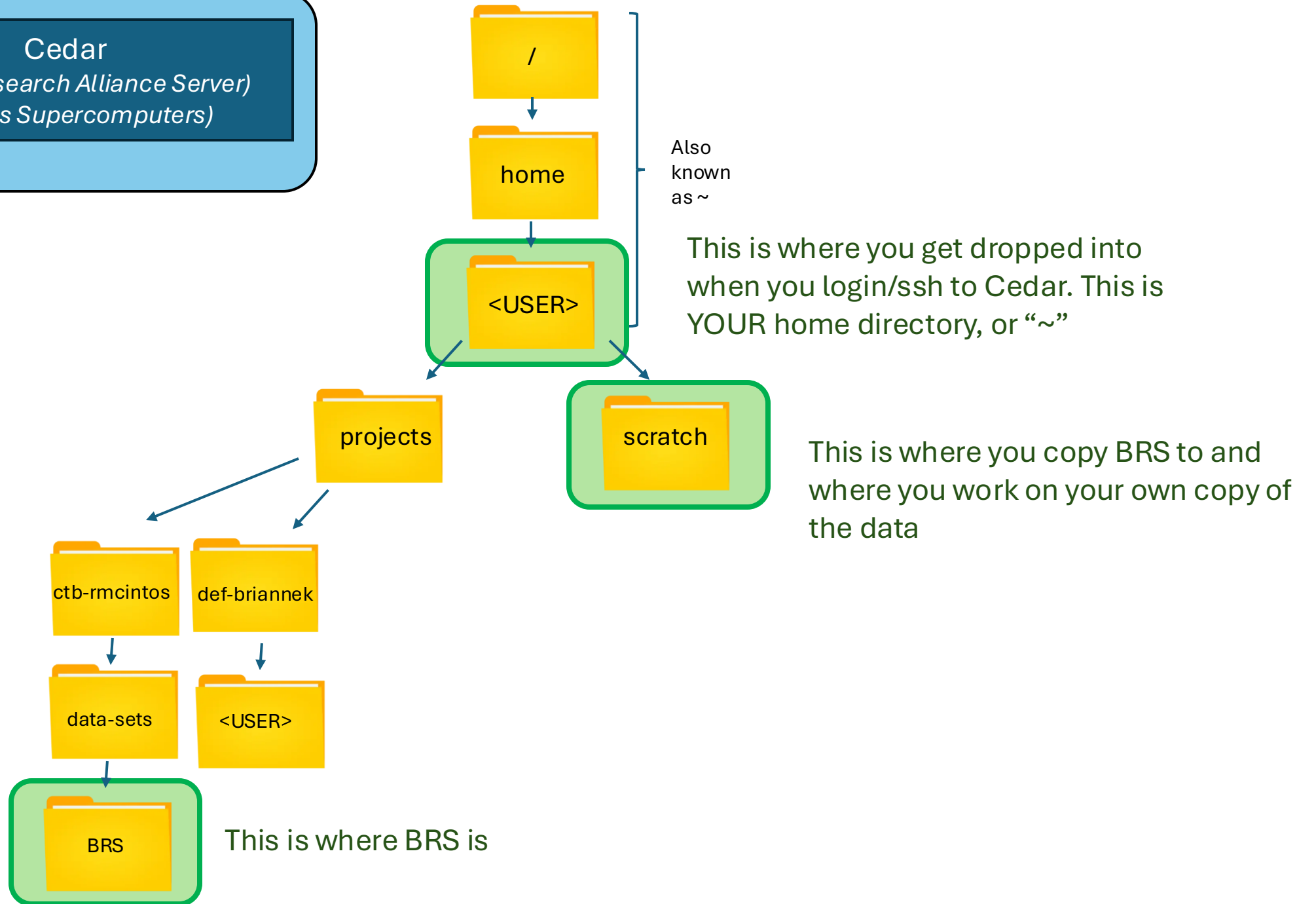
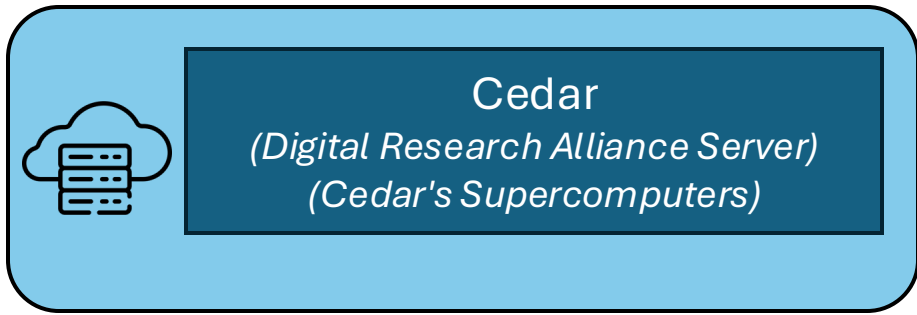
(Digital Research Alliance Server)
(Cedar's Supercomputers)






This is where you copy BRS to and where you work on your own copy of the data

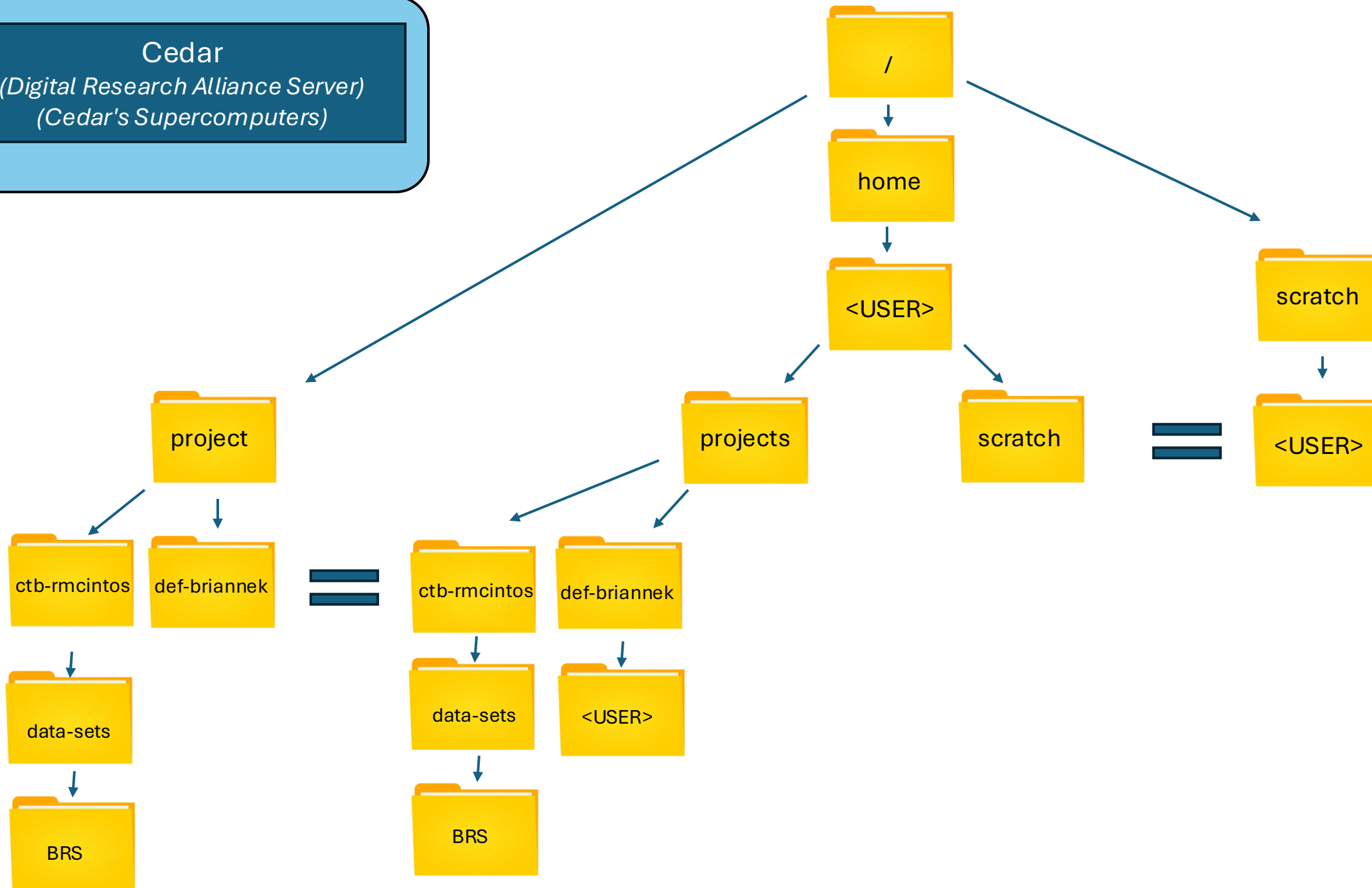
This is where BRS is






Cedar

(Digital Research Alliance Server)
(Cedar's Supercomputers)

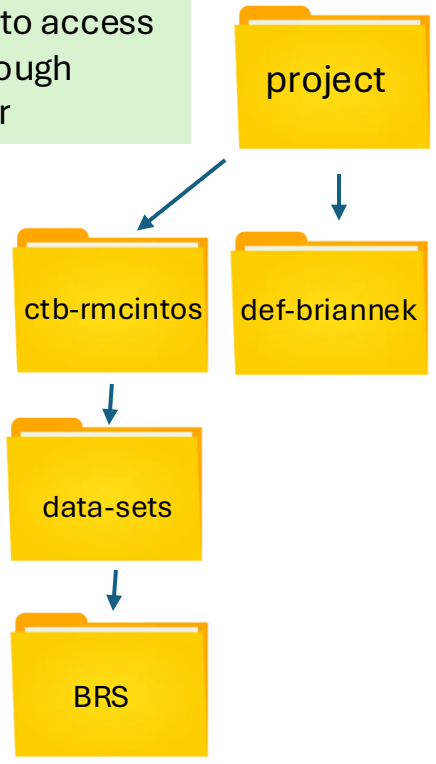




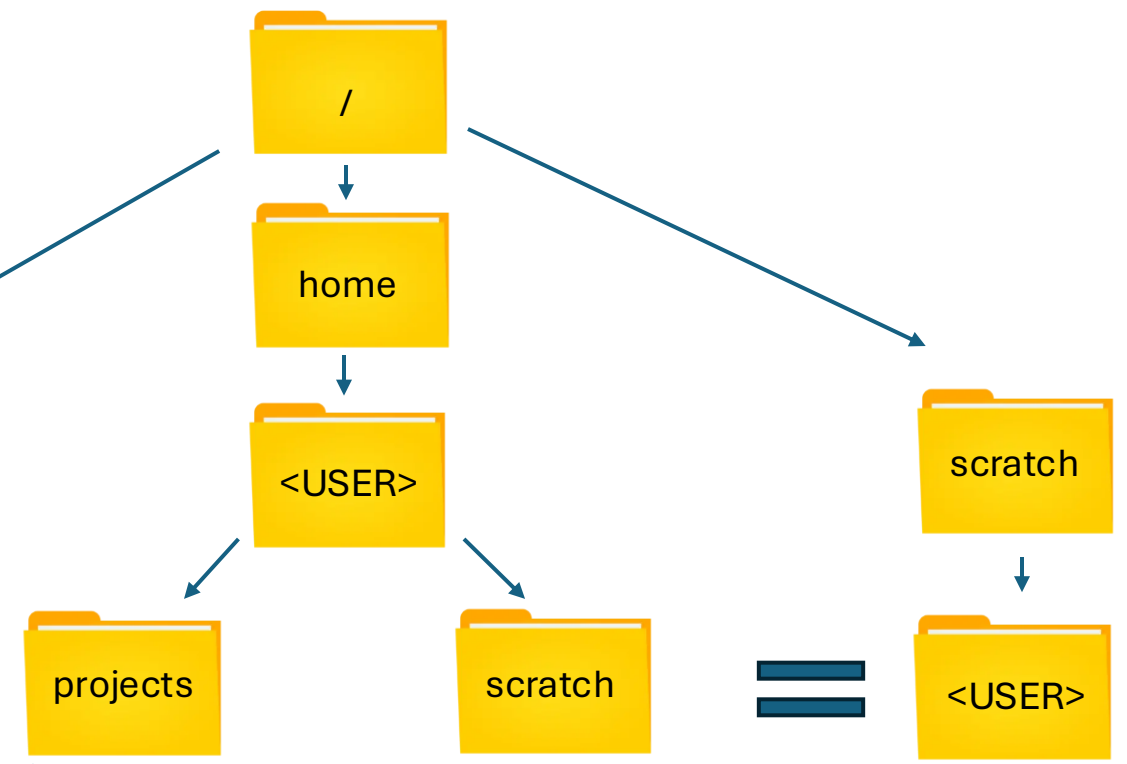
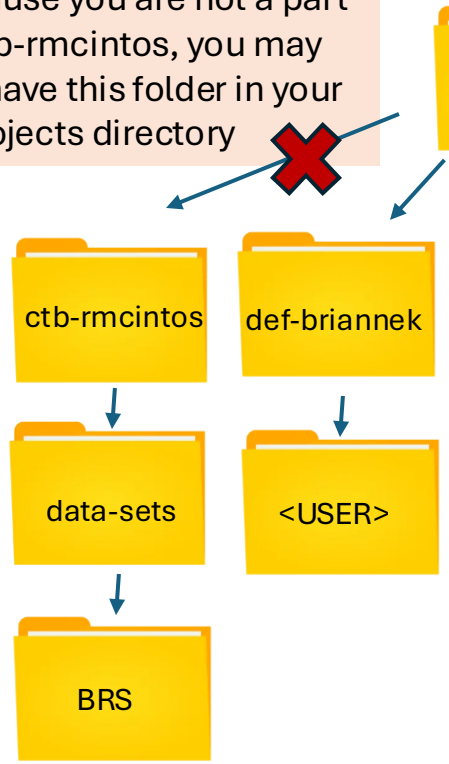
Cedar

(Digital Research Alliance Server)
(Cedar's Supercomputers)

You may be able to access ctb-rmcintos through /project, however



Because you are not a part of ctb-rmcintos, you may not have this folder in your ~/projects directory



Cedar: Demo

Tour of projects, home, scratch

Tour of ctb / BRS folder

Data Copying

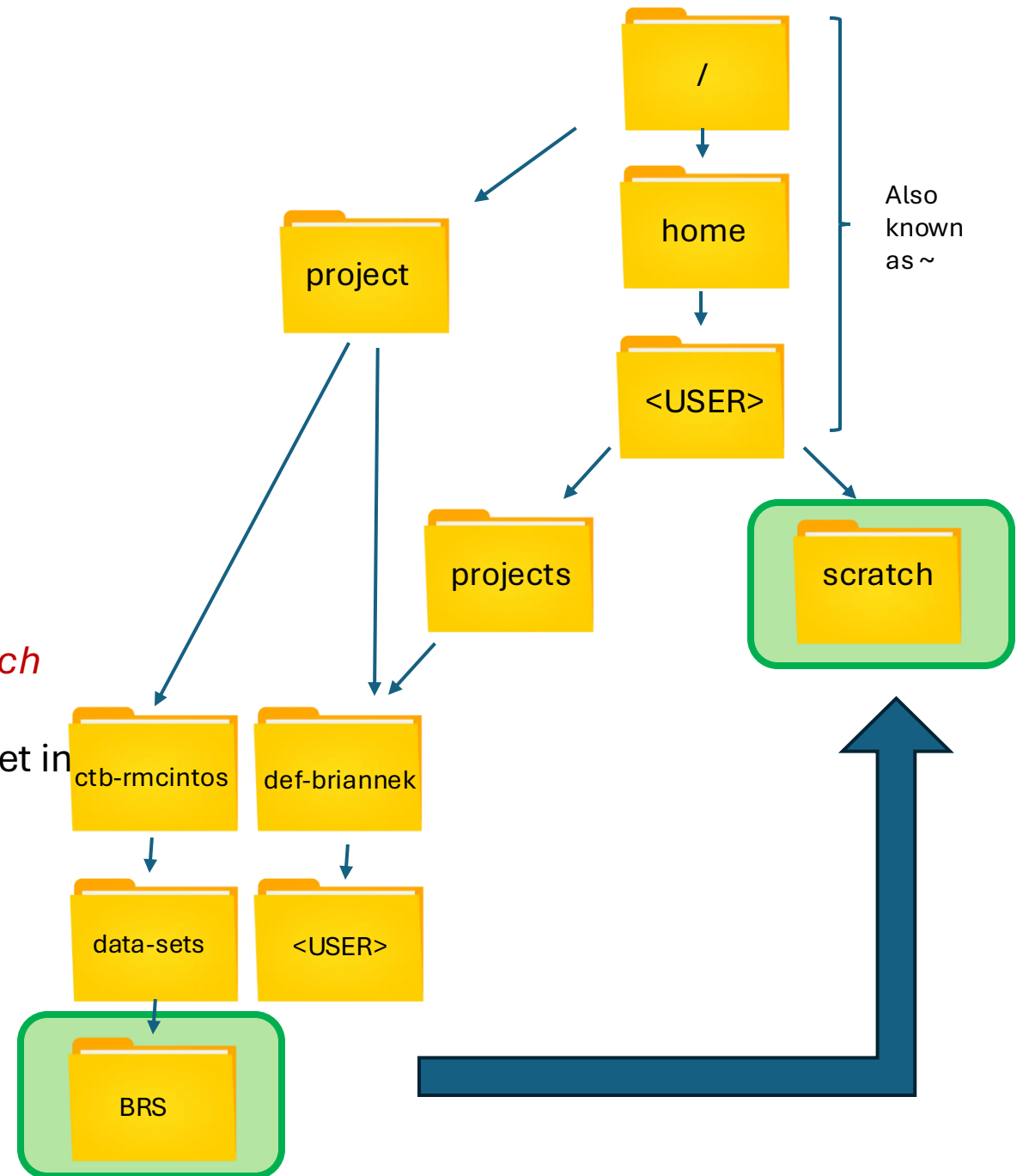
- *(scp & rsync)*
- *(from shared directory on projects to scratch & vice versa)*
- *(from local to Cedar & vice versa)*

Using BRS data?

Make a copy of the data in your scratch

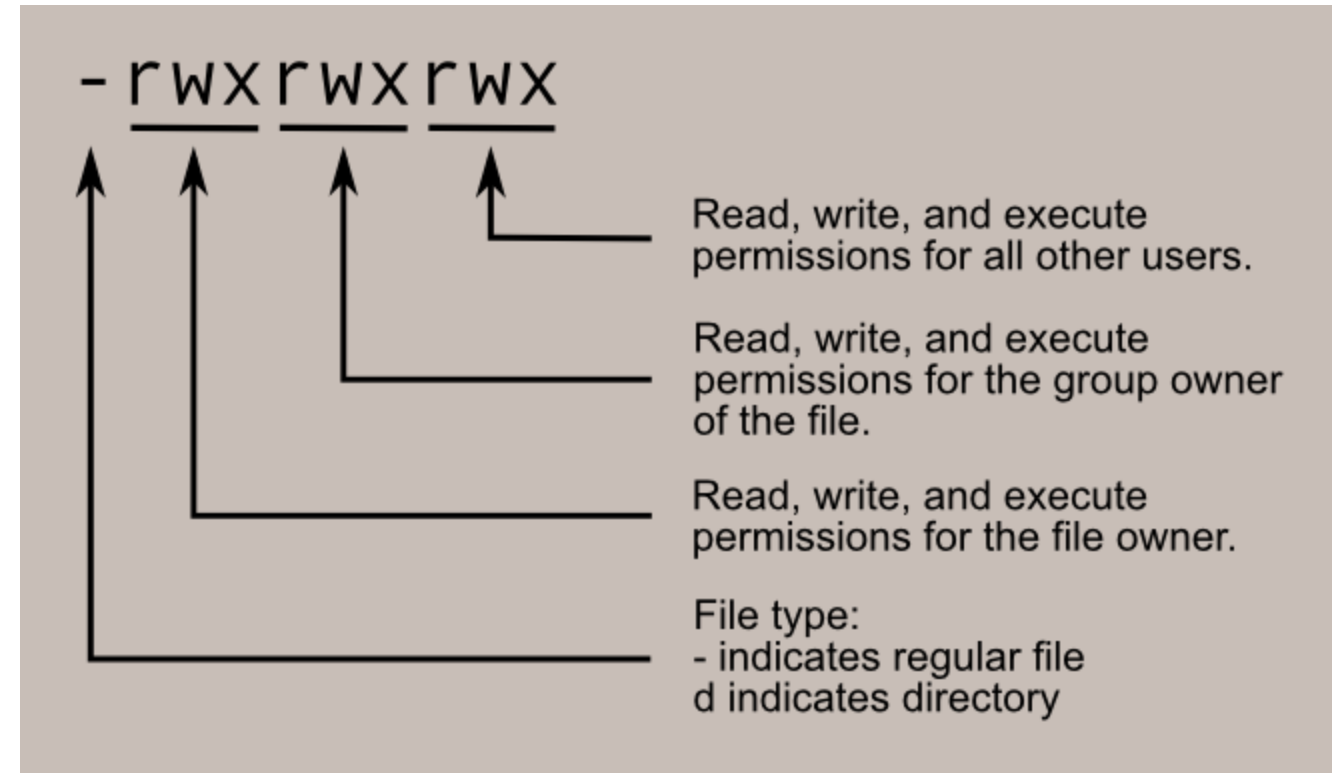
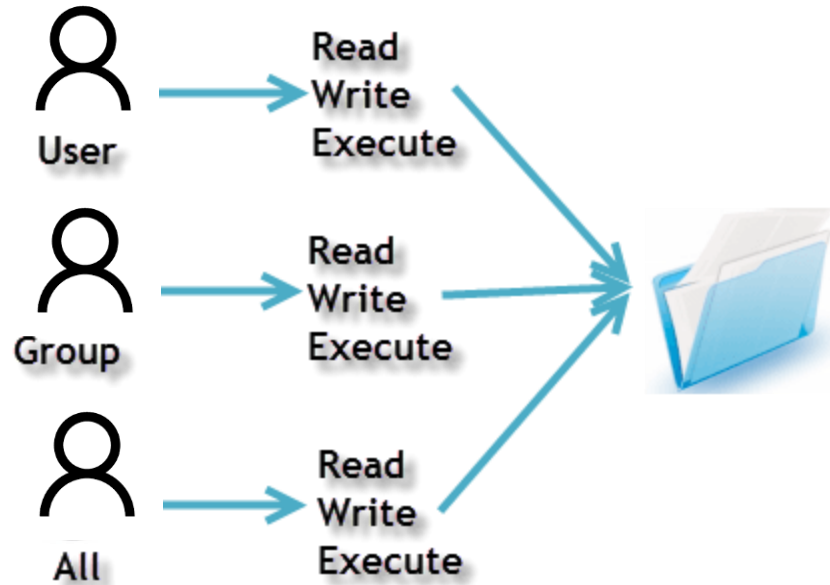
- ***TODO: correct these commands***
- *rsync -avzh --no-g --no-p /project/ctb-rmcintos/data-sets/BRS ~/scratch*
- *OR*
- *cp -R ~/projects/ctb-rmcintos/data-sets/BRS ~/scratch*

These commands will create a new copy of the BRS data-set in your scratch directory



Permissions

Owners assigned Permission On Every File and Directory



Files and folders belong to an owner and a group

`ls -l` - shows file/directory details, including permissions

`groups` - shows which groups you are a part of

`chmod` - allows the owner of the file to change its permissions

Sharing your BRS data?

Make a copy of your data on scratch, to your projects folder

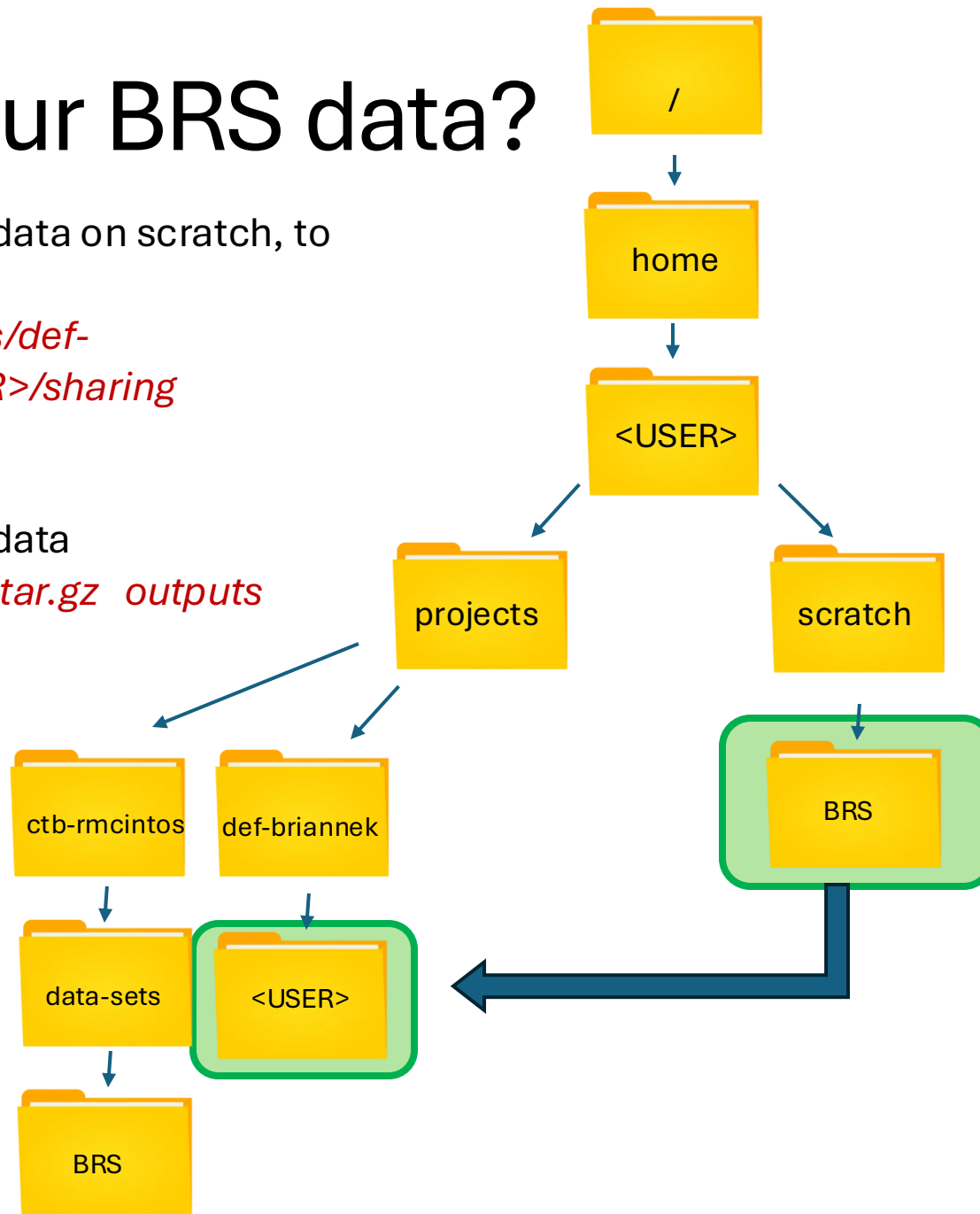
- `mkdir ~/projects/def-briannek/<USER>/sharing`

Tar (compress) your data

- `tar -czf outputs.tar.gz outputs`

Why?

- File limits!
- `diskusage_report`



- `rsync -avzh --no-g --no-p ~/scratch/BRS/outputs.tar.gz ~/projects/def-briannek/<USER>/sharing`
- OR
- `cp -R ~/scratch/BRS/outputs.tar.gz ~/projects/def-briannek/<USER>/sharing`
- `chgrp -R <group_name> <directory>`
- `chmod g+rx <directory>`
- `chmod g-w <directory>`

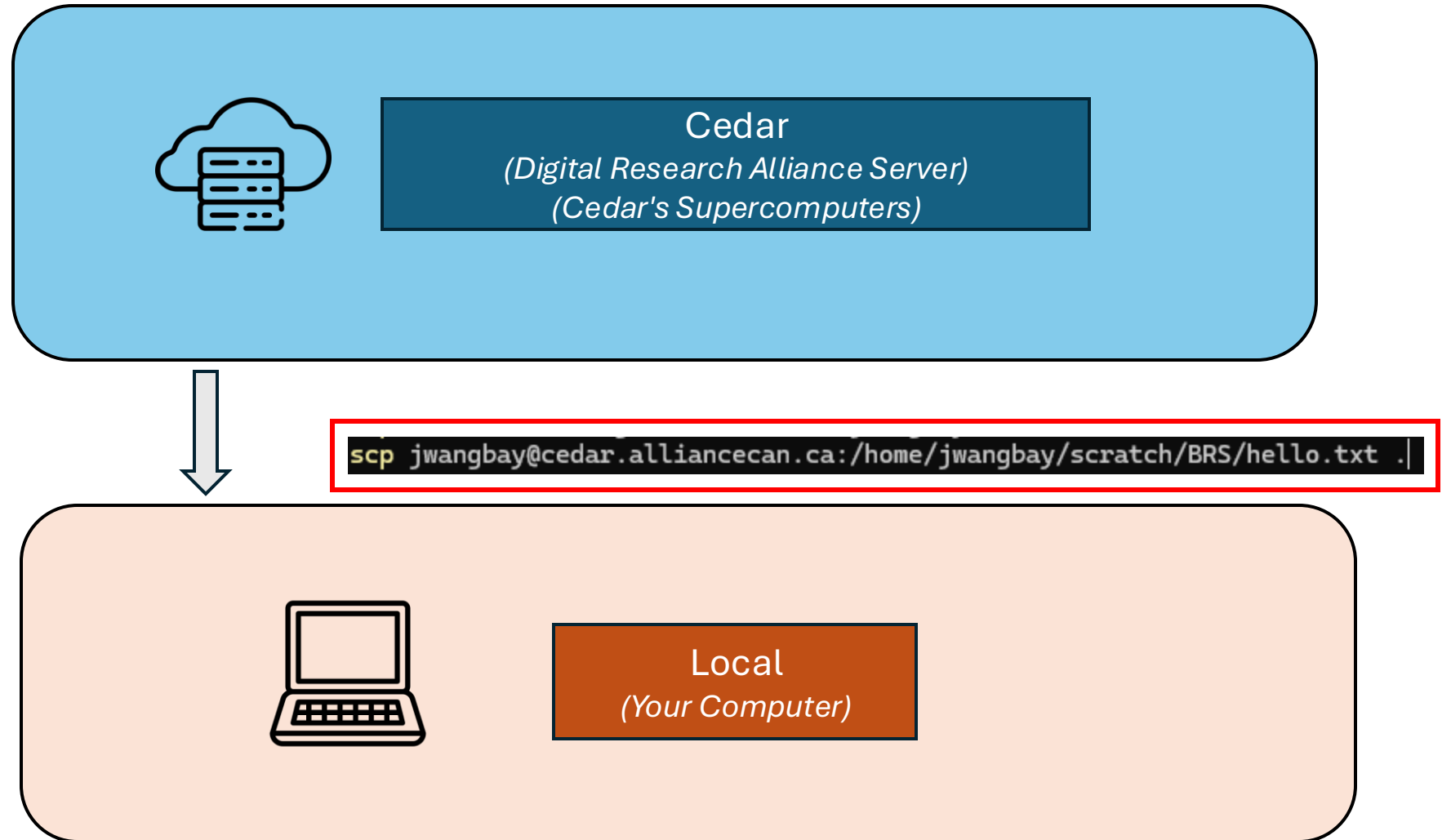
This will ensure members of def-briannek can see your shared files

Downloading data from Cedar to your computer

- *rsync -avzh --no-g --no-p source destination*
- *OR*
- *scp -R <source> <destination>*

Run this on terminal, from your local computer – not logged into Cedar

- If you're on Cedar, the *scp* command can't "see" your computer system



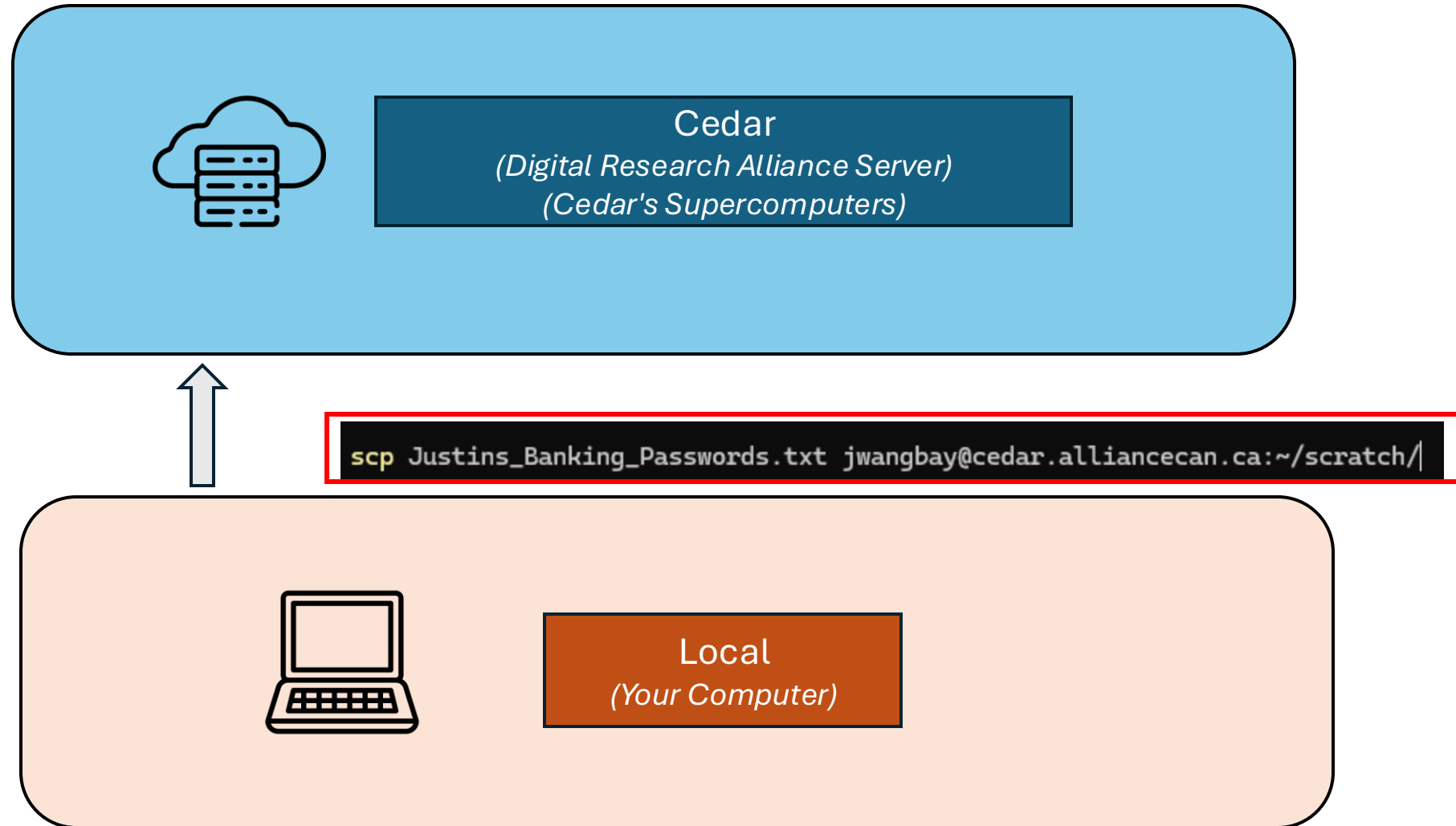
Uploading data to Cedar from your computer

- *rsync -avzh --no-g --no-p*
source destination
- OR
- *scp -R <source>*
<destination>

Run this on terminal, from your local computer – not logged into Cedar

- If you're on Cedar, the *scp* command can't "see" your computer's files

Same command but with source and destination flipped!



DO NOT DO LIST

- De-centralized storage
 - Multiple active/working copies of data, in multiple places, of multiple versions
- Difficult to reproduce workflows
 - Arbitrary filenames with no consistent format
 - Using Excel in Cedar
 - What does excel look like when you're writing code? There is no Microsoft Office on Cedar!
 - Solution: export spreadsheets as csv files and upload them to Cedar
 - Manual, untracked workflows
 - No tracking of code and how it's changed over time
 - No code, all changes are done manually
- Lack of data integrity safeguarding
 - No permission protection (important data may get deleted or overwritten)
 - Data is not in the correct place on Cedar for sharing, or inappropriate permissions may prevent visibility to other RAs
 - Directly editing the main, shared dataset
 - Solution: contact us if there are changes that need to be made to the main copy of BRS

Questions to ask yourself when trying to be FAIR

- Will this make things easier or harder the next time I or someone else tries to do the same thing?
 - Are my workflows recorded
 - Is the data easy to access
 - Am I doing things the roundabout way
 - Am I permanently, irreversibly altering the data
 - Am I doing things manually that could be automated
 - Could I make a mistake and not know it/be able to backtrack to it

What's next?

Make your own work FAIR:

- How to have reproducible code (Attend our GitHub workshop)
- How to automate certain processes (Attend our Python workshop)
- Look at how we've made BRS FAIR so far (Slide 5 in this PPT)
- Reach out if you have any questions or need support

Resources

- Brain Resilience Study data management plan
- [Alliance Can wiki](#)
- [Cedar status page](#)
- [Common errors, FAQ](#) *(to be updated)*
- Cedar will be [transitioned to Fir](#) sometime late-Spring