camh
Centre for Addiction and Mental Health

# Krembil Centre for Neuroinformatics

Using big data, artificial intelligence and brain modelling to fundamentally change our understanding of mental illness.

## SUMMER SCHOOL 2020

Day 1

Understanding clinical research questions and reproducible science

Afternoon: Best practices and tools for reproducible science

# Summer School Schedule

camh | Krembil Centre for Neuroinformatics

**Mon 05/07**
*Understanding clinical research questions and reproducible science.*
**S. Hill & E. Dickie**
**B. Jones & V. Tang**

**Tu 06/07**
**D. Buchman, L. Sikstrom & M Maslej**
*Applied ethics in machine learning and mental health*

**Wed 07/07**
*Genetics and transcriptomics*
*.*
**S.Tripathy,& D. Felsky**

**Wed 08/07**
**E. Hay & F. Mazza**
*Brain Microcircuit Simulations of Depression*

**Fri 09/07**
*Whole-Brain Modelling and Neuroimaging Connectomics*
**J. Griffiths & E Dickie**

**Mon 12/07**
**A Diaconescu + Lab**
*Bayesian Models of Learning and Integration of Neuroimaging Data*

**Tues 13/07**
*Digital Health and Population-based data resources*
**A. Pratap, J Yu & D. Felsky**

**Wed 14/07**
**A. Pratap & D. Felsky**
**Group Panel**
Integrative research methods
Panel Discussion

# Today’s Agenda

camh
Krembil Centre for Neuroinformatics

**Day 1:**
Welcome!
Understanding clinical research questions and reproducible science

| | |
|---|---|
| 9:00 am - 10:30 am | *Lecture 1: Welcome and Orientation + Neuroinformatics Across Scales*<br>Erin Dickie + Sean Hill |
| 10:45 am - 12:15 pm | *Problems and opportunities in the diagnosis and treatment of major depression*<br>Dr Victor Tang & Dr Brett Jones |
| 1:00 pm - 2:30 pm | *Workshop 1: Guiding principles for FAIR and open science*<br>Erin Dickie & Sejal Patel |
| 2:45 pm - 4:15 pm | *Workshop 2: Tools for Reproducible Science*<br>Erin Dickie & Sejal Patel] |

# This afternoon

camh | Krembil Centre for Neuroinformatics

*Guiding principles and tools for reproducible science*

**Erin Dickie Ph.D.**
Lead - Education and Knowledge Transfer
Krembil Centre for Neuroinformatics, Centre for Addiction and Mental Health, Toronto, Ontario
Twitter: @ErinWDickie Github: @edickie

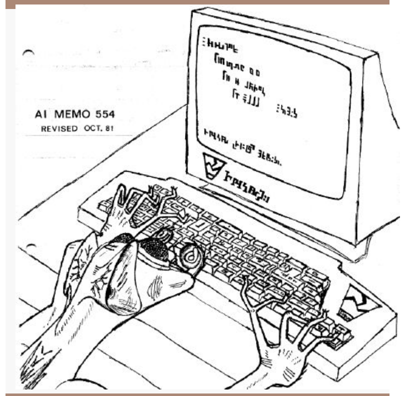
**Sejal Patel Ph.D.**
Post-doctoral Fellow - Whole Person Modeling
Krembil Centre for Neuroinformatics, Centre for Addiction and Mental Health, Toronto, Ontario
Github: @Sejal24

## Teaching Assistants for this section

camh | Krembil Centre for Neuroinformatics

Kevin Kadak
Whole Brain Modeling Lab
KCNI - CAMH

Taha Morshedzadeh
Whole Brain Modeling Lab
KCNI - CAMH



Kevin Witczak
CAMH Kimel TIGRlab
Github: @kimjetwav

# Remember - many ways to engage

camh | Krembil Centre for Neuroinformatics

(during sessions) Use the chat or the ask question!

You can always return to the session and re-watch the vidos after the session ends

come chat with us in KCNI Summer School Slack :)

virtually meet with us in gather.town

Tell us how the session went (post session survey): https://forms.gle/ji18qLMZEZ9L16Ln6

KCNISchool@camh.ca

# Outline for this afternoon

Why reproducible science?

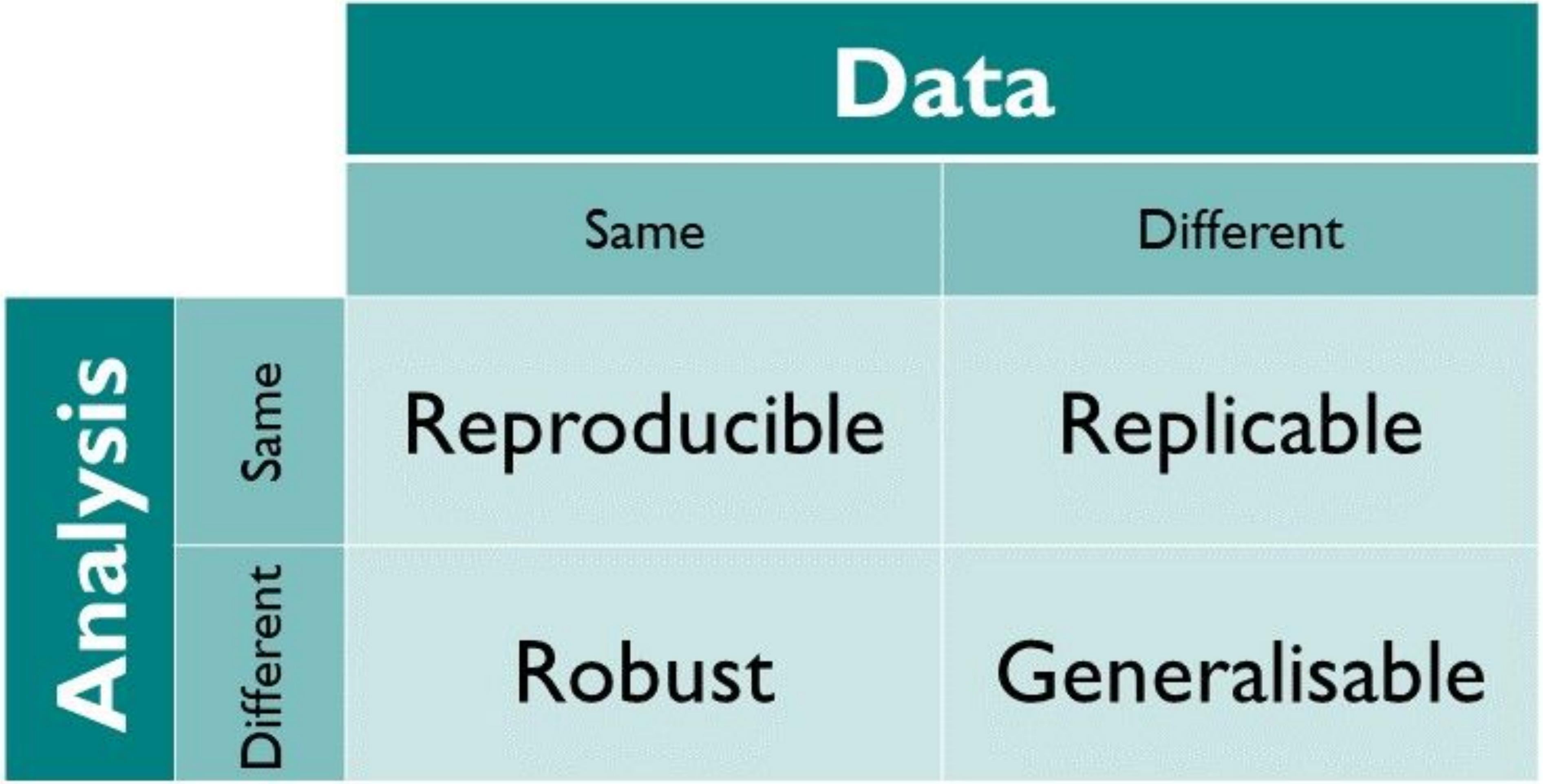
The things you need to know to get through this course:

- Versioning and publishing code (github)
- Versioning and publishing software (docker)
- R with Rmarkdown (walk through)
- Python in ipython notebooks (walk through)
- -also google colab

The fancy bits:

- building your own binder environments
- building your own containers (docker & singularity)

## What is reproducible science?



| | | Data | |
|---|---|---|---|
| | | Same | Different |
| Analysis | Same | Reproducible | Replicable |
| | Different | Robust | Generalisable |

**Fig. 5** How the Turing Way defines reproducible research

## Reproducible Neuroinformatics - What

camh
Krembil Centre for Neuroinformatics

An article about computational science in a scientific publication is not the scholarship itself, it is merely advertising of the scholarship. The actual scholarship is the complete software development environment and the complete set of instructions which generated the figures.

Buckheit and Donoho
WaveLab and Reproducible Research, 1995

## Reproducible Science - Why

For Yourself

For your colleagues

For science

## Reproducible Neuroinformatics - Why

camh Krembil Centre for Neuroinformatics

Your number one collaborator is yourself six months ago: And they don't answer emails

## For Science Friends

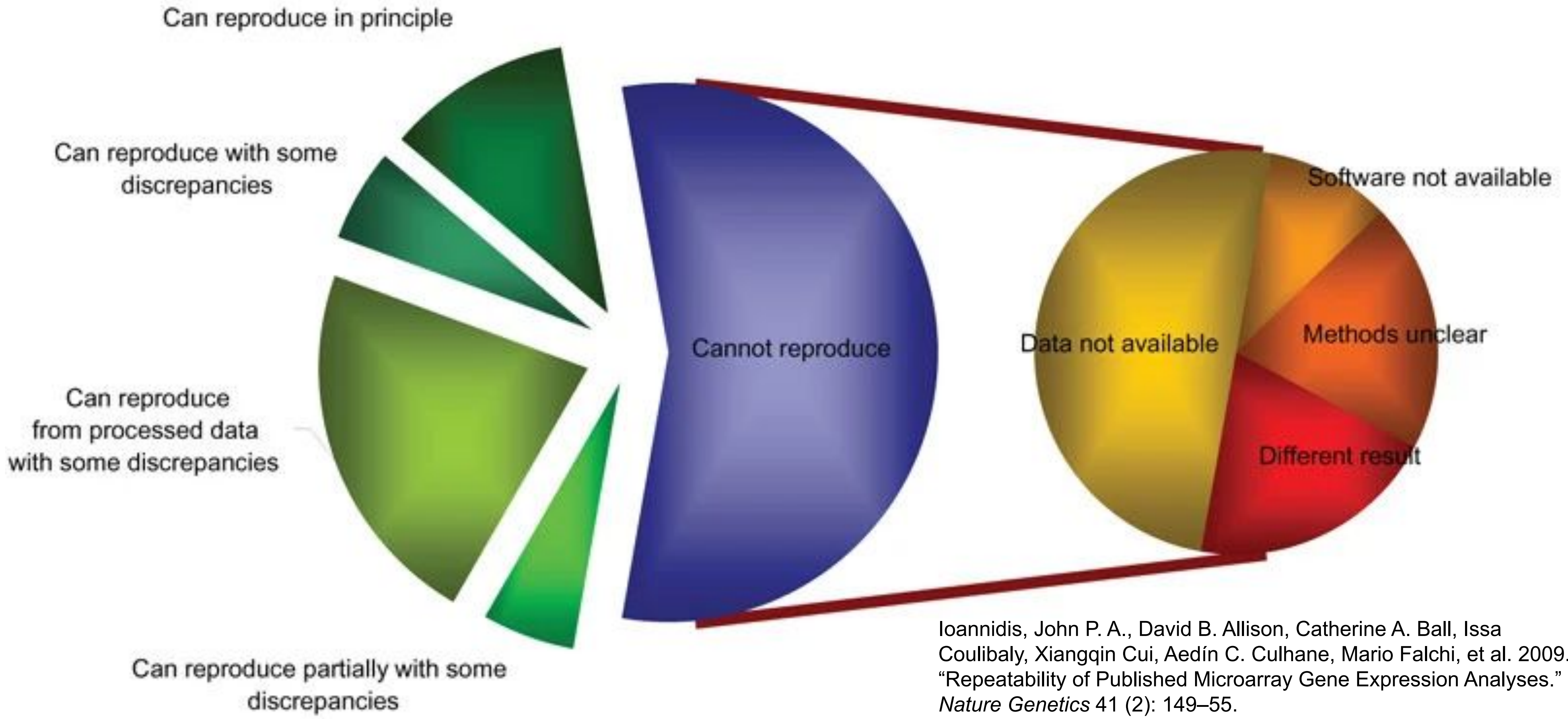
Think of reproducible science as **"tidy" code** and **data**. By keeping things tidy , labeled and organized you **create a space where you can invite guests**.

Reproducible science leads to **more meaningful collaborations** where collaborators can review your code, learn from it, and **build from you work**.

# Reproducible Science - for science

camh | Krembil Centre for Neuroinformatics

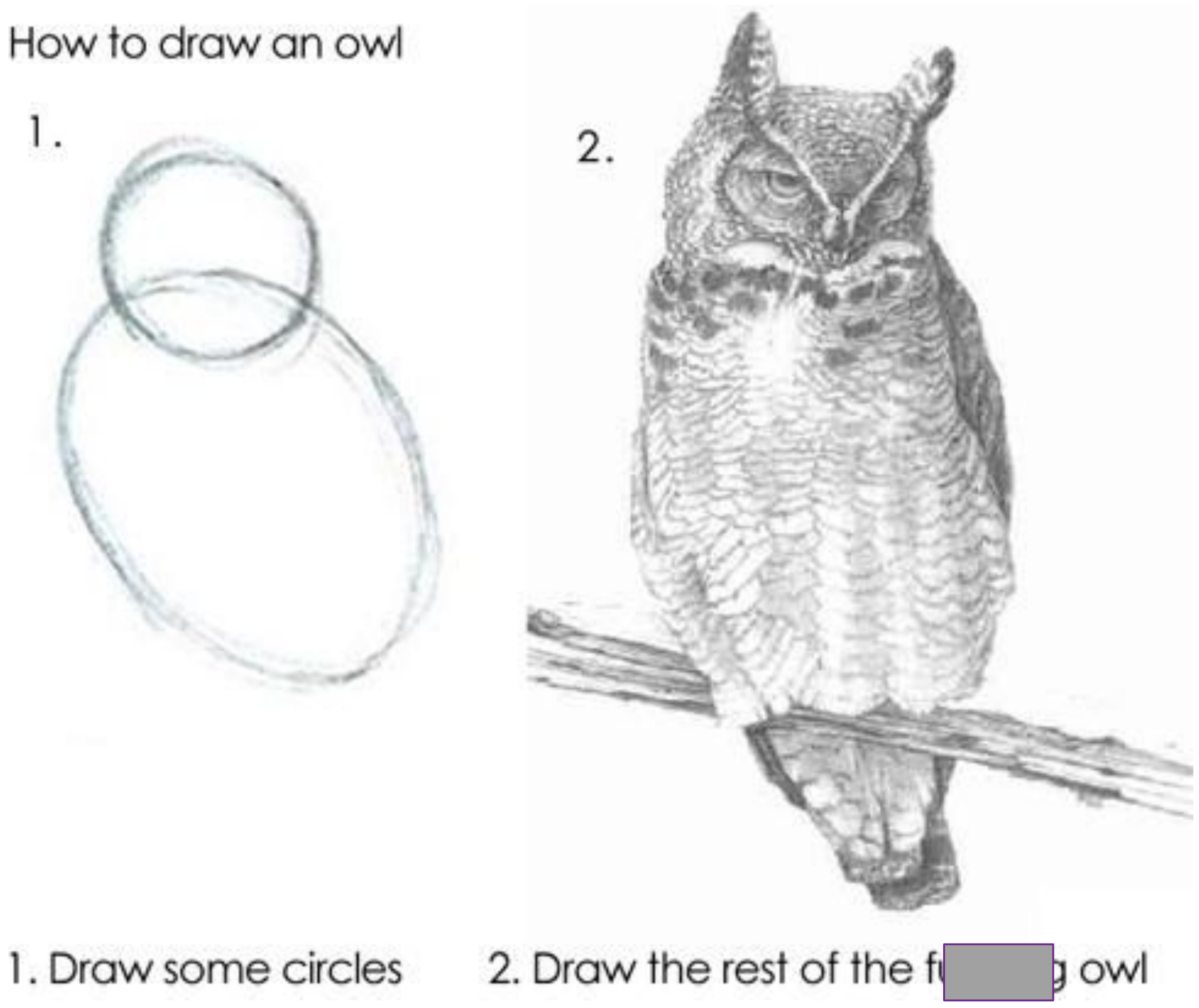
Tested the reproduction of data analyses in 18 articles on microarray-based gene expression profiling published in *Nature Genetics* in 2005–2006.



Ioannidis, John P. A., David B. Allison, Catherine A. Ball, Issa Coulibaly, Xiangqin Cui, Aedín C. Culhane, Mario Falchi, et al. 2009. “Repeatability of Published Microarray Gene Expression Analyses.” *Nature Genetics* 41 (2): 149–55.

## Communicating to your peers

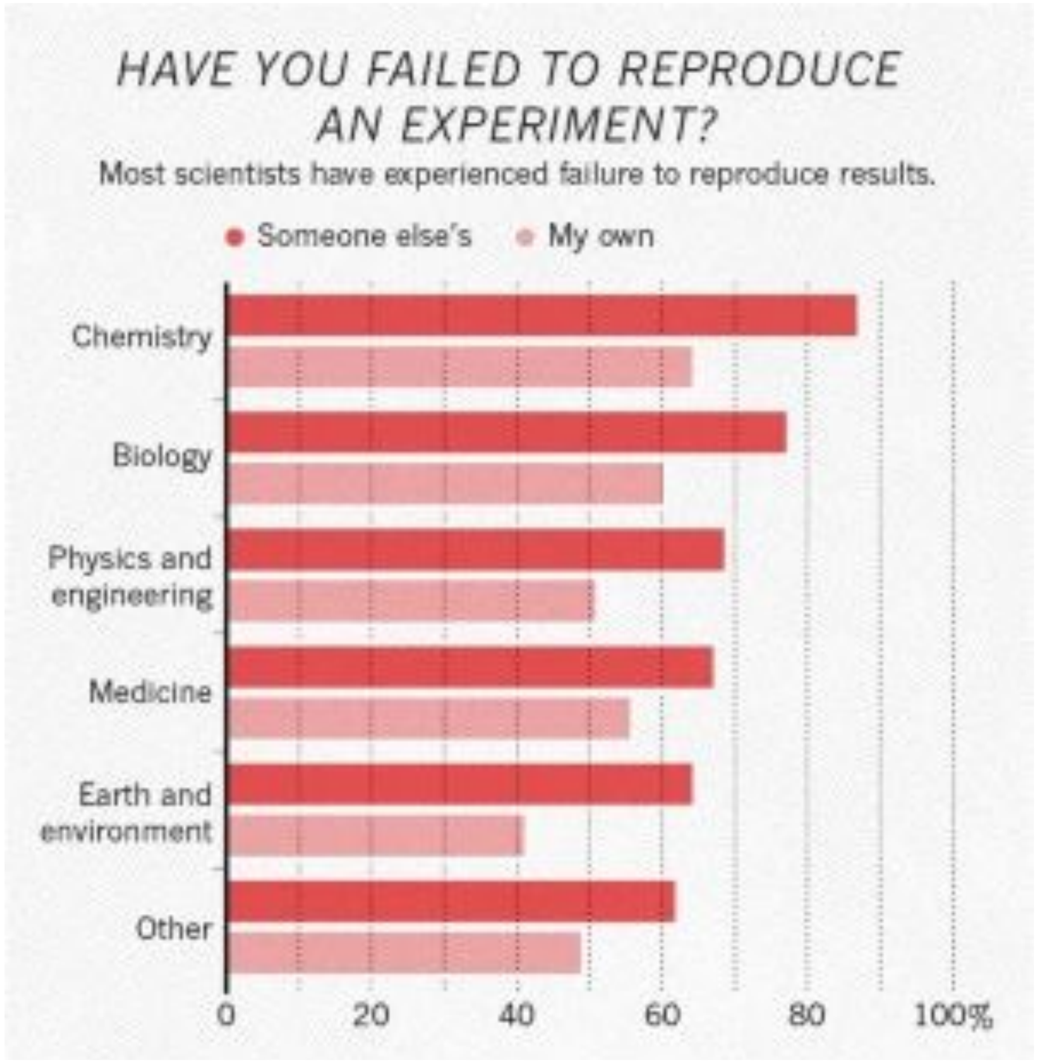
**Problem:** today's scientific papers can fail to communicate all details of the methods needed to reproduce the study.



# Reproducible Science - for science

camh | Krembil Centre for Neuroinformatics

Baker, Monya. 2016. “1,500 Scientists Lift the Lid on Reproducibility.” *Nature* 533 (7604): 452–54.



# Problems and Suggested Solutions

camh | Krembil Centre for Neuroinformatics

Poldrack, Russell A., Chris I. Baker, Joke Durnez, Krzysztof J. Gorgolewski, Paul M. Matthews, Marcus R. Munafò, Thomas E. Nichols, Jean-Baptiste Poline, Edward Vul, and Tal Yarkoni. 2017. “Scanning the Horizon: Towards Transparent and Reproducible Neuroimaging Research.” *Nature Reviews. Neuroscience* 18 (2): 115–26. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6910649/

| Problems | Suggested Solutions |
|---|---|
| Low statistical power | (one or many)  increasing the amount of available data for analysis |
| Flexibility and exploration in data analysis | pre-registration of methods and analysis plans |
| Multiple comparisons | sharing of both thresholded and un-thresholded result maps (for meta analysis) |
| Software errors<br>“As complexity of a software program increases, the likelihood of undiscovered bugs quickly reaches certainty” | avoid the trap of the ‘not invented here’ philosophy: when the problem at hand can be solved using software tools from a well-established project, these should be chosen instead of re-implementing the same method in custom code |
| Insufficient study reporting | Because the computer code is often necessary to understand exactly how a data set has been analysed, releasing the analysis code is particularly useful and should be standard practice. |

# The paper of the future/(is now?)

Krembil Centre for Neuroinformatics

## Towards the [neuroimaging] paper of the future

**"All code for data collection and analysis would be stored in a version-control system** and would **include software tests** to detect common problems"

"The repository would use **a continuous integration system** to ensure that each revision of the code passes appropriate software tests."

**"The entire analysis workflow** (including both successful and failed analyses) would be completely automated in a workflow engine and **packaged in a software container** or virtual machine **to ensure computational reproducibility**. "

# The many bits of a project

camh | Krembil Centre for Neuroinformatics

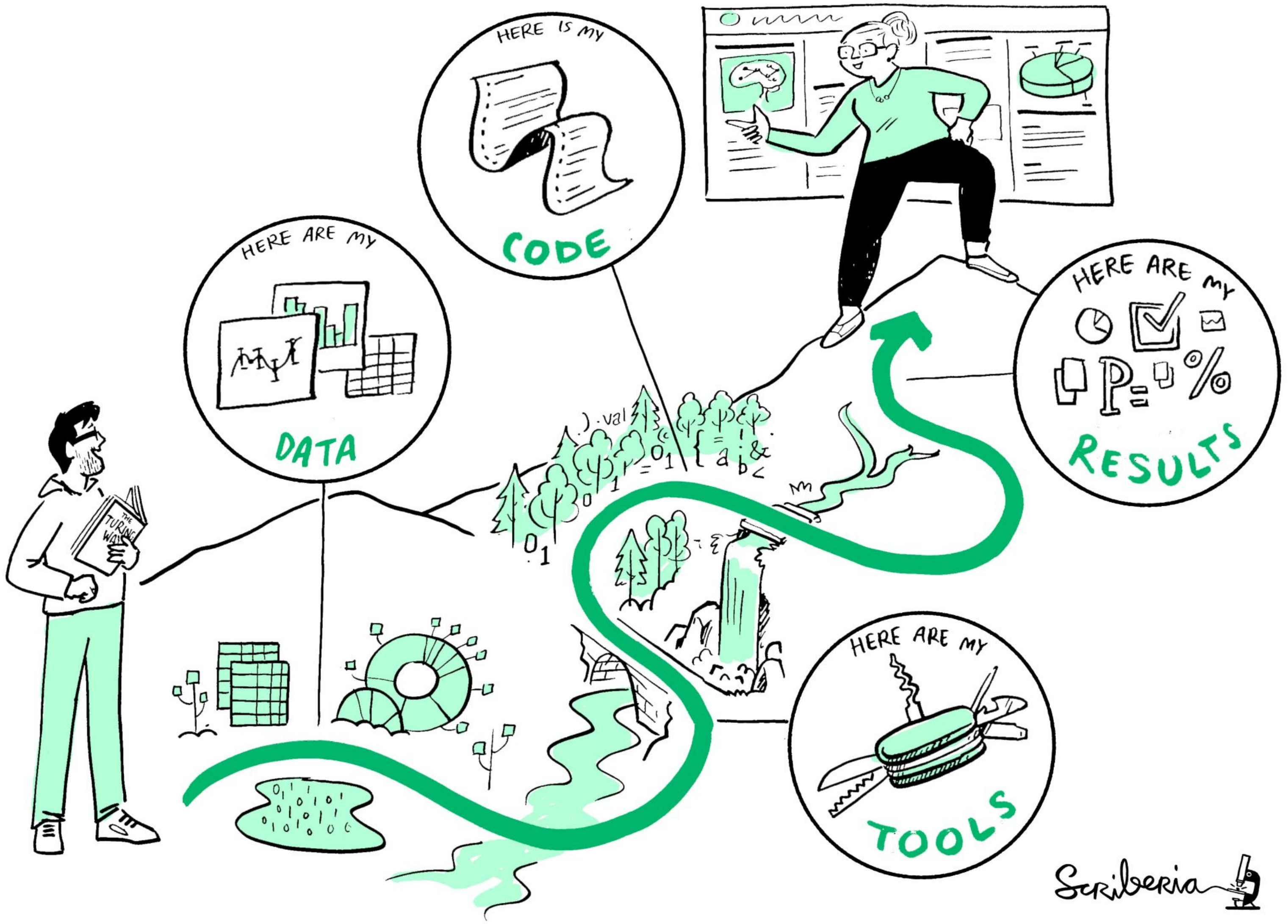


**Fig. 3** *The Turing Way* project illustration by Scriberia. Used under a CC-BY 4.0 licence. DOI: 10.5281/zenodo.3332807.¶

# FAIR data principles

camh | Krembil Centre for Neuroinformatics

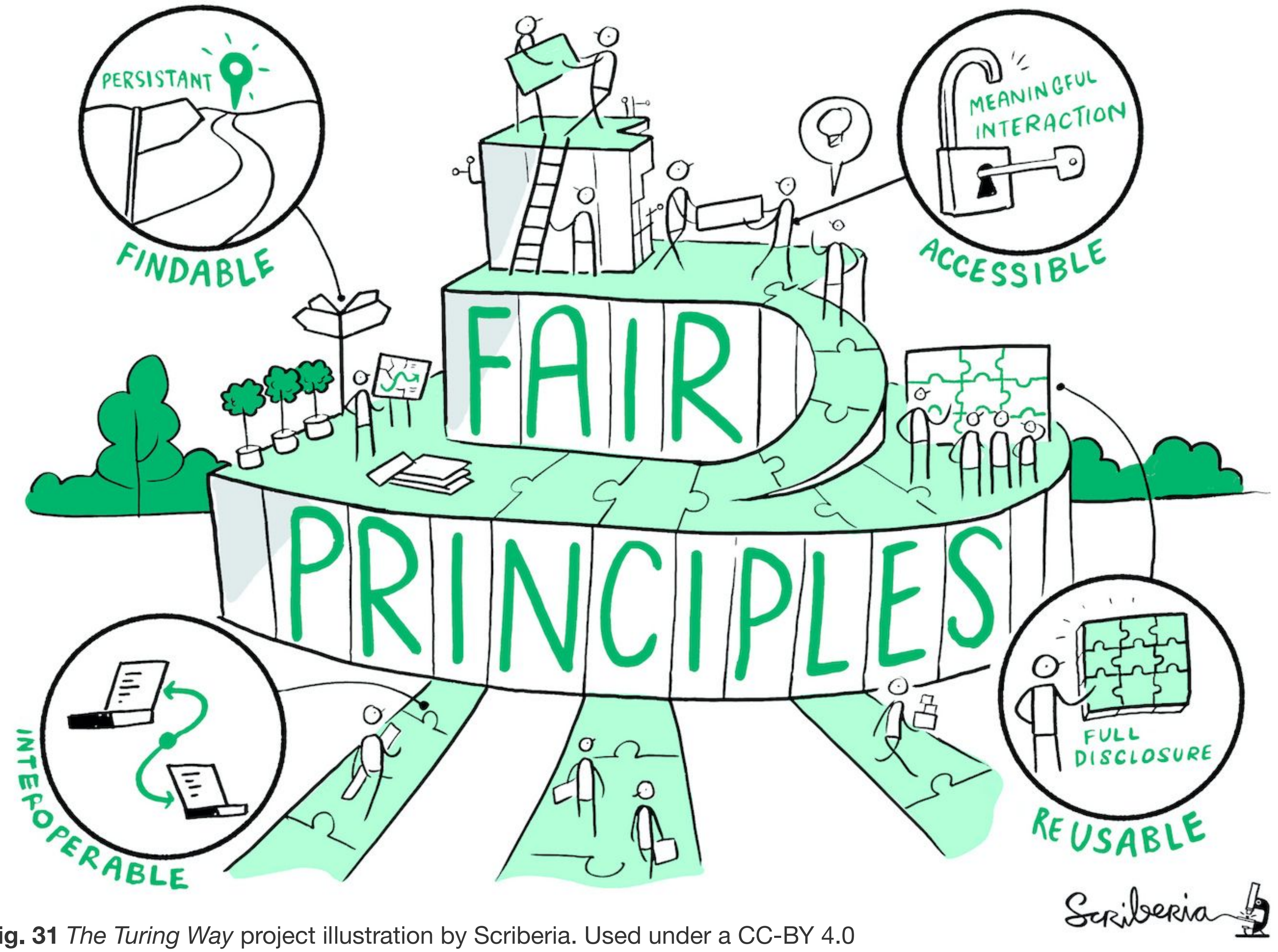


**Fig. 31** *The Turing Way* project illustration by Scriberia. Used under a CC-BY 4.0 licence. DOI: 10.5281/zenodo.3332807.¶

**Findable:** The first step in (re)using data is to find them! Descriptive metadata (information about the data such as keywords) are essential.

**Accessible:** Once the user finds the data and software they need to know how to access it. Data could be openly available but it is also possible that authentication and authorisation procedures are necessary.

**Interoperable:** Data needs to be integrated with other data and interoperate with applications or workflows.

**Reusable:** Data should be well-described so that they can be used, combined, and extended in different settings.

# Data management recommendations

camh | Krembil Centre for Neuroinformatics

0. **Organise your project as if you have no memory**
1. Raw data is read-only. (i.e. never modify your source data)
2. **Name files for both humans and computers to understand**
3. Always know where your data comes from
4. **Track all the changes the humans or computers make. Use version control.**

# Data Naming

Here are some tips for naming files within a research project, which are both human- and machine-readable [Cow20][Hod15]:

- Name your files consistently
- Keep it short but descriptive
- Avoid special characters or spaces to keep it machine-compatible
- Use capitals or underscores to keep it human-readable
- Use consistent date formatting, for example ISO 8601: YYYY-MM-DD to maintain default order
- Include a version number when applicable
- Share/establish a naming convention when working with collaborators
- Record a naming convention in your data management plan

[the turing way - https://the-turing-way.netlify.app/project-design/filenaming.html]

## Beware the WOMBAT

Krembil Centre for Neuroinformatics

**W** aste
**O** f
**M** oney
**B** rains
**A** nd
**T** ime

Do not reinvent the wheel

If a (well-maintain) tool or pipeline exists for your analysis - use it
- don’t recode your own

If a data naming standard exists for your data type - use it
- don’t invent you own

- **Michael Hanke @eknahm, BrainHack Leipzig 2012**

## If it exists - use a data standard!

Krembil Centre for Neuroinformatics

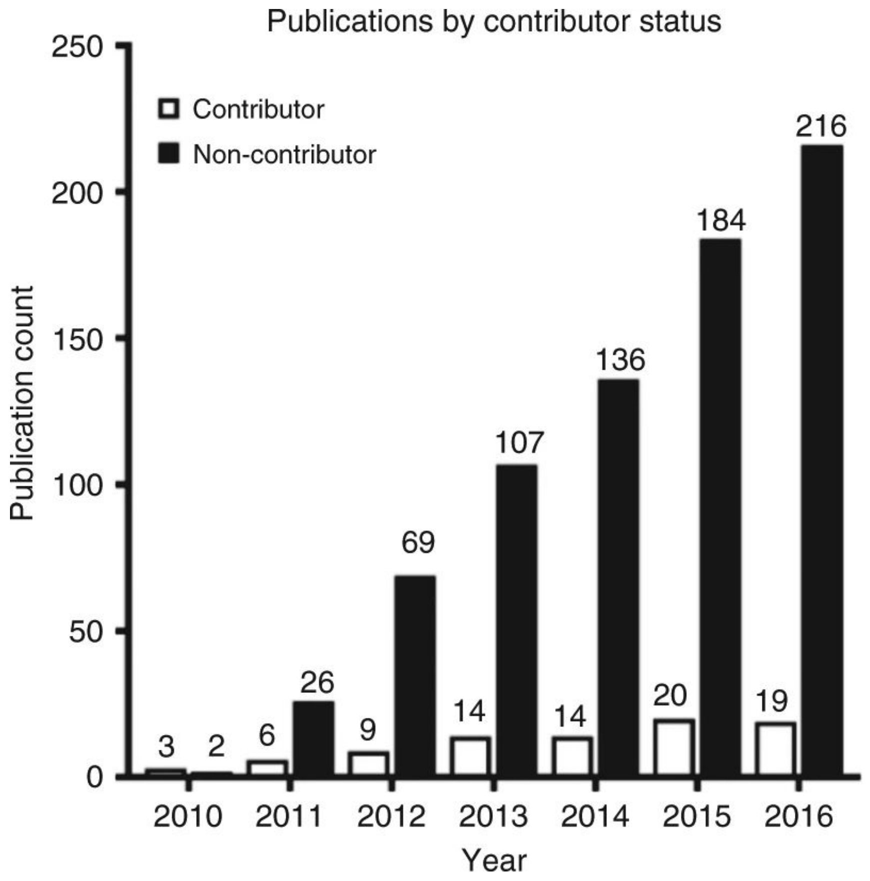
International Neuroinformatics Coordinating Facility (INCF) hosts a curated list of data standards that are useful our research

-https://www.incf.org/resources/sbps

- Brain Imaging Data Structure (BIDS) for neuroimaging and EEG
- neurodata without borders (NWB) for neurophysiology
- NeuroML for models

# The impact of open data

camh | Krembil Centre for Neuroinformatics

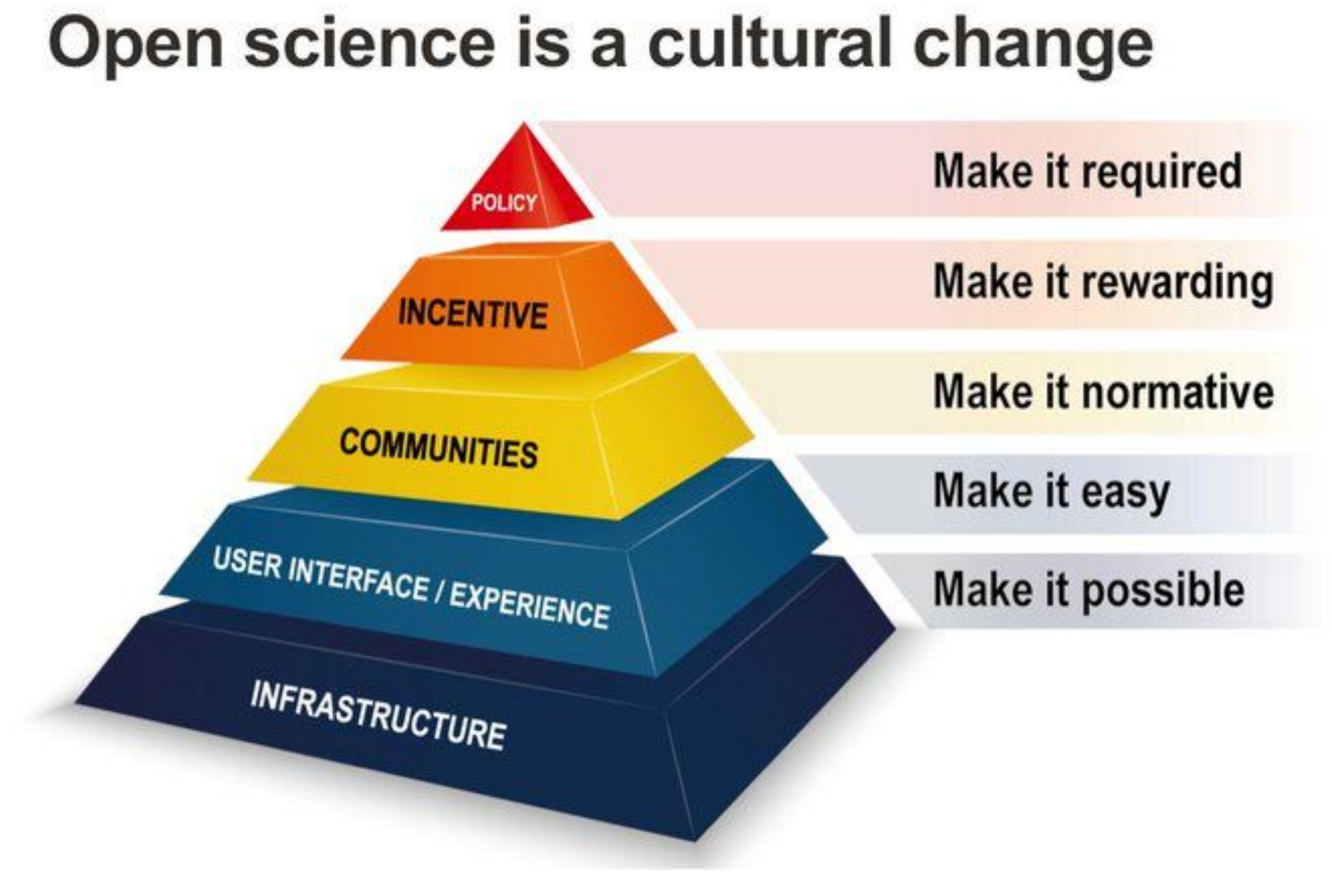


| Database | Cost/ subject | Pheno- typing | Pheno- typing | Clinical | Popu- lation | Difficulty | No. of publica- tions | No. of scans/ subject | $ Saved |
|---|---|---|---|---|---|---|---|---|---|
| | | Minimal | Compre -hensive | Low | Mod- erate | High | | | |
| FCP | $1000 | x | | | | | 308 | 1 | 101,003,000 |
| ADHD-200 | $2000–5000 | | | x | x | | 210 | 1 | 526,275,000 |
| NKI-RS | $3000 | | x | | | | 188 | 1 | 70,065,000 |
| ABIDE | $5000–10,000 | | | | x | x | 190 | 1 | 995,560,000 |
| CoRR | $2000 | x | | | | | 17 | 2 | 70,065,000 |

Milham, Michael P., R.et al. 2018. “**Assessment of the Impact of Shared Brain Imaging Data on the Scientific Literature**.” *Nature Communications* 9 (1): 2818.

# Modelling Best Practices

Krembil Centre for Neuroinformatics



These practices and tools can be difficult to incorporate into a full project with multiple stages of analysis...but they are especially useful in teaching

The materials in this course are examples of reproducible science tools.

# Outline for this afternoon

Why reproducible science?

**The things you need to know to get through this course:**

- **Versioning and publishing code (github)**
- Versioning and publishing software (docker)
- R with Rmarkdown (walk through)
- Python in ipython notebooks (walk through)
- -also google colab

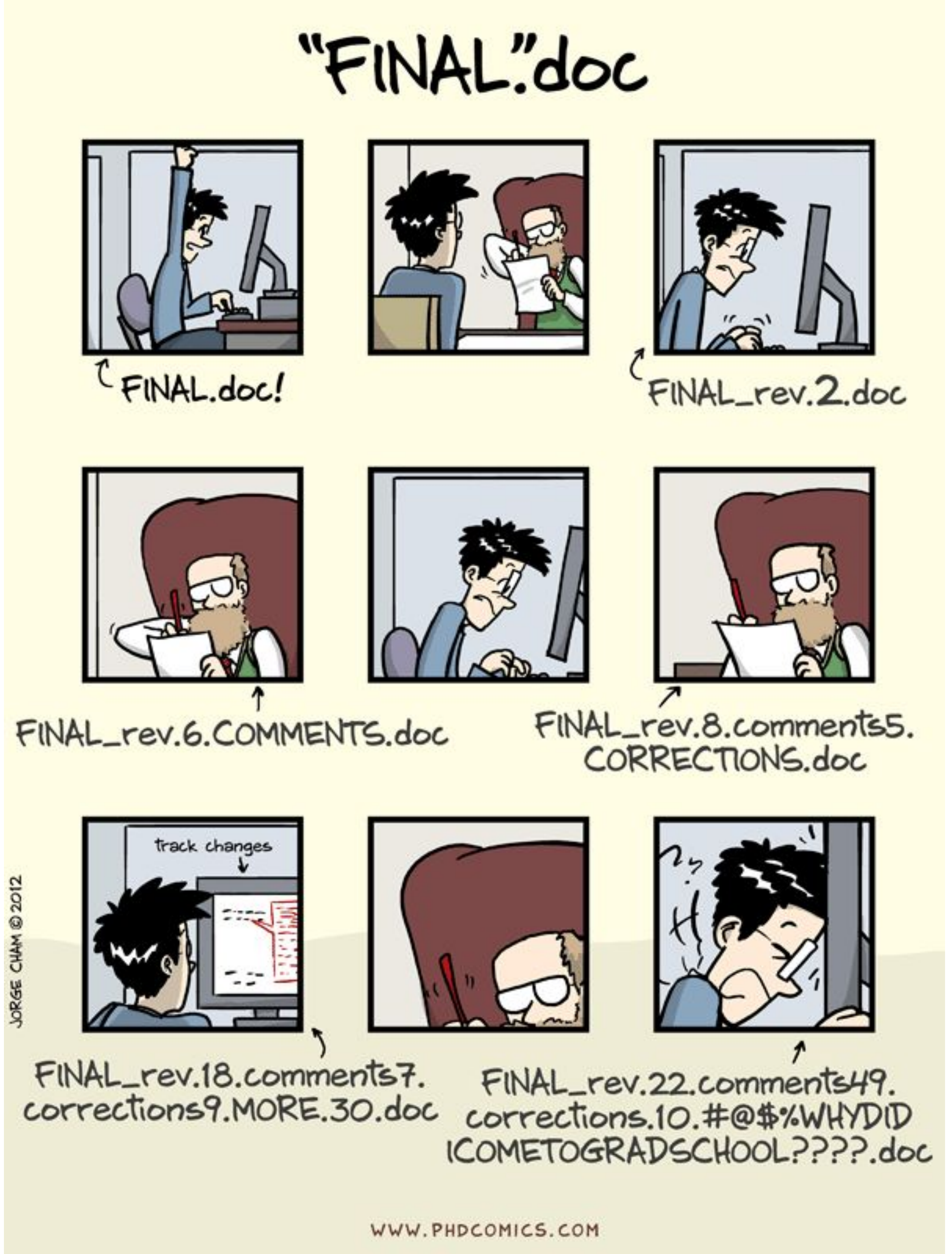
The fancy bits:

- building your own binder environments
- building your own containers (docker & singularity)

## Using version control

**version control: ('git')** is a tool for tracking what changes to a folder (usually a folder filled with code) when and by who..

- Like MS Word's "track changes"
  - ...but for code
  - ..and on steroids..


"FINAL".doc
FINAL.doc!
FINAL_rev.2.doc
FINAL_rev.6.COMMENTS.doc
FINAL_rev.8.comments5. CORRECTIONS.doc
track changes
FINAL_rev.18.comments7. corrections9.MORE.30.doc
FINAL_rev.22.comments49. corrections.10.#@$%WHYDID ICOMETOGRADSCHOOL????.doc
JORGE CHAM © 2012
WWW.PHDCOMICS.COM

## Git and Github - What and Why?

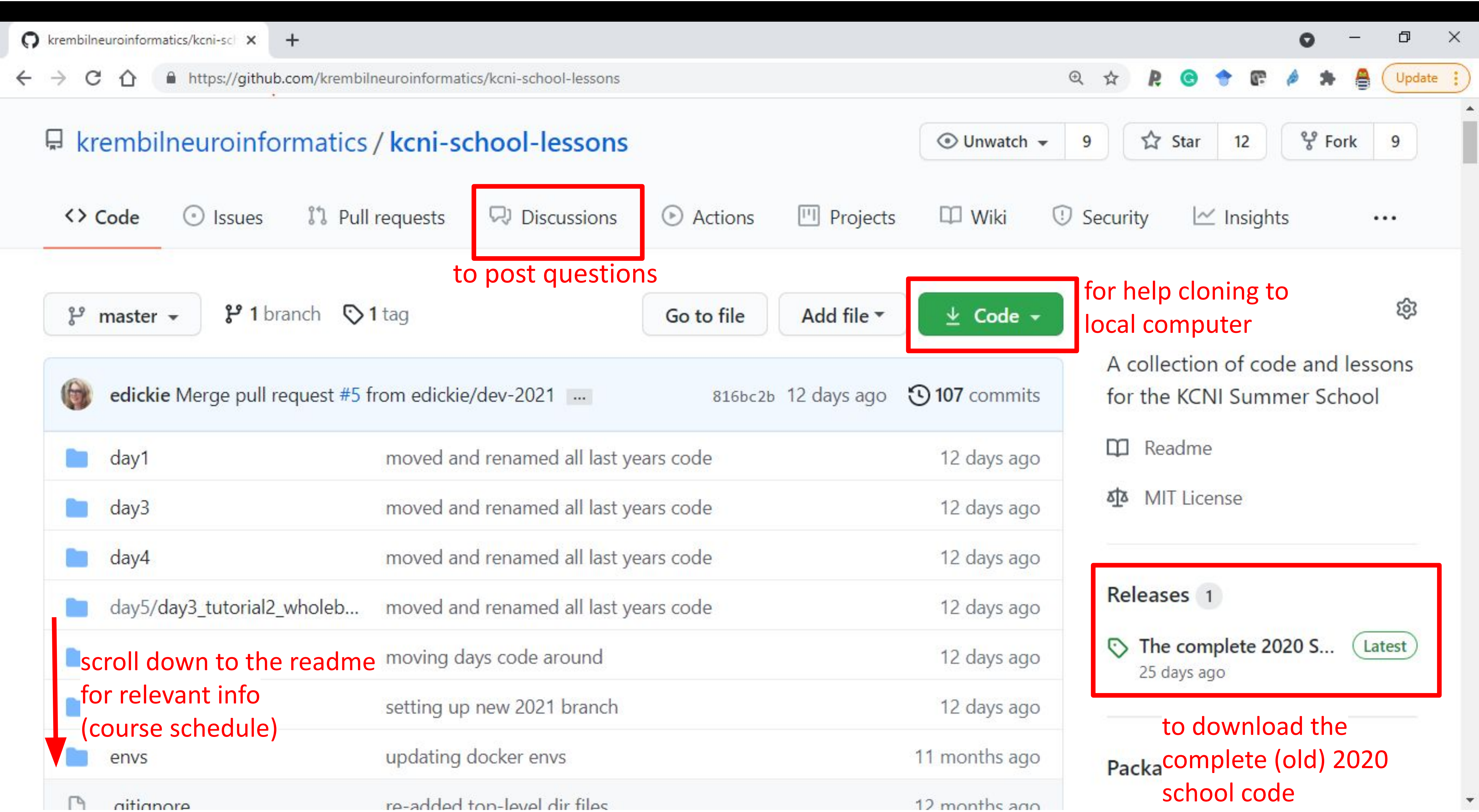
camh
Krembil Centre for Neuroinformatics

GitHub

**Github** is a website where everyone shares there code with themselves, their teams and with the world.

It has a a lot of useful features for:

- working with teams
- reading other people's code
- integrating with other platforms
  - continuous integration (CI) to test for bugs
  - Dockerhub
- hosting documentation websites and wiki's
- releasing versions



krembilneuroinformatics/kcni-sc

https://github.com/krembilneuroinformatics/kcni-school-lessons

krembilneuroinformatics / kcni-school-lessons

Unwatch 9 | Star 12 | Fork 9

Code | Issues | Pull requests | Discussions | Actions | Projects | Wiki | Security | Insights

to post questions

master | 1 branch | 1 tag | Go to file | Add file | Code

for help cloning to local computer

edickie Merge pull request #5 from edickie/dev-2021 … 816bc2b 12 days ago 107 commits

| | | |
|---|---|---|
| day1 | moved and renamed all last years code | 12 days ago |
| day3 | moved and renamed all last years code | 12 days ago |
| day4 | moved and renamed all last years code | 12 days ago |
| day5/day3_tutorial2_wholeb… | moved and renamed all last years code | 12 days ago |
| | moving days code around | 12 days ago |
| | setting up new 2021 branch | 12 days ago |
| envs | updating docker envs | 11 months ago |
| .gitignore | re-added top-level dir files | 12 months ago |

scroll down to the readme for relevant info (course schedule)

A collection of code and lessons for the KCNI Summer School

Readme

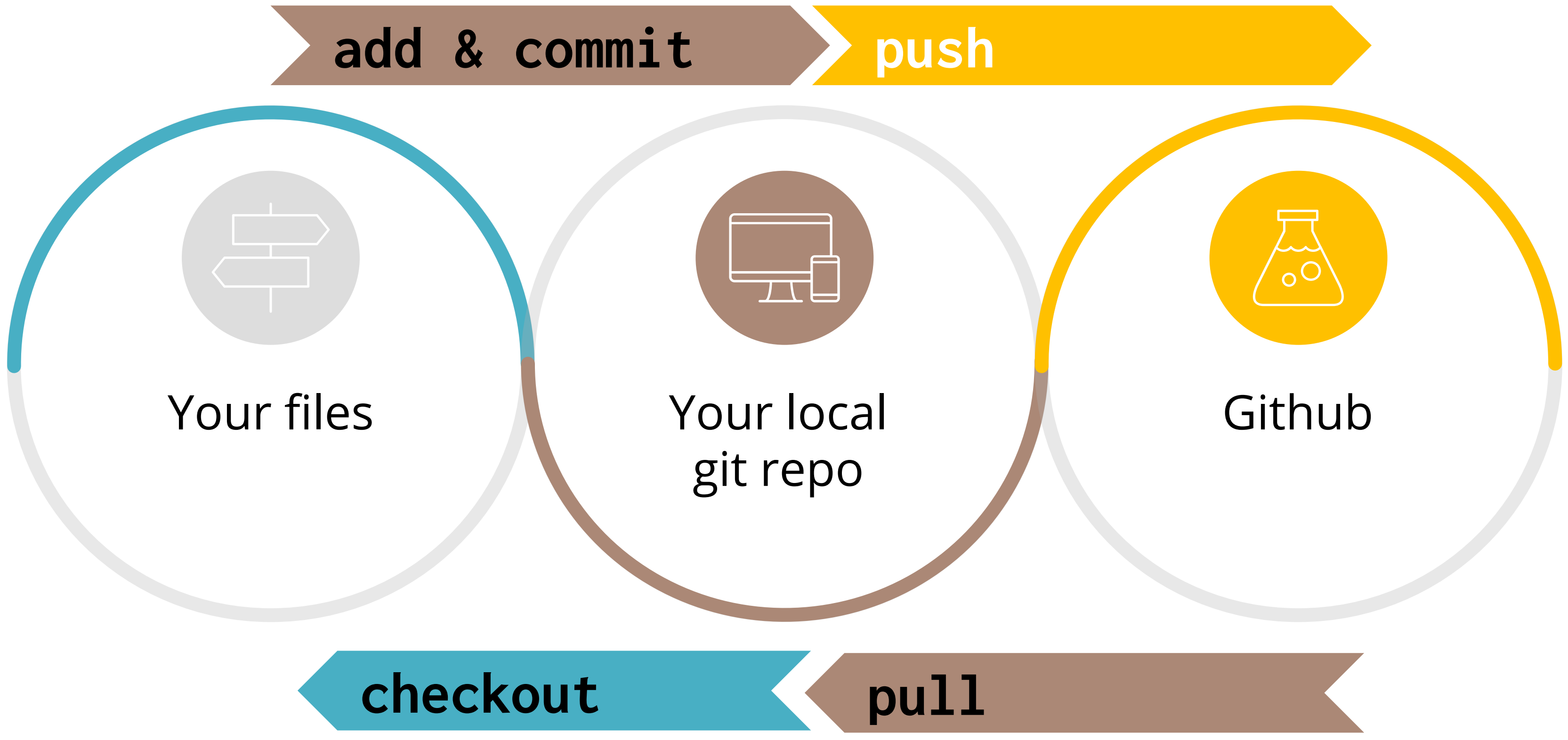
MIT License

Releases 1

The complete 2020 S… Latest
25 days ago

to download the complete (old) 2020 school code

Packa



**`git status`**
List what has not been committed (repo)

**`git diff`**
tells you want changes haven't been committed (file)

**`git init`**
starts a new repo

**`git clone`**
copies a repo from GitHub or GitLab to your local computer

## Git for your own project demo

# Git - make a repo

Step 1: build repo on github/gitlab

Step 2:

```
    cd ~/code
   git clone http:link/to/git/repo.git
```

| | |
|---|---|
| First thing in the morning | `git pull` |
| Before coffee break | `git commit bin/scriptx.py -m “I changed x”` |
| Before going to lunch | `git commit bin/scripty.py -m “modifying y”` |
| Done adding function foo | `git commit bin/sciprtx.py -m “now does foo”` |
| Before going home | `git push` |

" Git, the toothbrush of science"

# Git with friends..

camh | Krembil Centre for Neuroinformatics

Let go look at the lesson’s repo on github

- when working with larger groups - it’s usually better to put in a little more work to make sure that your changes don’t clash with other people’s changes
- some projects have “CONTRIBUTING.md” files that lay down some best practice steps

- **“fork”** the repo
  - make a copy under your own user space
- create a new **“branch”** that specific to you change
- make your changes inside the **“branch”**
- when you are done make a **“pull request”** back to the **“upstream”** repo

# Git submodules - what? and why?

**What?** - a git **"submodule"** is link from one git repository to another repository

- *think of the **kcni-school-lessons** repo as an **index** of lesson repos*
- on github you see the "submodule" as a hyperlink to another git repo ***at a specific commit***
- on your computer - a submodule looks like a git repo inside another git repo.

**Why?** - it can allow you to keep project **more modular**

- meaning you can use it to link to a common code base that is shared across multiple projects

- it can help git run faster (for very large projects)

## Git submodules - how - for everybody

camh | Krembil Centre for Neuroinformatics

**Get the tutorial scripts (with all the submodules) by typing**

```
git clone --recurse-submodules \
  https://github.com/krembilneuroinformatics/kcni-school-lessons.git
```

**Get the updated tutorial scripts (with all the submodules) after cloning**

```
cd kcni-school-lessons
git pull --recurse-submodules
```

## Git submodules - how - for everybody

camh | Krembil Centre for Neuroinformatics

**But what …**
**you might find, when navigating the kcni-school-lesson that the submodules folders are empty…this is because the submodule needs to he “updated”**

```
cd kcni-school-lessons
cd day1
git submodule init
git submodule update example-python-repo
```

# Git submodules - to create your own

camh | Krembil Centre for Neuroinformatics

A good tutorial for everything:

https://git-scm.com/book/en/v2/Git-Tools-Submodules

*For example here’s some code to add SPM12 toolbox code into kcni-school-lesson’s day 6 code.*

**Get add a submodule to your repo**

```
cd kcni-school-lessons/day6/toolboxes
git submodule add https:/github.com/spm/spm12.git
```

**Get updated the submodule content to the newest commit**

```
cd kcni-school-lessons
git submodule update --remote spm12
```

## Outline for this afternoon

Krembil Centre for Neuroinformatics

Why reproducible science?

The things you need to know to get through this course:

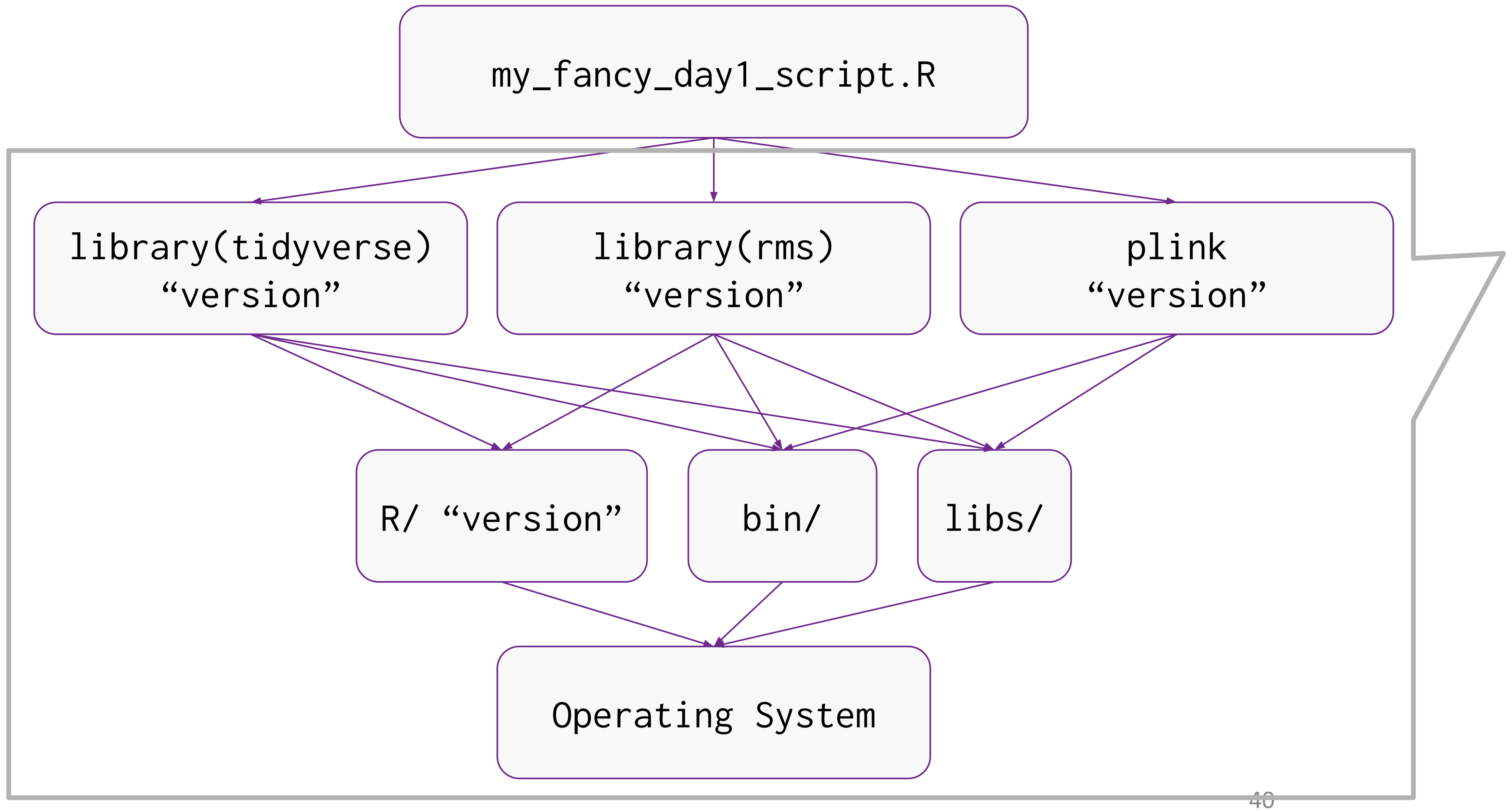
- Versioning and publishing code (github)
- **Versioning and publishing software (docker)**
- R with Rmarkdown (walk through)
- Python in ipython notebooks (walk through)
- -also google colab

The fancy bits:

- building your own binder environments
- building your own containers (docker & singularity)

# Containers - Why

camh
Krembil Centre for Neuroinformatics



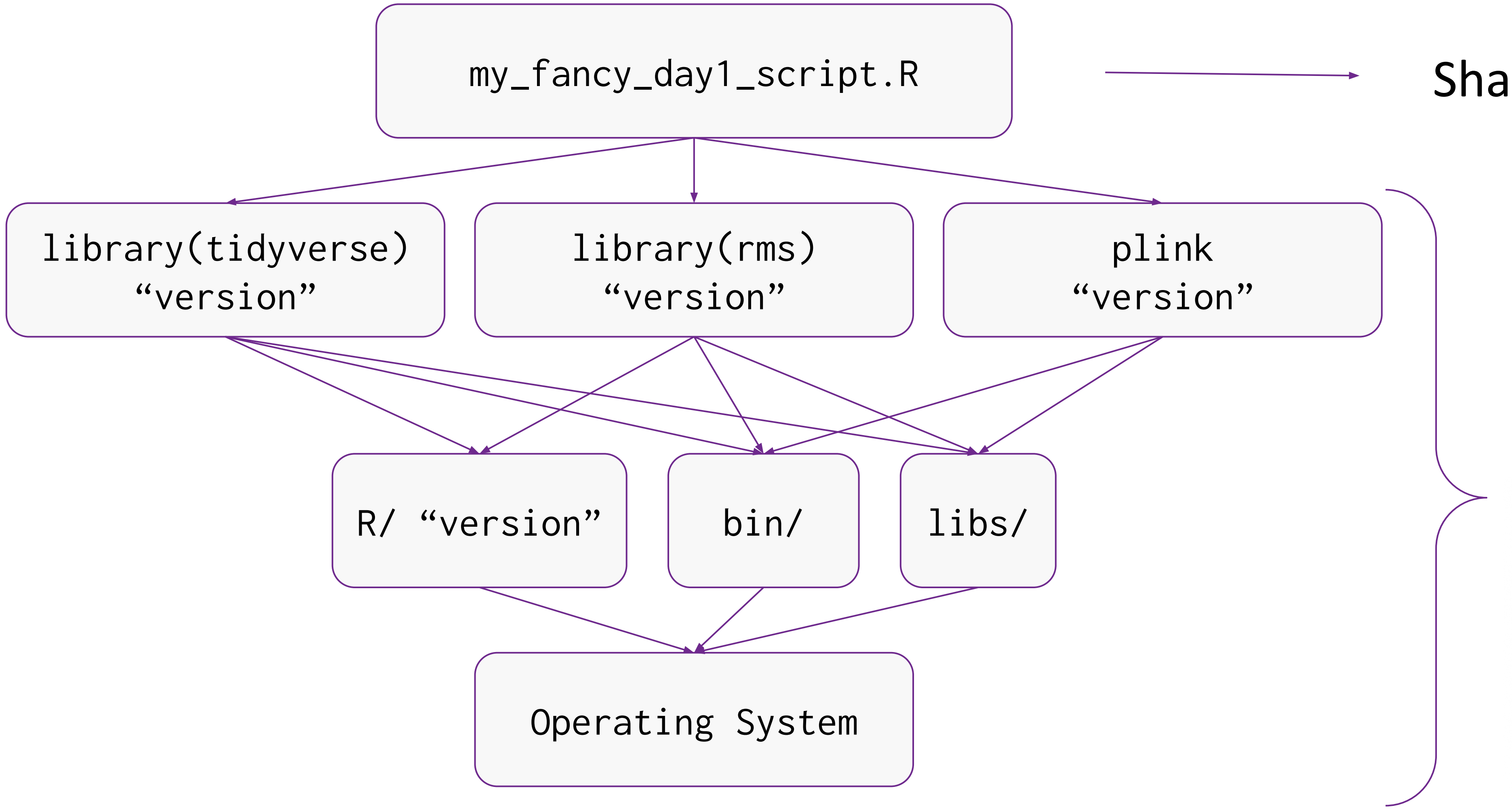
Every box here is something that could be installed differently (*or not at all*) by the next user

This will cause `my_fancy_day1_script.R` to:

1) crash/not work at all
2) produce unexpected/ different results
3) maybe still work?

# Reproducible Neuroinformatics - Solution

camh Krembil Centre for Neuroinformatics

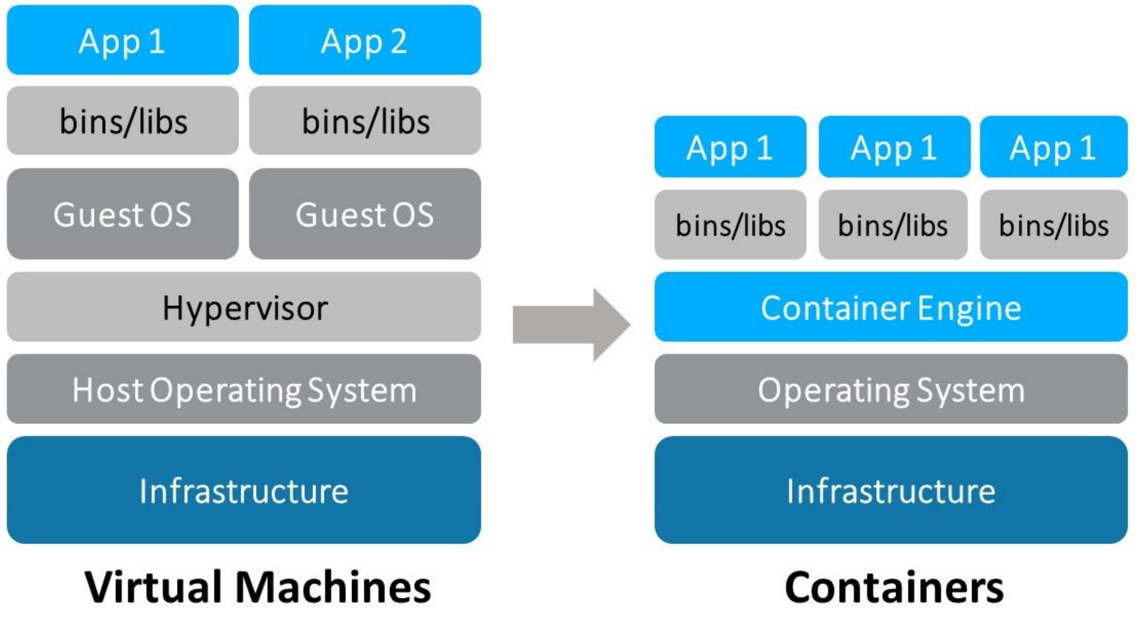


Share on Github

Share on Dockerhub

# Containers - what and why?

camh
Krembil Centre for Neuroinformatics



**Docker** - is a tool for sharing software + the dependencies

- the install instructions are stored script called "Dockerfile"
- it's like a virtual machine
  - without a display
  - that takes up a little less disk space
  - that can be installed in one line

Some Docker vocabulary

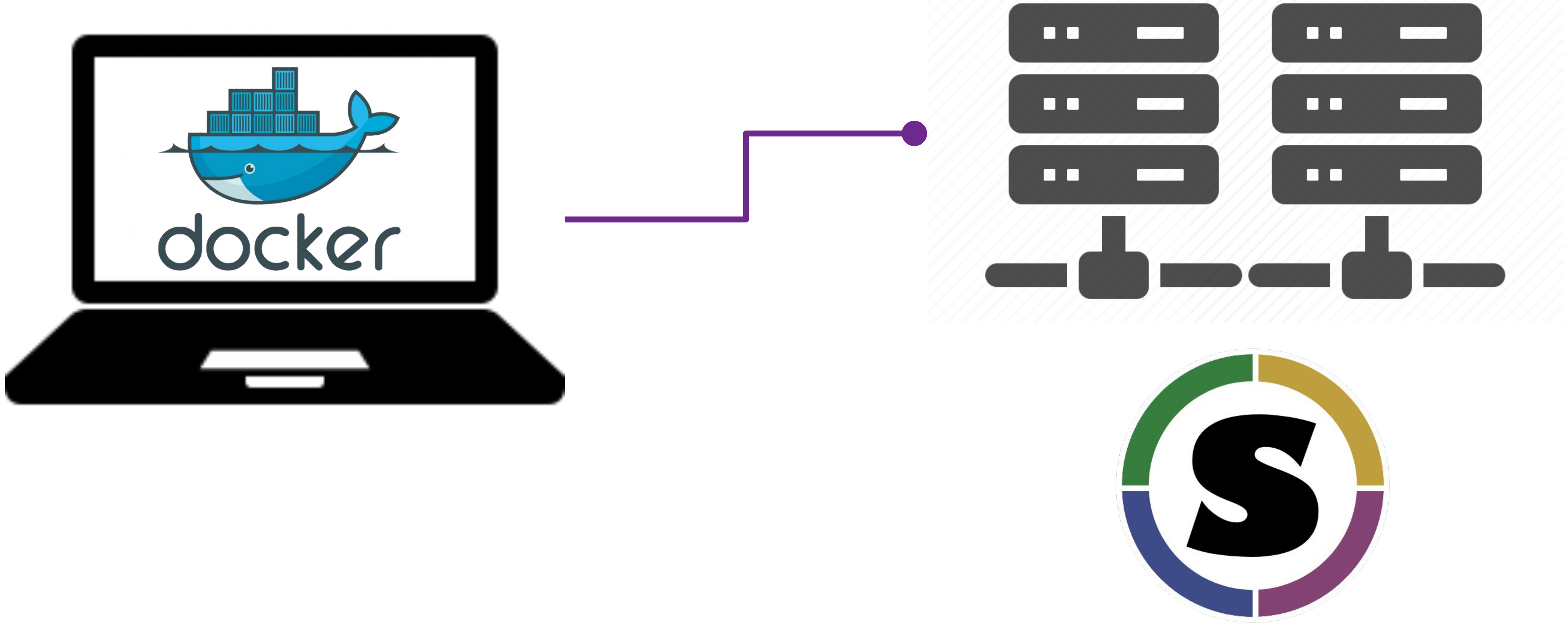
- **image**: your install of the software
- **container**: one instance of that software that is *usually* still running.

**Dockerhub** is a website that hosts docker images. So that anyone - anywhere is the world can run it!

# Containers for scaling up analysis

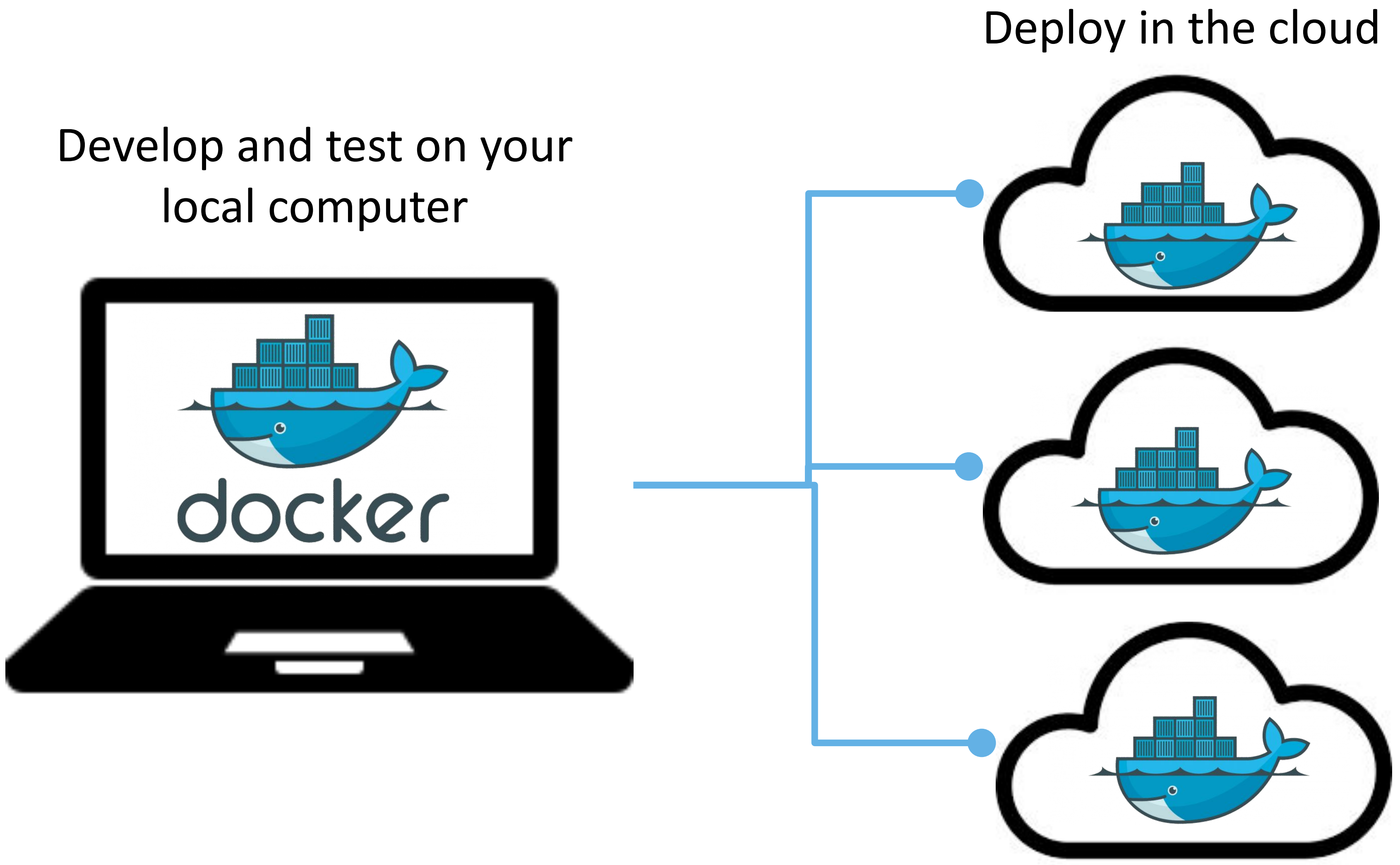
Develop and test on your local computer

Deploy software on a high performance cluster



# Containers for medical science

camh | Krembil Centre for Neuroinformatics



There is some speculation that (secure) cloud computing will become important of health research

# How to run our code

camh | Krembil Centre for Neuroinformatics

**Run on your Local Computer using Docker**

- will use less internet bandwidth while you watch the stream
- you will have a copy of the files locally
- requires installation of Docker Desktop

**Run on the web with Binder**

- no local installation needed
- may take some time to boot up
- limited resources for the computer

**Run on SciNet using a guest account**

- no local installation needed
- a few extra set-up steps needed
- good compute resources

# To follow along

camh | Krembil Centre for Neuroinformatics

## To interact with the school lesson code on your local computer you will need:

Docker Desktop

https://www.docker.com/products/docker-desktop

- installs on window, mac or linux

A terminal for pulling/cloning data from github.

-mac os terminal or WSL work

git bash

- On windows - gitbash is an option

**Also useful, but not necessary:**

VS Code

Desktop

# Step 1: Install Docker Desktop

Installing Docker *should* not be harder than installing any other program on your computer.

Download link and install instructions at: https://www.docker.com/products/docker-desktop.

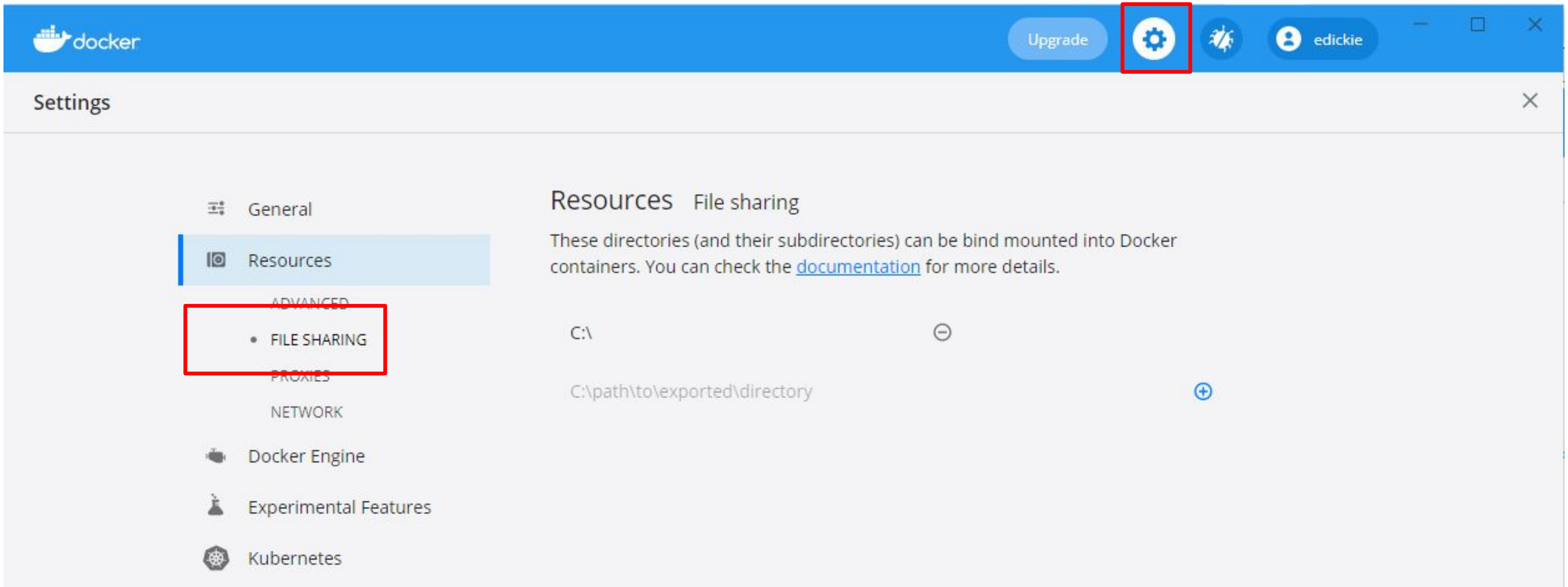
To check your install open up a terminal (in windows this is Powershell or WSL) and type:

```
docker run hello-world
```

# Docker Desktop install gotcha's

camh | Krembil Centre for Neuroinformatics

1. On Windows - you need to enable Hyper-V or WSL virtualization
2. You also need "Share the drive" with docker.
   a. Settings->Resources->File Sharing



Resources File sharing

These directories (and their subdirectories) can be bind mounted into Docker containers. You can check the documentation for more details.

C:\

C:\path\to\exported\directory

## If this fails! Fear not - we have a plan!

Krembil Centre for Neuroinformatics

***If you can't install Docker on your local computer (because you probably don't have enough administrative rights - or you don't have enough space on you home computer) We have a plan!***

***You can run the software on the binder instance or SciNet teach cluster!***

Instructions are available at: https://github.com/edickie/kcni-school-lessons

## Step 2: download and run the KCNIschool docker

camh
Krembil Centre for Neuroinformatics

In the same terminal window where you typed “docker pull hello-world” now type:

```
git clone --recurse-submodules \
  https://github.com/krembilneuroinformatics/kcni-school-lessons.git
cd kcni-school-lessons
docker compose up rstudio
```

Then you should see lots of things happening! What is happening? - *docker is downloading ~ 5G of software for our lessons into an “image”*

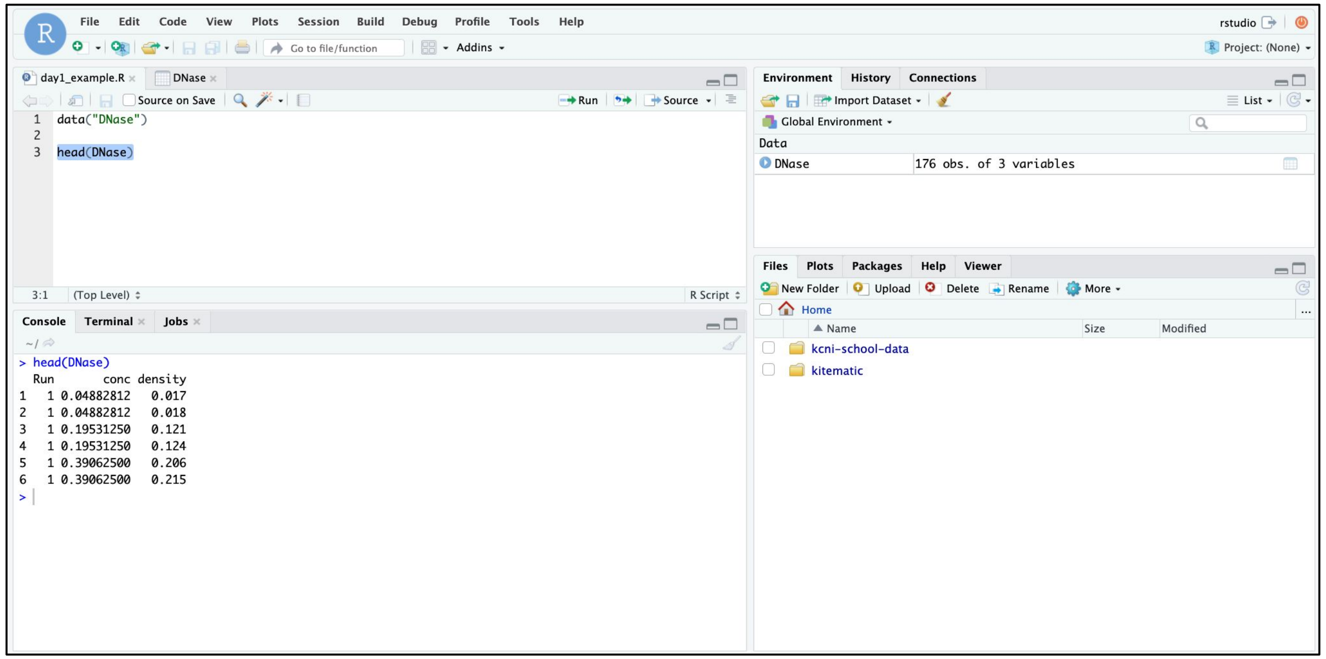
Copy and paste this line from:
https://github.com/edickie/kcni-school-lessons/tree/master/envs/README.md

# Step 3: open rstudio in browser

After typing

```
cd kcni-school-lessons
docker compose up rstudio
```

You will finally see the message `[services.d] done.`
point you browser to:
http://localhost:8787/



```
data("DNase")

head(DNase)
```

```
> head(DNase)
  Run       conc density
1   1 0.04882812   0.017
2   1 0.04882812   0.018
3   1 0.19531250   0.121
4   1 0.19531250   0.124
5   1 0.39062500   0.206
6   1 0.39062500   0.215
>
```

Data
DNase 176 obs. of 3 variables

kcni-school-data
kitematic

# Outline for this afternoon

Krembil Centre for Neuroinformatics

Why reproducible science?

The things you need to know to get through this course:

- Versioning and publishing code (github)
- Versioning and publishing software (docker)
- **R with Rmarkdown (walk through)**
- Python in ipython notebooks (walk through)
- -also google colab

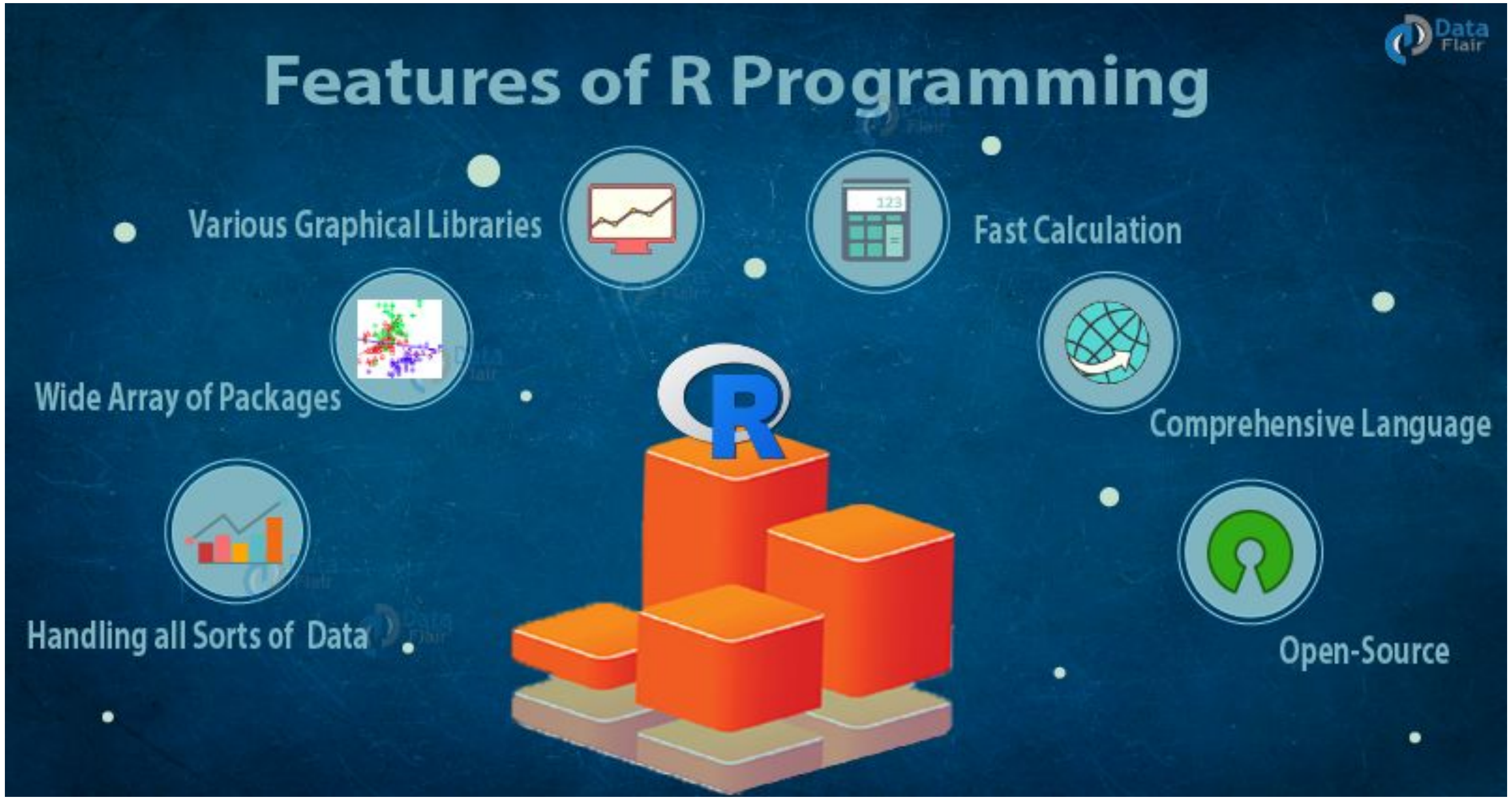
The fancy bits:

- building your own binder environments
- building your own containers (docker & singularity)

## R Programming

**R is a language and environment for statistical computing and graphics for data visualization**

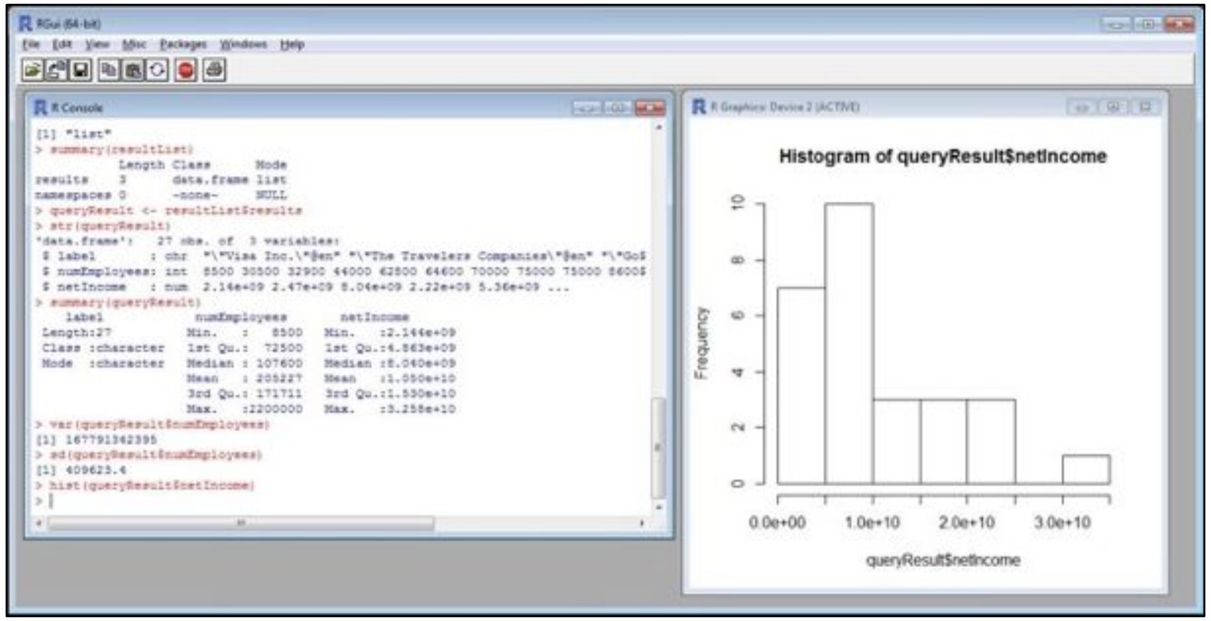
- Similar to S programming language and environment



# R GUI and RStudio

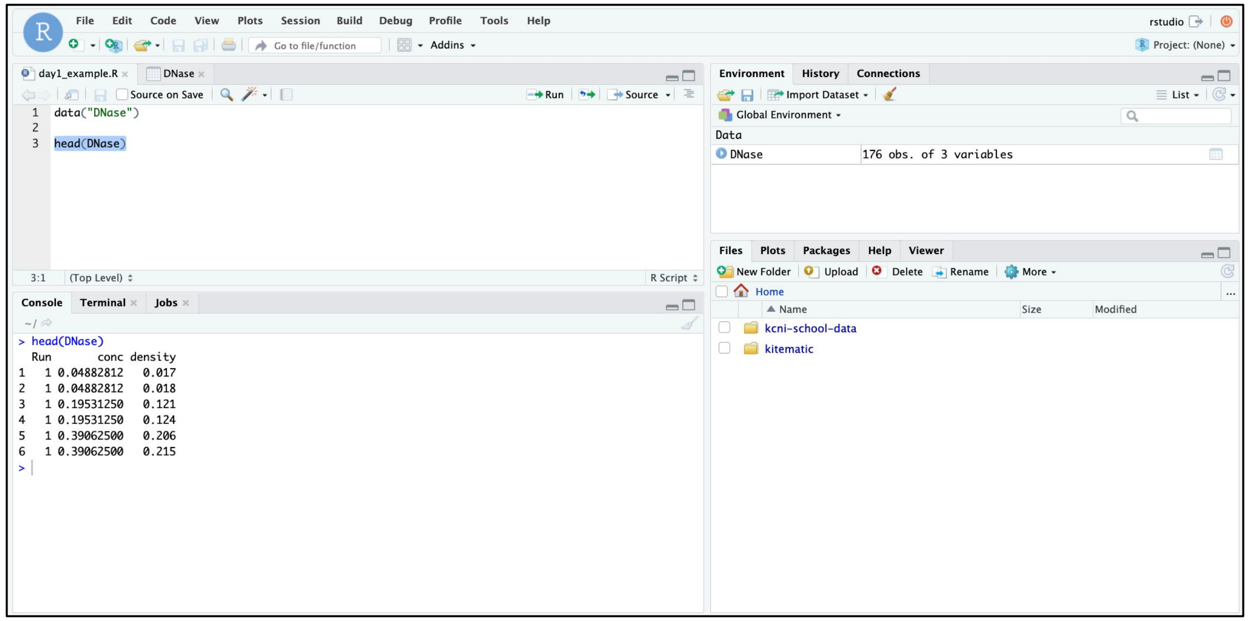
camh Krembil Centre for Neuroinformatics

## Old way of coding in R


RGui (64-bit)
File Edit View Misc Packages Windows Help
R Console
R Graphics: Device 2 (ACTIVE)
Histogram of queryResult$netIncome
Frequency
queryResult$netIncome

```
[1] "list"
> summary(resultList)
           Length Class      Mode
results    3      data.frame list
namespaces 0      -none-     NULL
> queryResult <- resultList$results
> str(queryResult)
'data.frame':   27 obs. of  3 variables:
 $ label       : chr  "\"Visa Inc.\"@en" "\"The Travelers Companies\"@en" "\"Go$
 $ numEmployees: int  8500 30500 32900 44000 62800 64600 70000 75000 75000 8600$
 $ netIncome   : num  2.14e+09 2.47e+09 8.04e+09 2.22e+09 5.36e+09 ...
> summary(queryResult)
    label             numEmployees       netIncome
 Length:27          Min.   :   8500   Min.   :2.144e+09
 Class :character   1st Qu.:  72500   1st Qu.:4.863e+09
 Mode  :character   Median : 107600   Median :8.040e+09
                    Mean   : 205227   Mean   :1.050e+10
                    3rd Qu.: 171711   3rd Qu.:1.530e+10
                    Max.   :2200000   Max.   :3.258e+10
> var(queryResult$numEmployees)
[1] 167791362395
> sd(queryResult$numEmployees)
[1] 409623.4
> hist(queryResult$netIncome)
> |
```

## New way of coding R



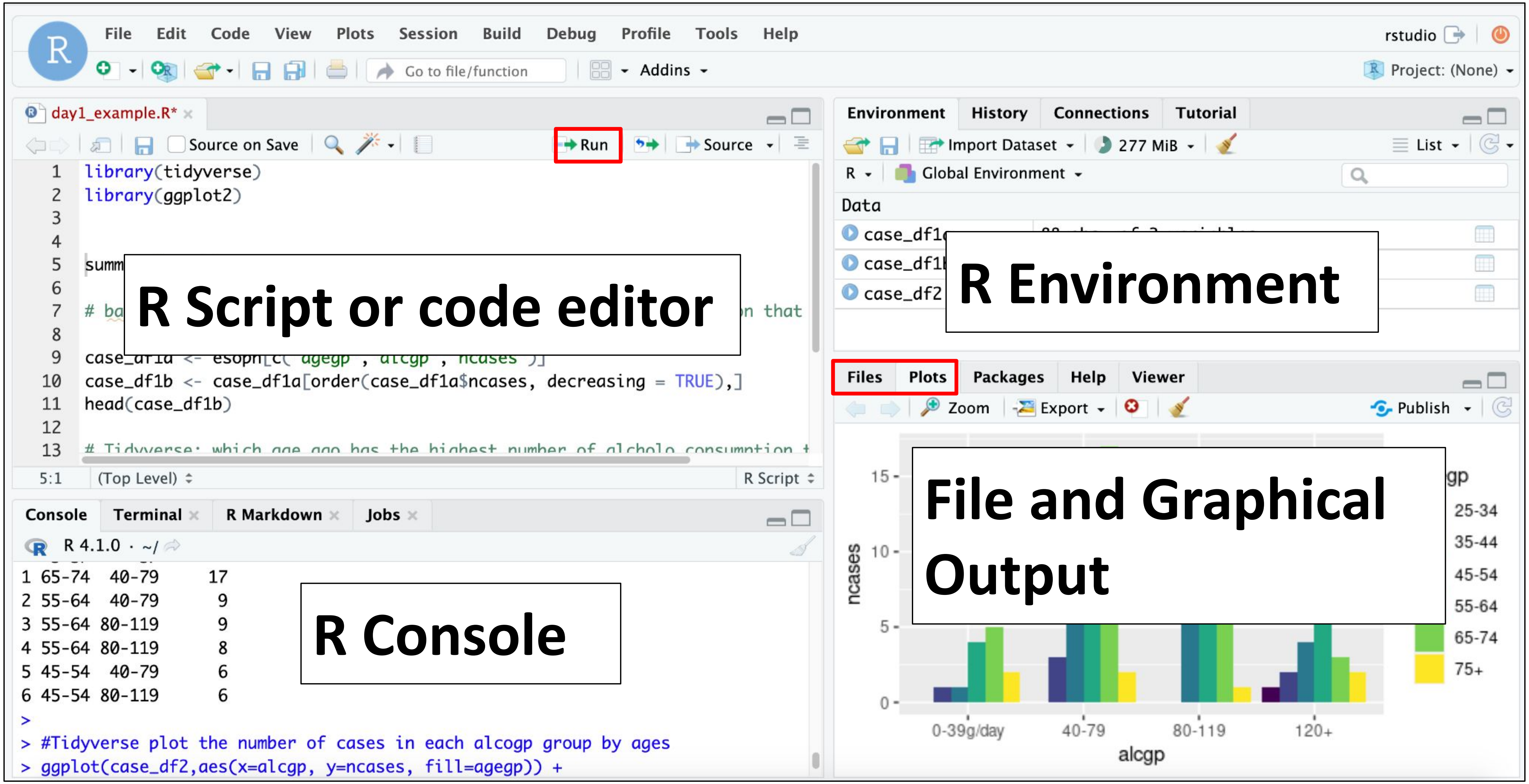
```
data("DNase")

head(DNase)
```

```
> head(DNase)
  Run       conc density
1   1 0.04882812   0.017
2   1 0.04882812   0.018
3   1 0.19531250   0.121
4   1 0.19531250   0.124
5   1 0.39062500   0.206
6   1 0.39062500   0.215
>
```

# Rstudio Interface and Demo

camh | Krembil Centre for Neuroinformatics



```r
library(tidyverse)
library(ggplot2)


summ

# ba                                                    n that

case_df1a <- esoph[c("agegp", "alcgp", "ncases")]
case_df1b <- case_df1a[order(case_df1a$ncases, decreasing = TRUE),]
head(case_df1b)

# Tidyverse: which age gap has the highest number of alcholo consumption t
```

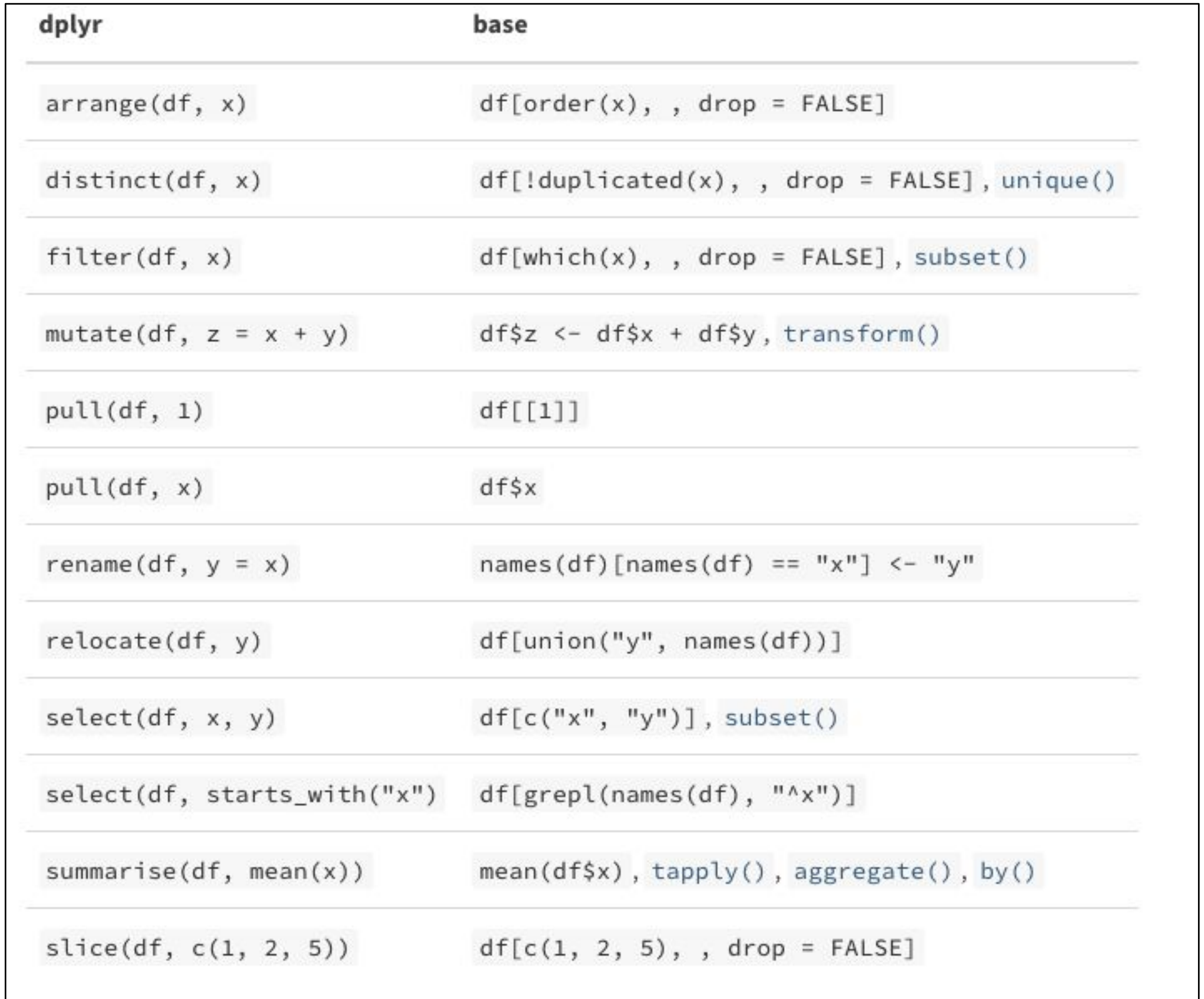
```
1 65-74  40-79      17
2 55-64  40-79       9
3 55-64  80-119      9
4 55-64  80-119      8
5 45-54  40-79       6
6 45-54  80-119      6
>
> #Tidyverse plot the number of cases in each alcogp group by ages
> ggplot(case_df2,aes(x=alcgp, y=ncases, fill=agegp)) +
```

# Tidyverse (dplyr package) vs Base R

camh | Krembil Centre for Neuroinformatics

- Here are few examples between base R function and the equivalent functions in the dplyr package found within tidyverse

- Note: through the different workshops you will encounter the use of both tidyverse and base R coding style



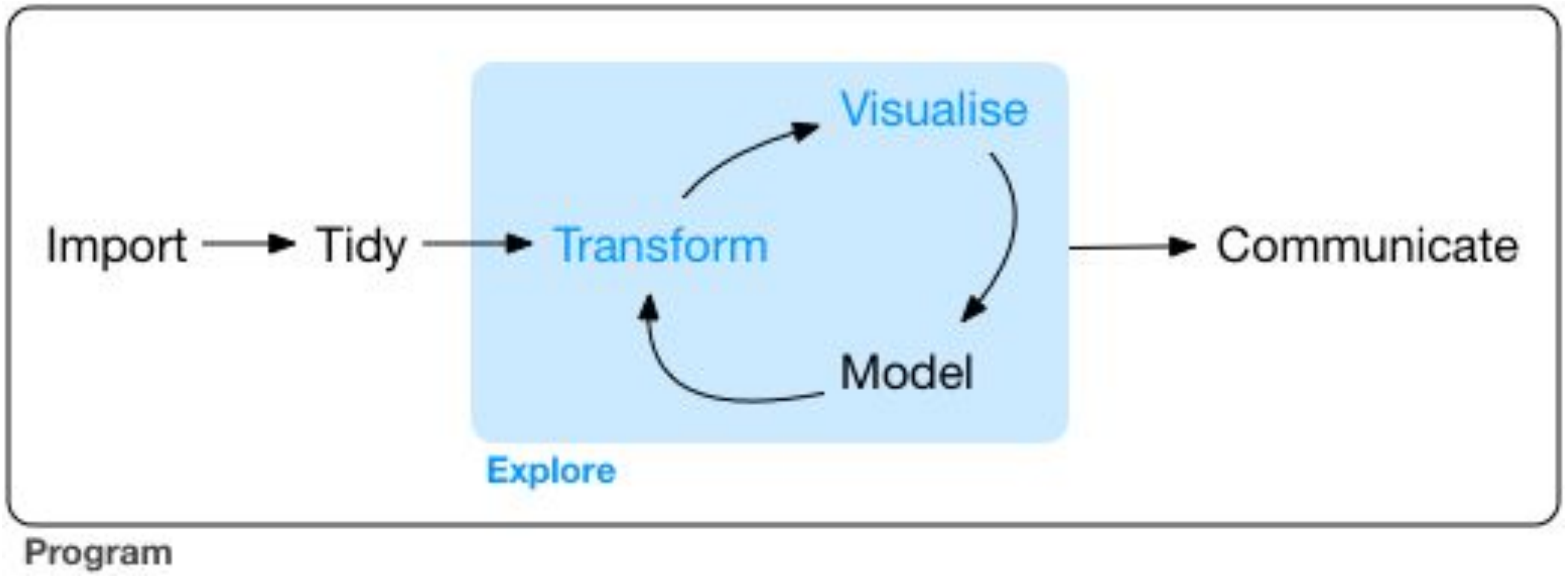
| dplyr | base |
|---|---|
| `arrange(df, x)` | `df[order(x), , drop = FALSE]` |
| `distinct(df, x)` | `df[!duplicated(x), , drop = FALSE]`, `unique()` |
| `filter(df, x)` | `df[which(x), , drop = FALSE]`, `subset()` |
| `mutate(df, z = x + y)` | `df$z <- df$x + df$y`, `transform()` |
| `pull(df, 1)` | `df[[1]]` |
| `pull(df, x)` | `df$x` |
| `rename(df, y = x)` | `names(df)[names(df) == "x"] <- "y"` |
| `relocate(df, y)` | `df[union("y", names(df))]` |
| `select(df, x, y)` | `df[c("x", "y")]`, `subset()` |
| `select(df, starts_with("x")` | `df[grepl(names(df), "^x")]` |
| `summarise(df, mean(x))` | `mean(df$x)`, `tapply()`, `aggregate()`, `by()` |
| `slice(df, c(1, 2, 5))` | `df[c(1, 2, 5), , drop = FALSE]` |

## Base R vs Tidyverse

Krembil Centre for Neuroinformatics

**Long debate in the R community of which way of R coding is better**

- Generally Tidyverse is easier to learn then Base R
- Tidyverse follow some logical flow when coding which is easier to understand



## R script vs R Notebook -

Krembil Centre for Neuroinformatics

### Two main ways to write and execute code

- R script file can be used to write your code and the file extension is .R
  - code and output are on two different panels

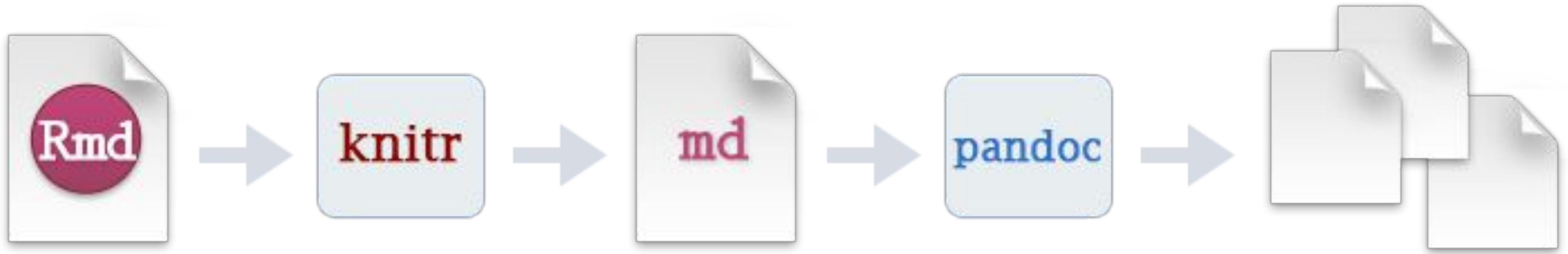
- However using R Notebooks is becoming used more often
  - The output of the code is below each code chunk
  - Documenting the code and reporting can be done beside the code using text elements to get a  fully formatted
  - Can use other programing language such as Bash or Python
  - Easy to share with collaborator
  - Work will with version control system
  - R Notebook is way to work with R Markdown files

# R Markdown

camh | Krembil Centre for Neuroinformatics

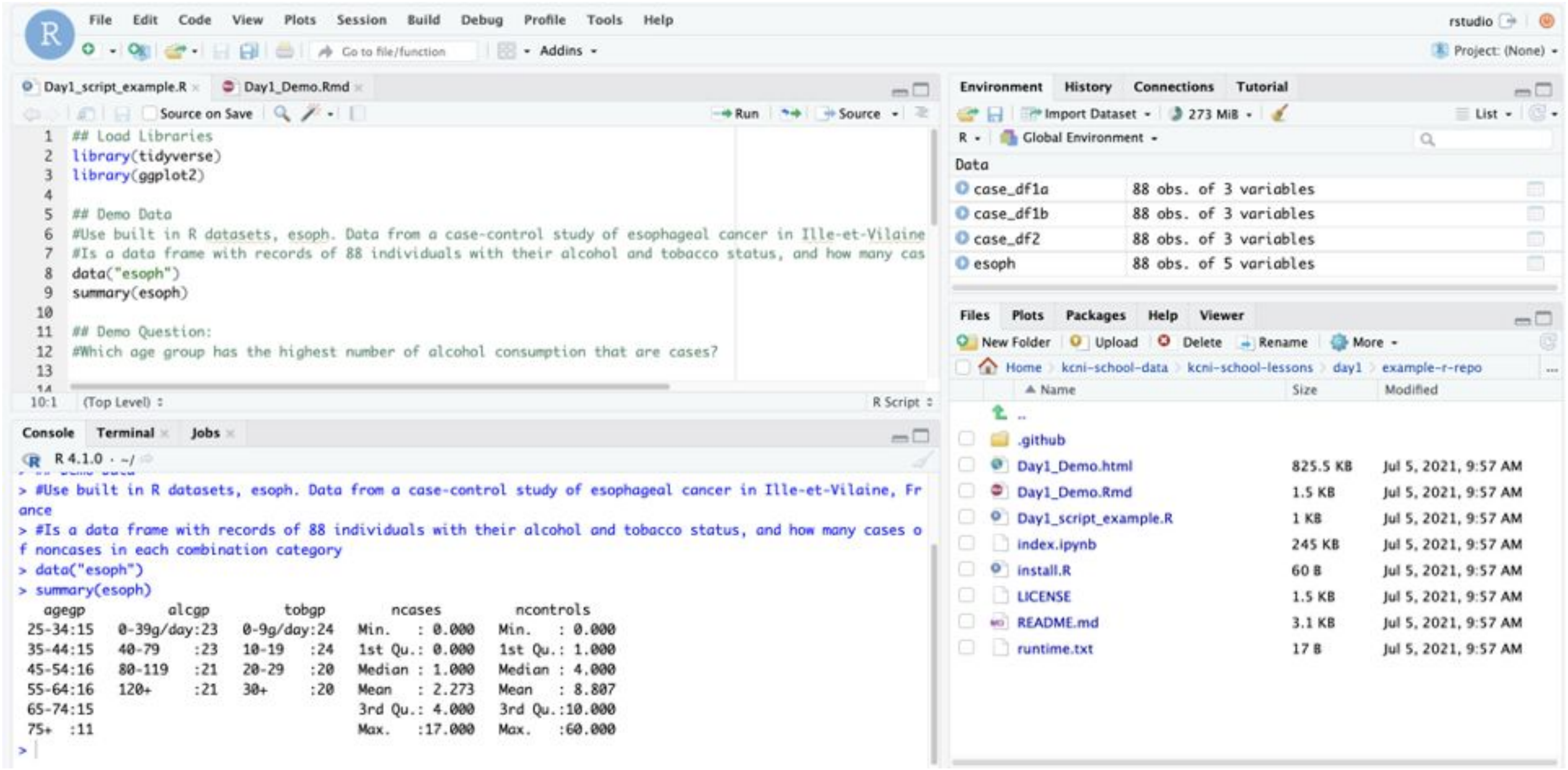
## R Markdown has a file extension .Rmd

- Provides an authoring framework for data science
- Fully formatted document into PDF, HTML or Word
- Combination of:
  - Written in plain text
  - Special characters for text formatting
  - R code within it to produce outputs such as table and plots
- Generate high quality reports that can be shared with an audience



# R Script Demo

camh | Krembil Centre for Neuroinformatics



```r
## Load Libraries
library(tidyverse)
library(ggplot2)

## Demo Data
#Use built in R datasets, esoph. Data from a case-control study of esophageal cancer in Ille-et-Vilaine
#Is a data frame with records of 88 individuals with their alcohol and tobacco status, and how many cas
data("esoph")
summary(esoph)

## Demo Question:
#Which age group has the highest number of alcohol consumption that are cases?
```

```
> #Use built in R datasets, esoph. Data from a case-control study of esophageal cancer in Ille-et-Vilaine, Fr
ance
> #Is a data frame with records of 88 individuals with their alcohol and tobacco status, and how many cases o
f noncases in each combination category
> data("esoph")
> summary(esoph)
   agegp          alcgp         tobgp        ncases          ncontrols
 25-34:15   0-39g/day:23   0-9g/day:24   Min.   : 0.000   Min.   : 0.000
 35-44:15   40-79    :23   10-19   :24   1st Qu.: 0.000   1st Qu.: 1.000
 45-54:16   80-119   :21   20-29   :20   Median : 1.000   Median : 4.000
 55-64:16   120+     :21   30+     :20   Mean   : 2.273   Mean   : 8.807
 65-74:15                                3rd Qu.: 4.000   3rd Qu.:10.000
 75+  :11                                Max.   :17.000   Max.   :60.000
>
```

Data

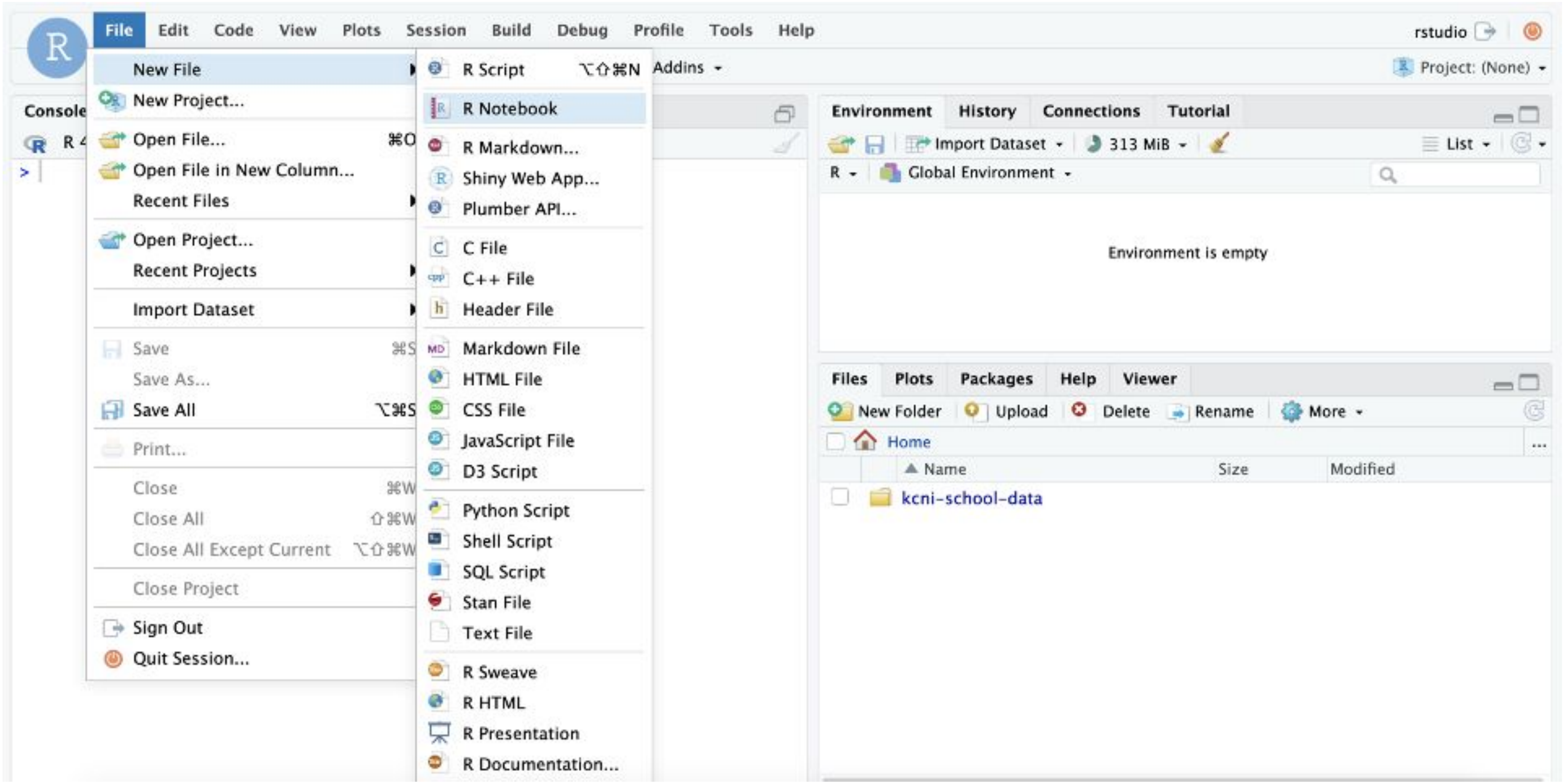
| | |
|---|---|
| case_df1a | 88 obs. of 3 variables |
| case_df1b | 88 obs. of 3 variables |
| case_df2 | 88 obs. of 3 variables |
| esoph | 88 obs. of 5 variables |

Home › kcni-school-data › kcni-school-lessons › day1 › example-r-repo

| Name | Size | Modified |
|---|---|---|
| .. | | |
| .github | | |
| Day1_Demo.html | 825.5 KB | Jul 5, 2021, 9:57 AM |
| Day1_Demo.Rmd | 1.5 KB | Jul 5, 2021, 9:57 AM |
| Day1_script_example.R | 1 KB | Jul 5, 2021, 9:57 AM |
| index.ipynb | 245 KB | Jul 5, 2021, 9:57 AM |
| install.R | 60 B | Jul 5, 2021, 9:57 AM |
| LICENSE | 1.5 KB | Jul 5, 2021, 9:57 AM |
| README.md | 3.1 KB | Jul 5, 2021, 9:57 AM |
| runtime.txt | 17 B | Jul 5, 2021, 9:57 AM |

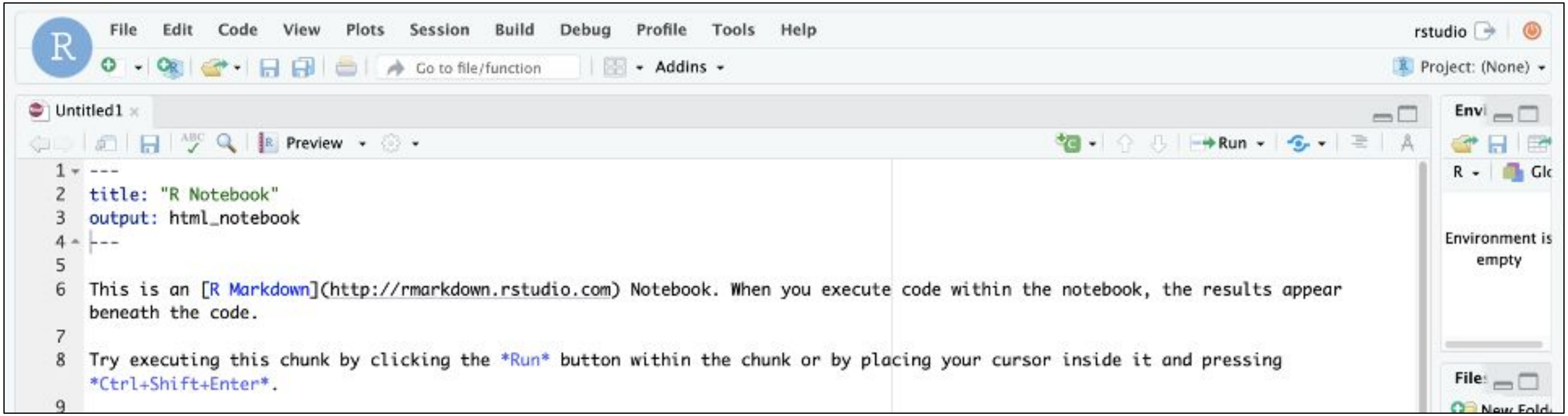
# R Notebook Example + Demo

camh | Krembil Centre for Neuroinformatics



# R Notebook Example + Demo

camh | Krembil Centre for Neuroinformatics



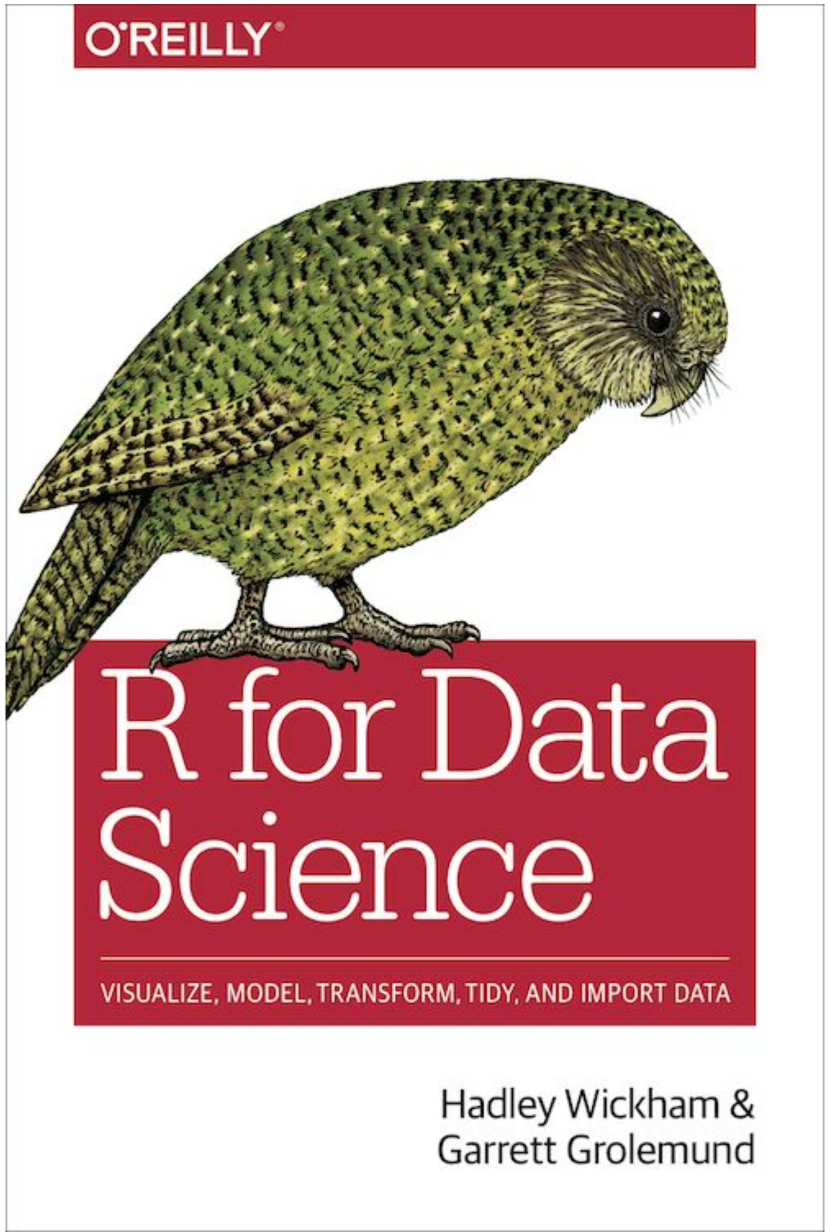
```
---
title: "R Notebook"
output: html_notebook
---

This is an [R Markdown](http://rmarkdown.rstudio.com) Notebook. When you execute code within the notebook, the results appear beneath the code.

Try executing this chunk by clicking the *Run* button within the chunk or by placing your cursor inside it and pressing *Ctrl+Shift+Enter*.
```

# R for Data Science

Krembil Centre for Neuroinformatics


O'REILLY

R for Data Science

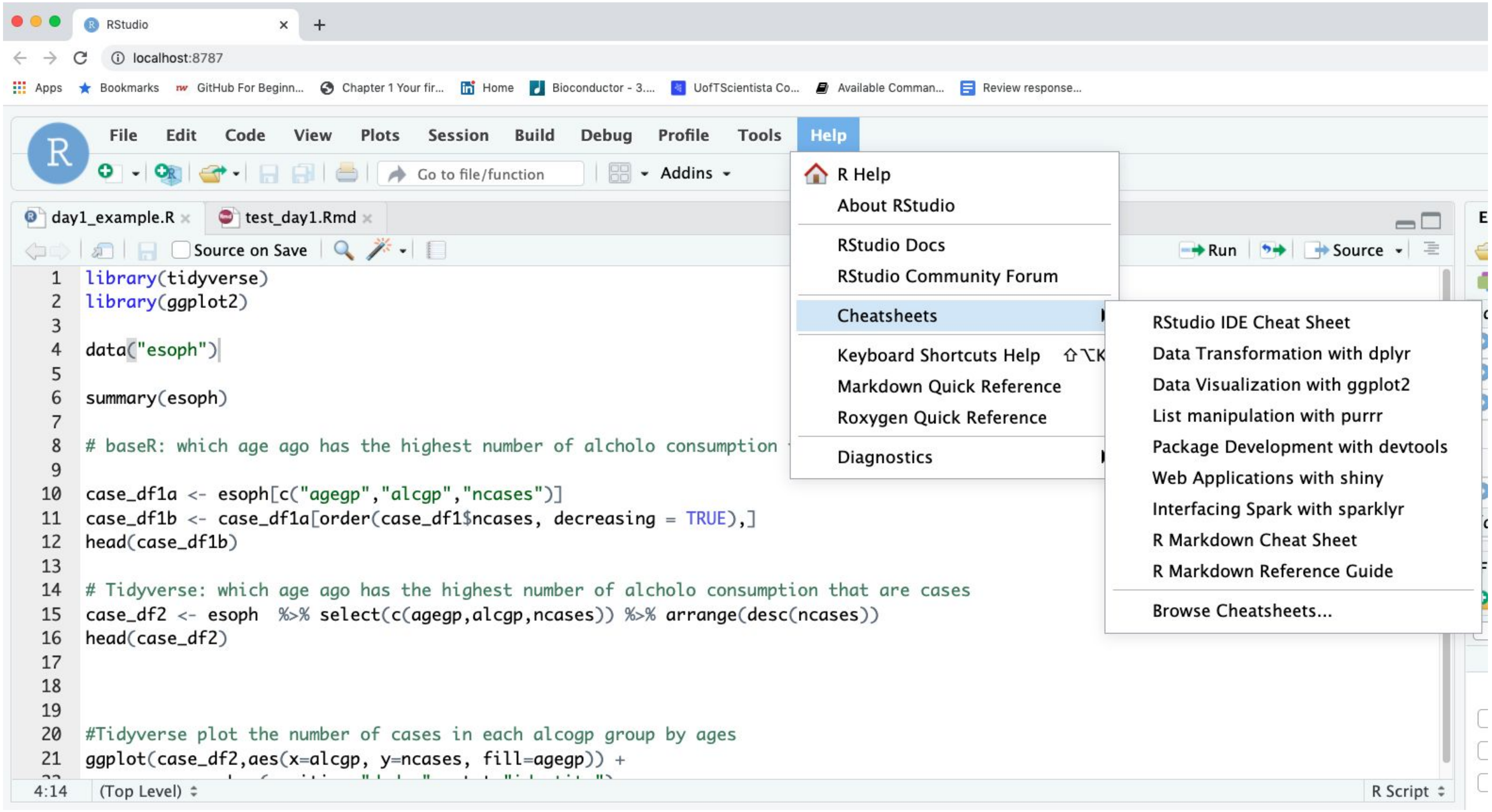
VISUALIZE, MODEL, TRANSFORM, TIDY, AND IMPORT DATA

Hadley Wickham & Garrett Grolemund

**https://r4ds.had.co.nz/index.html**

# R Cheatsheets!

camh | Krembil Centre for Neuroinformatics



```r
library(tidyverse)
library(ggplot2)

data("esoph")

summary(esoph)

# baseR: which age ago has the highest number of alcholo consumption

case_df1a <- esoph[c("agegp","alcgp","ncases")]
case_df1b <- case_df1a[order(case_df1$ncases, decreasing = TRUE),]
head(case_df1b)

# Tidyverse: which age ago has the highest number of alcholo consumption that are cases
case_df2 <- esoph  %>% select(c(agegp,alcgp,ncases)) %>% arrange(desc(ncases))
head(case_df2)



#Tidyverse plot the number of cases in each alcogp group by ages
ggplot(case_df2,aes(x=alcgp, y=ncases, fill=agegp)) +
```

# Outline for this afternoon

Krembil Centre for Neuroinformatics

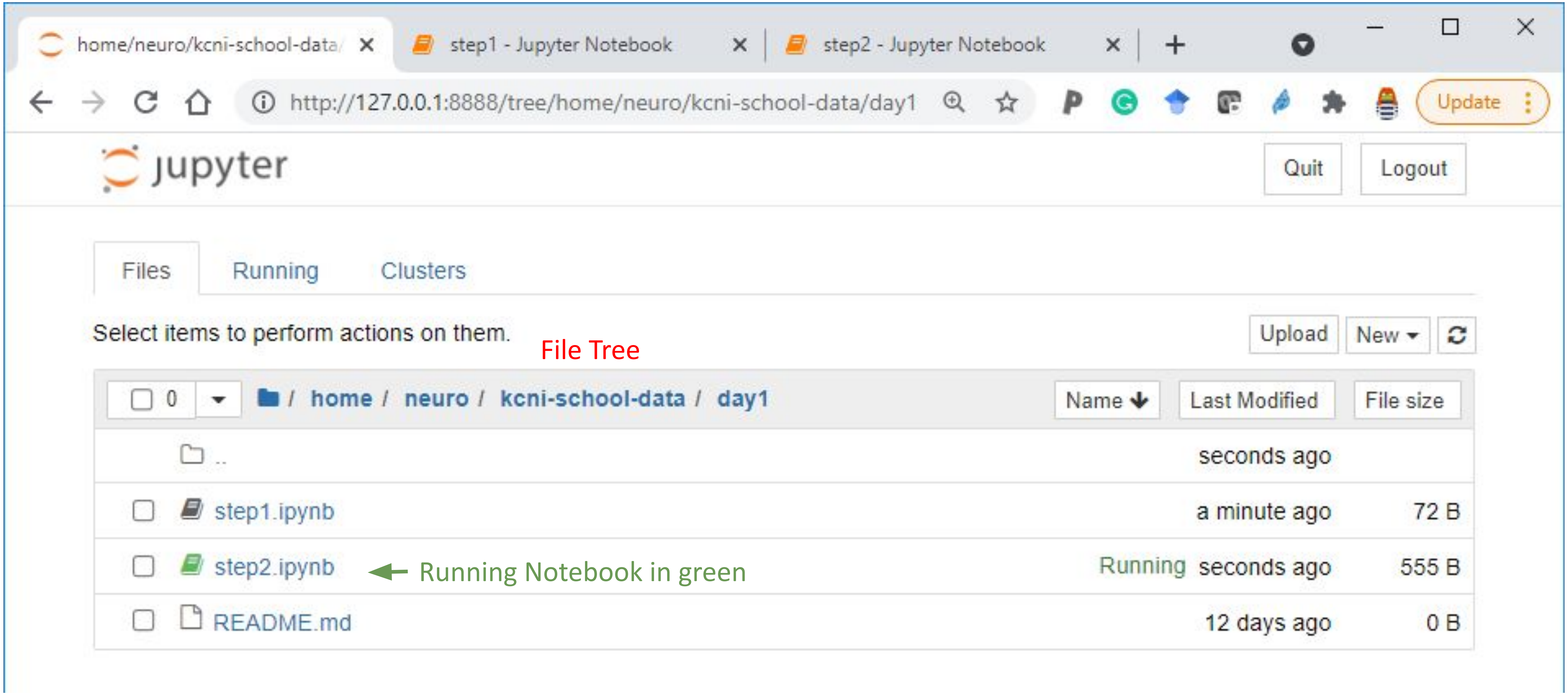
- Why reproducible science?
- The things you need to know to get through this course:
    - Versioning and publishing code (github)
    - Versioning and publishing software (docker)
    - R with Rmarkdown (walk through)
    - **Python in ipython notebooks (walk through)**
    - -also google colab
- The fancy bits:
    - building your own binder environments
    - building your own containers (docker & singularity)

## Jupyter notebooks - let’s open the one in our example-python-repo

```
docker compose up jupyter
```

# Jupyter file browser interface

camh Krembil Centre for Neuroinformatics



Select items to perform actions on them.

File Tree

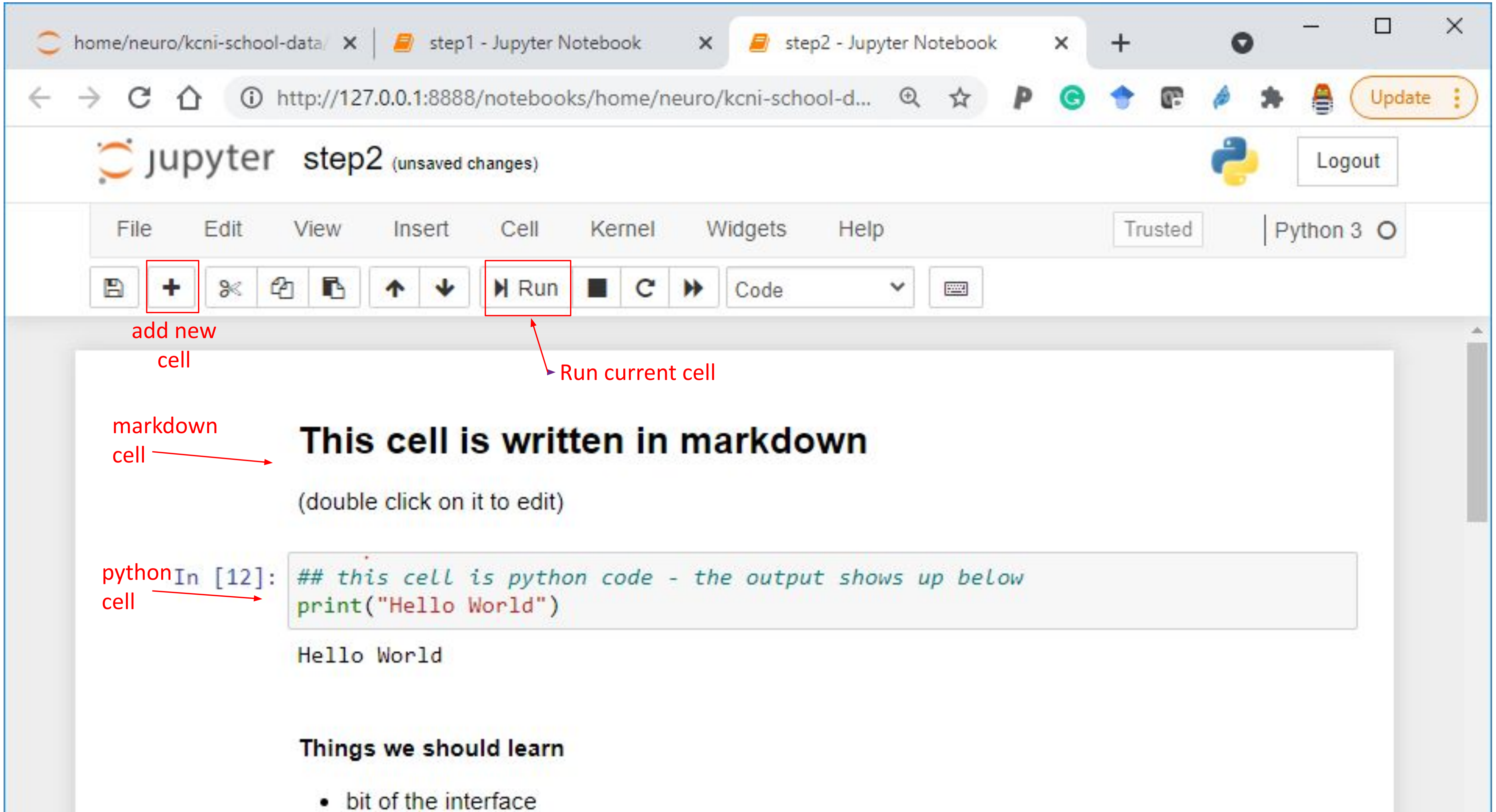
/ home / neuro / kcni-school-data / day1

| Name | Last Modified | File size |
|---|---|---|
| .. | seconds ago | |
| step1.ipynb | a minute ago | 72 B |
| step2.ipynb — Running | seconds ago | 555 B |
| README.md | 12 days ago | 0 B |

Running Notebook in green

# Jupyter file browser interface

camh | Krembil Centre for Neuroinformatics



This cell is written in markdown

(double click on it to edit)

```
In [12]: ## this cell is python code - the output shows up below
         print("Hello World")

Hello World
```

**Things we should learn**

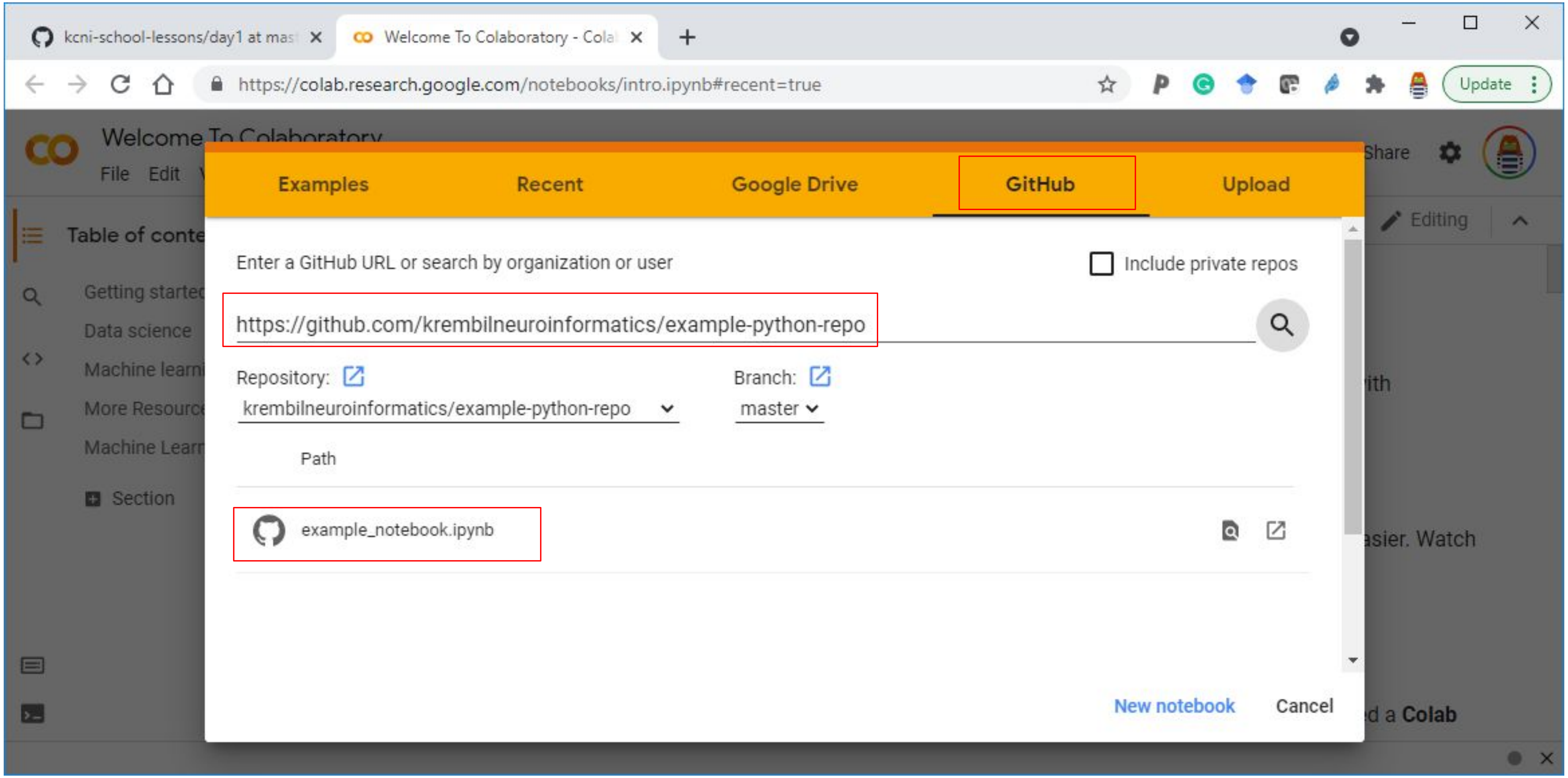
- bit of the interface

## Now let’s check out the same notebook on google colaboratory (colab)

Instructions at:
https://github.com/krembilneuroinformatics/kcni-school-lessons/tree/master/day1

# Recap: opening the notebook

camh | Krembil Centre for Neuroinformatics



Enter a GitHub URL or search by organization or user

Include private repos

https://github.com/krembilneuroinformatics/example-python-repo

Repository: krembilneuroinformatics/example-python-repo

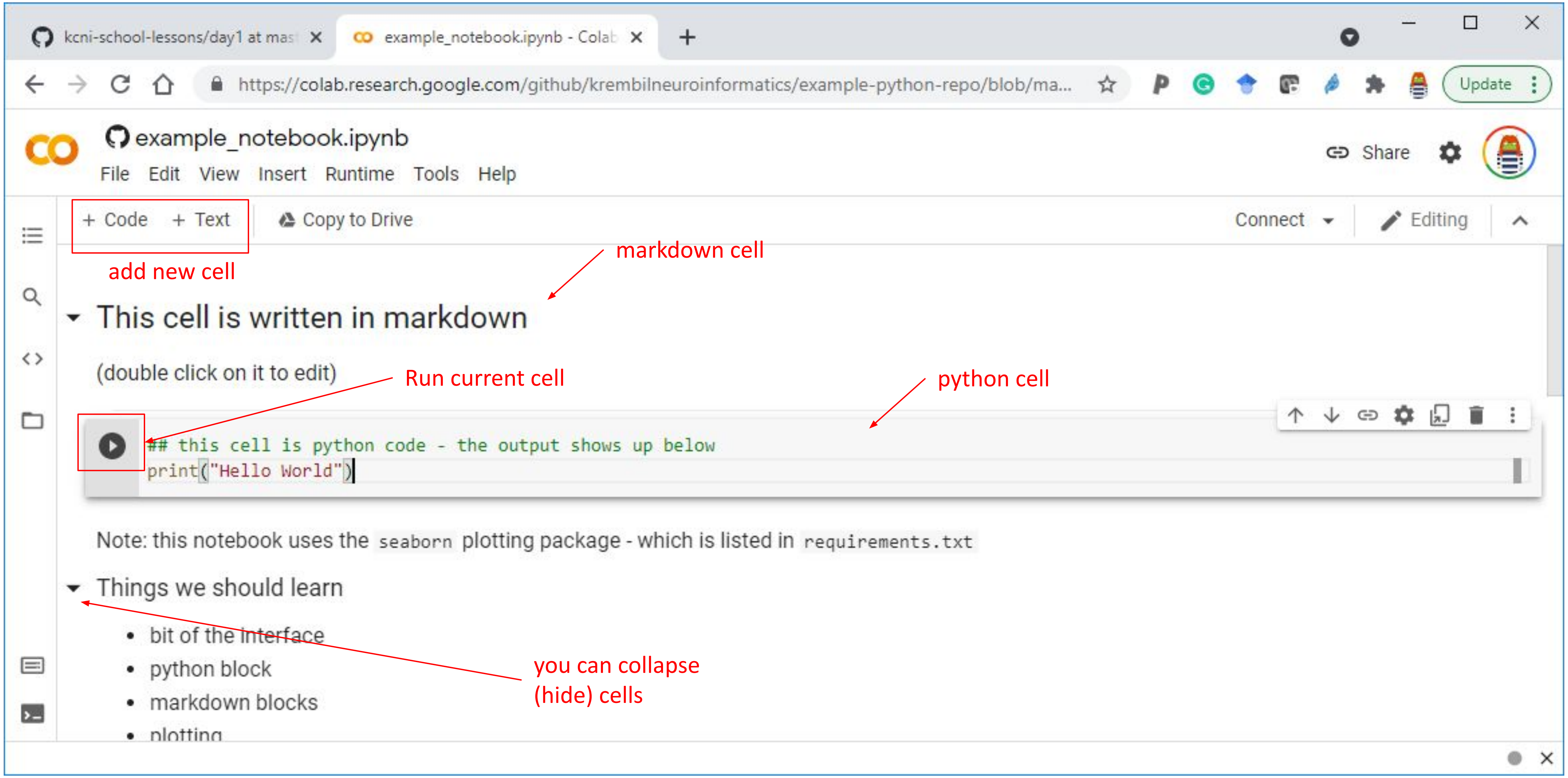
Branch: master

Path

example_notebook.ipynb

# The colab interface (README.md)

camh | Krembil Centre for Neuroinformatics


kcni-school-lessons/day1 at mas
example_notebook.ipynb - Colab
https://colab.research.google.com/github/krembilneuroinformatics/example-python-repo/blob/ma...
Update
example_notebook.ipynb
File
Edit
View
Insert
Runtime
Tools
Help
Share
+ Code
+ Text
Copy to Drive
Connect
Editing
add new cell
markdown cell
Run current cell
python cell
you can collapse (hide) cells

This cell is written in markdown

(double click on it to edit)

```
## this cell is python code - the output shows up below
print("Hello World")
```

Note: this notebook uses the `seaborn` plotting package - which is listed in `requirements.txt`

Things we should learn

- bit of the interface
- python block
- markdown blocks
- plotting

## colab strengths and caveates

Krembil Centre for Neuroinformatics

**strengths**

- more power, more diskspace and more RAM than binder
- easy to use
- can be left running for days (no timeouts, unlike binder)
- can connect to google drive for more cloud storage space.

**caveats**

**to remember for this course**

- each colab notebook is and island
  - it is not aware of the other data scripts or notebooks line same folder
- You always start with the same (cleanish) linux/python-3.7.1 environment
  - so you need to install all other software inside your notebook
  - the current env is old so some newer packages may not work

## Outline for this afternoon

Why reproducible science?

The things you need to know to get through this course:

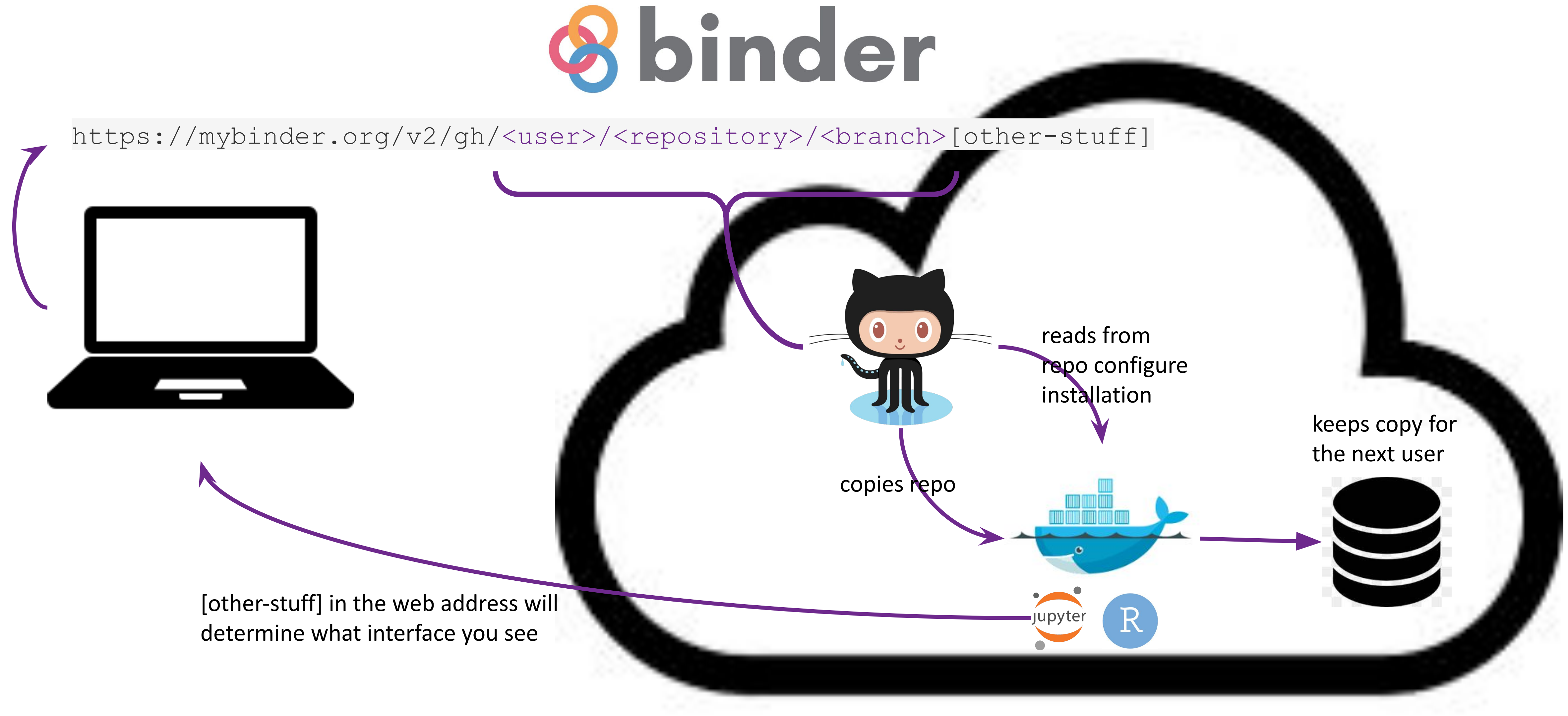
- Versioning and publishing code (github)
- Versioning and publishing software (docker)
- R with Rmarkdown (walk through)
- Python in ipython notebooks (walk through)
- -also google colab

**The fancy bits:**

- building your own binder environments
- building your own containers (docker & singularity)

# Binder - repros runable on the web

camh | Krembil Centre for Neuroinformatics


binder
https://mybinder.org/v2/gh/<user>/<repository>/<branch>[other-stuff]
reads from repo configure installation
copies repo
keeps copy for the next user
jupyter
R
[other-stuff] in the web address will determine what interface you see

## Binder how - the files

Krembil Centre for Neuroinformatics

For python (jupyter)

- `requirements.txt` : tells binder what python packages to install
- `runtime.txt` : tells binder what version of python to install

see example in:

day1/example-python-repo

For and R (with rstudio)

- `runtime.txt`: tells binder what version of R to use
- `install.R`: tells binder what R packages to install

see example in:

day1/example-r-repo

# Binder - how - the URL’s

Krembil Centre for Neuroinformatics

The standard format is:

`http://mybinder.org/v2/gh/<github-user>/<github-repo>/<branch>`

For this repo this is:

- http://mybinder.org/v2/gh/krembilneuroinformatics/example-python-repo/HEAD

But if directly to one particular python notebook. You could add the filepath to the end `?filepath=<filepath>`

- `Example:` https://mybinder.org/v2/gh/krembilneuroinformatics/example-python-repo.git/HEAD?filepath=example_notebook.ipynb

For the RStudio environment, we must add the following at the end of the URL: `?urlpath=rstudio`

- Example: http://mybinder.org/v2/gh/krembilneuroinformatics/example-r-repo/HEAD?urlpath=rstudio

Help at mybinder.org

## let’s look at two example repo’s in the kcni-school-lessons/day1

## Outline for this afternoon

Why reproducible science?

The things you need to know to get through this course:

- Versioning and publishing code (github)
- Versioning and publishing software (docker)
- R with Rmarkdown (walk through)
- Python in ipython notebooks (walk through)
- -also google colab

**The fancy bits:**

- building your own binder environments
- **building your own containers (docker & singularity)**

# Step 1: write a Dockerfile

camh | Krembil Centre for Neuroinformatics

In an empty folder (new github repo)

Create a file named:

**`Dockerfile`**

*Everything else in this folder will be copied inside the docker image.*

The `Dockerfile` contains the instructions for software installation. Commands are:

- **FROM** (at top line): points to another Docker image to start from
- **RUN**: will run an installation command)
  - -run is followed by shell install commands
  - can add && to package lines into the same "layer"
- **ENTRYPOINT**: will determine one command that is run "by default"

```
FROM <base-image>

# stuff to install
RUN <installation command>

# more stuff (layer 2)
RUN <installation line1>
   && <installation line2>
```

Browse https://hub.docker.com/ & kcni-school-lessons/envs for examples

## Step 1: write a Dockerfile

In an empty folder (new github repo)

Create a file named:

**`Dockerfile`**

*Everything else in this folder will be copied inside the docker image.*

The `Dockerfile` contains the instructions for software installation. Commands are:

- **FROM** (at top line): points to another Docker image to start from
- **RUN**: will run an installation command)
  - -run is followed by shell install commands
- **ENV**: can set environment variables
- **ENTRYPOINT**: will determine one command that is run "by default"

```
FROM rocker/verse:4.1.0

# adding plotting packages to from day 1 demo
RUN install2.r --error \
    --deps TRUE \
    ggrepel \
    ggthemes \
    here

## adding data grabbing packages
RUN apt-get update -qq \
     && apt-get -y --no-install-recommends \
     install wget curl git
```

Browse https://hub.docker.com/ & kcni-school-lessons/envs for examples

# Building your docker image

camh | Krembil Centre for Neuroinformatics

Use “docker build” to build and test your image on your local computer.
*Note “-t” is for tag - or the name you will give to your docker the last argument is the folder*

```
cd my_docker_folder/
docker build -t my_new_docker ./
```

You can connect you github repo to dockerhub and have dockerhub re-build your container everytime you make a push to your repo

docker hub

https://hub.docker.com/

## Running docker images

Krembil Centre for Neuroinformatics

**To run your docker:**

```
docker run [options] <dockerhub_user>/<image>:<version>
```

example: `docker run [options] edickie/rstudio-school:latest`

**Important options when you run and image:**

- **`--publish,-p`** : allows port forwarding from the inside of the docker to the outside
  - we need this to connect rstudio or the inside of the docker to your computer's browser
- **`--volume, -v`**: use this to connect data outside the docker readable/writable by the docker software.
  - Docker can only "see" data that is connected to it (not everything on your computer)

# Running the kcnischool-rstudio image

camh
Krembil Centre for Neuroinformatics

**The full command to run the KCNI rstudio image is:**

```
docker run --rm -it \
 -e DISABLE_AUTH=true \
 -p 127.0.0.1:8787:8787 \
 -v <path/to/your/data>:/home/rstudio/kcni-school-lessons \
 edickie/kcnischool-rstudio:latest
```

***where*** - `<path/to/my/data>` *is the name of the folder on your computer where you cloned the kcni-school-lessons repo.*

# docker-compose files save memory

camh | Krembil Centre for Neuroinformatics

We put a “docker-compose.yml” file at the base of the kcni-school-lessons repo. This file contains defaults of the docker options - so to run the docker you type

```
docker compose up rstudio
```

*Later in the week we will start jupyter with:*

```
docker compose up jupyter
```

```
version: '3'

services:
  rstudio:
    image: edickie/kcnischool-rstudio
    ports:
      - 8787:8787
    volumes:
      - ./:/home/rstudio/kcni-school-data
    environment:
      - DISABLE_AUTH=true

  jupyter:
    image: edickie/kcnischool-jupyter
    ports:
      - 8888:8888
    volumes:
      - ./:/home/neuro/kcni-school-data
```

# Docker vs Singularity

| | Pro | Cons |
|---|---|---|
| Docker | • Strong and building dev community<br>• can be pushed and pulled to dockerhub<br>• Works on any system (Windows, Mac, Linux)<br>• "Layers" decrease the hard disk space | **Needs "root" access** |
| Singularity | **Does not need "root" access.**<br><br>Can converted from Docker | • Smaller dev community<br>• Singularity hub exists, but is less used<br>• Only works on Linux |

## Docker vs Singularity

take home...

- we write a Docker spec
- we run singularity containers on High performance computers and shared servers (like the CAMH SCC or SciNet)

## Translating Docker Usage to Singularity

| | Docker | Singularity |
|---|---|---|
| Running | Docker run -it --rm | Singularity run |
| Mounting or binding a path | -v, --volume | -B, --bind |
| Attaching the “workdir” | -w, --workdir | -W |
| Port forwarding | -p, --publish | N/A |
| Removing the outside environment | N/A | -e, --cleanenv |
| Change the mount to $HOME | N/A | -H, --home |

## Building a singularity container from docker

### Direct from dockerhub to singularity

```
ssh <username>@teach.scinet.utoronto.ca

# note - in this example I am adding the new container to my $SCRATCH folder
mkdir $SCRATCH/test_sing_img

# singularity build <output-image>
docker://<dockerhubuser>/<dockername>:<version>
singularity build singularity build \
    edickie_kcnischool-jupyter_latest-2021-07-02.sif \
    docker://edickie/kcnischool-jupyter:latest
```

## A note about “latest”

Most dockers will allow you to download the latest version of their software by typing “latest” instead of a version number

Problem - for reporting and debugging...you need the version number

So  - for a real analysis - make sure to build a specific version and put the version number in the filename of the singularity image.

# Remember - many ways to engage

camh | Krembil Centre for Neuroinformatics

? (during sessions) Use the chat or the ask question!

You can always return to the session and re-watch the vidos after the session ends

come chat with us in KCNI Summer School Slack :)

virtually meet with us in gather.town

Tell us how the session went (post session survey):
https://forms.gle/ji18qLMZEZ9L16Ln6

KCNISchool@camh.ca

# Reproducibility crisis literature

camh | Krembil Centre for Neuroinformatics

7. Ioannidis JPA Why most published research findings are false. PLoS Med. 2, e124 (2005). This landmark paper outlines the ways in which common practices can lead to inflated levels of false positives.

8. Simmons JP, Nelson LD & Simonsohn U. False-positive psychology: undisclosed flexibility in data collection and analysis allows presenting anything as significant. Psychol. Sci 22, 1359–1366 (2011). This paper highlights the impact of common ‘questionable research practices’ on study outcomes and proposes a set of guidelines to prevent false-positive findings. [PubMed: 22006061]

9. Gelman A. & Loken E. The statistical crisis in science. American Scientist 102, 40 (2014).

10. Ioannidis JPA, Fanelli D, Dunne DD & Goodman SN Meta-research: evaluation and improvement of research methods and practices. PLoS Biol. 13, e1002264 (2015).

11. Collins FS & Tabak LA Policy: NIH plans to enhance reproducibility. Nature 505, 612–613 (2014). [PubMed: 24482835]