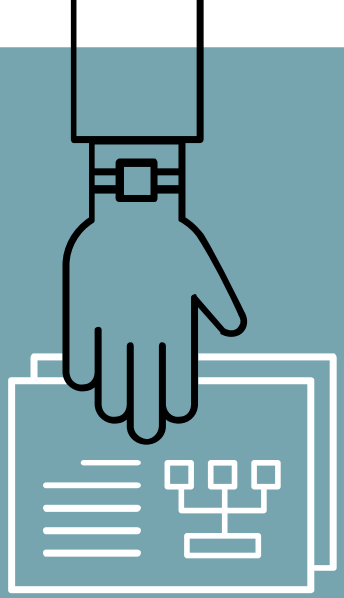
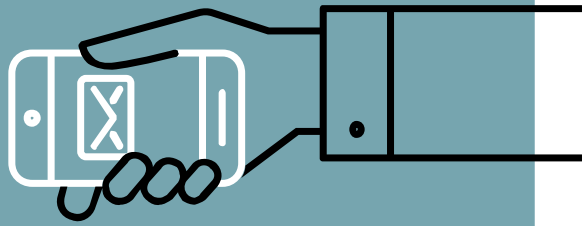
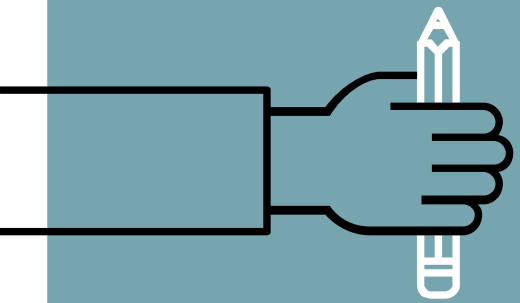
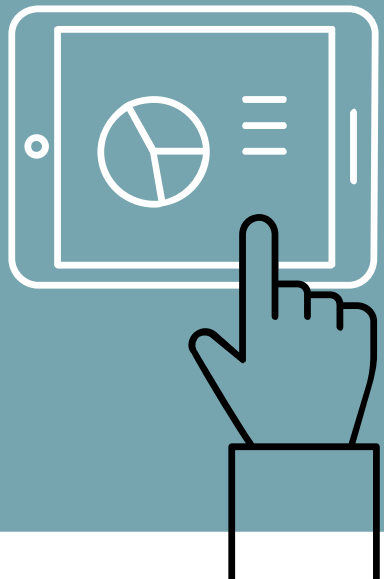




BST 270
Reproducible Data Science

Winter 2021
Session 7





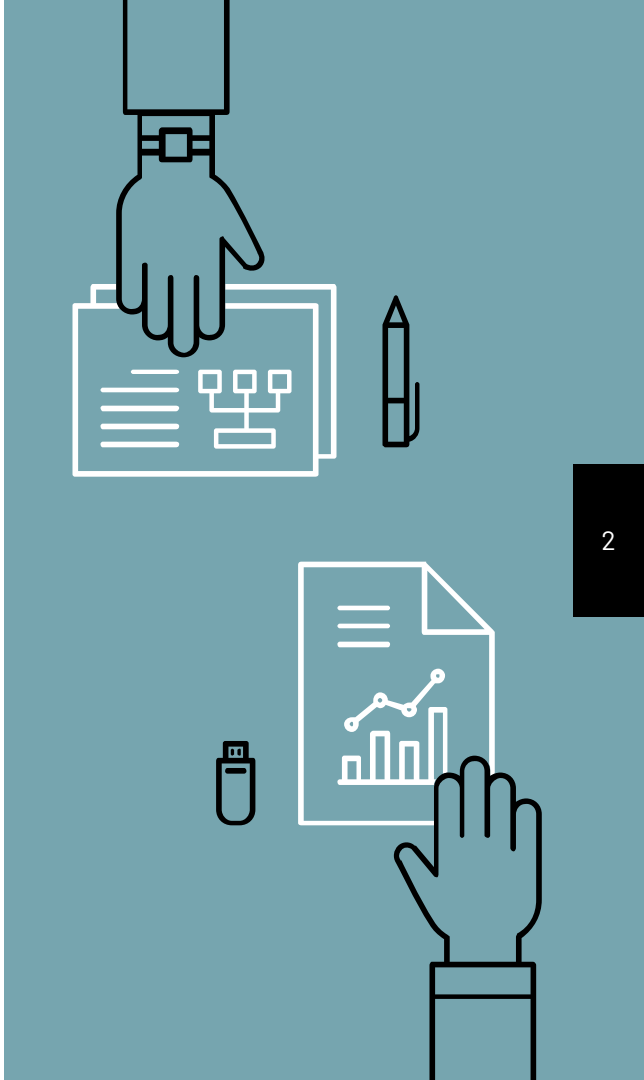
Module 6 Part 2 Discussion

I've taken a class before that used Rnw for knitting, but besides that I've only seen people working with Rmd. Are there particular reasons people don't really use Rnw as much, or why they would choose to use Rnw?

Rnw is a bit less intuitive (in my opinion). The syntax for code chunks for Rnw files is a little more complicated than Rmd files. Rnw is also used for LaTeX while Rmd can be used to create html files, word docs as well as tex files.

First, I wonder if we have a preference between Jupyter Notebook and R Markdown. Both of them look good to me and would be strong enough to handle daily tasks. Can we determine which one to use only according to our personal preferences? Besides, for example, since we already have R Markdown supporting Python, why would need to use Jupyter Notebook anymore?

In my experience researchers/data scientists tend to prefer to stick to one language and do prefer one over the other. Personally, R Markdown looks “nicer” to me and is easier to use. But, if I need to code in python I used a Jupyter Notebook. I could still use an R Markdown file with python, but for some reason I don’t like to.



Module 6 Part 2 Discussion

I'm actually really surprised I had not heard of the Harvard Dataverse. How common are Dataverses and how often are they actually used in research?

Before this class I hadn't heard of it either. I don't know how often they are used in research, but it is becoming more common if you are able to publish your data.

Some people I've worked with recommended using R Projects in Rstudio. Is that different from just creating a new directory with your code/files?

Yes. An R Project divides your work into multiple contexts, each with their own working directory, workspace, history, and source documents. You can also connect to a GitHub repository through an R Project to make version control easier. Check out more info here.

Module 6 Part 2 Discussion

Could you discuss a bit more the issue of reproducibility vs replicability? Not completely clear in the videos.

Reproducibility: given my data, code and documentation, can you reproduce my work and get the same results that I did?

Replicability: given a different dataset, my methods documentation (used for a different dataset) and your own code, can you get results similar to mine even though you used a completely different dataset?

How does using both R and Python in an Rmd file work? Can you use the Python variables you create in one chunk in a chunk with R code? How does rmd know how to convert between those variables?

You can use the reticulate package.

Module 6 Part 2 Discussion

Is there any danger with including sensitive data in data verses since you're not allowed to unsubmit data? Obviously you can restrict who gets access to it, but are there any complications?

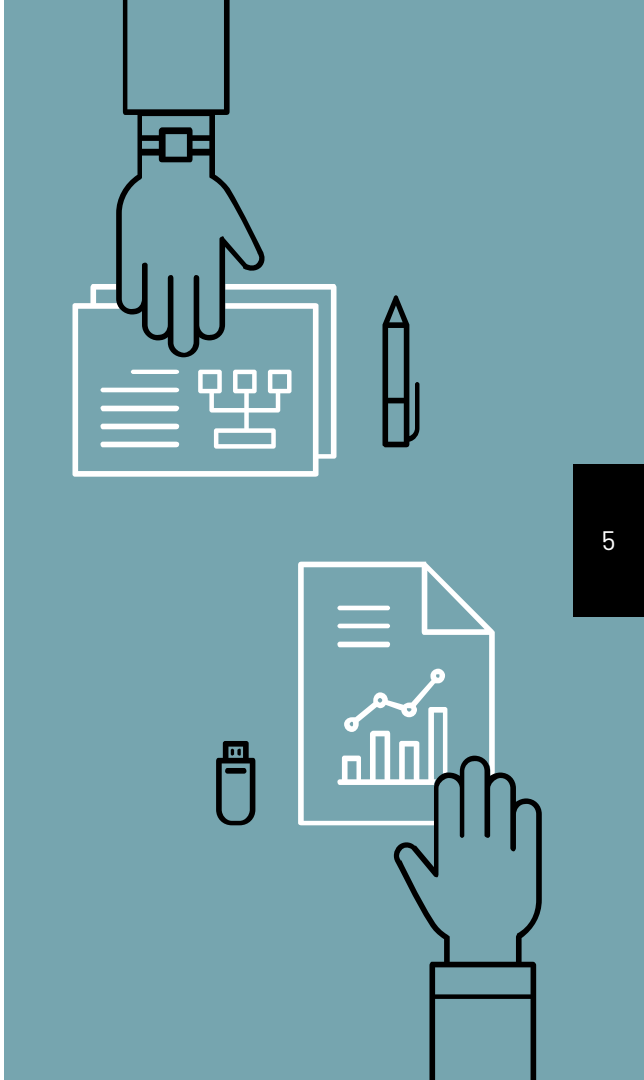
There is always a risk when sharing sensitive data. I don’t know of any specific examples but I’d imagine there would be a lot of paperwork and restrictions in place for highly sensitive data.

What's the benefit of using R Notebook versus R markdown?

Great explanation here.

As Sonia said in the video, data management projects such as Harvard Dataverse make data acquisition much easier than before. People may not have to come to Cambridge, MA to get access to the data stored in Harvard. I wonder if sharing the data internationally would cause any potential issues, such as copyright or the difference of the definition of "sensitive data" among different countries.

I believe the definition of “sensitive data” is held by the country the data originates from and all other potential users have to agree to their data use agreements. These can be extremely strict so I would hope there wouldn’t be any issues, but there is always risk involved. The dataset I worked with that had to live on a computer in JP’s office and couldn’t be connected to the Internet was from a European country and it took 18 months for JP to get access to the data.



Module 6 Part 2 Discussion

Regarding Rule 3: if the data being worked with is used for further scientific discoveries, how does not publishing the data take away credibility from the published work?

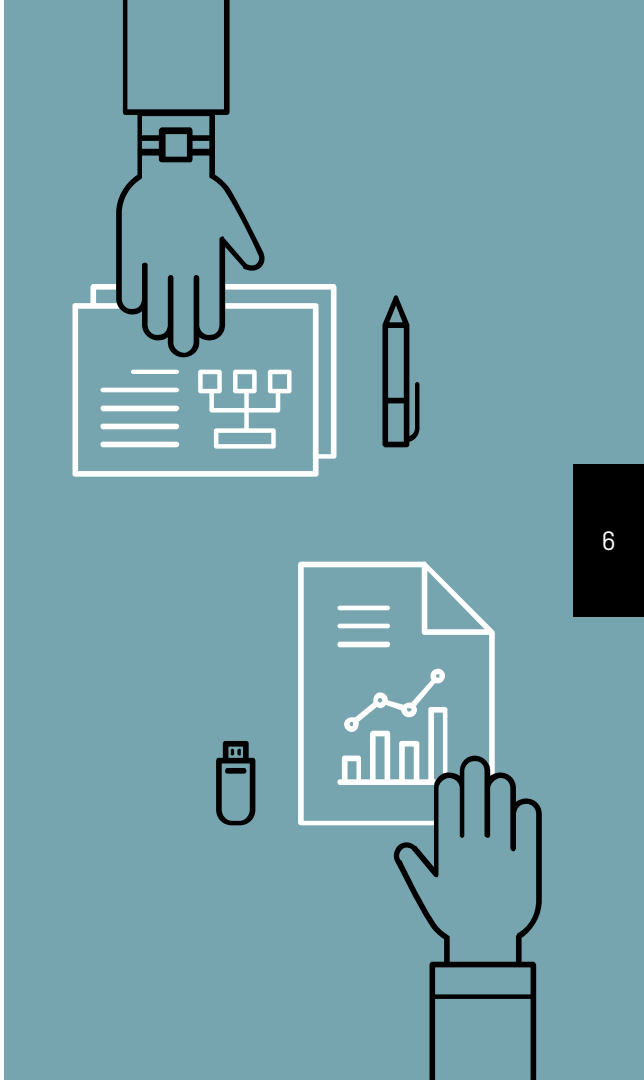
It makes it more difficult to reproduce the results and verify the scientific findings.

Does Jupyter Notebook have an equivalent to RMarkdown's cache option? Is there a way to easily run the entire notebook while caching the variables created in a single block?

Jupyter has built-in cell caching which automatically cache the contents of a code cell when it is executed, and keeps it in memory until the cell is re-run or the outcome is redefined. You can run all cells at once by choosing the ‘Run all’ button at the top of the notebook. Check out this post for more details.

Do we have API access in other programs such as SAS or Python?

I’m not sure and I couldn’t find anything very informative. You could ask Harvard IT.



Module 6 Part 2 Discussion

In one of Christopher Gandrud's modules, he recommended using plain text files to store both code and data, no matter which programming language you are using. Is this a common practice? I feel like I normally see R code stored in a .R or .Rmd file (as opposed to a .txt file), but maybe I have not been paying enough attention to file types.

In my experience, plain text or csv files are preferred because they can be used with (almost) any language/software easily. .R data is very specific to R and .dta is specific to Stata, so it depends.

For the Makefile, I am pretty confused that since it is a command that runs on a text file named Makefile, how exactly should we make such command in Windows 10 once we finished writing the context in that Makefile? Do we just run the file in the command prompt by some certain command or do we need other software to run the Makefile?

You just run the command in the command prompt - like bash for Windows and Terminal for Macs.

Module 6 Part 2 Discussion

Using a Make File seems like a very useful way to integrate different platforms into one analysis pipeline. Have you used a Makefile before? If so, how steep is the learning curve to make one? Do you recommend any other methods of integrating different code platforms within one analysis (Christine Choirat mentioned other options such as "Cmake")?

I haven't used a Makefile or Cmake before, but I have made a driver script that runs different files for one analysis. A fake example is below. There is a Cmake package in R that you can check out. And there are a few Makefile tutorials online.

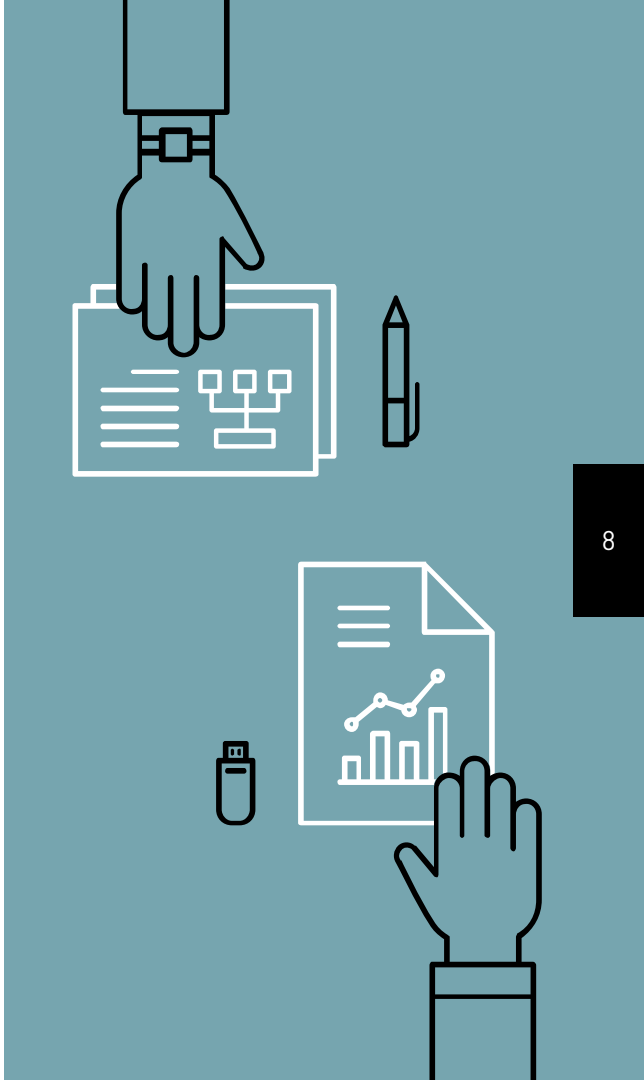
```
##############################
# Reproducible Project Driver Script  #
# Author: Jane Doe                    #
# Date: 01-01-2017                    #
##############################

# Set Directory #
cd /Users/heathermattie/Desktop/BST270/270finalproject

# Call data scraping file #
python data_scrapping.py

# Call data scrubbing file #
# Description of file input/output #
r data_scrubbing.R

# Call data analysis file #
# Description of file input/output #
r analysis.Rmd
```



Module 6 Part 2 Discussion

I am worried about a future where R loses its place in statistics and becomes an antiquated software (similar to SAS). Do you recommend we learn Python?

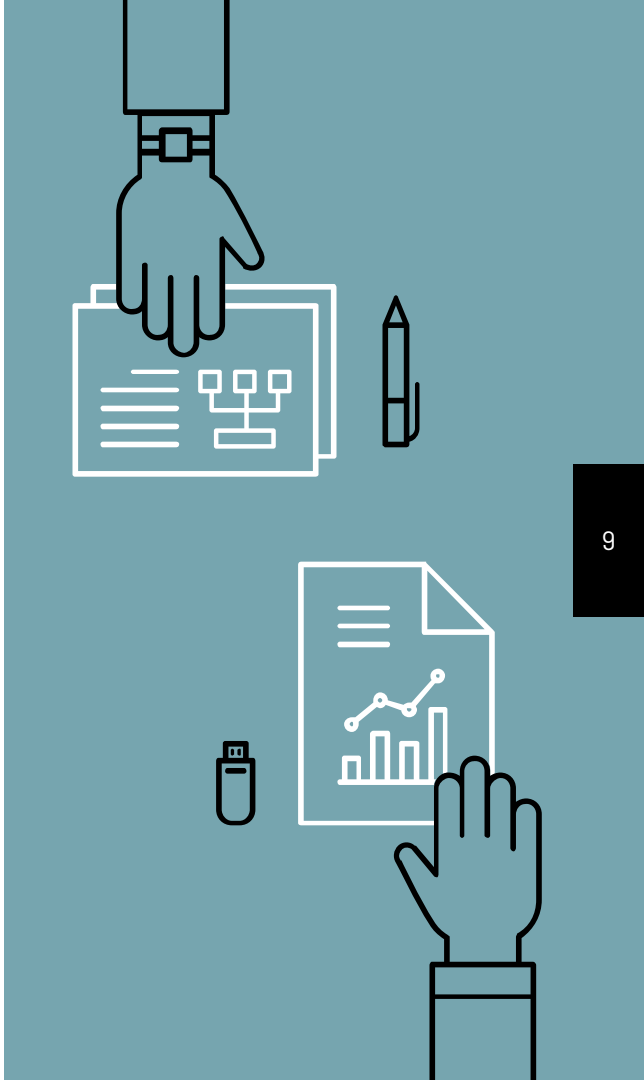
Regardless of R dying out, I recommend learning Python. It's used in industry a ton and has some advantages over R, like being able to handle bigger datasets.

Do you recommend we use Emacs?

If you have the time, yes. I personally haven't ever used Emacs and haven't ever had the need. But I think it would be a nice skill to have.

For the Harvard Dataverse, is there any requirement for people to create a dataverse or dataset or anyone can add a new dataset and dataverse? Are there any charges for people who preserved their data on the Harvard Dataverse?

I believe anyone can add a new dataset, as long as the data is legally allowed to be published. I don't think there are ever any charges for storing data on Harvard Dataverse (I couldn't find any information about pricing or payment).



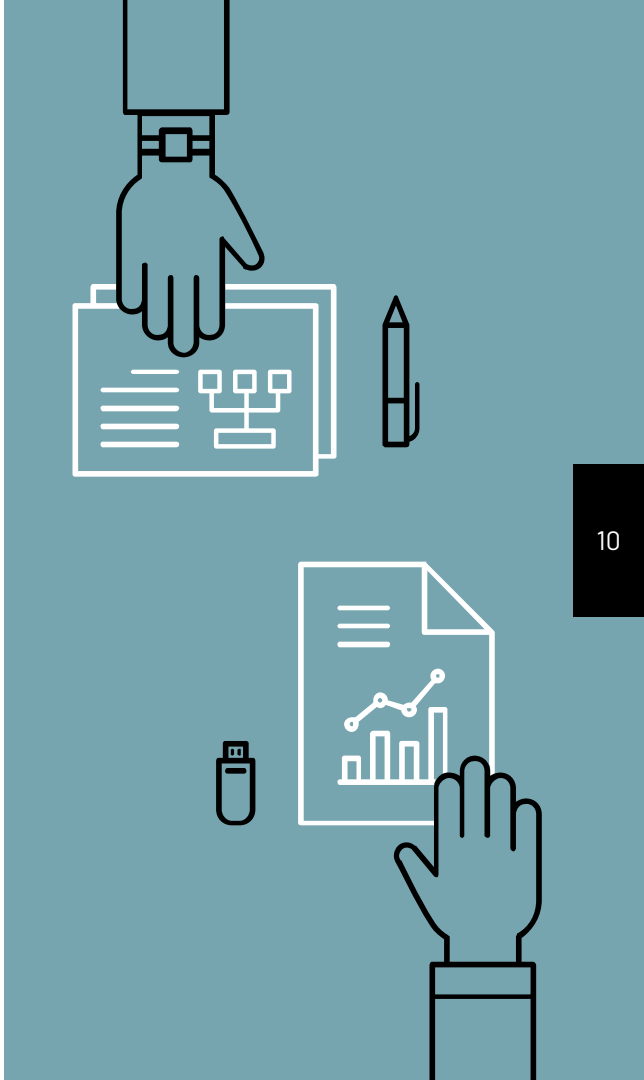
Module 6 Part 2 Discussion

What are the other most popular alternatives to the dataverse, particularly, do you know what's done in industry (is the dataverse exclusive to academic institutions)?

The NIH hosts a ton - check them out here. Science also has a list. I'm not sure about industry since a lot of their data is proprietary, but Kaggle is a great data/code sharing resource.

Is there a general preference for an rmd to be knitted as an html file or a pdf or something? Or is it largely context dependent?

It's largely context dependent. If my team/colleagues prefer a pdf of my work I will compile that. If they just want to see what I've done in a meeting I'll compile an html file to make it a bit easier / more esthetically pleasing to look at.



Module 6 Part 2 Discussion

Can you clarify what is meant by making sure your data can "map to other disciplines"?

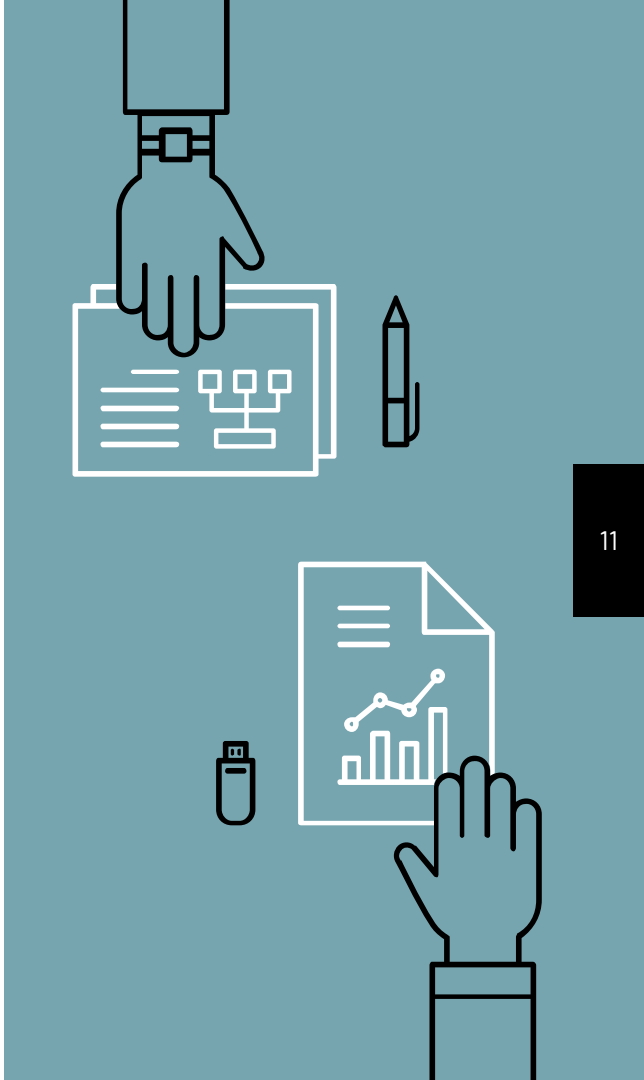
I'm not sure what this is referring to - if you remember which video mentions it let me know and I can watch it - but if I had to guess, it means making your data available in txt or csv files that anyone can use regardless of preferred software. And documenting your data thoroughly to help others understand how to use it.

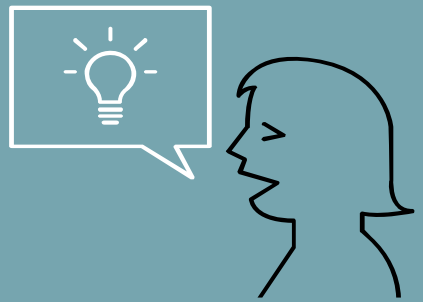
Is one example of sharing data online with a "persistent or permanent identifier," a single GitHub account containing all of your repositories--with each one corresponding to different research papers and named accordingly?

No, this is referring to a dedicated DOI number for a dataset.

Have you ever used Harvard Dataverse? Is it widely used in the department?

I haven't used it because I haven't used any data I could publish myself - it's all been either sensitive and couldn't be shared, or it was already publicly available somewhere else. In terms of the department, I'm not sure, but I think a few faculty members have used it.





Individual Project

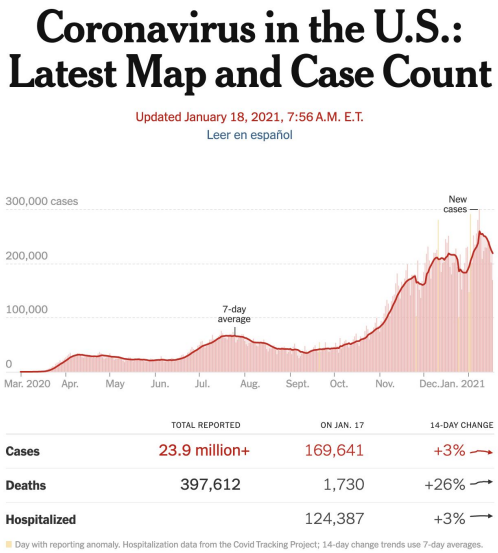
Reproducing a NYT Post

Data visualization is an incredibly powerful tool that can affect health policy decisions. Ensuring they are easy to interpret, and more importantly, showcase accurate insights from data is paramount for scientific transparency and the health of individuals. For this assignment you are tasked with reproducing COVID-19 visualizations and tables published by the New York Times.

Specifically, you will attempt to reproduce the following for January 17th, 2021:

1. New cases as a function of time with a rolling average plot - the first plot on the page (you don't need to recreate the colors or theme)

2. Table of cases, hospitalizations and deaths - the first table on the page



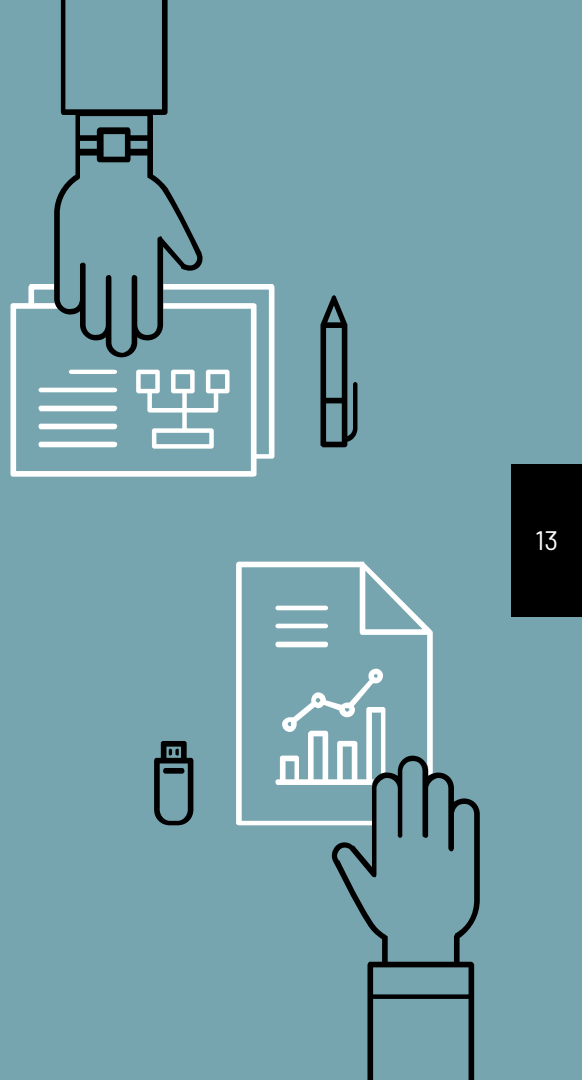
Coronavirus in the U.S.: Latest Map and Case Count

Updated January 18, 2021, 7:56 A.M. E.T.

Leer en español

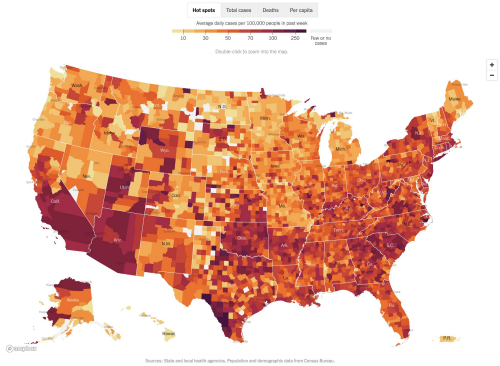
	TOTAL REPORTED	ON JAN. 17	14-DAY CHANGE
Cases	23.9 million+	169,641	+3%
Deaths	397,612	1,730	+26%
Hospitalized		124,387	+3%

Day with reporting anomaly. Hospitalization data from the Covid Tracking Project; 14-day change trends use 7-day averages.



Reproducing a NYT Post

3. (Optional) The county-level map for previous week ('Hot spots') - the second plot on the page (only the 'Hot Spots' plot)



4. Table of cases by state - the second table on the page (do not need to include per 100,000 or per capita columns)

Cases and deaths by state and county

This table is sorted by places with the most cases per 100,000 residents in the last seven days. Charts are colored to reveal when outbreaks emerged.

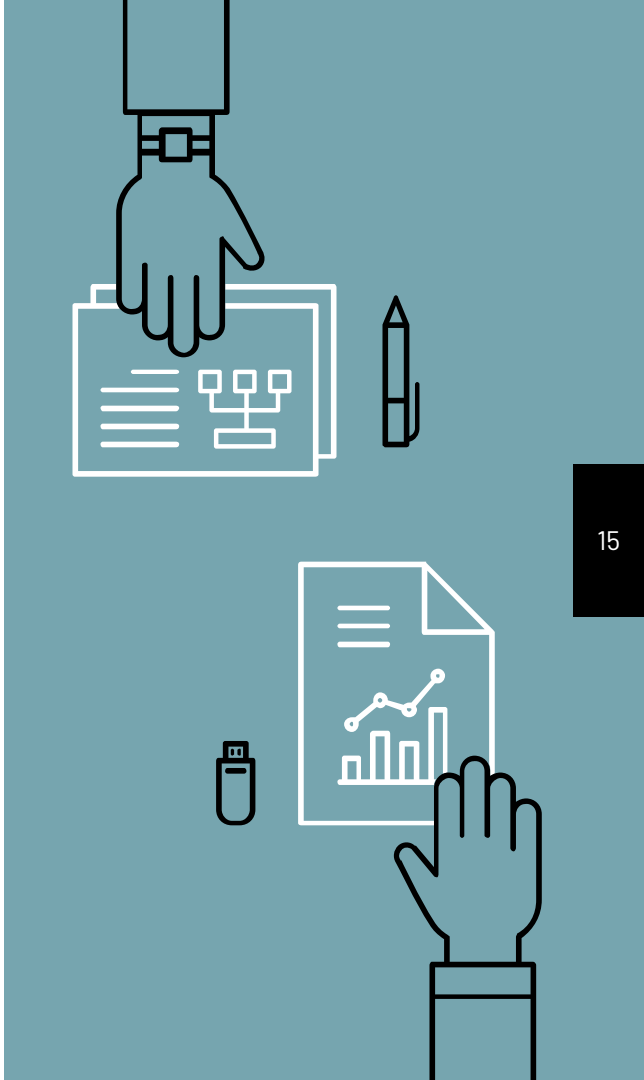
Cases | Deaths | Search counties

	TOTAL CASES	PER 100,000	DAILY AVG. IN LAST 7 DAYS	▼ PER 100,000	WEEKLY CASES PER CAPITA FEWER … MORE
+ Arizona MAP »	673,882	9,258	7,905	109	March 1 … Jan. 17
+ California MAP »	3,006,583	7,609	39,580	100	
+ South Carolina MAP »	388,184	7,539	4,808	93	
+ Rhode Island MAP »	104,443	9,859	976	92	
+ Oklahoma MAP »	354,979	8,971	3,374	85	
+ Georgia MAP »	791,322	7,453	8,457	80	
+ Utah MAP »	323,837	10,101	2,548	79	
+ Texas MAP »	2,127,334	7,337	22,782	79	
+ New York MAP »	1,242,818	6,389	15,281	79	
+ Massachusetts MAP »	470,140	6,821	5,336	77	

Reproducing a NYT Post

Notes:

- You can download the files here. You will need to create a repository (with README file) here.
- You don't need to make the plots look exactly the same as the post - just showcase the most important information
- You don't need to make the tables look pretty, but you do need to print out at least some of the rows to check the numbers against the NYT post
- Due: Monday, January 25th by 11:59pm EST
- You will have class time today and Thursday to work on it.
  - You don't have to stay on Zoom and can work on it at any time, but Matt and I will stay on Zoom until 11am each day if you have any questions or need help



Homework

- Watch Module 6 part 3 and Module 7 videos
  - 6.5.2 - 6.6
  - 7.1

- Submit Module 6 part 3 discussion points