HackHPC@ADMI22 | Training Session

Google Cloud

CloudyCluster by Omnibond

XSEDE
Extreme Science and Engineering
Discovery Environment

VOLTRON DATA

TACC
TEXAS ADVANCED COMPUTING CENTER

SGCI
Science Gateways
Community Institute

ADMI 2022

HackHPC@
ADMI
High Performance Computing
and Gateways 2022 Symposium
www.admiusa.org

Join the
HackHPC@ADMI22
Discord using this
QR Code!

Data to Dashboard

https://hackhpc.github.io/ADMI22/

HackHPC@ADMI22 | Training Session

Google Cloud

CloudyCluster
by Omnibond

XSEDE
Extreme Science and Engineering
Discovery Environment

VOLTRON DATA

TACC
TEXAS ADVANCED COMPUTING CENTER

SGCI
Science Gateways
Community Institute

AD MI 2022

HackHPC@
ADMI
High Performance Computing
and Gateways 2022 Symposium
www.admiusa.org

Data to Dashboard

https://hackhpc.github.io/ADMI22/

Agenda

1. Introductions
2. Hackathon Objective
3. Deliverables and Resources
4. General Information
5. Stages of Dashboard Implementation
6. Example Dashboard

Organizers

Linda Hayden - *ECSU/SGCI*
haydenl@mindspring.com

Amy Cannon - *Omnibond*
amycannon@omnibond.com

Alex Nolte - *University of Tartu*
alexander.nolte@ut.ee

Boyd Wilson - *Omnibond*
boyd@omnibond.com

Je'aime Powell - *TACC*
jpowell@tacc.utexas.edu

John Holly - *XSEDE*
jholly@sura.org

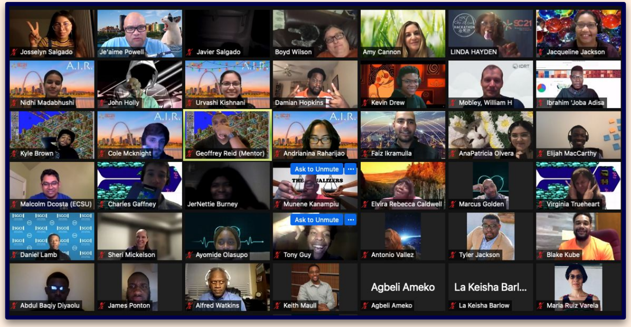
The Objective of HackHPC@ADMI

The hackathon aims to harness the resources, skills, and knowledge found in the HPC community in an effort to provide applied exposure towards students from 2-4 year post-secondary educational institutions. In short, the hackathon will provide HPC skills and training while targeting problems that directly affect the participants.

Develop knowledge about solutions to identified issues affecting them through application of data analysis/presentation or management.

Student Outcomes

- Increased familiarity with data science in the cloud
- Experience collaborative software engineering
- Develop professional communication skills



Student Deliverables and Resources

Deliverables:

- Source code Including Comments
- PDF of presentation
  - Team members with pictures
  - Use of HPC technology in the project
- Github Repository Link
  - README.md with project description

Resources:

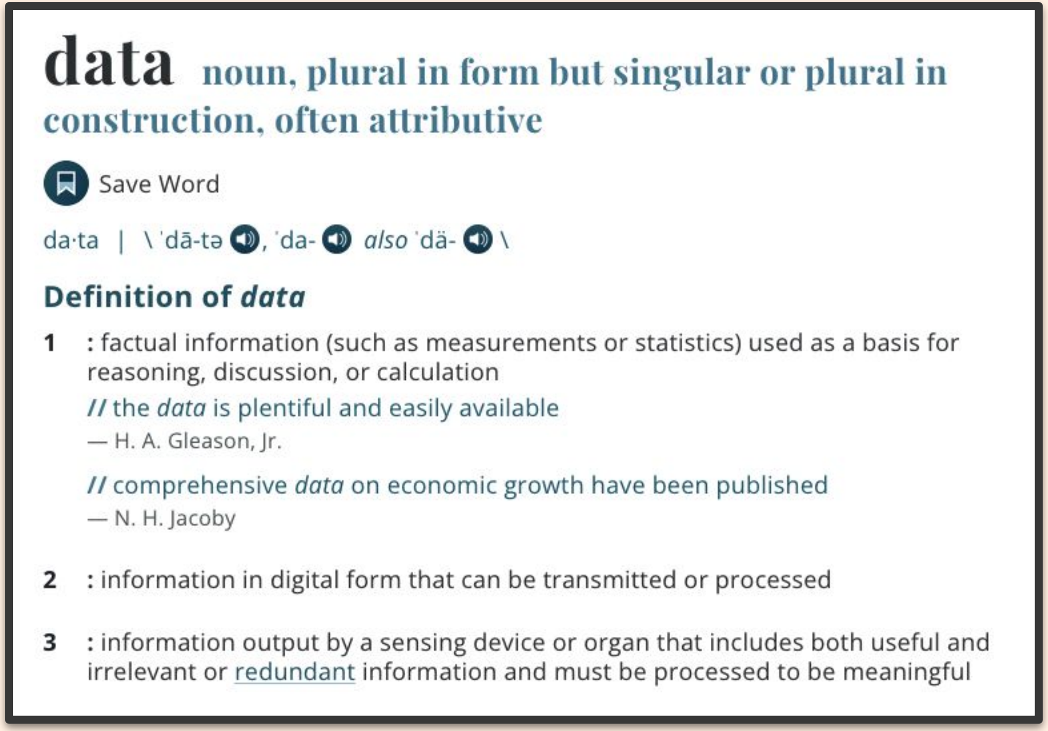
- Google Cloud (Provided Credits)
- Cloudy Cluster
- Most Commonly Used
  - Python
  - Jupyter Notebooks
  - Node.Js (JavaScript)
  - Repl.it (Collaborative Environment)
  - HTML
- Discord - https://discord.gg/ARg3vwWafF

General Information (the 3 T's)

- **Teams**
  - 4-5 Students
  - 1 Primary Mentor
  - 1 Technical Mentor
- **Time**
  - March 31st - April 4th
    - 3/31 @~7pm ET Event Start
      - *"The Draft"*
    - 4/[1-4] @ 11am ET & 7pm ET- Checkins
    - 4/4@6pm ET-Final Presentations

- **Topic Examples**
  - Data Analysis of COVID 19
  - Economic disparities and their effects on college participation
  - Genomics, Molecular Dynamics, or Weather Modeling in the Cloud.
  - Social Justice
  - AI-based Crowd Status
  - Public Data Management
  - Graduation Rates
  - Broadband Access
  - Insurance vs. Public Health Resilience

What is Data?



data noun, plural in form but singular or plural in construction, often attributive

Save Word

da·ta | \ ˈdā-tə , ˈda- *also* ˈdä- \

Definition of *data*

1 : factual information (such as measurements or statistics) used as a basis for reasoning, discussion, or calculation

// the *data* is plentiful and easily available

— H. A. Gleason, Jr.

// comprehensive *data* on economic growth have been published

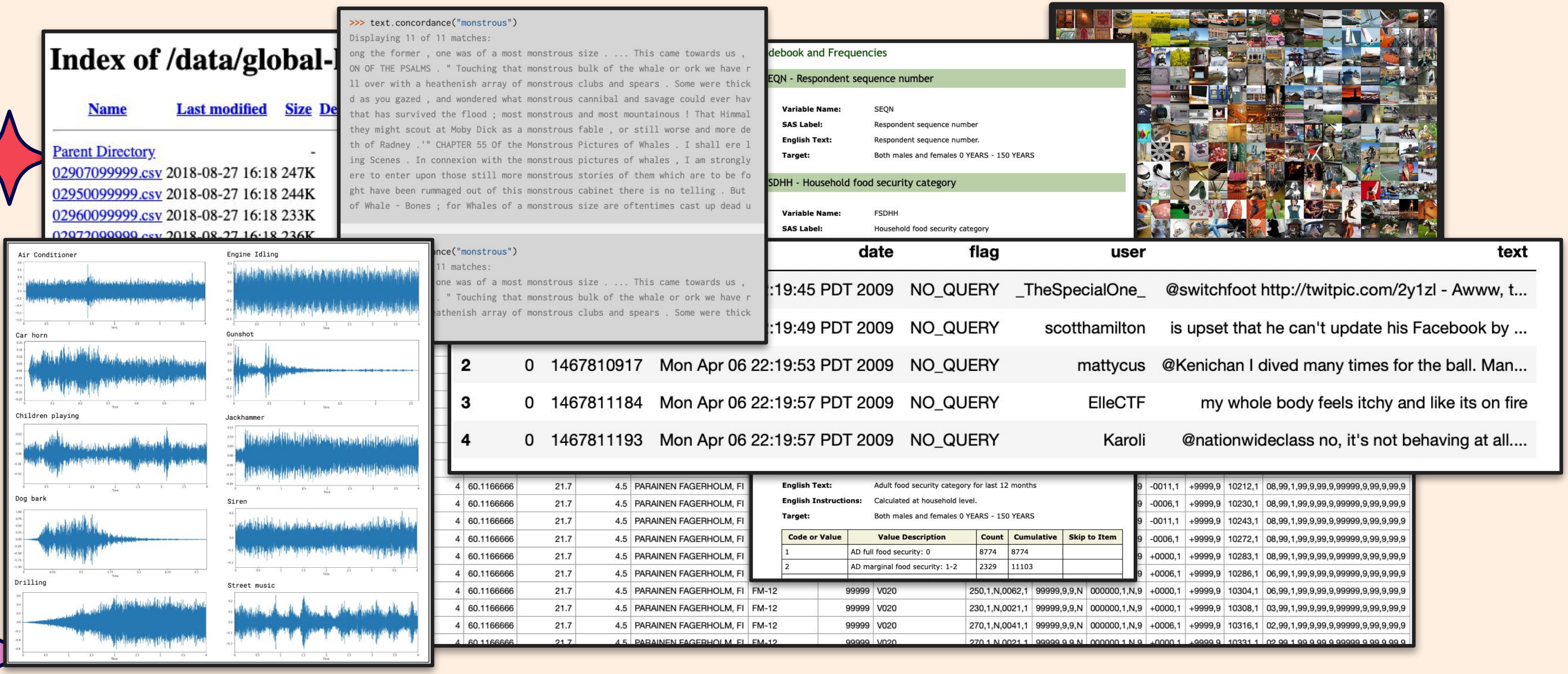
— N. H. Jacoby

2 : information in digital form that can be transmitted or processed

3 : information output by a sensing device or organ that includes both useful and irrelevant or redundant information and must be processed to be meaningful

Ref: *https://www.merriam-webster.com/dictionary/data*

OK, Yeah but... what is Data?



Index of /data/global-l

Name	Last modified	Size	De
Parent Directory		-	
02907099999.csv	2018-08-27 16:18	247K	
02950099999.csv	2018-08-27 16:18	244K	
02960099999.csv	2018-08-27 16:18	233K	
02972099999.csv	2018-08-27 16:18	236K	

```
>>> text.concordance("monstrous")
Displaying 11 of 11 matches:
ong the former , one was of a most monstrous size . ... This came towards us ,
ON OF THE PSALMS . " Touching that monstrous bulk of the whale or ork we have r
ll over with a heathenish array of monstrous clubs and spears . Some were thick
d as you gazed , and wondered what monstrous cannibal and savage could ever hav
that has survived the flood ; most monstrous and most mountainous ! That Himmal
they might scout at Moby Dick as a monstrous fable , or still worse and more de
th of Radney .'" CHAPTER 55 Of the Monstrous Pictures of Whales . I shall ere l
ing Scenes . In connexion with the monstrous pictures of whales , I am strongly
ere to enter upon those still more monstrous stories of them which are to be fo
ght have been rummaged out of this monstrous cabinet there is no telling . But
of Whale - Bones ; for Whales of a monstrous size are oftentimes cast up dead u
```

```
nce("monstrous")
 11 matches:
one was of a most monstrous size . ... This came towards us ,
. " Touching that monstrous bulk of the whale or ork we have r
eathenish array of monstrous clubs and spears . Some were thick
```

debook and Frequencies

EQN - Respondent sequence number

Variable Name: SEQN
SAS Label: Respondent sequence number
English Text: Respondent sequence number.
Target: Both males and females 0 YEARS - 150 YEARS

SDHH - Household food security category

Variable Name: FSDHH
SAS Label: Household food security category
English Text: Adult food security category for last 12 months
English Instructions: Calculated at household level.
Target: Both males and females 0 YEARS - 150 YEARS

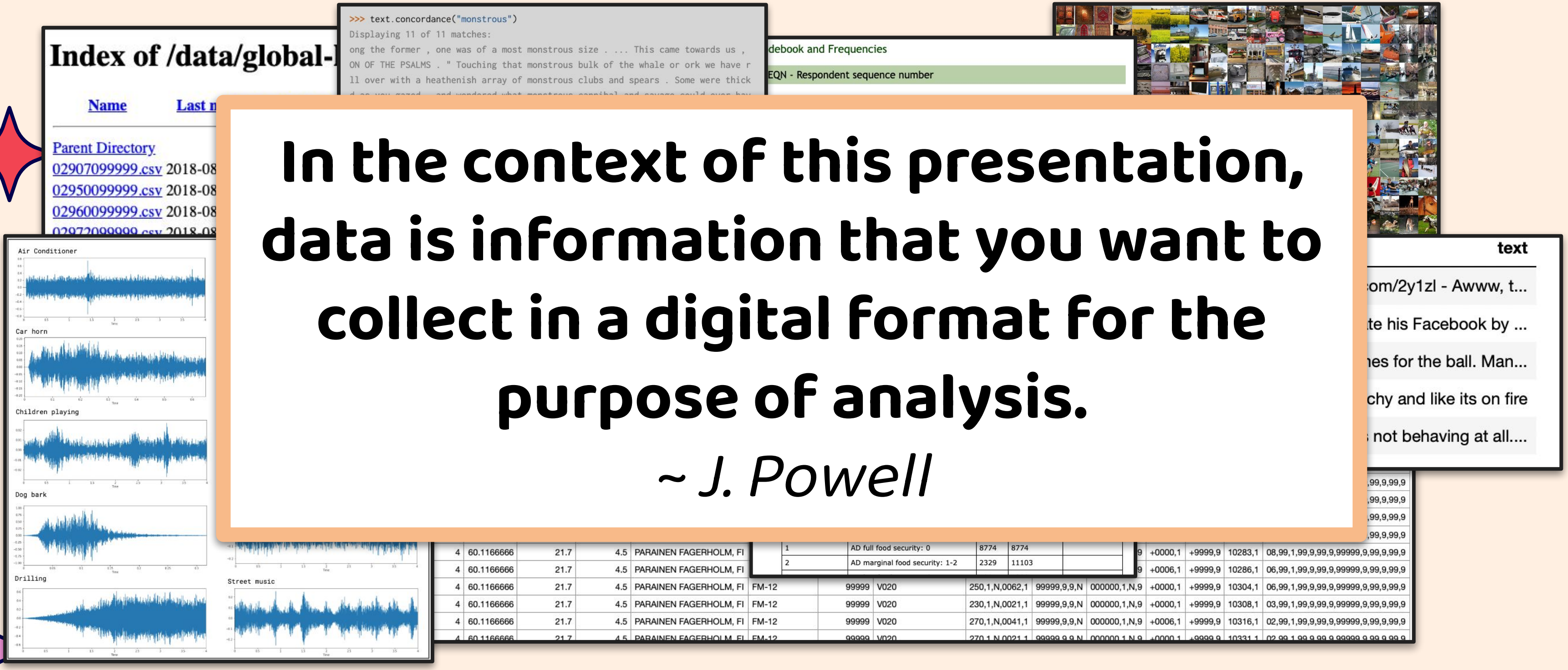
Code or Value	Value Description	Count	Cumulative	Skip to Item
1	AD full food security: 0	8774	8774	
2	AD marginal food security: 1-2	2329	11103	

			date	flag	user	text
			:19:45 PDT 2009	NO_QUERY	_TheSpecialOne_	@switchfoot http://twitpic.com/2y1zl - Awww, t...
			:19:49 PDT 2009	NO_QUERY	scotthamilton	is upset that he can't update his Facebook by ...
2	0	1467810917	Mon Apr 06 22:19:53 PDT 2009	NO_QUERY	mattycus	@Kenichan I dived many times for the ball. Man...
3	0	1467811184	Mon Apr 06 22:19:57 PDT 2009	NO_QUERY	ElleCTF	my whole body feels itchy and like its on fire
4	0	1467811193	Mon Apr 06 22:19:57 PDT 2009	NO_QUERY	Karoli	@nationwideclass no, it's not behaving at all....

4	60.1166666	21.7	4.5	PARAINEN FAGERHOLM, FI						9	-0011,1	+9999,9	10212,1	08,99,1,99,9,99,9,99999,9,99,9,99,9
4	60.1166666	21.7	4.5	PARAINEN FAGERHOLM, FI						9	-0006,1	+9999,9	10230,1	08,99,1,99,9,99,9,99999,9,99,9,99,9
4	60.1166666	21.7	4.5	PARAINEN FAGERHOLM, FI						9	-0011,1	+9999,9	10243,1	08,99,1,99,9,99,9,99999,9,99,9,99,9
4	60.1166666	21.7	4.5	PARAINEN FAGERHOLM, FI						9	-0006,1	+9999,9	10272,1	08,99,1,99,9,99,9,99999,9,99,9,99,9
4	60.1166666	21.7	4.5	PARAINEN FAGERHOLM, FI						9	+0000,1	+9999,9	10283,1	08,99,1,99,9,99,9,99999,9,99,9,99,9
4	60.1166666	21.7	4.5	PARAINEN FAGERHOLM, FI						9	+0006,1	+9999,9	10286,1	06,99,1,99,9,99,9,99999,9,99,9,99,9
4	60.1166666	21.7	4.5	PARAINEN FAGERHOLM, FI	FM-12	99999	V020	250,1,N,0062,1	99999,9,9,N	000000,1,N,9	+0000,1	+9999,9	10304,1	06,99,1,99,9,99,9,99999,9,99,9,99,9
4	60.1166666	21.7	4.5	PARAINEN FAGERHOLM, FI	FM-12	99999	V020	230,1,N,0021,1	99999,9,9,N	000000,1,N,9	+0000,1	+9999,9	10308,1	03,99,1,99,9,99,9,99999,9,99,9,99,9
4	60.1166666	21.7	4.5	PARAINEN FAGERHOLM, FI	FM-12	99999	V020	270,1,N,0041,1	99999,9,9,N	000000,1,N,9	+0006,1	+9999,9	10316,1	02,99,1,99,9,99,9,99999,9,99,9,99,9
4	60.1166666	21.7	4.5	PARAINEN FAGERHOLM, FI	FM-12	99999	V020	270,1,N,0021,1	99999,9,9,N	000000,1,N,9	+0000,1	+9999,9	10331,1	02,99,1,99,9,99,9,99999,9,99,9,99,9

https://hackhpc.github.io/ADMI22/

OK, Yeah but... what is Data?



Index of /data/global-

Name Last

Parent Directory
02907099999.csv 2018-08
02950099999.csv 2018-08
02960099999.csv 2018-08

```
>>> text.concordance("monstrous")
Displaying 11 of 11 matches:
ong the former , one was of a most monstrous size . ... This came towards us ,
ON OF THE PSALMS . " Touching that monstrous bulk of the whale or ork we have r
ll over with a heathenish array of monstrous clubs and spears . Some were thick
```

debook and Frequencies

EQN - Respondent sequence number

1	AD full food security: 0	8774	8774
2	AD marginal food security: 1-2	2329	11103

text
om/2y1zl - Awww, t...
te his Facebook by ...
es for the ball. Man...
chy and like its on fire
not behaving at all....

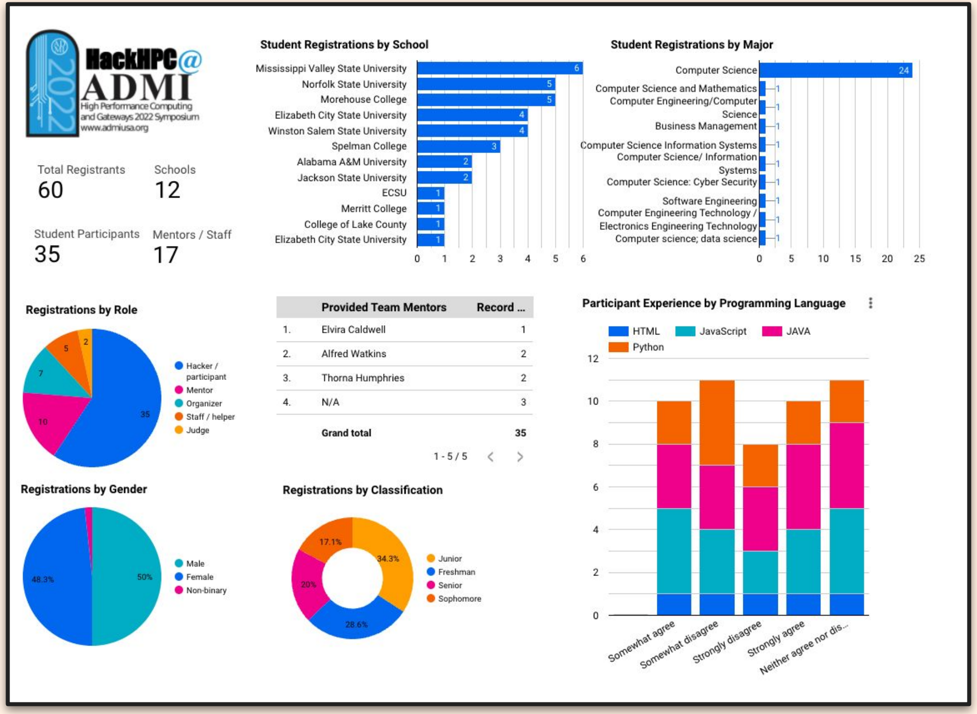
```
4  60.1166666  21.7  4.5  PARAINEN FAGERHOLM, FI                            +0000,1  +9999,9  10283,1  08,99,1,99,9,99,9,99999,9,99,9,99,9
4  60.1166666  21.7  4.5  PARAINEN FAGERHOLM, FI                            +0006,1  +9999,9  10286,1  06,99,1,99,9,99,9,99999,9,99,9,99,9
4  60.1166666  21.7  4.5  PARAINEN FAGERHOLM, FI  FM-12  99999  V020  250,1,N,0062,1  99999,9,9,N  000000,1,N,9  +0000,1  +9999,9  10304,1  06,99,1,99,9,99,9,99999,9,99,9,99,9
4  60.1166666  21.7  4.5  PARAINEN FAGERHOLM, FI  FM-12  99999  V020  230,1,N,0021,1  99999,9,9,N  000000,1,N,9  +0000,1  +9999,9  10308,1  03,99,1,99,9,99,9,99999,9,99,9,99,9
4  60.1166666  21.7  4.5  PARAINEN FAGERHOLM, FI  FM-12  99999  V020  270,1,N,0041,1  99999,9,9,N  000000,1,N,9  +0006,1  +9999,9  10316,1  02,99,1,99,9,99,9,99999,9,99,9,99,9
```

> In the context of this presentation, data is information that you want to collect in a digital format for the purpose of analysis.
>
> *~ J. Powell*

HackHPC@ADMI22 | Training Session

I'm Scared to ask but... What is a Dashboard?

A "*Dashboard*" frames a problem by telling a story using your data.



HackHPC@ ADMI High Performance Computing and Gateways 2022 Symposium www.admiusa.org

Total Registrants 60

Schools 12

Student Participants 35

Mentors / Staff 17

	Provided Team Mentors	Record ...
1.	Elvira Caldwell	1
2.	Alfred Watkins	2
3.	Thorna Humphries	2
4.	N/A	3
	Grand total	**35**

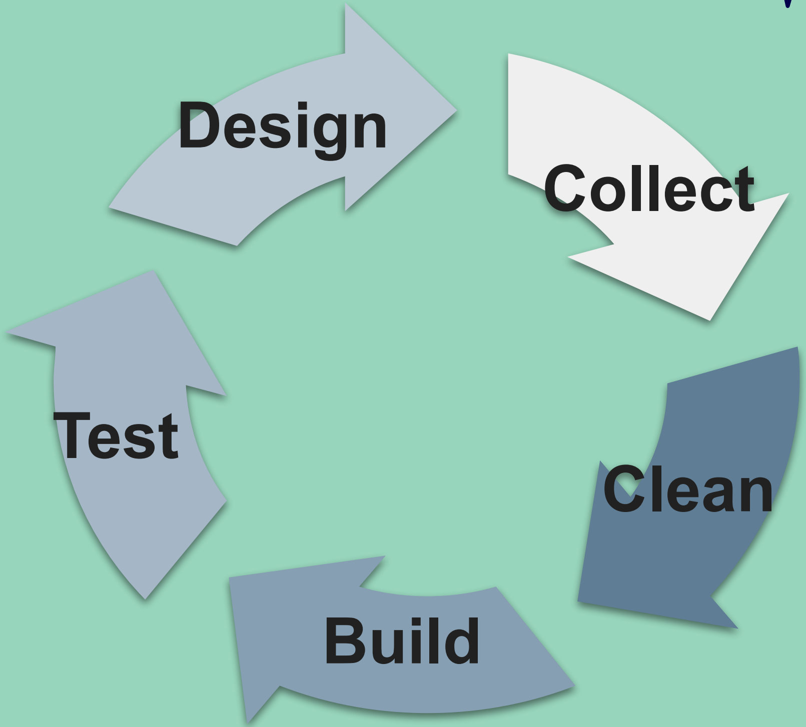
1 - 5 / 5

Student Registrations by School
Mississippi Valley State University 6
Norfolk State University 5
Morehouse College 5
Elizabeth City State University 4
Winston Salem State University 4
Spelman College 3
Alabama A&M University 2
Jackson State University 2
ECSU 1
Merritt College 1
College of Lake County 1
Elizabeth City State University 1
0 1 2 3 4 5 6
Student Registrations by Major
Computer Science 24
Computer Science and Mathematics 1
Computer Engineering/Computer Science 1
Business Management 1
Computer Science Information Systems 1
Computer Science/ Information Systems 1
Computer Science: Cyber Security 1
Software Engineering 1
Computer Engineering Technology / Electronics Engineering Technology 1
Computer science; data science 1
0 5 10 15 20 25
Registrations by Role
35
10
7
5
2
Hacker / participant
Mentor
Organizer
Staff / helper
Judge
Registrations by Gender
48.3%
50%
Male
Female
Non-binary
Registrations by Classification
17.1%
34.3%
20%
28.6%
Junior
Freshman
Senior
Sophomore
Participant Experience by Programming Language
HTML
JavaScript
JAVA
Python
0 2 4 6 8 10 12
Somewhat agree
Somewhat disagree
Strongly disagree
Strongly agree
Neither agree nor dis...

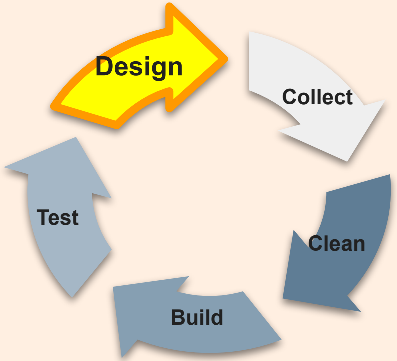
https://hackhpc.github.io/ADMI22/

The Dashboard Developmental Process

1. Dashboard Design
2. Collect Data
3. Piping/Cleaning
4. Build Dashboard
5. Feedback/Test



Dashboard Design

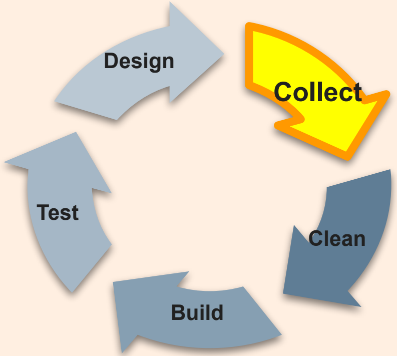


- **Who** is the audience
- **What** information should they get from your dashboard / **What** question are you answering?
- **When** is the temporal connection between the dashboard and data [dynamic vs static data]
- **Where** - the platform? (*desktop, server, kiosk*)
- **Why** - the goal for the whole project
  - Visualization - chart type
  - Pen and paper mockup

Output(s) from step:

- Site mockup
- Clearly defined question(s)
- Platform to be used
- Type of data needed for analysis

Collect Data



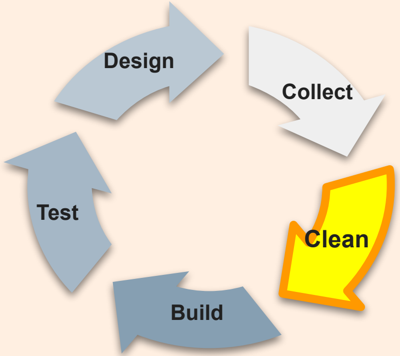
[Note: Third most time consuming process]

- Which datasets do you have access?
- What questions do you **WANT** to ask of the data?
- What questions **CAN** you answer from the available data you have?
  - Alternate analysis/indirect correlations

Output(s) from step:

- Dataset(s)
- Data Dictionaries
- Suggested analysis/correlation methods
- Dataset Documentation
- Database/Storage location(s)

Data Piping/Cleaning



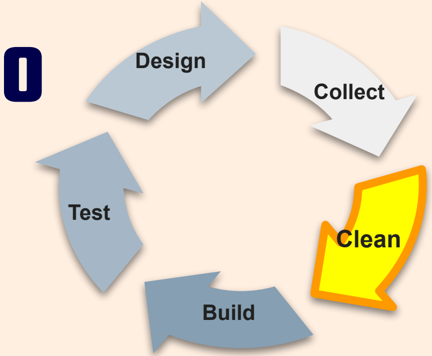
[Note: Most time consuming process!]

- Take raw data in
- Write scripts for necessary data transformations
  - *Python, R, JupyterNotebook*
- Identify data storage locations
- Handle moving data between locations
- Consider: data that changes over time

Output(s) from step:

- Clean Dataset(s)
- scripts for transformation
- output files
- database connections

Data Piping/Cleaning - GIGO



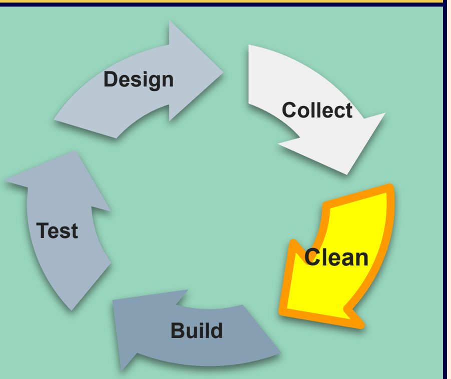
The reason this is the MOST time consuming process!

GIGO = Garbage In, Garbage Out

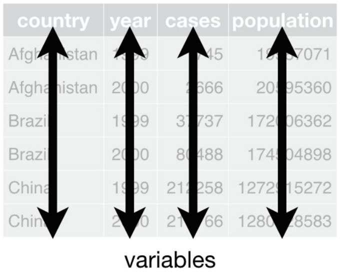
If your data is not properly organized and "transformed" the results will likely not make sense!

- Data Validation
- Proper/Non-Repeating Headers
- Proper databases
    - Georeference-enabled

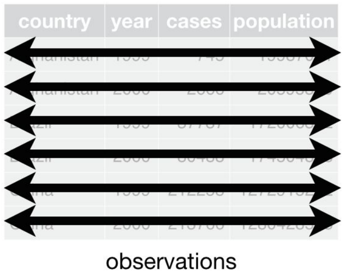
Data Piping/Cleaning - Tidy Data



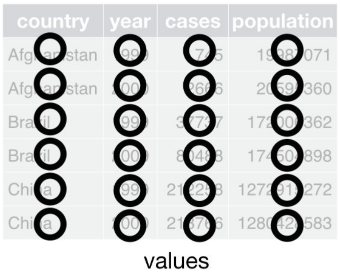
Each variable is a column.



Each observation is a row.



Each value is a cell.

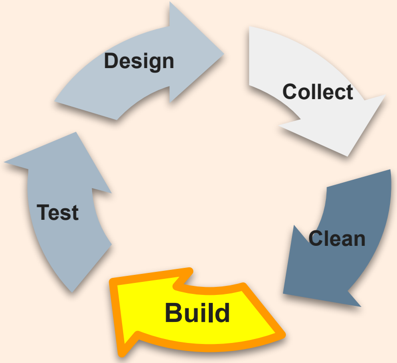


https://hackhpc.github.io/ADMI22/

Build Dashboard

[Note: Second most time consuming process]

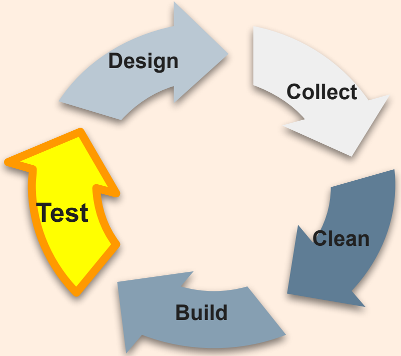
- Load outputs of data pipes/sources
- Code chart elements on page
- Code User interactivity
  - Data filters
  - Selection methods
  - Changing elements



Output(s) from step:

- Code used to build the dashboard
- Deployed dashboard locally or to a cloud service

Feedback/Testing



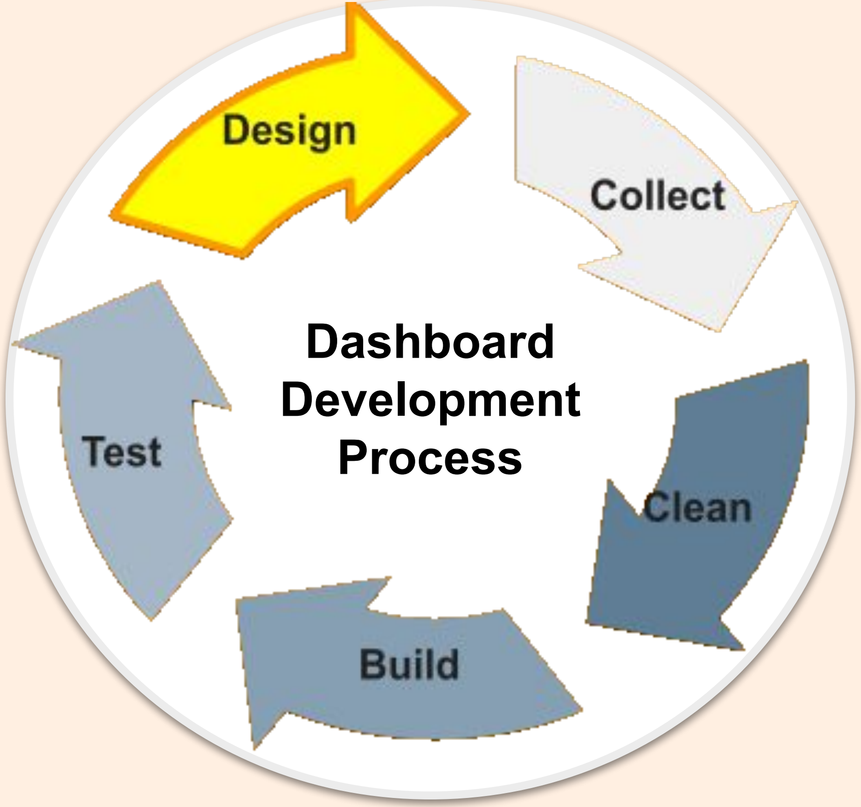
- **Demonstration to Client / Users**
  - Ideally a live deployed version
  - Screenshots / PDF better than nothing
- **Collect and Integrate feedback into next iterative development cycle**

DID DASHBOARD TELL THE STORY / AID THE DECISION / ANSWER THE QUESTION?

Output(s) from step:

- Documented feedback
- Informed tasking for the next iteration(s) of the design

Iterate the process until done!

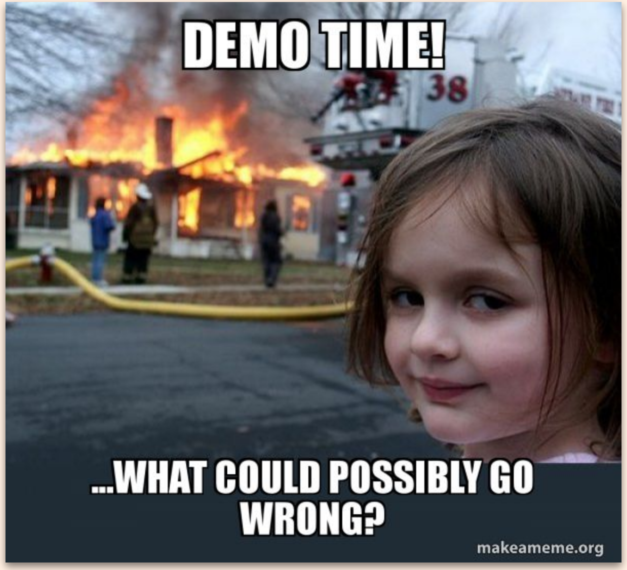


Dashboard Example

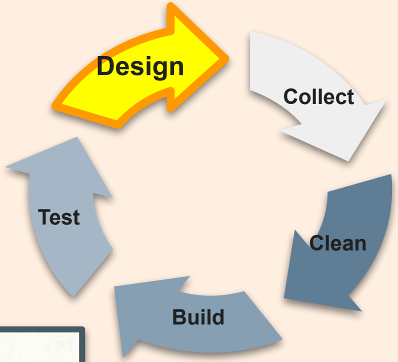
Demo Time!!

Example GitHub Repo:
https://github.com/mepearson/texas_congress

Deployed Heroku App:
https://texas-congress.herokuapp.com/



Dashboard Design

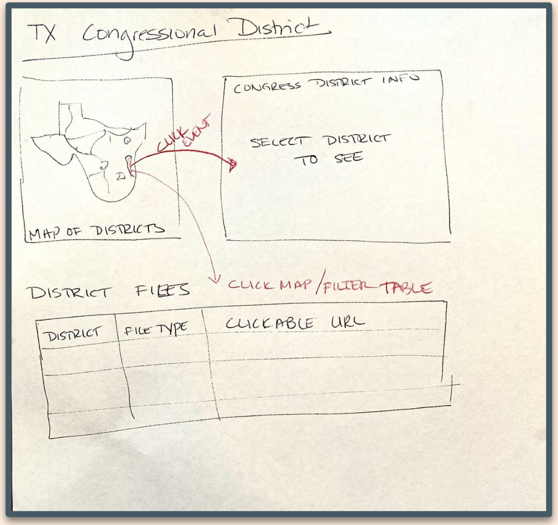


WHAT:

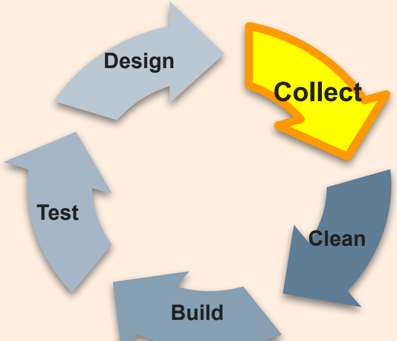
- Dashboard to link TX residents with information for their US Congressional District

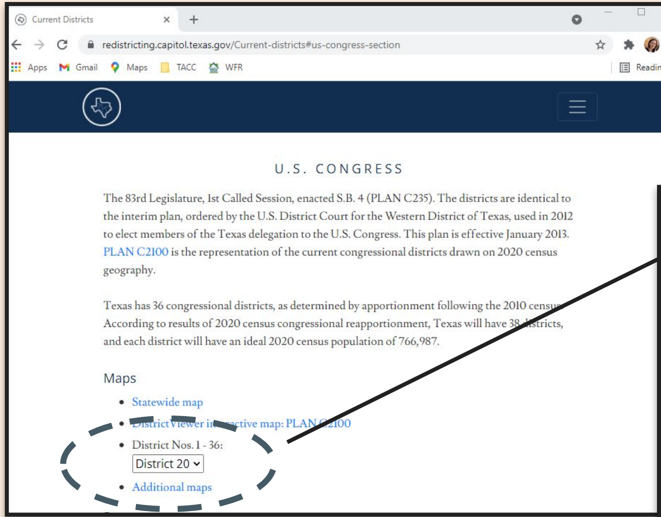
DESIRED ELEMENTS:

- Selectable map of Congressional Districts
- Display section for information related to selected District
- Table of clickable links to access District Information Files



Collect Data




Current Districts
redistricting.capitol.texas.gov/Current-districts#us-congress-section
Apps
Gmail
Maps
TACC
WFR
Reading list

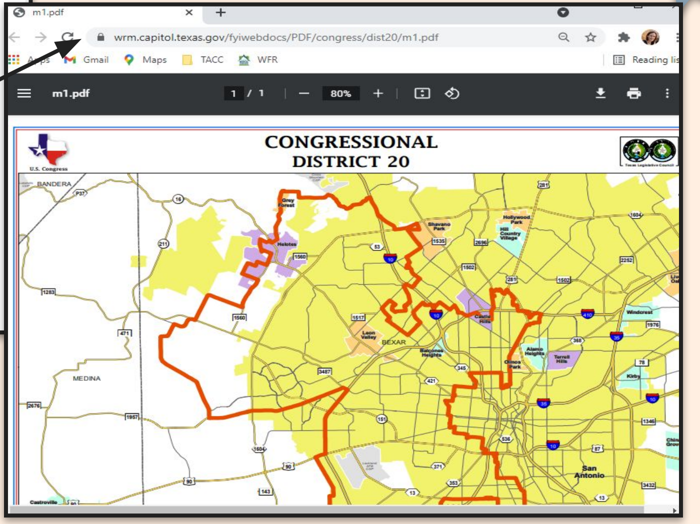
U.S. CONGRESS

The 83rd Legislature, 1st Called Session, enacted S.B. 4 (PLAN C235). The districts are identical to the interim plan, ordered by the U.S. District Court for the Western District of Texas, used in 2012 to elect members of the Texas delegation to the U.S. Congress. This plan is effective January 2013. PLAN C2100 is the representation of the current congressional districts drawn on 2020 census geography.

Texas has 36 congressional districts, as determined by apportionment following the 2010 census. According to results of 2020 census congressional reapportionment, Texas will have 38 districts, and each district will have an ideal 2020 census population of 766,987.

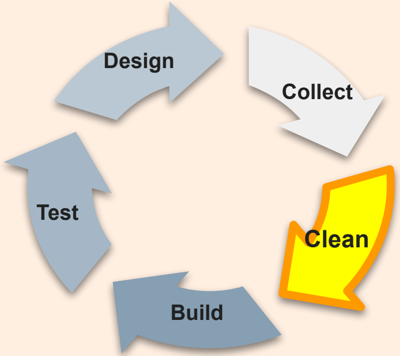
Maps

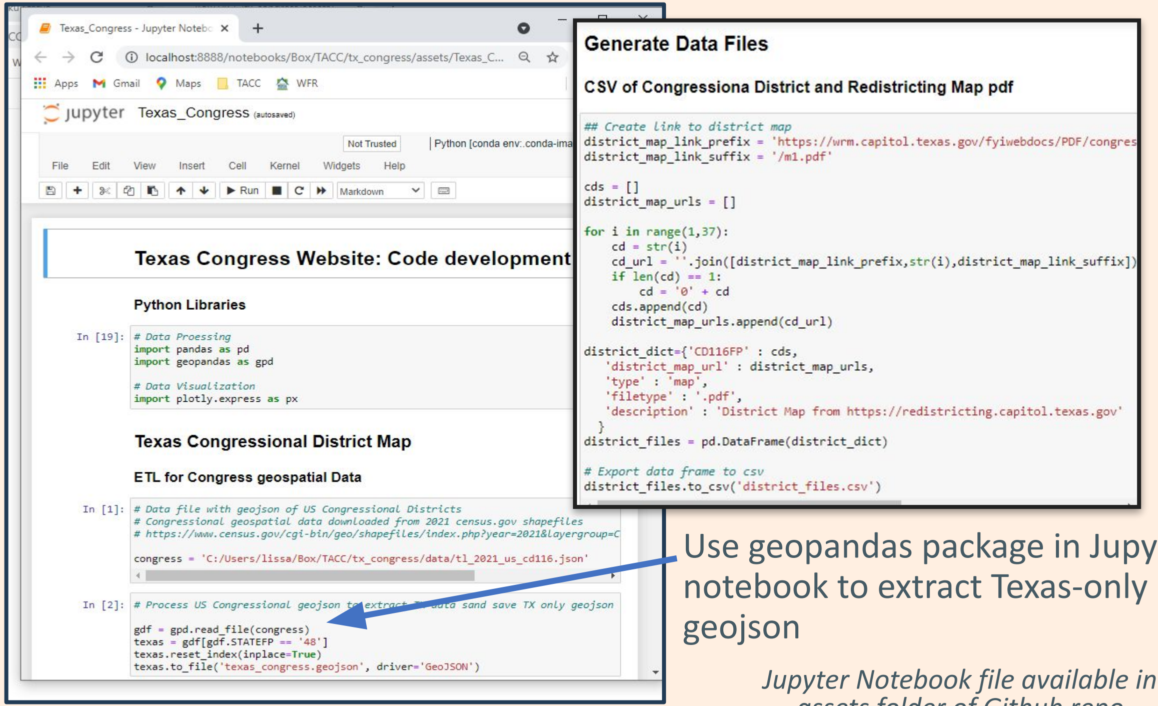
- Statewide map
- DistrictViewer interactive map: PLAN C2100
- District Nos. 1 - 36: District 20
- Additional maps



Get links to Congressional District maps from redistricting.capitol.texas.gov site

Data Piping/Cleaning




Texas_Congress - Jupyter Noteb
localhost:8888/notebooks/Box/TACC/tx_congress/assets/Texas_C...
Apps
Gmail
Maps
TACC
WFR
jupyter
Texas_Congress (autosaved)
Not Trusted
Python [conda env:.conda-ima
File
Edit
View
Insert
Cell
Kernel
Widgets
Help
Run
Markdown

Texas Congress Website: Code development

Python Libraries

```
In [19]: # Data Proessing
         import pandas as pd
         import geopandas as gpd

         # Data Visualization
         import plotly.express as px
```

Texas Congressional District Map

ETL for Congress geospatial Data

```
In [1]:  # Data file with geojson of US Congressional Districts
         # Congressional geospatial data downloaded from 2021 census.gov shapefiles
         # https://www.census.gov/cgi-bin/geo/shapefiles/index.php?year=2021&layergroup=C

         congress = 'C:/Users/lissa/Box/TACC/tx_congress/data/tl_2021_us_cd116.json'
```

```
In [2]:  # Process US Congressional geojson to extract TX data sand save TX only geojson

         gdf = gpd.read_file(congress)
         texas = gdf[gdf.STATEFP == '48']
         texas.reset_index(inplace=True)
         texas.to_file('texas_congress.geojson', driver='GeoJSON')
```

Generate Data Files

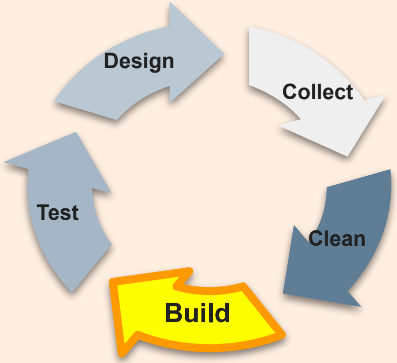
CSV of Congressiona District and Redistricting Map pdf

```
## Create link to district map
district_map_link_prefix = 'https://wrm.capitol.texas.gov/fyiwebdocs/PDF/congres
district_map_link_suffix = '/m1.pdf'

cds = []
district_map_urls = []

for i in range(1,37):
    cd = str(i)
    cd_url = ''.join([district_map_link_prefix,str(i),district_map_link_suffix])
    if len(cd) == 1:
        cd = '0' + cd
    cds.append(cd)
    district_map_urls.append(cd_url)

district_dict={'CD116FP' : cds,
   'district_map_url' : district_map_urls,
   'type' : 'map',
   'filetype' : '.pdf',
   'description' : 'District Map from https://redistricting.capitol.texas.gov'
  }
district_files = pd.DataFrame(district_dict)

# Export data frame to csv
district_files.to_csv('district_files.csv')
```

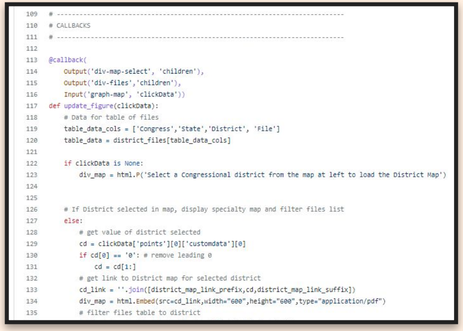
Use geopandas package in Jupyter notebook to extract Texas-only geojson

Jupyter Notebook file available in assets folder of Github repo

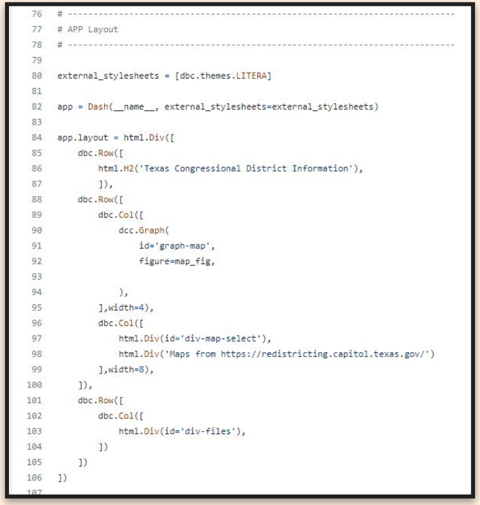
Build Dashboard



- **Write Dash code in IDE of choice**
- **Parts of App.py File:**
  - Python libraries
  - DATA Loading and DATA Visualizations
  - APP Layout – layout elements of page, similar to html
  - Callbacks – provide user interactivity / communication between elements
  - Run App



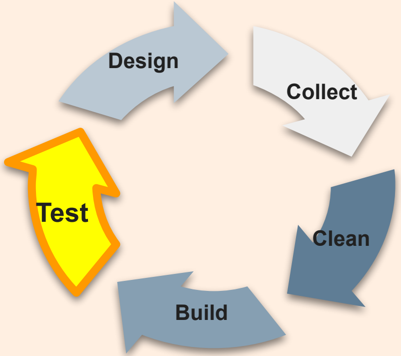
```python
# ---------------------------------------------------------------------------
# CALLBACKS
# ---------------------------------------------------------------------------

@callback(
    Output('div-map-select', 'children'),
    Output('div-files','children'),
    Input('graph-map', 'clickData'))
def update_figure(clickData):
    # Data for table of files
    table_data_cols = ['Congress','State','District', 'File']
    table_data = district_files[table_data_cols]

    if clickData is None:
        div_map = html.P('Select a Congressional district from the map at left to load the District Map')


    # If District selected in map, display specialty map and filter files list
    else:
        # get value of district selected
        cd = clickData['points'][0]['customdata'][0]
        if cd[0] == '0': # remove leading 0
            cd = cd[1:]
        # get link to District map for selected district
        cd_link = ''.join([district_map_link_prefix,cd,district_map_link_suffix])
        div_map = html.Embed(src=cd_link,width="600",height="600",type="application/pdf")
        # filter files table to district
```

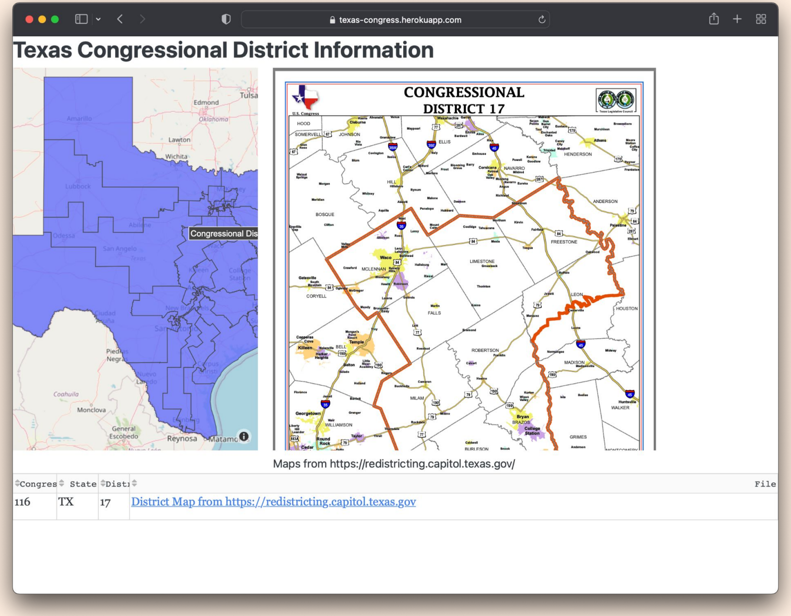


```python
# ---------------------------------------------------------------------------
# APP Layout
# ---------------------------------------------------------------------------

external_stylesheets = [dbc.themes.LITERA]

app = Dash(__name__, external_stylesheets=external_stylesheets)

app.layout = html.Div([
    dbc.Row([
        html.H2('Texas Congressional District Information'),
        ]),
    dbc.Row([
        dbc.Col([
            dcc.Graph(
                id='graph-map',
                figure=map_fig,

            ),
        ],width=4),
        dbc.Col([
            html.Div(id='div-map-select'),
            html.Div('Maps from https://redistricting.capitol.texas.gov/')
        ],width=8),
    ]),
    dbc.Row([
        dbc.Col([
            html.Div(id='div-files'),
        ])
    ])
])
```

https://hackhpc.github.io/ADMI22/

Feedback/Testing





Texas Congressional District Information

Maps from https://redistricting.capitol.texas.gov/

Congres	State	Distr	File
116	TX	17	District Map from https://redistricting.capitol.texas.gov

Deployed Heroku App:
https://texas-congress.herokuapp.com/

Additional References

Data Management

- R for Data Science. Code in R / concepts useful any language
  Welcome | R for Data Science (had.co.nz)
- Blog Overview (easy read): Tidy data for efficiency, reproducibility, and collaboration (openscapes.org)
- Original paper by Hadley Wickham (founder of R) who pioneered the concept of tidy data:
  - Official Paper: Tidy data (had.co.nz)
  - informal and example code heavy (in R) version: Tidy data • tidyr (tidyverse.org)

Data Visualization

- Chart Chooser — Juice Analytics - https://www.juiceanalytics.com/chartchooser
- Plotly graphing library - https://plotly.com/python/

Dash App

- Dash App documentation - https://dash.plotly.com/
- Deploy to Heroku
  - integration from github [https://devcenter.heroku.com/articles/github-integration]
  - Dash guidance / command line (scroll past Enterprise information to Heroku / free section) - https://dash.plotly.com/deployment

Questions and Concerns

Next Training Session:

- **Google / CloudyCluster -** [*6/26/22*]

Schedule:

https://hackhpc.github.io/ADMI22/schedule.html

Presenters Contact Information:

Je'aime Powell (*TACC*) - jpowell@tacc.utexas.edu