HPC in the City: St. Louis

SC21 St. Louis, MO | science & beyond.

HACKATHON

Data to Dashboard Training

Join the HPC in the City Discord using this QR Code!

intel

cloudyCluster by Omnibond

globus online

SGCI

Google Cloud

sighpc

TACC
TEXAS ADVANCED COMPUTING CENTER

XSEDE

HPC in the City: St. Louis

SC21 St. Louis, MO | science & beyond.

HACKATHON

Data to Dashboard Training

Agenda

- Introductions
- Hackathon Objective
- Deliverables and Resources
- General Information
- Data to Dashboard

Presenter: Je'aime Powell

Organizers

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

Boyd Wilson - *Omnibond*
boyd@omnibond.com

Amy Cannon - *Omnibond*
amycannon@omnibond.com

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

Linda Hayden - *ECSU*
haydenl@mindspring.com

The Objective of HPC in the City: St. Louis

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 St. Louis 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
  - Regional (St. Louis) implications of 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.com/invite/rSXasYKDwE**

Join the HPCHack Discord using this QR Code!

General Information (the 3 T's)

- **Teams**
  - 4-5 Students
  - 1 Primary Mentor
  - 1 Specialist/Staff
- **Time**
  - November 4th - 8th
    - 11/4@~6pm ET Event Start
      - Team formation
    - 11/[5-8] @ 11 ET & 6pm ET- Checkins
    - 11/8@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

Presenter: Melissa Pearson (TACC)

Creating a Data-based Dashboard Application

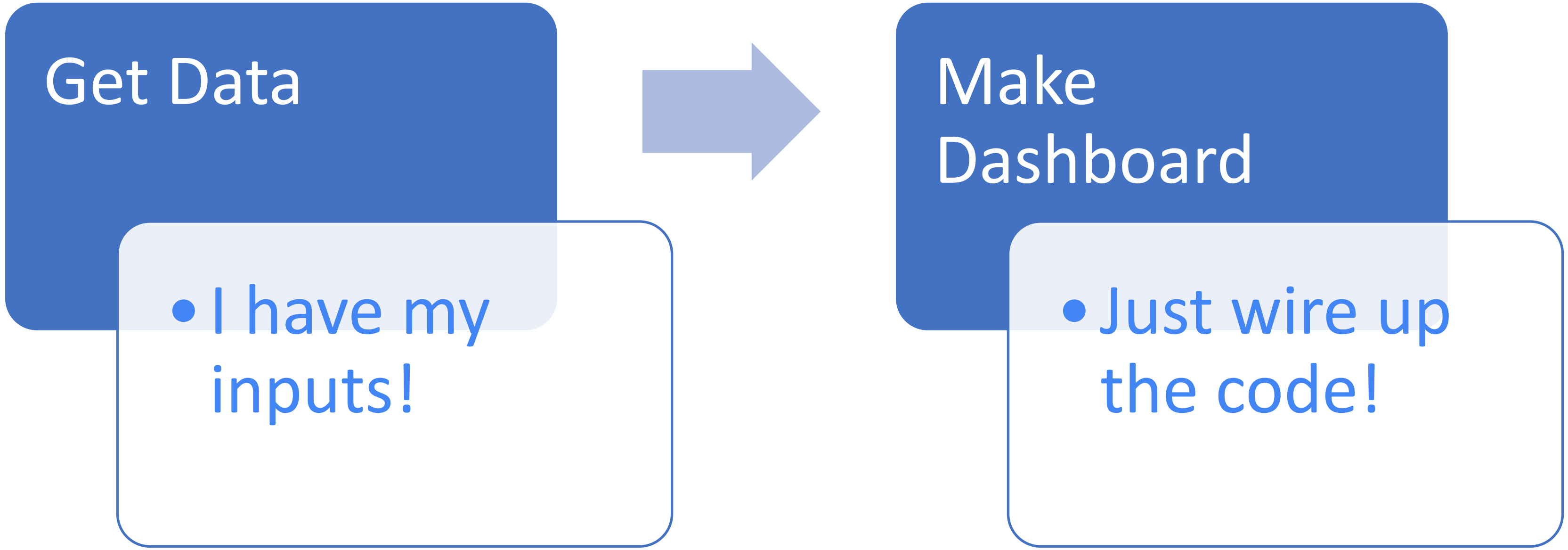
Talk Structure

BLUF: <u>B</u>ottom <u>L</u>ine <u>U</u>p <u>F</u>ront

Data to Dashboard Workflows	● Workflow: Perception.... ● .... vs. Reality ● Step by Step walkthrough
Example Dash App	● Simple application use case: TX Congressional District Info ● Example of workflow and products at each step ● Quick overview of deploying Github repo to Heroku
Data Piping: In Detail	● Data Cleaning and Wrangling ● <u>E</u>(xtraction), <u>T</u>(ransformation) and <u>L</u>(oad) ● "Tidy" Data

Data to Dashboards: Workflows

How people imagine the workflow….



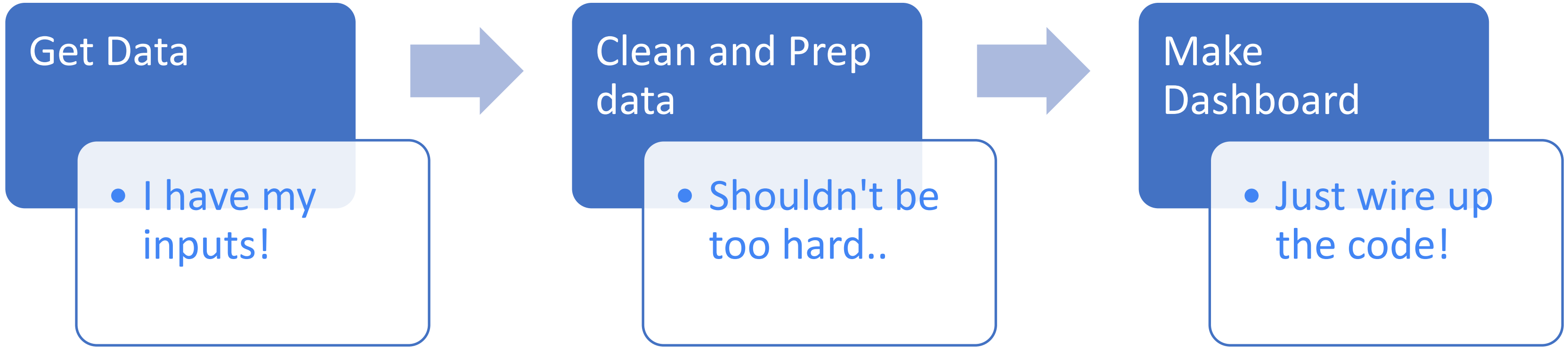
Get Data

- I have my inputs!

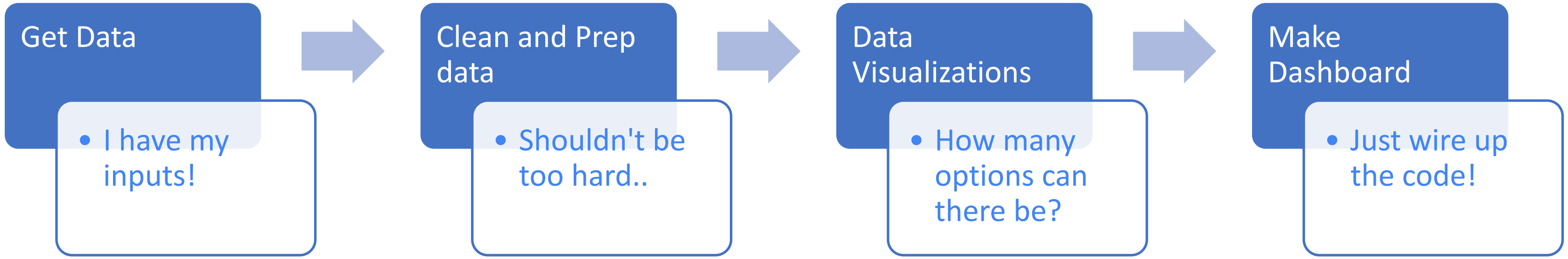
Make Dashboard

- Just wire up the code!

....when you point out the missing data cleaning step....

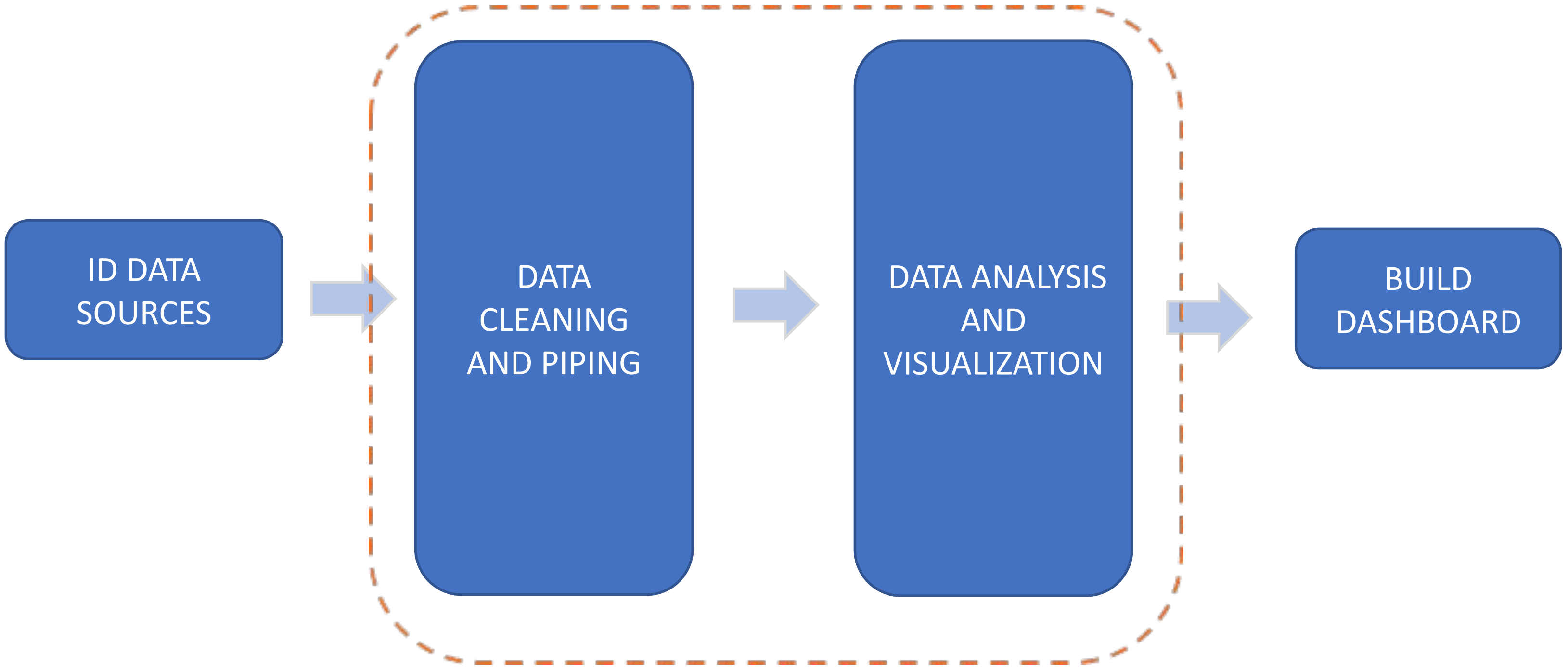


….and data visualization development….



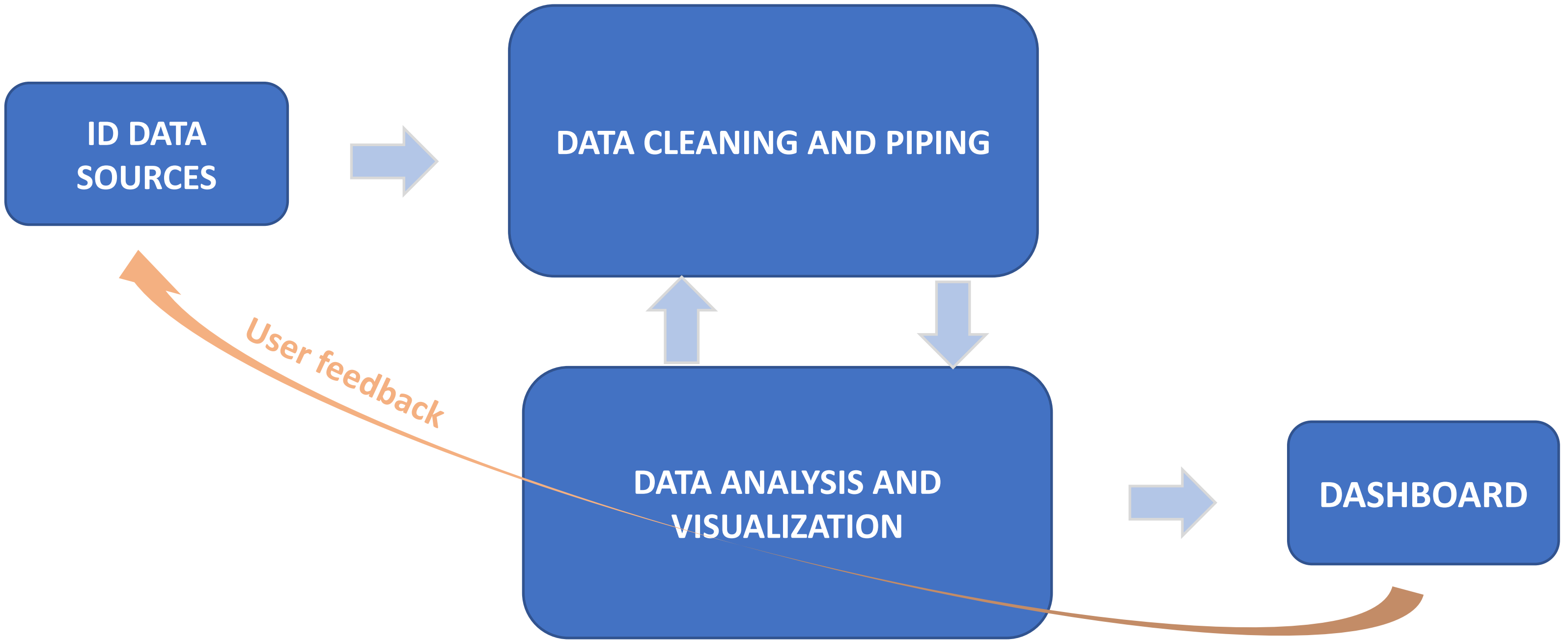
….and scale it to work effort….

Expect data wrangling = 80% of your project.

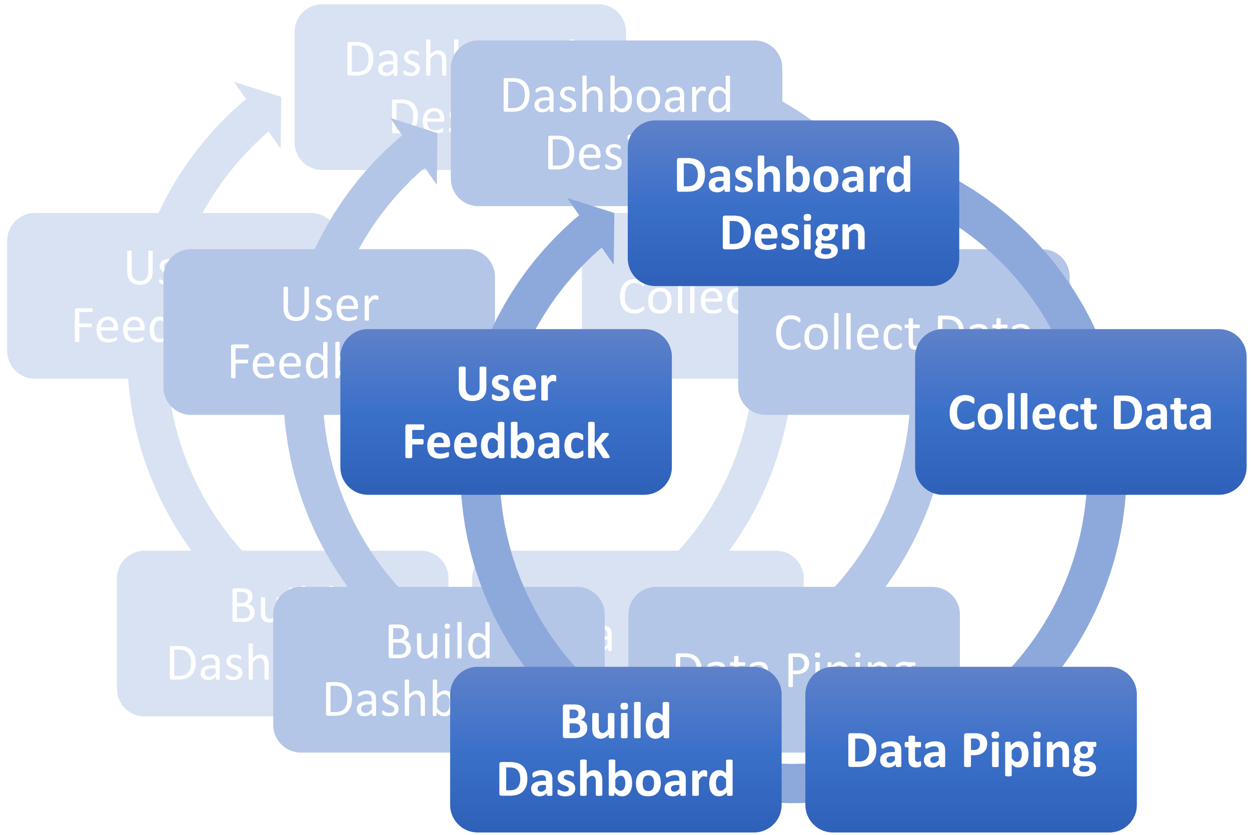


….and reflect iterative communication.

Good dashboard development is ***Agile***



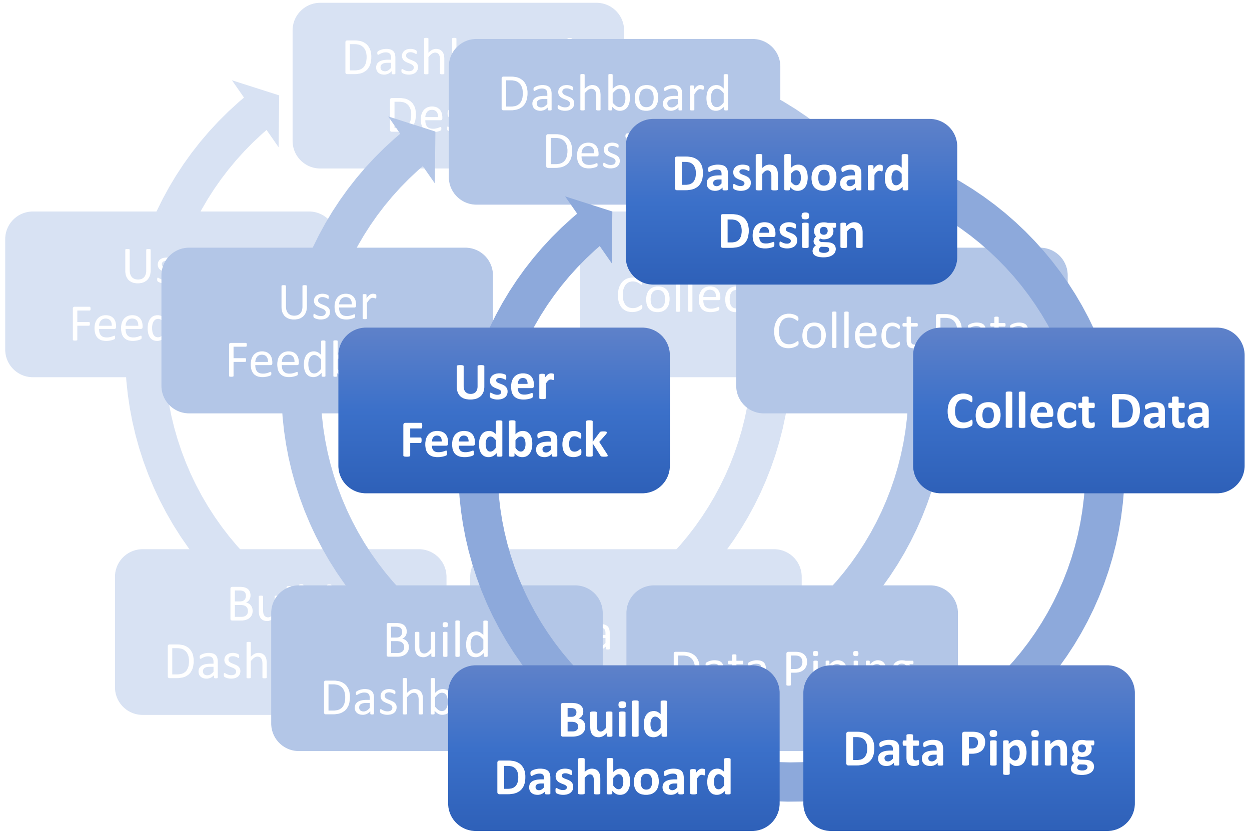
Development Cycle



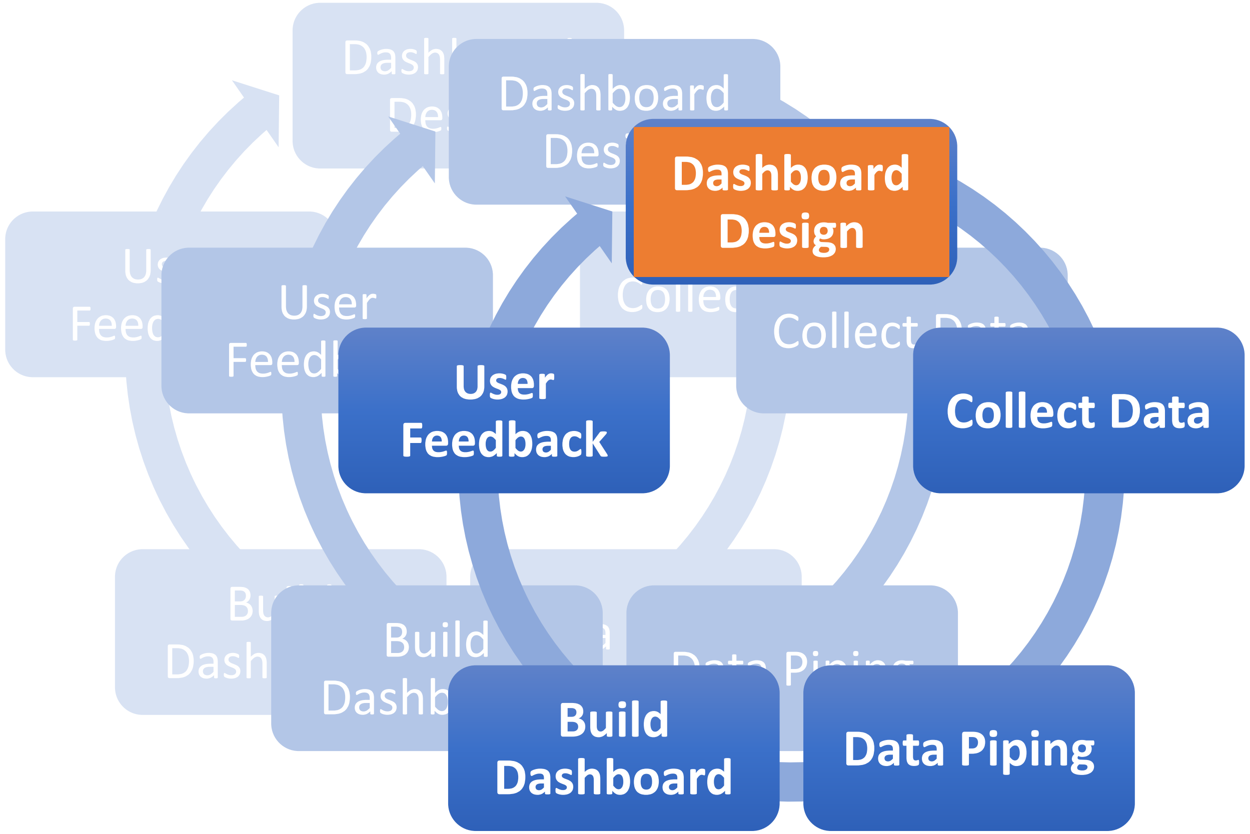
- Agile development is *iterative*.
- Quick cycling / rapid feedback.
- Keeps project from veering too far off course before a chance for correction.

Development Cycle

STEPS



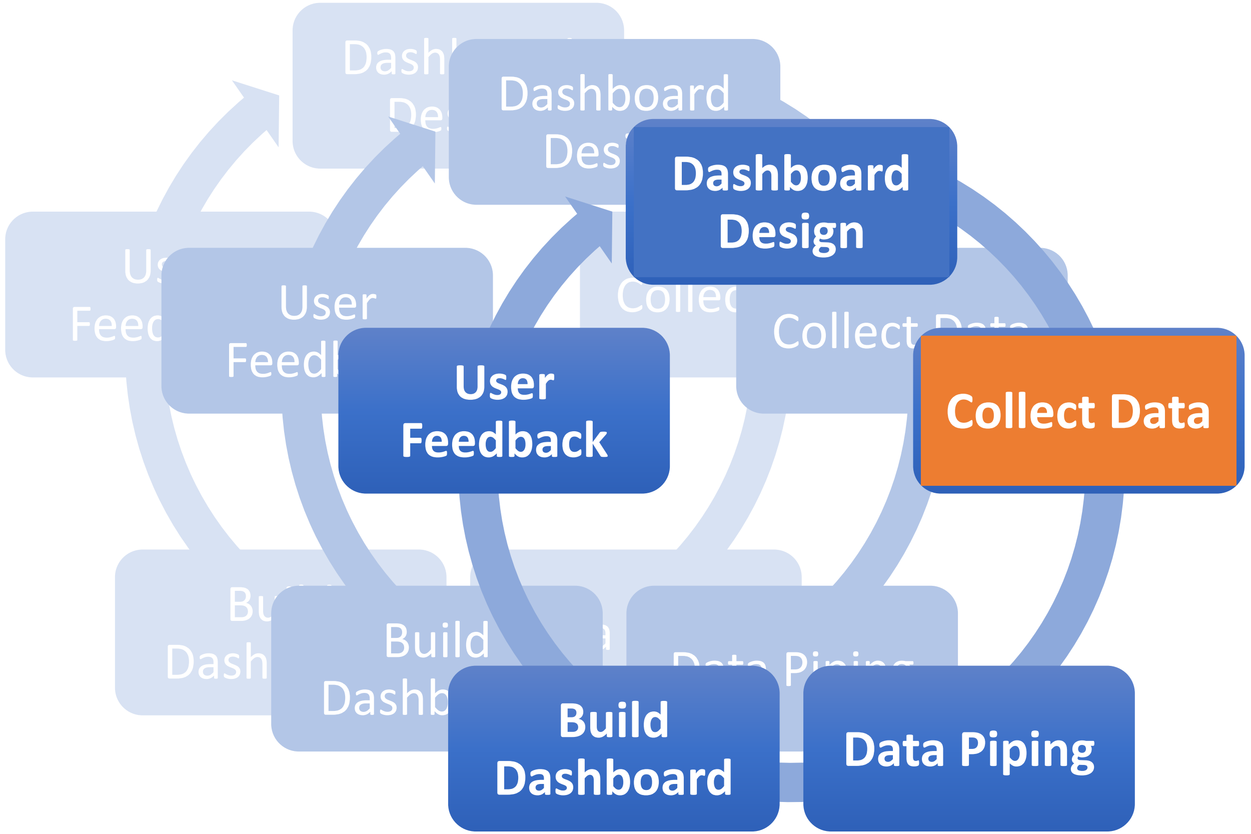
Development Cycle



STEPS

1. **Basic design / purpose**

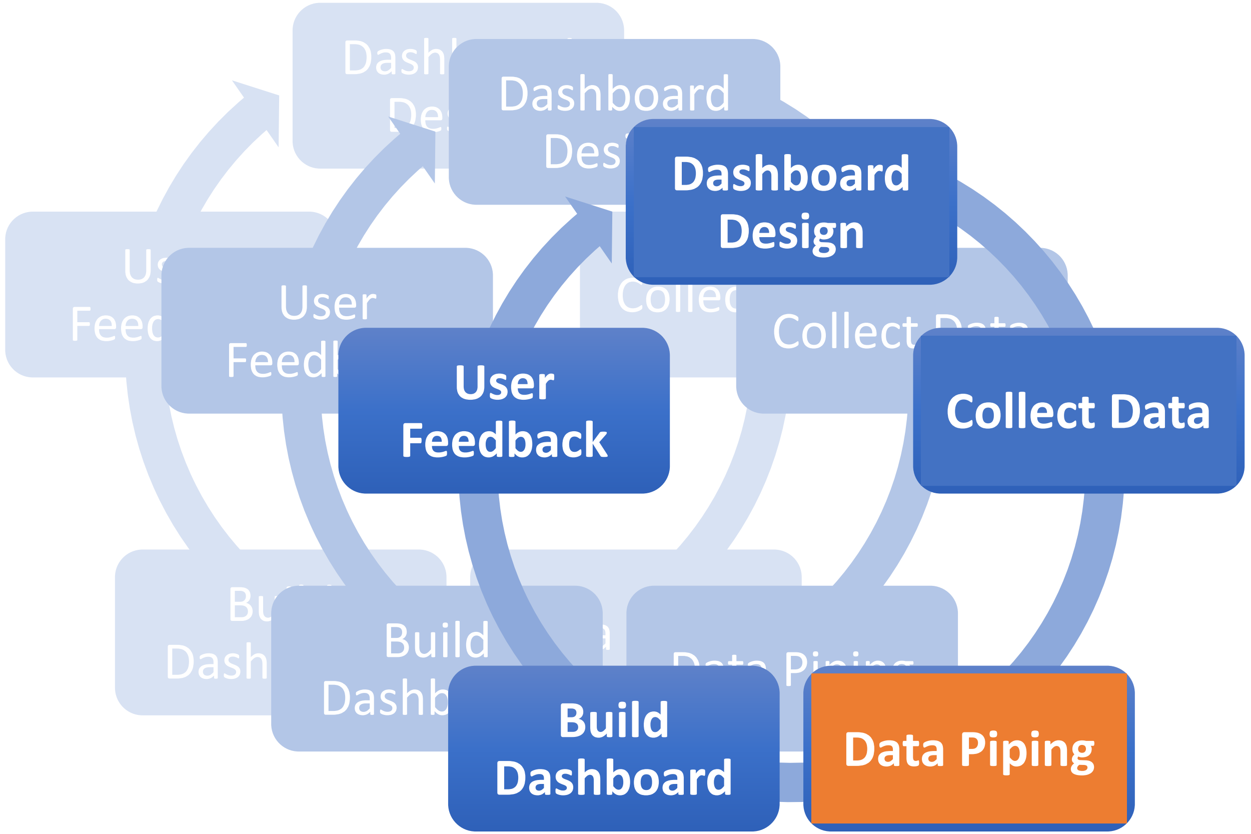
Development Cycle



STEPS

1. Basic design / purpose
2. Get your data

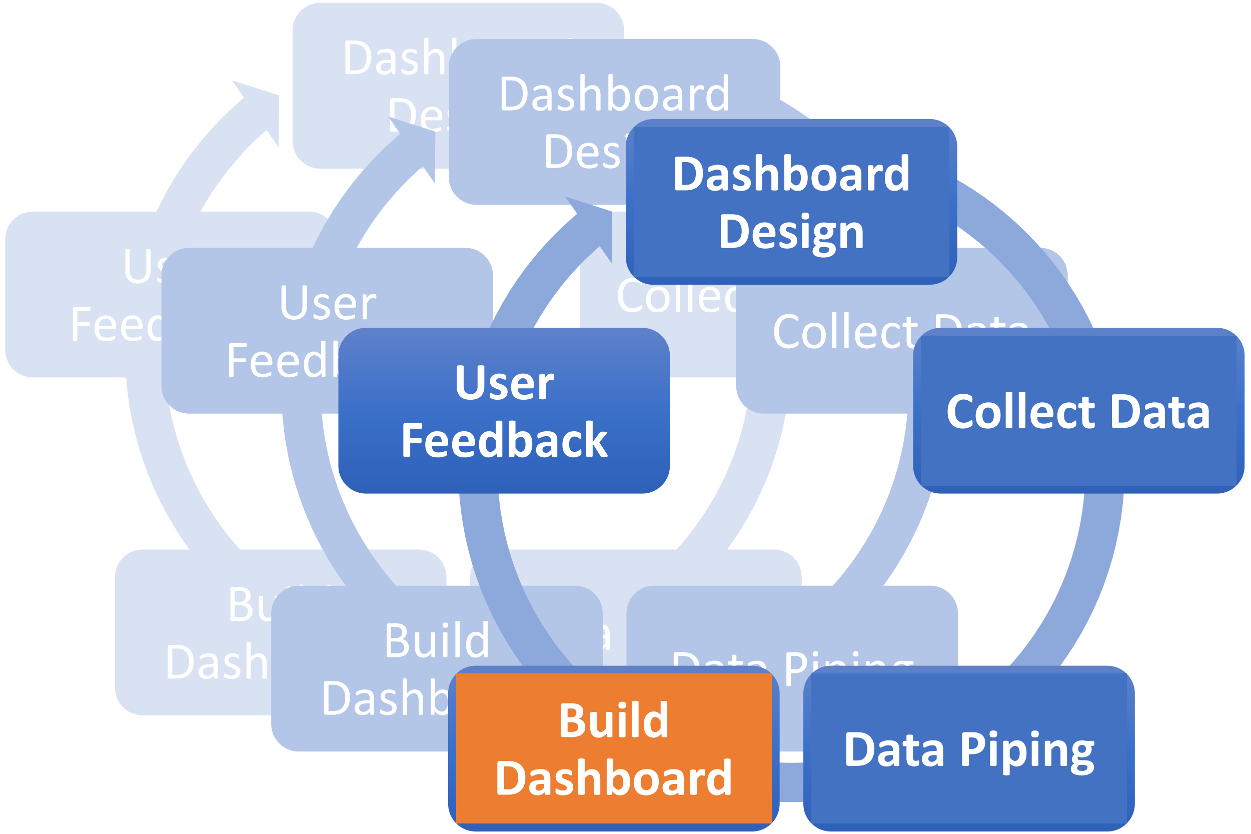
Development Cycle



STEPS

1. Basic design / purpose
2. Get your data
3. **Move, clean and shape the data.**

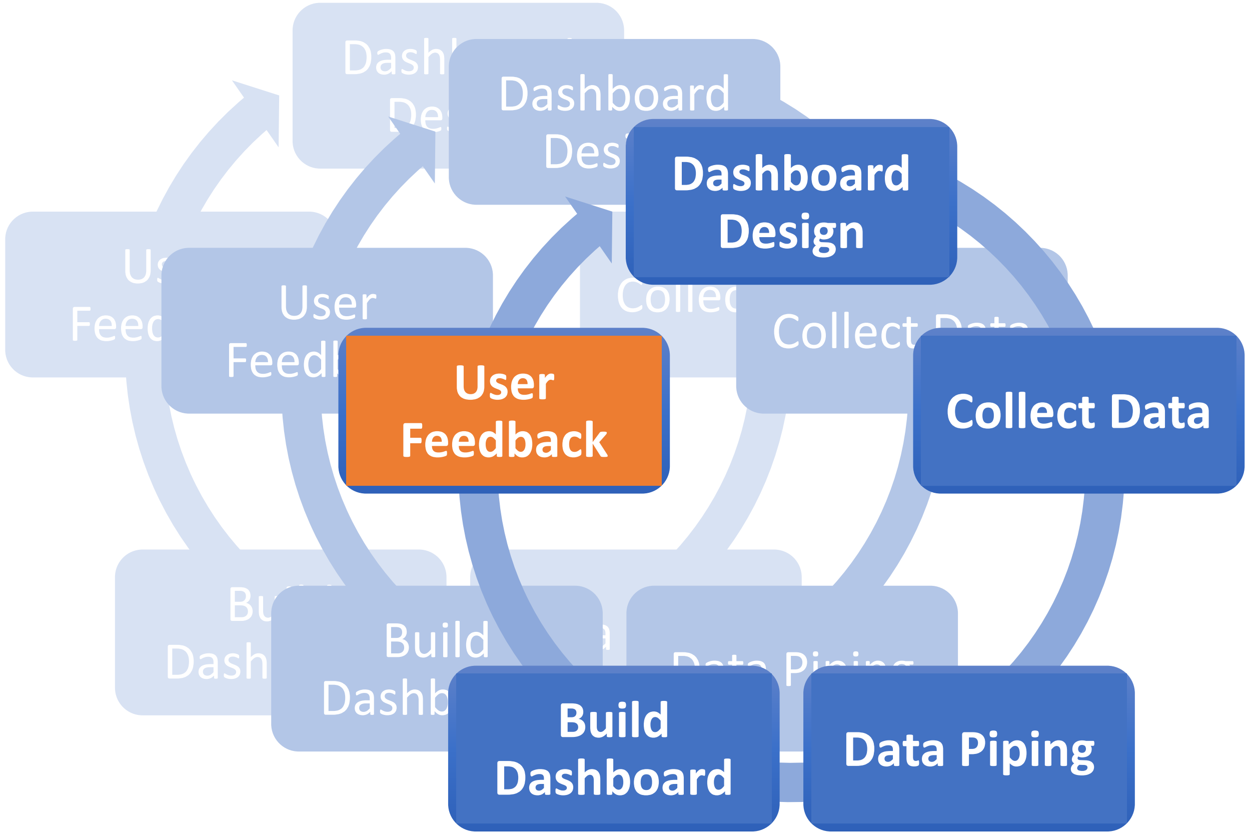
Development Cycle



STEPS

1.Basic design / purpose
2.Get your data
3.Move, clean and shape the data.
4.**Build and deploy data visualization dashboard**

Development Cycle



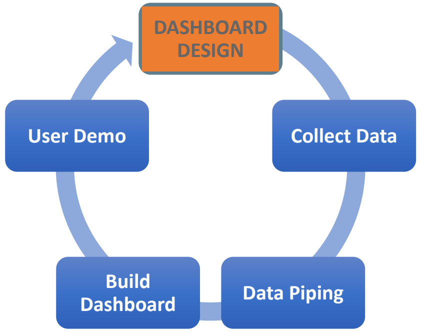
STEPS

1. Basic design / purpose
2. Get your data
3. Move, clean and shape the data.
4. Build and deploy data visualization dashboard
5. **Demo with User for feedback.**

...and Repeat

Data to Dashboards: Workflow Steps - *Detailed*

Dashboard Design

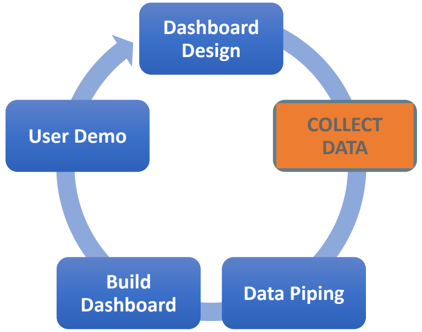


- A Dashboard frames the problem and tells the story in your data
  - Who – is the audience
  - What – information should they get from your dashboard
  - When – temporal connection between the dash and data [live vs not]
  - Where – platform, desktop vs. mobile
  - Why – goal for the whole project
- Data Visualization: the right chart for the need (see Resources)

OUTPUT FROM STEP: Site mockup

**WHITEBOARD
PEN AND PAPER**

Collect Data

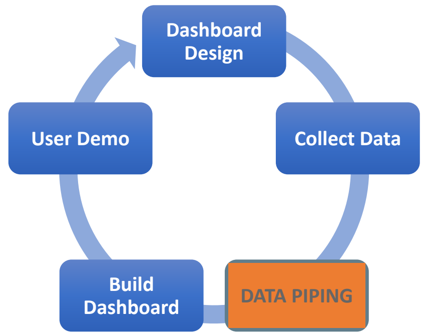


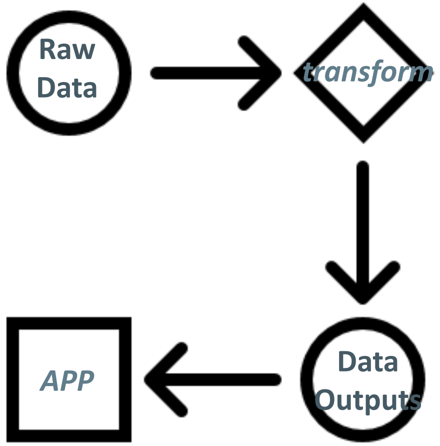
- What data do you have access to
- What questions do you *want* to ask of the data?
- What questions *can* you answer from the data you have?
- *"You can't always get what you want"*

OUTPUT FROM STEP: Documentation, data dictionaries, file directory

Github
Google Drive
Local File Directory

Data Piping



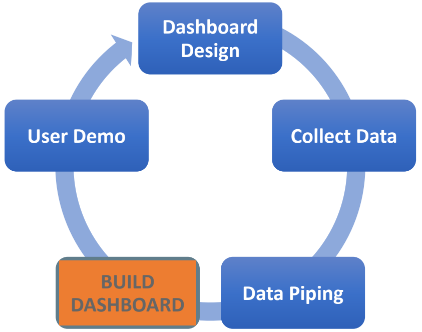


- Take raw data in
- Write scripts for necessary data transformations
- Identify data storage locations
- Handle moving data between locations
- *Consider: data that changes over time*

OUTPUT FROM STEP: scripts for transformation, output files, database connections

Github
Conda
Jupyter Notebook

Build Dashboard



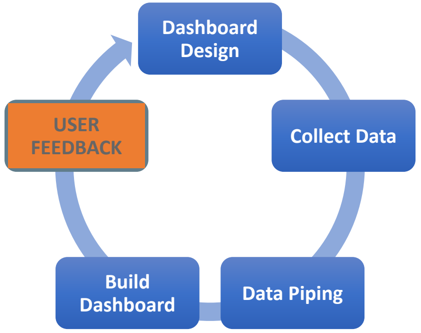
- Load outputs of data pipes into app
- Layout elements on page
- Wire-up User interactivity
  - Data filters
  - Selections
  - Changing elements

Form follows Function

OUTPUT FROM STEP: Code to build dashboard, deployed locally or to the cloud

Dash
IDE [Atom]
Github

User Feedback



- Demonstrate to Client / Users
  - Ideally deployed version that they can use
  - Try to WATCH user while they use system
  - Screenshots / PDF better than nothing
  - Collect feedback
- Integrate feedback into next iterative development cycle
- **KEY QUESTION**: DID DASHBOARD TELL THE STORY / AID THE DECISION

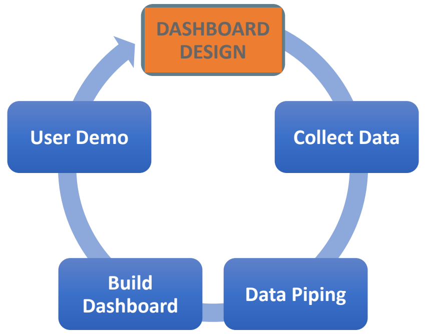
OUTPUT FROM STEP: Documented feedback to inform next iterations of design

Website [Heroku]
Powerpoint
PDF

Data to Dashboards: Dash App Example

Note: This example is written in Dash 2.0 and dash-bootstrap-components 1.0 (pre-release)

1. Dashboard Design



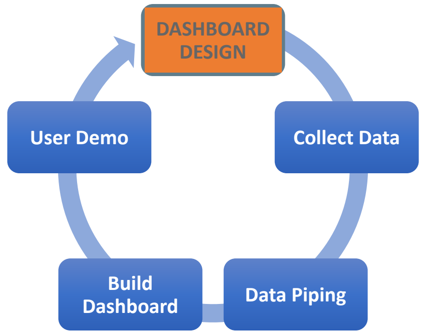
WHAT:

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

DESIRED ELEMENTS:

1. Selectable map of Congressional Districts
2. Display section for information related to selected District
3. Table of clickable links to access District Information Files

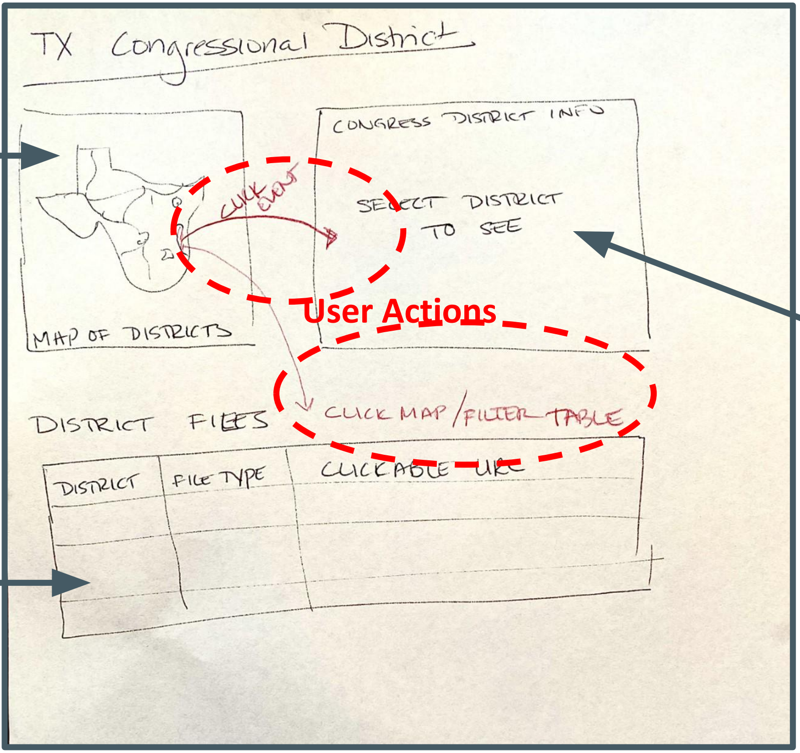
1. Dashboard Design: mock-up



1. Selectable Map

2. Table of District Files with linked URLs

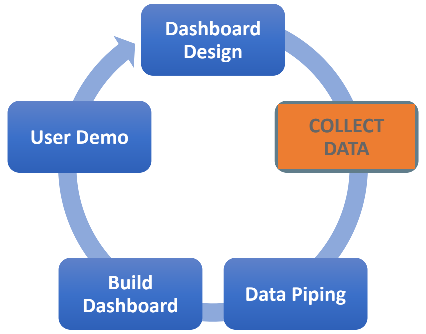
3. District Specific Information



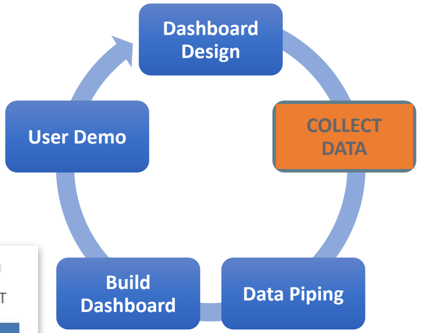
2. COLLECT DATA

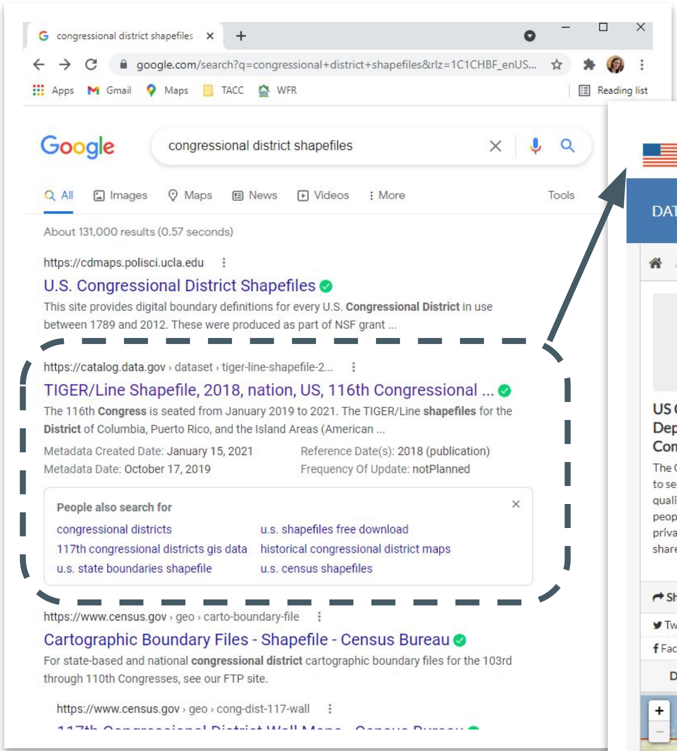
Identify online Data repositories.

- Geospatial data – Congressional District Geometries (map)
- PDF maps of individual districts
- Information on individual members
  [data source identified, not yet collected]



2. COLLECT DATA: Geospatial data





congressional district shapefiles
google.com/search?q=congressional+district+shapefiles&rlz=1C1CHBF_enUS...
Apps Gmail Maps TACC WFR Reading list
Google
All Images Maps News Videos More Tools

congressional district shapefiles

About 131,000 results (0.57 seconds)

https://cdmaps.polisci.ucla.edu

U.S. Congressional District Shapefiles

This site provides digital boundary definitions for every U.S. **Congressional District** in use between 1789 and 2012. These were produced as part of NSF grant ...

https://catalog.data.gov › dataset › tiger-line-shapefile-2...

TIGER/Line Shapefile, 2018, nation, US, 116th Congressional ...

The 116th **Congress** is seated from January 2019 to 2021. The TIGER/Line **shapefiles** for the **District** of Columbia, Puerto Rico, and the Island Areas (American ...

Metadata Created Date: January 15, 2021 Reference Date(s): 2018 (publication)
Metadata Date: October 17, 2019 Frequency Of Update: notPlanned

People also search for

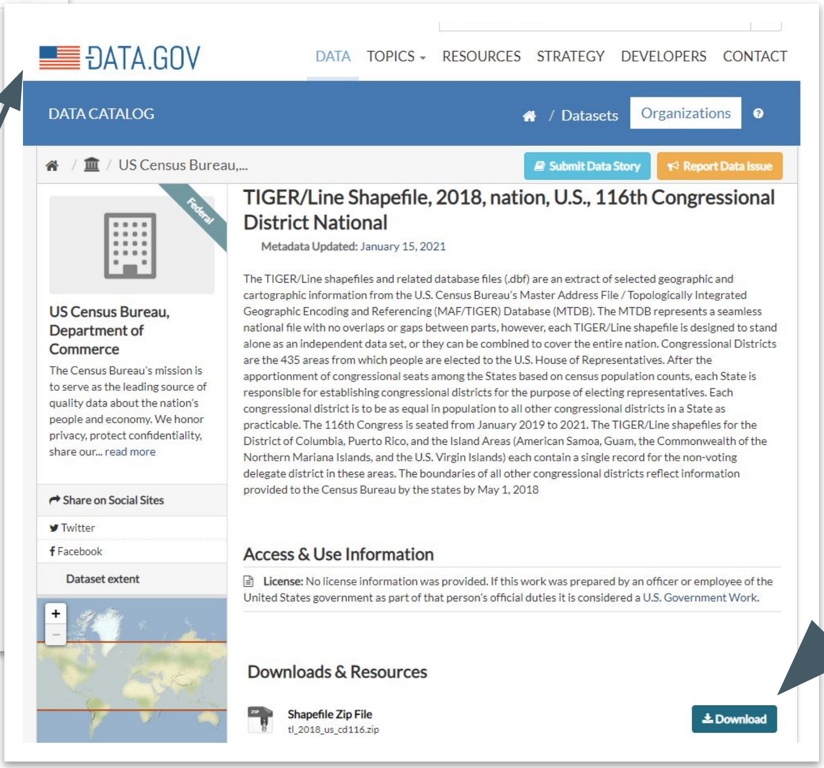
congressional districts	u.s. shapefiles free download
117th congressional districts gis data	historical congressional district maps
u.s. state boundaries shapefile	u.s. census shapefiles

https://www.census.gov › geo › carto-boundary-file

Cartographic Boundary Files - Shapefile - Census Bureau

For state-based and national **congressional district** cartographic boundary files for the 103rd through 110th Congresses, see our FTP site.

https://www.census.gov › geo › cong-dist-117-wall



TIGER/Line Shapefile, 2018, nation, U.S., 116th Congressional District National

Metadata Updated: January 15, 2021

US Census Bureau, Department of Commerce

The Census Bureau's mission is to serve as the leading source of quality data about the nation's people and economy. We honor privacy, protect confidentiality, share our... read more

The TIGER/Line shapefiles and related database files (.dbf) are an extract of selected geographic and cartographic information from the U.S. Census Bureau's Master Address File / Topologically Integrated Geographic Encoding and Referencing (MAF/TIGER) Database (MTDB). The MTDB represents a seamless national file with no overlaps or gaps between parts, however, each TIGER/Line shapefile is designed to stand alone as an independent data set, or they can be combined to cover the entire nation. Congressional Districts are the 435 areas from which people are elected to the U.S. House of Representatives. After the apportionment of congressional seats among the States based on census population counts, each State is responsible for establishing congressional districts for the purpose of electing representatives. Each congressional district is to be as equal in population to all other congressional districts in a State as practicable. The 116th Congress is seated from January 2019 to 2021. The TIGER/Line shapefiles for the District of Columbia, Puerto Rico, and the Island Areas (American Samoa, Guam, the Commonwealth of the Northern Mariana Islands, and the U.S. Virgin Islands) each contain a single record for the non-voting delegate district in these areas. The boundaries of all other congressional districts reflect information provided to the Census Bureau by the states by May 1, 2018

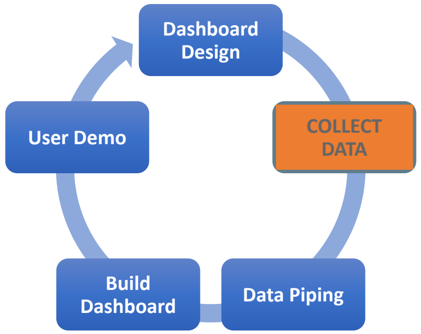
Access & Use Information

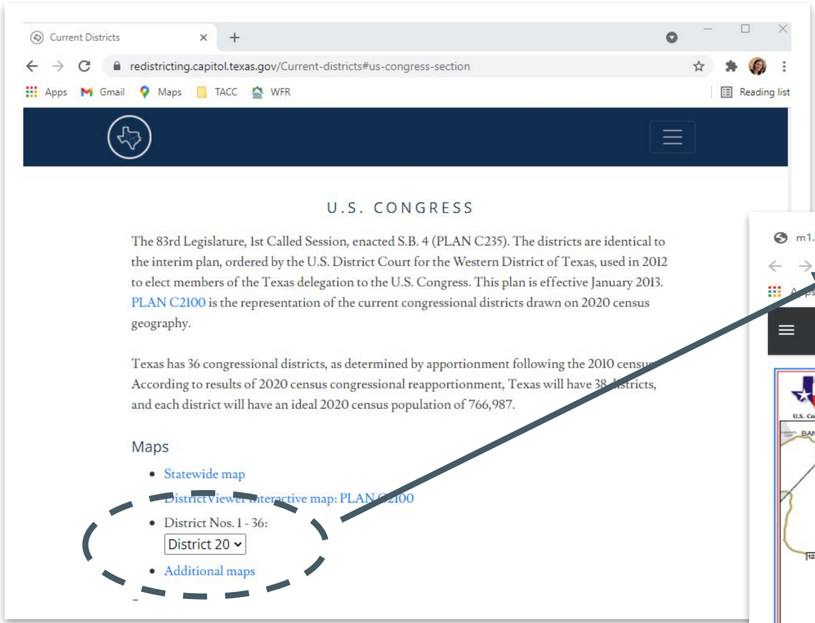
License: No license information was provided. If this work was prepared by an officer or employee of the United States government as part of that person's official duties it is considered a U.S. Government Work.

Downloads & Resources

Shapefile Zip File
tl_2018_us_cd116.zip

2. COLLECT DATA: District files




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

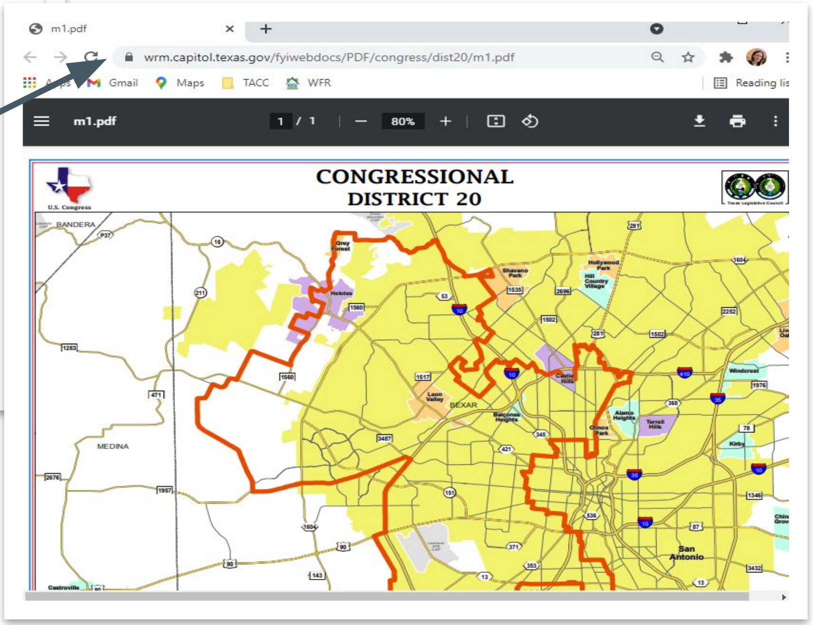
U.S. CONGRESS

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.

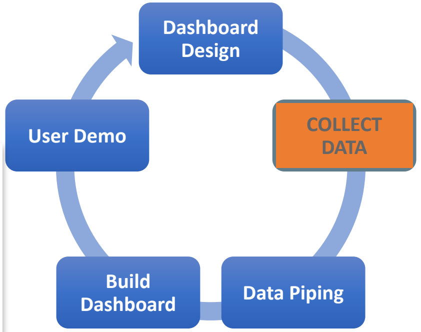
Maps

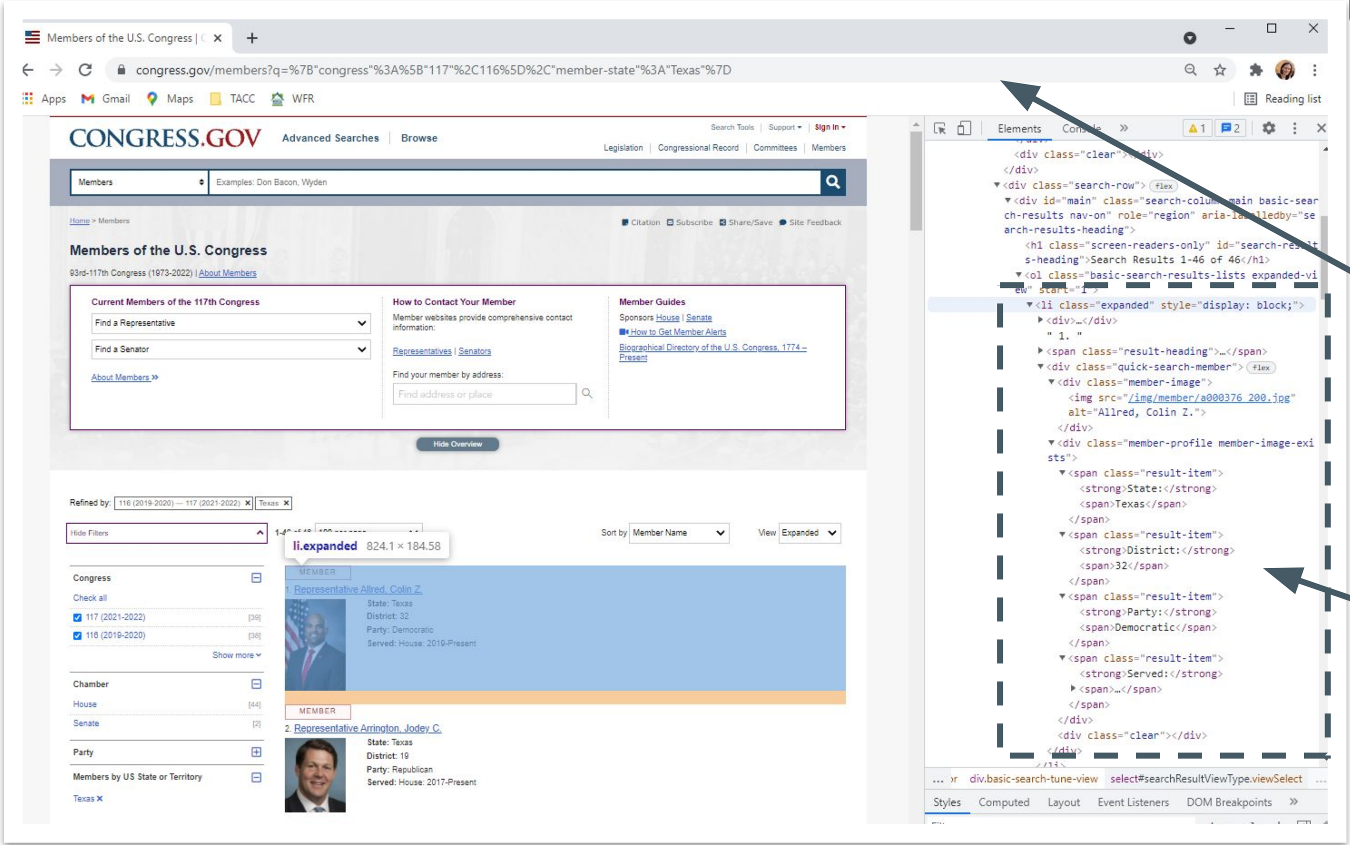
- Statewide map
- DistrictViewer interactive map: PLAN C2100
- District Nos. 1 - 36: District 20
- Additional maps


m1.pdf
wrm.capitol.texas.gov/fyiwebdocs/PDF/congress/dist20/m1.pdf
Gmail
Maps
TACC
WFR
1 / 1
80%
U.S. Congress
CONGRESSIONAL DISTRICT 20
BANDERA
MEDINA
BEXAR
San Antonio
Castroville

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

2. COLLECT DATA: member info *(to do)*





Members of the U.S. Congress

93rd-117th Congress (1973-2022) | About Members

Current Members of the 117th Congress

Find a Representative

Find a Senator

About Members »

How to Contact Your Member

Member websites provide comprehensive contact information:

Representatives | Senators

Find your member by address:

Find address or place

Member Guides

Sponsors House | Senate

How to Get Member Alerts

Biographical Directory of the U.S. Congress, 1774 – Present

Refined by: 116 (2019-2020) — 117 (2021-2022) ✕ Texas ✕

Congress

Check all

☑ 117 (2021-2022) [39]

☑ 116 (2019-2020) [38]

Show more

Chamber

House [44]

Senate [2]

Party

Members by US State or Territory

Texas ✕

MEMBER

1. Representative Allred, Colin Z.

State: Texas
District: 32
Party: Democratic
Served: House: 2019-Present

MEMBER

2. Representative Arrington, Jodey C.

State: Texas
District: 19
Party: Republican
Served: House: 2017-Present

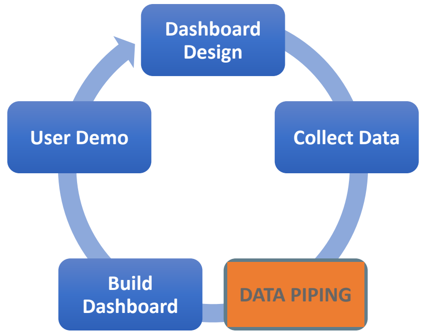
```
  <div class="clear"></div>
</div>
▼<div class="search-row"> flex
 ▼<div id="main" class="search-column main basic-sear
 ch-results nav-on" role="region" aria-labelledby="se
 arch-results-heading">
    <h1 class="screen-readers-only" id="search-result
    s-heading">Search Results 1-46 of 46</h1>
  ▼<ol class="basic-search-results-lists expanded-vi
   ew" start="1">
    ▼<li class="expanded" style="display: block;">
      ▶<div>…</div>
       " 1. "
      ▶<span class="result-heading">…</span>
      ▼<div class="quick-search-member"> flex
        ▼<div class="member-image">
           <img src="/img/member/a000376_200.jpg"
           alt="Allred, Colin Z.">
         </div>
        ▼<div class="member-profile member-image-exi
        sts">
          ▼<span class="result-item">
             <strong>State:</strong>
             <span>Texas</span>
           </span>
          ▼<span class="result-item">
             <strong>District:</strong>
             <span>32</span>
           </span>
          ▼<span class="result-item">
             <strong>Party:</strong>
             <span>Democratic</span>
           </span>
          ▼<span class="result-item">
             <strong>Served:</strong>
            ▶<span>…</span>
           </span>
         </div>
         <div class="clear"></div>
       </div>
     </li>
```

… div.basic-search-tune-view select#searchResultViewType.viewSelect …

House members website

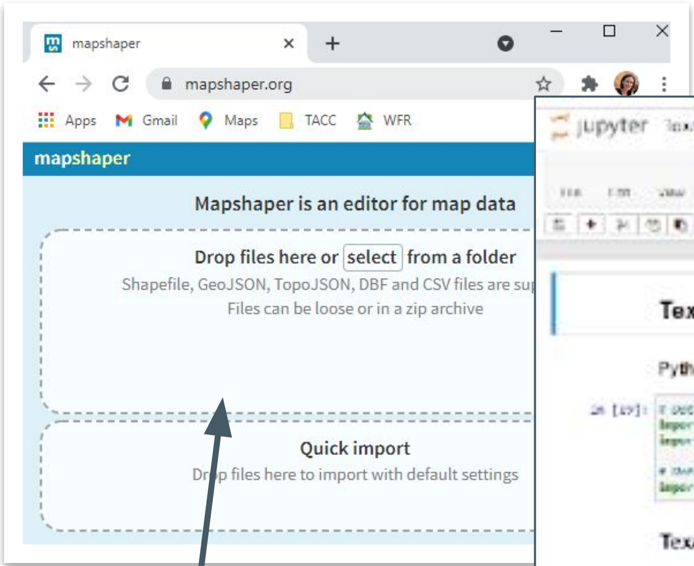
Examine content with developer tools to determine how to scrape site

3. DATA PIPING



- For data that needs transformation: move data where needed, transform, save new data output
- Load data files in local data directory
- Jupyter notebook to perform data transformations and develop data visualizations

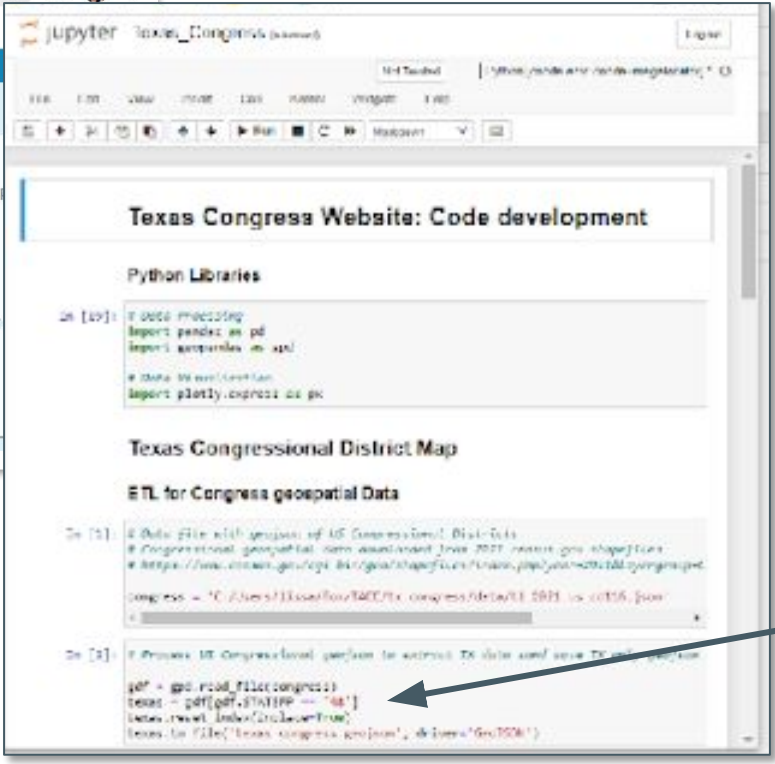
3. DATA PIPING: geospatial



Mapshaper is an editor for map data

Drop files here or select from a folder
Shapefile, GeoJSON, TopoJSON, DBF and CSV files are su
Files can be loose or in a zip archive

Quick import
Drop files here to import with default settings



Texas Congress Website: Code development

Python Libraries

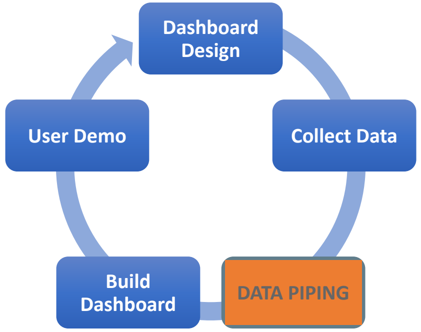
```
In [15]: # Data Processing
         import pandas as pd
         import geopandas as gpd

         # 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=[illegible]

         congress = 'C:/Users/[illegible]/TACC/tx congress/data/[illegible].json'
```

```
In [3]:  # Process US Congressional geojson to extract TX data and 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')
```

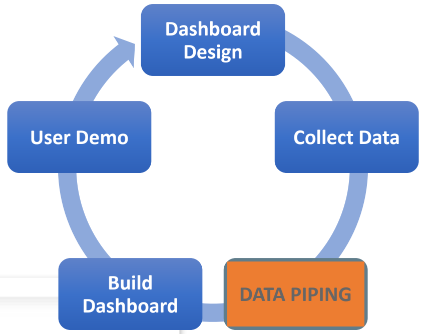


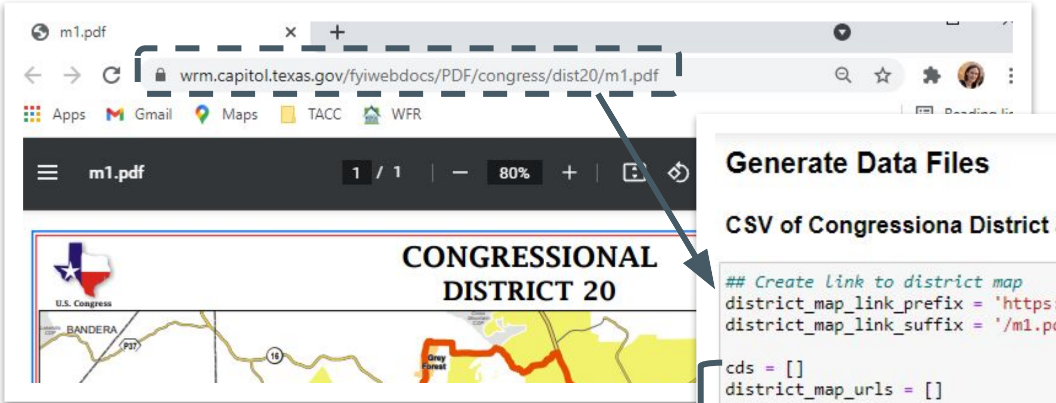
Import US district shapefile from Census to Mapshaper.org □ geojson

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

Jupyter Notebook file available in assets folder of Github repo

3. DATA PIPING: district maps file



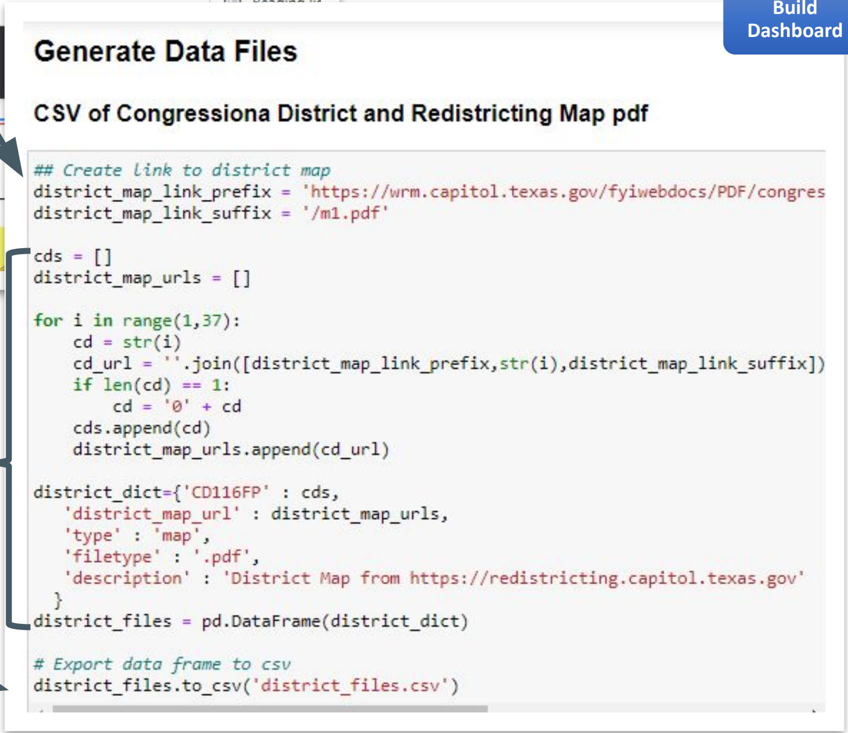


m1.pdf
wrm.capitol.texas.gov/fyiwebdocs/PDF/congress/dist20/m1.pdf
Apps
Gmail
Maps
TACC
WFR
m1.pdf
1 / 1
80%
CONGRESSIONAL DISTRICT 20
U.S. Congress
BANDERA

1. Determine pattern to generate map url for each district

2. Generate dataframe of congressional district and links to file urls.

3. Export dataframe to csv file

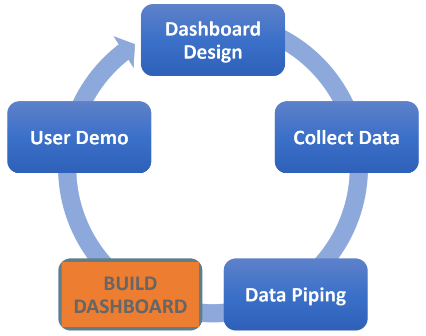


Generate Data Files

CSV of Congressiona District and Redistricting Map pdf

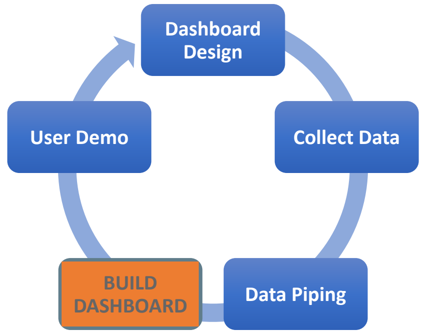
```
## Create link to district map
district_map_link_prefix = 'https://wrm.capitol.texas.gov/fyiwebdocs/PDF/congres
district_map_link_suffix = '/m1.pdf'

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')
```

4. BUILD DASHBOARD



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

4. BUILD DASHBOARD



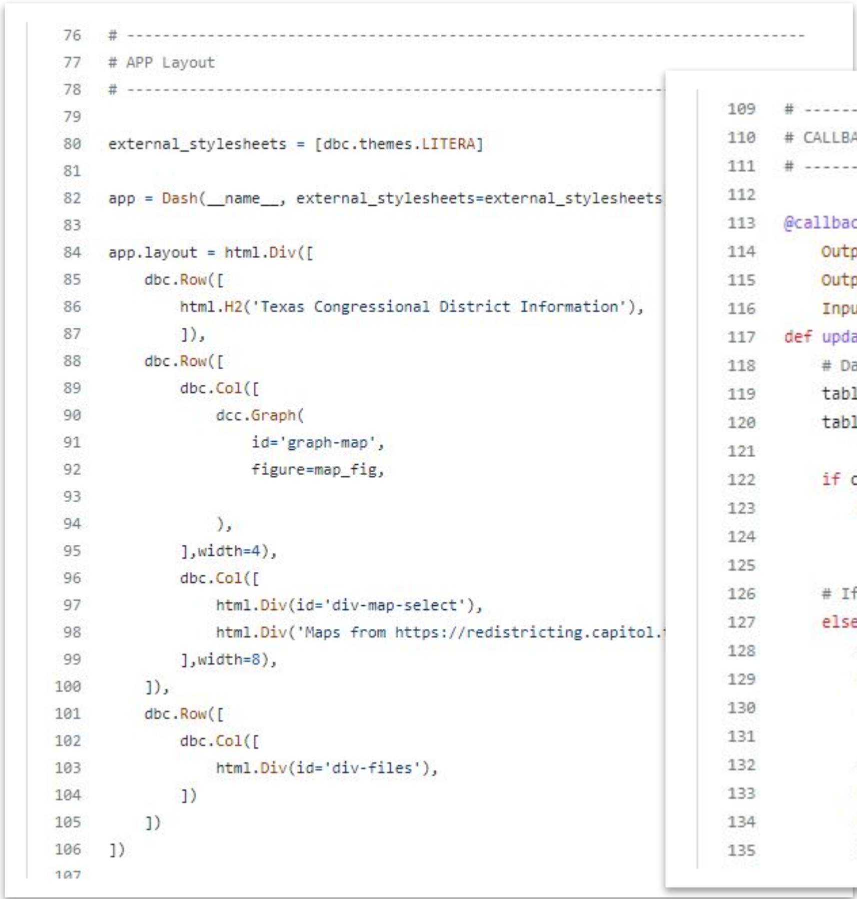
Develop Locally

1. Clone Github Repo [mepearson/texas_congress]
2. Create and launch virtual environment
3. Develop changes locally using IDE of choice [atom, vstudio, etc.]
4. Command Line: > python app.py

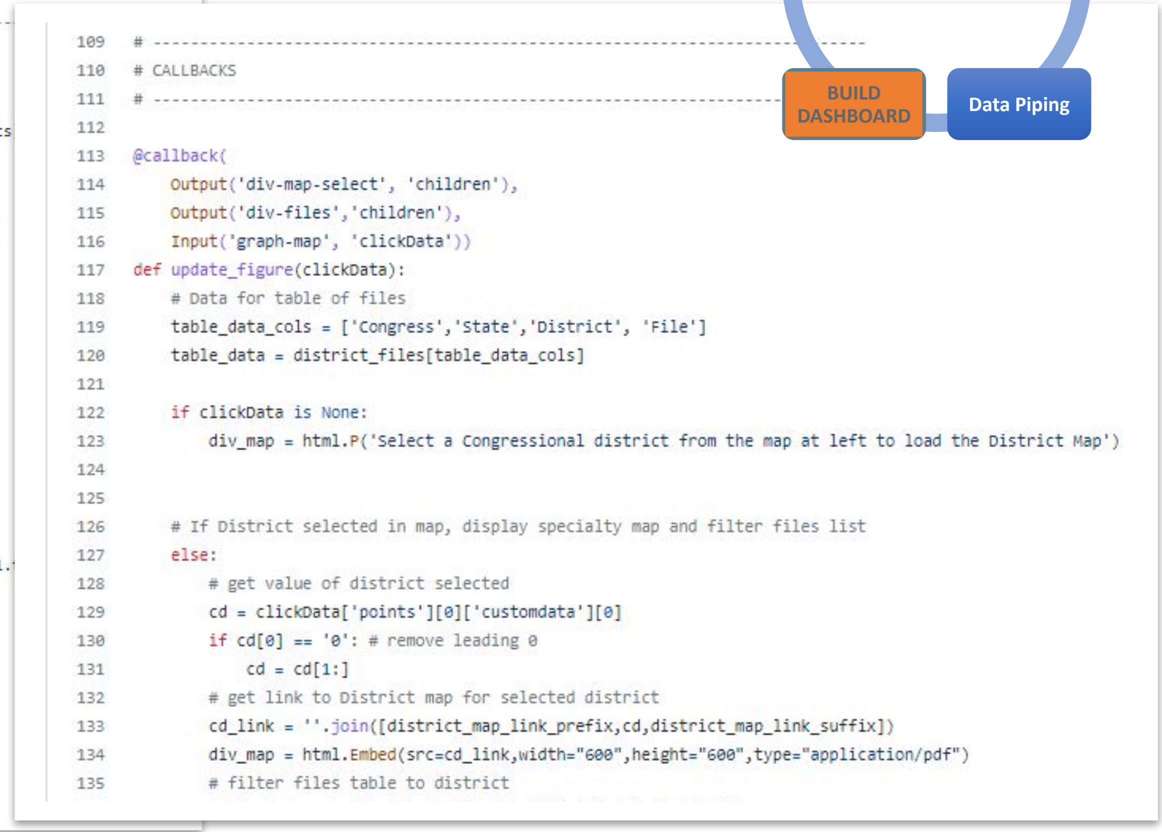
Deploy to Heroku

1. Push changes to Github
2. Create site on Heroku linked to Github repo
3. Manually Deploy

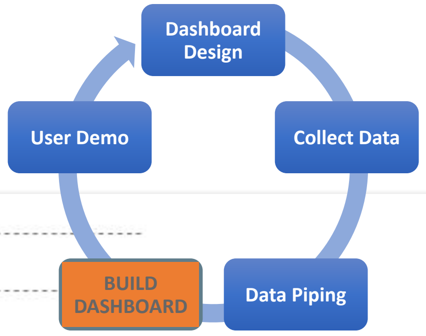
4. BUILD DASHBOARD: layout & callbacks



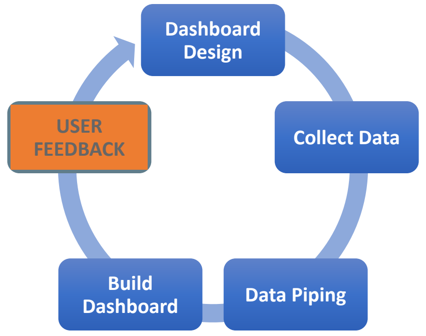
```
# ---------------------------------------------------------------------------
# APP Layout
# ------------------------------------------------------------------

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.
        ],width=8),
    ]),
    dbc.Row([
        dbc.Col([
            html.Div(id='div-files'),
        ])
    ])
])
```



```
# ------------------------------------------------------------------------
# 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
```



5. USER FEEDBACK



- Formal vs. Informal
- Decide how will collect feedback
- Set a process to track and prioritize feedback [spreadsheet, github issues, etc.]
- Use this to assign tasks
- Product Management is key. Tasks should have clear decision makers / authority for action.

Data to Dashboards: DATA PIPING

Data Cleaning and Wrangling

- "You can't always get what you need"
- Data arrives in many – usually Messy! - forms
- Data dashboards require standardized data
- Clean and wrangle what you get --> what you need
- What you need:
  - Quality controlled data…
  - Where you can access it..
  - In the right structure.

Garbage in --> Garbage out

Quality Control Data

- If collecting your own data: make sure it's good.
- Pre-existing data:
  - review for quality [assess provider, etc.]
  - handle missing values
- Cite your data. Track sources and date acquired.

Move and Shape your data [ETL]

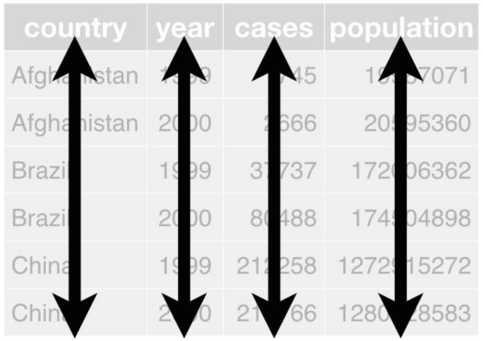
- ETL has specific meaning, but abstract concept useful here
- ETL = Extract, Transform, Load
- Extract data from data sources [websites, databases, files]
- Transform data into the necessary format
  - Clean – handle null values and bad data
  - Shape – wrangle data into required structure
- Load data into intermediate repository [database, file] or directly into dashboard

Shaping data: make it 'Tidy'

- Everything assumes clean structured data, usually in 'Tidy' format
- Concepts language agnostic – learn on what you know best
- See ‘Resources’ slides for in depth discussion and references

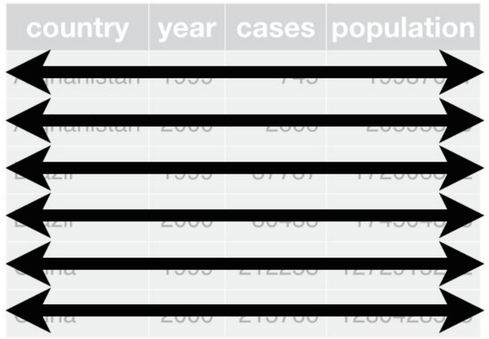
TIDY Data: what it is

Each variable is a column.



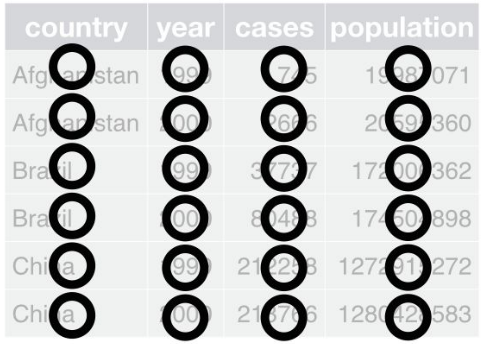
variables

Each observation is a row.



observations

Each value is a cell.



values

Shaping data: how to make it 'Tidy'

How to Tidy data: tools available

- R: Tidyverse packages
- Python: pandas and associated packages
- Excel: pivot tables and other built in functions to add text

Data to Dashboards: Resources

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

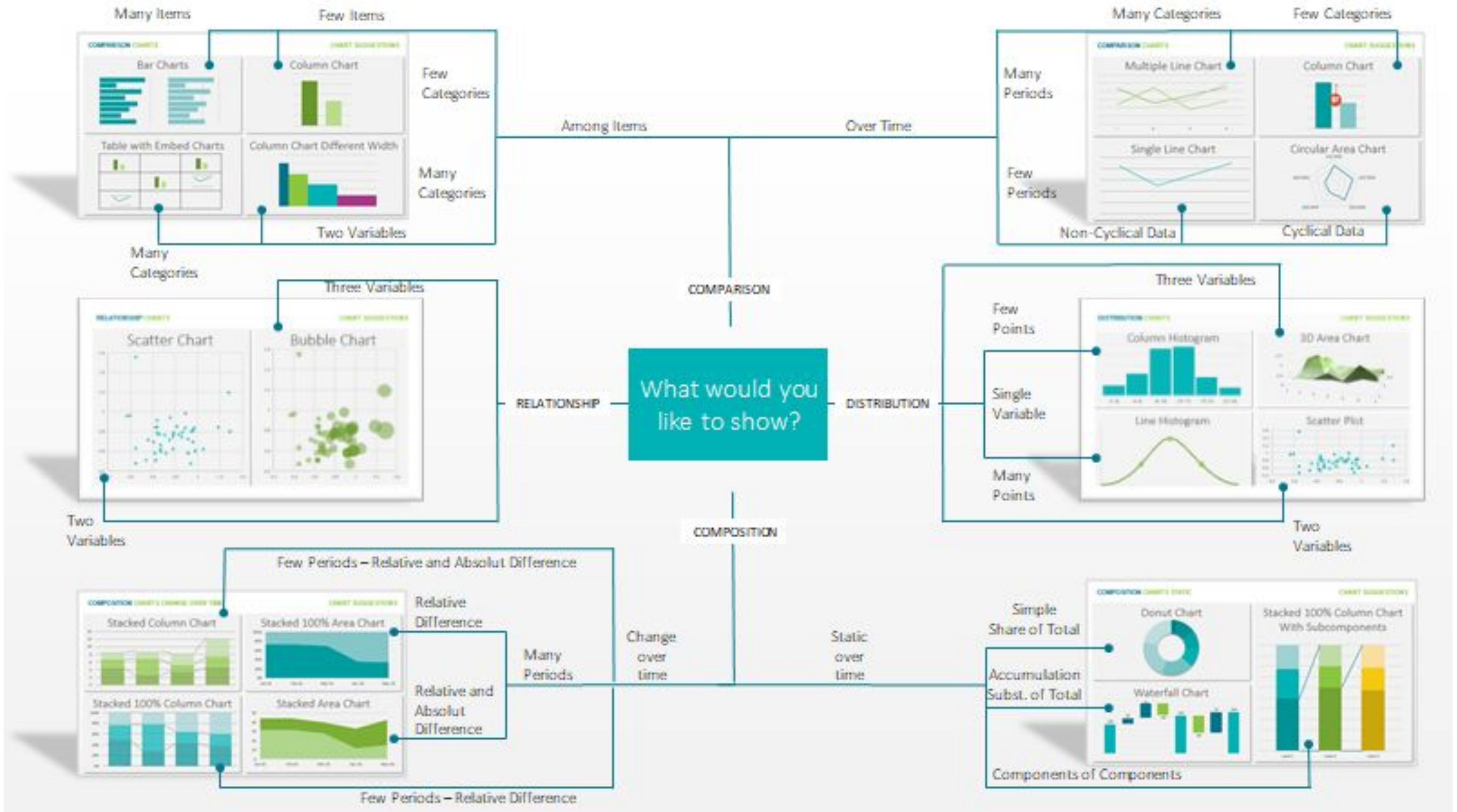
- Chart Chooser — Juice Analytics - https://www.juiceanalytics.com/chartchooser
- Plotly graphing library - https://plotly.com/python/

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

A.Abela chart from slidemodal.com template

Dashboard Design: Chart Selection



Questions and Concerns

Next Training Sessions:

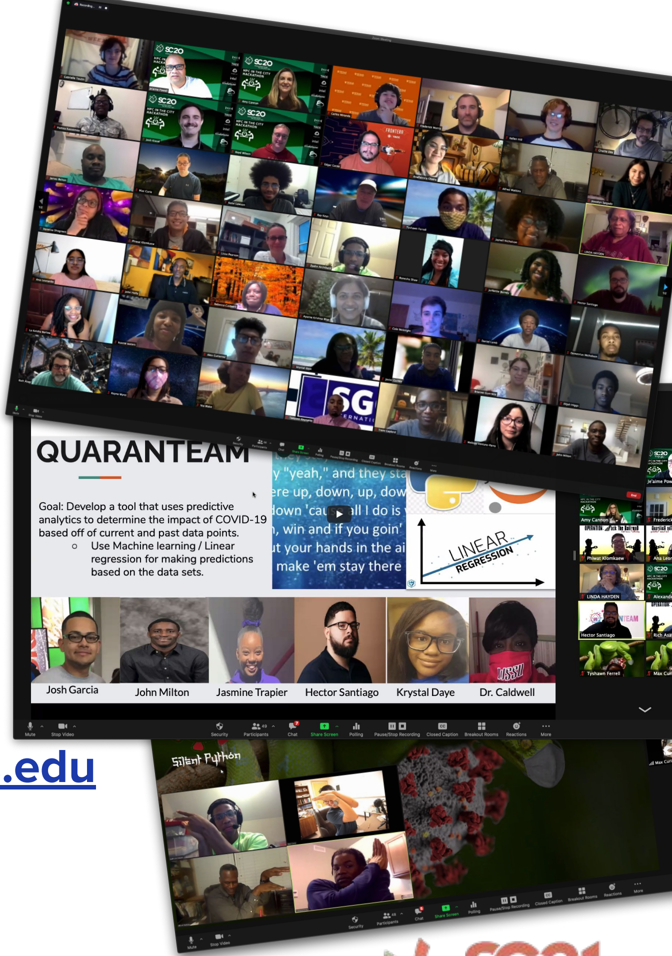
- Beginning to End Project Example - *[10/28/21]*

- **Kick-Off** - *[11/4/21]*

Schedule:
https://jeaimehp.github.io/HackHPC-HPCintheCity21/

Presenter Contact Information:

Melissa Pearson (TACC) - Lissa.Pearson@austin.utexas.edu


QUARANTEAM
Goal: Develop a tool that uses predictive analytics to determine the impact of COVID-19 based off of current and past data points.
Use Machine learning / Linear regression for making predictions based on the data sets.
LINEAR REGRESSION
Josh Garcia
John Milton
Jasmine Trapier
Hector Santiago
Krystal Daye
Dr. Caldwell
Silent Python

HPC in the City: St. Louis HACKATHON http://hackhpc.org/hpcinthecity SC21 St. Louis, MO | science & beyond.