Challenges to collaboration and reproducibility

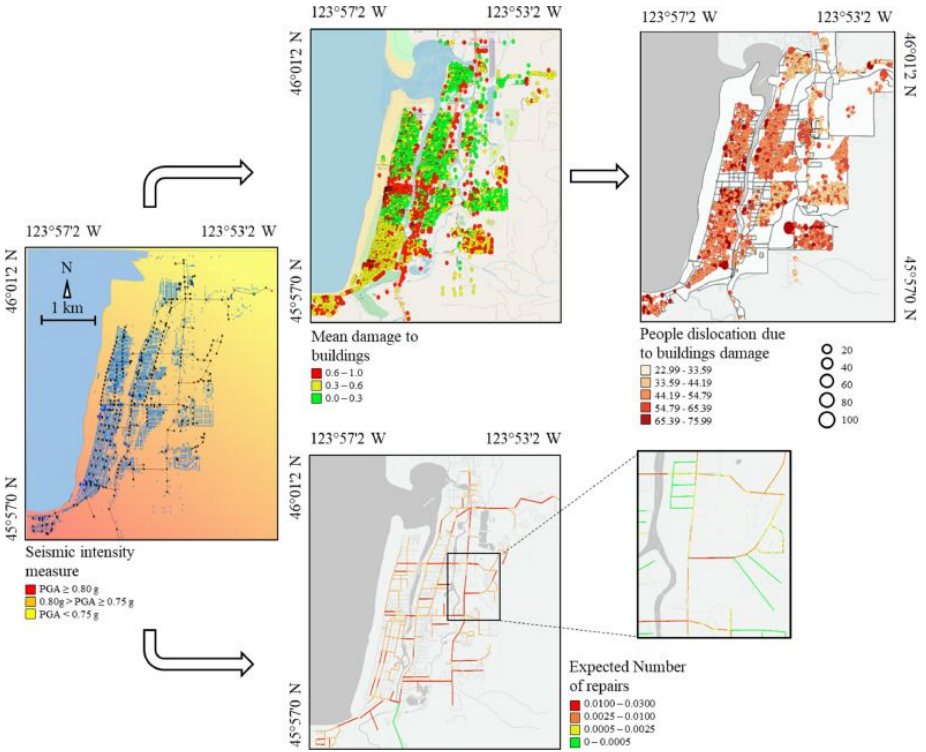
Nathanael Rosenheim, PhD

URSC 689
Feb 1, 2021

Challenges Collaborating Across Universities

Community resilience model that combines work completed by Urban Planning researchers at TAMU and Civil Engineers at the University of Illinois -Urbana-Champaign.

All data shared via email. Models were run independently without shared code or version control.



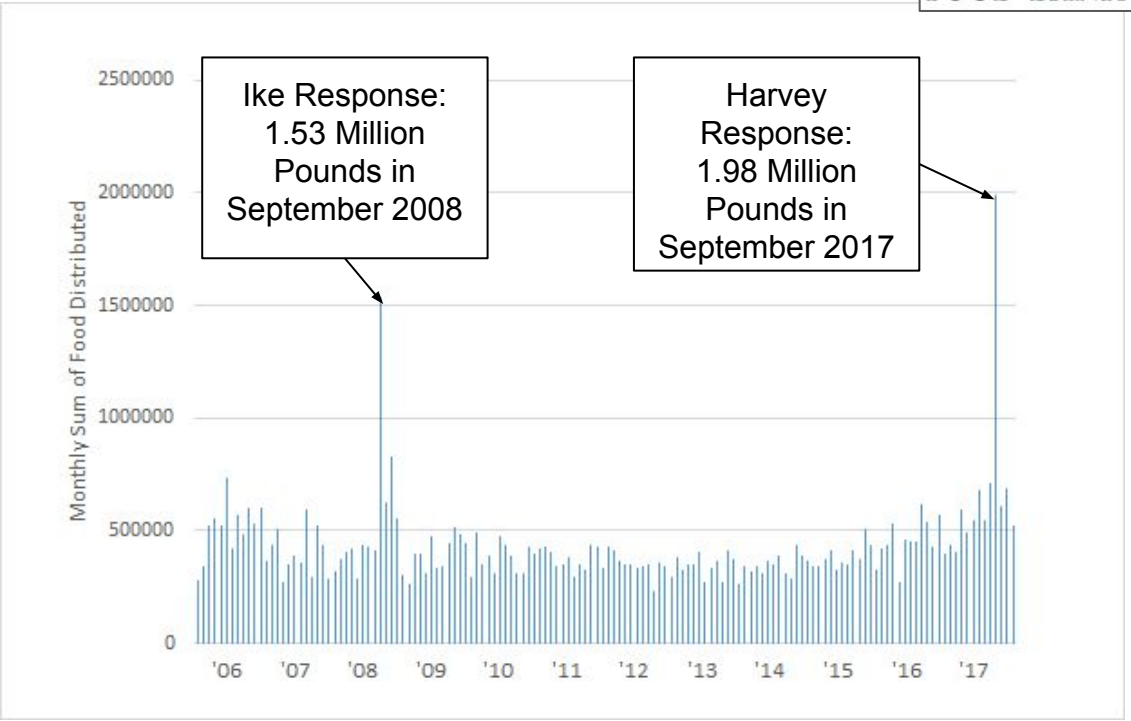
Guidotti, R., Gardoni, P., & Rosenheim, N. (2019). Integration of physical infrastructure and social systems in communities' reliability and resilience analysis. *Reliability Engineering & System Safety*. doi.org/10.1016/j.ress.2019.01.008

Challenges Collaborating with Communities

Models of post disaster food aid distributions.

Community partner had a massive SQL database, but only way to access information was with a point-and-click user interface.

Took graduate student 40 hours to download data.



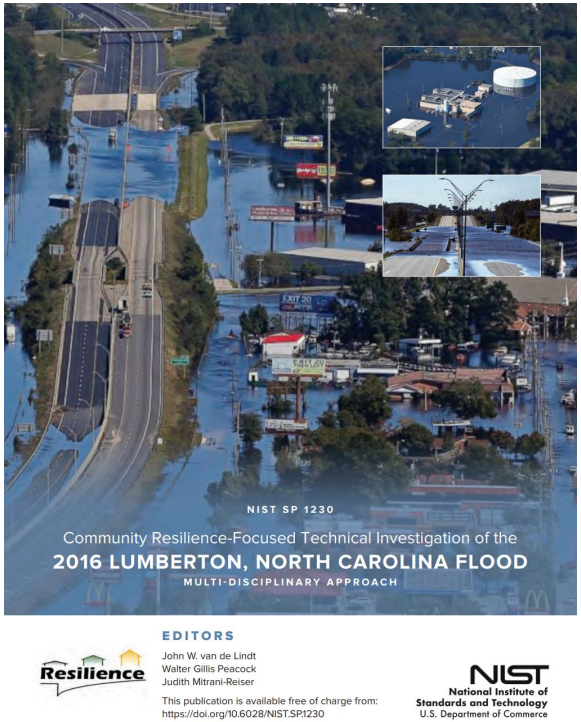
Source Data: Southeast Texas Food Bank Primarius Reports

Challenges Collaborating with Federal Agencies

Post disaster field studies that combine engineering damage assessments with social science household and business level surveys.

Multi-year collaboration between 11 universities through the National Institute for Standards and Technology (NIST).

Diverse range of data collection, required Institutional Review Board (IRB) approval. Requires storage of confidential data with limited means of sharing and citing data.



van de Lindt, J. et al (2018). *The Lumberton, North Carolina Flood of 2016: A Community Resilience Focused Technical Investigation* (No. Special Publication (NIST SP)-1230). https://doi.org/10.6028/NIST.SP.1230

Challenges Collaborating with yourself

The most important collaborator is yourself…

Make it easy for your future self to find files, programs, data, and to continue the work you started.

What is data?

Information obtained by scientific work and used for analysis.[1]

- Tabular Data
  - Survey responses
  - Administrative Data
- Metadata - Codebooks
- Relational Databases

	storeid	Q2_1
1	12	2. Manager
2	16	5. Other
3	41	5. Other
4	71	.
5	104	5. Other
6	123	2. Manager
7	125	5. Other
8	153	2. Manager
9	154	.
10	165	1. Owner
11	186	5. Other
12	202	5. Other
13	239	.
14	279	2. Manager
15	319	2. Manager
16	323	3. Owner and Manager
17	342	2. Manager
18	370	.
19	386	2. Manager
20	406	2. Manager
21	448	1. Owner
22	460	.
23	474	2. Manager

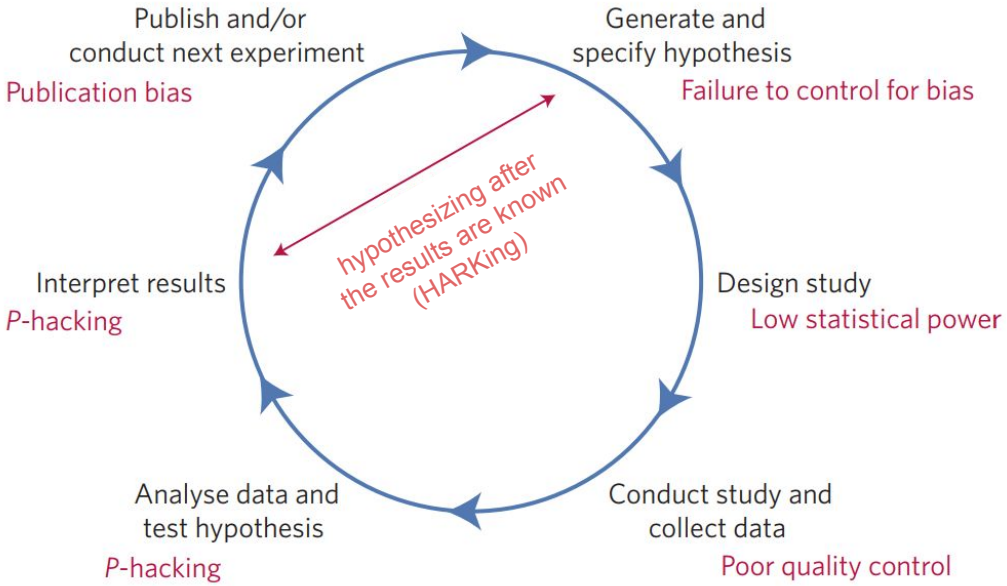
```
-------------------------------------------------------------------------------
storeid                                                             STOREID
-------------------------------------------------------------------------------

                  type:  numeric (int)

                 range:  [12,3605]                    units:  1
        unique values:  135                       missing .:  0/135

                  mean:   1716.04
             std. dev:   1061.55

           percentiles:        10%       25%       50%       75%       90%
                               279       832      1694      2564      3207

storeid:
  1.  [SETX Survey Text] Store ID
  2.  Primary Key - unique ID randomly assigned when the sample frame was set.
  3.  Use STOREID to merge Coverpage and Response Datasets.
  4.  [Citation] Rosenheim, N. et al 2018. Southeast Texas Food Retail Survey.
  5.  [Name of Saved Data File]
      RAPID17_1gv1_SNAP_SETX_RetailSurvey_2019-02-01/RAPID17_1gv1_SNAP_SETX_Reta
      > ilSurvey_2019-02-01.dta
  6.  [Program to replicate Data File]
      RAPID17_1gv1_SNAP_SETX_RetailSurvey_2019-02-01.do
  7.  [Date data file was created] 1 Feb 2019 16:48:00

-------------------------------------------------------------------------------
Q2_1                                                        Question: 1
-------------------------------------------------------------------------------

                  type:  numeric (byte)
                 label:  Q2_1lbl_r

                 range:  [1,5]                        units:  1
        unique values:  5                         missing .:  31/135

           tabulation:  Freq.   Numeric  Label
                          11         1  1. Owner
                          48         2  2. Manager
                           4         3  3. Owner and Manager
                           5         4  4. Assistant Manager
                          36         5  5. Other
                          31         .  Missing

Q2_1:
  1.  [SETX Survey Text] 1. What is your role with this business? - Selected
      Choice
  2.  [Citation] Rosenheim, N. et al 2018. Southeast Texas Food Retail Survey.
  3.  [Name of Saved Data File]
      RAPID17_1gv1_SNAP_SETX_RetailSurvey_2019-02-01/RAPID17_1gv1_SNAP_SETX_Reta
      > ilSurvey_2019-02-01.dta
  4.  [Program to replicate Data File]
      RAPID17_1gv1_SNAP_SETX_RetailSurvey_2019-02-01.do
  5.  [Date data file was created] 1 Feb 2019 16:48:00
```

Reference: 1. "data, n.". OED Online. December 2018. Oxford University Press. http://www.oed.com.lib-ezproxy.tamu.edu:2048/view/Entry/296948?rskey=c3az3E&result=1 (accessed February 08, 2019).

What is science?

"Science is an approach to knowledge… that strives to better approximate the state of nature by reducing errors in inferences."[1]

"Conceptualize science is a toolbox of… tools designed to minimize mistakes [or bias]."[1]



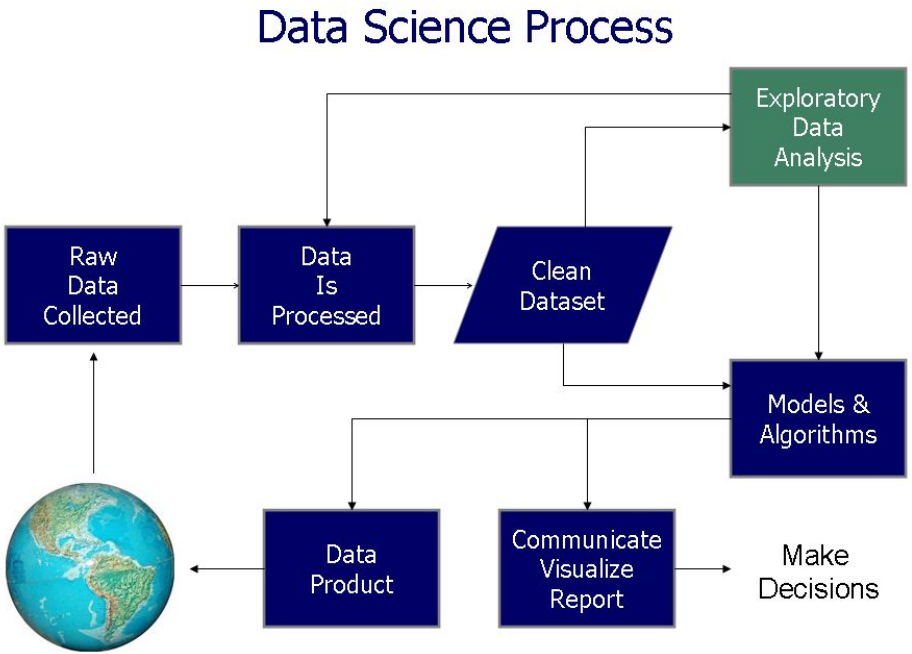
Scientific method with potential threats of bias.[2]

Reference: 1. Lilienfeld, S. O., Sauvigné, K. C., Lynn, S. J., Cautin, R. L., Latzman, R. D., & Waldman, I. D. (2015). Fifty psychological and psychiatric terms to avoid: a list of inaccurate, misleading, misused, ambiguous, and logically confused words and phrases. Frontiers in Psychology, 6, 1100. https://doi.org/10.3389/fpsyg.2015.01100
2. Munafò, M. R., Nosek, B. A., Bishop, D. V., Button, K. S., Chambers, C. D., du Sert, N. P., ... & Ioannidis, J. P. (2017). A manifesto for reproducible science. Nature Human Behaviour, 1, 0021. https://doi.org/10.1038/s41562-016-0021

What is Data Science?

Data science is a set of tools designed to minimize bias associated with the analysis of data. “The discipline of turning raw data into understanding.”1

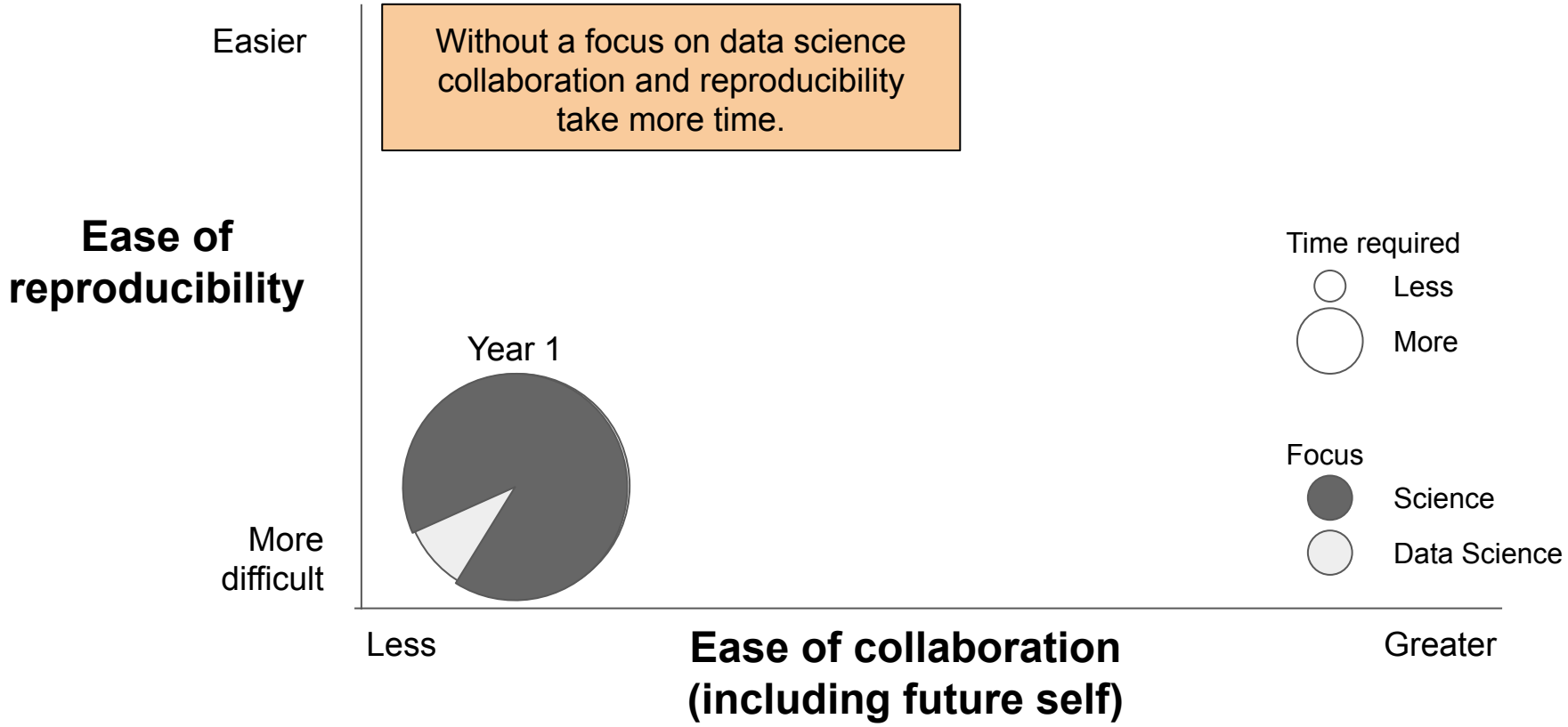
Example Tools/Concepts: Version Control, GitHub, Markdown [RMarkdown or Jupyter Notebook], Workflow, Repositories, Permanent Identifiers e.g. ‘‘handle’’ (hdl) or ‘‘digital object identifier’’ (doi)



2. Caldwell, J. (2016) A Data Science Solution to the Question "What is Data Science?" R-Bloggers

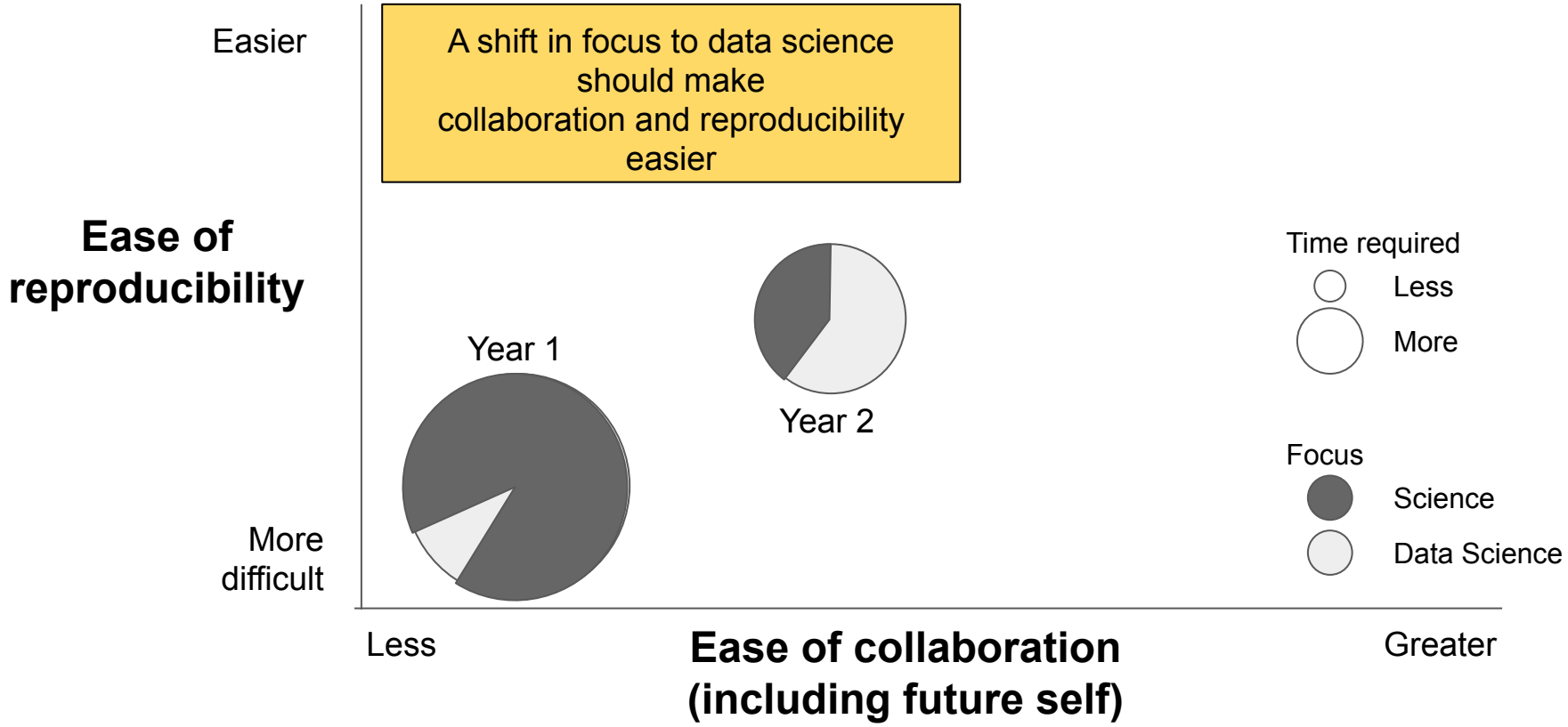
Reference: 1. Lowndes, J. S. S., Best, B. D., Scarborough, C., Afflerbach, J. C., Frazier, M. R., O'Hara, C. C., ... & Halpern, B. S. (2017). Our path to better science in less time using open data science tools. *Nature ecology & evolution*, *1*(6), 160. https://doi.org/10.1038/s41559-017-0160

Goal: Better science in less time



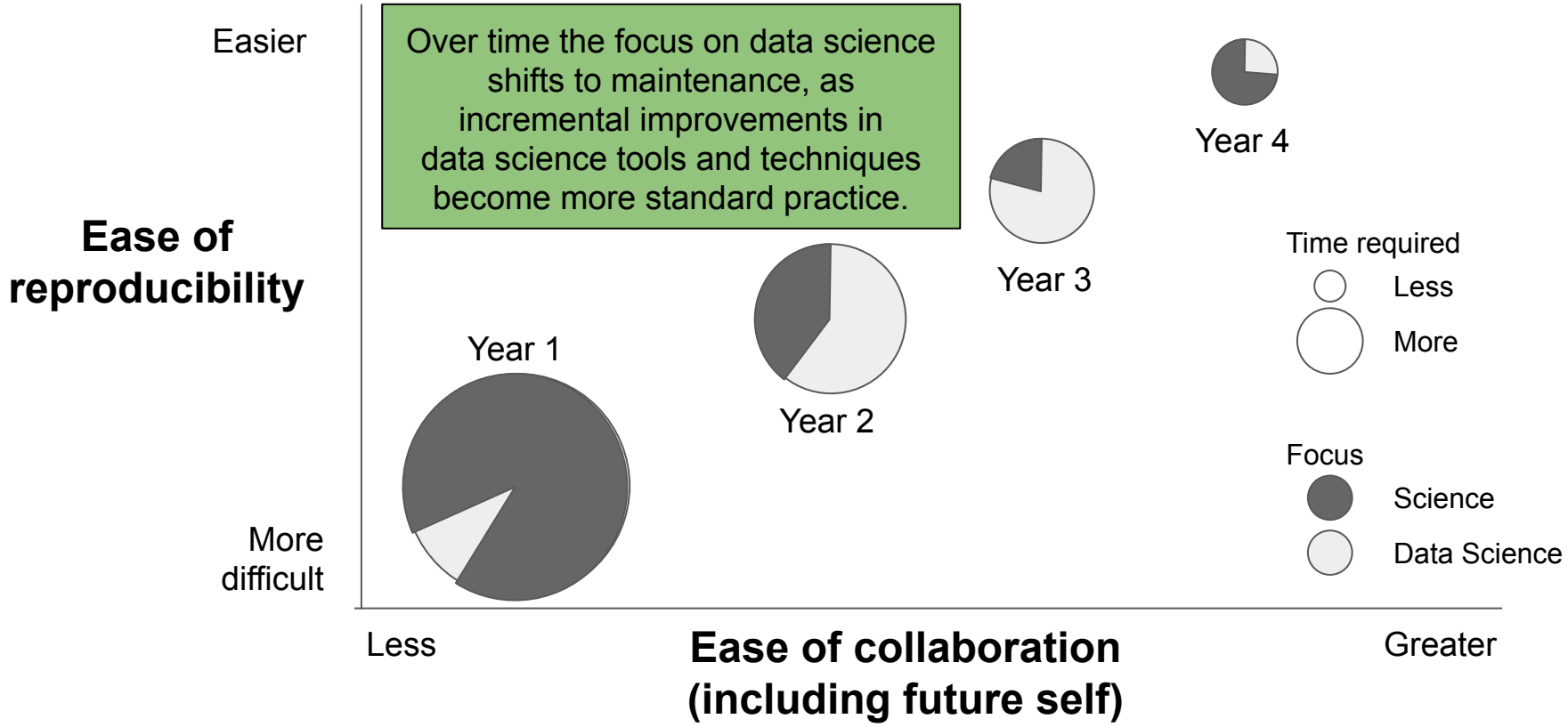
Reference: Lowndes, J. S. S., Best, B. D., Scarborough, C., Afflerbach, J. C., Frazier, M. R., O'Hara, C. C., ... & Halpern, B. S. (2017). Our path to better science in less time using open data science tools. *Nature ecology & evolution*, *1*(6), 160. https://doi.org/10.1038/s41559-017-0160

Goal: Better science in less time



Reference: Lowndes, J. S. S., Best, B. D., Scarborough, C., Afflerbach, J. C., Frazier, M. R., O'Hara, C. C., ... & Halpern, B. S. (2017). Our path to better science in less time using open data science tools. *Nature ecology & evolution*, *1*(6), 160. https://doi.org/10.1038/s41559-017-0160

Goal: Better science in less time



Reference: Lowndes, J. S. S., Best, B. D., Scarborough, C., Afflerbach, J. C., Frazier, M. R., O'Hara, C. C., ... & Halpern, B. S. (2017). Our path to better science in less time using open data science tools. *Nature ecology & evolution*, *1*(6), 160. https://doi.org/10.1038/s41559-017-0160

Replication Standard - Individual or Social Contract?

Individual Responsibility	Social Contract
If asked a researcher should be able to provide the files to replicate published results.	Files to replicate published results are submitted at time of publication.
Emphasis on trust.	Emphasis on transparency.
Faith in the author.	Focus on openness.
Reinforcement of status.	Distributes power and access.

Reference: Freese, J. (2007). Replication standards for quantitative social science: Why not sociology?. *Sociological Methods & Research*, *36*(2), 153-172.

Course objectives...

Help to overcome challenges to data sharing, documentation, publication, and analytics.

Help to develop a culture that supports a social contract for data replication.

To be a part of a community that bolsters data science and leads to more open, discoverable, reproducible research.

Motivating Examples

Rosenheim, Nathanael; Day, Wayne; Seong, Kijin (2021) “Automated Neighborhood Characteristics for Community Resilience Planning.” DesignSafe-CI. https://doi.org/10.17603/ds2-hj0p-bp40.

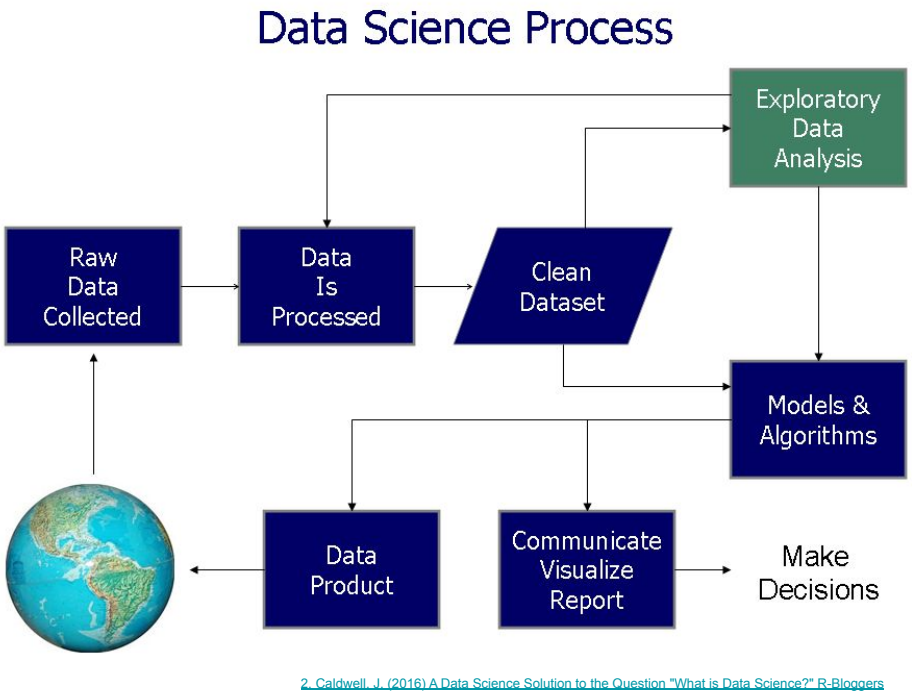
Roy, Malini; Rosenheim, Nathanael (2021) “Longitudinal Social Vulnerability Data Exploration for Harris County Census Tracts.” DesignSafe-CI. https://doi.org/10.17603/ds2-hn6r-dh03.

Gu, Donghwan, and Rosenheim, Nathanael. Demographic Analysis Workflow using Census API in Jupyter Notebook: 1990-2000 Population Size and Change. Ann Arbor, MI: Inter-university Consortium for Political and Social Research [distributor], 2020-07-30. https://doi.org/10.3886/E120381V1

Goodman, Cooper, Rosenheim, Nathanael, Day, Wayne, Gu, Donghwan, and Korukonda, Jayasaree. Population Distribution Workflow using Census API in Jupyter Notebook: Dynamic Map of Census Tracts in Boone County, KY, 2000. Ann Arbor, MI: Inter-university Consortium for Political and Social Research [distributor], 2020-07-31. https://doi.org/10.3886/E120382V1

Motivating Example

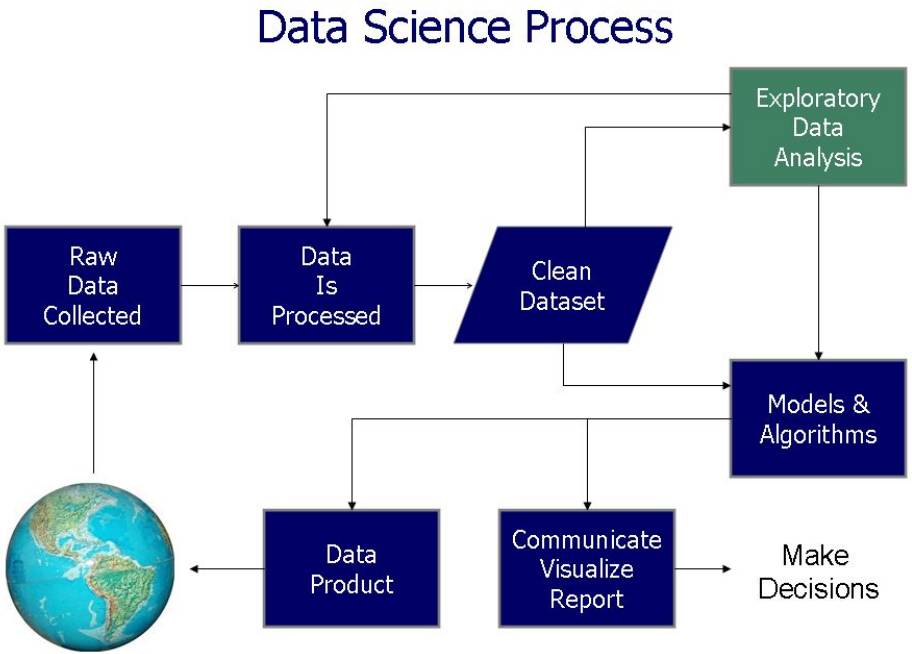
Rosenheim, Nathanael; Day, Wayne; Seong, Kijin (2021) “Automated Neighborhood Characteristics for Community Resilience Planning.” DesignSafe-CI. https://doi.org/10.17603/ds2-hj0p-bp40.



Reference: 1. Lowndes, J. S. S., Best, B. D., Scarborough, C., Afflerbach, J. C., Frazier, M. R., O'Hara, C. C., ... & Halpern, B. S. (2017). Our path to better science in less time using open data science tools. *Nature ecology & evolution*, *1*(6), 160. https://doi.org/10.1038/s41559-017-0160

Motivating Example

Motivation: Demonstrate the potential of Census API and Jupyter Notebooks for obtaining, cleaning, and exploring Census data.



2. Caldwell, J. (2016) A Data Science Solution to the Question "What is Data Science?" R-Bloggers

Rosenheim, Nathanael; Day, Wayne; Seong, Kijin (2021) “Automated Neighborhood Characteristics for Community Resilience Planning.” DesignSafe-CI. https://doi.org/10.17603/ds2-hj0p-bp40.

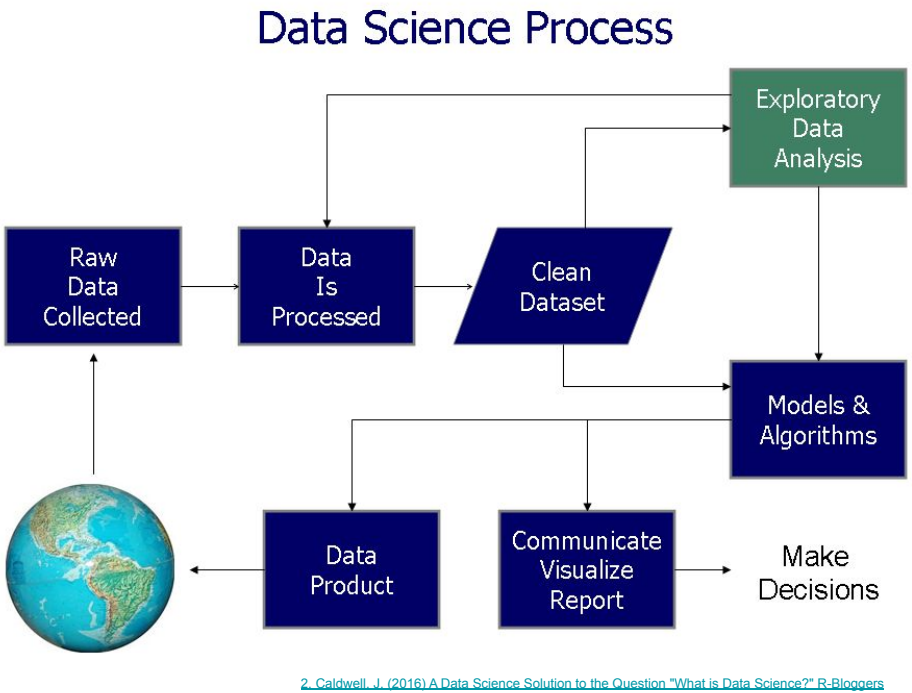
Reference: 1. Lowndes, J. S. S., Best, B. D., Scarborough, C., Afflerbach, J. C., Frazier, M. R., O'Hara, C. C., ... & Halpern, B. S. (2017). Our path to better science in less time using open data science tools. *Nature ecology & evolution*, *1*(6), 160. https://doi.org/10.1038/s41559-017-0160

Motivating Example

Raw data collection: 2010 Decennial Census data using Census API and TIGER Shapefiles for Block Group Boundaries

Program automates data collection:

IN-CORE_CENSUSAPI_DWNLD_JOPLIN MSA_2021-01-13.ipynb



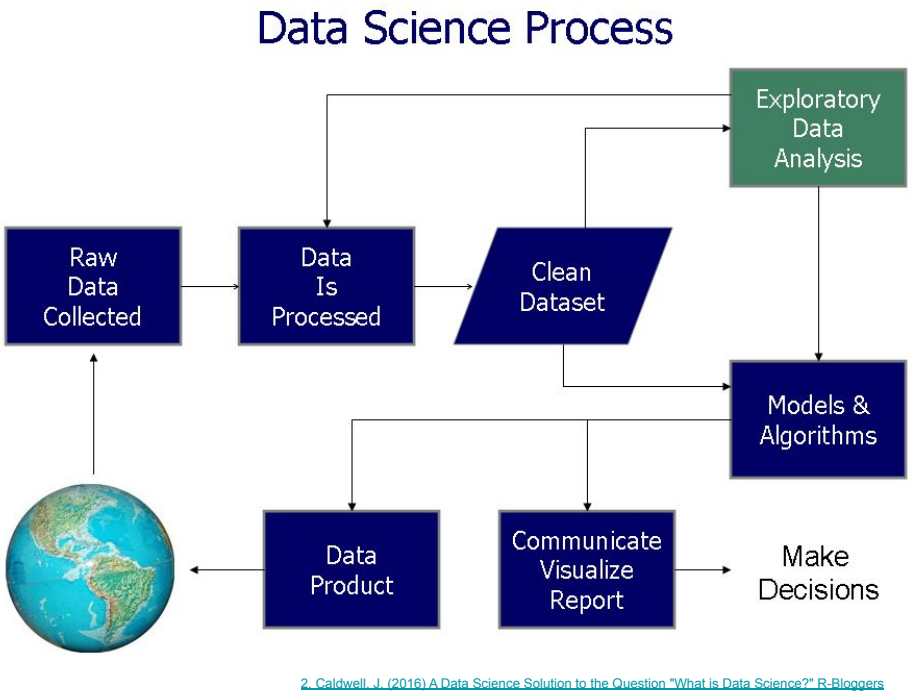
Reference: 1. Lowndes, J. S. S., Best, B. D., Scarborough, C., Afflerbach, J. C., Frazier, M. R., O'Hara, C. C., ... & Halpern, B. S. (2017). Our path to better science in less time using open data science tools. *Nature ecology & evolution*, *1*(6), 160. https://doi.org/10.1038/s41559-017-0160

Motivating Example

Data is Processed:

Jupyter Notebook

IN-CORE_CENSUSAPI_DWNLD_JOPLIN MSA_2021-01-13.ipynb



2. Caldwell, J. (2016) A Data Science Solution to the Question "What is Data Science?" R-Bloggers

Reference: 1. Lowndes, J. S. S., Best, B. D., Scarborough, C., Afflerbach, J. C., Frazier, M. R., O'Hara, C. C., ... & Halpern, B. S. (2017). Our path to better science in less time using open data science tools. *Nature ecology & evolution*, *1*(6), 160. https://doi.org/10.1038/s41559-017-0160

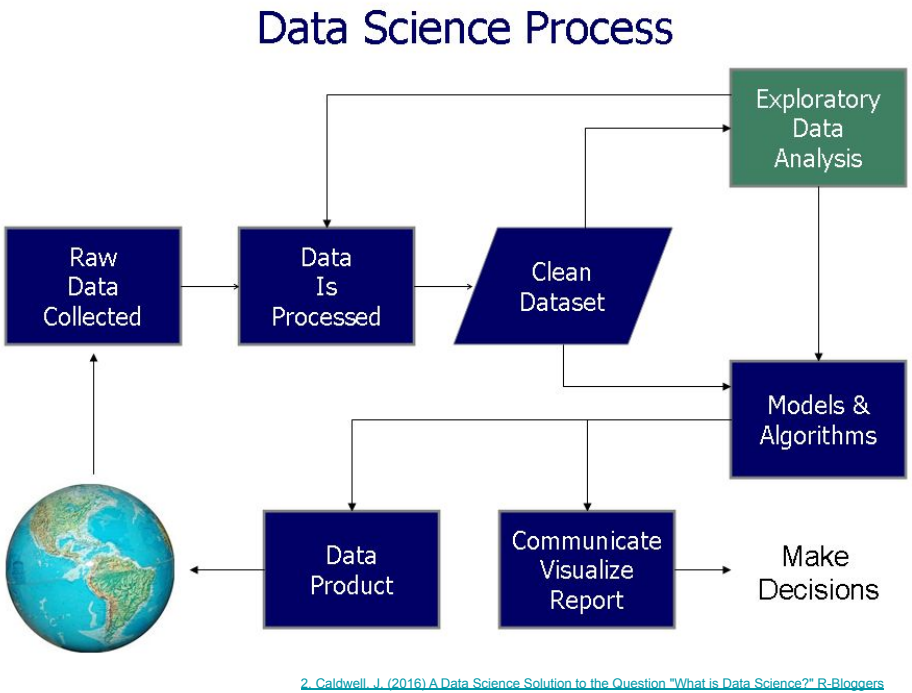
Motivating Example

Clean Dataset:

New variables are created and block group data is merged with shapefiles.

Program generates CSV, HTML, and Shp files

IN-CORE_CENSUSAPI_DWNLD_JOPLIN MSA_2021-01-11.csv



2. Caldwell, J. (2016) A Data Science Solution to the Question "What is Data Science?" R-Bloggers

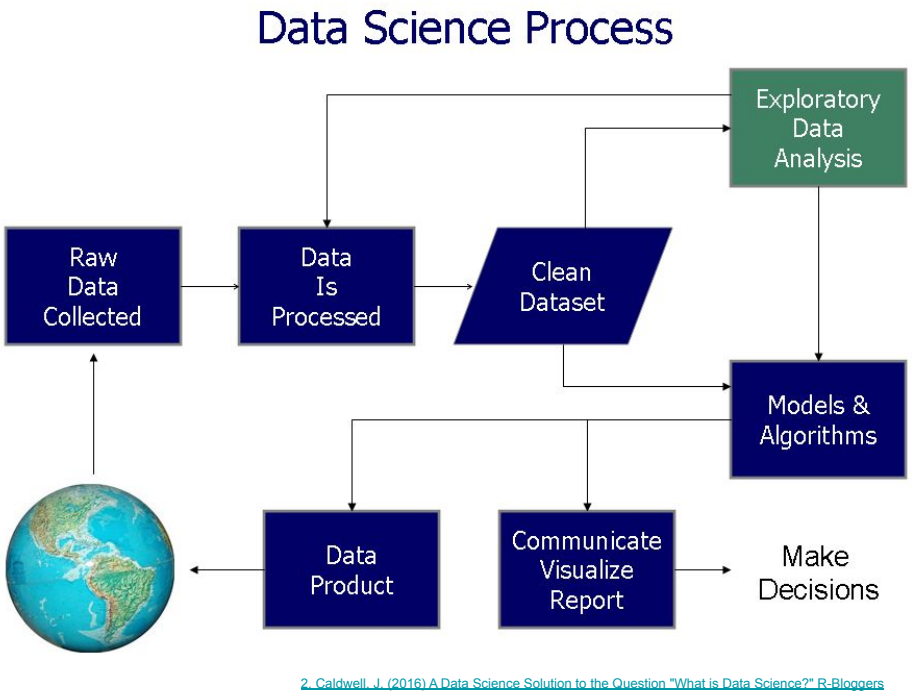
Reference: 1. Lowndes, J. S. S., Best, B. D., Scarborough, C., Afflerbach, J. C., Frazier, M. R., O'Hara, C. C., ... & Halpern, B. S. (2017). Our path to better science in less time using open data science tools. *Nature ecology & evolution*, *1*(6), 160. https://doi.org/10.1038/s41559-017-0160

Motivating Example

Exploratory Data Analysis:

Jupyter Notebook includes descriptive statistics, tables, and exploratory maps

IN-CORE_CENSUSAPI_DWNLD_JOPLIN MSA_2021-01-11_map.html



2. Caldwell, J. (2016) A Data Science Solution to the Question "What is Data Science?" R-Bloggers

Reference: 1. Lowndes, J. S. S., Best, B. D., Scarborough, C., Afflerbach, J. C., Frazier, M. R., O'Hara, C. C., ... & Halpern, B. S. (2017). Our path to better science in less time using open data science tools. *Nature ecology & evolution*, *1*(6), 160. https://doi.org/10.1038/s41559-017-0160

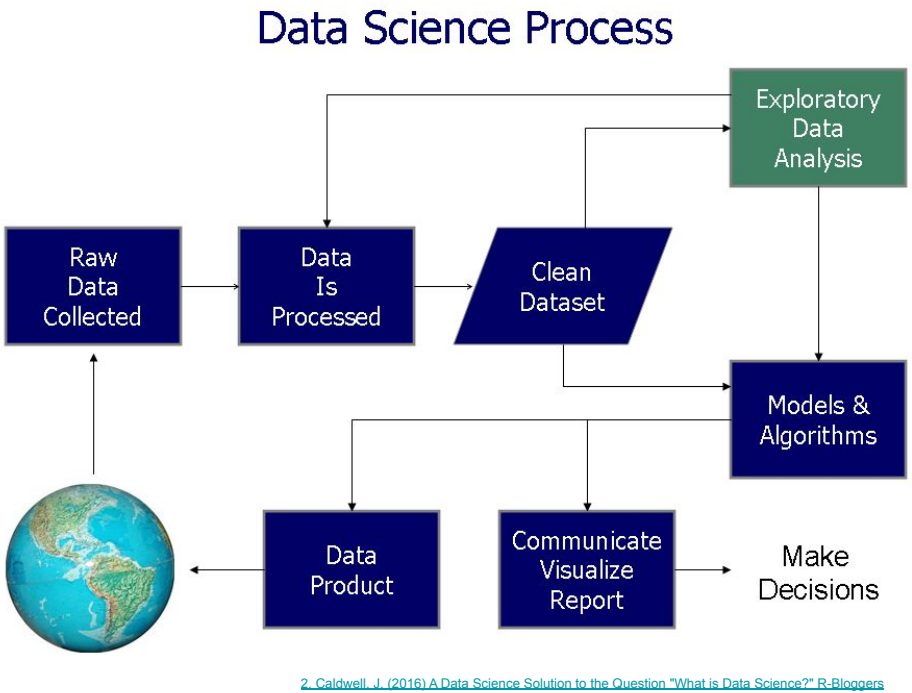
Motivating Example

Models and Algorithms:

Workflow does not include models

Algorithm saved as python program file:

IN-CORE_CENSUSAPI_BGMAP_2021-01-19T1403.py



2. Caldwell, J. (2016) A Data Science Solution to the Question "What is Data Science?" R-Bloggers

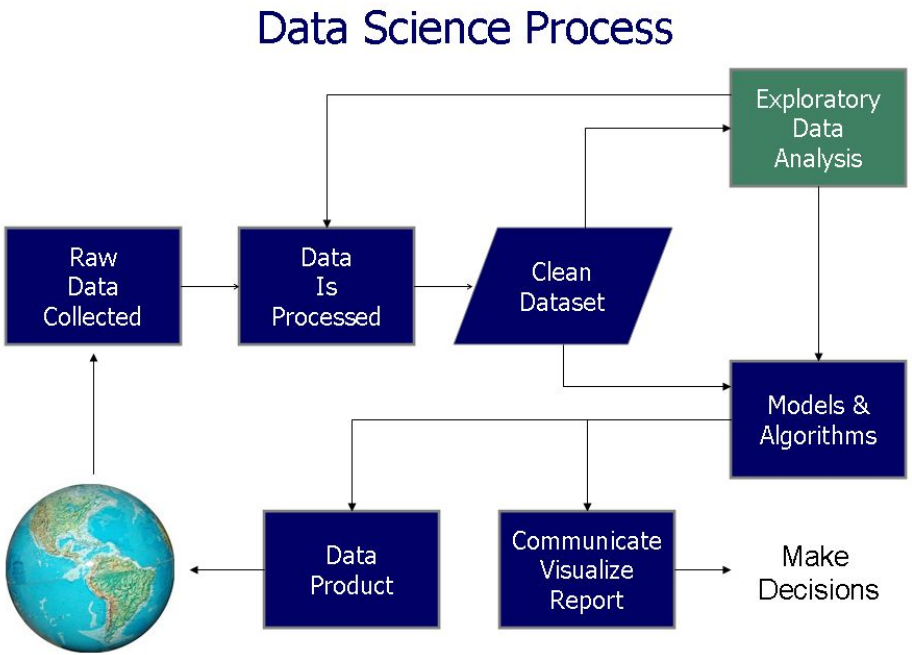
Reference: 1. Lowndes, J. S. S., Best, B. D., Scarborough, C., Afflerbach, J. C., Frazier, M. R., O'Hara, C. C., ... & Halpern, B. S. (2017). Our path to better science in less time using open data science tools. *Nature ecology & evolution*, *1*(6), 160. https://doi.org/10.1038/s41559-017-0160

Motivating Example

Communicate, Visualize, Report

_ReadMeFirst_CensusAPI_BlockGroup_2021-01-19.docx

Step by Step Instruction-How to Download Tables as a CSV file_2021-01-07.pdf



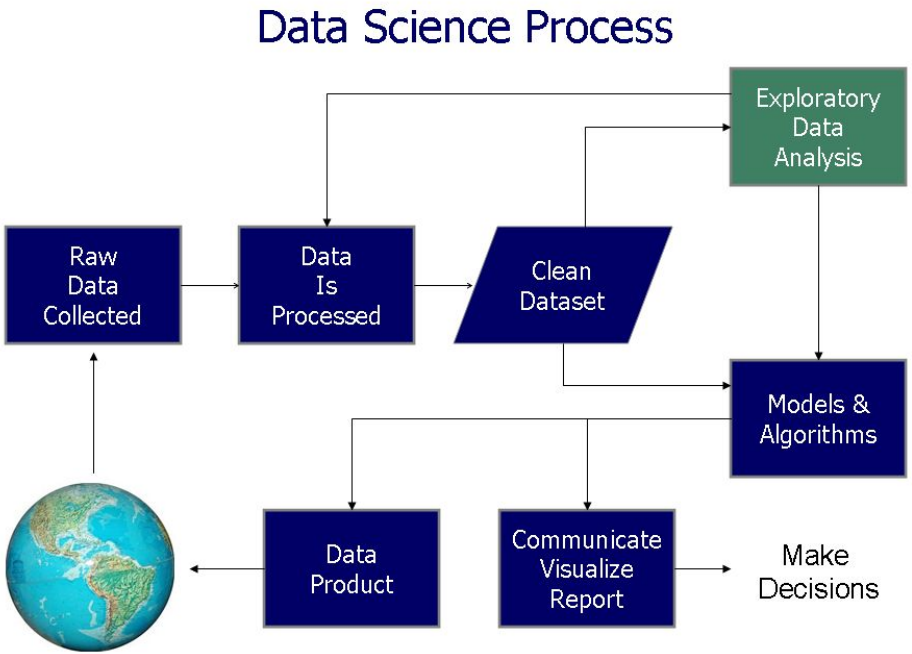
2. Caldwell, J. (2016) A Data Science Solution to the Question "What is Data Science?" R-Bloggers

Reference: 1. Lowndes, J. S. S., Best, B. D., Scarborough, C., Afflerbach, J. C., Frazier, M. R., O'Hara, C. C., ... & Halpern, B. S. (2017). Our path to better science in less time using open data science tools. *Nature ecology & evolution*, *1*(6), 160. https://doi.org/10.1038/s41559-017-0160

Motivating Example

Data Product:

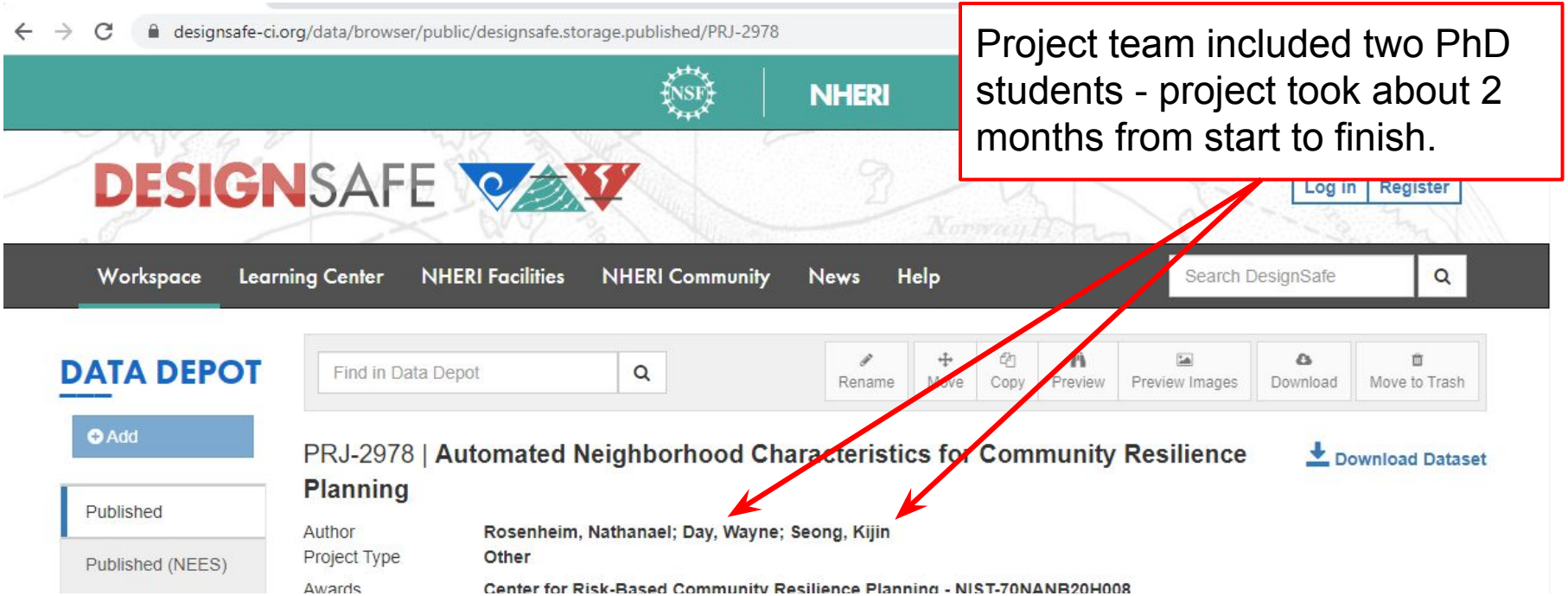
Rosenheim, Nathanael; Day, Wayne; Seong, Kijin (2021) “Automated Neighborhood Characteristics for Community Resilience Planning.” DesignSafe-CI. https://doi.org/10.17603/ds2-hj0p-bp40.



2. Caldwell, J. (2016) A Data Science Solution to the Question "What is Data Science?" R-Bloggers

Reference: 1. Lowndes, J. S. S., Best, B. D., Scarborough, C., Afflerbach, J. C., Frazier, M. R., O'Hara, C. C., ... & Halpern, B. S. (2017). Our path to better science in less time using open data science tools. *Nature ecology & evolution*, *1*(6), 160. https://doi.org/10.1038/s41559-017-0160

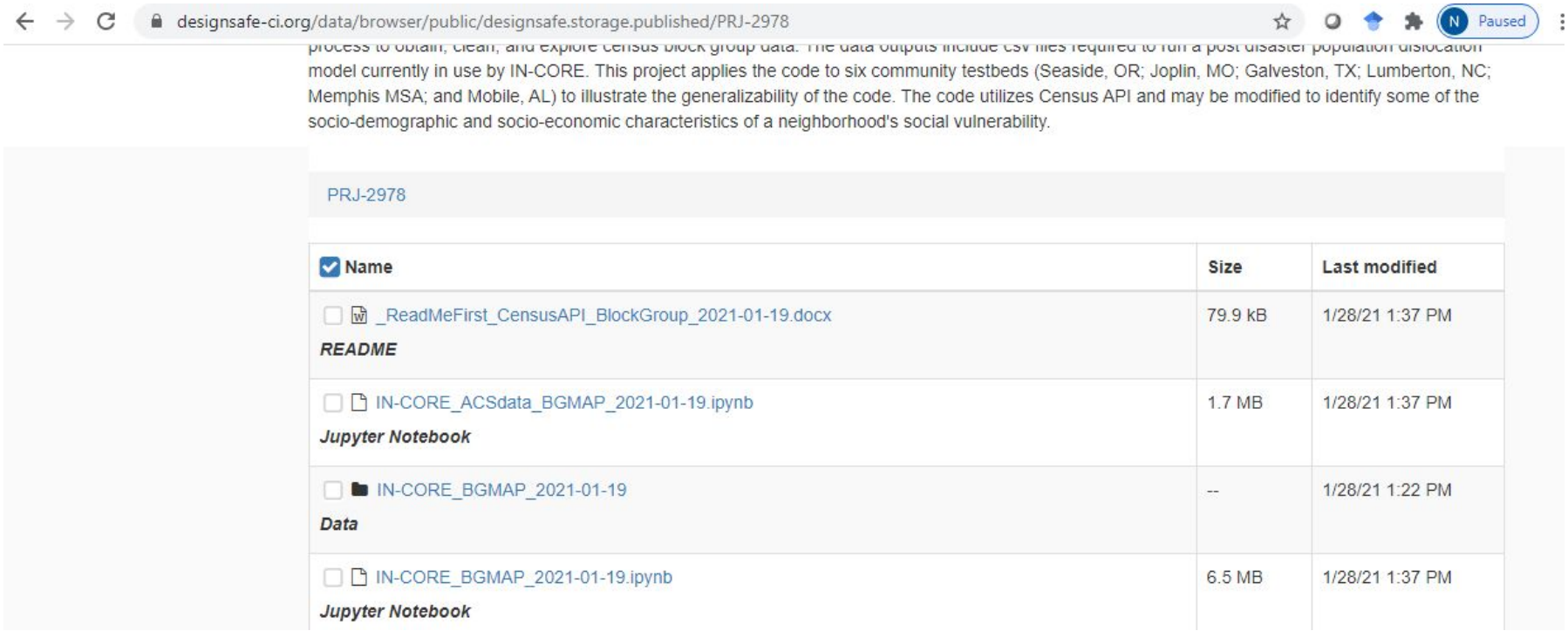
What the project looks like in DesignSafe



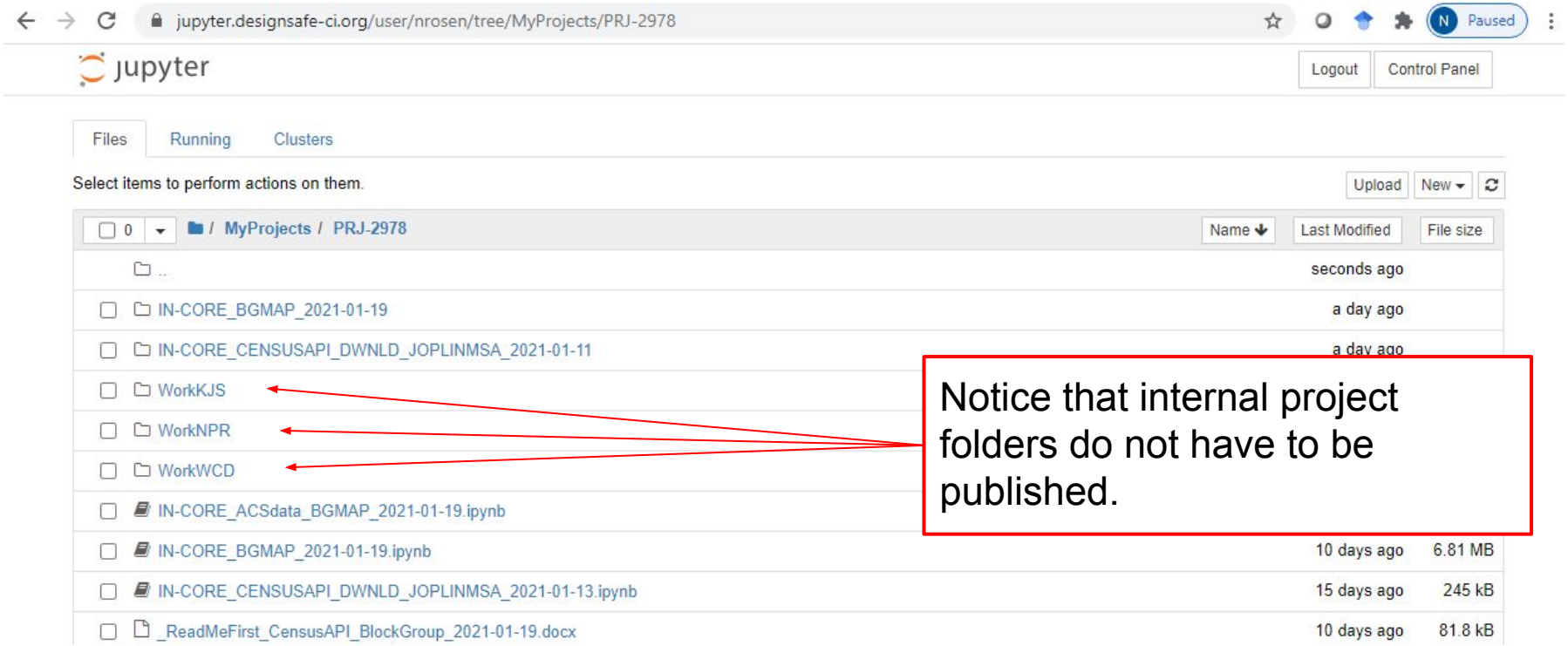
Project team included two PhD students - project took about 2 months from start to finish.

Rosenheim, Nathanael; Day, Wayne; Seong, Kijin (2021) “Automated Neighborhood Characteristics for Community Resilience Planning.” DesignSafe-CI. https://doi.org/10.17603/ds2-hj0p-bp40.

Project files developed and created in DesignSafe



Project Developed Within Jupyter-DesignSafe



Jupyter Notebook uses Python - Free and Open Source

jupyter IN-CORE_ACSdata_BGMAP_2021-01-19 Last Checkpoint: 01/19/2021 (unsaved changes)

Logout | Control Panel

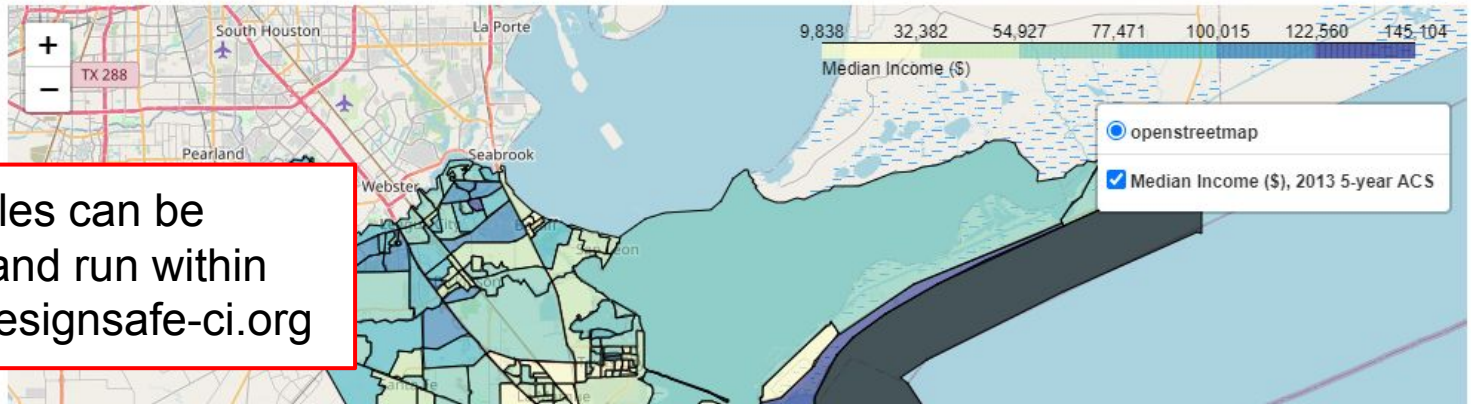
File Edit View Insert Cell Kernel Widgets Help | Trusted | Python 3

Run | Markdown

```
map.fit_bounds([sw_corner, ne_corner])

map.save(filename+'meidianincome.html')
#files.download(f'output_maplayers/{mapname}.html')
```

```
The center of the map data file is located at -94.96170002319587 29.394953239690718
The map data file is bounded by at [29.062738999999997, -95.233081] [29.598264, -94.369361]
```

In [24]:
```
display(map)
```



Python files can be opened and run within jupyter.designsafe-ci.org

Thank you!

Nathanael Rosenheim

nroseheim@arch.tamu.edu

Thoughts after the workshop

Data science may not be interested in reproducible, open, collaborative research.

What tools help make collaboration and replication easier - “Better science in less time”

Ideal dataset - provides individual level characteristics that identify a person or business. Provides information that helps distinguish and predict outcome. For example, which applicants to Texas A&M will decide to enroll? Need to have enough detail about each applicant to predict the outcome.

Ideal dataset - is linkable to other datasets that provide even more details and information.

Thoughts after the workshop

Some discussion about the drawbacks of synthetic, derived or sanitized data.

Discussion about how “hiding” data leads to bad data. Need to reward making data available and more usefull.

Thoughts about decision makers interpretation of data analytics:

- Each decision maker (example a football coach) will have different ways they want to see the interpretation of data. For example, some like dashboards, others want just a few numbers.
- Emotions in the room make a difference
- The person analyzing the data needs to be very familiar with the data - actually needs to see the data collected.
- Need to “read the audience” as the story is being told.

Thoughts after the workshop

Common theme across presentations “Real action comes from individual data”. Seemed to be true for the perspectives of public health, transportation, athletics, admissions.

When thinking about data collection start by planning to share the data.

An option for data sharing: Texas Data Repository https://data.tdl.org/