Challenges to collaboration and reproducibility for community resilience planning

Nathanael Rosenheim, PhD

Workshop on Operational Data Science
Texas A&M Institute of Data Science (TAMIDS)
February 11, 2019
Texas A&M MSC Rm 2406A

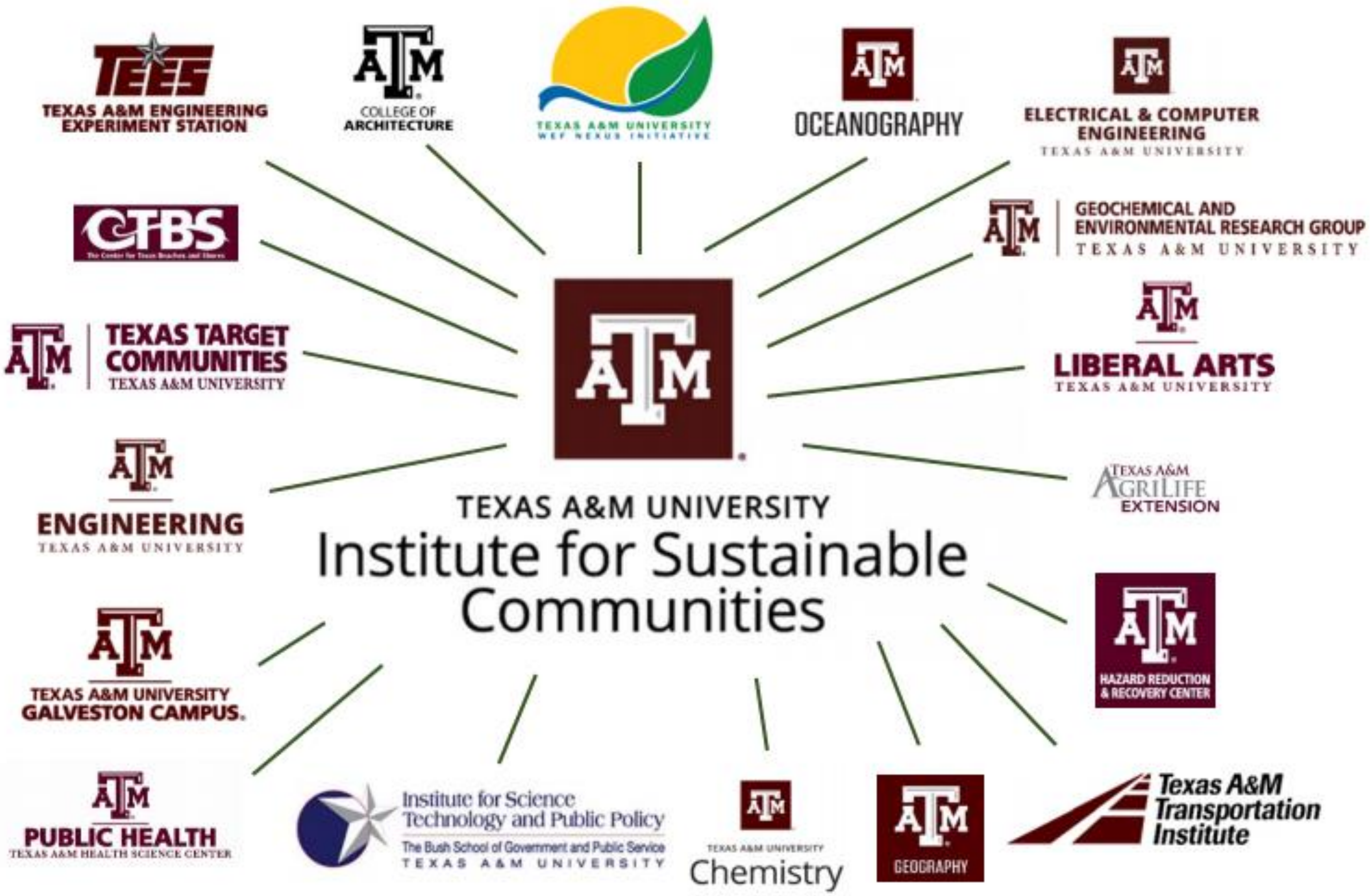
The IfSC and HRRC

- The university’s focal point for interdisciplinary sustainable community research.
- 30 faculty members from across the university
- Over 60 graduate students
- A focus on collaboration between disciplines and with communities

TEXAS A&M UNIVERSITY
Institute for Sustainable Communities

HAZARD REDUCTION & RECOVERY CENTER
TEXAS A&M UNIVERSITY

Collaborating Across Campus



TEXAS A&M UNIVERSITY

Institute for Sustainable Communities

TEES TEXAS A&M ENGINEERING EXPERIMENT STATION

ATM COLLEGE OF ARCHITECTURE

TEXAS A&M UNIVERSITY WEF NEXUS INITIATIVE

ATM OCEANOGRAPHY

ATM ELECTRICAL & COMPUTER ENGINEERING TEXAS A&M UNIVERSITY

CTBS The Center for Texas Beaches and Shores

ATM GEOCHEMICAL AND ENVIRONMENTAL RESEARCH GROUP TEXAS A&M UNIVERSITY

ATM TEXAS TARGET COMMUNITIES TEXAS A&M UNIVERSITY

ATM LIBERAL ARTS TEXAS A&M UNIVERSITY

ATM ENGINEERING TEXAS A&M UNIVERSITY

TEXAS A&M AGRILIFE EXTENSION

ATM TEXAS A&M UNIVERSITY GALVESTON CAMPUS.

ATM HAZARD REDUCTION & RECOVERY CENTER

ATM PUBLIC HEALTH TEXAS A&M HEALTH SCIENCE CENTER

Institute for Science Technology and Public Policy The Bush School of Government and Public Service TEXAS A&M UNIVERSITY

ATM TEXAS A&M UNIVERSITY Chemistry

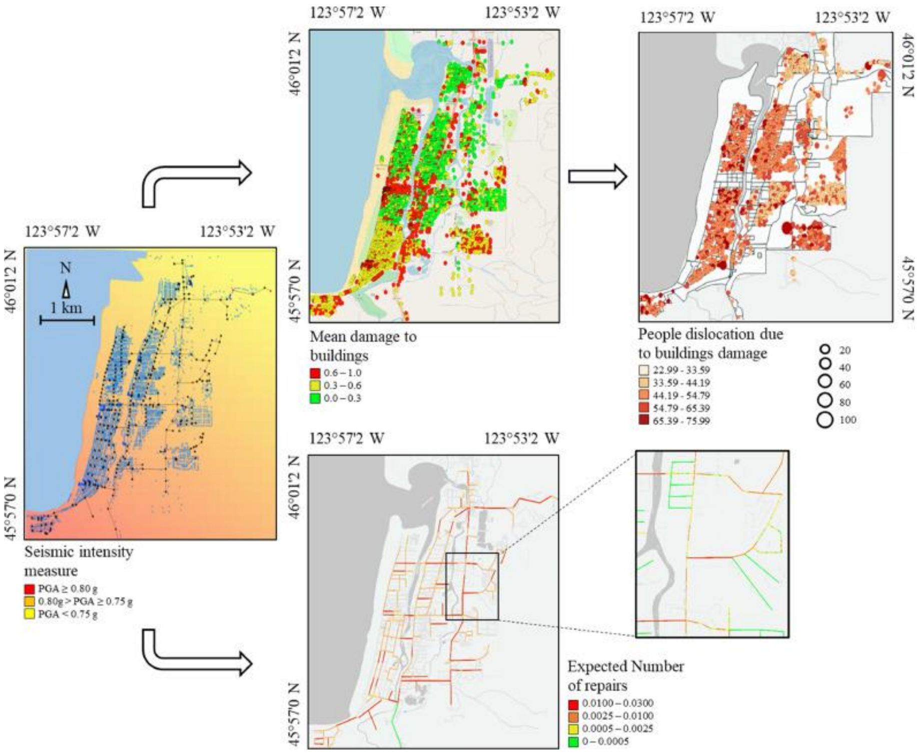
ATM GEOGRAPHY

Texas A&M Transportation Institute

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.



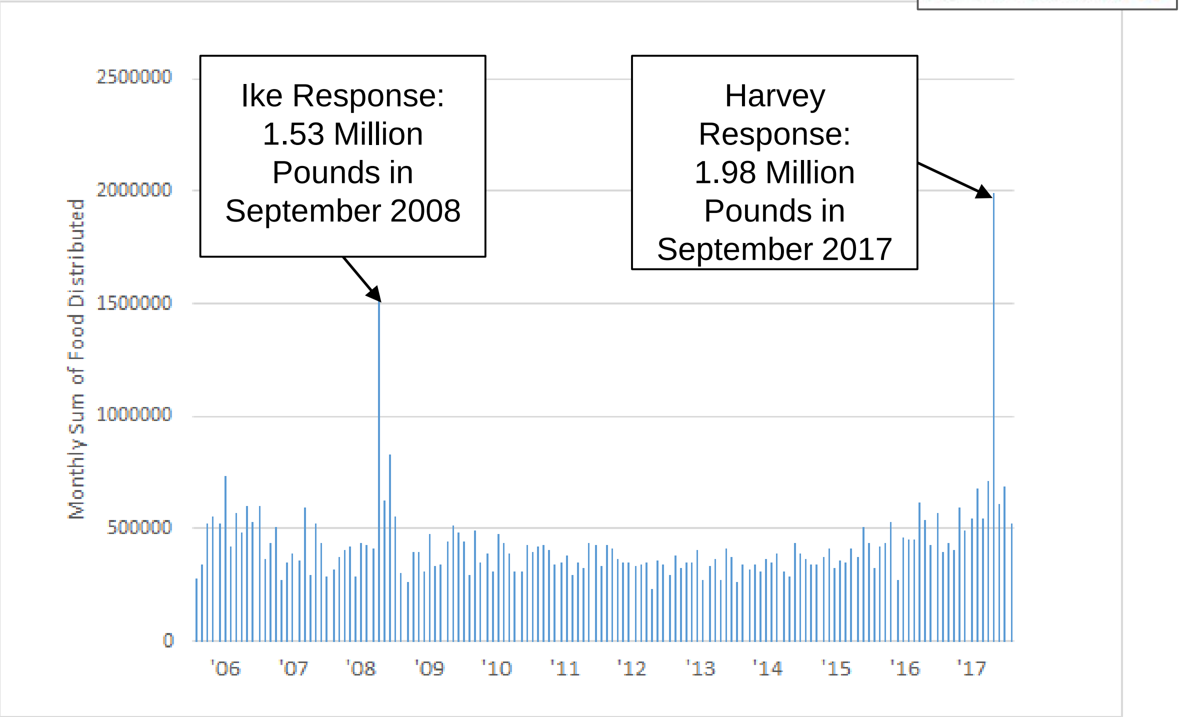
Guidotti, R., Gardoni, P., & Rosenheim, N. (2019). Integration of physical infrastructure and social systems in communities' reliability and resilience analysis. *Reliability Engineering & System Safety*. doi.org/10.1016/j.ress.2019.01.008

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.



Ike Response: 1.53 Million Pounds in September 2008

Harvey Response: 1.98 Million Pounds in September 2017

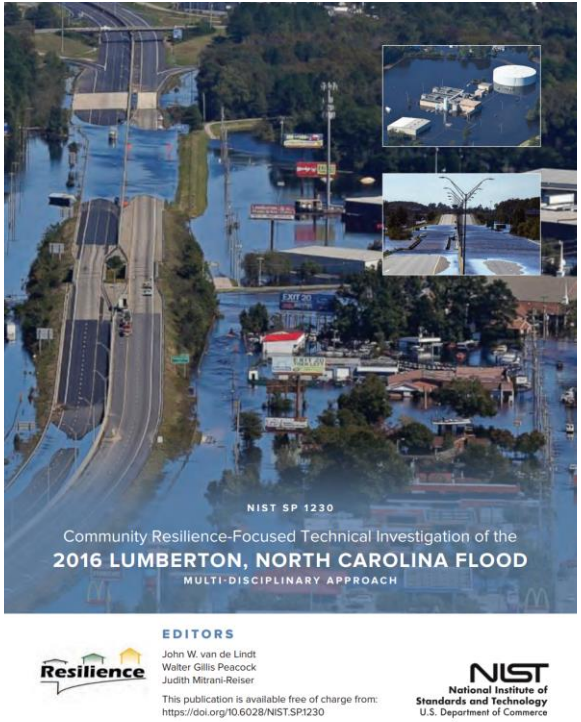
Source Data: Southeast Texas Food Bank Primarius Reports

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.



NIST SP 1230

Community Resilience-Focused Technical Investigation of the

2016 LUMBERTON, NORTH CAROLINA FLOOD

MULTI-DISCIPLINARY APPROACH

EDITORS

John W. van de Lindt
Walter Gillis Peacock
Judith Mitrani-Reiser

This publication is available free of charge from:
https://doi.org/10.6028/NIST.SP.1230

Resilience

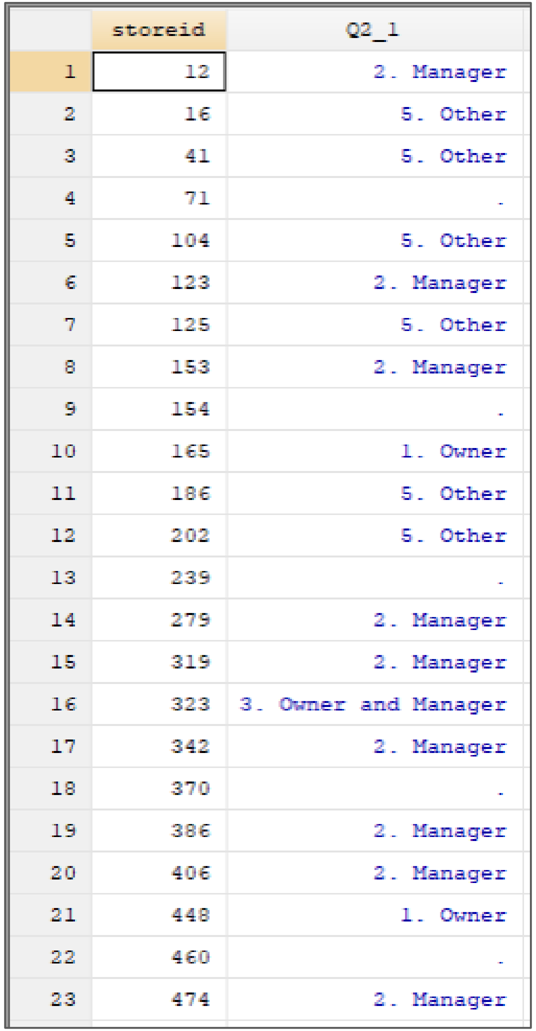
NIST
National Institute of Standards and Technology
U.S. Department of Commerce

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

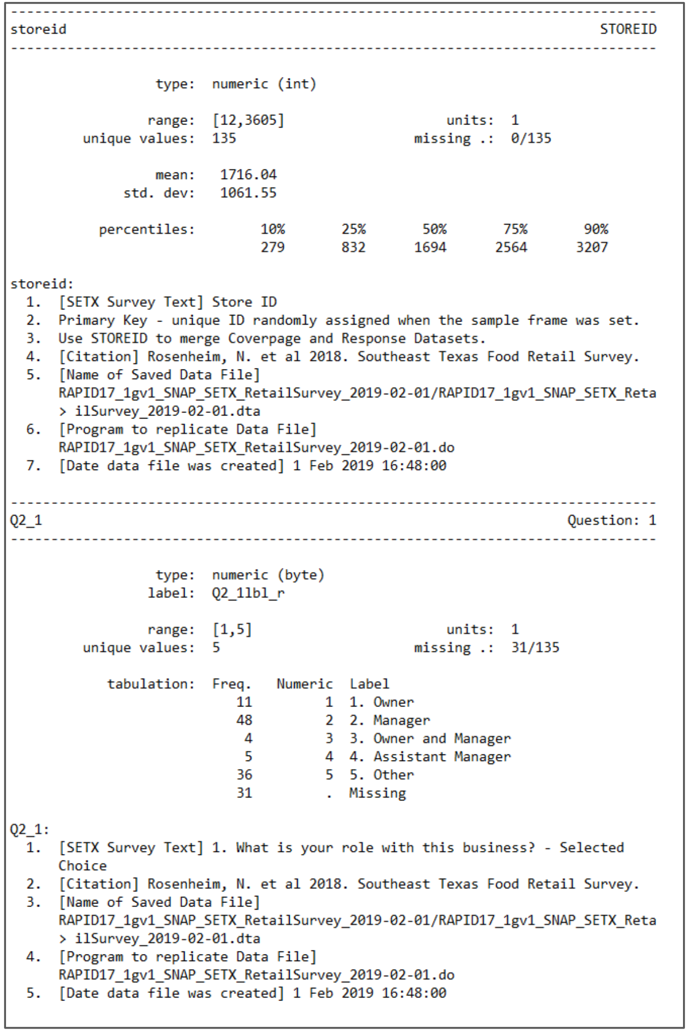
What is data?

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



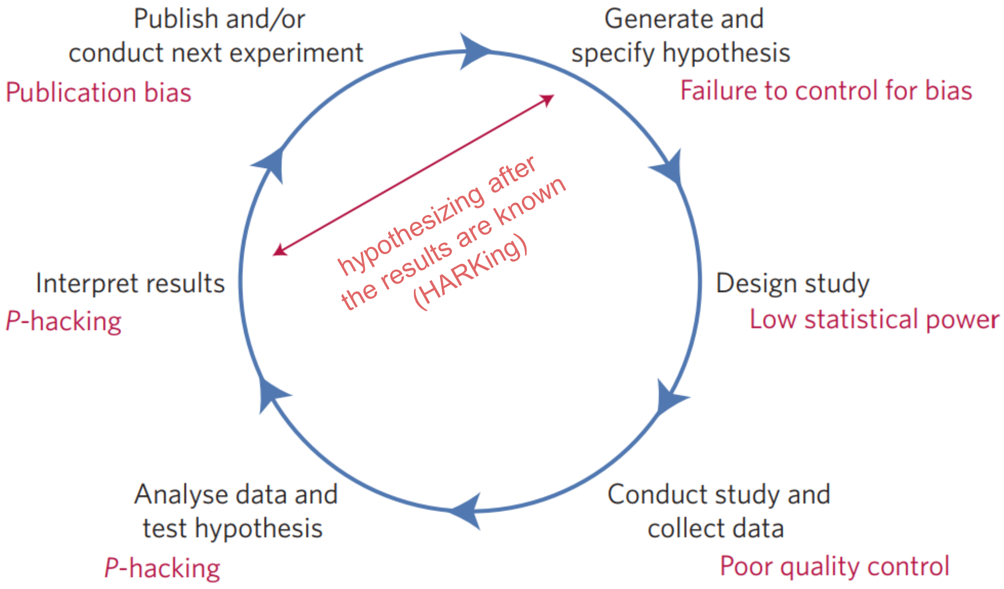
```
-------------------------------------------------------------------------------------------
storeid                                                                            STOREID
-------------------------------------------------------------------------------------------

                  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]


Generate and specify hypothesis
Failure to control for bias
Design study
Low statistical power
Conduct study and collect data
Poor quality control
Analyse data and test hypothesis
P-hacking
Interpret results
P-hacking
Publish and/or conduct next experiment
Publication bias
hypothesizing after the results are known (HARKing)

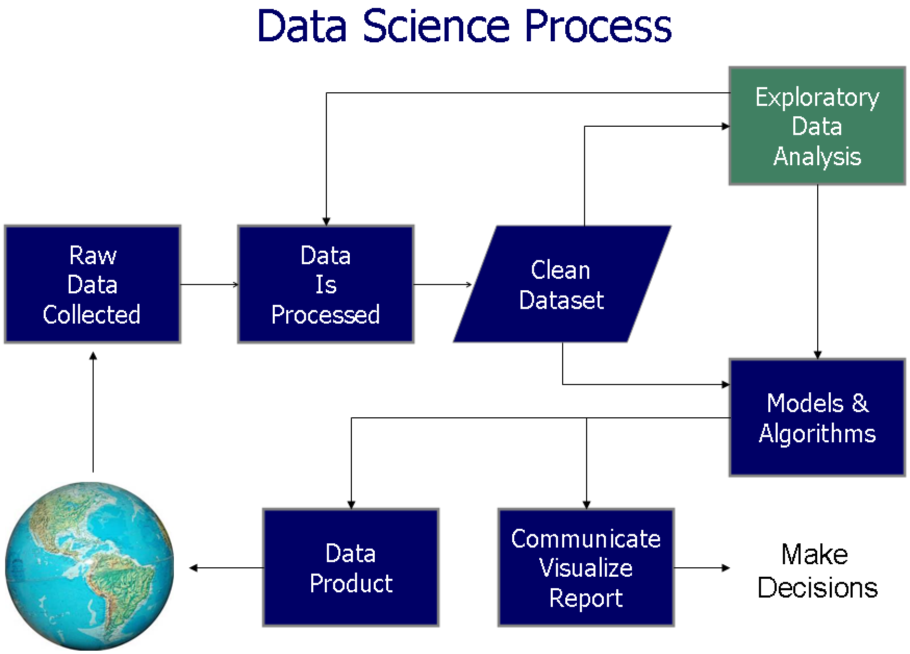
Scientific method with potential threats of bias.[2]

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

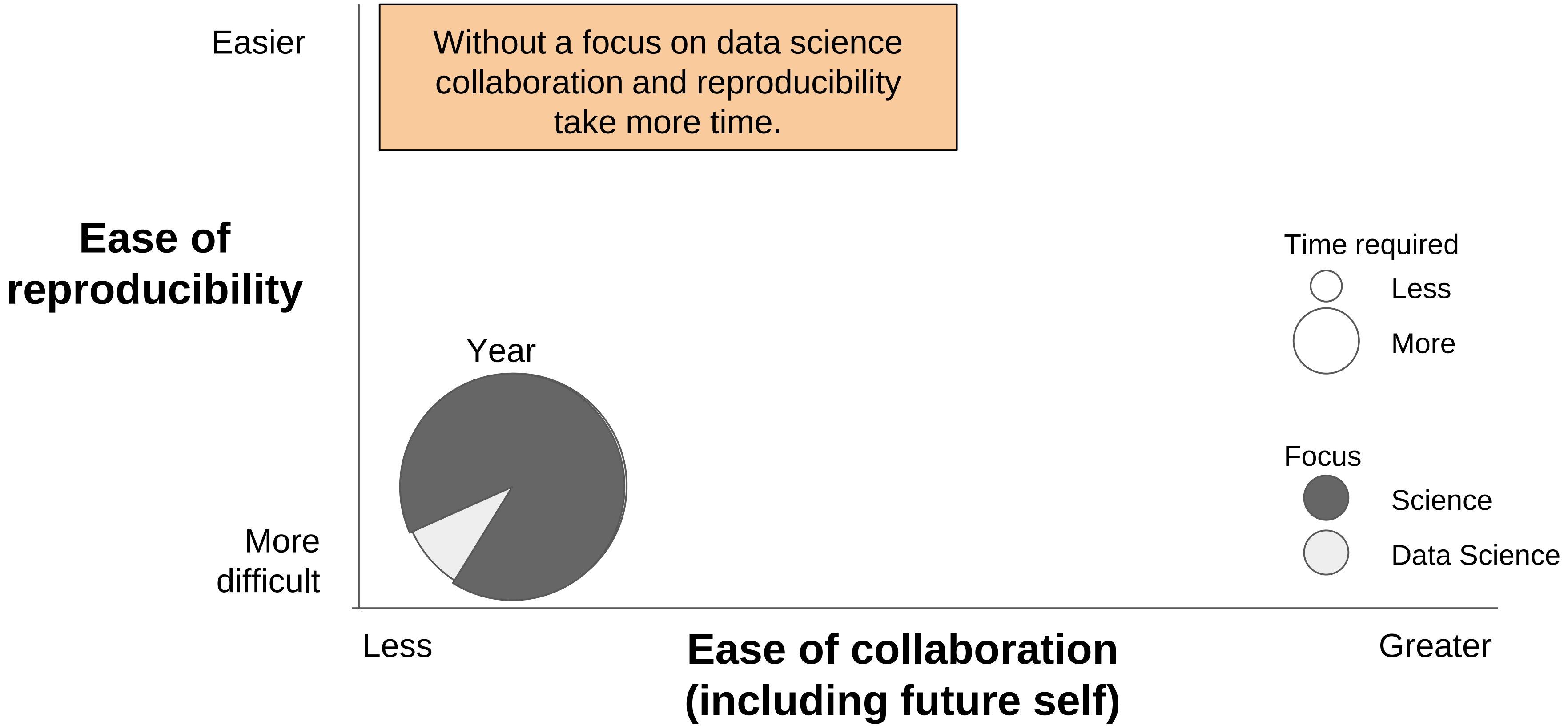
Example Tools/Concepts: Version Control, GitHub, Markdown [RMarkdown or Jupyter Notebook], Workflow, Repositories, Permanent Identifiers e.g. “handle” (hdl) or “digital object identifier” (doi)



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

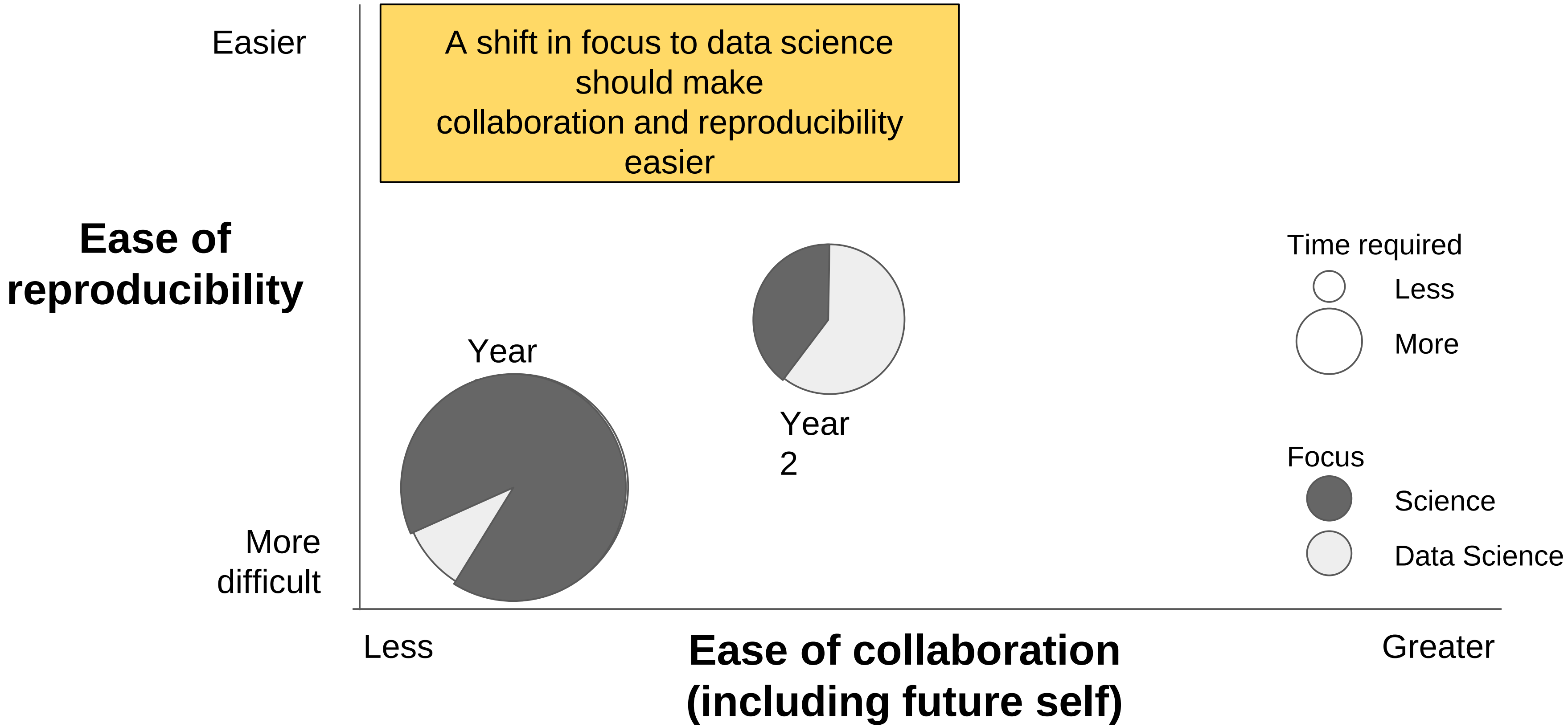
Goal: Better science in less time



Without a focus on data science collaboration and reproducibility take more 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

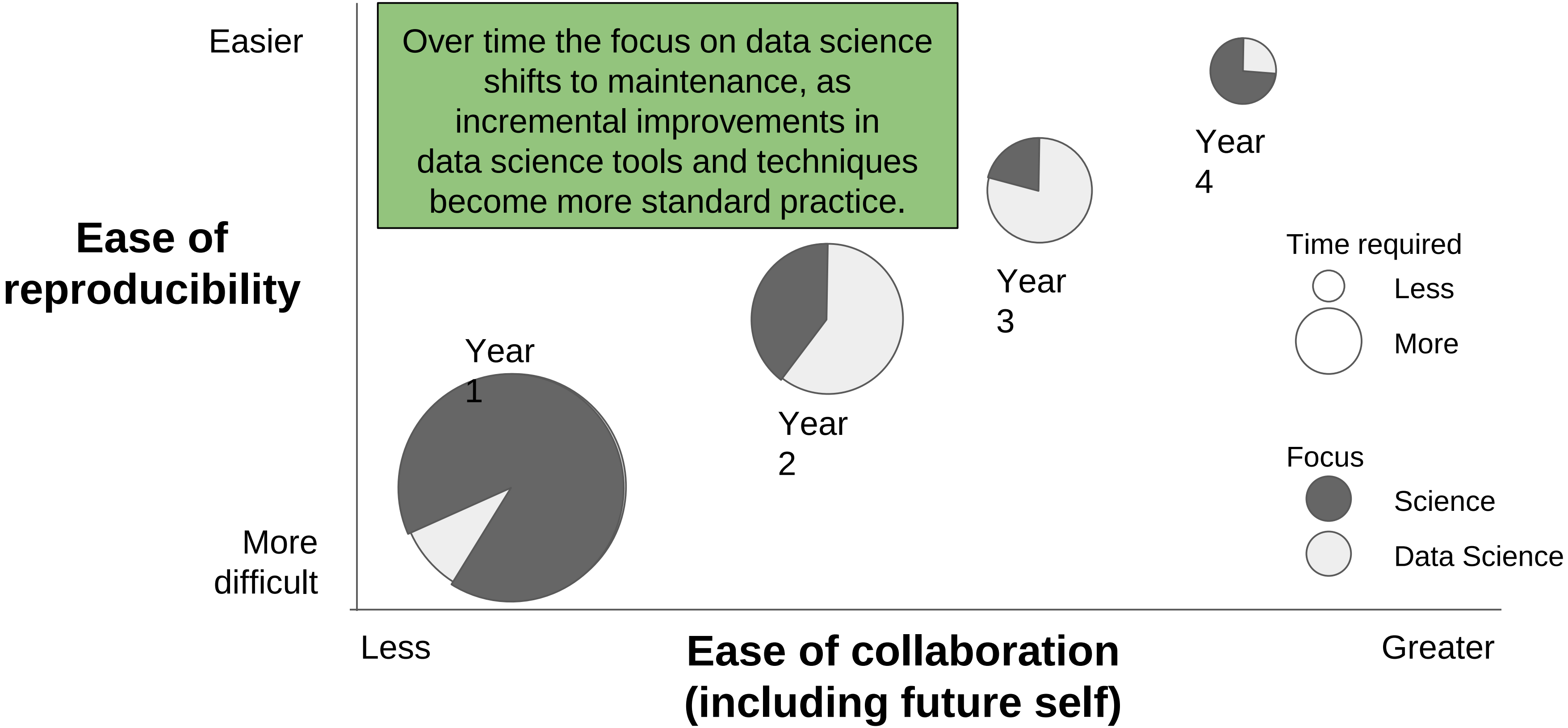
Goal: Better science in less time



A shift in focus to data science should make collaboration and reproducibility easier

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



Over time the focus on data science shifts to maintenance, as incremental improvements in data science tools and techniques become more standard practice.

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.

What we (IfSC and HRRC) need...

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.

TEXAS A&M UNIVERSITY
Institute for Sustainable Communities

HAZARD REDUCTION
& RECOVERY CENTER
TEXAS A&M UNIVERSITY

Thank you!

Nathanael Rosenheim

nroseheim@arch.tamu.edu

TEXAS A&M UNIVERSITY
Institute for Sustainable Communities

HAZARD REDUCTION
& RECOVERY CENTER
TEXAS A&M UNIVERSITY