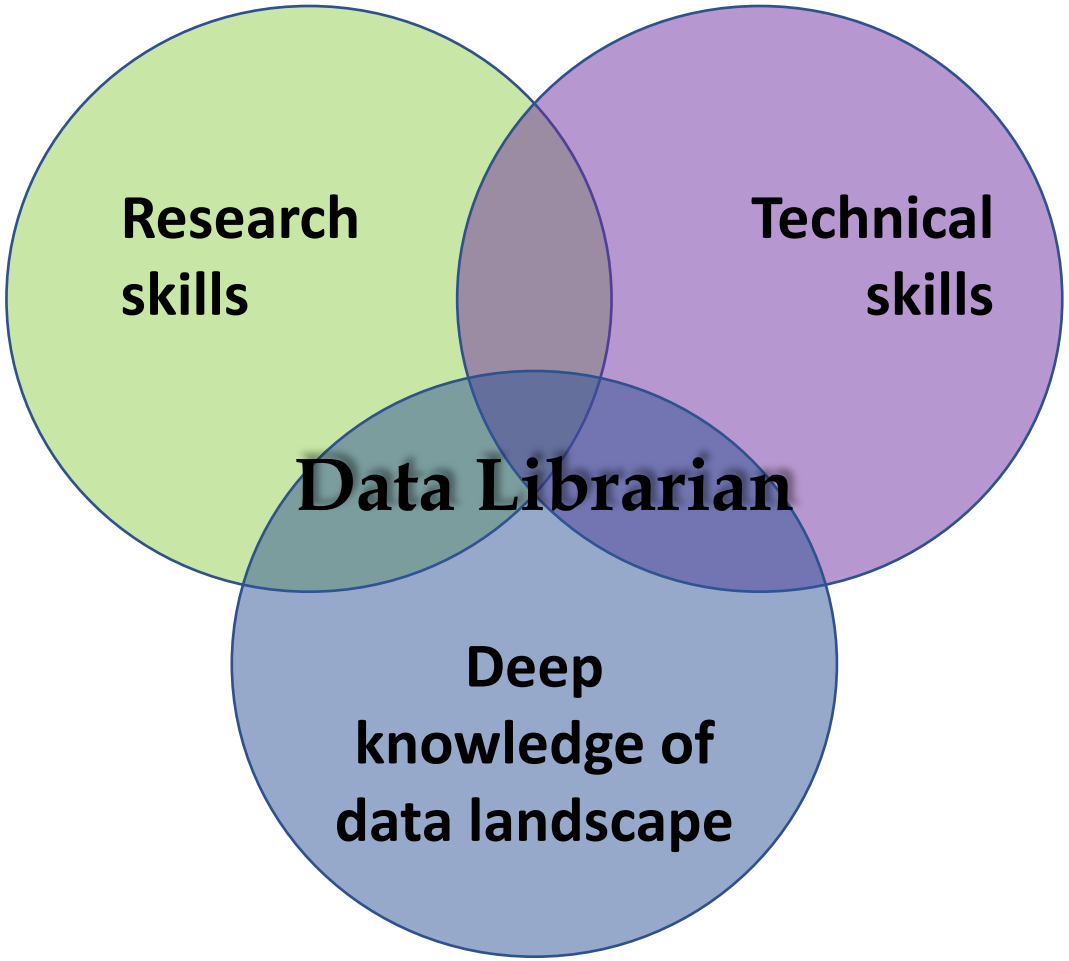
Natalia Festa, MD, MBA
Kaitlin Throgmorton, MLIS

Clin-STAR
Clinician – Scientists Transdisciplinary Aging Research Coordinating Center

Working with a Data Librarian

NIH National Institute on Aging | afar american federation for aging research

What's a Data Librarian?



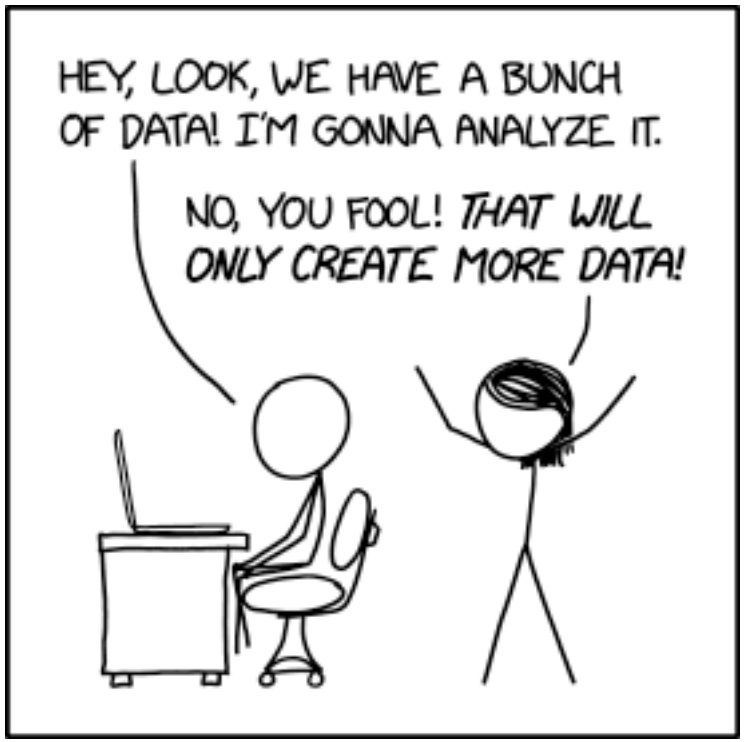


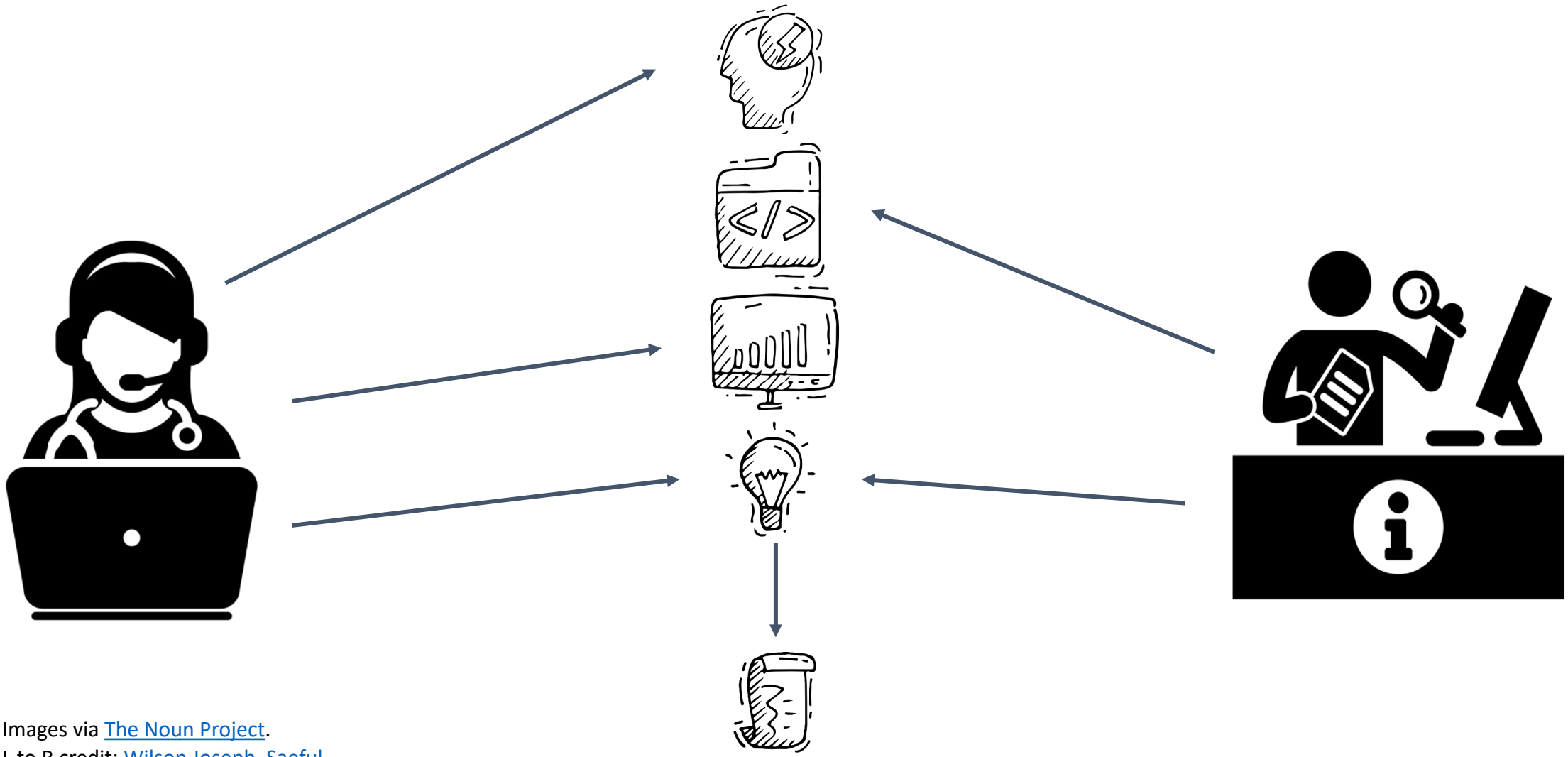
Image via xkcd

Current Work

- Supporting research
- Teaching
- Advising on data policy and infrastructure
  - *For example:* NIH Data Management and Sharing Policy
- Conducting research

Image by Terry Dagradi and courtesy of Cushing/Whitney Medical Library

Clinician-Librarian Collaboration



Images via The Noun Project.
L to R credit: Wilson Joseph, Saeful Muslim, and Miroslav KURDOV.

Example of original, collaborative research

Original Investigation | Geriatrics

January 6, 2023

Association of Nursing Home Exposure to Hurricane-Related Inundation With Emergency Preparedness [1]

Natalia Festa, MD, MBA[1,2]; Kaitlin F. Throgmorton, MLIS[3]; Nora Heaphy, BA[4]; Maureen Canavan, PhD, MPH[5]; Thomas M. Gill, MD[1]

» Author Affiliations | Article Information

JAMA Netw Open. 2023;6(1):e2249937. doi:10.1001/jamanetworkopen.2022.49937

Key Points

Question Are nursing homes exposed to potential hurricane-related inundation more likely to meet Centers for Medicare & Medicaid Services criteria for adequate emergency preparedness?

Findings In this cross-sectional study of 5914 nursing homes, a higher prevalence of emergency preparedness deficiencies among nursing homes exposed to hurricane-related inundation in the Mid-Atlantic region was observed. Exposure status remained positively associated with the presence and number of emergency preparedness deficiencies after adjustment for facility characteristics, with the converse for facilities within the Western Gulf Coast.

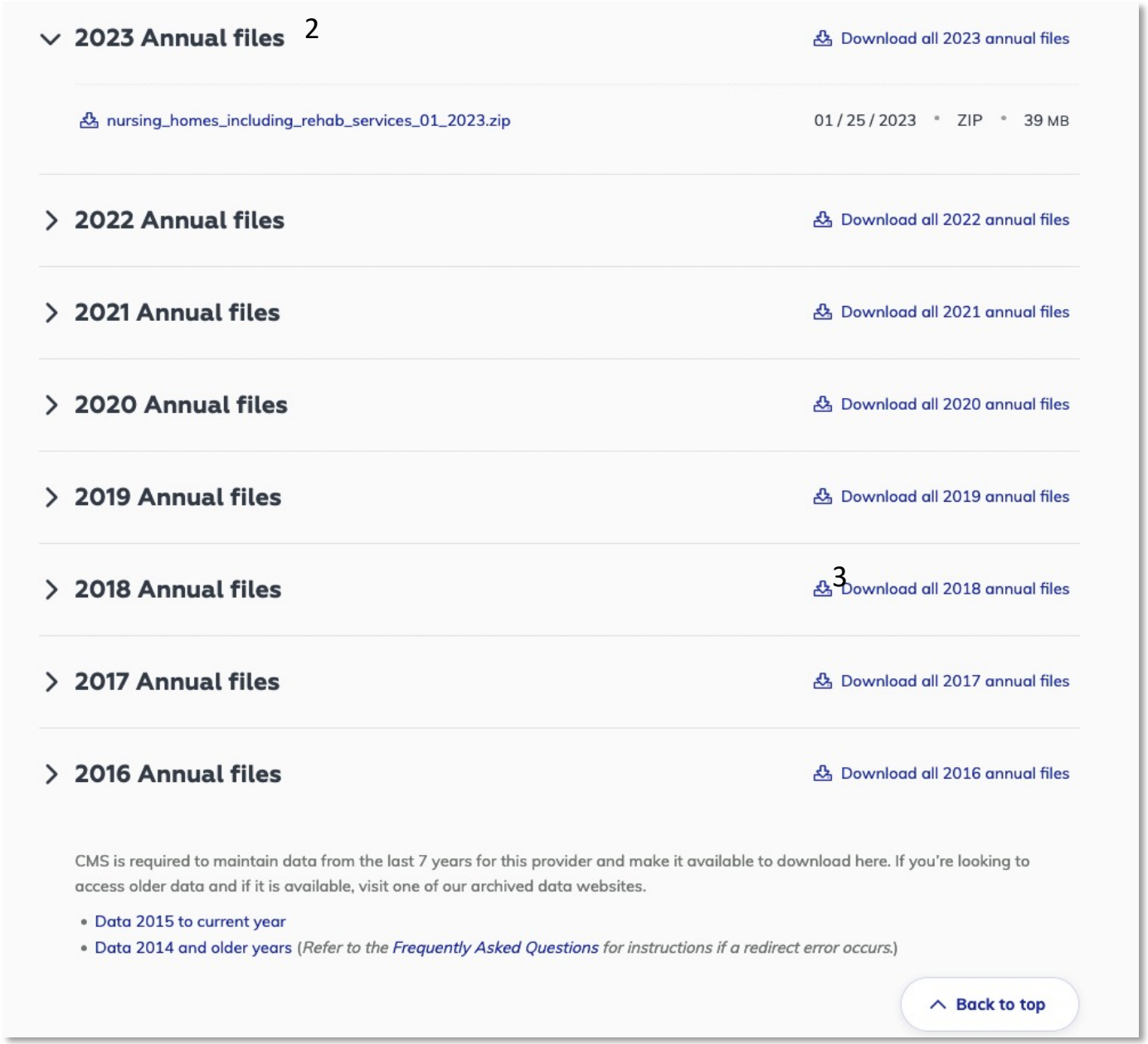
Meaning These findings suggest opportunities to reduce regional heterogeneity and improve the alignment of nursing home emergency preparedness with surrounding environmental risks.

Challenges Along the Way

- Access
- Management
- Processing
- Methodology

Approaches

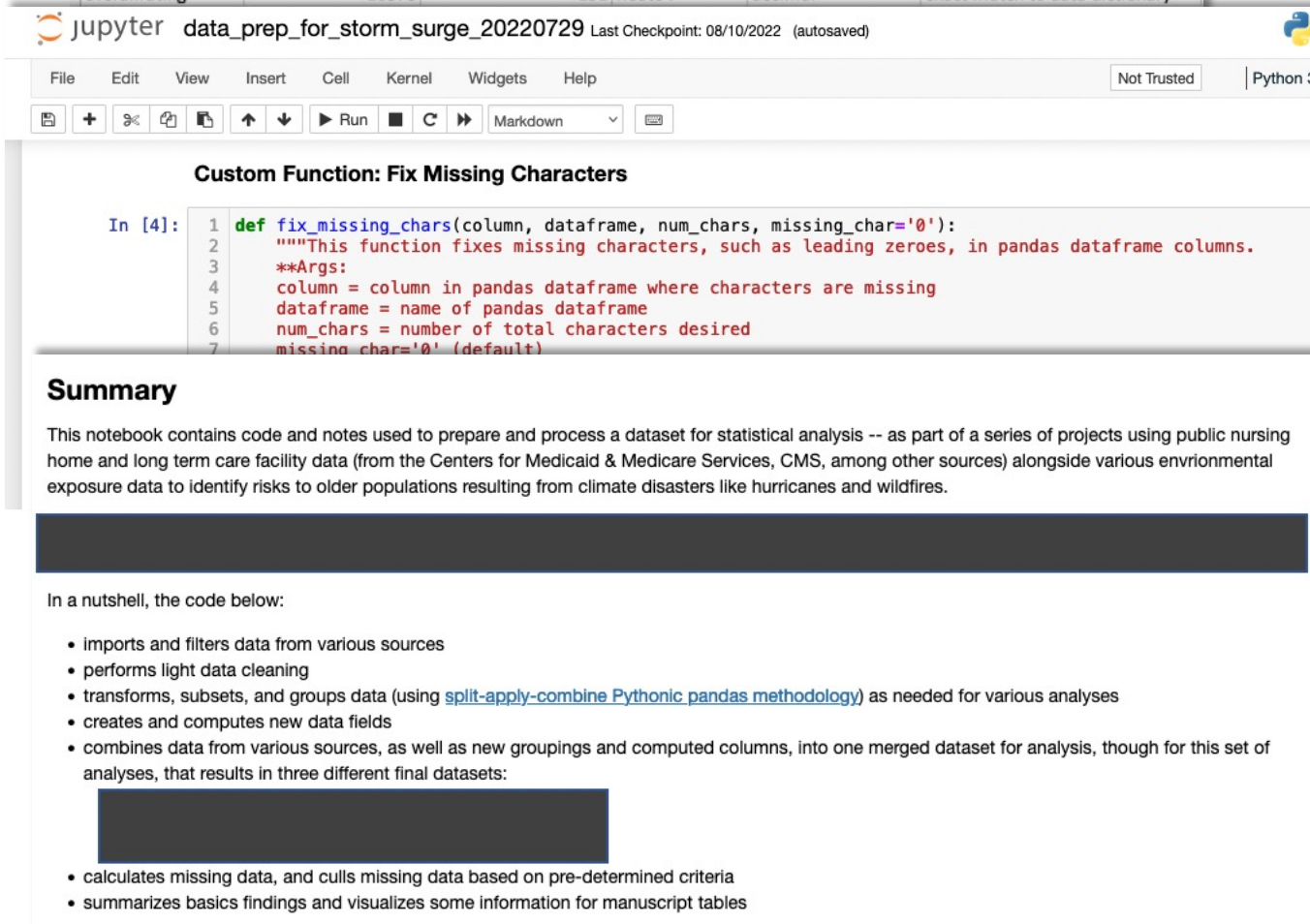
- Interpreting and assessing data retrieval methods, documentation and quality
- Tracking data sourcing and citing data appropriately



Approaches

- Making data dictionaries and missingness tables
- Shifting from tools like Stata to Python
- Converting hard-to-document Excel processes to replicable code
- Documenting work in multiple modalities with Jupyter Notebooks

Column	Non-Null Count	Missing Value Counts	Pandas Dtype	Conventional Type	Variable Description
federalprovidernum	16569	0	object	string	exact match to data dictionary
providername	16502	67	object	string	exact match to data dictionary
provideraddress	16502	67	object	string	exact match to data dictionary
providercity	16502	67	object	string	exact match to data dictionary
providerstate	16502	67	object	string	exact match to data dictionary
providerzipcode	16502	67	object	string	exact match to data dictionary
providercountyname	16502	67	object	string	exact match to data dictionary
ownershiptype	16502	67	object	string	exact match to data dictionary
numberofcertifiedb	16502	67	float64	decimal	exact match to data dictionary
averagenumberofre	16395	174	float64	decimal	exact match to data dictionary
providerresidesinho	16502	67	object	string	exact match to data dictionary
legalbusinessname	16502	67	object	string	exact match to data dictionary
mostrecenthealthin	16502	67	object	string	exact match to data dictionary
automaticsprinklers	16502	67	object	string	exact match to data dictionary
overallrating	16378	191	float64	decimal	exact match to data dictionary



Approaches

- Reviewing data literature and methodology
- Implementing data processing methodologies, such as:
  - tidy data[4]
  - split-apply-combine[5 6]

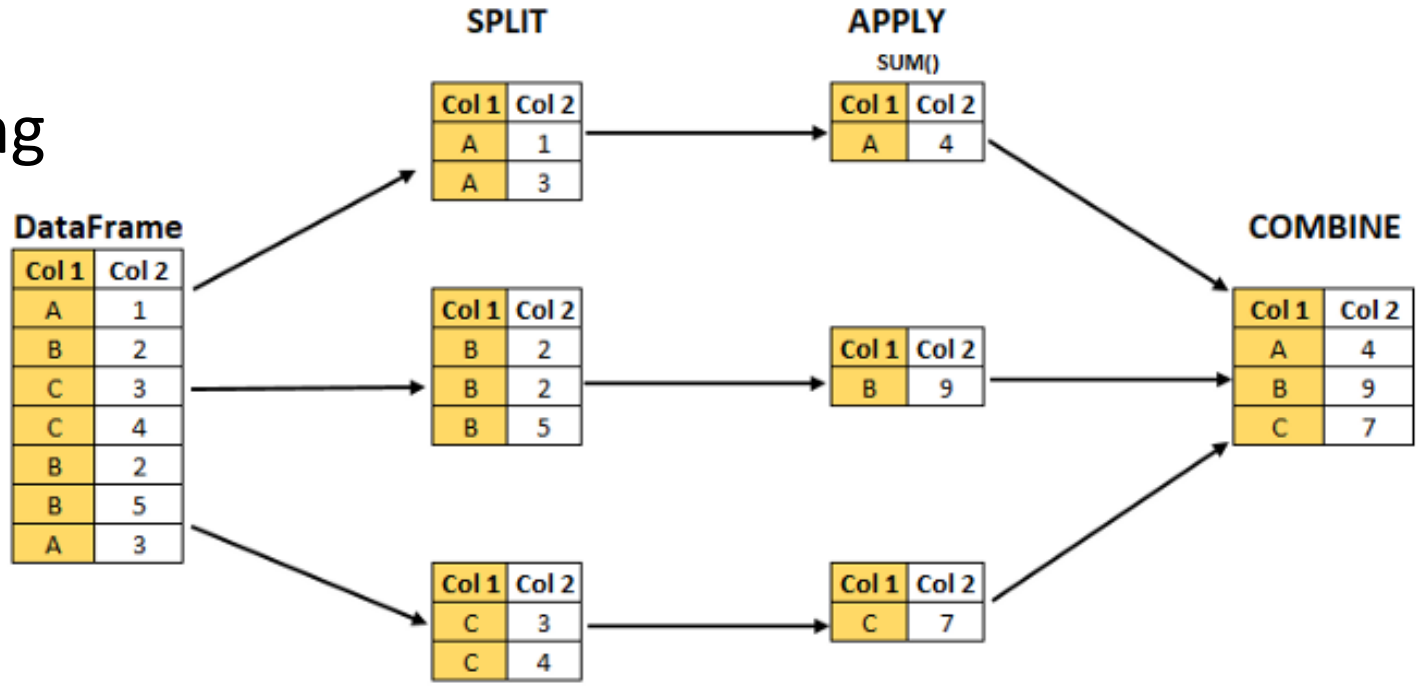


Image via Analytics Vidhya

What we've learned from each other

Clinician perspective:

- The utility of publicly available data to answer questions with clinical and health services relevance
- Public data can present many challenges to cleaning and merging across multiple data sources
  - Unstructured data
  - Data with inconsistent identifiers across years and data sources (healthcare facility identification numbers, for example)
  - Merging data with inconsistent or error-laden identifiers presents many challenges

What we've learned from each other

Clinician perspective:

- Many data sources can be more amenable to cleaning and organizing in languages other than those in which clinician researchers are trained
- Accurately logging procedures used in data pre-processing is essential to replicability
- Data Librarians have rich expertise in addressing each of the challenges above

What we've learned from each other

Librarian perspective:

- Growing clinician ambition to take on complex, multi-disciplinary data projects
- More support needed, especially in early clinician-researcher career, for data training
- Opportunity to:
  - Fully participate in research data lifecycle
  - See the power of open data in action

How to succeed as a clinician-librarian team

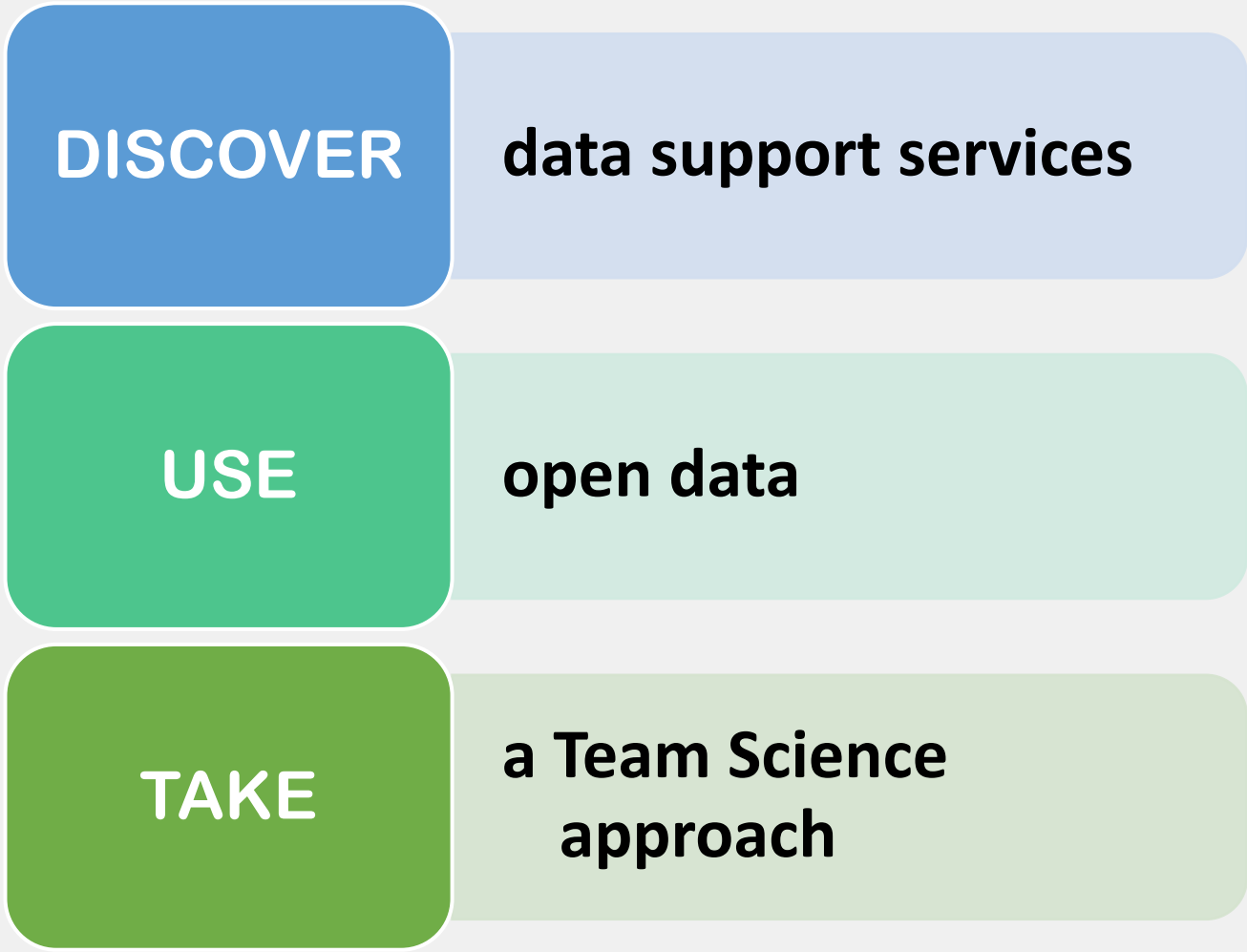
- Communicate early and often
- Grow project portfolio over time
- Teach and trade skills
- Seek expertise elsewhere when needed

Image via unsplash

Advice to clinicians on involving librarians

- Learn what data support is offered by your library
- Learn and practice principles that your data librarian encourages as foundational to data science integrity and reproducibility
- Establish your team and expectations regarding roles and responsibilities during early stages of the project
- Ensure that there are opportunities for learning and professional development for all team members

Summing Up


DISCOVER
data support services
USE
open data
TAKE
a Team Science approach

Resources

- **Data management planning guidance for NIH grants and beyond:**
  - Ten simple rules for maximizing the recommendations of the NIH data management and sharing plan | *PLOS Computational Biology*
  - The FAIR Guiding Principles for scientific data management and stewardship| *Scientific Data*
  - Ten simple rules for the care and feeding of scientific data | *PLOS Computational Biology*
- **Excellent data organization and cleaning principles:**
  - Data organization in spreadsheets | *The American Statistician*
- **For those interested in Python:**
  - Python for Non-Programmers | python.org

References

1. Festa, N., Throgmorton, K. F., Heaphy, N. M., Canavan, M., & Gill, T. M. (2023). Association of nursing home exposure to hurricane-related inundation with emergency preparedness. *JAMA Network Open 6*(1):e2249937. https://doi.org/10.1001/jamanetworkopen.2022.49937
2. Centers for Medicare and Medicaid. Provider Information. https://data.cms.gov/provider-data/dataset/4pq5-n9py
3. LTCFocus Public Use Data sponsored by the National Institute on Aging (P01 AG027296) through a cooperative agreement with the Brown University School of Public Health. Available at www.ltcfocus.org. https://doi.org/10.26300/h9a2-2c26
4. Wickham, H. (2014). Tidy Data. *Journal of Statistical Software*, *59*(10), 1–23. https://doi.org/10.18637/jss.v059.i10
5. Pandey, A. (2018). Split-Apply-Combine Strategy for Data Mining. *Medium*. https://medium.com/analytics-vidhya/split-apply-combine-strategy-for-data-mining-4fd6e2a0cc99
6. Pandas. (n.d.) Group by: split-apply-combine. https://pandas.pydata.org/pandas-docs/stable/user_guide/groupby.html