Evaluating, Analyzing and Framing Data

How to know if the math maths

Maggie Lee
Feb. 24, 2026
for the Ida B. Wells Society Investigative Reporting Fellowship

Links & sources at: https://github.com/maggie-lee/IBWS-2026-slides

Evaluating, Analyzing and Framing Data

What we will do:

- Discuss what where to find data and figure out if it's trustworthy
- Discuss why and how to be *constructively paranoid*
- Discuss data's role in research vs. in the published piece

What we will NOT do:

- Discuss specific software, scraping techniques, creating maps and graphs. (That's later on.)

Finding data

Anything that is...

- Inspected (restaurants, amusement parks, buildings)
- Licensed (doctors, taxis, pets, hunting)
- Enforced (speeding tickets, subway stops)
- Purchased (contracts, Medicare claims data, etc.)
- Reported (311 calls, etc.)

... will probably be in a government-held database somewhere and belongs to us, “the public.”

- Privately-gathered data can be harder to get and evaluate
- Making your own dataset

Knowing what to look for

- You need to figure out:
  - What kind of data might be helpful to your reporting project?
  - What data like that exists?
  - Who has it?
  - What is it called?
  - How can you get it?
- To answer those questions: REPORT!
  - Find and call sources
    - experts/academics/researchers
    - other reporters
    - government staff
    - people who work in the field (lawyers, teachers, business people, etc)
  - The internet (hold this thought)

Does a dataset exist?

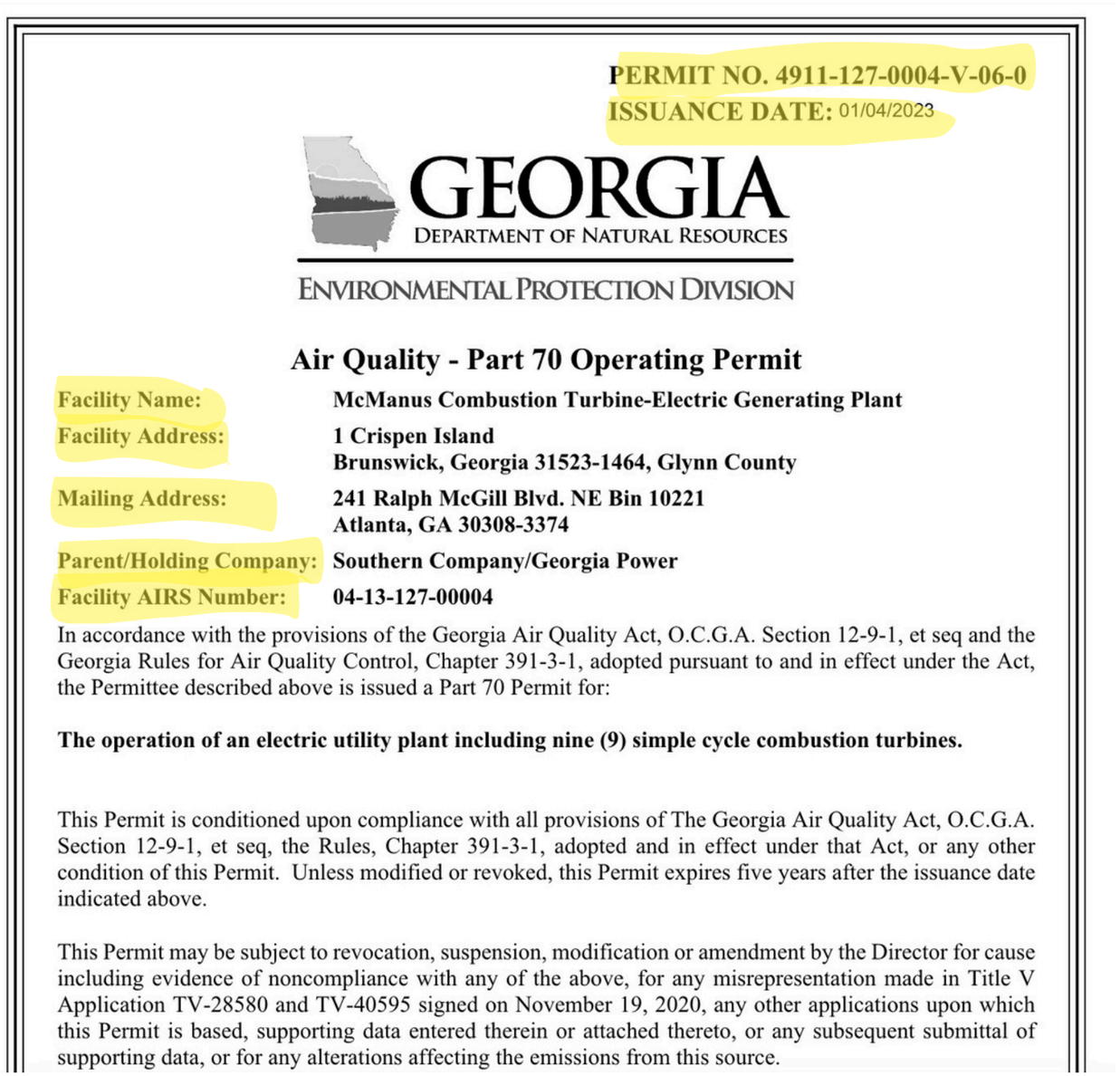
Evidence that a dataset exists:

- A form (digital or paper), with a specific title and/or form number

- An online lookup tool

- Published reports & research by academics/reporters/advocates etc.

A form with a title and a number



PERMIT NO. 4911-127-0004-V-06-0
ISSUANCE DATE: 01/04/2023

GEORGIA
DEPARTMENT OF NATURAL RESOURCES
ENVIRONMENTAL PROTECTION DIVISION

Air Quality - Part 70 Operating Permit

Facility Name:	**McManus Combustion Turbine-Electric Generating Plant**
Facility Address:	**1 Crispen Island** **Brunswick, Georgia 31523-1464, Glynn County**
Mailing Address:	**241 Ralph McGill Blvd. NE Bin 10221** **Atlanta, GA 30308-3374**
Parent/Holding Company:	**Southern Company/Georgia Power**
Facility AIRS Number:	**04-13-127-00004**

In accordance with the provisions of the Georgia Air Quality Act, O.C.G.A. Section 12-9-1, et seq and the Georgia Rules for Air Quality Control, Chapter 391-3-1, adopted pursuant to and in effect under the Act, the Permittee described above is issued a Part 70 Permit for:

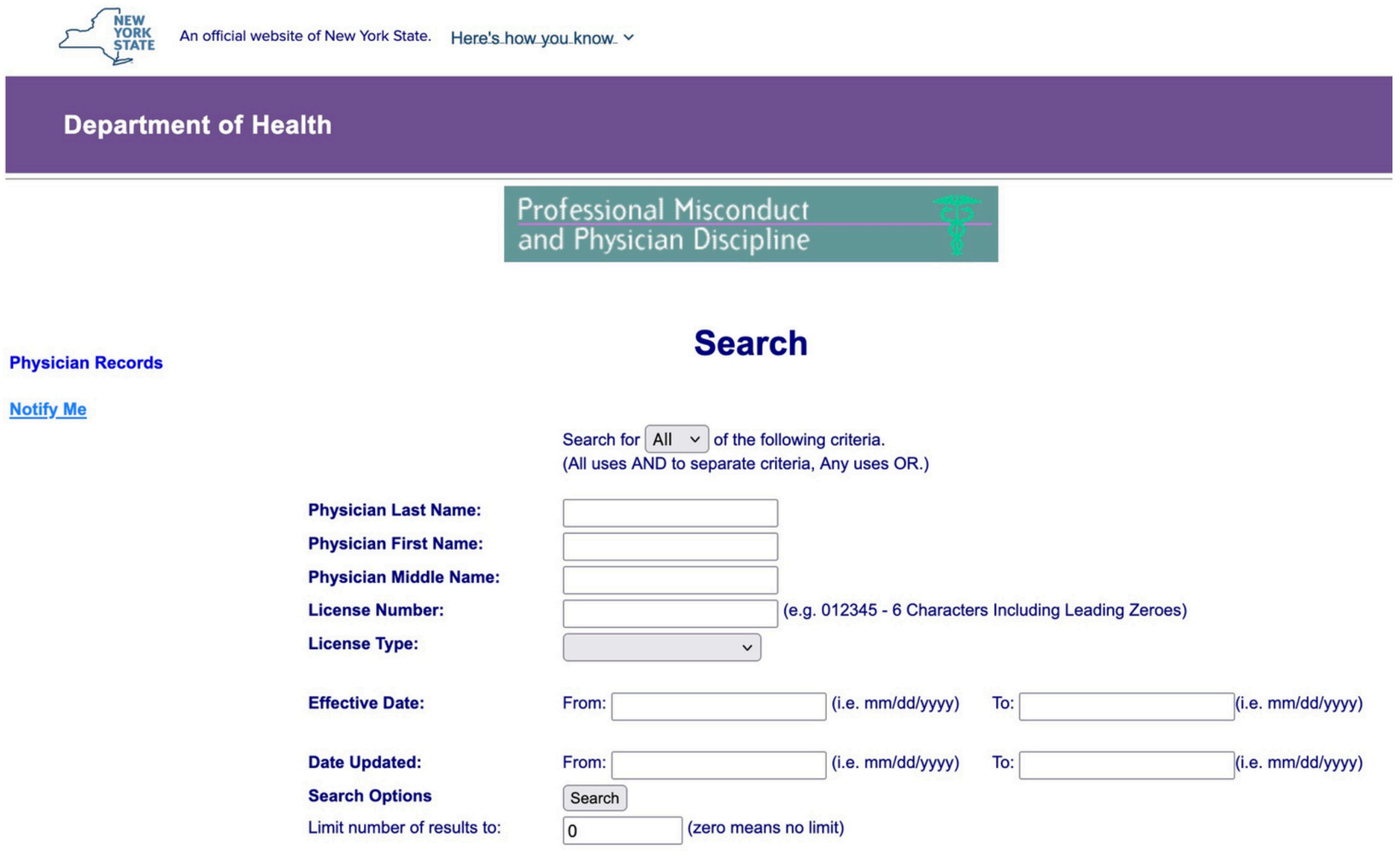
The operation of an electric utility plant including nine (9) simple cycle combustion turbines.

This Permit is conditioned upon compliance with all provisions of The Georgia Air Quality Act, O.C.G.A. Section 12-9-1, et seq, the Rules, Chapter 391-3-1, adopted and in effect under that Act, or any other condition of this Permit. Unless modified or revoked, this Permit expires five years after the issuance date indicated above.

This Permit may be subject to revocation, suspension, modification or amendment by the Director for cause including evidence of noncompliance with any of the above, for any misrepresentation made in Title V Application TV-28580 and TV-40595 signed on November 19, 2020, any other applications upon which this Permit is based, supporting data entered therein or attached thereto, or any subsequent submittal of supporting data, or for any alterations affecting the emissions from this source.

An online lookup tool

[NYS Physician Discipline]

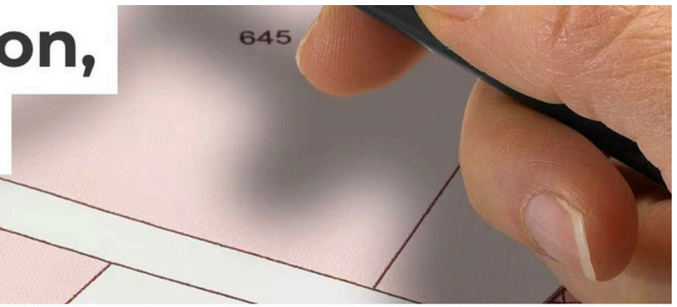


Published reports

The Conversation, 2026-02-20

Colorado has high levels of radon, which can cause lung cancer – here's how to lower your risk

Published: February 20, 2026 8:37am EST



Radon exposure is the leading cause of lung cancer for people who have never used tobacco. Francesco Scatena/iStock via Getty Images

In Colorado, as of 2025, about 500 people a year die from lung cancer as the result of radon gas exposure. Nationally, the number of lung cancer deaths attributed to radon is about 21,000 per year.

Author

Jan Lowery
Professor of Epidemiology, Colorado School of Public Health, University of Colorado Anschutz Medical Campus

Data on the internet. Is it legit?

Evaluate it like you would an article:

- Look for human names and email addresses that indicate an author
- Look for an “about” page
- Look for footnotes and definitions
- Are other media or experts citing this source?
- If it seems designed to make you angry, scared or open your wallet, that’s suspicious.

Very nice!

DC Inbox

About

DCinbox is academically derived project headed by me, Prof. Lindsey Cormack, with the assistance of 2 NYU undergrads, 8 Stevens Institute of Technology undergrads, and 1 professional Technical Lead.

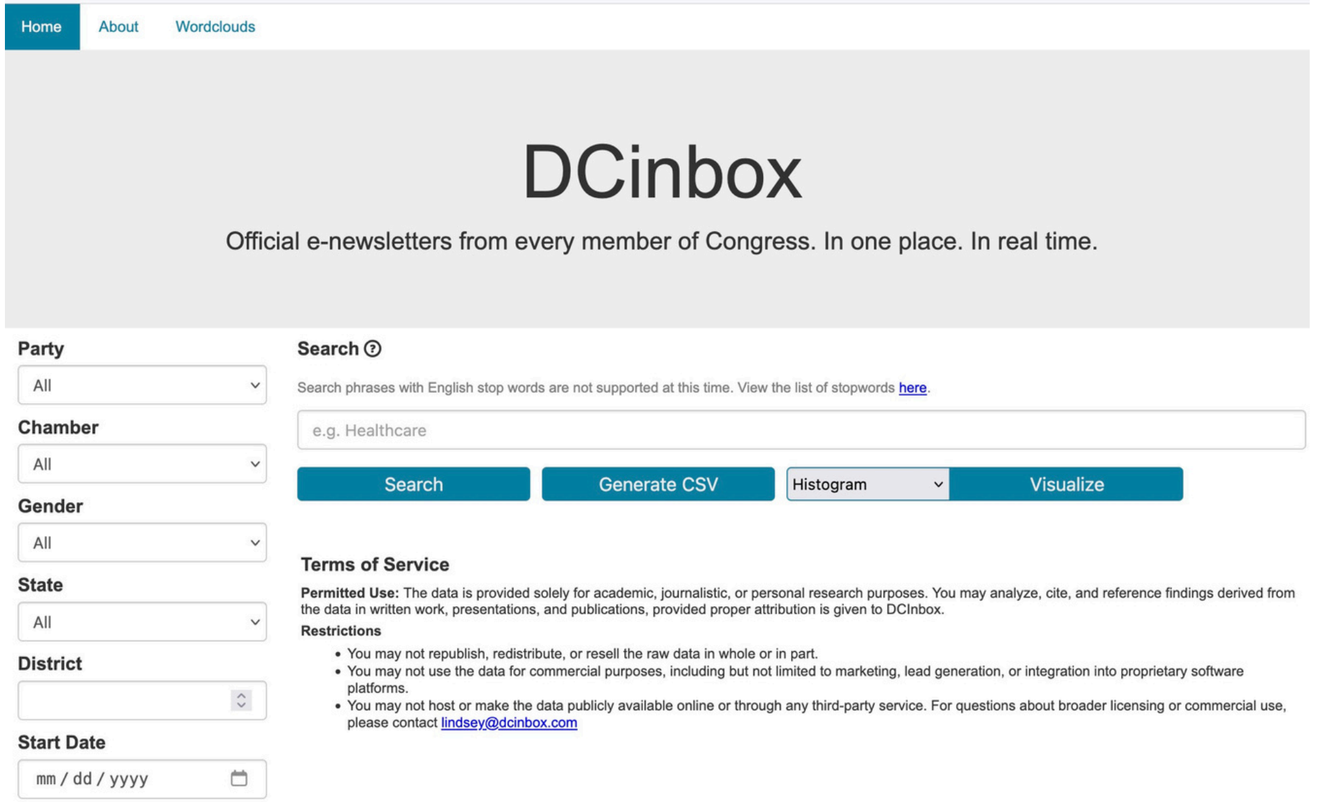
Starting in 2009, as a graduate student at the NYU Wilf family department of politics, I set up a small database to capture and house every official e-newsletter sent by sitting members of the U.S. House and Senate. For a detailed description of the ins and outs of the project, please see this from The Legislative Scholar.

These communications are sent by nearly every member of Congress, yet until 2009 there was no comprehensive set to use in research. This database is updated in real time, as members of Congress send constituent messages, and provides researchers with a systematic set of texts to examine the strategic choices that legislators and their staff make in crafting the content. Today, there are 213,320 unique e-newsletters.

If you are a researcher interested in these data please use what you like here. A static set of downloads of the data are available month by month here. If you ever have questions along the way, please feel free to email me at lcormack@stevens.edu. You can also check out @dcinbox on Twitter to see where this research is headed.

To date, these data have been used to explore:

- Partisan differences in attention to veterans as a topic in official communications. Republicans tend to talk way more about veterans in official communications than Democrats do, despite the fact that Democrats offer more veteran focused legislation
- How members of Congress try to shift their ideological appearances to constituents by attempting to moderate or extremize their communicated votes based on district preferences
- The differences in vote revelation strategies of men and women within Congress. Punch line: women discuss their votes in constituent communications more often than men do
- How members of different parties discuss environmental issues
- How members of Congress reacted to President Trump's trade policy

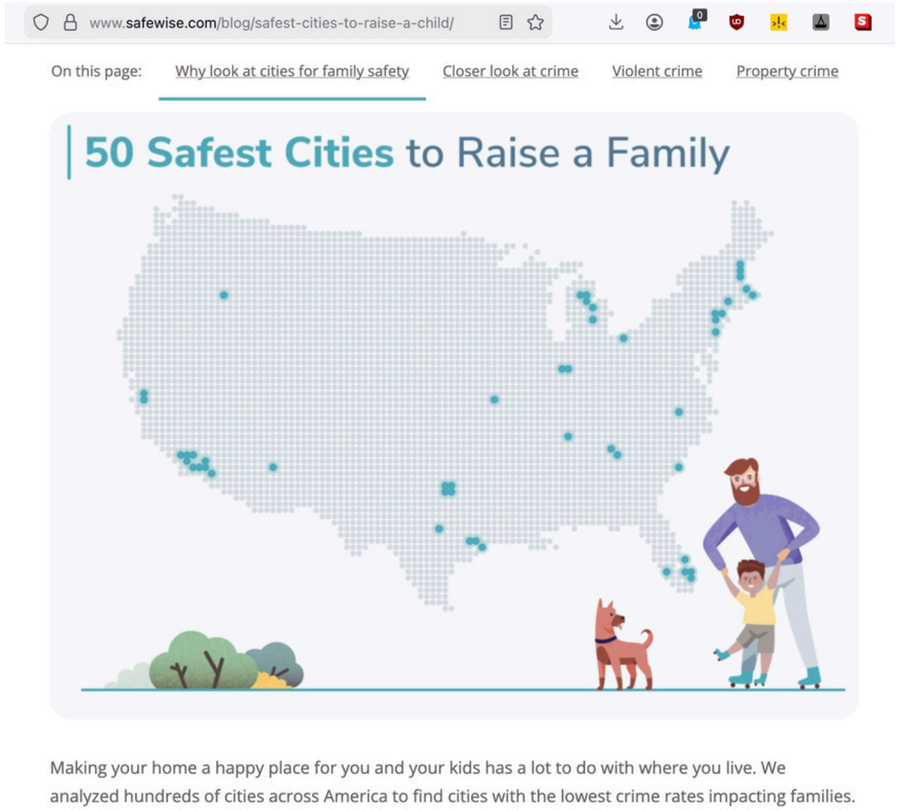


Background ! Contact info!
References ! ❤️

No!

"Safety" "rankings" from company selling home security cameras

- Fear is a lucrative business
- And the FBI does publish stats
- The FBI itself says don't use these stats for ranking. A responsible statistician wouldn't do it.
- FBI stats are missing important stuff anyway like drug trafficking and white collar crime
- Probably most crimes aren't reported



Constructive paranoia

Poking holes is job one

- Do Not Trust Your Data.
- Even (especially?) if you made it.
- Data comes from people.
  - People make mistakes.
  - What people choose to measure or not measure says something
- First law of data entry: the more you type, the more mistakes you make.

PUTTING NUMBERS IN A TABLE DOESN'T MAKE THEM TRUE

You must interview the data

Poking holes = “data validation”

Look out for:

- missing values, blanks, zeroes, N/A
- inconsistent data types (numbers where you expect letters & vice versa)
- out-of-range values (including dates)
- suspicious repeats / duplicates
- misspellings / typos
- truncation within cells
- entire rows truncated from your dataset
- blanks and zeroes and N/A (“Zero” is sometimes used for “I don’t know”)

The Quartz Guide to Bad Data is your friend!

Your sheriff boasts in a speech that his office “arrested 13 foreign nationals from January through April 2025 and ICE took them away.” He gives you this for receipts. What do you think?

	SO#	Name	Country	Arrest Date	Rel Date	Charge	Hold	notified	answered	Arrest Agcy
2	O21243	Holly Mahoney	Bahamas	1/17/27	1/31/25	battery				GAP
3	O24574	Mahoney	Bahamas	2/22/25	2/22/25	aggrevated assault	Y	Y	Y	ICE
4	O22548	Camden Mccann	Guam	3/4/25	3/6/25	reckless	7	Y	Y	GCSO
5	24680	Payton West	Chile	3/14/25	3/14/25	posession of a schedule	N	Y	N	GCPD
6	24681	Kristin Zavala	Chile	3/14/24	3/14/25	terroristic threats	N	Y	N	GCPD
7	O24394	Emerson Waller	China	1/2/25			Y			GCSO
8	O24796	Myah Houston	Colombia	4/7/25	4/7/25	no insurance				GCPD
9	O24632	Sofia Beck	Columbia		3/5/25	posession with intent	N	Y	N	GCPD
10	24695	Justin Bradford	Cuba	3/16/25	3/17/25	HL	N	Y	N	GCPD
11	O13197	Charlie Mccarty	Dominican Re	1/17/25	2/11/25	homicide	N	Y	Y	GCSO
12	O24511	Dahlia Santiago	El Salvador	1/17/25	2/9/25	erroneous release	N	N	N	GSP
13	O24512	Dahlia Santiaog	Guatemala	1/17/25	2/9/25	fugitive	N	N	N	MAR

It’s valid, now what does it mean?

- Speak to an expert such as:
  - The author of the data ⭐
  - A power-user of the data
    - Demographer
    - Prof
    - Reporter
    - Lawyer
    - Activist

You are working on aquifer use in Georgia. The state sends you the list of aquifer withdrawal permit holders as of 2024, the last full year for which data is available. I've excerpted ten large ones. What do you think?

GW Withdrawal Permit Number	Georgia Municipal & Industrial GW Withdrawal Permit Holder	Georgia County	Permit Status	Use1	Permitted Number of Wells	CURRENT Permit Limit Annual Avg (mgd)	CUURENT Permit Limit Monthly Avg (mgd)	Year	Calculated ANNUAL Avg (mgd)
151-0001	Rayonier Performance Fibers, LLC - Jesup Plant	Wayne	Active	Industrial	19	68.000	74.000	2024	58.584
063-0003	Brunswick Cellulose, LLC	Glynn	Active	Industrial	8	45.000	49.000	2024	31.418
025-0018	Savannah, City of	Chatham	Active	Municipal	48	17.962	38.050	2024	20.297
076-0006	Houston County Water - Feagin Mill / Elberta / Dunbar System	Houston	Active	Municipal	16	23.460	31.430	2024	17.761
047-0002	Albany, City of Utility Board	Dougherty	Active	Municipal	37	24.000	36.000	2024	12.891
025-0009	International Paper - Savannah Plant	Chatham	Active	Industrial	5	12.157	23.600	2024	11.449
092-0001	Packaging Corporation of America	Lowndes	Active	Industrial	6	13.700	14.600	2024	11.191
089-0001	Interstate Paper, LLC	Liberty	Active	Industrial	3	11.562	12.000	2024	10.863
076-0002	Warner Robins, City of	Houston	Active	Municipal	15	10.800	14.500	2024	8.829
092-0004	Valdosta, City of	Lowndes	Active	Municipal	10	15.300	19.100	2024	7.782

Why use data?

- When data or the lack thereof is the story
- To provide context
- To debunk
- To show your receipts & build your credibility