

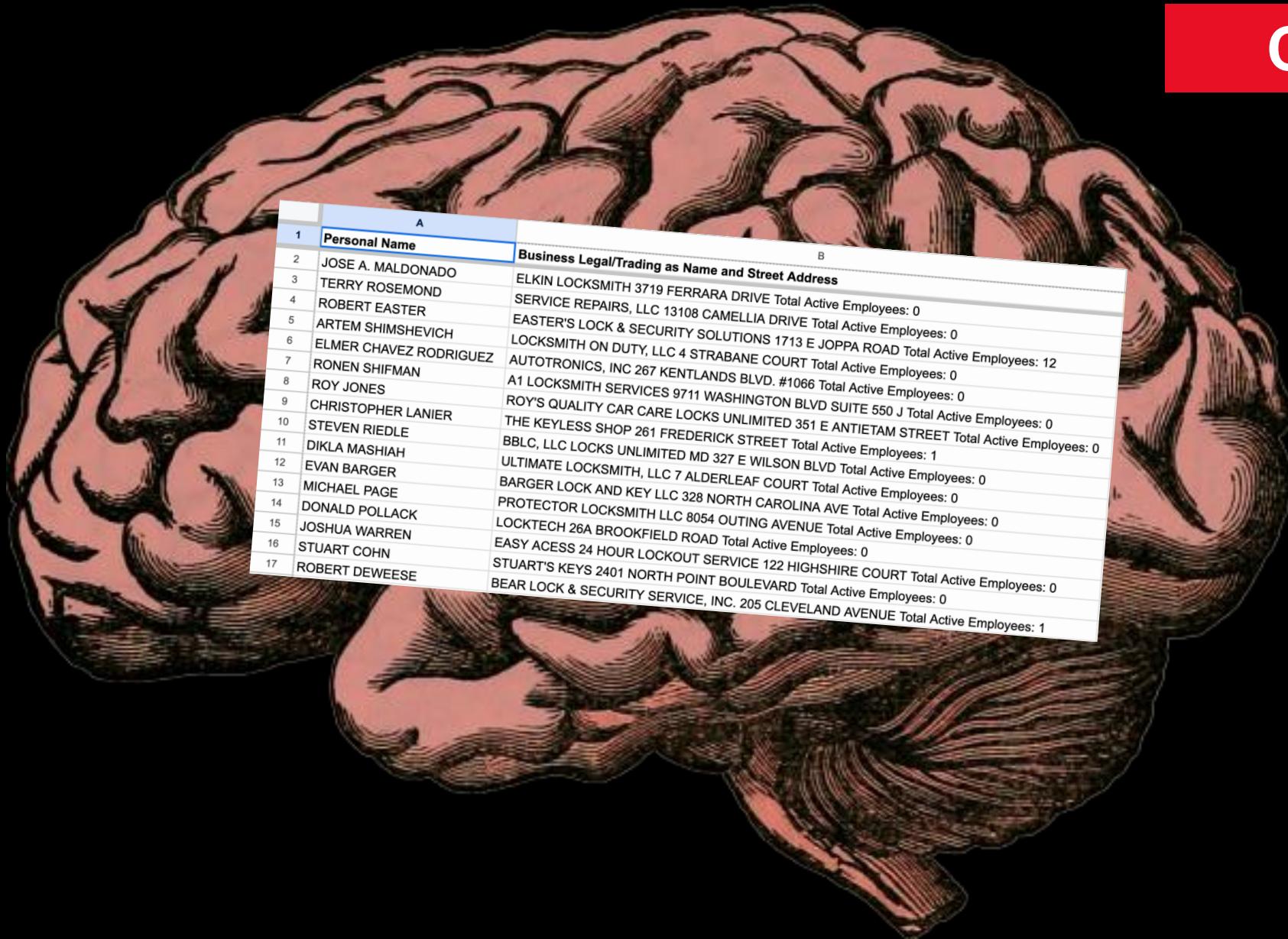
Jonathan Soma
Knight Chair in Data Journalism
Columbia Journalism School
js4571@columbia.edu

Visit bit.ly/ire25-cleaning

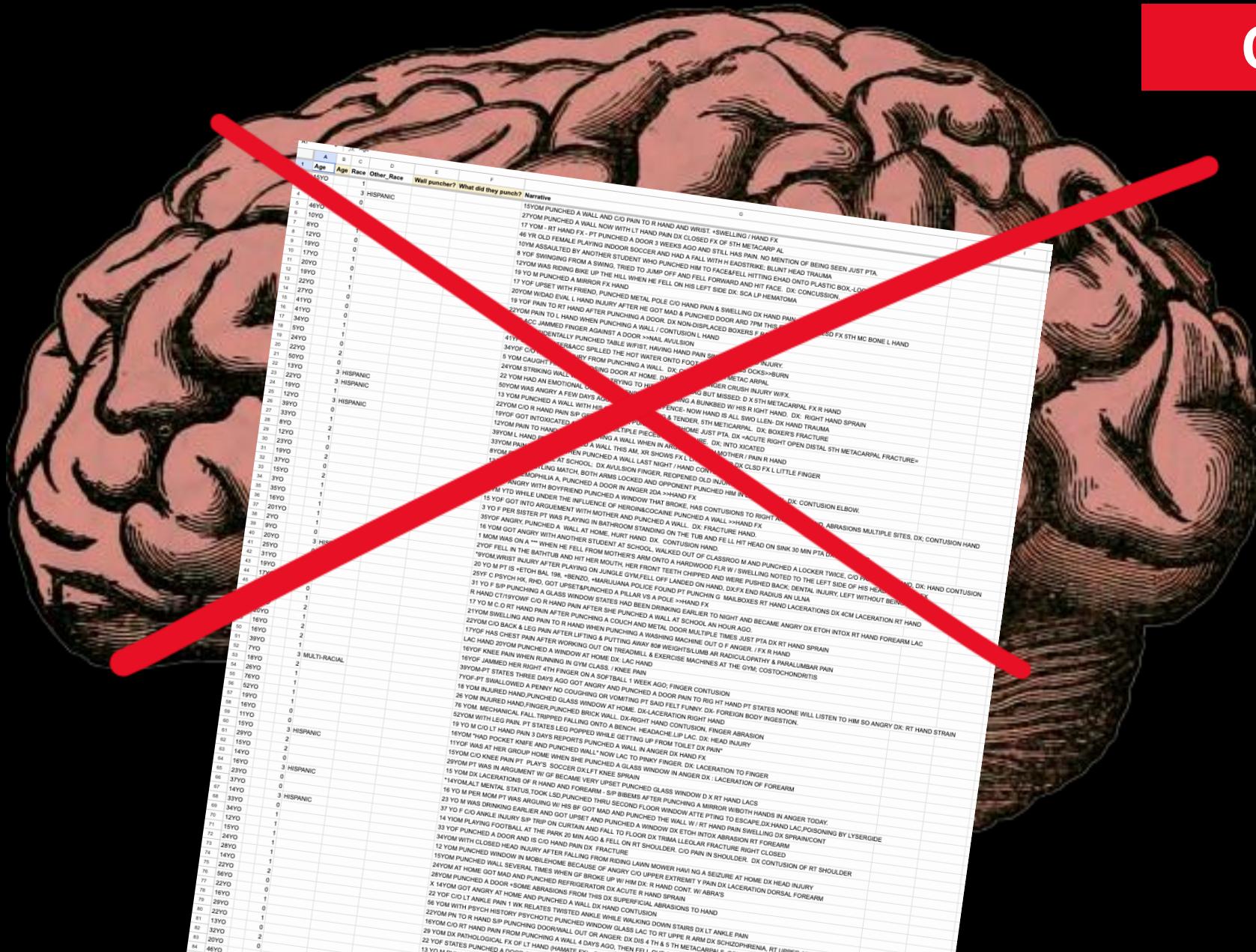
jonathansoma.com/everything

Why not AI?

Context window



Context window



Age	Race	Other_Race	E	F
1	3 HISPANIC			
2	0			
3	15YO			
4	15YD			
5	46YO			
6	12YO			
7	19YO			
8	12YO			
9	19YO			
10	17YO			
11	20YO			
12	19YO			
13	22YO			
14	27YO			
15	41YO			
16	14YO			
17	34YO			
18	5YO			
19	24YO			
20	22YO			
21	59YO			
22	13YO			
23	22YO			
24	3 HISPANIC			
25	19YO			
26	39YO			
27	33YO			
28	8YO			
29	12YO			
30	23YO			
31	19YO			
32	37YO			
33	15YO			
34	35YO			
35	36YO			
36	16YO			
37	201YO			
38	2YO			
39	9YO			
40	20YO			
41	25YO			
42	31YO			
43	18YO			
44	47YO			
45	0			
46	16YO			
47	16YO			
48	16YO			
49	7YO			
50	3 MULTI-RACIAL			
51	28YO			
52	54			
53	16YO			
54	18YO			
55	78YO			
56	52YO			
57	19YO			
58	16YO			
59	11YO			
60	15YO			
61	29YO			
62	3 HISPANIC			
63	18YO			
64	14YO			
65	23YO			
66	37YO			
67	14YO			
68	33YO			
69	34YO			
70	12YO			
71	13YO			
72	24YO			
73	28YO			
74	14YO			
75	22YO			
76	56YO			
77	22YO			
78	16YO			
79	28YO			
80	20YO			
81	13YO			
82	32YO			
83	20YO			
84	46YO			
85	13YO			
86	15YO			
87	15YO			
88	15YO			
89	15YO			
90	15YO			
91	15YO			
92	15YO			
93	15YO			
94	15YO			
95	15YO			
96	15YO			
97	15YO			
98	15YO			
99	15YO			
100	15YO			
101	15YO			
102	15YO			
103	15YO			
104	15YO			
105	15YO			
106	15YO			
107	15YO			
108	15YO			
109	15YO			
110	15YO			
111	15YO			
112	15YO			
113	15YO			
114	15YO			
115	15YO			
116	15YO			
117	15YO			
118	15YO			
119	15YO			
120	15YO			
121	15YO			
122	15YO			
123	15YO			
124	15YO			
125	15YO			
126	15YO			
127	15YO			
128	15YO			
129	15YO			
130	15YO			
131	15YO			
132	15YO			
133	15YO			
134	15YO			
135	15YO			
136	15YO			
137	15YO			
138	15YO			
139	15YO			
140	15YO			
141	15YO			
142	15YO			
143	15YO			
144	15YO			
145	15YO			
146	15YO			
147	15YO			
148	15YO			
149	15YO			
150	15YO			
151	15YO			
152	15YO			
153	15YO			
154	15YO			
155	15YO			
156	15YO			
157	15YO			
158	15YO			
159	15YO			
160	15YO			
161	15YO			
162	15YO			
163	15YO			
164	15YO			
165	15YO			
166	15YO			
167	15YO			
168	15YO			
169	15YO			
170	15YO			
171	15YO			
172	15YO			
173	15YO			
174	15YO			
175	15YO			
176	15YO			
177	15YO			
178	15YO			
179	15YO			
180	15YO			
181	15YO			
182	15YO			
183	15YO			
184	15YO			
185	15YO			
186	15YO			
187	15YO			
188	15YO			
189	15YO			
190	15YO			
191	15YO			
192	15YO			
193	15YO			
194	15YO			
195	15YO			
196	15YO			
197	15YO			
198	15YO			
199	15YO			
200	15YO			
201	15YO			
202	15YO			
203	15YO			
204	15YO			
205	15YO			
206	15YO			
207	15YO			
208	15YO			
209	15YO			
210	15YO			
211	15YO			
212	15YO			
213	15YO			
214	15YO			
215	15YO			
216	15YO			
217	15YO			
218	15YO			
219	15YO			
220	15YO			
221	15YO			
222	15YO			
223	15YO			
224	15YO			
225	15YO			
226	15YO			
227	15YO			
228	15YO			
229	15YO			
230	15YO			
231	15YO			
232	15YO			
233	15YO			
234	15YO			
235	15YO			
236	15YO			
237	15YO			
238	15YO			
239	15YO			
240	15YO			
241	15YO			
242	15YO			
243	15YO			
244	15YO			
245	15YO			
246	15YO			
247	15YO			
248	15YO			
249	15YO			
250	15YO			
251	15YO			
252	15YO			
253	15YO			
254	15YO			
255	15YO			
256	15YO			
257	15YO			
258	15YO			
259	15YO			
260	15YO			
261	15YO			
262	15YO			
263	15YO			
264	15YO			
265	15YO			
266	15YO			
267	15YO			
268	15YO			
269	15YO			
270	15YO			
271	15YO			
272	15YO			
273	15YO			
274	15YO			
275	15YO			
276	15YO			
277	15YO			
278	15YO			
279	15YO			
280	15YO			
281	15YO			
282	15YO			
283	15YO			
284	15YO			
285	15YO			
286	15YO			
287	15YO			
288	15YO			
289	15YO			
290	15YO			
291	15YO			
292	15YO			
293	15YO			
294	15YO			
295	15YO			
296	15YO			
297	15YO			
298	15YO			
299	15YO			
300	15YO			
301	15YO			
302	15YO			
303	15YO			
304	15YO			
305	15YO			
306	15YO			
307	15YO			
308	15YO			
309	15YO			
310	15YO			
311	15YO			
312	15YO			
313	15YO			
314	15YO			
315	15YO			
316	15YO			
317	15YO			
318	15YO			
319	15YO			
320	15YO			
321	15YO			
322	15YO			
323	15YO			
324	15YO			
325	15YO			
326	15YO			
327	15YO			
328	15YO			
329	15YO			
330	15YO			
331	15YO			
332	15YO			
333	15YO			
334	15YO			
335	15YO			
336	15YO	</td		

- It gets things wrong
- You won't double-check it
- It makes your brain rot away and you never learn new skills
- Okay but actually we WILL use AI
- It gets things wrong!!!

Basic concepts

replace and split are
99% of the game

66 years old

66 years old

66

convert to number

replace “years old”
with... nothing!

j soma, js4571@columbia.edu, @dangerscarf

j soma | js4571@columbia.edu | @dangerscarf

js4571@columbia.edu



j soma, js4571@columbia.edu, @dangerscarf

j soma | js4571@columbia.edu | @dangerscarf

take second

replace with nothing

js4571@columbia.edu

take second

columbia.edu

split

split on @

js4571 | columbia.edu

I have text like "j soma, js4571@columbia.edu, @dangerscarf" in a Google Sheet and I want to split it by ", ", take the second part, then replace the "@columbia.edu" with nothing. Just give me the formula.

**just ask!
then TEST IT**



You can use the following Google Sheets formula to achieve this:

plaintext

Copy code

```
=SUBSTITUTE(INDEX(SPLIT(A1, ", "), 2), "@columbia.edu", "")
```

Explanation:

1. `SPLIT(A1, ", ")` splits the text in cell `A1` by ", ".
2. `INDEX(SPLIT(A1, ", "), 2)` selects the second part of the split text.
3. `SUBSTITUTE(INDEX(SPLIT(A1, ", "), 2), "@columbia.edu", "")` removes the "@columbia.edu" part from the selected text.

*just ask!
then TEST IT*

I have text like "j soma, js4571@columbia.edu, @dangerscarf" in a Google Sheet and I want to extract the "@columbia.edu" part. Write me a formula to do this. Only show me the formula.



excel

Copy code

```
=RIGHT(A1,LEN(A1)-FIND("@",A1)+1)
```



Feel free to reach out at 555-555-5555



take the final 12 letters

555-555-5555



just ask!
then TEST IT

8

I have text like "j soma, js4571@columbia.edu, @dangerscarf" in a pandas dataframe column named "details" and I want to split it by ", ", take the second part, then replace the "@columbia.edu" with nothing. Try not to use .apply. Don't create the df, just give me the code.

< 6/8 >



Certainly! Here's the code to achieve that:

python

Copy code

```
# Split the 'details' column by ", " and take the second part
df['email'] = df['details'].str.split(', ').str[1]

# Replace "@columbia.edu" with nothing
df['email'] = df['email'].str.replace('@columbia.edu', '', regex=False)
```

This will split the column and then replace the domain part as required.



just ask!
then TEST IT

I have text like "j soma, js4571@columbia.edu, @dangerscarf" in a pandas dataframe column named "details" and I want to extract the "@columbia.edu" part. Try not to use .apply. Don't create the df, just give me the code.



You can use the `str.extract` method to extract the "@columbia.edu" part from the "details" column. Here's the code to do that:

python

Copy code

```
df['extracted'] = df['details'].str.extract(r'(\b@\w+\.\edu\b)')
```



Regular expressions

regex is the other
other 1%

66 years old

hug what we want
(capture group)

(66) years old

match anything

(.*) years old

I am 66 years old

I am (66) years old

I am (.*) years old

I am (\d*) years old

- match anything
- hug what we want
- demand numbers
- hug what we want

Contact me at 555-555-5555

Contact me at (.*)

Contact me at 555-555-5555 or
js4571@columbia.edu

Contact me at (\d\d\d-\d\d\d-\d\d\d\d)

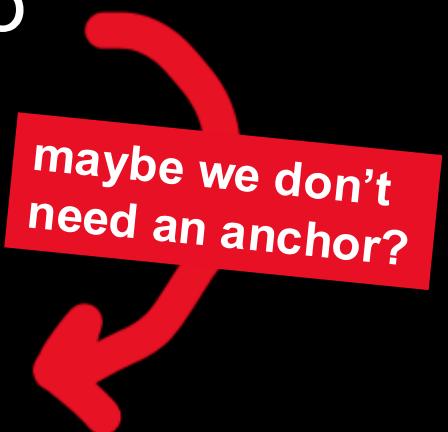
but maybe not!

hug what we want

just numbers

Contact me at 555-555-5555

(\d\d\d-\d\d\d-\d\d\d\d)



maybe we don't
need an anchor?

Contact me at 5555555555

(\d\d\d-?\d\d\d-?\d\d\d\d)



Contact me at 555+555+5555

(\d\d\d.?\d\d\d.?\d\d\d\d)

*any character is fine!
(and also optional)*

Google Sheets

```
REGEXMATCH("Spreadsheets", "S.r")
```

```
REGEXEXTRACT("I ate 24 cakes 2 days ago", "\d+")
```

```
REGEXREPLACE("Spreadsheets", "S.*d", "Bed")
```

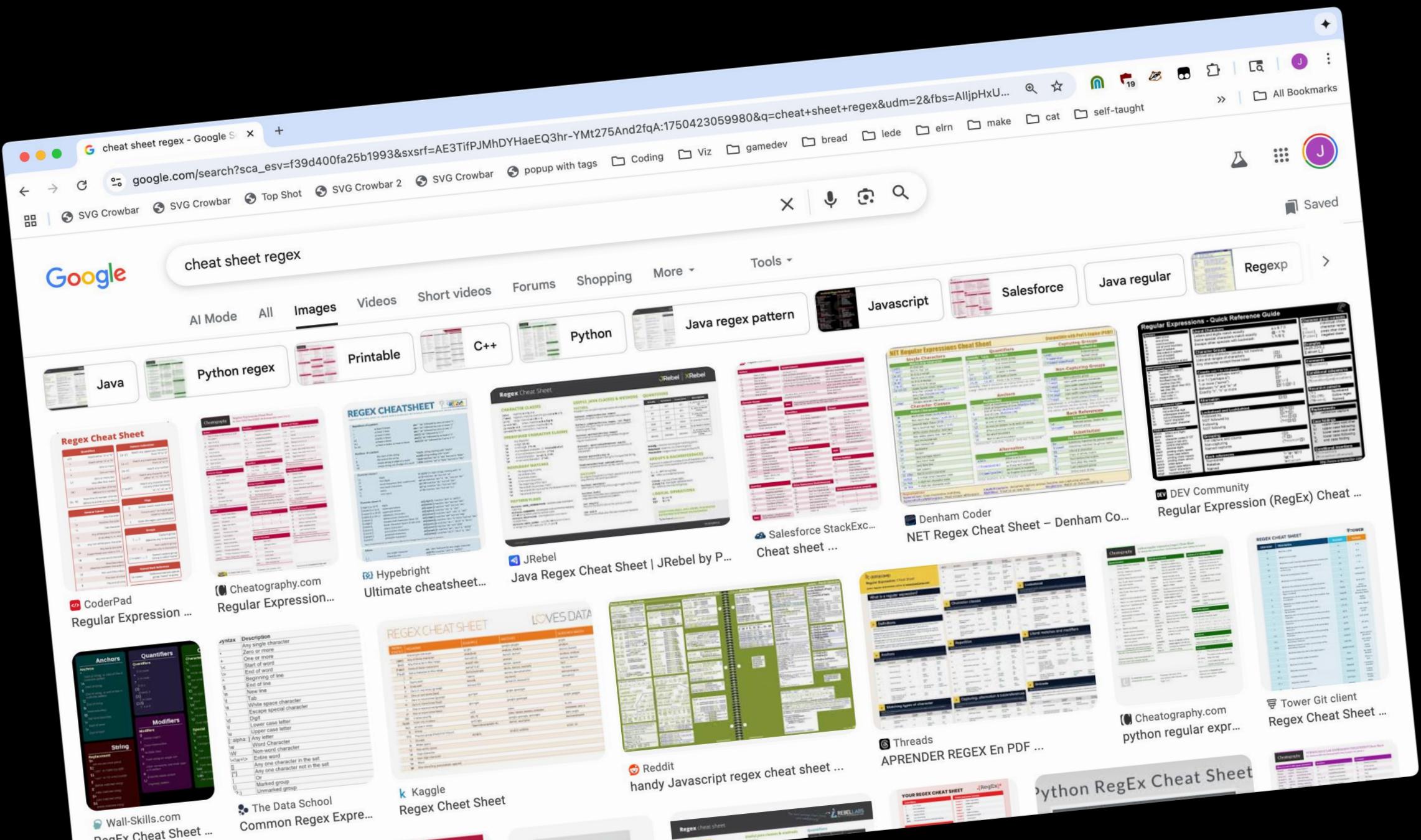
Python (pandas)

```
df['colname'].str.extract("(\\d*)")
```

(and many more)

Google Sheets

- `.*` - literally anything
- `\d` - digits
- `\w+` - more than one “word” characters
- `[A-Za-z]` - all letters
- `\s` - whitespace
- `\b\w{5}\b` - 5-letter words
- `^https?:\/\/\S+` - URLs



I have a table that looks like this. Write the SIMPLEST POSSIBLE regex to just extract the values for "Deforested area" in km. Maybe anchor around the year at the start of each line?

Deforestation of the Amazon rainforest[9][10][11]		
Year	Deforested area	Accumulated loss
2005	1,144 km2 (7,341 sq mi)	332,470 km2 (128,370 sq mi)
	6 km2 (5,516 sq mi)	346,756 km2 (133,883 sq mi)
	km2 (4,498 sq mi)	358,407 km2 (138,382 sq mi)
	1,111 km2 (4,985 sq mi)	371,318 km2 (143,367 sq mi)
2009	7,464 km2 (2,882 sq mi)	378,782 km2 (146,249 sq mi)
2010	7,000 km2 (2,700 sq mi)	385,782 km2 (148,951 sq mi)
2011	6,418 km2 (2,478 sq mi)	392,200 km2 (151,400 sq mi)
2012	4,571 km2 (1,765 sq mi)	396,771 km2 (153,194 sq mi)
2013	5,891 km2 (2,275 sq mi)	402,662 km2 (155,469 sq mi)
2014	4,848 km2 (1,872 sq mi)	407,510 km2 (157,340 sq mi)

just ask!
then TEST IT

< 5/5 >

To extract the "Deforested area" values in km² using a regex, you can use the following simple pattern:

scss

[Copy code](#)

```
^\d{4}\s+(\d{1,5})\s+km2
```

Here's a breakdown of the regex:

- `^` asserts the position at the start of the line.
- `\d{4}` matches exactly four digits (the year).
- `\s+` matches one or more whitespace characters.
- `(\d{1,5})` captures between 1 and 5 digits (the deforested area value) in a capturing group.
- `\s+` matches one or more whitespace characters.
- `km2` matches the literal string "km2".

This pattern will capture the deforested area in km² from each line.



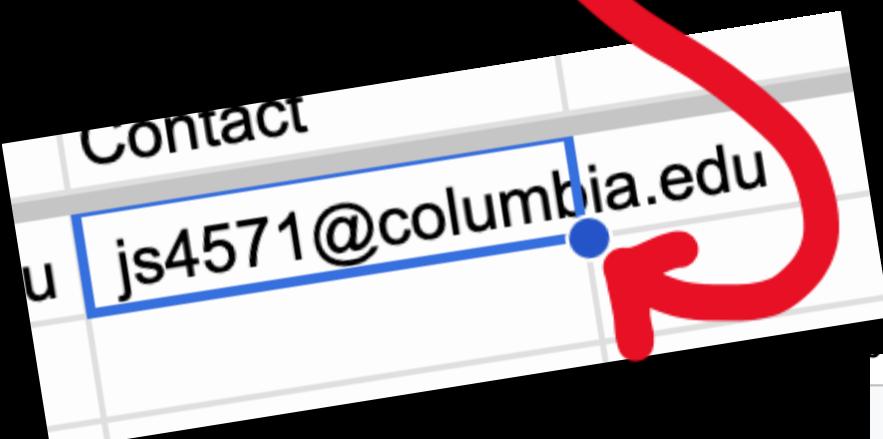
Doing things multiple times

=SUBSTITUTE(A2, "Contact me at", "")

A	B	C
Phrase	Contact	
Contact me at js4571@columbia.edu	js4571@columbia.edu	

click + drag

autofill



A	B	C	D	E
Phrase	Contact			
Contact me at js4571@columbia.edu	js4571@columbia.edu			
Contact me at 555-555-5555	555-555-5555			
Contact me at home	home			
Contact me at 555.555.5556	555.555.5556			
Contact me at 911	911			
Contact me at hello@example.com	hello@example.com			

AUTO FILL

Suggested autofill

⌘+Enter to Autofill. Show formula



=ARRAYFORMULA(....)

2

fx =ARRAYFORMULA(SUBSTITUTE(A2:A7, "Contact me at", ""))

A

B

D

Phrase

Contact

Contact me at js4571@columbia.edu

js4571@columbia.edu

Contact me at 555-555-5555

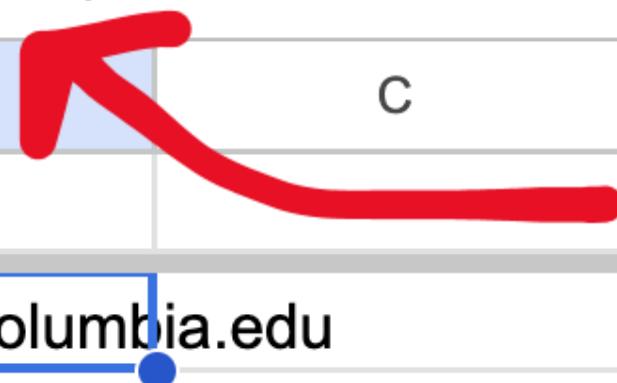
555-555-5555

Contact me at home

home

Contact me at 555.555.5556

555.555.5556



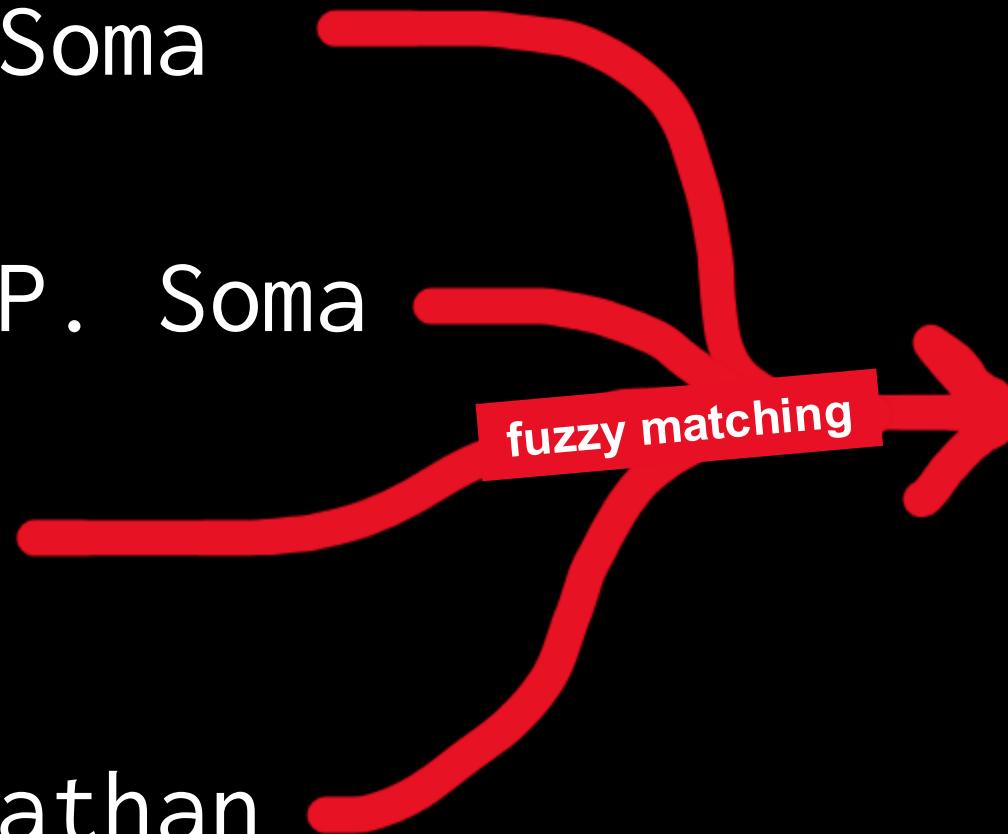
Fuzzy matching

Jonathan Soma

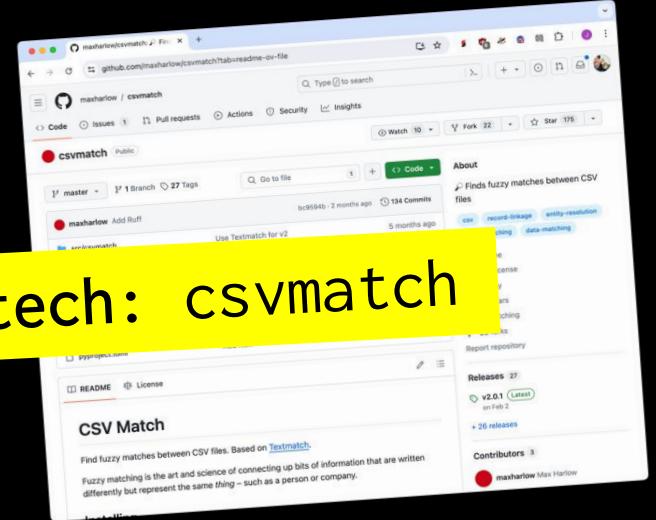
Jonathan P. Soma

Jon Soma

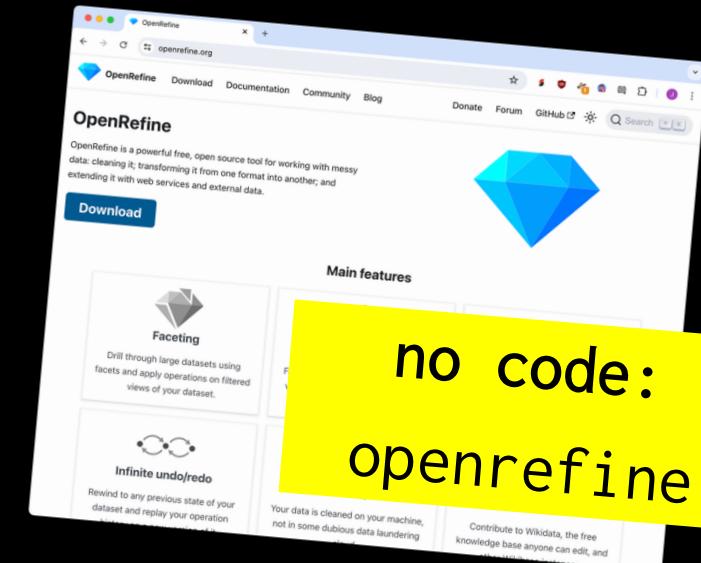
Soma, Jonathan



tech: csvmatch

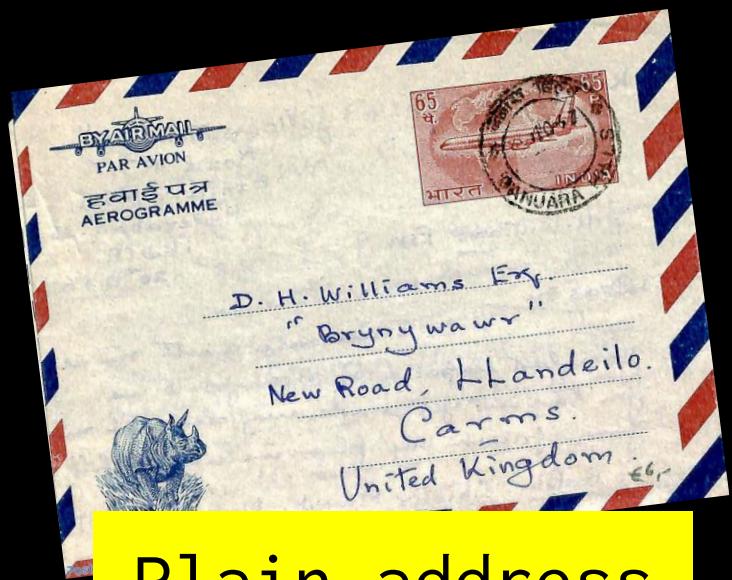


Jonathan Soma



no code:
openrefine

Addresses and maps



Plain address

Geocoding



Latitude/Longitude

Reverse geocoding

Hassle-free geocoding

UPLOAD SPREADSHEET GET AN API KEY

Straightforward and easy-to-use geocoding, reverse geocoding, and data matching for US and Canadian addresses.

drag and drop: geocod.io

Used by thousands of happy customers

ORACLE AI
Stanford PRO

Geocoder: Simple, Consistent

Release v1.38.1. [\(Installation\)](#)

Simple and consistent geocoding library written in Python. Many online providers such as Google & Bing have geocoding services, these providers do not include Python libraries and have different JSON responses between each other. It can be very difficult sometimes to parse a particular geocoding provider since each one of them have their own JSON schema.

Here is a [tiny example](#).

Lat & Lng from address

python: geocoder

```
    'com/maps/api/geocode/json'
    'address': 'Mountain View, CA'
  ]
  [
    'location'
  ]
```

v: master

Geocode by Awesome Table

Uninstall

Geocode is a tool that helps you get latitudes & longitudes from addresses in a Google Sheet to display them on a map you can share.

By: Talarian Listing updated: April 1, 2024

Works with:

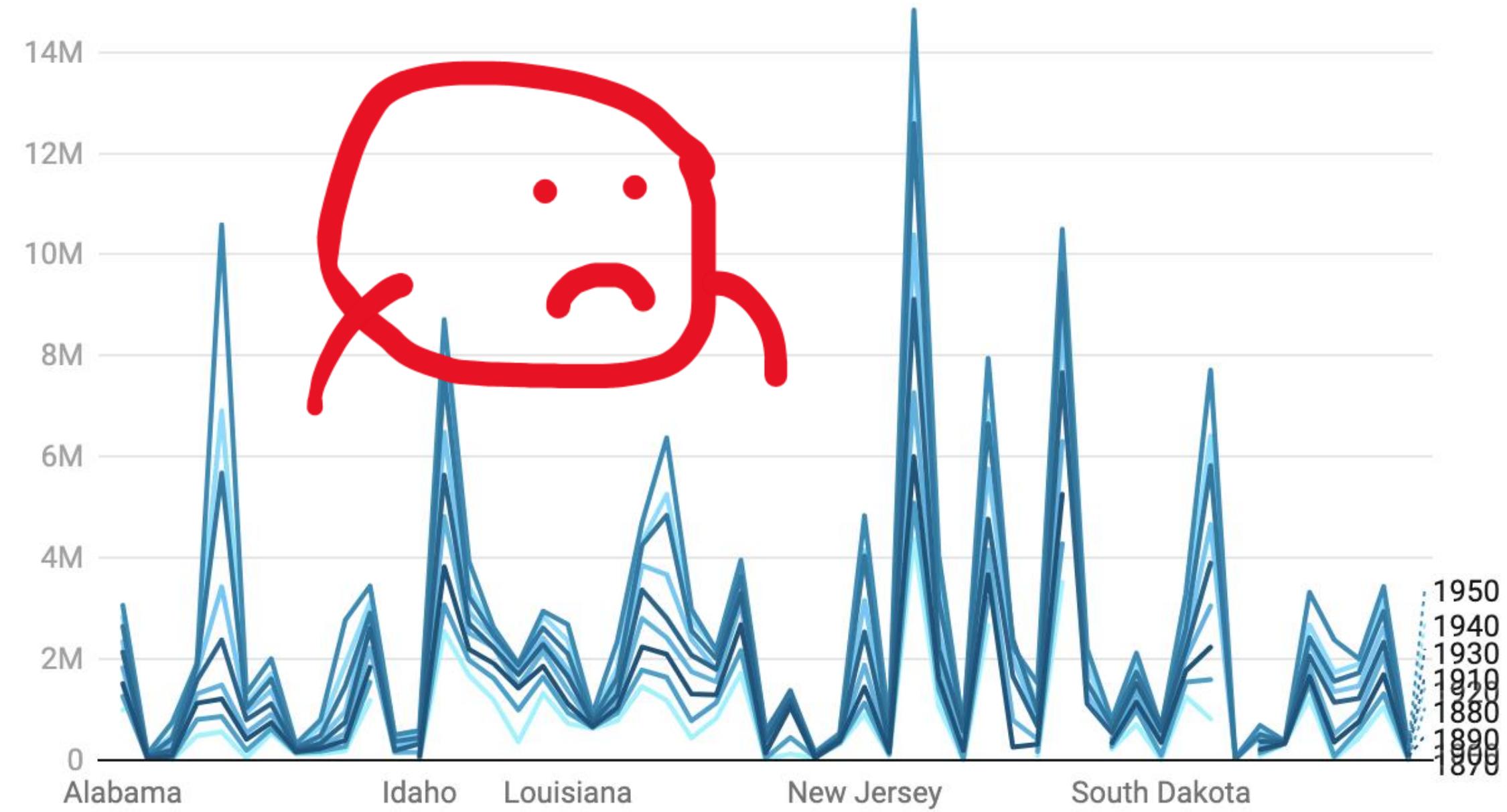
Overview Permissions Reviews

Custom tooltip

Address: Doha, United States Tokyo, Japan Shanghai, China Mexico, South America New York City, USA Taipei, Taiwan Singapore

Selected: - Press

Table formatting



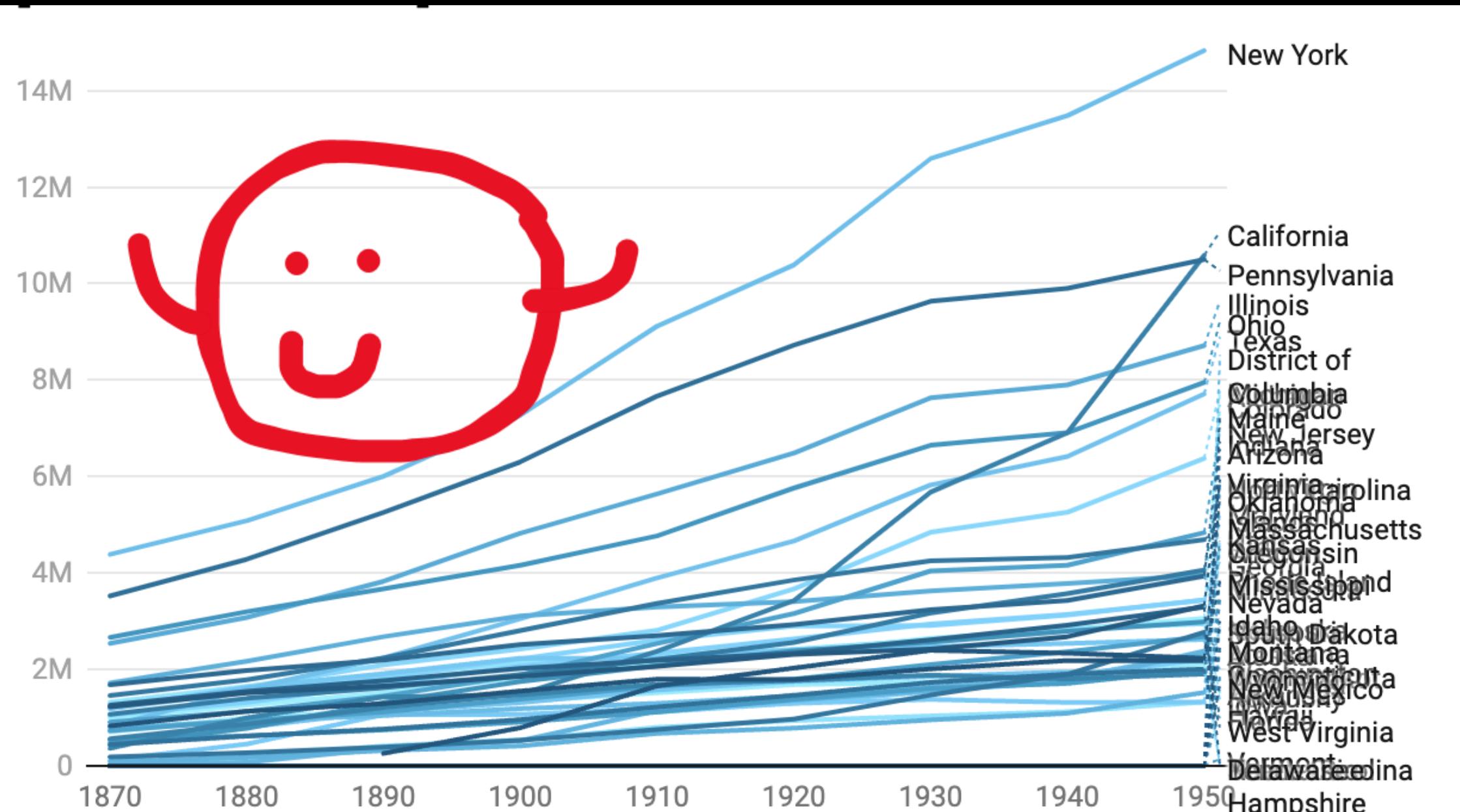
[Get the data](#) • Created with [Datawrapper](#)



	A	B	C	D	E	F	G	H	I	J
1	Name	1870	1880	1890	1900	1910	1920	1930	1940	1950
2	Alabama	996,992	1,262,505	1,513,401	1,828,697	2,138,093	2,348,174	2,646,248	2,832,961	3,061,743
3	Alaska	—	33,426	32.052	63,592	64,356	55,036	59,278	72,524	128,643
4	Arizona	9,658	40,440	88,243	122,931	204,354	334,162	435,573	499,261	749,587
5	Arkansas	484,471	560,247	802,525	1,128,211	1,574,449	1,752,204	1,854,482	1,949,387	1,909,511
6	California	560,247	622,700	746,258	799,024	2,377,549	3,426,861	5,677,251	6,907,387	10,586,223
7	Colorado	537,454	622,700	746,258	799,024	1,380,631	1,606,903	1,709,242	2,007,280	318,085
8	Connecticut	125,015	146,608	168,493	184,735	202,322	223,003	238,380	266,505	802,178
9	Delaware	—	—	—	278,718	331,069	437,571	486,869	663,091	—
10	District of Columbia	—	—	—	—	—	—	—	—	—

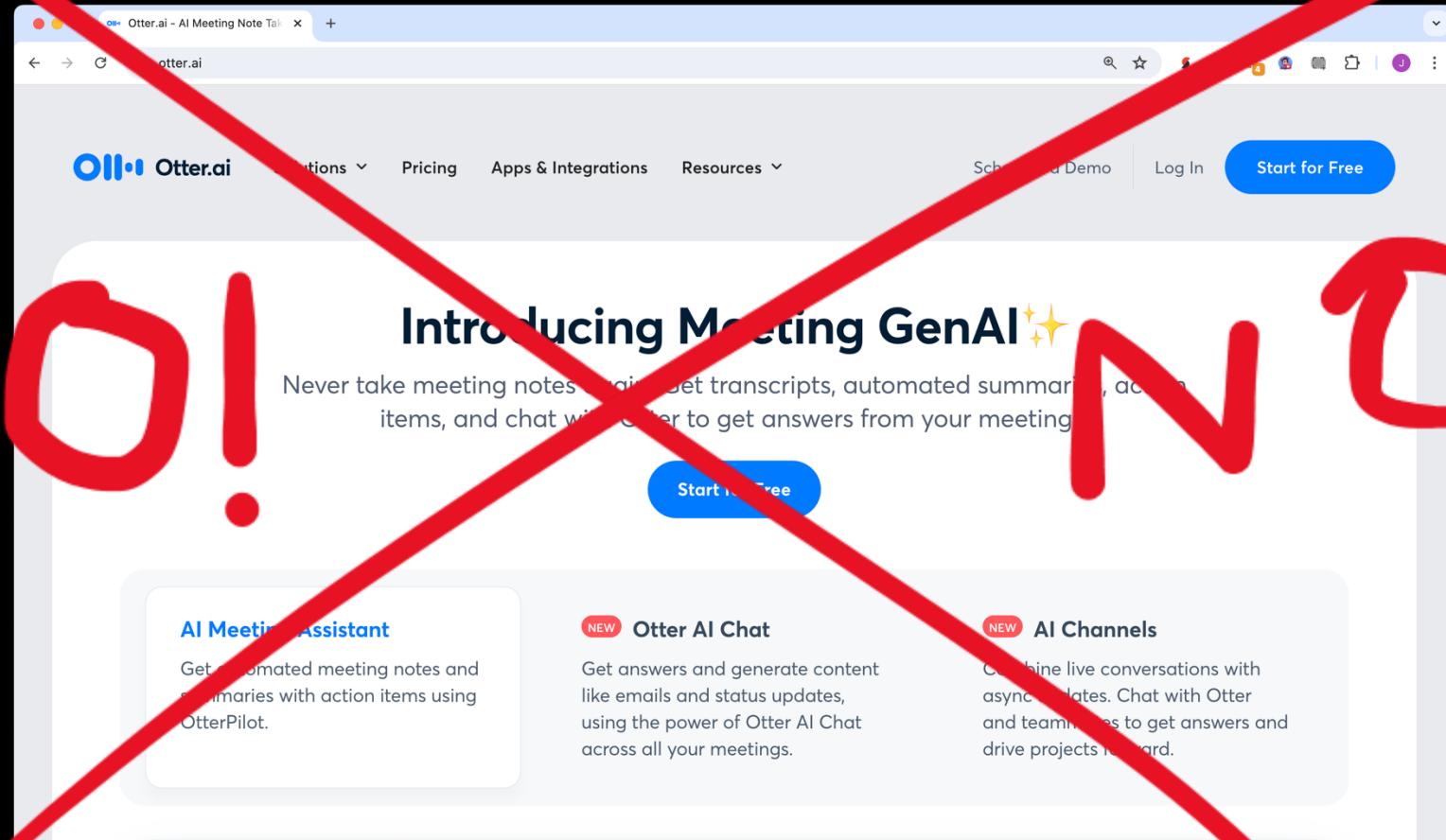
transpose
(swap rows/columns)

	A	B	C	D	E	F	G	H	I	J	K
1	Name	Alabama	Alaska	Arizona	Arkansas	California	Colorado	Connecticut	Delaware	District of Columbia	Florida
2	1870	996,992	—	9.658	484,471	560,247	39,864	537,454	125.015	131.7	187,74
3	1880	1,262,505	33.426	40.44	802,525	864,694	194,327	622,700	146.608	177.624	269,49
4	1890	1,513,401	32.052	88.243	1,128,211	1,213,398	413,249	746,258	168.493	230.392	391,42
5	1900	1,828,697	63.592	122.931	1,311,564	1,485,053	539,700	908,420	184.735	278.718	528,54
6	1910	2,138,093	64.356	204.354	1,574,449	2,377,549	799,024	1,114,756	202.322	331.069	752,61
7	1920	2,348,174	55.036	334.162	1,752,204	3,426,861	939,629	1,380,631	223.003	437.571	968,47
8	1930	2,646,248	59.278	435.573	1,854,482	5,677,251	1,035,791	1,606,903	238.38	486.869	1,468,21
9	1940	2,832,961	72.524	499.261	1,949,387	6,907,387	1,123,296	1,709,242	266.505	663.091	1,897,41
10	1950	3,061,743	128.643	749.587	1,909,511	10,586,223	1,325,089	2,007,280	318.085	802.178	2,771,30



Transcribing audio

NO!

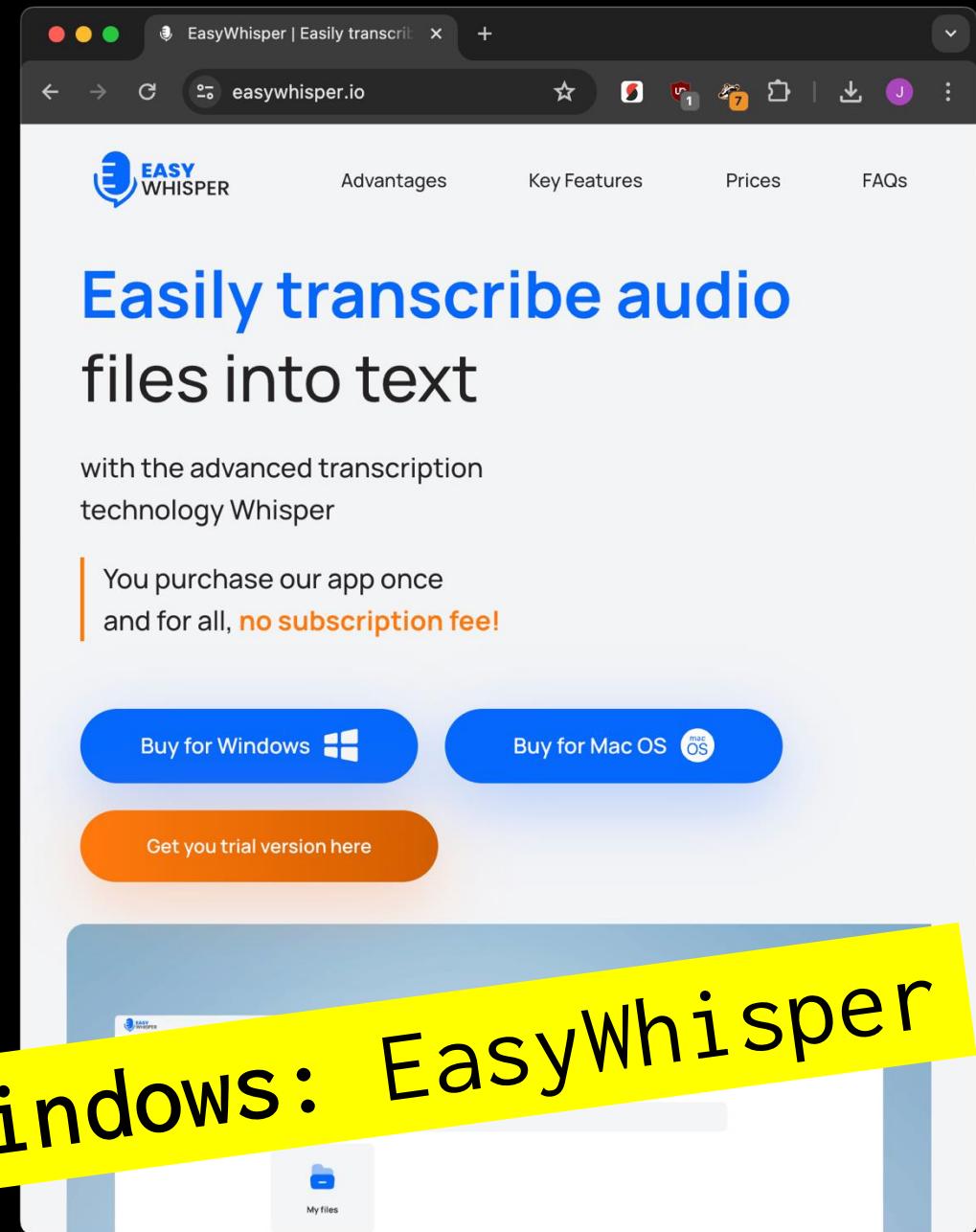


magic technology: OpenAI's Whisper

The image shows a screenshot of a web browser displaying the GitHub repository for OpenAI's Whisper. The repository page includes a search bar, navigation links for Code, Pull requests, Discussions, Actions, Security, and Insights, and a summary of 530 Watchers and 7.4k Forks. Below this, there are sections for the main branch, 7 branches, and 10 tags. A list of recent commits is shown, with the most recent being a merge from 'ryanheise and jongwook' skipping silence around hallucinations (#1838). The repository also contains .github/workflows, data, notebooks, tests, whisper, .flake8, .gitattributes, .gitignore, and .pre-commit-config.yaml files.

Below the repository details is a table comparing different Whisper model sizes based on size, parameters, English-only model variants, multilingual model variants, required VRAM, and relative speed.

Size	Parameters	English-only model	Multilingual model	Required VRAM	Relative speed
tiny	39 M	tiny.en	tiny	~1 GB	~32x
base	74 M	base.en	base	~1 GB	~16x
small	244 M	small.en	small	~2 GB	~6x
medium	769 M	medium.en	medium	~5 GB	~2x
large	1550 M	N/A	large	~10 GB	1x



Practical AI for Investigative J x +

youtube.com/playlist?list=PLewNEVDy7gq1_GPUaL0OQ31QsiHP5ncAQ

YouTube Search

Home Shorts Subscriptions

You > Your channel History Playlists Your videos Watch later Liked videos Your clips

Subscriptions Freya Holmér Home RenoVisio... 1kb construction Javier Mercedes Bill McClintock

Dearest friend, I daresay I have not partaken of food in ages. I'm positively famished

Practical AI for Investigative Journalism

Jonathan Soma 6 videos Public

A six session series held in April 2024 about real-life use cases for journalism in (mostly investigative) journalism. ...more

Play all

Sort

Dearest friend, I daresay I have not partaken of food in ages. I'm positively famished 2:09:32

Sorting documents (Practical AI for Investigative Journalism, Session 1)

Jonathan Soma • 1K views • Streamed 2 months ago

Large language models don't understand facts or concepts, they only know statistical probability 2:14:32

Structured, validated data from LLMs (Practical AI for Investigative Journalism, Session 2)

Jonathan Soma • 716 views • Streamed 2 months ago

Evaluating Verifiability in Generative Search Engines 2:14:02

Why generative AI is a dead end for responsible journalism (Practical AI for Journalism, Session 3)

Jonathan Soma • 572 views • Streamed 1 month ago

LLM citations are misinformation: 1:45:26

AI, Hugging Face and non-chatbot models (Practical AI for Journalism, Session 4)

Jonathan Soma • 810 views • Streamed 1 month ago

Context window 2:01:02

Local models/private AI (Practical AI for Investigative Journalism, Session 5)

Jonathan Soma • 444 views • Streamed 1 month ago

Transcription and audio models (Practical AI for Investigative Journalism, Session 6)

Jonathan Soma • 325 views • Streamed 1 month ago

Red annotations: A large red circle highlights the thumbnail of the first video in the playlist. A large red arrow points from the bottom right towards the bottom of the first video thumbnail.

***Generating structured data
with the awful power of LLMs***

FROM: Mulberry Peppertown
(mulbs@example.com)

When I pick up the cans of beans they are all so light! At first I thought they were empty, but it turns out they are just futuristic beans that are not heavy like the old style beans I was used to. It is incredible.

Mulberry Peppertown

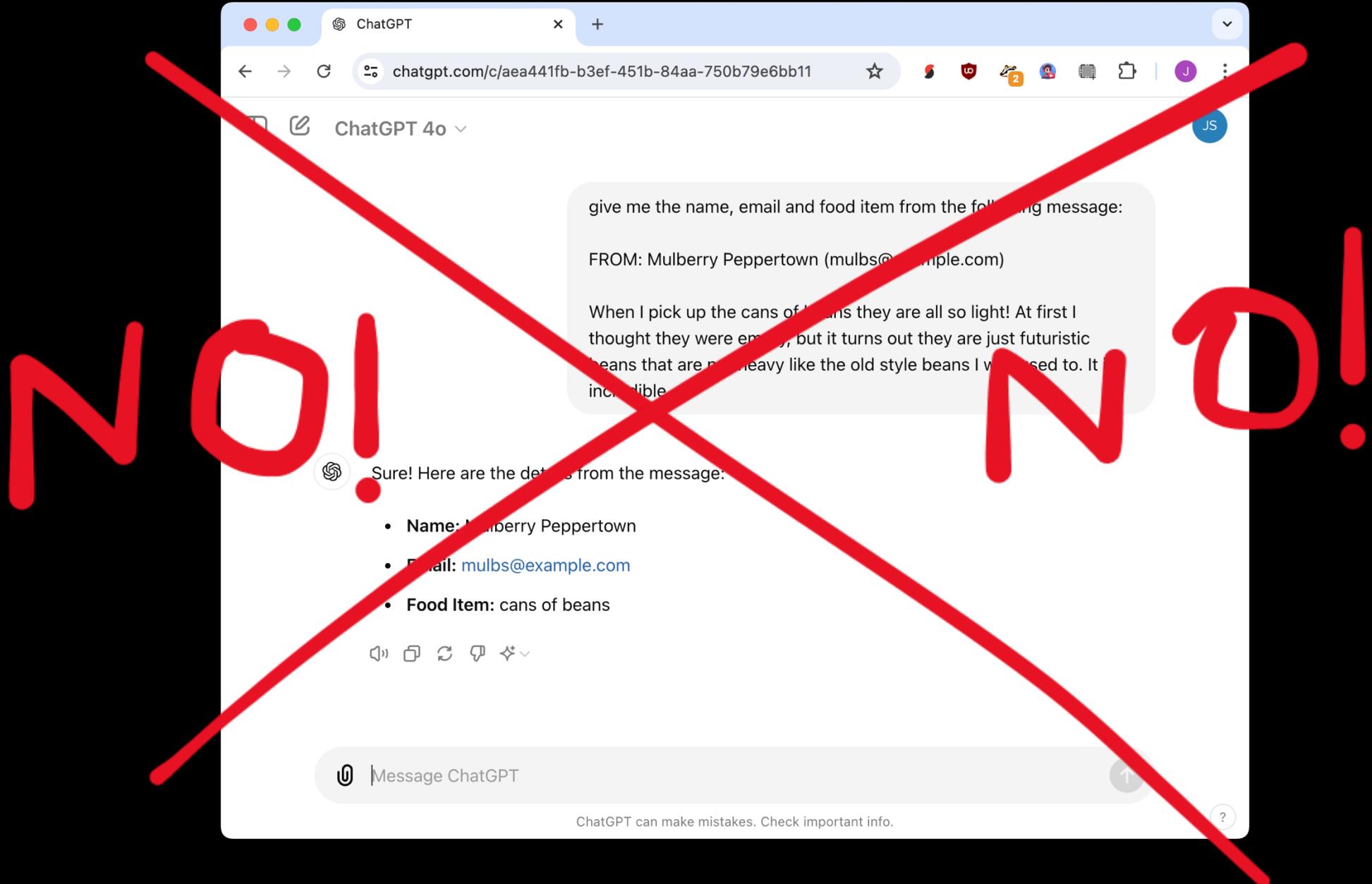
name

mulbs@example.com

email

beans

food item



email

name	email	product
FROM: Mulberry Peppertown (mulbs@example.com)		

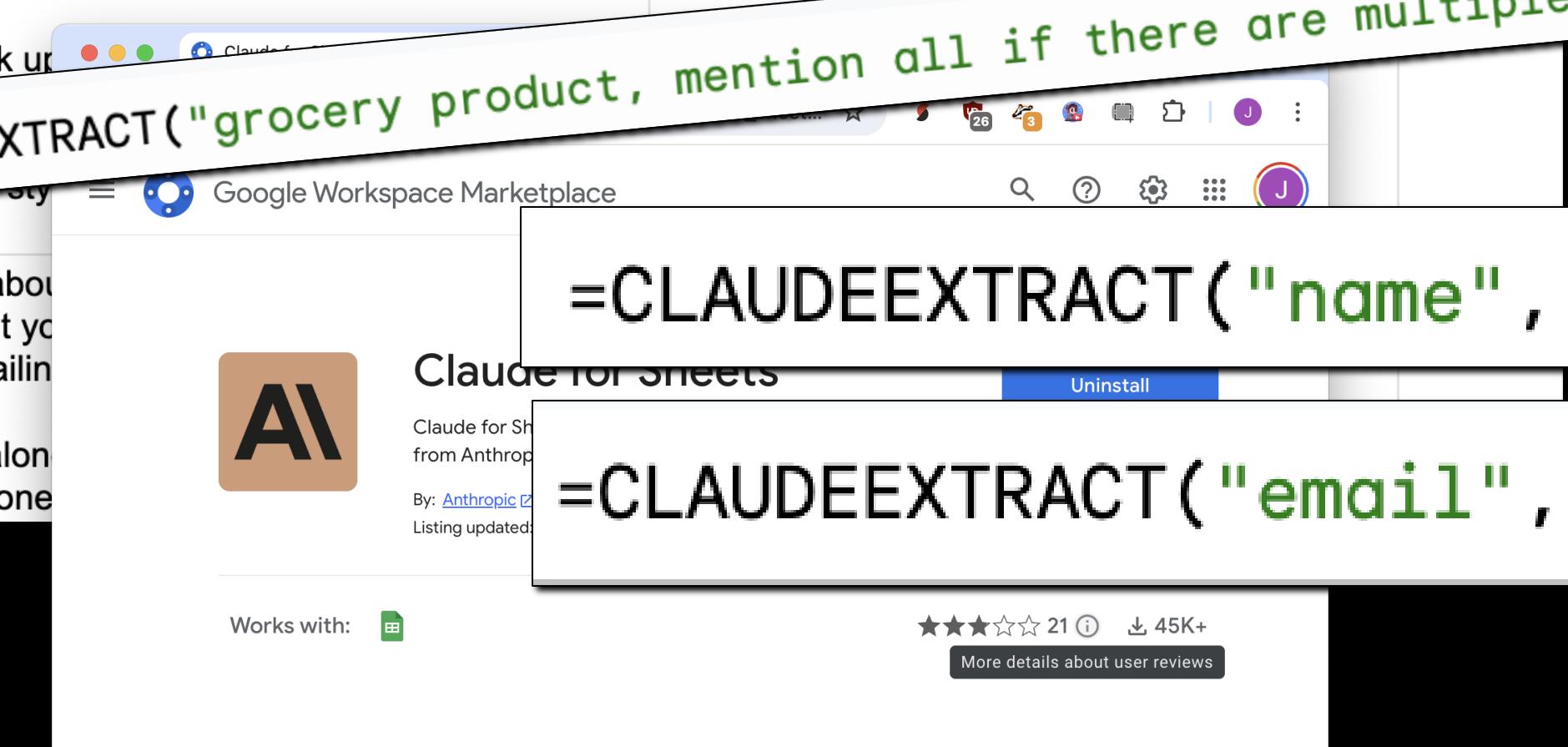
When I pick up my groceries, I can't remember what I bought. I'm thinking of creating a spreadsheet to keep track of my purchases. I've heard about Claude, a large language model, and I'm curious if it could help me extract product information from my grocery store receipt emails.

=CLAUDEEXTRACT("grocery product, mention all if there are multiple", A2)

incredible.

I am irate about shopping at your broccoli mailin

Jackary Balon jackary.balone



Google Sheets

Google Workspace Marketplace

Search

Uninstall

Claude for Sheets

By: Anthropic

Listing updated

Works with:

★★★★★ 21

45K+

More details about user reviews

Sheets: Claude for Sheets

```
=AI("extract the email address", A2)
```

=AI("extract the email address",

=AI("extract the email address")

Ch 8 - 10

name	email	product

ckary Baloney nose jackary.baloneynose@exam broccoli

Gen

Sheets: Gemini in Google Sheets

Python: Structured Outputs

```
class Comment(BaseModel):
    name: str = Field(description="Person who submitted the comment")
    email: Optional[str] = Field(description="Email address of commenter")
    food_item: str = Field(description="Food item the comment is about")
    emotion: Literal["positive", "negative", "uncertain"]
```

```
comment = """
FROM: Mulberry Peppertown, mulberry (at) example.co
When I pick up the cans of beans they are all so li
first I thought they were empty, but it turns out to
futuristic beans that are not heavy like the old s
I was used to. It is incredible.
"""

result = client.chat.completions.create(
    response_model=Comment,
    messages=[{"role": "user", "content": comment}],
```

```
{
    'name': 'Mulberry Peppertown',
    'email': 'mulberry@example.com',
    'food_item': 'cans of beans',
    'emotion': 'positive'
```

Time for exercises!!

Visit bit.ly/ire25-cleaning

Jonathan Soma
Knight Chair in Data Journalism
Columbia Journalism School
js4571@columbia.edu

jonathansoma.com/everything

Visit bit.ly/ire25-cleaning

File

Make a copy

IRE25 - Data cleaning worksheet (clean)   

File Edit View Insert Format Data Tools Extensions Help

New

Open

Import

Make a copy

Share

Email

Download

Rename

B7

00 → 123 | Default... ▾

Physical address
14367 GATEWAY WEST
3402 AGNES ST, CORP
1312 E COMMERCIAL S
8035 E R L THORNTON

Tab: tow trucks

A	B
Number of tow trucks	Tow trucks (as a number)
Number of tow trucks: 0	
Number of tow trucks: 4	
Number of tow trucks: 1	
Number of tow trucks: 12	

Tab: locksmiths

C	D	
Business Legal/Trading as Name and Street Address	Total employees	Address
ELKIN LOCKSMITH 3719 FERRARA DRIVE Total Active Employees: 0		
SERVICE REPAIRS, LLC 13108 CAMELLIA DRIVE Total Active Employees: 0		
EASTER'S LOCK & SECURITY SOLUTIONS 1713 E JOPPA ROAD Total Active Employees: 12		
LOCKSMITH ON DUTY, LLC 4 STRABANE COURT Total Active Employees: 0		
AUTOTRONICS, INC 267 KENTLANDS BLVD. #1066 Total Active Employees: 0		
A1 LOCKSMITH SERVICES 9711 WASHINGTON BLVD SUITE 550 J Total Active Employees: 0		
ROY'S QUALITY CAR CARE LOCKS UNLIMITED 351 E ANTIETAM STREET Total Active Employees: 0		
THE KEYLESS SHOP 261 FREDERICK STREET Total Active Employees: 1		
BBLC, LLC LOCKS UNLIMITED MD 327 E WILSON BLVD Total Active Employees: 0		
ULTIMATE LOCKSMITH, LLC 7 ALDERLEAF COURT Total Active Employees: 0		
BARGER LOCK AND KEY LLC 328 NORTH CAROLINA AVE Total Active Employees: 0		
PROTECTOR LOCKSMITH LLC 8054 OUTING AVENUE Total Active Employees: 0		
LOCKTECH 26A BROOKFIELD ROAD Total Active Employees: 0		
EASY ACES 24 HOUR LOCKOUT SERVICE 122 HIGHSHIRE COURT Total Active Employees: 0		
STUART'S KEYS 2401 NORTH POINT BOULEVARD Total Active Employees: 0		
BEAR LOCK & SECURITY SERVICE, INC. 205 CLEVELAND AVENUE Total Active Employees: 1		
TITAN TEAM LOCKSMITH LLC 8416 KAVANAGH ROAD Total Active Employees: 0		
GERMAN PERALTA LOCKSMITH 255 RIVERVIEW AVE Total Active Employees: 0		
MEGA LOCK AND KEY LLC 12315 TIMBER GROVE ROAD Total Active Employees: 0		

Tab: tow trucks 2

A	B	C	D	E	F	G
TDLR	page_contents	Company name	Phone number	Owners	DBA	Address
006534502C	<table align="center" width="" border="0" cellpadding="0" cellspacing="0"><tbody><tr><td bgcolor="#FFF8DC">Company Information:</td><td bgcolor="#FFF8DC"></td></tr><tr><td width="50%" align="left">Name: SKYHAWK T&R L.L.C.</td></tr>					
006546993C	<table align="center" width="" border="0" cellpadding="0" cellspacing="0"><tbody><tr><td bgcolor="#FFF8DC">Company Information:</td><td bgcolor="#FFF8DC"></td></tr><tr><td width="50%" align="left">Name: GULF COAST ASSET RECOVERY LLC</td></tr>					
006579452C	<table align="center" width="" border="0" cellpadding="0" cellspacing="0"><tbody><tr><td bgcolor="#FFF8DC">Company Information:</td><td bgcolor="#FFF8DC"></td></tr><tr>					
006532111C	<table align="center" width="" border="0" cellpadding="0" cellspacing="0"><tbody><tr><td bgcolor="#FFF8DC">Company Information:</td><td bgcolor="#FFF8DC"></td></tr><tr>					

Tab: tow trucks 2

A	B	C	D	E	F	G
TDLR	page_contents	Company name	Phone number	Owners	DBA	Address
006534502C	<table align="center" width="" border="0" cellpadding="0" cellspacing="0"><tbody><tr><td bgcolor="#FFF8DC">Company Information:</td><td bgcolor="#FFF8DC"></td></tr><tr><td width="50%" align="left">Name: SKYHAWK T&R L.L.C.</td></tr>					
006546993C	<table align="center" width="" border="0" cellpadding="0" cellspacing="0"><tbody><tr><td bgcolor="#FFF8DC">Company Information:</td><td bgcolor="#FFF8DC"></td></tr><tr><td width="50%" align="left">Name: GULF COAST ASSET RECOVERY LLC</td></tr>					
006579452C	<table align="center" width="" border="0" cellpadding="0" cellspacing="0"><tbody><tr><td bgcolor="#FFF8DC">Company Information:</td><td bgcolor="#FFF8DC"></td></tr><tr>					
006532111C	<table align="center" width="" border="0" cellpadding="0" cellspacing="0"><tbody><tr><td bgcolor="#FFF8DC">Company Information:</td><td bgcolor="#FFF8DC"></td></tr><tr>					

Tab: NEISS

A	B	C	D	E	F	
Age	Age	Race	Other_Race	Wall puncher?	What did they punch?	Narrative
15YO		1				15YOM PUNCHED A WALL AND C/O PAIN TO R HAND AND WRIST
27YO		3 HISPANIC				27YOM PUNCHED A WALL NOW WITH LT HAND PAIN DX CLOSE
17YO		0				17 YOM - RT HAND FX - PT PUNCHED A DOOR 3 WEEKS AGO A
46YO		0				46 YR OLD FEMALE PLAYING INDOOR SOCCER AND HAD A FA
10YO		0				10YM ASSAULTED BY ANOTHER STUDENT WHO PUNCHED HIM
8YO		1				8 YOF SWINGING FROM A SWING, TRIED TO JUMP OFF AND F
12YO		0				12YOM WAS RIDING BIKE UP THE HILL WHEN HE FELL ON HIS
19YO		0				19 YO M PUNCHED A MIRROR FX HAND
17YO		1				17 YOF UPSET WITH FRIEND, PUNCHED METAL POLE C/O HAN
20YO		0				20YOM W/DAD EVAL L HAND INJURY AFTER HE GOT MAD & PU
19YO		1				19 YOF PAIN TO RT HAND AFTER PUNCHING A DOOR. DX NON
22YO		1				22YOM PAIN TO L HAND WHEN PUNCHING A WALL / CONTUSIO
27YO		0				27YF ACC JAMMED FINGER AGAINST A DOOR >>NAIL AVULSION
41YO		0				41-YOF ACCIDENTALLY PUNCHED TABLE W/FIST, HAVING HAN
41YO		0				41YF BOILING WATER&ACC SPILLED THE HOT WATER ONTO F
34YO		1				34YOF C/O RT HAND INJURY FROM PUNCHING A WALL. DX; C
5YO		1				5 YOM CAUGHT FINGER IN CLOSING DOOR AT HOME. DX: R M
24YO		0				24YOM STRIKING WALL W R HAND W TRYING TO HIT PUNCHIN

Jonathan Soma
Knight Chair in Data Journalism
Columbia Journalism School
js4571@columbia.edu

Visit bit.ly/ire25-cleaning

jonathansoma.com/everything