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

Visit bit.ly/ire25-cleaning

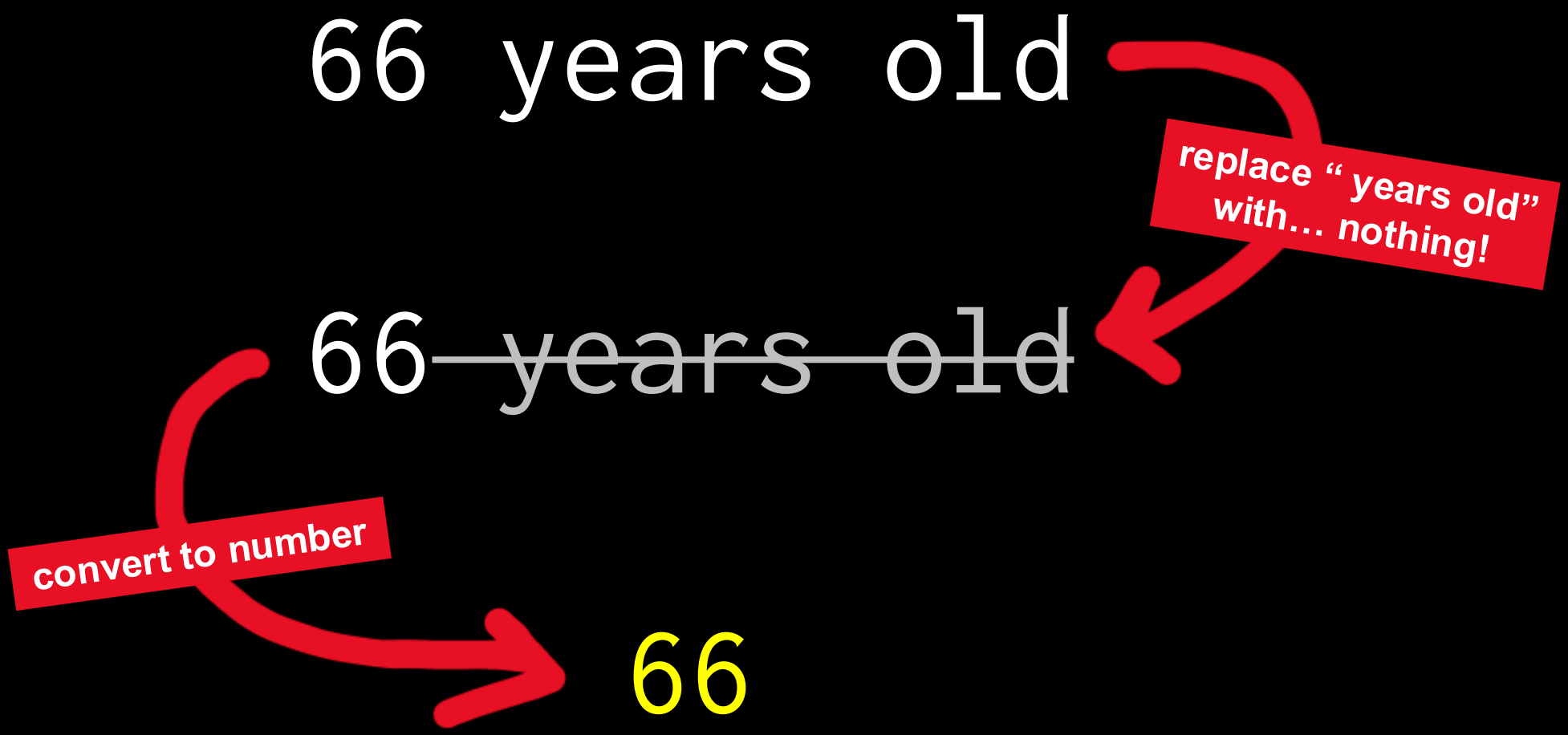
jonathansoma.com/everything

Why not AI?

- 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

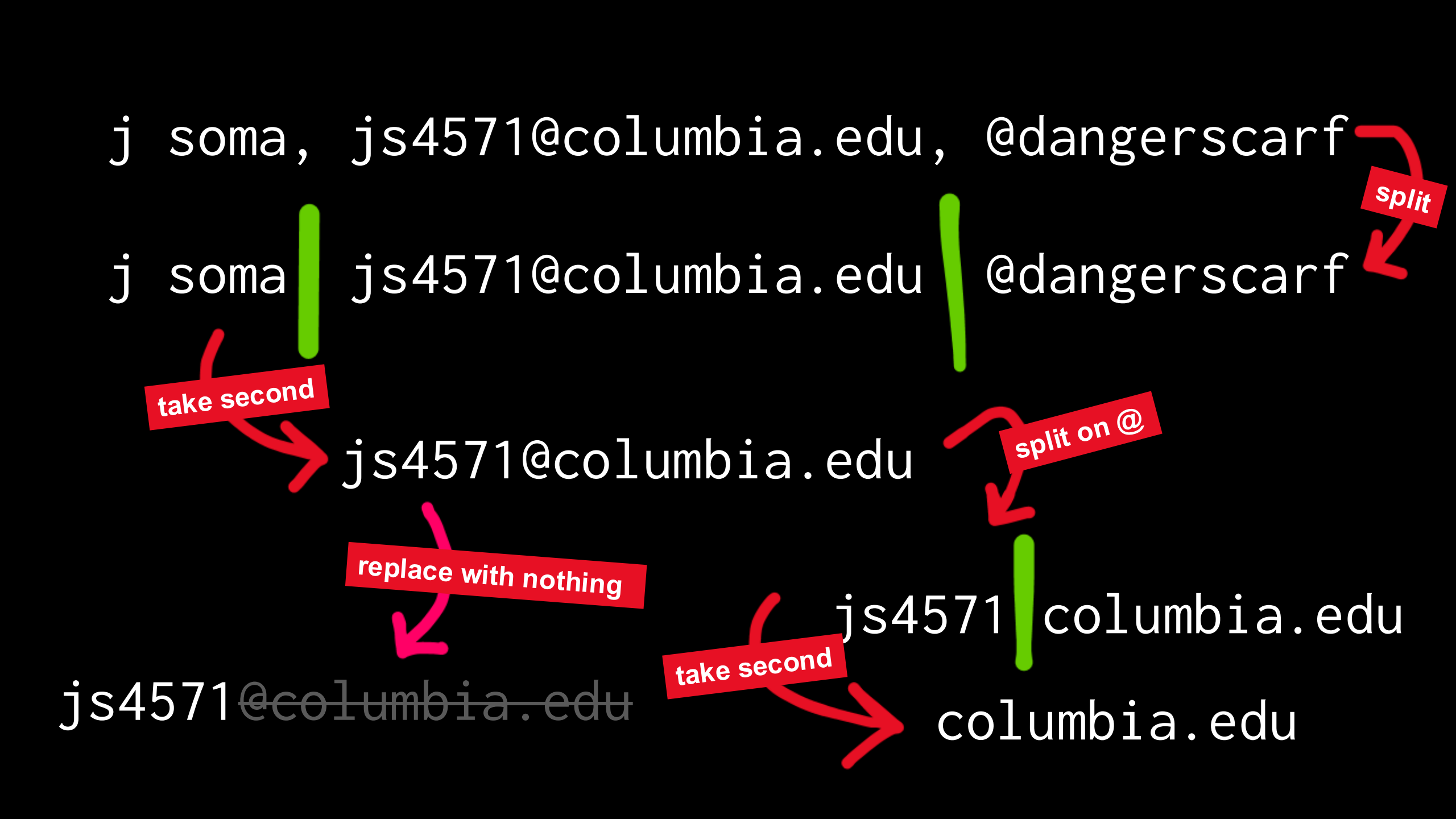
replace and split are 99% of the game


66 years old
replace " years old" with... nothing!
66 years old
convert to number
66

j soma, js4571@columbia.edu, @dangerscarf

j soma | js4571@columbia.edu | @dangerscarf

js4571@columbia.edu


j soma, js4571@columbia.edu, @dangerscarf
split
j soma
js4571@columbia.edu
@dangerscarf
take second
js4571@columbia.edu
split on @
replace with nothing
js4571
columbia.edu
take second
js4571@columbia.edu
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.

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.

just ask!
then TEST IT

excel

Copy code

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

Feel free to reach out at 555-555-5555

555-555-5555

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 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

```python
# 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.

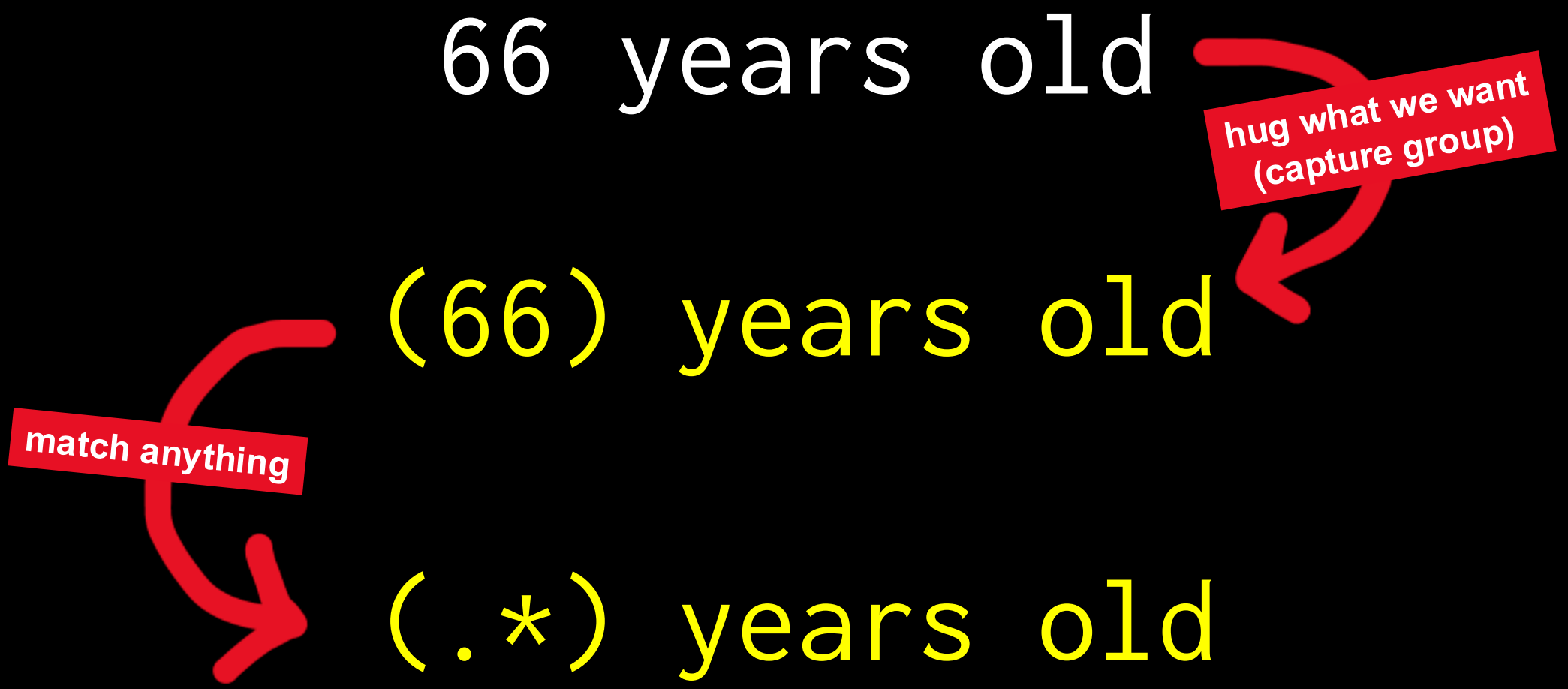
You can use the `str.extract` method to extract the "@columbia.edu" part from the "details" column. Here's the code to do that:

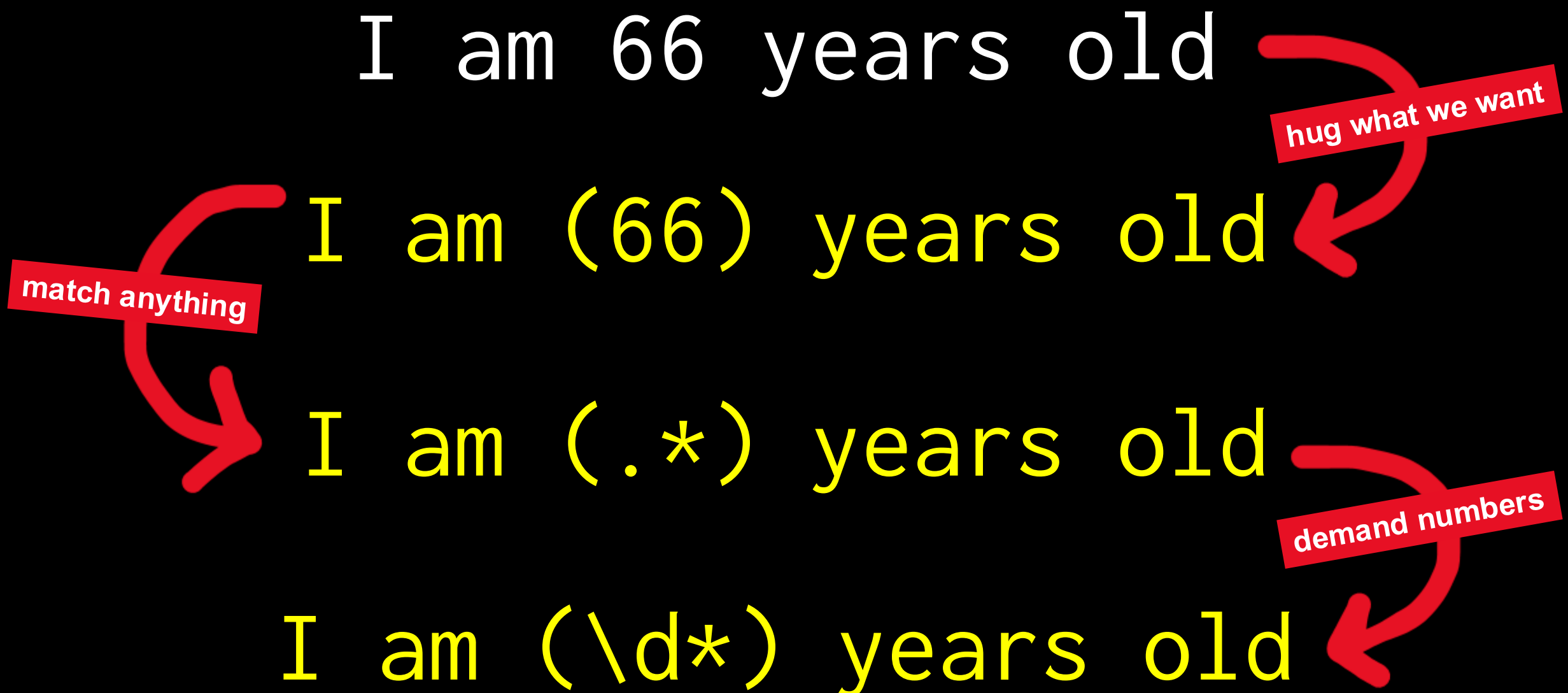
python Copy code

```python
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
hug what we want
I am (66) years old
match anything
I am (.*) years old
demand numbers
I am (\d*) years old

Contact me at 555-555-5555

Contact me at (.*)

hug what we want

but maybe not!

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

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

just numbers

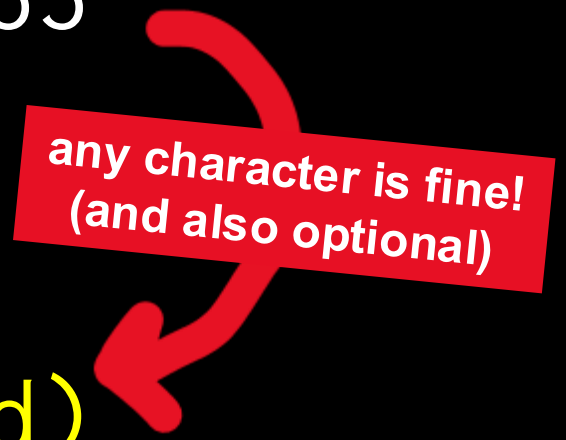
Contact me at 555-555-5555

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

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)

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)

just ask!
then TEST IT

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	4 km2 (7,341 sq mi)	332,470 km2 (128,370 sq mi)
	5 km2 (5,516 sq mi)	346,756 km2 (133,883 sq mi)
	km2 (4,498 sq mi)	358,407 km2 (138,382 sq mi)
	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)

< 5/5 >

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

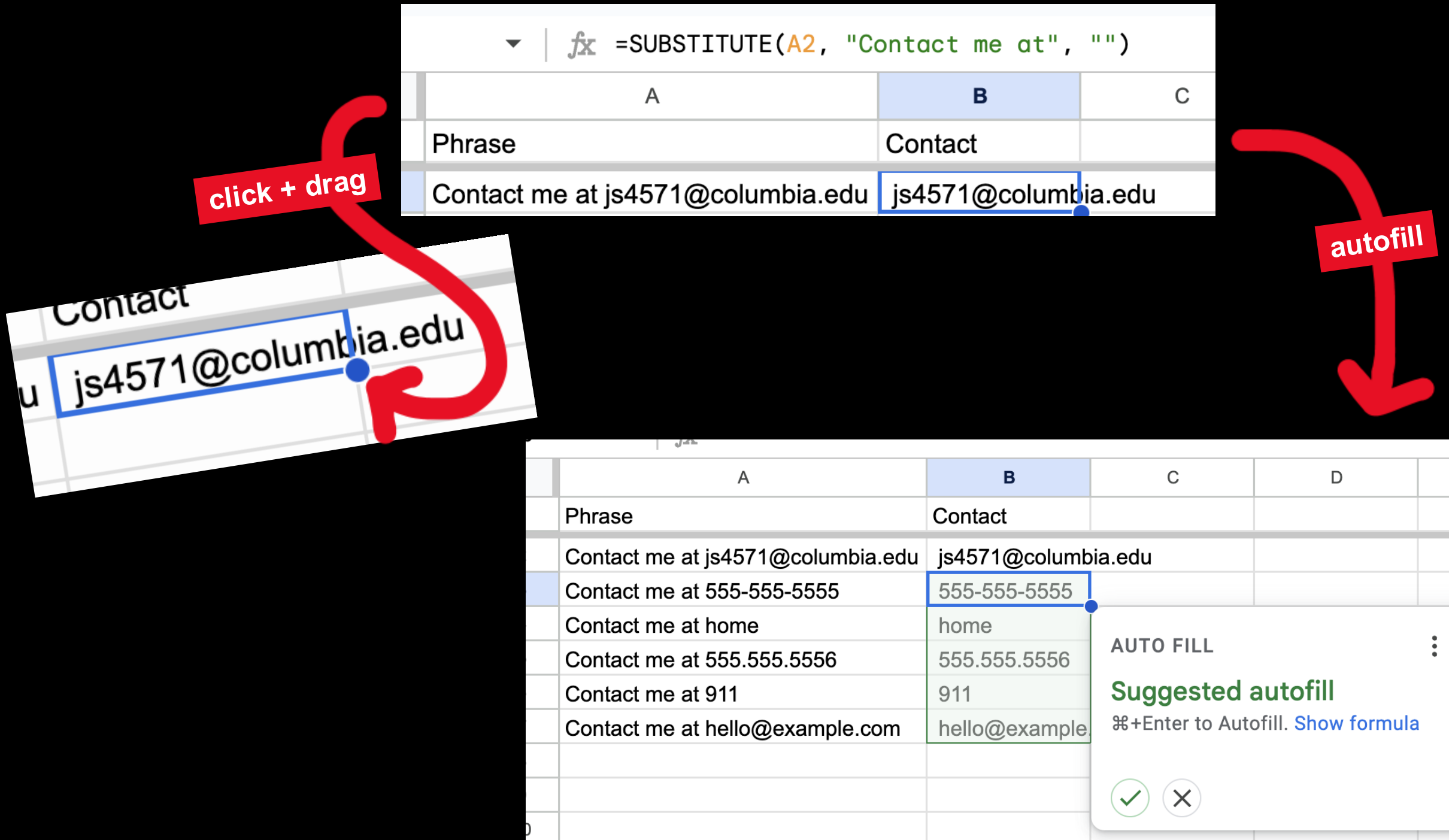
```scss
^\d{4}\s+(\d{1,5})\s+km2
```

Here's a breakdown of the regex:

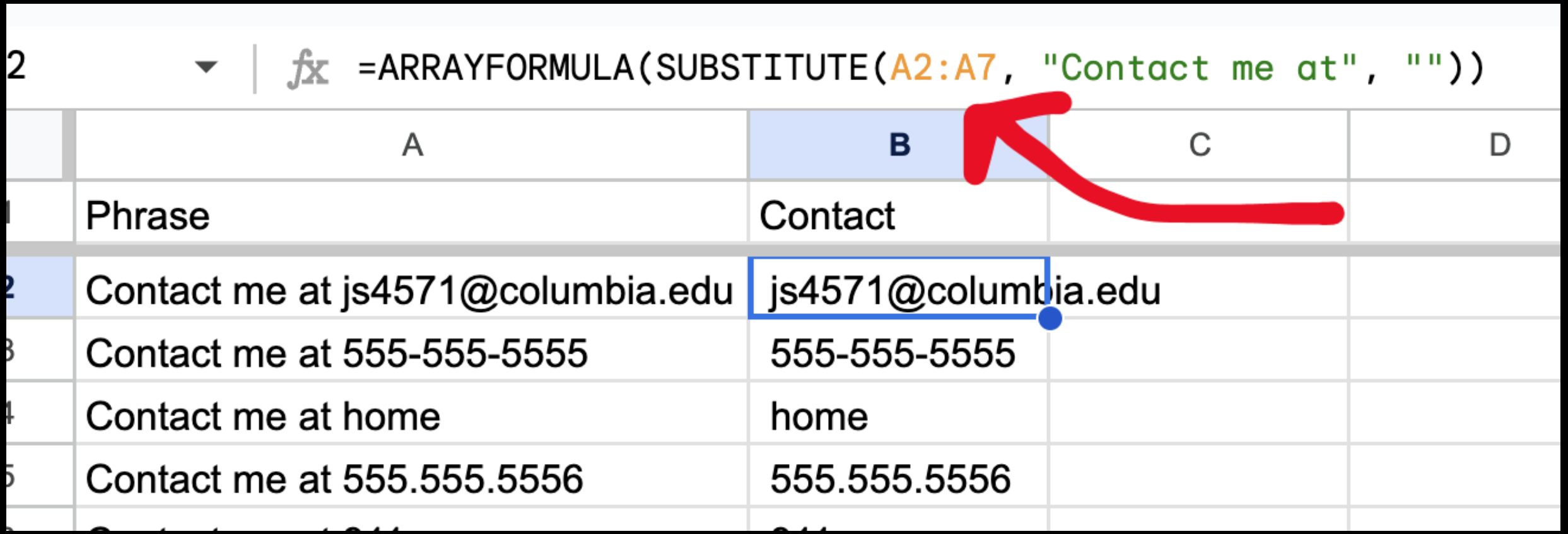
- `^` asserts the position at the start of the line.
- `\d{4}` matches exactly four digits (the year).
- `\s+` matches one or more whitespace characters.
- `(\d{1,5})` captures between 1 and 5 digits (the deforested area value) in a capturing group.
- `\s+` matches one or more whitespace characters.
- `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
Contact
u
js4571@columbia.edu
A
B
C
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
Contact me at 911
911
Contact me at hello@example.com
hello@example.
AUTO FILL
Suggested autofill
⌘+Enter to Autofill. Show formula

=ARRAYFORMULA(.....)



Fuzzy matching

Jonathan Soma

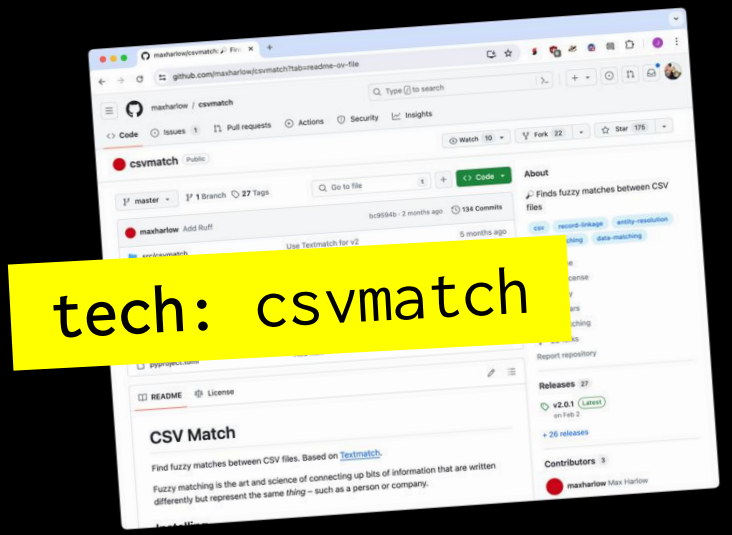
Jonathan P. Soma

Jon Soma

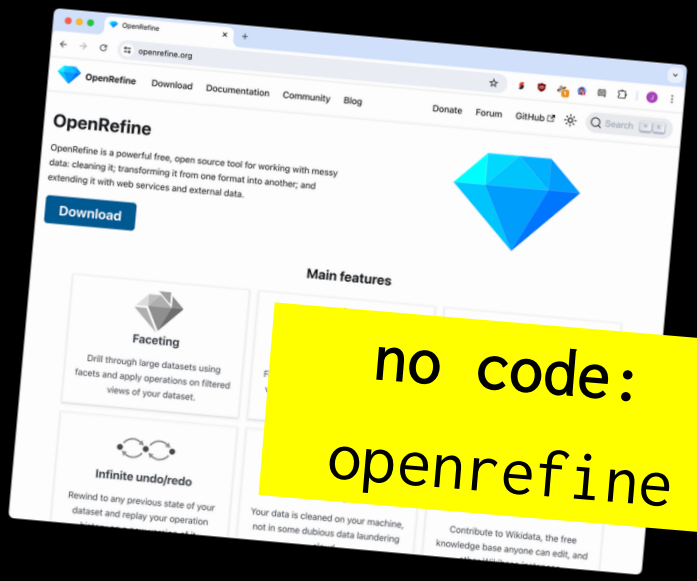
Soma, Jonathan

fuzzy matching

Jonathan Soma



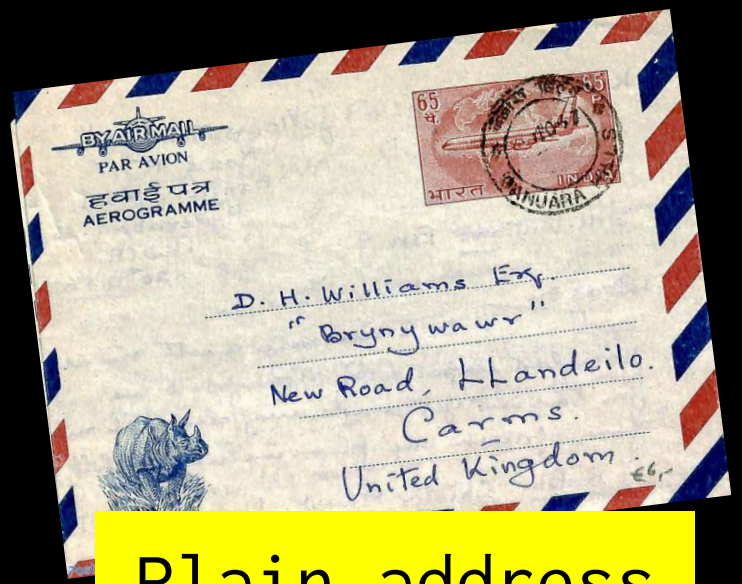
tech: csvmatch



no code: openrefine

Addresses and maps

Geocoding



Plain address

Latitude/Longitude

Reverse geocoding

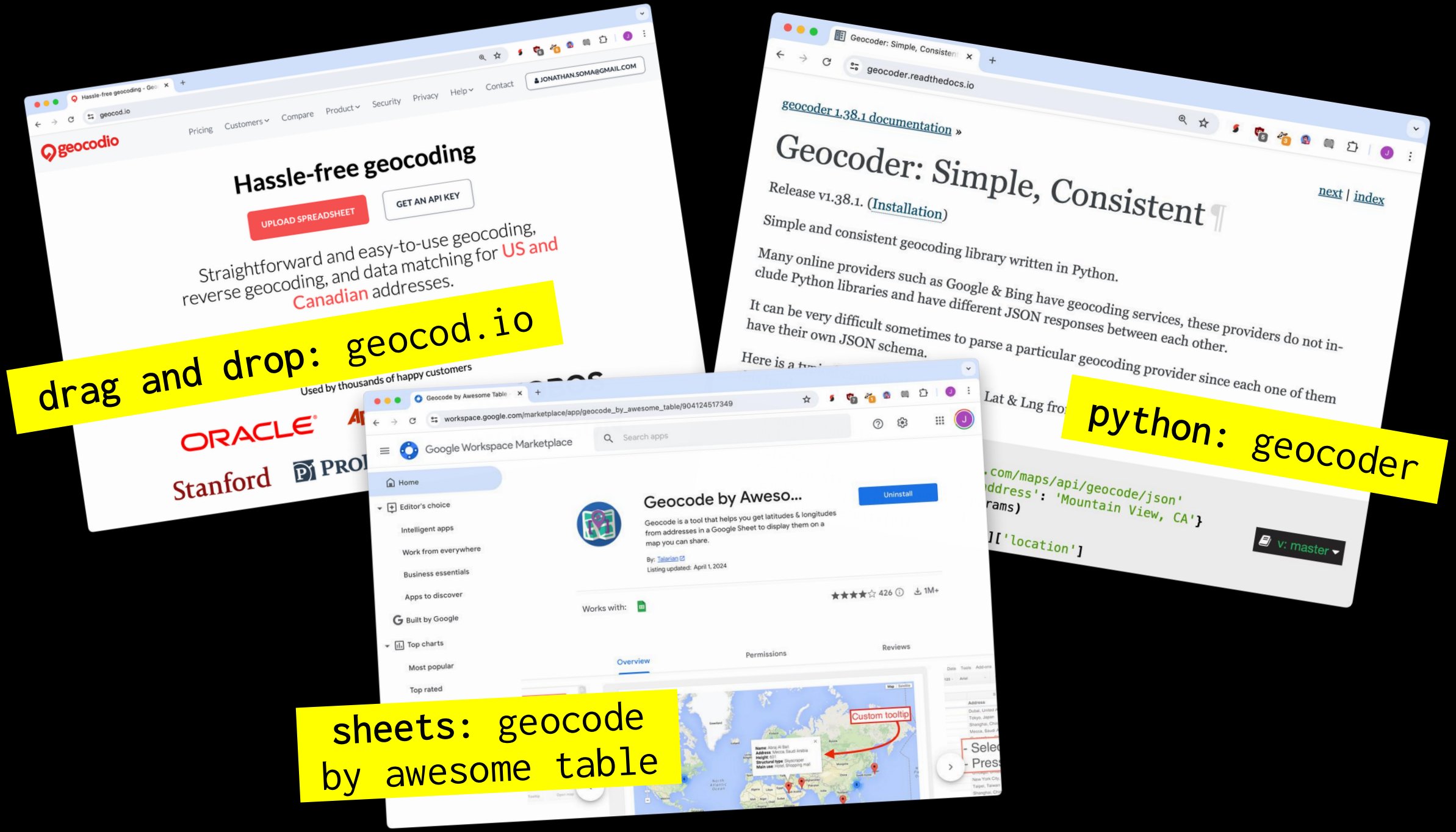
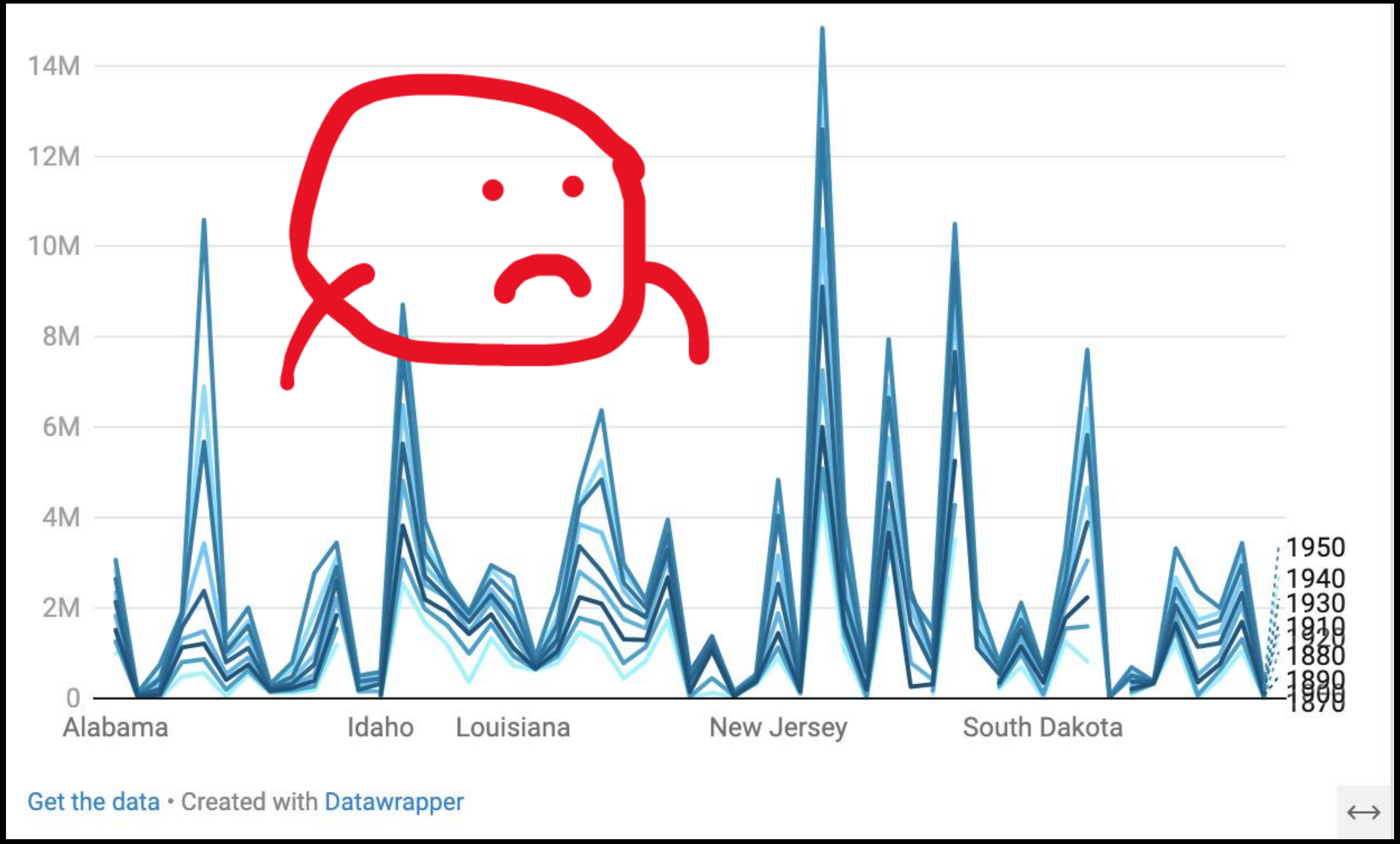

drag and drop: geocod.io
Hassle-free geocoding
UPLOAD SPREADSHEET
GET AN API KEY
Straightforward and easy-to-use geocoding, reverse geocoding, and data matching for US and Canadian addresses.
Used by thousands of happy customers
ORACLE
Stanford
python: geocoder
Geocoder: Simple, Consistent
geocoder 1.38.1 documentation »
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.
sheets: geocode by awesome table
Google Workspace Marketplace
Geocode by Aweso...
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.
Custom tooltip

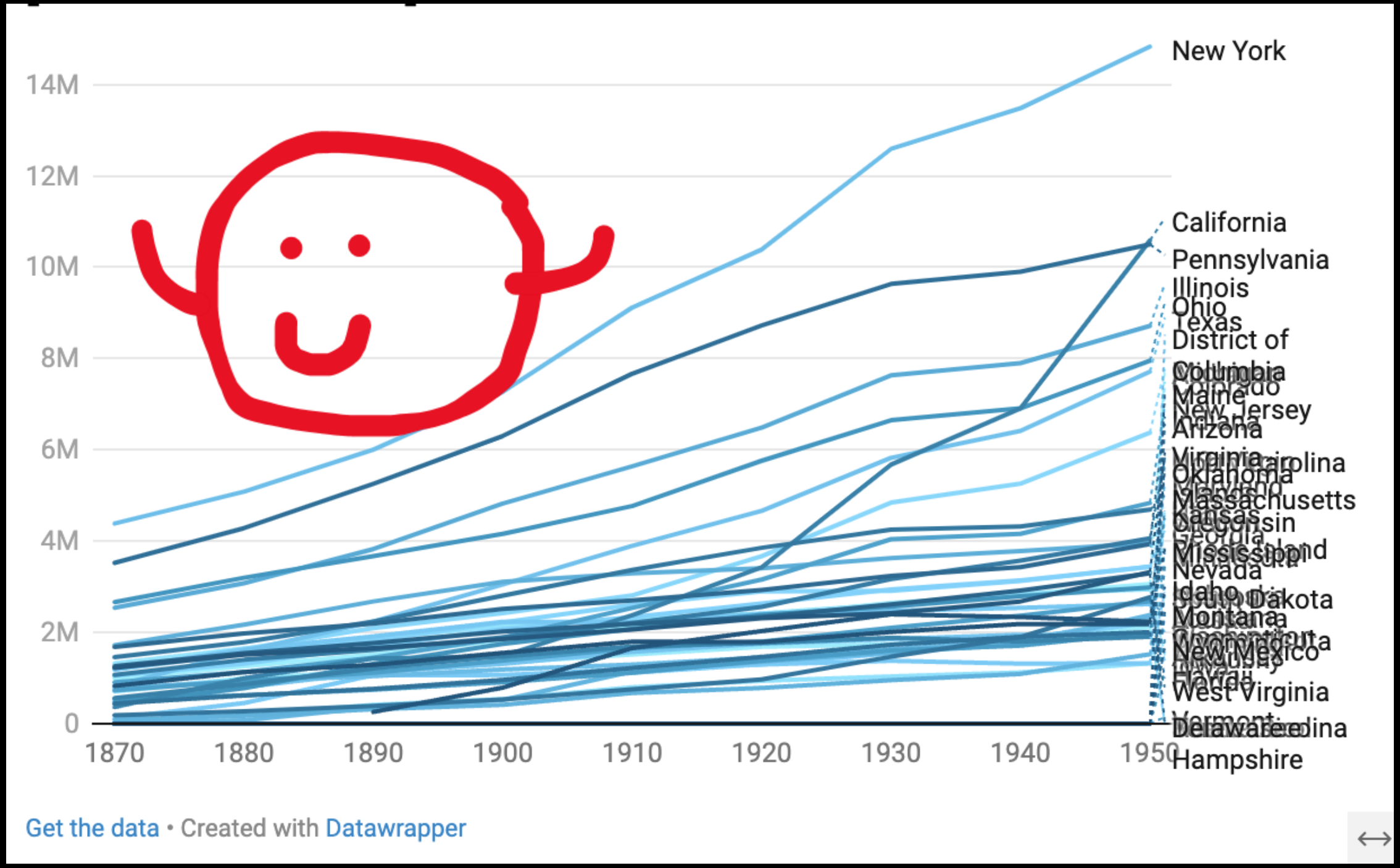
Table formatting


14M
12M
10M
8M
6M
4M
2M
0
Alabama
Idaho
Louisiana
New Jersey
South Dakota
1950
1940
1930
1910
1880
1890
1870
Get the data • Created with Datawrapper

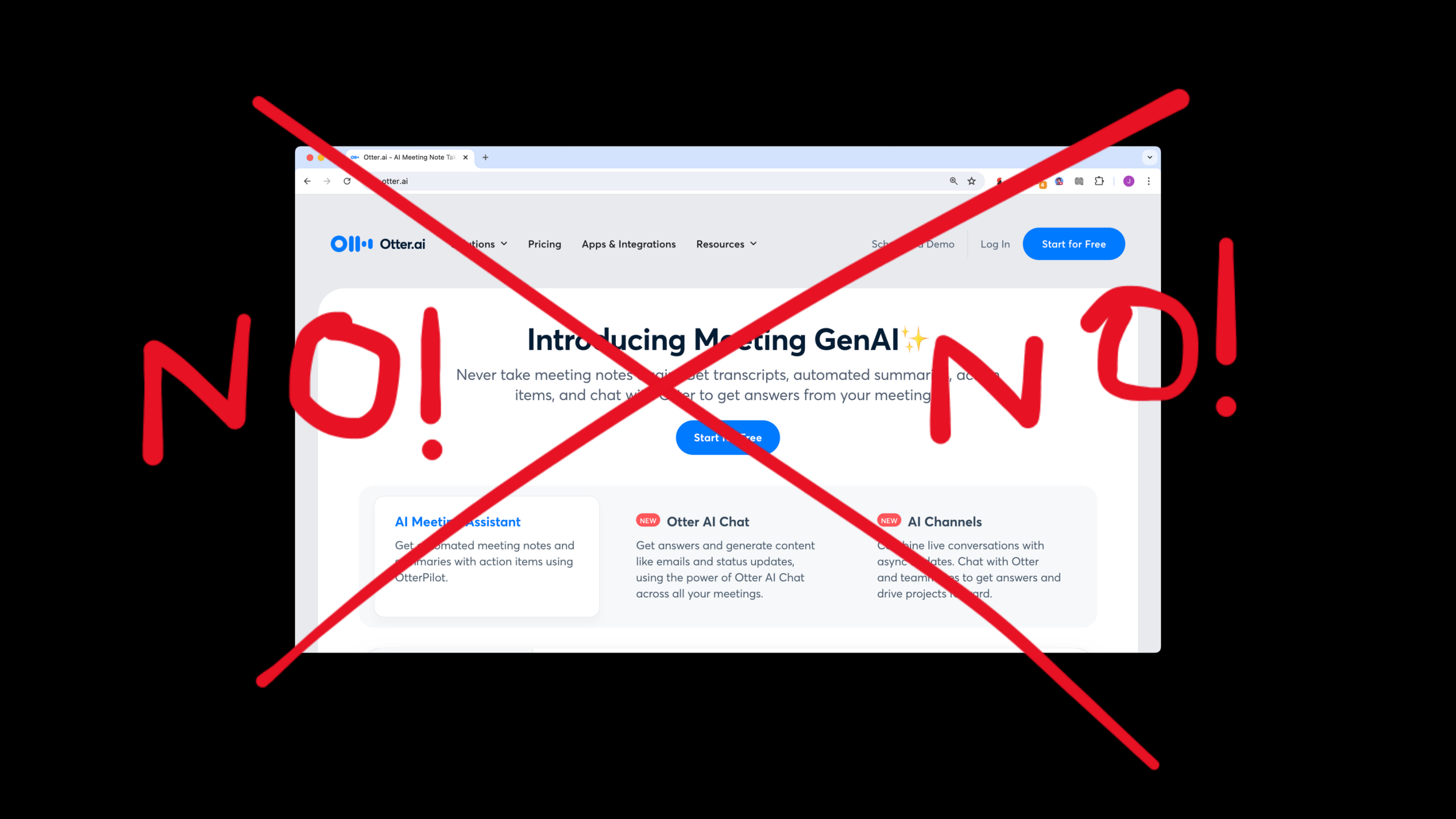
	A	B	C	D	E	F	G	H	I	J
1	Name	1870	1880	1890	1900	1910	1920	1930	1940	1950
2	Alabama	996,992	1,262,505	1,513,401	1,828,697	2,138,093	2,348,174	2,646,248	2,832,961	3,061,743
3	Alaska	–	33,426	32,052	63,592	64,356	55,036	59,278	72,524	128,643
4	Arizona	9,658	40,440	88,243	122,931	204,354	334,162	435,573	499,261	749,587
5	Arkansas	484,471	802,525	1,128,211	1,311,564	1,574,449	1,752,204	1,854,482	1,949,387	1,909,511
6	California	560,247	864,694	1,213,398	1,485,053	2,377,549	3,426,861	5,677,251	6,907,387	10,586,223
7	Colorado	39,864	194,327	413,249	539,700	799,024	939,629	1,035,791	1,123,296	1,325,089
8	Connecticut	537,454	622,700	746,258	908,420	1,114,756	1,380,631	1,606,903	1,709,242	2,007,280
9	Delaware	125,015	146,608	168,493	184,735	202,322	223,003	238,380	266,505	318,085
10	District of Columbia	131,700	177,624	230,392	278,718	331,069	437,571	486,869	663,091	802,178

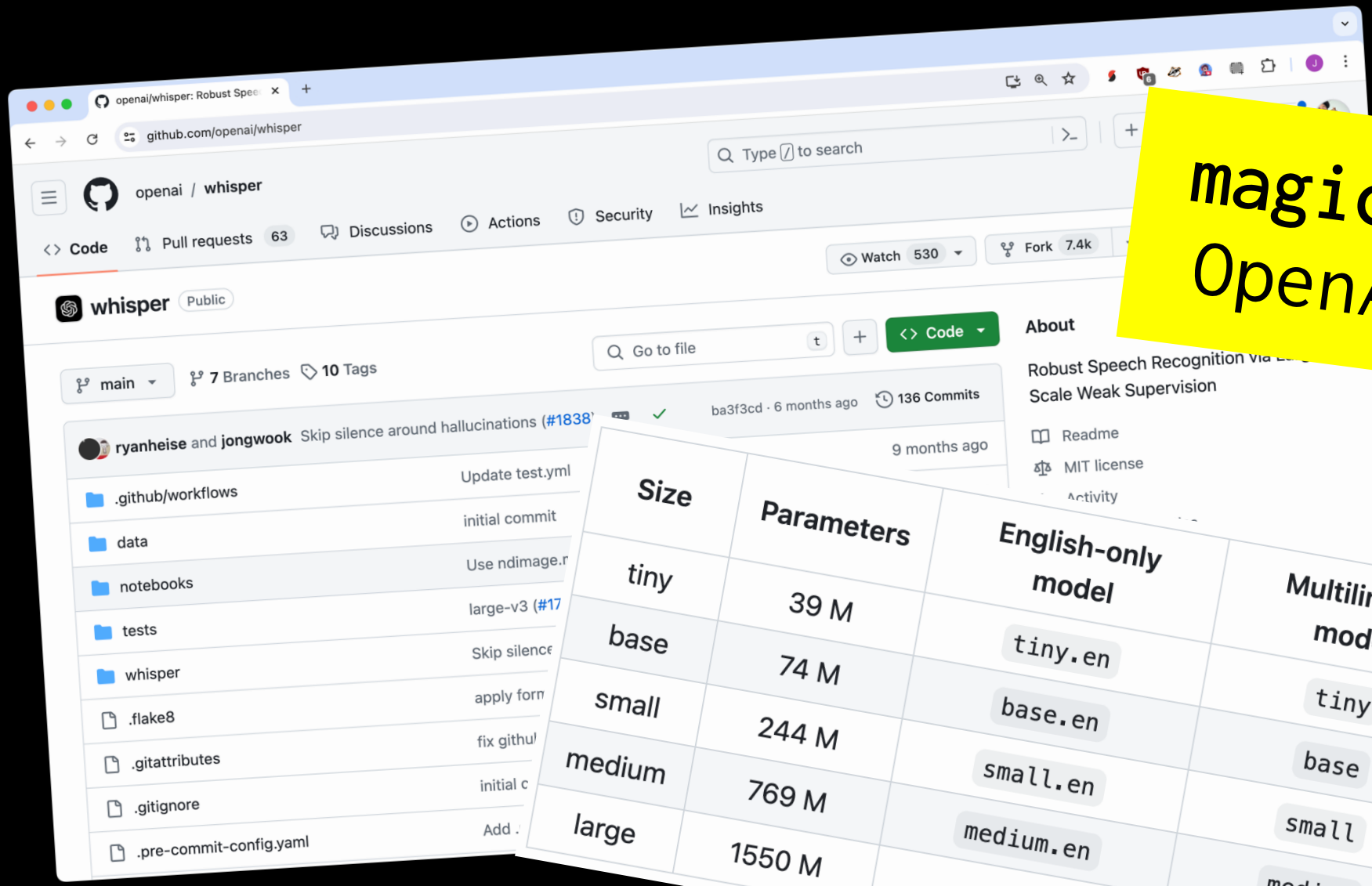
transpose
(swap rows/columns)

	A	B	C	D	E	F	G	H	I	J	K
1	Name	Alabama	Alaska	Arizona	Arkansas	California	Colorado	Connecticut	Delaware	District of Columbia	Florid
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


14M
12M
10M
8M
6M
4M
2M
0
1870
1880
1890
1900
1910
1920
1930
1940
1950
New York
California
Pennsylvania
Illinois
Ohio
Texas
District of
Columbia
Maine
New Jersey
Arizona
Virginia
Oklahoma
Massachusetts
Mississippi
Nevada
Idaho
Montana
New Mexico
West Virginia
Delaware
Hampshire
Get the data • Created with Datawrapper

Transcribing audio

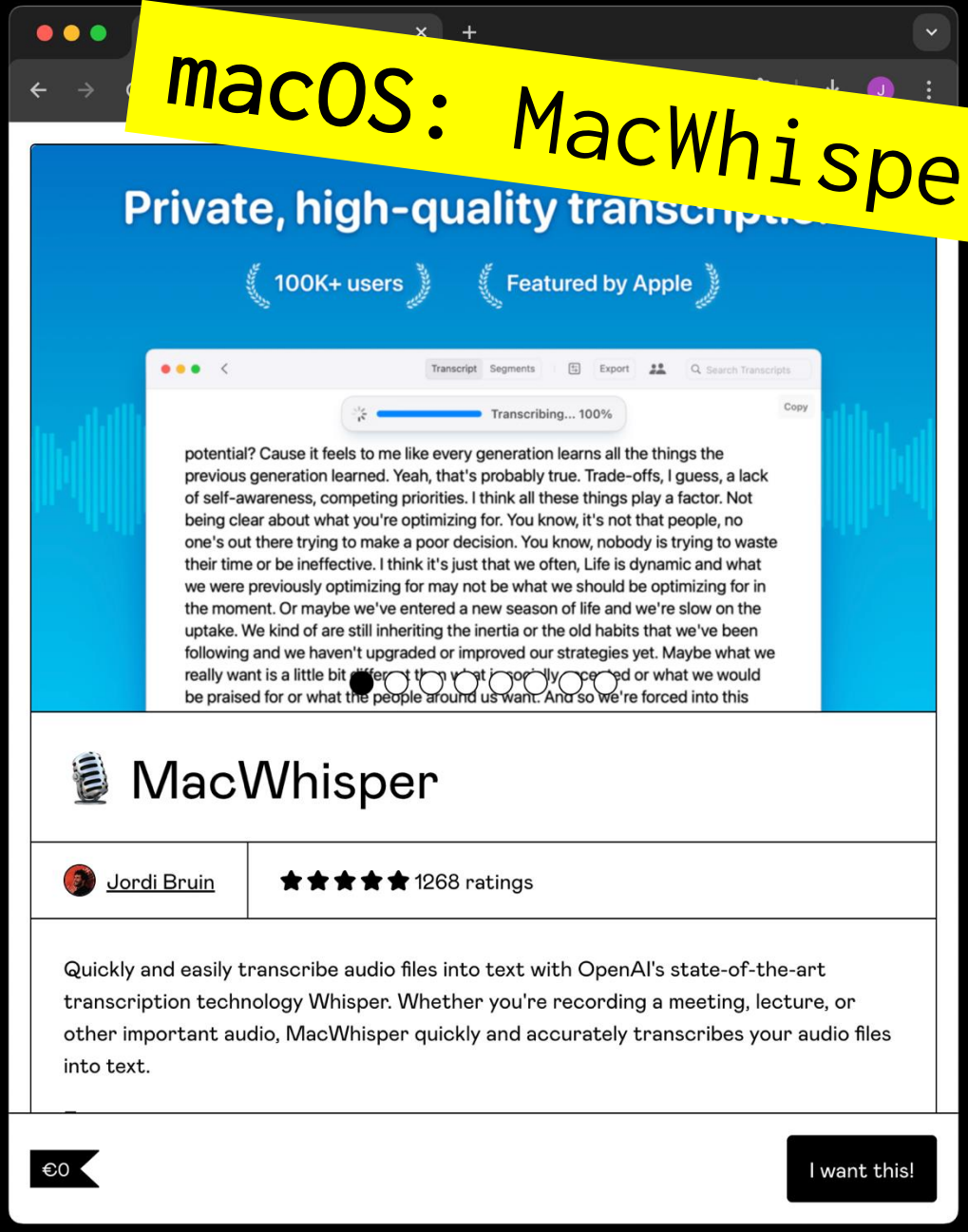

Otter.ai - AI Meeting Note Ta
otter.ai
Otter.ai
Pricing
Apps & Integrations
Resources
Log In
Start for Free
Introducing Meeting GenAI
Never take meeting notes again. Get transcripts, automated summaries, action items, and chat with Otter to get answers from your meeting.
Start for Free
AI Meeting Assistant
Get automated meeting notes and summaries with action items using OtterPilot.
NEW
Otter AI Chat
Get answers and generate content like emails and status updates, using the power of Otter AI Chat across all your meetings.
NEW
AI Channels
Combine live conversations with async updates. Chat with Otter and teammates to get answers and drive projects forward.
NO!
NO!



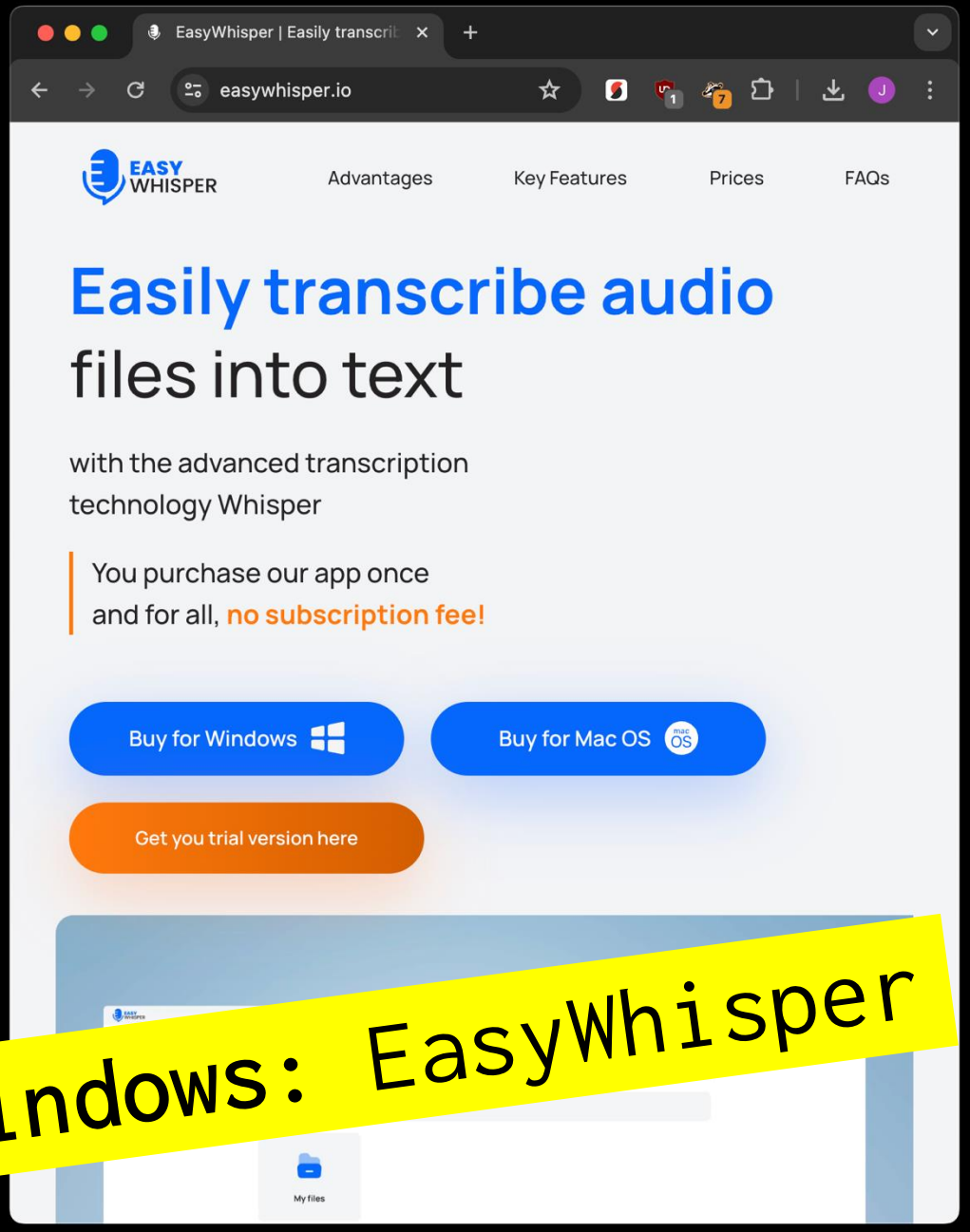
magic technology:
OpenAI's Whisper

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

macOS: MacWhisper



Windows: EasyWhisper



Practical AI for Investigative J

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

Practical AI for Investigative Journalism

Jonathan Soma

6 videos Public

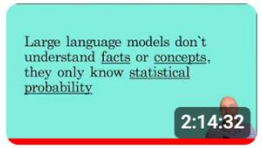
A six session series held in April 2024 about real-life use cases for journalism in (mostly investigative) journ ...more

Play all

Sort

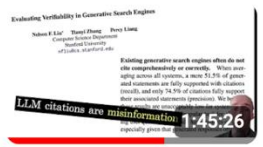
Sorting documents (Practical AI for Investigative Journalism, Session 1)

Jonathan Soma • 1K views • Streamed 2 months ago



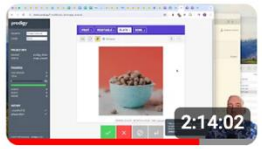
Structured, validated data from LLMs (Practical AI for Investigative Journalism, Session 2)

Jonathan Soma • 716 views • Streamed 2 months ago



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

Jonathan Soma • 572 views • Streamed 1 month ago

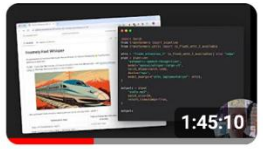


AI, Hugging Face and non-chatbot models (Practical AI for Journalism, Session 4)

Jonathan Soma • 810 views • Streamed 1 month ago

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

Generating structured data with the awful power of LLMs

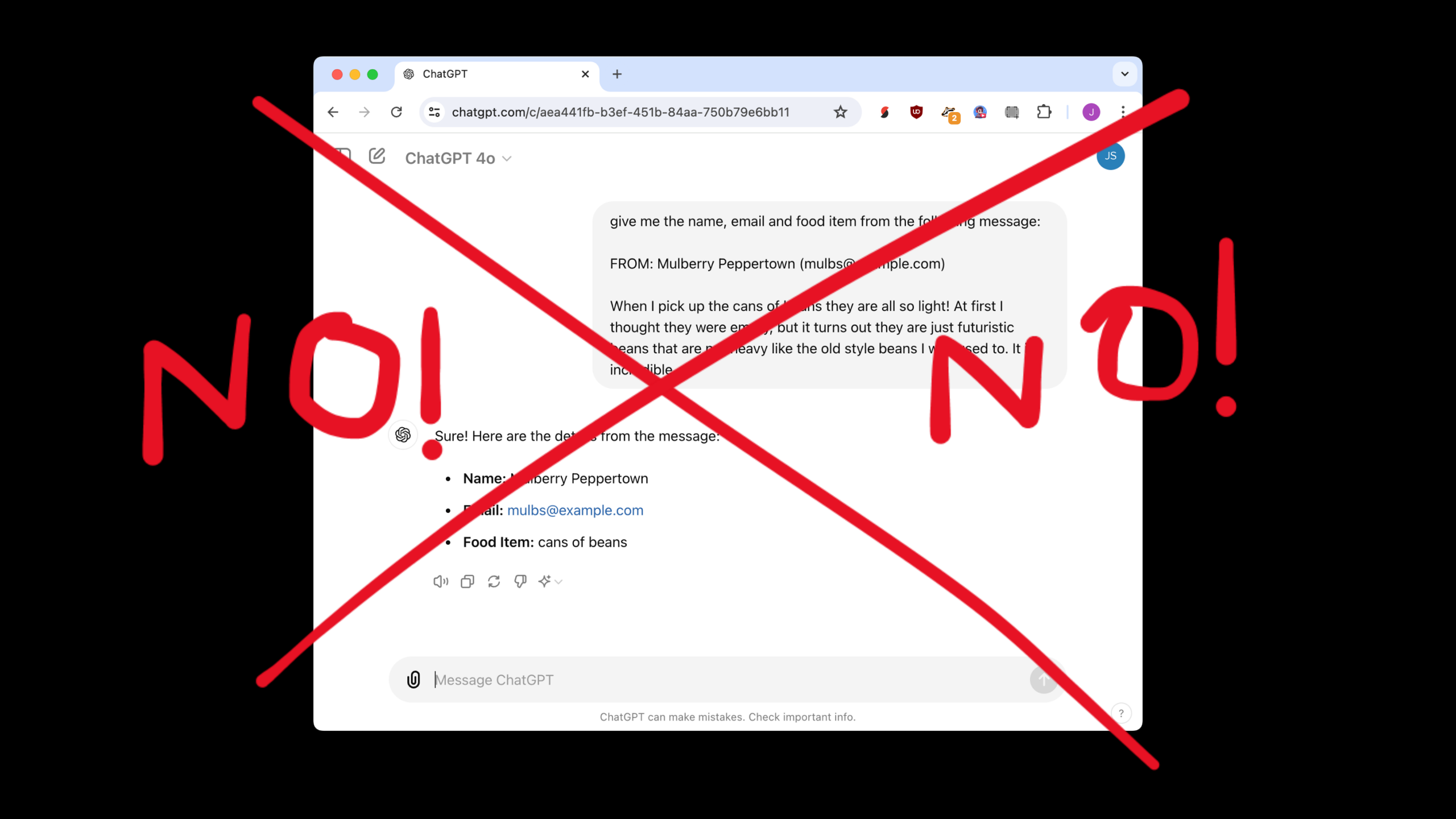
FROM: Mulberry Peppertown (mulbs@example.com)

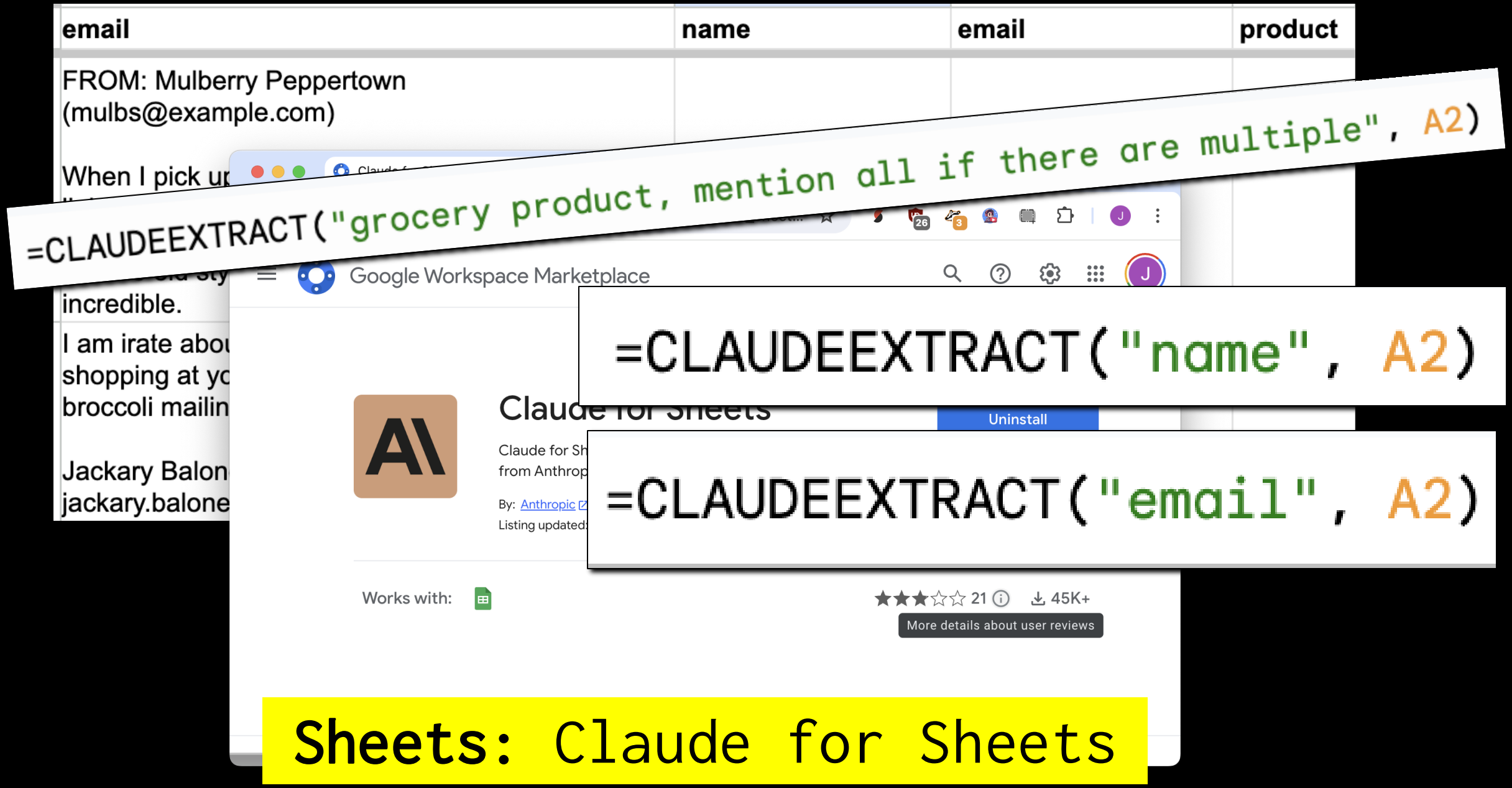
When I pick up the cans of beans they are all so light! At first I thought they were empty, but it turns out they are just futuristic beans that are not heavy like the old style beans I was used to. It is incredible.

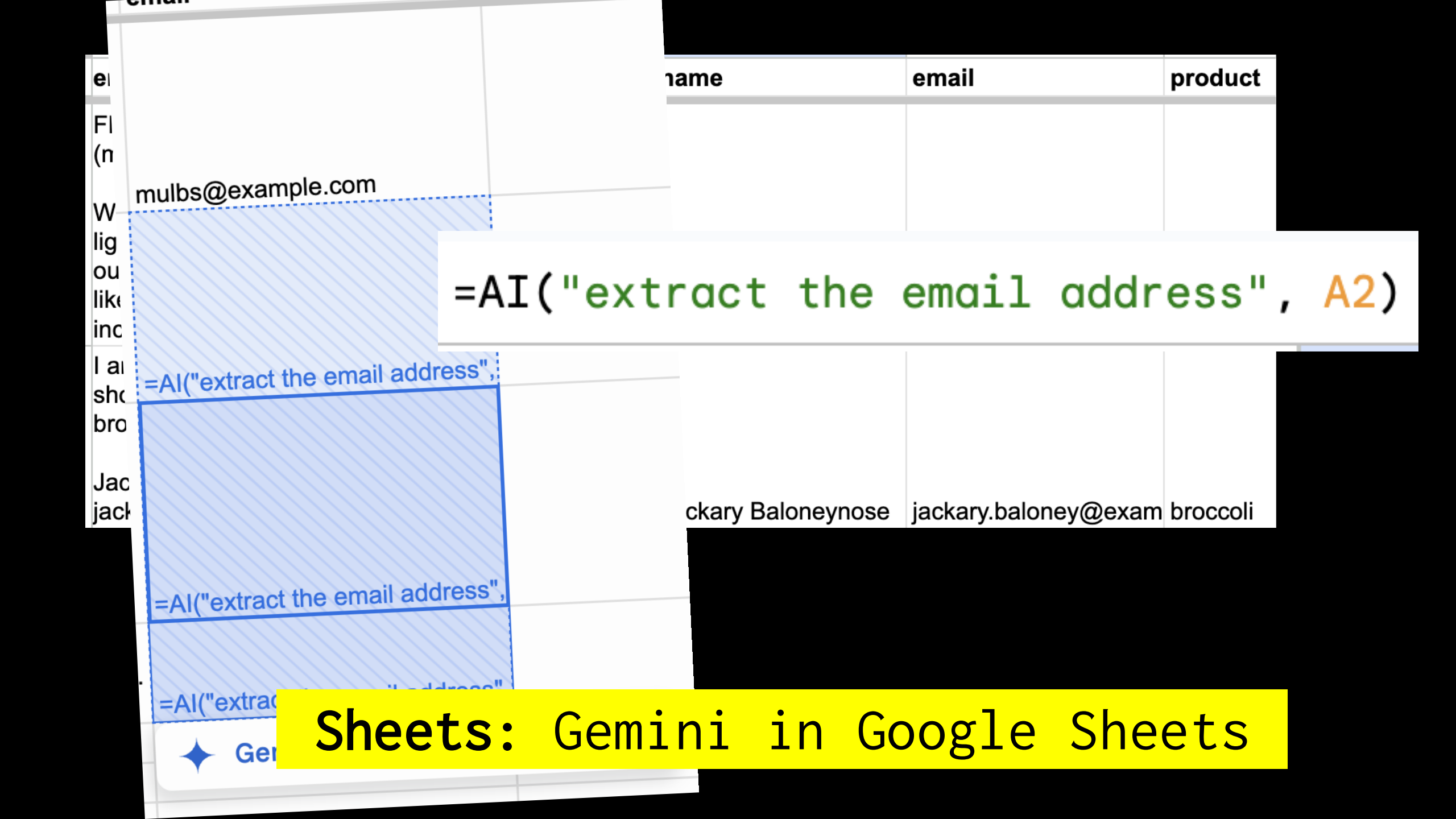
Mulberry Peppertown
name

mulbs@example.com
email

beans
food item


ChatGPT
chatgpt.com/c/aea441fb-b3ef-451b-84aa-750b79e6bb11
ChatGPT 4o
JS
give me the name, email and food item from the fo ng message:
FROM: Mulberry Peppertown (mulbs@ mple.com)
When I pick up the cans of ans they are all so light! At first I thought they were em , but it turns out they are just futuristic eans that are n eavy like the old style beans I w sed to. It inc dible
Sure! Here are the de s from the message:
Name: berry Peppertown
ail: mulbs@example.com
Food Item: cans of beans
Message ChatGPT
ChatGPT can make mistakes. Check important info.
NO!
NO!


email
name
email
product
FROM: Mulberry Peppertown (mulbs@example.com)
When I pick up
incredible.
I am irate abou
shopping at yo
broccoli mailin
Jackary Balon
jackary.balone
=CLAUDEEXTRACT("grocery product, mention all if there are multiple", A2)
Google Workspace Marketplace
=CLAUDEEXTRACT("name", A2)
Claude for Sheets
Uninstall
Claude for Sh
from Anthrop
By: Anthropic
Listing updated:
=CLAUDEEXTRACT("email", A2)
Works with:
21
45K+
More details about user reviews
Sheets: Claude for Sheets


name
email
product
mulbs@example.com
=AI("extract the email address", A2)
=AI("extract the email address",
=AI("extract the email address",
ckary Baloneynose
jackary.baloney@exam
broccoli
Sheets: Gemini in Google Sheets

Python: Structured Outputs

```python
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"]
```

```python
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, b it turns out
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", "conte : comment}],
)
```

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

Time for exercises!!

Visit bit.ly/ire25-cleaning

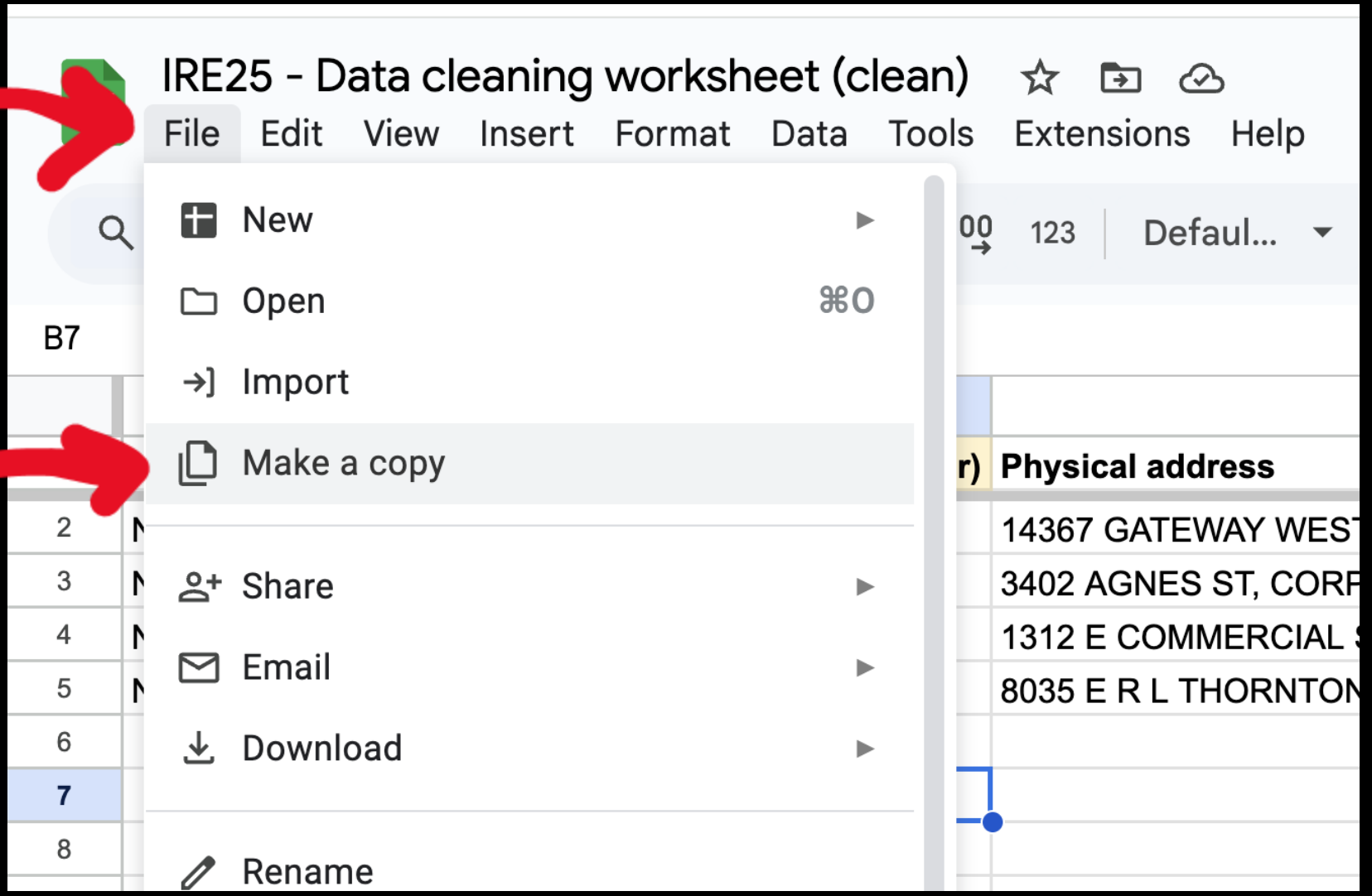
Jonathan Soma
Knight Chair in Data Journalism
Columbia Journalism School
js4571@columbia.edu

jonathansoma.com/everything

Visit bit.ly/ire25-cleaning

File

Make a copy



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 ACESS 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"><img src="SPACER.gif"></td></tr><tr><td width="50%" align="left">Name: SKYHAWK T&R L.L.C.</td>					
006546993C	<table align="center" width="" border="0" cellpadding="0" cellspacing="0"><tbody><tr><td bgcolor="#FFF8DC">Company Information:</td><td bgcolor="#FFF8DC"><img src="SPACER.gif"></td></tr><tr><td width="50%" align="left">Name: GULF COAST ASSET RECOVERY LLC</td>					
006579452C	<table align="center" width="" border="0" cellpadding="0" cellspacing="0"><tbody><tr><td bgcolor="#FFF8DC">Company Information:</td><td bgcolor="#FFF8DC"><img src="SPACER.gif"></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"><img src="SPACER.gif"></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"><img src="SPACER.gif"></td></tr><tr><td width="50%" align="left">Name: SKYHAWK T&R L.L.C.</td>					
006546993C	<table align="center" width="" border="0" cellpadding="0" cellspacing="0"><tbody><tr><td bgcolor="#FFF8DC">Company Information:</td><td bgcolor="#FFF8DC"><img src="SPACER.gif"></td></tr><tr><td width="50%" align="left">Name: GULF COAST ASSET RECOVERY LLC</td>					
006579452C	<table align="center" width="" border="0" cellpadding="0" cellspacing="0"><tbody><tr><td bgcolor="#FFF8DC">Company Information:</td><td bgcolor="#FFF8DC"><img src="SPACER.gif"></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"><img src="SPACER.gif"></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 WRIS
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 FI
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 AVULSIO
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