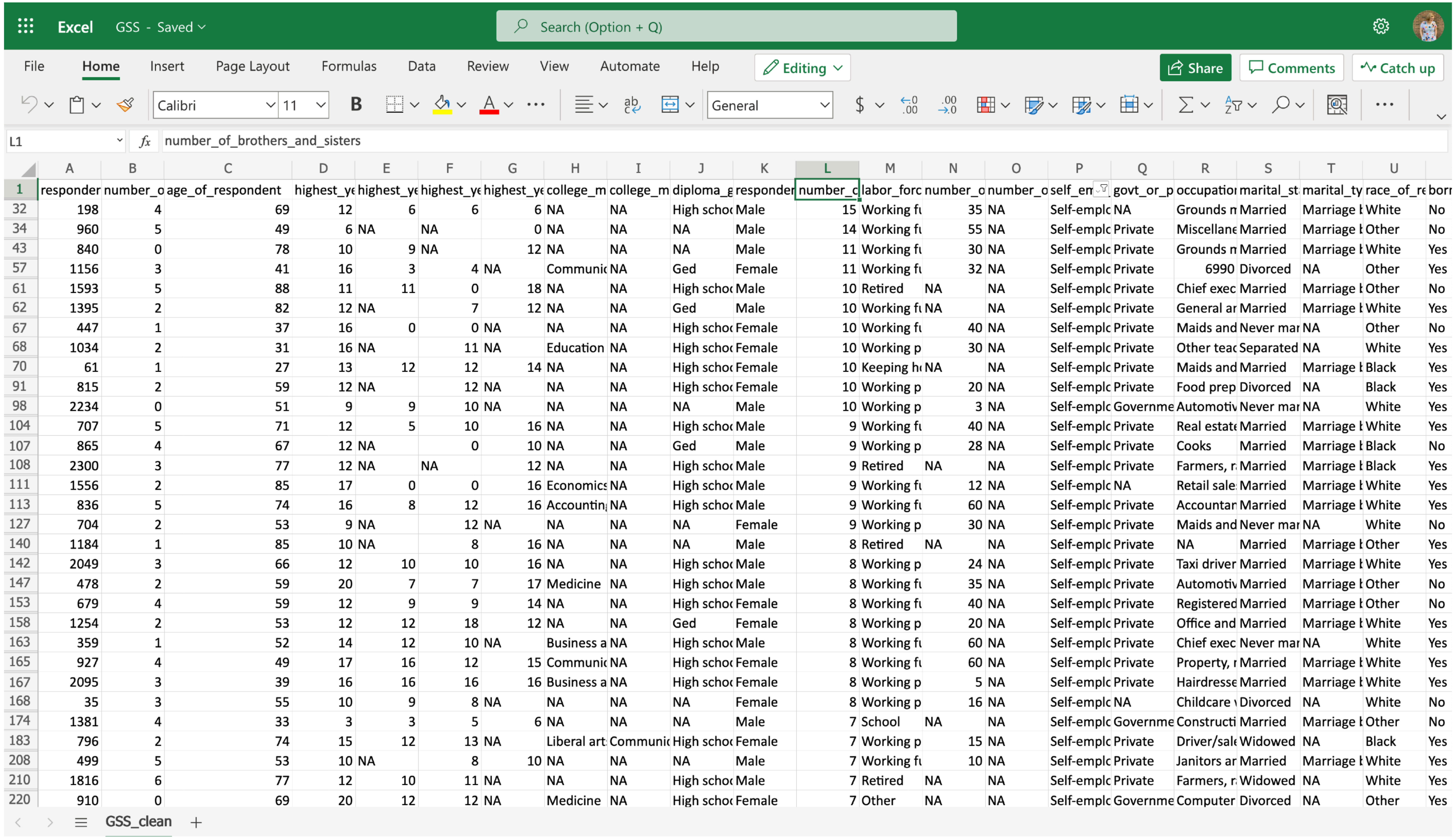
What is data?

Amelia McNamara

flickr: jannekestaaks

It is easy to think of data as spreadsheets



L1 | fx | number_of_brothers_and_sisters

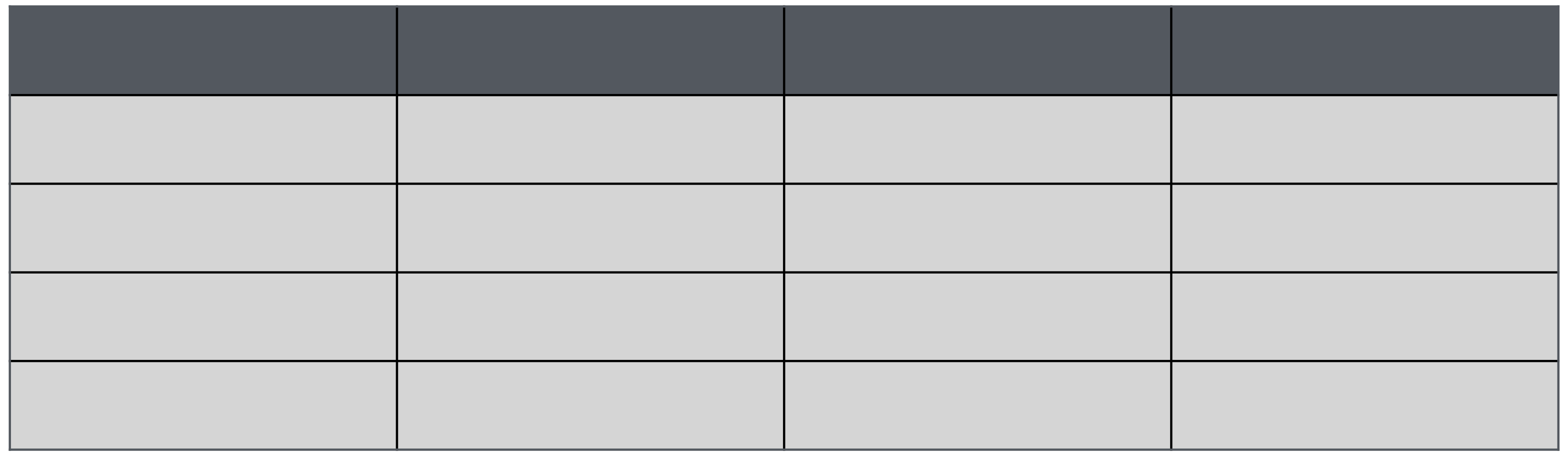
	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V
1	responder	number_o	age_of_respondent	highest_ye	highest_ye	highest_ye	highest_ye	college_m	college_m	diploma_g	responder	number_c	labor_forc	number_o	number_o	self_em	govt_or_p	occupation	marital_st	marital_ty	race_of_re	borr
32	198	4	69	12	6	6	6	NA	NA	High schoo	Male	15	Working fu	35	NA	Self-emplo	NA	Grounds m	Married	Marriage b	White	No
34	960	5	49	6	NA	NA	0	NA	NA	NA	Male	14	Working fu	55	NA	Self-emplo	Private	Miscellane	Married	Marriage b	Other	No
43	840	0	78	10	9	NA	12	NA	NA	NA	Male	11	Working fu	30	NA	Self-emplo	Private	Grounds m	Married	Marriage b	White	Yes
57	1156	3	41	16	3	4	NA	Communic	NA	Ged	Female	11	Working fu	32	NA	Self-emplo	Private	6990	Divorced	NA	Other	Yes
61	1593	5	88	11	11	0	18	NA	NA	High schoo	Male	10	Retired	NA	NA	Self-emplo	Private	Chief exec	Married	Marriage b	Other	No
62	1395	2	82	12	NA	7	12	NA	NA	Ged	Male	10	Working fu	NA	NA	Self-emplo	Private	General ar	Married	Marriage b	White	Yes
67	447	1	37	16	0	0	NA	NA	NA	High schoo	Female	10	Working fu	40	NA	Self-emplo	Private	Maids and	Never mar	NA	Other	No
68	1034	2	31	16	NA	11	NA	Education	NA	High schoo	Female	10	Working p	30	NA	Self-emplo	Private	Other teac	Separated	NA	White	Yes
70	61	1	27	13	12	12	14	NA	NA	High schoo	Female	10	Keeping h	NA	NA	Self-emplo	Private	Maids and	Married	Marriage b	Black	Yes
91	815	2	59	12	NA	12	NA	NA	NA	High schoo	Female	10	Working p	20	NA	Self-emplo	Private	Food prep	Divorced	NA	Black	Yes
98	2234	0	51	9	9	10	NA	NA	NA	NA	Male	10	Working p	3	NA	Self-emplo	Governme	Automotiv	Never mar	NA	White	Yes
104	707	5	71	12	5	10	16	NA	NA	High schoo	Male	9	Working fu	40	NA	Self-emplo	Private	Real estate	Married	Marriage b	White	Yes
107	865	4	67	12	NA	0	10	NA	NA	Ged	Male	9	Working p	28	NA	Self-emplo	Private	Cooks	Married	Marriage b	Black	No
108	2300	3	77	12	NA	NA	12	NA	NA	High schoo	Male	9	Retired	NA	NA	Self-emplo	Private	Farmers, r	Married	Marriage b	Black	Yes
111	1556	2	85	17	0	0	16	Economics	NA	High schoo	Male	9	Working fu	12	NA	Self-emplo	NA	Retail sale	Married	Marriage b	White	Yes
113	836	5	74	16	8	12	16	Accountin	NA	High schoo	Male	9	Working fu	60	NA	Self-emplo	Private	Accountan	Married	Marriage b	White	Yes
127	704	2	53	9	NA	12	NA	NA	NA	NA	Female	9	Working p	30	NA	Self-emplo	Private	Maids and	Never mar	NA	White	No
140	1184	1	85	10	NA	8	16	NA	NA	NA	Male	8	Retired	NA	NA	Self-emplo	Private	NA	Married	Marriage b	Other	Yes
142	2049	3	66	12	10	10	16	NA	NA	High schoo	Male	8	Working p	24	NA	Self-emplo	Private	Taxi driver	Married	Marriage b	White	Yes
147	478	2	59	20	7	7	17	Medicine	NA	High schoo	Male	8	Working fu	35	NA	Self-emplo	Private	Automotiv	Married	Marriage b	Other	No
153	679	4	59	12	9	9	14	NA	NA	High schoo	Female	8	Working fu	40	NA	Self-emplo	Private	Registered	Married	Marriage b	Other	No
158	1254	2	53	12	12	18	12	NA	NA	Ged	Female	8	Working p	20	NA	Self-emplo	Private	Office and	Married	Marriage b	White	Yes
163	359	1	52	14	12	10	NA	Business a	NA	High schoo	Male	8	Working fu	60	NA	Self-emplo	Private	Chief exec	Never mar	NA	White	Yes
165	927	4	49	17	16	12	15	Communic	NA	High schoo	Female	8	Working fu	60	NA	Self-emplo	Private	Property, r	Married	Marriage b	White	Yes
167	2095	3	39	16	16	16	16	Business a	NA	High schoo	Female	8	Working p	5	NA	Self-emplo	Private	Hairdresse	Married	Marriage b	White	Yes
168	35	3	55	10	9	8	NA	NA	NA	NA	Female	8	Working p	16	NA	Self-emplo	NA	Childcare w	Divorced	NA	White	No
174	1381	4	33	3	3	5	6	NA	NA	NA	Male	7	School	NA	NA	Self-emplo	Governme	Constructi	Married	Marriage b	Other	No
183	796	2	74	15	12	13	NA	Liberal art	Communic	High schoo	Female	7	Working p	15	NA	Self-emplo	Private	Driver/sale	Widowed	NA	Black	Yes
208	499	5	53	10	NA	8	10	NA	NA	NA	Male	7	Working fu	10	NA	Self-emplo	Private	Janitors ar	Married	Marriage b	White	Yes
210	1816	6	77	12	10	11	NA	NA	NA	High schoo	Male	7	Retired	NA	NA	Self-emplo	Private	Farmers, r	Widowed	NA	White	Yes
220	910	0	69	20	12	12	NA	Medicine	NA	High schoo	Female	7	Other	NA	NA	Self-emplo	Governme	Computer	Divorced	NA	Other	Yes

There's more to data

- Data is always generated by humans
- It can be numbers, categories, text, images, or any other type of record!
- The encoding of data was always a choice made by someone
- The most common way we characterize data in statistics is as a set of variables that capture various aspects of the world, and observations over those variables.

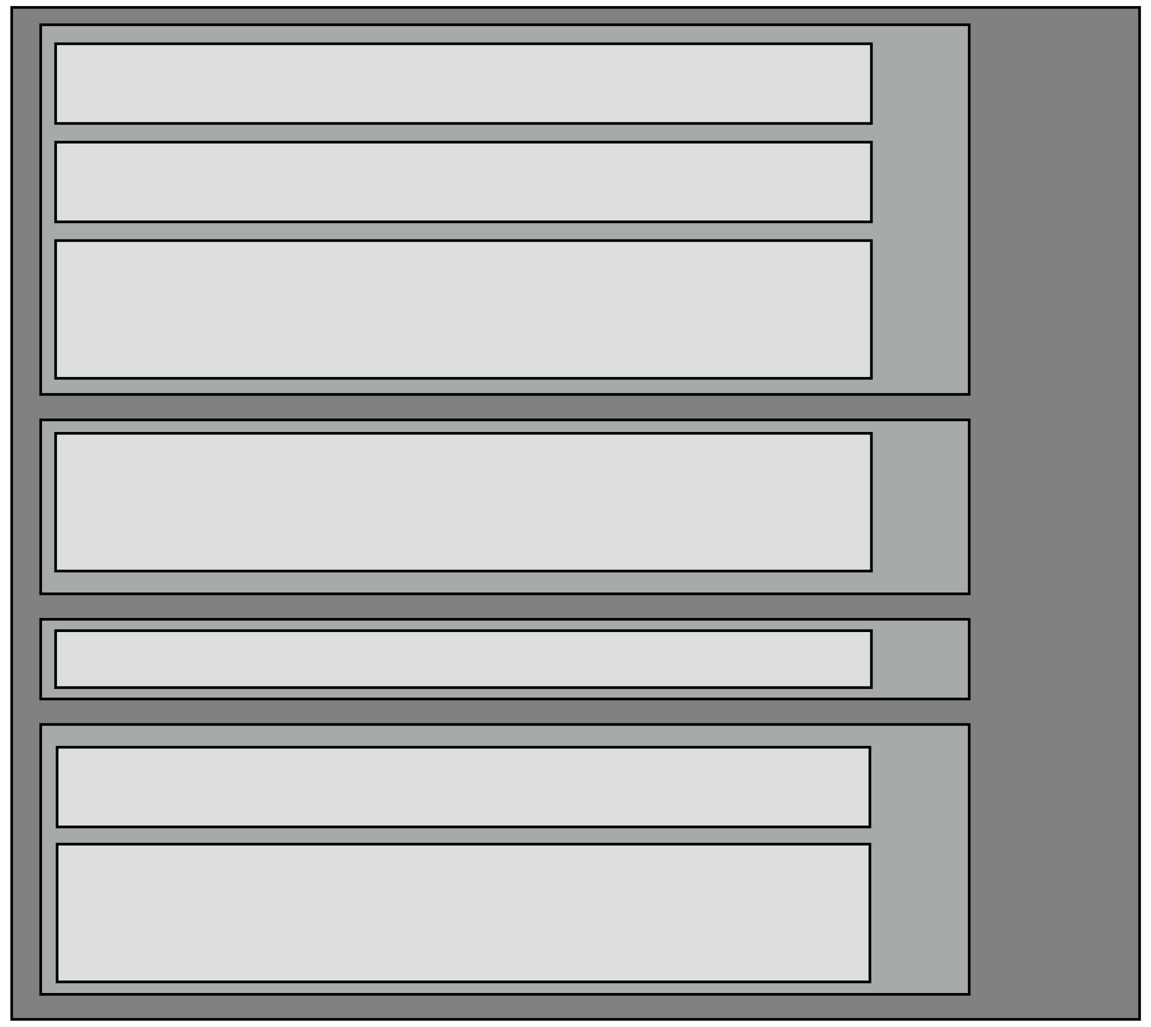
Tidy data

- Rows are observations (things we observe)
- Columns are variables (things that vary)



“Untidy” data can take other forms

- There is nothing inherently wrong about untidy data
- However, statistical methods expect tidy data so “wrangling” may be necessary



(Some very untidy data)

Joint work with students from my STAT 336 class, spring 2024

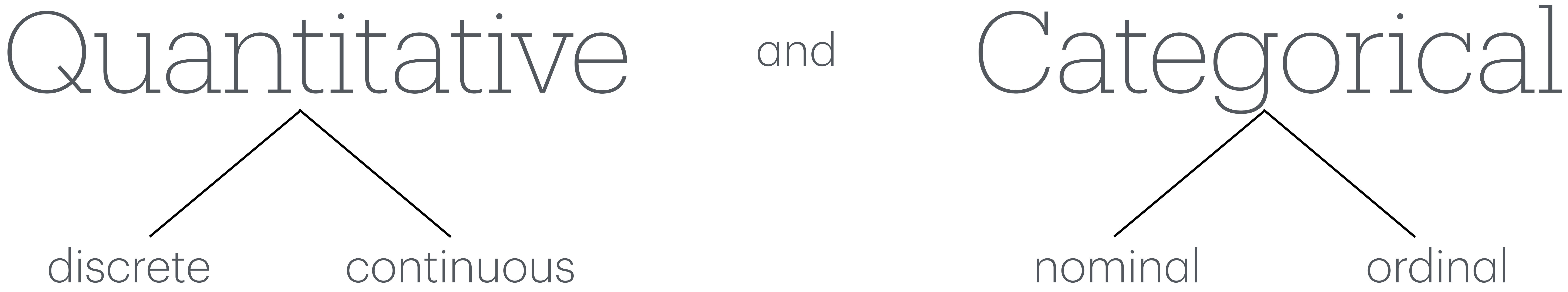
THE TECHNOLOGY OF TEACHING
being a teacher
Mentoring At-Risk Students through the Hidden Curriculum
SMITH
MCNAMARA

We often make the distinction between

Quantitative and Categorical

Let’s brainstorm some variables that could be recorded about us, and whether they are quantitative or categorical.

We often make the distinction between



Quantitative and Categorical

discrete continuous

nominal ordinal

Let’s brainstorm some variables that could be recorded about us, and whether they are quantitative or categorical.

Sometimes data is collected in a way we see

- The Census
- Pew Research surveys
- Science!
- ...and of course, many more

Sometimes data is generated for one reason and then used for another

- Health information about you at the doctor
- Location information from social media posts
- Emails (think Enron trove)
- …and more!

I often think of the data I unintentionally generate on a daily basis as “data exhaust.”
(Recall when D’Ignazio and Klein mentioned data being “the new oil”?)

Data science often serves “the three Ss: science (universities), surveillance (governments), and selling (corporations).”

Data Feminism, D’Ignzaio & Klein

Brainstorm: data exhaust

We generate data all the time, whether we're aware of it or not.

For example, I have a Withings watch, so I generate data every time I take a step. I consciously chose to wear this, but there are other times I am unconsciously generating data. It is incidental to what I'm doing, and streams off me as "data exhaust."

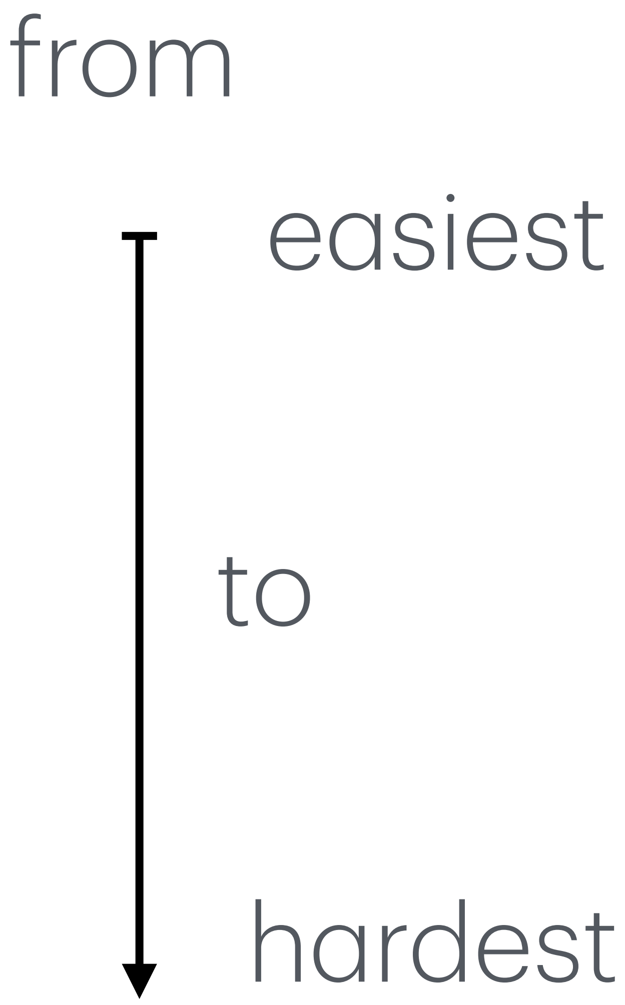
Take a few minutes and make a list of all the places you generated data today/this week/on a normal day.

flickr: severalseconds

“What gets counted, counts”

Joni Seager, via Data Feminism

Getting data



from

easiest

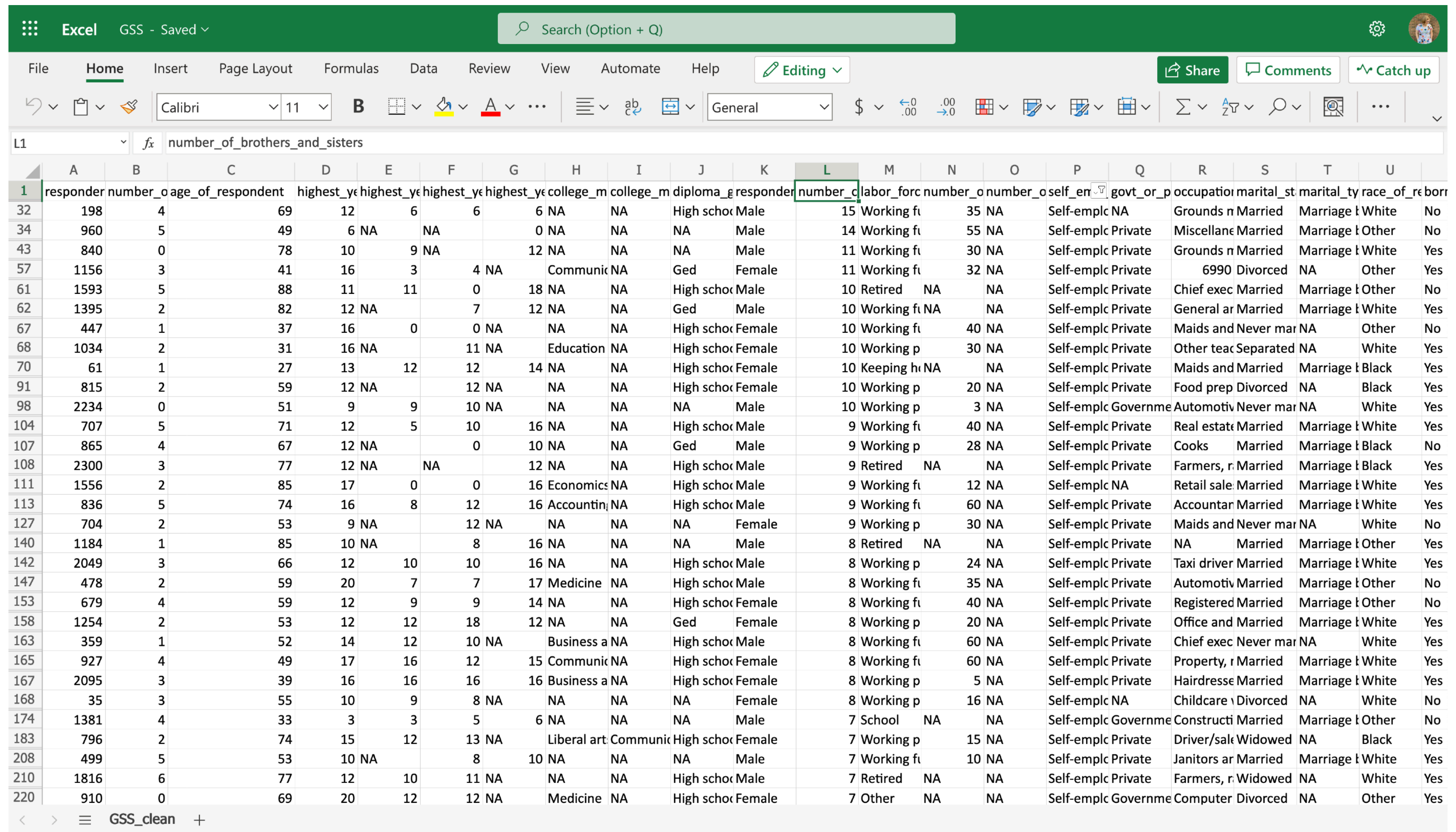
to

hardest

The very easiest

Data already in a nice format

- .csv
- Excel
- .txt
- .dat
- .sav



L1 *fx* number_of_brothers_and_sisters

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	
1	responder	number_o	age_of_respondent	highest_ye	highest_ye	highest_ye	highest_ye	college_m	college_m	diploma_g	responder	number_o	labor_forc	number_o	number_o	self_em	govt_or_p	occupation	marital_st	marital_ty	race_of_re	born
32	198	4	69	12	6	6	6	NA	NA	High schoo	Male	15	Working fu	35	NA	Self-emplo	NA	Grounds m	Married	Marriage	White	No
34	960	5	49	6	NA	NA	0	NA	NA	NA	Male	14	Working fu	55	NA	Self-emplo	Private	Miscellane	Married	Marriage	Other	No
43	840	0	78	10	9	NA	12	NA	NA	NA	Male	11	Working fu	30	NA	Self-emplo	Private	Grounds m	Married	Marriage	White	Yes
57	1156	3	41	16	3	4	NA	Communic	NA	Ged	Female	11	Working fu	32	NA	Self-emplo	Private	6990	Divorced	NA	Other	Yes
61	1593	5	88	11	11	0	18	NA	NA	High schoo	Male	10	Retired	NA	NA	Self-emplo	Private	Chief exec	Married	Marriage	Other	No
62	1395	2	82	12	NA	7	12	NA	NA	Ged	Male	10	Working fu	NA	NA	Self-emplo	Private	General ar	Married	Marriage	White	Yes
67	447	1	37	16	0	0	NA	NA	NA	High schoo	Female	10	Working fu	40	NA	Self-emplo	Private	Maids and	Never mar	NA	Other	No
68	1034	2	31	16	NA	11	NA	Education	NA	High schoo	Female	10	Working p	30	NA	Self-emplo	Private	Other teac	Separated	NA	White	Yes
70	61	1	27	13	12	12	14	NA	NA	High schoo	Female	10	Keeping h	NA	NA	Self-emplo	Private	Maids and	Married	Marriage	Black	Yes
91	815	2	59	12	NA	12	NA	NA	NA	High schoo	Female	10	Working p	20	NA	Self-emplo	Private	Food prep	Divorced	NA	Black	Yes
98	2234	0	51	9	9	10	NA	NA	NA	NA	Male	10	Working p	3	NA	Self-emplo	Governme	Automotiv	Never mar	NA	White	Yes
104	707	5	71	12	5	10	16	NA	NA	High schoo	Male	9	Working fu	40	NA	Self-emplo	Private	Real estate	Married	Marriage	White	Yes
107	865	4	67	12	NA	0	10	NA	NA	Ged	Male	9	Working p	28	NA	Self-emplo	Private	Cooks	Married	Marriage	Black	No
108	2300	3	77	12	NA	NA	12	NA	NA	High schoo	Male	9	Retired	NA	NA	Self-emplo	Private	Farmers, r	Married	Marriage	Black	Yes
111	1556	2	85	17	0	0	16	Economics	NA	High schoo	Male	9	Working fu	12	NA	Self-emplo	NA	Retail sale	Married	Marriage	White	Yes
113	836	5	74	16	8	12	16	Accountin	NA	High schoo	Male	9	Working fu	60	NA	Self-emplo	Private	Accountan	Married	Marriage	White	Yes
127	704	2	53	9	NA	12	NA	NA	NA	NA	Female	9	Working p	30	NA	Self-emplo	Private	Maids and	Never mar	NA	White	No
140	1184	1	85	10	NA	8	16	NA	NA	NA	Male	8	Retired	NA	NA	Self-emplo	Private	NA	Married	Marriage	Other	Yes
142	2049	3	66	12	10	10	16	NA	NA	High schoo	Male	8	Working p	24	NA	Self-emplo	Private	Taxi driver	Married	Marriage	White	Yes
147	478	2	59	20	7	7	17	Medicine	NA	High schoo	Male	8	Working fu	35	NA	Self-emplo	Private	Automotiv	Married	Marriage	Other	No
153	679	4	59	12	9	9	14	NA	NA	High schoo	Female	8	Working fu	40	NA	Self-emplo	Private	Registered	Married	Marriage	Other	No
158	1254	2	53	12	12	18	12	NA	NA	Ged	Female	8	Working p	20	NA	Self-emplo	Private	Office and	Married	Marriage	White	Yes
163	359	1	52	14	12	10	NA	Business a	NA	High schoo	Male	8	Working fu	60	NA	Self-emplo	Private	Chief exec	Never mar	NA	White	Yes
165	927	4	49	17	16	12	15	Communic	NA	High schoo	Female	8	Working fu	60	NA	Self-emplo	Private	Property, r	Married	Marriage	White	Yes
167	2095	3	39	16	16	16	16	Business a	NA	High schoo	Female	8	Working p	5	NA	Self-emplo	Private	Hairdresse	Married	Marriage	White	Yes
168	35	3	55	10	9	8	NA	NA	NA	NA	Female	8	Working p	16	NA	Self-emplo	NA	Childcare	Divorced	NA	White	No
174	1381	4	33	3	3	5	6	NA	NA	NA	Male	7	School	NA	NA	Self-emplo	Governme	Constructi	Married	Marriage	Other	No
183	796	2	74	15	12	13	NA	Liberal art	Communic	High schoo	Female	7	Working p	15	NA	Self-emplo	Private	Driver/sale	Widowed	NA	Black	Yes
208	499	5	53	10	NA	8	10	NA	NA	NA	Male	7	Working fu	10	NA	Self-emplo	Private	Janitors ar	Married	Marriage	White	Yes
210	1816	6	77	12	10	11	NA	NA	NA	High schoo	Male	7	Retired	NA	NA	Self-emplo	Private	Farmers, r	Widowed	NA	White	Yes
220	910	0	69	20	12	12	NA	Medicine	NA	High schoo	Female	7	Other	NA	NA	Self-emplo	Governme	Computer	Divorced	NA	Other	Yes

Slightly harder

Data in a computer format, not rectangular

- .json
- .xml

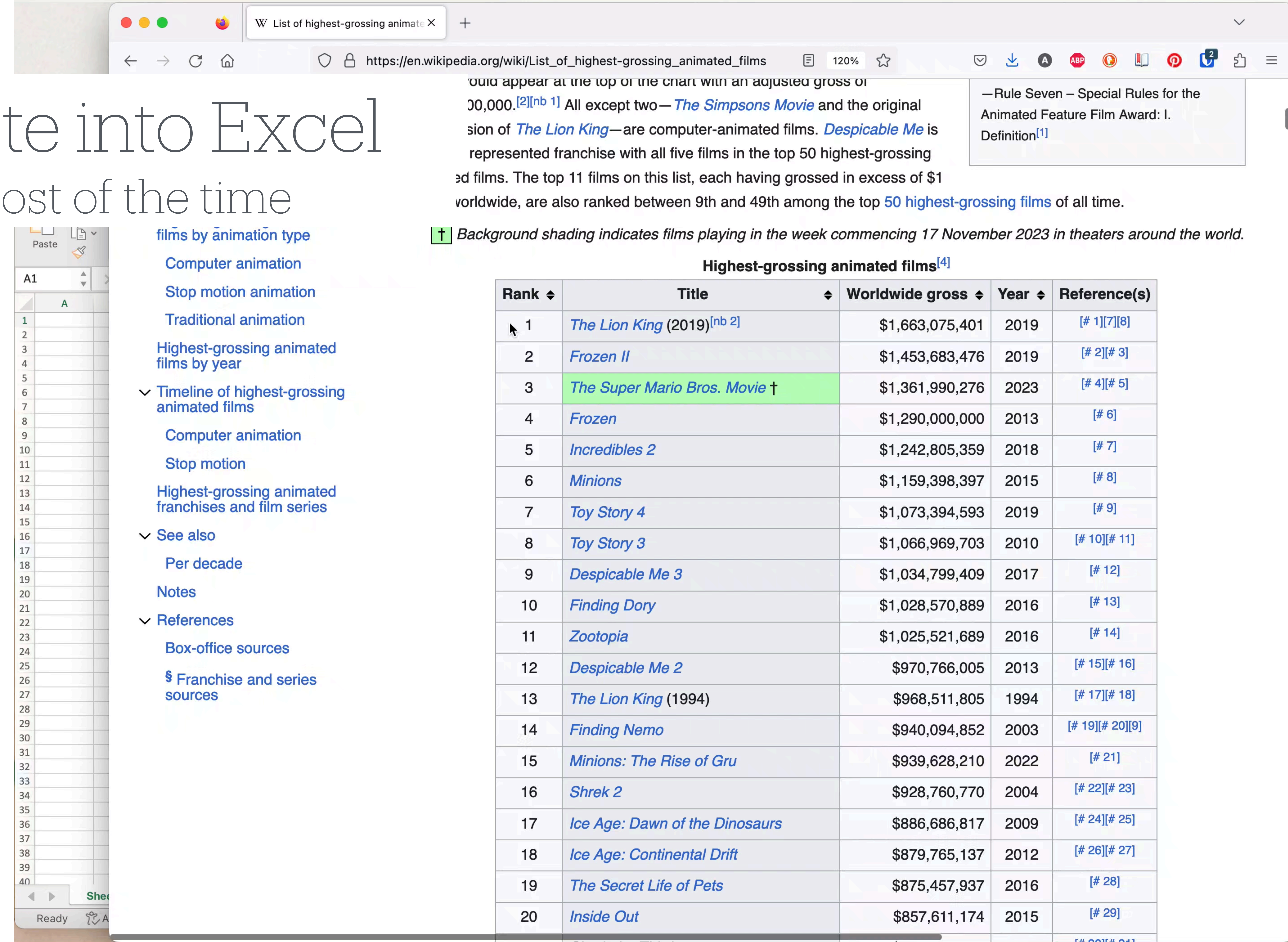
A little harder

The data is electronic, but not in a file

- Manual approach
  - Copy-paste (😱)
  - {datapasta} (😄)
- Automated approach
  - APIs
  - Scraping

Copy-paste into Excel

This works… most of the time


List of highest-grossing animate
https://en.wikipedia.org/wiki/List_of_highest-grossing_animated_films
120%
Paste
A1
A
Shee
Ready

ould appear at the top of the chart with an adjusted gross of
)0,000.[2][nb 1] All except two—*The Simpsons Movie* and the original
sion of *The Lion King*—are computer-animated films. *Despicable Me* is
represented franchise with all five films in the top 50 highest-grossing
ed films. The top 11 films on this list, each having grossed in excess of $1
vorldwide, are also ranked between 9th and 49th among the top 50 highest-grossing films of all time.

—Rule Seven – Special Rules for the Animated Feature Film Award: I. Definition[1]

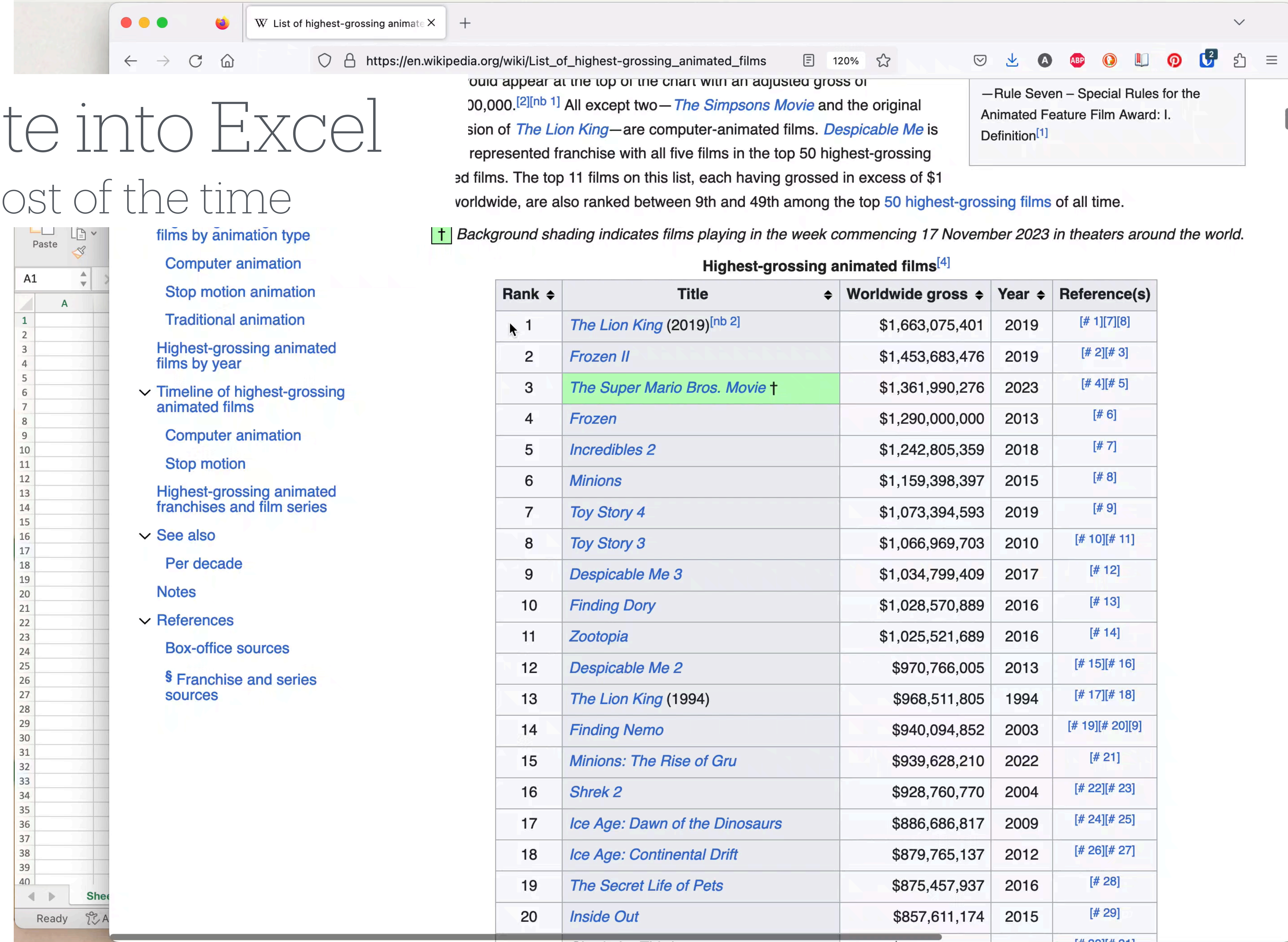
† *Background shading indicates films playing in the week commencing 17 November 2023 in theaters around the world.*

Highest-grossing animated films[4]

Rank	Title	Worldwide gross	Year	Reference(s)
1	*The Lion King* (2019)[nb 2]	$1,663,075,401	2019	[# 1][7][8]
2	*Frozen II*	$1,453,683,476	2019	[# 2][# 3]
3	*The Super Mario Bros. Movie* †	$1,361,990,276	2023	[# 4][# 5]
4	*Frozen*	$1,290,000,000	2013	[# 6]
5	*Incredibles 2*	$1,242,805,359	2018	[# 7]
6	*Minions*	$1,159,398,397	2015	[# 8]
7	*Toy Story 4*	$1,073,394,593	2019	[# 9]
8	*Toy Story 3*	$1,066,969,703	2010	[# 10][# 11]
9	*Despicable Me 3*	$1,034,799,409	2017	[# 12]
10	*Finding Dory*	$1,028,570,889	2016	[# 13]
11	*Zootopia*	$1,025,521,689	2016	[# 14]
12	*Despicable Me 2*	$970,766,005	2013	[# 15][# 16]
13	*The Lion King* (1994)	$968,511,805	1994	[# 17][# 18]
14	*Finding Nemo*	$940,094,852	2003	[# 19][# 20][9]
15	*Minions: The Rise of Gru*	$939,628,210	2022	[# 21]
16	*Shrek 2*	$928,760,770	2004	[# 22][# 23]
17	*Ice Age: Dawn of the Dinosaurs*	$886,686,817	2009	[# 24][# 25]
18	*Ice Age: Continental Drift*	$879,765,137	2012	[# 26][# 27]
19	*The Secret Life of Pets*	$875,457,937	2016	[# 28]
20	*Inside Out*	$857,611,174	2015	[# 29]

Copy-paste into Excel

This works… most of the time



List of highest-grossing animate
https://en.wikipedia.org/wiki/List_of_highest-grossing_animated_films
120%
Paste
A1
A
Shee
Ready

ould appear at the top of the chart with an adjusted gross of
00,000.[2][nb 1] All except two—*The Simpsons Movie* and the original
sion of *The Lion King*—are computer-animated films. *Despicable Me* is
represented franchise with all five films in the top 50 highest-grossing
ed films. The top 11 films on this list, each having grossed in excess of $1
worldwide, are also ranked between 9th and 49th among the top 50 highest-grossing films of all time.

—Rule Seven – Special Rules for the Animated Feature Film Award: I. Definition[1]

† *Background shading indicates films playing in the week commencing 17 November 2023 in theaters around the world.*

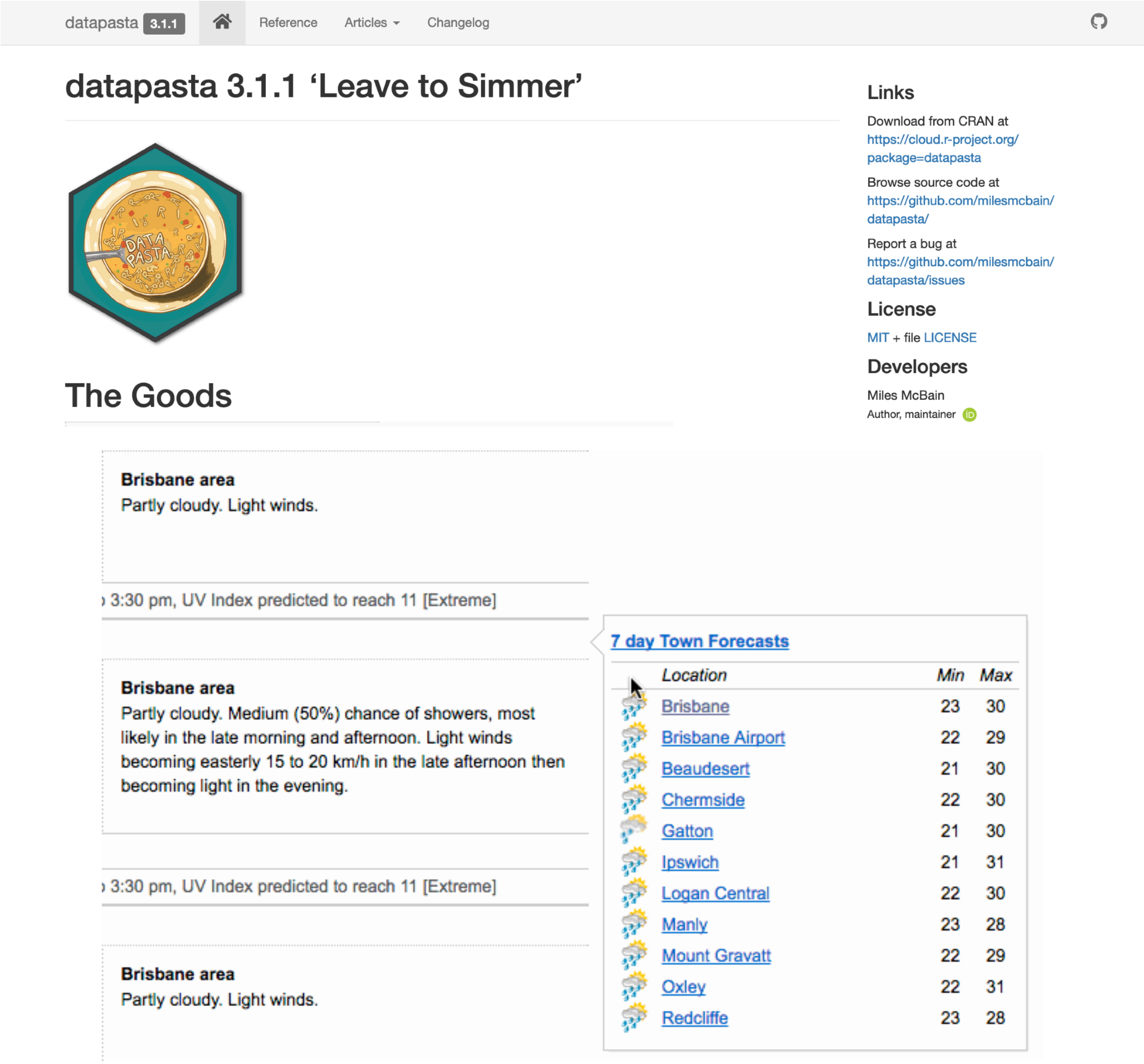
Highest-grossing animated films[4]

Rank	Title	Worldwide gross	Year	Reference(s)
1	*The Lion King* (2019)[nb 2]	$1,663,075,401	2019	[# 1][7][8]
2	*Frozen II*	$1,453,683,476	2019	[# 2][# 3]
3	*The Super Mario Bros. Movie* †	$1,361,990,276	2023	[# 4][# 5]
4	*Frozen*	$1,290,000,000	2013	[# 6]
5	*Incredibles 2*	$1,242,805,359	2018	[# 7]
6	*Minions*	$1,159,398,397	2015	[# 8]
7	*Toy Story 4*	$1,073,394,593	2019	[# 9]
8	*Toy Story 3*	$1,066,969,703	2010	[# 10][# 11]
9	*Despicable Me 3*	$1,034,799,409	2017	[# 12]
10	*Finding Dory*	$1,028,570,889	2016	[# 13]
11	*Zootopia*	$1,025,521,689	2016	[# 14]
12	*Despicable Me 2*	$970,766,005	2013	[# 15][# 16]
13	*The Lion King* (1994)	$968,511,805	1994	[# 17][# 18]
14	*Finding Nemo*	$940,094,852	2003	[# 19][# 20][9]
15	*Minions: The Rise of Gru*	$939,628,210	2022	[# 21]
16	*Shrek 2*	$928,760,770	2004	[# 22][# 23]
17	*Ice Age: Dawn of the Dinosaurs*	$886,686,817	2009	[# 24][# 25]
18	*Ice Age: Continental Drift*	$879,765,137	2012	[# 26][# 27]
19	*The Secret Life of Pets*	$875,457,937	2016	[# 28]
20	*Inside Out*	$857,611,174	2015	[# 29]

{datapasta}

This works…

even more of the time



datapasta 3.1.1 ‘Leave to Simmer’

The Goods

Brisbane area
Partly cloudy. Light winds.

› 3:30 pm, UV Index predicted to reach 11 [Extreme]

Brisbane area
Partly cloudy. Medium (50%) chance of showers, most likely in the late morning and afternoon. Light winds becoming easterly 15 to 20 km/h in the late afternoon then becoming light in the evening.

› 3:30 pm, UV Index predicted to reach 11 [Extreme]

Brisbane area
Partly cloudy. Light winds.

7 day Town Forecasts

Location	*Min*	*Max*
Brisbane	23	30
Brisbane Airport	22	29
Beaudesert	21	30
Chermside	22	30
Gatton	21	30
Ipswich	21	31
Logan Central	22	30
Manly	23	28
Mount Gravatt	22	29
Oxley	22	31
Redcliffe	23	28

Links

Download from CRAN at https://cloud.r-project.org/package=datapasta

Browse source code at https://github.com/milesmcbain/datapasta/

Report a bug at https://github.com/milesmcbain/datapasta/issues

License

MIT + file LICENSE

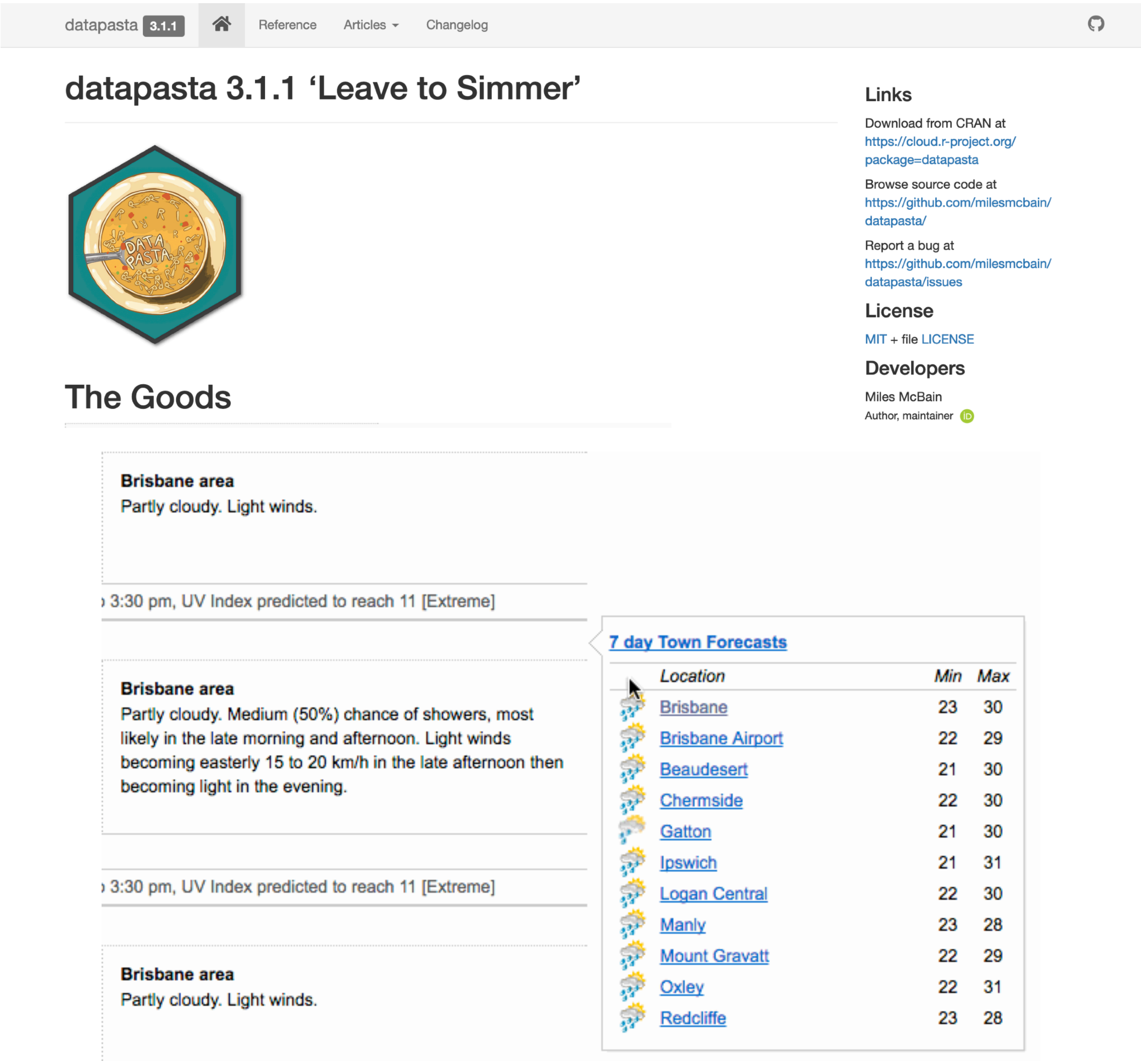
Developers

Miles McBain
Author, maintainer

{datapasta}

This works...

even more of the time



datapasta 3.1.1 ‘Leave to Simmer’

The Goods

Brisbane area
Partly cloudy. Light winds.

› 3:30 pm, UV Index predicted to reach 11 [Extreme]

Brisbane area
Partly cloudy. Medium (50%) chance of showers, most likely in the late morning and afternoon. Light winds becoming easterly 15 to 20 km/h in the late afternoon then becoming light in the evening.

› 3:30 pm, UV Index predicted to reach 11 [Extreme]

Brisbane area
Partly cloudy. Light winds.

[7 day Town Forecasts]

Location	*Min*	*Max*
Brisbane	23	30
Brisbane Airport	22	29
Beaudesert	21	30
Chermside	22	30
Gatton	21	30
Ipswich	21	31
Logan Central	22	30
Manly	23	28
Mount Gravatt	22	29
Oxley	22	31
Redcliffe	23	28

Links

Download from CRAN at https://cloud.r-project.org/package=datapasta

Browse source code at https://github.com/milesmcbain/datapasta/

Report a bug at https://github.com/milesmcbain/datapasta/issues

License

MIT + file LICENSE

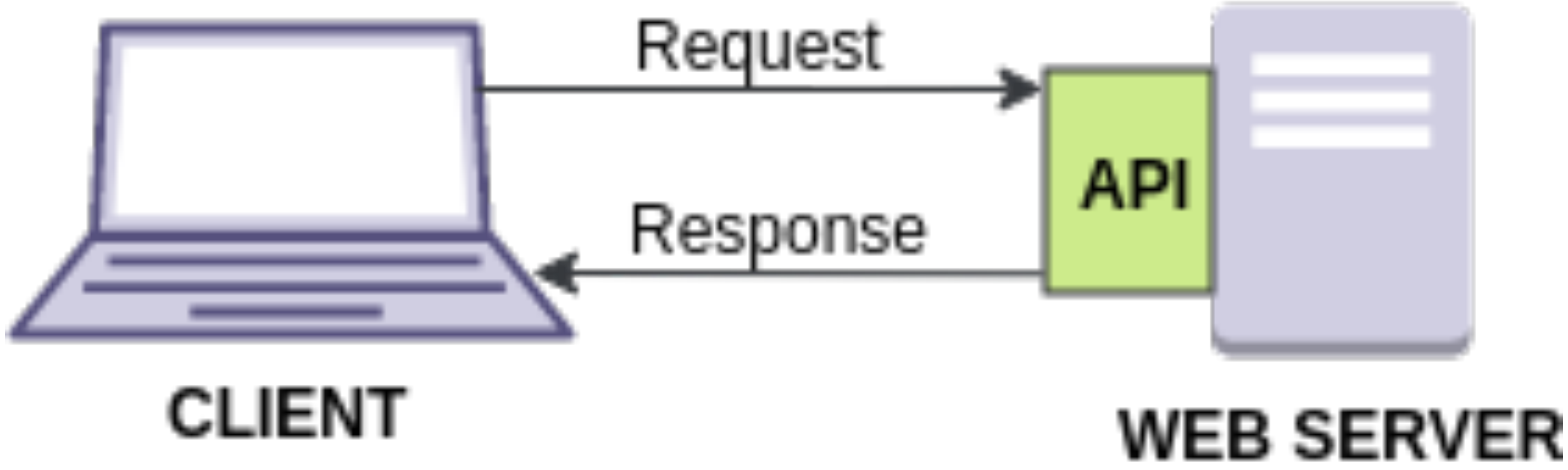
Developers

Miles McBain
Author, maintainer

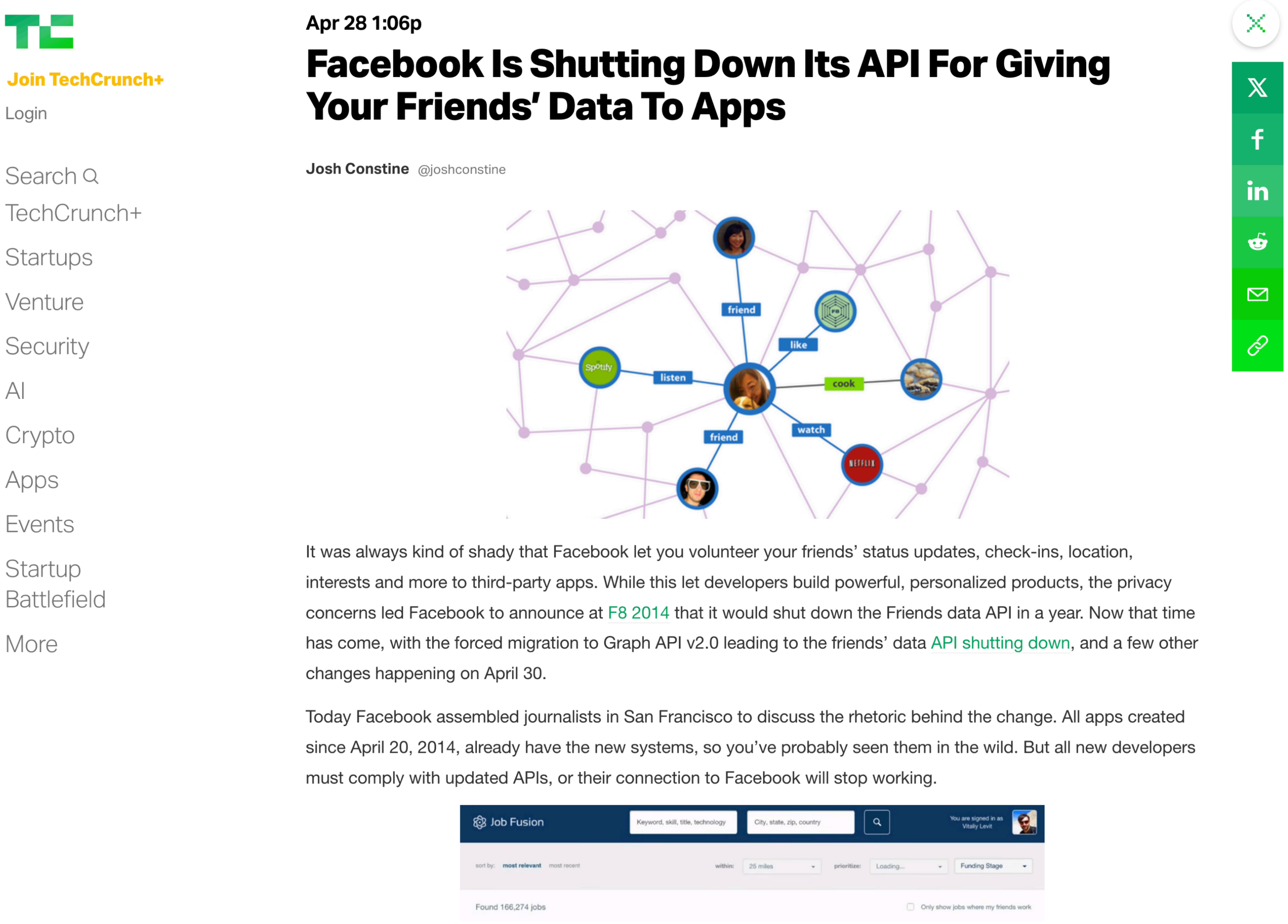
APIs

Application Programming Interfaces

- APIs are things that let computers talk to each other
- Frequently used to serve data on the web
- Sometimes require login/authentication



Brivadeneira, CC BY-SA 4.0, via Wikimedia Commons



Join TechCrunch+

Login

Search

TechCrunch+

Startups

Venture

Security

AI

Crypto

Apps

Events

Startup Battlefield

More

Apr 28 1:06p

Facebook Is Shutting Down Its API For Giving Your Friends' Data To Apps

Josh Constine @joshconstine

It was always kind of shady that Facebook let you volunteer your friends' status updates, check-ins, location, interests and more to third-party apps. While this let developers build powerful, personalized products, the privacy concerns led Facebook to announce at F8 2014 that it would shut down the Friends data API in a year. Now that time has come, with the forced migration to Graph API v2.0 leading to the friends' data API shutting down, and a few other changes happening on April 30.

Today Facebook assembled journalists in San Francisco to discuss the rhetoric behind the change. All apps created since April 20, 2014, already have the new systems, so you've probably seen them in the wild. But all new developers must comply with updated APIs, or their connection to Facebook will stop working.

Job Fusion

Keyword, skill, title, technology

City, state, zip, country

You are signed in as Vitaly Levit

sort by: most relevant most recent

within: 25 miles

prioritize: Loading...

Funding Stage

Found 166,274 jobs

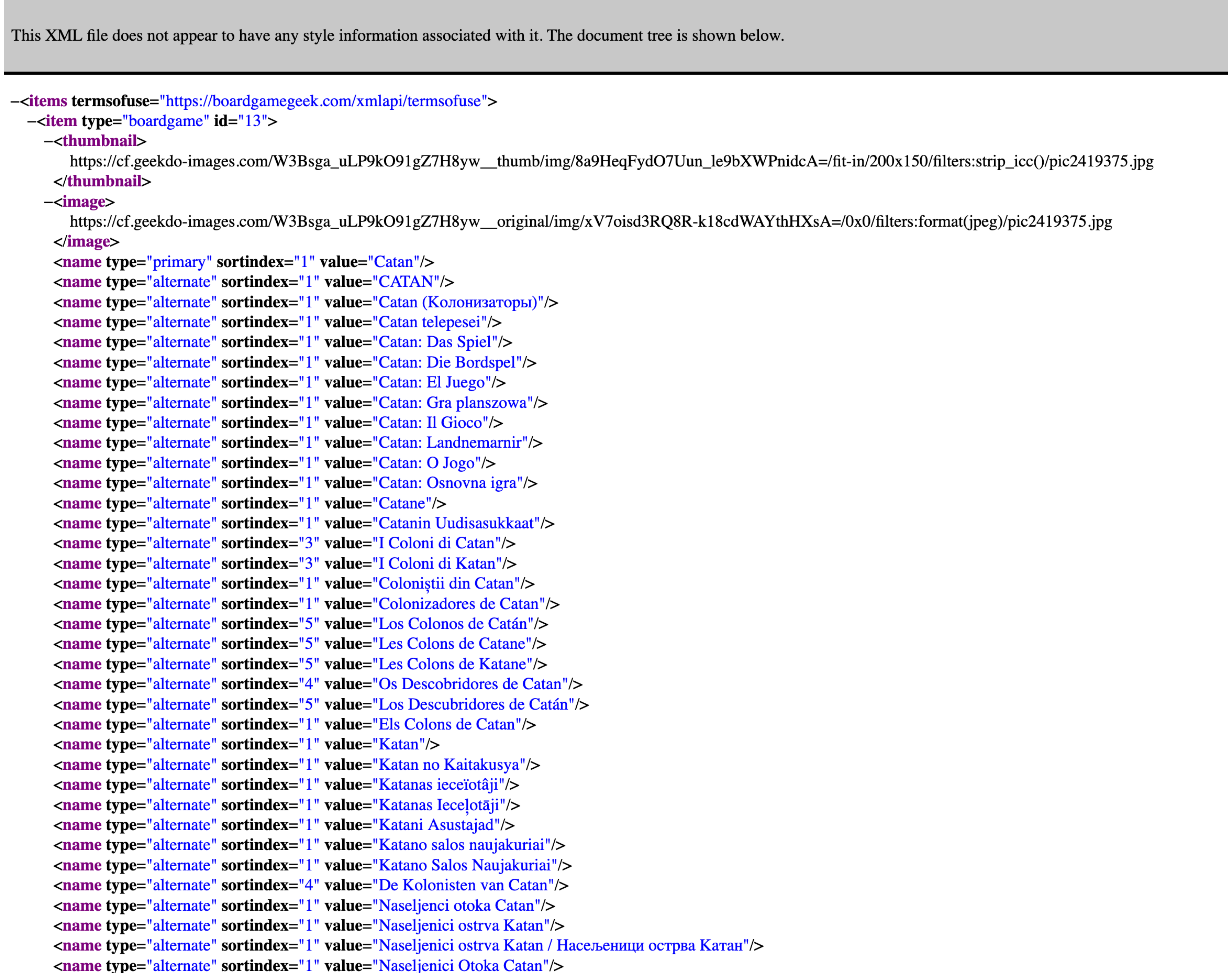
□ Only show jobs where my friends work

https://techcrunch.com/2015/04/28/facebook-api-shut-down/

Some cool APIs

- An API of Ice and Fire
- The Star Wars API
- The Rick and Morty API
- Native land API
- COVIDCast API
- BoardGame Geek API

https://boardgamegeek.com/xmlapi2/thing?id=13&type=boardgame&comments=1



This XML file does not appear to have any style information associated with it. The document tree is shown below.

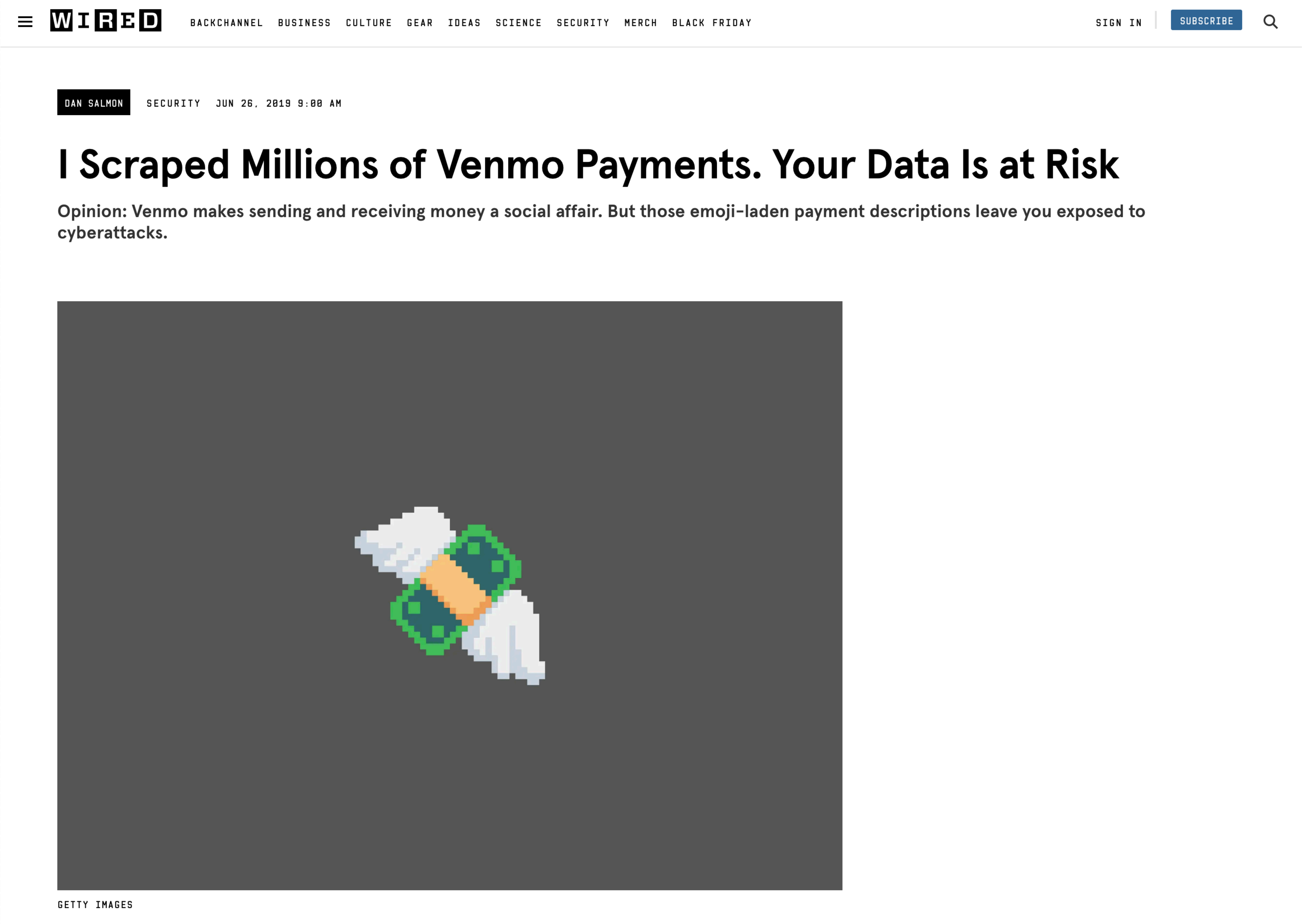
```
–<items termsofuse="https://boardgamegeek.com/xmlapi/termsofuse">
  –<item type="boardgame" id="13">
    –<thumbnail>
        https://cf.geekdo-images.com/W3Bsga_uLP9kO91gZ7H8yw__thumb/img/8a9HeqFydO7Uun_le9bXWPnidcA=/fit-in/200x150/filters:strip_icc()/pic2419375.jpg
     </thumbnail>
    –<image>
        https://cf.geekdo-images.com/W3Bsga_uLP9kO91gZ7H8yw__original/img/xV7oisd3RQ8R-k18cdWAYthHXsA=/0x0/filters:format(jpeg)/pic2419375.jpg
     </image>
     <name type="primary" sortindex="1" value="Catan"/>
     <name type="alternate" sortindex="1" value="CATAN"/>
     <name type="alternate" sortindex="1" value="Catan (Колонизаторы)"/>
     <name type="alternate" sortindex="1" value="Catan telepesei"/>
     <name type="alternate" sortindex="1" value="Catan: Das Spiel"/>
     <name type="alternate" sortindex="1" value="Catan: Die Bordspel"/>
     <name type="alternate" sortindex="1" value="Catan: El Juego"/>
     <name type="alternate" sortindex="1" value="Catan: Gra planszowa"/>
     <name type="alternate" sortindex="1" value="Catan: Il Gioco"/>
     <name type="alternate" sortindex="1" value="Catan: Landnemarnir"/>
     <name type="alternate" sortindex="1" value="Catan: O Jogo"/>
     <name type="alternate" sortindex="1" value="Catan: Osnovna igra"/>
     <name type="alternate" sortindex="1" value="Catane"/>
     <name type="alternate" sortindex="1" value="Catanin Uudisasukkaat"/>
     <name type="alternate" sortindex="3" value="I Coloni di Catan"/>
     <name type="alternate" sortindex="3" value="I Coloni di Katan"/>
     <name type="alternate" sortindex="1" value="Coloniştii din Catan"/>
     <name type="alternate" sortindex="1" value="Colonizadores de Catan"/>
     <name type="alternate" sortindex="5" value="Los Colonos de Catán"/>
     <name type="alternate" sortindex="5" value="Les Colons de Catane"/>
     <name type="alternate" sortindex="5" value="Les Colons de Katane"/>
     <name type="alternate" sortindex="4" value="Os Descobridores de Catan"/>
     <name type="alternate" sortindex="5" value="Los Descubridores de Catán"/>
     <name type="alternate" sortindex="1" value="Els Colons de Catan"/>
     <name type="alternate" sortindex="1" value="Katan"/>
     <name type="alternate" sortindex="1" value="Katan no Kaitakusya"/>
     <name type="alternate" sortindex="1" value="Katanas ieceïotâji"/>
     <name type="alternate" sortindex="1" value="Katanas Ieceļotāji"/>
     <name type="alternate" sortindex="1" value="Katani Asustajad"/>
     <name type="alternate" sortindex="1" value="Katano salos naujakuriai"/>
     <name type="alternate" sortindex="1" value="Katano Salos Naujakuriai"/>
     <name type="alternate" sortindex="4" value="De Kolonisten van Catan"/>
     <name type="alternate" sortindex="1" value="Naseljenci otoka Catan"/>
     <name type="alternate" sortindex="1" value="Naseljenici ostrva Katan"/>
     <name type="alternate" sortindex="1" value="Naseljenici ostrva Katan / Насељеници острва Катан"/>
     <name type="alternate" sortindex="1" value="Naseljenici Otoka Catan"/>
```

Scraping

Scraping allows you to access electronic data that does not have an API

Sites I've scraped include:

- IMDB
- GitHub (Counting Commits and Peer Code Review)
- Facebook (Deleting facebook)
- Wikipedia
- Pro football reference



DAN SALMON SECURITY JUN 26, 2019 9:00 AM

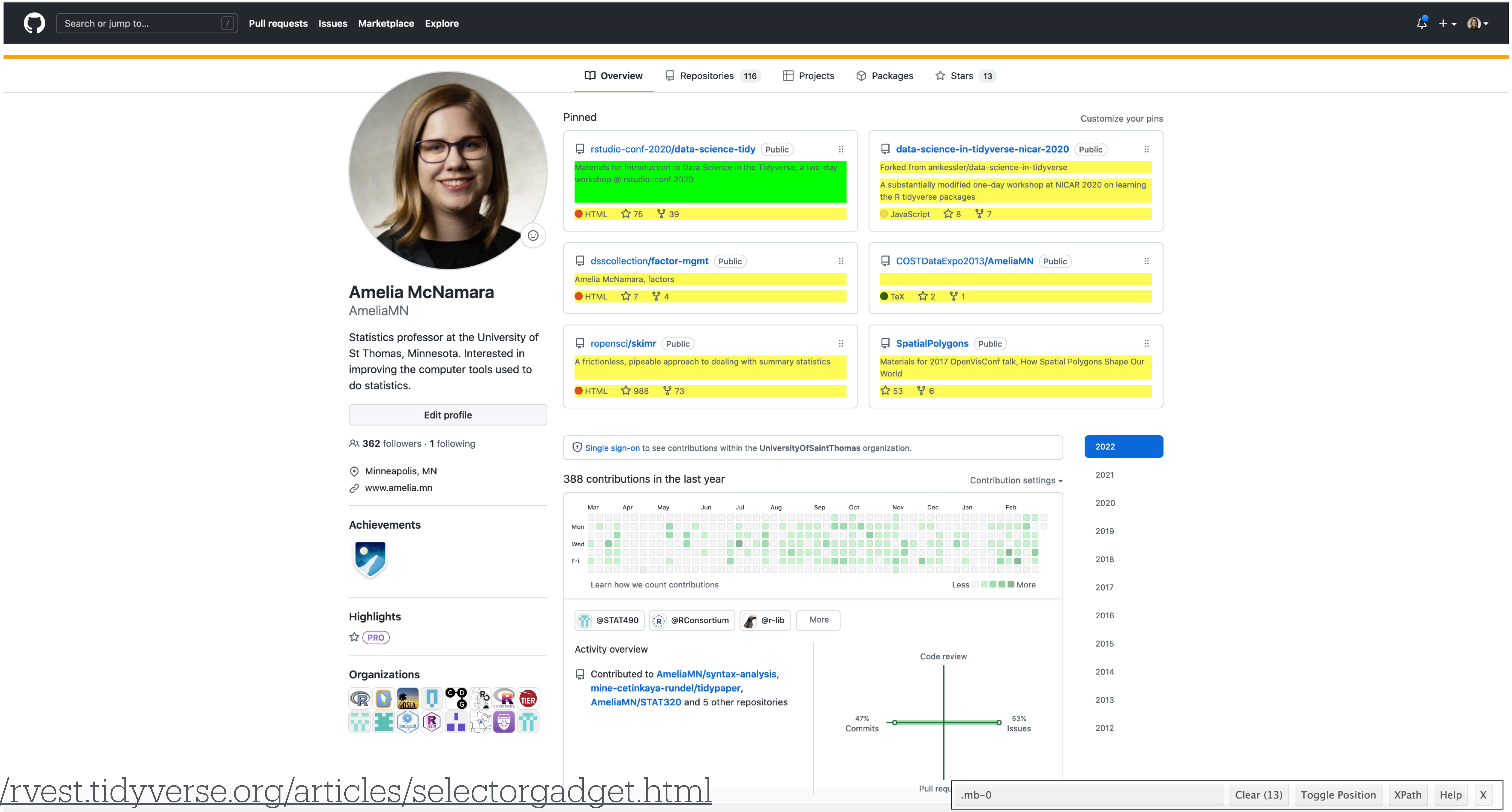
I Scraped Millions of Venmo Payments. Your Data Is at Risk

Opinion: Venmo makes sending and receiving money a social affair. But those emoji-laden payment descriptions leave you exposed to cyberattacks.

GETTY IMAGES

https://www.wired.com/story/i-scraped-millions-of-venmo-payments-your-data-is-at-risk/

Scraping: start with SelectorGadget



Search or jump to... / Pull requests Issues Marketplace Explore

Overview | Repositories 116 | Projects | Packages | Stars 13

Amelia McNamara
AmeliaMN

Statistics professor at the University of St Thomas, Minnesota. Interested in improving the computer tools used to do statistics.

Edit profile

362 followers · 1 following

Minneapolis, MN
www.amelia.mn

Achievements

Highlights
PRO

Organizations

Pinned | Customize your pins

rstudio-conf-2020/data-science-tidy Public
Materials for Introduction to Data Science in the Tidyverse, a two-day workshop @ rstudio::conf 2020
HTML ☆ 75 39

data-science-in-tidyverse-nicar-2020 Public
Forked from amkessler/data-science-in-tidyverse
A substantially modified one-day workshop at NICAR 2020 on learning the R tidyverse packages
JavaScript ☆ 8 7

dsscollection/factor-mgmt Public
Amelia McNamara, factors
HTML ☆ 7 4

COSTDataExpo2013/AmeliaMN Public
TeX ☆ 2 1

ropensci/skimr Public
A frictionless, pipeable approach to dealing with summary statistics
HTML ☆ 988 73

SpatialPolygons Public
Materials for 2017 OpenVisConf talk, How Spatial Polygons Shape Our World
☆ 53 6

Single sign-on to see contributions within the UniversityOfSaintThomas organization.

388 contributions in the last year | Contribution settings

Mar Apr May Jun Jul Aug Sep Oct Nov Dec Jan Feb
Mon
Wed
Fri
Learn how we count contributions
Less More

2022
2021
2020
2019
2018
2017
2016
2015
2014
2013
2012

@STAT490 @RConsortium @r-lib More

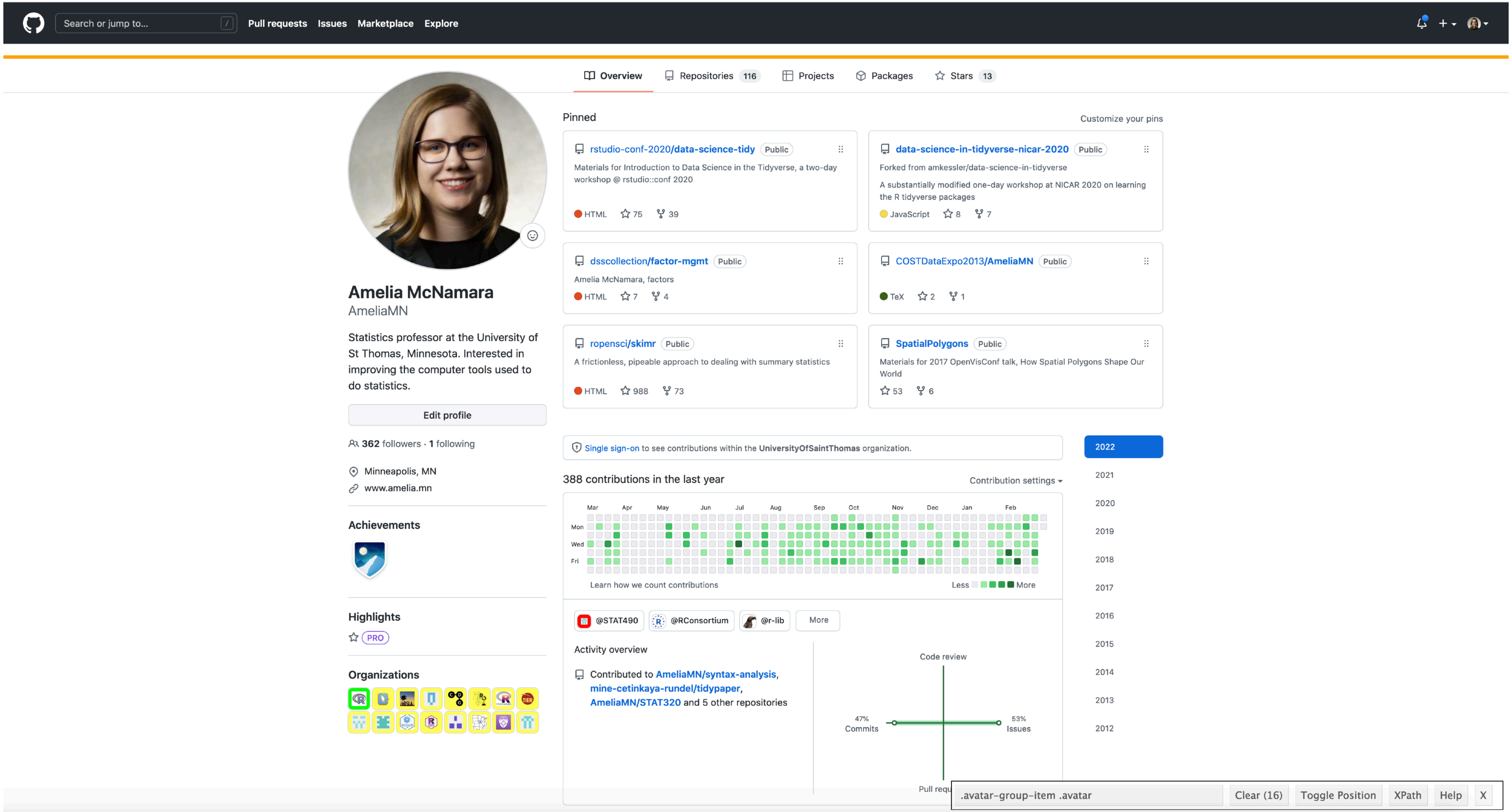
Activity overview

Contributed to AmeliaMN/syntax-analysis, mine-cetinkaya-rundel/tidypaper, AmeliaMN/STAT320 and 5 other repositories

.mb-0 | Clear (13) | Toggle Position | XPath | Help | X

https://rvest.tidyverse.org/articles/selectorgadget.html

Scraping



Search or jump to... / Pull requests Issues Marketplace Explore

Overview | Repositories 116 | Projects | Packages | Stars 13

Amelia McNamara
AmeliaMN

Statistics professor at the University of St Thomas, Minnesota. Interested in improving the computer tools used to do statistics.

Edit profile

362 followers · 1 following

Minneapolis, MN
www.amelia.mn

Achievements

Highlights

☆ PRO

Organizations

Pinned — Customize your pins

rstudio-conf-2020/data-science-tidy Public
Materials for Introduction to Data Science in the Tidyverse, a two-day workshop @ rstudio::conf 2020
● HTML ☆ 75 ⑂ 39

data-science-in-tidyverse-nicar-2020 Public
Forked from amkessler/data-science-in-tidyverse
A substantially modified one-day workshop at NICAR 2020 on learning the R tidyverse packages
● JavaScript ☆ 8 ⑂ 7

dsscollection/factor-mgmt Public
Amelia McNamara, factors
● HTML ☆ 7 ⑂ 4

COSTDataExpo2013/AmeliaMN Public
● TeX ☆ 2 ⑂ 1

ropensci/skimr Public
A frictionless, pipeable approach to dealing with summary statistics
● HTML ☆ 988 ⑂ 73

SpatialPolygons Public
Materials for 2017 OpenVisConf talk, How Spatial Polygons Shape Our World
☆ 53 ⑂ 6

Single sign-on to see contributions within the **UniversityOfSaintThomas** organization.

388 contributions in the last year — Contribution settings ▾

Learn how we count contributions — Less ▢🟩🟩🟩 More

2022 2021 2020 2019 2018 2017 2016 2015 2014 2013 2012

@STAT490 @RConsortium @r-lib More

Activity overview

Contributed to AmeliaMN/syntax-analysis, mine-cetinkaya-rundel/tidypaper, AmeliaMN/STAT320 and 5 other repositories

.avatar-group-item .avatar | Clear (16) | Toggle Position | XPath | Help | X

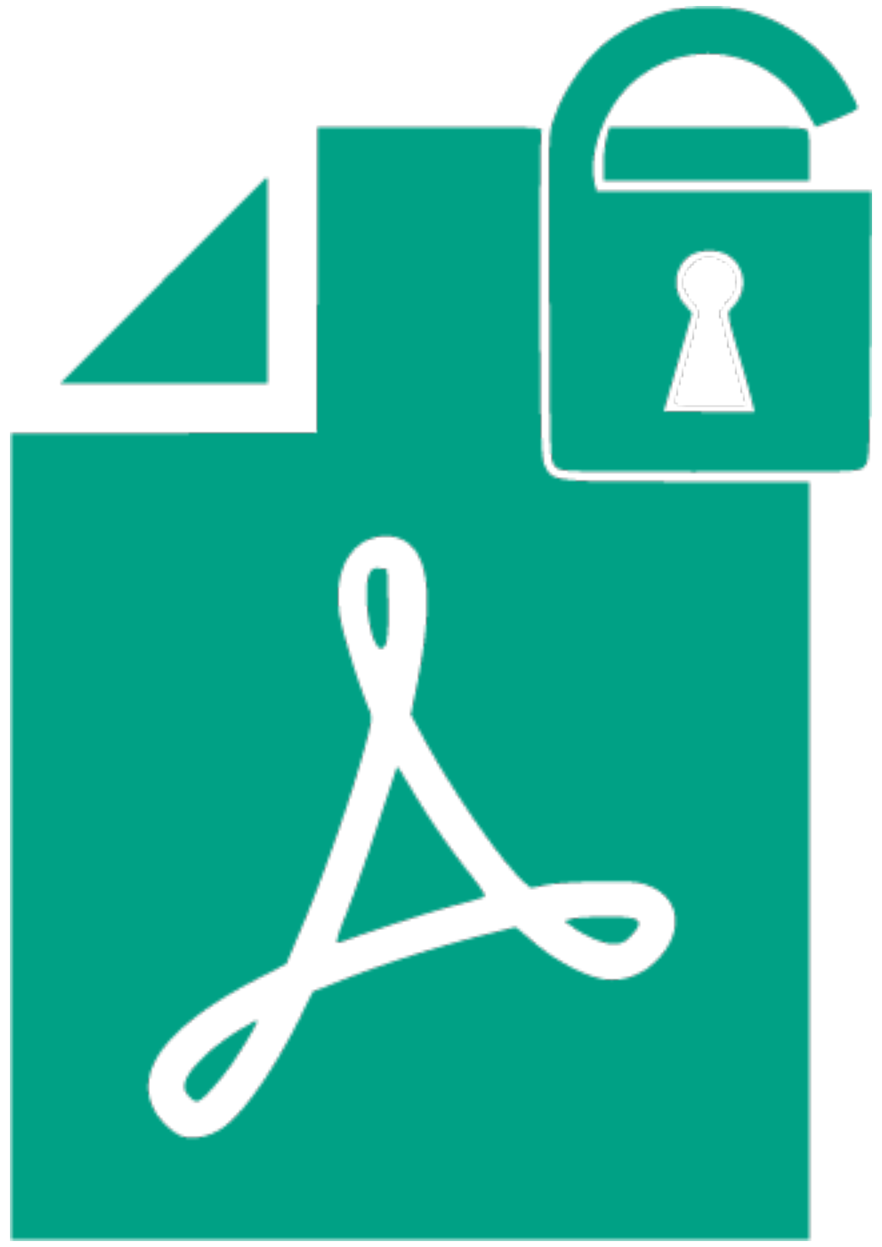
Hardest

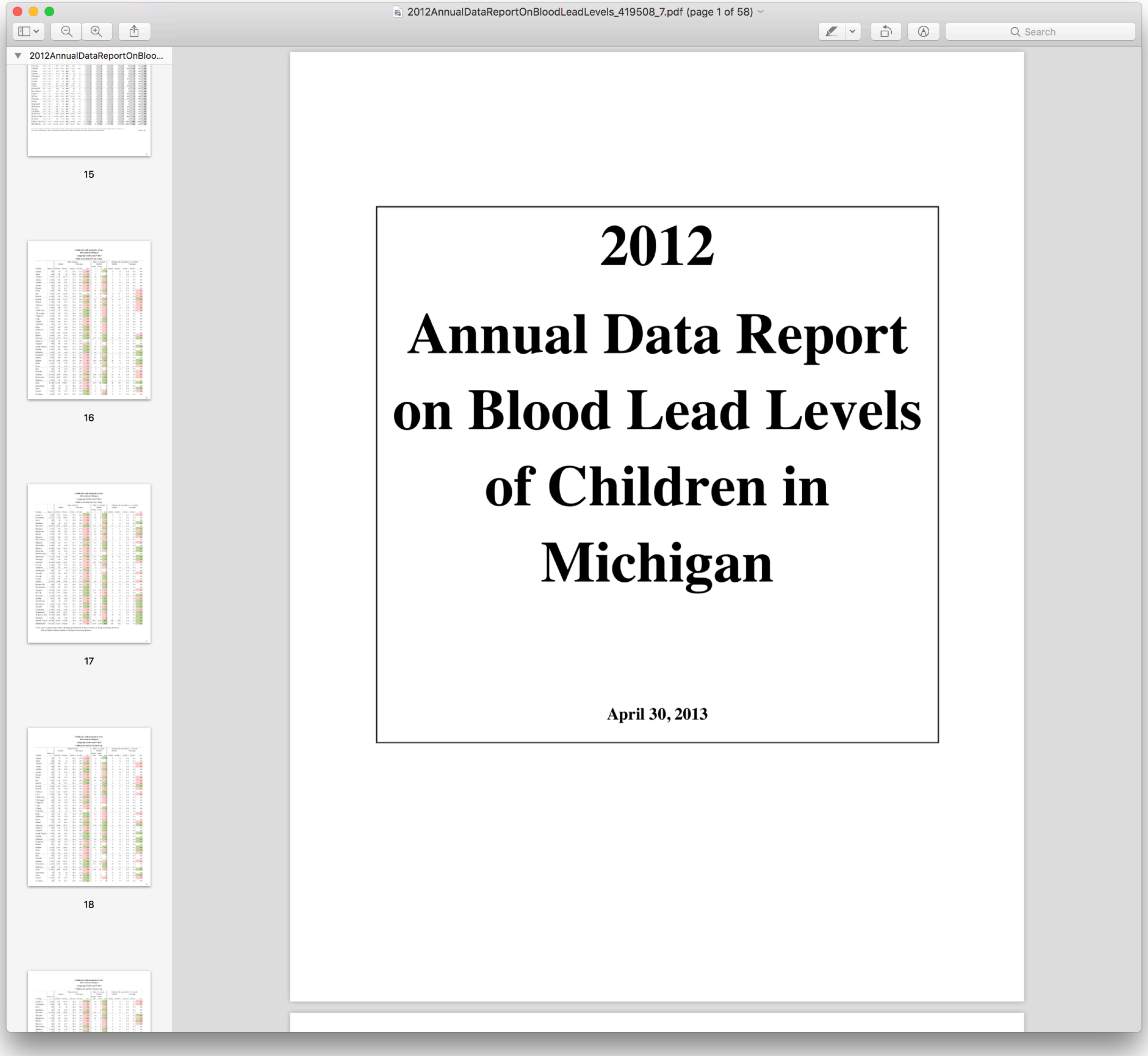
The data is not available electronically, or is locked in a bad format

- Data in PDFs
- Paper records
- Data in images, like JPEG or PNG

The easiest of these hard situations— data in PDFs

You can extract data from PDFS using Tabula





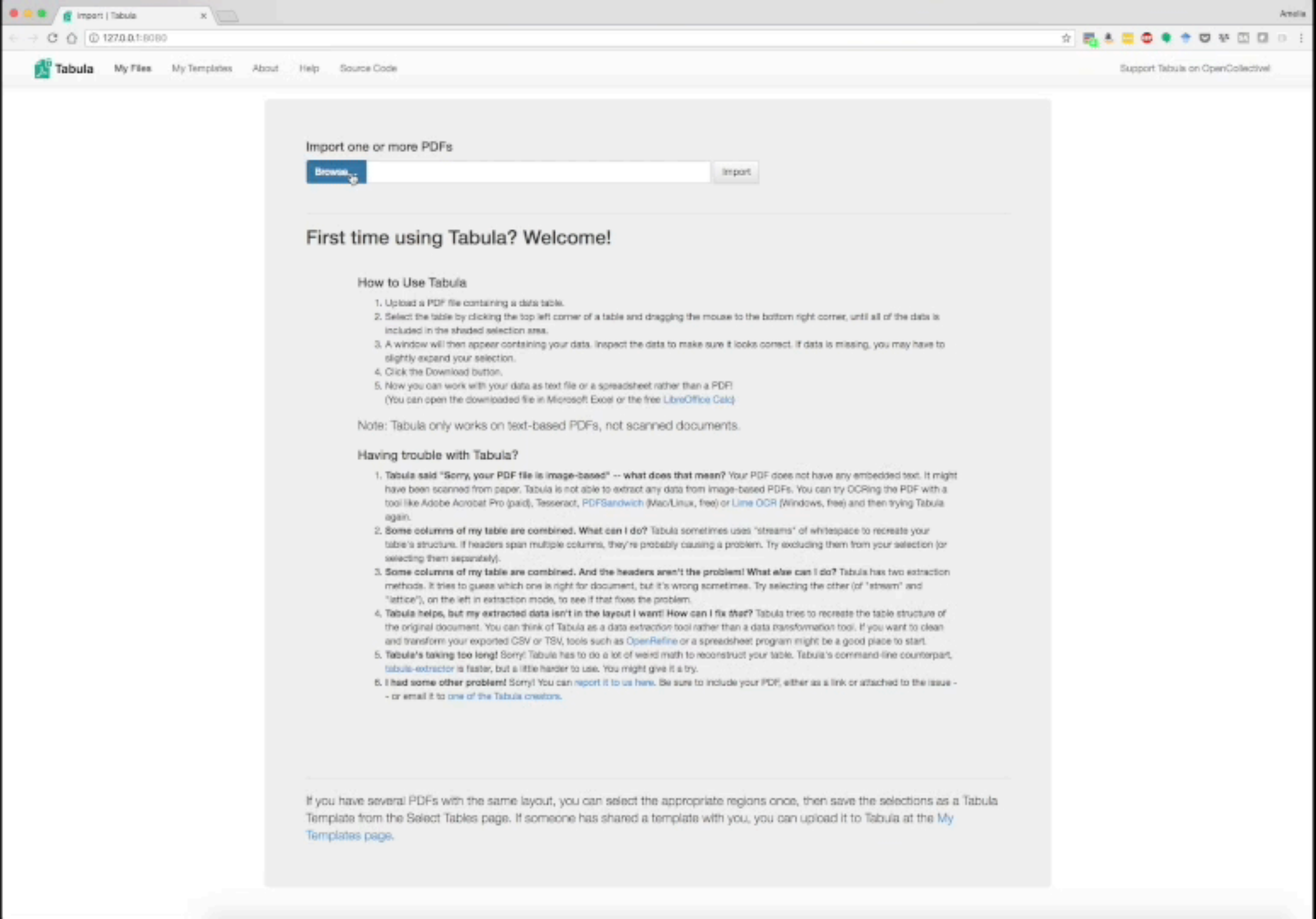
2012AnnualDataReportOnBloodLeadLevels_419508_7.pdf (page 1 of 58)
Search
2012AnnualDataReportOnBloo...
15
16
17
18

2012

Annual Data Report on Blood Lead Levels of Children in Michigan

April 30, 2013

Joint work with students from my Spring 2018 course SDS 236: Data Journalism, Smith College



Import one or more PDFs

First time using Tabula? Welcome!

How to Use Tabula

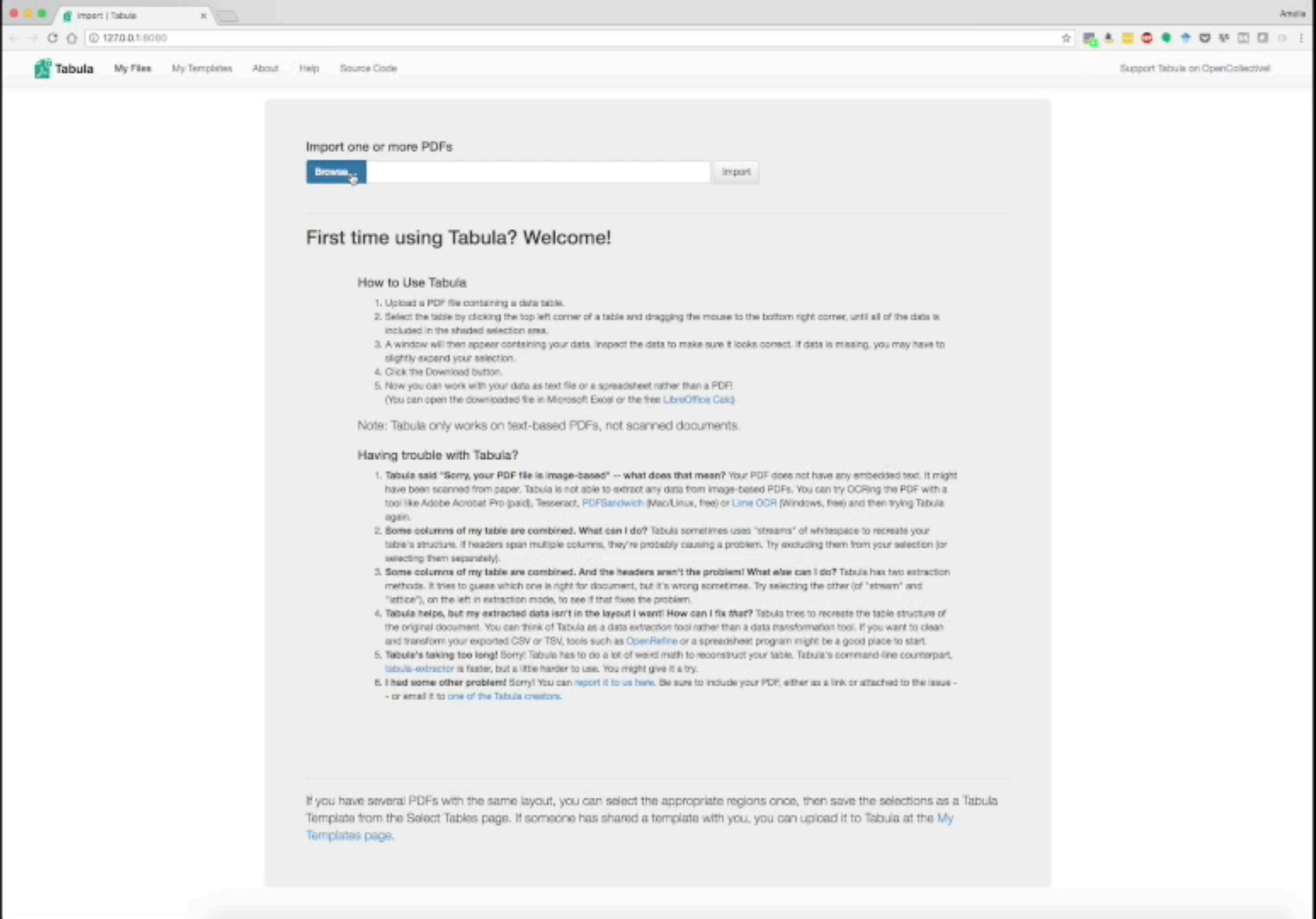
1. Upload a PDF file containing a data table.
2. Select the table by clicking the top left corner of a table and dragging the mouse to the bottom right corner, until all of the data is included in the shaded selection area.
3. A window will then appear containing your data. Inspect the data to make sure it looks correct. If data is missing, you may have to slightly expand your selection.
4. Click the Download button.
5. Now you can work with your data as text file or a spreadsheet rather than a PDF!
   (You can open the downloaded file in Microsoft Excel or the free LibreOffice Calc)

Note: Tabula only works on text-based PDFs, not scanned documents.

Having trouble with Tabula?

1. **Tabula said "Sorry, your PDF file is image-based" -- what does that mean?** Your PDF does not have any embedded text. It might have been scanned from paper. Tabula is not able to extract any data from image-based PDFs. You can try OCRing the PDF with a tool like Adobe Acrobat Pro (paid), Tesseract, PDFSandwich (Mac/Linux, free) or Lime OCR (Windows, free) and then trying Tabula again.
2. **Some columns of my table are combined. What can I do?** Tabula sometimes uses "streams" of whitespace to recreate your table's structure. If headers span multiple columns, they're probably causing a problem. Try excluding them from your selection (or selecting them separately).
3. **Some columns of my table are combined. And the headers aren't the problem! What else can I do?** Tabula has two extraction methods. It tries to guess which one is right for document, but it's wrong sometimes. Try selecting the other (of "stream" and "lattice"), on the left in extraction mode, to see if that fixes the problem.
4. **Tabula helps, but my extracted data isn't in the layout I want! How can I fix that?** Tabula tries to recreate the table structure of the original document. You can think of Tabula as a data extraction tool rather than a data transformation tool. If you want to clean and transform your exported CSV or TSV, tools such as OpenRefine or a spreadsheet program might be a good place to start.
5. **Tabula's taking too long!** Sorry! Tabula has to do a lot of weird math to reconstruct your table. Tabula's command-line counterpart, tabula-extractor is faster, but a little harder to use. You might give it a try.
6. **I had some other problem!** Sorry! You can report it to us here. Be sure to include your PDF, either as a link or attached to the issue -- or email it to one of the Tabula creators.

If you have several PDFs with the same layout, you can select the appropriate regions once, then save the selections as a Tabula Template from the Select Tables page. If someone has shared a template with you, you can upload it to Tabula at the My Templates page.



Import one or more PDFs

First time using Tabula? Welcome!

How to Use Tabula

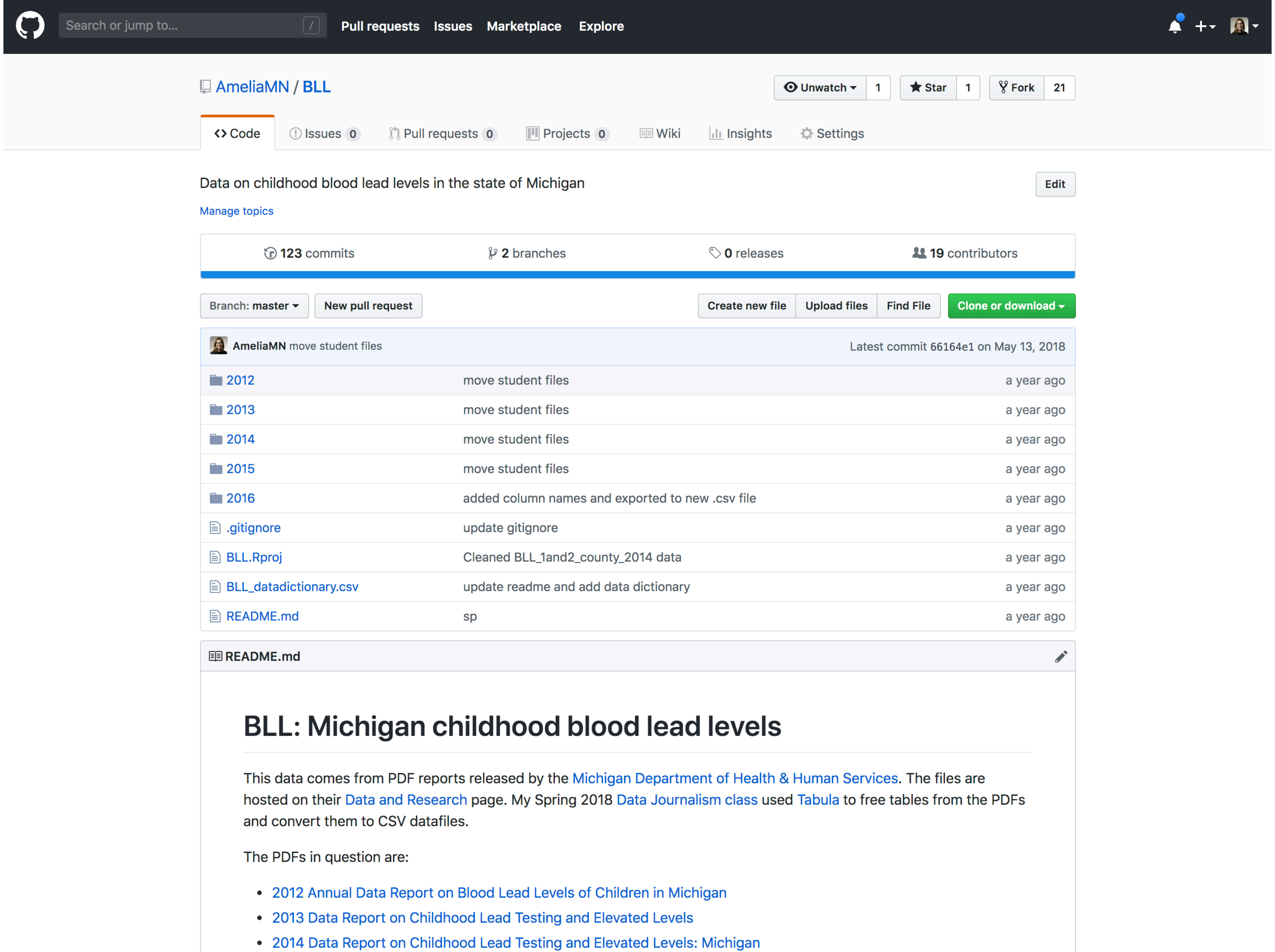
1. Upload a PDF file containing a data table.
2. Select the table by clicking the top left corner of a table and dragging the mouse to the bottom right corner, until all of the data is included in the shaded selection area.
3. A window will then appear containing your data. Inspect the data to make sure it looks correct. If data is missing, you may have to slightly expand your selection.
4. Click the Download button.
5. Now you can work with your data as text file or a spreadsheet rather than a PDF!
   (You can open the downloaded file in Microsoft Excel or the free LibreOffice Calc)

Note: Tabula only works on text-based PDFs, not scanned documents.

Having trouble with Tabula?

1. **Tabula said "Sorry, your PDF file is image-based" -- what does that mean?** Your PDF does not have any embedded text. It might have been scanned from paper. Tabula is not able to extract any data from image-based PDFs. You can try OCRing the PDF with a tool like Adobe Acrobat Pro (paid), Tesseract, PDFSandwich (Mac/Linux, free) or Lime OCR (Windows, free) and then trying Tabula again.
2. **Some columns of my table are combined. What can I do?** Tabula sometimes uses "streams" of whitespace to recreate your table's structure. If headers span multiple columns, they're probably causing a problem. Try excluding them from your selection (or selecting them separately).
3. **Some columns of my table are combined. And the headers aren't the problem! What else can I do?** Tabula has two extraction methods. It tries to guess which one is right for document, but it's wrong sometimes. Try selecting the other (of "stream" and "lattice"), on the left in extraction mode, to see if that fixes the problem.
4. **Tabula helps, but my extracted data isn't in the layout I want! How can I fix that?** Tabula tries to recreate the table structure of the original document. You can think of Tabula as a data extraction tool rather than a data transformation tool. If you want to clean and transform your exported CSV or TSV, tools such as OpenRefine or a spreadsheet program might be a good place to start.
5. **Tabula's taking too long!** Sorry! Tabula has to do a lot of weird math to reconstruct your table. Tabula's command-line counterpart, tabula-extractor is faster, but a little harder to use. You might give it a try.
6. **I had some other problem!** Sorry! You can report it to us here. Be sure to include your PDF, either as a link or attached to the issue -- or email it to one of the Tabula creators.

If you have several PDFs with the same layout, you can select the appropriate regions once, then save the selections as a Tabula Template from the Select Tables page. If someone has shared a template with you, you can upload it to Tabula at the My Templates page.



AmeliaMN / BLL

Data on childhood blood lead levels in the state of Michigan

123 commits · 2 branches · 0 releases · 19 contributors

AmeliaMN move student files — Latest commit 66164e1 on May 13, 2018

2012	move student files	a year ago
2013	move student files	a year ago
2014	move student files	a year ago
2015	move student files	a year ago
2016	added column names and exported to new .csv file	a year ago
.gitignore	update gitignore	a year ago
BLL.Rproj	Cleaned BLL_1and2_county_2014 data	a year ago
BLL_datadictionary.csv	update readme and add data dictionary	a year ago
README.md	sp	a year ago

README.md

BLL: Michigan childhood blood lead levels

This data comes from PDF reports released by the Michigan Department of Health & Human Services. The files are hosted on their Data and Research page. My Spring 2018 Data Journalism class used Tabula to free tables from the PDFs and convert them to CSV datafiles.

The PDFs in question are:

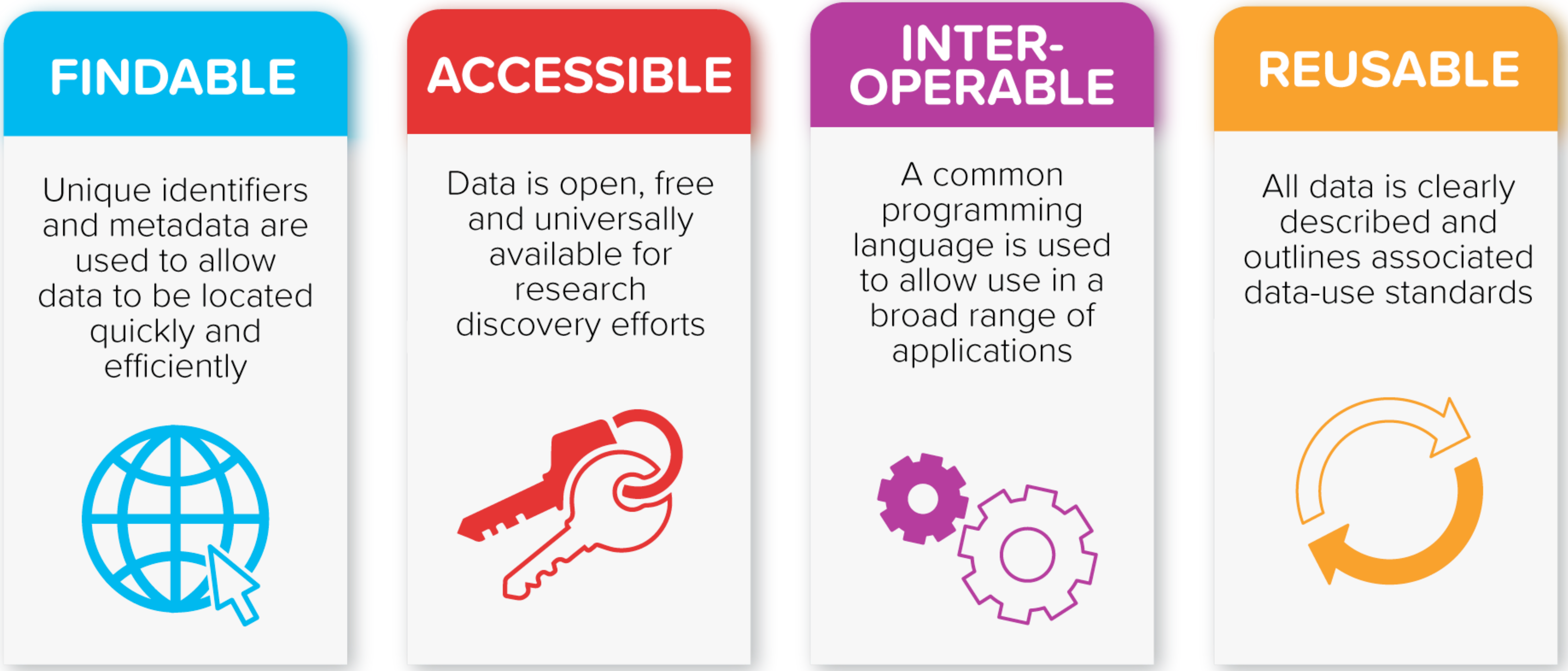
- 2012 Annual Data Report on Blood Lead Levels of Children in Michigan
- 2013 Data Report on Childhood Lead Testing and Elevated Levels
- 2014 Data Report on Childhood Lead Testing and Elevated Levels: Michigan

Joint work with students from my Spring 2018 course SDS 236: Data Journalism, Smith College

The hardest situation

Data is not available electronically, or in images

- Manual data entry (works best for small datasets)
- Optical Character Recognition (OCR)
    - Many tech companies have services
        - Microsoft has something built into OneNote
        - Google I think will do it with the Google Lens or something?
    - Of course there's a way to use R!



FINDABLE

Unique identifiers and metadata are used to allow data to be located quickly and efficiently

ACCESSIBLE

Data is open, free and universally available for research discovery efforts

INTER-OPERABLE

A common programming language is used to allow use in a broad range of applications

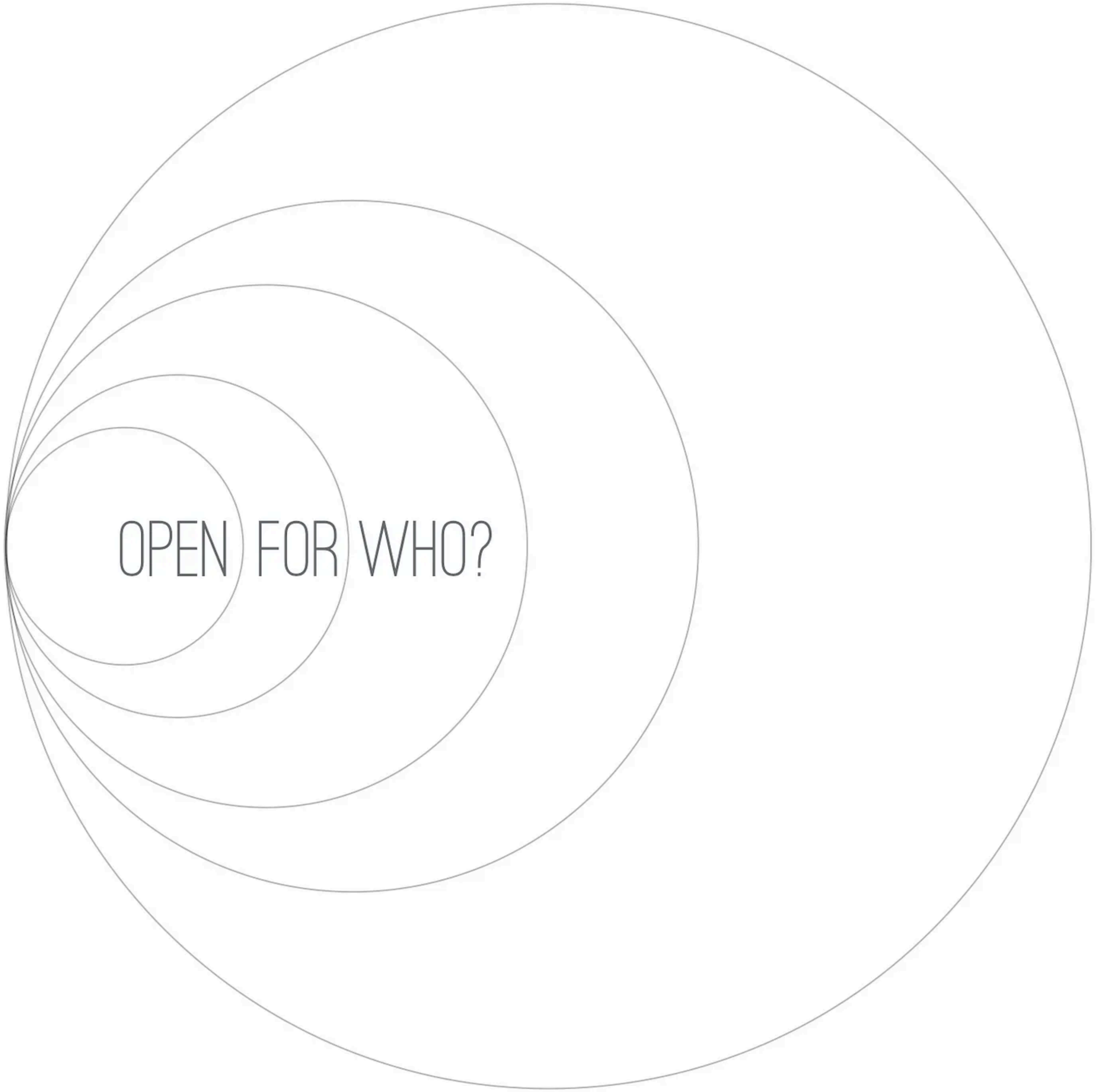
REUSABLE

All data is clearly described and outlines associated data-use standards

Source: https://kidsfirstdrc.org/about/drc_impact/
via Tulane libguide https://libguides.tulane.edu/datacuration/fair

> However, the act of sharing data implies the communication of something to a set of potentially unknown others. Moreover, the controversies surrounding the ethics of sharing data [...] and the methodological reasons for (not) doing so in the social sciences [...] as well as the natural sciences [...] indicate that how and what to communicate and to whom is more problematic than naïve accounts of scientific collaboration presume. Key recent studies in the field of e-science [...] have underlined how most of the obstacles to such data provision are less technological than social, ethical, legal, and institutional,

Samuel Carlson, Ben Anderson. What Are Data? The Many Kinds of Data and Their Implications for Data Re-Use (2007) https://doi.org/10.1111/j.1083-6101.2007.00342.x



Open for who?Jer Thorp

https://medium.com/memo-random/open-for-who-ce698a8de79c#.uxjqzre9b

Let’s try to find data in the easiest format(s) to work with