ROBERT GEBELOFF | @GEBELOFFNYT | DATAHARVEST 2021

HOW TO TURN ANYTHING INTO A DATABASE

HOW TO TURN ANYTHING INTO A DATABASE

WHO AM I?

- JOURNALIST WITH 30 YEARS OF PROFESSIONAL EXPERIENCE
- VERY TRADITIONAL U.S. REPORTING BACKGROUND
- ALWAYS INCORPORATED DATA IN MY WORK, SPECIALIZED IN DATA JOURNALISM STARTING IN 1997

HOW TO TURN ANYTHING INTO A DATABASE

MY APPROACH TO STORIES

DATA JOURNALISM IS NOT MAGIC, IT'S JUST ANOTHER FORM OF JOURNALISM

FOR ALL STORIES, WE NEED SOURCES:

Robert Gebeloff -- @gebeloffnyt- DATAHARVEST 2021

HOW TO TURN ANYTHING INTO A DATABASE

MY APPROACH TO STORIES

OBSERVATION – WHAT YOU SEE IS AS IMPORTANT AS ANYTHING… SET THE SCENE, JOT DOWN DETAILS, HEAR, SMELL, COUNT, ETC.

Robert Gebeloff -- @gebeloffnyt- DATAHARVEST 2021

HOW TO TURN ANYTHING INTO A DATABASE

MY APPROACH TO STORIES

PEOPLE – FINDING PEOPLE WHO KNOW THE INFO, GETTING THEM TO TELL YOU THE INFO YOU NEED

Robert Gebeloff -- @gebeloffnyt- DATAHARVEST 2021

HOW TO TURN ANYTHING INTO A DATABASE

MY APPROACH TO STORIES

DOCUMENTS – PAPER CAN BE BETTER THAN PEOPLE – FACTS IN BLACK AND WHITE – COURT RECORDS, AUDITS, LEAKS, ETC. YOUR EYES CAN DECEIVE YOU, PEOPLE CAN LIE TO YOU, BUT DOCS ARE DOCS...

Robert Gebeloff -- @gebeloffnyt- DATAHARVEST 2021

HOW TO TURN ANYTHING INTO A DATABASE

MY APPROACH TO STORIES

DATA – NUMERIC INFORMATION THAT ISN'T IN NARRATIVE FORM; RAW INFORMATION THAT HASN'T BEEN SUMMARIZED OR OTHERWISE PROCESSED – JOURNALISTS BECOME SOCIAL SCIENTISTS, FIND THE NEWS

Robert Gebeloff -- @gebeloffnyt- DATAHARVEST 2021

HOW TO TURN ANYTHING INTO A DATABASE

MY APPROACH TO STORIES

I RARELY START OUT WITH A DATASET AND SAY, "LET'S GO LOOKING FOR AN INVESTIGATIVE STORY!"

Robert Gebeloff -- @gebeloffnyt- DATAHARVEST 2021

HOW TO TURN ANYTHING INTO A DATABASE

MY APPROACH TO STORIES

HAVE "A DATA STATE OF MIND" -- @MARYJOWEBSTER.

Robert Gebeloff -- @gebeloffnyt- DATAHARVEST 2021

HOW TO TURN ANYTHING INTO A DATABASE

MY APPROACH TO STORIES

FOR WHATEVER TOPIC YOU'RE EXPLORING, ARE THERE THINGS YOU CAN BETTER QUANTIFY WITH DATA?

Robert Gebeloff -- @gebeloffnyt- DATAHARVEST 2021

HOW TO TURN ANYTHING INTO A DATABASE

MY APPROACH TO STORIES

FOR WHATEVER TOPIC YOU'RE EXPLORING, ARE THERE THINGS YOU CAN BETTER QUANTIFY WITH DATA?

WHAT DATA COULD YOU USE? FOIA A DATABASE? DOWNLOAD IT FROM THE WEB?

Robert Gebeloff -- @gebeloffnyt- DATAHARVEST 2021

HOW TO TURN ANYTHING INTO A DATABASE

MY APPROACH TO STORIES

FOR WHATEVER TOPIC YOU'RE EXPLORING, ARE THERE THINGS YOU CAN BETTER QUANTIFY WITH DATA?

WHAT DATA COULD YOU USE? FOIA A DATABASE? DOWNLOAD IT FROM THE WEB?

OR MAKE YOUR OWN OUT OF OTHER RECORDS…

HOW TO TURN ANYTHING INTO A DATABASE

MY APPROACH TO STORIES

ANOTHER USEFUL SLOGAN….

THINK IN ROWS AND COLUMNS

Robert Gebeloff -- @gebeloffnyt- DATAHARVEST 2021

HOW TO TURN ANYTHING INTO A DATABASE

MY APPROACH TO STORIES

ANOTHER USEFUL SLOGAN….

THINK IN ROWS AND COLUMNS

EACH ROW IS A **UNIQUE** ENTITY.

EACH COLUMN IS A **UNIQUE** CHARACTERISTIC OF THAT ENTITY.

Robert Gebeloff -- @gebeloffnyt- DATAHARVEST 2021

HOW TO TURN ANYTHING INTO A DATABASE

	A	B	C	D	E	F	G	H	I	J	K	L	M
1		**Jan**	**Feb**	**Mar**	**Apr**	**May**	**Jun**	**Jul**	**Aug**	**Sep**	**Oct**	**Nov**	**Dec**
2	**Item 1**	59	36	86	24	28	21	59	66	70	58	22	97
3	**Item 2**	80	8	10	87	78	93	24	51	37	86	58	10
4	**Item 3**	41	6	74	31	49	54	62	9	45	66	9	61
5	**Item 4**	89	56	45	8	55	86	25	98	6	23	1	10
6	**Item 5**	15	42	12	46	73	93	79	90	16	95	39	4
7	**Item 6**	13	89	57	66	11	17	14	10	68	17	25	60
8	**Item 7**	30	18	67	53	30	84	42	29	52	21	60	19
9	**Item 8**	7	56	95	64	56	50	84	9	38	7	98	23
10	**Item 9**	79	13	19	22	62	29	14	60	59	54	90	76
11	**Item 10**	75	55	13	6	30	15	86	2	71	65	82	98
12	**Item 11**	7	9	77	69	97	27	90	28	32	71	22	68

HOW TO TURN ANYTHING INTO A DATABASE

	A	B	C	D	E	F	G	H	I	J	K	L	M
1		Jan	Feb	Mar	Apr	May	Jun	Jul	Aug	Sep	Oct	Nov	Dec
2	Item 1	59	36	86	24	28	21	59	66	70	58	22	97
3	Item 2	80	8	10	87	78	93	24	51	37	86	58	10
4	Item 3	41	6	74	31	49	54	62	9	45	66	9	61
5	Item 4	89	56	45	8	55	86	25	98	6	23	1	10
6	Item 5	15	42	12	46	73	93	79	90	16	95	39	4
7	Item 6	13	89	57	66	11	17	14	10	68	17	25	60
8	Item 7	30	18	67	53	30	84	42	29	52	21	60	19
9	Item 8	7	56	95	64	56	50	84	9	38	7	98	23
10	Item 9	79	13	19	22	62	29	14	60	59	54	90	76
11	Item 10	75	55	13	6	30	15	86	2	71	65	82	98
12	Item 11	7	9	77	69	97	27	90	28	32	71	22	68

MY APPROACH TO STORIES

HAVING THIS STRUCTURE ALLOWS YOU TO ANALYZE:

-- WHAT HAPPENED IN MAY AMONG ALL ITEMS

-- WHAT HAPPENED TO ITEM 6 ACROSS ALL MONTHS?

-- COMPARE ALL MONTHS

-- COMPARE ALL ITEMS

HOW TO TURN ANYTHING INTO A DATABASE

MY APPROACH TO STORIES

WHEN YOU FOIA OR DOWNLOAD A DATABASE, THIS IS WHAT YOU GET: A COLLECTION OF ROWS AND COLUMNS, WITH DOCUMENTATION THAT TELLS YOU WHAT EACH ROW AND COLUMN REPRESENTS.

Robert Gebeloff -- @gebeloffnyt- DATAHARVEST 2021

HOW TO TURN ANYTHING INTO A DATABASE

MY APPROACH TO STORIES

WHEN YOU BUILD A DATABASE, **YOU'RE THE ARCHITECT**.

YOU DECIDE WHAT EACH ROW IS GOING TO REPRESENT.

AND THEN YOU DECIDE WHAT CHARACTERISTICS YOU'RE GOING TO CAPTURE IN COLUMNS.

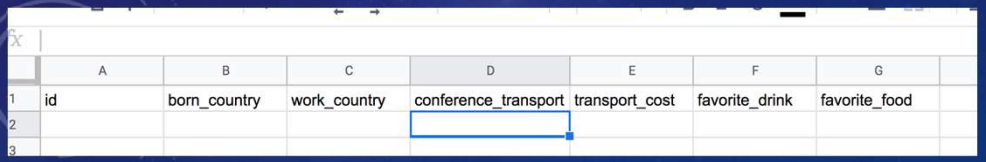
HOW TO TURN ANYTHING INTO A DATABASE

DATA JOURNALISM 101

WHEN I TEACH A BEGINNER'S COURSE IN DATA JOURNALISM, THE FIRST EXERCISE IN THE FIRST CLASS IS….

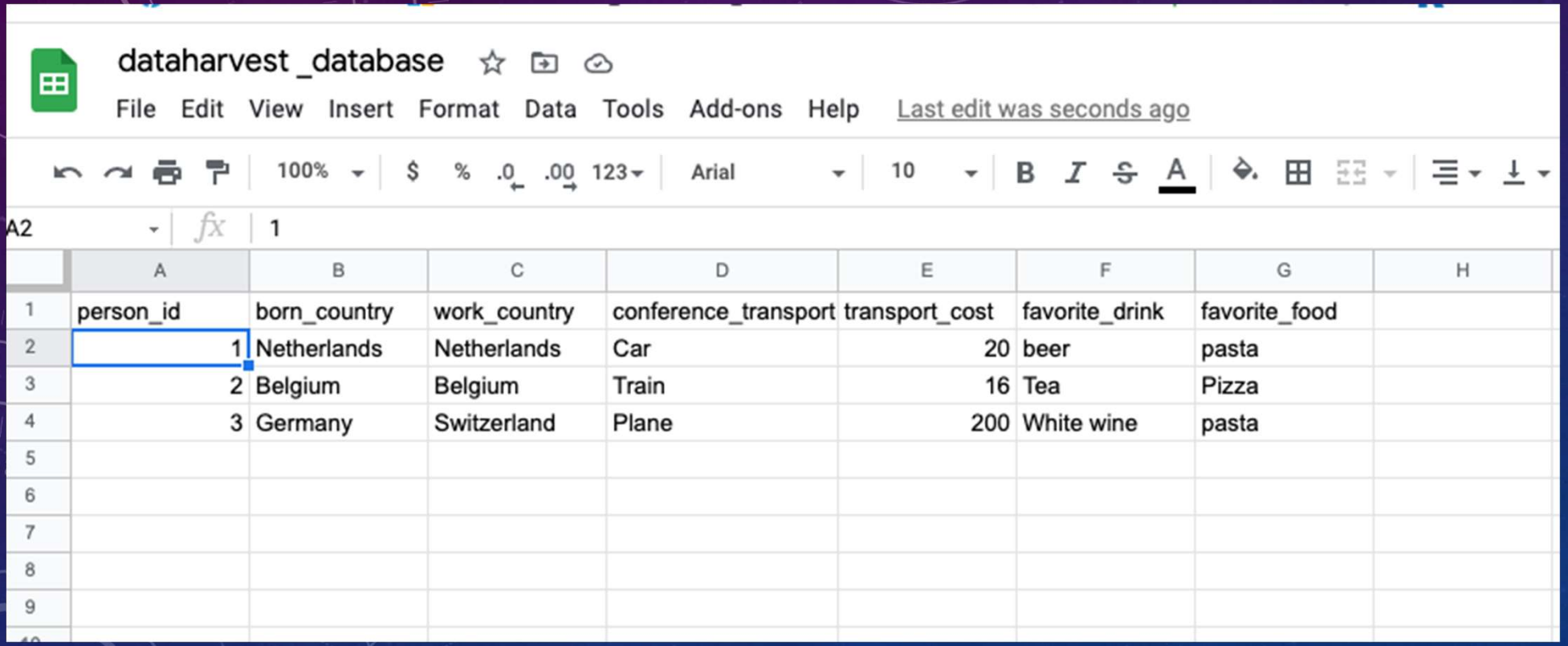
BUILDING A DATABASE.

LET'S TRY THIS OURSELVES… IMAGINING WE ARE STUDYING THE ATTENDEES OF THE DATAHARVEST CONFERENCE.



	A	B	C	D	E	F	G
1	id	born_country	work_country	conference_transport	transport_cost	favorite_drink	favorite_food
2							
3							

HOW TO TURN ANYTHING INTO A DATABASE



dataharvest _database

File Edit View Insert Format Data Tools Add-ons Help Last edit was seconds ago

100% $ % .0 .00 123 Arial 10

A2 | 1

	A	B	C	D	E	F	G	H
1	person_id	born_country	work_country	conference_transport	transport_cost	favorite_drink	favorite_food	
2	1	Netherlands	Netherlands	Car	20	beer	pasta	
3	2	Belgium	Belgium	Train	16	Tea	Pizza	
4	3	Germany	Switzerland	Plane	200	White wine	pasta	
5								
6								
7								
8								
9								

HOW TO TURN ANYTHING INTO A DATABASE

DATA JOURNALISM 101

LET'S TAKE A CLOSER LOOK AT THE COLUMNS WE CREATED AND WHAT WE MIGHT BE ABLE TO DO WITH THEM.

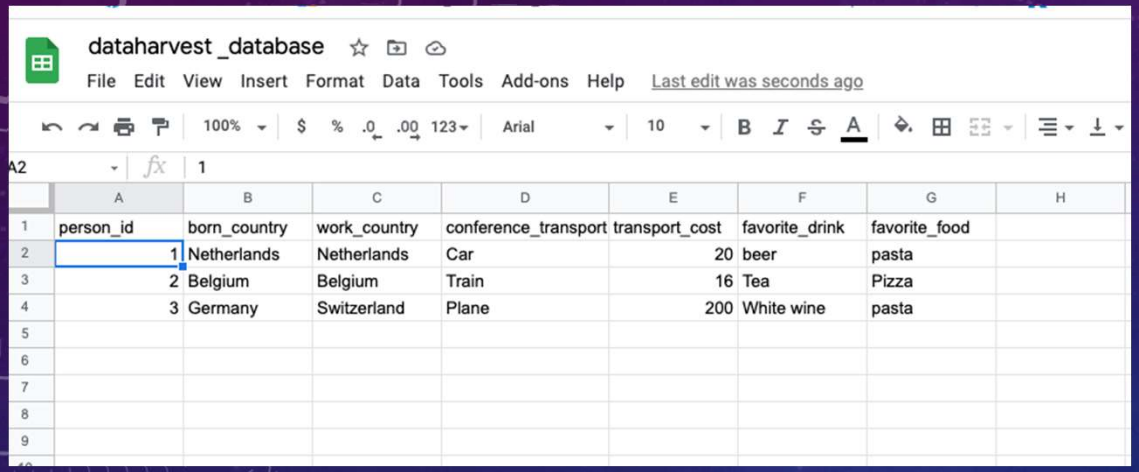
WE'LL ALSO PAY ATTENTION TO WHETHER WE'RE CREATING A VARIABLE THAT IS:

BINARY (YES/NO, A/B/C/D),

CONTINUOUS (UNLIMITED NUMBER,12, 48, 419 ETC)

OR FREE TEXT (UNLIMITED TEXT)

HOW TO TURN ANYTHING INTO A DATABASE

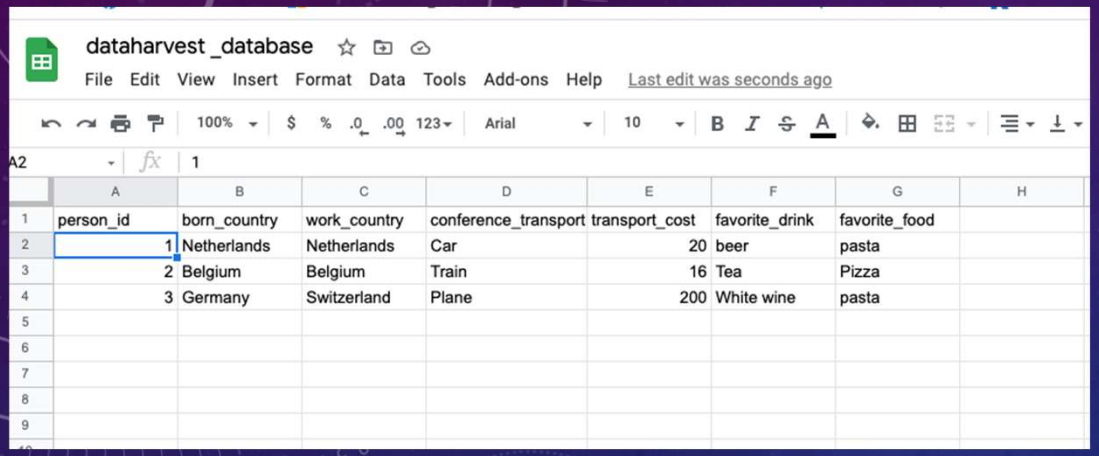


	A	B	C	D	E	F	G	H
1	person_id	born_country	work_country	conference_transport	transport_cost	favorite_drink	favorite_food	
2	1	Netherlands	Netherlands	Car	20	beer	pasta	
3	2	Belgium	Belgium	Train	16	Tea	Pizza	
4	3	Germany	Switzerland	Plane	200	White wine	pasta	
5								
6								
7								
8								
9								

DATA JOURNALISM 101

ID: WHENEVER YOU CREATE A DATABASE, THE FIRST THING YOU NEED IS A SYSTEM FOR IDENTIFYING EACH ENTRY. IT'S IMPORTANT THIS THIS COLUMN BE UNIQUE – SO WE OFTEN USE A SIMPLE 1,2,3 ETC NUMBERING SYSTEM.

HOW TO TURN ANYTHING INTO A DATABASE



dataharvest _database

File Edit View Insert Format Data Tools Add-ons Help Last edit was seconds ago

100% $ % .0 .00 123 Arial 10

A2 | 1

	A	B	C	D	E	F	G	H
1	person_id	born_country	work_country	conference_transport	transport_cost	favorite_drink	favorite_food	
2	1	Netherlands	Netherlands	Car	20	beer	pasta	
3	2	Belgium	Belgium	Train	16	Tea	Pizza	
4	3	Germany	Switzerland	Plane	200	White wine	pasta	
5								
6								
7								
8								
9								

DATA JOURNALISM 101

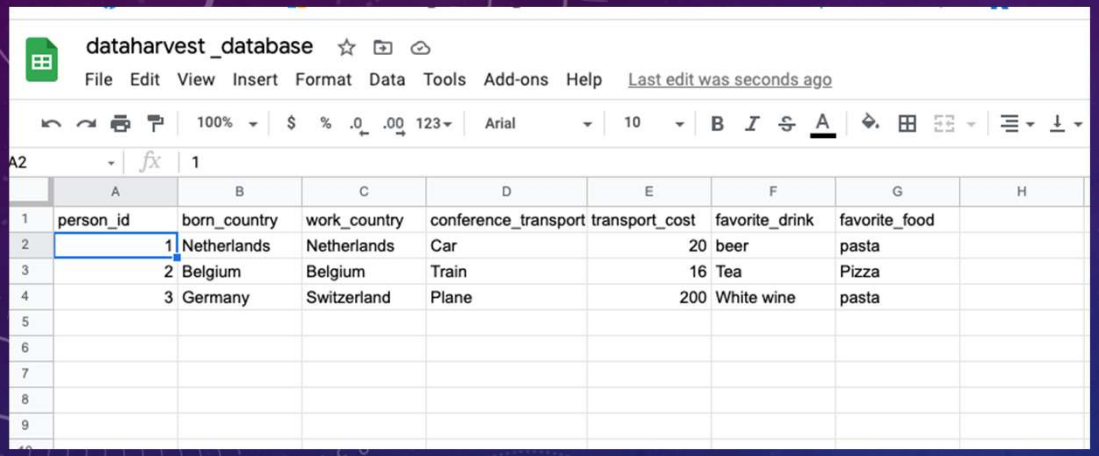
BORN_COUNTRY & WORK_COUNTRY

THESE ARE FREE TEXT VARIABLES BUT WITH LIMITED POSSIBLE ANSWERS. BUT NEED TO BE CAREFUL – ENGLAND, UK, U.K. ETC.

WHAT KINDS OF QUESTIONS CAN WE ANSWER FROM THESE COLUMNS?

HOW TO TURN ANYTHING INTO A DATABASE

DATA JOURNALISM 101



	A	B	C	D	E	F	G	H
1	person_id	born_country	work_country	conference_transport	transport_cost	favorite_drink	favorite_food	
2	1	Netherlands	Netherlands	Car	20	beer	pasta	
3	2	Belgium	Belgium	Train	16	Tea	Pizza	
4	3	Germany	Switzerland	Plane	200	White wine	pasta	
5								
6								
7								
8								
9								

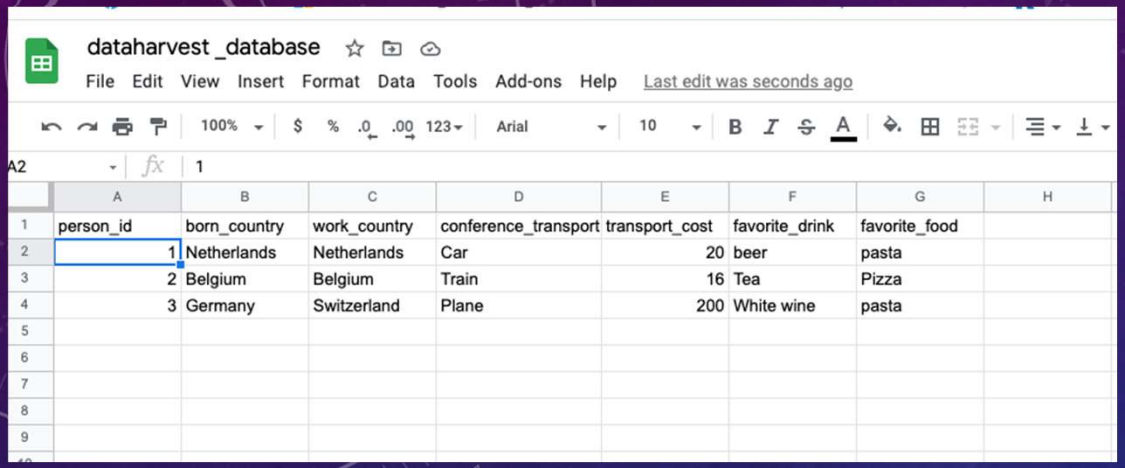
BORN_COUNTRY & WORK_COUNTRY

THESE ARE FREE TEXT VARIABLES BUT WITH LIMITED POSSIBLE ANSWERS. BUT NEED TO BE CAREFUL – ENGLAND, UK, U.K. ETC.

WHAT KINDS OF QUESTIONS CAN WE ANSWER FROM THESE COLUMNS?

How many attendees per country (born and work)? How many work in country of birth? How many from Europe? How many from Northern Europe?

HOW TO TURN ANYTHING INTO A DATABASE



dataharvest _database

File Edit View Insert Format Data Tools Add-ons Help Last edit was seconds ago

100% $ % .0 .00 123 Arial 10

A2 1

	A	B	C	D	E	F	G	H
1	person_id	born_country	work_country	conference_transport	transport_cost	favorite_drink	favorite_food	
2	1	Netherlands	Netherlands	Car	20	beer	pasta	
3	2	Belgium	Belgium	Train	16	Tea	Pizza	
4	3	Germany	Switzerland	Plane	200	White wine	pasta	
5								
6								
7								
8								
9								

DATA JOURNALISM 101

CONFERENCE_TRANSPORT

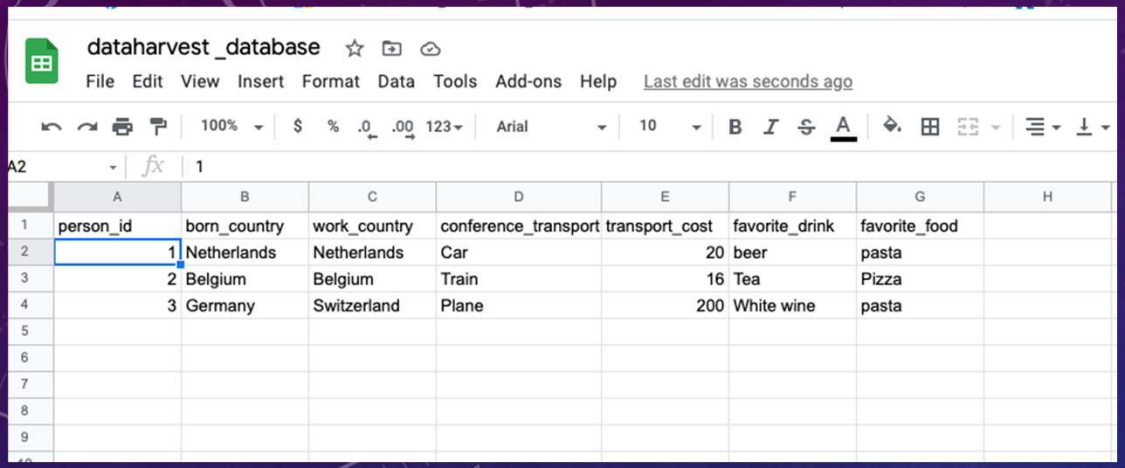
FREE TEXT TRANSPORT BUT ASSUMING LIMITED POSSIBILITY – PLANE, TRAIN, CAR. THOUGH SOMEBODY POSSIBLY WALKED.

TRANSPORT_COST

CONTINUOUS VARIABLE. EXPECT WIDE RANGE.

WHAT CAN WE LEARN FROM THESE VARIABLES. WHAT ARE SOME CAVEATS WE MIGHT NEED TO KEEP IN MIND?

HOW TO TURN ANYTHING INTO A DATABASE



dataharvest _database

File Edit View Insert Format Data Tools Add-ons Help Last edit was seconds ago

A2 | 1

	A	B	C	D	E	F	G	H
1	person_id	born_country	work_country	conference_transport	transport_cost	favorite_drink	favorite_food	
2	1	Netherlands	Netherlands	Car	20	beer	pasta	
3	2	Belgium	Belgium	Train	16	Tea	Pizza	
4	3	Germany	Switzerland	Plane	200	White wine	pasta	
5								
6								
7								
8								
9								

DATA JOURNALISM 101

CONFERENCE_TRANSPORT

FREE TEXT TRANSPORT BUT ASSUMING LIMITED POSSIBILITY – PLANE, TRAIN, CAR. THOUGH SOMEBODY POSSIBLY WALKED.

TRANSPORT_COST

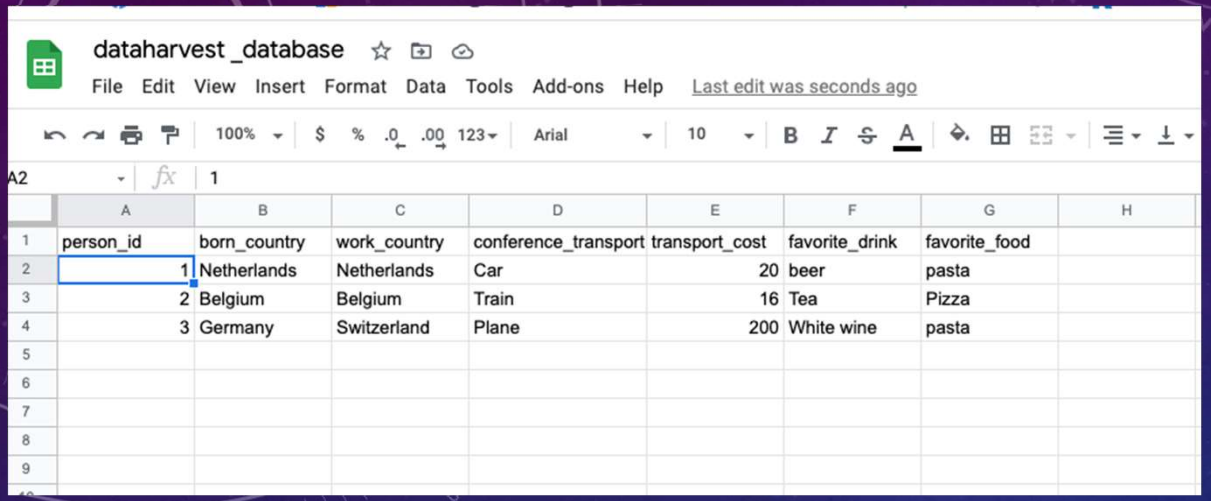
CONTINUOUS VARIABLE. EXPECT WIDE RANGE.

WHAT CAN WE LEARN FROM THESE VARIABLES. WHAT ARE SOME CAVEATS WE MIGHT NEED TO KEEP IN MIND?

What if somebody flew to Paris and took the train? Is the cost the gross cost or the cost after any reimbursements? Is the cost in Euros? USD?

HOW TO TURN ANYTHING INTO A DATABASE

DATA JOURNALISM 101



	A	B	C	D	E	F	G	H
1	person_id	born_country	work_country	conference_transport	transport_cost	favorite_drink	favorite_food	
2	1	Netherlands	Netherlands	Car	20	beer	pasta	
3	2	Belgium	Belgium	Train	16	Tea	Pizza	
4	3	Germany	Switzerland	Plane	200	White wine	pasta	
5								
6								
7								
8								
9								

FAVORITE_FOOD

FREE TEXT. COULD BE ANYTHING.

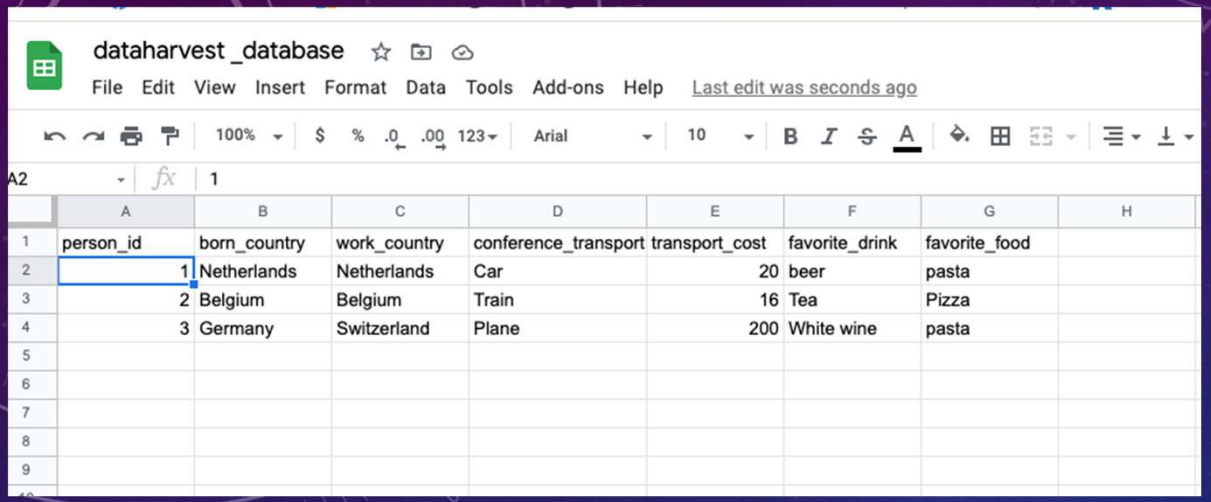
FAVORITE_DRINK

FREE TEXT COULD BE ANYTHING.

HOW MIGHT WE GROUP THESE VARIABLES?

HOW TO TURN ANYTHING INTO A DATABASE

DATA JOURNALISM 101



	A	B	C	D	E	F	G	H
1	person_id	born_country	work_country	conference_transport	transport_cost	favorite_drink	favorite_food	
2	1	Netherlands	Netherlands	Car	20	beer	pasta	
3	2	Belgium	Belgium	Train	16	Tea	Pizza	
4	3	Germany	Switzerland	Plane	200	White wine	pasta	
5								
6								
7								
8								
9								

FAVORITE_FOOD

FREE TEXT. COULD BE ANYTHING.

FAVORITE_DRINK

FREE TEXT COULD BE ANYTHING.

HOW MIGHT WE GROUP THESE VARIABLES?

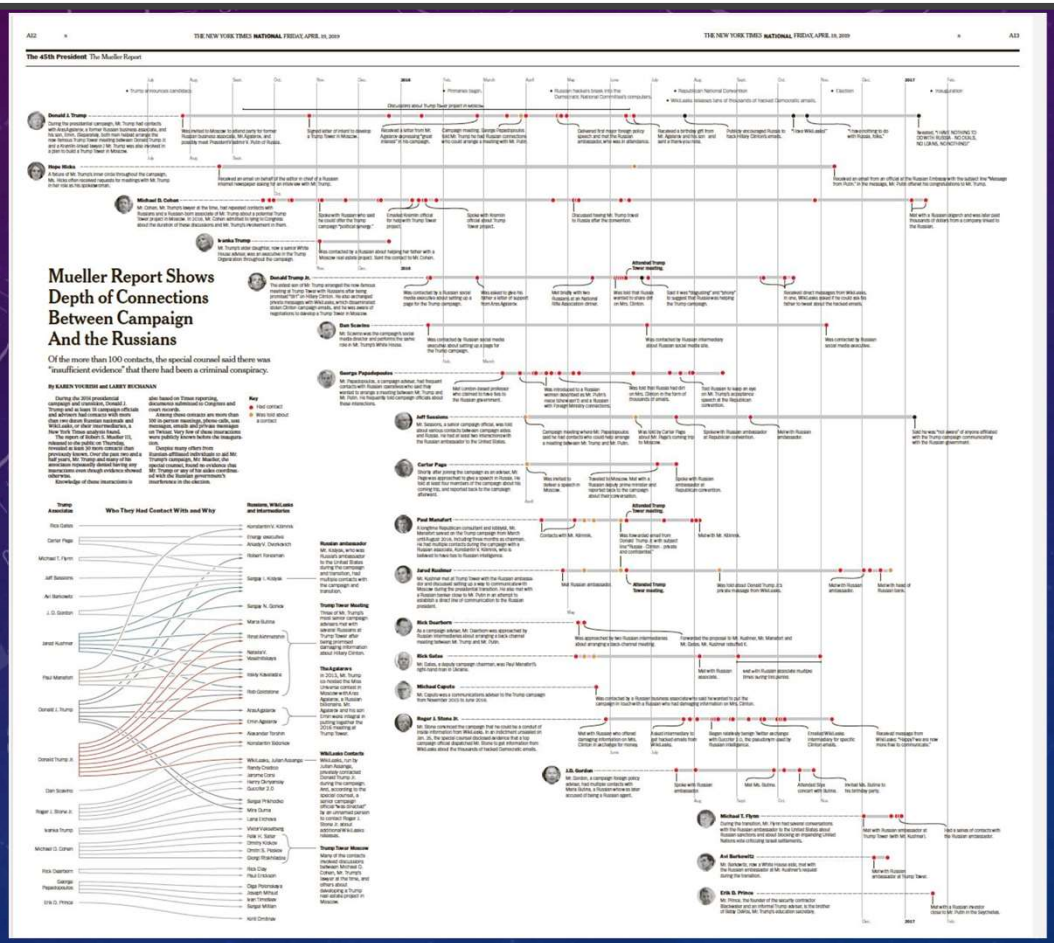
Binary Alcoholic or Not? Favorite food – by category (dinner, lunch, dessert) and/or by type (Italian, Chinese etc.)

HOW TO TURN ANYTHING INTO A DATABASE

CASE STUDIES

Robert Gebeloff -- @gebeloffnyt- DATAHARVEST 2021

HOW TO TURN ANYTHING INTO A DATABASE



THE NEW YORK TIMES **NATIONAL** FRIDAY, APRIL 19, 2019

The 45th President The Mueller Report

Mueller Report Shows Depth of Connections Between Campaign And the Russians

Of the more than 100 contacts, the special counsel said there was "insufficient evidence" that there had been a criminal conspiracy.

By KAREN YOURISH and LARRY BUCHANAN

HOW TO TURN ANYTHING INTO A DATABASE

MUELLER REPORT SHOWS DEPTH OF CONNECTIONS BETWEEN TRUMP CAMPAIGN AND RUSSIANS

BY KAREN YOURISH AND LARRY BUCHANAN

DOCUMENTED THE EXTENT TO WHICH TRUMP AND HIS ASSOCIATES HAD CONTACTS WITH RUSSIANS OR INTERMEDIARIES DURING THE CAMPAIGN AND THE TRANSITION -- AFTER CONTINUING TO DENY ANY SUCH THING HAPPENED.

Robert Gebeloff -- @gebeloffnyt- DATAHARVEST 2021

HOW TO TURN ANYTHING INTO A DATABASE

- CHRONOLOGY == POWER
- BIG MESS OF INFO – COURT FILINGS, REPORTED STORIES – ORGANIZED AND STANDARDIZED WITH A SPREADSHEET
- AT FIRST THEY WERE TRACKING CONTACTS – THEN REPORTING LED THEM TO ALSO DOCUMENT WHEN PEOPLE WERE TOLD ABOUT CONTACTS.
- 20 COLUMNS, 200 ROWS – EACH ROW SUMMARIZED IN TOOL TIP

HOW TO TURN ANYTHING INTO A DATABASE

Mass killings have claimed nearly 1,200 lives since '06

Jodi Upton, Paul Overberg and Meghan Hoyer, USA TODAY Published 2:56 p.m. ET Dec. 3, 2013 | Updated 3:44 p.m. ET Dec. 3, 2013

USA TODAY data journalists Meghan Hoyer and Paul Overberg share their comprehensive research into the nation's mass killings, and show viewers data visualization in the 'Behind the Bloodshed' interactive.

USA TODAY examined hundreds of documents and media reports.

(Photo: Shannon Hicks, AP/Newtown Bee)

CONNECT | TWEET | LINKEDIN | COMMENT | EMAIL | MORE

Since 2006, there have been more than 230 mass killings in America, and they are far more common than most people realize. They happen about every two weeks. In eight years, they have claimed nearly 1,200 lives, nearly a third of them children.

Share your feedback to help improve our site experience!

POPULAR STORIES

Most expensive disease in America could bankrupt Medicare

usatoday.com | a day ago

Trump, Supreme Court on 2020 collision course

usatoday.com | 11 hours ago

At least 40 dead in fiery Moscow jet landing

usatoday.com | 16 minutes ago

HOW TO TURN ANYTHING INTO A DATABASE

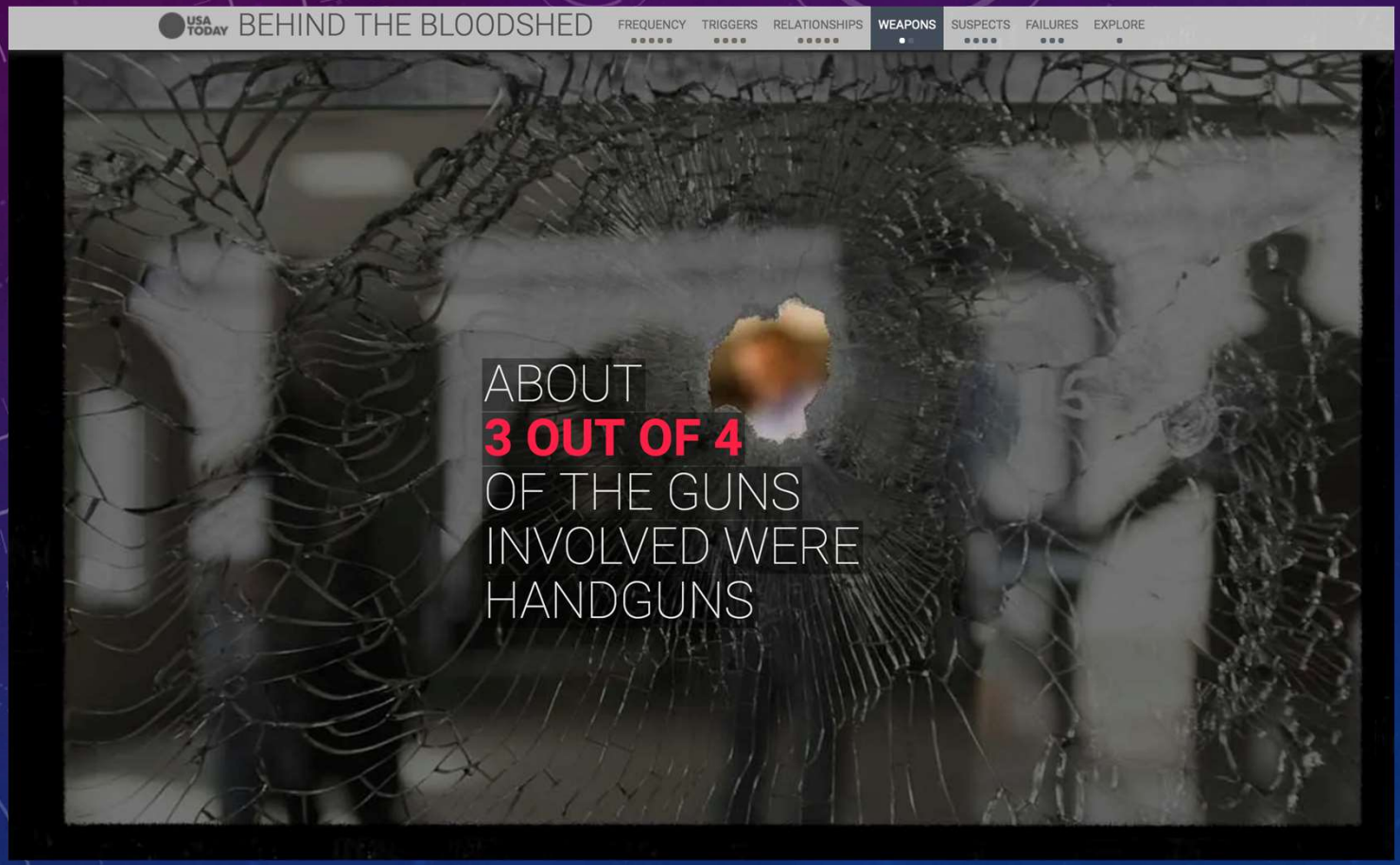
MASS KILLINGS HAVE CLAIMED NEARLY 1,200 LIVES SINCE '06

JODI UPTON, PAUL OVERBERG AND MEGHAN HOYER, USA TODAY

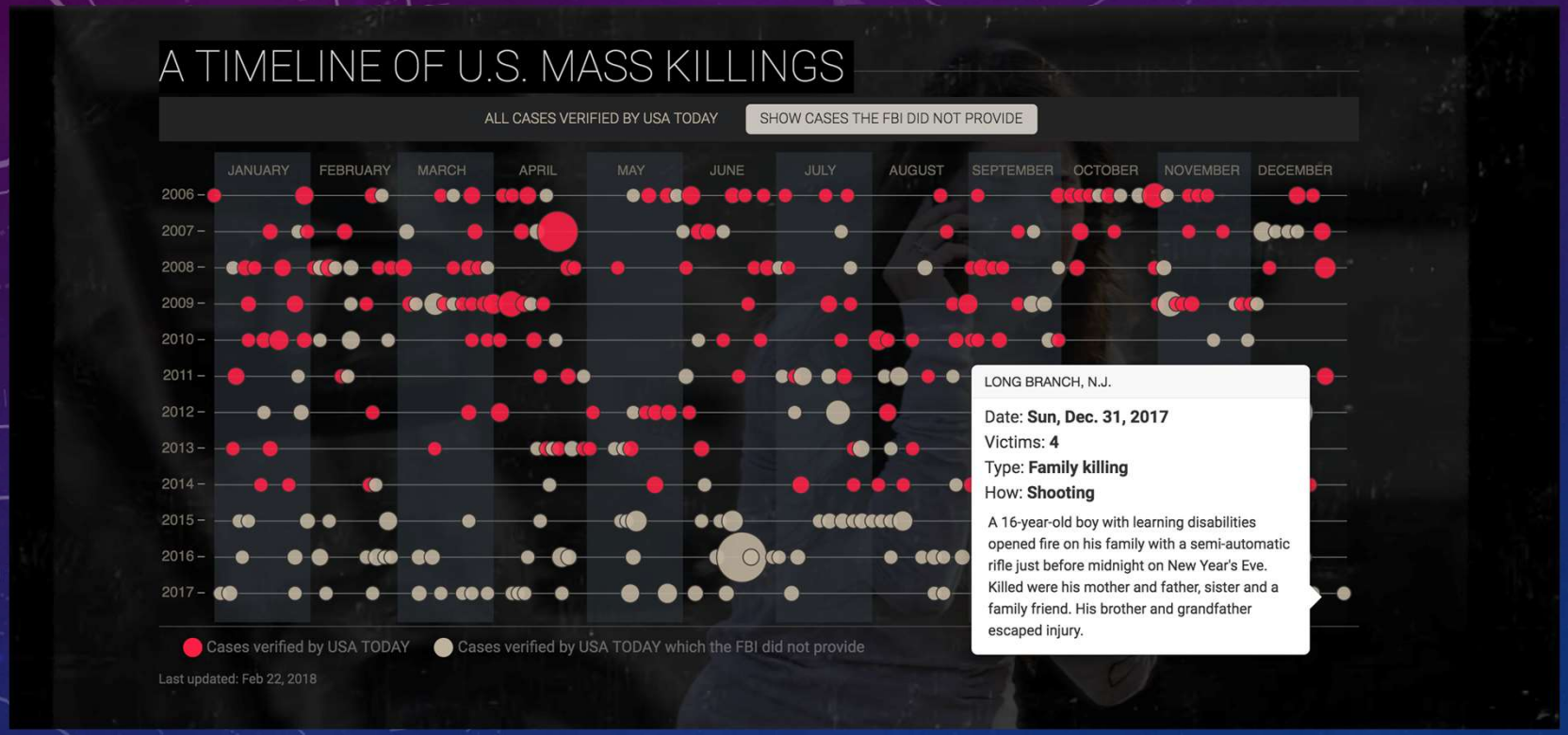
MASS KILLINGS IN THE U.S. HAPPEN FAR MORE OFTEN THAN MOST AMERICANS REALIZE. OFFICIAL FBI DATA IS INCOMPLETE AND NOT ALWAYS ACCURATE. SO REPORTERS STARTED WITH OFFICIAL DATABASE AND ADDED COLUMNS AND CASES

Robert Gebeloff -- @gebeloffnyt- DATAHARVEST 2021

HOW TO TURN ANYTHING INTO A DATABASE



HOW TO TURN ANYTHING INTO A DATABASE



Robert Gebeloff -- @gebeloffnyt- DATAHARVEST 2021

HOW TO TURN ANYTHING INTO A DATABASE

UPTON:

"A LOT OF OUTLETS … DO NOT INCLUDE WHAT POLICE DEEM DRUG OR GANG-RELATED SHOOTINGS, OR SHOOTINGS THAT ARE ENTIRELY WITHIN FAMILIES. **THE MORE CLOSELY WE LOOKED HOWEVER, THE LESS WE COULD DECLARE MOST CASES ALL ONE TYPE**.

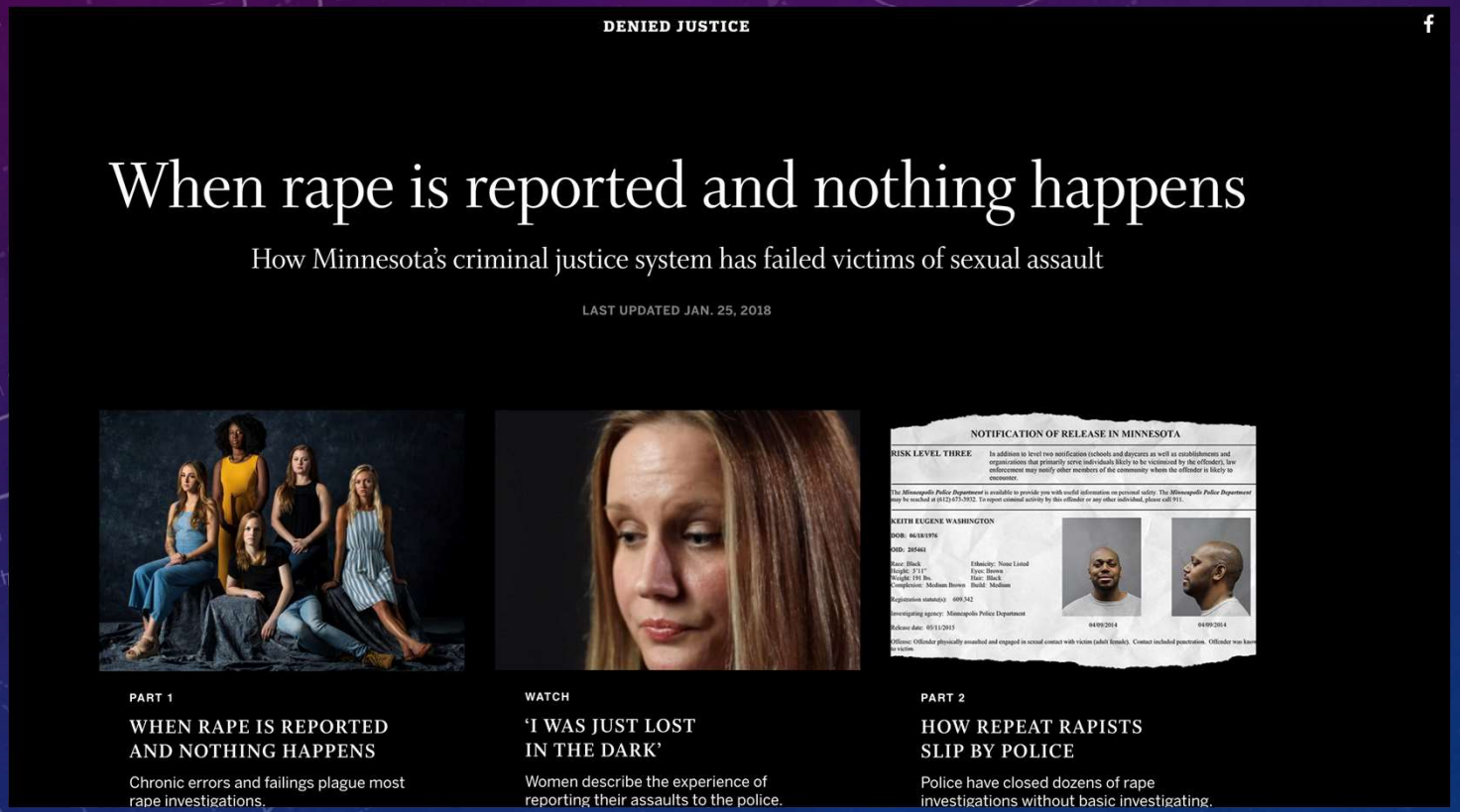
\

HOW TO TURN ANYTHING INTO A DATABASE

UPTON:

GUNS, KNIVES, ETC. ARE FAIRLY STRAIGHTFORWARD TO DETERMINE AN INTENT TO KILL, BUT WHEN IS A CAR THE WEAPON IN A MASS KILLING? OR FIRE? HOW DO WE DISTINGUISH THAT FROM ACCIDENTAL CASES? **IN THE END, WE ESTABLISHED VERY SPECIFIC CRITERIA FOR THOSE 'EDGE' CASES**

HOW TO TURN ANYTHING INTO A DATABASE



HOW TO TURN ANYTHING INTO A DATABASE

SEXUAL ASSAULT CASES IN THE TWIN CITIES AND ACROSS MINNESOTA ARE BEING INVESTIGATED POORLY OR NOT AT ALL, LEAVING MANY WOMEN FEELING BETRAYED BY A SYSTEM THEY ONCE TRUSTED

BY BRANDON STAHL, JENNIFER BJORHUS AND MARYJO WEBSTER

REPORTERS OBTAINED AND ANALYZED MORE THAN 1,500 POLICE REPORTS TO SHOWED THAT LESS THAN 1 IN 10 SEX ASSAULTS RESULTED IN A CONVICTION AND THAT IN ABOUT HALF THE CASES, POLICE DIDN'T TAKE BASIC INVESTIGATORY STEPS.

Robert Gebeloff -- @gebeloffnyt- DATAHARVEST 2021

HOW TO TURN ANYTHING INTO A DATABASE

WEBSTER: "WE NEEDED TO SHOW THAT THIS WAS A SYSTEMIC PROBLEM. THE ONLY WAY TO DO THAT WOULD **BE TO ANALYZE THOUSANDS OF CASES OURSELVES**. WE KNEW RIGHT FROM THE START THAT WE WANTED TO TRACK WHAT HAPPENS TO CASES -- DID THEY GET SENT TO PROSECUTOR? WERE CHARGES FILED? WAS THERE A CONVICTION?"

HOW TO TURN ANYTHING INTO A DATABASE

WEBSTER: "BUT THE HARDER THING TO FIGURE OUT WAS HOW TO MEASURE HOW WELL THE POLICE WERE INVESTIGATING THESE CASES. ULTIMATELY **WE CAME UP WITH A LIST OF BASIC POLICING STEPS THAT EXPERTS TOLD US POLICE SHOULD BE DOING ON EVERY CASE**. THE ANSWERS TO THOSE QUESTIONS BECAME THE BASIS FOR US SAYING POLICE MISHANDLING OF CASES WAS WIDESPREAD."

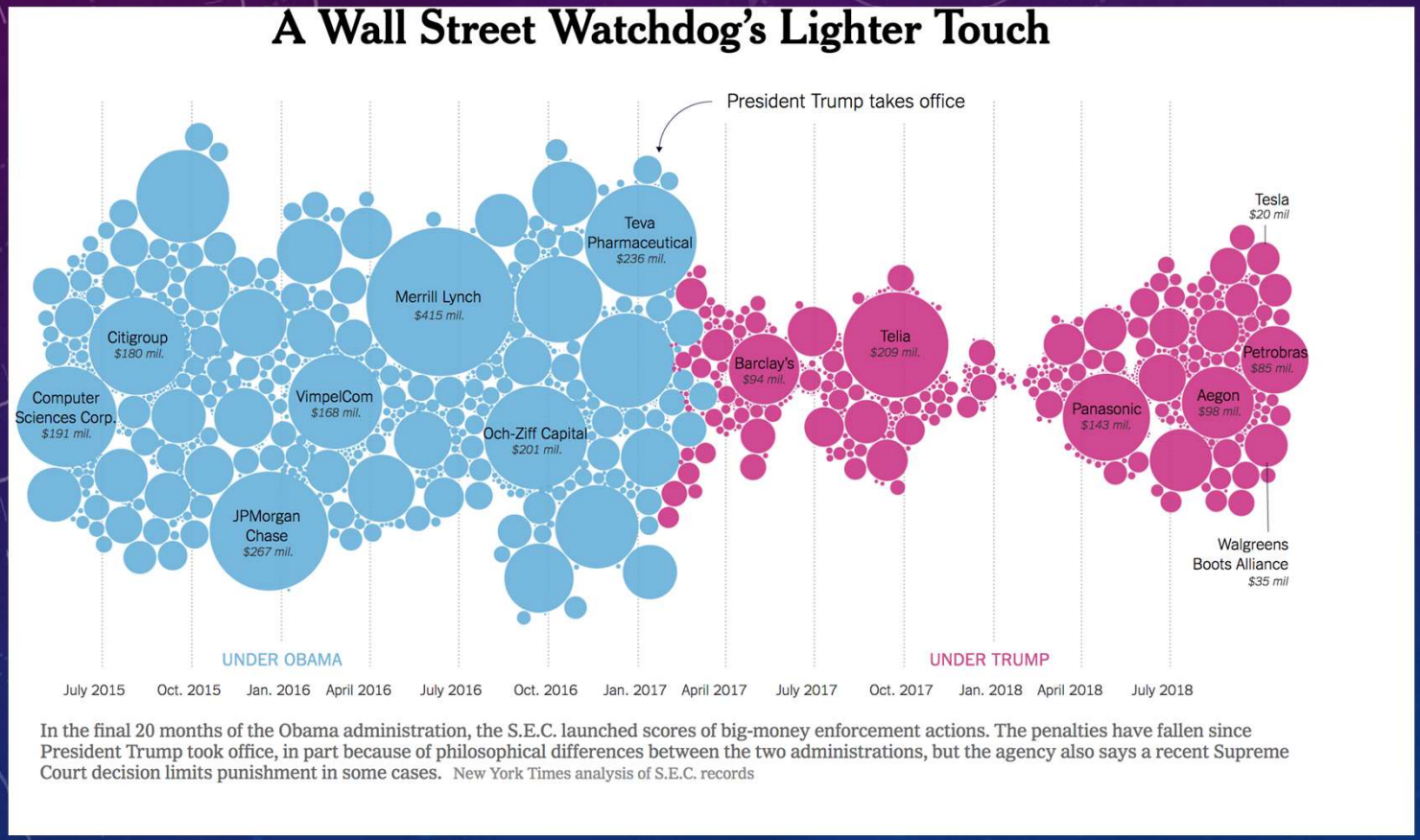
HOW TO TURN ANYTHING INTO A DATABASE

WEBSTER: "WE SPENT **THE BETTER PART OF A MONTH GOING THROUGH A TRIAL AND ERROR PROCESS**, FIGURING OUT WHICH FIELDS WERE WORKING, WHICH NEEDED TO BE TWEAKED, WHAT NEEDED TO BE ADDED OR REMOVED.."

HOW TO TURN ANYTHING INTO A DATABASE

WEBSTER: "ONE OF THE QUESTIONS ASKED IF THE VICTIM WAS INTOXICATED AT THE TIME OF THE INCIDENT. THE POLICE REPORTS OFTEN INDICATED THE VICTIM HAD BEEN AT A BAR OR AT A PARTY, **BUT WAS THAT ENOUGH TO SAY SHE WAS INTOXICATED?** WE ULTIMATELY DECIDED THAT WE NEEDED SOMETHING MORE -- HER SAYING THAT SHE PASSED OUT, OR HAD GOTTEN SICK, OR THAT SHE TOLD POLICE SHE WAS VERY INTOXICATED. **SOMETHING MORE THAN THAT SHE HAD MERELY BEEN DRINKING**.

HOW TO TURN ANYTHING INTO A DATABASE



HOW TO TURN ANYTHING INTO A DATABASE

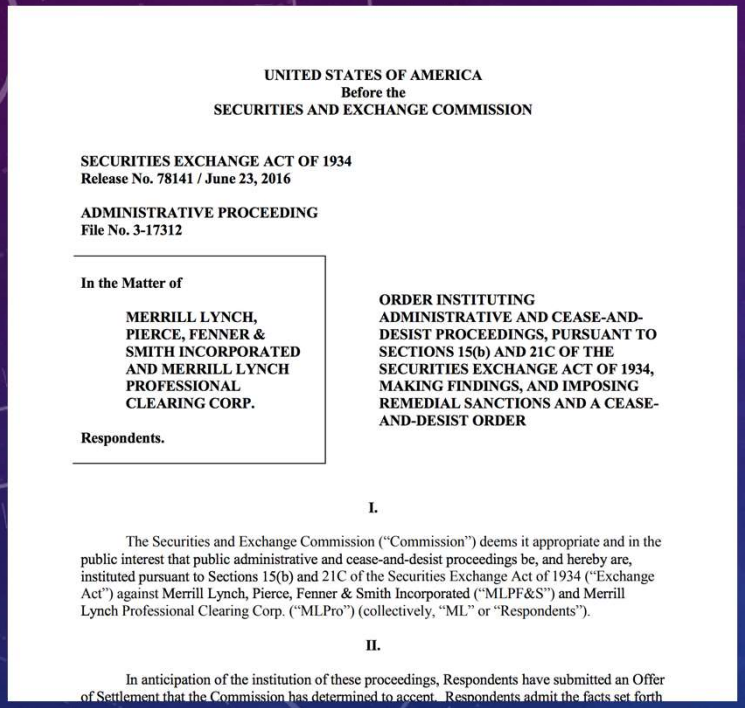
TRUMP ADMINISTRATION SPARES CORPORATE WRONGDOERS BILLIONS IN PENALTIES

BY BEN PROTESS, ROBERT GEBELOFF AND DANIELLE IVORY

REPORTERS SCRAPED AND ANALYZED 2,158 ENFORCEMENT ACTIONS AGAINST CORPORATE WRONG-DOERS AND FOUND A 62 PERCENT DROP IN PENALTIES LEVIED BY THE TRUMP ADMINISTRATION IN ITS FIRST 20 MONTHS.

Robert Gebeloff -- @gebeloffnyt- DATAHARVEST 2021

HOW TO TURN ANYTHING INTO A DATABASE



UNITED STATES OF AMERICA
Before the
SECURITIES AND EXCHANGE COMMISSION

SECURITIES EXCHANGE ACT OF 1934
Release No. 78141 / June 23, 2016

ADMINISTRATIVE PROCEEDING
File No. 3-17312

In the Matter of

MERRILL LYNCH, PIERCE, FENNER & SMITH INCORPORATED AND MERRILL LYNCH PROFESSIONAL CLEARING CORP.

Respondents.

ORDER INSTITUTING ADMINISTRATIVE AND CEASE-AND-DESIST PROCEEDINGS, PURSUANT TO SECTIONS 15(b) AND 21C OF THE SECURITIES EXCHANGE ACT OF 1934, MAKING FINDINGS, AND IMPOSING REMEDIAL SANCTIONS AND A CEASE-AND-DESIST ORDER

I.

The Securities and Exchange Commission ("Commission") deems it appropriate and in the public interest that public administrative and cease-and-desist proceedings be, and hereby are, instituted pursuant to Sections 15(b) and 21C of the Securities Exchange Act of 1934 ("Exchange Act") against Merrill Lynch, Pierce, Fenner & Smith Incorporated ("MLPF&S") and Merrill Lynch Professional Clearing Corp. ("MLPro") (collectively, "ML" or "Respondents").

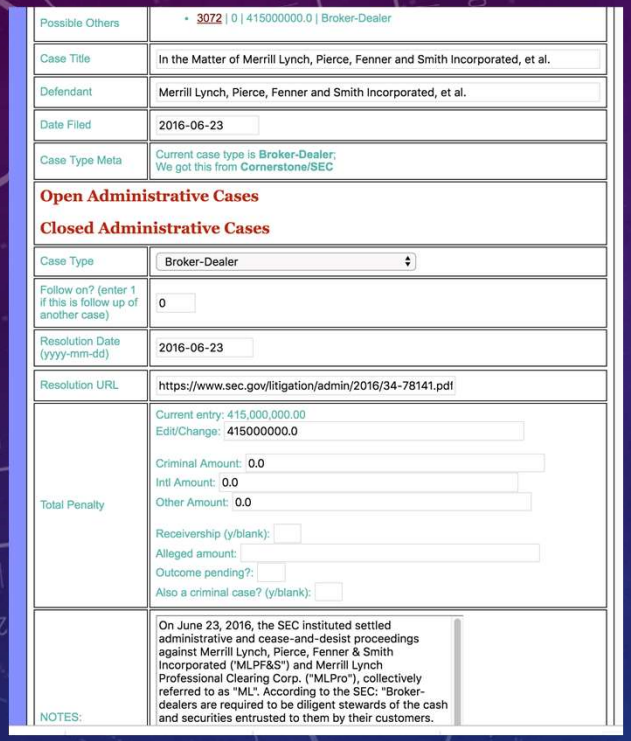
II.

In anticipation of the institution of these proceedings, Respondents have submitted an Offer of Settlement that the Commission has determined to accept. Respondents admit the facts set forth

There was no readily available database of all SEC cases. So we:

1) Scraped all "litigation releases", press releases, and administrative actions posted on the Web.
2) Read, and de-duped them.
3) Followed announcements through the court system.
4) Turned documents into data.

HOW TO TURN ANYTHING INTO A DATABASE

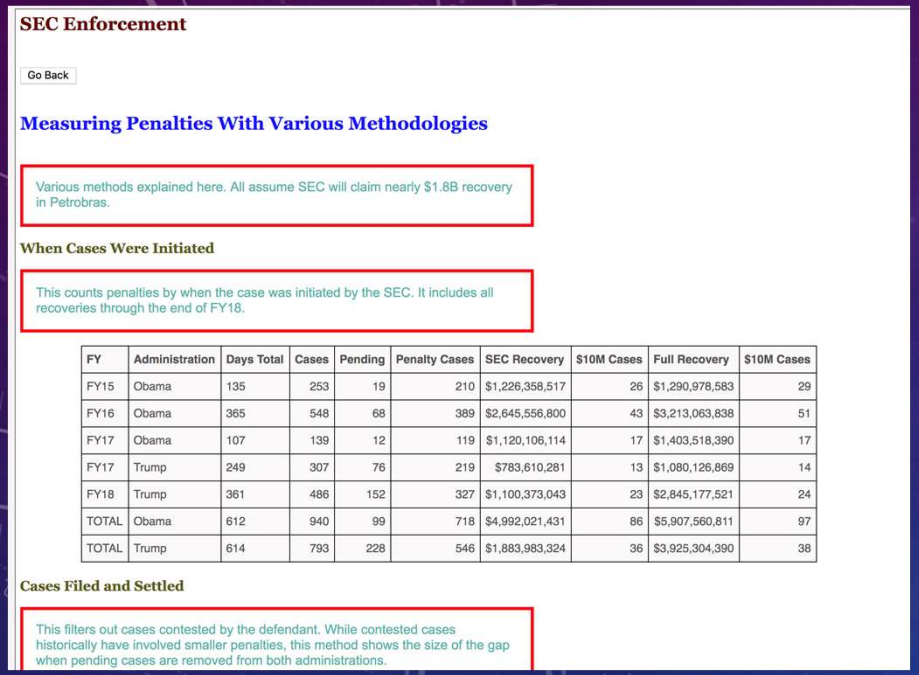


Possible Others	• 3072 | 0 | 415000000.0 | Broker-Dealer
Case Title	In the Matter of Merrill Lynch, Pierce, Fenner and Smith Incorporated, et al.
Defendant	Merrill Lynch, Pierce, Fenner and Smith Incorporated, et al.
Date Filed	2016-06-23
Case Type Meta	Current case type is **Broker-Dealer**; We got this from **Cornerstone/SEC**
Open Administrative Cases **Closed Administrative Cases**	
Case Type	Broker-Dealer
Follow on? (enter 1 if this is follow up of another case)	0
Resolution Date (yyyy-mm-dd)	2016-06-23
Resolution URL	https://www.sec.gov/litigation/admin/2016/34-78141.pdf
Total Penalty	Current entry: 415,000,000.00 Edit/Change: 415000000.0 Criminal Amount: 0.0 Intl Amount: 0.0 Other Amount: 0.0 Receivership (y/blank): Alleged amount: Outcome pending?: Also a criminal case? (y/blank):
NOTES:	On June 23, 2016, the SEC instituted settled administrative and cease-and-desist proceedings against Merrill Lynch, Pierce, Fenner & Smith Incorporated ("MLPF&S") and Merrill Lynch Professional Clearing Corp. ("MLPro"), collectively referred to as "ML". According to the SEC: "Broker-dealers are required to be diligent stewards of the cash and securities entrusted to them by their customers.

This is a little hard to read on this screen but among the things we captured were:

1) Case category
2) Amount of civil penalty
3) Amount of criminal penalty
4) Amount of penalty levied by international regulators in related cases
5) Notes about the case.
6) A “bookmark” function so we could save and review the most interesting cases.

HOW TO TURN ANYTHING INTO A DATABASE



SEC Enforcement

Go Back

Measuring Penalties With Various Methodologies

Various methods explained here. All assume SEC will claim nearly $1.8B recovery in Petrobras.

When Cases Were Initiated

This counts penalties by when the case was initiated by the SEC. It includes all recoveries through the end of FY18.

FY	Administration	Days Total	Cases	Pending	Penalty Cases	SEC Recovery	$10M Cases	Full Recovery	$10M Cases
FY15	Obama	135	253	19	210	$1,226,358,517	26	$1,290,978,583	29
FY16	Obama	365	548	68	389	$2,645,556,800	43	$3,213,063,838	51
FY17	Obama	107	139	12	119	$1,120,106,114	17	$1,403,518,390	17
FY17	Trump	249	307	76	219	$783,610,281	13	$1,080,126,869	14
FY18	Trump	361	486	152	327	$1,100,373,043	23	$2,845,177,521	24
TOTAL	Obama	612	940	99	718	$4,992,021,431	86	$5,907,560,811	97
TOTAL	Trump	614	793	228	546	$1,883,983,324	36	$3,925,304,390	38

Cases Filed and Settled

This filters out cases contested by the defendant. While contested cases historically have involved smaller penalties, this method shows the size of the gap when pending cases are removed from both administrations.

Because we built our own internal Web app to manage the data, we were also able to write “reports” that summed the data in real time.

As we reported, one thing we learned is that reasonable people could count penalties differently.

So in this report, we go over what our figures show under various methods of counting.

HOW TO TURN ANYTHING INTO A DATABASE

ırsca sec	urska nyt		Notes
$1,879,634	$1,879,634	Rob	Correct, she has typo in Ireeco Limited, ht x
$4,250,000	$4,250,000	Sylvia	Our total is correct. Urska did not count di x
$15,000,000	$15,000,000	Rob	We mis-added, changed to 15M
$150,000	$150,000	Sylvia	The case is split: 150,000 is correct for Anthony Andrade. Haven't changed in database because
$174,647	$174,647	Gilda	Urska is right. The total for three defendants is 174,647. Haven't changed in database because I
$117,152	$117,152	Sylvia	Her total is correct for that specific defendant (Ross). Our total reflects the defendants as a group
$489,554	$448,604	Gilda	We match Urska's SEC amount.
$8,858,285	$8,858,285	Rob	Changed in database. Ours was rounded.
$8,858,285	$8,858,285	Sylvia	Duplicate of above.
$10,063,464	$10,063,464	Gilda	Changed in database. Urska is right. We did not factor in the prejudgment interest: 126,697.94.
$5,419,800	$5,250,000	Rob	We were missing two minor defendants, changed in db
$4,928,301	$4,928,301	Sylvia	Changed in database. I had not taken the final judgment into account; instead relying solely on th
$50,000	$50,000	Gilda	I can only confirm Urska's figures. Not able to find case in Pacer to reverse engineer our figures.
$238,548	$238,548	Rob	we were missing two defendants, fixed in db
$95,559	$82,418	Sylvia	dupe
$30,000	$30,000	Gilda	Pierce's civil penalty is correctly noted in our database, where we note the total (not split).

About ¾ of the way through the project, we discovered that Professor Urksa Velikonja from Georgetown University also kept a complete database of cases which she used for her own research. She agreed to share her data so we could reconcile different interpretations of case results, which we did using a Google Sheet.

HOW TO TURN ANYTHING INTO A DATABASE

By going through the filings case by case, we developed some expertise on how penalties are counted – for example, we discovered that the SEC is very generous to itself when counting it's own impact:

SUMMARY SAYS: The Commission brought 821 actions (490 of which were "stand alone" actions) and obtained judgments and orders totaling more than **$3.9 billion** in disgorgement and penalties.

HOW TO TURN ANYTHING INTO A DATABASE

FOOTNOTE 17 SAYS: **Money ordered in FY 2018 includes** In the Matter of Petroleo Brasileiro S.A. – Petrobras, AP File No. 3-18843, Securities Exchange Act Release No. 34-84295 (Sept. 27, 2018), which orders payment of disgorgement and prejudgment interest totaling **$933,473,797** and a penalty of $853,200,000 and provides that certain of respondent's obligations for those amounts (all but $85,320,000) shall be deemed satisfied if, within one year, it makes payments in related actions involving the Department of Justice, a class action settlement fund, and Brazilian authorities

HOW TO TURN ANYTHING INTO A DATABASE

When the S.E.C. publicizes enforcement figures, it includes penalties that other government agencies have collected in related actions, and calculates them based on when the cases were resolved, not when they were filed. This method, which predates Mr. Trump, credits the S.E.C. under his administration with bringing in $1.7 billion that Petrobras, a Brazilian company, was ordered to pay to other government agencies and plaintiffs in a private lawsuit.

Tabulating penalties this way still leads to a $938 million gap between the two administrations, according to Urska Velikonja, a professor at Georgetown University Law Center who studies the S.E.C. (The Times analysis, which found a $3.1 billion gap, excluded payments to outside agencies and attributed penalties to each administration based on when a case was initiated, not settled.)

HOW TO TURN ANYTHING INTO A DATABASE

How a DIY Database Project Might Play out…

1) What is the target? Court records? Police reports? Inspections. Tweets.

HOW TO TURN ANYTHING INTO A DATABASE

How a DIY Database Project Might Play out…

1) What is the target? Court records? Police reports? Inspections. Tweets.
2) How to get it? FOIA? Downloading and parsing? Scraping?

HOW TO TURN ANYTHING INTO A DATABASE

How a DIY Database Project Might Play out…

1) What is the target? Court records? Police reports? Inspections. Tweets.
2) How to get it? FOIA? Downloading and parsing? Scraping?
3) Choosing the right technology:

EXCEL	Easy to use, hard to collaborate
GOOGLE SHEETS	Easy to use, easy to collaborate, but clunky with large entries.
AIRTABLE	Document and data management software, $20 US per month.
DIY PROGRAMMING	Customize exactly what you want, but need a programmer on board your project.

HOW TO TURN ANYTHING INTO A DATABASE

Data Entry With Airtable

For the data entry part, I initially thought we could use Google Forms. But it has one big limitation: It isn't very easy to search for and edit existing records. You have to wade into the spreadsheet and traverse your way through many rows and columns to find the bit you want to edit.

Alan Palazzolo introduced me to Airtable. It allows you to set up a data entry form that's even slicker than a Google Form, and then it also allows you to set up different "views" of your underlying data that make editing far easier.

Mary Jo Webster on Airtable

HOW TO TURN ANYTHING INTO A DATABASE

How a DIY Database Project Might Play out…

4) What are you going to capture from your records? – Plan it. Brainstorm ideas. What variables? Continuous, binary? Free text? Test with a few documents and revise.

HOW TO TURN ANYTHING INTO A DATABASE

How a DIY Database Project Might Play out…

4) What are you going to capture from your records? – Plan it. Brainstorm ideas. What variables? Continuous, binary? Free text? Test with a few documents and revise.

5) How are you going to track your work. Paper records and data entry need to have a corresponding ID so everything can be crosschecked.

HOW TO TURN ANYTHING INTO A DATABASE

How a DIY Database Project Might Play out…

4) What are you going to capture from your records? – Plan it. Brainstorm ideas. What variables? Continuous, binary? Free text? Test with a few documents and revise.

5) How are you going to track your work? Paper records and data entry need to have a corresponding ID so everything can be crosschecked.

6) How are you going to factcheck your work? You need to screen for unusual entries. You need to screen for unusual conclusions. You need to do this as you go so you nip any problems in the bud.

HOW TO TURN ANYTHING INTO A DATABASE

FINAL ADVICE:

Jodi Upton: **“**Think through as granular as possible… Have a system to debate criteria… Compare to as many outside incomplete/data sets as possible”

Mary Jo Webster: “There are 2 key elements -- the first is to spend a lot of time thinking ahead (don’t just dive in). We spent hours and hours thinking through what we might want to say in our stories and then translating that into fields we would need to have…. The second is to bake in some time for the trial and error process. Be prepared to have to go back and change things on the first dozen or two dozen records you’ve entered. “

HOW TO TURN ANYTHING INTO A DATABASE

HAPPY TO ANSWER ANY QUESTIONS AND DISCUSS ANY SITUATIONS YOU HAVE ENCOUNTERED IN YOUR OWN WORK…

Robert Gebeloff -- @gebeloffnyt- DATAHARVEST 2021