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

FOR ALL STORIES, WE NEED SOURCES:
MY APPROACH TO STORIES

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

PEOPLE – FINDING PEOPLE WHO KNOW THE INFO, GETTING THEM TO TELL YOU THE INFO YOU NEED
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- DATARVEST 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
MY APPROACH TO STORIES

I RARELY START OUT WITH A DATASET AND SAY, “LET’S GO LOOKING FOR AN INVESTIGATIVE STORY!”
HOW TO TURN ANYTHING INTO A DATABASE

MY APPROACH TO STORIES

HAVE “A DATA STATE OF MIND” -- @MARYJOWEBSTER.
MY APPROACH TO STORIES
FOR WHATEVER TOPIC YOU’RE EXPLORING, ARE THERE THINGS YOU CAN BETTER QUANTIFY WITH DATA?
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?
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- DATALAB 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

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

Robert Gebeloff -- @gebeloffnyt - DATAHARVEST 2021
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.
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.
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.

Robert Gebeloff -- @gebeloffnyt - DataHarvest 2021
# How to Turn Anything into a Database

This image shows a Google Sheets spreadsheet named `dataharvest_database`. The spreadsheet contains data about people's travel preferences and expenses. Here is a breakdown of the table:

<table>
<thead>
<tr>
<th>A</th>
<th>B</th>
<th>C</th>
<th>D</th>
<th>E</th>
<th>F</th>
<th>G</th>
<th>H</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>person_id</td>
<td>born_country</td>
<td>work_country</td>
<td>conference_transport</td>
<td>transport_cost</td>
<td>favorite_drink</td>
<td>favorite_food</td>
</tr>
<tr>
<td>2</td>
<td>1</td>
<td>Netherlands</td>
<td>Netherlands</td>
<td>Car</td>
<td>20</td>
<td>beer</td>
<td>pasta</td>
</tr>
<tr>
<td>3</td>
<td>2</td>
<td>Belgium</td>
<td>Belgium</td>
<td>Train</td>
<td>16</td>
<td>Tea</td>
<td>Pizza</td>
</tr>
<tr>
<td>4</td>
<td>3</td>
<td>Germany</td>
<td>Switzerland</td>
<td>Plane</td>
<td>200</td>
<td>White wine</td>
<td>pasta</td>
</tr>
</tbody>
</table>
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)
ID: WHENEVER YOU CREATE A DATABASE, THE FIRST THING YOU NEED IS A SYSTEM FOR IDENTIFYING EACH ENTRY. IT’S IMPORTANT THIS COLUMN BE UNIQUE – SO WE OFTEN USE A SIMPLE 1,2,3 ETC NUMBERING SYSTEM.
HOW TO TURN ANYTHING INTO A DATABASE

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?

Robert Gebeloff -- @gebeloffnyt- DATAHARVEST 2021
How to turn anything into a database

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 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

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?

Robert Gebeloff -- @gebeloffnyt- DATAHARVEST 2021
### 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**

<table>
<thead>
<tr>
<th>A</th>
<th>B</th>
<th>C</th>
<th>D</th>
<th>E</th>
<th>F</th>
<th>G</th>
<th>H</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>person_id</td>
<td>born_country</td>
<td>work_country</td>
<td>conference_transport</td>
<td>transport_cost</td>
<td>favorite_drink</td>
<td>favorite_food</td>
</tr>
<tr>
<td>2</td>
<td>Netherlands</td>
<td>Netherlands</td>
<td>Car</td>
<td>20</td>
<td>beer</td>
<td>pasta</td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>Belgium</td>
<td>Belgium</td>
<td>Train</td>
<td>16</td>
<td>Tea</td>
<td>Pizza</td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>Germany</td>
<td>Switzerland</td>
<td>Plane</td>
<td>200</td>
<td>White wine</td>
<td>pasta</td>
<td></td>
</tr>
</tbody>
</table>

---

Robert Gebeloff -- @gebeloffnyt- DATAHARVEST 2021
HOW TO TURN ANYTHING INTO A DATABASE

<table>
<thead>
<tr>
<th></th>
<th>A</th>
<th>B</th>
<th>C</th>
<th>D</th>
<th>E</th>
<th>F</th>
<th>G</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>person_id</td>
<td>born_country</td>
<td>work_country</td>
<td>conference_transport</td>
<td>transport_cost</td>
<td>favorite_drink</td>
<td>favorite_food</td>
</tr>
<tr>
<td>2</td>
<td>1</td>
<td>Netherlands</td>
<td>Netherlands</td>
<td>Car</td>
<td>20</td>
<td>beer</td>
<td>pasta</td>
</tr>
<tr>
<td>3</td>
<td>2</td>
<td>Belgium</td>
<td>Belgium</td>
<td>Train</td>
<td>16</td>
<td>Tea</td>
<td>Pizza</td>
</tr>
<tr>
<td>4</td>
<td>3</td>
<td>Germany</td>
<td>Switzerland</td>
<td>Plane</td>
<td>200</td>
<td>White wine</td>
<td>pasta</td>
</tr>
</tbody>
</table>

FAVORITE_FOOD
FREE TEXT. COULD BE ANYTHING.

FAVORITE_DRINK
FREE TEXT COULD BE ANYTHING.

HOW MIGHT WE GROUP THESE VARIABLES?

Robert Gebeloff -- @gebeloffnyt - DATAHARVEST 2021
HOW TO TURN ANYTHING INTO A DATABASE

Data Journalism 101

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.)

Robert Gebeloff -- @gebeloffnyt - DATAHARVEST 2021
HOW TO TURN ANYTHING INTO A DATABASE

CASE STUDIES

Robert Gebeloff -- @gebeloffnyt - DATAHARVEST 2021
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.
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.

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.
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
HOW TO TURN ANYTHING INTO A DATABASE

ABOUT 3 OUT OF 4 OF THE GUNS INVOLVED WERE HANDGUNS
HOW TO TURN ANYTHING INTO A DATABASE

A TIMELINE OF U.S. MASS KILLINGS

LONG BRANCH, N.J.
Date: Sun, Dec. 31, 2017
Victims: 4
Type: Family killing
How: Shooting
A 16-year-old boy with learning disabilities opened fire on his family with a semi-automatic rifle just before midnight on New Year’s Eve. Killed were his mother and father, sister and a family friend. His brother and grandfather escaped injury.

Robert Gebeloff -- @gebeloffnyt- DATAHARVEST 2021
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**
When rape is reported and nothing happens

How Minnesota’s criminal justice system has failed victims of sexual assault

LAST UPDATED: JAN. 25, 2018

PART 1
WHEN RAPE IS REPORTED AND NOTHING HAPPENS
Chronic errors and failings plague most rape investigations.

PART 2
HOW REPEAT RAPISTS SLIP BY POLICE
Police have closed dozens of rape investigations without basic investigating.

WATCH
'I WAS JUST LOST IN THE DARK'
Women describe the experience of reporting their assaults to the police.
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.
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?”
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.”
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.”
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."
In the final 20 months of the Obama administration, the S.E.C. launched scores of big-money enforcement actions. The penalties have fallen since President Trump took office, in part because of philosophical differences between the two administrations, but the agency also says a recent Supreme Court decision limits punishment in some cases.  

*New York Times analysis of S.E.C. records*
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
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

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.

---

**Open Administrative Cases**

**Closed Administrative Cases**

- **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
- **Resolution Date**: 2016-06-23

**Total Penalty**

- Current entry: $1,000,000,000
- Amount of: $0
- Other Amount: $0
- Civil Penalty: $0
- Criminal Penalty: $0
- Other: $0
- Financial Penalty (if any): $0
- Alleged amount: $0
- Outcome pending:
- Also a criminal case? (yes/no): No

---

**NOTES:**

On June 30, 2016, the SEC instituted settled administrative and cease-and-desist proceedings against Merrill Lynch, Pierce, Fenner & Smith Incorporated, and its executive team, including David G. Driscoll, the firm’s President and Chief Executive Officer, and Thomas F. Underwood, its Executive Vice President, among others, for their roles in a classic ‘corporate fraud’ scheme. The SEC’s complaint alleged that Merrill Lynch, Pierce, Fenner & Smith Incorporated (collectively referred to as “M”) and its employees (collectively referred to as “M’s dealers”) defrauded U.S. and non-U.S. customers by, among other things, misappropriating cash and securities entrusted to them by their customers.
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.
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.
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 its 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.
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.
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...

HOW TO TURN ANYTHING INTO A DATABASE

How a DIY Database Project Might Play out...

2) How to get it? FOIA? Downloading and parsing? Scraping?
How a DIY Database Project Might Play out...

2) How to get it? FOIA? Downloading and parsing? Scraping?
3) Choosing the right technology:

<table>
<thead>
<tr>
<th>Tool</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>EXCEL</td>
<td>Easy to use, hard to collaborate</td>
</tr>
<tr>
<td>GOOGLE SHEETS</td>
<td>Easy to use, easy to collaborate, but clunky with large entries.</td>
</tr>
<tr>
<td>AIRTABLE</td>
<td>Document and data management software, $20 US per month.</td>
</tr>
<tr>
<td>DIY PROGRAMMING</td>
<td>Customize exactly what you want, but need a programmer on board your project.</td>
</tr>
</tbody>
</table>
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.
HOW TO TURN ANYTHING INTO A DATABASE

How a DIY Database Project Might Play out...

How a DIY Database Project Might Play out…


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 a DIY Database Project Might Play out...


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...