Collaborative Journalism Resource Hub

There’s data!

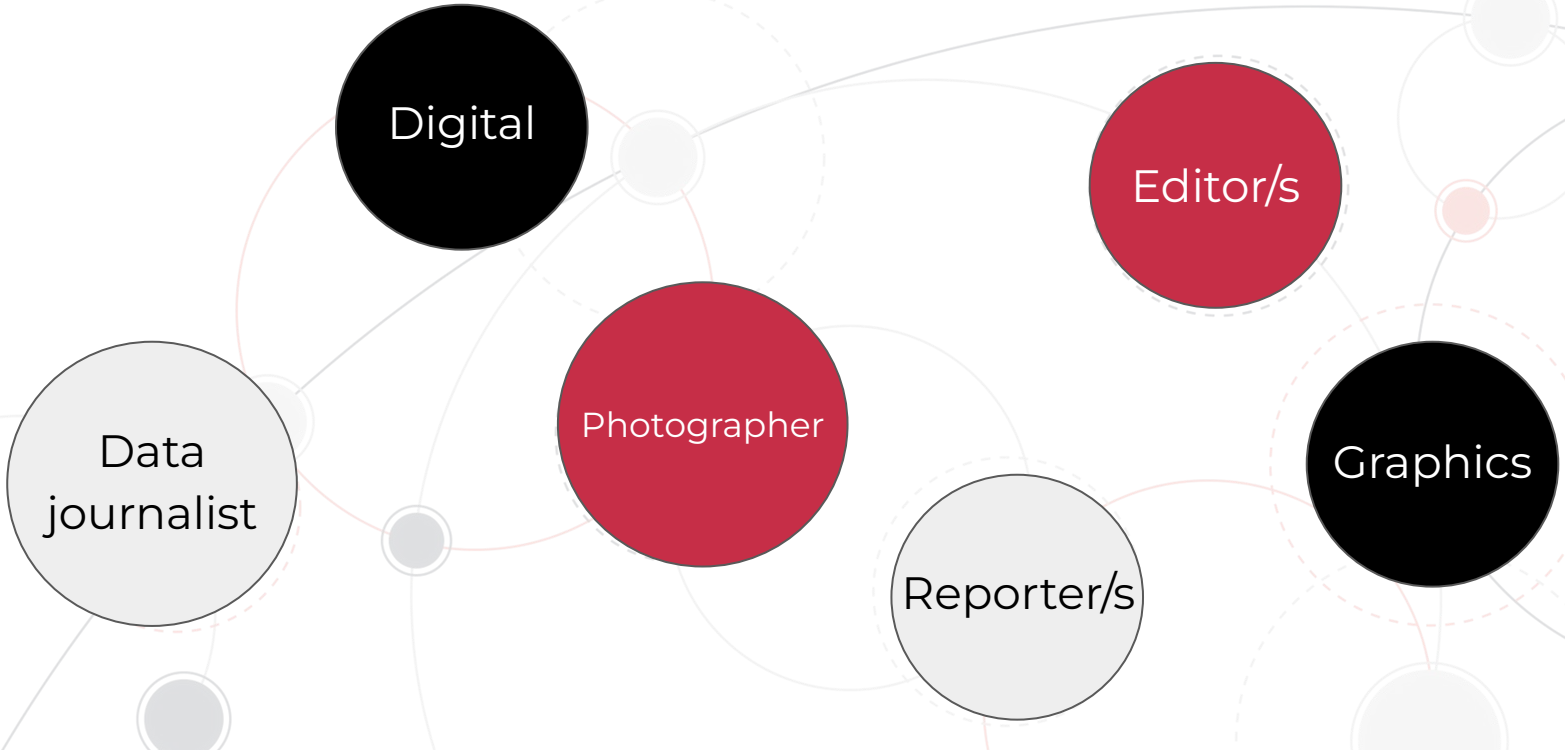
Let’s collaborate!

Collaborating on data journalism projects

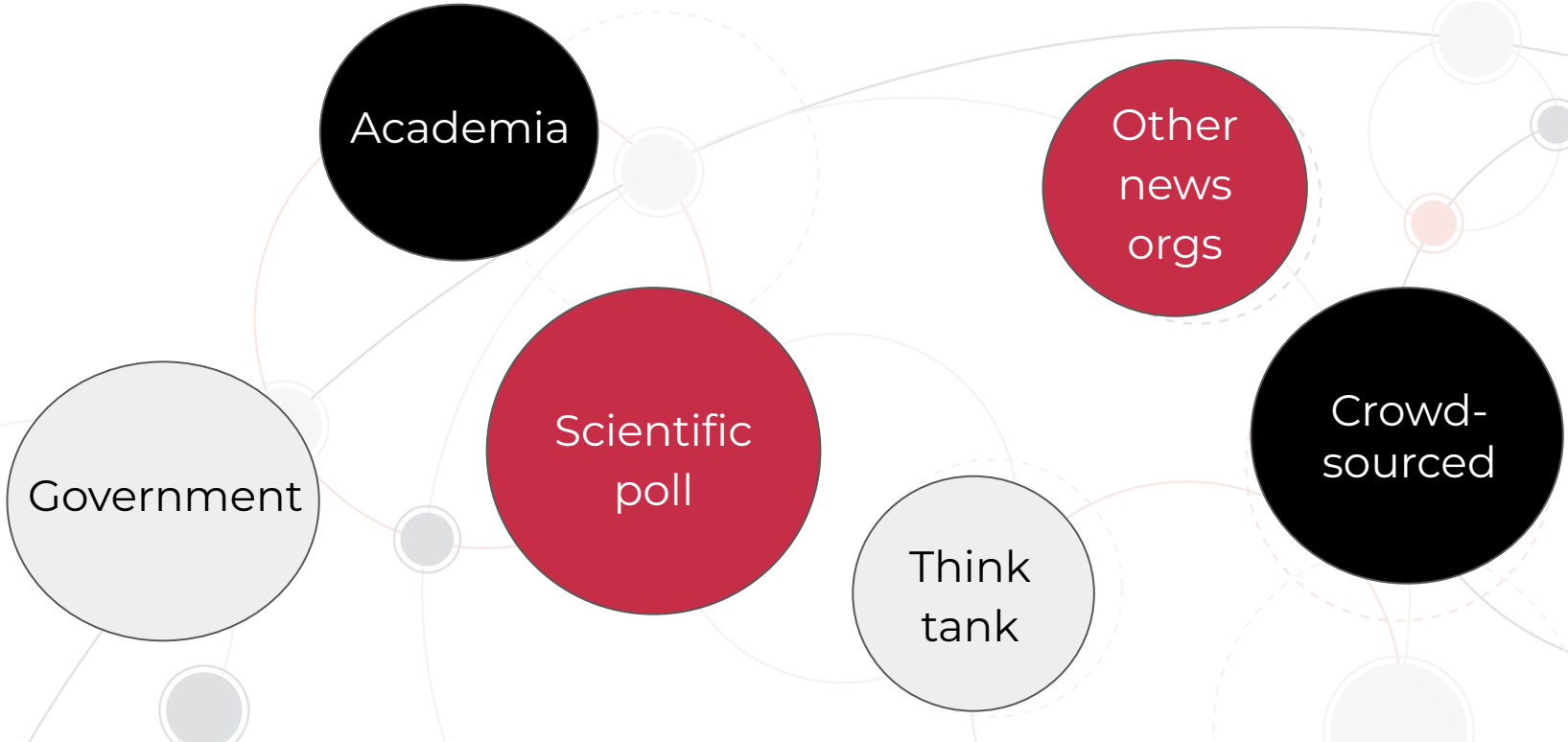
- ❑ Defining roles/tasks/duties
- ❑ Knowing the data
- ❑ Analyzing the data
- ❑ Documenting your analysis/transparency
- ❑ Reporting beyond the data
- ❑ Presentation (not my strong suit, as you'll see!)

A collaboration example

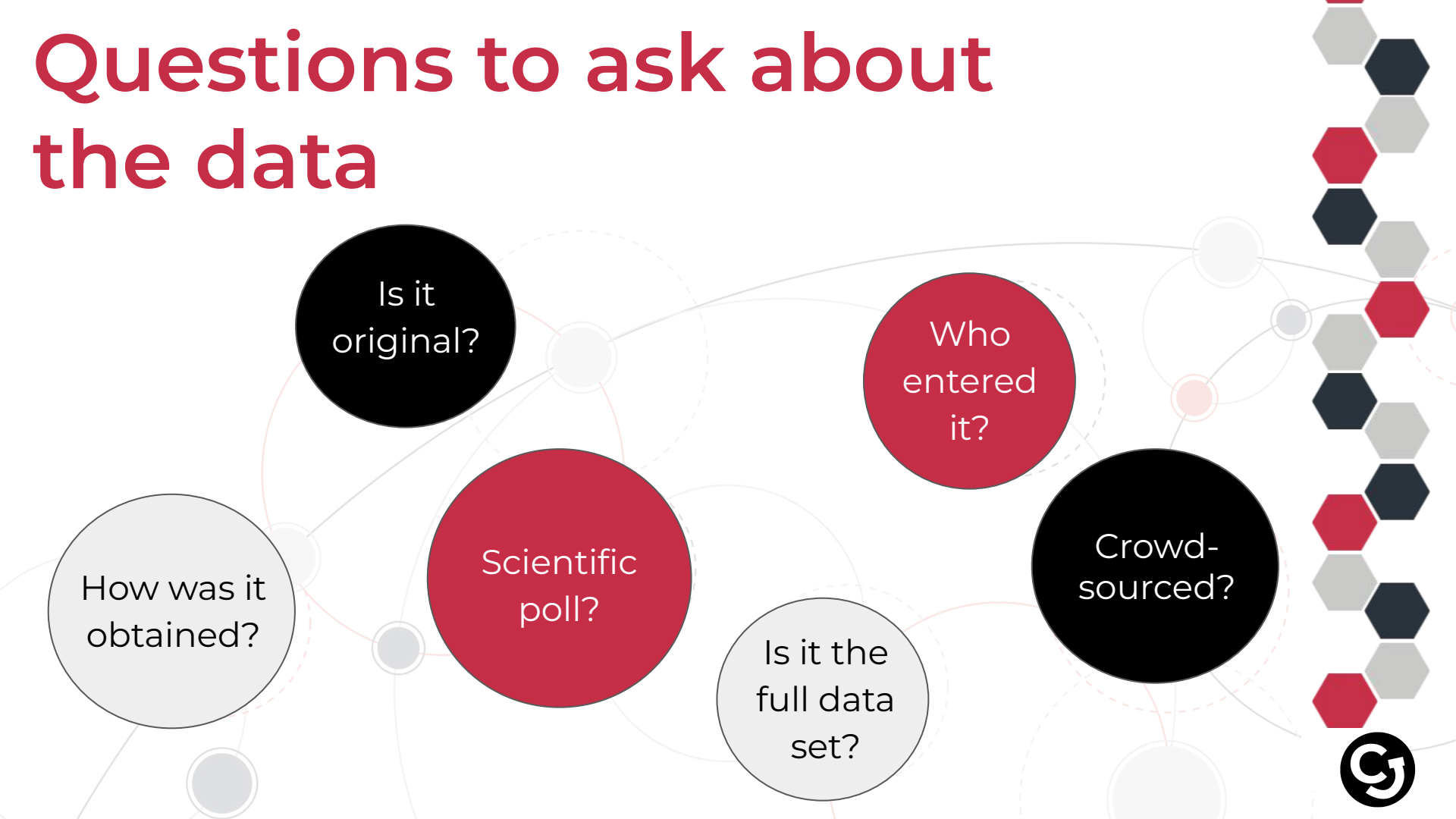
Who is collaborating? What are their roles?



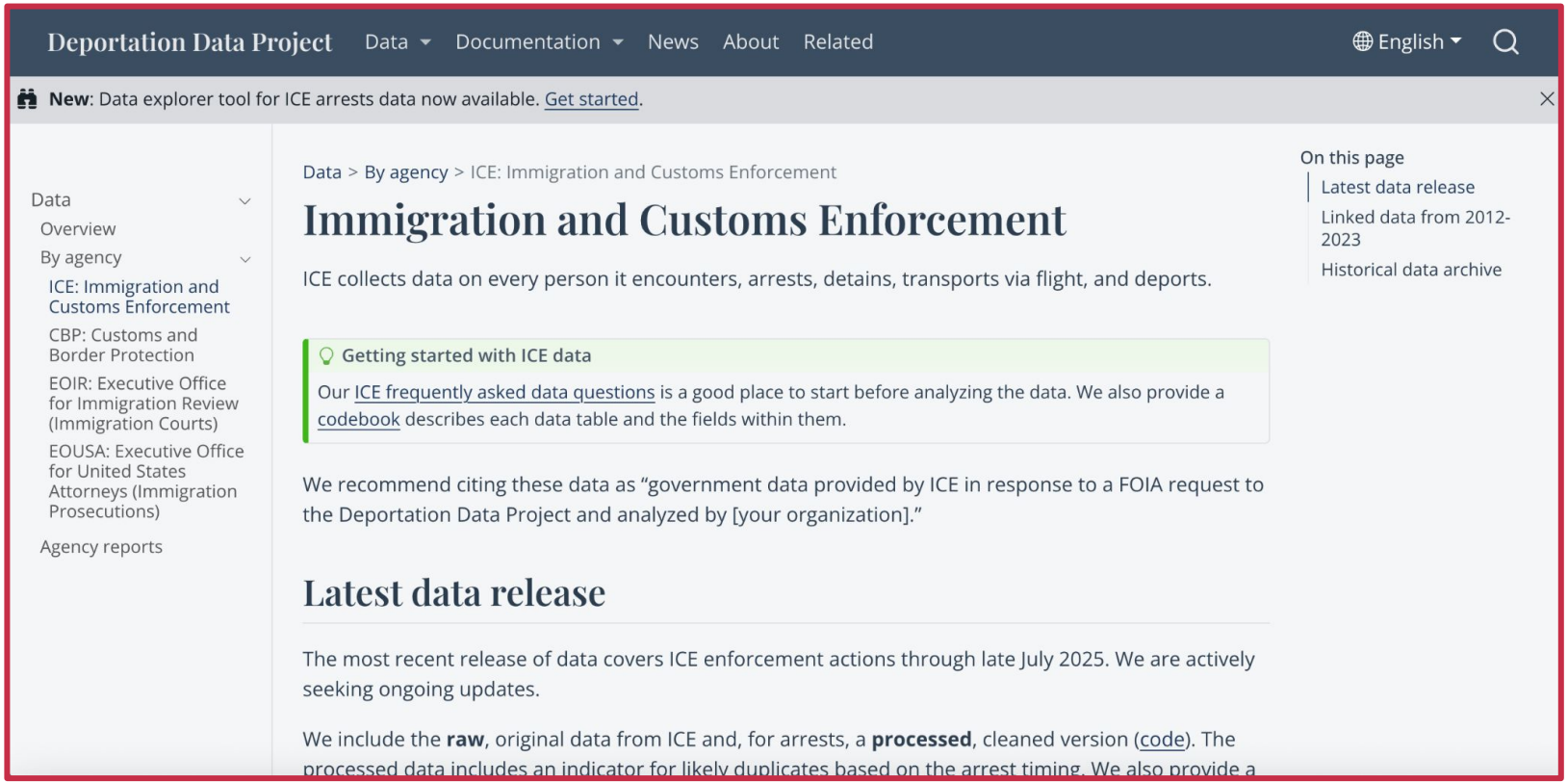
Where did the data come from?



Questions to ask about the data



Our data source — academic project via FOIA



Caveats galore!

⚠ We are not posting a removals dataset in this release because we found that it included far fewer records in 2025 than the previous release (in late June). We remain concerned that the late June removals table may not include all relevant records, and that associated fields in the other tables, such as the departure date, may also create an incomplete picture of removals. These concerns lead us to advise caution when performing any analysis of removals in any ICE data release in 2025. If you have insight into these problems or find other concerns, please let us know. We are actively seeking updated and corrected data.

- I noticed that there are some duplicates in the arrests data. How should I account for those?
- How can I identify courtroom arrests in the data?
- How can I identify raids in communities in the data, as opposed to arrests at check-ins or a jail or prison?
- How can I tell whether detainers are honored by local and state jails and prisons?
- Do these data include all immigration arrests, detentions, and removals by the US government?
- How can I identify removals to third countries?

How is the data being analyzed?

Judgment calls!

Keep a diary/record

What software is being used?

Does the data need cleaning?

Are there duplicates?

Excel issues...

Shit Management Says
@ShitMgmtSays

Important to highlight that Excel's row limit brought down Charlie Ja

cc: @ryxcommar

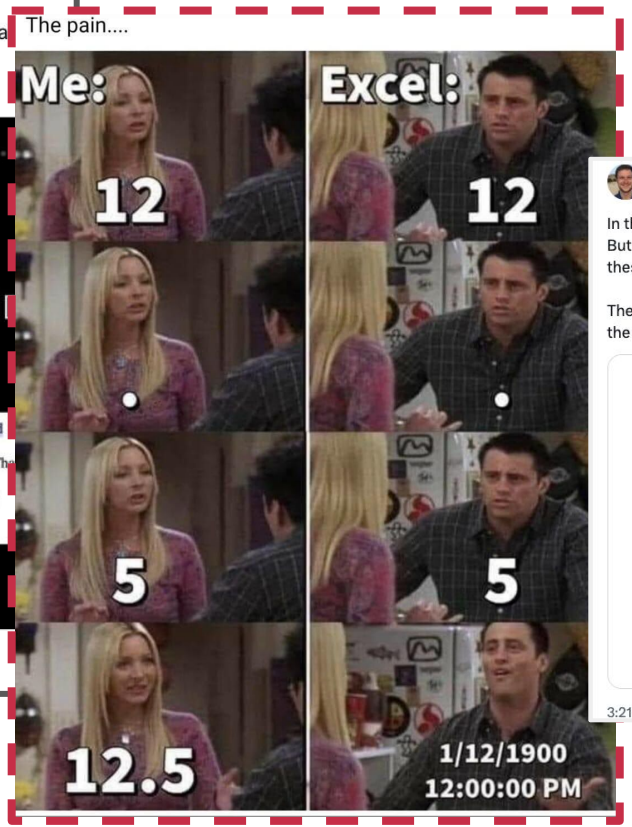
Senior PowerPoint Engineer
@ryxcommar

OH MY GOD. NO WAY. MICROSOFT EXCEL'S ROW LIMIT MADE THE FRAU UNRAVEL LMFAOOOOOOOoooo

173. JPMC first noticed irregularities with the list when a JPMC employee observed that the list contained exactly 1,048,576 rows, the maximum permitted by Microsoft Excel. Th number also did not correspond to any number of customer accounts previously identified by Javice or others at Frank for any given time period.

10:58 AM · 1/14/23

12:10 PM · Jan 14, 2023 · **868.5K** Views



Max Roser
@MaxCRoser

In the UK the number of cases rose rapidly.
But the public – and authorities – are only learning this now because these cases were only published now as a backlog.

The reason was apparently that the database is managed in Excel and the number of columns had reached the maximum.

A furious blame game is under way today after 16,000 coronavirus cases were missed due to a computer glitch - meaning thousands more potentially infected contacts were not traced.

The extraordinary meltdown is believed to have been caused by an Excel spreadsheet containing lab results reaching its maxium size, and failing to update.

Some 15,841 cases between September 25 and October 2 were not uploaded to the government dashboard.

As well as underestimating the scale of the outbreak in the UK, critically the details were not passed to contact tracers, meaning people exposed to the virus were not tracked down.

Work and Pensions Secretary Therese Coffey conceded this morning that more Britons 'may well' have been infected due to the blunder.

The shambolic situation sparked an immediate backlash against

3:21 AM · Oct 5, 2020

Replication is good... this is a Python/Pandas example

Launcher | ice_analysis_715.ipynb

Code | Python 3 (ipykernel)

[1]:
```
#imports
import pandas as pd
import numpy as np
```

[2]:
```
#making decimal points only two
pd.options.display.float_format = '{:20,.2f}'.format
```

The Data

This comes from the Deportation Data Project at the University of California Berkeley, School of Law.
https://deportationdata.org/data/ice.html

You'd want to download the data files into your Google drive to process them. i typically download the zip fil
computer, these are text files and i save them as .csv files, then upload to google.

Also, it appears the unique identifiers, which are key, change with each release.

Arrests

This is the main table we are working with to start. From the description: "Records every time ICE arrests sor
whether or not that arrest results in a decision to detain the person; or issues a Notice to Appear (NTA), the document
that starts a deportation case. Note that other agencies also issue NTAs, and those would not appear as arrests in this
table. We treat "apprehensions," "arrests," and "administrative arrests" as synonyms."

[4]:
```
#This gets arrests with Denver area of responsibility and/or the state of Colorado/Wyoming
co_arrests = arrests[
    (arrests['Apprehension AOR'] == 'Denver Area of Responsibility') |
    (arrests['Apprehension State'] == 'Colorado') | (arrests['Apprehension State'] == 'Wyoming')
]
co_arrests.info()
```

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 4582 entries, 34 to 265199
Data columns (total 23 columns):
 #   Column                  Non-Null Count  Dtype
---  ------                  --------------  -----
 0   Apprehension Date       4582 non-null   object
 1   Apprehension State      3562 non-null   object
 2   Apprehension County     0 non-null      float64
 3   Apprehension AOR        4582 non-null   object
 4   Final Program           4582 non-null   object
 5   Final Program Group     4582 non-null   object
 6   Apprehension Method     4582 non-null   object
 7   Apprehension Criminality 4582 non-null  object
 8   Case Status             4479 non-null   object
 9   Case Category           4479 non-null   object
 10  Departed Date           2662 non-null   object
 11  Departure Country       2662 non-null   object
 12  Final Order Yes No      4479 non-null   object
 13  Final Order Date        3126 non-null   object
 14  Birth Date              4582 non-null   object
 15  Birth Year              4582 non-null   int64
 16  Citizenship Country     4582 non-null   object
```

Dr Kareem Carr, Statistics Person
@kareem_carr

lawyer: and what percent of time did you spend fitting the model?
me, a data scientist: 2%
lawyer: and the other 98%?!
me: cleaning the data
lawyer: CLEANING THE DATA!!!
judge: *gasp*
jury: *gasp*
audience:*gasp*
me: Well ya know what guys, I'm not super happy about it either!

4:53 PM · Aug 2, 2019

42 589 3.4K 57

A couple of examples …

87059
2625 SAN PABLO ST NE,ALBUQUERQUE
2625 San Pablo St NEAlbuquerque
Abq
Abuquerque
Al;buquerque
ALB
Albquerque
Albququerque
Albrightsville
Albu
Albuqeruque
Albuqueque
Albuquerque
Albuquerque (Bernalillo)
ALBUQUERQUE NM 87106
Albuquerque,
Albuqueruque
Alburtis
ALBUUQUERQUE
T Or C
Truth or Consequences
Truth or Consequesces

ARMED SERVICES - PACIFIC	1
CALIFORNIA	21
COLORADO	3154
FLORIDA	2
LOUISIANA	1
MARYLAND	1
MASSACHUSETTS	1
UTAH	1
WYOMING	380
(blank)	999

Asking others for review …

Debunking your favorite NFL officiating conspiracy theories

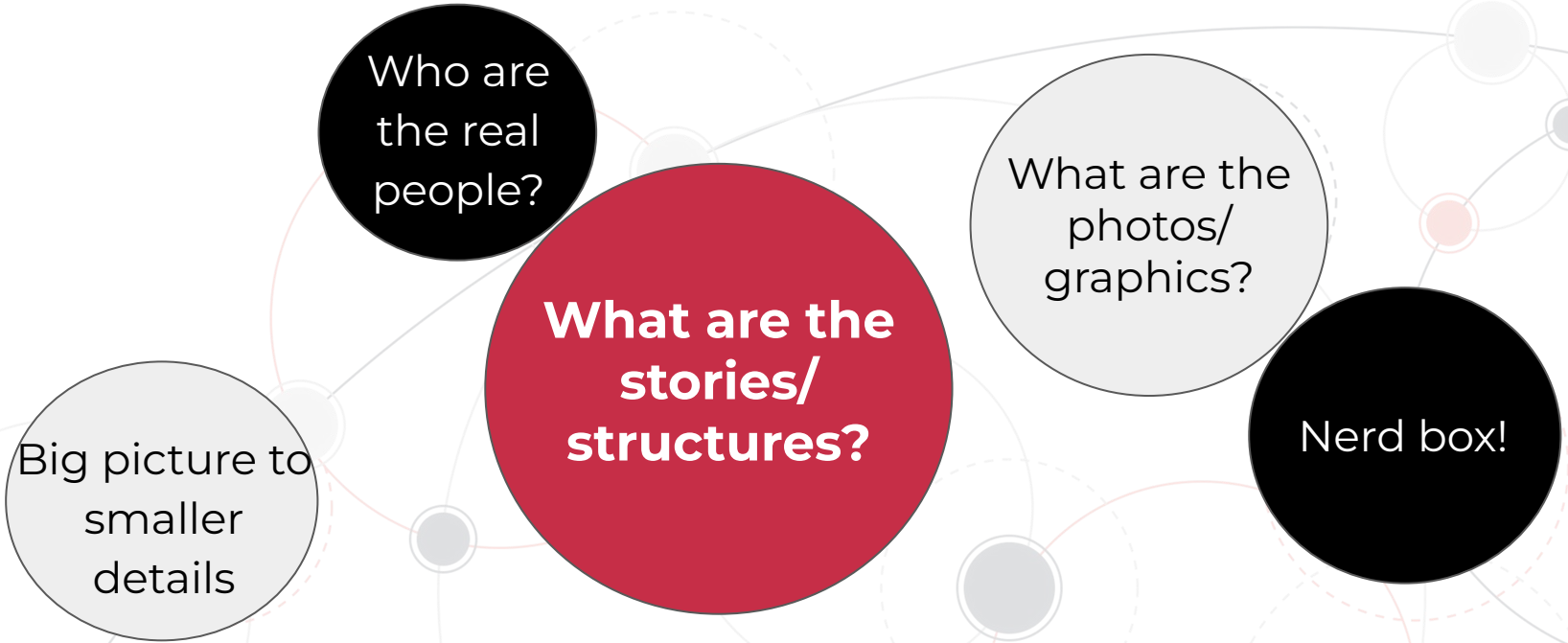
Throw the flag, coach! How do Andy Reid and Nick
challenges? Illustration by ESPN

Paula Lavigne and Shwetha Surendran

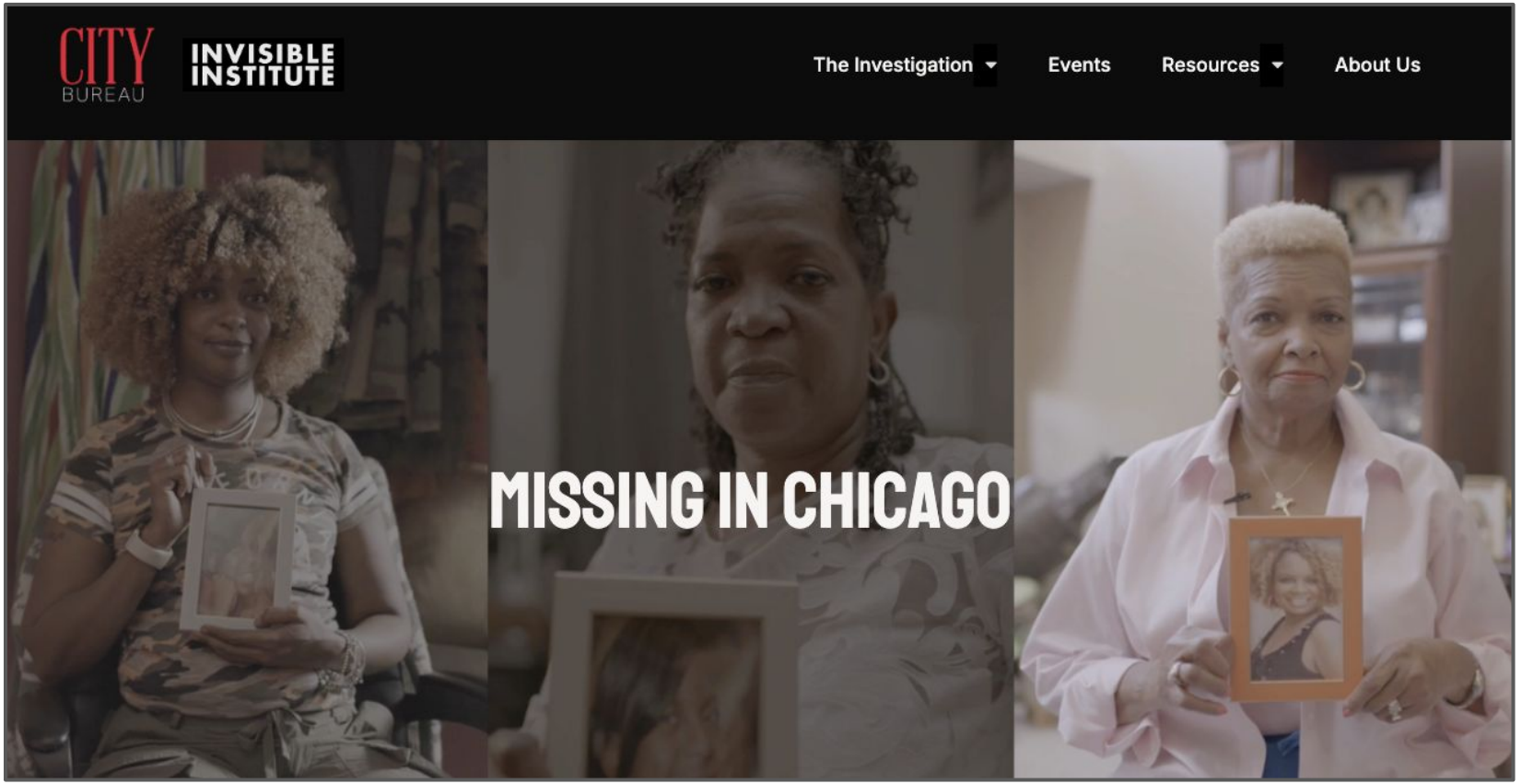
Aug 29, 2025, 07:11 AM ET

To assess whether officiating is routinely impacting game outcomes, ESPN compiled every penalty that referees called in regular-season and postseason games between 2001 and 2024, the type of penalty assessed, when it occurred during the game, and which team went on to win. ESPN also compiled every challenge a head coach made in the regular season from 2010 to 2024 as noted in official NFL Game Books. Dr. James J. Cochran, professor of applied statistics in the Culverhouse College of Business at the University of Alabama, helped ESPN run an analysis.

What story does the data tell?



A great digital example ...



One initial story, with follow ups

IMMIGRATION

What happens when ICE takes away a Colorado family? A teammate disappears. A colleague misses work. Neighbors are gone.

Taylor Dolven *4:25 AM MDT on Sep 5, 2025*

IMMIGRATION

Colorado towns are in the dark about whether ICE will reopen private prisons in their communities

Taylor Dolven *3:59 AM MDT on Jul 28, 2025*

POLITICS AND GOVERNMENT

Chaffee County commission urged to do more to protect immigrants from ICE arrest

Taylor Dolven *3:50 PM MDT on Sep 9, 2025*

POLITICS AND GOVERNMENT

Colorado governor won't say if state handed over residents' personal information to ICE

Taylor Dolven *3:25 AM MDT on Jul 22, 2025*

New Wyoming law asks law enforcement to act like ICE, generating concern over constitutional violations

by **Andrew Graham** July 23, 2025

Detaining Jackson Hole immigrants, ICE agents target western Wyoming

by **Jeannette Boner, Jackson Hole News&Guide, Andrew Graham, Cali O'Hare, The Pinedale Roundup** and **Joy Ufford** July 30, 2025

IMMIGRATION

Colorado gave records to ICE four times since February in response to subpoenas, including once by mistake

Taylor Dolven *1:23 PM MDT on Jul 22, 2025*

Wyoming Highway Patrol chief describes limited rollout of ICE agreement. Advocate warns it could undermine public safety.

by **Andrew Graham** July 31, 2025

Gordon to deploy Wyoming Nat Guard to support ICE in Trump's mass deportation effort

by **Andrew Graham** August 4, 2025

Without a country to take him, Cuban immigrant spends month in limbo inside Wyoming jail

by **Andrew Graham** August 29, 2025

POLITICS AND GOVERNMENT

Colorado agencies have received at least 9 subpoenas from ICE since Trump took office in January

Taylor Dolven *6:04 PM MDT on Jul 18, 2025*

Nerd Box! Transparency! How we did this:

Methods:
This story is based on data from the Deportation Data Center at the University of California, Berkeley School of Law.

Here's how we performed the analysis:

- Saved the tables as comma separated value (CSV) files.
- Imported them into a Jupyter notebook using the Python library Pandas.
- Narrowed the tables to data that included the Denver area of responsibility, Colorado or Wyoming.
- In the case of detention data, a list of detention facilities in Colorado and Wyoming were used.
- Using unique identifiers to match with detention data, added the most serious conviction charge to the arrest data.
- Using unique identifiers to match with detainer data, added conviction dates and additional state data to the arrest data.
- Narrowed the arrest data to the time frame of Jan. 20 to June 26 for both 2024 and 2025.
- Removed 41 records where the state could not be determined.
- Removed 28 duplicate records. Some duplicates in which individuals were arrested in both 2024 and 2025 remain in the data.
- Removed 25 records that appear to have come from other states.
- Used Excel pivot tables to analyze the remaining 2,718 records.

Some arrested people who were designated as having pending criminal charges at the time of their arrest had a conviction associated with them in the detention dataset. The Colorado Sun and WyoFile included those people, 17 in total, as having pending charges, so it is possible that our total number of convicted arrestees is an undercount.

It is worth noting that there is no way to gauge the accuracy of this data. For instance, three of the records removed listed locations in Colorado or Wyoming, but apprehension states of Florida and Utah.

WyoFile and The Colorado Sun are publishing a spreadsheet containing the data analyzed here.

Questions?

Contact me: fishnette@gmail.com