Data Munging for Justice

Helping public defenders and investigative journalists identify wrongful arrest and police misconduct

Rachel Warren

rwarren2@uci.edu
@warre_n_peace

PHD Student: UC Irvine Bren School of Informatics

Formerly MIMS at Berkeley School of Information

Formerly Machine Learning Engineer & Data Scientists at Salesforce and others

'22 EPIC Research Fellow

Meaningful transparency is somewhere in here …

Road Map:

Case Study 1: Interview study of public defenders

1. Background on public defenders & data in the criminal justice system
2. Public defenders challenges with surveillance data
3. Opportunities to build tech for public defenders
4. Larger themes

Case Study 2: Building technology for investigative journalists

Trial by File Formats

Exploring **public defenders'** technical needs in working with novel surveillance data

RESEARCH-ARTICLE OPEN ACCESS

Trial by File Formats: Exploring Public Defenders' Challenges Working with Novel Surveillance Data

Authors: Rachel B. Warren, Niloufar Salehi Authors Info & Claims

Proceedings of the ACM on Human-Computer Interaction, Volume 6, Issue CSCW1 • April 2022 • Article No.: 67, pp 1–26 • https://doi.org/10.1145/3512914

Online: 07 April 2022 Publication History

Abstract

In the United States, public defenders (lawyers assigned to people accused of crimes who cannot afford a private attorney) serve as an essential bulwark against wrongful arrest and incarceration for low-income and marginalized people. Public defenders have long been overworked and under-

Collaborators

Niloufar Salehi: Assistant Professor, UC Berkeley School of Information

Tiffany Pham, Jyen Yiee Wong, Sneha Chowdhury

Funding From Center for Technology, Society, and Policy (CTSP)

Road Map

Case Study 1: Interview study of public defenders

1. **Background on public defenders & surveillance data in the criminal justice system**
2. Public defenders' challenges with surveillance data
3. Opportunities to build for public defenders
4. Larger themes

Case Study 2: Building technology for investigative journalists

Interview Study

22 semi-structured interviews with members of the U.S. public defense community

Felony Public Defenders (including capital public defenders) & Misdemeanor Public Defenders

Federal Public Defender & Local Public Defenders

Paralegals working in Public Defense Offices

Tech employees responding to subpoena requests

7 States

Research Questions

How does an increase in digital surveillance data impact the quality of representation that public defenders feel they can provide their clients?

What specific barriers complicate public defenders' ability to use or refute surveillance data in court?

Public defenders (PDs) are **public employees** who represent **poor people** charged with **crimes**

What do we know about public defenders?

Highly utilized: 80% of people charged with felonies*

Underfunded: 72% work over 150 felony or 400 misdemeanor cases a year*

Heterogeneously administered

* 2010 Census of the Public Defender, Bureau of Justice Statistics (BJS).

Highly utilized and under resourced

80% of Americans charged with felonies qualify as indigent (i.e., poor) [1]

<2% of the $295 billion/year spent on criminal justice in the U.S. goes to PDs [2]

72% of PDs work more than the legal limit of 400 misdemeanor cases or 150 felony cases per year [2]

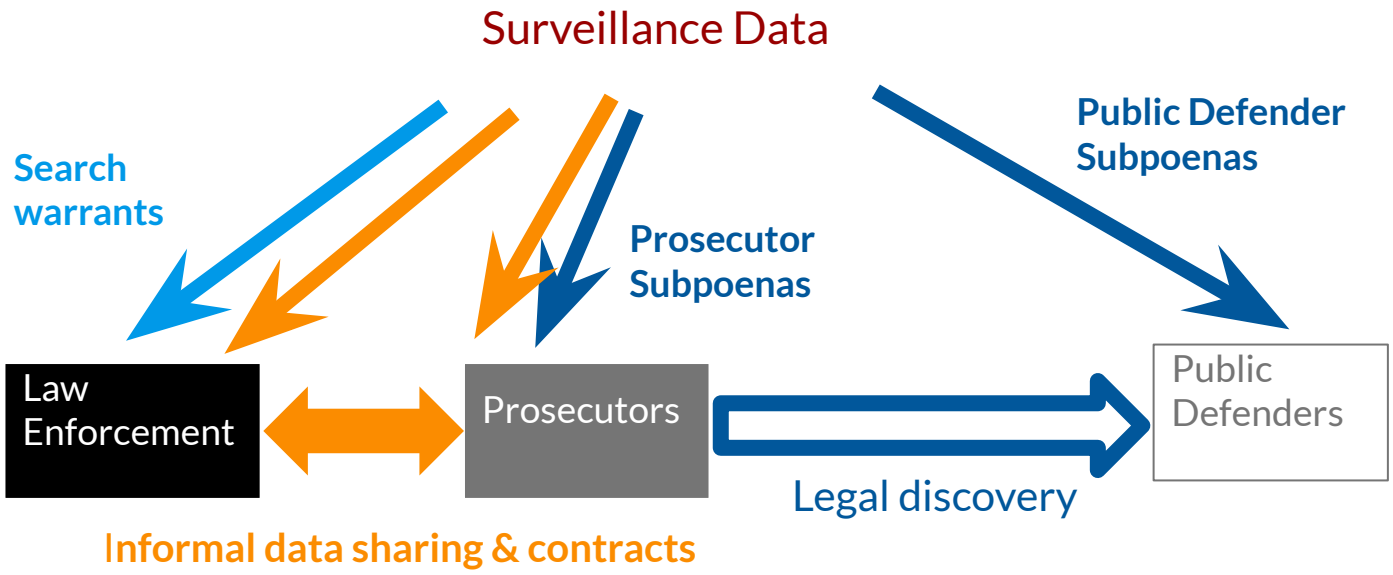
1.Harlow. 2000. Defense Counsel in Criminal Cases. Technical Report. Bureau of Justice Statistics (BJS).
2. 2010 Census of the Public Defender, Bureau of Justice Statistics (BJS).

What do we know about surveillance data in the criminal justice system?

There is more and more of it!

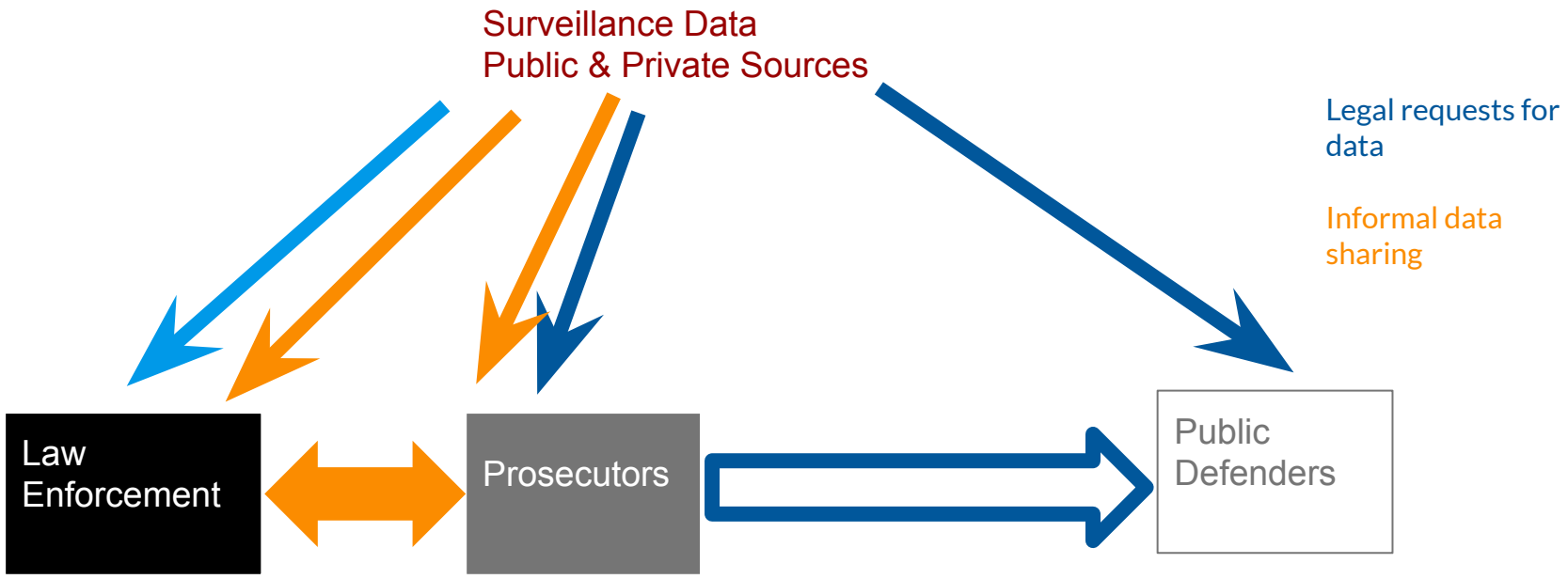
Public surveillance systems: license plate readers, body camera footage, public CCTV cameras, drones, surveillance in jails & public housing

Data from private sources: call detail (CDR) records, social media feeds, GPS data, search history

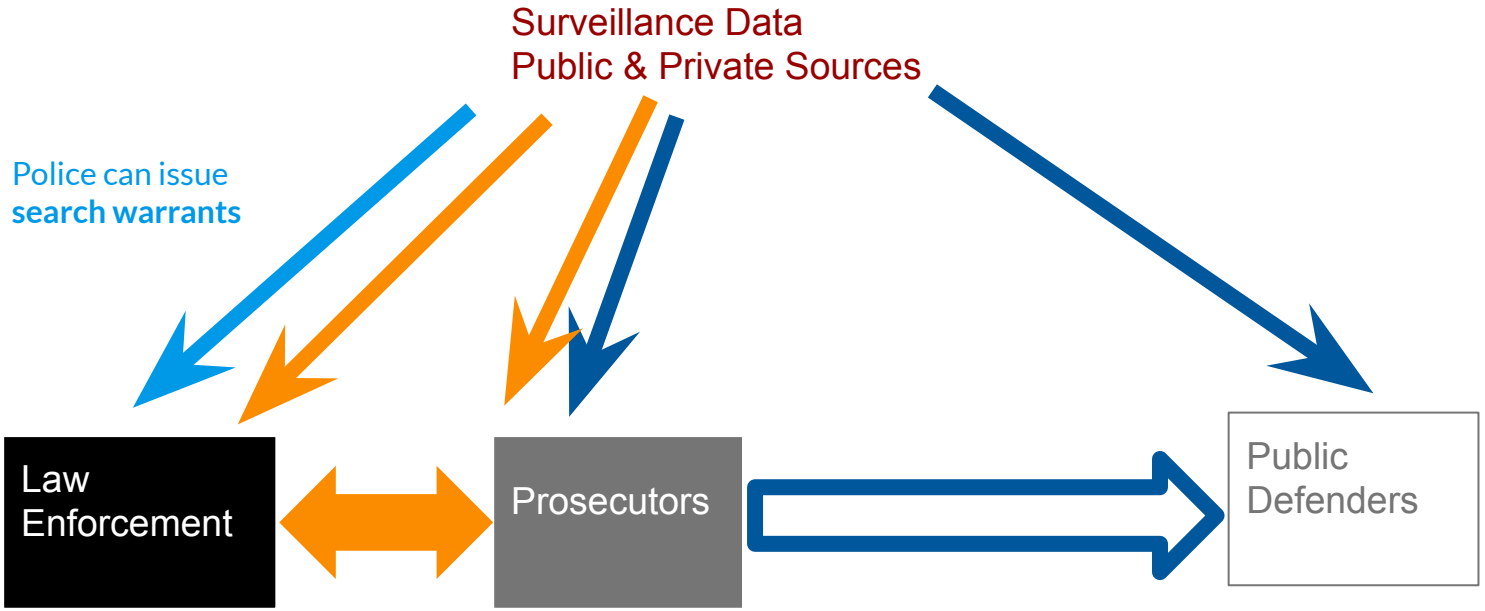


Prosecutors get data through both subpoenas and informal relationships with data brokers and law enforcement.

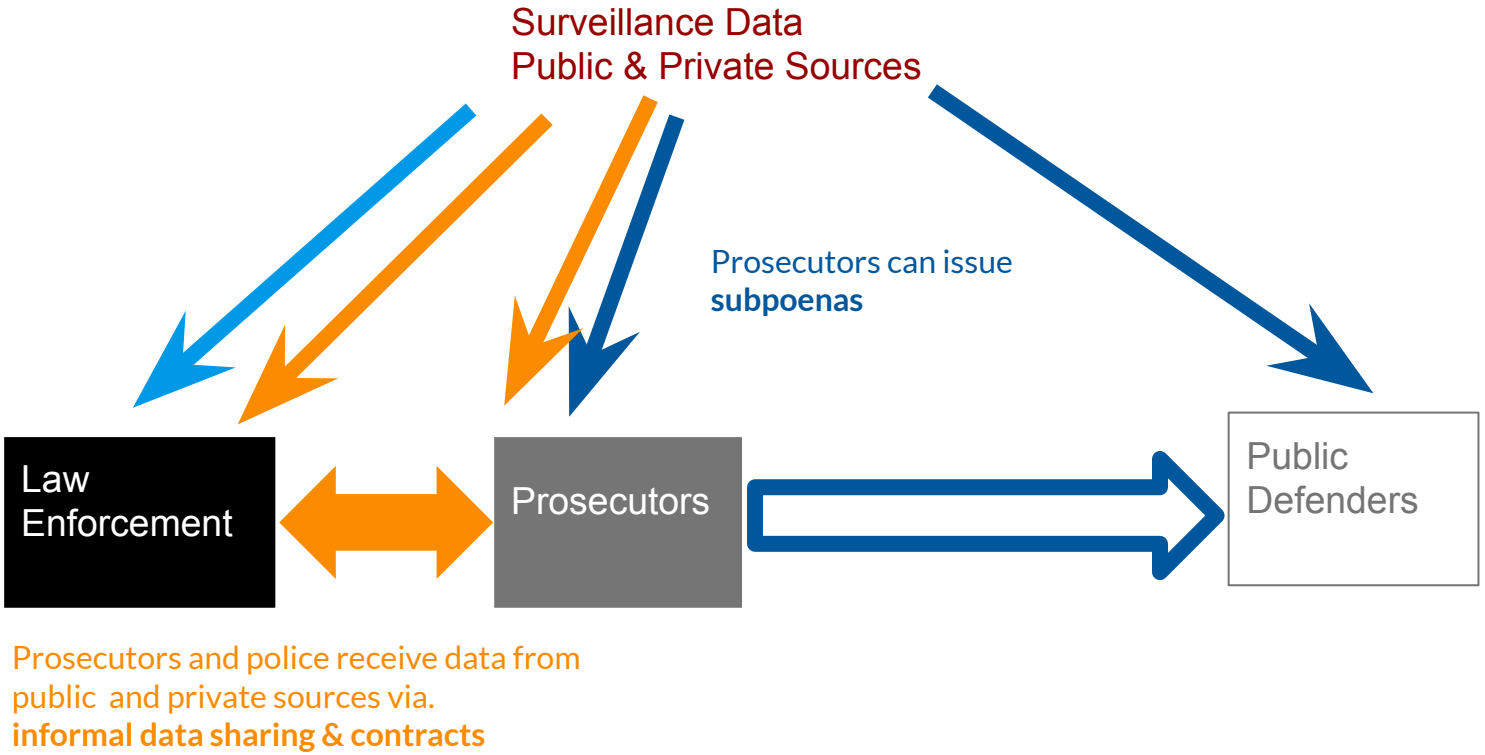
Public defenders get data through discovery from prosecutors or subpoenas from private companies



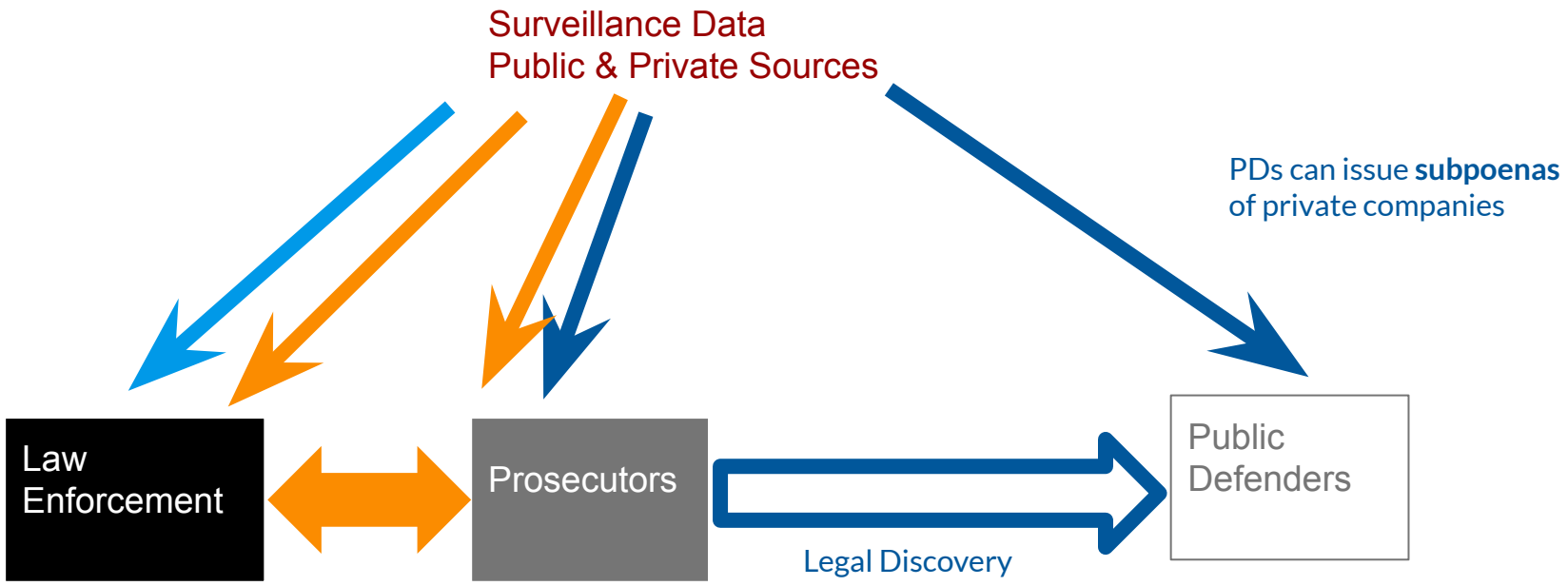
How data moves through the criminal justice system



How data moves through the criminal justice system



How surveillance data moves through the criminal justice system



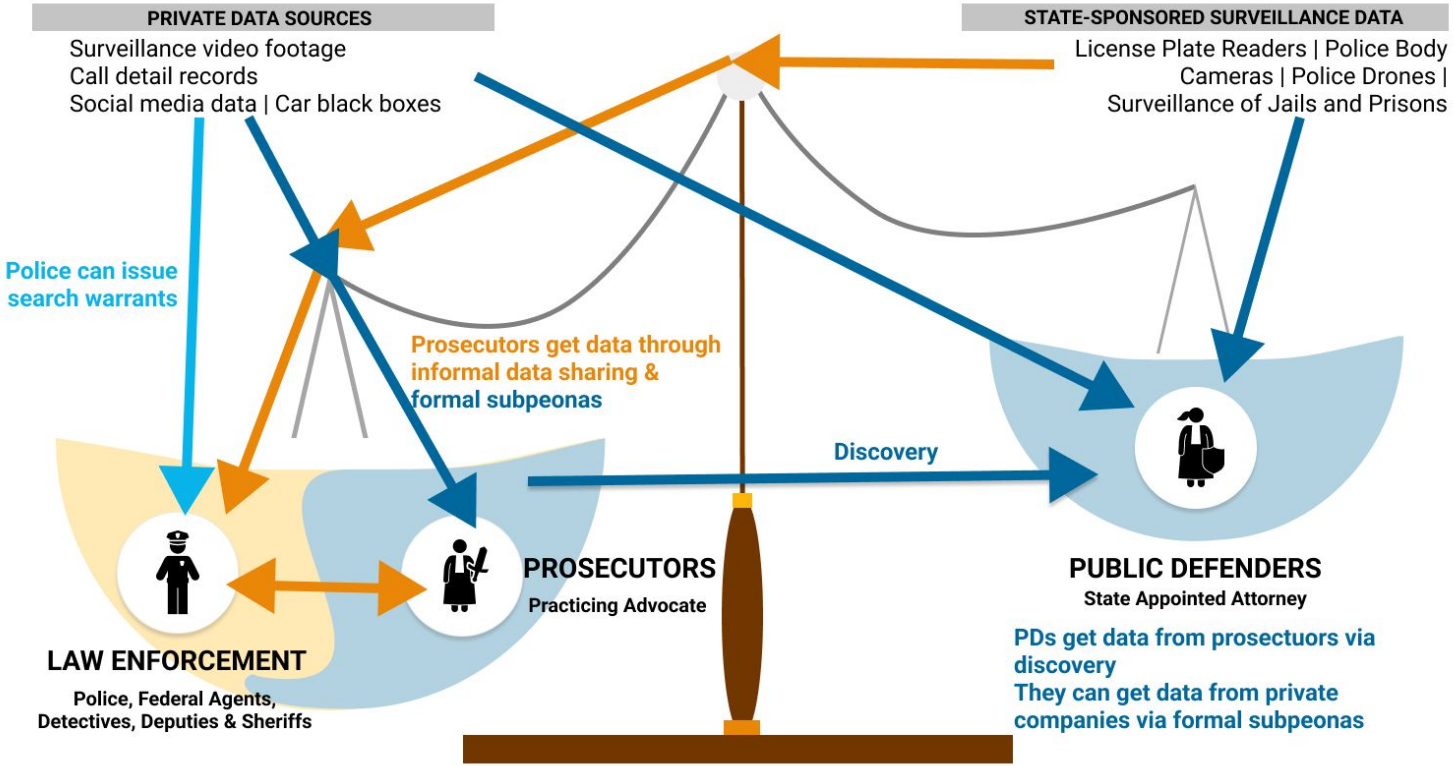
How surveillance data moves through the criminal justice system

Why study public defenders?

Prevent wrongful incarceration due to surveillance

Legal requests for data

Informal data sharing



How data moves through the criminal justice system

—

How does an increase in digital surveillance data impact the quality of representation that PDS feel they can provide their clients?

Road Map

Case Study 1: Interview study of public defenders

1. Background on public defenders & data in the criminal justice system
2. Public defenders' challenges with surveillance data
3. Opportunities to build tech for public defenders
4. Larger themes

Case Study 2: Building technology for investigative journalists

Research Questions

1. What kinds of data do PDs receive in discovery?
2. How do PDs use or examine surveillance data?
3. What organizational and technical factors complicate PDs' ability to review surveillance data in order to defend their clients?

Interview Study

22 semi-structured interviews with members of the U.S. public defense community

Federal & misdemeanor PDs

Federal & state PDs

Investigators in PD offices

Paralegal in PD offices

PDs in 7 states

People who respond to subpoena requests at tech companies

Findings

1. Public defenders **lack the time and resources** to adequately process this surveillance data
2. Public defenders are **structurally disadvantaged** by the way that data moves through the criminal justice system
3. In some instances, **privacy laws actually disadvantage** public defenders

Public defenders are overwhelmed by the volume of surveillance data

1. Felony cases can include **100 + hours of video**
2. Videos in dozens of **file formats**
3. Social media feeds as **10k+ page PDFS**
4. Lack of access to basic technology such as **in-office wifi, working laptops**
5. Lack of access to more sophisticated technology, **money to hire experts**

Public defenders are overwhelmed by the volume of surveillance data

1. Felony cases can include **100 + hours of video**
2. Videos in dozens of **file formats**
3. Social media feeds as **10k+ page PDFS**
4. Lack of access to basic technology such as **in-office wifi, working laptops**
5. Lack of access to more sophisticated technology, **money to hire experts**

—

“We get a lot of cases where the feds have gotten a warrant to Instagram, and they will send us a [25,000 page PDF] which is not usable.”

- Public Defender

“I spend 3-7 hours a day just standardizing file formats”

- Paralegal in a Public Defense Office

Why does examining surveillance data matter?

PDs may be able to challenge a narrative put forward by police by reinterpreting or recontextualizing surveillance collected by prosecutors

Examples: surveillance video & text data

Why review video data?

Discrepancies in police reports

Identify new witnesses

Uncover civil rights violations and police misconduct

—

“It’s a significant amount of video [. . .] And you are required to watch it. It can break a case.”

- Public Defender

Barriers to reviewing video data

Body camera in short unlableled clips

Unplayable formats

Duplicates

No transcription software

AutoSave OFF

BWC

Home Insert Draw Page Layout Formulas Data Review View Developer Tell me

Share Comments

PERSONAL.XLSB!Macro1

Paste Cut Copy Format

Calibri (Body) 12

Wrap Text Merge & Center

General

Conditional Formatting Format as Table

Normal Bad Good Neutral Calculation Check Cell Explanatory T... Input Linked Cell Note

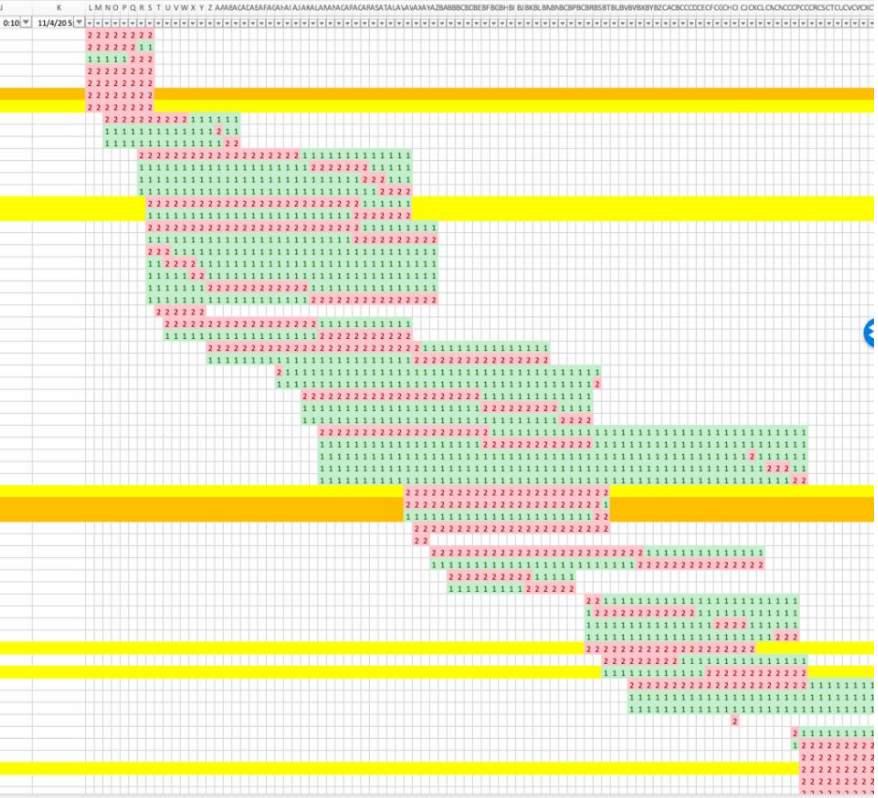
Insert Delete Format

AutoSum Fill Clear Sort & Filter Find & Select

Open recovered workbooks? Your recent changes were saved. Do you want to continue working where you left off? Yes No

A1 File name

File name	Officer	X No	Start time	Duration	PST start time	PST end time	Officer's first appearance	Officer's last appearance
[redacted]			11/4/20 13:21:39	1:10:03	11/4/20 5:21:39	11/4/20 6:31:42	11/4/20 5:21:39	11/4/20 6:31:42
			11/4/20 13:22:19	0:55:36	11/4/20 5:22:19	11/4/20 6:17:55	11/4/20 5:22:19	11/4/20 6:31:47
			11/4/20 14:18:12	0:13:35	11/4/20 6:18:12	11/4/20 6:31:47	11/4/20 5:22:19	11/4/20 6:31:47
			11/4/20 13:23:15	1:10:59	11/4/20 5:23:15	11/4/20 6:34:14	11/4/20 5:23:15	11/4/20 6:34:14
			11/4/20 13:23:50	1:10:13	11/4/20 5:23:50	11/4/20 6:34:03	11/4/20 5:23:50	11/4/20 6:34:03
			11/4/20 13:23:57	1:10:03	11/4/20 5:23:57	11/4/20 6:34:00	11/4/20 5:23:57	11/4/20 6:34:00
			11/4/20 13:23:59	1:12:02	11/4/20 5:23:59	11/4/20 6:36:01	11/4/20 5:23:59	11/4/20 6:36:01
			11/4/20 13:40:00	1:35:58	11/4/20 5:40:00	11/4/20 7:15:58	11/4/20 5:40:00	11/4/20 8:11:40
			11/4/20 15:52:25	0:06:59	11/4/20 7:52:25	11/4/20 7:59:24	11/4/20 5:40:00	11/4/20 8:11:40
			11/4/20 16:04:55	0:06:45	11/4/20 8:04:55	11/4/20 8:11:40	11/4/20 5:40:00	11/4/20 8:11:40
			11/4/20 14:25:49	3:00:59	11/4/20 6:25:49	11/4/20 9:26:48	11/4/20 6:25:49	11/4/20 11:39:08
			11/4/20 17:46:49	1:00:55	11/4/20 9:46:49	11/4/20 10:47:44	11/4/20 6:25:49	11/4/20 11:39:08
			11/4/20 18:48:29	0:11:41	11/4/20 10:48:29	11/4/20 11:00:10	11/4/20 6:25:49	11/4/20 11:39:08
			11/4/20 19:03:14	0:35:54	11/4/20 11:03:14	11/4/20 11:39:08	11/4/20 6:25:49	11/4/20 11:39:08
			11/4/20 14:30:58	4:00:02	11/4/20 6:30:58	11/4/20 10:31:00	11/4/20 6:30:58	11/4/20 11:39:22
			11/4/20 18:31:01	1:08:21	11/4/20 10:31:01	11/4/20 11:39:22	11/4/20 6:30:58	11/4/20 11:39:22
			11/4/20 14:30:58	4:00:02	11/4/20 6:30:58	11/4/20 10:31:00	11/4/20 6:30:58	11/4/20 12:02:20
			11/4/20 18:31:02	1:31:18	11/4/20 10:31:02	11/4/20 12:02:20	11/4/20 6:30:58	11/4/20 12:02:20
			11/4/20 14:32:06	0:26:31	11/4/20 6:32:06	11/4/20 6:58:37	11/4/20 6:32:06	11/4/20 12:09:47
			11/4/20 14:58:38	0:27:57	11/4/20 6:58:38	11/4/20 7:26:35	11/4/20 6:32:06	11/4/20 12:09:47
			11/4/20 15:29:20	0:10:31	11/4/20 7:29:20	11/4/20 7:39:51	11/4/20 6:32:06	11/4/20 12:09:47
			11/4/20 15:49:23	1:43:57	11/4/20 7:49:23	11/4/20 9:33:20	11/4/20 6:32:06	11/4/20 12:09:47
			11/4/20 17:42:41	2:27:06	11/4/20 9:42:41	11/4/20 12:09:47	11/4/20 6:32:06	11/4/20 12:09:47
			11/4/20 14:47:02	0:50:23	11/4/20 6:47:02	11/4/20 7:37:25	11/4/20 6:47:02	11/4/20 7:37:25
			11/4/20 14:50:28	2:58:12	11/4/20 6:50:28	11/4/20 9:48:40	11/4/20 6:50:28	11/4/20 11:30:23
			11/4/20 17:51:04	1:39:19	11/4/20 9:51:04	11/4/20 11:30:23	11/4/20 6:50:28	11/4/20 11:30:23
			11/4/20 15:42:11	4:00:02	11/4/20 7:42:11	11/4/20 11:42:13	11/4/20 7:42:11	11/4/20 14:17:05
			11/4/20 19:42:14	2:34:51	11/4/20 11:42:14	11/4/20 14:17:05	11/4/20 7:42:11	11/4/20 14:17:05
			11/4/20 17:05:26	0:04:15	11/4/20 9:05:26	11/4/20 9:09:41	11/4/20 9:05:26	11/4/20 15:19:50
			11/4/20 23:10:15	0:09:35	11/4/20 15:10:15	11/4/20 15:19:50	11/4/20 9:05:26	11/4/20 15:19:50
			11/4/20 17:33:14	3:23:58	11/4/20 9:33:14	11/4/20 12:57:12	11/4/20 9:33:14	11/4/20 15:07:43
			11/4/20 21:06:41	1:14:40	11/4/20 13:06:41	11/4/20 14:21:21	11/4/20 9:33:14	11/4/20 15:07:43
			11/4/20 22:30:38	0:37:05	11/4/20 14:30:38	11/4/20 15:07:43	11/4/20 9:33:14	11/4/20 15:07:43
			11/4/20 17:56:15	3:07:53	11/4/20 9:56:15	11/4/20 13:04:08	11/4/20 9:56:15	11/4/20 19:18:56
			11/4/20 21:04:09	2:05:15	11/4/20 13:04:09	11/4/20 15:09:24	11/4/20 9:56:15	11/4/20 19:18:56
			11/5/20 2:15:55	0:00:38	11/4/20 18:15:55	11/4/20 18:16:33	11/4/20 9:56:15	11/4/20 19:18:56
			11/5/20 2:37:26	0:21:32	11/4/20 18:37:26	11/4/20 18:58:58	11/4/20 9:56:15	11/4/20 19:18:56
			11/5/20 3:04:44	0:14:12	11/4/20 19:04:44	11/4/20 19:18:56	11/4/20 9:56:15	11/4/20 19:18:56
			11/4/20 19:36:50	3:49:32	11/4/20 11:36:50	11/4/20 15:26:22	11/4/20 11:36:50	11/4/20 15:26:22
			11/4/20 19:37:38	3:34:50	11/4/20 11:37:38	11/4/20 15:12:28	11/4/20 11:37:38	11/4/20 15:26:23
			11/4/20 23:18:14	0:08:09	11/4/20 15:18:14	11/4/20 15:26:23	11/4/20 11:37:38	11/4/20 15:26:23
			11/4/20 19:43:22	3:43:22	11/4/20 11:43:22	11/4/20 15:26:44	11/4/20 11:43:22	11/4/20 15:26:44
			11/4/20 19:49:27	0:02:50	11/4/20 11:49:27	11/4/20 11:52:17	11/4/20 11:49:27	11/4/20 11:52:17
			11/4/20 20:00:43	4:00:02	11/4/20 12:00:43	11/4/20 16:00:45	11/4/20 12:00:43	11/4/20 18:24:27
			11/5/20 0:00:46	2:23:41	11/4/20 16:00:46	11/4/20 18:24:27	11/4/20 12:00:43	11/4/20 18:24:27
			11/4/20 20:24:26	1:26:05	11/4/20 12:24:26	11/4/20 13:50:31	11/4/20 12:24:26	11/4/20 14:40:52
			11/4/20 21:57:35	0:43:17	11/4/20 13:57:35	11/4/20 14:40:52	11/4/20 12:24:26	11/4/20 14:40:52
			11/4/20 23:07:19	0:02:53	11/4/20 15:07:19	11/4/20 15:10:12	11/4/20 15:07:19	11/4/20 19:01:34
			11/4/20 23:17:48	1:44:31	11/4/20 15:17:48	11/4/20 17:02:19	11/4/20 15:07:19	11/4/20 19:01:34
			11/5/20 1:31:47	0:36:20	11/4/20 17:31:47	11/4/20 18:08:07	11/4/20 15:07:19	11/4/20 19:01:34
			11/5/20 2:47:42	0:13:52	11/4/20 18:47:42	11/4/20 19:01:34	11/4/20 15:07:19	11/4/20 19:01:34
			11/4/20 23:09:08	3:10:00	11/4/20 15:09:08	11/4/20 18:19:08	11/4/20 15:09:08	11/4/20 18:19:08
			11/4/20 23:26:41	1:20:11	11/4/20 15:26:41	11/4/20 16:46:52	11/4/20 15:26:41	11/4/20 19:18:07
			11/5/20 1:28:00	1:50:07	11/4/20 17:28:00	11/4/20 19:18:07	11/4/20 15:26:41	11/4/20 19:18:07
			11/4/20 23:52:55	3:25:52	11/4/20 15:52:55	11/4/20 19:18:47	11/4/20 15:52:55	11/4/20 23:22:43
			11/5/20 7:17:27	0:00:34	11/4/20 23:17:27	11/4/20 23:18:01	11/4/20 15:52:55	11/4/20 23:22:43
			11/5/20 7:22:10	0:00:33	11/4/20 23:22:10	11/4/20 23:22:43	11/4/20 15:52:55	11/4/20 23:22:43
			11/5/20 1:52:20	0:06:49	11/4/20 17:52:20	11/4/20 17:59:09	11/4/20 17:52:20	11/4/20 17:59:09
			11/5/20 3:02:14	0:05:39	11/4/20 19:02:14	11/4/20 19:07:53	11/4/20 19:02:14	11/4/20 21:40:01
			11/5/20 3:18:18	2:21:43	11/4/20 19:18:18	11/4/20 21:40:01	11/4/20 19:02:14	11/4/20 21:40:01
			11/5/20 3:18:46	2:22:25	11/4/20 19:18:46	11/4/20 21:41:11	11/4/20 19:18:46	11/4/20 21:41:11
			11/5/20 3:18:48	2:21:41	11/4/20 19:18:48	11/4/20 21:40:29	11/4/20 19:18:48	11/4/20 21:40:29
			11/5/20 3:18:54	2:22:47	11/4/20 19:18:54	11/4/20 21:41:41	11/4/20 19:18:54	11/4/20 21:41:41



Sheet1

87%

Unstructured PDFs

Social media feeds (often includes every transaction)

Text histories

Device forensics: PDF of everything on a desktop

Why do PDs need to review social media or communications data?

Alibis, identifying new witnesses

Putting statements in context

→ 4 interview participants talked about battles over the meaning of emojis

Police (esp. gang task forces) are known to monitor social media and use it to open investigations (Patton 2017)

Protecting my family 💯🔫🔥🔫

Do these emojis mean the defendant was declaring they were “armed 100% of the time” ….
or did he/she just use these emojis for emphasis?

See me tonight and we can square up

Does a square mean the bulge of a gun in a waistband
… or just to settle a bill?

Public defenders also have a harder time getting data

Privacy asymmetries in the law

Lack of informal relationship with private companies / public surveillance systems

Lack of ability to enforce subpoenas

“It’s not privacy laws or technical hurdles . . . It’s Facebook being dicks . . . they will provide all of this information to law enforcement without a warrant, but **they will not respond to our subpoenas very often**”

- Public Defender

PSA for Technologists!

Make a plan for responding to warrant and subpoena requests in a fair way that is easy to interpret for both law enforcement and public defenders!

Road Map

Case Study 1: Interview study of public defenders

1. Background on public defenders & data in the criminal justice system
2. Public defenders' challenges with surveillance data
3. Opportunities to build for public defenders
4. Larger themes

Case Study 2: Building technology for investigative journalists

Opportunities to Build Tools for Public Defenders

- Video: automatic video sorting or labeling & transcription software
- PDF processing: better search, extracting posts or content in a date range or from a particular person
- Social tools so PDs can compare experts, science resources

Constraints

- Low technical literacy and access to technical resources
- Offices are very different
- Legal & organizational regulations
- “Automatic” or ML processing is hard to explain in court
- Adversarial environment

Ideal Technical Partnerships with Public Defenders

- Work directly with public defenders
- Keep tools simple, avoid over reliance on existing storage systems
- Don’t rule out policy solutions!

—

“I feel [with some of the technical solutions] it’d be great to have a system where people are texted before their court date, but even better would be a system where 90 percent of the court dates don’t happen because they’re totally useless”

- Public Defender

Road Map

Case Study 1: Interview study of public defenders

1. Background on public defenders & data in the criminal justice system
2. PDs' challenges with surveillance data
3. Opportunities to build for public defenders
4. Larger themes

Case Study 2: Building technology for investigative journalists

Policy Solutions are also critical

1. Reform privacy asymmetries (hopefully by limiting data to law enforcement rather than expanding data for everyone) [Wexler 2019, Wexler 2017]
2. Transparency in police technology acquisition and oversight into police technology budgets [Joh 2017]
3. Reduce bureaucracy in the criminal justice system

Broader Implications

1. Tools of interpretation are required for meaningful transparency
2. What data attorneys can legal access > practical technical and social reality of what data they can meaningfully use
3. Justice is not only about more or less access, but also about *parity* in access

—

The most oppressive regime is not necessarily the one with the most information, but one that has a monopoly on constructing the narratives from that information that will define its citizens' freedoms and choices.

Part 2: Investigative Journalists

Searching for patterns &
indexing unstructured electronic data

Collaborators

Niloufar Salehi: Assistant Professor UCB School of Information

Aditya Parameswaran: Associate Professor UCB EECS

Lisa Pickoff-White: Reporter KQED & California Reporting Project

Road Map

Case Study 1: Interview study of public defenders

Case Study 2: Building technology for investigative journalists

1. Background: California Reporting Project and public record requests
2. Needfinding & prototype
3. Evaluating and using the prototype
4. Next steps and larger themes

California Reporting Project

Coordinated effort to gather data on police misconduct

2018 California Passes SB 1421 “Right to Know Act”

Allows citizens to submit FOIA request from every law enforcement agency in California about any incident of police use of force or sustained misconduct

California Reporting Project

Coordinated effort to gather data on police misconduct

100 + public record requests from different police departments

Data has audio & video

Different formats and organizations from different police departments

Public Record Requests

- Very similar problems to subpoena responses or discovery
- Legally regulated but often adversarial process
- Agencies often fail to provide what they said they would (often due to internal disorganization)
- Very common to have a back-and-forth

Special Challenges of CPR

- 100 + FOIA of many different agencies, often involve months long litigation
- Dozens of journalists uploading results of these requests to their system
- Requests include video, audio & documents
- Police departments store / organize their data in different ways

Road Map

Case Study 1: Interview study of public defenders

Case Study 2: Building technology for investigative journalists

1. Background: California Reporting Project and public record requests
2. Needfinding & prototype development
3. Developing & evaluating a prototype
4. Next steps and larger themes

What could I do to help

Unsupervised clustering to identify topics and patterns in the police reports

→ unsupervised generally not meaningful, since most of the data is superfluous

→ hard because initial structure is so unknown

First we need to just understand what is in the data we have

What information is in this public records request?

Did the agency fail to provide what was legally required?

Pain points

Nested zip files / Corrupt zip files

Unnamed and unstructured files

Cases can be in many small or one large PDF

Constraints

Run in the same system where journalists have access to the data

Easy to run and interpret so that it can be used iteratively

Easy to run and interpret so that it can be used iteratively

System Constraints

Run in the same system where journalists have access to the data

Be able to process tens of thousands of files quickly

Easy to run and interpret so that it can be used iteratively

Handle different people working in parallel

InspectFiles.ipynb

File Edit View Insert Runtime Tools Help All changes saved

Comment Share R

+ Code + Text Connect Editing

Part 1 Get Access to Google Drive and Configure Output Paths

Give the notebook access to your google drive (click yes to all the prompts) After you click yes, this will takea little while to run.

```
[ ] import sys
    import pandas as pd
    from google.colab import drive
    drive.mount('/content/drive/')
```

```
Mounted at /content/drive/
```

Give the Google API access to your credentials (This is needed in order to get the last modified and created user and date.

```
[ ] from google.colab import files, auth
    from google.colab import auth
    auth.authenticate_user()
    from googleapiclient.discovery import build
    drive_service = build('drive', 'v3')
```

Set the value of input data to directory that you would like to inspect. Before doing this, you will have to add a shortcut to the google drive folder to your home directory.

```
[ ] # Enter the path of the directory you want to inspect
    import os

    INPUT_DATA = '/content/drive/MyDrive/narative_watch_code/Case Fi
    print(INPUT_DATA)

    if os.path.exists(INPUT_DATA):
      print(f'Input path {INPUT_DATA} is valid')
    else:
      print(f"ERROR!! {INPUT_DATA} does not exists")
```

INPUT_DATA: "/content/drive/MyDrive/narative"

File Indexer

Google Colab Notebook to create an index of files

	A	B	C	D	E	F	G	H	I	J	K	L	M
1	Jurisdiction	Agency	Total Files	# W Valid Case Date	# Unique Case Dates	Missing Docs	Missing Case Date	Largest File Name	Largest File Size	Earliest Case	Latest Case	document files: Total Files	document files: Extensions
2	Alameda County	Oakland Police Department	7288	7115	20	0	173	/Alameda County/Oaklan	3.64 GB	2007-12-31	2021-08-25	5458	['.pdf' '.PDF' '.do
3	Santa Clara County	San Jose Police Department	13452	3824	83	1	9628	/Santa Clara County/San	6.74 GB	2013-09-22	2021-05-31	4949	['.gdoc' '.pdf']
4	State agencies	Calif. Dept. of Corrections and Rehabilitation	5266	3086	236	0	2180	/State agencies/Calif. De	1.90 GB	2011-10-25	2019-10-30	4287	['.gdoc' '.pdf' '.PD
5	State agencies	Calif. Dept. of Justice	3609	2749	95	0	860	/State agencies/Calif. De	513.57 MB	2009-10-27	2018-11-30	3608	.pdf
6	Contra Costa County	Richmond Police Department	7178	3968	134	0	3210	/Contra Costa County/Ri	3.99 GB	2014-01-07	2021-04-28	3000	['.pdf' '.gdoc']
7	Orange County	Irvine Police Department	1466	1438	8	0	28	/Orange County/Irvine Pc	582.75 MB	2005-09-09	2013-09-03	1466	.pdf
8	Orange County	Anaheim Police Department	1318	0	0	0	1318	/Orange County/Anaheim	1.90 GB	0	0	976	['.pdf' '.PDF']
9	Los Angeles County	Los Angeles Police Department	858	130	49	0	728	/Los Angeles County/Los	114.01 MB	2007-07-06	2018-04-20	858	['.pdf' '.PDF']
10	Sonoma County	Santa Rosa Police Department - KQED BANG	4381	3031	19	0	1350	/Sonoma County/Santa F	1.87 GB	2013-06-08	2019-12-07	835	['.pdf' '.gdoc']
11	Orange County	Santa Ana Police Department	3400	1338	18	4	2062	/Orange County/Santa A	2.78 GB	2008-05-18	2018-08-23	805	['.pdf' '.PDF' '.do
12	Riverside County	Hemet	7284	24	8	5	7260	/Riverside County/Hemet	4.19 GB	2013-06-26	2020-02-11	789	['.pdf' '.docx']
13	Fresno County	Fresno Police Department	1658	1132	44	3	526	/Fresno County/Fresno F	3.84 GB	2012-07-12	2018-11-12	698	['.pdf' '.gdoc']
14	San Bernardino County	Ontario Police Department	1306	400	12	0	906	/San Bernardino County/	800.11 MB	2009-10-14	2017-04-09	648	.pdf
15	Kern County	Kern County Sheriff	1518	84	19	0	1434	/Kern County/Kern Count	1.64 GB	2013-05-08	2018-12-12	603	.pdf
16	Los Angeles County	Burbank Police Dept.	1946	0	0	0	1946	/Los Angeles County/Bur	2.22 GB	0	0	576	['.pdf' '.docx']
17	State agencies	California Highway Patrol (MAIN)	4076	4019	118	6	57	/State agencies/Californi	4.21 GB	2009-09-26	2021-10-05	508	['.gdoc' '.pdf' '.PD
18	Merced County	UC Merced Police Department	3622	3622	1	0	0	/Merced County/UC Merc	128.82 MB	2015-11-04	2015-11-04	486	.pdf
19	San Diego County	El Cajon Police Department	505	505	4	0	0	/San Diego County/El Ca	93.38 MB	2014-05-15	2017-01-01	452	.pdf
20	Riverside County	Riverside County District Attorney	446	175	111	0	271	/Riverside County/Rivers	360.83 MB	2011-08-03	2017-10-31	445	['.pdf' '.gdoc' '.do
21	Riverside County	Palm Springs Police Department	1307	108	18	0	1199	/Riverside County/Palm S	444.56 MB	2005-06-16	2019-09-09	370	['.pdf' '.gdoc']
22													

Outputs a spreadsheet of files by subfolder

Features

Recursively unzips files

Counts files of various types

Uses the Google Drive API to determine who edited different files and when

Extracts the date and some other information from file paths to guess at how many unique cases are represented in the data

Matches data of different file formats to a case

—

oc/irvine/interview/

alexander_08_09_2020_7.4.5.2021
alexander_08_09_2020_7.4.5.2021
smith_12.19.2020_7.1.2021.mp3
smith_12.19.2020_7.2.2021.mp3
warren_12.07.2020_6.1.2021.mp3
warren_12.07.2020_6.2.2021.mp3
warren_12.07.2020_6.3.2021.mp3

oc/irvine/report/

alexander_08_09_2020.pdf
alexander_08_09_2020_Part2.pdf
alexander_08_09_2020_Part3.pdf
alexander_08_09_2020_Part4.pdf
warren_Appendix-A-20-12-07.pdf
warren_Appendix-B-20-12-07.pdf
warren_20-12-07.pdf

We want to match cases across types

Road Map

Case Study 1: Interview study of public defenders

Case Study 2: Building technology for investigative journalists

1. Background: California Reporting Project and public record requests
2. Needfinding & prototype
3. Evaluating the prototype
4. Next steps and larger themes

User evaluation

Observing journalists using the output of the tool

Identify Missing Data

- Difference between number of unique dates in docs vs. audio file → missing data type
- Find missing date ranges

Prioritize jurisdictions for review

- Number of unique dates in → high number of cases
- Large files suggest → internal investigations, compex cases

–

“Productionizing”

Training 5 people with little technical experience to use the tool

Challenges

Using the Notebook In Practice

- Required administrator involvement to authorize running colab notebooks in organizations drive
- Version control and concurrency
- Required better error handling (what if it takes to long to unzip a file, if its corrupted)

“Productionizing”: Training 5 people with little technical experience to use the tool!

Challenges

- Required administrator involvement to authorize running colab notebooks in organizations drive
- Version control and concurrency
- Required better error handling (what if it takes to long to unzip a file, if its corrupted)

Successes

- Ultimately everyone was able to run it
- The first coding experience for a few people
- Possible to make slight modifications
- Used iteratively, on new data that came in

Success Stories!

Unzipping and deduplicating is a simple time saver

Used to support negotiations with two agencies, one where police reports were missing and one where a chunk of files from a certain time frame was missing

Help prioritize work of going through data

Will be used to manage a next round of 100+ data requests

Road Map

Case Study 1: Interview study of public defenders

Case Study 2: Building technology for investigative journalists

1. Background: California Reporting Project and public record requests
2. Needfinding & prototype
3. Evaluating the prototype
4. Next steps and larger themes

Next Steps

1. Evaluate tool as new requests come in
2. Pilot the tool with other investigative journalists
3. Tool improvements → pdf inspection, better pattern recognition

Larger Technical Problem: pattern recognition from semi-organized data

oc/irvine/interview/

alexander_08_09_2020_7.4.5.2021.mp3
alexander_08_09_2020_7.4.5.2021.mp3
smith_12.07.2020_7.1.2021.mp3
smith_12.07.2020_7.2.2021.mp3
warren_12.07.2020_6.1.2021.mp3
warren_12.07.2020_6.2.2021.mp3
warren_12.07.2020_6.3.2021.mp3

oc/irvine/report/

Alexander_08_09_2020.pdf
alexander_08_09_2020.pdf
alexander_08_09_2020_Part2.pdf
alexander_08_09_2020_Part3.pdf
alexander_08_09_2020_Part4.pdf
warren_Appendix-A-20-12-07.pdf
warren_Appendix-B-20-12-07.pdf
warren_20-12-07.pdf

What if we could extrapolate from known pattern?

oc/irvine/interview/

alexander_08_09_2020_7.4.5.2021.mp3
alexander_08_09_2020_7.4.5.2021.mp3
smith_12.07.2020_7.1.2021.mp3
smith_12.07.2020_7.2.2021.mp3
warren_12.07.2020_6.1.2021.mp3
warren_12.07.2020_6.2.2021.mp3
warren_12.07.2020_6.3.2021.mp3

oc/irvine/report/

Alexander_08_09_2020.pdf
alexander_08_09_2020.pdf
alexander_08_09_2020_Part2.pdf
alexander_08_09_2020_Part3.pdf
alexander_08_09_2020_Part4.pdf
warren_Appendix-A-20-12-07.pdf
warren_Appendix-B-20-12-07.pdf
warren_20-12-07.pdf

oc/irvine/alexander_08_09_2020/

interview_7.4.5.2021.mp3
Interview_7.4.5.2021.mp3

oc/irvine/smith_12_19_2020/

oc/irvine/warren_12_07_2020/

smith_12.19.2020_7.1.2021.mp3
smith_12.19.2020_7.2.2021.mp3
warren_12.07.2020_6.1.2021.mp3
warren_12.07.2020_6.2.2021.mp3
warren_12.07.2020_6.3.2021.mp3

oc/irvine/report/

alexander_08_09_2020.pdf
alexander_08_09_2020_Part2.pdf
alexander_08_09_2020_Part3.pdf
alexander_08_09_2020_Part4.pdf
warren_Appendix-A-20-12-07.pdf
warren_Appendix-B-20-12-07.pdf
warren_20-12-07.pdf

Larger Themes & Future Directions

1. Emerging problem, iterative relationship and data management
2. The most important problems might not come with ready to use clean data!
3. Value in simple tools that fit the problem

Thank You!

rwarren2@uci.edu