

Hellina Hailu Nigatu

Oct 26, 2022

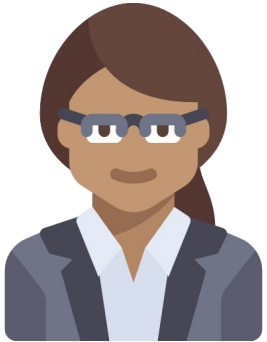
Sarah Chasins & John Canny

DOT: Building a Document Organization Tool for Domain Experts



Berkeley
UNIVERSITY OF CALIFORNIA

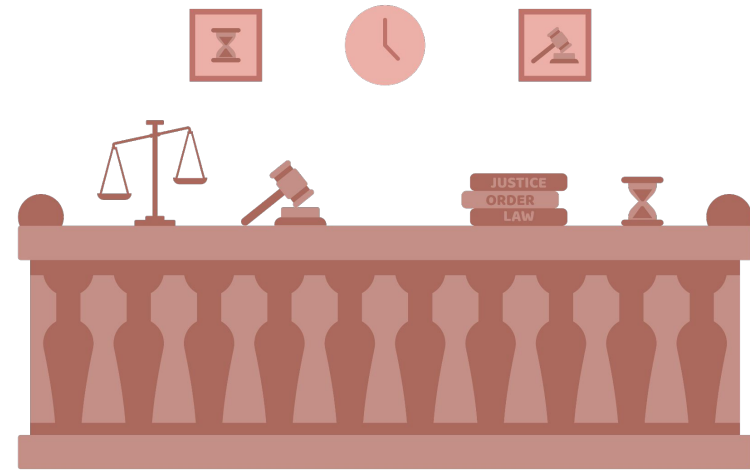
Introduction



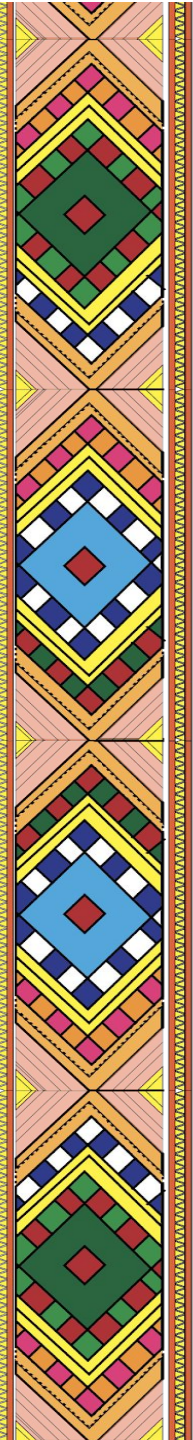
Public Defender

Prior Cases involving Officer Tom

- Patterns
- Misconduct filings

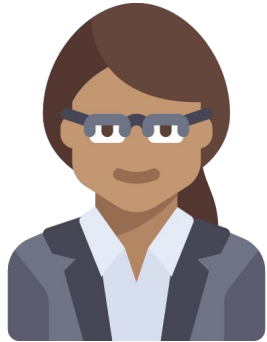


Court



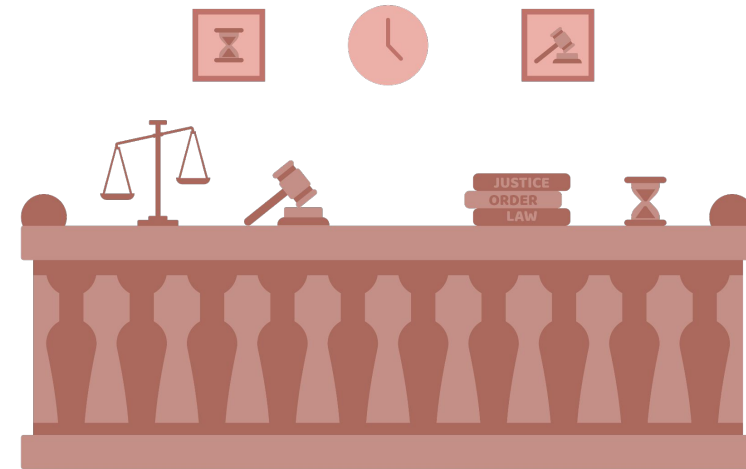
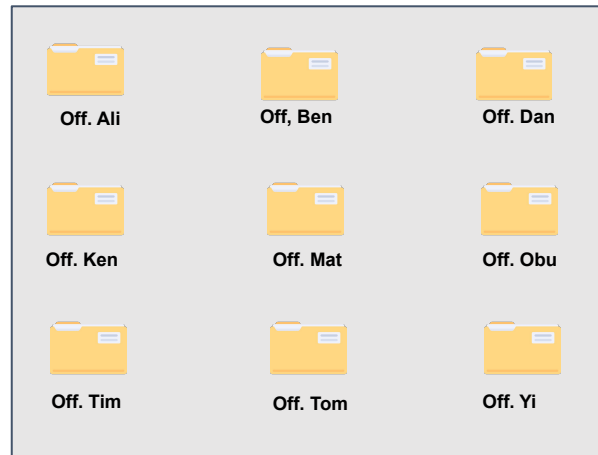
Introduction

In an ideal scenario...

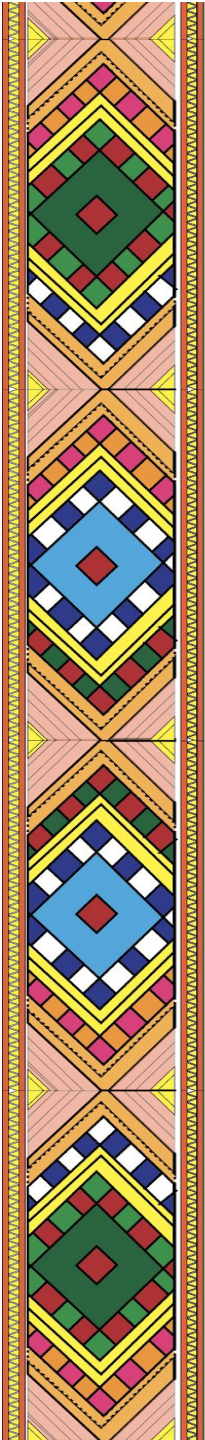


Public Defender

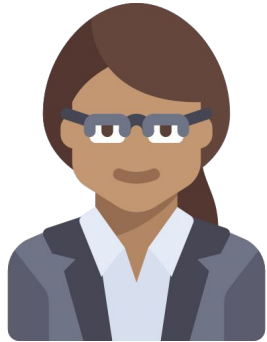
Well Organized Documents



Court

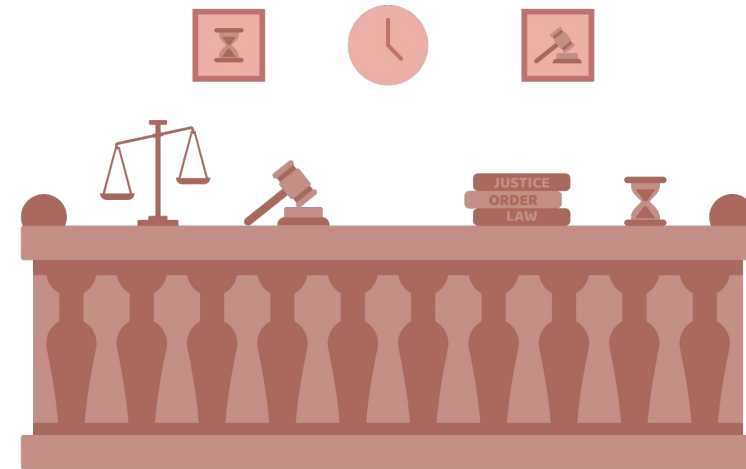
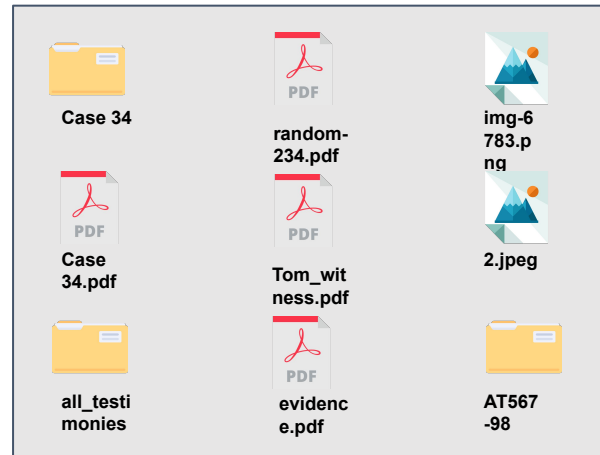


Introduction

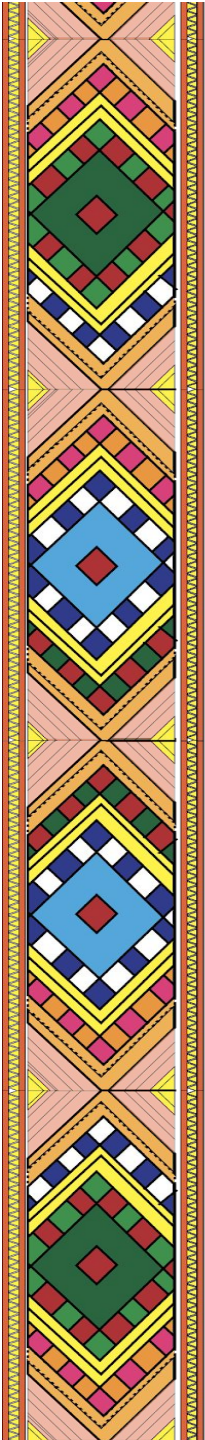


Public Defender

Document Dumps

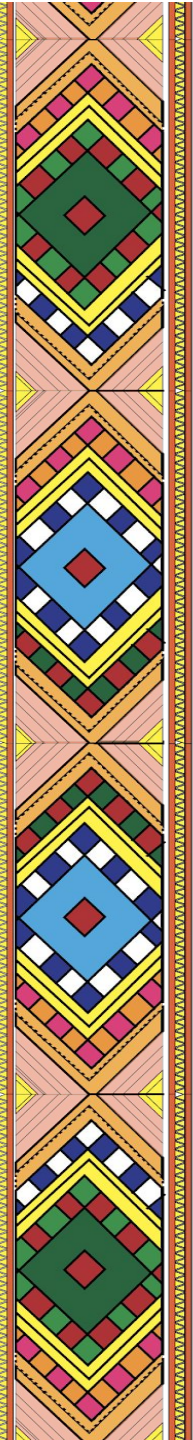
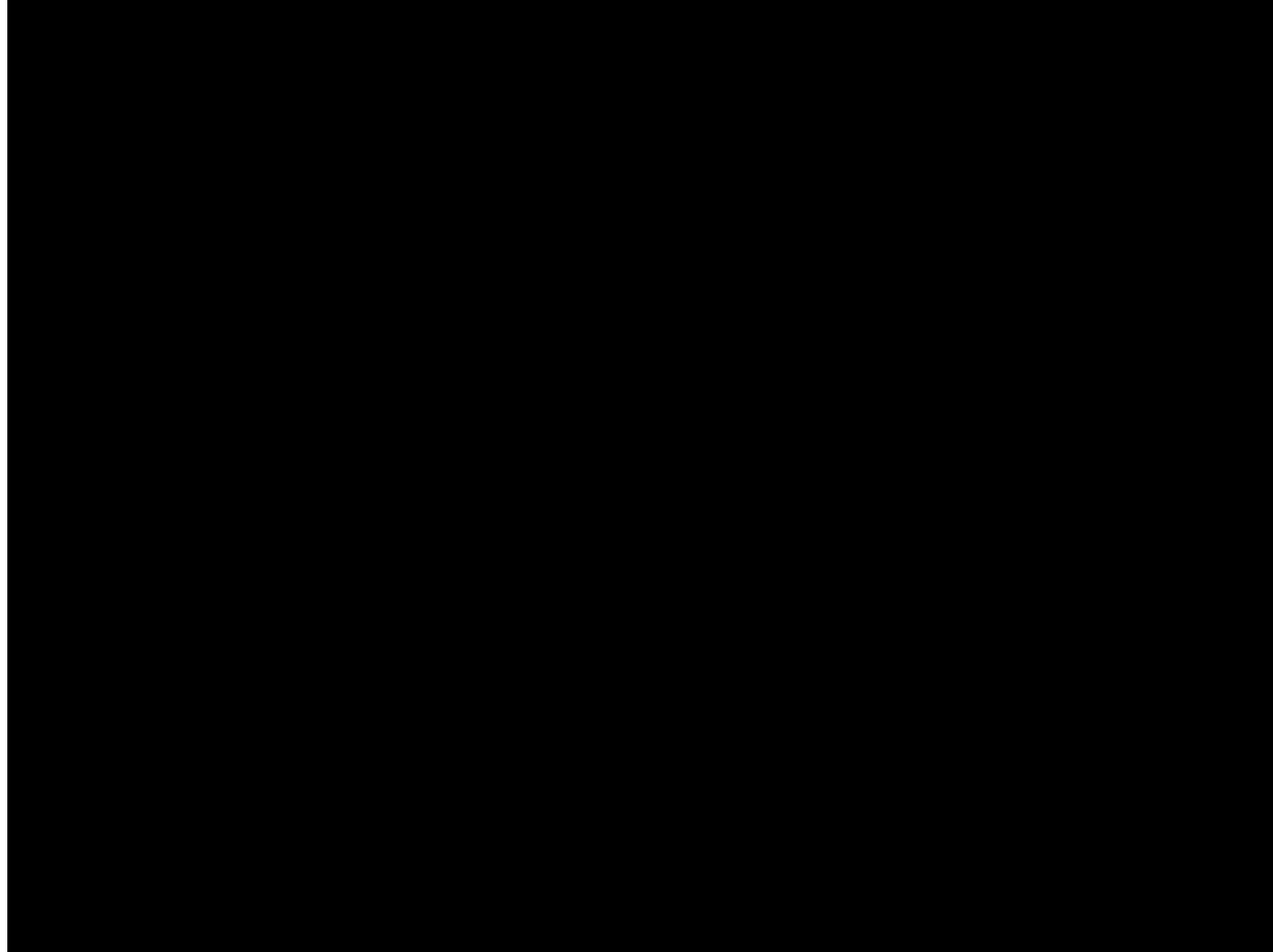


Court



Introduction

1. **Volume** of data makes it hard to work through manually.



Introduction

1. Volume

2. Duplicate files exist with in the data, wasting manual efforts in processing.

MURRIETA POLICE DEPARTMENT
2 Front Street, Murrieta, CA 92562
Phone (951) 204-3227 Fax (951) 494-3008 CAD Case # 8

INCIDENT REPORT

Case # 1802M-4663

DATE: 08/27/2018 TIME: 02:00 PM LOCATION: 22004
OFFICER: [REDACTED] DISPATCHER: [REDACTED]

TYPE: [REDACTED] SUBTYPE: [REDACTED]

REPORTING OFFICER: [REDACTED]

SUSPECT: BALDORNICA, ERYN DANIEL

STATE OF CALIFORNIA

Case_#_1802M-4663.pdf

MURRIETA POLICE DEPARTMENT
2 Front Street, Murrieta, CA 92562
Phone (951) 204-3227 Fax (951) 494-3008 CAD Case # 8

INCIDENT REPORT

Case # 1802M-4663

DATE: 08/27/2018 TIME: 02:00 PM LOCATION: 22004
OFFICER: [REDACTED] DISPATCHER: [REDACTED]

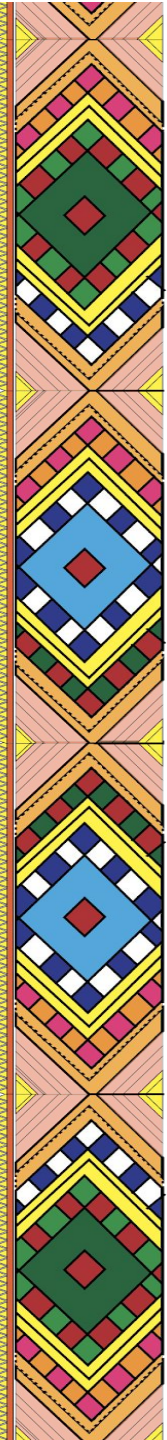
TYPE: [REDACTED] SUBTYPE: [REDACTED]

REPORTING OFFICER: [REDACTED]

SUSPECT: BALDORNICA, ERYN DANIEL

STATE OF CALIFORNIA

MurrietaPoliceDeptMatthe...



Introduction

- 1. Volume.
- 2. Duplicates.

3. Data exhibits different page types, making it hard to find the data points of interest fast.

NEVADA COUNTY SHERIFF'S OFFICE KEITH ROYAL SHERIFF/CORONER PUBLIC ADMINISTRATOR

RE: NOTICE OF ADMINISTRATIVE INVESTIGATION AND INTERVIEW
 DATE: March 5, 2018
 TO: Corporal Rory Sommer
 FROM: Lieutenant Walsh/Sergeant Tyrner

COPY JA 2018-001

An investigation of suspected misconduct by an employee or member of this department is being conducted. You have been identified as:

X The employee suspected of misconduct. Employees accused or suspected of misconduct should be aware that subsequent investigation/testimony may be used against them in administrative proceedings and may be discoverable pursuant to the Evidence Code. Employees so accused or suspected are entitled to specific procedural rights under Government Code Section 3300 et seq. or pursuant to labor agreements between their bargaining unit and the county. These rights include a right to representation by a person of their choosing who is not involved in the same investigation either as a witness or suspect. If desired, it is the employee's responsibility to secure representation prior to the date and time of the scheduled interview.

A witness to the incident. Personnel interviewed as witnesses are neither the subject of the investigation nor is a punitive action proposed or contemplated against the employee as a result of the alleged incident at this time. As this is an administrative inquiry, you are being ordered by the Sheriff to cooperate and answer all questions asked of you truthfully, fully, and completely. Questions truthfully will be deemed as insubordination, and punitive action will be taken. If during the investigation information is discovered wherein punitive action is possible or contemplated, you will be advised of that fact and offered all of those rights afforded to an employee suspected of misconduct. As a witness, you do have the right to representation by a person of your choosing who is not involved in the same investigation either as a witness or suspect. If desired, it is the employee's responsibility to secure representation prior to the date and time of the scheduled interview.

In order to complete this investigation, you are required to furnish a detailed account of your action and/or observations as soon as possible. You are hereby instructed not to discuss the substance of the investigation with anyone with the exception of your attorney or representative.

Contact this investigator as soon as possible, but no later than _____ at telephone number (330) _____ for an appointment for an interview.

(Interview TBD you will be notified in writing.)

Print Date: 01/07/16 Page 12
 Rev. Date: 01/1/2016

INCIDENT REPORT NEVADA COUNTY SHERIFF'S OFFICE 180002

1101	INCIDENT	SEARCHED	INDEXED	FILED	RECEIVED
1102	ADDITIONAL INFORMATION	SEARCHED	INDEXED	FILED	RECEIVED
1103	ADDITIONAL INFORMATION	SEARCHED	INDEXED	FILED	RECEIVED
1104	ADDITIONAL INFORMATION	SEARCHED	INDEXED	FILED	RECEIVED
1105	ADDITIONAL INFORMATION	SEARCHED	INDEXED	FILED	RECEIVED
1106	ADDITIONAL INFORMATION	SEARCHED	INDEXED	FILED	RECEIVED
1107	ADDITIONAL INFORMATION	SEARCHED	INDEXED	FILED	RECEIVED
1108	ADDITIONAL INFORMATION	SEARCHED	INDEXED	FILED	RECEIVED
1109	ADDITIONAL INFORMATION	SEARCHED	INDEXED	FILED	RECEIVED
1110	ADDITIONAL INFORMATION	SEARCHED	INDEXED	FILED	RECEIVED
1111	ADDITIONAL INFORMATION	SEARCHED	INDEXED	FILED	RECEIVED
1112	ADDITIONAL INFORMATION	SEARCHED	INDEXED	FILED	RECEIVED
1113	ADDITIONAL INFORMATION	SEARCHED	INDEXED	FILED	RECEIVED
1114	ADDITIONAL INFORMATION	SEARCHED	INDEXED	FILED	RECEIVED
1115	ADDITIONAL INFORMATION	SEARCHED	INDEXED	FILED	RECEIVED
1116	ADDITIONAL INFORMATION	SEARCHED	INDEXED	FILED	RECEIVED
1117	ADDITIONAL INFORMATION	SEARCHED	INDEXED	FILED	RECEIVED
1118	ADDITIONAL INFORMATION	SEARCHED	INDEXED	FILED	RECEIVED
1119	ADDITIONAL INFORMATION	SEARCHED	INDEXED	FILED	RECEIVED
1120	ADDITIONAL INFORMATION	SEARCHED	INDEXED	FILED	RECEIVED

NARRATIVE
 INFORMATIONAL REPORT
 On October 13 of 2018, Sergeant [redacted] and I responded to the storage facility where the suspect vehicle was being stored in the attached case. Several items from the searched vehicle had been identified by [redacted] belonging to [redacted] as a victim in a burglary in Nevada City (NVC) on 08/20/18.
 Several items belonging to [redacted] were collected and booked into the Sheriff's Property Unit. Refer to property section of report for further details.

https://webmail.aecounty.gov/

Pursuit/018- Jeff Tyrner Page 1 of 1

Pursuit/018-2/22/18

Jeff Tyrner
 Tm 3:30:19 AM

Greetings,

I am conducting an Internal Affairs investigation on behalf of the Nevada County Sheriff's Office in reference to the Pursuit which ended in a multiple Officer Involved Shooting dated February 22, 2018. My records indicate that you responded to the scene.

I am interested in setting up an interview with you as a witness to the investigation and anticipate that it will be brief. Please contact me at the below number.

Sergeant Jeff Tyrner
 Nevada County Sheriff's Office
 930 Main Ave.
 Nevada City, CA 95959
 jeff.tyrner@necounty.ca.us
 (530) 265-1328 Business
 (530) 265-8451 Fax

https://webmail.aecounty.gov/ 3/20/2018

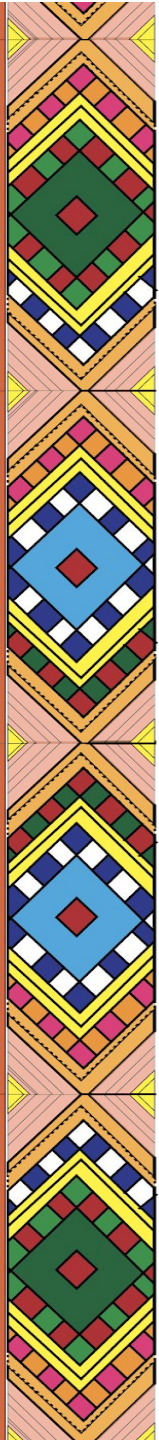
for the items being upon preliminary investigation by authorities certain of you lost these charges without reporting it appears to be nothing but a copy I keep with a date of entry in the gutter notes are so just address it TMC to Mac

Well you like me to send you some of the items that were missing or have gone missing but you didn't report them - grand old folks money, I got a good appraisal on the phone at 8100,000 for the jewelry and some digital equipment for a total value of 250,000,000 I got 250,000 for it - well, but the items the top was factory and so the items were in the evidence room I don't know for sure what you want to do I think I can get you some more information about the items that were missing and I can help you with that

I'll send you some more information about the items that were missing and I can help you with that

I'll send you some more information about the items that were missing and I can help you with that

I'll send you some more information about the items that were missing and I can help you with that



Introduction

1. Volume.
2. Duplicates.

3. Data exhibits different page types, making it hard to find the data points of interest fast.

NEVADA COUNTY SHERIFF'S OFFICE
KEITH ROYAL, SHERIFF/PROBATIONER PUBLIC ADMINISTRATOR

RE: NOTICE OF ADMINISTRATIVE INVESTIGATION AND INTERVIEW
DATE: March 6, 2018
TO: Corporal Rory Springer
FROM: Lieutenant Walsh/Sergeant Tyner

COPY 2A 2018 - 004

An investigation of suspected misconduct by an employee or member of this department is being conducted. You have been identified as:

A witness to the incident. Personnel interviewed as witnesses are neither the subject of the investigation nor is a punitive action proposed or contemplated against the employee as a result of the alleged incident at this time. As this is an exploratory inquiry, you are being advised by the Sheriff to cooperate and answer all questions asked of you truthfully. Failure to cooperate or answer questions truthfully will be deemed as insubordination, and punitive action will be taken. If during the investigation information is discovered wherein punitive action is possible or contemplated, you will be advised of that fact and offered all of the rights afforded to an employee suspected of misconduct. As a witness, you do have the right to representation by a person of your choosing who will not be involved in the same investigation either as a witness or suspect. If desired, it is the employee's responsibility to secure representation prior to the date and time of the scheduled interview.

If you wish to be interviewed as a witness, please contact the investigator assigned to your case at the time and place of the investigation as soon as possible. You are hereby instructed not to discuss the substance of the investigation with anyone with the exception of your attorney or representative.

Contact this investigator as soon as possible, but no later than, _____ at telephone number (530) _____ for an appointment for an interview.

(Interview TBD you will be notified in writing.)

Print Date 01/07/16 Page 12
Rev Date 01/12/04

Pursuit/OIS 2/22/18 - Jeff Tyner Page 1 of 1

Pursuit/OIS 2/22/18

Jeff Tyner
764 628918 vcm am

Greetings,

I am conducting an Internal Affairs investigation on behalf of the Nevada County Sheriff's Office in reference to the Pursuit which ended in a multiple Officer Involved Shooting dated February 22, 2018. My records indicate that you responded to the scene.

I am interested in setting up an interview with you as a witness to the investigation and anticipate that it will be brief. Please contact me at the below number.

Sergeant Jeff Tyner
Nevada County Sheriff's Office
950 Main Ave
Nevada City, CA 95959
Jeff.Tyner@dc-nv.com
(530) 245-1210 Business
(530) 245-8451 Fax

<https://webmail.nccosheriff.net/owa/> 3/20/2018

INCIDENT REPORT		NEVADA COUNTY SHERIFF'S OFFICE		1188302	
118	119	120	121	122	123
114	115	116	117	118	119
110	111	112	113	114	115
106	107	108	109	110	111
102	103	104	105	106	107
98	99	100	101	102	103
94	95	96	97	98	99
90	91	92	93	94	95
86	87	88	89	90	91
82	83	84	85	86	87
78	79	80	81	82	83
74	75	76	77	78	79
70	71	72	73	74	75
66	67	68	69	70	71
62	63	64	65	66	67
58	59	60	61	62	63
54	55	56	57	58	59
50	51	52	53	54	55

NARRATIVE

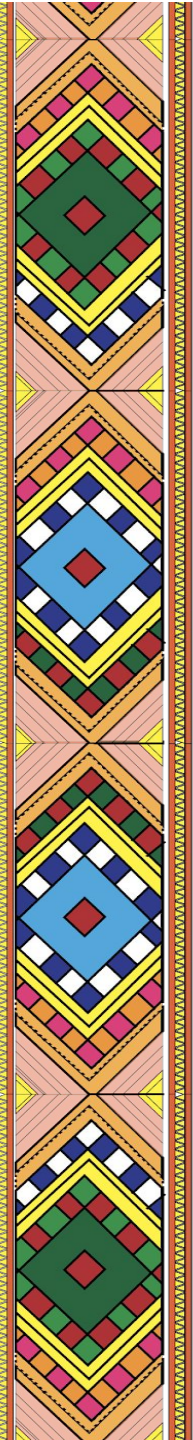
On October 11, 2018, Sergeant [REDACTED] and I responded to the storage facility where the suspect vehicle was being stored in the incident case. Several items from the searched vehicle had been identified by [REDACTED] as belonging to [REDACTED].

Several items belonging to [REDACTED] were collected and booked into the Sheriff's Property Unit. Refer to property section of report for further details.

for the time being, upon preliminary investigation for, actually contain up can look these charges without repeating the apparatus, the policy just to copy. I don't see a letter I will copy, the particular case so just address it.

TMC to Mac

What you like me to send you some of the less one for preparing a case pending, not only relevant, it goes to ground call some myself, I got a rough reporting on the place, all 1993, 2000, I just probably have some things to give to Bill your way, fairly good report. \$20,000 or so. Will be able to bring the body very quickly, had no equipment, not that weekend. I did a job for preparing a copy or computer, and then the 11th to 12th, I had a chance to think I wasn't a happy day, for preparing, hoping to see into a certain little islands, some things that I want to bring. I don't see a police report, I don't see anything, that makes 60 pages or longer, report to get from into the night, I don't know, the hand, since that time it D, many gaps, have your, thank you, [REDACTED]



Introduction

1. Volume.
2. Duplicates.
3. Page Types.

4. Data comes in different levels of quality, making data extraction extremely difficult.

UNCLASSIFIED
~~SECRET~~


(V21)

BIOGRAPHIC DATA

CHILE
Gen Augusto PINOCHET Ugarte
January 1975

(U) NAME: Gen Augusto Pinochet Ugarte (pee-noh-CHET), Army.

(U) POSITION: President (chief of state since 12 Sept 1973; position officially named President of the Government Junta, 12 Sept 1973-June 1974; Supreme Chief of the Nation and Head of the Executive Branch June-Dec 1974; President since 18 Dec 1974); and Commander in Chief of the Army (since 24 Aug 1973).



(U) 1973

(U) SIGNIFICANCE: Gen Pinochet, an intelligent, ambitious, professionally competent and experienced Infantry officer, is widely admired and respected by fellow officers. He became President and the strongest member of the Government Junta (composed of the four service commanders) following the 11 Sept 1973 military coup, the first in Chile since 1931, which overthrew the government of Marxist-Socialist Salvador Allende Gossens (President, 1970-1973). In June 1974, the Junta structure changed and Pinochet became head of the executive branch of the government, while continuing as head of the Junta, which became the legislative branch. Gen Pinochet would have preferred that the Armed Forces, and particularly the Army, remain in their traditional role as a professional, apolitical force that does not involve itself with partisan politics. The deteriorating economic and political situation, however, forced Pinochet reluctantly to join in the military intervention. The Junta abolished Congress and all political parties but claims to be moving towards a return to democracy. It is most concerned with rebuilding Chile, especially the economy; obtaining foreign arms purchases and making other preparations against the threat of war with Peru; and improving Chile's world image regarding human rights.

(U) POLITICS:

(U) International: Anti-Communist and anti-Cuban, Gen Pinochet has always spoken favorably of, and desires to keep close ties with, the United States. He has twice travelled to the U.S. He favors the acquisition of U.S. equipment and the training of Chilean military personnel in U.S. service schools. He shares the common concern of most Chilean Army officers over the threat of a possible invasion of Chile by Peru. Pinochet has served as an Instructor at the Ecuadorian Army War College and has travelled to Mexico and the Canal Zone.

UNCLASSIFIED
NO FOREIGN DISSEM
~~SECRET~~

Declassified by DIA

PAGE 1 OF 3


~~SECRET~~

BIOGRAPHIC DATA

CHILE
Gen Augusto PINOCHET Ugarte
January 1975

(U) NAME: Gen Augusto Pinochet Ugarte (pee-noh-CHET), Army.

(U) POSITION: President (chief of state since 12 Sept 1973; position officially named President of the Government Junta, 12 Sept 1973-June 1974; Supreme Chief of the Nation and Head of the Executive Branch June-Dec 1974; President since 18 Dec 1974); and Commander in Chief of the Army (since 24 Aug 1973).



(U) 1973

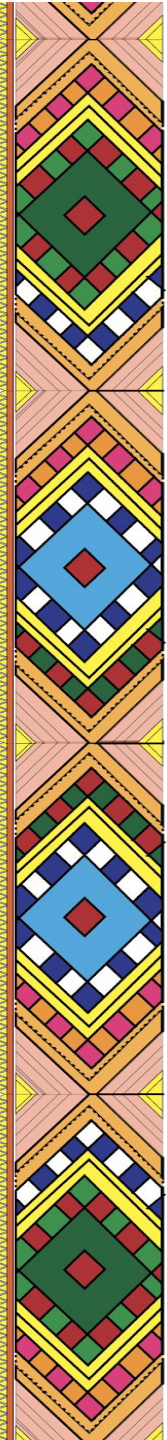
(U) He became President and the strongest member of the Government Junta (composed of the four service commanders) following the 11 Sept 1973 military coup, the first in Chile since 1931, which overthrew the government of Marxist-Socialist Salvador Allende Gossens (President, 1970-1973). In June 1974, the Junta structure changed and Pinochet became head of the executive branch of the government, while continuing as head of the Junta, which became the legislative branch.

The Junta abolished Congress and all political parties but claims to be moving towards a return to democracy. It is most concerned with rebuilding Chile, especially the economy.

(U) POLITICS:

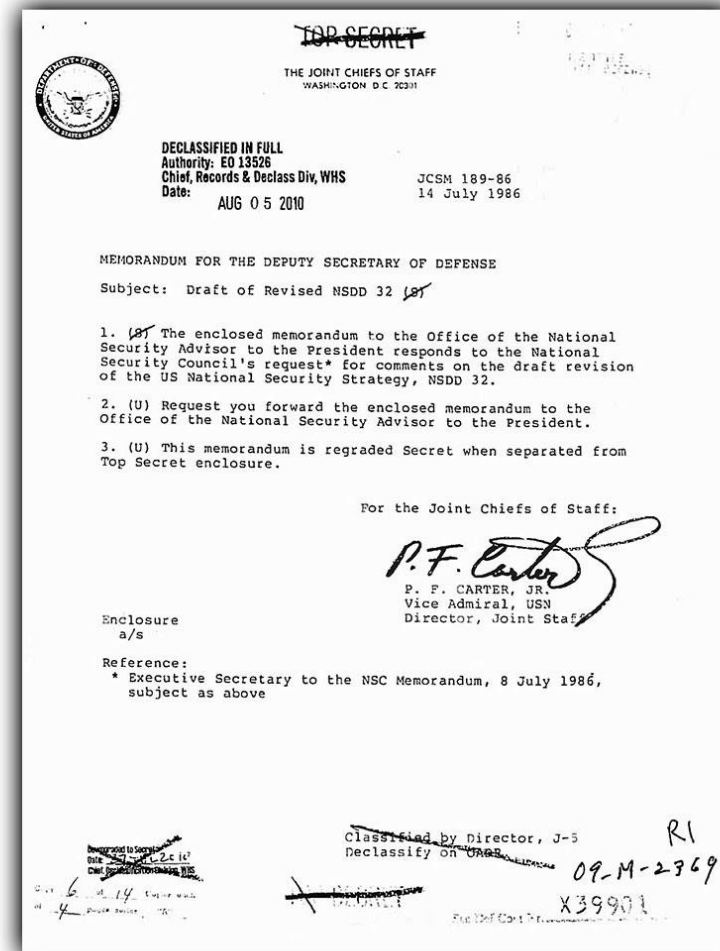
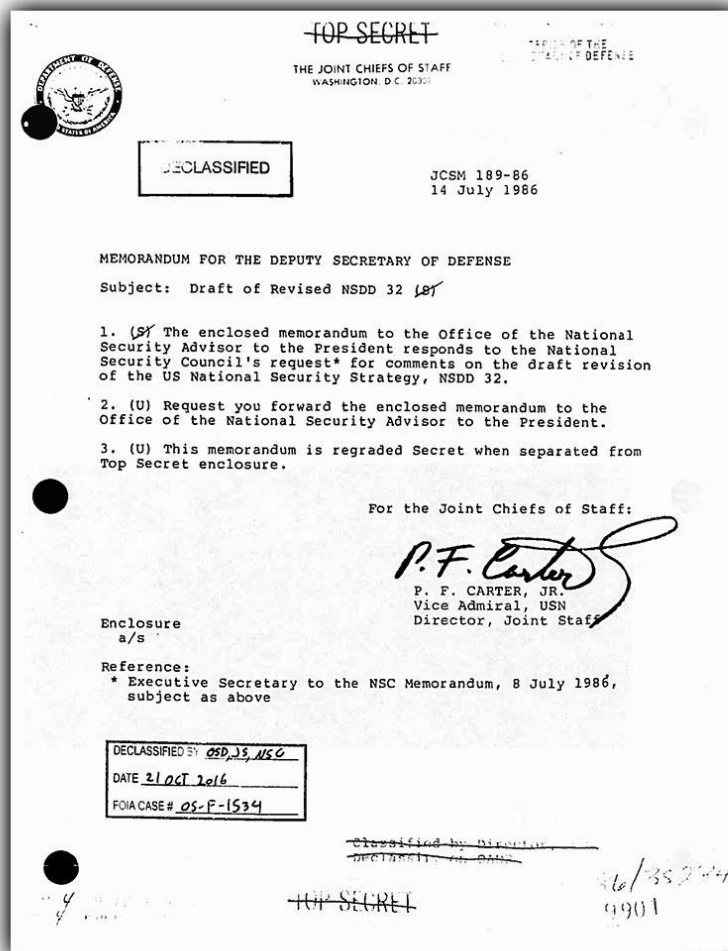
(U) International: He shares the common concern of most Chilean Army officers over the threat of a possible invasion of Chile by Peru. Pinochet has served as an Instructor at the Ecuadorian Army War College and has travelled to Mexico and the Canal Zone.

NO FOREIGN DISSEM
~~SECRET~~



Introduction

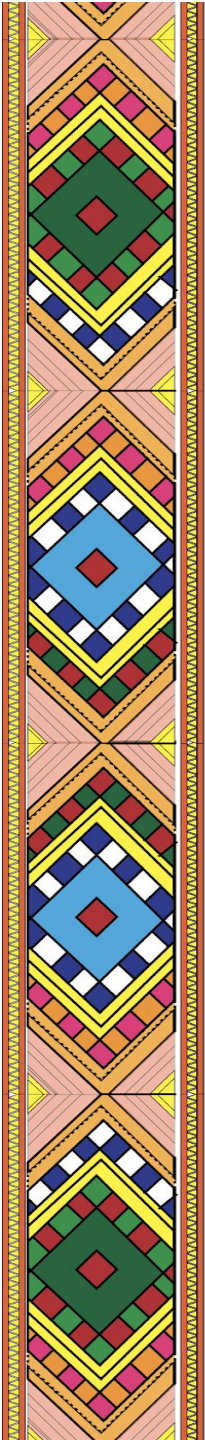
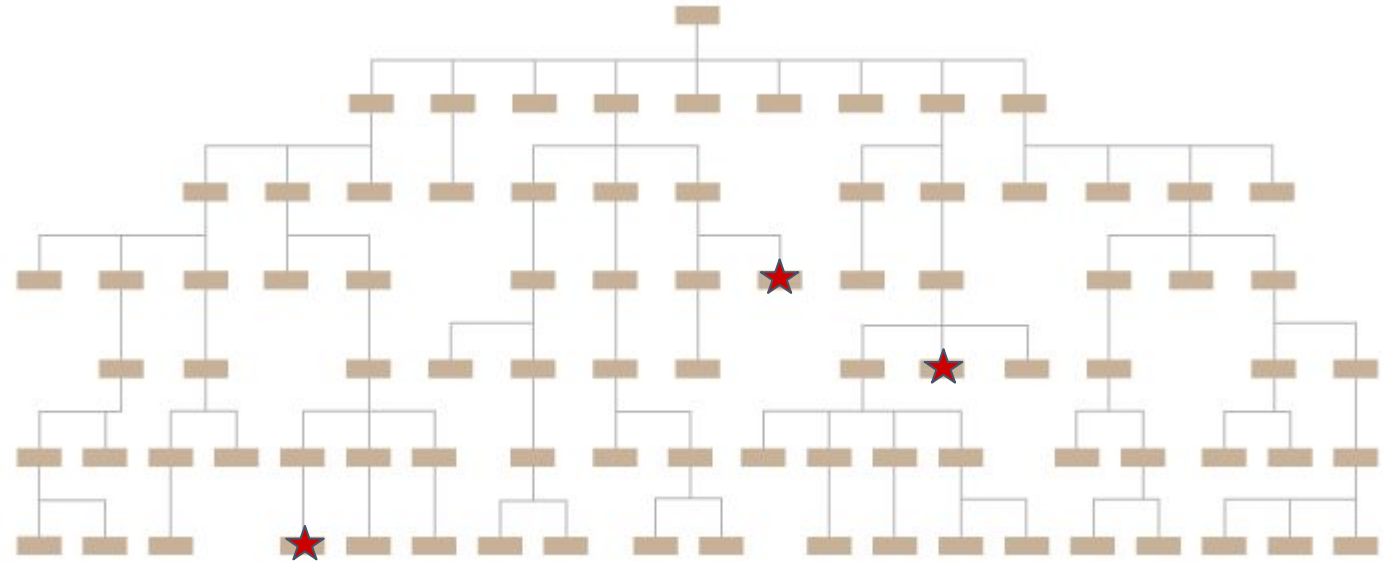
1. Volume.
2. Duplicates.
3. Page Types.
4. Data comes in different levels of quality, making data extraction extremely difficult.



Introduction

1. Volume.
2. Duplicates.
3. Page Types.
4. Quality of pages.

5. Files of the same case **spread across several folders.**



Introduction

1. Volume.
2. Duplicates.
3. Page Types.
4. Quality of pages.
5. One case; several files.

6. Multiple cases might be found in a single PDF.

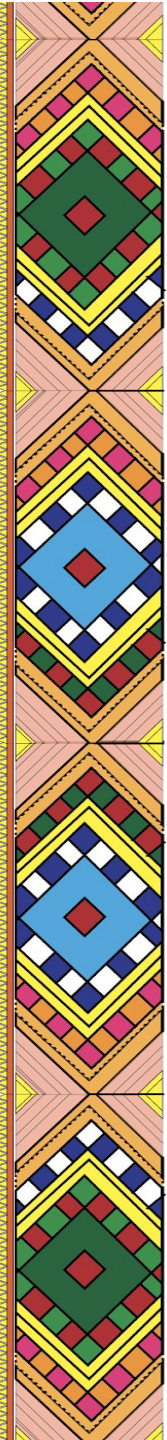


15-24439 Supplement No
0001

Reported Date
07/01/2015
Rpt/Incident Typ
288
Member#

15-34683 Supplement No
0003

Reported Date
09/13/2015
Rpt/Incident Typ
148
Member#

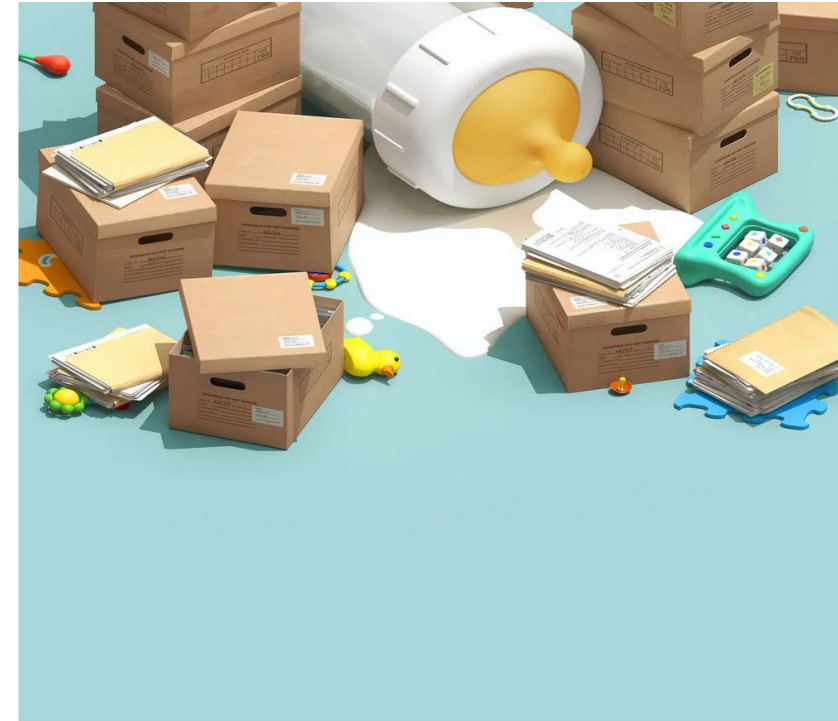


Introduction

Not just Public Defenders...

How Abbott Kept Sick Babies From Becoming a Scandal

Abbott's lawyers at Jones Day negotiated secret settlements and used scorched earth tactics with families whose infants fell ill after consuming powdered formula.



Max Guther

10 Gift Articles [Share](#) [Bookmark](#) [Comments](#) 505

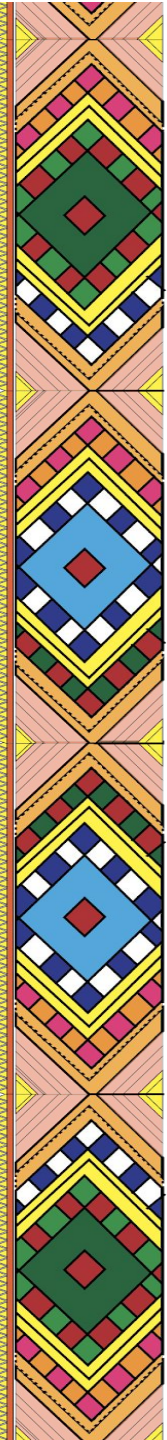


By David Enrich

David Enrich, the business investigations editor for The New York Times, is the author of the forthcoming book, "Servants of the Damned: Giant Law Firms, Donald Trump, and the Corruption of Justice," from which this article is adapted.

Published Sept. 6, 2022 Updated Sept. 8, 2022

Early on a Saturday morning in 2013, Mark Bennett, a federal judge, walked into his chambers in the courthouse in Sioux City, Iowa. He'd been out of town for a speaking engagement and was hoping to catch up on work. A surprise awaited him as he entered his office: Cardboard boxes were stacked everywhere. His



Introduction

Not just Public Defenders...

How Abbott Kept Sick Babies From Becoming a Scandal

Abbott's lawyers at Jones Day negotiated secret settlements and used scorched earth tactics with families whose infants fell ill after consuming powdered formula.



Max Guther

10 Gift Articles

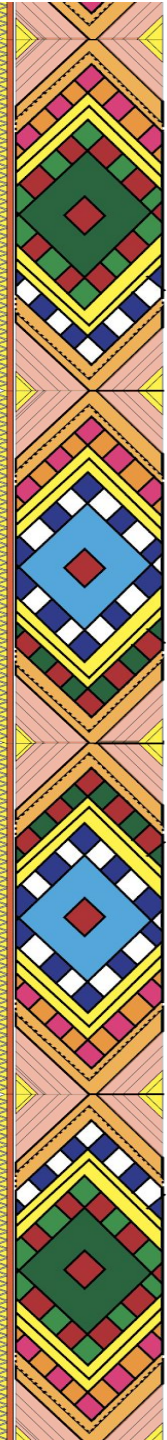
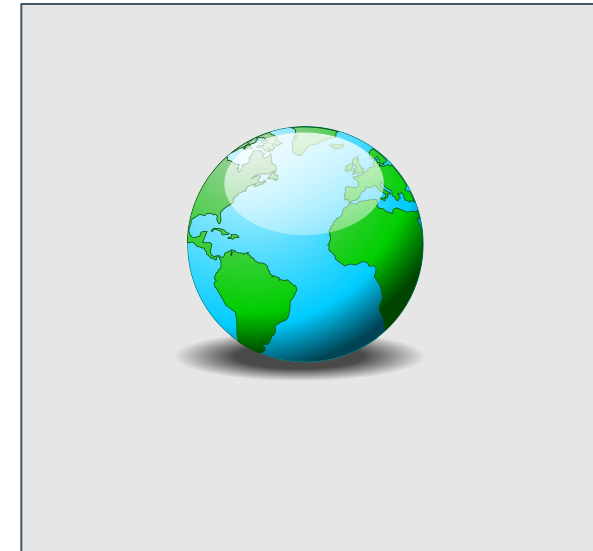


By David Enrich

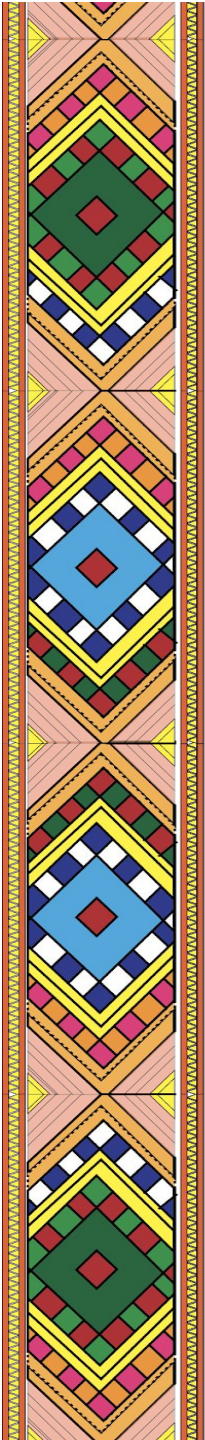
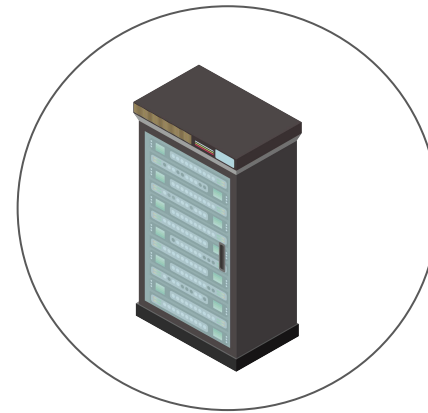
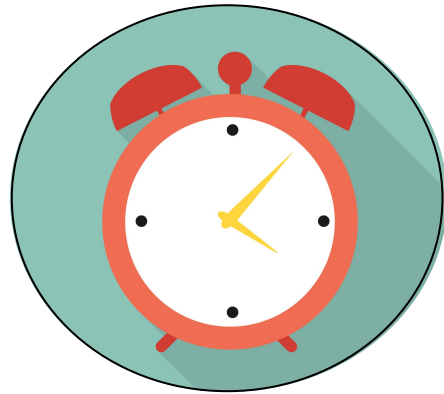
David Enrich, the business investigations editor for The New York Times, is the author of the forthcoming book, "Servants of the Damned: Giant Law Firms, Donald Trump, and the Corruption of Justice," from which this article is adapted.

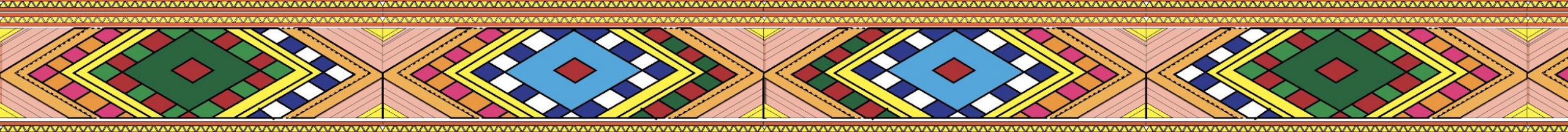
Published Sept. 6, 2022 Updated Sept. 6, 2022

Early on a Saturday morning in 2013, Mark Bennett, a federal judge, walked into his chambers in the courthouse in Sioux City, Iowa. He'd been out of town for a speaking engagement and was hoping to catch up on work. A surprise awaited him as he entered his office: Cardboard boxes were stacked everywhere. His

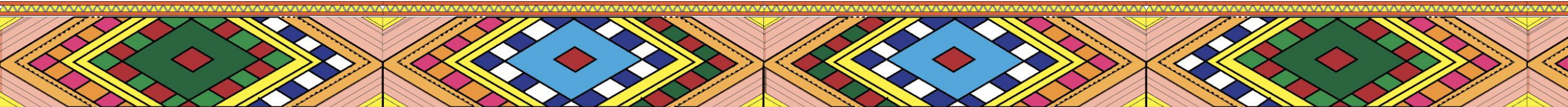


Introduction





There is a clear need for a data organization and cleaning tool for such domain experts who work with limited time, budget and technical expertise.



Exact Duplicate Detection

```
# import the required functions from the library.
from dotlibrary.exact_duplicate_functions import HashPages, threshold_by_percent, threshold_by_number_of_matched_pages
# run HashPages Function to get the duplicate info on file level, a dataframe for visualization and a dataframe with representative pages selected for processing.
file_level_df, reduced_df=HashPages(dataset_path="/home/alice/wonderland dataset/")

#set conditions for filtering exact duplicates
condition_1=threshold_by_percent(file_level_df, min_page_in_file=2, max_page_in_file=20,min_percentage_value=40)
condition_2=threshold_by_number_of_matched_pages(file_level_df, max_page_in_file=10, max_page_in_file_match=5)

# import functions for visualizing
from dotlibrary.exact_duplicate_functions import visualize_file_pairs, print_duplicate_info

print_duplicate_info((condition_2))

path_1="/home/alice/wonderland dataset/pages/main.pdf"
path_2="/home/alice/wonderland dataset/pages/wolf_11.pdf"
file_a, file_b=visualize_file_pairs(path_1, path_2)
```

Near Duplicate Detection

```
# import the classifier and pair correlation function and pass the dataset path along with the name of
# the page types we want to find near duplicates for.
from dotlibrary.near_duplicate_detection_functions import ClassifyAndGetPairCorrelation
correlation_df=ClassifyAndGetPairCorrelation('/home/alice/wonderland dataset/', ['form', 'image', 'narrative'])

from dotlibrary.near_duplicate_detection_functions import plot_pairs_of_pages
plot_pairs_of_pages(correlation_df, ['form'], 10, threshold=0.6, random=True, sort_ascending=False)
plot_pairs_of_pages(correlation_df, ['image'], 3, threshold=0, random=True, sort_ascending=False)

from dotlibrary.near_duplicate_detection_functions import set_correlation_threshold, print_near_duplicate_information
filtered_df=set_correlation_threshold(correlation_df, {'form': 0.9, 'image': 0.77})

print_near_duplicate_information(filtered_df)

from dotlibrary.near_duplicate_detection_functions import visualize_file_pairs
sample_8, sample_14=visualize_file_pairs('sample 8.pdf', 'sample 14.pdf')
sample_8
```

Data Extraction

```
from dotlibrary.data_extraction_functions import extract_with_bbox, clean_entity_list
listOfNames=get_named_entities(extracted_df['Text'], 'PERSON')
listOfDates=get_named_entities(extracted_df['Text'], 'DATE')
listOfLocations=get_named_entities(extracted_df['Text'], 'LOCATION')
listOfCaseNumbers=extract_with_bbox([100, 700, 200, 400], encoded_dataset, processor, 'form')

cleanNames=clean_entity_list(cleanNameList, listOfNames)
cleanDates=clean_entity_list(cleanDateList, listOfDates)
cleanLocations=clean_entity_list(['drive', 'highway', 'state', 'park', 'city'], listOfLocations, minLen=15, maxLen=100)

set_entity_breakdown(extracted_df, 'CaseNumber', listOfCaseNumbers)
set_entity_whole(extracted_df, 'Location', cleanLocations)
set_entity_whole(extracted_df, 'Date', cleanDates)
set_entity_whole(extracted_df, 'Names', cleanLocations)

set_entities_doc(extracted_df, doc_data, 'CaseNumber')
set_entities_doc(extracted_df, doc_data, 'Names')
set_entities_doc(extracted_df, doc_data, 'Date')
set_entities_doc(extracted_df, doc_data, 'Location')
```

Data Organization

```
demo=connections()
for ind in doc_data.index:
    singlepage=document(doc_data['File'][ind], doc_data['CaseNumber'][ind], doc_data['Names'][ind], doc_data['Date'][ind])
    demo.add_vertices(singlepage)
demo.find_connections()
con=demo.print_connections()
con=dict(sorted(con.items(), key=lambda item: len(item[1]), reverse=True))
key=list(con.keys())[3]
doc_data.sort_values(by='CaseNumber').drop(['Text', 'Names', 'Location'],
axis=1).style.apply(lambda x: ['background: lightgreen' if (con[key].intersection(set([x.File])))
else '' for i in x], axis=1)
```

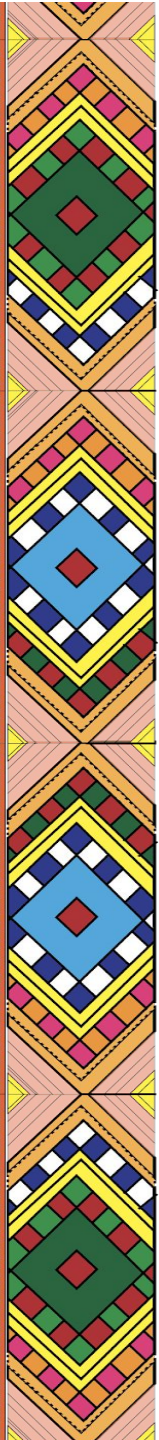
	CaseNumber	Date	File	highlightedcase
6	[17-47052, '2017-47052']	['10/14/17', '10/14/2017', '10/18/2017', '10/23/2017', '10/24/2017', '11/08/2017', '18/14/13']	Greg_Lambert0001	nan
4	[17-47052, '2017-47052']	['10/14/17', '10/14/2017', '10/18/2017', '10/23/2017', '10/24/2017', '11/08/2017', '18/14/13']	Jeff_Christian0001	nan
5	[17-47052, '2017-47052']	['10/14/17', '10/14/2017', '10/18/2017', '10/23/2017', '10/24/2017', '11/08/2017', '18/14/13']	Mark_Sudachan0001	nan
1	[17-47052, '2017-47052']	['10/14/17', '10/14/2017', '10/18/2017', '10/23/2017', '10/24/2017', '11/08/2017', '18/14/13']	Marc_Becerra00001	nan
0	[2004-31278]	['10/26/04']	Jse_Nnez00001	nan
2	[17-47052, '2017-47052']	['10/14/17', '10/14/2017', '10/18/2017', '10/23/2017', '10/24/2017', '11/08/2017', '18/14/13']	John_Donohue0001	nan
7	[2013-00065675, '2013-65675']	['04/20/13', '06/20/13', '09/14/2013', '09/16/13', '09/17/13', '09/20/13', '09/26/2013', '10/01/2013']	Rmer_Aberin00001	nan
3	nan	[]	Brian_Bnn_00001	nan

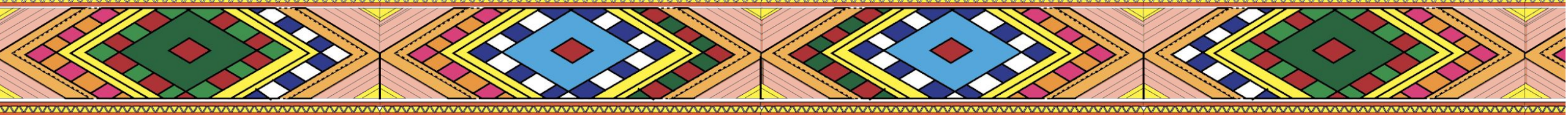
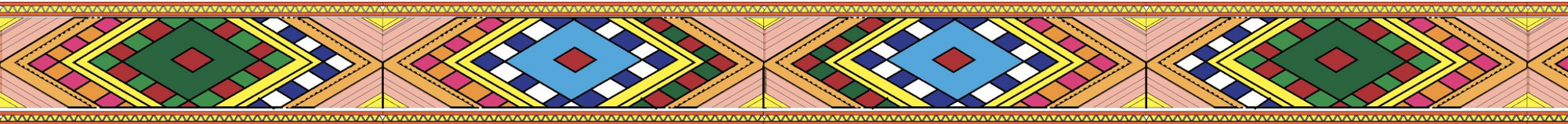
```
ca1=get_similar_casenumbers(doc_data, 13)
ca2=get_similar_casenumbers(doc_data, 6)
ca3=get_similar_casenumbers(doc_data, 12)
ca4=get_similar_casenumbers(doc_data, 3)

case1=caseFiles(highlight_text(doc_data, 'highlightedcase', 'CaseNumber', ca1))
case2=caseFiles(highlight_text(doc_data, 'highlightedcase', 'CaseNumber', ca2))
case3=caseFiles(highlight_text(doc_data, 'highlightedcase', 'CaseNumber', ca3))
case4=caseFiles(highlight_text(doc_data, 'highlightedcase', 'CaseNumber', ca4))

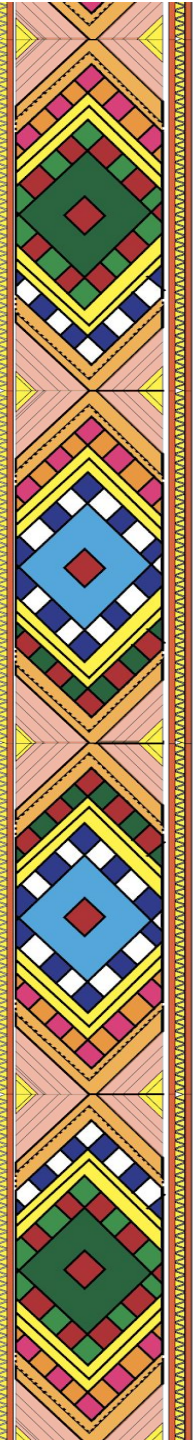
caseList=[]
caseList.append(case1)
caseList.append(case2)
caseList.append(case3)
caseList.append(case4)
```



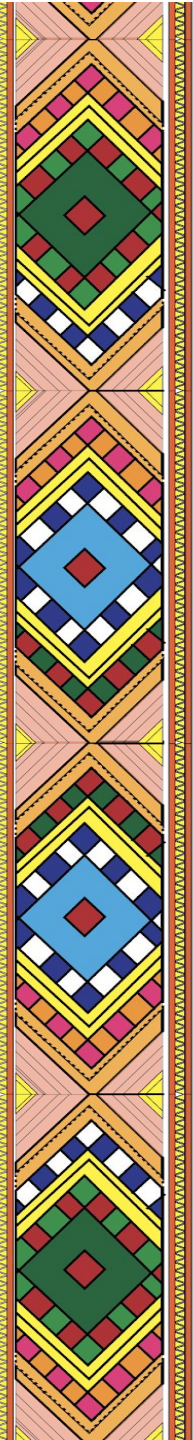
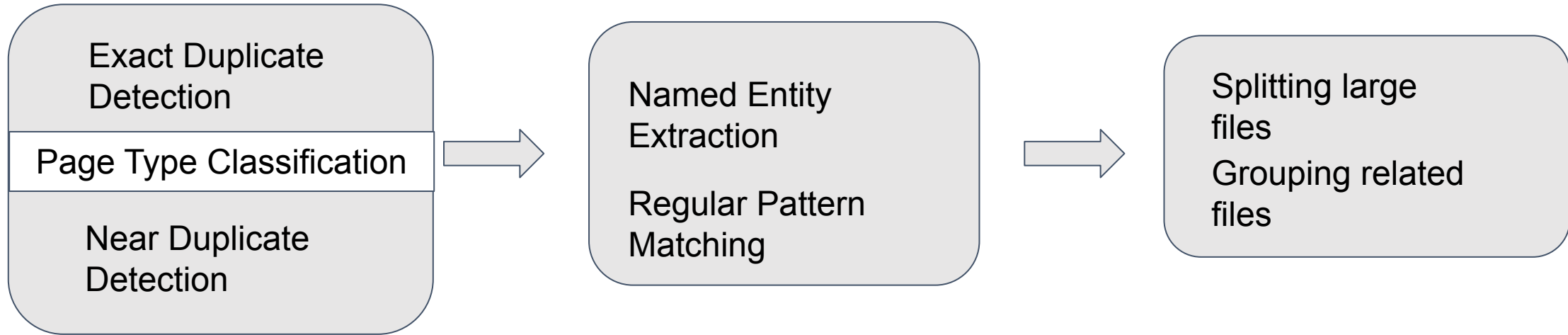


- 
- Implementation
 - Programming paradigm
- 

Methodology



Methodology



Methodology

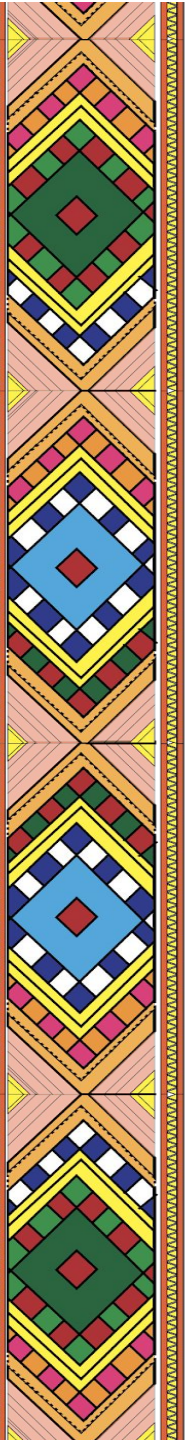
Data Cleaning



Data Extraction



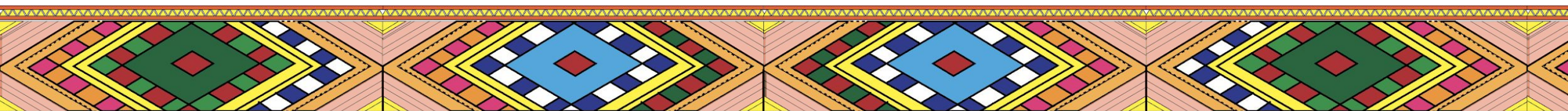
Data Organization



Problem Statement

Document scanning, redaction, handling, and storing practices result in data duplication in large document dumps.

- Save compute resources.
- Train models efficiently.
- Humans are manually extracting data; identifying duplicates to avoid repeated processing.
- Know the status of their data.
- Information on how departments are sharing data.



Challenges

Data Cleaning

Exact Duplicates

Exact pixel-for-pixel copies
of the same page.

May appear in same or
different folders.

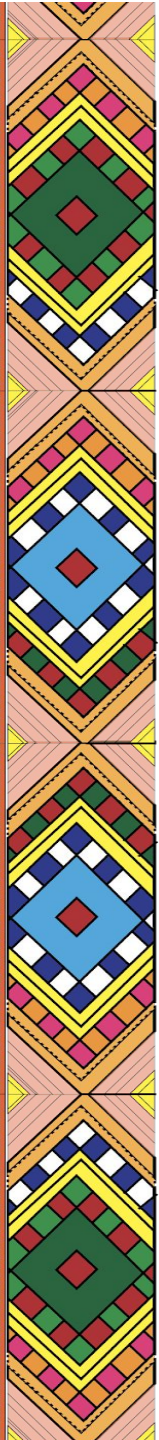
Smaller document may
be included in larger
document.

Near Duplicates

Same physical page in
the real world.

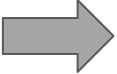
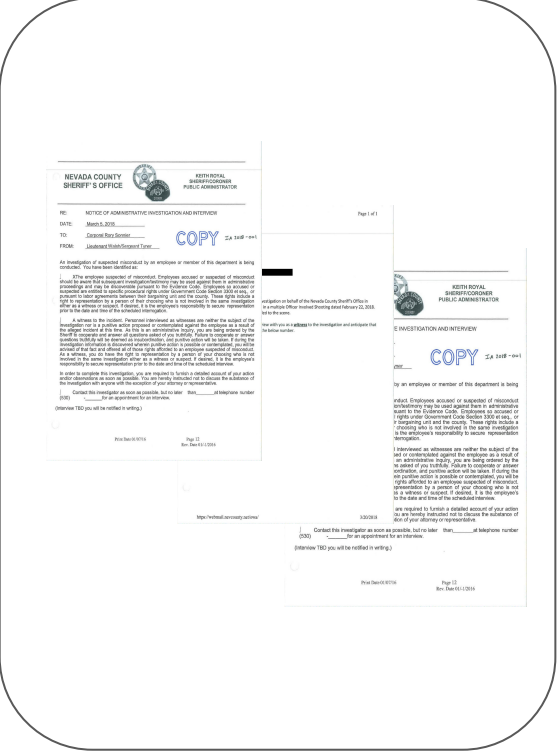
May include different
types/levels of
redaction.

May come from different
scans.



Methodology

Exact Duplicate Detection



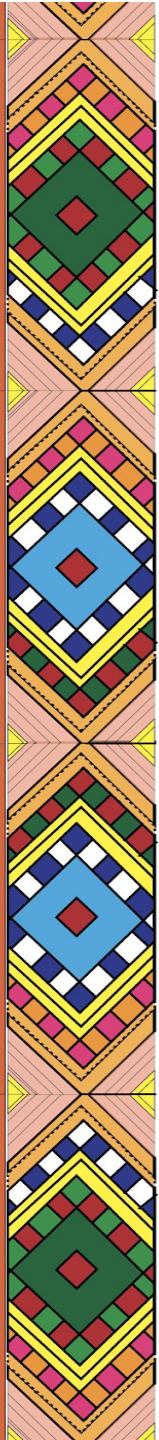
Hashing



YTU56ASF

DSAFDSAF

YTU56ASF



Exact Duplciate Detection

```
# import the required fuctions from the library.
from dotlibrary.exact_duplicate_functions import HashPages, threshold_by_percent, threshold_by_number_of_matched_pages
# run HashPages Function to get the duplicate info on file level, a datafmae for visualization and a dataframe with representative pages selected for processing.
file_level_df, reduced_df=HashPages(dataset_path="/home/alice/wonderland dataset/")

#set conditions for filtering exact duplciates
condition_1=threshold_by_percent(file_level_df, min_page_in_file=2, max_page_in_file=20,min_percentage_value=40)
condition_2=threshold_by_number_of_matched_pages(file_level_df, max_page_in_file=10, max_page_in_file_match=5)

# import functions for visualizing
from dotlibrary.exact_duplicate_functions import visualize_file_pairs, print_duplicate_info

print_duplicate_info([condition_2])

path_1='/home/alice/wonderland dataset/pages/main.pdf'
path_2='/home/alice/wonderland dataset/pages/wolf_11.pdf'
file_a, file_b=visualize_file_pairs(path_1, path_2)
```

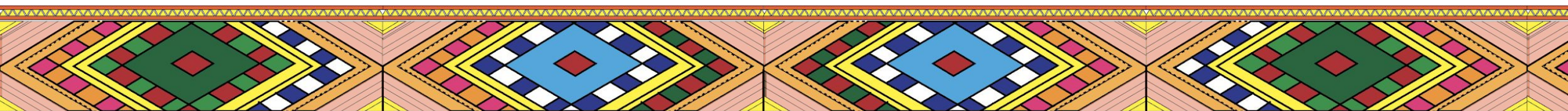
[1]

Problem Statement

PDF files have different types of pages such as forms, narratives, interviews, and letters.

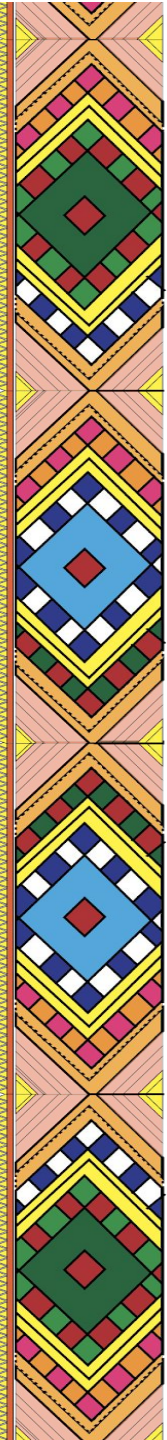
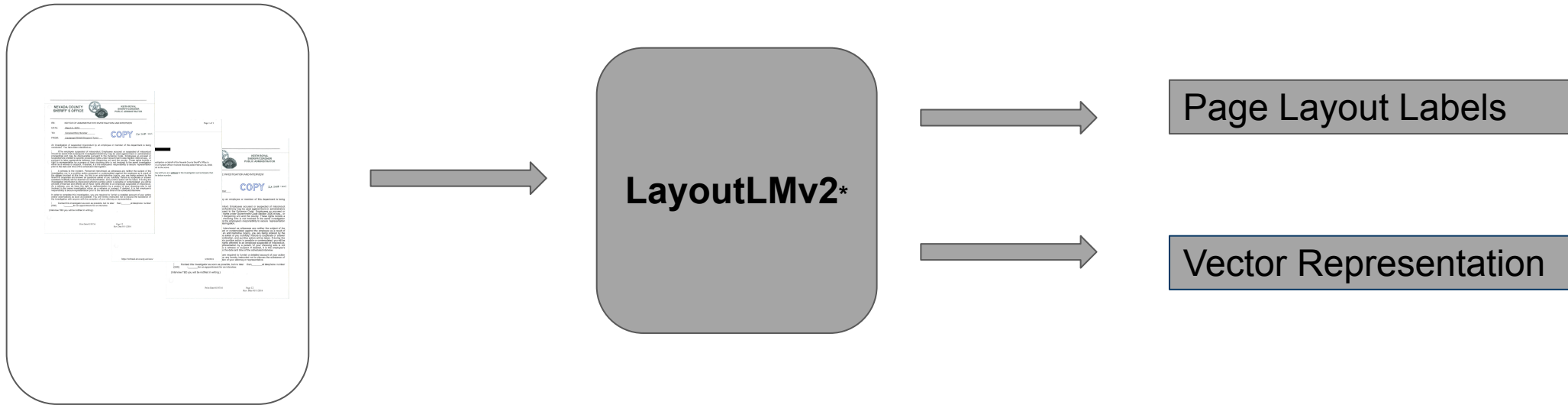
Apply different strategy for different layouts

Sometimes they are only interested in one format.



Methodology

Page Type Classification



Methodology

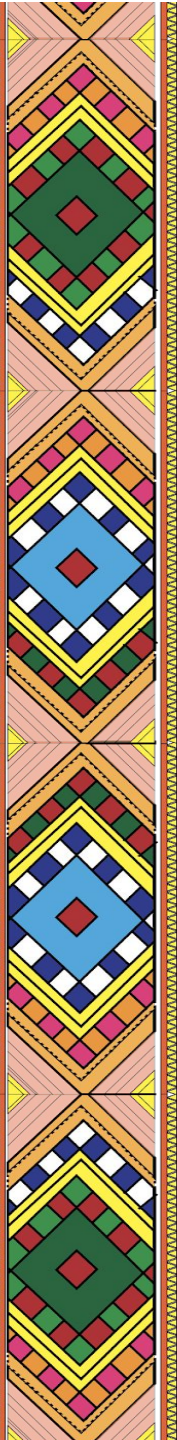
Page Type Classification

Fine Tuning with 6k pages manually annotated:

- Training Accuracy: 83.78%
- Validation Accuracy: 84.11%
- Test Accuracy: 83.52%

Time for training: around 6 hours.

Time for labeling: around 6 hours.



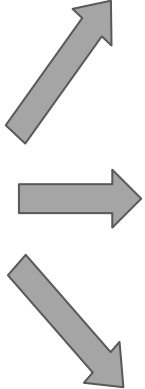
Methodology

Near Duplicate Detection

Vector Embeddings
Page0, Page1,
Page2, Page3,
Page4, Page5,
Page6, Page7,
Page8



K-means Clustering

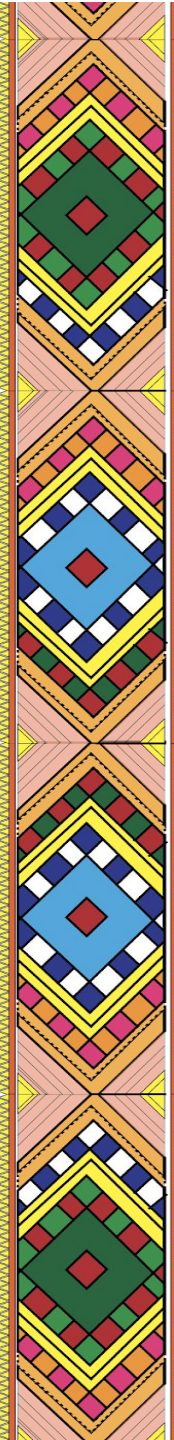


Cluster 1
Page2, Page6,
Page7

Cluster 2
Page0, Page4,
Page8

Cluster 3
Page1, Page3,
Page5

Cluster 1
Corr(Page2, Page6)=0.6
Corr(Page2, Page7)=0.2
Corr(Page6, Page7)=0.9



Near Duplicate Detection

```
# import the classifier and pair correlation function and pass the dataset path along with the name of
# the page types we want to find near duplciates for.
from dotlibrary.near_duplicate_detection_functions import ClassifyAndGetPairCorrelation
correlation_df=ClassifyAndGetPairCorrelation('/home/alice/wonderland dataset/', ['form', 'image', 'narrative'])

from dotlibrary.near_duplicate_detection_functions import plot_pairs_of_pages
plot_pairs_of_pages(correlation_df, ['form'], 10, threshold=0.6, random=True, sort_acending=False)
plot_pairs_of_pages(correlation_df, ['image'], 3 ,threshold=0, random=True, sort_acending=False)

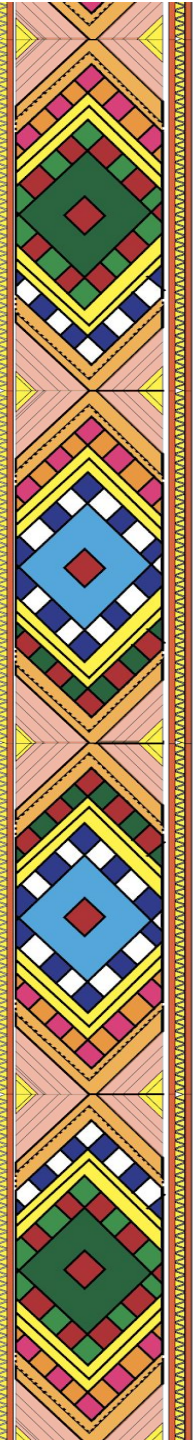
from dotlibrary.near_duplicate_detection_functions import set_correlation_threshold, print_near_duplciate_information
filtered_df=set_correlation_threshold(correlation_df, {'form': 0.9, 'image': 0.77})

print_near_duplciate_information(filtered_df)

from dotlibrary.near_duplicate_detection_functions import visualize_file_pairs
sample_8, sample_14=visualize_file_pairs('sample 8.pdf', 'sample 14.pdf')
sample_8
```

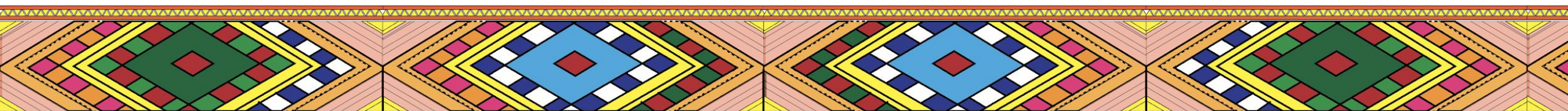
[]

Methodology



Problem Statement

Different page types and institution formats result in difference of how data points are extracted.



Challenges

Data Extraction

Administrative Information							
Agency	DR	Supplement No	Reported Date	Reported Time			
STOCKTON POLICE DEPARTMENT	15-24439	0003	08/05/2015	10:32			
CAD Call No	Status	Rpt/Incident Typ					
151820168	RTF INVESTIGATIONS	LEWD AND LASCIVIOUS CONDUCT					
[REDACTED]							City
[REDACTED]							Stockton
ZIP Code	Rep Dist	District	Sector	From Date	From Time	To Date	To Time
95212	0363	VA	VS	05/28/2015	20:00	05/28/2015	20:00
Member#	Assignment						
2610/ Doe, John	CHILD ABUSE/SEXUAL ASSAULT PHASE 1						
Entered By	Assignment	RMS Transfer	Prop Trans Stat				
2610	CHILD ABUSE/SEXUAL ASSAULT PHASE 1	Successful	Successful				
Approving Officer	Approval Date	Approval Time					
1357	08/07/2015	09:06:43					

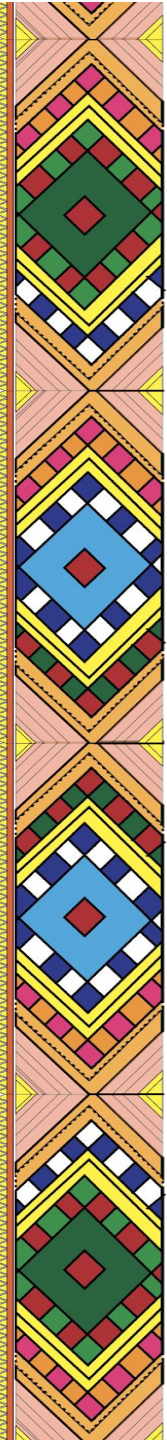
Data Points of Interest

- Names of officers
- Date of Incidence
- Case Number
- Location

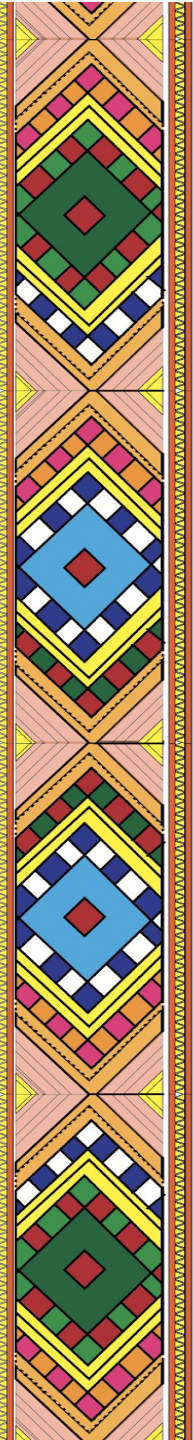
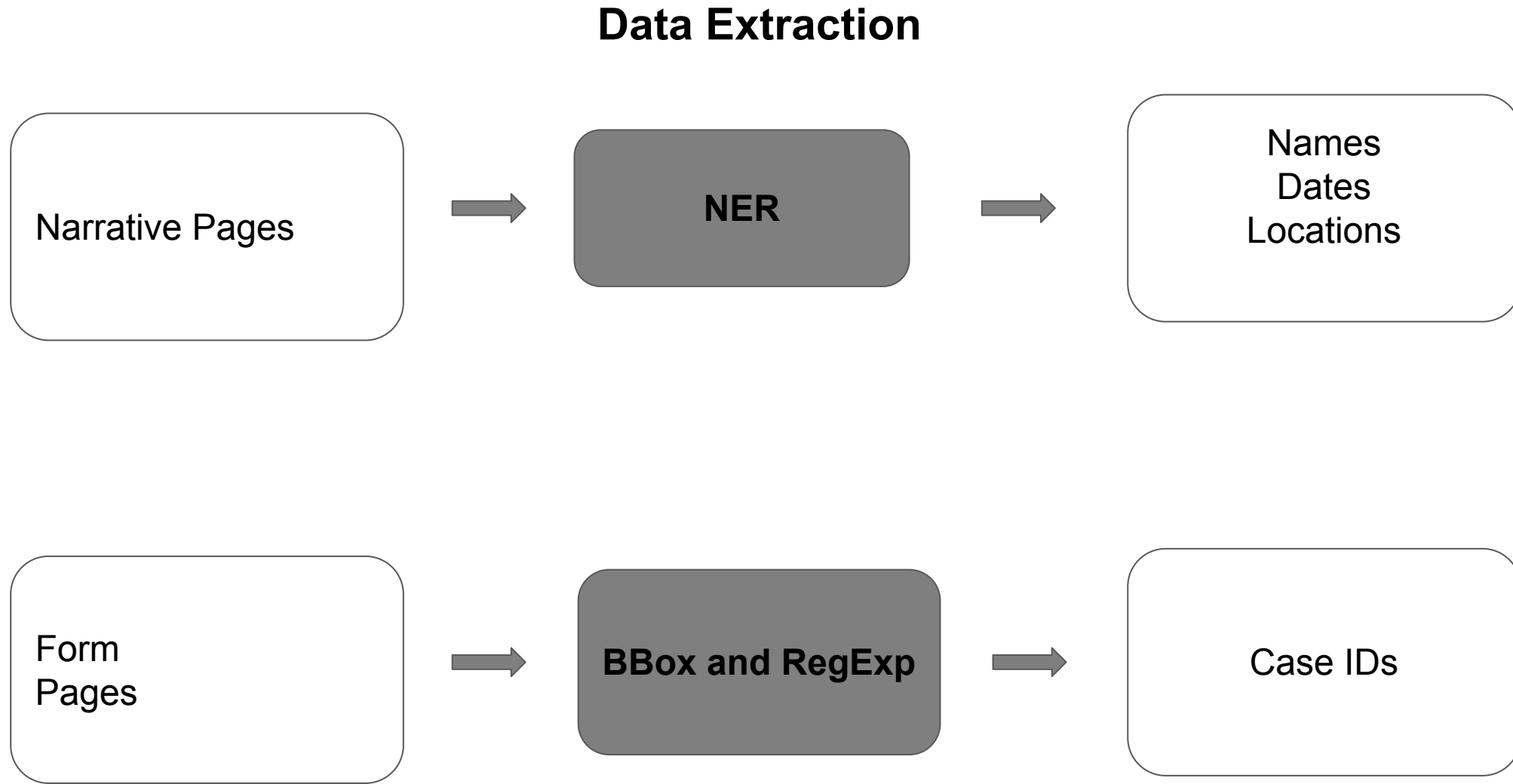
On August 4, 2015 I went back to [REDACTED] address but did not locate [John's] vehicle. I called [REDACTED]. [REDACTED] advised me that [John] was not staying with her but that she had seen the week prior. She said she did not know where he was and did not know what was going on between him and [REDACTED]. She said she did not want to know what was going on. I told her to have [John] come to the police department to speak with me.

When I returned to the office I received a phone call from a man who identified himself as [John]. [John] gave me the following summarized statement:

The following interview was digitally recorded and has been summarized below. The purpose of the summary is to render an overview of what was described; however, they are not transcriptions of the recorded interview. The summary should not be viewed as containing only the important facts of the interview. They are intended to provide the reader with a general understanding of what was said; therefore, refer to the recording itself for the exact wording.



Methodology



Data Extraction

```
from dotlibrary.data_extraction.functions import extract_with_bbox, clean_entity_list
listOfNames=get_named_entites(extracted_df['Text'], 'PERSON')
listOfDates=get_named_entites(extracted_df['Text'], 'DATE')
listOfLocations=get_named_entites(extracted_df['Text'], 'LOCATION')
listOfCaseNumbers= extract_with_bbox([100, 700, 200, 400], encoded_dataset, processor, 'form')

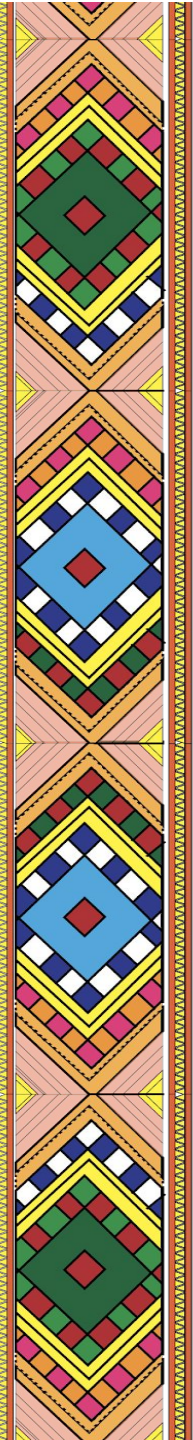
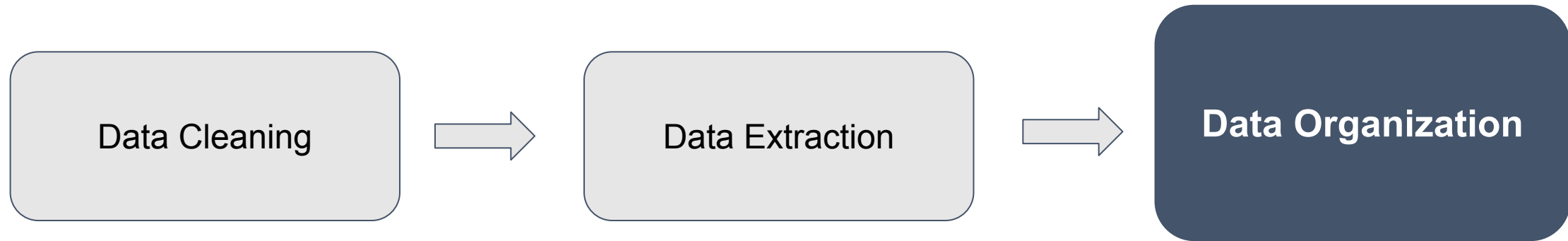
cleanNames=clean_entity_list(cleanNameList, listOfNames)
cleanDates=clean_entity_list(cleanDateList, listOfDates)
cleanLocations=clean_entity_list(['drive', 'highway', 'state', 'park', 'city'], listOfLocations, minLen=15, maxLen=100)

set_entity_breakdown(extracted_df, 'CaseNumber', listOfCaseNumbers)
set_entity_whole(extracted_df, 'Location', cleanLocations)
set_entity_whole(extracted_df, 'Date', cleanLocations)
set_entity_whole(extracted_df, 'Names', cleanLocations)

set_entities_doc(extracted_df, doc_data, 'CaseNumber')
set_entities_doc(extracted_df, doc_data, 'Names')
set_entities_doc(extracted_df, doc_data, 'Date')
set_entities_doc(extracted_df, doc_data, 'Location')
```

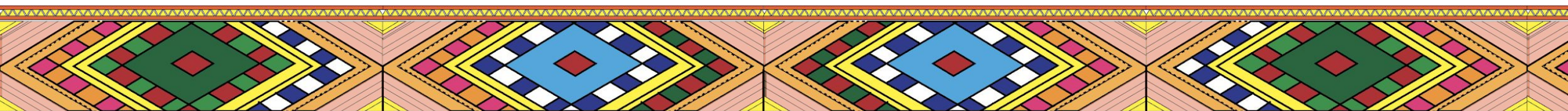
	Names	CaseNumber	Date	Location	File	Text
9	[1s13, address, admin leave, al garza, aldo se...	[01-07-001014, CR07-2908, VCM097845]	[05/03/07, 05/03/2007, 05/07/2007, 05/09/2007]	[1515 east 4th street benicia, apartment build...	Sevens_2007_Redaced3100001	CAD Operations Report SOLANO COUNTY DISPATCH C...
15	[1s13, address, anthony, anthony evans, anthon...	[0227308, CR04-8569, CR04-8969]	[01/02/2003, 05/08/2003, 11/28/2004]	[12 sandy beach vallejo, sandy beach road, san...	Samle_2004_Redaced260001	CONFIDENTIAL DEPARTMENT OF MOTOR VEHICLES DMV...
25	[autry, bryan glenn, citizen, cont, denton, en...	[17000412]	[02/26/2017]	[solano county sheriff]	Kennedy_Ble_Team_(Redaced)10001	11 12 2019 Incident Summary FRONT BACK 1 A 2 ...
7	[address, appel, bigarani, chase, citizen, con...	[17000412, CR17-1007]	[02/26/17, 02/26/2017]	[01 jcdf booking housing, 01 jcdf booking hous...	KENNEDY_Michael_(Redaced)00001	POBLETE SCJ436 OLANO COUNTY SHERIFF S OFFI 02...
23	[address, anthony, anthony mah, arabia, autry, ...	[18-0251]	[01/17/2018]	[1955 grande circle, solano county sheriff]	Mah_Ble_Team_(Redaced)00001	11 12 2019 Incident Summary Pursuit Distance T...
21	[address, anthony, anthony mah, arabia, caller...	[18-0251, CR18-0251]	[01/17/2018, 2018/01/17]	[1955 grande circle, e80 // leisure town, sola...	Mah_Anhny_Redaced_+_CC10001	CAD Operations Report 155 SOLANO COUNTY SHERIF...
3	[address, arabia, autry, citizen, cont, demare...	[18-3027]	[]	NaN	K9_Use_f_Frce_R_06.05.18_(redaced)1_Redaced10001	Page 1 Use of Force Control FN2018 0105 Re...
1	[address, alex sanchez, alfred smith, arabia, ...	[18-3027, CR18-3027]	[06/05/2018, 06/07/18, 09/18/2018, 12/10/18, 2...	[kilkeny byrnes vacaville california 95688, k...	Dran-Garcia_Salvadr_Redaced+CC00001	SOLANO COUNTY SHERIFF S OFFICE Page C CA0480...
10	[chapman, cont, gojkovich, kamman, left, offic...	[18002415]	[12/05/2018, 12/06/2018]	[01 jcdf booking male dressout, solano county ...	Raygza_Lis10001	KAMMAN SCJ311 SOLANO COUNTY SHERIFF S OFFICE 1...
18	[bill elbert, bradley, citizen, cont, elbert, ...	[18002415]	[12/05/2018]	[solano county sheriff]	Raygza_Ble_Team_Redaced00001	FRONT BACK 1 A 2 7 1 C 2 3 C 3 8 4 H 4 D CO...
26	[address, brian, brook byerley, bryan braker, ...	[849-1497-eCI-228, 899-1497-CCI-228, 899-7497-...	[08/20/1999, 08/20/99, 08/23/99, 08/24/99]	[2401 port street, solano county sheriff]	899-1497-CCI-228_Clark_Dglas_Preliminary_Rer_R...	Coroner and Public Administrator SOLANO JAMES ...
2	[bryan braker, clark, doug, douglas, fairfield...	[899-1497-CCI-228]	[06/23/99, 08/20/99]	NaN	899-1497-CCI-228_Clark_Dglas_Release_Frm_Orig...	BRYAN BRAKER FUNERAL HOME AND FUNERAL DIRECTOR...
13	[clark, doug, douglas, fairfield, gary faulkner]	[899-1497-CCI-228]	[08/20/99]	NaN	899-1497-CCI-228_Clark_Dglas_Fingerrin_Cmaris...	AUTOMATED LATENT PRINT SECTION 530 Union Avenu...
27	[clark, cont, doug, douglas, locker, solano, s...	[899-1497-CCT-228]	[08/20/99, 08/23/99]	NaN	899-1497-CCI-228_Clark_Dglas_Mrge_Cnrl_Frm_Or...	SOLANO COUNTY CORONER S OFFICE JAMES E O BRIE...
12	[1154, 1s13, address, angry, arabia, baffico, ...	[AI99-50, CR99-5601, CR99-5601 (DA 99-205)]	[08/10/19, 08/20/99, 08/21/99, 08/23/99, 08/24...	[5053 noonan lane fairfield, 5053 noonan lane f...	Clark_1999_Redaced350001	COPIES TO SOLANO COUNTY SHERIFF S OFFICE CASE...
6	[1s13, 94571, aaron dillon, aaron dillon page, ...	[CR02-4759, CR12-4759]	[10/13/2012, 10/15/2012, 10/16/2012, 10/18/12...	[collins ville rdc fire house rd, fire truck r...	Clinge_2012_Redaced340001	CA04800 SOLANO COUNTY SHERIFF S OFFICE Page 5 ...
8	[1154, aaron dillon, aaron dillon page, adante...	[CR12-0689]	[02/01/1300, 02/12/12, 02/14/12, 02/15/12, 02/...	[apartment building, ronpway 1fo 900 pase, sol...	Elber_2012_Redaced280001	CONFIDENTIAL DEPARTMENT OF MOTOR VEHICLES DMV...
24	[arabia, autry, christopher cavazos, citizen, ...	[CR18-3027]	[06/05/2018]	[solano county sheriff]	Dran-Garcia_Ble_Team_(Redaced)20001	11 12 2019 Incident Summary Officer Assessment...
17	[anthony, anthony kasper, autry, bush, charles...	[CR18-4574]	[09/03/2018]	[solano county sheriff]	Threadgill_Ble_Team_(Wrking_Cy)00001	11 12 2019 Incident Summary FRONT BACK 1 A 2 ...
20	[anthony, autry, bush, charles, charles thread...	[CR18-4574]	[]	NaN	K9_Use_f_Frce_R_09.03.18_(redaced)1_Redaced10001	Page 3 and taken to surgery Threadgill was s...
16	[address, anthony, bush, charles, charles thre...	[CR18-4574]	[09/03/18]	[solano county sheriff]	Threadgill_Charles_Redaced_+CC20001	CONFIDENTIAL DEPARTMENT OF MOTOR VEHICLES DMV...
11	[address, angulo, app02 ps, arabia, axon body, ...	[CR18-6148]	[10/18/12, 12/10/18, 12/10/2018, 2018/12/10]	[solano county sheriff]	Angl_Ral_Redaced+CC00001	SOLANO COUNTY SHERIFF 911 COMMUNICATIONS Call ...
14	[angulo, arabia, autry, axon body, bush, chase...	[CR18-6148]	[12/10/2018]	[solano county sheriff]	Angl_Ble_Team_(Redaced)10001	11 12 2019 Incident Summary Solano County Sher...
0	[bryan braker, carolyn clark, charles, charles...	NaN	[08/20/1999, 08/23/1999]	NaN	899-1497-CCI-228_Clark_Dglas_Deah_Cerificae_R...	Aug 24 99 03 08P Bryan Braker F H 7074257352 ...
4	[anthony, chase, citizen, cont, fairfield, fer...	NaN	[]	[solano county sheriff]	Jseh_Michael_12-24-2007_redaced120001	Alternatively you may appeal the final decisi...
5	[brian, brian peterson, clark, doug, douglas, ...	NaN	[]	NaN	899-1497-CCI-228_Clark_Dglas_Saemen_f_Fac_Orig...	STATE OF CALIFORNIA ss County of Solano In L...
19	[clark, cont, cover, doug, douglas, douglas cl...	NaN	[]	[solano county sheriff]	899-1497-CCI-228_Clark_Dglas_Case_Invenry_Orig...	SOLANO COUNTY SHERIFF CORONER S OFFICE CORONER...
22	[chase, citizen, cont, fairfield, ferrara, gar...	NaN	[]	[solano county sheriff]	Jseh_Michael_11-30-2007_redaced140001	DC RECEIVED JF Gary R Stanton 707 42 1 70...
28	[chase, citizen, cont, fairfield, ferrara, gar...	NaN	[]	[solano county sheriff]	Jseh_Michael_12-04-2007_redaced20001	DC JF a Gary R Stanton 707 42 1 7000 T She...

Methodology



Problem Statement

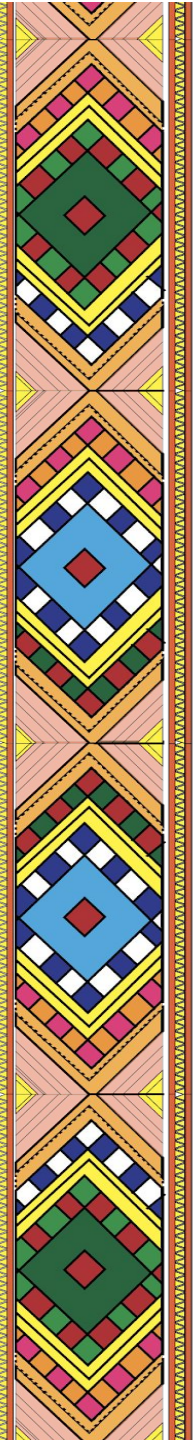
A particular case might be spread across several PDFs or multiple cases might be together in one PDF.



Methodology

Data Organization

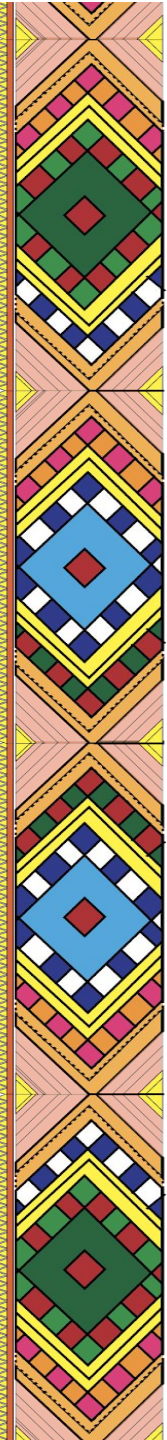
	Names	CaseNumber	Date	Location	File	Text
9	[1s13, address, admin leave, al garza, aldo se...	[01-07-001014, CR07-2908, VCM097845]	[05/03/07, 05/03/2007, 05/07/2007, 05/09/2007]	[1515 east 4th street benicia, apartment build...	Sevens_2007_Redaced3100001	CAD Operations Report SOLANO COUNTY DISPATCH C...
15	[1s13, address, anthony, anthony evans, anthon...	[0227308, CR04-8569, CR04-8969]	[01/02/2003, 05/08/2003, 11/28/2004]	[12 sandy beach vallejo, sandy beach road, san...	Samle_2004_Redaced260001	CONFIDENTIAL DEPARTMENT OF MOTOR VEHICLES DMV...
25	[autry, bryan glenn, citizen, cont, denton, en...	[17000412]	[02/26/2017]	[solano county sheriff]	Kennedy_Ble_Team_(Redaced)10001	11 12 2019 Incident Summary FRONT BACK 1 A 2 ...
7	[address, appel, bigarani, chase, citizen, con...	[17000412, CR17-1007]	[02/26/17, 02/26/2017]	[01 jcdf booking housing, 01 jcdf booking hous...	KENNEDY_Michael_(Redaced)00001	POBLETE SCJ436 OLANO COUNTY SHERIFF S OFFI 02...
23	[address, anthony, anthony mah, arabia, autry...	[18-0251]	[01/17/2018]	[1955 grande circle, solano county sheriff]	Mah_Ble_Team_(Redaced)00001	11 12 2019 Incident Summary Pursuit Distance T...
21	[address, anthony, anthony mah, arabia, caller...	[18-0251, CR18-0251]	[01/17/2018, 2018/01/17]	[1955 grande circle, e80 // leisure town, sola...	Mah_Anhny_Redaced_+_CC10001	CAD Operations Report 155 SOLANO COUNTY SHERIF...
3	[address, arabia, autry, citizen, cont, demare...	[18-3027]	[]	NaN	K9_Use_f_Frce_R_06.05.18_(redaced)1_Redaced10001	Page 1 Use of force Control FN2018 0105 Re...
1	[address, alex sanchez, alfred smith, arabia, ...	[18-3027, CR18-3027]	[06/05/2018, 06/07/18, 09/18/2018, 12/10/18, 2...	[kilkenny byrnes vacaville california 95688, k...	Dran-Garcia_Salvadr_Redaced+CC00001	SOLANO COUNTY SHERIFF S OFFICE Page c CA0480...
10	[chapman, cont, gojkovich, kamman, left, offic...	[18002415]	[12/05/2018, 12/06/2018]	[01 jcdf booking male dressout, solano county ...	Raygza_Lis10001	KAMMAN SCJ311 SOLANO COUNTY SHERIFF S OFFICE 1...
18	[bill elbert, bradley, citizen, cont, elbert, ...	[18002415]	[12/05/2018]	[solano county sheriff]	Raygza_Ble_Team_Redaced00001	FRONT BACK 1 A 2 7 1 C 2 3 C 3 B 4 H 4 D CO...
26	[address, brian, brook byerley, bryan braker, ...	[849-1497-eCI-228, 899-1497-CCI-228, 899-7497-...	[08/20/1999, 08/20/99, 08/23/99, 08/24/99]	[2401 port street, solano county sheriff]	899-1497-CCI-228_Clark_Dglas_Preliminary_Rer_R...	Coroner and Public Administrator SOLANO JAMES ...
2	[bryan braker, clark, doug, douglas, fairfield...	[899-1497-CCI-228]	[06/23/99, 08/20/99]	NaN	899-1497-CCI-228_Clark_Dglas_Release_Frm_Orig...	BRYAN BRAKER FUNERAL HOME AND FUNERAL DIRECTOR...
13	[clark, doug, douglas, fairfield, gary faulkner]	[899-1497-CCI-228]	[08/20/99]	NaN	899-1497-CCI-228_Clark_Dglas_Fingerrin_Cmaris...	AUTOMATED LATENT PRINT SECTION 530 Union Avenu...
27	[clark, cont, doug, douglas, locker, solano, s...	[899-1497-CCT-228]	[08/20/99, 08/23/99]	NaN	899-1497-CCI-228_Clark_Dglas_Mrge_Cnrl_Frm_Or...	SOLANO COUNTY CORONER S OFFICE JAMES E O BRIE...
12	[1154, 1s13, address, angry, arabia, baffico, ...	[AI99-50, CR99-5601, CR99-5601 (DA 99-205)]	[08/10/19, 08/20/99, 08/21/99, 08/23/99, 08/24...	[5053 noonan lane faifield, 5053 noonan lane f...	Clark_1999_Redaced350001	COPIES TO SOLANO COUNTY SHERIFF S OFFICE CASE...
6	[1s13, 94571, aaron dillon, aaron dillon page...	[CR02-4759, CR12-4759]	[10/13/2012, 10/15/2012, 10/16/2012, 10/18/12...	[collins ville rdc fire house rd, fire truck r...	Cllinge_2012_Redaced340001	CA04800 SOLANO COUNTY SHERIFF S OFFICE Page 5 ...
8	[1154, aaron dillon, aaron dillon page, adante...	[CR12-0689]	[02/01/1300, 02/12/12, 02/14/12, 02/15/12, 02/...	[apartment building, ronpway 1fo 900 pase, sol...	Elber_2012_Redaced280001	CONFIDENTIAL DEPARTMENT OF MOTOR VEHICLES DMV...
24	[arabia, autry, christopher cavazos, citizen, ...	[CR18-3027]	[06/05/2018]	[solano county sheriff]	Dran-Garcia_Ble_Team_(Redaced)20001	11 12 2019 Incident Summary Officer Assessment...
17	[anthony, anthony kasper, autry, bush, charles...	[CR18-4574]	[09/03/2018]	[solano county sheriff]	Threadgill_Ble_Team_(Wrking_Cy)00001	11 12 2019 Incident Summary FRONT BACK 1 A 2 ...
20	[anthony, autry, bush, charles, charles thread...	[CR18-4574]	[]	NaN	K9_Use_f_Frce_R_09.03.18_(redaced)1_Redaced10001	Page 3 and taken to surgery Threadgill was s...
16	[address, anthony, bush, charles, charles thre...	[CR18-4574]	[09/03/18]	[solano county sheriff]	Threadgill_Charles_Redaced_+CC20001	CONFIDENTIAL DEPARTMENT OF MOTOR VEHICLES DMV...
11	[address, angulo, app02 ps, arabia, axon body...	[CR18-6148]	[10/18/12, 12/10/18, 12/10/2018, 2018/12/10]	[solano county sheriff]	Angl_Ral_Redaced+CC00001	SOLANO COUNTY SHERIFF 911 COMMUNICATIONS Call ...
14	[angulo, arabia, autry, axon body, bush, chase...	[CR18-6148]	[12/10/2018]	[solano county sheriff]	Angl_Ble_Team_(Redaced)10001	11 12 2019 Incident Summary Solano County Sher...
0	[bryan braker, carolyn clark, charles, charles...	NaN	[08/20/1999, 08/23/1999]	NaN	899-1497-CCI-228_Clark_Dglas_Deah_Cerifica_R...	Aug 24 99 03 08P Bryan Braker F H 7074257352 ...
4	[anthony, chase, citizen, cont, fairfield, fer...	NaN	[]	[solano county sheriff]	Jseh_Michael_12-24-2007_redaced120001	Alternatively you may appeal the final decis...
5	[brian, brian peterson, clark, doug, douglas, ...	NaN	[]	NaN	899-1497-CCI-228_Clark_Dglas_Saemen_f_Fac_Orig...	STATE OF CALIFORNIA ss County of Solano In t...
19	[clark, cont, cover, doug, douglas, douglas cl...	NaN	[]	[solano county sheriff]	899-1497-CCI-228_Clark_Dglas_Case_Invenry_Orig...	SOLANO COUNTY SHERIFF CORONER S OFFICE CORONER...
22	[chase, citizen, cont, fairfield, ferrara, gar...	NaN	[]	[solano county sheriff]	Jseh_Michael_11-30-2007_redaced140001	DC RECEIVED JF Gary R Stanton 707 42 1 70...
28	[chase, citizen, cont, fairfield, ferrara, gar...	NaN	[]	[solano county sheriff]	Jseh_Michael_12-04-2007_redaced20001	DC JF a Gary R Stanton 707 42 1 7000 T She...



Methodology

Data Organization

	CaseNumber	Date	File
9	['01-07-001014', 'CR07-2908', 'VCM097845']	['05/03/07', '05/03/2007', '05/07/2007', '05/09/2007']	Sevens_2007_Redaced3100001
15	['0227308', 'CR04-8569', 'CR04-8969']	['01/02/2003', '05/08/2003', '11/28/2004']	Samle_2004_Redaced260001
25	['17000412']	['02/26/2017']	Kennedy_Ble_Team_(Redaced)10001
7	['17000412', 'CR17-1007']	['02/26/17', '02/26/2017']	KENNEDY_Michael_(Redaced)00001
23	['18-0251']	['01/17/2018']	Mah_Ble_Team_(Redaced)00001
21	['18-0251', 'CR18-0251']	['01/17/2018', '2018/01/17']	Mah_Anhny_Redaced+_CC10001
3	['18-3027']	[]	K9_Use_f_Frce_R_06.05.18_(redaced)1_Redaced10001
1	['18-3027', 'CR18-3027']	['06/05/2018', '06/07/18', '09/18/2018', '12/10/18', '2018/06/05']	Dran-Garcia_Salvadr_Redaced+CC00001
10	['18002415']	['12/05/2018', '12/06/2018']	Raygza_Lis10001
18	['18002415']	['12/05/2018']	Raygza_Ble_Team_Redaced00001
26	['849-1497-eCI-228', '899-1497-CCI-228', '899-7497-CCT 228', 'CR99-5601']	['08/20/1999', '08/20/99', '08/23/99', '08/24/99']	899-1497-CCI-228_Clark_Dglas_Preliminary_Rer_Redaced00001
2	['899-1497-CCI-228']	['06/23/99', '08/20/99']	899-1497-CCI-228_Clark_Dglas_Release_Frm_Original10001
13	['899-1497-CCI-228']	['08/20/99']	899-1497-CCI-228_Clark_Dglas_FIngerrin_Cmarissn_Rer_Redaced_CC110001
27	['899-1497-CCT-228']	['08/20/99', '08/23/99']	899-1497-CCI-228_Clark_Dglas_Mrge_Cnrl_Frm_Original30001
12	['AI99-50', 'CR99-5601', 'CR99-5601 (DA 99-205)']	['08/10/19', '08/20/99', '08/21/99', '08/23/99', '08/24/99', '08/25/99', '09/01/99']	Clark_1999_Redaced350001
6	['CR02-4759', 'CR12-4759']	['10/13/2012', '10/15/2012', '10/16/2012', '10/18/12', '10/18/16', '12/10/18', '12/28/2010']	Cllinge_2012_Redaced340001
8	['CR12-0689']	['02/01/1300', '02/12/12', '02/14/12', '02/15/12', '02/20/2012', '02/21/12']	Elber_2012_Redaced280001
24	['CR18-3027']	['06/05/2018']	Dran-Garcia_Ble_Team_(Redaced)20001
17	['CR18-4574']	['09/03/2018']	Threadgill_Ble_Team_(Wrking_Cy)00001
20	['CR18-4574']	[]	K9_Use_f_Frce_R_09.03.18_(redaced)1_Redaced10001
16	['CR18-4574']	['09/03/18']	Threadgill_Charles_Redaced+_CC20001
11	['CR18-6148']	['10/18/12', '12/10/18', '12/10/2018', '2018/12/10']	Angl_Ral_Redaced+CC00001
14	['CR18-6148']	['12/10/2018']	Angl_Ble_Team_(Redaced)10001



Methodology

Data Organization

Each **Case** object has:

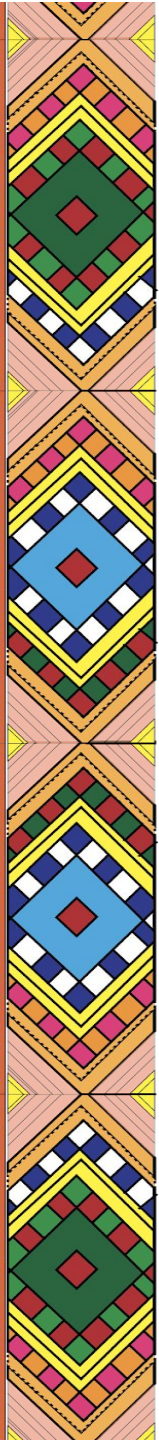
- dates
- files
- names
- locations
- paths

```
case1.dates
```

```
{ '06/23/99',  
  '08/10/19',  
  '08/20/1999',  
  '08/20/99',  
  '08/21/99',  
  '08/23/99',  
  '08/24/99',  
  '08/25/99',  
  '09/01/99' }
```

```
case1.files
```

```
{ '899-1497-CCI-228_Clark_Dglas_Deah_Cerificae_Redaced_CC130001',  
  '899-1497-CCI-228_Clark_Dglas_FIngerrin_Cmarissn_Rer_Redaced_CC110001',  
  '899-1497-CCI-228_Clark_Dglas_Mrge_Cnrl_Frm_Original30001',  
  '899-1497-CCI-228_Clark_Dglas_Release_Frm_Original10001',  
  '899-1497-CCI-228_Clark_Dglas_Case_Invenry_Original90001',  
  '899-1497-CCI-228_Clark_Dglas_Preliminary_Rer_Redaced00001',  
  '899-1497-CCI-228_Clark_Dglas_Saemen_f_Fac_Original70001',  
  'Clark_1999_Redaced350001' }
```

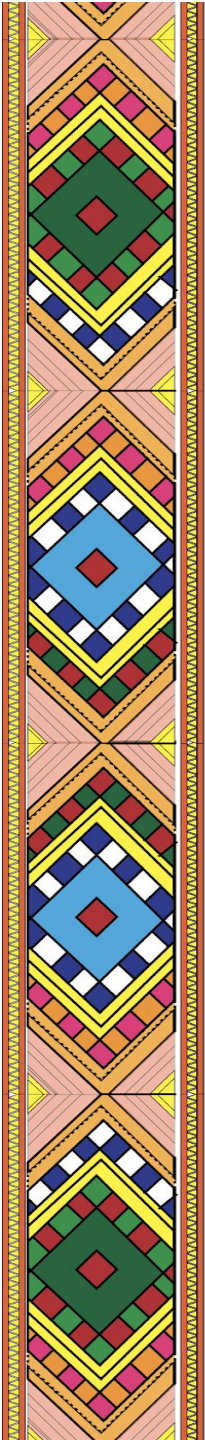


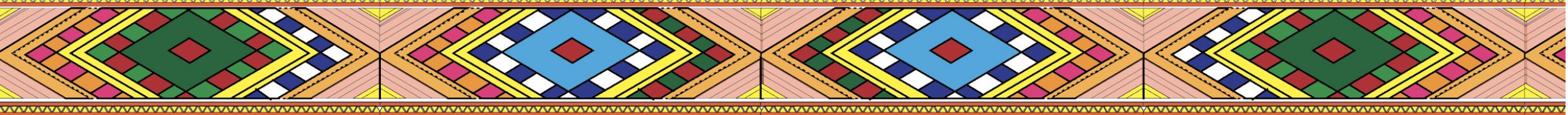
Methodology

Data Organization

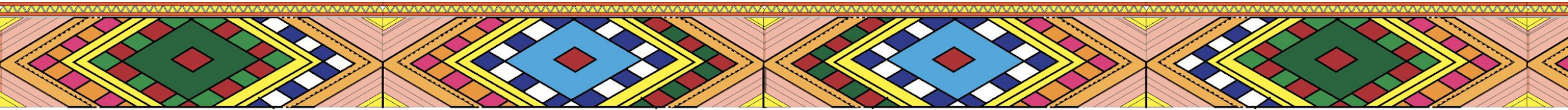
Final output: list of the files belonging to individual cases

```
.. 899-1497-CCI-228_Clark, Dglas_Deah_Cerificae_Redaced_CC130001
899-1497-CCI-228_Clark_Dglas_Case_Invenry_Original90001
Clark_1999_Redaced350001
899-1497-CCI-228_Clark_Dglas_Preliminary_Rer_Redaced00001
899-1497-CCI-228_Clark, Dglas_Mrge_Cnrl_Frm_Original30001
899-1497-CCI-228_Clark_Dglas_Saemen_f_Fac_Original70001
899-1497-CCI-228_Clark, Dglas_Release_Frm_Original10001
899-1497-CCI-228_Clark, Dglas_Fingerrin_Cmarissn_Rer_Redaced_CC110001
*****
Threadgill, Charles_Redaced_+CC20001
Threadgill_Ble_Team_(Wrking_Cy)00001
K9_Use_f_Frce_R_09.03.18_(redaced)1_Redaced10001
*****
Dran-Garcia, Salvadr_Redaced+CC00001
K9_Use_f_Frce_R_06.05.18_(redaced)1_Redaced10001
Dran-Garcia_Ble_Team_(Redaced)20001
*****
KENNEDY, Michael_(Redaced)00001
Kennedy_Ble_Team_(Redaced)10001
*****
Raygza_Lis10001
Raygza_Ble_Team_Redaced00001
*****
Mah, Anhny_Redaced_+_CC10001
Mah_Ble_Team_(Redaced)00001
*****
Angl, _Ral_Redaced+CC00001
Angl_Ble_Team_(Redaced)10001
*****
Sevens_2007_Redaced3100001
*****
Samle_2004_Redaced260001
*****
Cllinge_2012_Redaed340001
*****
Elber_2012_Redaced280001
*****
```



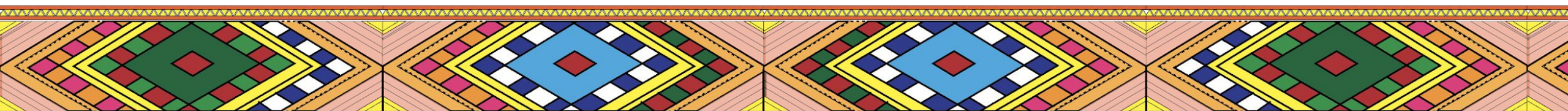


Programming Paradigm Design



Programming Paradigm Design

Some domain experts with limited time and budget became self-taught programmers in an attempt to automate their data organization tasks.



Programming Paradigm Design

DOT Python Library

Visual

PBE

Text-Based

NEAR DUPLICATE PAGE DETECTION

/home/hellina/tutorial/dataset/ Find Near Duplicates

Add Threshold

Set Form Threshold

Set Image Threshold

Set Narrative Threshold

Generate Output

EXACT DUPLICATE PAGE DETECTION

/home/hellina/tutorial/dataset/ Find Exact Duplicates

Doc A Doc B

Percentage Match: 16.00%

Page Ratio: 26:162

Accept Reject

Please give 9 more examples to see generated rules.

jupyter NearDuplicateDetection Last Checkpoint: 08:0:2022 (auto-save)

Python 3 (ipykernel)

Near Duplicate Page Detection

Example

Step 1: Classify the Pages in PDPs based on their type and calculate correlation between pairs of pages.

```
In [1]: # Import the classifier and page correlation function and pass the dataset path along with the name of
# the page pairs to use to find near duplicates. The
from dotlibrary_near_duplicate_detection.functions import ClassifierAndPairCorrelation
correlation_of_page_pairs = correlation_of_page_pairs(path='./dataset', type = 'image', 'narrative')

In [2]: correlation_of_page_pairs()
```

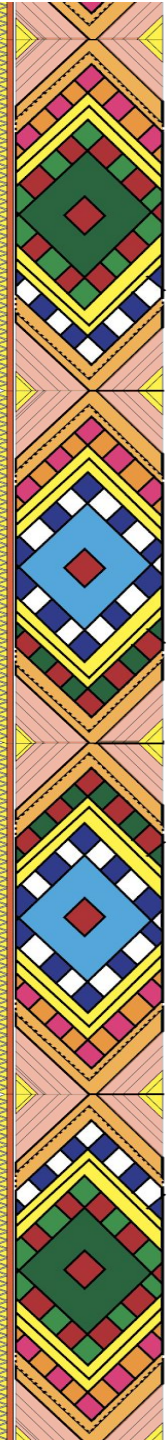
path page 0	path page 1	correlation value	page type
./home/hellina/tutorial/dataset/image_sample_1	./home/hellina/tutorial/dataset/image_sample_2	0.970000	image
./home/hellina/tutorial/dataset/image_sample_1	./home/hellina/tutorial/dataset/image_sample_3	0.974000	image
./home/hellina/tutorial/dataset/image_sample_1	./home/hellina/tutorial/dataset/image_sample_4	0.980000	image
./home/hellina/tutorial/dataset/image_sample_1	./home/hellina/tutorial/dataset/image_sample_5	0.990000	image
./home/hellina/tutorial/dataset/image_sample_1	./home/hellina/tutorial/dataset/image_sample_6	0.976000	image
./home/hellina/tutorial/dataset/image_sample_1	./home/hellina/tutorial/dataset/image_sample_7	0.999000	image
./home/hellina/tutorial/dataset/image_sample_1	./home/hellina/tutorial/dataset/image_sample_8	0.989000	image
./home/hellina/tutorial/dataset/image_sample_1	./home/hellina/tutorial/dataset/image_sample_9	0.992000	image
./home/hellina/tutorial/dataset/image_sample_1	./home/hellina/tutorial/dataset/image_sample_10	0.999000	image

Step 2: Visualize Correlation Value in Pairs and Identify Threshold.

```
In [3]: from dotlibrary_near_duplicate_detection.functions import plot_pairs_of_page
plot_pairs_of_page(correlation_of_page_pairs, 'path='./dataset', x_name='page', y_name='correlation')
```

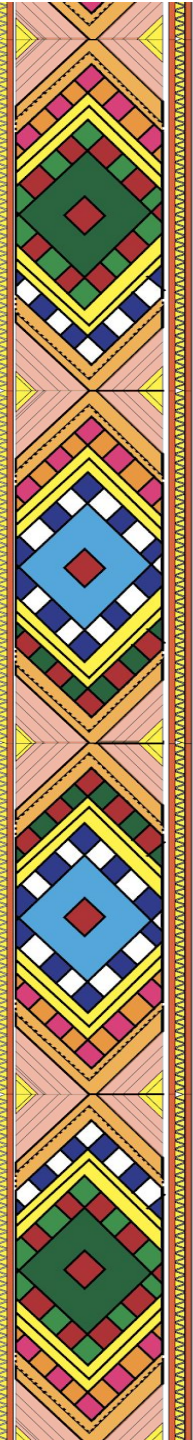
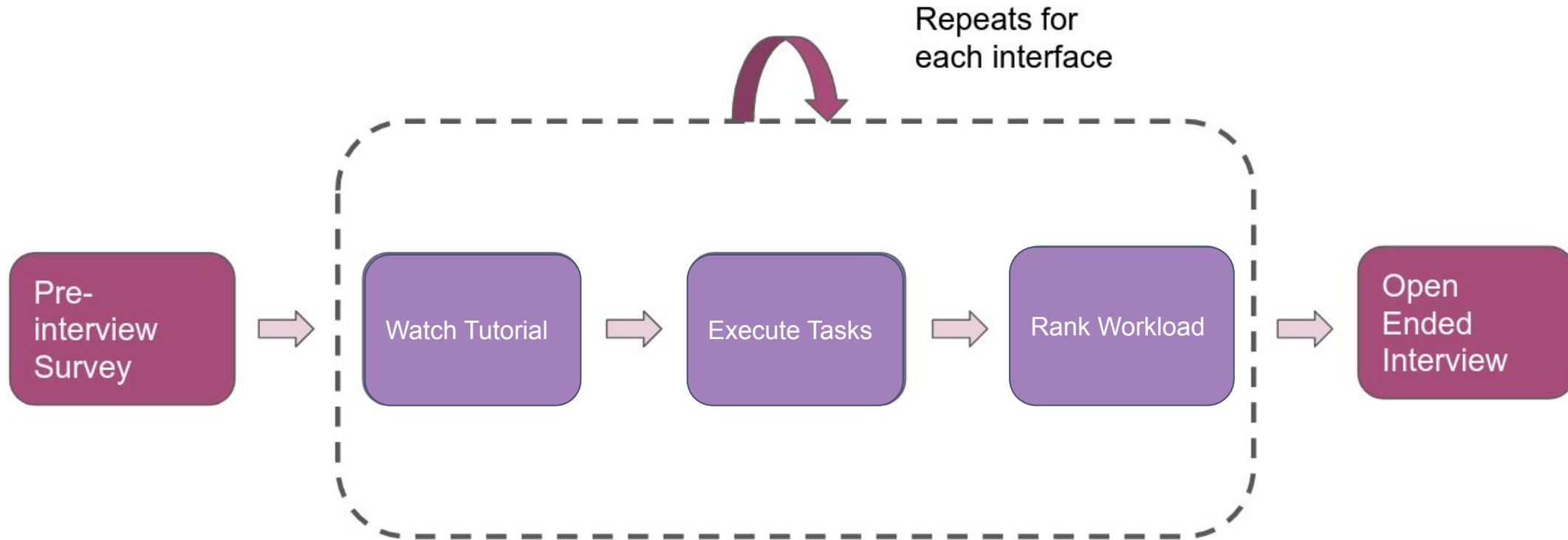
Showing correlation between images...

Central Central



User Study Design

- Within subject study with n=12 participants.



Research Questions

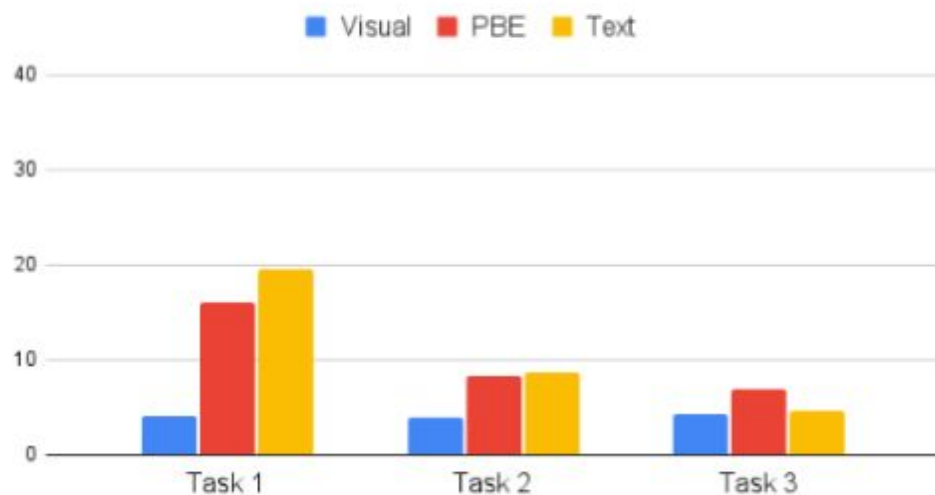
- In the document processing domain, are experts from non-technical domains more successful with Visual, Programming-By-Example or Text-Based programming paradigms?
- For a non-technical audience, what are the relative strengths of each of the programming paradigms?

Results

► RQ1: Are experts from non-technical domains more successful with Visual, Programming-By-Example or Text-Based programming paradigms?

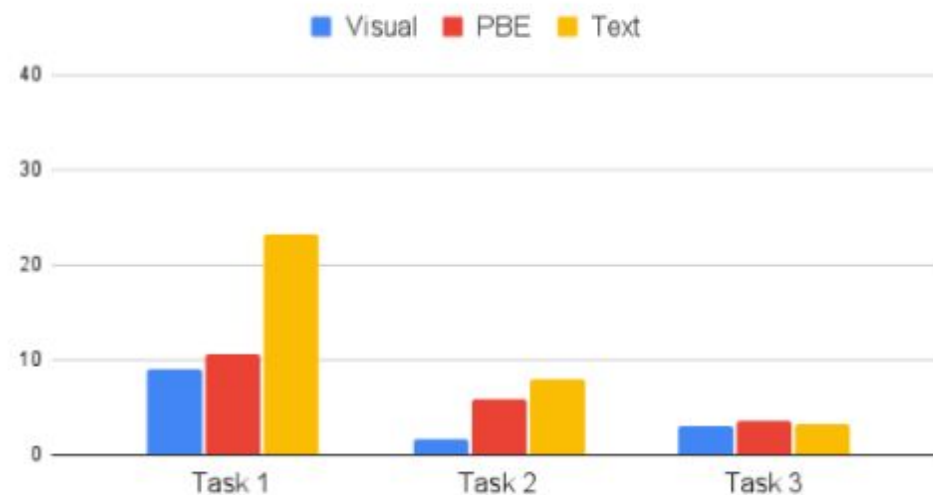
Overall Time

Overall time taken in Exact Duplicate Detection

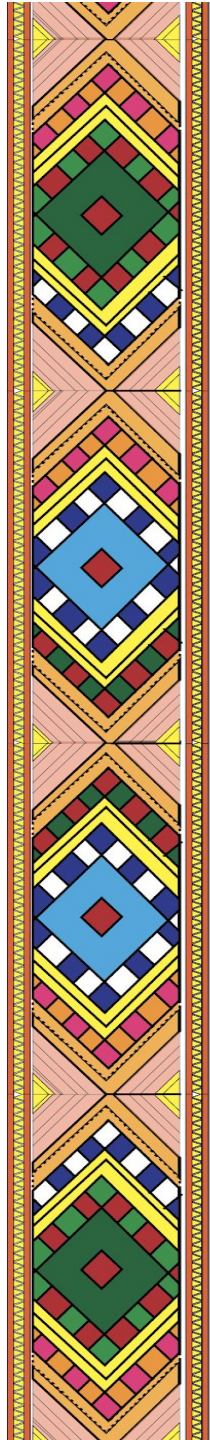


(a) Average time it took participants to complete each of the tasks in Exact Duplicate Detection.

Overall time taken for Near Duplicate Detection



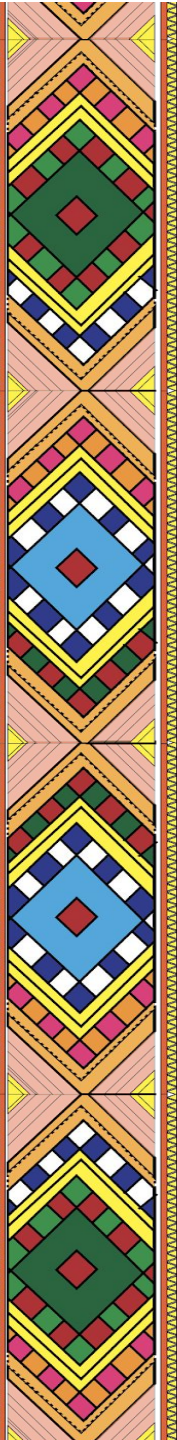
(b) Average time it took participants to complete each of the tasks in Near Duplicate Detection.



Results

▶ **RQ2: For a non-technical audience, what are the relative strengths of each of the programming paradigms?**

- Text-Based programming offers low-level control and flexibility to explore outside of the designer-provided abstractions.
- Visual programming gives tool designers an opportunity to offer information by default that programmers might not think to uncover themselves.
- The PBE paradigm puts the focus on the data rather than the program structure.



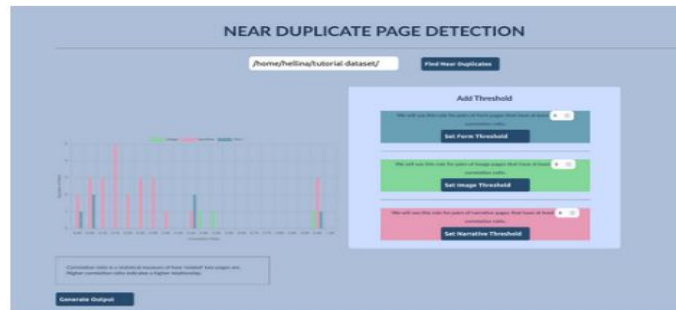
Conclusion

- Domain experts dealing with large document dumps need a data cleaning and data organization tool...We built DOT!



- Domain experts have tried to automate parts of their data organization tasks; so we conducted a formative study to figure out what paradigm works for our set of programmers.

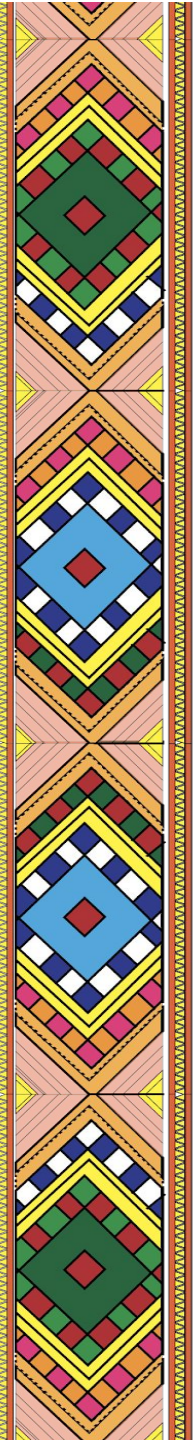
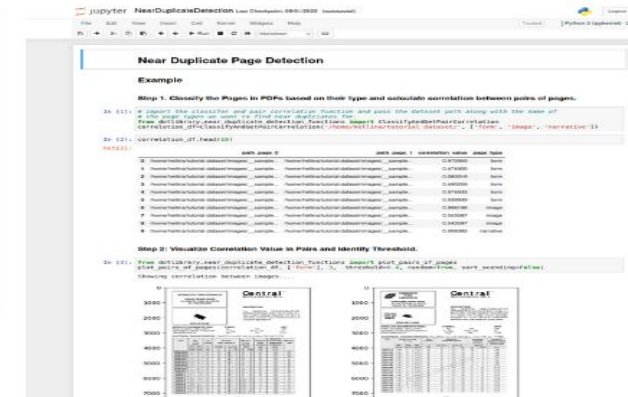
Visual

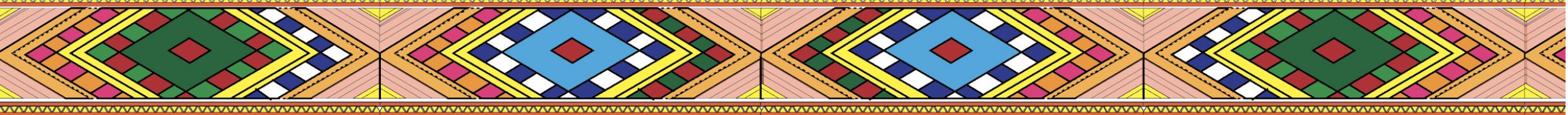


PBE

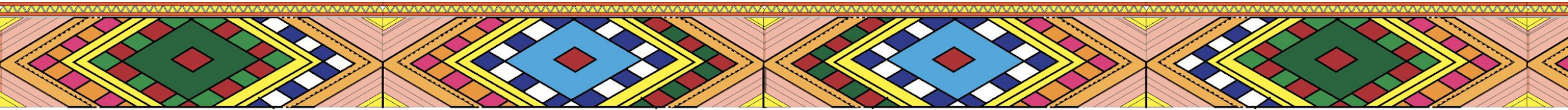


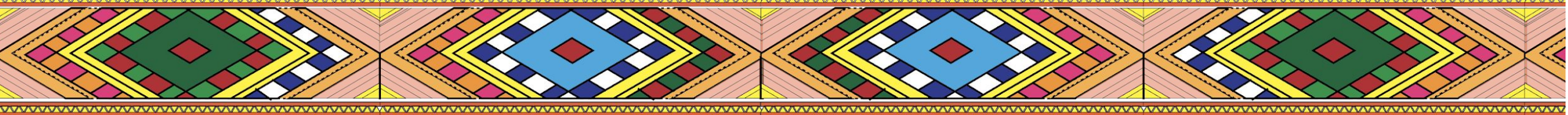
Text-Based



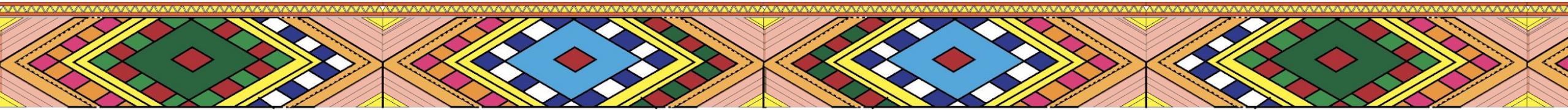


Questions?

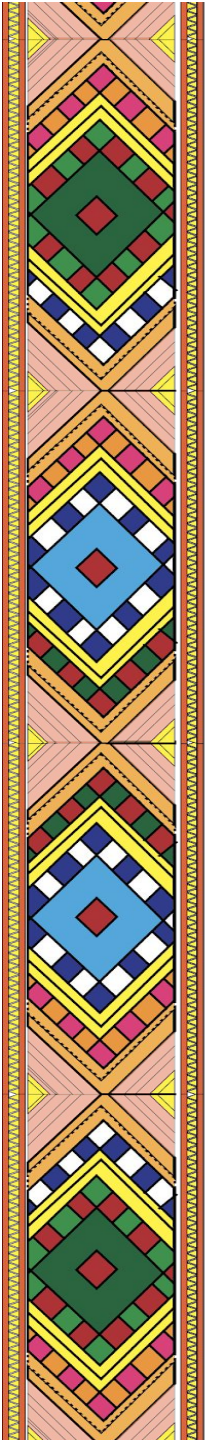
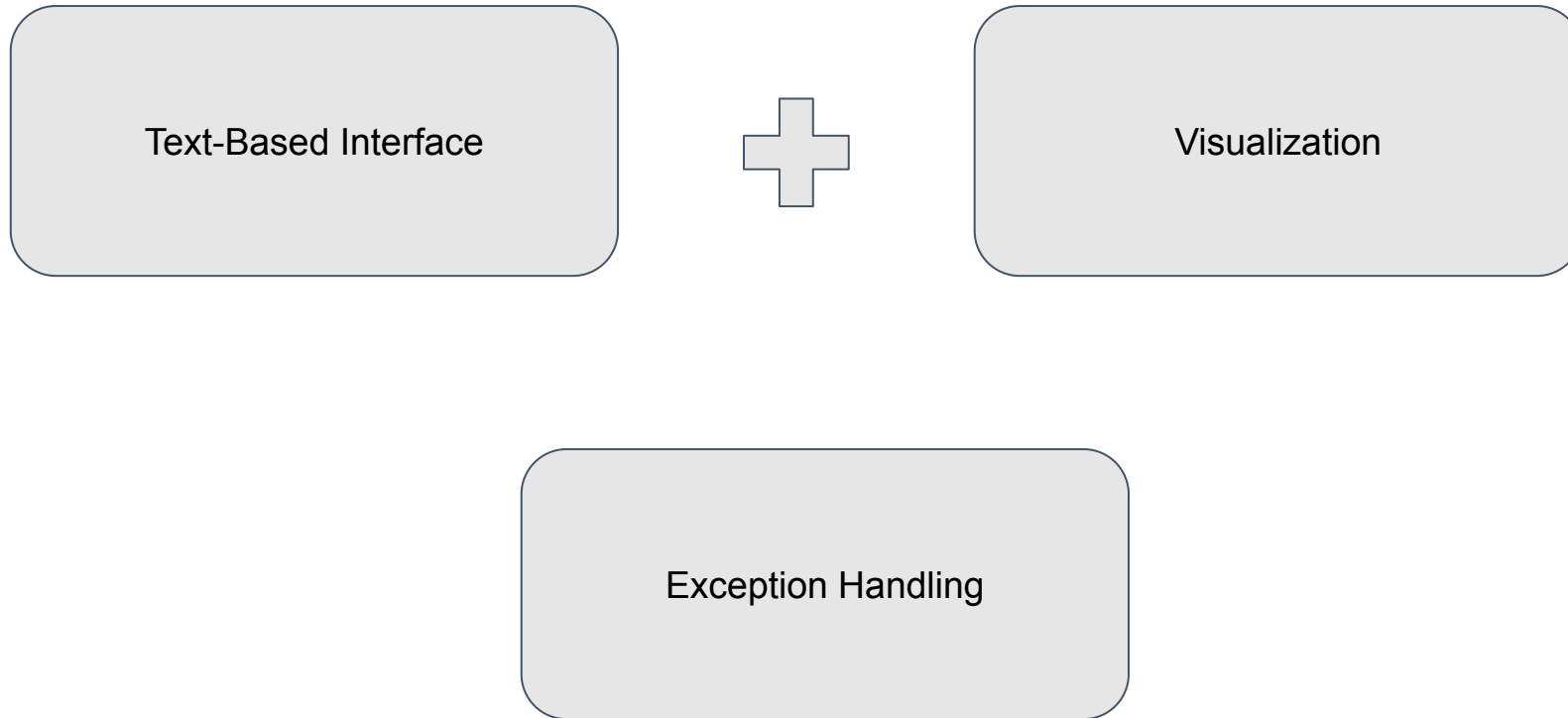




Current Work



Current Work



Methodology

Data Organization

1. File splitting.

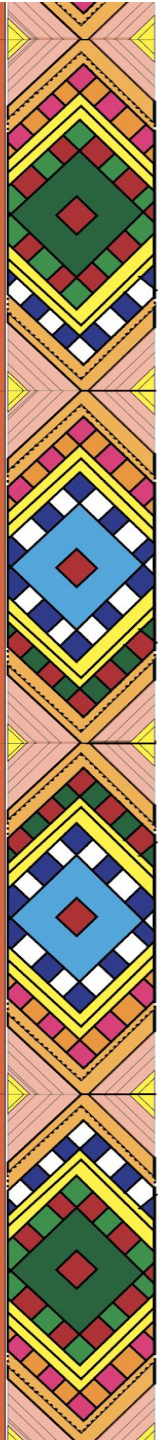
2. File matching with case number.

3. File matching with name and date.

Case Numbers >1?

Select Page Type

Chop PDF at the first
page of the specified
type.



Methodology

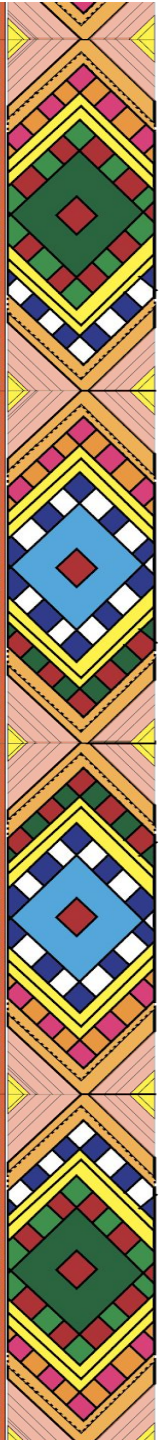
Data Organization

1. File splitting.

2. File matching with case number.

3. File matching with name and date.

8	['CR12-0689']	['02/01/1300', '02/12/12', '02/14/12', '02/15/12', '02/20/2012', '02/21/12']	Elber_2012_Redaced280001
24	['CR18-3027']	['06/05/2018']	Dran-Garcia_Ble_Team_(Redaced)20001
16	['CR18-4574']	['09/03/18']	Threadgill,_Charles_Redaced_+CC20001
17	['CR18-4574']	['09/03/2018']	Threadgill_Ble_Team_(Wrking_Cy)00001
20	['CR18-4574']	[]	K9_Use_f_Frce_R_09.03.18_(redaced)1_Redaced10001
11	['CR18-6148']	['10/18/12', '12/10/18', '12/10/2018', '2018/12/10']	Angl_Ral_Redaced+CC00001
14	['CR18-6148']	['12/10/2018']	Angl_Ble_Team_(Redaced)10001



Methodology

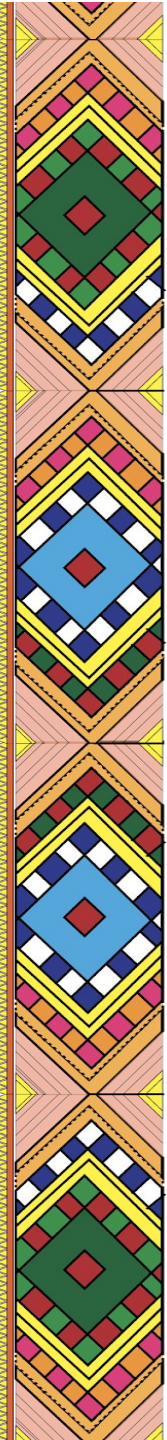
Data Organization

1. File splitting.

2. File matching with case number.

3. File matching with name and date.

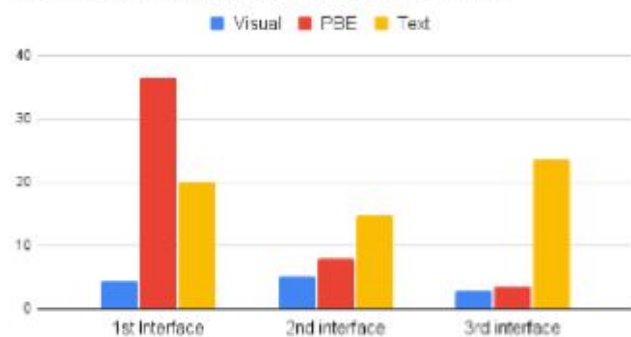
	Names	Date	File
0	['bryan braker', 'carolyn clark', 'charles', 'charles clark', 'clark', 'cont', 'doug', 'douglas', 'douglas clifford', 'fairfield', 'ford', 'jean marcotte', 'knowl 115', 'montana', 'mumber', 'noonan', 'ryan', 'solano', 'typed']	['08/20/1999', '08/23/1999']	899-1497-CCI-228_Clark_Dglas_Deah_Cerifcae_Redaced_CC130001
4	['anthony', 'chase', 'citizen', 'cont', 'fairfield', 'ferrara', 'jose', 'joseph', 'lee axelrad', 'michael', 'sheriffs', 'solano', 'thomas ferrara', 'vallejo']	[]	Jseh_Michael_12-24-2007_redaced120001
5	['brian', 'brian peterson', 'clark', 'doug', 'douglas', 'douglas clifford', 'fairfield', 'ford', 'noonan', 'peterson', 'solano']	[]	899-1497-CCI-228_Clark_Dglas_Saemen_f_Fac_Original70001
19	['clark', 'cont', 'cover', 'doug', 'douglas', 'douglas clark', 'items', 'page', 'respond', 'solano']	[]	899-1497-CCI-228_Clark_Dglas_Case_Invenry_Original90001
22	['chase', 'citizen', 'cont', 'fairfield', 'ferrara', 'gary stanton', 'jose', 'joseph', 'lee axelrad', 'michael', 'respond', 'solano', 'vallejo']	[]	Jseh_Michael_11-30-2007_redaced140001
28	['chase', 'citizen', 'cont', 'fairfield', 'ferrara', 'gary stanton', 'jose', 'joseph', 'lee axelrad', 'michael', 'respond', 'solano', 'vallejo']	[]	Jseh_Michael_12-04-2007_redaced20001



Results

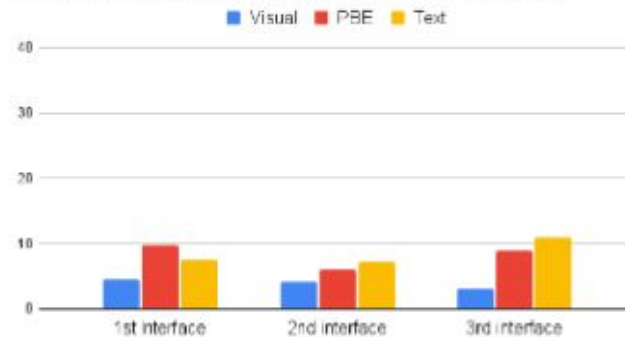
Exact Duplicate Detection

Task 1: Set minimum percentage of match



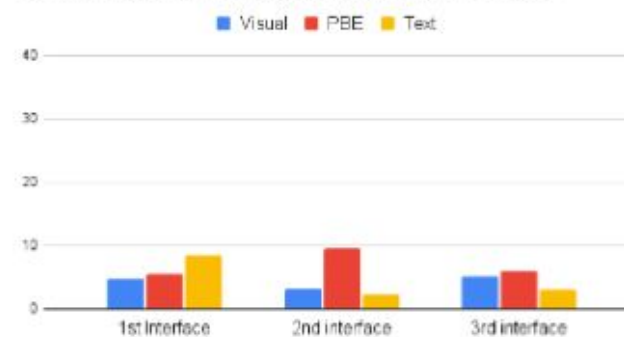
(a) Time taken by participants for Fixed Task of setting percentage threshold.

Task 2: Set page range and minimum matched

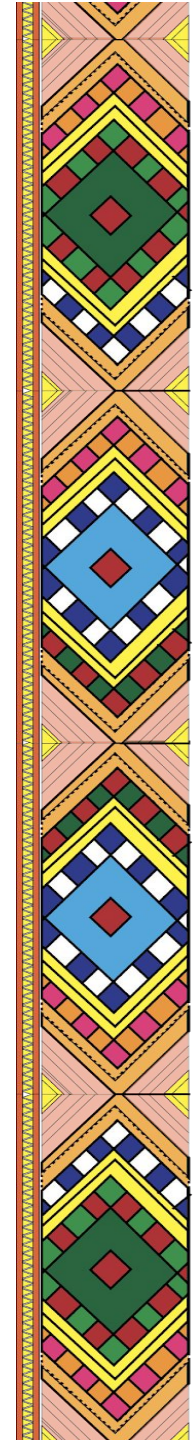


(b) Time taken by participants for Fixed Task of setting number of pages in file and number of matched pages.

Task 3: Explore and set your own parameters



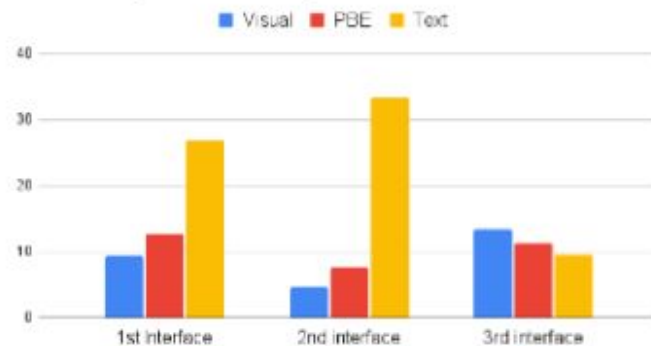
(c) Time taken by participants for Exploration Task.



Results

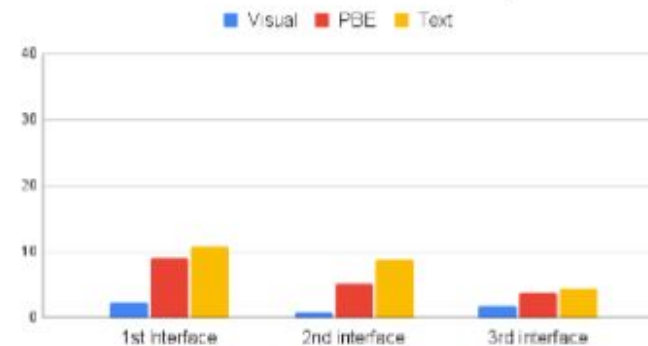
Near Duplicate Detection

Task 1: Explore and set your own threshold



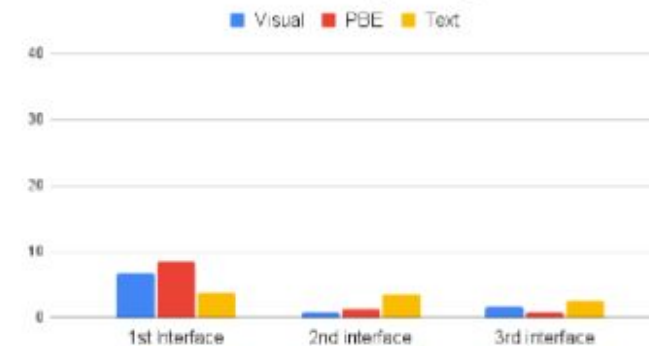
(a) Time taken by participants for Exploration Task.

Task 2: Set threshold for forms and images

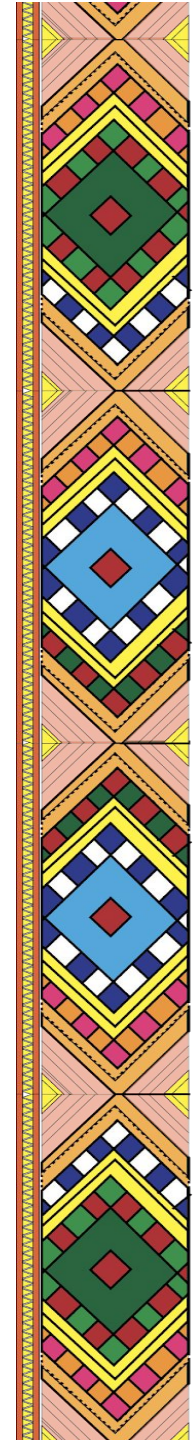


(b) Time taken by participants for Fixed Task of setting correlation threshold for both image and form page types.

Task 3: Set threshold for images only



(c) Time taken by participants for Fixed Task of setting correlation threshold for just image page types.



Introduction

Data comes in different levels of quality, making data extraction extremely difficult.

SOLANO COUNTY SHERIFF'S OFFICE
 530 UNION AVENUE, SUITE 100
 FAIRFIELD, CA 94533
 (707) 421-7000
 CAMAROSSO

UTILITY REPORT

Case No: CR18-4874
 Date: 08/15/18
 Officer: [Redacted]

Vehicle: 2008 Acura Integra EX-L
 License: 1A123456789
 Make: Acura
 Model: Integra
 Year: 2008
 Color: Silver

Location: 1234 Main St, Fairfield, CA 94533

Officer: T. Pierce, 1152/1K5

RIVERSIDE COUNTY SHERIFF'S DEPARTMENT

Case No: [Redacted]

Defendant Information:
 Name: [Redacted]
 DOB: [Redacted]
 Sex: [Redacted]
 Race: [Redacted]

Arrest Information:
 Agency: [Redacted]
 Date: [Redacted]

Booking Charges:
 Charge: [Redacted]

Domestic Violence Notification:
 YES/NO [Redacted]

AUTOMATED LATENT PRINT SECTION
 530 Union Avenue
 Fairfield, CA 94533
 (707) 421-7062

CASE SUBMISSION FORM

PLEASE PROVIDE ELIMINATION PRINTS

REQUESTING AGENCY: Solano County
 DATE OF REQUEST: 8/21/19

AGENCY CASE # 94-001-CR-208 OFFENSE

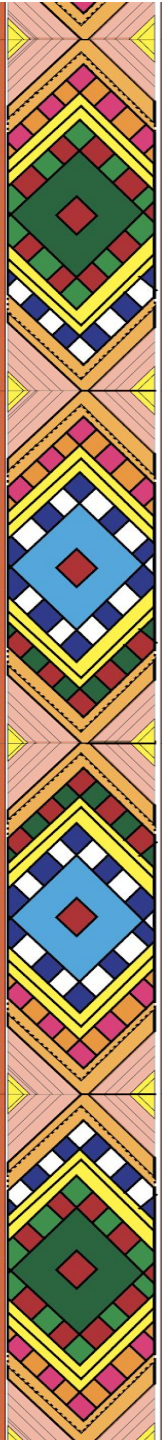
SUBMITTING OFFICER: James Jim Bunker
 PHONE #

EVIDENCE DESCRIPTION AND SPECIAL INSTRUCTIONS
 Compare prints of Solano County Case #19-1497-CR-E-228
 with RT of CR# [Redacted]

CASE RESULTS AND DISPOSITION OF MATERIALS
 The right thumb print of CR# [Redacted] in the name Douglas Clifford Clark was issued on 1/15/92, and the right thumb print of Solano County case # 899-1497-CR-E-228 are the thumb print of the same person.

Sam Murphy, Latent Fingerprint Examiner
 8/26/19

CR# number from PEO Gary Faulkner 8/26/19 1/800
 Compare prints from James Jim Bunker 8/26/19 3/800



Introduction

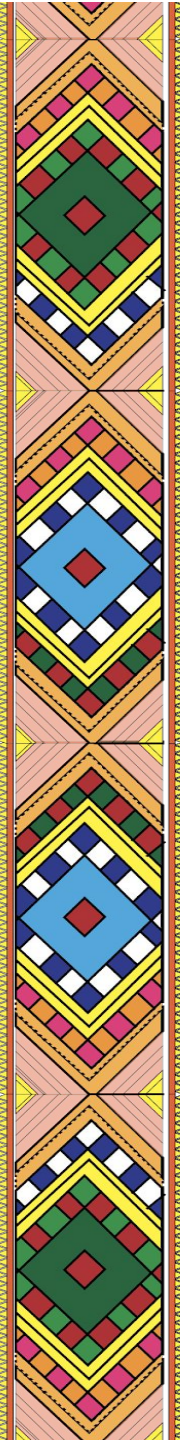
Data comes in different levels of quality, making data extraction extremely difficult.

A yellow form from the Riverside County Sheriff's Department. It contains sections for 'DEFENDANT INFORMATION', 'ARREST INFORMATION', and 'DOMESTIC VIOLENCE NOTIFICATION'. Handwritten in the top right corner are 'SW' and 'Danno'.

A form from the Solano County Sheriff's Office titled 'UTILITY REPORT'. It includes fields for 'TO/FR TO', 'AGENCY', 'CASE NO.', and 'DATE'. The form is mostly filled out with handwritten information.

A small form at the bottom left with handwritten entries. It includes fields for 'NAME', 'DATE', and 'TIME'. The name 'T. Pierce, 1152' is visible.

An 'AUTOMATED LATENT PRINT SECTION' case submission form. It includes a police badge logo and fields for 'REQUESTING AGENCY', 'DATE OF REQUEST', 'AGENCY CASE #', and 'OFFENSE'. Handwritten text describes a case involving 'Compare prints of Solano Co case 899-1497-CCE-228 with RT at CR#'. It also includes a signature: 'Sam Murphy, Latent Fingerprint Examiner 8/26/99'.



Introduction

Data comes in different levels of quality, making data extraction extremely difficult.

AUTOMATED LATENT PRINT SECTION
530 Union Avenue
Fairfield, CA 94533
(707) 421-7062

RICHARD D. HULSE
Sheriff

CASE SUBMISSION FORM

PLEASE PROVIDE ELIMINATION PRINTS
REQUESTING AGENCY Solano County DATE OF REQUEST 8/20/99

AGENCY CASE # 94-497-CC-228 OFFENSE _____

SUBMITTING OFFICER James Sim Bunker PHONE# _____
EVIDENCE DESCRIPTION AND SPECIAL INSTRUCTIONS
Compare prints of Solano County Case 899-1497-CC-228
with PT of CR# 23 _____

CASE RESULTS AND DISPOSITION OF MATERIALS
The right thumb print of CR# _____ in the name
Douglas Clifford Clark was _____ issued on 1/18/92,
and the right thumb print of Solano County
Case# 899-1497-CC-228, date the thumb print of
the same person.

Jean Murphy, Latent Fingerprint Examiner
8/20/99

CDL Sender from P.O. Gary Faulkner 8/20/99 1800
Compare prints from James Sim Bunker 8/20/99 1810

SOLANO COUNTY SHERIFF'S OFFICE
530 UNION AVENUE, SUITE 100
FAIRFIELD, CA 94533
(707) 421-7062
CAD48000

UTILITY REPORT

Deputy T. Pience, 11.52/1K5
See Attached

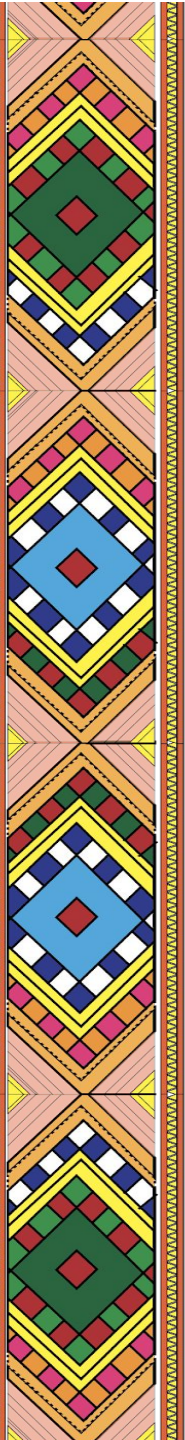
RIVERSIDE COUNTY SHERIFF'S DEPARTMENT
DEFENDANT INFORMATION

ARREST INFORMATION

BOOKING CHARGES

DOMESTIC VIOLENCE NOTIFICATION - PER 846.30 PC

Deputy T. Pience, 11.52/1K5



Introduction

Files belonging to the same case might be **spread across several folders**, creating a challenge for working with a particular case.

