Cleaning And Merging Crowdsourced data on Police Brutality

Jainul Vaghasia

13 May, 2018

Motivation

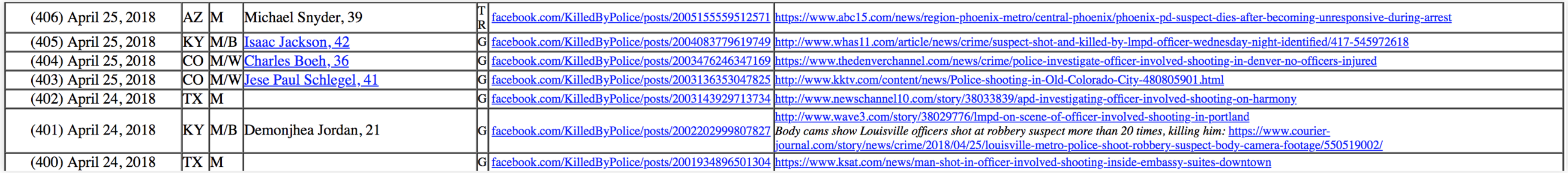
- Would like to analyze data, Maddi's paper
- But wait…

The problem

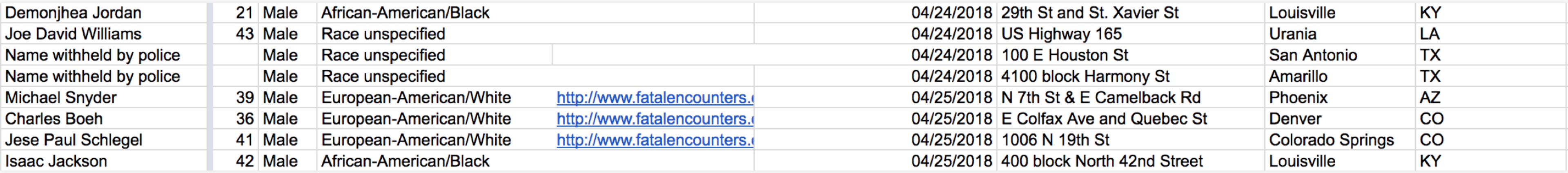
Some beginning problems:

- Structure of the tables
  - Different format of data storage in columns
  - Rows not flat; they contain multiple records in same row
  - Columns contain multiple attributes in same column
- Differing information
  - Conflicting information recorded
  - Data present in one dataset, while absent in other

SnapShot of datasets (April 24, 2018 - April 25, 2018)



(406) April 25, 2018	AZ	M	Michael Snyder, 39	T R	facebook.com/KilledByPolice/posts/2005155559512571	https://www.abc15.com/news/region-phoenix-metro/central-phoenix/phoenix-pd-suspect-dies-after-becoming-unresponsive-during-arrest
(405) April 25, 2018	KY	M/B	Isaac Jackson, 42	G	facebook.com/KilledByPolice/posts/2004083779619749	http://www.whas11.com/article/news/crime/suspect-shot-and-killed-by-lmpd-officer-wednesday-night-identified/417-545972618
(404) April 25, 2018	CO	M/W	Charles Boeh, 36	G	facebook.com/KilledByPolice/posts/2003476246347169	https://www.thedenverchannel.com/news/crime/police-investigate-officer-involved-shooting-in-denver-no-officers-injured
(403) April 25, 2018	CO	M/W	Jese Paul Schlegel, 41	G	facebook.com/KilledByPolice/posts/2003136353047825	http://www.kktv.com/content/news/Police-shooting-in-Old-Colorado-City-480805901.html
(402) April 24, 2018	TX	M		G	facebook.com/KilledByPolice/posts/2003143929713734	http://www.newschannel10.com/story/38033839/apd-investigating-officer-involved-shooting-on-harmony
(401) April 24, 2018	KY	M/B	Demonjhea Jordan, 21	G	facebook.com/KilledByPolice/posts/2002202999807827	http://www.wave3.com/story/38029776/lmpd-on-scene-of-officer-involved-shooting-in-portland *Body cams show Louisville officers shot at robbery suspect more than 20 times, killing him:* https://www.courier-journal.com/story/news/crime/2018/04/25/louisville-metro-police-shoot-robbery-suspect-body-camera-footage/550519002/
(400) April 24, 2018	TX	M		G	facebook.com/KilledByPolice/posts/2001934896501304	https://www.ksat.com/news/man-shot-in-officer-involved-shooting-inside-embassy-suites-downtown



Demonjhea Jordan	21	Male	African-American/Black		04/24/2018	29th St and St. Xavier St	Louisville	KY
Joe David Williams	43	Male	Race unspecified		04/24/2018	US Highway 165	Urania	LA
Name withheld by police		Male	Race unspecified		04/24/2018	100 E Houston St	San Antonio	TX
Name withheld by police		Male	Race unspecified		04/24/2018	4100 block Harmony St	Amarillo	TX
Michael Snyder	39	Male	European-American/White	http://www.fatalencounters.	04/25/2018	N 7th St & E Camelback Rd	Phoenix	AZ
Charles Boeh	36	Male	European-American/White	http://www.fatalencounters.	04/25/2018	E Colfax Ave and Quebec St	Denver	CO
Jese Paul Schlegel	41	Male	European-American/White	http://www.fatalencounters.	04/25/2018	1006 N 19th St	Colorado Springs	CO
Isaac Jackson	42	Male	African-American/Black		04/25/2018	400 block North 42nd Street	Louisville	KY

The problem quantified

- On two datasets FE and KBP, we consider the missing data counts for perfectly matching (intersecting fields match exactly except the possibly missing field in consideration) records
  - Note Race is the most missing data

Race

	Present in KBP	Absent in KBP
Present in FE	2206	243
Absent in FE	21	135

Age

	Present in KBP	Absent in KBP
Present in FE	2206	10
Absent in FE	4	2

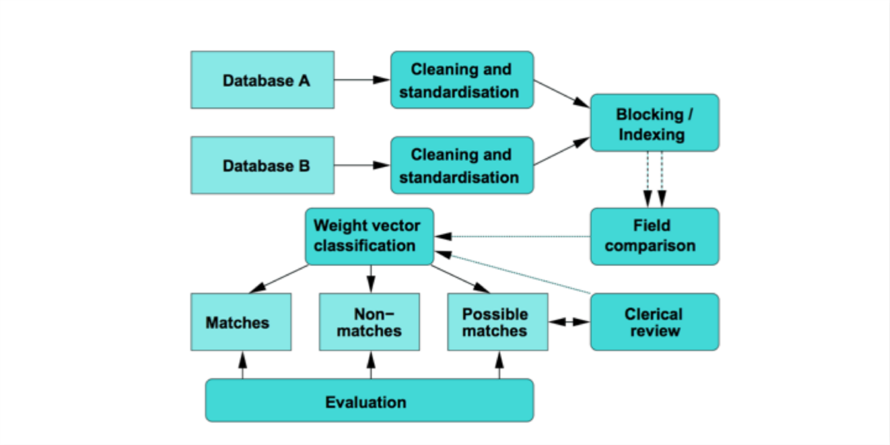
Gender

	Present in KBP	Absent in KBP
Present in FE	2206	0
Absent in FE	1	0

- Matching and merging data helps in extracting information that is present in one dataset but absent in another.
- Using learning algorithm further allows us to match records that are same but do not match perfectly

Our Approach

1. Clean the datasets and standardize them based on the understanding of structure
   - Reformat fields so that they hold the same format in the intersecting fields
   - Flatten the rows to only contain one record per row
   - Partition columns with multiple attributes into multiple columns
   - Standard technique in the field
2. Use the information in the well maintained records according to multiple datasets to extract information from the differing records
   - Train a learning algorithm to classify the matching and non-matching records between datasets
   - Based on the algorithm's classifications, merge the records to create a complete dataset



Steps

- Choose a similarity metric to define how similar two records are to each other
  - We use String edit distance based on Jaro-Wrinkler in order to account for clerical errors

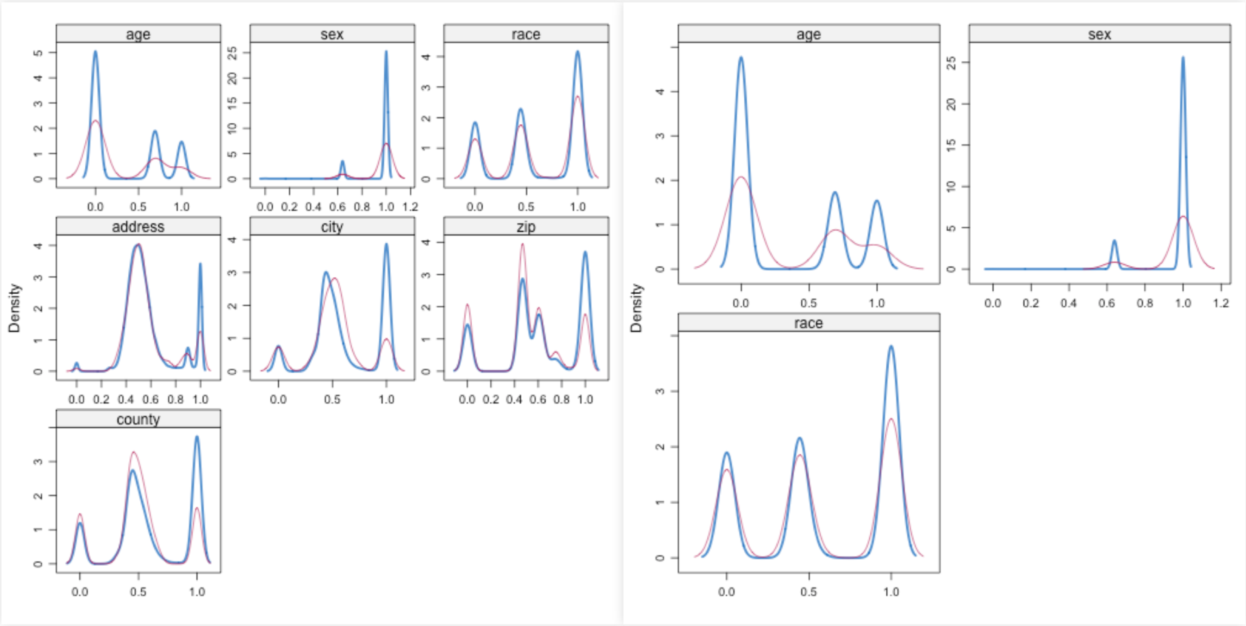
	id1	id2	name	age	sex	race	dateDMY	address	city	state	zip	county
94610	1893	2773	0.5282187	0	1	1	0.86	0.4821869	1	1	1	1

name	age	sex	race	dateDMY	address	city	state	zip	county
Winfield Carlton Fisher III	32	Male	Black	2014-03-18	2765 N Salisbury Blvd	Salisbury	MD	21801	Wicomico

name	age	sex	race	dateDMY	address	city	state	zip	county
Fednel Rhinvil	25	Male	Black	2015-03-03	East Road and Olivia Street	Salisbury	MD	21801	Wicomico

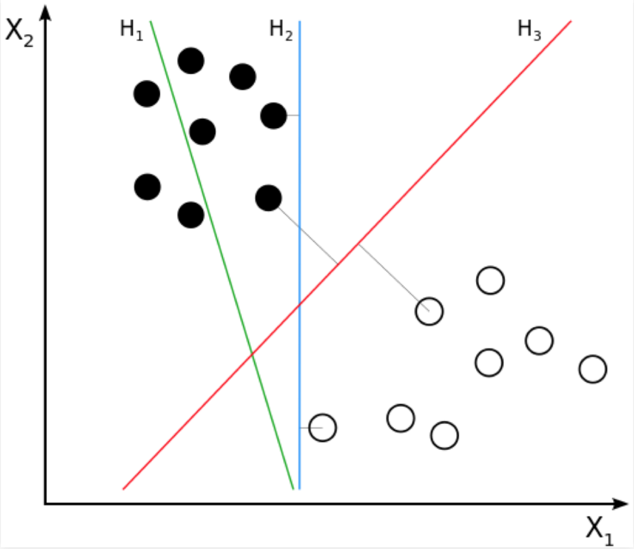
Steps

- To calculate the similarity measure for record pairs that are missing data, we use regression to imputate placeholder value
  - Note that the imputed data (pink) follows a roughly similar distribution of densities as observed data (blue)



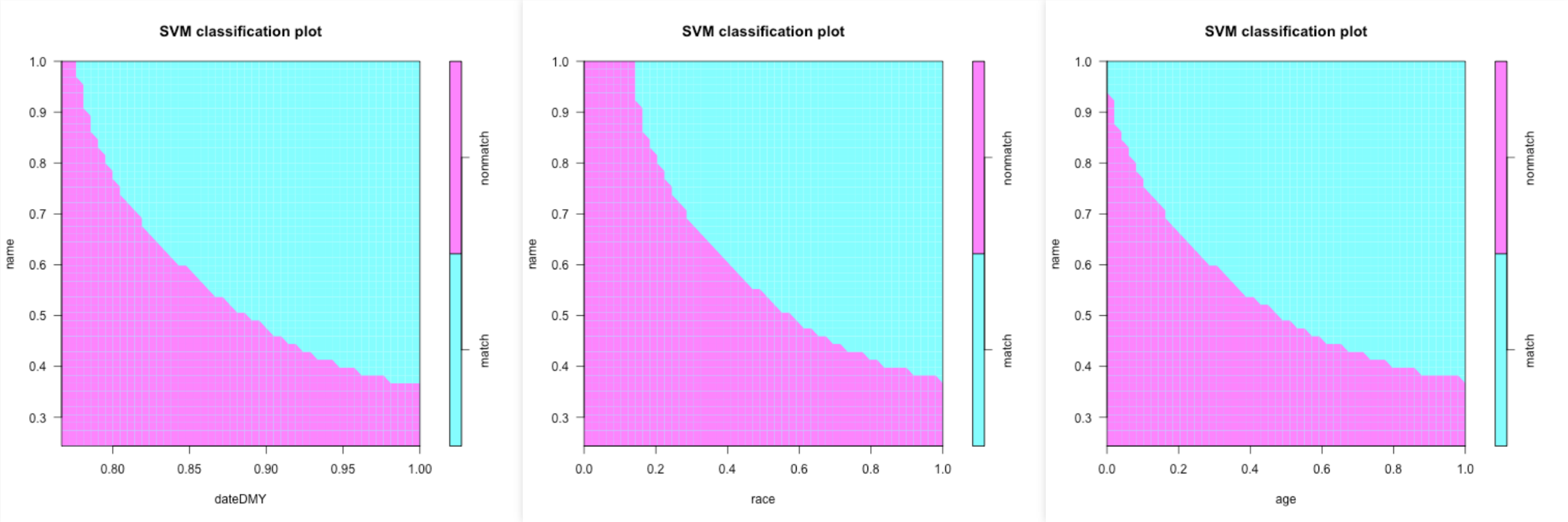
Steps

- Create a seed training data that consists of the perfect matches and non-matches
- Use this training data to train a classification algorithm
  - We use the Support Vector Machine algorithm



Decision Boundaries

- Boundaries sliced at points that match except at the attributes plotted



Bounds on Accuracy

Future Work

- Use text analytics to extract information from the news articles linked in the datasets
- Explore methods to handle data that is missing in all datasets

References