MATCHING HETEROGENOUS DATASETS USING SEEDED CLASSIFICATION

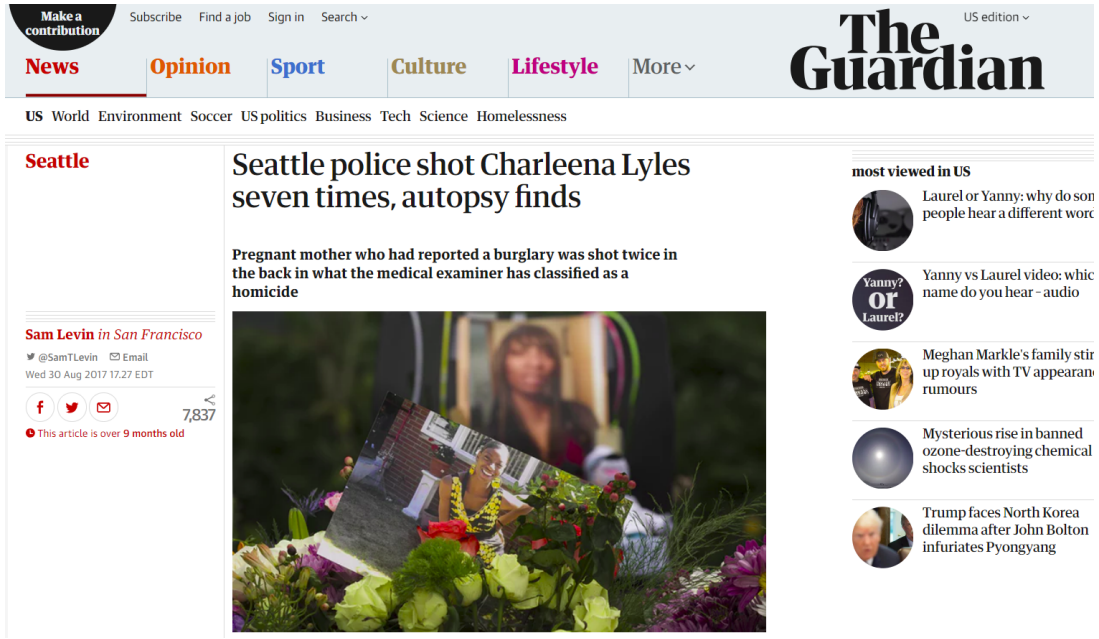
Using machine learning to merge crowdsourced datasets on fatal encounters with police

Jainul Vaghasia

Mentors: Martina Morris, Ben Marwick

Linked Project: "Fatal Encounters with Police: Improving Public Access to Exploratory Data Analytics" by Madeline Cummins

Motivation



- Fatal encounters with Police in the US have drawn widespread national and international interest
- What are the trends? The states with the highest rates? Are there disparities by race?
- It turns out the data are not easy to find, or analyze

Data on fatal shootings

- Not collected by the Federal Government
  - Or the States
  - Some localities do (but this is not always publicly available)

- Two active crowd-sourced datasets available online:

	Years covered	Number of records
Fatal Encounters	2000-2018	~ 21,000
Killed by Police	2013-2018	~ 6,000

All other online data sources (e.g., Washington Post) are derivative of these

These data sets have different information

- They have X pieces of information in common
  - Name, address, age, sex, race
- But each also has unique information
- So, we want to merge them to create a single comprehensive file
  - Merge using the fields they have common

Merge Challenges (1)

- The data are stored and accessed differently

Snapshot of data (April 24, 2018–April 25, 2018)

KBP data

(406) April 25, 2018	AZ	M	Michael Snyder, 39	TR	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

FE data

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

- For example: In KBP, Name and Age are in the same field, sex and race are in the same field

Merge Challenges (2)

- The information in the shared fields is sometimes missing
  - Especially for Race:
    $\sim 32\%$ missing in FE
    $\sim 14\%$ missing in KBP

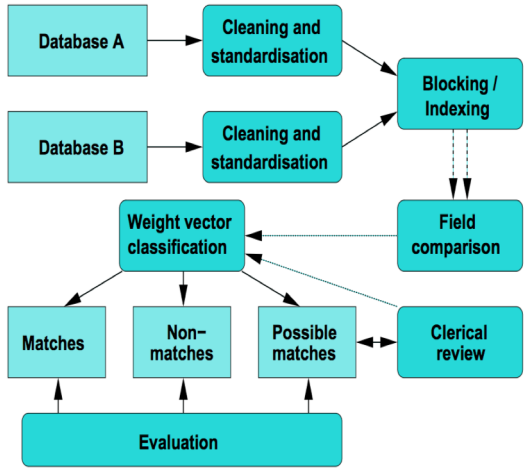
- The information in the shared fields differs
  - Typos
  - One has middle name, the other doesn’t

Our approach has three steps

1. Pre-processing
   - 1.1 Scrape and clean the data
   - 1.2 Deal with missing data (transfer or impute)

2. Merge using Machine Learning algorithm
   - 2.1 Training the classifier
   - 2.2 Applying the trained classifier to the data

3. Assess the performance of the merge algorithm
   - 3.1 Sensitivity and specificity



How the matching algorithm works

- Calculates a “distance” between every pair of records

- Distance is a function of:
  - Difference in length (1)
  - Number of matching characters (4)
  - Transposition (12 vs 21)

0A1234
1B21345

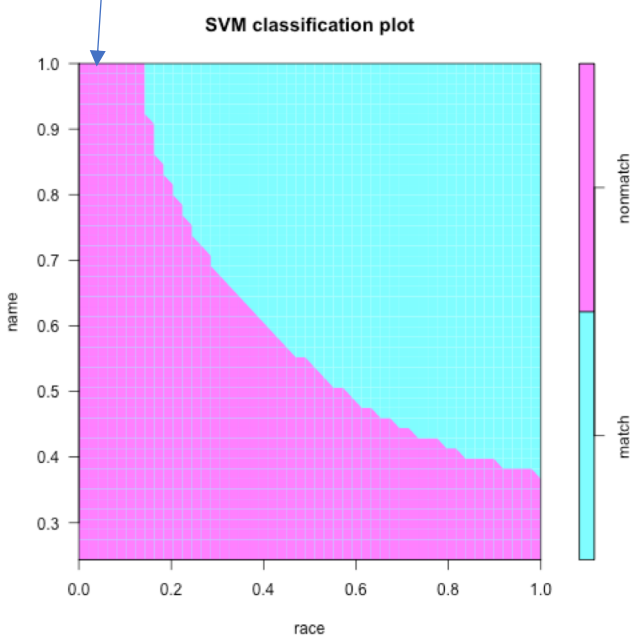
How the matching algorithm works

- Jaro Similarity: $sim = \begin{cases} 0, & if\ m = 0 \\ \frac{1}{3}\left(\frac{m}{|s_1|} + \frac{m}{|s_2|} + \frac{m-t}{m}\right), & otherwise \end{cases}$
  - m = # matching characters
  - $|s_i|$ = length of string s_i
  - t = ½ the number of transpositions
- Jaro-Winkler uses sim and accounts for matching initial characters

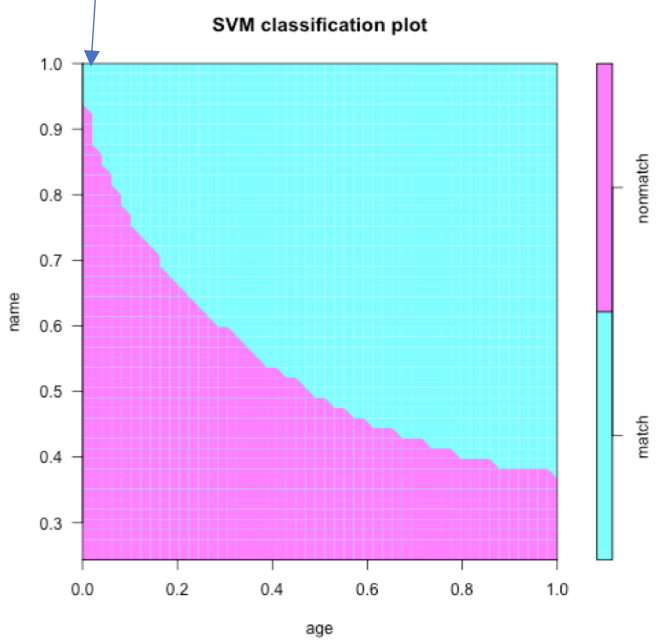
Training and using the algorithm

- Train by using on a subset of known records
  - matches + non-matches
- This generates a set of "**decision boundaries**" for deciding matches
  - Matches are more sensitive to some fields than others

Same name, different race = no match



Small differences in name = match, even if age is very different



Assess the accuracy of the classifier

- Based on clerical (manual) review of 25 matches, and 25 non-matches

	Algorithm Match	Algorithm Non-Match
True Match	25	1
True Non-Match	0	24

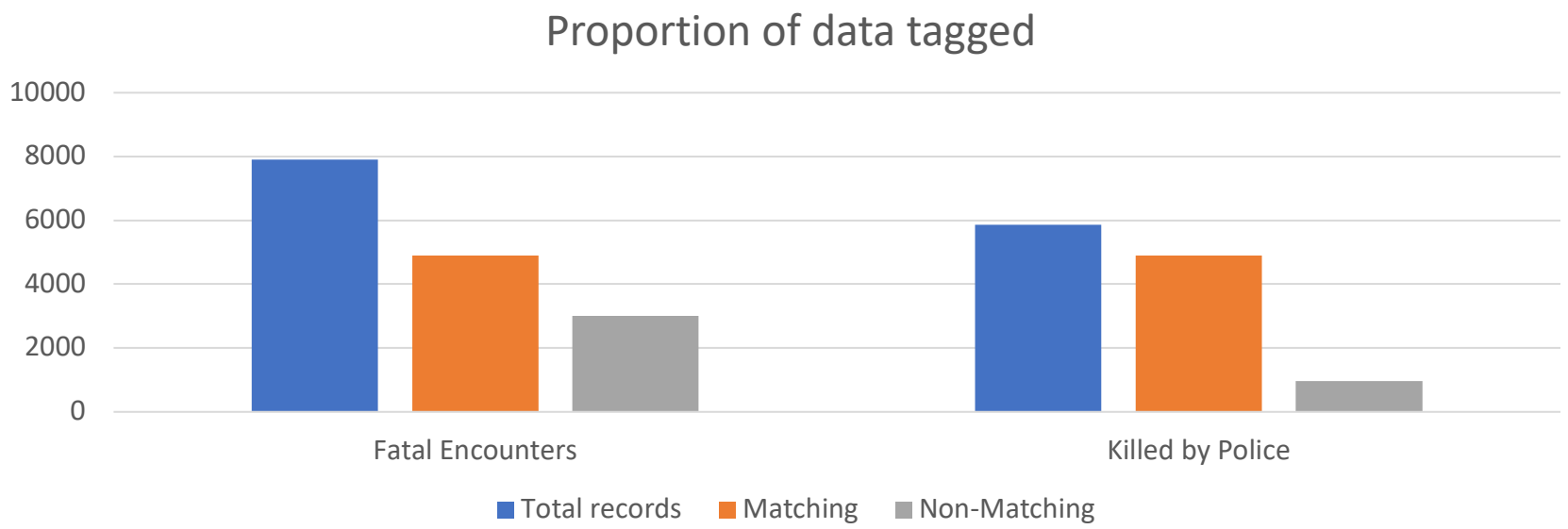
- Sensitivity: $25/26 = 96\%$ of true matches are correctly identified
- Specificity: $24/24 = 100\%$ of true non-matches are correctly identified

- Positive Predictive Validity: 100% of the algorithm matches are correct
- Negative Predictive Validity: 96% of the algorithm non-matches are correct

Final Results

	Fatal Encounters	Killed by Police
Records in [2013, 2018]	7905	5863

Matching records found	4902



Future work

- Use text analytics to extract information from articles linked in the datasets
- Explore methods to handle data that is missing in both datasets
- Evaluate the performance of imputation methods