Information Extraction

Chapter 17 of the Jurafsky and Martin textbook

Information Extraction (IE)

- Information extraction (IE), turns the unstructured text information into structured data
- Populate a relational database to enable further processing, support queries

Template Filling

Citing high fuel prices, United Airlines said Friday it has increased fares by $6 per round trip on flights to some cities also served by lower- cost carriers. American Airlines, a unit of AMR Corp., immediately matched the move, spokesman Tim Wagner said. United, a unit of UAL Corp., said the increase took effect Thursday and applies to most routes where it competes against discount carriers, such as Chicago to Dallas and Denver to San Francisco.

FARE-RAISE ATTEMPT:	LEAD AIRLINE:	UNITED AIRLINES
	AMOUNT:	$6
	EFFECTIVE DATE:	2006-10-26
	FOLLOWER:	AMERICAN AIRLINES

Steps in IE

- NER and co-reference resolution
- **relation** extraction
  - *spouse-of, child-of, employer-of, part-of, membership-in, located-in*
- event extraction
- temporal expression normalization
- template filling

Named Entity Recognition

Citing high fuel prices, [ORG United Airlines] said [TIME Friday] it has increased fares by [MONEY $6] per round trip on flights to some cities also served by lower-cost carriers. [ORG American Airlines], a unit of [ORG AMR Corp.], immediately matched the move, spokesman [PER Tim Wagner] said. [ORG United], a unit of [ORG UAL Corp.], said the increase took effect [TIME Thursday] and applies to most routes where it competes against discount carriers, such as [LOC Chicago] to [LOC Dallas] and [LOC Denver] to [LOC San Francisco].

Type	Tag	Sample Categories	Example sentences
People	PER	people, characters	**Turing** is a giant of computer science.
Organization	ORG	companies, sports teams	The **IPCC** warned about the cyclone.
Location	LOC	regions, mountains, seas	The **Mt. Sanitas** loop is in **Sunshine Canyon**.
Geo-Political Entity	GPE	countries, states, provinces	**Palo Alto** is raising the fees for parking.
Facility	FAC	bridges, buildings, airports	Consider the **Tappan Zee Bridge**.
Vehicles	VEH	planes, trains, automobiles	It was a classic **Ford Falcon**.

Category Ambiguity in NER

Name	Possible Categories
Washington	Person, Location, Political Entity, Organization, Vehicle
Downing St.	Location, Organization
IRA	Person, Organization, Monetary Instrument
Louis Vuitton	Person, Organization, Commercial Product

Examples of type ambiguities in the use of the name *Washington:*

[$_{\text{PER}}$ Washington] was born into slavery on the farm of James Burroughs.
[$_{\text{ORG}}$ Washington] went up 2 games to 1 in the four-game series.
Blair arrived in [$_{\text{LOC}}$ Washington] for what may well be his last state visit.
In June, [$_{\text{GPE}}$ Washington] passed a primary seatbelt law.
The [$_{\text{VEH}}$ Washington] had proved to be a leaky ship, every passage I made...

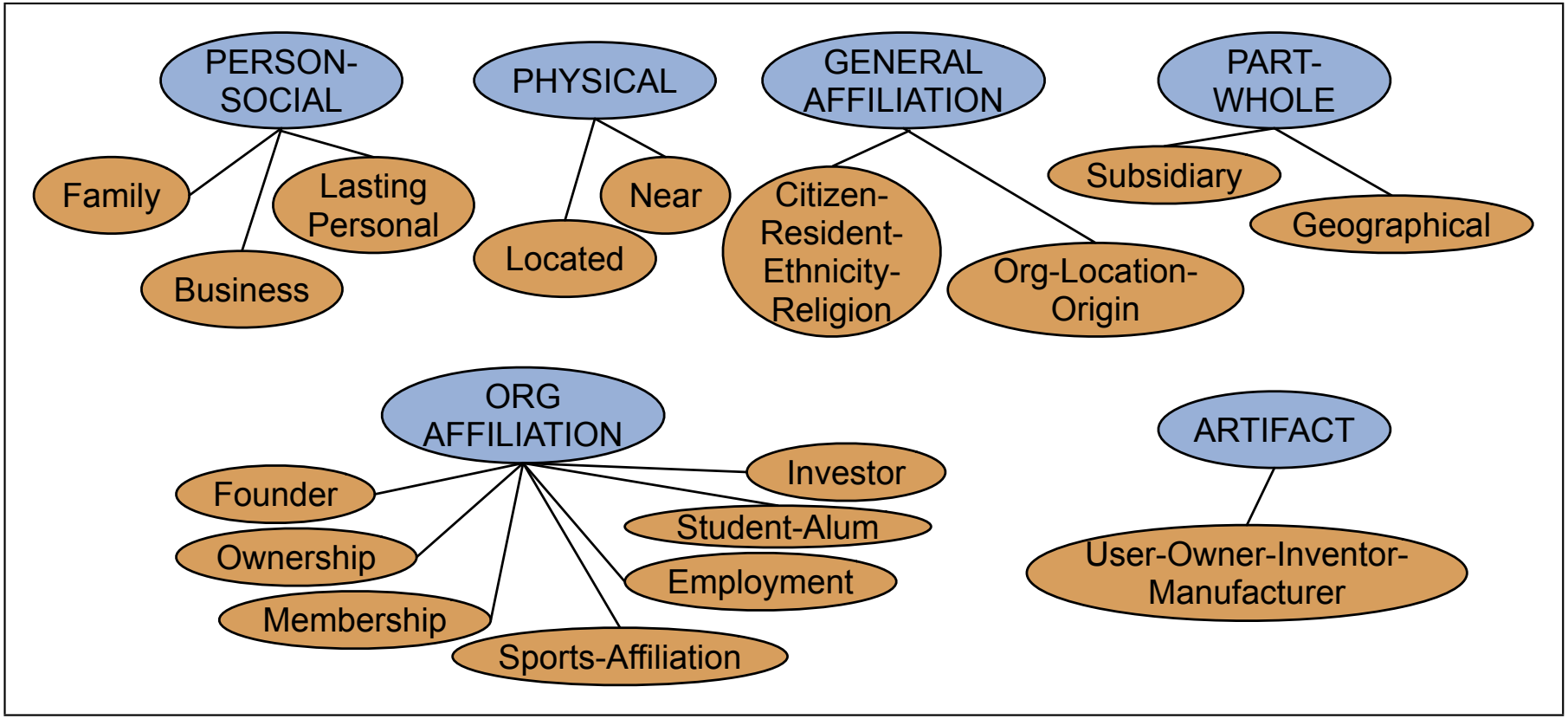
Relation Extraction

Citing high fuel prices, [ORG United Airlines] said [TIME Friday] it has increased fares by [MONEY $6] per round trip on flights to some cities also served by lower-cost carriers. [ORG American Airlines], a unit of [ORG AMR Corp.], immediately matched the move, spokesman [PER Tim Wagner] said. [ORG United], a unit of [ORG UAL Corp.], said the increase took effect [TIME Thursday] and applies to most routes where it competes against discount carriers, such as [LOC Chicago] to [LOC Dallas] and [LOC Denver] to [LOC San Francisco].

Tim Wagner **is a spokesman for** American Airlines

United **is a unit of** UAL Corp

American **is a unit of** AMR



Relations	Types	Examples
Physical-Located	PER-GPE	**He** was in **Tennessee**
Part-Whole-Subsidiary	ORG-ORG	**XYZ**, the parent company of **ABC**
Person-Social-Family	PER-PER	**Yoko**'s husband **John**
Org-AFF-Founder	PER-ORG	**Steve Jobs**, co-founder of **Apple**...

Unified Medical Language System

Entity	Relation	Entity
Injury	disrupts	Physiological Function
Bodily Location	location-of	Biologic Function
Anatomical Structure	part-of	Organism
Pharmacologic Substance	causes	Pathological Function
Pharmacologic Substance	treats	Pathologic Function

Given a medical sentence like:

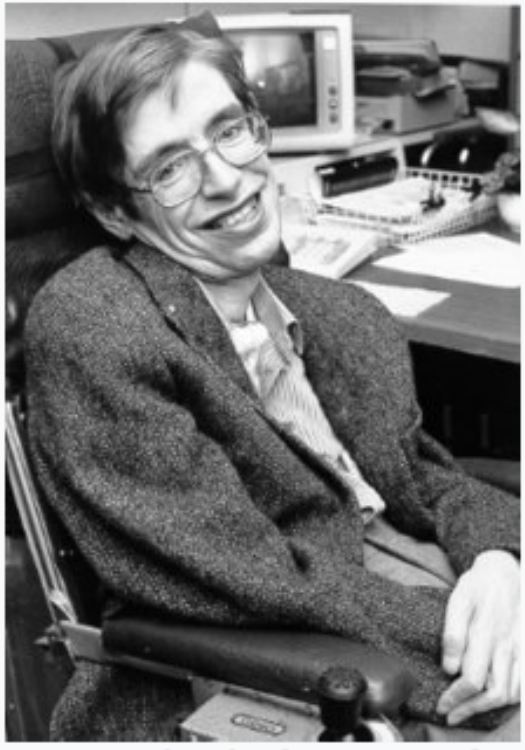
Doppler echocardiography can be used to diagnose left anterior descending artery stenosis in patients with type 2 diabetes

Extract the UMLS relation:

Echocardiography (Doppler) **Diagnoses** *Acquired stenosis*

Wikipedia info boxes

Stephen Hawking
CH CBE FRS FRSA



Hawking at NASA's StarChild Learning Center, ca. 1999

Born	Stephen William Hawking 8 January 1942 Oxford, England
Died	14 March 2018 (aged 76) Cambridge, England
Education	St Albans School, Hertfordshire
Alma mater	University of Oxford (BA) University of Cambridge (MA, PhD)
Known for	Hawking radiation Penrose–Hawking theorems Bekenstein–Hawking formula Hawking energy Gibbons–Hawking ansatz Gibbons–Hawking effect Gibbons–Hawking space Gibbons–Hawking–York boundary term Thorne–Hawking–Preskill bet
Spouse(s)	Jane Wilde (m. 1965; div. 1995) Elaine Mason (m. 1995; div. 2006)
Children	3, including Lucy
Awards	Adams Prize (1966) Eddington Medal (1975) Maxwell Medal and Prize

RDF Triples

- Resource Description Framework
- **RDF triple** is a tuple of
  - entity-relation-entity, aka
  - subject-predicate-object expression
- **subject** University of Pennsylvania
  predicate location
  object Philadelphia, PA

Zachary G. Ives

FOLLOW

Professor of Computer and Information Science, University of Pennsylvania
Verified email at cis.upenn.edu - Homepage

Databases data integration distributed systems web data management

TITLE	CITED BY	YEAR
Dbpedia: A nucleus for a web of open data S Auer, C Bizer, G Kobilarov, J Lehmann, R Cyganiak, Z Ives The semantic web, 722-735	3510	2007

DBpedia is a community effort to extract structured information from Wikipedia and to make this information available on the Web. DBpedia allows you to ask sophisticated queries against datasets derived from Wikipedia and to link other datasets on the Web to Wikipedia data. We describe the extraction of the DBpedia datasets, and how the resulting information is published on the Web for human-and machine-consumption. We describe some emerging applications from the DBpedia community and show how website authors can facilitate DBpedia content within their sites. Finally, we present the current status of interlinking DBpedia with other open datasets on the Web and outline how DBpedia could serve as a nucleus for an emerging Web of open data.

Freebase, WordNet, other ontologies

- **Freebase** relations:
  people/person/nationality
  location/location/contains
  people/person/place-of-birth
  biology/organism classification
- **WordNet** relations:
  is-a, instance-of
  hypernyms/hyponyms
  Giraffe is-a ruminant is-a ungulate is-a mammal is-a vertebrate is-a animal…

Strategies for relation extraction

- hand-written patterns
- supervised machine learning
- semi-supervised machine learning
- unsupervised machine learning

Hearst Patterns

Agar is a substance prepared from a mixture of red algae, such as Gelidium, for laboratory or industrial use.

She suggests that the following **lexico-syntactic pattern**

$$NP_0 \textit{ such as } NP_1\{,NP_2\ldots,(and|or)NP_i\}, i \geq 1$$

implies the following semantics

$$\forall NP_i, i \geq 1, \text{hyponym}(NP_i, NP_0)$$

allowing us to infer

$$\text{hyponym}(\text{Gelidium}, \text{red algae})$$

NP {, NP}* {,} (and|or) other NP_H	temples, treasuries, and other important civic buildings
NP_H such as {NP,}* {(or|and)} NP	red algae such as Gelidium
such NP_H as {NP,}* {(or|and)} NP	such authors as Herrick, Goldsmith, and Shakespeare
NP_H {,} including {NP,}* {(or|and)} NP	common-law countries, including Canada and England
NP_H {,} especially {NP}* {(or|and)} NP	European countries, especially France, England, and Spain

Machine learning techniques

- Supervised: training corpus annotated with manually annotated with fixed set of relations and entities
- Semi-supervised: high-precision seed patters, or seed tuples, are used to bootstrap more examples
- Distant supervision: start with a huge number of seeds, learn noisy pattern fields (e.g. 100k examples of birth-place-of from infoboxes, help learn the corresponding text patterns)

Open IE

- Unsupervised relation extraction
- Find all *strings of words* that satisfy the tripe relation.
  United has a hub in Chicago, which is the headquarters of United Continental Holdings.
  r1: <United, has a hub in, Chicago>
  r2: <Chicago, is the headquarters of, United Continental Holdings>

Temporal Expression Extraction

Absolute	Relative	Durations
April 24, 1916	yesterday	four hours
The summer of ’77	next semester	three weeks
10:15 AM	two weeks from yesterday	six days
The 3rd quarter of 2006	last quarter	the last three quarters

Lexical triggers for temporal expressions:

Category	Examples
Noun	*morning*, *noon*, *night*, *winter*, *dusk*, *dawn*
Proper Noun	*January, Monday, Ides, Easter, Rosh Hashana, Ramadan, Tet*
Adjective	*recent, past, annual, former*
Adverb	*hourly, daily, monthly, yearly*

Event Extraction

[EVENT Citing] high fuel prices, United Airlines [EVENT said] Friday it has [EVENT increased] fares by $6 per round trip on flights to some cities also served by lower-cost carriers. American Airlines, a unit of AMR Corp., immediately [EVENT matched] [EVENT the move], spokesman Tim Wagner [EVENT said]. United, a unit of UAL Corp., [EVENT said] [EVENT the increase] took effect Thursday and [EVENT applies] to most routes where it [EVENT competes] against discount carriers, such as Chicago to Dallas and Denver to San Francisco.

In English events are typically expressed verbs like *exploded,* and sometimes nouns like *explosion.*

Temporal ordering of events

Delta Air Lines earnings soared 33% to a record in the fiscal first quarter, bucking the industry trend toward declining profits.

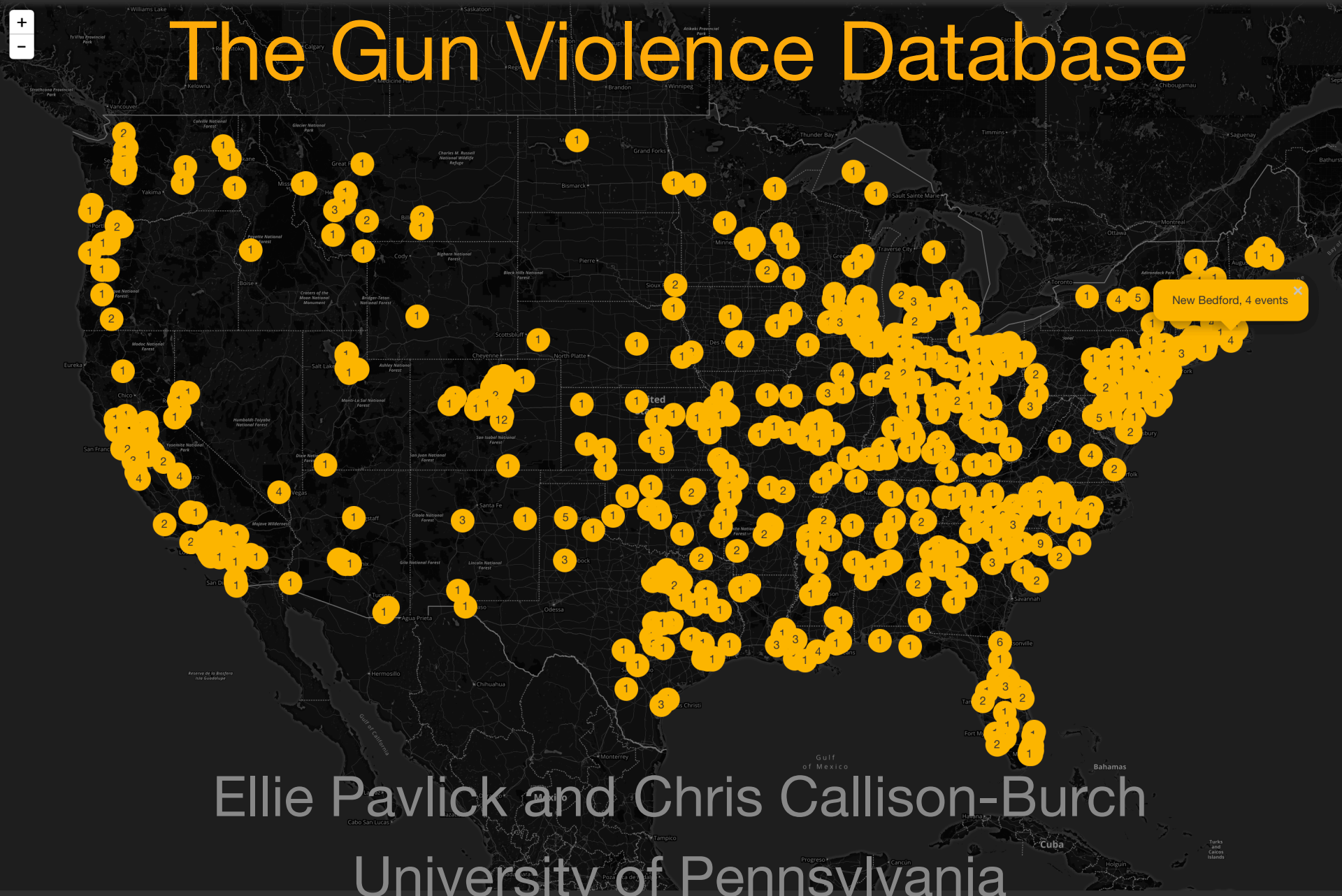
- Soaring$_{e1}$ is **included in** the fiscal first quarter$_{t58}$
- Soaring$_{e1}$ is **simultaneous with** the bucking$_{e3}$
- Declining$_{e4}$ **includes** soaring$_{e1}$

Template Filling

- Scripts are common, stereotypical situations
- Templates represent scripts with a fixed set of slots that take slot-filler values

FARE-RAISE ATTEMPT:	LEAD AIRLINE:	UNITED AIRLINES
	AMOUNT:	$6
	EFFECTIVE DATE:	2006-10-26
	FOLLOWER:	AMERICAN AIRLINES

The Gun Violence Database



Ellie Pavlick and Chris Callison-Burch
University of Pennsylvania

Goals of the GVDB

Collect data about gun violence in the US to facilitate **public health research**

Draw sample from local newspapers and television stations that publish online

Use machine learning and crowdsourcing to **extract structured data from text**

An engraving of the Mechanical Turk, the 18th century chess-playing automaton

Crowdsourcing and human computation are emerging fields that sit squarely at the intersection of economics and computer science. They examine how people can be used to solve complex tasks that are currently beyond the capabilities of artificial intelligence algorithms. Online marketplaces like Mechanical Turk and CrowdFlower provide an infrastructure that allows micropayments to be given to people in return for completing human intelligence tasks. This opens up previously unthinkable possibilities like people being used as function calls in software. We will investigate how crowdsourcing can be used for computer science applications like machine learning, next-generation interfaces, and data mining. Beyond these computer science aspects, we will also delve into topics like the sharing economy, prediction markets, how businesses can capitalize on collective intelligence, and the fundamental principles that underlie democracy and other group decision-making processes.

Course number
NETS 213 - students from all majors are welcome!
Instructors
Chris Callison-Burch and Ellie Pavlick
Teaching Assistants
Course Staff
Discussion Forum
Piazza
Time and place
Spring 2016, MWF 2-3PM, LRSM Auditorium
Office Hours
See calendar page
Prerequisites
CIS 120 or prior programming experience
Course Readings
Each lecture has an accompanying set of academic papers
Grading
This is a project-based course. Instead of exams, you will do a series of hands-on assignments and a final project.

http://crowdsourcing-class.org

- Weekly assignments (45%)
- Final project (45%)
- Peer grading (5%)
- Participation (5%)

Chicago Tribune

CRIME IN CHICAGO

Chicago shooting victims

The map below shows where people were shot in Chicago, broken down by community area. The darker the shade of blue, the larger the number of victims. Read our special report on shootings. **This data was last updated July 2.**

1,043 shooting victims in 2013

VICTIMS BY MONTH • 2013 • 2012

---

theguardian

The Counted

People killed by police in the US

SEND A TIP | DATABASE | ABOUT | READ ARTICLES | JOIN US:

PEOPLE KILLED IN 2016: **793**

RACE & ETHNICITY (PER MILLION / TOTAL)

- 5.49 Native American
- 4.86 Black
- 2.3 Hispanic/Latino
- 1.96 White
- 0.72 Asian/Pacific Islander

STATE (NAME / PER CAPITA / TOTAL)

NEW MEXICO

Total killed:	21
Population:	2,085,109
Rank (per capita):	1 / 51
Rank (total):	7 / 51

2016 | 2015 | List | Map

Search by name: Filter by:

Fatal Encounters

A step toward creating an impartial, comprehensive and searchable national database of people killed during interactions with law enforcement

Search

Deadspin Police-Shooting Database Update: We're Still Going

CRIME | MURDER, THEFT, AND OTHER WICKEDNESS. | SEPT. 16 2013 3:34 PM

Slate

How Many People Have Been Killed by Guns Since Newtown?

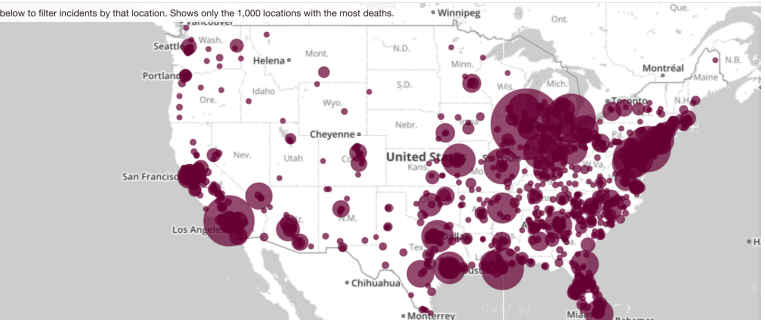
Slate partners with @GunDeaths for an interactive, crowdsourced tally of the toll firearms have taken since Dec. 14.

122k 20

By Chris Kirk and Dan Kois

The answer to the simple question in that headline is surprisingly hard to come by. So *Slate* is collecting data for our crowdsourced interactive. This data is necessarily incomplete (**click here to see why, and to learn more about @GunDeaths, the Twitter user who helped us create this interactive**). But the more people who are paying attention, the better the data will be. You can help us draw a more complete picture of gun violence in America. If you know about a gun death in your community that isn't represented here, please email a link to a news report to **slatedata@gmail.com**. And if you'd like to use this data yourself for your own projects, it's open. You can download it **here**.

Update, Dec. 31, 2013: After **a year of gun deaths**, *Slate* is retiring this project. The count is being picked up by Michael Klein's **Gun Violence Archive project**, launching soon. Thank you to all who volunteered to make the data as comprehensive and accurate as possible.



GUN VIOLENCE Archive

LOGIN | CONTACT US | Search Database

HOME | CHARTS & MAPS | LAST 72 HOURS | REPORTS | ANALYSIS/OPINION | ABOUT US | PAST YEARS | CONGRESSIONAL REPORTS

GUN VIOLENCE Archive 2016

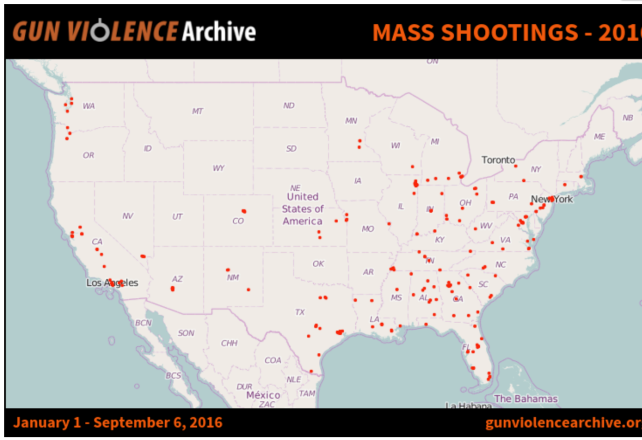
Total Number of Incidents	41,309
Number of Deaths[1]	10,599
Number of Injuries[1]	22,023
Number of Children (age 0-11) Killed or Injured[1]	477
Number of Teens (age 12-17) Killed or Injured[1]	2,252
Mass Shooting[2]	285
Officer Involved Incident Officer Shot or Killed[2]	233
Officer Involved Incident Subject-Suspect Shot or Killed[2]	1,359
Home Invasion[2]	1,724
Defensive Use[2]	1,288

MISSION

Gun Violence Archive (GVA) is a not for profit corporation formed in 2013 to provide free online public access to accurate information about gun-related violence in the United States. GVA will collect and check for accuracy, comprehensive information about gun-related violence in the U.S. and then post and disseminate it online.

Read More

CHARTS AND MAPS



THE NEW OLD AGE
As Their Numbers Grow, Home Care Aides Are Stuck at $10.11

Will Women Play Major League Baseball? (And Not Just on TV)

WELL
Too Old to Donate Blood?

HEALTH | VITAL SIGNS: CAUSE AND EFFECT

VITAL SIGNS: CAUSE AND EFFECT; Linking Guns and Gun Violence

By ERIC NAGOURNEY MAY 27, 2003

People with guns in their homes are almost twice as likely to be killed by guns as people who do not keep them at home, researchers reported yesterday in The Annals of Emergency Medicine.

And, the researchers found, people with guns are 16 times as likely to commit suicide using guns.

The explanation may lie in the unforgiving nature of firearms, said the author of the study, Dr. Douglas J. Wiebe, who conducted the research at the University of California at Los Angeles and is now at the University of Pennsylvania.

"People who are shot are substantially more likely to die than people injured with nongun weapons," Dr. Wiebe said.

The study was based on a review of the deaths of 1,720 homicide victims and 1,959 suicide victims and a sampling of American adults.

It found that most of the victims, over 56 percent, knew their assailants. A fifth of the homicides occurred during robberies, 6 percent during drug deals and about 15 percent during family arguments.

The study also found that women were significantly more likely than men to be victims of gun homicides. "This likely reflects the singular danger faced by women in abusive relationships," Dr. Wiebe wrote.

Douglas Wiebe
Professor of Epidemiology
University of Pennsylvania
Perelman School of Medicine

Congressman Who Restricted Gun Violence Research Has Regrets

The politics of gun control were as divisive in the 1990s as they are today. Republicans had won big in the '94 elections by campaigning against President Bill Clinton's gun control legislation. And in the spring of 1996, the National Rifle Association and its allies set their sights on the Centers for Disease Control and Prevention for funding increasingly assertive studies on firearms ownership and the effects on public health. The gun rights advocates claimed the research veered toward advocacy and covered such logical ground as to be effectively useless.

At first, the House tried to close down the CDC's entire, $46 million National Center for Injury Prevention. When that failed, Dickey stepped in with an alternative: strip $2.6 million that the agency had spent on gun studies that year. The money would eventually be re-appropriated for studies unrelated to guns. But the far more damaging inclusion was language that stated, "None of the funds made available for injury prevention and control at the Centers for Disease Control and Prevention may be used to advocate or promote gun control."

Dickey proclaimed victory — an end, he said at the time, to the CDC's attempts "to raise emotional sympathy" around gun violence. But the agency spent the subsequent years petrified of doing *any* research on gun violence, making the costs of the amendment clear even to Dickey himself.

Congressman Who Restricted Gun Violence Research Has Regrets

"Compared to five years ago, the funding picture for a few of us who have done this work for a long time is rosy," Wintemute said. "Compared to what it requires, it is still bleak. We have lost 20 years of concentrated effort."

Others have found the field fairly difficult to traverse. Dr. Douglas Wiebe, an associate professor of epidemiology at the Perelman School of Medicine at the University of Pennsylvania, worked on a 2009 study on the link between gun possession and gun assault that is believed to have sparked Congress' interest in applying the Dickey amendment to the NIH. He called the restriction of funds "not fatal" to his field, "but very close to it." Investigators, he explained, are being forced toward less-politically contentious studies, which makes it close to impossible to conduct sound epidemiological research.

Why the CDC Hasn't Launched a Comprehensive Gun Study in 15 Years

By JULIE BARZILAY, DR. LAURA JOHNSON and GILLIAN MOHNEY •
Jun 16, 2016, 4:37 PM ET

Scott Olson/Getty I

WATCH | American Medical Association Calls Gun Violence a 'Public Health Crisis'

The U.S. Centers for Disease Control and Prevention studies a variety of public health threats every year, from infectious diseases to automobile safety. But for 15 years, the CDC has avoided comprehensive research on one of the top causes of death in the U.S.: firearms.

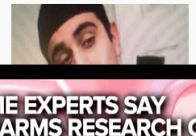
Why the
Launched
Study in

Feds Invi
to 'Consu
Pipeline

What Co
Know Ab
Security

Teen Wh
Sprayed

Dentist A
Harming
Making M



'I Did the
Orlando

National Institutes of Health

Table 1 Major NIH research awards and cumulative morbidity for select conditions in the US, 1973–2002

Condition	Total cases	NIH research awards
Cholera	373	101
Diphtheria	1337	54
Polio	266	106
Rabies	55	59
Total of four diseases	**2031**	**320**
Firearm injuries	>3000000	3

Branas, Wiebe, Schwab, Richmond. Getting past the “f” word in federally funded public health research. Injury Prevention 2005; 11:191-192

Want to answer questions like

Too much emphasis on mass shootings has a cost

America's gun control debate continues to revolve around
xact circumstances of the shooting that is currently on
ews. Is a new gun law worth it, or not? That depends on
her it might have prevented this particular shooting.
e this is an understandable, human response, it is a
le way to go about saving lives.

hock and horror that follows mass shootings has led to
sessive focus on the dangers of military-style rifles -
though rifles of any kind were used in less than 3% of
gun murders in 2014, according to FBI data.

A tunnel focus on mass shootings has also fueled the public

How many deaths result from mass shootings compared to other gun crimes? How has this changed over time?

Want to answer questions like

PUBLICITY AND PUBLIC HEALTH

The Science Behind Suicide Contagion

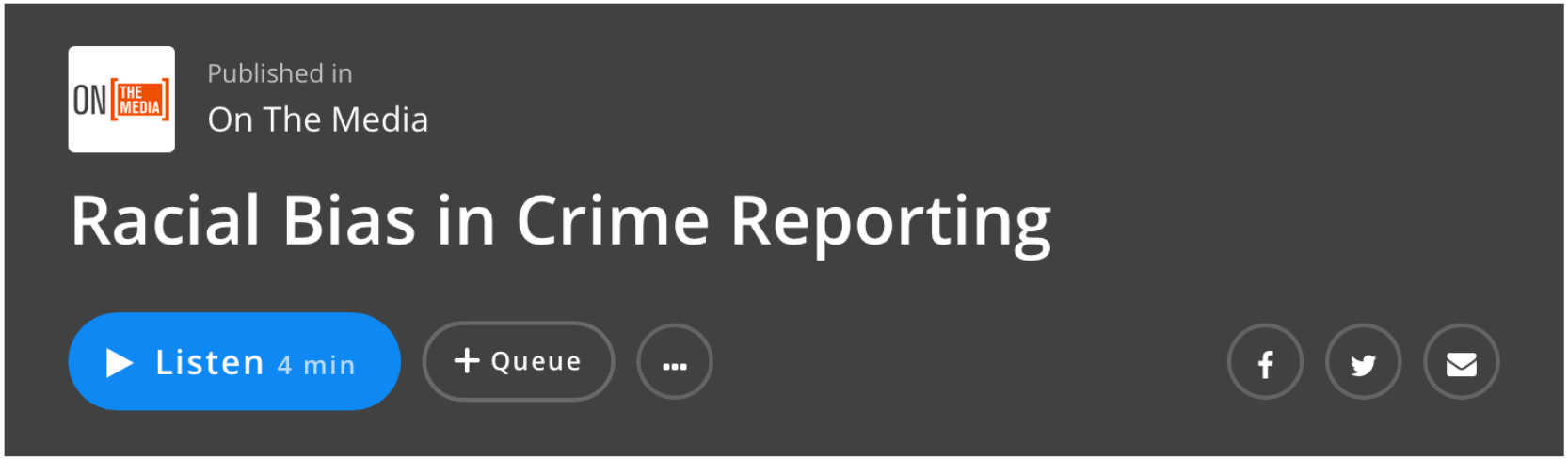
Margot Sanger-Katz @sangerkatz AUG. 13, 2014

When Marilyn Monroe died in August 1962, with the cause listed as probable suicide, the nation reacted. In the month was extensive news coverage, widespread sorrow suicides. According to one study, the suicide rate jumped by 12 percent compared with the same mo year.

Mental illness is not a communicable disease, but of evidence that suicide is still contagious. Publici suicide has been repeatedly and definitively linke increase in suicide, especially among young people. Analysis suggests that at least 5 percent of youth suicides are influenced by contagion.

How strong is the effect of suicide contagion? Does it change with age, gender? Is it effected by the style of reporting?

Want to answer questions like



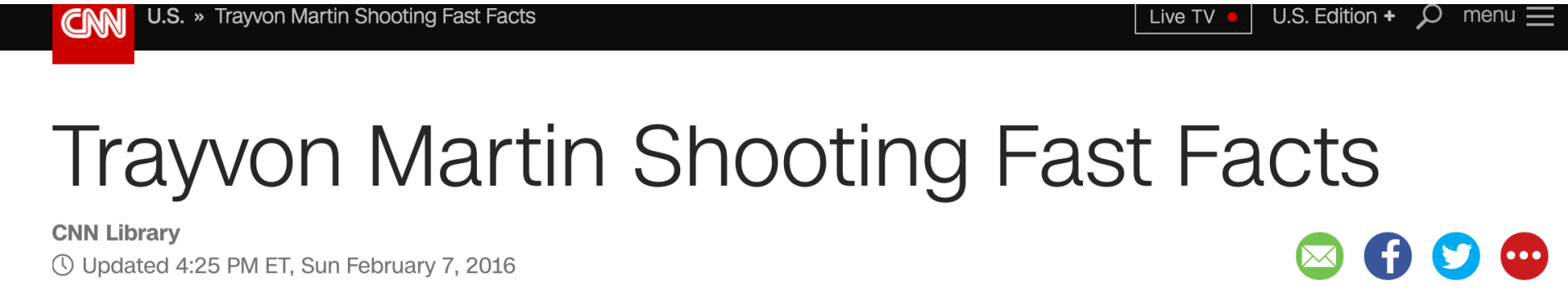
Jun 5, 2015

Does the media portray African-Americans differently than Whites in reporting on gun violence?

Summary Transcript

Research shows the media disproportionately depict African-Americans as criminals, and whites as victims. Brooke speaks with Nazgol Ghandnoosh, research analyst at The Sentencing Project, about her study, "Race and Punishment: Racial Perceptions of Crime and Support for Punitive Policies," which details how media distortions feed our own implicit biases. (And you can take Harvard's Implicit Association Test yourself here.)

Want to answer questions like



Can we predict events that will become politically relevant touchstone events?

The New York Times

Freddie Gray Case Ends With No Convictions of Any Police Officers

Require detailed, local data

Time and Place
- City
- State
- Other details (home, school, etc.)
- Date
- Clock Time
- Time of day

Alleged Shooter(s)
- Name
- Gender
- Age
- Race

Victim(s)
- Name
- Gender
- Age
- Race
- Was the victim injured?
- Was the victim hospitalized?
- Was the victim killed?

Circumstances of shooting
- Type of gun
- Number of shots fired
- Answer Yes/No/Not able to determine
- The shooter and the victim knew each other.
- The incident was a case of domestic violence.
- The firearm was used during another crime.
- The firearm was used in self defense.
- Alcohol was involved.
- Drugs (other than alcohol) were involved.
- The shooting was self-directed.
- The shooting was a suicide or suicide attempt.
- The shooting was unintentional.
- The shooting was by a police officer.
- The shooting was directed at a police officer.
- The firearm was stolen.
- The firearm was owned by the victim/victim's family

The Gun Report

The Opinion Pages

Joe Nocera

Go to Joe Nocera Home

GUN REPORT

The Gun Report: May 30, 2014

MAY 30, 2014 3:32 PM 314 Comments

The Kalashnikov family of assault rifles. Alexander Vasilkov/Wikimedia Commons

Recent shootings involving children have rocked two American cities.

Michael Day, 13, died after being caught in the crossfire between two groups in the Edison Neighborhood of Kalamazoo, Mich., on Memorial Day. This wasn't even the first time Day had been a victim of gun violence: On April 6, he was shot in the back while leaving a party. He told police he was walking when he heard a gunshot and realized he had been hit.

Victor Manuel Garay, 15, has been accused of firing the shot that killed Day. Police had been called earlier in the day to break up the large brawl, but as soon as they left, the fighting continued. If charged as an adult, Garay could face life in prison without the possibility of parole.

Published in
On The Media

The End of the Gun Report

Listen 10 min + Queue ...

Jennifer Mascia described how she wrote the Gun Report in an NPR interview

JENNIFER MASCIA: Well, I would google "**shooting**," "**man shot**," "**woman shot**," "**child shot**," "**teen shot**" and "**accidentally shot**". You know, this was all day one coverage of shootings, so a lot of times the details aren't flushed out. If there was no name and scant details, I had to skip over those. So each day, there'd be about 35 to 40 shootings that I would present.

Title	Description	Url	Source	Phrase
Man Shot Near Pierce Park in Coral Gables: Police	A man is in the hospital recovering after police say he was shot multiple times near a park in Coral Gables. The incident happened just before 3 a.m. Thursday outside Pierce Park, located at 101 Oak Avenue. Coral Gables police say the victim, a man in his ...	http://www.nbcmiami.com/news/local/Man-Shot-Near-Pierce-Park-in-Coral-Gables-Police-360424171.html	NBC Universal Media	man shot
I'm Voting for Hillary Clinton Because She's a Woman	"New drinking game: take a shot every time Hillary says 'as a woman' or 'as the first woman president,'" quips a straight white male on Facebook. This comment was part of a larger thread of young male Democrats discussing why Bernie Sanders is a better ...	http://www.huffingtonpost.com/jillian-gutowitz/im-voting-for-hillary-clinton-because-shes-a-woman_b_8684910.html	The Huffington Post	woman shot
Child attends school despite allegedly being stabbed by guardian - FOX10 News | WALA	... police said a 12-year-old child ran away from home after originally saying the child had been abandoned. A suspect was killed by officers after police said he shot at them while trying escape authorities in downtown Atlanta. A suspect was killed by ...	http://www.fox10tv.com/story/30649388/child-attends-school-despite-allegedly-being-stabbed-by-guardian-mother	WALA-TV FOX10	child shot
Palestinian who attacked soldier shot dead	The organisation has not claimed responsibility for killing the Henkins, who were shot in front of their young children as they drove on a West Bank road between the northern colonies of Itamar and Elon Moreh. The October 1 shooting was followed two days ...	http://gulfnews.com/news/mena/palestine/palestinian-who-attacked-soldier-shot-dead-1.1631076	Gulf News	child shot
12-year-old child abandoned at Atlanta police station	A 6-year-old child is hospitalized after allegedly being stabbed in the face by his mother Wednesday morning. A suspect was killed by officers after police said he shot at them while trying escape authorities in downtown Atlanta.More >> A suspect was ...	http://www.walb.com/story/30654553/12-year-old-child-abandoned-at-atlanta-police-station	WALB 10 News	child shot

The Gun Report: Training data

http://www.mlive.com/news/kalamazoo/index.ssf/2014/04/kalamazoo_teenager_13_shot_and.html

http://www.jsonline.com/news/crime/new-developments-in-playground-shooting-to-be-announced-at-430-pm-b99278118z1-260682381.html

http://www.mlive.com/news/kalamazoo/index.ssf/2014/05/fighting_led_up_to_fatal_shoot.html

http://www.mlive.com/news/kalamazoo/index.ssf/2014/05/michael_day_kalamazoo.html

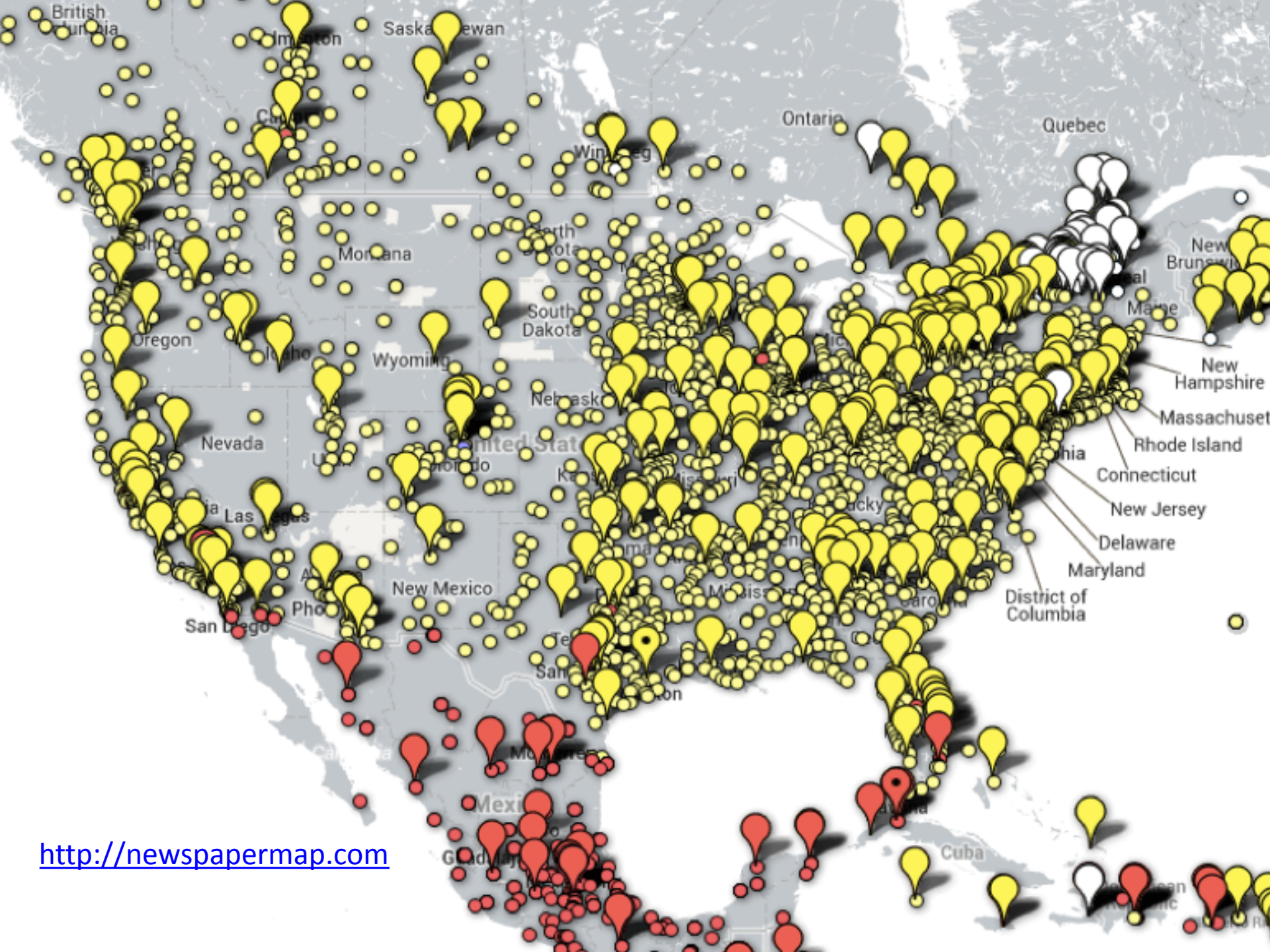
http://www.mlive.com/news/kalamazoo/index.ssf/2014/05/15-year-old_charged_with_murde.html

http://www.jsonline.com/news/crime/girl-10-on-life-support-after-being-hit-in-playground-shootout-b99275748z1-260251491.html

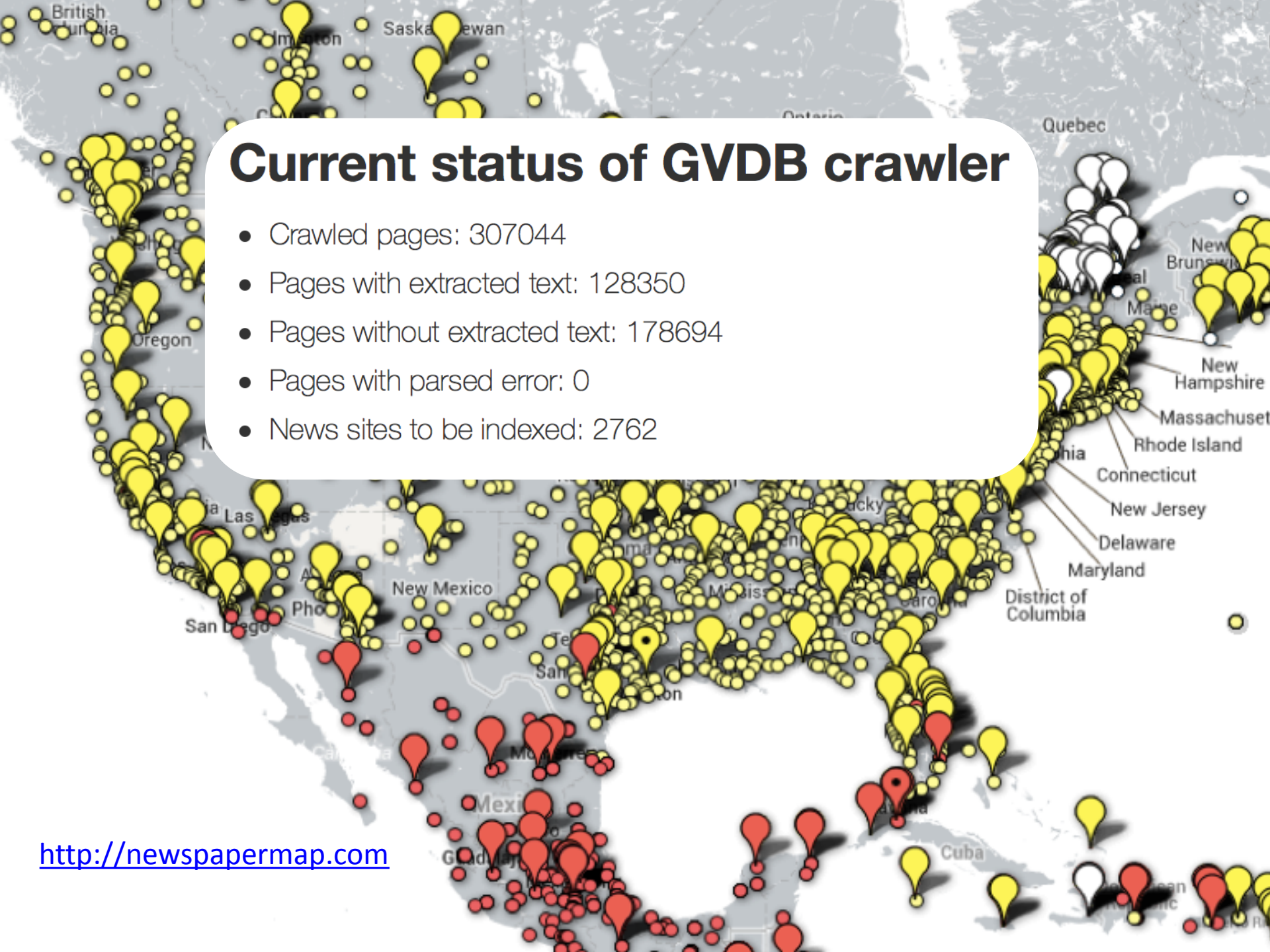
classifier.py

```python
#!/bin/python

import os
import sys
import string
import random
import operator
from sklearn.tree import export_graphviz
from sklearn.tree import DecisionTreeClassifier
from sklearn.naive_bayes import MultinomialNB
from sklearn.linear_model import LogisticRegres
from sklearn.preprocessing import LabelEncoder
from sklearn.feature_extraction import DictVect
from sklearn.cross_validation import train_test
from sklearn.externals.six import StringIO

#read in raw data from file and return a list o
def get_data(filename):
    data = [line.strip().split('\t') for line i
    random.shuffle(data)
    return data

#this function builds the feature matrix for th
def get_dtree_features(X) :
    features = []
    feature_list = []
    wordCounts = {}

    for x in X :
        f = {}
        for w in [word.strip(string.punctuation
            if not w in wordCounts :
                wordCounts[w] = 0
            wordCounts[w] += 1
    for w in wordCounts:
        if wordCounts[w] > 5000:
```

```
Statistical classification
2.891467        shooting
2.560138        accidentally
2.342422        shot
2.012679        gun
1.925938        accidental
1.706036        subscriber
1.673353        guns
1.626867        skip
1.597359        discharged
1.505322        rifle
1.449963        homicides
1.419679        hunting
1.418678        bullet
1.324652        ad
1.279014        log
1.225336        source
1.214517        chest
1.211558        reserved
1.204958        gunshot
1.147636        detroit
1.142863        innovative
1.136201        penny
1.117853        leg
1.108006        unlimited
1.098322        township
1.095423        update
1.094982        firearms
1.092020        marion
```



http://newspapermap.com


Current status of GVDB crawler
Crawled pages: 307044
Pages with extracted text: 128350
Pages without extracted text: 178694
Pages with parsed error: 0
News sites to be indexed: 2762
http://newspapermap.com
Quebec
Oregon
New Hampshire
Rhode Island
Connecticut
New Jersey
Delaware
Maryland
District of Columbia
New Mexico
Cuba

Work on tasks

Read more about our tasks and select how you want to contribute.

Extract information

We want data that is easy for researchers to search and study. We need your help reading articles about gun violence and extracting key pieces of information (such as the location of the shooting or the name and age of the victim).

Total submitted tasks: 15526

Total available tasks: 5341

Go to task

Scan the Headlines

Read headlines and tell us which ones describe incidents of gun violence.

Total submitted tasks: 5408

Total available tasks: 882

Go to task

Combine records

We want to have as complete a database as possible. When there are multiple reports of the same incident, we want to combine the information from all the articles so nothing is left out. You can help by comparing two records and deciding which information is best to keep in the database.

Total submitted tasks: 297

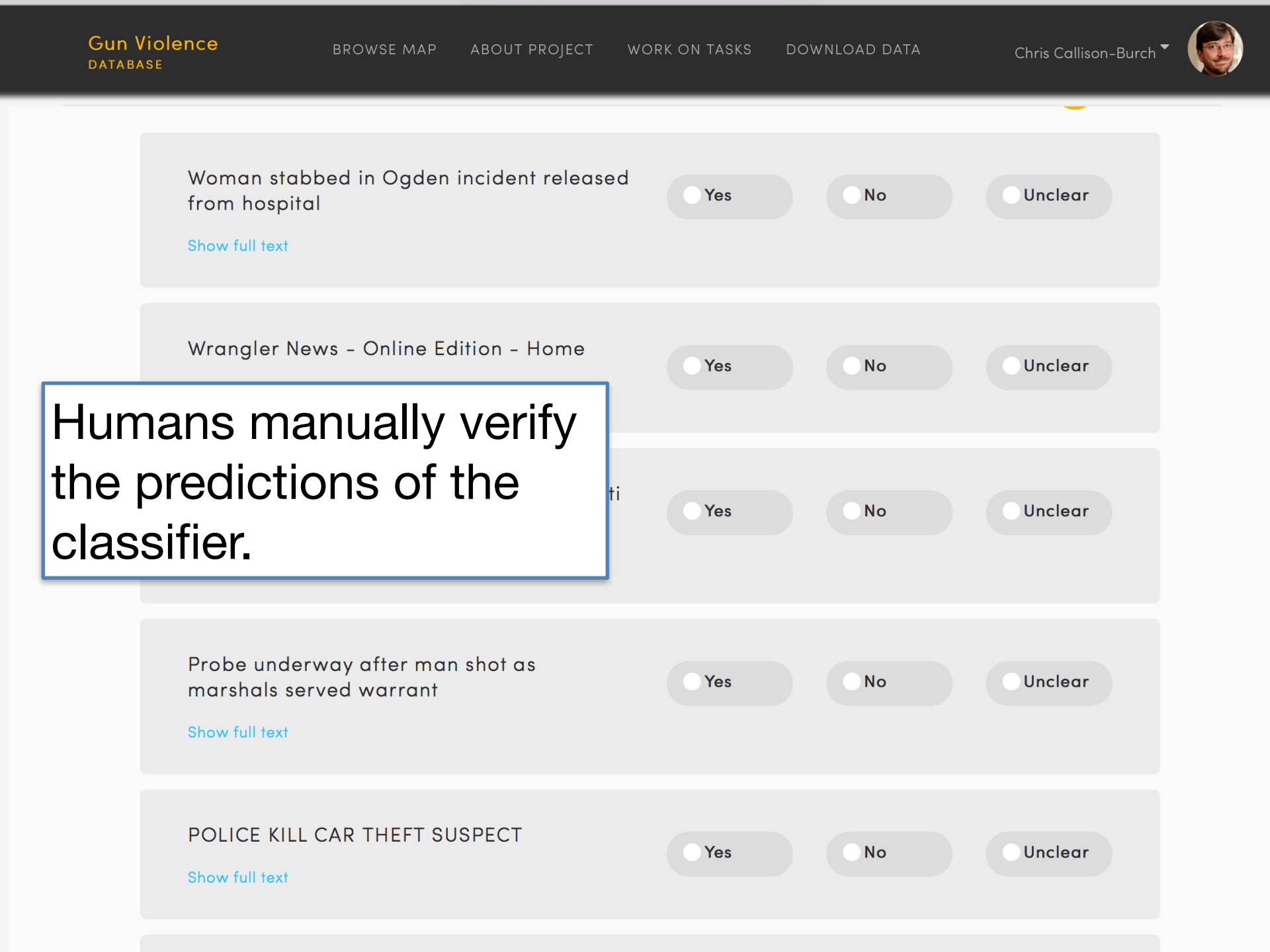
Total available tasks: 7

Go to task

Identify articles

Compare Articles

Identify People


Gun Violence
DATABASE
BROWSE MAP
ABOUT PROJECT
WORK ON TASKS
DOWNLOAD DATA
Chris Callison-Burch
Woman stabbed in Ogden incident released from hospital
Show full text
Yes
No
Unclear
Wrangler News - Online Edition - Home
Yes
No
Unclear
Humans manually verify the predictions of the classifier.
Yes
No
Unclear
Probe underway after man shot as marshals served warrant
Show full text
Yes
No
Unclear
POLICE KILL CAR THEFT SUSPECT
Show full text
Yes
No
Unclear

Woman stabbed in Ogden incident released from hospital

Show full text

Yes No Unclear

Wrangler Nev

Show full text

Unclear

LIVE Budget s
clash in Rajy

Show full text

Unclear

Probe underway after man shot as marshals served warrant

Show full text

Yes No Unclear

POLICE KILL CAR THEFT SUSPECT

Show full text

Yes No Unclear

Wrangler News - Online Edition - Home

http://www.wranglernews.com/053108.htm

If your kids are enrolled in the Hoops Star camp this summer at Kiwanis Park, you can expect they?ll be learning basketball fundamentals and shooting skills from a pair of coaches who really know the game, namely the two Sam Duanes, senior and junior. Calendar | Classifieds | Contact Us | Home | Make a Payment | Media Kit | Online Advertising | Online Map | Online Pages | Previous Issues | Submit Your Ad Copyright ? 2008 Wrangler News

PUBLICATION DATE: AUGUST 1, 2016

Police: Officer shoots, wounds shoplifting suspect outside Conroe Wal-Mart - Houston Chronicle

Conroe police say an officer shot and wounded a suspected shoplifter in the parking lot of a Wal-Mart store Monday - the second officer-related shooting of someone suspected of taking merchandise from a city Wal-Mart store in less than three years. According to investigators, Fillmore was observed concealing several items of merchandise and then leaving the store when a Wal-Mart employee attempted to stop him. A Conroe police officer who responded to the incident feared for the safety of the Wal-Mart employee and other citizens in the busy parking lot and fired one shot, wounding the suspect in his left shoulder. Fillmore has convictions dating to 1977 for various offenses including home burglaries, thefts and illegal drug possession, public records show.

First answer a series of binary questions about the circumstances of the shooting….

Please read the text carefully, and then select an answer for all questions. Please base your answers only on information that is explicitly stated or can be confidently inferred from the text of the article.

The shooting was unintentional.

Yes | No | Not Mentioned

The shooting was by a police officer.

Yes | No | Not Mentioned

The shooting was directed at a police officer.

Yes | No | Not Mentioned

The firearm was stolen.

Yes | No | Not Mentioned

PUBLICATION DATE: JULY 26, 2016

Man shot in North Baltimore and checks himself into hospital

A 32-year-old man was shot in North Baltimore and checked himself into a hospital Wednesday evening, police said. Detectives determined the man was shot on the 4600 block of Midwood Ave. in the Winston-Govans neighborhood, police said. Officers had been called e did not release the man's condition. Anyone with in -396-2221.

Clear this highlight

Clock Time (1p.m., 2:37a.m)

Additional Location Details

…then extract structured information from text to populate the database.

First, try to figure out the date of the described event, and select it by clicking on the calendar icon. The publication date and the day of week mentioned in article are helpful in determining the date of the shooting. Next, click on parts of the text that correspond to the other information listed below, if that information is present in the article. When you highlight a passage of text in the article, you will get a dropdown menu that lets you select which question it answers.

Date

2016-07-20

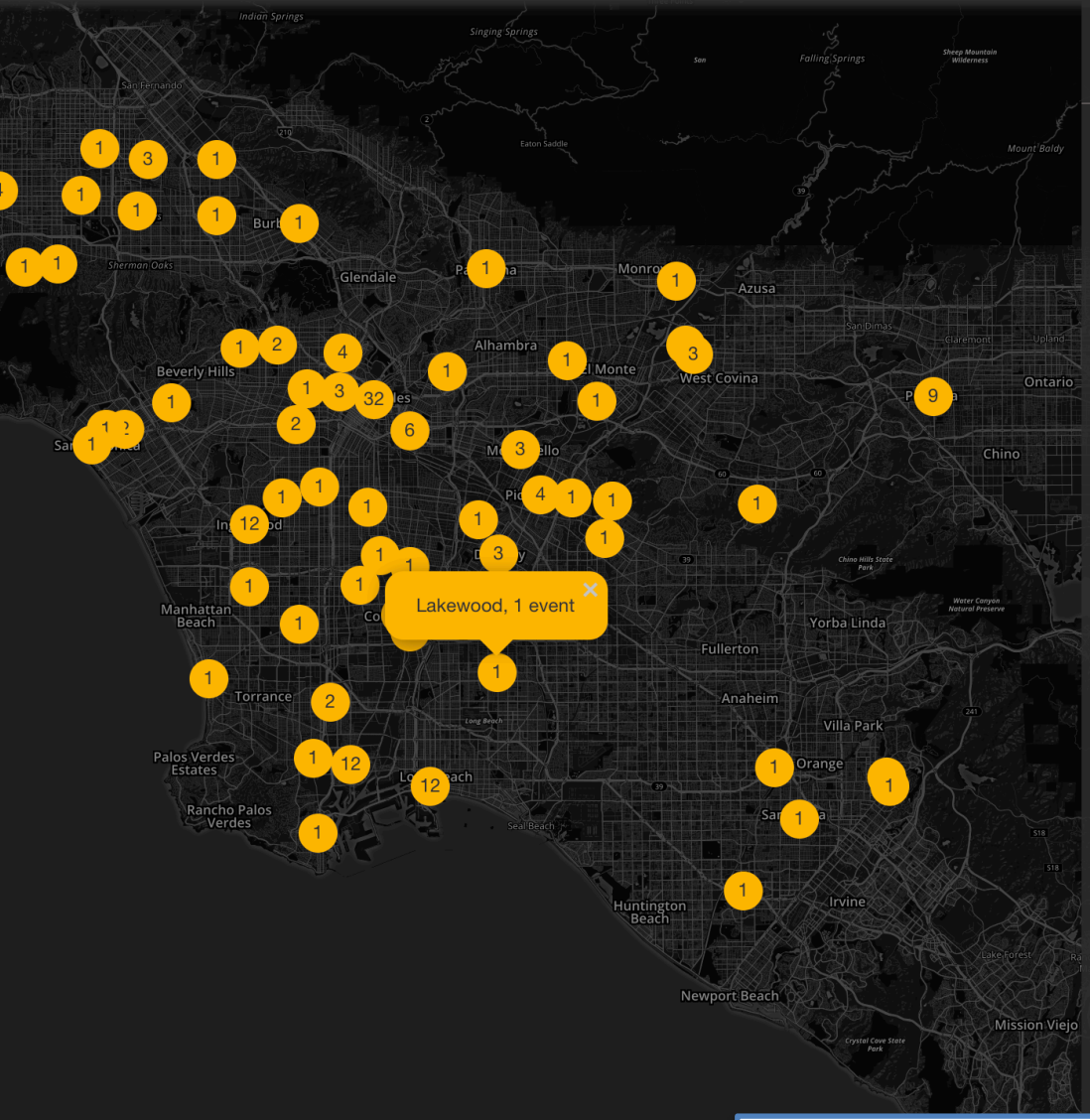
State

MD - Maryland

City

Baltimore

Clock Time (1p.m., 2:37a.m)



›› Lakewood

‹ 1 events

"Richard Wayne Van Heyningen, a 47-year-old white man, was shot and killed Tuesday, March 31 near Hardwick Street and Fanwood Avenue in Lakewood, according to the Los Angeles County coroner's office. The victim was riding his bicycle on Hardwick Street when a pickup truck driven by his brother came up behind him and ran into his bicycle, causing him to fall into the street, according to a news release from the Los Angeles County Sheriff's Department."

Read the article

Number of victims: 1

VICTIM #1

NAME: Richard Wayne Van Heyningen
AGE: 47
RACE: white
GENDER: Male
VICTIM WAS: killed

Work on task

We plot the extracted information on a map

Automating The Pipeline

Chicago Police release Laquan McDonald shooting video | National News

Three seconds. On a dashcam video clock, that's the amount of time between the moment when two officers have their guns drawn and the point when Laquan McDonald falls to the ground. The video, released to the public for the first time late Tuesday, is a key piece of evidence in a case that's sparked protests in Chicago and has landed an officer behind bars. The 17-year-old McDonald was shot 16 times on that day the video shows in October 2014. Chicago police Officer Jason Van Dyke was charged Tuesday with first-degree murder….

Incident #1053

City	
Date	
Shooter	
Victim	
Victim Killed	

Person #1014

Name	
Gender	
Age	
Race	

Automating The Pipeline

Chicago Police release Laquan McDonald shooting video | National News

Three seconds. On a dashcam video clock, that's the amount of time between the moment when two officers have their guns drawn and the point when Laquan McDonald falls to the ground. The video, released to the public for the first time late Tuesday, is a key piece of evidence in a case that's sparked protests in Chicago and has landed an officer behind bars. The 17-year-old McDonald was shot 16 times on that day the video shows in October 2014. Chicago police Officer Jason Van Dyke was charged Tuesday with first-degree murder….

Named Entity Recognition

Person #1014

Name	Laquan McDonald
Gender	
Age	
Race	

- Named Entity Recognition without Gazetteers. Mikheev et al (1999).
- Neural Architectures for Named Entity Recognition. Lample et al (2016).

Automating The Pipeline

Chicago Police release Laquan McDonald shooting video | National News

Three seconds. On a dashcam video clock, that's the amount of time between the moment when two officers have their guns drawn and the point when Laquan McDonald falls to the ground. The video, released to the public for the first time late Tuesday, is a key piece of evidence in a case that's sparked protests in Chicago and has landed an officer behind bars. The 17-year-old McDonald was shot 6 times on that day the video shows in October 2014. Chicago police Officer Jason Van Dyke was charged Tuesday with first-degree murder….

- Joint event extraction via structured prediction. Li et al (2013).
- Event Detection and Domain Adaptation with Convolutional Neural Networks. Nguyen and Grishman (2015).

Event Detection

Person #1014

Name	Laquan McDonald
Gender	
Age	
Race	

Incident #1053

City	
Date	
Shooter	
Victim	
Victim Killed	

Automating The Pipeline

Chicago Police release Laquan McDonald shooting video | National News

Three seconds. On a dashcam video clock, that's the amount of time between the moment when two officers have their guns drawn and the point when Laquan McDonald falls to the ground. The video, released to the public for the first time late Tuesday, is a key piece of evidence in a case that's sparked protests in Chicago and has landed a police officer behind bars. The 17-year-old McDonald was shot 6 times on that day the video shows in October 2014. Chicago police Officer Jason Van Dyke was charged Tuesday with first-degree murder….

- Semantic Role Labeling. Palmer et al (2010).
- Joint event extraction via structured prediction with global features. Li et al (2013).

Semantic Role Labeling

Person #1014

Name	Laquan McDonald
Gender	
Age	
Race	

Incident #1053

City	
Date	
Shooter	
Victim	McDonald
Victim Killed	

Automating The Pipeline

Entity Coreference

Chicago Police release Laquan McDonald shooting video | National News

Three seconds. On a dashcam video clock, that's the amount of time between the moment when two officers have their guns drawn and the point when Laquan McDonald falls to the ground. The video, released to the public for the first time late Tuesday, is a key piece of evidence in a case that's sparked protests in Chicago and has landed an officer behind bars. The 17-year-old McDonald was shot 16 times on that day the video shows in October 2014. Chicago police Officer Jason Van Dyke was charged Tuesday with first-degree murder….

Person #1014

Name	Laquan McDonald
Gender	
Age	
Race	

Incident #1053

City	
Date	
Shooter	
Victim	Person #1014
Victim Killed	

- Coreference Resolution in a Modular, Entity-Centered Model. Haghighi and Klein (2010).
- Using Wikitology for Cross-Document Entity Coreference Resolution. Finin et al (2009).

Automating The Pipeline

Event Coreference

Chicago Police release Laquan McDonald shooting video | National News

Three seconds. On a dashcam video clock, that's the amount of time between the moment when two officers have their guns drawn and

Police release video of officer shooting teen | Oklahoma City

Protesters took to the streets of Chicago late Tuesday after police released a video showing an officer shooting 17-year-old Laquan McDonald…McDonald was a black teenager. The officer who shot him, Jason Van Dyke, is white. He was charged Tuesday with first-degree murder in McDonald's death…

- Cross-document Event Coreference Resolution based on Cross-media Features. Zhang et al (2015).

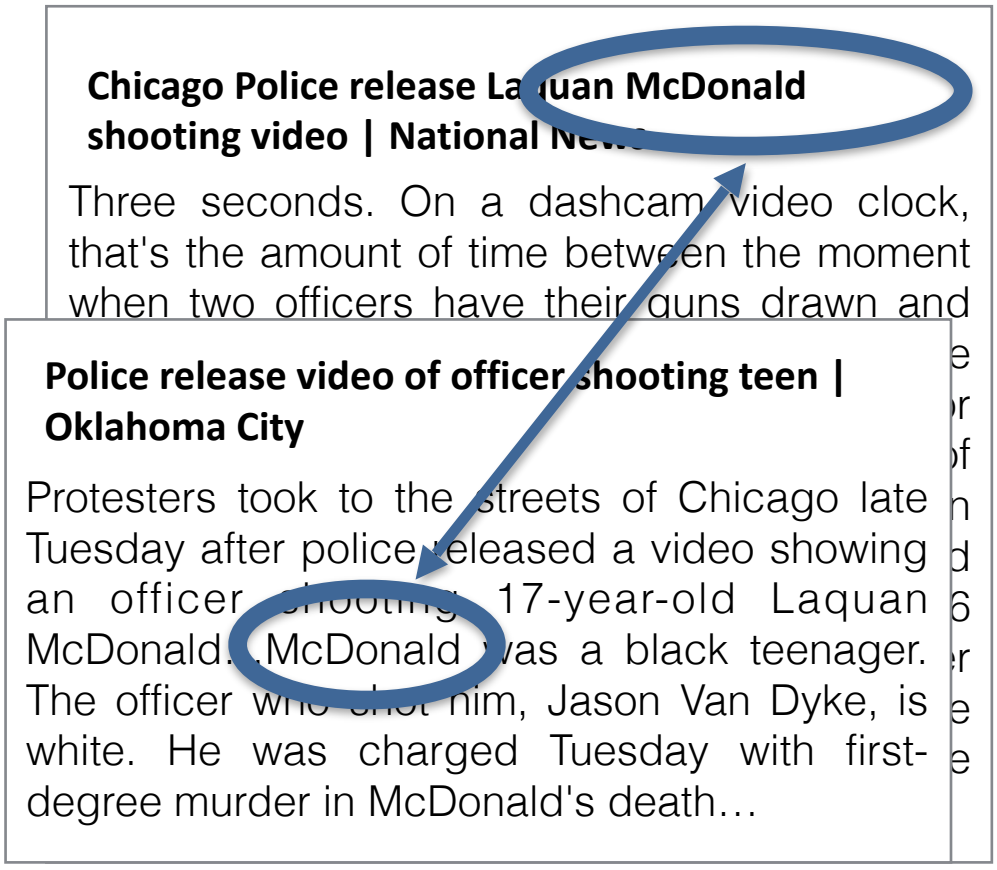
Person #1014

Name	Laquan McDonald
Gender	
Age	
Race	

Incident #1053

City	
Date	
Shooter	
Victim	Person #1014
Victim Killed	

Automating The Pipeline



- Cross-document Event Coreference Resolution based on Cross-media Features. Zhang et al (2015).

Cross-Document Entity Coref.

Person #1014

Name	Laquan McDonald
Gender	
Age	
Race	

Incident #1053

City	
Date	
Shooter	
Victim	Person #1014
Victim Killed	

Automating The Pipeline

Chicago Police release Laquan McDonald shooting video | National News

Three seconds. On a dashcam video clock, that's the amount of time between the moment when two officers have their guns drawn and

Police release video of officer shooting teen | Oklahoma City

Protesters took to the streets of Chicago late Tuesday after police released a video showing an officer shooting 17-year-old Laquan McDonald.. McDonald was a black teenager. The officer who shot him, Jason Van Dyke, is white. He was charged Tuesday with first-degree murder in McDonald's death…

- Cross-document Event Coreference Resolution based on Cross-media Features. Zhang et al (2015).

Semantic Parsing

Person #1014

Name	Laquan McDonald
Gender	
Age	17
Race	Black

Incident #1053

City	
Date	
Shooter	
Victim	Person #1014
Victim Killed	

Automating The Pipeline

Chicago Police release Laquan McDonald shooting video | National News

Three seconds. On a dashcam video clock, that's the amount of time between the moment when two officers have their guns drawn and

Police release video of officer shooting teen | Oklahoma City

Protesters took to the streets of Chicago late Tuesday after police released a video showing an officer shooting 17-year-old Laquan McDonald.. McDonald was a black teenager. The officer who shot him, Jason Van Dyke, is white. He was charged Tuesday with first-degree murder in McDonald's death…

- Cross-document Event Coreference Resolution based on Cross-media Features. Zhang et al (2015).

Semantic Parsing

Person #1014

Name	Laquan McDonald
Gender	
Age	17
Race	Black

Incident #1053

City	
Date	
Shooter	
Victim	Person #1014
Victim Killed	TRUE

NLP Systems Require Training Data

- Most existing training datasets consist of factoid and pop-culture questions

NLP Systems Require Training Data

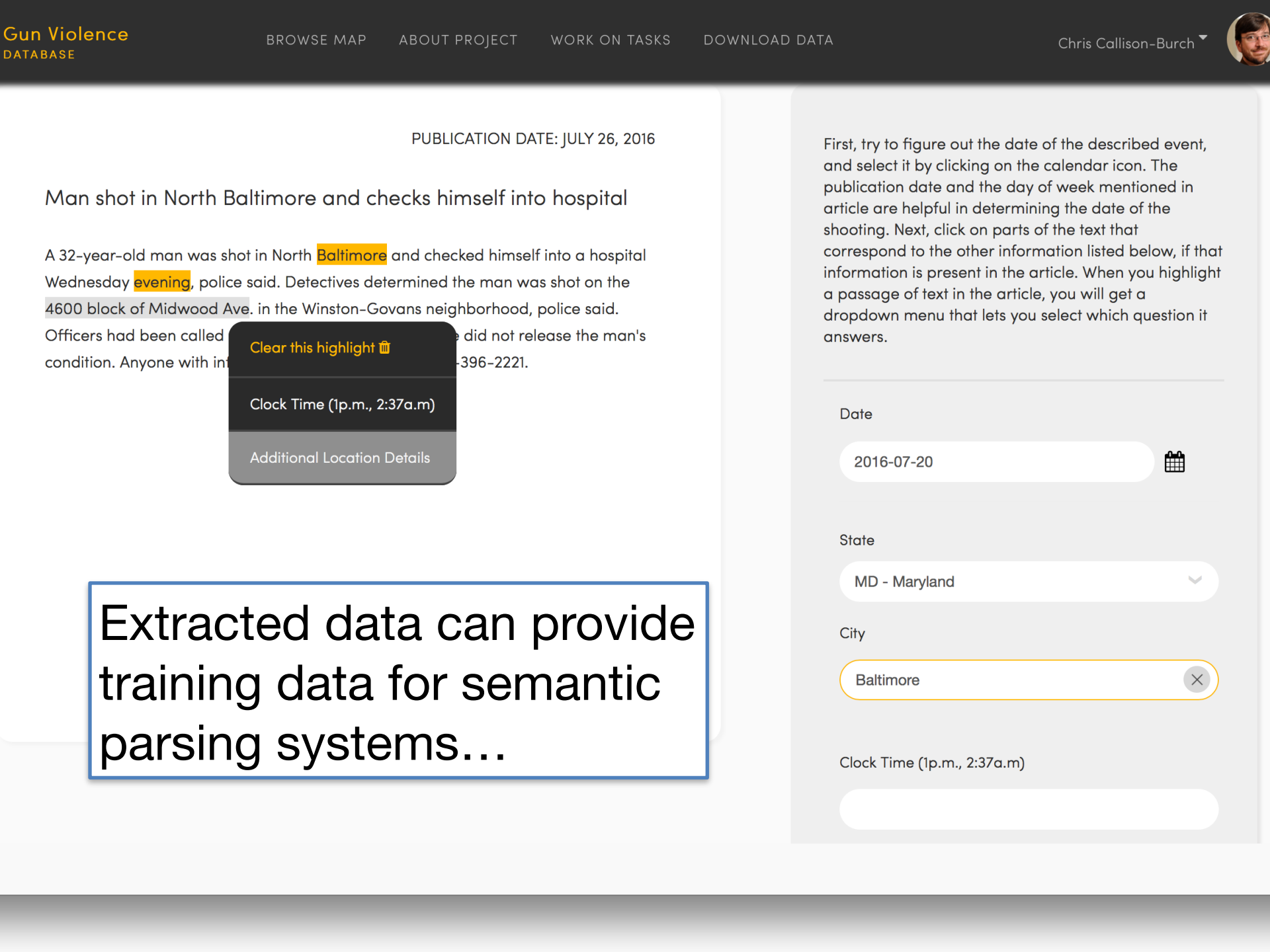
- Most existing training datasets consist of factoid and pop-culture questions
  - *What is the cover price of the X-Men comic book? (Berant et al., 2013)*
  - *How are Kim Kardashian and Kanye West related? (Wijaya, et al., 2015)*

NLP Systems Require Training Data

- Most existing training datasets consist of factoid and pop-culture questions
  - *What is the cover price of the X-Men comic book? (Berant et al., 2013)*
  - *How are Kim Kardashian and Kanye West related? (Wijaya, et al., 2015)*
- Usually assume entities are celebrities who have Wikipedia pages. *(Singh et al., 2012)*

NLP Systems Require Training Data

- Most existing training datasets consist of factoid and pop-culture questions
  - *What is the cover price of the X-Men comic book? (Berant et al., 2013)*
  - *How are Kim Kardashian and Kanye West related? (Wijaya, et al., 2015)*
- Usually assume entities are celebrities who have Wikipedia pages. *(Singh et al., 2012)*
- This leads makes it hard to extract the necessary data for in-depth social science research


Gun Violence
DATABASE
BROWSE MAP
ABOUT PROJECT
WORK ON TASKS
DOWNLOAD DATA
Chris Callison-Burch
PUBLICATION DATE: JULY 26, 2016
Man shot in North Baltimore and checks himself into hospital
A 32-year-old man was shot in North Baltimore and checked himself into a hospital Wednesday evening, police said. Detectives determined the man was shot on the 4600 block of Midwood Ave. in the Winston-Govans neighborhood, police said. Officers had been called
did not release the man's condition. Anyone with in
-396-2221.
Clear this highlight
Clock Time (1p.m., 2:37a.m)
Additional Location Details
First, try to figure out the date of the described event, and select it by clicking on the calendar icon. The publication date and the day of week mentioned in article are helpful in determining the date of the shooting. Next, click on parts of the text that correspond to the other information listed below, if that information is present in the article. When you highlight a passage of text in the article, you will get a dropdown menu that lets you select which question it answers.
Date
2016-07-20
State
MD - Maryland
City
Baltimore
Clock Time (1p.m., 2:37a.m)
Extracted data can provide training data for semantic parsing systems…

Gun Violence
DATABASE

BROWSE MAP ABOUT PROJECT WORK ON TASKS DOWNLOAD DATA Chris Callison-Burch

Read the texts below. Are both related to the same even?

?

Memphis Officer Shot During Traffic Stop | National News - WCVB Home

CRAWLED DATE: 8/3/2015

Authorities say Officer Sean Bolton died after he was shot during a traffic stop in Memphis Saturday night.
Authorities say Officer Sean Bolton died after he was shot during a traffic stop in Memphis Saturday night.

Police Identify Suspect In Memphis Officer's Killing

CRAWLED DATE: 8/3/2015

Police in Memphis identified a 29-year-old convicted bank robber as the man who allegedly killed a city cop during a traffic stop on Saturday. Police director Toney Armstrong on Sunday evening told reporters that officer Sean Bolton apparently interrupted a drug deal and that after "some type of physical altercation," Bolton was shot and killed by a passenger in the car. Armstrong said police were searching for 29-year-old Tremaine Wilbourn and had issued a warrant for his arrest. Wilbourn had been free on supervised release by the U.S. Western

….event coreference…

Gun Violence
DATABASE

BROWSE MAP ABOUT PROJECT WORK ON TASKS DOWNLOAD DATA Chris Callison-Burch

Below you can see list of victims/shooter that we got from two different articles. Please match people from the first article with people from the second article.

?

Read the article

Read article

Victim #1a

✓
No match
Victim #2a Officer Sean Bolton

NAME: Sean Bolton

AGE: unknown

RACE: unknown

GENDER: Male

VICTIM WAS: killed

Victim #2a

NAME: Officer Sean Bolton

AGE: 33

RACE: unknown

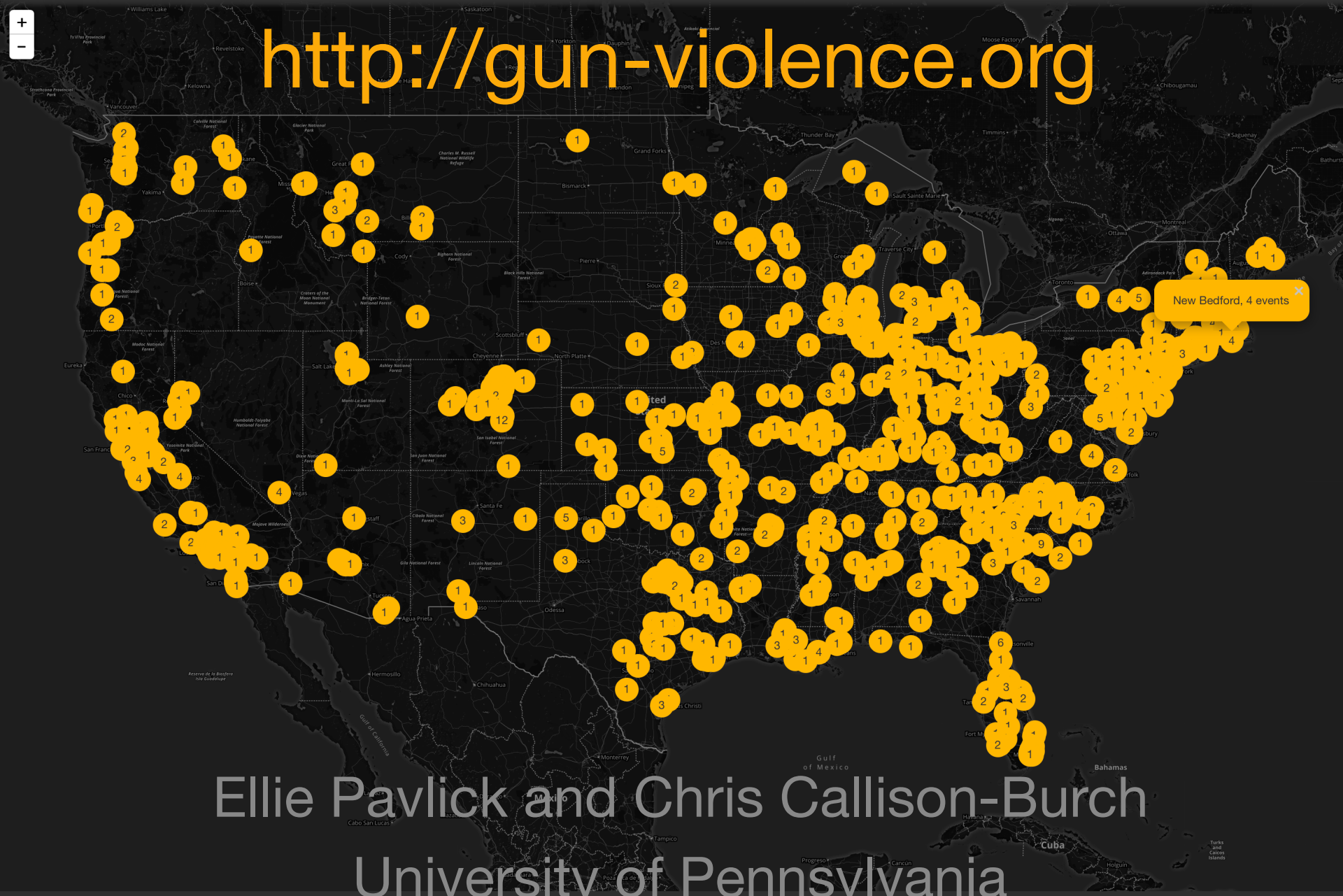
GENDER: Male

VICTIM WAS: killed

…and entity coreference.

The Gun Violence Database

- Use crowdsourcing to collected detailed local data to support research about gun violence in the United States
- Collect data that can be used to train state-of-the-art NLP systems to automate the pipeline, enabling databases to be updated in real time
- Constantly growing with new data and richer annotation



Ellie Pavlick and Chris Callison-Burch
University of Pennsylvania