

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
<i>Washington</i>	Person, Location, Political Entity, Organization, Vehicle
<i>Downing St.</i>	Location, Organization
<i>IRA</i>	Person, Organization, Monetary Instrument
<i>Louis Vuitton</i>	Person, Organization, Commercial Product

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

[PER Washington] was born into slavery on the farm of James Burroughs.

[ORG Washington] went up 2 games to 1 in the four-game series.

Blair arrived in [LOC Washington] for what may well be his last state visit.

In June, [GPE Washington] passed a primary seatbelt law.

The [VEH Washington] had proved to be a leaky ship, every passage I made...

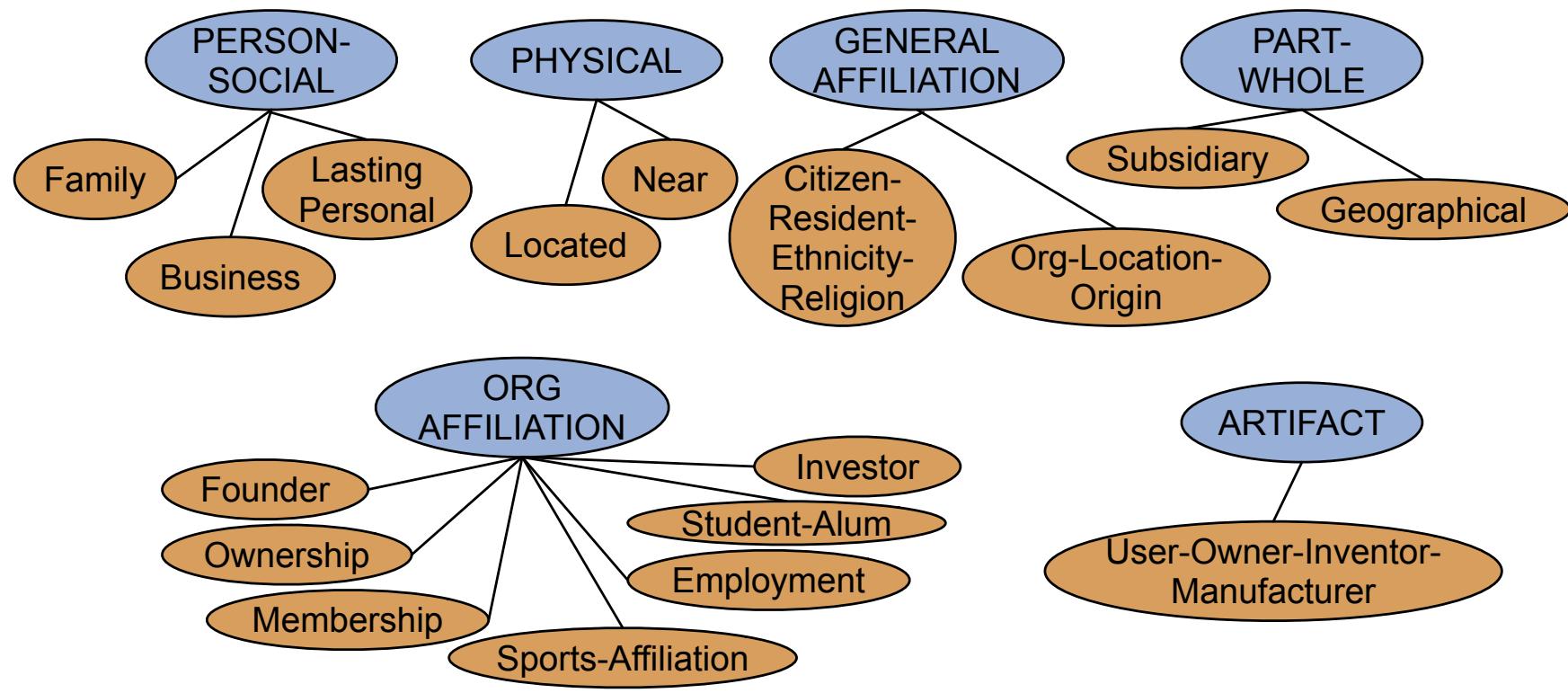
Relation Extraction

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

Robert Mueller



**Special Counsel for the
United States Department of Justice**

In office

May 17, 2017 – March 22, 2019

Appointed by Rod Rosenstein

6th Director of the Federal Bureau of Investigation

In office

September 4, 2001 – September 4, 2013

President George W. Bush

Barack Obama

Deputy Thomas J. Pickard

Bruce J. Gebhardt

John S. Pistole

Timothy P. Murphy

Sean M. Joyce

Preceded by Louis Freeh

Succeeded by James Comey

Acting United States Deputy Attorney General

In office

January 20, 2001 – May 10, 2001

President George W. Bush

Preceded by Eric Holder

Succeeded by Lanny Thompson

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](#)

[3510](#)

2007

S Auer, C Bizer, G Kobilarov, J Lehmann, R Cyganiak, Z Ives

The semantic web, 722-735

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**

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

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 patterns, 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)

Semisupervised Relation Extraction

function BOOTSTRAP(*Relation R*) **returns** *new relation tuples*

tuples \leftarrow Gather a set of seed tuples that have relation *R*  <Ryanair, hub, Charleroi>

iterate

sentences \leftarrow find sentences that contain entities in *tuples*

patterns \leftarrow generalize the context between and around entities in *sentences*

newpairs \leftarrow use *patterns* to grep for more tuples

newpairs \leftarrow *newpairs* with high confidence

tuples \leftarrow *tuples* + *newpairs*

return *tuples*

Budget airline Ryanair, which uses Charleroi as a hub, scrapped all weekend flights out

/ [ORG], which uses [LOC] as a hub /

All flights in and out of Ryanair's Belgian hub at Charleroi airport were grounded on F

/ [ORG]'s hub at [LOC] /

A spokesman at Charleroi, a main hub for Ryanair, estimated that 8000 passengers had

/ [LOC] a main hub for [ORG] /

Distant Supervision

place-of-birth

function DISTANT SUPERVISION(*Database D, Text T*) **returns** *relation classifier C*

foreach relation *R*

foreach tuple (e_1, e_2) of entities with relation *R* in *D*

sentences \leftarrow Sentences in *T* that contain *e1* and *e2*

f \leftarrow Frequent features in *sentences*

observations \leftarrow *observations* + new training tuple (e_1, e_2, f, R)

C \leftarrow Train supervised classifier on *observations*

return *C*

<Edwin Hubble, Marshfield>,
<Albert Einstein, Ulm>, ...
and 100,000 more

...Hubble was born in Marshfield...

...Einstein, born (1879), Ulm...

...Hubble's birthplace in Marshfield...

Open IE

- Unsupervised relation extraction
- Find all *strings of words* that satisfy the triple 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	<i>morning, noon, night, winter, dusk, dawn</i>
Proper Noun	<i>January, Monday, Ides, Easter, Rosh Hashana, Ramadan, Tet</i>
Adjective	<i>recent, past, annual, former</i>
Adverb	<i>hourly, daily, monthly, yearly</i>

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.

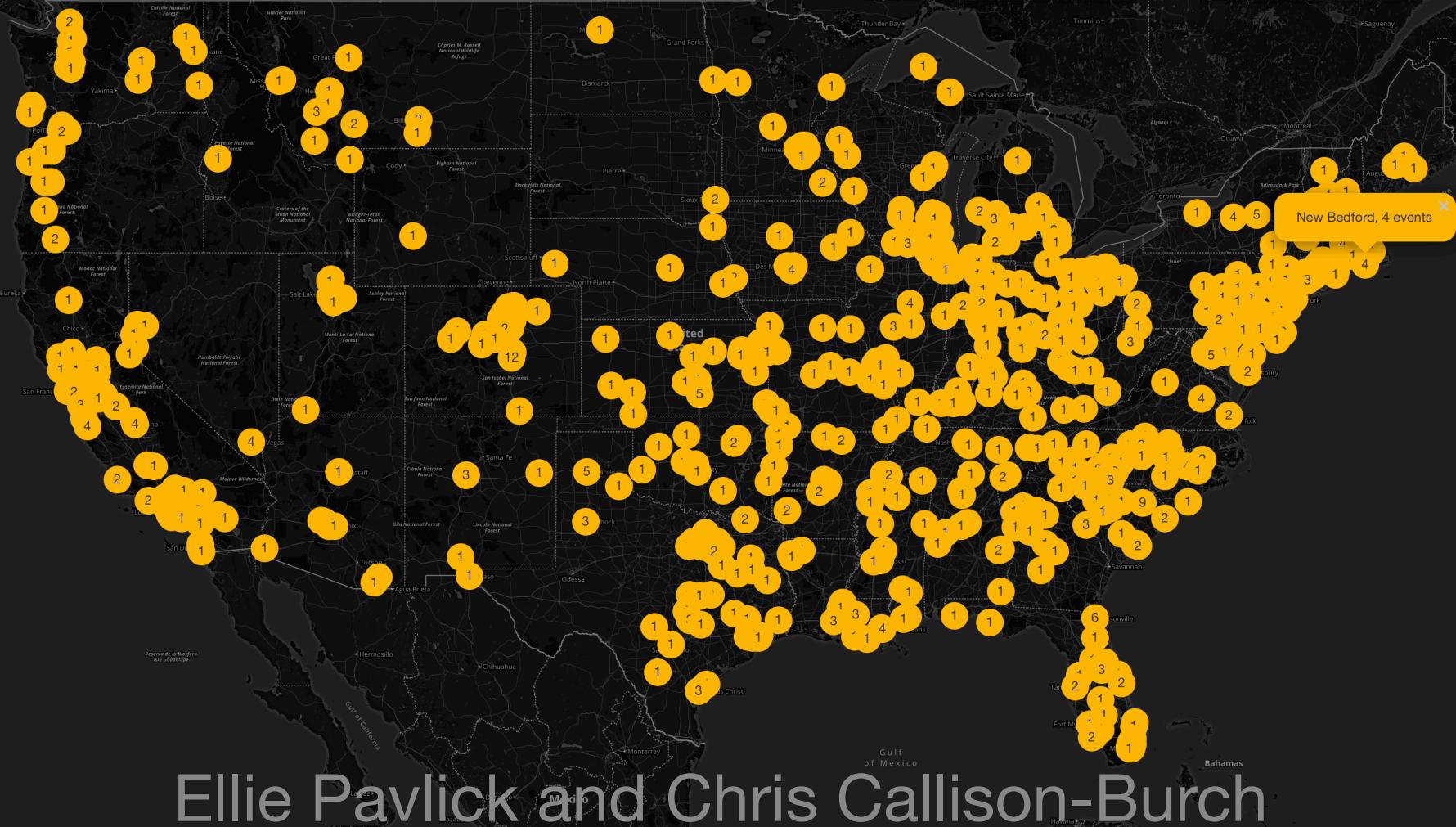
- Soaring_{e1} is **included in** the fiscal first quarter_{t58}
- Soaring_{e1} is **simultaneous with** the bucking_{e3}
- Declining_{e4} **includes** soaring_{e1}

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](#)

<http://crowdsourcing-class.org>

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.

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

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

Fatal Encounters

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

Search

CRIME MURDER, THEFT, AND OTHER WICKEDNESS.

SEPT. 16 2013 3:34 PM

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.



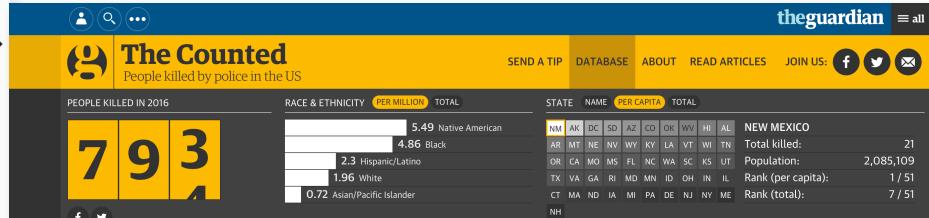
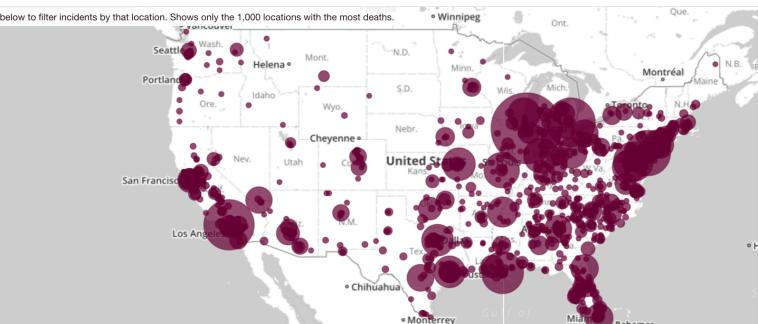
122k

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.



2016 2015 List Map

Search by name:

Filter by:

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

LOGIN CONTACT US Search Database

GUN VIOLENCE Archive



PHOTO CREDIT: DAVID LASSMAN, 2007

GUN VIOLENCE Archive 2016

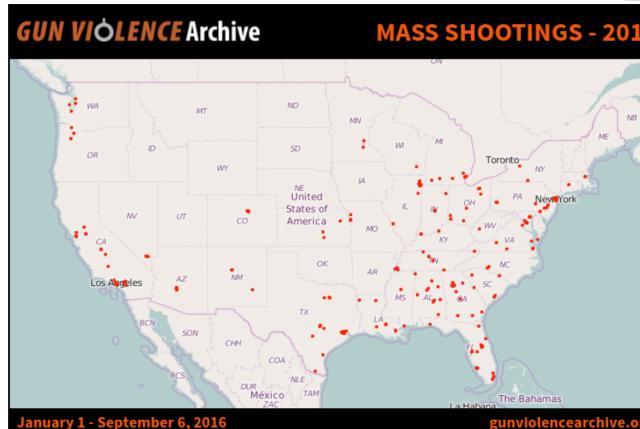
Total Number of Incidents	41,309
Number of Deaths ¹	10,599
Number of Injuries ¹	22,023
Number of Children (age 0-11) Killed or Injured ¹	477
Number of Teens (age 12-17) Killed or Injured ¹	2,252
Mass Shooting ²	285
Officer Involved Incident Officer Shot or Killed ²	233
Officer Involved Incident Subject-Suspect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

H

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

Share with Facebook

Share with Twitter



Scott Olson/Getty I

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

1K

SHARES

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 Inv
to 'Consu
Pipeline I



What Co
Know Ab
Security



Teen Wh
Sprayed



Dentist A
Harming
Making M



'I Did the
Orlando I





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

Want to answer questions like

Gun control - a Guardian investigation



America's gun problem is so
much bigger than mass
shootings

Too much emphasis on mass shootings has a cost

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

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

The shock and horror that follows mass shootings has led to a obsessive focus on the dangers of military-style rifles – even 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

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 that followed, there was extensive news coverage, widespread sorrow and a number of suicides. According to one study, the suicide rate in the United States [jumped by 12 percent](#) compared with the same month the previous year.

Mental illness is not a communicable disease, but there is a growing body of evidence that [suicide is still contagious](#). Publicized cases of suicide have been repeatedly and definitively linked to an 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

The screenshot shows a dark-themed podcast player interface. In the top left corner is the 'ON [THE MEDIA]' logo. To its right, the text 'Published in On The Media'. The main title 'Racial Bias in Crime Reporting' is centered in large white font. Below the title are three buttons: a blue 'Listen 4 min' button, a grey '+ Queue' button, and a grey '...' button. To the right of these are three social media sharing icons: Facebook (f), Twitter (t), and Email (e).

Jun 5, 2015

Summary Transcript

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

Research shows the media disproportionately depict African-Americans as criminals, and whites as victims. Brooke speaks with [Nazgol Ghandoosh](#), 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



U.S. » Trayvon Martin Shooting Fast Facts

Live TV •

U.S. Edition +



menu

Trayvon Martin Shooting Fast Facts

CNN Library

🕒 Updated 4:25 PM ET, Sun February 7, 2016



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.

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.nbciami.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



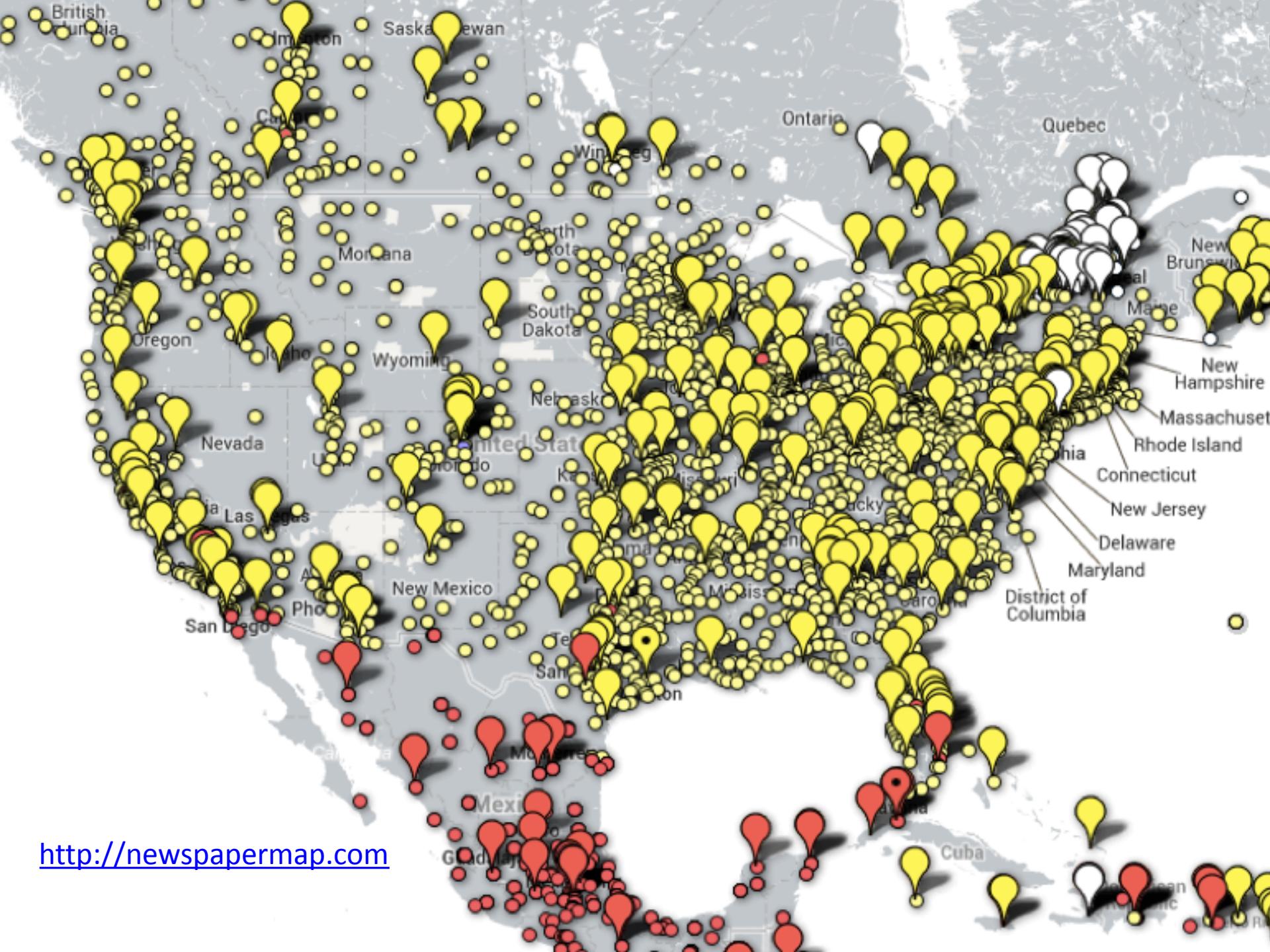
http://www.mlive.com/news/kalamazoo/index.ssf/2014/05/15-year-old_charged_with_murde.html



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



classifier.py		classifier.p
1	<code>#!/bin/python</code>	Statistical classification
2		2.891467 shooting
3	<code>import os</code>	2.560138 accidentally
4	<code>import sys</code>	2.342422 shot
5	<code>import string</code>	2.012679 gun
6	<code>import random</code>	1.925938 accidental
7	<code>import operator</code>	1.706036 subscriber
8	<code>from sklearn.tree import export_graphviz</code>	1.673353 guns
9	<code>from sklearn.tree import DecisionTreeClassifier</code>	1.626867 skip
10	<code>from sklearn.naive_bayes import MultinomialNB</code>	1.597359 discharged
11	<code>from sklearn.linear_model import LogisticRegres</code>	1.505322 rifle
12	<code>from sklearn.preprocessing import LabelEncoder</code>	1.449963 homicides
13	<code>from sklearn.feature_extraction import DictVect</code>	1.419679 hunting
14	<code>from sklearn.cross_validation import train_test</code>	1.418678 bullet
15	<code>from sklearn.externals.six import StringIO</code>	1.324652 ad
16		1.279014 log
17	<code>#read in raw data from file and return a list o</code>	1.225336 source
18	<code>def get_data(filename):</code>	1.214517 chest
19	<code> data = [line.strip().split('\t') for line i</code>	1.211558 reserved
20	<code>n random.shuffle(data)</code>	1.204958 gunshot
21	<code> return data</code>	1.147636 detroit
22		1.142863 innovative
23	<code>#this function builds the feature matrix for th</code>	1.136201 penny
24	<code>def get_dtreet_features(X) :</code>	1.117853 leg
25	<code> features = []</code>	1.108006 unlimited
26	<code> feature_list = []</code>	1.098322 township
27	<code> wordCounts = {}</code>	1.095423 update
28		1.094982 firearms
29	<code> for x in X :</code>	1.092020 marion
30	<code> f = {}</code>	
31	<code> for w in [word.strip(string.punctuation</code>	
32	<code> if not w in wordCounts :</code>	
33	<code> wordCounts[w] = 0</code>	
34	<code> wordCounts[w] += 1</code>	
35	<code> for w in wordCounts:</code>	
36	<code> if wordCounts[w] > 5000:</code>	



<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>



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



Woman stabbed in Ogden incident released from hospital

Yes

No

Unclear

[Show full text](#)

Wrangler News - Online Edition - Home

Yes

No

Unclear

Humans manually verify the predictions of the classifier.

Probe underway after man shot as marshals served warrant

Yes

No

Unclear

[Show full text](#)

POLICE KILL CAR THEFT SUSPECT

Yes

No

Unclear

[Show full text](#)



Woman stabbed in Ogden incident released from hospital

[Show full text](#) Yes No Unclear

Wrangler News

[Show full text](#)

LIVE Budget clash in Rajya

[Show full text](#)

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

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

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 because the man did not release the man's condition. Anyone with information should call 410-396-2221.

[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



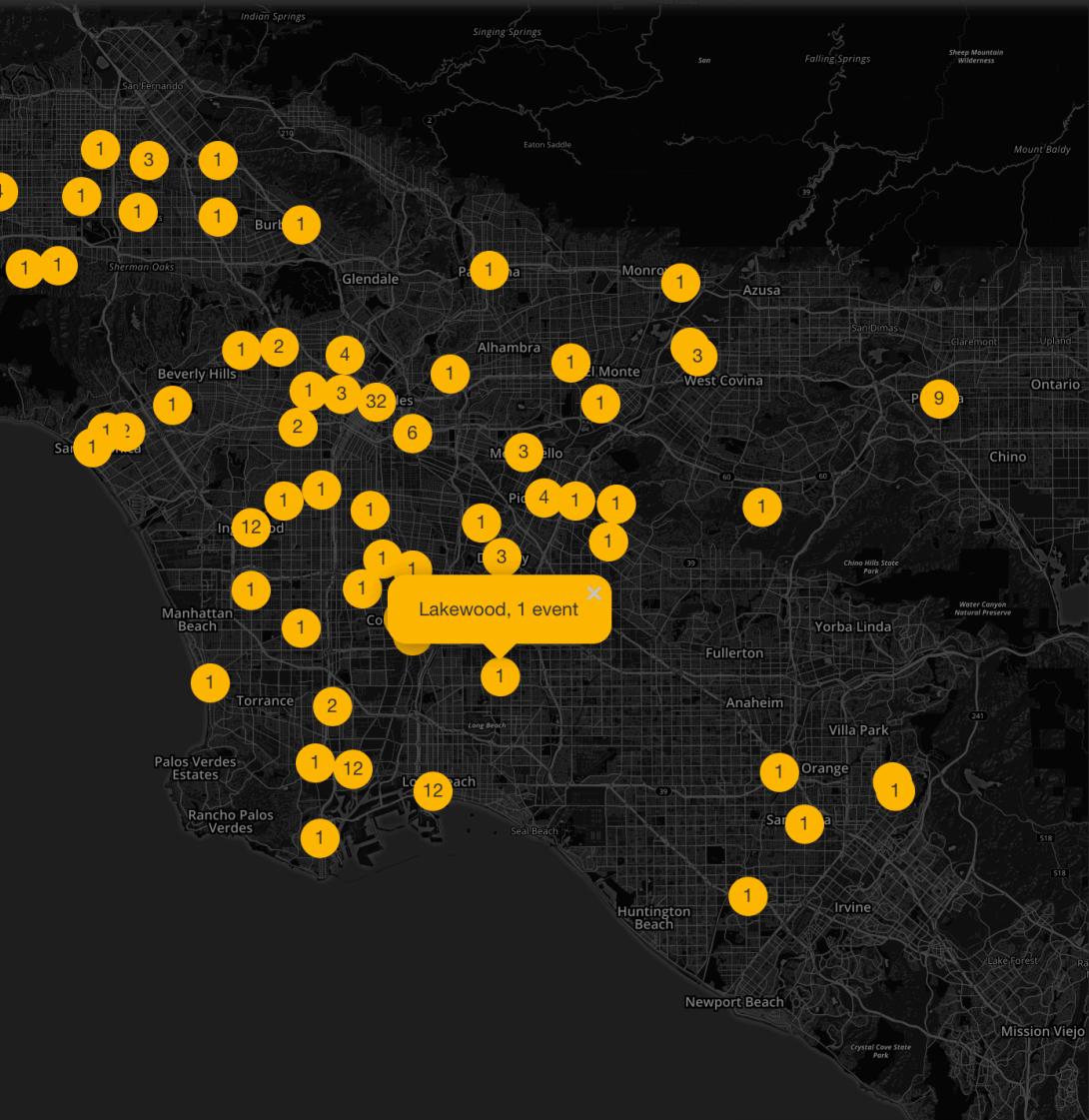
City

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

We plot the extracted information on a map

Work on task

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

Named Entity

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 where 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....

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

Event

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....

Detection

Person #1014

Name	Laquan McDonald
Gender	
Age	
Race	

Incident #1053

City	
Date	
Shooter	
Victim	
Victim Killed	

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

Automating The Pipeline

Semantic Role

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....

Labeling

Person #1014

Name	Laquan McDonald
Gender	
Age	
Race	

Incident #1053

City	
Date	
Shooter	
Victim	McDonald
Victim Killed	

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

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 where 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. Haghghi 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...

Person #1014

Name	Laquan McDonald
Gender	
Age	
Race	

Incident #1053

City	
Date	
Shooter	
Victim	Person #1014
Victim Killed	

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

Automating The Pipeline

Cross-Document

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...

Entity Coref.
Person #1014

Name	Laquan McDonald
Gender	
Age	
Race	

Incident #1053

City	
Date	
Shooter	
Victim	Person #1014
Victim Killed	

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

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 a 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...

Semantic Parsing

Person #1014

Name	Laquan McDonald
Gender	
Age	17
Race	Black

Incident #1053

City	
Date	
Shooter	
Victim	Person #1014
Victim Killed	

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

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 a 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...

Semantic Parsing

Person #1014

Name	Laquan McDonald
Gender	
Age	17
Race	Black

Incident #1053

City	
Date	
Shooter	
Victim	Person #1014
Victim Killed	TRUE

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

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



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 to the scene but did not release the man's condition. Anyone with information can call 410-396-2221.

[Clear this highlight](#) [Clock Time \(1p.m., 2:37a.m\)](#)[Additional Location Details](#)

Extracted data can provide training data for semantic parsing systems...

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)





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...



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](#) ↗

Victim #1a

NAME: Sean Bolton

AGE: unknown

RACE: unknown

GENDER: Male

VICTIM WAS: killed



No match
Victim #2a Officer Sean Bolton

[Read article](#) ↗

Victim #2a

NAME: Officer Sean Bolton

AGE: 33

RACE: unknown

GENDER: Male

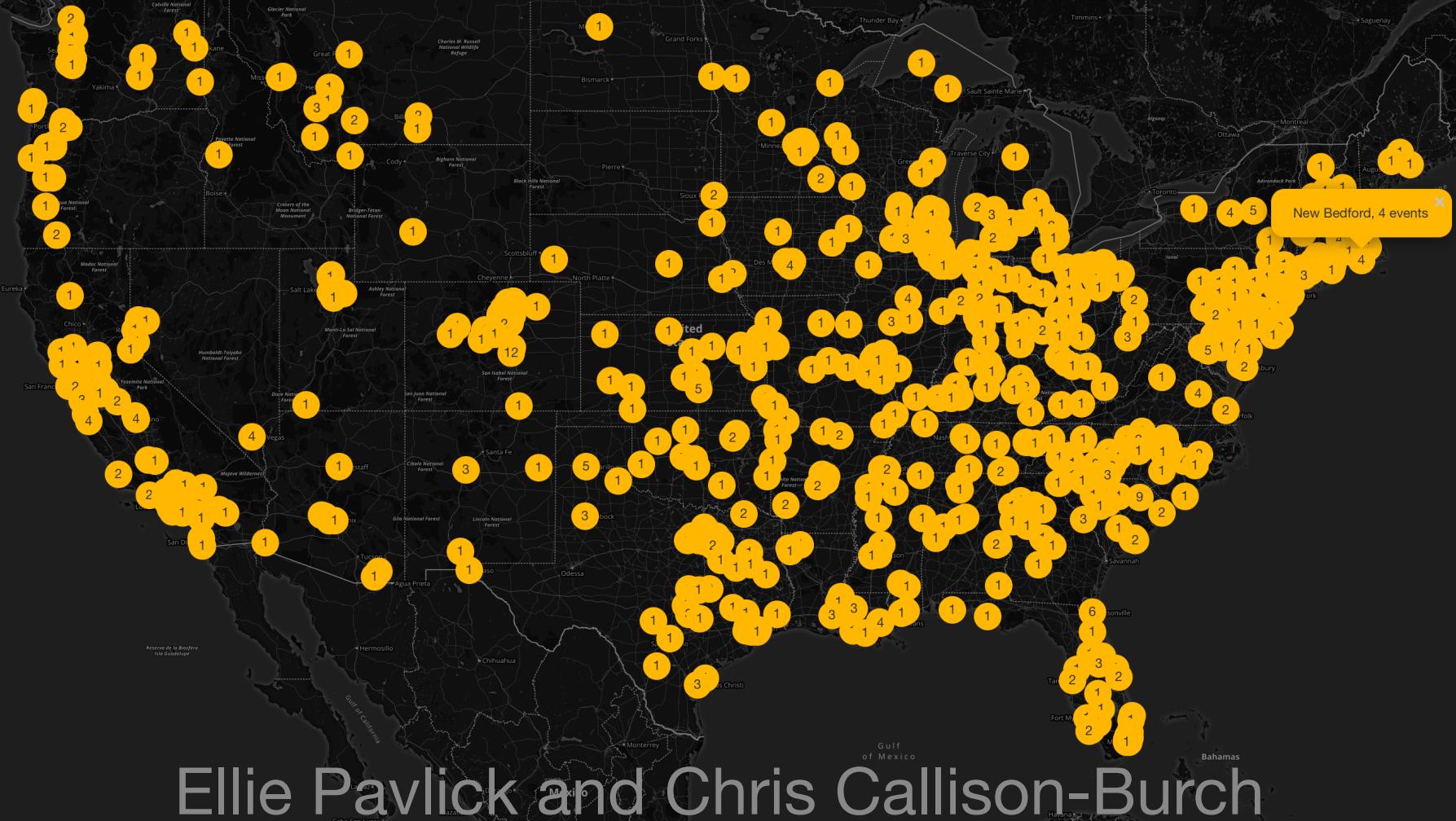
VICTIM WAS: killed

...and entity coreference.

The Gun Violence Database

- Use crowdsourcing to collect 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

<http://gun-violence.org>



Ellie Pavlick and Chris Callison-Burch
University of Pennsylvania