

NETS 213: CROWDSOURCING
AND HUMAN COMPUTATION

Introduction to Machine Learning

Professor Chris Callison-Burch





What is Machine Learning?

How can we build computer systems that automatically improve with experience and what are the fundamental laws that govern all learning processes?

-Tom Mitchell, CMU

The age of machine learning is now

- In the past 10 years there has been a revolution in machine learning / artificial intelligence because of the resurgence of neural network architectures
- We now have access to much better computing infrastructure, with GPUs optimized to the operations used by neural nets
- We have so much data that we can barely store it, and it provides great opportunities for analysis.

What can we do with machine learning?

- Find groups of related things via “clustering”. Used for recommendations by Amazon, Netflix, etc
- Are two items the same? Named entity disambiguation
- Classification: Is this YouTube comment offensive? What language is this web page written in? Whose face is shown in a picture?

Shop by
Department ▾

Search

All ▾

Go

Hello, Chris
Your Account ▾Your
Prime ▾ 0 Cart ▾Wish
List ▾

Your Amazon.com

Your Browsing History

Recommended For You

Rate Items You Like

Improve Your Recommendations

Your Profile

Learn More

Your Amazon.com > Recommended for You

(If you're not Chris Callison-Burch, click here.)

Just For Today[Browse Recommended](#)**Recommendations**[Amazon Instant Video](#)[Amazon MP3 Store](#)[Appliances](#)[Appstore for Android](#)[Arts, Crafts & Sewing](#)[Automotive](#)[Baby](#)[Beauty](#)[Books](#)[Books on Kindle](#)[Camera & Photo](#)[Cell Phones & Accessories](#)[Clothing & Accessories](#)[Computers](#)[Electronics](#)[Grocery & Gourmet Food](#)[Health & Personal Care](#)[Home & Kitchen](#)[Home Improvement](#)[Industrial & Scientific](#)[Jewelry](#)[Kitchen & Dining](#)[Magazine Subscriptions](#)[Magazines on Kindle](#)These recommendations are based on [items you own](#) and more.view: [All](#) | [New Releases](#) | [Coming Soon](#)[More results](#) 

1.

**[Carter's Keep Me Dry Waterproof Fitted Quilted Crib Pad, White](#)**

by Kids Line (December 11, 2009)

Average Customer Review:  (383)

In Stock

List Price: \$12.99**Price:** \$12.79[Add to Cart](#)[Add to Wish List](#)[17 used & new from \\$11.02](#) I own it Not interested Rate this itemRecommended because you added **Summer Infant Contoured Changing Pad Amazon Frustration F...** to your Shopping Cart and more ([Fix this](#))

2.

**[Nosefrida The Snotsucker Nasal Aspirator](#)**

by FridaBaby (April 1, 2010)

Average Customer Review:  (1,859)

In Stock

List Price: \$16.99**Price:** \$14.78[Add to Cart](#)[Add to Wish List](#)[43 new from \\$9.86](#) I own it Not interested Rate this itemRecommended because you added **Summer Infant Infant Character Change Pad Cover, Safari S...** to your Shopping Cart and more ([Fix this](#))

3.

**[Safety 1st Heavenly Dreams White Crib Mattress](#)**

by Dorel Home Products (December 11, 2010)

Average Customer Review:  (627)

In Stock

List Price: \$54.99**Price:** \$52.00[Add to Cart](#)[Add to Wish List](#)



Shop by
Department ▾

Chris...

Search

Your Amazon.com

Your Brow...

Your Amazon.com > Recommended for You
(If you're not Chris Callison-Burch, click here)

Just For Today

[Browse Recommended](#)

Recommendations

[Amazon Instant Video](#)

[Amazon MP3 Store](#)

[Appliances](#)

[Appstore for Android](#)

[Arts, Crafts & Sewing](#)

[Automotive](#)

[Baby](#)

[Beauty](#)

[Books](#)

[Books on Kindle](#)

[Camera & Photo](#)

[Cell Phones & Accessories](#)

[Clothing & Accessories](#)

[Computers](#)

[Electronics](#)

[Grocery & Gourmet Food](#)

[Health & Personal Care](#)

[Home & Kitchen](#)

[Home Improvement](#)

[Industrial & Scientific](#)

[Jewelry](#)

[Kitchen & Dining](#)

[Magazine Subscriptions](#)

[Magazines on Kindle](#)

Amazon.com: Why is this recommended for you?

amazon.com

Help | Close window

Recommended for You

[Nosefrida The Snotsucker Nasal Aspirator](#)
by FridaBaby (April 1, 2010)
In Stock
List Price: \$16.99
Price: \$14.78
43 new from \$9.86

Rate this item
 I own it
 Not interested

[Add to Cart](#) [Add to Wish List](#)

Because you purchased...

[GE 51386 Metal Shade With Flower Design Incandescent Night Light](#)
 This was a gift
 Don't use for recommendations

[Munchkin Arm & Hammer Diaper Pail Refill Bags, 30 Count](#)
by Munchkin
 This was a gift
 Don't use for recommendations

[My Brest Friend Original Pillow, Bluebells](#)
by Zenoff Products
 This was a gift
 Don't use for recommendations

Because your Wish List includes...

[WiFi Baby 2.0 \(2013 Model\) - iPhone, iPad, Android, Baby Monitor & Nanny Cam DVR, Video, Audio, Recording, Anywhere. Same Look, New Features \(WFB2013\)](#)
by WiFi Baby
 Don't use for recommendations

Cart ▾ [Wish List ▾](#)

More results ↗

Halloween Shop
Shop now

Cart ▾ [Wish List ▾](#)

More results ↗

Wish

Add to Wish List

your Shopping Cart and more (Fix this)

Add to Wish List

your Shopping Cart and more (Fix this)

Add to Wish List



amazon.com

Recommended for you, Chris

Buy It Again in Grocery
3 ITEMS

Children's Books
100 ITEMS

Buy it Again in Household Supplies
5 ITEMS

Buy It Again in Pets
3 ITEMS

Boys' Apparel
41 ITEMS

Arts & Crafts Supplies
88 ITEMS

Literature & Fiction
100 ITEMS

Computer & Technology Books
100 ITEMS

Swimming Equipment
65 ITEMS

New Releases

Movies & TV

Science Fiction & Fantasy Books



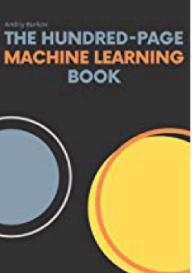
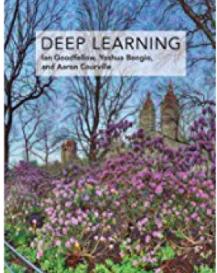
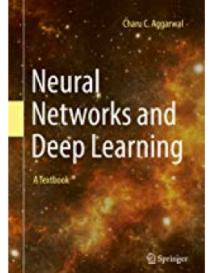
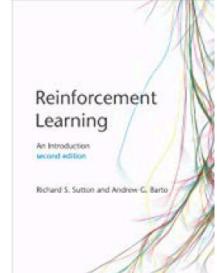
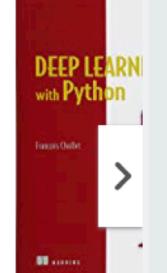
amazon.com

Recommended for you, Chris



Computer & Technology Books [View All & Manage](#)

Page 1 of 20 X

 <p>The Hundred-Page Machine Learning Book Andriy Burkov ★★★★★ 9 \$33.65</p>	 <p>Foundations of Machine Learning (Adaptive... Mehryar Mohri ★★★★★ 2 \$64.56</p>	 <p>Deep Learning (Adaptive Computation and Machine... Ian Goodfellow ★★★★★ 184 \$65.51</p>	 <p>Neural Networks and Deep Learning: A Textbook Charu C. Aggarwal ★★★★★ 14 \$53.12</p>	 <p>Reinforcement Learning: An Introduction (Adaptive... Richard S. Sutton ★★★★★ 12 \$63.00</p>	 <p>Deep Learning with Python Francois Chollet ★★★★★ 88 \$18.99 prime</p>
---	---	--	---	--	--



Swimming Equipment
65 ITEMS



New Releases



Movies & TV



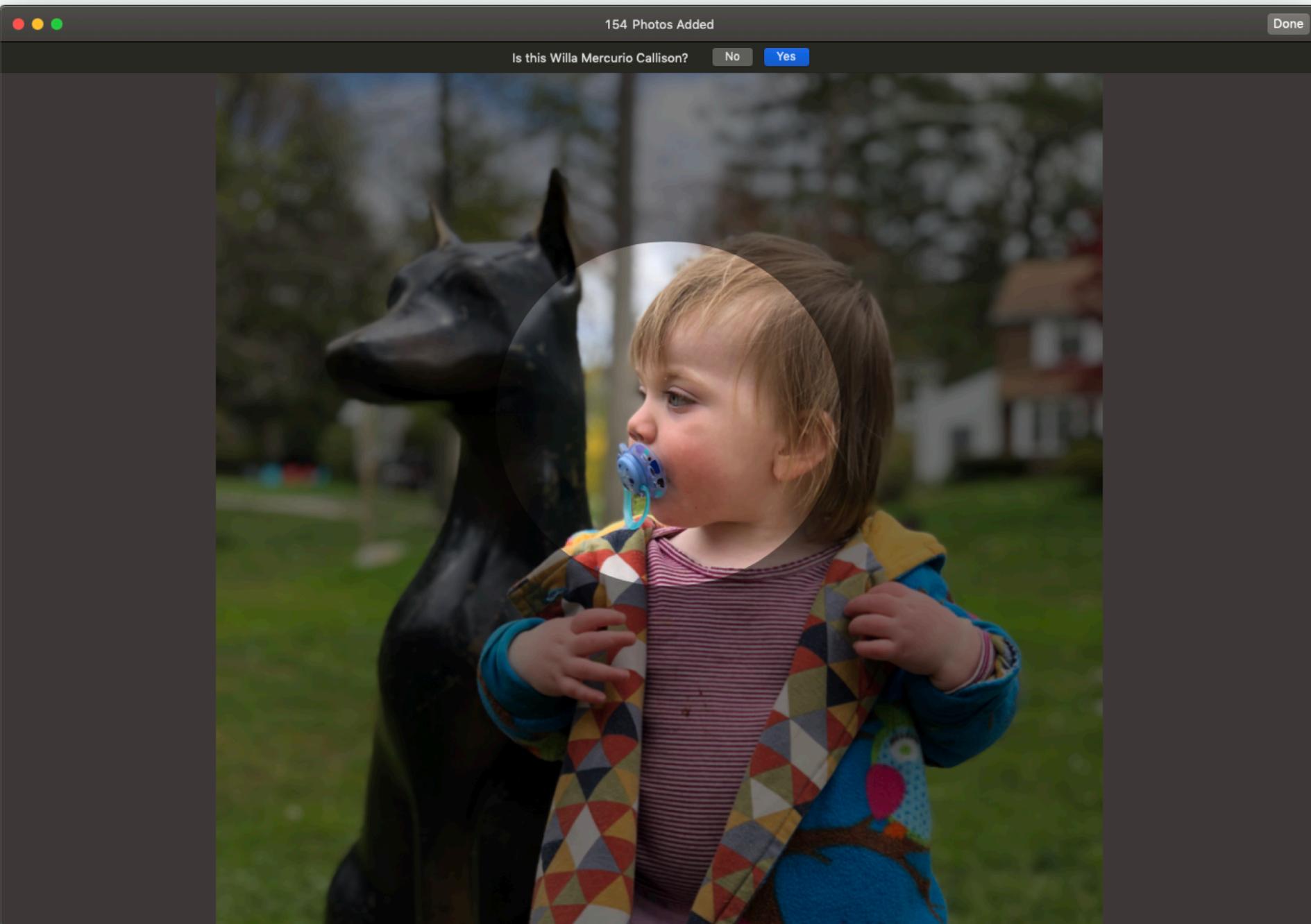
Science Fiction & Fantasy Books



People

18 People





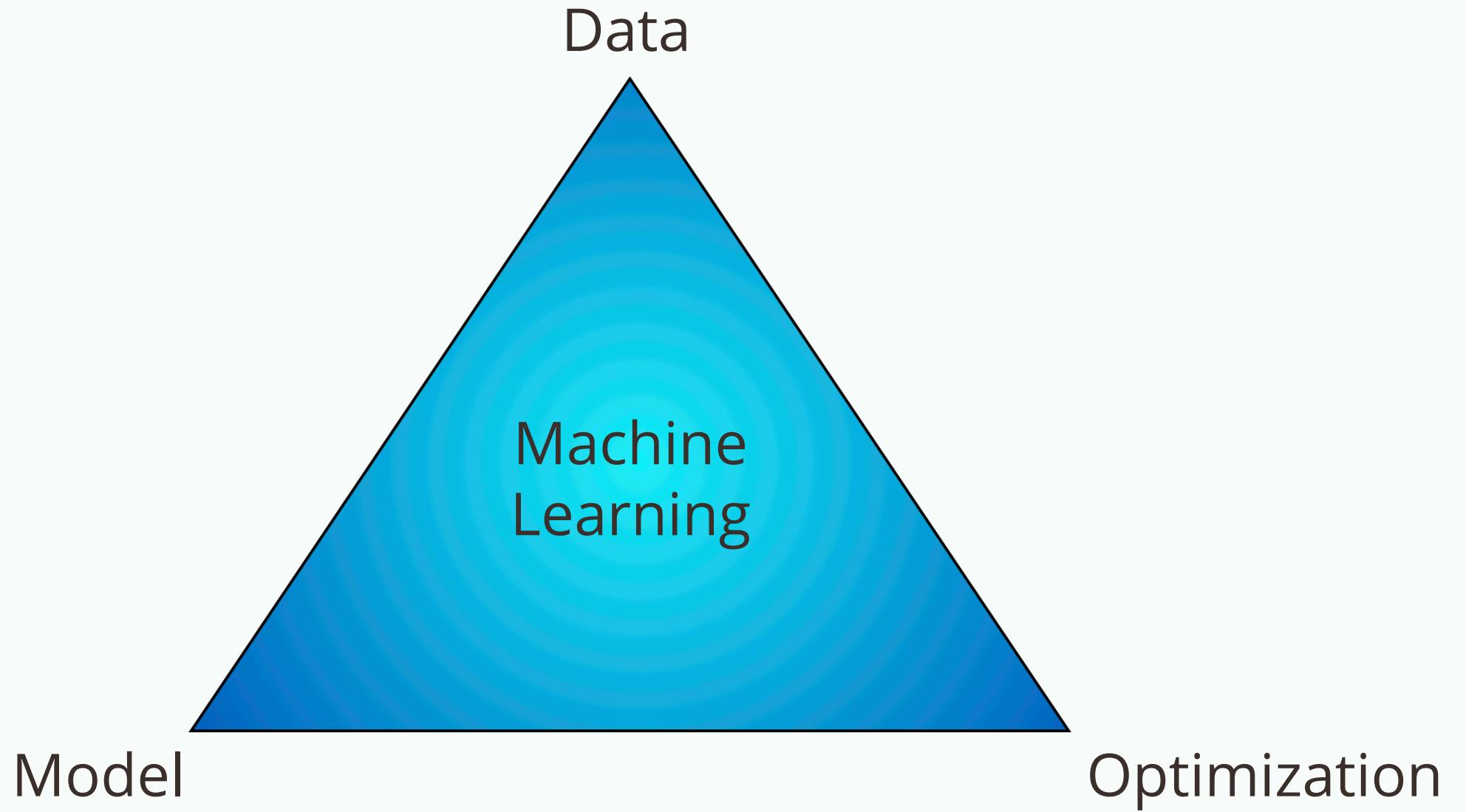
154 Photos Added

Done

Is this Willa Mercurio Callison?

No

Yes



Supervised v. Unsupervised Learning

- In supervised learning you are starting with a labeled training set of data
- In unsupervised learning you don't (yet) have labels for your data

Kinds of data?

- Text and speech
- Images and video
- Geographic information
- Time series information
- Transaction data from customers
- Climate data
- Census data

Where does data come from?

- Some datasets are available for free:
<http://crowdsourcing-class.org/resources.html>
- Some are owned by companies
- Sometimes you can assemble it yourself
- Crowdsourcing!

Your Amazon.com | Yelp Dataset | Datasets | Kaggle

yelp Dataset

Dataset Challenge Documentation

Yelp Dataset Challenge

Discover what insights lie hidden in our data.



What is the dataset challenge?

The challenge is a chance for students to conduct research or analysis on our data and share their discoveries with us. Whether you're trying to figure out how food trends start or identify the impact of different connections from the local graph, you'll have a chance to win cash prizes for your work! See some of the [past winners](#) and [hundreds of academic papers written](#) using the dataset.

The Challenge

We challenge students to use our data in innovative ways and break ground in research. Here are some examples of topics we find interesting, but remember these are only to get you thinking and we welcome novel approaches!

Photo Classification

Maybe you've heard of our ability to [identify hot dogs \(and other foods\)](#) in photos. Or how we can tell you if your photo will be [beautiful or not](#). Can you do better?



Natural Language Processing & Sentiment Analysis

What's in a review? Is it positive or negative? Our reviews contain a lot of metadata that can be mined and used to infer meaning, business attributes, and sentiment.

Graph Mining

We recently launched our [Local Graph](#) but can you take the graph further? How do user's relationships define their usage patterns? Where are the trend setters eating before it becomes popular?

Screenshot of the Data.gov website showing the Data Catalog page.

The page features a search bar at the top with the placeholder "Search Data.Gov". Below the search bar is a navigation menu with links to DATA, TOPICS, IMPACT, APPLICATIONS, DEVELOPERS, and CONTACT.

The main content area is titled "DATA CATALOG" and includes a search bar for datasets, a filter by location map, and a dropdown for ordering results by popularity.

A large heading indicates "304,468 datasets found". Below this, several datasets are listed:

- Pittsburgh Wards Map** (2527 recent views)
Allegeny County / City of Pittsburgh / Western PA Regional Data Center — Allows users to look up City of Pittsburgh Wards.
[HTML](#)
- National Student Loan Data System** (1977 recent views)
Department of Education — The National Student Loan Data System (NSLDS) is the national database of information about loans and grants awarded to students under Title IV of the Higher...
[XLSX](#) [XLS](#) [XLS](#) [XLS](#) [XLS](#) [XLS](#) 11 more in dataset
- ZIP Code Data** (1809 recent views)
Department of the Treasury — This study provides detailed tabulations of individual income tax return data at the state and ZIP code level.
[HTML](#)
- Demographic Statistics By Zip Code** (1723 recent views)
City of New York — Demographic statistics broken down by zip code
[CSV](#) [RDF](#) [JSON](#) [XML](#)
- Crimes - 2001 to present** (1589 recent views)
City of Chicago — This dataset reflects reported incidents of crime (with the exception of murders where data exists for each victim) that occurred in the City of Chicago from 2001 to...
[CSV](#) [RDF](#) [JSON](#) [XML](#)
- College Scorecard** (1250 recent views)
Department of Education — College Scorecards make it easier for students to search for a college that is a good fit for them. They can use the College Scorecard to find out more about a...
[HTML](#) [Zipped CSV](#) [CSV](#) [CSV](#) [CSV](#) [CSV](#) 1 more in dataset

On the right side of the page, there are three teal-colored callout boxes with white text:

- County
- Federal
- City

At the bottom of the page, there is a link to "U.S. Crime Data Repository (CDR)".

kaggle Search Competitions Datasets Kernels Discussion Learn ... Sign In

Datasets

Documentation New Dataset

Public Sort by Most Votes

14,474 Datasets Sizes File types Licenses Tags Search datasets

Rank	Dataset Name	Description	Tags	File Types	Size	Comments	Created
2508	Credit Card Fraud Detection	Anonymized credit card transactions labeled as fraudulent or genuine	crime finance	CSV ODbL	66 MB	37 1m	Machine Learning Group - ULB updated a year ago (Version 3)
1590	European Soccer Database	25k+ matches, players & teams attributes for European Professional Football	association... europe	SQLite ODbL	34.4 MB	88 558k	Hugo Mathien updated 2 years ago (Version 10)
1373	TMDB 5000 Movie Dataset	Metadata on ~5,000 movies from TMDB	film	CSV Other	9.3 MB	51 564k	The Movie Database (TMDB) updated a year ago (Version 2)
1147	Global Terrorism Database	More than 180,000 terrorist attacks worldwide, 1970-2017	crime terrorism internation...	CSV Other	27.9 MB	667 12 273k	START Consortium updated 5 months ago (Version 3)
1117	Wine Reviews	130k wine reviews with variety, location, winery, price, and description	critical the... food and dr...	CSV CC4	50.9 MB	19 229k	zackthoutt updated a year ago (Version 4)
1056	Bitcoin Historical Data	Bitcoin data at 1-min intervals from select exchanges, Jan 2012 to November 2018	history finance	CSV CC4	110.8 MB	108 25 264k	Zielak updated 3 months ago (Version 15)
934	Google Play Store Apps	Web scraped data of 10k Play Store apps for analysing the Android market.	video games computer s... internet mobile web	CSV Other	1.9 MB	192 22 216k	Lavanya Gupta updated 9 days ago (Version 6)
815	Trending YouTube Video Statistics	Daily statistics for trending YouTube videos	languages popular cul... statistics	CSV CCO	1991 MB	94 21 172k	Mitchell J updated 3 months ago (Version 114)
801	Pokemon with stats	721 Pokemon with stats and types	popular cul... games and ... video games	CSV CCO	14.7 KB	1k 15 191k	Alberto Barradas updated 2 years ago (Version 2)
791	Data Science for Good: Kiva Crowdfunding	Use Kernels to assess welfare of Kiva borrowers for \$30k in prizes	geography finance lending	CSV CCO	41.9 MB	240 47 143k	Kiva updated a year ago (Version 5)
786	2018 Kaggle ML & DS Survey Challenge	Explore the 2018 Kaggle ML & Data Science Survey for \$28,000 in cash prizes	survey anal...	CSV CC4	3.9 MB	255 17 302k	Kaggle updated 3 months ago

kaggle Search Competitions Datasets Kernels Discussion Learn ... Sign In

Dataset • Released Under CC BY-NC-SA 4.0

Wine Reviews

130k wine reviews with variety, location, winery, price, and description

 zackthoutt · updated a year ago (Version 4)

Data Overview Kernels (1,687) Discussion (19) Activity Download (51 MB) New Kernel

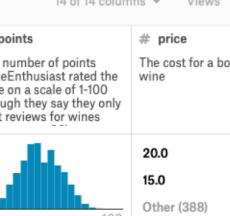
Data (51 MB)

Data Sources		About this file	Columns
<ul style="list-style-type: none"> winemag-data-130k... 130k x 14 winemag-data_first15... 151k x 11 winemag-data-130k-v2.json 	<p>Here is a CSV version of the data I scraped. This dataset has three new fields --Title (which you can parse the vintage from), Taster Name, and Taster Twitter Handle. This should also fix the duplicate entries problem in the first version of the dataset and add ~25k unique reviews to play with.</p>	<p># A country The country that the wine is from A description A designation The vineyard within the winery where the grapes that made the wine are from # points The number of points WineEnthusiast rated the wine on a scale of 1-100 (though they say they</p>	

winemag-data-130k-v2.csv (50.46 MB)

14 of 14 columns Views

A country	A description	A designation	# points	# price	A province
US	42%	Reserve	2%	20.0	California
France	17%	Estate	1%	15.0	Washington
Other (41)	41%	Other (37977)	97%	Other (388)	Other (423)
Italy	Aromas include tropical fruit, broom, brimstone and dried herb. The palate isn't overly expressive, offering unripened apple, citrus and dried sage alongside brisk acidity.	Vulka Bianco	87		Sicily & Sardegna
Portugal	This is ripe and fruity, a wine that is smooth while still structured. Firm tannins are filled out with juicy red berry fruits and freshened with acidity. It's already drinkable, although it will cellar...	Avidagos	87	15.0	Douro
US	Tart and snappy, the		87	14.0	Oregon



Classification

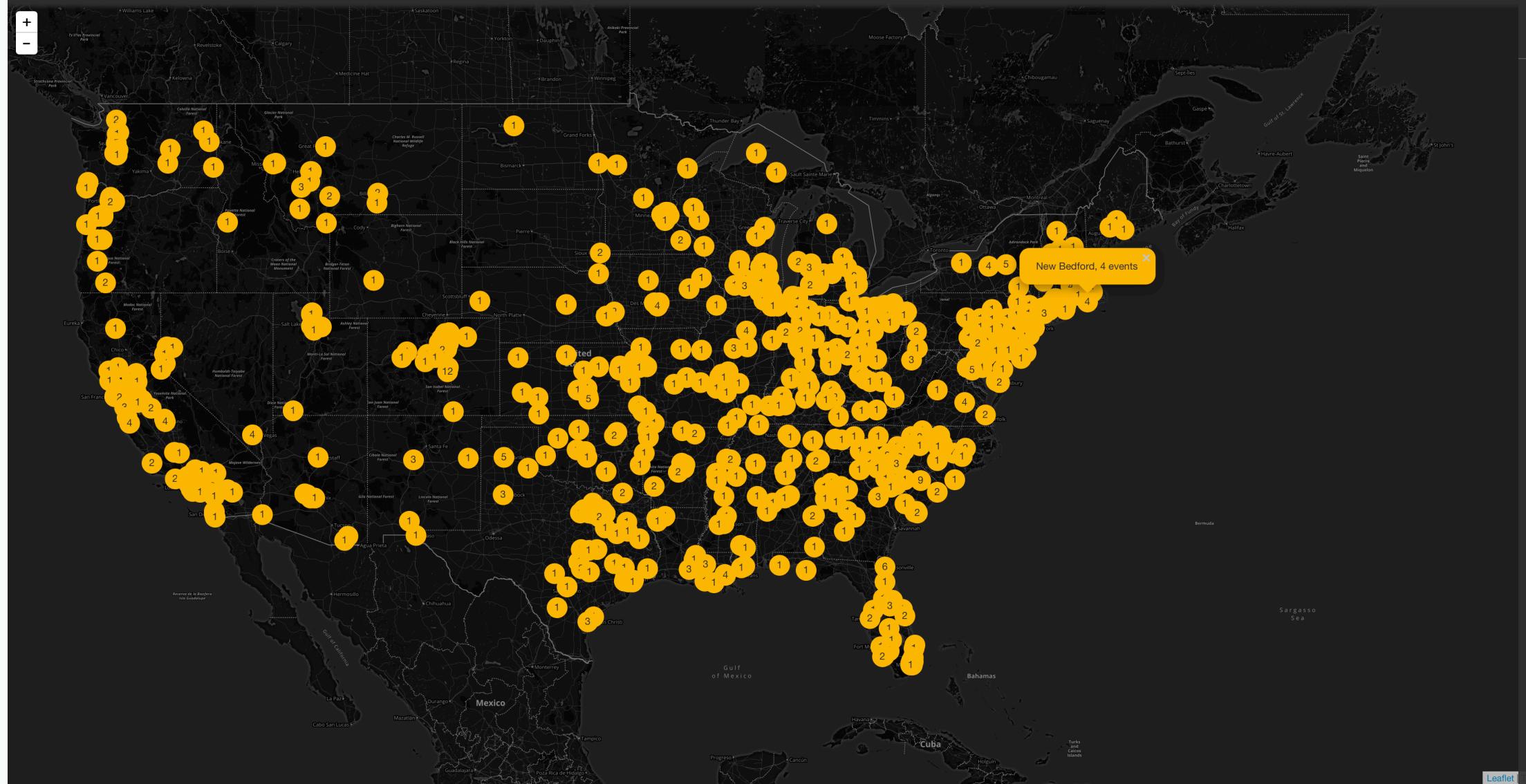
- Classification is the assignment of a label to unlabeled input based on previously seen data
- Learn $f(x)$...
- that outputs a label ...
- along with a probability that that label is true



Gun Violence DATABASE

BROWSE MAP ABOUT PROJECT WORK ON TASKS DOWNLOAD DATA

Chris Callison-Burch ▾



© Gun Violence Database 2016

[Leaflet](#) [Terms & Feedback](#)

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.

- 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.

Fatal Encounters

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

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

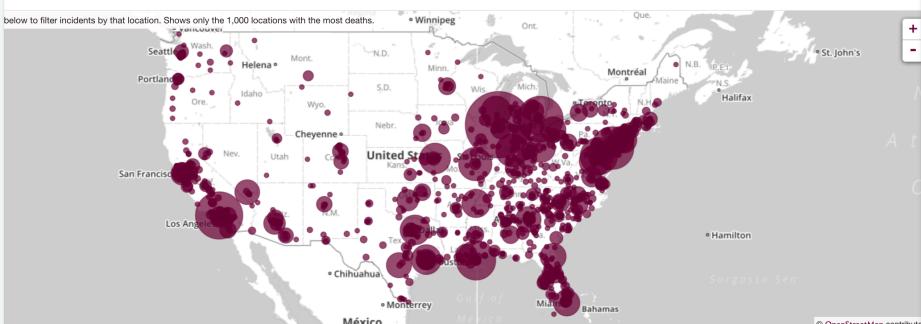
How Many People Have Been Killed by Guns Since Newtown?

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

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.



PEOPLE KILLED IN 2016

7 9 3 4

RACE & ETHNICITY PER MILLION

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

STATE NAME PER CAPITA TOTAL

STATE	NAME	PER CAPITA	TOTAL
NM	AK DC SD AZ CO OK WV HI AL		NEW MEXICO

Total killed: 21
Population: 2,085,109

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

LOGIN CONTACT US Search Database

GUN VIOLENCE Archive

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



GUN VIOLENCE Archive 2016

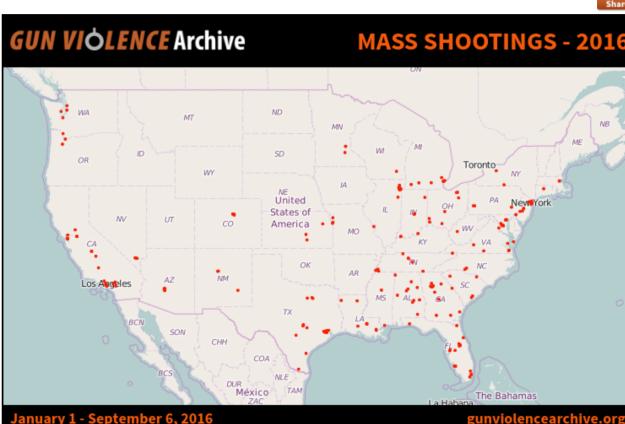
Total Number of Incidents	41,309
Number of Deaths ¹	10,599
Number of Injuries ¹	22,023
Number of Children (age 0-11) Killed or Injured ¹	477
Number of Teens (age 12-17) Killed or Injured ¹	2,252
Mass Shooting ²	285
Officer Involved Incident Officer Shot or Killed ²	233
Officer Involved Incident Subject-Suspect Shot or Killed ²	1,359
Home Invasion ²	1,724
Defensive Use ²	1,288
Accidental Shooting ²	1,538

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



impossibly ambitious project: cataloguing gun violence in America over the last three years. After all, we could have imagined.

100 incidents of police-involved shootings have been recorded and have at least one incident logged on our database for four years. We owe everyone who's been involved in this when this is all over—and for anyone

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

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.

“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.



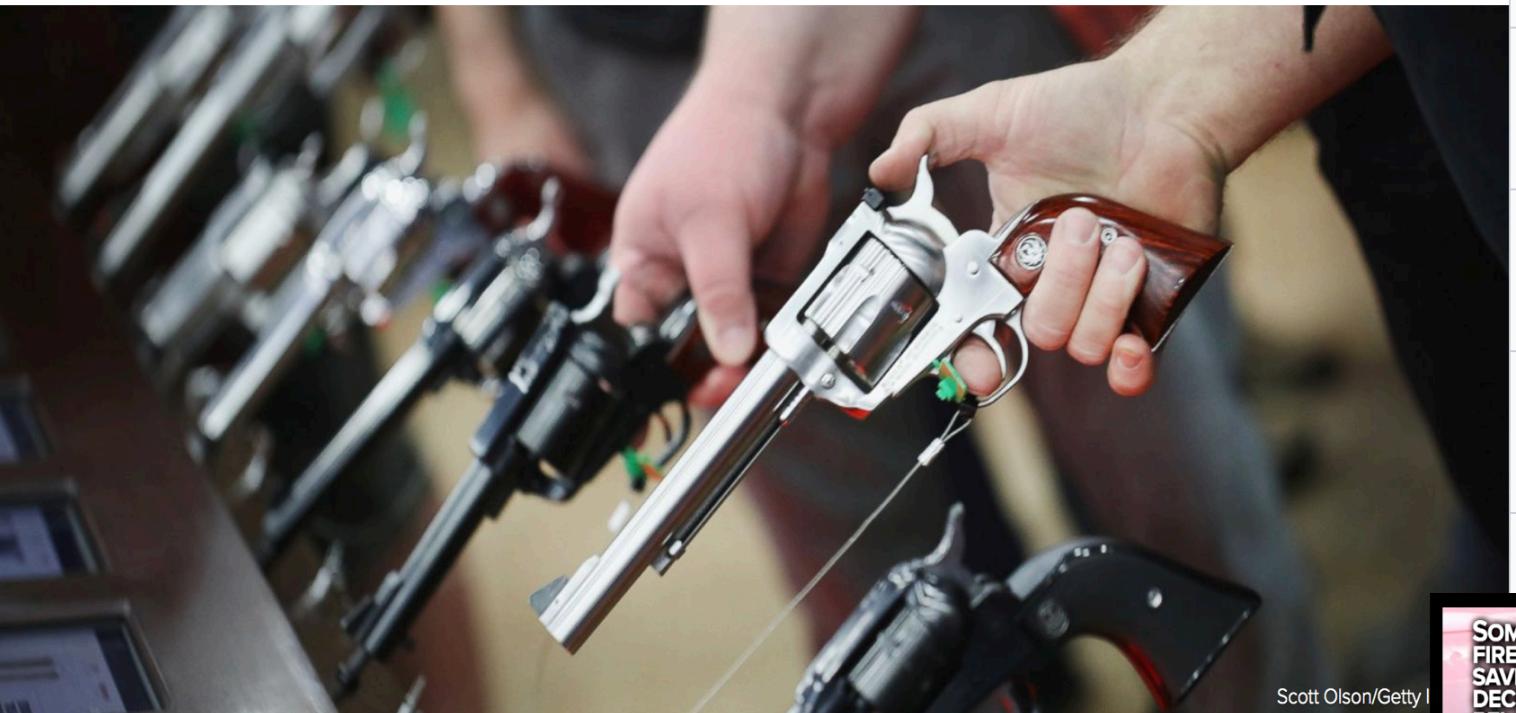
Why the CDC Hasn't Launched a Comprehensive Gun Study in 15 Years

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

[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 CDC
Launched a
Study in 15



Feds Invite
to 'Consult'
Pipeline Pro



What Consu
Know Abou
Security Bre



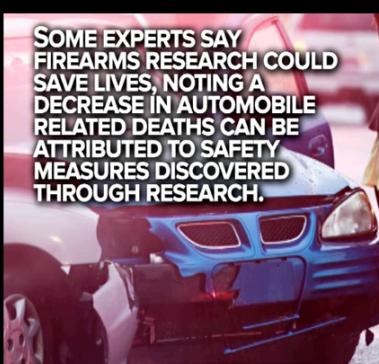
Teen Who V
Sprayed Sh



Dentist Acc
Harming Ch
Making Mill



'I Did the Sh
Orlando Nig



SOME EXPERTS SAY
FIREARMS RESEARCH COULD
SAVE LIVES, NOTING A
DECREASE IN AUTOMOBILE
RELATED DEATHS CAN BE
ATTRIBUTED TO SAFETY
MEASURES DISCOVERED
THROUGH RESEARCH.



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

Gun control - a Guardian investigation



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

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

Too much emphasis on mass shootings has a cost

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 terrible way to go about saving lives.

The shock and horror that follows mass shootings has led to an obsessive focus on the dangers of military-style rifles – even though rifles of any kind were used in less than 3% of

Want to answer questions like

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

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

Mental illness is not a communicable disease, but there's a strong body of evidence that [suicide is still contagious](#). Publicity surrounding a suicide has been repeatedly and definitively linked to a subsequent increase in suicide, especially among young people. Analysis suggests that at least 5 percent of youth suicides are influenced by contagion.

Want to answer questions like

The screenshot shows a dark-themed web page for 'ON [THE MEDIA]'. At the top left is the 'ON [THE MEDIA]' logo. To its right, the text 'Published in On The Media'. Below this is the title 'Racial Bias in Crime Reporting'. Underneath the title are three buttons: a blue rounded rectangle with a play icon and the text 'Listen 4 min', a white button with a plus sign and the text '+ Queue', and a small grey circle with three dots. To the right of these buttons are three social media icons: a Facebook 'f', a Twitter bird, and an envelope. The main content area below the title has a dotted horizontal line separating it from the sidebar.

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 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



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.

Kalamazoo County Prosecutor Jeff Getting revealed his anguish at a press conference Thursday afternoon. "To talk about the death of a 13-year-old who was shot on one of our streets, allegedly by a 15-year-old, and to think about those as eighth graders and ninth graders...It has an effect on me. I think it has an effect on everyone: it should."



Penn Engineering

The End of the Gun Report

▶ Listen 10 min

+ Queue

...

f

Twitter icon

Email icon



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 Henkis, 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	http://gulfnews.com/news/mena/palestine/palestinian-who-attacked-soldier-shot-dead-1.1631076	Gulf News	child shot

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](#)

Woman stabbed in Ogden incident released from hospital

Yes

No

Unclear

[Show full text](#)

Humans manually verify the predictions of the classifier.

Wrangler News - Online Edition - Home

Yes

No

Unclear

[Show full text](#)

LIVE Budget session: Smriti Irani, Mayawati clash in Rajya Sabha over Vemula suicide

Yes

No

Unclear

[Show full text](#)

Probe underway after man shot as marshals served warrant

Yes

No

Unclear

[Show full text](#)

Woman stabbed in Ogden incident released from hospital

Yes

No

Unclear

Show full text

Wrangler News

Show full text

LIVE Budget clash in Rajya

Show full text

Probe underway after man shot as marshals served warrant

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 to the scene about 1:30 a.m. and 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

**...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

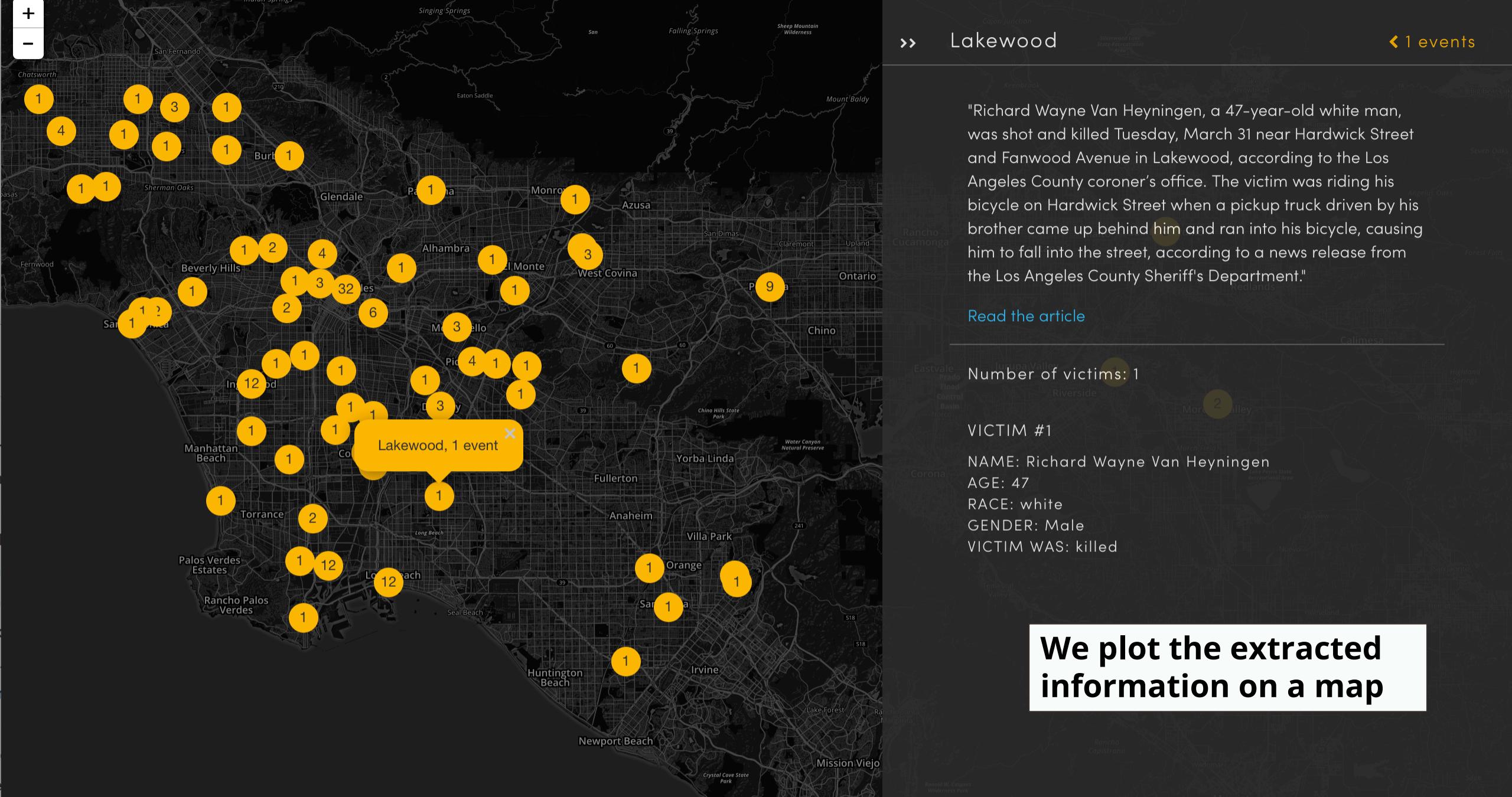


City

Baltimore



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



Gun violence classifier

- We'd like to automate the creation of the gun violence database.
- Let's start by creating a classifier that will tell us whether a newspaper article describes an incident of gun violence or not.

Labeled training data



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.

Kalamazoo County Prosecutor Jeff Getting revealed his anguish at a **press conference** Thursday afternoon. "To talk about the death of a 13-year-old who was shot on one of our streets, allegedly by a 15-year-old, and to think about those as eighth graders and ninth graders...It has an effect on me. I think it has an effect on everyone; it should."

Collecting data from the web

```
import urllib
import urllib2
from cookielib import CookieJar

def compile_gunreport_urls:
    for year in ["2014", "2013"]:
        for month in range(1, 13):
            for day in range(1, 32):
                url = "http://nocera.blogs.nytimes.com/%s/%s/%s/" %
                      (year, month, day)
            try:
                cj = CookieJar()
                opener = urllib2.build_opener
                    (urllib2.HTTPCookieProcessor(cj))
                site = opener.open(url).read()
```

Collecting data from the web

```
import lxml.etree
import lxml.html
import re

def extract_external_links():
    for url in gunreport_urls:
# The NYTimes redirects you if you don't have cookies set.
        cj = CookieJar()
        opener = urllib2.build_opener(urllib2.HTTPCookieProcessor(cj))
        site = opener.open(url).read()

        doc = lxml.etree.HTML(site)

        result = doc.xpath("//div[@class='entry-content']//p")
        link = re.compile('href="(.*?)"')
        for item in result:
            source = lxml.html.tostring(item)
            if link.search(source):
                print link.search(source).group(1)
```

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	
http://fox6now.com/2014/05/29/fund-created-for-sierra-guyton-victim-of-shooting-near-playground/	

Extracting web page text

≡ Menu  Set Weather ▾ **m** LIVE Michigan ▾ [Subscribe](#) ▾  Sign In  Search

Kalamazoo teenager, 13, shot and injured late Saturday while leaving party, police say



By [Alex Mitchell](#) | amitche5@mlive.com

on April 06, 2014 at 9:38 AM, updated April 06, 2014 at 1:01 PM

KALAMAZOO, MI — A 13-year-old Kalamazoo juvenile was shot in the back and injured late Saturday while leaving a party on the south side of Kalamazoo, police say.

Officers responded around 11:40 p.m. to the area of Lake Street and Maywood Avenue to a report of a subject that had been shot and discovered a teenage male suffering from a gunshot wound in his back, according to a press release issued by the [Kalamazoo Department of Public Safety](#).

The victim, whose name has not been released, told police he had just left a party in the 600 block of Carr Street. The teen said he was walking near Lake and Maywood when he heard a gunshot and realized he had been struck in the back, police said.



Gazette File

High School Football



Follow the latest prep football news from around Michigan

- Statewide schedules
- Recruiting news
- State rankings
- Full high school sports coverage



Newspaper3k: Article scraping & curation

[pypi package](#) [0.2.8](#) [build](#) [passing](#) [coverage](#) [unknown](#)

Inspired by [requests](#) for its simplicity and powered by [lxml](#) for its speed:

“Newspaper is an amazing python library for extracting & curating articles.” –
[tweeted by Kenneth Reitz](#), Author of [requests](#)

“Newspaper delivers Instapaper style article extraction.” – [The Changelog](#)

Newspaper is a Python3 library! [View on Github here](#), or, view our [deprecated and buggy Python2 branch](#)

Useful Links

[Newspaper @ GitHub](#)

[Newspaper @ PyPI](#)

[Issue Tracker](#)

This Page

[Show Source](#)

Quick search

A Glance:

```
>>> from newspaper import Article  
  
>>> url = 'http://fox13now.com/2013/12/30/new-year-new-laws-obamacare-pot-gur  
>>> article = Article(url)
```

```
>>> article.download()  
  
>>> article.html  
'<!DOCTYPE HTML><html itemscope itemtype="http://...'
```

```
>>> article.parse()  
  
>>> article.authors  
['Leigh Ann Caldwell', 'John Honway']
```

```
>>> from newspaper import Article  
>>> url = 'http://www.mlive.com/news/kalamazoo/index.ssf/2014/04/kalamazoo_teenage  
r_13_shot_and.html'  
>>> article.download()  
>>> article.parse()  
>>> print(article.text.strip())  
- A 13-year-old Kalamazoo juvenile was shot in the back and injured late Saturday  
while leaving a party on the south side of Kalamazoo, police say.
```

Officers responded around 11:40 p.m. to the area of Lake Street and Maywood Avenue to a report of a subject that had been shot and discovered a teenage male suffering from a gunshot wound in his back, according to a press release issued by the Kalamazoo Department of Public Safety.

The victim, whose name has not been released, told police he had just left a party in the 600 block of Carr Street. The teen said he was walking near Lake and Maywood when he heard a gunshot and realized he had been struck in the back, police said.

While waiting for an ambulance to arrive, the victim was transported to Bronson Methodist Hospital for treatment of non-life threatening injuries by an acquaintance, officers said. He is currently in stable condition.

Representing data with Features

- In machine learning, we represent the training data as a vector of labels (\mathbf{y}) and a matrix of training items (\mathbf{X})
- Each training item is itself represented as a vector
- The vector specifies what **features** that item has

Representing data with Features



X?

Pretty awful - very soft and commercial. Confectioned.

An absolute star that could even benefit from another year or two. Tremendous weight, and concentrated minerality but all in balance. Fantastic. Top

Very classy, pure, blackberry and apple fruit. Demanding but ripe tannins, very succulent. Really good Dolcetto.

Good Syrah character, fruit-driven but not to the point of undrinkability. Pleasant. Scruples

Thin and completely uninspiring.

Fragrant, dry and long. More mineral and complex than the other Ogier wines. Really lovely and should be drunk on its own away from the Gentaz wines that tend to upstage it.

Representing data with Features

raw input

Pretty awful - very **soft** and commercial.
Confected.

An absolute **star** that could even **benefit** from another year or two. **Tremendous** weight, and concentrated minerality but all in balance. **Fantastic. Top**

Very **classy, pure**, blackberry and apple fruit. **Demanding** but ripe tannins, very succulent. Really **good** Dolcetto.

Good Syrah character, fruit-driven but not to the point of undrinkability. **Pleasant**. Scrapes

Thin and completely **uninspiring**.

Fragrant, dry and long. More mineral and complex than the other Ogier wines. Really lovely and should be drunk on its own away from the Gentaz wines that tend to upstage it.

subjectivity lexicon

Word	Polarity	Strength
abandoned	negative	weak
abandonment	negative	weak
abandon	negative	weak
abase	negative	strong
abasement	negative	strong
abash	negative	strong
abate	negative	weak
abdicate	negative	weak
aberration	negative	strong
aberration	negative	strong
...		...
zest	positive	strong

Representing data with Features

raw input

Pretty awful - very **soft** and commercial.
Confected.

An absolute **star** that could even **benefit** from another year or two. **Tremendous** weight, and concentrated minerality but all in balance.

Fantastic. Top

Very **classy, pure**, blackberry and apple fruit.
Demanding but ripe tannins, very succulent.
Really **good** Dolcetto.

Good Syrah character, fruit-driven but not to the point of undrinkability. **Pleasant**. Scrapes

Thin and completely **uninspiring**.

Fragrant, dry and long. More mineral and complex than the other Ogier wines. Really lovely and should be drunk on its own away from the Gentaz wines that tend to upstage it.

feature matrix X

Strong Neg	Neg	Pos	Strong Pos
1	1		1
		2	3
	1	3	
		2	
	1		

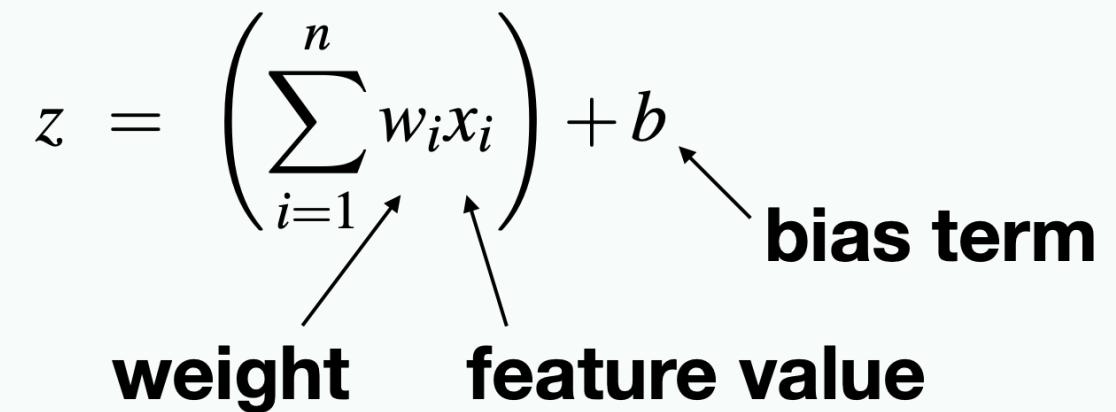
Classification via logistic regression

- To make a decision on a test instance, we multiply each x_i by a weight w_i that is automatically learned.
- Then we sum these together and add a bias term b

$$z = w \cdot x + b$$

$$z = \left(\sum_{i=1}^n w_i x_i \right) + b$$

weight **feature value** **bias term**

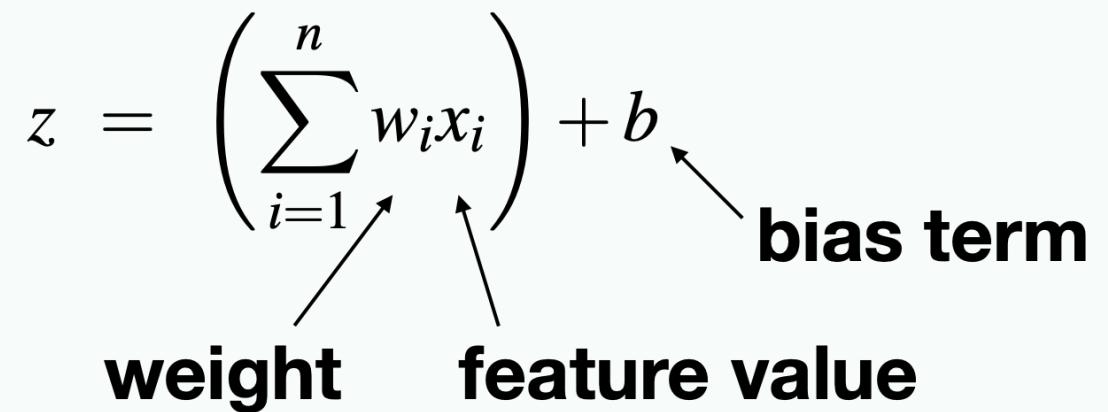


What's wrong with z?

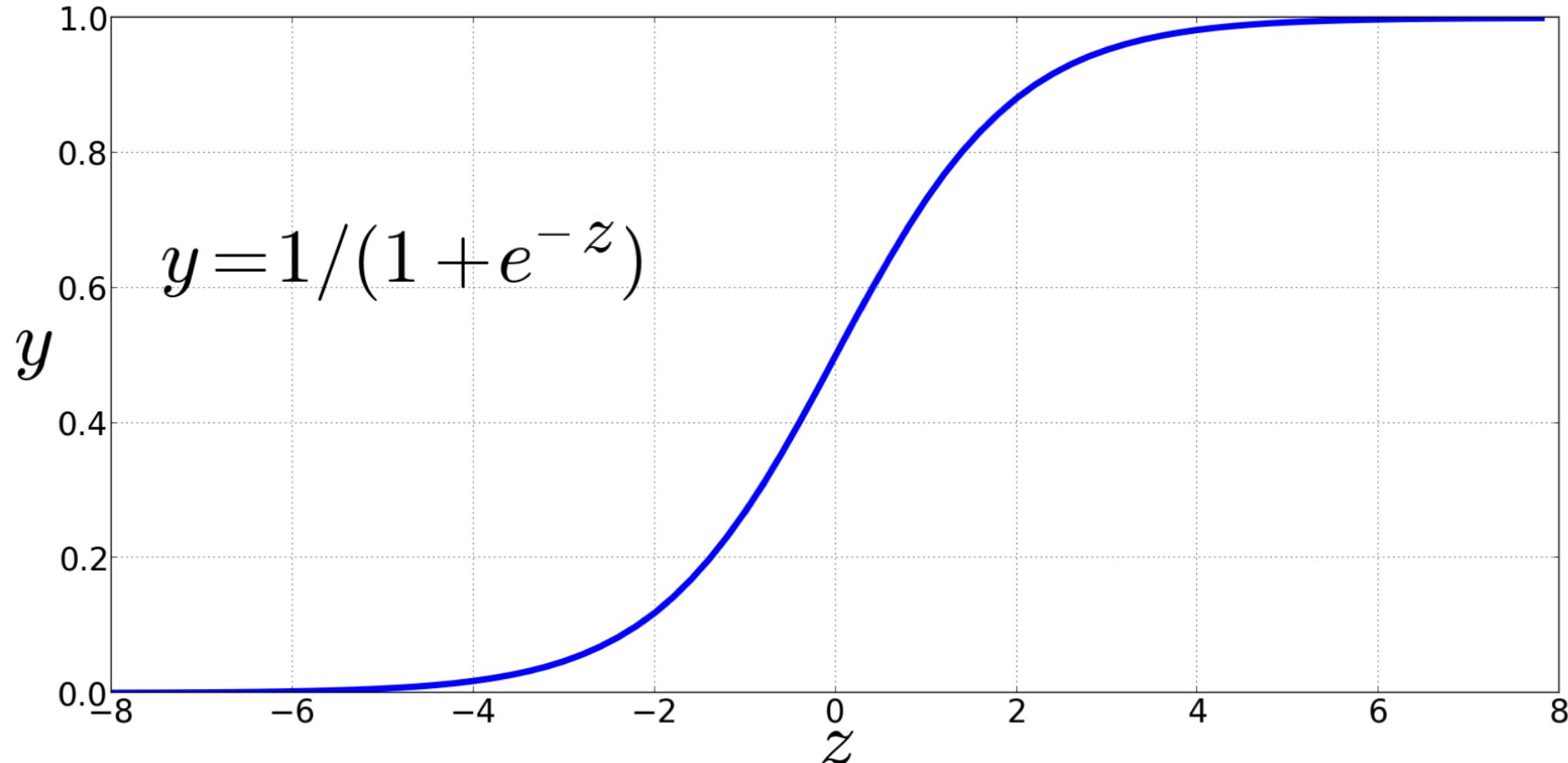
- It isn't a valid probability!
- Weights can be real valued, and might even be negative.
- z ranges from $-\infty$ to $+\infty$

$$z = \left(\sum_{i=1}^n w_i x_i \right) + b$$

weight **feature value** **bias term**



Logistic function aka the sigmoid



Sigmoid to probability

$$\begin{aligned} P(y = 1) &= \sigma(w \cdot x + b) \\ &= \frac{1}{1 + e^{-(w \cdot x + b)}} \end{aligned}$$

Decision boundary

$$\hat{y} = \begin{cases} 1 & \text{if } P(y=1|x) > 0.5 \\ 0 & \text{otherwise} \end{cases}$$

Representing data with Features

raw input

Pretty awful - very **soft** and commercial.
Confected.

An absolute **star** that could even **benefit** from another year or two. **Tremendous** weight, and concentrated minerality but all in balance.

Fantastic. Top

Very **classy, pure**, blackberry and apple fruit.
Demanding but ripe tannins, very succulent.
Really **good** Dolcetto.

Good Syrah character, fruit-driven but not to the point of undrinkability. **Pleasant**. Scrapes

Thin and completely **uninspiring**.

Fragrant, dry and long. More mineral and complex than the other Ogier wines. Really lovely and should be drunk on its own away from the Gentaz wines that tend to upstage it.

feature matrix X

Strong Neg	Neg	Pos	Strong Pos
1	1		1
		2	3
	1	3	
		2	
	1		

Learning weights and bias

- Logistic regression is an instance of supervised learning
- We know the correct label y for each training observation x
- The system produces
- We want to learn parameters to make as close as possible to y

function STOCHASTIC GRADIENT DESCENT($L()$, $f()$, x , y) **returns** θ

where: L is the loss function

f is a function parameterized by θ

x is the set of training inputs $x^{(1)}$, $x^{(2)}$, ..., $x^{(n)}$

y is the set of training outputs (labels) $y^{(1)}$, $y^{(2)}$, ..., $y^{(n)}$

$\theta \leftarrow 0$

repeat T times

For each training tuple $(x^{(i)}, y^{(i)})$ (in random order)

Compute $\hat{y}^{(i)} = f(x^{(i)}; \theta)$ # What is our estimated output \hat{y} ?

Compute the loss $L(\hat{y}^{(i)}, y^{(i)})$ # How far off is $\hat{y}^{(i)}$ from the true output $y^{(i)}$?

$g \leftarrow \nabla_{\theta} L(f(x^{(i)}; \theta), y^{(i)})$ # How should we move θ to maximize loss ?

$\theta \leftarrow \theta - \eta g$ # go the other way instead

return θ

Training classifiers in Python

```
1 #!/bin/python
2
3 import os
4 import sys
5 import string
6 import random
7 import operator
8 from sklearn.tree import export_graphviz
9 from sklearn.tree import DecisionTreeClassifier
10 from sklearn.naive_bayes import MultinomialNB
11 from sklearn.linear_model import LogisticRegression
12 from sklearn.preprocessing import LabelEncoder
13 from sklearn.feature_extraction import DictVectorizer
14 from sklearn.cross_validation import train_test_split
15 from sklearn.externals.six import StringIO
16
17 #read in raw data from file and return a list of (label, article) tuples
18 def get_data(filename):
19     data = [line.strip().split('\t') for line in open(filename).readlines()]
20     random.shuffle(data)
21     return data
22
23 #this function builds the feature matrix for the Decision Tree.
24 def get_dtreet_features(X) :
25     features = []
26     feature_list = []
27     wordCounts = {}
28
29     for x in X :
```

```
1 #!/bin/python
2
3 import os
4 import sys
5 import string
6 import random
7 import operator
8 from sklearn.tree import export_graphviz
9 from sklearn.tree import DecisionTreeClassifier
10 from sklearn.naive_bayes import MultinomialNB
11 from sklearn.linear_model import LogisticRegression
12 from sklearn.preprocessing import LabelEncoder
13 from sklearn.feature_extraction import DictVectorizer
14 from sklearn.cross_validation import train_test_split
15 from sklearn.externals.six import StringIO
16
17 #read in raw data from file and return a list of (label, a
18 def get_data(filename):
19     data = [line.strip().split('\t') for line in open(file
20     random.shuffle(data)
21     return data
22
23 #this function builds the feature matrix for the Decision
24 def get_dtreet_features(X) :
25     features = []
26     feature_list = []
27     wordCounts = {}
28
29     for x in X :
```

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

Experimental design in machine learning

- Splitting data into training / test sets
- Baselines
- Evaluation

Training/test split

- Typically we have a fixed set of labeled data that we run experiments on
- In our experiments we typically split the data into a training set, and a disjoint test set
- Why?

It is generalization that counts

- The fundamental goal of machine learning is to generalize beyond the examples in the training set
- No matter how much data we have, at test time we are unlikely to see exactly the same items

The problem of overfitting

- Sometimes our classifier *overfits* the data
- It encodes random quirks of the data instead of learning good generalizations
- Symptom: your learner creates a classifier that is 100% accurate on the training data but only 50% accurate on test data

Our data may be too easy

- Jennifer Mascia described how she wrote the Gun Report for the NYTimes 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.

n-fold cross validation

- Splitting the data reduces the amount of available data for training
- Mitigated through *cross-validation*: randomly dividing your training data into 10 pieces, train on 9 test on 1, average results

Precision and Recall

		Actual Class
Predicted class	True positive: Correct result	False positive: Unexpected result
	False negative: Missing result	True negative: Correct absence of result

$$Precision = \frac{tp}{tp + fp}$$

$$Recall = \frac{tp}{tp + fn}$$

Precision and Recall

		Actual Class
Predicted class	True positive: Correct result	False positive: Unexpected result
	False negative: Missing result	True negative: Correct absence of result

$$Precision = \frac{tp}{tp + fp}$$

Precision and Recall

		Actual Class
Predicted class	True positive: Correct result	False positive: Unexpected result
	False negative: Missing result	True negative: Correct absence of result

$$Recall = \frac{tp}{tp + fn}$$

Accuracy

		Actual Class
Predicted class	True positive: Correct result	False positive: Unexpected result
	False negative: Missing result	True negative: Correct absence of result

$$Accuracy = \frac{tp + tn}{tp + tn + fp + fn}$$

Evaluating a system

- Say your system gets 93% accuracy, is that good?

Baselines

- Our training data is imbalanced:
 - 8973 positive examples
 - 62811 negative examples
- A system that always guessed "not a gun related article" would get 87% accuracy
- This is the "majority class baseline"
- The rule-based system that guess + iff "shooting" occurs in the article and - otherwise gets 93%

Learning curves

