

# Intro to Machine Learning

**Crowdsourcing and Human Computation**

**Instructor: Chris Callison-Burch**

**Website: [crowdsourcing-class.org](http://crowdsourcing-class.org)**

# 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 ▾

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](#)These recommendations are based on [items you own](#) and more.view: [All](#) | [New Releases](#) | [Coming Soon](#)[More results](#) ▾

### 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](#)

#### [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[17 used & new from \\$11.02](#)[Add to Cart](#)[Add to Wish List](#) 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](#))

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

by FridaBaby (April 1, 2010)

Average Customer Review: ★★★★★ (1,859)

In Stock

**List Price:** \$15.99**Price:** \$14.78[43 new from \\$9.86](#)[Add to Cart](#)[Add to Wish List](#) 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](#))

#### [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:** \$52.99[Add to Cart](#)[Add to Wish List](#)

Shop by  
Department ▾

Search

All ▾

Your Amazon.com

Your Browsing History

[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: \$15.99**  
**Price: \$14.78**  
[43 new from \\$9.86](#)

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

Rate this item  
 I own it  
 Not interested

### Because you purchased...



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



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



[My Brest Friend Original Pillow, Bluebells](#)  
by Zenoff Products  
 5 star rating  
 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  
 5 star rating  
 Don't use for recommendations

[More results](#) [Wish List](#)[Shopping Cart and more \(Fix this\)](#)[Wish List](#)[Shopping Cart and more \(Fix this\)](#)

# Recommended for you, Chris



Buy It Again in Grocery

3 ITEMS



Children's Books

100 ITEMS



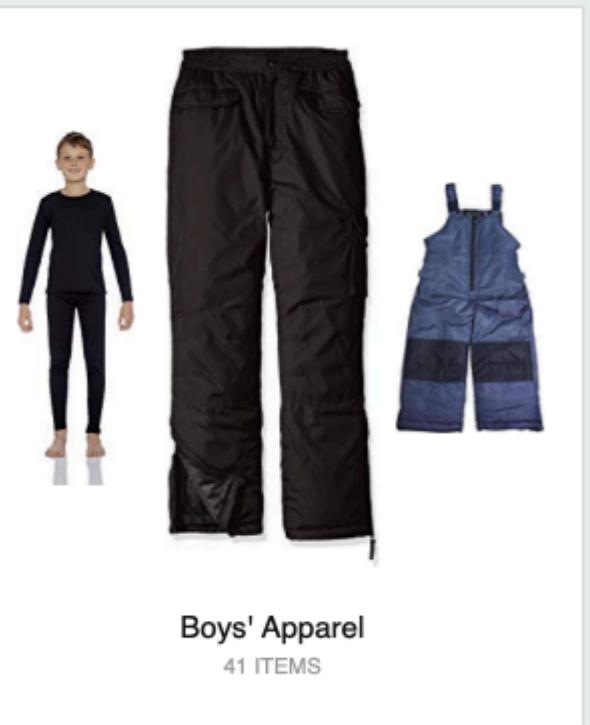
Buy it Again in Household Supplies

5 ITEMS



Buy It Again in Pets

3 ITEMS



Boys' Apparel

41 ITEMS



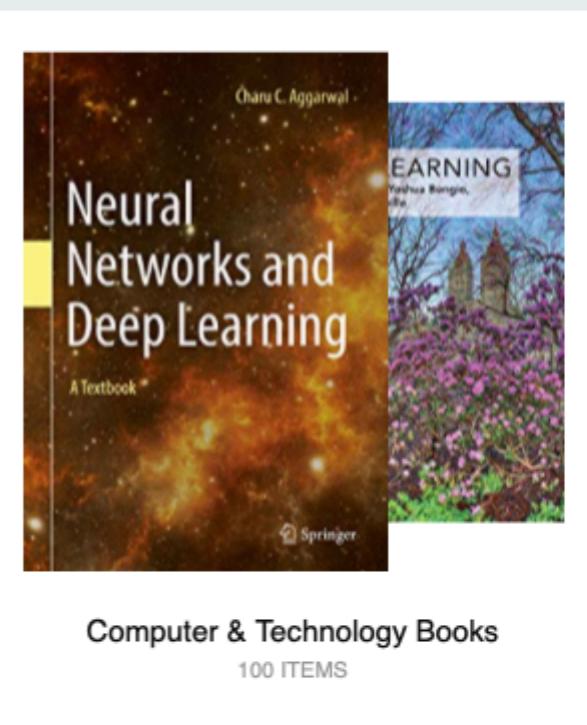
Arts & Crafts Supplies

68 ITEMS



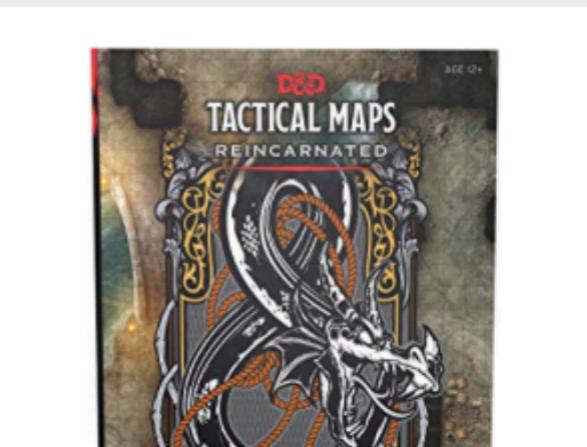
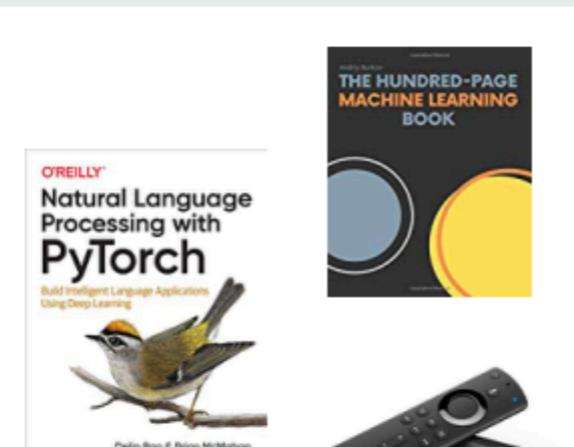
Literature & Fiction

100 ITEMS

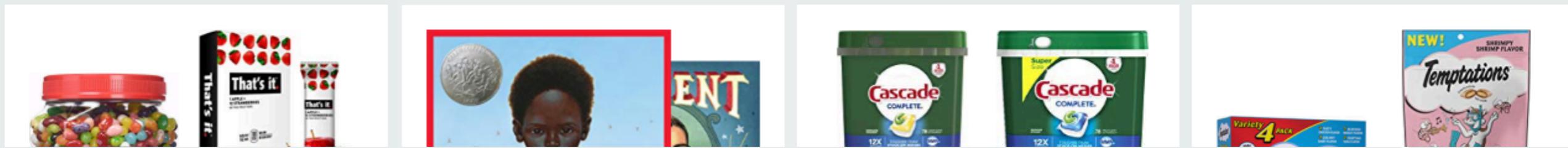


Computer & Technology Books

100 ITEMS



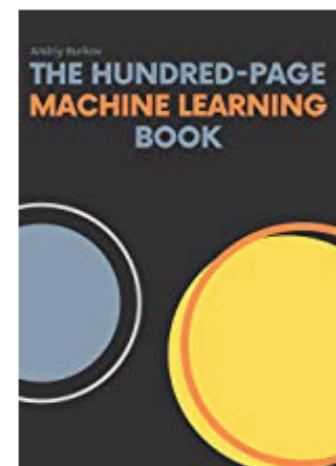
# Recommended for you, Chris



## Computer & Technology Books

[View All & Manage](#)

Page 1 of 20



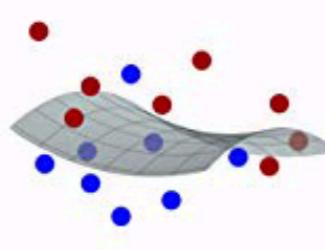
The Hundred-Page Machine Learning Book

Andriy Burkov

★★★★★ 9

\$33.65

Foundations of Machine Learning second edition



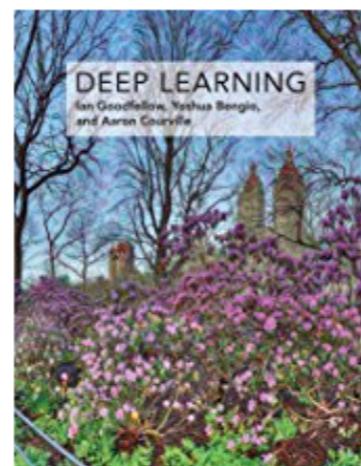
Mehryar Mohri,  
Afshin Rostamizadeh,  
and Ameet Talwalkar

Foundations of Machine Learning (Adaptive...

Mehryar Mohri

★★★★★ 2

\$64.56

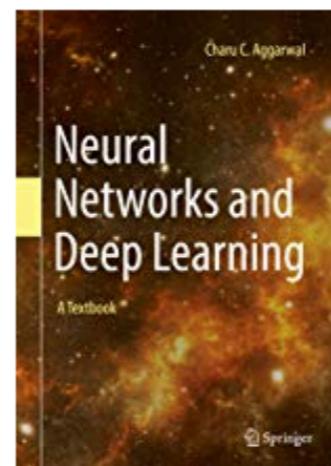


Deep Learning (Adaptive Computation and Machine...

Ian Goodfellow

★★★★★ 184

\$65.51

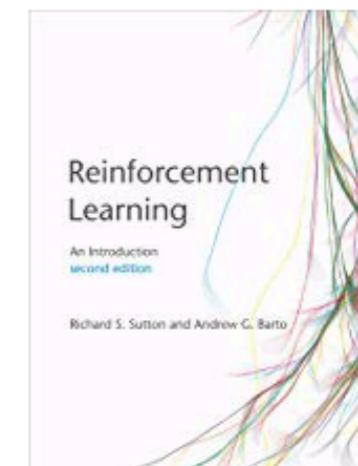


Neural Networks and Deep Learning: A Textbook

Charu C. Aggarwal

★★★★★ 14

\$53.12

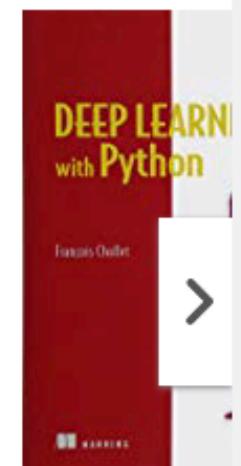


Reinforcement Learning: An Introduction (Adaptive...

Richard S. Sutton

★★★★★ 12

\$63.00



Deep Learning with Python  
Francois Chollet

★★★★★ 88

\$18.99 prime

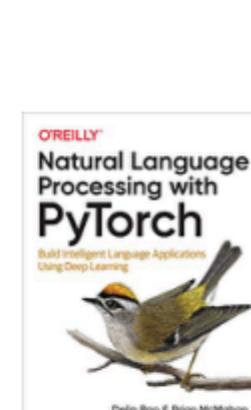
## Boys' Apparel

41 ITEMS



## Arts & Crafts Supplies

68 ITEMS



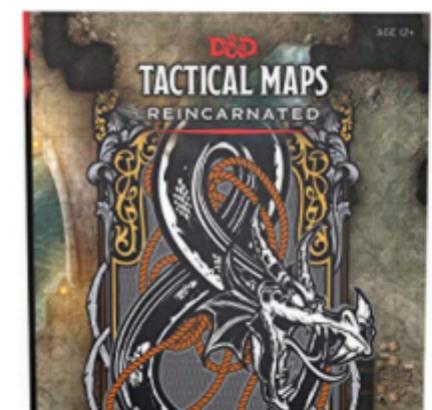
## Literature & Fiction

100 ITEMS



## Computer & Technology Books

100 ITEMS



# People

18 People



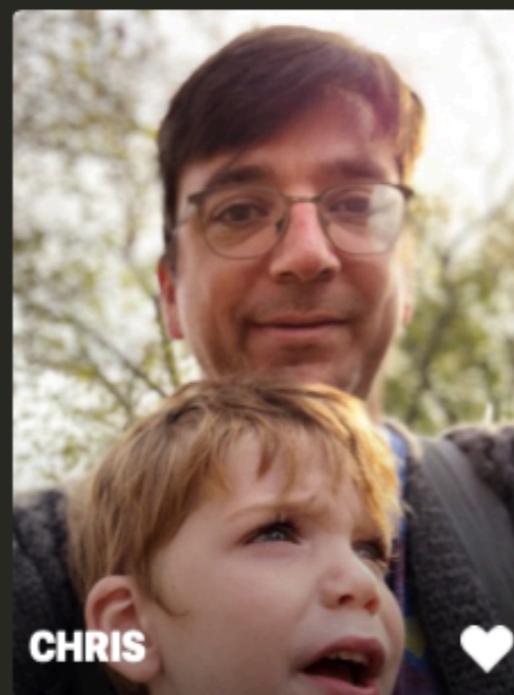
REED



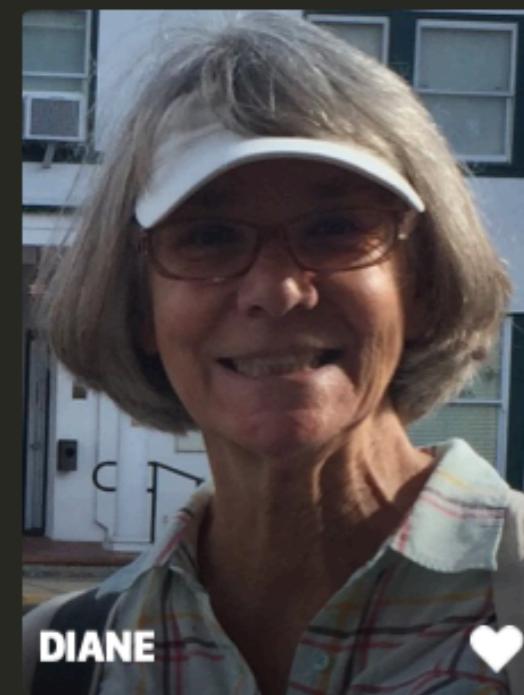
WILLA



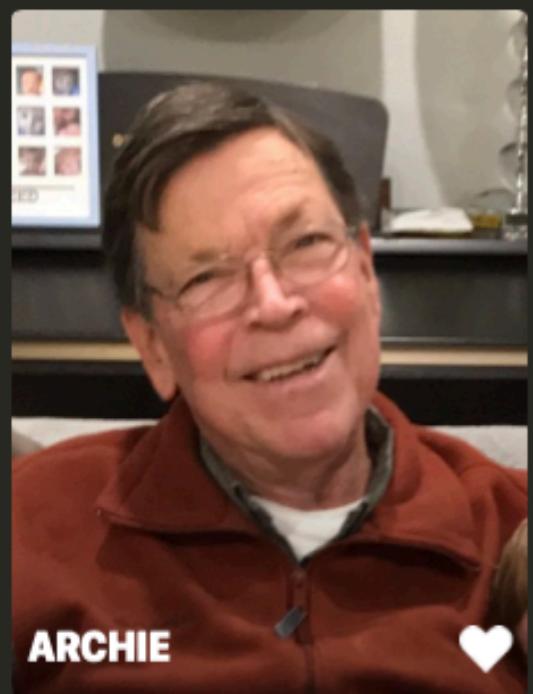
DAWN



CHRIS



DIANE



ARCHIE



KAREN



MARGARETTA



ALEXA



VANESSA CALLISON-... ❤️



154 Photos Added

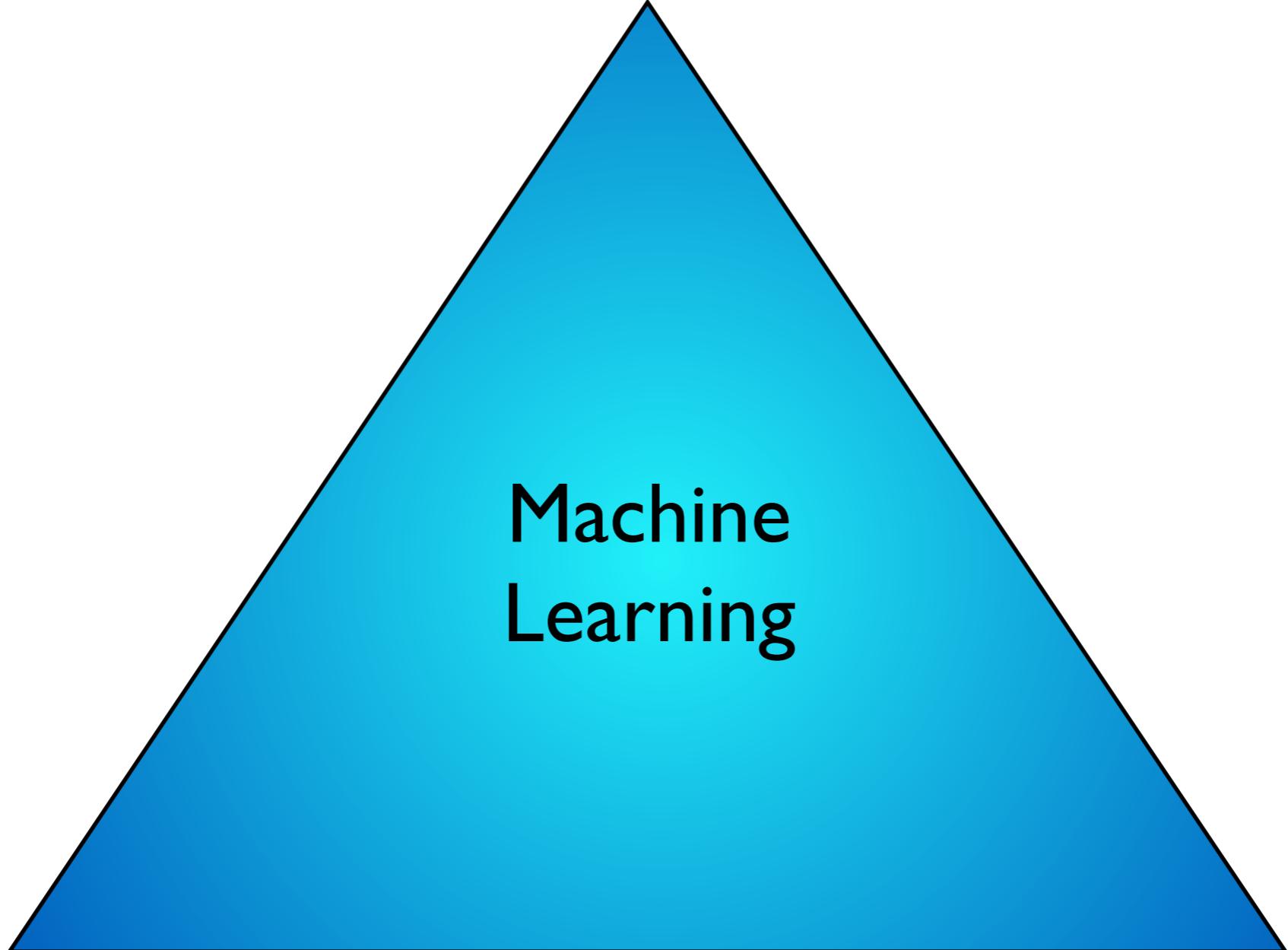
Is this Willa Mercurio Callison?

No

Yes



Data



Machine  
Learning

Model

Optimization

# 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](http://crowdsourcing-class.org/resources.html)
- Some are owned by companies
- Sometimes you can assemble it yourself
- Crowdsourcing!

# Yelp Dataset Challenge

Discover what insights lie hidden in our data.



## What is the dataset challenge?

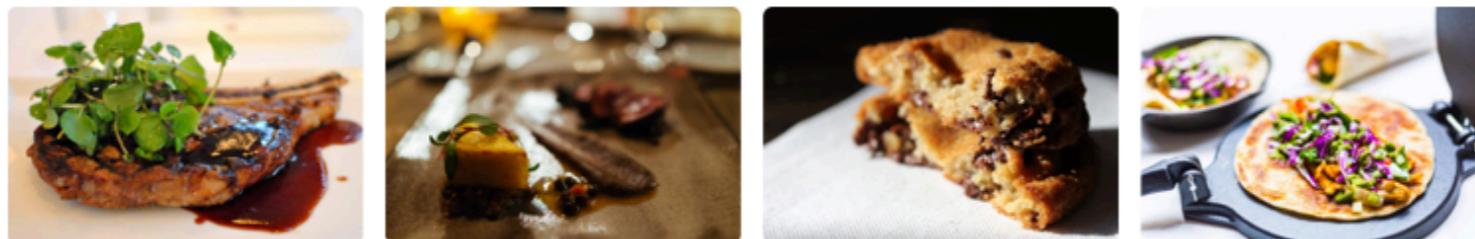
The challenge is a chance for students to conduct research or analysis on our data and share their discoveries with us. Whether you're trying to figure out how food trends start or identify the impact of different connections from the local graph, you'll have a chance to win cash prizes for your work! 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?

Search datasets...



Order by:

Popular

Datasets ordered by Popular

## Filter by location

Clear

Enter location...

Map tiles & Data by [OpenStreetMap](#), under [CC BY SA](#).

## Topics

A-Z

1-9

Clear All

Local Government (17888)

Agriculture (517)

Education (428)

AAPI (421)

Climate (400)

Show More Topics

## Topic Categories

A-Z

1-9

Clear All

Asian (199)

Total Energy (191)

Pacific Islander (100)

## 304,468 datasets found

**Pittsburgh Wards Map** 2527 recent views

*Allegheny County / City of Pittsburgh / Western PA Regional Data Center* – Allows users to look up City of Pittsburgh Wards

[HTML](#)

County

**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

Federal

**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](#)

Federal

**Demographic Statistics By Zip Code** 1723 recent views

*City of New York* – Demographic statistics broken down by zip code

[CSV](#) [RDF](#) [JSON](#) [XML](#)

City

**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...

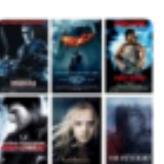
[CSV](#) [RDF](#) [JSON](#) [XML](#)

City

**College Scorecard** 1250 recent views

*Department of Education* – College Scorecards make it easier for students to search for a college that is

Feder

2508		<b>Credit Card Fraud Detection</b> Anonymized credit card transactions labeled as fraudulent or genuine Machine Learning Group - ULB updated a year ago (Version 3)	crime finance	 CSV  66 MB  ODbL	</> 1k 37 1m
1590		<b>European Soccer Database</b> 25k+ matches, players & teams attributes for European Professional Football Hugo Mathien updated 2 years ago (Version 10)	association... europe	 SQLite  34.4 MB  ODbL	</> 1k 88 558k
1373		<b>TMDB 5000 Movie Dataset</b> Metadata on ~5,000 movies from TMDb The Movie Database (TMDb) updated a year ago (Version 2)	film	 CSV  9.3 MB  Other	</> 1k 51 564k
1147		<b>Global Terrorism Database</b> More than 180,000 terrorist attacks worldwide, 1970-2017 START Consortium updated 5 months ago (Version 3)	crime terrorism internation...	 CSV  27.9 MB  Other	</> 667 12 273k
1117		<b>Wine Reviews</b> 130k wine reviews with variety, location, winery, price, and description zackthoutt updated a year ago (Version 4)	critical the... food and dr...	 CSV  50.9 MB  CC4	</> 1k 19 229k
1056		<b>Bitcoin Historical Data</b> Bitcoin data at 1-min intervals from select exchanges, Jan 2012 to November 2018 Zielak updated 3 months ago (Version 15)	history finance	 CSV  110.8 MB  CC4	</> 108 25 264k
934		<b>Google Play Store Apps</b> Web scraped data of 10k Play Store apps for analysing the Android market. Lavanya Gupta updated 9 days ago (Version 6)	video games computer s... internet mobile web	 CSV  1.9 MB  Other	</> 192 22 216k
815		<b>Trending YouTube Video Statistics</b> Daily statistics for trending YouTube videos Mitchell J updated 3 months ago (Version 114)			
801		<b>Pokemon with stats</b> 721 Pokemon with stats and types Alberto Barradas updated 2 years ago (Version 2)			

kaggle™

# 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

- winemag-data-130k... 130k x 14
- winemag-data\_first15... 151k x 11
- winemag-data-130k-v2.json

## About this file

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.

## Columns

- #
- country The country that the wine is from
- description
- 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

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

14 of 14 columns ▾

Views

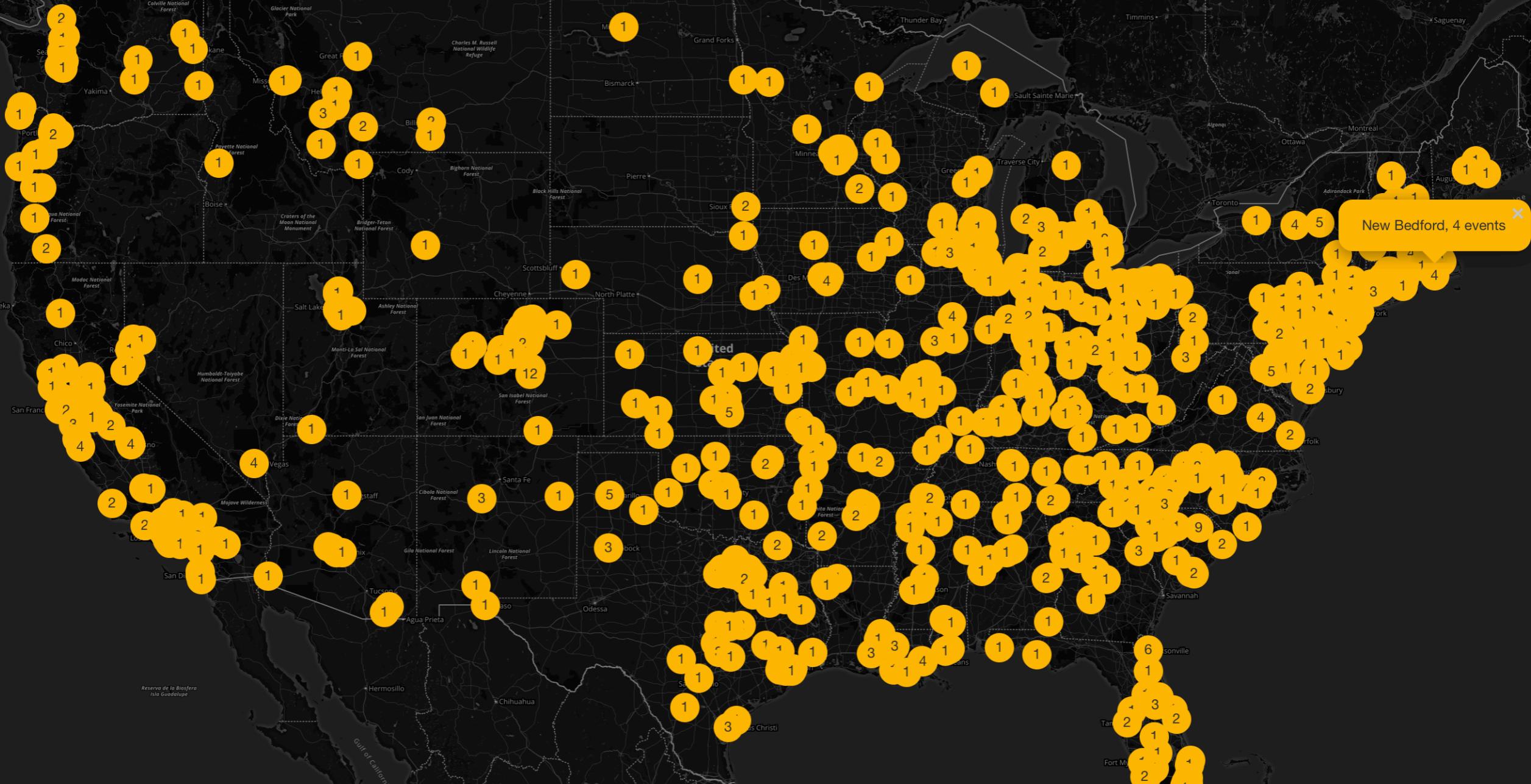


A country	A description	A designation	# points	# price	A province
The country that the wine is from		The vineyard within the winery where the grapes that made the wine are from	The number of points WineEnthusiast rated the wine on a scale of 1-100 (though they say they only post reviews for wines)	The cost for a bottle of the wine	The province or the wine is from
US 42%		Reserve 2%		20.0 5%	California
France 17%	119955 unique values	Estate 1%		15.0 5%	Washington
Other (41) 41%		Other (37977) 97%		Other (388) 90%	Other (423)
Italy	Aromas include tropical fruit, broom, brimstone and dried herb. The	Vulkà Bianco	87		Sicily & Sar

# 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

# 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

CRIME

MURDER, THEFT, AND OTHER WICKEDNESS.

SEPT. 16 2013 3:34 PM

Slate

# How Many People Have Been Killed by Guns Since Newtown?



122k

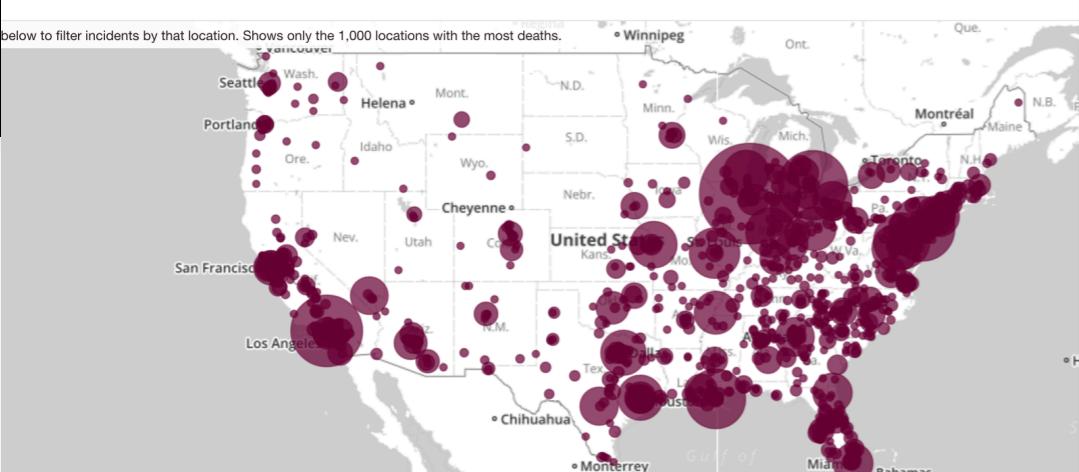
20

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](mailto: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.



## The Counted

People killed by police in the US

SEND A TIP DATABASE ABOUT READ ARTICLES JOIN US: [f](#) [t](#) [e](#)

PEOPLE KILLED IN 2016

7 | 9 | 3  
4



2016 2015 List Map

Search by name:

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
AR	MT	NE NV KY LA VT WI TN	Total killed: 21
OR	CA MO MS FL NC WA SC KS UT	Population: 2,085,109	
TX	VA GA RI MD MN ID OH IN IL	Rank (per capita): 1 / 51	
CT	MA ND IA MI PA DE NJ NY ME	Rank (total): 7 / 51	
NH			

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

LOGIN CONTACT US [Search Database](#) [f](#) [t](#)

### GUN VIOLENCE Archive



PHOTO CREDIT: DAVID LASSMAN, 2007

### GUN VIOLENCE Archive 2016

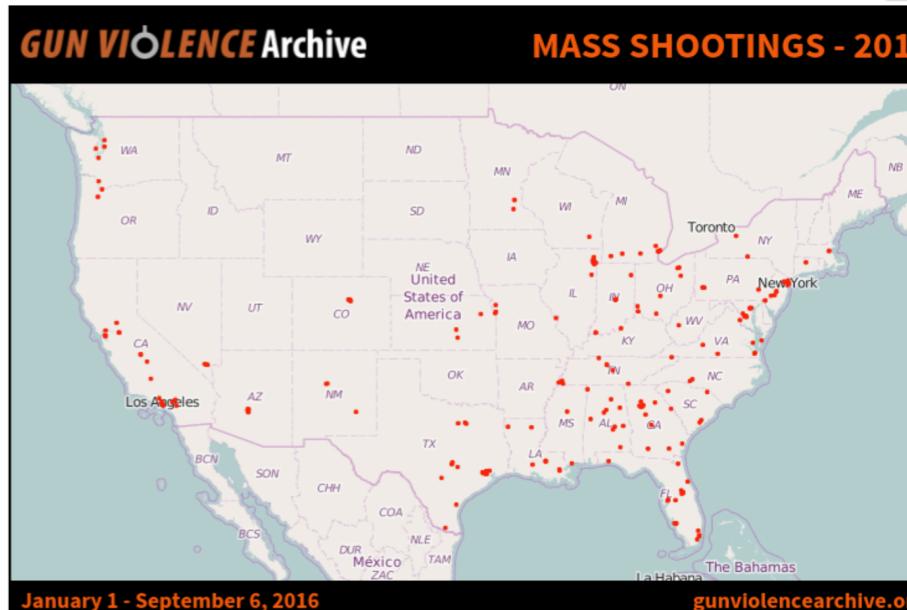
Total Number of Incidents	41,309
Number of Deaths <sup>1</sup>	10,599
Number of Injuries <sup>1</sup>	22,023
Number of Children (age 0-11) Killed or Injured <sup>1</sup>	477
Number of Teens (age 12-17) Killed or Injured <sup>1</sup>	2,252
Mass Shooting <sup>2</sup>	285
Officer Involved Incident	233
Officer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



**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 1



Feds Inv  
to 'Consu  
Pipeline I



What Co  
Know Ab  
Security I



Teen Wh  
Sprayed S



Dentist A  
Harming  
Making M



'I Did the  
Orlando I



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
<b>Total of four diseases</b>	<b>2031</b>	<b>320</b>
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 terrible 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



THE UPSHOT | The Science Behind Suicide Contagion

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 sharp increase in 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 a subsequent 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. 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' in large white font. Underneath 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 sharing icons: Facebook, Twitter, and Email.

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

# 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 ...	<a href="http://www.nbciami.com/news/local/Man-Shot-Near-Pierce-Park-in-Coral-Gables-Police-360424171.html">http://www.nbciami.com/news/local/Man-Shot-Near-Pierce-Park-in-Coral-Gables-Police-360424171.html</a>	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 ...	<a href="http://www.huffingtonpost.com/jillian-gutowitz/im-voting-for-hillary-clinton-because-shes-a-woman_b_8684910.html">http://www.huffingtonpost.com/jillian-gutowitz/im-voting-for-hillary-clinton-because-shes-a-woman_b_8684910.html</a>	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 ...	<a href="http://www.fox10tv.com/story/30649388/child-attends-school-despite-allegedly-being-stabbed-by-guardian-mother">http://www.fox10tv.com/story/30649388/child-attends-school-despite-allegedly-being-stabbed-by-guardian-mother</a>	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 ...	<a href="http://gulfnews.com/news/mena/palestine/palestinian-who-attacked-soldier-shot-dead-1.1631076">http://gulfnews.com/news/mena/palestine/palestinian-who-attacked-soldier-shot-dead-1.1631076</a>	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 ...	<a href="http://www.walb.com/story/30654553/12-year-old-child-abandoned-at-atlanta-police-station">http://www.walb.com/story/30654553/12-year-old-child-abandoned-at-atlanta-police-station</a>	WALB 10 News	child shot



## 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

## Wrangler News - Online Edition - Home

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

LIVE Budget s  
clash in Rajya

Show full text

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 to the scene. Police 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](#)

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



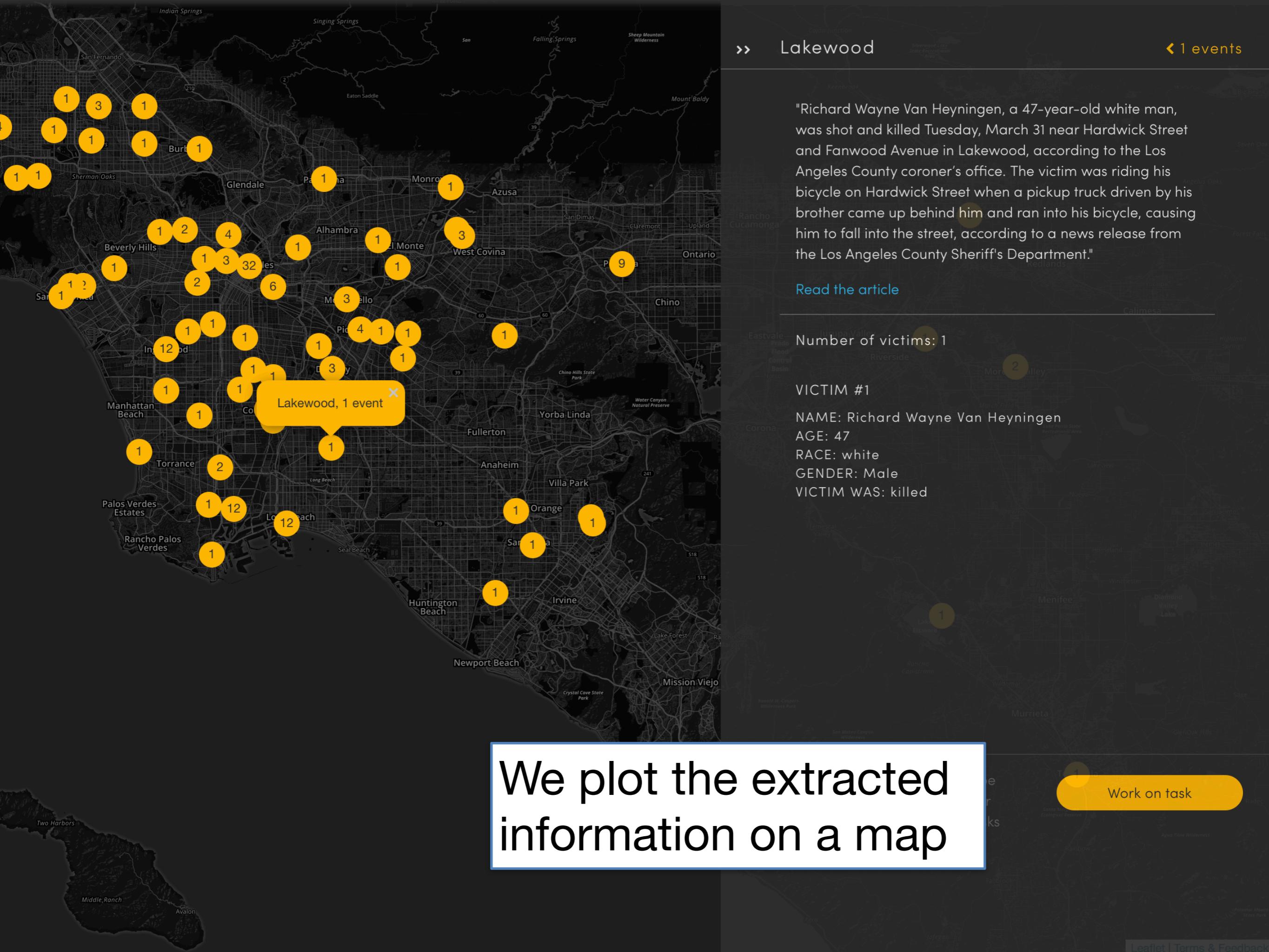
City

Baltimore



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

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



# 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 New York Times

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."

Meanwhile, on a playground at a school in Milwaukee last Wednesday, 10-year-old Sierra Guyton was caught in the crossfire of a shootout. She is in stable condition, but as of yesterday, she is not responsive. A fund has been created for her family, and a rally was held in her honor on Memorial Day.

The suspect is an 18-year-old with a long criminal record, who had been wounded by gunfire a week before he allegedly shot Sierra. A witness said 20 shots were fired in the direction of the playground. The suspect told police that he fired his gun until it was empty, then

# 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.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/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/michael_day_kalamazoo.html)



[http://www.mlive.com/news/kalamazoo/index.ssf/2014/05/15-year-old\\_charged\\_with\\_murde.html](http://www.mlive.com/news/kalamazoo/index.ssf/2014/05/15-year-old_charged_with_murde.html)



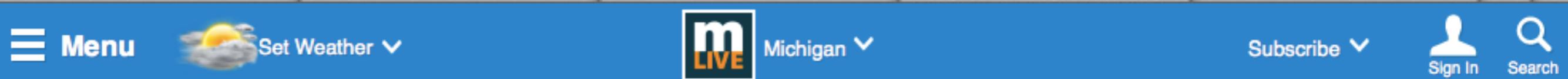
<http://www.jsonline.com/news/crime/girl-10-on-life-support-after-being-hit-in-playground-shootout-b99275748z1-260251491.html>



<http://fox6now.com/2014/05/29/fund-created-for-sierra-guyton-victim-of->



# Extracting web page text



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

5  
comments



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



Gazette File

### High School Football



**Follow the latest prep football news from around Michigan**

- Statewide schedules
- Recruiting news
- State rankings



# Newspaper3k: Article scraping & curation

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

## Useful Links

[Newspaper @ GitHub](#)

[Newspaper @ PyPI](#)

[Issue Tracker](#)

## This Page

[Show Source](#)

## Quick search

Go

## 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']  
  
>>> article.publish_date  
datetime.datetime(2013, 12, 30, 0, 0)
```

```
>>> 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.

Police said no arrests related to this incident have been made at this time.

This was the second reported shooting in Kalamazoo Saturday. A 21-year-old Kalamazoo man was also shot in the back while leaving a party in the Northside neighborhood around 2:30 a.m. and was treated at Bronson Hospital for non-life threatening injuries, police said.

```
>>> print(article.title)
```

# Representing data with Features

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

# Representing data with Features

## X?

y



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 Contez 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.**

Thin and completely **uninspiring**.

## subjectivity lexicon

Word	Polarity	Strength
abandoned	negative	weak
abandonmen	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.**

Thin and completely **uninspiring**.

## feature matrix X

Strong Neg	Neg	Pos	Strong Pos
------------	-----	-----	------------

1	1		1
---	---	--	---

2	3
---	---

1	3
---	---

2
---

# 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  $b$

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

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

The diagram illustrates the components of the logistic regression equation. The equation is  $z = w \cdot x + b$ . The first part,  $w \cdot x$ , is shown as a summation from  $i=1$  to  $n$  of  $w_i x_i$ . Arrows point from the labels "weight" and "feature value" to the  $w_i$  and  $x_i$  terms respectively. An arrow points from the label "bias term" to the constant  $b$ .

# What's wrong with z?

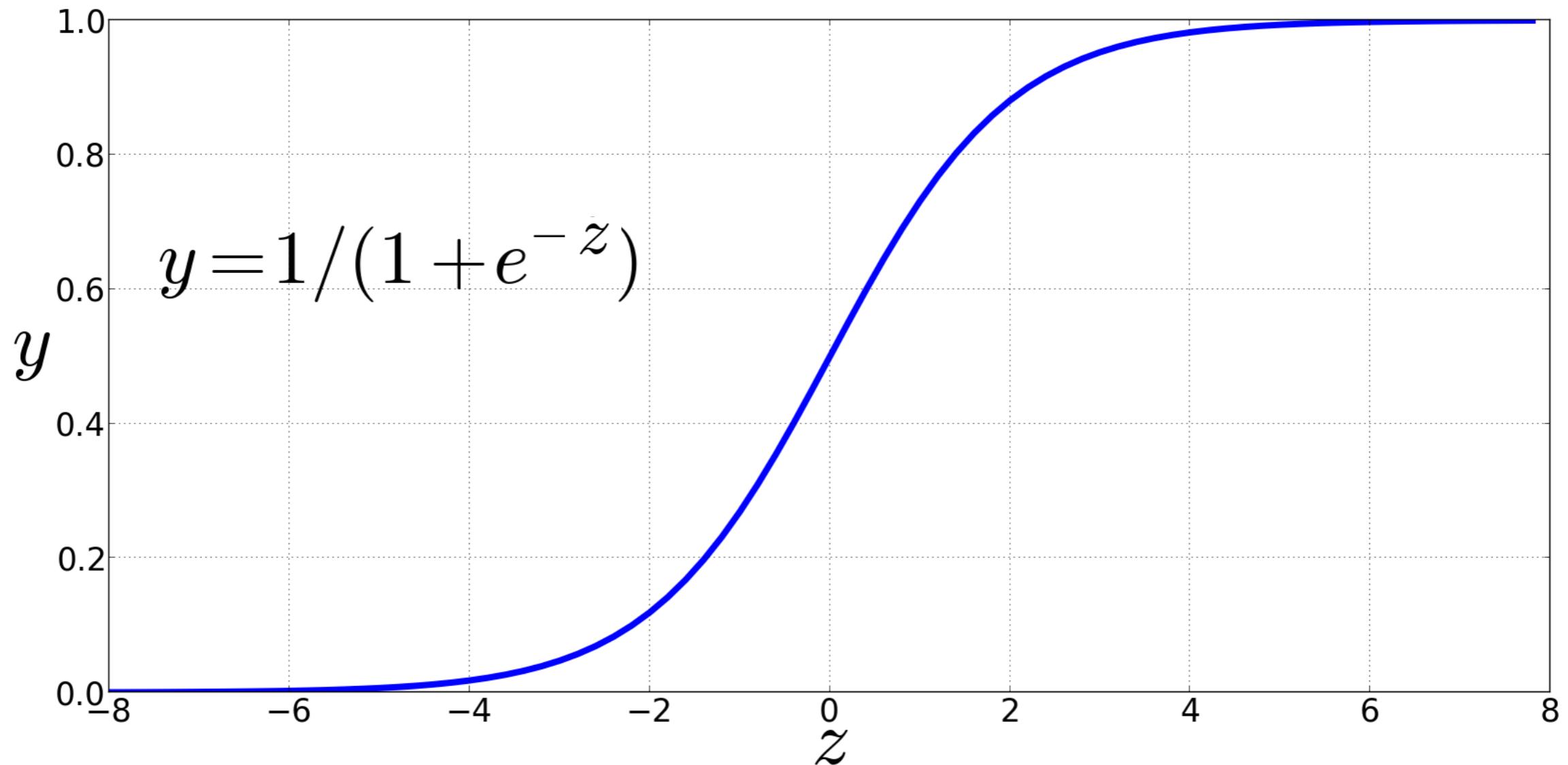
- It isn't a legal probability!
- Weights can be real valued, and might even be negative.
- $z$  ranges from  $-\infty$  to  $+\infty$

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

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

The diagram shows the mathematical expression for a weighted sum. A large bracket encloses the summation part, with an arrow pointing to it labeled "weight". Another arrow points to the individual terms within the summation, labeled "feature value". A final arrow points to the constant term "b", which is labeled "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

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

# Once I have my features, what else do I need?

## 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.**

Thin and completely **uninspiring**.

## feature matrix X

Strong Neg	Neg	Pos	Strong Pos
------------	-----	-----	------------

1	1		1
---	---	--	---

		2	3
--	--	---	---

	1	3	
--	---	---	--

		2	
--	--	---	--

# 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  $\hat{y}$
- We want to learn parameters to make  $\hat{y}$  as close as possible to  $y$

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

# where: L is the loss function

# f is a function parameterized by  $\theta$

# 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  $\theta$  to maximize loss ?

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

**return**  $\theta$

# 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 :
30         f = {}
31         for w in [word.strip(string.punctuation) for word in x.split()] :
32             if not w in wordCounts :
33                 wordCounts[w] = 0
34                 wordCounts[w] += 1
35         for w in wordCounts:
36             if wordCounts[w] > 5000:
```

1	#!/bin/python	2.891467
2		2.560138
3	import os	2.342422
4	import sys	2.012679
5	import string	1.925938
6	import random	1.706036
7	import operator	1.673353
8	from sklearn.tree import export_graphviz	1.626867
9	from sklearn.tree import DecisionTreeClassifier	1.597359
10	from sklearn.naive_bayes import MultinomialNB	1.505322
11	from sklearn.linear_model import LogisticRegression	1.449963
12	from sklearn.preprocessing import LabelEncoder	1.419679
13	from sklearn.feature_extraction import DictVectorizer	1.418678
14	from sklearn.cross_validation import train_test_split	1.324652
15	from sklearn.externals.six import StringIO	1.279014
16		1.225336
17	#read in raw data from file and return a list	1.214517
18	def get_data(filename):	1.211558
19	data = [line.strip().split('\t') for line in open(filename)]	1.204958
20	random.shuffle(data)	1.147636
21	return data	1.142863
22		1.136201
23	#this function builds the feature matrix for	1.117853
24	def get_dtreet_features(X) :	1.108006
25	features = []	1.098322
26	feature_list = []	1.095423
27	wordCounts = {}	1.094982
28		1.092020
29	for x in X :	1.086457
30	f = {}	
31	for w in [word.strip(string.punctuation) for word in x]:	
32	if not w in wordCounts :	
33	wordCounts[w] = 0	
34	wordCounts[w] += 1	
35	for w in wordCounts:	
36	if wordCounts[w] > 5000:	

# 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

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

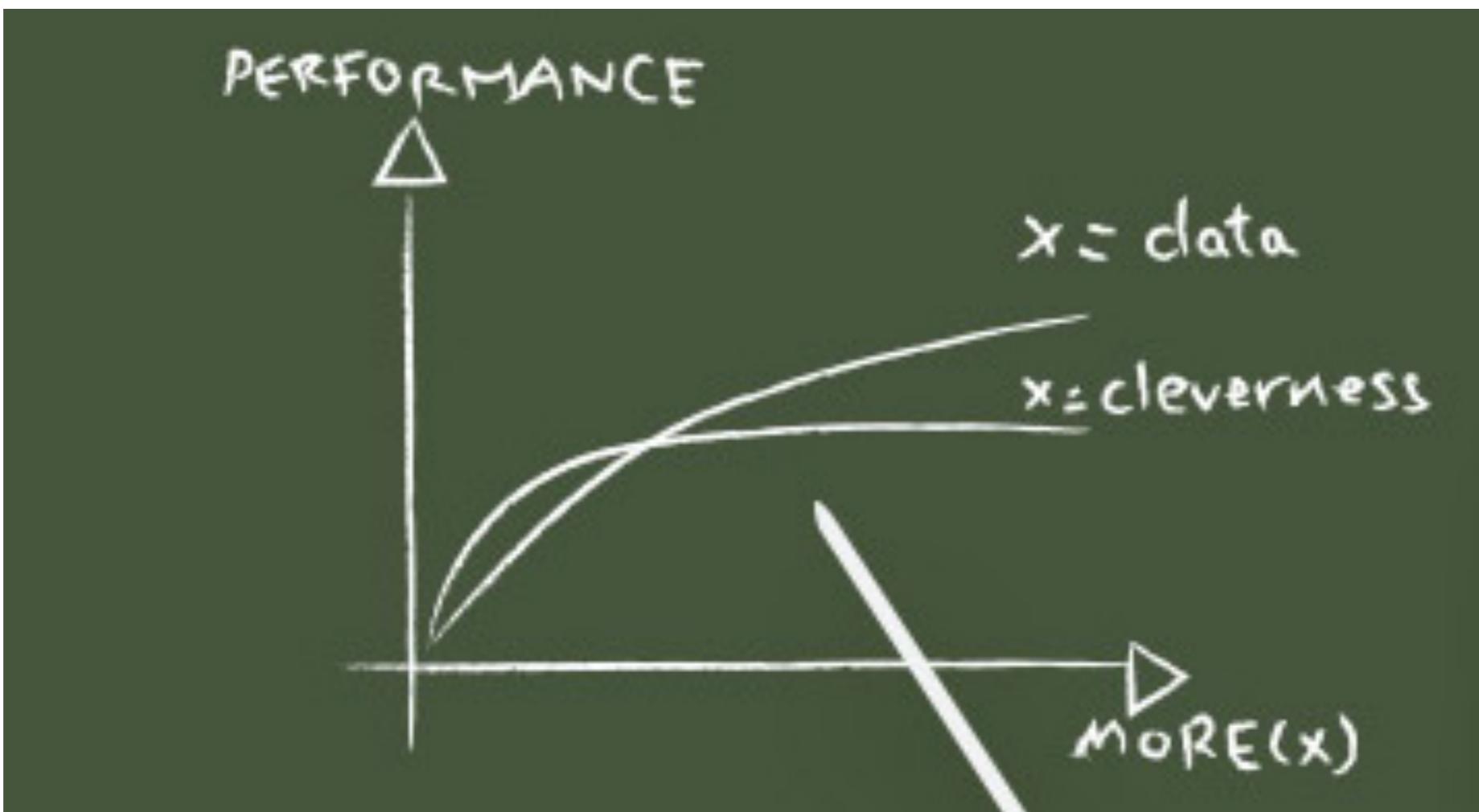
# Evaluating a system

- 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



# Your Next Homework

newfoundland



american\_pit\_bull\_terminer



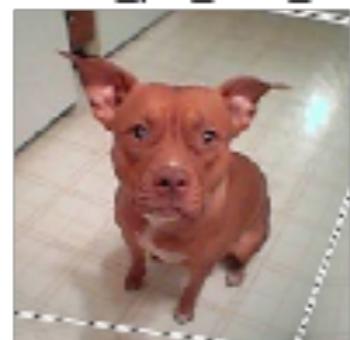
Bombay



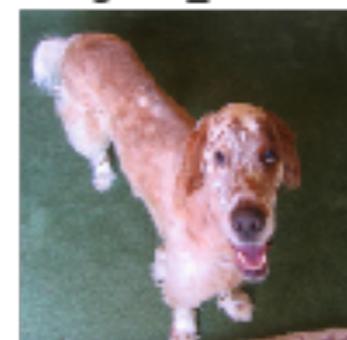
Abyssinian



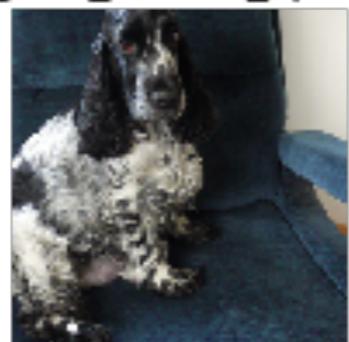
american\_pit\_bull\_terminer



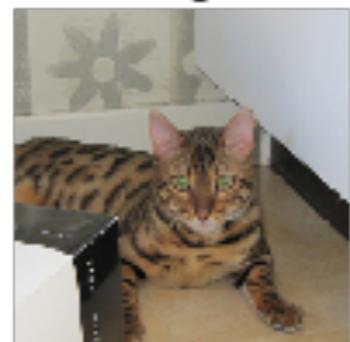
english\_setter



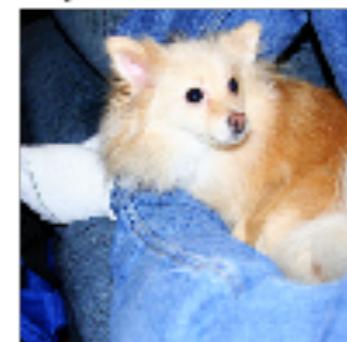
english\_cocker\_spaniel



Bengal



pomeranian



**fast.ai**

Making neural nets  
uncool again