Intro to Machine Learning

Crowdsourcing and Human Computation

Instructor: Chris Callison-Burch

Website: 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 90s we rediscovered probabilistic models and statistics and applied it to CS and machine learning
- We now have access to much better computing infrastructure
- 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 email spam? What language is this web page written in? Whose face is shown in a picture?

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: ★★★★½ (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 item

Recommended 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: ★★★★½ (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 item

Recommended 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: ★★★★½ (627)
In Stock

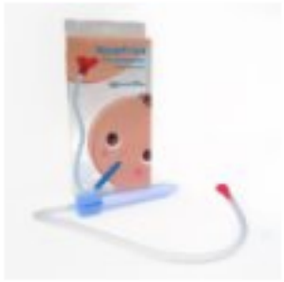
List Price: ~~$54.99~~

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

☒ ☆☆☆☆☆
☐ 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

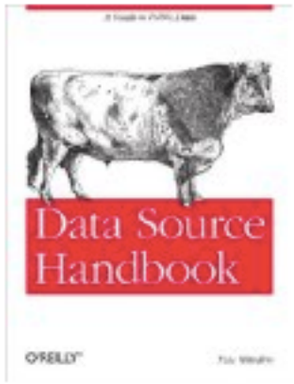
Because your Wish List includes...

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

More Items to Consider

You viewed

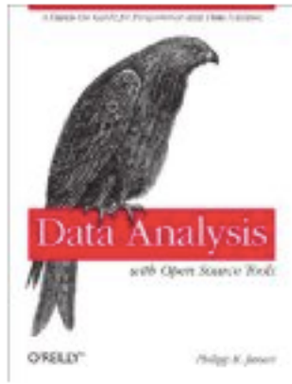


Data Source Handbook
› Pete Warden
Paperback
★★☆☆☆ (8)
~~$29.99~~ $20.78

Customers who viewed this also viewed

View or edit your browsing history ›

Big Data Glossary
› Pete Warden
Paperback
★★★☆☆ (5)
~~$19.99~~ $14.78



Data Analysis with Open Source Tools
› Philipp K. Janert
Paperback
★★★★☆ (35)
~~$39.99~~ $24.22

1. ASICS Women's GT-1000 Running Shoe
ASICS

2. ASICS Women's GEL-Noosa Tr Running Shoe
ASICS

3. ASICS Women's Gel-Nimbus 14 Running Shoe
ASICS

4. ASICS Women's Gel-Kayano 19 Running Shoe
ASICS

5. ASICS Women's GEL-Cumulus Running Shoe
ASICS

›See all Best Sellers in Women's Running

Related to Items You've Viewed

You viewed

Bose QuietComfort 20i Acoustic Noise...
★★★★½ (144)

Customers who viewed this also viewed

View or edit your browsing history ›

Bose IE2 Audio Headphones
★★★★½ (1,153)
$99.95

Bose® MIE2i Mobile Headset
★★★★☆ (632)
$129.95

Best Sellers

Gift Cards Store

Updated hourly

1. Amazon Gift Card - E-mail
Amazon
$50.00

2. Amazon Gift Card - E-mail - Hap Birthday (Candles)
Amazon
$50.00

3. Amazon Gift Card - E-mail - Tha You (Note)
Amazon
$50.00

4. Amazon Gift Card Upload Your Photo - Gift for You
Amazon
$50.00

5. Amazon.com Gift Cards - E-mail

More Items to Consider

You viewed

Customers who viewed this also viewed

View or edit your browsing history

1. ASICS Women's GT-1000 Running Shoe
ASICS

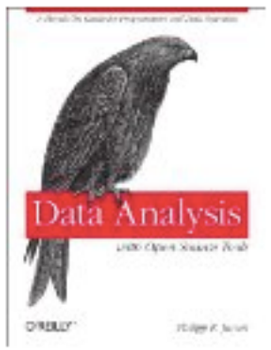
2. ASICS Women's GEL-Noosa T
Running Shoe
ASICS

3. ASICS Women's Gel-Nimbus 1
Running Shoe
ASICS

4. ASICS Women's Gel-Kayano 1

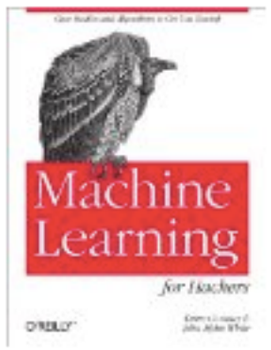
Continue Shopping: Customers Who Bought Items in Your Recent History Also Bought

Page 1 of 13

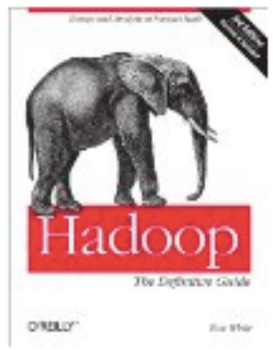


Data Analysis with Open Source Tools
› Philipp K. Janert
(35)
Paperback
$24.22 Prime

Python for Data Analysis
› Wes McKinney
(34)
Paperback
$24.68 Prime



Machine Learning for Hackers
› Drew Conway
(22)
Paperback
$33.66 Prime



Hadoop: The Definitive Guide
› Tom White
(36)
Paperback
$29.99 Prime

Bose QuietComfort 20i Acoustic Noise...
(144)

Bose IE2 Audio Headphones
(1,153)
$99.95

Bose® MIE2i Mobile Headset
(632)
$129.95

$50.00

3. Amazon Gift Card - E-mail - Tha You (Note)
Amazon
$50.00

4. Amazon Gift Card Upload Your Photo - Gift for You
Amazon
$50.00

5. Amazon.com Gift Cards - E-mail

Does Anne Hathaway News Drive Berkshire Hathaway's Stock?

ALEXIS C. MADRIGAL | MAR 18 2011, 10:50 AM ET

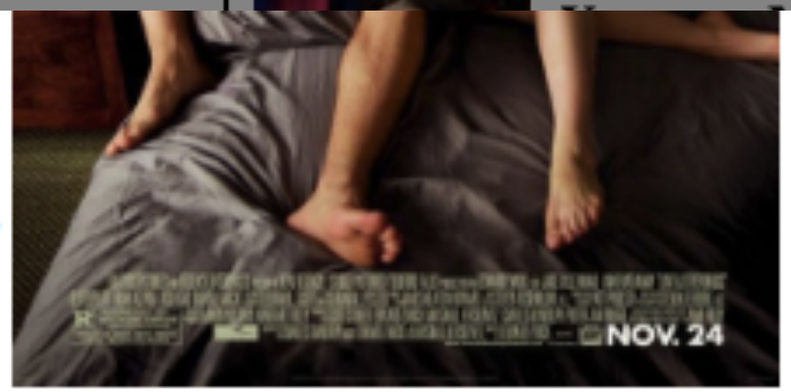
Given the awesome correlating powers of today's stock trading computers, the idea may not be as far-fetched as you think.

A couple weeks ago, Huffington Post blogger Dan Mirvish noted a funny trend: when Anne Hathaway was in the news, Warren Buffett's Berkshire Hathaway's shares went up. He pointed to six dates going back to 2008 to show the correlation. Mirvish then suggested a mechanism to explain the trend: "automated, robotic trading programming are picking up the same chatter on the Internet about 'Hathaway' as the IMDb's StarMeter, and they're applying it to the stock market."

The idea seems ridiculous. But the more I thought about the strange behavior of algorithmic trading systems and the news that Twitter sentiment analysis could be used by stock market analysts and the fact that many computer programs are simply looking for tradeable correlations, I really started to

Does Anne Hathaway News Drive Berkshire Hathaway's Stock?

ALEXIS C.



On the Friday before the Oscars, Berkshire shares rose a whopping 2.02%. And on the Monday just after the Academy Awards, they rose again, this time 2.94%. But it's not just an Oscar bounce, or something Warren Buffett may have said in the newspaper, or even necessarily something the company itself is doing (i.e. rumors afoot to buy Costco). Just look back at some other landmark dates in Anne Hathaway's still young career:

Oct. 3, 2008 - *Rachel Getting Married* opens: BRK.A up .44%
Jan. 5, 2009 - *Bride Wars* opens: BRK.A up 2.61%
Feb. 8, 2010 - *Valentine's Day* opens: BRK.A up 1.01%
March 5, 2010 - *Alice in Wonderland* opens: BRK.A up .74%
Nov. 24, 2010 - *Love and Other Drugs* opens: BRK.A up 1.62%
Nov. 29, 2010 - Anne announced as co-host of the Oscars: BRK.A up .25%

My guess is that all those automated, robotic trading programming are picking up the same chatter on the internet about "Hathaway" as the IMDb's StarMeter, and they're applying it to the stock market. Of course, this isn't necessarily bad news for the investor. After all, can you imagine what might have happened to Berkshire stock if Warren Buffett had appeared nude in *Love and Other Drugs* rather than Anne Hathway? Perhaps it's best if we don't think about it.

of algorithmic trading systems and the news that Twitter sentiment analysis could be used by stock market analysts and the fact that many computer programs are simply looking for tradeable correlations, I really started to

Faces

Search

Faces

47 Faces


Dawn Mercurio


Chris Callison-Burch


Vanessa Callison-Burch


Archie Burch


Reed Callison


Alexa Callison-...

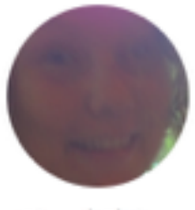

Matt Fisher


Diane Callison


Colin Bannard


Eric Minor


Geof


Sherri Fisher


Alex Klementiev


Sue


TJ


Juri Ganitkevit...


Iris Fisher


Ben Van Durme


Eilidh Macdon...


Moni


Keith Hall


Ondrej Bojar


Sara


Xuchen Yao


Megan Shanho...

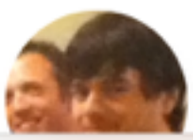

Frank Ferraro

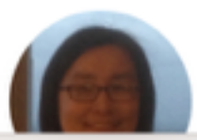

Gabor


Omar F. Zaidan


Ann Irvine


Ken Heafield





Suggested Faces

Reed Callison

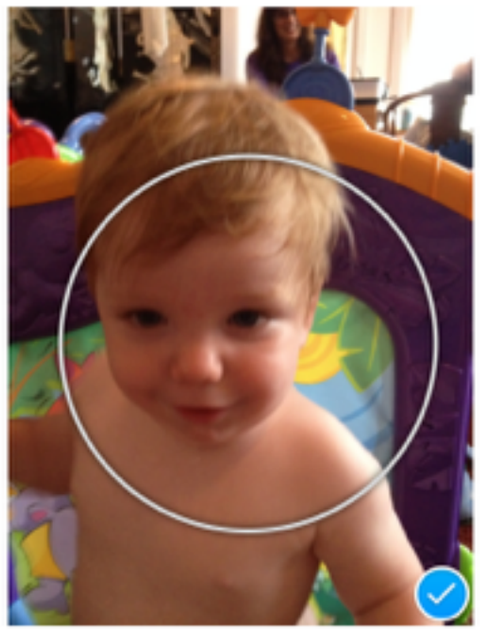
f Reed Callison is not highlighted in the photo, click to reject it.

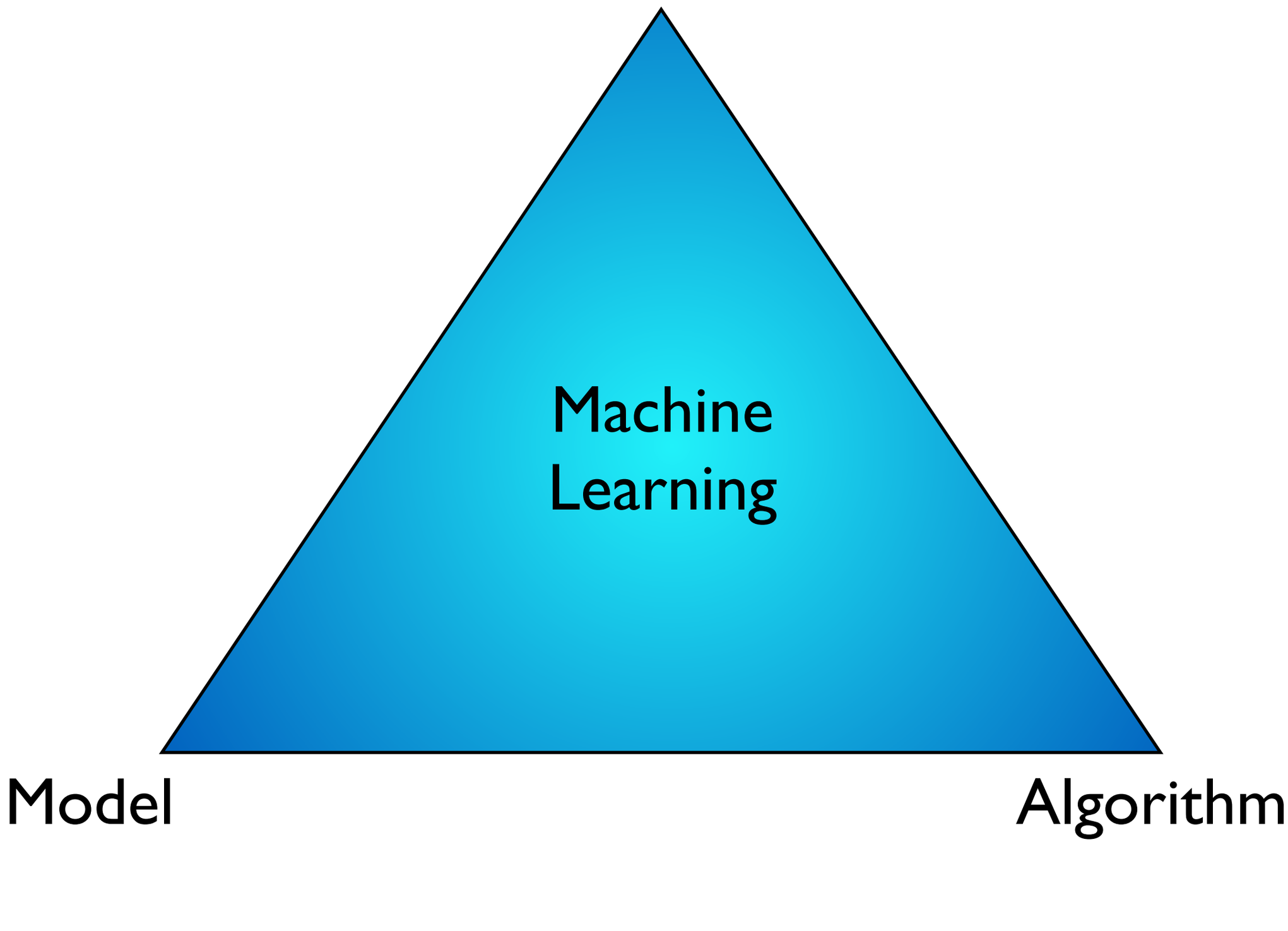
Finish Later | Add and Continue








Data
Machine
Learning
Model
Algorithm

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!

Yelp Dataset Challenge

Yelp is proud to introduce a deep dataset for research-minded academics from our wealth of data. If you've used our Academic Dataset and want something richer to train your models on and use in publications, this is it. Tired of using the same standard datasets? Want some real-world relevance in your research project? This data is for you!

The Challenge Dataset includes data from **Phoenix**, **Las Vegas**, **Madison**, **Waterloo** and **Edinburgh**:

- **42,153** businesses
- **320,002** business attributes
- **31,617** check-in sets
- **252,898** users
- **955,999** edge social graph
- **403,210** tips
- **1,125,458** reviews

Get the Data

The Challenge

Not only would we like to give you our data, we'd also like to announce the fourth round of the **Yelp Dataset Challenge**. We challenge you to use this data in an innovative way and break ground in research.

How well can you guess a review's rating from its text alone? Can you

The Awards

If you are a student and come up with an appealing project, you'll have the opportunity to win one of ten Yelp Dataset Challenge awards for $5,000. Yes, that's $5,000 for showing us how you use our data in insightful, unique, and compelling ways.

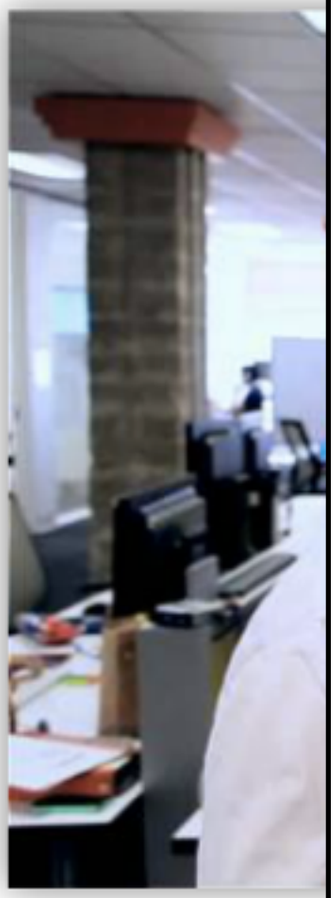
yelp

Find pizza, pub, Fox & Hound | Near London | Sign Up

Home | About Me | Write a Review | Find Friends | Messages | Talk | Events | Log In

Yelp Dataset Challenge

Yelp is proud to introduce a deep dataset for research-minded academics from our wealth of data. If you've used our Academic D… Tired of using the same … or you!



…ludes data **Madison,**

The Challenge

Not only would we like to give you our data, we'd also like to announce the fourth round of the **Yelp Dataset Challenge**. We challenge you to use this data in an innovative way and break ground in research.

How well can you guess a review's rating from its text alone? Can you take all of the reviews of a business and predict when it will be the most busy, or when the business is open? Can you predict if a business is good for kids? Has Wi-Fi? Has Parking? What makes a review useful, funny, or cool? Can you figure out which business a user is likely to review next? How much of a business's success is really just location, location, location? What businesses deserve their own subcategory (i.e., Szechuan or Hunan versus just "Chinese restaurants"), and can you learn this from the review text? What makes a tip useful? What are the differences between the cities in the dataset? There is a myriad of deep, machine learning questions to tackle with this rich dataset.

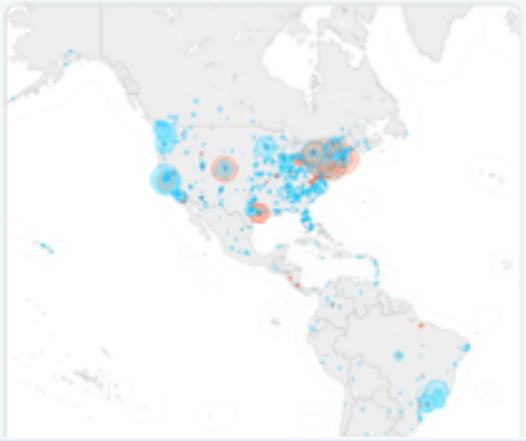
The Challenge

Not only would we like to give you our data, we'd also like to announce the fourth round of the **Yelp Dataset Challenge**. We challenge you to use this data in an innovative way and break ground in research.

How well can you guess a review's rating from its text alone? Can you

The Awards

If you are a student and come up with an appealing project, you'll have the opportunity to win one of ten Yelp Dataset Challenge awards for $5,000. Yes, that's $5,000 for showing us how you use our data in insightful, unique, and compelling ways.



Welcome to Kaggle Datasets

The best place to discover and seamlessly analyze publicly available data.

Dig In

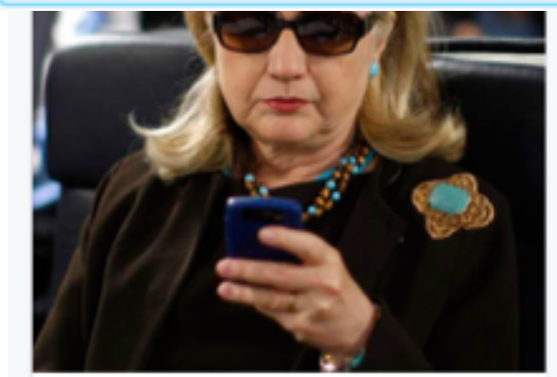
Explore a dataset with our in-browser analytics tool, Kaggle Scripts. You can also download it in an easy to read format.

Build

Create your data science portfolio. Publish insights and code with Kaggle Scripts and it will be saved to your profile.

Connect

Engage with other data scientists. Share feedback on other Kagglers' scripts, or ask a question in a dataset's forum.



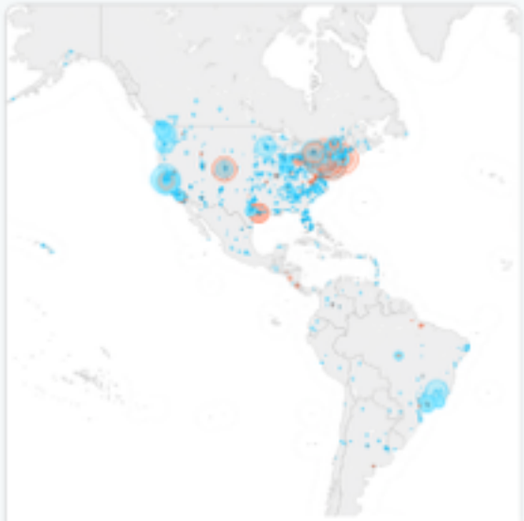
Hillary Clinton's Emails
362 Scripts · 20 Topics
22

Health Insurance Market...
25 Scripts · 0 Topics
12

First GOP Debate Twitter ...
43 Scripts · 3 Topics
7

US Baby Names
106 Scripts · 5 Topics
17

Amazon Fine Food
27 Scripts · 2 Topics



Meta Kaggle
222 Scripts · 27 Topics
17

World Food Facts
50 Scripts · 1 Topic
10

Ocean Ship Logbooks (175...
117 Scripts · 24 Topics
9

US Dept of Education: Col...
310 Scripts · 17 Topics
19

2015 Notebook UX
5 Scripts · 0 Topics

Iris
110 Scripts · 4 Topics
7

May 2015 Reddit Comments
775 Scripts · 32 Topics
7

SF Salaries
88 Scripts · 6 Topics
16

2013 American Communi...
736 Scripts · 48 Topics
10

Twitter US Airline
32 Scripts · 3 Topics

Hillary Clinton's Emails
362 Scripts · 20 Topics
22

Health Insurance Market...
25 Scripts · 0 Topics
12

First GOP Debate Twitter ...
43 Scripts · 3 Topics
7

US Baby Names
106 Scripts · 5 Topics
17

Amazon Fine Food
27 Scripts · 2 Topics

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

Example classification tasks

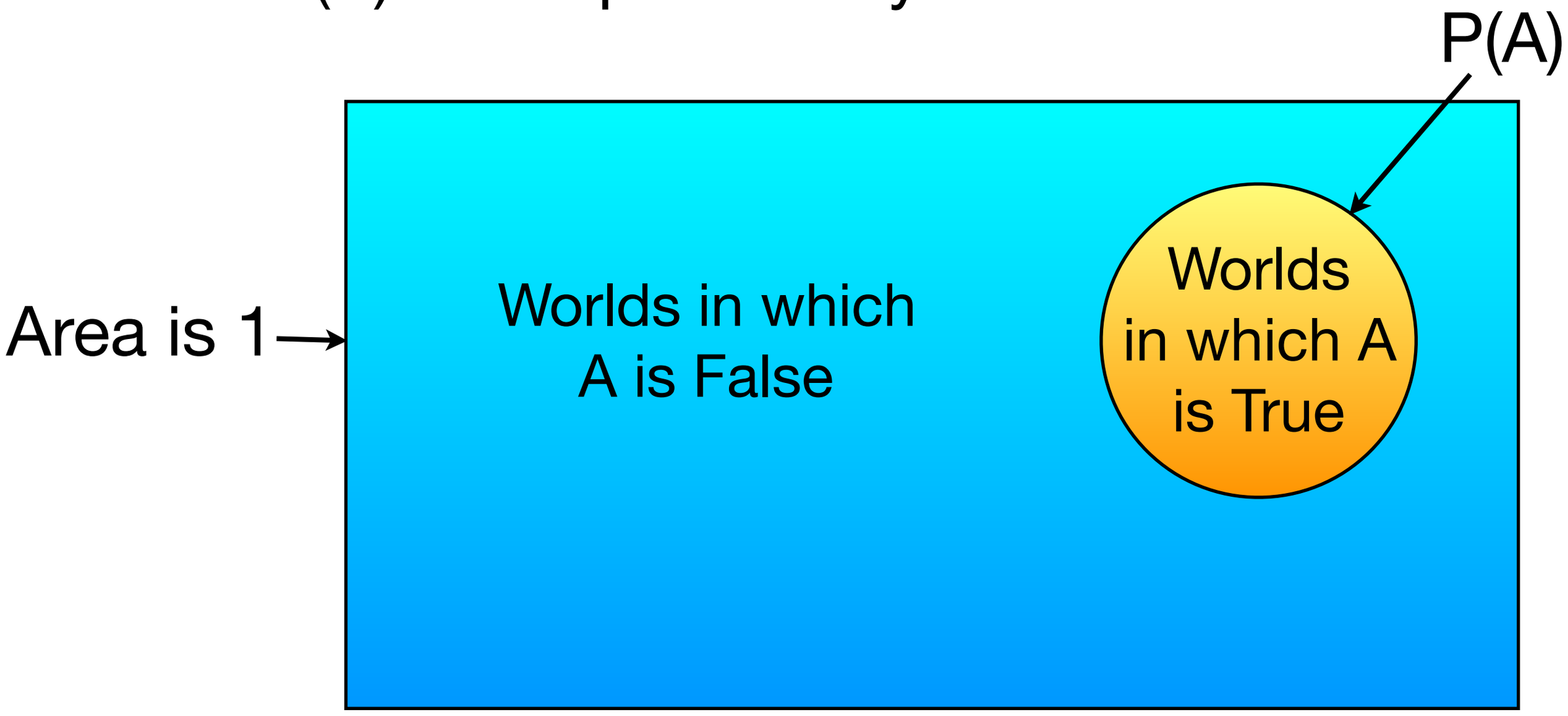
- Spam v. not Spam
- Face detection
- Language Identification

Naive Bayes

- First used for Spam filters in the 1990s
- Math is just counting, dividing and multiplying
- No calculus

Probability

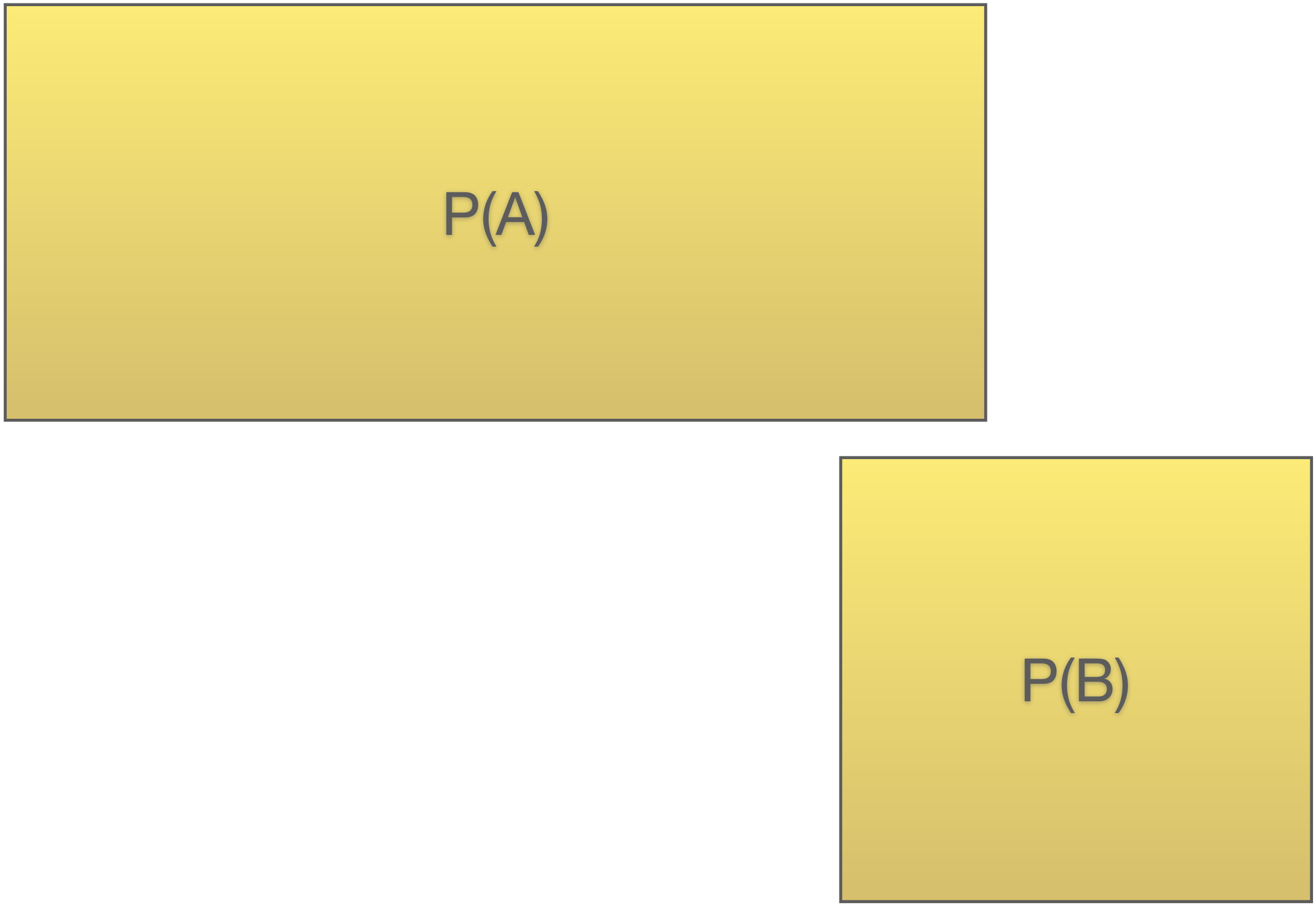
P(A) is the probability that A is true



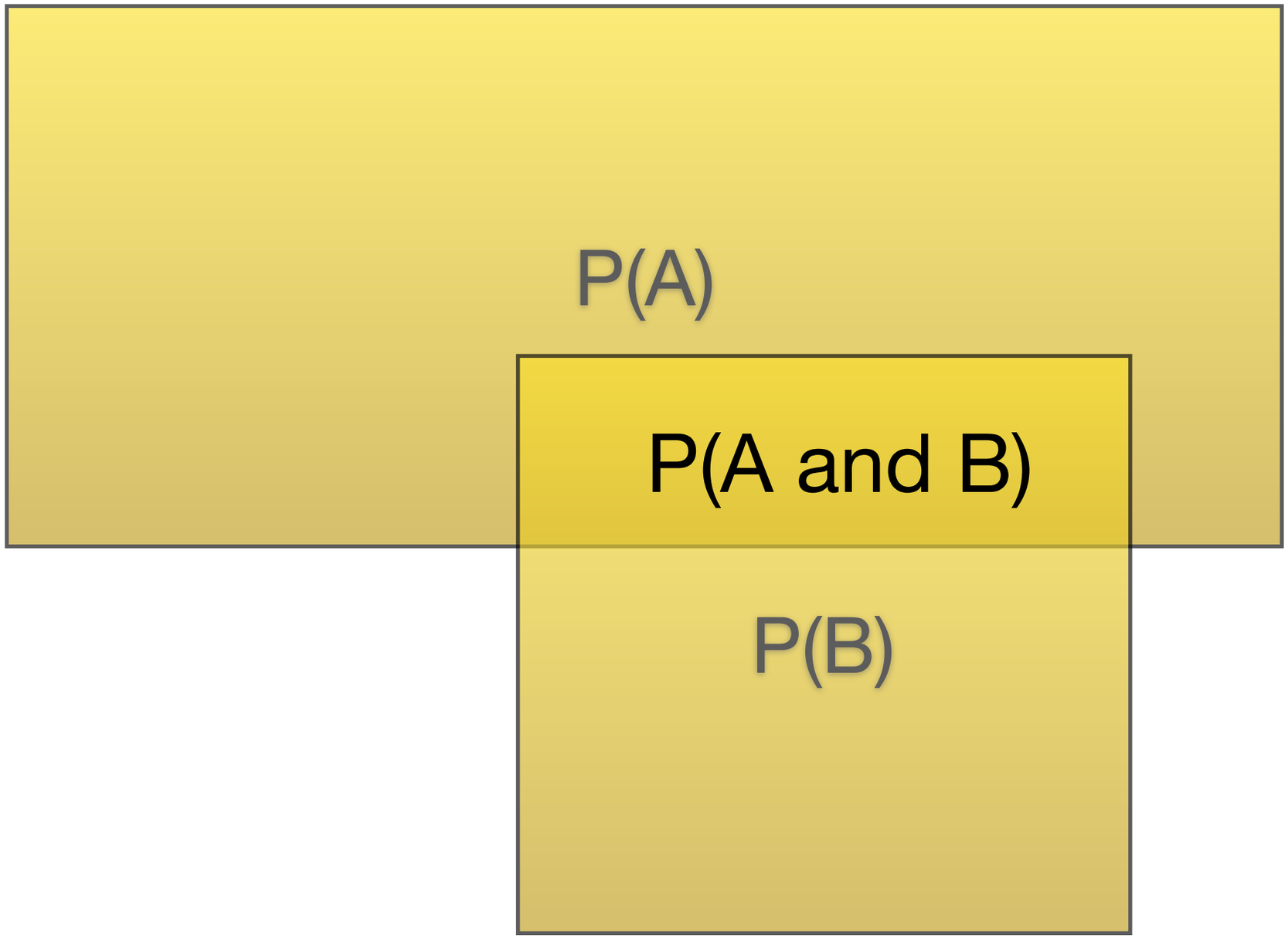
Axioms of Probability

- P(True) = 1
- P(False) = 0
- $0 \leq P(A) \leq 1$
- P(A or B) = P(A) + P(B) - P(A and B)

P(A or B) = P(A) + P(B) - P(A and B)



P(A or B) = P(A) + P(B) - P(A and B)



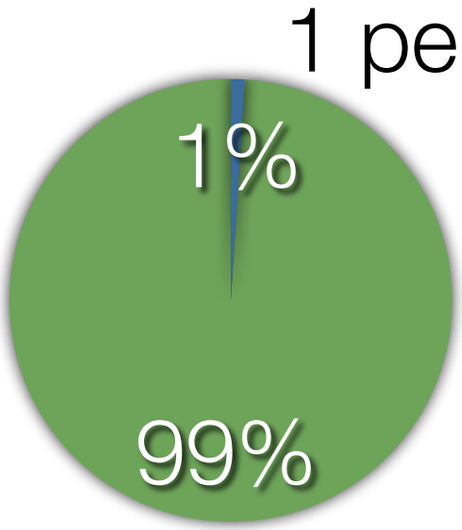
Bayes Law

$$P(B \mid A) = P(A \mid B) * P(B) / P(A)$$

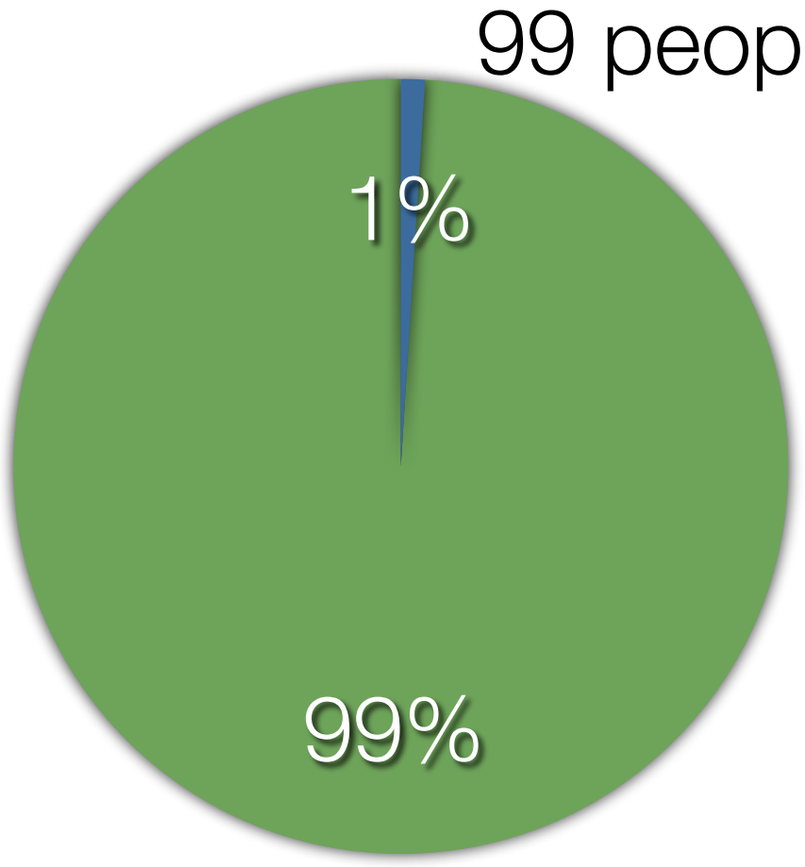
Example

- In a group of 10,000 people
- 1% of them have a rare disease
- There is a test that is 99% effective
  - 99% of sick patients test positive
  - 99% of healthy patients test negative
- Given a positive result, what is the probability that the patient are sick?

Sick population (100 people)



Healthy population (9900 people)



False positives
True positives

False negative
True negatives

Disease diagnosis

- 99 sick patients test positive, 99 healthy patients test positive
- Give a positive test result, there is a 50% chance that a patient has the test result

Bayesian disease

$$p(\text{sick} \mid \text{test\_pos}) = \frac{p(\text{test\_pos} \mid \text{sick})\, p(\text{sick})}{p(\text{test\_pos})} = \frac{99}{198} = \frac{1}{2}$$

99/100 * 1/100

99/10,000 + 99/10,000 = 198/10,000

(true positives + false positives)

We know the effectiveness of test (probability of testing positive given being sick), and the prior probability of being sick.

Machine learning uses these insights to learn how to classify data

Building a classifier

- Your first programming assignment will be to build a classifier from labeled training data
- Wednesday: in-class Python bootcamp
- Friday: Learn how to use a Python machine learning package called scikit learn, which you will use in the upcoming HW

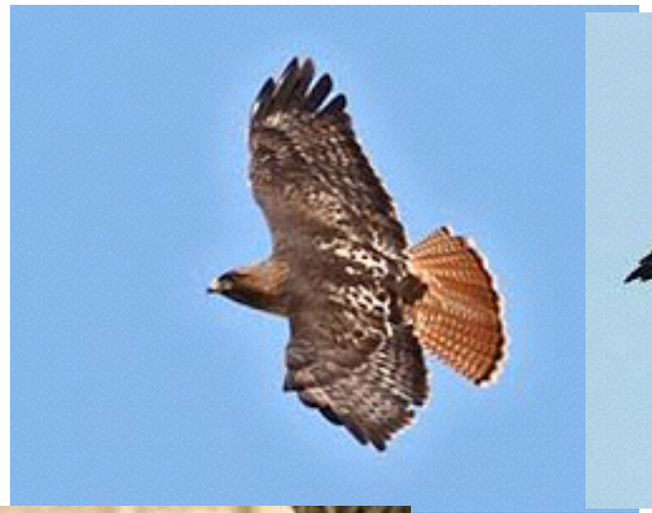
Labeled training data

Pretty awful - very soft and commercial. Confected.	★
Thin and completely uninspiring.	★
Good Syrah character, fruit-driven but not to the point of undrinkability. Pleasant. Scrapes	★★
Very classy, pure, blackberry and apple fruit. Demanding but ripe tannins, very succulent. Really good Dolcetto.	★★★
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.	★★★★
An absolute star that could even benefit from another year or two. Tremendous weight, and concentrated minerality but all in balance. Fantastic. Top	★★★★★

Labeled training data

"cockatoo"

"red tailed hawk"



Previous programming assignments sequence

- Harvest Tweets that mention a company or multiple companies
- Design a MTurk HIT to have Turkers categorize the sentiment (neutral/+/-)
- Automatically grade the Turkers and analyze their quality / trustworthiness
- **Train a machine learning classifier based on the data that you collected**

Gun Violence Database

- Designed to have social impact
- Create a structured database about all reports of gun violence in local newspapers in the USA
- In consultation with a Professor of Epidemiology who is an affiliate of the Firearm and Injury Center at Penn
- **Train a text classifier to predict whether an article describes gun violence or not**

Why is this important?

- Congress prohibited agencies like the CDC and the NIH, from offering grants to study anything that could be used to promote gun control
- In 2012, the CDC spent only $100,000 of its $5.6 billion budget on gun injury research

The Congressman Who Restricted Gun Violence Research Has Regrets

Rep. Jay Dickey (R-Ark.) authored the controversial 1996 amendment that remains in place. He wishes Congress would change it.

10/06/2015 08:03 am ET | **Updated** Oct 06, 2015

Sam Stein
Senior Politics Editor, The Huffington Post

THE WASHINGTON POST VIA GETTY IMAGES

WASHINGTON -- Looking back, nearly 20 years later, Jay Dickey is apologetic.

He is gone from Congress, giving him space to reflect on his namesake amendment that, to this day, continues to define the rigid politics of gun policy. When he helped pass a restriction of federal funding for gun violence research in 1996, the goal wasn't to be so suffocating, he insisted. But the measure was just that, dampening federal research for years and discouraging researchers from entering the field.

Well over half a million people have died by firearms since 1996, when the ban on gun violence research was enacted. The Dickey Amendment was supposed to tamp down funding for what the NRA and other critics claimed was anti-gun advocacy research by the CDC's National Center for Injury Prevention. In effect, it stopped federal gun violence research almost entirely.

"Gun violence is probably the only thing in this country that kills so many people, injures so many people, that we are not actually doing sufficient research on," Dr. Alice Chen, the executive director of Doctors for America

Labeled training data

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."

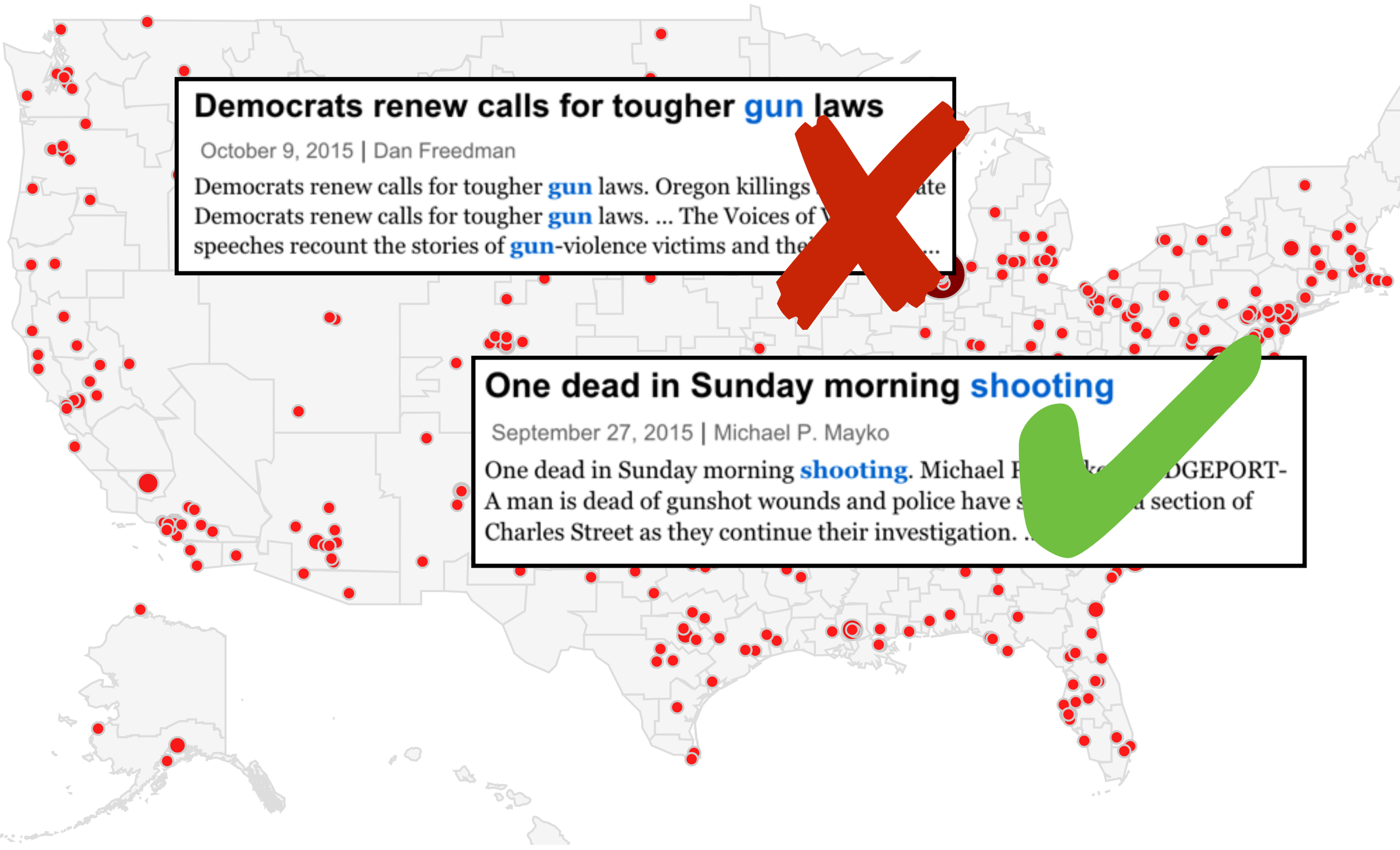
Meanwhile, on a playground at a school in Milwaukee last Wednesday, 10-year-old Sierra Guyton was caught in the crossfire of a **shootout**. She is in stable condition, but as of yesterday, she is not responsive. A **fund** has been created for her family, and a **rally** was held in her honor on Memorial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

Labeled 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/	🔫

Large Scale Classification



Machine Learning + Crowdsourcing

- Validate your classifier's predictions using workers from CrowdFlower
- Use the workers to perform information extraction tasks that are not currently reliably done using ML alone

Information Extraction

Bridgeport, Connecticut -- Christopher Pettway, the city's ninth homicide victim this year, was recalled Wednesday as ""a man with morals and values'' by a family friend. Ernest Newton III, the former state senator, said he knew Pettway's family well, "
They are religious people.'' Pettway died in the hospital as a re
near Reservoir and Trumbull avenues at about 2:30 p.m. Tuesc
Pettway was only 26 years old when he passed. ""I grew up w
Newton said. He had not spok stop
the family friend said. The Pett arges
talked to the dead man's cous ery ha
another gunshot death. We ha n't b
Haven.'' Kaempffer said the in g. So
fired using a 22 caliber pistol. Beyond that, Kaempffer said onl
very strong leads,"" but did not release more details about what led up to the shooting or what

Clear this highlight

Age

Race

Victim #1

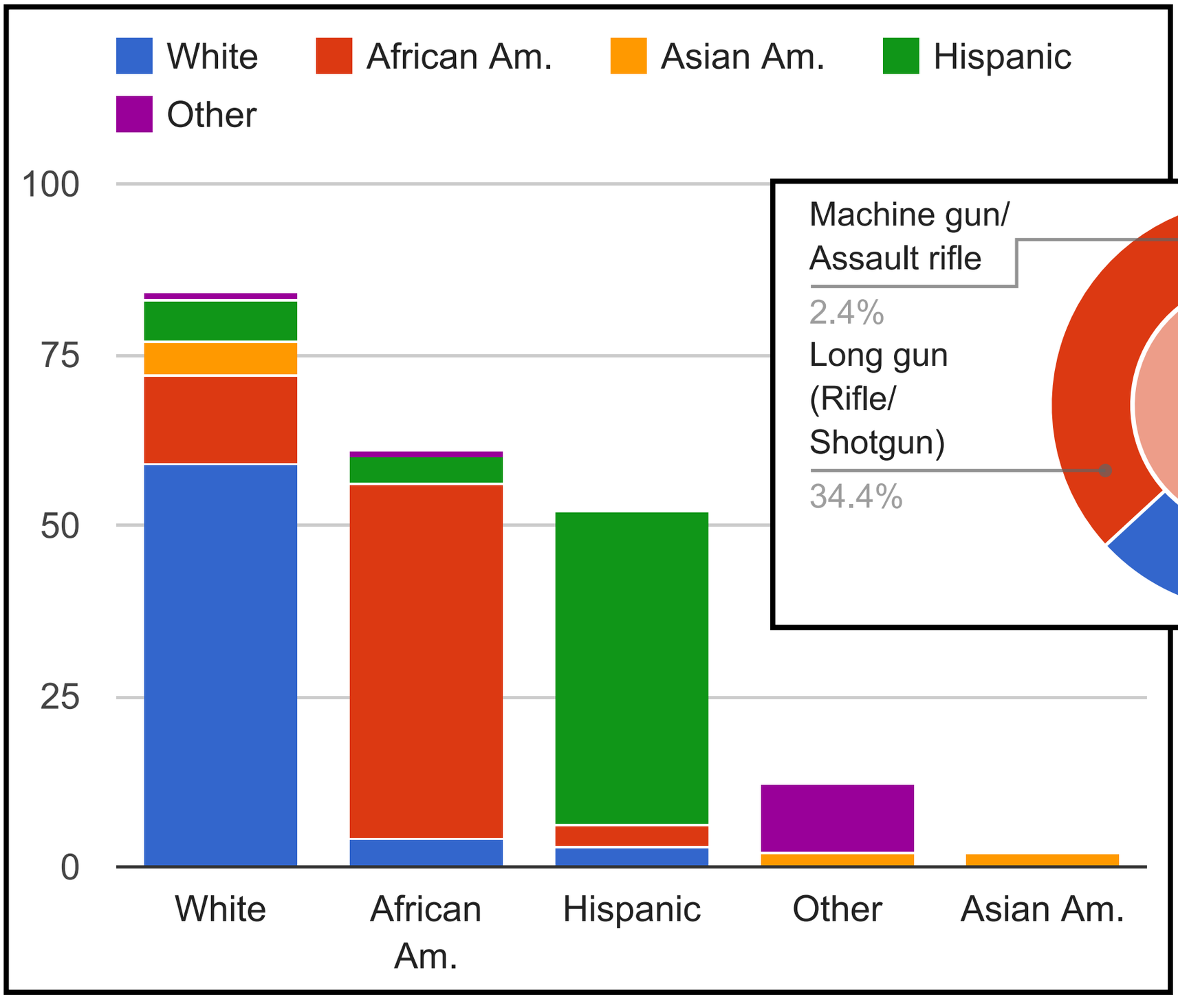
Name: Christopher Pettway

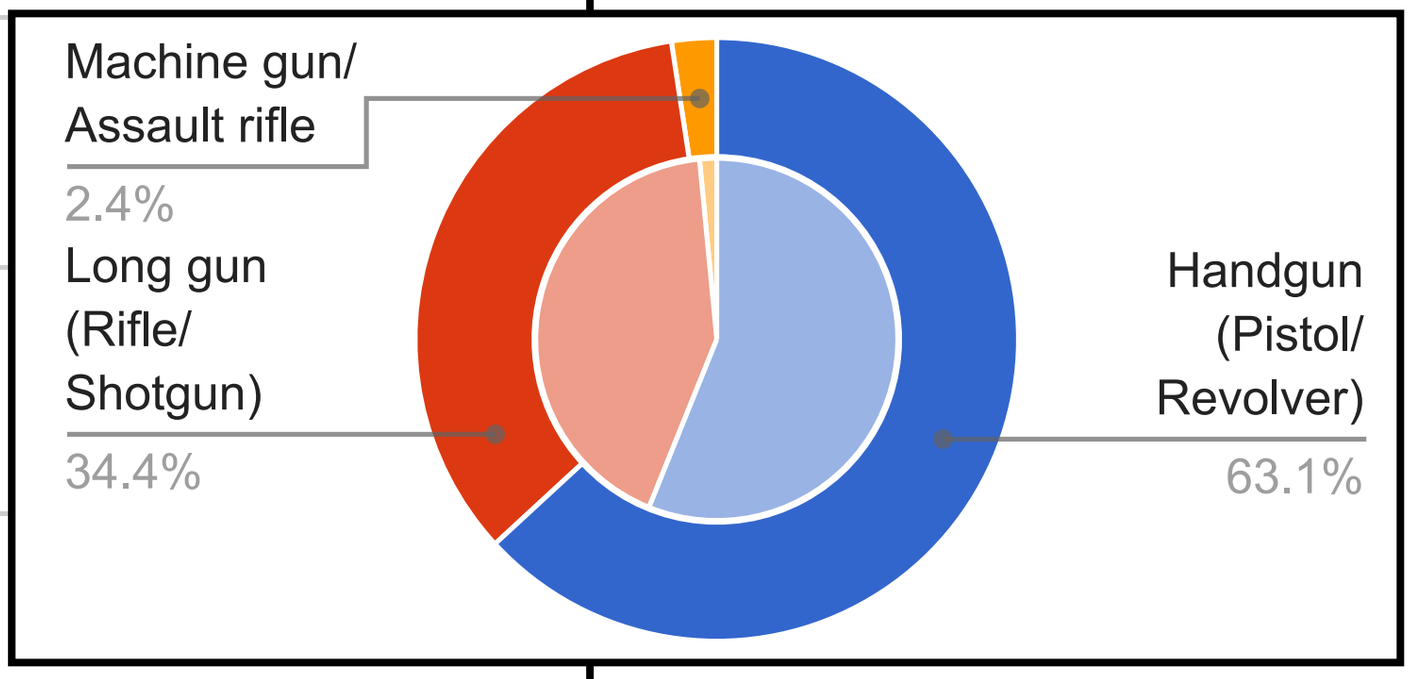
Age: 26 years old

Gender:

Race:

Analyze the Data





For Wednesday:

Be sure you have IPython Notebook installed on your computer. We will do an in-class Python bootcamp (counts towards your final grade).