Intro to Machine Learning

Crowdsourcing and Human Computation
Lecture 5

Instructor: Chris Callison-Burch
TA: Ellie Pavlick

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

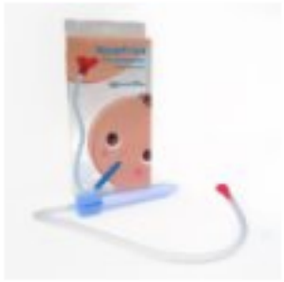
List Price: ~~$54.99~~

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

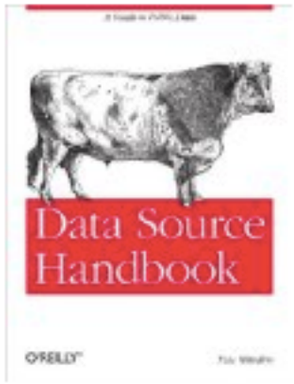
Because your Wish List includes...

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

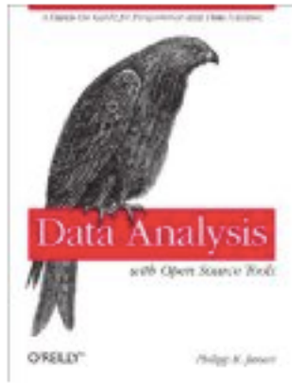


Data Source Handbook
› Pete Warden
Paperback
★★☆☆☆ (8)
~~$29.99~~ $20.78

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

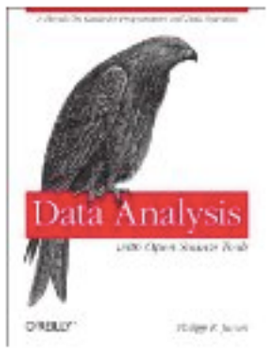
2. ASICS Women's GEL-Noosa T
Running Shoe
ASICS

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

4. ASICS Women's Gel-Kayano 1

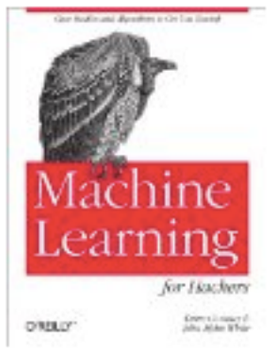
Continue Shopping: Customers Who Bought Items in Your Recent History Also Bought

Page 1 of 13

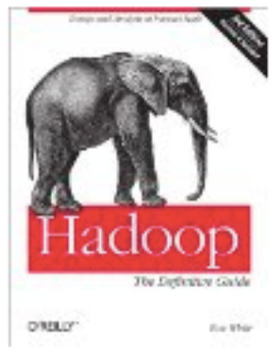


Data Analysis with Open Source Tools
› Philipp K. Janert
(35)
Paperback
$24.22 Prime

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

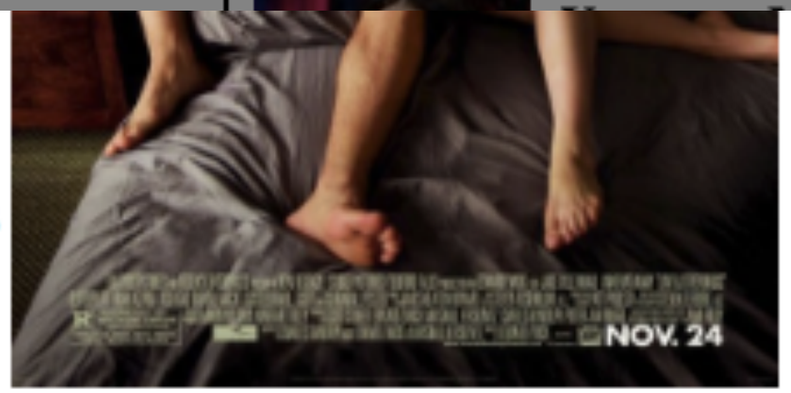
Given the awesome correlating powers of today's stock trading computers, the idea may not be as far-fetched as you think.

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

iPhoto
LIBRARY
Events
Photos
Faces
Places
RECENT
Oct 3, 2009
Last 12 Months
Last Import
Flagged
Trash
WEB
Photo Stream
callison-burch
DEVICES
ccb's iPhone
EVENTS
ALBUMS
Last 12 Months
2nd floor bathroom
3029 Elm Ave
Alex Klementiev
Alexa Callison-Burch
Ann Irvine
Archie Burch
Charley Chan
Chris Callison-Burch
Chris Landers
Colin Bannard
Danielle Matthews
Dave Hawkey
Dawn Mercurio
Diane Callison
Search
Zoom
Slideshow
Find Faces

LIBRARY
- Events
- Photos
- Faces
- Places

RECENT
- Oct 3, 2009
- Last 12 Months
- Last Import
- Flagged
- Trash

WEB
- Photo Stream
- callison-burch

DEVICES
- ccb's iPhone

EVENTS

ALBUMS
- Last 12 Months
- 2nd floor bathroom
- 3029 Elm Ave

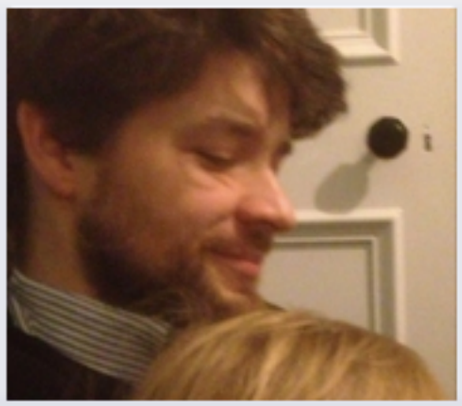

Chris Callison-Burch


Chris Landers


Colin Bannard


Courtney Napoles


Danielle Matthews

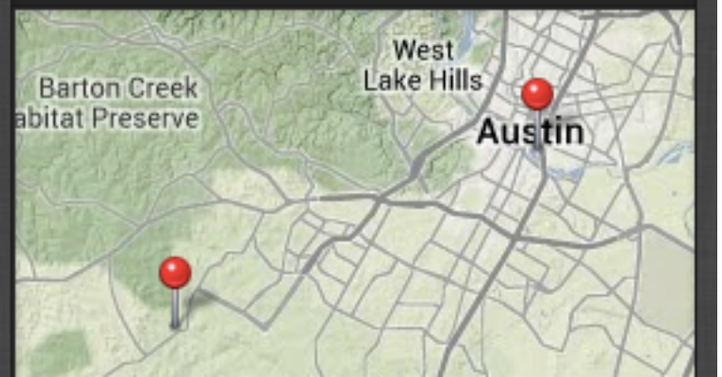

Dave Hawkey


Colin Bannard

Colin Bannard
bannard@mail.utexas.edu

12/6/12 - 12/8/12
3 photos
34 unconfirmed matches

Texas



iPhoto

Colin Bannard 37 photos

Photos


Colin Bannard


Colin Bannard


Colin Bannard

Unconfirmed Faces


click to confirm


click to confirm


click to confirm

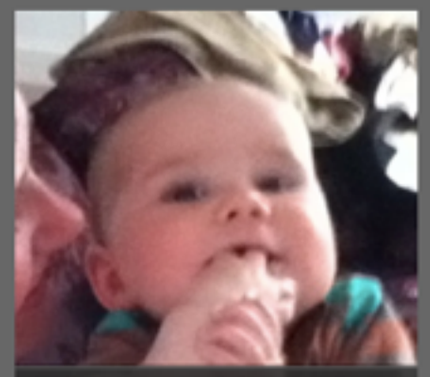

click to confirm


click to confirm

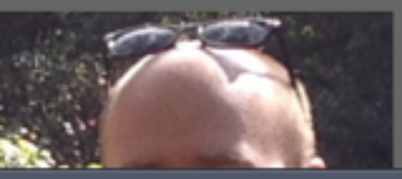

click to confirm

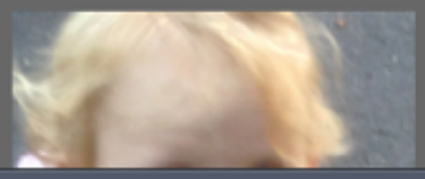

click to confirm

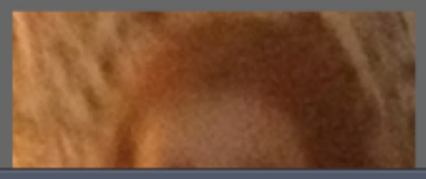

click to confirm

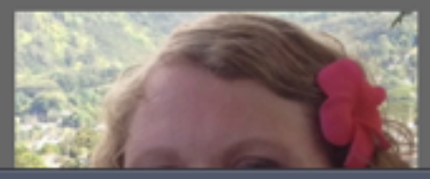

click to confirm

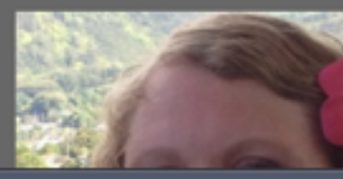

click to confirm





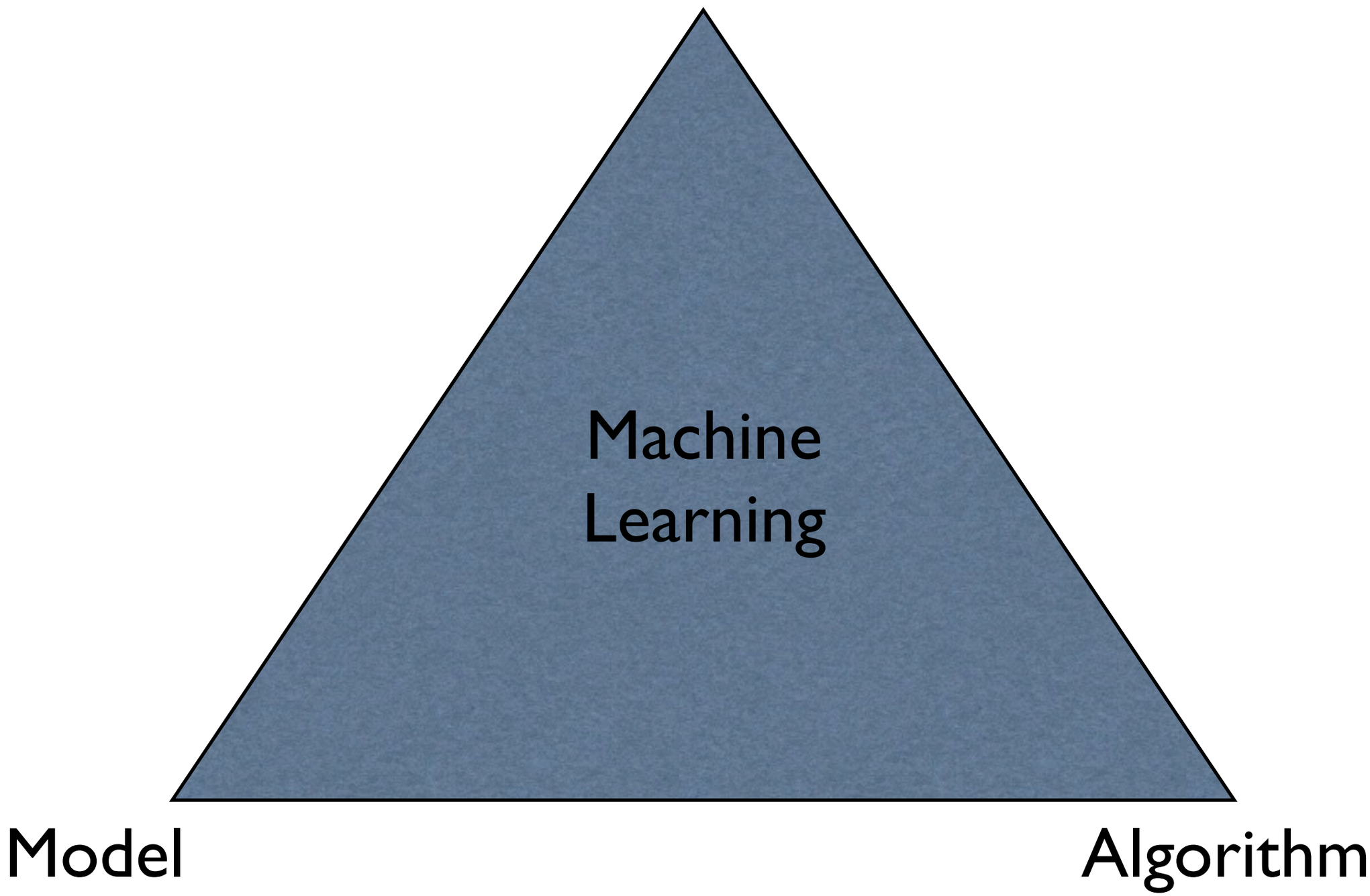






Click to confirm or Option-click to reject photos

Cancel


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://delicious.com/pskomoroch/ dataset
- Some is 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.

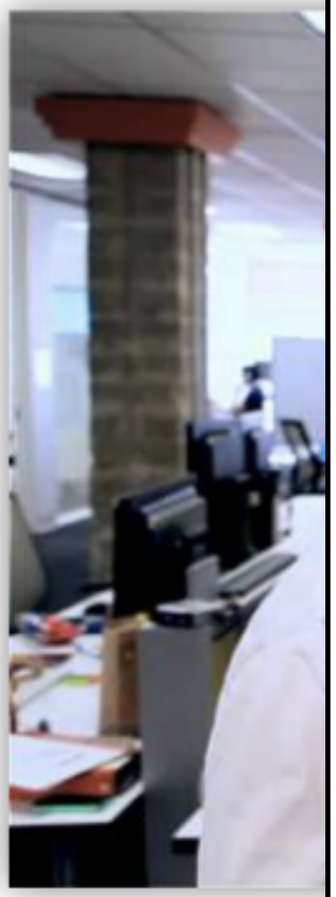
yelp

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.

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

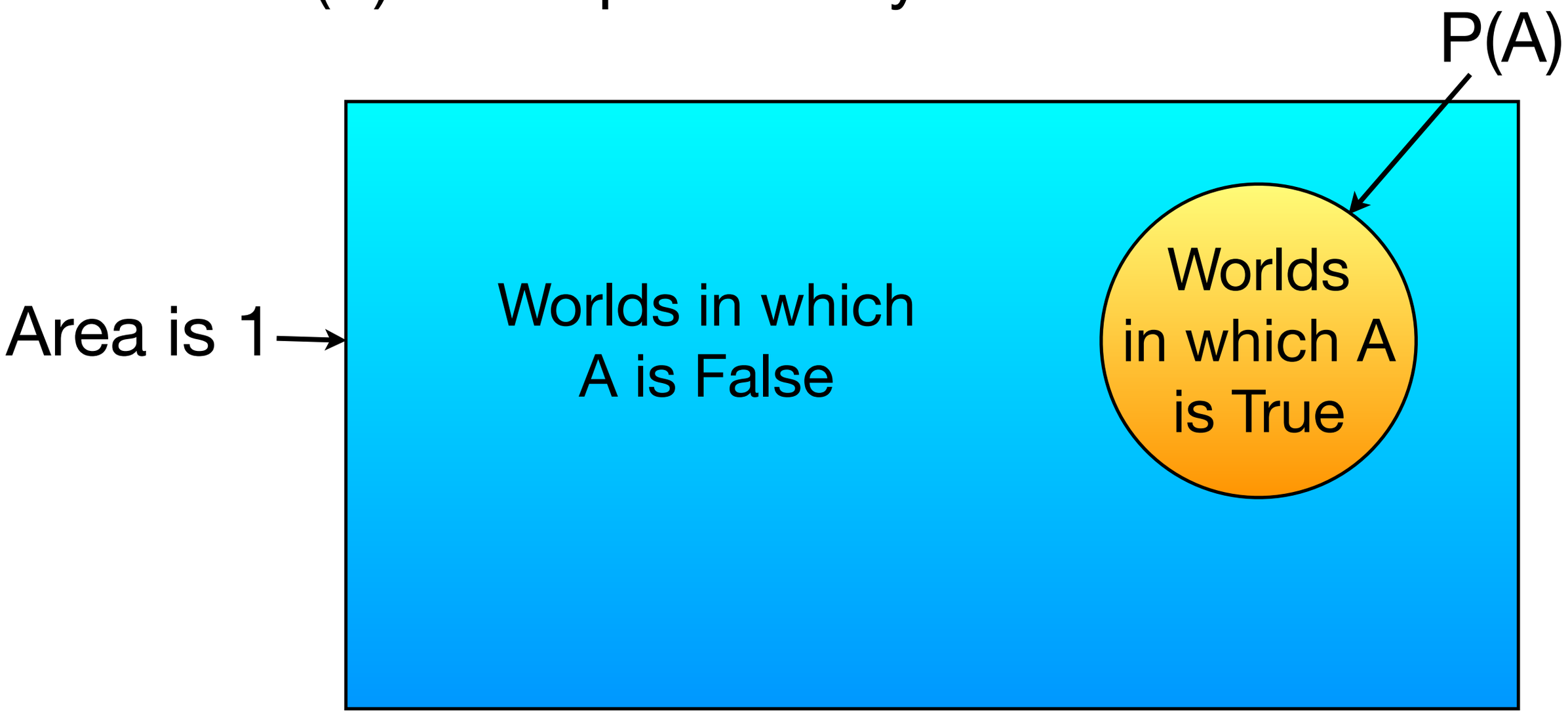
- Spam v. not Spam
- Face detection
- Language Identification

Naive Bayes

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

Probability

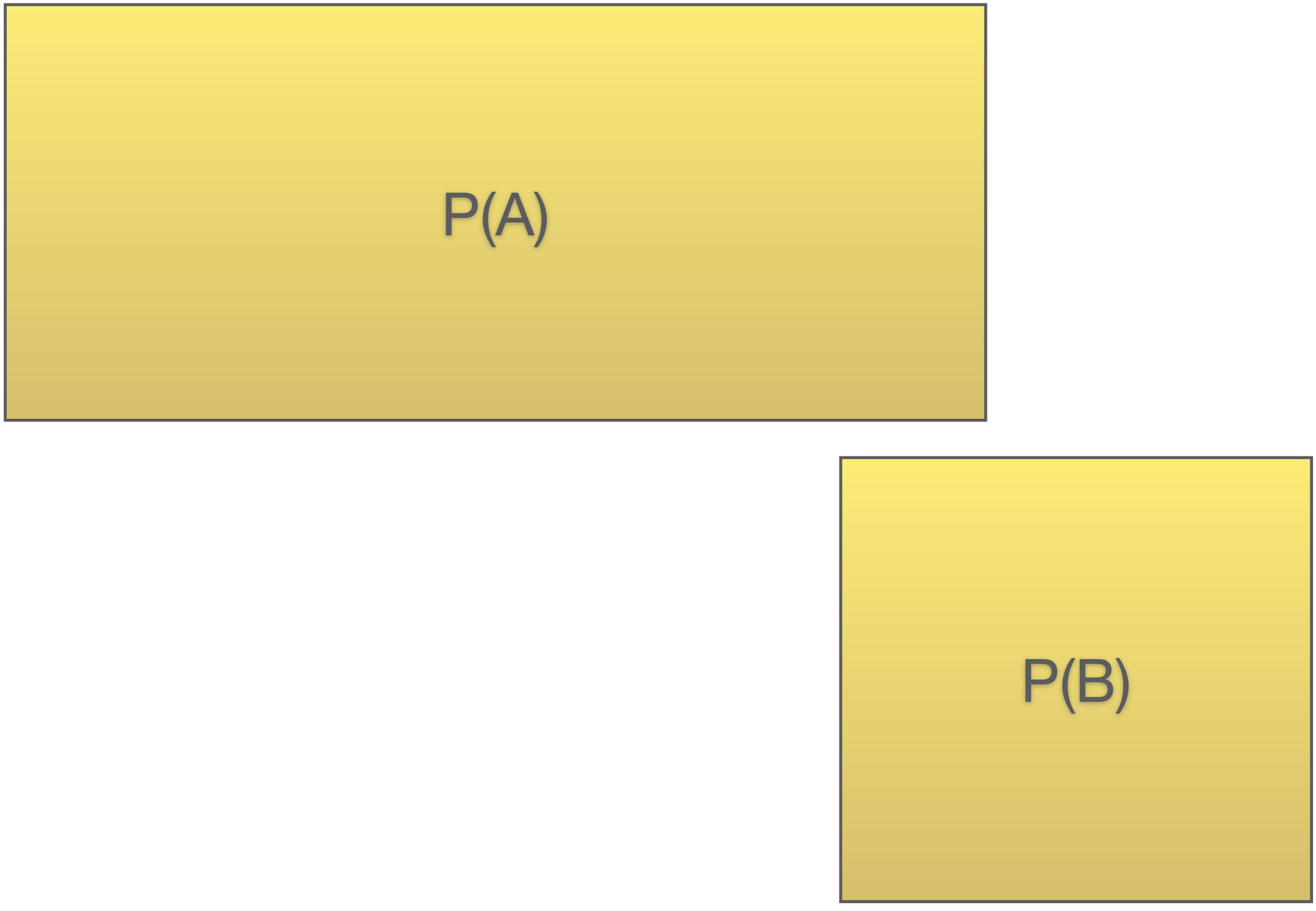
P(A) is the probability that A is true



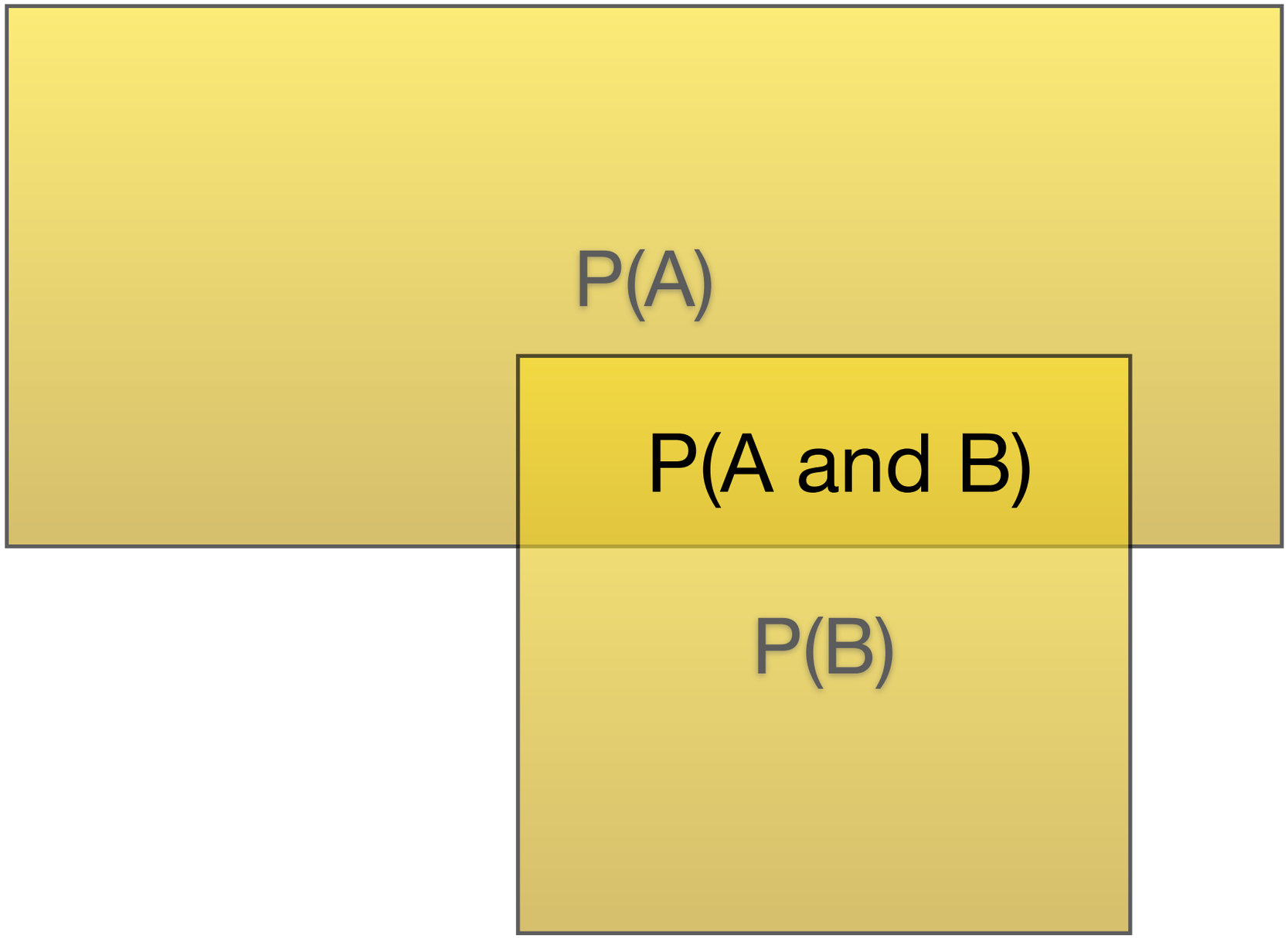
Axioms of Probability

- 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)



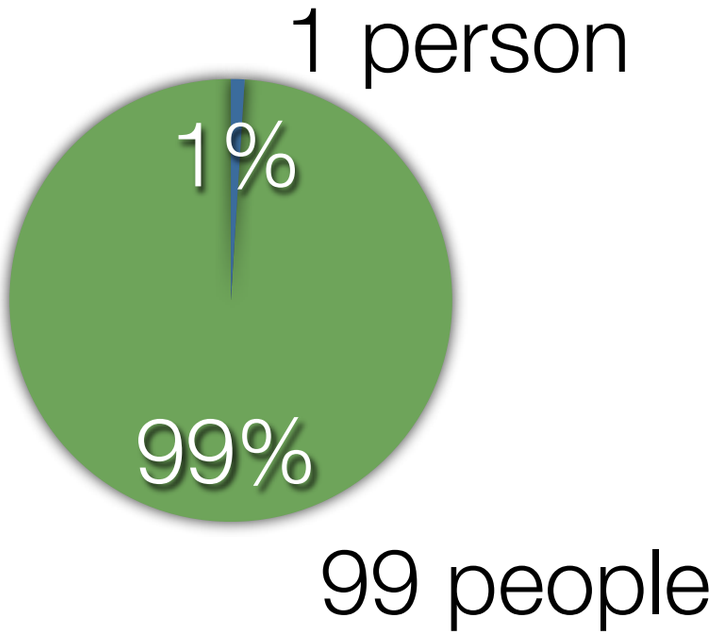
Bayes Law

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

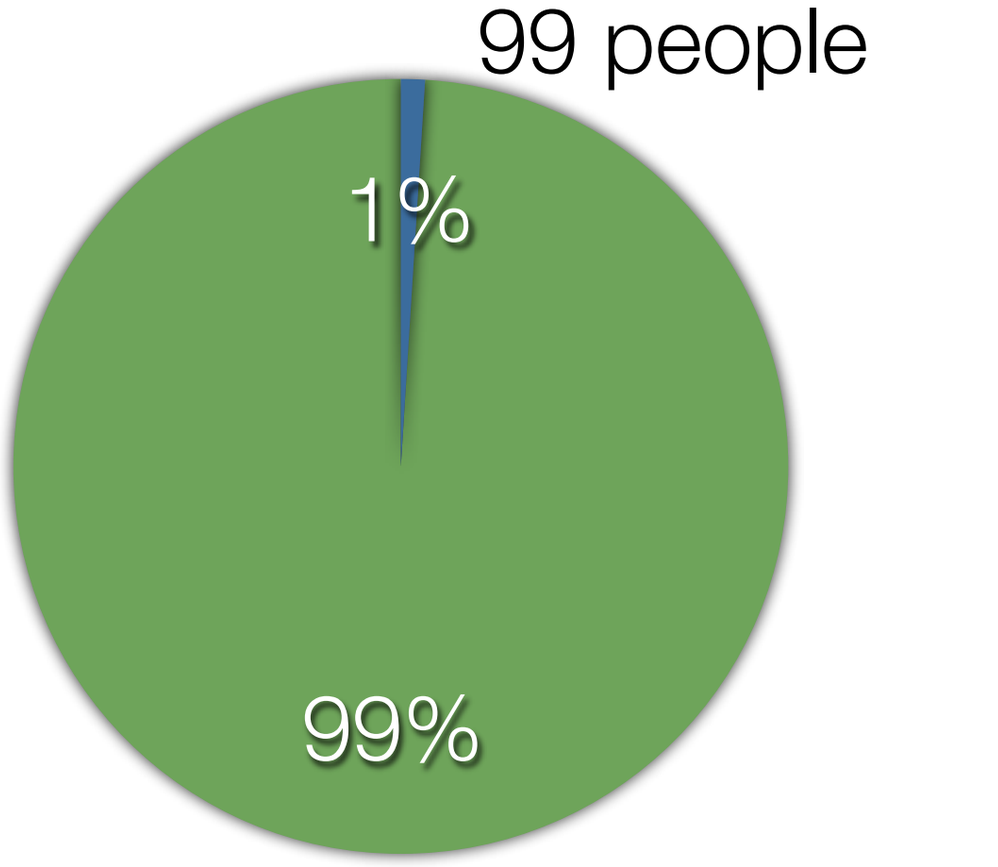
Example

- 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

$$p(\text{sick} \mid \text{test\_pos}) = \frac{\overbrace{p(\text{test\_pos} \mid \text{sick})}^{99/100} \overbrace{p(\text{sick})}^{1/100}}{p(\text{test\_pos})} = \frac{99}{198} = \frac{1}{2}$$

$$99/10{,}000 + 99/10{,}000 = 198/10{,}000$$

(true positives + false positives)

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

Next time: How do we use it for machine learning?

Last year's 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

This year’s assignments

- Re-designed to have social impact
- We are going to be looking into creating 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

Sequence of assignments

- Train a text classifier to predict whether an article describes gun violence or not
  - Run your classifier over 1 million articles collected from 2000 local newspapers around the country
- Use CrowdFlower to validate whether the predictions of your classifier are correct

Sequence of assignments

- Perform quality control on the labels that you collected via crowdsourcing
- Include gold-standard controls
- Perform analysis of errors that crowd-workers made
- Automate the approval/rejection process

Sequence of assignments

- Create a structured database from the unstructured text data:
    - Who was the victim?
    - What was their age, race, gender?
    - Were they killed or injured?
    - Who was the alleged perpetrator?
    - Where did the incident take place?
    - What were the circumstances? Was alcohol involved? Was it a domestic violence incident? Was it a police shooting? Was it carried out during another crime?

Sequence of assignments

- Design an interface for Crowd Workers to extract the data
- Use natural language processing tools through the AlchemyAPI to highlight named entities (persons, places)
- Evaluate the goodness of the interface through the speed and accuracy of the annotation

Sequence of assignments

- Analysis of data
  - How much would it cost for us to manually annotate every reported instance of gun violence in the USA?
  - What sort of research questions can we answer once we have a structured database about gun violence?

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
- Guest lecture on Monday 9/29 on Gun Violence Research and Prevention, and the importance of data

Carroll authorities update details from Sunday shooting

Last updated: January 27. 2014 11:45AM - 1864 Views

Pamela Denise Peaks

Story Tools:

Email Print QR Favorite

The Carroll County Sheriff's Office has charged 47-year-old Pamela Denise Peaks after a shooting Sunday at 2472 Carrollton Pike that appears to have stemmed from a domestic dispute.

According to investigator Donnie Spangler, Pamela Denise Peaks, 47, has been charged with aggravated malicious wounding and use of a firearm in the commission of a felony.

At approximately 2 p.m. Sunday, the Carroll County Sheriff's Office received a call for a possible gunshot wound at 2472 Carrollton Pike. When deputies arrived on scene, they found James Parnell Peaks, 48, lying on the side of the road with a gunshot wound. The victim was airlifted to Wake Forest Baptist Medical Center in Winston-Salem, N.C.

"He was going in for exploratory surgery this morning. That

Any questions about the python bootcamp HW, or the videos?