30: Wrap-up

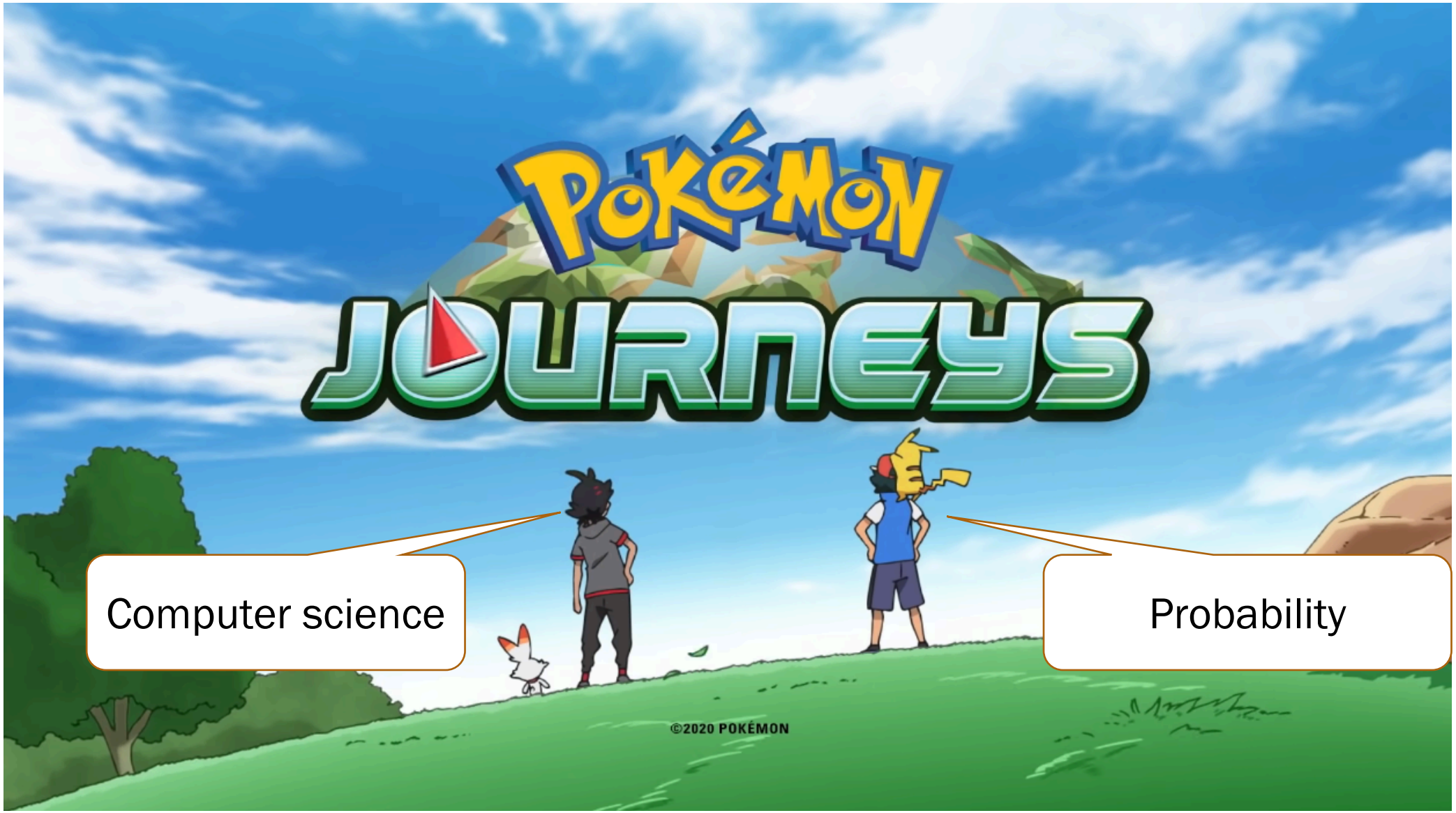
Lisa Yan and Jerry Cain
November 20, 2020

Quick slide reference

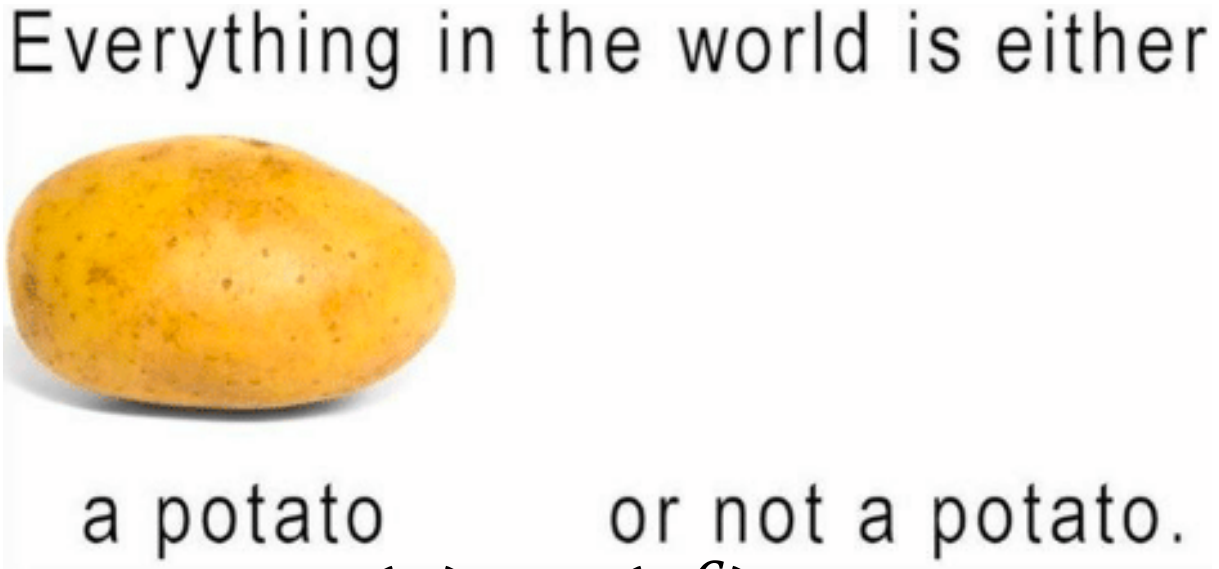
3	CS109 Wrap-Up	LIVE

What have we learned in CS109?

A wild journey



From combinatorics to probability…

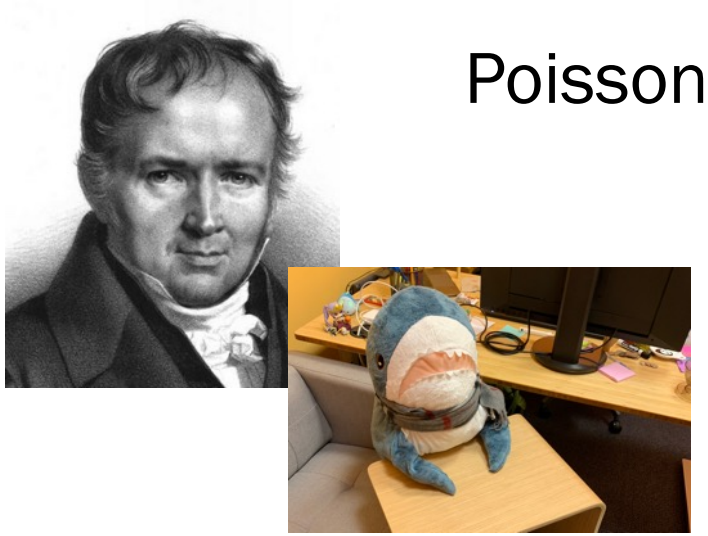


$$P(E) + P(E^C) = 1$$

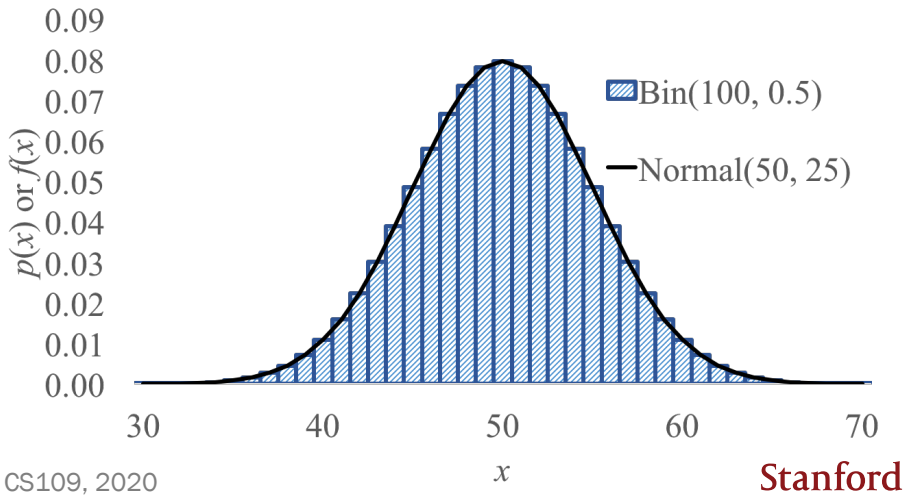
...to random variables and the Central Limit Theorem...

Bernoulli

Gaussian

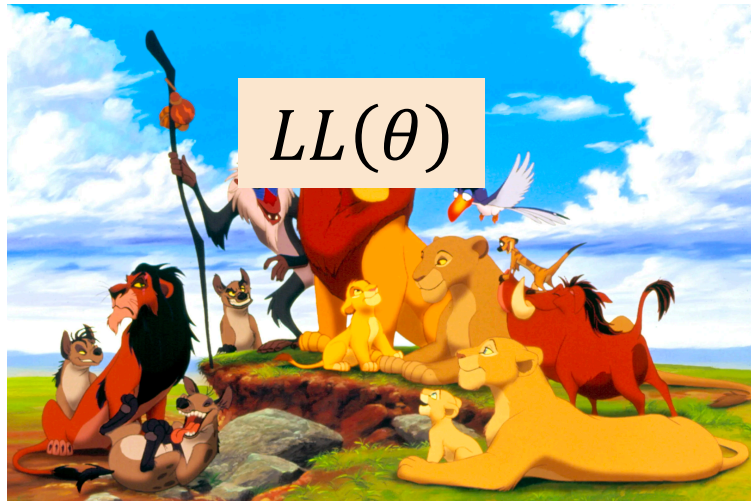


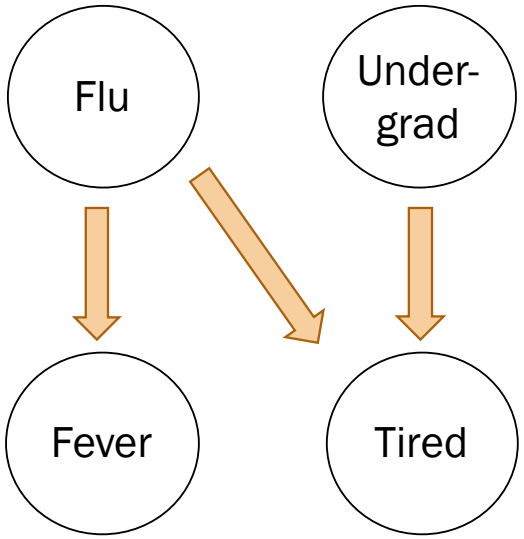
Poisson



...to statistics, parameter estimation, and machine learning

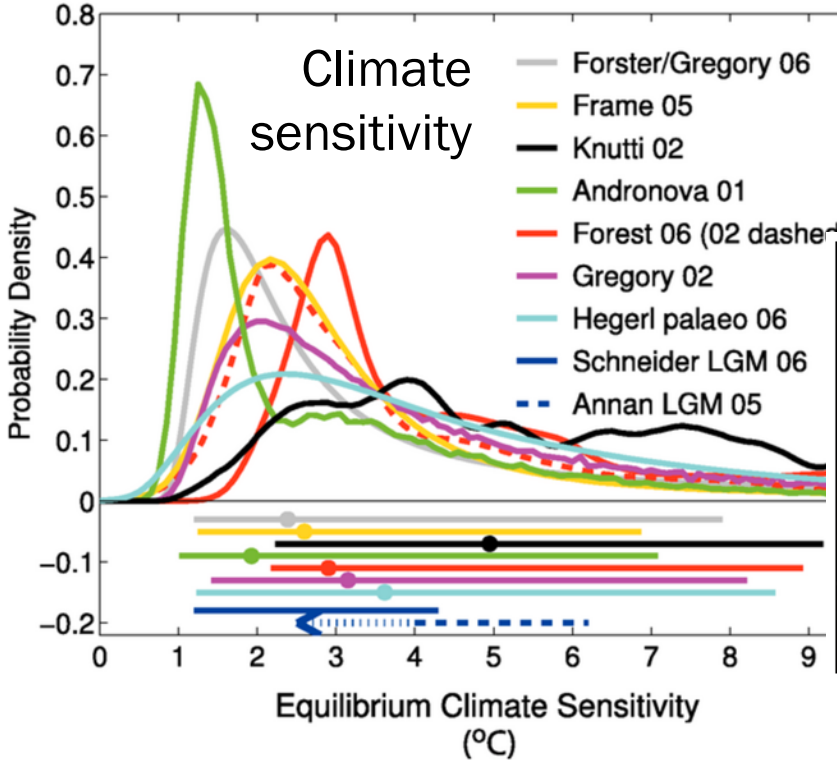
A happy
Bhutanese person

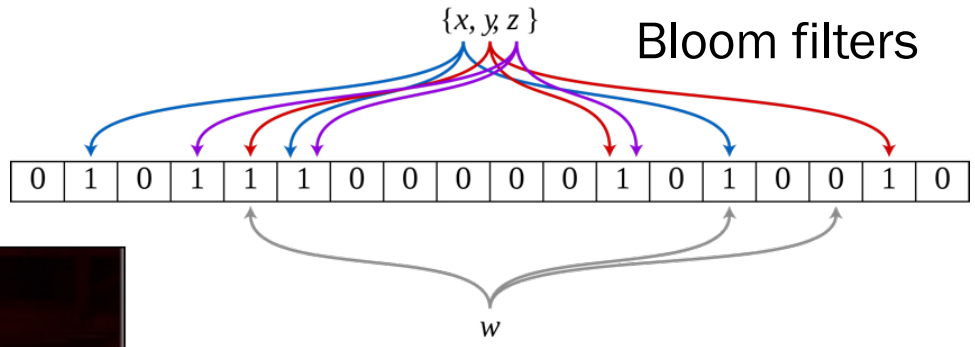




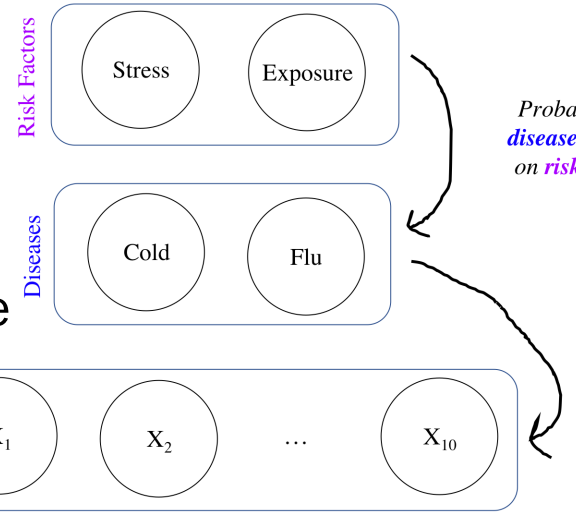
NETFLIX
and Learn

Lots and lots of analysis

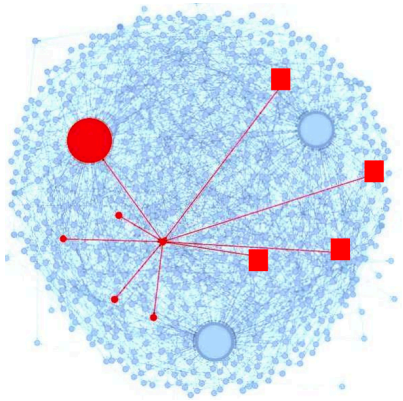




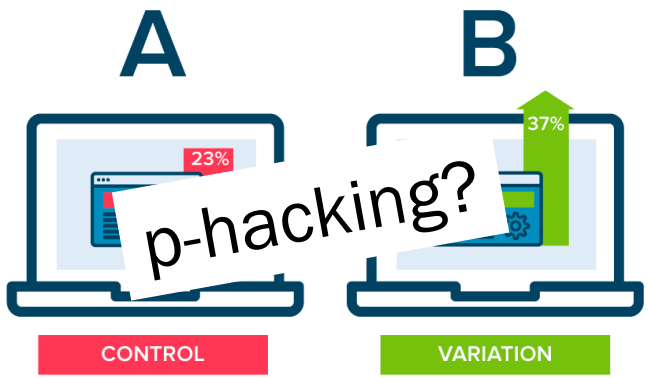
Why does he write like he's running out of time?



Web MD inference



Peer grading

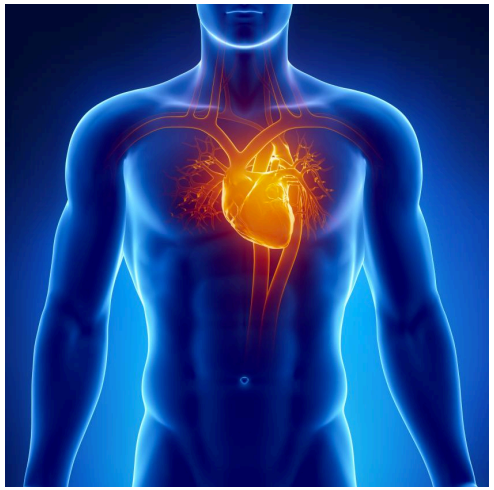


Coursera A/B testing

Lisa Yan and Jerry Cain, CS109, 2020

Lots and lots of analysis

Heart



Ancestry

23andMe

Netflix

What have we done together this quarter?

The CS109 teaching team

Discussion, synchronous and asynchronous

PS4Q8-Q12 [Federalist Papers]

Problem Sets – Ps... Lisa Y... INSTRUCTOR 1mth 34 1

Suggested classes to take with a social impact focus

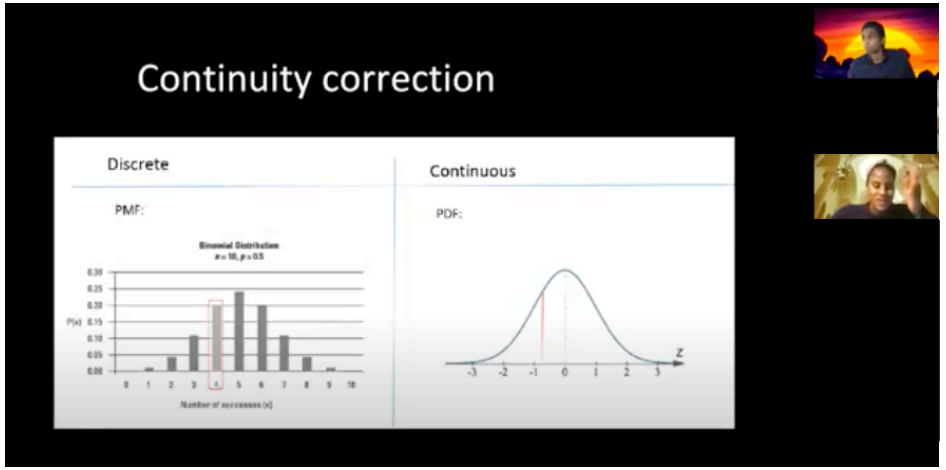
to continue working on projects that apply data analysis to social issues, and I was wondering what classes ... For example, would you recommend taking classes from the POLISCI 150 series or the SOC 180 series, or ... do you recommend to continue taking classes from the CS department? ... (Or a combination of all — do the concepts in the

Monday 16 November in General

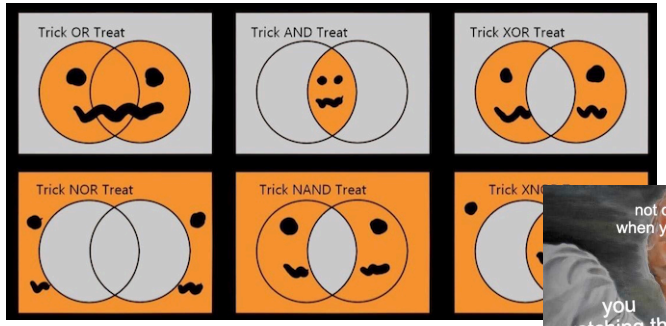
Suggested next class to take

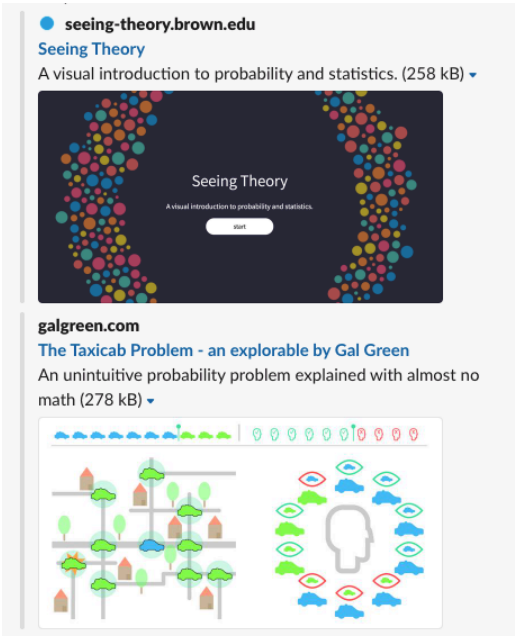
Hi everybody, What are the suggested math classes to t
the probability line before diving into machine learning

Thursday 5 November in General



I want to feel like I'm really alive	35:56
j'integrate donc je suis...?	36:21
Pretty sure that's the quote ^	36:45
definitely	36:56





Wow	39:15
I love these jokes	39:21
It couldn't find the other side	39:21
LOLLLL	39:23
Gottem	39:24
all the jokes were solid haha	39:26
Lol	39:26
ONE MORE	39:28
Hahaha	39:31
chat is NOT quiet	39:31
More!	39:33

slack

What have <u>you</u> learned from CS109?

What do you want to remember in 5 years?

What do you want to remember in the next 5 years?

Lisa's collection of plushies. (all the stats modeling stuff was cool too)

Definitely Bayes' Theorem for sure!

The Federalist Papers analysis was by far my favourite thing to do in this class! I will remember it forever!

The basic process of taking a word problem (some situation) and translating it into math/probability/etc.

poisson shark

Expectation

the applications in machine learning!

Inference and ML techniques are so cool!

I will remember how data/information can be connected and interpreted with probabilities and random variables. I will also remember the struggle of learning the equations.

The endless jokes and interludes, of course :)

I want to remember how funny Lisa and Jerry were.

Is this the end? After nine weeks of sharks, Pokémon, Doris, breakout rooms, operating system W, BestJokesEver.com, gift card giveaways, fish sticks, and so much more, I can't quite believe it's over. Actually, let me rephrase that—I'm can believe it's over. I'm just sad that it is. :(

Jerry's jokes!

That Bayes' Theorem is truly bae.

Lisa & Jerry's jokes also the central limit theorem

Lisa's Jokes! and Bayes' Thereom

the no idear joke

But so as not to fail this concept check, I should answer the question. I'll remember the friends I made during a challenging quarter. I'll remember the effort Lisa, Jerry, and all the TAs always put in for us students. And most of all, I'll remember how much fun attending class was—it's something so undescribably special when a required-for-your-major Zoom course has the liveliest atmosphere of any math class you've ever taken. :)

Everything I can! But also Lisa's great references

I love the machine learning content :)

P(I remember CS109 | I am alive) = .999

I will fondly remembered when Jerry demoed Rejection Sampling for analyzing disease symptoms as well as Lisa's iconic visual lecture models (:

I'll want to remember why MLE is so important.

a

I don't know what I'll be doing 5 years from now, so I don't know.

How to do bootstrapping!

Some of the top notch jokes from live lecture :)

I want to remember how the content from 109 saves my butt in the next 5 years

What can you do with CS109 material?

Your interests

everyday probability questions.

Being able to apply the same foundational concepts to a massive variety of problems

determining the ethicalness/fairness of an algorithm

into my longterm memory. I think in 5 years I'll be referencing th ese same concepts as I (potentially) work in data science.

I hope to use the Python skills I learned in a comp bio lab

Want to get much deeper into Machine Learning for my work (music composition) and h opefully will remember that in 5 years!

I want to work in biology and the example about false positives really struck me as interesting. Thank you all for a great year!

data analysis of what I will be working in the health field!!!

Fall 2020 contest winners

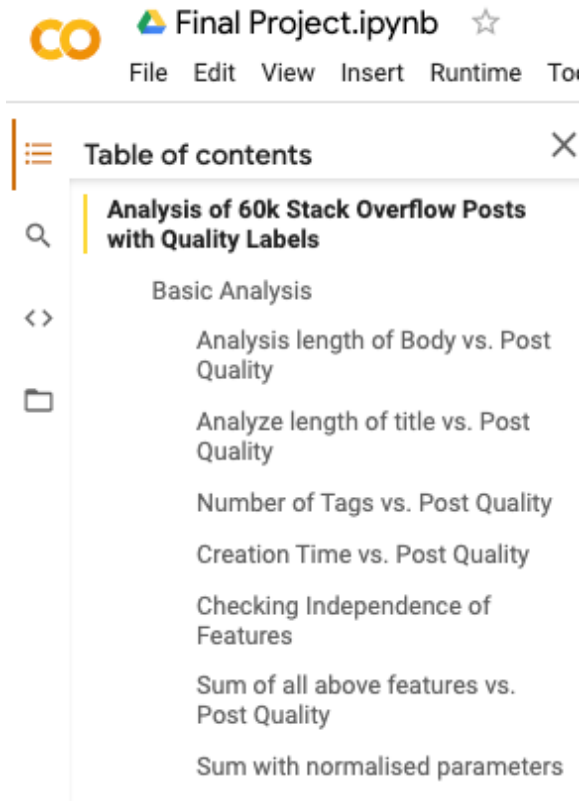
Grand Prize Winners:

- Ian Chang, Classifying Art using Probability
- Sohit Gatiganti and Chris Kim, Stack Overflow: A Deep Dive into Post Quality Analysis

Runners-Up:

- Valexa Orelien, The Book Matcher
- Edward Park, From the First to the Last, From Cradle to Grave
- Anna Quinlan, Improving Virtual Diabetes Patient Simulations with Bayesian Networks
- Erika Hunting and Wes Peisch, Metro Mania

Thus, the Naive Bayes MAP correctly predicts the era of the painting 70% of the time while the Naive Bayes MLE correctly predicts the era of the painting 80% of the time. This is pretty decent considering that random choice would only correctly predict era 25% of the time. The reasoning behind the mistakes also seems pretty obvious considering the conditional independence. Naive Bayes MLE being more accurate then MAP is probably due to the prior having a large effect on the small dataset. Let us look at the Naive Bayes MLE



After CS109

Theory

CS161 – Algorithmic analysis
Stats 217 – Stochastic Processes
CS238 – Decision Making Under Uncertainty
CS228 – Probabilistic Graphical Models

Statistics

Stats 200 – Statistical Inference
Stats 208 – Intro to the Bootstrap
Stats 209 – Group Methods/Causal Inference

After CS109

AI

CS 221 – Intro to AI
CS 229 – Machine Learning
CS 230 – Deep Learning

CS 224N – Natural Language Processing
CS 231N – Conv Neural Nets for Visual Recognition
CS 234 – Reinforcement Learning

Linear algebra:
Math 104/ENGR 108

Applications

CS 279 – Bio Computation
Literally any class with numbers in it

I hope I will continue to be skeptical of statistics

and data!

Correlation does not imply causation

X, Y independent → implies → $\mathrm{Cov}(X, Y) = 0$, $\rho(X, Y) = 0$

But the converse is not necessarily true!

Published: 13 May 1999

Myopia and ambient lighting at night

Graham E. Quinn, Chai H. Shin, Maureen G. Maguire & Richard A. Stone

Nature **399**, 113–114(1999) | Cite this article

- Child myopia correlated to sleeping with light on

Published: 09 March 2000

Vision

Myopia and ambient night-time lighting

Karla Zadnik, Lisa A. Jones, Brett C. Irvin, Robert N. Kleinstein, Ruth E. Manny, Julie A. Shin & Donald O. Mutti

Nature **404**, 143–144(2000) | Cite this article

- Parental myopia correlated with child myopia
- Myopic parents correlated with leaving light on

Not all correlations should be dismissed as spurious

Published: 30 August 1958

Cancer and Smoking

RONALD A. FISHER

Nature **182**, 596(1958) | Cite this article

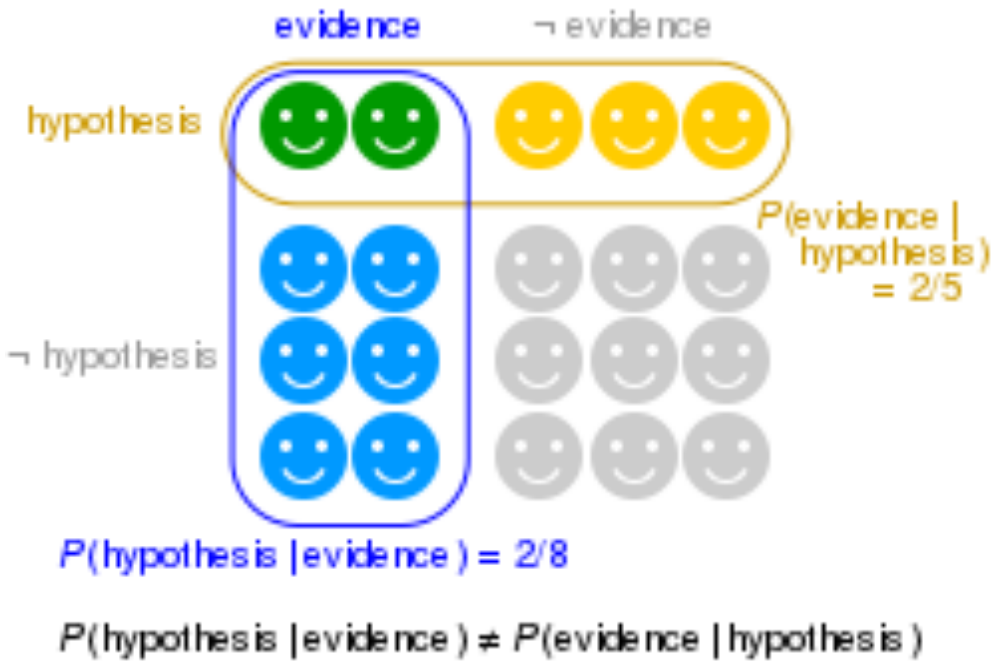
1504 Accesses | **101** Citations | **30** Altmetric | Metrics

Abstract

THE curious associations with lung cancer found in relation to smoking habits do not, in the minds of some of us, lend themselves easily to the simple conclusion that the products of combustion reaching the surface of the bronchus induce, though after a long interval, the development of a cancer. If, for example, it were possible to infer that smoking cigarettes is a cause of this disease, it would equally be possible to infer on exactly similar grounds that inhaling cigarette smoke was a practice of considerable prophylactic value in preventing the disease, for the practice of inhaling is rarer among patients with cancer of the lung than with others.

- Fisher–Tippett–Gnedenko theorem Fis
- Fisher–Tippett distribution
- Von Mises–Fisher distribution[81]
- Inverse probability, a term Fisher used
- Fisher's permutation test
- Fisher's inequality[83]
- Sufficient statistic, when a statistic is *s* associated unknown parameter if "no c provides any additional information as
- Fisher's noncentral hypergeometric dis distribution, where sampling probabiliti
- Student's *t*-distribution, widely used in

Understanding Bayes' Rule

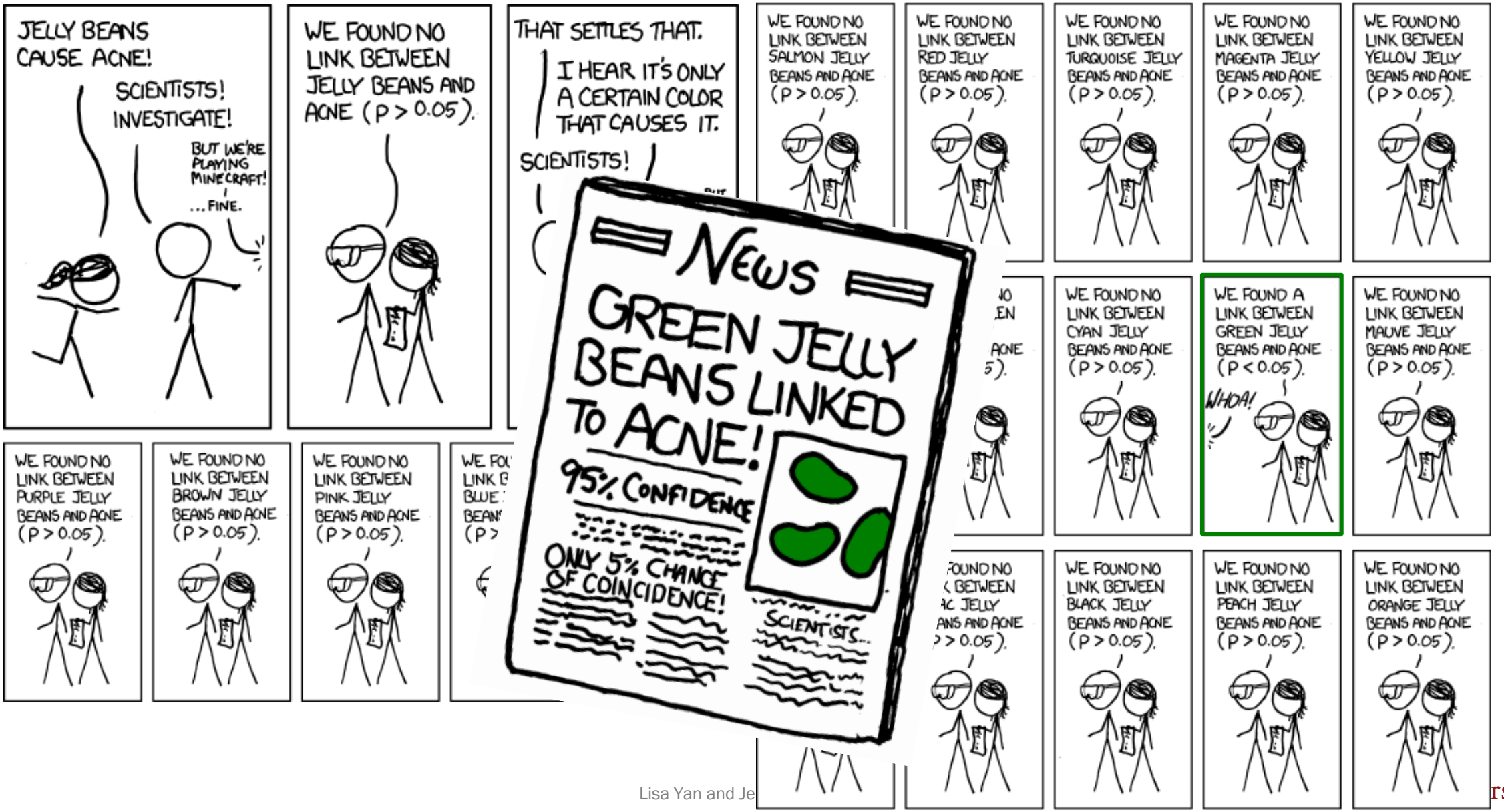


Prosecutor's Fallacy:

"The odds of finding this evidence on an innocent man are so small that the jury can safely disregard the possibility that this defendant is innocent."

https://en.wikipedia.org/wiki/Prosecutor%27s_fallacy

Mishandling of p-values



Ethics and datasets

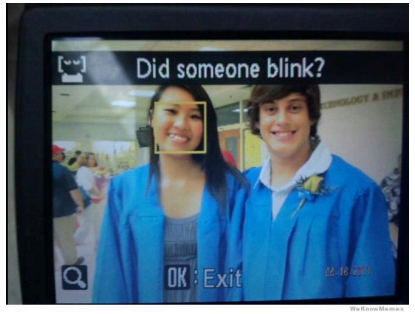
Sometimes machine learning feels universally unbiased.

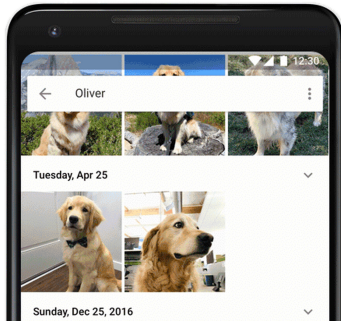
We can even prove our estimators are “unbiased” (mathematically).

Google/Nikon/HP have had biased datasets.

"HP has been informed of a potential issue with facial-tracking software. Consistent with other webcams, proper foreground lighting is required for the product to effectively track any person and their movements,” [HP, 2009]

“We’re appalled and genuinely sorry that this happened…We are taking immediate action to prevent this type of result from appearing. There is still clearly a lot of work to do with automatic image labelling, and we’re looking at how we can prevent these types of mistakes from happening in the future.” [Google, 2015]





Face-tracking, auto-tagging

Should your data be unbiased?

Dataset: Google News

$$\overrightarrow{\text{man}} - \overrightarrow{\text{woman}} \approx \overrightarrow{\text{king}} - \overrightarrow{\text{queen}}$$

$$\overrightarrow{\text{man}} - \overrightarrow{\text{woman}} \approx \overrightarrow{\text{computer programmer}} - \overrightarrow{\text{homemaker}}.$$

Should our unbiased data collection reflect society's systemic bias?

Bolukbasi et al., Man is to Computer Programmer as Woman is to Homemaker? Debiasing Word Embeddings. NIPS 2016

How can we explain decisions?

If your task is image classification, reasoning about high-level features is relatively easy.

Everything can be visualized.

What if you are trying to classify social outcomes?

- Criminal recidivism
- Job performance
- Policing
- Terrorist risk
- At-risk kids

Why study probability + CS?

Why study probability + CS?

Fastest growing occupations: 20 occupations with the highest percent change of employment between 2018-28.

Click on an occupation name to see the full occupational profile.

OCCUPATION	GROWTH RATE, 2018-28	2018 MEDIAN PAY
Physician assistants	31%	$108,610 per year
Nurse practitioners	28%	$107,030 per year
Software developers, applications	26%	$103,620 per year
Mathematicians	26%	$101,900 per year
Information security analysts	32%	$98,350 per year
Health specialties teachers, postsecondary	23%	$97,370 per year
Statisticians	31%	$87,780 per year
Operations research analysts	26%	$83,390 per year
Genetic counselors	27%	$80,370 per year

Source: US Bureau of Labor Statistics

Why study probability + CS?

Interdisciplinary

Closest thing to magic

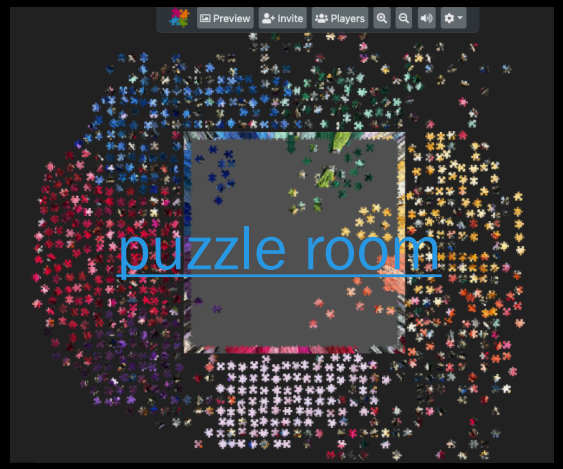
Why study probability + CS?

Everyone is welcome!

Technology magnifies. What do we want magnified?

You are all one step closer to improving the world.

(all of you!)



The end

See you soon… ☺