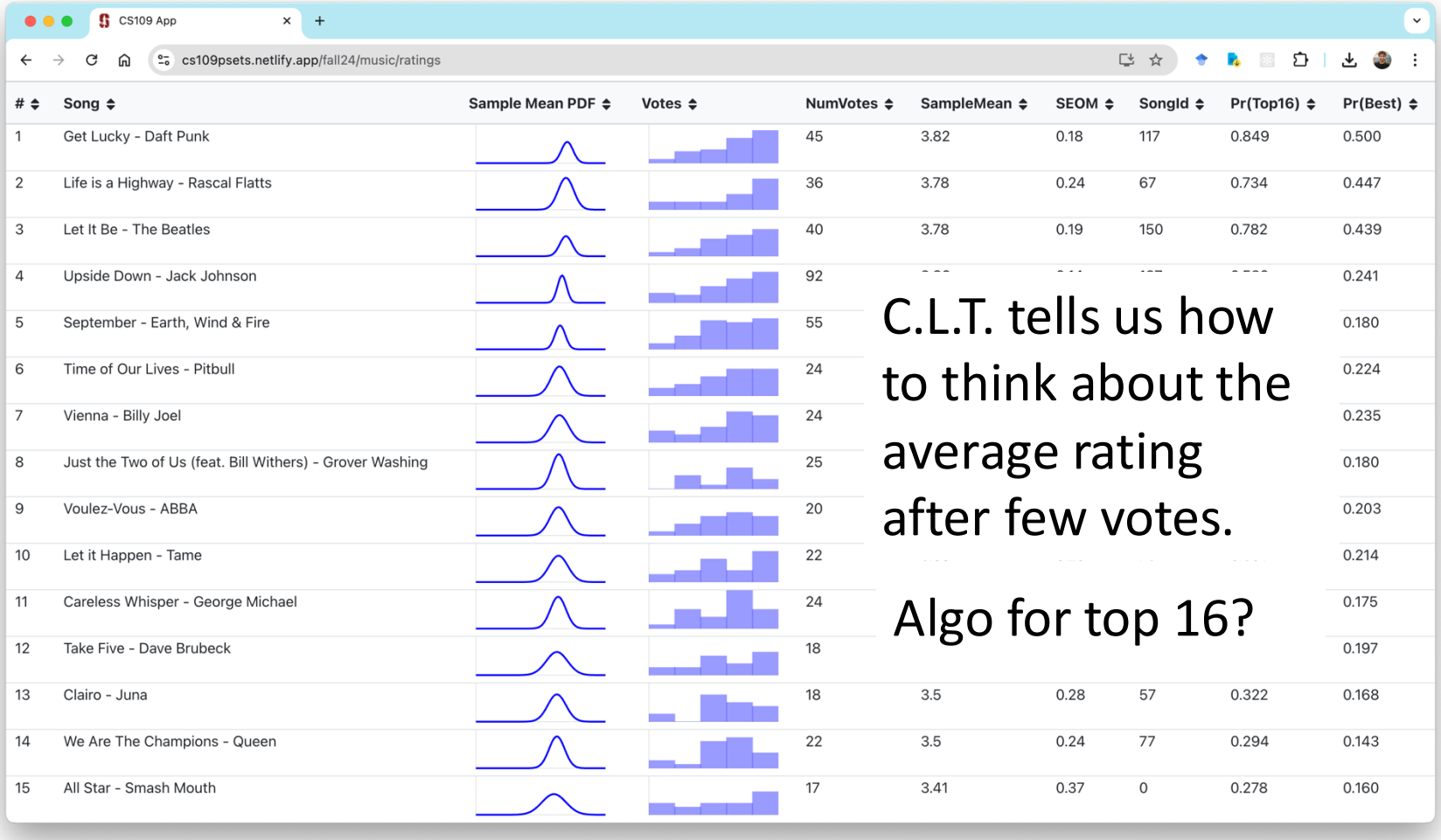
Beyond CS109

Chris Piech
CS109, Stanford University

Class Theme Song



#	Song	Sample Mean PDF	Votes	NumVotes	SampleMean	SEOM	SongId	Pr(Top16)	Pr(Best)
1	Get Lucky - Daft Punk			45	3.82	0.18	117	0.849	0.500
2	Life is a Highway - Rascal Flatts			36	3.78	0.24	67	0.734	0.447
3	Let It Be - The Beatles			40	3.78	0.19	150	0.782	0.439
4	Upside Down - Jack Johnson			92					0.241
5	September - Earth, Wind & Fire			55					0.180
6	Time of Our Lives - Pitbull			24					0.224
7	Vienna - Billy Joel			24					0.235
8	Just the Two of Us (feat. Bill Withers) - Grover Washing			25					0.180
9	Voulez-Vous - ABBA			20					0.203
10	Let it Happen - Tame			22					0.214
11	Careless Whisper - George Michael			24					0.175
12	Take Five - Dave Brubeck			18					0.197
13	Clairo - Juna			18	3.5	0.28	57	0.322	0.168
14	We Are The Champions - Queen			22	3.5	0.24	77	0.294	0.143
15	All Star - Smash Mouth			17	3.41	0.37	0	0.278	0.160

C.L.T. tells us how to think about the average rating after few votes.

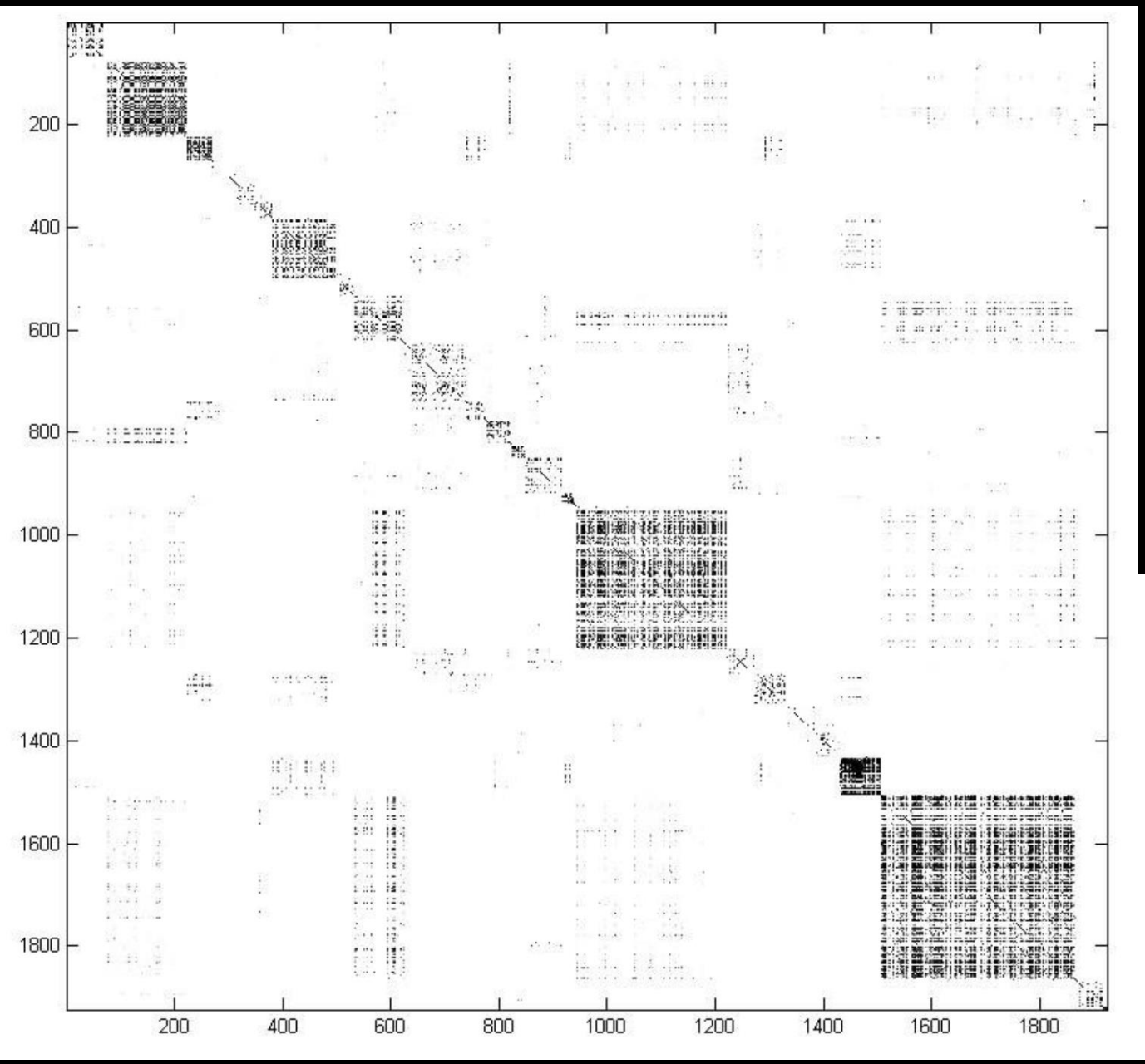
Algo for top 16?

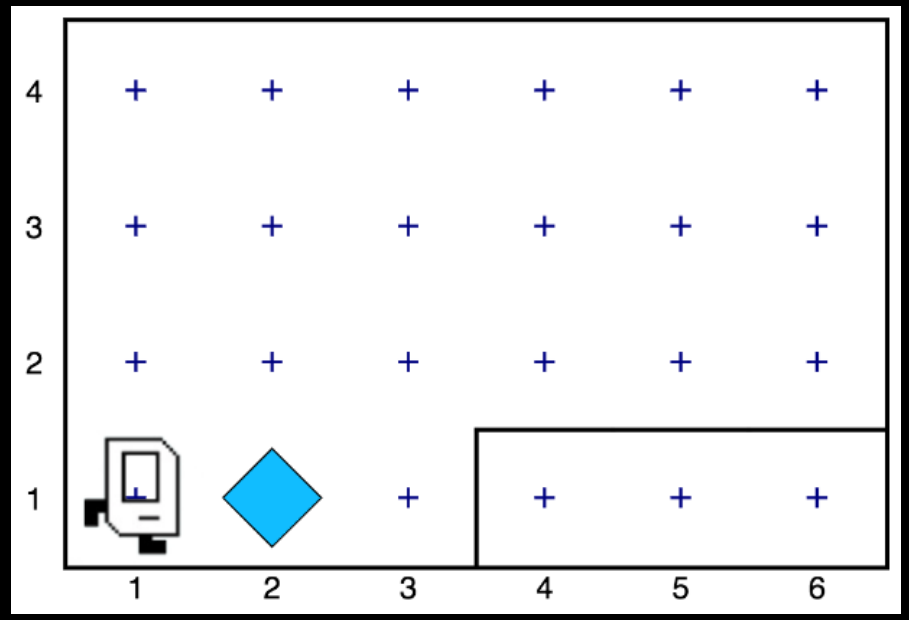
This quarter

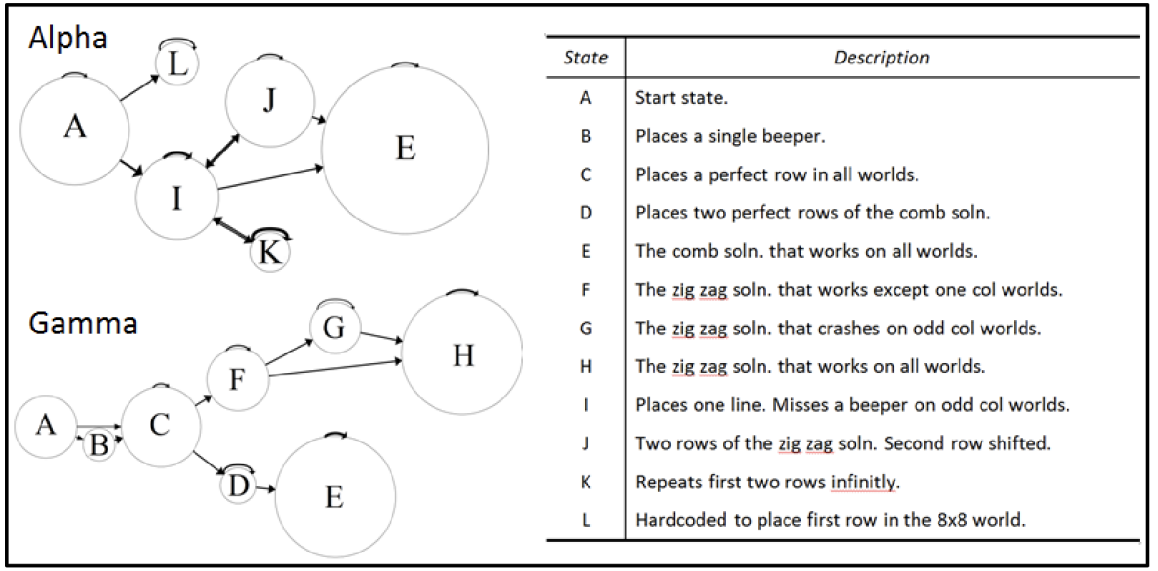
1	Sereia - Mundo Azul
2	Smart - Le Sserafim
3	Viva La Vida - Coldplay
4	Take Me Home, Country Roads - John Denver
5	Real Love Baby - Father John Misty
6	Sports car - Tate Mcrea
7	Something Just Like This - The Chainsmokers
8	My Shot - Lin-Manuel Miranda
9	Young Folks - Peter Bjorn and John
10	Don't Let Me Down - The Beatles
11	Feel Again - With Heartbeats - OneRepublic :D
12	Parce que je t'aime - Elodie Frege
13	Golden - KPOP Demon Hunters
14	Oghniat Al Wadaa - Fairuz
15	Tokyo Daylight - Lyn
16	Tampa - Filho do Zua

Reflection: I love music. Should we focus more on love of songs instead of average?

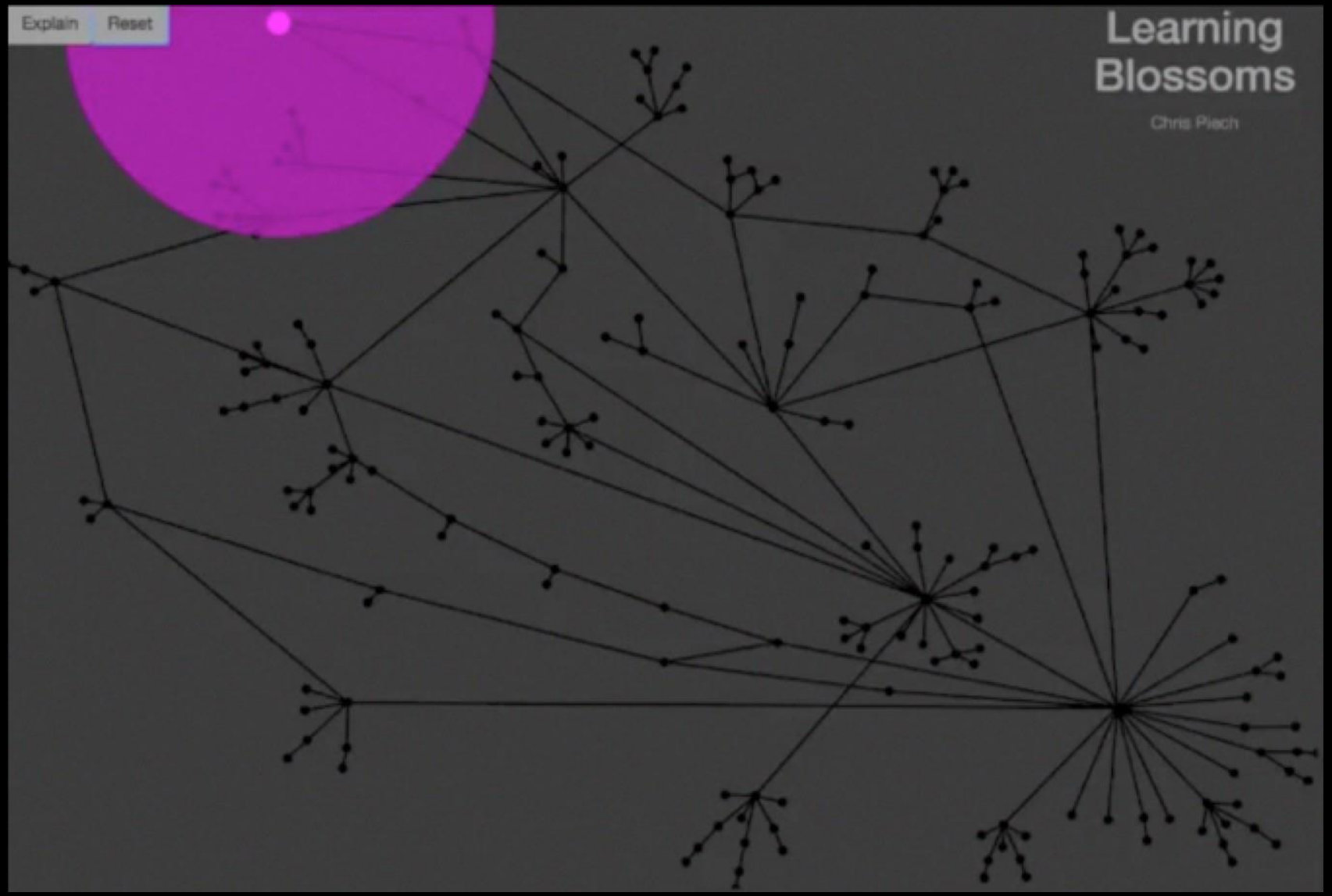
Probability gives you a new lens on the world

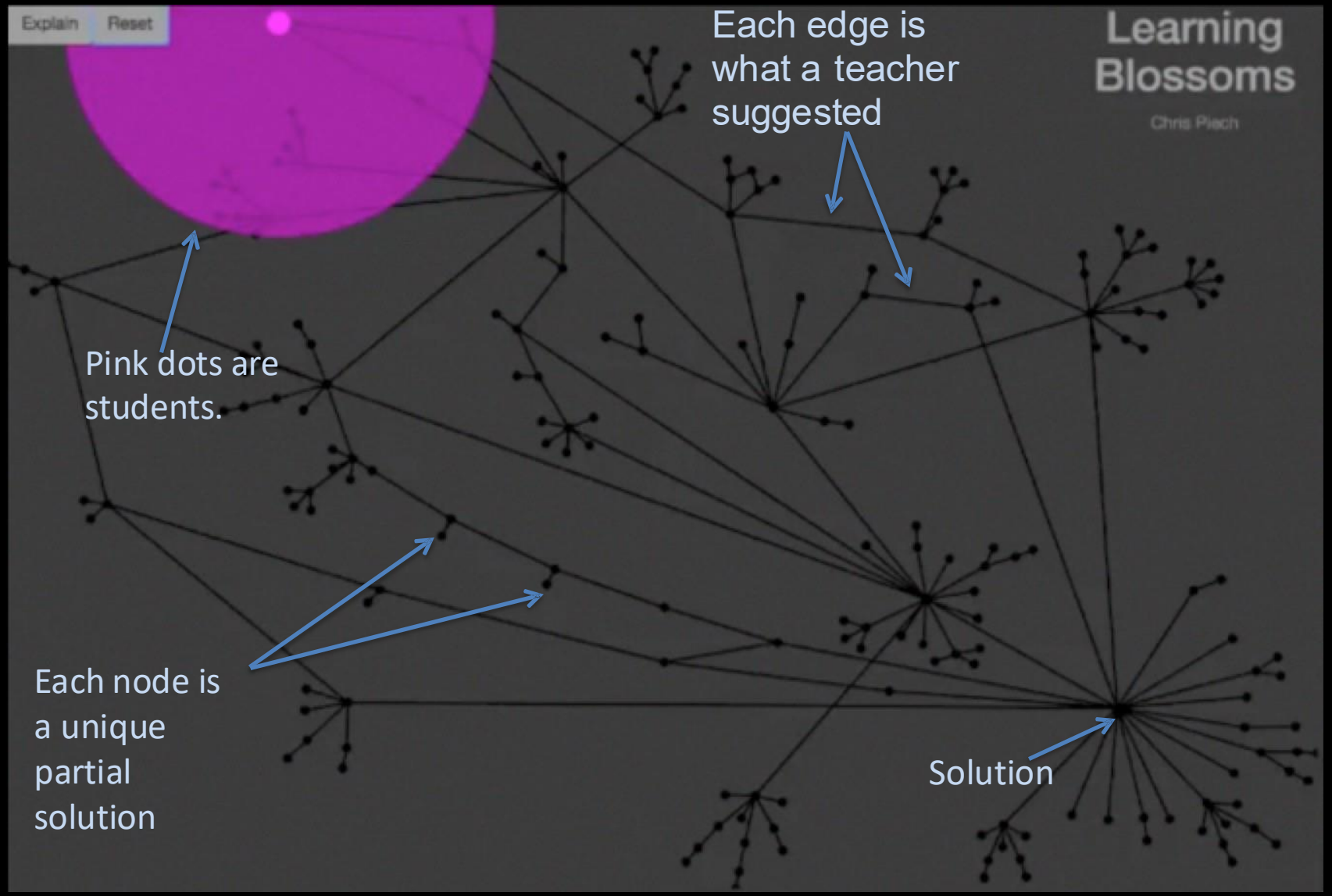


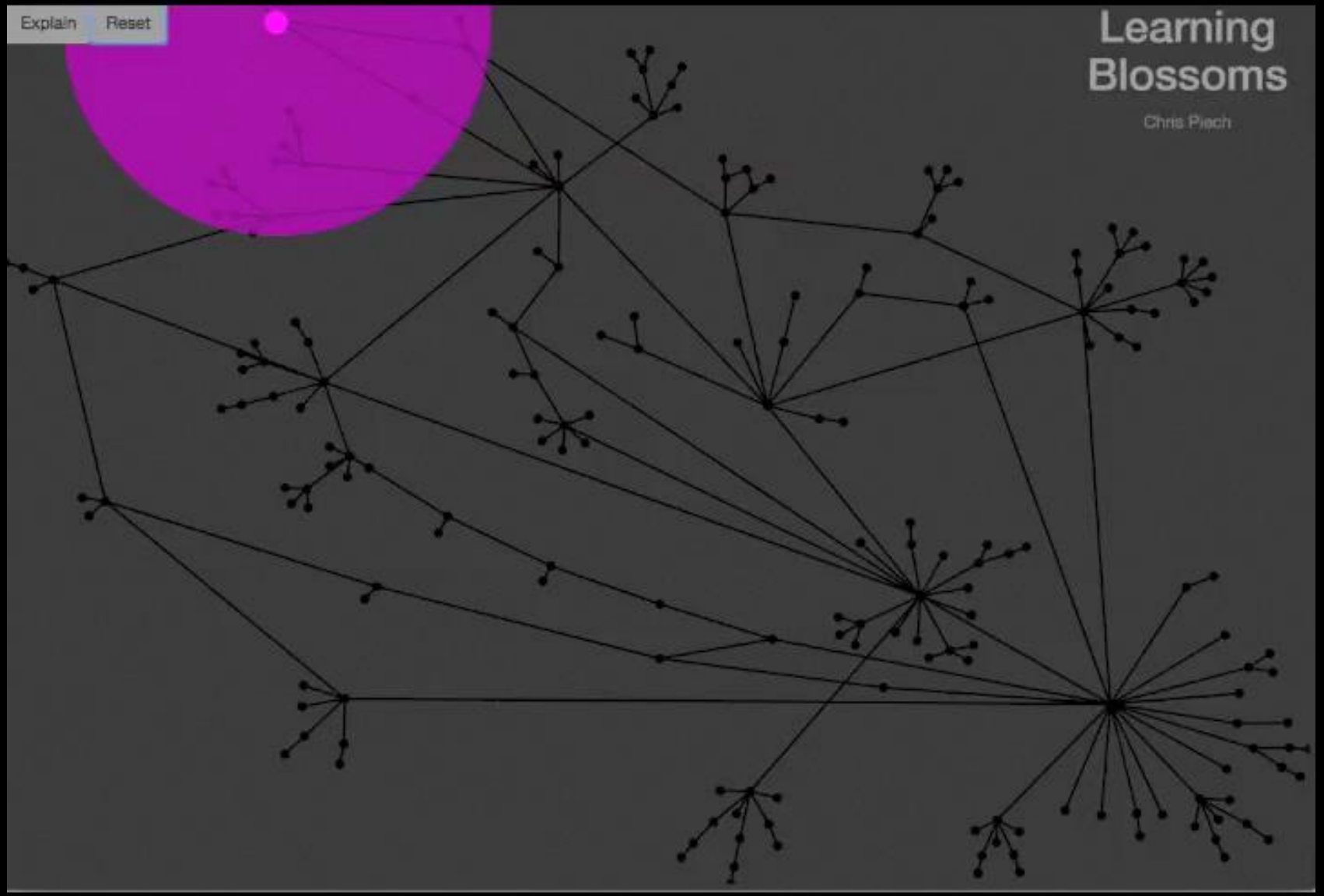




State	Description
A	Start state.
B	Places a single beeper.
C	Places a perfect row in all worlds.
D	Places two perfect rows of the comb soln.
E	The comb soln. that works on all worlds.
F	The zig zag soln. that works except one col worlds.
G	The zig zag soln. that crashes on odd col worlds.
H	The zig zag soln. that works on all worlds.
I	Places one line. Misses a beeper on odd col worlds.
J	Two rows of the zig zag soln. Second row shifted.
K	Repeats first two rows infinitly.
L	Hardcoded to place first row in the 8x8 world.


Explain
Reset
Learning
Blossoms
Chris Piech


Explain
Reset
Each edge is what a teacher suggested
Learning Blossoms
Chris Piech
Pink dots are students.
Each node is a unique partial solution
Solution


Explain
Reset
Learning
Blossoms
Chris Piech

The Crowd is Un-wise

Temporal methods tried:
Shortest path
Min Time
Expected Success
Reinforcement learning
Most Common Next
Most Popular Path

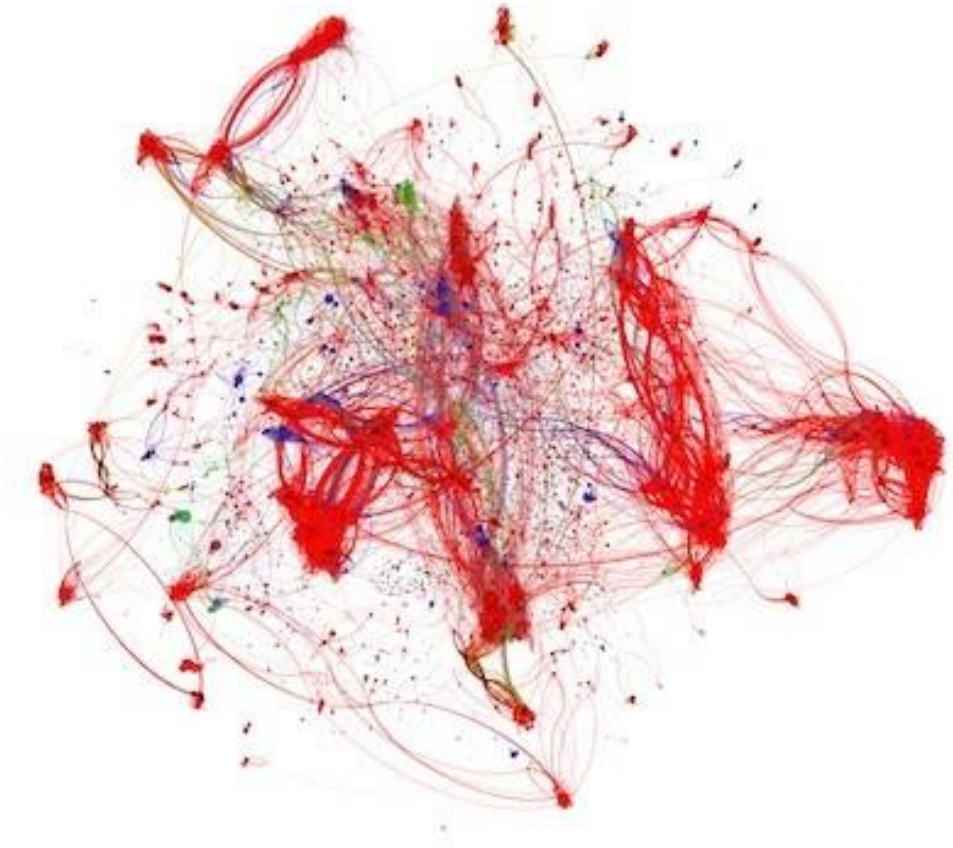
18%

45%

12%

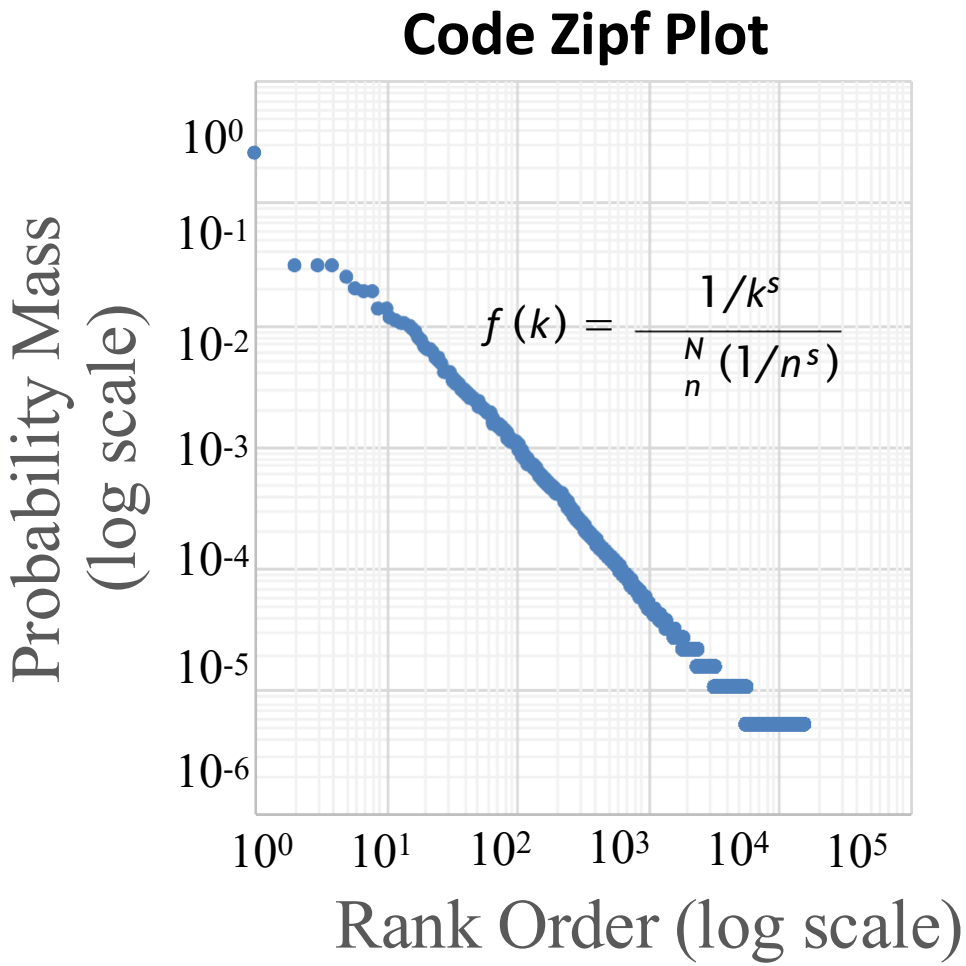
Hard Problem!

Brute force solution?



1 million unique solutions to programming Linear Regression

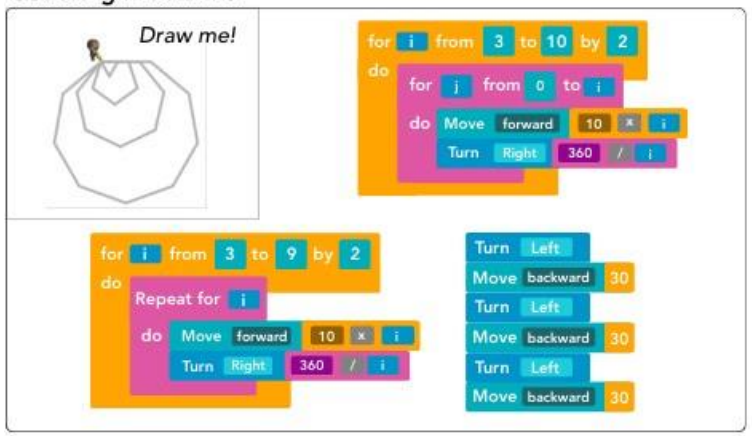
WWW 2014



They are all Zipf!

(a) Datasets in Computational Education

Code.org Problem 8



Powergrading P13

What is one reason the original colonists came to America?

- Religuous freedom
- For religious freedom
- Freedom

- declared our independence from england
- religeous freedom
- as a criminal punishment

- to create a new colony
- to find better economic prospects
- to break away from the church in great britain

CS1: Liftoff

Write a Java Program to print the numbers 10 down to 1 and then write liftoff. You must use a loop.

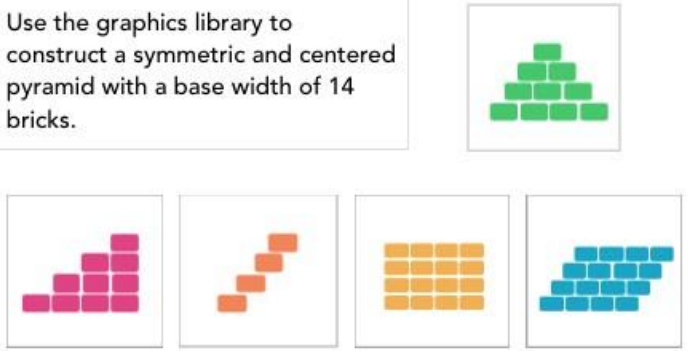
```
public void run() {
  for (int i=START; i>0; i -=1)
  {
    println(i);
  }
  println("Liftoff");
}
```

```
public void run() {
  for(int i = START; i>0; i--)
  {
    println(i);
    pause(1000);
  }
  println("Liftoff!");
}
```

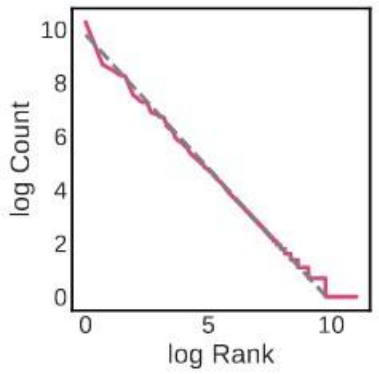
```
public void run() {
  int x = START;
  int y = 1;
  int z = 9;
  while (x>=1) {
    println(x);
    x=z;
    z=x-y;
  }
  println("Liftoff");
}
```

PyramidSnapshot

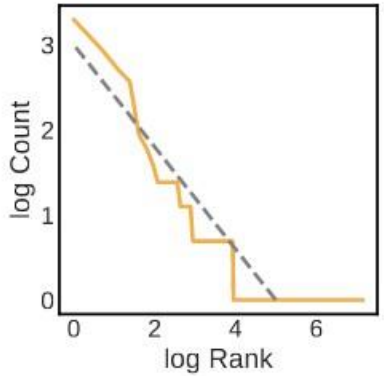
Use the graphics library to construct a symmetric and centered pyramid with a base width of 14 bricks.



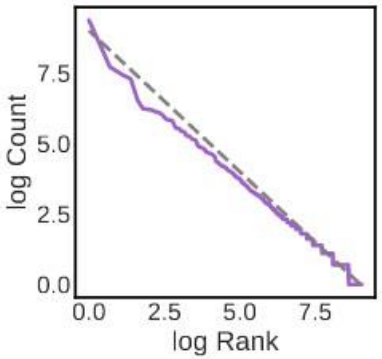
(b) Code.org P8



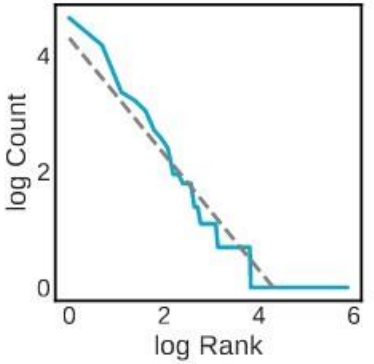
(c) CS1: Liftoff



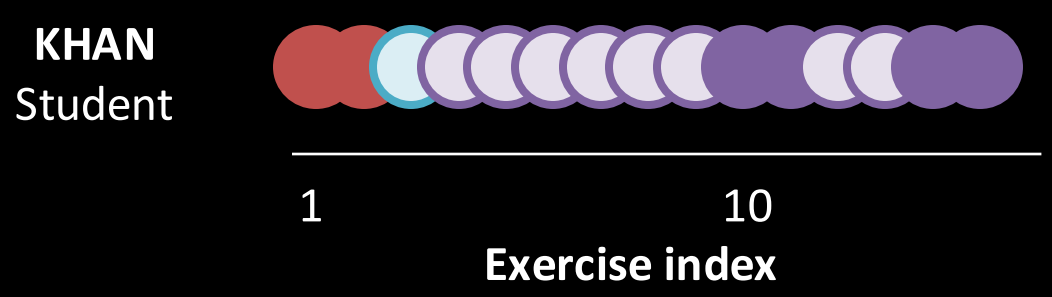
(d) Pyramid



(e) Powergrading



Original deep learning for education paper



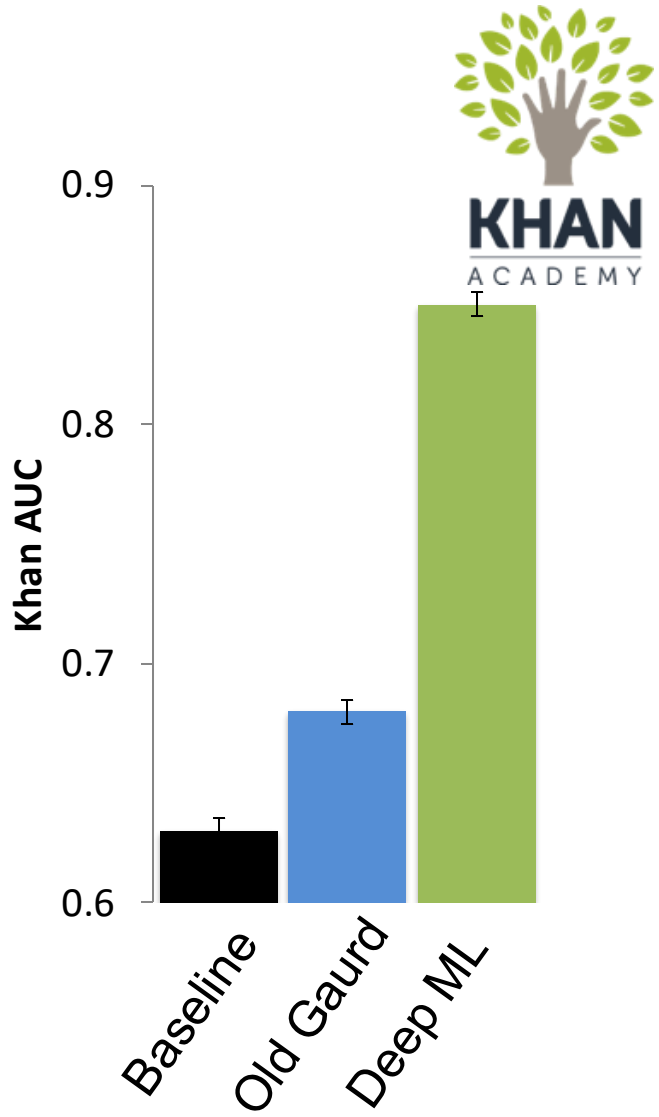
Exercise Type:

- Solving for x-intercept
- Solving for y-intercept
- Square roots

Answer:

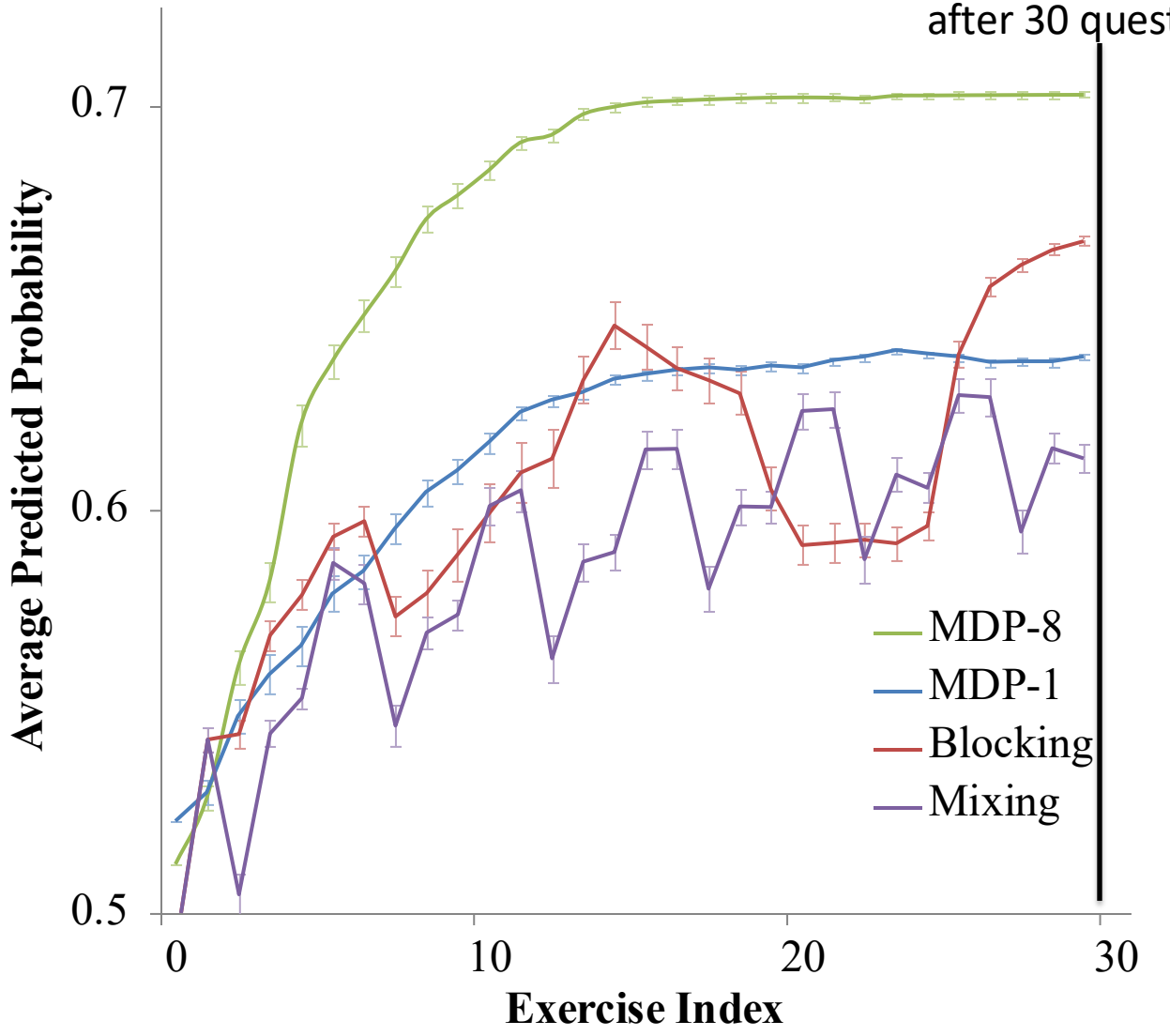
- Correct
- Incorrect

Powers DuoLingo Bird Brain (as of 2023)



Optimal Teaching

Maximize knowledge after 30 questions

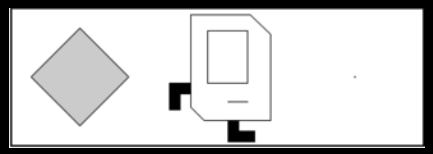


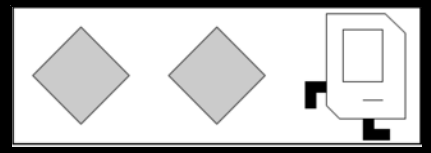
Used in DuoLingo Bird Brain (last checked in 2023)

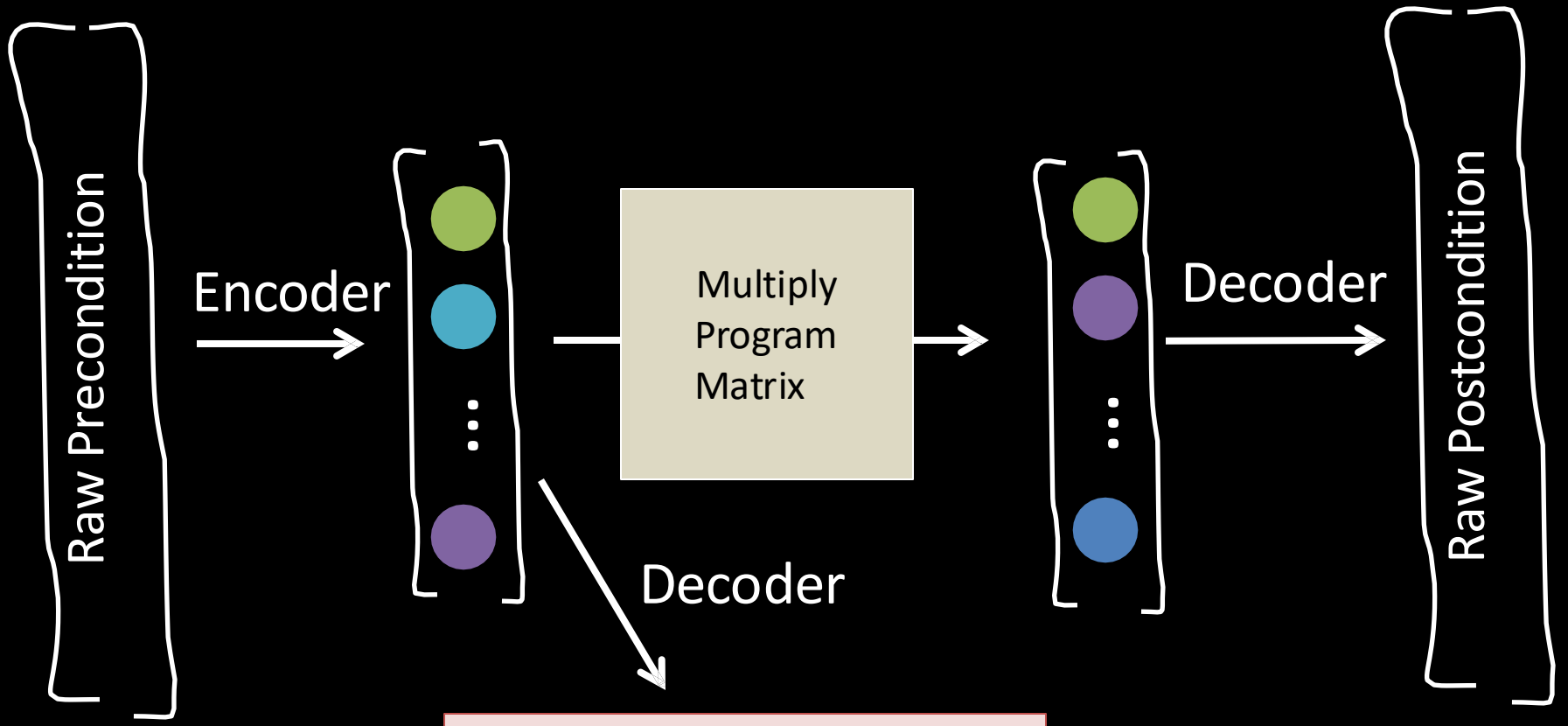
Back to coding!

Neural Network for Programs

```
method step() {
    putBeeper();
    move();
}
```



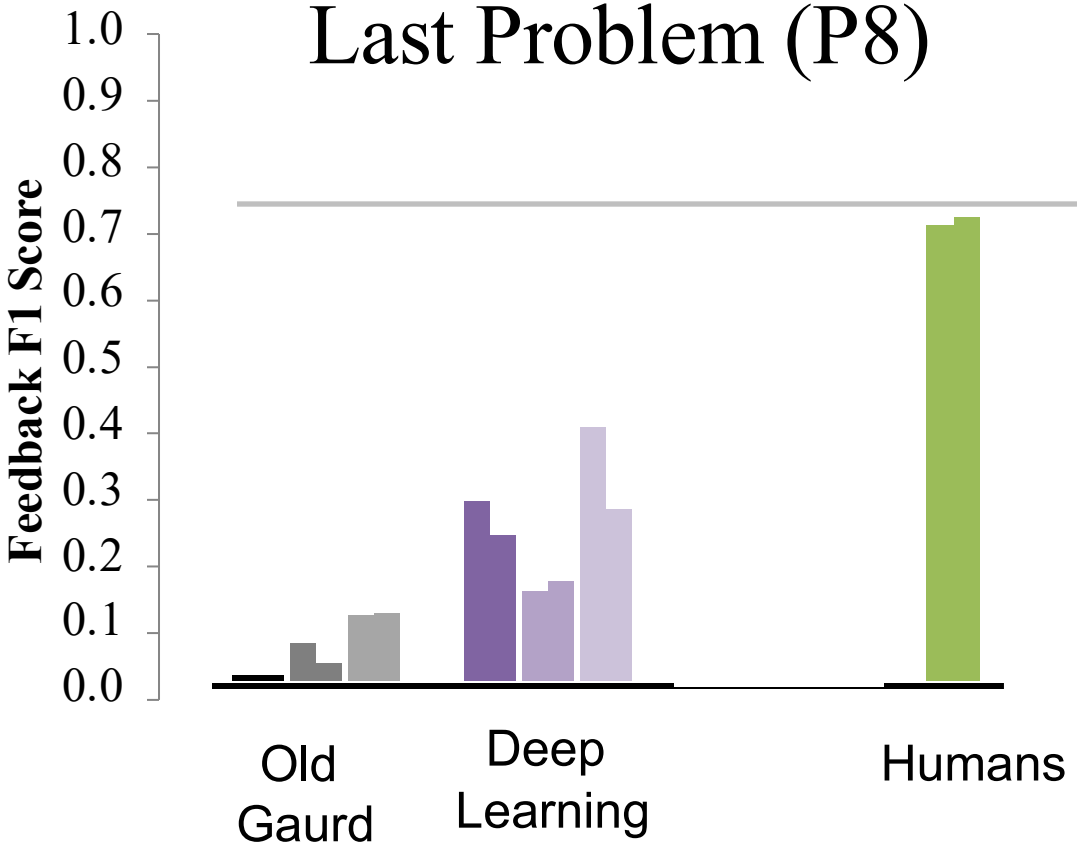


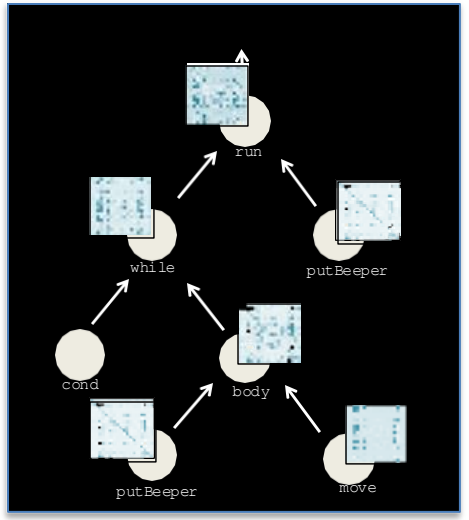


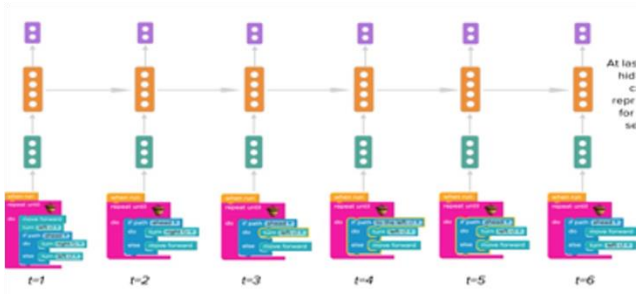
*coded pre-tensor flow

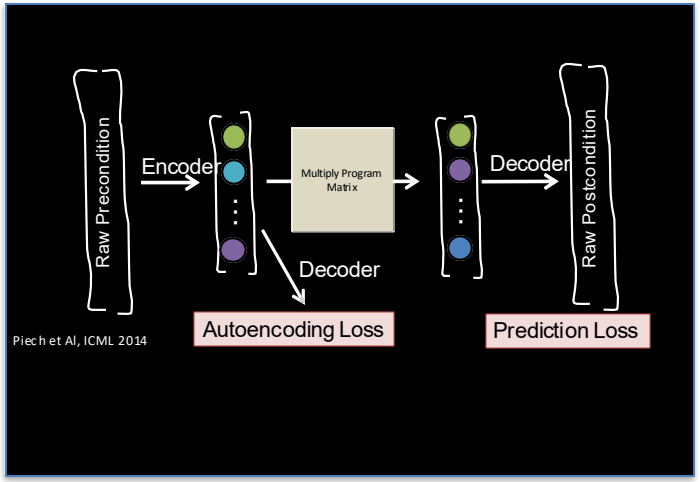
Inaccurate, Uninterpretable, and Data Hungry

Label student code



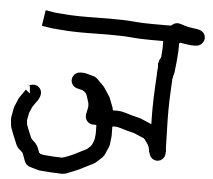






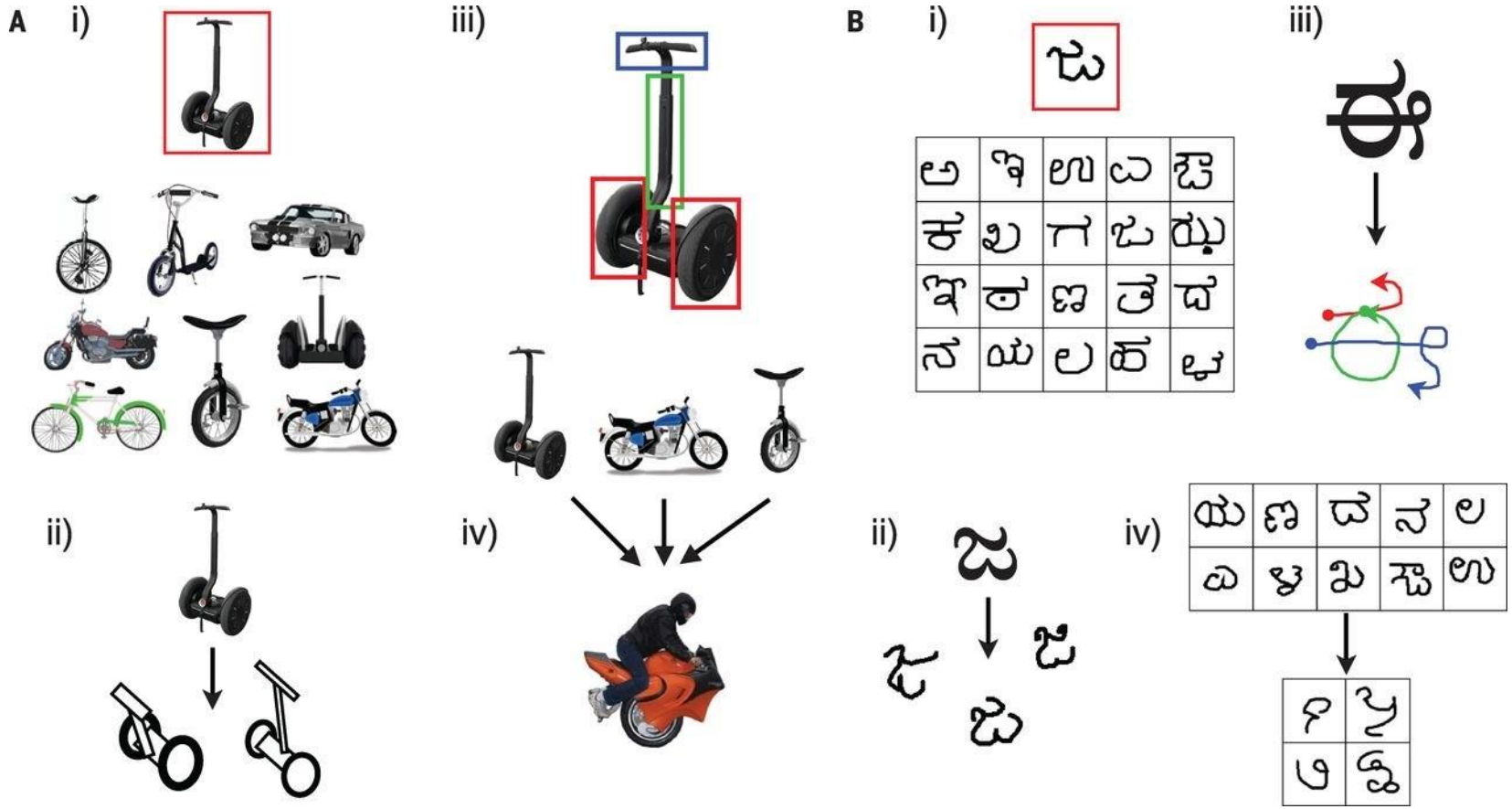
Humans Don't Need Much Data

Single training example:



Test set:

Fig. 1 People can learn rich concepts from limited data.

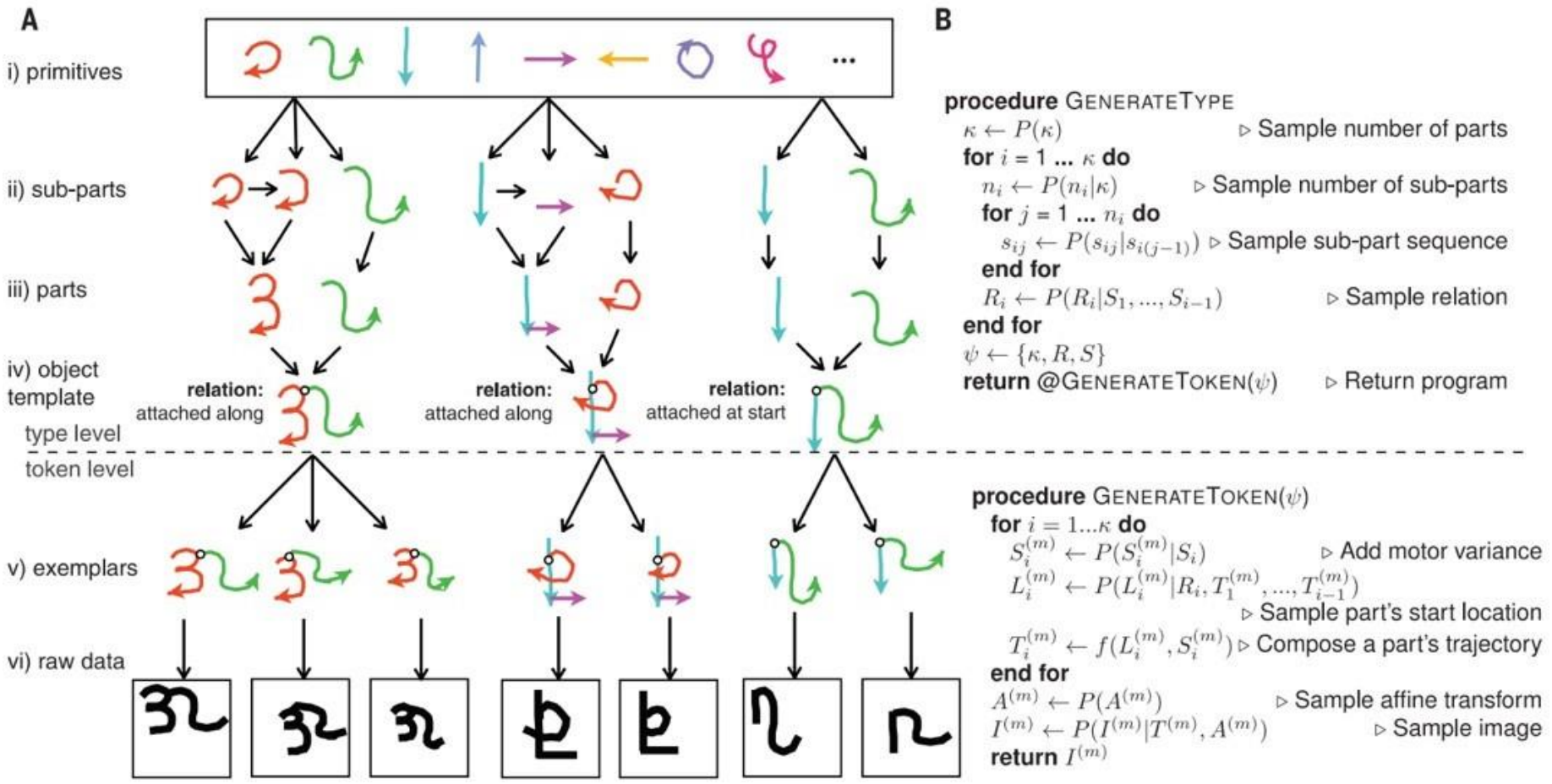


Brenden M. Lake et al. Science 2015;350:1332-1338

Fig. 2 Simple visual concepts for comparing human and machine learning.

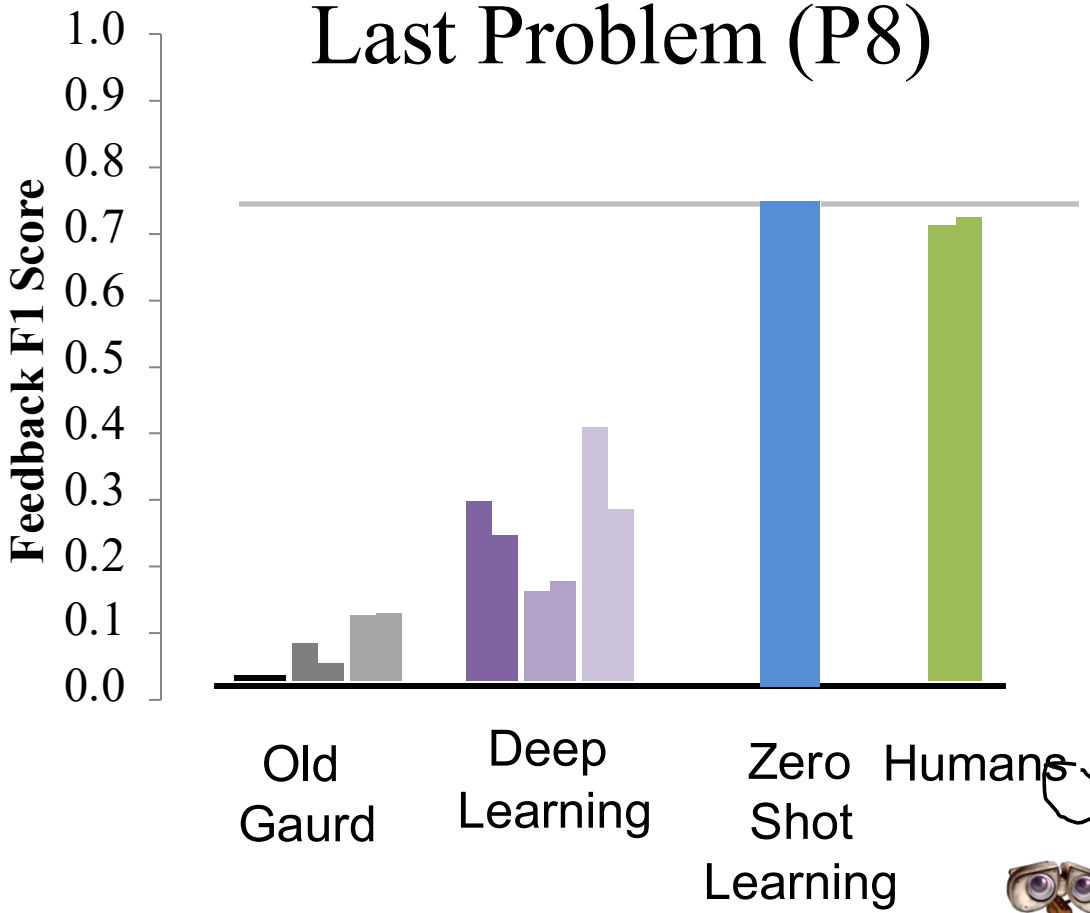
Brenden M. Lake et al. Science 2015;350:1332-1338

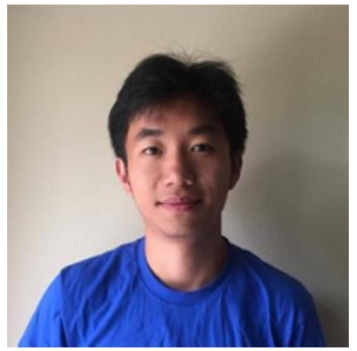
Bayesian Program Learning



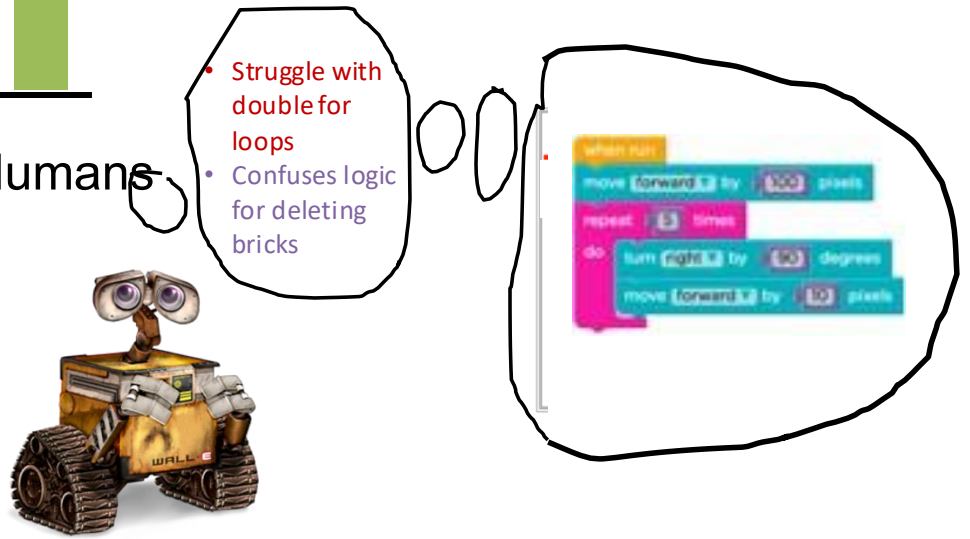
Generative Understanding

Label student code

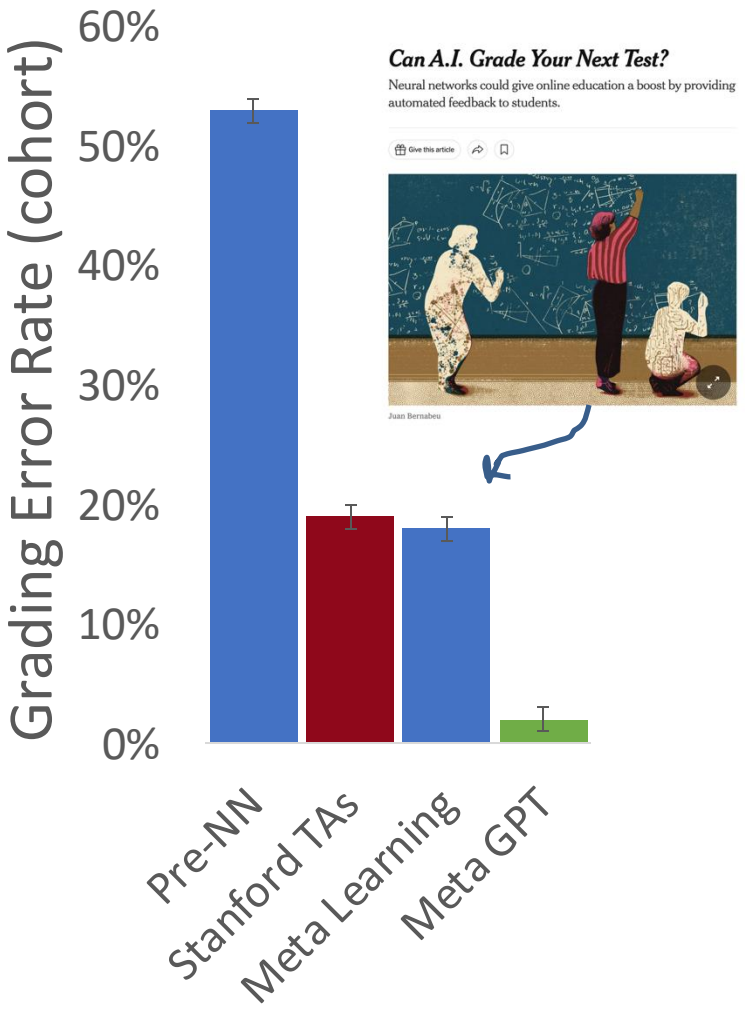




Outstanding Student paper award, AAAI 2019



Rubric Level Accuracy on Few-Shot Grading a Novel Question



Now lives in code.org's teaching assistant

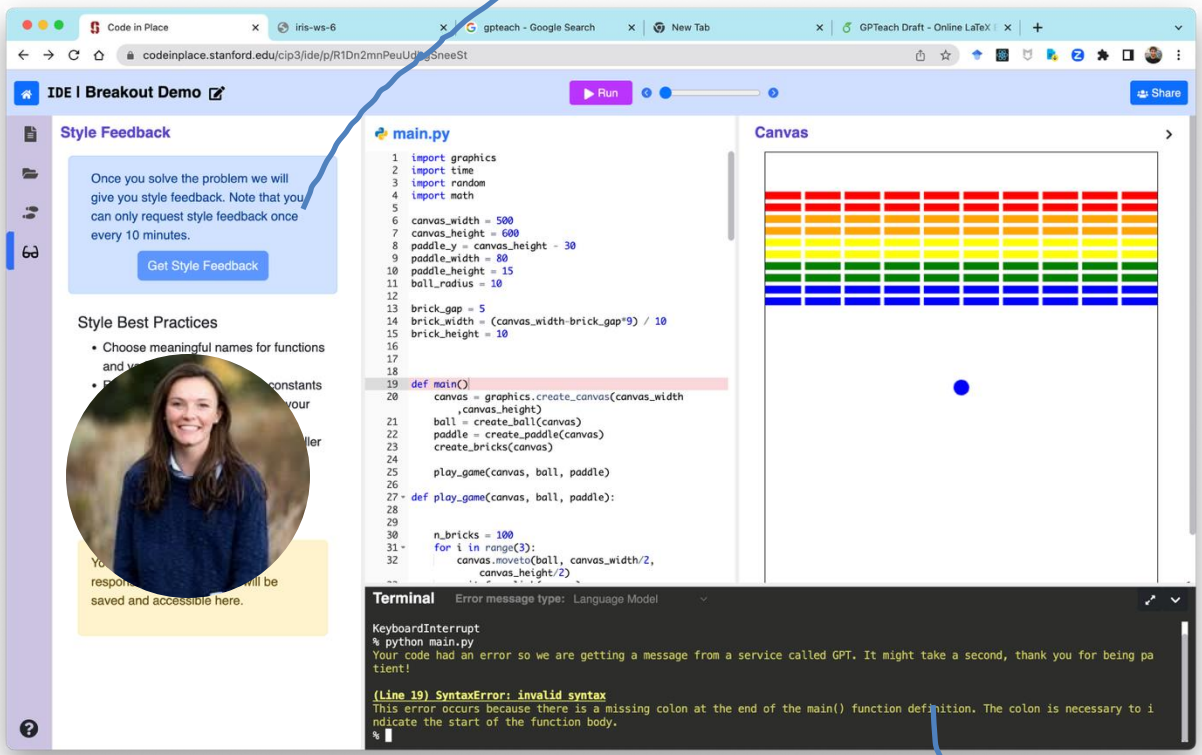
Stanford Code in Place:

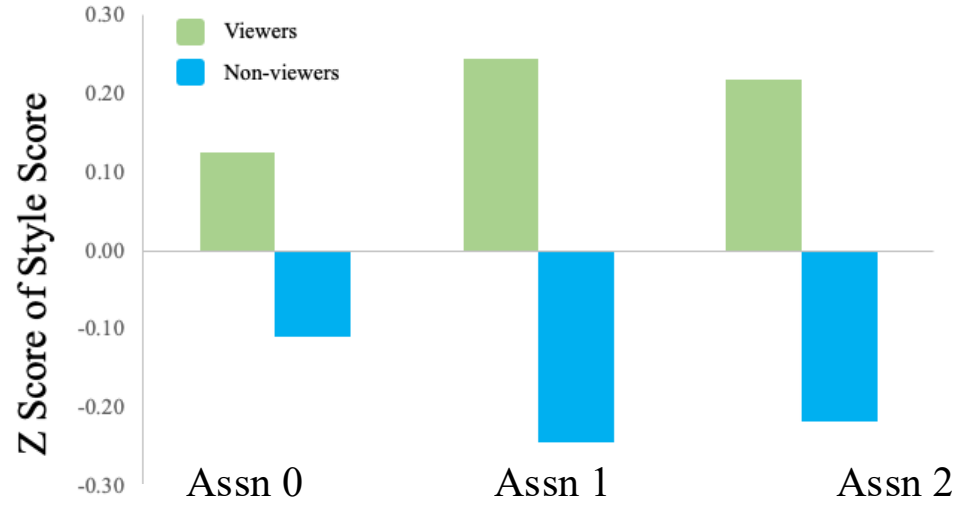
5000+ section leaders teach
60,000+ students
CS106A
As Community Service

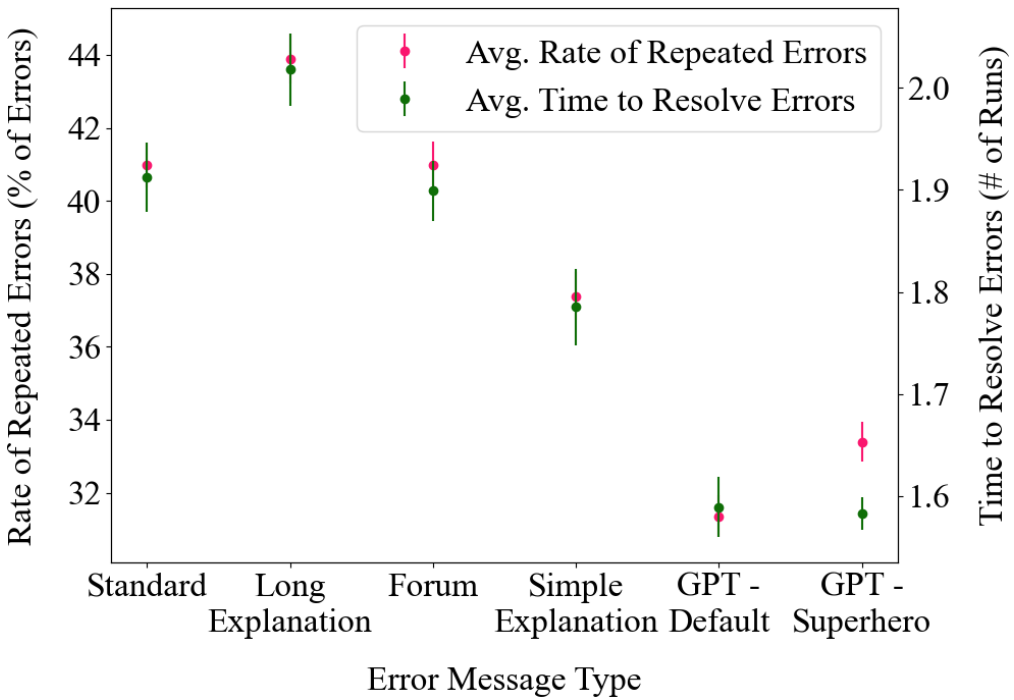
Featured in

SCIENTIFIC AMERICAN

AI Realtime Feedback

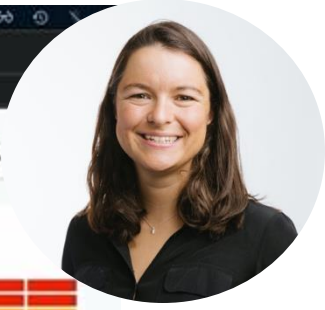






Both Accepted: SIGCSE 2024

Grading Creative Projects



Chrome File Edit View History Bookmarks Profiles Tab Window Help

DreamApp Grading Interface

Not Secure | iris-ws-6:3000/grading_display#

demo_student_2 Not Graded

Assignment: Breakout

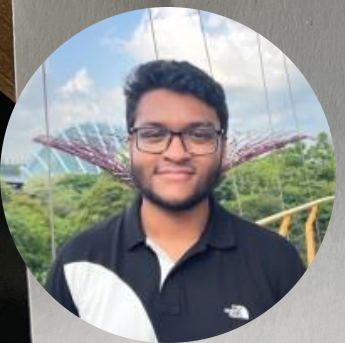
Mouse movement	4/4	
Brick drawing	5/5	
Paddle drawing	4/4	
Ball drawing	4/4	
Constants	0/2	
Wall bouncing	3/3	
Paddle bouncing	2/2	
Paddle skewering	0/1	
Ball does not become skewered on paddle (1)		
New life	4/4	
Total Score	37/40	

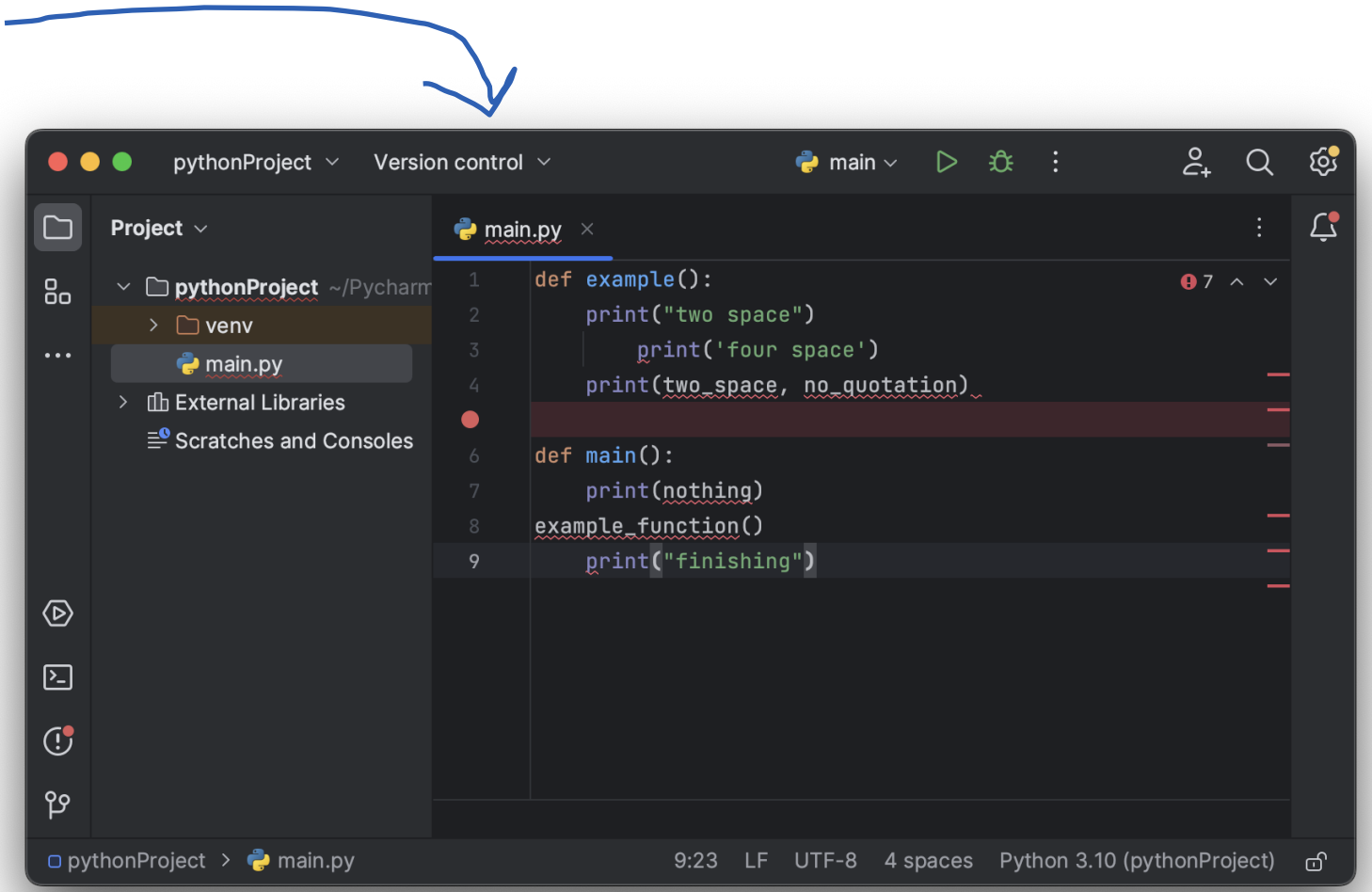
?

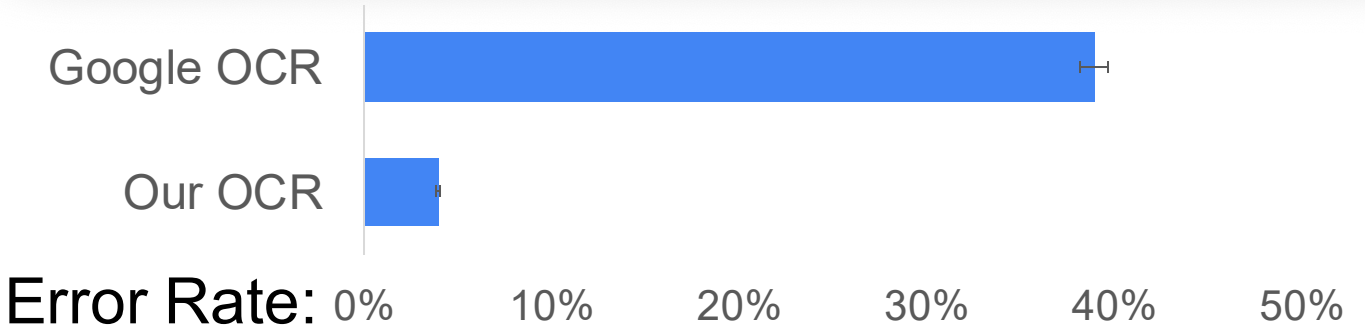
Video Code Demo

SIGCSE 2024

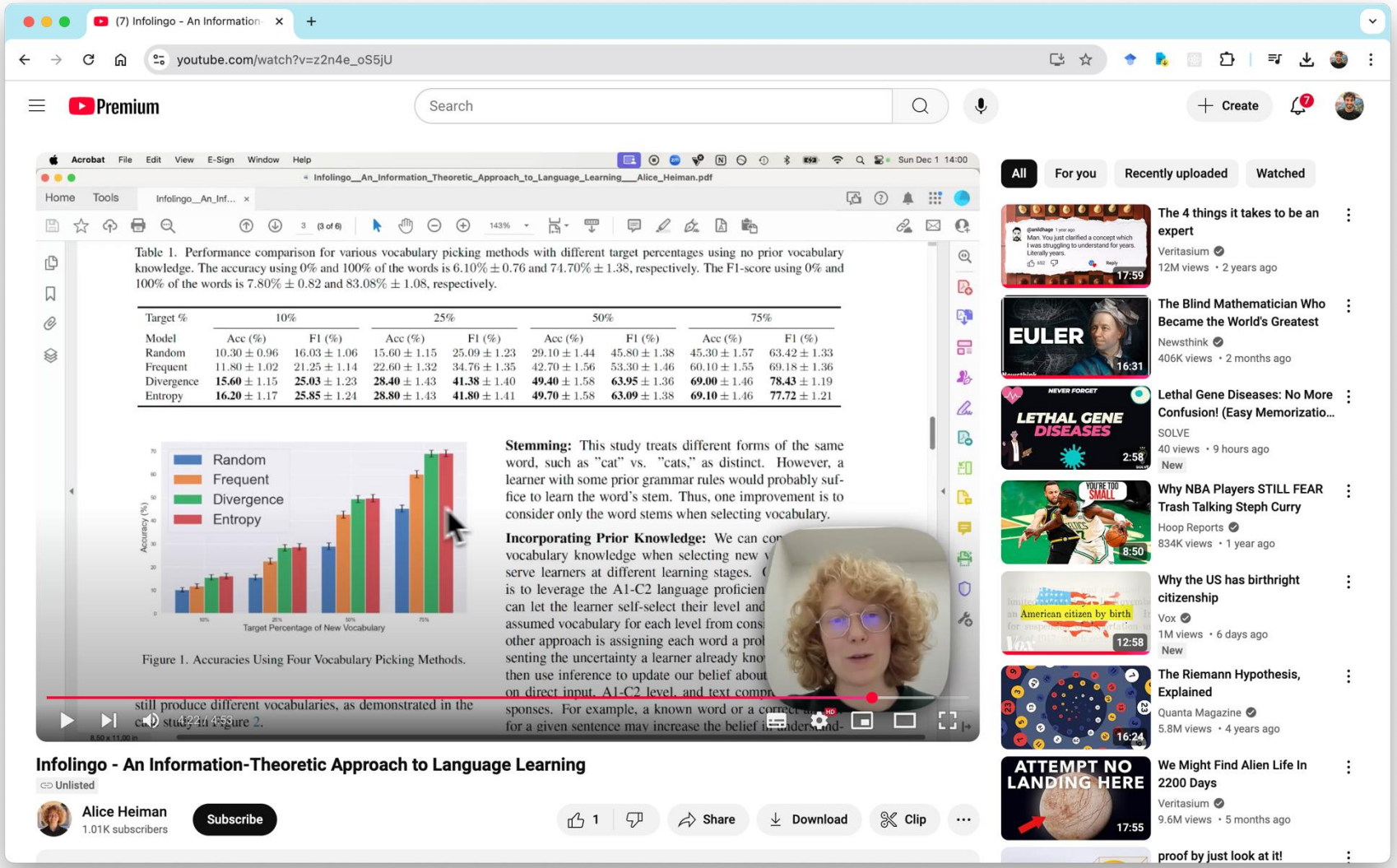
```
def example():
    -> print("two space")
    -> -> print('four space')
    -> print(two space, no quotation)

def main():
    -> print(nothing)
example_function()
    -> print("finishing")

if __name__ == "__main__":
    main()
```





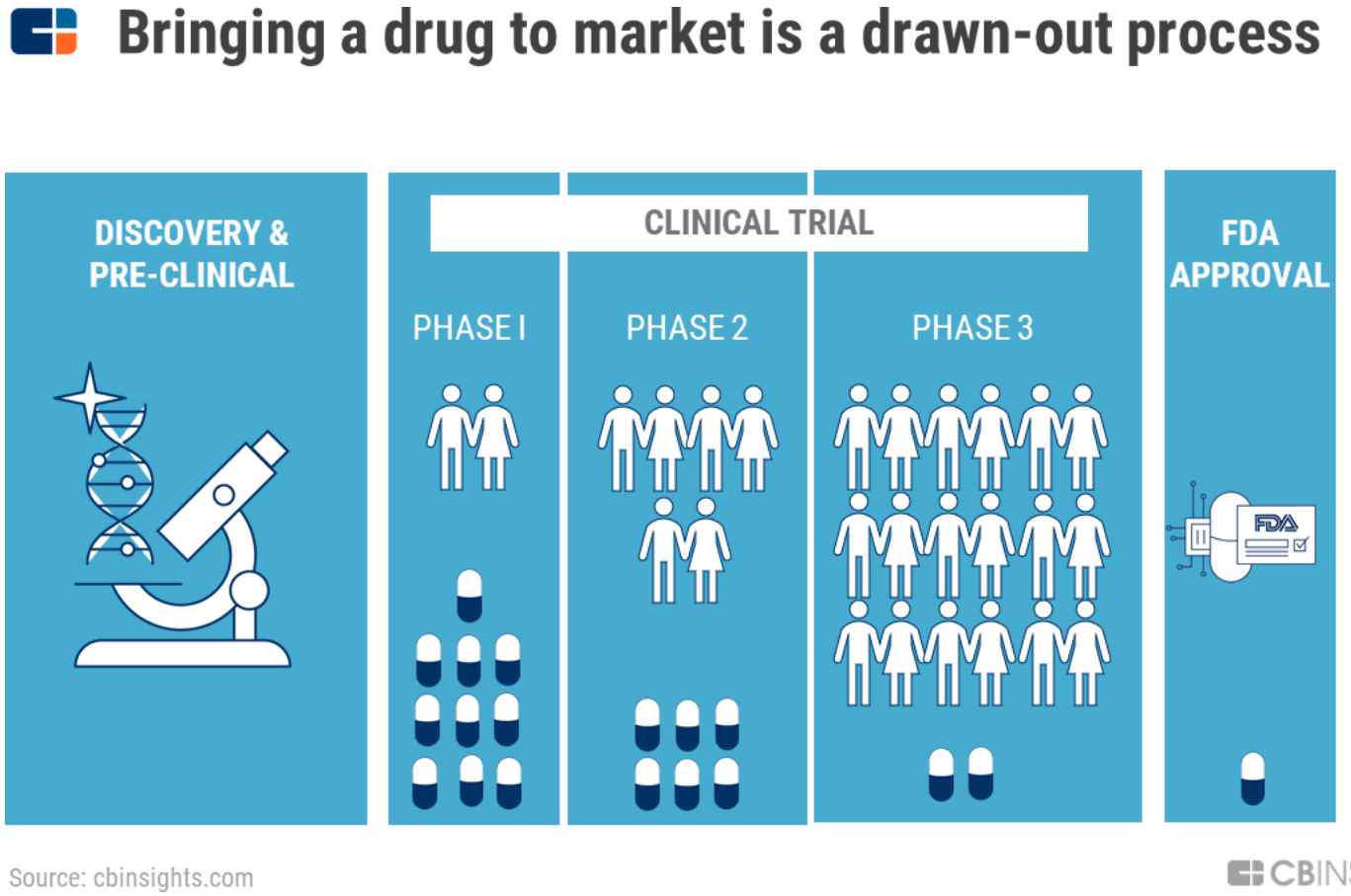


Information Theoretic Learning!

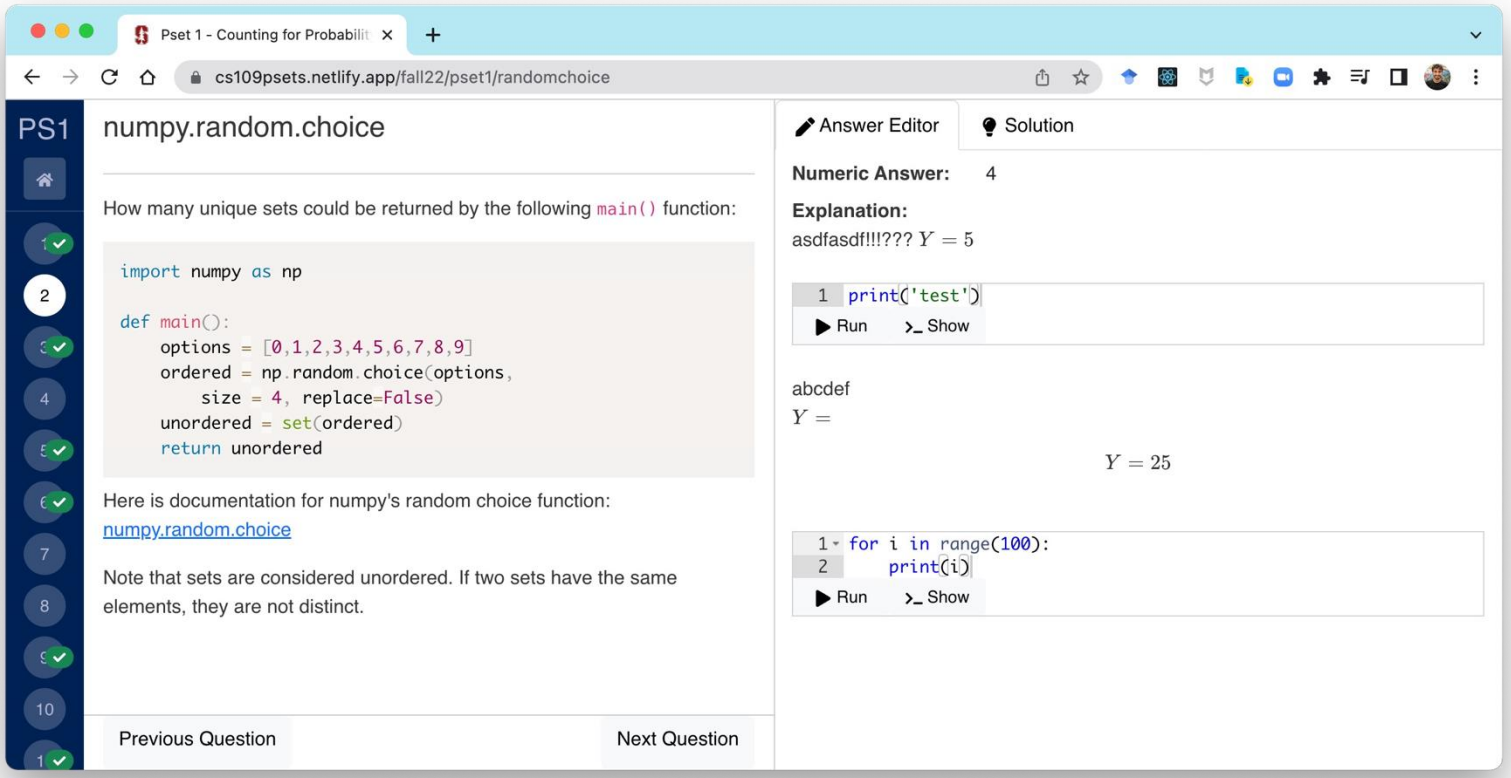


More than Education

More than education



Chose k examples from a dataset of lots of student work

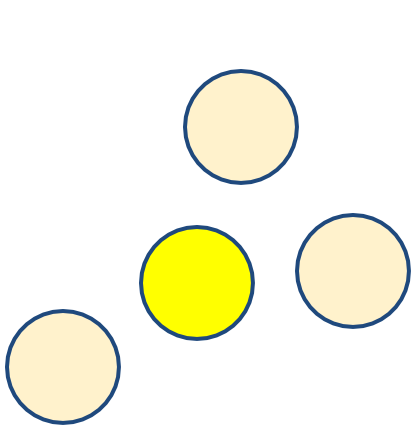


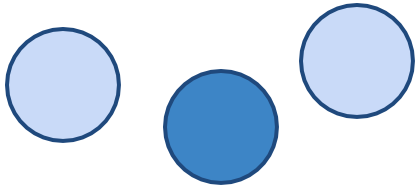
300 solutions

Find the 10 solutions which are most representative

K Medioids: A Classic Algorithm

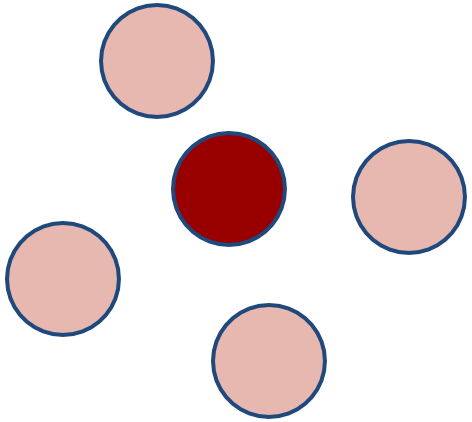
Chose the k nodes such that the sum of minimized distances is as small as possible





Before $O(n^2)$

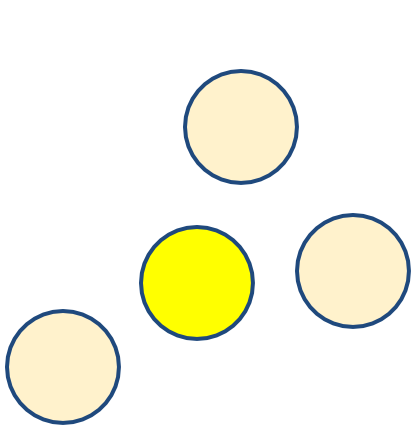
After $O(n \log n)$

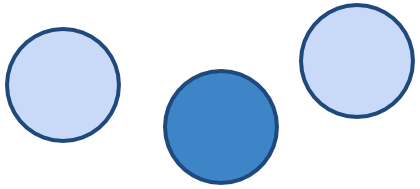


For any two points you can query the distance between them. May be non-Euclidean

Step 1: Chose the node closest to the rest

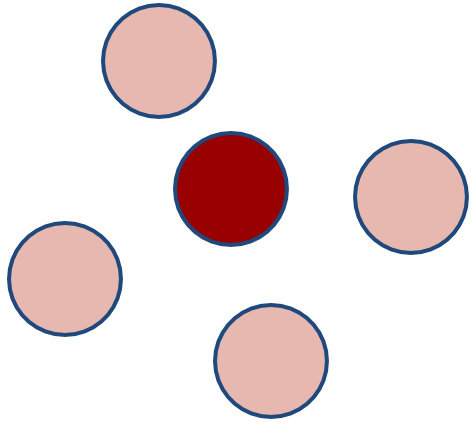
Chose the k nodes such that the sum of minimized distances is as small as possible





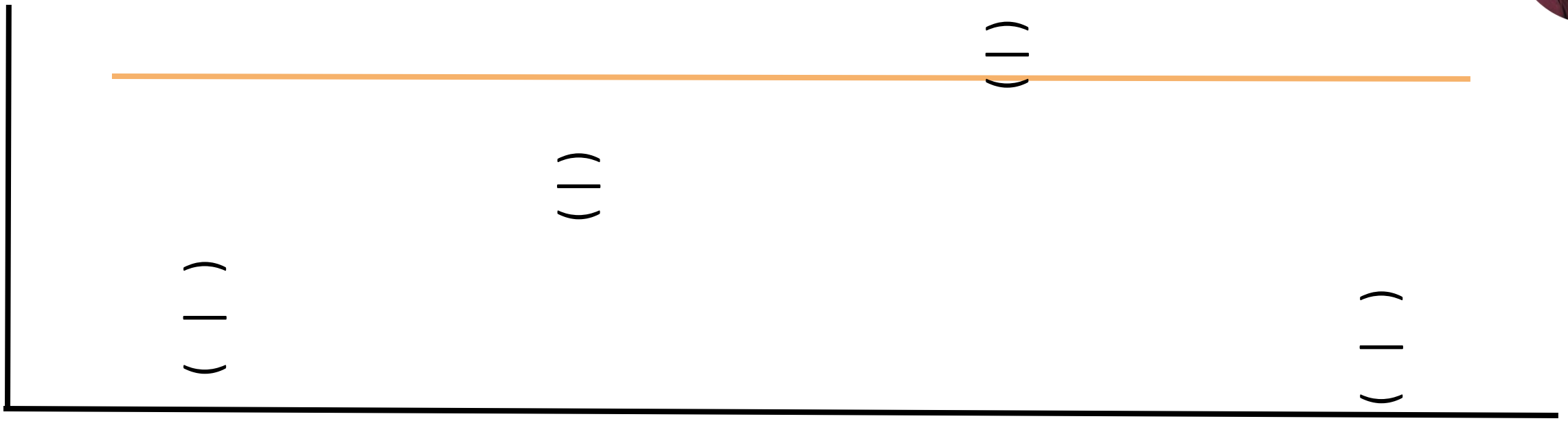
Before $O(n^2)$

After $O(n \log n)$



For any two points you can query the distance between them. May be non-Euclidean

The inner loop can be thought of as Thompson Sampling



3, 4

10, 11, 9, 8

12, 11, 10, 14

2, 0

Allowed us to revisit several core algorithms

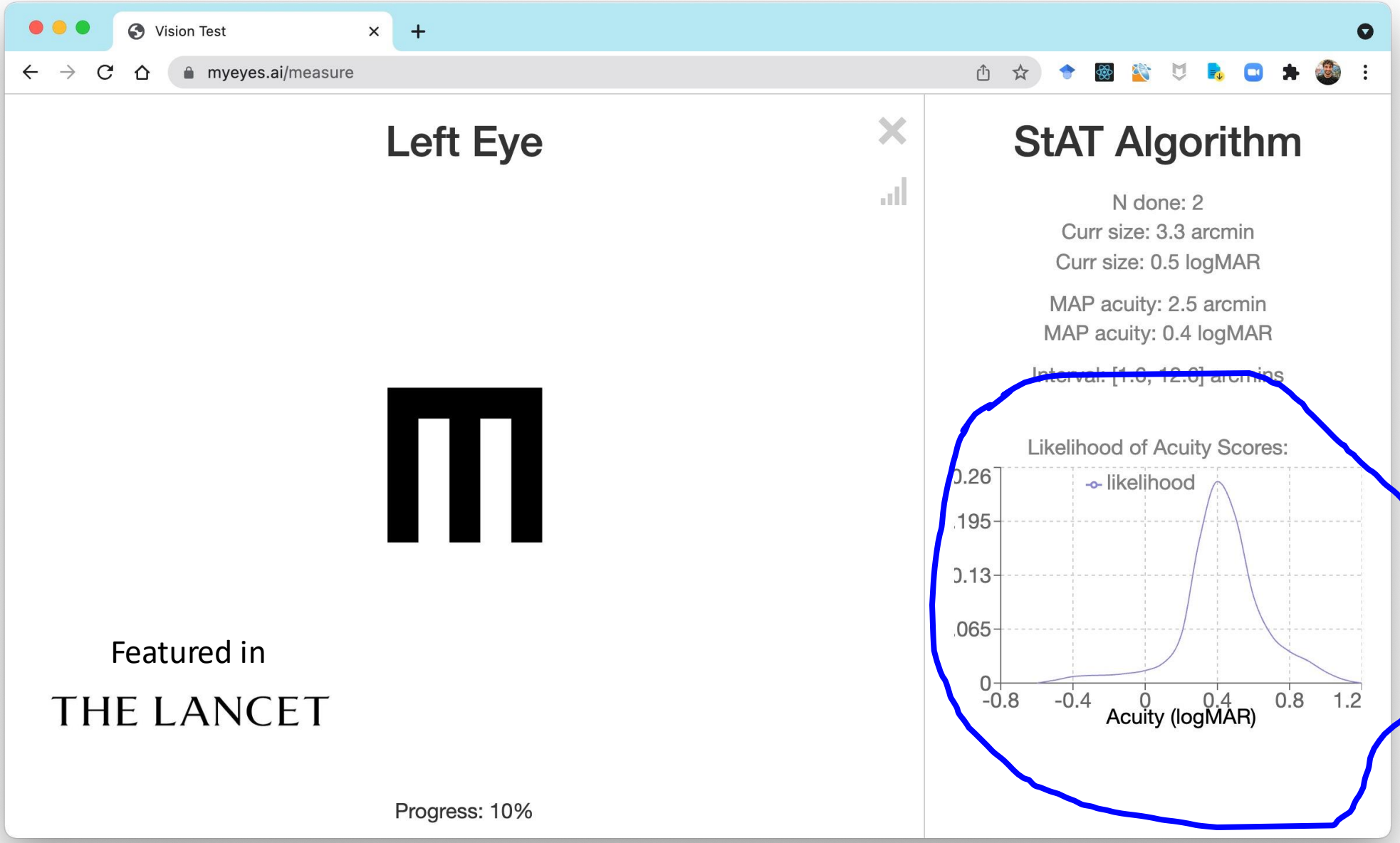
- *k*-Medoids
  - *BanditPAM: Almost Linear Time k-medoids Clustering via Multi-Armed Bandits*", NeurIPS 2020
- Random Forests
  - "*MABSplit: Faster Forest Training Using Multi- Armed Bandits*", NeurIPS 2022
- Maximum Inner Product Search
  - "Faster Inner Product Search in High Dimensions", NeurIPS 2023

CS109 TA



People just miss the random variables

More than education



What else should be a **random variable**?

Chris 2017:Ability to See??

What else should be a **random variable**?

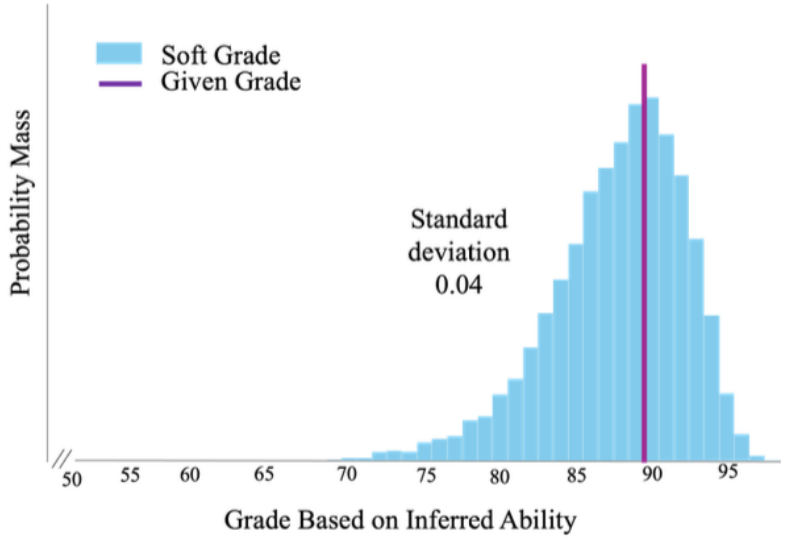
Chris 2023: Grades??

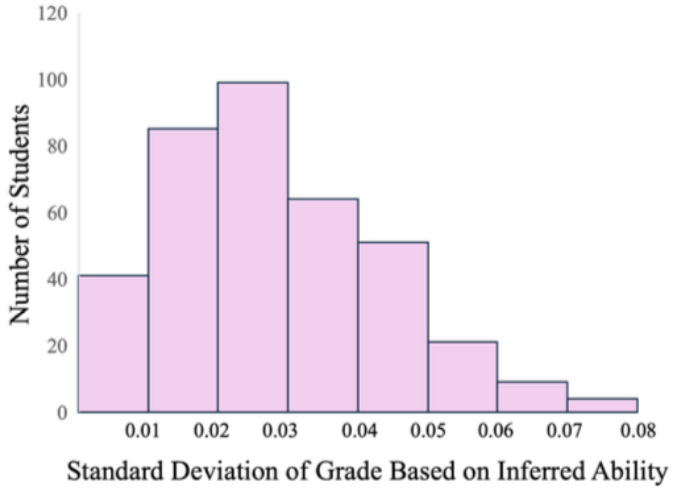
What else should be a **random variable**?

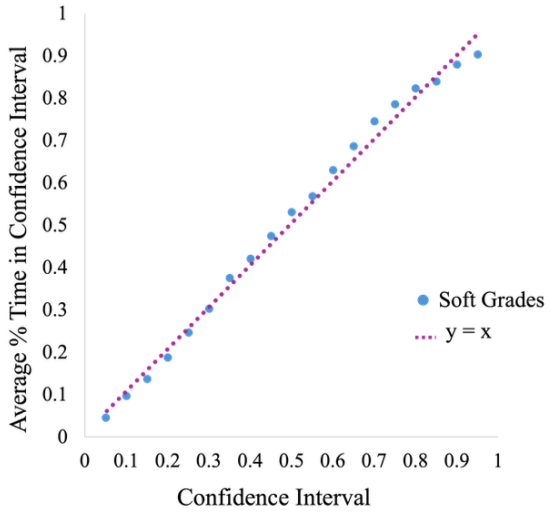
Juliette 2024: Grades!

Our Head TA!

Soft Grades. Accepted for publication in 2025







State of the Art imputation of grades, among other uses

	C1	C2	OULAD-1	OULAD-2	OULAD-3	OULAD-4	OULAD-5	OULAD-6	OULAD-7
Soft Grades (RMSE)	**0.042**	**0.058**	**0.069**	**0.067**	**0.062**	**0.089**	**0.071**	**0.046**	**0.076**
CRM Baseline (RMSE)	0.060	0.264	0.079	0.221	0.071	0.145	0.080	0.401	0.349

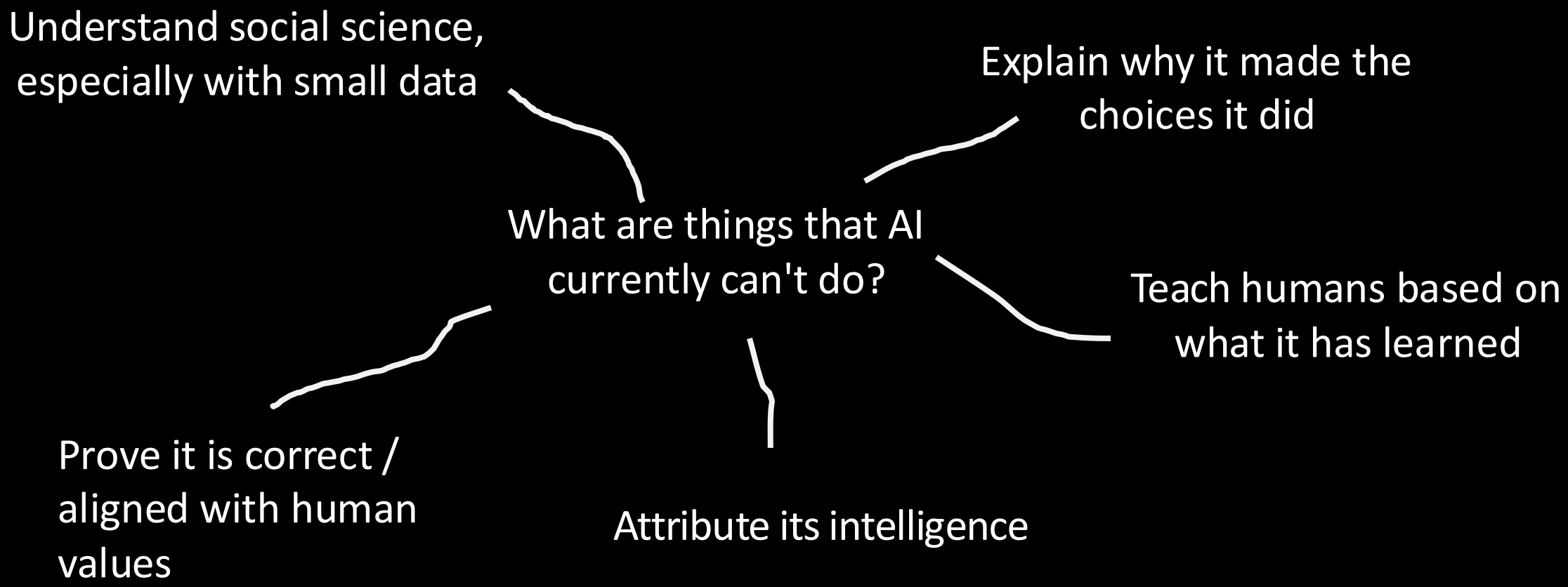
Table 3. Root Mean Square Error (RMSE) comparison between Soft Grades and CRM Baseline across all courses. Lower RMSE is better.

Juliette Woodrow

What else should be a **random variable**?

PSet Timing???
Travel Timing???
Stock Prices???
(You could do better!)

Application -> Theory



What should you do next?

Go solve amongst the abundance of important problems

Final Project | AA228/CS238

web.stanford.edu/clas...

Stanford University

AA228/CS238

Decision Making under Uncertainty

MENU

Final Project

The objective of the final project is to explore topics in decision making under uncertainty in greater depth than is permitted in class. The choice of topic is up to you, but it should be related to the general themes of the course. As part of the project you should:

- *describe* an approach (existing or newly developed),
- *apply* the approach to a problem of interest (which may or may not be related to aerospace), and
- *analyze* the performance of the approach according to a set of metrics.

CS221: Artificial Intelligence: P

stanford-cs221.github...

CS221: Artificial Intelligence: Principles and Techniques

Stanford / Autumn 2022-2023

[Calendar] [Modules] [Coursework] [Schedule]

- Lectures: Mon/Wed 1:30-2:50pm in NVIDIA Auditorium.
- Problem sessions: Fridays 1:30-2:20pm in Huang 018.
- Office hours, homework parties: see the Calendar.
- To contact the teaching staff, please use Ed; for more personal/sensitive matters, email cs221-aut22-23-lead-staff@lists.stanford.edu.

Teaching Staff

Percy Liang
Instructor

Dorsa Sadigh
Instructor

CS229: Machine Learning

cs229.stanford.edu

CS229

CS229: Machine Learning

Instructors

Andrew Ng

Moses Charikar

Carlos Guestrin

Course Description This course provides a broad introduction to machine learning and statistical pattern recognition. Topics include: supervised learning (generative/discriminative learning, parametric/non-parametric learning, neural networks, support vector machines); unsupervised learning (clustering, dimensionality reduction, kernel methods); learning theory

CS 228 - Probabilistic Graphic

ermongroup.github.io/...

CS 228 - Probabilistic Graphical Models

Winter 2021-22

Ed | Calendar | Course Notes

Logistics | Course Info | Syllabus | Other Resources

Logistics

- **Lectures**: Tue, Thu, 9:45am-11:15am, Nvidia Auditorium
- **Office Hours and Sections**: Google Calendar

Statistics 200: Introduction to

web.stanford.edu/clas...

Statistics 200: Introduction to Statistical Inference

Zhou Fan, Stanford University, Autumn 2016

Home

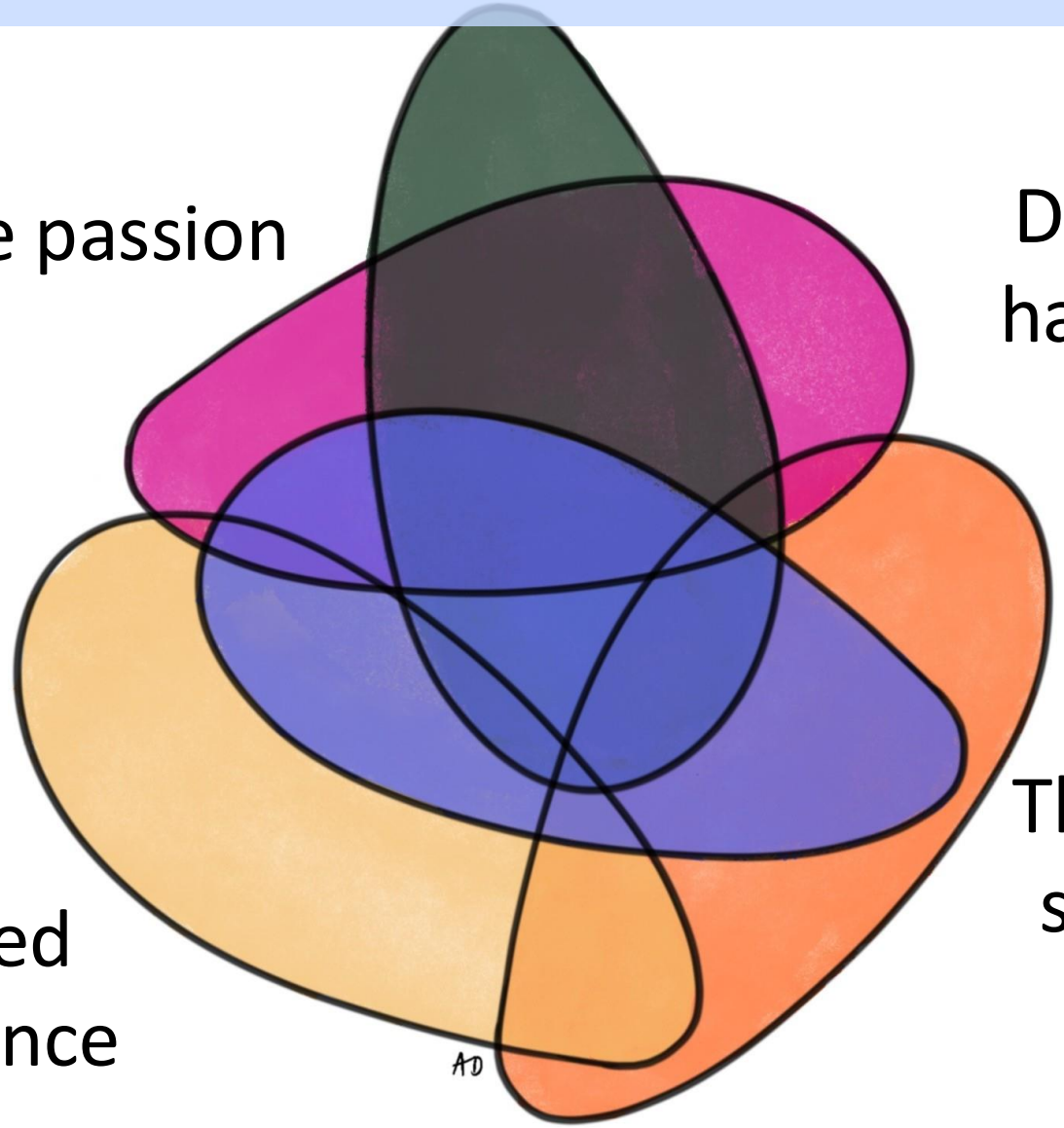
Lectures

Homework

Grades

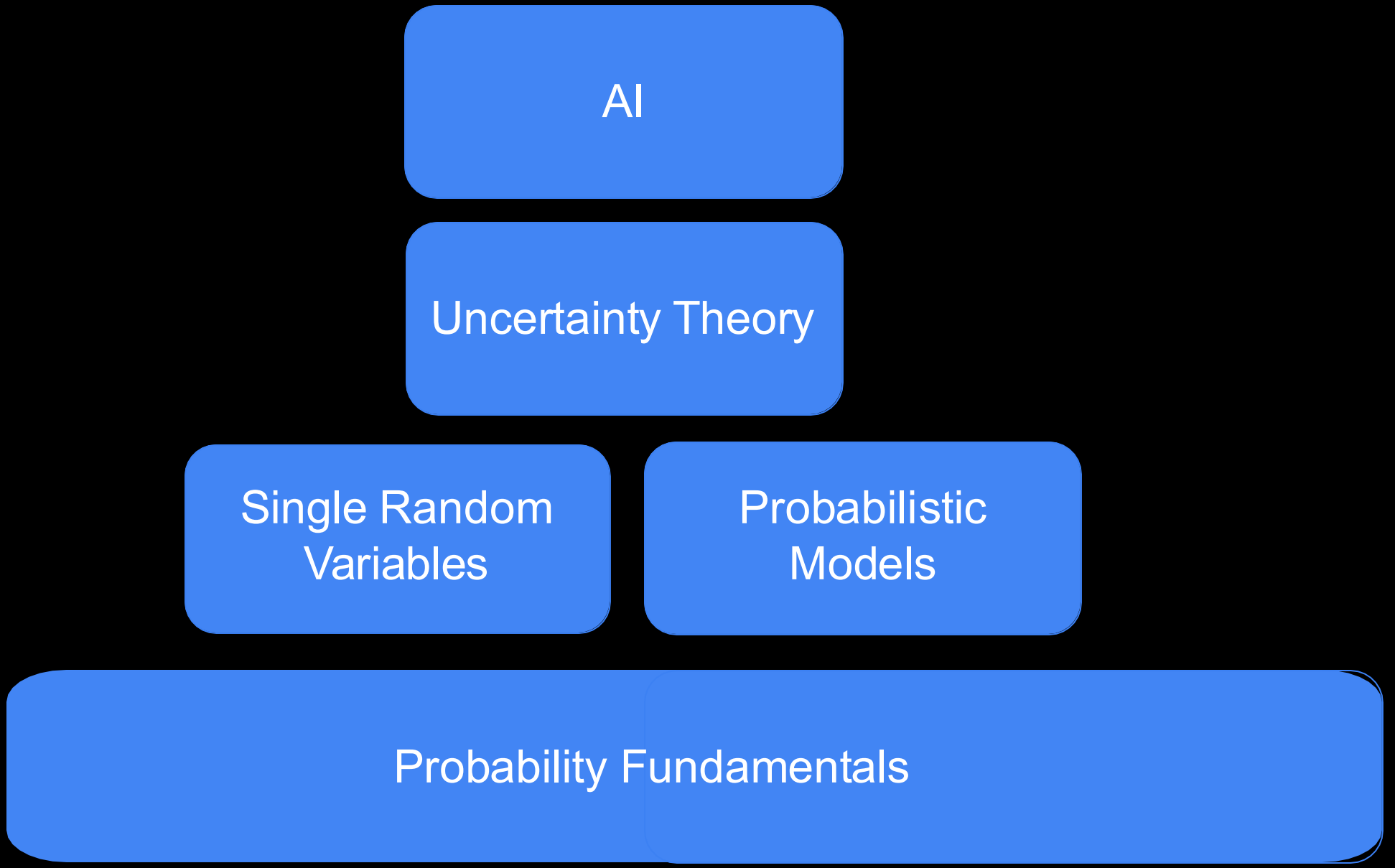
Piazza

Think about intersectionality



Last Class…

CS109


AI
Uncertainty Theory
Single Random Variables
Probabilistic Models
Probability Fundamentals

COURSE VALUES

Everyone is welcome. Intellectual joy. Be kind. Be humane. Social connection. Learn by doing. Thrill of building. Adapt to new contexts.

Fantastic Teaching Team







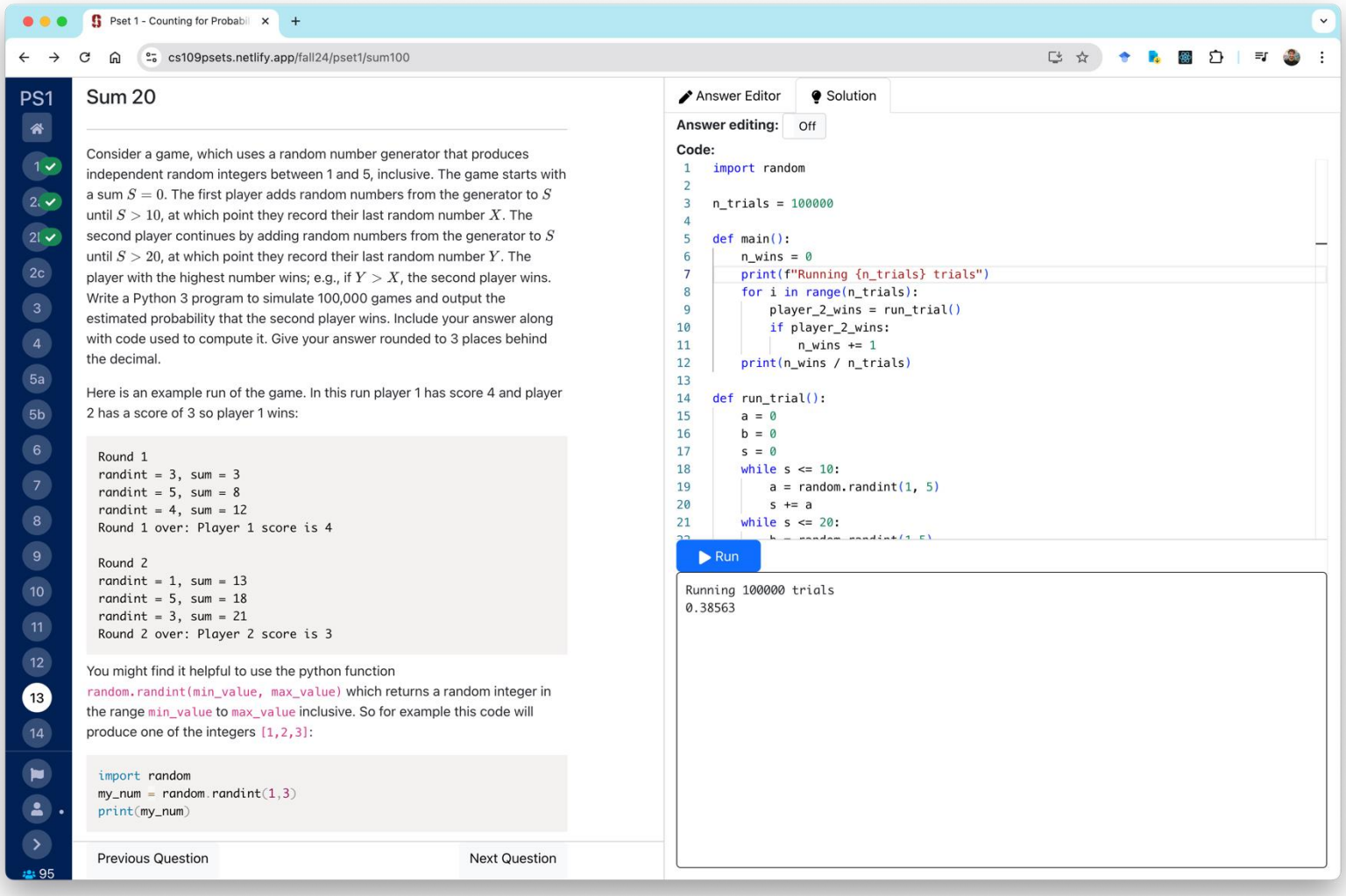




What is a Probability?

$$P(E) = \lim_{n \to \infty} \frac{n(E)}{n}$$

What is a Probability



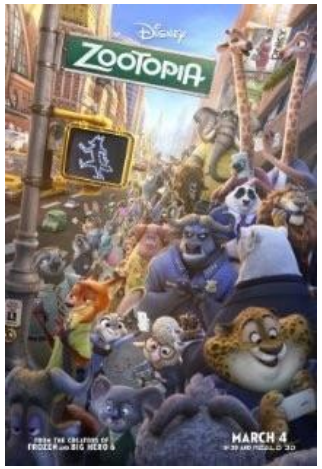
Netflix and Learn

$$P(E|F) = \frac{P(EF)}{P(F)}$$ **Definition of Cond. Probability**

- Let E be the event that a user watches the given movie.
- Let F be the event that the same user watches CODA (2021).

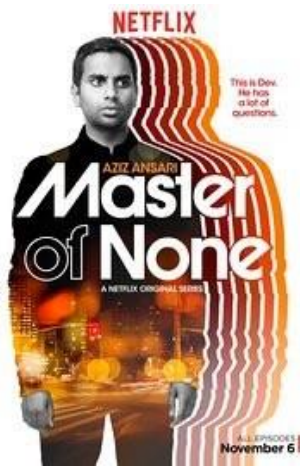
$P(E) = 0.19$

$P(E|F) = 0.14$



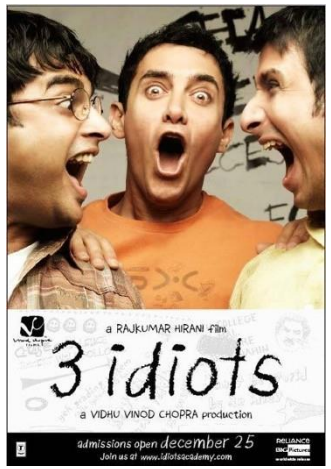
$P(E) = 0.32$

$P(E|F) = 0.35$



$P(E) = 0.20$

$P(E|F) = 0.20$



$P(E) = 0.09$

$P(E|F) = 0.72$

$P(E) = 0.20$

$P(E|F) = 0.42$

Montey Hall Problem

Marilyn discovers the Probability Bug

WHEN YOU MEET YOUR BEST FRIEND

Somewhere you didn't expect to.

Trailing the dovetail shuffle to it’s lair – Persi Diaconosis

Zika Test

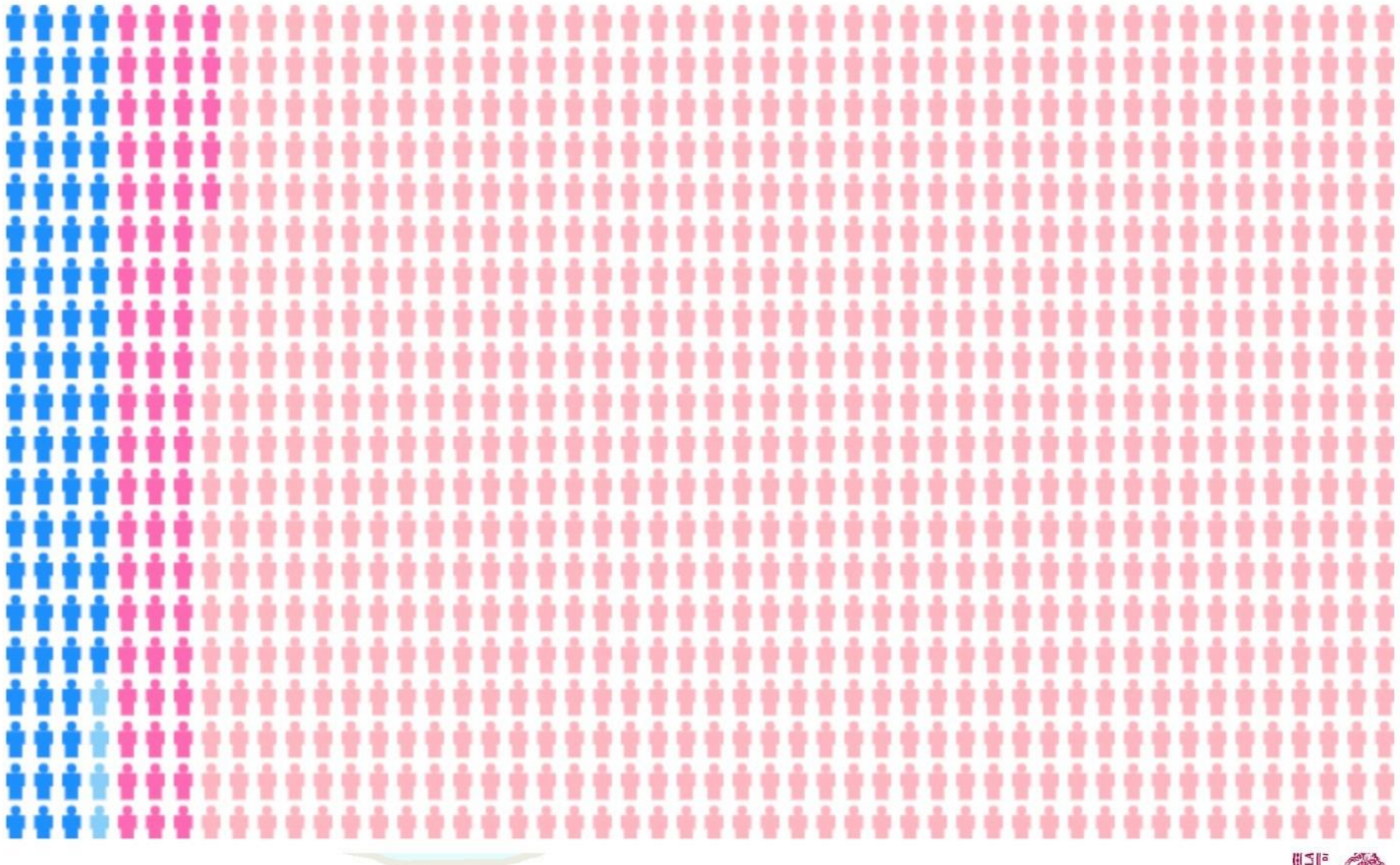
Positive Zika.

What is the probability of zika?

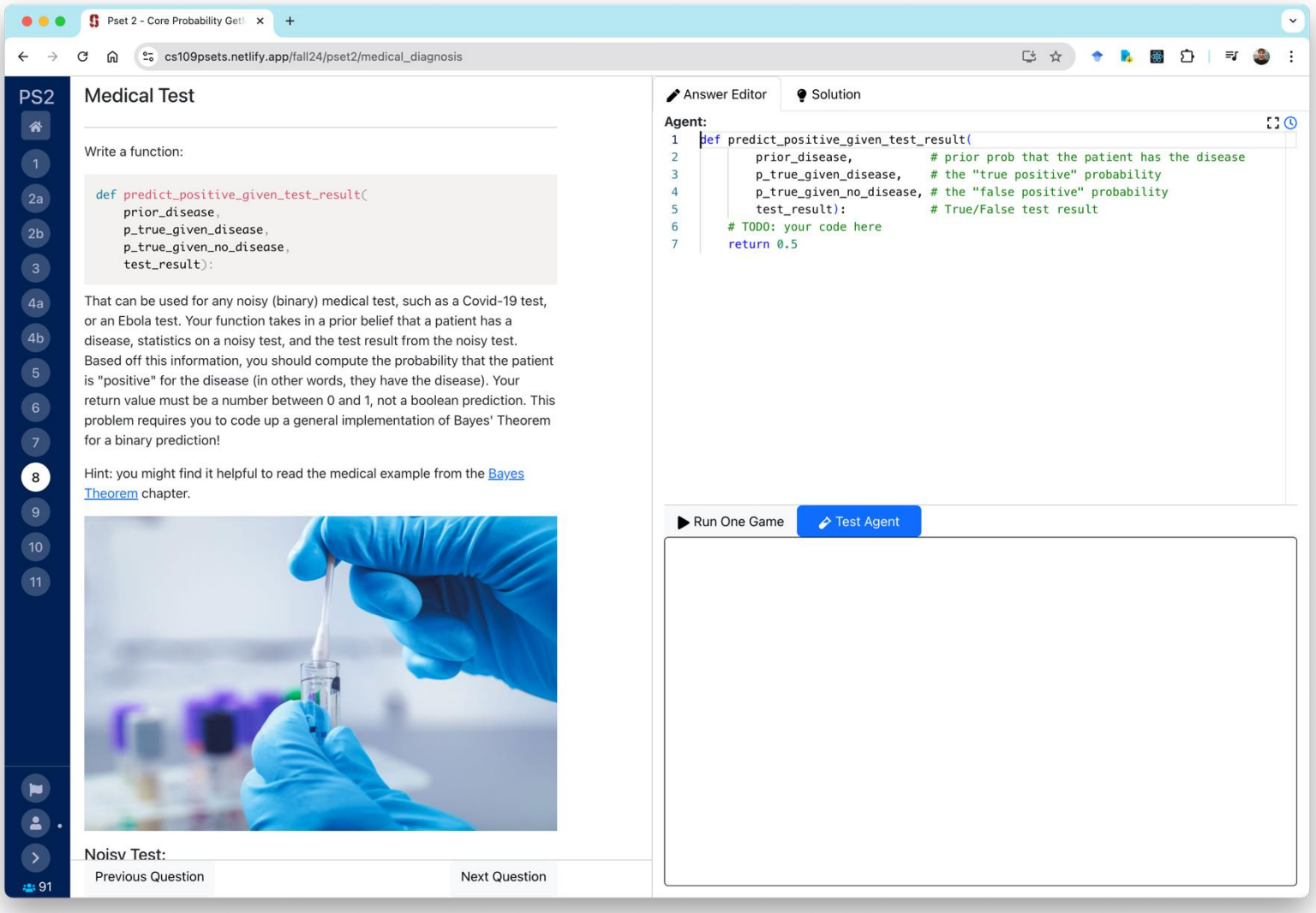
- *0.1% of people have zika*
- *90% positive rate for people with zika*
- *7% positive rate for people without zika*

The right answer is 1%

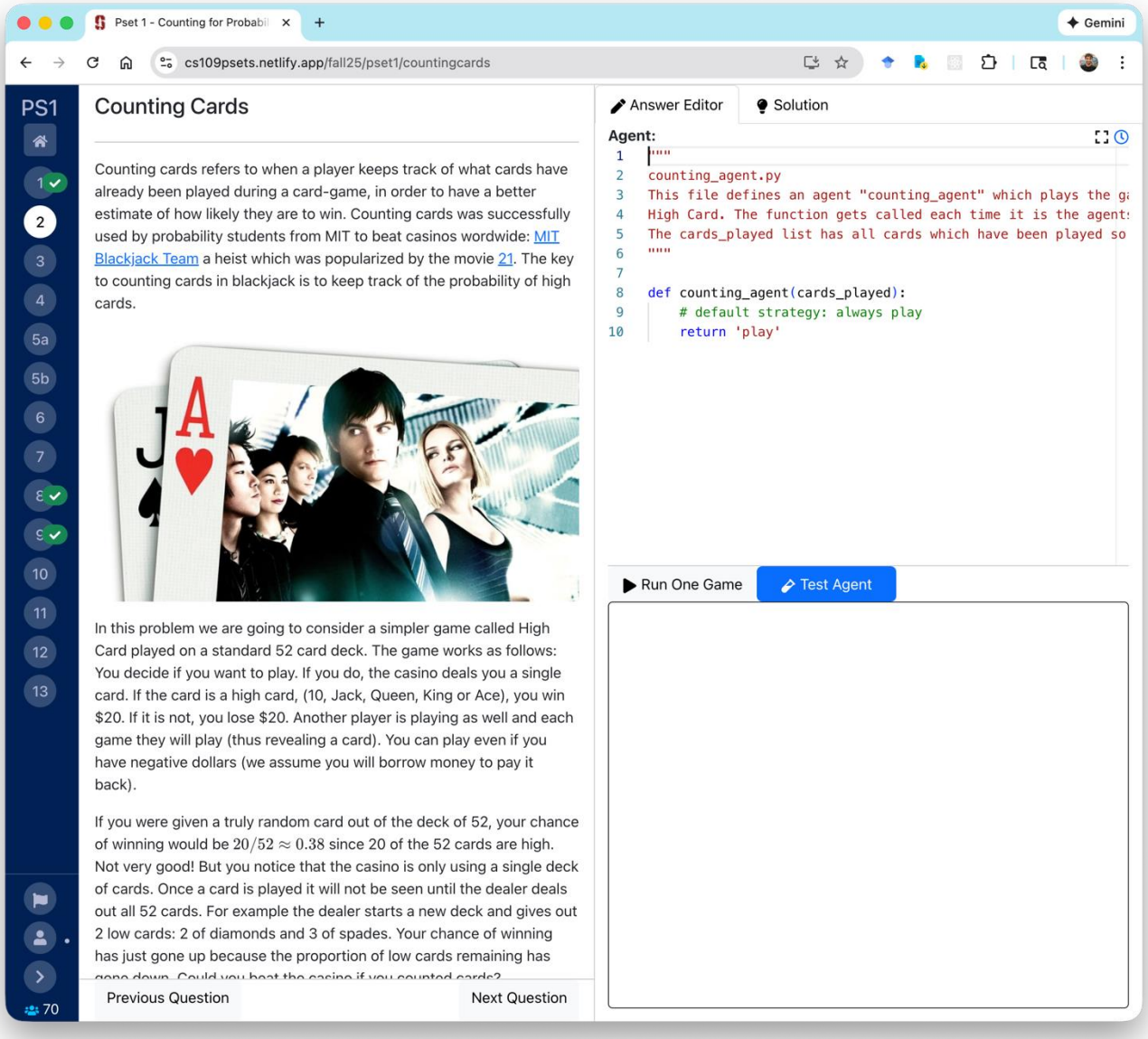
Bayes Theorem Intuition



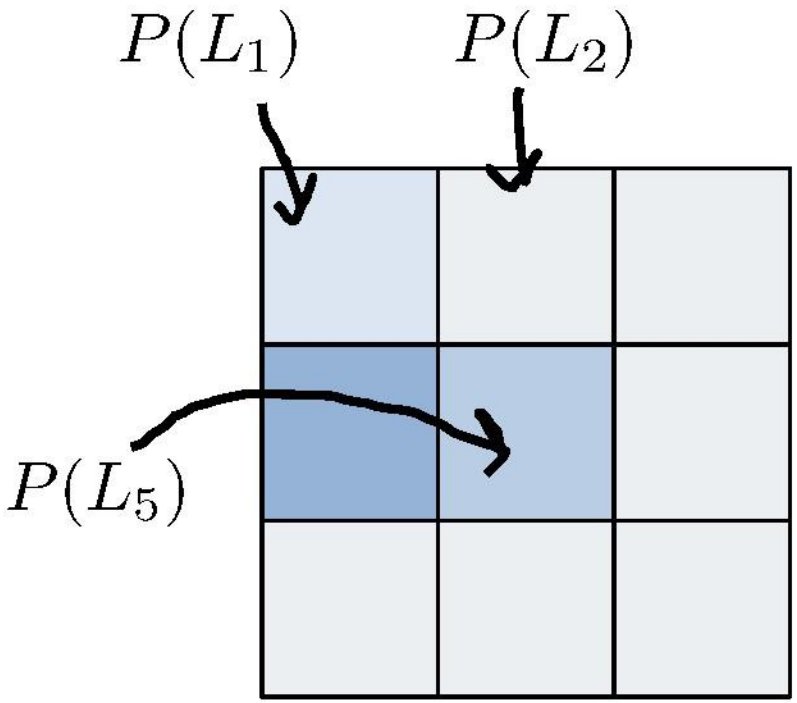
Program the General Version



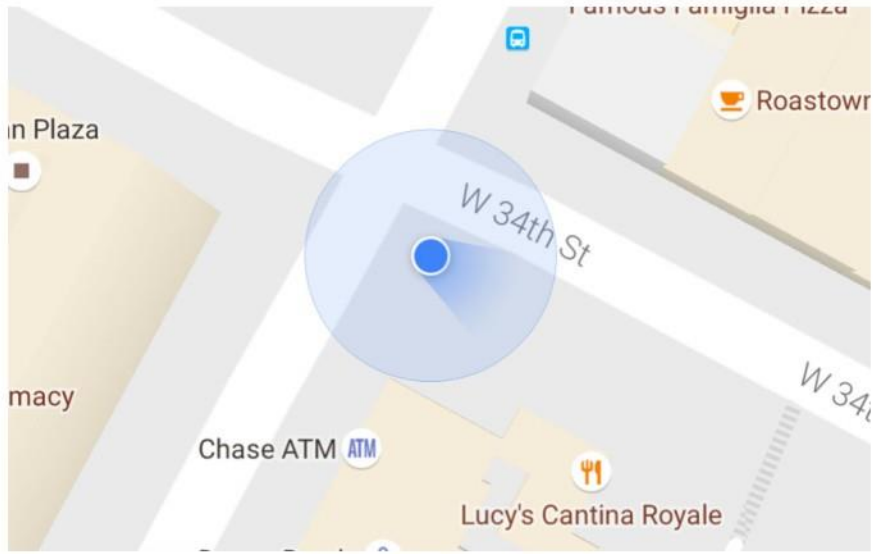
Counting Cards



Update Belief

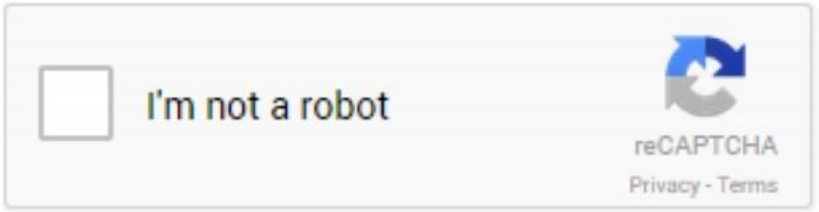


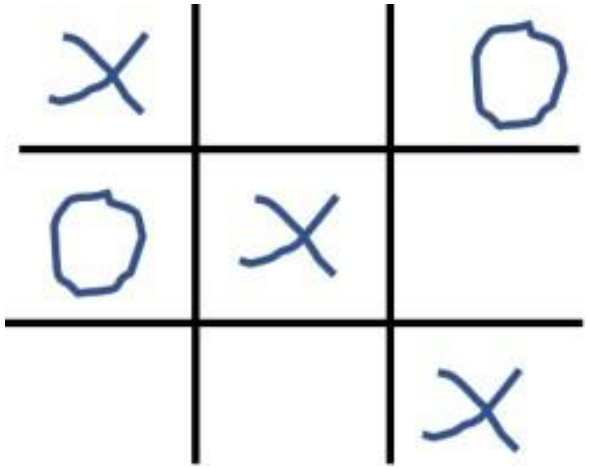
Before Observation



Recall our Ebola Bats

Fourth Year of Sections





Time to Start Flippin Coins

Exactly k heads

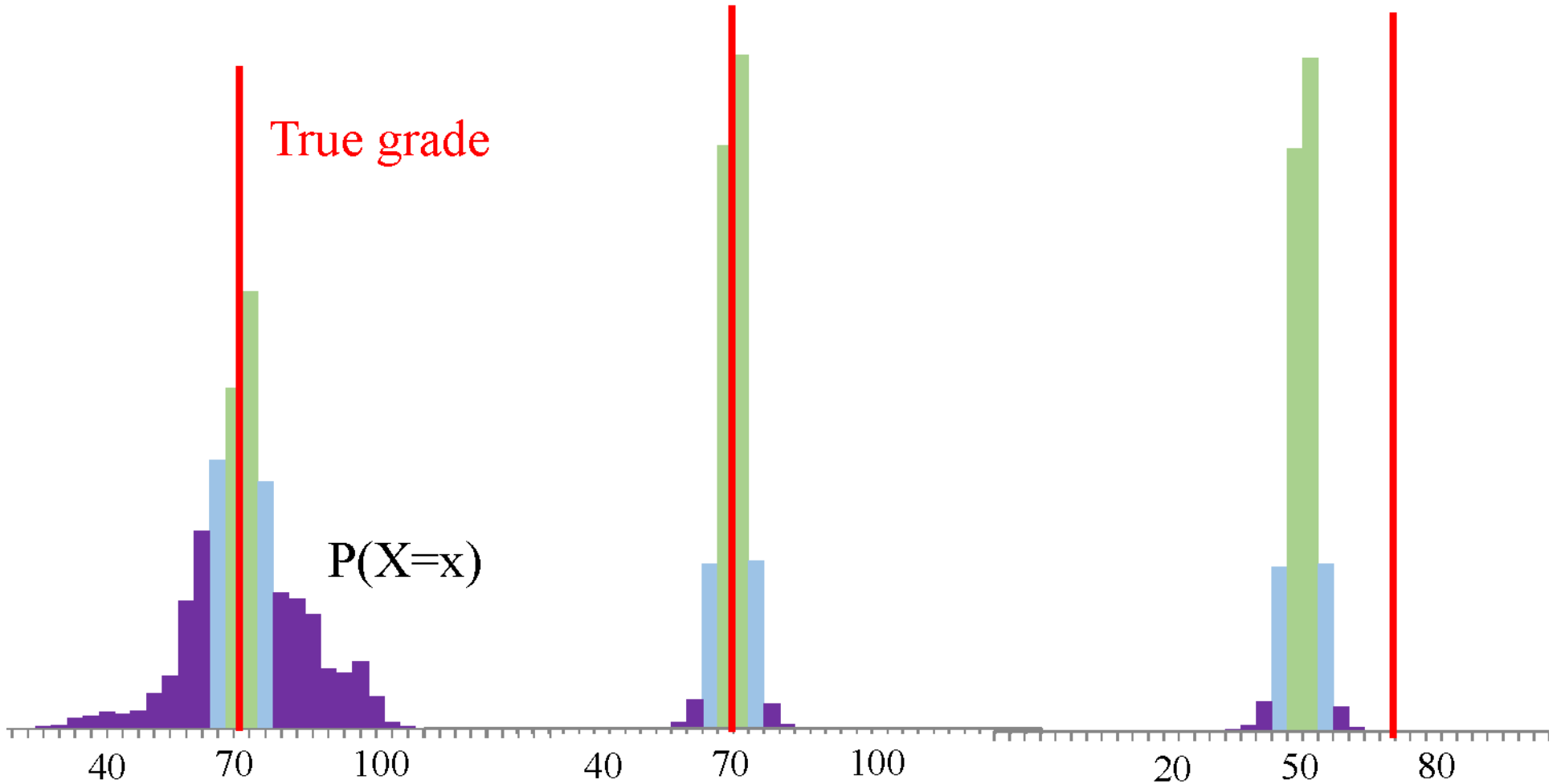
Next lets try to figure out the probability of exactly k heads in the n flips. Importantly we don't care where in the n flips that we get the heads, as long as there are k of them. Note that this question is different than the question of first k heads and then $n-k$ tails which requires that the k heads come first! That particular result does generate exactly k coin flips, but there are others.

There are many others! Let's ask the computer to list the ways we could generate exactly k heads within n coin flips. The output region is scrollable:

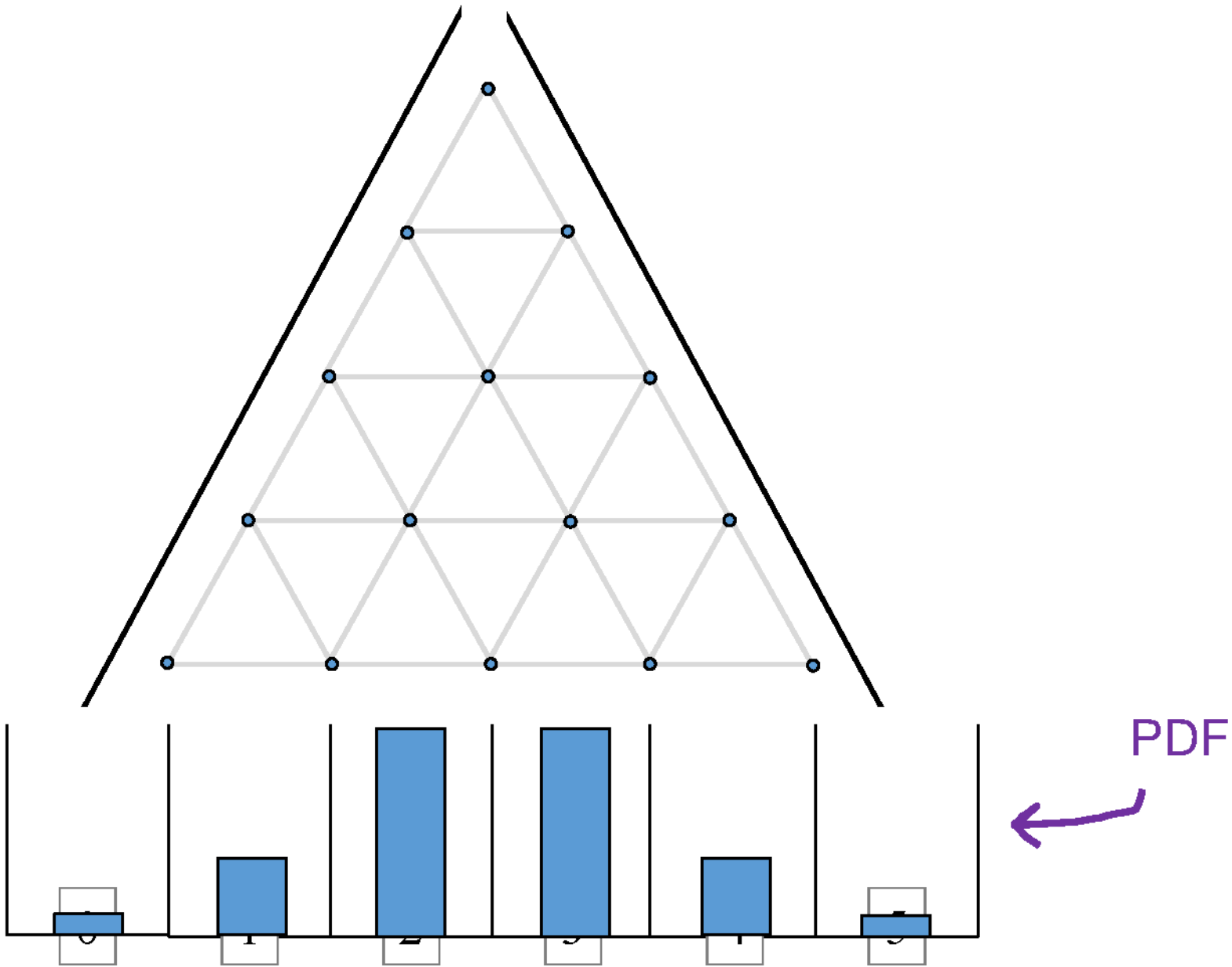
```
(H, H, H, H, T, T, T, T, T, T)
(H, H, H, T, H, T, T, T, T, T)
(H, H, H, T, T, H, T, T, T, T)
(H, H, H, T, T, T, H, T, T, T)
(H, H, H, T, T, T, T, H, T, T)
(H, H, H, T, T, T, T, T, H, T)
(H, H, H, T, T, T, T, T, T, H)
(H, H, T, H, H, T, T, T, T, T)
(H, H, T, H, T, H, T, T, T, T)
(H, H, T, H, T, T, H, T, T, T)
(H, H, T, H, T, T, T, H, T, T)
(H, H, T, H, T, T, T, T, H, T)
(H, H, T, H, T, T, T, T, T, H)
(H, H, T, T, H, H, T, T, T, T)
(H, H, T, T, H, T, H, T, T, T)
(H, H, T, T, H, T, T, H, T, T)
(H, H, T, T, H, T, T, T, H, T)
(H, H, T, T, H, T, T, T, T, H)
(H, H, T, T, T, H, H, T, T, T)
(H, H, T, T, T, H, T, H, T, T)
```

Random Variables

X is the score a peer grader gives to an assignment submission



Binomial



Geometric

Sequence 1:

TTHHTHTTHTTTHTTTTHTTTHTTHTHHTH

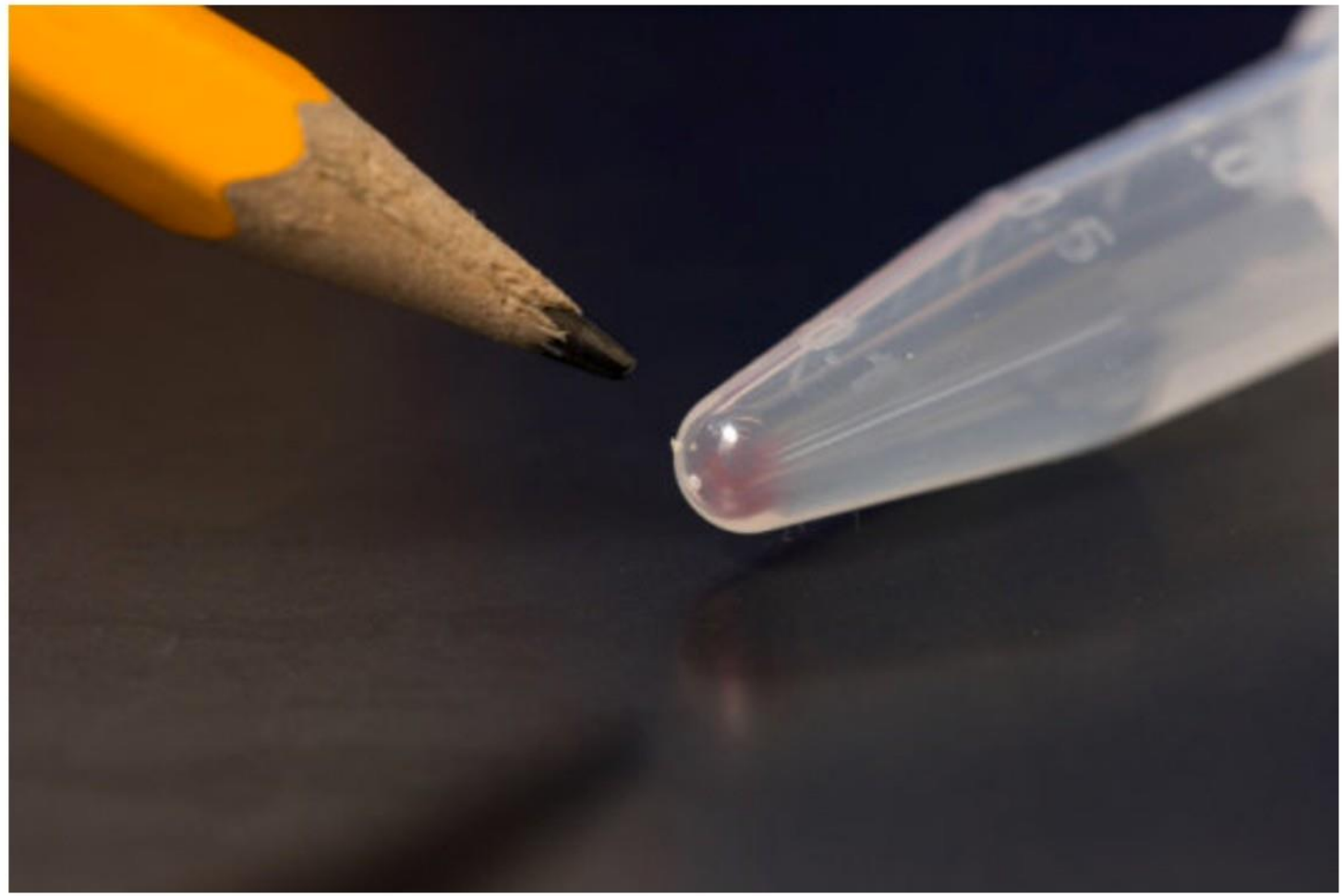
HTTHHTTHHHTHHTHTTHTHTTHHTHHHT

HTHHTHHHTHTHTTHTTHHTHTHTHTTHH

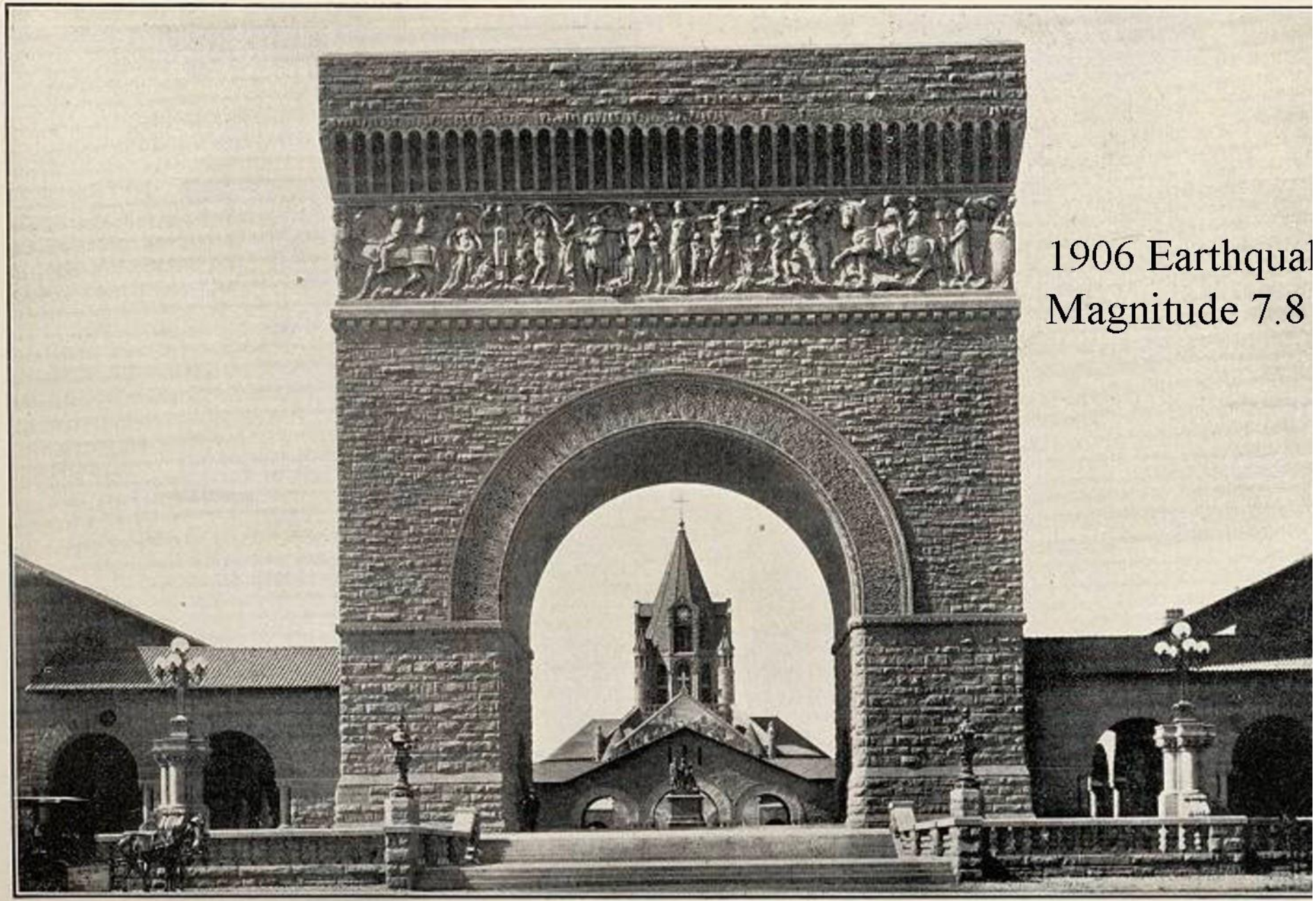
TTHHTHTTHTHTHTHTHTHTHHHTHTHTH

TTHHTHTHTHTHHHTTHHTHTTTHTHHTH

Storing Data on DNA

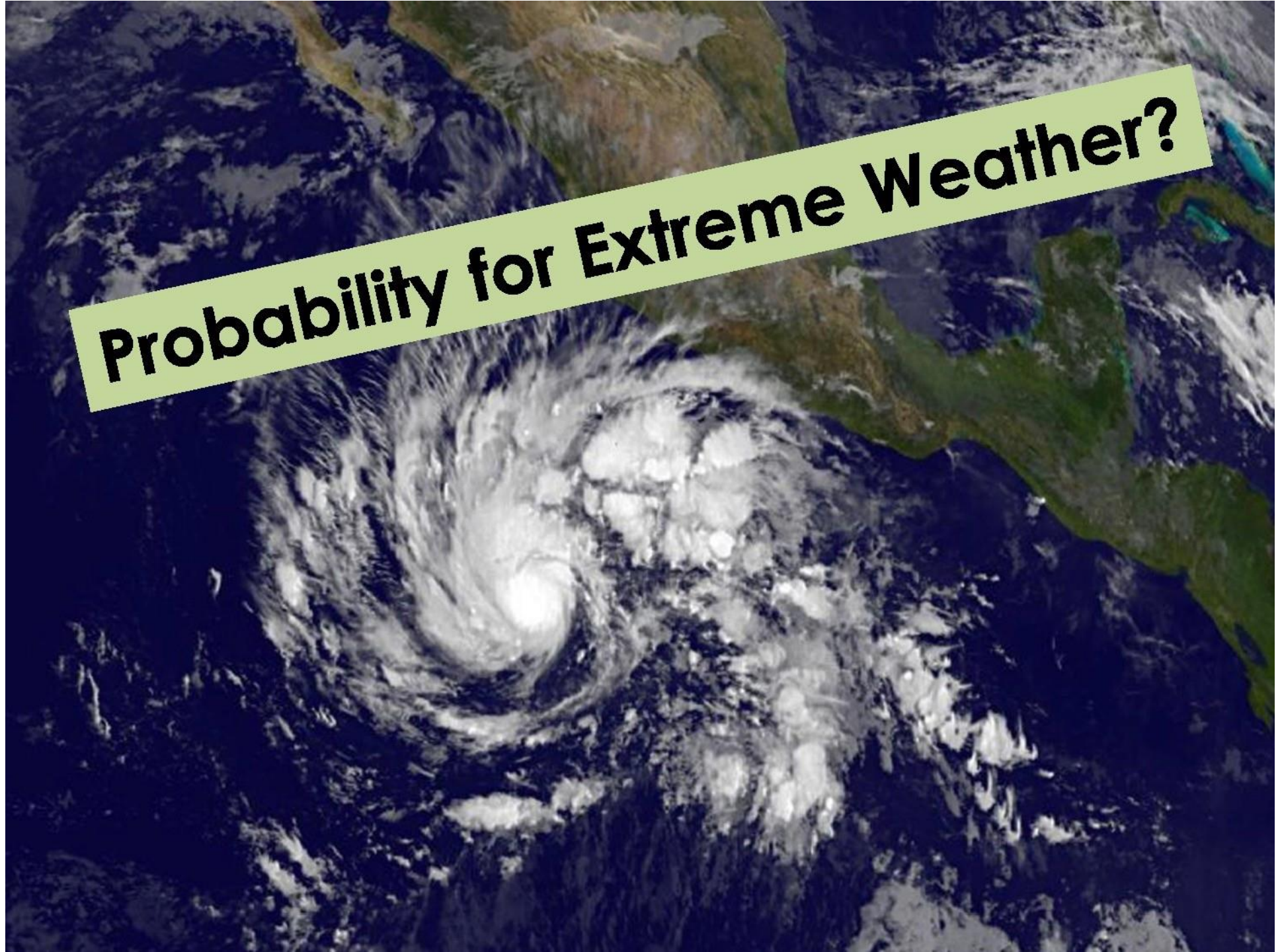


All the movies, images, emails and other digital data from more than 600 smartphones (10,000 gigabytes) can be stored in the faint pink smear of DNA at the end of this test tube.



1906 Earthqual
Magnitude 7.8

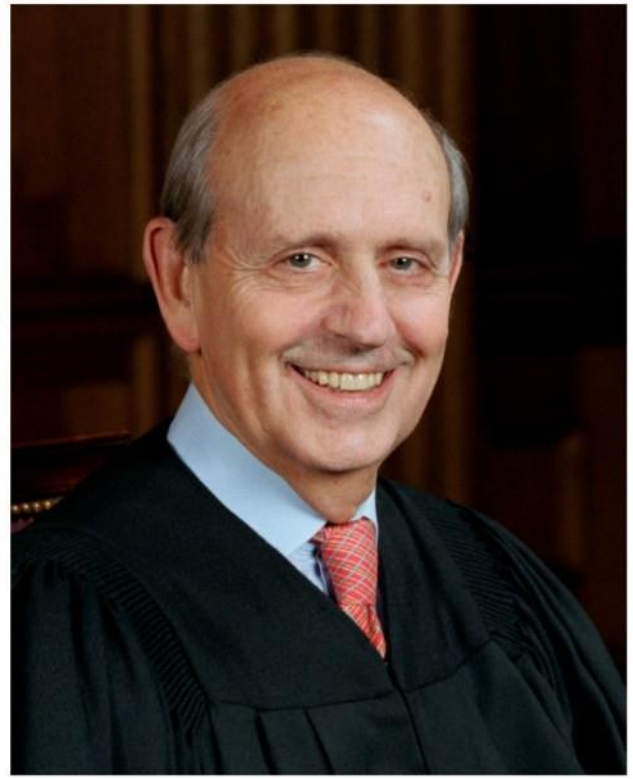
ILL. No. 65. MEMORIAL ARCH, WITH CHURCH IN BACKGROUND, STANFORD UNIVERSITY, SHOWING TYPES OF CARVED WO
WITH THE SANDSTONE.


Probability for Extreme Weather?

Bit Coin Mining

You "mine a bitcoin" if, for given data D, you find a number N such that Hash(D, N) produces a string that starts with g zeroes.

Representative Juries



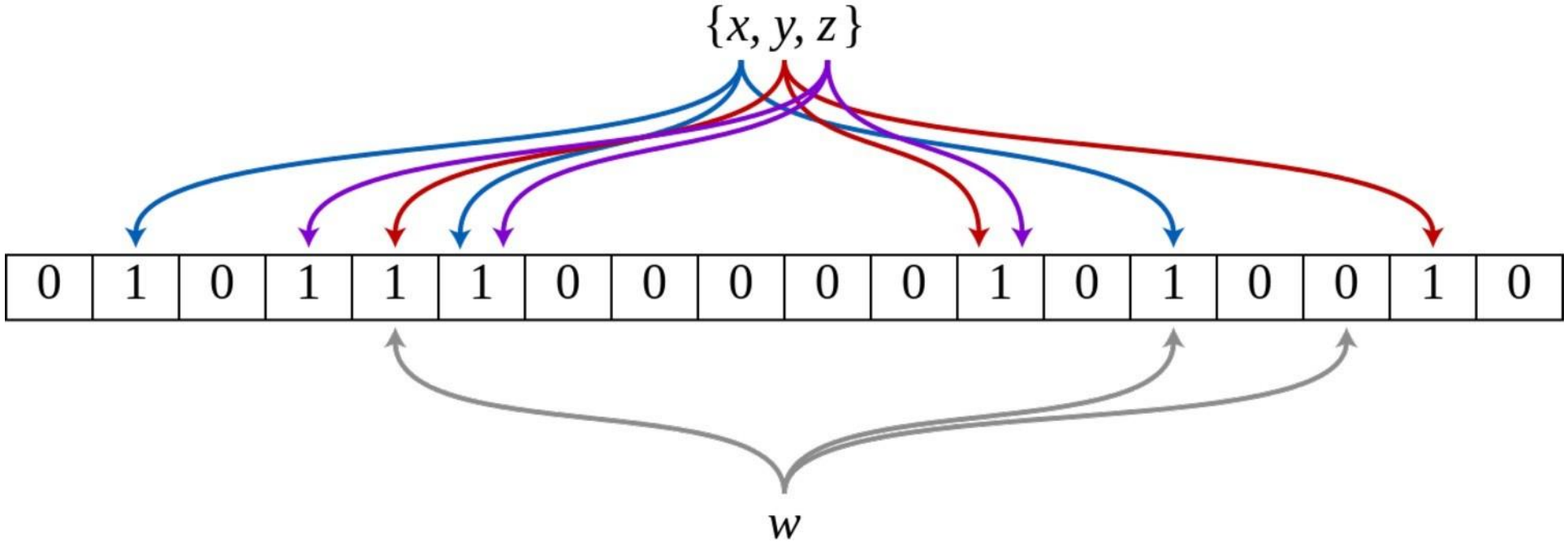
Simulate

Simulation:

Dating at Stanford

Each person you date has a 0.2 probability of being someone you spend your life with. What is the average number of people one will date? What is the standard deviation?

Bloom Filter

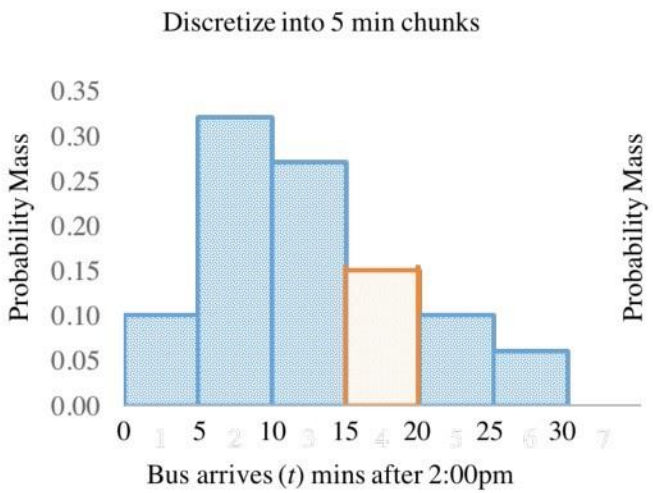


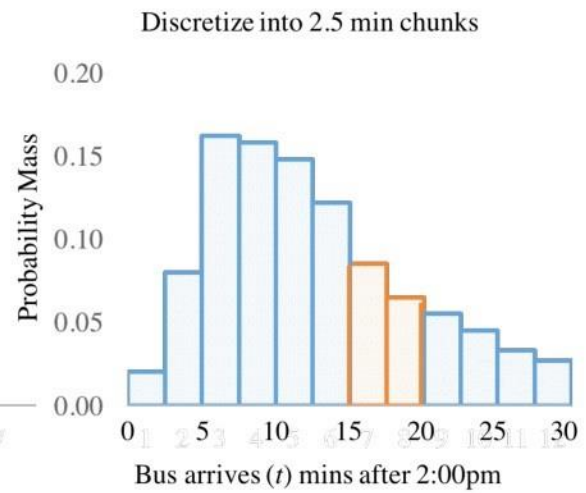
random()?

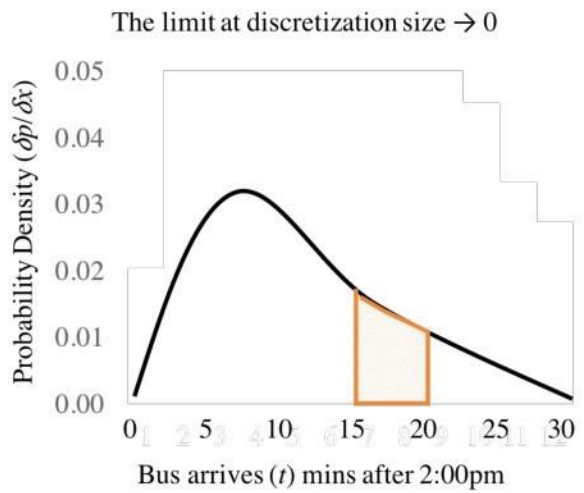
Riding the Marguerite

You are running to the bus stop. You don't know exactly when the bus arrives. You arrive at 2:20pm.

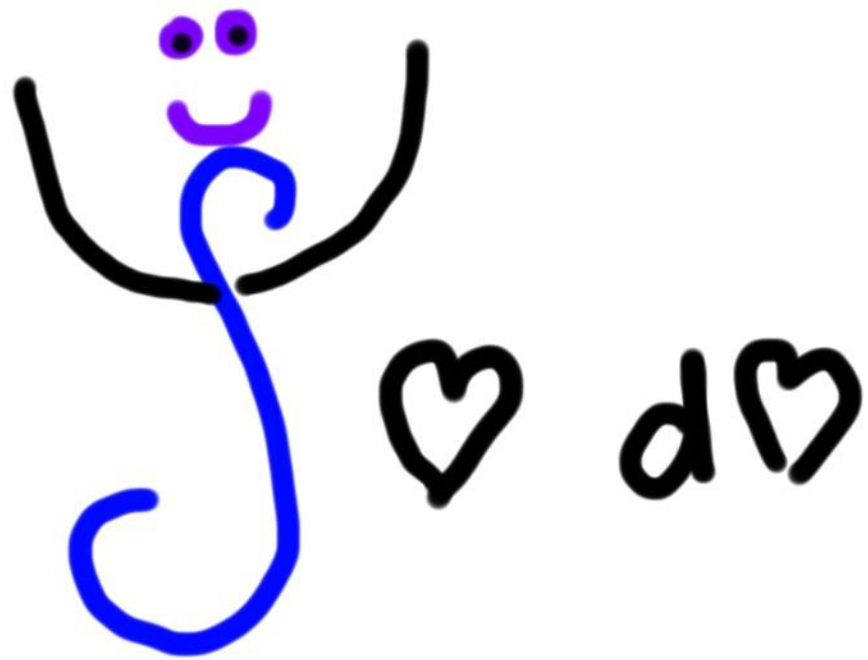
What is P(wait < 5 min)?







Integrals

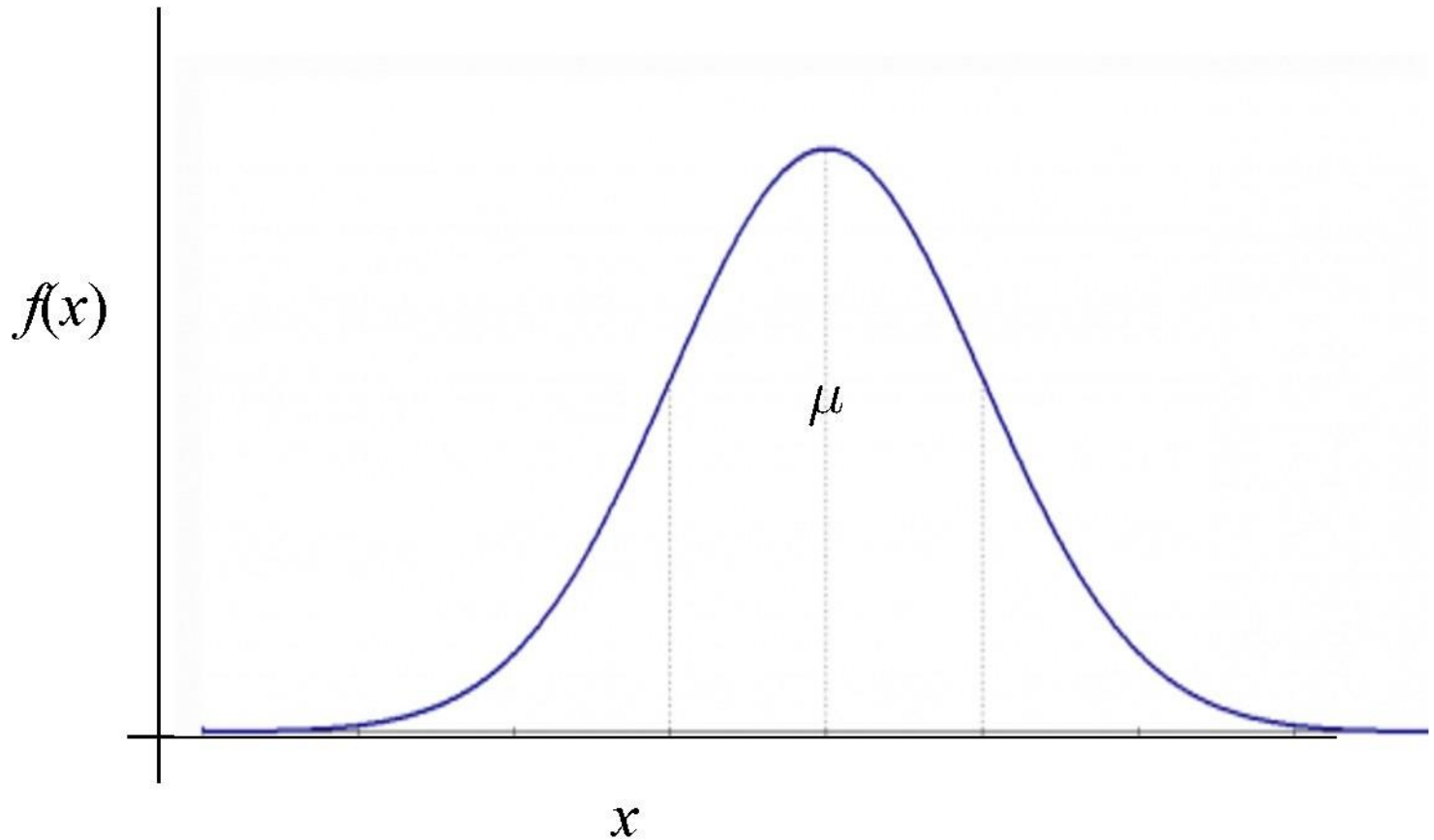


*loving, not scary

Probability Density Function

$$\mathcal{N}(\mu, \sigma^2)$$

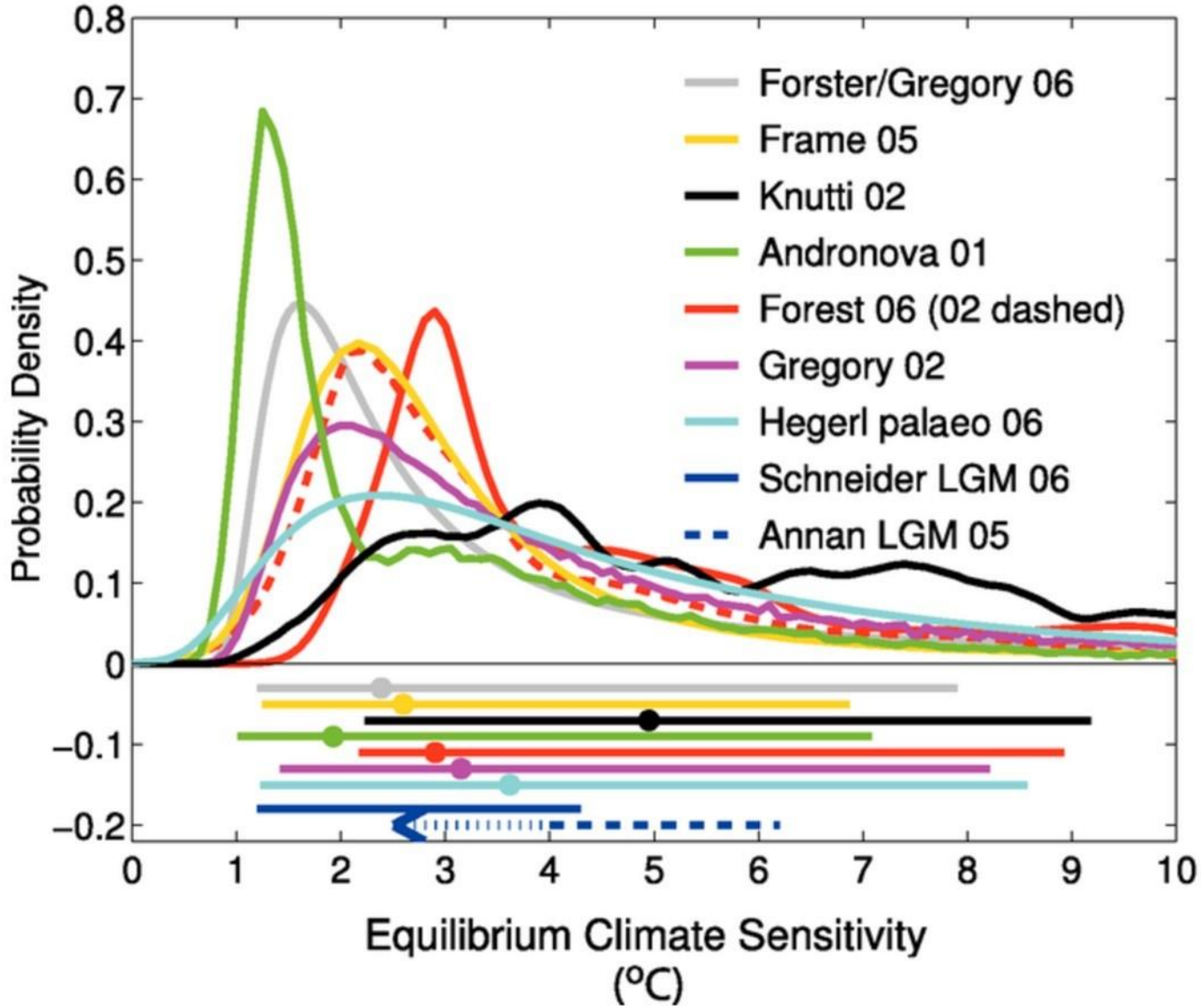
$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{\frac{-(x-\mu)^2}{2\sigma^2}}$$



What do you get if you integrate over a probability *density* function?

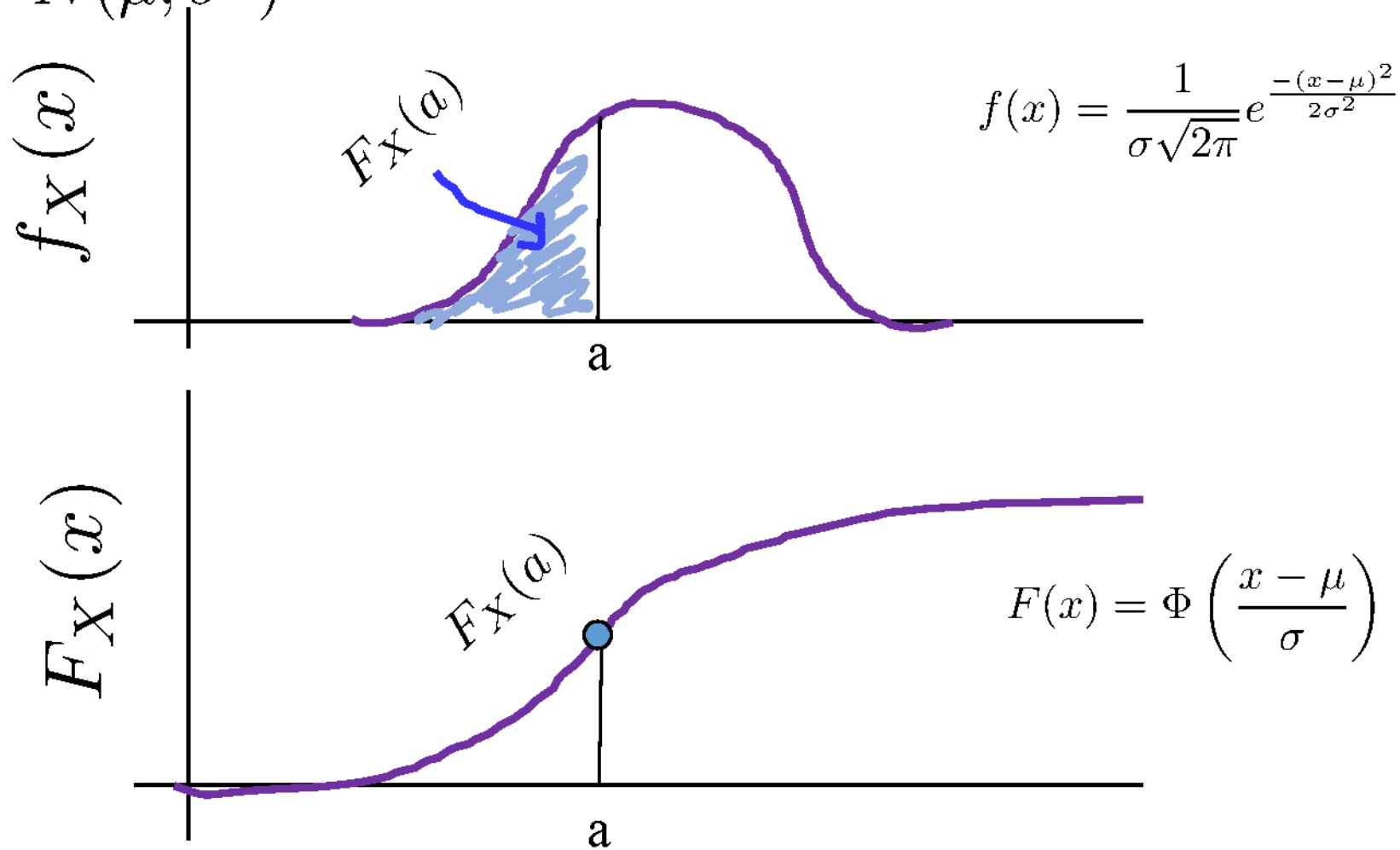
A probability!

Climate Sensitivity



PDF and CDF of a Normal

$X \sim N(\mu, \sigma^2)$



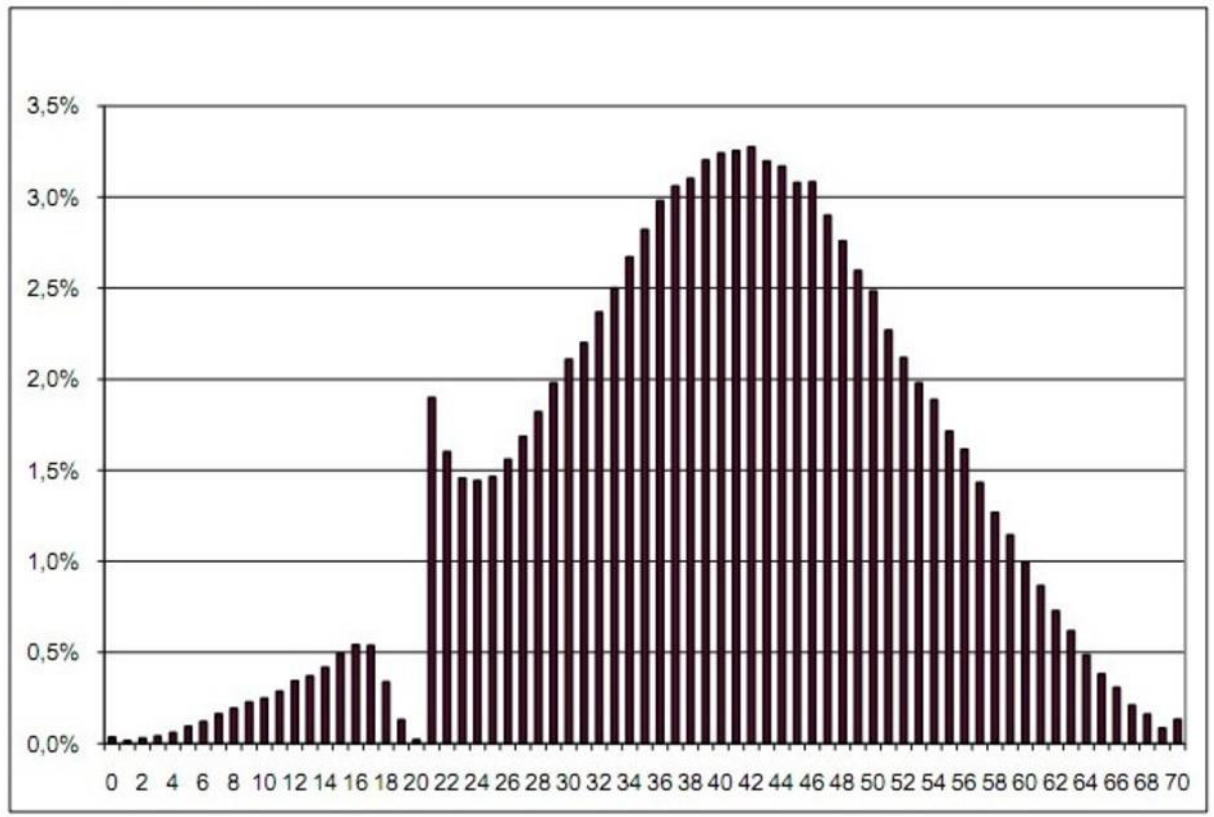
A CDF is the integral from –infinity to x of the PDF

Altruism?

Scores for a standardized test that students in Poland are required to pass before moving on in school

See if you can guess the minimum score to pass the test.

2.1. Poziom podstawowy



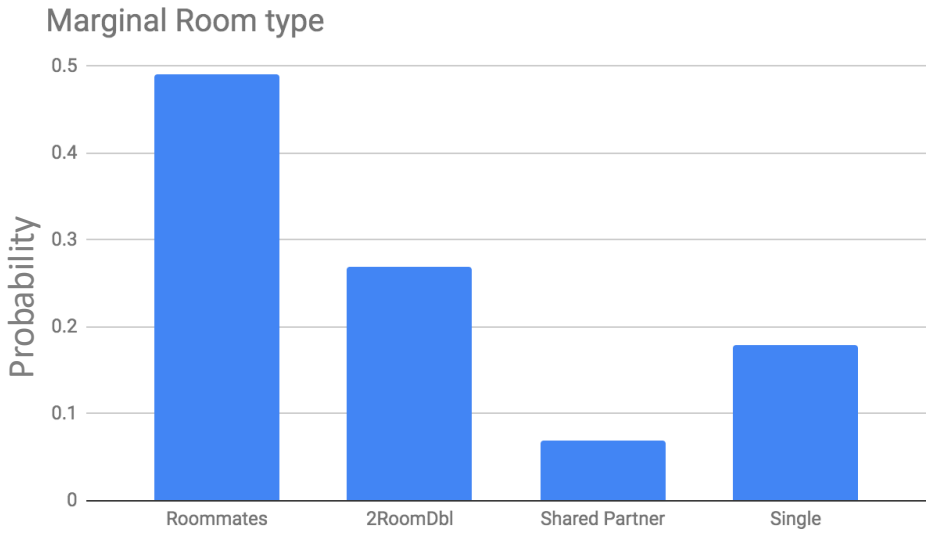
Wykres 1. Rozkład wyników na poziomie podstawowym

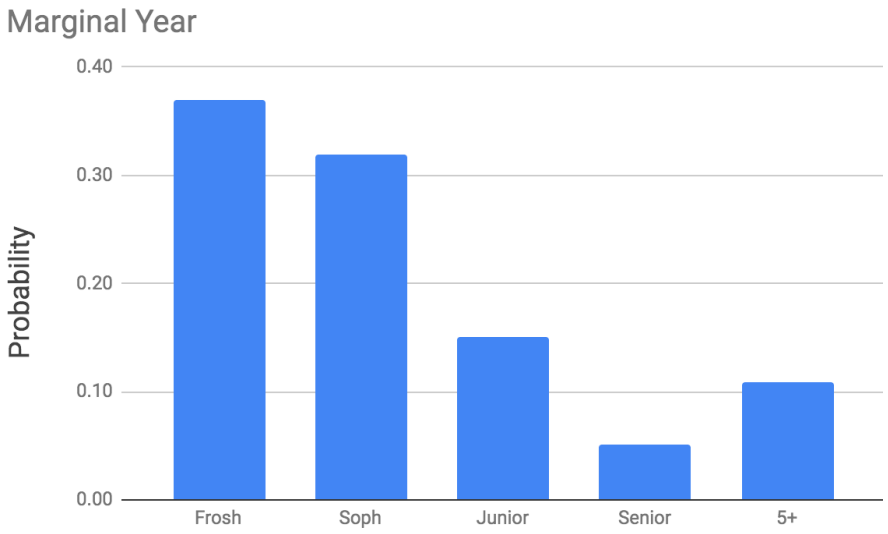
http://freakonomics.com/2011/07/07/another-case-of-teacher-cheating-or-is-it-just-altruism/comment-page-2/

Probabilistic Models

Joint Probability Table

	Roommates	2RoomDbl	Shared Partner	Single	
Frosh	0.30	0.07	0.00	0.00	0.37
Soph	0.12	0.18	0.00	0.03	0.32
Junior	0.04	0.01	0.00	0.10	0.15
Senior	0.01	0.02	0.02	0.01	0.05
5+	0.02	0.00	0.05	0.04	0.11
	0.49	0.27	0.07	0.18	1.00





Inference

Inference *noun*

Updating one's belief about a random variable (or multiple) based on conditional knowledge regarding another random variable (or multiple) in a probabilistic model.

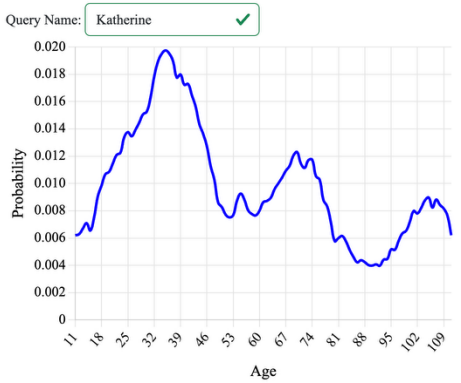
TLDR: conditional probability with random variables.

Inference

Age from C14

Updated Delivery Prob

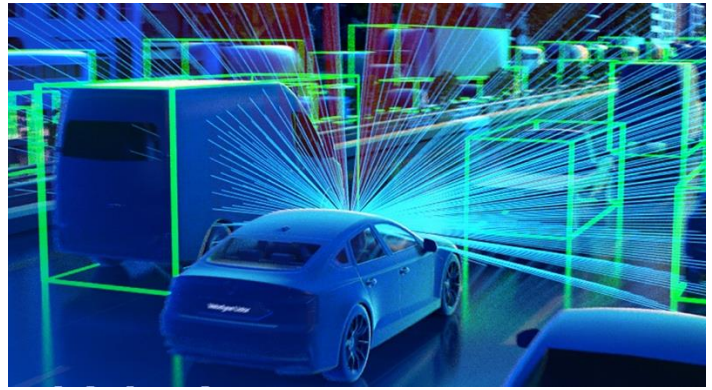
Age from Name



Hidden Chambers

Stanford Eye Test

Updating Lidar Belief



```
def update_belief_carbon_dating(m = 900):
  # pr_A[i] is P(Age = i| m = 900).
  pr_A = {}
  for i in range(100,10000+1):
    prior = 1 / n_years # P(A = i)
    likelihood = calc_likelihood(m, i) #P(M=m|A=i)
    pr_A[i] = likelihood * prior
  # implicitly computes the normalization constant
  normalize(pr_A)
  return pr_A
```

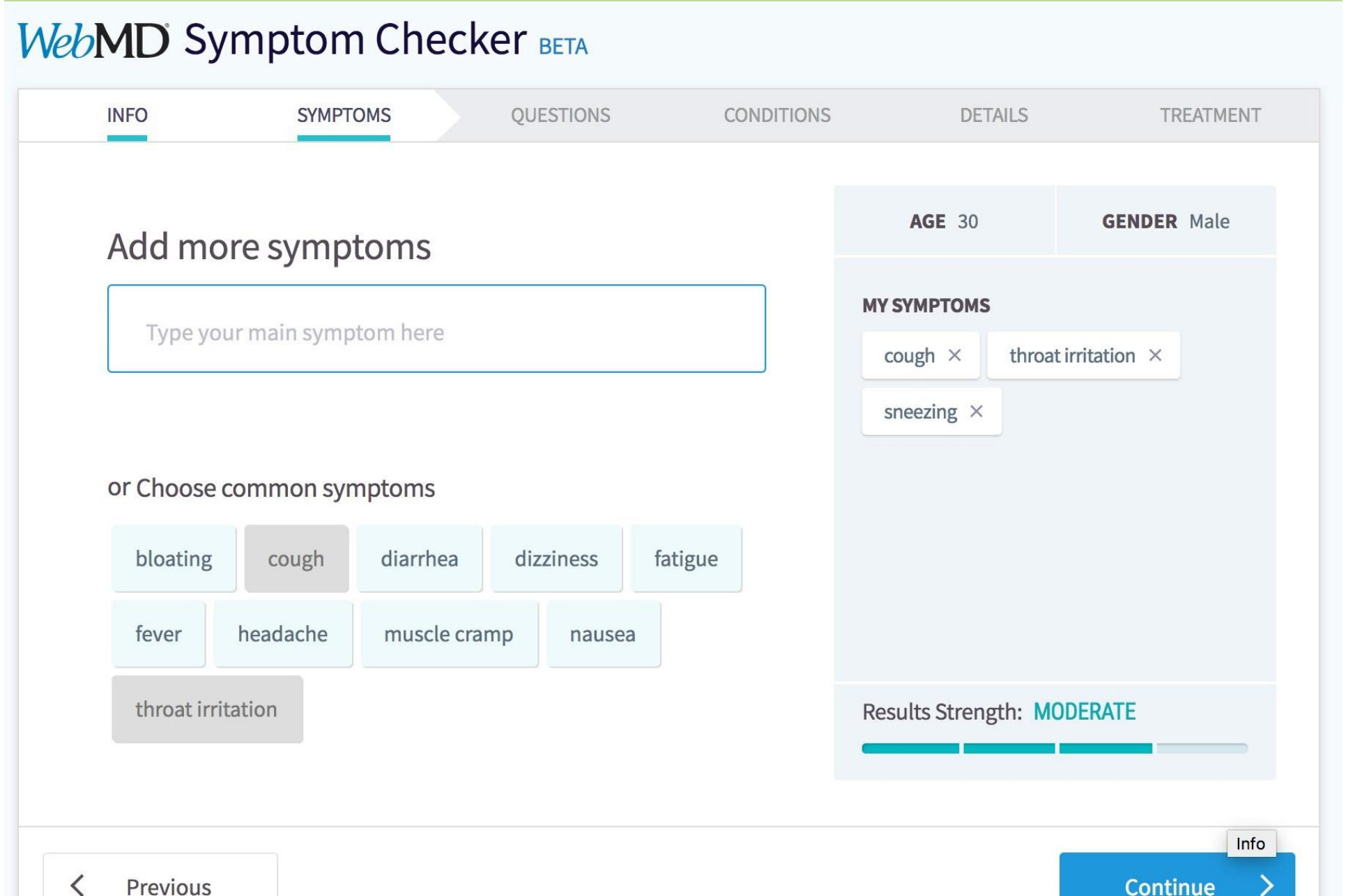
```
def update_belief_baby(prior, today = 10):
  # pr_D[i] is P(D = i| No Baby Yet).
  pr_D = {}
  for i in range(-50,25):
    # P(NoBaby | D = i)
    likelihood = 0 if i < today else 1
    pr_D[i] = likelihood * prior[i]
  # implicitly computes the LOTP
  normalize(pr_D)
  return pr_D
```

```
def update_belief_name_to_age(name = 'Laura'):
  # pr_age[i] is P(Age = i| name).
  # prob_name_and_age is just a counting from the US
  # Social Security database.
  pr_age = {}
  for i in range(10,110):
     pr_age[i] = calc_prob_name_and_age(name, i)
  # implicitly computes the normalization constant
  normalize(pr_age)
  return pr_age
```

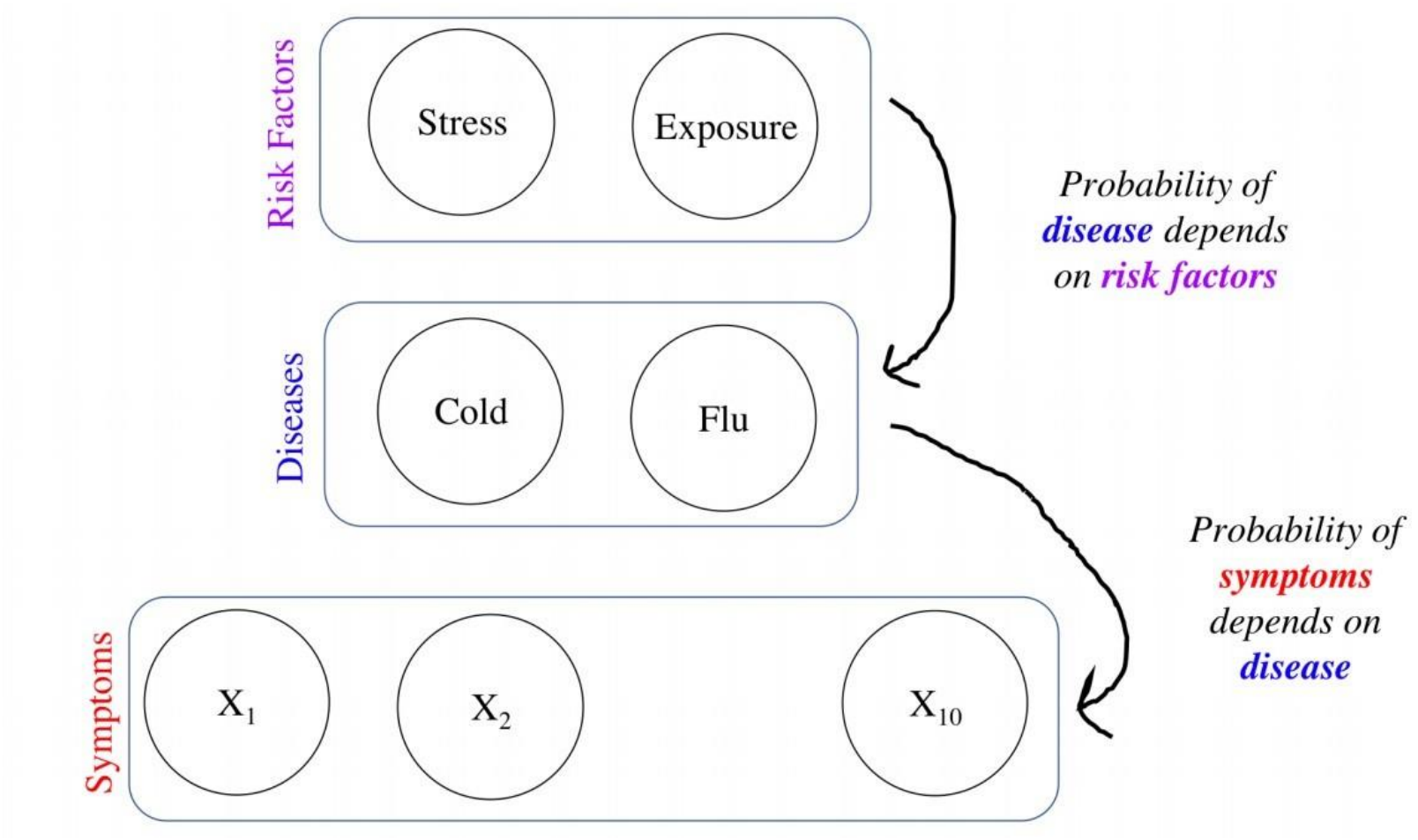
What do you notice is the same. What is different?

General “Inference”

General “Inference”



Huge Joint Models

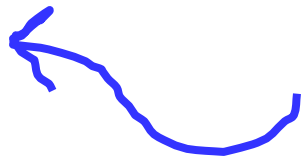


```
 3  N_SAMPLES = 100000
 4
 5  # Program: Joint Sa
 6  # ----------------
 7  # we can answer any
 8  # with multivariate
 9  # where conditioned
10  def main():
11      obs = getObserv
12      print 'Observat
13
14      samples = sampl
15      prob = probFluG
16      print 'Pr(Flu)
```

webMd — -bash — 38×22

```
[0, 0, 0, 0]
[0, 1, 0, 1]
[1, 0, 1, 0]
[1, 1, 1, 1]
[0, 1, 0, 1]
[0, 1, 0, 0]
[0, 0, 0, 0]
[0, 1, 1, 1]
[0, 1, 0, 0]
[0, 1, 0, 1]
[0, 1, 0, 0]
[0, 1, 0, 1]
[0, 1, 0, 1]
[0, 0, 0, 0]
[1, 1, 1, 1]
[0, 0, 0, 0]
[0, 0, 0, 0]
[1, 1, 1, 1]
[0, 1, 0, 0]
Observation =  [None, None, None, 1]
Pr(Flu | Obs) =  0.140635888502
>
```

Each one of these is one posterior sample:



[Flu, Ugrad, Fever, Tired]

Multinomial

Example document:
"Pay for Viagra with a credit-card. Viagra is great.
So are credit-cards. Risk free Viagra. Click for free."
n = 18

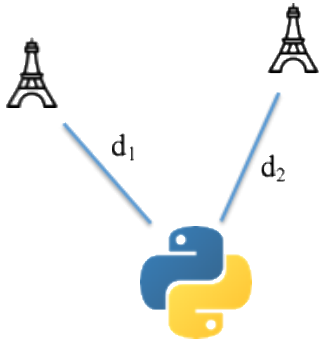
$$P\left(\begin{array}{l}\text{Viagra} = 2\\ \text{Free} = 2\\ \text{Risk} = 1\\ \text{Credit-card: } 2\\ \dots\\ \text{For} = 2\end{array} \middle| \text{spam}\right) = \frac{n!}{2!2!\dots 2!} p_{\text{viagra}}^2 p_{\text{free}}^2 \cdots p_{\text{for}}^2$$

It's a Multinomial!

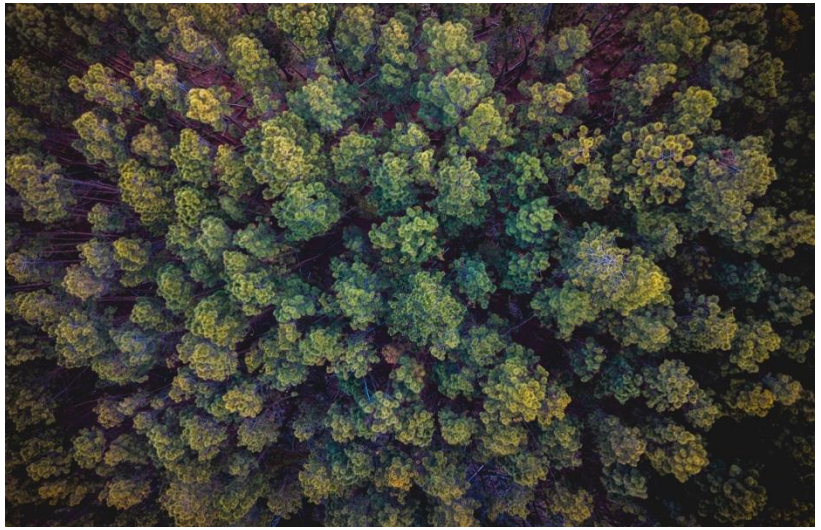
Probability of seeing this document | spam

The probability of a word in spam email being viagra

Hard Midterm, Great Job




```
def simulation_1(p=0.4, k=0.01):
    X = bern(p)                 # sa
    Y = bern(p+k)               # in
    return Y - X

def simulation_2(p=0.4, k=0.01):
    X = bern(p)
    if X == 1:
        Y = 1
    else:
        Y = bern(k/(1-p))
    return Y - X
```



PEP

New last year

Personalized Exam Prep Signup: Final

PEP is back for the final. The only difference is that Final PEP is 15 mins long.

What is PEP?
This quarter we are trying something **new in CS109**! In the past we have talked to many students *after* the final to get feedback on how they studied and what they found challenging. We often have some good insights for students, but it can feel like those insights are a few weeks too late. This quarter we are trying to get you those insights ***before*** the final so you can master the material more effectively. We call these 1:1s **Personalized Exam Prep**.

You meet in-person with a TA a week before the final for **15 mins**. You don't need to prepare or bring anything. The TA gets to know you and, after the session, sends you home with a draft of a study plan. Participating will get you an automatic 4 points on the final. If you can't participate, that is fine, your final will be graded as usual. You likely will not get your section TA, but it is possible (what is the probability???).

View My Personalized Guide

Reserve a Time

Tuesday, Dec 3

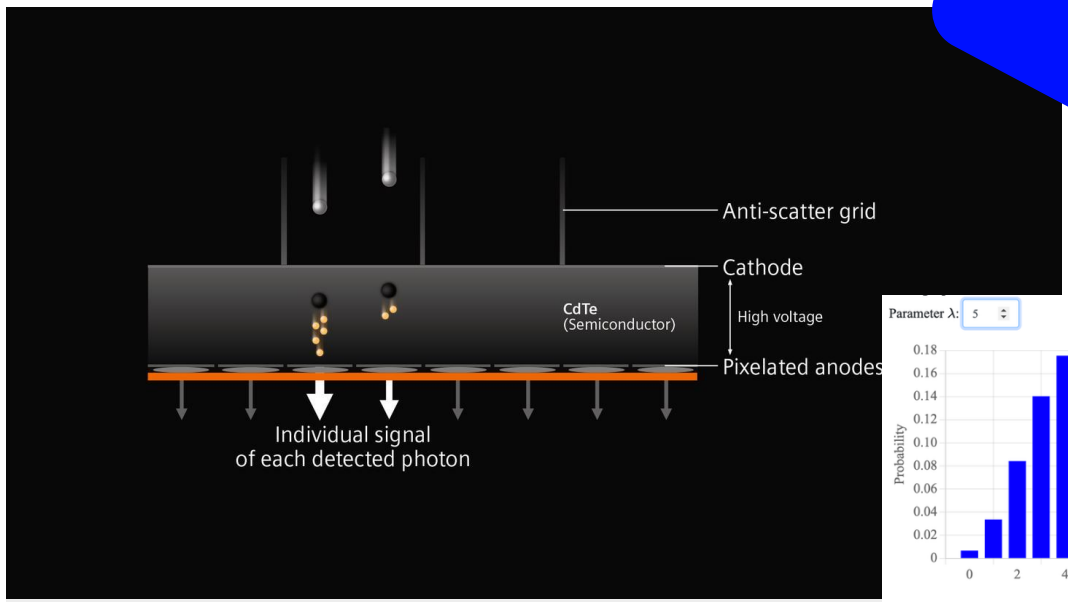
Select a time slot

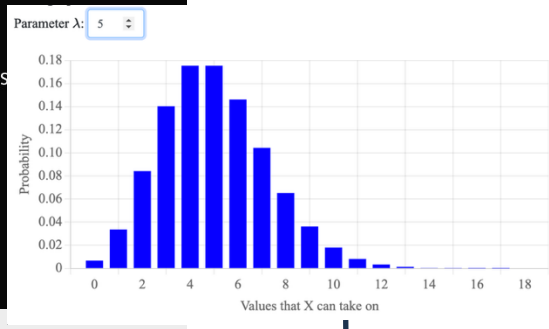
Wednesday, Dec 4

✓ Select a time slot
Wednesday, Dec 4 9:15 AM
Wednesday, Dec 4 9:30 AM
Wednesday, Dec 4 12:00 PM
Wednesday, Dec 4 12:15 PM

288 Midterm, 290 Final 120 Hours of Extra TA time

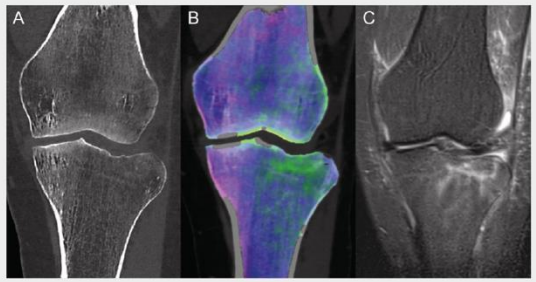
Learning Goal: Be fluent in the language of probability





Poisson

Bone marrow edema at photon-counting CT



A 56-year-old female patient with knee injury. (A) Diagnostic CT image acquired with photon-counting CT (PCCT) shows a nondisplaced fracture of the lateral tibial plateau. The associated bone marrow edema (BME) is shown on (B) the BME map reconstructed from the unenhanced PCCT scan and (C) the fat-suppressed T2-weighted MRI scan.

"A 56-year-old female patient presented with severe pain after a ground-level fall on the left knee. Antero-posterior and lateral view radiographs of the knee demonstrated no evident abnormalities."

"Photon-counting CT (PCCT) demonstrated a nondisplaced fracture of the lateral tibial plateau. Furthermore, the bone marrow edema (BME) map reconstructed from the unenhanced PCCT image showed associated BME (Figure, B). This finding correlated well with BME seen on the fat-suppressed T2-weighted MRI scan."

↓

Inference

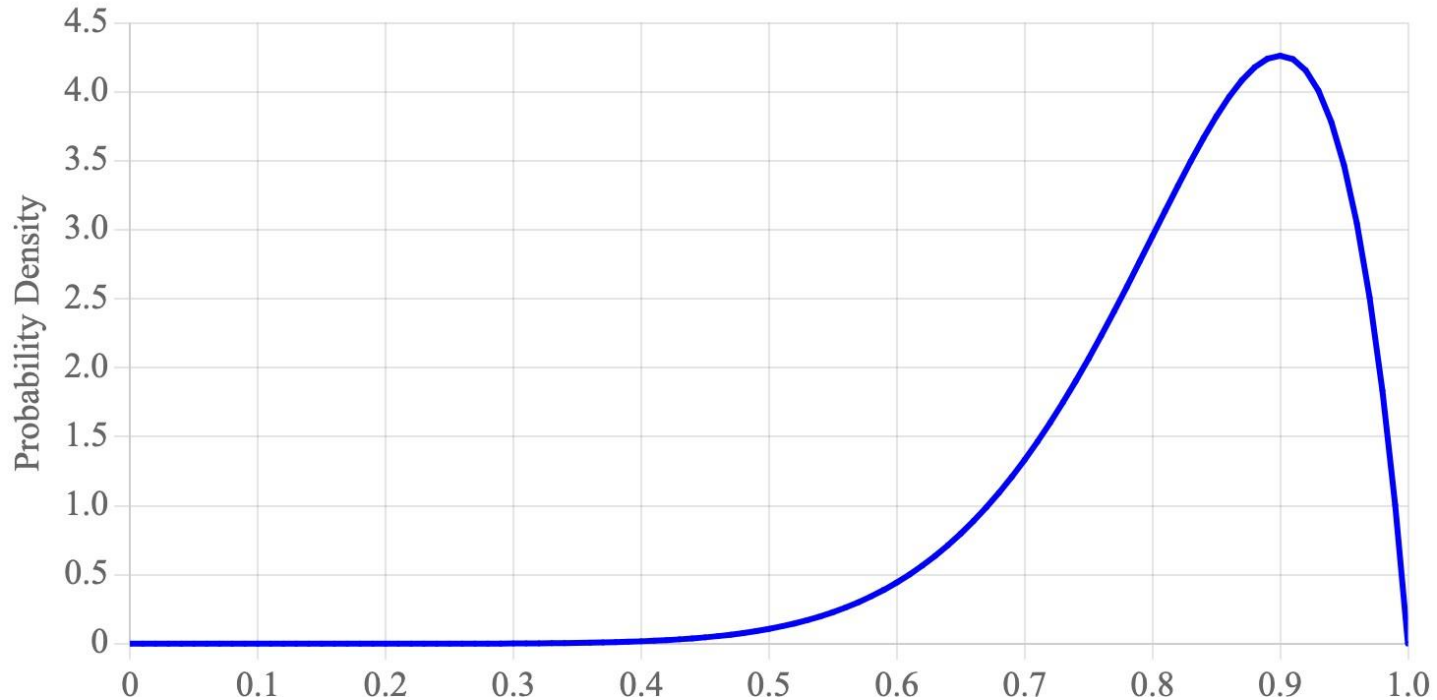
↓

Information Theory

Uncertainty Theory

$$
\begin{aligned}
& f(X = x | H = 9, T = 1) \\
&= \frac{P(H = 9, T = 1 | X = x) \cdot f(X = x)}{P(H = 9, T = 1)} && \text{Bayes Theorem} \\
&= \frac{\binom{10}{9} x^9 (1 - x)^1 \cdot f(X = x)}{P(H = 9, T = 1)} && \text{Binomial PMF} \\
&= \frac{\binom{10}{9} x^9 (1 - x)^1 \cdot 1}{P(H = 9, T = 1)} && \text{Uniform PDF} \\
&= \frac{\binom{10}{9}}{P(H = 9, T = 1)} x^9 (1 - x)^1 && \text{Constants to front} \\
&= K \cdot x^9 (1 - x)^1 && \text{Rename constant}
\end{aligned}
$$

Lets take a look at that function. For now we can let $K = \frac{1}{110}$. Regardless of K we will get the same shape, just scaled:



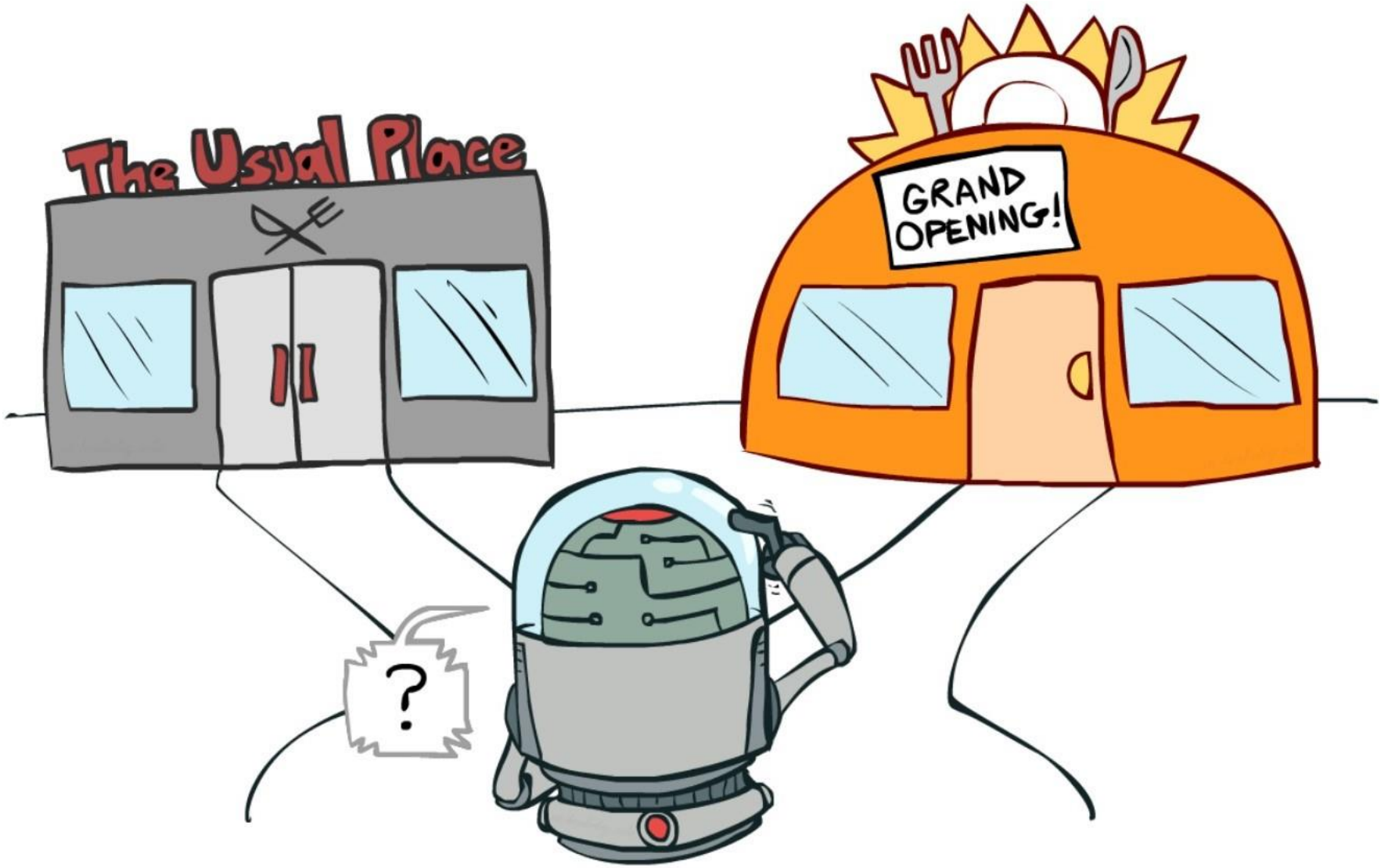
Let’s Play

Drug A

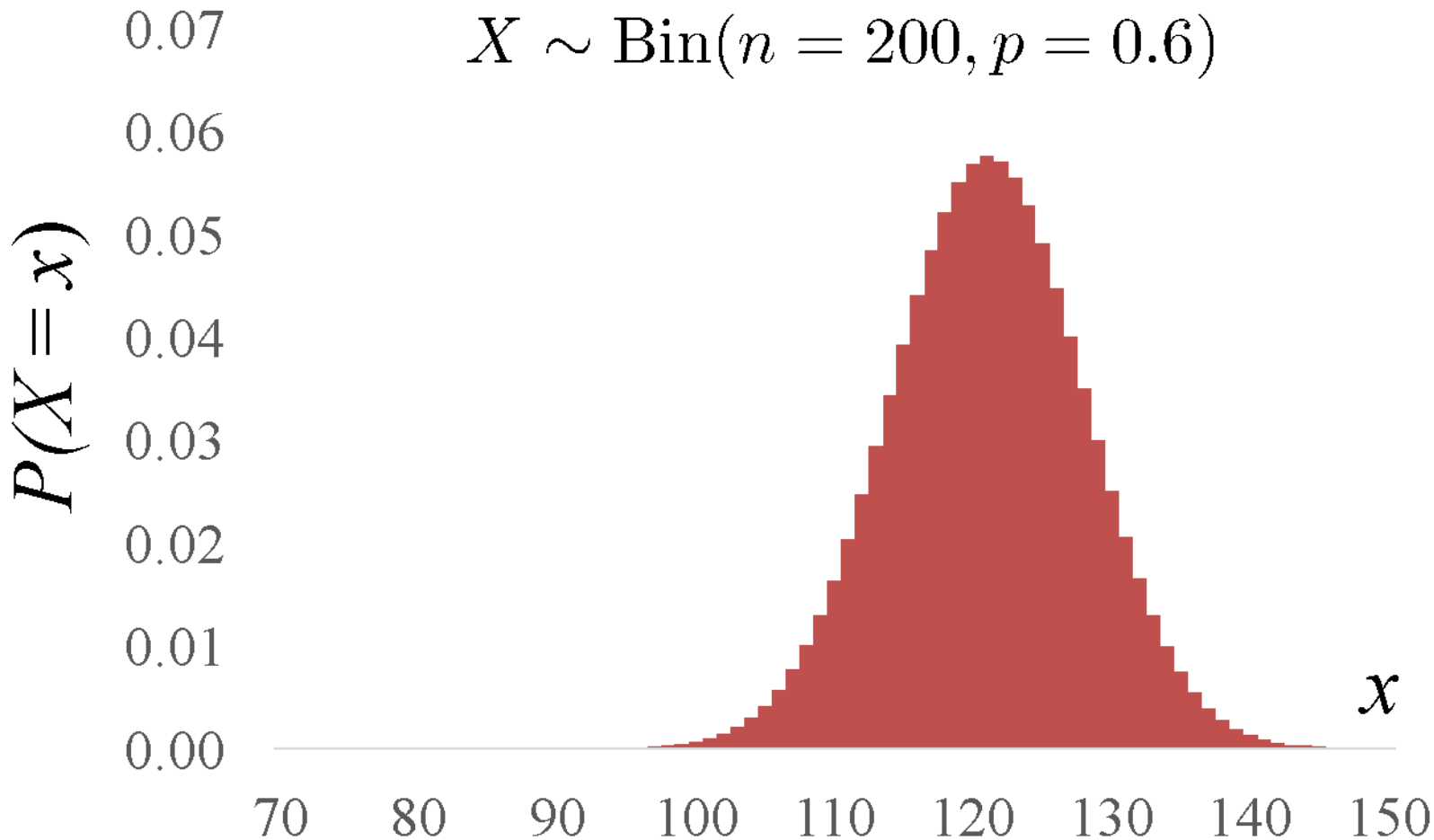
Drug B

Which one do you give to a patient?

Thompson Sampling

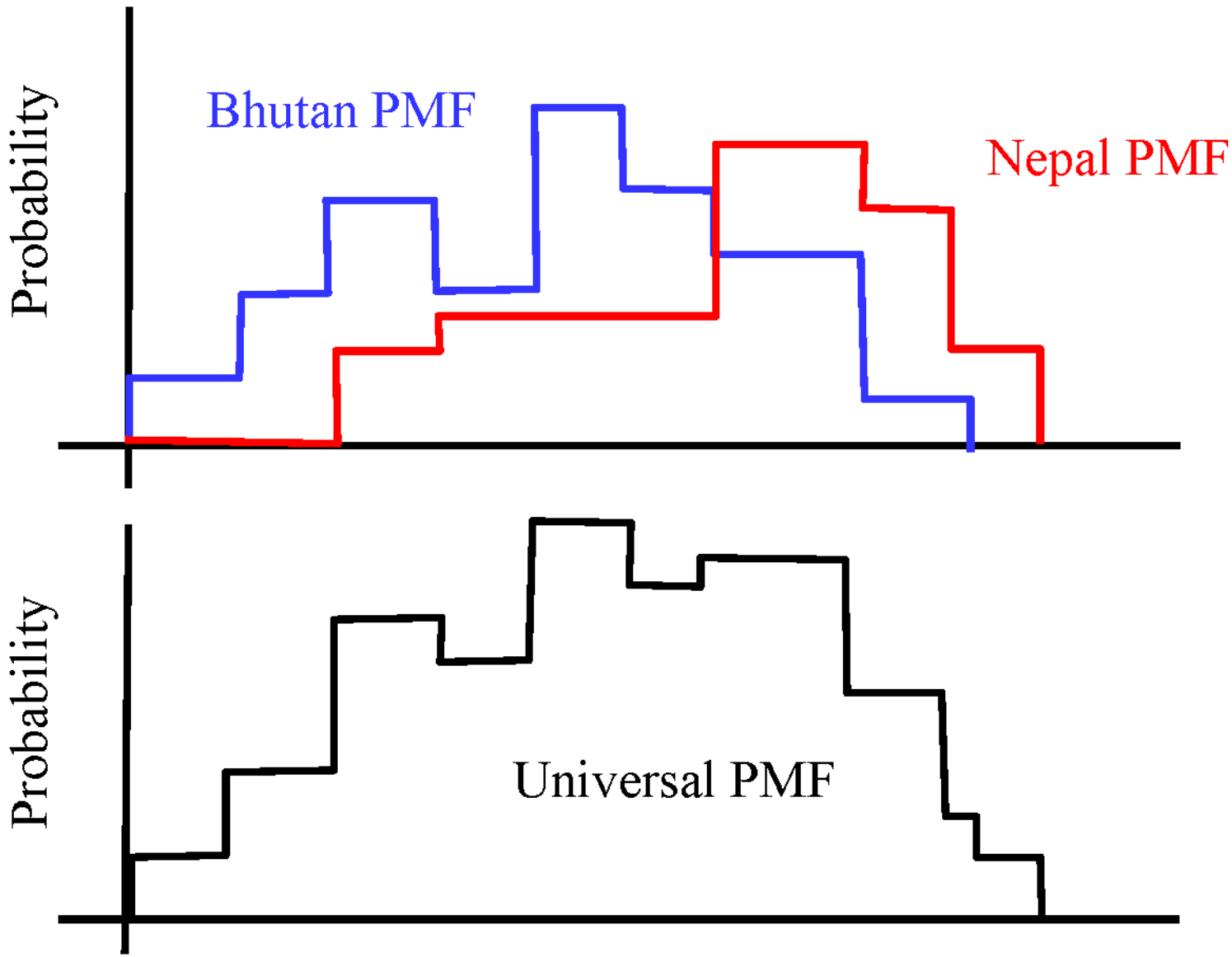


C.L.T. Explains This

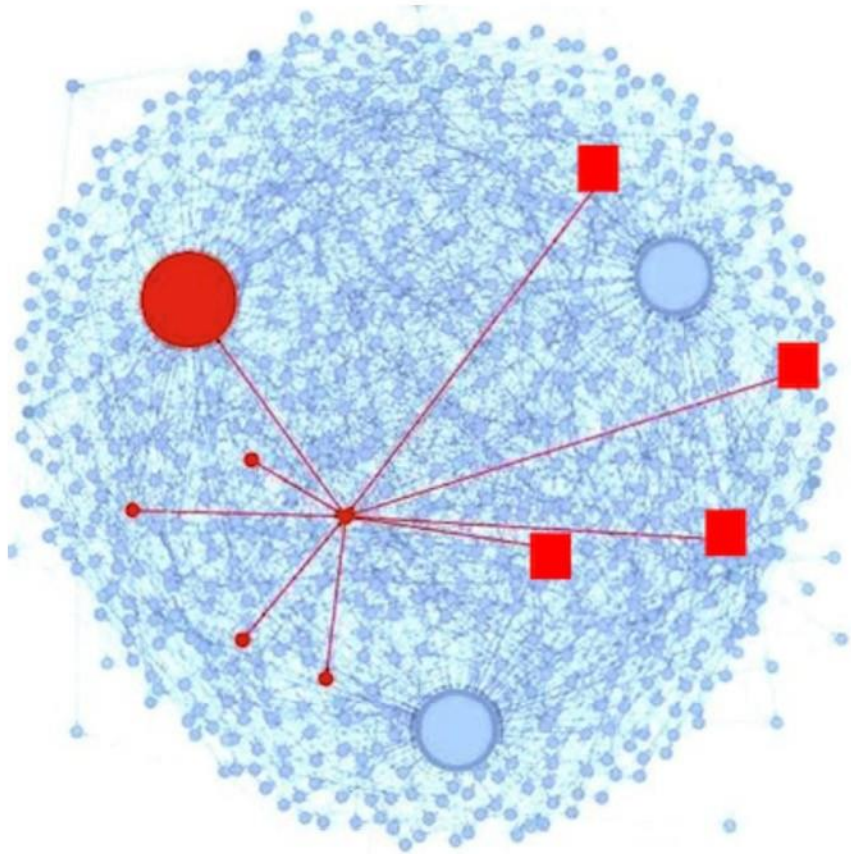


Bootstrap

Universal Sample



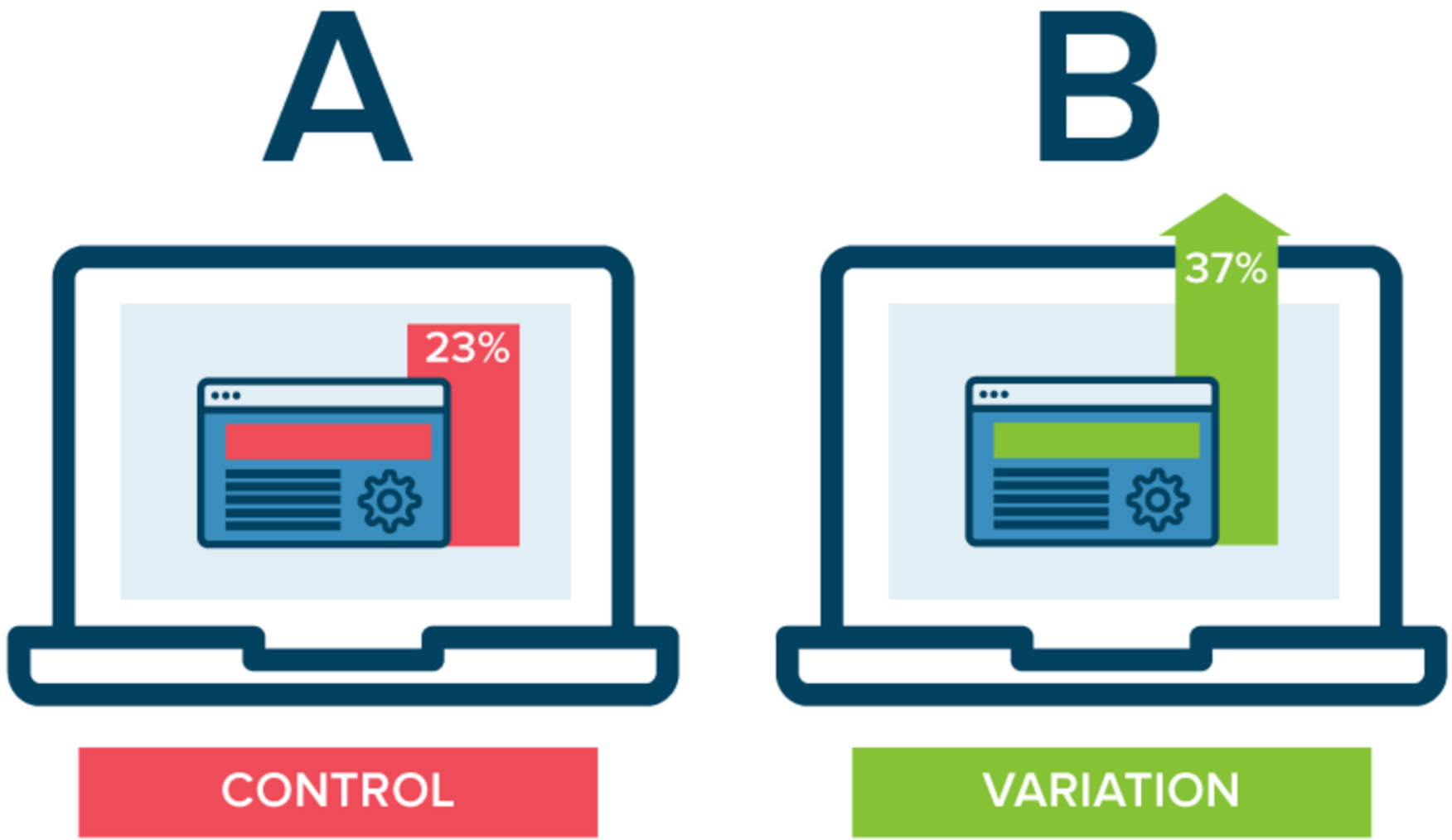
Peer Grading



Peer Grading on Coursera HCI.

31,067 peer grades for 3,607 students.

A/B Testing



Information Theory

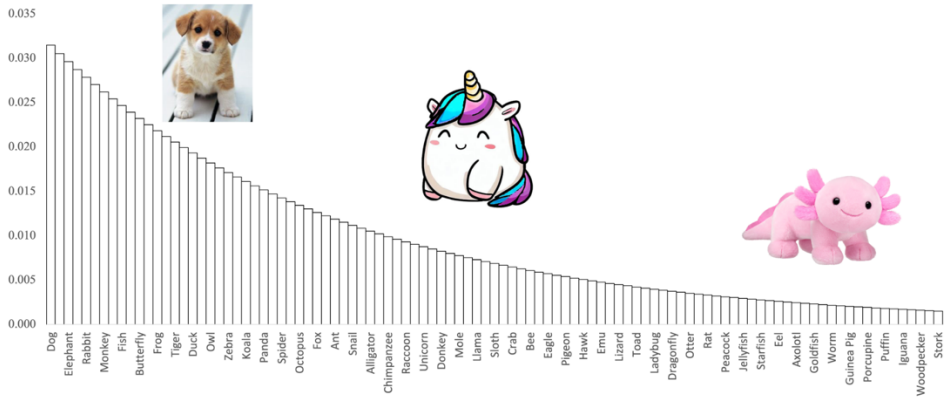
New Last Year

Information Theory

Information theory is an incredibly powerful perspective which plays a central role in a ton of algorithms, including Decision Trees, the WordleBot, Adaptive Tests, Optimal Poker Play and even ompression of data (like Huffman Encoding or even Jpeg files)! The goal of this chapter is to balance showing off the awesome power of Information Theory while also keeping things as straight forward as possible. To that end, a great place to start is thinking about how you could write a bot that can play the question answering game of, "Think of an Animal".

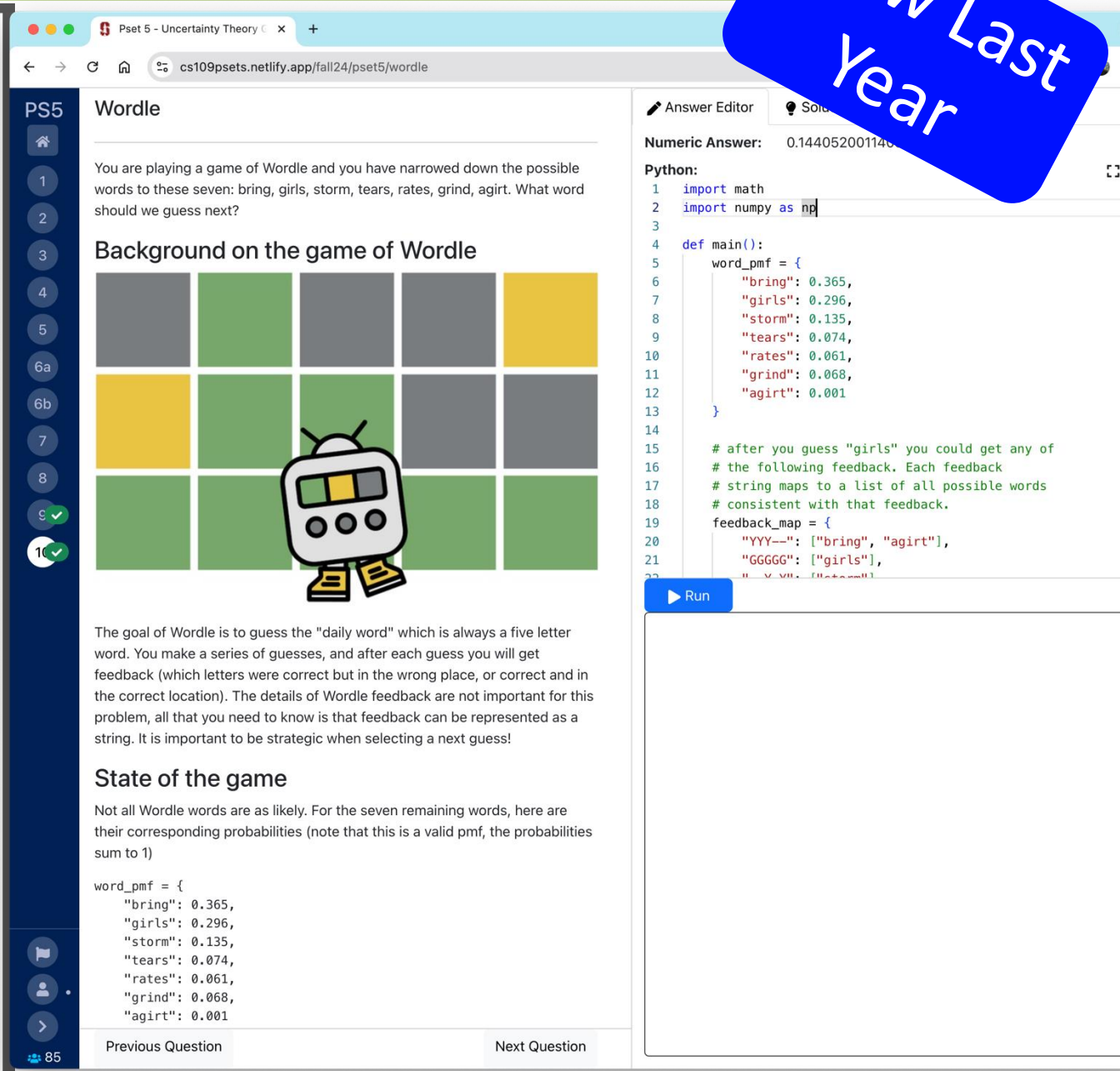
Think of an Animal!

The game of "Think of an Animal" goes like this: The human is going to be thinking of an animal. We assume that the distribution of how often they chose an animal is known (based off how popular the animal is to four year olds):



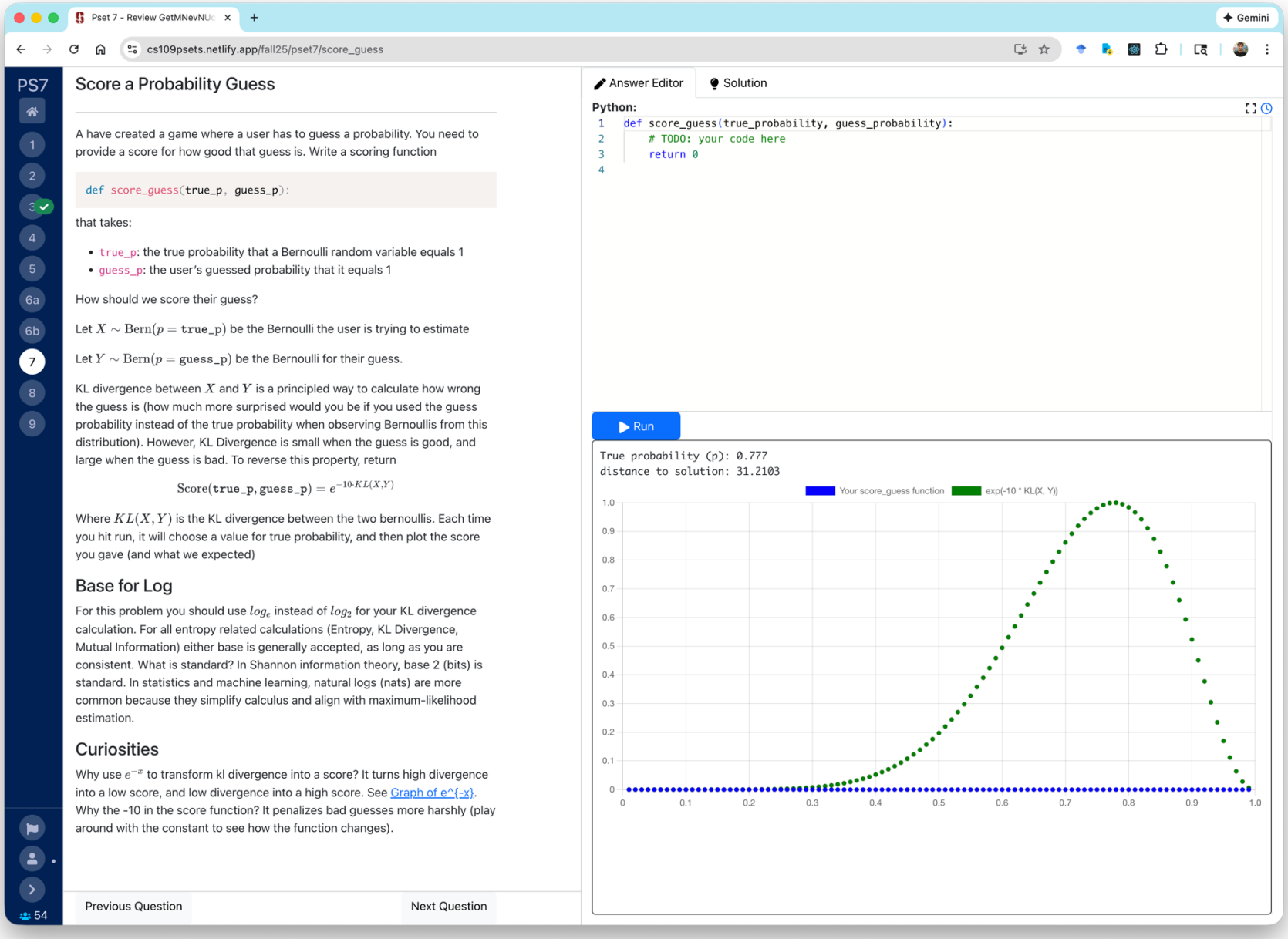
The task of your algorithm is to select which question to ask next. Assume you are given a bank of yes or no questions which include classics like:

- Is it a pet?
- Does it live in the water?
- Are you thinking of a dog?



KL Divergence

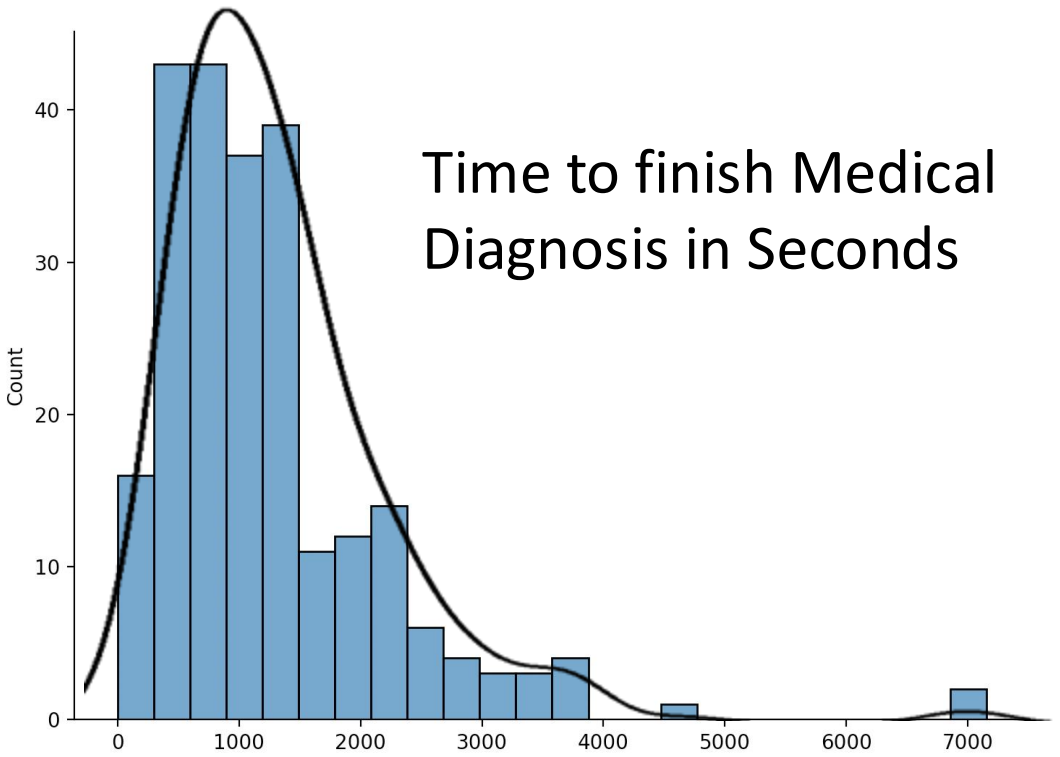
New!



Machine Learning

MLE for Choosing Params

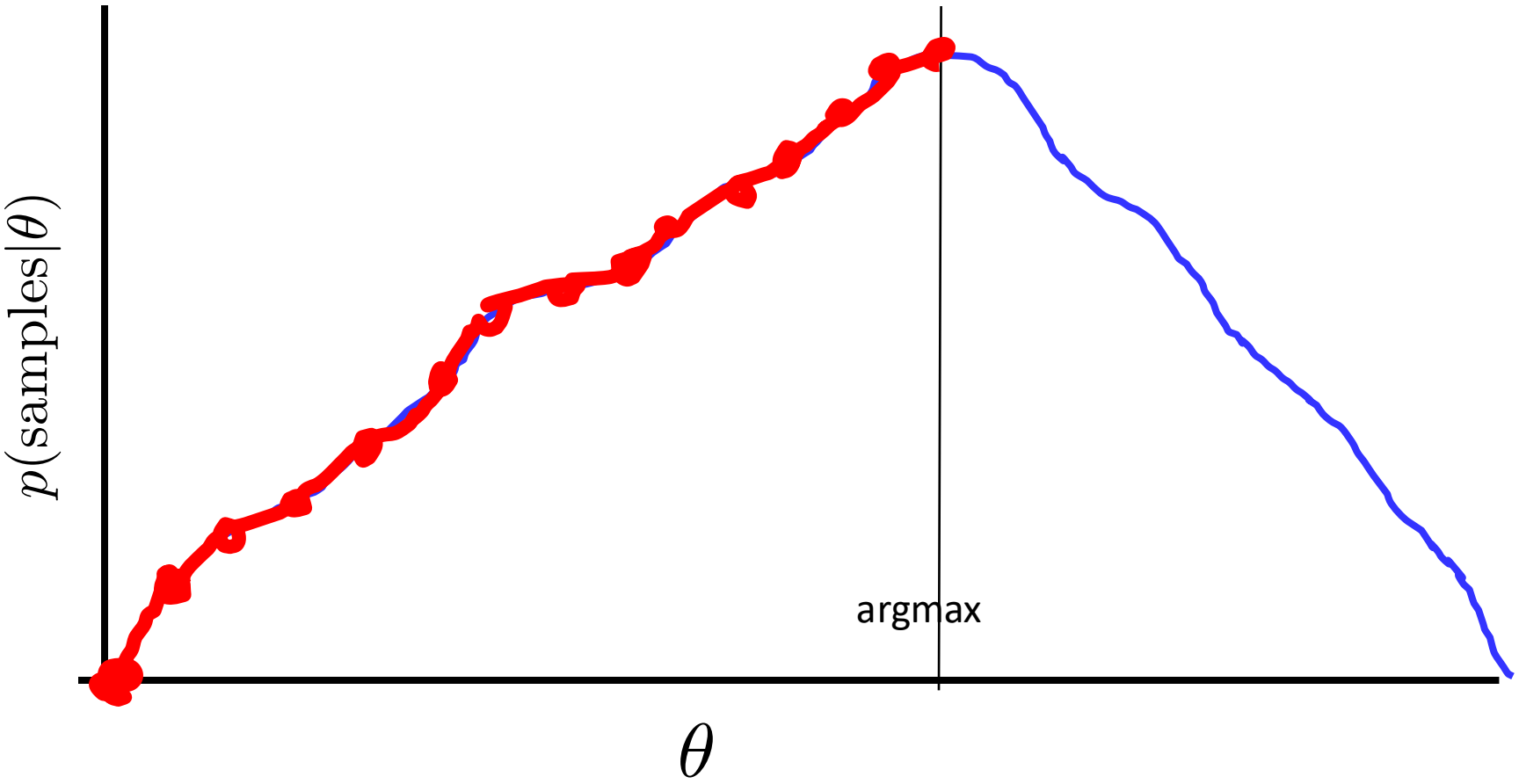
```
[3.002, 0.983, 2.186, 1.624, 3.997, 1.777,
2.809, 0.42, 0.515, 1.582, 0.948, 0.458, 1.
066, 0.8, 2.398, 0.794, 2.561, 2.61, 0.
595, 3.897, 1.852, 1.182, 3.043, 0.905, 1.
45, 0.405, 0.445, 2.103, 1.425, 3.12, 0.
973, 1.056, 3.715, 2.952, 1.817, 2.686, 4.
173, 0.358, 2.185, 2.581, 7.134, 0.206, 2.
049, 0.896, 2.095, 4.39, 2.199, 3.434, 5.
696, 0.819, 0.416, 1.571, 1.337, 2.79, 2.
701, 3.061, 4.677, 0.671, 1.594, 3.586, 2.
708, 1.417, 1.799, 1.137, 1.771, 2.12, 0.
93, 6.835, 3.213, 2.541, 2.505, 1.257, 1.
99, 1.5, 0.014, 3.856, 0.979, 2.413, 2.
596, 1.653, 0.881, 4.457, 0.717, 3.305, 2.
456, 3.462, 1.737, 0.968, 0.528, 0.18, 1.
626, 2.224, 1.466, 1.6, 1.572, 0.12, 2.86,
1.062, 2.139, 1.217]
```



Time to finish Medical Diagnosis in Seconds

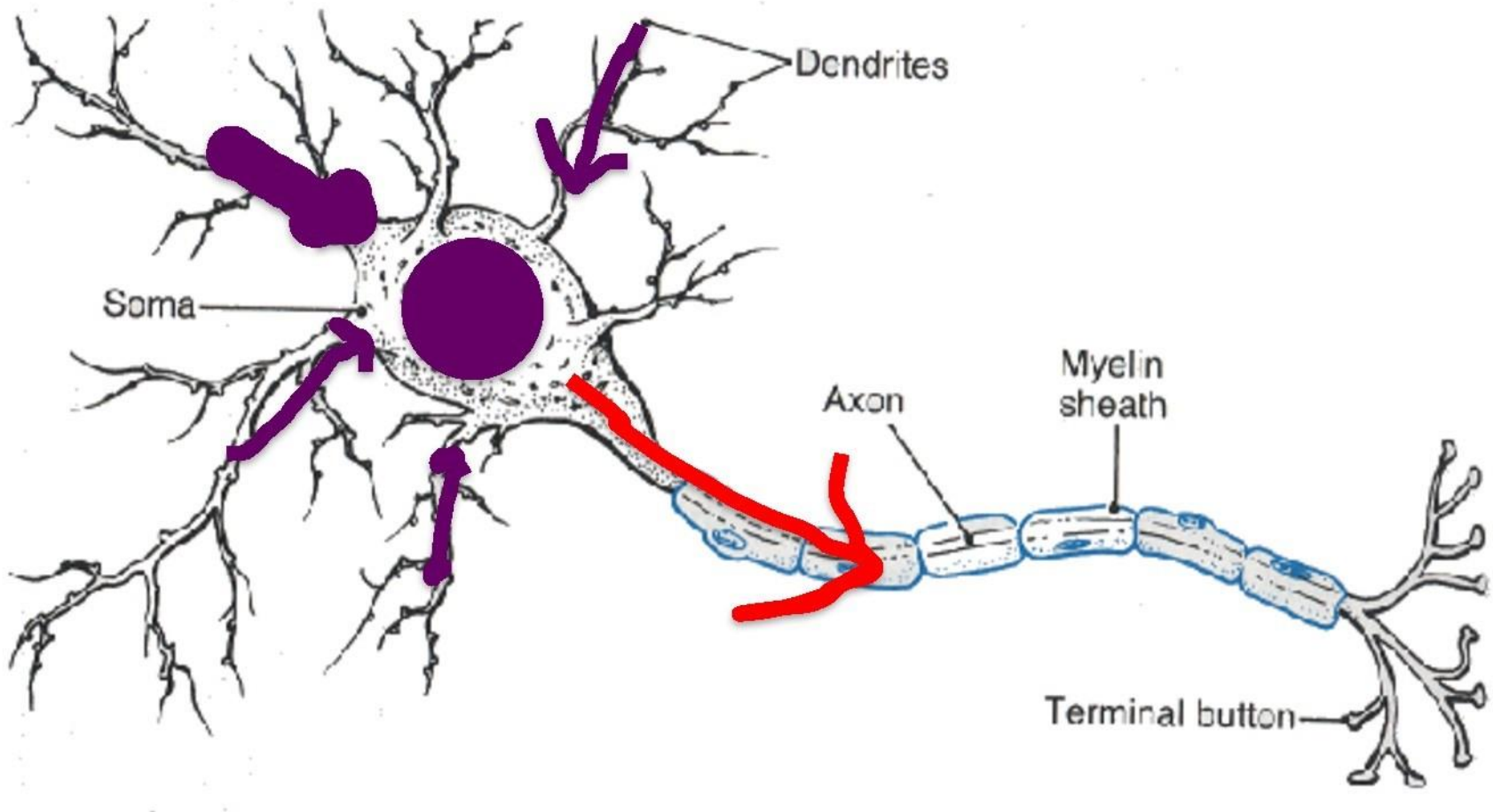
$$f(x) = \frac{\lambda^k x^{k-1} e^{-\lambda x}}{(k-1)!} = \frac{\lambda^k x^{k-1} e^{-\lambda x}}{\Gamma(k)}$$

Gradient Ascent



Walk uphill and you will find a local maxima
(if your step size is small enough)

Logistic Regression



Machine Learning

Heart

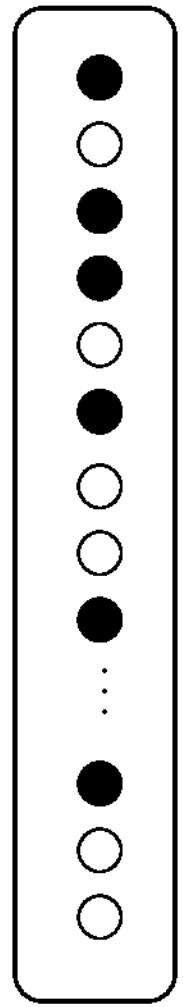
Ancestry

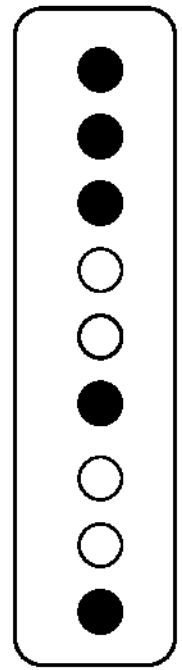
23andMe

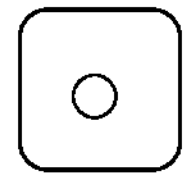
Netflix

NETFLIX

Deep Learning







$$LL(\theta) = y \log \hat{y} + (1 - y) \log[1 - \hat{y}]$$

$$\hat{y} = \sigma\left(\sum_{j=0}^{m_h} \mathbf{h}_j \theta_j^{(\hat{y})}\right)$$

$$\mathbf{h}_j = \sigma\left(\sum_{i=0}^{m_x} \mathbf{x}_i \theta_{i,j}^{(h)}\right)$$

Calibration

New!

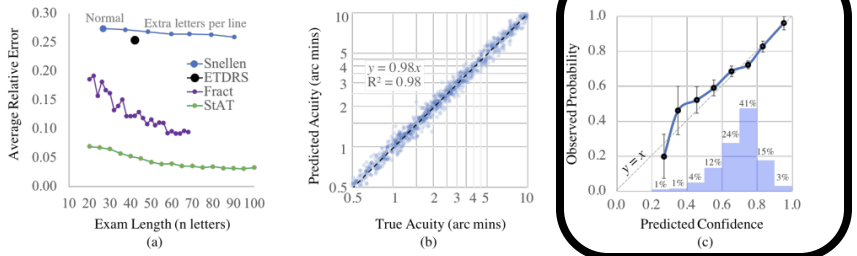


Figure 4: (a) The tradeoff between length of exam and error for the different algorithms. (b) A visualization of the predictions made by StACT. (c) Calibration test: StACT confidences correspond to how often it is correct.

4.2 Baseline Acuity Tests

We use the following baselines and prior algorithms to compare against the StACT algorithm.

Const Policy. This policy always predicts the most common visual acuity in our data i.e. the mode of the visual acuity prior. This serves as a true null model because it doesn't take patient responses into account at all.

Snellen and ETDRS. The Revised 2000 Series ETDRS charts and the Traditional Snellen Eye Chart were programmed so that we could simulate their response to different virtual patients. Both exams continue until the user incorrectly answers questions for more than half of the letters on a line. ETDRS has a function for predicted acuity score that takes into account both the last line passed, and how many letters were read on the last line not-passed. Both charts use 19 unique optotypes.

FrACT. We use an implementation of the FrACT algorithm (Bach and others 1996), with the help of code graciously shared by the original author. We also included the ability to learn the "s" parameter as suggested by the 2006 paper (Bach 2006), and verified that it improved performance.

	μ Acuity Error	μ Test length
Const	0.536	0
Snellen†	0.264	27
ETDRS†	0.254	42
FrACT	0.212	20
StACT	**0.069**	20
StACT-noSlip	0.150	20
StACT-greedyMAP	0.132	20
StACT-logistic	0.125	20
StACT-noPrior	0.090	20
StACT-goodPrior	**0.047**	20
StACT-star	**0.038**	63

Table 1: Average relative error for each algorithm. Except for Snellen each test was allowed 20 letters. Results are average relative error after 1000 tests. † Snellen and ETDRS used 19 unique optotypes.

5 Results and Evaluation

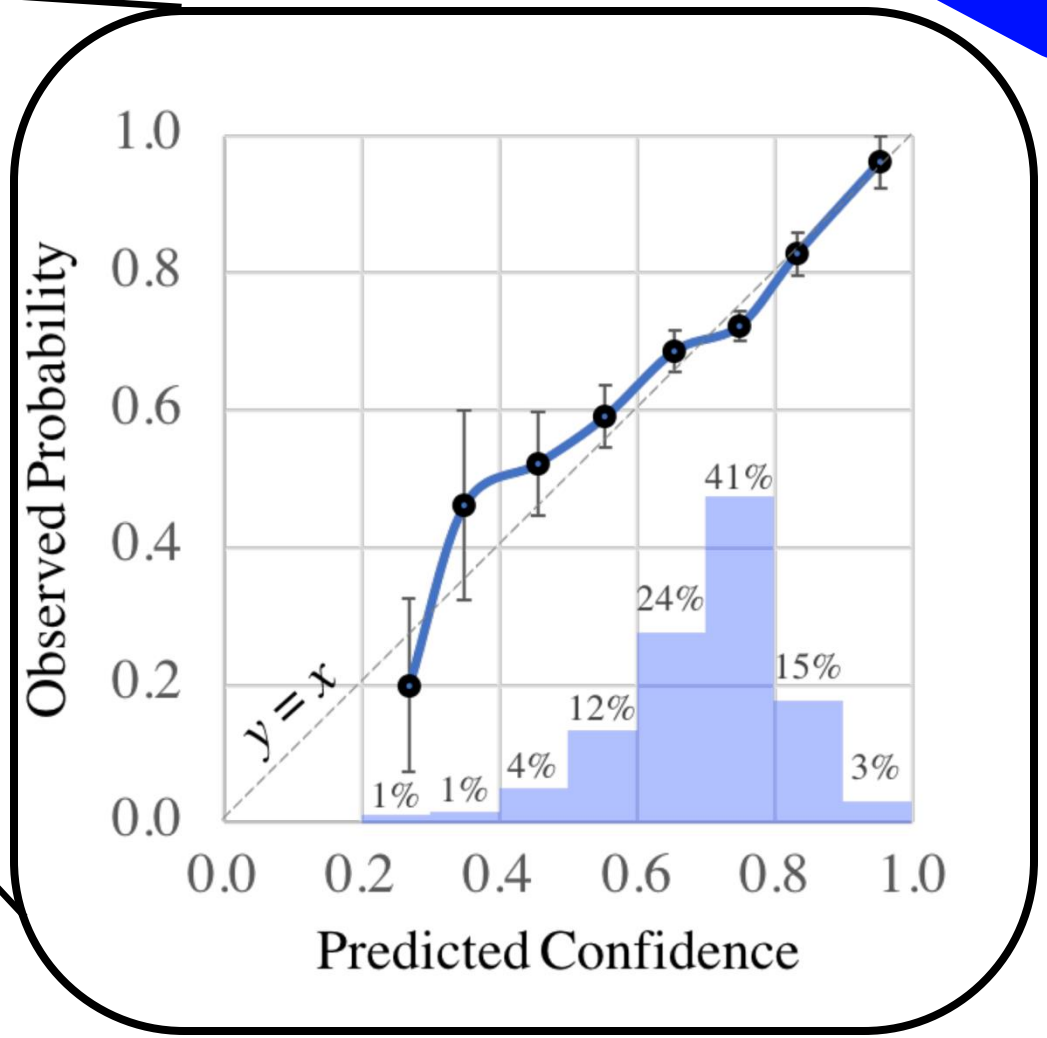
The results of the experiments can be seen in Table 1.

Accuracy and error. As can be seen from Table 1, the StACT test has substantially less error than all the other baselines. After 20 optotype queries, our algorithm is capable of predicting acuity with an average relative error of 0.069. This prediction is a 74% reduction in error from our implementation of the ubiquitous Snellen test (average error = 0.276), as well as a 67% reduction in error from the FrACT test (average error = 0.212). One possible reason for the improvement over FrACT is that the simulations used in our evaluations are based off the Floored-Exponential model that StACT uses. However, even when we evaluate StACT on simulations drawn from the FrACT logistic assumption we still achieve a 41% reduction. The improved accuracy of the StACT algorithm suggests our Bayesian approach to measuring acuity is a fruitful proposal both because of our introduction of the floored exponential as well as our Thompson-sampling inspired algorithm to chose a next letter size.

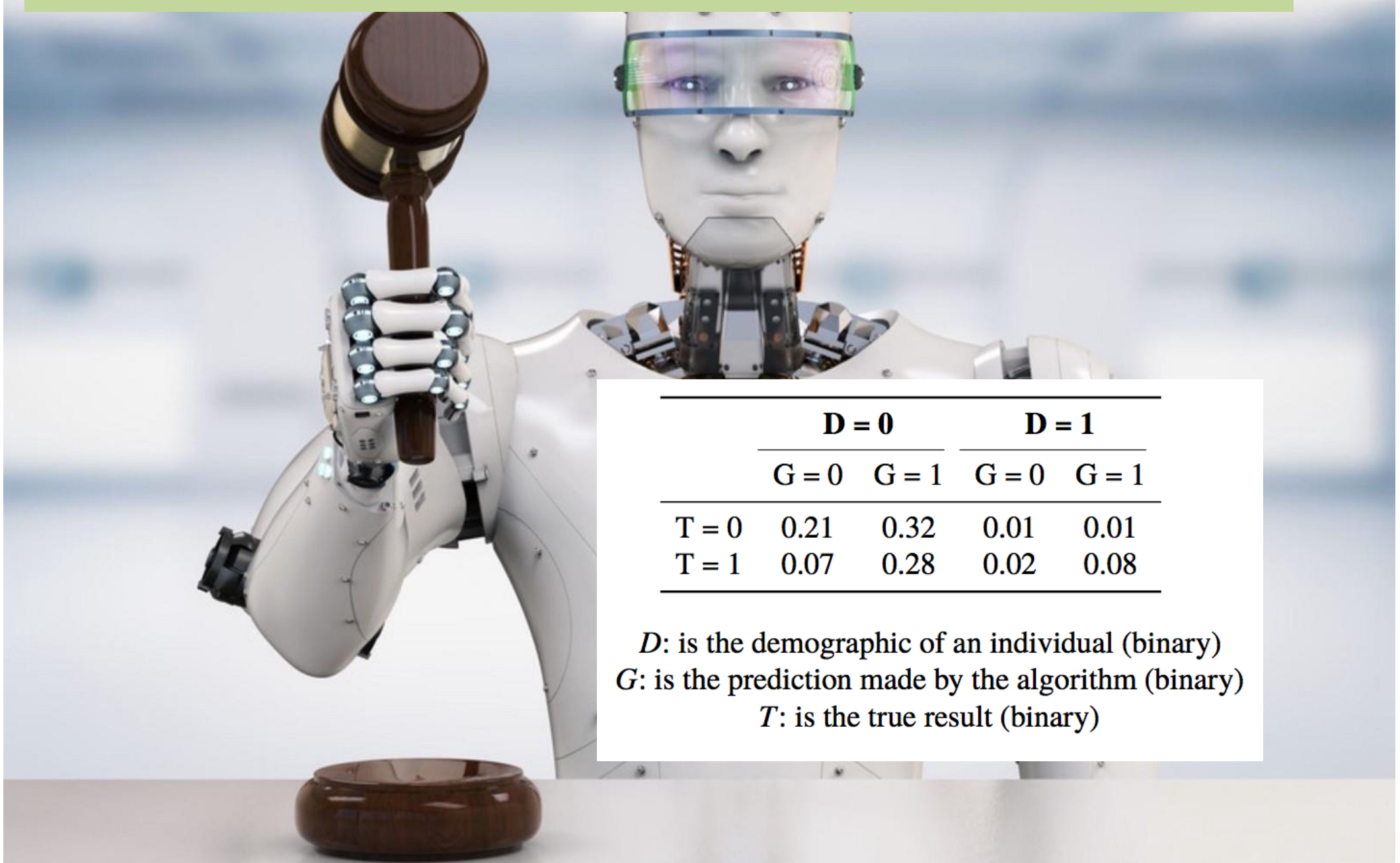
Figure 4 (b) visualizes what StACT's small relative error means in terms of predictions. Each point in the plot is a single patient. The x-axis is the true acuity of the patient and the y-axis is the predicted accuracy. We can qualitatively observe that the predictions are often accurate, there are no truly erroneous predictions, and that the exam is similarly accurate for patients of all visual acuities.

Moreover, as seen in Figure 4 (a), StACT's significant improvement in error rate holds even when the length of the exam is increased. It is also evident that increasing exam length reduces our error rate: if we increase the exam length to 200 letters, the average error of StACT falls to 0.020. While this is highly accurate, its far too long an exam, even for patients who need to know their acuity to high precision.

StACT Star Exam. Our primary experiments had a fixed



Algorithmic Fairness



	D = 0		D = 1	
	G = 0	G = 1	G = 0	G = 1
T = 0	0.21	0.32	0.01	0.01
T = 1	0.07	0.28	0.02	0.08

D: is the demographic of an individual (binary)
G: is the prediction made by the algorithm (binary)
T: is the true result (binary)

Other Tasks, Other Models

New!

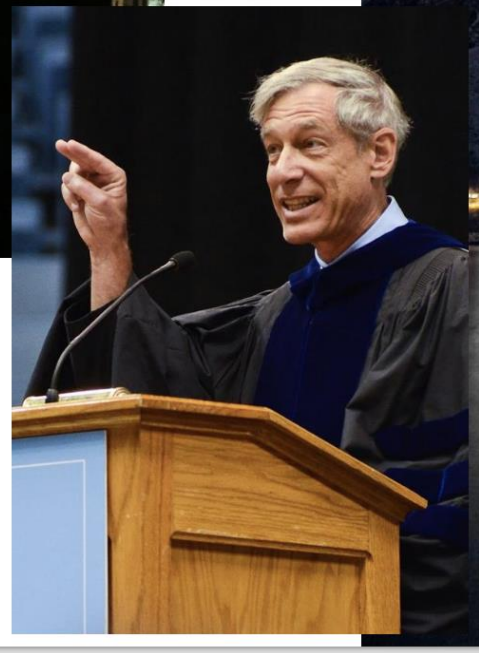
Model	Train Accuracy	Test Accuracy
Baseline	0.6138	0.6300
Logistic Regression	0.7300	0.7200
Naive Bayes	0.7275	0.7200
Decision Tree	0.7975	0.6150
Random Forest	0.7950	0.7100
Gradient Boosting	0.7738	**0.7250**
AdaBoost	0.7588	0.7100

The Kaggle Champion

Reinforcement!

Night Sight

New!



Mark Levoy, Stanford Emeritus Professor

https://static.googleusercontent.com/media/hdrplusdata.org/en//hdrplus.pdf

Wisdom of the Crowds

New!

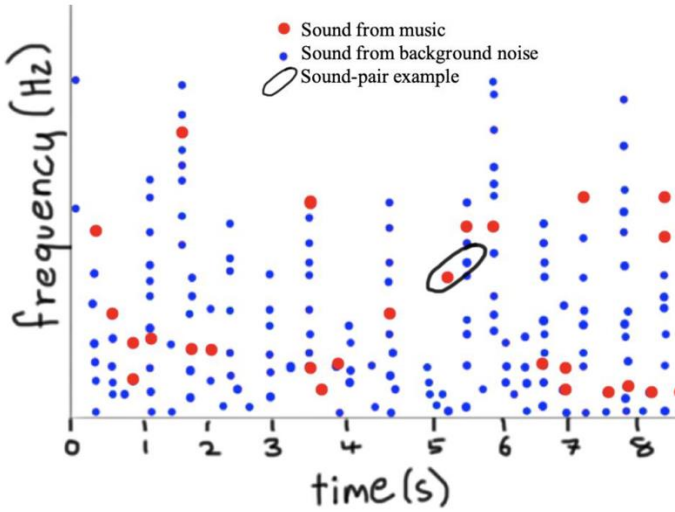
In 1999, what animal was taken off the U.S. Endangered species list aster 29 years?

A:

B: Peregrine Falcon

C: Humpback Whale

D:



Review PSet

New!

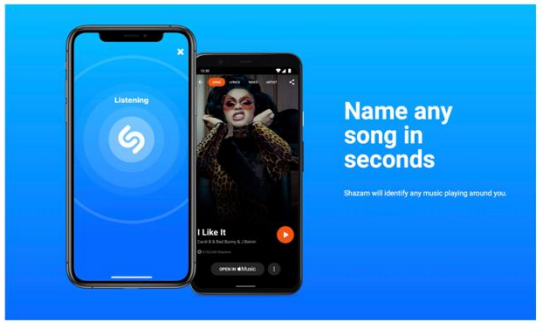
Pset 7 - Review GetMNevNUc

cs109psets.netlify.app/fall25/pset7/shazam

PS7

Shazam

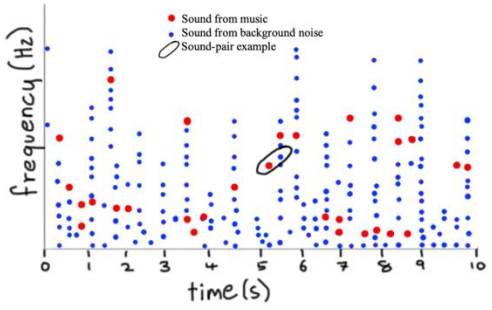
Shazam is trying to detect which song is playing in a noisy bar. We will explore how "wisdom of the crowds" effects allow it to accurately identify the song even when there is far more noise than music.



When Shazam runs its acoustic fingerprinting algorithm on an audio recording, it extracts many short "notes", represented as (frequency, time) tuples, from the recording.

You are in a really loud restaurant. In the current audio recording, there are 5000 notes from background noise and **only** 25 notes from the song. Is it still possible to identify the song?

Shazam lets every pair of notes vote for which song it thinks is playing. There are (5025 choose 2) such pairs.



Previous Question

Next Question

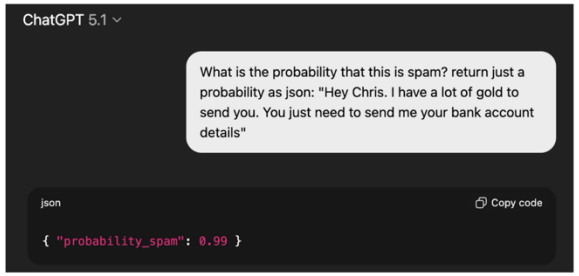
Pset 7 - Review GetMNevNUc

cs109psets.netlify.app/fall25/pset7/calibration

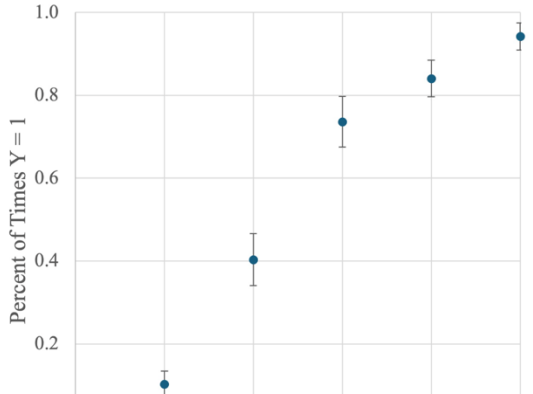
PS7

Calibrating ChatGPT

Imagine you use ChatGPT (though you could use any prediction for that matter) to make a prediction for a datapoint with features **x** and ChatGPT responds to your prompt with a probability that the label is a 1, $P(Y = 1|X = x)$. You hesitate. Are those probabilities trustworthy? In a situation like this, it might be really important to also know how *calibrated* the probabilities are. For example, if ChatGPT says that $P(Y = 1|X = x) = 0.8$ does that mean there really is an 80% chance the label is a 1?



More formally, how can we check if the probabilities output from a model are "calibrated"? To evaluate if a model is calibrated we first group datapoints together based on the $P(Y = 1|X = x)$ probability output given by the model. Then, among those groups we check how often the label is 1. Here is an example of calibration curve which plots probability groupings on the x-axis and fraction of times that the label was 1 on the y-axis:



Pset 7 - Review GetMNevNUc

cs109psets.netlify.app/fall25/pset7/flo_bayesian

PS7

Bayesian Flo

As we saw in section Flo models each user's cycle lengths! In this problem we are going to revisit the same task, but this time we are going to approach parameter estimation from a Bayesian perspective.

We will slightly simplify the formula to make the math less long. Assume that the length of someones period X is given by the following pdf which is based off a single parameter β:

$$f(X = x \mid \mathrm{B} = \beta) = \beta(x - 27)^{\beta-1}e^{-(x-27)^{\beta}}$$

For a new user, you only observe **two** cycle lengths $X_1 = 28$ and $X_2 = 29$, which means that you don't have enough data to estimate β using MLE (in section we used MLE, but we had many more cycles)! How can we estimate β in this low data situation? Inference. Just like in the Beta distribution derivation.

Good news! Based on historical data, we have a strong prior on the values that β can take! For thousands of past users, we have learned individual values for β. We then treat those values as samples from a random variable B. After analysing all the data, we found that B is a Gamma. We then used MLE to fit the Gamma distribution which resulted in the following nice equation for our prior belief in B:

$$P(\mathrm{B} = \beta) = \frac{1}{2}\beta^2 e^{-\beta}$$

Given the two observations, use inference to compute your updated belief in B

$$P(\mathrm{B} = \beta \mid X_1 = 28, X_2 = 29)$$

Allow for a normalization constant in your answer. You can assume that the two observations (X_1 and X_2) are independent if you know the value B.

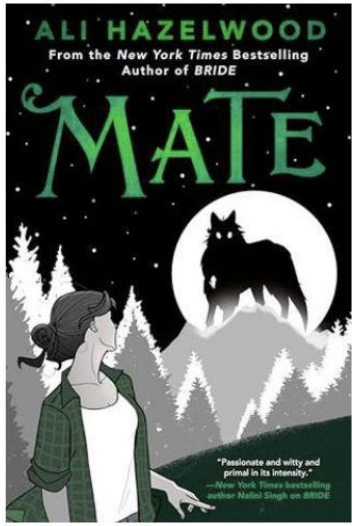
Answer Editor

Numeric Answer:

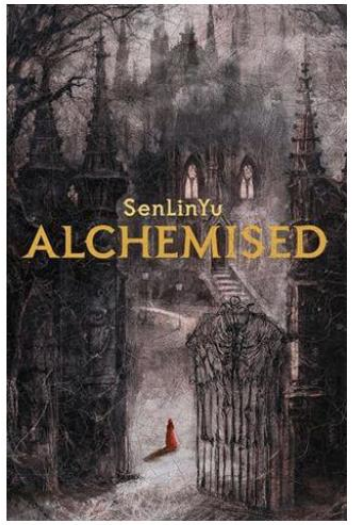
Explanation:

Block LaTeX

Which Book?



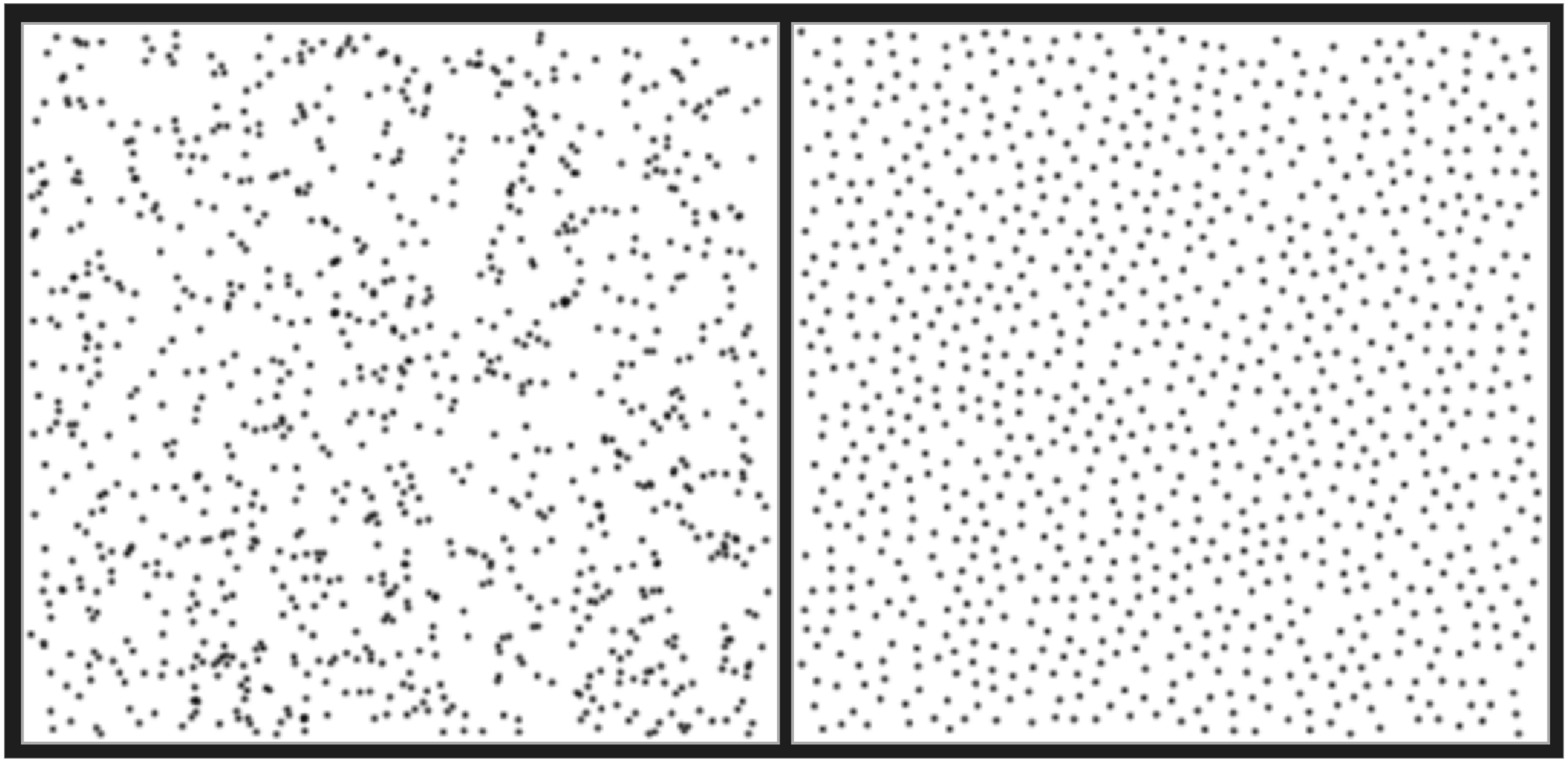
Rating	1	2	3	4	5
Count	0	0	0	1	4



Rating	1	2	3	4	5
Count	0	1	2	20	200

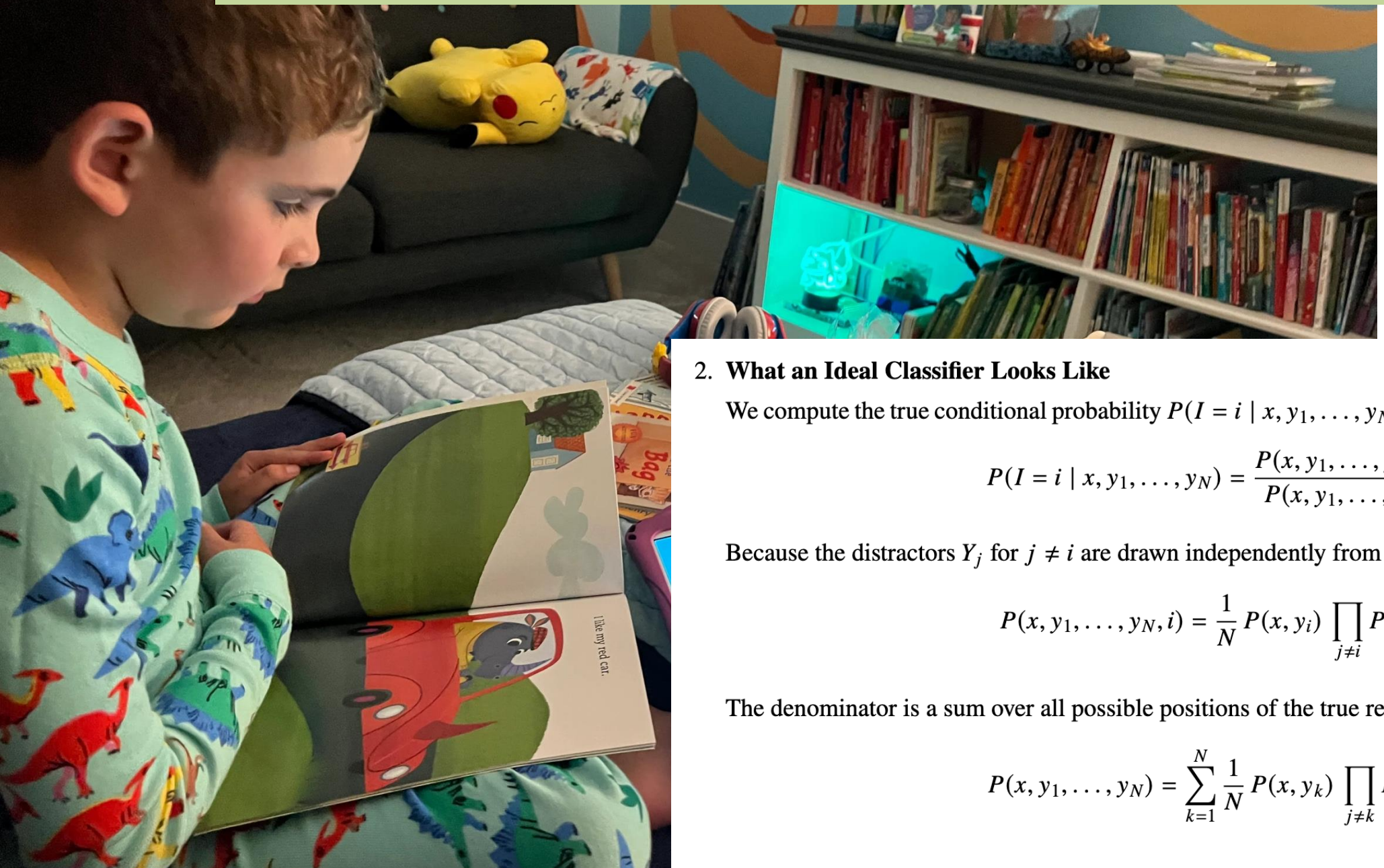
Which is Poisson?

New!



Math -> Prompt

New!



2. **What an Ideal Classifier Looks Like**

We compute the true conditional probability $P(I = i \mid x, y_1, \ldots, y_N)$. Using conditional probability,

$$P(I = i \mid x, y_1, \ldots, y_N) = \frac{P(x, y_1, \ldots, y_N, i)}{P(x, y_1, \ldots, y_N)}.$$

Because the distractors Y_j for $j \neq i$ are drawn independently from $P(Y)$ and independently of X,

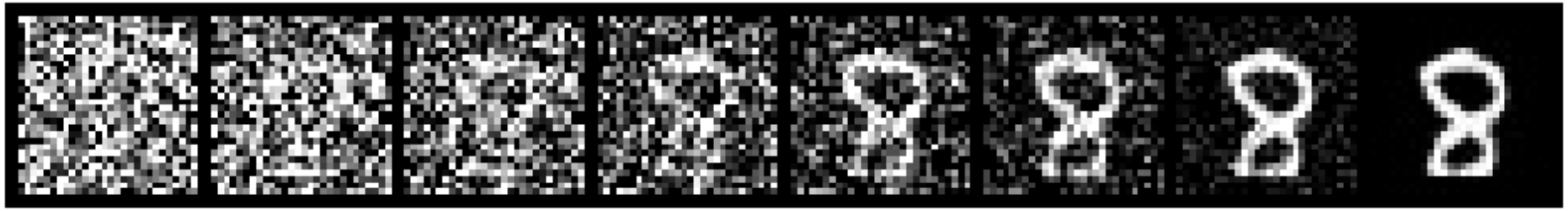
$$P(x, y_1, \ldots, y_N, i) = \frac{1}{N} P(x, y_i) \prod_{j \neq i} P(y_j).$$

The denominator is a sum over all possible positions of the true record:

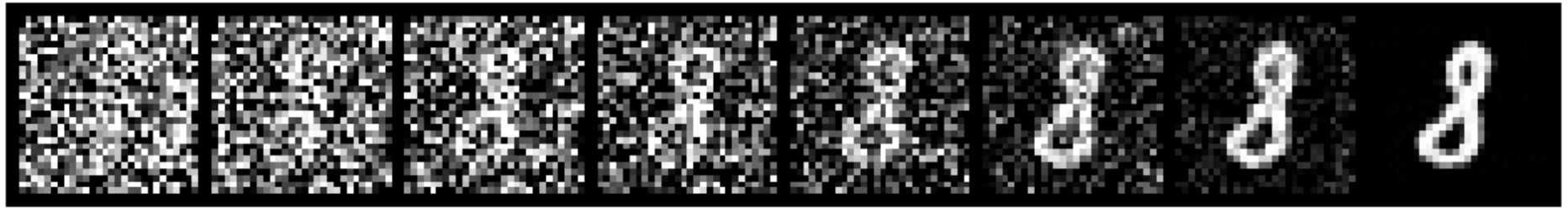
$$P(x, y_1, \ldots, y_N) = \sum_{k=1}^{N} \frac{1}{N} P(x, y_k) \prod_{j \neq k} P(y_j).$$

Diffusion!

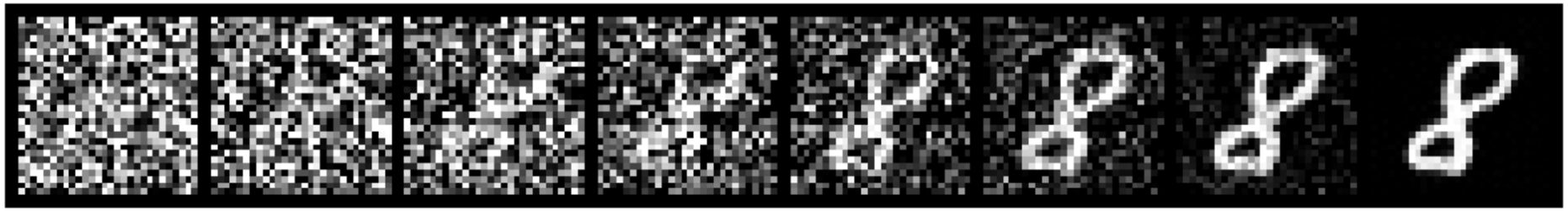
Reverse denoising (class 8)



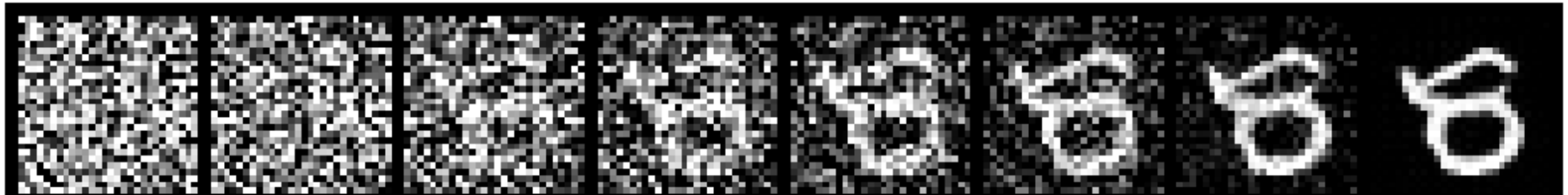
Reverse denoising (class 8)



Reverse denoising (class 8)



Reverse denoising (class 8)



By the numbers

~50 Major Keys

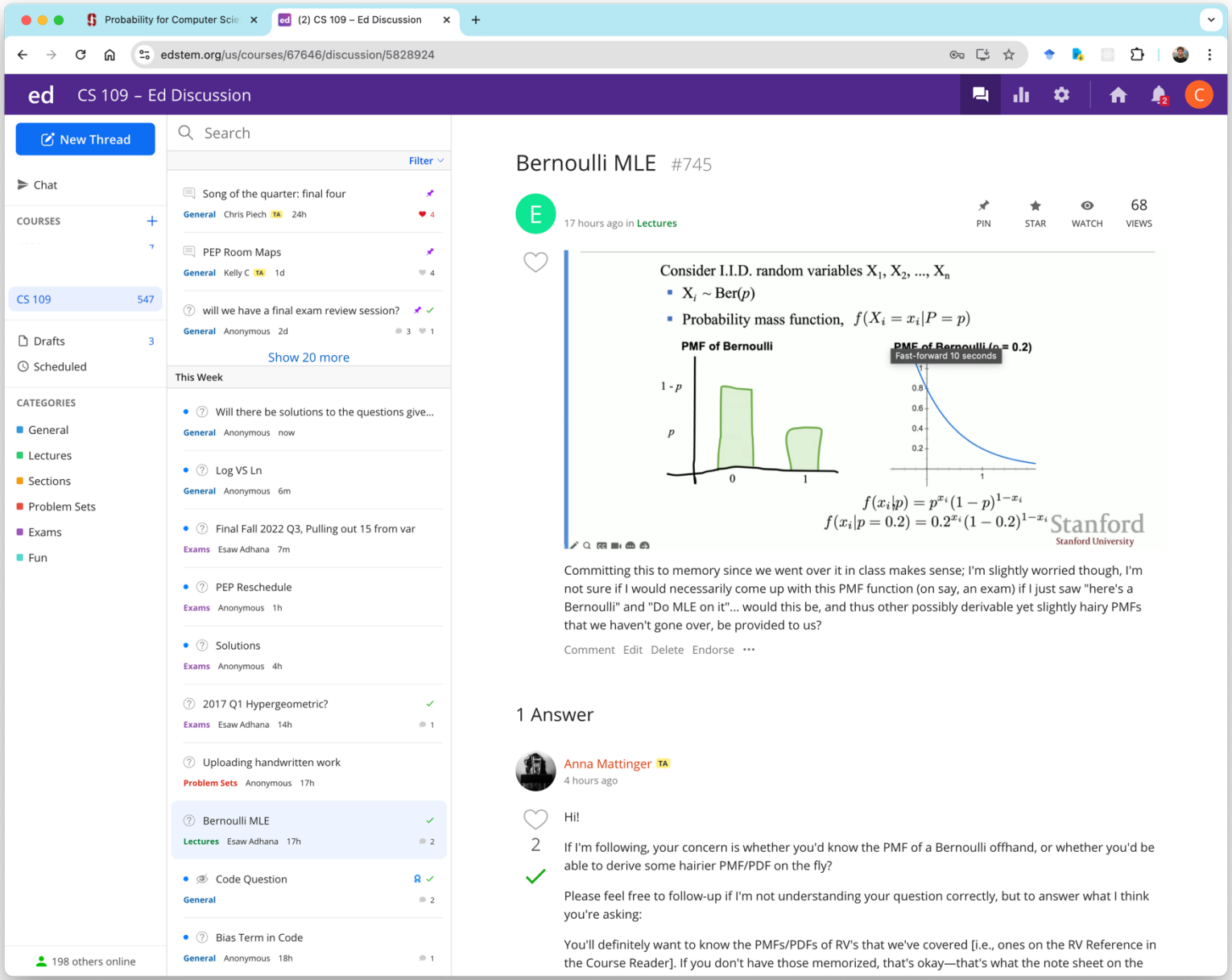
By the Central Limit Theorem, the mean of IID variables are distributed normally. As *n* -> ∞

$$\bar{X} \sim N(\mu, \frac{\sigma^2}{n})$$

~600 Fruit

0 Fruit Related Injuries

1173 EdStem Answers



15 New Course Reader Chapters

New Chapters

Probability for Computer Scie

probabilitycoders.stanford.edu/fall25/dart_logo

CS109 Logo

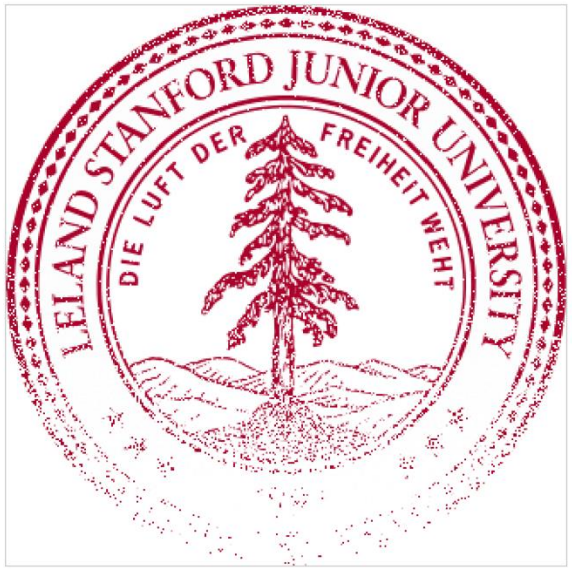
id you know that we used a stochastic process to generate the CS109 logo? Throw 300,000 darts at a picture of the Stanford seal and only keep the pixels that are hit by at least one dart. Each dart has it's x-pixel and y-pixel chosen independently at random from gaussian distributions:

Let S and be a constant that equals the size of the logo (its width is equal to its height).
For the Stanford seal $S = 300$.

Let X be a random variable which represent the x-pixel. $X \sim \mathcal{N}\left(\mu = \frac{S}{2}, \sigma^2 = (\frac{S}{2})^2\right)$

Let Y be a random variable which represents the y-pixel. $Y \sim \mathcal{N}\left(\mu = \frac{S}{3}, \sigma^2 = (\frac{S}{5})^2\right)$

You can re-run the process by pressing the Play animation button:



Play animation

Throw 5,000 darts

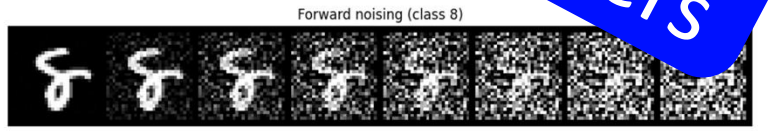
Reset animation

Darts thrown: 300,000 / 300,000

Here is the probability density function of X (left) and Y (right) on their own:

probabilitycoders.stanford.edu/fall25/diffusion

Here is what adding noise looked like for my hand draw 8s:



The Reverse Process

We have now redefined our task. Could we learn to take a square image with randomly selected pixels, and learn to reverse the noise adding task until we are left with a picture of a tree? To reverse this process, we need to learn the conditional distribution $P(x_{t-1}|x_t)$. Here's the surprising part:

Diffusion Critical Fact: Reverse Process is also Gaussian

If the noise variance σ^2 is small enough, the distribution of $X_t|X_{t+1}$ can be approximated as:

$$X_t|X_{t+1} \sim N(\mu, \sigma^2),$$

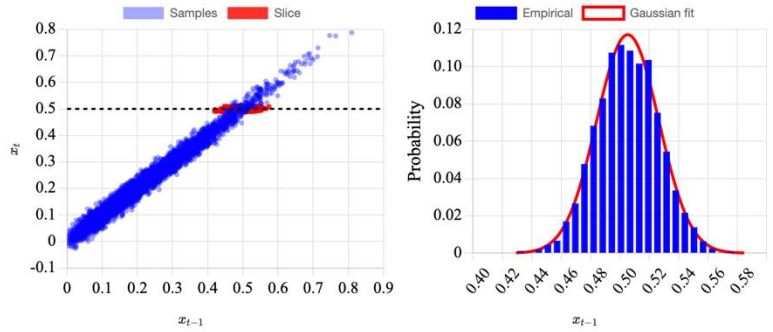
where:

- $\mu =$ is the mean of the Gaussian, which is not known (but we assume is a complex function of the image)
- σ^2 is the variance of the noise that we selected when adding in Gaussian noise to our images

Here is a demo of the Diffusion Critical Fact. In this demo I first sample half a million values from $X_t \sim \text{Beta}(2, 7)$ (the critical fact doesn't depend on the distribution of X_t). For each point I then chose a value X_{t+1} using the diffusion process: $X_{t+1} = X_t + N_t$ where $N_t \sim N(0, \sigma^2)$.

On the left is the joint of X_t and X_{t+1}. On the right is the posterior of X_t after observing that $X_{t+1} = k$. Notice how Normal it looks when σ is small. Then try setting σ to be 2 and see that the Normal shape goes away.

Joint $P(X_{t-1} = x_{t-1}, X_t = x_t)$ **Posterior** $P(X_{t-1} = x_{t-1} \mid X_t \approx k)$



k: 0.5 σ: 0.02 Slice width: 0.01 Samples in slice: **3600**

A proof sketch of the Diffusion Critical Fact is provided later on in this chapter. It uses inference (and Taylor Series

52+ Personal Challenges

1 New Counter on the Way

I had an important job!
I hope you think I did it justice

Wild time. Incredible school at which to study probability


thank you