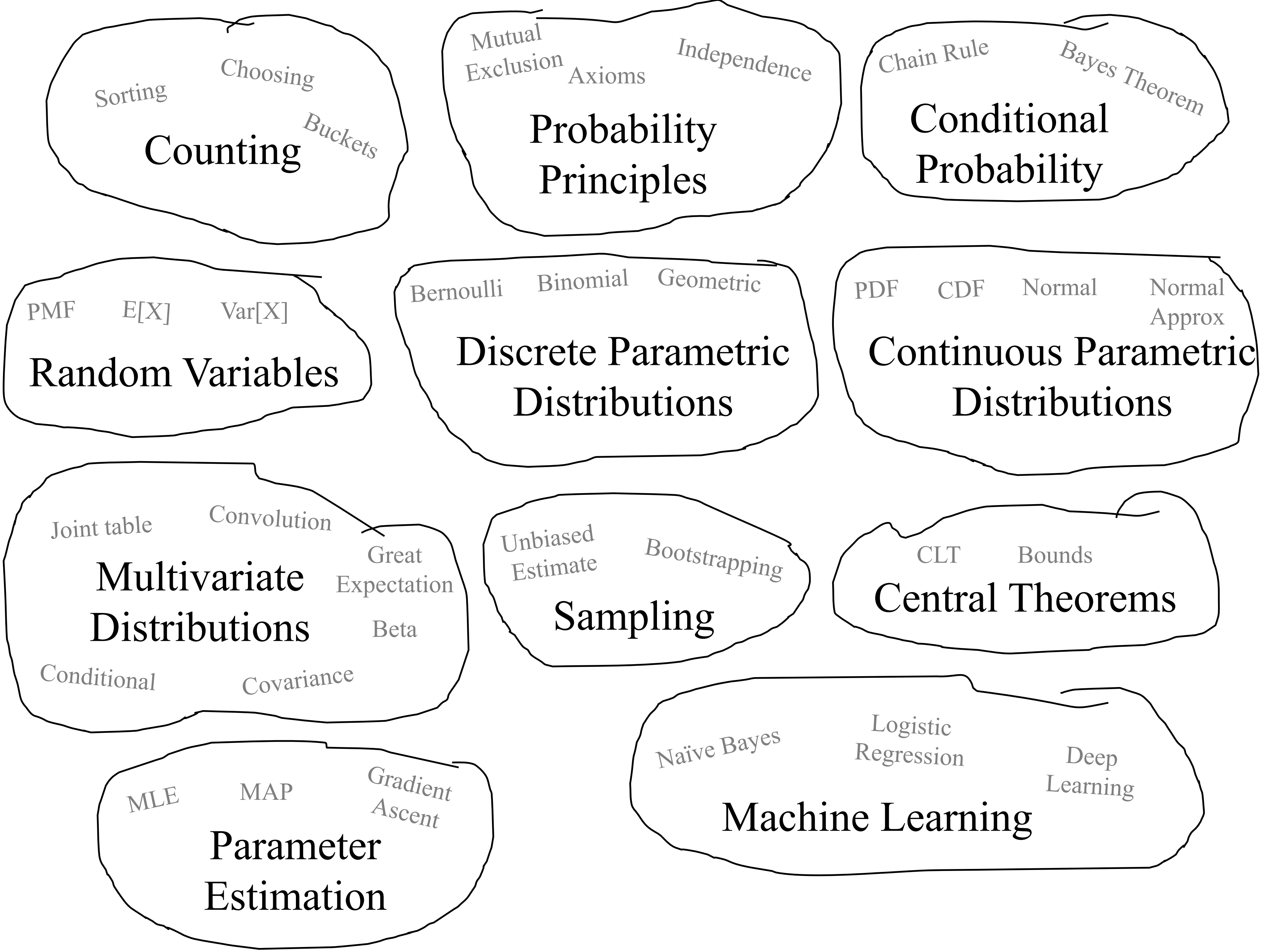
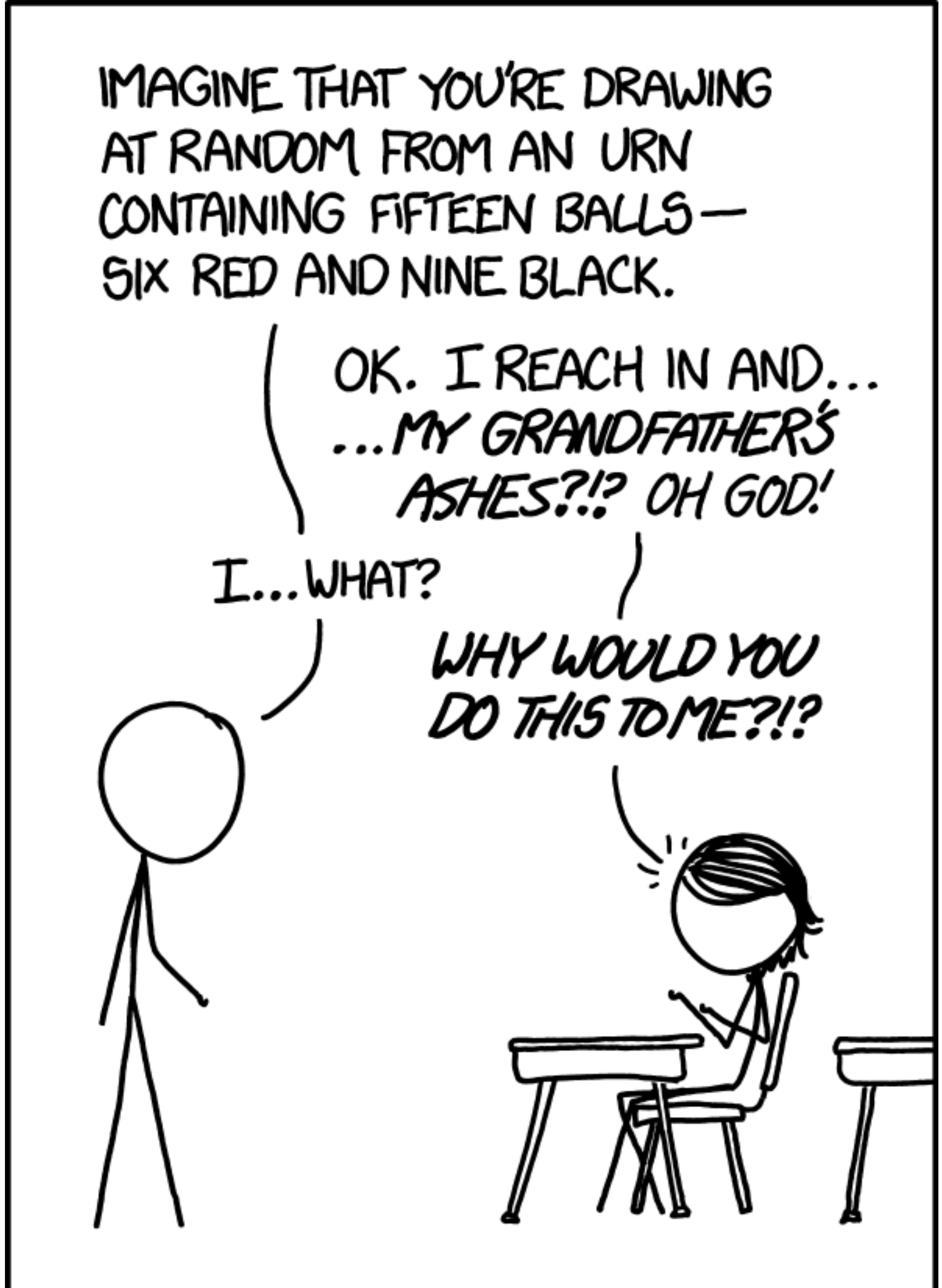
CS109 Summary

Chris Piech
CS109, Stanford University


Counting
Sorting
Choosing
Buckets
Probability Principles
Mutual Exclusion
Axioms
Independence
Conditional Probability
Chain Rule
Bayes Theorem
Random Variables
PMF
E[X]
Var[X]
Discrete Parametric Distributions
Bernoulli
Binomial
Geometric
Continuous Parametric Distributions
PDF
CDF
Normal
Normal Approx
Multivariate Distributions
Joint table
Convolution
Great Expectation
Beta
Conditional
Covariance
Sampling
Unbiased Estimate
Bootstrapping
Central Theorems
CLT
Bounds
Parameter Estimation
MLE
MAP
Gradient Ascent
Machine Learning
Naïve Bayes
Logistic Regression
Deep Learning


IMAGINE THAT YOU'RE DRAWING AT RANDOM FROM AN URN CONTAINING FIFTEEN BALLS—SIX RED AND NINE BLACK.
OK. I REACH IN AND...
...MY GRANDFATHER'S ASHES?!? OH GOD!
I...WHAT?
WHY WOULD YOU DO THIS TO ME?!?

Counting Rules

Counting operations on *n* objects

Sort (order matters) (perms)

- Distinct: $n!$
- Some Distinct: $\dfrac{n!}{n_1! n_2! \dots}$

Choose *k* (combinations)

- Distinct: $\binom{n}{k} = \dfrac{n!}{k!(n-k)!}$

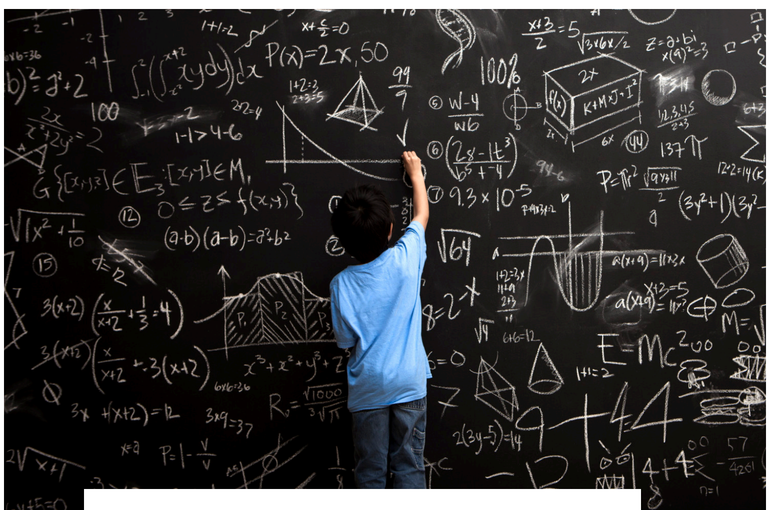
Put in *r* buckets

- Distinct: r^n
- None Distinct: $\dfrac{(n+r-1)!}{n!(r-1)!}$

Sources of Probability

1. Experimentation

2. Dataset

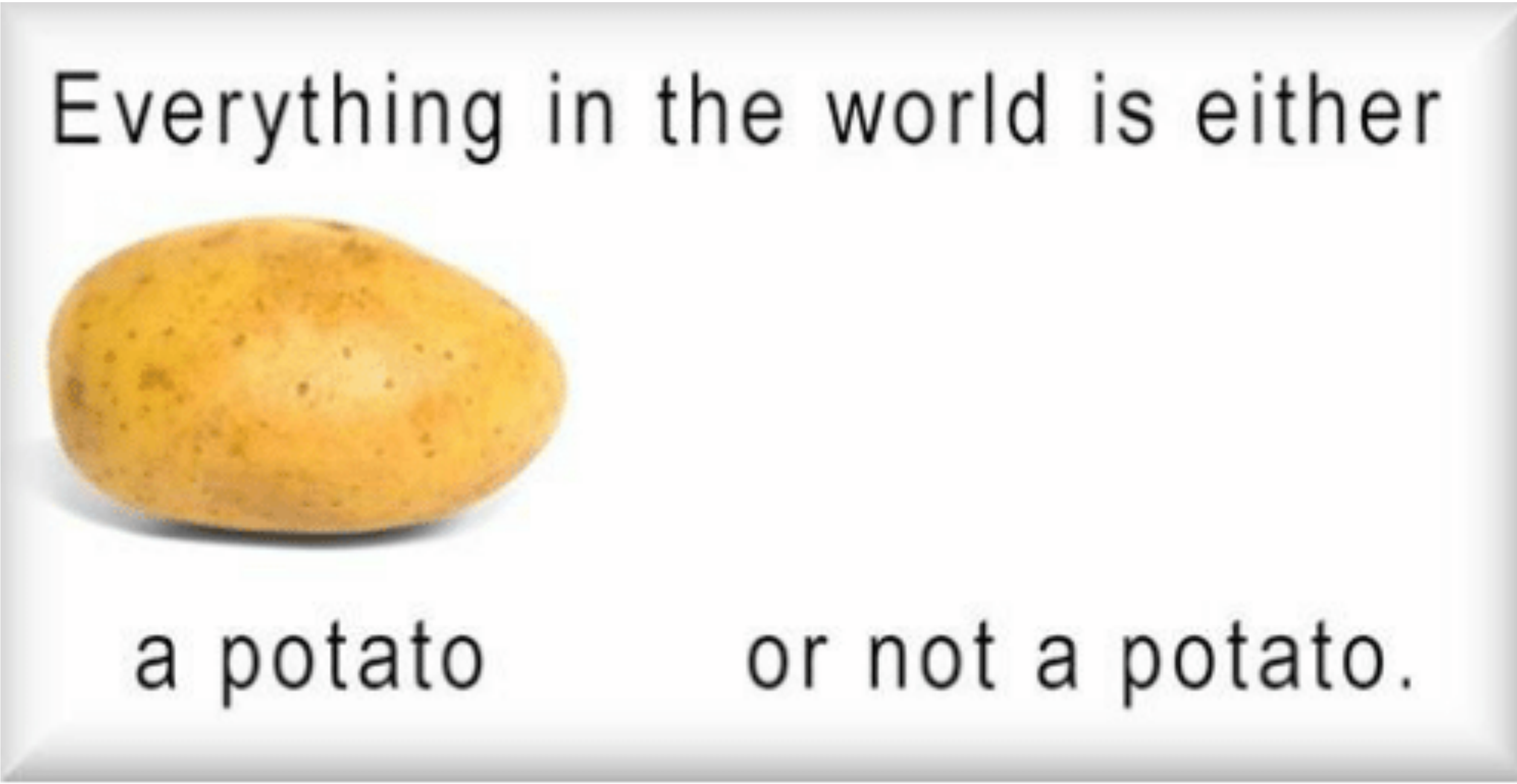


3. Analytic Solution

4. Expert Opinion

What is a Probability?

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



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

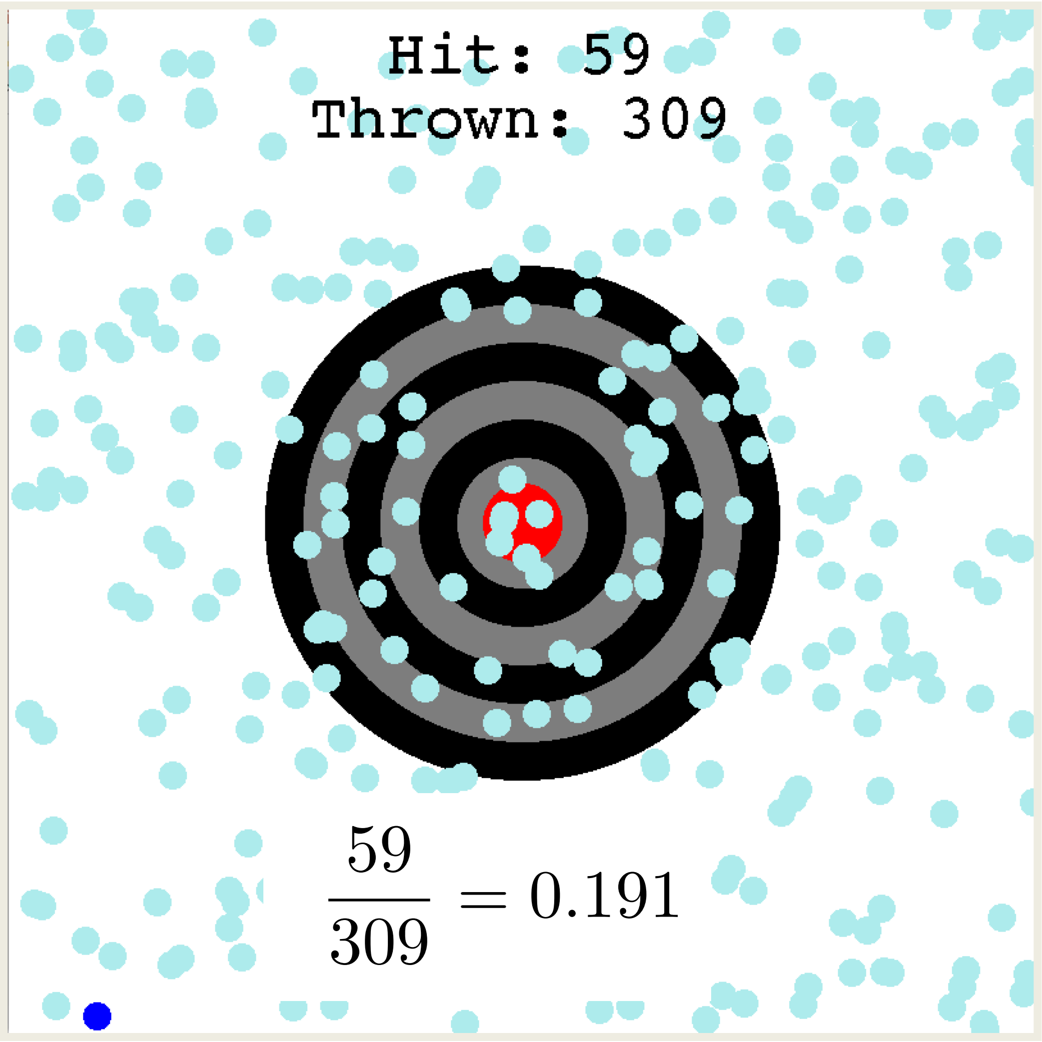
Sending Bit Strings

- Bit string with *m* 0's and *n* 1's sent on network
  - All distinct arrangements of bits equally likely
  - E = first bit received is a 1
  - F = *k* of first *r* bits received are 1's

P(E|F)?

*Think of the bits as distinct so that all outcomes are equally likely

Target Revisited



$$\frac{59}{309} = 0.191$$

Screen size = 800x800
Radius of target = 200

The dart is equally likely to land anywhere on the screen.

What is the probability of hitting the target?

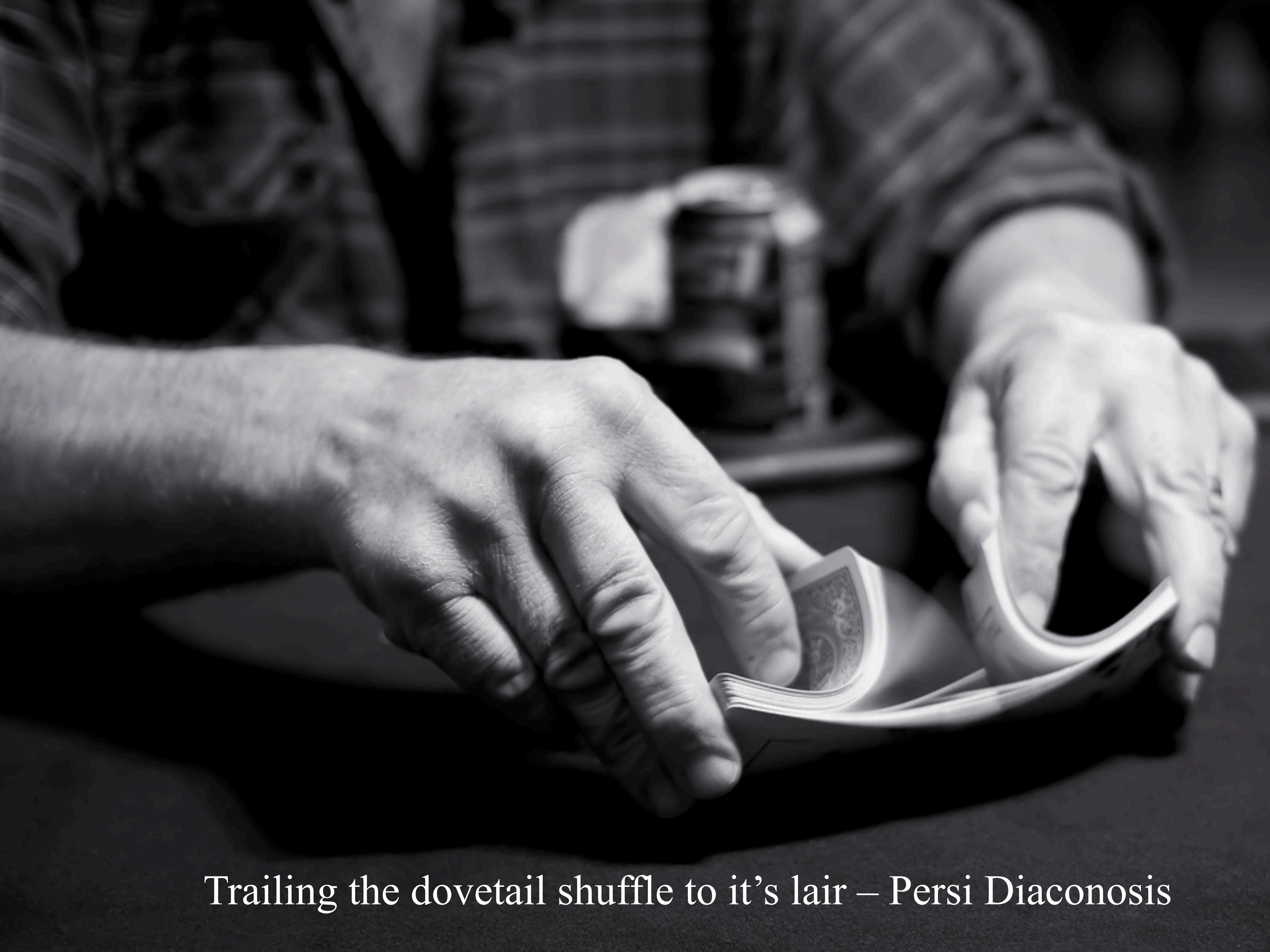
$$|S| = 800^2$$

$$|E| = \pi 200^2$$

$$p(E) = \frac{\pi \cdot 200^2}{800^2} \approx 0.1963$$

WHEN YOU MEET YOUR BEST FRIEND

Somewhere you didn't expect to.



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

Netflix and Learn

What is the probability that a user will watch Life is Beautiful?

$P(E)$

$$P(E) = \lim_{n \to \infty} \frac{n(E)}{n} \approx \frac{\#\text{people who watched movie}}{\#\text{people on Netflix}}$$

$$P(E) = 10{,}234{,}231 \,/\, 50{,}923{,}123 = 0.20$$

Netflix and Learn

What is the probability that a user will watch Life is Beautiful, given they watched Amelie?

$P(E|F)$

$$P(E|F) = \frac{P(EF)}{P(F)} = \frac{\#\text{people who watched both}}{\#\text{people who watched } F}$$

$P(E|F) = 0.42$

Let's Make a Deal

- Game show with 3 doors: A, B, and C

  - Behind one door is prize (equally likely to be any door)
  - Behind other two doors is nothing
  - We choose a door
  - Then host opens 1 of other 2 doors, revealing nothing
  - We are given option to change to other door
- Should we?
  - Note: If we don't switch, P(win) = 1/3   (random)

First Ever Sections!

Lecturer: Chris Piech
piech@cs.stanford.edu
Wed 1-3pm
Gates 193

Head TA: Ben Ulmer
ulmerb@stanford.edu

TA: Ana-Maria Istrate
aistrate@stanford.edu

TA: Gobi Dasu
gdasu@stanford.edu

TA: Julia Daniel
jdaniel7@stanford.edu

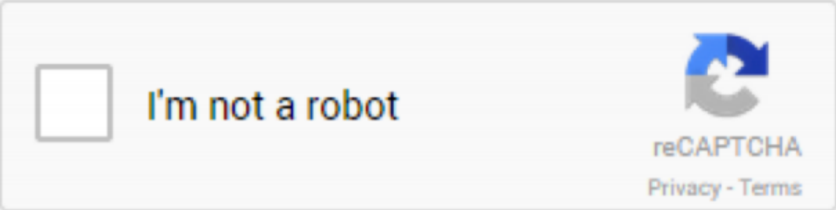
TA: Luke Johnston
lukej@stanford.edu

TA: Elliot Chartock
elboy@stanford.edu

TA: Eric Redondo
eredondo@stanford

TA: Yuling Liu
yulingl@stanford.edu

TA: Brendan Corcoran
bmc2016@stanford.edu



Bayes Theorem

$$P(F|E) = \frac{P(E|F)P(F)}{P(E)}$$

Zika Test

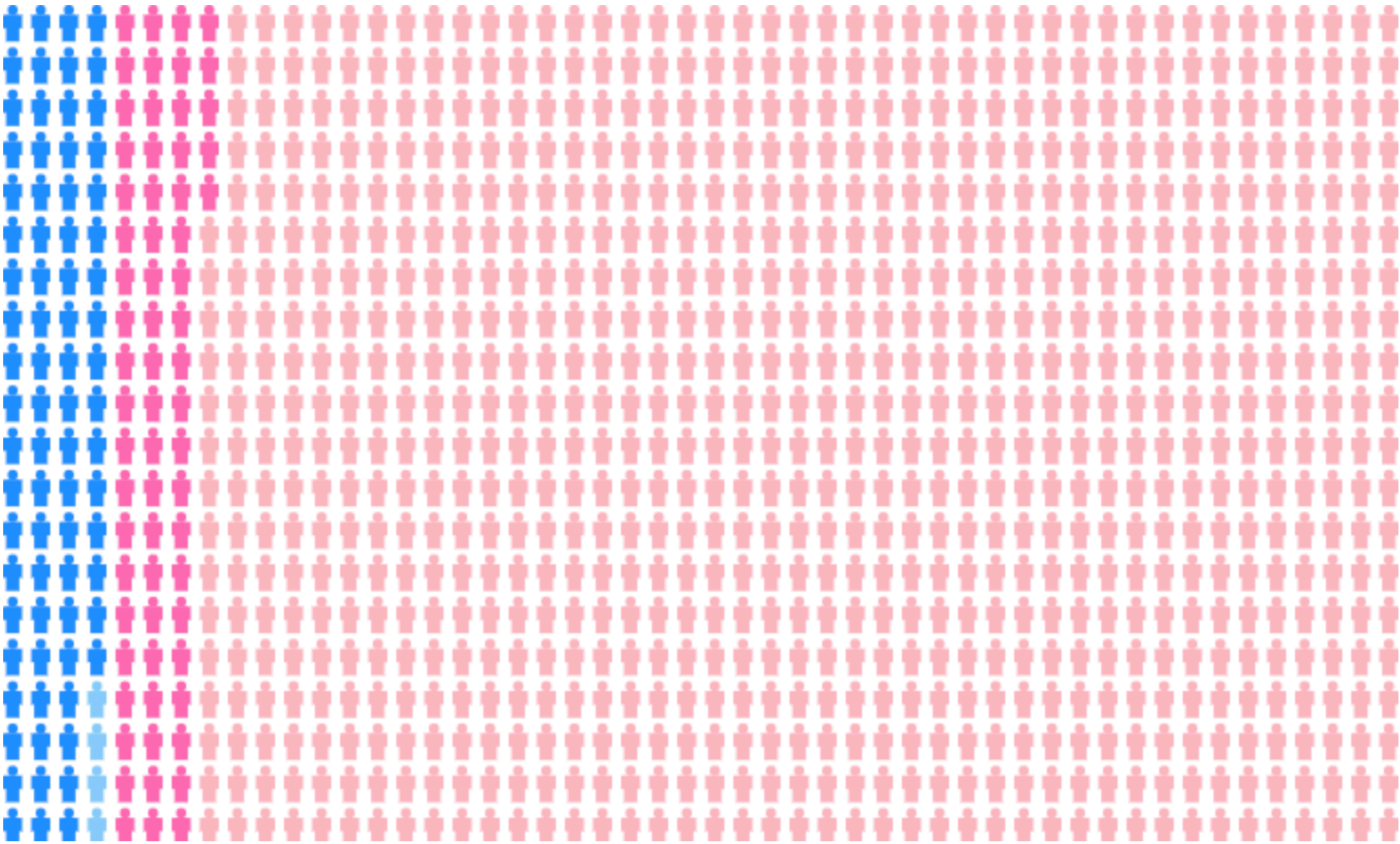
Positive Zika.

What is the probability of zika?

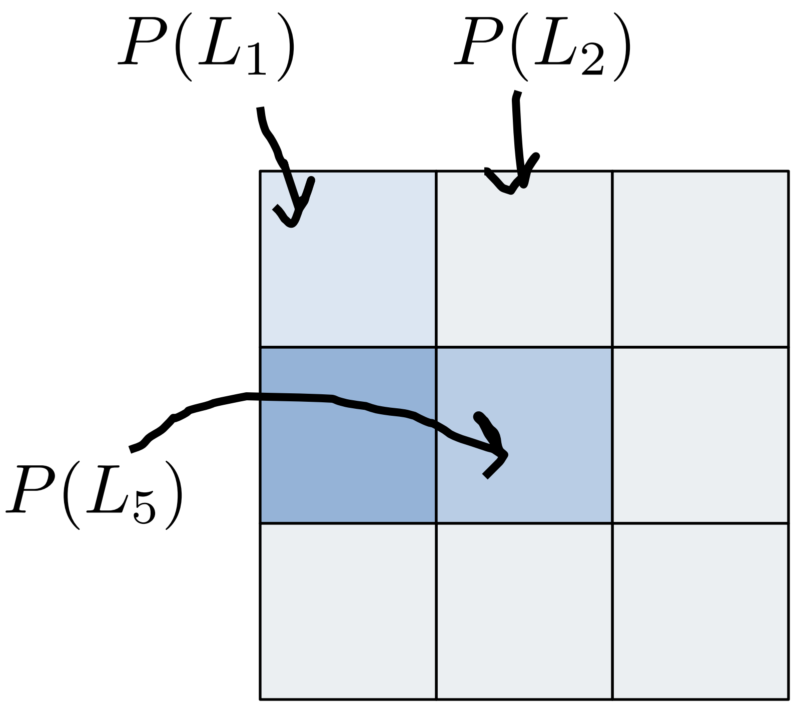
- *0.1% of people have zika*
- *90% positive rate for people with zika*
- *7% positive rate for people without zika*

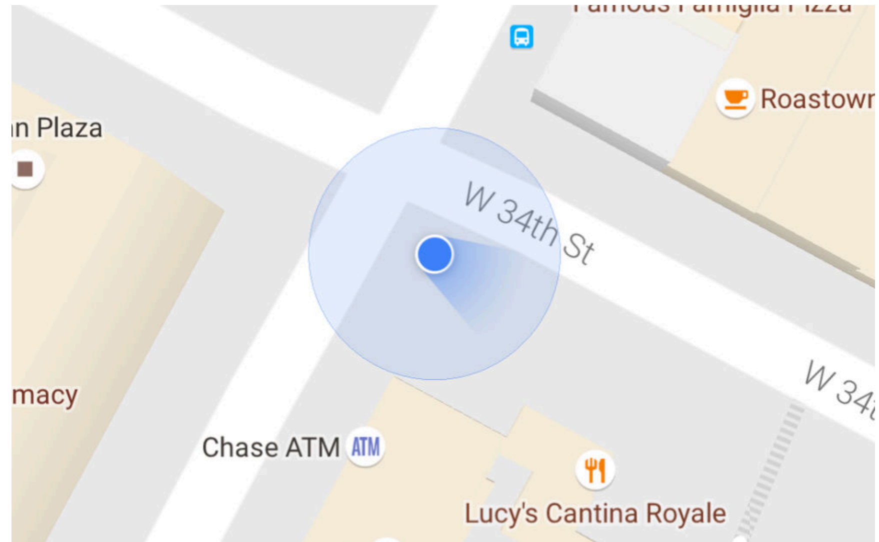
The right answer is 1%

Bayes Theorem Intuition



Update Belief



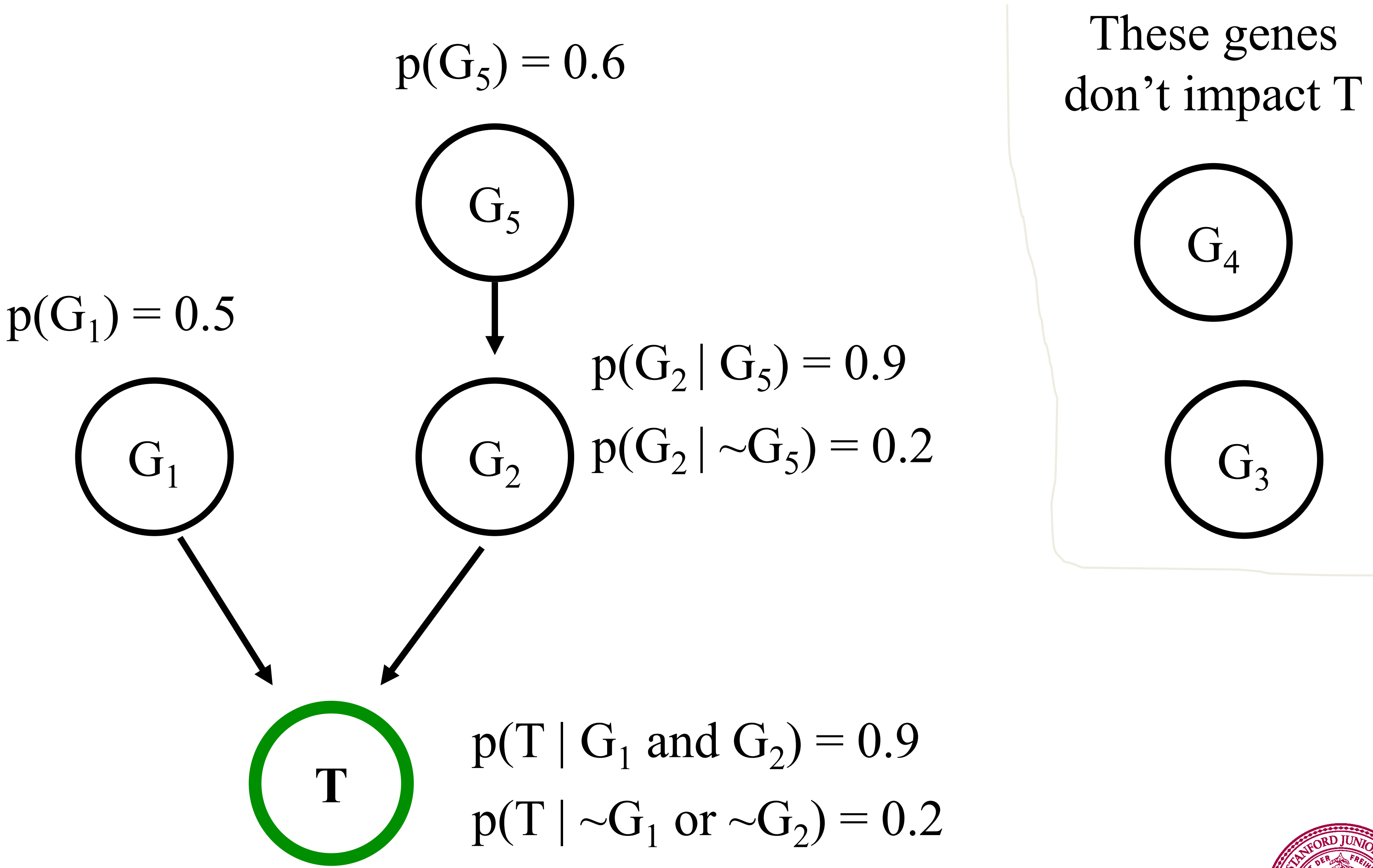


Before Observation

Independence

Recall our Ebola Bats

Discovered Pattern



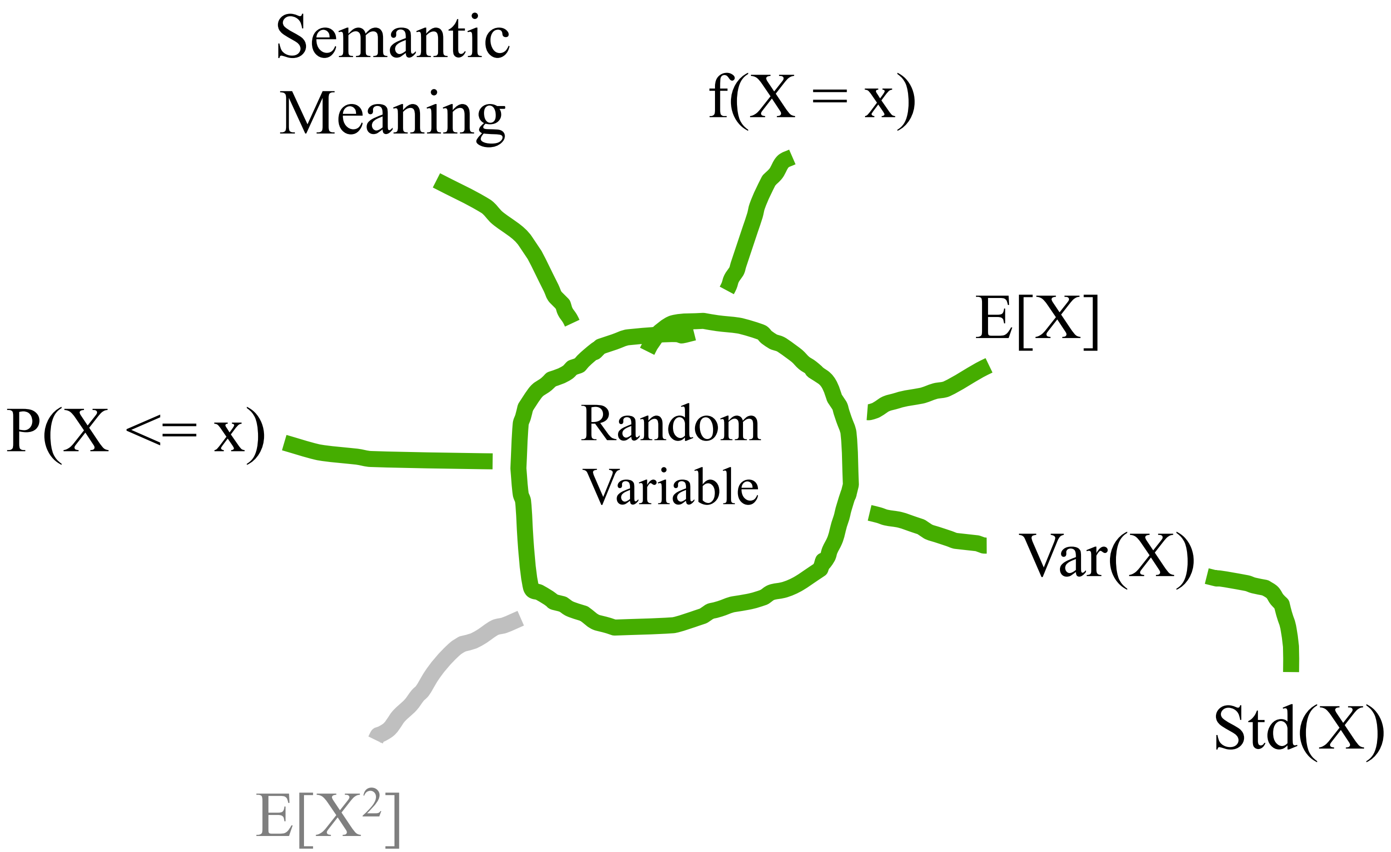
Random Variables

$$P(Y = k)$$

This is a function

For example Y is the number of heads in 5 coin flips

Fundamental Properties



Expectation

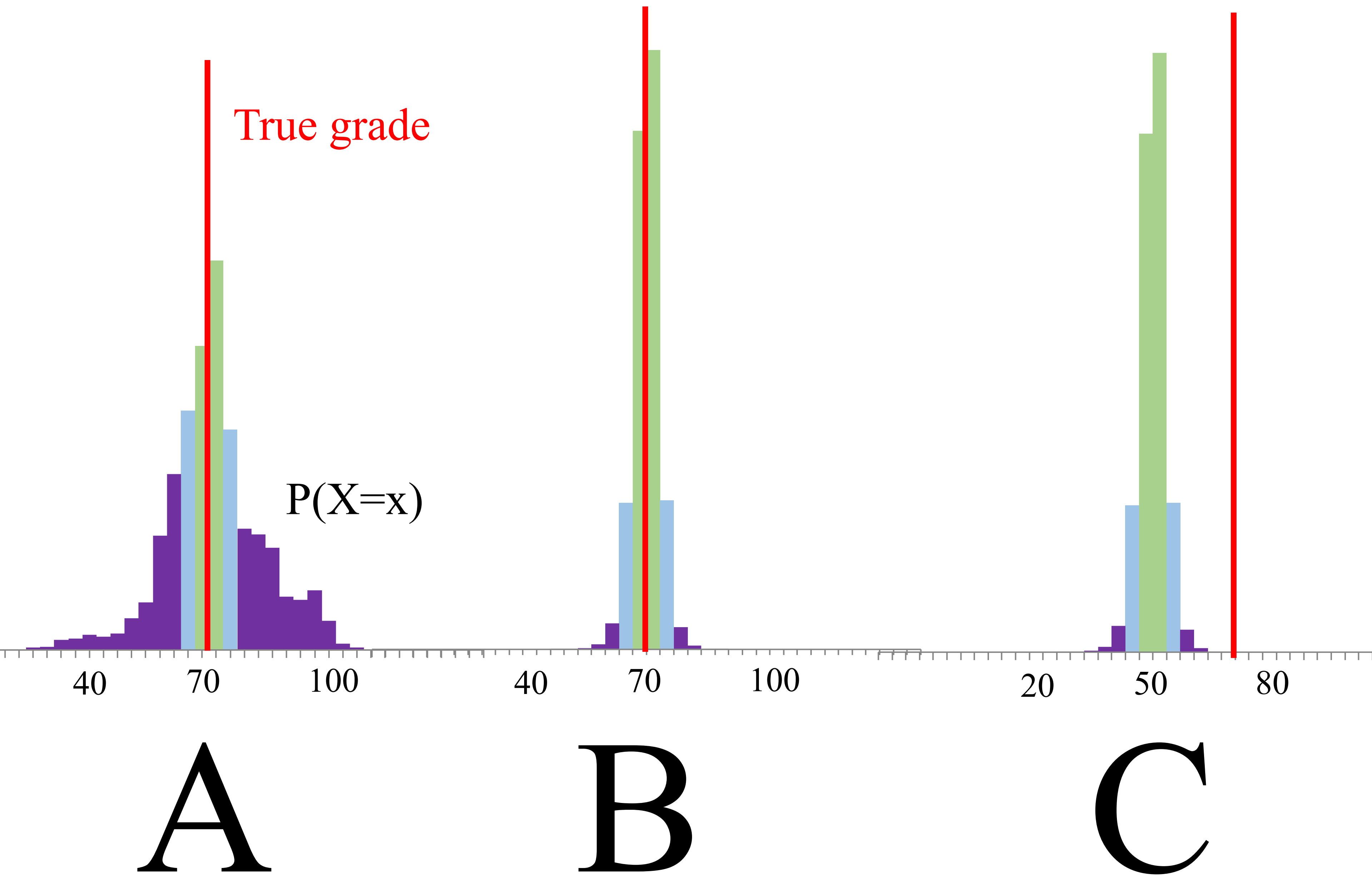
Big deal lemma: first
stated without proof

$$E[X + Y] = E[X] + E[Y]$$

Generalized: $$E\left[\sum_{i=1}^{n} X_i\right] = \sum_{i=1}^{n} E[X_i]$$

Holds regardless of dependency between X_i's

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



St Petersburg

- Game set-up
  - We have a fair coin (come up “heads” with p = 0.5)
  - Let n = number of coin flips (“heads”) before first “tails”
  - You win $\$2^n$

- How much would you pay to play?

Bernoulli

Binomial

Our random variable

Num trials

Probability of success on each trial

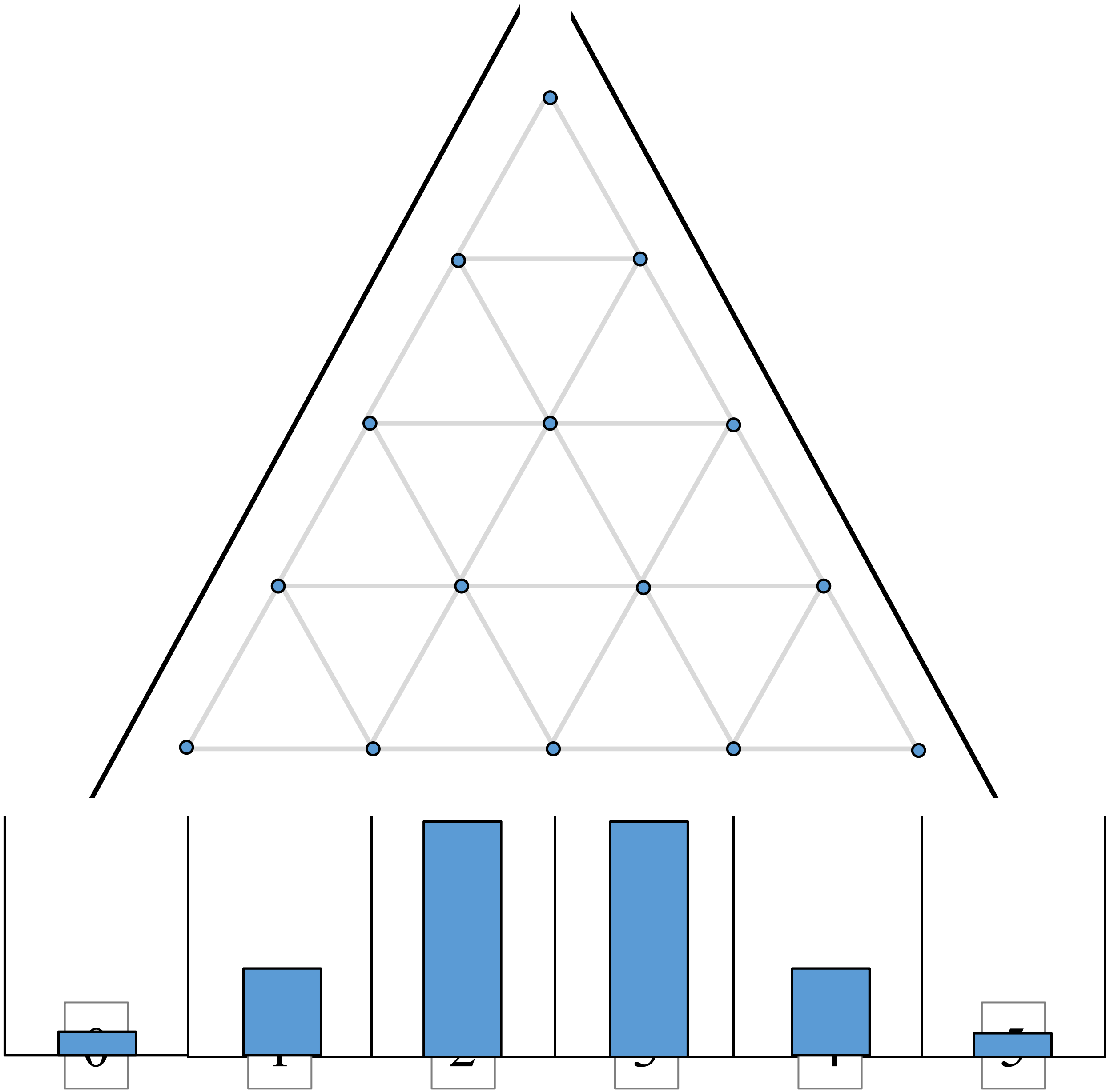
$$X \sim \mathrm{Bin}(n, p)$$

Is distributed as a

Binomial

With these parameters

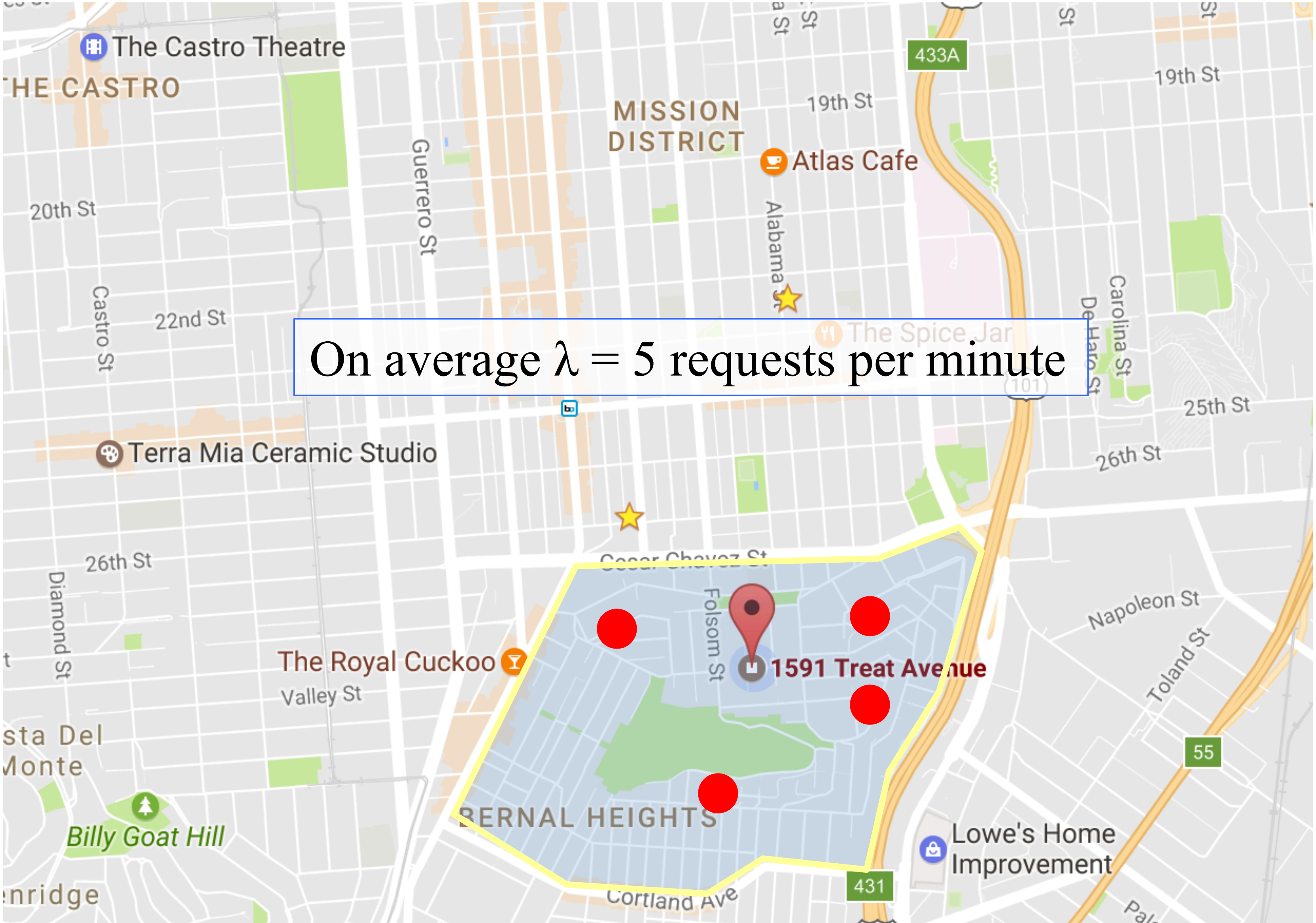
Binomial



PDF


1001

Poisson



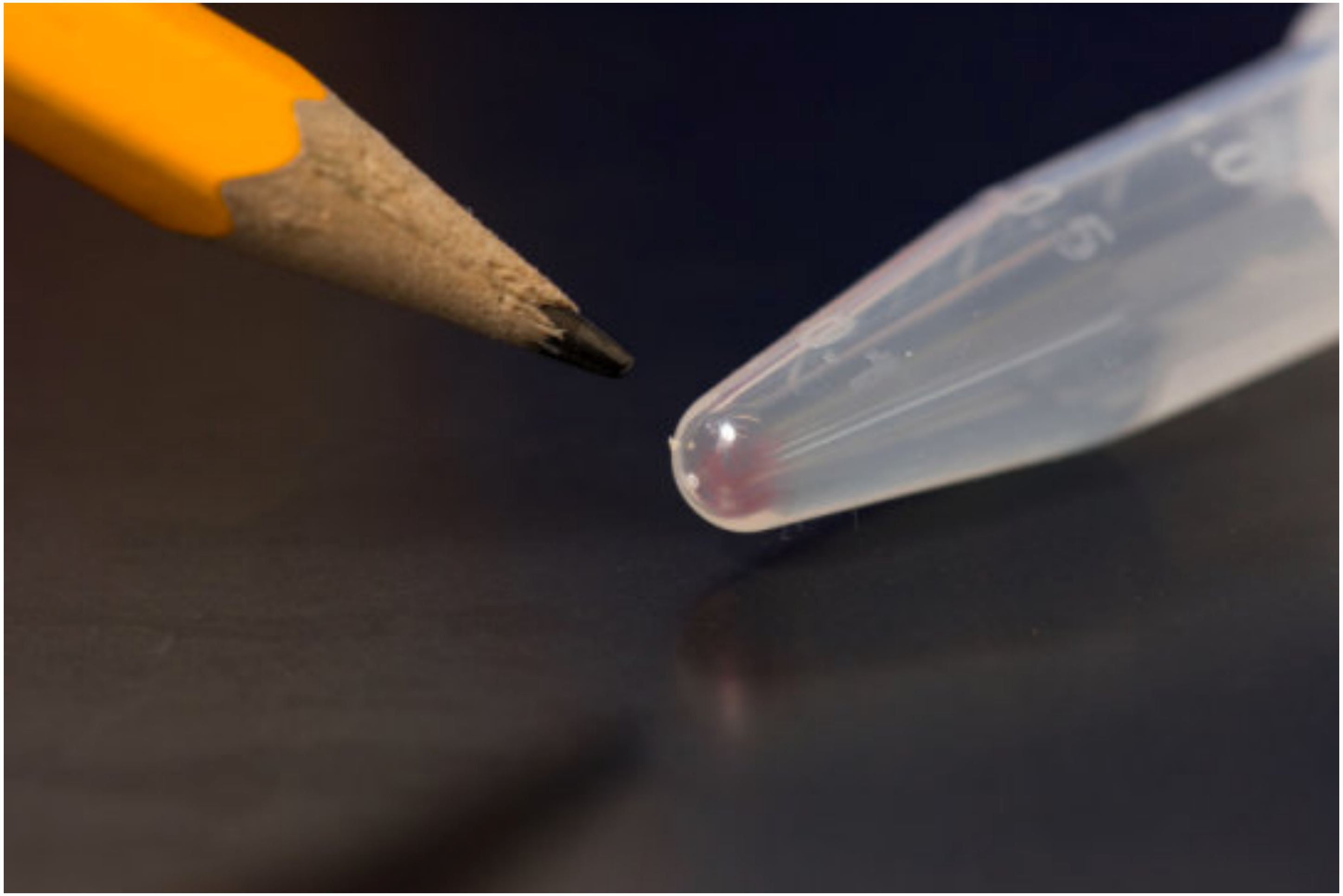
On average $\lambda = 5$ requests per minute

Geometric

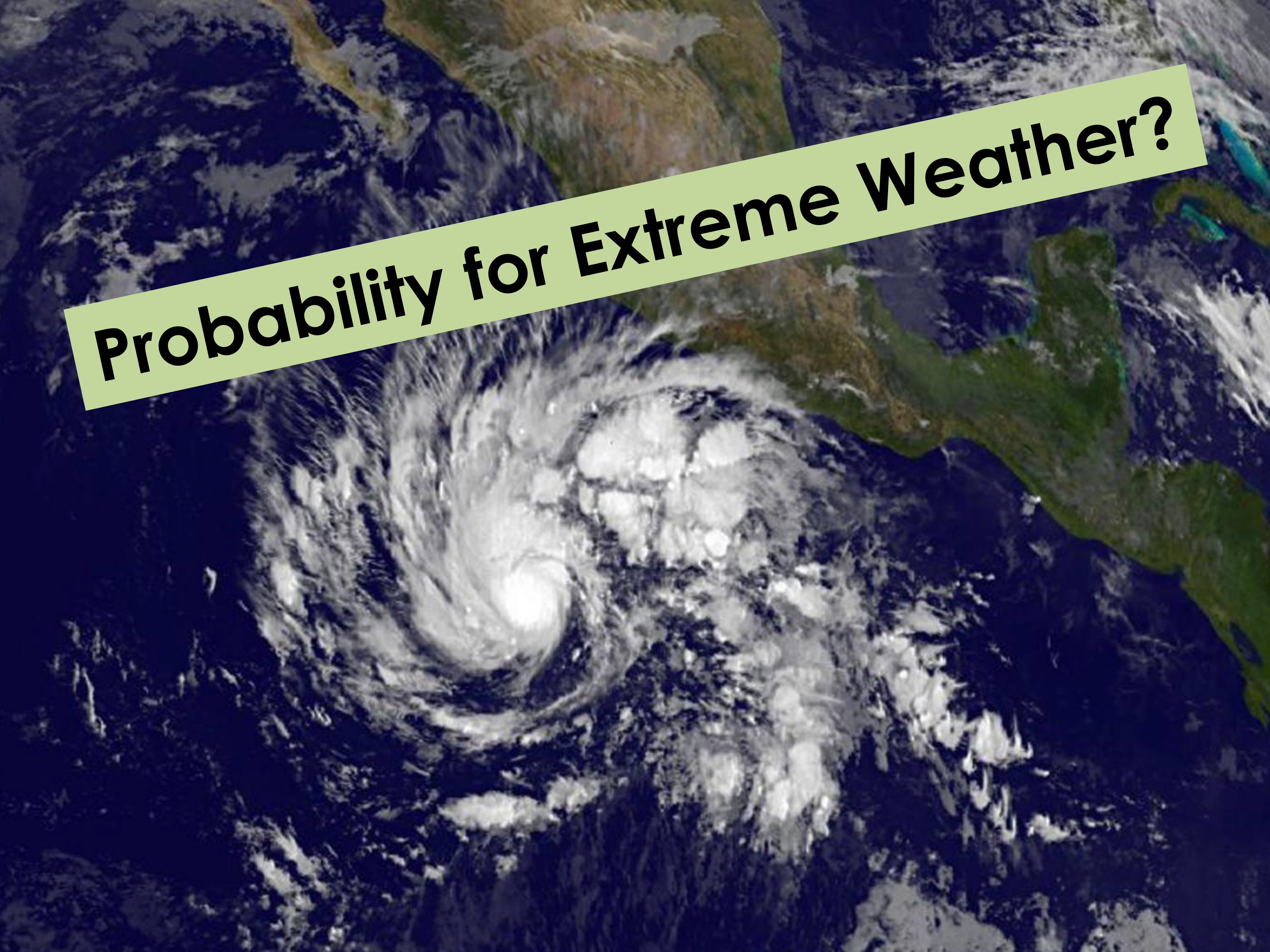
Sequence 1:

TTHHTHTTHTTTHTTTHTTTHTTHTHHT
HTTHHTTHHHTHHTHTTHTHTTHHTHHH
HTHHTHHHTHTHTTHTTHHTHTHTHTTH
TTHHTHTTHTHTHTHTHTHTHHHTHTHT
TTHHTHTHTHTHHHTTHHTHTTTHTHHT

Storing Data on DNA



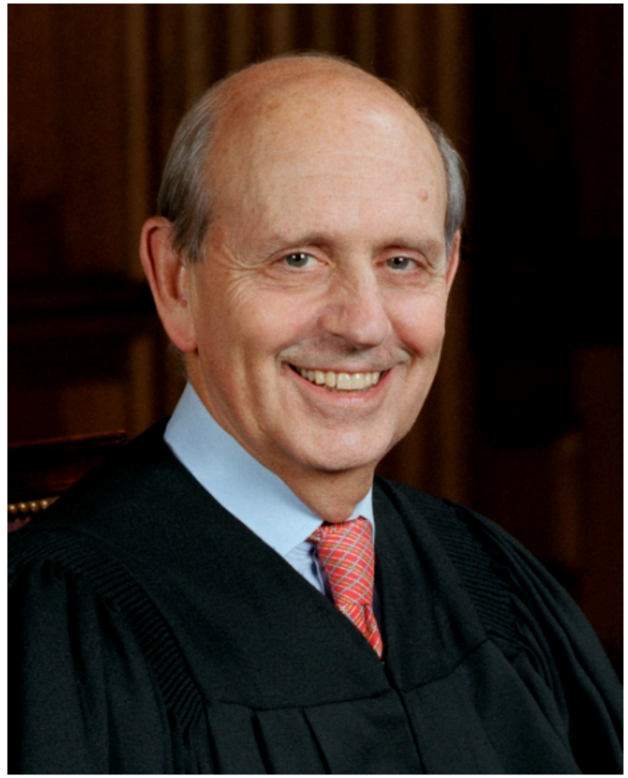
All the movies, images, emails and other digital data from more than 600 smartphones (10,000 gigabytes) can be stored in the faint pink smear of DNA at the end of this test tube.


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



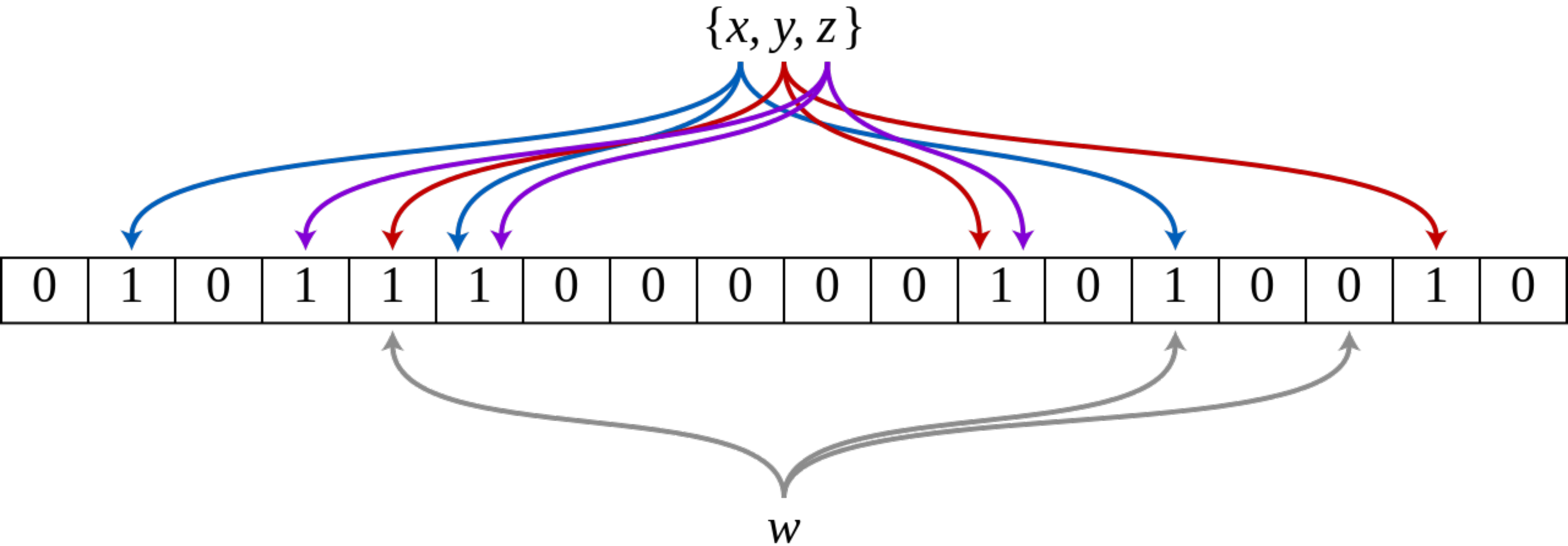
Simulate

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?

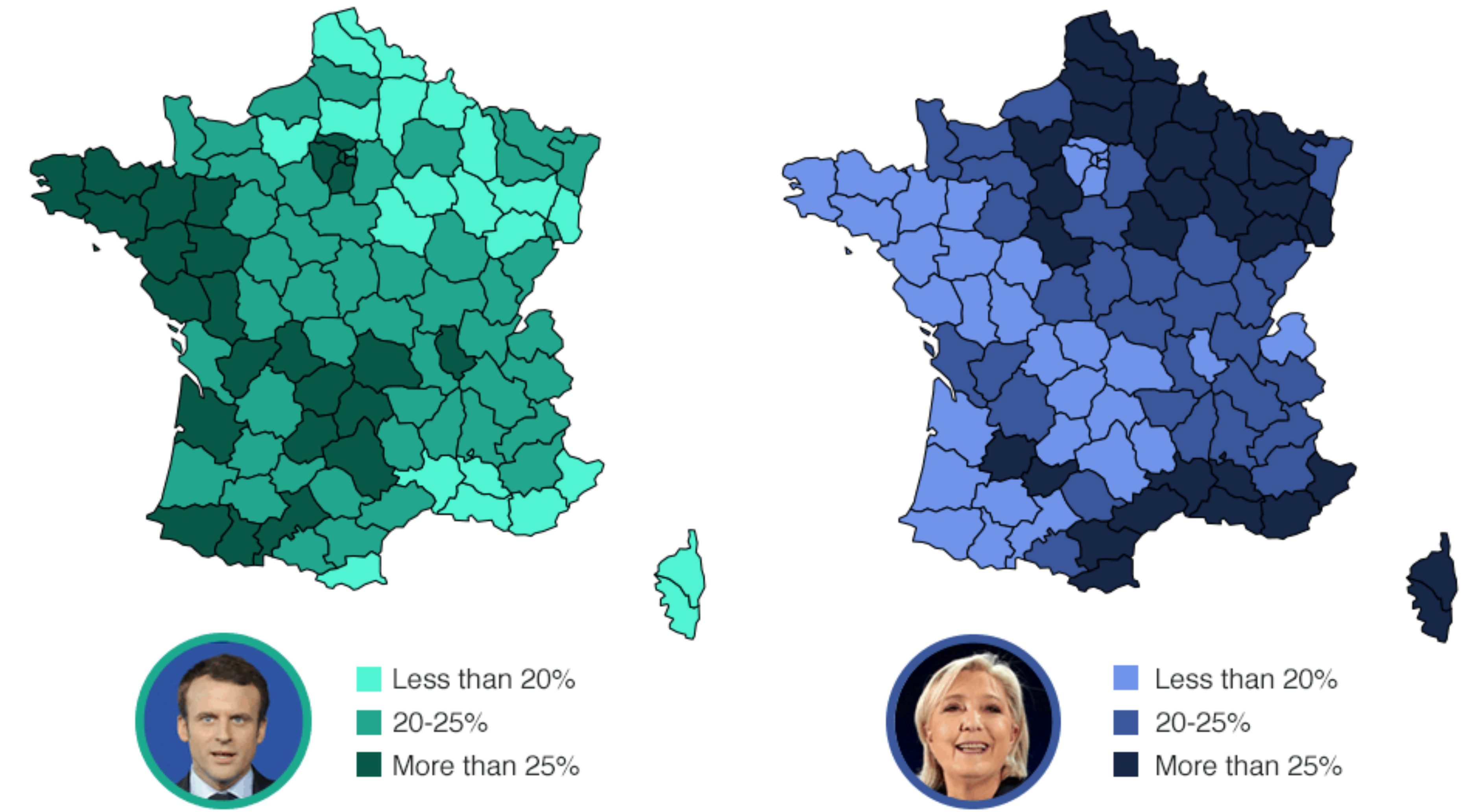
Bloom Filter



Election Prediction

Macron polled strongly in the west, while Le Pen fared better in the north-east and south coast

The two remaining presidential candidates' vote share mapped



Source: French government

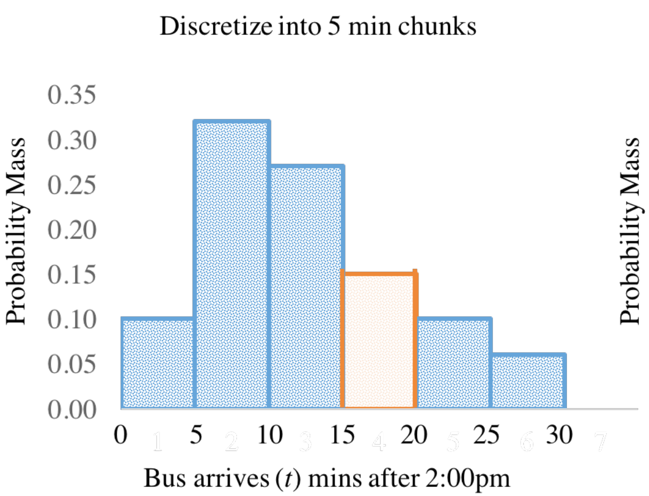
BBC

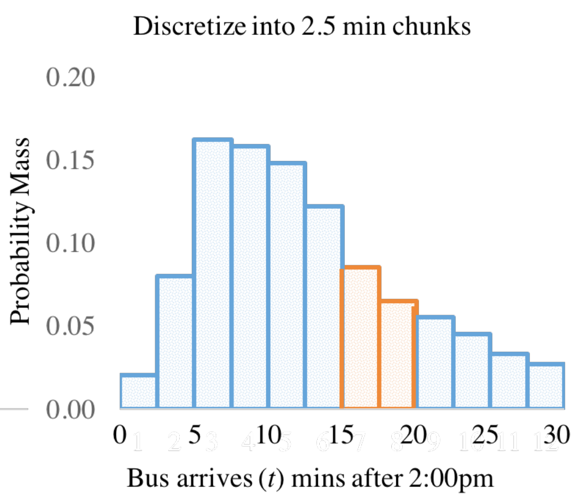
`random()?`

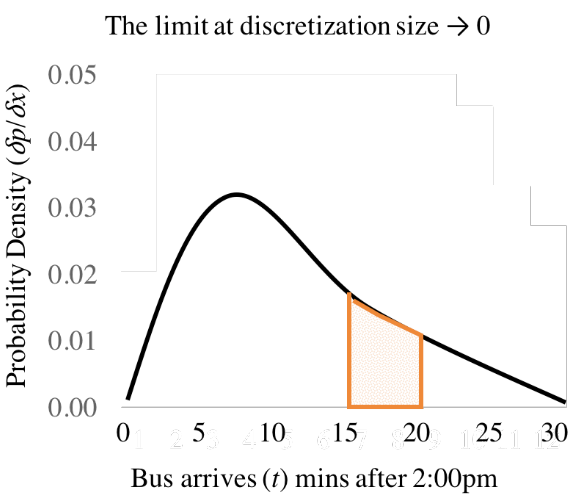
Riding the Marguerite

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

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

A probability!

Simple Example

Consider a random 5000x5000 matrix, where each element in the matrix is Uniform(0,1). What is the probability that a selected eigenvalue (λ) of the matrix is greater than 0?*

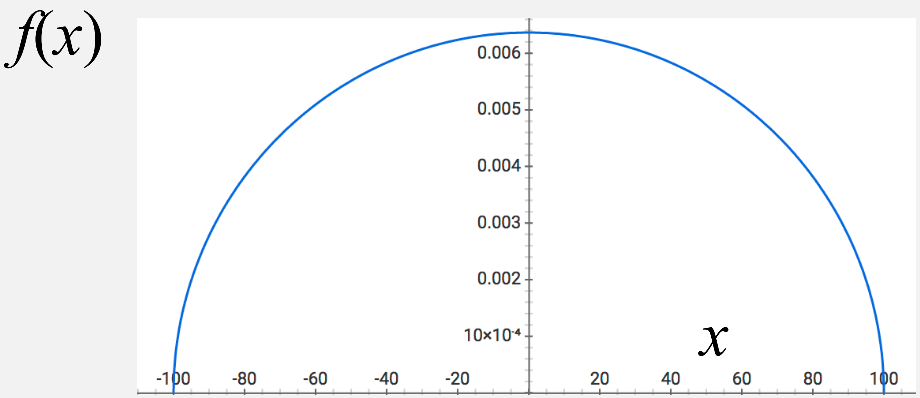
* With help from Wigner, Chris is going to rephrase this problem

Simple Example from Quantum Physics

Let X be a continuous random variable:

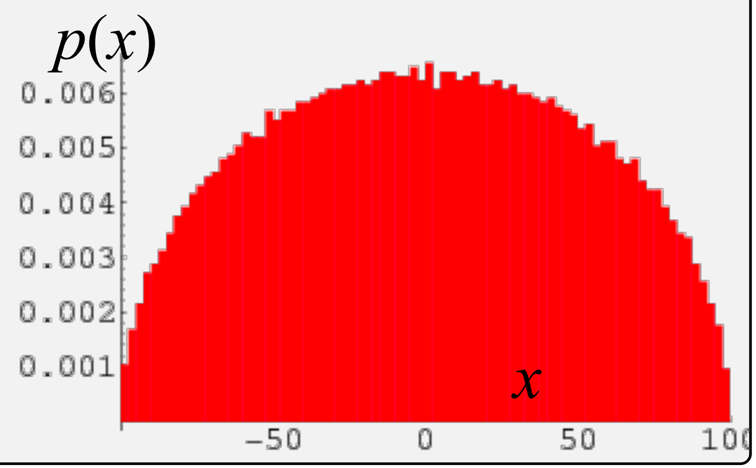
Theory

$$f(x) = \frac{1}{15708}\sqrt{100^2 - x^2}$$



Practice

From simulations



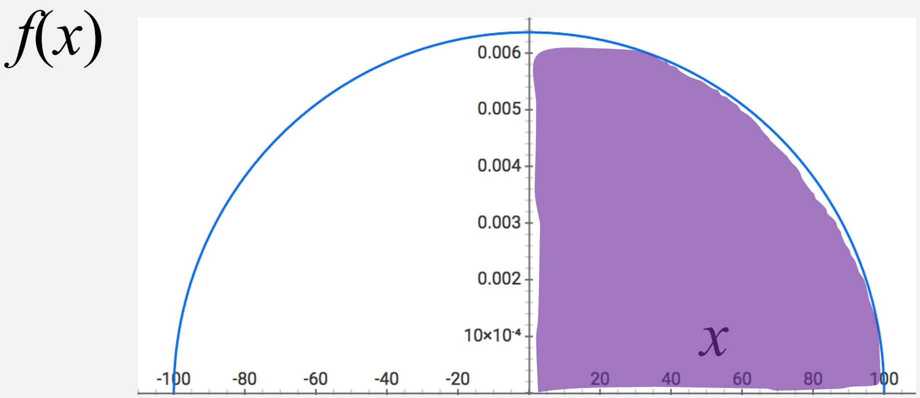
$$P(X > 0) =?$$

Simple Example from Quantum Physics

Let X be a continuous random variable:

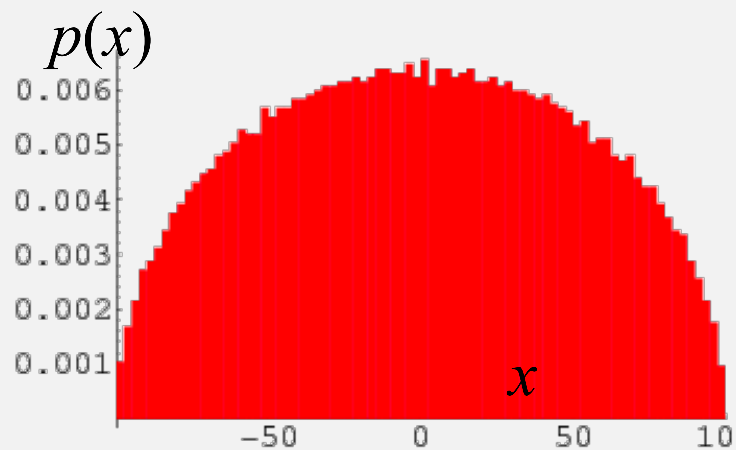
Theory

$$f(x) = \frac{1}{15708}\sqrt{100^2 - x^2}$$



Practice

From simulations



Approach #1: Integrate over the PDF

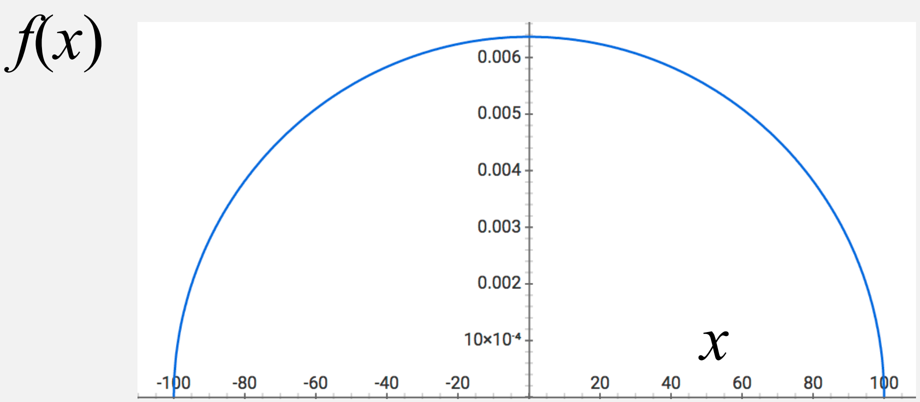
$$P(X > 0) = \int_{0}^{100} f_X(x)\ dx$$

Simple Example from Quantum Physics

Let X be a continuous random variable:

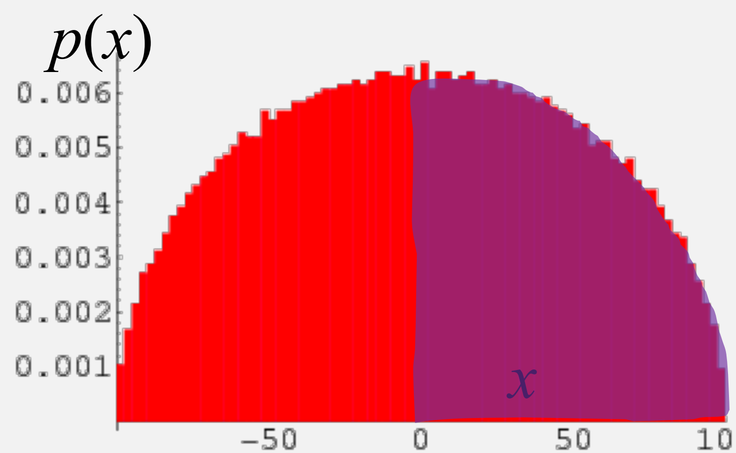
Theory

$$f(x) = \frac{1}{15708}\sqrt{100^2 - x^2}$$



Practice

From simulations



Approach #2: Discrete Approximation

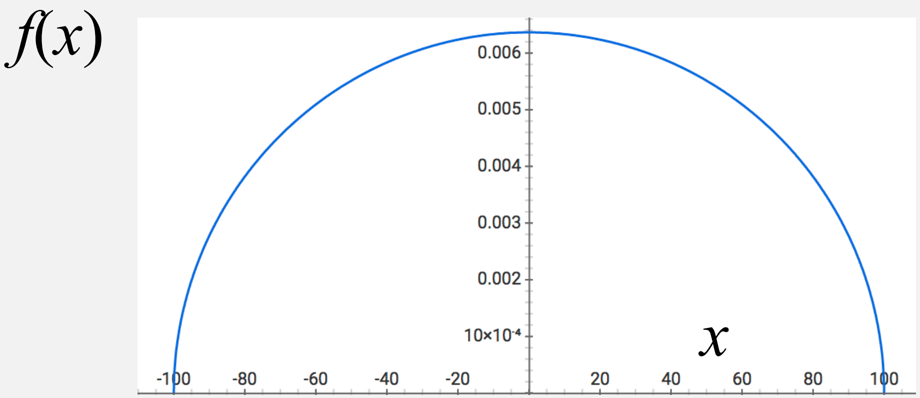
$$P(X > 0) \approx \sum_{i=0}^{100} P(X = i)$$

Simple Example from Quantum Physics

Let X be a continuous random variable:

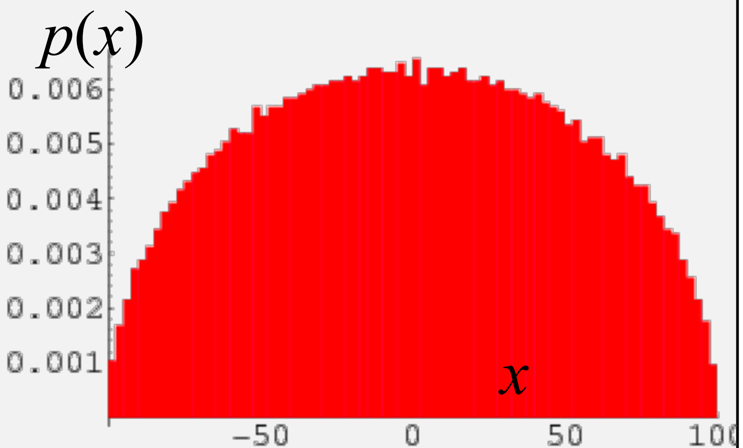
Theory

$$f(x) = \frac{1}{15708}\sqrt{100^2 - x^2}$$



Practice

From simulations



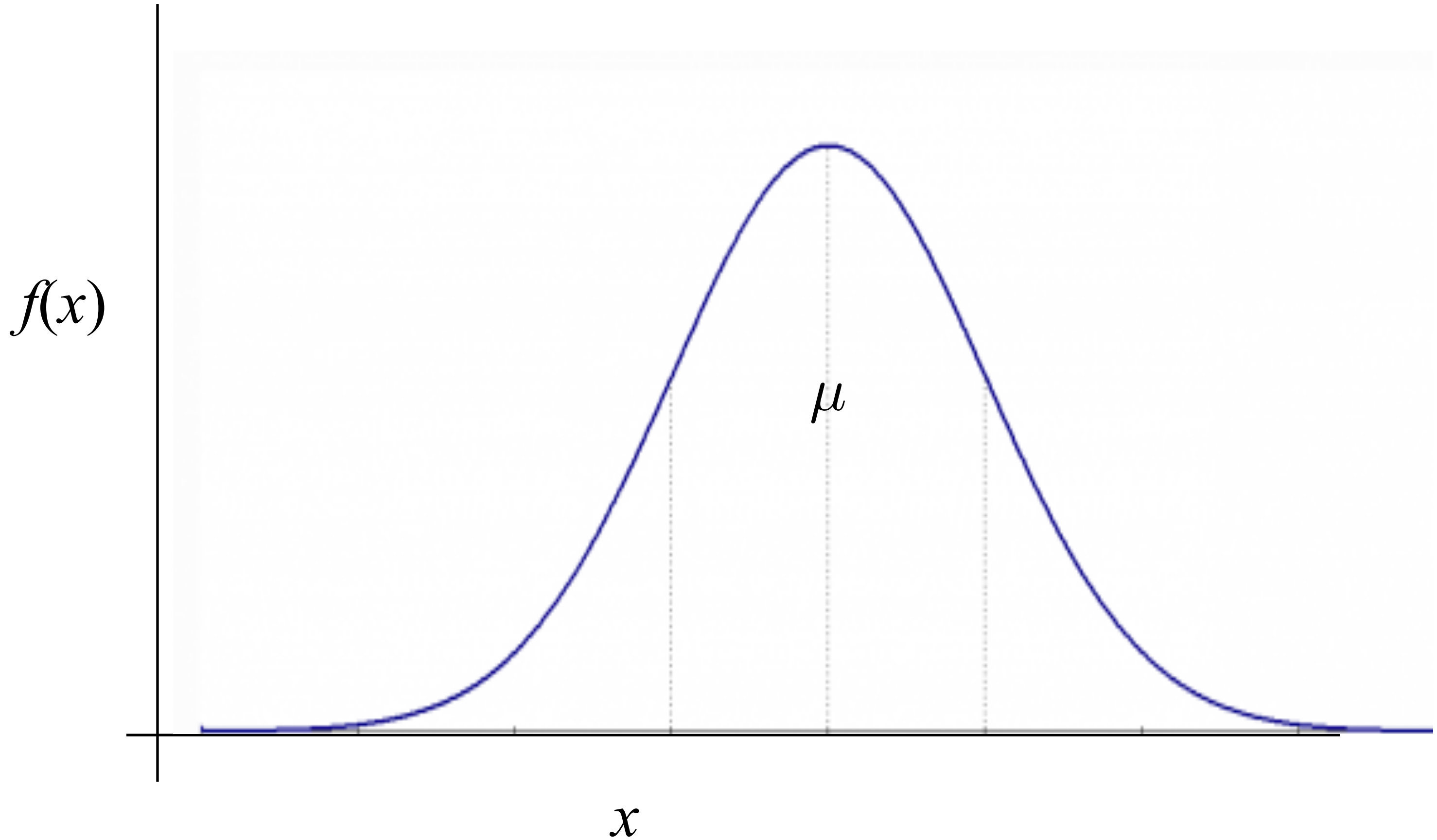
Approach #3: Know Semi-Circles

$$P(X > 0) = \frac{1}{2}$$

Probability Density Function

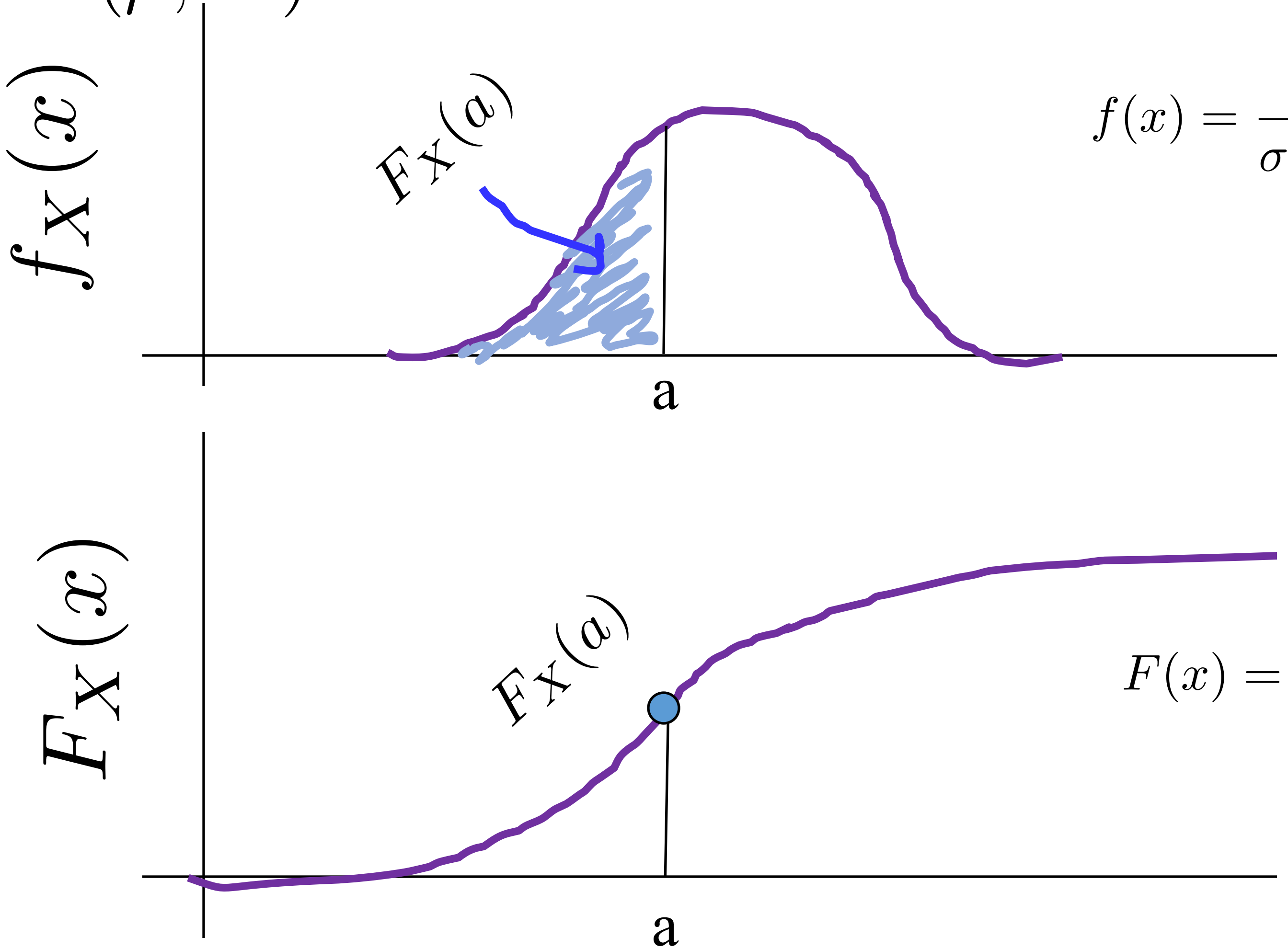
$$\mathcal{N}(\mu, \sigma^2)$$

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



PDF and CDF of a Normal

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



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

$$F(x) = \Phi\left(\frac{x-\mu}{\sigma}\right)$$

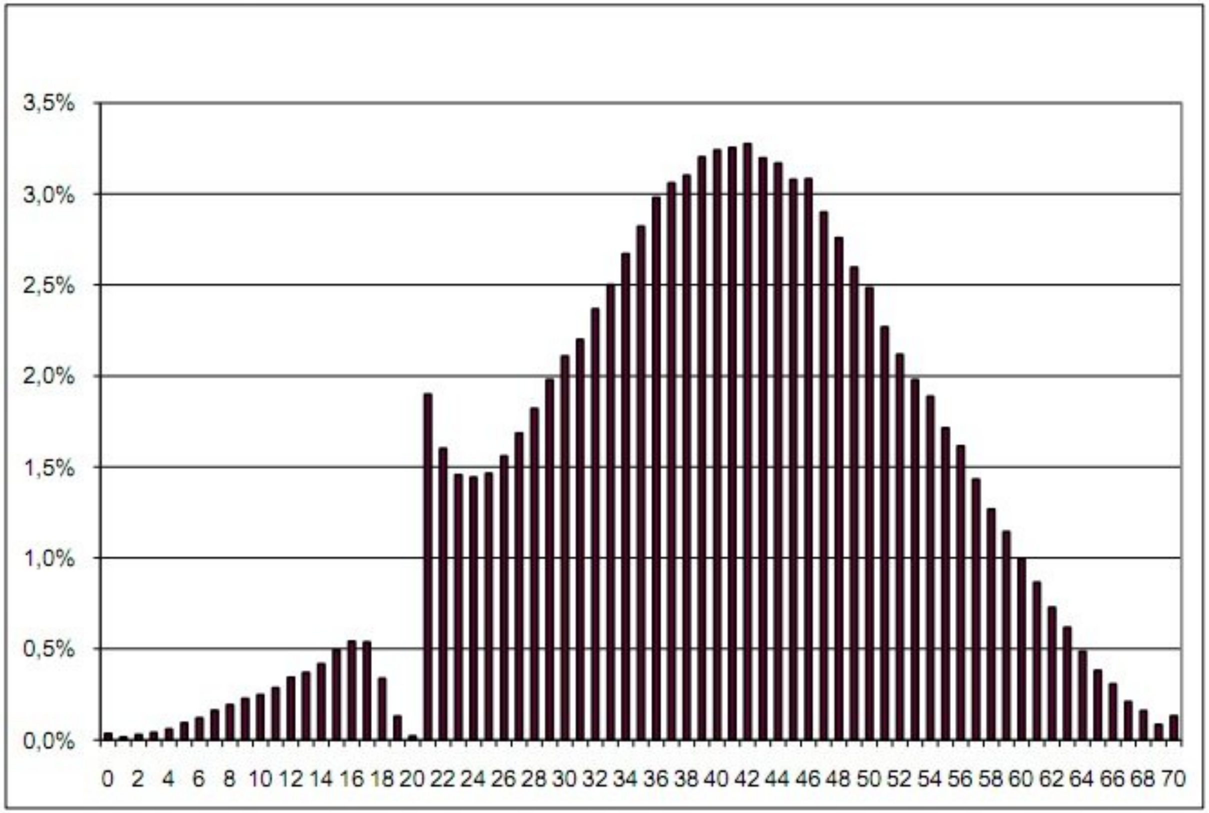
A CDF is the integral from –infinity to x of the PDF

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.

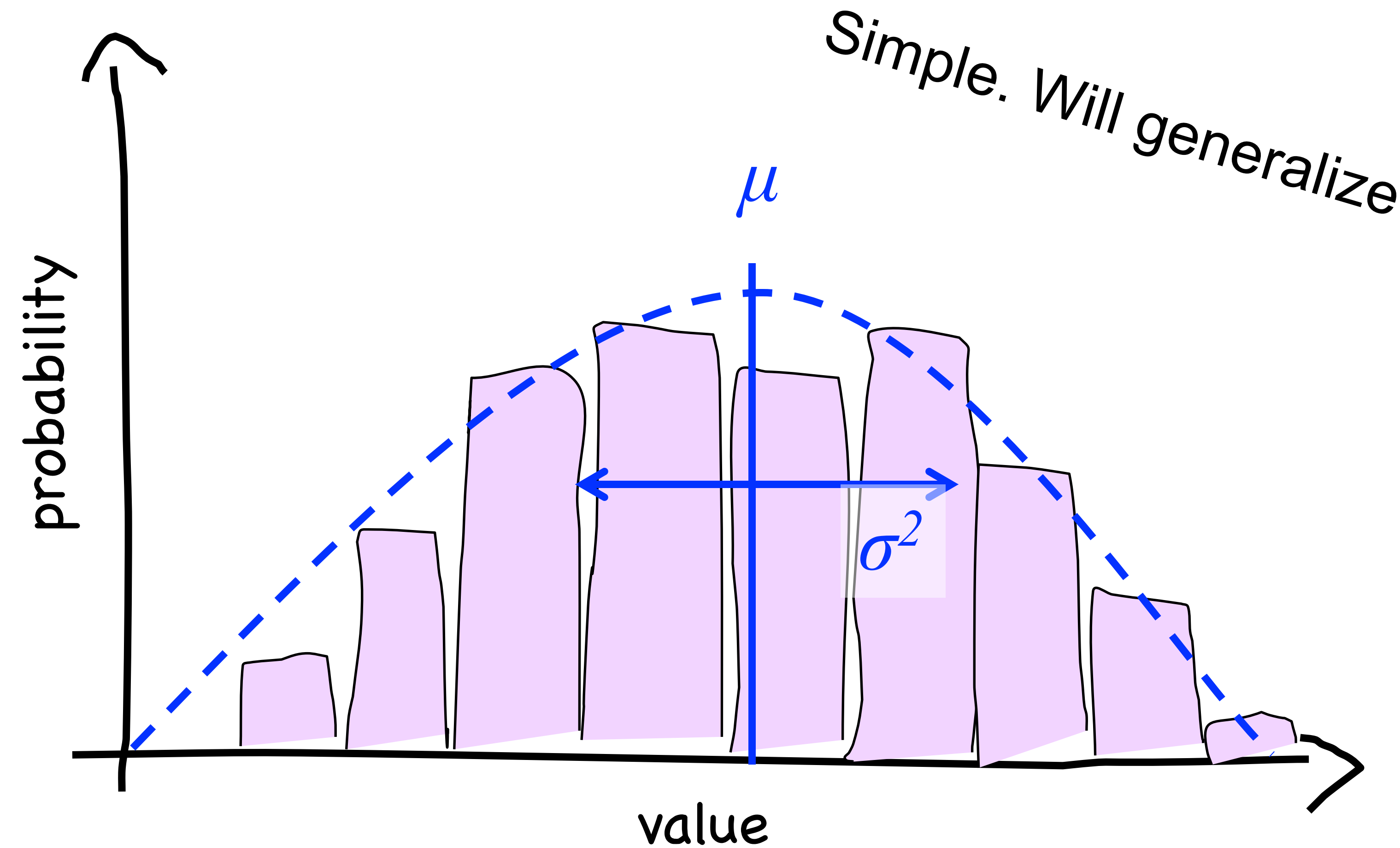
2.1. Poziom podstawowy



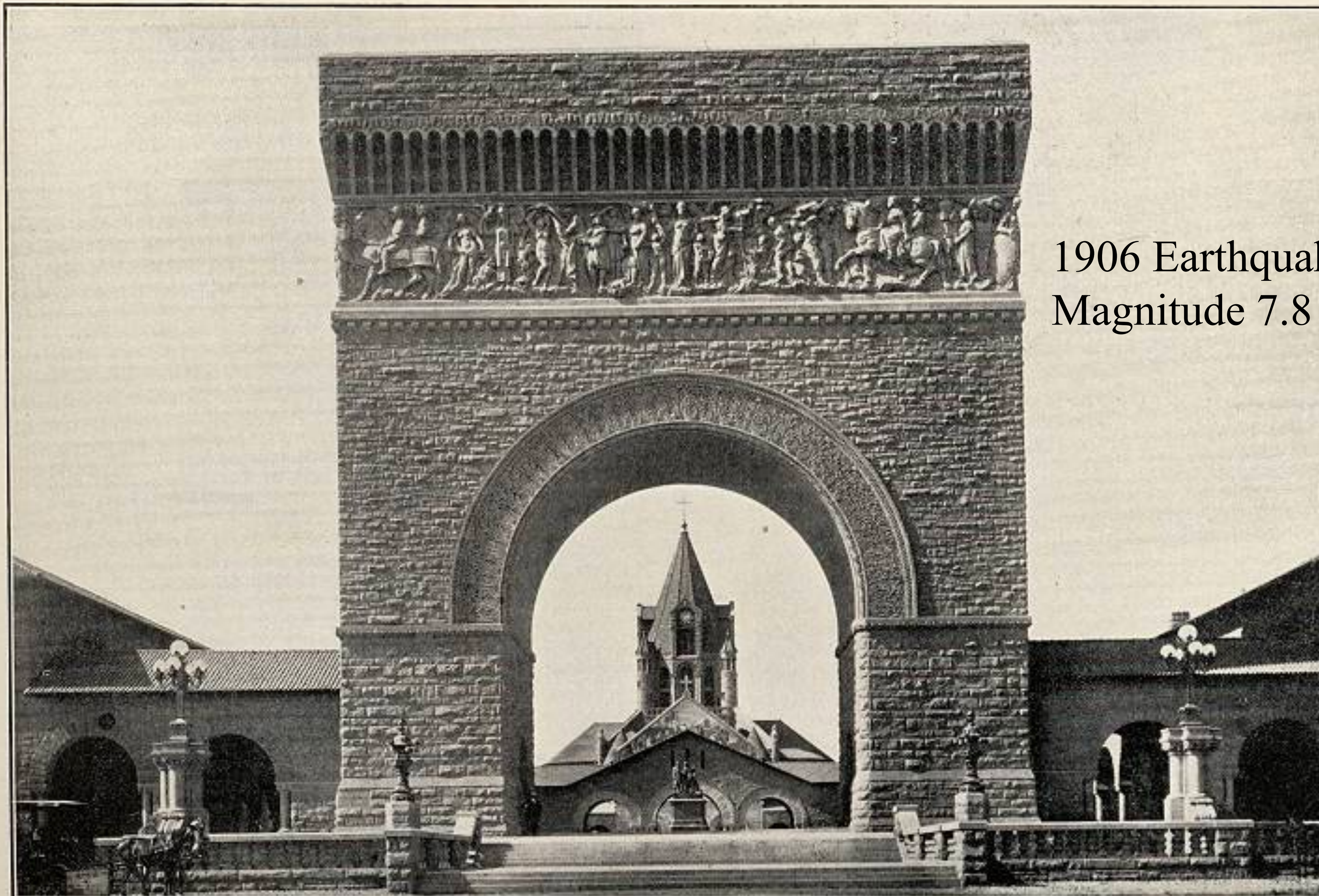
Wykres 1. Rozkład wyników na poziomie podstawowym

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

Simplicity is Humble



* A Gaussian maximizes entropy for a given mean and variance



1906 Earthqual
Magnitude 7.8

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

Will the Warriors Win?

What is the probability that the Warriors beat the Blazers?

How do you model zero sum games?

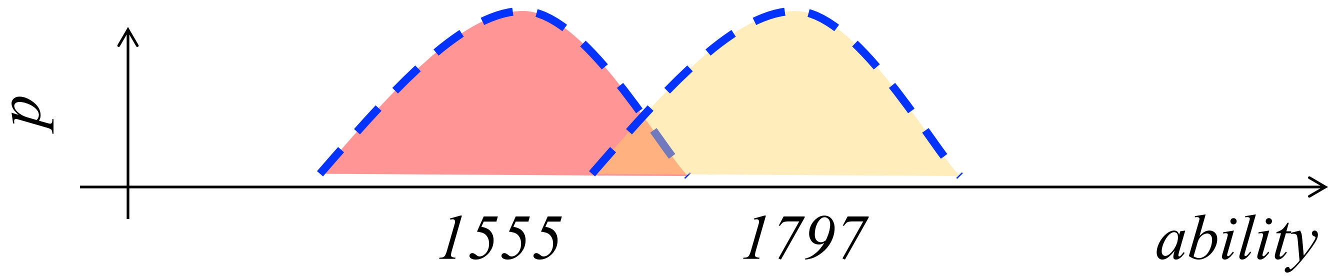
ELO Ratings

How it works:

- Each team has an "ELO" score *S*, calculated based on their past performance.
- Each game, the team has ability $A \sim N(S, 200^2)$
- The team with the higher sampled ability wins.

Arpad Elo

$A_B \sim \mathcal{N}(1555, 200^2)$ $A_W \sim \mathcal{N}(1797, 200^2)$

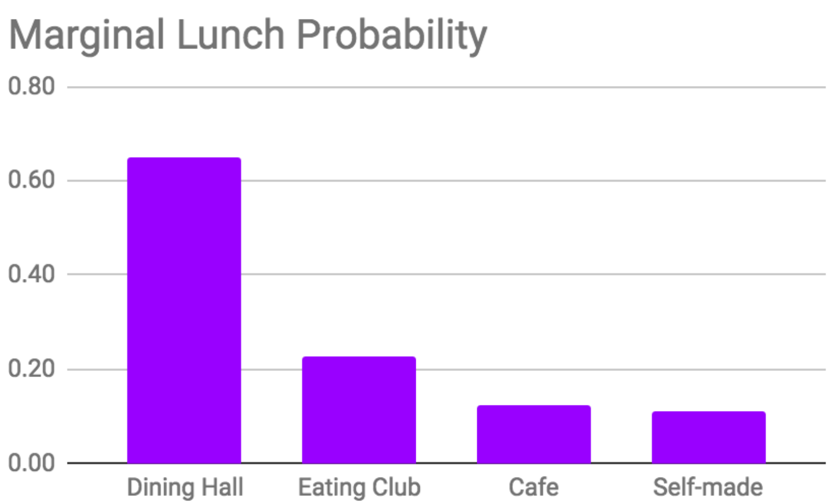


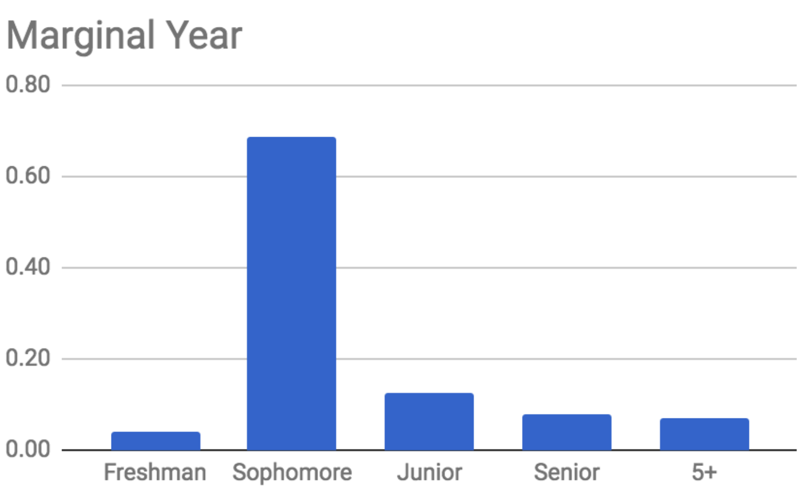
$$P(\text{Warriors win}) = P(A_W > A_B)$$

Joint Distributions

Joint Probability Table

Joint Probability Table					
	Dining Hall	Eating Club	Cafe	Self-made	**Marginal Year**
Freshman	0.02	0.00	0.02	0.00	0.04
Sophomore	0.51	0.15	0.03	0.03	0.69
Junior	0.08	0.02	0.02	0.02	0.13
Senior	0.02	0.05	0.01	0.01	0.08
5+	0.02	0.01	0.05	0.05	0.07
Marginal Status	0.65	0.23	0.13	0.11	



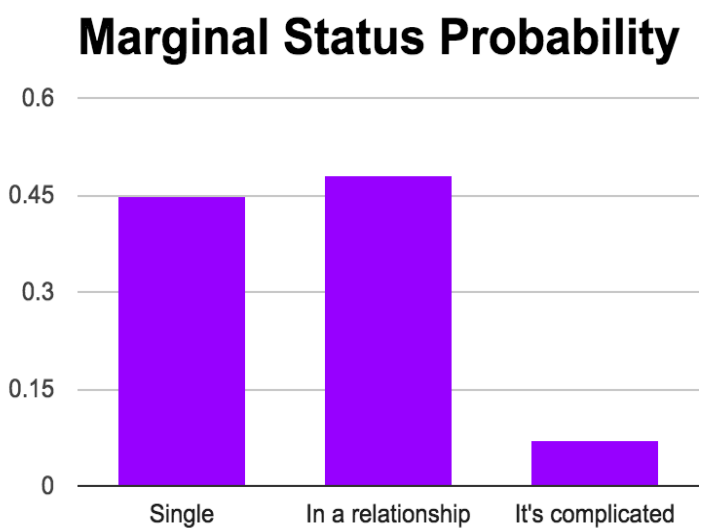


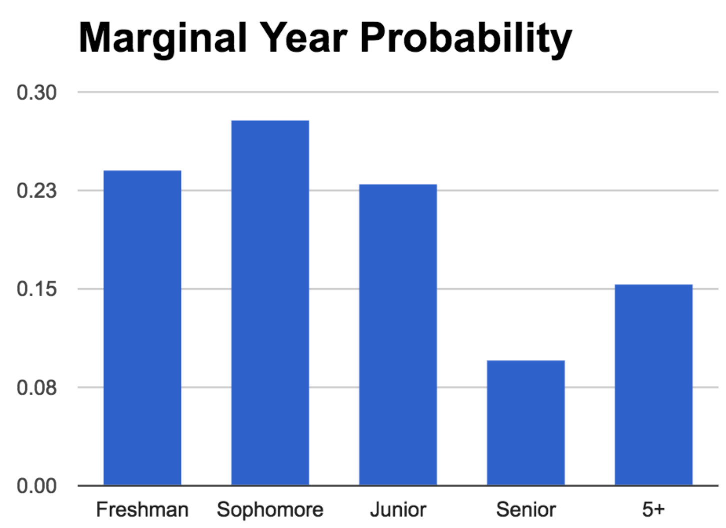
It’s Complicated Demo

Go to this URL: **https://goo.gl/jCMY18**

Probability Table

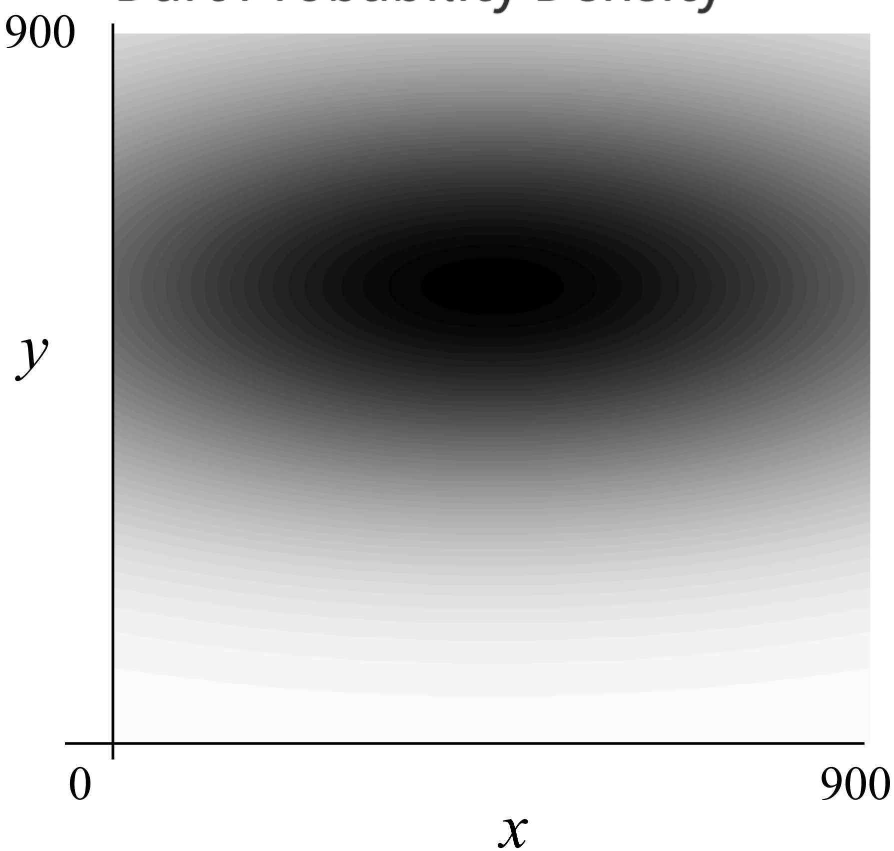
Joint Probability Table				
	Single	In a relationship	It's complicated	**Marginal Year**
Freshman	0.13	0.09	0.02	0.24
Sophomore	0.16	0.10	0.02	0.28
Junior	0.12	0.10	0.02	0.23
Senior	0.01	0.09	0.00	0.10
5+	0.03	0.12	0.01	0.15
Marginal Status	0.45	0.48	0.07	





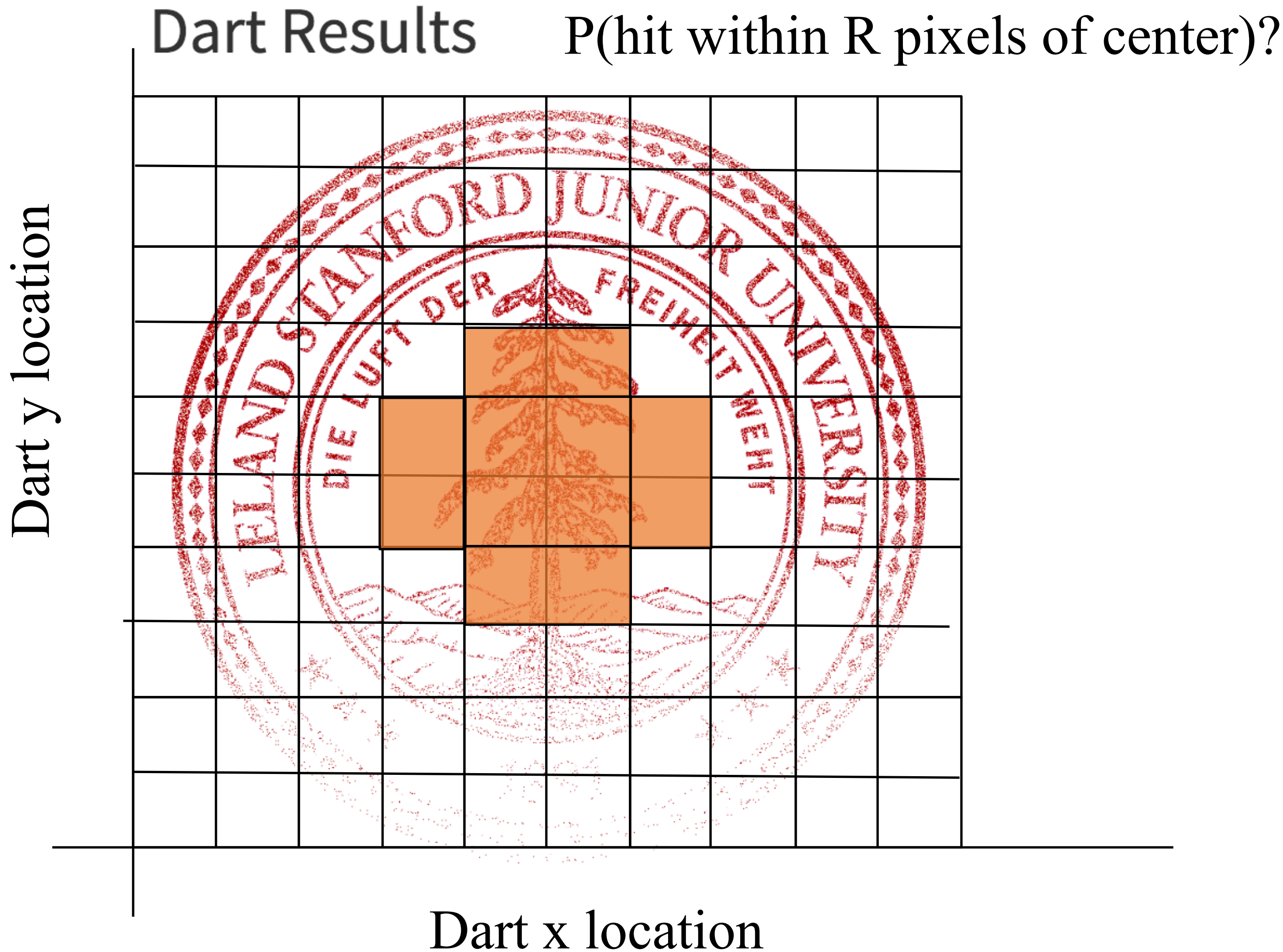
Joint Dart Distribution

Dart Probability Density

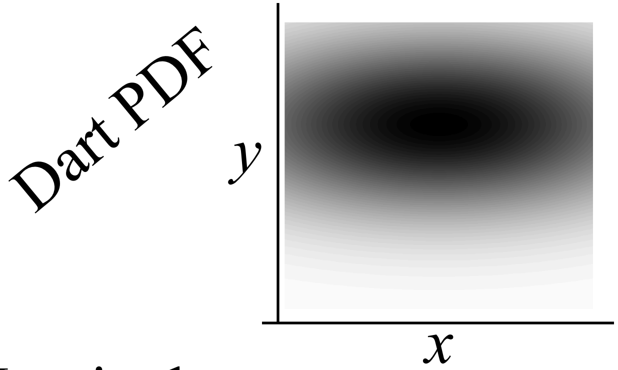


Dart Results

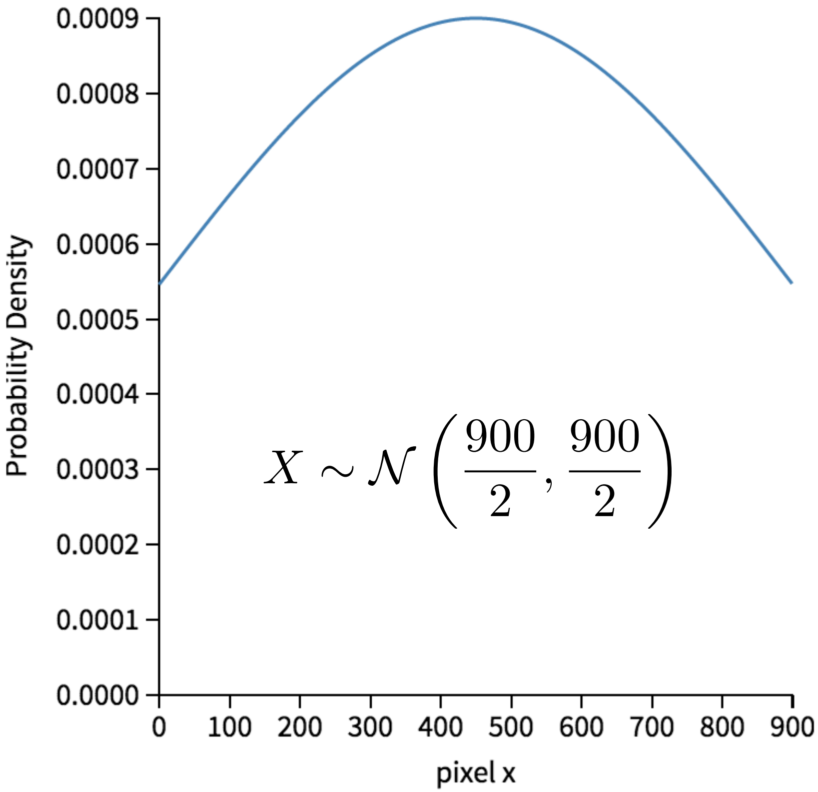
Joint Dart Distribution



Darts!

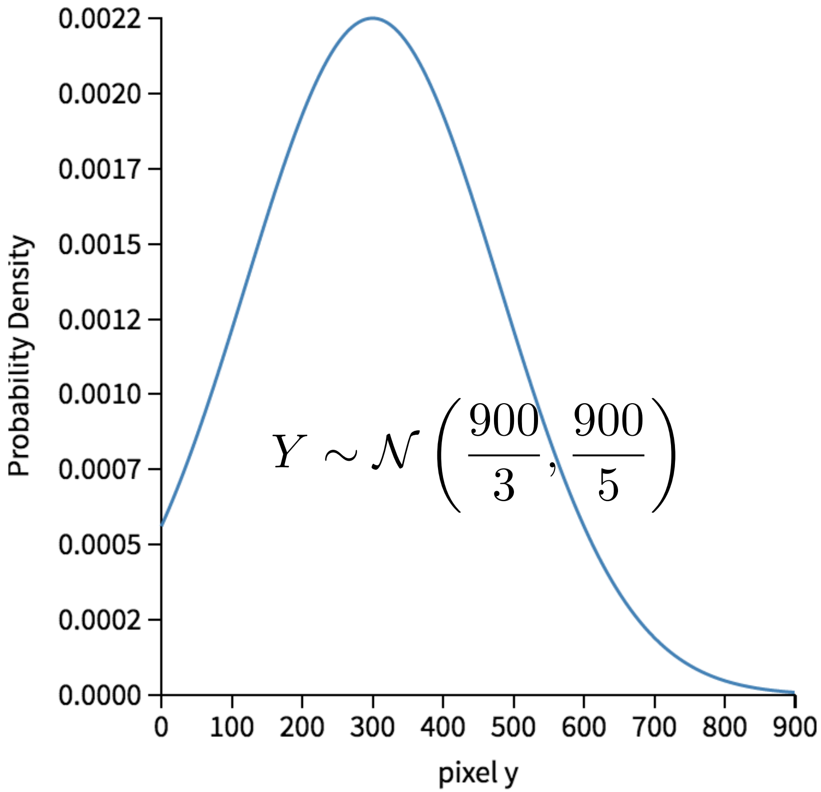


X-Pixel Marginal



$$X \sim \mathcal{N}\left(\frac{900}{2}, \frac{900}{2}\right)$$

Y-Pixel Marginal



$$Y \sim \mathcal{N}\left(\frac{900}{3}, \frac{900}{5}\right)$$

Gaussian Blur

In image processing, a Gaussian blur is the result of blurring an image by a Gaussian function. It is a widely used effect in graphics software, typically to reduce image noise.

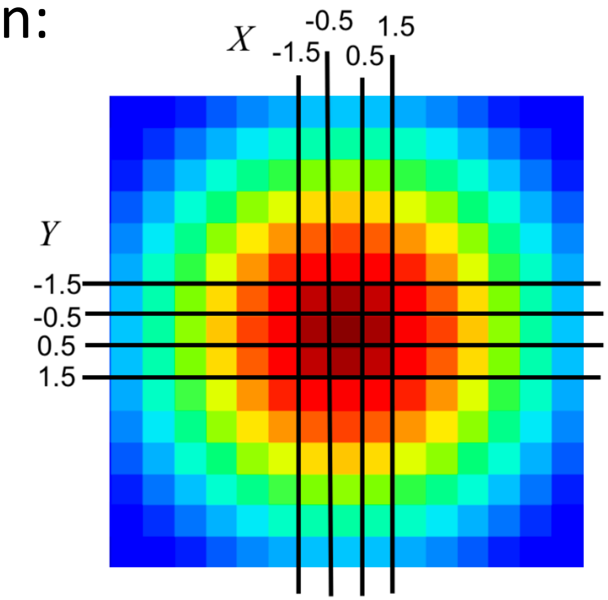
Gaussian, typically with StDev = 3, is based on a joint probability distribution:

Joint PDF

$$f_{X,Y}(x,y) = \frac{1}{2\pi \cdot 3^2} e^{-\frac{x^2+y^2}{2\cdot 3^2}}$$

Joint CDF

$$F_{X,Y}(x,y) = \Phi\left(\frac{x}{3}\right) \cdot \Phi\left(\frac{y}{3}\right)$$



Used to generate this weight matrix

Multinomial

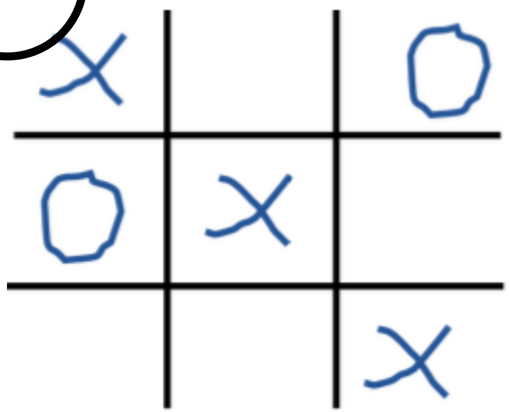
Example document:
"Pay for Viagra with a credit-card. Viagra is great. So are credit-cards. Risk free Viagra. Click for free."
n = 18

$$\underbrace{P\left(\begin{array}{l}\text{Viagra} = 2\\ \text{Free} = 2\\ \text{Risk} = 1\\ \text{Credit-card: } 2\\ \dots\\ \text{For} = 2\end{array} \middle| \text{spam}\right)}_{\text{Probability of seeing this document | spam}} = \overbrace{\frac{n!}{2!2!\dots2!} p_{\text{viagra}}^2 p_{\text{free}}^2 \dots p_{\text{for}}^2}^{\text{It's a Multinomial!}}$$

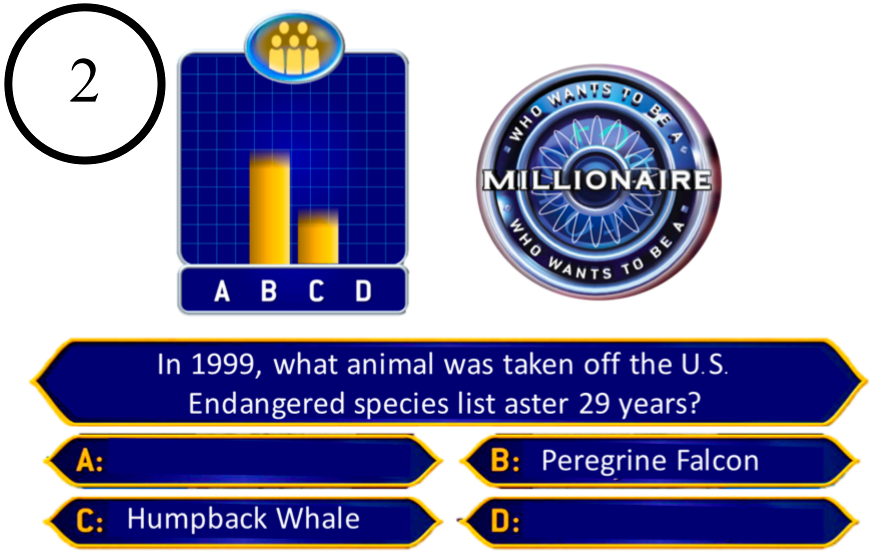
The probability of a word in spam email being viagra

Midterm (part 1)

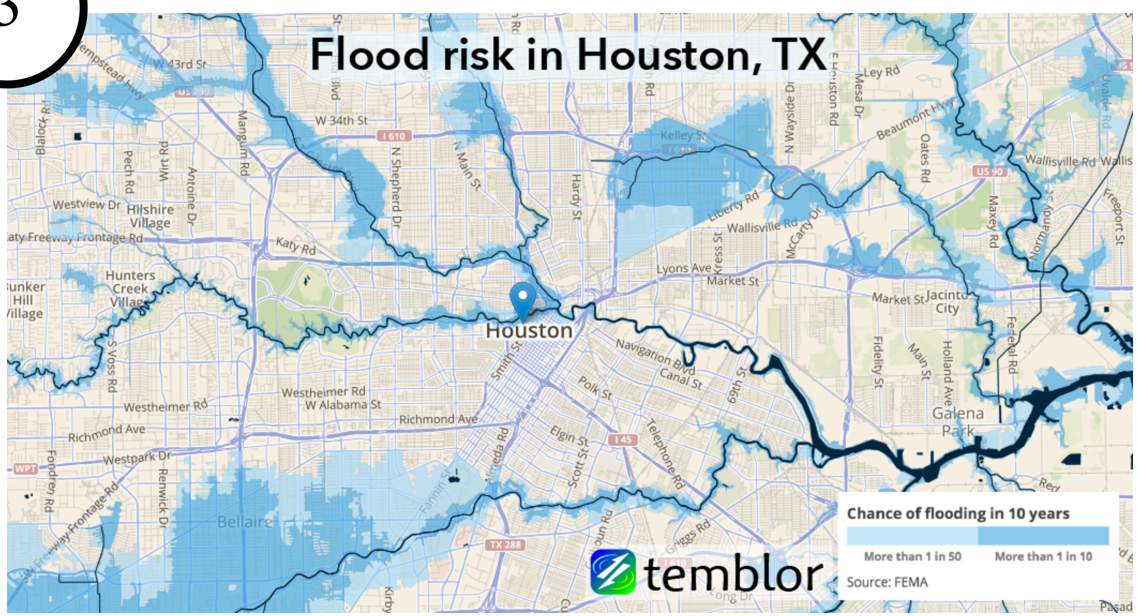
1



2

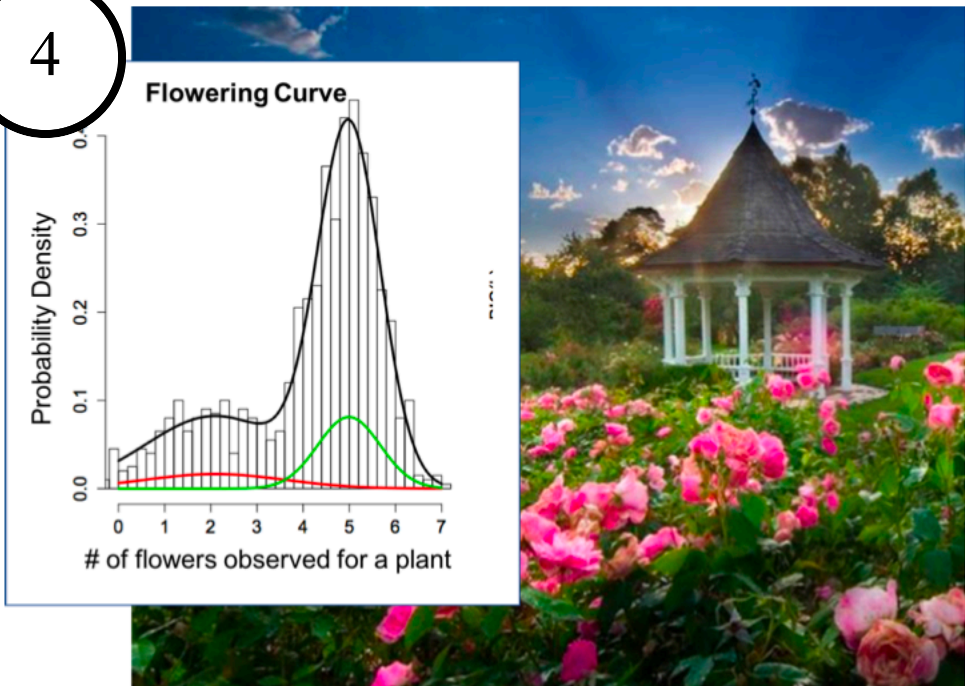


3

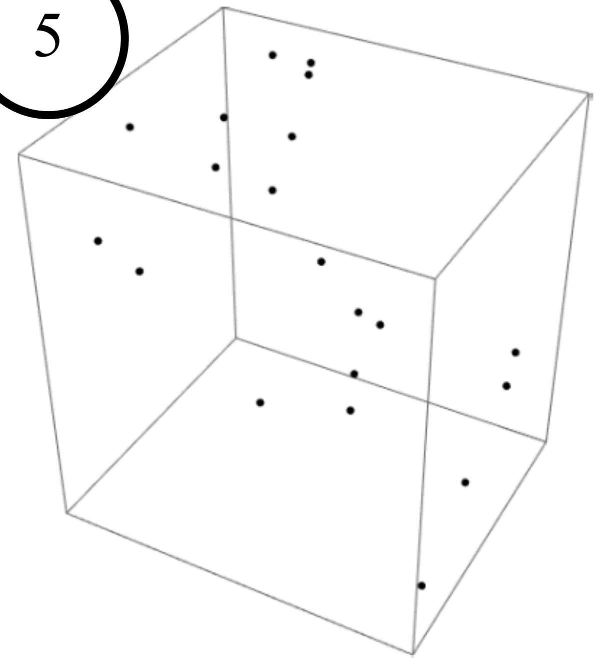


Midterm (part 2)

4



5



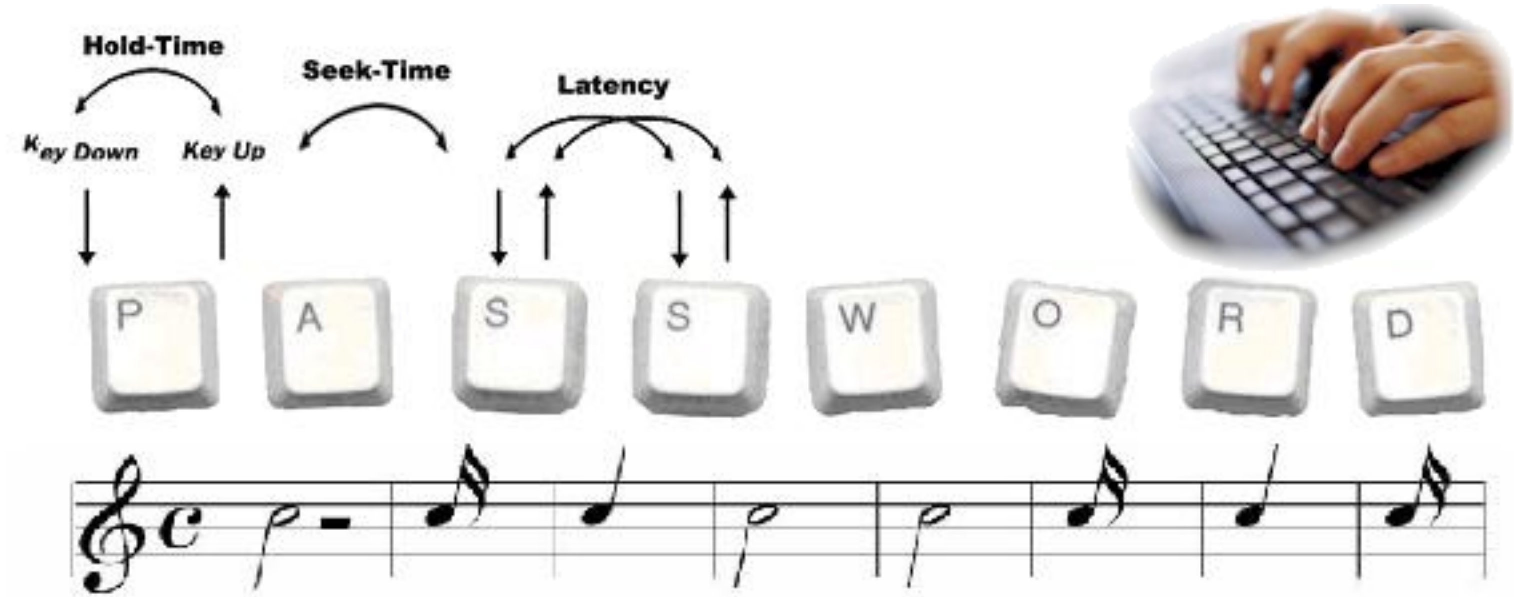
6

Generate N values $(X_1, X_2, \ldots, X_N)$ uniformly sampled over a range (a, b). We can approximate the integral of a function h over (a, b) as:

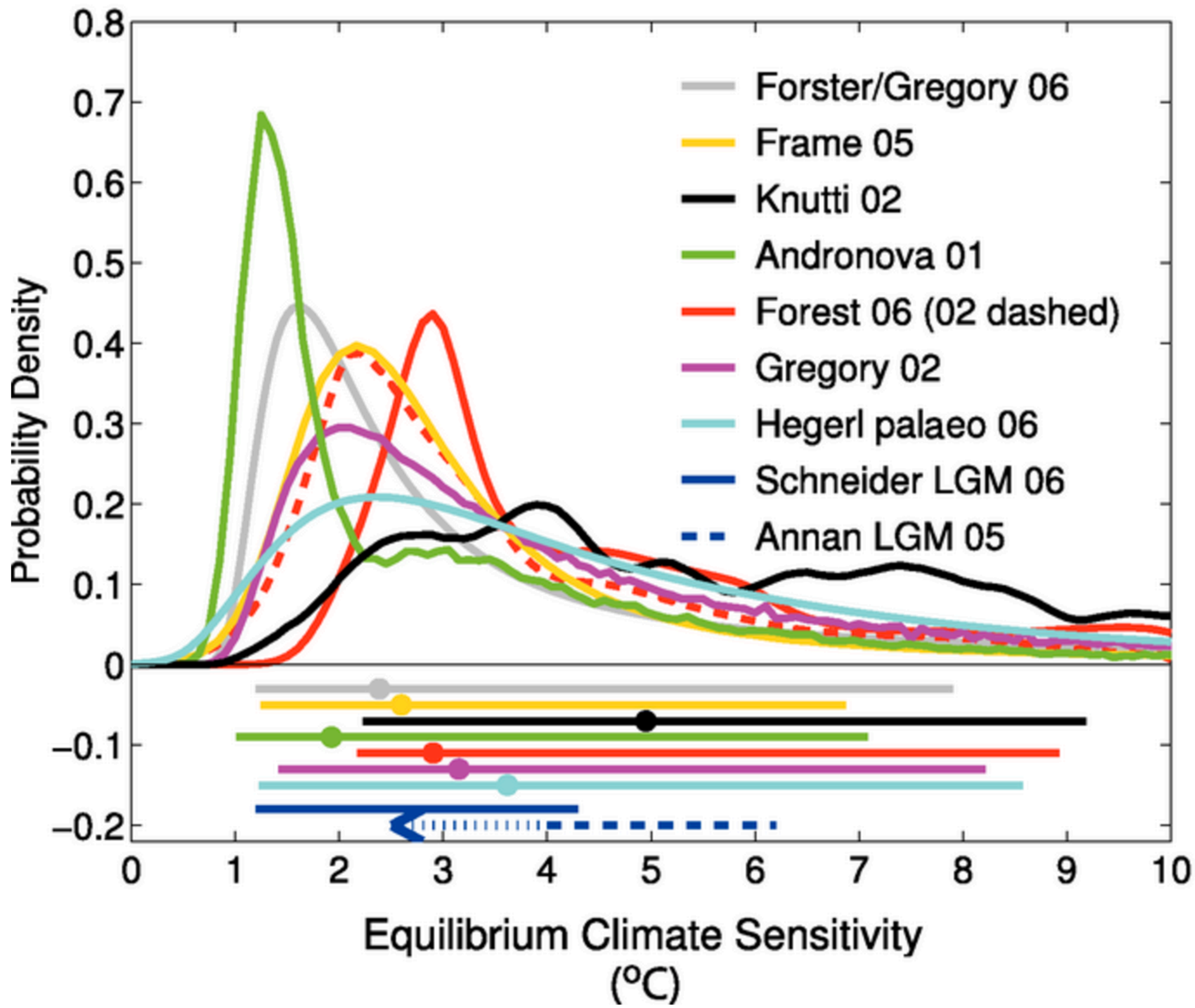
$$\int_a^b h(x)\,dx \approx \frac{(b-a)}{N} \sum_{i=1}^{N} h(X_i)$$

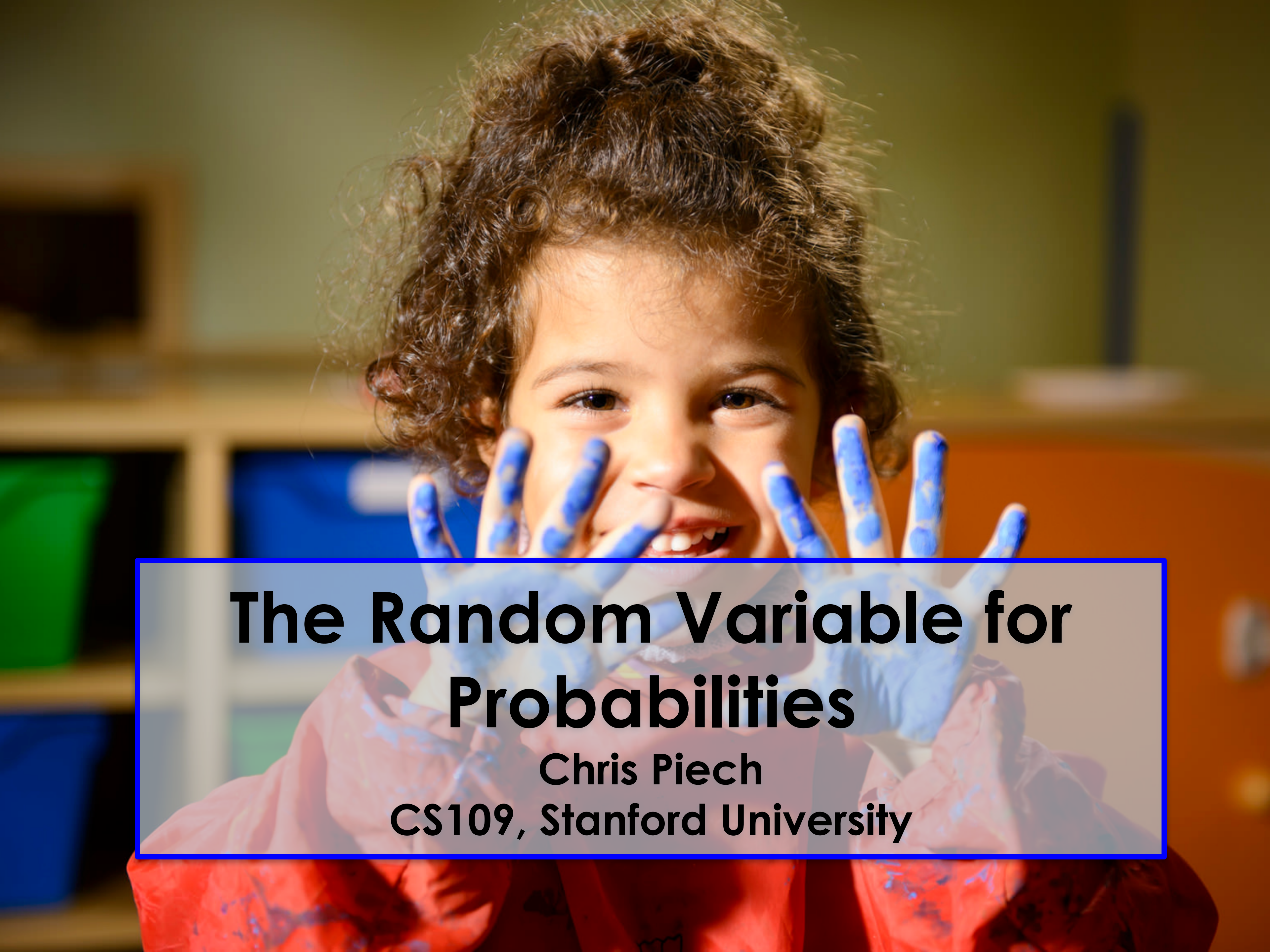
Monte Carlo Integration

Biometric Keystroke



Climate Sensitivity





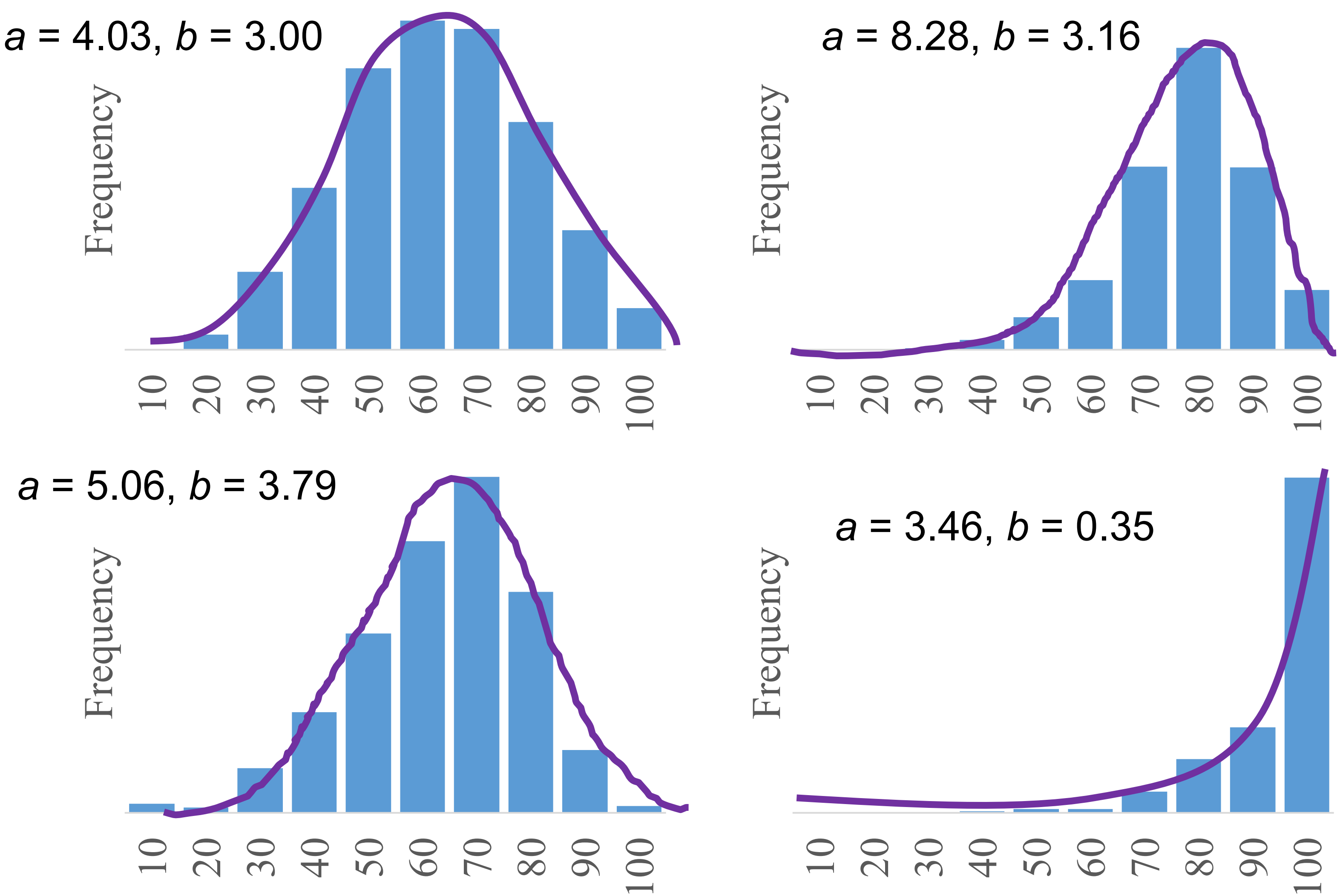
The Random Variable for Probabilities

Chris Piech
CS109, Stanford University

Flip a Coin With Unknown Probability

Demo

Assignment Grades



We have 2055 assignment distributions from grade scope

Recursive Insight

2	3	1	4	5	6	8	7

Partition array so:

- everything smaller than pivot is on left
- everything greater than or equal to pivot is on right
- pivot is in-between

Machine Learning Example

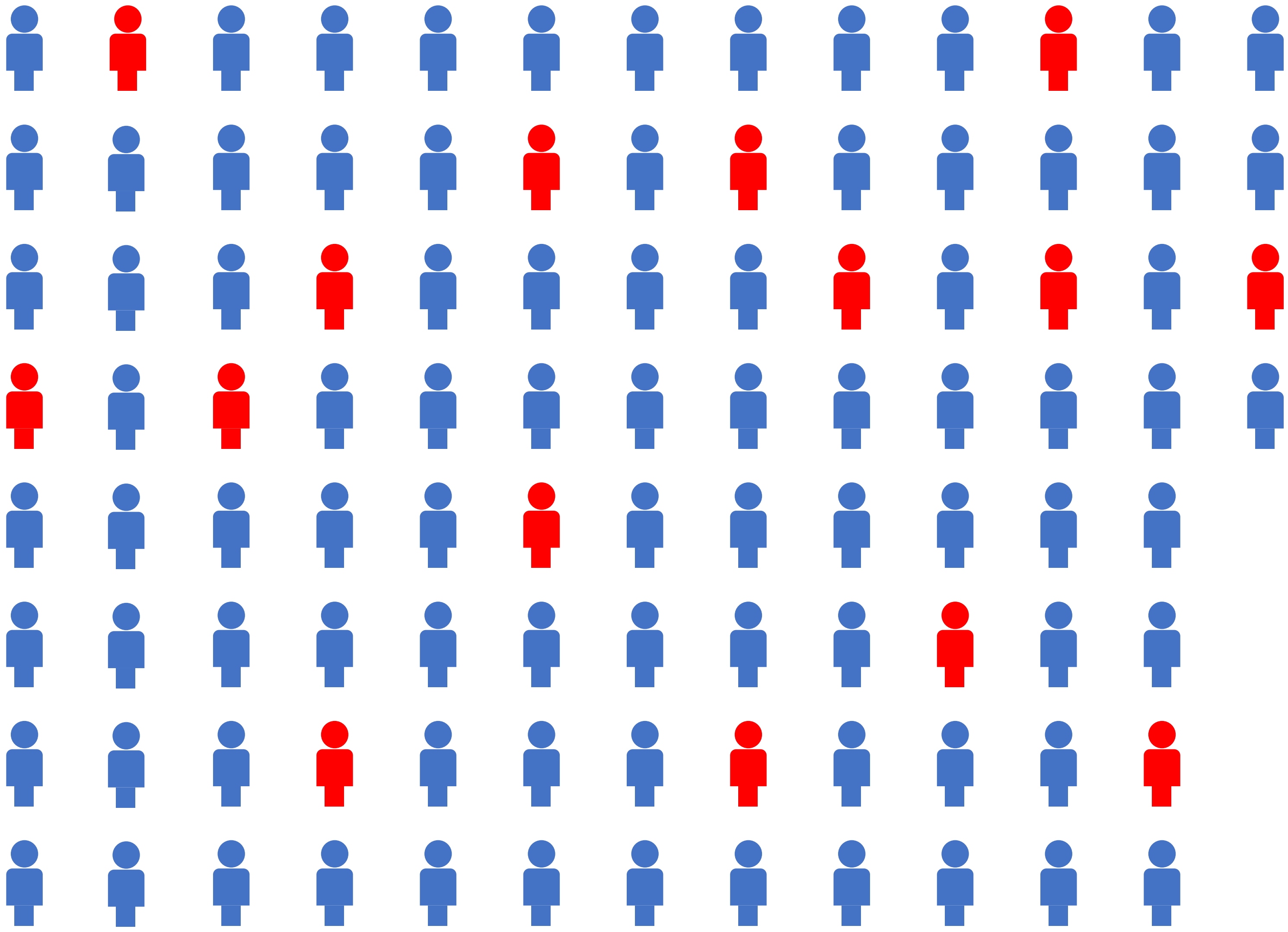
- You want to know the true mean and variance of happiness in Buthan
  - But you can't ask everyone.
  - Randomly sample 200 people.
  - Your data looks like this:

Happiness = {72, 85, 79, 91, 68, … , 71}

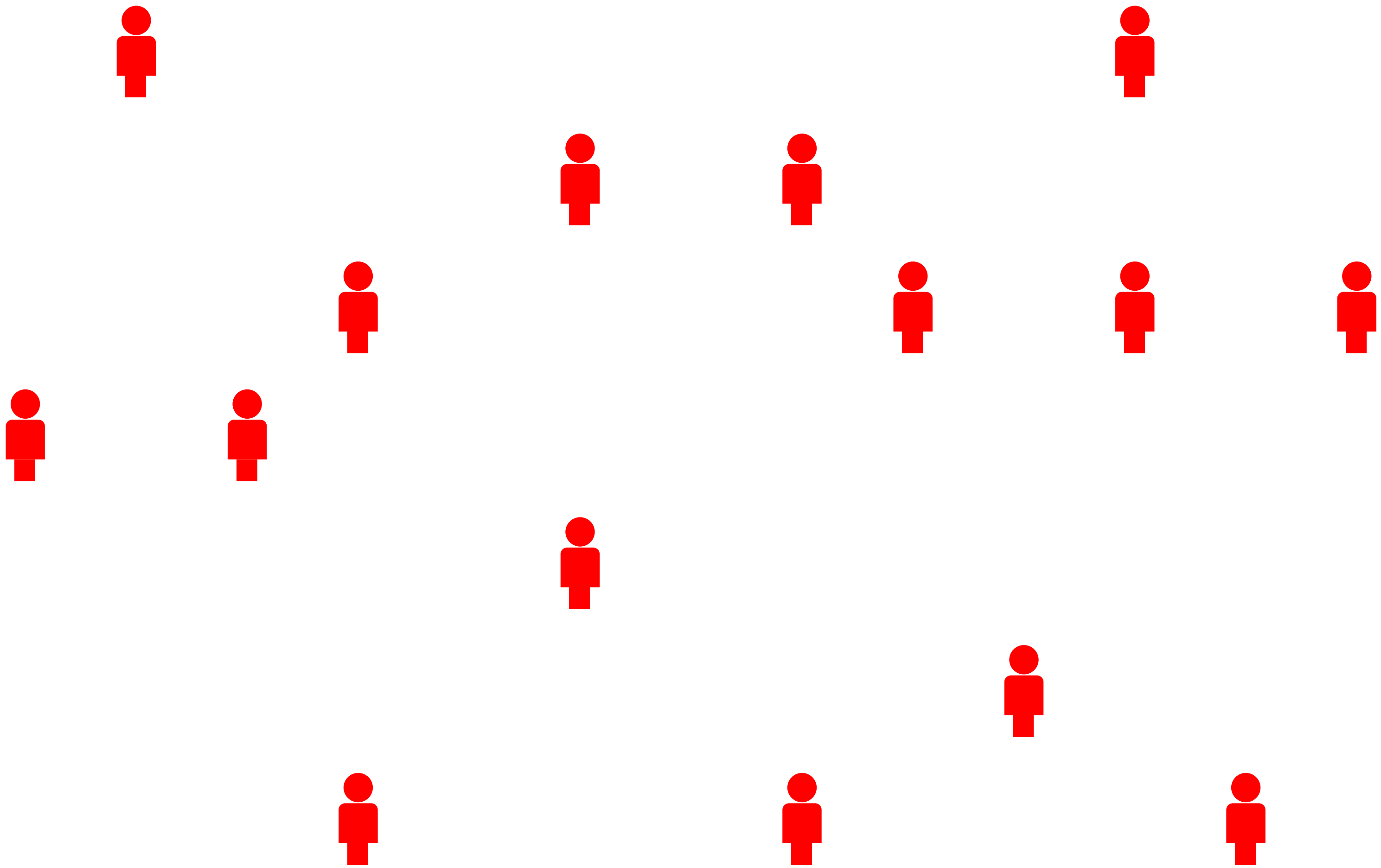
  - The mean of all of those numbers is 83. Is that the true average happiness of Bhutanese people?

Population

Sample

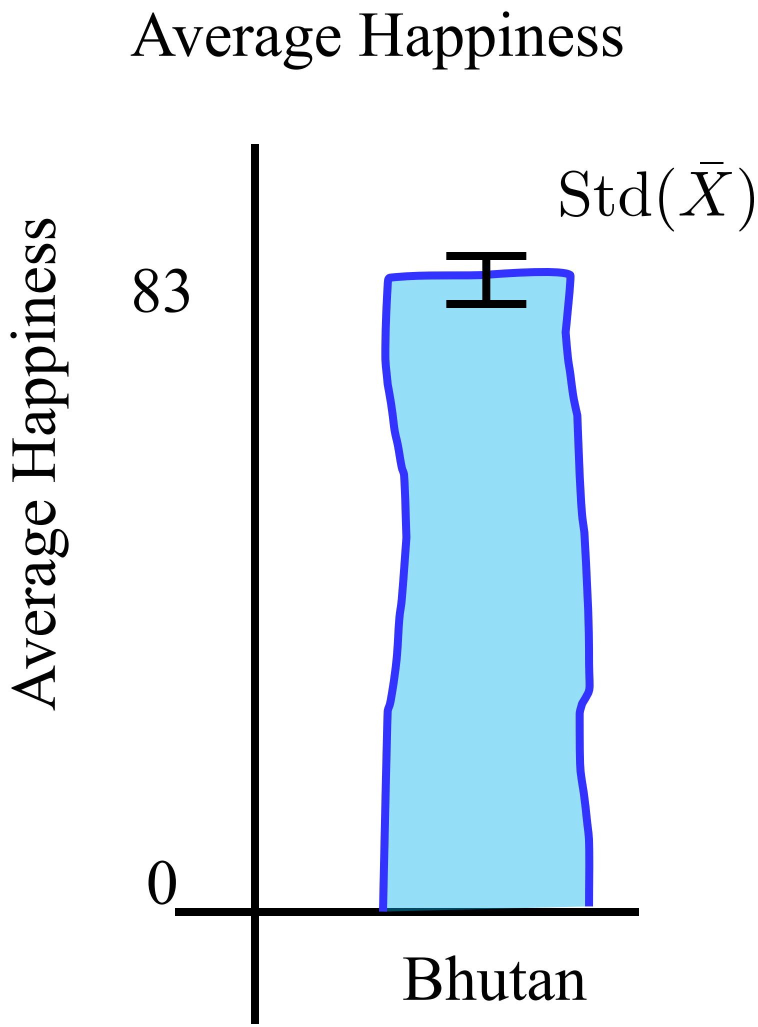


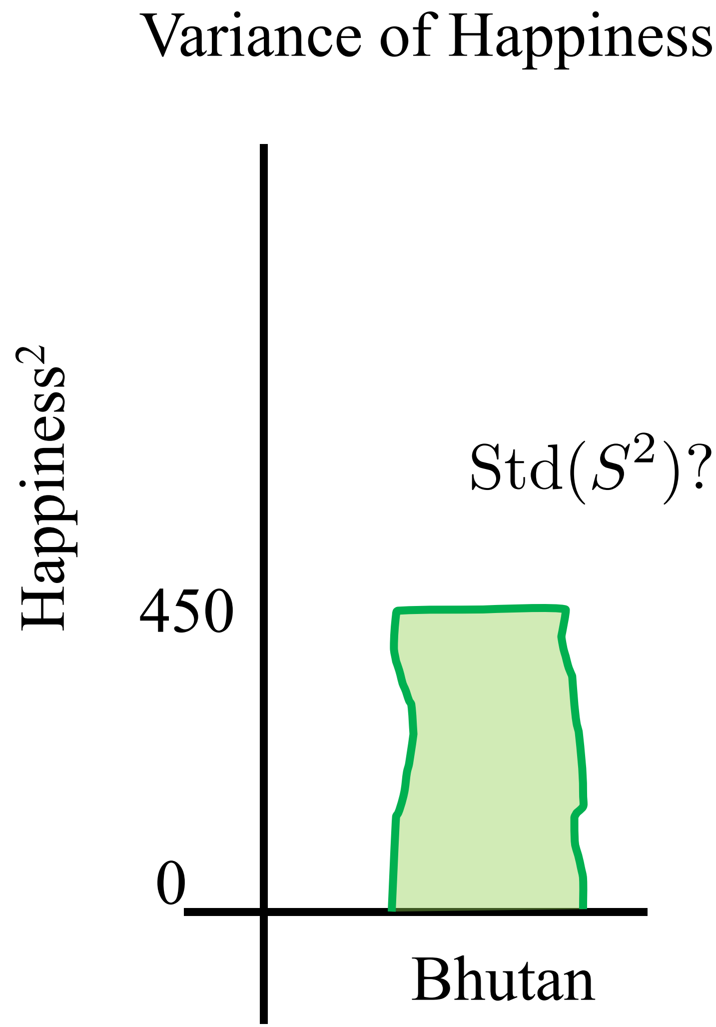
Sample



Collect one (or more) numbers from each person

Sample Mean



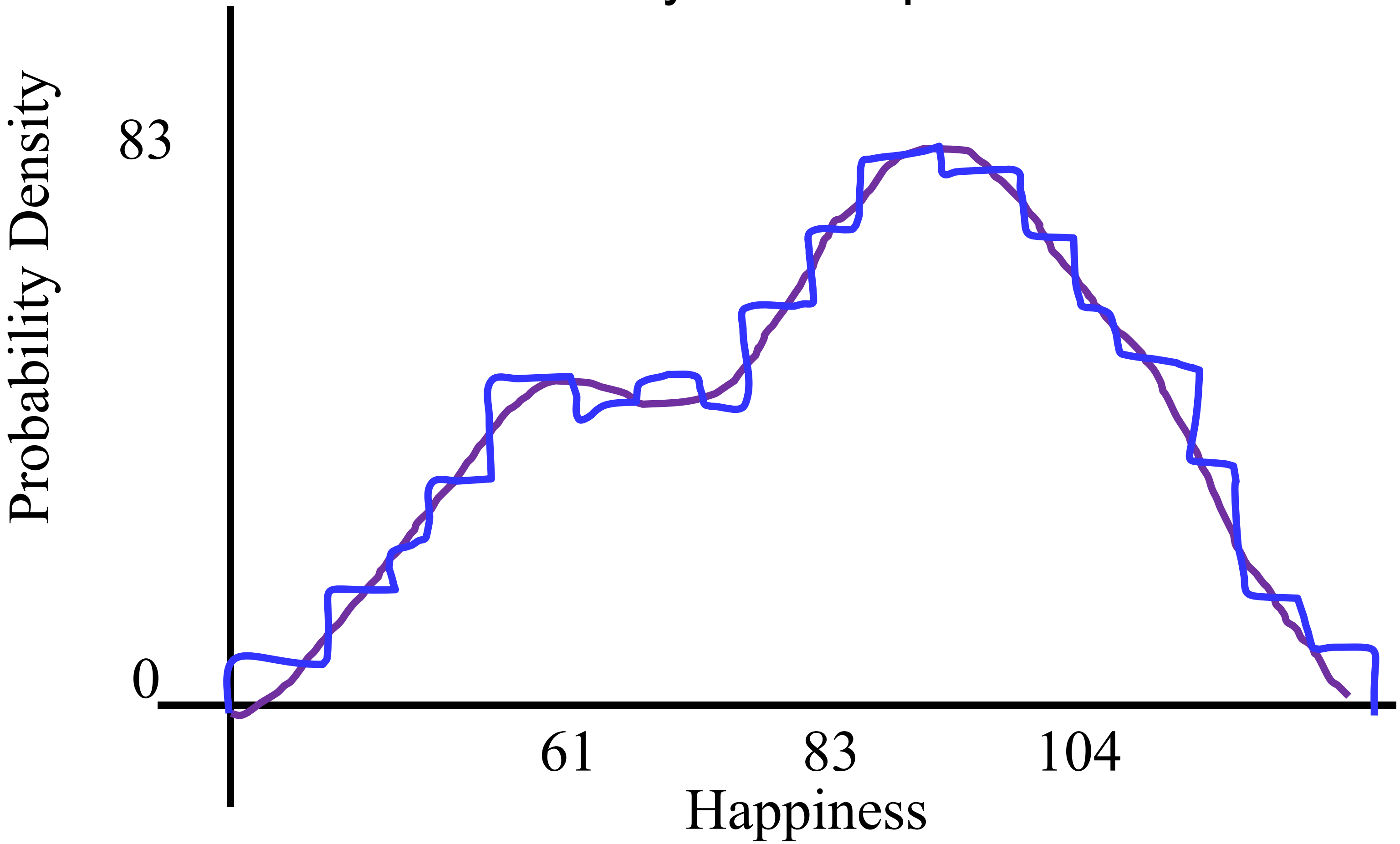


Claim: The average happiness of Bhutan is 83 ± 2

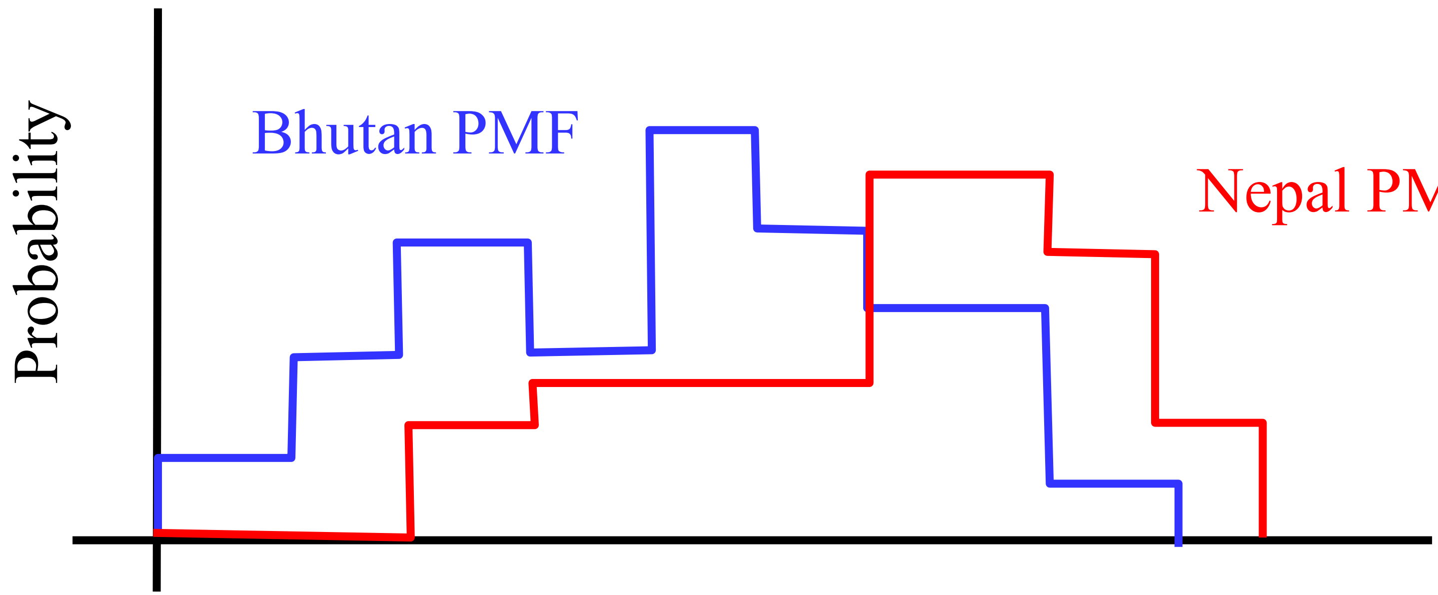
Bootstrap

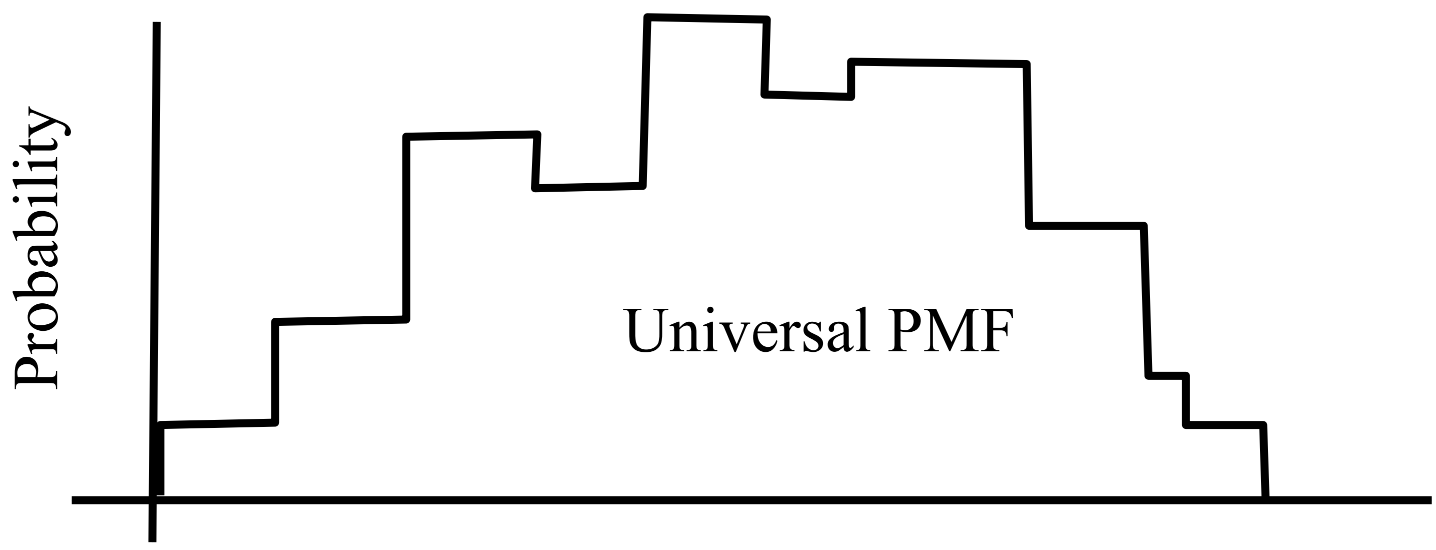
Key Insight

You can estimate the PMF of the underlying distribution, from your sample.



Universal Sample





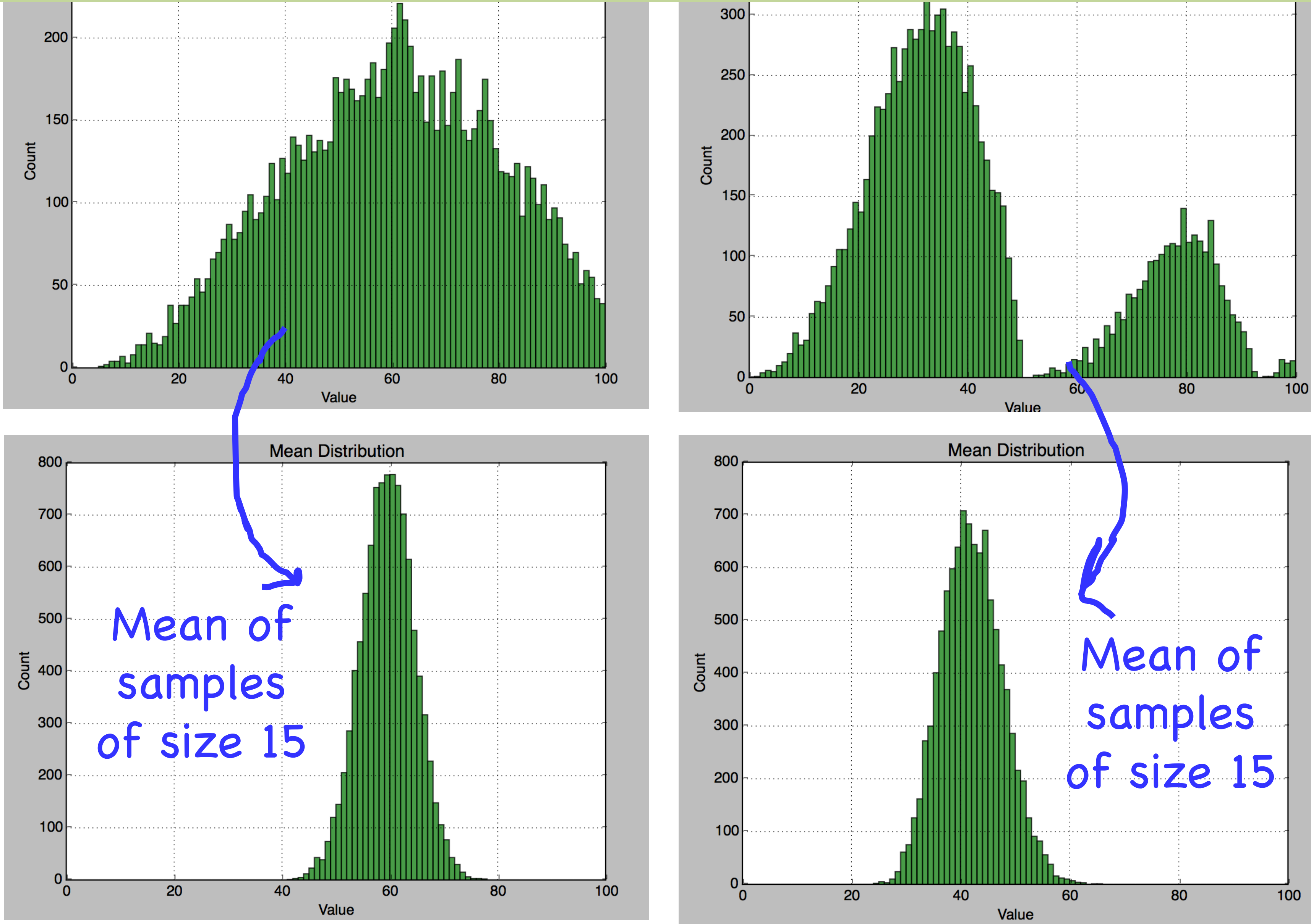
Inequalities

Mystery: Why is Binomial Normal?

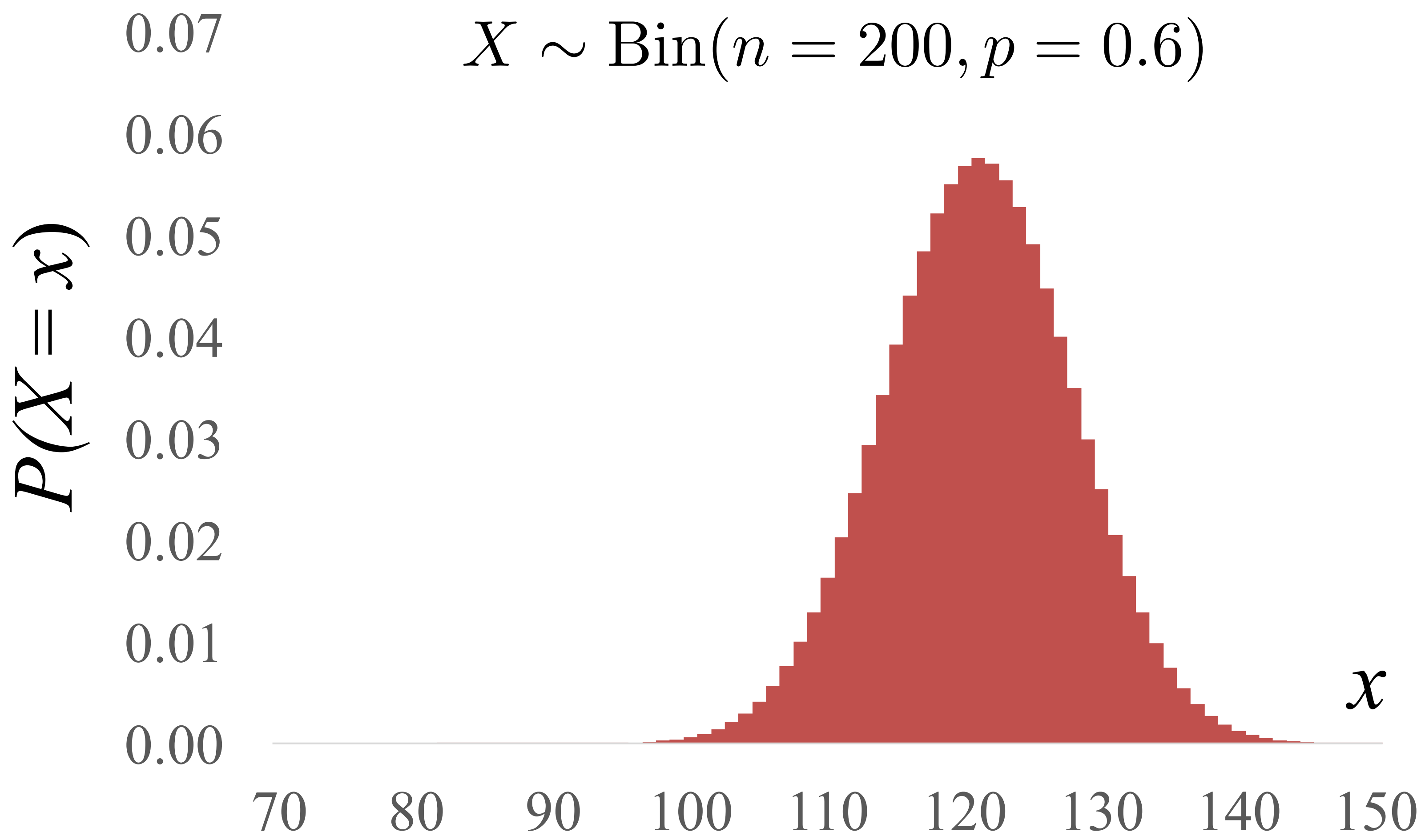
Mystery: Why is the sum of IID uniforms normal?

Mystery: Why is the mean of IID vars normal?

C.L.T. Explains This

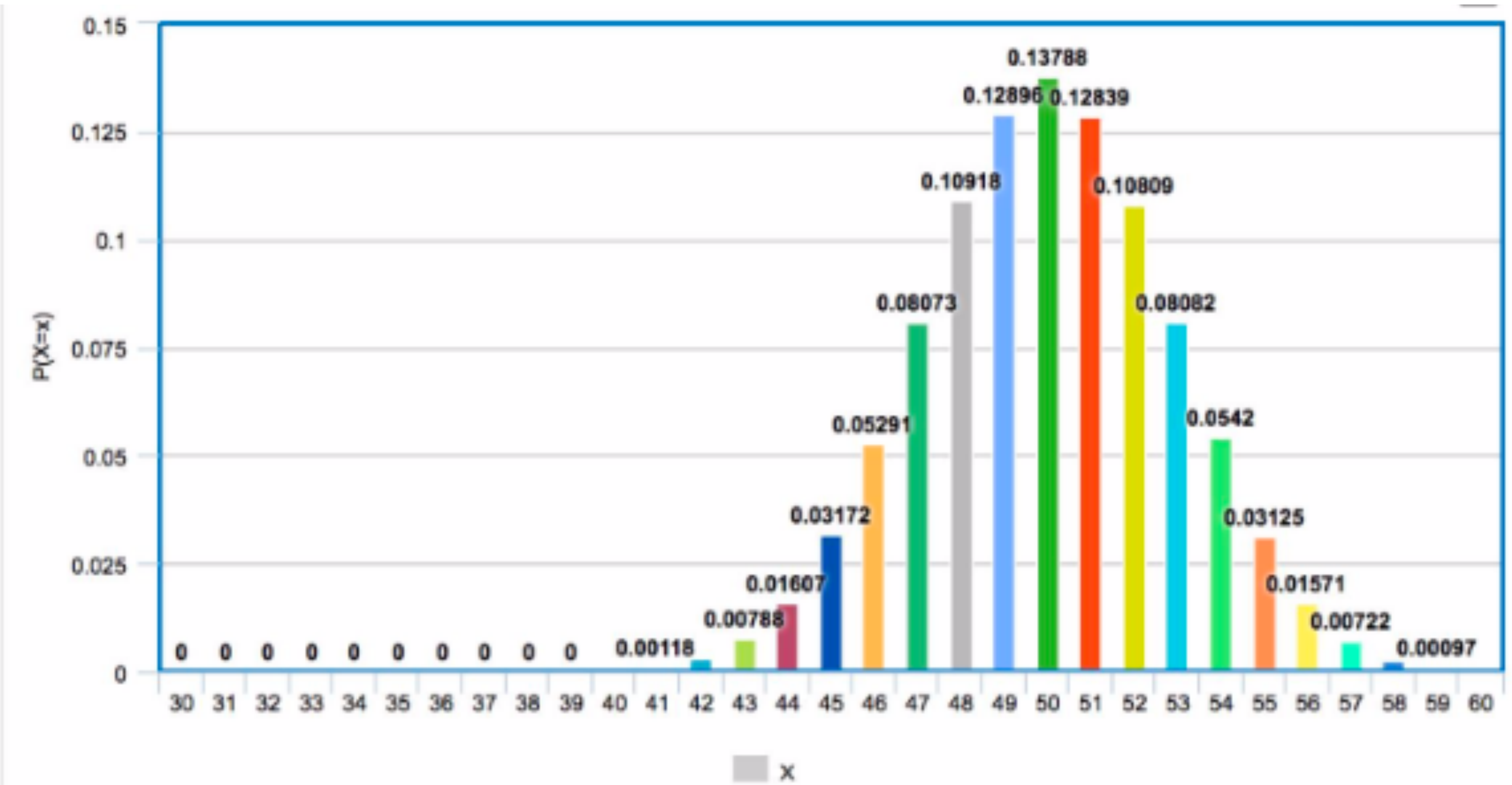


C.L.T. Explains This

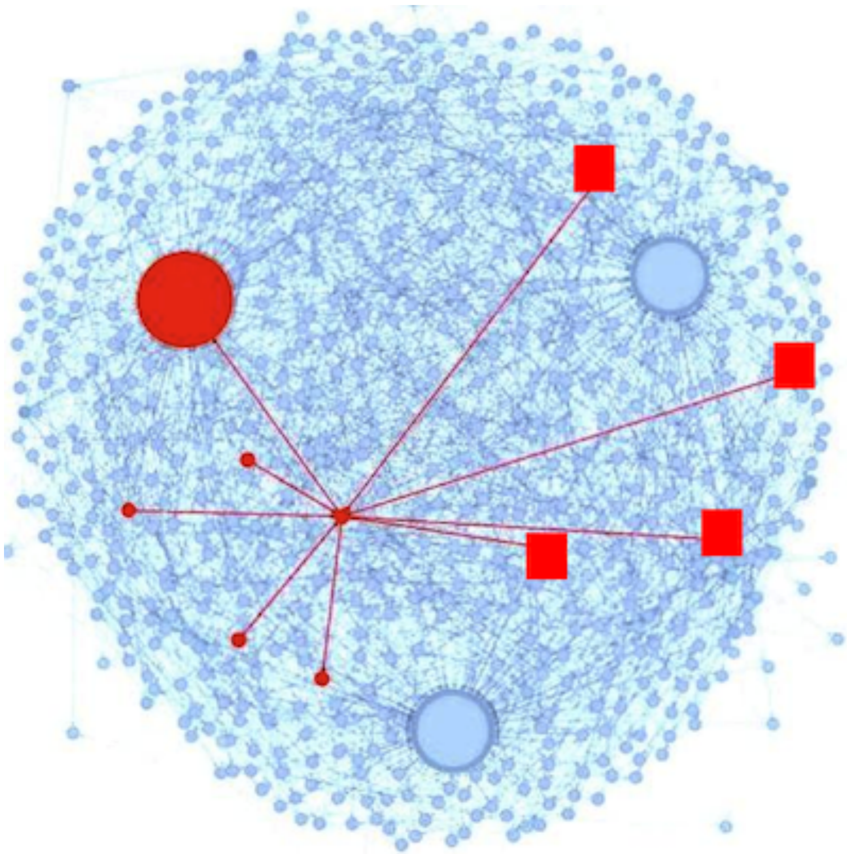


C.L.T. Explains This

Problem set 5: What is the sum of IID uniforms?



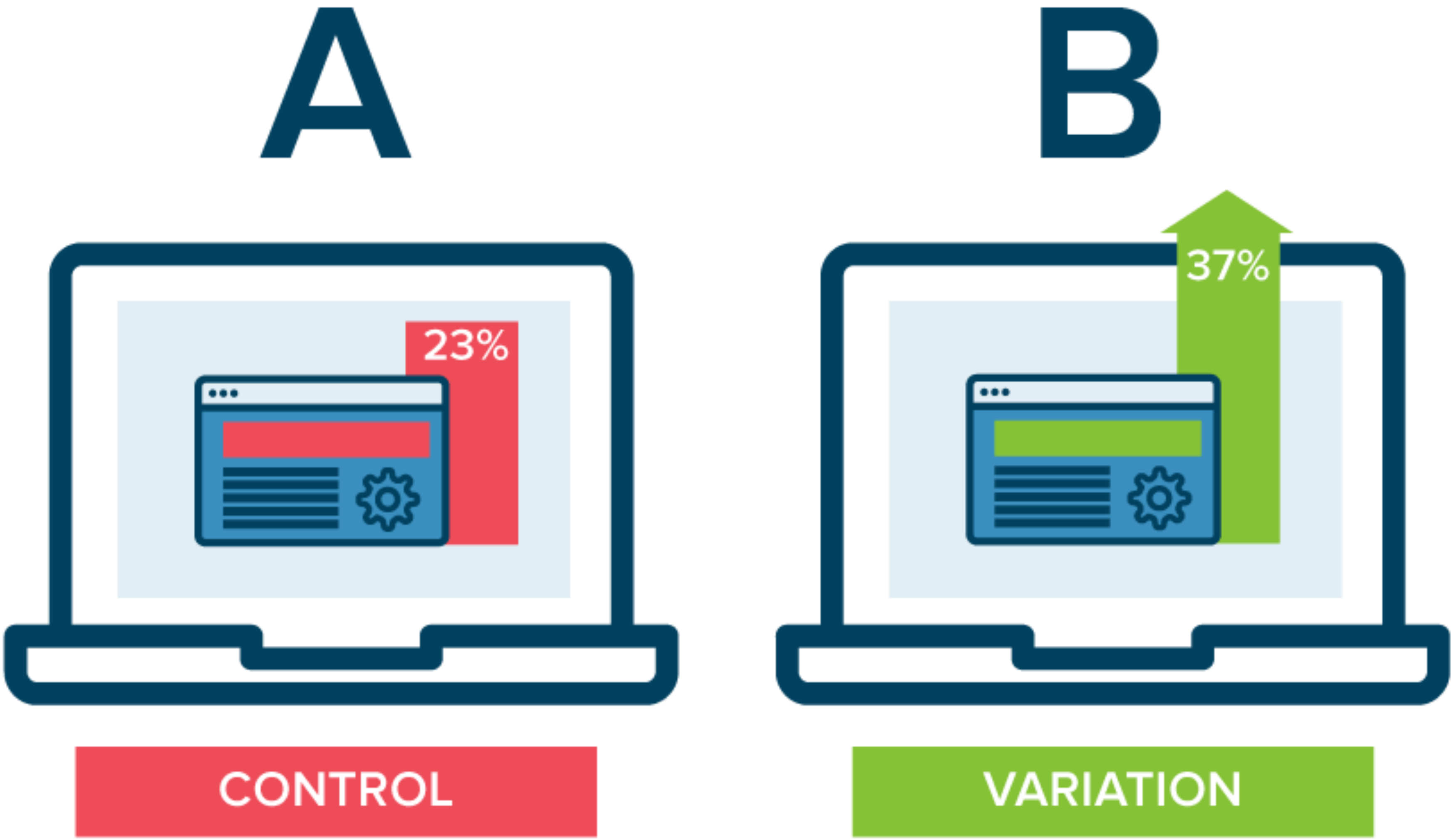
Peer Grading



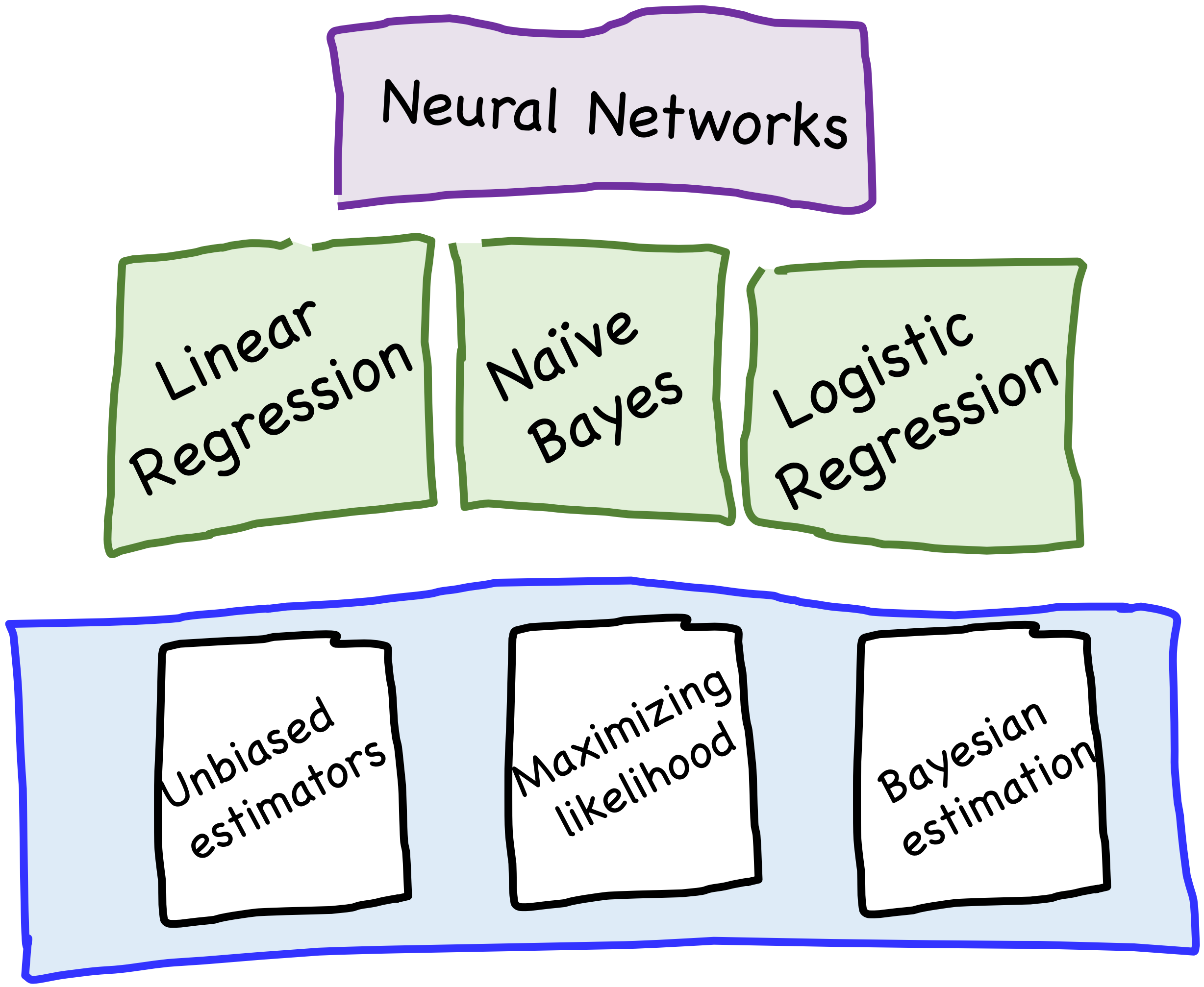
Peer Grading on Coursera HCI.

31,067 peer grades for 3,607 students.

A/B Testing

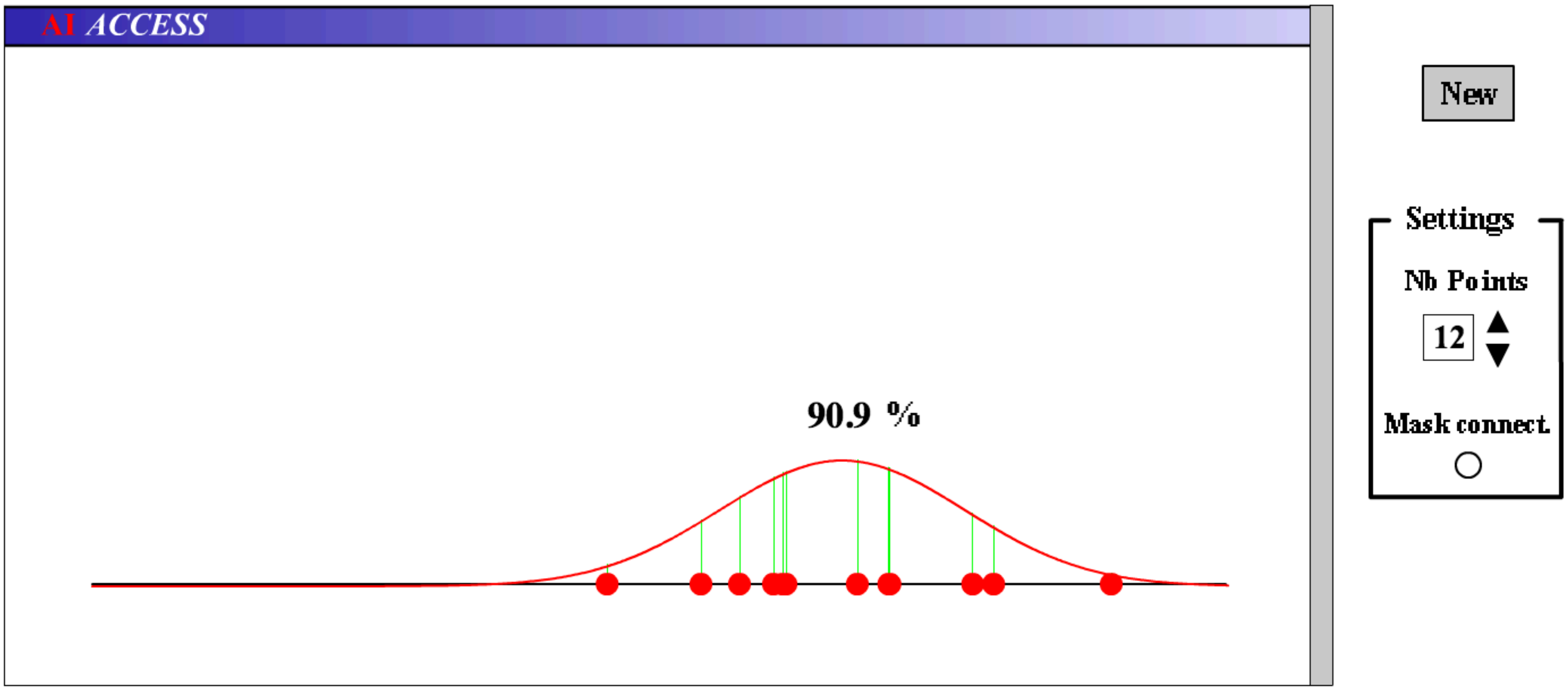


Towards Machine Learning

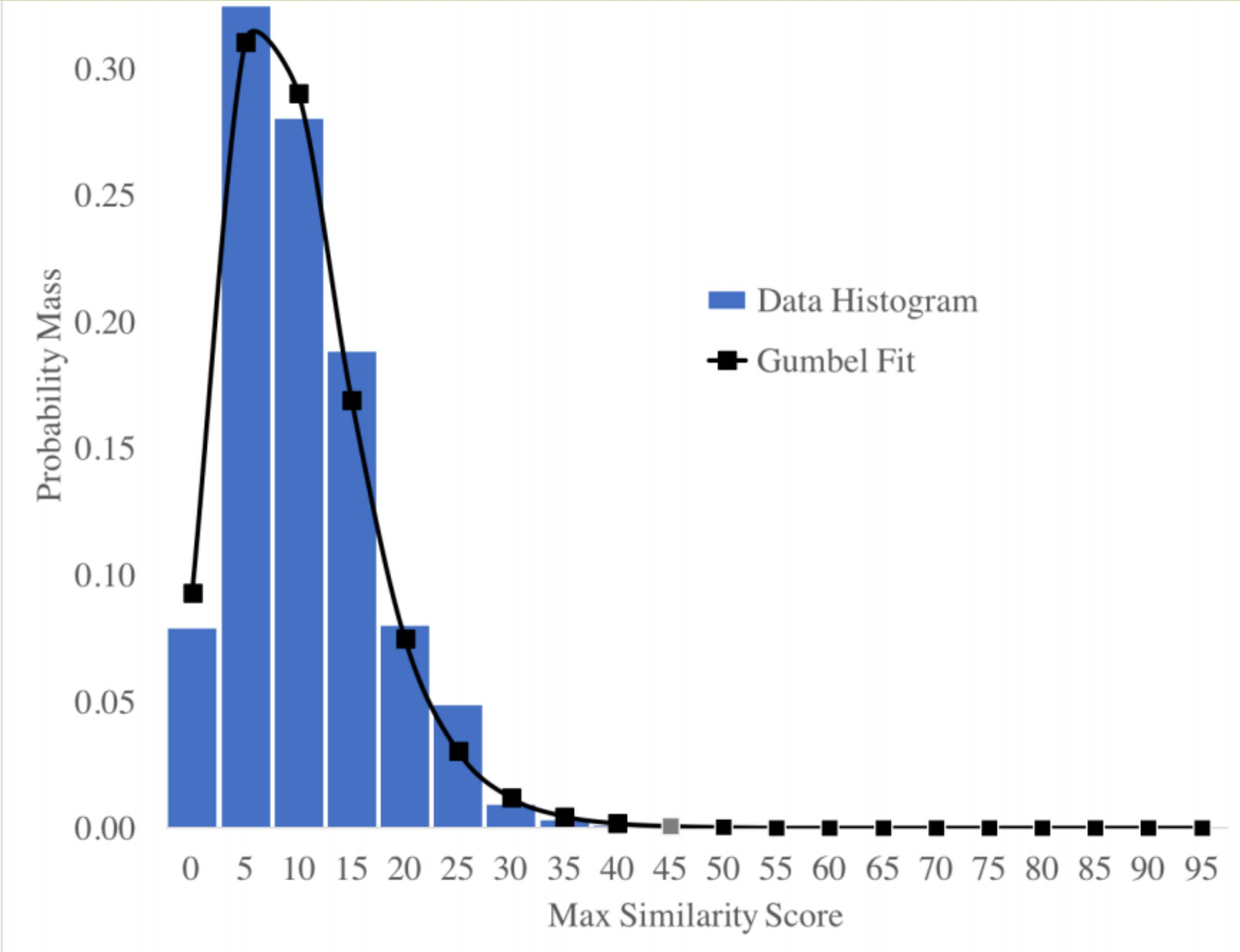


MLE: Likelihood of Data

Likelihood of Data from a Normal

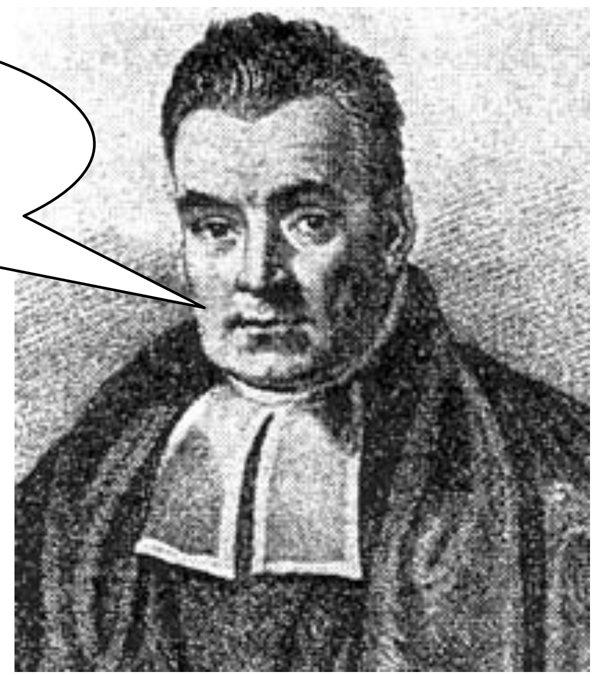


Gumbel Fit


0.30
0.25
0.20
0.15
0.10
0.05
0.00
Probability Mass
Data Histogram
Gumbel Fit
0 5 10 15 20 25 30 35 40 45 50 55 60 65 70 75 80 85 90 95
Max Similarity Score

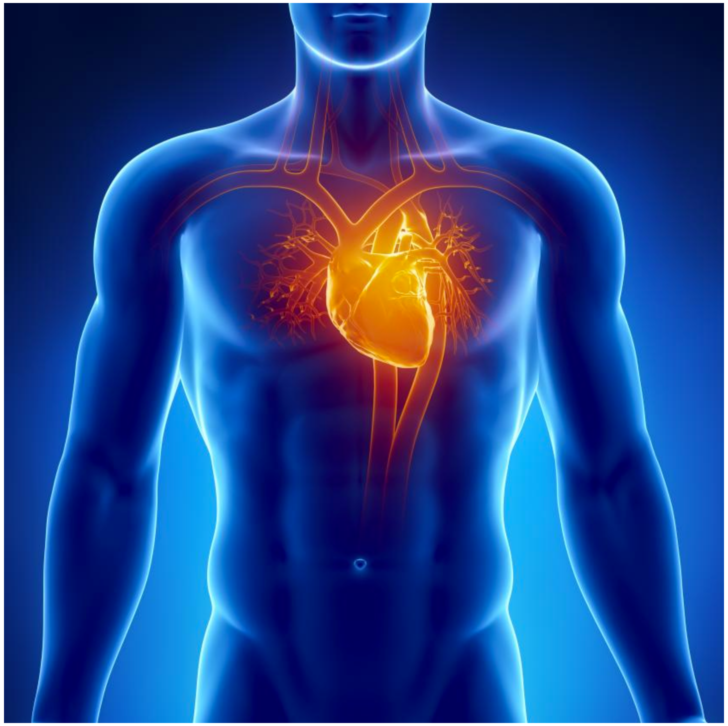
MAP: Most Probable Parameter

So good to see
you again!



Machine Learning

Heart



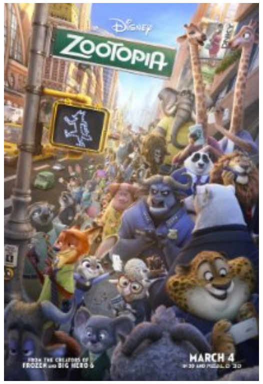
Ancestry

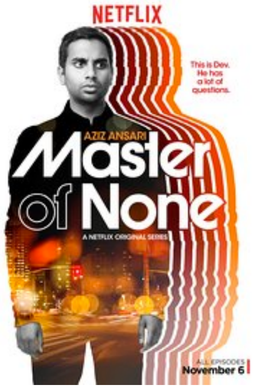
23andMe

Netflix

NETFLIX

Naïve Bayes



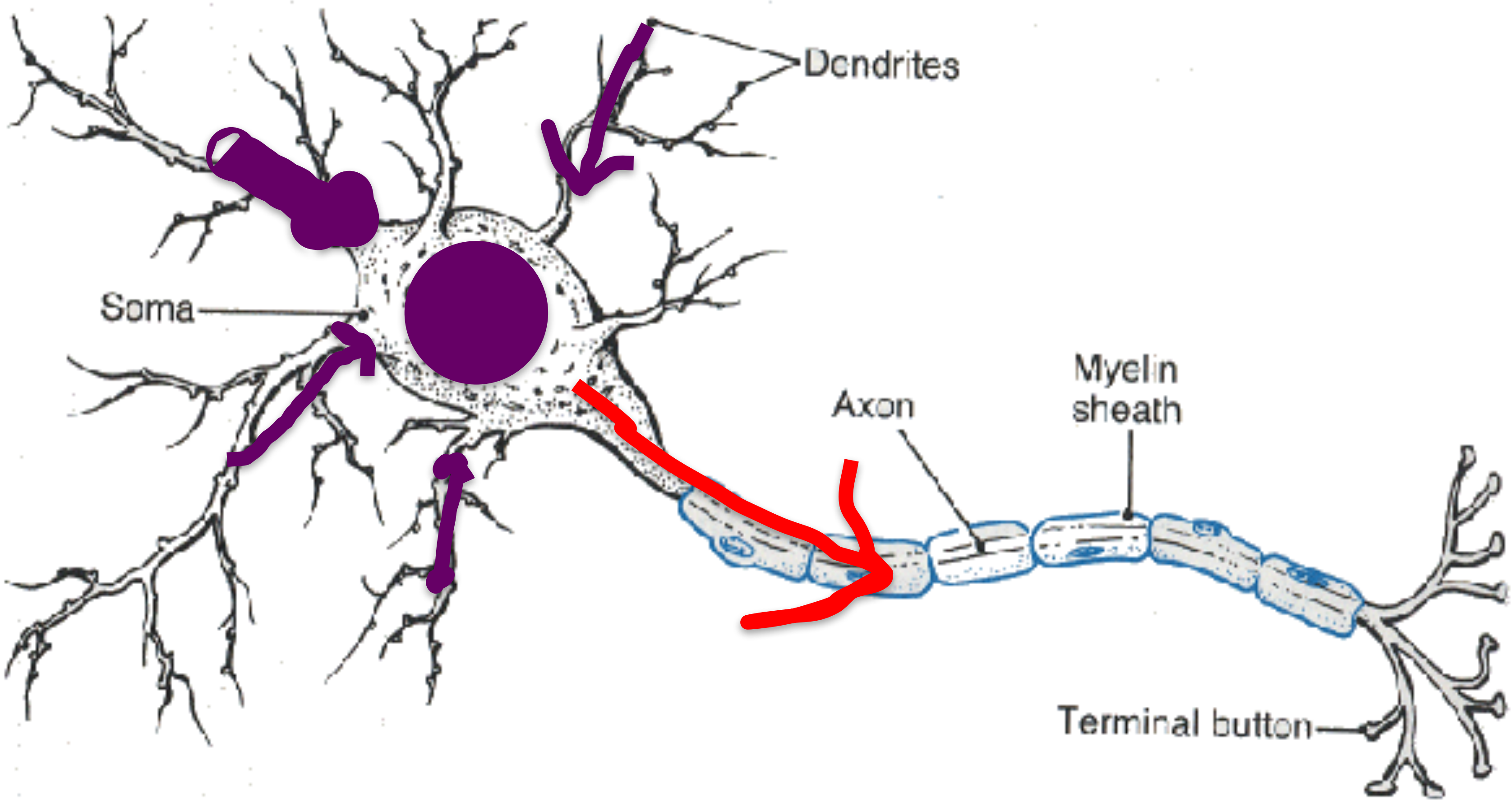


	x_1	x_2	x_3	y
User 1	1	0	1	1
User 2	1	0	1	0
				$\vdots$
User n	0	1	1	1

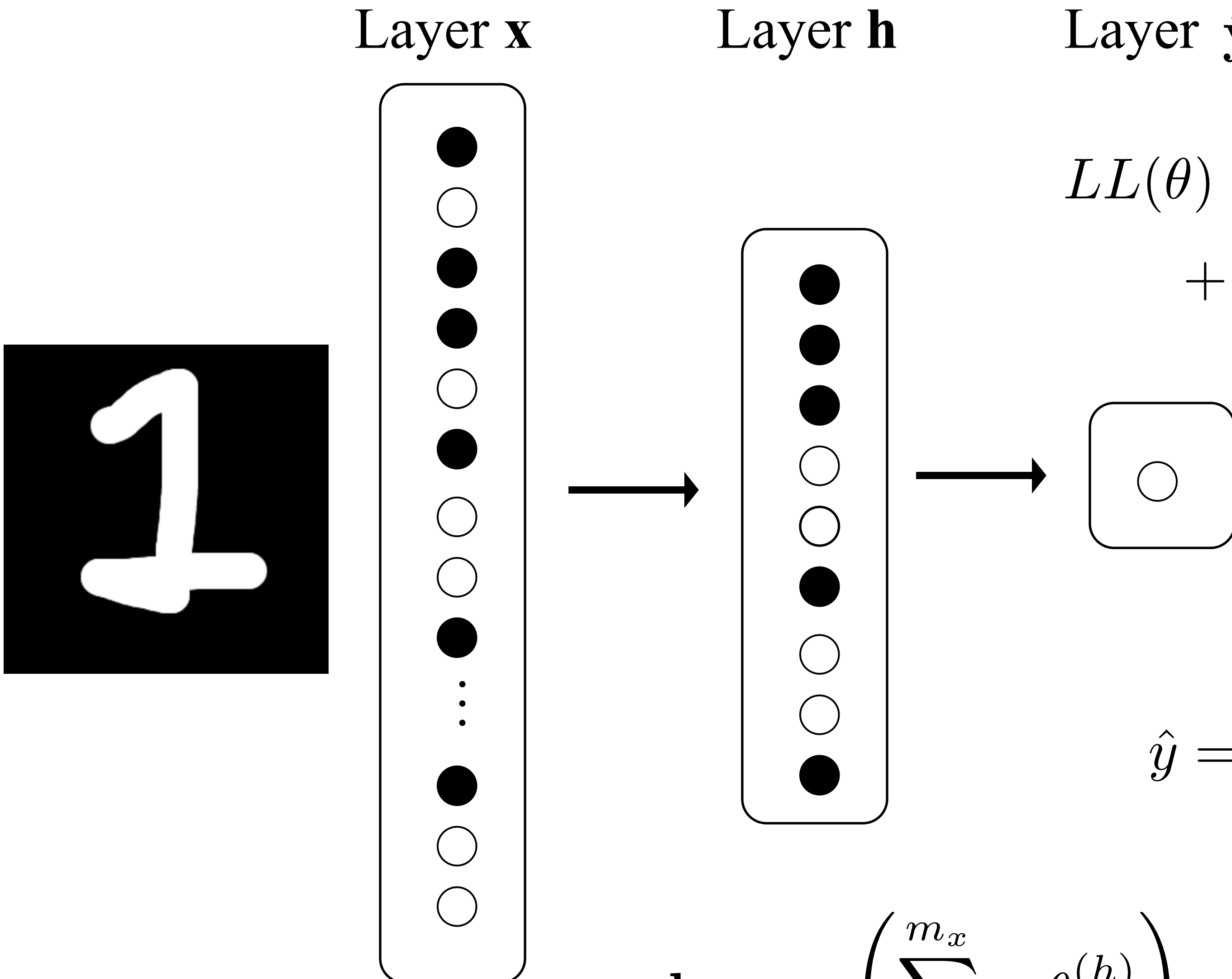
Logistic Regression

$$\mathbf{x}$$

Logistic Regression



Deep Learning



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

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

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

Deep Dream

By the numbers

~600 Fruit


Cuties
seedless
CALIFORNIA MANDARINS
easy peel
Cuties
seedless
CALIFORNIA MANDARINS
easy peel

~ 30 Major Keys

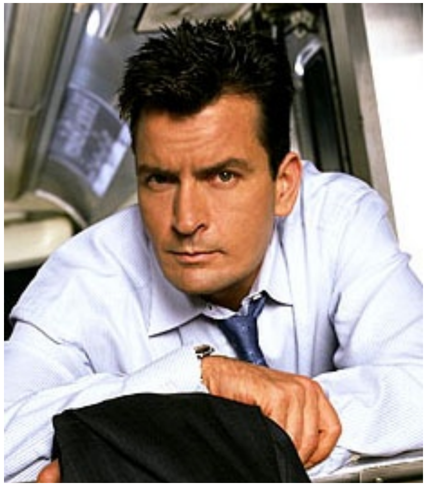
Naïve Bayes Assumption:

$$P(\mathbf{x}|y) = \prod_i P(x_i|y)$$

1 Contest

Thomas Bayes

- Rev. Thomas Bayes (1702 –1761) was a British mathematician and Presbyterian minister



- He looked remarkably similar to Charlie Sheen
  - But that’s not important right now...

Jacob Bernoulli

- Jacob Bernoulli (1654-1705), also known as "James", was a Swiss mathematician

- One of many mathematicians in Bernoulli family
- The Bernoulli Random Variable is named for him
- He is my *academic* great^{12}-grandfather
- Same eyes as Ice Cube

Simeon-Denis Poisson

- Simeon-Denis Poisson (1781-1840) was a prolific French mathematician

- Published his first paper at 18, became professor at 21, and published over 300 papers in his life
  - He reportedly said *"Life is good for only two things, discovering mathematics and teaching mathematics."*
- I'm going with French Martin Freeman

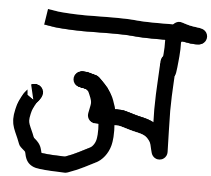
Carl Friedrich Gauss

- Carl Friedrich Gauss (1777-1855) was a remarkably influential German mathematician

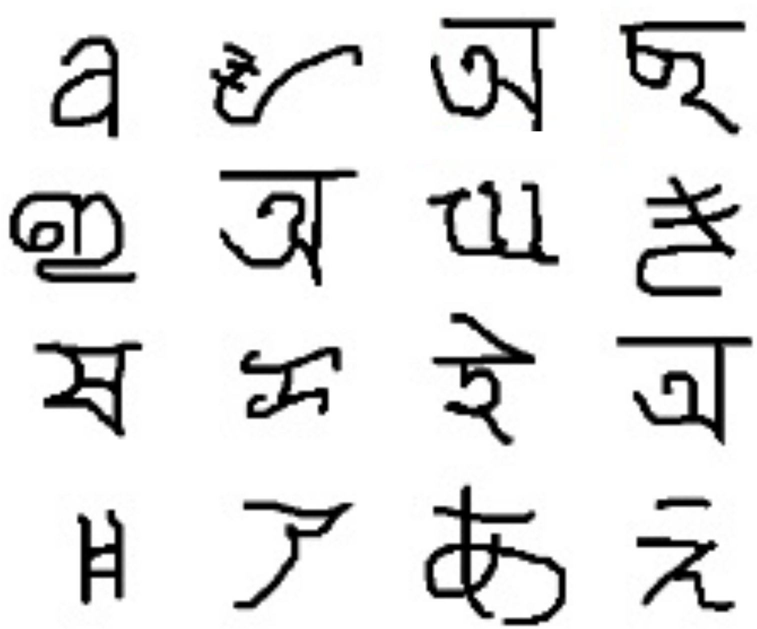
- Started doing groundbreaking math as teenager
  - Did not invent Normal distribution, but popularized it
- He looked like Martin Sheen
  - Who is, of course, Charlie Sheen’s father

Open Problems

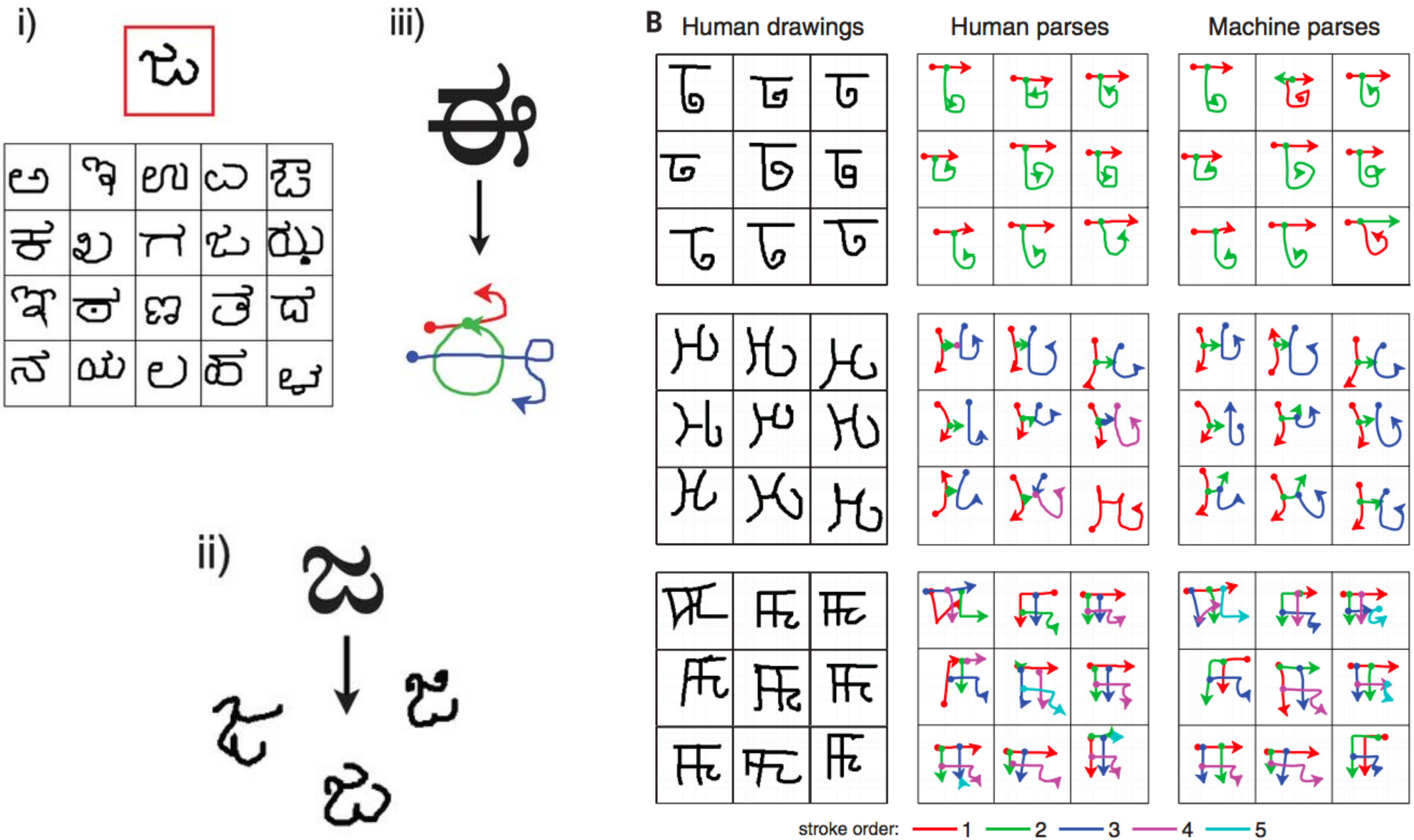
One Shot Learning

Single training example: 

Test set:

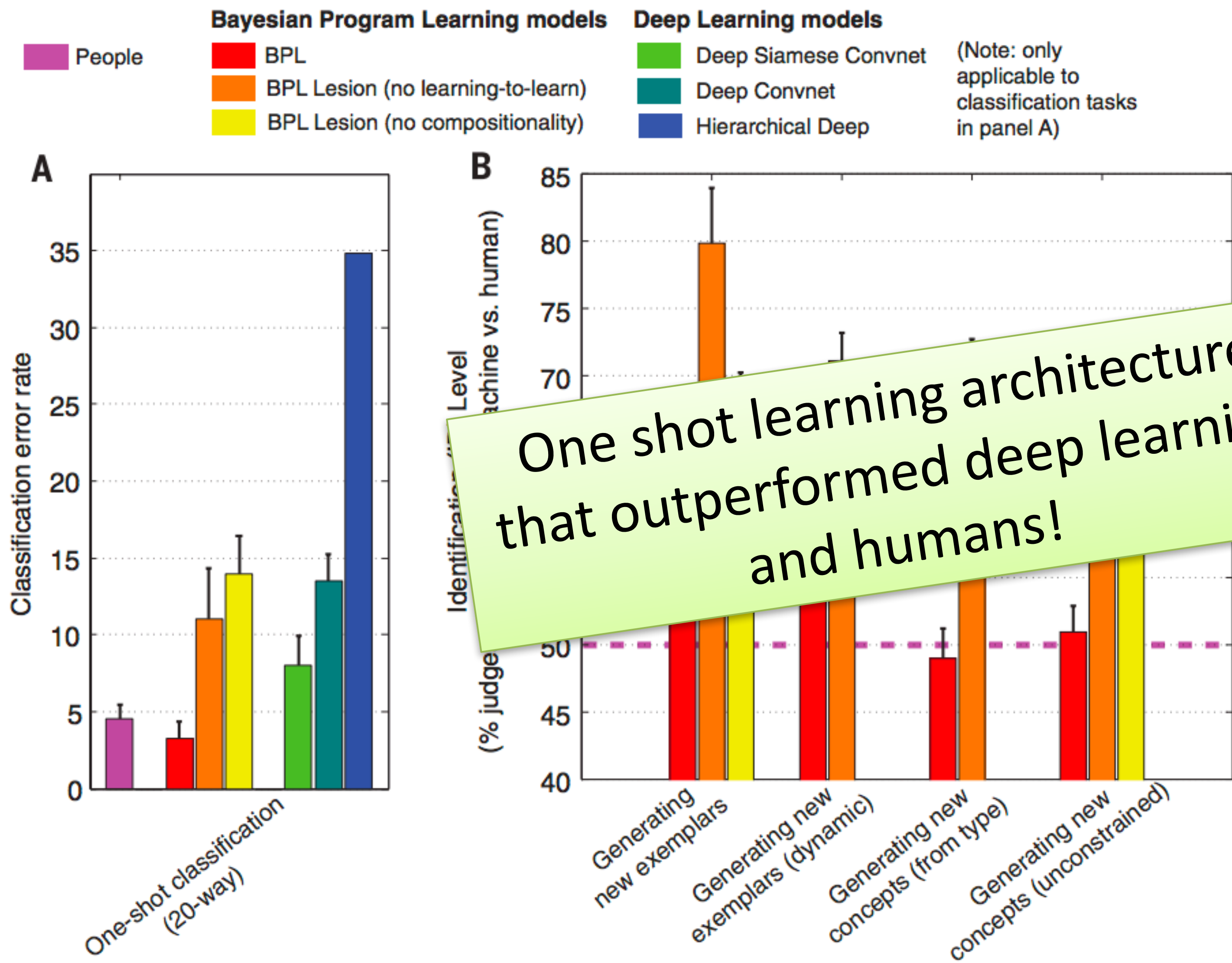


Bayesian Program Learning



Lake et al. Human-level concept learning through probabilistic program induction

Bayesian Program Learning

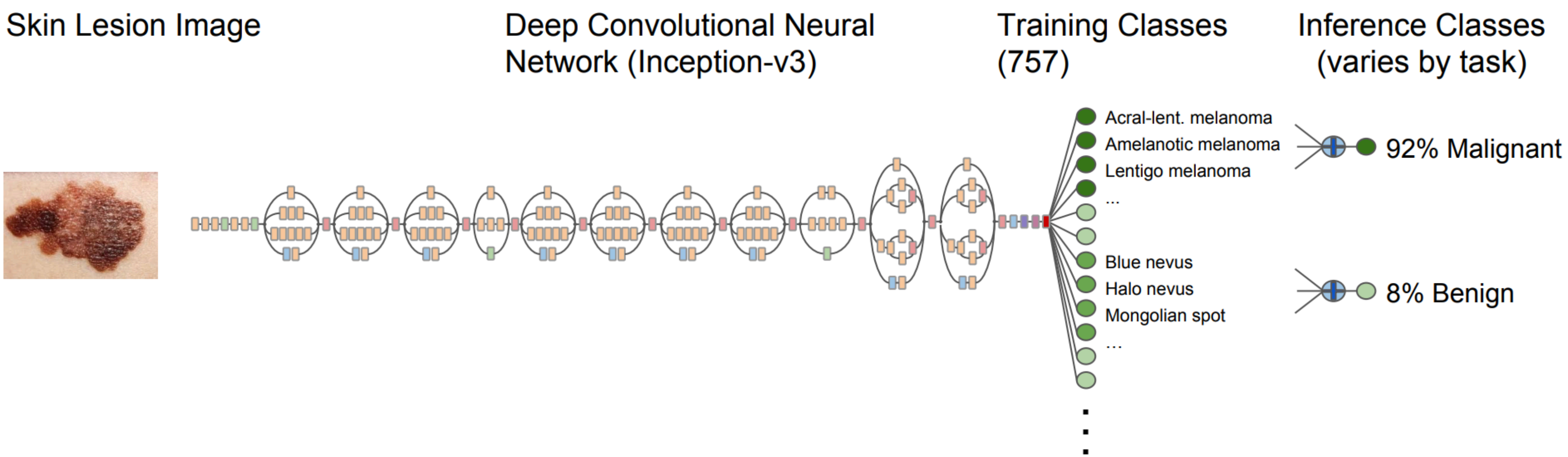


One shot learning architecture that outperformed deep learning and humans!

Lake et al. Human-level concept learning through probabilistic program induction

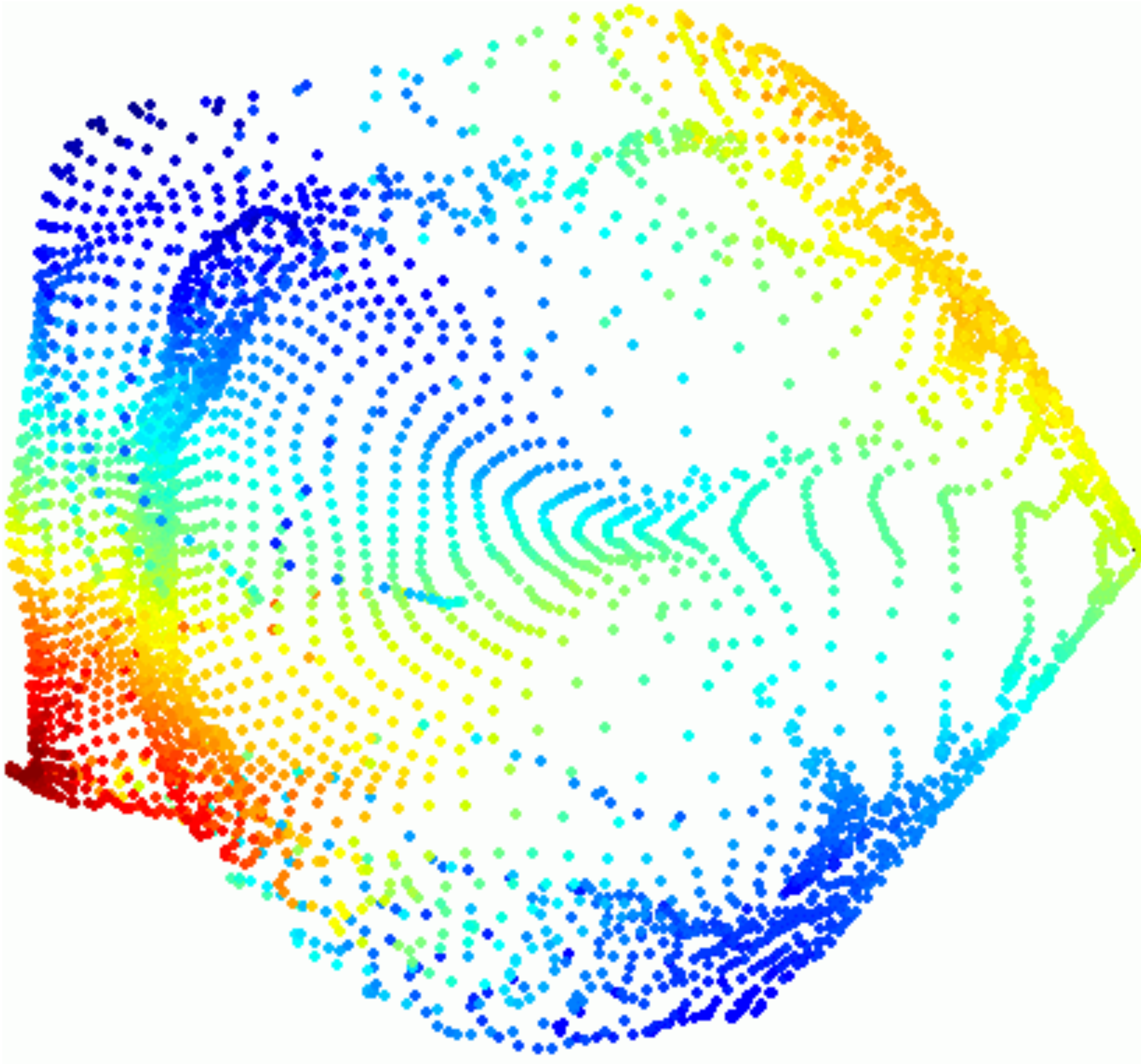
Natural Language

AI for Medicine



Esteva, Andre, et al. "Dermatologist-level classification of skin cancer with deep neural networks." *Nature* 542.7639 (2017): 115-118.

Theoretical Deep Learning



Honorable Mentions:
Smart Grids
Differential Privacy
General AI
Priors and Deep Learning

Ethics and AI

After CS109

Theory
CS161 – Algorithmic analysis
Stats 217– Stochastic Processes
CS 238 – Decision Making Under Uncertainty
CS 228 – Probabilistic Graphical Models

AI
CS 221 – Intro to AI
CS 229 – Macine Learning
CS 230 – Deep Learning
CS 224N – Natural Language Processing
CS 234 – Reinforcement Learning

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

Technology magnifies.
What do we want magnified?