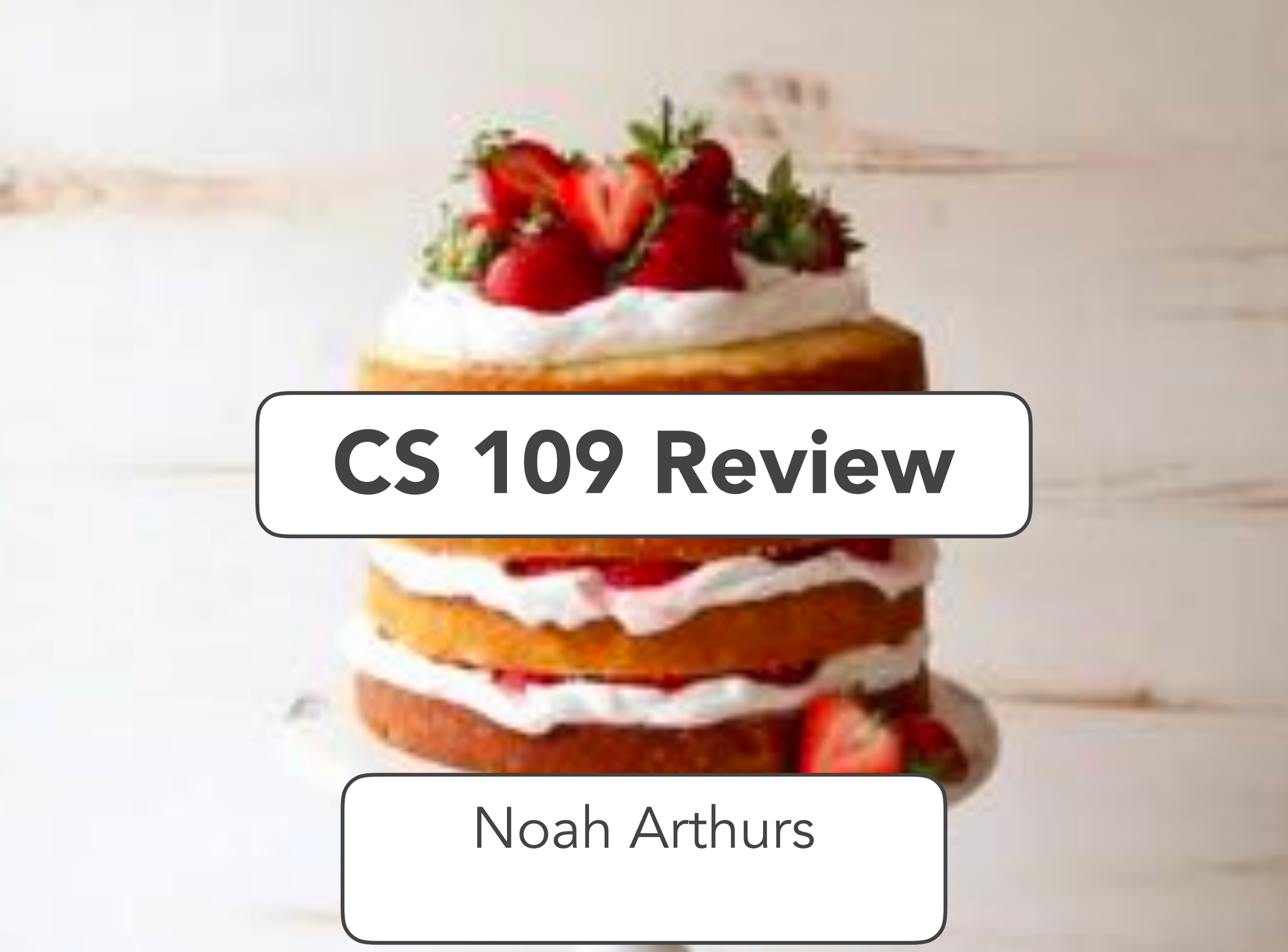


CS 109 Review

Noah Arthurs

CS 109

topics

machine learning

sampling, making conclusions from data

random variables / distributions

core probability fundamentals

methods

examples

demos

problem-solving

stories and memes!

skills

interpreting word problems into math

analyzing and producing code

CS 109

topics

machine learning

sampling, making conclusions from data

random variables / distributions

core probability fundamentals

skills

interpreting word problems into math

analyzing and producing code

Solving a CS109 problem

Word problem

Math expression of question

this is usually what students focus on

this is often the hard part!

Application of formulas

maybe: numerical answer

Step 1: Defining Your Terms

1. Let X represent _____
2. X~_____
3. We want to know _____

Step 1: Defining Your Terms

Problem: If you roll 100 dice, what is the probability of getting less than 30 2's and 5's?

1. Let X represent the number of 2's and 5's we roll
2. X~Binom(100, 2/6)
3. We want to know P(X < 30)

Translating English to Probability

What the problem asks:	What you should immediately think:
"What's the probability of _____ "	P( )
"___ given ___", "___ if ___"	___ | ___
"at least ___"	*could we use what we know about everything less than __?*
"approximate ____."	*use an approximation!*
"How many ways…"	*combinatorics*

these are just a few, and these are why practice is the best way to prepare for the exam!

CS 109

topics

machine learning

sampling, making conclusions from data

random variables / distributions

core probability fundamentals

skills

interpreting word problems into math

analyzing and producing code

Code in CS 109

Analysis

Expectation of binary tree depth

Bloom Filter Analysis

Expectation of recursive die roll game

Implementation

Dithering

CO2 Levels

Biometric Keystrokes

Titanic

Peer Grading

Thompson Sampling

CS 109

topics

machine learning

sampling, making conclusions from data

random variables / distributions

core probability fundamentals

counting

conditional probability

probability principles

Counting

Sum Rule	Inclusion-Exclusion Principle
$outcomes = \|A\| + \|B\|$ $if \|A \cap B\| = 0$	$\|A\| + \|B\| - \|A \cap B\|$ $for\ any\ \|A \cap B\|$
Product Rule	**Pigeonhole Principle**
$outcomes = \|A\| \times \|B\|$ **if all outcomes of B are possible regardless of the outcome of A**	**If m objects are placed into n buckets, then at least one bucket has at least *ceiling(m / n)* objects.**

Combinatorics: Arranging Items

	Permutations (ordered)	**Combinations (unordered)**
Distinct	$n!$	$\binom{n}{k} = \frac{n!}{k!(n-k)!}$
Indistinct	$\frac{n!}{k_1!k_2!\ldots k_n!}$	$\binom{n+r-1}{r-1}$ the divider method!

Probability basics

$$P(E) = \lim_{x->\infty} \frac{n(E)}{n}$$

in the general case

$$\text{Probability} = \frac{\text{Event space}}{\text{Sample space}}$$

if all outcomes are equally likely!

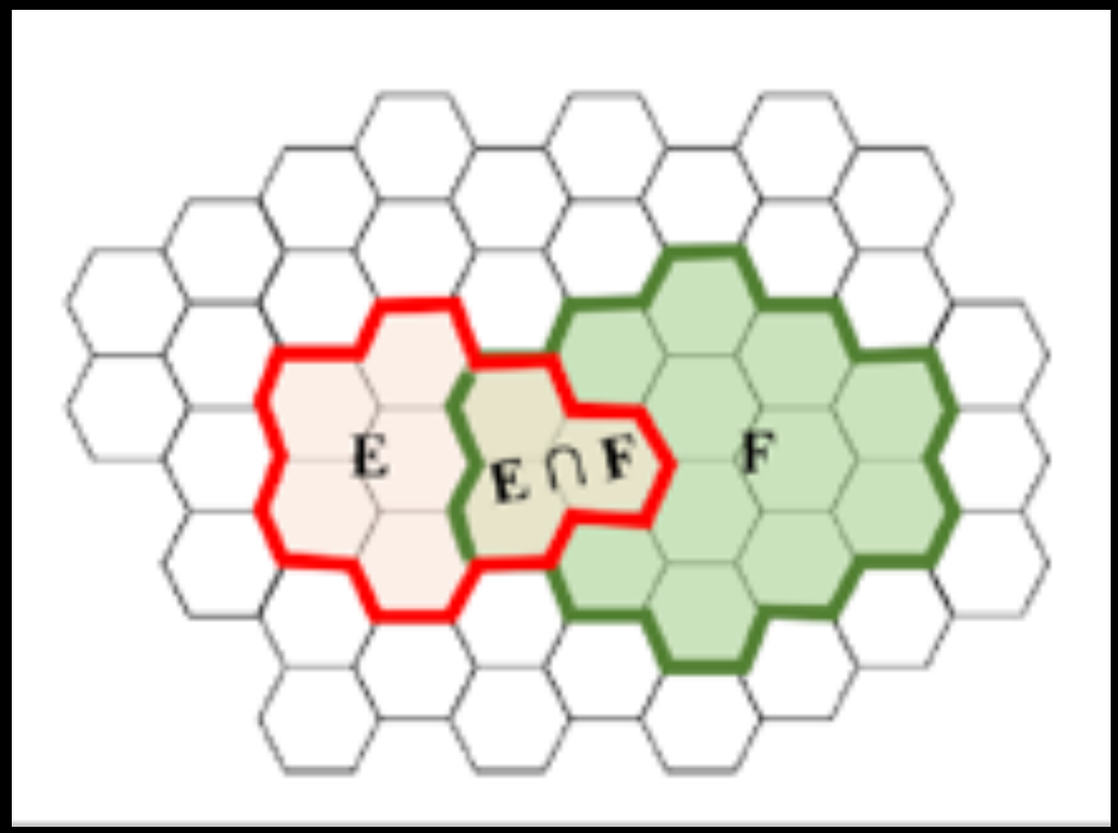
(use counting with distinct objects)

Axioms: $0 \leq P(E) \leq 1$ $P(S) = 1$ $P(E^C) = 1 - P(E)$

Conditional Probability

definition:

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



* $P(EF) = P(E \cap F)$

Chain Rule:

$$P(EF) = P(E|F)P(F)$$

Law of Total Probability

Let's say we don't know P(A), but we do know the probability of A given any value of B:

$$P(A) = P(AB) + P(AB^C)$$

$$= P(A \mid B)P(B) + P(A \mid B^C)P(B^C)$$

If B can take on any value in S:

$$P(A) = \sum_{b \in S} P(A, B = b)$$

$$P(A) = \sum_{b \in S} P(A \mid B = b)P(B = b)$$

Bayes' Rule

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

Bayes' Rule

posterior

likelihood

prior

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

normalization constant

Bayes' Rule

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

$$P(F|E)P(E) + P(F|E^C)P(E^C)$$

divide the event F into all the possible ways it can happen; use LoTP

Which rule when?

$$P(EF) = P(E|F)P(F)$$

- Goes from an “and” to a conditional or vice versa
- Think about which event you want to condition on

$$P(A) = P(A|B)P(B) + P(A|B^C)P(B^C)$$

- We don’t know about A but we do know about A|B
- Don’t forget about the “and” version and “summation” version

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

- Good for when E|F is hard but F|E is not so hard
- Common mistake: not trying chain rule first

Old Principles, New Tricks

Name of Rule	Original Rule	Conditional Rule
First axiom of probability	$0 \le P(E) \le 1$	$0 \le P(E \mid G) \le 1$
Complement Rule	$P(E) = 1 - P(E^C)$	$P(E \mid G) = 1 - P(E^C \mid G)$
Chain Rule	$P(EF) = P(E \mid F)P(F)$	$P(EF \mid G) = P(E \mid FG)P(F \mid G)$
Bayes Theorem	$P(E \mid F) = \frac{P(F\|E)P(E)}{P(F)}$	$P(E \mid FG) = \frac{P(F\|EG)P(E\|G)}{P(F\|G)}$

Combining Events

$$P(ABC) = ?$$

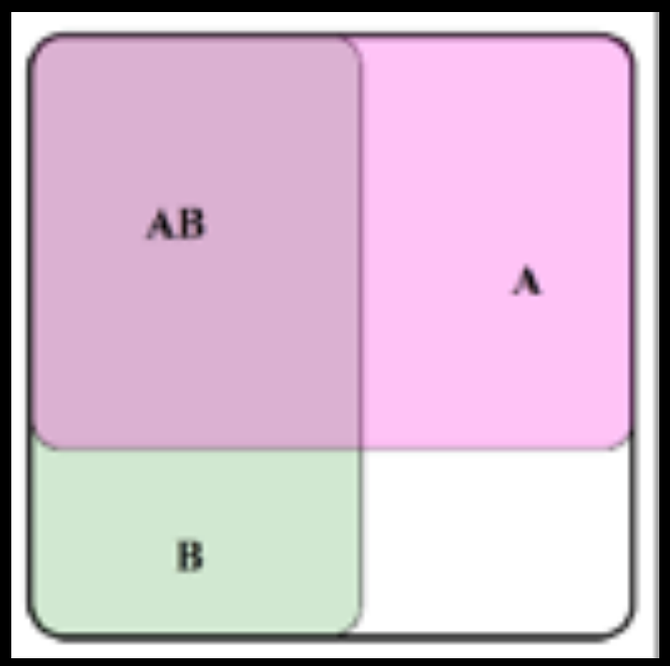
Let X = AC:

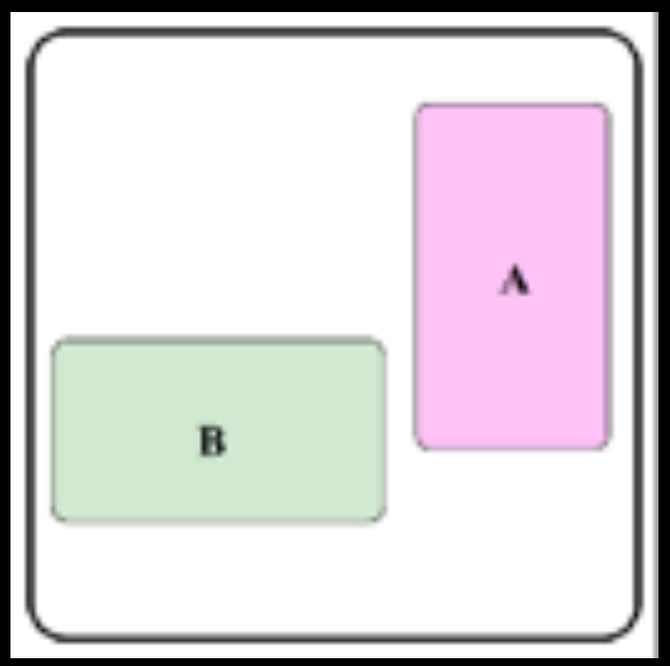
$$P(ABC) = P(BX) = P(B|X)P(X) = P(B|AC)P(AC)$$

There are three correct ways to apply chain rule to P(ABC)!

Independence

Independence	Mutual Exclusion
$P(EF) = P(E)P(F)$	$\lvert E \cap F \rvert = 0$
“AND”	“OR”





Independence

Independence	Conditional Independence
$P(EF) = P(E)P(F)$	$P(EF \mid G) = P(E \mid G)P(F \mid G)$ $P(E \mid FG) = P(E \mid G)$
“AND”	**“AND [if]”**

If E and F are independent…..

…..that does not mean they’ll be independent if another event happens!

& vice versa

CS 109

topics

machine learning

sampling, making conclusions from data

multivariate distributions

random variables / distributions

discrete RVs

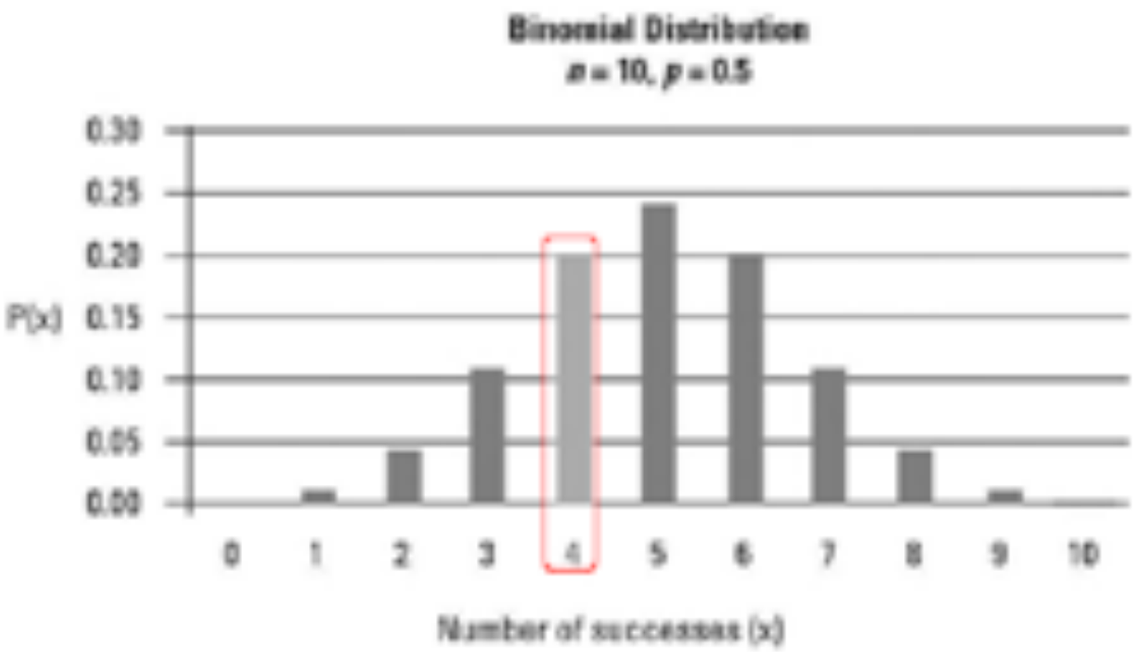
continuous RVs

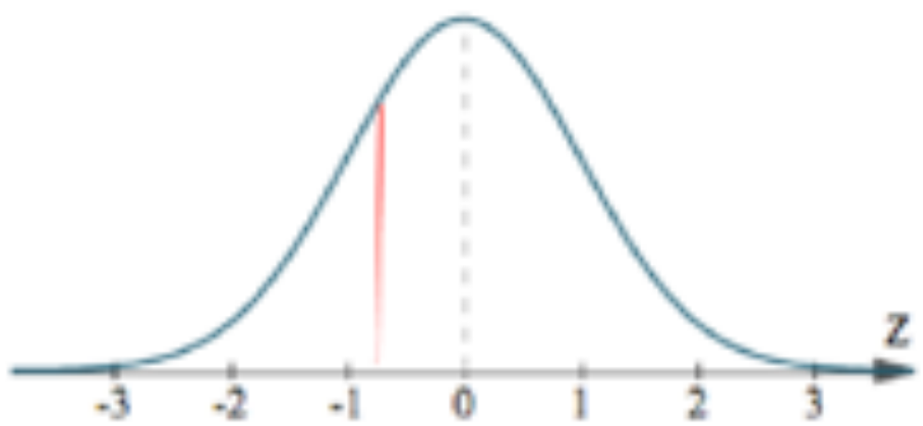
properties of RVs

core probability fundamentals

Probability Distributions

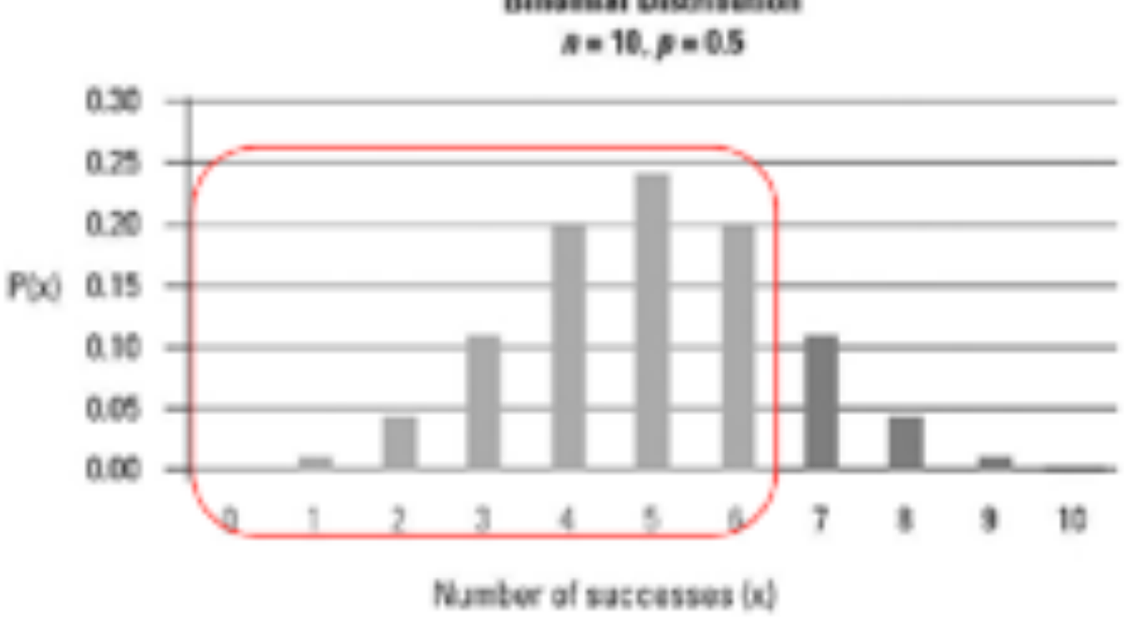
Discrete	Continuous
PMF:	PDF:





Probability Distributions

Discrete	Continuous
CDF:	CDF:



Expectation & Variance

Discrete definition

$$E[X] = \sum_{x:P(x)>0} x * P(x)$$

Continuous definition

$$E[X] = \int_{x} x * f_X(x)dx$$

Properties of Expectation

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

$$E[aX + b] = aE[X] + b$$

$$E[g(X)] = \sum_{x} g(x) * p_X(x)$$

Properties of Variance

$$Var(X) = E[X^2] - E[X]^2$$

$$Var(aX + b) = a^2 Var(X)$$

If X and Y are independent:

$$Var(X + Y) = Var(X) + Var(Y)$$

All our (discrete) friends

Ber(p)	Bin(n, p)	Poi(λ)	Geo(p)	NegBin (r, p)
P(X) = p	$\binom{n}{k} p^k (1-p)^{n-k}$	$\frac{\lambda^k e^{-\lambda}}{k!}$	$(1-p)^{k-1} p$	$\binom{k-1}{r-1} p^r (1-p)^{k-r}$
E[X] = p	E[X] = np	E[X] = λ	E[X] = 1 / p	E[X] = r / p
Var(X) = p(1-p)	Var(X) = np(1-p)	Var(X) = λ	$\frac{1-p}{p^2}$	$\frac{r(1-p)}{p^2}$
Getting candy or not at a random house	# houses out of 20 that give out candy	# houses in an hour that give out candy	# houses to visit before getting candy	# houses to visit before getting candy 3 times

All our (continuous) friends

Uni(α, β)	Exp(λ)	N(μ, σ)
$f(x) = \frac{1}{\beta - \alpha}$	$f(x) = \lambda e^{-\lambda x}$	$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{\frac{-(x-\mu)^2}{2\sigma^2}}$
$P(a \leq X \leq b) = \frac{b-a}{\beta - \alpha}$	$F(x) = 1 - e^{-\lambda x}$	$F(x) = \Phi(\frac{x-\mu}{\sigma})$
$E(x) = \frac{\alpha + \beta}{2}$	E[x] = 1 / λ	E[x] = μ
$Var(x) = \frac{(\beta - \alpha)^2}{12}$	$Var(x) = \frac{1}{\lambda^2}$	$Var(x) = \sigma^2$
thickness of sidewalk pavement between houses	time until feet get too sore to trick or treat	weight of filled candy baskets

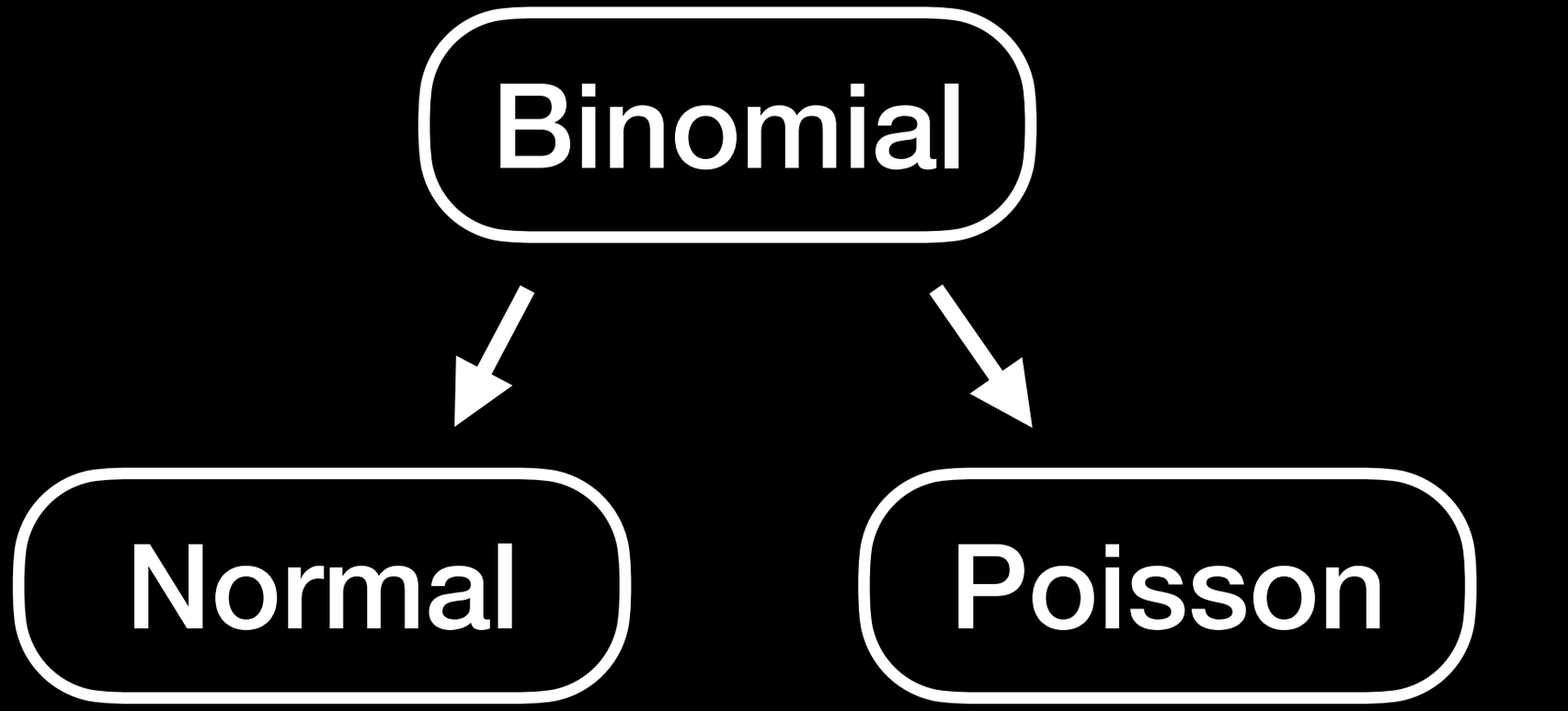
Discrete vs Continuous

Discrete	Continuous
$E[X] = \sum_{x:P(x)>0} x * P(x)$	$E[X] = \int_x x * f_X(x)dx$
$P(EF) = P(E \mid F)P(F)$	$f(E = e, F = f) = f(E = e \mid F = f)f(F = f)$
$P(A) = \sum_{b \in S} P(A \mid B = b)P(B = b)$	$P(A) = \int_b P(A \mid B = b)f_B(b)db$
$P(E \mid F) = \frac{P(F \mid E)P(E)}{P(F)}$	$f(E = e \mid F) = \frac{P(F \mid E = e)f(E = e)}{P(F)}$

Approximations

When can we approximate a binomial?

n is large

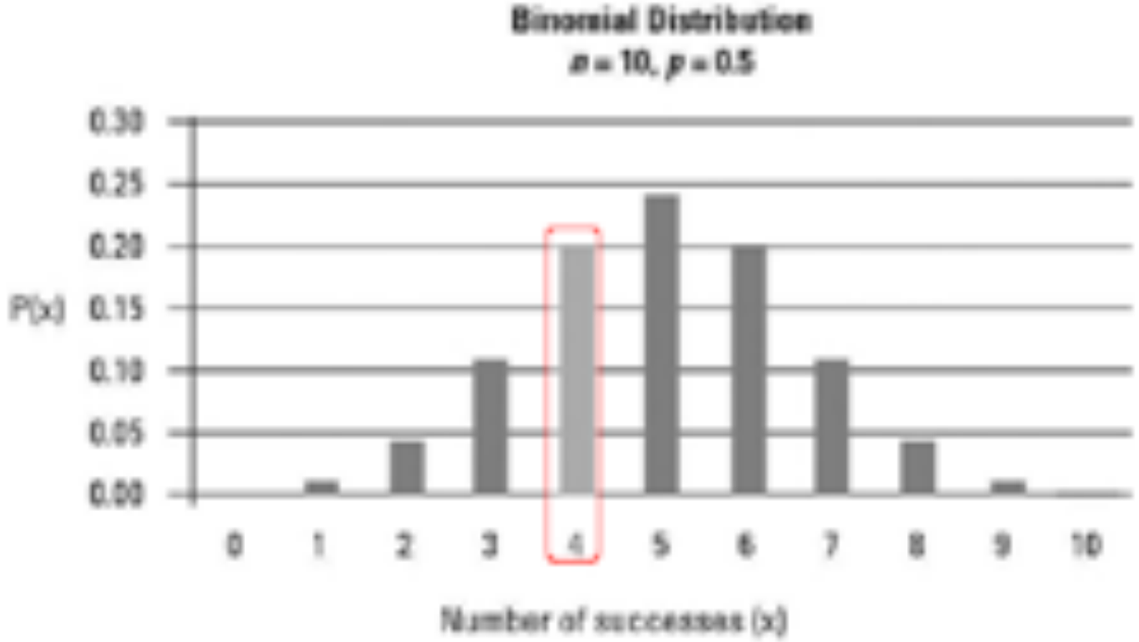


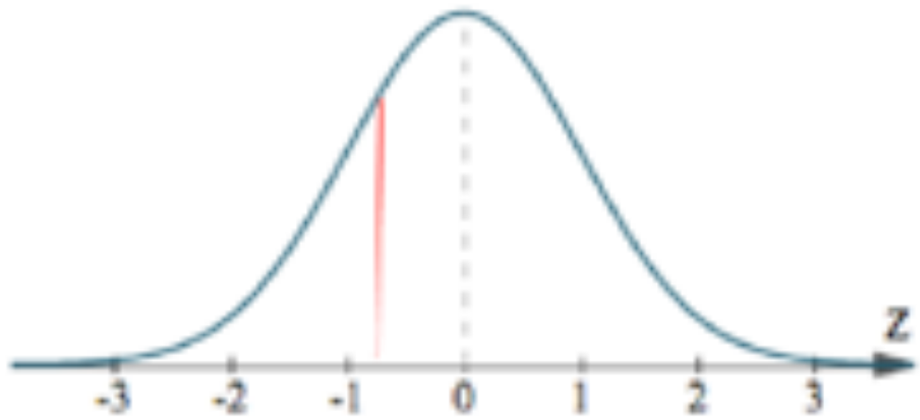
p is moderate

p is small

Continuity Correction

Discrete	Continuous
PMF:	PDF:





Only applies to PDF - why?

Joint Distributions

- Discrete case:

$$p_{x,y}(a,b) = P(X = a, Y = b)$$

$$P_x(a) = \sum_y P_{x,y}(a,y)$$

- Continuous case:

$$P(a_1 < x \leq a_2, b_1 < y \leq b_2) = \int_{a_1}^{a_2} \int_{b_1}^{b_2} f_{X,Y}(x,y)dydx$$

$$f_X(a) = \int_{-\infty}^{\infty} f_{X,Y}(a,y)dy$$

Joint Distributions

- Discrete case:

$$p_{x,y}(a,b) = P(X = a, Y = b) \qquad P_x(a) = \sum_y P_{x,y}(a,y)$$

- Continuous case:

$$P(a_1 < x \leq a_2, b_1 < y \leq b_2) = \int_{a_1}^{a_2} \int_{b_1}^{b_2} f_{X,Y}(x,y)dydx$$

$$f_X(a) = \int_{-\infty}^{\infty} f_{X,Y}(a,y)dy$$ **This is just marginalization!**

Convolutions

$$X \sim Bin(n_1, p), Y \sim Bin(n_2, p) => X + Y \sim Bin(n_1 + n_2, p)$$

$$X \sim Poi(\lambda_1), Y \sim Poi(\lambda_2) => X + Y \sim Poi(\lambda_1 + \lambda_2)$$

$$X \sim \mathrm{N}(\mu_1, \sigma_1^2), Y \sim \mathrm{N}(\mu_2, \sigma_2^2) => X + Y \sim \mathrm{N}(\mu_1 + \mu_2, \sigma_1^2 + \sigma_2^2)$$

$$f_{X+Y}(a) = \int_{y=-\infty}^{\infty} f_X(a-y) f_Y(y) dy$$ **(if X, Y are indep.)**

Deriving Convolution

Discrete Case:

$$P(A + B = z) = \sum_{b} P(A + B = z \,|\, B = b) P(B = b)$$

Expanded Law of Total Probability!

Deriving Convolution

Discrete Case:

$$P(A + B = z) = \sum_{b} P(A + B = z \,|\, B = b) P(B = b)$$

$$P(A + B = z) = \sum_{b} P(A = b - z \,|\, B = b) P(B = b)$$

Deriving Convolution

Discrete Case:

$$P(A + B = z) = \sum_b P(A + B = z \,|\, B = b) P(B = b)$$

$$P(A + B = z) = \sum_b P(A = b - z \,|\, B = b) P(B = b)$$

If A and B are independent:

$$P(A + B = z) = \sum_b P(A = b - z) P(B = b)$$

Deriving Convolution

Discrete Case:

$$P(A+B=z)=\sum_{b} P(A+B=z|B=b)P(B=b)$$

$$P(A+B=z)=\sum_{b} P(A=b-z|B=b)P(B=b)$$

If A and B are independent:

$$P(A+B=z)=\sum_{b} P(A=b-z)P(B=b)$$

Continuous Case:

$$f(A+B=z)=\int_{b} f(A=b-z|B=b)f(B=b)db$$

Deriving Convolution

Discrete Case:

$$P(A+B=z)=\sum_b P(A+B=z|B=b)P(B=b)$$

$$P(A+B=z)=\sum_b P(A=b-z|B=b)P(B=b)$$

If A and B are independent:

$$P(A+B=z)=\sum_b P(A=b-z)P(B=b)$$

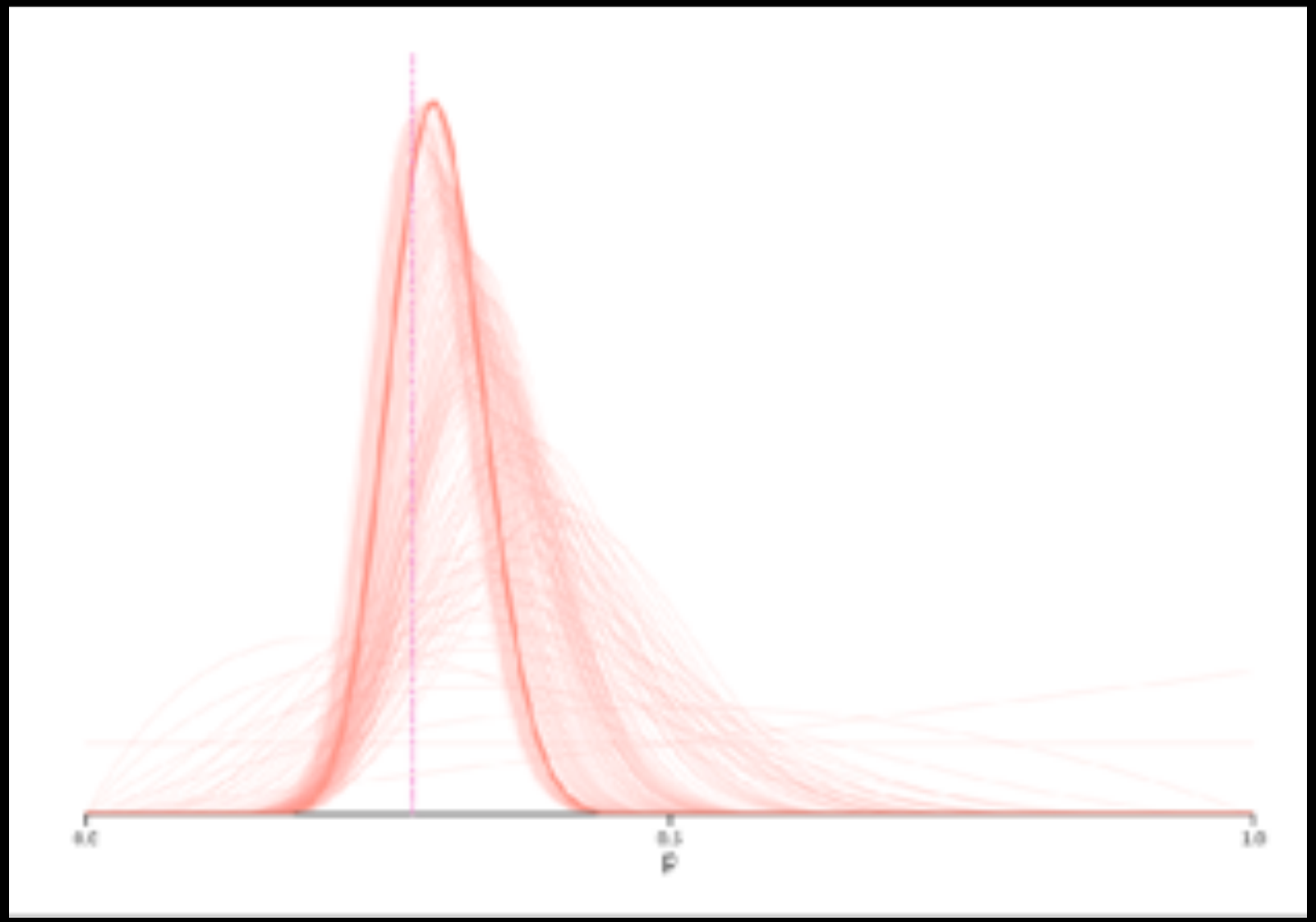
Continuous Case:

If A and B are independent

$$f(A+B=z)=\int_b f(A=b-z|\cancel{B=b})f(B=b)db$$

Beta

Our first look at the concept of estimating parameters by observing data!



https://seeing-theory.brown.edu/bayesian-inference/index.html#section3

CS 109

topics

machine learning

general inference

sampling, making conclusions from data

bootstrapping

CLT

unbiased estimators

random variables / distributions

core probability fundamentals

Sampling From Populations

Challenge: we want to know what the distribution of happiness looks like in Bhutan, but we have limited time and resources and the landscape looks like this:

climb every mountain!

Sampling From Populations

violating data collection norms so that it's unreasonable to assume that a sample is representative of the population

only asking people in Thimphu, e.g.

using statistical methods to draw reasonable conclusions about the population based on data from a random sample

understanding how your results might differ if you sample from the same population multiple times

being an omniscient entity who knows the true population distribution

Taking One Sample

Pick a random sample

if sample size is large enough and sampling methodology is good enough, you can consider it representative of the population!

Take measurements

we have handy equations for the sample mean and sample variance, which are **unbiased estimators** of the population mean and variance

Report estimate uncertainty

we can use the data from one sample to report our uncertainty about how our estimate of the mean might compare to the true mean (error bars!)

$$\bar{X} = \sum_{i=1}^{n} \frac{X_i}{n} \qquad S^2 = \sum_{i=1}^{n} \frac{(X_i - \bar{X})^2}{n-1}$$

makes the estimate unbiased

$$Std(\bar{X}) \approx \sqrt{\left(\frac{S^2}{n}\right)}$$

Sample vs True

True mean and variance are properties of the underlying distribution. They are platonic ideals, completely unattainable!

$$\mu = E[X] \qquad \sigma^2 = Var(X)$$

Sample mean and variance are unbiased estimates of true mean and variance based on a single IID sample.

$$\bar{X} = \sum_{i=1}^{n} \frac{X_i}{n} \qquad S^2 = \sum_{i=1}^{n} \frac{(X_i - \bar{X})^2}{n-1}$$

Variance of sample mean tells us our uncertainty about how good of an estimate sample mean is.

$$Var(\bar{X}) = Var\left(\sum_{i=1}^{n} \frac{X_i}{n}\right) = \sum_{i=1}^{n} var\left(\frac{X_i}{n}\right) = \sum_{i=1}^{n} \frac{Var(X)}{n^2} \approx \frac{S^2}{n}$$

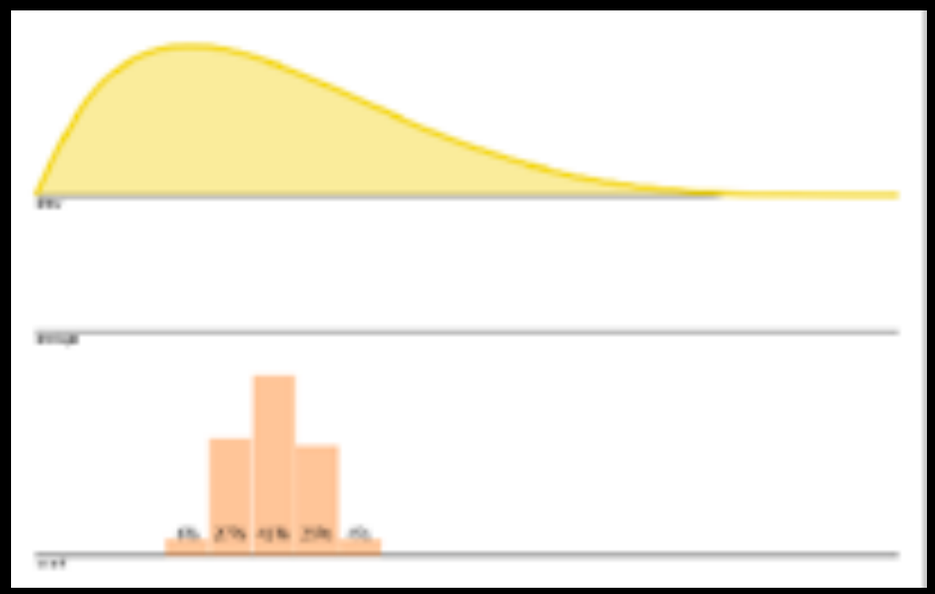
Taking Many Samples

Unbiased Estimators

the expected value of the estimated statistic is the value of the true population statistic (if many samples were to be taken)

Central Limit Theorem

if you sample from the same population a bunch of times, the mean and sum of all your samples (or any IID RVs) will be normally distributed no matter what your distribution looks like!



https://seeing-theory.brown.edu/probability-distributions/index.html#section3

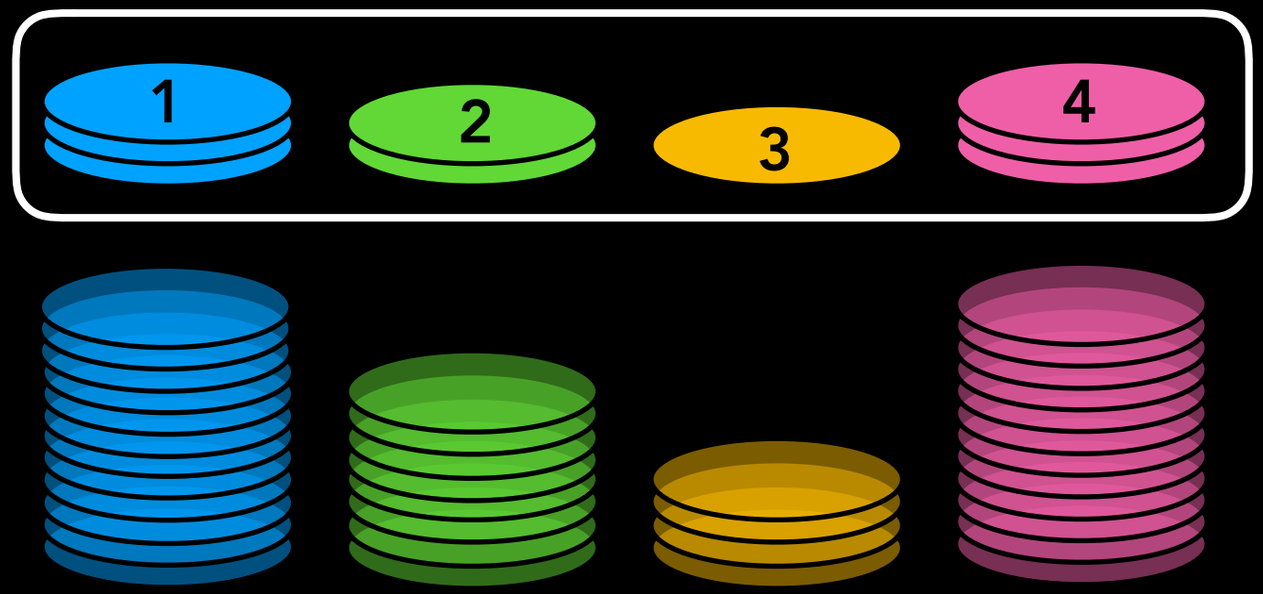
Bootstrapping: Simulating Many Samples From One

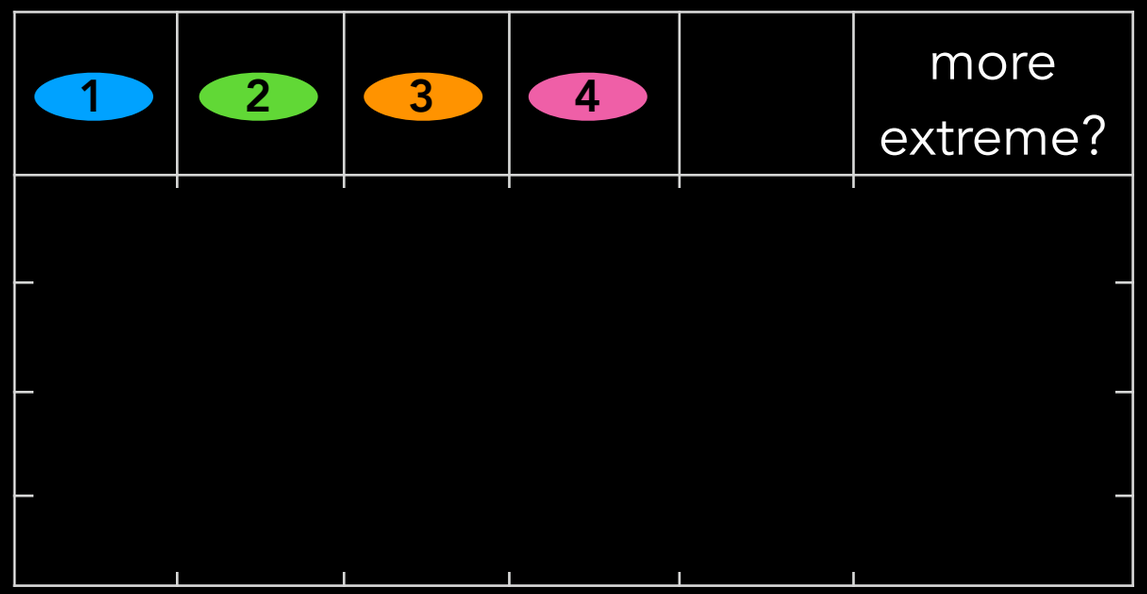
challenge

we want to find the probability that the data results we saw were due to chance, but we only have one sample of data

insight

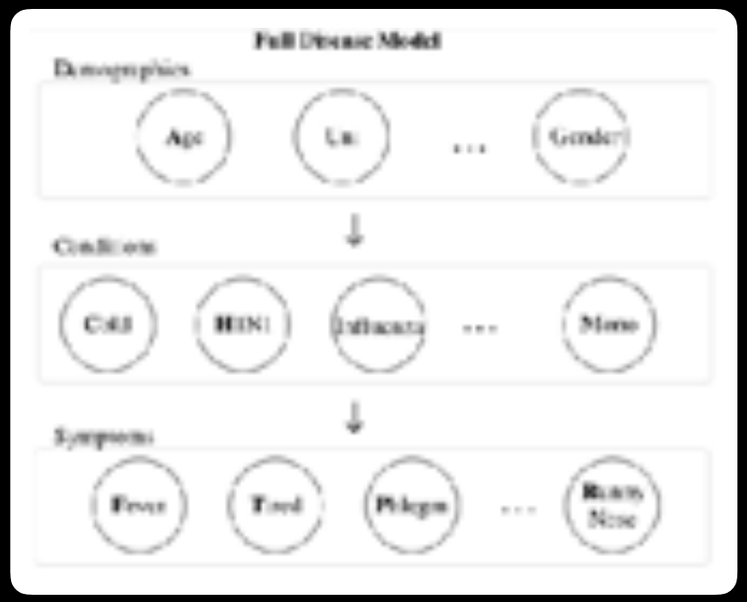
since our sample represents our population, we can sample from the data we have and it's as if we had gone out and collected more





We sample **with replacement** from our data and calculate our statistic of interest each time, ending up with many estimates for our statistic of interest. We can even use this data to assess whether our observations are due to chance based on our p-value of choice.

General Inference: Sampling from a Bayesian Network to Find Joint Probability



Joint Sampling

generate many “particles” by tracing through the network, generating values for children based on their parents

Calculate Conditional Probability

we can calculate any conditional probability of specific variable assignments by simply counting the particles that match what we’re looking for

$$P(\mathbf{X} = \mathbf{a} \mid \mathbf{Y} = \mathbf{b}) = \frac{N(\mathbf{X} = \mathbf{a}, \mathbf{Y} = \mathbf{b})}{N(\mathbf{Y} = \mathbf{b})}$$

CS 109

topics

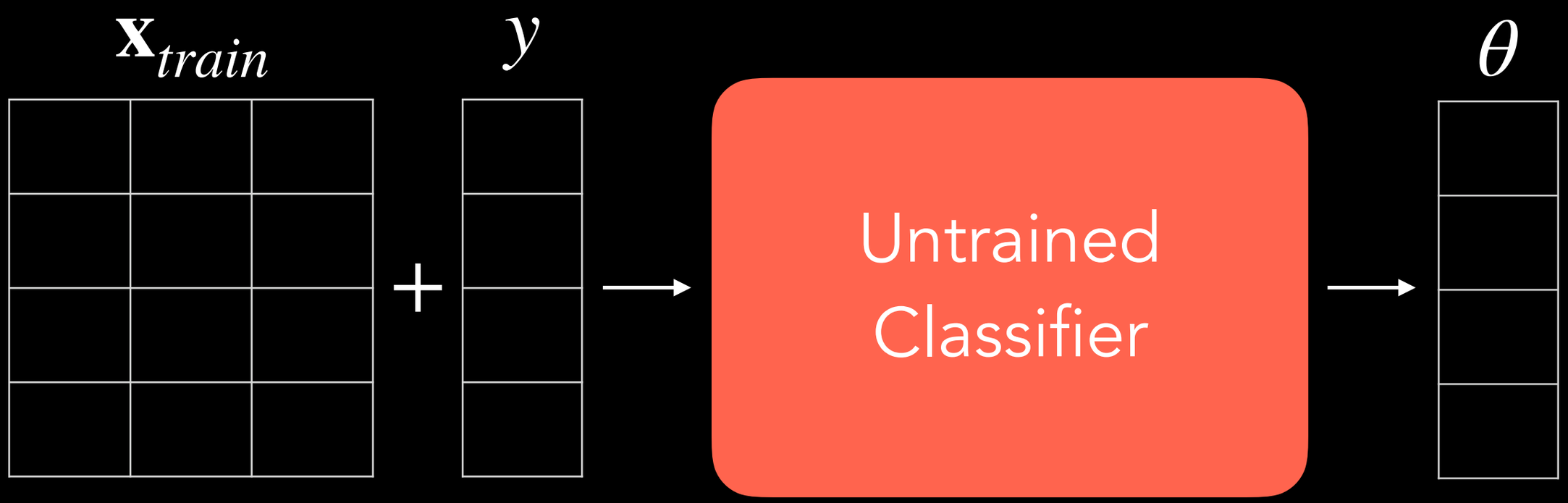
machine learning | parameter estimation | classifiers | deep learning

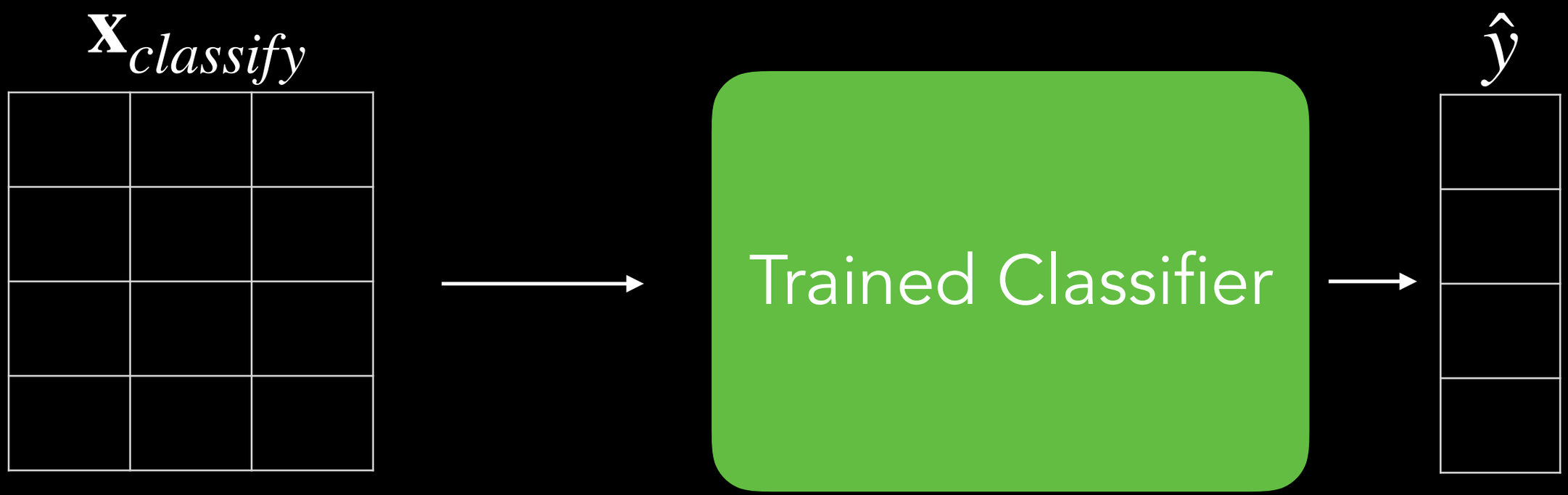
sampling, making conclusions from data

random variables / distributions

core probability fundamentals

Classifiers





Parameter Estimation

Maximum Likelihood Estimation

1. Find likelihood: product of likelihoods of each sample/ datapoint given theta
2. Take the log of that expression
3. Take the derivative of that with respect to the parameters
4. Either set to 0 and solve
   (if it's a simple case with closed form solution)
   or plug into gradient ascent to find a value for theta that maximizes your likelihood

Maximum A Posteriori

1. Find likelihood: product of likelihoods of each sample/ datapoint given theta, times your prior likelihood of that theta
2. - 4. same as above

MLE vs. MAP

MLE:

$$\underset{\theta}{argmax}(P(data\,|\,\theta)) = \underset{\theta}{argmax}\left(\prod_{i=1}^{n} P(x^{(i)}\,|\,\theta)\right) = \underset{\theta}{argmax}\left(\sum_{i=1}^{n} \log P(x^{(i)}\,|\,\theta)\right)$$

MLE vs. MAP

MLE:

$$\underset{\theta}{argmax}(P(data \mid \theta)) = \underset{\theta}{argmax}\left(\prod_{i=1}^{n} P(x^{(i)} \mid \theta)\right) = \underset{\theta}{argmax}\left(\sum_{i=1}^{n} \log P(x^{(i)} \mid \theta)\right)$$

MAP:

$$\underset{\theta}{argmax}(P(\theta \mid data)) = \underset{\theta}{argmax}\left(\frac{P(data \mid \theta)P(\theta)}{P(data)}\right)$$

MLE vs. MAP

MLE:

$$\underset{\theta}{argmax}(P(data \mid \theta)) = \underset{\theta}{argmax}\left(\prod_{i=1}^{n} P(x^{(i)} \mid \theta)\right) = \underset{\theta}{argmax}\left(\sum_{i=1}^{n} \log P(x^{(i)} \mid \theta)\right)$$

MAP:

$$\underset{\theta}{argmax}(P(\theta \mid data)) = \underset{\theta}{argmax}\left(\frac{P(data \mid \theta)P(\theta)}{\cancel{P(data)}}\right)$$

MLE vs. MAP

MLE:

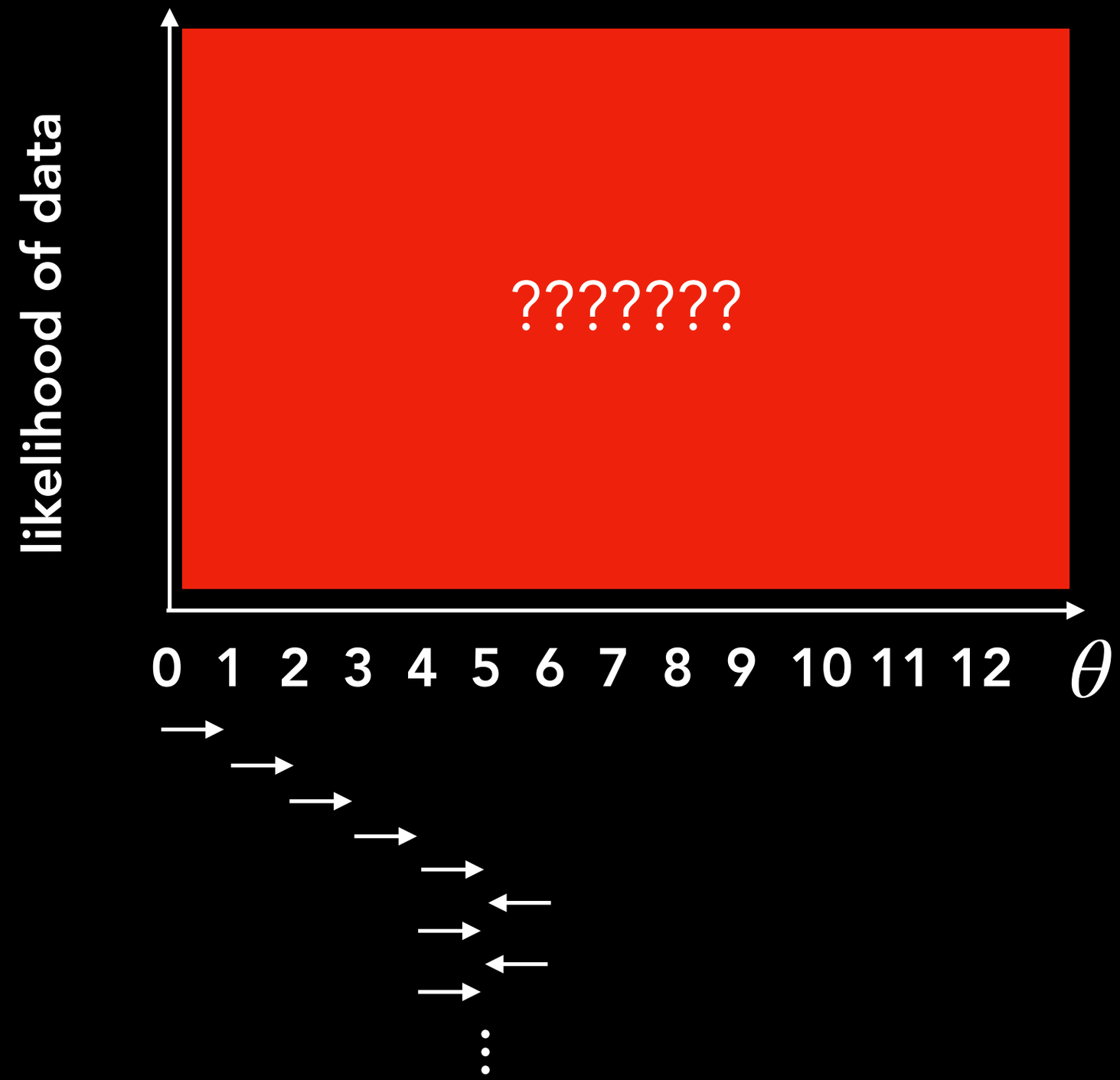
$$\underset{\theta}{argmax}(P(data\,|\,\theta)) = \underset{\theta}{argmax}\left(\prod_{i=1}^{n} P(x^{(i)}\,|\,\theta)\right) = \underset{\theta}{argmax}\left(\sum_{i=1}^{n} \log P(x^{(i)}\,|\,\theta)\right)$$

MAP:

$$\underset{\theta}{argmax}(P(\theta\,|\,data)) = \underset{\theta}{argmax}\left(\frac{P(data\,|\,\theta)P(\theta)}{P(data)}\right) = \underset{\theta}{argmax}\left(P(\theta)\prod_{i=1}^{n} P(x^{(i)}\,|\,\theta)\right)$$

$$= \underset{\theta}{argmax}\left(\log P(\theta) + \sum_{i=1}^{n} \log P(x^{(i)}\,|\,\theta)\right)$$

Gradient Ascent



step size

$$\eta = 1$$

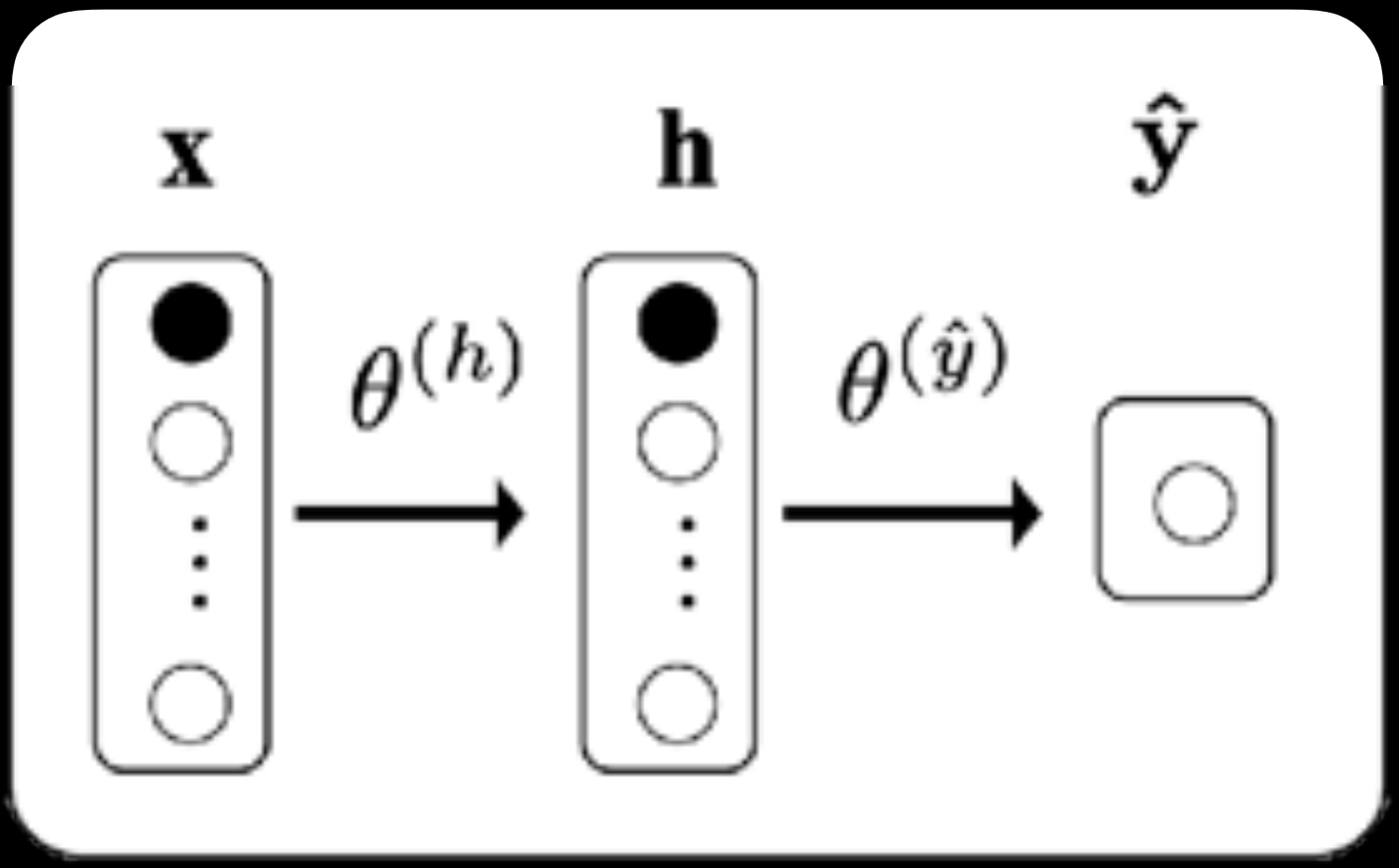
step direction

$$= \textbf{sign}\left[\frac{\partial\ \textbf{prob}}{\partial\theta}\right]$$

Classifier Algorithms

Naïve Bayes	Algorithm	Logistic Regression
All features in **x** are conditionally independent given classification	Assumption	Sigmoid gives us the probability of class 1
Whether y = 0 or y = 1 maximizes the probability of our data	What are we optimizing/ figuring out?	The value(s) for θ such that the probability of our data is maximized
Learn (from data) estimates for $\hat{P}(Y = y), \hat{P}(X_i = x_i \mid Y = y)$: $\hat{P}(x_i \mid y) = \frac{(\textbf{ex. where } X_i = x_i \text{ and } Y = y) + 1}{(\textbf{ex. where } Y = y) + 2}$ $\hat{P}(Y = y) = \frac{\textbf{ex. where } Y = y}{\textbf{total examples}}$	How do we do that mathematically?	Probability of 1 datapoint $P(y \mid \mathbf{x}) = \sigma(\theta^T \mathbf{x})^y \cdot [1 - \sigma(\theta^T \mathbf{x})]^{1-y}$ Use data & gradient ascent to improve thetas $LL(\theta) = \sum_{i=1}^{n} y^{(i)} \textbf{ log } \sigma(\theta^T \mathbf{x}^{(i)}) + (1 - y^{(i)}) \textbf{ log } \left[1 - \sigma(\theta^T \mathbf{x}^{(i)})\right] x_j$ $\frac{\partial LL(\theta)}{\partial \theta_j} = \sum_{i=1}^{n} \left[y^{(i)} - \sigma(\theta^T \mathbf{x}^{(i)})\right] x_j^{(i)}$

Neural Networks



1. Make deep learning assumption: $P(Y = y \mid \mathbf{X} = \mathbf{x}) = (\hat{y})^{y}(1 - \hat{y})^{1-y}$
2. Calculate log likelihood for all data: $LL(\theta) = \sum_{i=0}^{n} y^{(i)}(\textbf{log}\ \hat{y}^{(i)}) + (1 - \hat{y}^{(i)})\ \textbf{log}\ [1 - \hat{y}^{(i)}]$
3. Find partial derivative of LL with respect to each theta:

   use the chain rule!

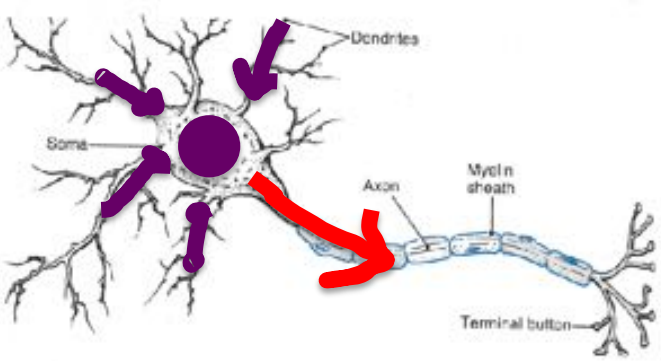
$$\frac{\partial LL(\theta)}{\partial \theta_j^{(\hat{y})}} = \frac{\partial LL(\theta)}{\partial \hat{y}} \cdot \frac{\partial \hat{y}}{\partial \theta_j^{(\hat{y})}} \qquad \frac{\partial LL(\theta)}{\partial \theta_{i,j}^{(h)}} = \frac{\partial LL(\theta)}{\partial \hat{y}} \cdot \frac{\partial \hat{y}}{\partial \mathbf{h}_j} \cdot \frac{\partial \mathbf{h}_j}{\partial \theta_{i,j}^{(h)}}$$

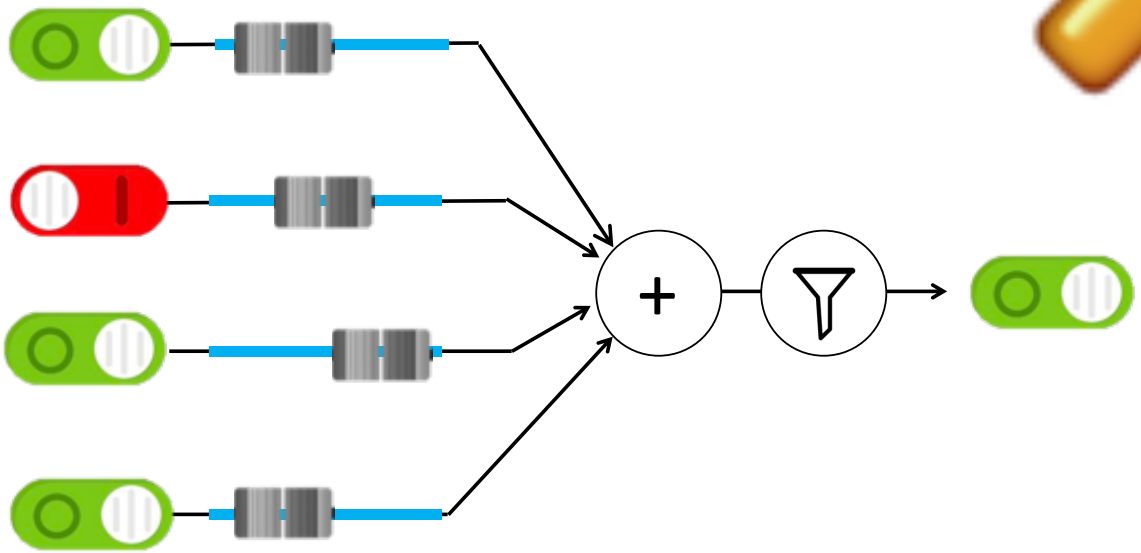
Good luck on the final!

Stretch!

Review

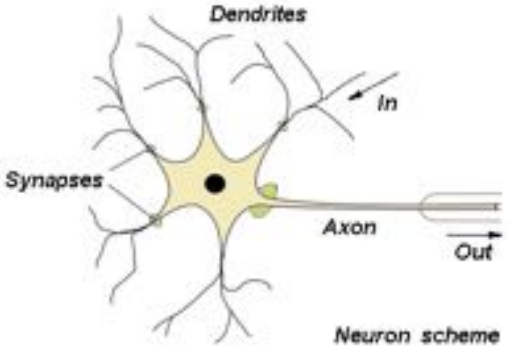
Artificial Neurons

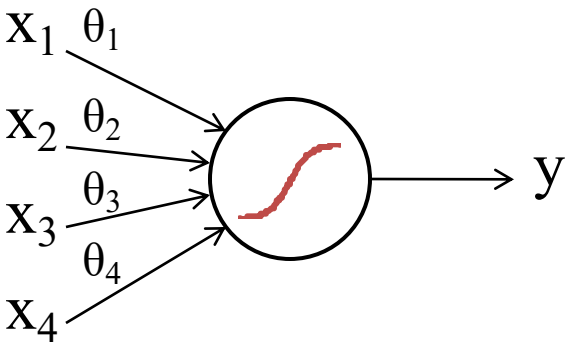




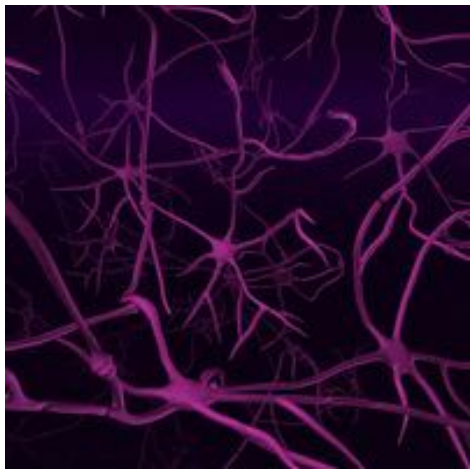
Biological Basis for Neural Networks

- A neuron

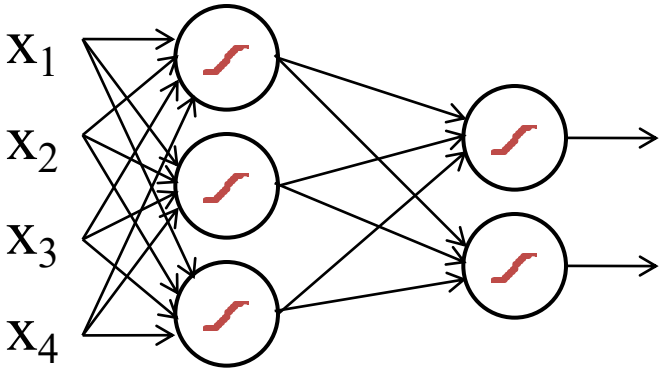




- Your brain



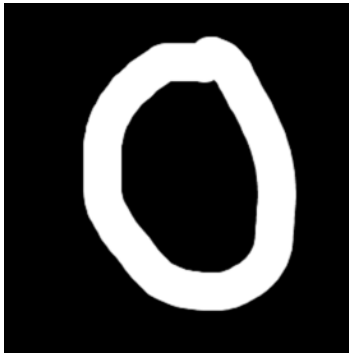
Actually, it's probably someone else's brain

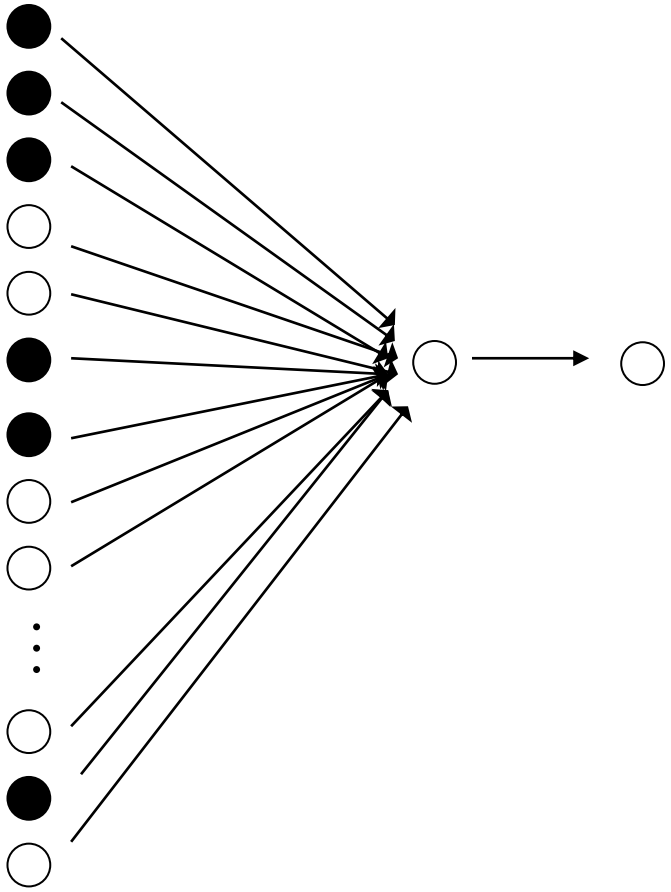


(aka Neural Networks)

Deep learning is (at its core) many logistic regression pieces stacked on top of each other.

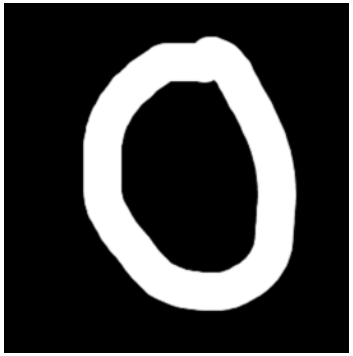
Logistic Regression

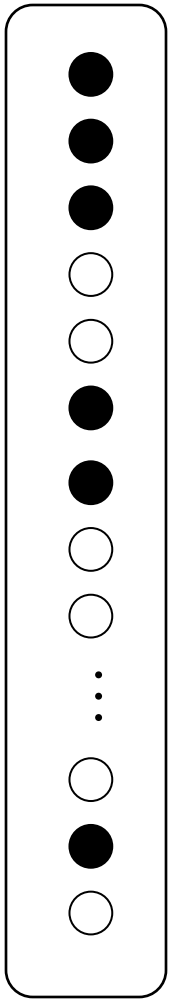




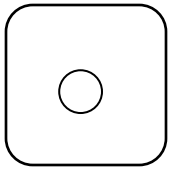
This means it predicts a 0

Logistic Regression



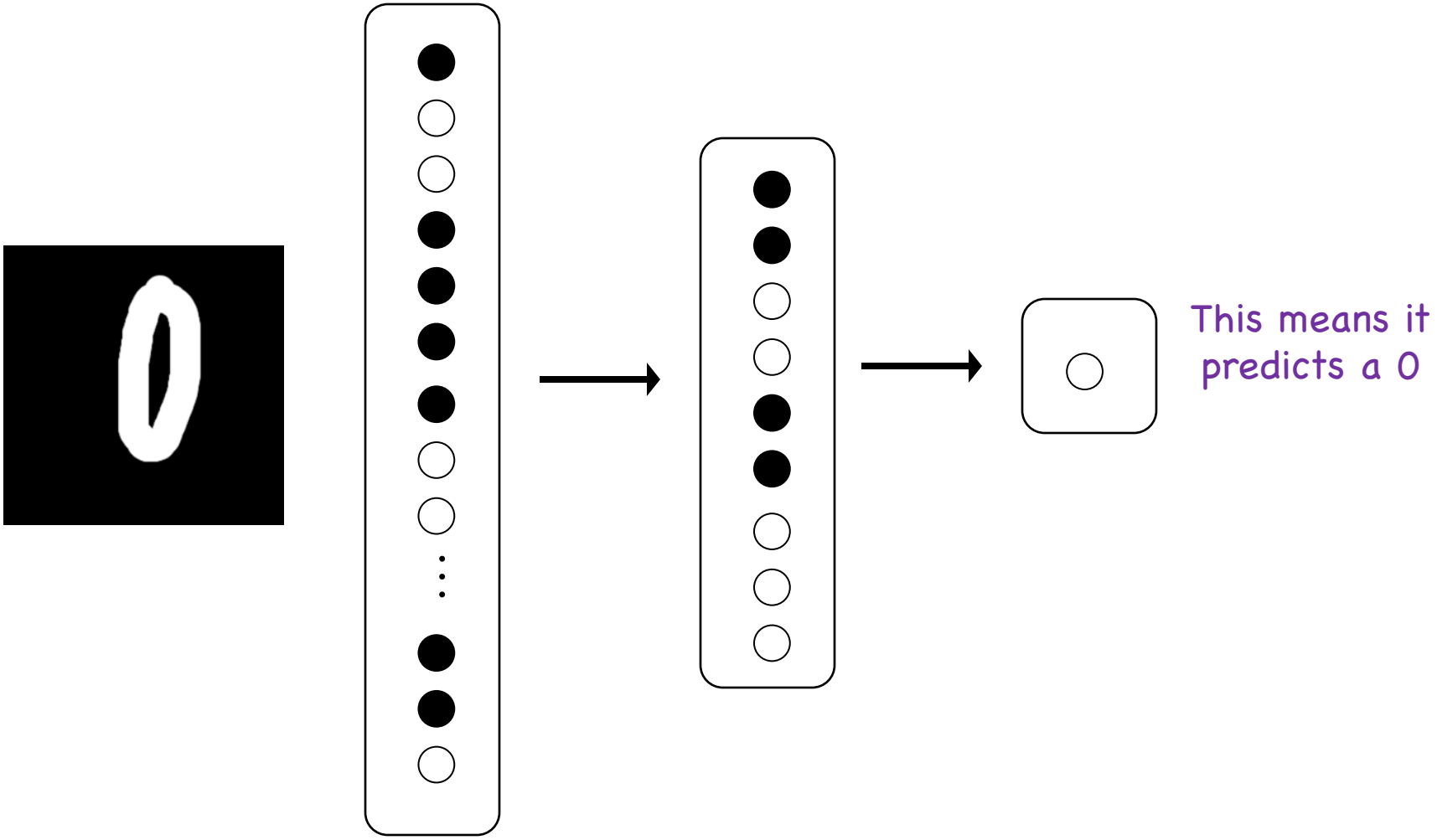


Indicates logistic regression connection

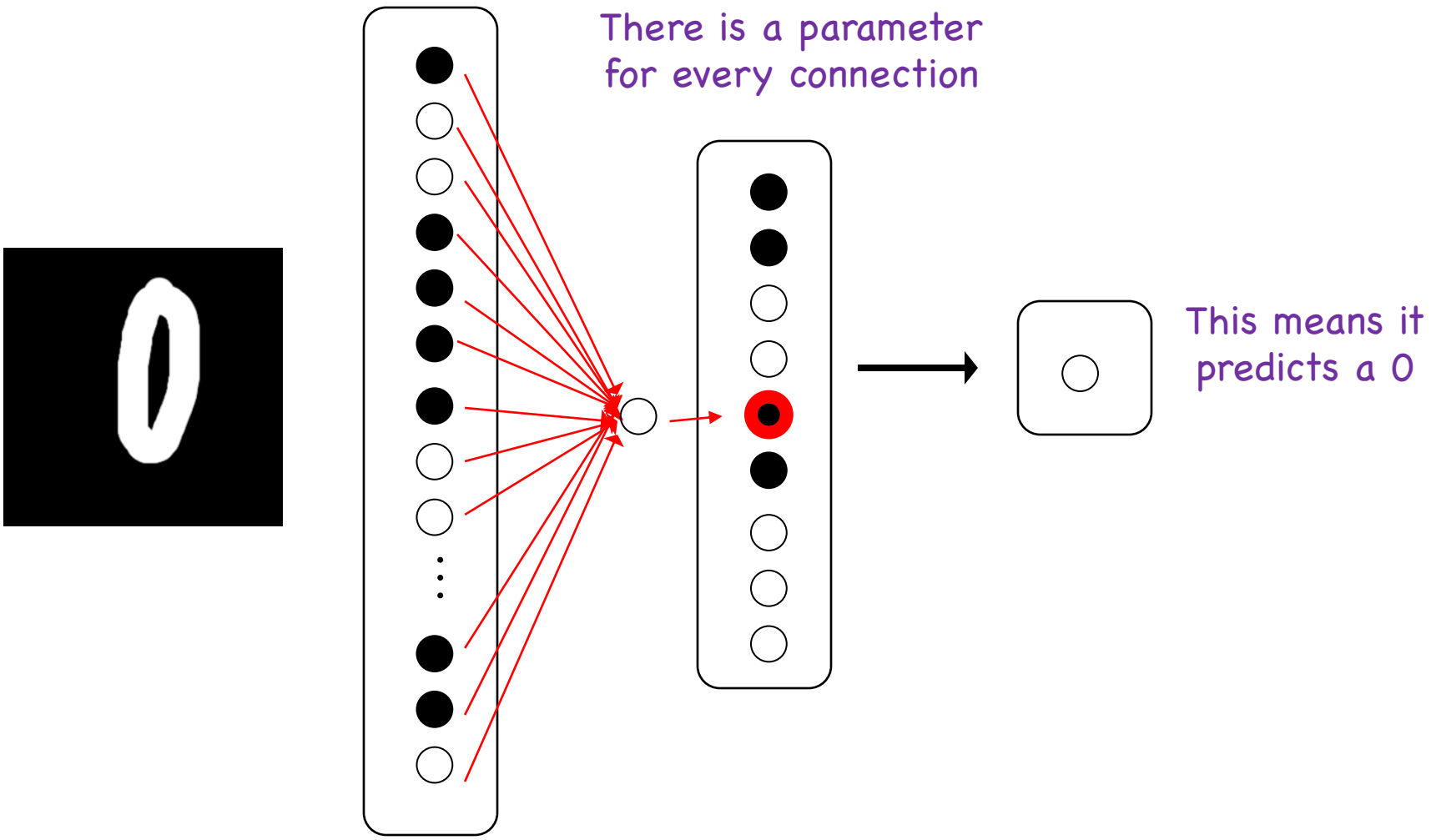


This means it predicts a 0

We Can Put Neurons Together

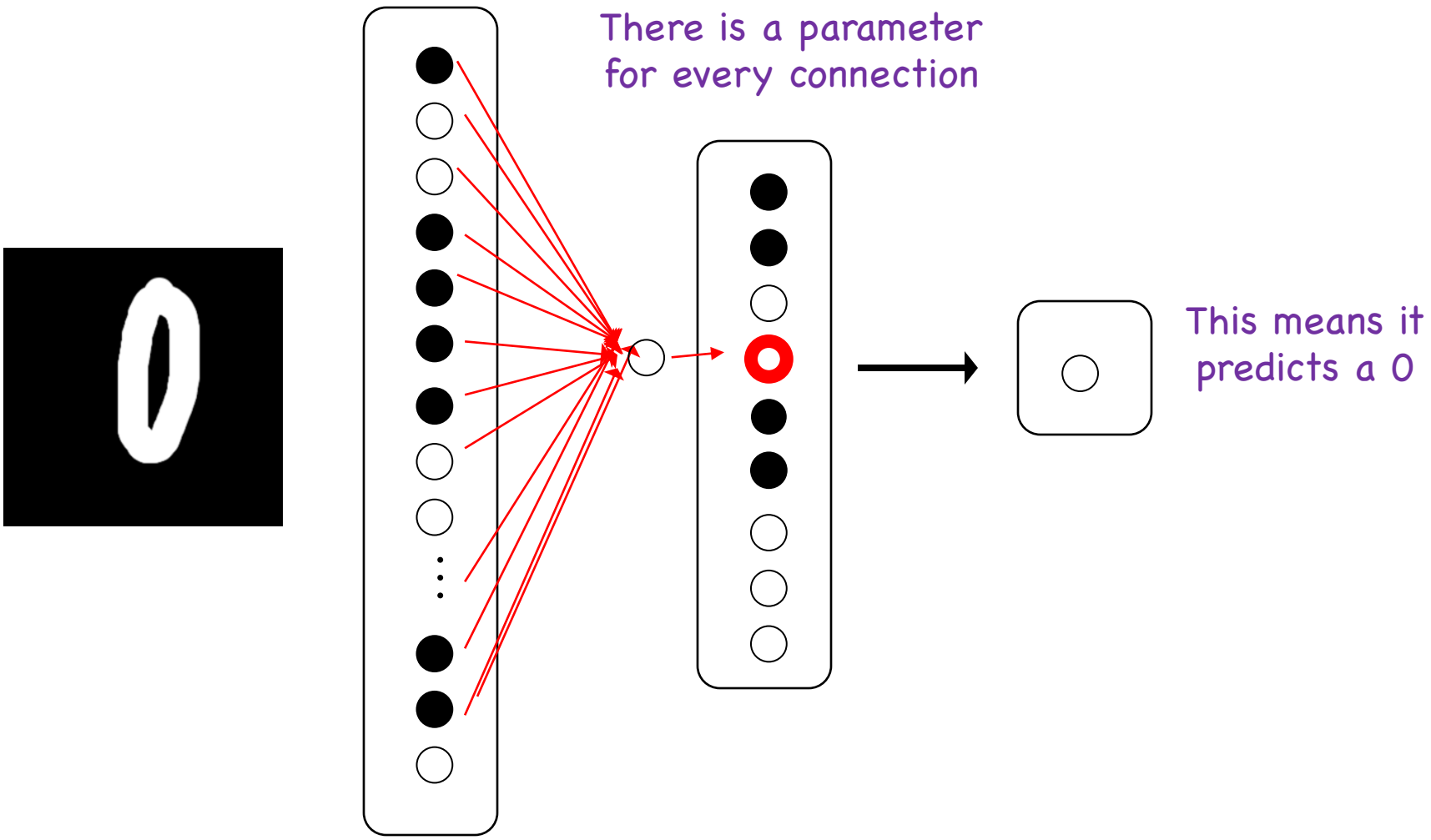


We Can Put Neurons Together



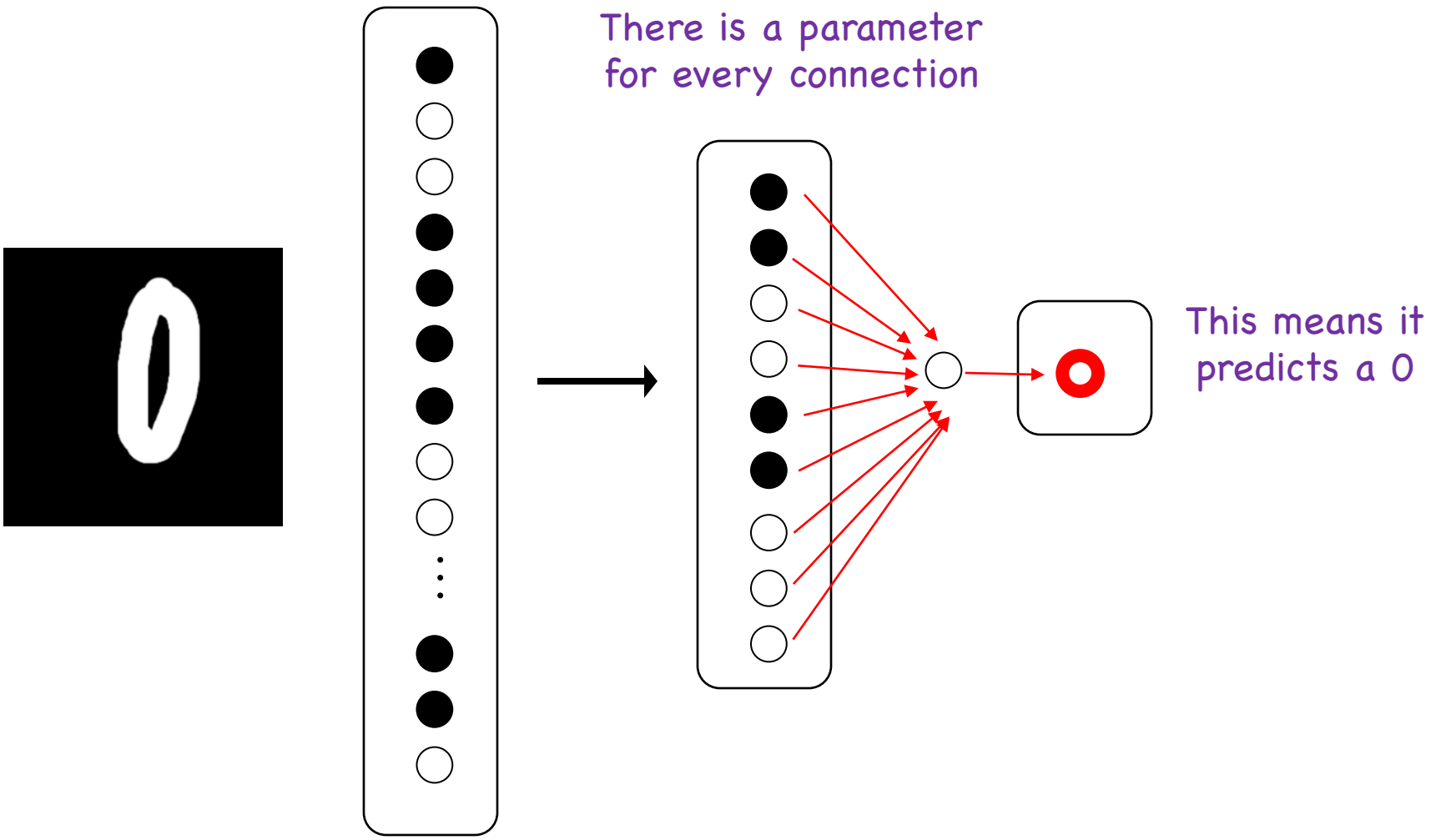
Look at a single “hidden” neuron

We Can Put Neurons Together



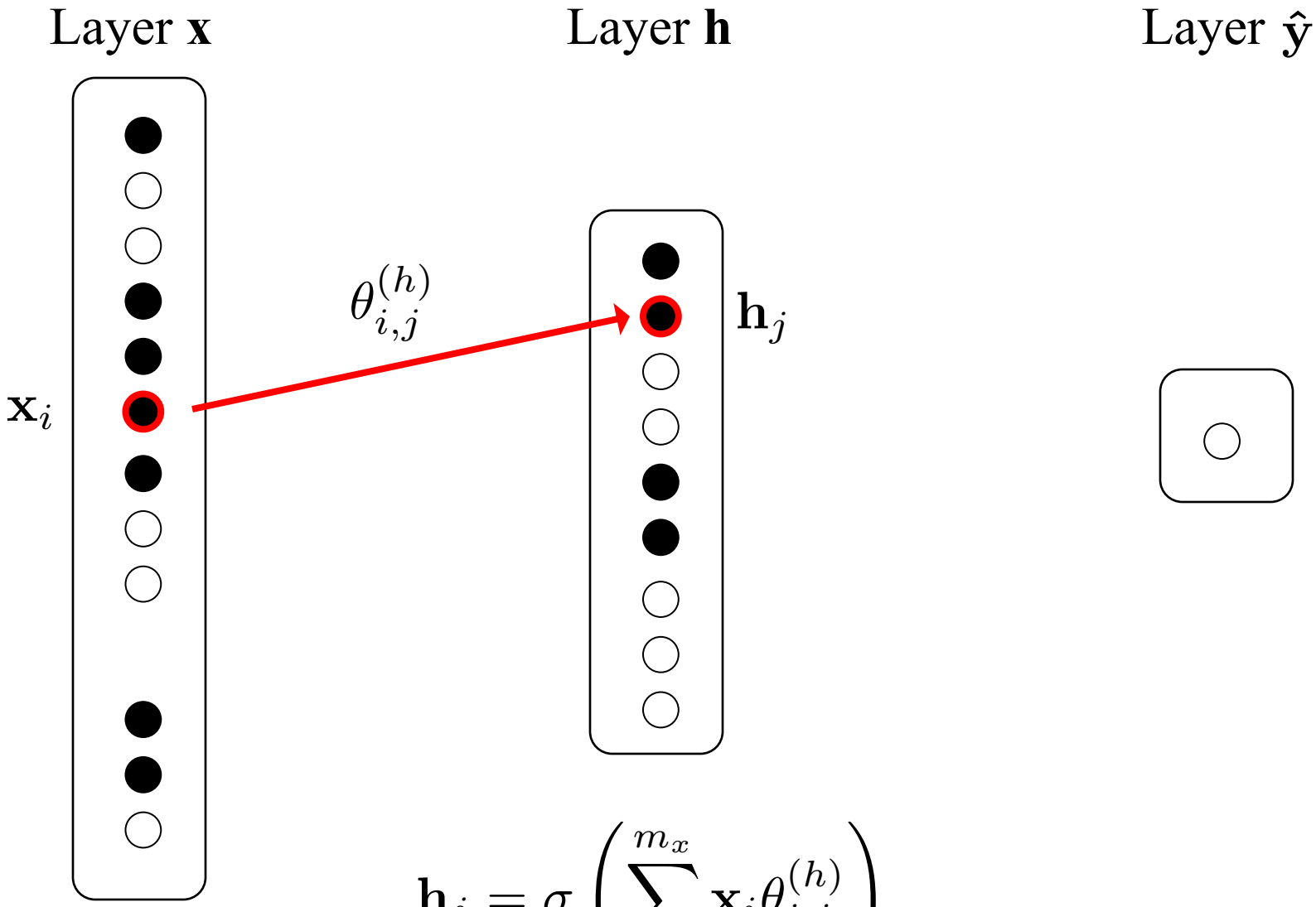
Look at another “hidden” neuron

We Can Put Neurons Together



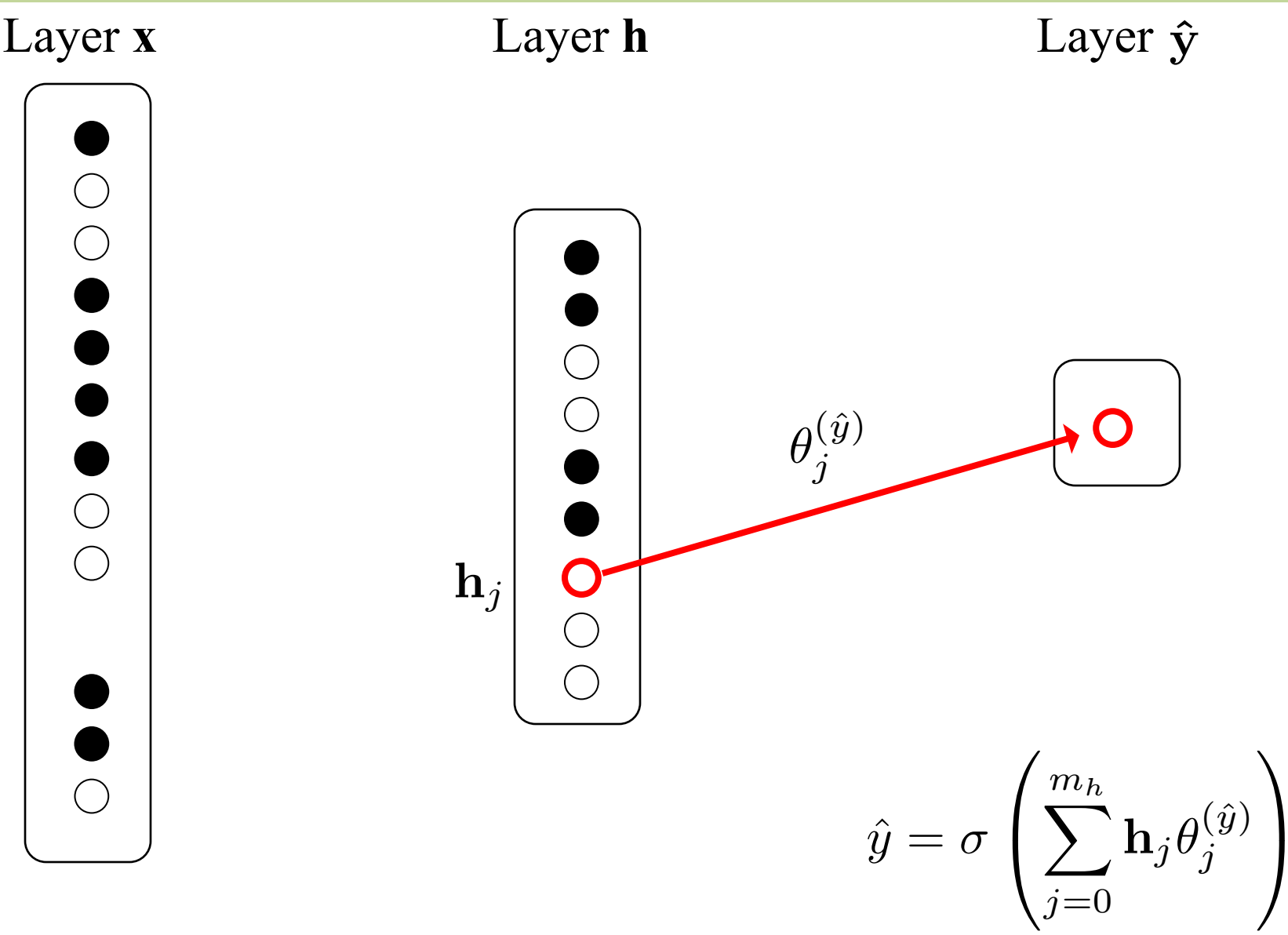
Look at another neuron

New Notation



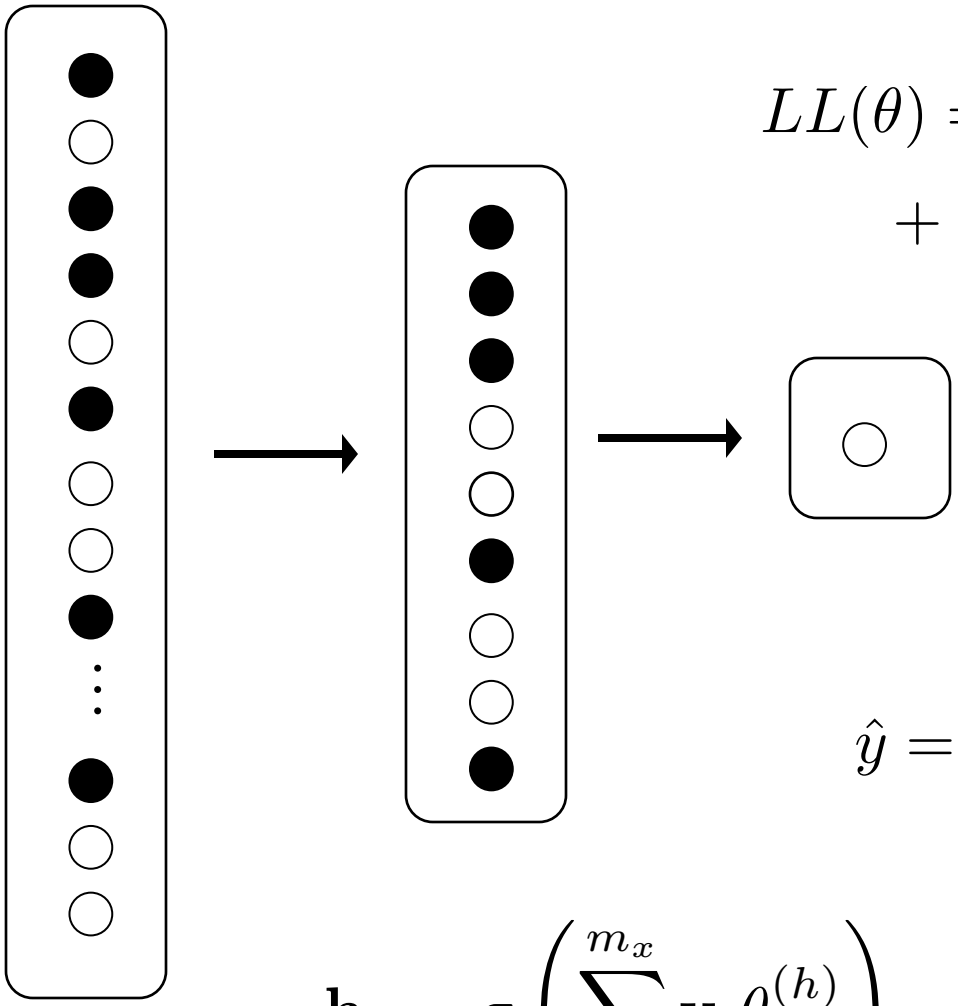
$$\mathbf{h}_j = \sigma\left(\sum_{i=0}^{m_x} \mathbf{x}_i \theta_{i,j}^{(h)}\right)$$

New Notation



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

Forward Pass



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

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

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

Same Assumption, Same LL

$$P(Y = 1|X = \mathbf{x}) = \hat{y} \qquad \hat{y} = \sigma(\theta^T \mathbf{x})$$

For one datum

$$P(Y = y|\mathbf{X} = \mathbf{x}) = (\hat{y})^y (1 - \hat{y})^{1-y} \qquad Y \sim \text{Bern}(\hat{y})$$

For IID data

$$L(\theta) = \prod_{i=1}^{n} P(Y = y^{(i)}|X = \mathbf{x}^{(i)})$$

$$= \prod_{i=1}^{n} (\hat{y}^{(i)})^{y^{(i)}} \cdot \left[1 - (\hat{y}^{(i)})\right]^{(1-y^{(i)})}$$

Take the log

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

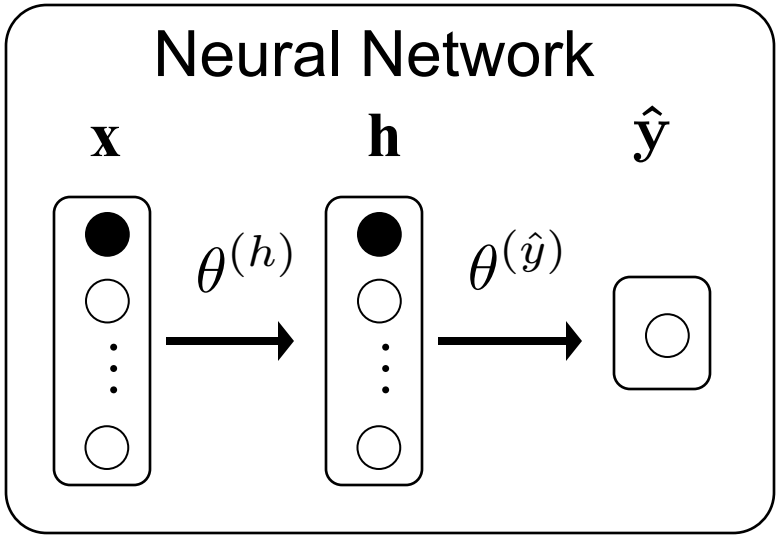
Derivative Goals

Loss with respect to output layer params

$$\frac{\partial LL(\theta)}{\partial \theta_i^{(\hat{y})}}$$

Loss with respect to hidden layer params

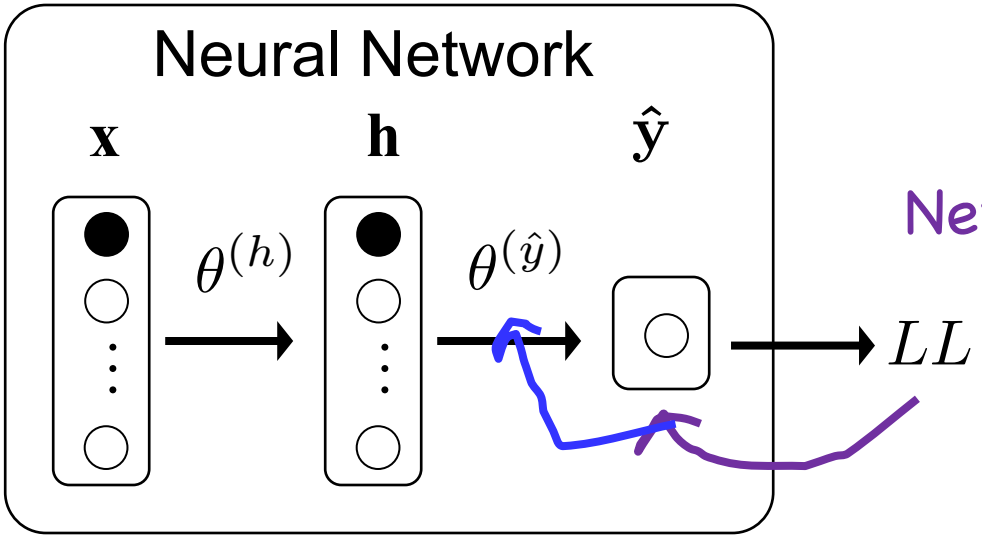
$$\frac{\partial LL(\theta)}{\partial \theta_{i,j}^{(h)}}$$



Chain Rule Example 1

$$\frac{\partial LL(\theta)}{\partial \theta_i^{(\hat{y})}}$$

Goal



Network

$$\frac{\partial LL(\theta)}{\partial \theta_i^{(\hat{y})}} = \frac{\partial LL}{\partial \hat{y}} \cdot \frac{\partial \hat{y}}{\partial \theta_i^{(\hat{y})}}$$

Decomposition

Make it Simple

$$\frac{\partial LL(\theta)}{\partial \theta_i^{(\hat{y})}} = \text{  } \cdot \text{  }$$

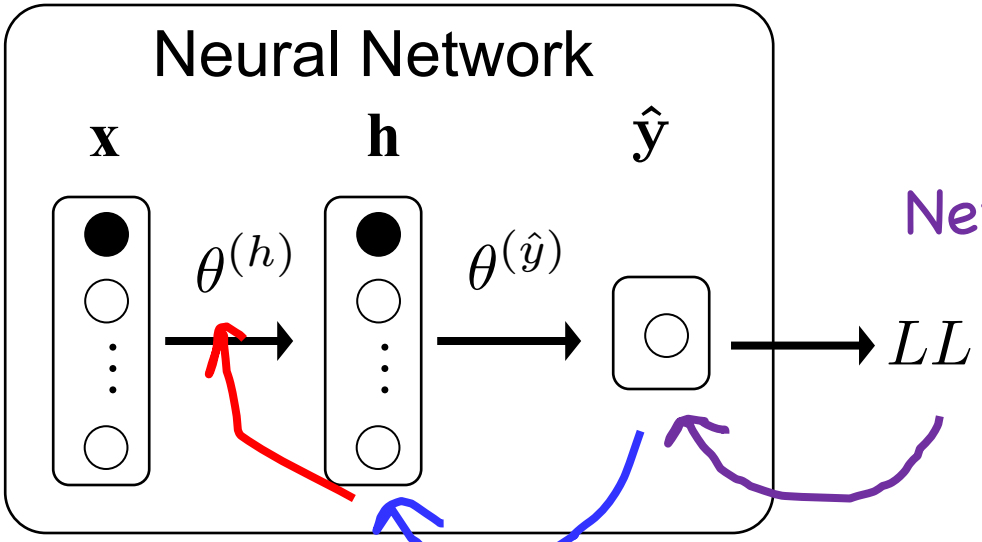
$$\text{  } = \frac{y}{\hat{y}} - \frac{(1-y)}{(1-\hat{y})}$$

$$\text{  } = \hat{y}[1-\hat{y}] \cdot h_i$$

Chain Rule Example 2

$$\frac{\partial LL(\theta)}{\partial \theta_{i,j}^{(h)}}$$

Goal

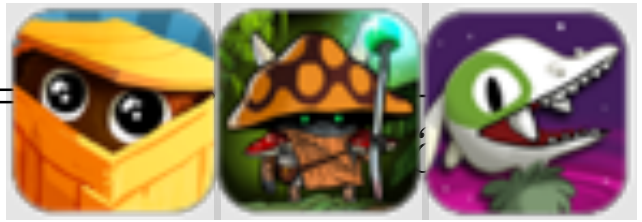


Network

$$\frac{\partial LL(\theta)}{\partial \theta_{i,j}^{(h)}} = \frac{\partial LL}{\partial \hat{y}} \cdot \frac{\partial \hat{y}}{\partial \mathbf{h}_j} \cdot \frac{\partial \mathbf{h}_j}{\partial \theta_{i,j}^{(h)}}$$

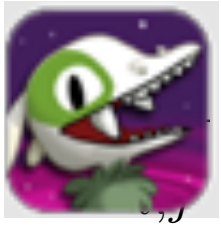
Decomposition

Make it Simple

$$\frac{\partial LL(\theta)}{\partial \theta_{i,j}^{(h)}} = \text{  }$$

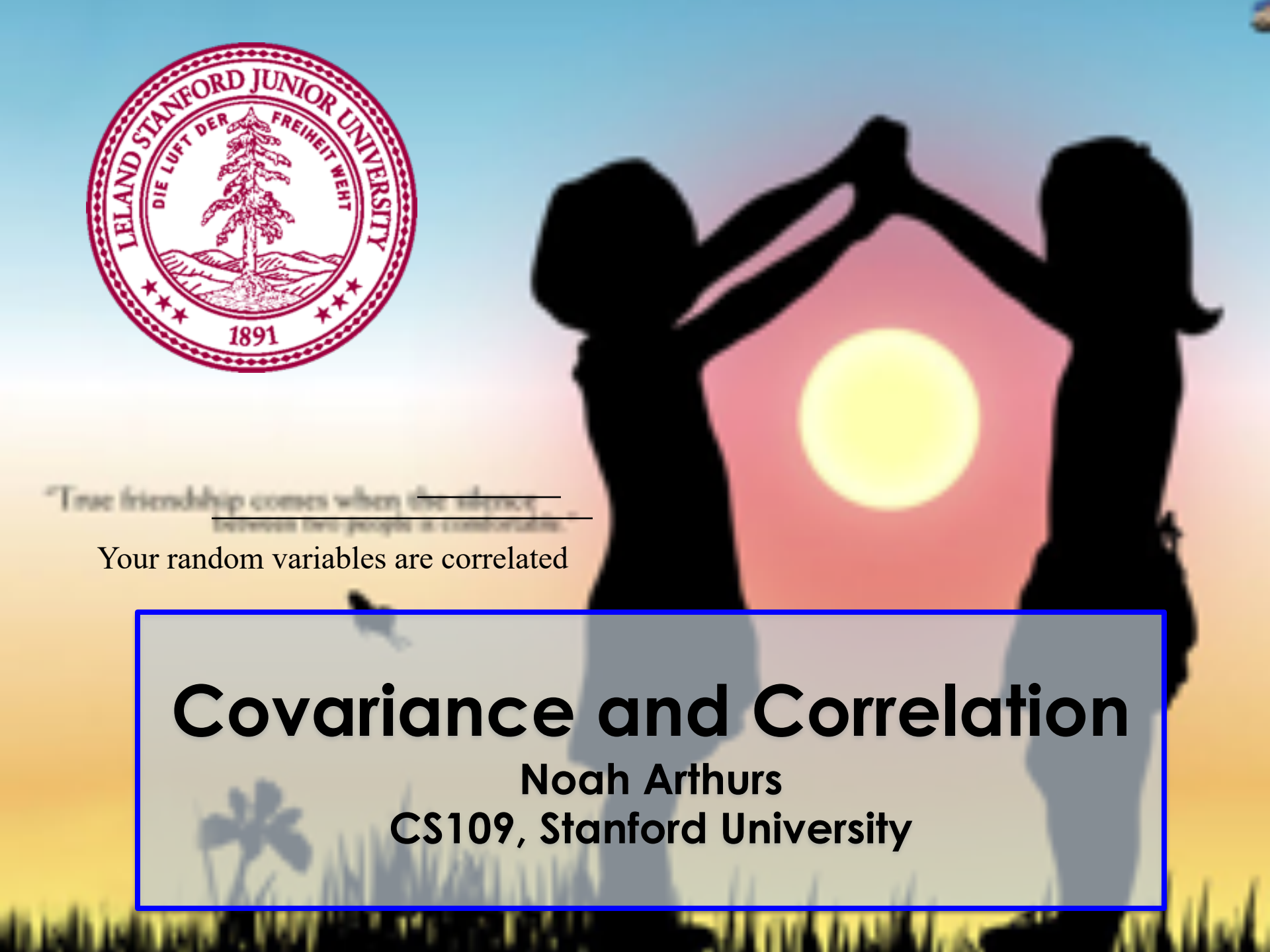
$$= \frac{y}{\hat{y}} - \frac{(1-y)}{(1-\hat{y})}$$

$$= \hat{y}[1-\hat{y}]\theta_j^{(\hat{y})}$$



$$= \mathbf{h}_j[1-\mathbf{h}_j]\mathbf{x}_j$$

End Review



Your random variables are correlated

Covariance and Correlation

Noah Arthurs
CS109, Stanford University

Recall our Ebola Bats

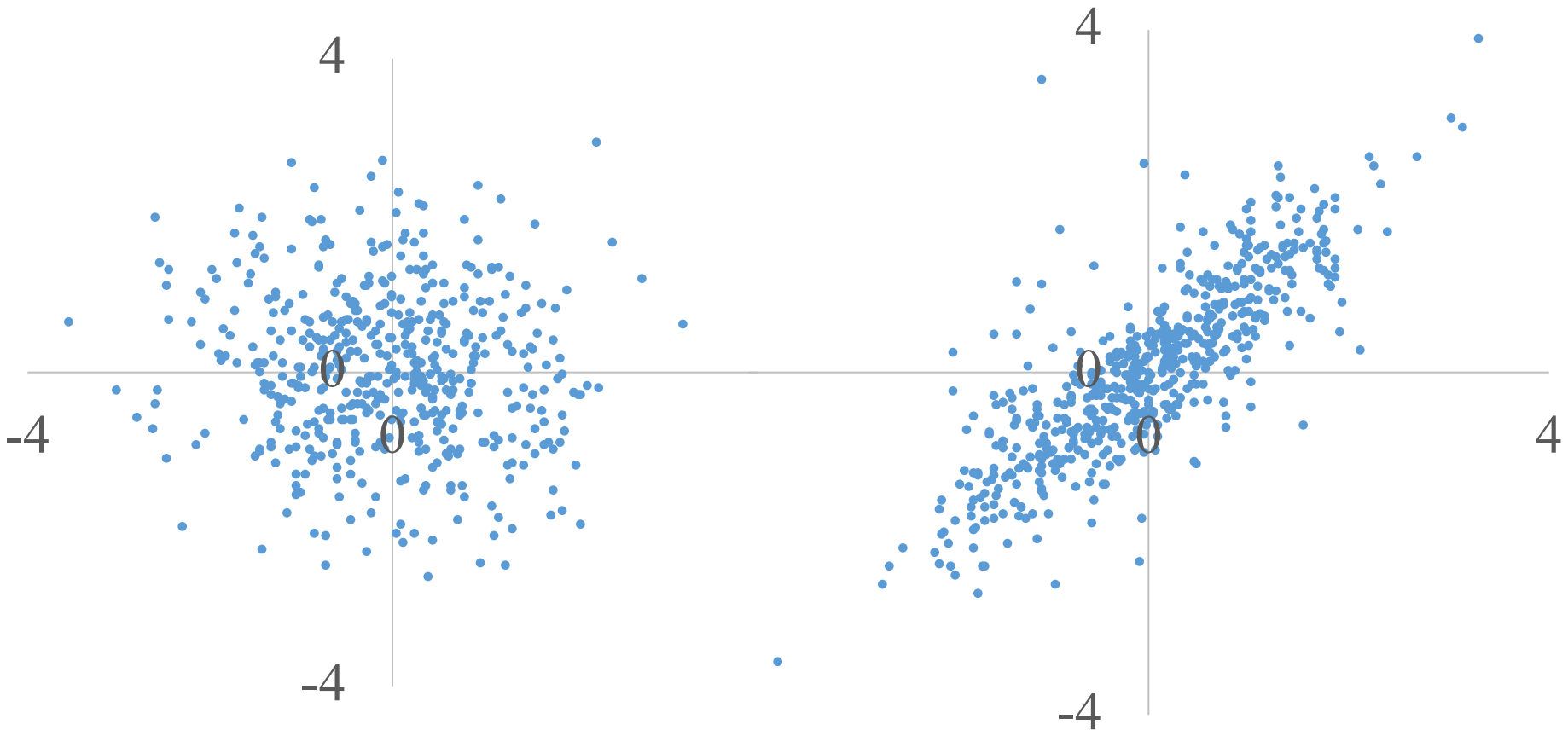
Bat Data

Gene1	Gene2	Gene3	Gene4	Gene5	Trait
TRUE	FALSE	TRUE	TRUE	FALSE	FALSE
FALSE	FALSE	TRUE	TRUE	TRUE	TRUE
TRUE	FALSE	TRUE	FALSE	FALSE	FALSE
TRUE	FALSE	TRUE	TRUE	TRUE	FALSE
FALSE	TRUE	TRUE	TRUE	TRUE	TRUE
FALSE	FALSE	FALSE	TRUE	FALSE	FALSE
TRUE	FALSE	FALSE	TRUE	FALSE	FALSE
TRUE	FALSE	FALSE	TRUE	FALSE	FALSE
TRUE	FALSE	TRUE	FALSE	FALSE	FALSE
FALSE	TRUE	FALSE	TRUE	FALSE	FALSE
TRUE	TRUE	FALSE	TRUE	FALSE	FALSE
TRUE	FALSE	FALSE	TRUE	FALSE	FALSE
TRUE	FALSE	TRUE	TRUE	TRUE	FALSE
FALSE	FALSE	TRUE	TRUE	FALSE	FALSE
TRUE	FALSE	FALSE	TRUE	FALSE	FALSE
TRUE	FALSE	FALSE	TRUE	FALSE	FALSE
		...			
TRUE	FALSE	FALSE	TRUE	FALSE	FALSE

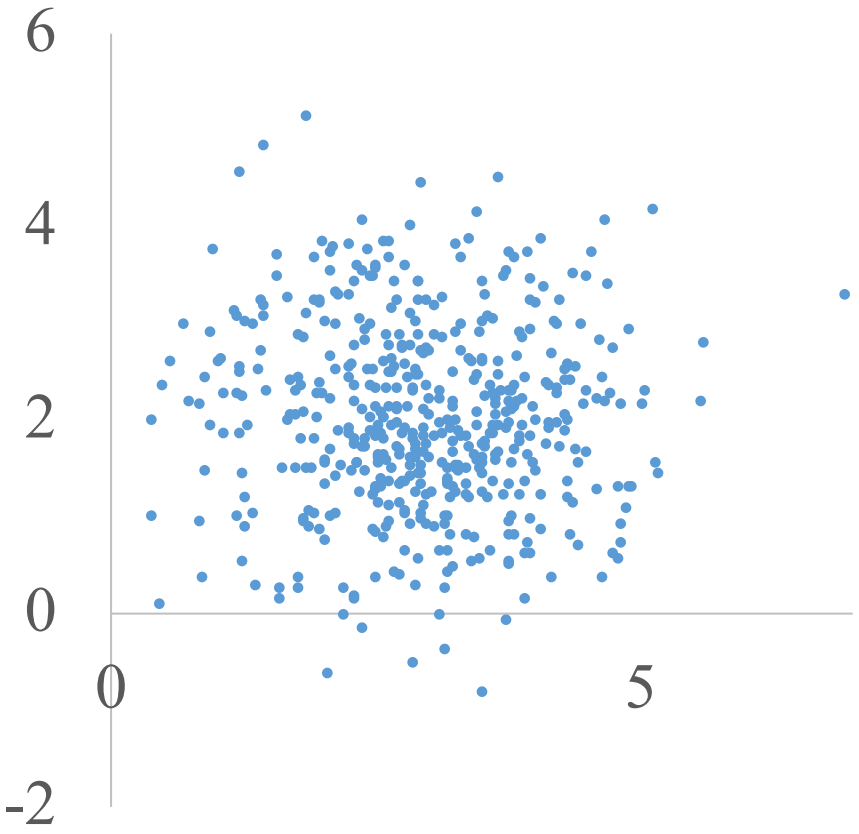
Expression Amount

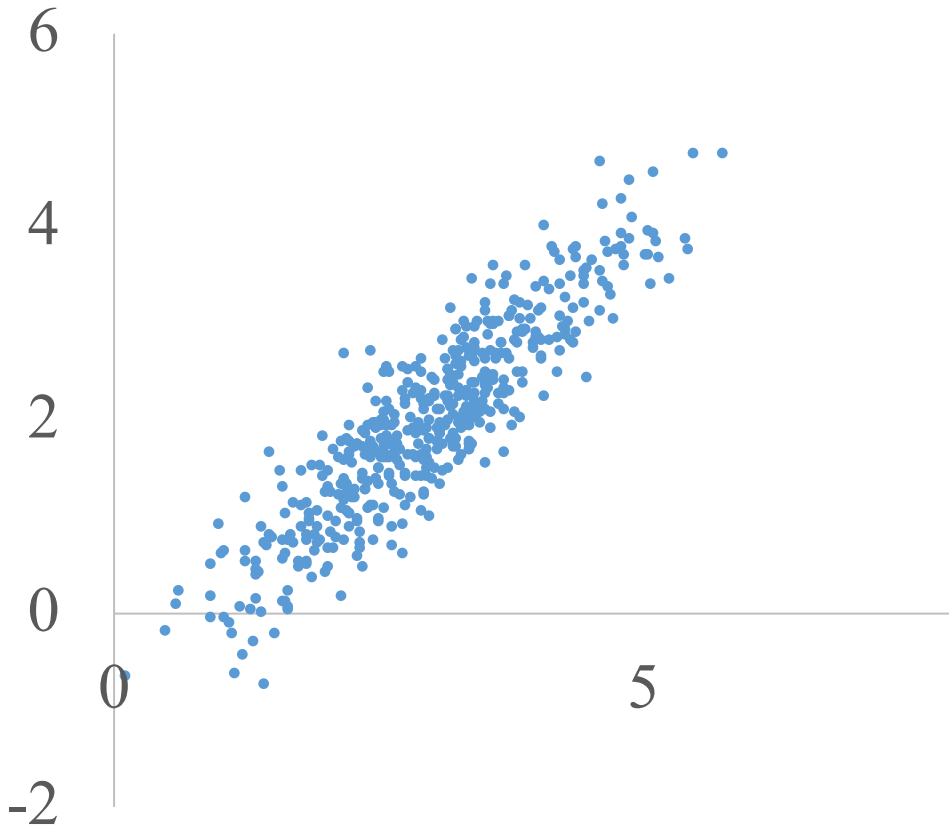
Gene5	Trait
0.76	0.83
0.94	0.85
0.82	0.03
0.94	0.32
0.50	0.10
0.40	0.53
0.90	0.67
0.29	0.71
0.72	0.25
0.15	0.24
0.79	0.98
0.68	0.77
0.71	0.37
0.36	0.18
0.62	0.08
0.59	0.38
0.82	0.76

Spot The Difference

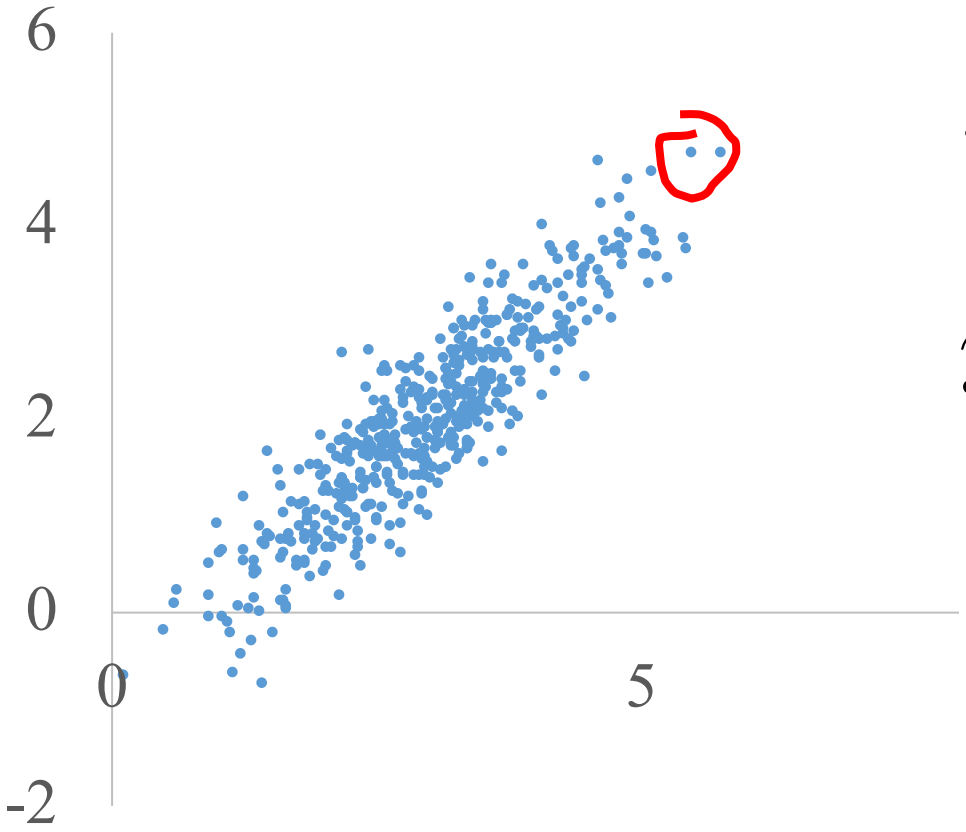


Spot The Difference





Vary Together

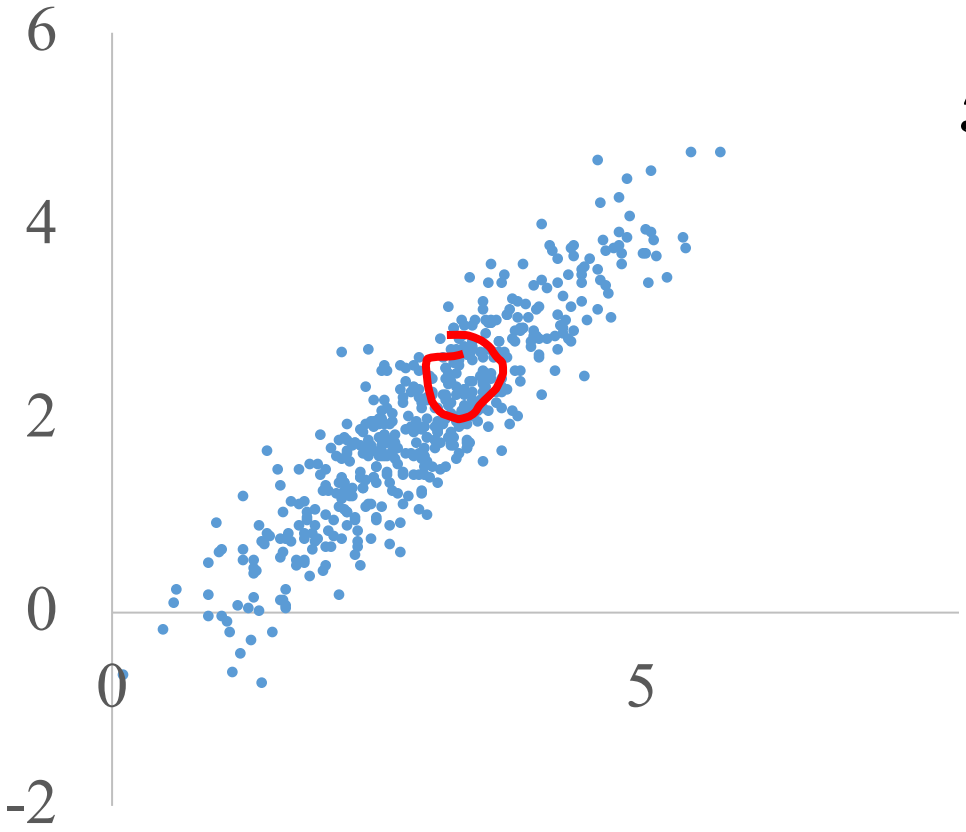


$$x - E[x] = 3$$

$$y - E[y] = 2.6$$

$$(x - E[x])(y - E[y]) = 7.8$$

Vary Together

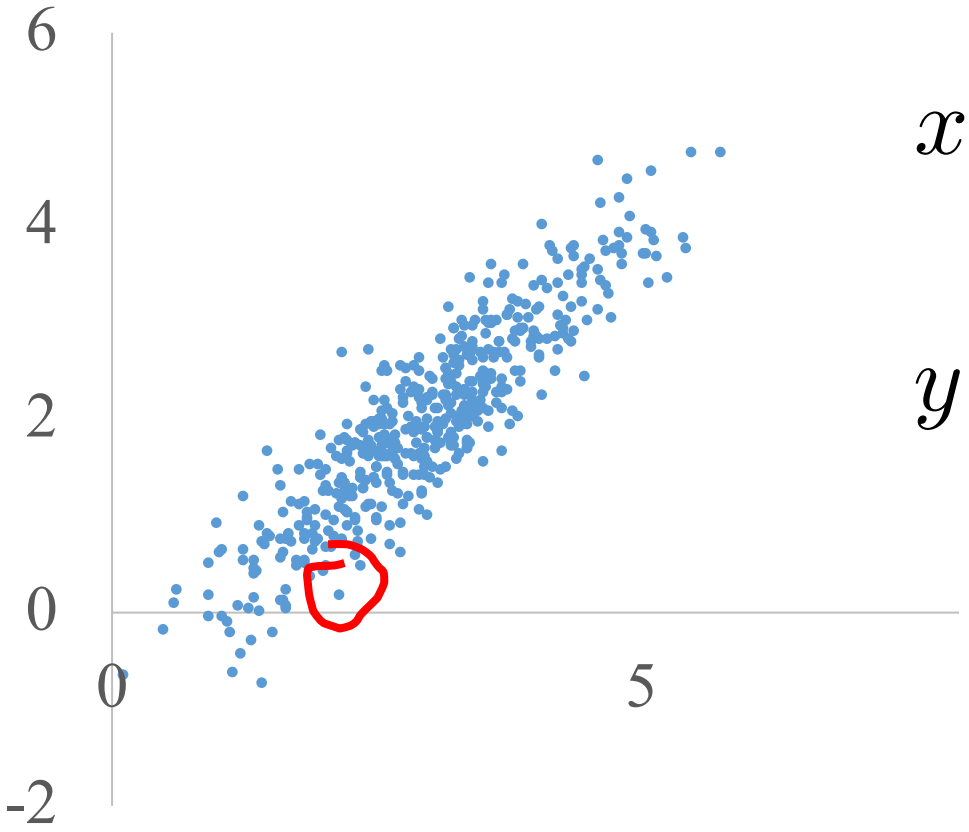


$$x - E[x] \approx 0$$

$$y - E[y] \approx 0$$

$$(x - E[x])(y - E[y]) = 0$$

Vary Together

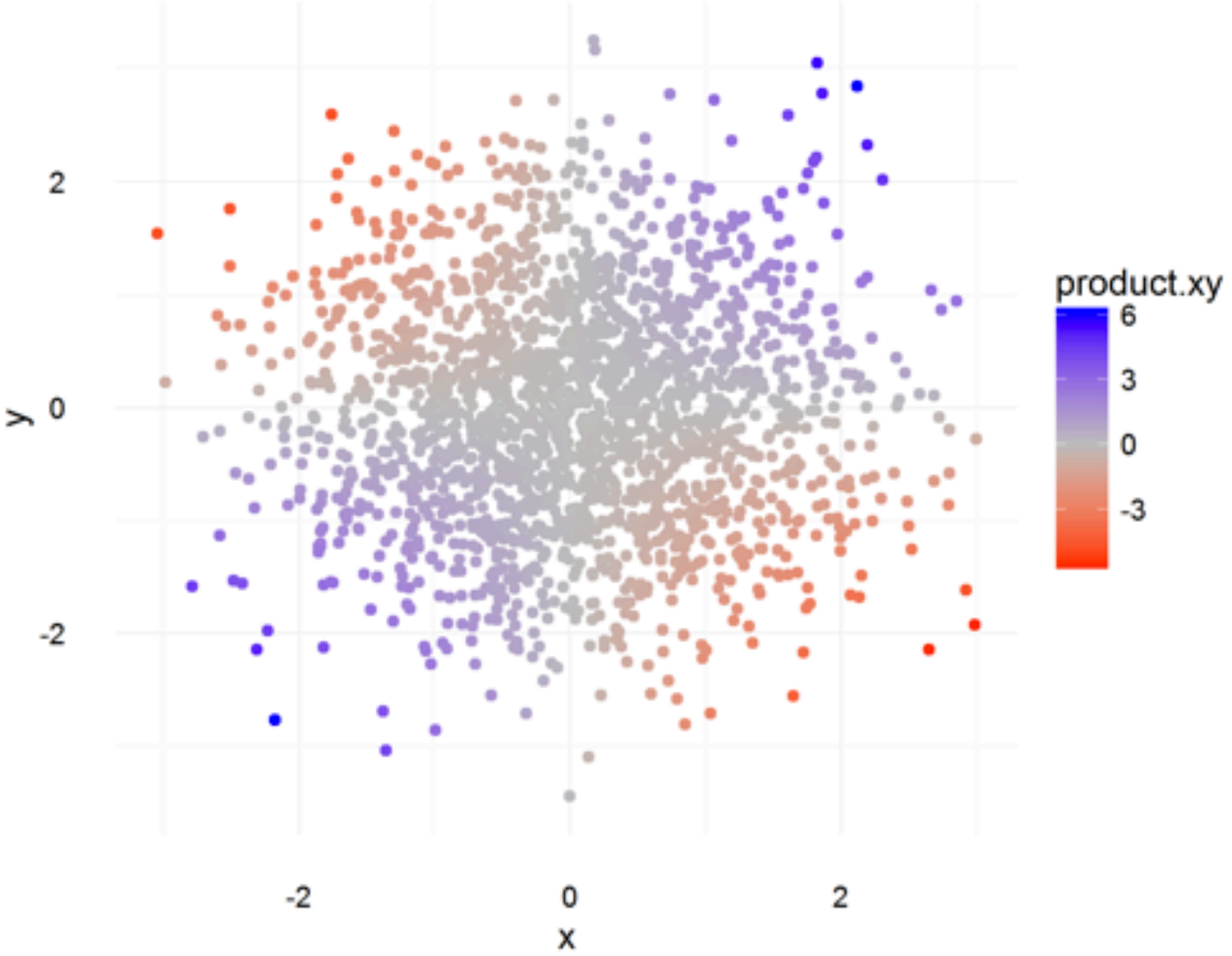


$$x - E[x] = -1.1$$

$$y - E[y] = -2.8$$

$$(x - E[x])(y - E[y]) \approx 3.1$$

Understanding Covariance



The Dance of the Covariance

- Say X and Y are arbitrary random variables
- Covariance of X and Y:

$$\mathrm{Cov}(X,Y) = E[(X - E[X])(Y - E[Y])]$$

x	y	$(x - E[X])(y - E[Y])p(x,y)$
Above mean	**Above mean**	**Positive**
Bellow mean	**Bellow mean**	**Positive**
Bellow mean	**Above mean**	**Negative**
Above mean	**Bellow mean**	**Negative**

The Dance of the Covariance

- Say X and Y are arbitrary random variables
- Covariance of X and Y:

$$\mathrm{Cov}(X,Y) = E[(X - E[X])(Y - E[Y])]$$

- Equivalently:

$$\begin{aligned}\mathrm{Cov}(X,Y) &= E[XY - E[X]Y - XE[Y] + E[Y]E[X]] \\ &= E[XY] - E[X]E[Y] - E[X]E[Y] + E[X]E[Y] \\ &= E[XY] - E[X]E[Y]\end{aligned}$$

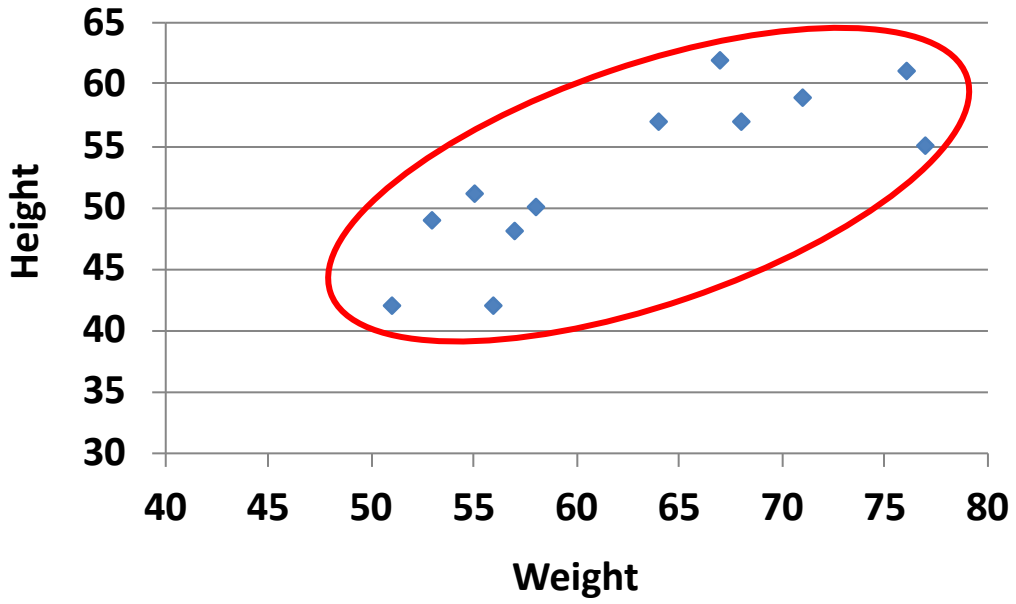
  - X and Y independent, E[XY] = E[X]E[Y] → Cov(X,Y) = 0
  - But Cov(X,Y) = 0 does **not** imply X and Y independent!

Covariance and Data

- Consider the following data:

Weight	Height	Weight * Height
64	57	3648
71	59	4189
53	49	2597
67	62	4154
55	51	2805
58	50	2900
77	55	4235
57	48	2736
56	42	2352
51	42	2142
76	61	4636
68	57	3876

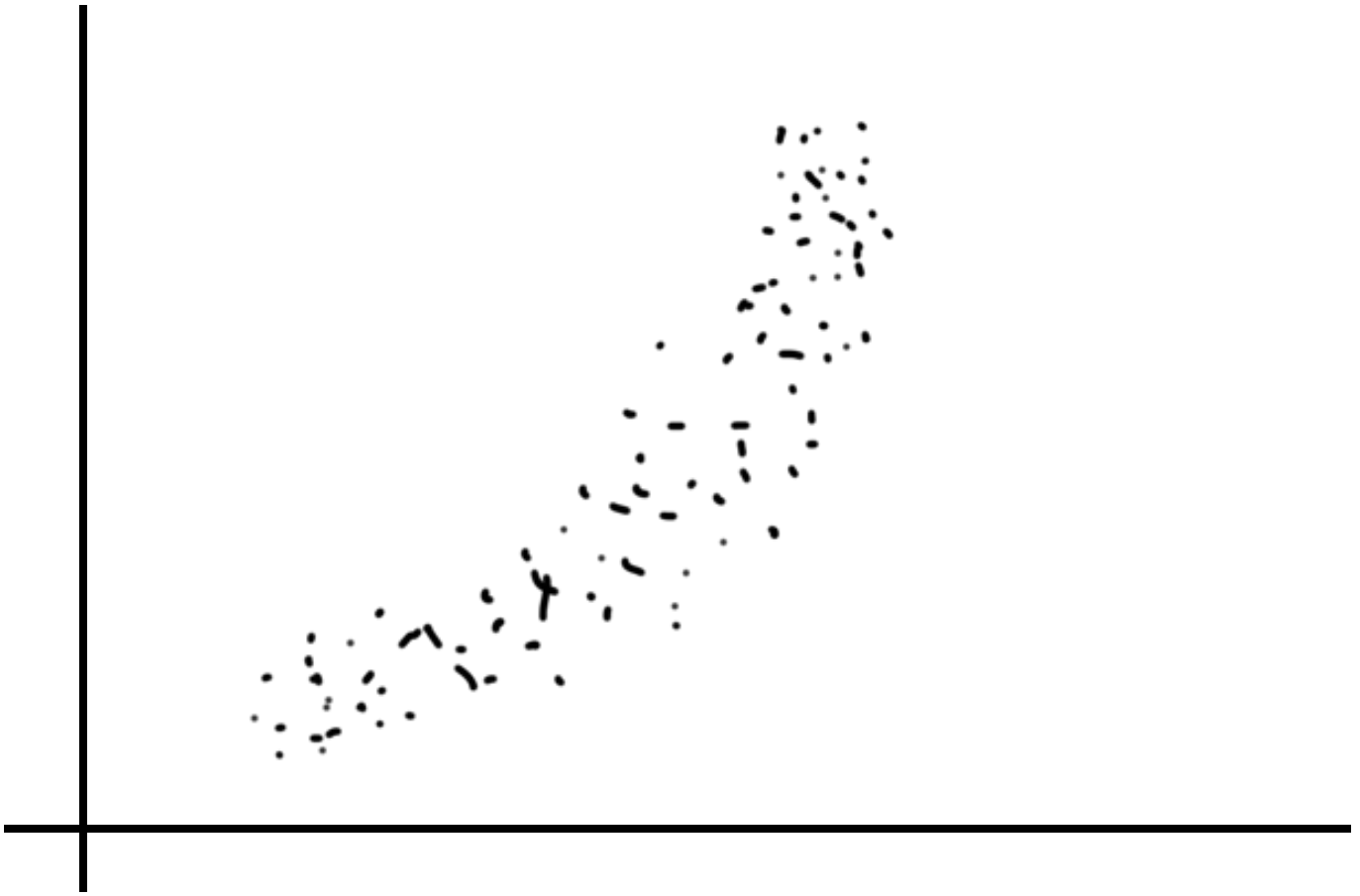
E[W]	E[H]	E[W*H]
= 62.75	= 52.75	= 3355.83



$$\begin{aligned} Cov(W, H) &= E[W*H] - E[W]E[H] \\ &= 3355.83 - (62.75)(52.75) \\ &= 45.77 \end{aligned}$$

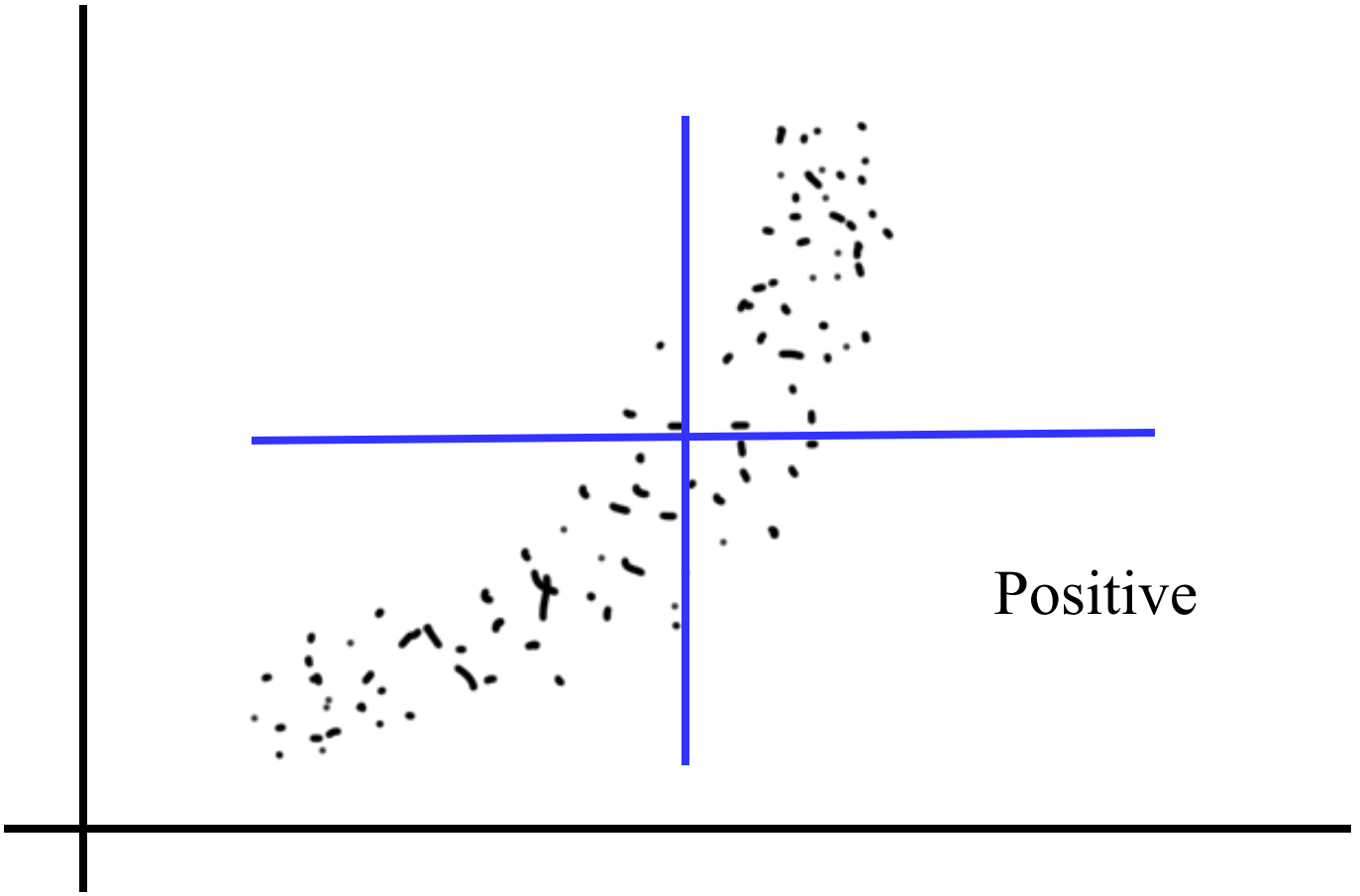
Covariance

Poll: (a) positive, (b) negative, (c) zero



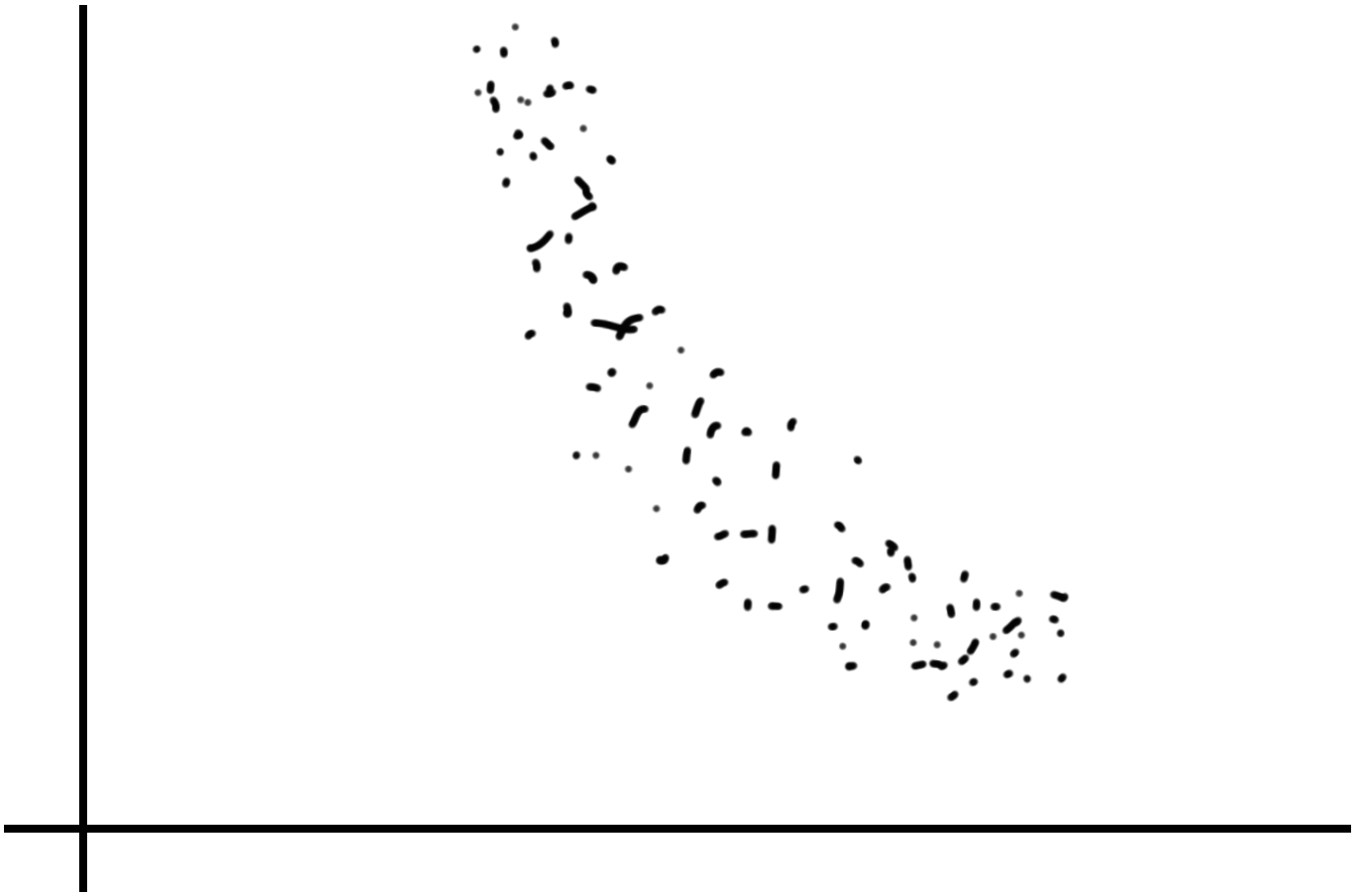
Covariance

Is the Covariance: (a) positive, (b) negative, (c) zero



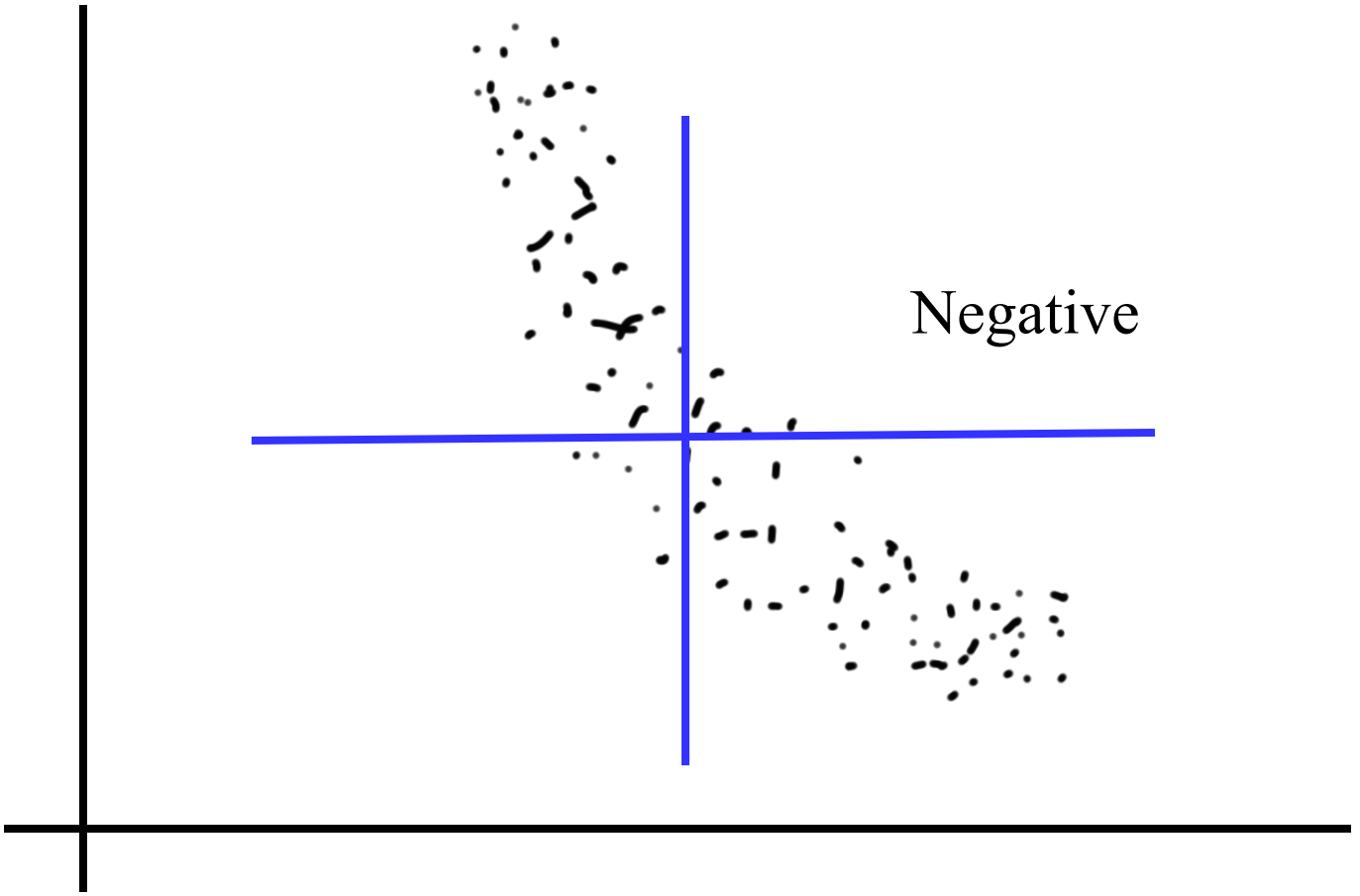
Covariance

Is the Covariance: (a) positive, (b) negative, (c) zero



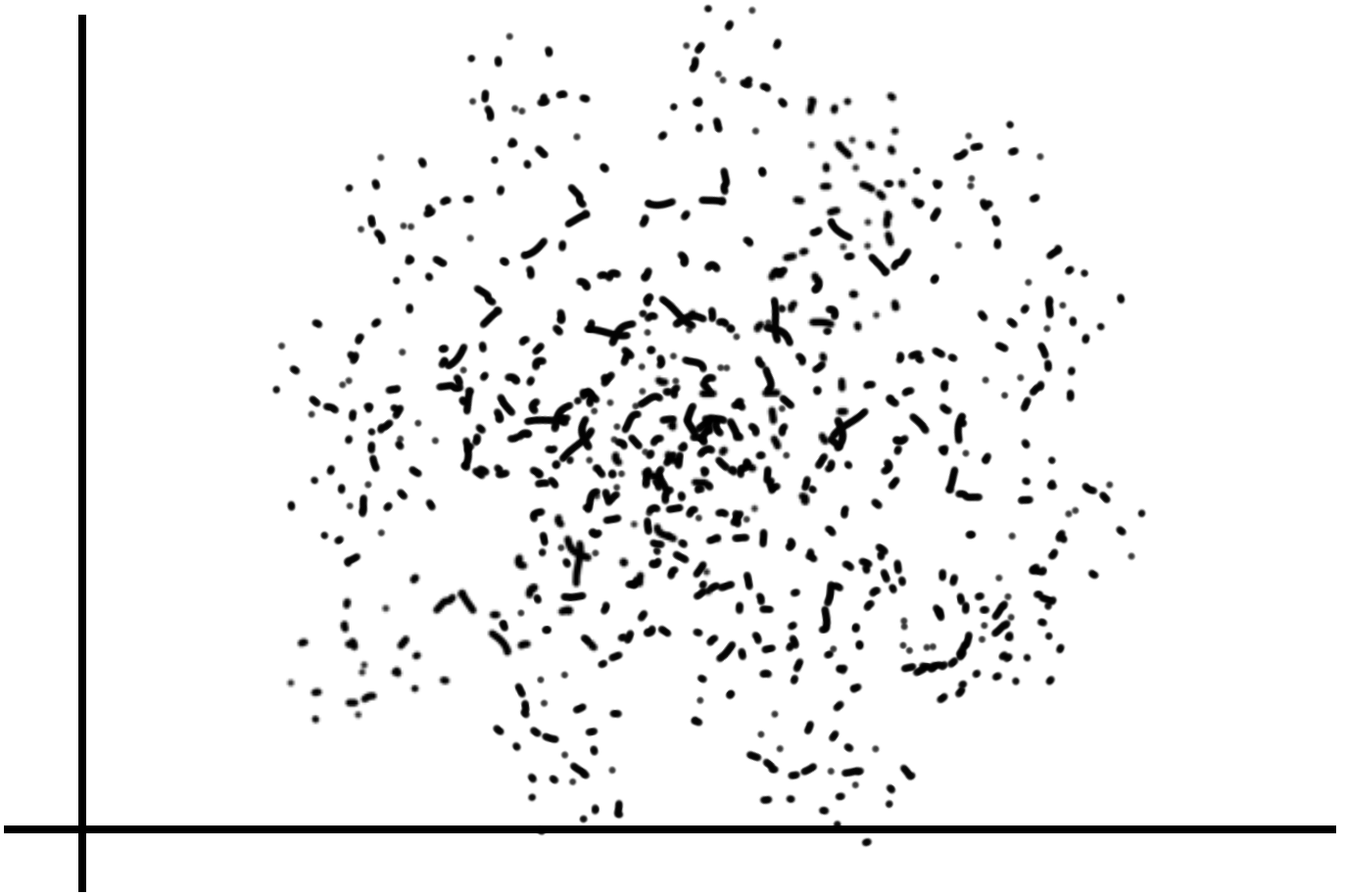
Covariance

Is the Covariance: (a) positive, (b) negative, (c) zero



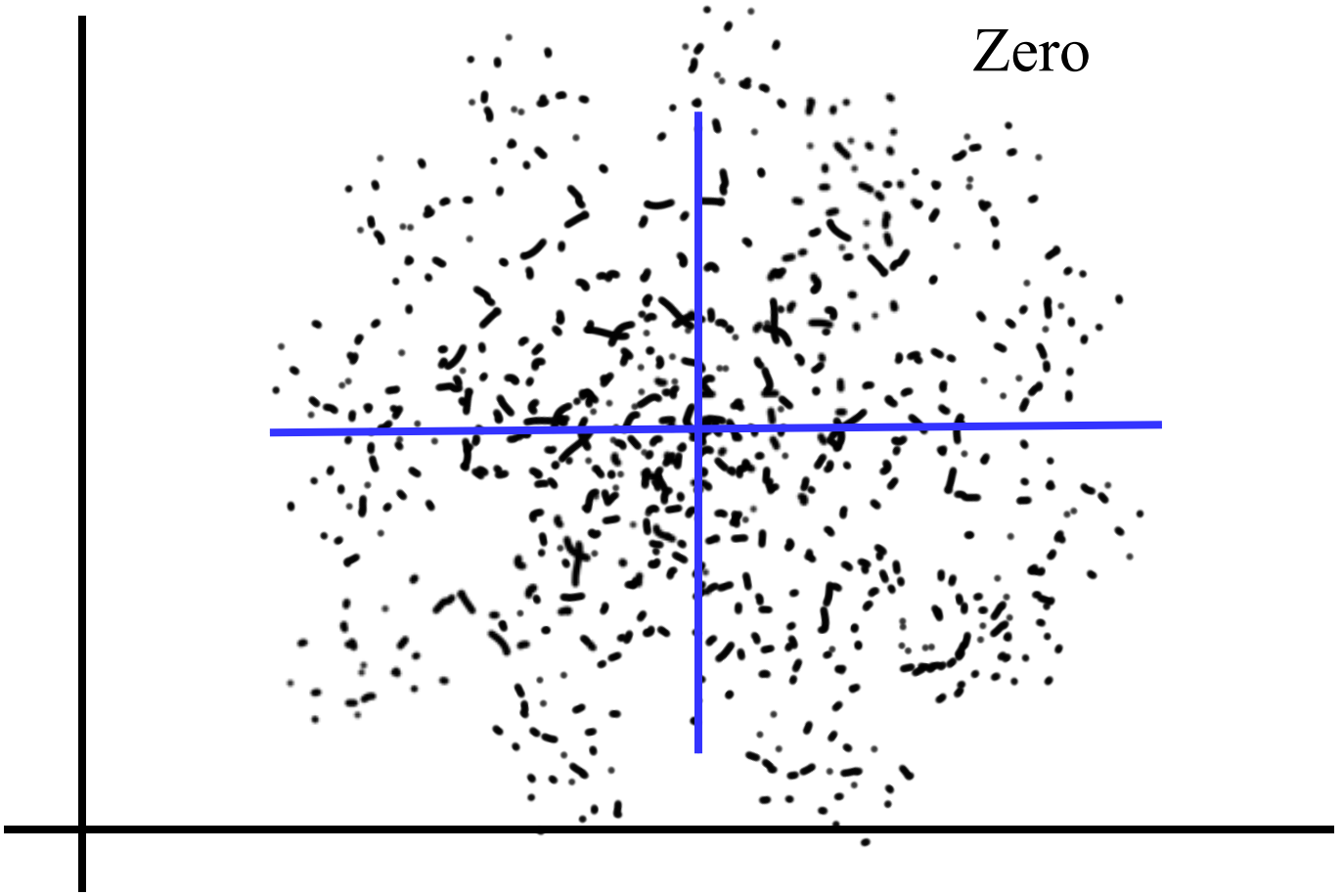
Covariance

Is the Covariance: (a) positive, (b) negative, (c) zero



Covariance

Is the Covariance: (a) positive, (b) negative, (c) zero



Independence and Covariance

- X and Y are random variables with PMF:

Y \ X	-1	0	1	$p_Y(y)$
0	1/3	0	1/3	2/3
1	0	1/3	0	1/3
$p_X(x)$	1/3	1/3	1/3	1

$$Y = \begin{cases} 0 & \text{if } X \neq 0 \\ 1 & \text{otherwise} \end{cases}$$

  - E[X] = -1(1/3) + 0(1/3) + 1(1/3) = 0
  - E[Y] = 0(2/3) + 1(1/3) = 1/3
  - Since XY = 0, E[XY] = 0
  - Cov(X, Y) = E[XY] – E[X]E[Y] = 0 – 0 = 0
- But, X and Y are clearly dependent!

Properties of Covariance

- Say X and Y are arbitrary random variables
  - $\mathrm{Cov}(X,Y) = \mathrm{Cov}(Y,X)$
  - $\mathrm{Cov}(X,X) = E[X^2] - E[X]E[X] = \mathrm{Var}(X)$
  - $\mathrm{Cov}(aX+b,Y) = a\mathrm{Cov}(X,Y)$
- Covariance of sums of random variables
  - $X_1, X_2, \ldots, X_n$ and $Y_1, Y_2, \ldots, Y_m$ are random variables
  - $\mathrm{Cov}\left(\sum_{i=1}^{n} X_i, \sum_{j=1}^{m} Y_j\right) = \sum_{i=1}^{n}\sum_{j=1}^{m} \mathrm{Cov}(X_i, Y_j)$

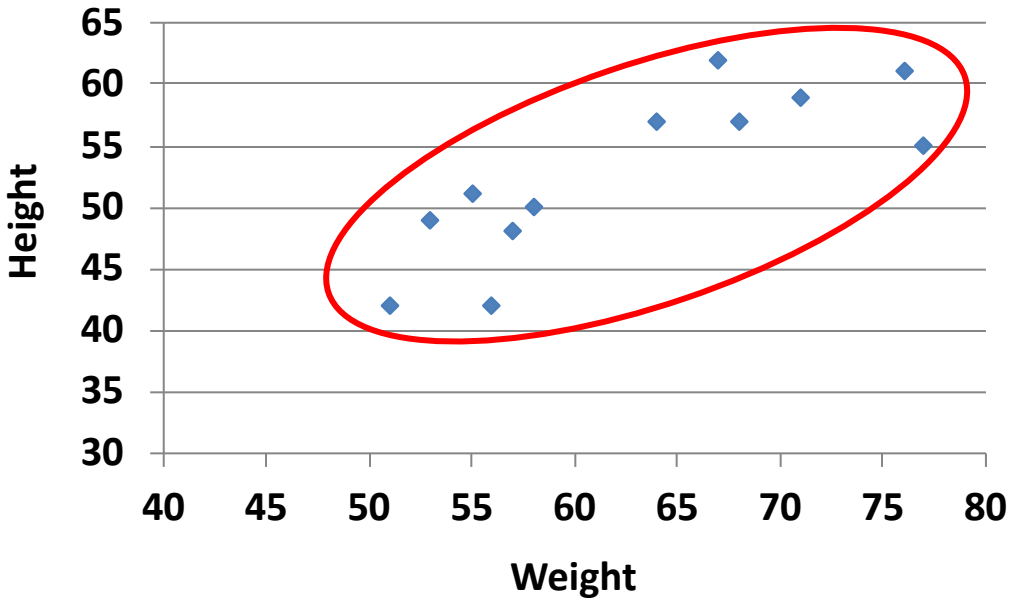
Correlation

What is Wrong With This?

- Consider the following data:

Weight	Height	Weight * Height
64	57	3648
71	59	4189
53	49	2597
67	62	4154
55	51	2805
58	50	2900
77	55	4235
57	48	2736
56	42	2352
51	42	2142
76	61	4636
68	57	3876

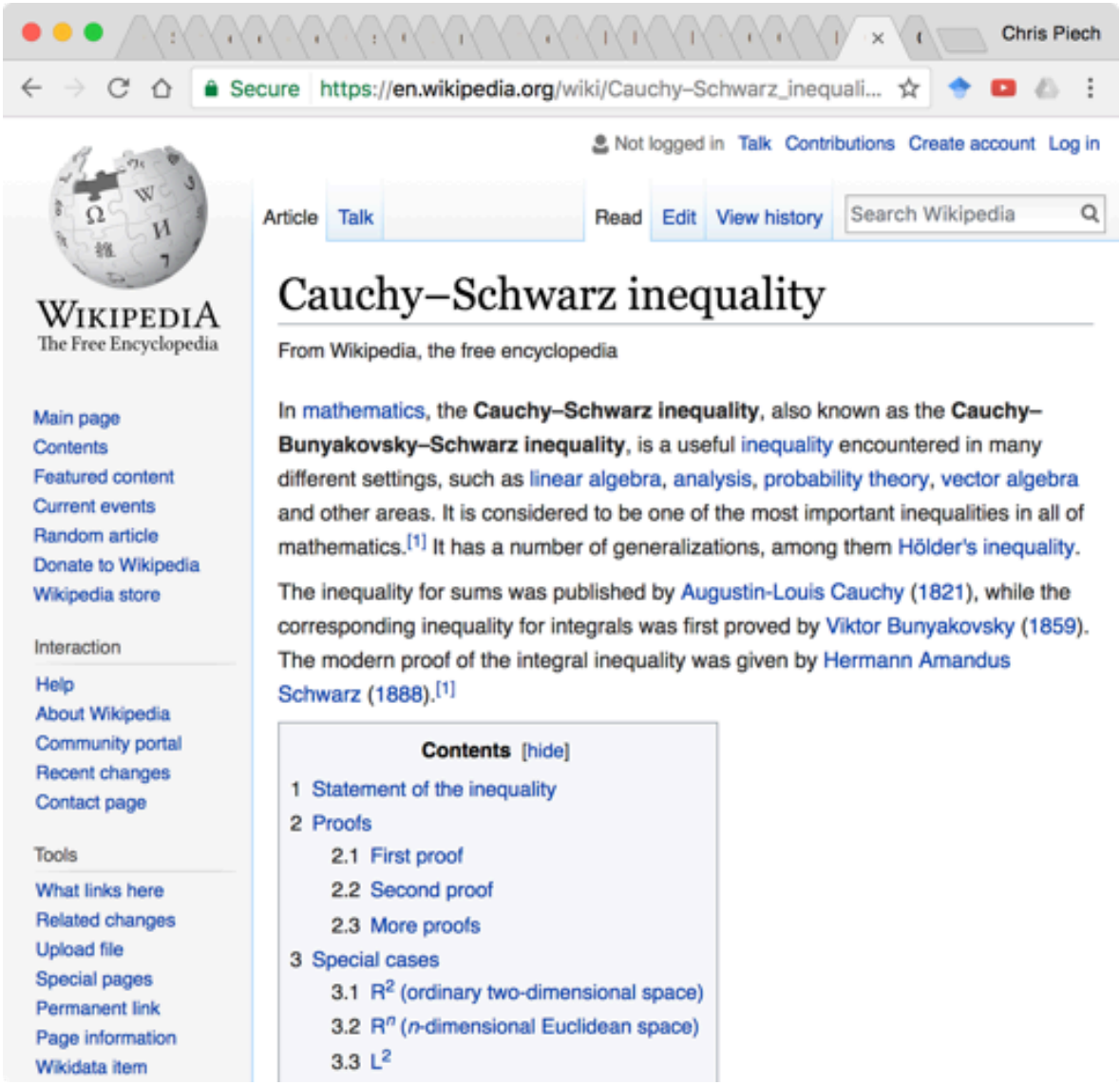
E[W]	E[H]	E[W*H]
= 62.75	= 52.75	= 3355.83



Cov(W, H) = E[W*H] – E[W]E[H]

= 3355.83 – (62.75)(52.75)

= 45.77



$$-\text{Std}(X)\text{Std}(Y) \leq \text{Cov}(X, Y) \leq \text{Std}(X)\text{Std}(Y)$$

Viva La Correlatión

- Say X and Y are arbitrary random variables
  - Correlation of X and Y, denoted $\rho(X, Y)$:

$$\rho(X,Y) = \frac{\mathrm{Cov}(X,Y)}{\sqrt{\mathrm{Var}(X)\mathrm{Var}(Y)}}$$

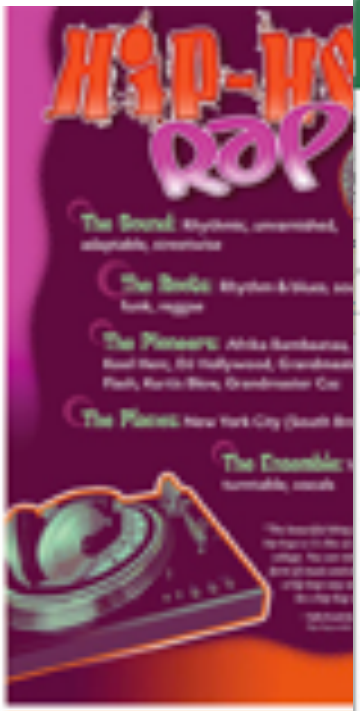
  - Note: $-1 \leq \rho(X, Y) \leq 1$
  - Correlation measures <u>linearity</u> between X and Y
  - $\rho(X, Y) = 1 \quad \Rightarrow$ Y = aX + b where $a = \sigma_y/\sigma_x$
  - $\rho(X, Y) = -1 \quad \Rightarrow$ Y = aX + b where $a = -\sigma_y/\sigma_x$
  - $\rho(X, Y) = 0 \quad \Rightarrow$ absence of <u>linear</u> relationship
    - But, X and Y can still be related in some other way!
  - If $\rho(X, Y) = 0$, we say X and Y are “uncorrelated”
    - Note: Independence implies uncorrelated, but **<u>not</u>** vice versa!

Viva La Correlatión

- Say X and Y are arbitrary random variables
  - Correlation of X and Y, denoted $\rho(X, Y)$:

$$\rho(X,Y) = \frac{\mathrm{Cov}(X,Y)}{\sqrt{\mathrm{Var}(X)\mathrm{Var}(Y)}}$$

- Say Y = cX. Correlation should be 1.



AutoSave OFF

Search Sheet

Home | Insert | Page Layout | Formulas | Data | Share

Clipboard | Font | Alignment | Number | Conditional Formatting | Format as Table | Cell Styles | Cells | Editing

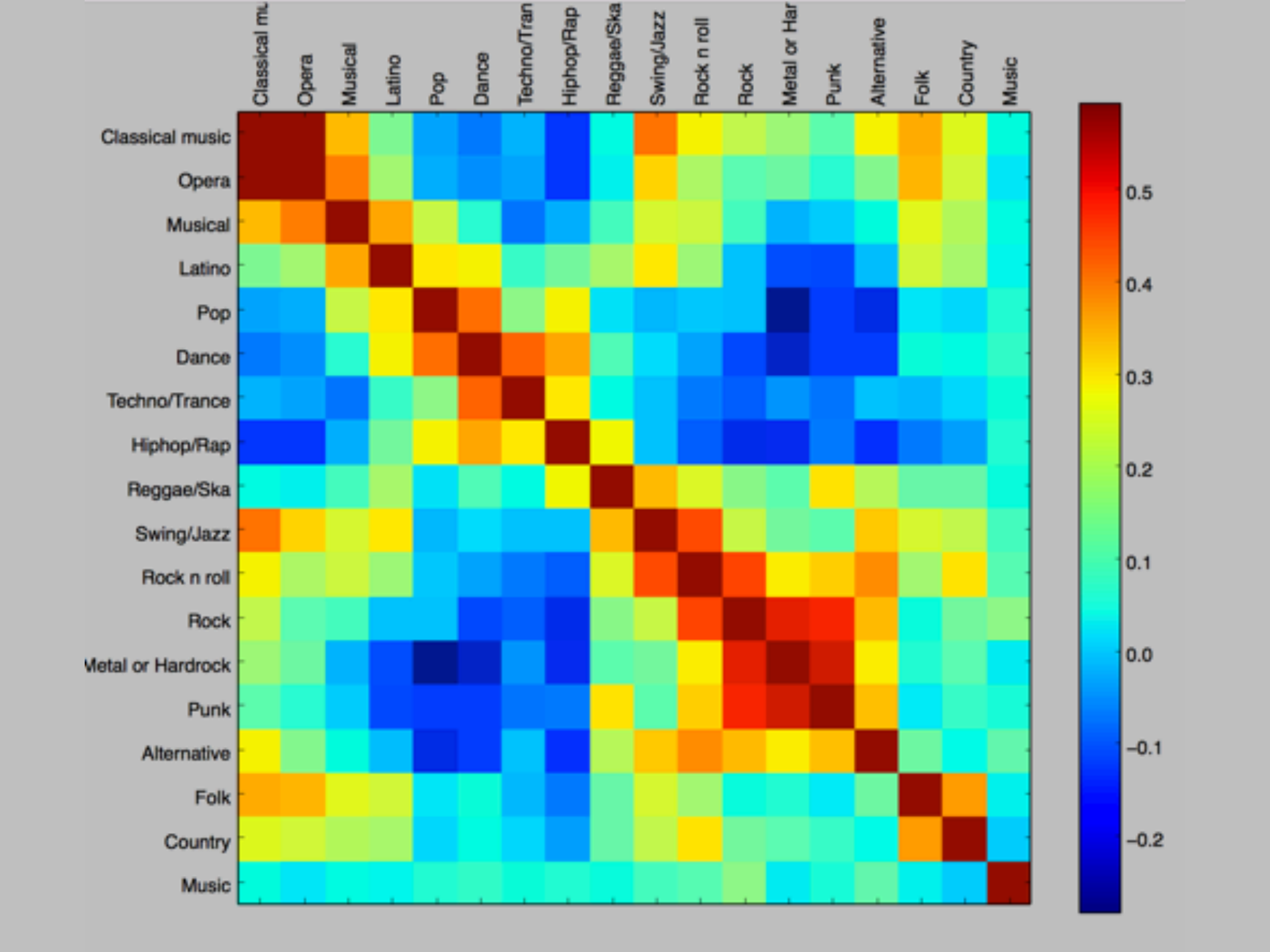
C15 | fx | 3

	A	B	C	D	E	F	G	H
1	Music	Dance	Folk	Country	Classical music	Musical	Pop	Rock
2	5	2	1	2	2	1	5	5
3	4	2	1	1	1	2	3	5
4	5	2	2	3	4	5	3	5
5	5	2	1	1	1	1	2	2
6	5	4	3	2	4	3	5	3
7	5	2	3	2	3	3	2	5
8	5	5	3	1	2	2	5	3
9	5	3	2	1	2	2	4	5
10	5	3	1	1	2	4	3	5
11	5	2	5	2	2	5	3	5
12	5	3	2	1	2	3	4	3
13	5	1	1	1	4	1	2	5
14	5	1	2	1	4	3	3	5
15	5	5	3	2	1	5	5	2
16	5	2	1	1	2	3	4	5
17	1	2	2	3	4	3	3	5
18	5	3	1	1	1	2	4	4
19	5	3	3	3	2	2	4	4
20	5	5	4	3	4	5	5	4
21	5	3	3	2	4	2	2	4
22	5	3	2	3	4	3	2	5
23	5	1	1	3	2	2	2	5
24	5	3	2	3	3	3	4	
25	5	4	2	2	2	4	4	5
26	5	3	1	1	4	3	3	5
27	5	4	2	1	2	3	5	1
28	5	5	5	4	5	3	4	4
29	4	3	4	1	3	2	2	4
30	5	5	1	1	1	1	3	4
31	5	3	4	2	3	3	3	4
32	4	4	3	3	3	3	4	4
33	4	4	1	3	2	3	5	3
34	5	3	1	3	2	3	3	4
35	5	2	2	3	4	5	4	3

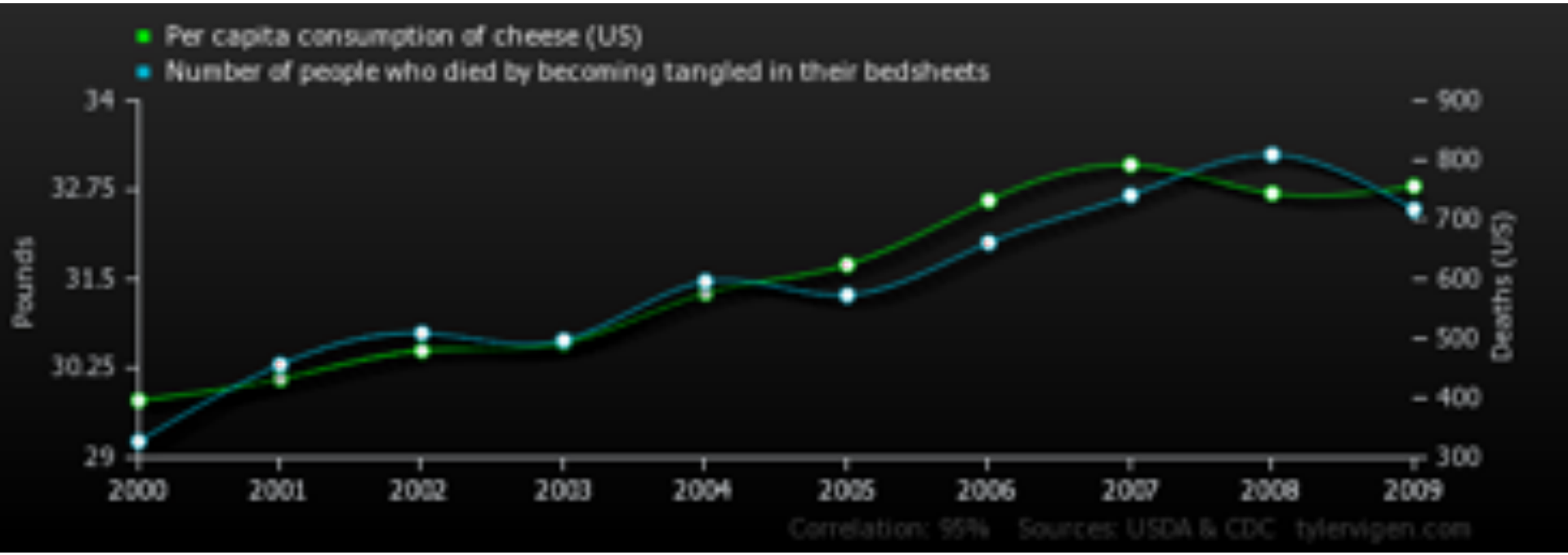
music

Ready

100%


Classical mu
Opera
Musical
Latino
Pop
Dance
Techno/Tran
Hiphop/Rap
Reggae/Ska
Swing/Jazz
Rock n roll
Rock
Metal or Har
Punk
Alternative
Folk
Country
Music
Classical music
Opera
Musical
Latino
Pop
Dance
Techno/Trance
Hiphop/Rap
Reggae/Ska
Swing/Jazz
Rock n roll
Rock
Metal or Hardrock
Punk
Alternative
Folk
Country
Music
0.5
0.4
0.3
0.2
0.1
0.0
−0.1
−0.2

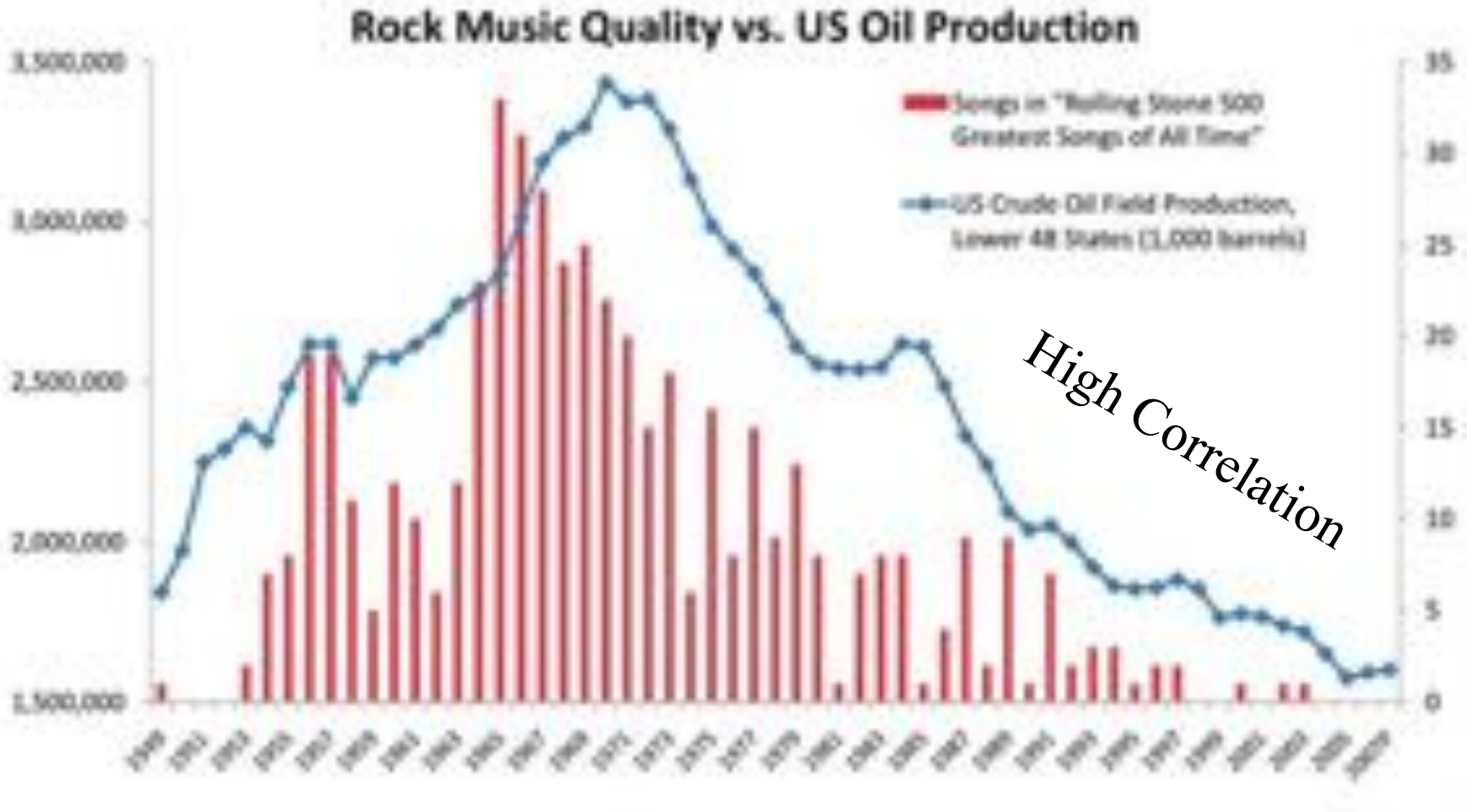
Tell your friends!



	2000	2001	2002	2003	2004	2005	2006	2007	2008	2009
Per capita consumption of cheese (US) *Pounds (USDA)*	29.8	30.1	30.5	30.6	31.3	31.7	32.6	33.1	32.7	32.8
Number of people who died by becoming tangled in their bedsheets *Deaths (US) (CDC)*	327	456	509	497	596	573	661	741	809	717

Correlation: 0.947091

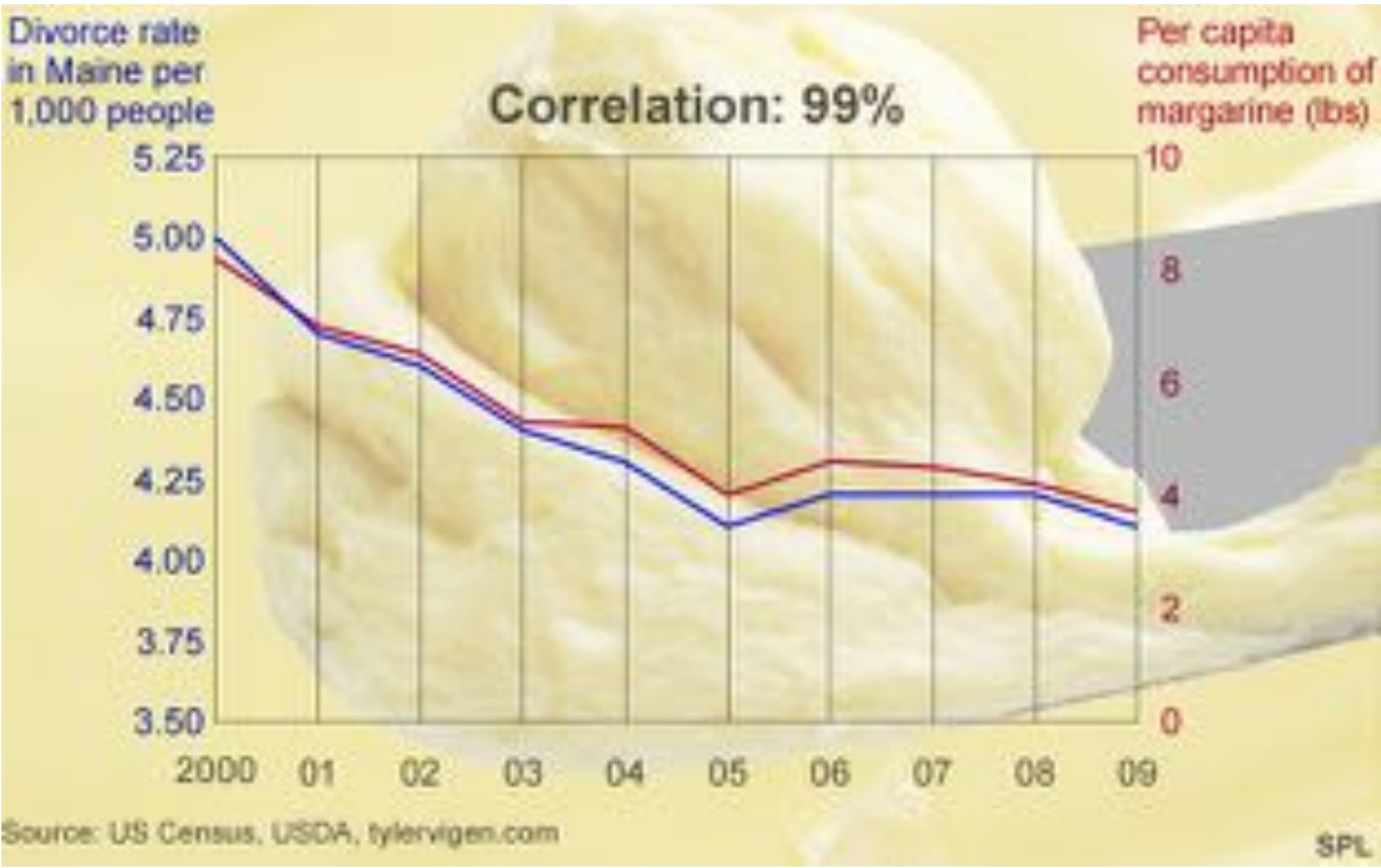
Rock Music Vs Oil?



Hubbert Peak Theory

http://www.aei.org/publication/blog/

Divorce Vs Butter?



http://www.bbc.com/news/magazine-27537142

Great Expectations

Noah Arthurs
CS109, Stanford University

Conditional Expectation

Conditional Expectation

- X and Y are jointly discrete random variables
  - Recall conditional PMF of X given Y = y:

$$p_{X|Y}(x \mid y) = P(X = x \mid Y = y) = \frac{p_{X,Y}(x, y)}{p_Y(y)}$$

- Define conditional expectation of X given Y = y:

$$E[X \mid Y = y] = \sum_x x P(X = x \mid Y = y) = \sum_x x p_{X|Y}(x \mid y)$$

- Analogously, jointly continuous random variables:

$$f_{X|Y}(x \mid y) = \frac{f_{X,Y}(x, y)}{f_Y(y)} \qquad E[X \mid Y = y] = \int_{-\infty}^{\infty} x f_{X|Y}(x \mid y)\, dx$$

Rolling Dice

- Roll two 6-sided dice D_1 and D_2
  - X = value of D_1 + D_2 Y = value of D_2
  - What is E[X | Y = 6]?

$$E[X \mid Y = 6] = \sum_x x P(X = x \mid Y = 6)$$

$$= \left(\frac{1}{6}\right)(7+8+9+10+11+12) = \frac{57}{6} = 9.5$$

  - Intuitively makes sense: 6 + E[value of D_1] = 6 + 3.5

Properties of Conditional Expectation

- X and Y are jointly distributed random variables

$$E[g(X) \mid Y = y] = \sum_x g(x)\, p_{X|Y}(x \mid y) \quad \text{or} \quad \int_{-\infty}^{\infty} g(x)\, f_{X|Y}(x \mid y)\, dx$$

- Expectation of conditional sum:

$$E\left[\sum_{i=1}^{n} X_i \mid Y = y\right] = \sum_{i=1}^{n} E[X_i \mid Y = y]$$

Conditional Expectation Functions

This is a number:

$$E[X]$$

This is a function of y:

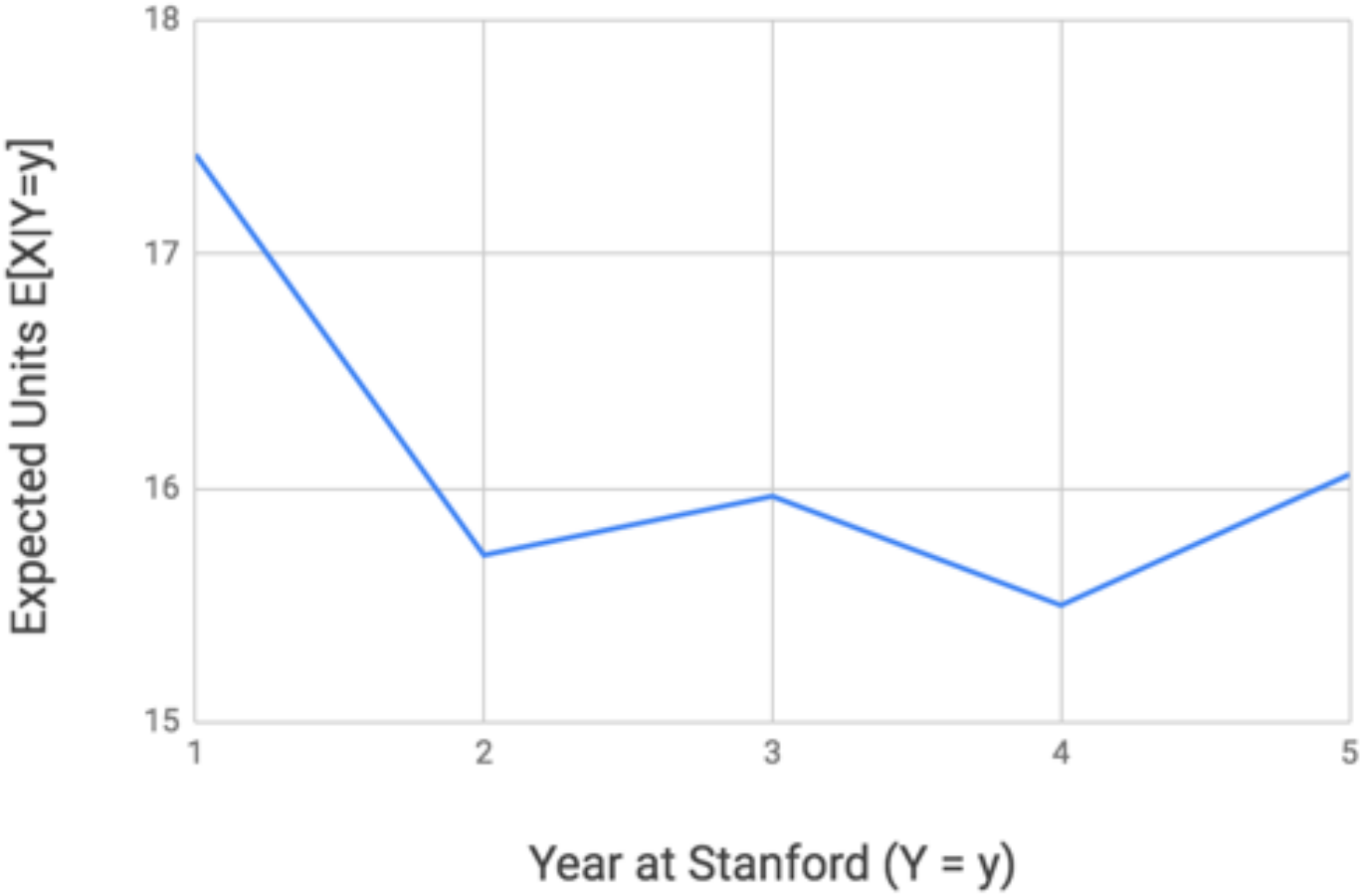
$$E[X|Y=y]$$

$E[X=5]$ Doesn't make sense. Take expectation of random variables, not events

Conditional Expectation Functions

X = units in fall quarter
Y = year in school

E[X | Y] ?



Law of Total Expectation

For any random variable X and any discrete random variable Y

$$E[X] = \sum_y E[X|Y = y]P(Y = y)$$

Analyzing Recursive Code

```
int Recurse() {
   int x = randomInt(1, 3);  // Equally likely values

   if (x == 1) return 3;
   else if (x == 2) return (5 + Recurse());
   else return (7 + Recurse());
}
```

- Let Y = value returned by `Recurse()`. What is E[Y]?

$$E[Y] = E[Y \mid X = 1]P(X = 1) + E[Y \mid X = 2]P(X = 2) + E[Y \mid X = 3]P(X = 3)$$

$$E[Y \mid X = 1] = 3$$

$$E[Y \mid X = 2] = E[5 + Y] = 5 + E[Y]$$

$$E[Y \mid X = 3] = E[7 + Y] = 7 + E[Y]$$

$$E[Y] = 3(1/3) + (5 + E[Y])(1/3) + (7 + E[Y])(1/3) = (1/3)(15 + 2E[Y])$$

$$E[Y] = 15$$

Protip: do this in CS161