CS 109 Review

Julia Daniel

Dec. 3, 2018

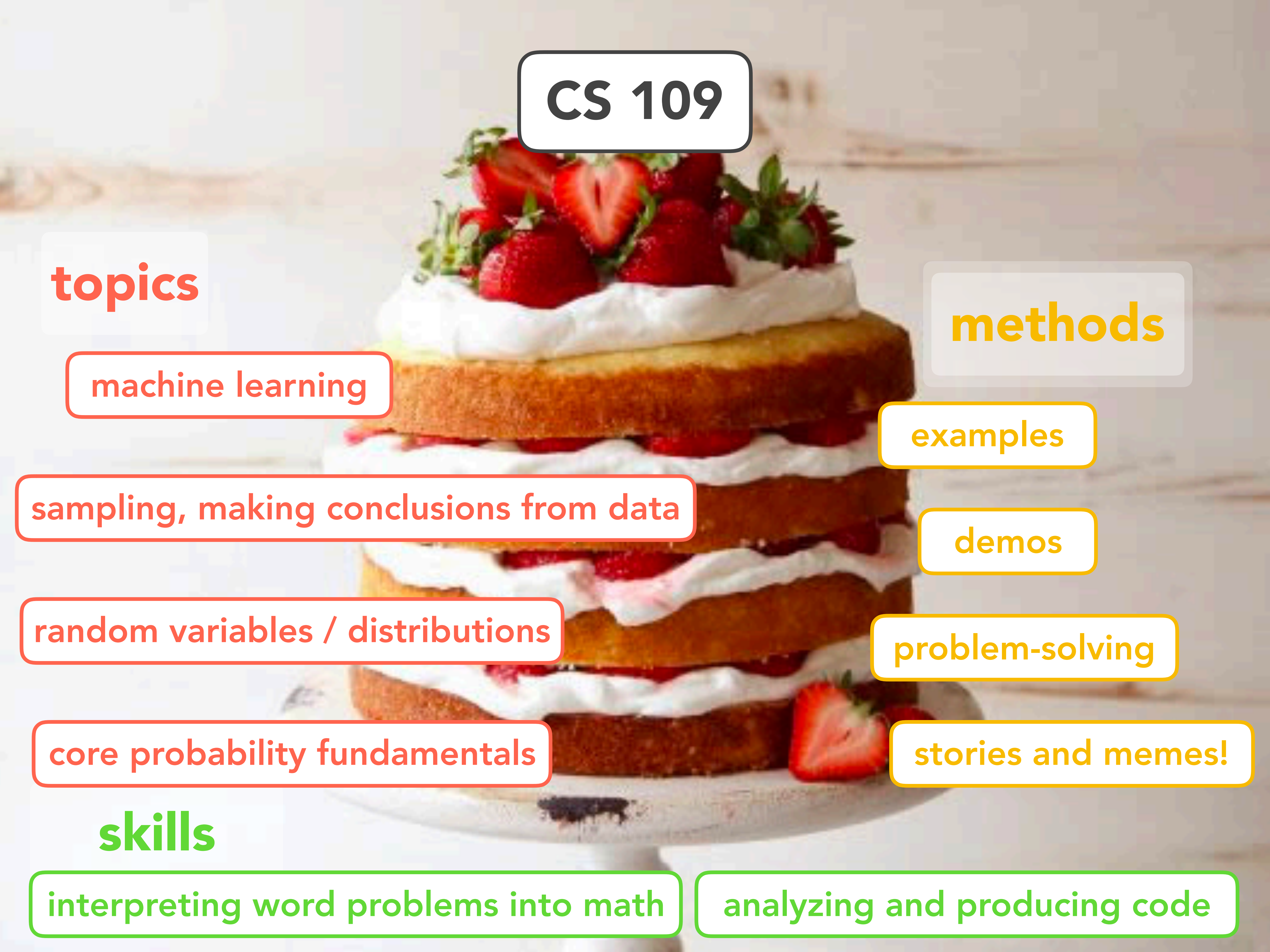
Where we're at

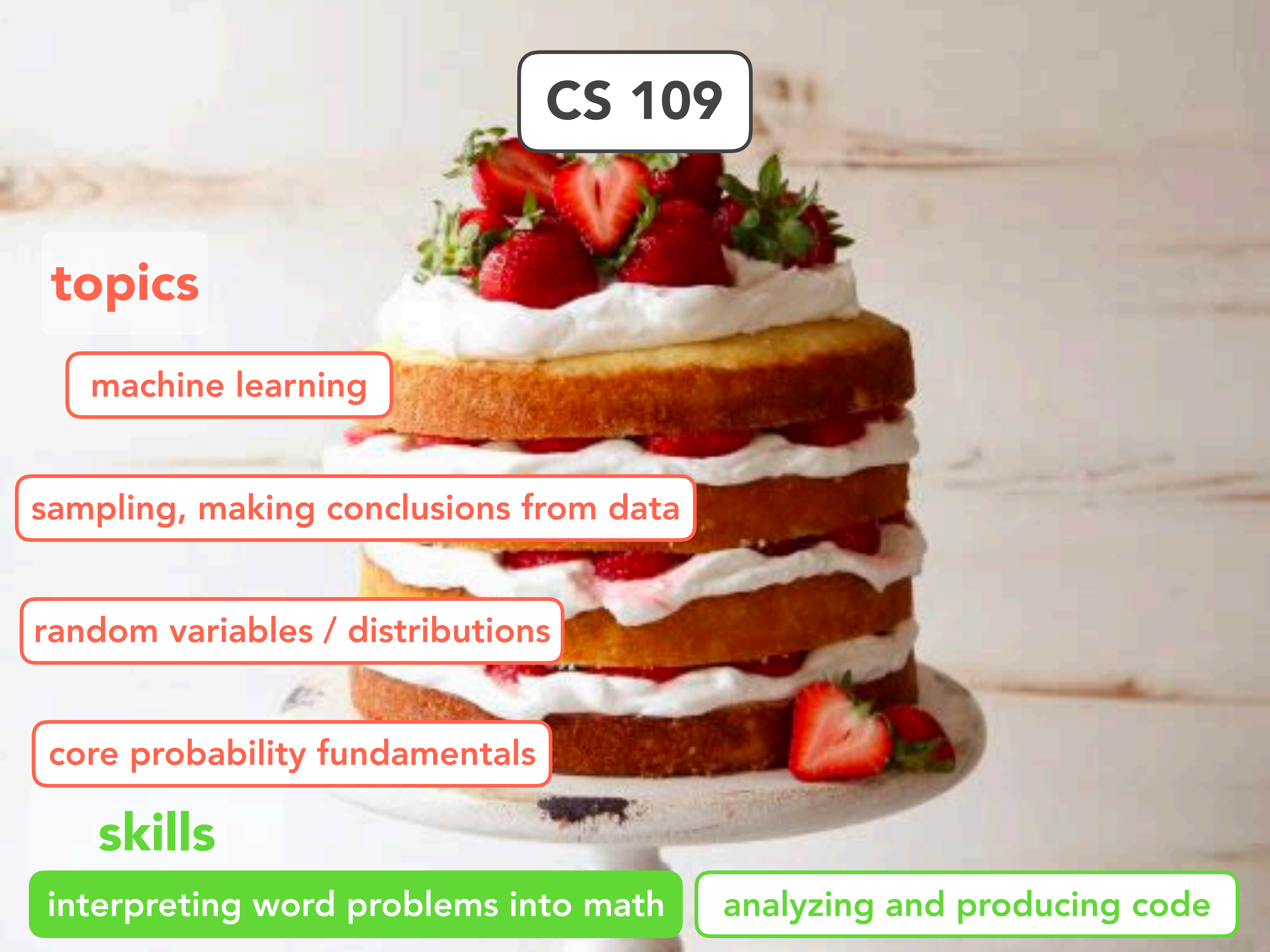
Last week: ML wrap-up, theoretical background for modern ML

This week: course overview, open questions after CS 109

Section: machine learning theory & practice

Next week: final exam Wednesday!


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

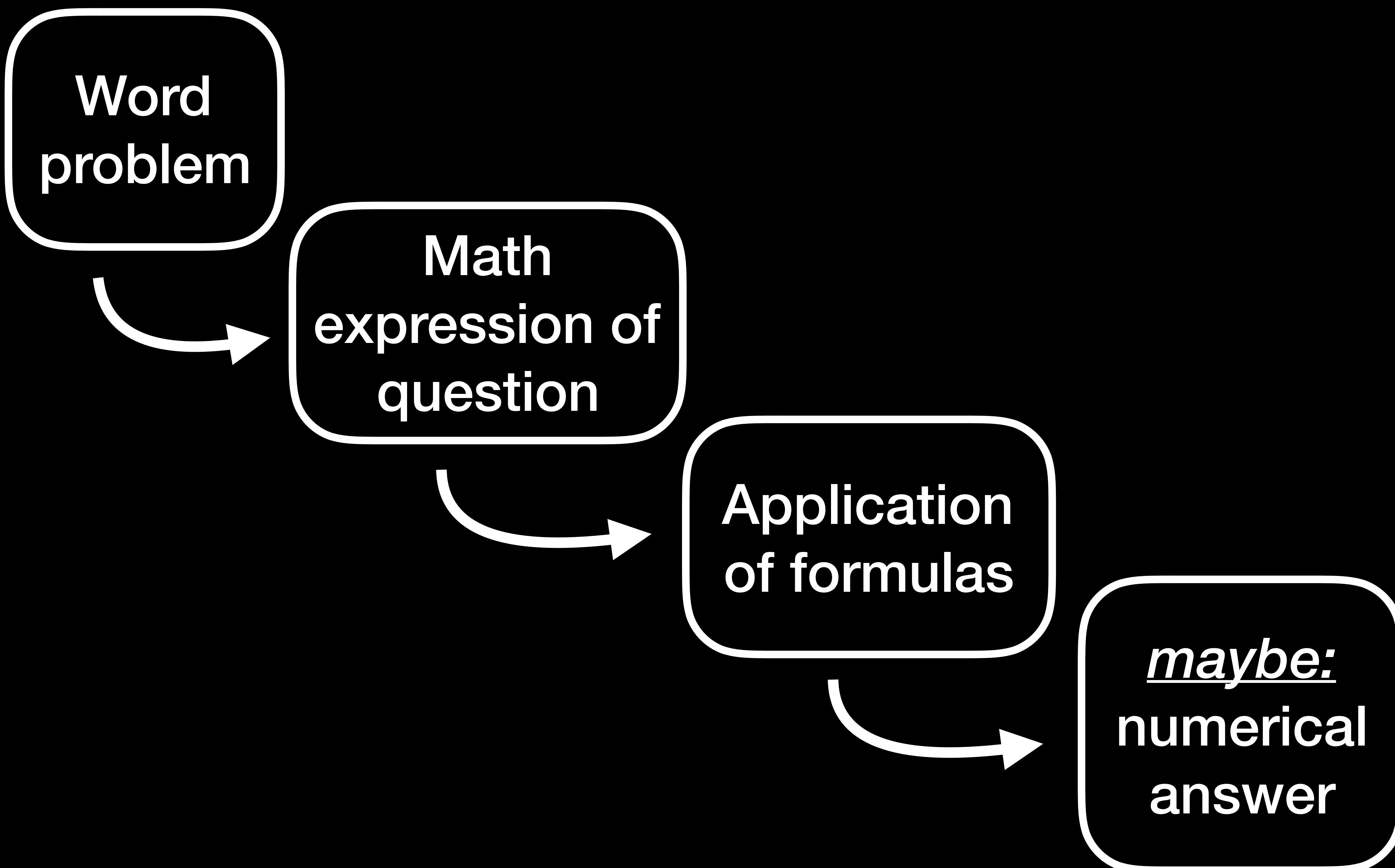
core probability fundamentals

skills

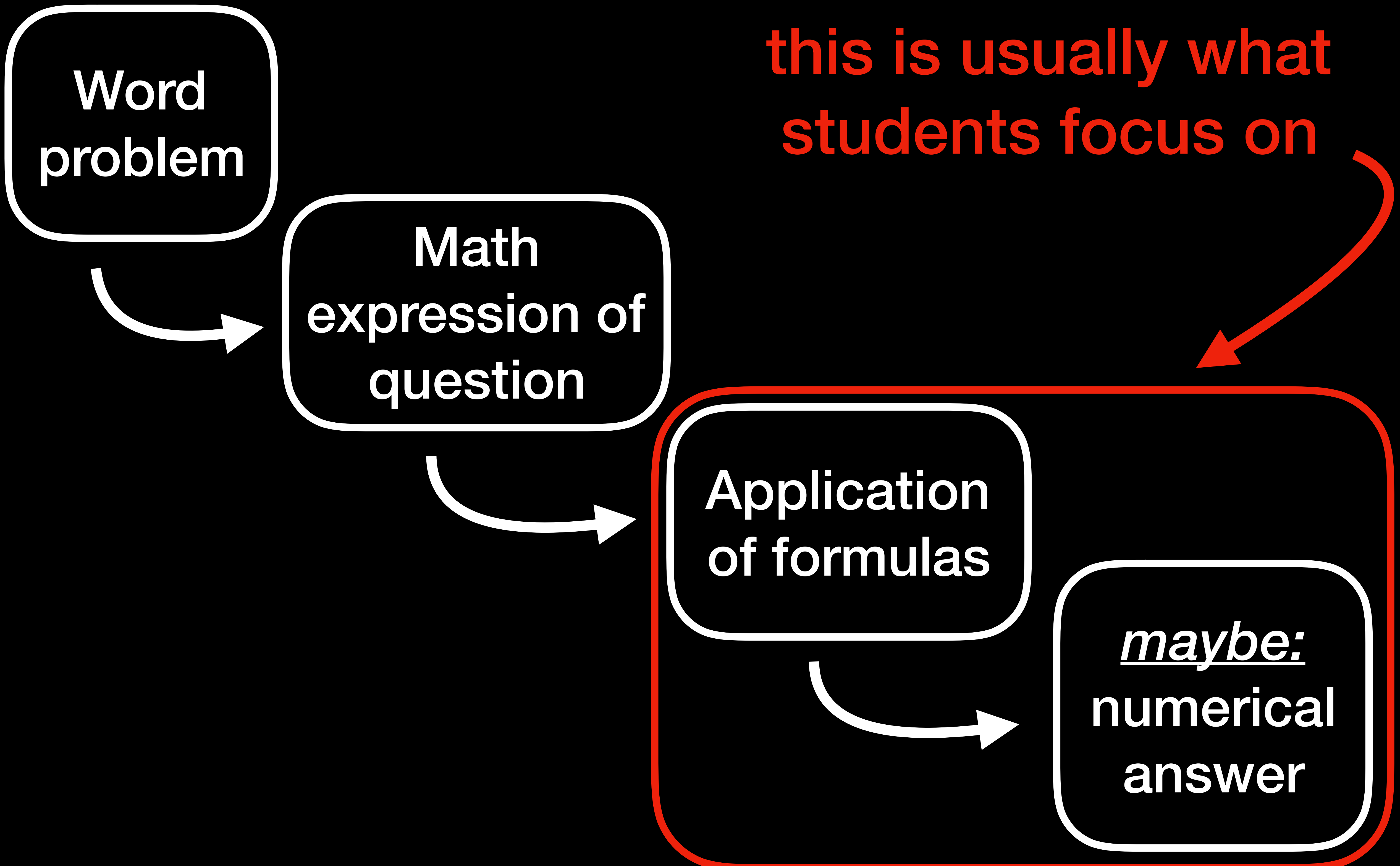
interpreting word problems into math

analyzing and producing code

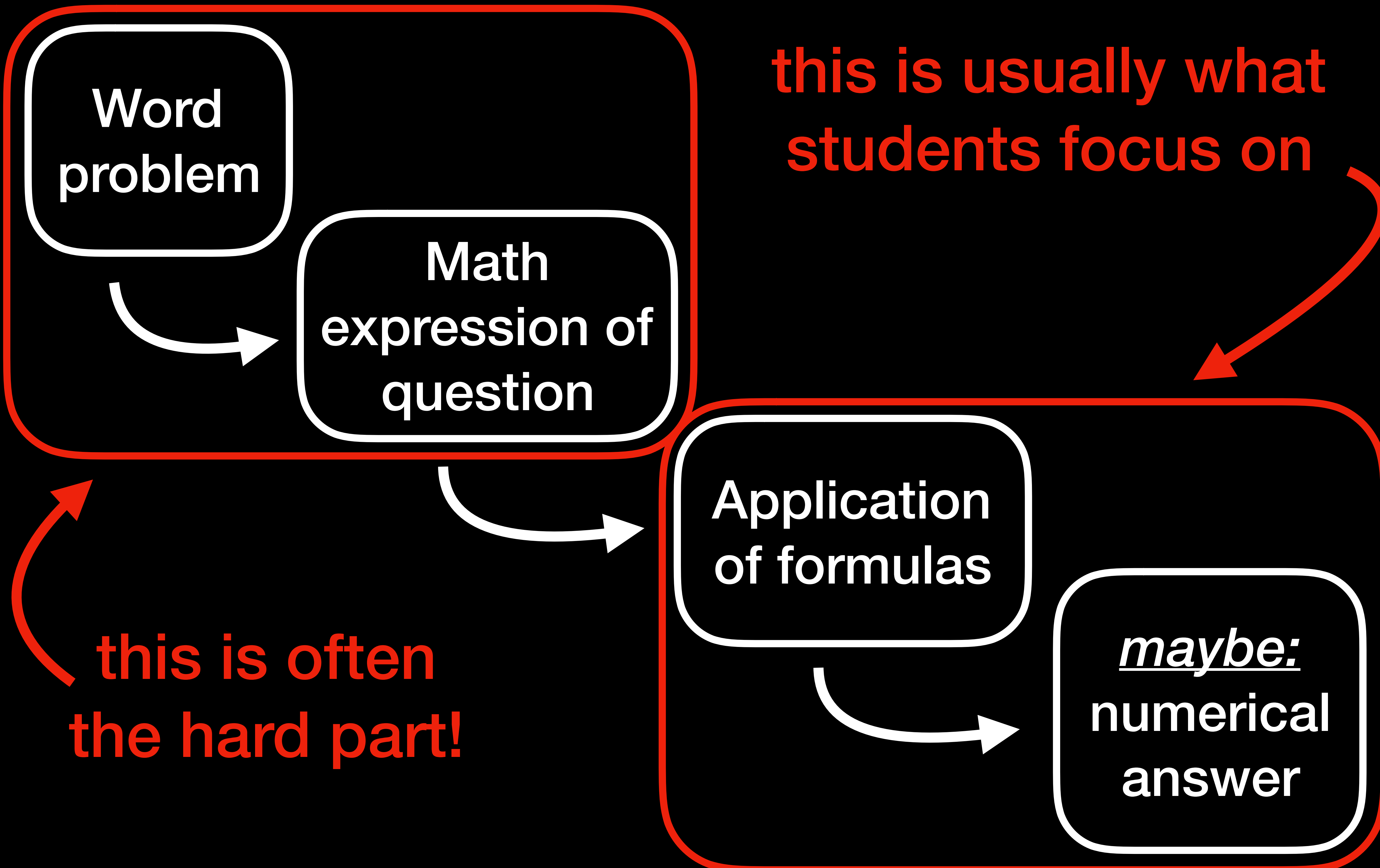
Solving a CS109 problem



Solving a CS109 problem



Solving a CS109 problem



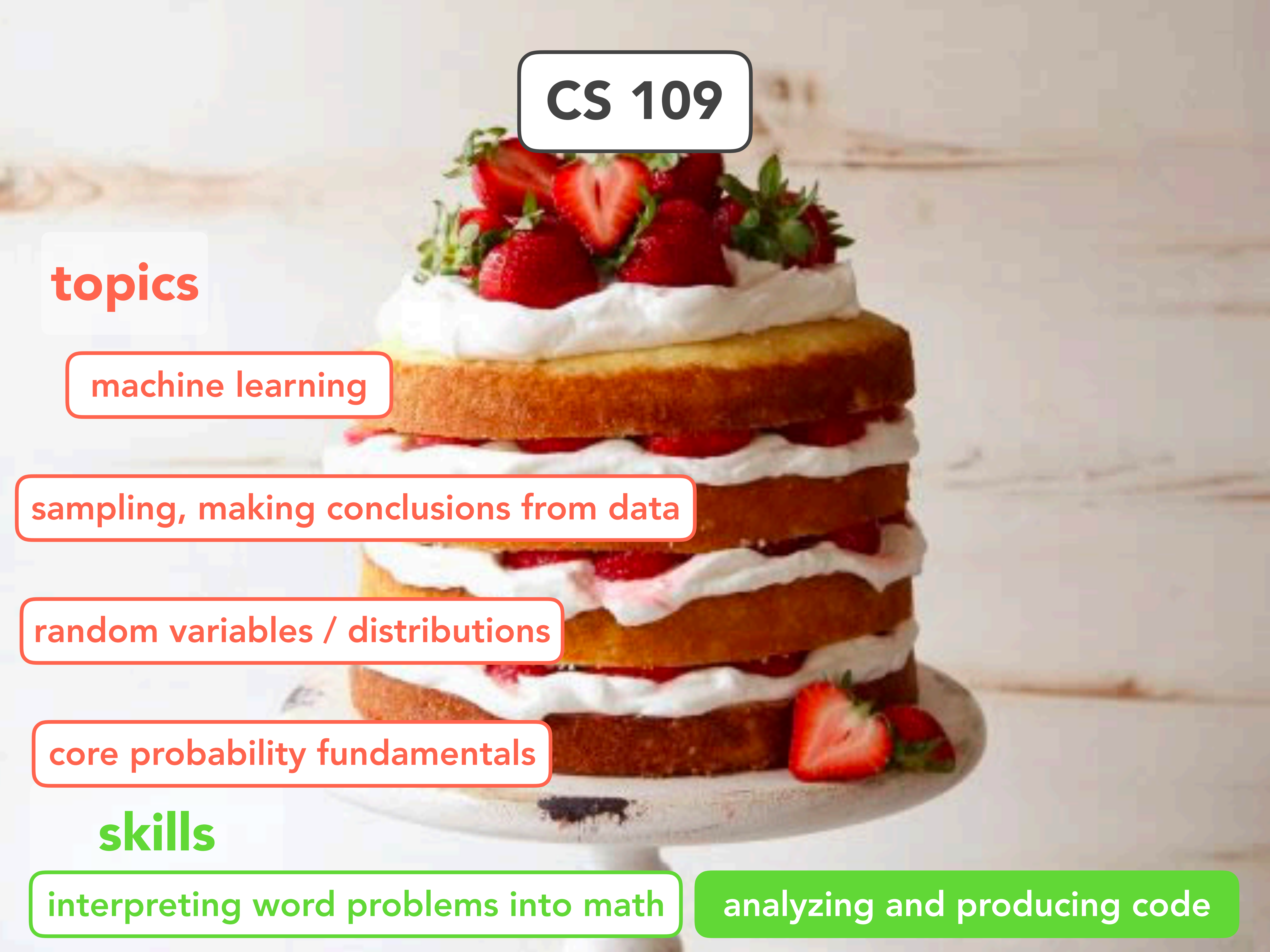
Step 1: Defining Your Terms

- What's a 'success'? What's the sample space?
- What does each random variable actually represent, in English? Every definition of an event or a random variable should have a **verb** in it. (' = ' is a verb)
- Make sure units match - particularly important for λ

Translating English to Probability

<u>What the problem asks:</u>	<u>What you should immediately think:</u>
"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

CO_2 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

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)

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!

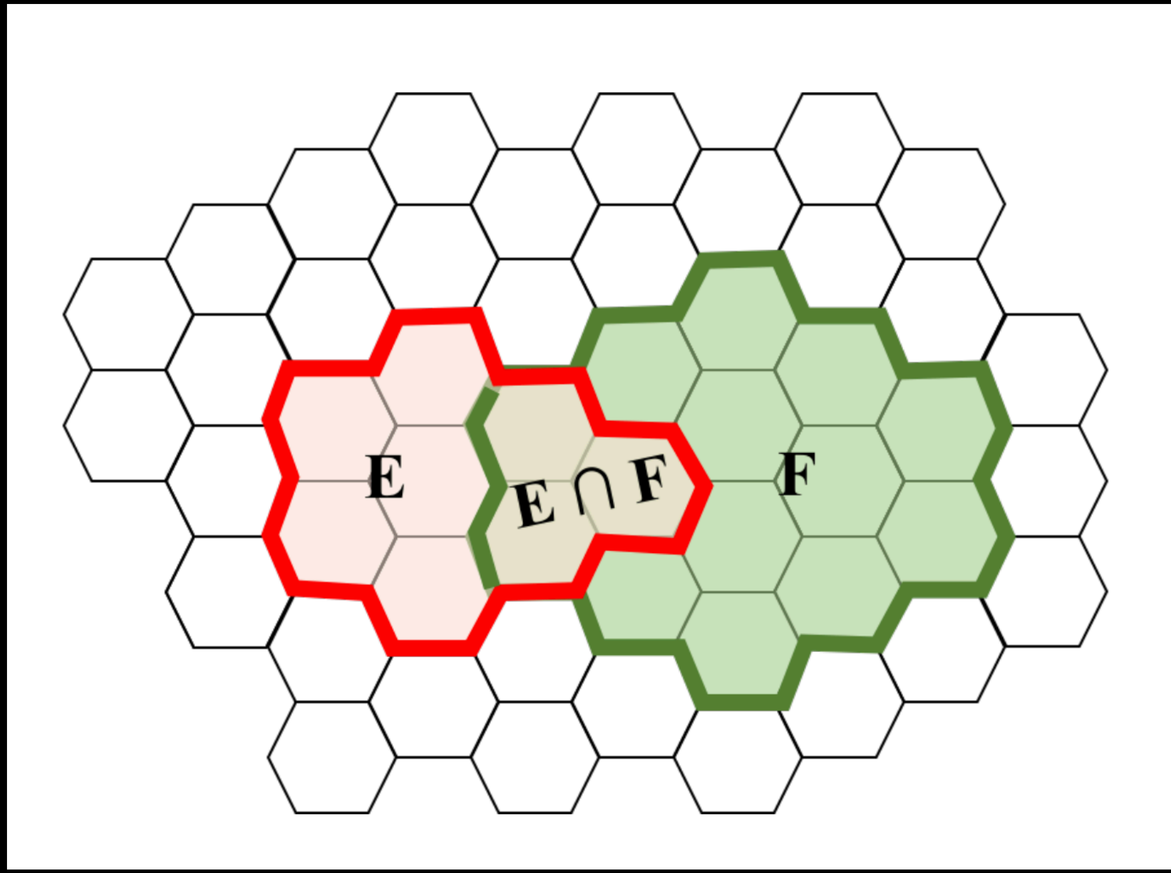
(use counting with distinct objects)

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

Conditional Probability

definition:

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



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

Chain Rule:

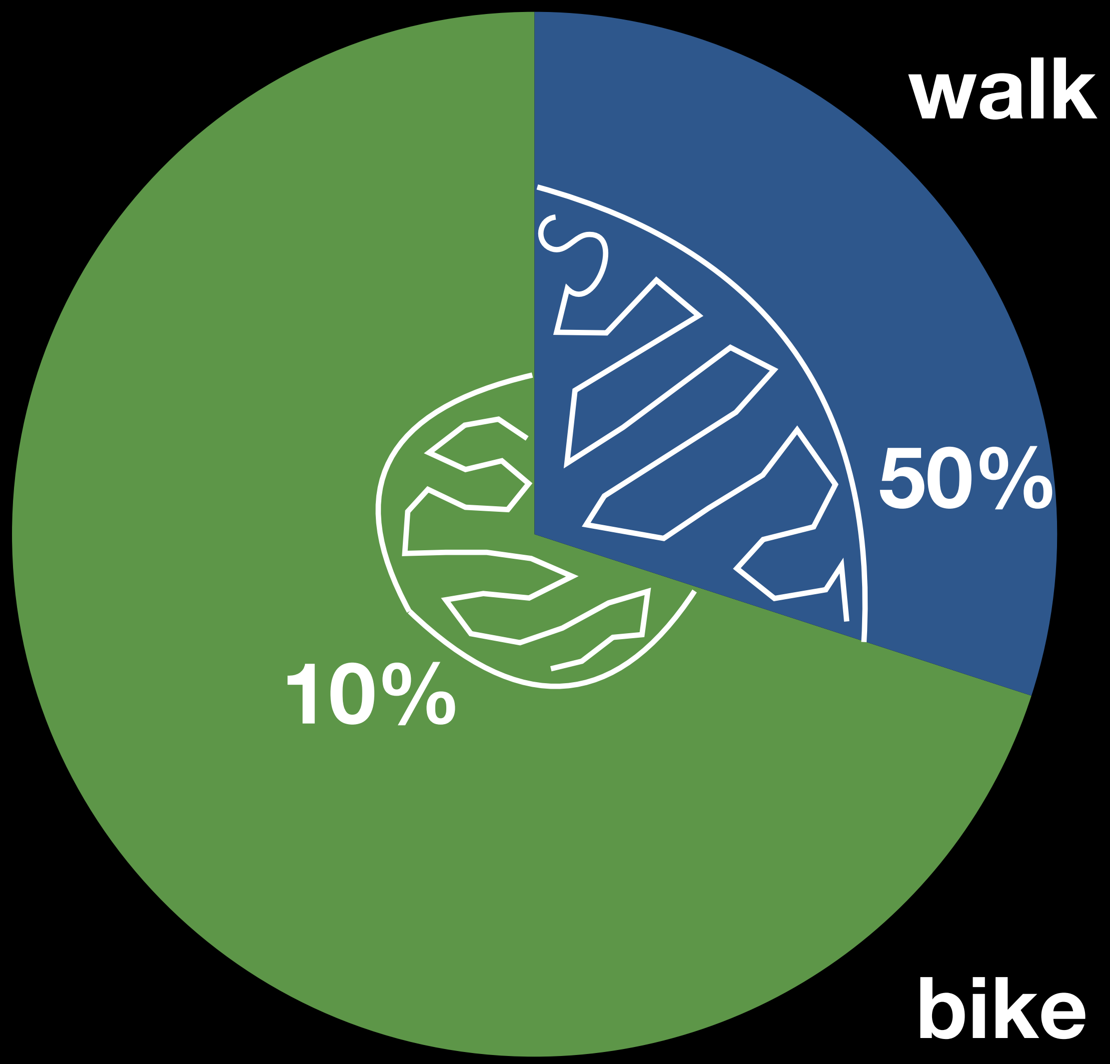
$$P(EF) = P(E \mid F)P(F)$$

Law of Total Probability

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

Event W = we walk to class. Event B = we bike = W^C.
Event L = we are late to class.
P(L | W) = 0.5, P(L | B) = 0.1.
P(W) = 0.3.
P(L) = ?

total shaded = ?% of whole

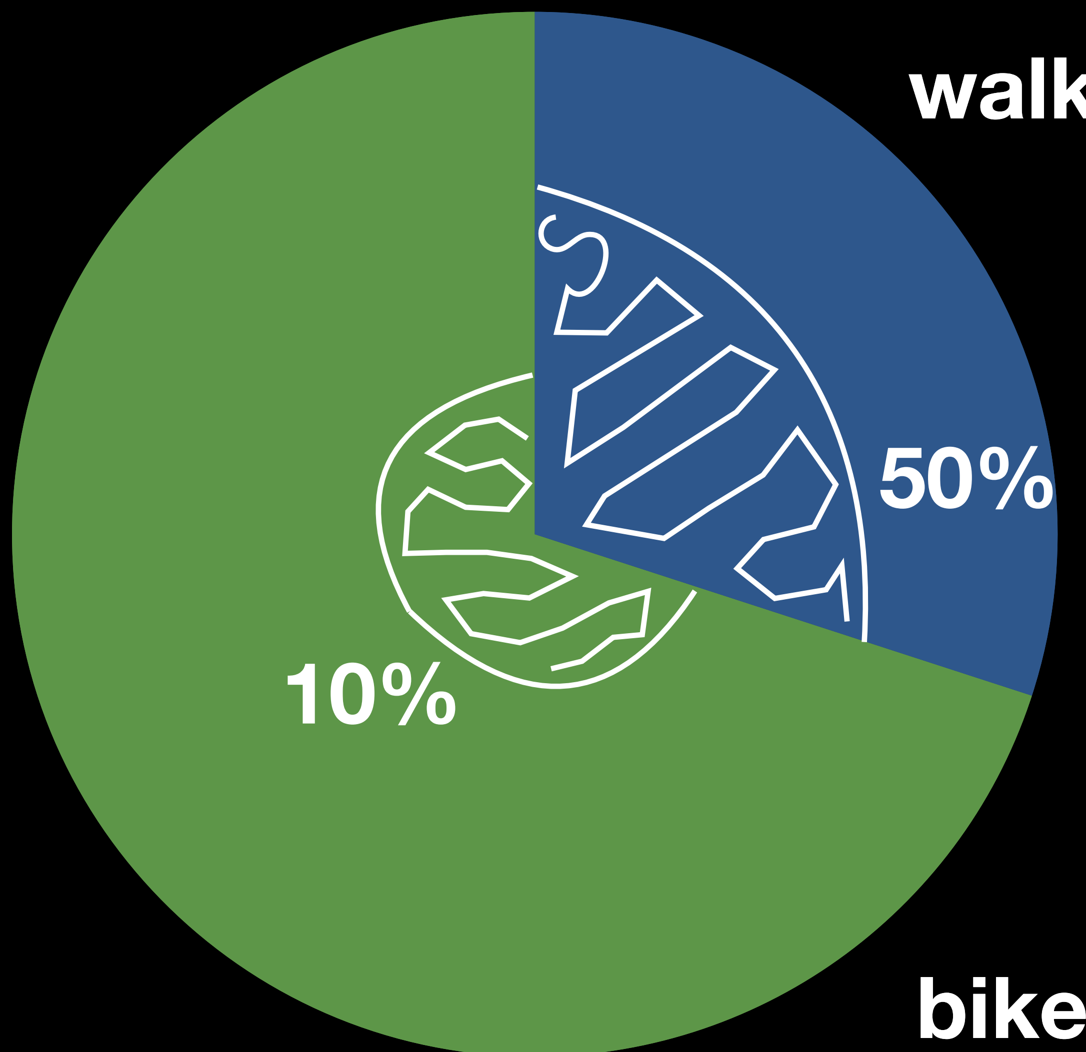


Law of Total Probability

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

Event W = we walk to class. Event B = we bike = W^C.
Event L = we are late to class.
P(L | W) = 0.5, P(L | B) = 0.1.
P(W) = 0.3.
P(L) = ?

total shaded = ?% of whole



$$P(L) = P(L \mid W)P(W) + P(L \mid W^C)P(W^C)$$

$$= (0.5)(0.3) + (0.1)(0.7)$$

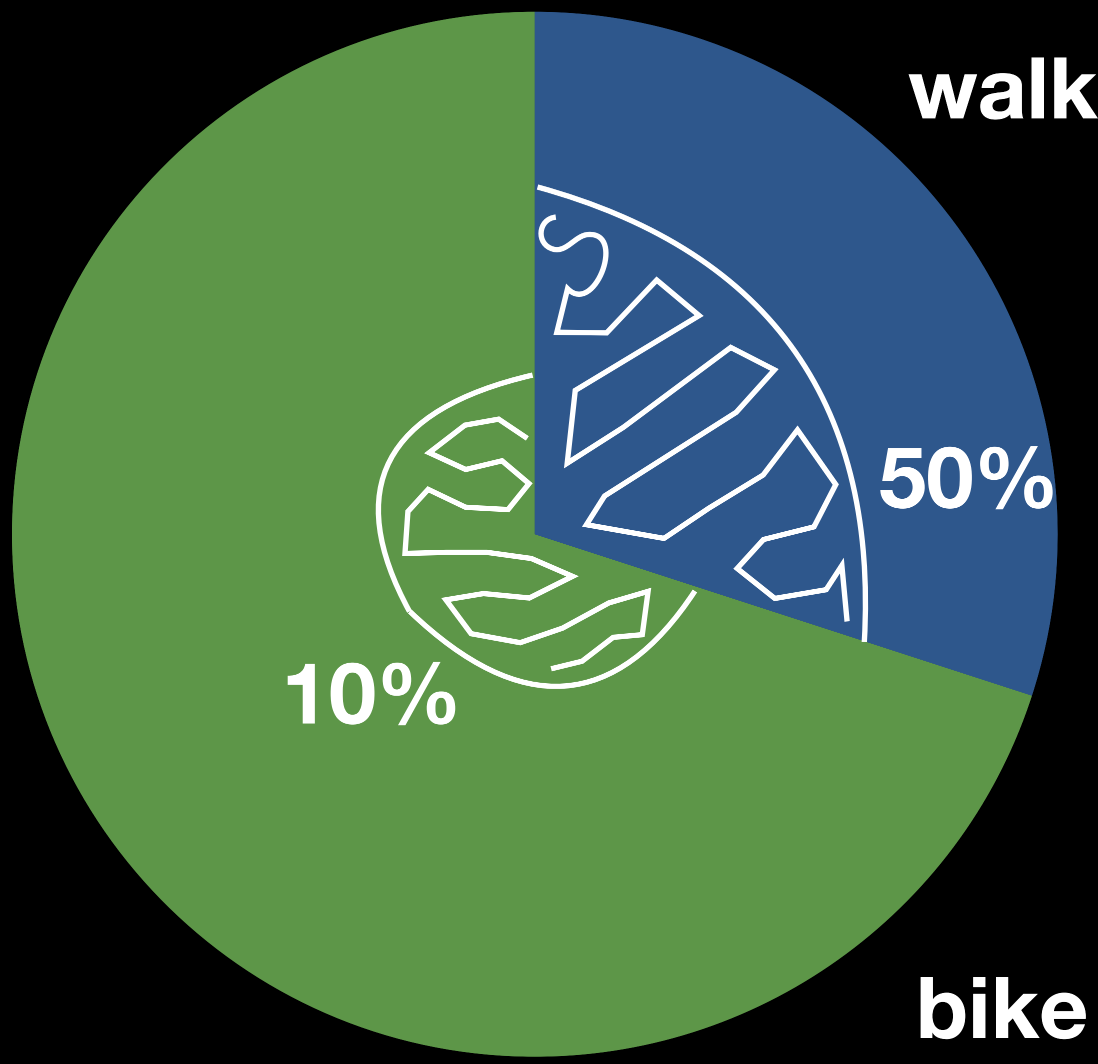
$$= 0.22$$

Law of Total Probability

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

Event W = we walk to class. Event B = we bike = W^C.
Event L = we are late to class.
P(L | W) = 0.5, P(L | B) = 0.1.
P(W) = 0.3.
P(L) = ?

what if we can bike, walk, or
take the Marguerite (> 2 options)?



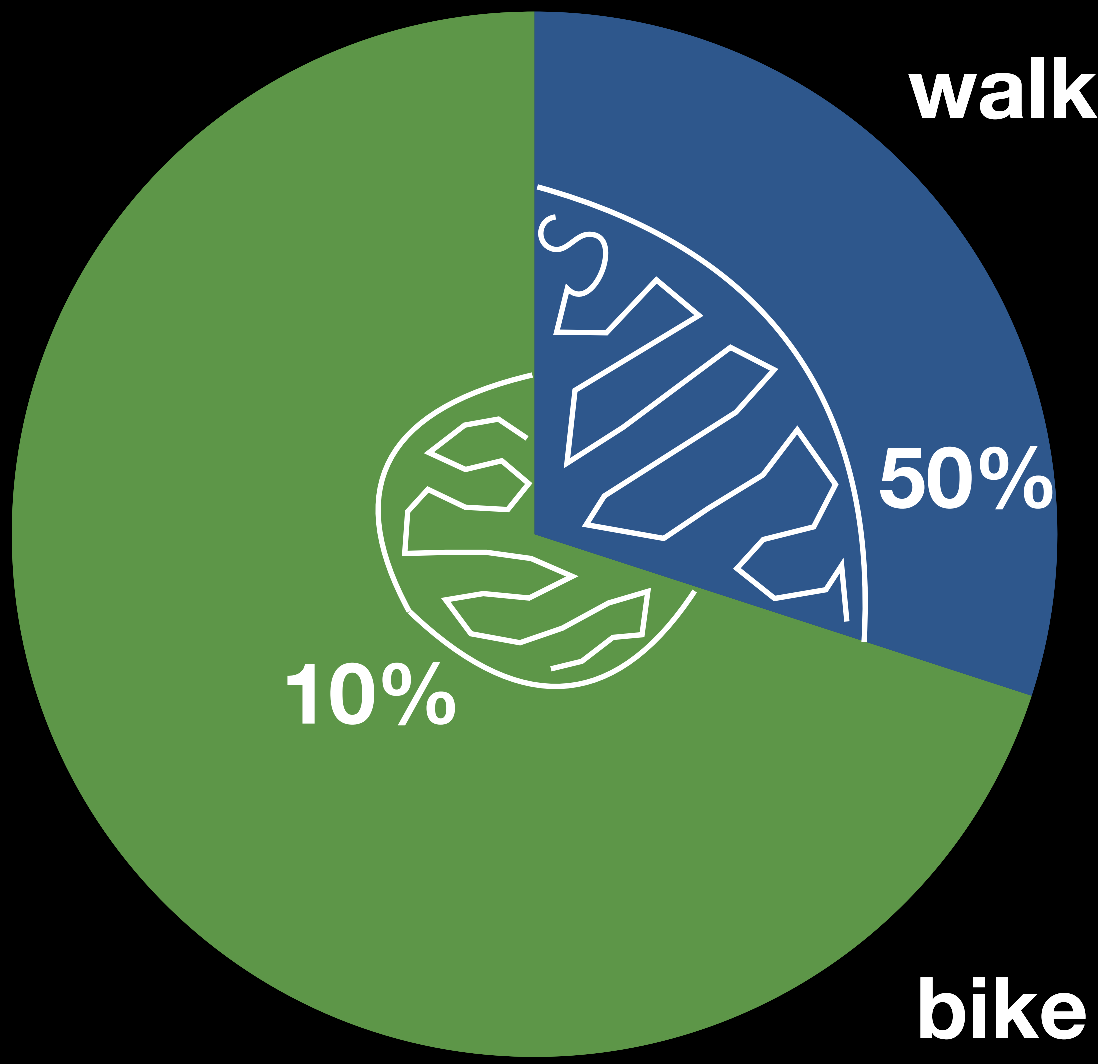
Law of Total Probability

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

Event W = we walk to class. Event B = we bike = W^C.
Event L = we are late to class.
P(L | W) = 0.5, P(L | B) = 0.1.
P(W) = 0.3.
P(L) = ?

what if we can bike, walk, or
take the Marguerite (> 2 options)?

events for “scale factors” must be:
- **mutually exclusive**, and
- **exhaustive**



Bayes' Rule

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

Bayes' Rule

posterior — likelihood — prior

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

normalization constant

Bayes' Rule

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

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

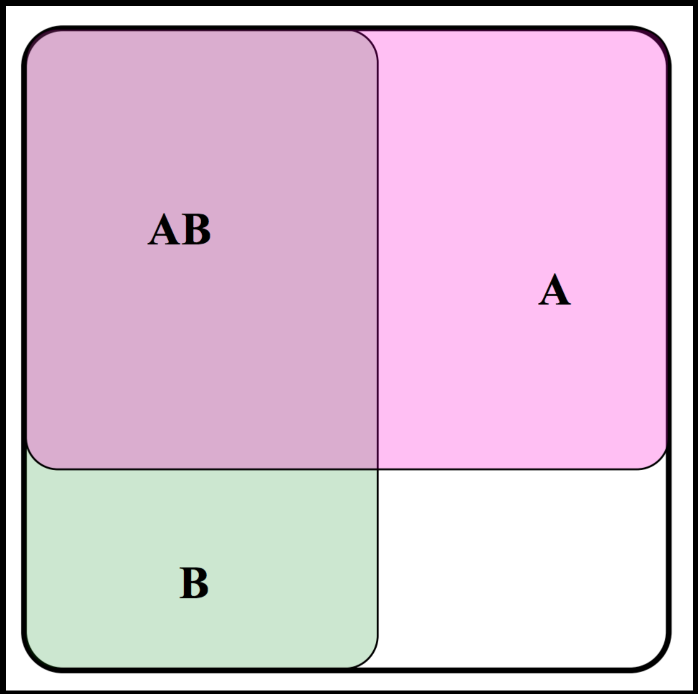
divide the event F into all the possible ways it can happen; use LoTP

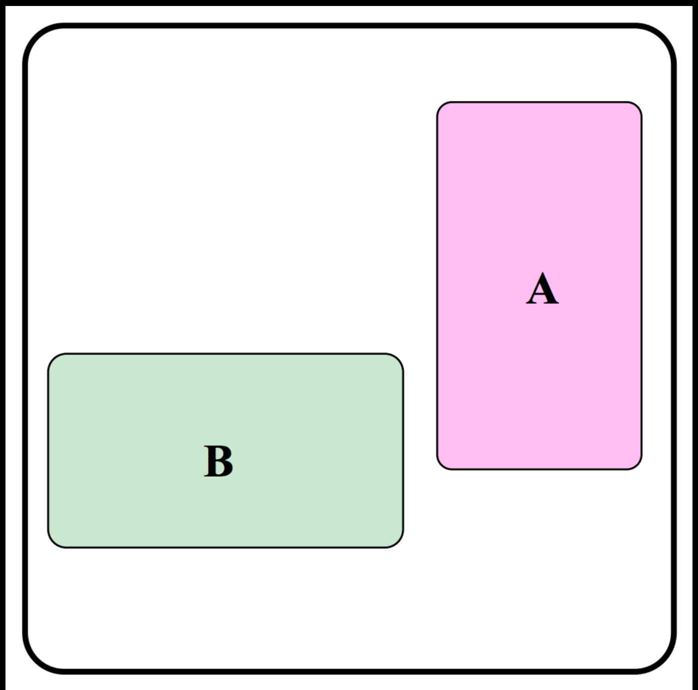
Old Principles, New Tricks

Name of Rule	**Original Rule**	**Conditional Rule**
First axiom of probability	$0 \leq P(E) \leq 1$	$0 \leq P(E \mid G) \leq 1$
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	$P(E \mid F) = \frac{P(F\|E)P(E)}{P(F)}$	$P(E \mid FG) = \frac{P(F\|EG)P(E\|G)}{P(F\|G)}$

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!

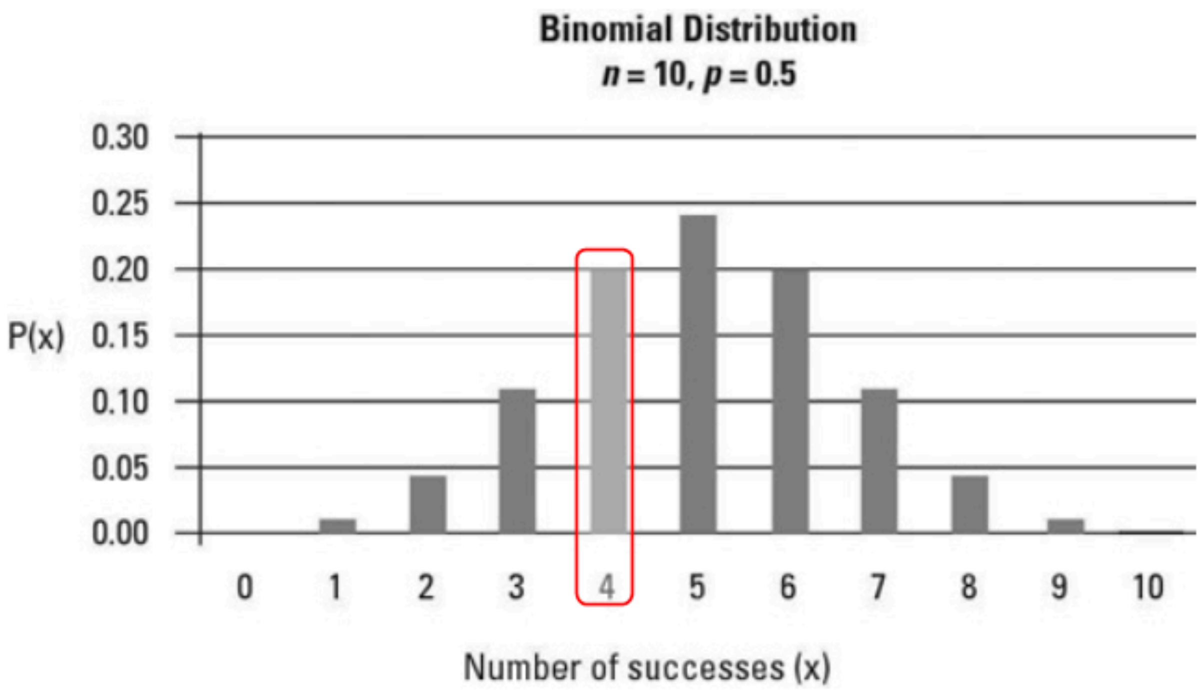
& vice versa


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

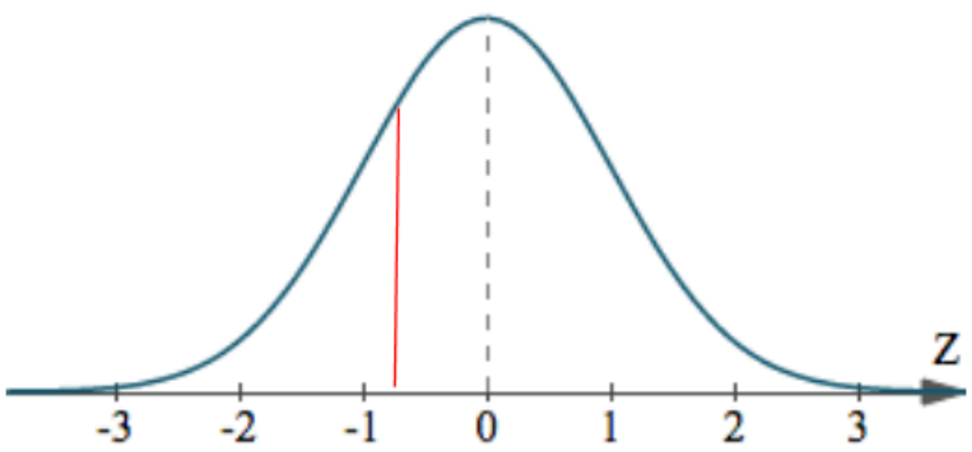
Discrete

PMF:



Continuous

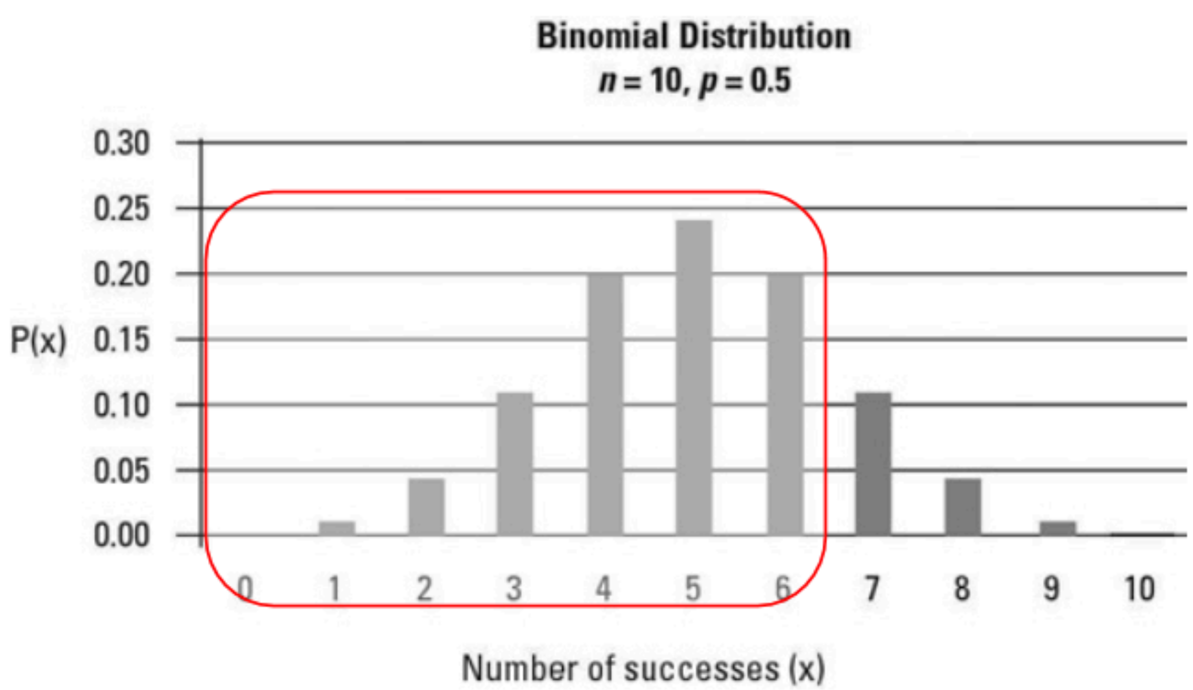
PDF:



Probability Distributions

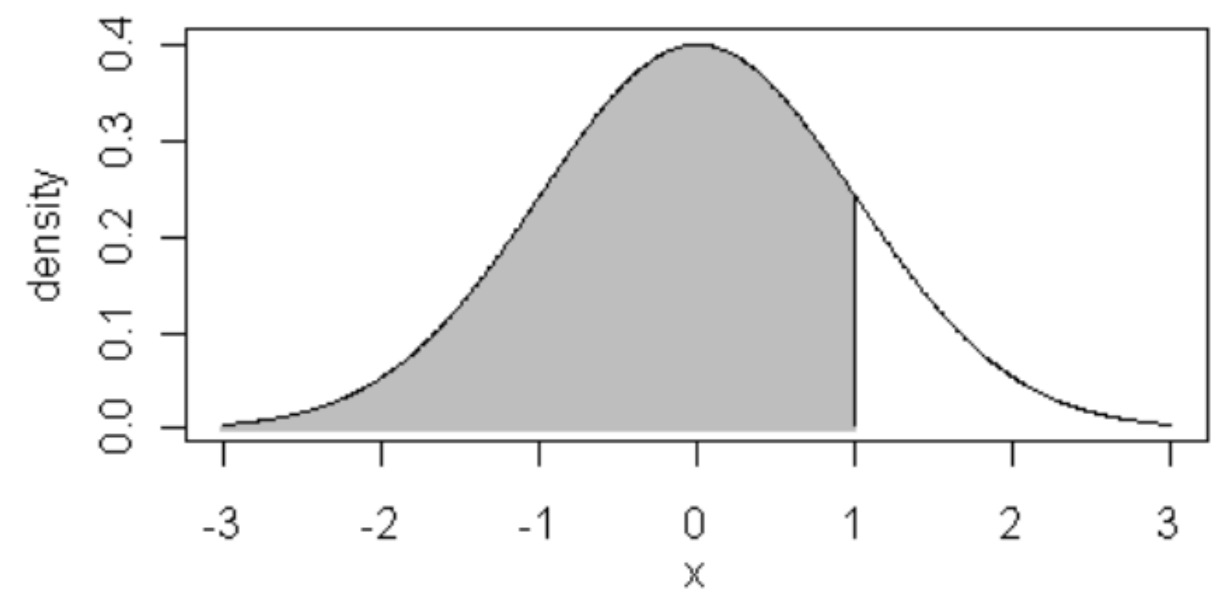
Discrete

CDF:



Continuous

CDF:



Expectation & Variance

Discrete definition

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

Continuous definition

$$E[X] = \int_x x * p(x)dx$$

Expectation & Variance

Discrete definition

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

Continuous definition

$$E[X] = \int_x x * p(x)dx$$

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

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

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

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

All our (discrete) friends

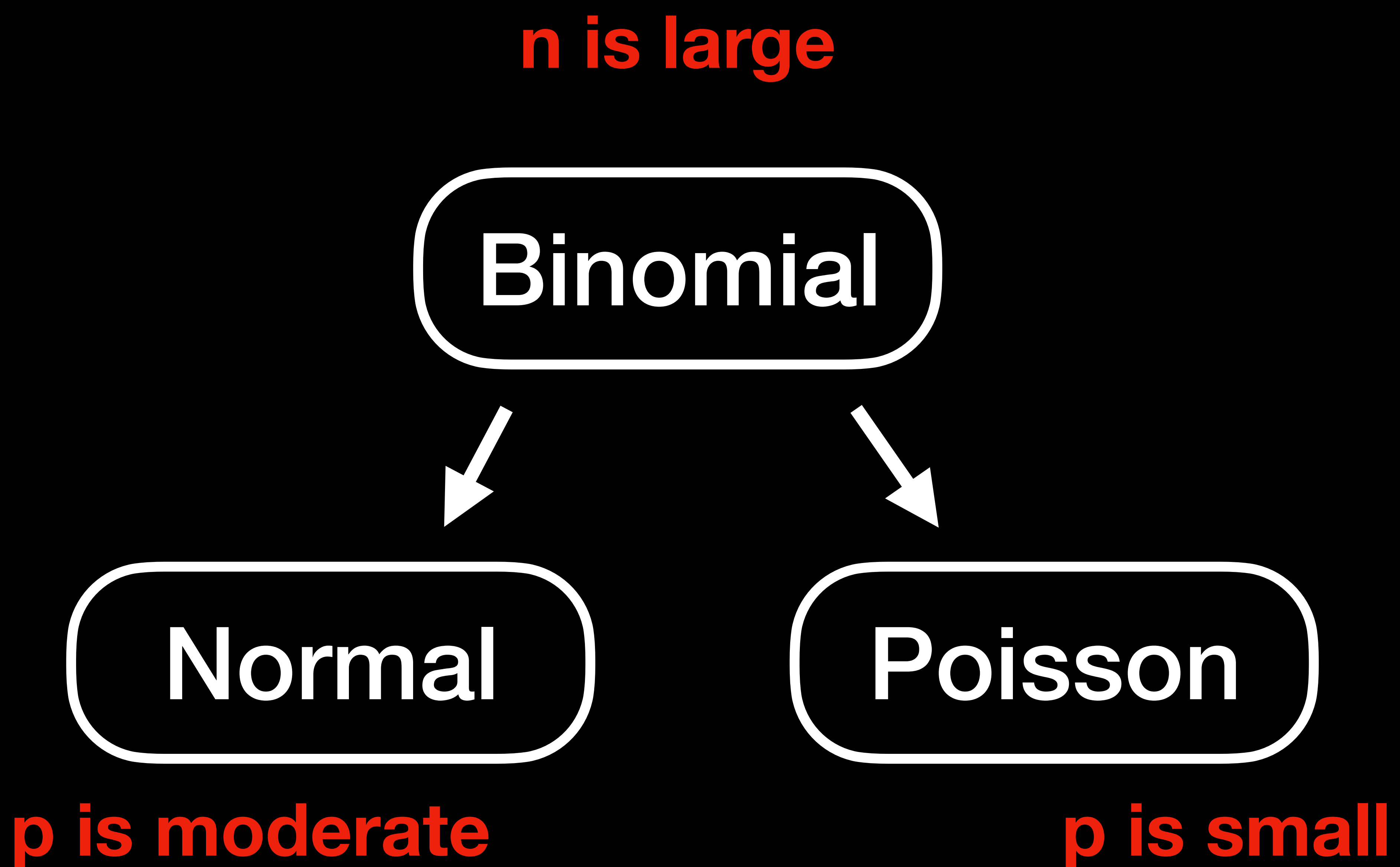
Ber(p)	Bin(n, p)	Poi(λ)	Geo(p)	NegBin (r, p)
P(X) = p	$\binom{n}{k} p^k (1-p)^{n-k}$	$\frac{\lambda^k e^{-\lambda}}{k!}$	$(1-p)^{k-1} p$	$\binom{k-1}{r-1} p^r (1-p)^{k-r}$
E[X] = p	E[X] = np	E[X] = λ	E[X] = 1 / p	E[X] = r / p
Var(X) = p(1-p)	Var(X) = np(1-p)	Var(X) = λ	$\frac{1-p}{p^2}$	$\frac{r(1-p)}{p^2}$
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

Approximations

When can we approximate a binomial?

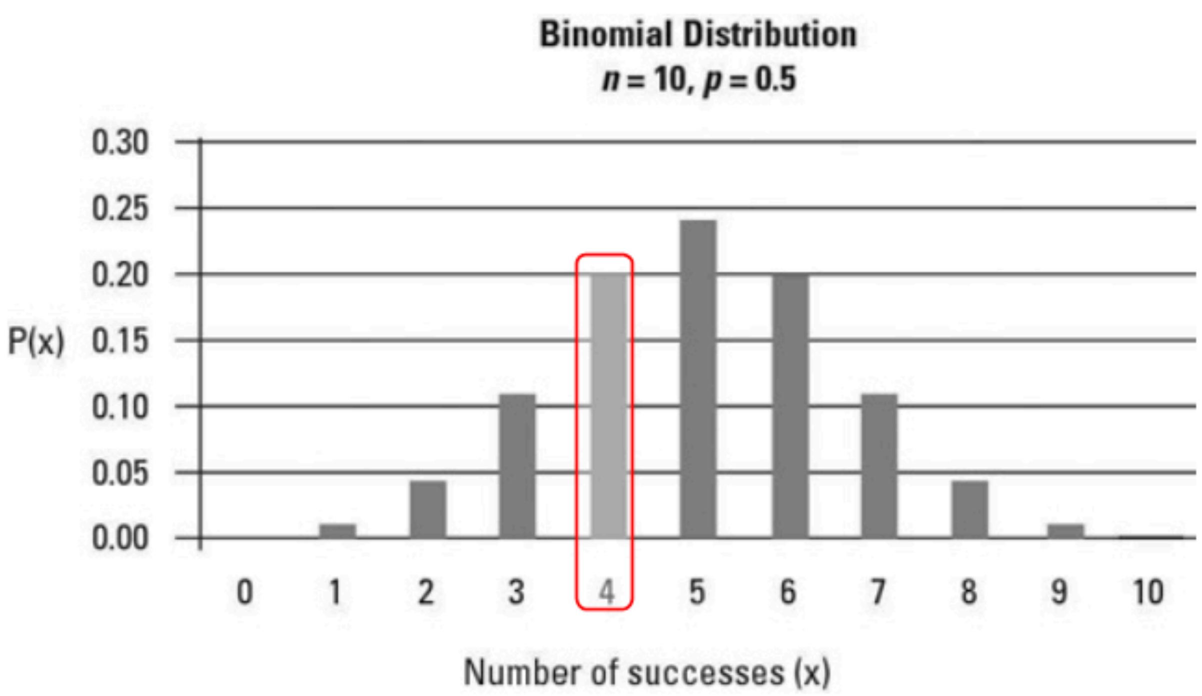


Continuity Correction

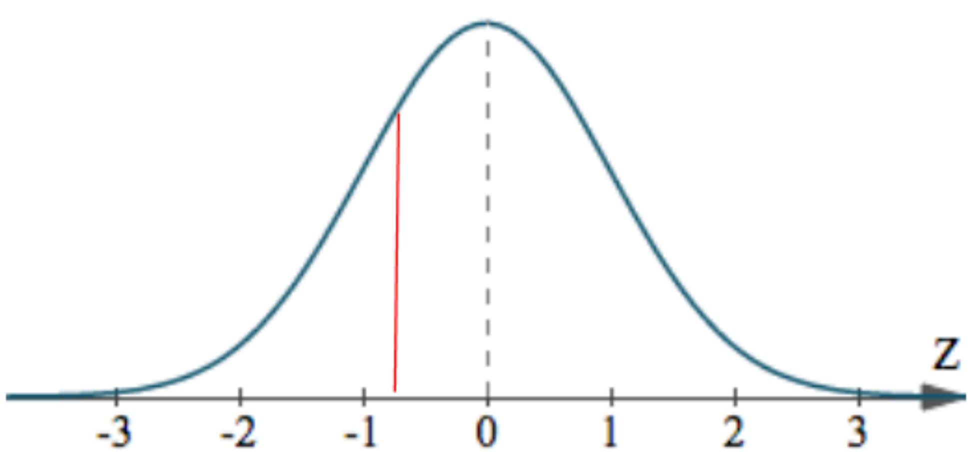
Discrete

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:

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

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

- For joint distributions to be independent, both their joint probability density function must be factorable and the bounds of the variables must be separable.

Convolutions

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

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

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

$$f_{X+Y}(a) = \int_{y=-\infty}^{\infty} f_X(a-y) f_Y(y) dy$$ **(general case)**

Relationships Between Random Variables

Covariance

the extent to which the deviation of one variable from its mean matches the deviation of the other from its mean

$$Cov(X, Y) = E[XY] - E[Y]E[X]$$

Correlation

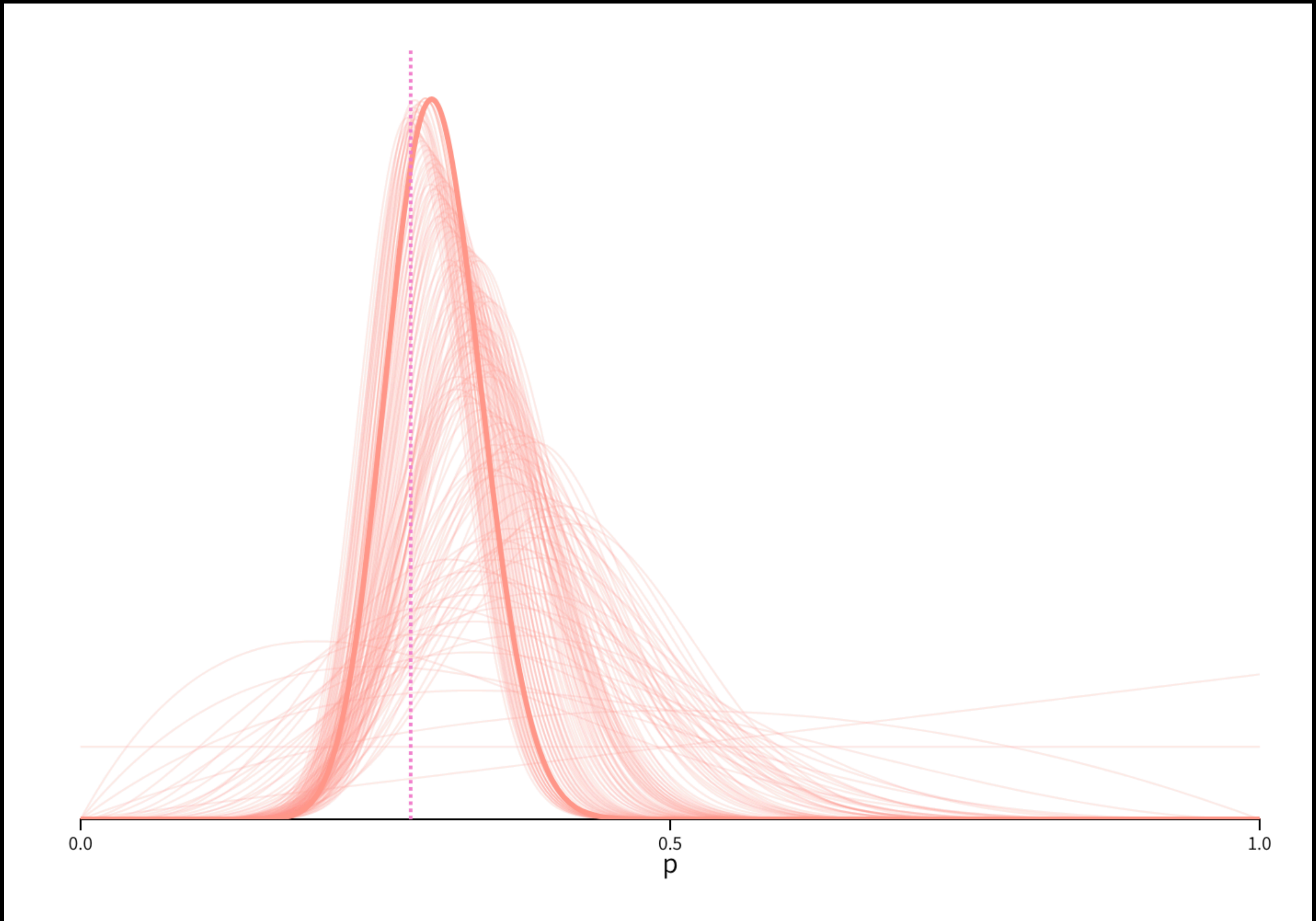
covariance normalized by the variance of each variable
(cancels the units out)

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

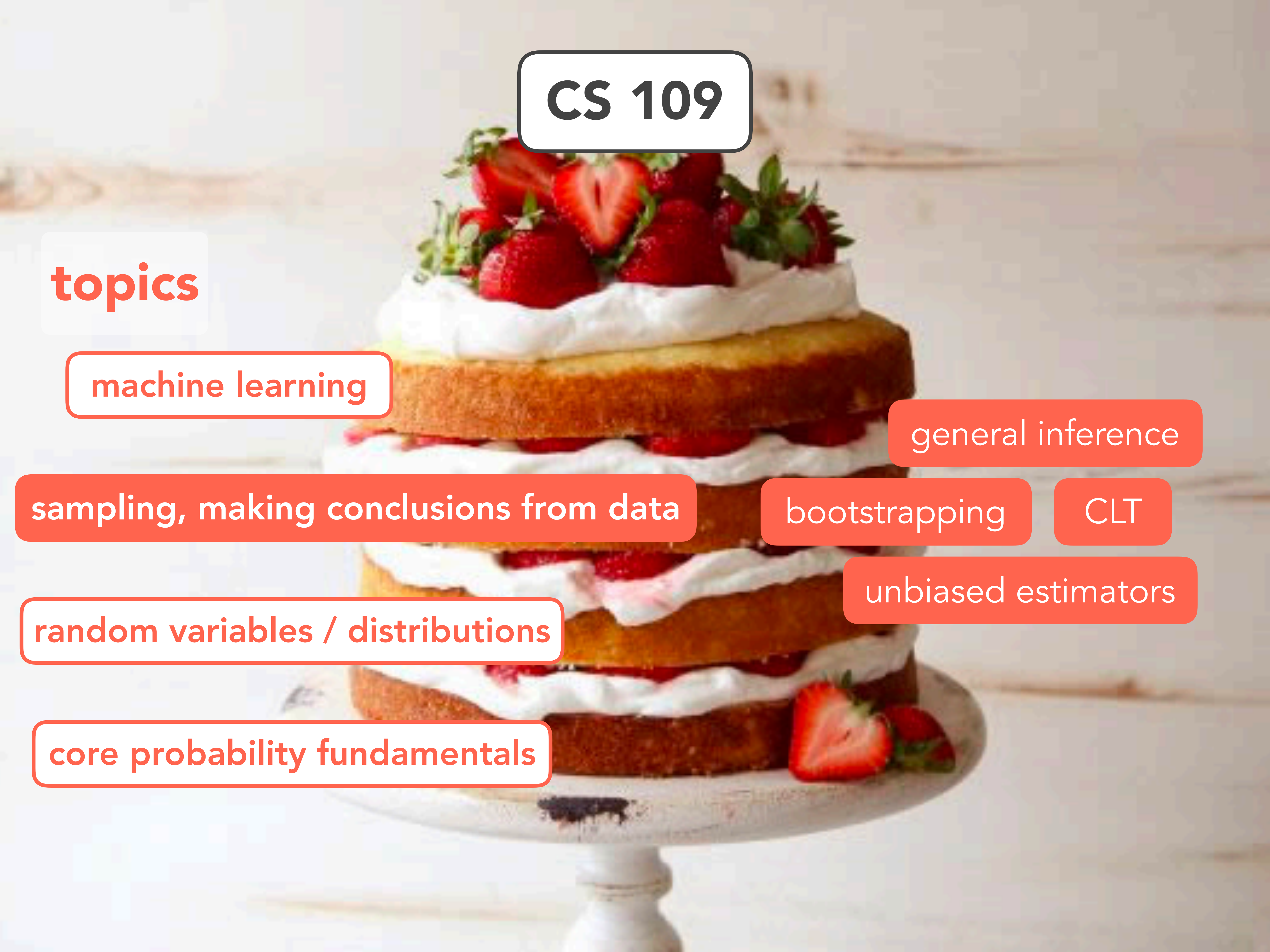
if two random variables have a covariance of 0, they are independent
(but not necessarily true the other way around!)

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

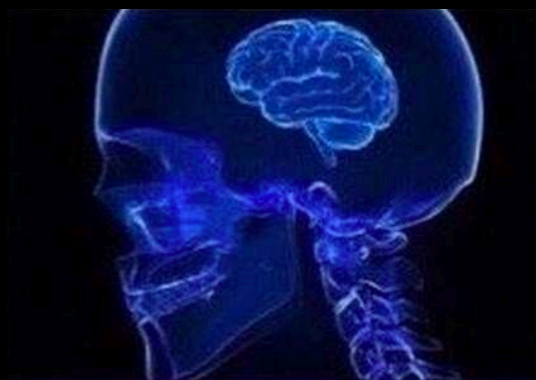
Sampling From Populations

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….

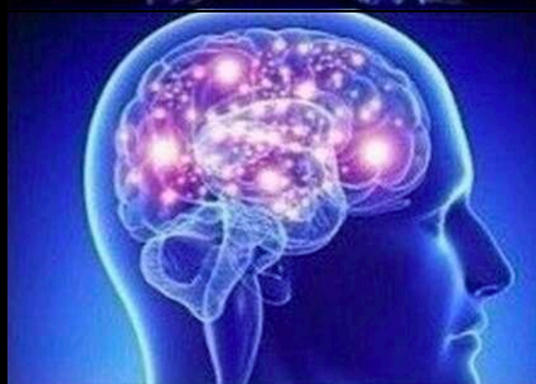
uh no

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!

If we assume the underlying distribution is normal

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!)

$$\bar{X} = \sum_{i=1}^{n} \frac{X_i}{n}$$

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

makes the estimate unbiased

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

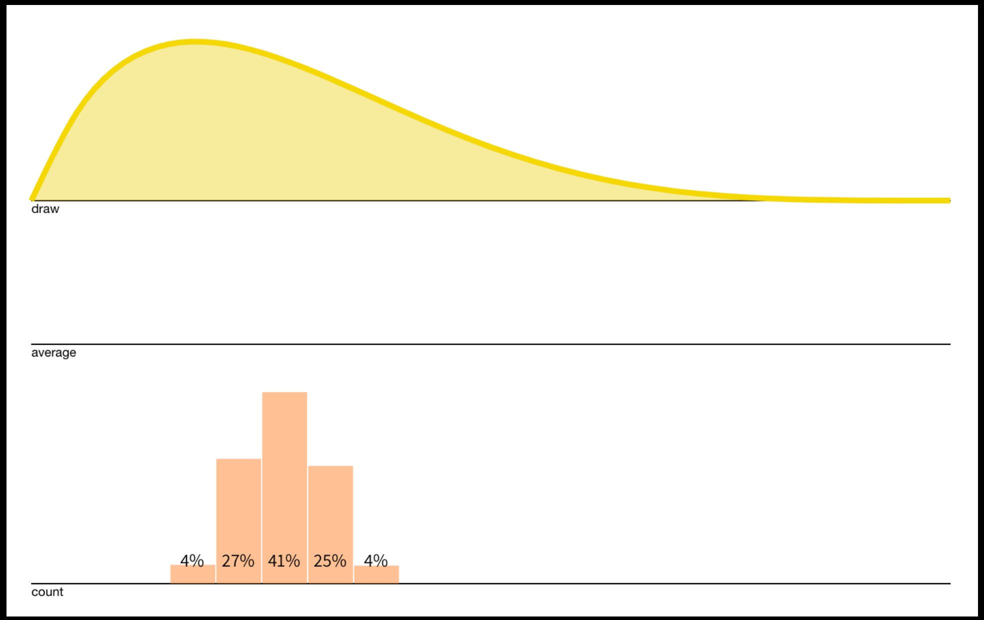
Taking Many Samples

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

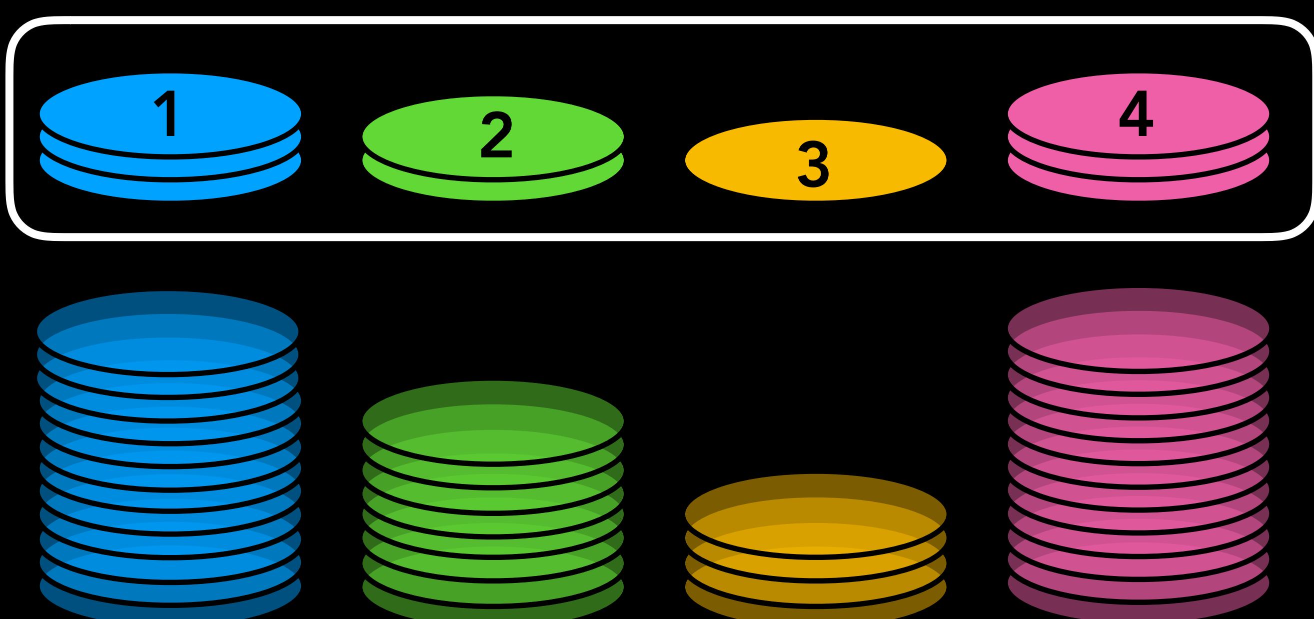
Bootstrapping: Simulating Many Samples From One

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

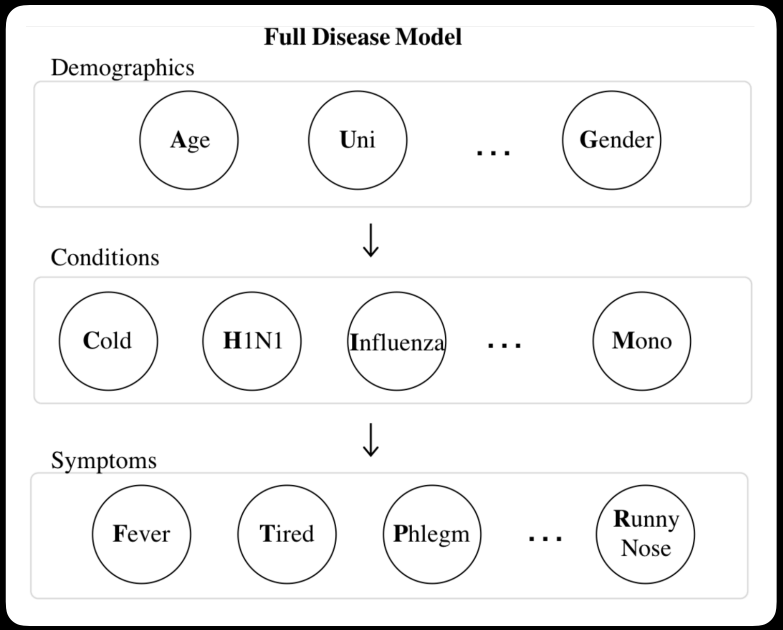
insight

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

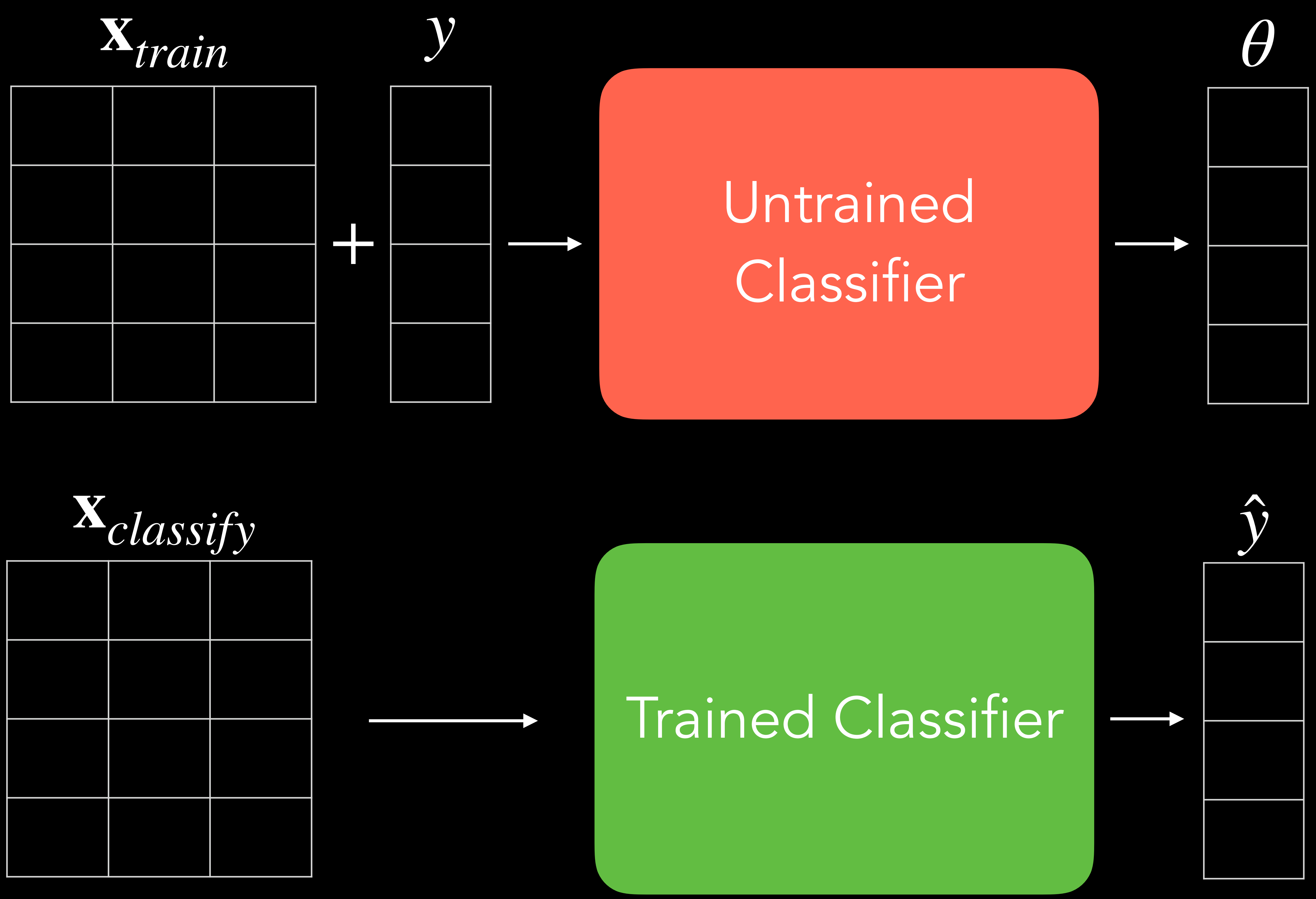
we can calculate any conditional probability of specific variable assignments by simply counting the particles that match what we're looking for

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

we can also generate samples where we hold some values fixed (MCMC)


CS 109
topics
machine learning
parameter estimation
classifiers
deep learning
sampling, making conclusions from data
random variables / distributions
core probability fundamentals

Classifiers



Parameter Estimation

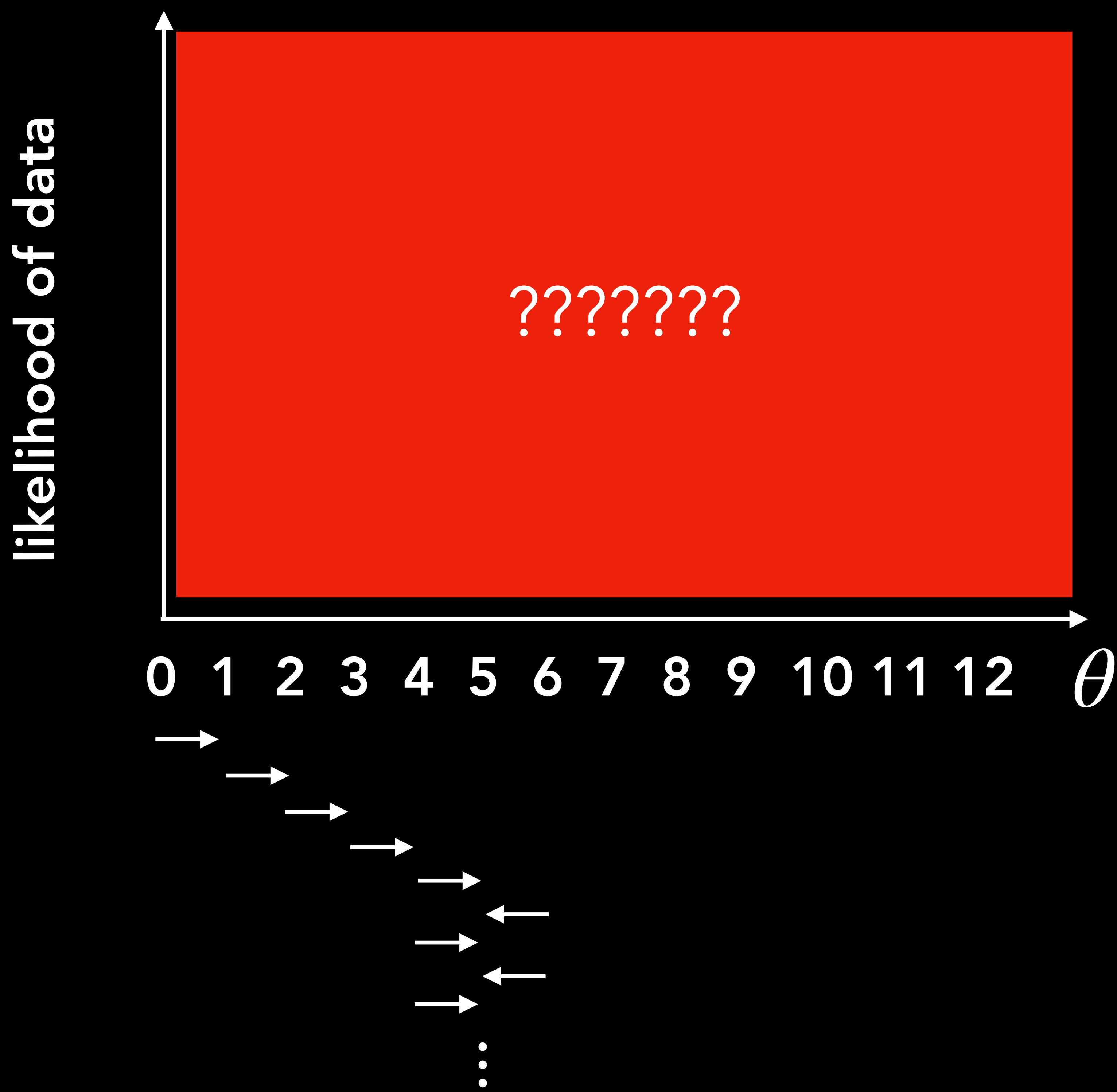
Maximum Likelihood Estimation

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

Gradient Ascent



step size

$$\eta = 1$$

step direction

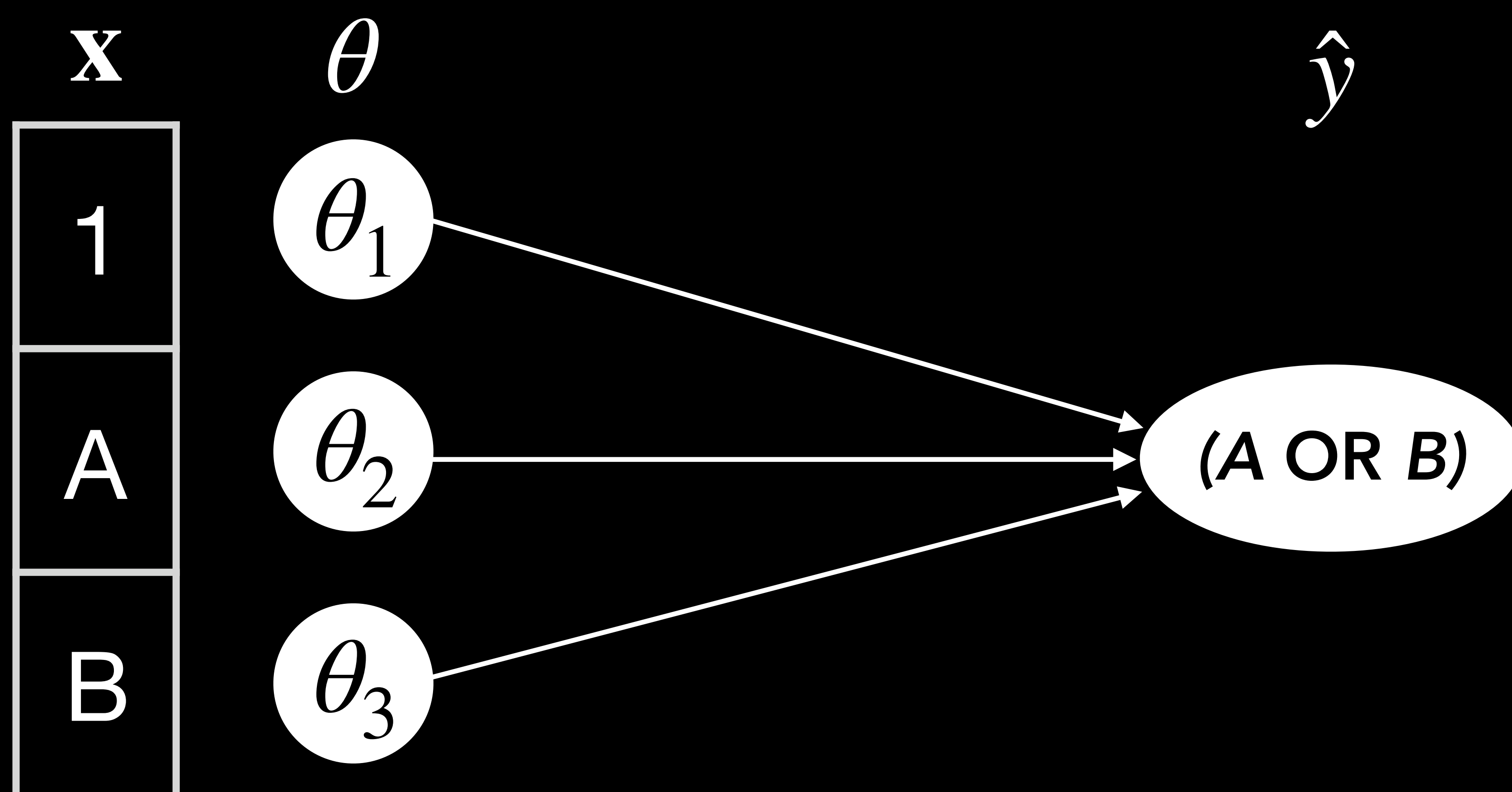
$$= \textbf{sign}\left[\frac{\partial\ \textbf{prob}}{\partial\theta}\right]$$

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 $\boldsymbol{\theta}$ such that the probability of our data is maximized
Learn (from data) estimates for $\hat{P}(Y=y), \hat{P}(X_i = x_i \mid Y = y)$: $\hat{P}(x_i \mid y) = \frac{(\textbf{ex. where } X_i = x_i and Y = y) + 1}{(\textbf{ex. where } Y = y) + 2}$ $\hat{P}(Y=y) = \frac{\textbf{ex. where } Y = y}{\textbf{total examples}}$	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

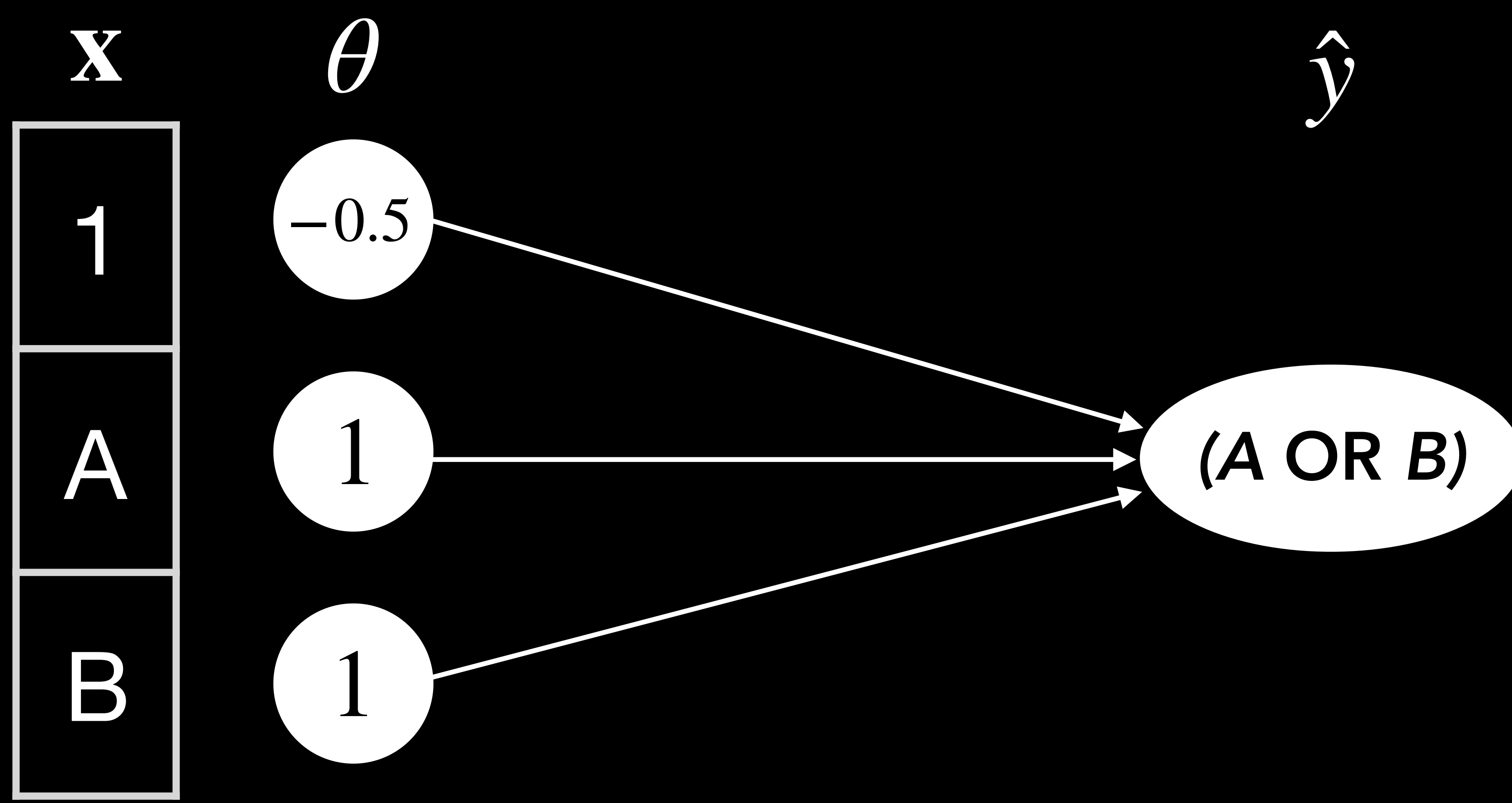
one neuron (logistic regression model)



What weights do we have to learn for θ_1, θ_2, θ_3 to perfectly classify data of the form (*A* OR *B*)?

Neural Networks

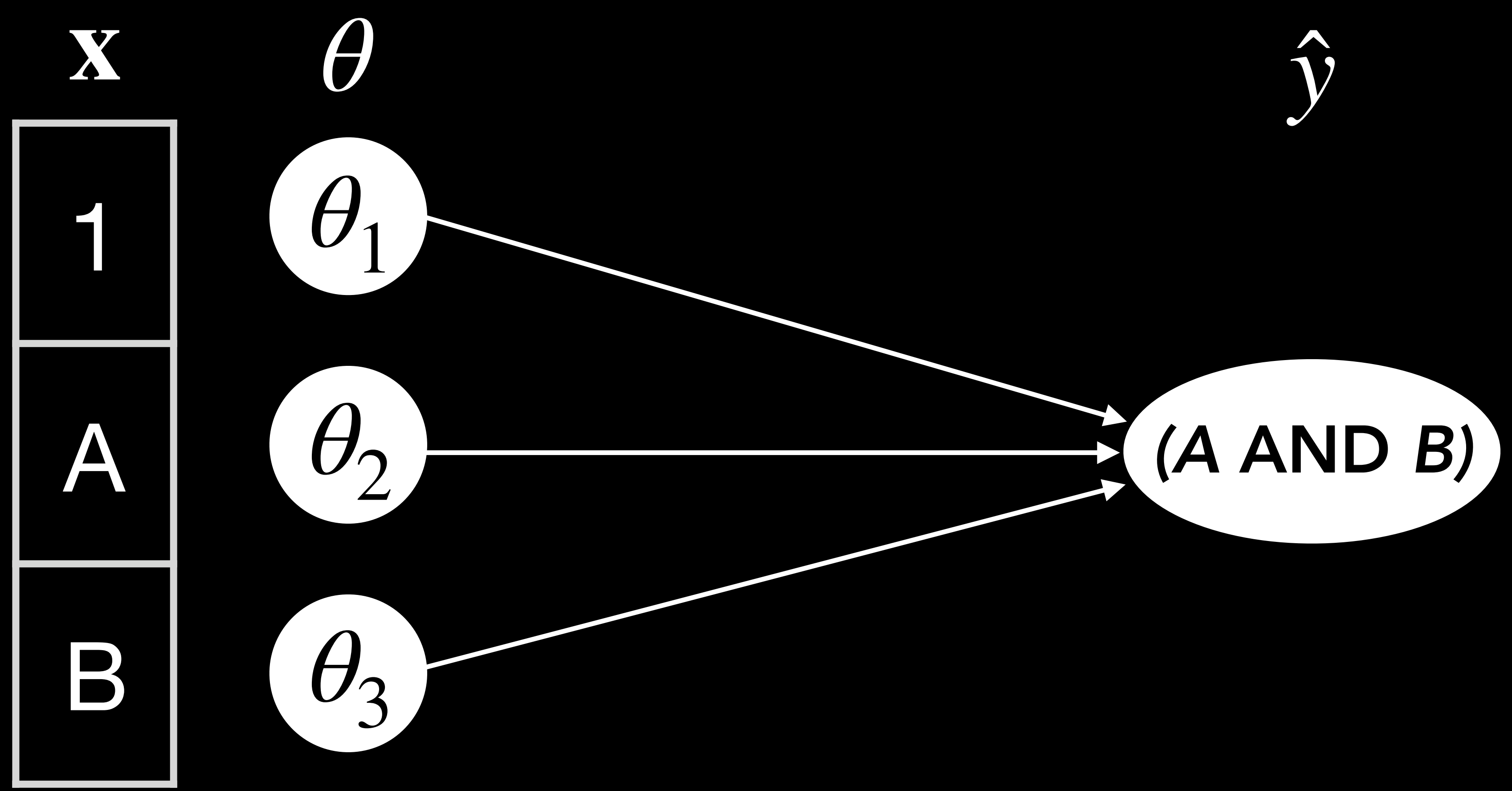
one neuron (logistic regression model)



What weights do we have to learn for θ_1, θ_2, θ_3 to perfectly classify data of the form (*A* OR *B*)?

Neural Networks

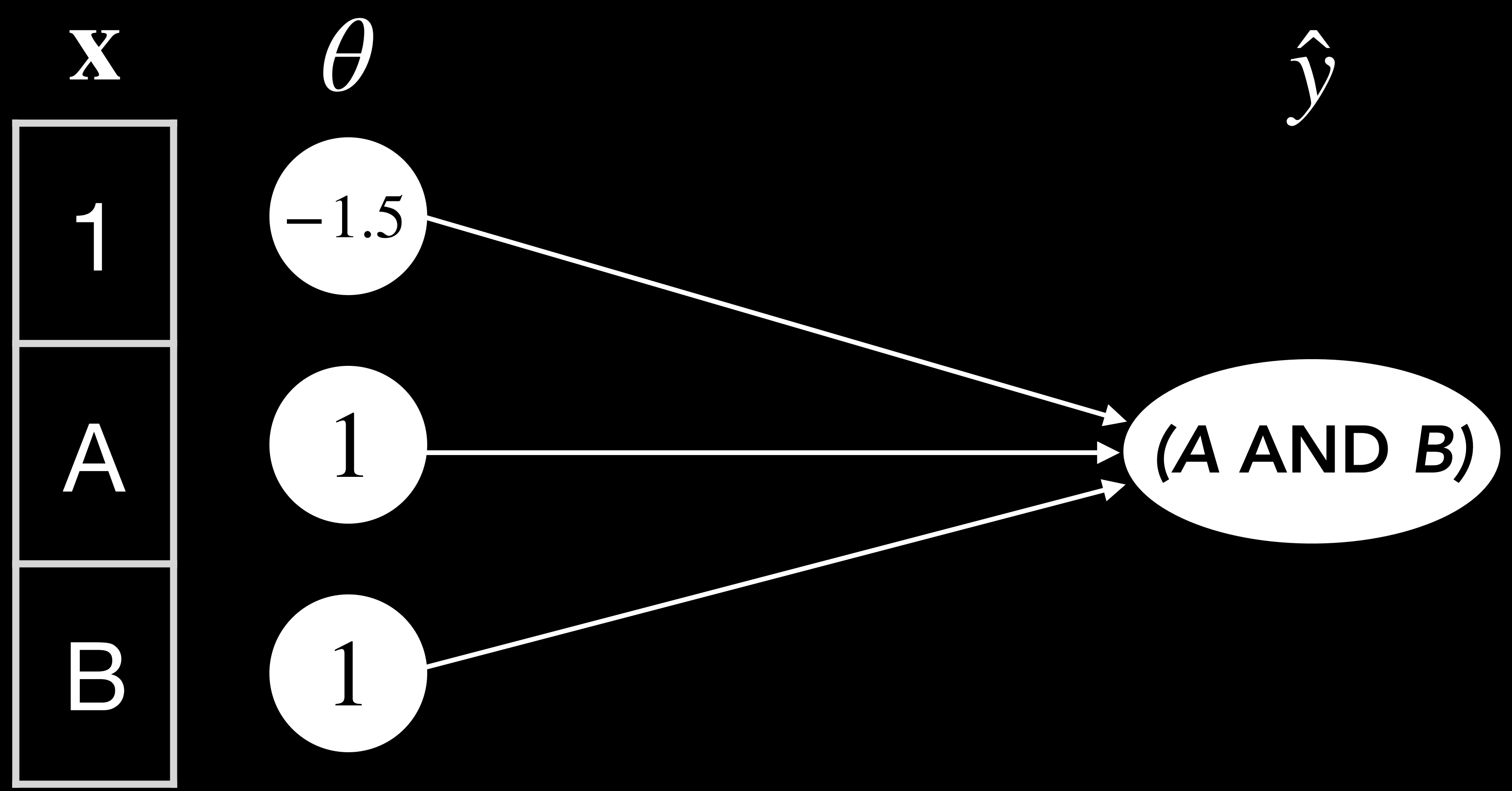
one neuron (logistic regression model)



What weights do we have to learn for θ_1, θ_2, θ_3 to perfectly classify data of the form (*A* AND *B*)?

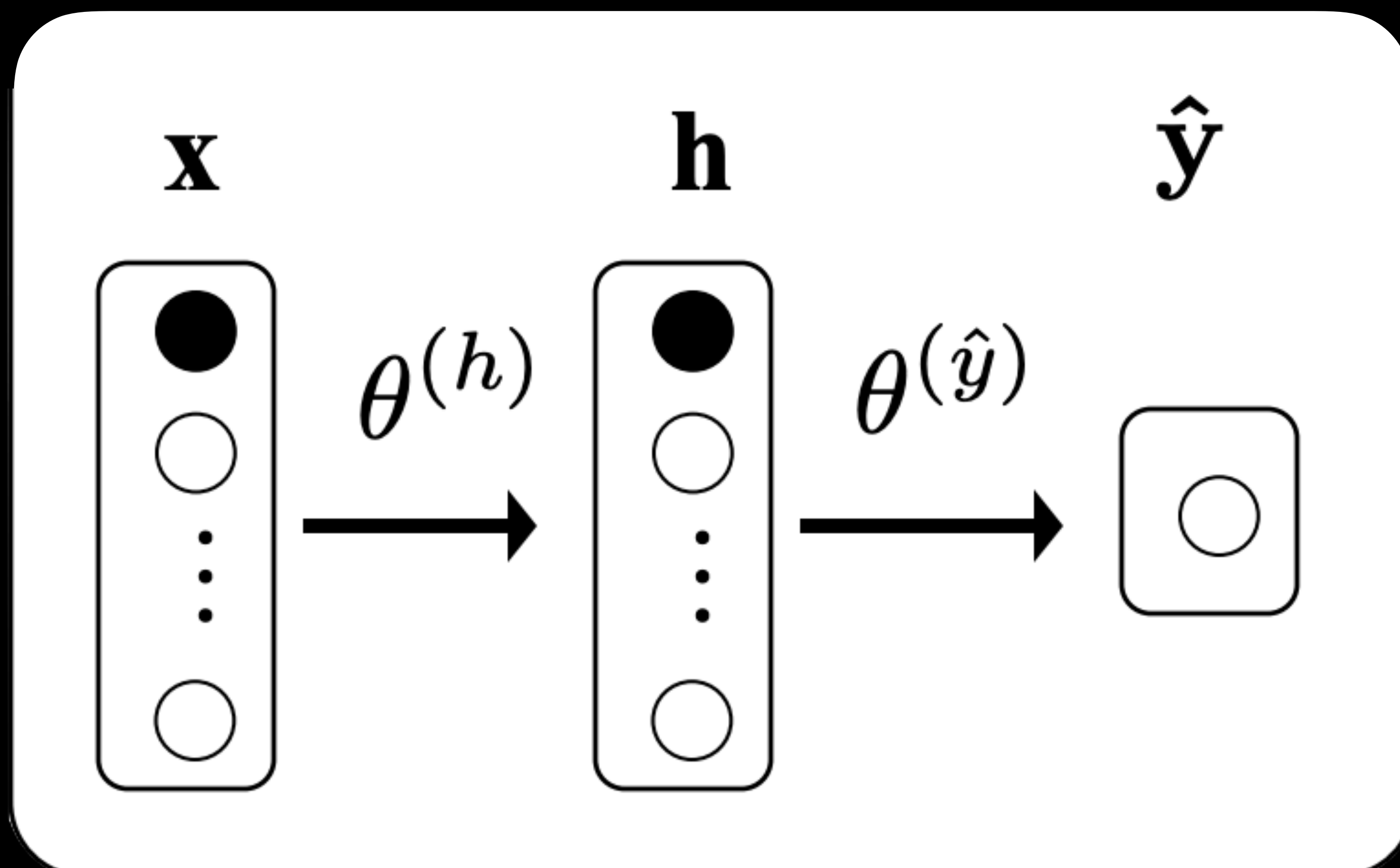
Neural Networks

one neuron (logistic regression model)



What weights do we have to learn for θ_1, θ_2, θ_3 to perfectly classify data of the form (*A* AND *B*)?

Neural Networks

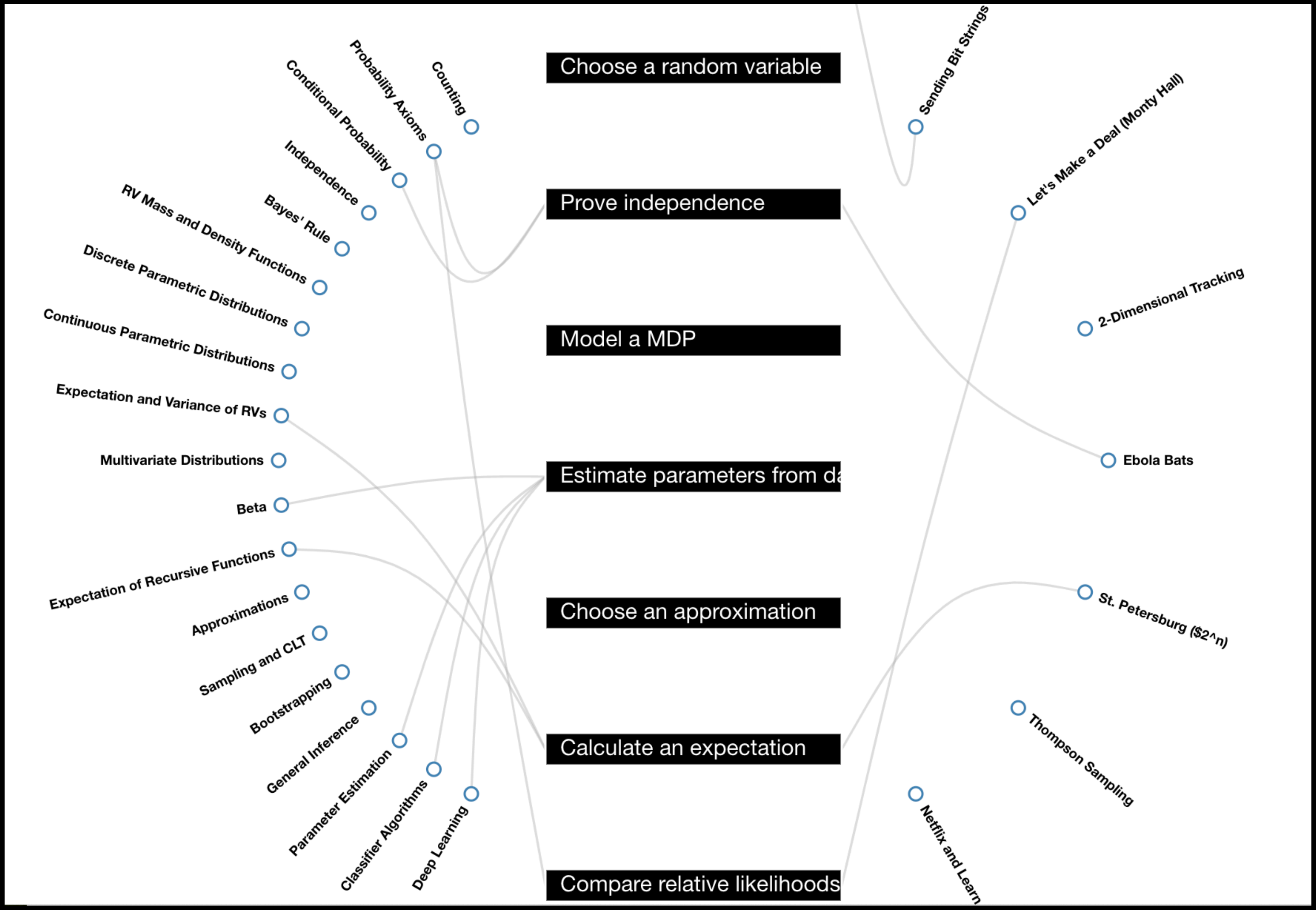


TensorFlow

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: $LL(\theta) = \sum_{i=0}^{n} y^{(i)}(\log \hat{y}^{(i)}) + (1 - \hat{y}^{(i)}) \log [1 - \hat{y}^{(i)}]$
3. Find partial derivative of LL with respect to each theta: $\frac{\partial LL(\theta)}{\partial \theta_j^{(\hat{y})}} = \frac{\partial LL(\theta)}{\partial \hat{y}} \cdot \frac{\partial \hat{y}}{\partial \theta_j^{(\hat{y})}}$ $\quad \frac{\partial LL(\theta)}{\partial \theta_{i,j}^{(h)}} = \frac{\partial LL(\theta)}{\partial \hat{y}} \cdot \frac{\partial \hat{y}}{\partial \mathbf{h}_j} \cdot \frac{\partial \mathbf{h}_j}{\partial \theta_{i,j}^{(h)}}$

   use the chain rule!

Concept Organizer



Check out cs109.stanford.edu > Handouts > Big Picture! (live later tonight)


Good luck on the final!