Announcements

Lecture

Today:	Course review session
Wednesday:	Beyond CS109
Friday:	No lecture

Homework

Contest (optional) due:	Tonight, 11:59pm
PS6 on-time deadline:	Wednesday 12/4, 1pm
PS6 late deadline:	Friday 12/6, 1pm

Participation

Concept check:	Tomorrow 1pm
Section:	regularly scheduled

Final exam announcement

Final exam

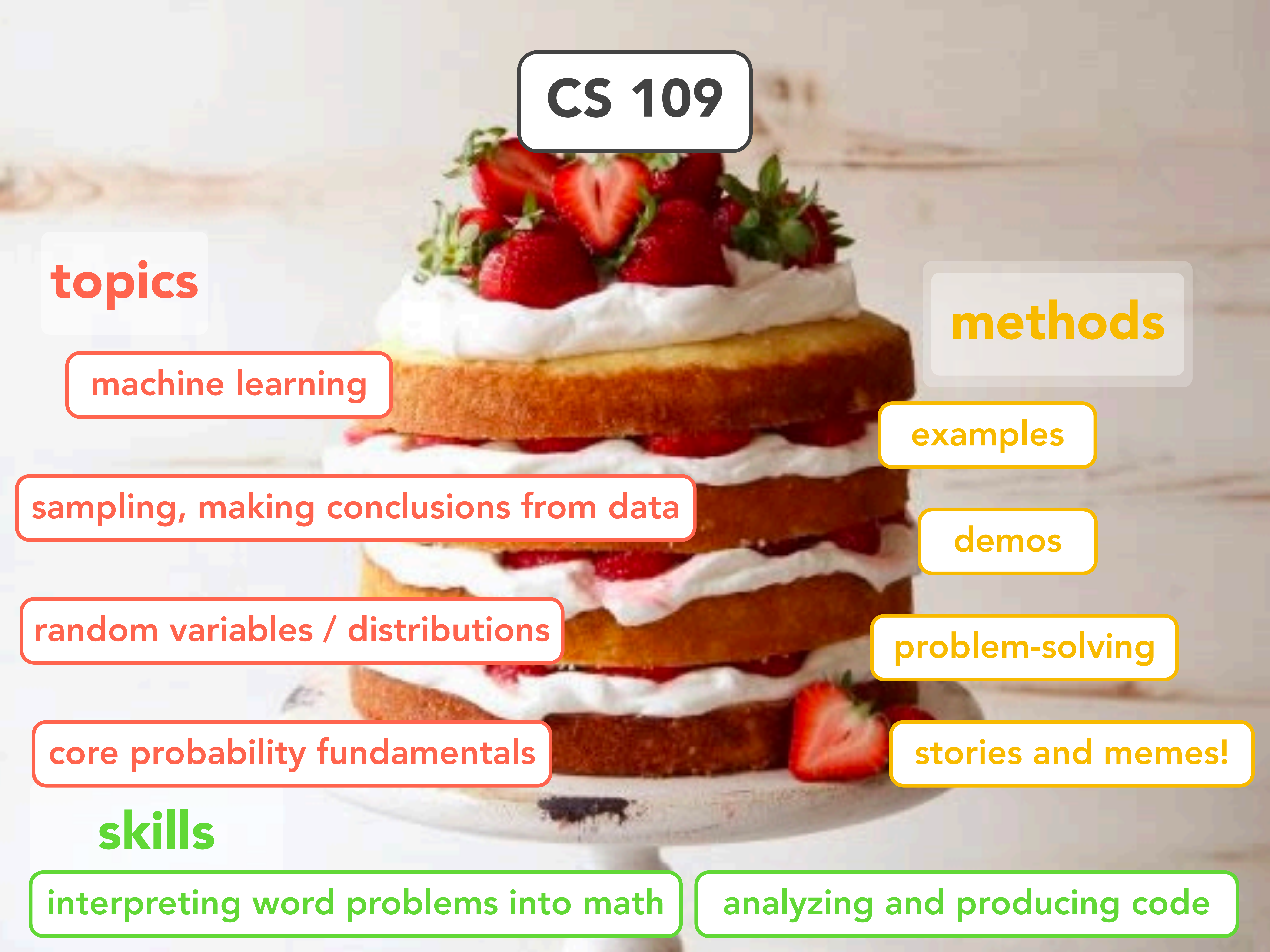
When:	Wednesday, December 11th, 3:30pm-6:30pm
Where:	CEMEX Auditorium
Not permitted:	book/computer/calculator
Permitted:	six 8.5"x11" double-sided sheets of notes
Covers:	Everything (up to Lecture 27)
Practice:	http://web.stanford.edu/class/cs109/exams/final.html
Review session:	today

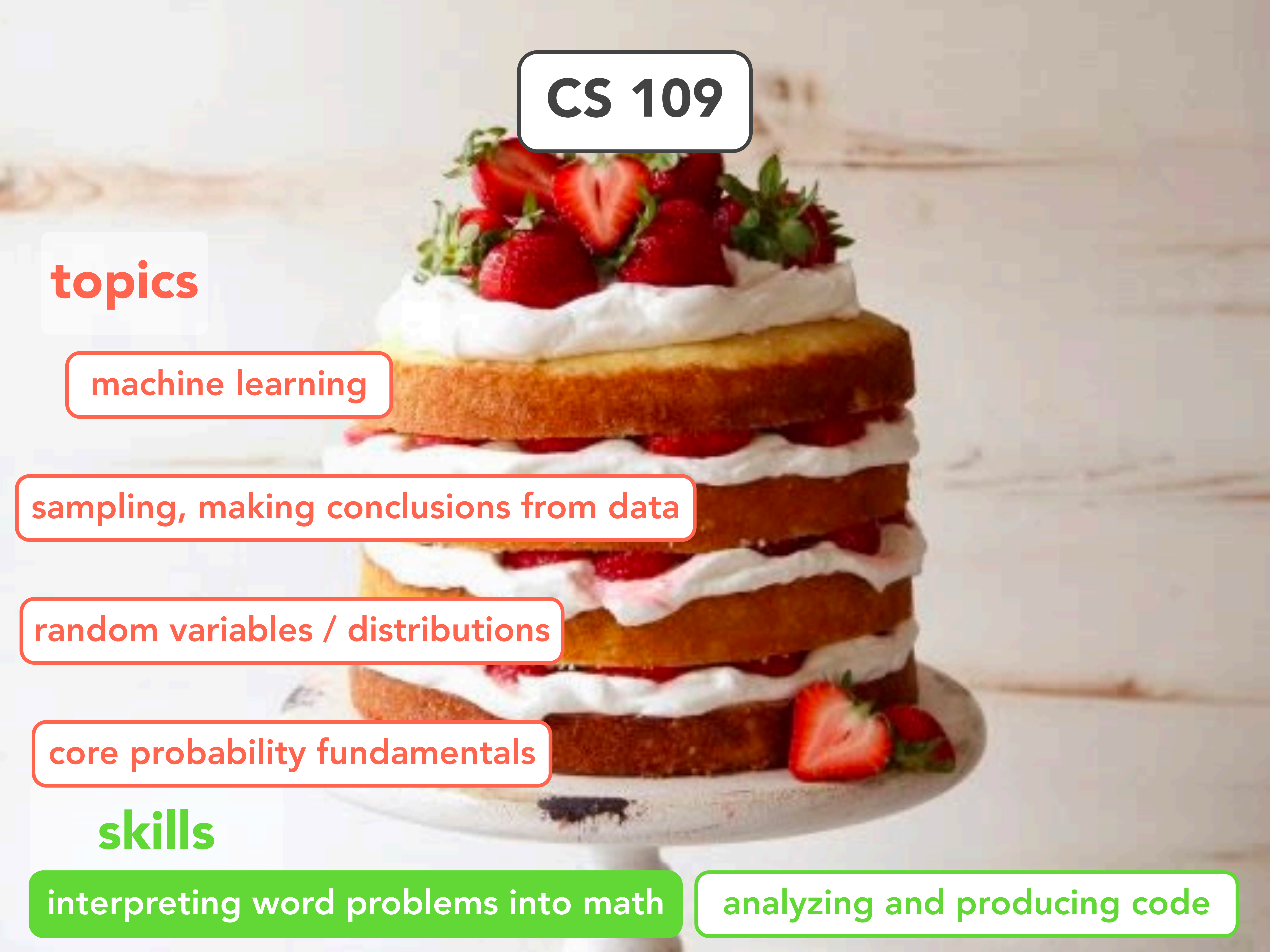
CS 109 Review

Cooper Raterink

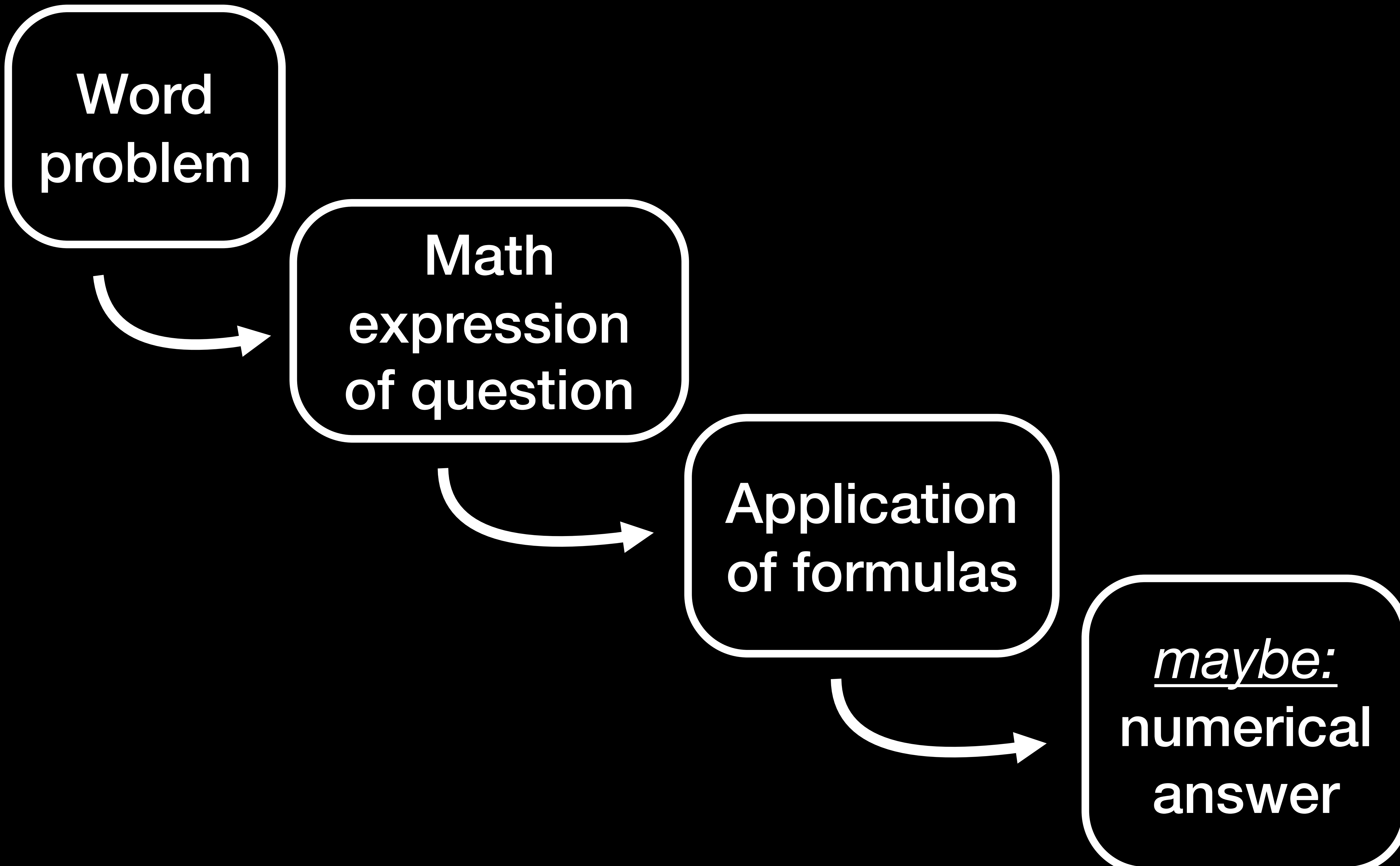
Dec. 2, 2019

Adapted from slides by Julia Daniel

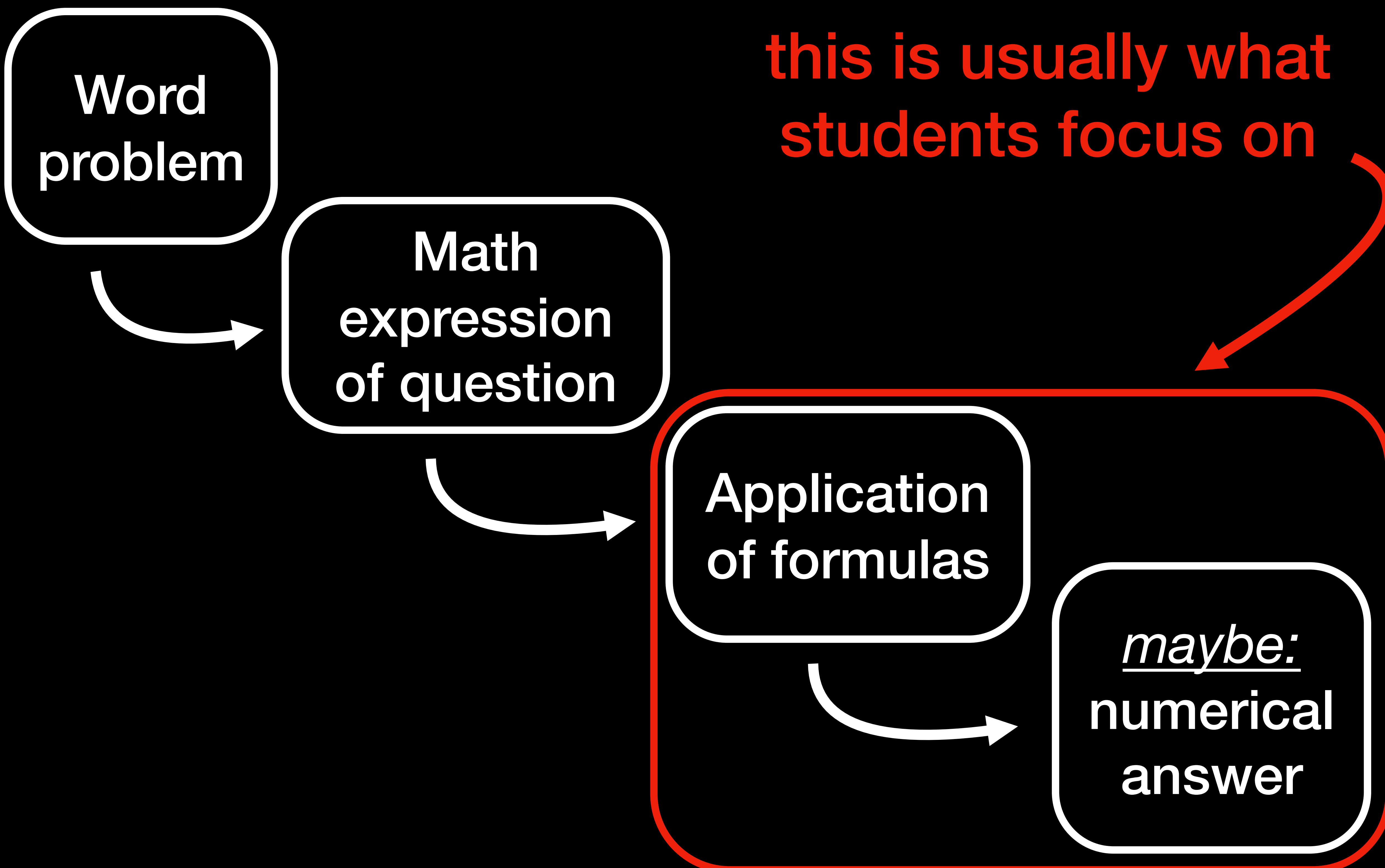

CS 109
topics
machine learning
sampling, making conclusions from data
random variables / distributions
core probability fundamentals
methods
examples
demos
problem-solving
stories and memes!
skills
interpreting word problems into math
analyzing and producing code


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

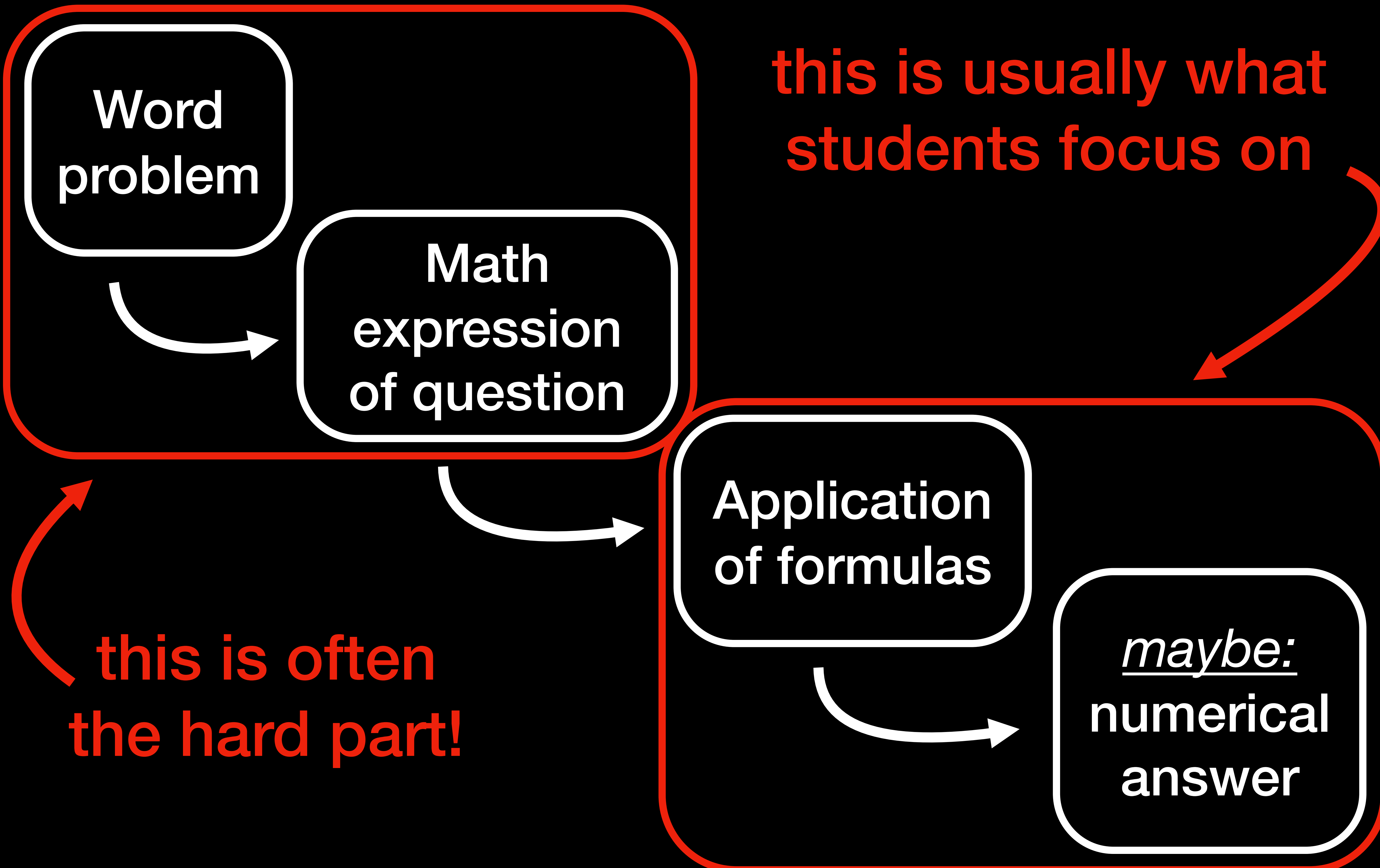
Solving a CS109 problem



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

What the problem asks:	What you should immediately think:
"What's the probability of _____ "	**P( )**
"___ given ___", "___ if ___"	**___ | ___**
"at least ___"	***Flip it: 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!

2 Medical Testing [24 points]

In medicine, there are many circumstances where we would like to detect the presence of a disease in a large population. Suppose that we would like to identify the number of individuals who have measles in a population of 1000 people using a blood test. The test is completely accurate: that is, if there are traces of measles in the blood sample, the test will return true 100% of the time and will otherwise return false. The probability that an individual has measles is 1% for everyone, independently of others.

a. (8 points) Suppose that we use a blood test on each person, in order, for a total of 1000 blood tests. What is the probability that the tenth test is the first positive test (i.e., the first person we identify with measles)?

2 Medical Testing [24 points]

Example

In medicine, there are many circumstances where we would like to detect the presence of a disease in a large population. Suppose that we would like to identify the number of individuals who have measles in a population of 1000 people using a blood test. The test is completely accurate: that is, if there are traces of measles in the blood sample, the test will return true 100% of the time and will otherwise return false. The probability that an individual has measles is 1% for everyone, independently of others.

a. (8 points) Suppose that we use a blood test on each person, in order, for a total of 1000 blood tests. What is the probability that the tenth test is the first positive test (i.e., the first person we identify with measles)?

(preamble)

1. 1000 people
2. P(test_positive | person_has_measles) = 1
3. P(person_has_measles) = 0.01

(part a)

1. 1000 independent tests
2. P(tests 1-9 negative and test 10 is positive)

2 Medical Testing [24 points]

Example

In medicine, there are many circumstances where we would like to detect the presence of a disease in a large population. Suppose that we would like to identify the number of individuals who have measles in a population of 1000 people using a blood test. The test is completely accurate: that is, if there are traces of measles in the blood sample, the test will return true 100% of the time and will otherwise return false. The probability that an individual has measles is 1% for everyone, independently of others.

a. (8 points) Suppose that we use a blood test on each person, in order, for a total of 1000 blood tests. What is the probability that the tenth test is the first positive test (i.e., the first person we identify with measles)?

(preamble)
1.1000 people
2.P(test_positive | person_has_measles) = 1
3.P(person_has_measles) = 0.01

(part a)
1.1000 independent tests
2.P(tests 1-9 negative and test 10 is positive)

What is actually important?

2 Medical Testing [24 points]

Example

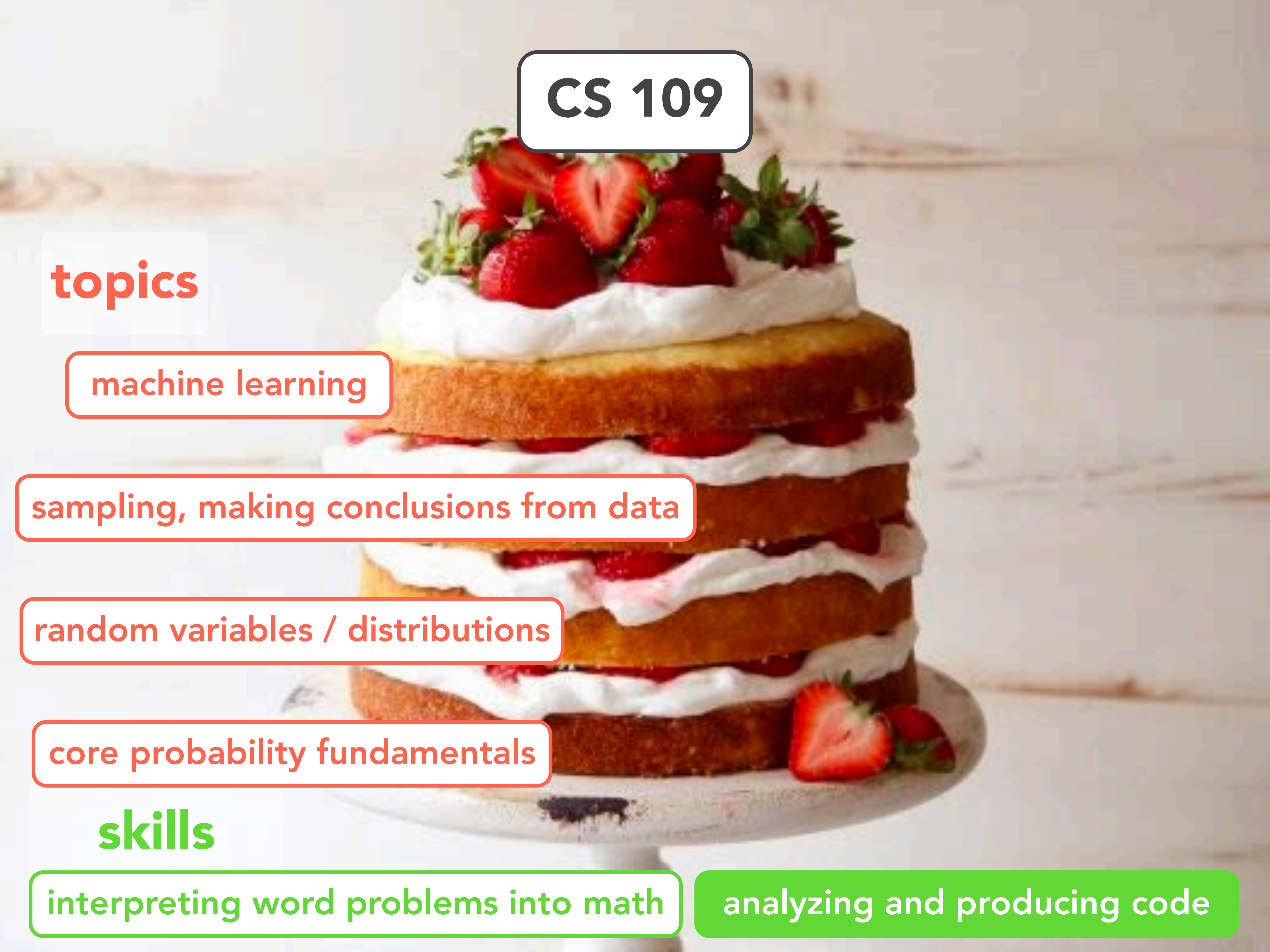
In medicine, there are many circumstances where we would like to detect the presence of a disease in a large population. Suppose that we would like to identify the number of individuals who have measles in a population of 1000 people using a blood test. The test is completely accurate: that is, if there are traces of measles in the blood sample, the test will return true 100% of the time and will otherwise return false. The probability that an individual has measles is 1% for everyone, independently of others.

a. (8 points) Suppose that we use a blood test on each person, in order, for a total of 1000 blood tests. What is the probability that the tenth test is the first positive test (i.e., the first person we identify with measles)?

(part a solution)

1. Independent tests/trials
2. P(test positive) = 0.01
3. P(10 trials until "success")
4. Implies use Geometric

5. Answer is: (0.99^9)*(0.01^1)


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

Code Analysis

Expectation of
binary tree depth
("recursive" expectation)

Bloom Filter Analysis

Expectation of
recursive die roll game

Coding Applications

Dithering

CO_2 Levels

Biometric Keystrokes

Titanic

Peer Grading

Thompson Sampling

Code in CS 109

Example

```
int fairRandom() {
    int r1, r2;
    while (true) {
        r1 = unknownRandom();
        r2 = unknownRandom();
        if (r1 != r2) break;
    }
    return r2;
}
```

a. Show mathematically that `fairRandom` does indeed return a 0 or a 1 with equal probability.

Code in CS 109

Example

```
int fairRandom() {
    int r1, r2;
    while (true) {
        r1 = unknownRandom();
        r2 = unknownRandom();
        if (r1 != r2) break;
    }
    return r2;
}
```

a. Show mathematically that `fairRandom` does indeed return a 0 or a 1 with equal probability.

1. **Need to prove P(fairR() = 1) = 0.5**
2. **<u>We know the function returns, so we break into cases</u>:**

3. **Case 1: r1 = 1 and r2 = 0 => likelihood p*(1-p)**
4. **Case 2: r1 = 0 and r2 = 1 => likelihood p*(1-p)**
5. **These are equal => equally likely for r2 to be 0 or 1**


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**	

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	$\frac{n!}{k_1!k_2!\ldots k_n!}$	the divider method!

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

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

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

Probability basics

Example

7. If we assume that all possible poker hands (comprised of 5 cards from a standard 52 card deck) are equally likely, what is the probability of being dealt:

   a. a flush? (A hand is said to be a flush if all 5 cards are of the same suit. Note that this definition means that *straight flushes* (five cards of the same suit in numeric sequence) are also considered flushes.)
   b. two pairs? (This occurs when the cards have numeric values a, a, b, b, c, where a, b and c are all distinct.)
   c. three of a kind? (This occurs when the cards have numeric values a, a, a, b, c, where a, b and c are all distinct.)

Part a:

Probability basics

Example

7. If we assume that all possible poker hands (comprised of 5 cards from a standard 52 card deck) are equally likely, what is the probability of being dealt:
   a. a flush? (A hand is said to be a flush if all 5 cards are of the same suit. Note that this definition means that *straight flushes* (five cards of the same suit in numeric sequence) are also considered flushes.)
   b. two pairs? (This occurs when the cards have numeric values a, a, b, b, c, where a, b and c are all distinct.)
   c. three of a kind? (This occurs when the cards have numeric values a, a, a, b, c, where a, b and c are all distinct.)

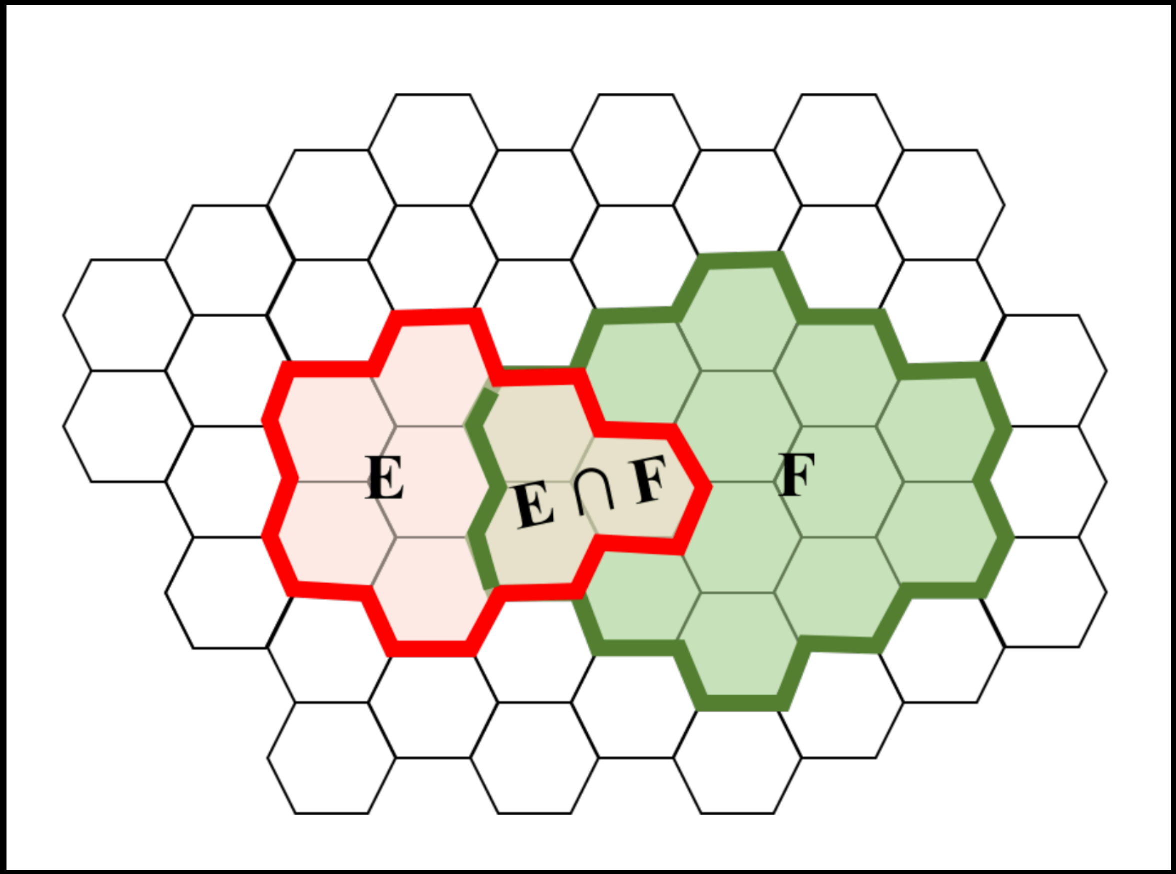
Part a:

1. **Hand rearrangement OK => use unordered sample space**
2. **Sample space => 52C5**
3. **For event space: choose suit, choose cards => 4C1 * 13C5**
4. **Put it together: P(a flush) = 4C1 * 13C5 / 52C5**

Conditional Probability

definition:

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



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

Chain Rule:

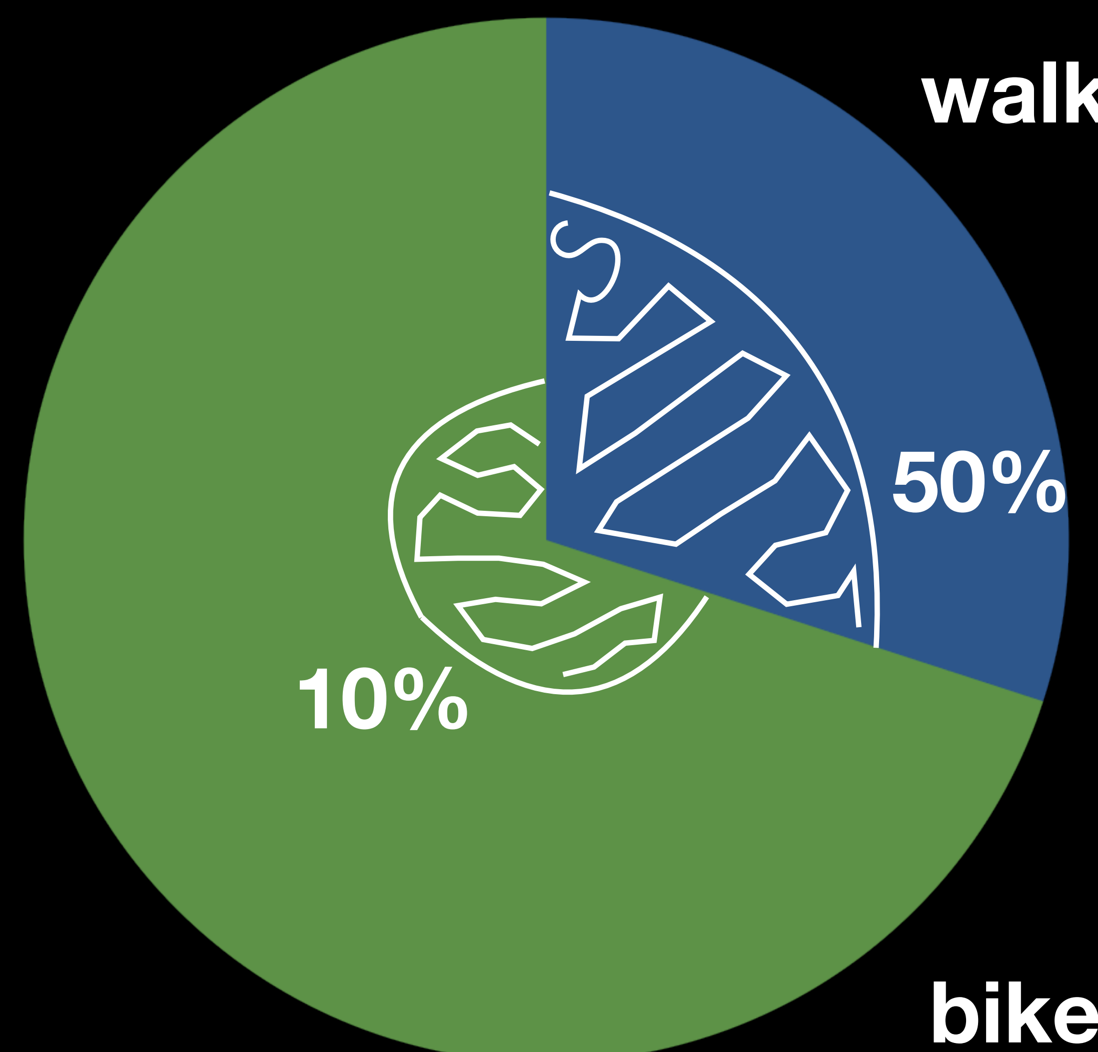
$$P(EF) = P(E \mid F)P(F)$$

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

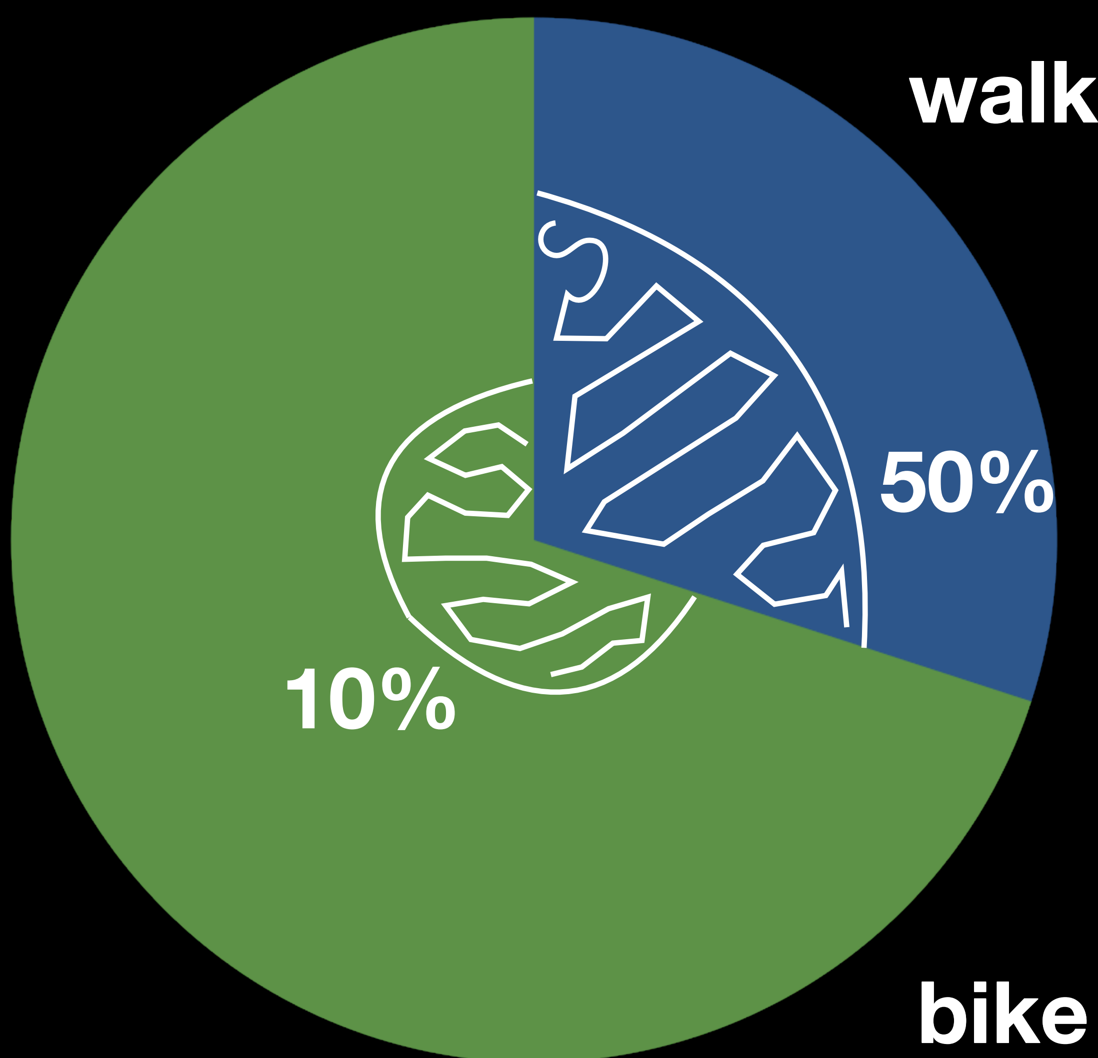


Law of Total Probability

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

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



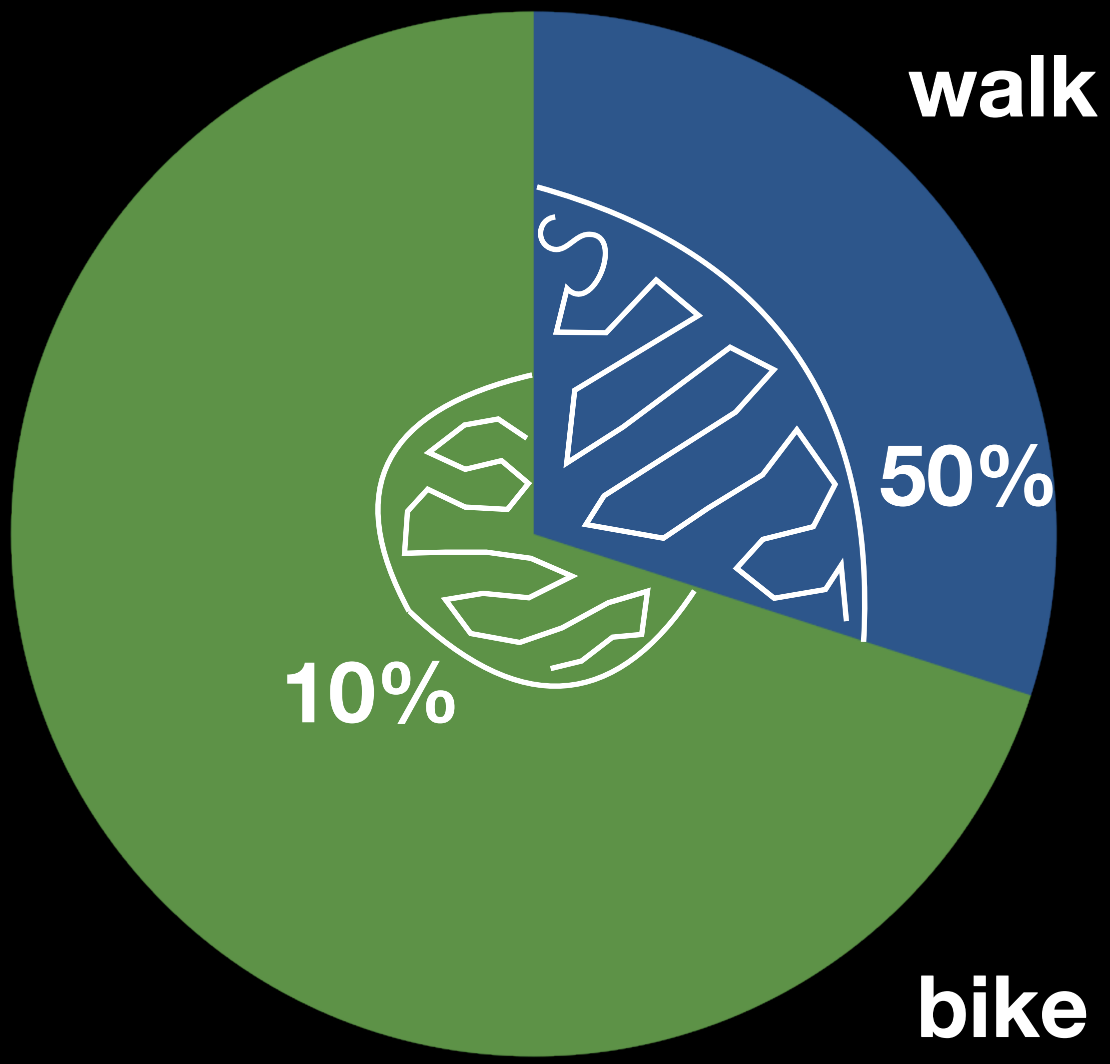
$$\begin{aligned} P(L) &= P(L \mid W)P(W) + P(L \mid W^C)P(W^C) \\ &= (0.5)(0.3) + (0.1)(0.7) \\ &= 0.22 \end{aligned}$$

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) = ?

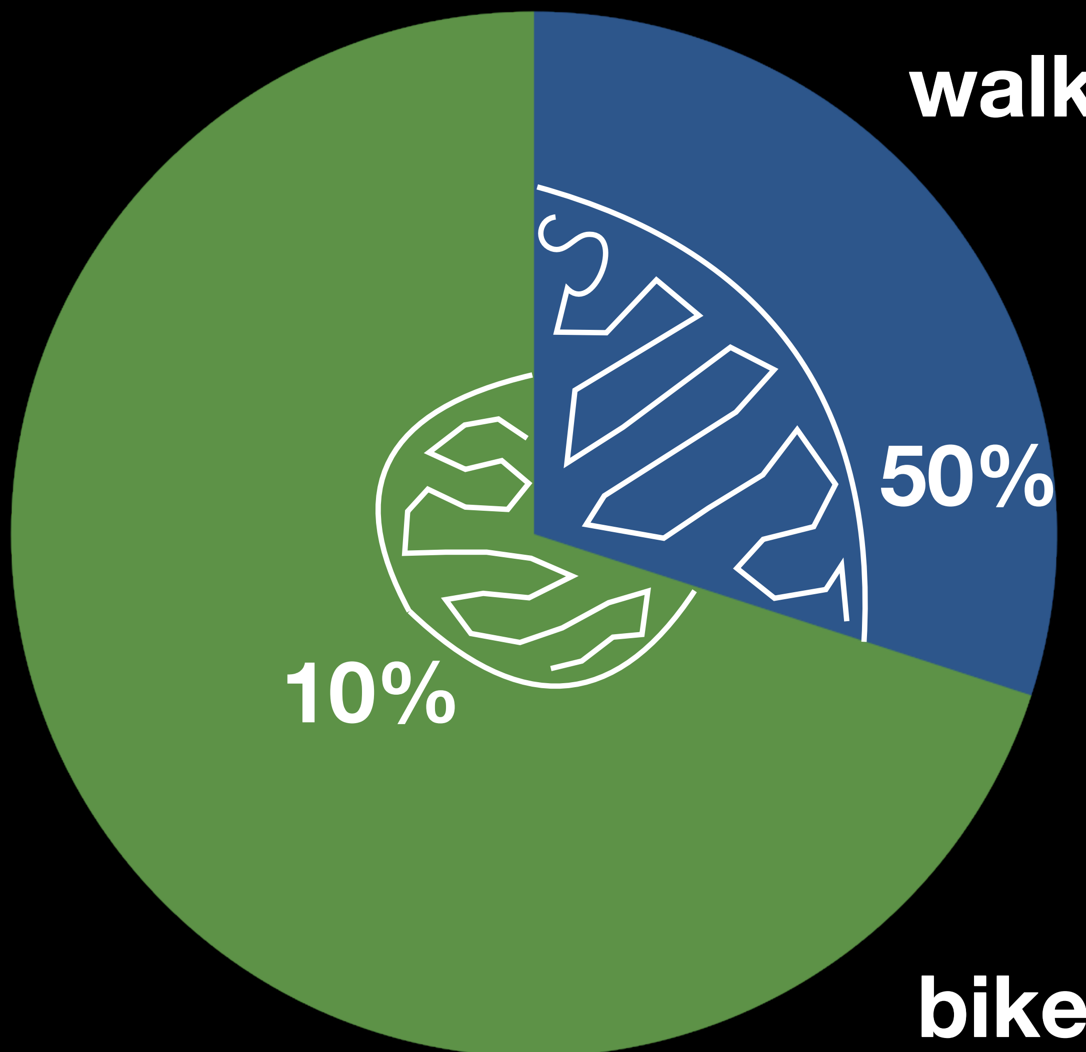
what if we can bike, walk, or
take the Marguerite (> 2 options)?



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 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

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

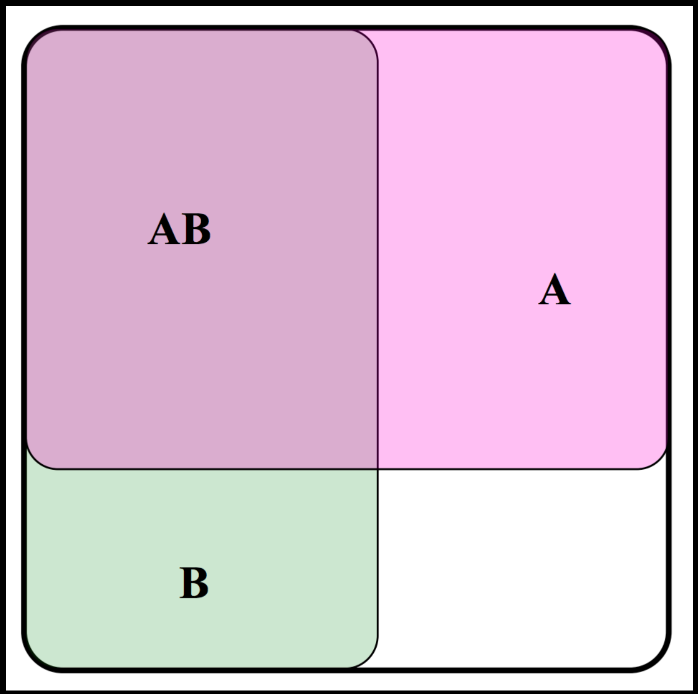
$$P(F \mid E)P(E) + P(F \mid E^C)P(E^C)$$

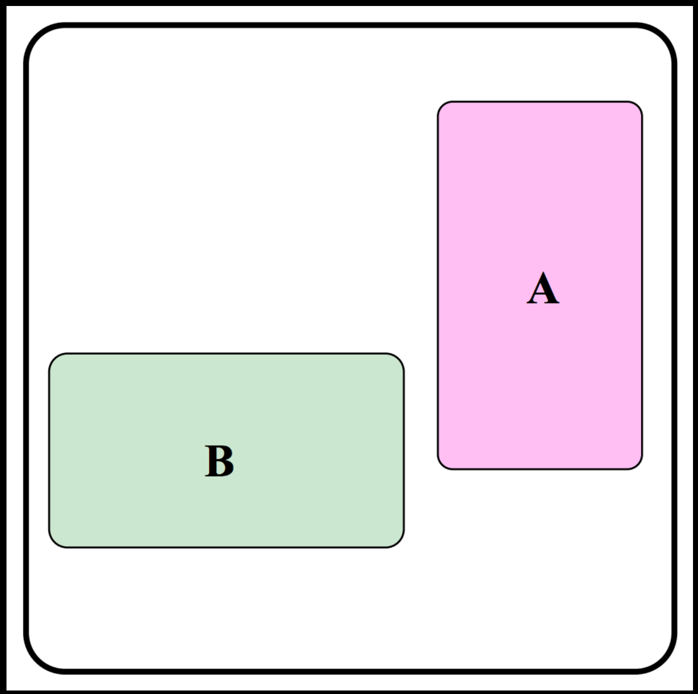
divide the event F into all the possible ways it can happen; use LoTP

Old Principles, New Tricks

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

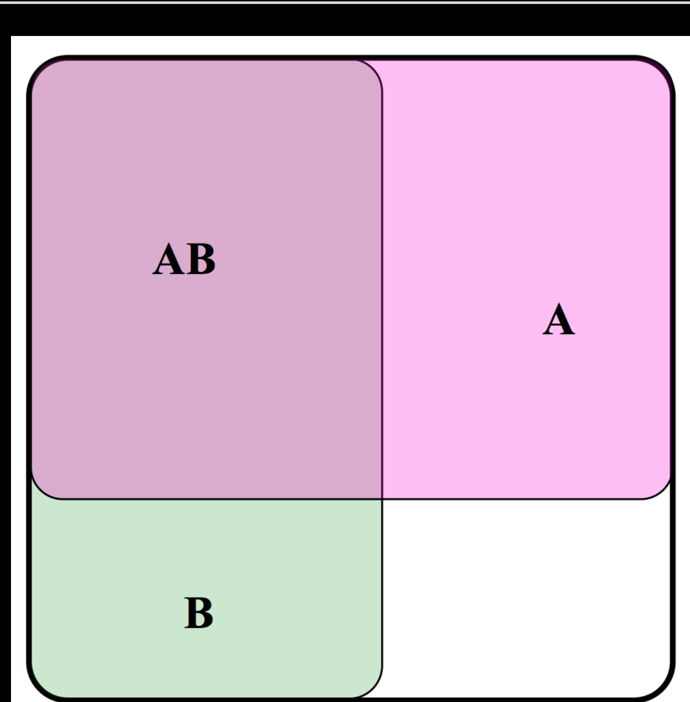
Independence

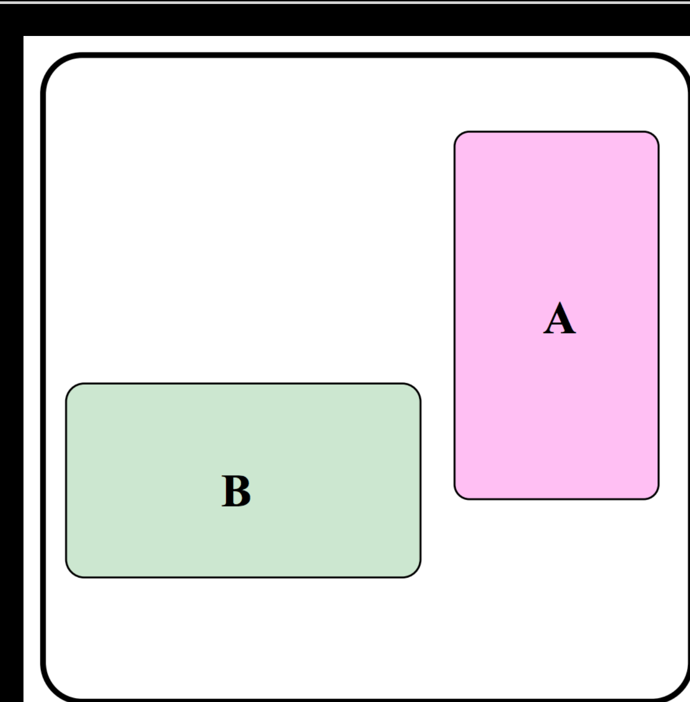




Independence

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





Independence

Independence	Conditional Independence
$P(EF) = P(E)P(F)$ $P(E\|F) = P(E)$	$P(EF\|G) = P(E\|G)P(F\|G)$ $P(E\|FG) = P(E\|G)$
"AND"	"AND [if]"

If E and F are independent…..

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

& vice versa

Beyond the basics

Example

7. Consider a hash table with 15 buckets, of which 9 are empty (have no strings hashed to them) and the other 6 buckets are non-empty (have at least one string hashed to each of them already). Now, 2 new strings are independently hashed into the table, where each string is equally likely to be hashed into any bucket. Later, another 2 strings are hashed into the table (again, independently and equally likely to get hashed to any bucket). What is the probability that both of the final 2 strings are each hashed to empty buckets in the table?

Beyond the basics

Example

7. Consider a hash table with 15 buckets, of which 9 are empty (have no strings hashed to them) and the other 6 buckets are non-empty (have at least one string hashed to each of them already). Now, 2 new strings are independently hashed into the table, where each string is equally likely to be hashed into any bucket. Later, another 2 strings are hashed into the table (again, independently and equally likely to get hashed to any bucket). What is the probability that both of the final 2 strings are each hashed to empty buckets in the table?

How do you begin to break down this problem?

Beyond the basics

Example

7. Consider a hash table with 15 buckets, of which 9 are empty (have no strings hashed to them) and the other 6 buckets are non-empty (have at least one string hashed to each of them already). Now, 2 new strings are independently hashed into the table, where each string is equally likely to be hashed into any bucket. Later, another 2 strings are hashed into the table (again, independently and equally likely to get hashed to any bucket). What is the probability that both of the final 2 strings are each hashed to empty buckets in the table?

How do you begin to break down this problem?

Let event A = first of initial two strings hashed to empty bucket.
Let event B = second of initial two strings hashed to empty bucket.
Let event C = first of final two strings hashed to empty bucket.
Let event D = second of final two strings hashed to empty bucket.

Define events

Example

Beyond the basics

7. Consider a hash table with 15 buckets, of which 9 are empty (have no strings hashed to them) and the other 6 buckets are non-empty (have at least one string hashed to each of them already). Now, 2 new strings are independently hashed into the table, where each string is equally likely to be hashed into any bucket. Later, another 2 strings are hashed into the table (again, independently and equally likely to get hashed to any bucket). What is the probability that both of the final 2 strings are each hashed to empty buckets in the table?

How do you begin to break down this problem?

Let event A = first of initial two strings hashed to empty bucket.
Let event B = second of initial two strings hashed to empty bucket.
Let event C = first of final two strings hashed to empty bucket.
Let event D = second of final two strings hashed to empty bucket.

Define events

We compute $P(CD)$ as follows:

$$\begin{aligned} P(CD) &= P(CD \mid AB)P(AB) + P(CD \mid A^C B)P(A^C B) + \\ &\quad P(CD \mid AB^C)P(AB^C) + P(CD \mid A^C B^C)P(A^C B^C) \\ &= (7/15)(6/15)(9/15)(8/15) + (8/15)(7/15)(6/15)(9/15) + \\ &\quad (8/15)(7/15)(9/15)(7/15) + (9/15)(8/15)(6/15)(6/15) \\ &= 12168/154 \approx 0.2404 \end{aligned}$$

What is the question asking?

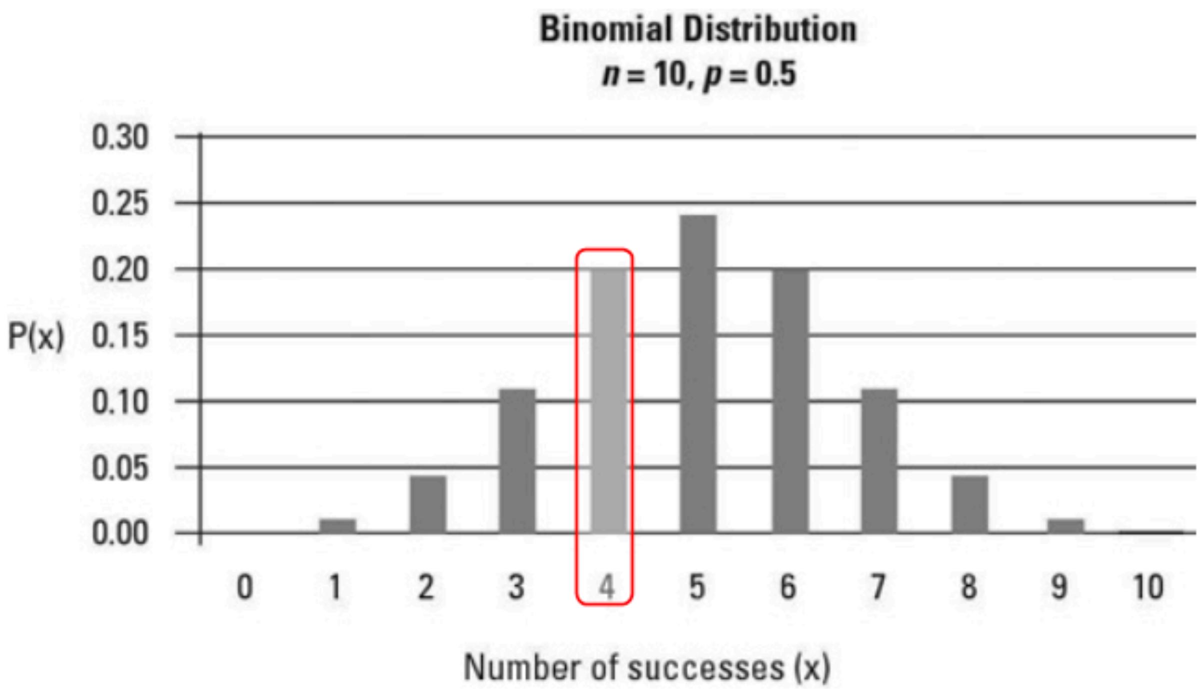
Use LOTP


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

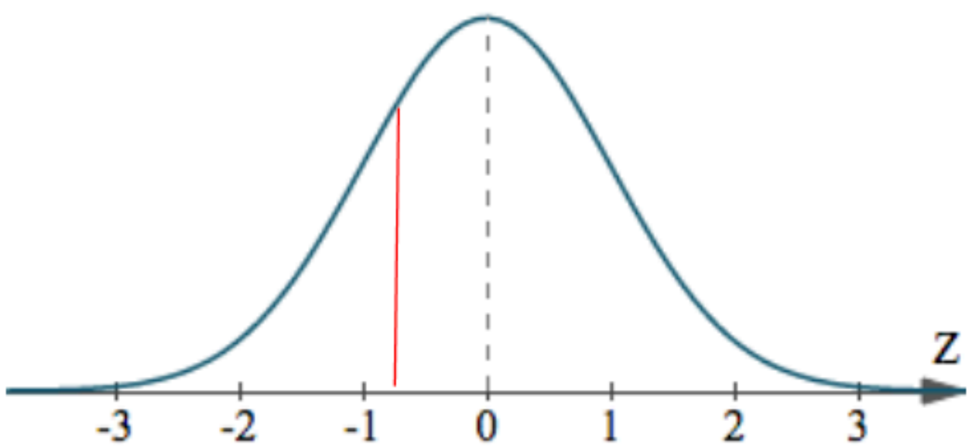
Discrete

PMF:



Continuous

PDF:

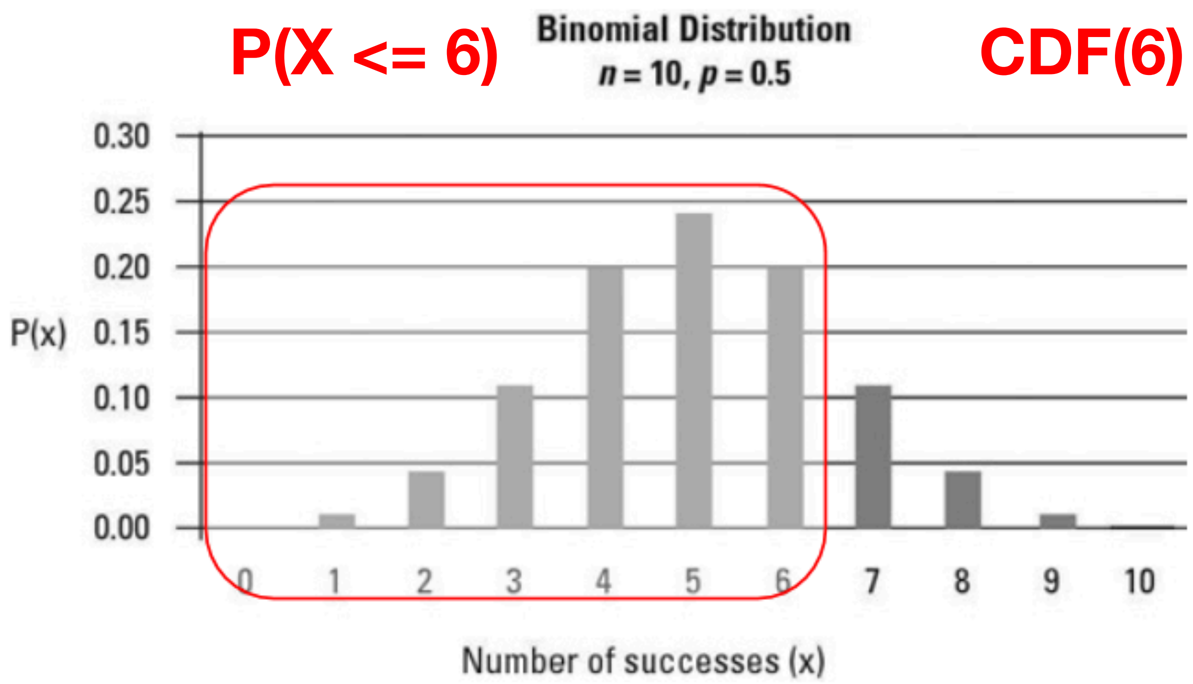


Probability Distributions

Discrete

CDF: **P(X <= x)**

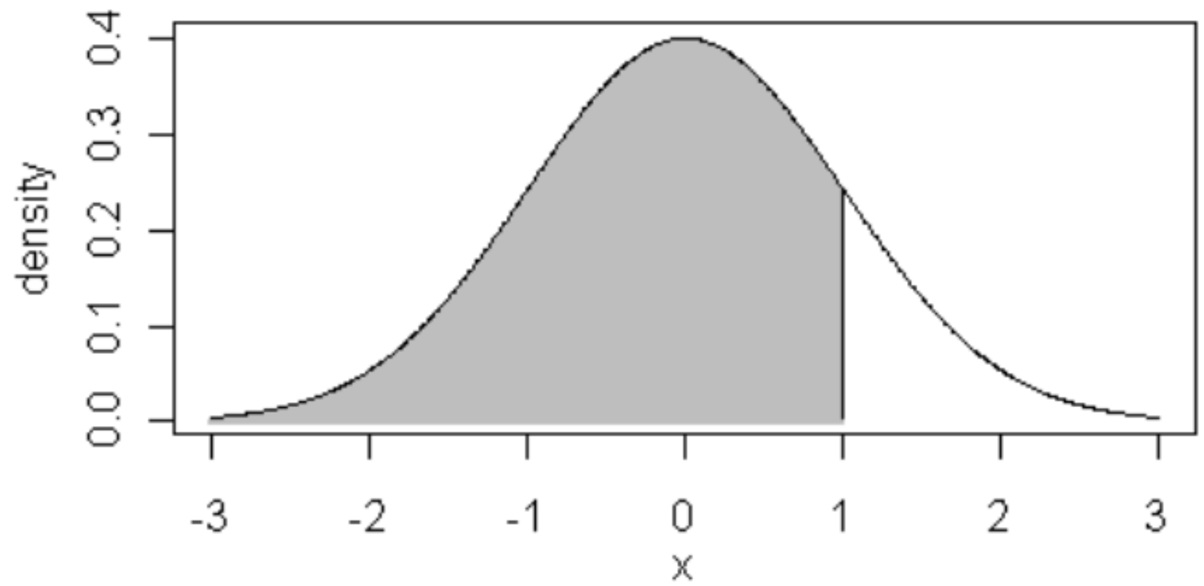
P(X <= 6) **CDF(6)**



Continuous

CDF: **P(X <= x)**

P(X <= 1) **CDF(1)**



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

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

(mu)

Continuous definition

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

(mu)

Properties of Expectation

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

Expectation & Variance

Discrete definition

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

(mu)

Continuous definition

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

(mu)

Properties of Expectation

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

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

$$E[g(X)] = \sum_x g(x) * p_X(x)$$

Expectation & Variance

Discrete definition

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

(mu)

Continuous definition

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

(mu)

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

(mu)

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

Expectation & Variance

Discrete definition

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

(mu)

Continuous definition

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

(mu)

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

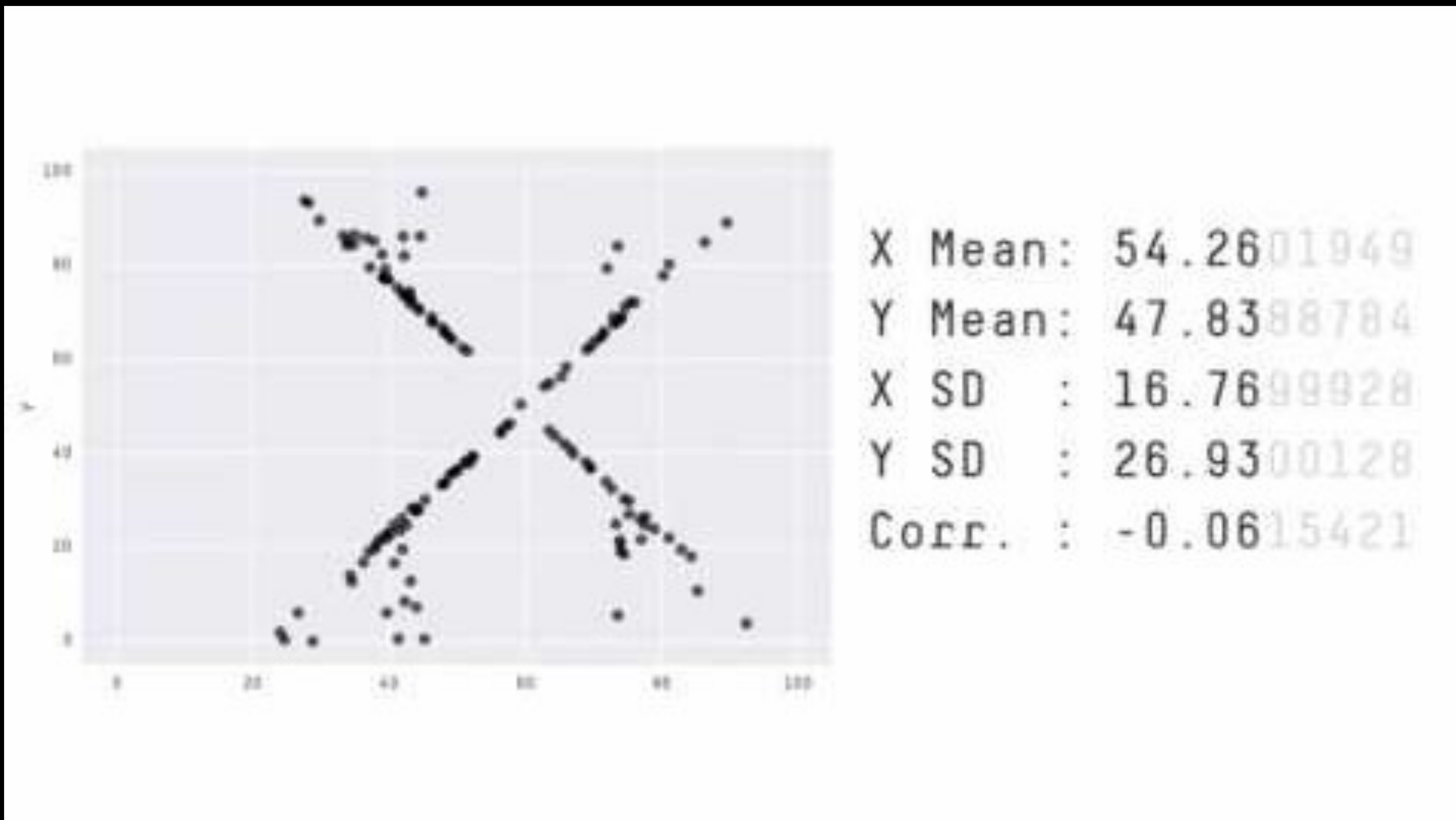
(mu)

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

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

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

Extras



All our (discrete) friends

Ber(p)	Bin(n, p)	Poi(λ)	Geo(p)	NegBin(r, p)
P(X) = p	$\binom{n}{k}p^k(1-p)^{n-k}$	$\frac{\lambda^k e^{-\lambda}}{k!}$	$(1-p)^{k-1}p$	$\binom{k-1}{r-1}p^r(1-p)^{k-r}$
Getting candy or not at a random house				

All our (discrete) friends

Ber(p)	Bin(n, p)	Poi(λ)	Geo(p)	NegBin(r, p)
P(X) = p	$\binom{n}{k} p^k (1-p)^{n-k}$	$\frac{\lambda^k e^{-\lambda}}{k!}$	$(1-p)^{k-1} p$	$\binom{k-1}{r-1} p^r (1-p)^{k-r}$
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 (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)
$f(x) = \frac{1}{\beta - \alpha}$	$f(x) = \lambda e^{-\lambda x}$	$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{\frac{-(x-\mu)^2}{2\sigma^2}}$
$P(a \leq X \leq b) = \frac{b - a}{\beta - \alpha}$	$F(x) = 1 - e^{-\lambda x}$	$F(x) = \Phi(\frac{x - \mu}{\sigma})$
thickness of sidewalk pavement between houses		

All our (continuous) friends

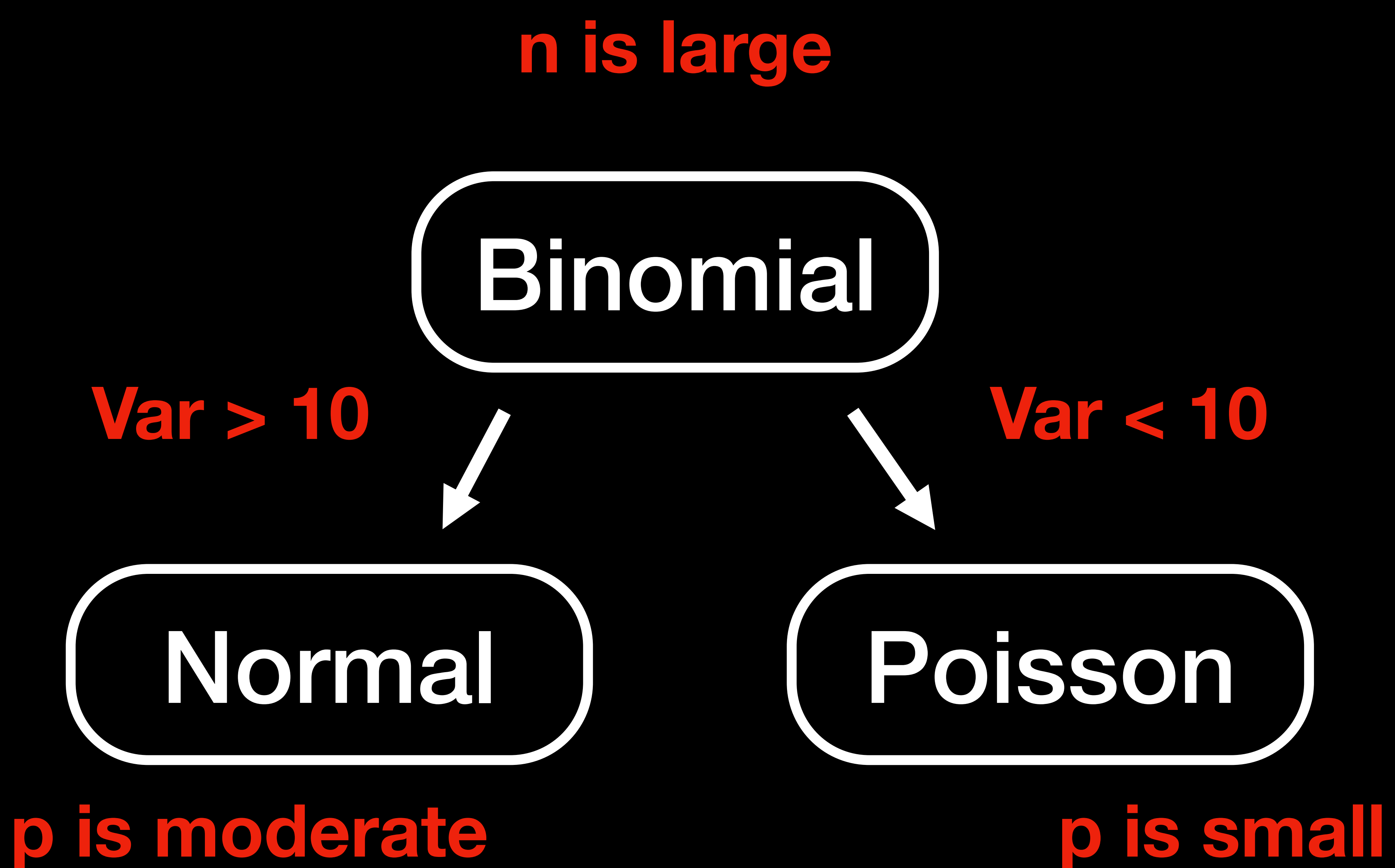
Uni(α, β)	Exp(λ)	N(μ, σ)
$f(x) = \frac{1}{\beta - \alpha}$	$f(x) = \lambda e^{-\lambda x}$	$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{\frac{-(x-\mu)^2}{2\sigma^2}}$
$P(a \leq X \leq b) = \frac{b-a}{\beta - \alpha}$	$F(x) = 1 - e^{-\lambda x}$	$F(x) = \Phi(\frac{x-\mu}{\sigma})$
thickness of sidewalk pavement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

All our (continuous) friends

Uni(α, β)	Exp(λ)	N(μ, σ)
$f(x) = \frac{1}{\beta - \alpha}$	$f(x) = \lambda e^{-\lambda x}$	$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{\frac{-(x-\mu)^2}{2\sigma^2}}$
$P(a \le X \le b) = \frac{b-a}{\beta - \alpha}$	$F(x) = 1 - e^{-\lambda x}$	$F(x) = \Phi(\frac{x-\mu}{\sigma})$
$E(x) = \frac{\alpha + \beta}{2}$	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?



Approximations

When can we approximate a binomial?

n is large

Binomial

Var > 10 → Normal (Continuity Correction)

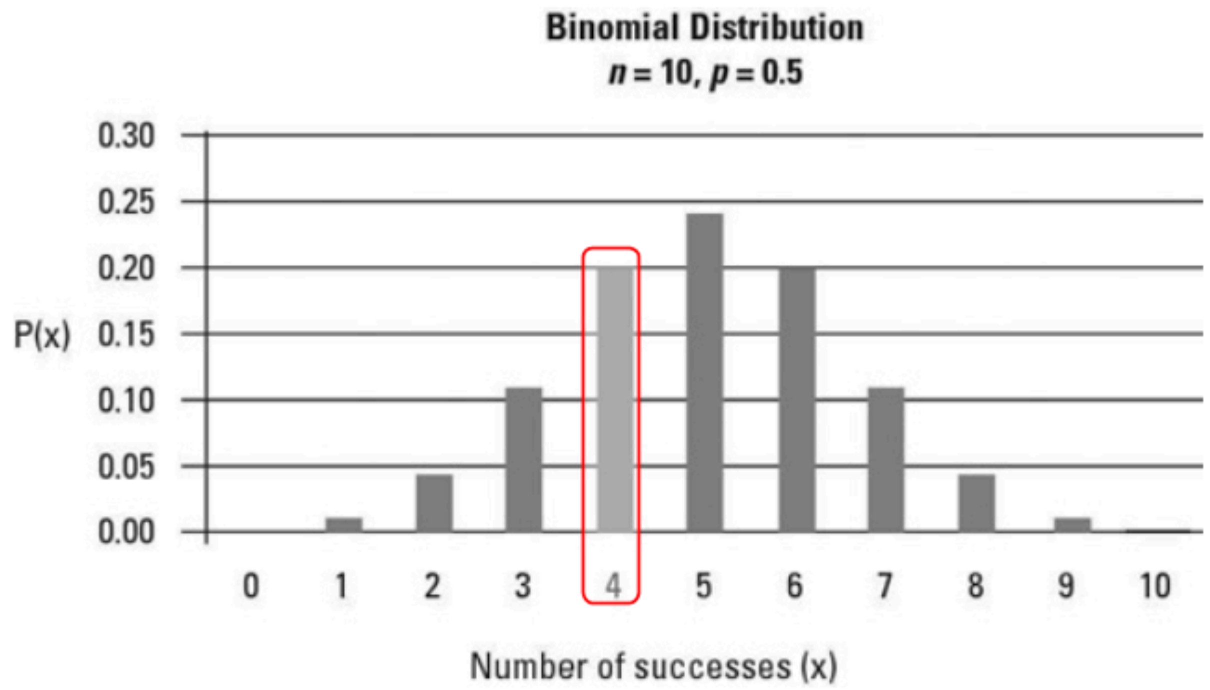
Var < 10 → Poisson

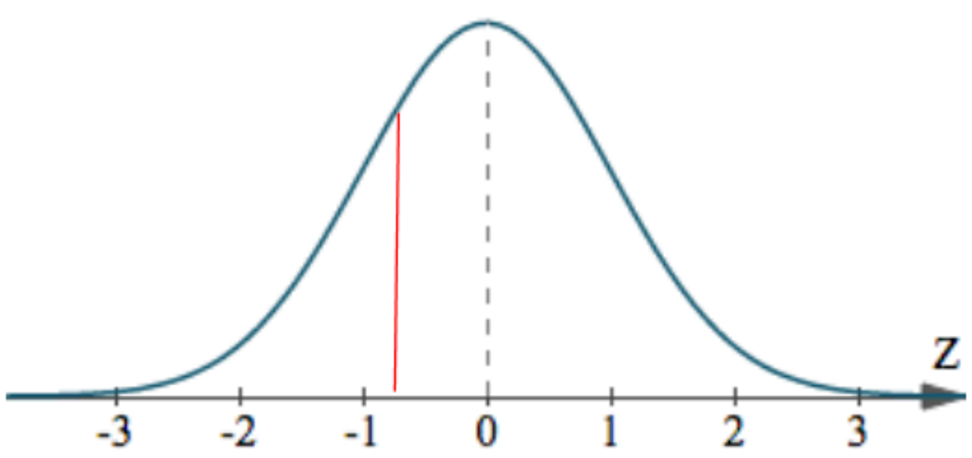
p is moderate

p is small

Continuity Correction

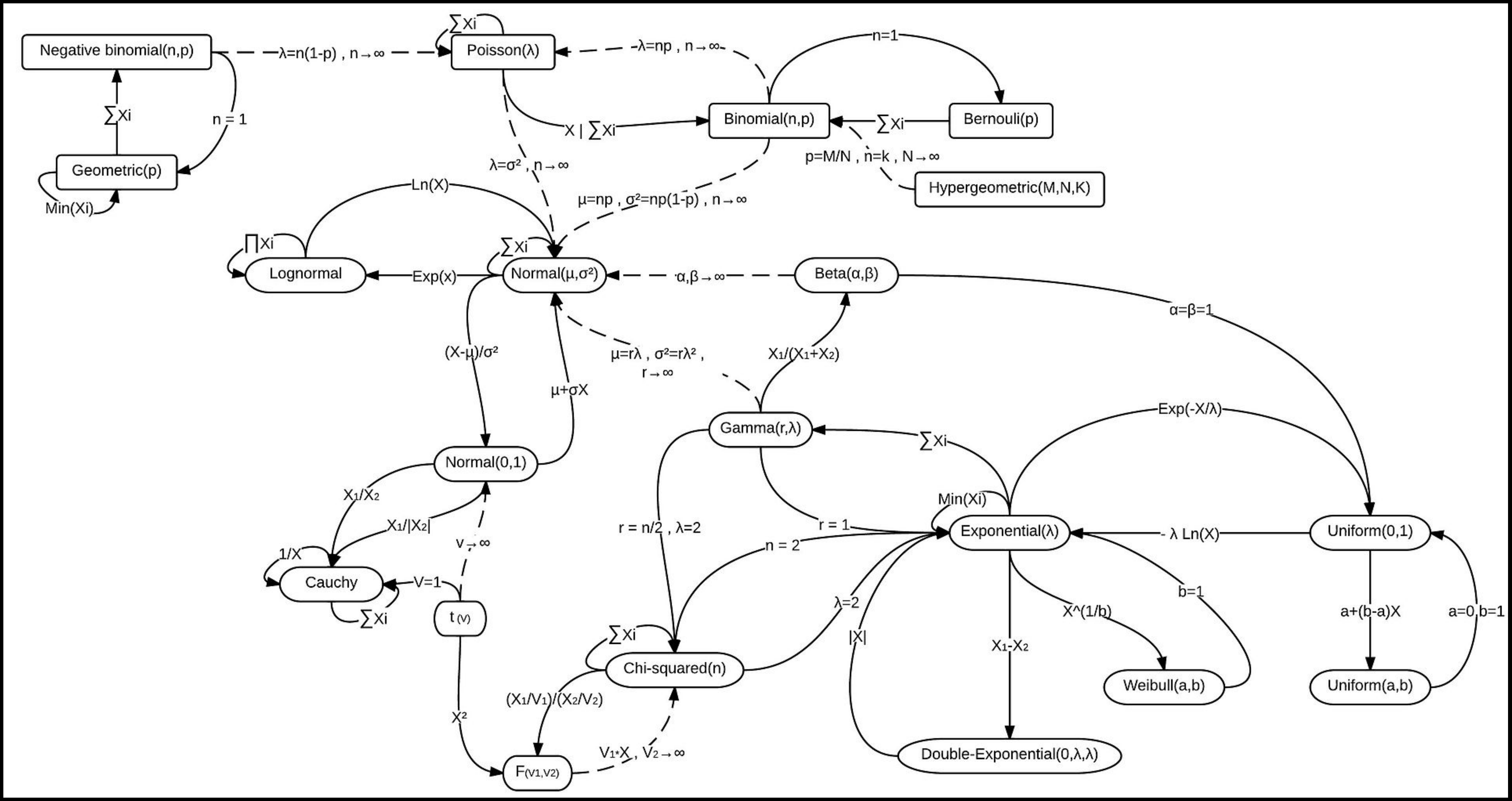
Discrete	Continuous
PMF:	PDF:





Only applies to continuous RVs approximating discrete RVs - why?

Extras



Negative binomial(n,p)
λ=n(1-p) , n→∞
Poisson(λ)
∑Xi
λ=np , n→∞
n=1
Binomial(n,p)
Bernouli(p)
∑Xi
n = 1
X | ∑Xi
p=M/N , n=k , N→∞
Hypergeometric(M,N,K)
Geometric(p)
Min(Xi)
λ=σ² , n→∞
μ=np , σ²=np(1-p) , n→∞
Ln(X)
∏Xi
Lognormal
Exp(x)
Normal(μ,σ²)
α,β→∞
Beta(α,β)
α=β=1
(X-μ)/σ²
μ+σX
μ=rλ , σ²=rλ² , r→∞
X₁/(X₁+X₂)
Gamma(r,λ)
Exp(-X/λ)
Normal(0,1)
X₁/X₂
X₁/|X₂|
v→∞
1/X
Cauchy
V=1
t (V)
r = n/2 , λ=2
r = 1
Min(Xi)
Exponential(λ)
- λ Ln(X)
Uniform(0,1)
n = 2
b=1
X^(1/b)
a+(b-a)X
a=0,b=1
λ=2
|X|
X₁-X₂
Chi-squared(n)
Weibull(a,b)
Uniform(a,b)
(X₁/V₁)/(X₂/V₂)
X²
F(V1,V2)
V₁*X , V₂→∞
Double-Exponential(0,λ,λ)

Distribution onslaught!

Example

Coin flip is heads

Distribution onslaught!

Example

Number of heads in 10 coin flips

Distribution onslaught!

Example

Coin flips until a heads

Distribution onslaught!

Example

Chance of CS109 student sleeping in class is 70%

Number of CS109 students sleeping in class right now?

Distribution onslaught!

Example

Chance of CS109 student sleeping is 70%

Number of CS109 students sleeping right now?
(approximate)

Distribution onslaught!

Example

You look around and see 120 out of 150 CS109 students asleep.

What is your belief distribution for P(CS109 student asleep)?

Distribution onslaught!

Example

CS109 students fall asleep on average once a minute.

Time until a CS109 student falls asleep?

Distribution onslaught!

Example

CS109 students fall asleep on average once a minute.

Number of CS109 students who fall asleep in the next 10 minutes?

Joint Distributions

- Discrete case: $p_{x,y}(a,b) = P(X = a, Y = b)$

  - Marginalize a variable out: $P_x(a) = \sum_y P_{x,y}(a,y)$

- Continuous case:

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

  - Marginalize a variable out:

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

Joint Distributions

- Discrete case: $p_{x,y}(a,b) = P(X = a, Y = b)$

  - Marginalize a variable out: $P_x(a) = \sum_y P_{x,y}(a,y)$

- Continuous case:

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

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

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

Sums of Indep. RVs

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

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

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

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

Sums of Indep. RVs

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

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

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

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

Caveat: These rules only work for independent X and Y!

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

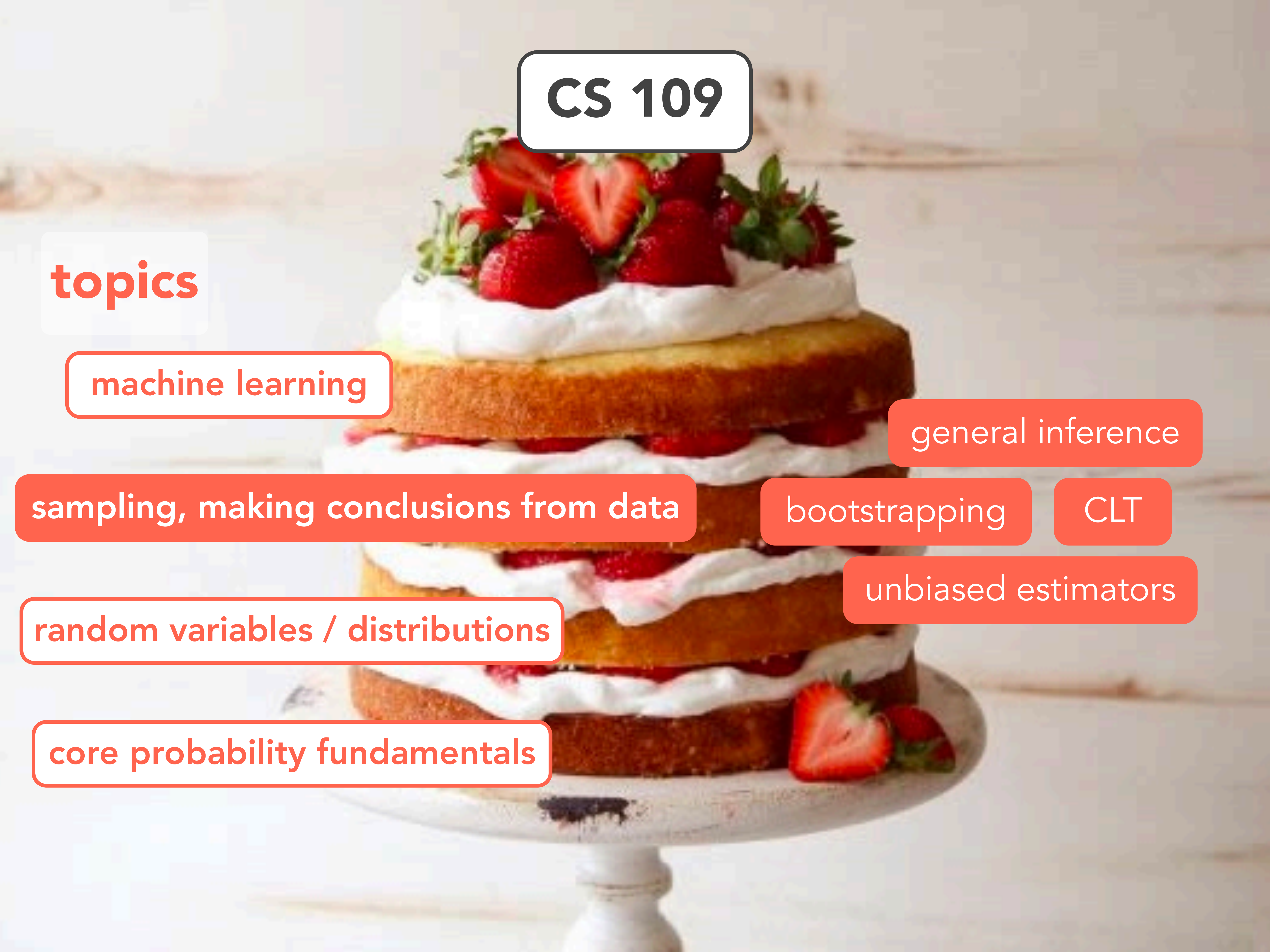
$$Cov(X, Y) = E[XY] - E[Y]E[X]$$

Correlation

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

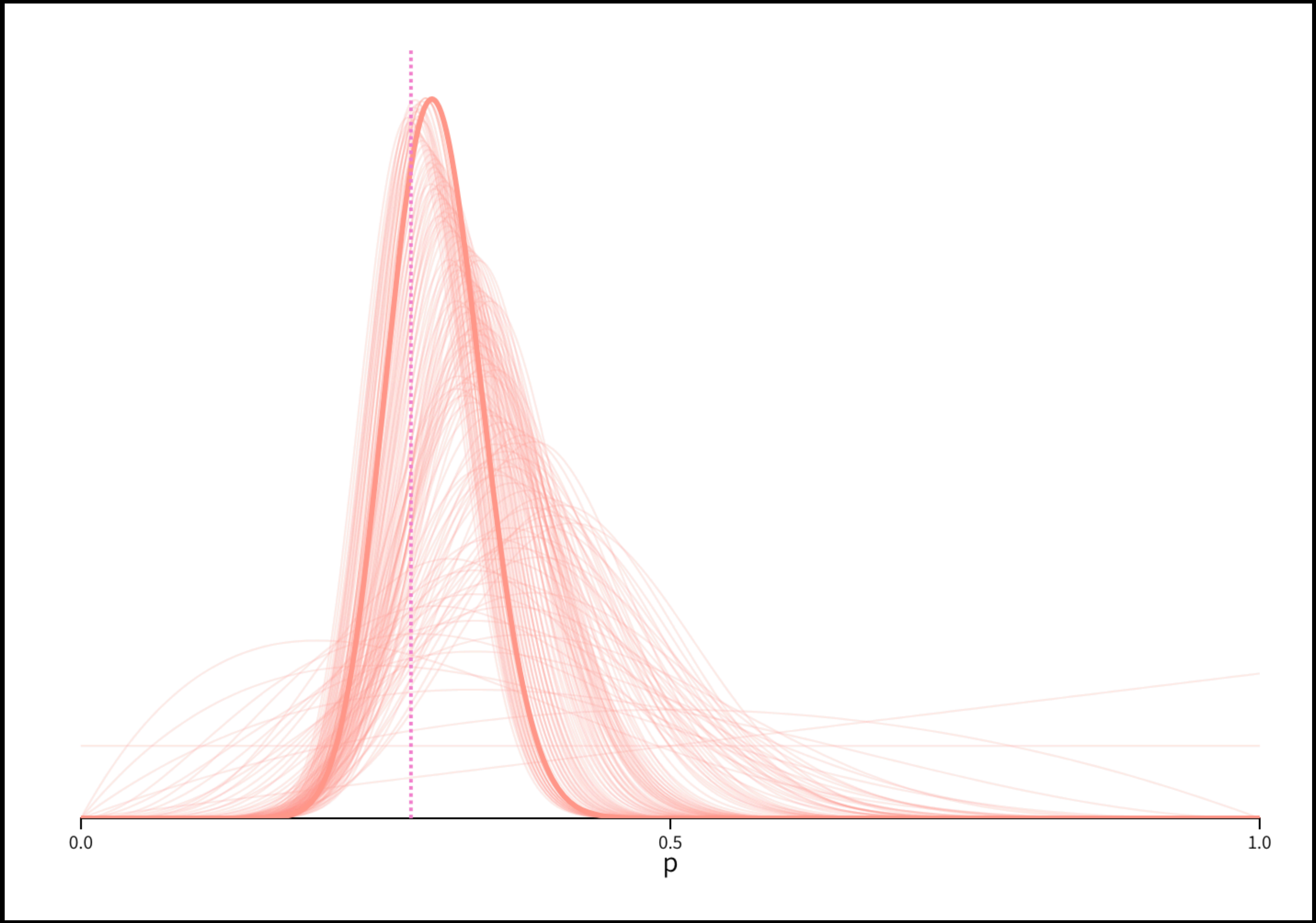
$$\rho(X, Y) = \frac{Cov(X, Y)}{\sqrt{Var(X)Var(Y)}}$$

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


CS 109
topics
machine learning
general inference
sampling, making conclusions from data
bootstrapping
CLT
unbiased estimators
random variables / distributions
core probability fundamentals

Beta

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



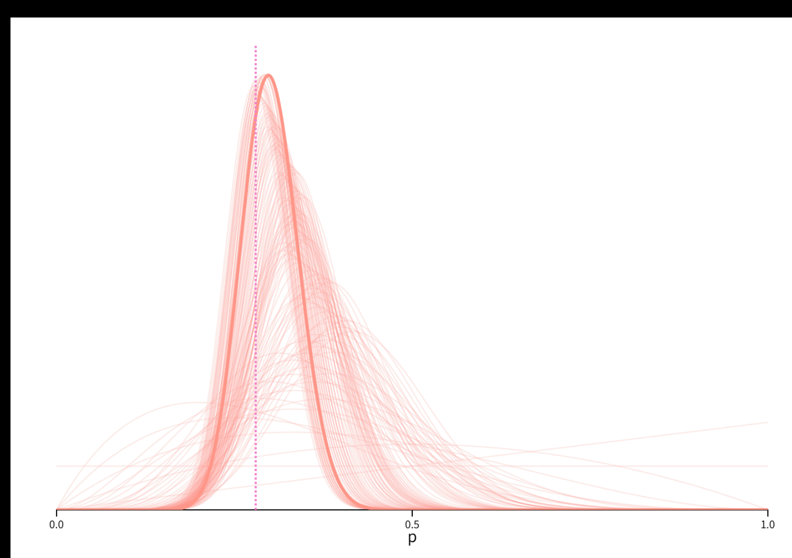
https://seeing-theory.brown.edu/bayesian-inference/index.html#section3

Beta

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

Bern(p)

1001010101000101010010101 00 ....



Updating belief about Bernoulli parameter p

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 – Conceptual principles

Take a representative sample as large as you can

Sampling – Conceptual principles

Take a representative sample as large as you can

Sample statistics can be helpful in understanding the population

Sampling – Conceptual principles

Take a representative sample as large as you can

Sample statistics can be helpful in understanding the population

Be careful in assuming things about population from sample statistics (you can bootstrap to better understand your population and statistics)

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!

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

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

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

makes the estimate unbiased

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

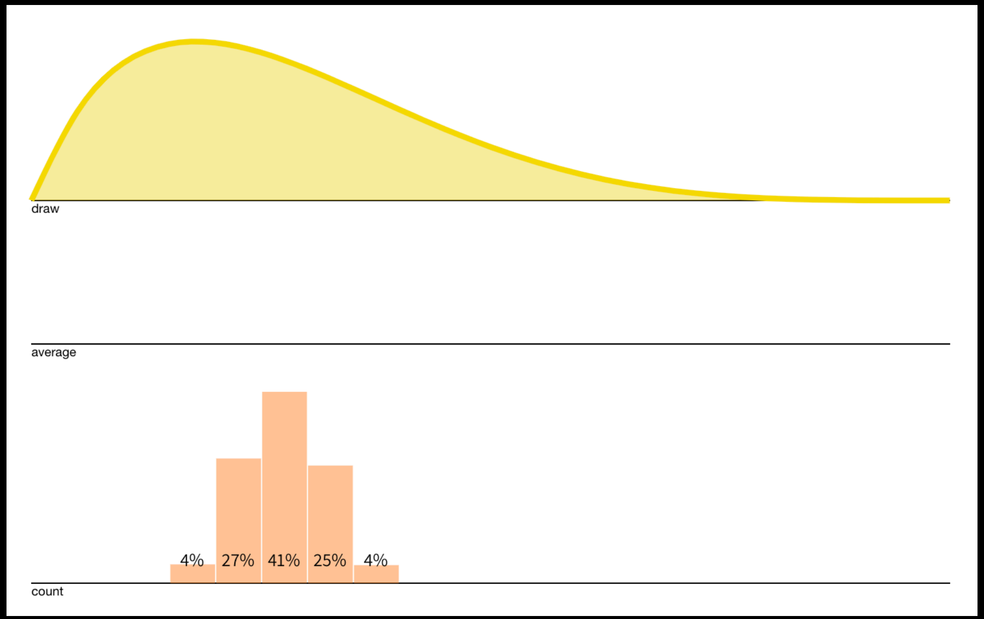
Taking Many Samples

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

Central Limit Theorem

As $n \to \infty$

$$\sum_{i=1}^{n} X_i \sim \mathcal{N}(n\mu, n\sigma^2)$$

Extras

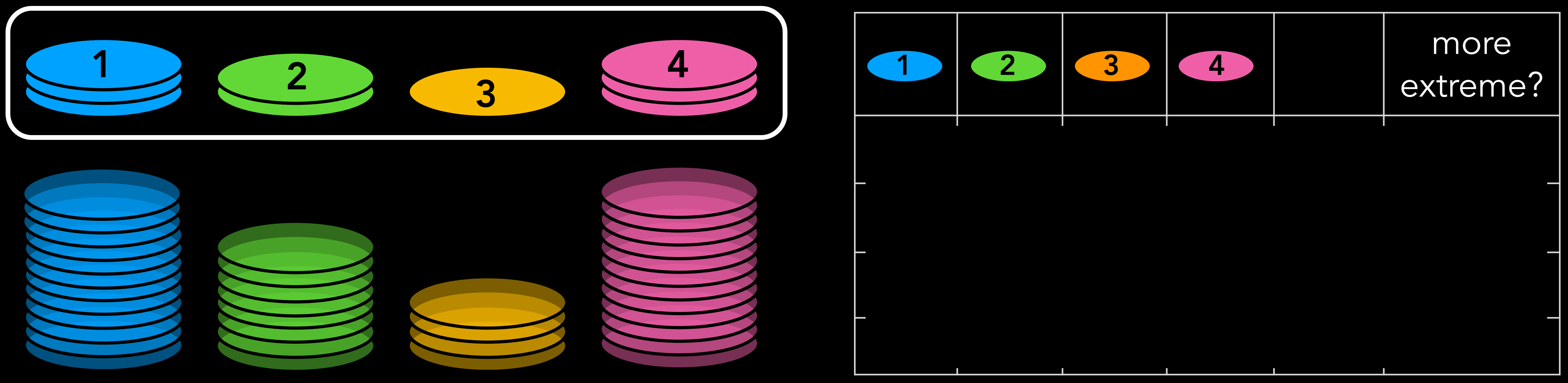
Bootstrapping: Simulating Many Samples From One

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.

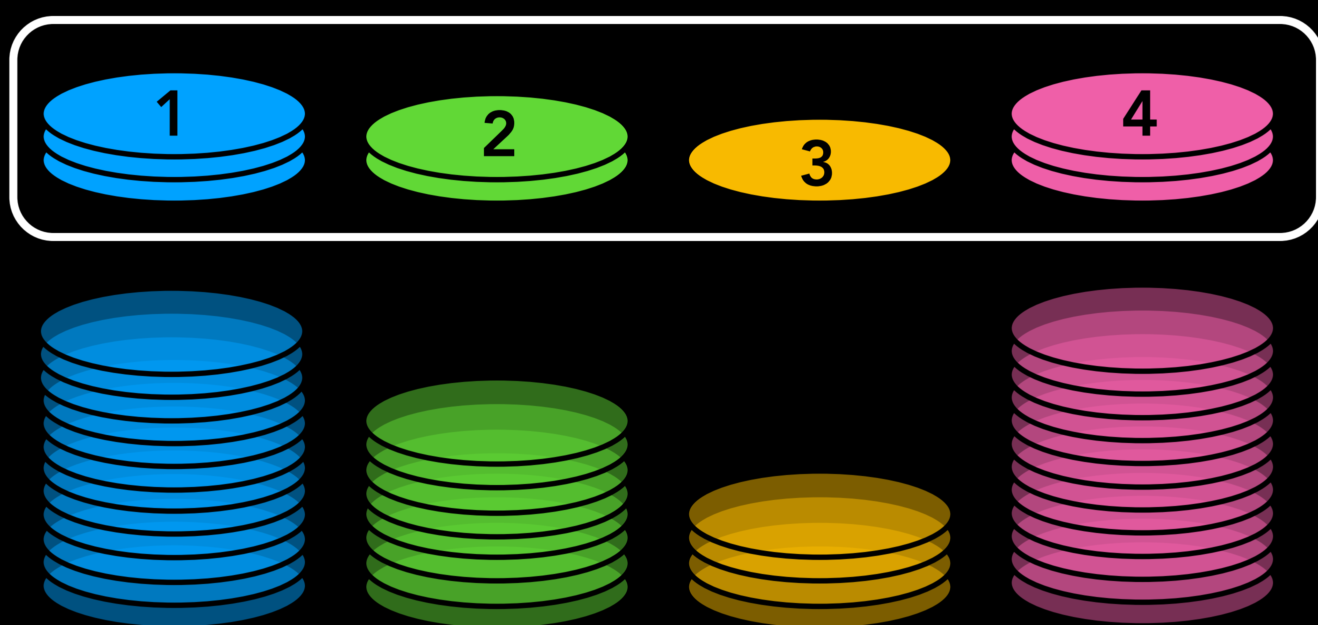
Bootstrapping: Simulating Many Samples From One

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



1	2	3	4		more extreme?
4	1	2	2		
3	3	3	0		
2	1	1	5		
…	…	…	…		

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.

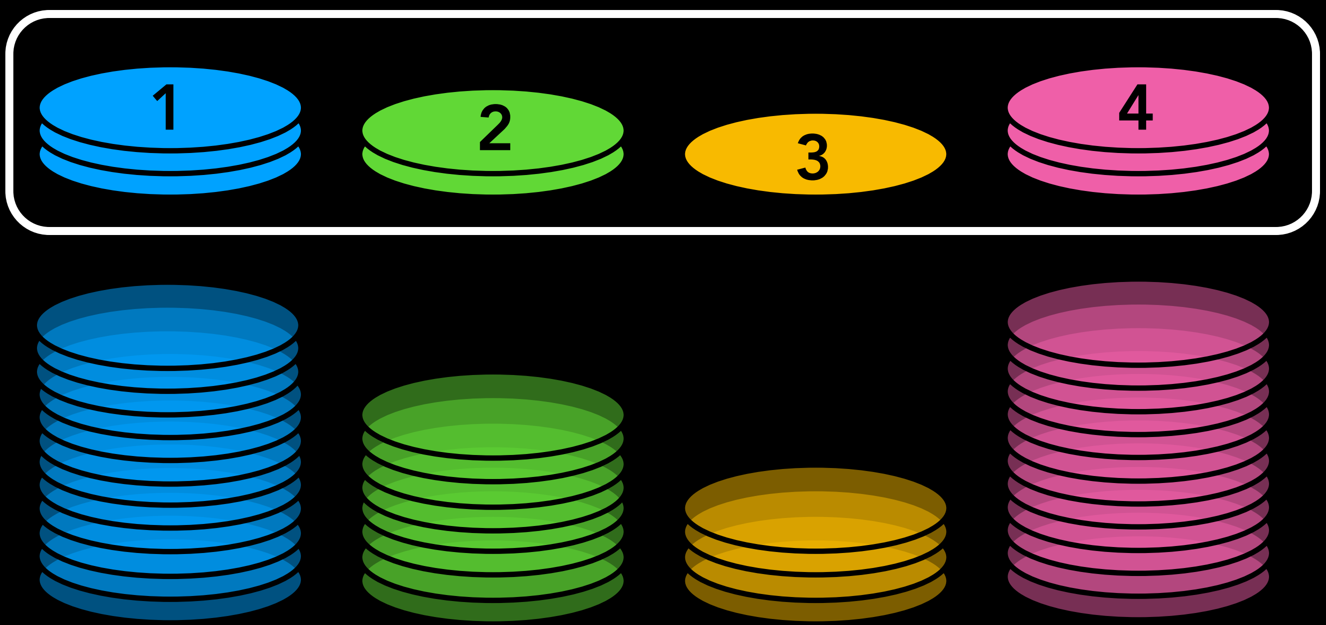
Bootstrapping: Simulating Many Samples From One

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



1	2	3	4	s^2	more extreme?
4	1	2	2	2.1	
3	3	3	0	0.8	
2	1	1	5	2.6	
…	…	…	…	…	

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.

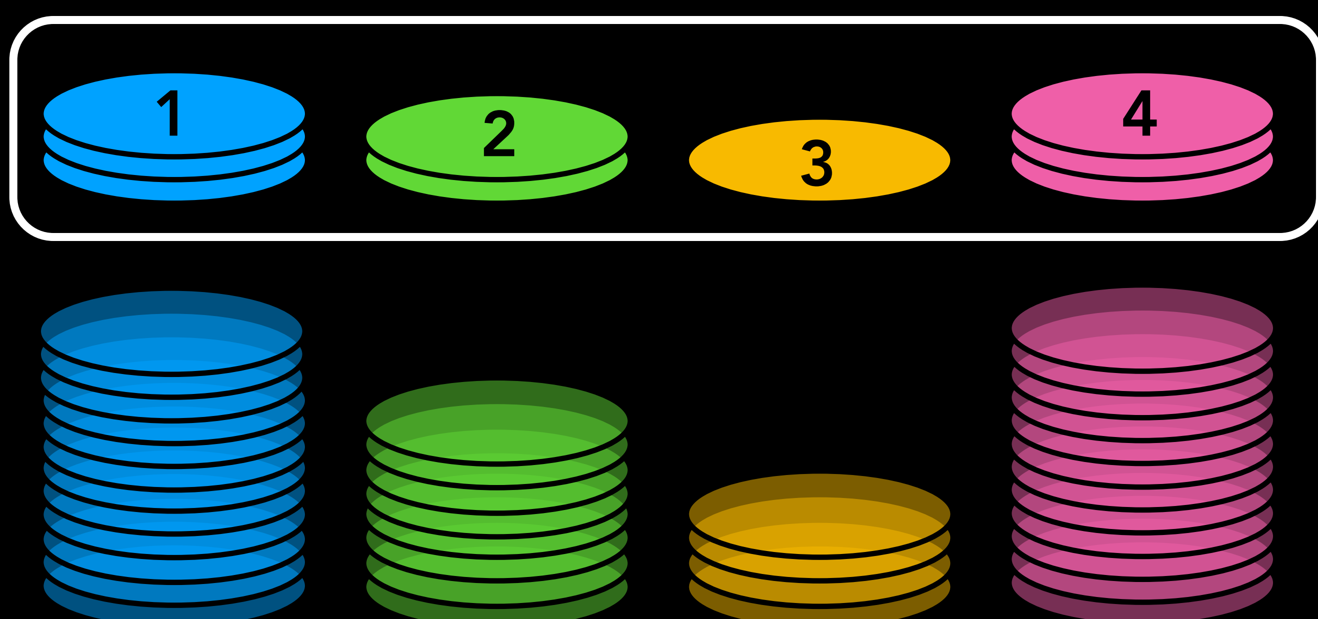
Bootstrapping: Simulating Many Samples From One

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

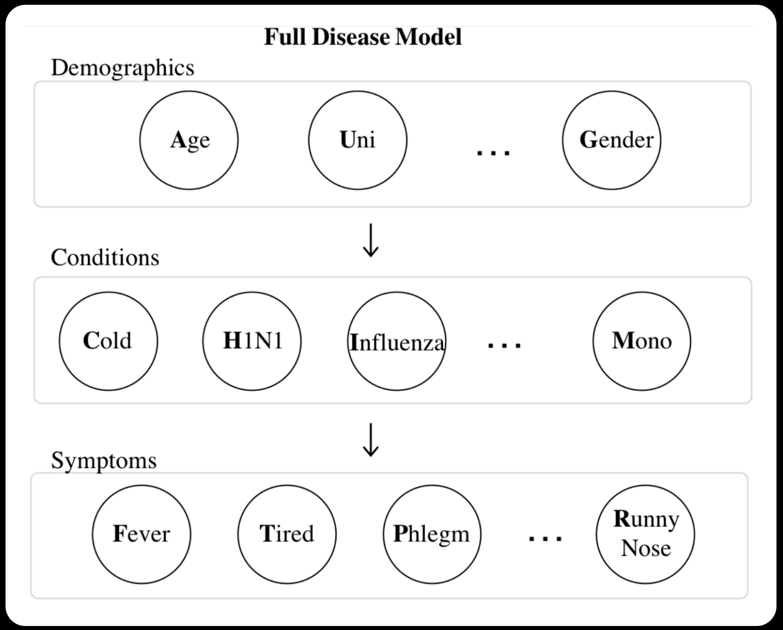
since our sample represents our population, we can sample from the data we have and it's as if we had gone out and collected more



1	2	3	4	S^2	more extreme?
4	1	2	2	2.1	no
3	3	3	0	0.8	no
2	1	1	5	2.6	yes
…	…	…	…	…	…

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

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

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

Think: What is reasonable to ask on a test about these topics?

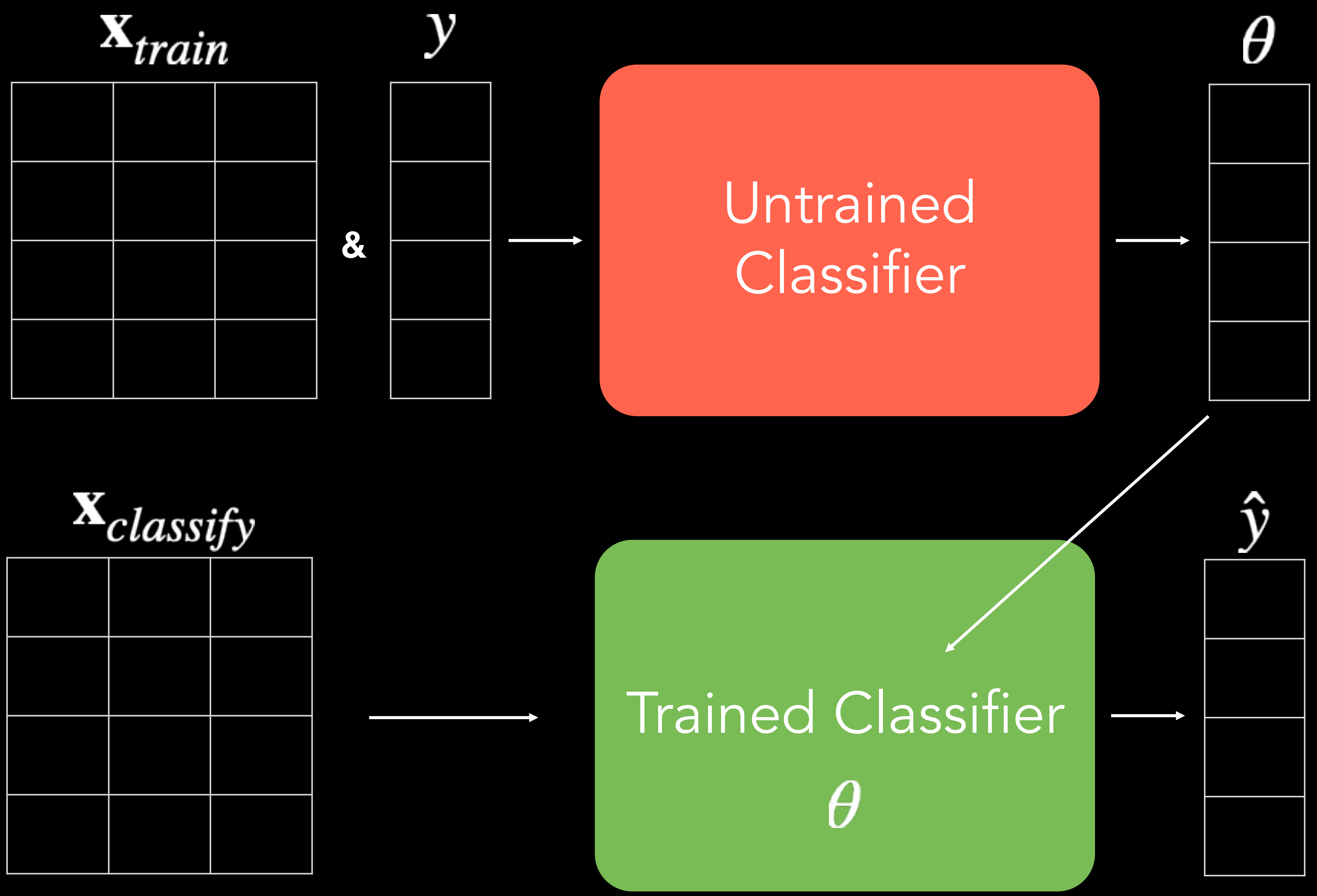
Remember the ideas of the algorithms, and practice turning them into high-level pseudocode:

Computing sample statistics
Boot-strapping p values

Joint Sampling
Rejection Sampling
Thompson Sampling


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

Classifiers



Parameter Estimation

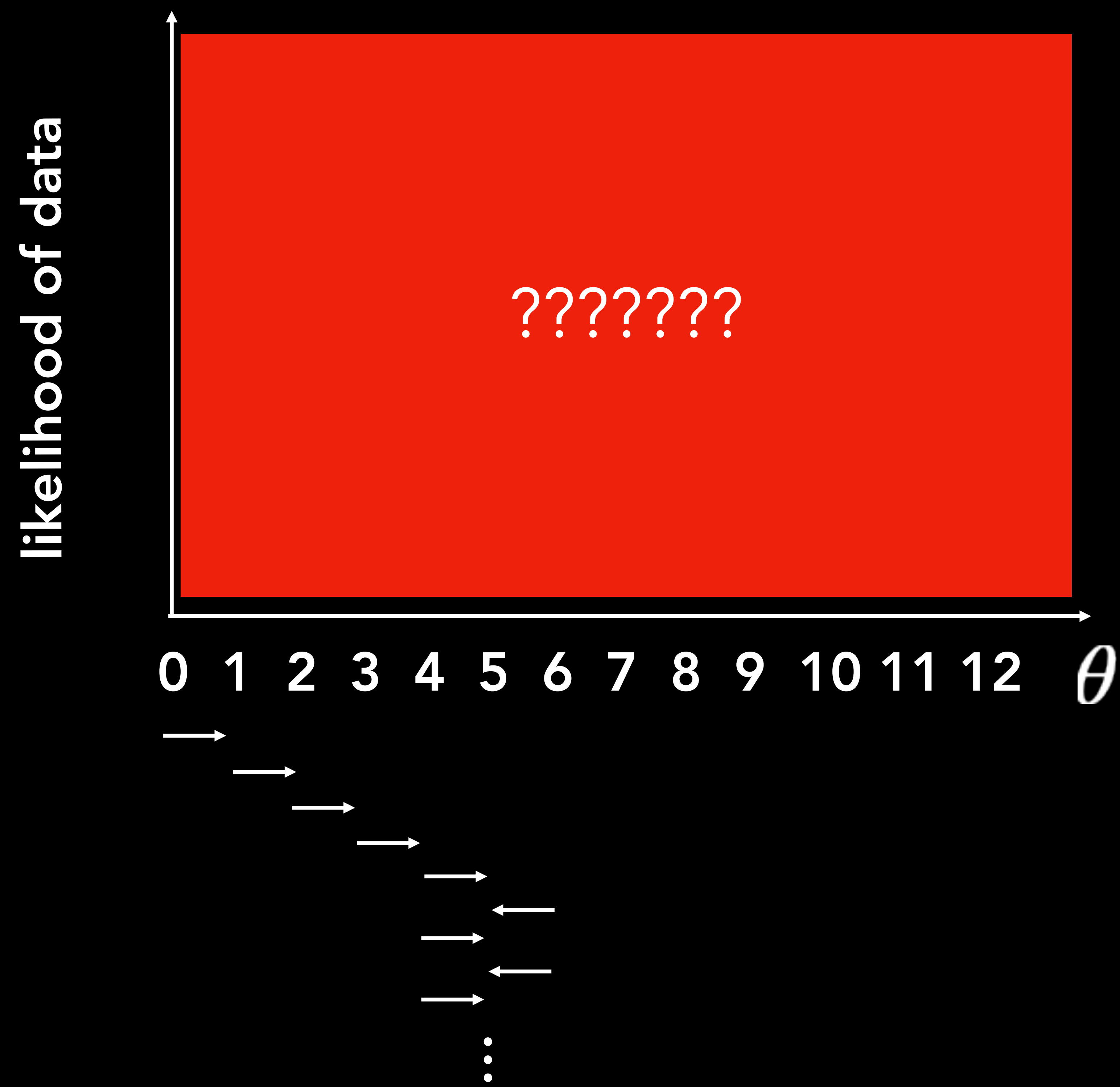
Maximum Likelihood Estimation

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



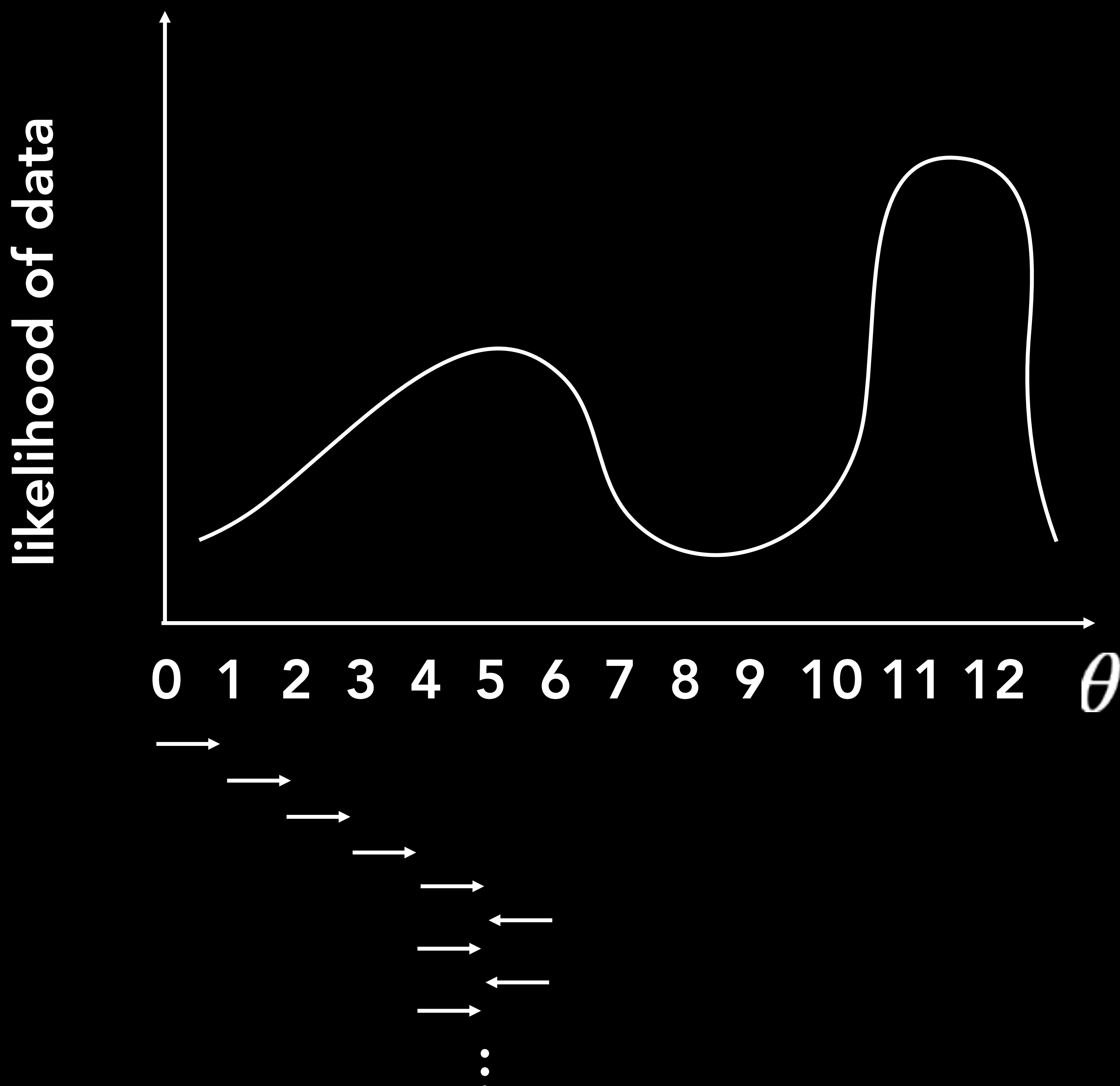
step size

$$\eta = 1$$

step direction

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

Gradient Ascent



step size

$$\eta = 1$$

step direction

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

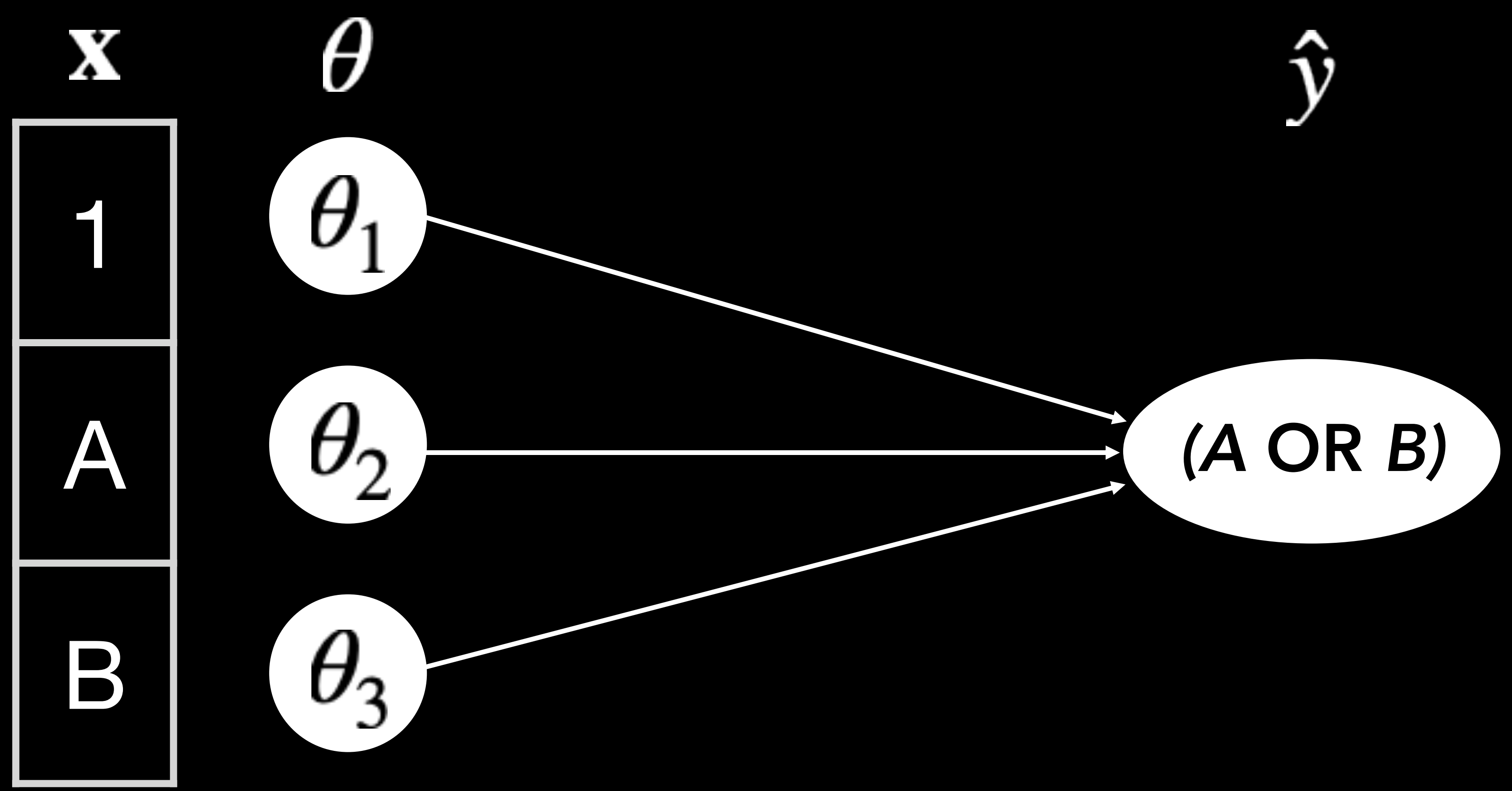
Classifier Algorithms

<u>Naïve Bayes</u>	Algorithm	<u>Logistic Regression</u>
All features in **x** are conditionally independent given classification	Assumption	Sigmoid gives us the probability of class 1
At train: Best estimates for prior on y and conditional likelihood of data	What are we optimizing/figuring out? At test: Whether y=0 or y=1 is the best guess	At train: The value(s) for **θ** such that the probability of our data is maximized
Learn (from data) estimates for $\hat{P}(Y=y), \hat{P}(X_i = x_i \mid Y=y)$: $\hat{P}(x_i \mid y) = \frac{(\textbf{ex. where } X_i = x_i and Y = y) + 1}{(\textbf{ex. where } Y = y) + 2}$ $\hat{P}(Y=y) = \frac{\textbf{ex. where } Y = y}{\textbf{total examples}}$	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 $LL(\theta) = \sum_{i=1}^{n} y^{(i)} \log \sigma(\theta^T\mathbf{x}^{(i)}) + (1-y^{(i)}) \log \left[1-\sigma(\theta^T\mathbf{x}^{(i)})\right] x_j$ $\frac{\partial LL(\theta)}{\partial \theta_j} = \sum_{i=1}^{n} \left[y^{(i)} - \sigma(\theta^T\mathbf{x}^{(i)})\right] x_j^{(i)}$

Logistic Regression

Example

one neuron (logistic regression model)

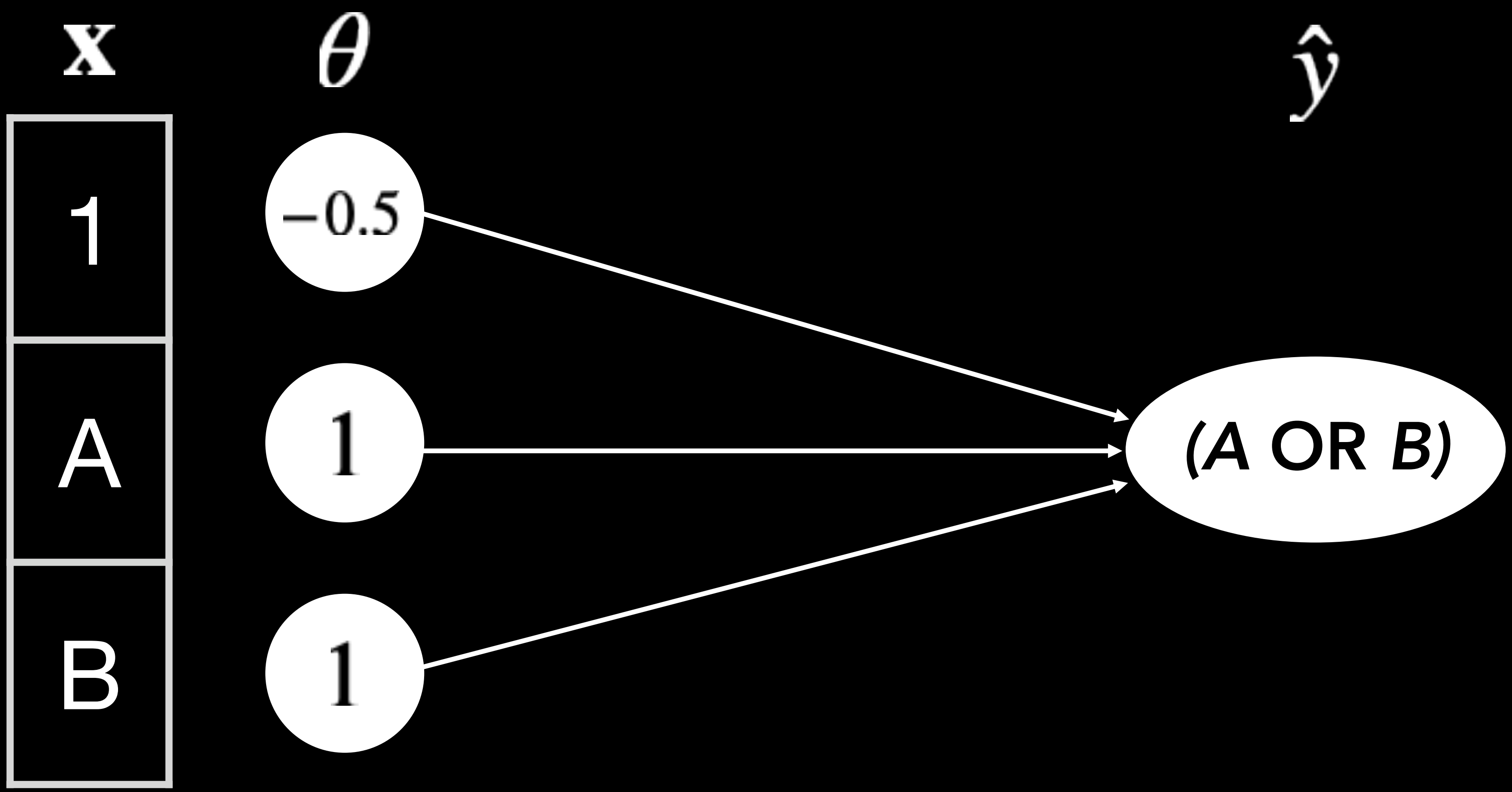


What weights do we have to learn for **θ**1, **θ**2, **θ**3 to perfectly classify data of the form (*A* OR *B*)?

Logistic Regression

Example

one neuron (logistic regression model)

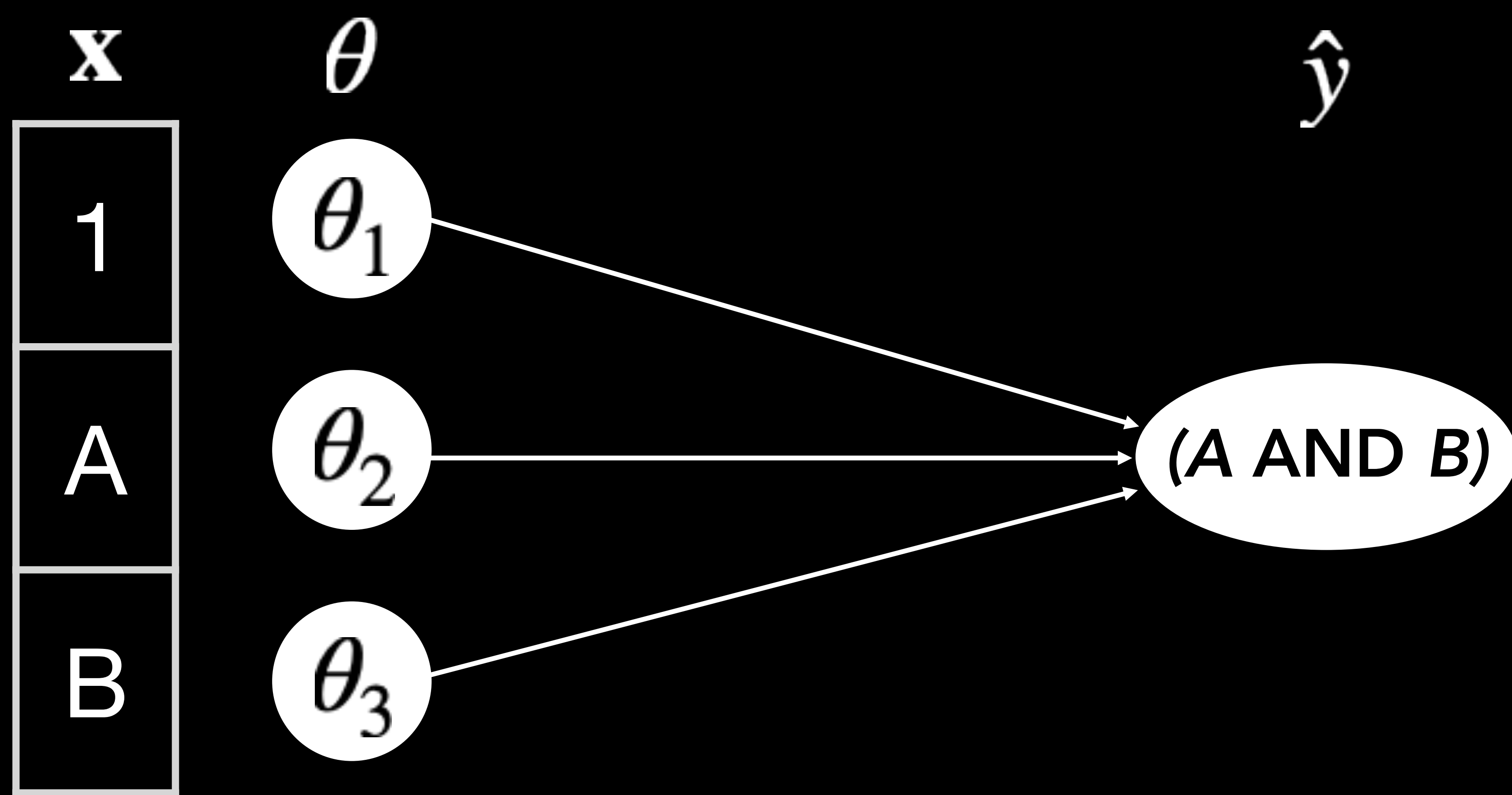


What weights do we have to learn for θ_1, θ_2, θ_3 to perfectly classify data of the form (*A* OR *B*)?

Example

Logistic Regression

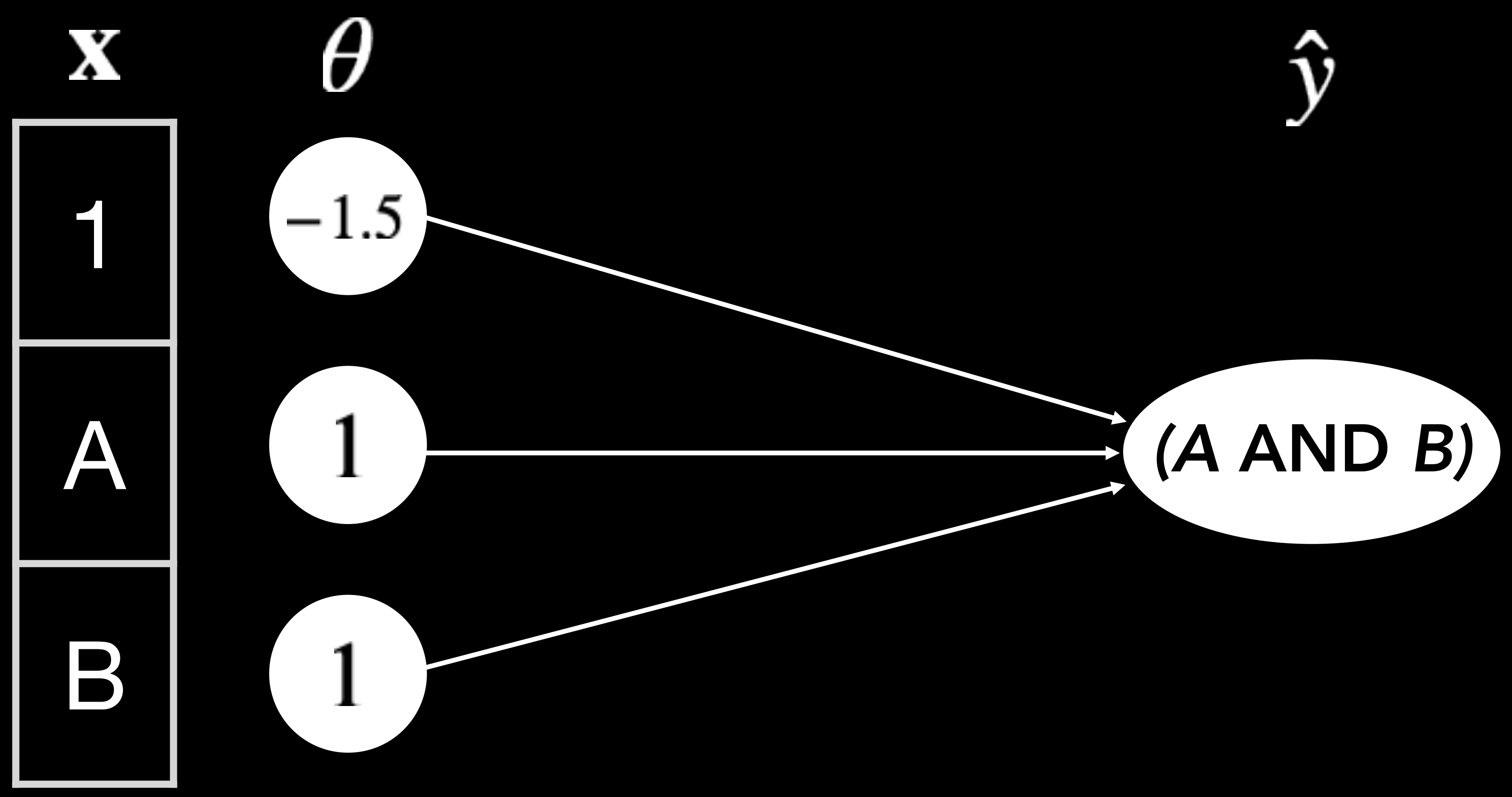
one neuron (logistic regression model)



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

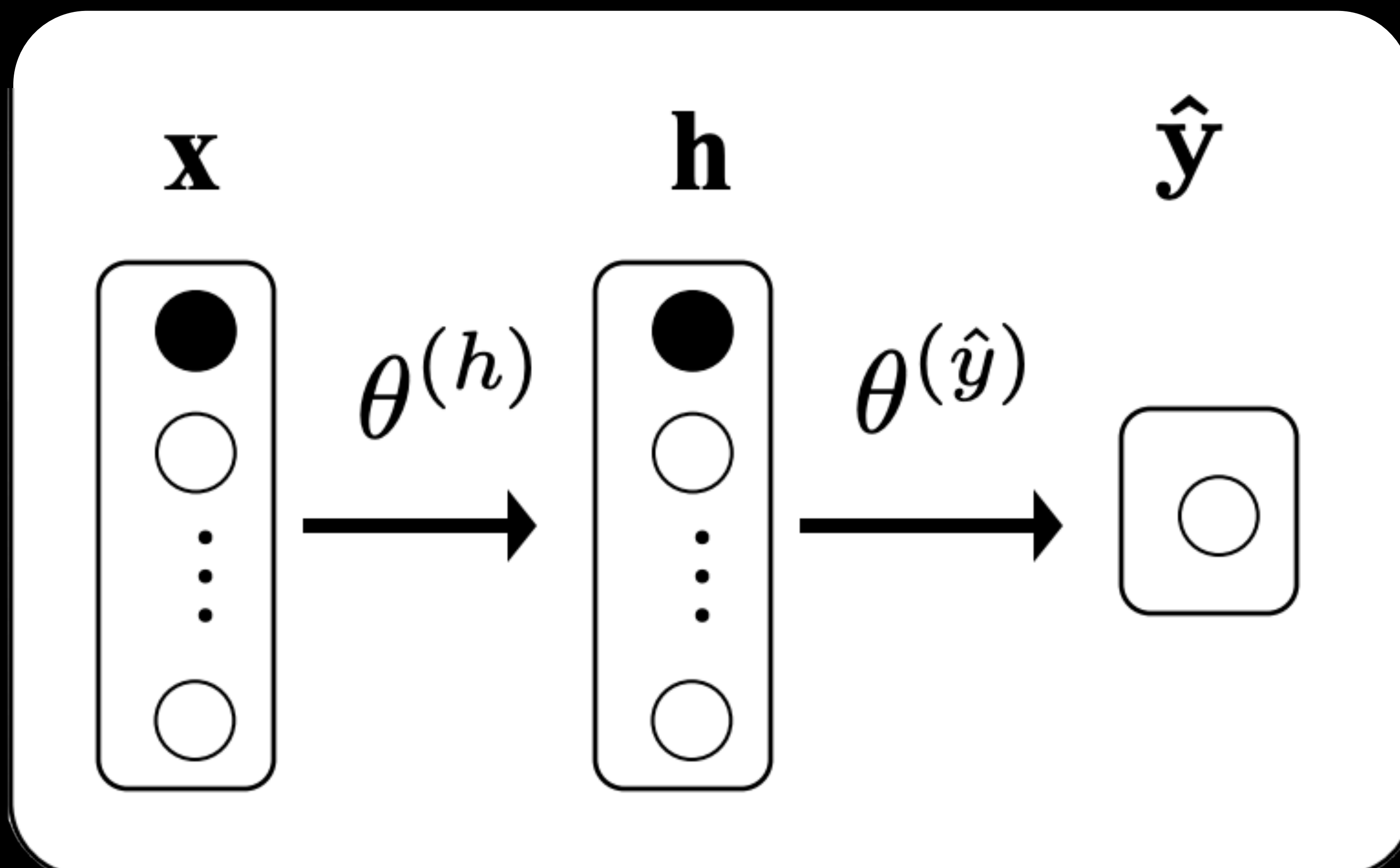
Logistic Regression

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

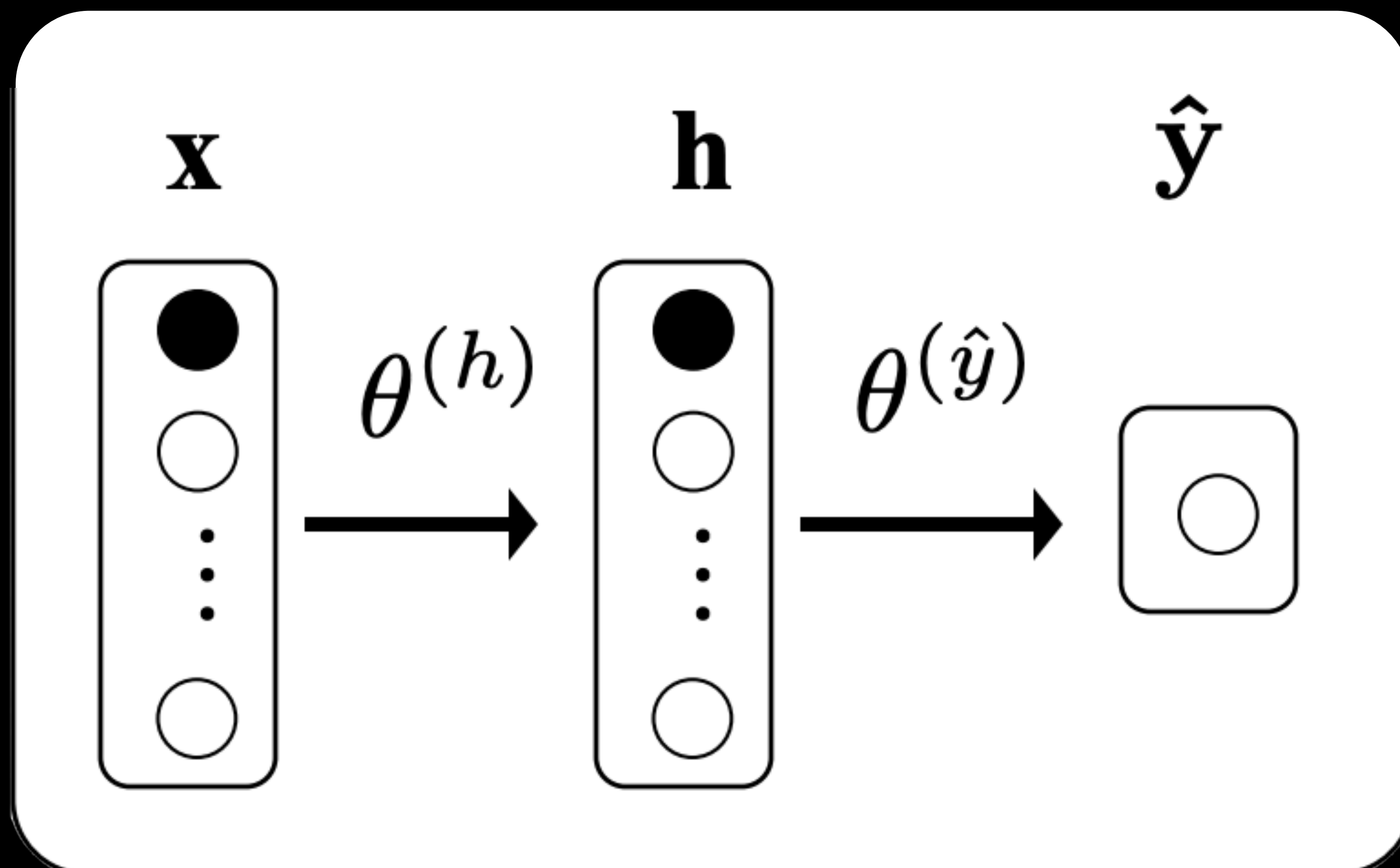


1. Make deep learning assumption: $P(Y = y \mid \mathbf{X} = \mathbf{x}) = (\hat{y})^y (1 - \hat{y})^{1-y}$
2. Calculate log likelihood for all data: $LL(\theta) = \sum_{i=0}^{n} y^{(i)} (\log \hat{y}^{(i)}) + (1 - \hat{y}^{(i)}) \log [1 - \hat{y}^{(i)}]$
3. Find partial derivative of LL with respect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

   use the chain rule!

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

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:

   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!