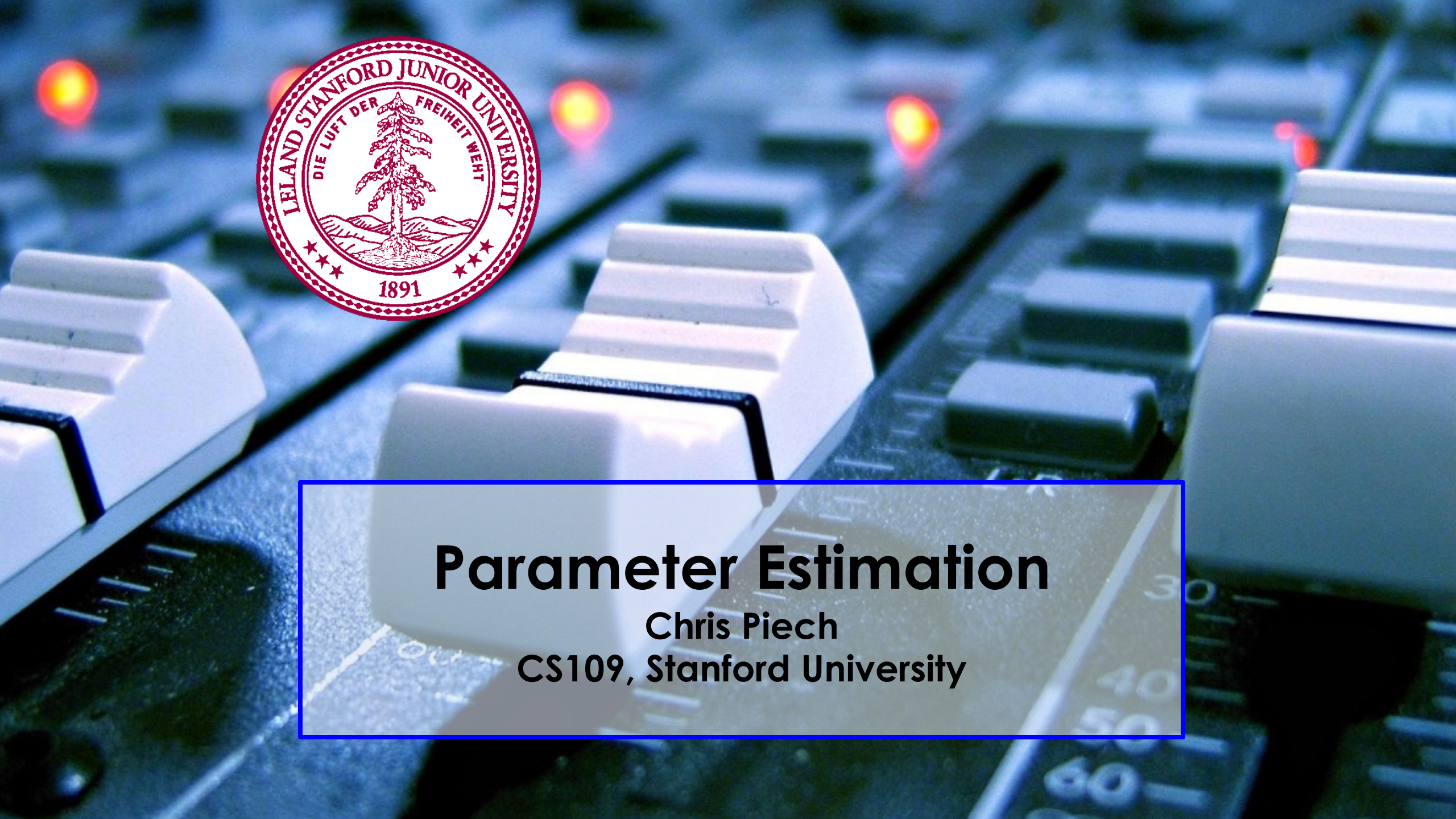


Parameter Estimation

Chris Piech
CS109, Stanford University

MACHINE LEARNING

Finishing up!

Expectation

The probability that X takes on that value

$$\mathrm{E}[X] = \sum_{x} x \cdot P(X = x)$$

All the values that X can take on

Expectation of a Function

Law of unconscious statistician

$$\mathrm{E}[g(X)] = \sum_x g(x) \cdot P(X = x)$$

So for example...

$$\mathrm{E}[X^2] = \sum_x x^2 \cdot P(X = x)$$

Conditional Expectation

The probability that X takes on that value

$$\mathrm{E}[X|Y=y] = \sum_{x} x \cdot P(X=x|Y=y)$$

All the values that X can take on

Function or Number?

$$E[X|Y=1]$$

Number

Function or Number?

$$E[X = 2]$$

Error.

Expectation is defined on a random variable

Function or Number?

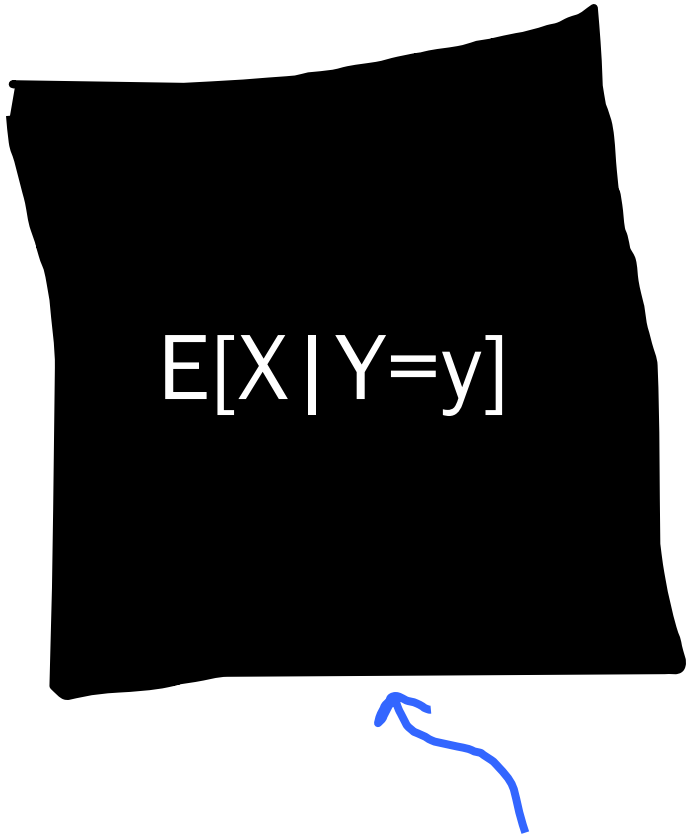
$$E[X|Y = y]$$

Function

Conditional Expectation

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

This is just function of Y

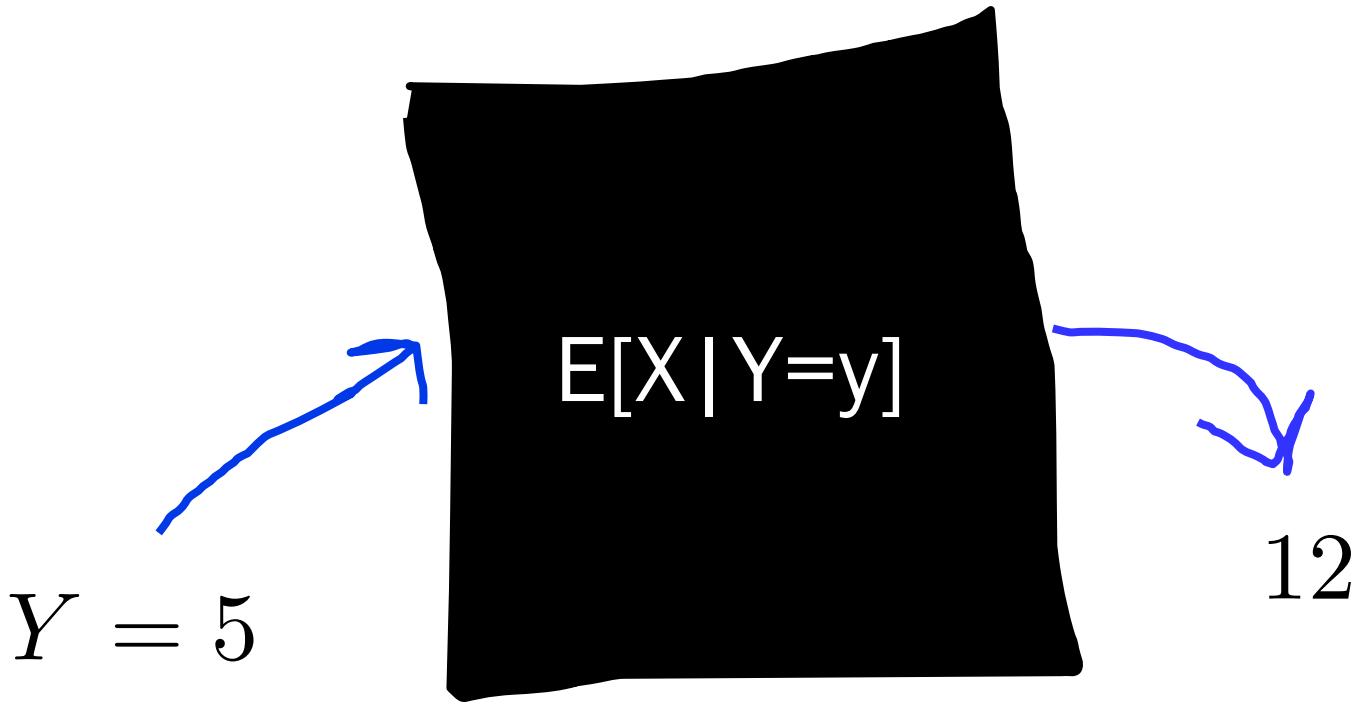


This is a function with Y as input

Conditional Expectation

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

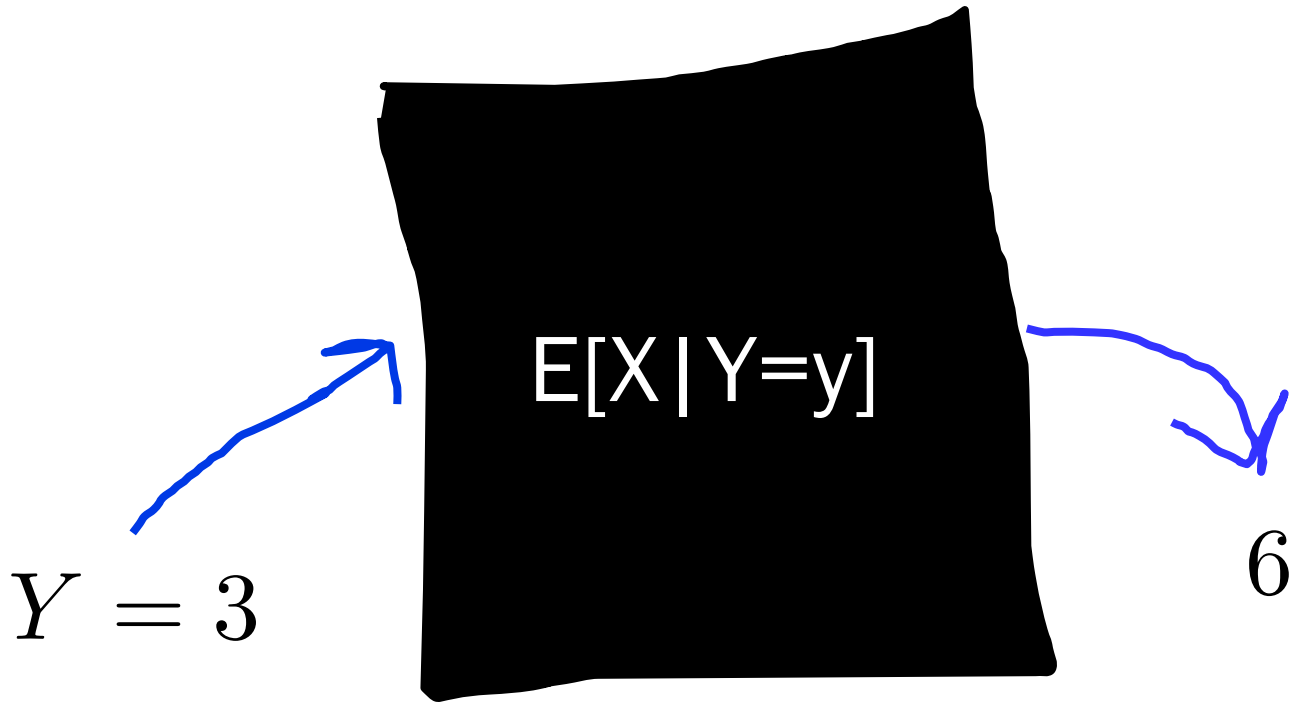
This is just function of Y



Conditional Expectation

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

This is just function of Y



Conditional Expectation

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

This is a number:

$$E[X]$$

This is a function of y:

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

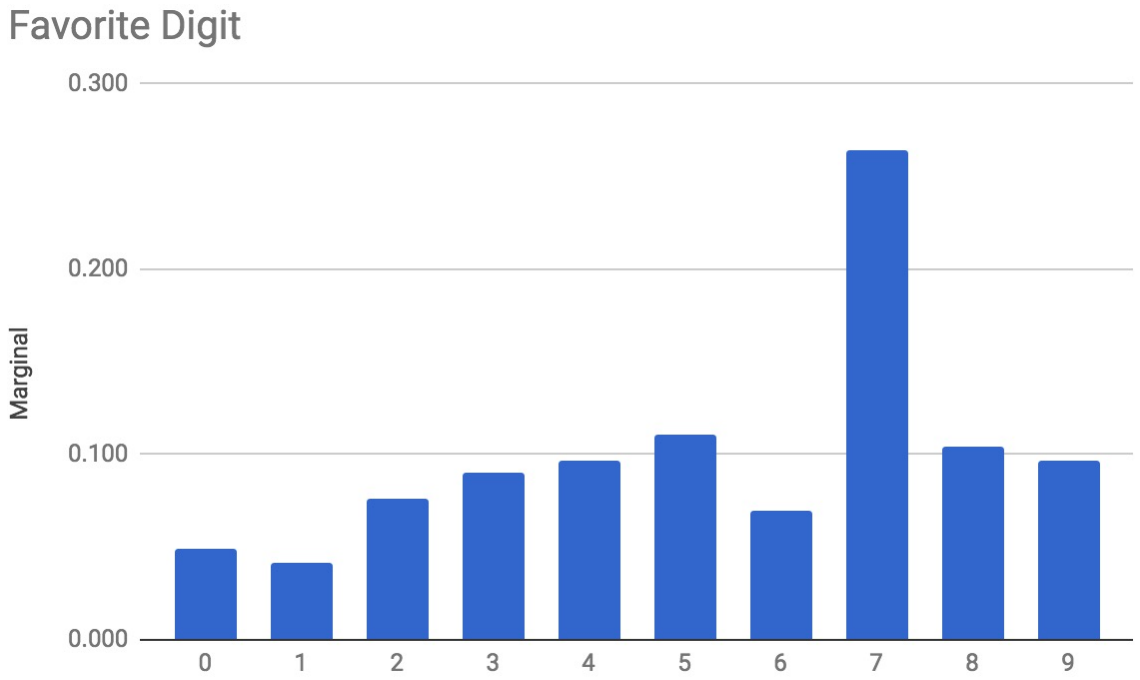
$E[X=5]$

Doesn't make sense. Take expectation of random variables, not events

Expectation

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

X = favorite number
Y = year in school



E[X] = 0 * 0.05 + ... + 9 * 0.10 = 5.38

Conditional Expectation

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

X = favorite number
Y = year in school

E[X | Y] ?

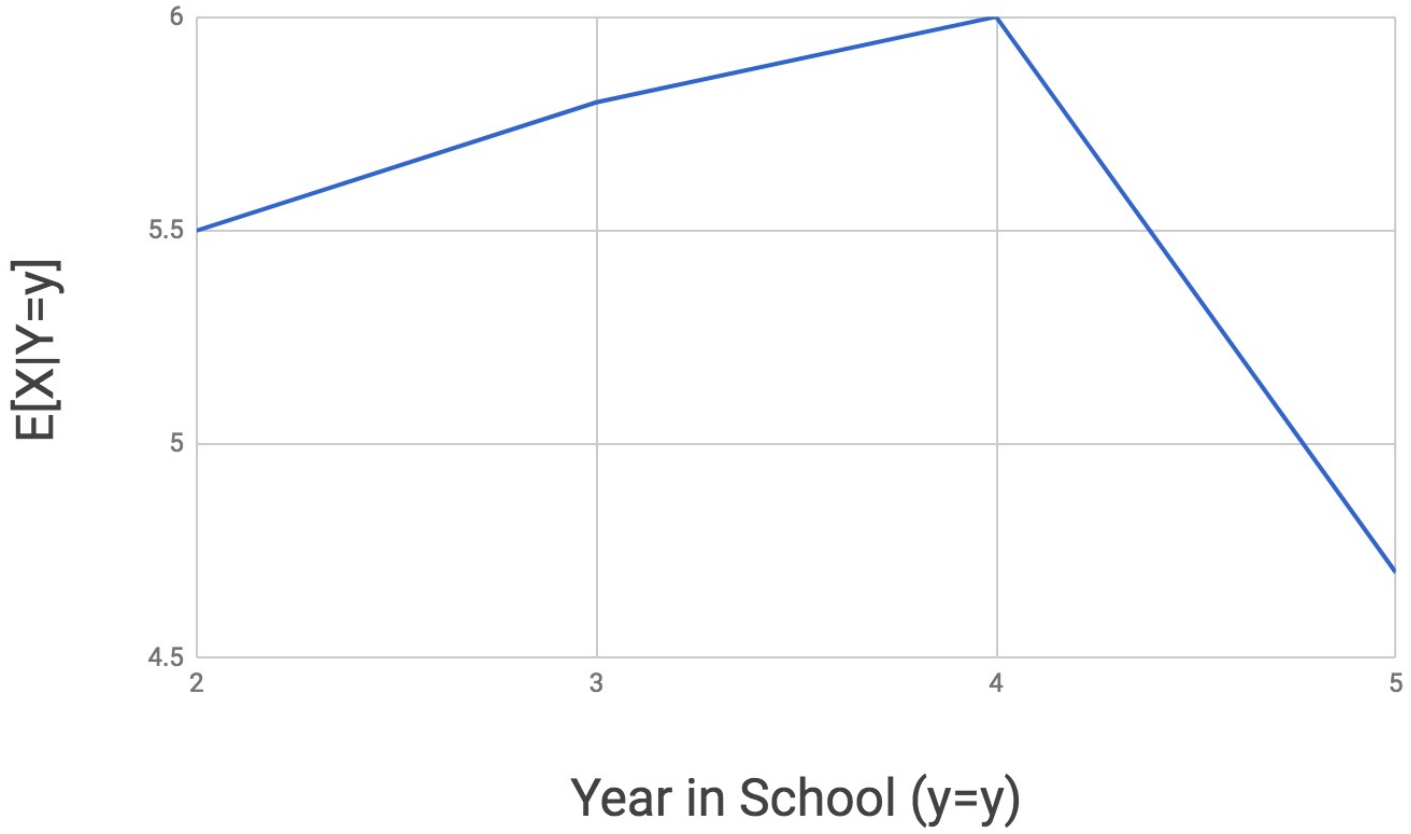
Year in school, Y = y	E[X | Y = y]
2	5.5
3	5.8
4	6.0
5	4.7

Conditional Expectation

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

X = favorite number
Y = year in school

E[X | Y] ?

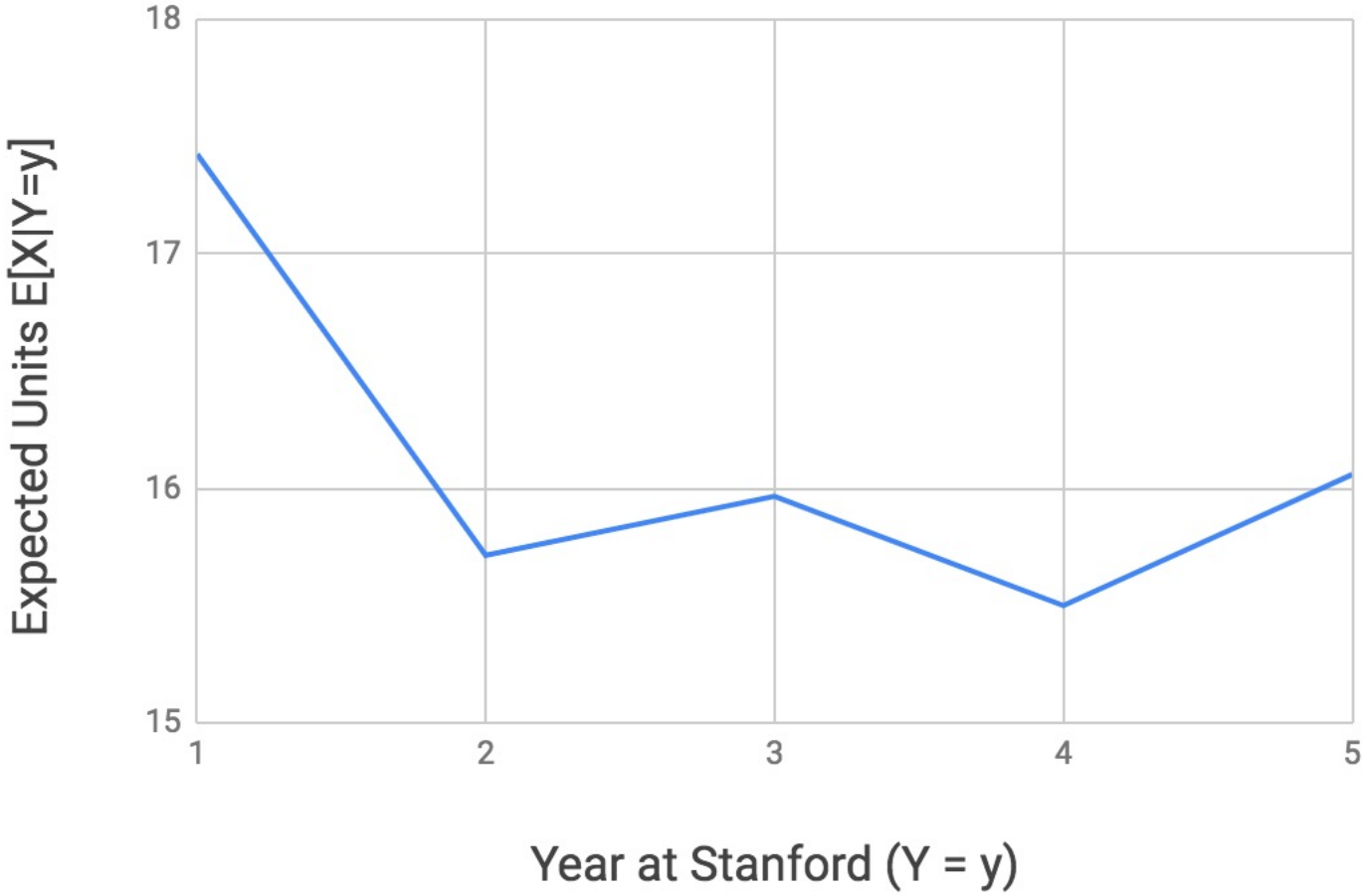


Conditional Expectation

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

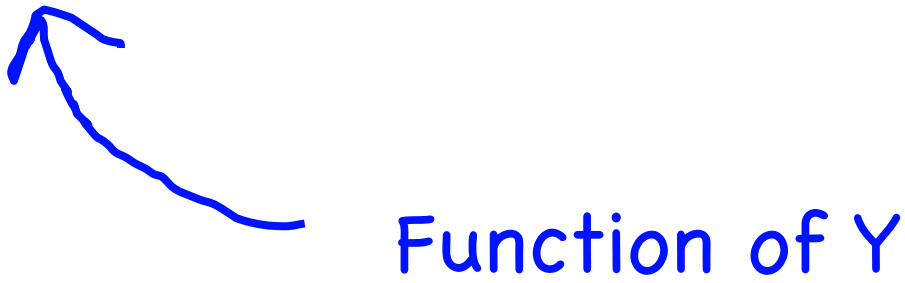
X = units in fall quarter
Y = year in school

E[X | Y] ?



What is this???

$$E[E[X|Y]]$$



Law of Total Expectation

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

Law of unconscious statistician

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

$$= \sum_y \sum_x xP(X=x|Y=y)P(Y=y)$$ Def of E[X|Y]

$$= \sum_y \sum_x xP(X=x, Y=y)$$ Chain rule!

$$= \sum_x \sum_y xP(X=x, Y=y)$$ I switch the order of the sums

$$= \sum_x x \sum_y P(X=x, Y=y)$$ Move that x outside the y sum

$$= \sum_x xP(X=x)$$ Marginalization

$$= E[X]$$ Def of E[X]

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]?

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

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

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

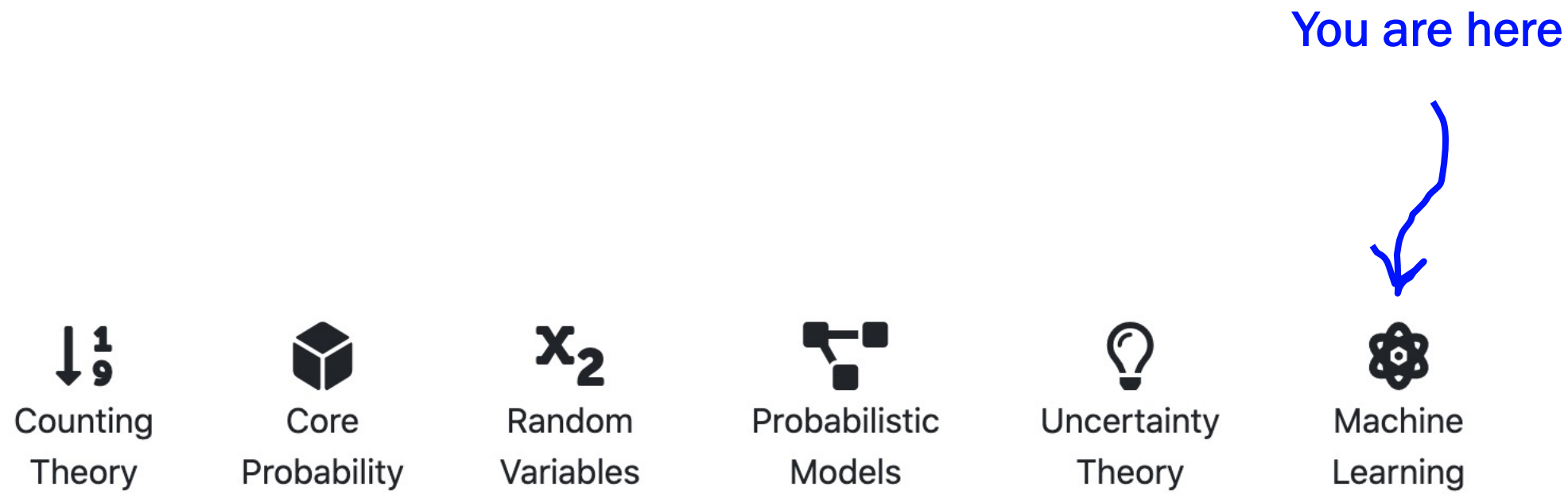
$$E[Y \mid X = 3] = E[7 + Y] = 7 + E[Y]$$

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

$$E[Y] = 15$$

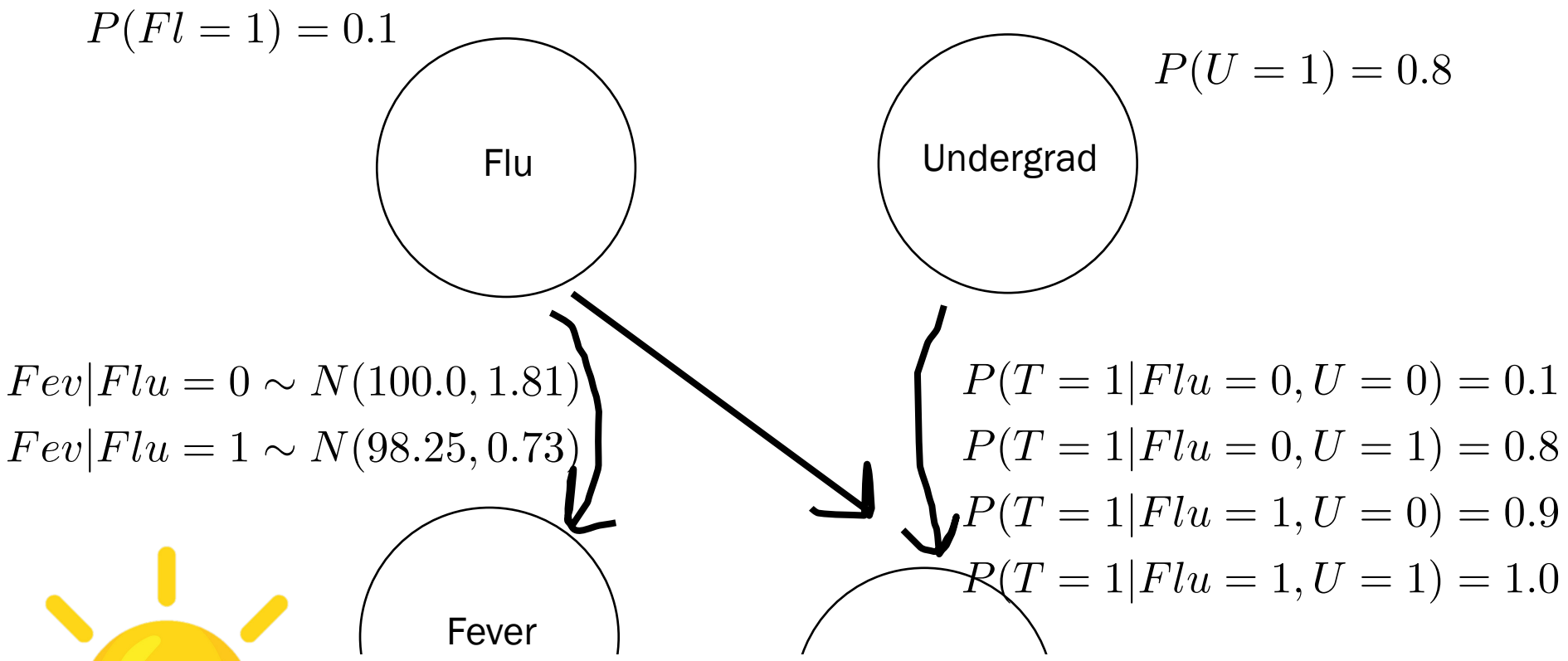
Protip: do this in CS161

Where are we in CS109?



General “Inference”

Probabilistic Model



If you know the probability of each random variables given the ones that directly cause it, you can joint sample!

But where do those numbers come from?

Suspense

At this point, if you are given a *model*, with all the involved probabilities, you can make predictions

But what if you want to *learn* the probabilities in the model?

But what if you want to *learn* the probabilities in the model?

Oh can we also learn the *structure* of the model too?

But what if you want to *learn* the probabilities in the model?

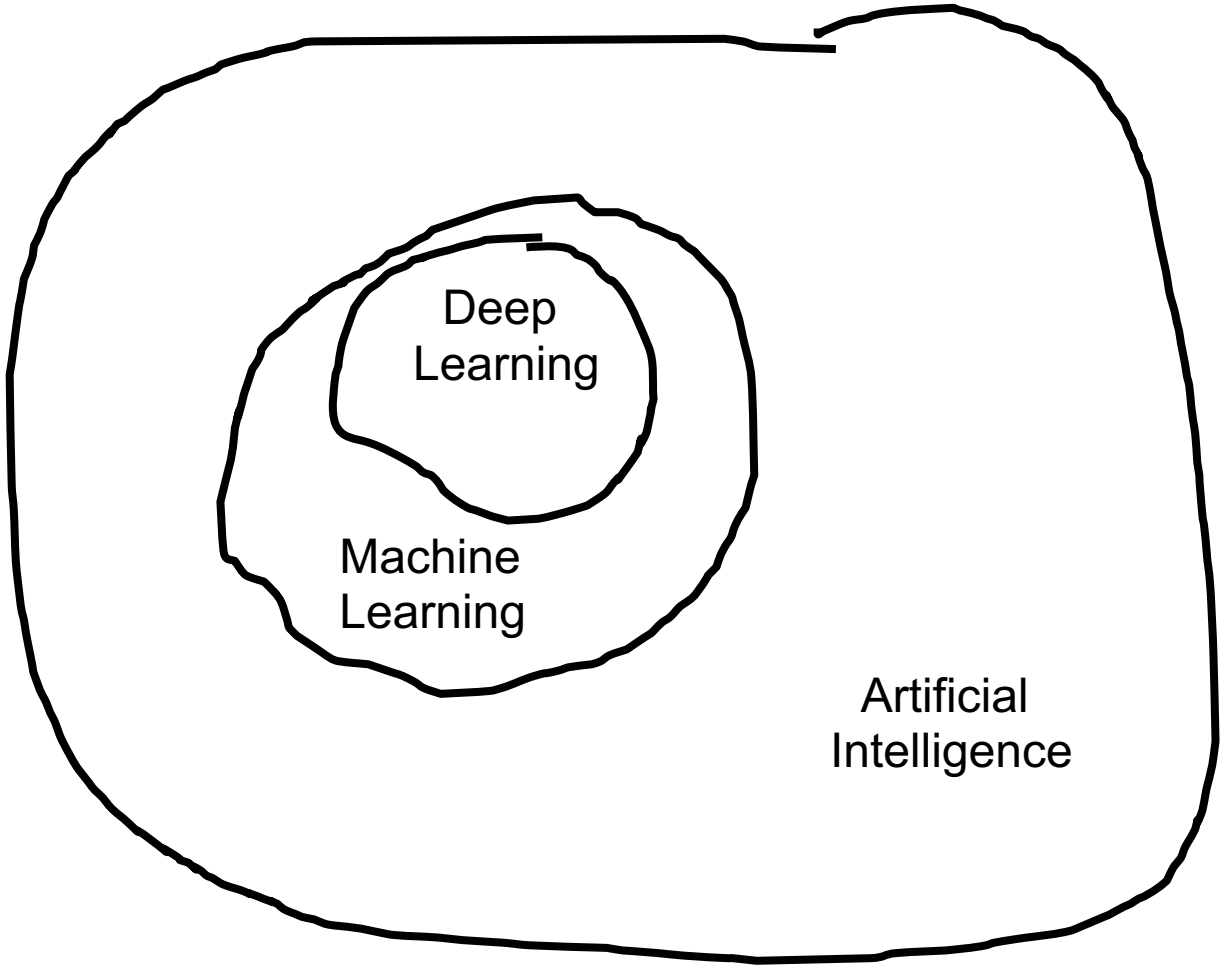
~~Oh can we also learn the *structure* of the model too?~~

I wish. Another day ☺

But what if you want to *learn* the probabilities in the model?

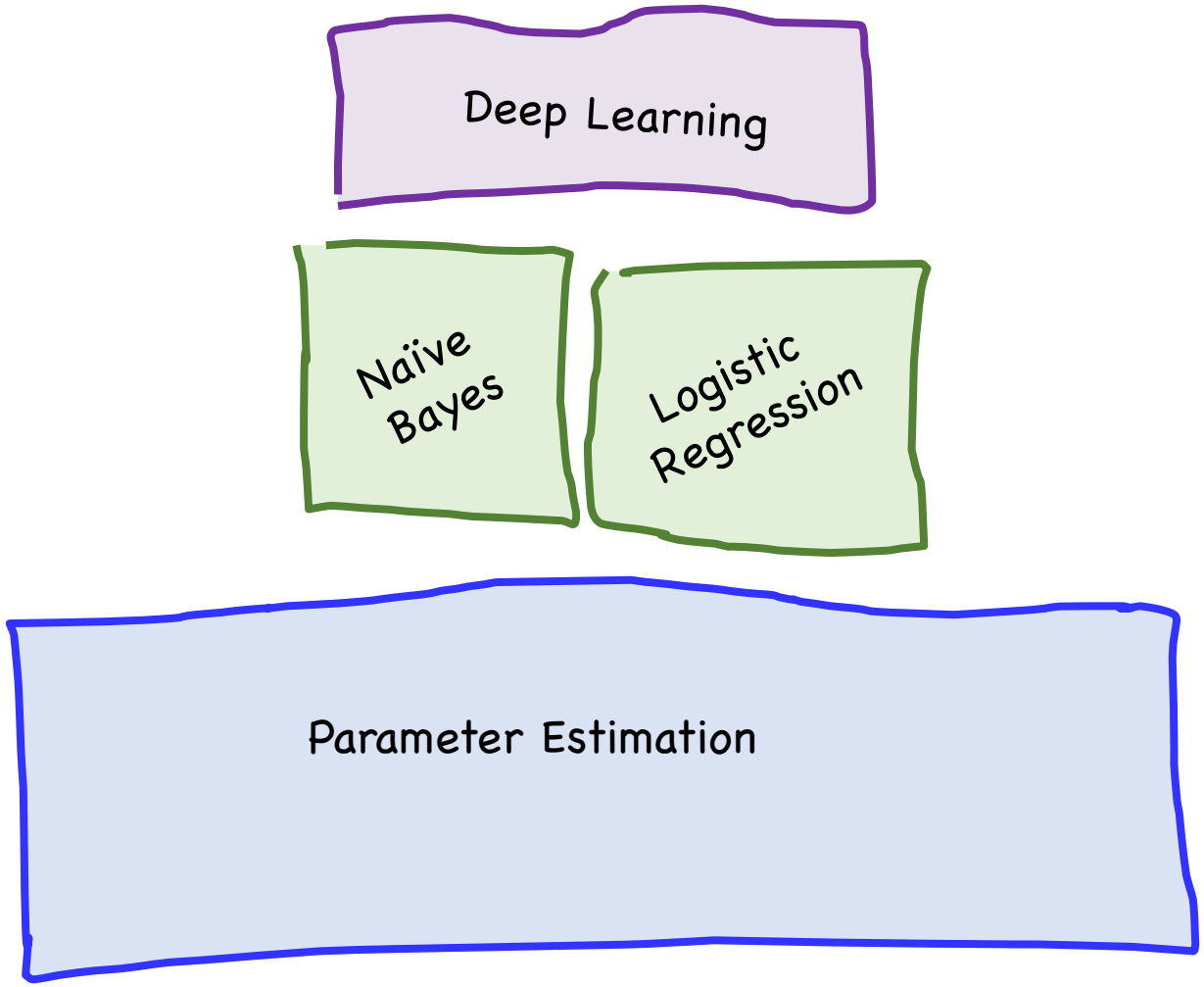
Machine Learning

AI and Machine Learning

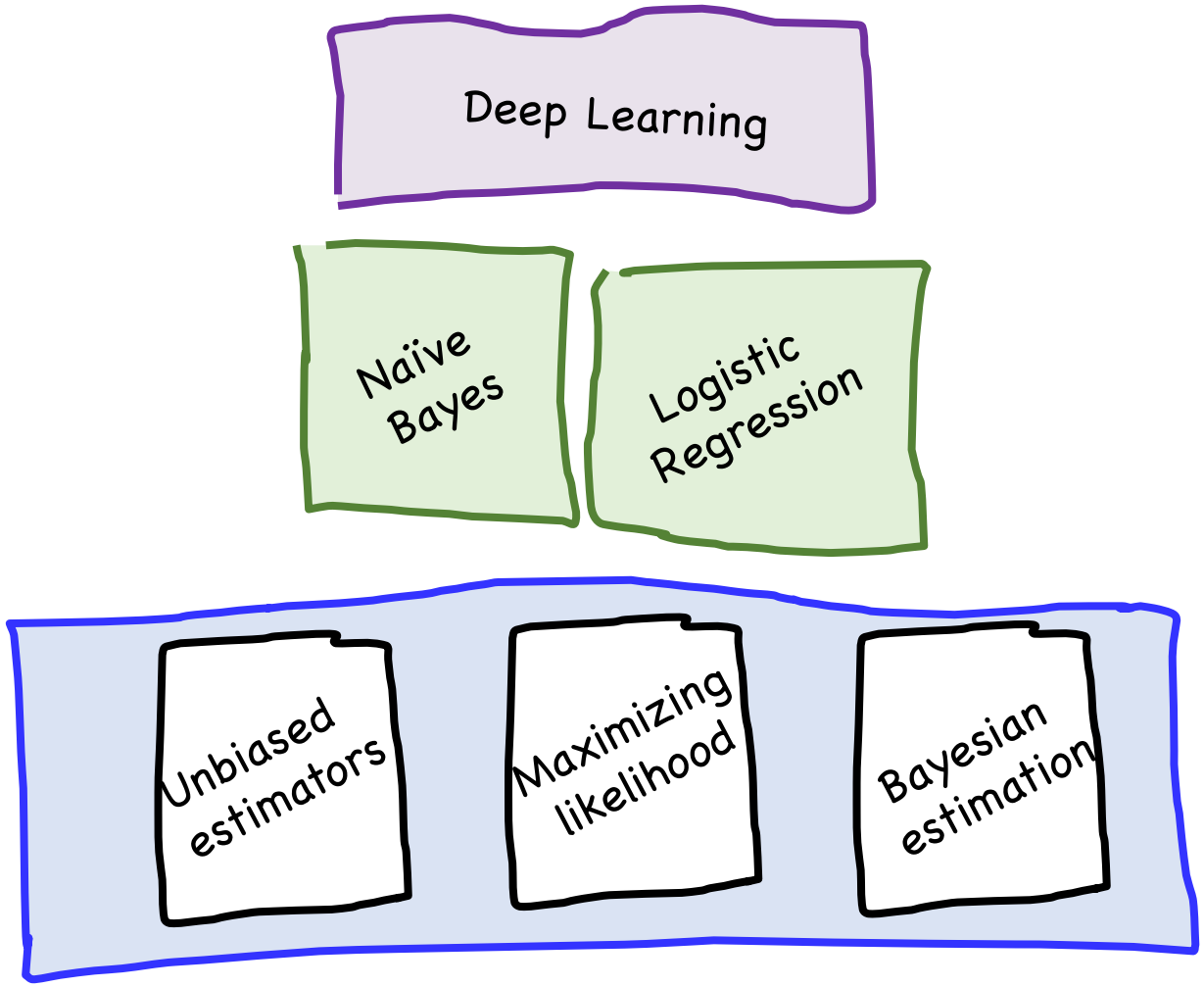


ML: Rooted in probability theory

Our Path


Deep Learning
Naïve Bayes
Logistic Regression
Parameter Estimation

Our Path


Deep Learning
Naïve Bayes
Logistic Regression
Unbiased estimators
Maximizing likelihood
Bayesian estimation

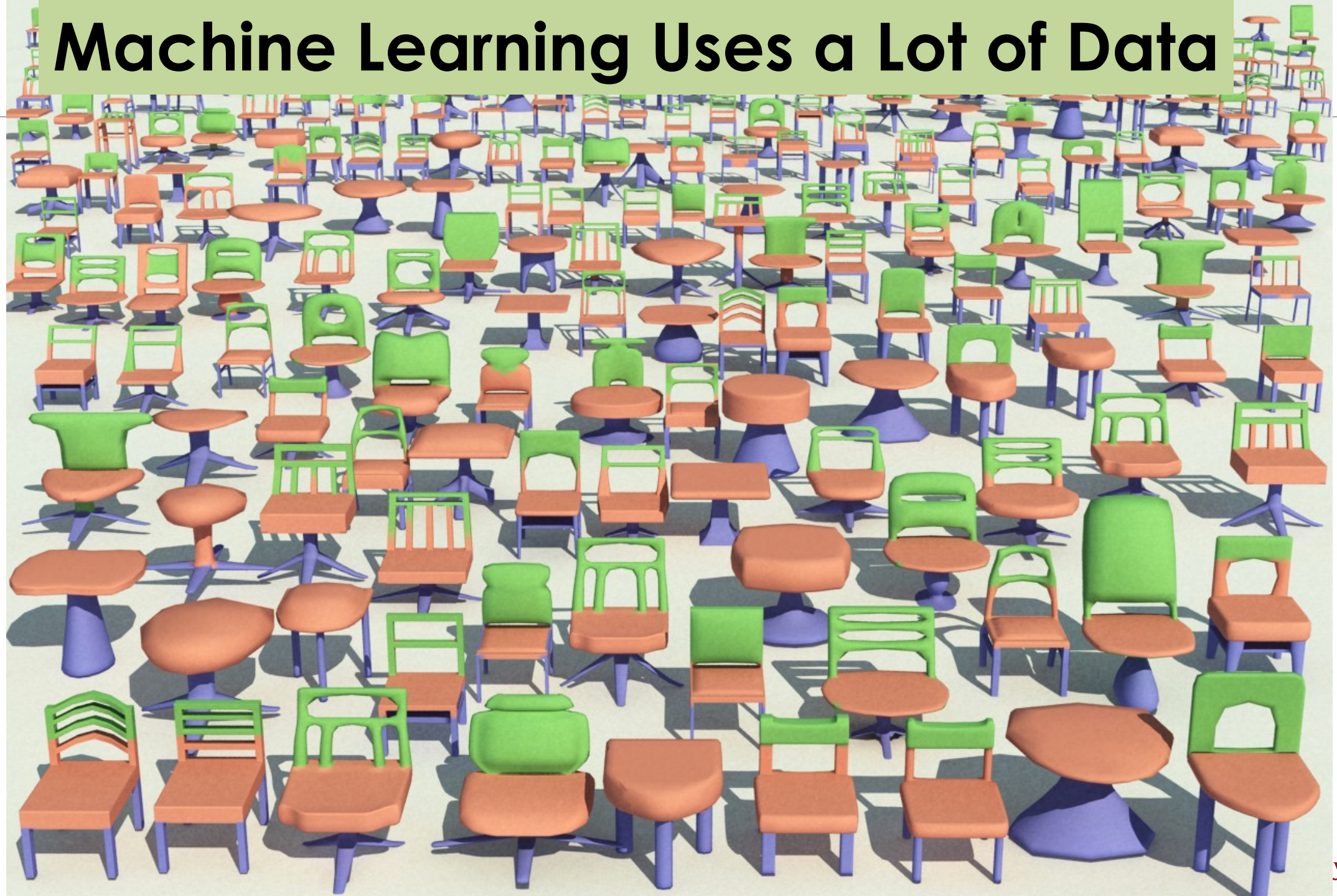
Jump Straight to Deep Learning?

Jump Straight to Deep Learning?

Understand the theory to help you debug

But another reason…

Machine Learning Uses a Lot of Data

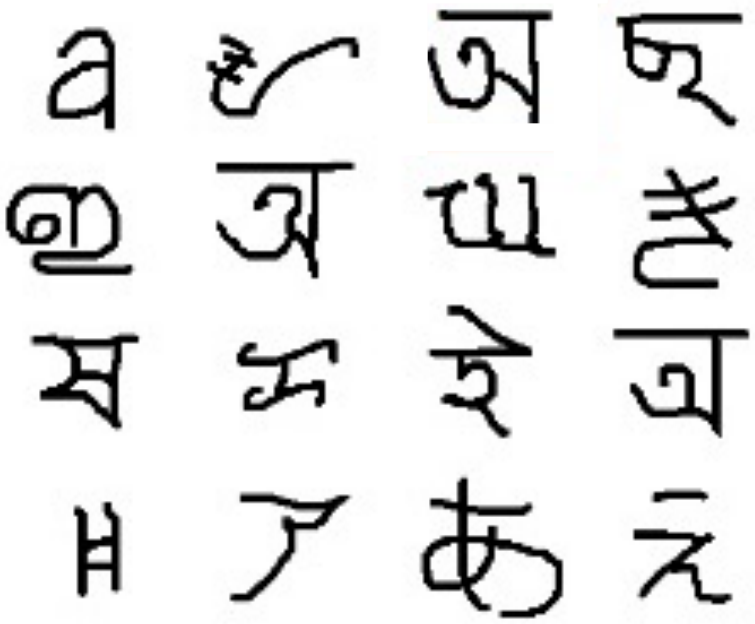


y

One Shot Learning

Single training example:

Test set:



One Shot Learning

Single training example:

Computers struggle...

… especially for **human** problems.

Understand the theory
to push on the **grand challenges**


WALT DISNEY
PICTURES

Once upon a time…

…there was parameter estimation

What are Parameters?

Consider some probability distributions:

- Ber(p)
- Poi(λ)
- Uni(α, β)
- Normal(μ, σ^2)
- $Y = mX + b$
- etc…

$\theta = p$

$\theta = \lambda$

$\theta = (\alpha, \beta)$

$\theta = (\mu, \sigma^2)$

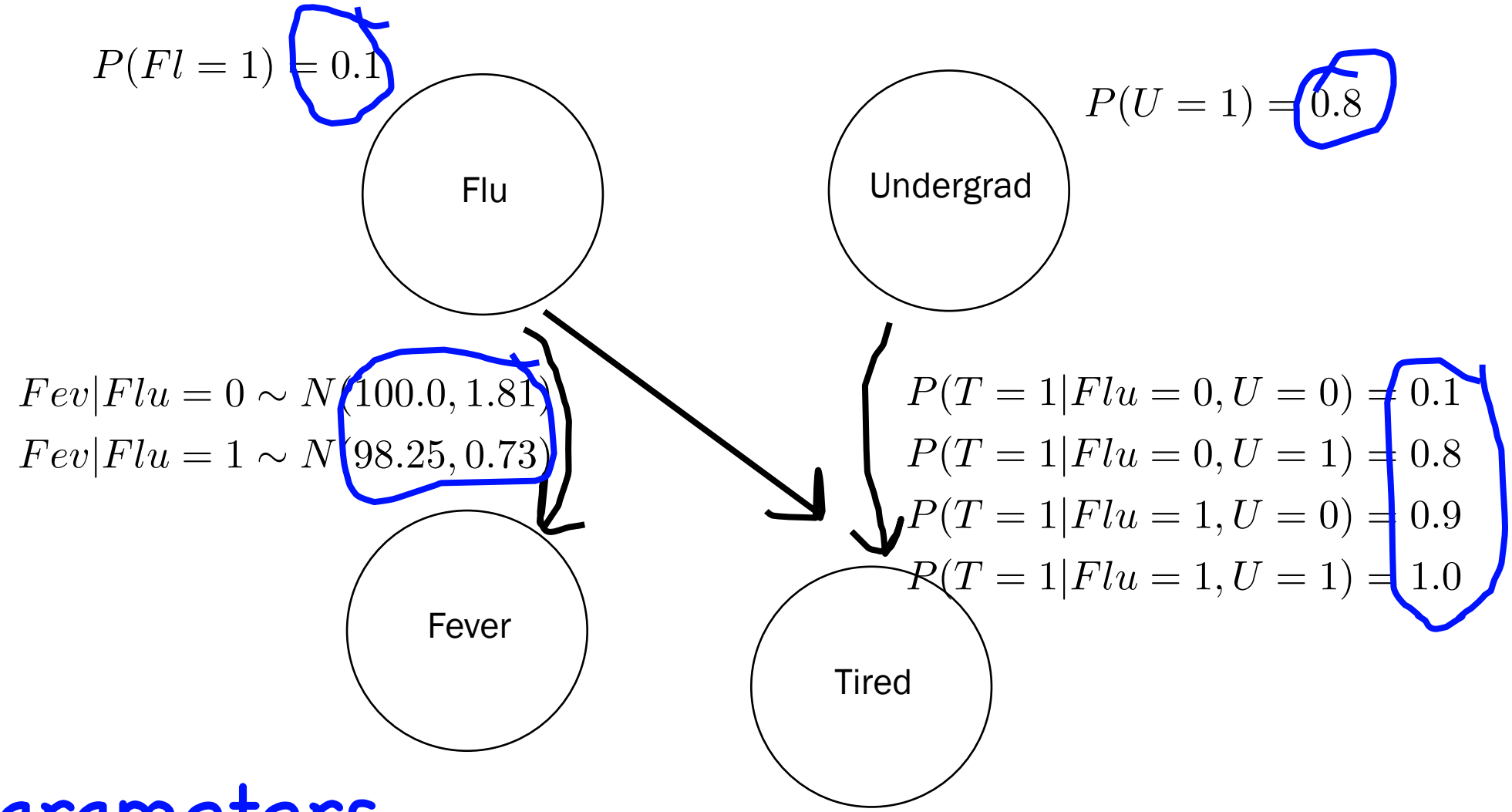
$\theta = (m, b)$

Call these “parametric models”

Given model, **parameters** yield actual distribution

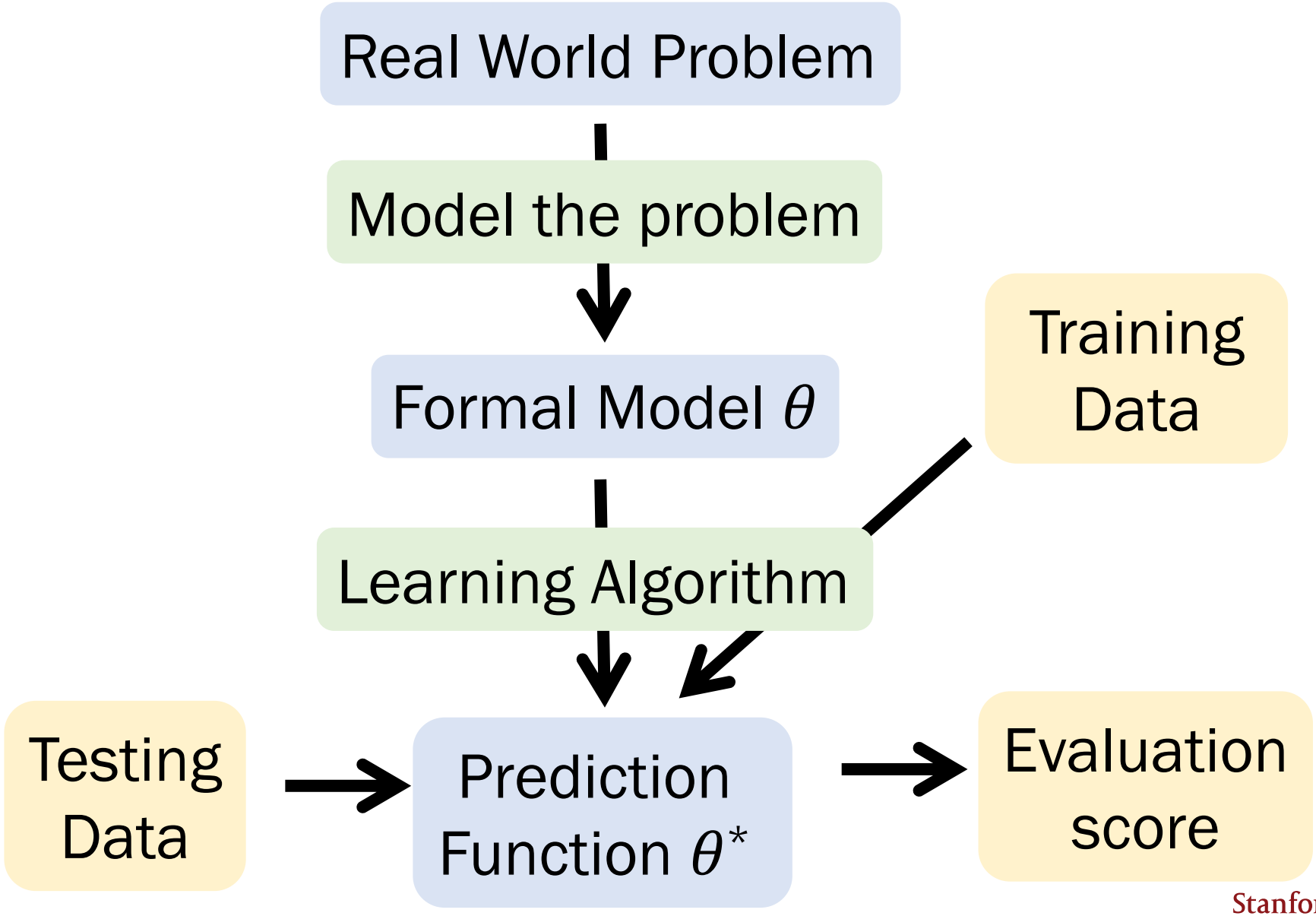
- Usually refer to parameters of distribution as θ
- Note that θ that can be a vector of parameters

What are Parameters?

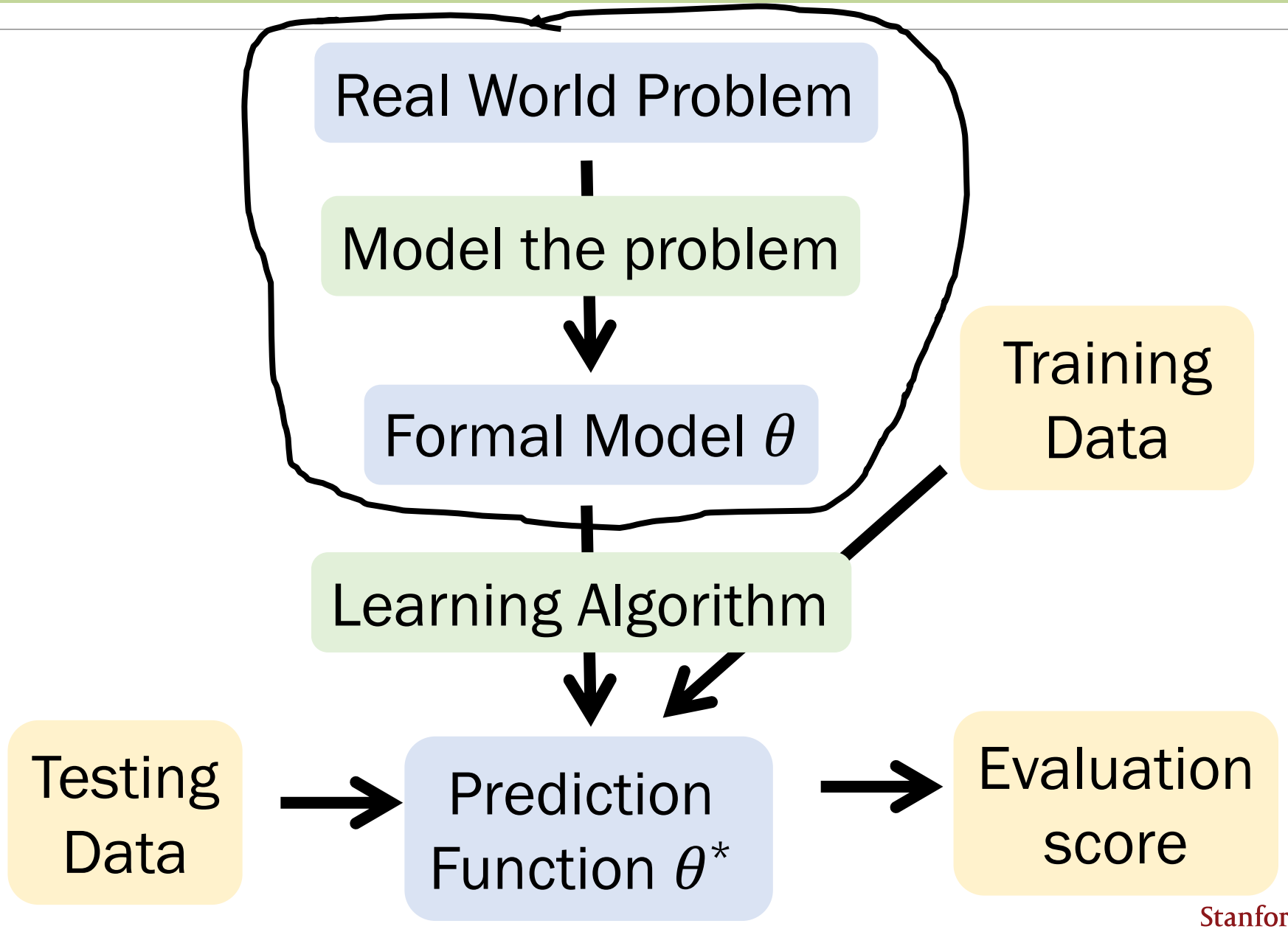


Parameters

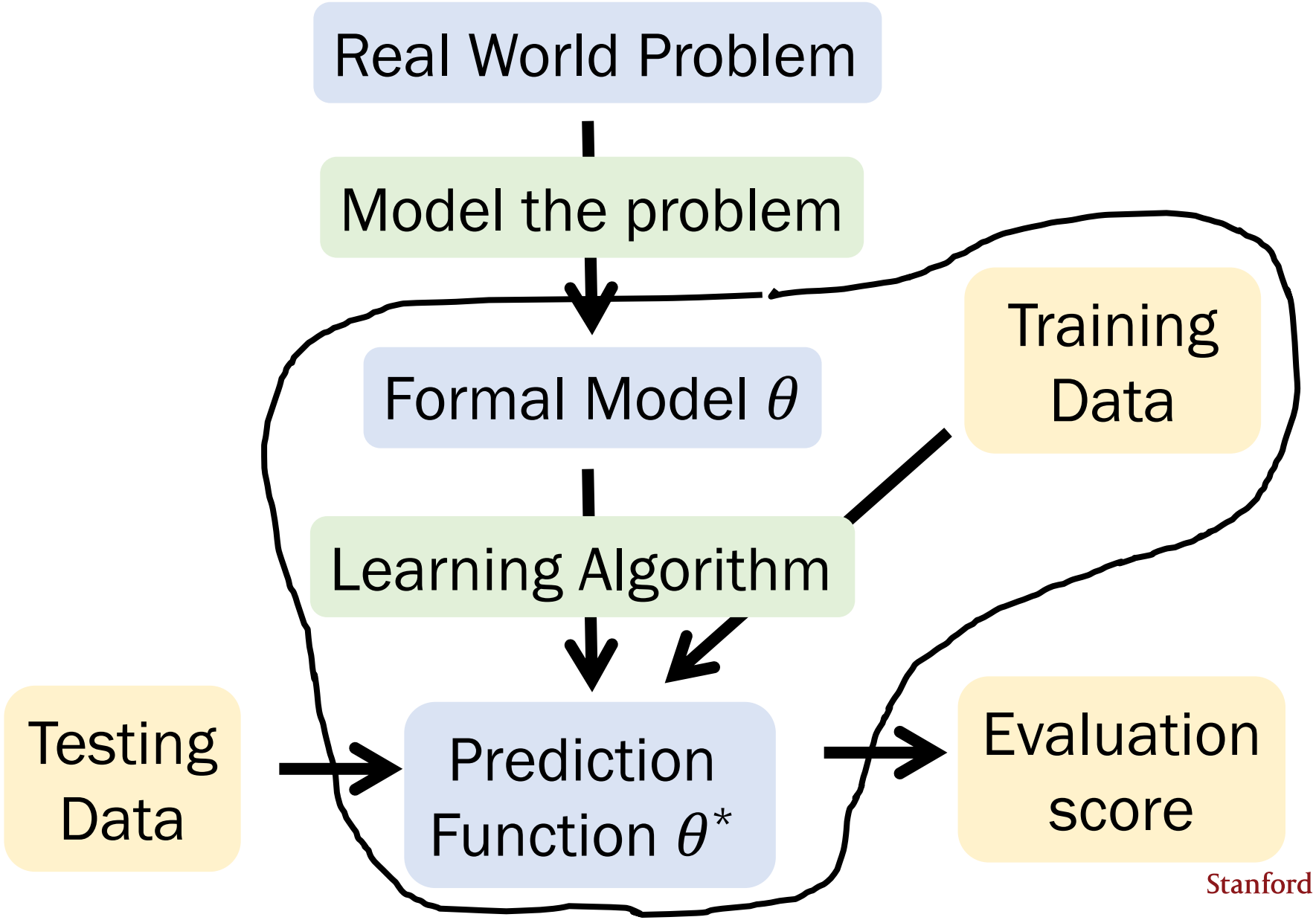
Why Do We Care?



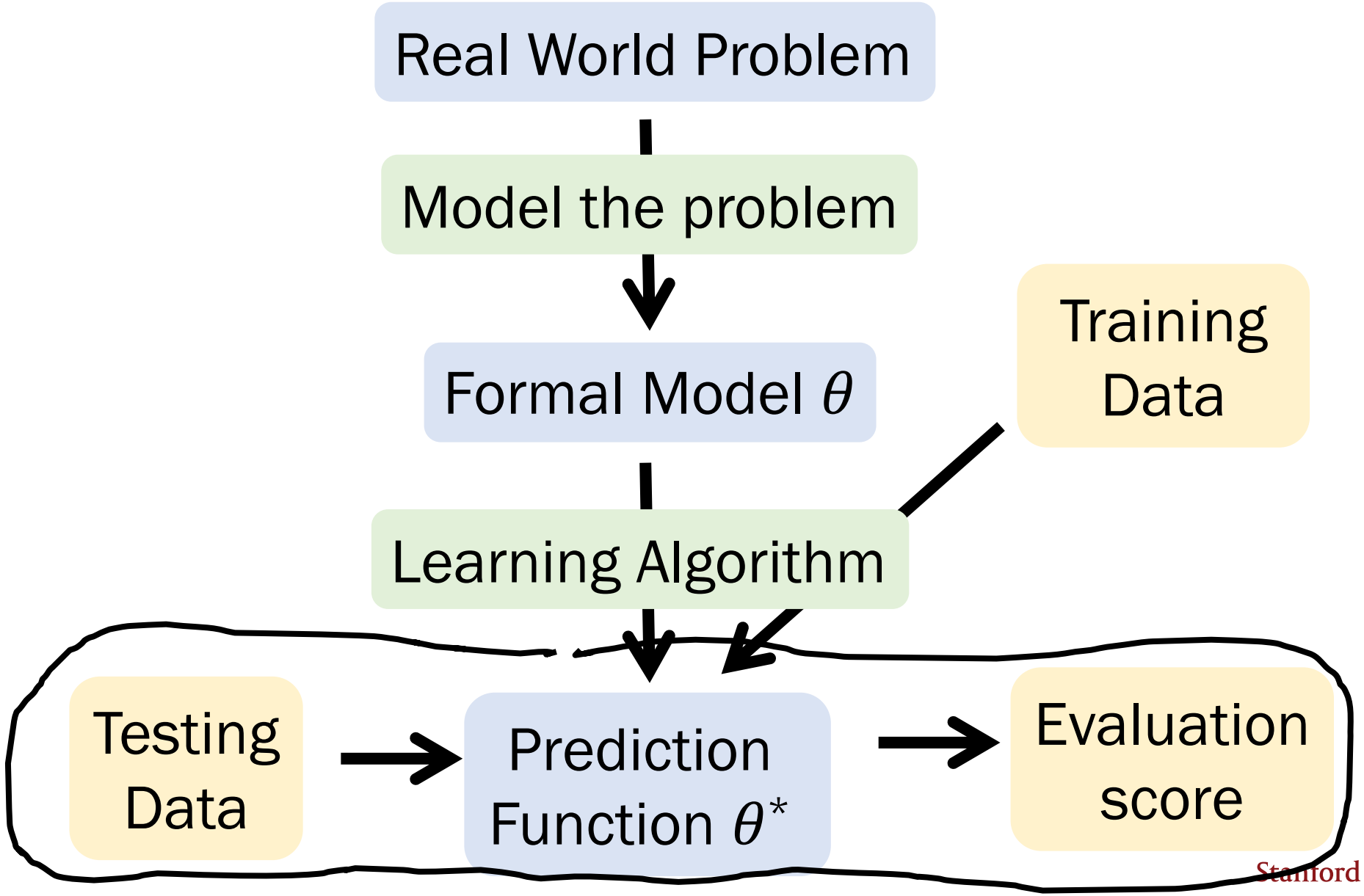
Modelling


Real World Problem
Model the problem
Formal Model θ
Training Data
Learning Algorithm
Testing Data
Prediction Function θ^*
Evaluation score

Parameter Estimation (aka Training)

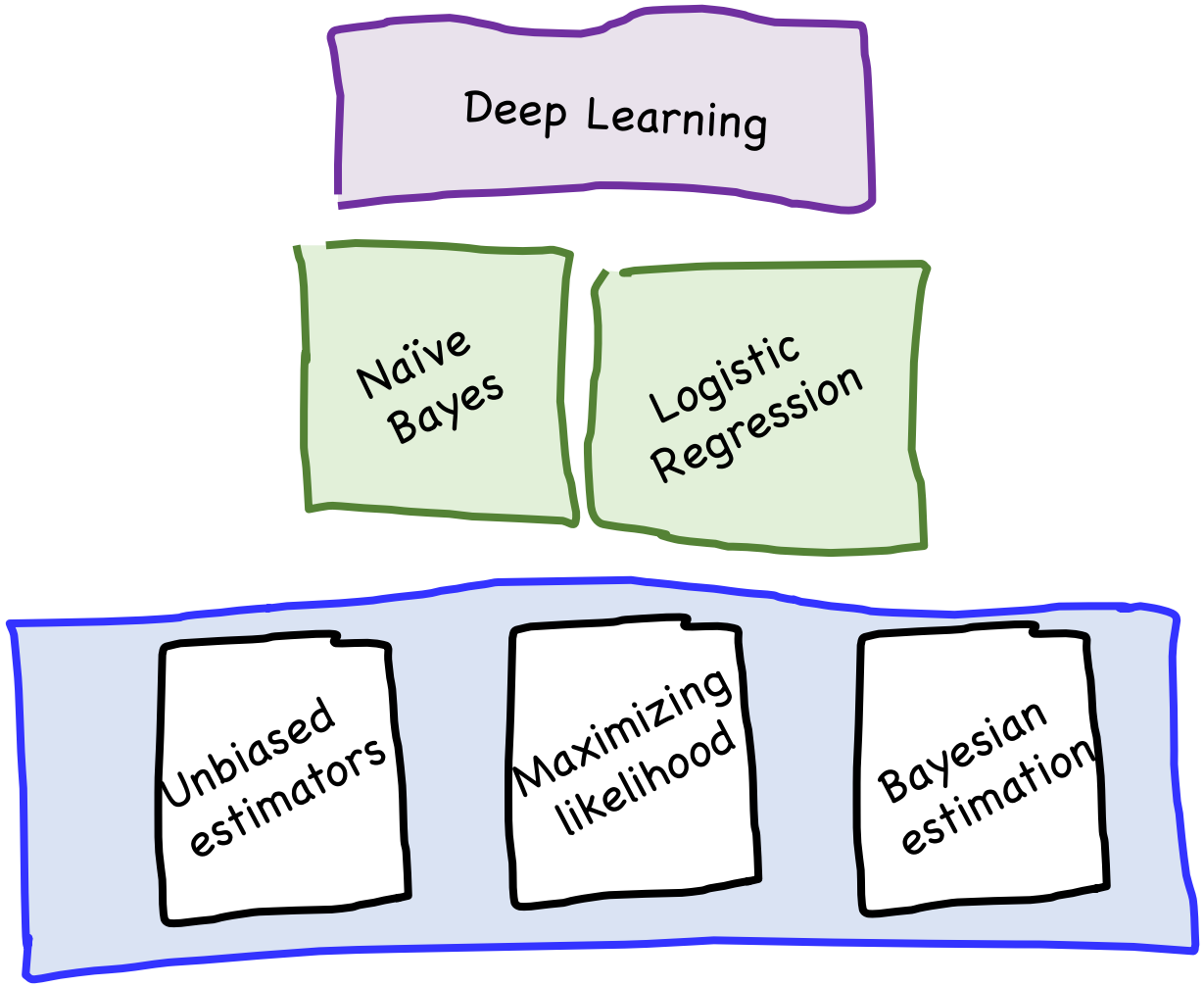


Testing

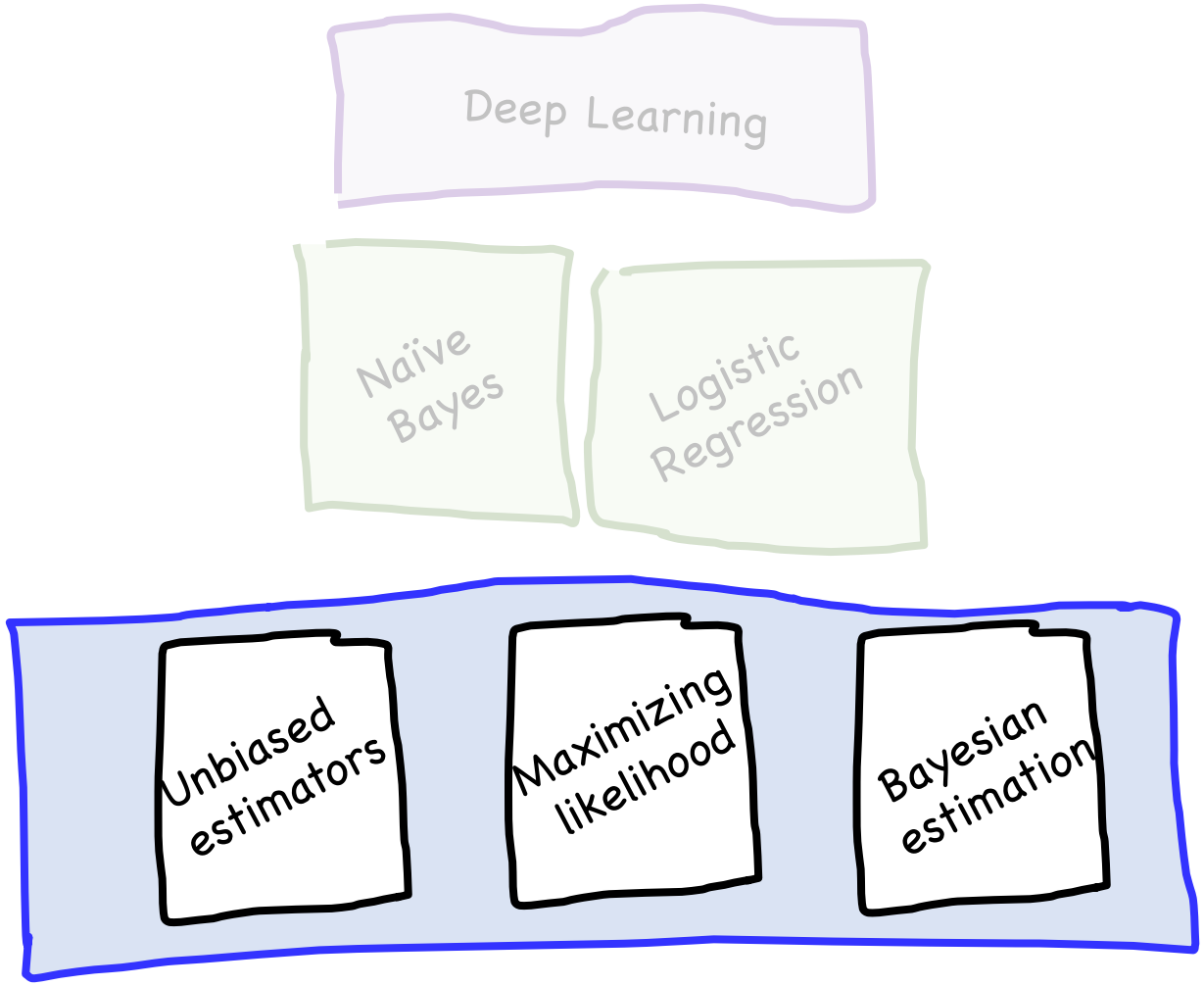

Real World Problem
Model the problem
Formal Model θ
Training Data
Learning Algorithm
Testing Data
Prediction Function θ^*
Evaluation score

Basis for learning from data

Our Path


Deep Learning
Naïve Bayes
Logistic Regression
Unbiased estimators
Maximizing likelihood
Bayesian estimation

Parameter Estimation



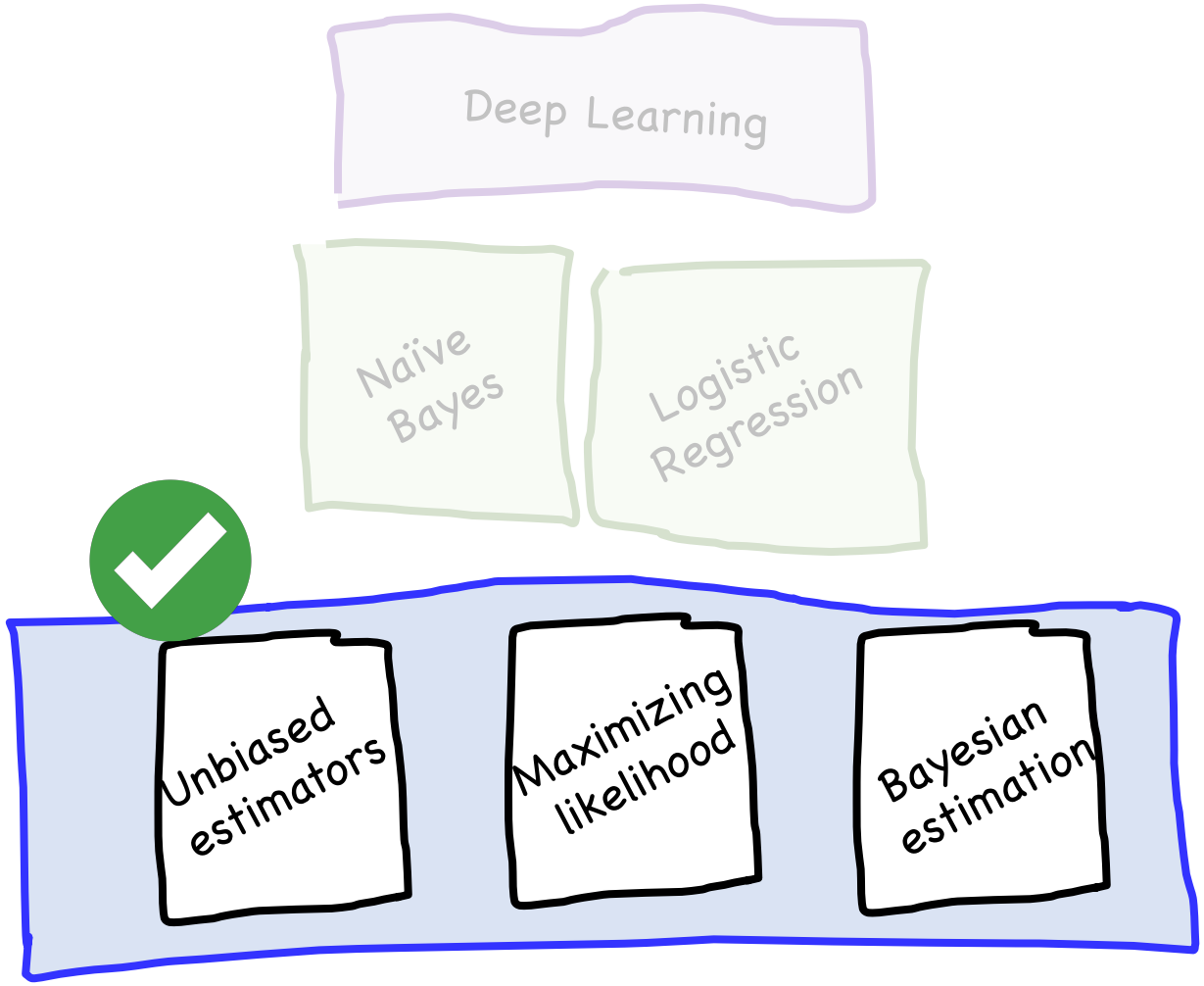
We've already seen some estimations

$X_1, X_2, \ldots, X_n$ are n i.i.d. random variables,
where X_i drawn from distribution F with $E[X_i] = \mu, \text{Var}(X_i) = \sigma^2$.

Sample mean: $$\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$$ unbiased estimate of μ

Sample variance: $$S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2$$ unbiased estimate of σ^2

Parameter Estimation



Limited tool: how could we use that for fitting a “Mixture of Gaussians”?

Great idea in Machine Learning

Demo: Likelihood of Data

Data = [6.3 , 5.5 , 5.4, 7.1, 4.6, 6.7, 5.3 , 4.8, 5.6, 3.4, 5.4, 3.4, 4.8, 7.9, 4.6, 7.0, 2.9, 6.4, 6.0 , 4.3]

Estimate the Parameters

Parameter μ: 5.6 Parameter σ: 1.4

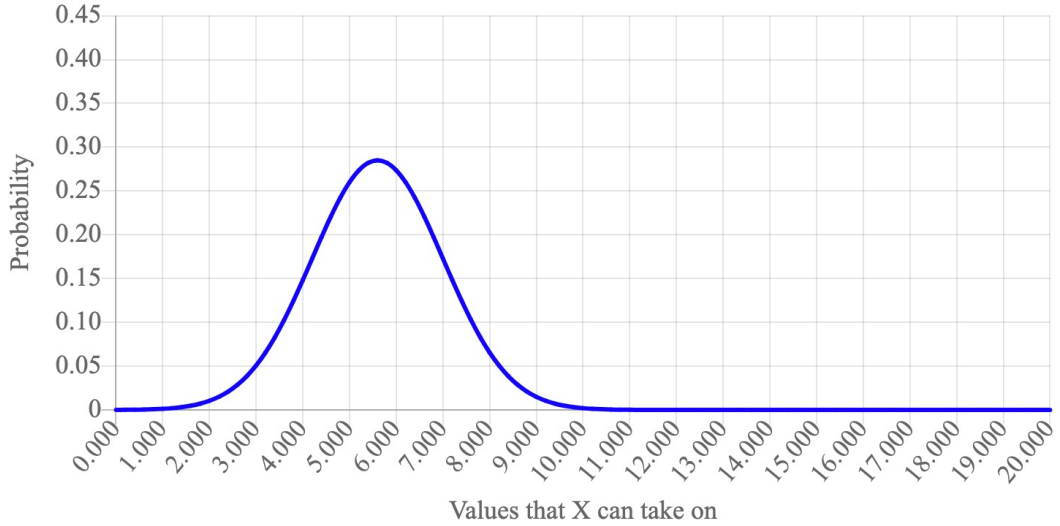
Likelihood

Likelihood: 1.9542923784106326e-15

Log Likelihood: -301.9

Best Seen: -301.9

PDF Graph




L-R
30
40
50
60

Maximum Likelihood Algorithm

1. Decide on a model for the distribution of your samples.
   Define the PMF / PDF for your sample.

2. Write out the log likelihood function.

3. State that the optimal parameters are
   the argmax of the log likelihood
   function.

4. Use an optimization algorithm to calculate argmax

The Likelihood Function

n I.I.D. data points $X_1, X_2, \ldots, X_n$

$$L(\theta) = \prod_{i=1}^{n} f(X_i \mid \theta)$$

This is just a product since X_i are I.I.D.

We explicitly specify parameter θ of distribution

Likelihood (of data given parameters):

$$L(\theta) = \prod_{i=1}^{n} f(X_i \mid \theta)$$

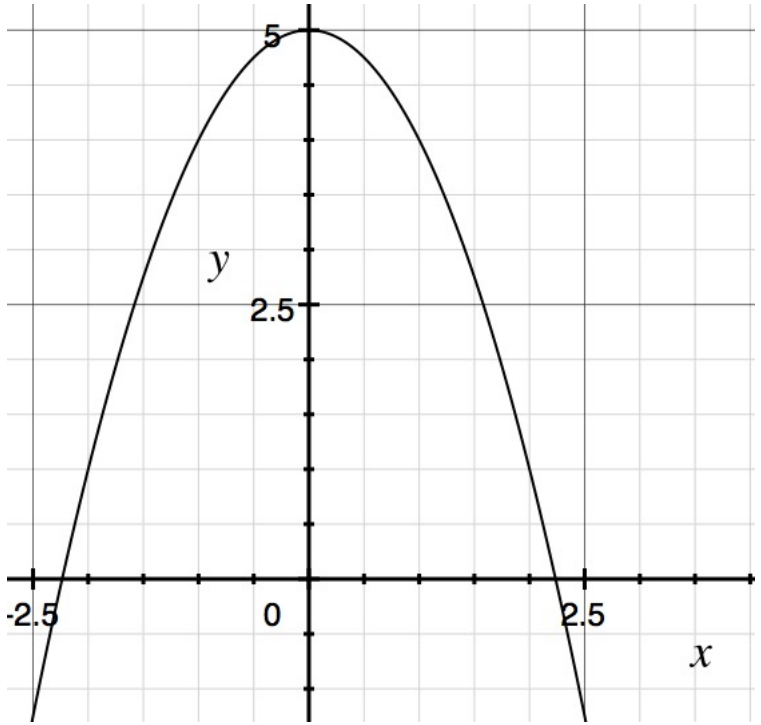
Either the
PDF (continuous) or
PMF (discrete), or
joint if multiple variables per datapoint

Argmax

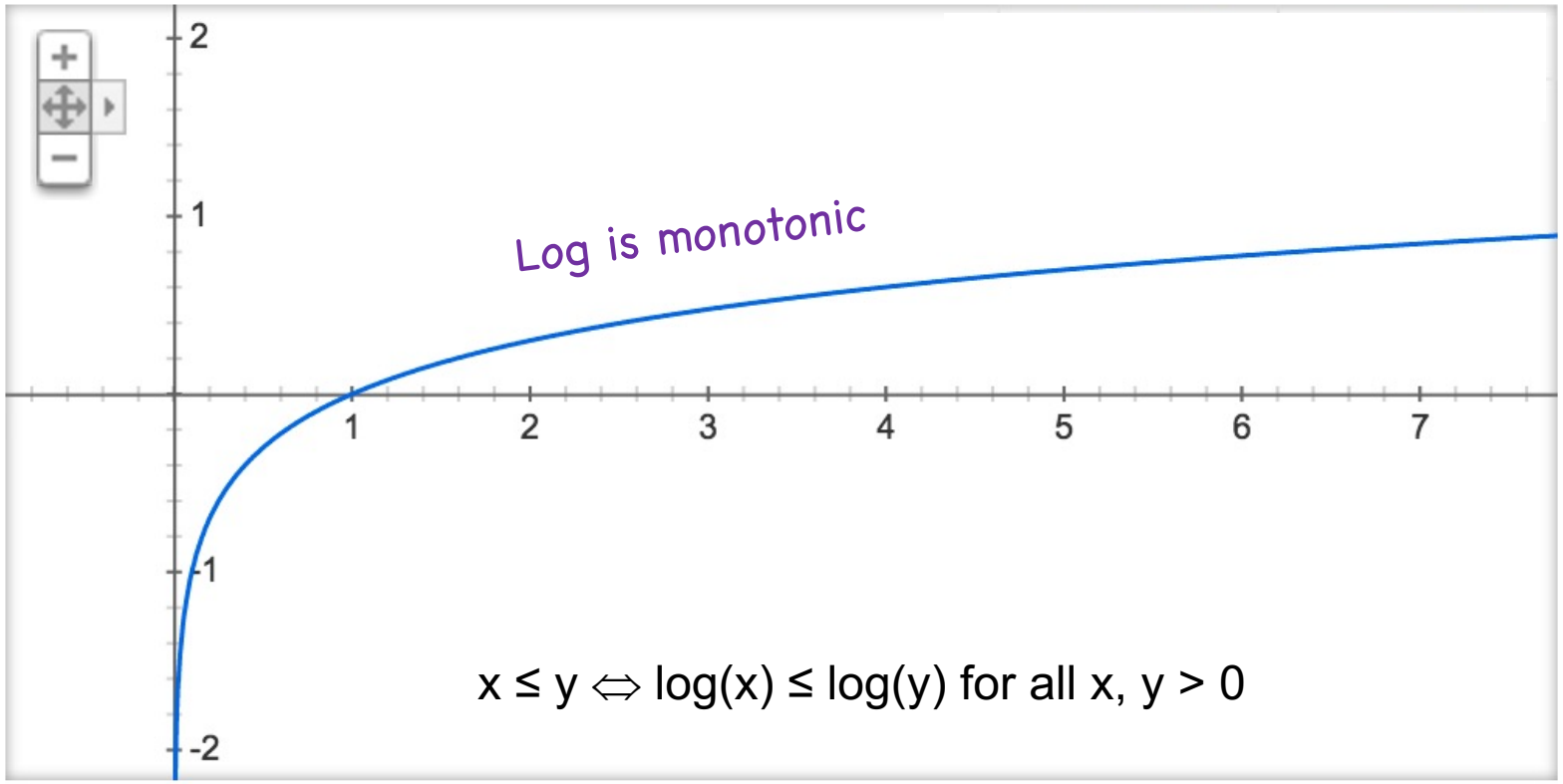
$$f(x) = -x^2 + 5$$

$$\max_x \; -x^2 + 5 = 5$$

$$\operatorname*{argmax}_x \; -x^2 + 5 = 0$$



Argmax of Log



Claim: $\underset{x}{\operatorname{argmax}}\ f(x) = \underset{x}{\operatorname{argmax}}\ \log f(x)$

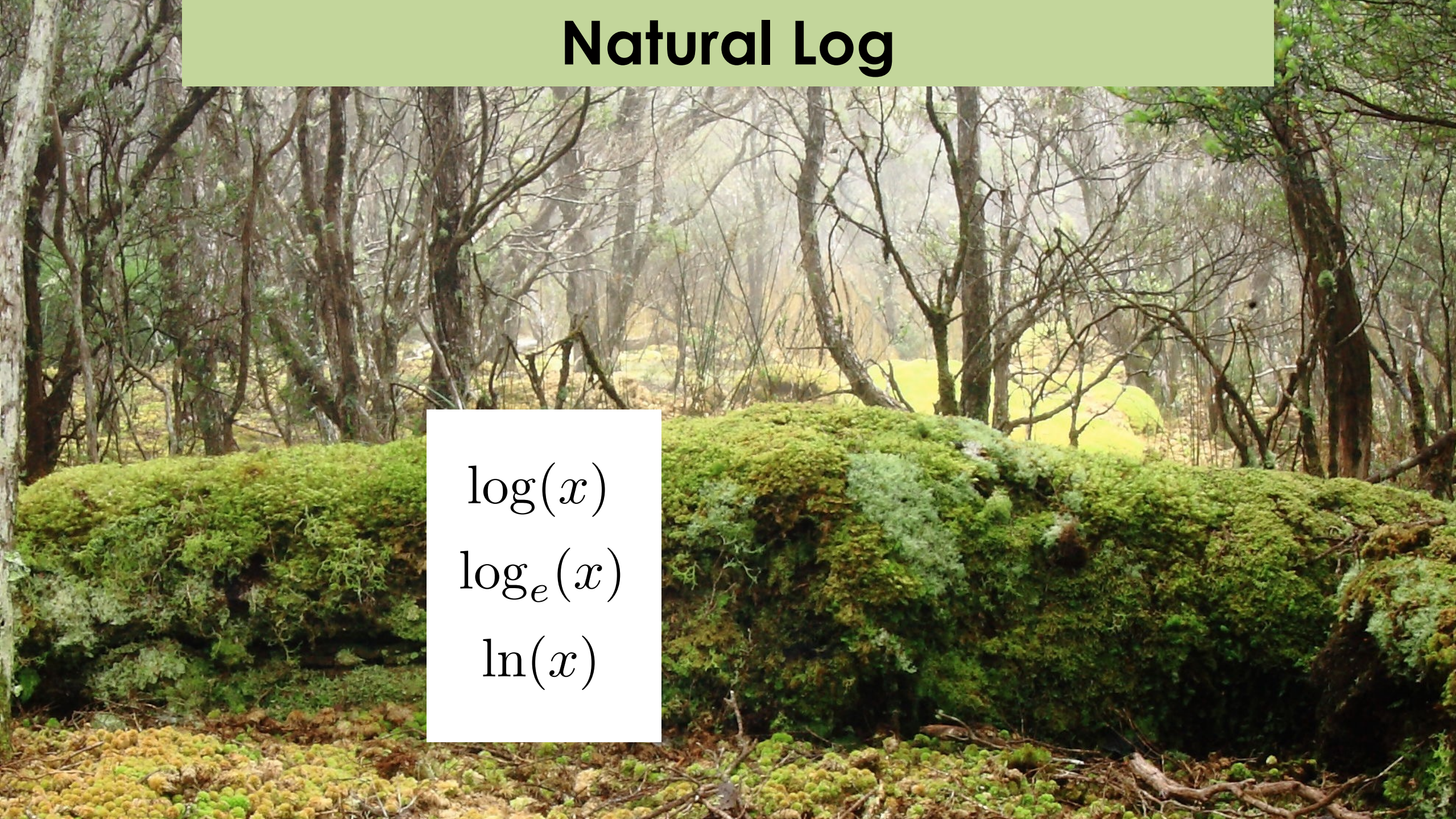
Argmax of Log

$$\underset{x}{\operatorname{argmax}}\ f(x) = \underset{x}{\operatorname{argmax}}\ \log f(x)$$

Log I Love You

$$\log(ab) = \log(a) + \log(b)$$

Natural Log



$$\log(x)$$

$$\log_e(x)$$

$$\ln(x)$$

Maximum Likelihood Algorithm

1. Decide on a model for the distribution of your samples. Define the PMF / PDF for your sample.

2. Write out the log likelihood function.

3. State that the optimal parameters are the argmax of the log likelihood function.

4. Use an optimization algorithm to calculate argmax

Story so far: We can chose parameters by finding the argmax of the log likelihood of our data

Maximum Likelihood

$$L(\theta) = \prod_{i=1}^{n} f(X_i \mid \theta)$$

$$LL(\theta) = \sum_{i=1}^{n} \log f(X_i|\theta)$$

$$\hat{\theta} = \underset{\theta}{\operatorname{argmax}}\ LL(\theta)$$


arg max

But how do we compute argmax?

Option #1: Straight optimization

Finding the argmax with calculus

$$\hat{x} = \arg\max_x f(x)$$

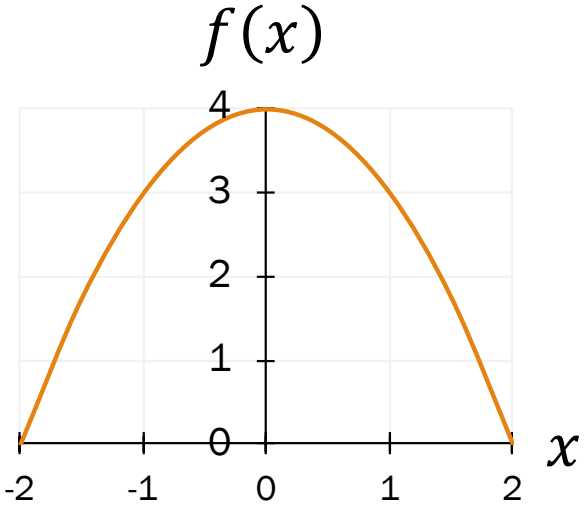
Let $f(x) = -x^2 + 4$, where $-2 < x < 2$.

Differentiate w.r.t. argmax's argument

$$\frac{d}{dx} f(x) = \frac{d}{dx}(x^2 + 4) = 2x$$

Set to 0 and solve

$$2x = 0 \quad \Rightarrow \quad \hat{x} = 0$$



Make sure $\hat{x}$ is a maximum

- Check $f(\hat{x} \pm \epsilon) < f(\hat{x})$
- Generally ignored in expository derivations
- We'll ignore it here too (and won't require it in class)
- arg min is defined similarly, relevant for gradient descent

Computing the MLE

$$\theta_{MLE} = \arg\max_{\theta} LL(\theta)$$

General approach for finding θ_{MLE} , the MLE of θ:

1. Determine formula for $LL(\theta)$
2. Differentiate $LL(\theta)$ w.r.t. (each) θ
3. Solve

$$LL(\theta) = \sum_{i=1}^{n} \log f(X_i|\theta)$$

$$\frac{\partial LL(\theta)}{\partial \theta}$$

To maximize: $\frac{\partial LL(\theta)}{\partial \theta} = 0$

(algebra or computer)

4. Technically, make sure derived $\hat{\theta}_{MLE}$ is a maximum
   - Check $LL(\theta_{MLE} \pm \epsilon) < LL(\theta_{MLE})$
   - Often ignored in expository derivations
   - We'll ignore it here too (and won't require it in class)

$LL(\theta)$ is often easier to differentiate than $L(\theta)$.

Maximizing Likelihood with Bernoulli

Consider I.I.D. random variables $X_1, X_2, ..., X_n$

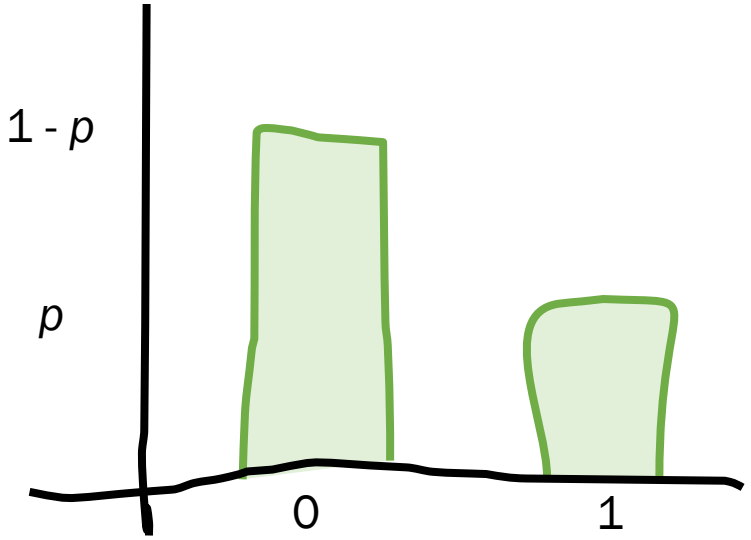
- $X_i \sim \text{Ber}(p)$
- Probability mass function, $f(X_i \mid p)$:

Maximizing Likelihood with Bernoulli

Consider I.I.D. random variables X_1, X_2, ..., X_n

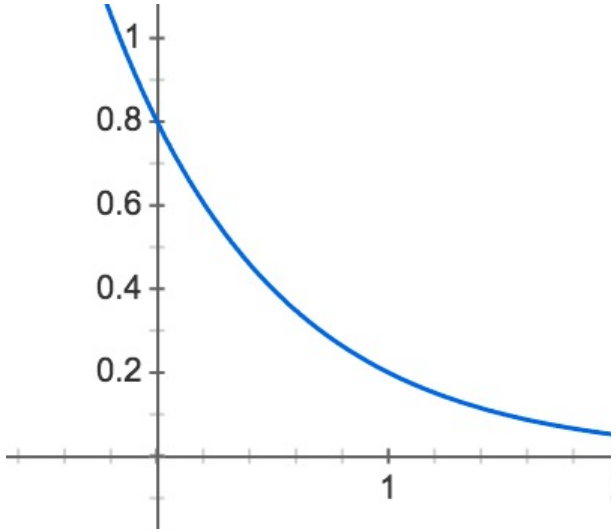
- X_i ~ Ber(p)
- Probability mass function, $f(X_i \mid p)$:

PMF of Bernoulli



$$f(X_i \mid p) = p^{x_i}(1-p)^{1-x_i}$$

PMF of Bernoulli (p = 0.2)



$$f(x) = 0.2^x(1-0.2)^{1-x}$$

Bernoulli PMF

$$X \sim \text{Ber}(p)$$

$$f(X = x|p) = p^x(1-p)^{1-x}$$

Maximum Likelihood with Bernoulli

Consider a sample of n i.i.d. RVs $X_1, X_2, \ldots, X_n$.

What is $\theta_{MLE} = p_{MLE}$?

- Let $X_i \sim \text{Ber}(p)$.

1. Determine formula for $LL(\theta)$

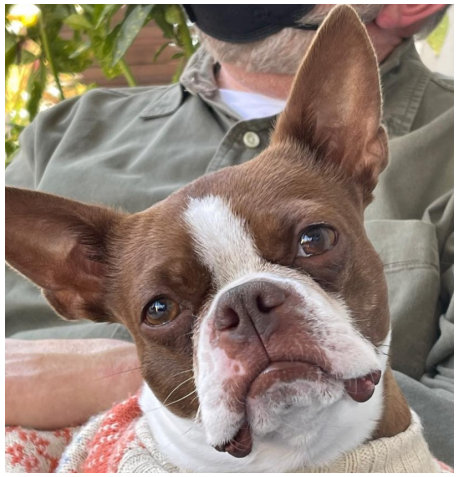
$$LL(\theta) = \sum_{i=1}^{n} \log f(X_i|p)$$

?

$$f(X_i|p) = \begin{cases} p & \text{if } X_i = 1 \\ 1-p & \text{if } X_i = 0 \end{cases}$$

2. Differentiate $LL(\theta)$ w.r.t. (each) θ, set to 0

3. Solve resulting equations



Maximum Likelihood with Bernoulli

Consider a sample of n i.i.d. RVs $X_1, X_2, \ldots, X_n$.

What is $\theta_{MLE} = p_{MLE}$?

- Let $X_i \sim \text{Ber}(p)$.
- $f(X_i|p) = p^{X_i}(1-p)^{1-X_i}$

1. Determine formula for $LL(\theta)$

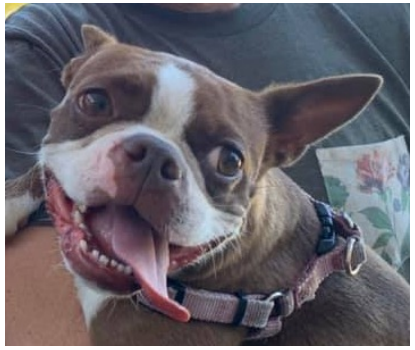
$$LL(\theta) = \sum_{i=1}^{n} \log f(X_i|p)$$

$$f(X_i|p) = \begin{cases} p & \text{if } X_i = 1 \\ 1-p & \text{if } X_i = 0 \end{cases}$$

$$f(X_i|p) = p^{X_i}(1-p)^{1-X_i} \text{ where } X_i \in \{0,1\}$$

2. Differentiate $LL(\theta)$ w.r.t. (each) θ, set to 0

3. Solve resulting equations



✅

- Is differentiable with respect to p
- Valid PMF over discrete domain

Maximum Likelihood with Bernoulli

Consider a sample of n i.i.d. RVs $X_1, X_2, \ldots, X_n$.

What is $\theta_{MLE} = p_{MLE}$?

- Let $X_i \sim \text{Ber}(p)$.
- $f(X_i|p) = p^{X_i}(1-p)^{1-X_i}$

1. Determine formula for $LL(\theta)$

$$LL(\theta) = \sum_{i=1}^{n} \log f(X_i|p) = \sum_{i=1}^{n} \log(p^{X_i}(1-p)^{1-X_i})$$

2. Differentiate $LL(\theta)$ w.r.t. (each) θ, set to 0

$$= \sum_{i=1}^{n} [X_i \log p + (1-X_i)\log(1-p)]$$

3. Solve resulting equations

$$= Y(\log p) + (n-Y)\log(1-p), \text{ where } Y = \sum_{i=1}^{n} X_i$$

Maximum Likelihood with Bernoulli

Consider a sample of n i.i.d. RVs $X_1, X_2, \ldots, X_n$.

What is $\theta_{MLE} = p_{MLE}$?

- Let $X_i \sim \text{Ber}(p)$.
- $f(X_i|p) = p^{X_i}(1-p)^{1-X_i}$

1. Determine formula for $LL(\theta)$

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

$$= Y(\log p) + (n - Y) \log(1-p), \text{ where } Y = \sum_{i=1}^{n} X_i$$

2. Differentiate $LL(\theta)$ w.r.t. (each) θ, set to 0

$$\frac{\partial LL(\theta)}{\partial p} = Y\frac{1}{p} + (n-Y)\frac{-1}{1-p} = 0$$

3. Solve resulting equations

Maximum Likelihood with Bernoulli

Consider a sample of n i.i.d. RVs $X_1, X_2, \ldots, X_n$.

What is $\theta_{MLE} = p_{MLE}$?

- Let $X_i \sim \text{Ber}(p)$.
- $f(X_i|p) = p^{X_i}(1-p)^{1-X_i}$

1. Determine formula for $LL(\theta)$

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

$$= Y(\log p) + (n - Y) \log(1-p), \text{ where } Y = \sum_{i=1}^{n} X_i$$

2. Differentiate $LL(\theta)$ w.r.t. (each) θ, set to 0

$$\frac{\partial LL(\theta)}{\partial p} = Y\frac{1}{p} + (n-Y)\frac{-1}{1-p} = 0$$

3. Solve resulting equations

Maximum Likelihood with Bernoulli

Consider a sample of n i.i.d. RVs $X_1, X_2, \ldots, X_n$.

What is $\theta_{MLE} = p_{MLE}$?

- Let $X_i \sim \text{Ber}(p)$.
- $f(X_i|p) = p^{X_i}(1-p)^{1-X_i}$

1. Determine formula for $LL(\theta)$

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

$$= Y(\log p) + (n - Y) \log(1-p), \text{ where } Y = \sum_{i=1}^{n} X_i$$

2. Differentiate $LL(\theta)$ w.r.t. (each) θ, set to 0

$$\frac{\partial LL(\theta)}{\partial p} = Y\frac{1}{p} + (n-Y)\frac{-1}{1-p} = 0$$

3. Solve resulting equations

$$p_{MLE} = \frac{1}{n}Y = \frac{1}{n}\sum_{i=1}^{n} X_i$$

MLE of the Bernoulli parameter, p_{MLE}, is the unbiased estimate of the mean, $\bar{X}$ (sample mean)

Isn’t that the same as
unbiased estimator?

Yes. For Bernoulli.

MLE of Bernoulli is the sample mean

Quick check

- You draw n i.i.d. random variables $X_1, X_2, \ldots, X_n$ from the distribution F, yielding the following sample:

$$[0, 0, 1, 1, 1, 1, 1, 1, 1, 1] \qquad (n = 10)$$

- Suppose distribution $F = \text{Ber}(p)$ with unknown parameter p.

1. What is p_{MLE}, the MLE of the parameter p?

   A. 1.0
   B. 0.5
   C. 0.8
   D. 0.2
   E. None/other

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

Quick check

- You draw n i.i.d. random variables $X_1, X_2, \ldots, X_n$ from the distribution F, yielding the following sample:

$$[0, 0, 1, 1, 1, 1, 1, 1, 1, 1] \qquad (n = 10)$$

- Suppose distribution $F = \text{Ber}(p)$ with unknown parameter p.

1. What is p_{MLE}, the MLE of the parameter p?

A. 1.0
B. 0.5
C. 0.8
D. 0.2
E. None/other

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

Quick check

- You draw n i.i.d. random variables $X_1, X_2, \ldots, X_n$ from the distribution F, yielding the following sample:

$$[0, 0, 1, 1, 1, 1, 1, 1, 1, 1] \qquad (n = 10)$$

- Suppose distribution $F = \text{Ber}(p)$ with unknown parameter p.

1. What is p_{MLE}, the MLE of the parameter p? C. 0.8
2. What is the likelihood $L(\theta)$ of this particular sample?

$$f(X_i|p) = p^{X_i}(1-p)^{1-X_i} \text{ where } X_i \in \{0,1\}$$

$$L(\theta) = \prod_{i=1}^{n} f(X_i|p) \quad \text{where } \theta = p$$

$$= p^8(1-p)^2$$

Maximum Likelihood Algorithm

1. Decide on a model for the distribution of your samples. Define the PMF / PDF for your sample.

2. Write out the log likelihood function.

3. State that the optimal parameters are the argmax of the log likelihood function.

4. Use an optimization algorithm to calculate argmax

MLE for Poisson

$$X \sim \text{Poi}(\lambda)$$

Maximum Likelihood with Poisson

Consider a sample of n i.i.d. RVs $X_1, X_2, \ldots, X_n$.

What is $\theta_{MLE} = \lambda_{MLE}$?

- Let $X_i \sim \text{Poi}(\lambda)$.
- PMF: $f(X_i|\lambda) = \frac{e^{-\lambda}\lambda^{X_i}}{X_i!}$

1. Determine formula for $LL(\theta)$

$$LL(\theta) = \sum_{i=1}^{n} \log\left(\frac{e^{-\lambda}\lambda^{X_i}}{X_i!}\right) = \sum_{i=1}^{n} (-\lambda \log e + X_i \log \lambda - \log X_i!)$$

$$= -n\lambda + \log(\lambda) \sum_{i=1}^{n} X_i - \sum_{i=1}^{n} \log(X_i!)$$

(using natural log, $\ln e = 1$)

Maximum Likelihood with Poisson

Consider a sample of n i.i.d. RVs $X_1, X_2, \ldots, X_n$.

What is $\theta_{MLE} = \lambda_{MLE}$?

- Let $X_i \sim \text{Poi}(\lambda)$.
- PMF: $f(X_i|\lambda) = \frac{e^{-\lambda}\lambda^{X_i}}{X_i!}$

1. Determine formula for $LL(\theta)$

$$LL(\theta) = \sum_{i=1}^{n} \log\left(\frac{e^{-\lambda}\lambda^{X_i}}{X_i!}\right) = \sum_{i=1}^{n}(-\lambda \log e + X_i \log\lambda - \log X_i!)$$

$$= -n\lambda + \log(\lambda)\sum_{i=1}^{n} X_i - \sum_{i=1}^{n}\log(X_i!)$$

(using natural log, $\ln e = 1$)

2. Differentiate $LL(\theta)$ w.r.t. (each) θ, set to 0

$$\frac{\partial LL(\theta)}{\partial \lambda} = ?$$

A. $-n + \frac{1}{\lambda}\sum_{i=1}^{n} X_i + n\log\lambda - \sum_{i=1}^{n}\frac{1}{X_i!}\cdot\frac{\partial X_i!}{\partial X_i}$

B. $-n + \frac{1}{\lambda}\sum_{i=1}^{n} X_i$

C. None/other/ don't know

Maximum Likelihood with Poisson

Consider a sample of n i.i.d. RVs $X_1, X_2, \ldots, X_n$.

What is $\theta_{MLE} = \lambda_{MLE}$?

- Let $X_i \sim \text{Poi}(\lambda)$.
- PMF: $f(X_i|\lambda) = \frac{e^{-\lambda}\lambda^{X_i}}{X_i!}$

1. Determine formula for $LL(\theta)$

$$LL(\theta) = \sum_{i=1}^{n} \log\left(\frac{e^{-\lambda}\lambda^{X_i}}{X_i!}\right) = \sum_{i=1}^{n} (-\lambda \log e + X_i \log \lambda - \log X_i!)$$

$$= -n\lambda + \log(\lambda)\sum_{i=1}^{n} X_i - \sum_{i=1}^{n} \log(X_i!)$$

(using natural log, $\ln e = 1$)

2. Differentiate $LL(\theta)$ w.r.t. (each) θ, set to 0

$$\frac{\partial LL(\theta)}{\partial \lambda} = ?$$

A. $-n + \frac{1}{\lambda}\sum_{i=1}^{n} X_i + n \log \lambda - \sum_{i=1}^{n} \frac{1}{X_i!} \cdot \frac{\partial X_i!}{\partial X_i}$

B. $-n + \frac{1}{\lambda}\sum_{i=1}^{n} X_i$

C. None/other/ don't know

Maximum Likelihood with Poisson

Consider a sample of n i.i.d. RVs $X_1, X_2, \ldots, X_n$.

What is $\theta_{MLE} = \lambda_{MLE}$?

- Let $X_i \sim \text{Poi}(\lambda)$.
- PMF: $f(X_i|\lambda) = \frac{e^{-\lambda}\lambda^{X_i}}{X_i!}$

1. Determine formula for $LL(\theta)$

$$LL(\theta) = \sum_{i=1}^{n} \log\left(\frac{e^{-\lambda}\lambda^{X_i}}{X_i!}\right) = \sum_{i=1}^{n} (-\lambda \log e + X_i \log \lambda - \log X_i!)$$

$$= -n\lambda + \log(\lambda) \sum_{i=1}^{n} X_i - \sum_{i=1}^{n} \log(X_i!)$$

(using natural log, $\ln e = 1$)

2. Differentiate $LL(\theta)$ w.r.t. (each) θ, set to 0

$$\frac{\partial LL(\theta)}{\partial \lambda} = -n + \frac{1}{\lambda} \sum_{i=1}^{n} X_i = 0$$

3. Solve resulting equations

$$\lambda_{MLE} = \frac{1}{n} \sum_{i=1}^{n} X_i$$

Maximum Likelihood with Poisson

Consider a sample of n i.i.d. RVs $X_1, X_2, \ldots, X_n$.

What is $\theta_{MLE} = \lambda_{MLE}$?

- Let $X_i \sim \text{Poi}(\lambda)$.
- PMF: $f(X_i|\lambda) = \frac{e^{-\lambda}\lambda^{X_i}}{X_i!}$

1. Determine formula for $LL(\theta)$

$$LL(\theta) = \sum_{i=1}^{n} \log\left(\frac{e^{-\lambda}\lambda^{X_i}}{X_i!}\right) = \sum_{i=1}^{n} (-\lambda \log e + X_i \log \lambda - \log X_i!)$$

$$= -n\lambda + \log(\lambda)\sum_{i=1}^{n} X_i - \sum_{i=1}^{n} \log(X_i!)$$

(using natural log, $\ln e = 1$)

2. Differentiate $LL(\theta)$ w.r.t. (each) θ, set to 0

$$\frac{\partial LL(\theta)}{\partial \lambda} = -n + \frac{1}{\lambda}\sum_{i=1}^{n} X_i = 0$$

3. Solve resulting equations

$$\lambda_{MLE} = \frac{1}{n}\sum_{i=1}^{n} X_i$$

MLE of the Poisson parameter, λ_{MLE}, is the unbiased estimate of the mean, $\bar{X}$ (sample mean)

Its so general!

MLE for Gaussian

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

Maximum Likelihood with Normal

Consider a sample of n i.i.d. random variables $X_1, X_2, \ldots, X_n$.

- Let $X_i \sim \mathcal{N}(\mu, \sigma^2)$.

$$f(X_i|\mu, \sigma^2) = \frac{1}{\sqrt{2\pi}\sigma} e^{-(X_i-\mu)^2/(2\sigma^2)}$$

What is $\theta_{MLE} = (\mu_{MLE}, \sigma^2_{MLE})$?

1. Determine formula for $LL(\theta)$
2. Differentiate $LL(\theta)$ w.r.t. (each) θ, set to 0
3. Solve resulting equations

$$LL(\theta) = \sum_{i=1}^{n} \log\left(\frac{1}{\sqrt{2\pi}\sigma} e^{-(X_i-\mu)^2/(2\sigma^2)}\right) = \sum_{i=1}^{n} \left[-\log\left(\sqrt{2\pi}\sigma\right) - (X_i-\mu)^2/(2\sigma^2)\right]$$

(using natural log)

$$= -\sum_{i=1}^{n} \log\left(\sqrt{2\pi}\sigma\right) - \sum_{i=1}^{n} \left[(X_i-\mu)^2/(2\sigma^2)\right]$$

Maximum Likelihood with Normal

Consider a sample of n i.i.d. random variables $X_1, X_2, \ldots, X_n$.

- Let $X_i \sim \mathcal{N}(\mu, \sigma^2)$.

$$f(X_i|\mu,\sigma^2) = \frac{1}{\sqrt{2\pi}\sigma} e^{-(X_i-\mu)^2/(2\sigma^2)}$$

What is $\theta_{MLE} = (\mu_{MLE}, \sigma^2_{MLE})$?

1. Determine formula for $LL(\theta)$
2. Differentiate $LL(\theta)$ w.r.t. (each) θ, set to 0
3. Solve resulting equations

$$LL(\theta) = -\sum_{i=1}^{n} \log\left(\sqrt{2\pi}\sigma\right) - \sum_{i=1}^{n} [(X_i - \mu)^2/(2\sigma^2)]$$

with respect to μ

$$\frac{\partial LL(\theta)}{\partial \mu} = \sum_{i=1}^{n} [2(X_i - \mu)/(2\sigma^2)]$$

$$= \frac{1}{\sigma^2} \sum_{i=1}^{n} (X_i - \mu) = 0$$

Maximum Likelihood with Normal

Consider a sample of n i.i.d. random variables $X_1, X_2, \ldots, X_n$.

- Let $X_i \sim \mathcal{N}(\mu, \sigma^2)$.

$$f(X_i|\mu,\sigma^2) = \frac{1}{\sqrt{2\pi}\sigma} e^{-(X_i-\mu)^2/(2\sigma^2)}$$

What is $\theta_{MLE} = (\mu_{MLE}, \sigma^2_{MLE})$?

1. Determine formula for $LL(\theta)$
2. Differentiate $LL(\theta)$ w.r.t. (each) θ, set to 0
3. Solve resulting equations

$$LL(\theta) = -\sum_{i=1}^{n} \log\left(\sqrt{2\pi}\sigma\right) - \sum_{i=1}^{n} [(X_i-\mu)^2/(2\sigma^2)]$$

with respect to μ

$$\frac{\partial LL(\theta)}{\partial \mu} = \sum_{i=1}^{n} [2(X_i-\mu)/(2\sigma^2)]$$

$$= \frac{1}{\sigma^2}\sum_{i=1}^{n}(X_i-\mu) = 0$$

with respect to σ

$$\frac{\partial LL(\theta)}{\partial \sigma} = -\sum_{i=1}^{n}\frac{1}{\sigma} + \sum_{i=1}^{n} 2(X_i-\mu)^2/(2\sigma^3)$$

$$= -\frac{n}{\sigma} + \frac{1}{\sigma^3}\sum_{i=1}^{n}(X_i-\mu)^2 = 0$$

Maximum Likelihood with Normal

Consider a sample of n i.i.d. random variables $X_1, X_2, \ldots, X_n$.

- Let $X_i \sim \mathcal{N}(\mu, \sigma^2)$.

$$f(X_i|\mu, \sigma^2) = \frac{1}{\sqrt{2\pi}\sigma} e^{-(X_i-\mu)^2/(2\sigma^2)}$$

What is $\theta_{MLE} = (\mu_{MLE}, \sigma^2_{MLE})$?

3. Solve resulting equations

Two equations, two unknowns:

$$\frac{1}{\sigma^2}\sum_{i=1}^{n}(X_i - \mu) = 0 \qquad -\frac{n}{\sigma} + \frac{1}{\sigma^3}\sum_{i=1}^{n}(X_i - \mu)^2 = 0$$

First, solve for μ_{MLE}:

$$\frac{1}{\sigma^2}\sum_{i=1}^{n} X_i - \frac{1}{\sigma^2}\sum_{i=1}^{n}\mu = 0 \quad \Rightarrow \quad \sum_{i=1}^{n} X_i = n\mu \quad \Rightarrow \quad \mu_{MLE} = \frac{1}{n}\sum_{i=1}^{n} X_i$$

unbiased

Maximum Likelihood with Normal

Consider a sample of n i.i.d. random variables $X_1, X_2, \ldots, X_n$.

- Let $X_i \sim \mathcal{N}(\mu, \sigma^2)$.

$$f(X_i|\mu,\sigma^2) = \frac{1}{\sqrt{2\pi}\sigma} e^{-(X_i-\mu)^2/(2\sigma^2)}$$

What is $\theta_{MLE} = (\mu_{MLE}, \sigma^2_{MLE})$?

3. Solve resulting equations

Two equations, two unknowns:

$$\frac{1}{\sigma^2}\sum_{i=1}^{n}(X_i-\mu) = 0 \qquad -\frac{n}{\sigma} + \frac{1}{\sigma^3}\sum_{i=1}^{n}(X_i-\mu)^2 = 0$$

First, solve for μ_{MLE}:

$$\frac{1}{\sigma^2}\sum_{i=1}^{n} X_i - \frac{1}{\sigma^2}\sum_{i=1}^{n}\mu = 0 \quad \Rightarrow \sum_{i=1}^{n} X_i = n\mu \quad \Rightarrow \mu_{MLE} = \frac{1}{n}\sum_{i=1}^{n} X_i$$

unbiased

Next, solve for σ_{MLE}:

$$\frac{1}{\sigma^3}\sum_{i=1}^{n}(X_i-\mu)^2 = \frac{n}{\sigma} \quad \Rightarrow \sum_{i=1}^{n}(X_i-\mu)^2 = \sigma^2 n \quad \Rightarrow \sigma^2_{MLE} = \frac{1}{n}\sum_{i=1}^{n}(X_i-\mu_{MLE})^2$$

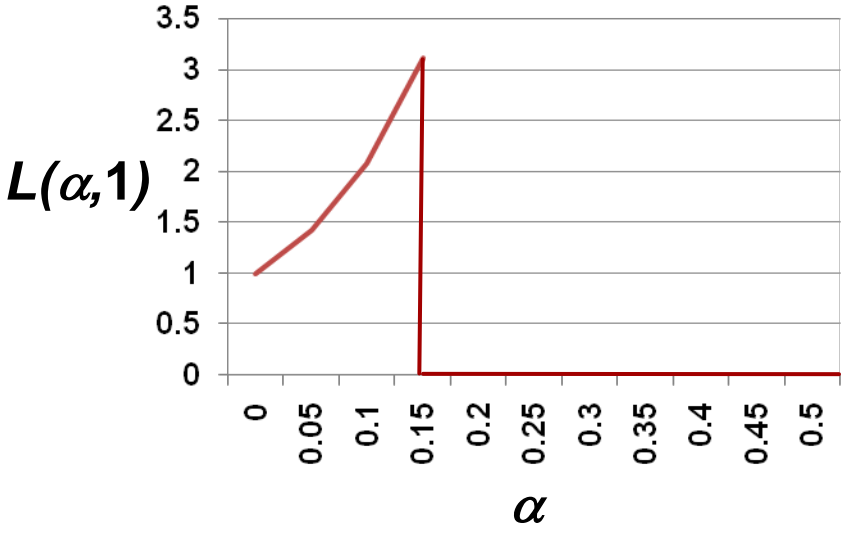
biased

Understanding MLE with Uniform

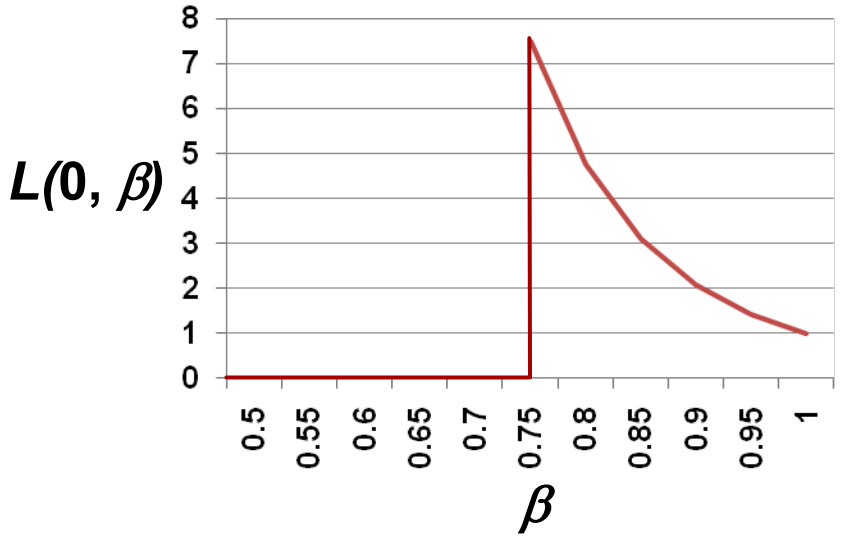
Consider I.I.D. random variables $X_1, X_2, \ldots, X_n$

- $X_i \sim \text{Uni}(0, 1)$
- Observe data:
  - 0.15, 0.20, 0.30, 0.40, 0.65, 0.70, 0.75

Likelihood: L(α,1)



Likelihood: L(0, β)



Small Samples = Problems

How do small samples affect MLE?

- In many cases, $\mu_{MLE} = \frac{1}{n}\sum_{i=1}^{n} X_i$ = sample mean
  - Unbiased. Not too shabby…
- As seen with Normal, $\sigma^2_{MLE} = \frac{1}{n}\sum_{i=1}^{n}(X_i - \mu_{MLE})^2$
  - Biased. Underestimates for small *n* (e.g., 0 for n = 1)
- As seen with Uniform, $\alpha_{MLE} \geq \alpha$ and $\beta_{MLE} \leq \beta$
  - Biased. Problematic for small *n* (e.g., $\alpha = \beta$ when n = 1)
- Small sample phenomena intuitively make sense:
  - Maximum likelihood $\Rightarrow$ best explain data we've seen
  - Does not attempt to generalize to unseen data

Properties of MLE

Maximum Likelihood Estimators are generally:

- **Consistent**: $\lim_{n\to\infty} P(|\hat{\theta} - \theta| < \varepsilon) = 1$ for $\varepsilon > 0$
- Potentially biased (though asymptotically less so)
- **Asymptotically optimal**
  - Has smallest variance of "good" estimators for large samples
- **Often used in practice** where sample size is large relative to parameter space
  - But be careful, there are some very large parameter spaces

Machine Learning:
Learn parameters (mostly with MLE) for probabilistic models.