



Parameter Estimation

Chris Piech

CS109, Stanford University

MACHINE LEARNING



GG



<Review>

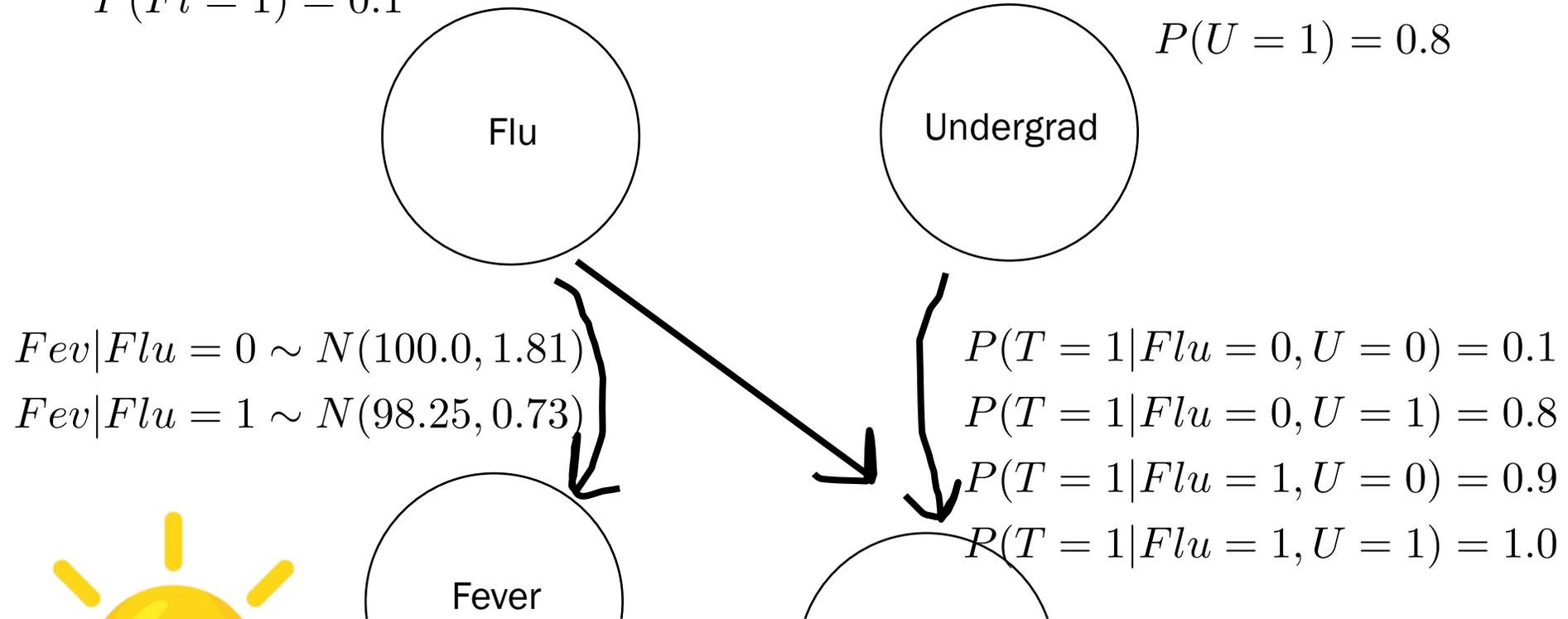
General “Inference”



Probabilistic Model

$$P(Fl = 1) = 0.1$$

$$P(U = 1) = 0.8$$



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

General Inference

Conditions that match your symptoms

UNDERSTANDING YOUR RESULTS 

Influenza (flu) adults



Moderate match



Pneumococcal Infections



Moderate match



H1N1 Flu Virus (Swine Flu)



Moderate match



Bacterial Pneumonia



Moderate match



Sepsis (blood infection)



Moderate match



Gender **Male**

Age **30**

[Edit](#)

My Symptoms

[Edit](#)

fever 103f to 104f, dizziness,

throat irritation, migraine headache



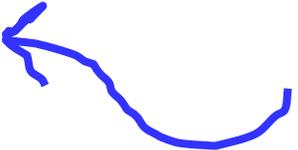
Start Over

Alg #1: Rejection Sampling

```
3 N_SAMPLES = 100000
4
5 # Program: Joint Sa
6 # -----
7 # we can answer any
8 # with multivariate
9 # where conditioned
10 def main():
11     obs = getObserv
12     print 'Observat
13
14     samples = sampl
15     prob = probFluG
16     print 'Pr(Flu)
```

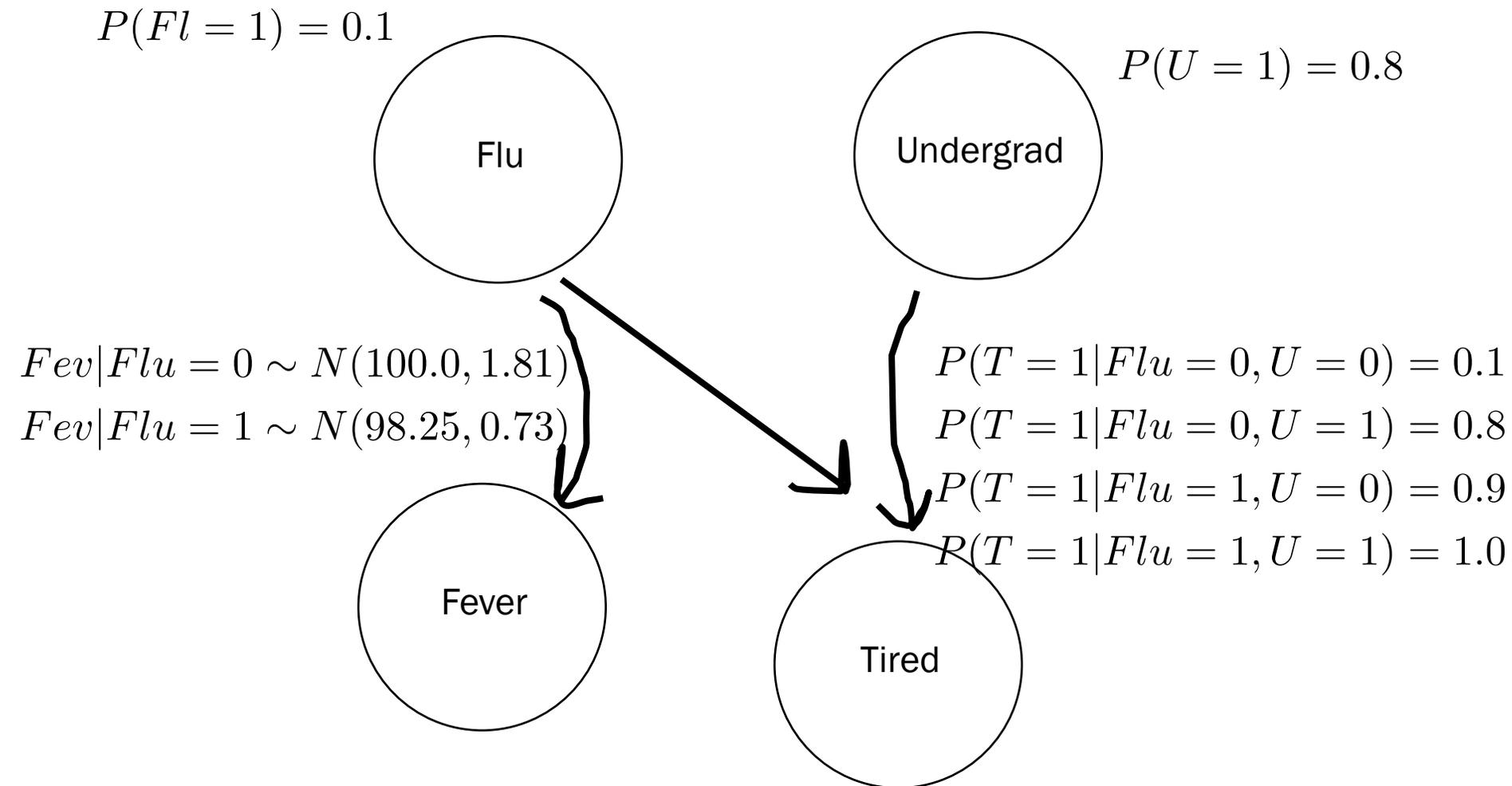
```
webMd -- -bash -- 38x22
[0, 0, 0, 0]
[0, 1, 0, 1]
[1, 0, 1, 0]
[1, 1, 1, 1]
[0, 1, 0, 1]
[0, 1, 0, 0]
[0, 0, 0, 0]
[0, 1, 1, 1]
[0, 1, 0, 0]
[0, 1, 0, 1]
[0, 1, 0, 0]
[0, 1, 0, 1]
[0, 1, 0, 1]
[0, 0, 0, 0]
[0, 0, 0, 0]
[1, 1, 1, 1]
[0, 0, 0, 0]
[0, 0, 0, 0]
[1, 1, 1, 1]
[0, 1, 0, 0]
Observation = [None, None, None, 1]
Pr(Flu | Obs) = 0.140635888502
>
```

Each one of these is one posterior sample:



[Flu, Ugrad, Fever, Tired]

Where Do The Numbers Come From?



End Review

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?

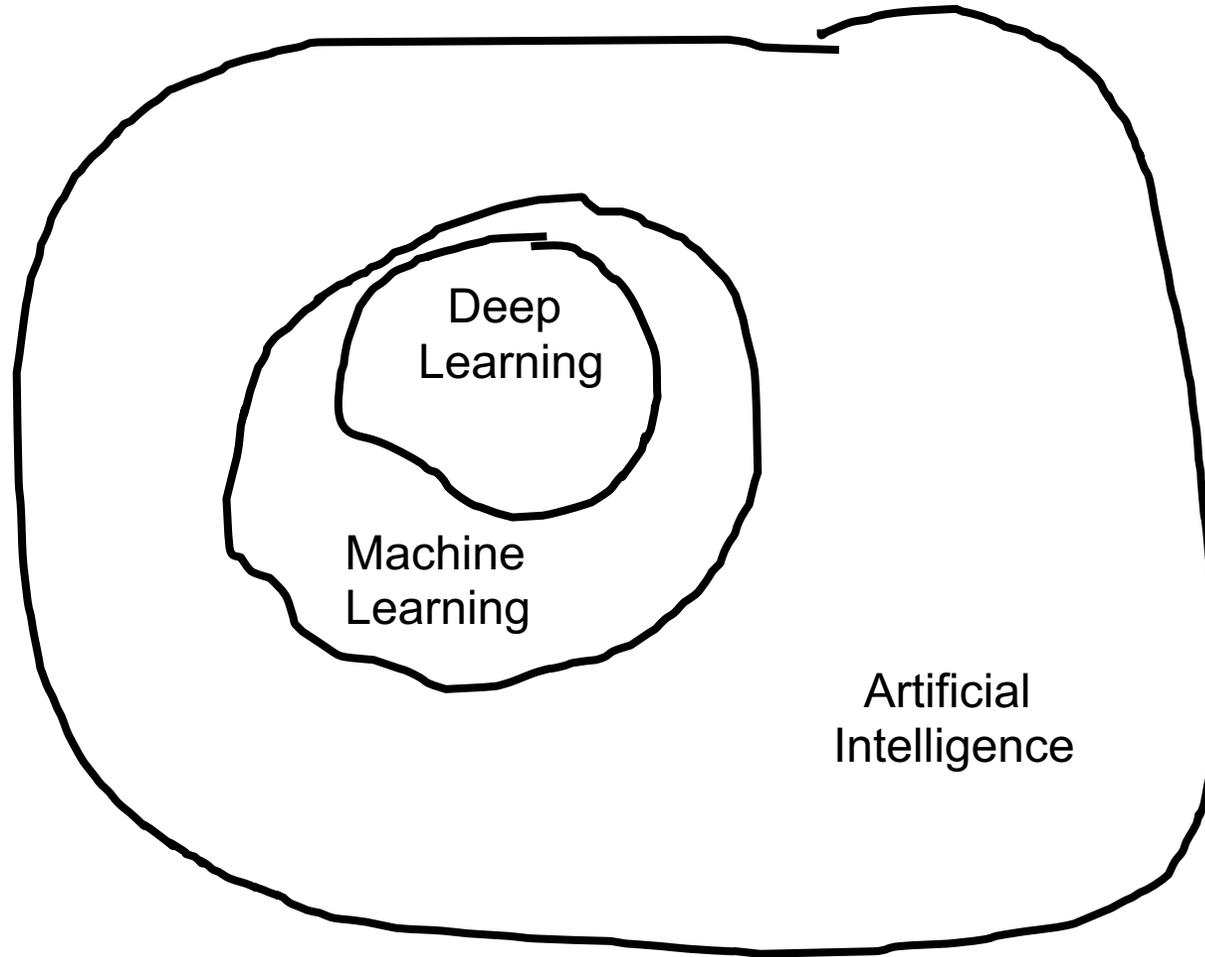
~~Oh can we also learn the *structure* of the model too?~~

I wish. Another day 😊

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

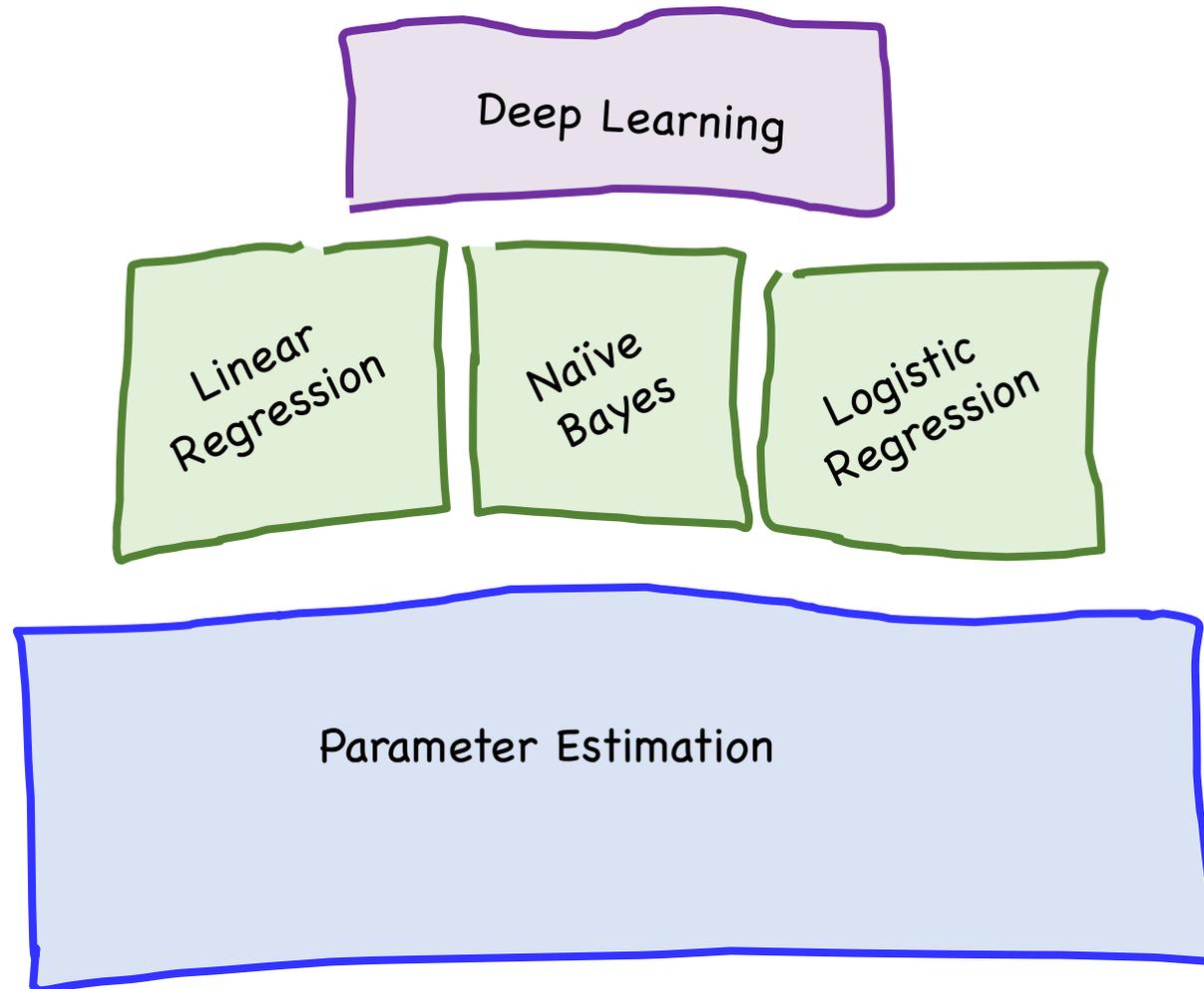
Machine Learning

AI and Machine Learning

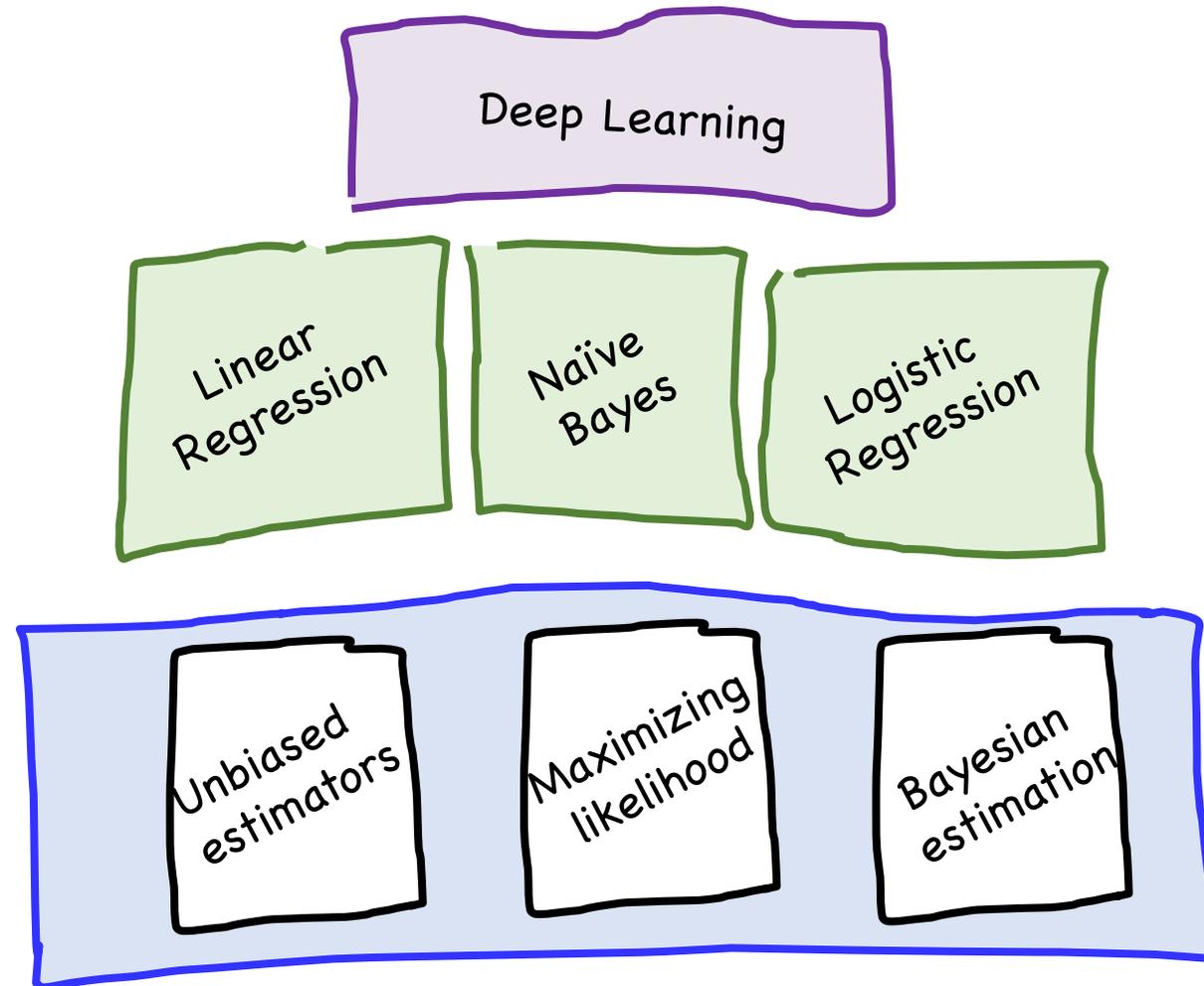


ML: Rooted in probability theory

Our Path



Our Path



Jump Straight to Deep Learning?

Tensor Flow



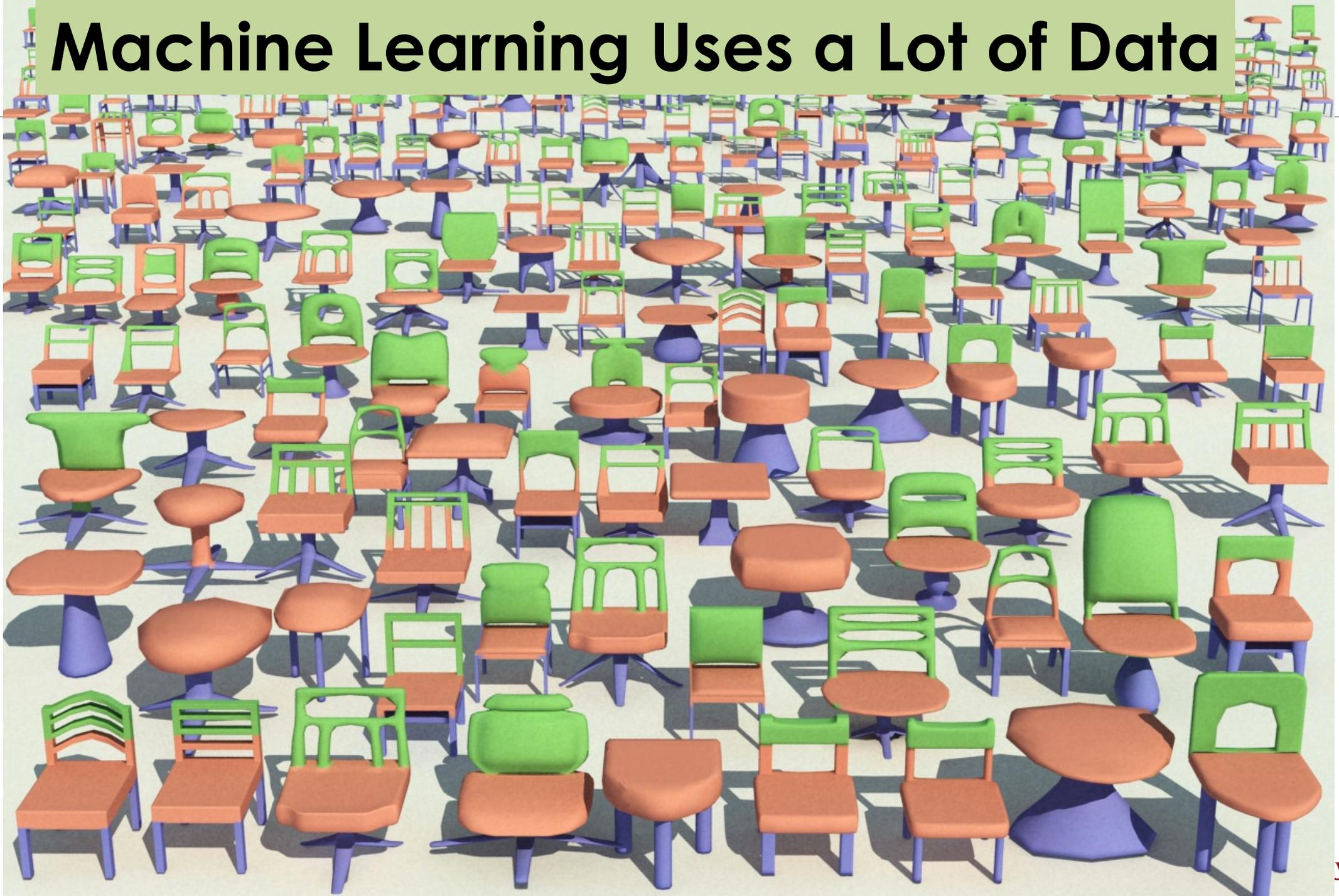
Jump Straight to Deep Learning?



Understand the theory to help you debug

But another reason...

Machine Learning Uses a Lot of Data

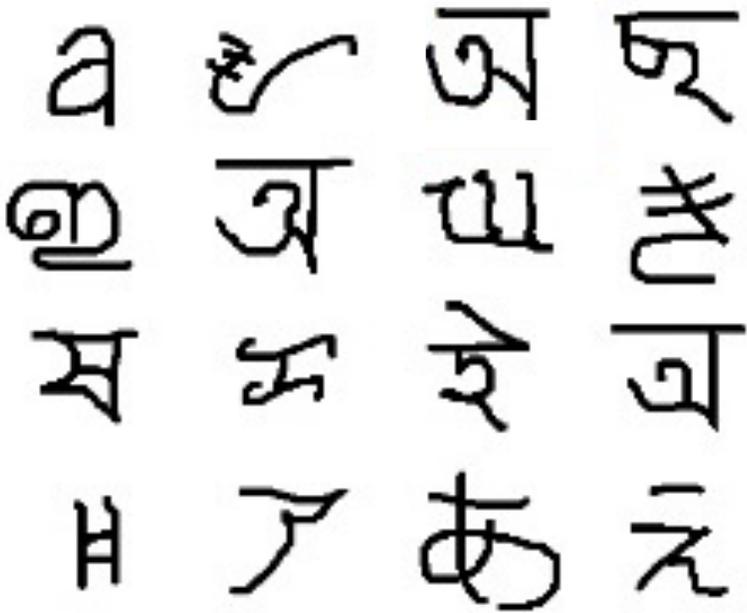


One Shot Learning

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**

The image features the iconic Walt Disney Pictures logo centered over a scene of a castle at night. The castle, with its multiple spires and towers, is brightly lit from within, casting a warm glow. The sky is a deep, dark blue, filled with numerous small white stars and streaks of light, suggesting a magical or dreamlike atmosphere. The castle is situated on a small island or peninsula, with a body of water in the foreground that reflects the lights from the castle and the stars in the sky. The overall composition is symmetrical and visually striking, capturing the essence of Disney's storytelling and imagination.

WALT DISNEY
PICTURES



Once upon a time...

...there was parameter estimation

What are Parameters?

Consider some probability distributions:

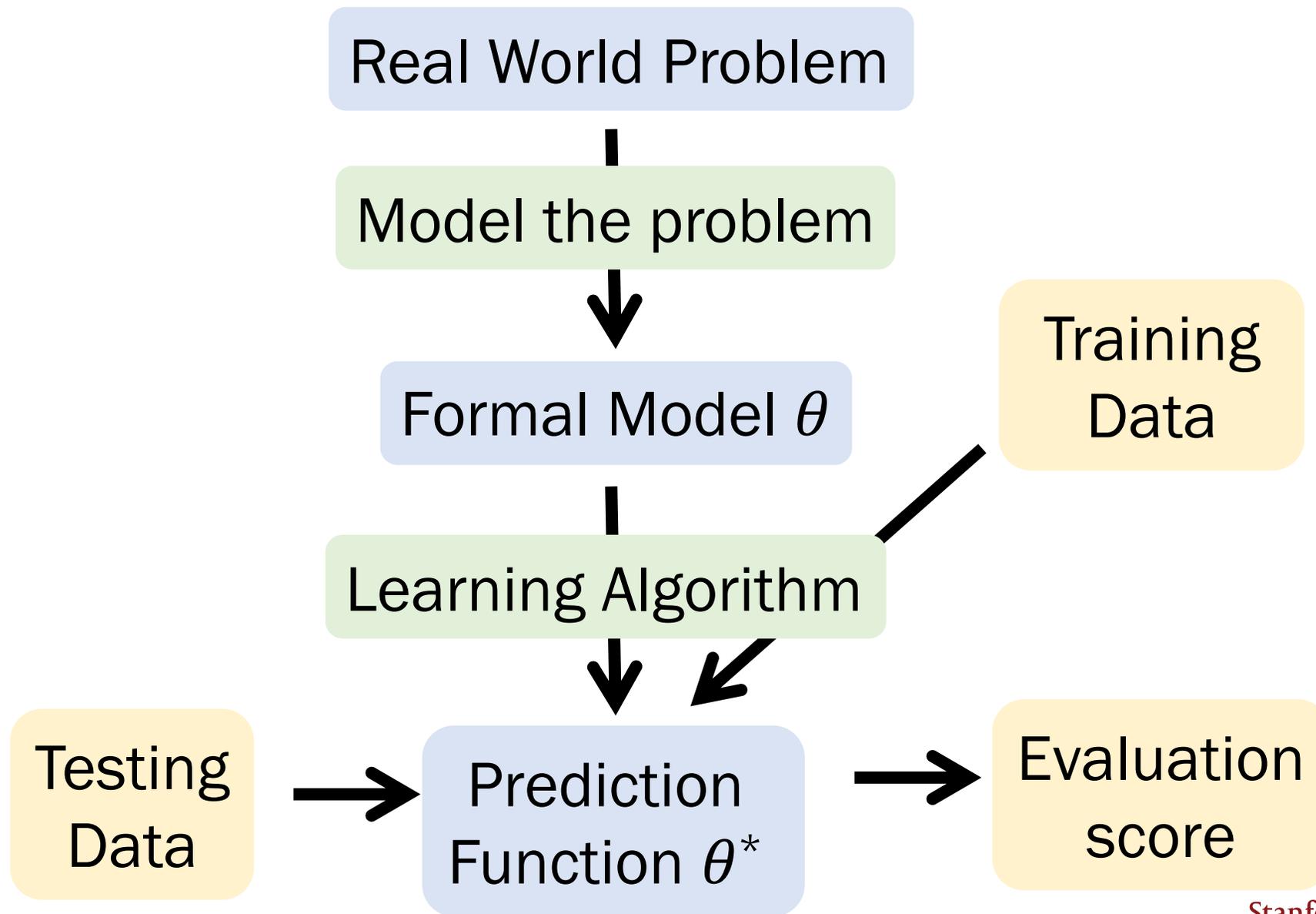
- $\text{Ber}(p)$ $\theta = p$
- $\text{Poi}(\lambda)$ $\theta = \lambda$
- $\text{Uni}(\alpha, \beta)$ $\theta = (\alpha, \beta)$
- $\text{Normal}(\mu, \sigma^2)$ $\theta = (\mu, \sigma^2)$
- $Y = mX + b$ $\theta = (m, b)$
- etc...

Call these “parametric models”

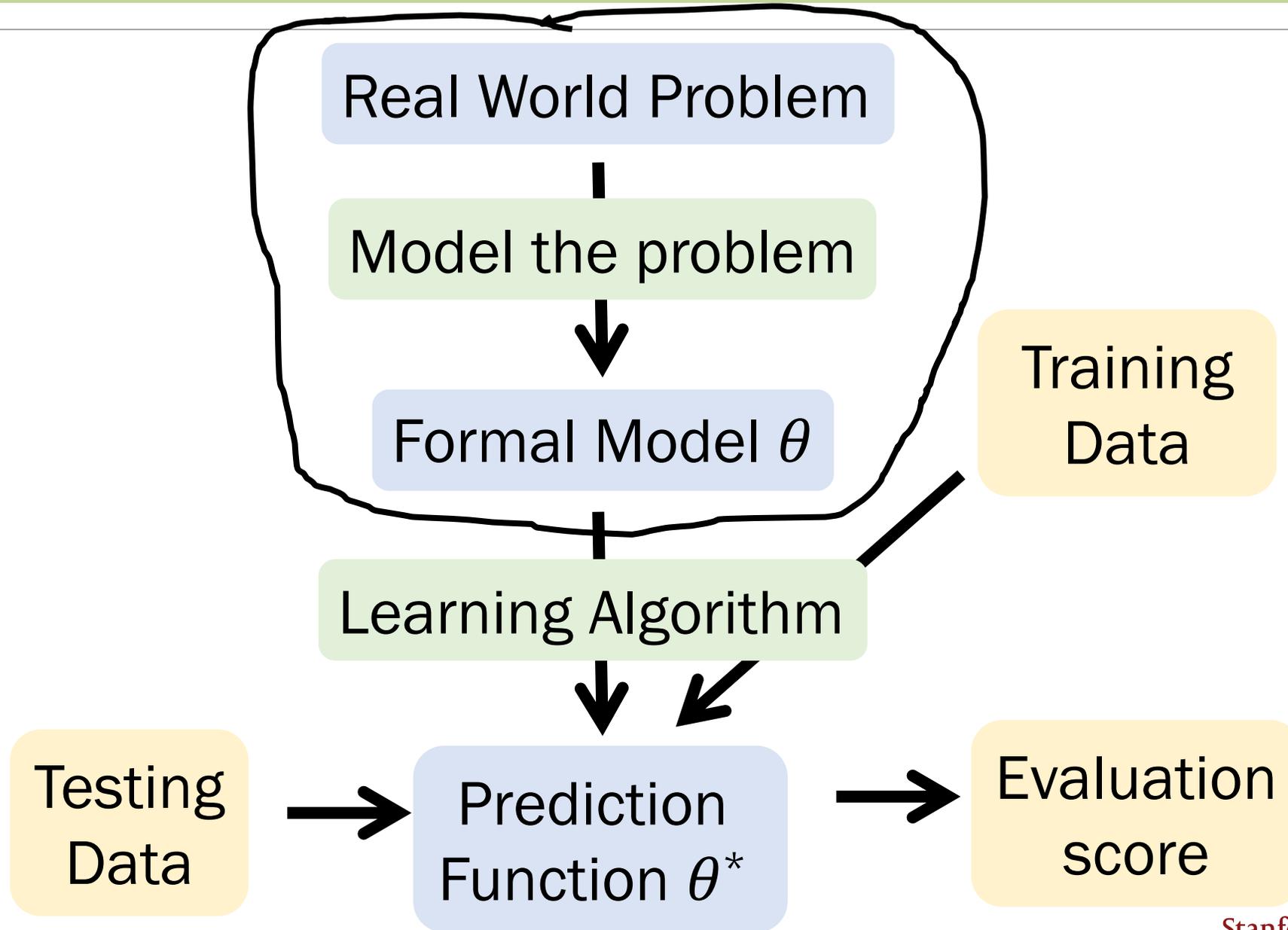
Given model, **parameters** yield actual distribution

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

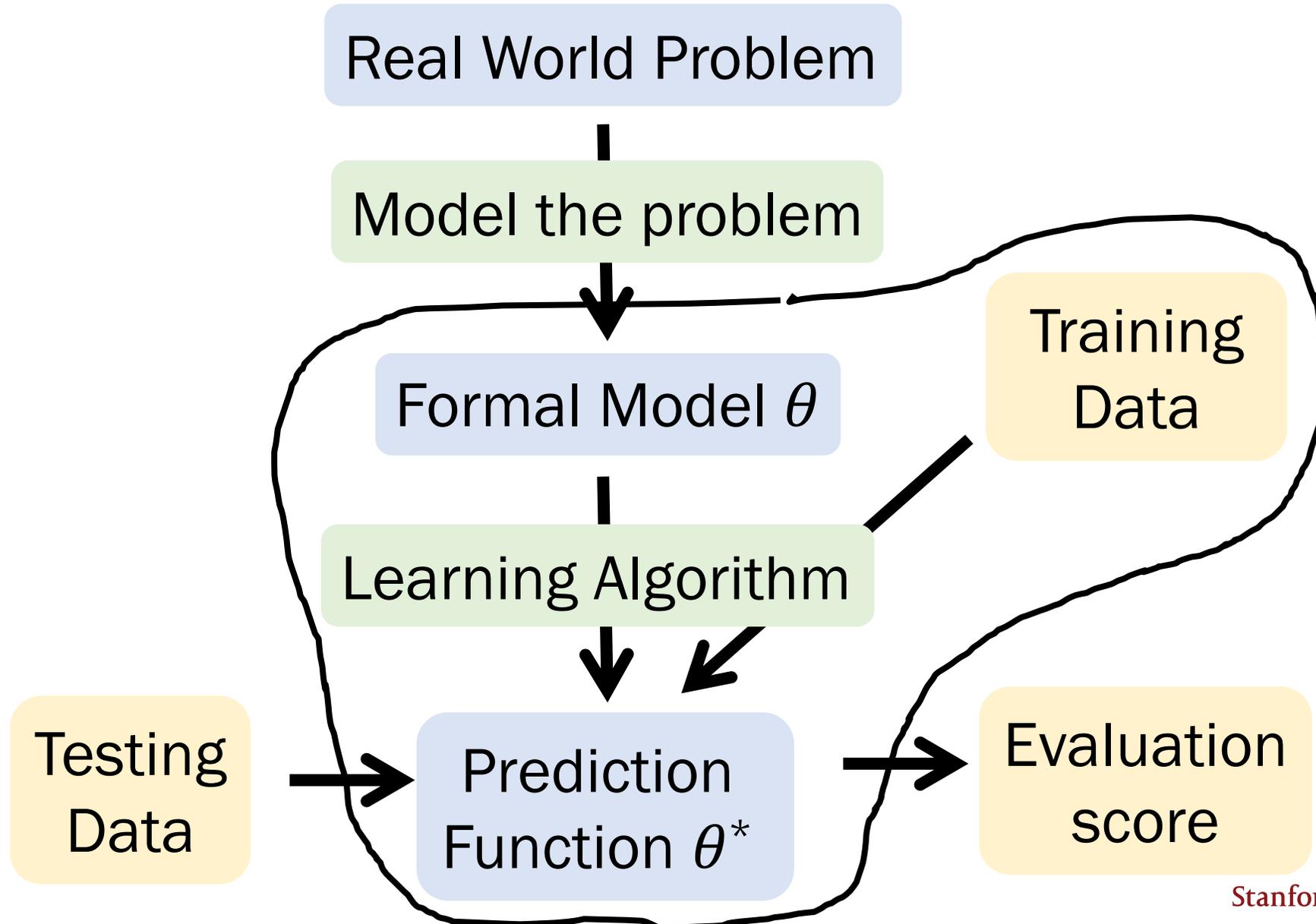
Why Do We Care?



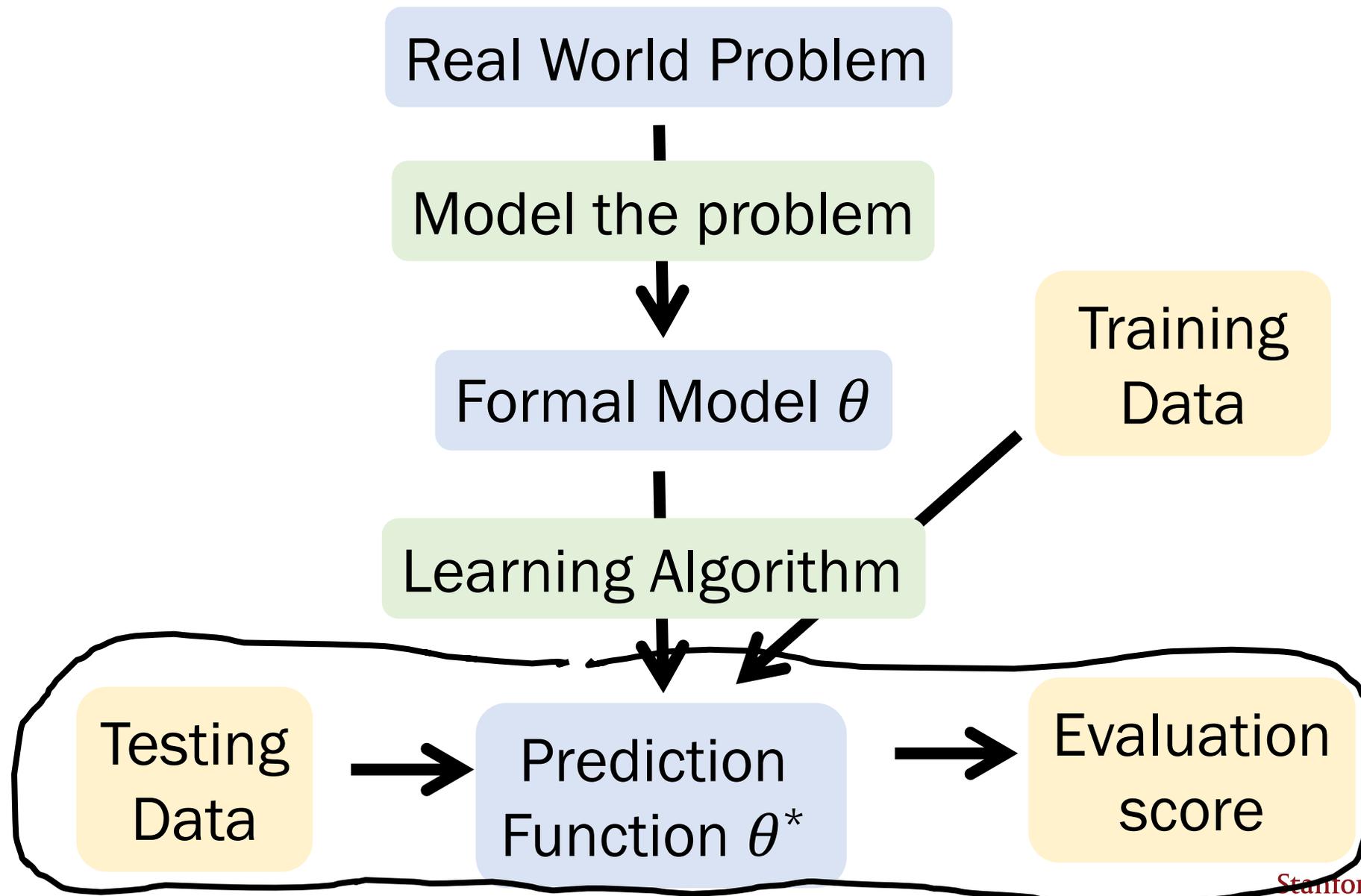
Modelling



Parameter Estimation (aka Training)

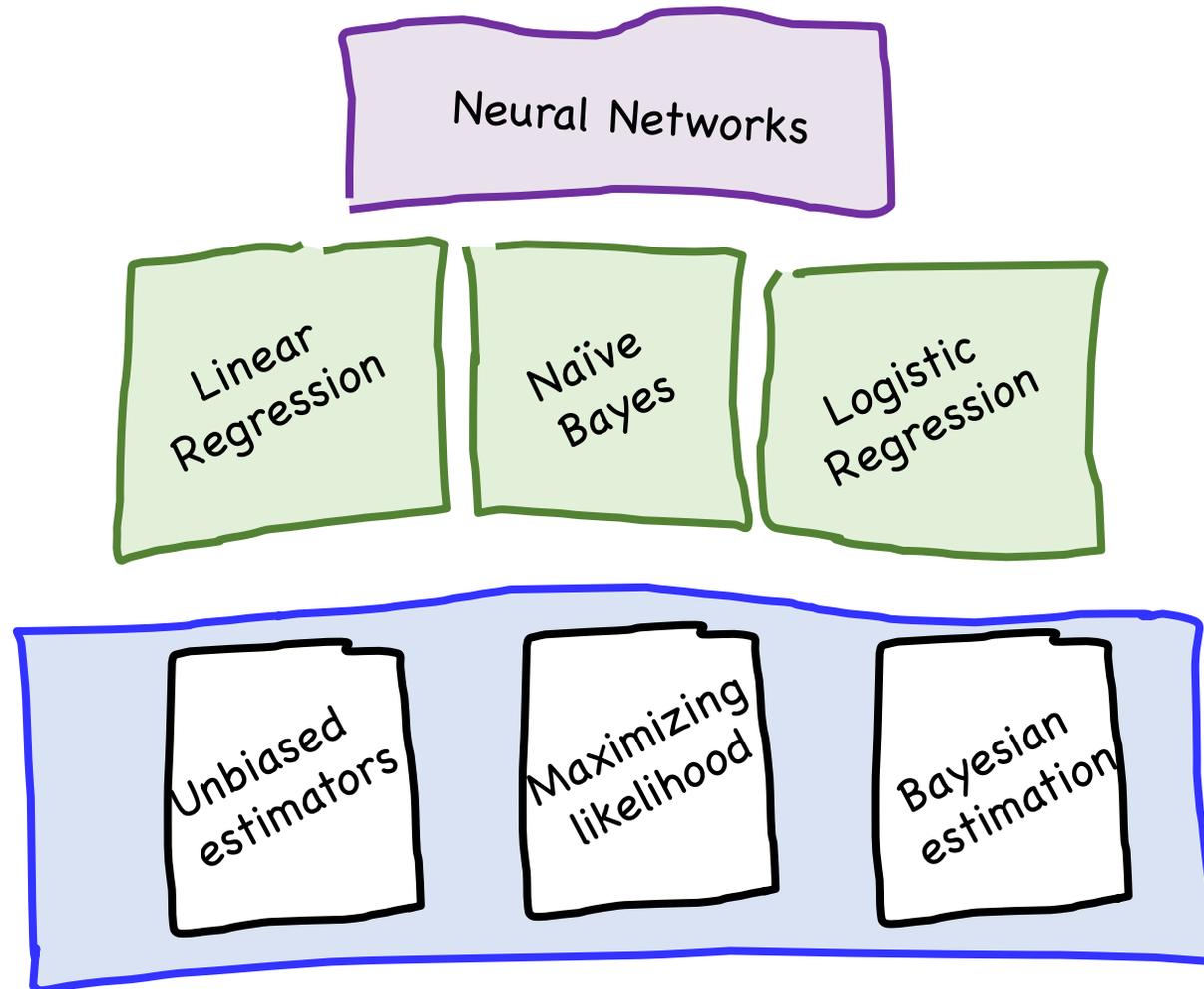


Testing

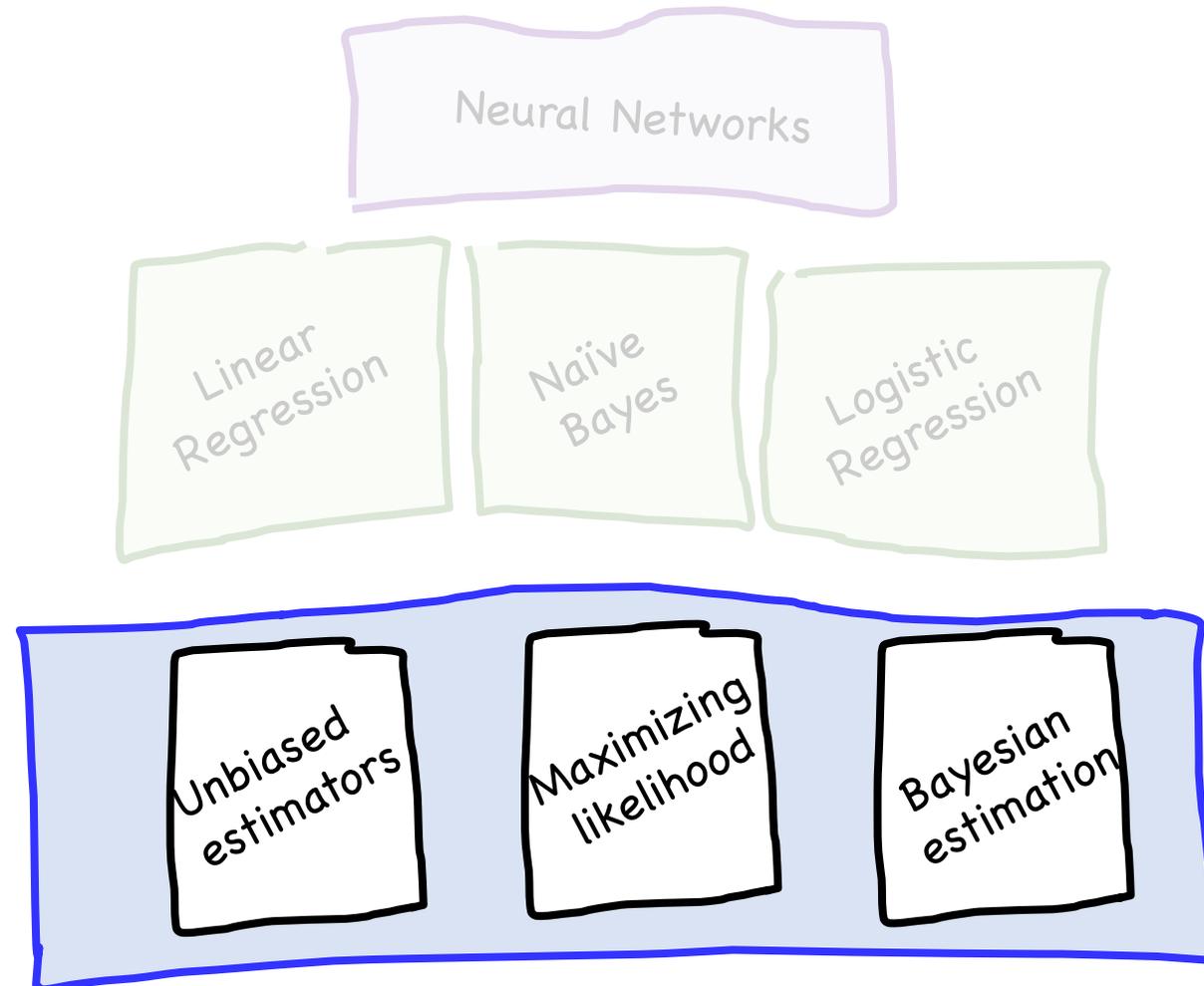


Basis for learning from data

Our Path



Parameter Estimation



We've already seen some estimations

X_1, X_2, \dots, 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

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 :

Parameter σ :

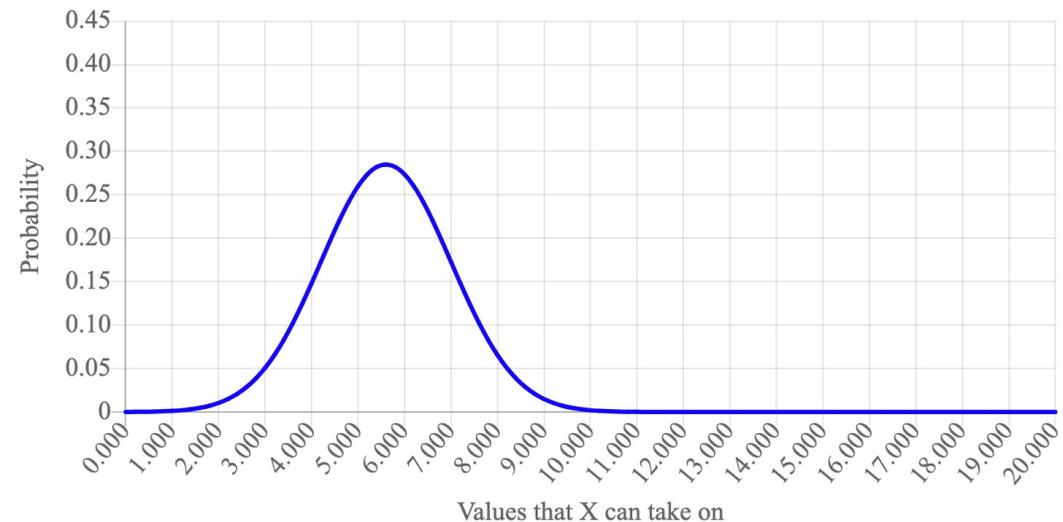
Likelihood

Likelihood: 1.9542923784106326e-15

Log Likelihood: -301.9

Best Seen: -301.9

PDF Graph







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, \dots, X_n



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

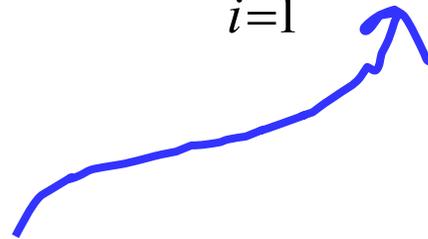
This is just a product since X_i are I.I.D.

We explicitly specify parameter θ of distribution



Likelihood (of data given parameters):

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



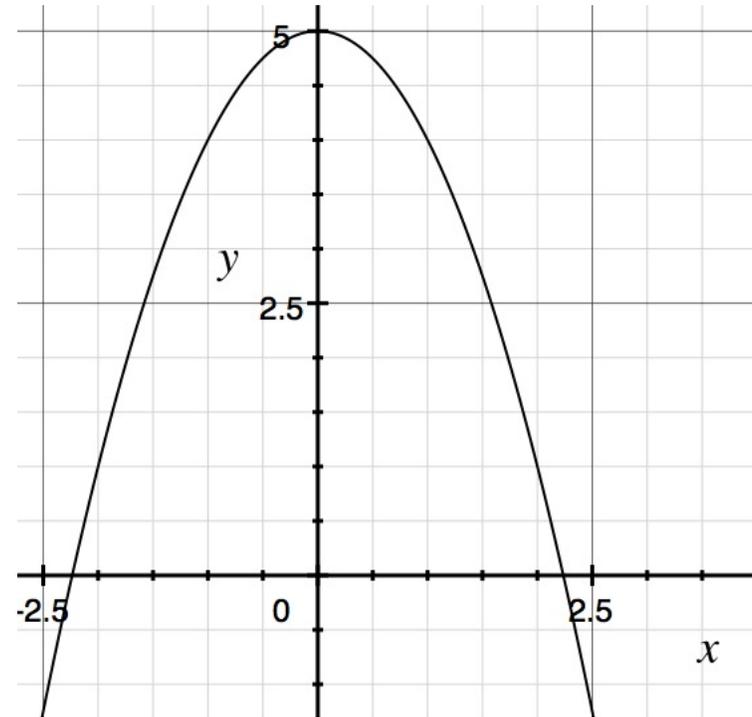
Either the
PDF (continuous) or
PMF (discrete), or
joint if multiple variables per datapoint

Argmax

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

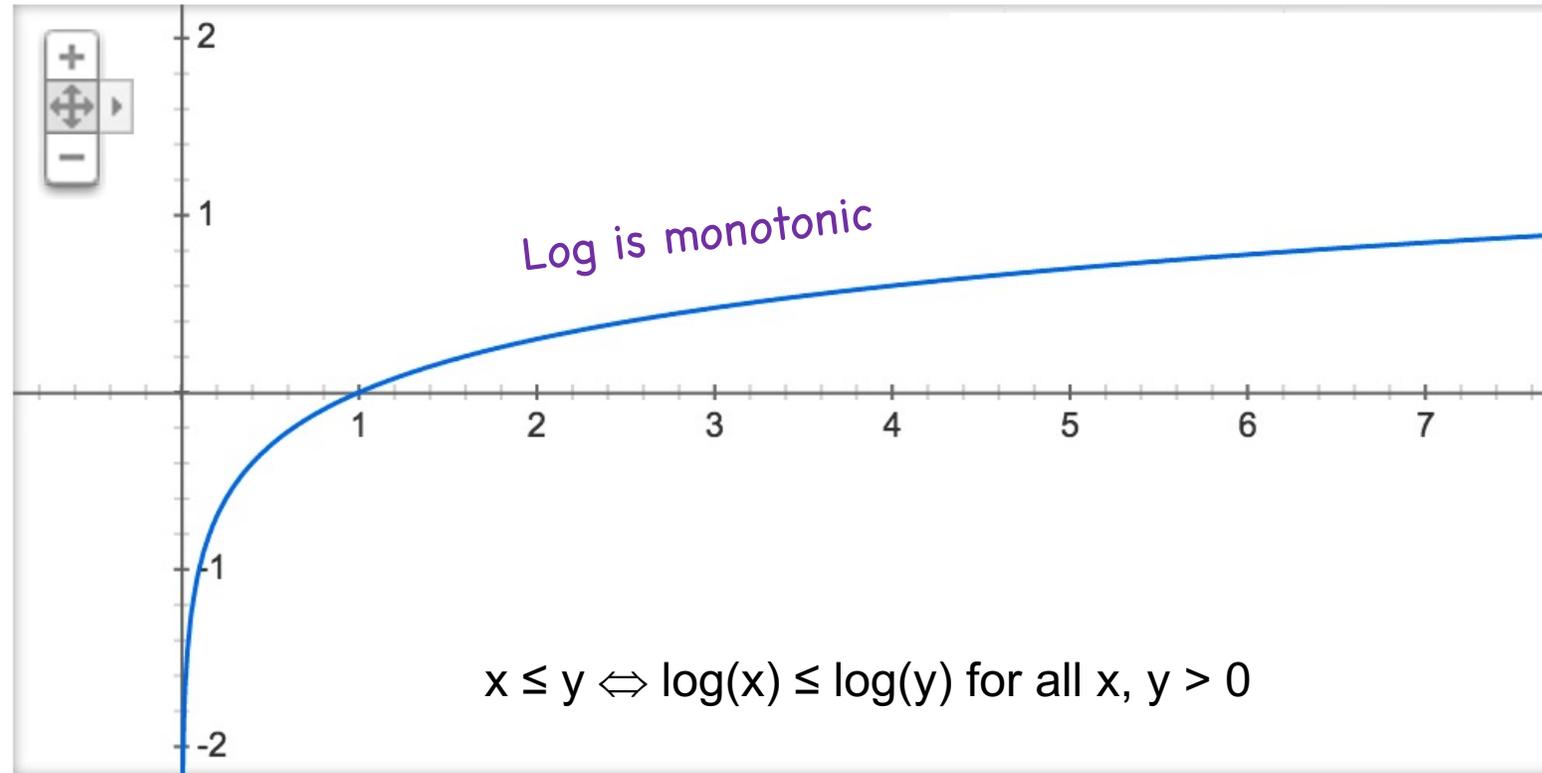
$$\max_x -x^2 + 5 = 5$$

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



Argmax of Log

Graph for $\log(x)$



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

Argmax of Log



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

Log I Love You

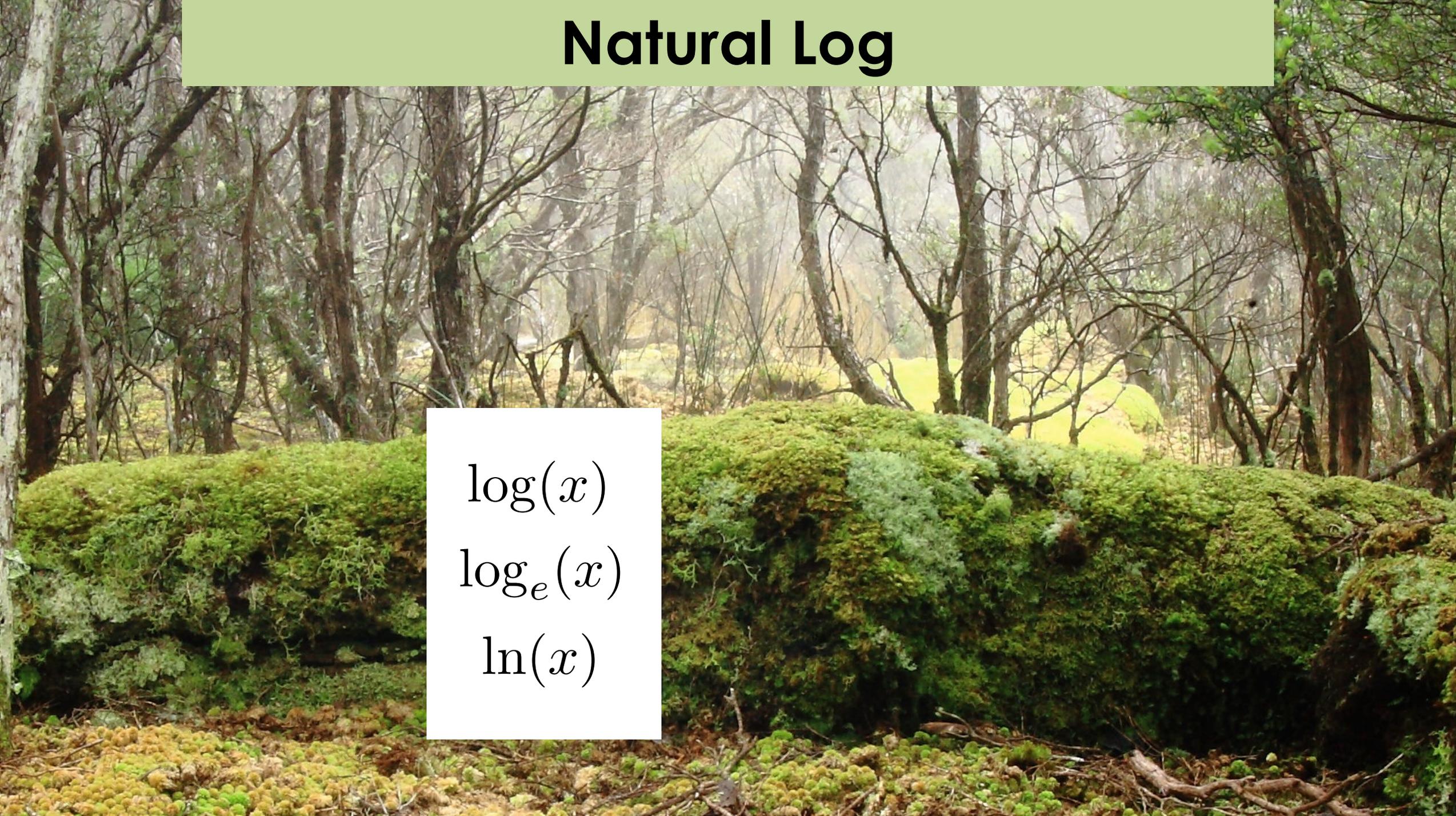
$$\log(ab) = \log(a) + \log(b)$$

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 choose parameters by finding the argmax of the log likelihood of our data



Maximum Likelihood

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

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

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

A cartoon illustration of a lion's face, likely from the Disney movie 'The Lion King'. The lion has orange fur, a black mane, and a single glowing green eye. A white rectangular text box with the text 'arg max' is overlaid on the lion's face, partially covering its eye and nose. The background is a dark blue, textured surface.

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

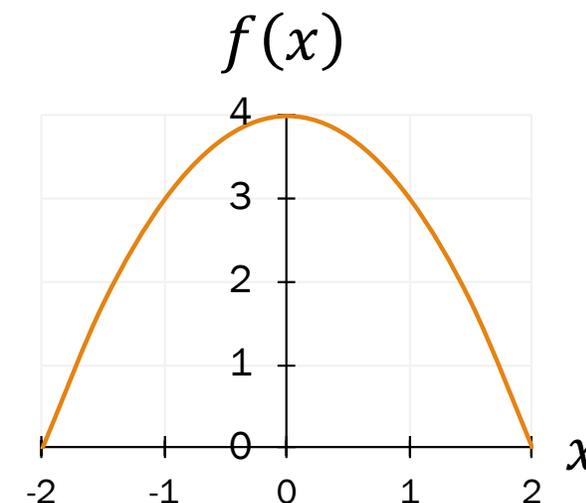
$$\frac{d}{dx} f(x) = \frac{d}{dx} (x^2 + 4) = 2x$$

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 $LL(\theta)$

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

2. Differentiate $LL(\theta)$ w.r.t. (each) θ

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

3. Solve

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

(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, \dots, X_n

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

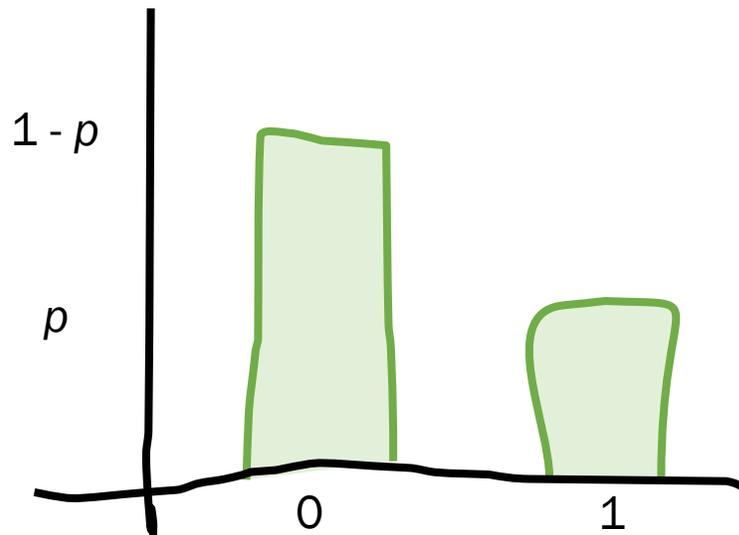


Maximizing Likelihood with Bernoulli

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

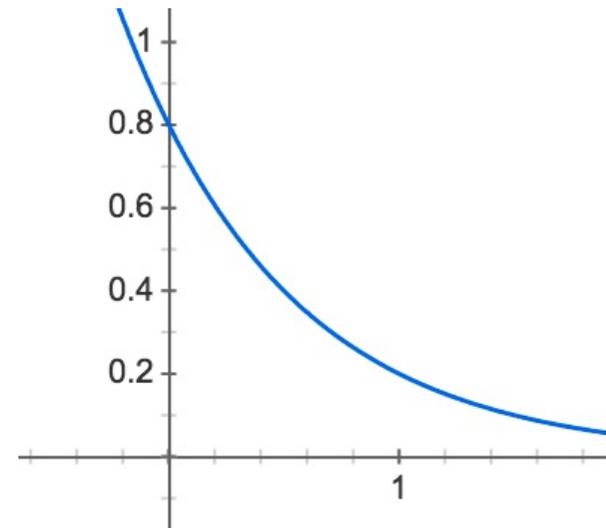
- $X_i \sim \text{Ber}(p)$
- Probability mass function, $f(X_i | p)$:

PMF of Bernoulli



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

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, \dots, 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, \dots, 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 $LL(\theta)$

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

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

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

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

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, \dots, 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, \dots, 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 $LL(\theta)$

$$\begin{aligned} LL(\theta) &= \sum_{i=1}^n [X_i \log p + (1 - X_i) \log(1 - p)] \\ &= Y(\log p) + (n - Y) \log(1 - p), \text{ where } Y = \sum_{i=1}^n X_i \end{aligned}$$

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, \dots, 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 $LL(\theta)$

$$\begin{aligned} LL(\theta) &= \sum_{i=1}^n [X_i \log p + (1 - X_i) \log(1 - p)] \\ &= Y(\log p) + (n - Y) \log(1 - p), \text{ where } Y = \sum_{i=1}^n X_i \end{aligned}$$

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, \dots, 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, \dots, X_n from the distribution F , yielding the following sample:

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

- 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, \dots, X_n from the distribution F , yielding the following sample:

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

- 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, \dots, X_n from the distribution F , yielding the following sample:

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

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

- What is p_{MLE} , the MLE of the parameter p ? C. 0.8
- 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, \dots, 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 $LL(\theta)$

$$\begin{aligned} LL(\theta) &= \sum_{i=1}^n \log \left(\frac{e^{-\lambda} \lambda^{X_i}}{X_i!} \right) = \sum_{i=1}^n (-\lambda \log e + X_i \log \lambda - \log X_i!) \\ &= -n\lambda + \log(\lambda) \sum_{i=1}^n X_i - \sum_{i=1}^n \log(X_i!) \quad (\text{using natural log, } \ln e = 1) \end{aligned}$$

Maximum Likelihood with Poisson

Consider a sample of n i.i.d. RVs X_1, X_2, \dots, 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 $LL(\theta)$

$$\begin{aligned} LL(\theta) &= \sum_{i=1}^n \log \left(\frac{e^{-\lambda} \lambda^{X_i}}{X_i!} \right) = \sum_{i=1}^n (-\lambda \log e + X_i \log \lambda - \log X_i!) \\ &= -n\lambda + \log(\lambda) \sum_{i=1}^n X_i - \sum_{i=1}^n \log(X_i!) \quad (\text{using natural log, } \ln e = 1) \end{aligned}$$

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, \dots, 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 $LL(\theta)$

$$\begin{aligned} LL(\theta) &= \sum_{i=1}^n \log \left(\frac{e^{-\lambda} \lambda^{X_i}}{X_i!} \right) = \sum_{i=1}^n (-\lambda \log e + X_i \log \lambda - \log X_i!) \\ &= -n\lambda + \log(\lambda) \sum_{i=1}^n X_i - \sum_{i=1}^n \log(X_i!) \quad (\text{using natural log, } \ln e = 1) \end{aligned}$$

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, \dots, 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 $LL(\theta)$

$$\begin{aligned} LL(\theta) &= \sum_{i=1}^n \log \left(\frac{e^{-\lambda} \lambda^{X_i}}{X_i!} \right) = \sum_{i=1}^n (-\lambda \log e + X_i \log \lambda - \log X_i!) \\ &= -n\lambda + \log(\lambda) \sum_{i=1}^n X_i - \sum_{i=1}^n \log(X_i!) \quad (\text{using natural log, } \ln e = 1) \end{aligned}$$

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, \dots, 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 $LL(\theta)$

$$\begin{aligned} LL(\theta) &= \sum_{i=1}^n \log \left(\frac{e^{-\lambda} \lambda^{X_i}}{X_i!} \right) = \sum_{i=1}^n (-\lambda \log e + X_i \log \lambda - \log X_i!) \\ &= -n\lambda + \log(\lambda) \sum_{i=1}^n X_i - \sum_{i=1}^n \log(X_i!) \quad (\text{using natural log, } \ln e = 1) \end{aligned}$$

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

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

Maximum Likelihood with Normal

Consider a sample of n i.i.d. random variables X_1, X_2, \dots, 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 $\theta_{MLE} = (\mu_{MLE}, \sigma_{MLE}^2)$?

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

$$\begin{aligned} LL(\theta) &= \sum_{i=1}^n \log\left(\frac{1}{\sqrt{2\pi}\sigma} e^{-(X_i-\mu)^2/(2\sigma^2)}\right) = \sum_{i=1}^n \left[-\log(\sqrt{2\pi}\sigma) - (X_i - \mu)^2/(2\sigma^2)\right] \\ & \hspace{20em} \text{(using natural log)} \\ &= -\sum_{i=1}^n \log(\sqrt{2\pi}\sigma) - \sum_{i=1}^n [(X_i - \mu)^2/(2\sigma^2)] \end{aligned}$$

Maximum Likelihood with Normal

Consider a sample of n i.i.d. random variables X_1, X_2, \dots, 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 $\theta_{MLE} = (\mu_{MLE}, \sigma_{MLE}^2)$?

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

with respect to μ

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

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

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

Maximum Likelihood with Normal

Consider a sample of n i.i.d. random variables X_1, X_2, \dots, 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 $\theta_{MLE} = (\mu_{MLE}, \sigma_{MLE}^2)$?

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

with respect to μ $LL(\theta) = -\sum_{i=1}^n \log(\sqrt{2\pi}\sigma) - \sum_{i=1}^n [(X_i - \mu)^2/(2\sigma^2)]$ with respect to σ

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

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

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

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

Maximum Likelihood with Normal

Consider a sample of n i.i.d. random variables X_1, X_2, \dots, 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 $\theta_{MLE} = (\mu_{MLE}, \sigma_{MLE}^2)$?

3. Solve resulting equations

Two equations, two unknowns:

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

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

First, solve for μ_{MLE} :

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

$$\Rightarrow \sum_{i=1}^n X_i = n\mu$$

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

unbiased

Maximum Likelihood with Normal

Consider a sample of n i.i.d. random variables X_1, X_2, \dots, 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 $\theta_{MLE} = (\mu_{MLE}, \sigma_{MLE}^2)$?

3. Solve resulting equations

Two equations, two unknowns:

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

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

First, solve for μ_{MLE} :

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

$$\Rightarrow \sum_{i=1}^n X_i = n\mu$$

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

unbiased

Next, solve for σ_{MLE} :

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

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

$$\Rightarrow \sigma_{MLE}^2 = \frac{1}{n} \sum_{i=1}^n (X_i - \mu_{MLE})^2$$

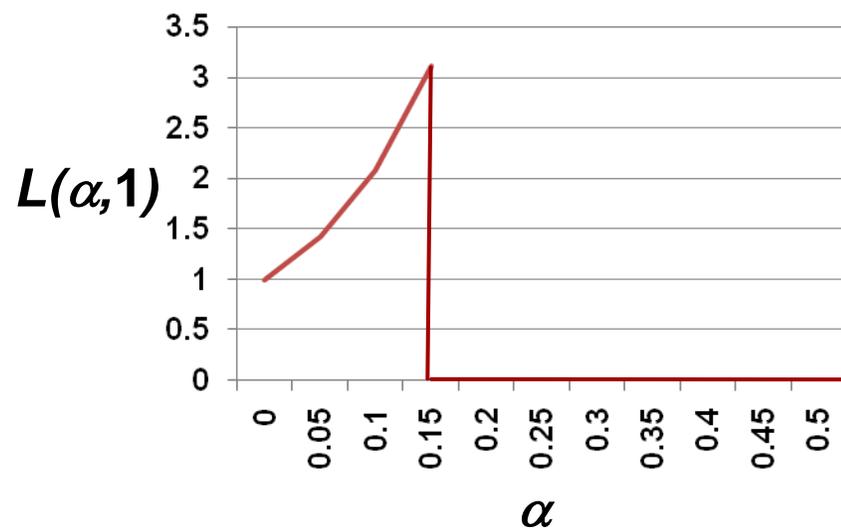
biased

Understanding MLE with Uniform

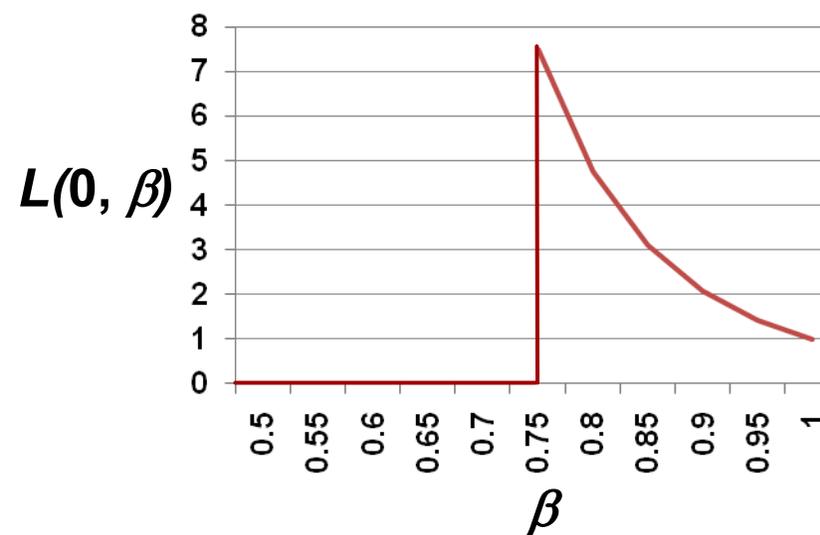
Consider I.I.D. random variables X_1, X_2, \dots, X_n

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

Likelihood: $L(\alpha, 1)$



Likelihood: $L(0, \beta)$



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, $\sigma_{MLE}^2 = \frac{1}{n} \sum_{i=1}^n (X_i - \mu_{MLE})^2$
 - 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:** $\lim_{n \rightarrow \infty} P(|\hat{\theta} - \theta| < \varepsilon) = 1$ 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.

