

# Parameter Estimation

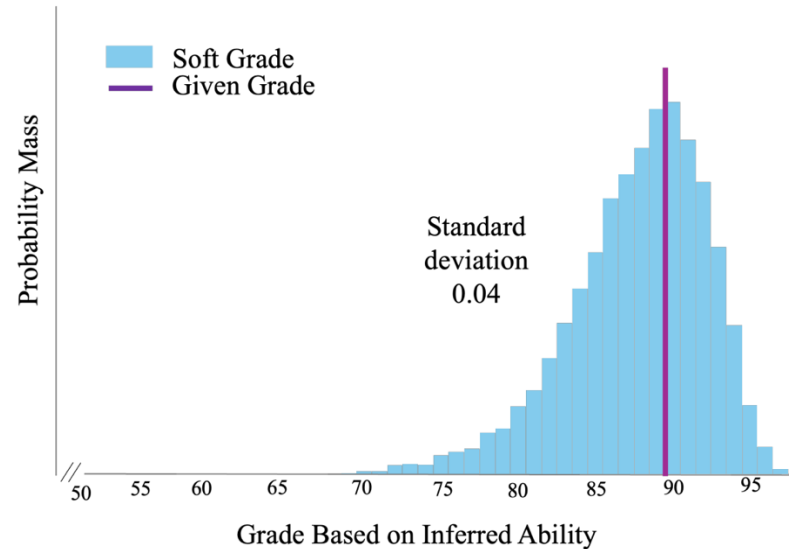
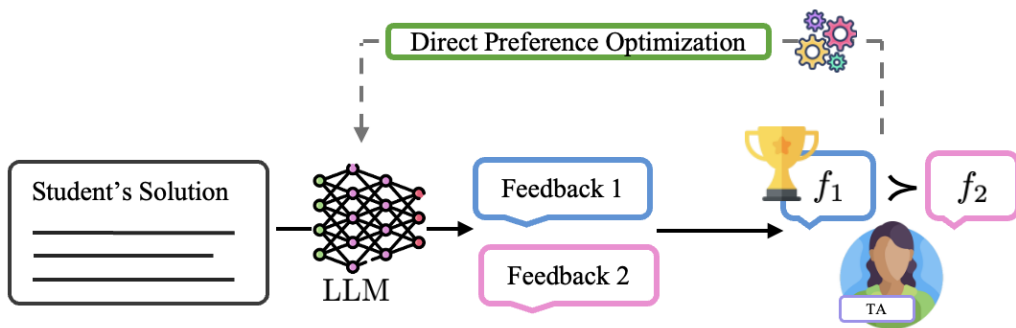
Juliette Woodrow  
CS109, Stanford University



Grew up in Colorado



Help build and run Code in Place

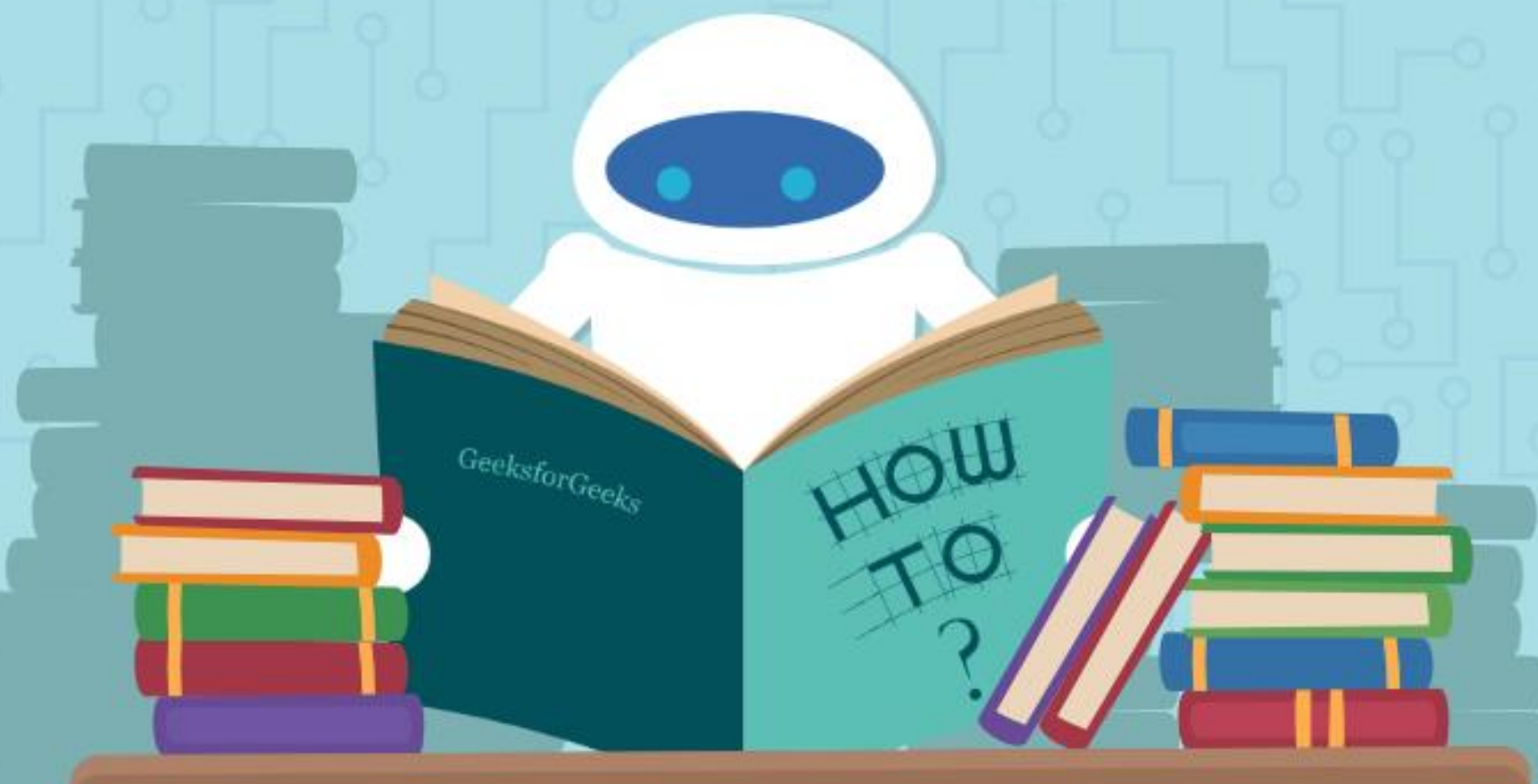


Research how we can use AI to improve teaching and learning!!!

# Where are we in CS109?



# MACHINE LEARNING

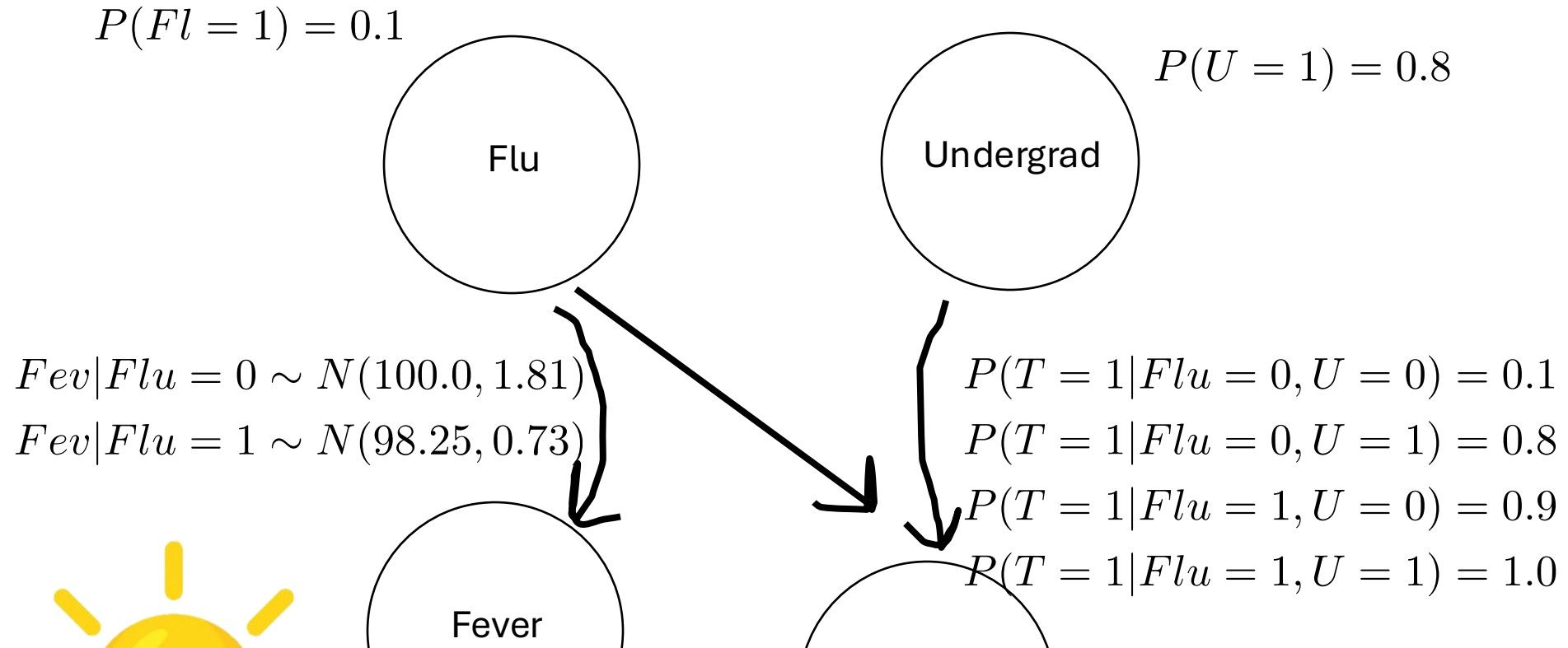




# General “Inference”



# 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?

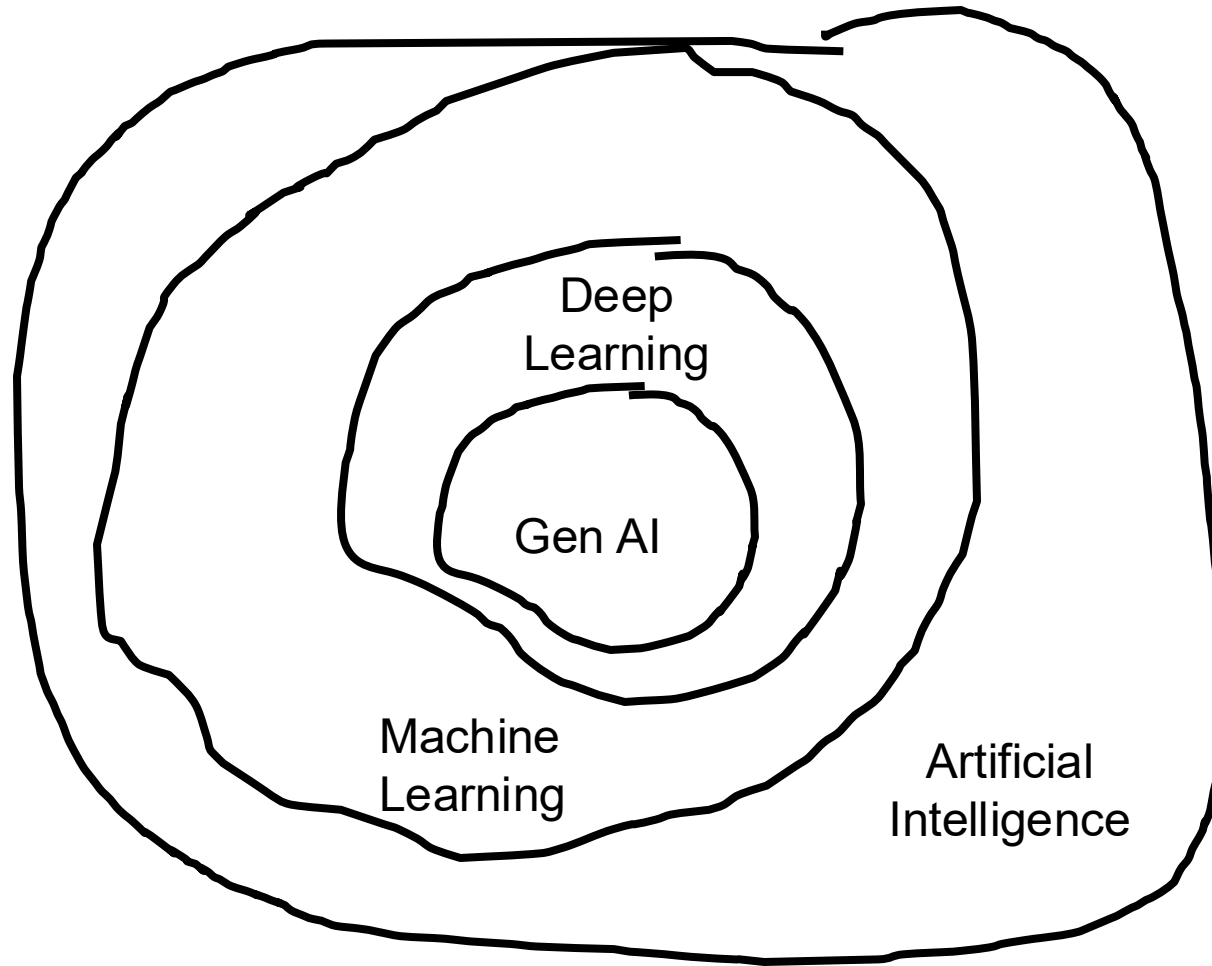
Suspense

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?

# Machine Learning

# AI and Machine Learning



ML: Rooted in probability theory

# Our Path

Deep Learning

Core Algorithms

Parameter Estimation

# Our Path

Deep Learning

Core Algorithms

Unbiased  
estimators

Maximizing  
likelihood

The image features the iconic Walt Disney Pictures logo centered over a scene of Cinderella Castle at night. The castle is brightly lit with warm yellow and orange lights, set against a deep blue twilight sky with wispy clouds and a few stars. The logo itself is rendered in a classic, white, cursive script with a subtle glow and a slight shadow, giving it a three-dimensional appearance. Below the main logo, the word "PICTURES" is written in a smaller, white, all-caps serif font. The entire scene is reflected in a body of water in the foreground, which also shows some light trails from the sky.

WALT DISNEY  
PICTURES

Review

# Shorthand for Equality Events

## Our shorthand notation

$$f(x|\theta)$$

$x$  Is shorthand for the event  $X = x$

$\theta$  Is shorthand for the event  $\Theta = \theta$

## Full Notation

$$f(X = x | \Theta = \theta)$$

End Review



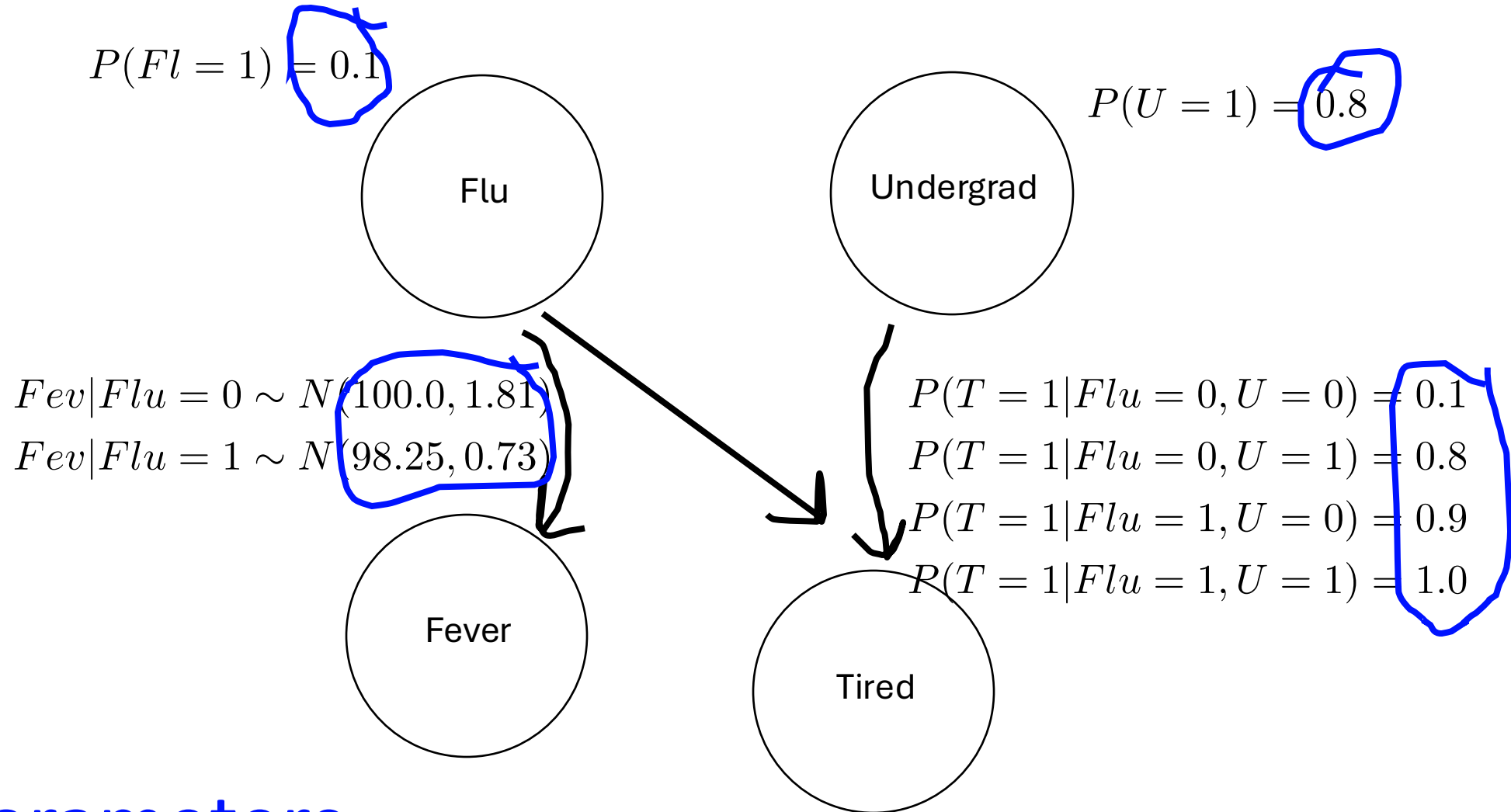
Once upon a time...

...there was parameter estimation

# What are Parameters?

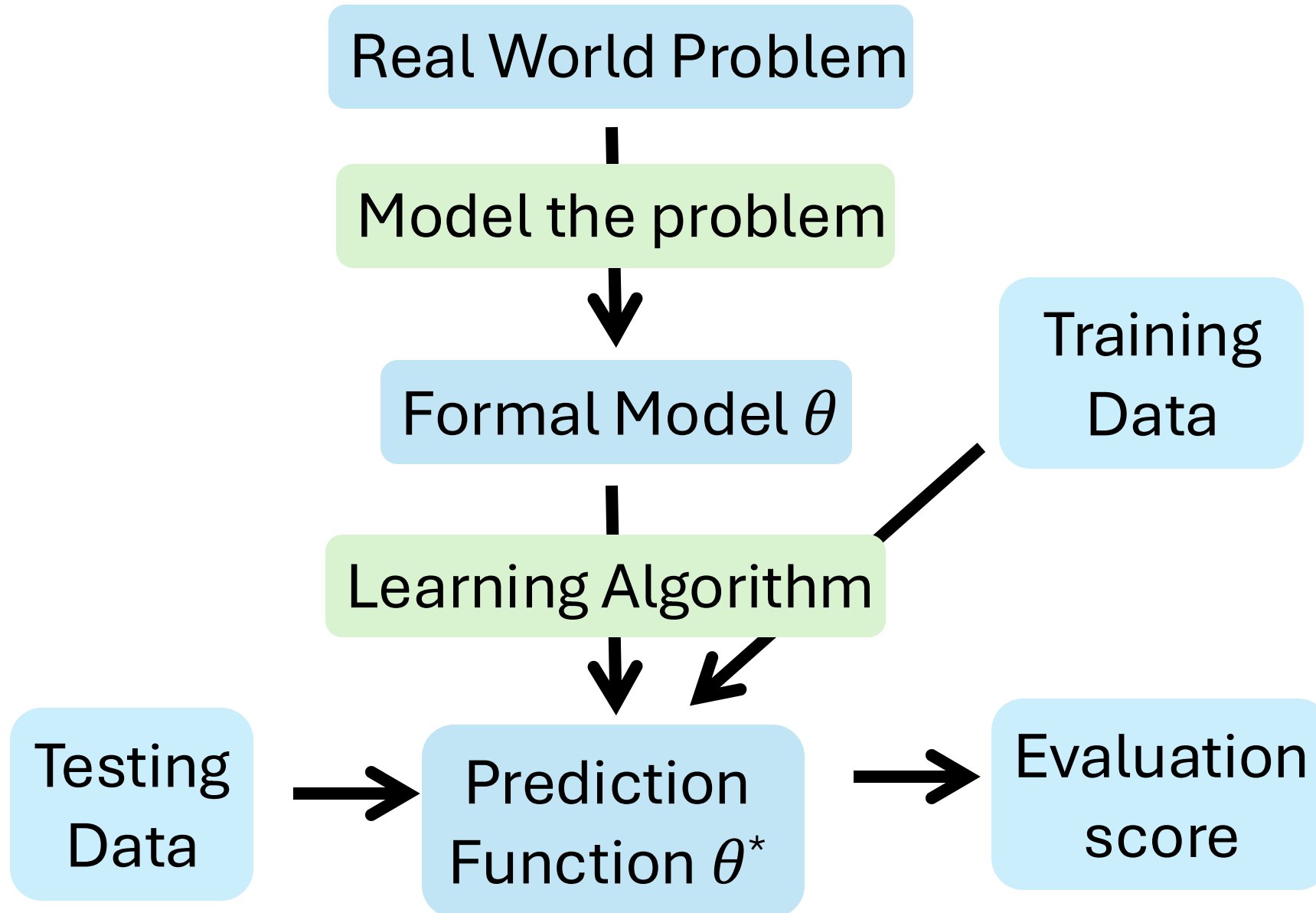
- Consider some probability distributions:
  - Ber( $p$ )  $\theta = p$
  - Poi( $\lambda$ )  $\theta = \lambda$
  - Uni( $\alpha, \beta$ )  $\theta = (\alpha, \beta)$
  - Normal( $\mu, \sigma^2$ )  $\theta = (\mu, \sigma^2)$
  - $Y = \mathbf{m}X + \mathbf{b}$   $\theta = (m, b)$
  - etc...
- Call these “parametric models”
- Given model, **parameters** yield actual distribution
  - Usually refer to parameters of distribution as  $\theta$
  - Note that  $\theta$  that can be a vector of parameters

# What are Parameters?

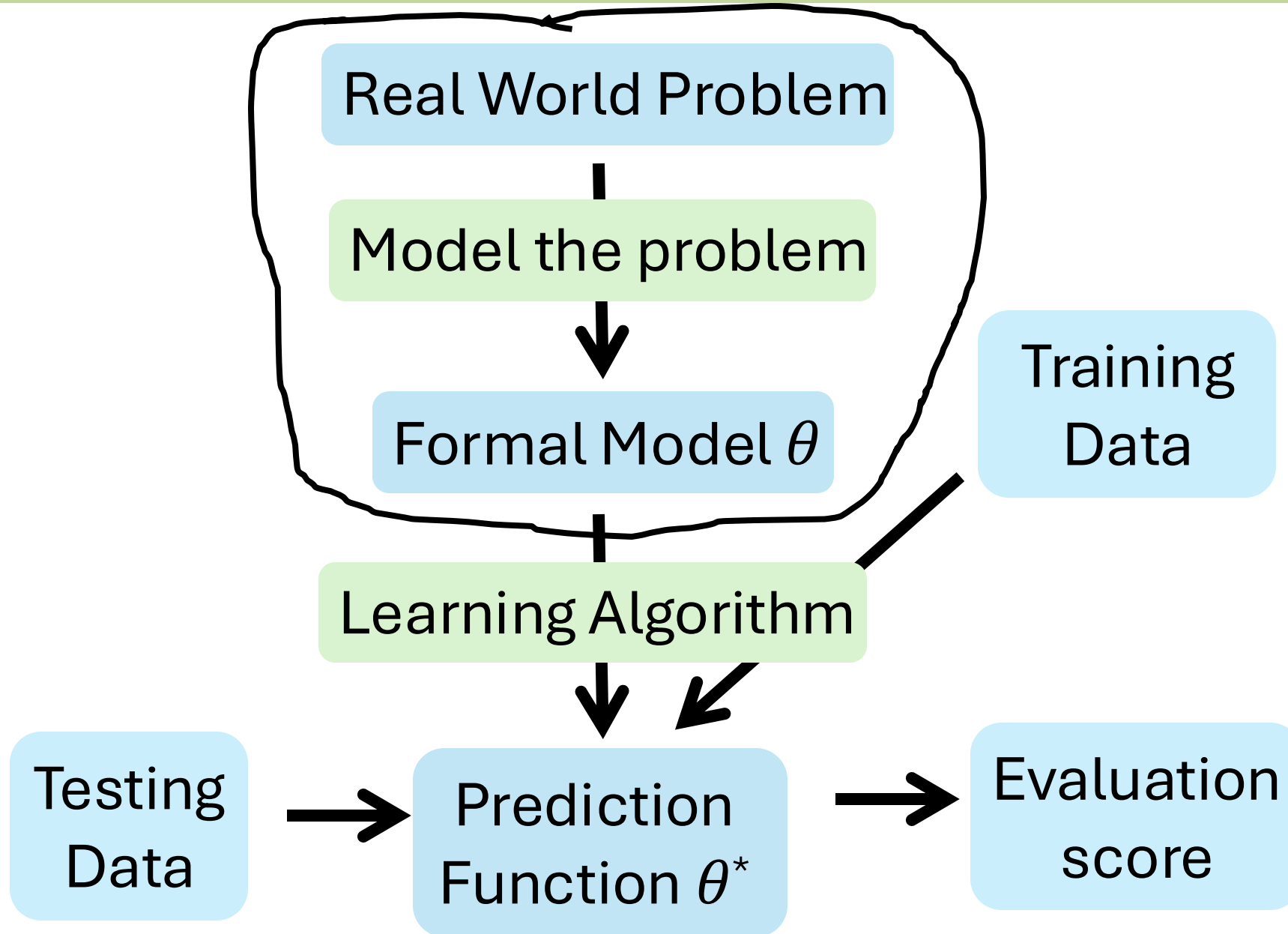


Parameters

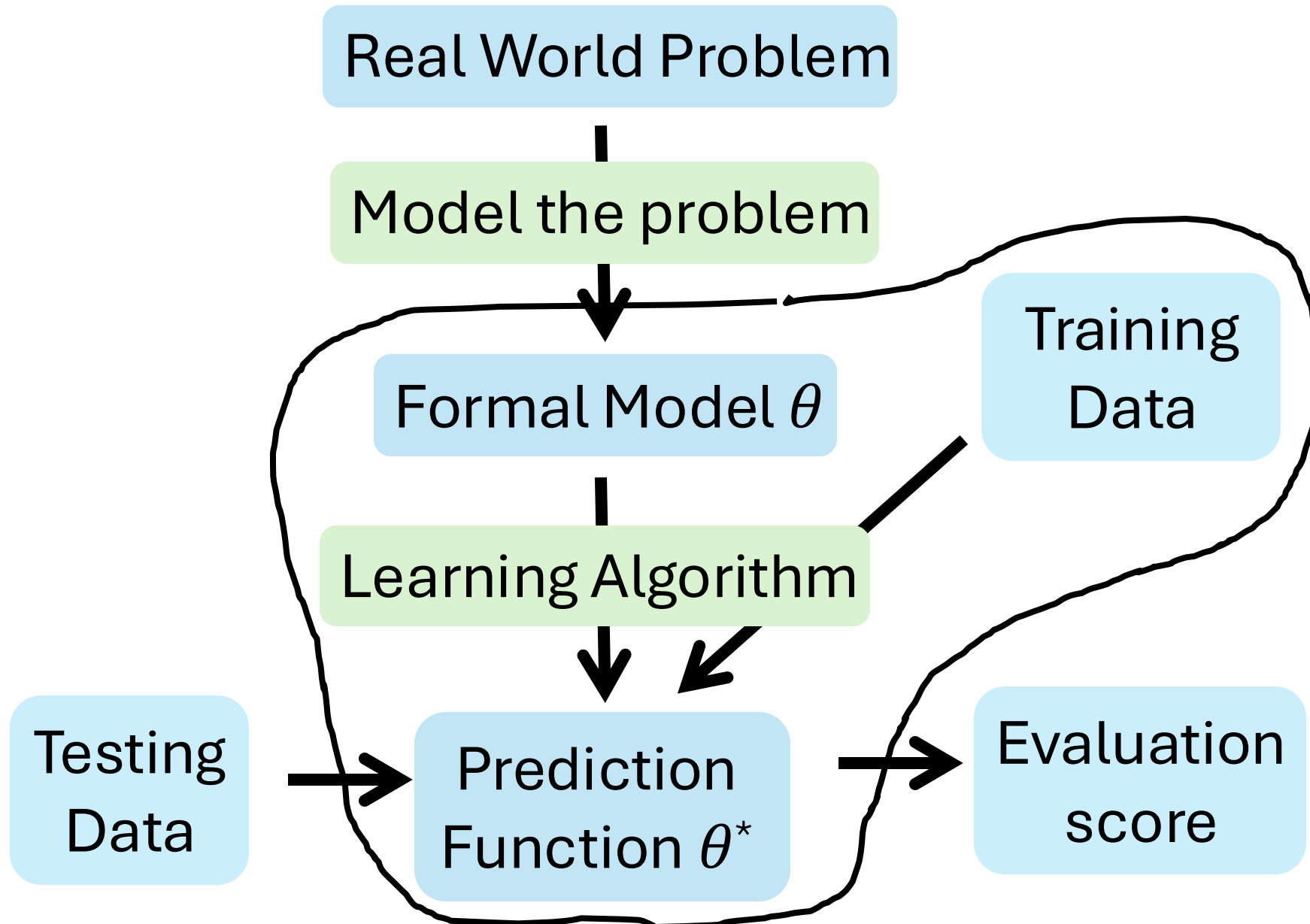
# Why Do We Care?



# Modelling



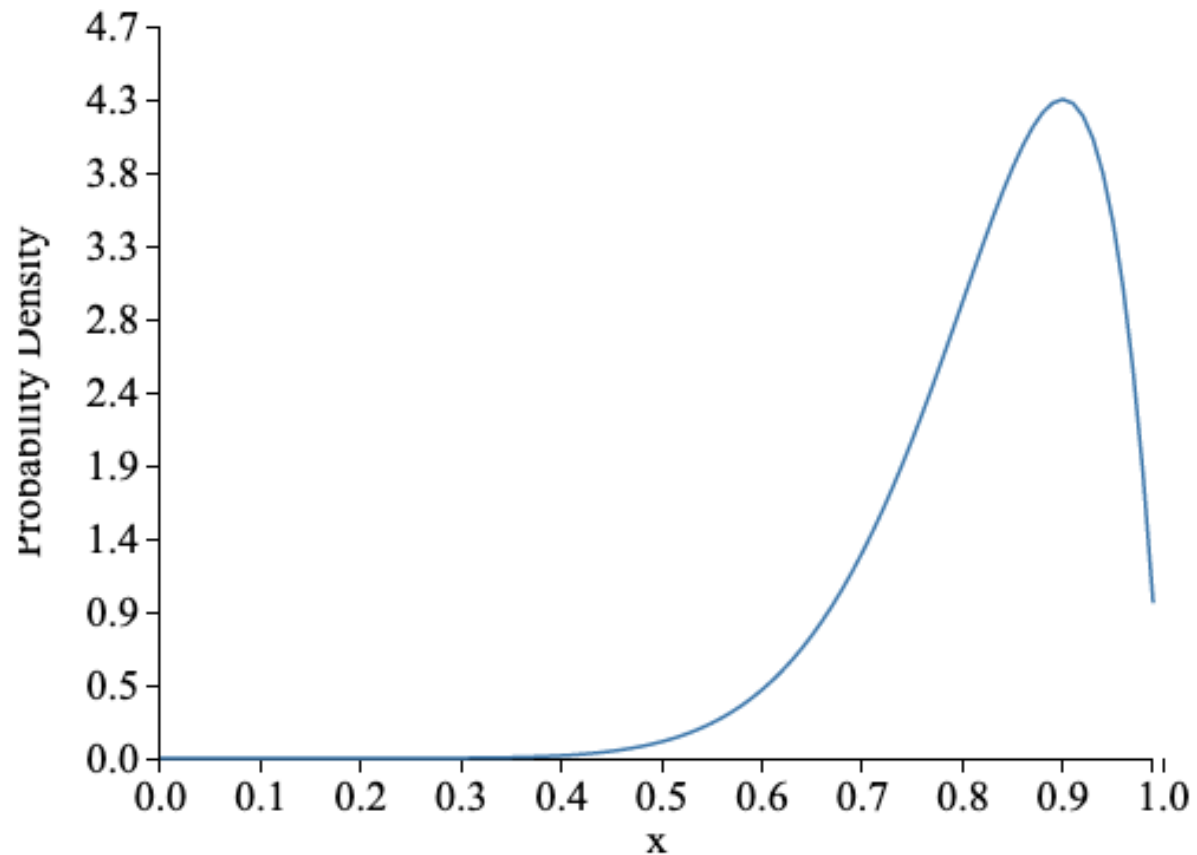
# Parameter Estimation (aka Training)



We have already seen some  
parameter estimators!

9 Heads out of 10 Flips. What is your Belief in  $p$ ?

$$f(X = x | H = 9, T = 1)$$



# Unbiased Estimators of Mean and Variance

- $X_1, X_2, \dots, X_n$  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  $\mu$

- Sample variance:

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

unbiased **estimate** of  $\sigma^2$

# Our Path

Deep Learning

Core Algorithms



Unbiased  
estimators

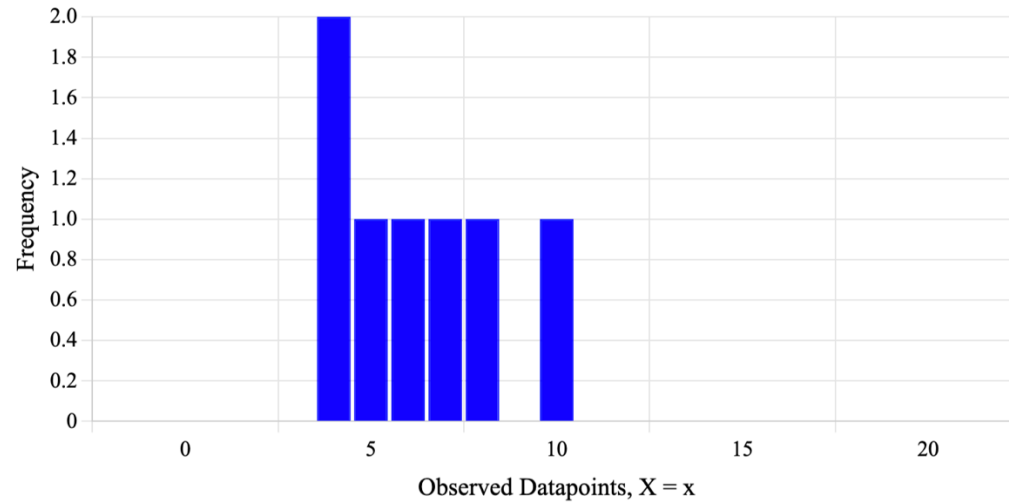
Maximizing  
likelihood

Unbiased Estimation is a limited tool:  
how could we use that for fitting WebMD?

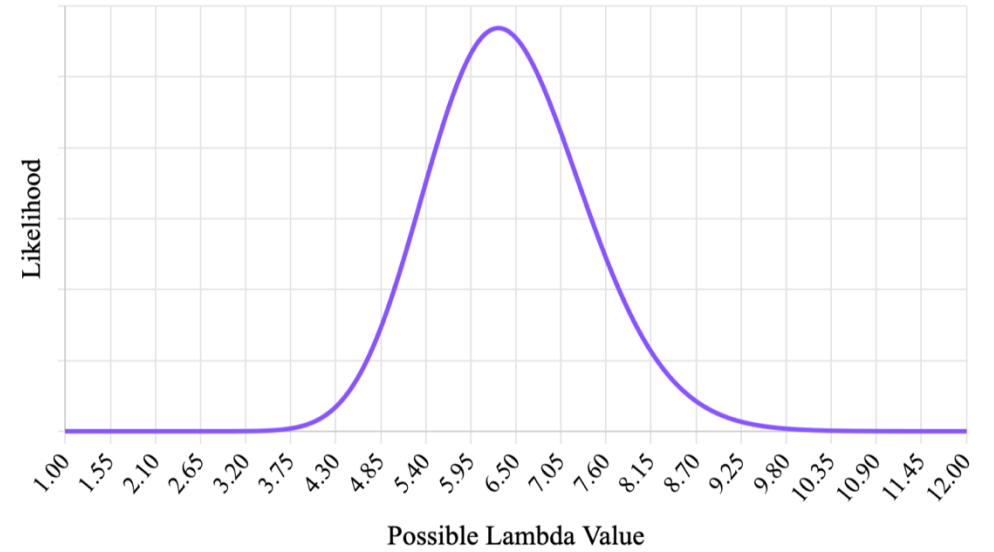
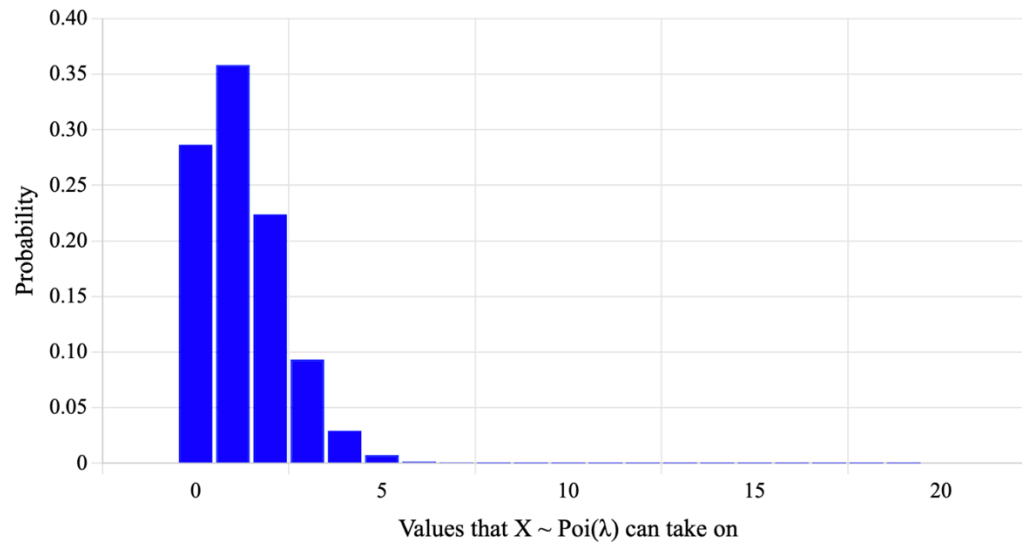
Great idea in Machine Learning

# To the Course Reader!

Histogram of Data



Parameter  $\lambda$ :



# How to Choose the “Best” Parameters: MLE



We want to choose the parameter value that maximizes the probability of the data.

# How to Choose the “Best” Parameters: MLE

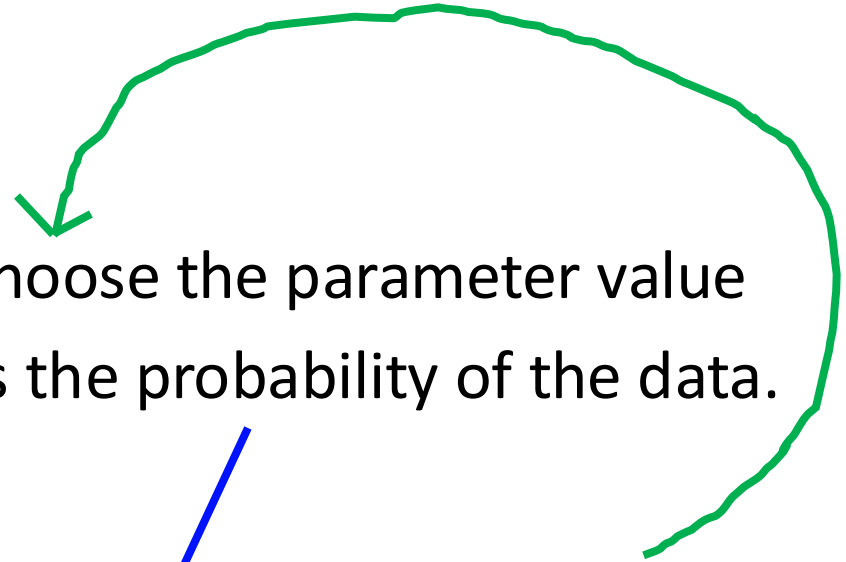


We want to choose the parameter value that maximizes the probability of the data.

Maximum

Likelihood

Estimation!





“I feel seen”



# How to Choose the “Best” Parameters: MLE



We want to choose the parameter value that maximizes the probability of the data.

Maximum

Likelihood

Estimation!

How do we quantify “probability of the data”?

# Likelihood Definition

Wikipedia:

## Likelihood function

[Article](#) [Talk](#)

From Wikipedia, the free encyclopedia

The **likelihood function** (often simply called the **likelihood**) is the [joint probability](#) (or probability density) of [observed data](#) viewed as a function of the [parameters](#) of a [statistical model](#).<sup>[1] [2] [3]</sup>

A generalized term for “PDF / PMF / Joint”  
of data as a function of parameters

# The Likelihood Function

**Definition:** The probability of our observed data if our parameters were  $\theta$ .

$$L(\theta) = P(\text{data}|\theta)$$



$\theta$  is shorthand for parameter(s)  
(if we have a Poisson,  $\theta = \lambda$ )

# The Likelihood Function

**Definition:** The probability of our observed data if our parameters were  $\theta$ .

If we had a single observation,  $X = x$ :

$$\begin{aligned}L(\theta) &= P(X = x | \theta) \\ &= P(x | \theta)\end{aligned}$$

(in our example, this would just be the Poisson PMF)

# The Likelihood Function

**Definition:** The *joint* probability of our observed data if our parameters were  $\theta$ .

For a list of observations,  $[x_1, x_2, \dots, x_n]$ :

$$L(\theta) = P(x_1, x_2, \dots, x_n | \theta)$$

# The Likelihood Function

**Definition:** The *joint* probability of our observed data if our parameters were  $\theta$ .

For a list of observations,  $[x_1, x_2, \dots, x_n]$ :

$$L(\theta) = P(x_1, x_2, \dots, x_n | \theta)$$

We assume that data  
points are I.I.D.



# The Likelihood Function

**Definition:** The *joint* probability of our observed data if our parameters were  $\theta$ .

For a list of observations,  $[x_1, x_2, \dots, x_n]$ :

$$L(\theta) = P(x_1, x_2, \dots, x_n | \theta)$$

We assume that data  
points are I.I.D.



$$= \prod_{i=1}^n P(x_i | \theta)$$

# The Likelihood Function

**Definition:** The *joint* probability of our observed data if our parameters were  $\theta$ .

For a list of observations,  $[x_1, x_2, \dots, x_n]$ :

$$L(\theta) = P(x_1, x_2, \dots, x_n | \theta)$$

We assume that data points are I.I.D.



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

We always use  $f$  for likelihood in MLE (even for discrete) 😊

# The Likelihood Function

$n$  I.I.D. data points  $x_1, x_2, \dots, x_n$



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

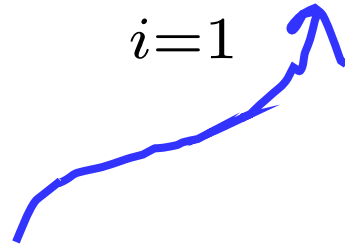
This is just a product since  $X_i$  are I.I.D.

We explicitly specify parameter  $\theta$  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

# How to Choose the “Best” Parameters: MLE



Likelihood

We want to choose the parameter value that maximizes the probability of the data:

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

# How to Choose the “Best” Parameters: MLE



Likelihood

We want to choose the parameter value that maximizes the probability of the data:

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

To put words into math:

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

Our best estimate

# How to Choose the “Best” Parameters: MLE



We want to choose the parameter value that maximizes the probability of the data:

Likelihood

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

Log Likelihood

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

To put words into math:

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

↑  
Our best estimate

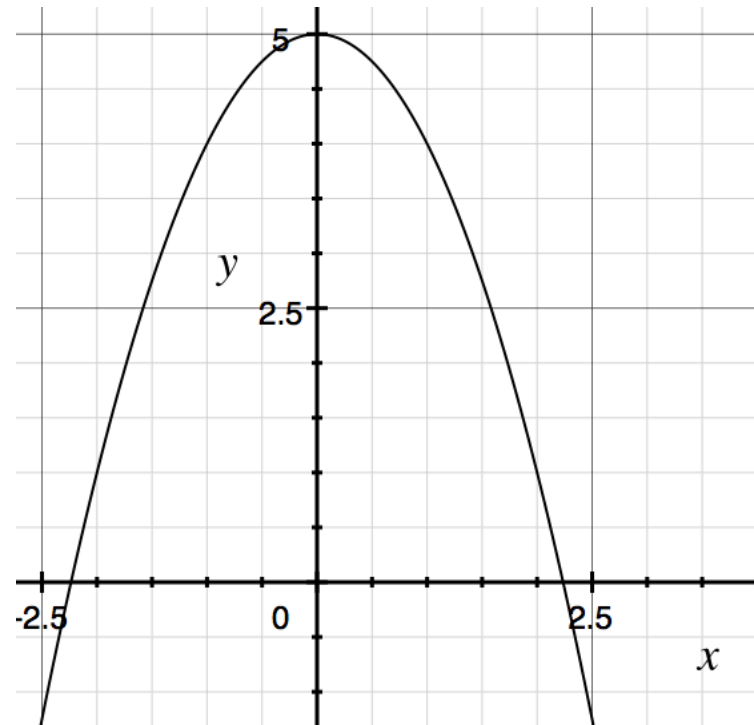
Sidequest: Review  $\operatorname{argmax}$

# Argmax

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

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

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



A close-up of Scar from Disney's The Lion King. He has a dark orange face, a black mane, and a single glowing green eye. A white rectangular text box is overlaid on his face, containing the text "arg max". The background is a dark blue, textured surface.

arg max

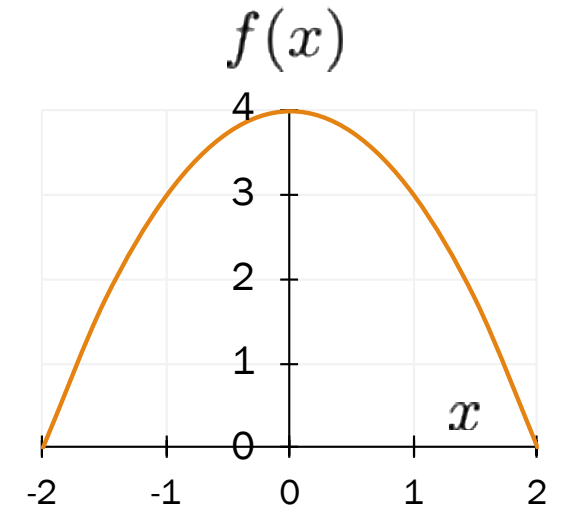
But how do we compute  $\operatorname{argmax}$ ?

Option #1: Straight optimization

# Finding the Argmax with Calculus

$$f(x) = -x^2 + 4$$

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



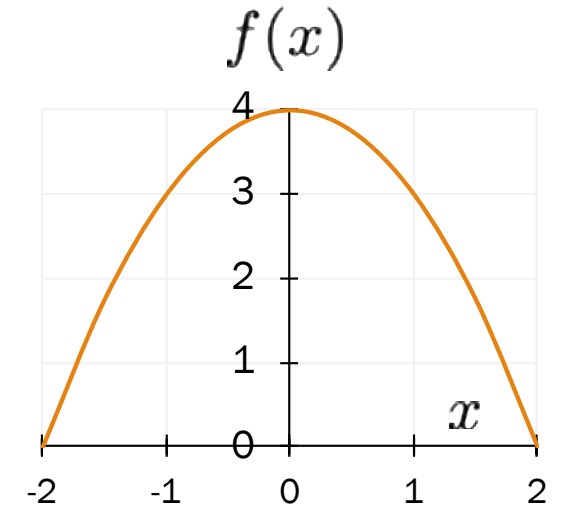
# Finding the Argmax with Calculus

$$f(x) = -x^2 + 4$$

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

Differentiate w.r.t.  
argmax's argument

$$\frac{\partial}{\partial x} f(x) = \frac{\partial}{\partial x} [-x^2 + 4] = -2x$$



# Finding the Argmax with Calculus

$$f(x) = -x^2 + 4$$

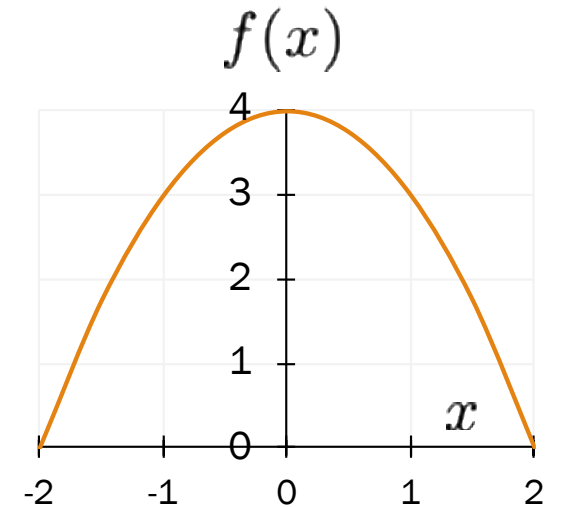
$$\hat{x} = \arg \max_x f(x)$$

Differentiate w.r.t.  
argmax's argument

$$\frac{\partial}{\partial x} f(x) = \frac{\partial}{\partial x} [-x^2 + 4] = -2x$$

Set to 0 and solve

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



# Finding the Argmax with Calculus

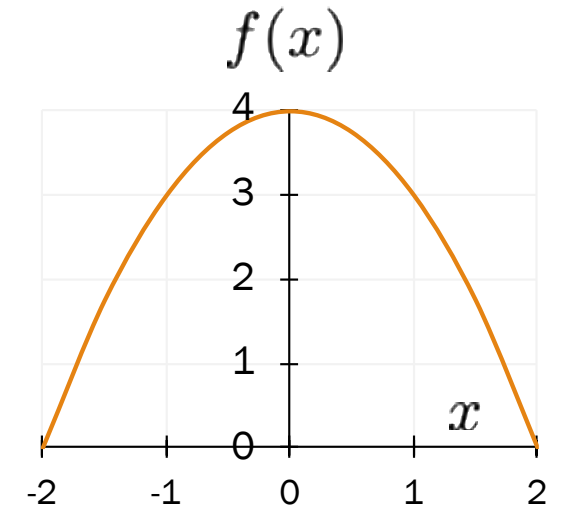
$$f(x) = -x^2 + 4 \qquad \hat{x} = \arg \max_x f(x)$$

Differentiate w.r.t.  
argmax's argument

$$\frac{\partial}{\partial x} f(x) = \frac{\partial}{\partial x} [-x^2 + 4] = -2x$$

Set to 0 and solve

$$-2\hat{x} = 0 \qquad \Rightarrow \qquad \hat{x} = 0$$

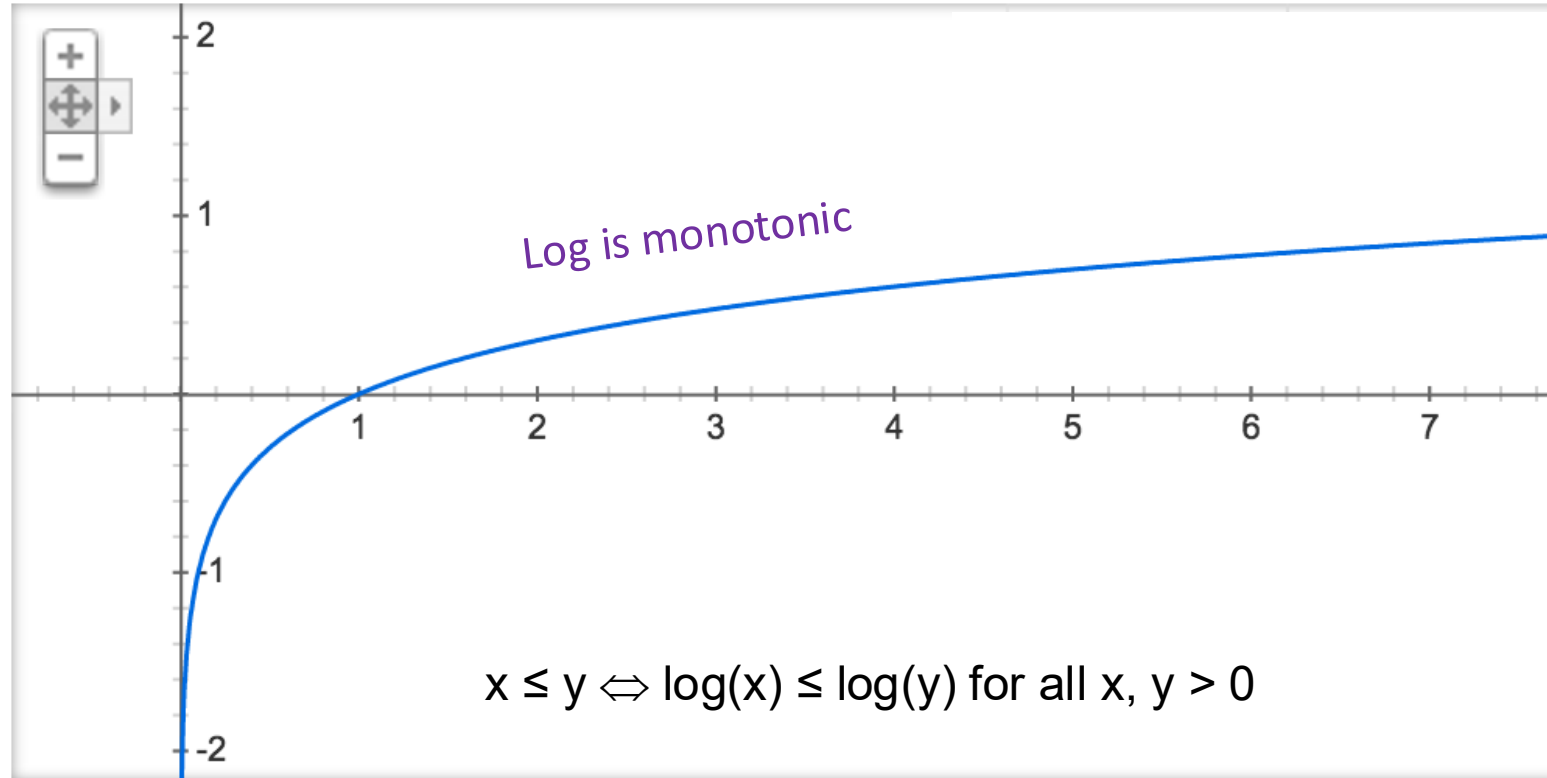


Make sure  $\hat{x}$   
is a maximum

- Check the second derivative
- Generally ignored in expository derivations
- arg min is defined similarly, relevant for gradient descent

# 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

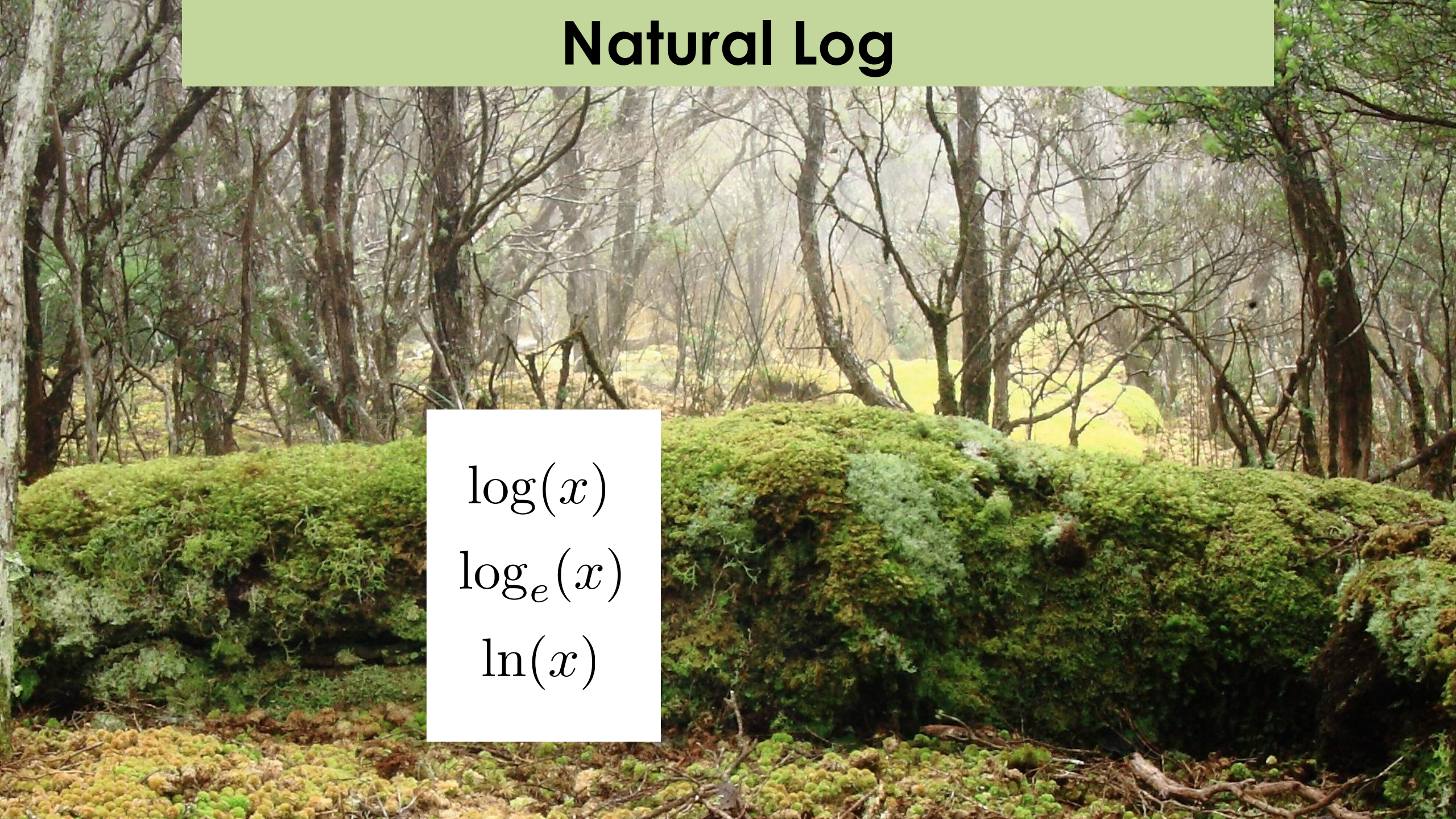
$$\log(ab) = \log(a) + \log(b)$$

# Natural Log

$\log(x)$

$\log_e(x)$

$\ln(x)$



End Sidequest



# MLE For Poisson

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

# MLE For Poisson

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

We observed the following samples:

[6, 1, 2, 1, 2, 3, 3, 2, 1, 3, 1, 3]

$x_i$

What is lambda, ?  $\lambda$

# MLE For Poisson

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

$X_i \sim \text{Poi}(\lambda)$     **Use Maximum Likelihood to estimate  $\lambda$**

1. What is the likelihood of one  $X_i$

2. What is the likelihood of all the *data*

3. What is the log-likelihood all the *data*

4. Find the value of  $\lambda$  which maximizes log likelihood

# MLE For Poisson

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

$X_i \sim \text{Poi}(\lambda)$     **Use Maximum Likelihood to estimate  $\lambda$**

- Probability mass function can be written as:  $f(x_i|\lambda) = \frac{e^{-\lambda} \lambda^{x_i}}{x_i!}$

2. What is the likelihood of all the *data*

3. What is the log-likelihood all the *data*

4. Find the value of  $\lambda$  which maximizes log likelihood

# MLE For Poisson

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

$X_i \sim \text{Poi}(\lambda)$     **Use Maximum Likelihood to estimate  $\lambda$**

- Probability mass function can be written as:  $f(x_i|\lambda) = \frac{e^{-\lambda} \lambda^{x_i}}{x_i!}$
- Likelihood:  $L(\lambda) = f(x_1 \dots x_n|\lambda) = \prod_{i=1}^n f(x_i|\lambda) = \prod_{i=1}^n \frac{e^{-\lambda} \lambda^{x_i}}{x_i!}$
- Log-likelihood:

3. What is the log-likelihood all the *data*

4. Find the value of  $\lambda$  which maximizes log likelihood

# MLE For Poisson

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

$X_i \sim \text{Poi}(\lambda)$     **Use Maximum Likelihood to estimate  $\lambda$**

- Probability mass function can be written as:  $f(x_i|\lambda) = \frac{e^{-\lambda} \lambda^{x_i}}{x_i!}$

- Likelihood:  $L(\lambda) = f(x_1 \dots x_n|\lambda) = \prod_{i=1}^n f(x_i|\lambda) = \prod_{i=1}^n \frac{e^{-\lambda} \lambda^{x_i}}{x_i!}$

- Log-likelihood:

$$LL(\lambda) = \log \prod_{i=1}^n \frac{e^{-\lambda} \lambda^{x_i}}{x_i!} = \sum_{i=1}^n \log \frac{e^{-\lambda} \lambda^{x_i}}{x_i!} = \sum_{i=1}^n -\lambda + x_i \log \lambda - \log x_i!$$

- Differentiate w.r.t  $\lambda$  and set to 0

4. Find the value of  $\lambda$  which maximizes log likelihood

# MLE For Poisson

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

$X_i \sim \text{Poi}(\lambda)$     **Use Maximum Likelihood to estimate  $\lambda$**

- Probability mass function can be written as:  $f(x_i|\lambda) = \frac{e^{-\lambda} \lambda^{x_i}}{x_i!}$
- Likelihood:  $L(\lambda) = f(x_1 \dots x_n|\lambda) = \prod_{i=1}^n f(x_i|\lambda) = \prod_{i=1}^n \frac{e^{-\lambda} \lambda^{x_i}}{x_i!}$
- Log-likelihood:

$$LL(\lambda) = \log \prod_{i=1}^n \frac{e^{-\lambda} \lambda^{x_i}}{x_i!} = \sum_{i=1}^n \log \frac{e^{-\lambda} \lambda^{x_i}}{x_i!} = \sum_{i=1}^n -\lambda + x_i \log \lambda - \log x_i!$$

- Differentiate w.r.t.  $\lambda$ , and set to 0:

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

$$\lambda = \frac{1}{n} \sum_{i=1}^n x_i$$

Isn't that the same as  
the sample mean?

Yes. For Poisson.

# MLE For Poisson



# MLE For Pareto

```
observations = [1.677, 3.812, 1.463, 2.641, 1.256, 1.678, 1.157,  
1.146, 1.323, 1.029, 1.238, 1.018, 1.171, 1.123, 1.074, 1.652,  
1.873, 1.314, 1.309, 3.325, 1.045, 2.271, 1.305, 1.277, 1.114,  
1.391, 3.728, 1.405, 1.054, 2.789, 1.019, 1.218, 1.033, 1.362,  
1.058, 2.037, 1.171, 1.457, 1.518, 1.117, 1.153, 2.257, 1.022,  
1.839, 1.706, 1.139, 1.501, 1.238, 2.53, 1.414, 1.064, 1.097,  
1.261, 1.784, 1.196, 1.169, 2.101, 1.132, 1.193, 1.239, 1.518,  
2.764, 1.053, 1.267, 1.015, 1.789, 1.099, 1.25, 1.253, 1.418,  
1.494, 1.015, 1.459, 2.175, 2.044, 1.551, 4.095, 1.396, 1.262,  
1.351, 1.121, 1.196, 1.391, 1.305, 1.141, 1.157, 1.155, 1.103,  
1.048, 1.918, 1.889, 1.068, 1.811, 1.198, 1.361, 1.261, 4.093,  
2.925, 1.133, 1.573]
```

```
def estimate_alpha(observations):  
    print('your code here')
```



We know sand is distributed as a pareto with PDF

$$f(x) = \frac{\alpha}{x^{\alpha+1}}$$

# MLE for a Pareto

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

- $X_i \sim \text{Pareto}(\alpha)$ . **Use Maximum Likelihood to estimate  $\alpha$ .**

1. What is the likelihood of all the *data*

2. What is the log-likelihood all the *data*

3. Find the value of  $\alpha$  which maximizes log likelihood

# MLE for a Pareto

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

- $X_i \sim \text{Pareto}(\alpha)$ . **Use Maximum Likelihood to estimate  $\alpha$ .**
- Likelihood:

$$L(\alpha) = \prod_{i=1}^n \frac{\alpha}{x_i^{\alpha+1}}$$

2. What is the log-likelihood all the *data*

3. Find the value of  $\alpha$  which maximizes log likelihood

# MLE for a Pareto

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

- $X_i \sim \text{Pareto}(\alpha)$ . **Use Maximum Likelihood to estimate  $\alpha$ .**

- Likelihood:

$$L(\alpha) = \prod_{i=1}^n \frac{\alpha}{x_i^{\alpha+1}}$$

- Log-likelihood:

$$LL(\alpha) = \sum_{i=1}^n \log \alpha - (\alpha + 1) \log x_i = n \log \alpha - (\alpha + 1) \sum_{i=1}^n \log x_i$$

3. Find the value of  $\alpha$  which maximizes log likelihood

# MLE for a Pareto

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

- $X_i \sim \text{Pareto}(\alpha)$ . **Use Maximum Likelihood to estimate  $\alpha$ .**

- Likelihood:

$$L(\alpha) = \prod_{i=1}^n \frac{\alpha}{x_i^{\alpha+1}}$$

- Log-likelihood:

$$LL(\alpha) = \sum_{i=1}^n \log \alpha - (\alpha + 1) \log x_i = n \log \alpha - (\alpha + 1) \sum_{i=1}^n \log x_i$$

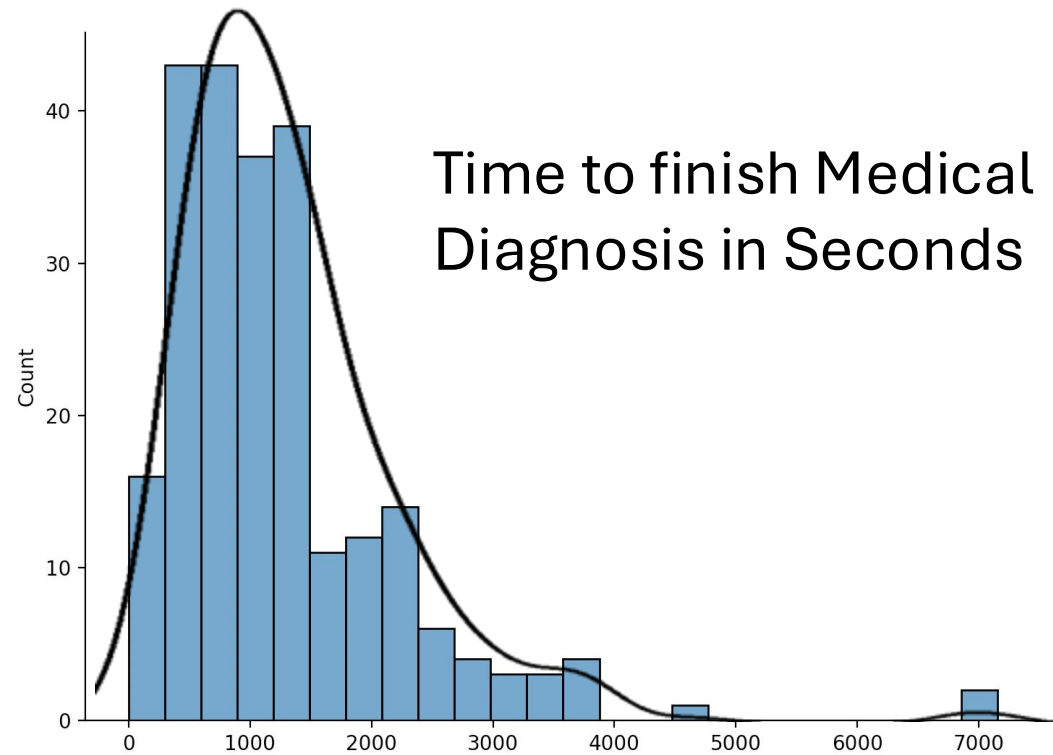
- Chose  $\alpha$  to be the argmax of LL:

$$\frac{\partial LL(\alpha)}{\partial \alpha} = \frac{n}{\alpha} - \sum_{i=1}^n \log x_i$$

Pause

# MLE for Erlang

```
[3.002, 0.983, 2.186, 1.624, 3.997, 1.777,  
2.809, 0.42, 0.515, 1.582, 0.948, 0.458, 1.  
066, 0.8, 2.398, 0.794, 2.561, 2.61, 0.  
595, 3.897, 1.852, 1.182, 3.043, 0.905, 1.  
45, 0.405, 0.445, 2.103, 1.425, 3.12, 0.  
973, 1.056, 3.715, 2.952, 1.817, 2.686, 4.  
173, 0.358, 2.185, 2.581, 7.134, 0.206, 2.  
049, 0.896, 2.095, 4.39, 2.199, 3.434, 5.  
696, 0.819, 0.416, 1.571, 1.337, 2.79, 2.  
701, 3.061, 4.677, 0.671, 1.594, 3.586, 2.  
708, 1.417, 1.799, 1.137, 1.771, 2.12, 0.  
93, 6.835, 3.213, 2.541, 2.505, 1.257, 1.  
99, 1.5, 0.014, 3.856, 0.979, 2.413, 2.  
596, 1.653, 0.881, 4.457, 0.717, 3.305, 2.  
456, 3.462, 1.737, 0.968, 0.528, 0.18, 1.  
626, 2.224, 1.466, 1.6, 1.572, 0.12, 2.86,  
1.062, 2.139, 1.217]
```

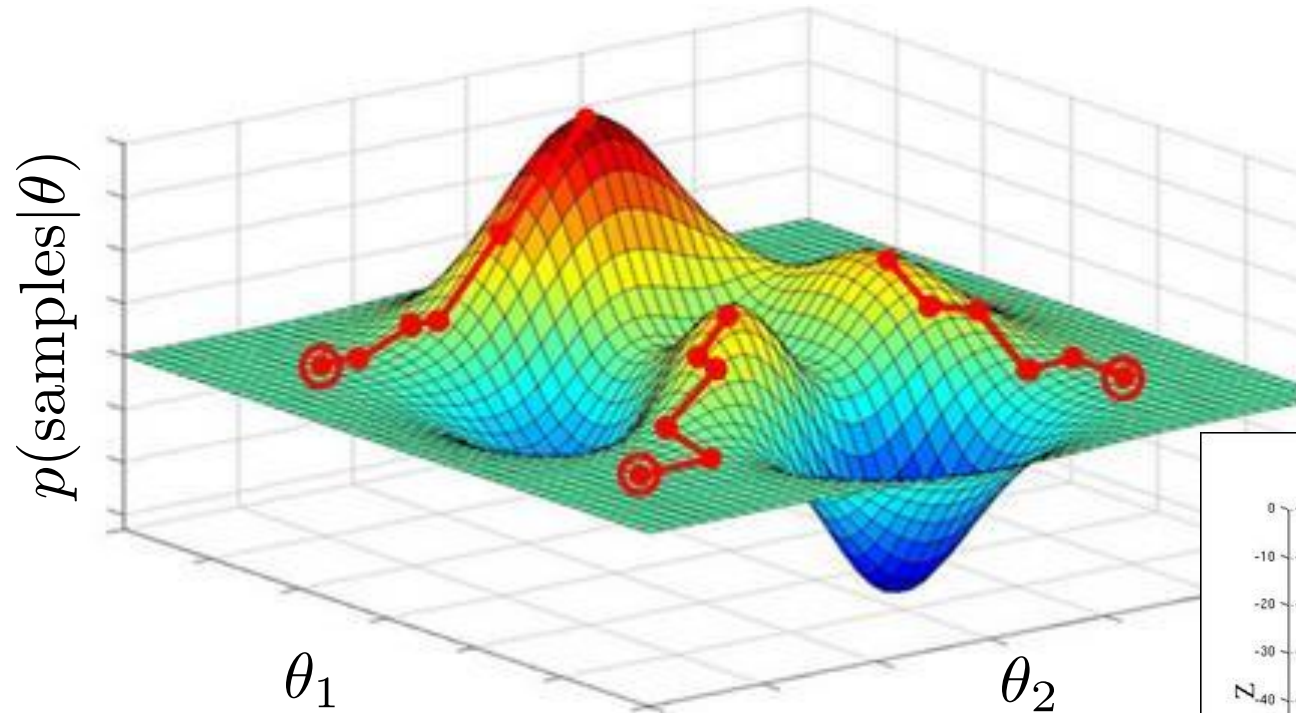


$$f(x) = \frac{\lambda^k x^{k-1} e^{-\lambda x}}{(k-1)!} = \frac{\lambda^k x^{k-1} e^{-\lambda x}}{\Gamma(k)}$$

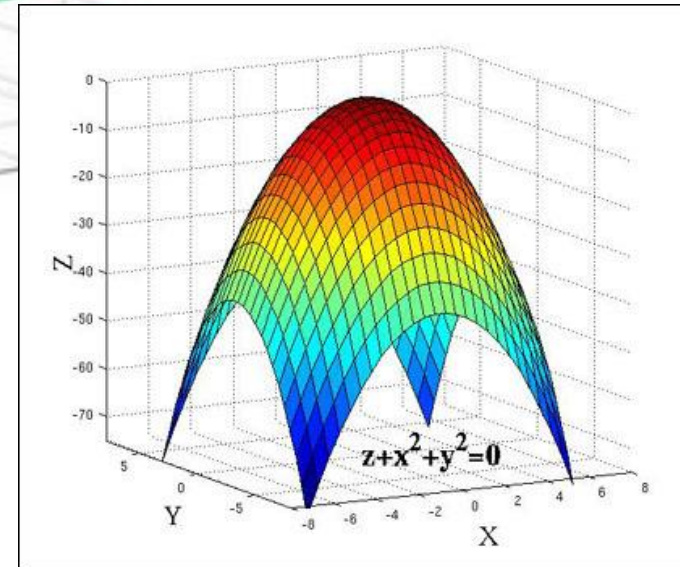
A close-up of Scar from Disney's The Lion King. He has a dark orange face, a black mane, and a single glowing green eye. A white rectangular text box with the text "arg max" is overlaid on his face, partially covering his eyes and nose. The background is a dark blue, textured surface.

arg max

# Gradient Ascent



Especially good if function is convex



Walk uphill and you will find a local maxima  
(if your step size is small enough)

# Gradient Ascent

Repeat many times

$$\theta_j^{\text{new}} = \theta_j^{\text{old}} + \eta \cdot \frac{\partial LL(\theta^{\text{old}})}{\partial \theta_j^{\text{old}}}$$



Step size constant

This is some **profound** life philosophy

Walk uphill and you will find a local maxima  
(if your step size is small enough)

# Gradient Ascent

Initialize:  $\theta_j = \text{random}$  for all  $0 \leq j \leq m$

Repeat many times:

*Calculate all gradient[j]'s based on data*

$\theta_j += \eta * \text{gradient}[j]$  for all  $0 \leq j \leq m$

To the code!

# Gradient Ascent for MLE of Erlang

```
def fit_erlang(observations, n_iterations=10, initial_step_size=0.0001):
    step_size = initial_step_size
    k = 1.0 # Initial guess for shape parameter
    lambda_ = 2.0 # Initial guess for rate parameter
    n = len(observations)

    for i in range(n_iterations):
        # To debug: Calculate log-likelihood
        # ll = calc_log_likelihood_erlang(observations, k, lambda_)
        # print(f"Log-Likelihood at iteration {k, lambda_}: {ll}")

        # Calculate gradients
        # gradient_lambda = (n * k / lambda_) - sum(observations)
        gradient_lambda = 0
        for x_i in observations:
            gradient_lambda += k / lambda_ - x_i
        gradient_k = n * math.log(lambda_) + sum(math.log(x) for x in observations) - n * digamma(k)

        # Update parameters
        lambda_ += step_size * gradient_lambda
        k += step_size * gradient_k

    return k, lambda_
```

Derived these partial derivatives of log likelihood

Pause

# MLE For Bernoulli

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

Don't we already have the Beta?

Yes! But this example is critical for developing  
towards deep learning.

# MLE For Bernoulli

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

$X_i \sim \text{Bern}(p)$       **Use Maximum Likelihood to estimate  $p$**

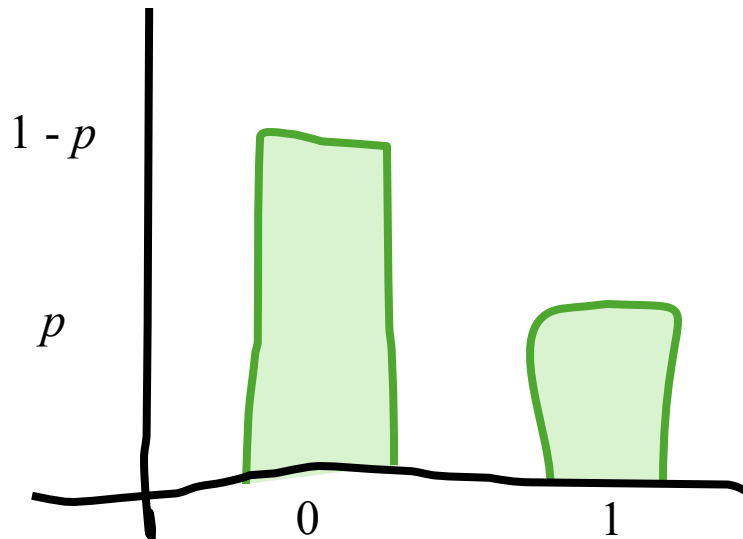
- Probability mass function can be written as: 
$$f(x_i|p) = \begin{cases} p & \text{if } x_i = 1 \\ 1 - p & \text{if } x_i = 0 \end{cases}$$



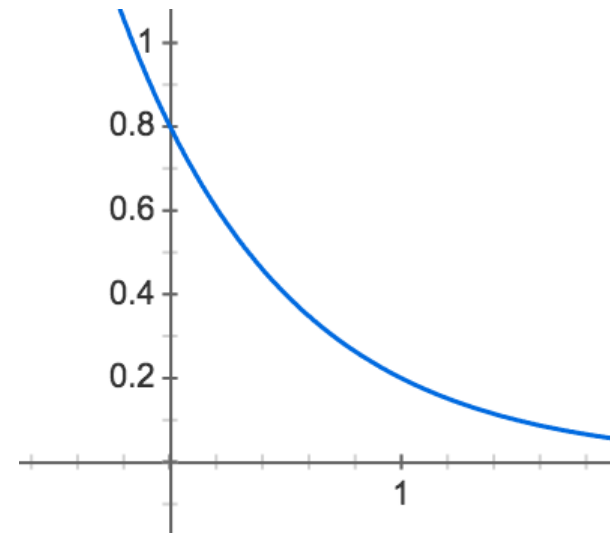
# Differentiable PMF for Bernoulli

- Consider I.I.D. random variables  $X_1, X_2, \dots, X_n$ 
  - $X_i \sim \text{Ber}(p)$
  - Probability mass function  $f(X_i = x_i | P = p)$

PMF of Bernoulli



PMF of Bernoulli ( $p = 0.2$ )



$$f(x_i | p) = p^{x_i} (1 - p)^{1 - x_i}$$
$$f(x_i | p = 0.2) = 0.2^{x_i} (1 - 0.2)^{1 - x_i}$$

# Bernoulli PMF

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



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

# Maximum Likelihood For Bernoulli

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

$$X_i \sim \text{Bern}(p)$$

Use Maximum Likelihood to estimate  $p$

1. What is the likelihood of one  $X_i$

2. What is the likelihood of all the *data*

3. What is the log-likelihood all the *data*

4. Find the value of  $p$  which maximizes log likelihood

# Maximum Likelihood For Bernoulli

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

$X_i \sim \text{Bern}(p)$       **Use Maximum Likelihood to estimate  $p$**

- Probability mass function can be written as:  $f(x_i|p) = p^{x_i}(1-p)^{1-x_i}$

2. What is the likelihood of all the *data*

3. What is the log-likelihood all the *data*

4. Find the value of  $p$  which maximizes log likelihood

# Maximum Likelihood For Bernoulli

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

$X_i \sim \text{Bern}(p)$       **Use Maximum Likelihood to estimate  $p$**

- Probability mass function can be written as:  $f(x_i|p) = p^{x_i}(1-p)^{1-x_i}$

- Likelihood: 
$$L(p) = \prod_{i=1}^n f(x_i|p) = \prod_{i=1}^n p^{x_i}(1-p)^{1-x_i}$$

- Log-likelihood:

3. What is the log-likelihood all the *data*

4. Find the value of  $p$  which maximizes log likelihood

# Maximum Likelihood For Bernoulli

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

$X_i \sim \text{Bern}(p)$       **Use Maximum Likelihood to estimate  $p$**

- Probability mass function can be written as:  $f(x_i|p) = p^{x_i}(1-p)^{1-x_i}$

- Likelihood: 
$$L(p) = \prod_{i=1}^n f(x_i|p) = \prod_{i=1}^n p^{x_i}(1-p)^{1-x_i}$$

- Log-likelihood:

$$LL(p) = \sum_{i=1}^n x_i \log p + (1 - x_i) \log(1 - p)$$

4. Find the value of  $p$  which maximizes log likelihood

Take Derivative:  $LL(p) = \sum_{i=1}^n x_i \log p + (1 - x_i) \log(1 - p)$

$$\frac{\partial LL(p)}{\partial p} = \frac{\partial}{\partial p} \sum_{i=1}^n x_i \log p + (1 - x_i) \log(1 - p)$$

Take the derivative wrt p

$$= \sum_{i=1}^n \frac{\partial}{\partial p} \left[ x_i \log p + (1 - x_i) \log(1 - p) \right]$$

Derivative of a sum!

$$= \sum_{i=1}^n \left[ \frac{\partial}{\partial p} x_i \log p \right] + \frac{\partial}{\partial p} (1 - x_i) \log(1 - p)$$

Derivative of a sum!

$$= \sum_{i=1}^n \frac{x_i}{p} + \frac{\partial}{\partial p} (1 - x_i) \log(1 - p)$$

Derivative of log p

$$= \sum_{i=1}^n \frac{x_i}{p} - \frac{1 - x_i}{1 - p}$$

Derivative of log (1-p)

Set to Zero:  $\frac{\partial LL(p)}{\partial p} = \sum_{i=1}^n \frac{x_i}{p} - \frac{1-x_i}{1-p}$

$$\begin{aligned} 0 &= \sum_{i=1}^n \frac{x_i}{\hat{p}} - \frac{1-x_i}{1-\hat{p}} \\ &= \sum_{i=1}^n \frac{x_i}{\hat{p}} - \sum_{i=1}^n \frac{1-x_i}{1-\hat{p}} \\ &= \frac{y}{\hat{p}} - \frac{n-y}{1-\hat{p}} \end{aligned}$$

$$\frac{n-y}{1-\hat{p}} = \frac{y}{\hat{p}}$$

$$\hat{p}(n-y) = y(1-\hat{p})$$

$$\hat{p}n - \hat{p}y = y - \hat{p}y$$

$$\hat{p}n = y$$

Let  $\sum_{i=1}^n x_i = y$  To make life easier

And  $\sum_{i=1}^n 1 - x_i = \sum_{i=1}^n 1 - \sum_{i=1}^n x_i = n - y$

$$\begin{aligned} \hat{p} &= \frac{1}{n}y \\ &= \frac{1}{n} \sum_{i=1}^n x_i \end{aligned}$$

Isn't that the same as  
unbiased estimator?

Yes. For Bernoulli.

# MLE for Bernoulli is Sample Mean



# MLE vs. Beta

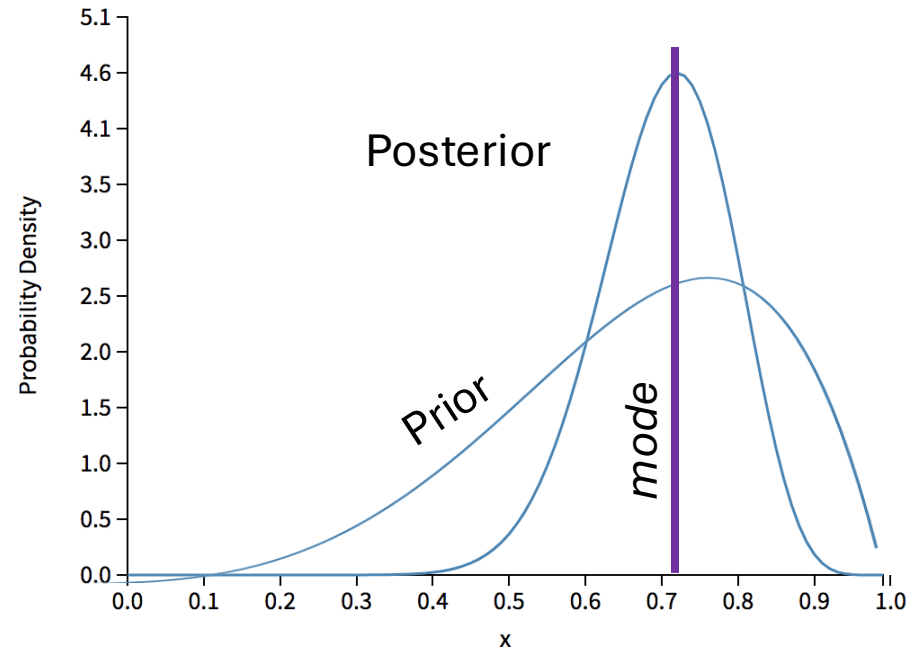
The medicine is tried on 20 patients. It “works” for 14 and “doesn’t work” for 6. What is your new belief that the drug works?

In other words I have 20 IID samples from a Bernoulli. Estimate  $p$ . The data is [1,1,1,1,1,1,1,1,1,1,1,1,1,1,0,0,0,0,0,0]

MLE estimate:

$$p \approx \frac{14}{20} = 0.7$$

Beta estimate:





Think about the difference between a **point estimate** and a **distribution**

$$p = 0.75$$

$$p =$$

