Parameter Estimation

Chris Piech
CS109, Stanford University

Where are we in CS109?

You are here

Counting Theory

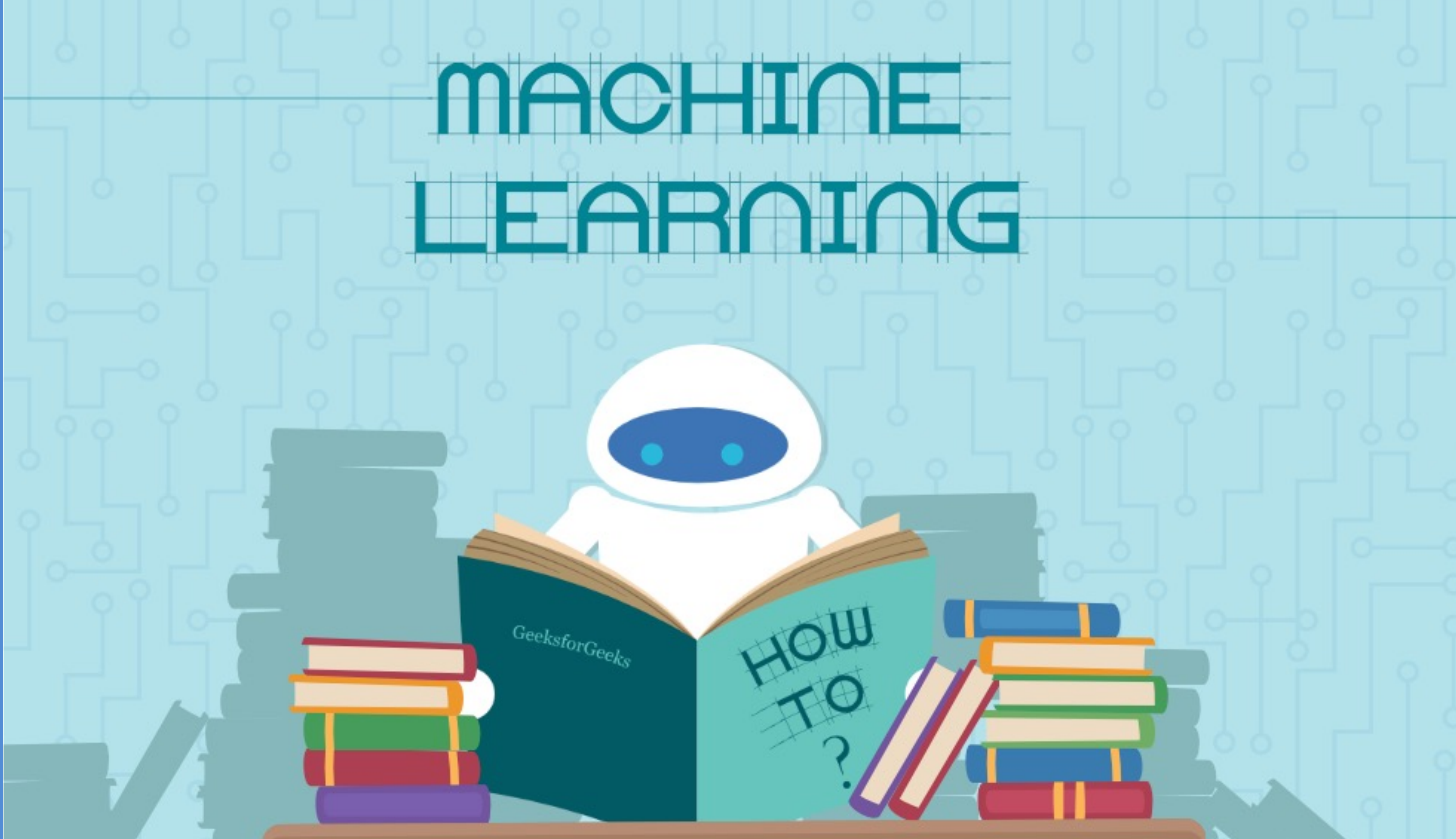
Core Probability

Random Variables

Probabilistic Models

Uncertainty Theory

Machine Learning


MACHINE
LEARNING
GeeksforGeeks
HOW
TO
?

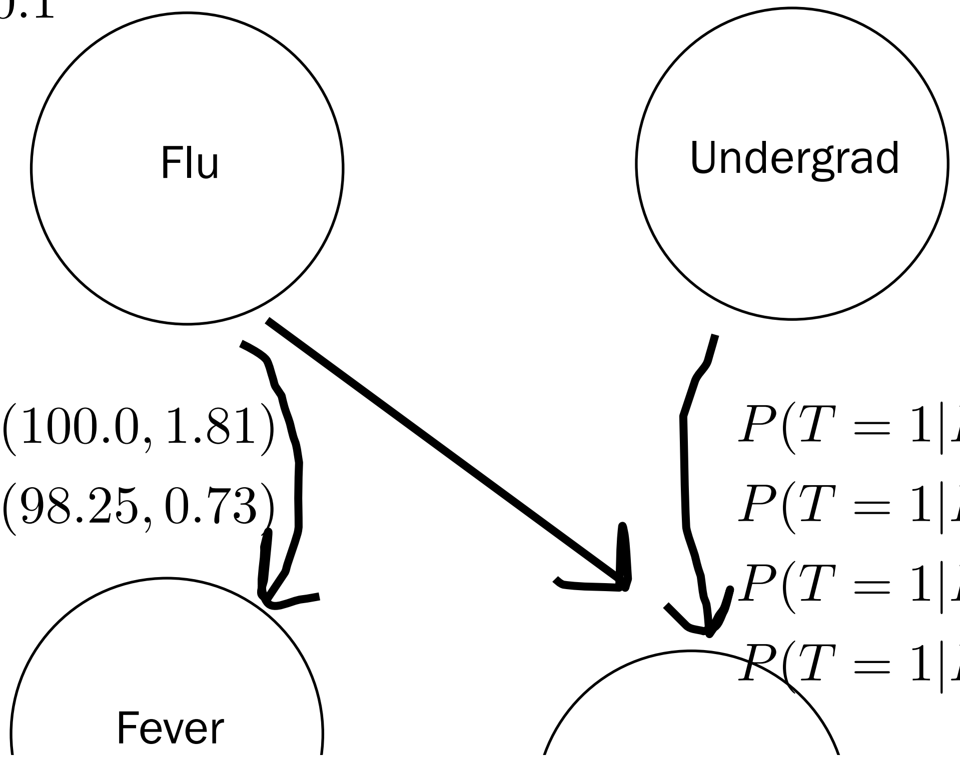
General “Inference”

WebMD HEALTH NOW

Probabilistic Model

$P(Fl = 1) = 0.1$

$P(U = 1) = 0.8$



$Fev | Flu = 0 \sim N(100.0, 1.81)$

$Fev | Flu = 1 \sim N(98.25, 0.73)$

$P(T = 1 | Flu = 0, U = 0) = 0.1$

$P(T = 1 | Flu = 0, U = 1) = 0.8$

$P(T = 1 | Flu = 1, U = 0) = 0.9$

$P(T = 1 | Flu = 1, U = 1) = 1.0$

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?

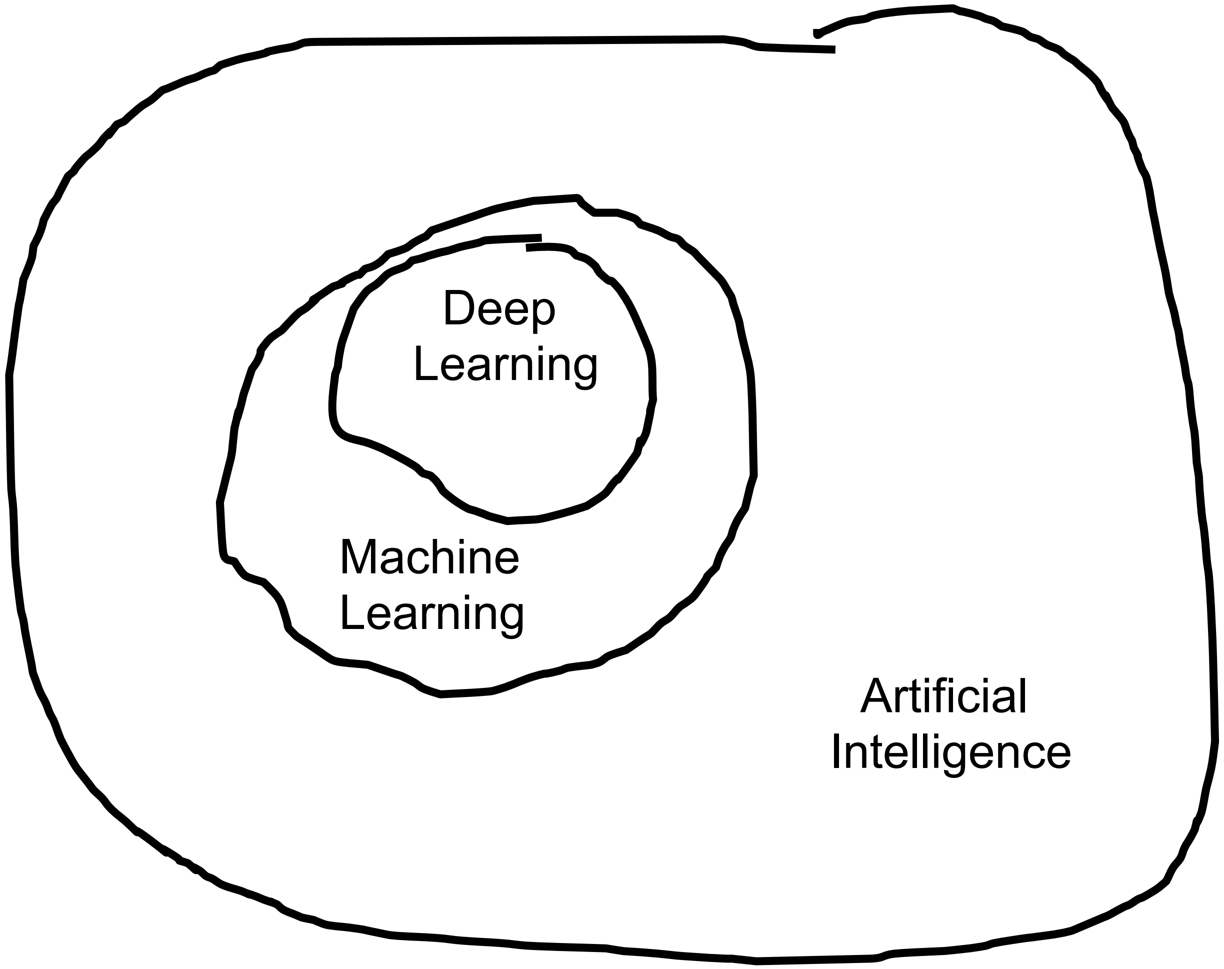
~~Oh can we also learn the *structure* of the model too?~~

I wish. Another day ☺

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

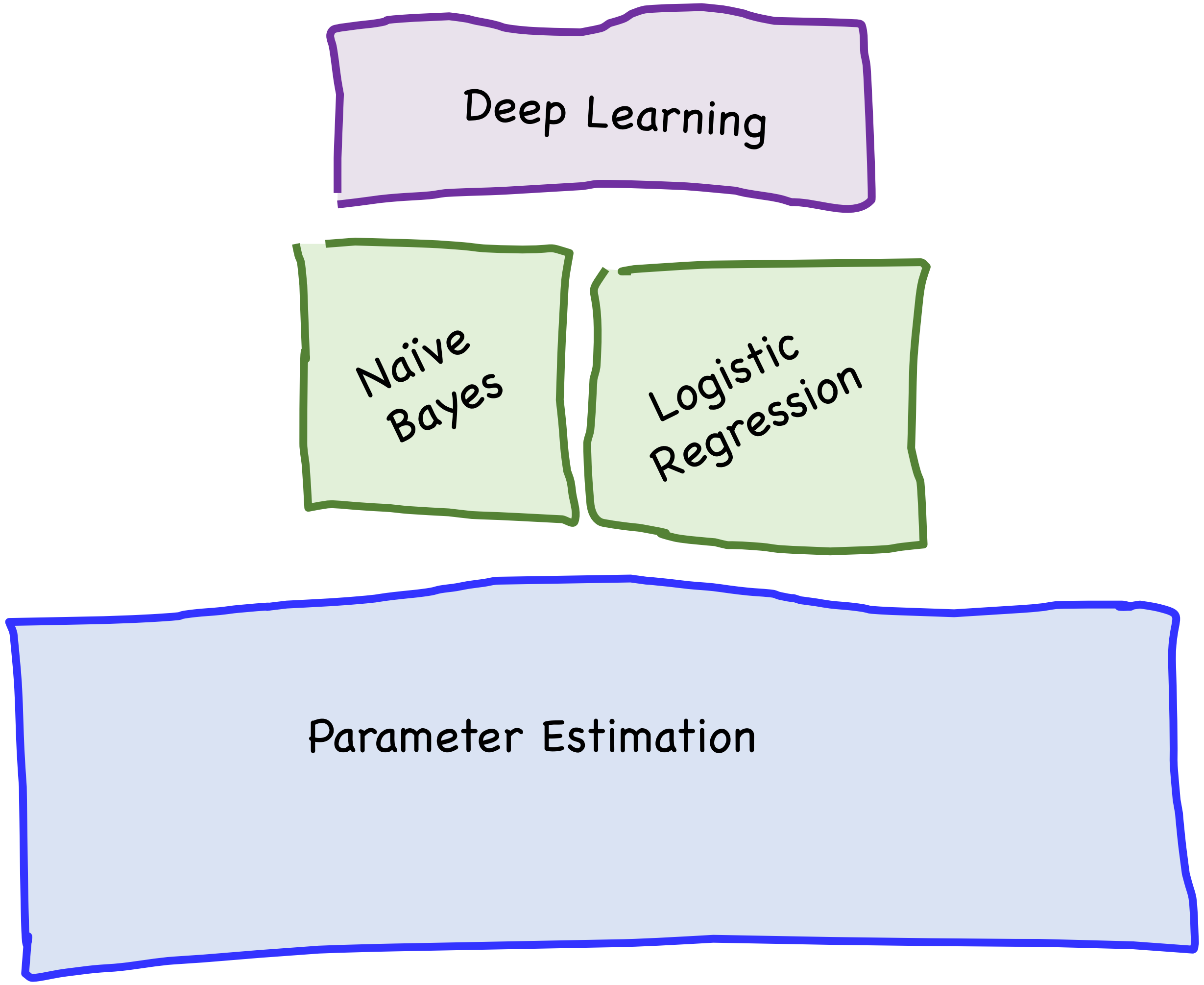
Machine Learning

AI and Machine Learning

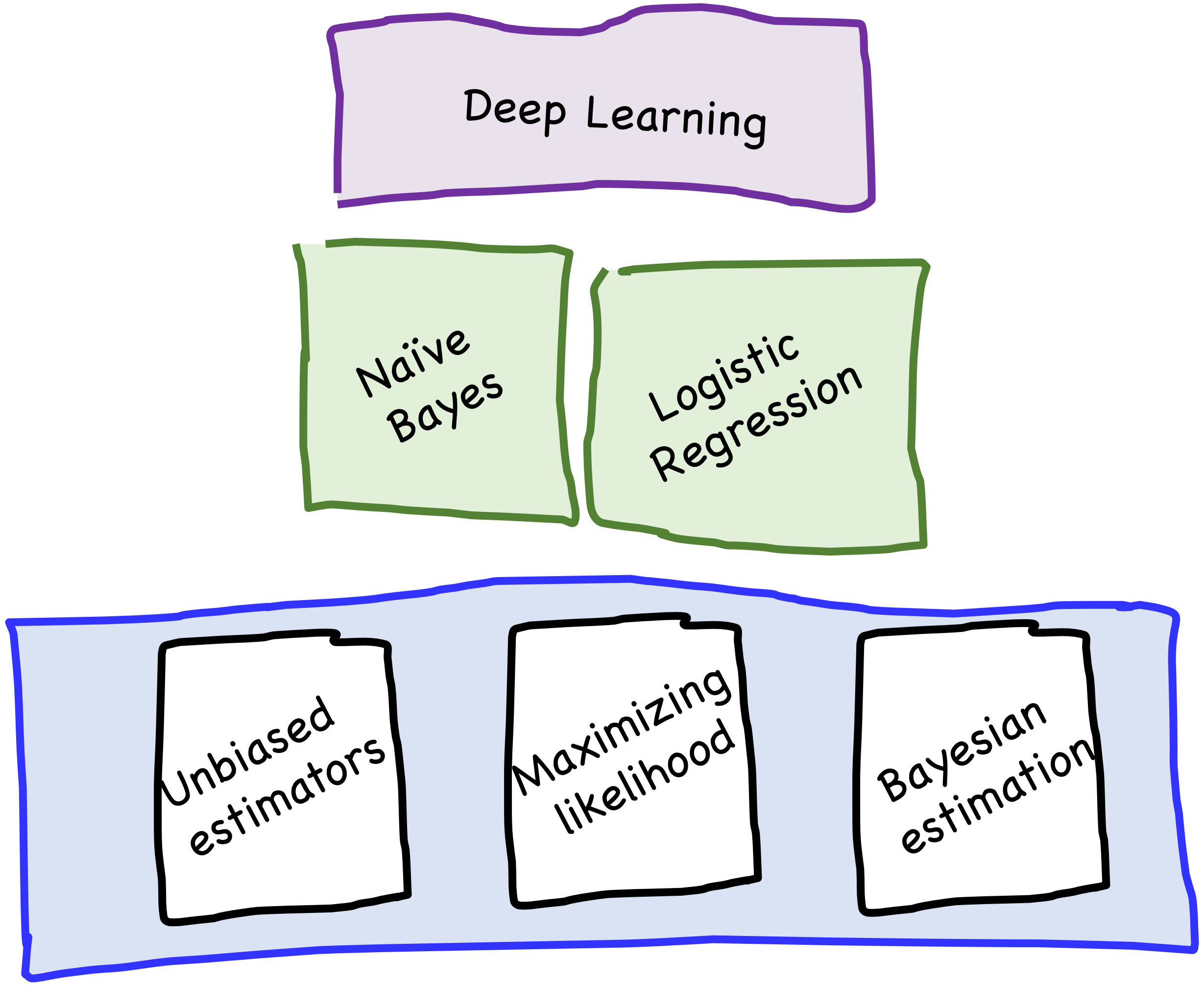


ML: Rooted in probability theory

Our Path

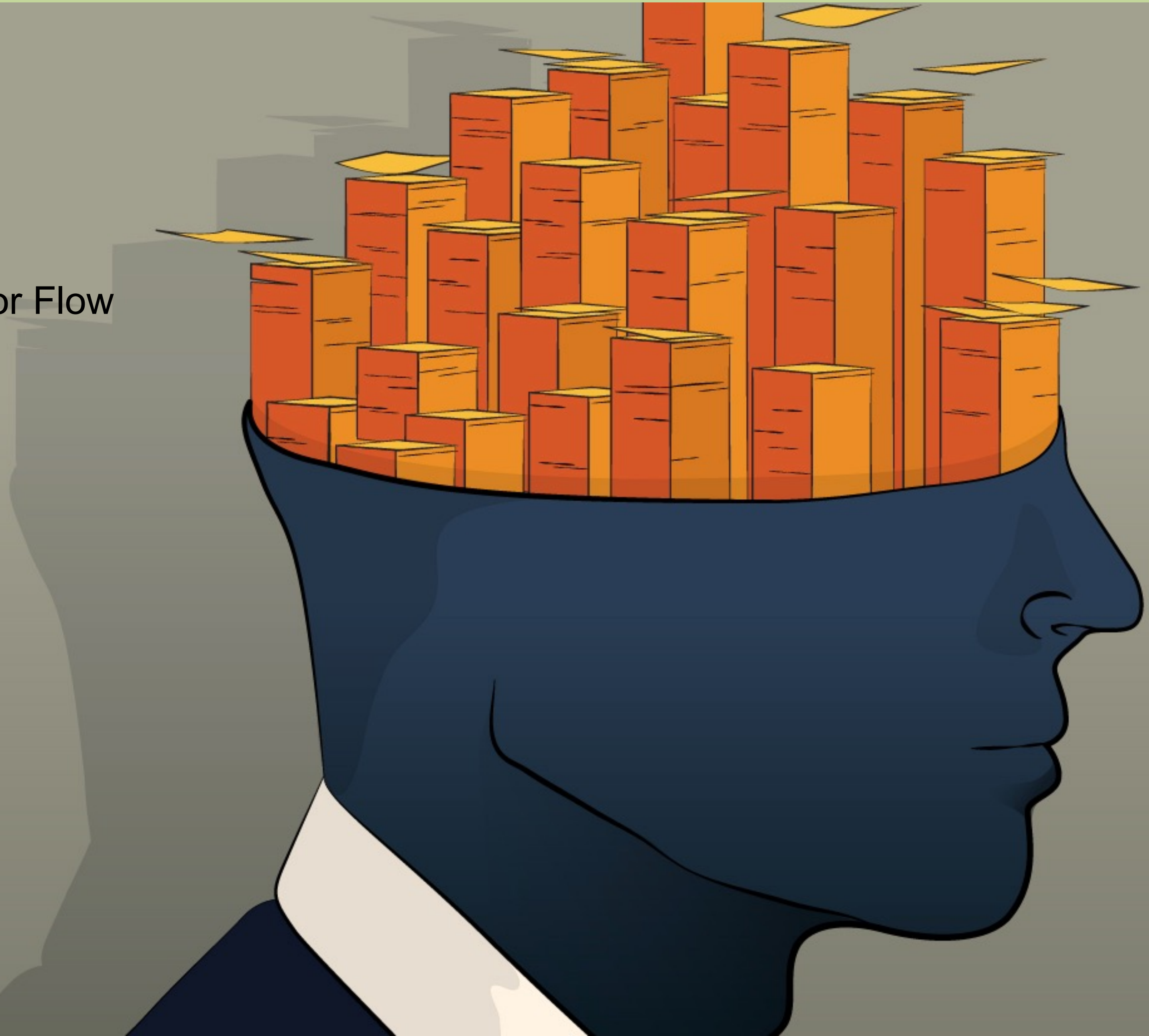

Deep Learning
Naïve Bayes
Logistic Regression
Parameter Estimation

Our Path


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

Jump Straight to Deep Learning?

Tensor Flow

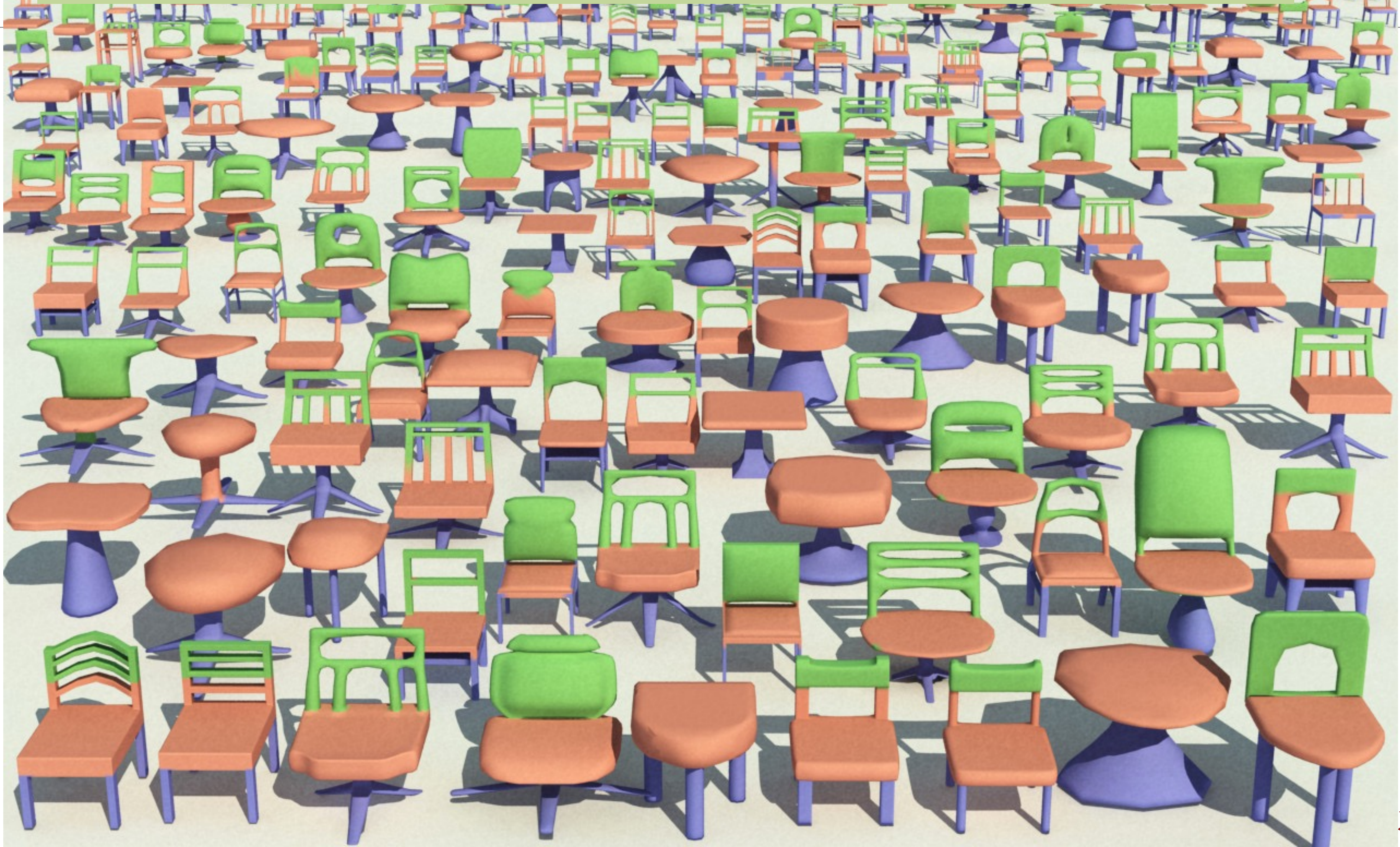


Jump Straight to Deep Learning?

Understand the theory to help you debug

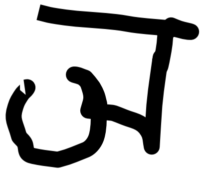
But another reason…

Machine Learning Uses a Lot of Data

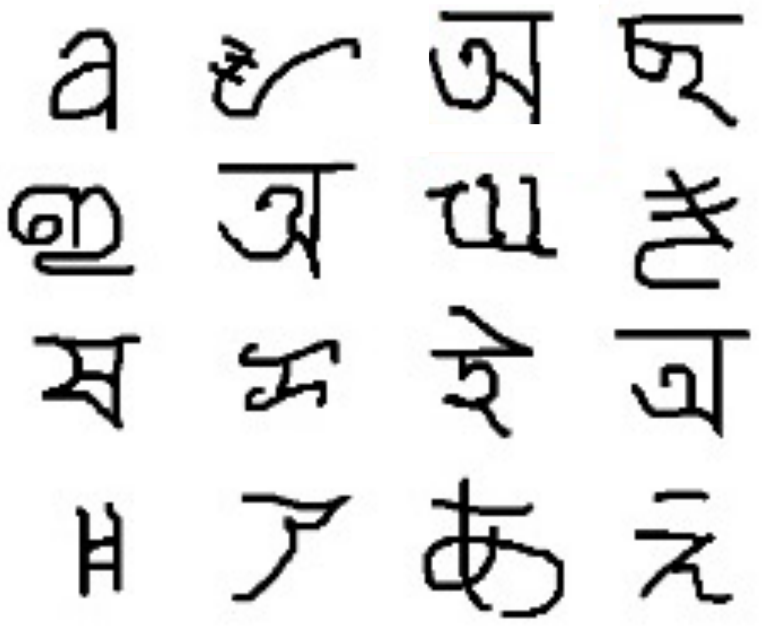


One Shot Learning

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**


Walt Disney
PICTURES

Once upon a time…

…there was parameter estimation

What are Parameters?

Consider some probability distributions:

- Ber(p) — $\theta = p$
- Poi(λ) — $\theta = \lambda$
- Uni(α, β) — $\theta = (\alpha, \beta)$
- Normal(μ, σ^2) — $\theta = (\mu, \sigma^2)$
- $Y = mX + b$ — $\theta = (m, b)$
- etc…

Call these "parametric models"

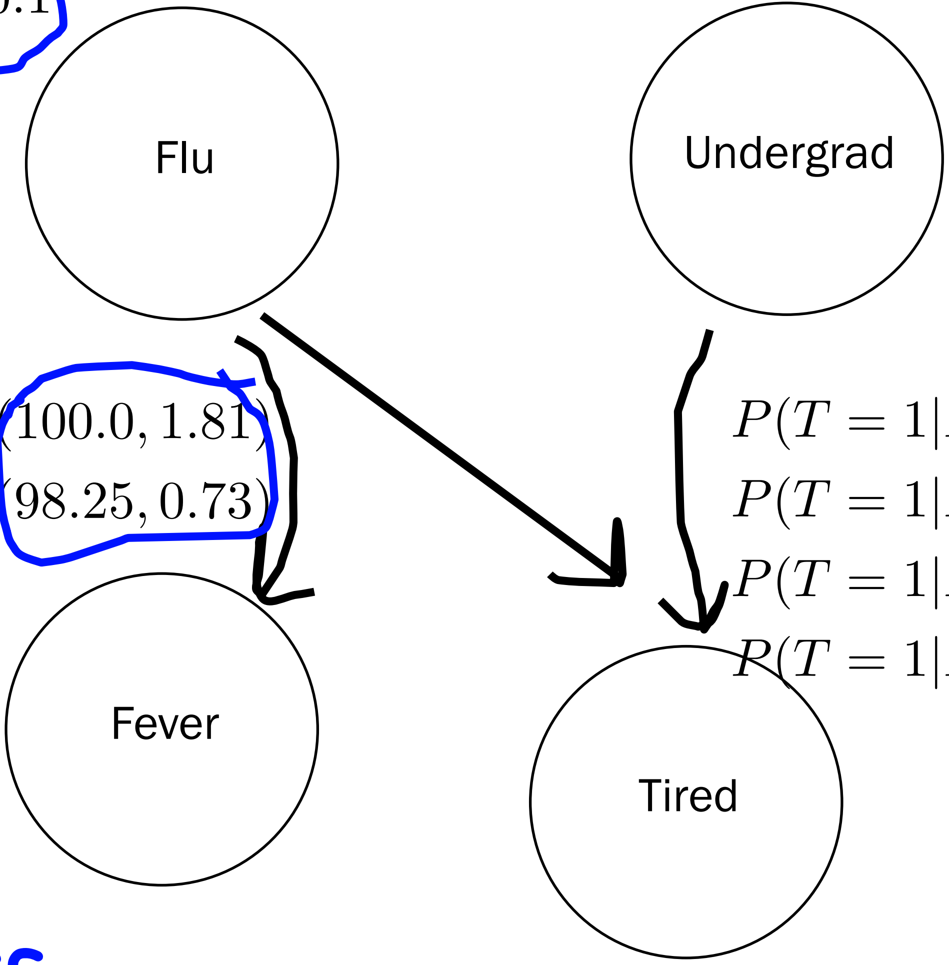
Given model, **parameters** yield actual distribution

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

What are Parameters?

$P(Fl = 1) = 0.1$

$P(U = 1) = 0.8$



$Fev|Flu = 0 \sim N(100.0, 1.81)$

$Fev|Flu = 1 \sim N(98.25, 0.73)$

$P(T = 1|Flu = 0, U = 0) = 0.1$

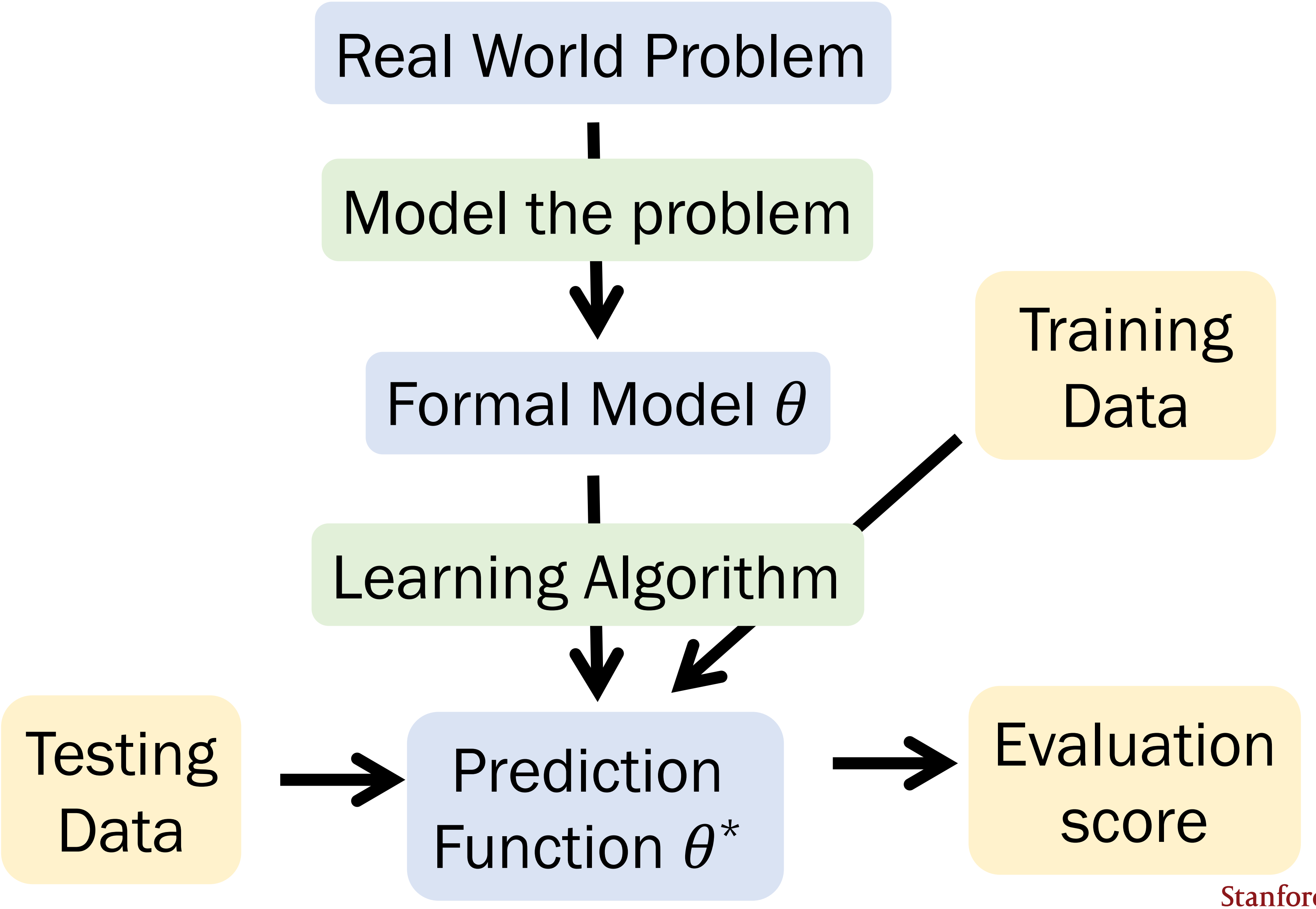
$P(T = 1|Flu = 0, U = 1) = 0.8$

$P(T = 1|Flu = 1, U = 0) = 0.9$

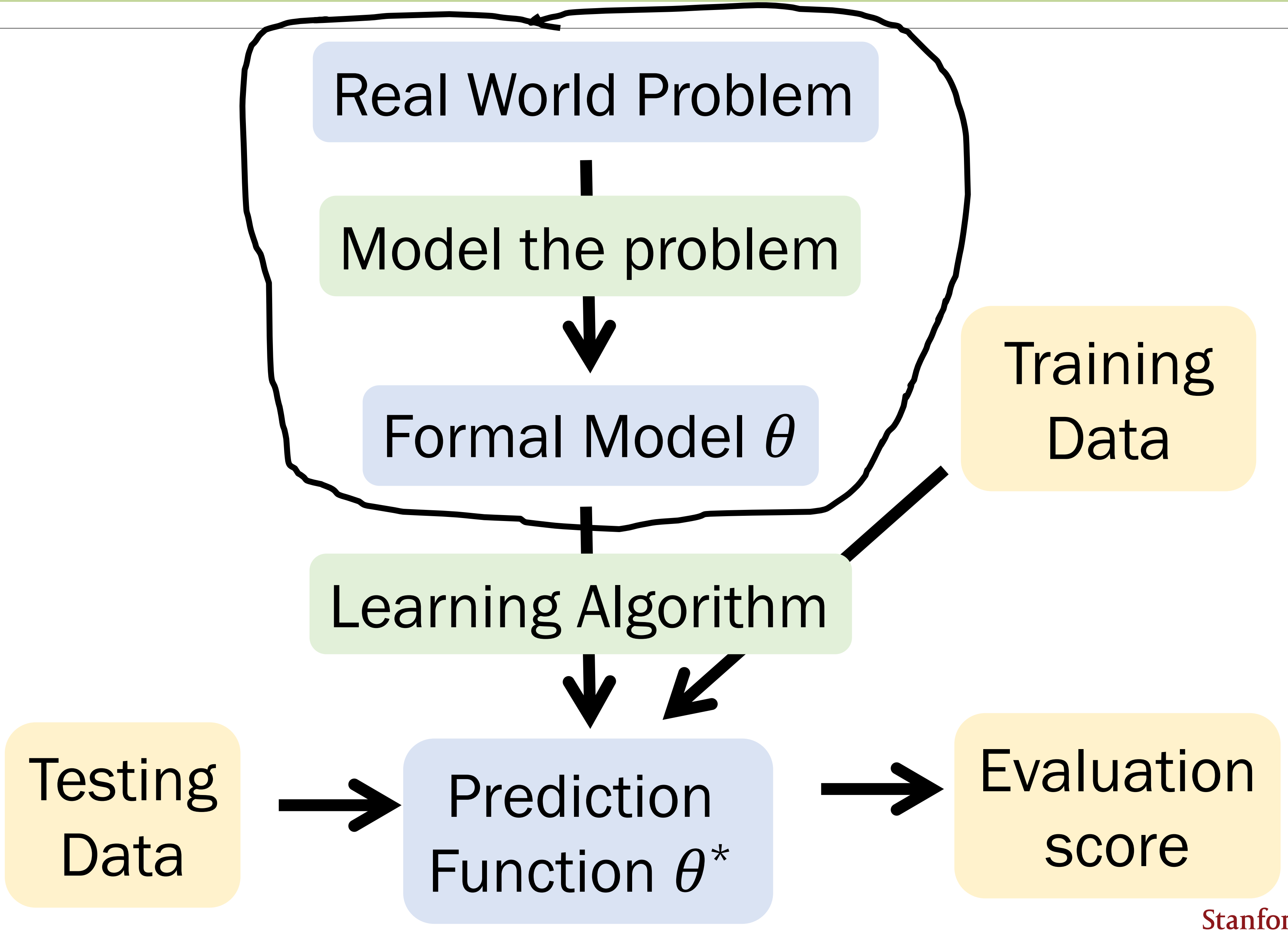
$P(T = 1|Flu = 1, U = 1) = 1.0$

Parameters

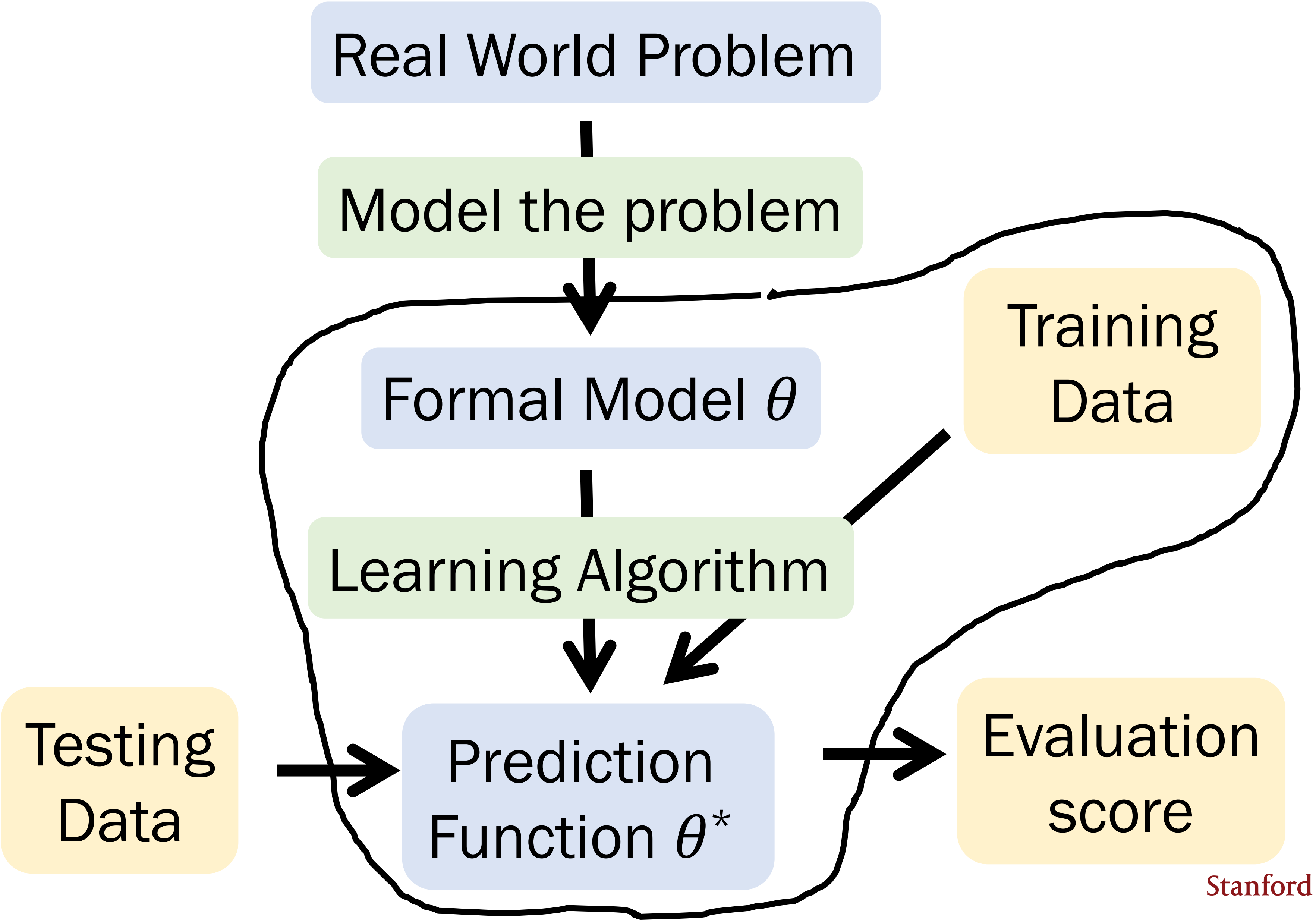
Why Do We Care?



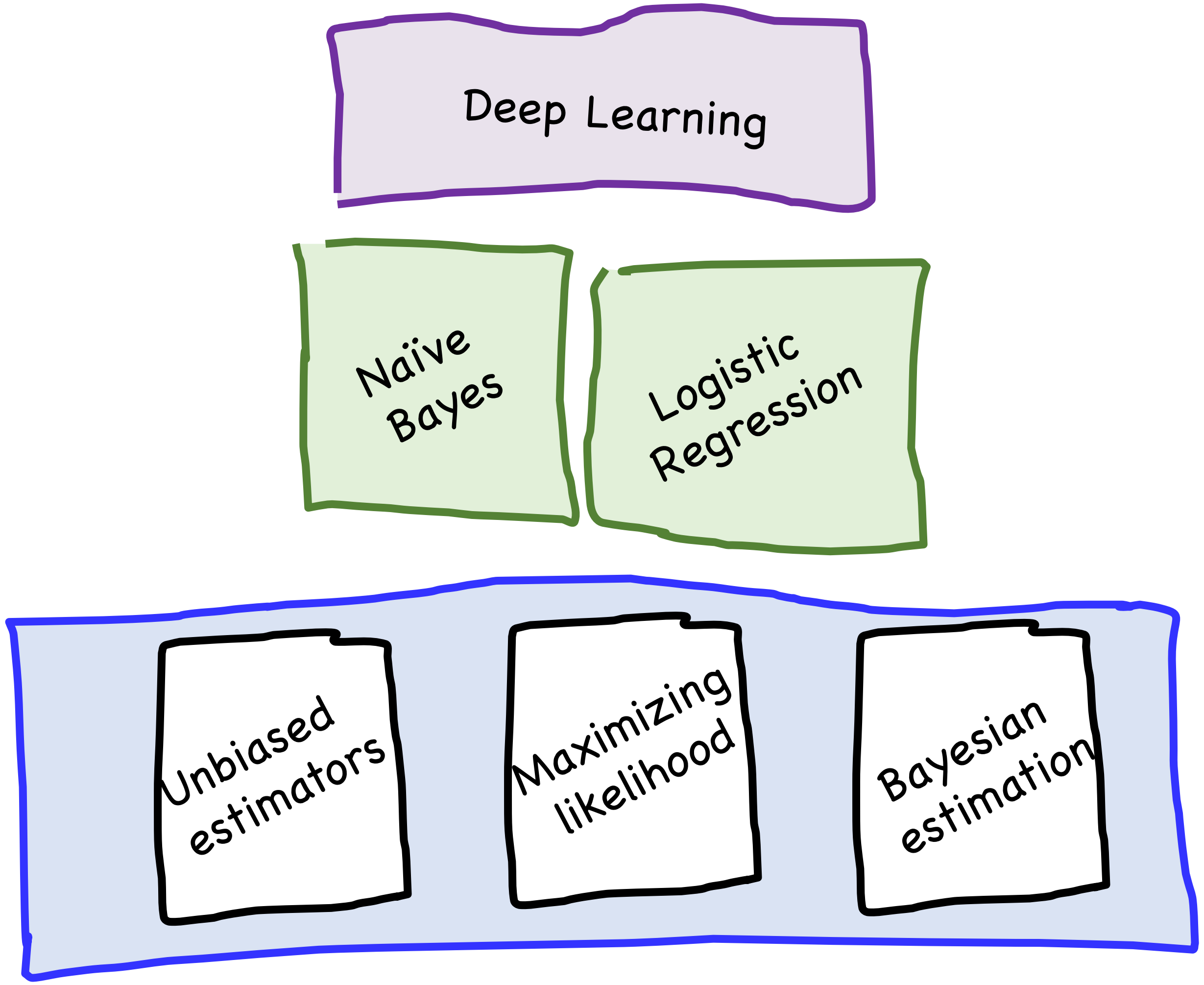
Modelling


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

Parameter Estimation (aka Training)

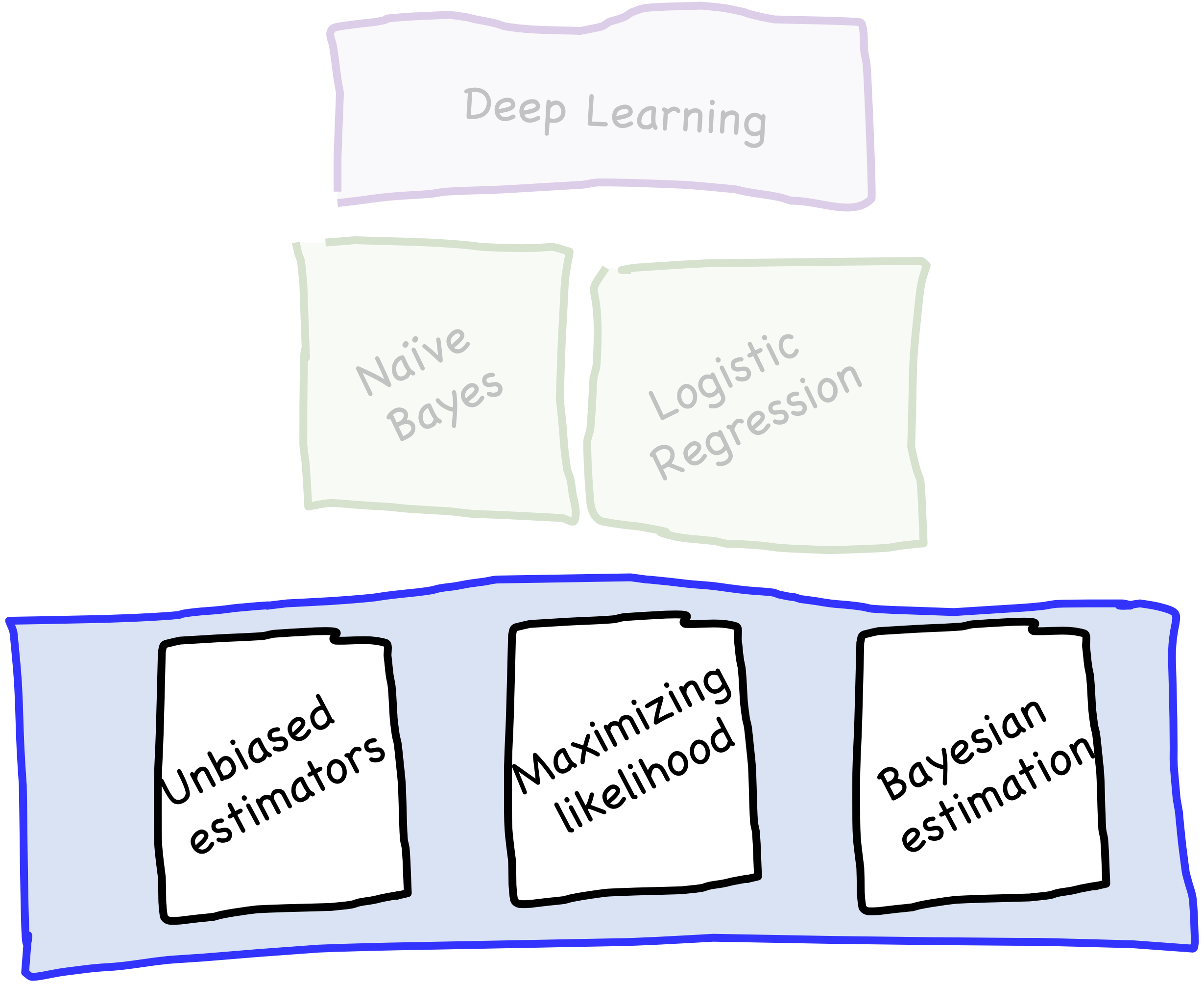


Our Path



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

Parameter Estimation



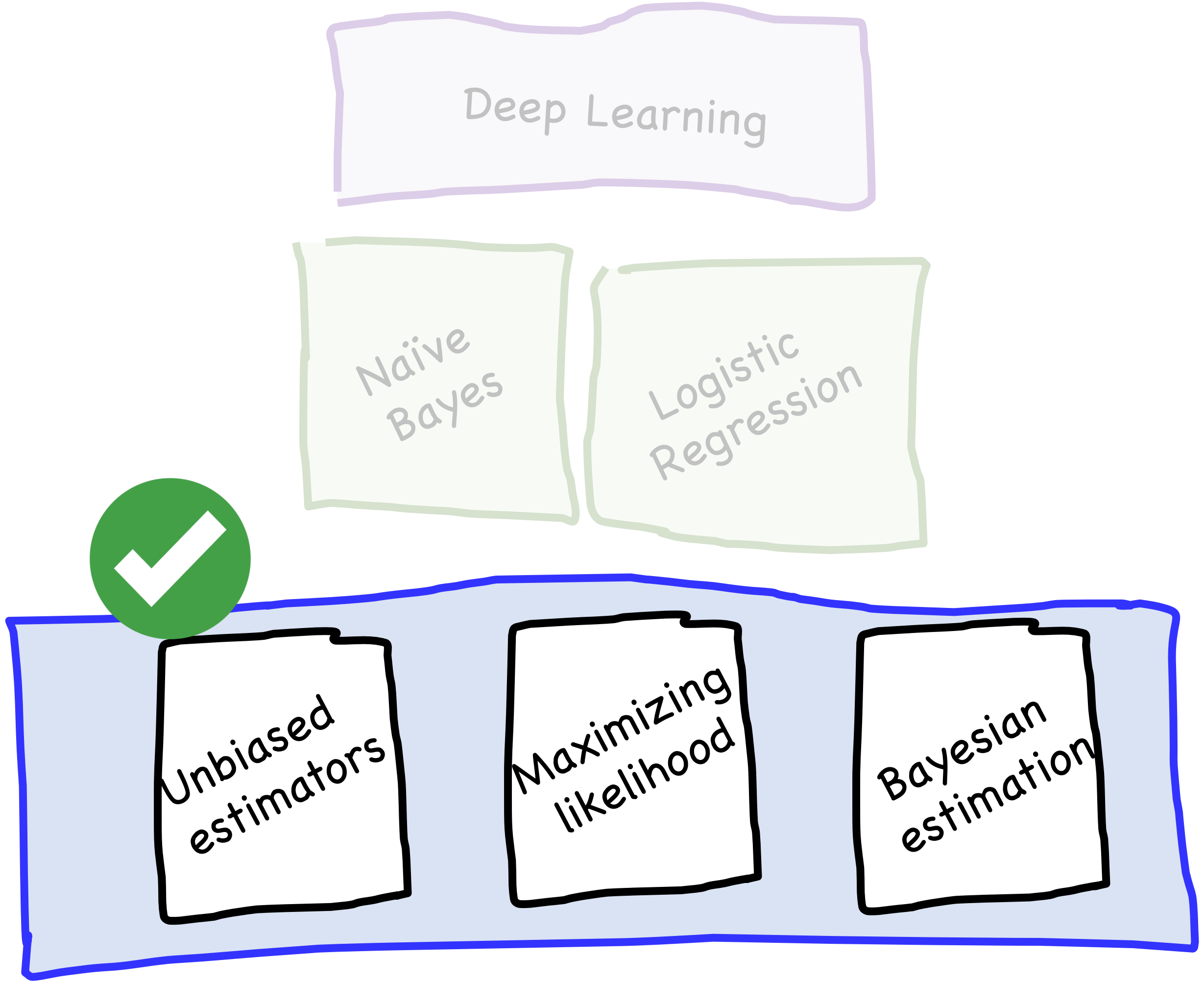
We've already seen some estimations

$X_1, X_2, \ldots, X_n$ are n i.i.d. random variables,
where X_i drawn from distribution F with $E[X_i] = \mu$, 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

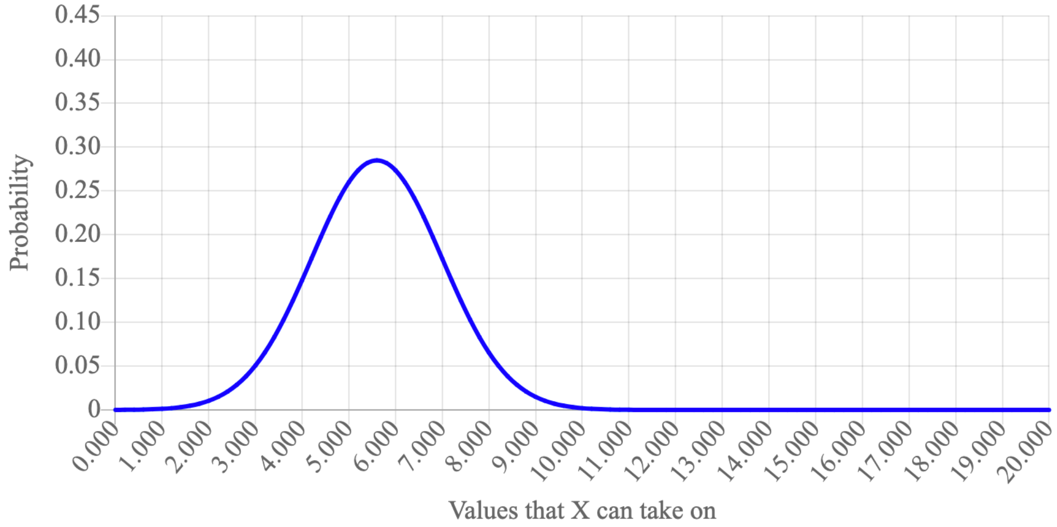
Likelihood

Likelihood: 1.9542923784106326e-15

Log Likelihood: -301.9

Best Seen: -301.9

PDF Graph




L-R
30
40
50
60

Insight: find the arguments that maximize measure of likelihood

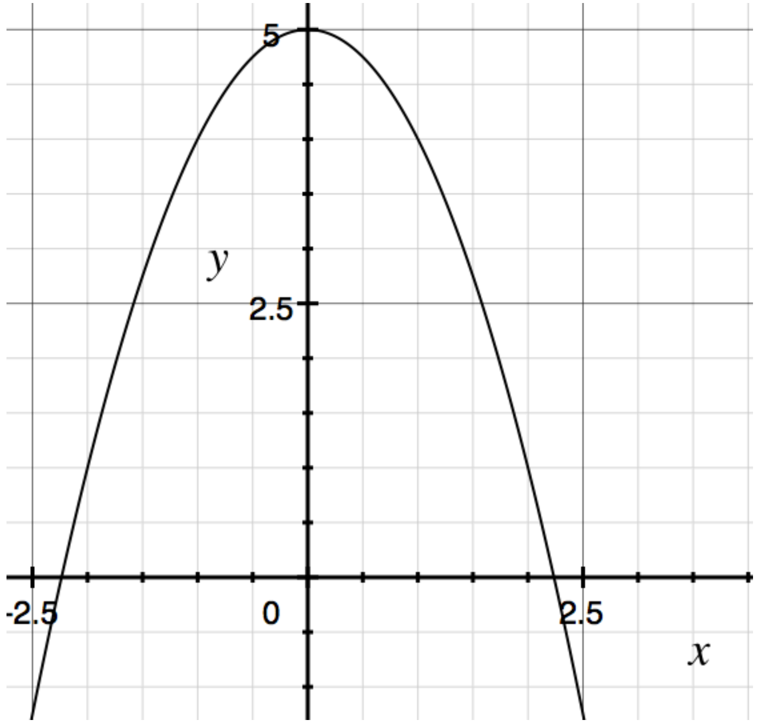
`argmax`

Argmax

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

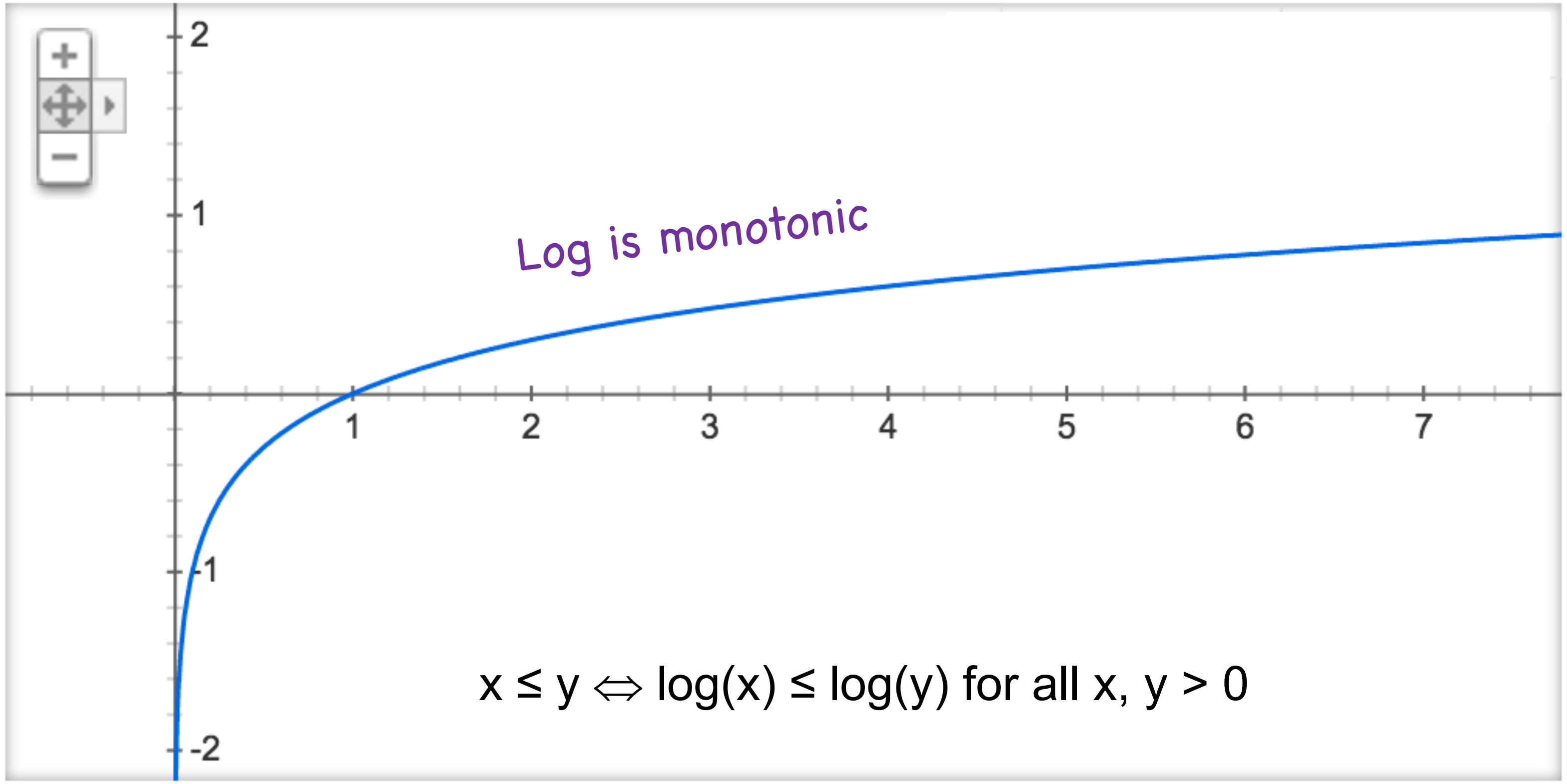
$$\max_x \; -x^2 + 5 = 5$$

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



Argmax of Log

Graph for log(x)



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

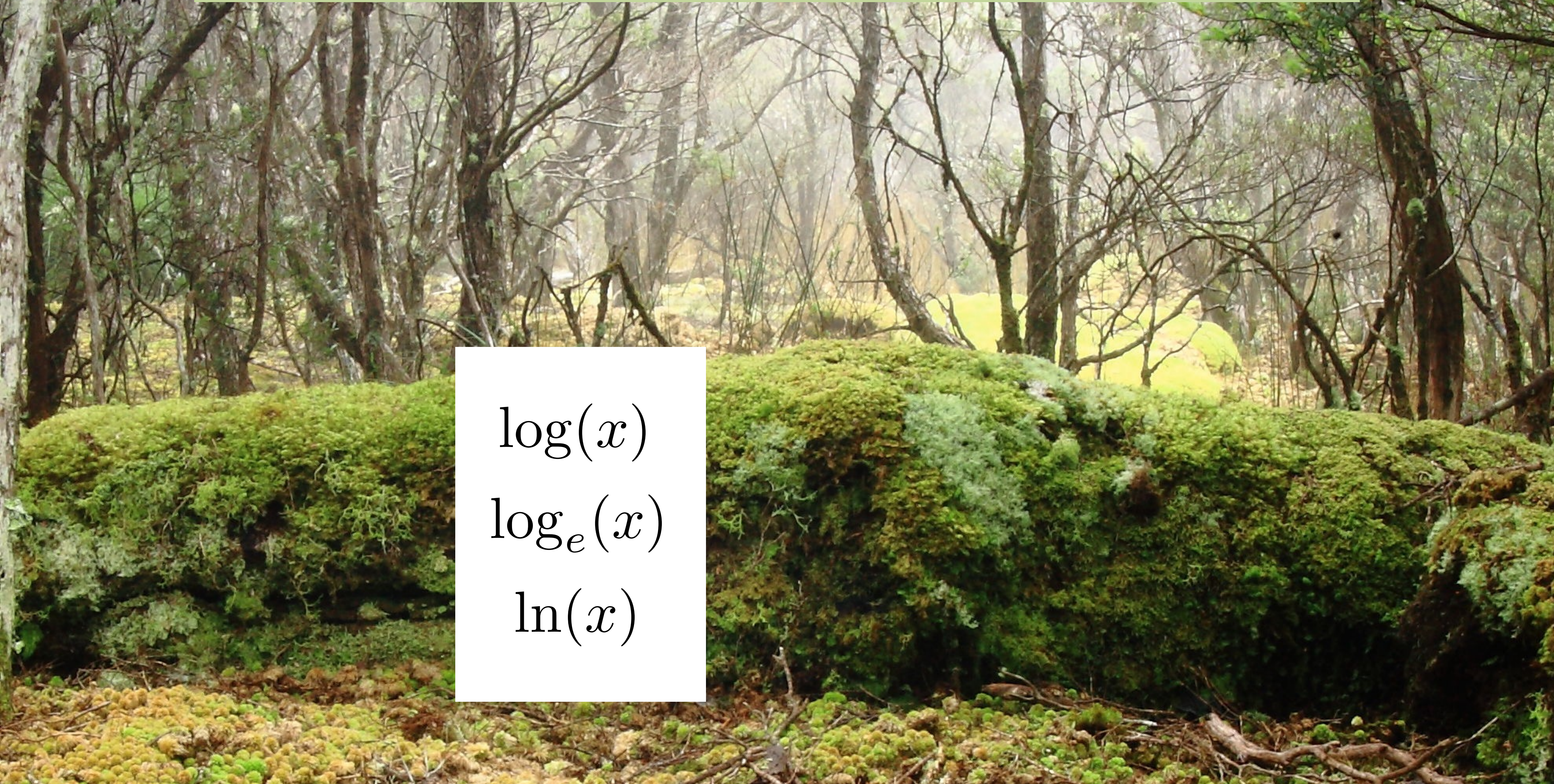
Argmax of Log

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

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

The Likelihood Function

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

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

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

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

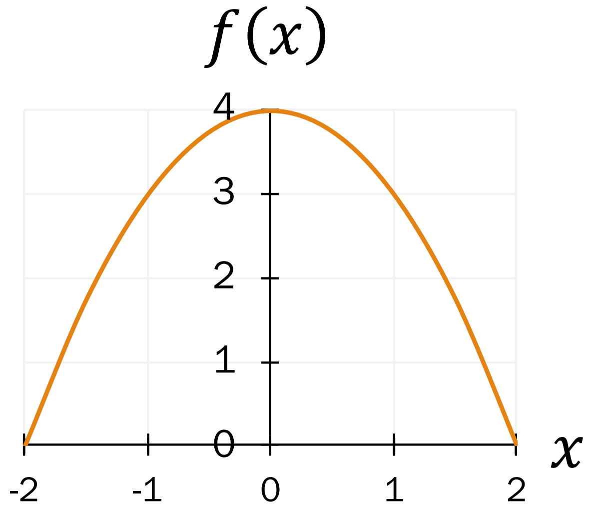
But how do we compute argmax?

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

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

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

General MLE Formula

Consider I.I.D. data: $X_1, X_2, \ldots, X_n$. Assume a model.

Use Maximum Likelihood to estimate parameters

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 λ which maximizes log likelihood

MLE for Poisson

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

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?
```

Maximum Likelihood with Poisson

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

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

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 λ which maximizes log likelihood

Maximum Likelihood with Poisson

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

- $X_i \sim \text{Poi}(\lambda)$ **Use Maximum Likelihood to estimate λ**
- 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 λ which maximizes log likelihood

Maximum Likelihood with Poisson

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

- $X_i \sim \text{Poi}(\lambda)$ **Use Maximum Likelihood to estimate λ**
- Probability mass function can be written as: $$f(x_i|\lambda) = \frac{e^{-\lambda}\lambda^{x_i}}{x_i!}$$
- Likelihood: $$L(\lambda) = f(x_1 \ldots x_n|\lambda) = \prod_{i=1}^{n} f(x_i|\lambda) = \prod_{i=1}^{n} \frac{e^{-\lambda}\lambda^{x_i}}{x_i!}$$

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

4. Find the value of λ which maximizes log likelihood

Maximum Likelihood with Poisson

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

- $X_i \sim \text{Poi}(\lambda)$ **Use Maximum Likelihood to estimate λ**
- Probability mass function can be written as: $f(x_i|\lambda) = \frac{e^{-\lambda}\lambda^{x_i}}{x_i!}$
- Likelihood: $L(\lambda) = f(x_1 \ldots x_n|\lambda) = \prod_{i=1}^{n} f(x_i|\lambda) = \prod_{i=1}^{n} \frac{e^{-\lambda}\lambda^{x_i}}{x_i!}$
- Log-likelihood:

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

4. Find the value of λ which maximizes log likelihood

Maximum Likelihood with Poisson

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

- $X_i \sim \text{Poi}(\lambda)$ **Use Maximum Likelihood to estimate λ**
- Probability mass function can be written as: $f(x_i|\lambda) = \frac{e^{-\lambda}\lambda^{x_i}}{x_i!}$
- Likelihood: $L(\lambda) = f(x_1 \ldots x_n|\lambda) = \prod_{i=1}^{n} f(x_i|\lambda) = \prod_{i=1}^{n} \frac{e^{-\lambda}\lambda^{x_i}}{x_i!}$
- 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. λ, and set to 0:

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

Isn’t that the same as
the sample mean?

Yes. For Poisson.

MLE of Poisson is the sample mean

MLE for Bernoulli

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

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

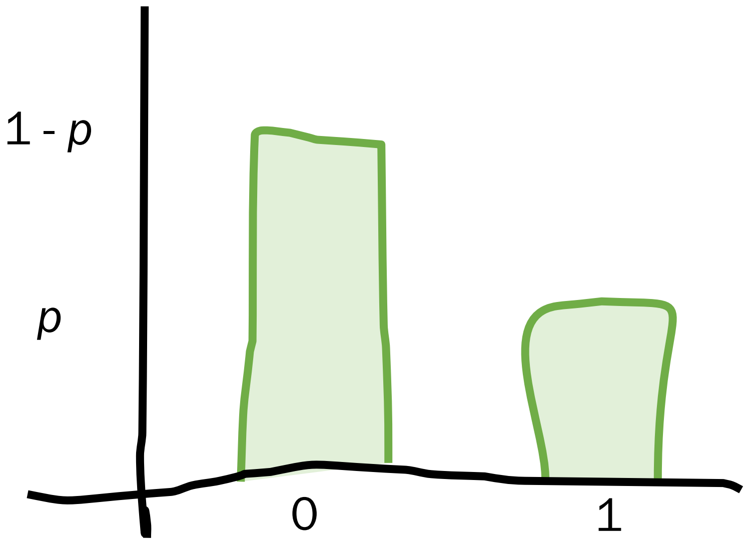
3. Solve resulting equations

Differentiable PMF for Bernoulli

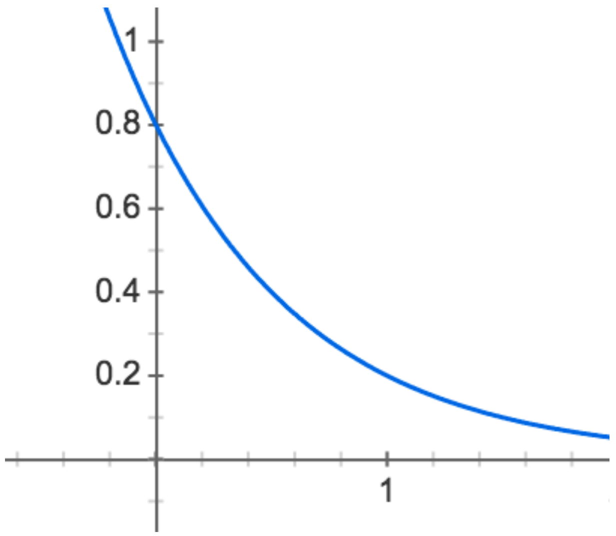
Consider I.I.D. random variables $X_1, X_2, \ldots, X_n$

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

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

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

Maximizing Likelihood with Bernoulli

- Consider I.I.D. random variables $X_1, X_2, \ldots, X_n$
  - $X_i \sim \mathrm{Ber}(p)$. **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

Maximizing Likelihood with Bernoulli

- Consider I.I.D. random variables $X_1, X_2, \ldots, X_n$
  - $X_i \sim \text{Ber}(p)$. **Use Maximum Likelihood to estimate *p*.**
  - Probability mass function, $f(X_i \mid p)$, can be written as:

$$f(X_i \mid p) = p^{x_i}(1-p)^{1-x_i} \quad \text{where } x_i = 0 \text{ or } 1$$

n

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

Maximizing Likelihood with Bernoulli

- Consider I.I.D. random variables $X_1, X_2, ..., X_n$
  - $X_i \sim \text{Ber}(p)$. **Use Maximum Likelihood to estimate *p*.**
  - Probability mass function, $f(X_i \mid p)$, can be written as:

    $$f(X_i \mid p) = p^{x_i}(1-p)^{1-x_i} \quad \text{where } x_i = 0 \text{ or } 1$$

  - Likelihood: $L(\theta) = \prod_{i=1}^{n} p^{X_i}(1-p)^{1-X_i}$

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

4. Find the value of p which maximizes log likelihood

Maximizing Likelihood with Bernoulli

- Consider I.I.D. random variables $X_1, X_2, \ldots, X_n$
  - $X_i \sim \text{Ber}(p)$. **Use Maximum Likelihood to estimate *p*.**
  - Probability mass function, $f(X_i \mid p)$, can be written as:

$$f(X_i \mid p) = p^{x_i}(1-p)^{1-x_i} \quad \text{where } x_i = 0 \text{ or } 1$$

  - Likelihood: $L(\theta) = \prod_{i=1}^{n} p^{X_i}(1-p)^{1-X_i}$
  - Log-likelihood:

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

4. Find the value of p which maximizes log likelihood

Maximizing Likelihood with Bernoulli

- Consider I.I.D. random variables $X_1, X_2, \ldots, X_n$
  - $X_i \sim \text{Ber}(p)$. **Use Maximum Likelihood to estimate *p*.**
  - Probability mass function, $f(X_i \mid p)$, can be written as:

    $$f(X_i \mid p) = p^{x_i}(1-p)^{1-x_i} \quad \text{where } x_i = 0 \text{ or } 1$$

  - Likelihood: $L(\theta) = \prod_{i=1}^{n} p^{X_i}(1-p)^{1-X_i}$
  - Log-likelihood:

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

  - Differentiate w.r.t. p, and set to 0:

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

    $$\frac{\partial LL(p)}{\partial p} = Y\frac{1}{p} + (n-Y)\frac{-1}{1-p} = 0 \quad \Rightarrow \quad p_{MLE} = \frac{Y}{n} = \frac{1}{n}\sum_{i=1}^{n} X_i$$

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:

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

- 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?

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

Its so general!

MLE for Gaussian

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

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]

What are the parameters?

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

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

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 $X_1, X_2, \ldots, X_n$.

- 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

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

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}:

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

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

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}:

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

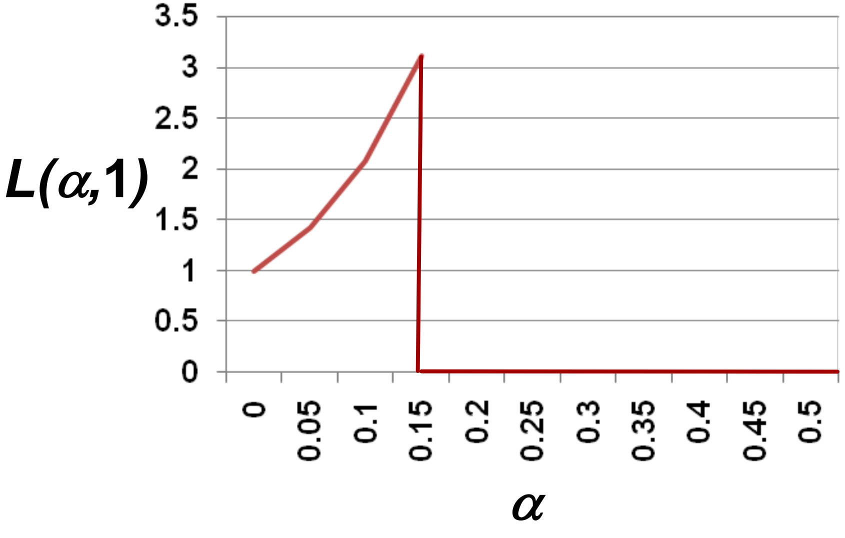
biased

Understanding MLE with Uniform

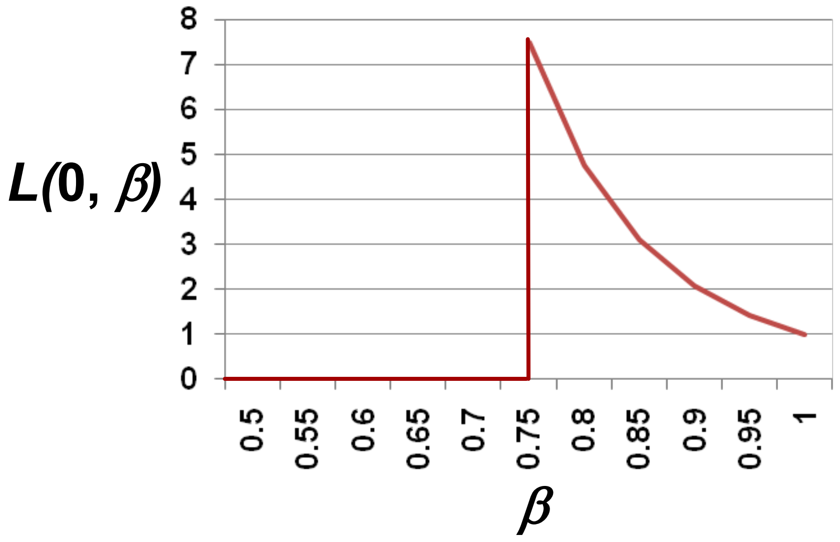
Consider I.I.D. random variables $X_1, X_2, ..., X_n$

- $X_i \sim \text{Uni}(0, 1)$
- 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, $\sigma^2_{MLE} = \frac{1}{n}\sum_{i=1}^{n} (X_i - \mu_{MLE})^2$
  - 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:

- **Asymptotically optimal** $\lim_{n \to \infty} P(|\hat{\theta} - \theta| < \varepsilon) = 1$ for $\varepsilon > 0$
- **Potentially biased** (though asymptotically less so)
- **Often used in practice**

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

MLE of the Wind

Climate sensitivity suggests that there is a fierce urgency to developing clean energy solutions. Wind is a powerful yet unpredictable source of clean energy and thus requires probability theory. The speed of the wind at a windfarm is a random variable that varies as a *Rayliegh Distribution*. A Rayliegh distribution is parameterized by a single scale parameter θ and has the following probability density function.

$$f_X(x) = \begin{cases} \frac{x}{\theta} e^{-x^2/2\theta} & x \geq 0 \\ 0 & else \end{cases}$$

We wish to model the wind speed on a wind farm. To this end we collect N independent measurements of wind speeds $w_1, w_2, \cdots, w_N$.

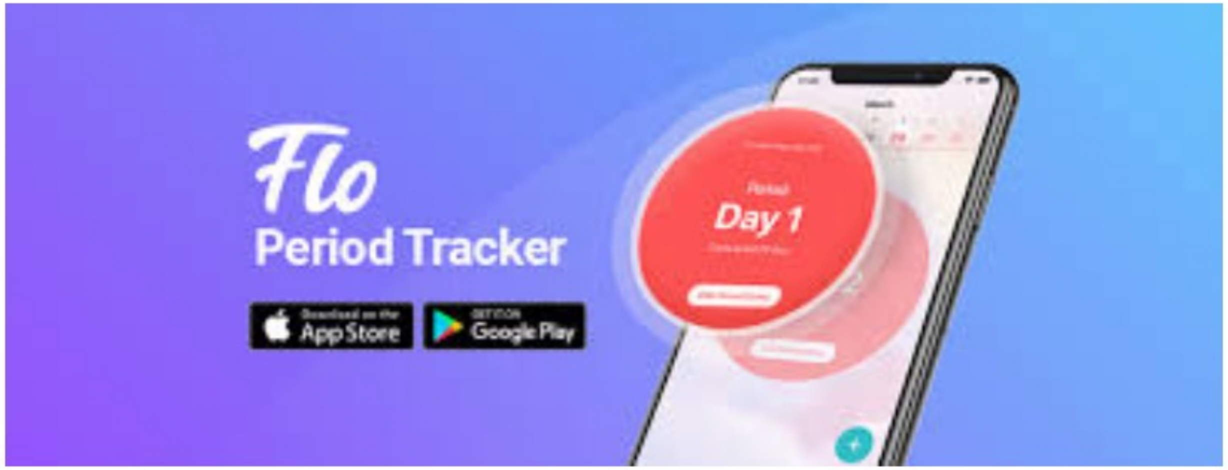
Your Task: Derive an equation for the maximum likelihood estimate of θ if we are modeling the wind speed as coming from a Rayleigh distribution. Make sure to include the equation in your answer. Then use the equation to estimate θ for observed 10 speeds:

```
[7.55, 8.15, 8.91, 1.17, 6.77, 3.03, 8.43, 5.56, 3.26, 2.55]
```

Give your answer to three decimal places

MLE Will Certainly Show up on the Final

2. **Flo. Tracking Menstrual Cycles**



Let X represent the length of a menstrual cycle: the number of days, as a continuous value, between the first moment of one period to the first moment of the next, for a given person. X is parameterized by α and β with probability density function:

$$f(X = x) = \beta \cdot (x - \alpha)^{\beta-1} \cdot e^{-(x-\alpha)^2}$$

MLE Will Certainly Show up on the Final

5 Reliability engineering (23 points)

The “reliability distribution" is a random variable parameterized by a with PDF:

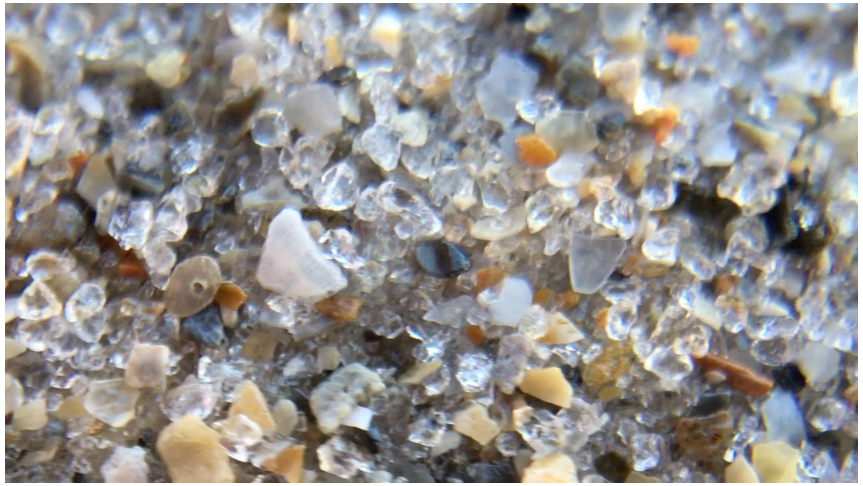
$$f(X = x) = \frac{1}{a^2} x^{a-1} e^{-\frac{x^2}{a^2}}$$

We wish to model how long a particular model of phone will function before it breaks. We are going to use a reliability distribution. To this end we collect N independent measurements of how long the type of phone functions before it breaks: $x_1, x_2, \ldots, x_N$. Explain, in words, how you would choose parameter a using the maximum likelihood estimation framework, and provide any necessary derivatives.

Can you learn a parameter from data?

```python
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]
```

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



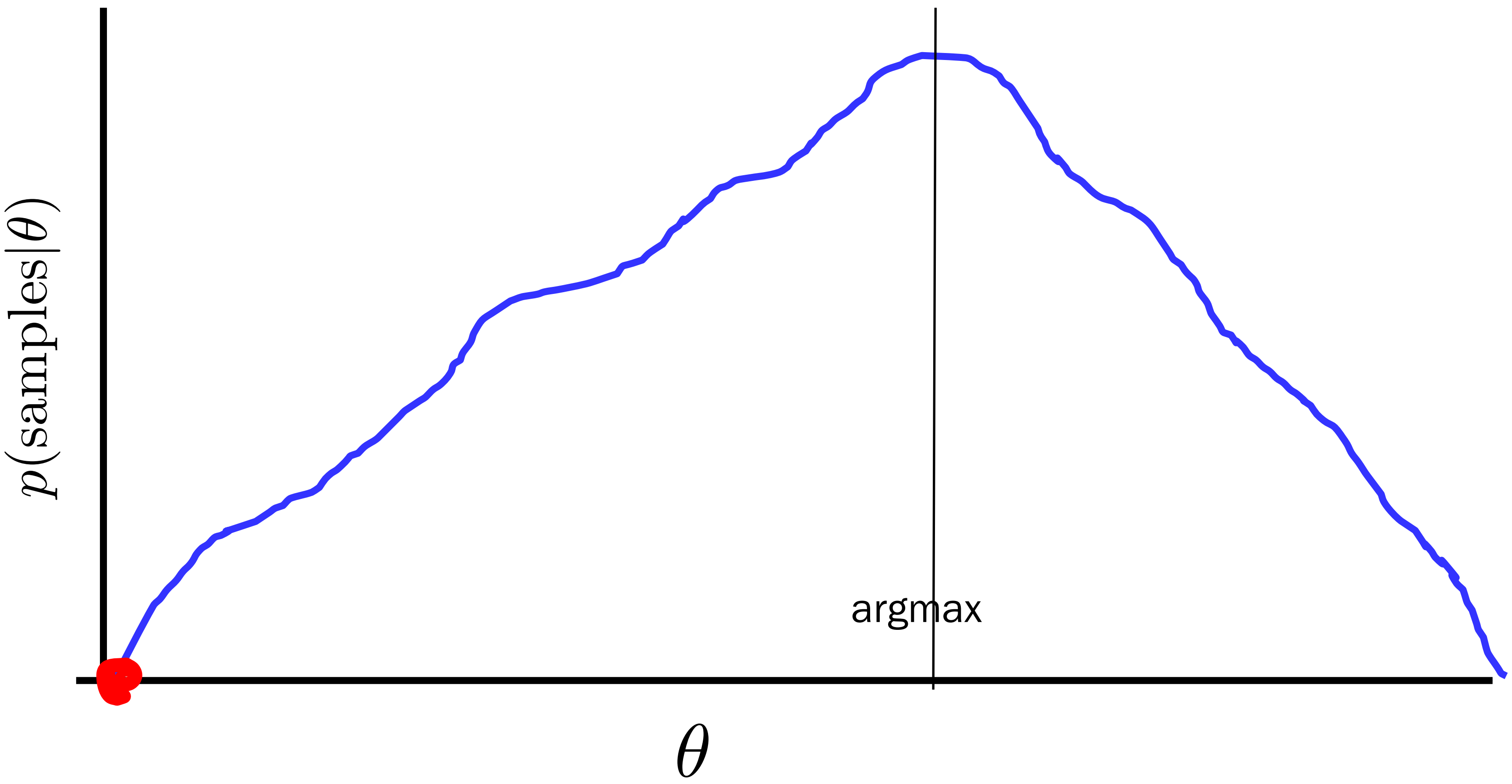
We know sand is distributed as a pareto with PDF

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


Disney
THE
LION KING II
SIMBA'S • PRIDE

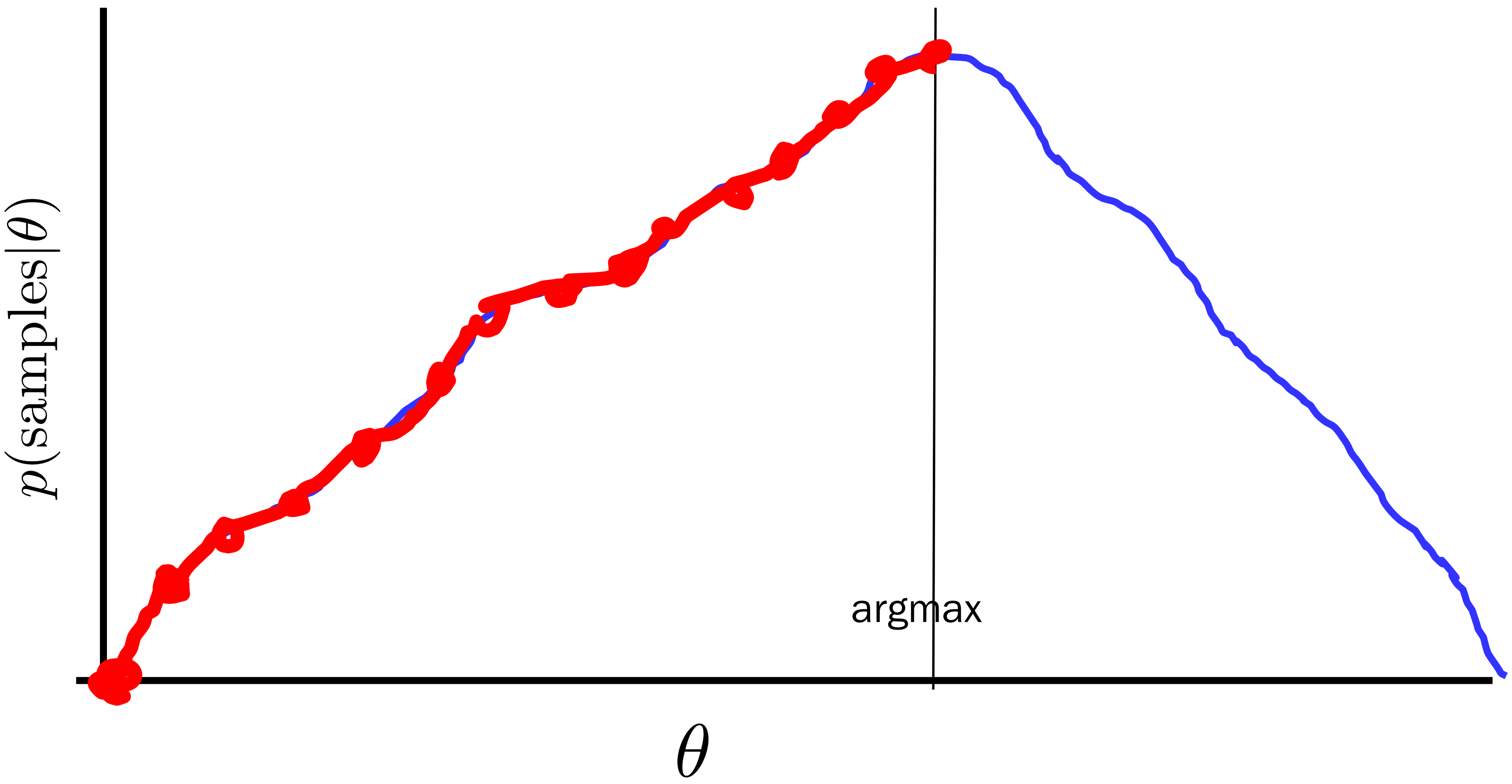
Optimization (argmax)
Option #2: Gradient Descent

Gradient Ascent



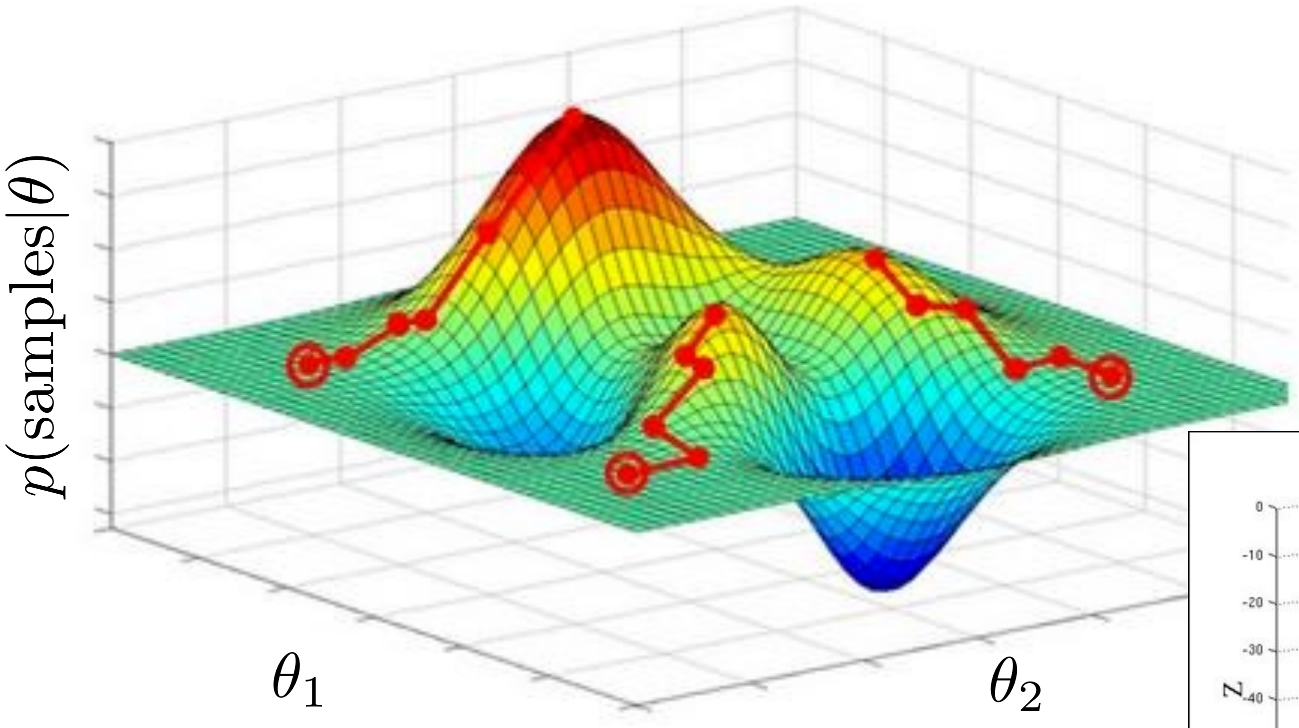
Walk uphill and you will find a local maxima
(if your step size is small enough)

Gradient Ascent

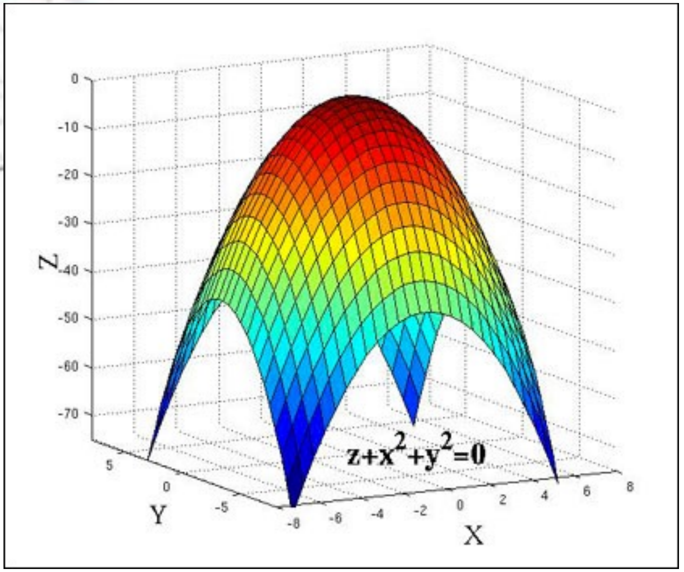


Walk uphill and you will find a local maxima
(if your step size is small enough)

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

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

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: θ_j = random for all $0 \leq j \leq m$

Calculate all θ_j

Gradient Ascent

Initialize: θ_j = random for all $0 \le j \le m$

Repeat many times:

gradient[j] = 0 for all $0 \le j \le m$

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

θ_j -= η * gradient[j] for all $0 \le j \le m$

Gradient Ascent

Initialize: θ_j **= random for all 0 ≤** j **≤** m

Repeat many times:

`gradient[j] = 0 for all 0 ≤` j ≤ m

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

$$\frac{dLL(\vec{\theta})}{d\mu_a} = \sum_i^n \frac{d}{d\mu_a}\left[-\frac{1}{2}\left(\frac{x_i-\mu_a}{\sigma_a}\right)^2\right]$$
$$= \sum_i^n 2\left(\frac{x_i-\mu_a}{\sigma_a}\right)\frac{1}{\sigma_a}$$

θ_j `-= η * gradient[j] for all 0 ≤` j ≤ m

Gradient Ascent

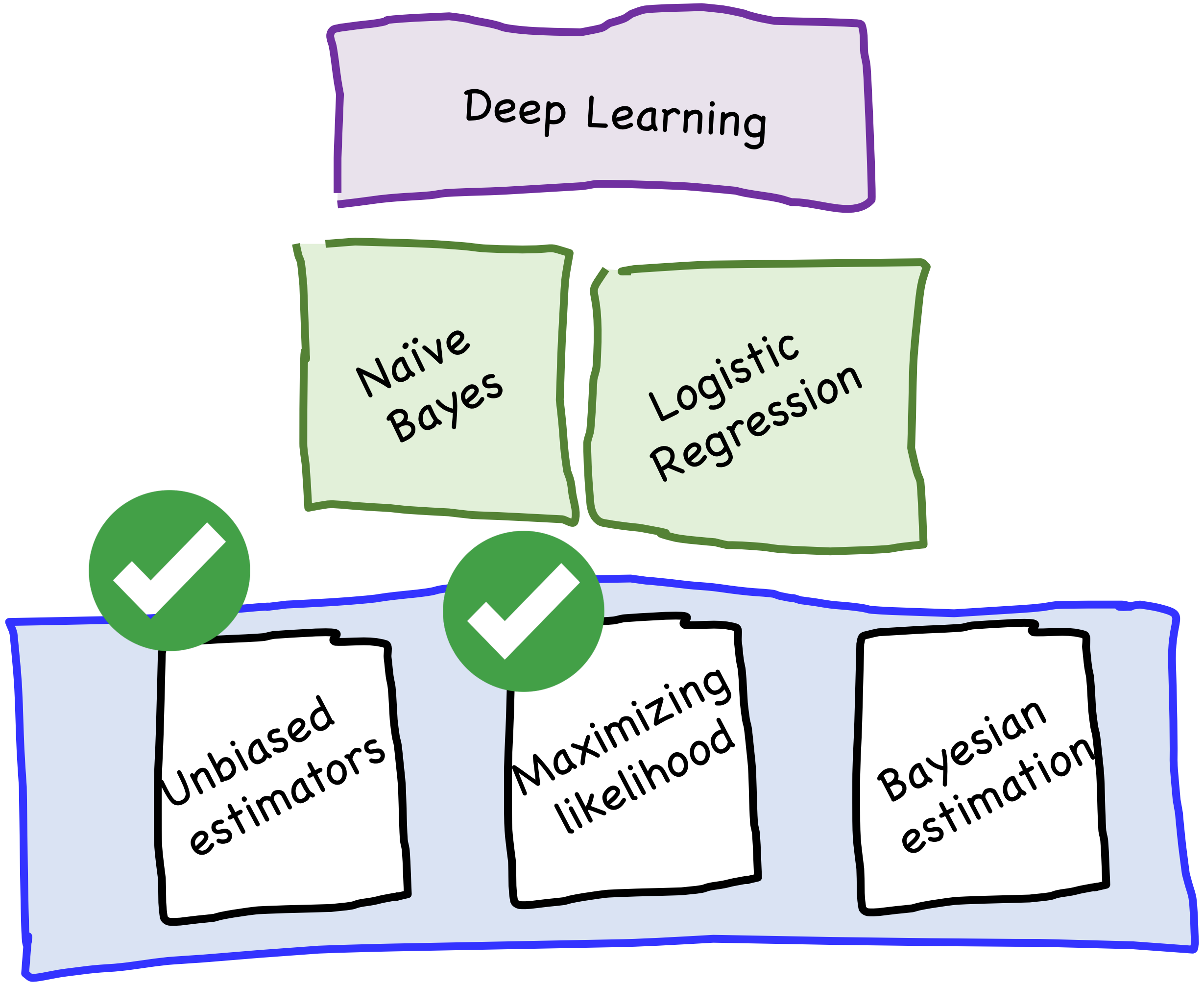
Repeat many times

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

This is some **profound** life philosophy

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

Our Path


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

Gradient **descent** is your bread and butter algorithm for optimization
(use argmin of neg LL)

Next Level MLE Example: Mixture of Gaussians

Data = [6.47, 5.82, 8.7, 4.76, 7.62, 6.95, 7.44, 6.73, 3.38, 5.89, 7.81, 6.93, 7.23, 6.25, 5.31, 7.71, 7.42, 5.81, 4.03, 7.09, 7.1, 7.62, 7.74, 6.19, 7.3, 7.37, 6.99, 2.97, 3.3, 7.08, 6.23, 3.67, 3.05, 6.67, 6.5, 6.08, 3.7, 6.76, 6.56, 3.61, 7.25, 7.34, 6.27, 6.54, 5.83, 6.44, 5.34, 7.7, 4.19, 7.34]

Parameter t:	Parameter μ_a:	Parameter σ_a:	Parameter μ_b:	Parameter σ_b:
0.8	6.8	0.7	3.5	0.7