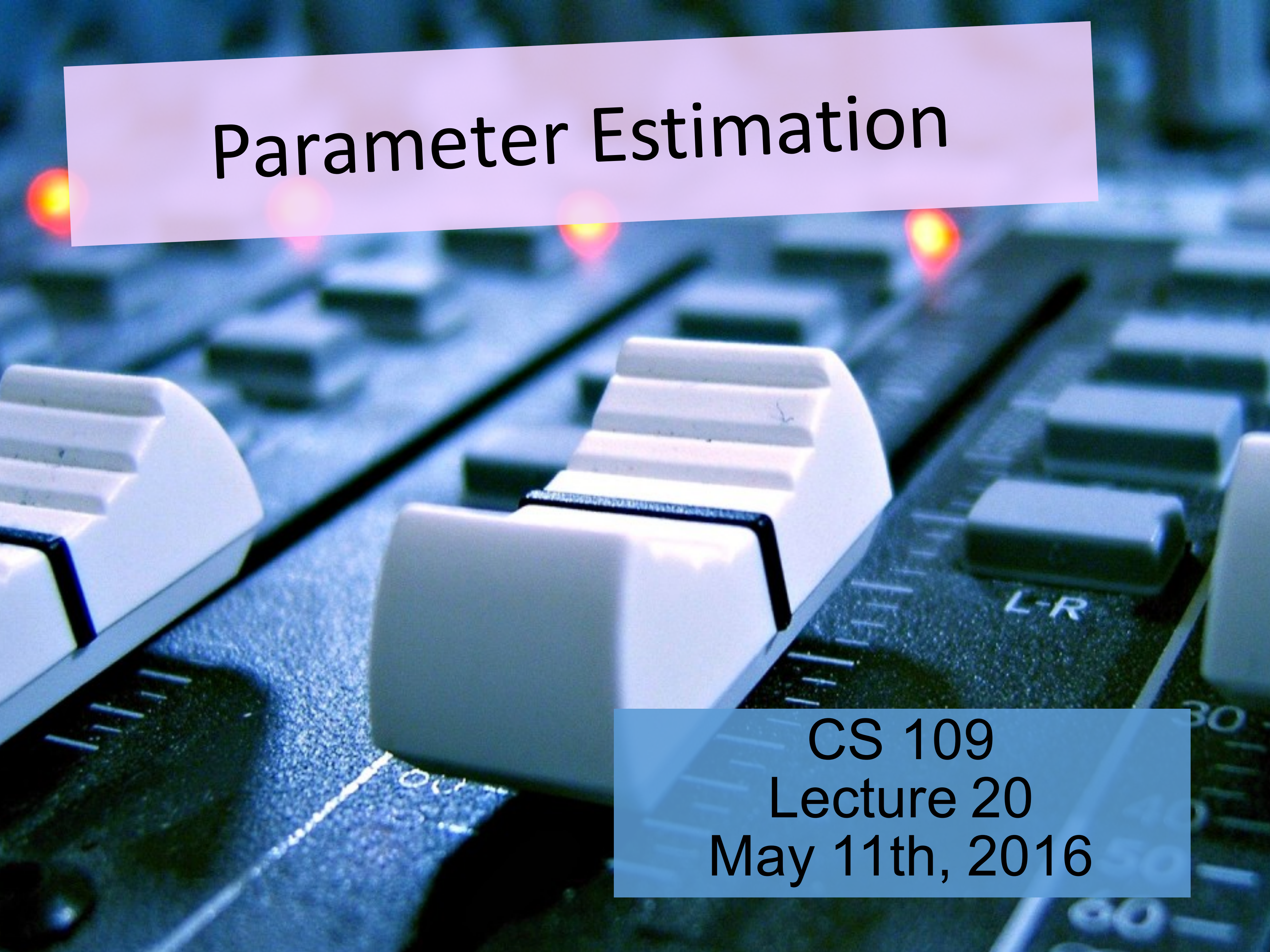
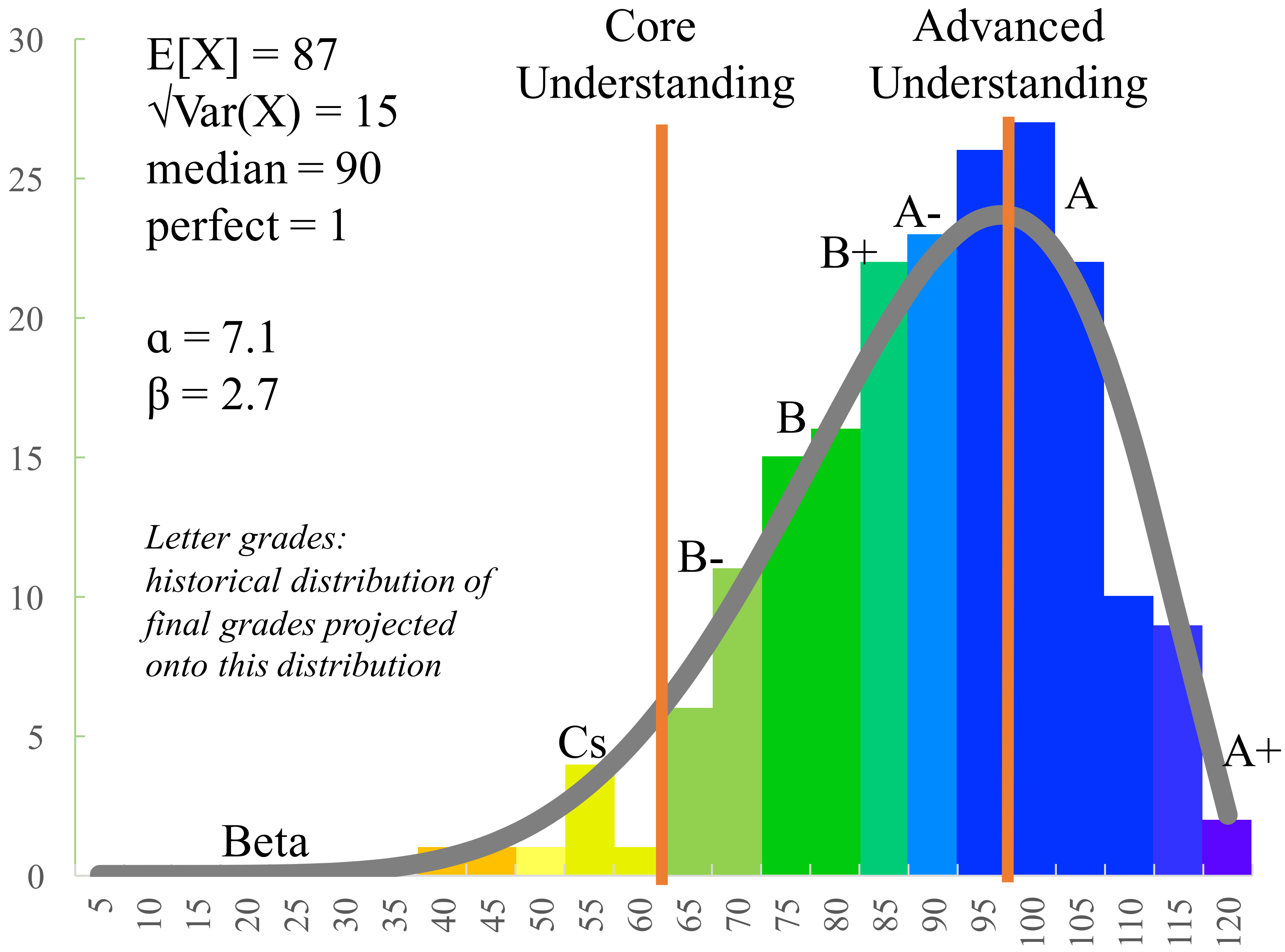
Parameter Estimation



CS 109
Lecture 20
May 11th, 2016



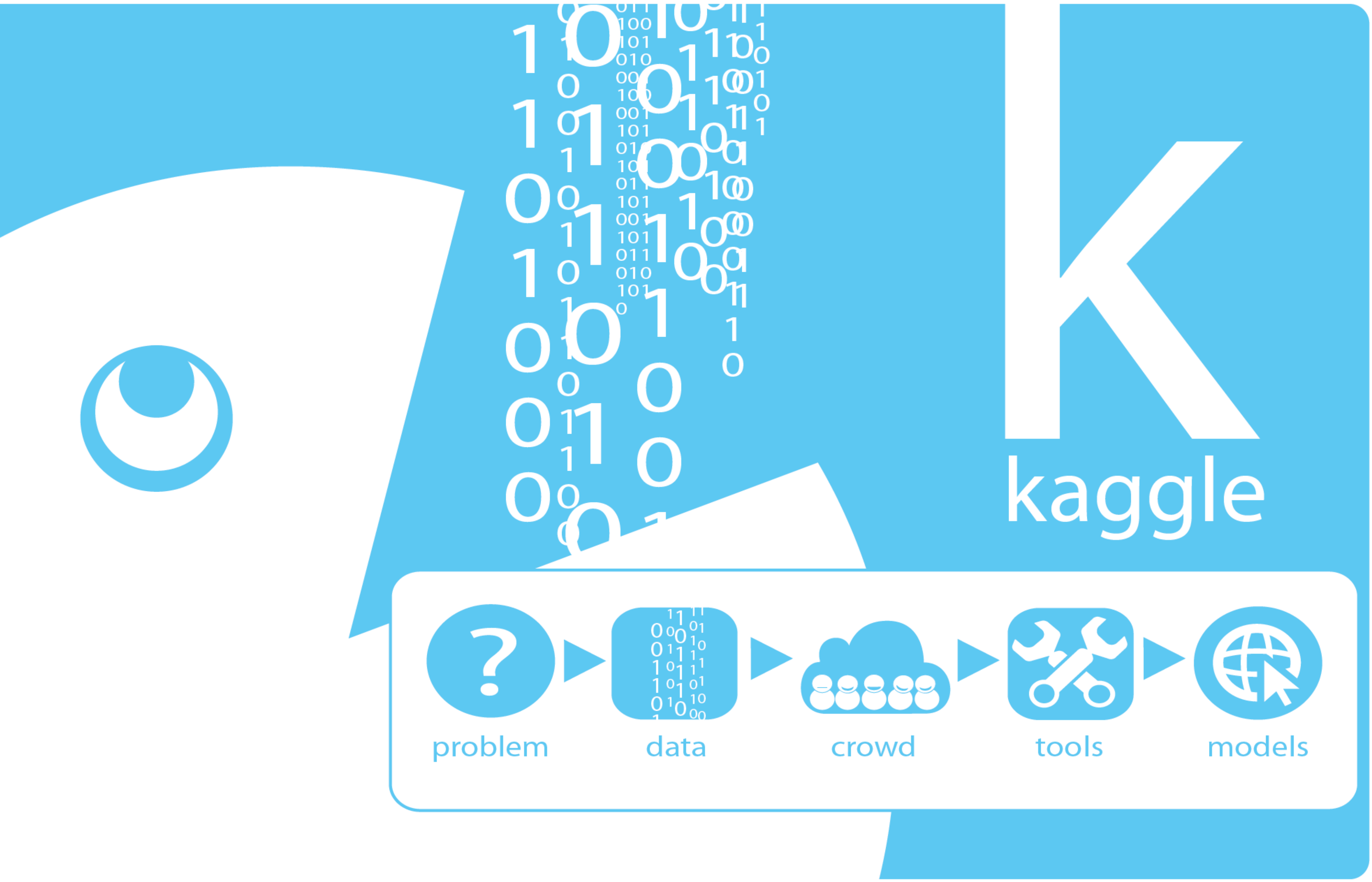
Core Understanding
Advanced Understanding
E[X] = 87
√Var(X) = 15
median = 90
perfect = 1
α = 7.1
β = 2.7
Letter grades: historical distribution of final grades projected onto this distribution
Beta
Cs
B-
B
B+
A-
A
A+
0
5
10
15
20
25
30
5 10 15 20 25 30 35 40 45 50 55 60 65 70 75 80 85 90 95 100 105 110 115 120

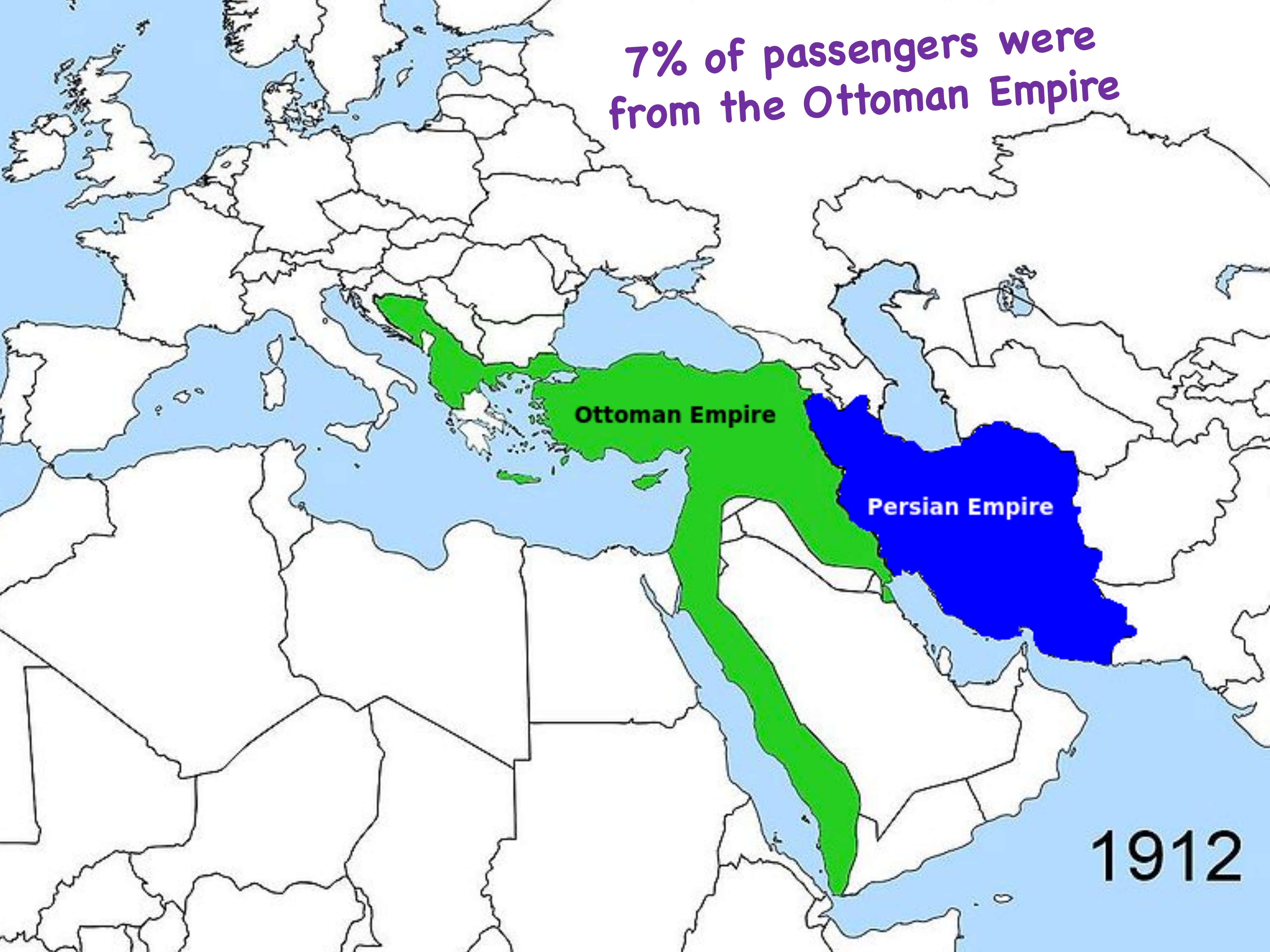
titanic.csv — website

overview.html | problem12.html | titanic.csv | index.html

```
Survived,Pclass,Name,Sex,Age,Siblings/Spouses Aboard,Parents/Children Aboard,Fare
0,3,"Braund, Mr. Owen Harris",male,22,1,0,7.25
1,1,"Cumings, Mrs. John Bradley (Florence Briggs Thayer)",female,38,1,0,71.2833
1,3,"Heikkinen, Miss. Laina",female,26,0,0,7.925
1,1,"Futrelle, Mrs. Jacques Heath (Lily May Peel)",female,35,1,0,53.1
0,3,"Allen, Mr. William Henry",male,35,0,0,8.05
0,3,"Moran, Mr. James",male,27,0,0,8.4583
0,1,"McCarthy, Mr. Timothy J",male,54,0,0,51.8625
0,3,"Palsson, Master. Gosta Leonard",male,2,3,1,21.075
1,3,"Johnson, Mrs. Oscar W (Elisabeth Vilhelmina Berg)",female,27,0,2,11.1333
1,2,"Nasser, Mrs. Nicholas (Adele Achem)",female,14,1,0,30.0708
1,3,"Sandstrom, Miss. Marguerite Rut",female,4,1,1,16.7
1,1,"Bonnell, Miss. Elizabeth",female,58,0,0,26.55
0,3,"Saundercock, Mr. William Henry",male,20,0,0,8.05
0,3,"Andersson, Mr. Anders Johan",male,39,1,5,31.275
0,3,"Vestrom, Miss. Hulda Amanda Adolfina",female,14,0,0,7.8542
1,2,"Hewlett, Mrs. (Mary D Kingcome) ",female,55,0,0,16
0,3,"Rice, Master. Eugene",male,2,4,1,29.125
1,2,"Williams, Mr. Charles Eugene",male,23,0,0,13
0,3,"Vander Planke, Mrs. Julius (Emelia Maria Vandemoortele)",female,31,1,0,18
1,3,"Masselmani, Mrs. Fatima",female,22,0,0,7.225
0,2,"Fynney, Mr. Joseph J",male,35,0,0,26
1,2,"Beesley, Mr. Lawrence",male,34,0,0,13
1,3,"McGowan, Miss. Anna ""Annie""",female,15,0,0,8.0292
1,1,"Sloper, Mr. William Thompson",male,28,0,0,35.5
0,3,"Palsson, Miss. Torborg Danira",female,8,3,1,21.075
1,3,"Asplund, Mrs. Carl Oscar (Selma Augusta Emilia Johansson)",female,38,1,5,31.3875
0,3,"Emir, Mr. Farred Chehab",male,26,0,0,7.225
0,1,"Fortune, Mr. Charles Alexander",male,19,3,2,263
1,3,"O'Dwyer, Miss. Ellen ""Nellie""",female,24,0,0,7.8792
0,3,"Todoroff, Mr. Lalio",male,23,0,0,7.8958
0,1,"Uruchurtu, Don. Manuel E",male,40,0,0,27.7208
1,1,"Spencer, Mrs. William Augustus (Marie Eugenie)",female,48,1,0,146.5208
1,3,"Glynn, Miss. Mary Agatha",female,18,0,0,7.75
0,2,"Wheadon, Mr. Edward H",male,66,0,0,10.5
```

Survived	Pclass	Name	Sex	Age	Siblings/Spouses Aboar	Parents/Children Aboard	Fare
0	3	Braund, Mr. Owen Harris	male	22	1	0	7.25
1	1	Cumings, Mrs. John Bradley (Florence	female	38	1	0	71.2833
1	3	Heikkinen, Miss. Laina	female	26	0	0	7.925
1	1	Futrelle, Mrs. Jacques Heath (Lily Ma	female	35	1	0	53.1
0	3	Allen, Mr. William Henry	male	35	0	0	8.05
0	3	Moran, Mr. James	male	27	0	0	8.4583
0	1	McCarthy, Mr. Timothy J	male	54	0	0	51.8625


kaggle
problem
data
crowd
tools
models


7% of passengers were
from the Ottoman Empire
Ottoman Empire
Persian Empire
1912


THE OTTOMAN EMPIRE IN 1683
Ottoman Beylik, 1300
Acquisitions, 1300 - 1359
Acquisitions, 1359 - 1451
Acquisitions, 1451 - 1481 (Mehmed II)
Acquisitions, 1512 - 1520 (Selim I)
Acquisitions, 1520 - 1566 (Suleiman the Magnificent)
Acquisitions, 1566 - 1683
Europe
Kiev
Paris
France
Vienna
Budapest
Jassy
Christian vassal states
Venice
Belgrade
Bucharest
Caffa
Marseille
Italy
Varna
Black Sea
Barcelona
Rome
Prizren
Sofia
Constantinople
Caspian Sea
Baku
Spain
Naples
Salonica
Sicily
Athens
Tunis
Algiers
Algeria
Rhodes
Alexandretta
Mediterranean Sea
Damascus
Baghdad
Tripoli
Alexandria
Jerusalem
Lybia
Cairo
Egypt
750 miles
1000 km
Africa
Arabia
Medina
Mecca
Red Sea
Kassala
Zeila

Review

The Central Limit Theorem

- Consider I.I.D. random variables X_1, X_2, ...
  - X_i have distribution with $E[X_i] = \mu$ and $Var(X_i) = \sigma^2$
  - Let: $\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$
  - Central Limit Theorem:

$$\bar{X} \sim N(\mu, \frac{\sigma^2}{n}) \quad \text{as} \quad n \to \infty$$

Demo

http://onlinestatbook.com/stat_sim/sampling_dist/

The Central Limit Theorem

- Consider I.I.D. random variables X_1, X_2, ...
  - X_i have distribution with $E[X_i] = \mu$ and $Var(X_i) = \sigma^2$
  - Let: $\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$ $\quad \bar{X} \sim N(\mu, \frac{\sigma^2}{n})$ as $n \to \infty$
  - Recall $Z = \frac{\bar{X} - \mu}{\sqrt{\sigma^2 / n}}$ where $Z \sim N(0, 1)$:

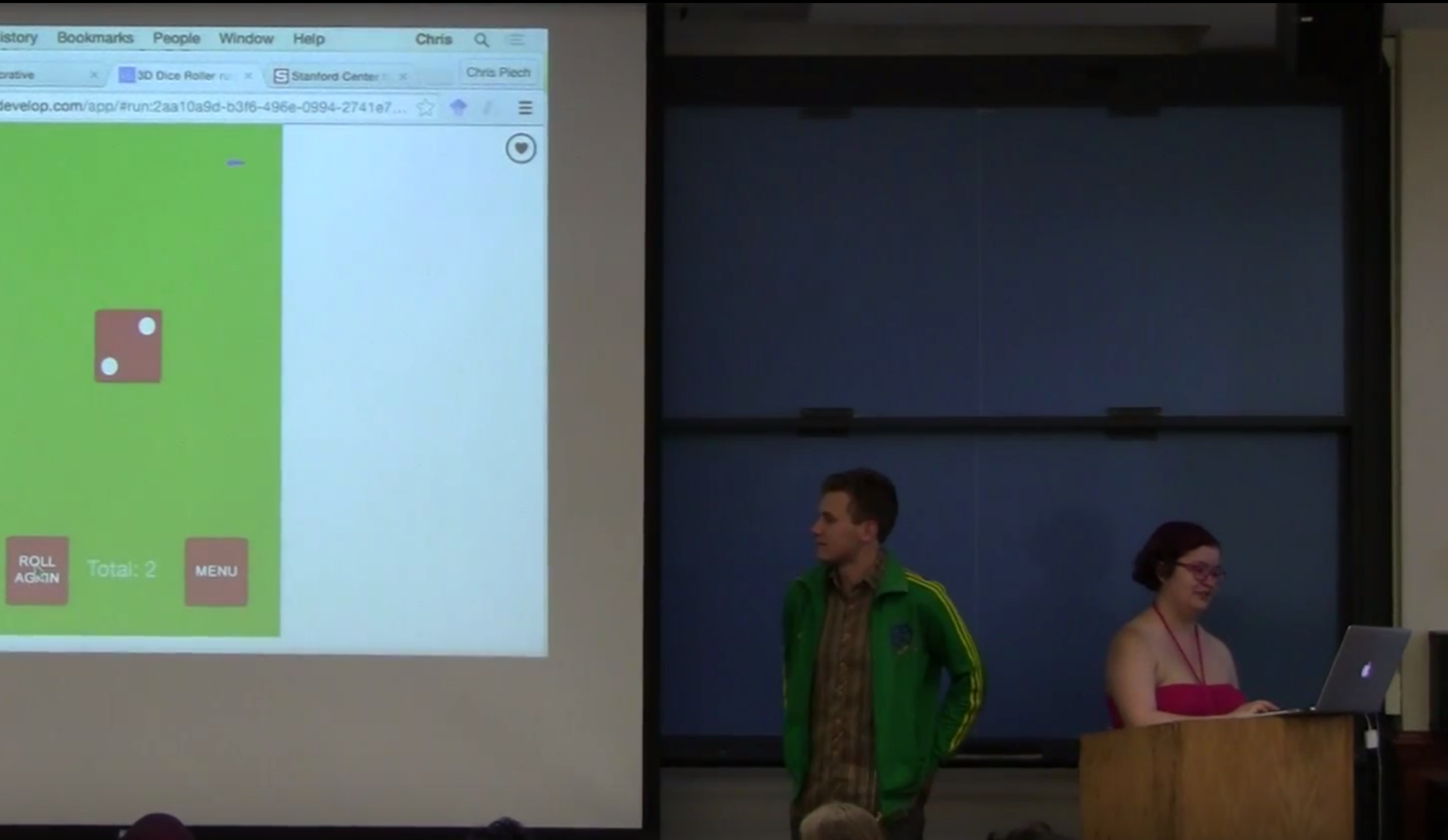
$$Z = \frac{\frac{1}{n}\left(\sum_{i=1}^{n} X_i\right) - \mu}{\sqrt{\sigma^2 / n}} = \frac{n\left[\frac{1}{n}\left(\sum_{i=1}^{n} X_i\right) - \mu\right]}{n\sqrt{\sigma^2 / n}} = \frac{\left(\sum_{i=1}^{n} X_i\right) - n\mu}{\sigma\sqrt{n}}$$

$$\frac{X_1 + X_2 + \ldots + X_n - n\mu}{\sigma\sqrt{n}} \sim N(0,1) \quad \text{as} \quad n \to \infty$$

Another form of the Central Limit Theorem

Thinking about play time!

Last Class we Played Sum of Dice



Sum of Dice

- You will roll 10 6-sided dice (X_1, X_2, …, X_{10})
  - X = total value of all 10 dice = $X_1 + X_2 + \ldots + X_{10}$
  - Win if: $X \le 25$ or $X \ge 45$
  - Roll!
- And now the truth (according to the CLT)…

Sum of Dice

- You will roll 10 6-sided dice (X_1, X_2, …, X_{10})
  - X = total value of all 10 dice = $X_1 + X_2 + \ldots + X_{10}$
  - Win if: $X \leq 25$ or $X \geq 45$
- Recall CLT: $\frac{X_1 + X_2 + \ldots + X_n - n\mu}{\sigma\sqrt{n}} \rightarrow N(0,1)$ as $n \rightarrow \infty$
  - Determine P(X ≤ 25 or X ≥ 45) using CLT:

$$\mu = E[X_i] = 3.5 \qquad \sigma^2 = \text{Var}(X_i) = \frac{35}{12}$$

$$1 - P(25.5 \leq X \leq 44.5) = 1 - P\left(\frac{25.5 - 10(3.5)}{\sqrt{35/12}\sqrt{10}} \leq \frac{X - 10(3.5)}{\sqrt{35/12}\sqrt{10}} \leq \frac{44.5 - 10(3.5)}{\sqrt{35/12}\sqrt{10}}\right)$$

$$\approx 1 - (2\Phi(1.76) - 1) \approx 2(1 - 0.9608) = 0.0784$$

Crashing Your Website

- Number visitors to web site/minute: X ~ Poi(100)
  - Server crashes if ≥ 120 requests/minute
  - What is P(crash in next minute)?
  - Exact solution: $P(X \geq 120) = \sum_{i=120}^{\infty} \frac{e^{-100}(100)^i}{i!} \approx 0.0282$
  - Use CLT, where $\text{Poi}(100) \sim \sum_{i=1}^{n} \text{Poi}(100/n)$ (all I.I.D)

$$P(X \geq 120) = P(Y \geq 119.5) = P\left(\frac{Y-100}{\sqrt{100}} \geq \frac{119.5-100}{\sqrt{100}}\right) = 1 - \Phi(1.95) \approx 0.0256$$

- Note: Normal can be used to approximate Poisson

Wonderful Form of Cosmic Order

I know of scarcely anything so apt to impress the imagination as the wonderful form of cosmic order expressed by the ”[Central limit theorem]". The law would have been personified by the Greeks and deified, if they had known of it. It reigns with serenity and in complete self-effacement, amidst the wildest confusion. The huger the mob, and the greater the apparent anarchy, the more perfect is its sway. It is the supreme law of Unreason. Whenever a large sample of chaotic elements are taken in hand and marshalled in the order of their magnitude, an unsuspected and most beautiful form of regularity proves to have been latent all along.

-Sir Francis Galton

End Review

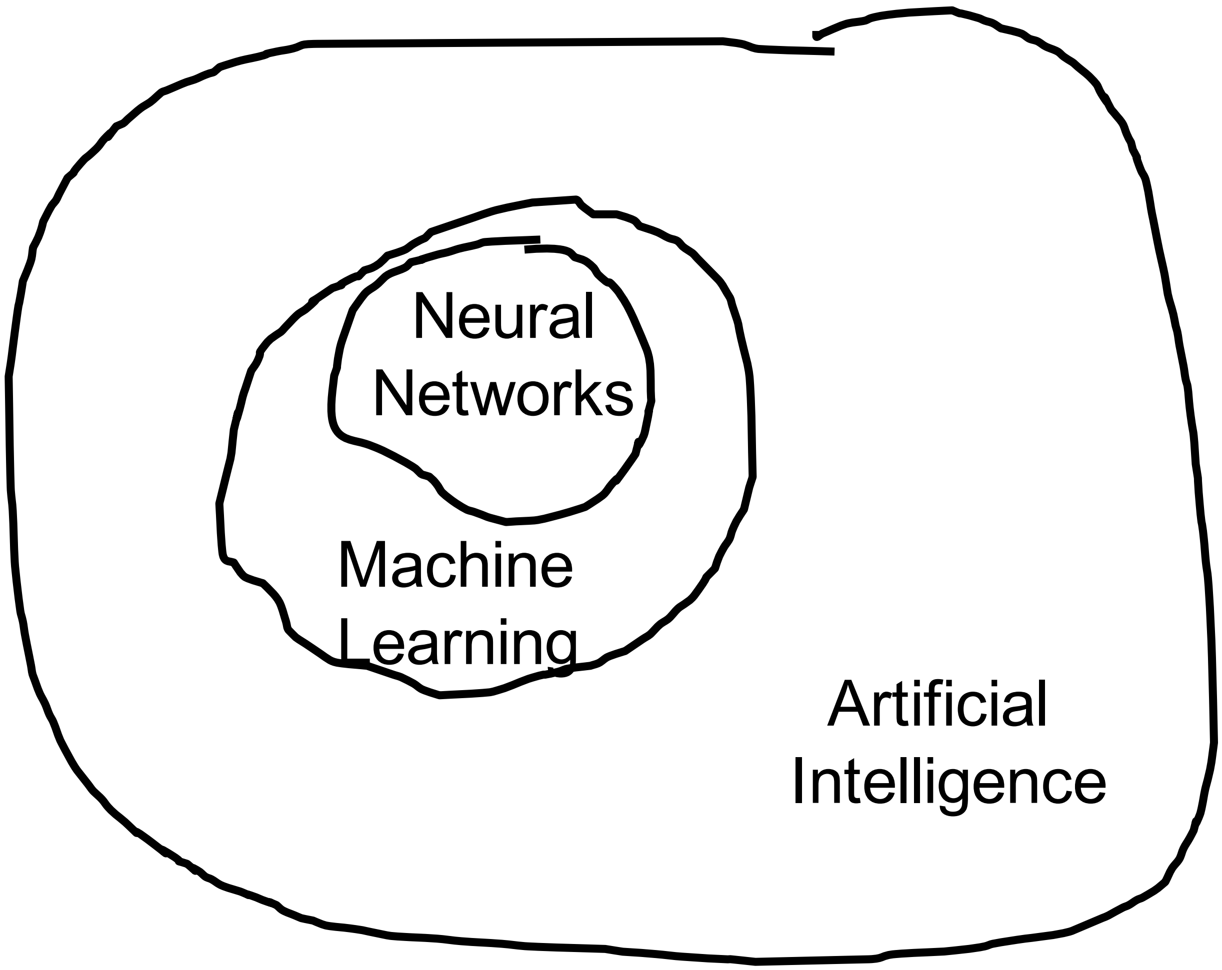
What is AI?

[suspense]

AI: The study and design of intelligent **agents**

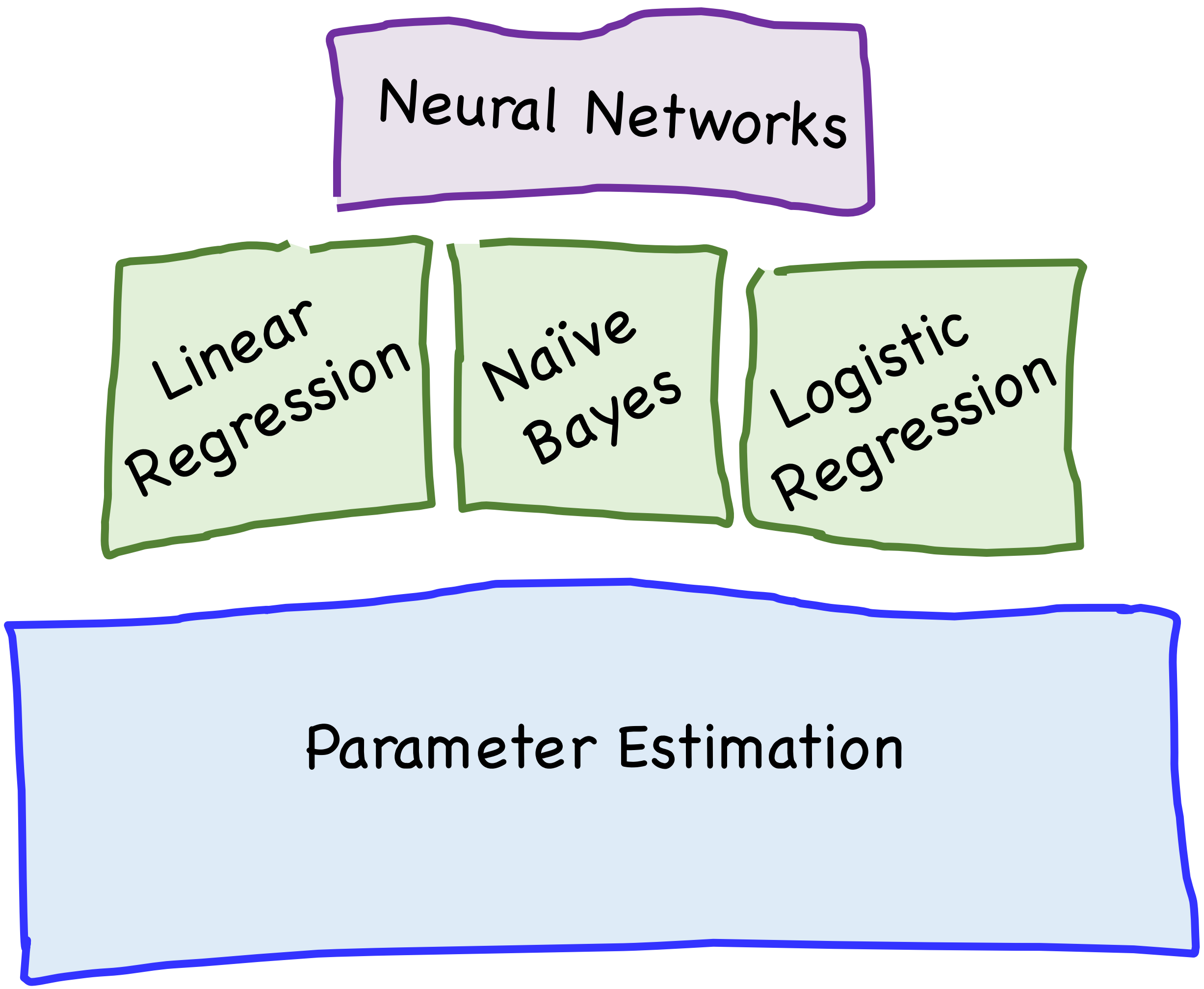
Volunteer

AI and Machine Learning

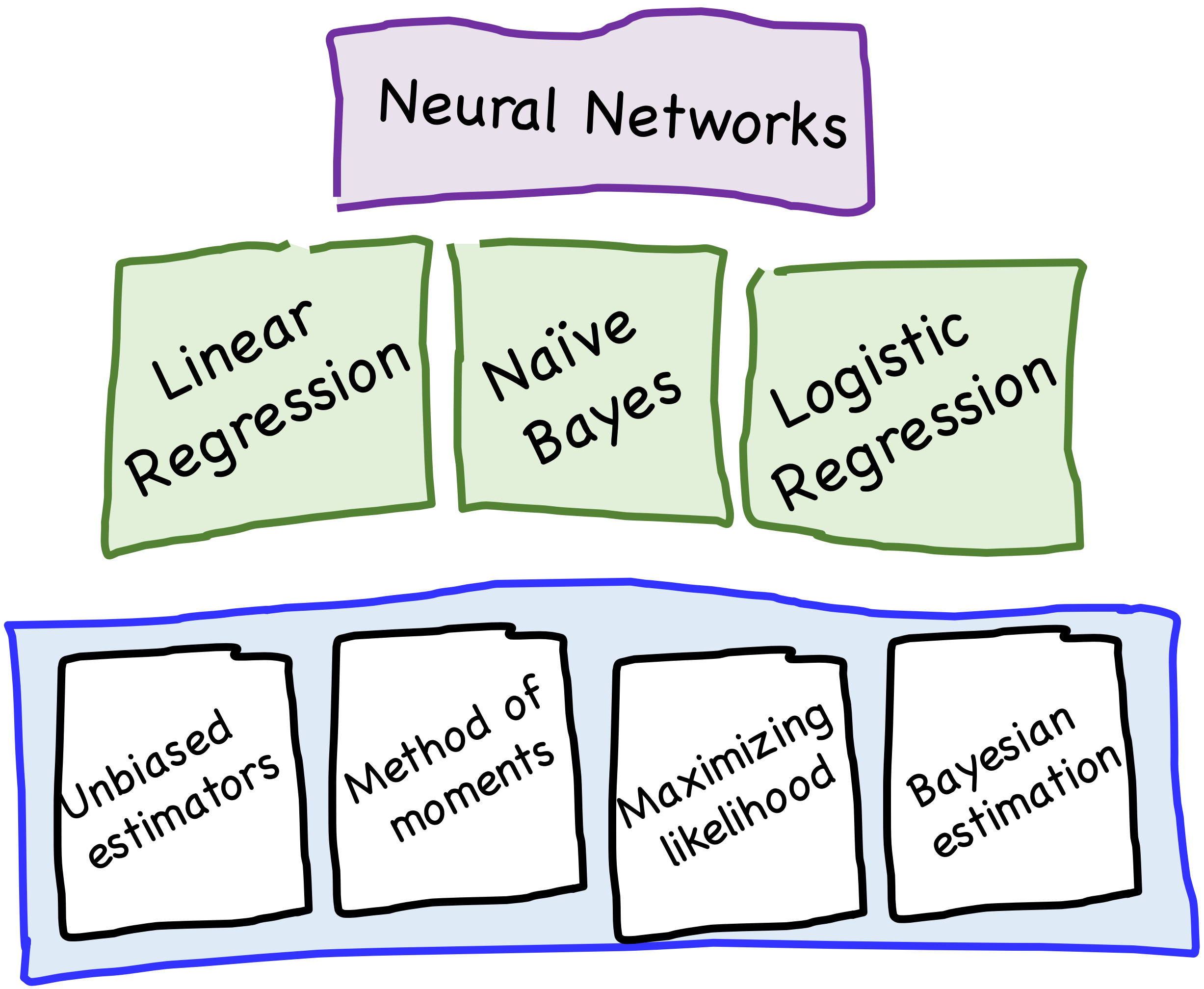


ML: Rooted in probability theory

Our Path


Neural Networks
Linear Regression
Naïve Bayes
Logistic Regression
Parameter Estimation

Our Path


Neural Networks
Linear Regression
Naïve Bayes
Logistic Regression
Unbiased estimators
Method of moments
Maximizing likelihood
Bayesian estimation

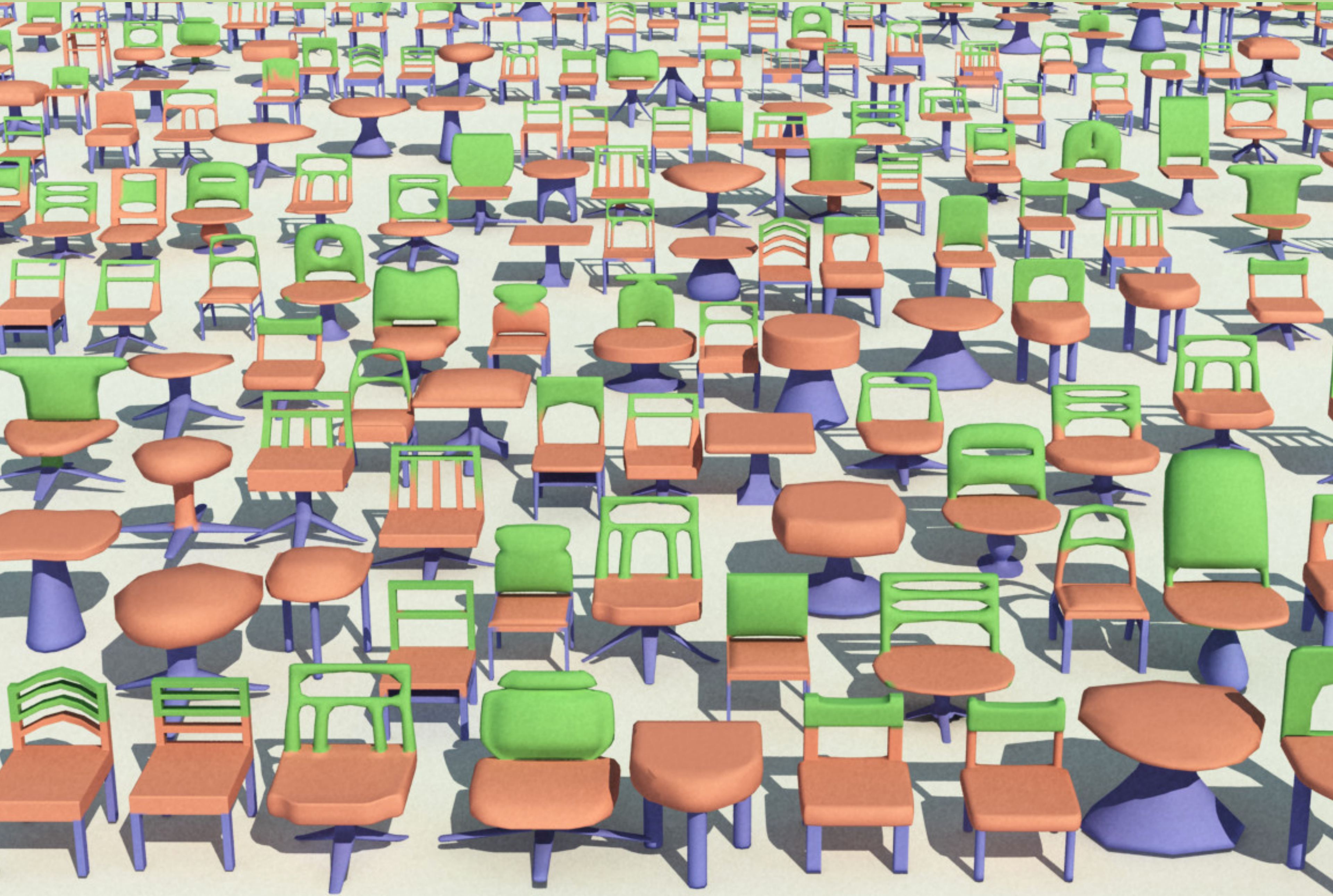
Jump Straight to Neural Networks?

Tensor Flow

Understand the theory to help you debug

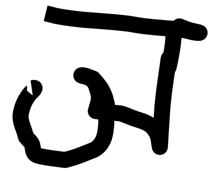
But another reason…

Machine Learning Uses a Lot of Data

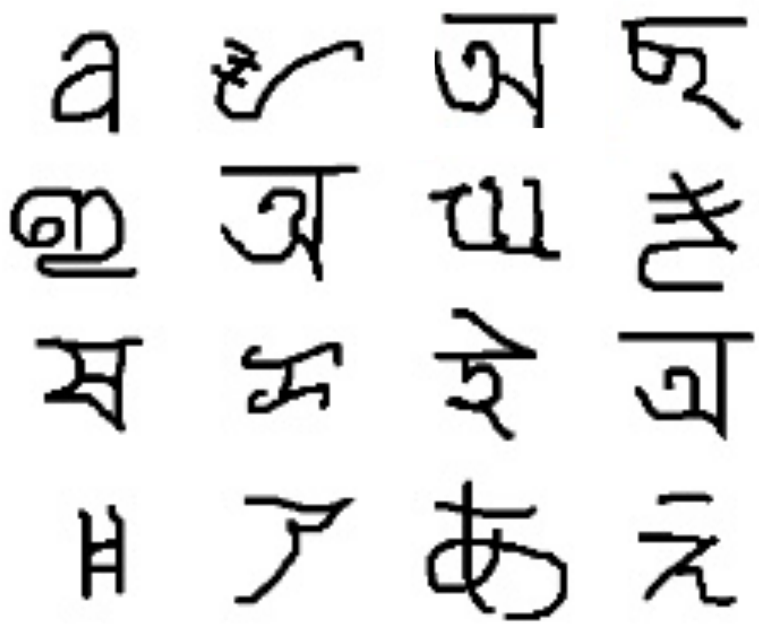


One Shot Learning

Single training example:



Test set:



One Shot Learning

Single training example:

Computers can’t do that.

Understand the theory to push on the grand challenges


Walt Disney
PICTURES

Once upon a time…

…there was parameter estimation

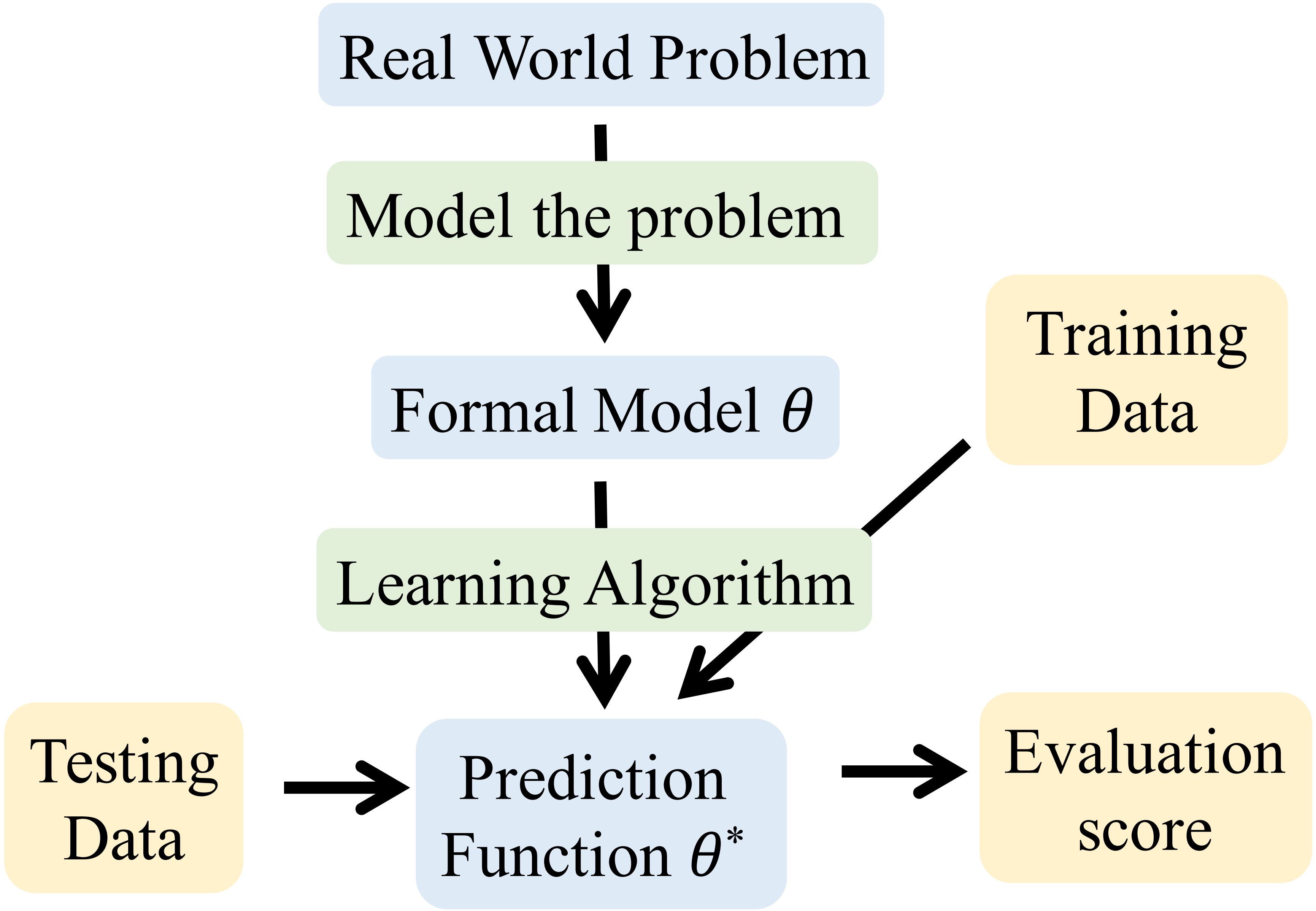
What are Parameters?

- 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

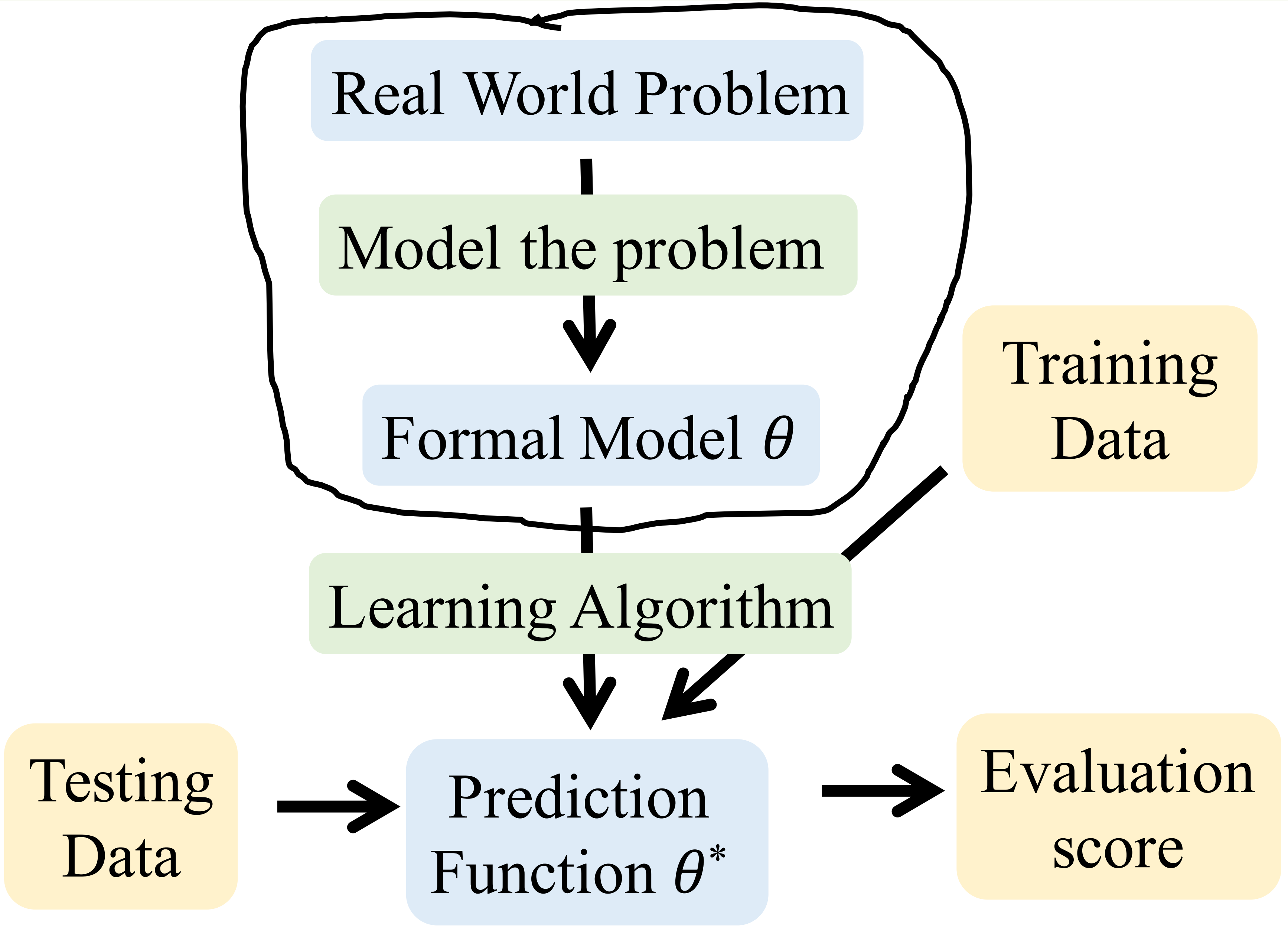
Why Do We Care?

- In real world, don't know "true" parameters
  - But, we do get to observe data
    - E.g., number of times coin comes up heads, lifetimes of disk drives produced, number of visitors to web site per day, etc.
  - Need to estimate model parameters from data
  - "Estimator" is random variable estimating parameter
- Estimate of parameters allows:
  - Better understanding of process producing data
  - Future predictions based on model
  - Simulation of processes

Supervised Learning

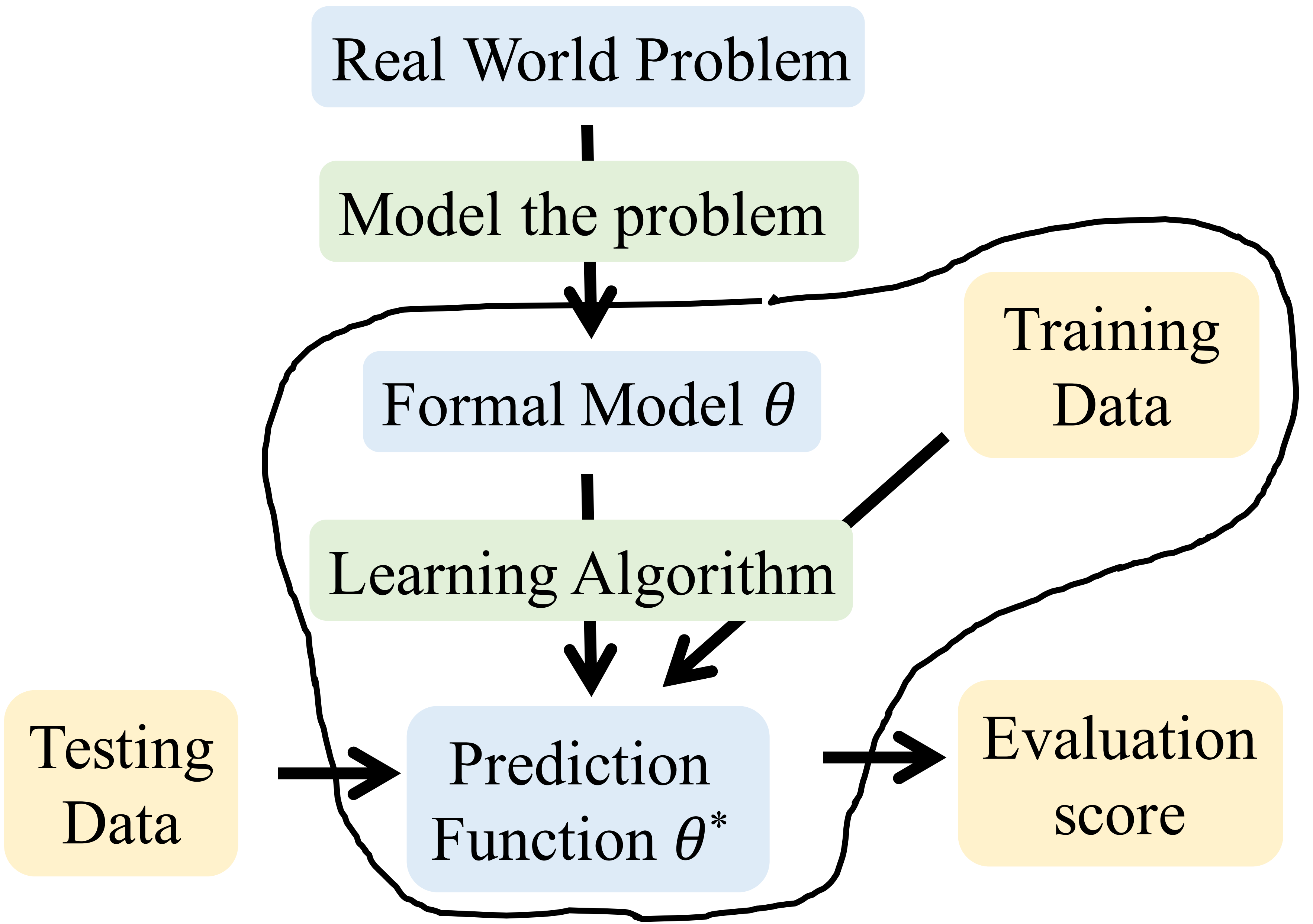


Modelling

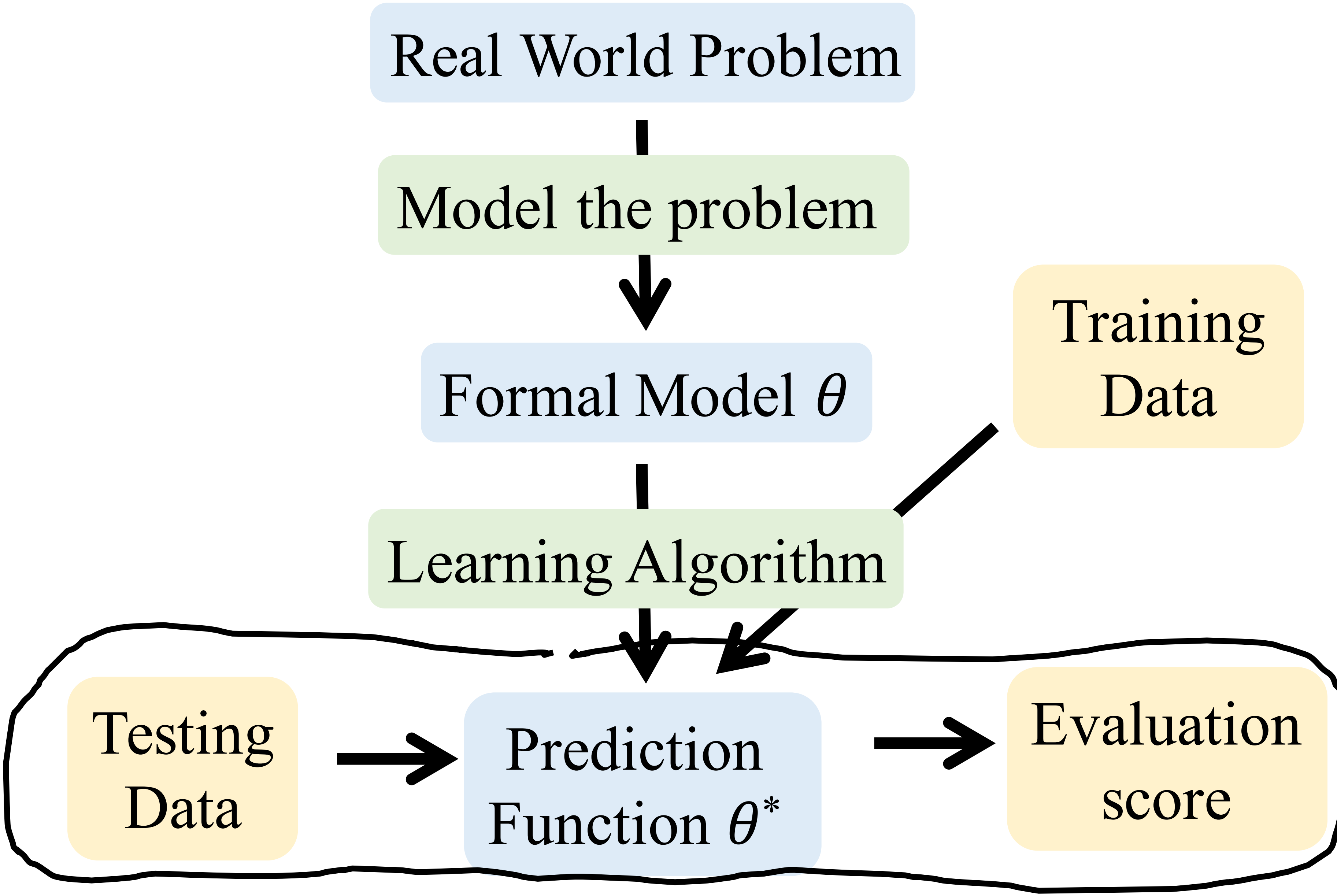


Real World Problem
Model the problem
Formal Model θ
Training Data
Learning Algorithm
Testing Data
Prediction Function θ^*
Evaluation score

Training


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

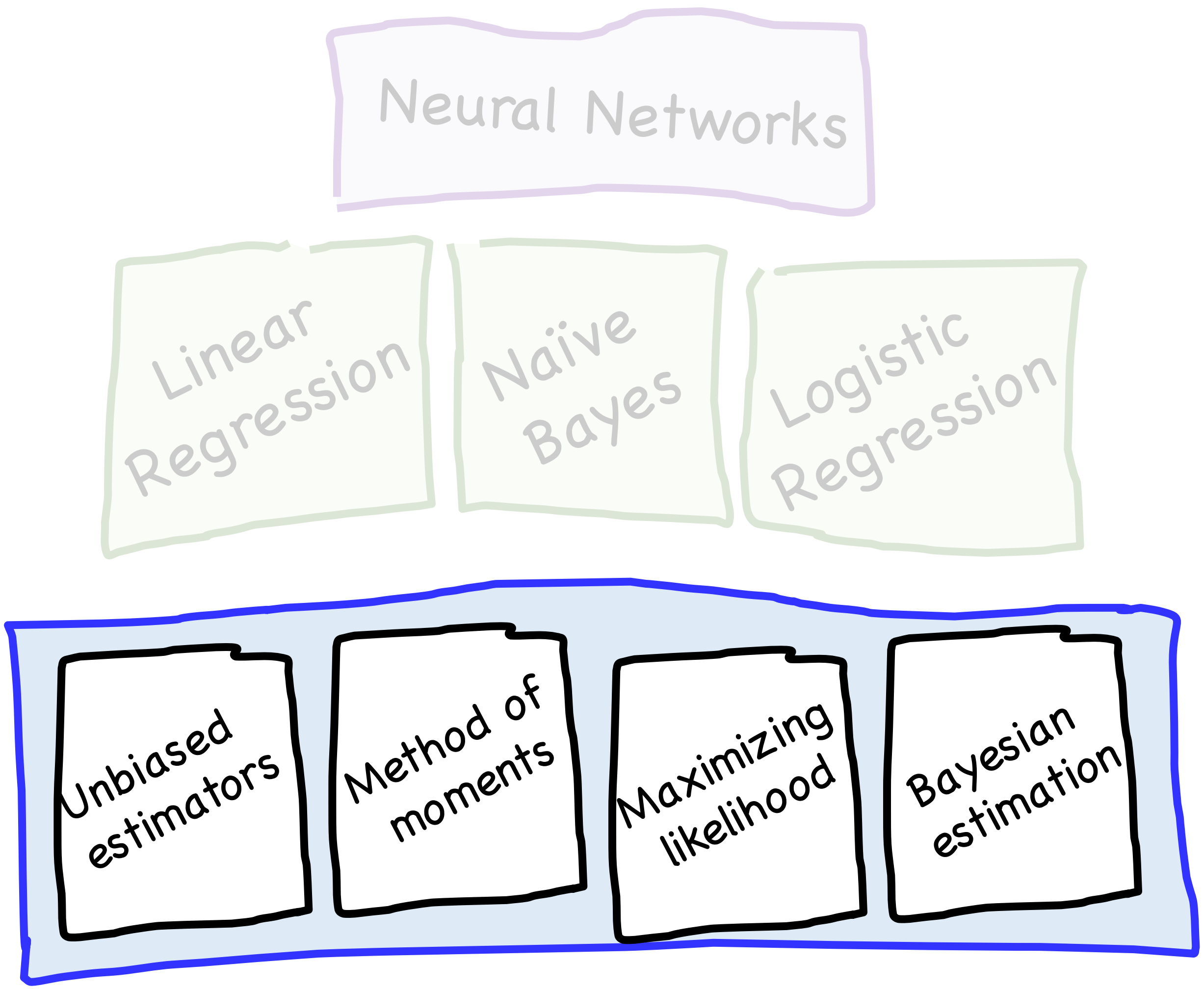
Testing



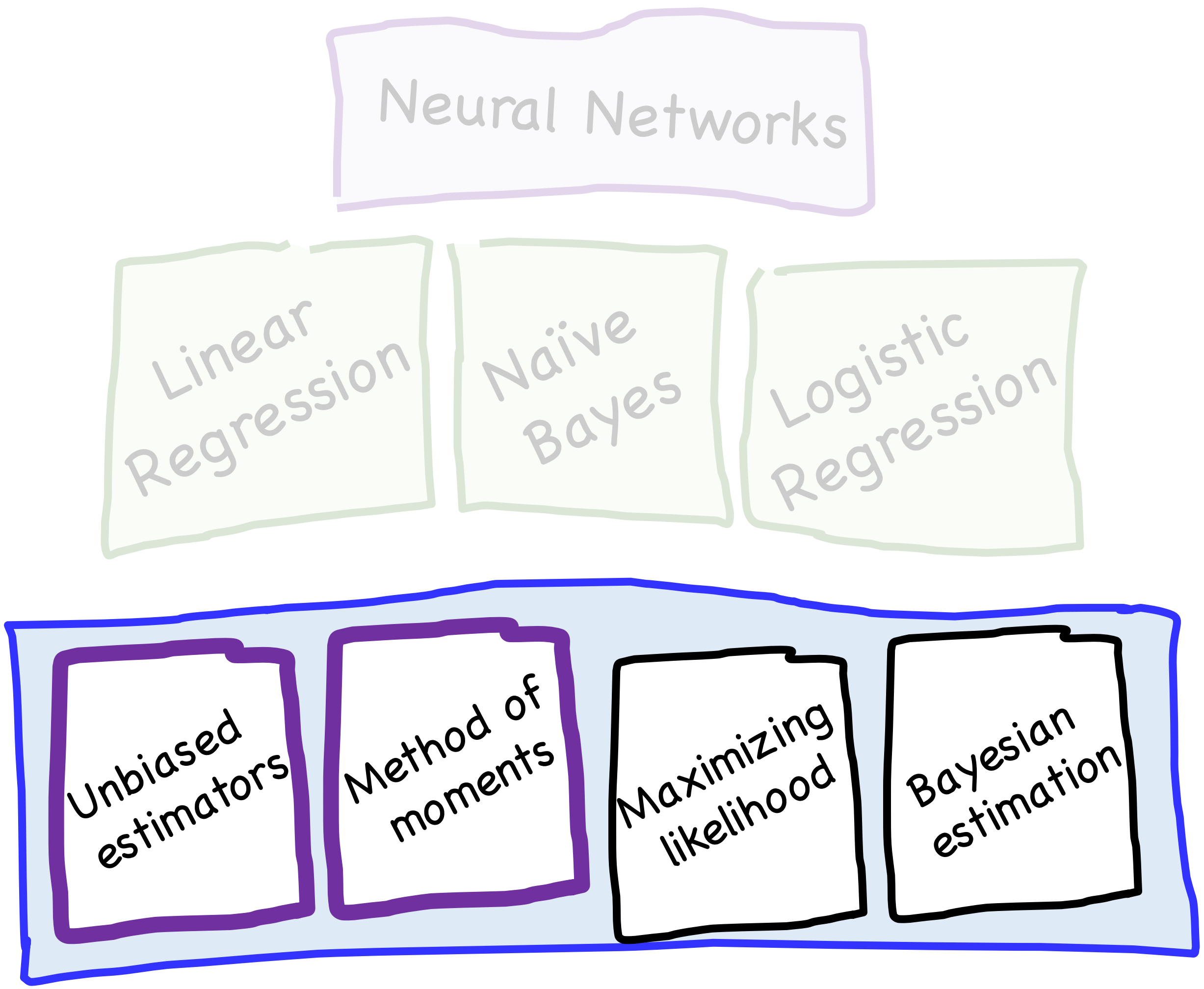
Real World Problem
Model the problem
Formal Model θ
Training Data
Learning Algorithm
Testing Data
Prediction Function θ^*
Evaluation score

Basis for learning from data

Parameter Estimation



Parameter Estimation



Recall Sample Mean + Variance?

- Consider *n* I.I.D. random variables $X_1, X_2, \ldots X_n$
  - X_i have distribution *F* with $E[X_i] = \mu$ and $Var(X_i) = \sigma^2$
  - We call sequence of X_i a **sample** from distribution *F*
  - Recall sample mean: $\bar{X} = \sum_{i=1}^{n} \frac{X_i}{n}$ where $E[\bar{X}] = \mu$

    $\bar{X} \sim N(\mu, \frac{\sigma^2}{n})$ as $n \to \infty$
  - Recall sample variance:

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

Estimate parameters for Bernoulli and Normal

Method of Moments

- Recall: *n*-th moment of distribution for variable X:

$$m_n = E[X^n]$$

- Consider I.I.D. random variables $X_1, X_2, \ldots, X_n$
  - X_i have distribution *F*
  - Let $\hat{m}_1 = \frac{1}{n}\sum_{i=1}^{n} X_i \qquad \hat{m}_2 = \frac{1}{n}\sum_{i=1}^{n} X_i^{\,2} \quad \ldots \quad \hat{m}_k = \frac{1}{n}\sum_{i=1}^{n} X_i^{\,k}$
  - $\hat{m}_i$ are called the "sample moments"
    - Estimates of the moments of distribution based on data
- Method of moments estimators
  - Estimate model parameters by equating "true" moments to sample moments: $m_i \approx \hat{m}_i$

Examples of Methods of Moments

- Recall the sample mean: $\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i = \hat{m}_1 \approx E[X]$
  - This is method of moments estimator for E[X]
- Method of moments estimator for variance
  - Estimate second moment: $\hat{m}_2 = \frac{1}{n}\sum_{i=1}^{n} X_i^{\,2}$
  - $\text{Var}(X) = E[X^2] - (E[X])^2$
  - Estimate: $\text{Var}(X) \approx \hat{m}_2 - (\hat{m}_1)^2$

$$= \left(\frac{1}{n}\sum_{i=1}^{n} X_i^{\,2}\right) - \overline{X}^2 = \frac{1}{n}\sum_{i=1}^{n} X_i^{\,2} - \frac{1}{n}\sum_{i=1}^{n} \overline{X}^2 = \frac{\sum_{i=1}^{n}(X_i^{\,2} - \overline{X}^2)}{n}$$

  - Recall sample variance:

$$S^2 = \sum_{i=1}^{n} \frac{(X_i - \overline{X})^2}{n-1} = \sum_{i=1}^{n} \frac{(X_i^{\,2} - 2X_i\overline{X} + \overline{X}^2)}{n-1} = \frac{\sum_{i=1}^{n}(X_i^{\,2} - \overline{X}^2)}{n-1} = \frac{n}{n-1}(\hat{m}_2 - (\hat{m}_1)^2)$$

Small Samples = Problems

- What is difference between sample variance and MOM estimate for variance?
  - Imagine you have a sample of size $n = 1$
  - What is sample variance?

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

  - i.e., don't really know variability of data
  - What is MOM estimate of variance?

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

  - i.e., have complete certainty about distribution!
    - There is no variance

Estimator Bias

- Bias of estimator: $E[\hat{\theta}] - \theta$
  - When bias = 0, we call the estimator "unbiased"
  - A biased estimator is not necessarily a bad thing
  - Sample mean $\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$ is unbiased estimator
  - Sample variance $S^2 = \sum_{i=1}^{n} \frac{(X_i - \bar{X})^2}{n-1}$ is unbiased estimator
  - MOM estimator of variance = $\frac{n-1}{n} S^2$ is biased
    - Asymptotically less biased as *n* → ∞
  - For large *n*, either sample variance or MOM estimate of variance is fine.

Estimator Consistency

- Estimator "consistent": $\lim_{n \to \infty} P(|\hat{\theta} - \theta| < \varepsilon) = 1$ for $\varepsilon > 0$
  - As we get more data, estimate should deviate from true value by at most a small amount
  - This is actually known as "weak" consistency
  - Note similarity to weak law of large numbers:

$$\lim_{n \to \infty} P(|\bar{X} - \mu| \geq \varepsilon) \to 0$$

  - Equivalently:

$$\lim_{n \to \infty} P(|\bar{X} - \mu| < \varepsilon) \to 1$$

  - Establishes sample mean as consistent estimate for μ
  - Generally, MOM estimates are consistent

Method of Moments with Bernoulli

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

$$p = E[X_i] \approx \hat{m}_1 = \overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i = \hat{p}$$

  - Can use estimate of *p* for X ~ Bin(*n*, *p*)
  - If you know what *n* is, you don’t need to estimate that

Conditional Bernoulli

- Consider I.I.D. random variables $X_1|Y$, $X_2|Y$, ..., $X_n|Y$
  - $X_i|Y \sim \text{Ber}(p)$
- Estimate p

Count of successes

$$p = E[X_i|Y] \approx \hat{m}_1 = \bar{X}|Y = \frac{1}{n}\sum_{i=1}^{n} X_i|Y = \hat{p}$$

Count of samples

Isn’t that the same as unbiased estimator?

Yes. For Bernoulli.

Conditional Bernoulli

- Let S be survived, X is fare paid in British Pounds
- P(S = true | X > 100)?
- Consider I.I.D. random variables $S_1|X$, $S_2|X$, ..., $S_n|X$
  - $S_i|X \sim \text{Ber}(p)$
- Estimate p

Count of successes

$$p = E[S_i|X] \approx \hat{m}_1 = \bar{S}|X = \frac{1}{n}\sum_{i=1}^{n} S_i|X = \hat{p}$$

$$= \frac{39}{53} = 0.74$$

Count of samples

Technically Machine Learning

But really it’s a building block

Method of Moments with Poisson

- Consider I.I.D. random variables $X_1, X_2, ..., X_n$
  - $X_i \sim \text{Poi}(\lambda)$
- Estimate λ

$$\lambda = E[X_i] \approx \hat{m}_1 = \overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i = \hat{\lambda}$$

  - But note that for Poisson, $\lambda = \text{Var}(X_i)$ as well!
  - Could also use method of moments to estimate:

$$\lambda = E[X_i^2] - E[X_i]^2 \approx \hat{m}_2 - (\hat{m}_1)^2 = \frac{\sum_{i=1}^{n}(X_i^2 - \overline{X}^2)}{n} = \hat{\lambda}$$

  - Usually, use first moment estimate
  - More generally, use the one that's easiest to compute

Method of Moments with Normal

- Consider I.I.D. random variables $X_1, X_2, ..., X_n$
  - $X_i \sim N(\mu, \sigma^2)$
- Estimate μ

$$\mu = E[X_i] \approx \hat{m}_1 = \overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i = \hat{\mu}$$

- Now estimate σ^2

$$\sigma^2 \approx \hat{m}_2 - (\hat{m}_1)^2$$

$$= \left(\frac{1}{n}\sum_{i=1}^{n} X_i^{\,2}\right) - \hat{\mu}^2 = \frac{1}{n}\sum_{i=1}^{n} X_i^{\,2} - \frac{1}{n}\sum_{i=1}^{n} \overline{X}^2 = \frac{\sum_{i=1}^{n}(X_i^{\,2} - \overline{X}^2)}{n}$$

Method of Moments with Uniform

- Consider I.I.D. random variables $X_1, X_2, \ldots, X_n$
  - $X_i \sim \text{Uni}(\alpha, \beta)$
  - Estimate mean:

$$\mu \approx \hat{m}_1 = \frac{1}{n}\sum_{i=1}^{n} X_i = \hat{\mu}$$

  - Estimate variance:

$$\sigma^2 \approx \hat{m}_2 - (\hat{m}_1)^2 = \frac{\sum_{i=1}^{n}(X_i^{\,2} - \overline{X}^2)}{n} = \hat{\sigma}^2$$

  - For $\text{Uni}(\alpha, \beta)$, know that: $\mu = \dfrac{\alpha+\beta}{2}$ and $\sigma^2 = \dfrac{(\beta-\alpha)^2}{12}$
  - Solve (two equations, two unknowns):
    - Set $\beta = 2\mu - \alpha$, substitute into formula for σ^2 and solve:

$$\hat{\alpha} = \overline{X} - \sqrt{3}\hat{\sigma} \quad \text{and} \quad \hat{\beta} = \overline{X} + \sqrt{3}\hat{\sigma}$$

Can we think of parameters as random variables?