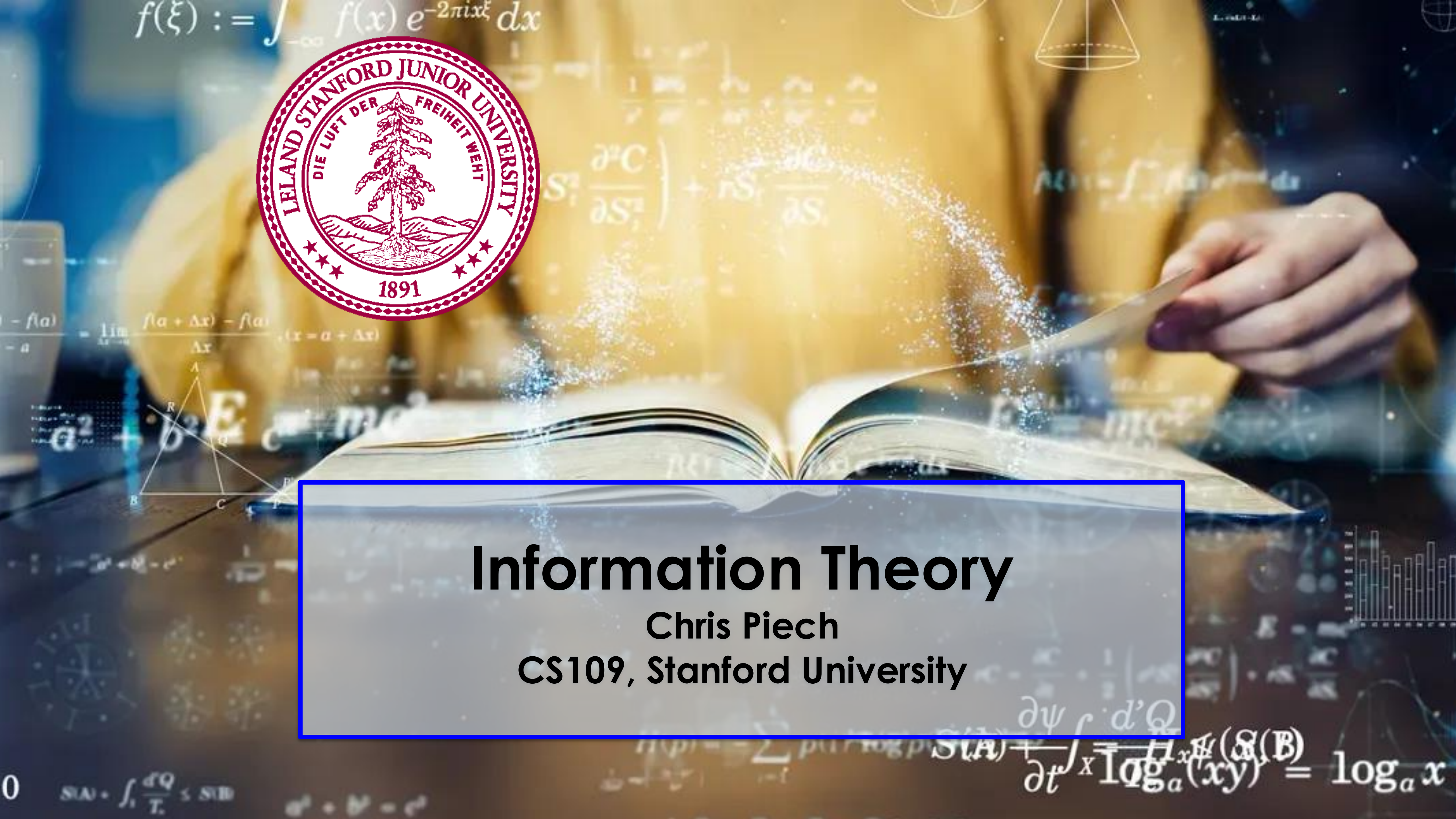


Information Theory

Chris Piech

CS109, Stanford University

Learning Goals

1. Calculate information gain
2. Make choices that maximize information gain
3. Numerically score how similar two distributions are

Uncertainty Theory

Beta Distributions

Thompson Sampling

Adding Random Vars

Central Limit Theorem

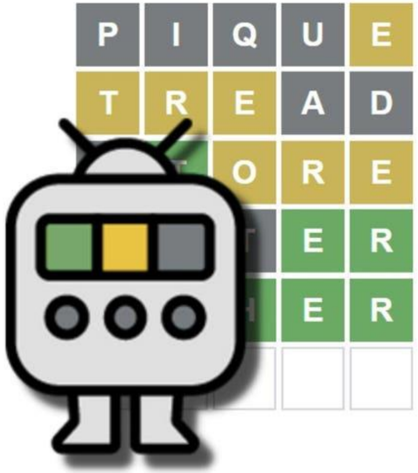
Sampling

Bootstrapping

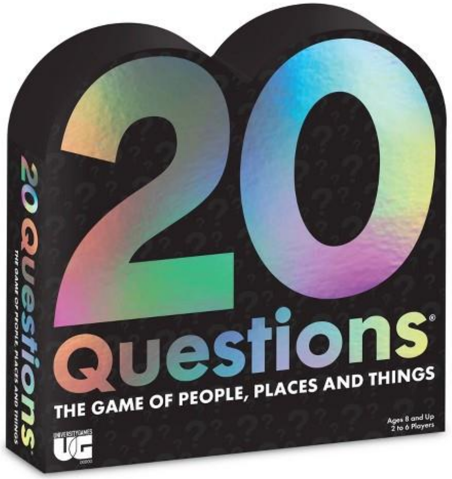
Algorithmic Analysis

Information Theory

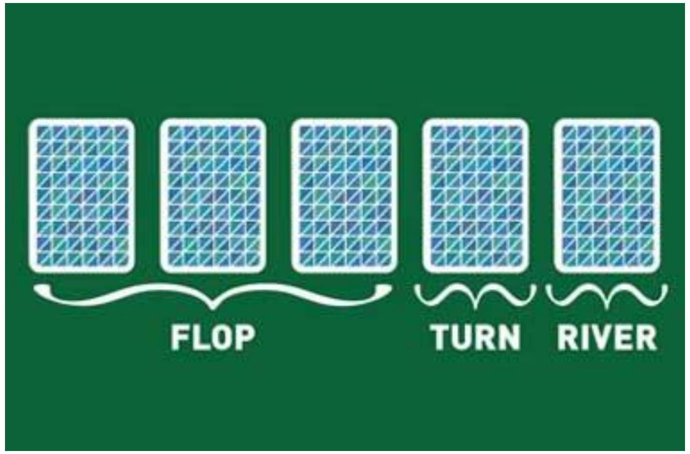
WorldeBot



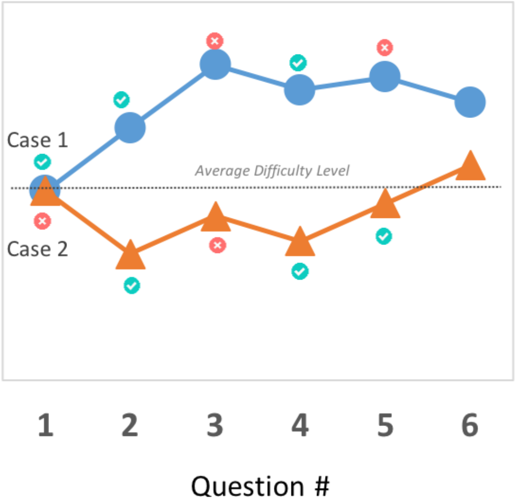
Decision Trees



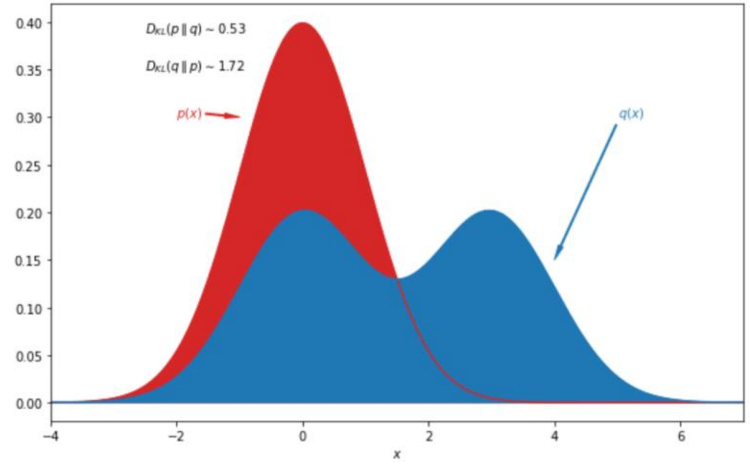
Value of Info in Poker



Adaptive Tests



Comparing Distributions



Compression of Data

Plot Line

1. We want to **chose questions** in **think of an animal** (Let X be the animal random var)

2. Idea! Select the question which most **reduces our "uncertainty"** in X

3. We can **measure "uncertainty"** as "expected amount of surprise when we find out X"

4. We can **measure "surprise"** in an assignment as log 1/P(X=x)

5. This measure of **uncertainty of X** is **super helpful** for lots of problems!

Review

Expectation

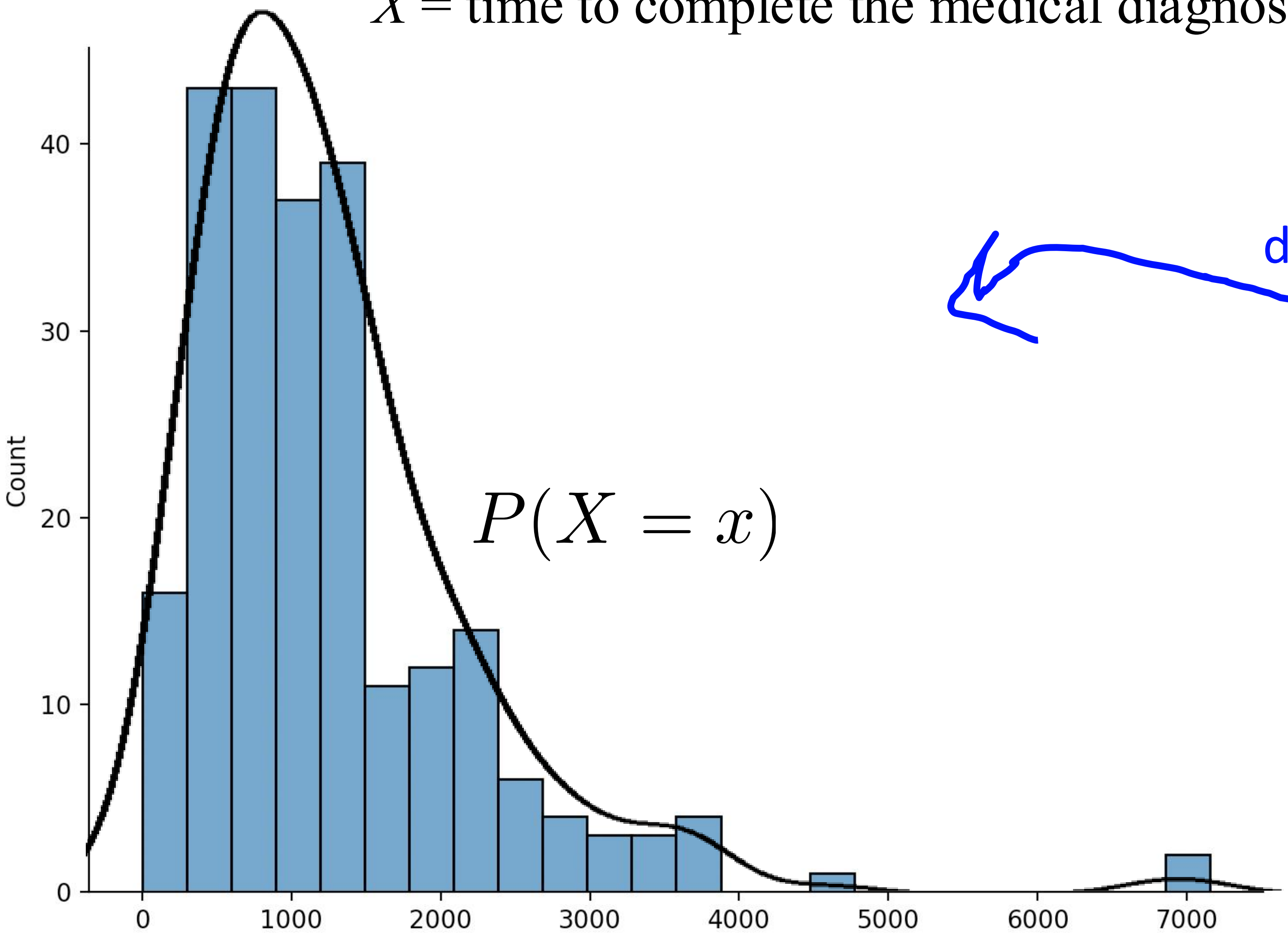
The probability that X takes on that value

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

All the values that X can take on

Limitation of Expectation

X = time to complete the medical diagnosis problem (in seconds)



$$P(X = x)$$

Summarize the distribution in one number

$$\mathrm{E}[X] = 1260$$

Expectation of a Function

Law of unconscious statistician

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

So for example…

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

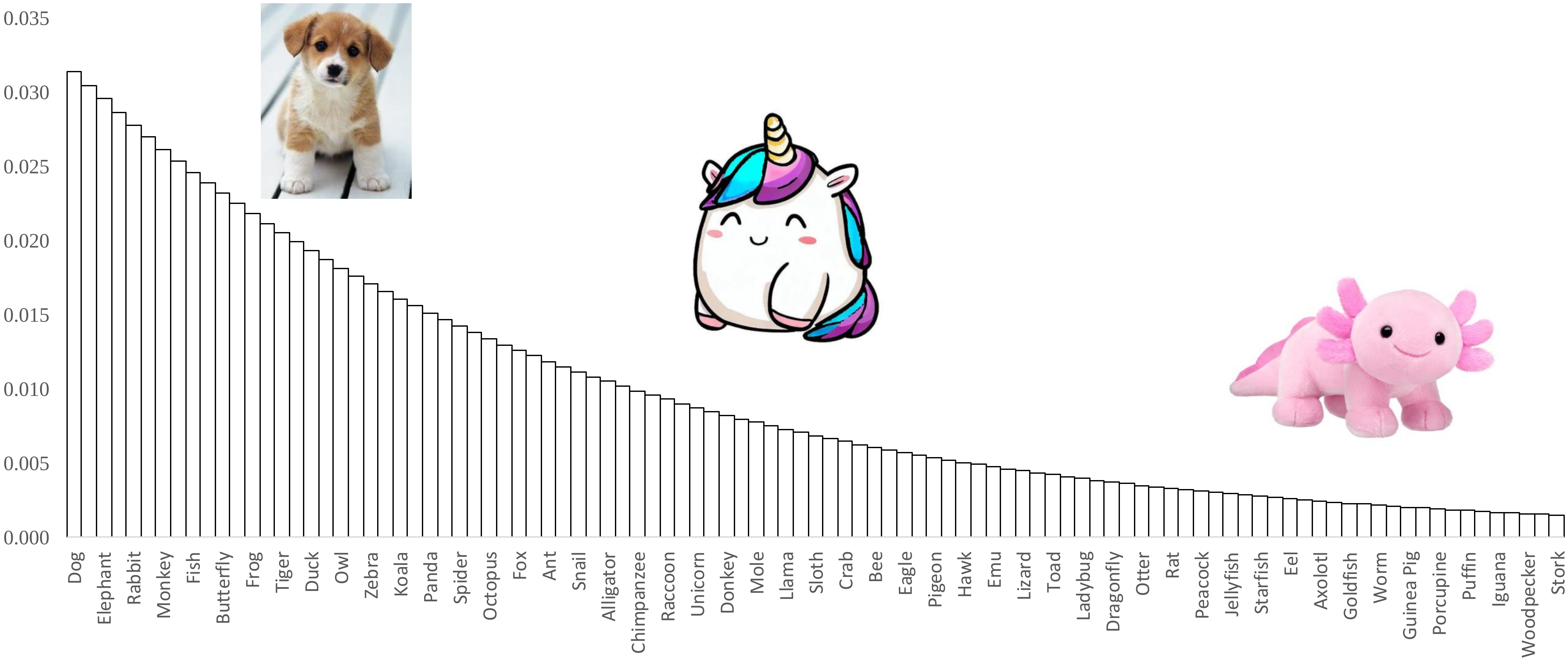
Def: Law of Total Expectation

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

End Review

I am thinking of an **animal**…

I am thinking of an animal



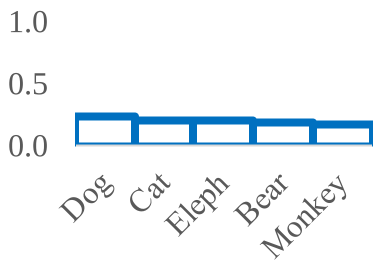
What is the **best question** to ask?

Question Choosing Algorithms

Algorithm	Average Questions	Standard Error of the Mean
Random Questions	61.34	1.41
Binary Search	6.79	0.02
Information Theory Search	5.33	0.04

Hey what's that?

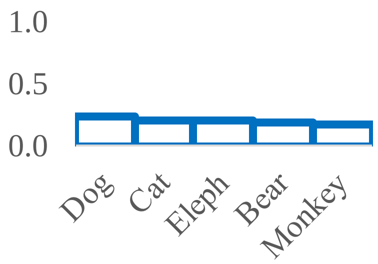
Which Question is Better?



Is it a pet?

Is it a Dog?

Which Question is Better?



Is it a pet?

yes
p = 0.44

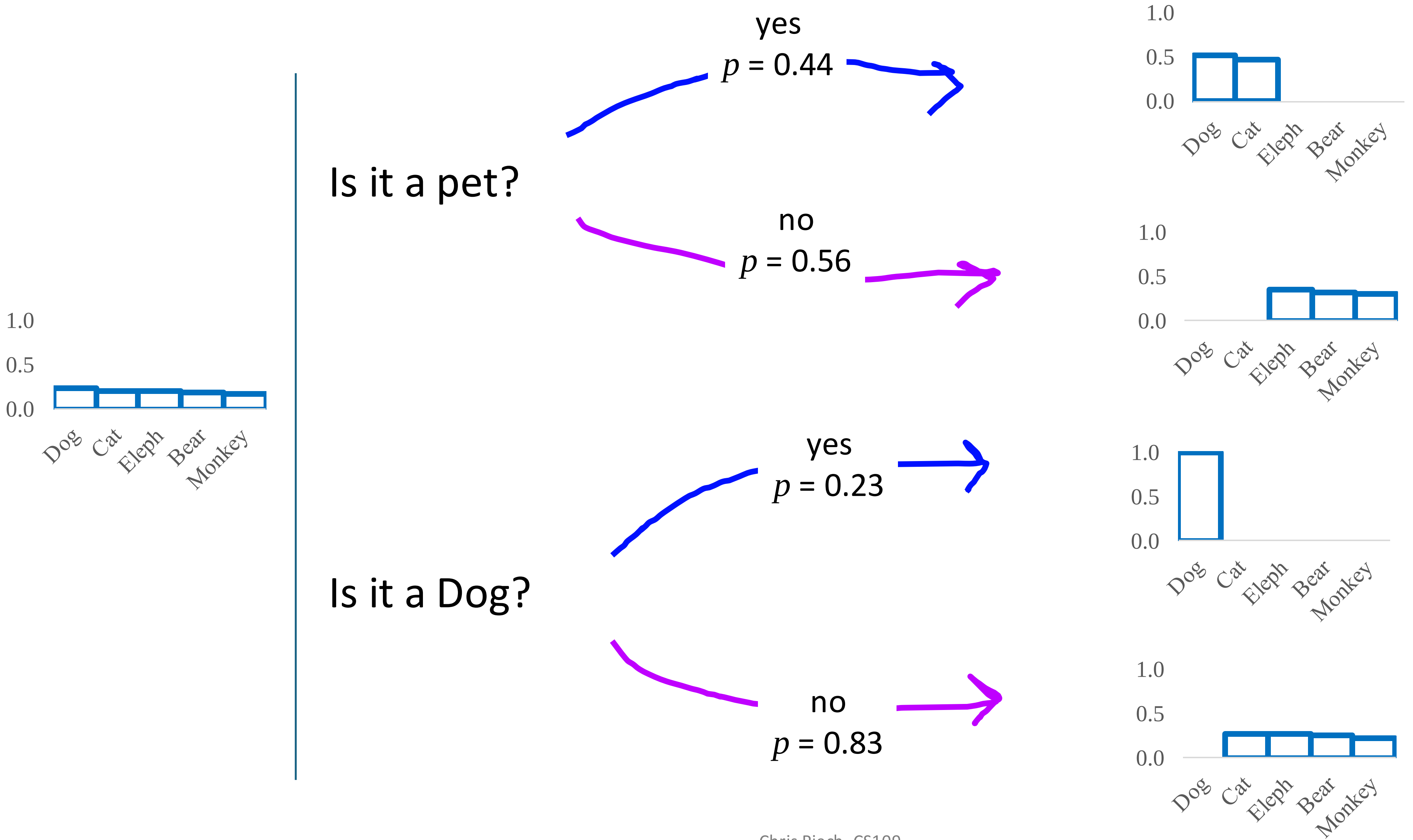
no
p = 0.56

Is it a Dog?

yes
p = 0.23

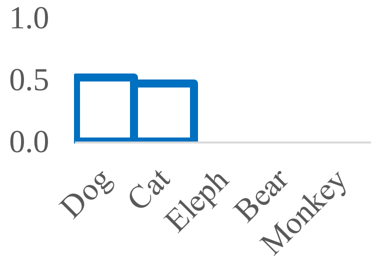
no
p = 0.83

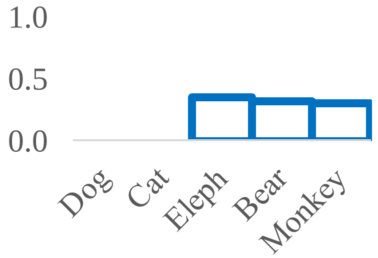
Which Question is Better?

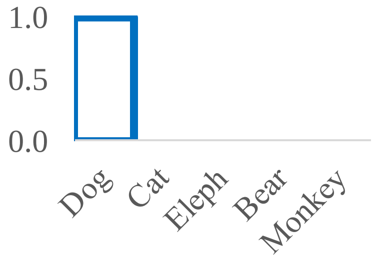


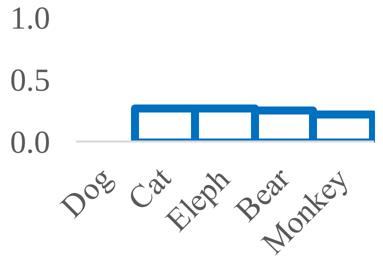
Which Question is Better

Can you "score" how **bad** each of these PMFs are?





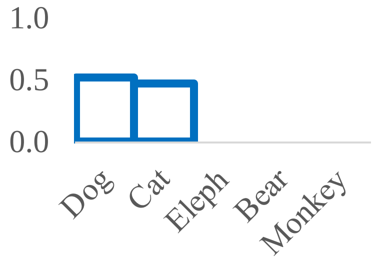


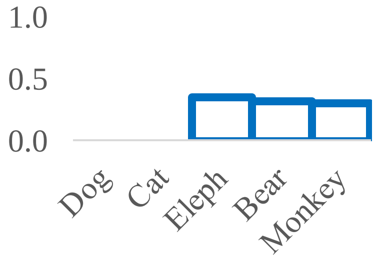


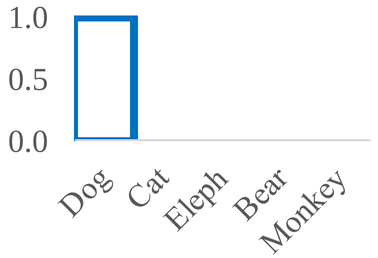
Which Question is Better

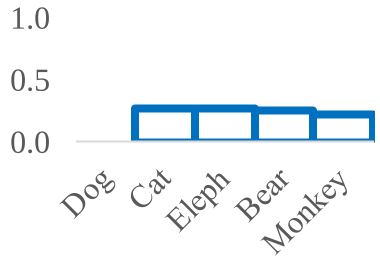
uncertain

Can you "score" how ~~bad~~ each of these PMFs are?





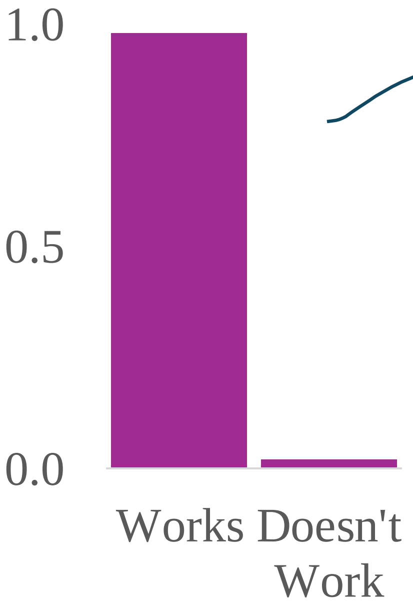




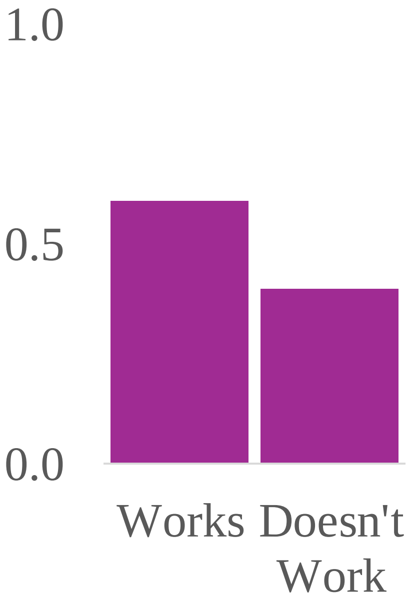
What we really need is a **measure** of our **uncertainty** in a random variable

Uncertainty of a Random Variable

Let X be any random variable. We can calculate a statistic, "**Uncertainty**" to express how much we don't know about X



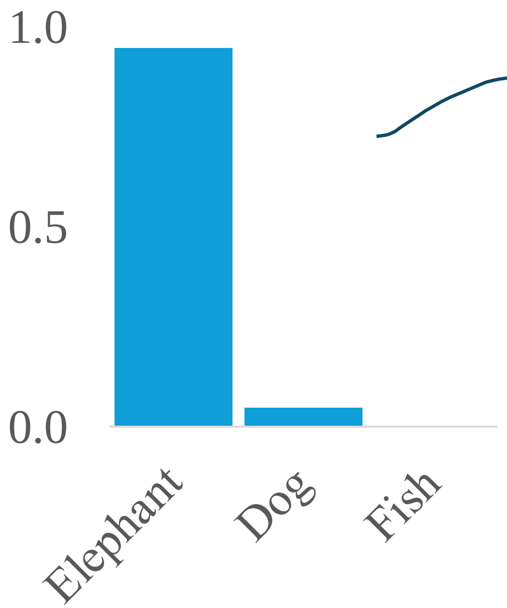
Low uncertainty



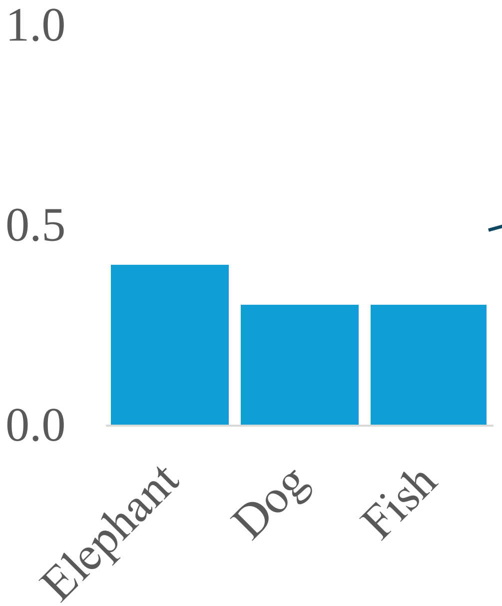
High uncertainty

Uncertainty of a Random Variable

Let X be any random variable. We can calculate a statistic, "**Uncertainty**" to express how much we don't know about X



Low uncertainty

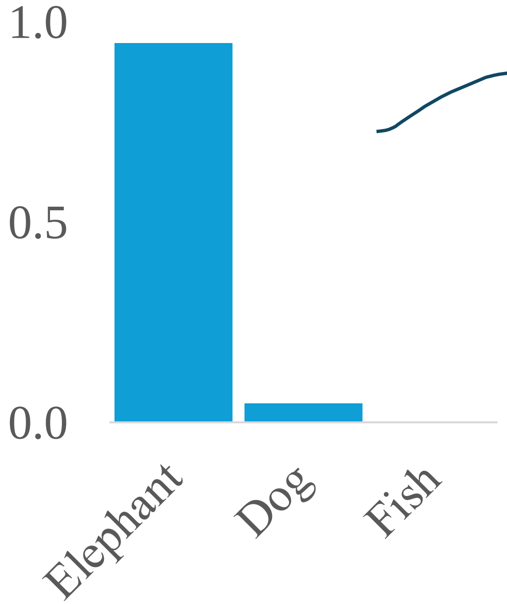


High uncertainty

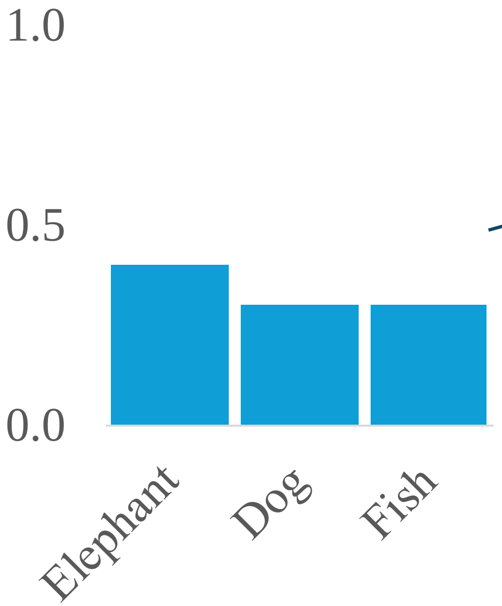
Uncertainty of a Random Variable

Let X be any random variable. We can calculate a statistic, "**Uncertainty**" to express how much we don't know about X

$$\text{Uncertainty}(X) = \text{Expected "Surprise" when I observe X}$$



Low uncertainty



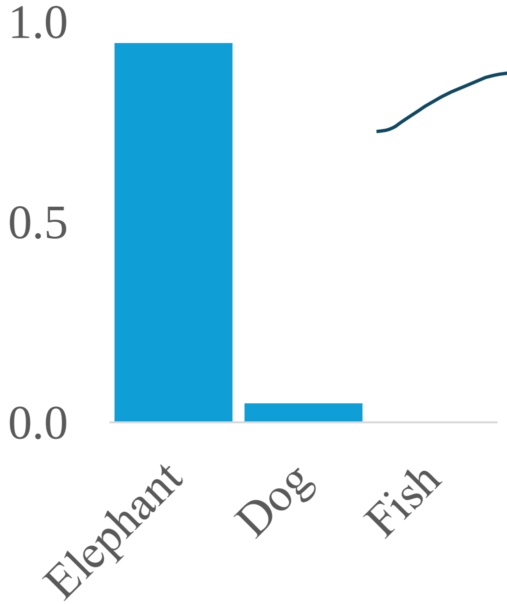
High uncertainty

Uncertainty of a Random Variable

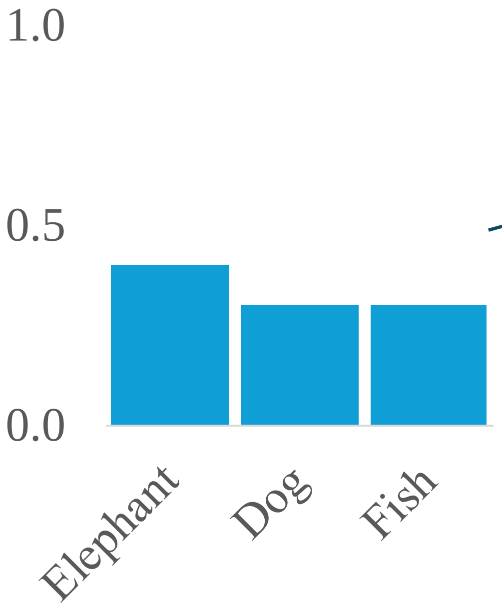
Let X be any random variable. We can calculate a statistic, "**Uncertainty**" to express how much we don't know about X

$$\text{Uncertainty}(X) = \sum_{x \in X} \text{Surprise}(X = x) \cdot P(X = x)$$

Uncertainty is expected Surprise



Low uncertainty



High uncertainty

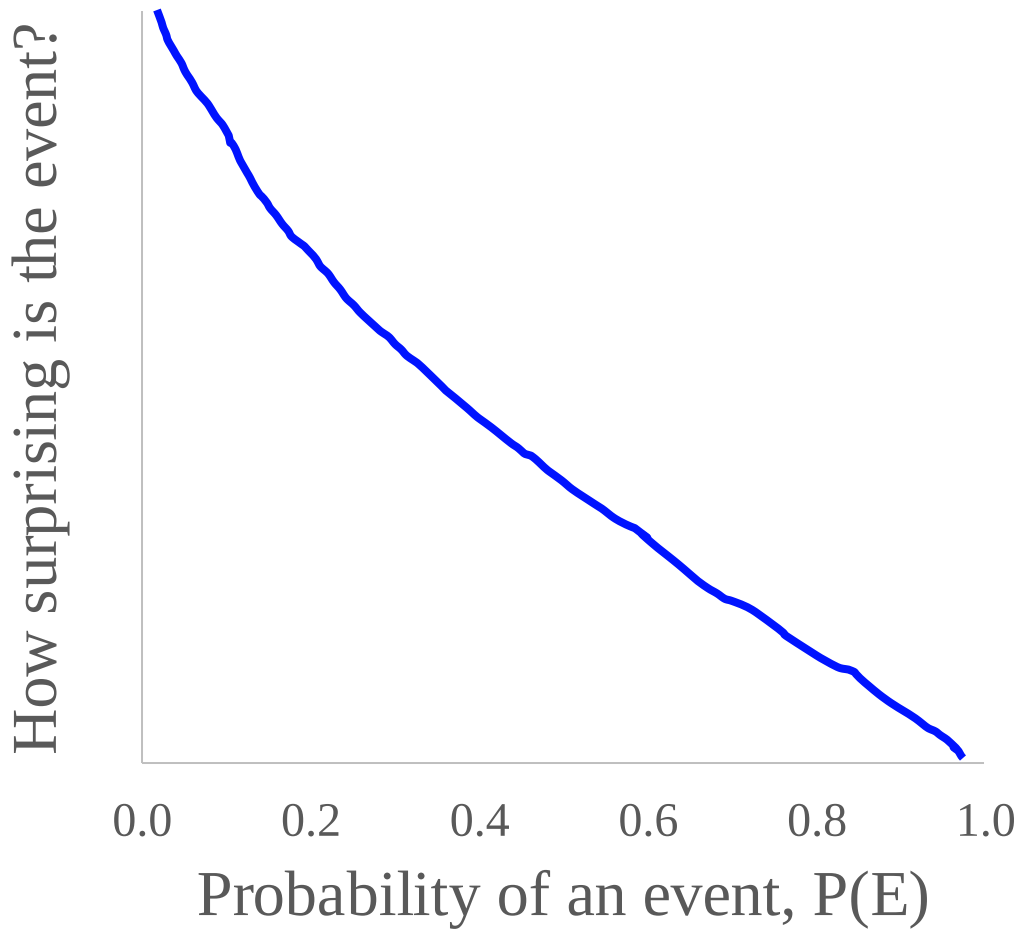
Ok, but then what is our measure of **Surprise**?

Surprise of an Event

High probability events are **not surprising**

Low probability events are **surprising**

Relationship should be monotonic



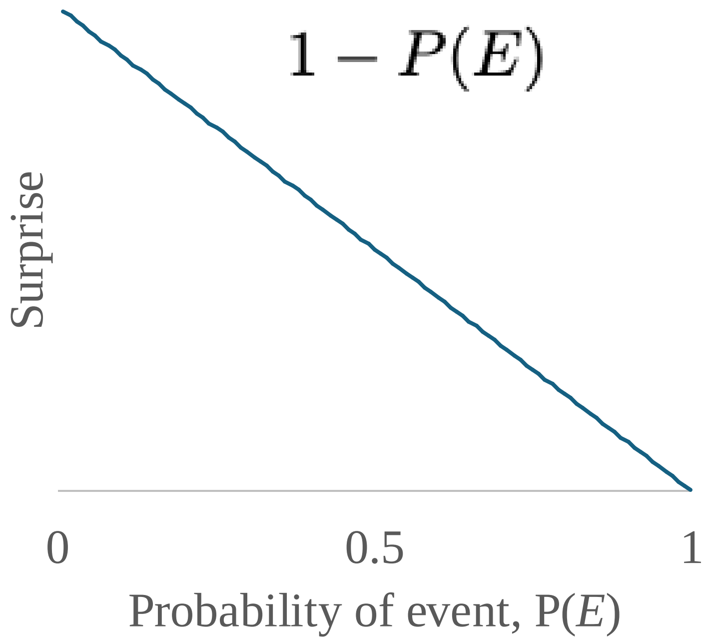
Surprise of an Event

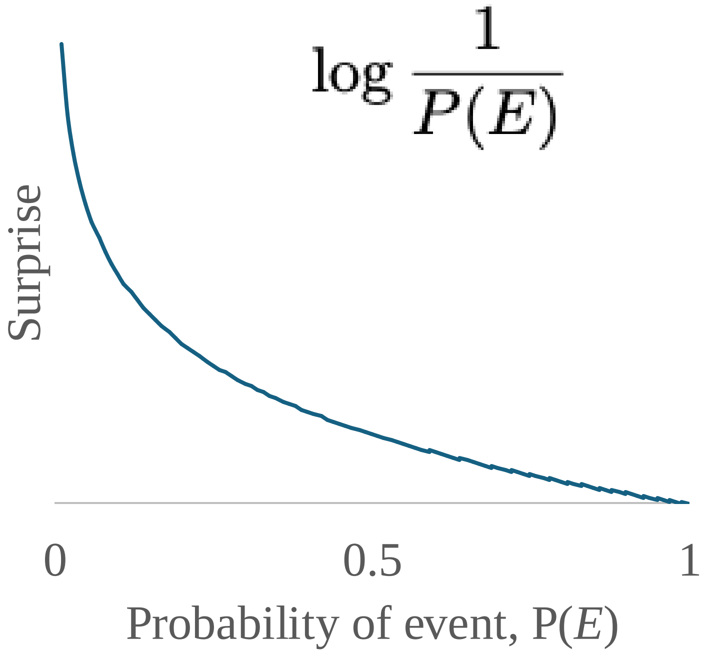
High probability events are **not surprising**

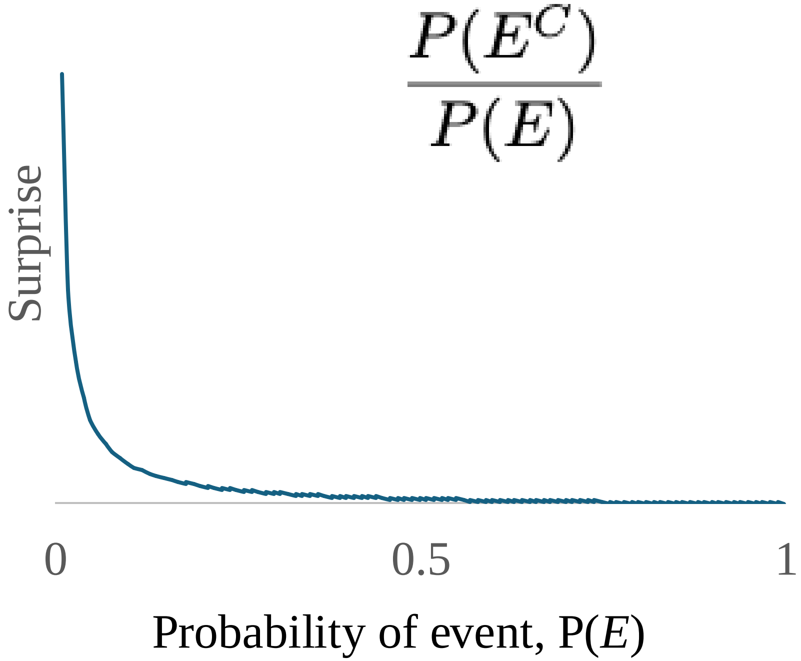
Low probability events are **surprising**

Relationship should be monotonic

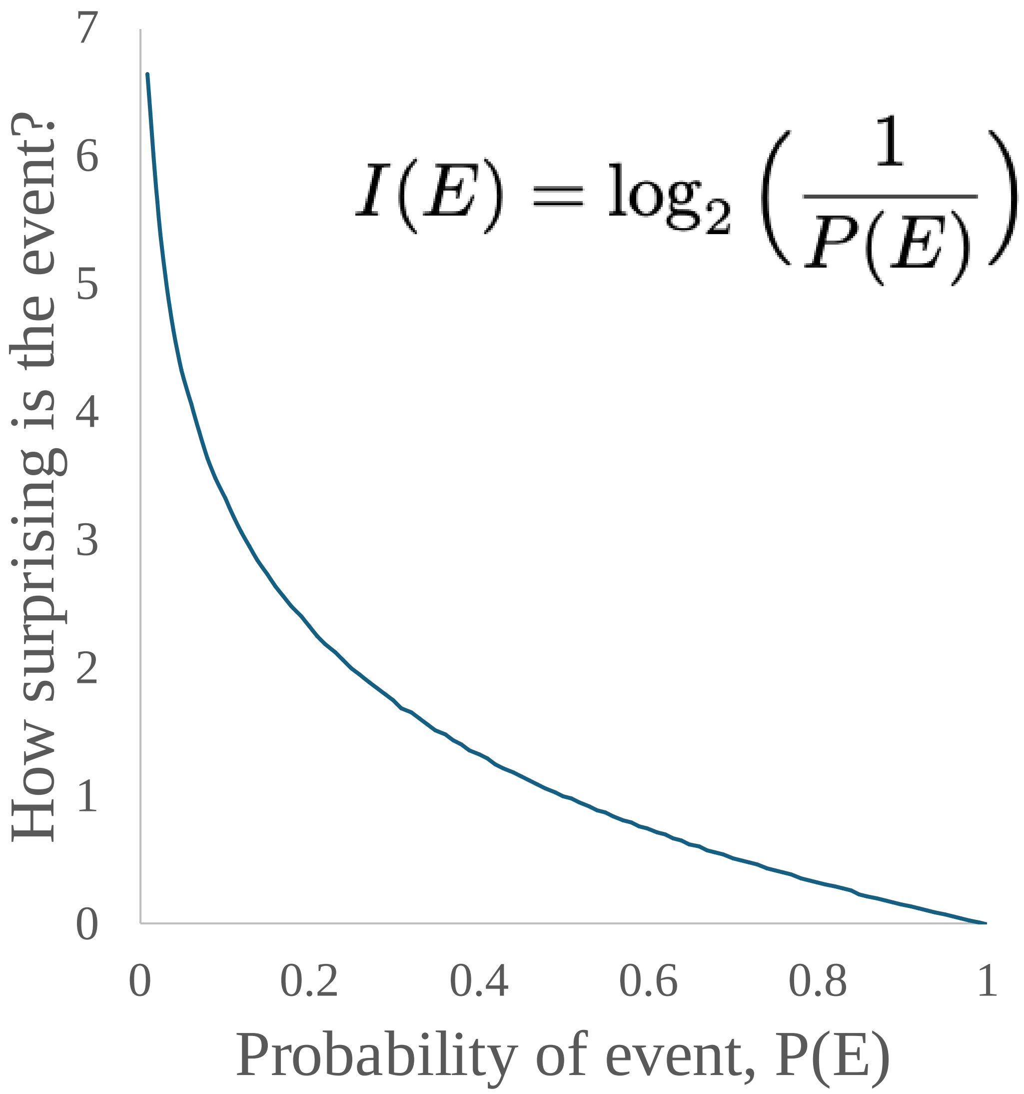
Here are three reasonable options







Surprise of an Event, I(E)

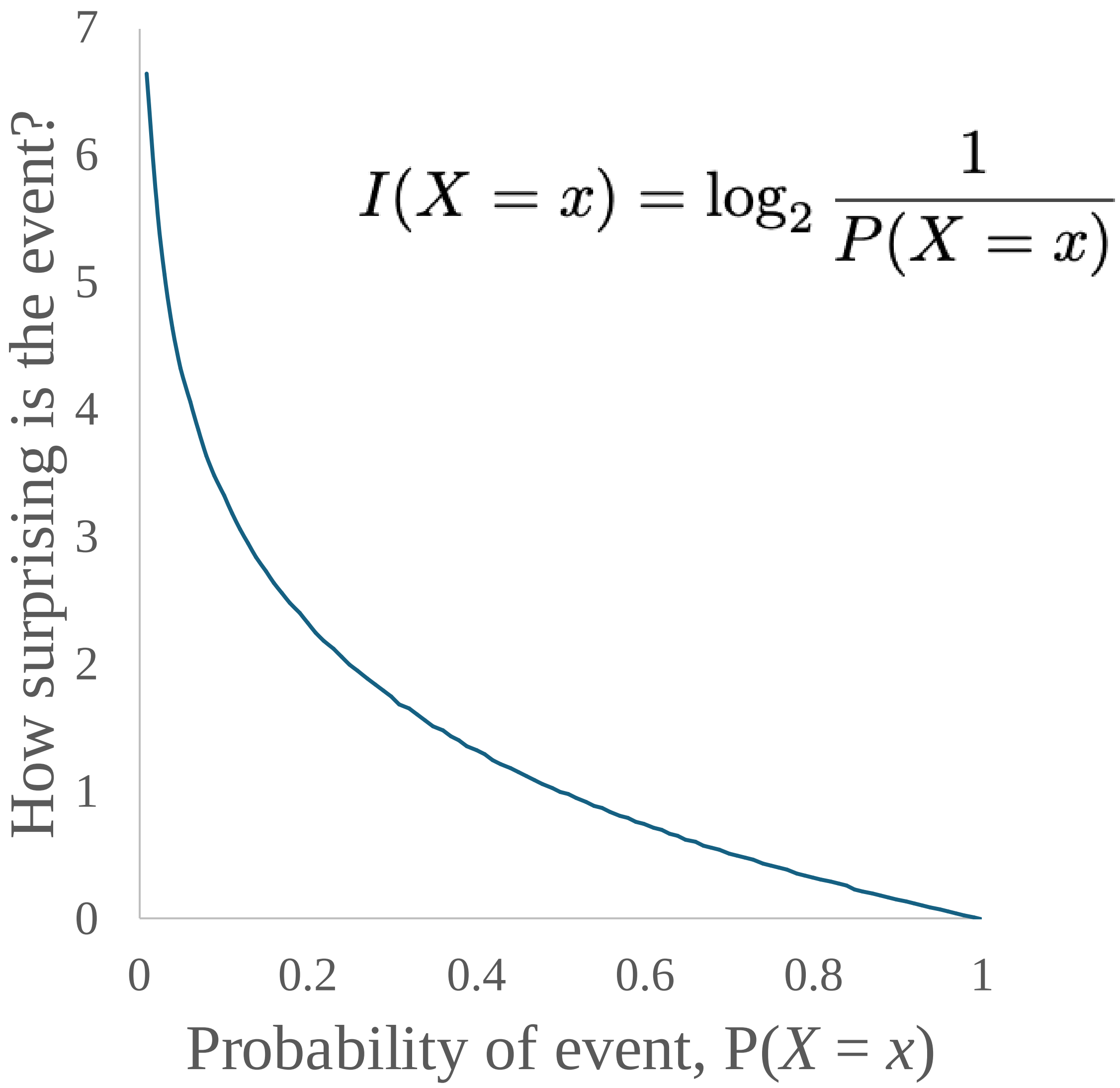


$$I(E) = \log_2\left(\frac{1}{P(E)}\right)$$

Probability of Event $P(E)$	Surprise of Event $I(E)$
1	0
½	1
1/4	2
1/8	3
1/16	4
1/32	5
1/64	6

$I(E)$ stands for "Information Content" aka "Surprisal" aka "Self-Information"

Surprise of an Event, $\mathrm{I}(X = x)$



Probability of Event $P(X = x)$	Surprise of Event $I(X = x)$
1	0
½	1
1/4	2
1/8	3
1/16	4
1/32	5
1/64	6

$I(X = x)$ stands for "Information Content" aka "Surprisal" aka "Self-Information"

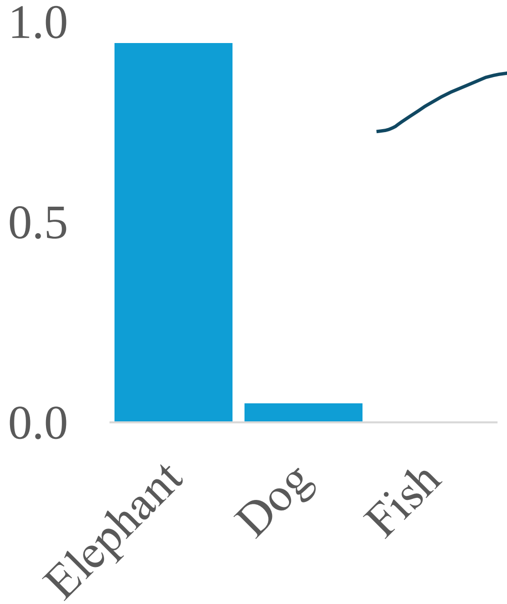
Back to the **measure** of **uncertainty** in the outcome of a random variable

Uncertainty of a Random Variable

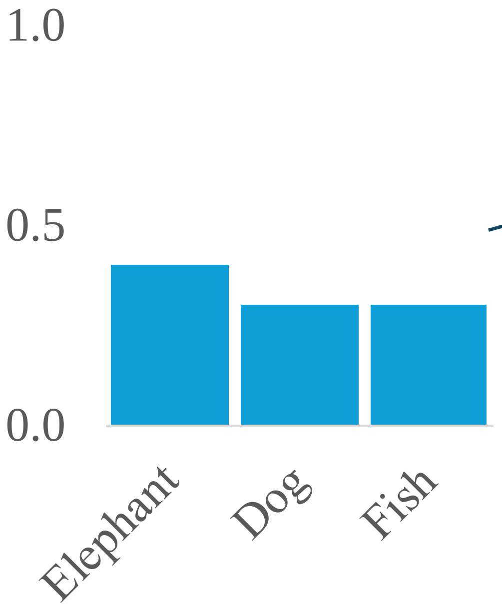
Let X be any random variable. We can calculate a statistic, "**Uncertainty**" to express how much we don't know about X

$$\text{Uncertainty}(X) = \sum_{x \in X} \text{Surprise}(X = x) \cdot P(X = x)$$

Uncertainty is expected Surprise



Low uncertainty



High uncertainty

Uncertainty of a Random Variable

Let X be any random variable. We can calculate a statistic, "**Uncertainty**" to express how much we don't know about X

$$\text{Uncertainty}(X) = \sum_{x \in X} \text{Surprise}(X = x) \cdot P(X = x)$$

Uncertainty is expected Surprise

Uncertainty of a Random Variable

Let X be any random variable. We can calculate a statistic, "**Uncertainty**" to express how much we don't know about X

$$\text{Uncertainty}(X) = \sum_{x \in X} \text{Surprise}(X = x) \cdot P(X = x)$$

Uncertainty is expected Surprise

$$= \sum_{x \in X} \log_2 \frac{1}{P(X = x)} \cdot P(X = x)$$

Our favorite measure of Surprise

$$= \sum_{x \in X} \log_2 P(X = x)^{-1} \cdot P(X = x)$$

1/x is the same as x^{-1}

$$= \sum_{x \in X} -\log_2 P(X = x) \cdot P(X = x)$$

Log of a power (here -1)

$$= -\sum_{x \in X} \log_2 P(X = x) \cdot P(X = x)$$

Pull the negative out

Uncertainty of a Random Variable (Entropy)

Let X be any random variable. We can calculate a statistic, "**Uncertainty**" to express how much we don't know about X

Calculates expected surprise

$$H(X) = -\sum_{x \in X} \log_2 P(X = x) \cdot P(X = x)$$

H is the symbol for the uncertainty statistic

$$\text{Surprise}(X = x) = \log_2 \frac{1}{P(X = x})$$

By the way…

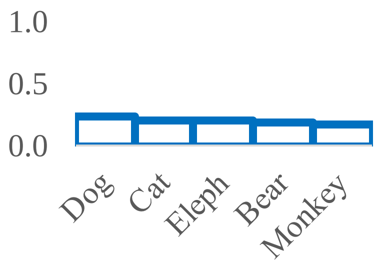
$$\mathrm{H}(X)$$

Is called **Shannon Entropy** to scare students and impress Physics people

Back to

I am thinking of an **animal**…

Which Question is Better?

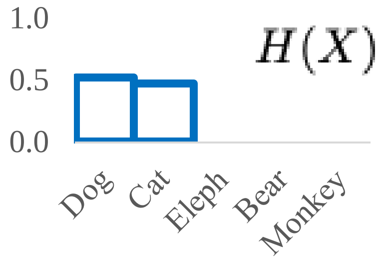


$H(X) = 2.3$

Is it a pet?

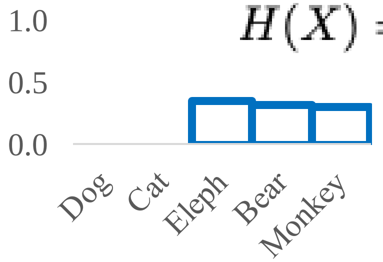
$E[H(X)] = 1.3$

yes
p = 0.44



$H(X) = 1$

no
p = 0.56

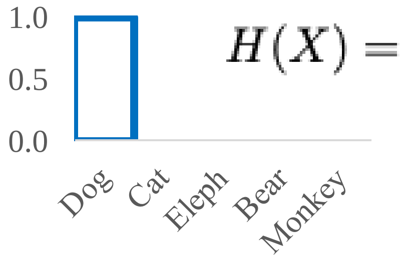


$H(X) = 1.6$

Is it a Dog?

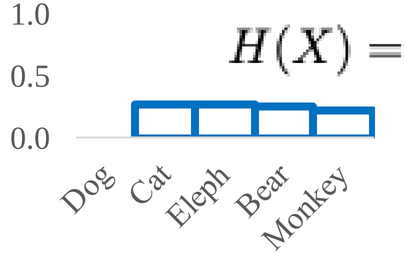
$E[H(X)] = 1.7$

yes
p = 0.23



$H(X) = 0$

no
p = 0.83



$H(X) = 2$

Uncertainty (aka Entropy) in code

```
def calc_uncertainty(pmf):
  # this calculates the entropy of the distribution
  # aka the uncertainty
  uncertainty = 0
  for x in pmf:
    p_x = pmf[x]
    # skip zero probabilities
    if p_x == 0: continue
    suprise_x = np.log2(1/p_x)
    uncertainty += suprise_x * p_x
  return uncertainty
```

$$\overbrace{\underbrace{H(X)}_{\text{Uncertainty of X}} = \sum_{x \in X} \underbrace{\log_2 \frac{1}{P(X=x)}}_{\text{Surprise(x)}} \cdot P(X=x)}^{\text{Expected Surprise}}$$

Entropy in the sum of two dice.

What is more informative:

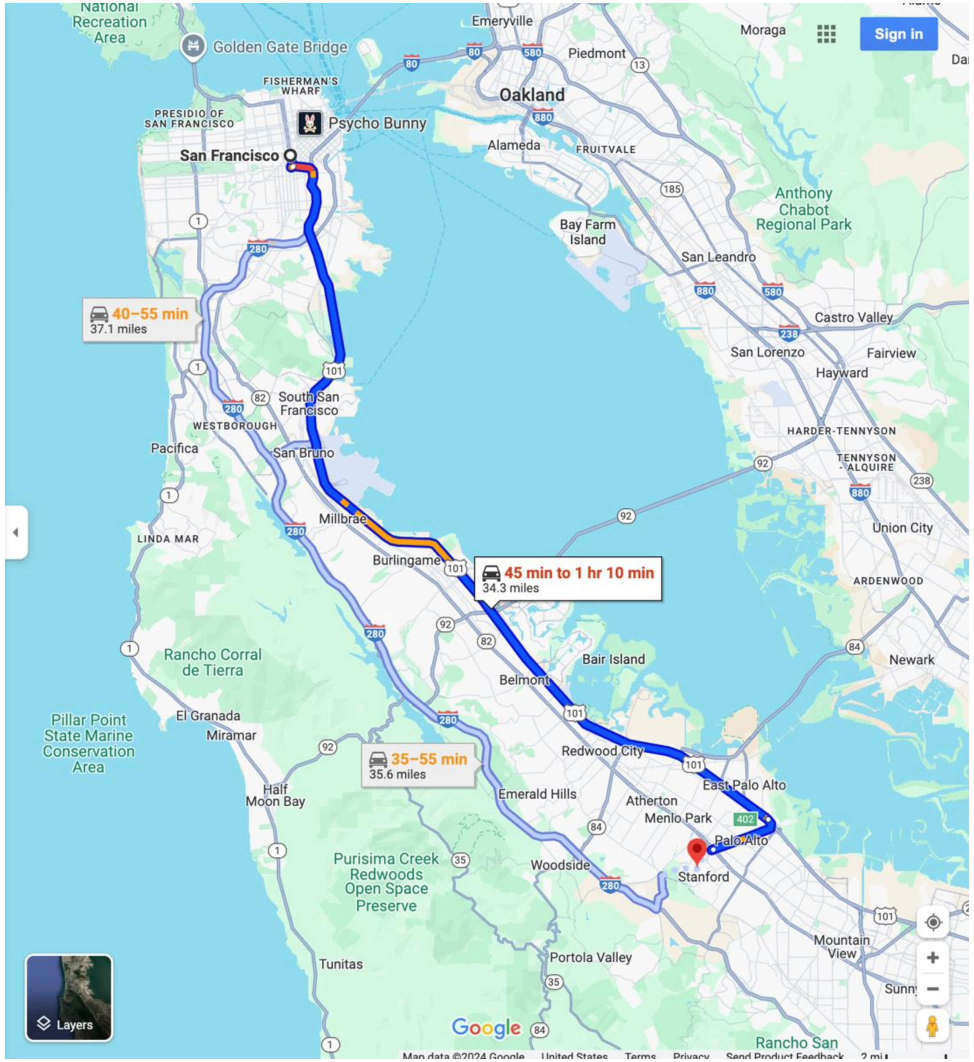
I tell you that the **sum is odd**

I tell you the value of the **first dice**

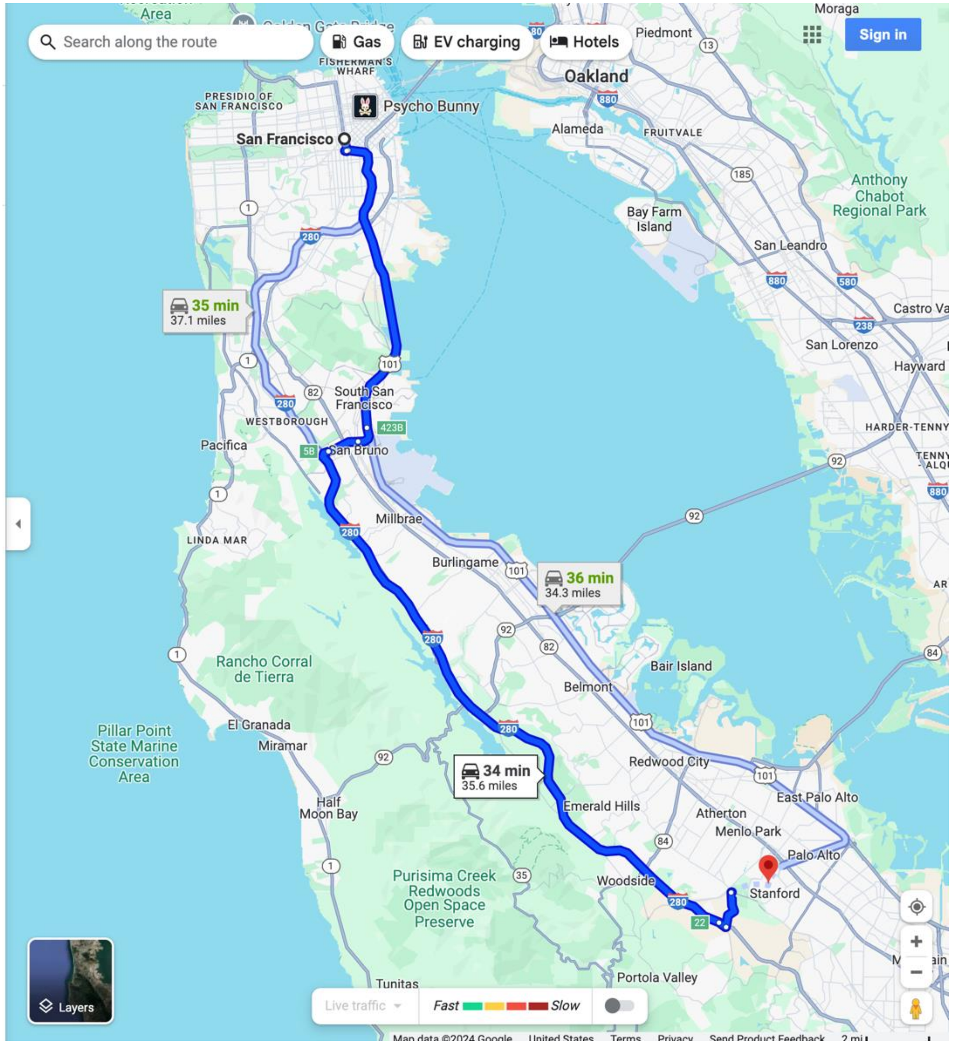
Traffic Predictions Over Time

Should You Wait to Plan Your Route?

Prediction 1 day in advance

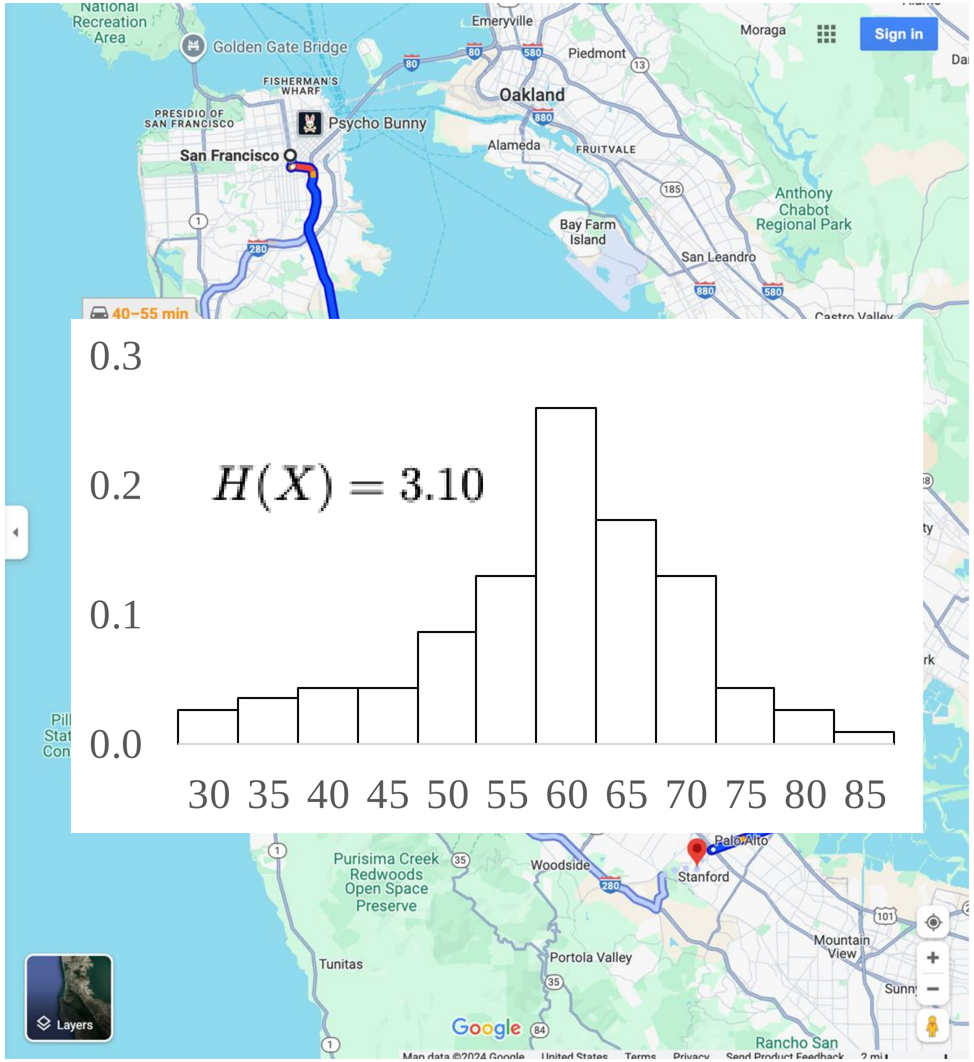


Prediction 30 mins in advance

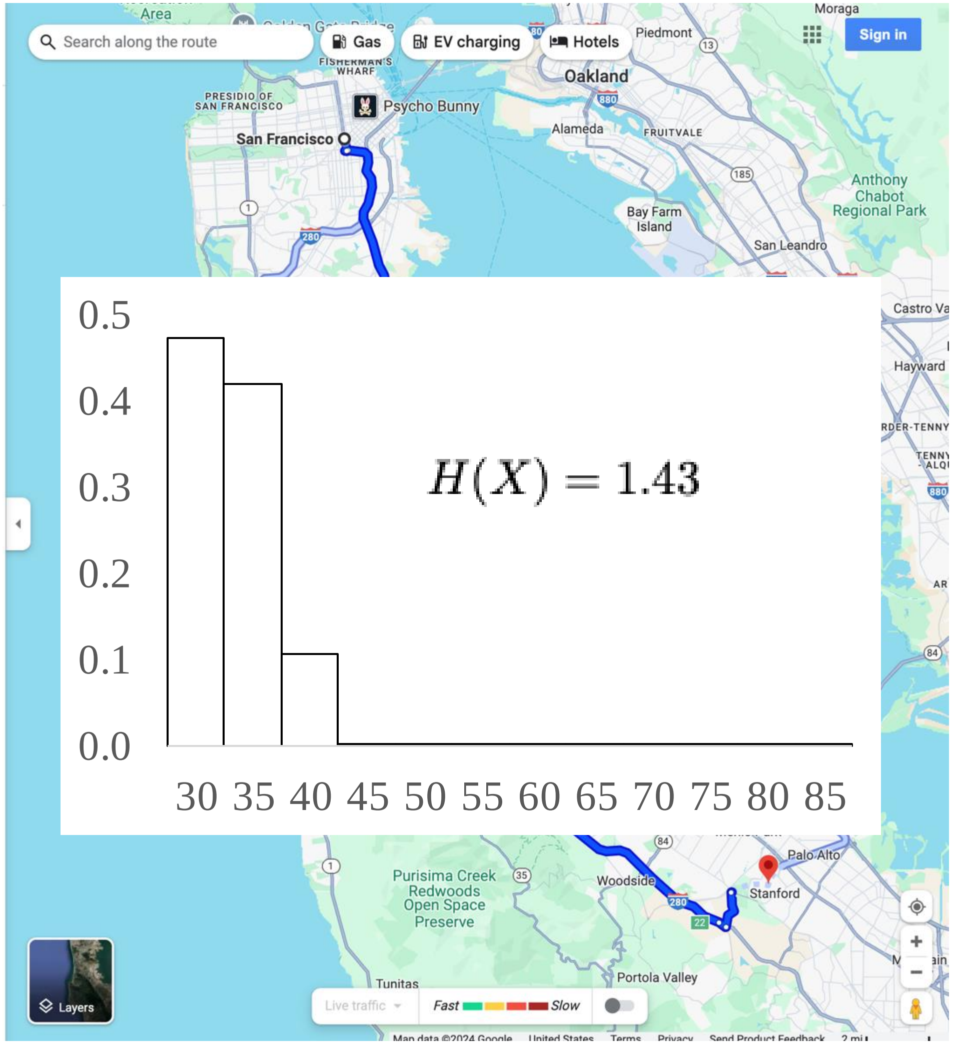


Should You Wait to Plan Your Route?

Prediction 1 day in advance

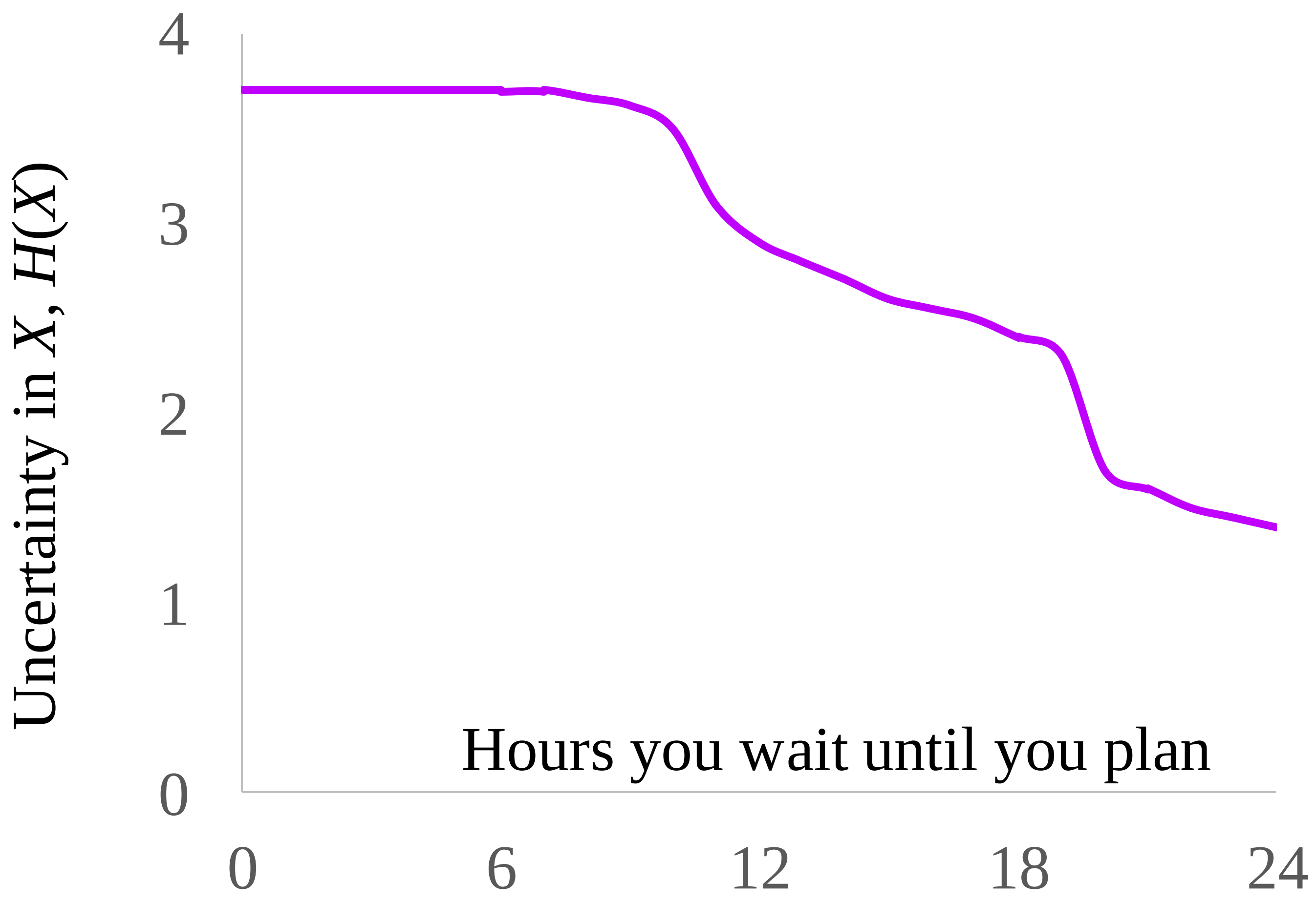


Prediction 30 mins in advance



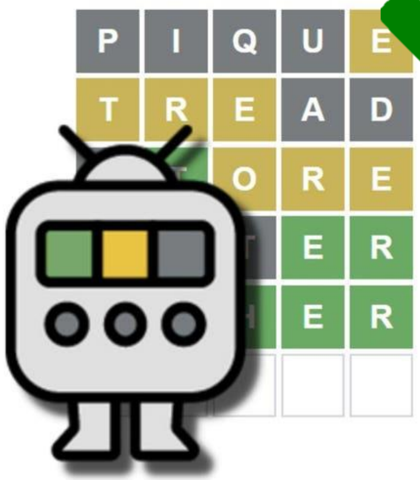
Should You Wait to Plan Your Route?

Let X be the amount of time to drive from SF to Stanford

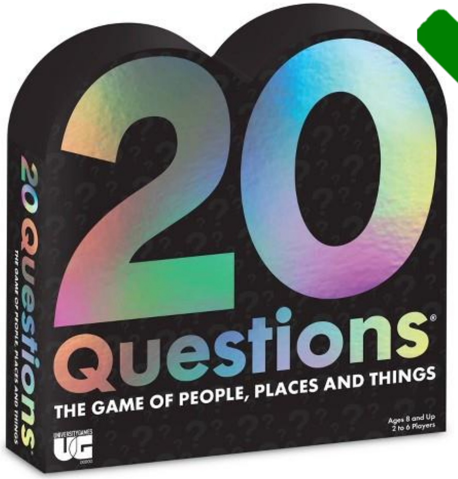


Limitations of Entropy for Decision Making?

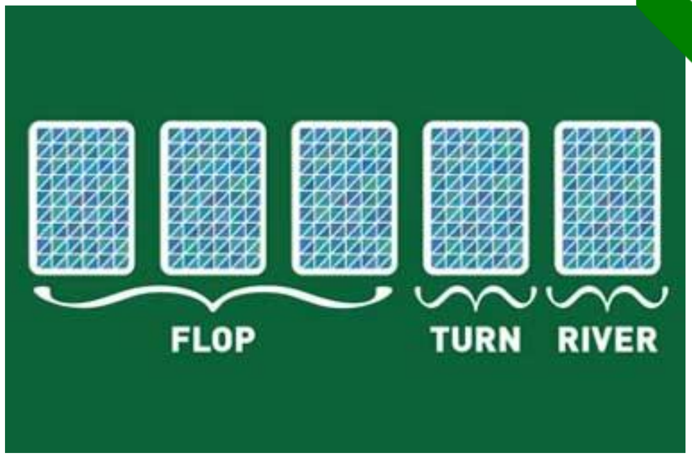
WorldeBot



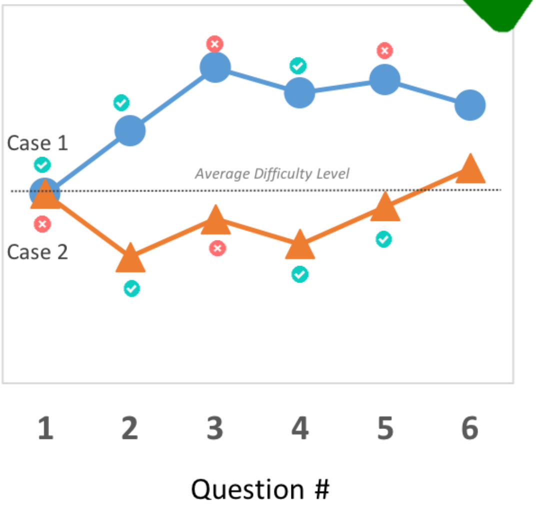
Decision Trees



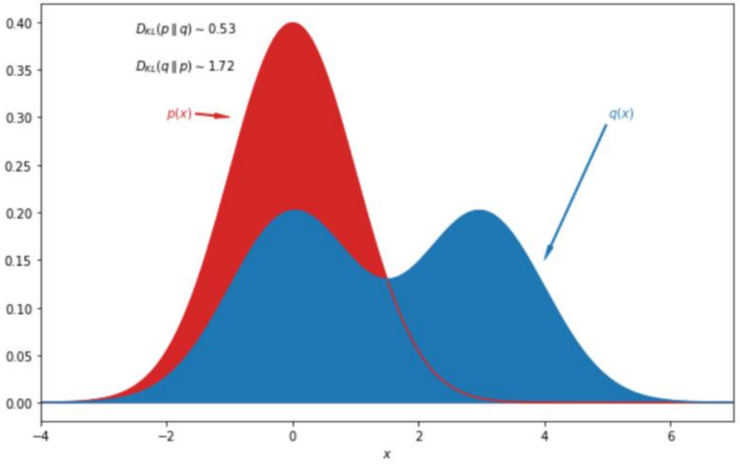
Value of Info in Poker



Adaptive Tests



Comparing Distributions

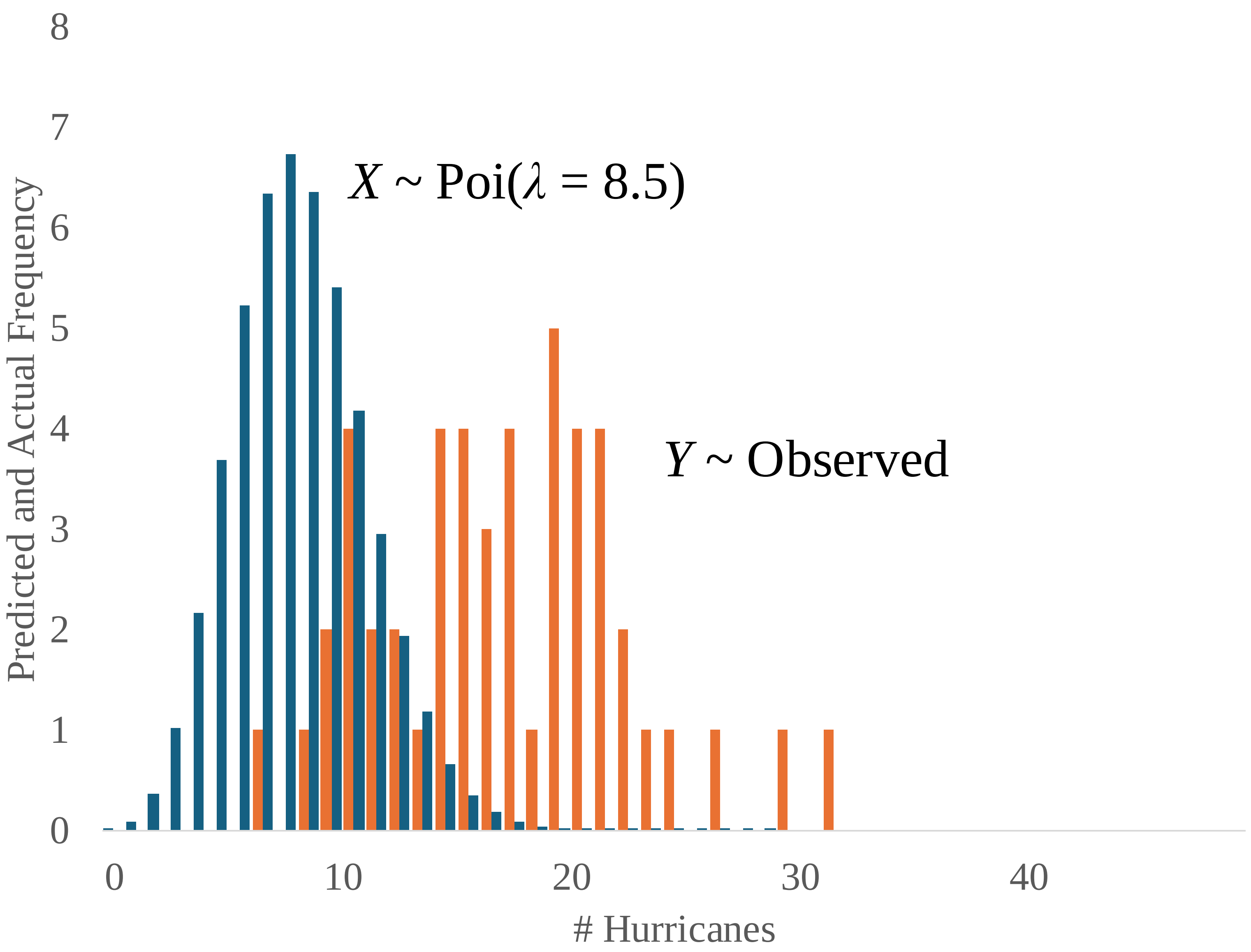


Compression of Data

DALL·E 3

Distance Between Two Distributions

Recall this

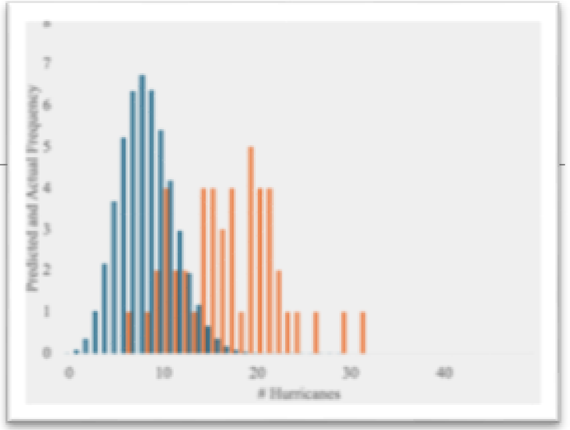


$X \sim \text{Poi}(\lambda = 8.5)$
$Y \sim$ Observed
Predicted and Actual Frequency
Hurricanes
0
1
2
3
4
5
6
7
8
0
10
20
30
40

Distance Between Two Distributions

Let poisson prediction be X

Let real data be Y



Three reasonable ideas

Total Variation (TV)

Loop over all possible values and calculate the ***absolute difference*** *in probability*

$$TV(X,Y) = \sum_i |P(X=i) - P(Y=i)|$$

Earth Movers (EMD)

Imagine one distribution is a ***lump of dirt****. How much work would it take to make it look just like the other?*

Solved using an LP Solver

$O(n^3 \log n)$

Kullback Leibler (KL)

Expected ***excess surprise*** *from using Y as a model instead of X when the actual distribution is X.*

$$KL(X,Y) = \sum_x \log \frac{P(X=x)}{P(Y=x)} \cdot P(X=x)$$

KL Divergence Without Tears

$$\mathrm{KL}(X,Y) = \sum_{x \in X} \mathrm{ExcessSurprise}(x) \cdot P(X=x)$$

How much more surprising is x under Y than X?

$$= \sum_{x \in X} \left[\mathrm{Surprise}_Y(x) - \mathrm{Surprise}_X(x)\right] \cdot P(X=x)$$

Surprise according to?

$$= \sum_{x \in X} \left[\log_2 \frac{1}{P(Y=x)} - \log_2 \frac{1}{P(X=x)}\right] \cdot P(X=x)$$

Surprise!

$$= \sum_{x \in X} -\log_2 P(Y=x) + \log_2 P(X=x) \cdot P(X=x)$$

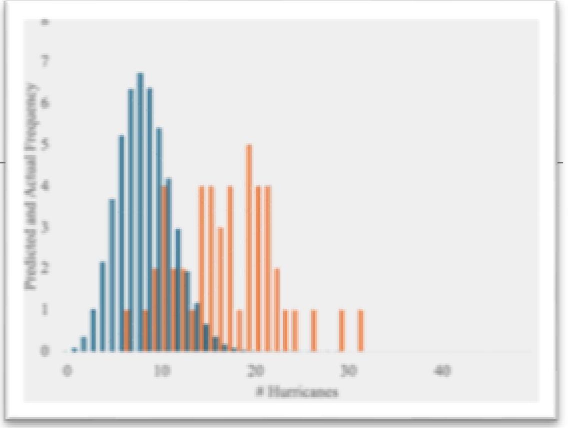
$1/x = x^{-1}$

$$= \sum_{x \in X} \log_2 \frac{P(X=x)}{P(Y=x)} \cdot P(X=x)$$

Log rules

People often use natural log

KL Divergence in Code



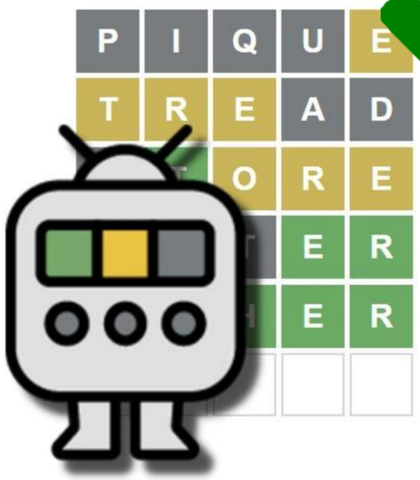
Let poisson prediction be X

Let real data be Y

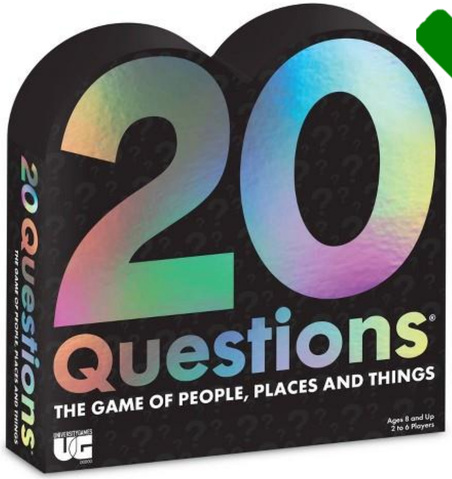
```
from scipy import stats
import math

def kl_divergence(predicted_lambda, observed_pmf):
  """
  We predicted that the number of hurricanes would be
  X ~ Poisson(predicted_lambda) and observed a real world
  number of hurricanes Y ~ observed_pmf
  """
  X = stats.poisson(predicted_lambda)
  divergence = 0
  # loop over all the values of hurricanes
  for i in range(0, 40):
    pr_X_i = X.pmf(i)
    pr_Y_i = observed_pmf[i]
    excess_surprise_i = math.log(pr_X_i / pr_Y_i)
    divergence += excess_surprise_i * pr_X_i
  return divergence
```

$$KL(X, Y) \approx 0.376$$

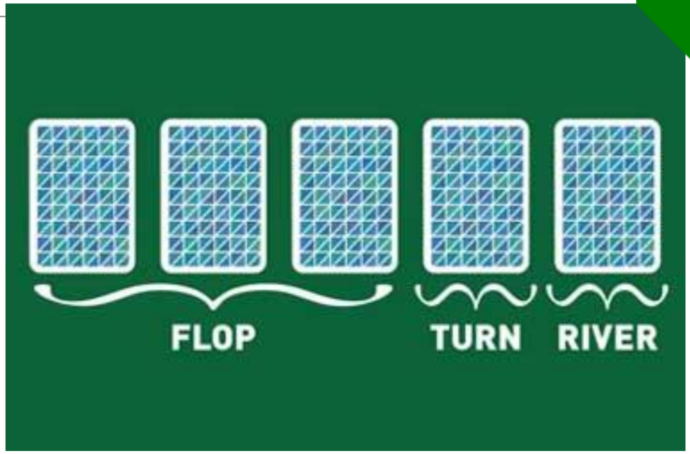
WorldeBot



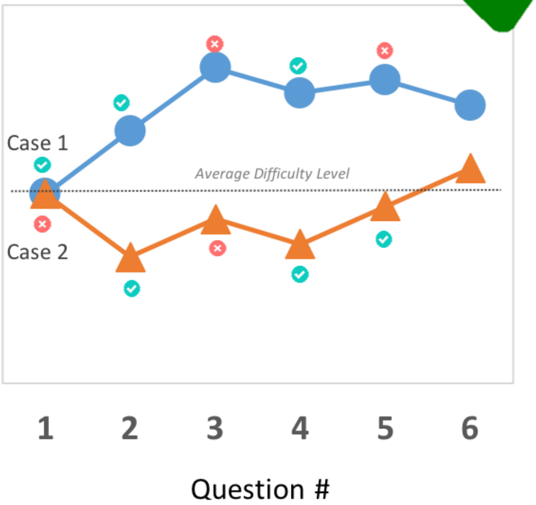
Decision Trees



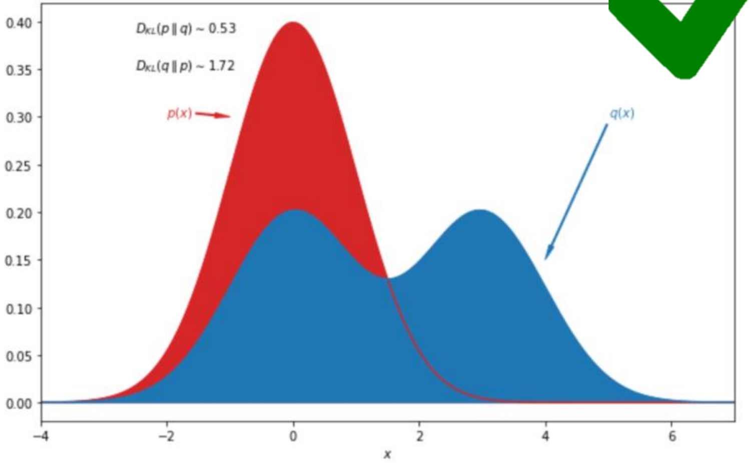
Value of Info in Poker



Adaptive Tests



Comparing Distributions



Compression of Data

Where are we in CS109?

On Monday…

Counting Theory	Core Probability	Random Variables	Probabilistic Models	Uncertainty Theory	Machine Learning