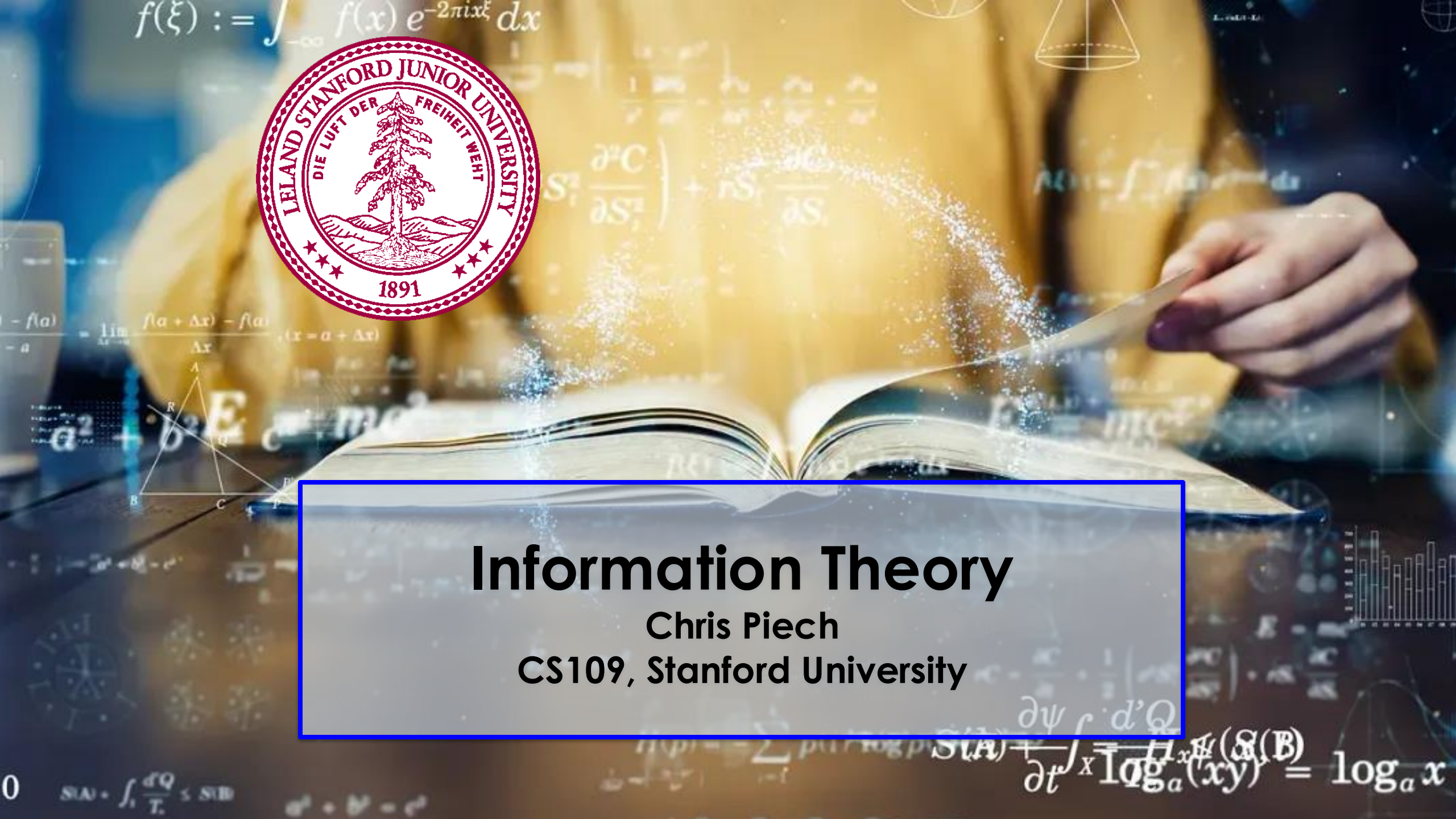
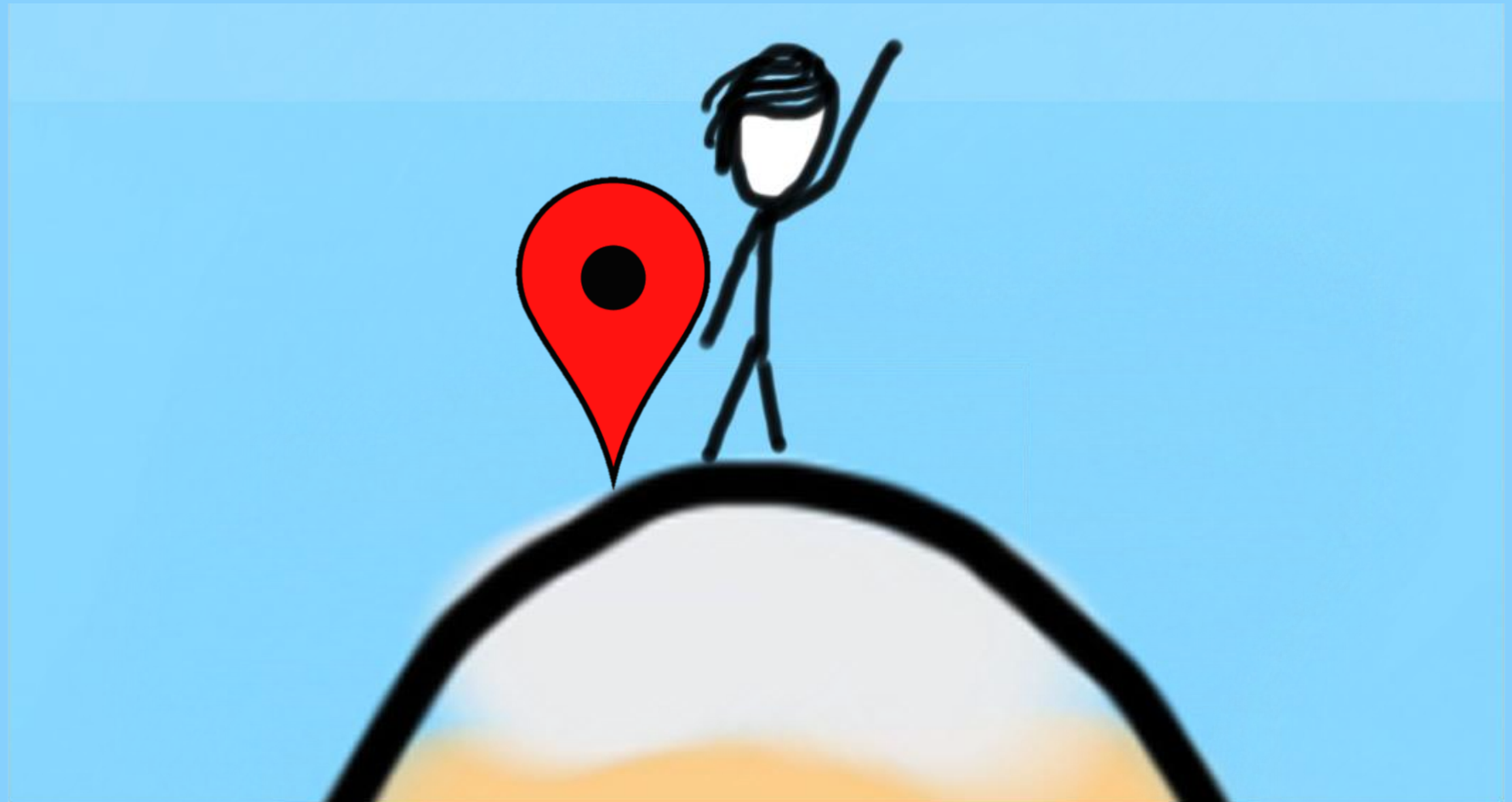


Information Theory

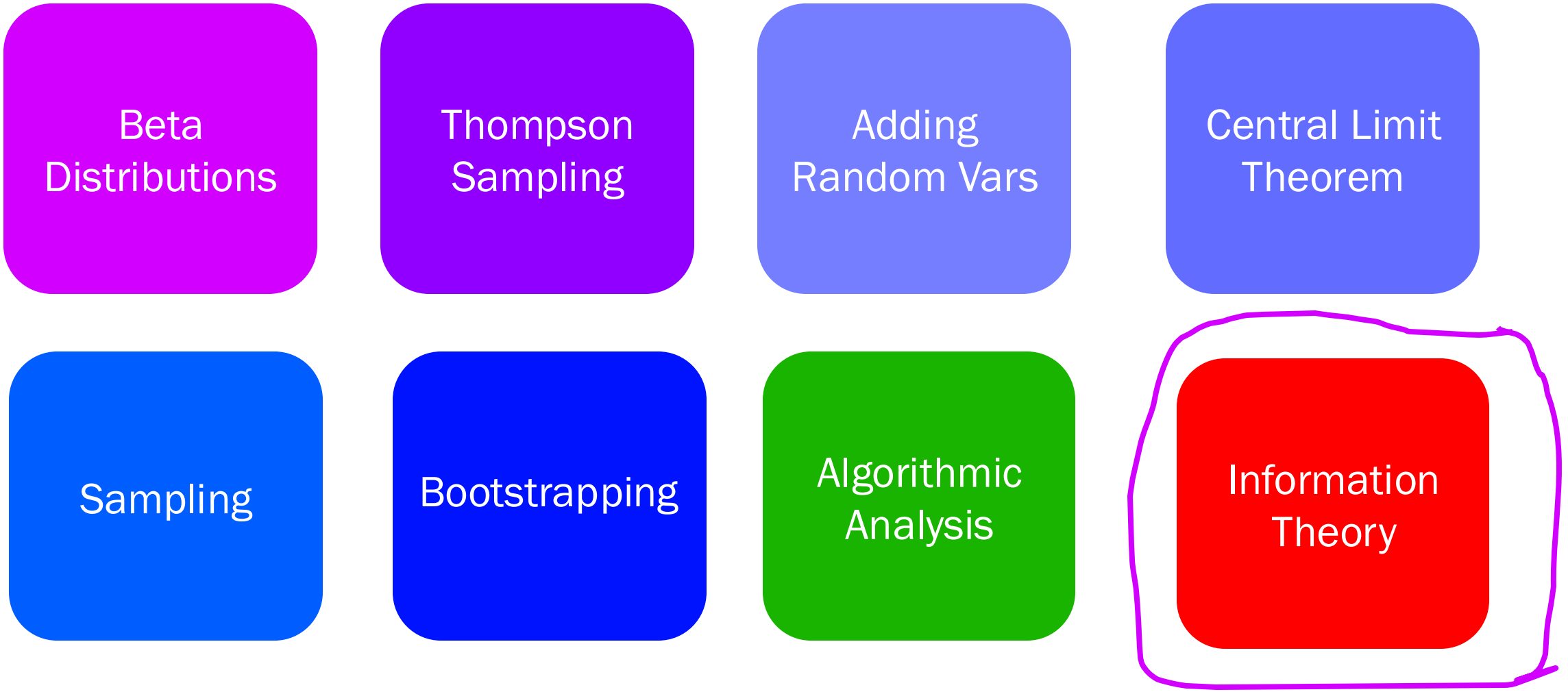
Chris Piech
CS109, Stanford University

Learning Goals

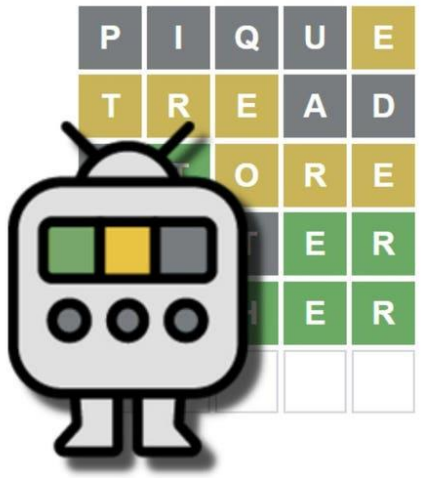
1. Calculate information gain
2. Make choices that maximize information gain
3. Numerically score how similar two distributions are



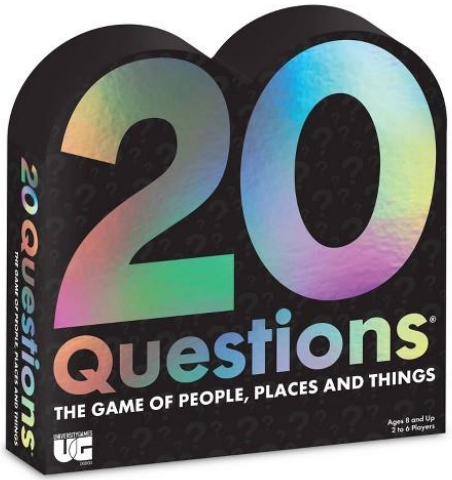
Uncertainty Theory



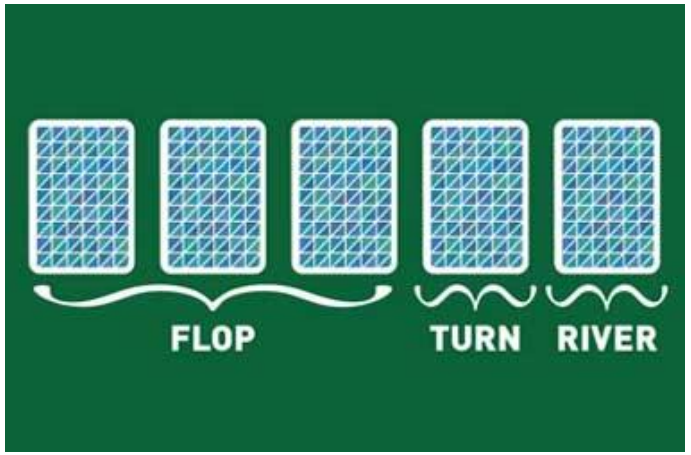
WorldeBot



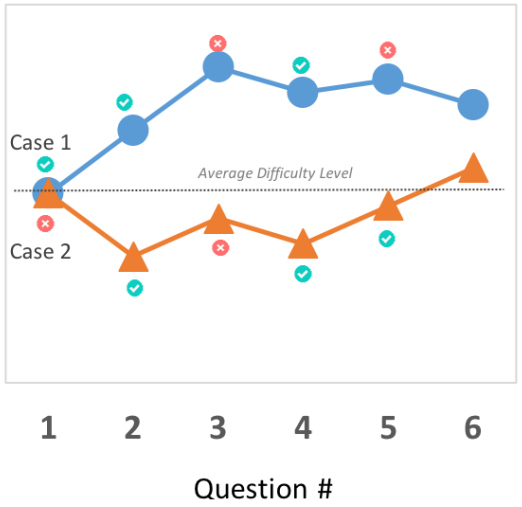
Decision Trees



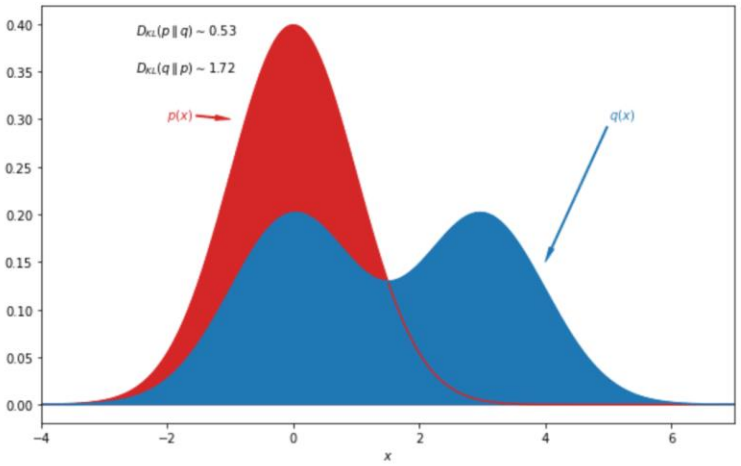
Value of Info in Poker



Adaptive Tests



Comparing Distributions

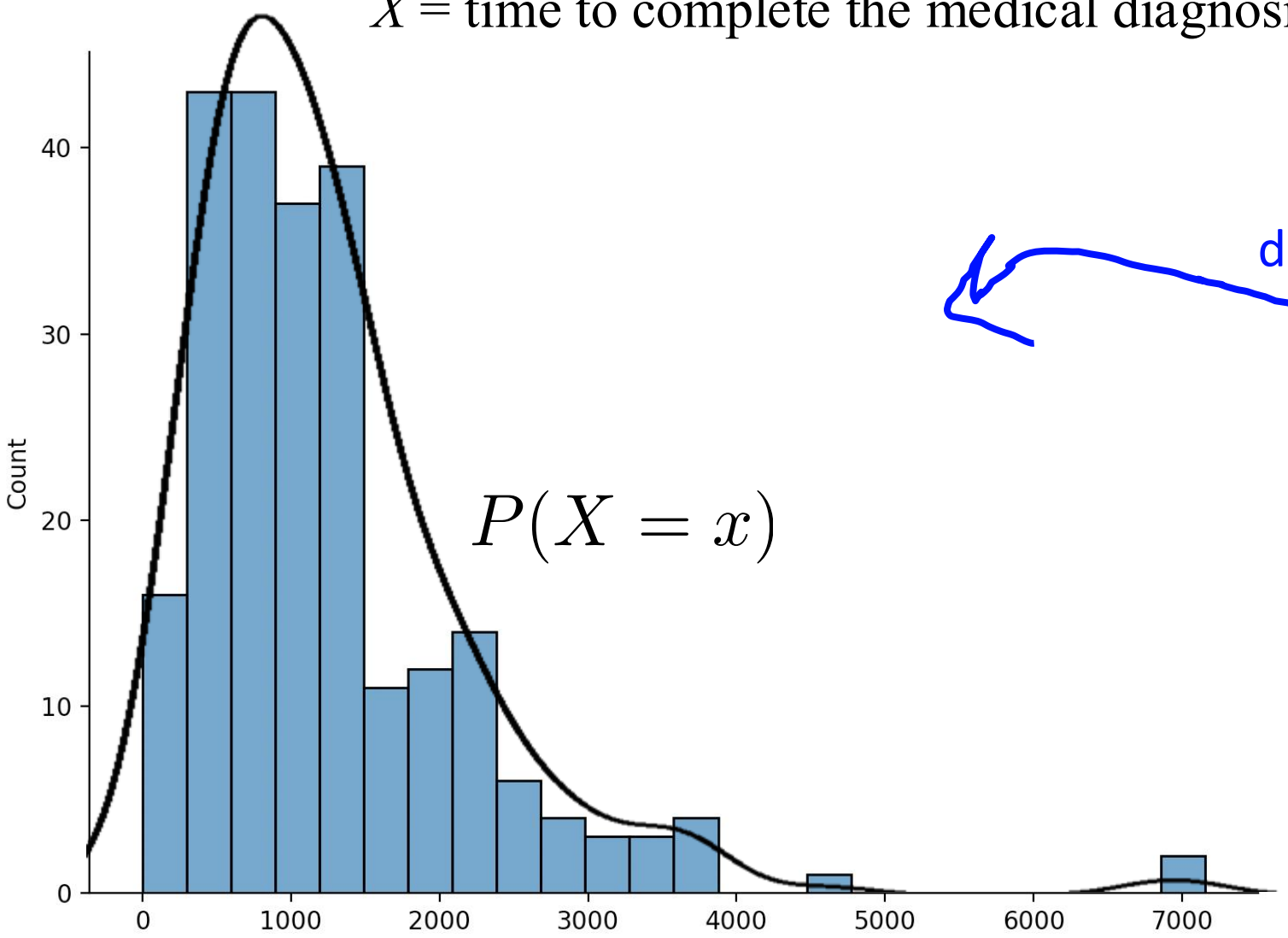


Compression of Data

Review

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

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

$$E[X] = \sum_i E[Y_i] \quad \text{if} \quad X = \sum_i Y_i$$

$$E[X] = \sum_y E[X|Y = y]P(Y = y) \quad \text{For any } Y$$

Expectation of a Function

Law of unconscious statistician

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

So for example…

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

End Review

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 expected "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!

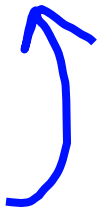
I am thinking of an **animal**…

I am thinking of an animal

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

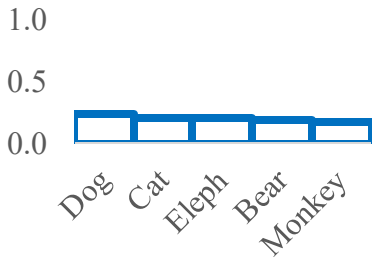
Question Choosing Algorithms

Algorithm	Average Questions	Standard Error of the Mean
Random Questions	61.34	1.41
Binary Search	7.79	0.02
Information Theory Search	5.33	0.04



Hey what's that?

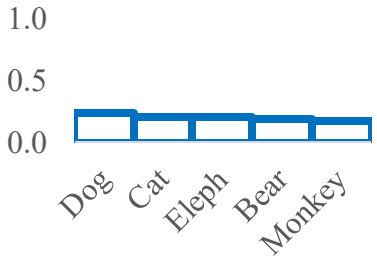
Which Question is Better?



Is it a pet?

Is it a Dog?

Which Question is Better?



Is it a pet?

yes
$p = 0.44$

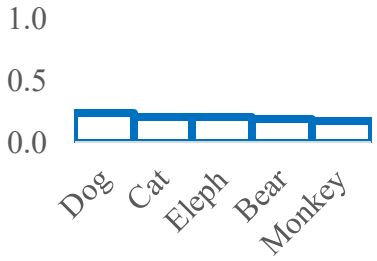
no
$p = 0.56$

Is it a Dog?

yes
$p = 0.23$

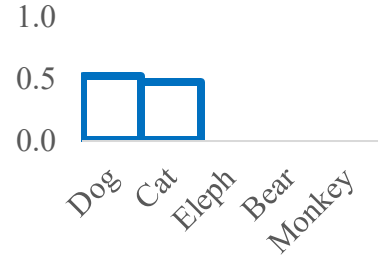
no
$p = 0.83$

Which Question is Better?

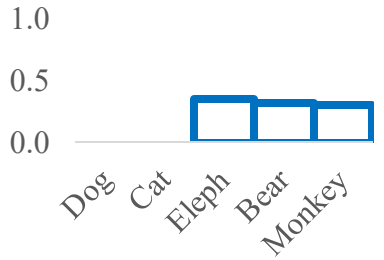


Is it a pet?

yes
$p = 0.44$

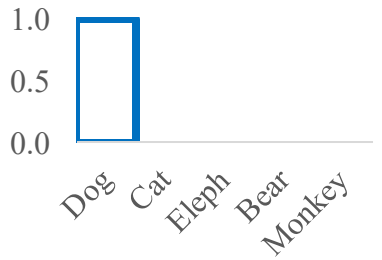


no
$p = 0.56$

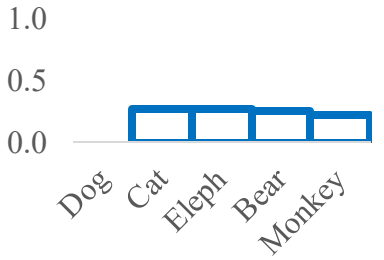


Is it a Dog?

yes
$p = 0.23$

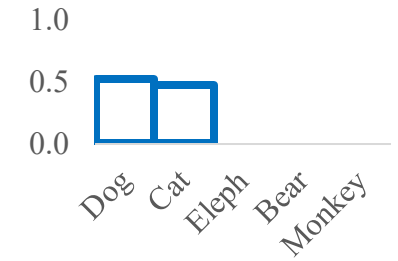


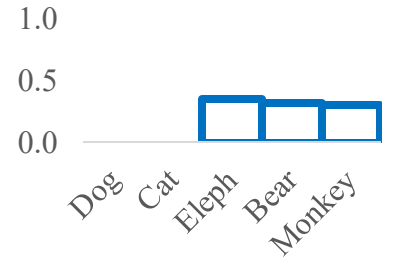
no
$p = 0.83$
77

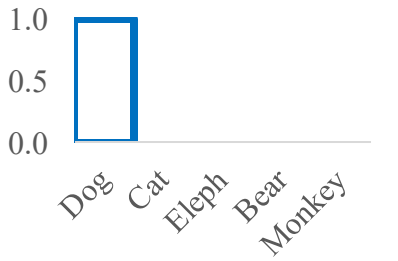


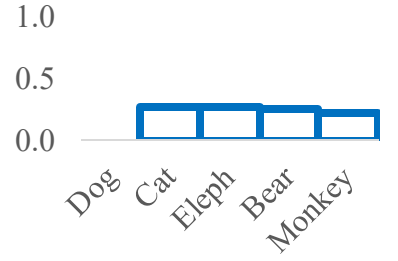
Which Question is Better

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





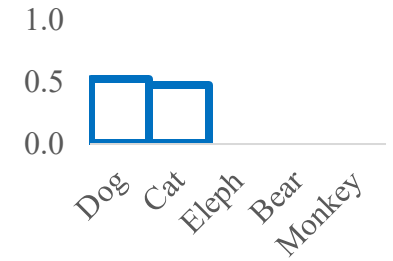


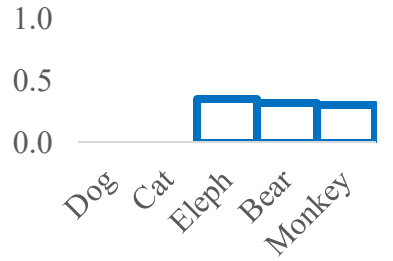


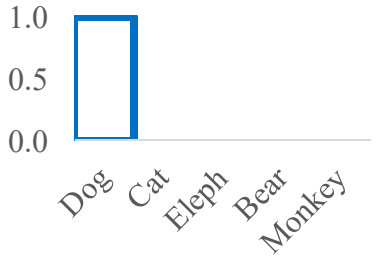
Which Question is Better

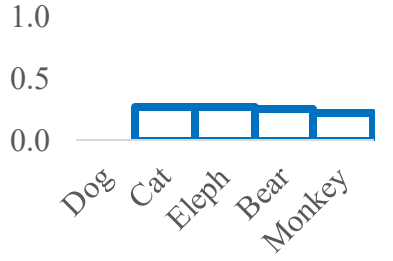
uncertain

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





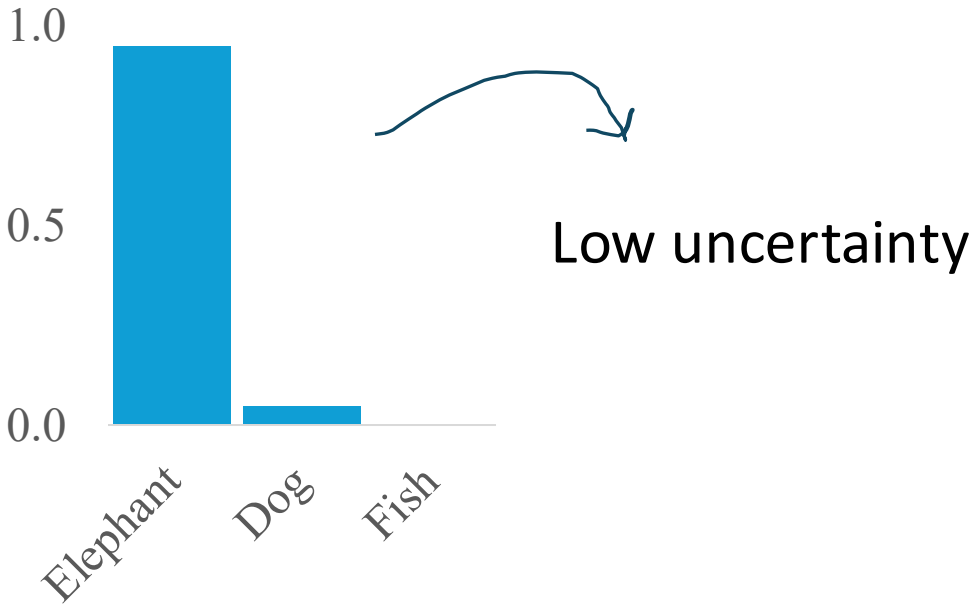




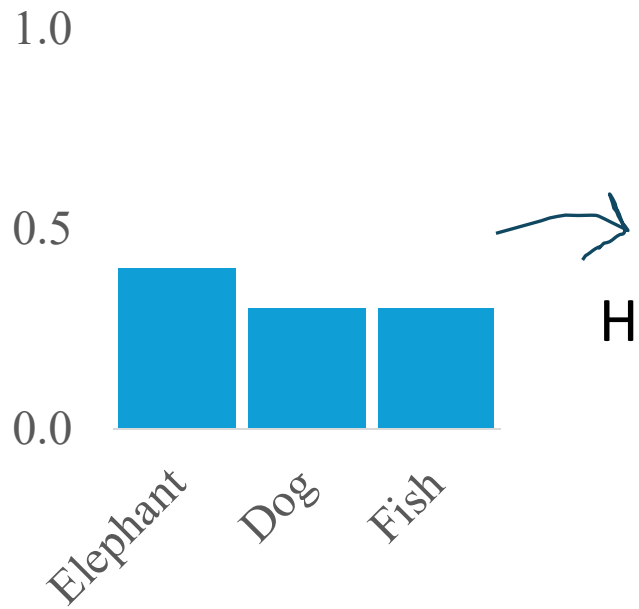
What we really need is a **measure** of our **uncertainty** in a random variable

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

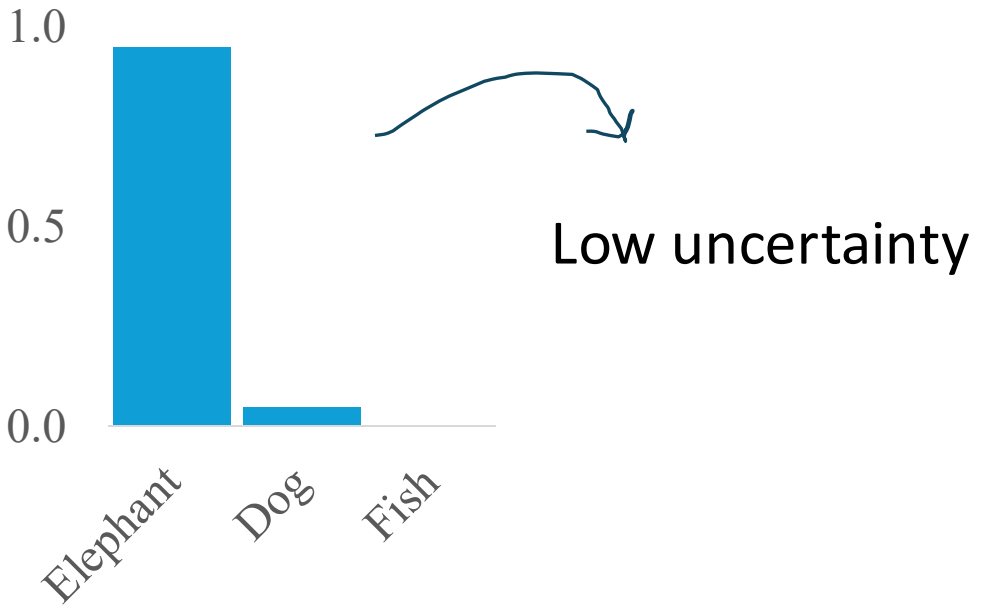


High uncertainty

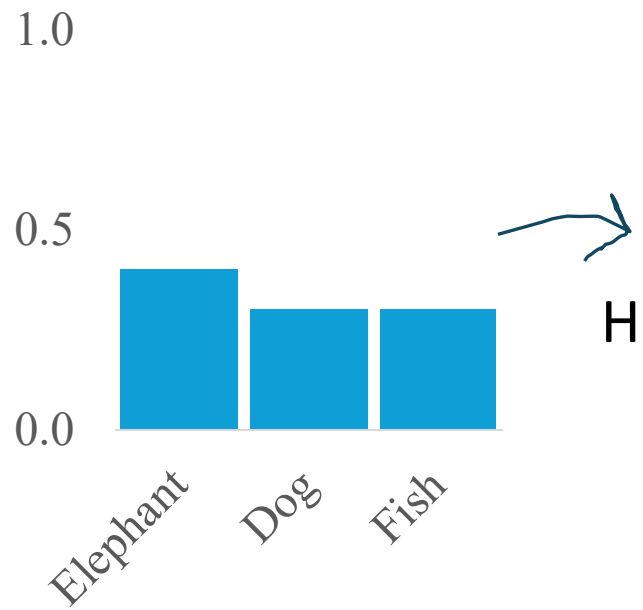
Uncertainty of a Random Variable

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

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



Low uncertainty



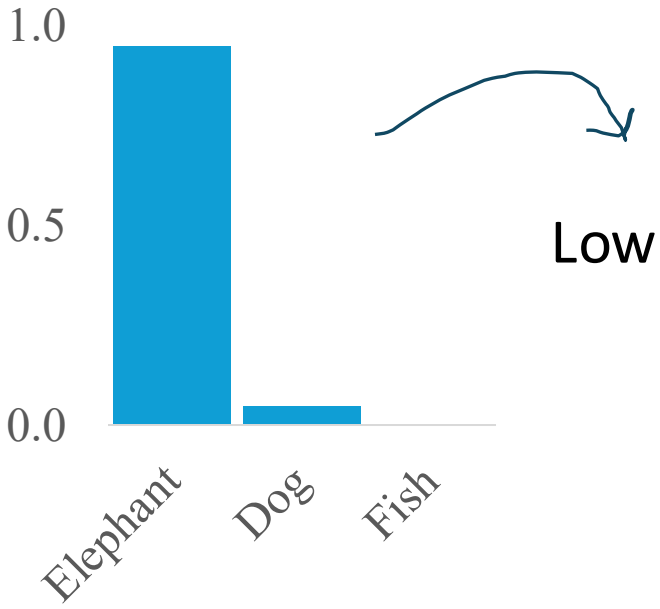
High uncertainty

Uncertainty of a Random Variable

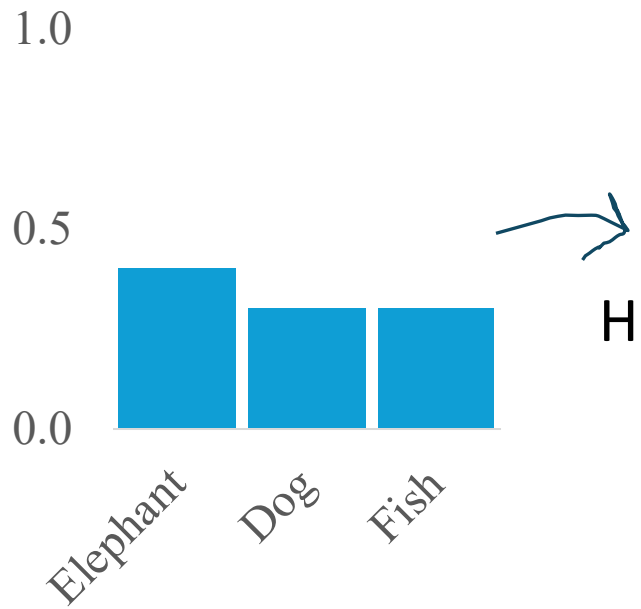
Let X be any random variable. We can calculate a statistic, "**Uncertainty**" to express how much we don't know about X

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

Uncertainty is expected Surprise



Low uncertainty



High uncertainty

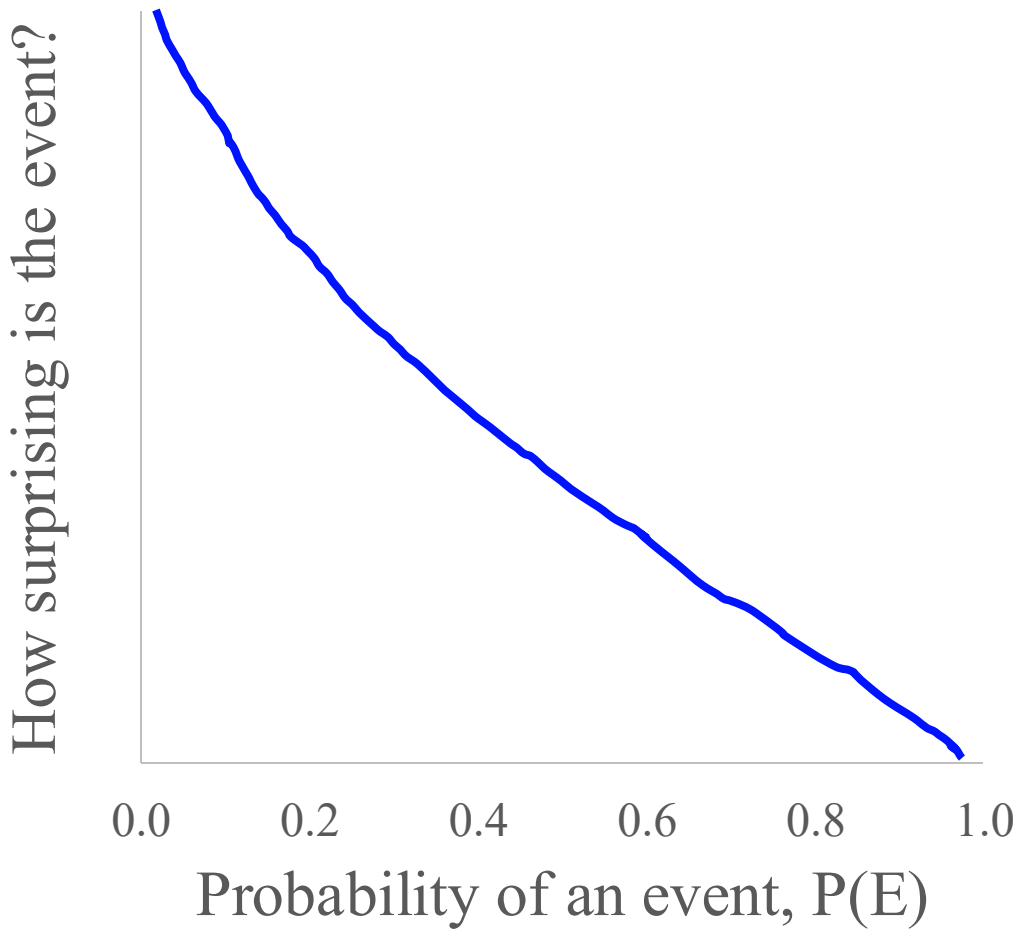
Ok, but then what is our measure of **Surprise**?

Surprise of an Event

High probability events are **not surprising**

Low probability events are **surprising**

Relationship should be monotonic



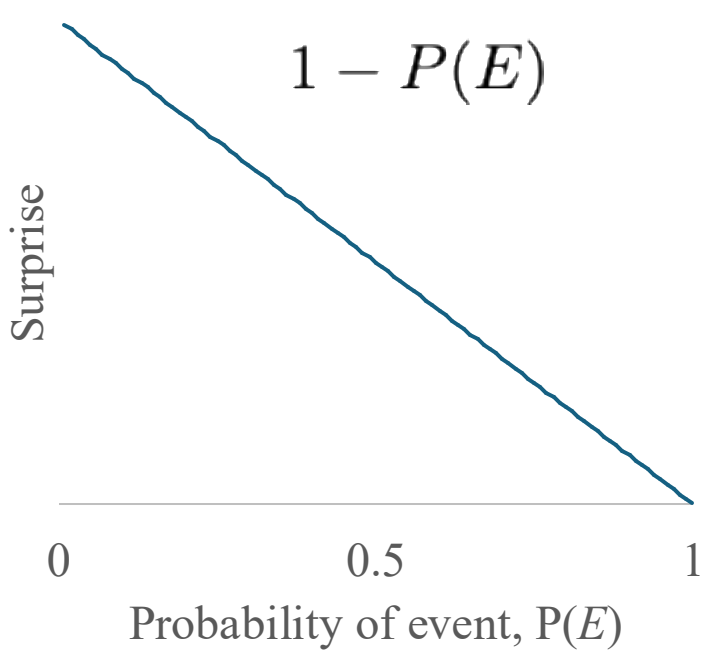
Surprise of an Event

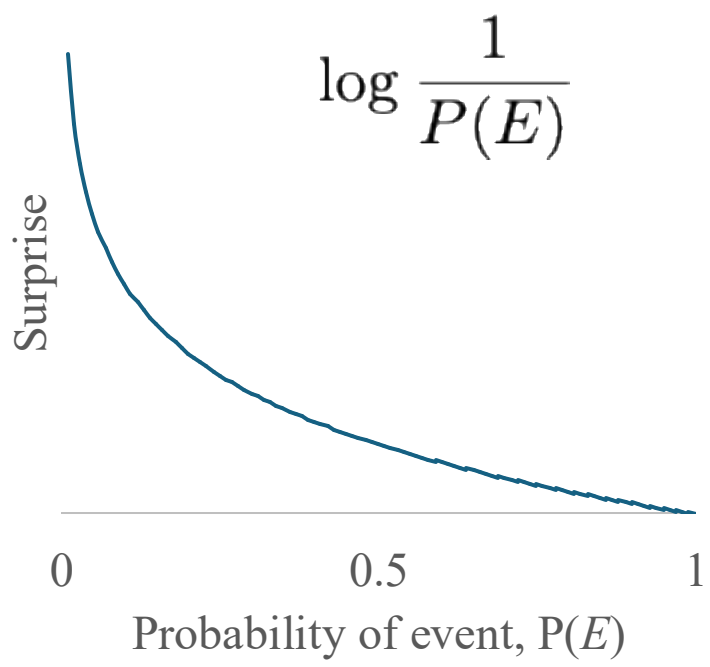
High probability events are **not surprising**

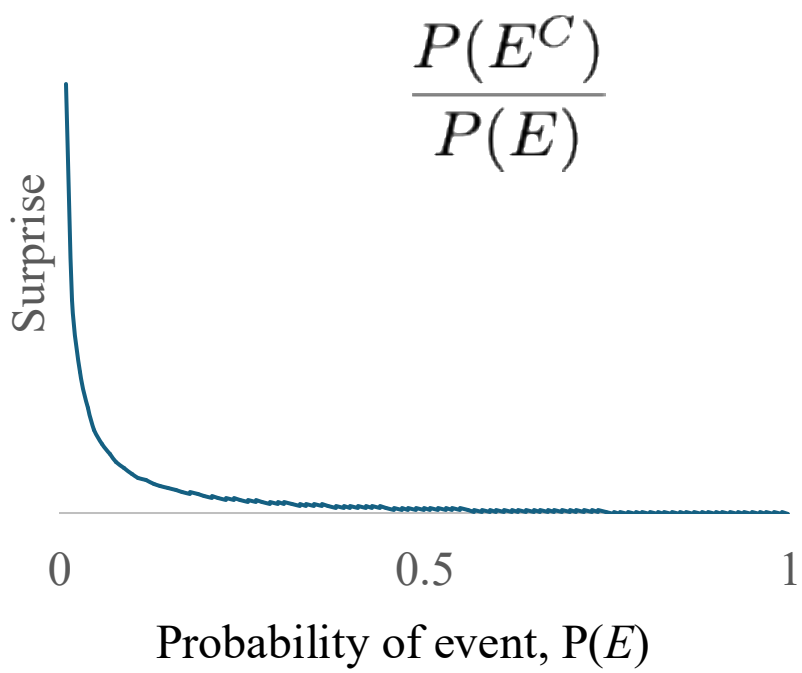
Low probability events are **surprising**

Relationship should be monotonic

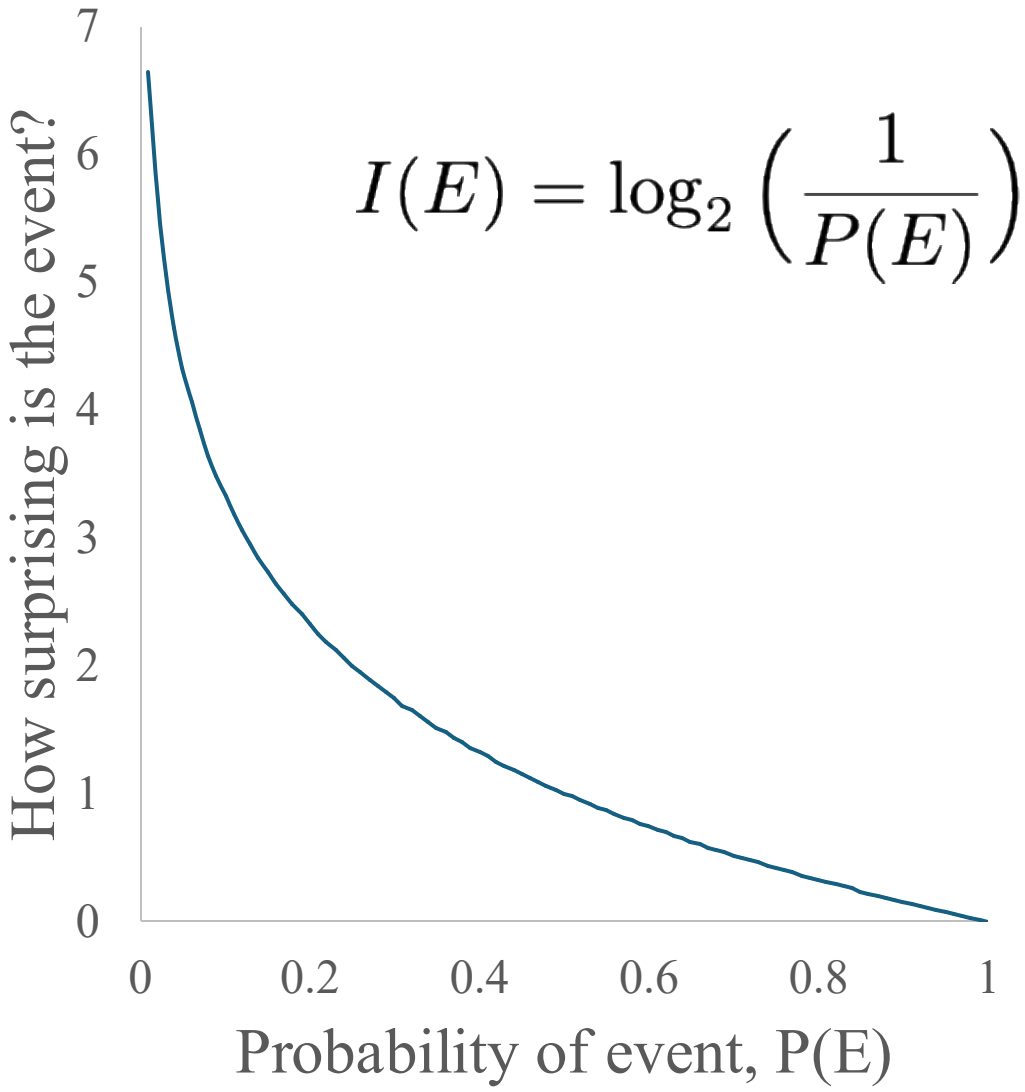
Here are three reasonable options







Surprise of an Event, I(E)

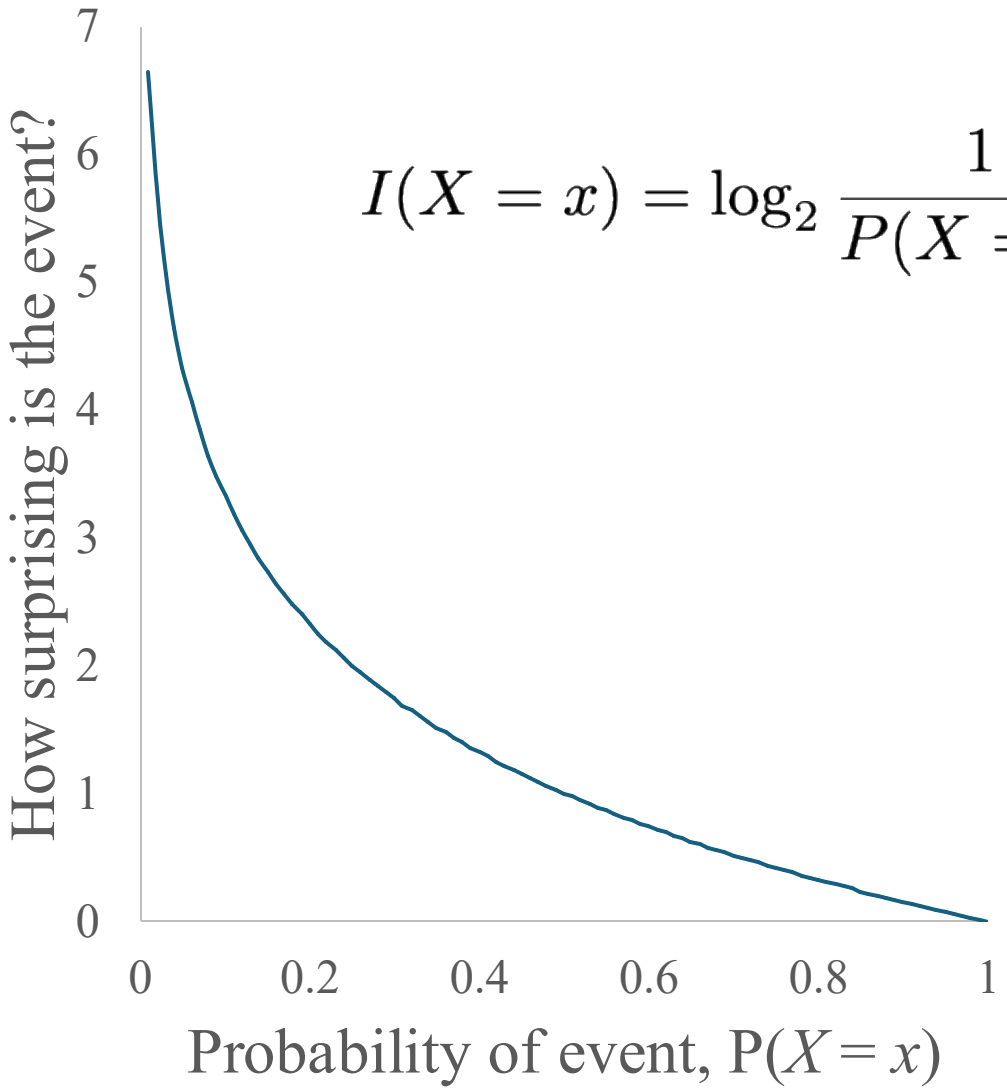


$$I(E) = \log_2\left(\frac{1}{P(E)}\right)$$

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, $\mathrm{I}(X = x)$



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

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"

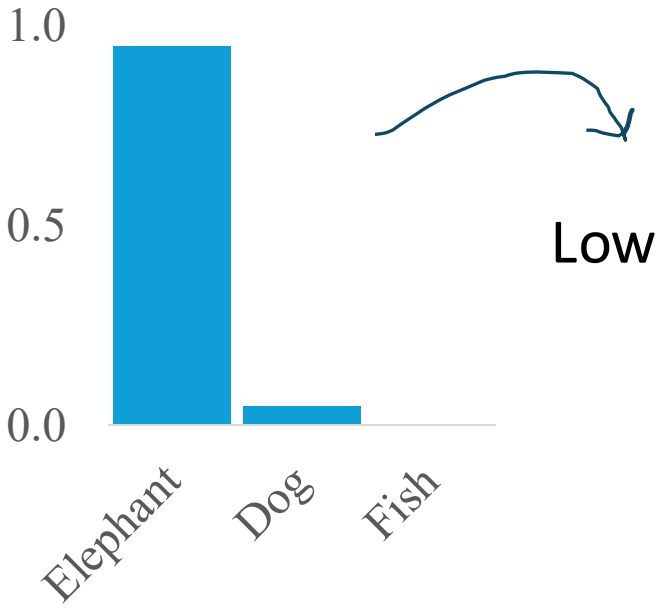
Back to the **measure** of **uncertainty** in the outcome of a random variable

Uncertainty of a Random Variable

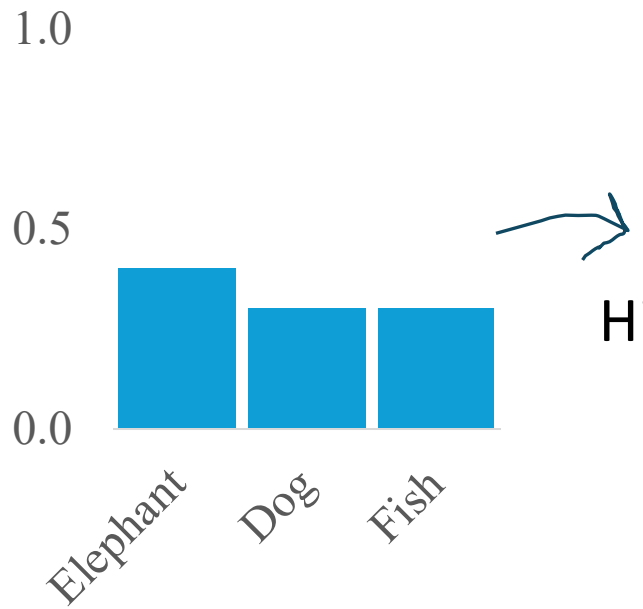
Let X be any random variable. We can calculate a statistic, "**Uncertainty**" to express how much we don't know about X

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

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

$$\begin{aligned}
\text{Uncertainty}(X) &= \sum_{x \in X} \text{Surprise}(X = x) \cdot P(X = x) && \text{Uncertainty is expected Surprise} \\
&= \sum_{x \in X} \log_2 \frac{1}{P(X = x)} \cdot P(X = x) && \text{Our favorite measure of Surprise} \\
&= \sum_{x \in X} \log_2 P(X = x)^{-1} \cdot P(X = x) && \text{1/x is the same as } x^{-1} \\
&= \sum_{x \in X} -\log_2 P(X = x) \cdot P(X = x) && \text{Log of a power (here -1)} \\
&= -\sum_{x \in X} \log_2 P(X = x) \cdot P(X = x) && \text{Pull the negative out}
\end{aligned}$$

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

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

My preferred name for "entropy" aka H(X)

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

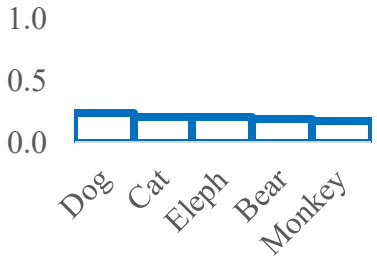
By the way…

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

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

Back to
I am thinking of an **animal**…

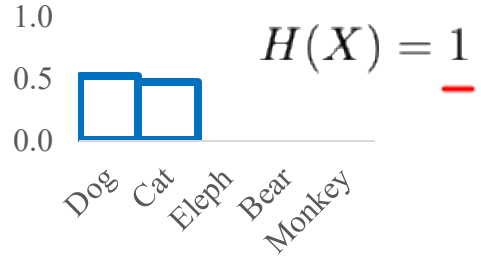
Which Question is Better?



$H(X) = 2.3$

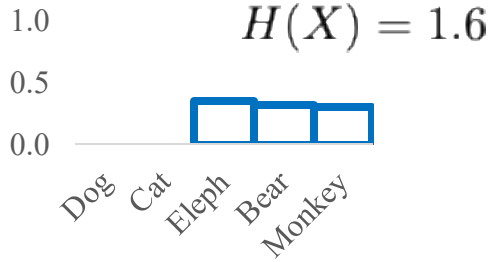
Is it a pet?

yes
p = 0.44



$H(X) = 1$

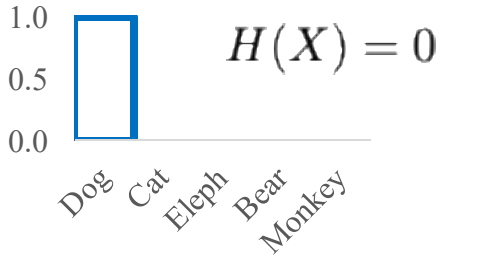
no
p = 0.56



$H(X) = 1.6$

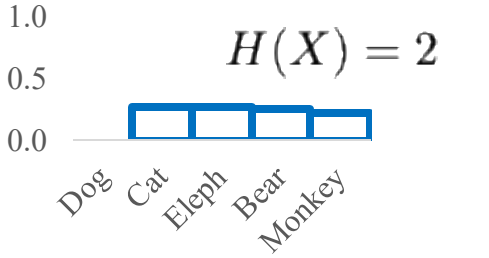
Is it a Dog?

yes
p = 0.23



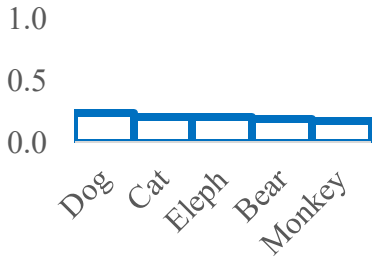
$H(X) = 0$

no
p = 0.83



$H(X) = 2$

Which Question is Better?

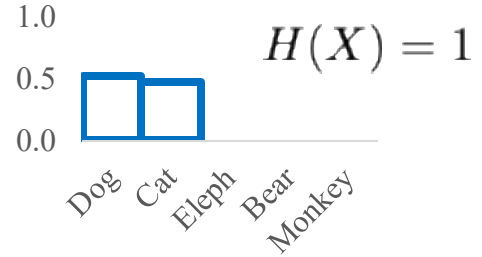


$H(X) = 2.3$

Is it a pet?

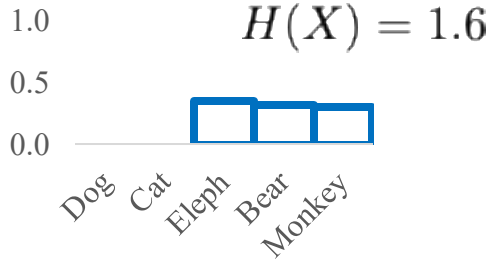
$E[H(X)] = 1.3$

yes
p = 0.44



$H(X) = 1$

no
p = 0.56

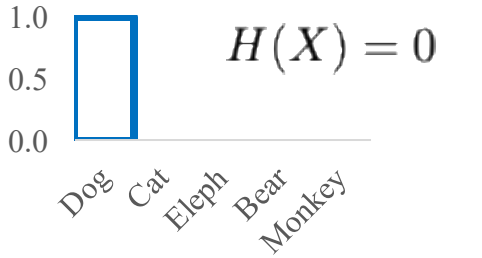


$H(X) = 1.6$

Is it a Dog?

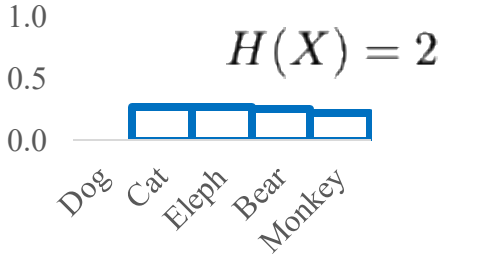
$E[H(X)] = 1.7$

yes
p = 0.23



$H(X) = 0$

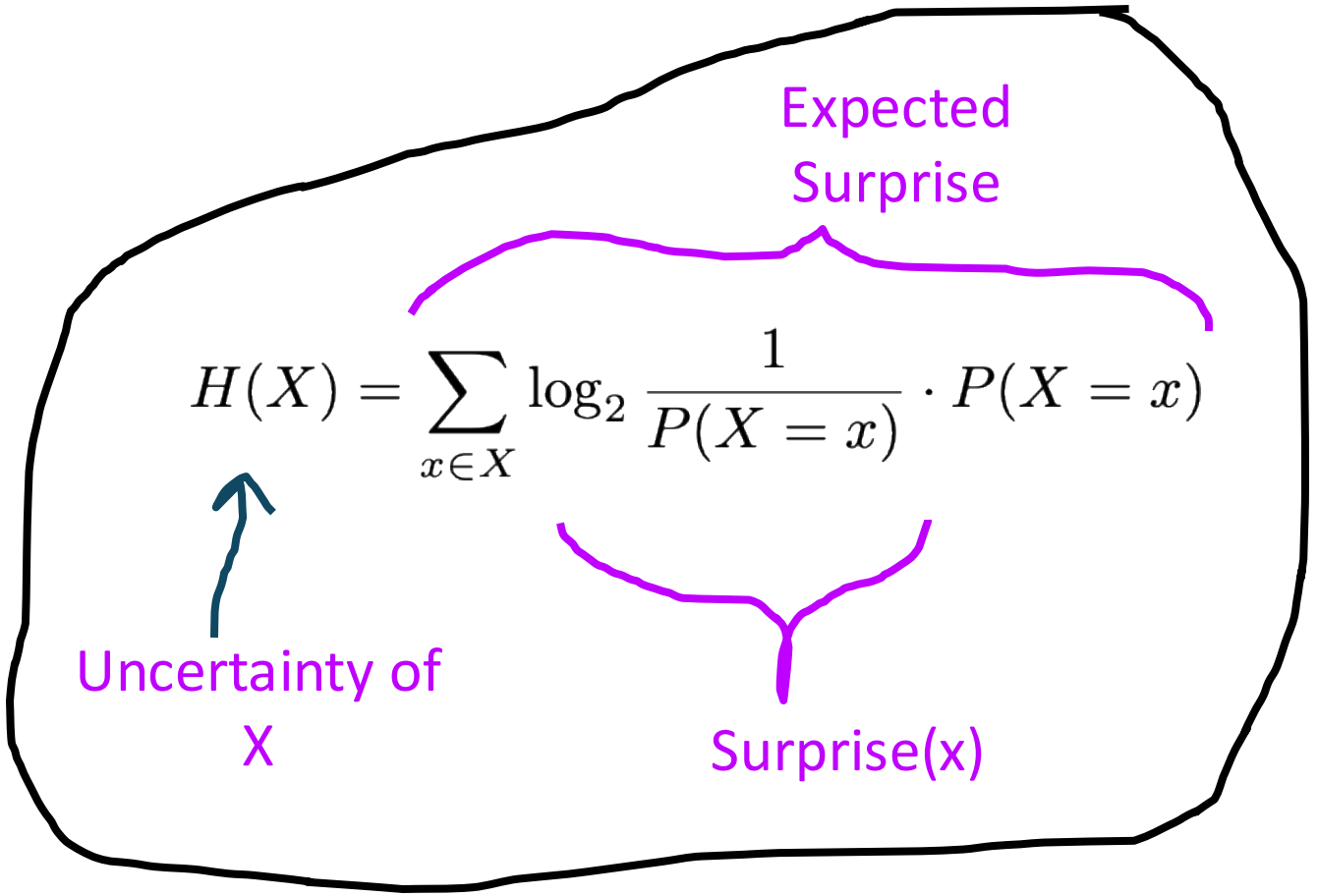
no
p = 0.83



$H(X) = 2$

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
    surprise_x = np.log2(1/p_x)
    uncertainty += surprise_x * p_x
  return uncertainty
```



To the code!

Limitations of Expected Entropy for Decision Making?

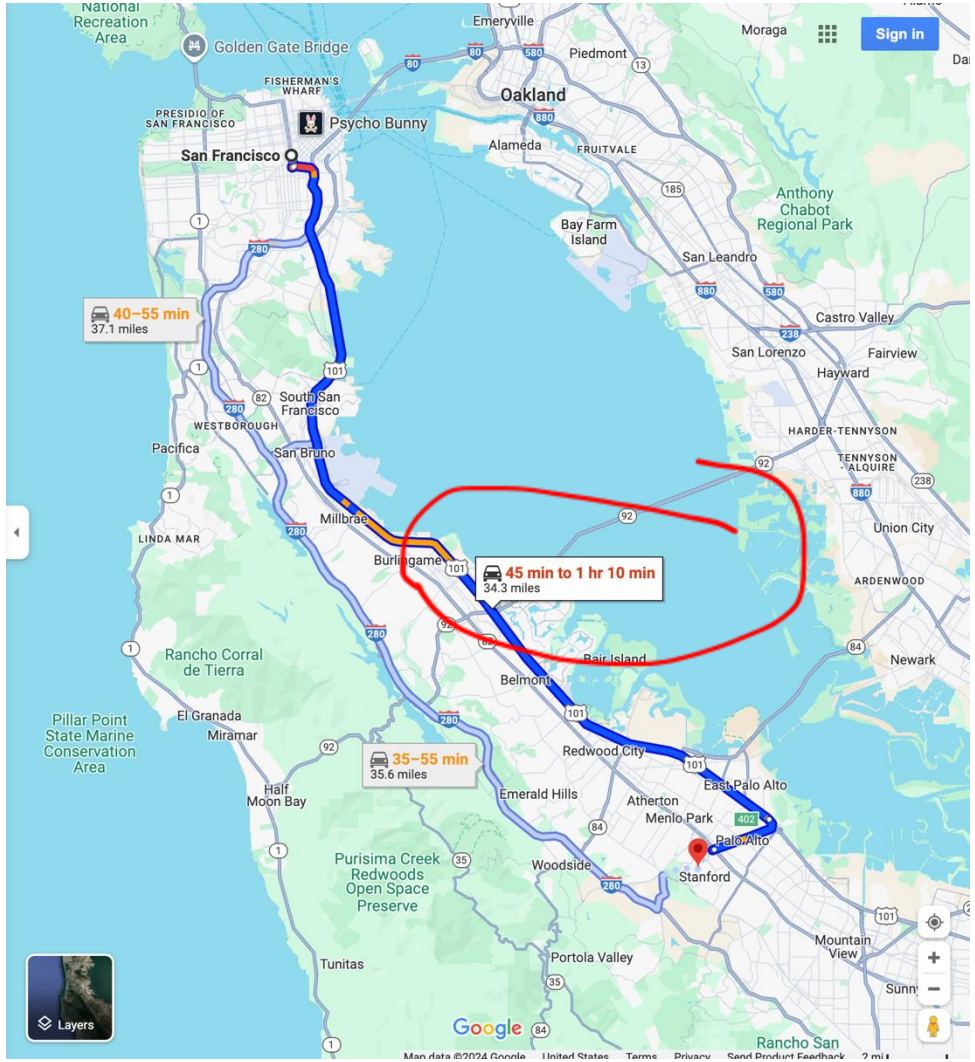
Entropy in the sum of two dice.

What is more informative:
I tell you that the **sum is odd**
I tell you the value of the **first dice is 1**

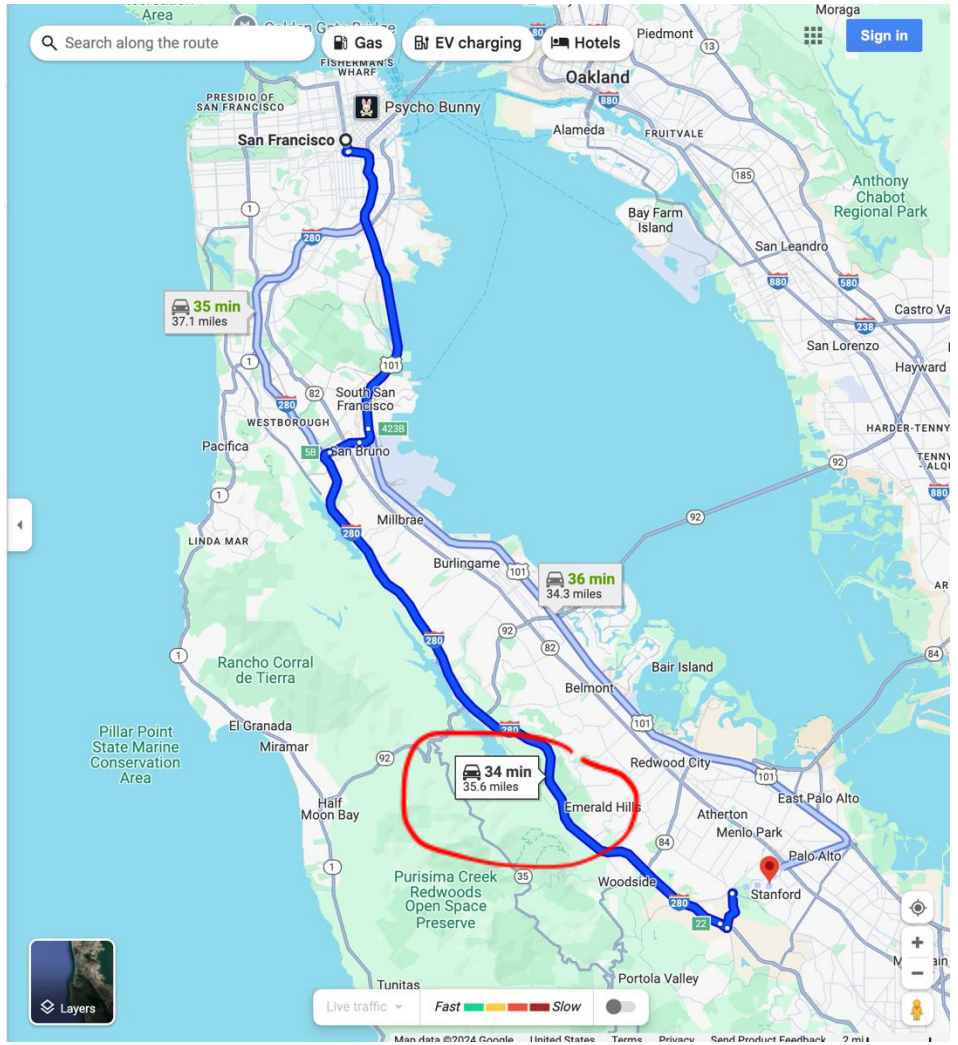
Traffic Predictions Over Time

Should You Wait to Plan Your Route?

Prediction 1 day in advance

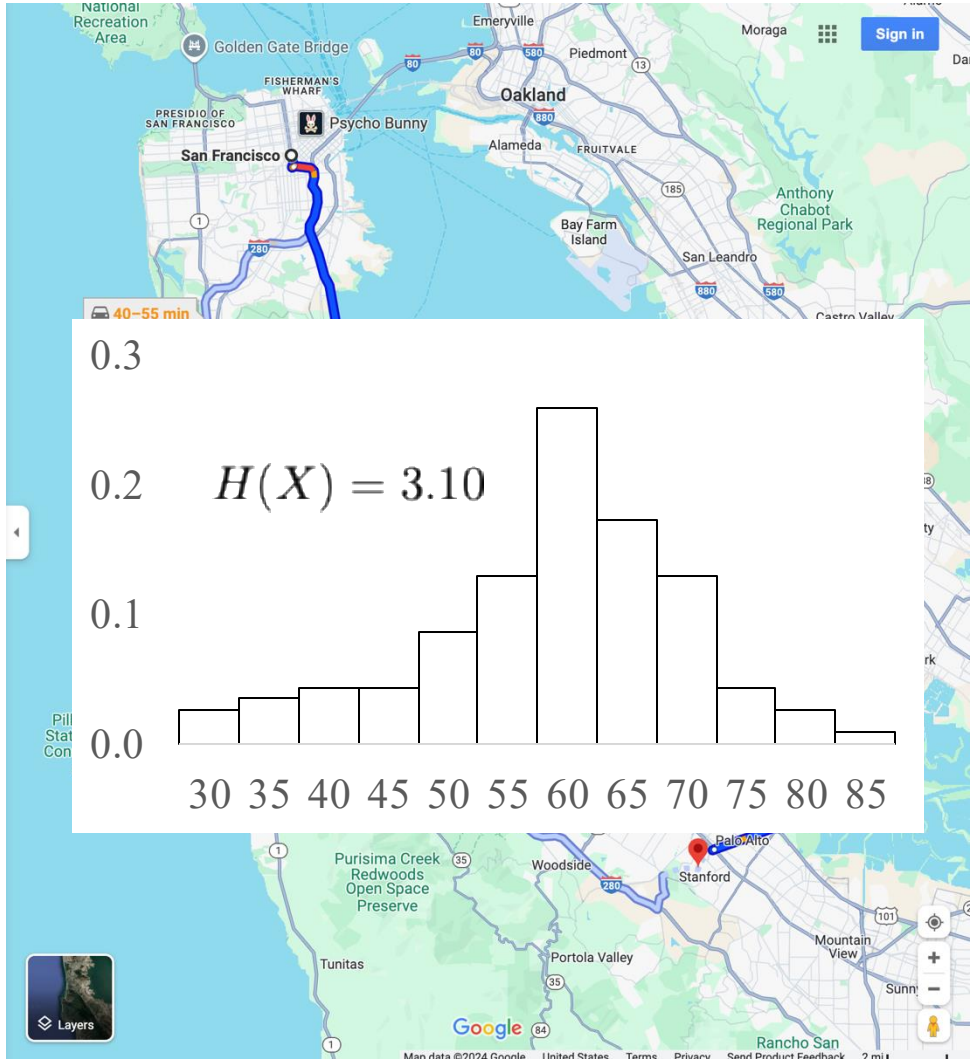


Prediction 30 mins in advance

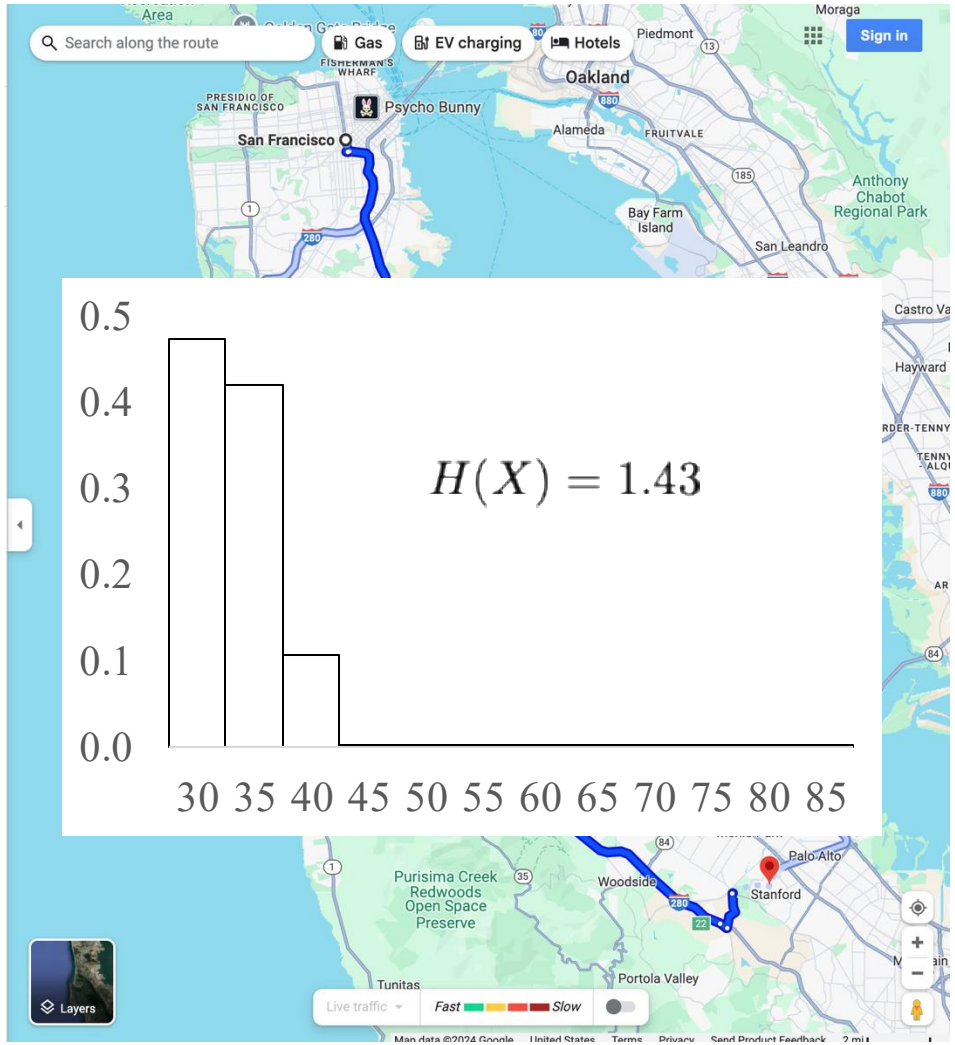


Should You Wait to Plan Your Route?

Prediction 1 day in advance

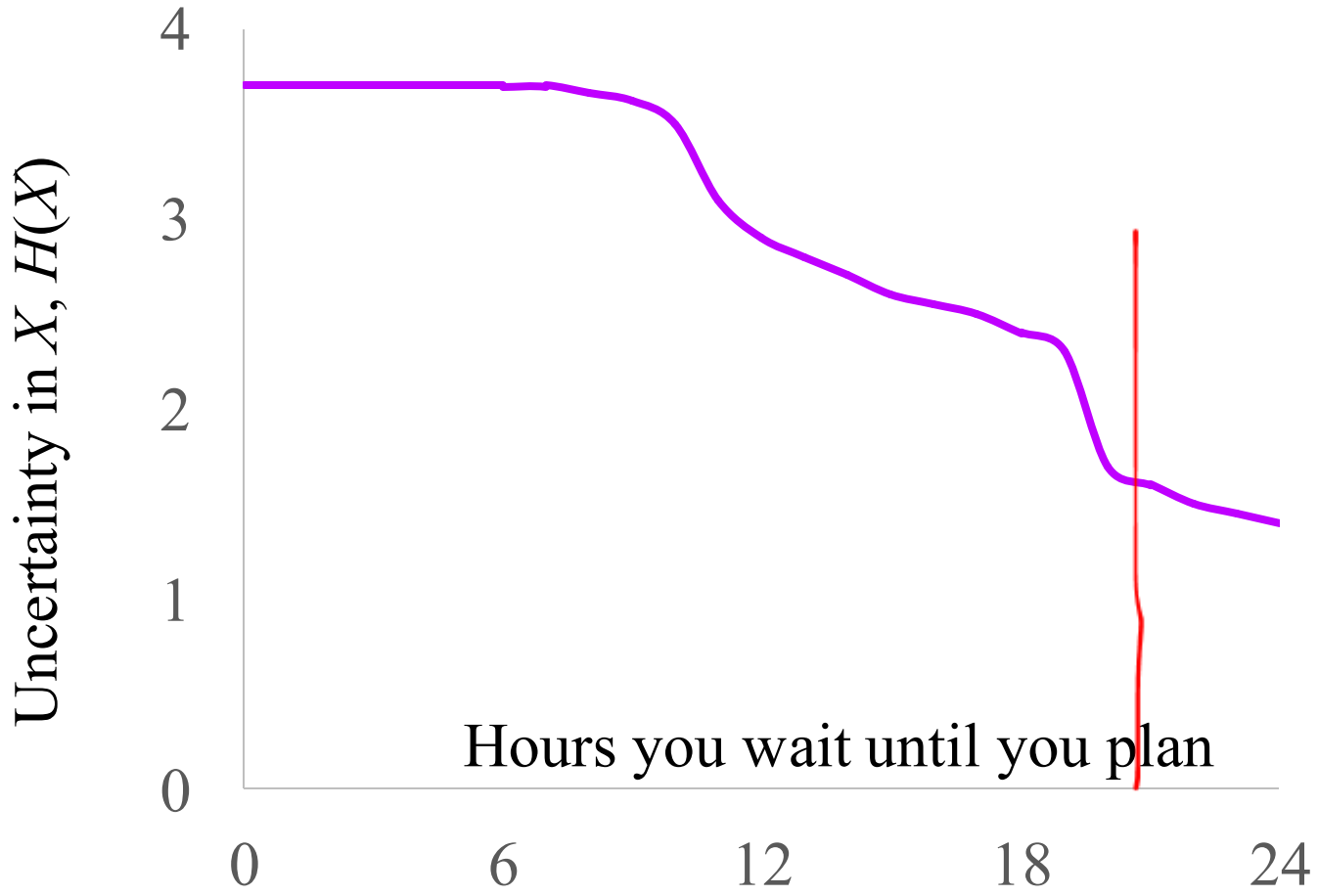


Prediction 30 mins in advance



Should You Wait to Plan Your Route?

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



If time:

Indus Valley Script. A language?

The Indus civilization - one of the world's earliest urban societies - emerged 5,300 years ago. The script they used is one of the last remaining undeciphered alphabets.

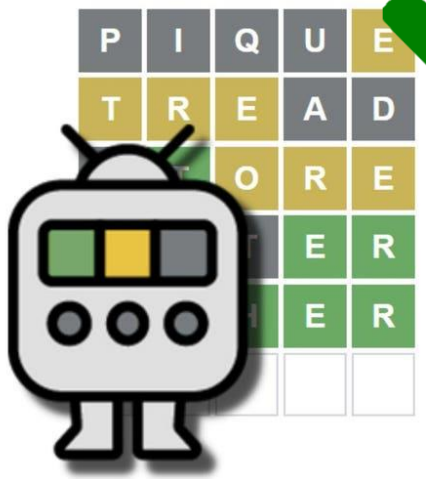
Let X be the first char
Let Y be the second char

a) What is $P(Y = y \mid X = x)$?
b) $E[H(Y \mid X = x)]$

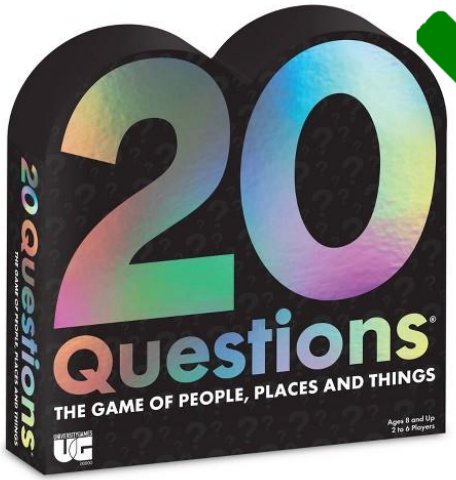
We are using entropy analysis to show that the script represents a true linguistic system. Let **`all_examples`** be the list of all recorded examples of the script, where each item in the list is one example string:

```
all_examples = ['𓀀𓀁𓀂𓀃',
                '𓀄𓀅𓀆𓀇𓀈',
                '𓀉𓀊',
                ...]
```

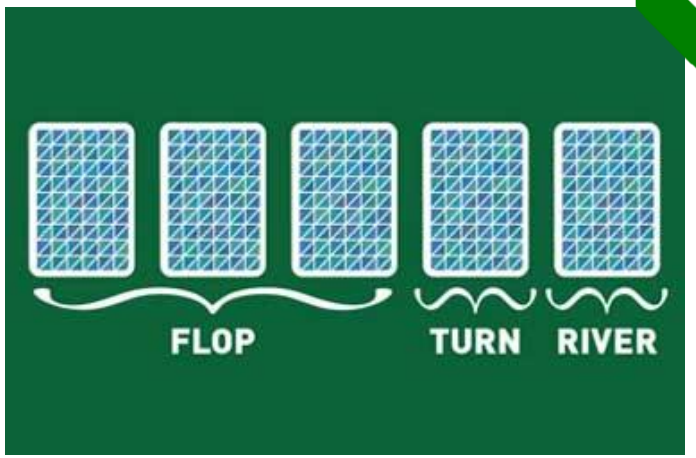
WorldeBot



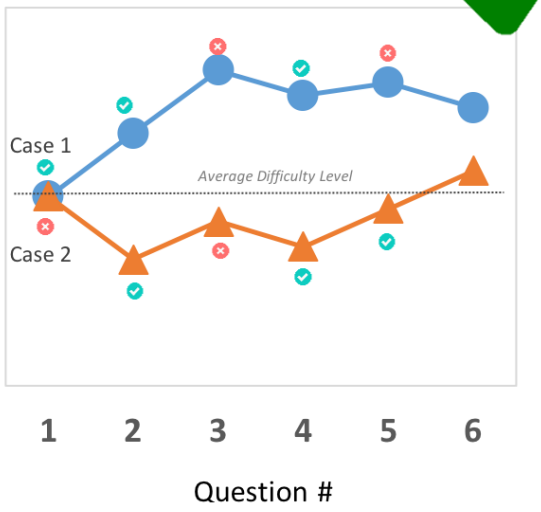
Decision Trees



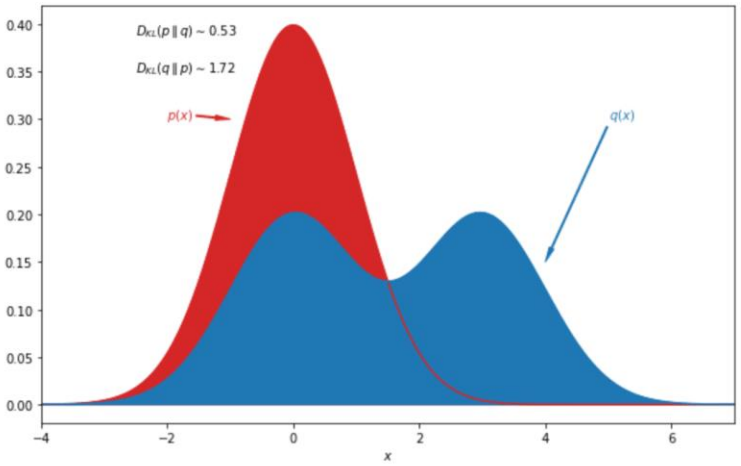
Value of Info in Poker



Adaptive Tests



Comparing Distributions

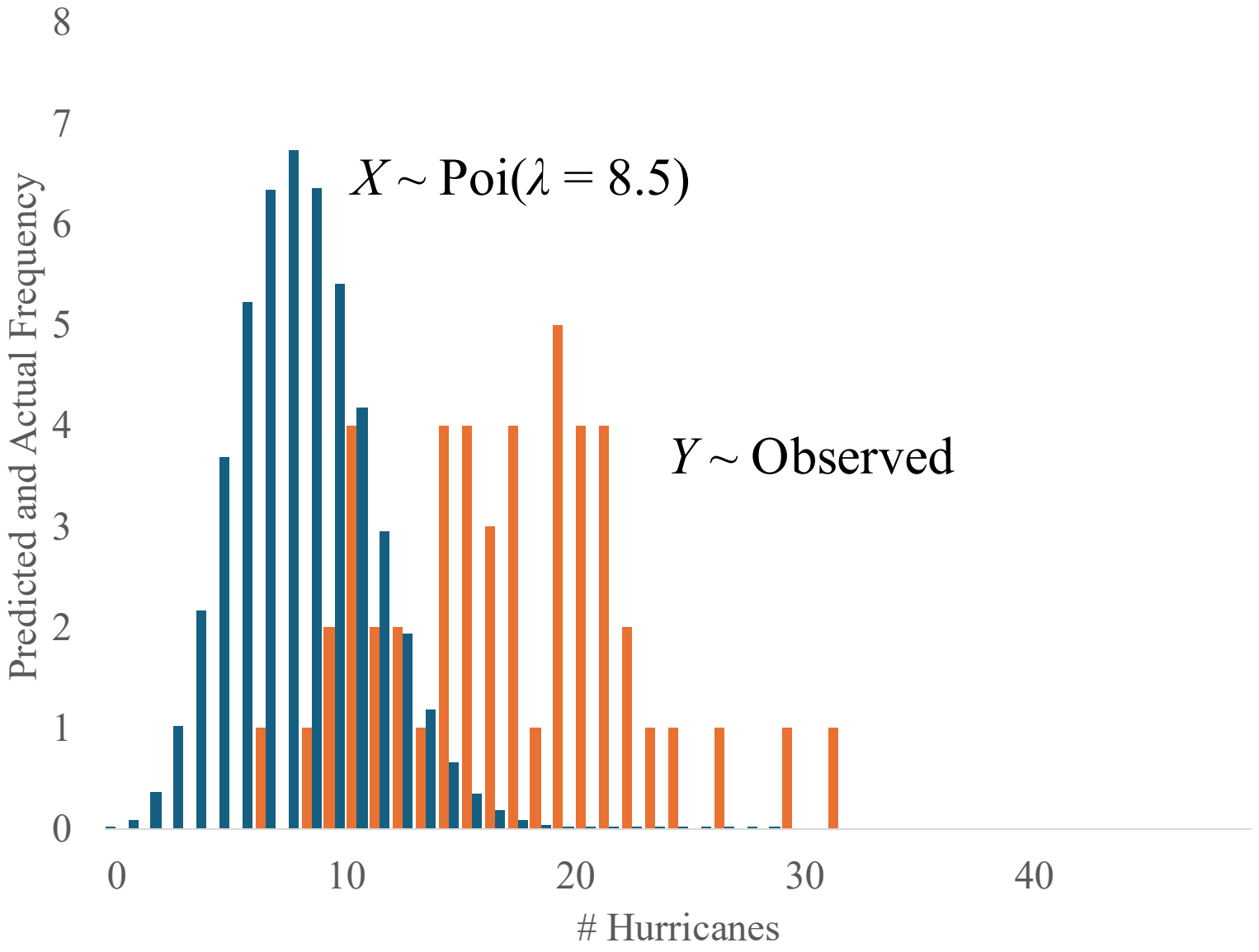


Compression of Data

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