Lecture 04
Information Theory

Chao Zhang
Georgia Tech

These slides are based on slides from Le Song, Roni Rosenfeld, and Maneesh Sahani.

Outline

- Motivation
- Entropy
- Conditional Entropy and Mutual Information
- Cross-Entropy and KL-Divergence

Uncertainty and Information

- information $\neq$ knowledge
  Concerned with abstract possibilities, not their meaning
- information: reduction in uncertainty

Imagine:

\#1 you're about to observe the outcome of a coin flip

\#2 you're about to observe the outcome of a die roll

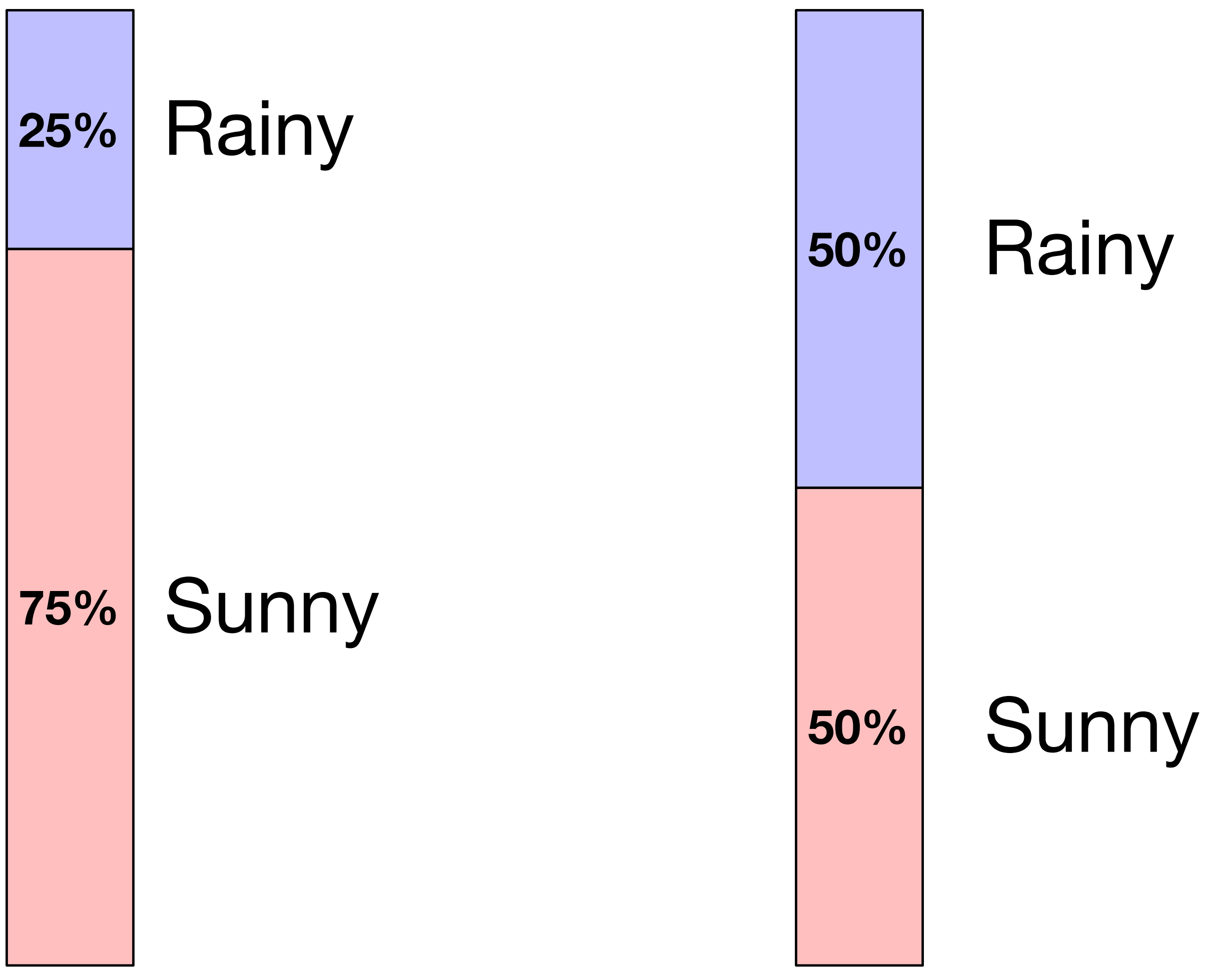
There is more uncertainty in #2

Next:

1. You observed outcome of #1 $\rightarrow$ uncertainty reduced to zero.
2. You observed outcome of #2 $\rightarrow$ uncertainty reduced to zero.

$\Longrightarrow$ more information was provided by the outcome in #2

Uncertainty and Information

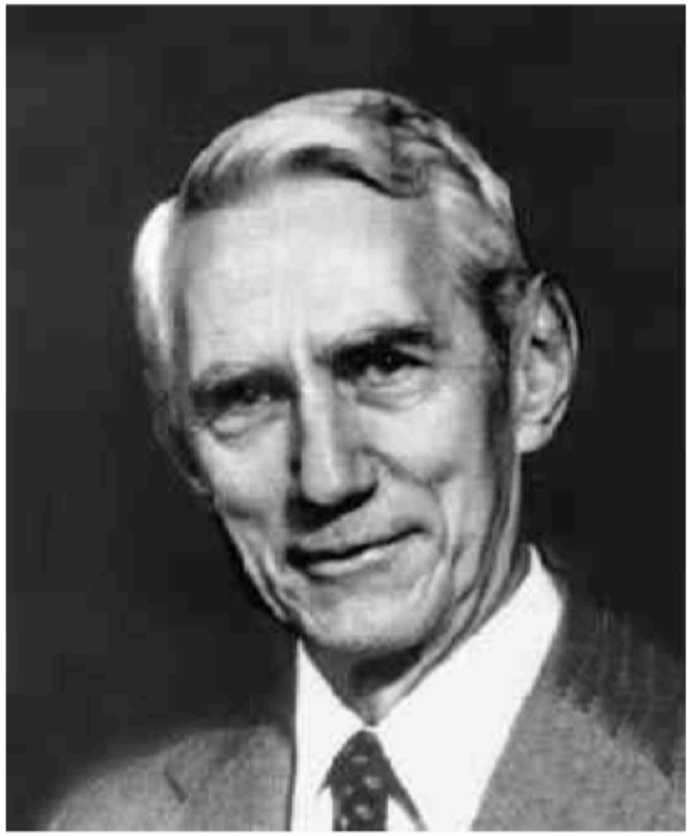


Which day is more uncertain?

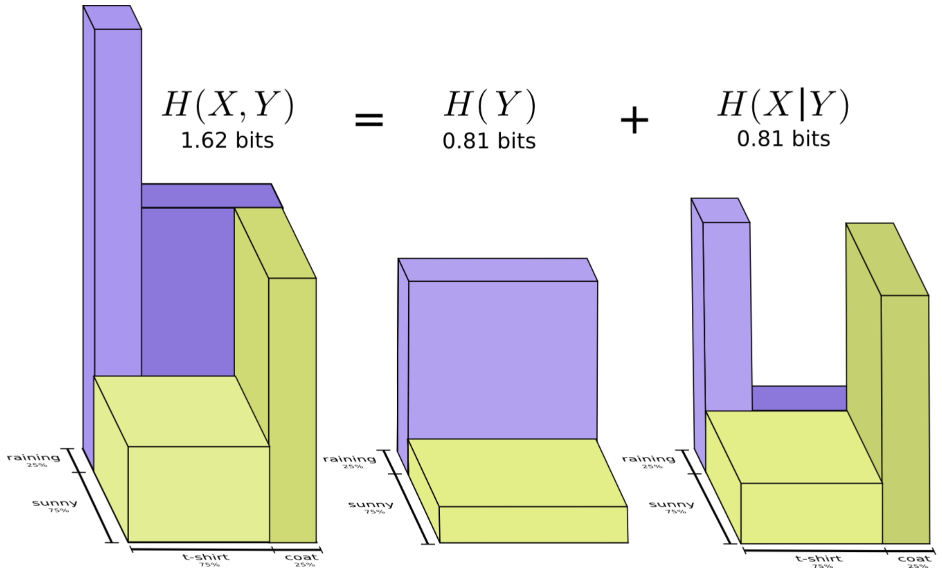
How do we quantify uncertainty?

Information Theory

- Information theory is a mathematical framework which addresses questions like:
  - How much information does a random variable carry about?
  - How efficient is a hypothetical code, given the statistics of the random variable?
  - How much better or worse would another code do?
  - Is the information carried by different random variables complementary or redundant?



Claude Shannon



Outline

- Motivation
- Entropy
- Conditional Entropy and Mutual Information
- Cross-Entropy and KL-Divergence

Entropy

- Entropy $H(Y)$ of a random variable Y

$$H(Y) = -\sum_{k=1}^{K} P(y = k) \log_2 P(y = k)$$

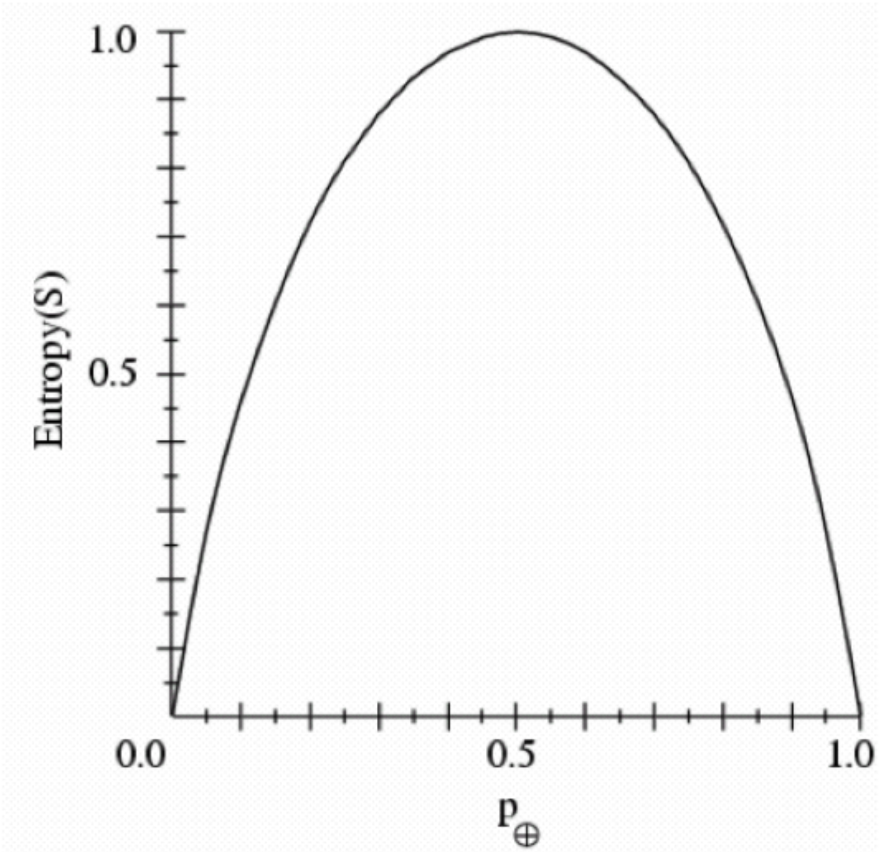
- $H(Y)$ is the expected number of bits needed to encode a randomly drawn value of Y (under most efficient code)

- Information theory:

  Most efficient code assigns $-\log_2 P(Y = k)$ bits to encode the message $Y = k$, So, expected number of bits to code one random Y is:

$$-\sum_{k=1}^{K} P(y = k) \log_2 P(y = k)$$

Entropy



- S is a sample of coin flips
- p_+ is the proportion of heads in S
- p_- is the proportion of tails in S
- Entropy measure the uncertainty of S

$$H(S) \equiv -p_+ \log_2 p_+ - p_- \log_2 p_-$$

Entropy Computation: An Example

$$H(S) \equiv -p_+ \log_2 p_+ - p_- \log_2 p_-$$

head	0
tail	6

P(h) = 0/6 = 0   P(t) = 6/6 = 1

Entropy = – 0 log 0 – 1 log 1 = – 0 – 0 = 0

head	1
tail	5

P(h) = 1/6   P(t) = 5/6

Entropy = – (1/6) $\log_2$ (1/6) – (5/6) $\log_2$ (1/6) = 0.65

head	2
tail	4

P(h) = 2/6   P(t) = 4/6

Entropy = – (2/6) $\log_2$ (2/6) – (4/6) $\log_2$ (4/6) = 0.92

Properties of Entropy

$$H(P) \;=\; \sum_i p_i \cdot \log \frac{1}{p_i}$$

1. Non-negative: $H(P) \geq 0$
2. Invariant wrt permutation of its inputs:
   $H(p_1, p_2, \ldots, p_k) = H(p_{\tau(1)}, p_{\tau(2)}, \ldots, p_{\tau(k)})$
3. For any *other* probability distribution $\{q_1, q_2, \ldots, q_k\}$:
   $$H(P) \;=\; \sum_i p_i \cdot \log \frac{1}{p_i} \;<\; \sum_i p_i \cdot \log \frac{1}{q_i}$$
4. $H(P) \leq \log k$, with equality iff $p_i = 1/k \;\; \forall i$
5. The further P is from uniform, the lower the entropy.

Outline

- Motivation
- Entropy
- Conditional Entropy and Mutual Information
- Cross-Entropy and KL-Divergence

Joint Entropy

	cold	mild	hot	
low	0.1	0.4	0.1	0.6
high	0.2	0.1	0.1	0.4
	0.3	0.5	0.2	1.0

- $H(T) = H(0.3, 0.5, 0.2) = 1.48548$
- $H(M) = H(0.6, 0.4) = 0.970951$
- $H(T) + H(M) = 2.456431$
- **Joint Entropy**: consider the space of (t, m) events $H(T, M) = \sum_{t,m} P(T = t, M = m) \cdot \log \frac{1}{P(T=t, M=m)}$
  $H(0.1, 0.4, 0.1, 0.2, 0.1, 0.1) = 2.32193$

Notice that $H(T, M) < H(T) + H(M)$!!!

Conditional Entropy

$$P(T = t|M = m)$$

	cold	mild	hot	
low	1/6	4/6	1/6	1.0
high	2/4	1/4	1/4	1.0

Conditional Entropy:

- $H(T|M = low) = H(1/6, 4/6, 1/6) = 1.25163$
- $H(T|M = high) = H(2/4, 1/4, 1/4) = 1.5$
- **Average Conditional Entropy** (aka equivocation):
  $H(T/M) = \sum_m P(M = m) \cdot H(T|M = m) =$
  $0.6 \cdot H(T|M = low) + 0.4 \cdot H(T|M = high) = 1.350978$

Conditional Entropy

$$P(M = m|T = t)$$

	cold	mild	hot
low	1/3	4/5	1/2
high	2/3	1/5	1/2
	1.0	1.0	1.0

Conditional Entropy:

- $H(M|T = cold) = H(1/3, 2/3) = 0.918296$
- $H(M|T = mild) = H(4/5, 1/5) = 0.721928$
- $H(M|T = hot) = H(1/2, 1/2) = 1.0$
- Average Conditional Entropy (aka Equivocation):
  $H(M/T) = \sum_t P(T = t) \cdot H(M|T = t) =$
  $0.3 \cdot H(M|T = cold) + 0.5 \cdot H(M|T = mild) + 0.2 \cdot H(M|T = hot) = 0.8364528$

Conditional Entropy

- Conditional entropy $H(Y|X)$ of a random variable Y given X_i

$$H(Y|X_i) = -\int \left(\sum_{k=1}^{K} P(y = k|x_i) \log_2 P(y = k) \right) p(x_i) dx_i$$

- Quantify the uncerntainty in Y after seeing feature X_i

- $H(Y)$ is the expected number of bits needed to encode a randomly drawn value of Y
  - given X_i, and
  - average over the likelihood of seeing particular value of x_i

Mutual Information

- Mutual information: quantify the reduction in uncertainty in Y after seeing feature X_i

$$I(X_i, Y) = H(Y) - H(Y|X_i)$$

- The more the reduction in entropy, the more informative a feature.

- Mutual information is symmetric
  - $I(X_i, Y) = I(Y, X_i) = H(X_i) - H(X_i|Y)$
  - $I(Y, X_i) = \int \sum_k^K p(x_i, y = k) \log_2 \frac{p(x_i, y=k)}{p(x_i)p(y=k)} \, dx_i$
  - $= \int \sum_k^K p(x_i|y = k)p(y = k) \log_2 \frac{p(x_i|y = k)}{p(x_i)} \, dx_i$

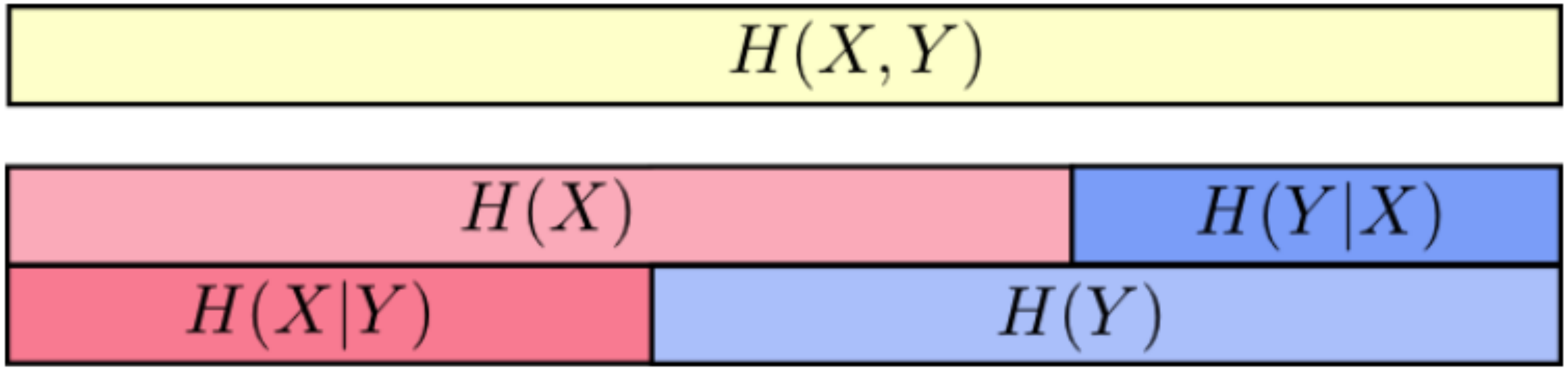
Properties of Mutual Information

$$
\begin{aligned}
I(X;Y) &= H(X) - H(X/Y) \\
&= \sum_x P(x) \cdot \log \frac{1}{P(x)} - \sum_{x,y} P(x,y) \cdot \log \frac{1}{P(x|y)} \\
&= \sum_{x,y} P(x,y) \cdot \log \frac{P(x|y)}{P(x)} \\
&= \sum_{x,y} P(x,y) \cdot \log \frac{P(x,y)}{P(x)P(y)}
\end{aligned}
$$

Properties of Average Mutual Information:

- Symmetric (but $H(X) \neq H(Y)$ and $H(X/Y) \neq H(Y/X)$)
- Non-negative (but $H(X) - H(X/y)$ may be negative!)
- Zero iff X, Y independent

CE and MI: Visual Illustration



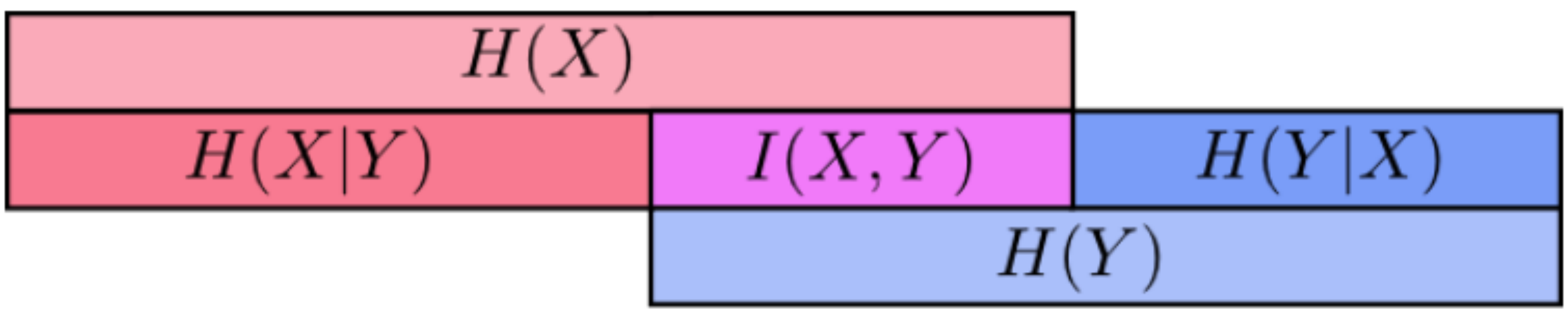


Image Credit: Christopher Olah.

Outline

- Motivation
- Entropy
- Conditional Entropy and Mutual Information
- Cross-Entropy and KL-Divergence

Cross Entropy

Cross Entropy: The expected number of bits when a wrong distribution Q is assumed while the data actually follows a distribution P

$$H(p,q) = -\sum_{x \in \mathcal{X}} p(x)\, \log q(x)$$

This is because:

$$H(p,q) = \mathrm{E}_p[l_i] = \mathrm{E}_p\left[\log \frac{1}{q(x_i)}\right]$$

$$H(p,q) = \sum_{x_i} p(x_i)\, \log \frac{1}{q(x_i)}$$

$$H(p,q) = -\sum_{x} p(x)\, \log q(x).$$

Kullback-Leibler Divergence (Relative Entropy)

Another useful information theoretic quantity measures the difference between two distributions.

$$\begin{aligned}\mathbf{KL}[P(S)\|Q(S)] &= \sum_s P(s)\log\frac{P(s)}{Q(s)} \\ &= \underbrace{\sum_s P(s)\log\frac{1}{Q(s)}}_{\text{cross entropy}} - \mathbf{H}[P]\end{aligned}$$

Excess cost in bits paid by encoding according to Q instead of P.

$$\begin{aligned}-\mathbf{KL}[P\|Q] &= \sum_s P(s)\log\frac{Q(s)}{P(s)} \\ &\leq \log\sum_s P(s)\frac{Q(s)}{P(s)} \qquad \text{by Jensen} \\ &= \log\sum_s Q(s) = \log 1 = 0\end{aligned}$$

So $\mathbf{KL}[P\|Q] \geq 0$. Equality iff $P = Q$

Take-Home Messages

- Entropy
  - A measure for uncertainty
  - Why it is defined in this way (optimal coding)
  - Its properties
- Joint Entropy, Conditional Entropy, Mutual Information
  - The physical intuitions behind their definitions
  - The relationships between them
- Cross Entropy, KL Divergence
  - The physical intuitions behind them
  - The relationships between entropy, cross-entropy, and KL divergence