Information Theory can be used to characterize probability distributions or quantify similarity between probability distributions.

• Likely events should have low information content, and in the extreme case, events that are guaranteed to happen should have no information content whatsoever.

• Less likely events should have higher information content.

• Independent events should have additive information. For example, finding out that a tossed coin has come up as heads twice should convey twice as much information as finding out that a tossed coin has come up as heads once.

In order to satisfy all three of these properties, we define the self-information of an event x = x to be

$$I(x) = -\log P(x).$$

The definition of I(x) is therefore written in unit of **nats**. One nat is the amount of information gained by observing an event of probability 1/e.

Self-information deals only with a single outcome. We can quantify the amount of uncertainty in an entire probability distribution using the Shannon entropy:

$$H(\mathrm{x}) = \mathbb{E}_{\mathrm{x}\sim P}[I(x)] = -\mathbb{E}_{\mathrm{x}\sim P}[\log P(x)].$$

If we have two separate probability distributions P(x) and Q(x) over the same random variable x, we can measure how different these two distributions are using the Kullback-Leibler (KL) divergence:

$$D_{\mathrm{KL}}(P\|Q) = \mathbb{E}_{\mathrm{x}\sim P}\left[\log \frac{P(x)}{Q(x)}\right] = \mathbb{E}_{\mathrm{x}\sim P}\left[\log P(x) - \log Q(x)\right]$$