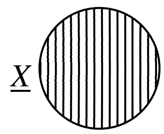
Mutual Information

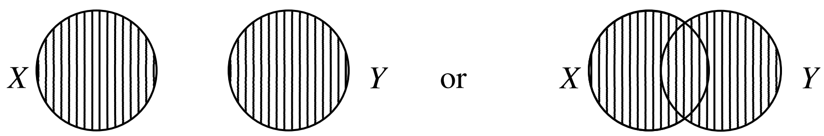
Given a random variable X, we can represent the information $H(X)$ by a diagram showing X's information as a region of the space of all information.

$H(X)$



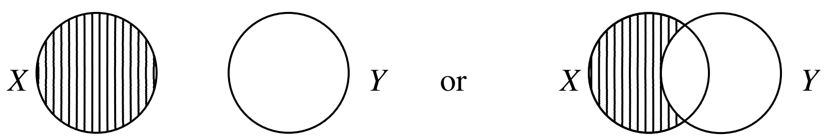
If we have two random variables (independent or not), the information conveyed jointly by them $H(X,Y)$ as the set union of the information conveyed by each.

$H(X,Y)$



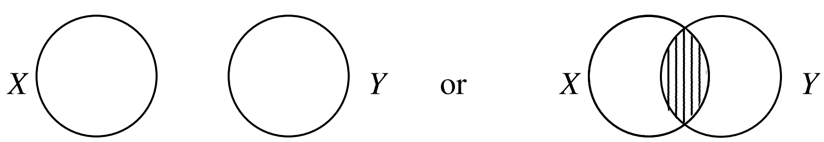
The information conveyed by X given Y, $H(X|Y)$, is represented by the set difference (not symmetric difference) of X and Y, because, if Y is known, all the information that was conveyed by Y is now given, and no longer conveys any information.

$H(X|Y)$



The mutual information of random variables X and Y is defined to be $I(X,Y) = H(X) - H(X|Y)$. In this setting, – once again is represented by set difference in our diagram.

$I(X,Y)$



So the information *mutual* to X and Y is that which is shared by both of them. This can also be computed as follows: $I(X,Y) = H(X) + H(Y) - H(X,Y)$. In this case, + is represented by union in our diagram if we consider the diagram to represent multisets, that is, sets where the multiplicity of an element can be greater than 1. From $H(X)$, we get multiplicity 1 for each element of X. From $H(Y)$, we add multiplicity 1 to each element of Y, making the elements common to X and Y have multiplicity 2. Finally, we subtract multiplicity one from each element that belongs to either X or Y, thus leaving the elements common to X and Y with multiplicity 1 and all others with multiplicity 0.

We can extend this idea of mutual information to what is known as *interaction information* when more than two variables are involved. In such a case, we define $I(X,Y,Z) = -H(X) - H(Y) - H(Z) + H(X,Y) + H(X,Z) + H(Y,Z) - H(X,Y,Z)$. It is the information that is common to X,Y, and Z. Information appearing only in X is assigned multiplicity $-1 - 0 - 0 + 1 + 1 + 0 - 1 = 0$ by the expression above. Information belonging to X and Y but not Z is assigned multiplicity $-0 - 0 - 0 + 1 + 0 + 0 - 1 = 0$. Information belonging to X and Y and Z is assigned $-1 - 1 - 1 + 1 + 1 + 1 - 1 = -1$, that is, a negative value! This is one of the oddities of interaction information. The general form for the interaction information of a set of random variables $V = \{X_1,..., X_n\}$ is given by

$$I(V) = -\sum_{S \subseteq V} -1^{|V|-|S|} H(S).$$