Conditional entropy of an ensemble X, given that y = bj

measures the uncertainty remaining about random variable X after specifying that random variable Y has taken on a particular value y = bj. It is defined naturally as the entropy of the probability distribution
p(x|y = bj):

$$H(X|y=b_j) = \sum_x p(x|y=b_j) \log \frac{1}{p(x|y=b_j)} \tag{7}$$

If we now consider the above quantity *averaged* over all possible outcomes that Y might have, each weighted by its probability p(y), then we arrive at the...

Conditional entropy of an ensemble X, given an ensemble Y:

$$H(X|Y) = \sum_y p(y) \left[\sum_x p(x|y) \log \frac{1}{p(x|y)} \right] \tag{8}$$

and we know from the Sum Rule that if we move the p(y) term from the outer summation over y, to inside the inner summation over x, the two probability terms combine and become just p(x, y) summed over all x, y. Hence a simpler expression for this conditional entropy is:

$$H(X|Y) = \sum_{x,y} p(x,y) \log \frac{1}{p(x|y)} \tag{9}$$

This measures the average uncertainty that remains about X, when Y is known.

Chain Rule for Entropy
The joint entropy, conditional entropy, and marginal entropy for two
ensembles X and Y are related by:

$$H(X, Y) = H(X) + H(Y|X) = H(Y) + H(X|Y) \tag{10}$$

It should seem natural and intuitive that the joint entropy of a pair of random variables is the entropy of one plus the conditional entropy of the other (the uncertainty that it adds once its dependence on the first one has been discounted by conditionalizing on it). You can derive the Chain Rule from the earlier definitions of these three entropies. Corollary to the Chain Rule: If we have three random variables X, Y, Z, the conditionalizing of the joint distribution of any two of them, upon the third, is also expressed by a Chain Rule:

$$H(X, Y|Z) = H(X|Z) + H(Y|X,Z) \tag{11}$$

“Independence Bound on Entropy”
A consequence of the Chain Rule for Entropy is that if we have many different random variables X1, X2, ..., Xn, then the sum of all their individual entropies is an upper bound on their joint entropy:

$$H(X_1, X_2, ..., X_n) \leq \sum_{i=1}^{n} H(X_i) \tag{12}$$

Their joint entropy only reaches this upper bound if all of the random variables are independent.

Mutual Information between X and Y

The *mutual information* between two random variables measures the amount of information that one conveys about the other. Equivalently, it measures the average reduction in uncertainty about X that results from learning about Y . It is defined:

$$I(X;Y) = \sum_{x,y} p(x,y) \log \frac{p(x,y)}{p(x)p(y)} \tag{13}$$

Clearly X says as much about Y as Y says about X. Note that in case X and Y are independent random variables, then the numerator inside the logarithm equals the denominator. Then the log term vanishes, and the mutual information equals zero, as one should expect. Non-negativity: mutual information is always ≥ 0. In the event that the two random variables are perfectly correlated, then their mutual information is the entropy of either one alone. (Another way to say this is: I(X; X) = H(X): the mutual information of a random variable with itself is just its entropy. For this reason, the entropy H(X) of a random variable X is sometimes referred to as its *self-information*.) These properties are reflected in three equivalent definitions for the mutual information between X and Y :

$$I(X; Y) = H(X) - H(X|Y) \tag{14}$$

$$I(X; Y) = H(Y) - H(Y |X) = I(Y ;X) \tag{15}$$

$$I(X; Y) = H(X) + H(Y) - H(X, Y) \tag{16}$$

In a sense the mutual information I(X; Y) is the intersection between H(X) and H(Y), since it represents their statistical dependence. In the Venn diagram given at the top of page 18, the portion of H(X) that does not lie within I(X; Y) is just H(X|Y). The portion of H(Y) that does not lie within I(X; Y) is just H(Y |X).

Distance D(X, Y) between X and Y

The amount by which the joint entropy of two random variables exceeds their mutual information is a measure of the "*distance*" between them:

$$D(X, Y) = H(X, Y) - I(X; Y) \tag{17}$$

Note that this quantity satisfies the standard axioms for a distance:

D(X, Y) ≥ 0, D(X,X) = 0, D(X, Y) = D(Y,X), and D(X,Z) ≤ D(X, Y) + D(Y,Z).