Expectation	$\mathbb{E}[cX] = c\mathbb{E}[X]$	$\mathbb{E}[X+Y] = \mathbb{E}[X] + \mathbb{E}[X]$	$\mathbb{E}[XY] = \mathbb{E}[X]\mathbb{E}[Y]$ if $X \perp Y$

Conditioning	$P_{Y\|X}(y\|x) = \frac{P_{XY}(x,y)}{P_X(x)}$	$P_{Y\|X}(y\|x) = \frac{P_Y(y)P_{X\|Y}(x\|y)}{P_X(x)}$

Properties of logarithms	$\log 1/x = -\log x$	$\log y/x = \log y - \log x$
$\log xy = \log x + \log y$	$\log x^c = c \log x$	$\log_a x = \frac{\log_b x}{\log_b a}$

Axioms for Information of an Event $\psi(p)$	
Non-negative	$\psi(p) \geq 0$
Zero for discrete events	$\psi(1) = 0$
Monotonicity	$p \leq p' \Longrightarrow \psi(p) \geq \psi(p')$
Continuity	$\psi(p)$ is continuous
Additivity under independence	$\psi(p_1 p_2) = \psi(p_1) + \psi(p_2)$ if $p_1 \perp p_2$

Information Entropy	$\psi(p) = \log_b \frac{1}{p}$ for $b > 0$
Shannon Entropy	$H(X) = \mathbb{E}_{X \sim P_X}\left[\log_2 \frac{1}{P_X(x)}\right] = \sum_x P_X(x) \log_2 \frac{1}{P_X(x)}$
Binary Distribution	$H(X) = p \log_2 \frac{1}{p} + (1-p) \log_2 \frac{1}{1-p}$
Uniform Distribution	$H(X) = \log_2 \|X\|$
Joint Entropy	$H(X,Y) = \sum_{x,y} P_{XY}(x,y) \log_2 \frac{1}{P_{XY}(x,y)}$
Conditional Entropy	$H(Y\|X) = \sum_{x,y} P_{XY}(x,y) \log_2 \frac{1}{P_{Y\|X}(y\|x)} = \sum_x P_X(x) H(Y\|X=x)$ $H(Y\|X=x) = \sum_y P_{Y\|X}(y\|x) \log_2 \frac{1}{P_{Y\|X}(y\|x)}$
Axiom	
Continuity	$\psi(p)$ is continuous.
Uniform case	$\psi(p)$ is increasing with N if $p_i = \frac{1}{N}; i \in 1,2,\dots,N$
Successive decisions	$\psi(p_1, p_2, \dots, p_1) = \psi(p_1 + p_2, p_3, \dots, p_1) + (p_1 + p_2)\psi\left(\frac{p_1}{p_1+p_2}, \frac{p_2}{p_1+p_2}\right)$
Non-negative	$H(X) \geq 0$
Upper bound	$H(X) \leq \log_2 \|X\|$
Chain rule (two variables)	$H(X,Y) = H(X) + H(Y\|X)$
KL Divergence (relative entropy)	$D(P \parallel Q) = \sum_x P_X(x) \log_2 \frac{P_X(x)}{Q_X(x)}$

Mutual Information	$I(X;Y) = H(Y) - H(Y\|X)$
Joint mutual information	$I(X_1, X_2; Y_1, Y_2) = H(Y_1, Y_2) - H(Y_1, Y_2 \| X_1, X_2)$
Conditional mutual information	$I(X;Y\|Z) = H(Y\|Z) - H(Y\|X,Z)$
Axiom	
Alternative forms	$I(X;Y) = D(P_{XY} \parallel P_X P_Y)$ $I(X;Y) = \sum_{x,y} P_{XY}(x,y) \log_2 \frac{P_{XY}(x,y)}{P_X(x)P_Y(y)}$ $I(X;Y) = \sum_{x,y} P_{XY}(x,y) \log_2 \frac{P_{Y\|X}(y\|x)}{P_Y(y)}$
Symmetry	$I(X;Y) = H(X) + H(Y) - H(X,Y)$ $I(X;Y) = I(Y;X) \Rightarrow I(X;Y) = H(X) - H(X\|Y)$
Non-negative	$I(X;Y) \geq 0$
Upper bounds	$I(X;Y) < H(X) \leq \log_2 \|X\|$
Chain rule	$I(X_1, X_2, \dots, X_n; Y) = \sum_{i=1}^{n} I(X_i; Y \| X_1, \dots, X_{i-1})$
Data processing in equality	$I(X;Z) \leq I(X;Y)$ if $X \to Y \to Z$ $I(W;Z) \leq I(X;Y)$ if $W \to X \to Y \to Z$
Partial sub-additivity	$I(X_1, \dots, X_n; Y_1, \dots, Y_n) \leq \sum_{i=1}^{n} I(X_i; Y_i)$ if $(Y_1, \dots, Y_n)$ are conditionally independent and Y_i depends on $(X_1, \dots, X_n)$ only through X_i

Symbol-Wise Coding	$L(C) = \sum_{x\in X} P_X(x)\ell(x)$ $L(C)$ average length of code $C(x)$ $\ell(x)$ length of this sequence
Kraft's Inequality	Any prefix-free code that maps each $x \in X$ to a word of length $\ell(x)$ must satisfy $\sum_{x\in X} 2^{-\ell(x)} \le 1$
Entropy Bound	For $X \sim P_X$ and any prefix-free code, $L(C) \ge H(X)$ with equality if and only if $\forall x \in X, P_X(x) = 2^{-\ell(x)} \Leftrightarrow \ell(x) = \log_2 \frac{1}{P_X(x)}$
Shannon-Fano Code	Rounds the ideal lengths up to the nearest integer. $\ell(x) = \left\lceil \log_2 \frac{1}{P_X(x)} \right\rceil$ $\sum_{x\in X} 2^{-\ell(x)} = \sum_{x\in X} 2^{-\left\lceil \log_2 \frac{1}{P_X(x)} \right\rceil} \le \sum_{x\in X} 2^{-\log_2 \frac{1}{P_X(x)}} = \sum_{x\in X} P_X(x) = 1$ $L(C) = \sum_{x\in X} P_X(x) \left\lceil \log_2 \frac{1}{P_X(x)} \right\rceil < \sum_{x\in X} P_X(x) \left(\log_2 \frac{1}{P_X(x)} + 1 \right) = H(X) + 1$ $H(X) \le L(C) < H(X) + 1$
Huffman Code	Huffman code has the smallest possible average length. Construct a tree as follows: 1. List the symbols of X from highest probability from highest to lowest. 2. Draw a branch connecting the two symbols with the lowest probability and label the merged point with the sum of the two associated probabilities. 3. Repeat the first two steps (with the two original probabilities replaced by the merged probability) until everything has merged to a single point with total probability 1.