# Shannon Entropy

# Shannon

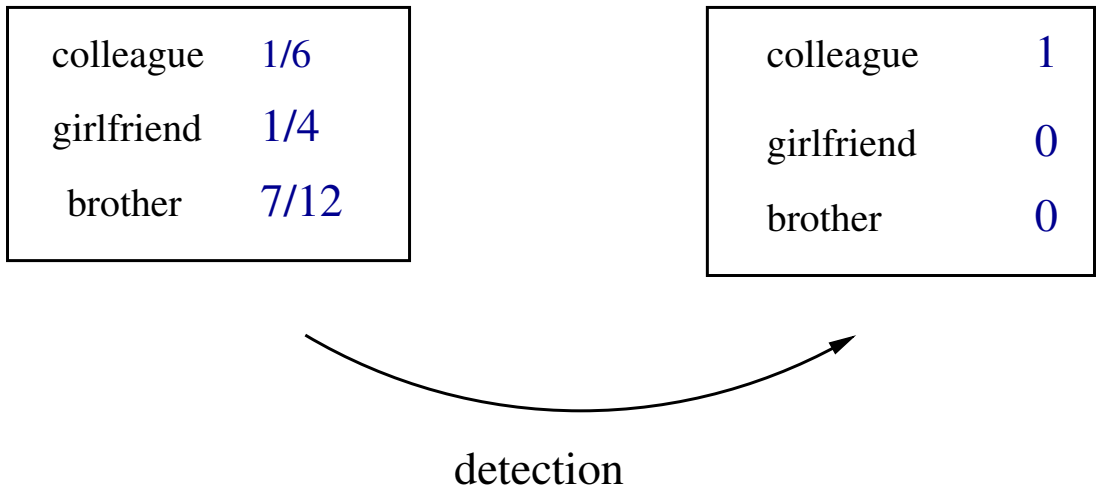
Claude Shannon

A Mathematical Theory of Communication (1948)

# Uncertainty and the Ideal Policeman

"knowing" $\equiv$ "certainty"
"information to learn" $\equiv$ "uncertainty"

- dead man
- list of suspects: resentful colleague, abandoned girlfriend, envious brother



# Uncertainty and Information

- information should be additive
- information in an event should measure "reduction in uncertainty" when the event occurs
- low probability $\Rightarrow$ high reduction in uncertainty
- highest when every possible event is equally likely

# Entropy (Information Quantity)

- uncertainty reduction when an event $a \in A$ occurs is $log_2 \frac{1}{p(a)}$
  - $\frac{1}{p(a)}$: low probability $\Rightarrow$ high reduction in uncertainty
  - $log_2$: information should be additive
  - 2: base 2 produces information "bits"
- get weighted average over all events: sum uncertainty reduction for each event weighted by the probability of each event

**Entropy of a set of events**

$$\mathcal{H}(A) = \sum_{a \in A} p(a) \log_2 \frac{1}{p(a)}$$

# Entropy calculations

| | |
|---|---|
| colleague | 1 |
| girlfriend | 0 |
| brother | 0 |

$$\mathcal{H} = 1.log_2 1 + 2.0.log_2 0 = 0$$

| | |
|---|---|
| colleague | 1/6 |
| girlfriend | 1/4 |
| brother | 7/12 |

$$\mathcal{H} = \frac{1}{6}.log_2 6 + \frac{1}{4}.log_2 4 + \frac{7}{12}.log_2 \frac{12}{7} = 1.3844$$

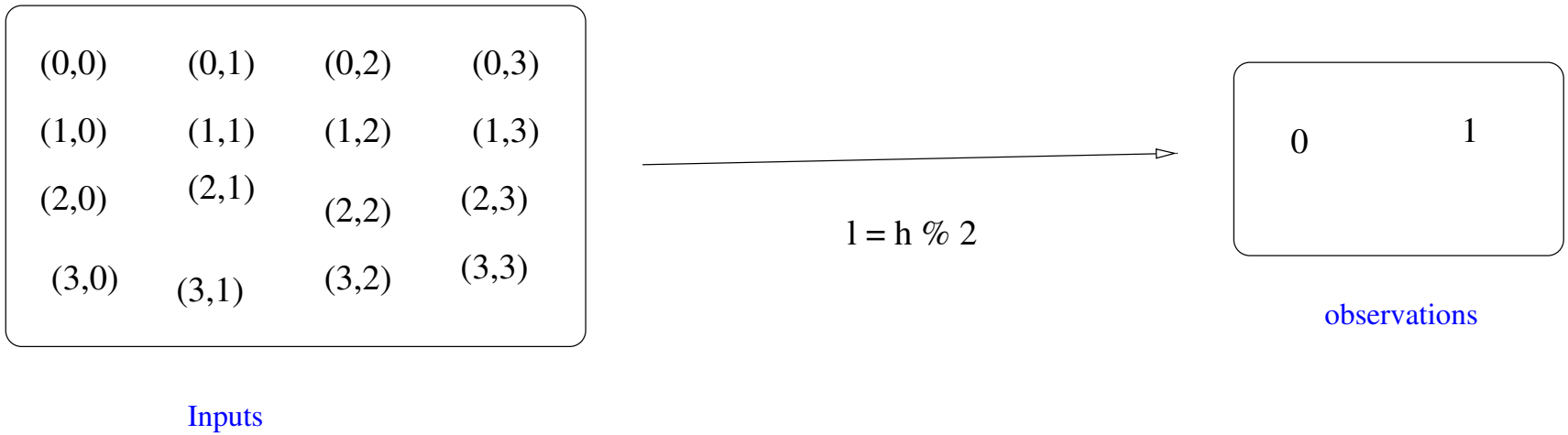
| | |
|---|---|
| colleague | 1/3 |
| girlfriend | 1/3 |
| brother | 1/3 |

$$\mathcal{H} = 3.\frac{1}{3}.log_2 3 = 1.585$$

# Example Program

## l = h % 2;

- Store is two 2-bit variables, *h* and *l*
- h is confidential, *l* is public
- stores representation is ordered pairs in $h \times l$



# A formal treatment

- A random variable (or discrete random element in this case) is a total function $X : D \to R$. $D$ and $R$ are finite sets, $D$ has a probability distribution.
- joint random variable: $(X, Y)$ defined as $\langle X, Y \rangle$
- Entropy of a random variable $X$:

$$\mathcal{H}(X) = \sum_{x \in R} p(x) \log \frac{1}{p(x)}$$

- Associate random variables with expressions; program variables; states at program points within a program.
- Of interest are observations of values of variables and states at $\iota$ (the *entry point*) and the special node $\omega$ (the *exit point*).

# Conditional Entropy

- $P((X \restriction (Y = y)) = x) = P(X = x | Y = y)$, where

$$P(X = x | Y = y) = \frac{p(x, y)}{p(y)}$$

- $$\mathcal{H}(X|Y) = \sum_y p(y)\mathcal{H}(X \restriction (Y = y))$$

- A key property of conditional information is that $\mathcal{H}(X|Y) \leq \mathcal{H}(X)$, with equality iff $X$ and $Y$ are independent.

# The Chain rule of Entropy

$$\mathcal{H}(A, B) = \mathcal{H}(A) + \mathcal{H}(B|A)$$

Entropy of the joint variation of a pair of variables is the entropy of one plus the conditional entropy of the other

$$p(x, y) = p(x).p(y|x)$$

# Mutual Information

Given two random variables $X$ and $Y$, the mutual information between $X$ and $Y$, written $\mathcal{I}(X;Y)$ is defined as follows:

$$\mathcal{I}(X;Y) = \sum_x \sum_y p(x,y) \log \frac{p(x,y)}{p(x)p(y)}$$

# Symmetry of Mutual Information

Use the chain rule to get three equivalent definitions:

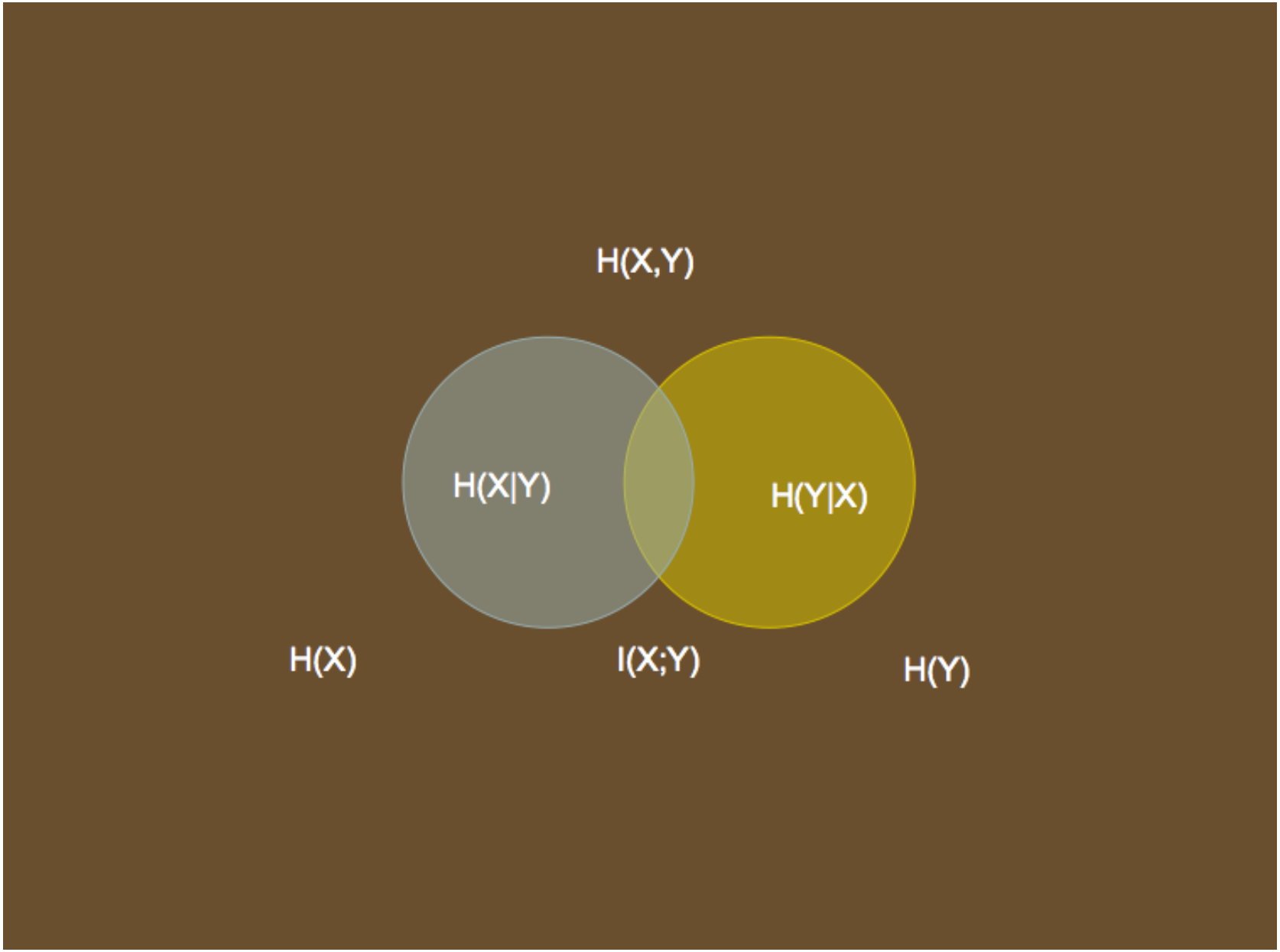
$$\mathcal{I}(X;Y) = \mathcal{H}(X) + \mathcal{H}(Y) - \mathcal{H}(X,Y)$$

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

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

This quantity is a direct measure of the amount of information carried by $X$ which can be learned by observing $Y$ (or vice versa).

# Yeung's diagram



# Channel Capacity

- Channel capacity in information theory is the maximum rate at which information can flow along a given communication channel.
- A fundamental result in information theory is that the channel capacity can be given as the least upper bound of the mutual information between inputs and outputs over all possible probability distributions on the input.

$$\bigsqcup_{\sigma_I} \mathcal{I}(I;O)$$

- In the case of deterministic programs

$$\mathcal{I}(I;O) = \mathcal{H}(O) - \mathcal{H}(O|I) = \mathcal{H}(O)$$

since $O$ is a function of $I$.

- So

$$\bigsqcup_{\sigma_I} \mathcal{I}(I;O) = \bigsqcup_{\sigma_I} \mathcal{H}(O)$$

- The maximum that this quantity can possibly be is

$$log_2|O|$$