Information!

Information ...

Sources of Information:

Apparent randomness:
Uncontrolled initial conditions
Actively generated: Deterministic chaos

Hidden regularity:
Ignorance of forces
Limited capacity to model structure

Information ...

Issues:

- What is information?
- How do we measure unpredictability
- How do we quantify structure?
- Information $\neq$ Energy

History of information:

- Boltzmann (19th Century):
  - Equilibrium in large-scale systems
- Hartley-Shannon-Wiener (Early 20th):
  - Communication & Cryptography
- Current threads (late 20th century):
  - Coding, Statistics, Dynamics, and Learning

Information ...

Information as uncertainty and surprise:

Observe something unexpected:

Gain information

Bateson: "A difference that makes a difference"

Information ...

Information as uncertainty and surprise ...

How to formalize?

Shannon's approach:

A measure of surprise.

Connection with Boltzmann's thermodynamic entropy

Self-information of an event $\propto -\log \Pr(\text{event})$.

Predictable: No surprise $-\log 1 = 0$

Completely unpredictable: Maximally surprised

$$-\log \frac{1}{\text{Number of Events}} = \log(\text{Number of Events})$$

Information ...

Shannon Entropy: $X \sim P$ $\quad x \in \mathcal{X} = \{1, 2, \ldots, k\}$

$$H(X) = -\sum_{x \in \mathcal{X}} p(x) \log_2 p(x)$$

Note: $0 \log 0 = 0$

$$H(X) = \langle - \log_2 p(x) \rangle$$

Units:

Log base 2: $H(X) = [\mathrm{bits}]$

Natural log: $H(X) = [\mathrm{nats}]$

Properties:

1. Positivity: $H(X) \geq 0$
2. Predictive: $H(X) = 0 \;\Leftrightarrow\; p(x) = 1$ for one and only one x
3. Random: $H(X) = \log_2 k \;\Leftrightarrow\; p(x) = U(x) = 1/k$

Information ...

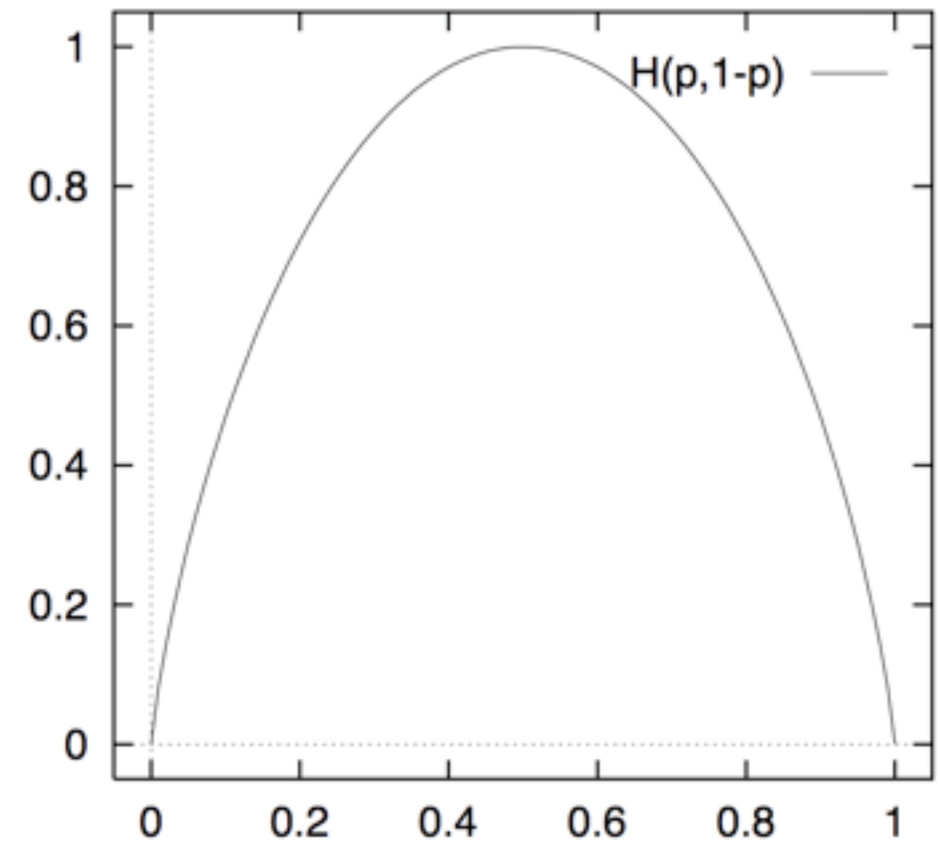
Examples: Binary random variable X

$$\mathcal{X} = \{0, 1\} \qquad \Pr(1) = p \ \& \ \Pr(0) = 1 - p$$

$H(X)$?

Binary entropy function:

$$H(p) = -p \log_2 p - (1 - p) \log_2 (1 - p)$$



Fair coin: $p = \frac{1}{2}$

$$H(p) = 1 \text{ bit}$$

Completely biased coin: $p = 0 \ (\text{or } 1)$

$$H(p) = 0 \text{ bits}$$

Recall: $0 \cdot \log 0 = 0$

Example: IID Process over four events

$$\mathcal{X} = \{a, b, c, d\} \qquad \Pr(X) = (\tfrac{1}{2}, \tfrac{1}{4}, \tfrac{1}{8}, \tfrac{1}{8})$$

Entropy: $H(X) = \frac{7}{4}$ bits

Number of questions to identify the event?

x = a? (must always ask at least one question)

x = b? (this is necessary only half the time)

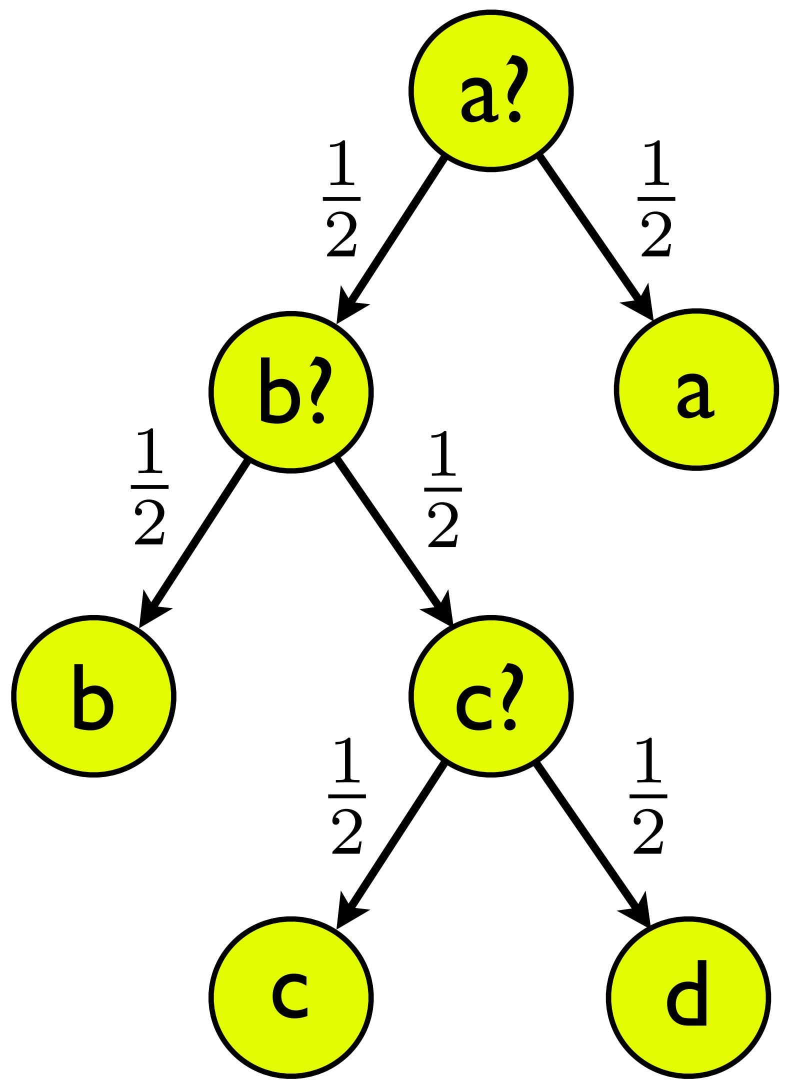
x = c? (only get this far a quarter of the time)

Average number: $1 \cdot 1 + 1 \cdot \frac{1}{2} + 1 \cdot \frac{1}{4} = 1.75$ questions

Interpretation? Optimal way to ask questions.

Example: IID Process over four events ...

Average number: $1 \cdot 1 + 1 \cdot \frac{1}{2} + 1 \cdot \frac{1}{4} = 1.75$ questions



$$\Pr(X) = \left(\tfrac{1}{2}, \tfrac{1}{4}, \tfrac{1}{8}, \tfrac{1}{8}\right)$$

Information ...

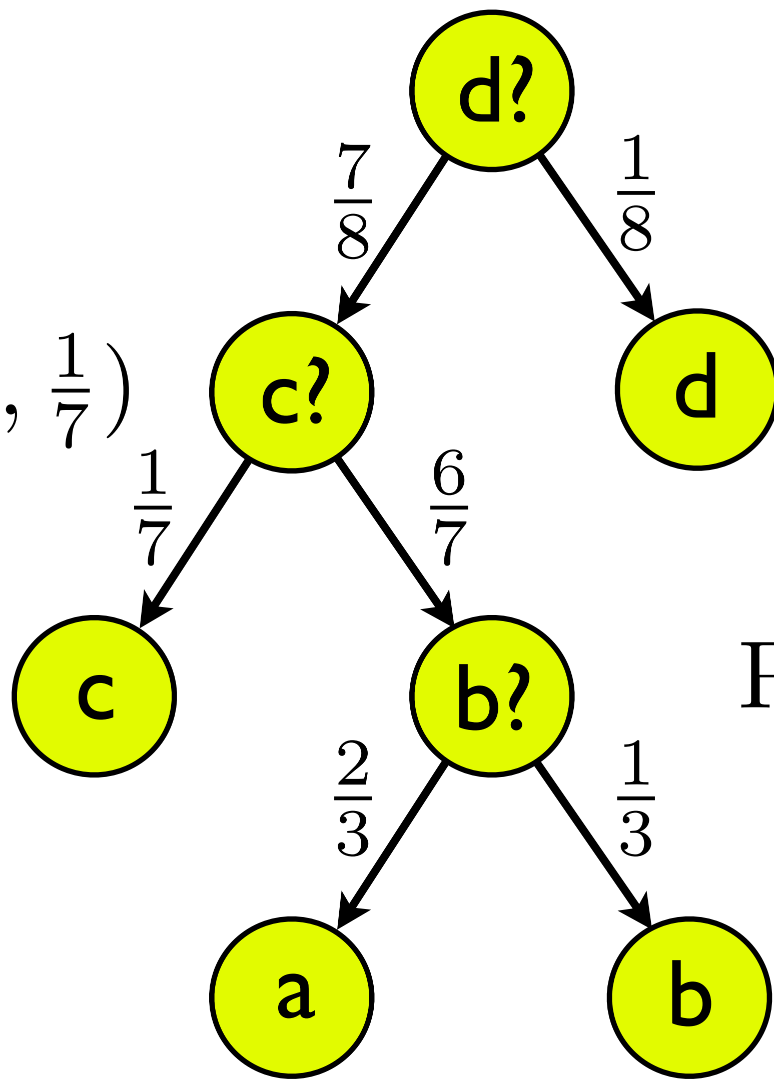
Example: IID Process over four events ...

Query in a different order:

Average number: $1 \cdot 1 + 1 \cdot \frac{7}{8} + 1 \cdot \frac{6}{7} \approx 2.7$ questions

$$\Pr(X) = (\tfrac{1}{2}, \tfrac{1}{4}, \tfrac{1}{8}, \tfrac{1}{8})$$

$$\Pr(a, b, c) = (\tfrac{4}{7}, \tfrac{2}{7}, \tfrac{1}{7})$$



$$\Pr(a, b) = (\tfrac{2}{3}, \tfrac{1}{3})$$

Example: IID Process over four events

Entropy: $H(X) = \frac{7}{4}$ bits

At each stage, ask questions that are most informative.

Choose partitions of event space that give “most random” measurements.

Theorem:
Entropy gives the smallest number of questions to identify an event, on average.

Information ...

Interpretations of Shannon Entropy:

Observer's *degree of surprise* in outcome of a random variable

Uncertainty *in* random variable

Information required to *describe* random variable

A measure of *flatness* of a distribution

Two random variables: $(X, Y) \sim p(x, y)$

Joint Entropy: Average uncertainty in X and Y occurring

$$H(X, Y) = -\sum_{x \in \mathcal{X}} \sum_{y \in \mathcal{Y}} p(x, y) \log_2 p(x, y)$$

Independent:

$$X \perp Y \Rightarrow H(X, Y) = H(X) + H(Y)$$

Conditional Entropy: Average uncertainty in X, knowing Y

$$H(X|Y) = -\sum_{x \in \mathcal{X}} \sum_{y \in \mathcal{Y}} p(x,y) \log_2 p(x|y)$$

$$H(X|Y) = H(X,Y) - H(Y)$$

Not symmetric: $H(X|Y) \neq H(Y|X)$

Example: Dining on campus

Food served at cafeteria is a random process:
Random variables:

Dinner one night: $D \in \{\text{Pizza}, \text{Meat w/Vegetable}\} = \{P, M\}$
Lunch the next day: $L \in \{\text{Casserole}, \text{Hot Dog}\} = \{C, H\}$

After many meals, estimate:

$$\Pr(P) = \frac{1}{2} \ \& \ \Pr(M) = \frac{1}{2}$$

$$\Pr(C) = \frac{3}{4} \ \& \ \Pr(H) = \frac{1}{4}$$

Entropies:

$$H(D) = 1 \text{ bit}$$

$$H(L) = H(\tfrac{3}{4}) \approx 0.81 \text{ bits}$$

Example: Dining on campus ...

Also, after many meals, estimate the joint probabilities:

$$\Pr(P, C) = \frac{1}{4} \ \& \ \Pr(P, H) = \frac{1}{4}$$

$$\Pr(M, C) = \frac{1}{2} \ \& \ \Pr(M, H) = 0$$

Joint Entropy: $H(D, L) = 1.5 \text{ bits}$

Dinner and Lunch are not independent:

$$H(D, L) = 1.5 \text{ bits} \neq H(D) + H(L) = 1.81 \text{ bits}$$

Suspect something's correlated: What?

Information ...

Example: Dining on campus ...

Conditional entropy of lunch given dinner:

$$\Pr(C|P) = \Pr(P, C)/\Pr(P) = \frac{1}{2}$$

$$\Pr(H|P) = \Pr(P, H)/\Pr(P) = \frac{1}{2}$$

$$\Pr(C|M) = \Pr(M, C)/\Pr(M) = 1$$

$$\Pr(H|M) = \Pr(M, H)/\Pr(M) = 0$$

$H(L|P) = 1 \text{ bit}$ Lunch unpredictable, if dinner was Pizza

$H(L|M) = 0 \text{ bits}$ Lunch predictable, if dinner was Meat w/Veg

Average uncertainty about lunch, given dinner:

$$H(L|D) = \frac{1}{2} \text{ bit}$$

Example: Dining on campus ...

Other way around?

Conditional entropy of dinner given lunch:

$$\Pr(P|C) = \Pr(P,C)/\Pr(C) = \tfrac{1}{3}$$

$$\Pr(M|C) = \Pr(M,C)/\Pr(C) = \tfrac{2}{3}$$

$$\Pr(P|H) = \Pr(P,H)/\Pr(H) = 1$$

$$\Pr(M|H) = \Pr(M,H)/\Pr(H) = 0$$

$$H(D|C) = H(\tfrac{2}{3}) \approx 0.92 \text{ bits}$$

$$H(D|H) = 0 \text{ bits}$$

Average uncertainty about dinner, given lunch:

$$H(D|L) = \tfrac{3}{4} H(\tfrac{2}{3}) \approx 0.69 \text{ bits}$$

Note: $H(D|L) \neq H(L|D)$. In fact, $H(D|L) > H(L|D)$.

Common Information Between Two Random Variables:

$$X \sim p(x) \ \& \ Y \sim p(y)$$

$$(X, Y) \sim p(x, y)$$

Mutual Information:

$$I(X;Y) = \mathcal{D}(P(x,y)||P(x)P(y))$$

$$I(X;Y) = \sum_{(x,y) \in \mathcal{X} \times \mathcal{Y}} p(x,y) \log_2 \frac{p(x,y)}{p(x)p(y)}$$

Information ...

Mutual Information ...

Properties:

(1) $I(X;Y) \geq 0$

(2) $I(X;Y) = I(Y;X)$

(3) $I(X;Y) = H(X) - H(X|Y)$

(4) $I(X;Y) = H(X) + H(Y) - H(X,Y)$

(5) $I(X;X) = H(X)$

(6) $X \perp Y \Rightarrow I(X;Y) = 0$

Interpretations:

Information one variable has about another
Information shared between two variables
Measure of dependence between two variables

Example: Dining on campus ...

Mutual information:

Reduction in uncertainty about lunch, given dinner:

$$I(D;L) = H(L) - H(L|D)$$
$$= H(\tfrac{3}{4}) - \tfrac{1}{2} \approx 0.31 \text{ bits}$$

Reduction in uncertainty about dinner, given lunch:

$$I(D;L) = H(D) - H(D|L)$$
$$= 1 - H(\tfrac{2}{3}) \approx 1 - 0.69 = 0.31 \text{ bits}$$

Shared information between what's served for dinner & lunch.

Information ...

Example: Dining on campus ...

Mutual information ...

What is the shared information?

Information ...

Example: Dining on campus ...

Mutual information ...

What is the shared information?

Further inquiry:

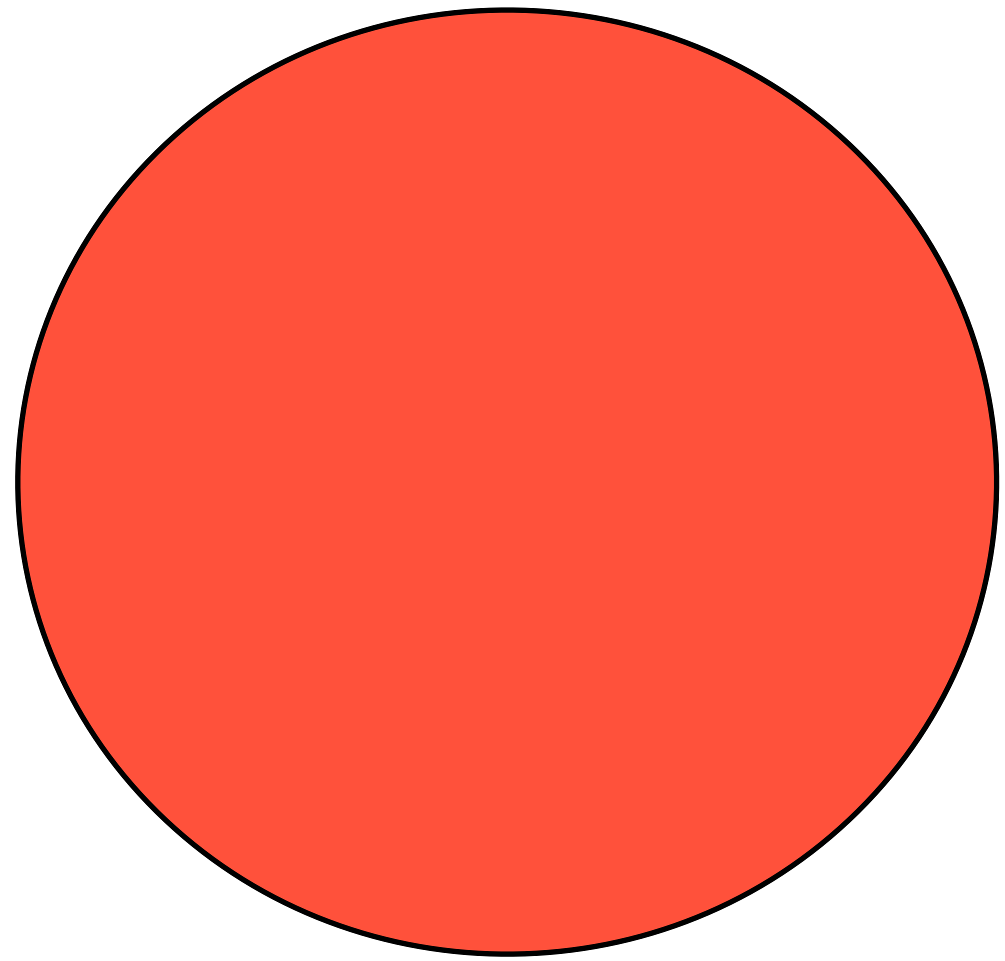
Vegetable served with dinner (Meat + Veg)

appears in lunch's casserole!

Event Space Relationships of Information Quantifiers:

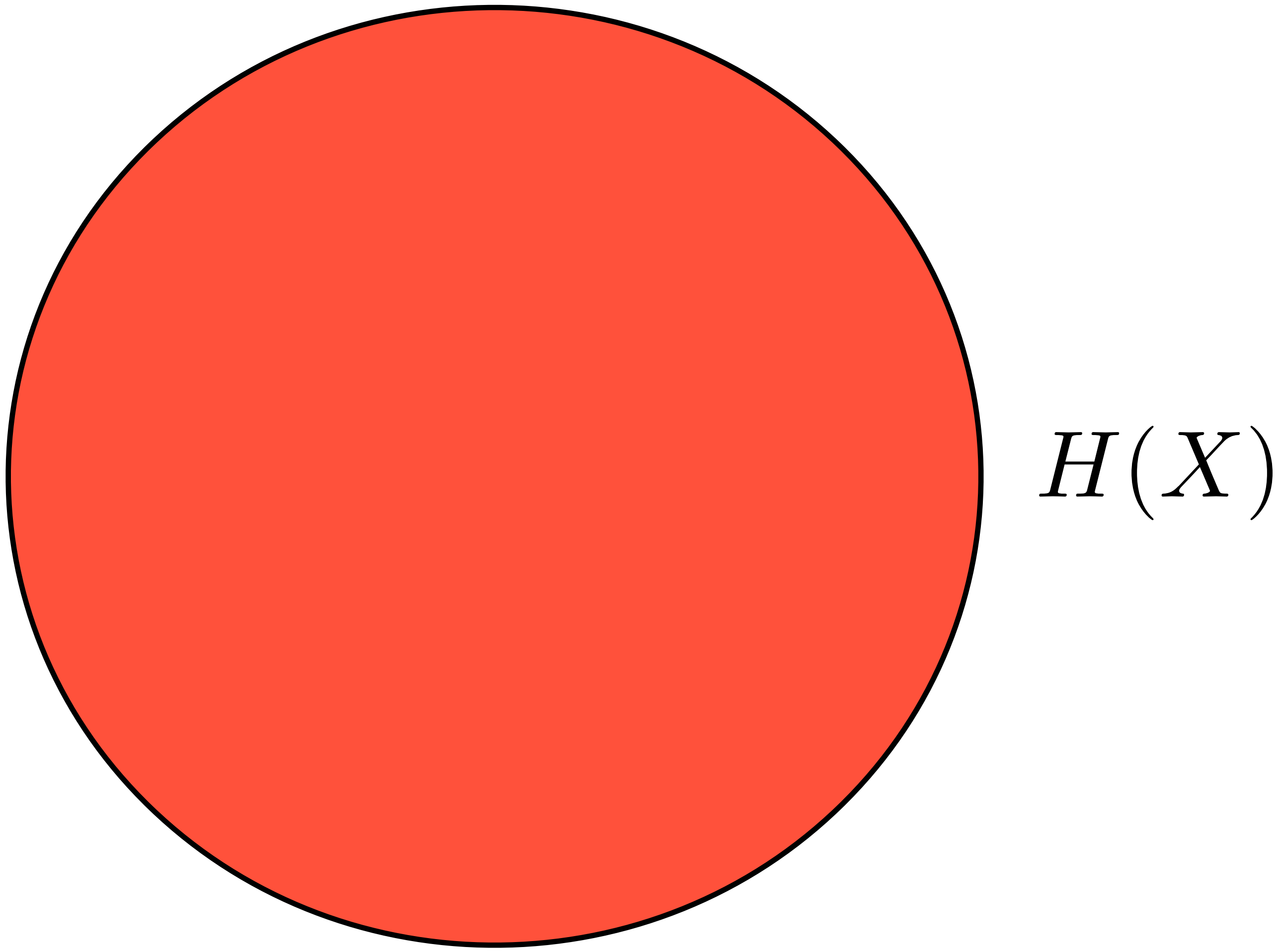
Information ...

Event Space Relationships of Information Quantifiers:



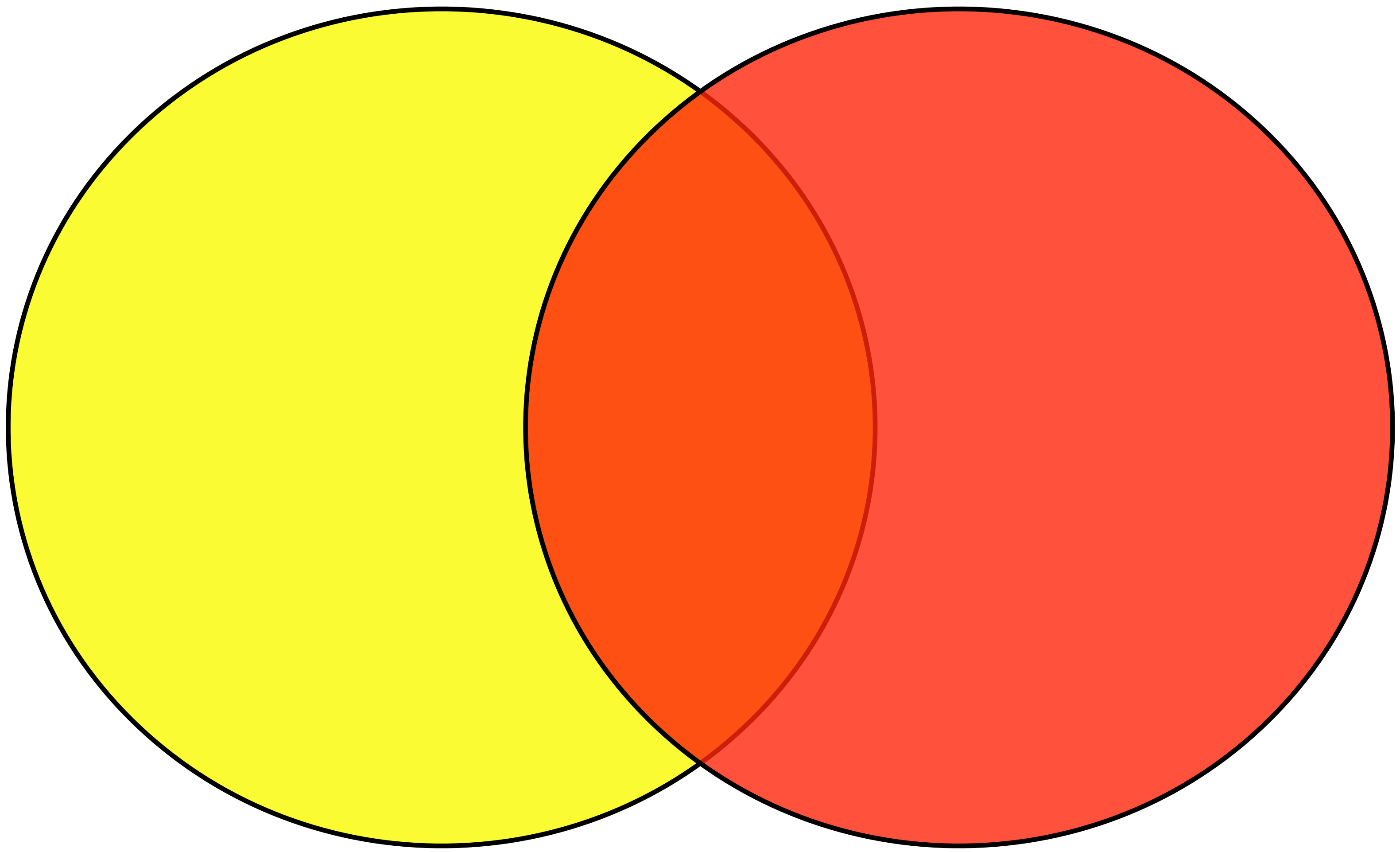
Information ...

Event Space Relationships of Information Quantifiers:



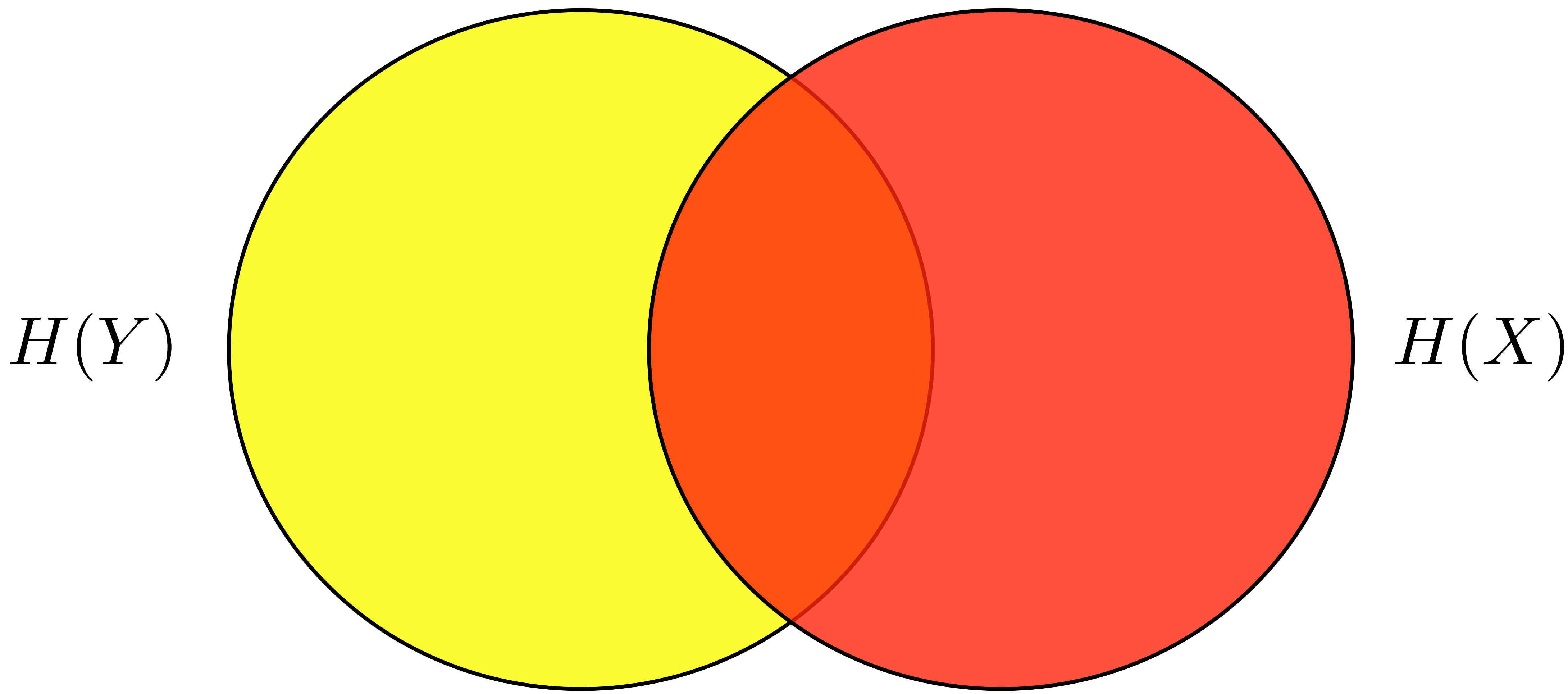
Information ...

Event Space Relationships of Information Quantifiers:



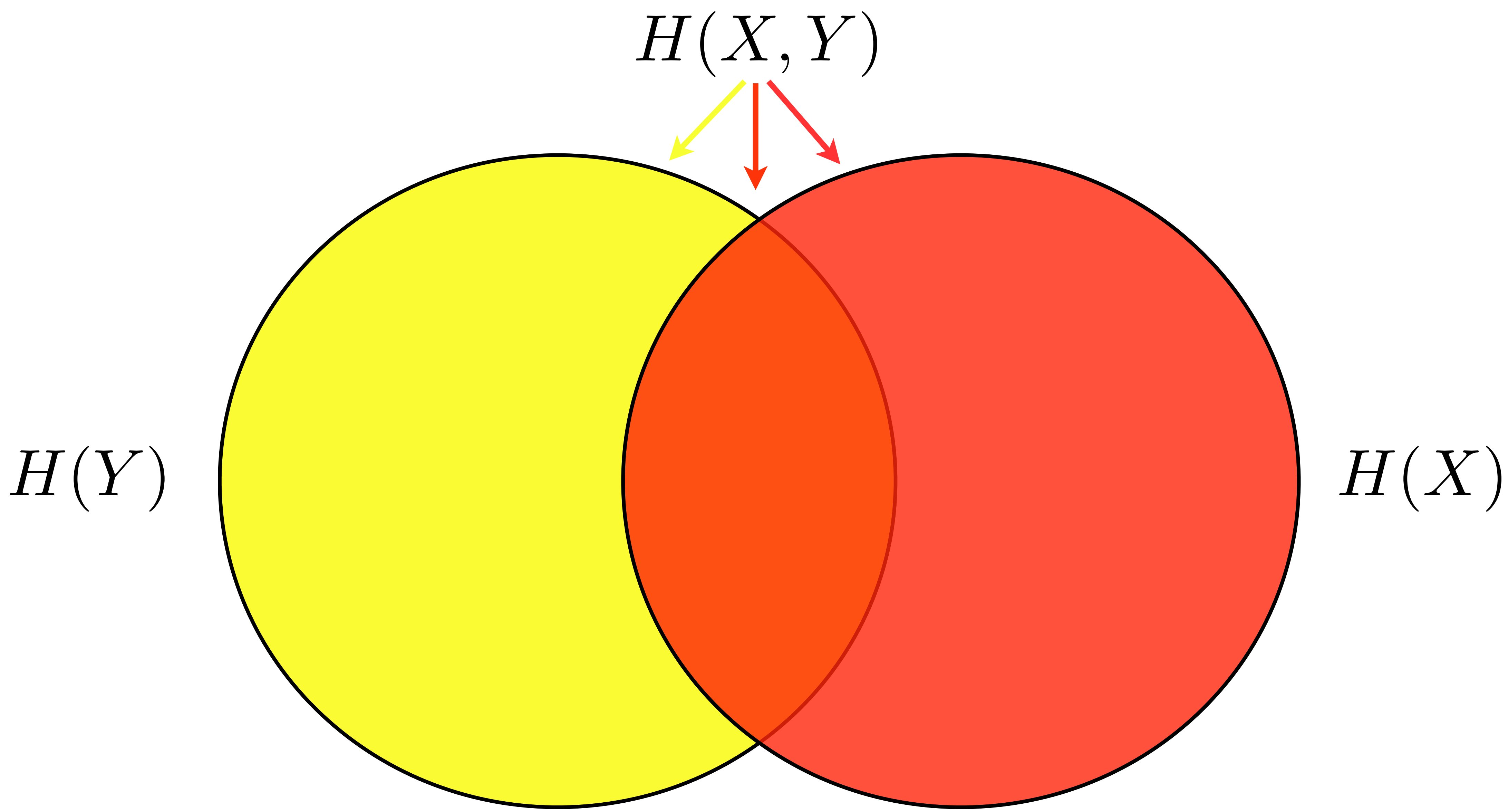
$H(X)$

Event Space Relationships of Information Quantifiers:



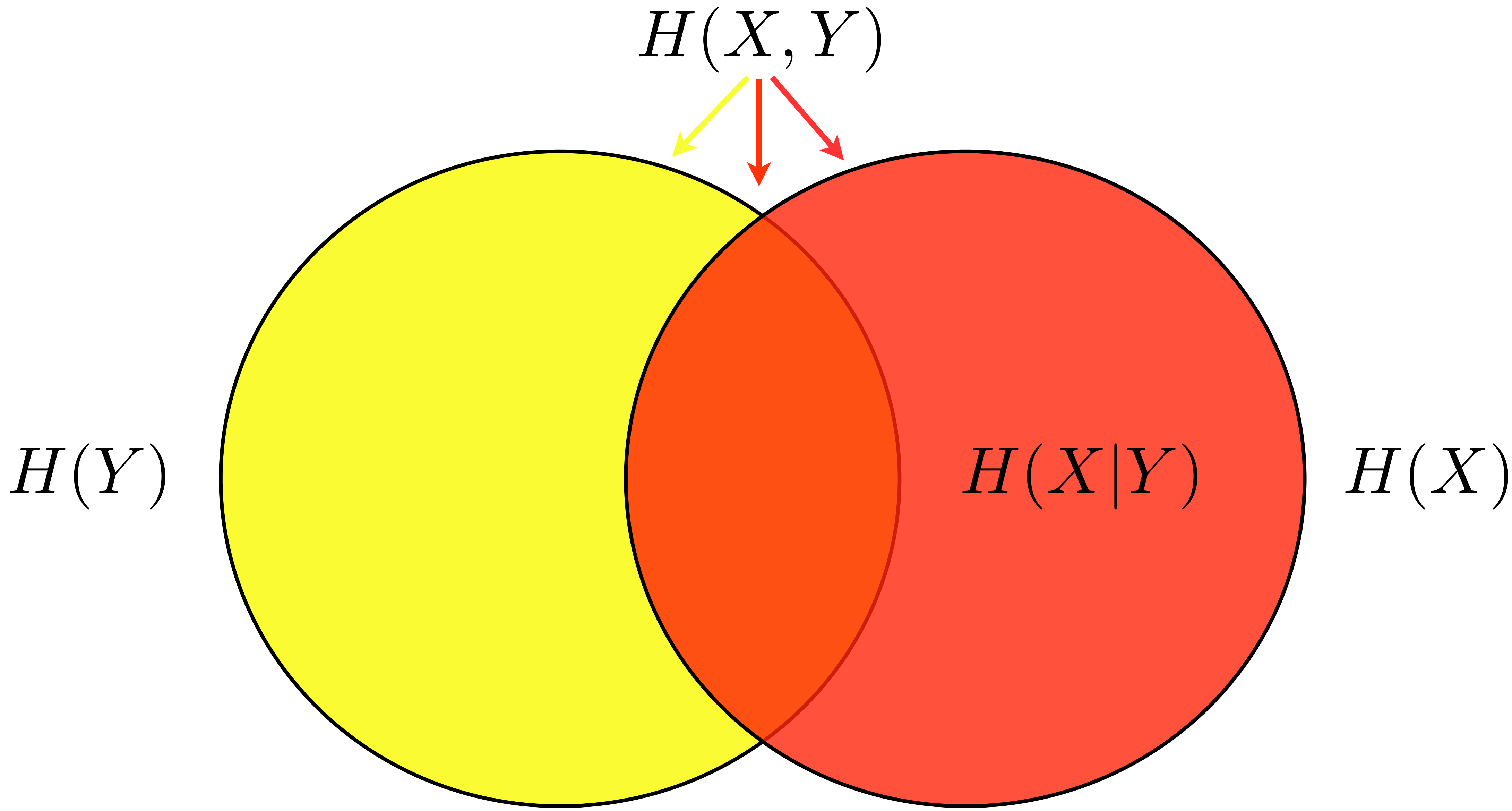
Information ...

Event Space Relationships of Information Quantifiers:



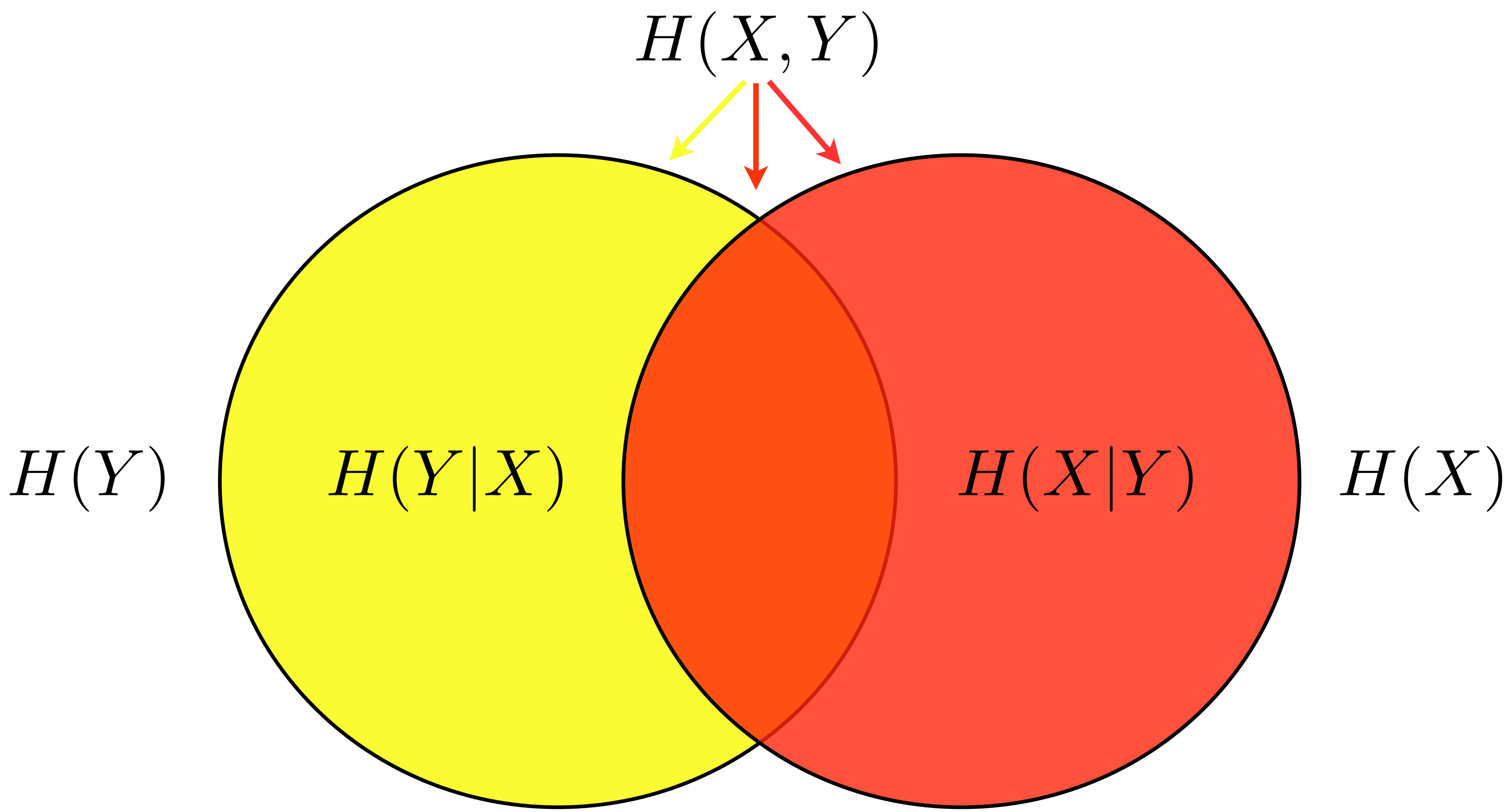
Information ...

Event Space Relationships of Information Quantifiers:



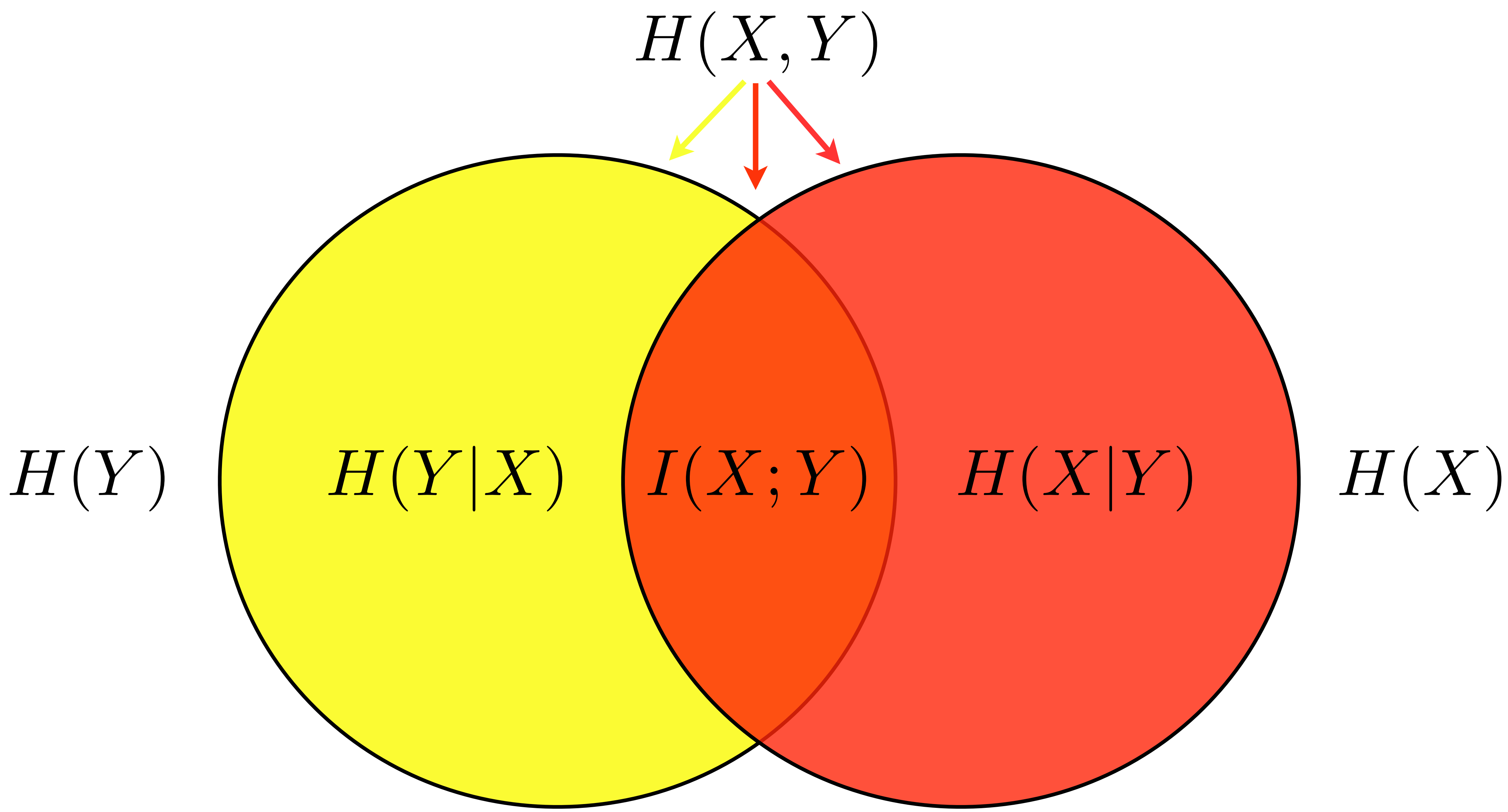
Information ...

Event Space Relationships of Information Quantifiers:



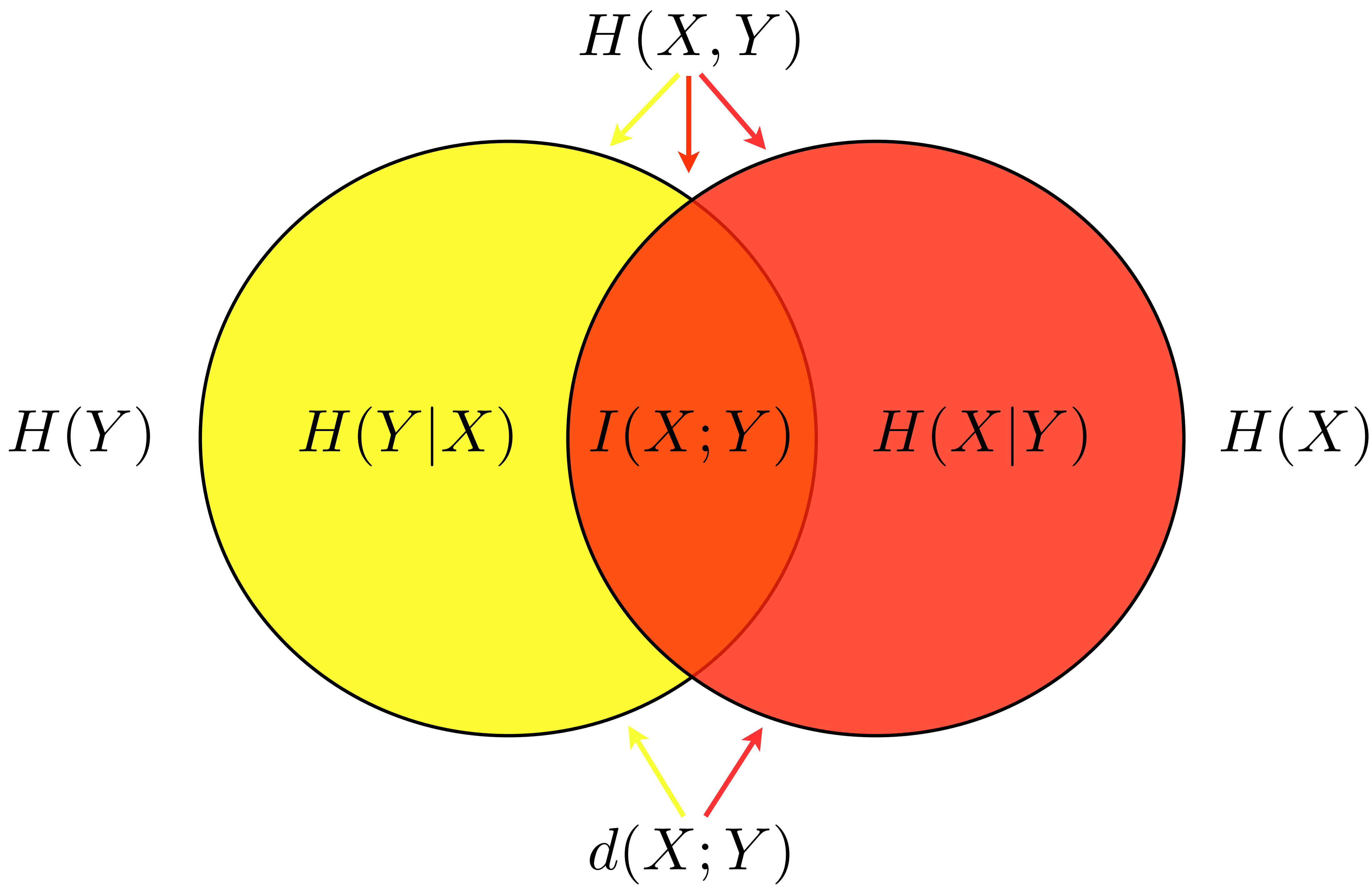
Information ...

Event Space Relationships of Information Quantifiers:



Information ...

Event Space Relationships of Information Quantifiers:



Information ...

Bounds:

Uniform Distribution:

$$X \sim U(x) = 1/k$$

$$H(X) = \log |\mathcal{X}|$$

Generally: $H(X) \leq \log |\mathcal{X}|$

In fact: $H(X) = \log |\mathcal{X}| - \mathcal{D}(P(x)||U(x))$

Bounds ...

Conditioning Reduces Entropy:

$$H(X|Y) \leq H(X)$$

Independence:

$$H(X_1, \ldots, X_n) \leq \sum_{i=1}^{n} H(X_i)$$

Three random variables: $(X, Y, Z) \sim p(x, y, z)$

Markov Chain: $X \rightarrow Y \rightarrow Z$

$$p(x, z|y) = p(x|y)p(z|y) \quad \text{or} \quad I(X; Z|Y) = 0$$

Y shields X and Z from each other: $X \perp_Y Z$

Properties:

(1) $X \rightarrow Y \rightarrow Z \Rightarrow Z \rightarrow Y \rightarrow X$

(2) $Z = f(Y) \Rightarrow X \rightarrow Y \rightarrow Z$

Data Processing Inequality:

$$X \to Y \to Z \Rightarrow I(X;Y) \geq I(X;Z)$$

Corollary:

$$Z = g(Y) \Rightarrow I(X;Y) \geq I(X;g(Y))$$

Manipulation *cannot* increase information about X.

Information ...

Dining example:

Hidden variable was "leftovers".

Knowing this, lunch and dinner are independent:

$$\text{Dinner} \perp_{\text{leftovers}} \text{Lunch}$$

Markov chain:

$$\text{Dinner} \rightarrow \text{leftovers} \rightarrow \text{Lunch}$$

Now:

How to compress a process:

Can't do better than H(X)

(Shannon's First Theorem)

How to communicate a process's data:

Can transmit error-free at rates up to channel capacity

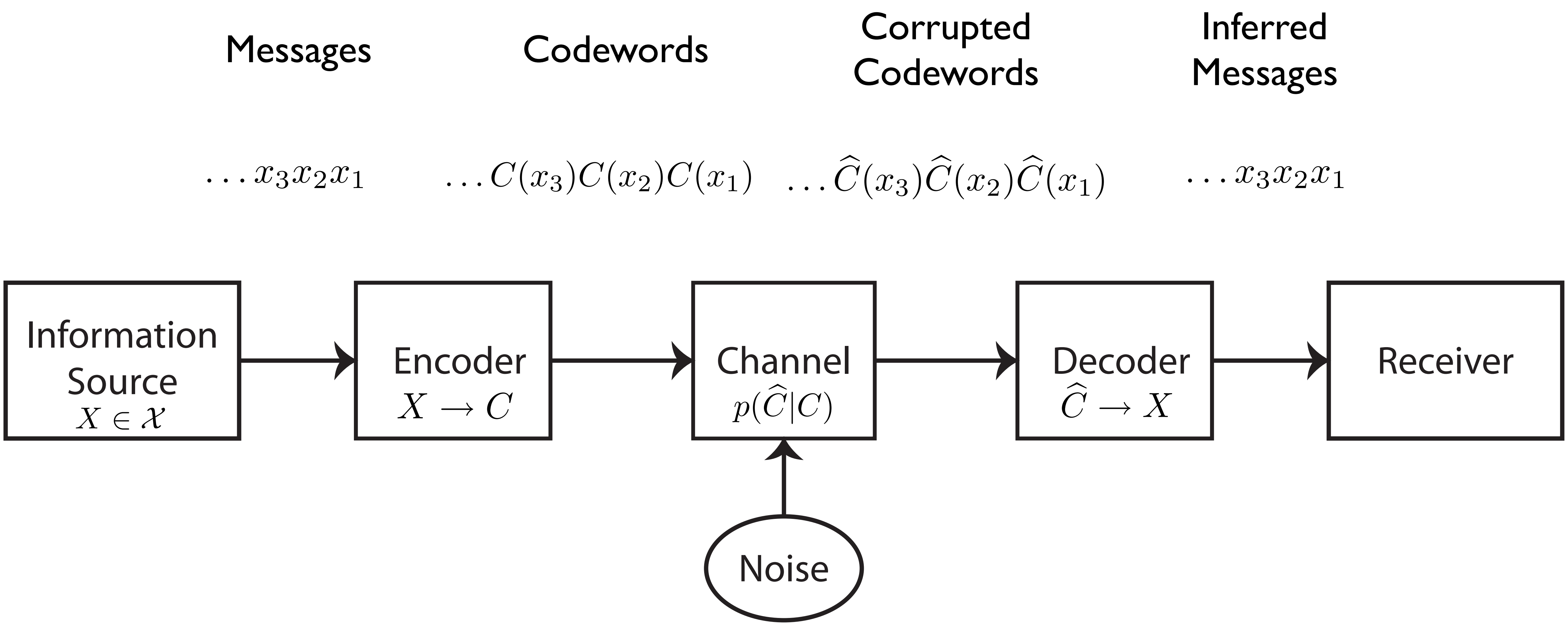
(Shannon's Second Theorem)

Both results give operational meaning to entropy.

Previously: entropy motivated as a measure of surprise.

Information in Processes ...

Communication channel:



Information in Processes ...

Example: $\mathcal{X} = \{a, b, c, d\}$

$X \sim p(x)$

Distribution:
$$\begin{aligned} p(a) &= \frac{1}{2} \\ p(b) &= \frac{1}{4} \\ p(c) &= \frac{1}{8} \\ p(d) &= \frac{1}{8} \end{aligned}$$

$$H(X) = 1.75 \text{ bits}$$

Codebook:
$$\begin{aligned} C(a) &= 0 \\ C(b) &= 10 \\ C(c) &= 110 \\ C(d) &= 111 \end{aligned}$$

Average code length:

$$R(C) = 1.75 \text{ bits}$$

Kinds of codes ...

Example (continued):

Codebook:
$$\begin{aligned} C(a) &= 0 \\ C(b) &= 10 \\ C(c) &= 110 \\ C(d) &= 111 \end{aligned}$$

Encoding:

$$acdbac \to 0110111100110$$

Decoding:

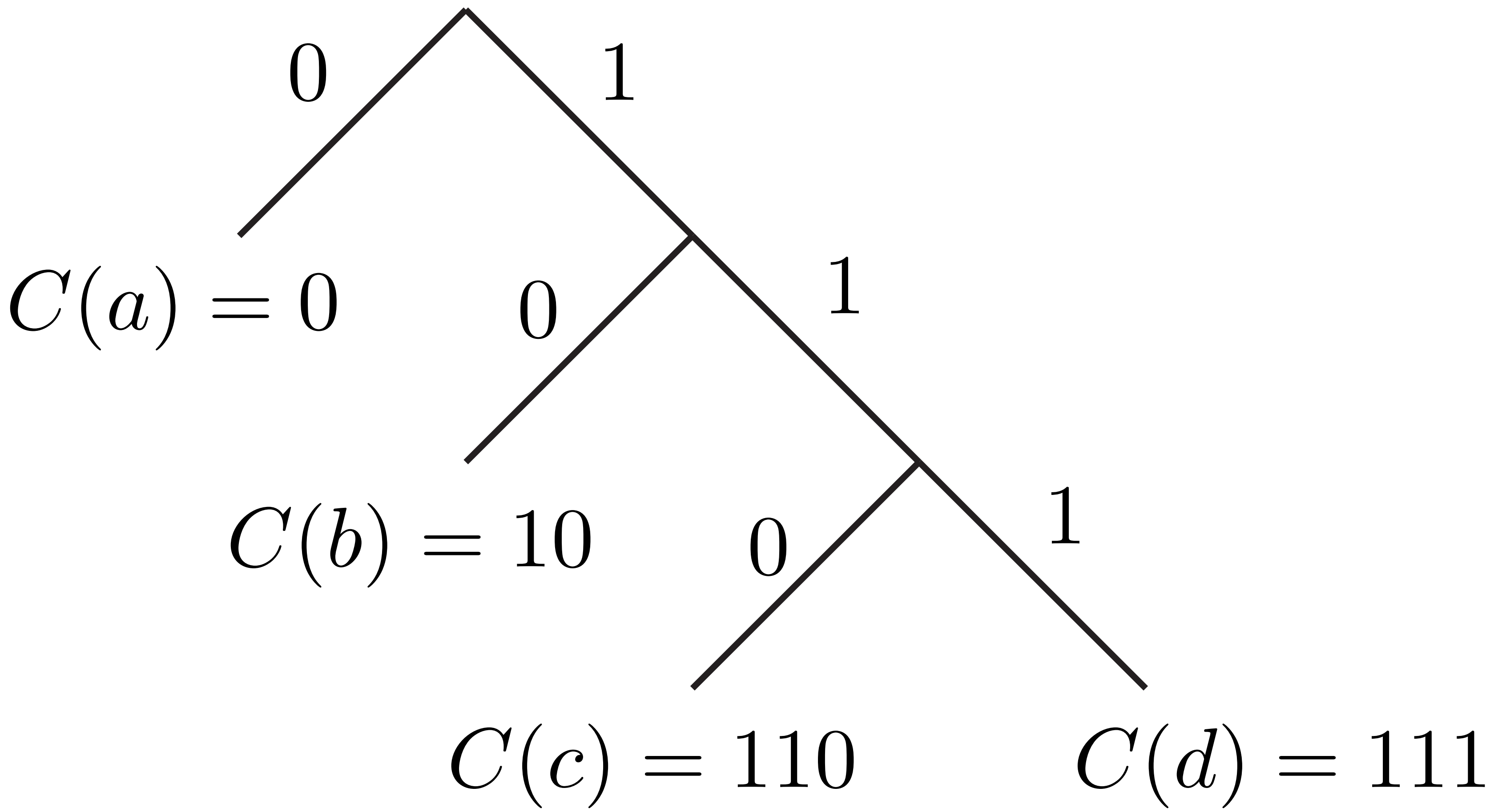
$$0110111100110 \to \overbrace{0}^{a}\;\overbrace{110}^{c}\;\overbrace{111}^{d}\;\overbrace{10}^{b}\;\overbrace{0}^{a}\;\overbrace{110}^{c}$$

$$0110111100110 \to acdbac$$

A prefix code.

Information in Processes ...

Example (continued):



$$
\begin{aligned}
C: \sum &= 2^{-l(a)} + 2^{-l(b)} + 2^{-l(c)} + 2^{-l(d)} \\
&= 2^{-1} + 2^{-2} + 2^{-3} + 2^{-3} \\
&= 1
\end{aligned}
$$

Optimal codes:

Given an information source, find codebook such that

1. Minimize expected code length:

$$R = \langle l(x) \rangle = \sum_{x \in \mathcal{X}} p(x) l(x)$$

2. Subject to constraint of decodability:

$$\sum_{x \in \mathcal{X}} 2^{-l(x)} \leq 1$$

Answer: optimal code words has lengths

$$l(x) = -\log_2 p(x)$$

And, average codebook length:

$$\langle l(x) \rangle = H(X)$$

Data Compression Theorem (Shannon's First Theorem):

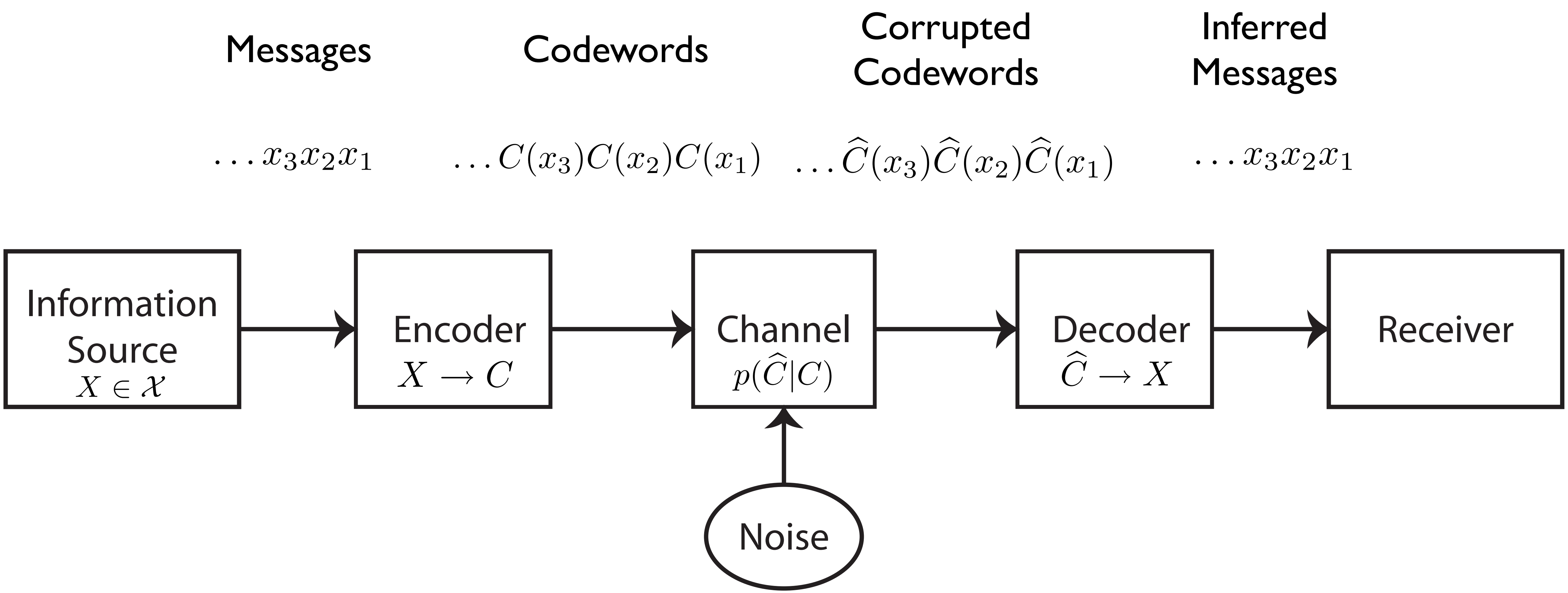
$$R(C) \geq H(X)$$

Cannot compress source below its entropy rate.

Operational meaning of entropy: fundamental limit.

Information in Processes ...

Communication channel:



Reliable transmission through noisy channel: Possible?

How to code in presence of distorted codewords?

Coding for Communication Channels:

Kinds of channel:

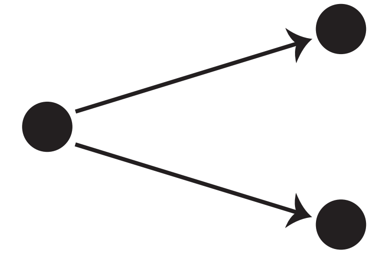
- Phone line, ftp transfer, monologue, ...
- Dynamical system at time t and t+1
- Spin system at one site and another
- Measurement channel

Information in Processes ...

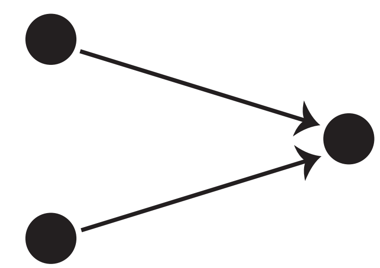
Coding for Communication Channels ...

Channel coding problem is to overcome errors:

Equivocation:
Same input sequence leads to different outputs



Ambiguity:
Two different inputs lead to same output



Strategy:
Find channel inputs that are *least ambiguous* given distortion properties.
Codebook: Map information source onto those inputs.

Coding for Communication Channels ...

Discrete channel:

Input: $X \sim p(x)$

Output: $Y \sim p(y)$

Channel: $p(y|x)$

Memoryless channel:

$$p(y_t|x_t x_{t-1} \cdots) = p(y_t|x_t)$$

Channel Capacity:

$$\mathcal{C} = \max_{p(x)} I(X;Y)$$

Highest rate one can transmit over channel.

Coding for Communication Channels ...

Extremes of no communication:

No info to send: $H(X) = 0$

$$I(X;Y) = H(X) - H(X|Y) = 0 - 0 = 0$$

Complete distortion:

Output independent of input: $X \perp Y$

$$I(X;Y) = 0$$

Duality:

Compression removes redundancy to give smallest description.

Encoding adds redundancy to compensate channel errors.

Channel Coding Theorem (Shannon's Second Theorem):

(1) Capacity is the maximum reliable transmission rate.
(2) Error-free codes exist if $R < \mathcal{C}$.

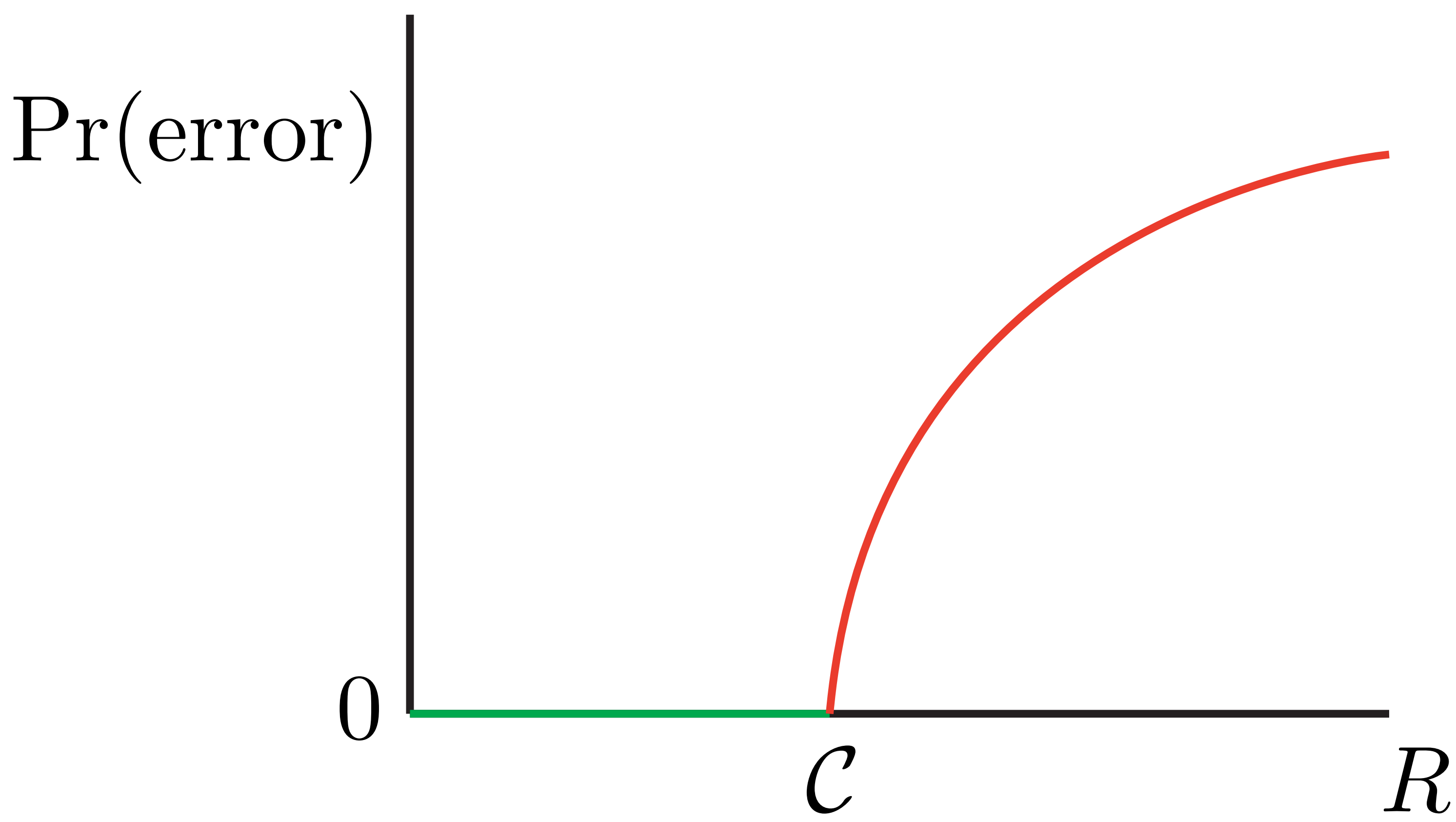
Idea:
Model as noisy channel with non-overlapping outputs.

Strategy:
Code long block lengths: $|\mathcal{X}^L| \approx 2^{LH(X)}$
Choose codewords (channel inputs) that
produce non-overlapping outputs.

Information in Processes ...

Channel Coding Theorem ...

What happens when transmitting above capacity, $R > \mathcal{C}$?



(Typical of measurement systems?)