Introduction to Information Theory

Information theory

- What is it?
  - formal way of counting “information” bits
- Why do we need it?
  - often used in crypto
  - "quality" of keys
  - best way to talk about physical sources of randomness

Notation for probabilities

- Capital letter (uppercase): name of the experiment
- Lowercase: outcome

Example 1: Roll a fair six-sided die.
For all $x \in \{1,2,3,4,5,6\}$ the probability of outcome x is $\Pr[X=x] = 1/6$.

Example 2: Biased coin. $x \in \{0,1\}$.
$\Pr[X=0] = 0.2$, $\Pr[X=1] = 0.8$

Often used notations:
$X \in \{\text{some set}\}$
$p_x = \Pr[X=x]$

Probability distributions

Let $X \in \{x_1, x_2, \ldots, x_n\}$.

Notation $X \sim (a_1, \ldots, a_n)$ means " X has distribution $(a_1, \ldots, a_n)$ ".

$\Pr[X=x_i] = a_i$.

The numbers must satisfy $a_i \geq 0$ and $\Sigma_i\, a_i = 1$.

For $X \in$ {discrete set} it is called a **Probability Mass Function**.

Fot $X \in$ continuum: Probability *Density* Function.

Cumulative Distribution Function (cdf):

$\Pr[X \leq$ something$]$

In above example (with increasing x_i),

$\Pr[X \leq x_1] = a_1$; $\Pr[X \leq x_2] = a_1+a_2$; $\Pr[X \leq x_3] = a_1+a_2+a_3$; ...

Probability distributions (continued)

Distributions can be given a name.

Example

$X \in \{0,1,2,3\}$ $X \sim \mathbb{P}$ $\mathbb{P} = (1/2, 1/6, 1/6, 1/6)$

Notation $\mathbb{P}(x)$ stands for $\Pr[X=x]$.

Joint distributions

Two experiments, X and Y, which may influence each other.
Joint distribution $(X,Y) \sim \mathbb{P}$.
Notation XY for the combined experiment.
$p_{xy} = \Pr[X=x \text{ and } Y=y]$

Marginals
$p_y = \Pr[Y=y] = \Sigma_x\, p_{xy}$. $p_x = \Pr[X=x] = \Sigma_y\, p_{xy}$.

Chain rule
$p_{xy} = p_x\, p_{y|x} = p_y\, p_{x|y}$.
$p_{y|x} = \Pr[Y=y \text{ given that } X=x]$

Measuring ignorance

Experiment X with unpredictable outcome

- known possible outcomes x_i
- known probabilities p_i.

We want to measure the "unpredictability" of the experiment

- should increase with #outcomes
- should be maximal for uniform p_i

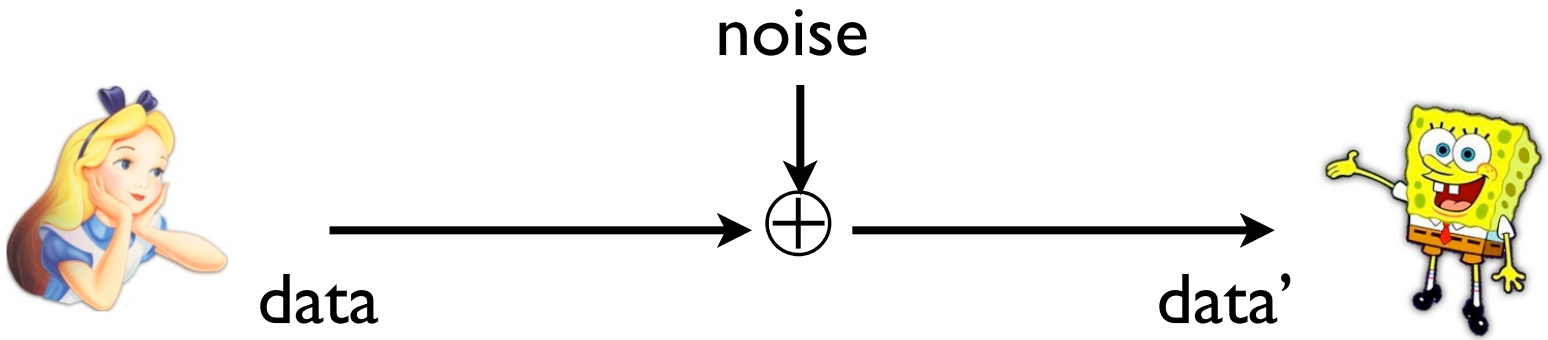
Measuring ignorance

plaintext → Encrypt → ciphertext → Decrypt → plaintext

ciphertext → Eve

How much does Eve know about the plaintext given that she has seen the ciphertext?

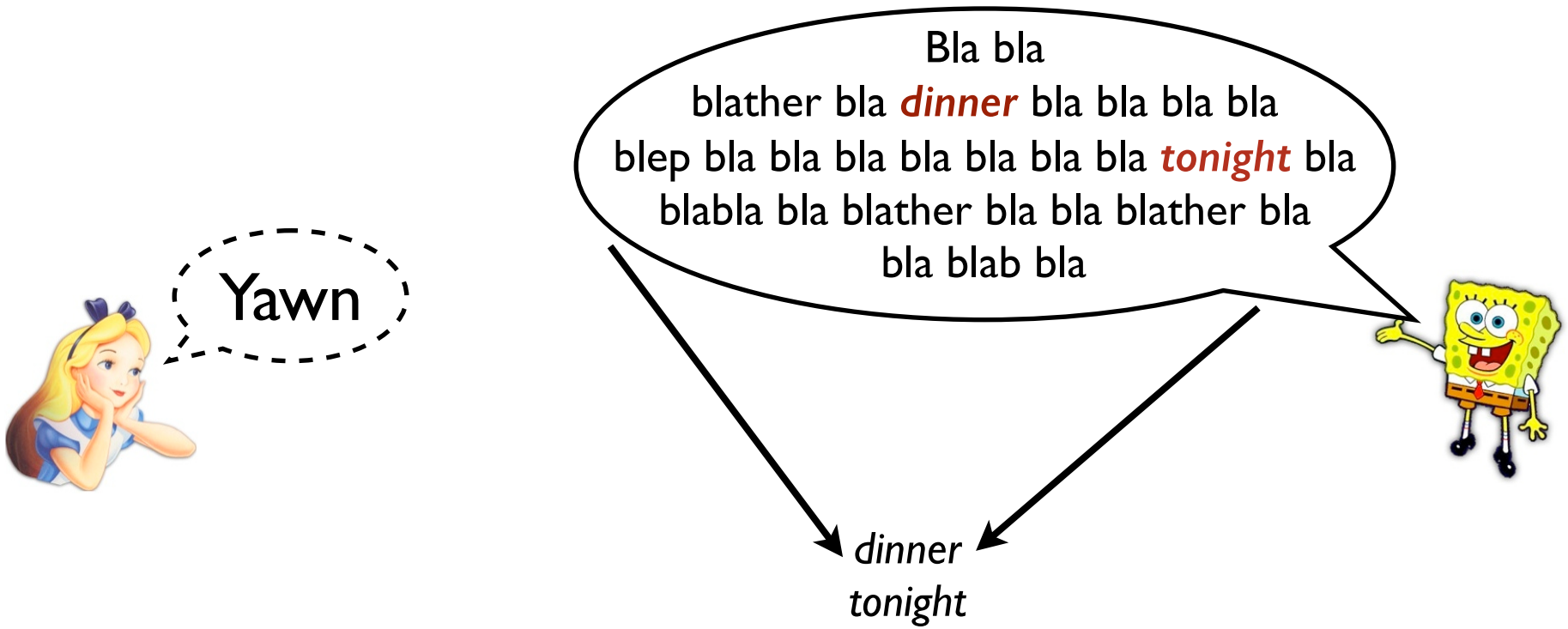
Measuring common knowledge



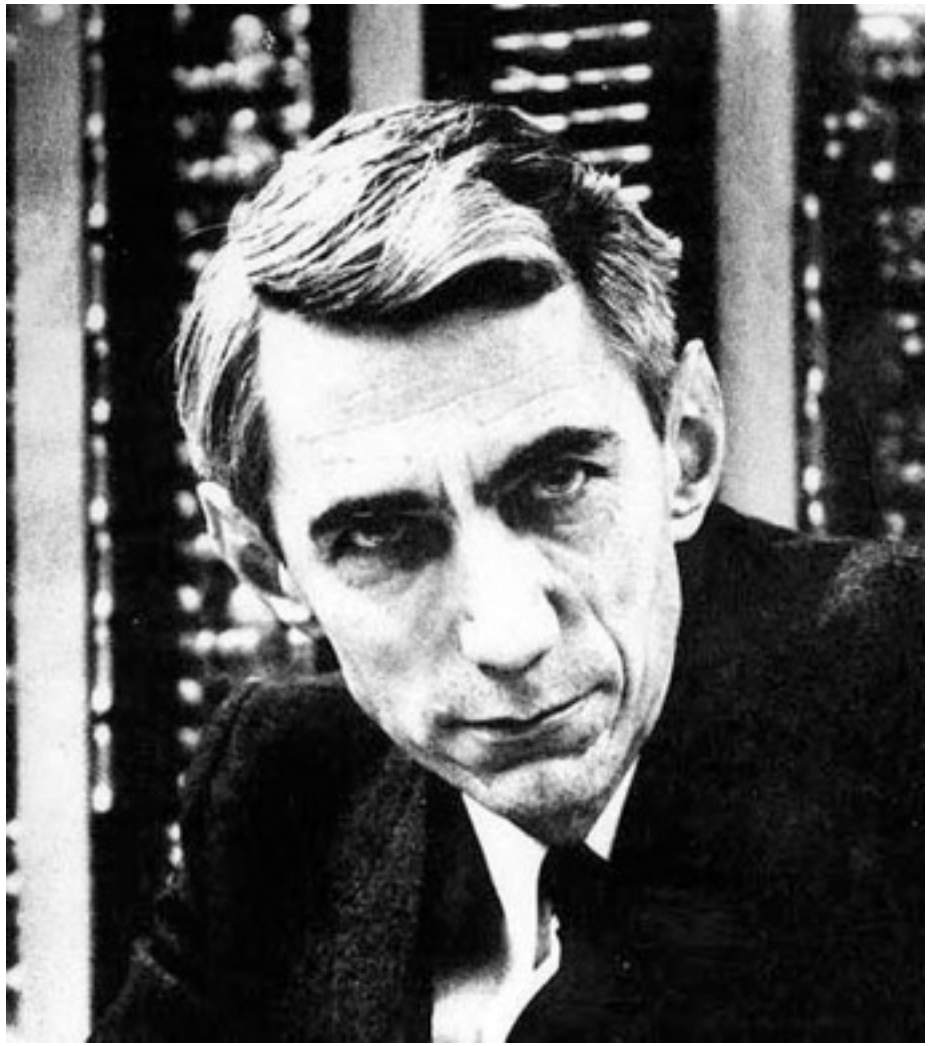
How much of the info that Alice sends does actually reach Bob?

("channel capacity")

Data compression



How far can a body of data be compressed?



Claude Shannon (1916 - 2001)

Groundbreaking work in 1948

Shannon entropy

Notation: H(X) or H($\{p_1, p_2, \ldots\}$)

Formal requirements for counting information bits

- Sub-additivity: $H(X,Y) \leq H(X) + H(Y)$
  - equality only when X and Y independent
- Expansibility: extra outcome x_j with $p_j=0$ has no effect
- Normalization:
  - entropy of (1/2, 1/2) is 1 bit; of (1,0,0, ...) is zero.

Unique expression satisfying all requirements:

$$H(X) = -\Sigma_x\, p_x \log_2(p_x)$$

Examples of Shannon entropy

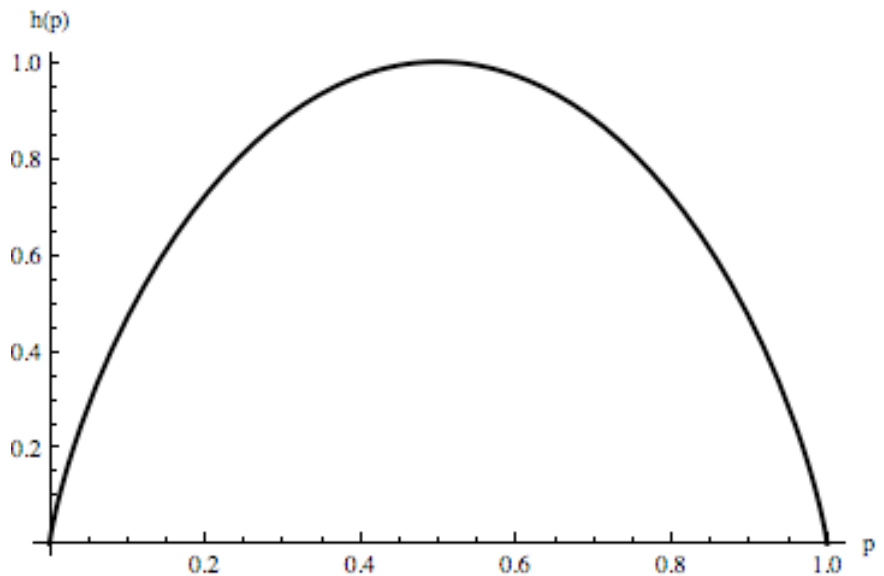
$$H(X) = -\Sigma_x p_x \log(p_x)$$

- Uniform distribution X~(1/n, ..., 1/n).

  H(X) = log n.

- Binary entropy function. X~(p, 1-p).

  h(p) = h(1-p) = −p log p − (1-p) log(1-p)



Mini quiz

$$H(X) = -\Sigma_x p_x \log(p_x)$$

X~(1/2, 1/4, 1/4). Compute H(X).

- Which yes/no questions would you ask to quickly determine the outcome of the experiment?
- How many questions do you need on average?

Interpretation of Shannon entropy

1) Average # of binary questions needed to decide on outcome of experiment (lower bound)
2) Lower bound on compression size
3) Theory of 'typicality':
   - Given pmf $\mathbb{S} = (s_1,\ldots,s_q)$; N independent experiments;
   - t = some sequence with exactly $n_i = N\, s_i$ occurrences of i. ("typical sequence")

$$\Pr[X = t] = \prod_{i=1}^{q} s_i^{n_i} = \prod_{i=1}^{q} s_i^{N s_i} = 2^{\sum_i N s_i \log s_i} = 2^{-N\mathrm{H}(\mathbb{S})}$$

Relative entropy

- Measure of distance between distributions
  - asymmetric
  - non-negative
  - zero only when distr. are identical

$$D(\mathbb{P}||\mathbb{Q}) = \sum_{x \in \mathcal{X}} \mathbb{P}(x) \log \frac{\mathbb{P}(x)}{\mathbb{Q}(x)}$$

- Interpretation
  - when you think distr. is $\mathbb{Q}$, but actually it is $\mathbb{P}$,

    #questions ≥ H($\mathbb{P}$) + D($\mathbb{P}$ || $\mathbb{Q}$).

Time for an exercise

Prove that relative entropy is non-negative,

$$D(P \,||\, Q) \geq 0.$$

Prove sub-additive property,

$$H(X,Y) \leqslant H(X) + H(Y)$$

Conditional entropy

- Joint distribution $(X,Y) \sim P$.
- Conditional probability: $p_{y|x} = p_{xy} / p_x$.
- Entropy of Y for given X=x:

$$H(Y|X=x) = -\Sigma_y\, p_{y|x} \log(p_{y|x})$$

- Conditional entropy of Y given X:

$$\boxed{H(Y|X) = \mathbb{E}_x[H(Y|X=x)]} = -\Sigma_{xy}\, p_{xy} \log(p_{y|x})$$

Conditional entropy

Watch out:

Y|X is *not* a RV, but a set of RVs.

Y|X=x *is* a RV

<u>Chain rule</u>:

$p_{xy} = p_x\, p_{y|x} = p_y\, p_{x|y}$.

$$H(X,Y) = H(X) + H(Y|X)$$
$$= H(Y) + H(X|Y).$$

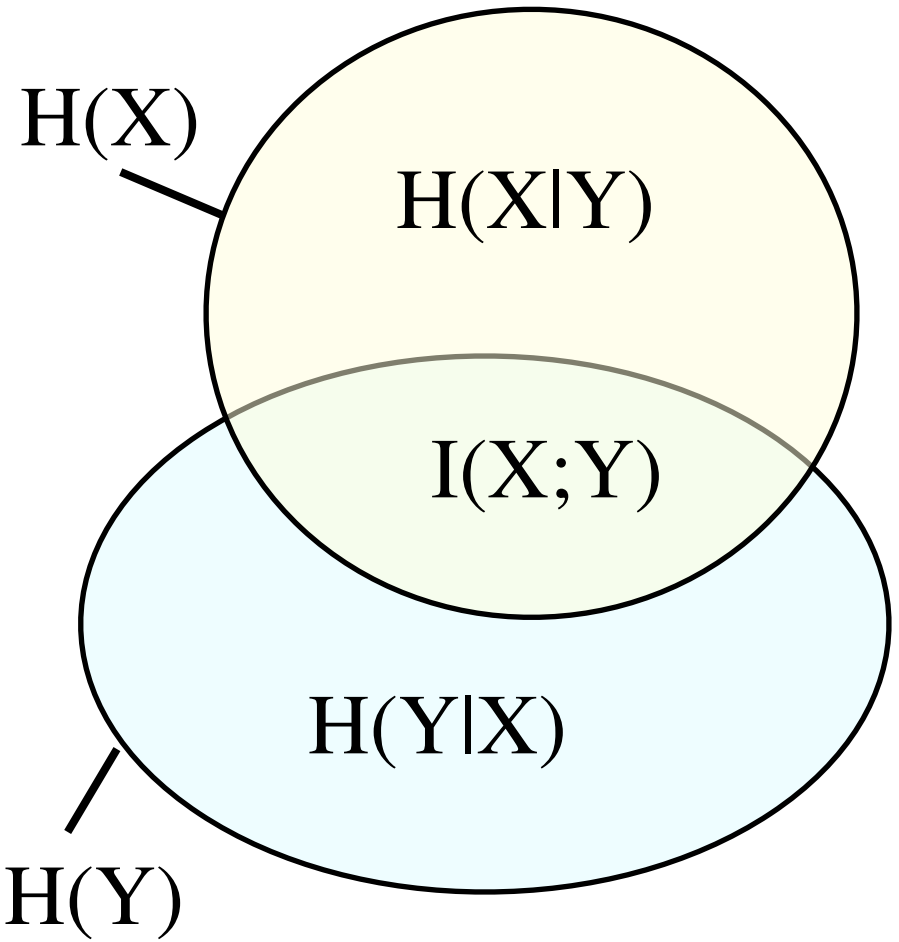
Some properties

Conditioning reduces entropy: $H(X|Y) \leq H(X)$

Proof:

- chain rule $H(X|Y) = H(X,Y) - H(Y)$
- sub-additive property $H(X,Y) \leq H(X) + H(Y)$

Mutual information

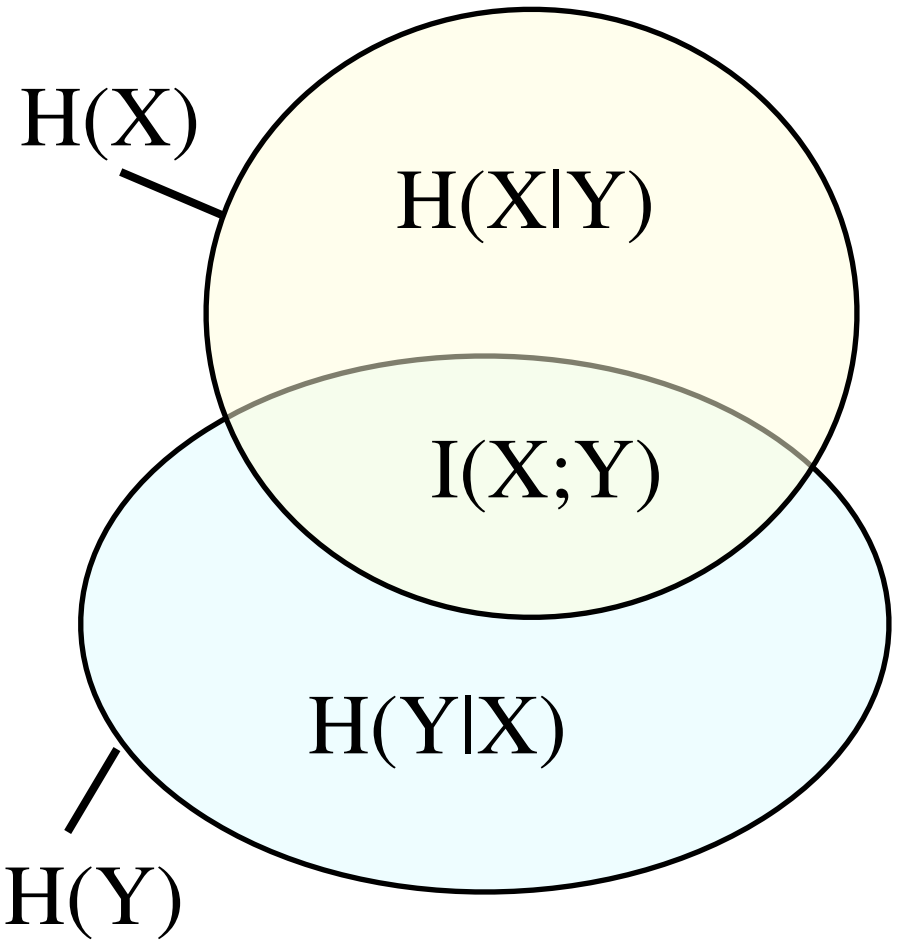


Mutual information I(X;Y)
- overlap between X and Y
- non-negative
- binary questions in common

Total area is H(X,Y)

I(X;X) = H(X)

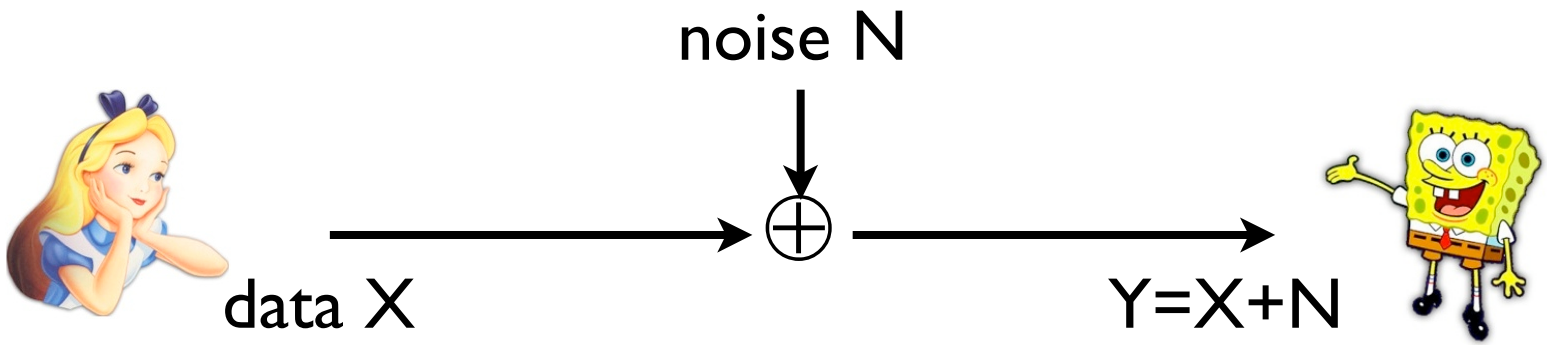
Mutual information



$$\mathbf{I(X;Y)}$$
$$= H(X) - H(X|Y)$$
$$= H(Y) - H(Y|X)$$
$$= H(X,Y) - H(X|Y) - H(Y|X)$$
$$= H(X) + H(Y) - H(X,Y)$$
$$= D(\, XY \,||\, X \times Y)$$

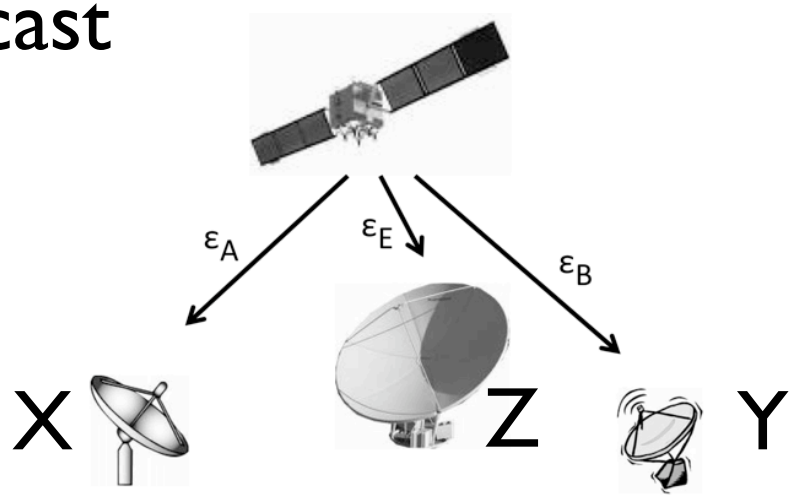
$$I(X;Y) = \Sigma_{xy}\, p_{xy} \log(p_{xy} / p_x\, p_y)$$

Uses of mutual information



Channel capacity I(X;Y)

Noisy broadcast



Secret key generation capacity I(X;Y|Z)

Uses of mutual information / cond. entropies

- Communication theory (error correction, ...)
- Crypto
- Biometrics
- Statistics (classification, hypothesis testing, ...)
- Image processing
- Statistical physics
- Econometrics
- ...

Other information measures

Min-entropy

- $H_{min}(X) = -\log(p_{max})$
- also called guessing entropy
- crypto: p_{max} is Prob[correct guess in one go]

Rényi entropy

- $\alpha > 1$
- collision entropy
- we will encounter $\alpha = 2$
- $\lim_{\alpha \to \infty} H_\alpha(X) = H_{min}(X)$
- $\lim_{\alpha \to 1} H_\alpha(X) = H(X)$

$$H_\alpha(X) = \frac{-1}{\alpha - 1} \log \sum_{x \in \mathcal{X}} p_x^\alpha$$

...