Joint Probability and the Markov Assumption

Joint Probability calculation when A and B are independent events

$$p(AB) = p(A)p(B)$$

$$LHS\ written\ also \quad p(A,B)\ or\ p(A \cap B)$$

Conditional Probability Calculation

$$p(A\,|\,B) = \frac{p(AB)}{p(B)}$$

Joint Conditional Probability Calculation (Chain Rule)

$$p(ABC) = p(A \mid BC)\,p(BC)$$

$$but \quad p(BC) = p(B \mid C)\,p(C)$$

$$so \quad p(ABC) = p(A \mid BC)\,p(B \mid C)\,p(C)$$

While the chain rule *can* be represented succinctly by the expression $p\left(\bigcap_{i=1}^{N} \quad\right) \quad {}_{i=1}^{N} \quad \left(\quad \bigcap_{j=1}^{N-1} \quad\right)$

where $\bigcap$ he intersection of probabilities (joint probability),

the expansion of the expression needed for calculation can get very messy. Here is the expansion for 6 variables. Imagine how messy it would get with more.

$$p(x_1, x_2, x_3, x_4, x_5, x_6) =$$

$$p(x_1 \mid x_2, x_3, x_4, x_5, x_6)\,p(x_2 \mid x_3, x_4, x_5, x_6)\,p(x_3 \mid x_4, x_5, x_6)\,p(x_4 \mid x_5, x_6\}\,p(x_5 \mid x_6)\,p(x_6)$$

The Markov Assumption cleans up the mess

$$p(x) =$$

$$p(x_1 \mid x_2, \cancel{x_3}, \cancel{x_4}, \cancel{x_5}, \cancel{x_6})\, \cancel{p(x_2 \mid x_3, x_4, x_5, x_6)}\, \cancel{p(x_3 \mid x_4, x_5, x_6)}\, \cancel{p(x_4 \mid x_5, x_6\}}\, \cancel{p(x_5 \mid x_6)}\, p(\cancel{x_6})$$

⇨

$$p(x) = p(x_1 \mid x_2)$$

There is no memory of previous events. X_1 is the future, x_2 is the now.
The x_3,x_4 …are the past; they are irrelevant

ENTROPY

Writings of Friar William of Occam….

"Pluralitas non est ponenda sine neccesitate"

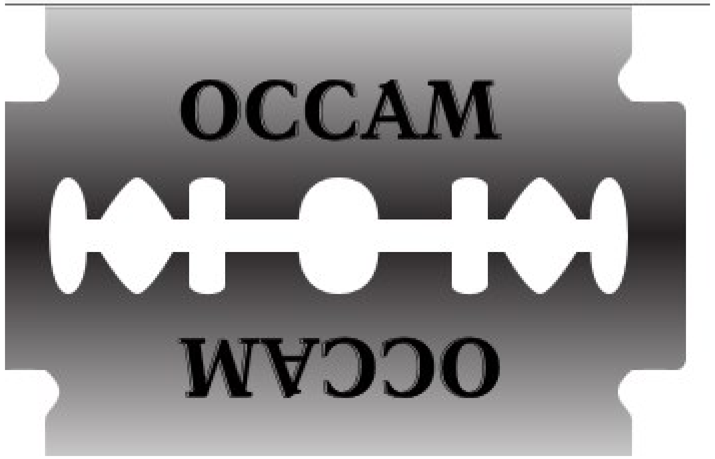
also

Frusta fit per plura quod potest fieri per pauciora.

OCCAM’S RAZOR

When there are 2 or more possible explanations for an observed event, we want to choose the explanation with the fewest and simplest assumptions

This is OCCAM’S razor: it ‘cuts’ away the excess verbiage, conditions, complications, and unnecessary logic



So, what is ‘simple’?

- We will always assume that a system will seek its lowest energy state

Equivalently

- We will always assume that a system will always seek its highest entropy state

Equivalently

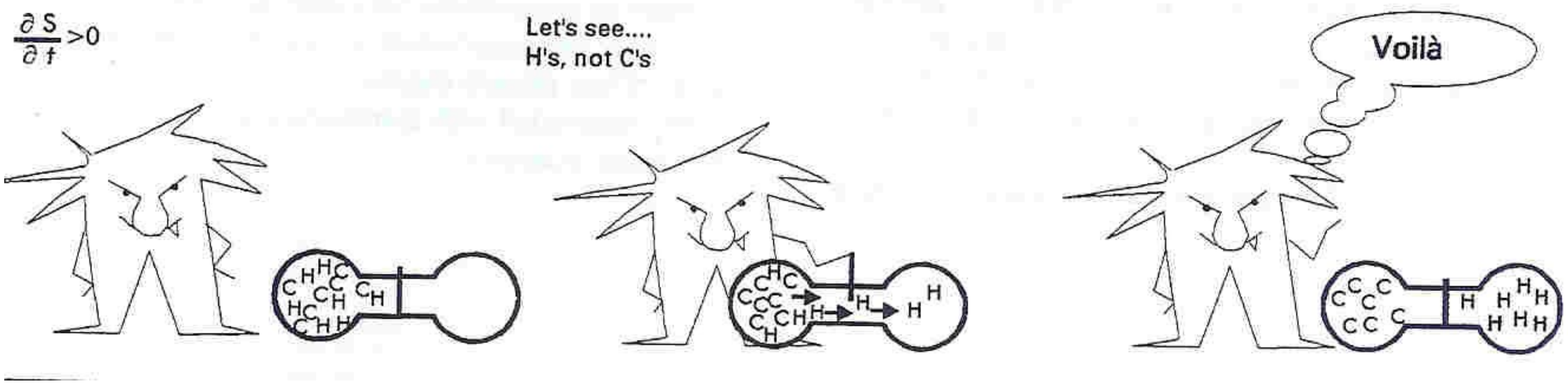
- The least (Kolmogorov) complexity is the best

Entropy

Entropy (real number $\geq$0) is a measure of disorder

- High entropy
  - High disorder
  - Much information needed to specify all the states
- Low entropy
  - Well organized
  - Little or no information needed to specify all the states

MAXWELL'S DEMON



A Demon operates a frictionless, weightless gate, and whenever a hot particle comes to the gate, he opens the gate, letting the hot particle through, then closes the gate.

In this fashion, he expends no Energy but drives Entropy down, in violation of the Second Law, and divorced from Energy and Enthalpy, in violation of the First Law

Shannon Entropy

- A bridge between statistical thermodynamics and information

- If you can know (or guess) about each particle in the system (say, a gas), you can determine the entropy of the system

Shannon Entropy

- Likewise, you can measure the information in a message by knowing (or guessing) the probability of each element of the message.

- Information relates to entropy through probability as:

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

  where S is entropy, p(x) is the probability of event x

Shannon Entropy

Shannon generalized this for a set of events in a system and for letters in an alphabet.

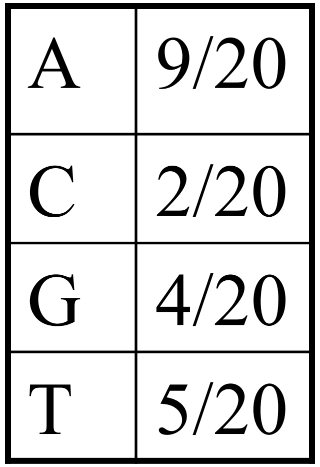
$$\boldsymbol{S} = -\sum_i p_i \log_2 p_i$$

Shannon Entropy

Example:

Given AATGATGCTGCAAATAAGTA

The frequencies (probabilities) of the bases are



A	9/20
C	2/20
G	4/20
T	5/20

Shannon Entropy

$$S =$$

$$-\left[.45\log_2 .45 \quad +.1\log_2 .1 \quad +.2\log_2 .2 \quad +.25\log_2 .25\right]$$

-1.81498

.45(-1.152)+.1(-3.32)+.2(-2.32)+.25(-2)

=1.815

Information

A measure of entropy reduction by sending a signal

Difference of entropy before and after

Relative Information

We can compare sequence data with background data, if we know the probability densities for both, using this notion of information, expressed as a ‘distance’.

For example, if we look at a character at position in a sequence, its distance to the background characters, say, for DNA, would be the Kullback-Liebler distance*

$$KL_{p\ relative\ to\ q} = \sum_{a \in \{A,C,G,T\}} p_{ai} \log_2\left(\frac{p_{ai}}{q_a}\right)$$

where q is the probability of the character in the sequence density and p is the probability in the background

*Not a real distance. Better called the K-L divergence. Distance is a metric; K-L is not;

$KL_{p||q} \neq KL_{q||p}$