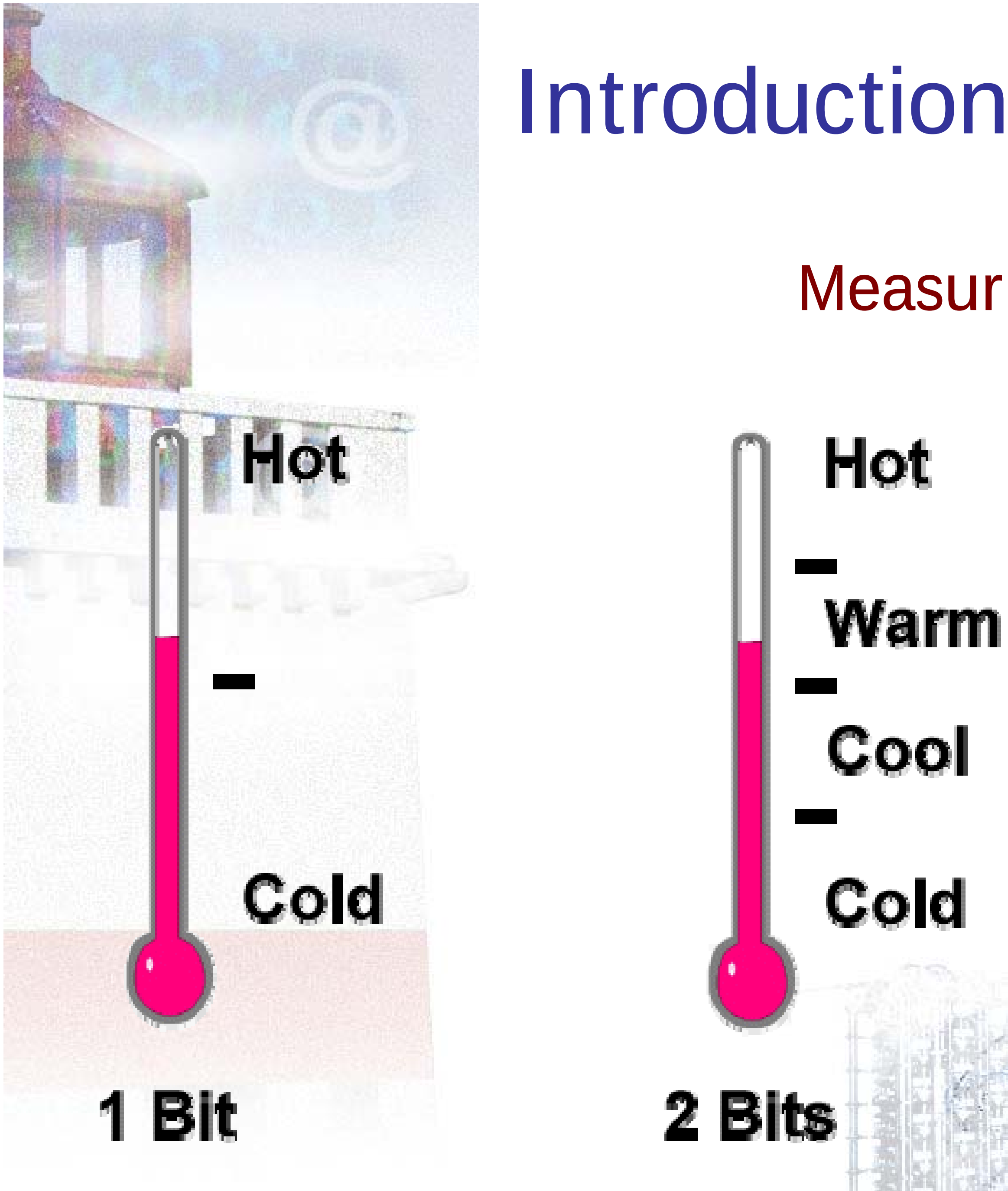
Introduction to Informatics

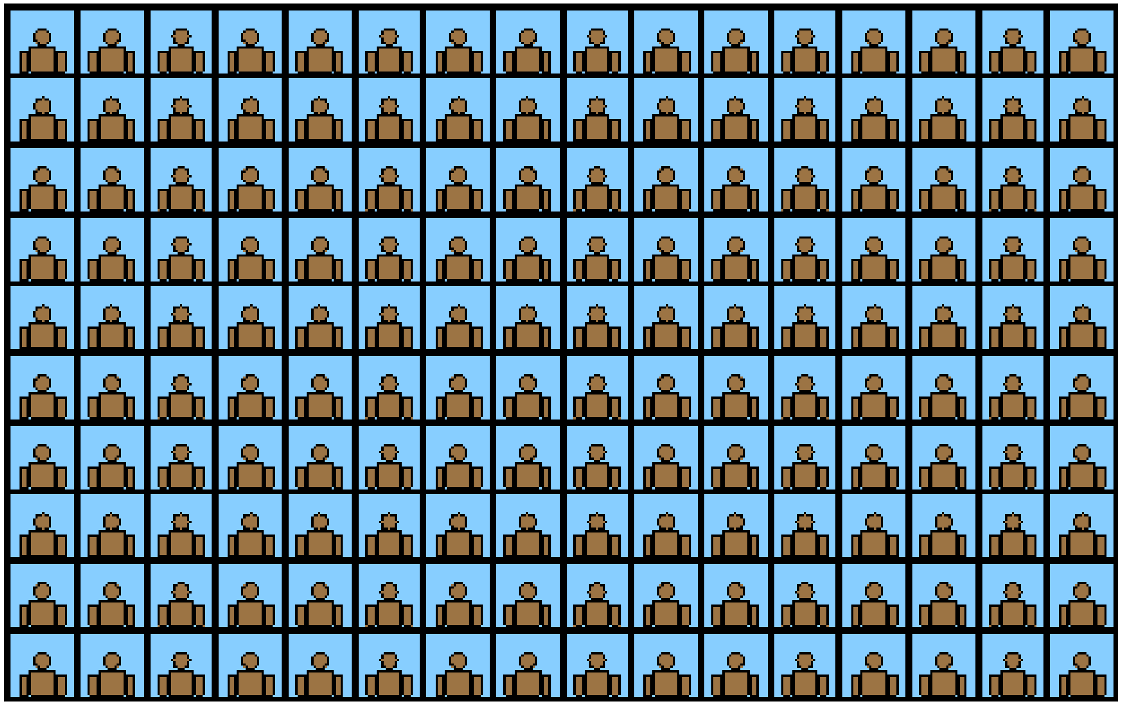
Lecture 21:

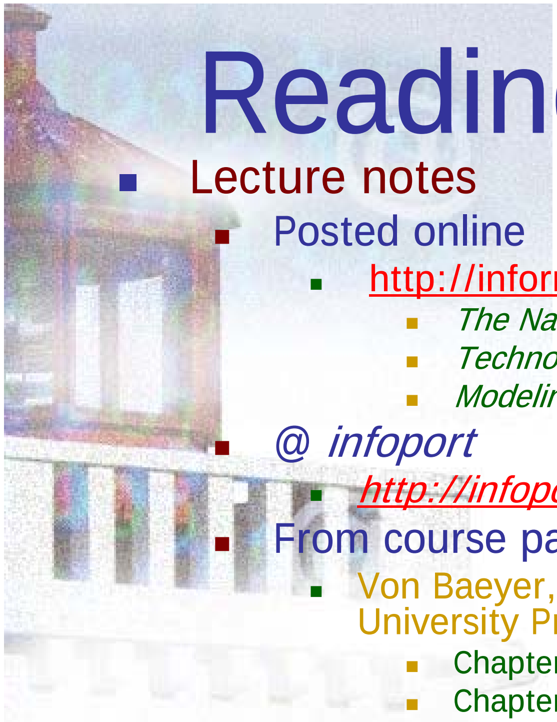
Measuring Information with Uncertainty



Luis M.Rocha and Santiago Schnell

NO LAB THIS WEEK !!!

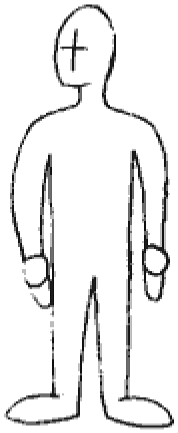




Readings until now

- Lecture notes
  - Posted online
    - http://informatics.indiana.edu/rocha/i101
      - *The Nature of Information*
      - *Technology*
      - *Modeling the World*
  - *@ infoport*
    - *http://infoport.blogspot.com*
  - From course package
    - Von Baeyer, H.C. [2004]. *Information: The New Language of Science*. Harvard University Press.
      - Chapters 1, 4 (pages 1-12)
      - Chapter 10 (pages 13-17)
    - From Andy Clark's book "*Natural-Born Cyborgs*“
      - Chapters 2 and 6 (pages 19 - 67)
    - From Irv Englander's book “*The Architecture of Computer Hardware and Systems Software*“
      - Chapter 3: Data Formats (pp. 70-86)
    - Klir, J.G., U. St. Clair, and B.Yuan [1997]. Fuzzy Set Theory: foundations and Applications. Prentice Hall
      - Chapter 2: Classical Logic (pp. 87-97)
      - Chapter 3: Classical Set Theory (pp. 98-103)
    - Norman, G.R. and D.L. Streinrt [2000]. *Biostatistics: The Bare Essentials*.
      - Chapters 1-3 (pages 105-129)
      - OPTIONAL: Chapter 4 (pages 131-136)
      - Chapter 13 (pages 147-155)
      - Chapter 5 (pages 141-144)
    - Igor Aleksander, "Understanding Information Bit by Bit"
      - Pages 157-166

Assignment Situation



- Labs
  - Past
    - Lab 1: Blogs
      - Closed (Friday, January 19): Grades Posted
    - Lab 2: Basic HTML
      - Closed (Wednesday, January 31): Grades Posted
    - Lab 3: Advanced HTML: Cascading Style Sheets
      - Closed (Friday, February 2): Grades Posted
    - Lab 4: More HTML and CSS
      - Closed (Friday, February 9): Grades Posted
    - Lab 5: Introduction to Operating Systems: Unix
      - Closed (Friday, February 16): Grades Posted
    - Lab 6: More Unix and FTP
      - Closed (Friday, February 23): Grades Posted
    - Lab 7: Logic Gates
      - Closed (Friday, March 9): Grades Posted
    - Lab 8: Intro to Statistical Analysis using Excel
      - Closed (Friday, March 30): being graded
    - Lab 9: Data analysis with Excel (linear regression)
      - Due Friday, April 6
  - Next: Lab 10
    - Lab 10: Simple programming in Excel and Measuring Uncertainty
      - April 12 and 13, Due April 20

- Assignments
  - Individual
    - First installment
      - Closed: February 9: Grades Posted
    - Second Installment
      - Past: March 2: Grades Posted
    - Third installment
      - Past: Being Graded
    - Fourth Installment
      - Presented April 10th, Due April 20th
  - Group
    - First Installment
      - Past: March 9th, Being graded
    - Second Installment
      - March 29; Due Friday, April 6

Group Assignment

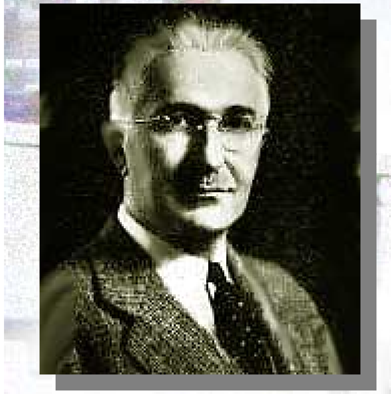
- Second Installment: Given the text of "Lottery of Babylon" by Jorge Luis Borges
  - Measures of central tendency and dispersion of letter frequency
  - Probability of a letter being a vowel
  - Probability of a letter being a consonant
  - Conditional probability of letters 'e' and 'u'
    - P(e|♥) where ♥ is the letter occurring before 'e'
    - P(u|♥) where ♥ is the letter occurring before 'u'
    - Compute for all letters (not space)
    - Produce histogram of P(e|♥), for all ♥.
    - Produce histogram of P(u|♥), for all ♥.
    - Discuss the independence of 'e' and 'u' from other letters
  - Upload to Oncourse

$$P(e\,|\,h) = \frac{|h \wedge e|}{|h|} = \frac{|'he'|}{|h|}$$

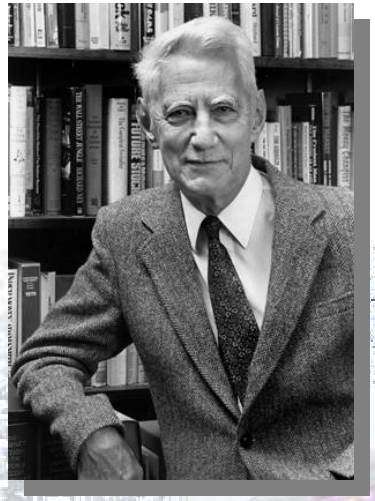
$$P(e) = \frac{|e|}{N}$$

Why are we dealing with uncertainty in Informatics?

- Information is transmitted through noisy communication channels
  - Ralph Hartley and Claude Shannon (at Bell Labs), the fathers of Information Theory, worked on the problem of efficiently transmitting information; i. e. ***decreasing the uncertainty*** in the transmission of information!



Hartley, R.V.L., "Transmission of Information", *Bell System Technical Journal*, July 1928, p.535.



C. E. Shannon, ``A mathematical theory of communication,'' *Bell System Technical Journal*, vol. 27, pp. 379-423 and 623-656, July and October, 1948.

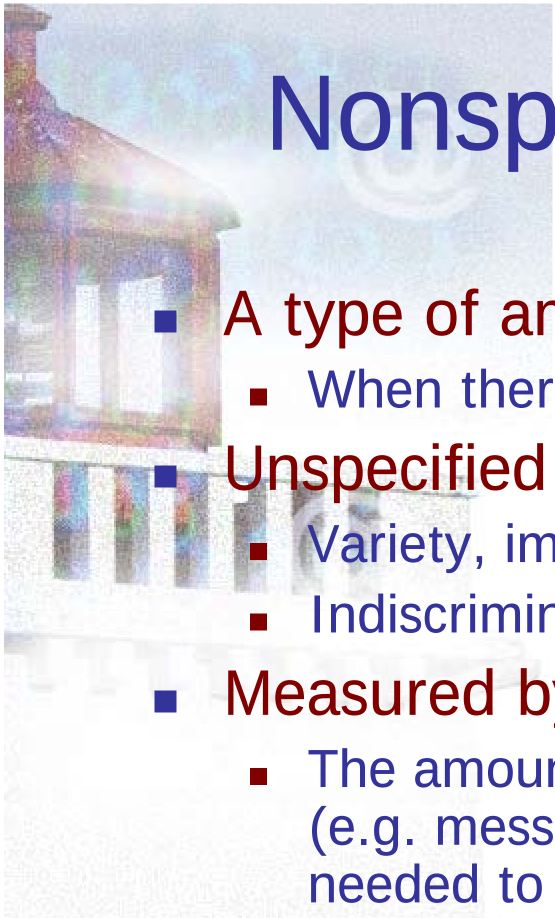
Uncertainty-based Information

- In a problem-solving or decision-making activity
  - Uncertainty is the result of some information deficiency
- *Information* is defined as “a measure of the freedom from *choice* with which a message is *selected* from the set of all possible messages”
  - Bit (short for *binary digit*) is the most elementary choice one can make between **two equally likely choices**
    - Between two items: “0’ and “1”, “heads” or “tails”, “true” or “false”, etc.
      - Example, if we know that a coin is to be tossed, but are unable to see it as it falls, a message telling whether the coin came up heads or tails gives us one bit of information
  - Therefore the ***fundamental unit of information***

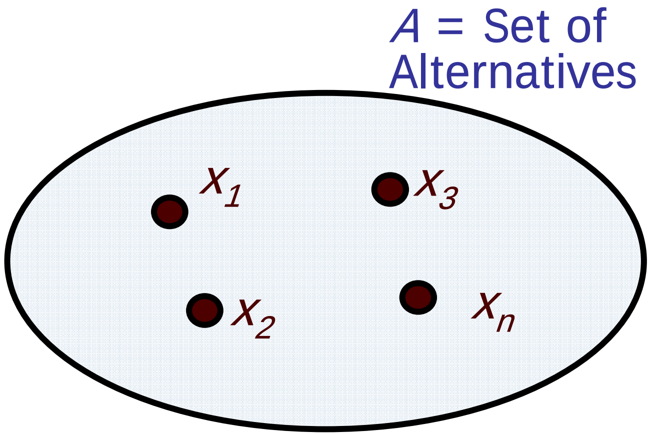
[Klir and Weirman, “Uncertainty-based information”]

Let’s talk about choices

- Multiplication Principle
  - “If some choice can be made in M different ways, and some subsequent choice can be made in N different ways, then there are M x N different ways these choices can be made in succession” [Paulos]
    - 3 shirts and 4 pants = 3 x 4 = 12 outfit choices



Nonspecificity



- A type of ambiguity
  - When there are choices
- Unspecified distinctions between several alternatives
  - Variety, imprecision
  - Indiscriminate choices
- Measured by Hartley measure
  - The amount of uncertainty associated with a set of alternatives (e.g. messages) is measured by the amount of information needed to remove the uncertainty

$$H(A) = \log_2 |A|$$

Measured in bits

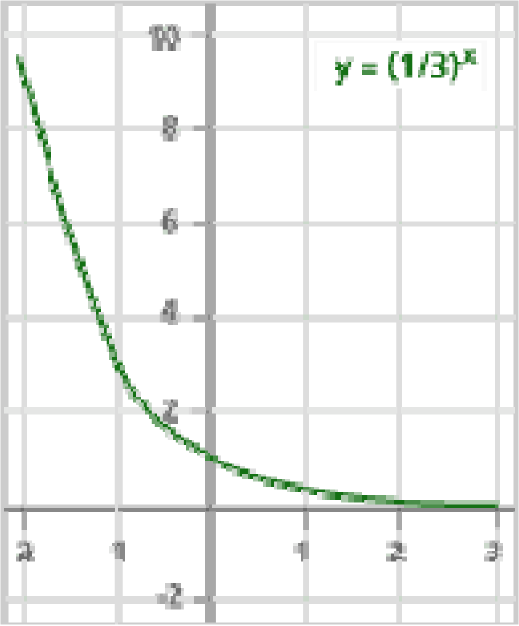
Number of Choices

Exponential Function

$$f(x) = b^x$$

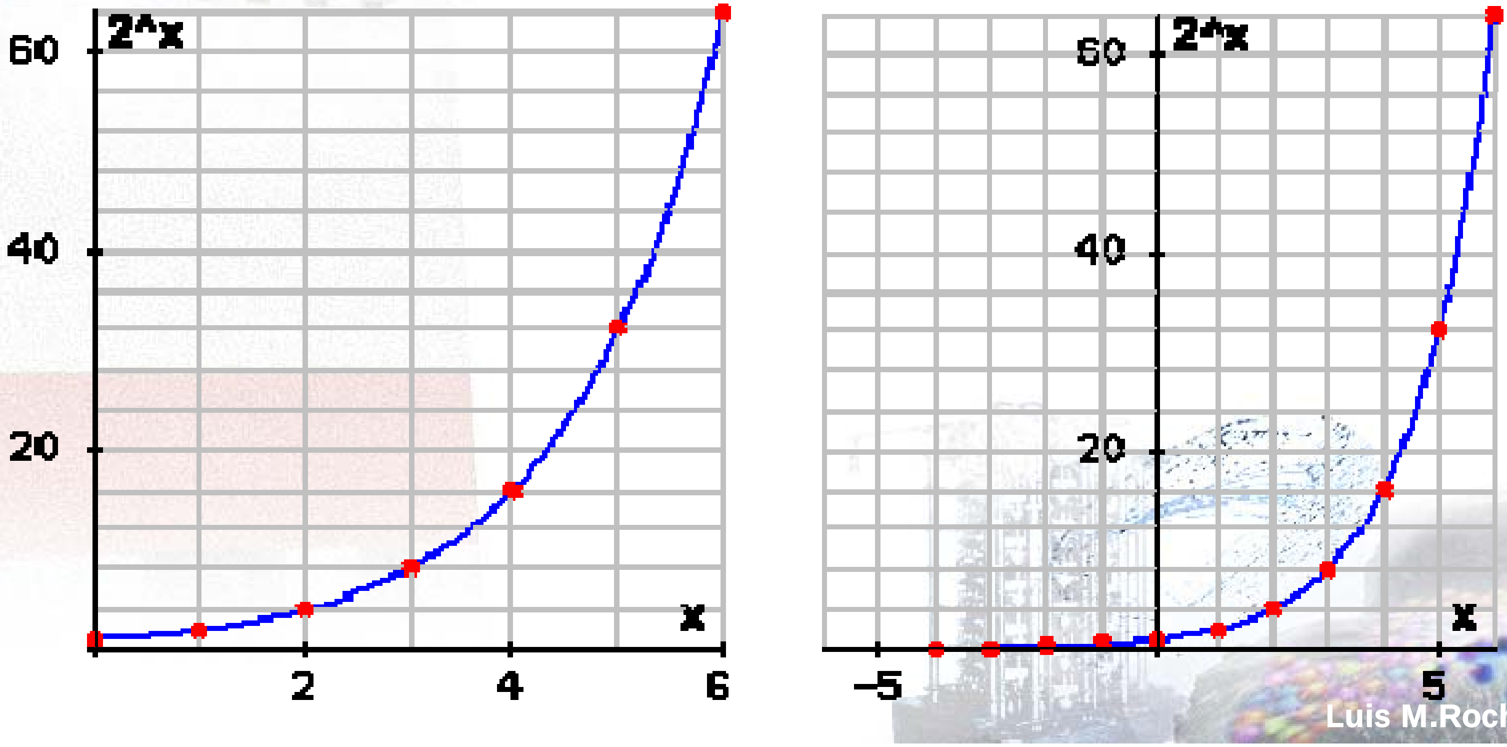
with base b

Positive real number

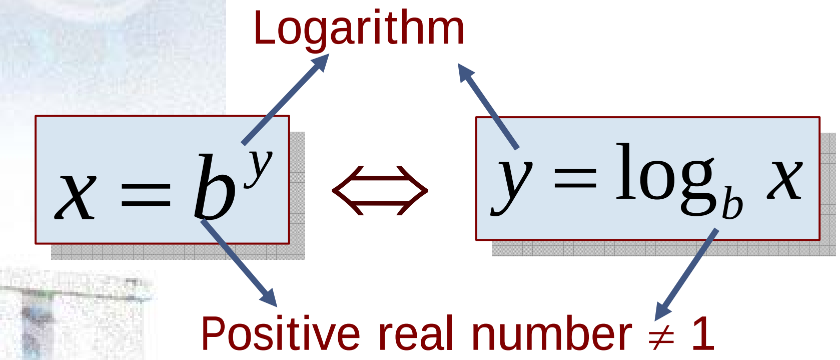


Decay: 0<b<1

Growth: b>1



Logarithm Function



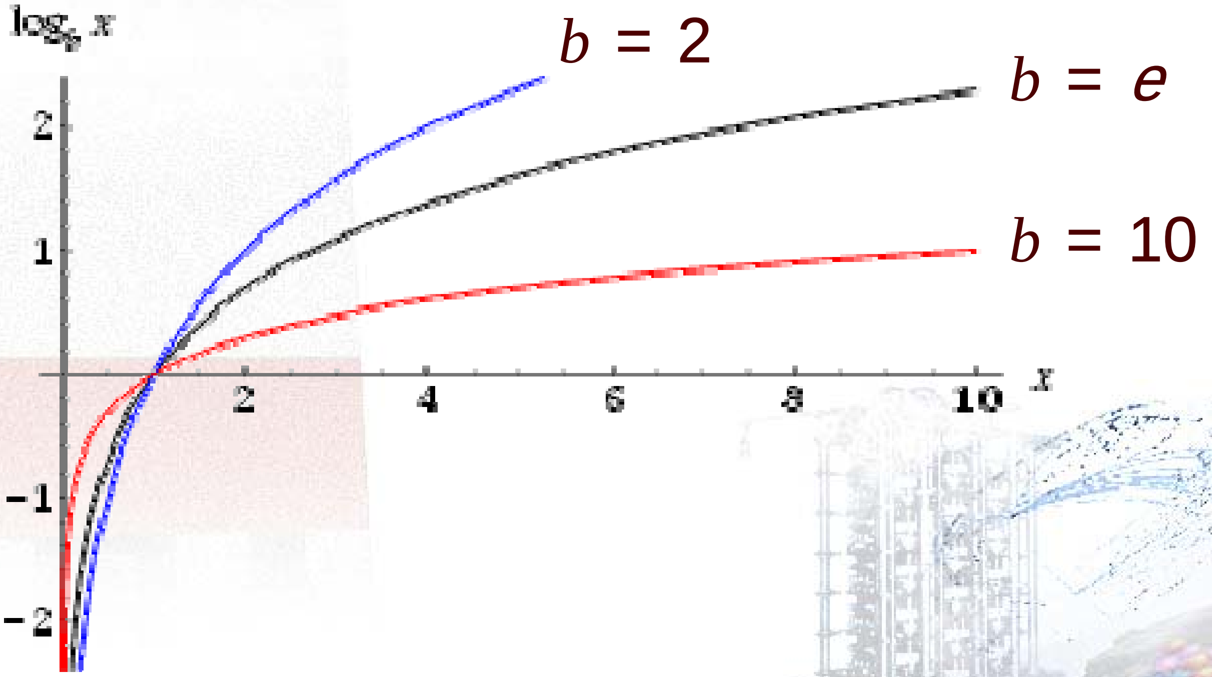
Logarithm

$x = b^y \Leftrightarrow y = \log_b x$

Positive real number ≠ 1

with base b

Example: $\log_2 8 = 3$ because $2^3 = 8$



$\left(1+\frac{1}{n}\right)^n, n \to \infty$

2.71828.....

Properties of Logarithms

$$x = b^y \Leftrightarrow y = \log_b x$$

$$\log_b b = 1 \qquad \log_b 1 = 0$$

b=2, computes the uncertainty of 2 choices as 1: the bit

$$\log_b(M.N) = \log_b M + \log_b N$$

Converts multiplication into sum. Easier to deal with accounting choices

$$\log_b\left(N^r\right) = r.\log_b N$$

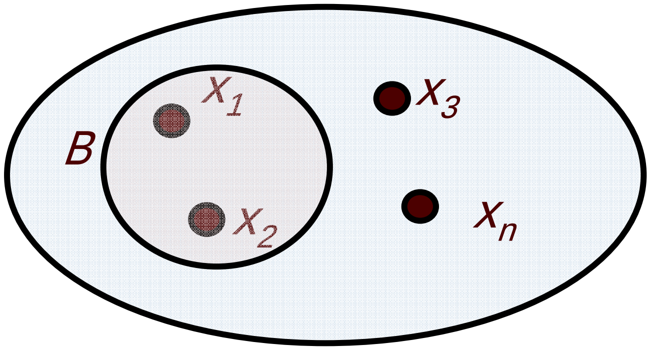
$$\log_b\left(\frac{M}{N}\right) = \log_b M - \log_b N$$

$$\log_b\left(b^x\right) = x$$

$$\log_b x = \frac{\log_a x}{\log_a b}$$

Hartley Uncertainty

A = Set of Alternatives



- Nonspecificity
  - Hartley measure
    - The amount of uncertainty associated with a set of alternatives (e.g. messages) is measured by the amount of information needed to remove the uncertainty

Quantifies how many yes-no questions need to be asked to establish what the correct alternative is

$$H(A) = \log_2 |A|$$

Measured in bits

Number of Choices

Elementary Choice is between 2 alternatives: 1 bit

$$H(B) = \log_2(2) = 1$$

$$\log_2(4) = 2$$

$$2^2 = 4$$

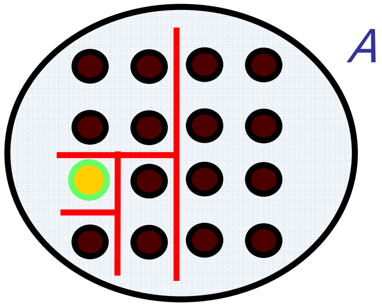
$$\log_2(16) = 4$$

$$2^4 = 16$$

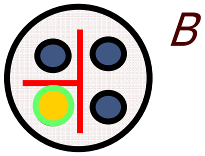
$$\log_2(1) = 0$$

Hartley Uncertainty

- Example
  - Menu Choices
    - A = 16 Entrees
    - B = 4 Desserts
  - How many dinner combinations?
    - 16 x 4 = 64



$$H(A) = \log_2(16) = 4$$



$$H(B) = \log_2(4) = 2$$

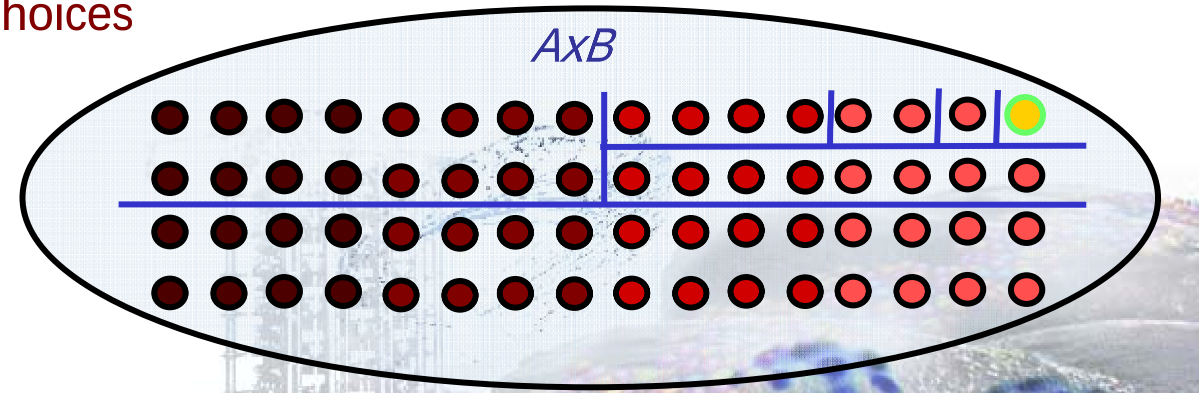
$$H(A \times B) = \log_2(16 \times 4) =$$
$$= \log_2(16) + \log_2(4) = 6$$

$$H(A) = \log_2 |A|$$

Measured in bits

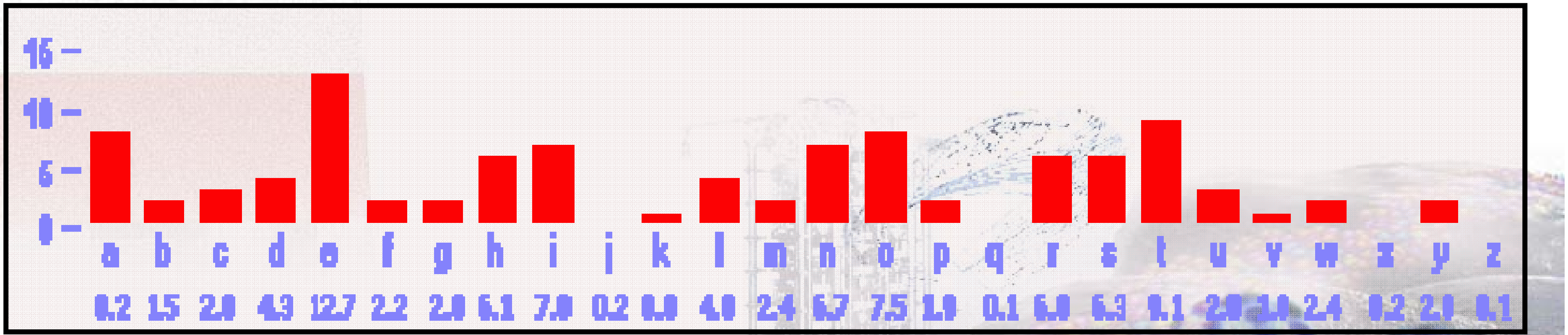
Number of Choices

Quantifies how many yes-no questions need to be asked to establish what the correct alternative is



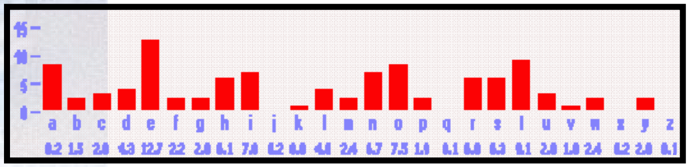
What about probability?

- Some alternatives may be more probable than others!
- A different type of ambiguity
  - Alternatives are distinct
    - **Conflict**, strife, discord
- Measured by Shannon's ***entropy*** measure
  - The amount of uncertainty associated with a set of alternatives (e.g. messages) is measured by the ***average*** amount of information needed to remove the uncertainty

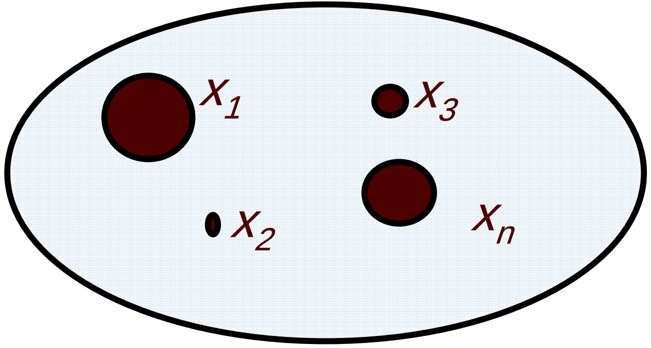


Luis M.Rocha and Santiago Schnell

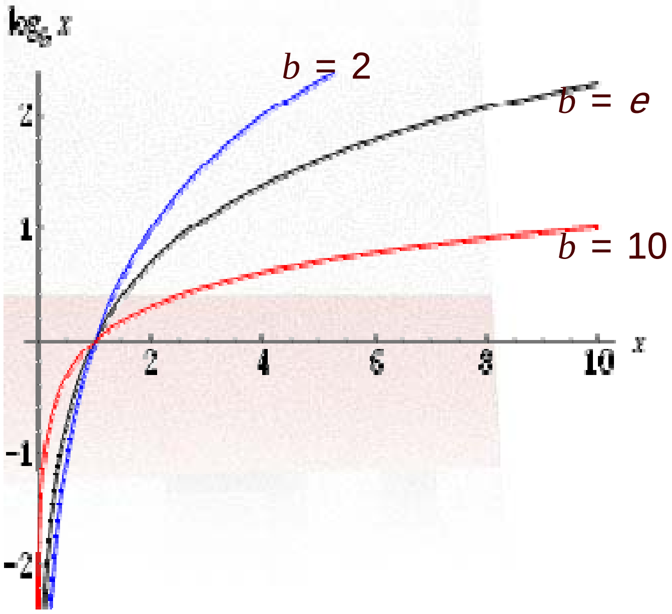
Entropy



A = Set of weighted Alternatives



- Shannon’s measure
  - The ***average*** amount of uncertainty associated with a set of ***weighted*** alternatives (e.g. messages) is measured by the ***average*** amount of information needed to remove the uncertainty

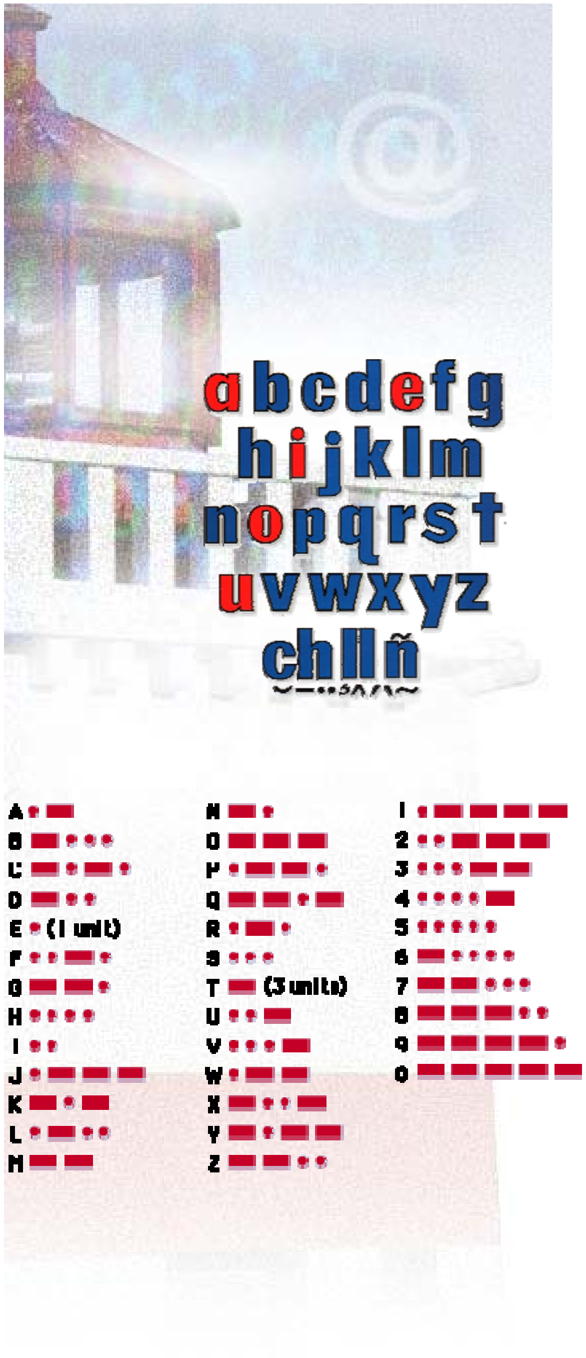


$$H_S(A) = -\sum_{i=1}^{n} p(x_i)\log_2\left(p(x_i)\right)$$

Probability of alternative

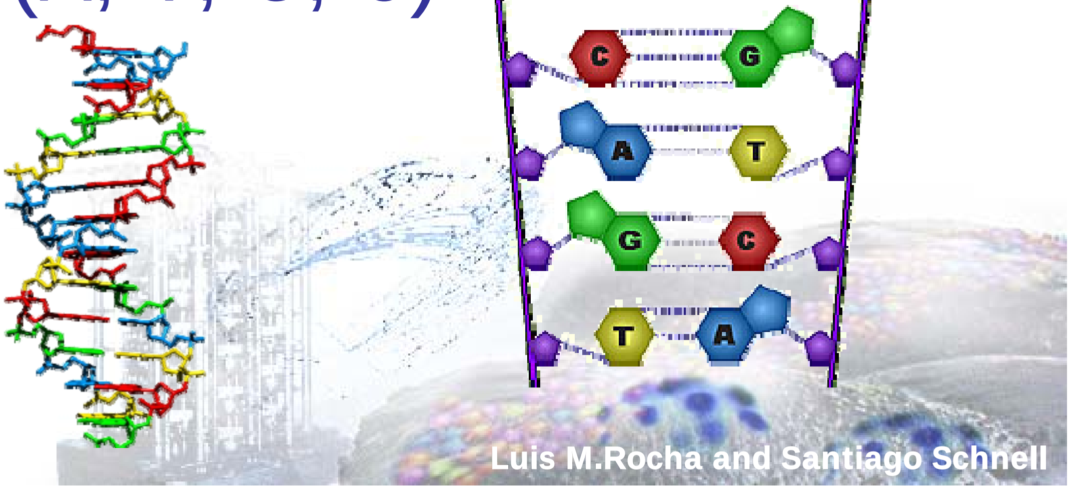
Measured in bits

Entropy of a message



Message encoded in an alphabet of *n* symbols, for example:

- English (26 letters + space + punctuations)
- Morse code (dot, dash, space)
- DNA (A, T, G, C)



Luis M.Rocha and Santiago Schnell

Shannon’s entropy formula

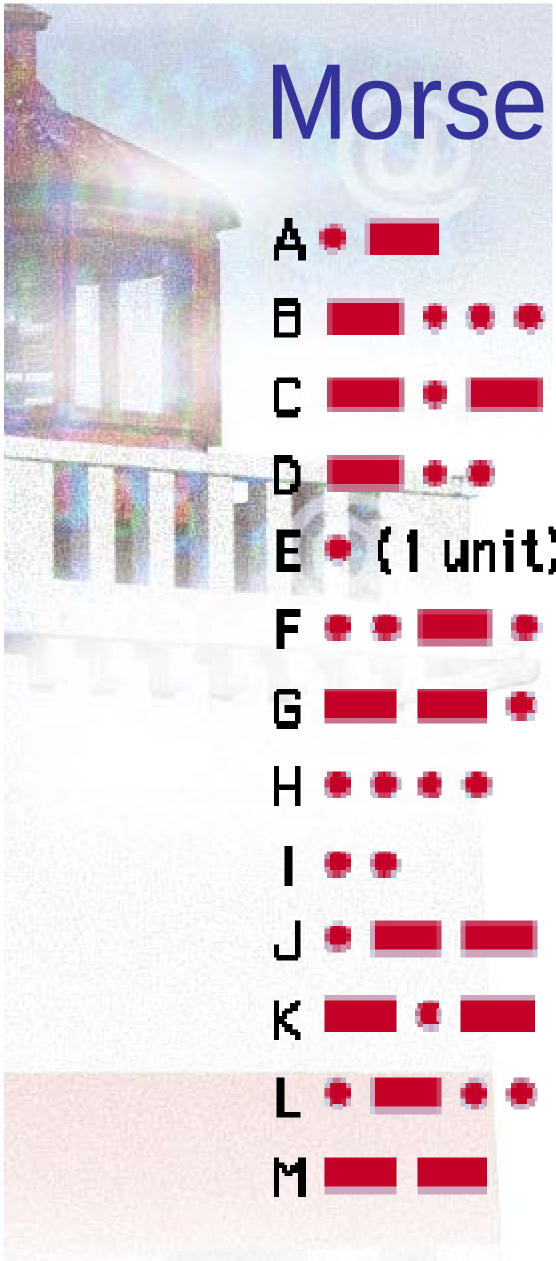
Shannon formulated the following problem:

Let’s us define a quantity that *measures*

- *missing information*, how much information is needed to establish what the symbol is, or
- *uncertainty* about what the symbol is, or
- ***on average***, how many *yes-no* questions need to be asked to establish what the symbol is.

$$H_S(A) = -\sum_{i=1}^{n} p(x_i)\log_2\left(p(x_i)\right)$$

Morse Code



A	● ▬	N	▬ ●	1	● ▬ ▬ ▬ ▬
B	▬ ● ● ●	O	▬ ▬ ▬	2	● ● ▬ ▬ ▬
C	▬ ● ▬ ●	P	● ▬ ▬ ●	3	● ● ● ▬ ▬
D	▬ ● ●	Q	▬ ▬ ● ▬	4	● ● ● ● ▬
E	● (1 unit)	R	● ▬ ●	5	● ● ● ● ●
F	● ● ▬ ●	S	● ● ●	6	▬ ● ● ● ●
G	▬ ▬ ●	¯	▬ (3 units)	7	▬ ▬ ● ● ●
H	● ● ● ●	U	● ● ▬	8	▬ ▬ ▬ ● ●
I	● ●	V	● ● ● ▬	9	▬ ▬ ▬ ▬ ●
J	● ▬ ▬ ▬	W	● ▬ ▬	0	▬ ▬ ▬ ▬ ▬
K	▬ ● ▬	X	▬ ● ● ▬		
L	● ▬ ● ●	Y	▬ ● ▬ ▬		
M	▬ ▬	Z	▬ ▬ ● ●		

dot, dash, space

Examples – Morse code

$$H_S(A) = -\sum_{i=1}^{n} p(x_i)\log_2(p(x_i))$$

$$H_S = -(p_1 \log_2(p_1) + p_2 \log_2(p_2) + p_3 \log_2(p_3))$$

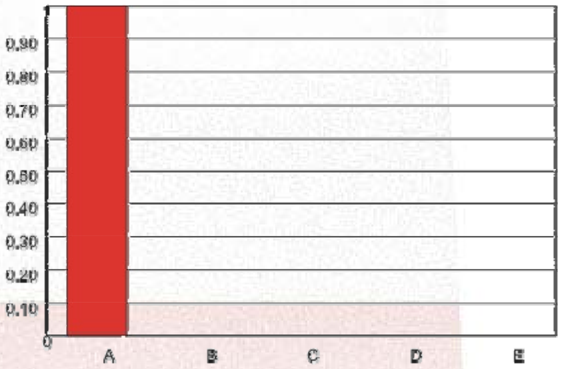
dot, dash, space

- All dots: $p_1 = 1$, $p_2 = p_3 = 0$.
  - Take any symbol – it's a dot; no uncertainty, no question needed, no missing information, $H_S = -1.\log_2(1) = 0$.
- 50-50 chance that it's a dot or a dash: $p_1 = p_2 = 1/2$, $p_k = 0$.
  - Given the *probabilities*, need to ask one question
  - one piece of missing information
    - $H_S = -(1/2.\log_2(1/2) + 1/2.\log_2(1/2)) = -1.\log_2(1/2) = -(\log_2(1) - \log_2(2)) = \log_2(2) = 1$ bit
- Uniform: all symbols equally likely, $p_1 = p_2 = p_3 = 1/3$.
  - Given the *probabilities,* need to ask as many as 2 questions - 2 pieces of missing information, $H_S = -\log_2(1/3) = -(\log_2(1) - \log_2(3)) = \log_2(3) = 1.59$ bits

Example English

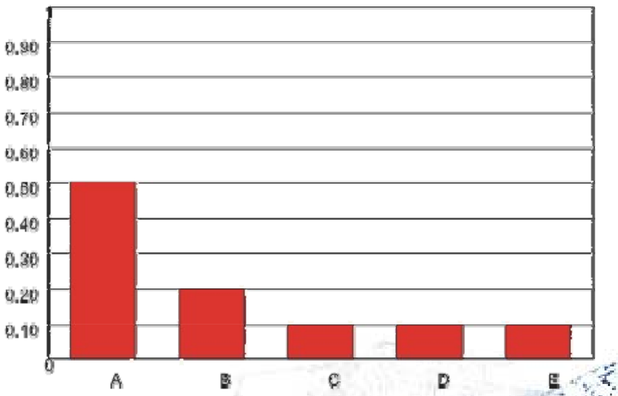
- Given a symbol set {A,B,C,D,E}
  - And occurrence probabilities P_A, P_B, P_C, P_D, P_E,
- The Shannon entropy is
  - The average minimum number of bits needed to represent a symbol

$$H_S = -\left(p_A \log_2\left(p_A\right) + p_B \log_2\left(p_B\right) + p_C \log_2\left(p_C\right) + p_D \log_2\left(p_D\right) + p_E \log_2\left(p_E\right)\right)$$



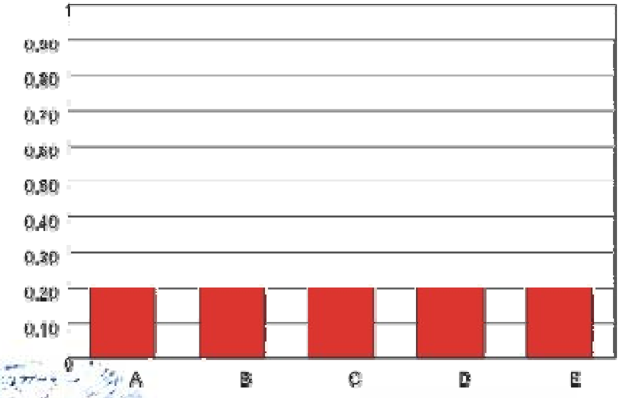
H_S = 0

0 questions



H_S = 1.96

≈2 questions



H_S = 2.32

Shannon’s entropy

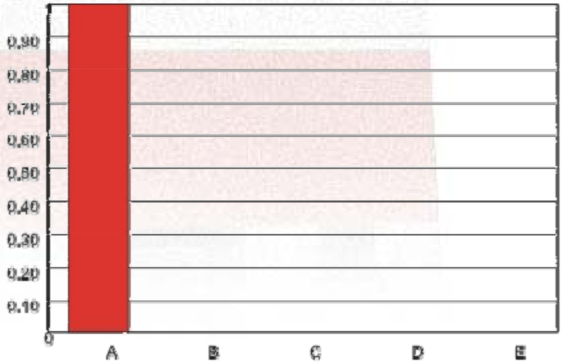
on average, how many *yes-no* questions need to be asked to establish what the symbol is.

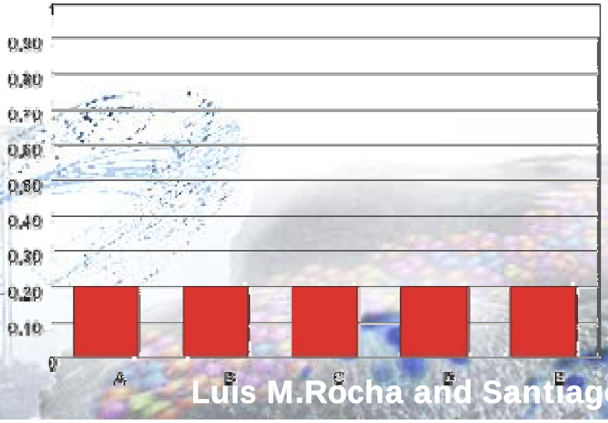
$$H_S(A) = -\sum_{i=1}^{n} p(x_i)\log_2\big(p(x_i)\big)$$

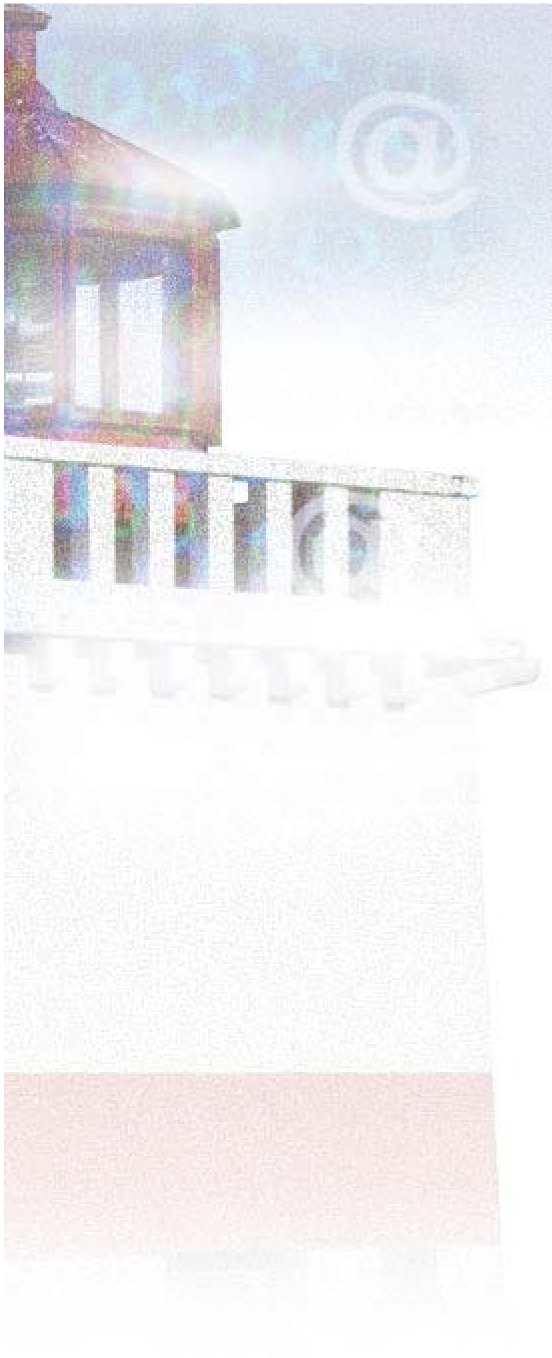
$$H_S \in \big[0, \log_2 |X|\big]$$

For one alternative

Uniform distribution

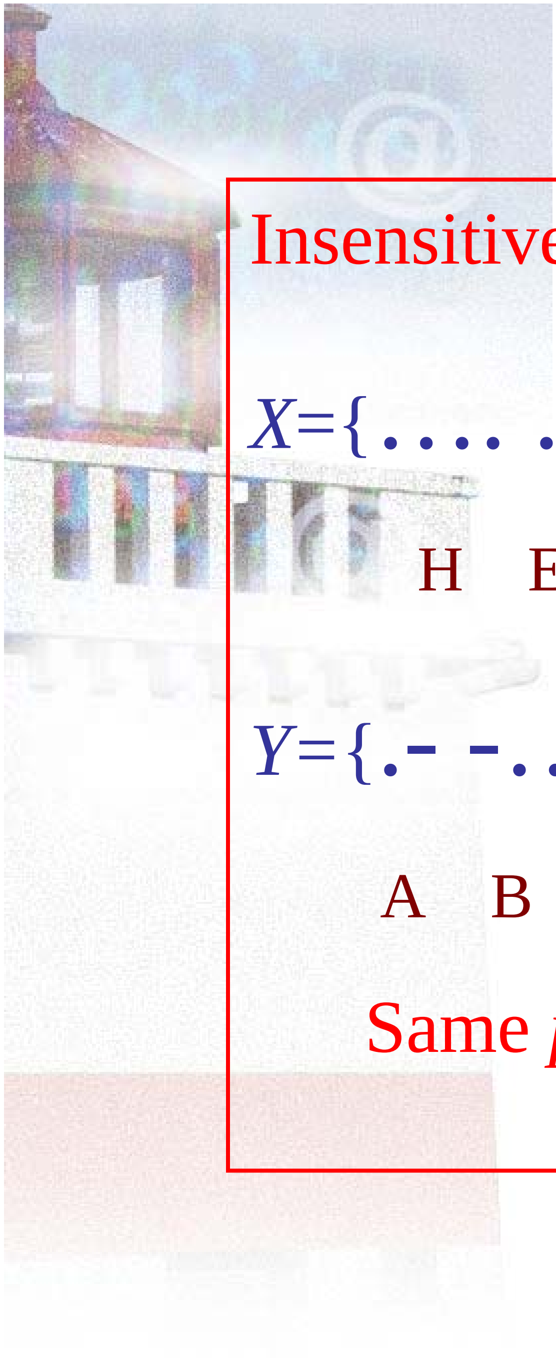






Critique of Shannon’s communication theory

- The entropy formula as a measure of information is arbitrary
- Shannon’s theory measures quantities of information, but it does not consider information content
- In Shannon’s theory, the semantic aspects of information are irrelevant to the engineering problem



Insensitive to meaning: Morse revisited

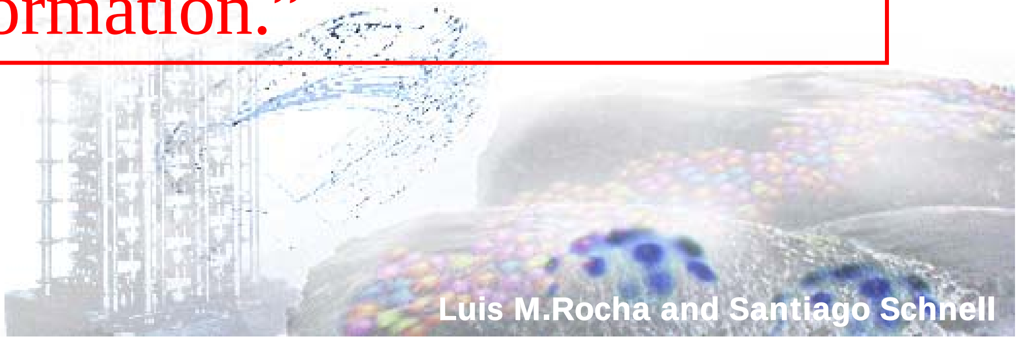
X={.... . .-.. .-.. --- .-- --- .-. .-.. -..}

H E L L O W O R L D

Y={.- -... -.-. -.. . ..-. --. ....--- -.-}

A B C D E F G H I J M

Same p_k's, same entropies – same “missing information.”

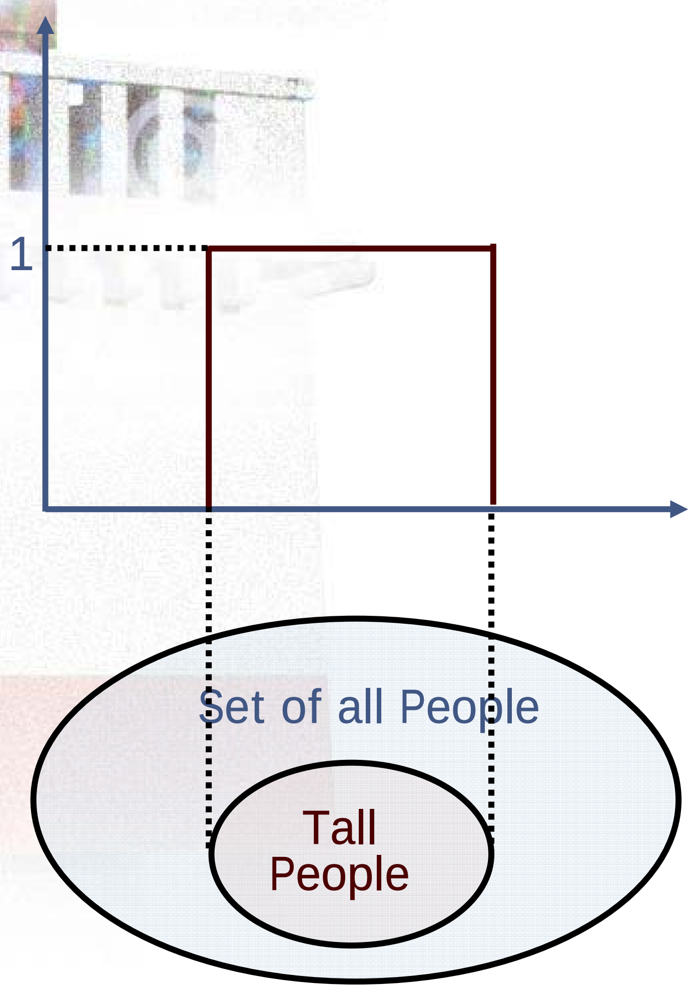


Other Forms of Uncertainty

- Vagueness or fuzziness
  - Simultaneously being “True” and “False”
  - Fuzzy Logic and Fuzzy Set Theory

"Me, ambivalent?...Well, yes and no..."

From Crisp to Fuzzy Sets

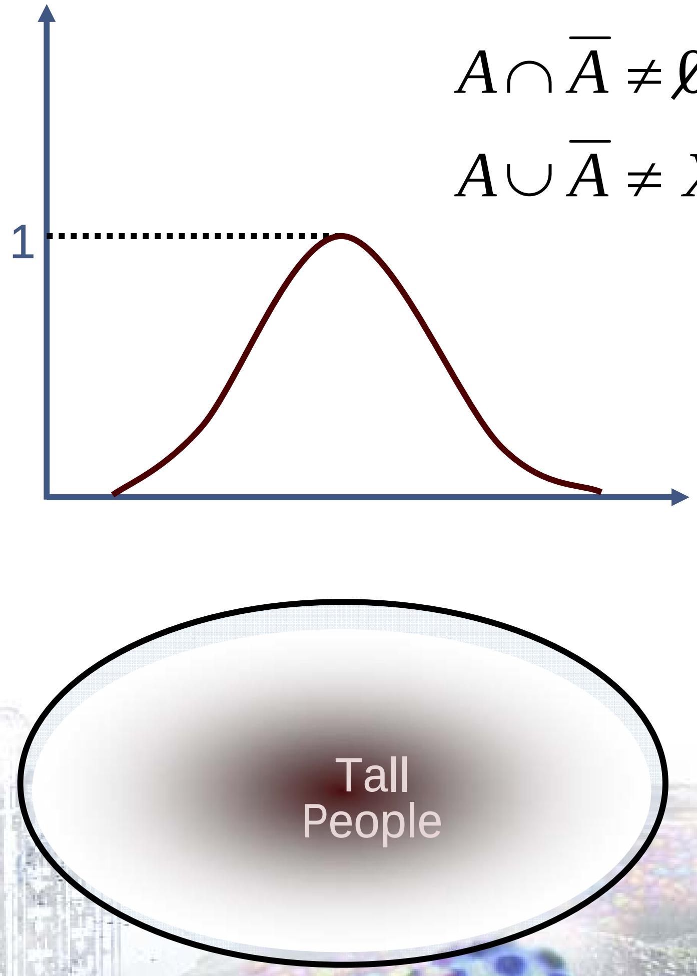


Fuzziness: Being and Not Being

Laws of Contradiction and Excluded Middle are Broken

$$A \cap \overline{A} \neq \emptyset$$

$$A \cup \overline{A} \neq X$$



Luis M.Rocha and Santiago Schnell

Frequency Analysis and Cryptography

- Cryptography
  - Derived from the Greek word *Kryptos*: hidden
- See Simon Singh's The Code Book CD-ROM
  - Enigma

Next Class!

- Topics
  - Algorithms
- Readings for Next week
  - *@ infoport*
  - From course package
    - Norman, G.R. and D.L. Streinrt [2000]. *Biostatistics: The Bare Essentials*.
      - Chapters 1-3 (pages 109-134)
      - OPTIONAL: Chapter 4 (pages 135-140)
      - Chapter 13 (pages 151-159)
      - Chapter 5 (pages 141-144)
    - Von Baeyer, H.C. [2004]. *Information: The New Language of Science*. Harvard University Press.
      - Chapter 10 (pages 13-17)
    - Igor Aleksander, "Understanding Information Bit by Bit"
      - Pages 157-166
- No Lab this week!!!