CS-341 Lecture 3

February 6, 2001

Information Encoding

- Binary Numbers
  - Binary Digits: bits (Tukey)
  - Uncertainty Reduction (Shannon)
- Encoding Information
  - Instructions
  - Text
  - Numbers
    - Discrete, continuous
  - Analogs of physical information
    - Aural, visual

Information Theory

- John Tukey of Bell Laboratories coined the word "bit" meaning "binary digit.
- Claude Shannon, also of Bell Labs, used the bit as a unit of information, the amount of uncertainty that is reduced by answering one yes/no question, in 1948.
- How many bits are needed …
  - To encode *n* different numbers?
    - $log_2 n$
  - To encode the days of the week?
  - To encode the faces of a die?
    - "From Shannon to Tannen!"

Encoding Instructions

- Operation Code
- Location of operands
  - Stack
  - Registers
  - Memory
    - Address
    - Index Register
    - Base/Segment Register

Encoding Text: ASCII

- American Standard Code for Information Interchange (ASCII)
  - 7 bits per character.
  - 0x00 through 0x1F are control codes
  - American National Standards Institute (ANSI) standard X.64 defines *escape sequences* for extended terminal control, using ASCII characters.

Encoding Text: ISO Latin-1

- International Standards Organization (ISO)
- 8-bits per character
- First 128 characters are the same as ASCII 0x00 through 0x7F.
- Can represent European languages with diacritical marks (à á â ã ä å, etc.)
- Also known as ISO-8859-1
- Other ISO Latin-*n* codes for other sets of alphabets.

Encoding Text: Unicode

- 16 bits per character
- First 256 codes (0x0000 through 0x00FF) are the same as ISO Latin -1.
- Used by Java, Windows NT, …
- Version 3.0.1 of the standard has 49,194 different characters.
  - ISO/IEC 10646: ISO name for the standard.
  - Karit, Manipuri, Moso, Pahawh Hmong, Rong, Tai Lu, Tai Mau, and Tifinagh are not yet supported!
- http://www.unicode.org has information about the standard, the consortium that is developing it, etc.