Information, Compression, and Encoding

Instructor: Dmitri A. Gusev

Fall 2007

CS 210: Computing and Culture

Lecture 6, October 8, 2007

Information Theory

- Claude Shannon (1948): “A Mathematical Theory of Communication”.
- Source Coding Theorem: On average, the number of bits needed to represent the result of an uncertain event is given by its *entropy.*
- Noisy-Channel Coding Theorem: Reliable communication is possible over noisy channels provided that the rate of communication is below a certain threshold called the *channel capacity*.

Data Compression

Save storage space; speed up transmission.

Bandwidth: Bits (bytes) per second

Compression ratio: $\frac{size\_of\_the\_compressed\_data}{size\_of\_the\_uncompressed\_data}$

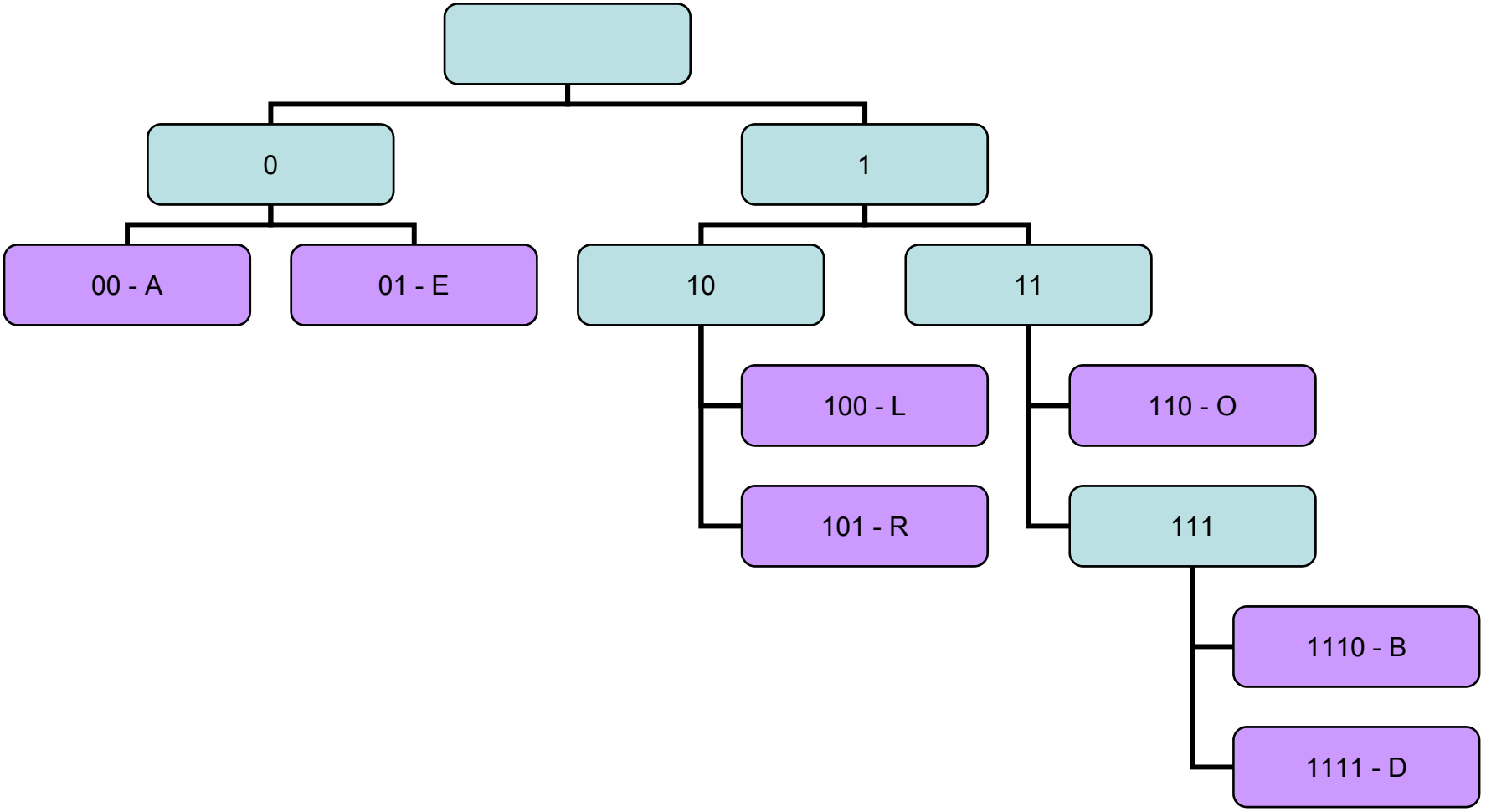
Lossless vs. *lossy* compression

Keyword encoding: Replace a popular word with a shorter code (“with” → “w/”, “without” → “w/o”)

Run-length encoding: AAAAAA → A6

Can combine the two.

Huffman Encoding



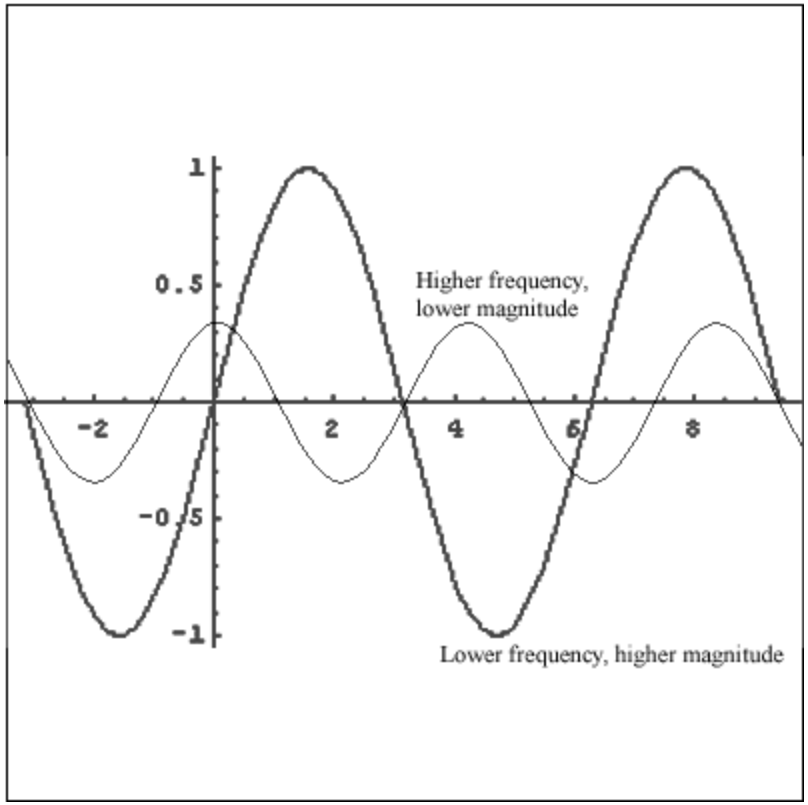
Why would anyone accept *lossy* compression?

Audio formats: WAV, MP3…

Video *codecs* (**co**mpressors / **dec**ompressors): MPEG-2, MPEG-4, DiVX… *Temporal compression* and *spatial compression*

Raster image formats: JPEG, JPEG2000, GIF, PCX, BMP, RAW…

Vector graphics: Images are described in terms of lines and geometric shapes



Representing Text

- Encoding characters vs. *formatting* (fonts, margins, tables, color, etc.)
- A *character set* is a list of characters and the codes used to represent them. How many characters do we need?..
- *ASCII* (American Standard Code for Information Interchange): Originally allowed 128 unique characters. The eighth bit was a *check bit*. *Latin-1 Extended ASCII character set*: 256 characters.

The Unicode Character Set

- 16 bits per character. 2^{16}=65536 unique characters can be represented.
- The first 256 characters in the Unicode set correspond to those of the extended ASCII character set. ("*Backward compatibility*".)

Encryption

- Cryptography: The field of study related to encoded information
- Encryption: The process of converting plaintext into ciphertext
- Decryption: convert back
- Substitution cipher. Example: Caesar cipher
- Transposition cipher. Example: Route cipher
- Cryptanalysis. Frequency analysis

Public-Key Cryptography

- Encrypt with a *public key*
- *Decrypt* with a *private key*

Error-Detecting And Error-Correcting Codes

- *Reed–Solomon error correction* is an *error-correcting code* that works by oversampling a polynomial constructed from the data. The data symbols are represented by the coefficients of a polynomial. The polynomial is evaluated at several points, and these values are sent or recorded. By sampling the polynomial more often than is necessary, the polynomial is *over-determined*.
- Used in disk drives, CDs, telecommunication and digital broadcast protocols.