encoding
compression
encryption

- ASCII utf-8 utf-16
- zip mpeg jpeg
- AES RSA diffie-hellman

Expressing characters ...

ASCII and Unicode,
conventions of how characters
are expressed in bits.

ASCII (7 bits) - 128 characters
00 - 7F

ASCII Code Chart

	0	1	2	3	4	5	6	7	8	9	A	B	C	D	E	F
0	NUL	SOH	STX	ETX	EOT	ENQ	ACK	BEL	BS	HT	LF	VT	FF	CR	SO	SI
1	DLE	DC1	DC2	DC3	DC4	NAK	SYN	ETB	CAN	EM	SUB	ESC	FS	GS	RS	US
2		!	"	#	$	%	&	'	(	)	*	+	,	-	.	/
3	0	1	2	3	4	5	6	7	8	9	:	;	<	=	>	?
4	@	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O
5	P	Q	R	S	T	U	V	W	X	Y	Z	[	\	]	^	_
6	`	a	b	c	d	e	f	g	h	i	j	k	l	m	n	o
7	p	q	r	s	t	u	v	w	x	y	z	{	|	}	~	DEL

Expressing characters ...

ASCII and Unicode,
conventions of how characters
are expressed in bits.

ASCII (7 bits) - 128 characters
00 - 7F

ASCII Code Chart

	0	1	2	3	4	5	6	7	8	9	A	B	C	D	E	F
0	NUL	SOH	STX	ETX	EOT	ENQ	ACK	BEL	BS	HT	LF	VT	FF	CR	SO	SI
1	DLE	DC1	DC2	DC3	DC4	NAK	SYN	ETB	CAN	EM	SUB	ESC	FS	GS	RS	US
2		!	"	#	$	%	&	'	(	)	*	+	,	-	.	/
3	0	1	2	3	4	5	6	7	8	9	:	;	<	=	>	?
4	@	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O
5	P	Q	R	S	T	U	V	W	X	Y	Z	[	\	]	^	_
6	`	a	b	c	d	e	f	g	h	i	j	k	l	m	n	o
7	p	q	r	s	t	u	v	w	x	y	z	{	|	}	~	DEL

Expressing characters ...

ASCII and Unicode,
conventions of how characters
are expressed in bits.

ASCII (7 bits) - 128 characters
00 - 7F

Unicode designed to encode any language
more than 109,000 characters

e.g. Chinese, 20,902 ideogram characters

Room for expansion: 1,114,112 code points
in the range 0_{hex} to $10FFFF_{hex}$
various encodings UTF-8 UTF-16

Basic Multilingual Plane
0000 - FFFF

v · d · e	Unicode planes and code point (character) ranges					[hide]
Basic	**Supplementary**					
0000–FFFF	**10000–1FFFF**	**20000–2FFFF**	**30000–DFFFF**	**E0000–EFFFF**	**F0000–10FFFF**
Plane 0: Basic Multilingual Plane	**Plane 1:** Supplementary Multilingual Plane	**Plane 2:** Supplementary Ideographic Plane	**Planes 3–13:** Unassigned	**Plane 14:** Supplementary Special-purpose Plane	**Planes 15–16:** Private Use Area
BMP	**SMP**	**SIP**	**—**	**SSP**	**PUA**
0000–0FFF	8000–8FFF	10000–10FFF		20000–20FFF	28000–28FFF		E0000–E0FFF	*15: PUA-A*
1000–1FFF	9000–9FFF	11000–11FFF		21000–21FFF	29000–29FFF			F0000–FFFFF
2000–2FFF	A000–AFFF	12000–12FFF		22000–22FFF	2A000–2AFFF			
3000–3FFF	B000–BFFF	13000–13FFF	1B000–1BFFF	23000–23FFF	2B000–2BFFF			*16: PUA-B*
4000–4FFF	C000–CFFF			24000–24FFF				100000–10FFFF
5000–5FFF	D000–DFFF		1D000–1DFFF	25000–25FFF				
6000–6FFF	E000–EFFF	16000–16FFF		26000–26FFF				
7000–7FFF	F000–FFFF		1F000–1FFFF	27000–27FFF	2F000–2FFFF			

All code points in the BMP are accessed as a single code unit in UTF-16 encoding and can be encoded in one, two or three bytes in UTF-8. Code points in Planes 1 through 16 (**supplementary planes**, or, informally, **astral planes**) are accessed as surrogate pairs in UTF-16 and encoded in four bytes in UTF-8.

Code point	Binary code point	UTF-8 bytes	Example
U+0000 to U+007F	0xxxxxxx	0xxxxxxx	'$' U+0024 = 00100100 → 00100100 → 0x24
U+0080 to U+07FF	00000yyy yyxxxxxx	110yyyyy 10xxxxxx	'¢' U+00A2 = 00000000 10100010 → 11000010 10100010 → 0xC2 0xA2
U+0800 to U+FFFF	zzzzyyyy yyxxxxxx	1110zzzz 10yyyyyy 10xxxxxx	'€' U+20AC = 00100000 10101100 → 11100010 10000010 10101100 → 0xE2 0x82 0xAC
U+010000 to U+10FFFF	000wwwzz zzzzyyyy yyxxxxxx	11110www 10zzzzzz 10yyyyyy 10xxxxxx	'𤭢' U+024B62 = 00000010 01001011 01100010 → 11110000 10100100 10101101 10100010 → 0xF0 0xA4 0xAD 0xA2

UTF-8 : first 128 characters (US-ASCII) need one byte;

; next 1,920 characters need two bytes to encode.

Saturday, 3 December 2011

In UTF-8 :
first 128 characters (00-7F US-ASCII) need one byte;
next 1,920 characters (80-7FF) need two bytes to encode;
next (800-FFFF) each need two bytes to encode;
next (10000-10FFFF) each need four bytes.

Good for english and european texts - not so good for others.
Cyrillic and Greek alphabet pages in UTF-8 may be double the size, Thai and Devanagari, (Hindi) letters triple the size, compared with an encoding adapted to these character sets.

GB18030 is another encoding form for Unicode, from the Standardization Administration of China. It is the official character set of the People's Republic of China (PRC).
GB abbreviates Guójiā Biāozhǔn (国家标准), which means national standard in Chinese.

Huffman encoding (1952)

letter frequencies in English text

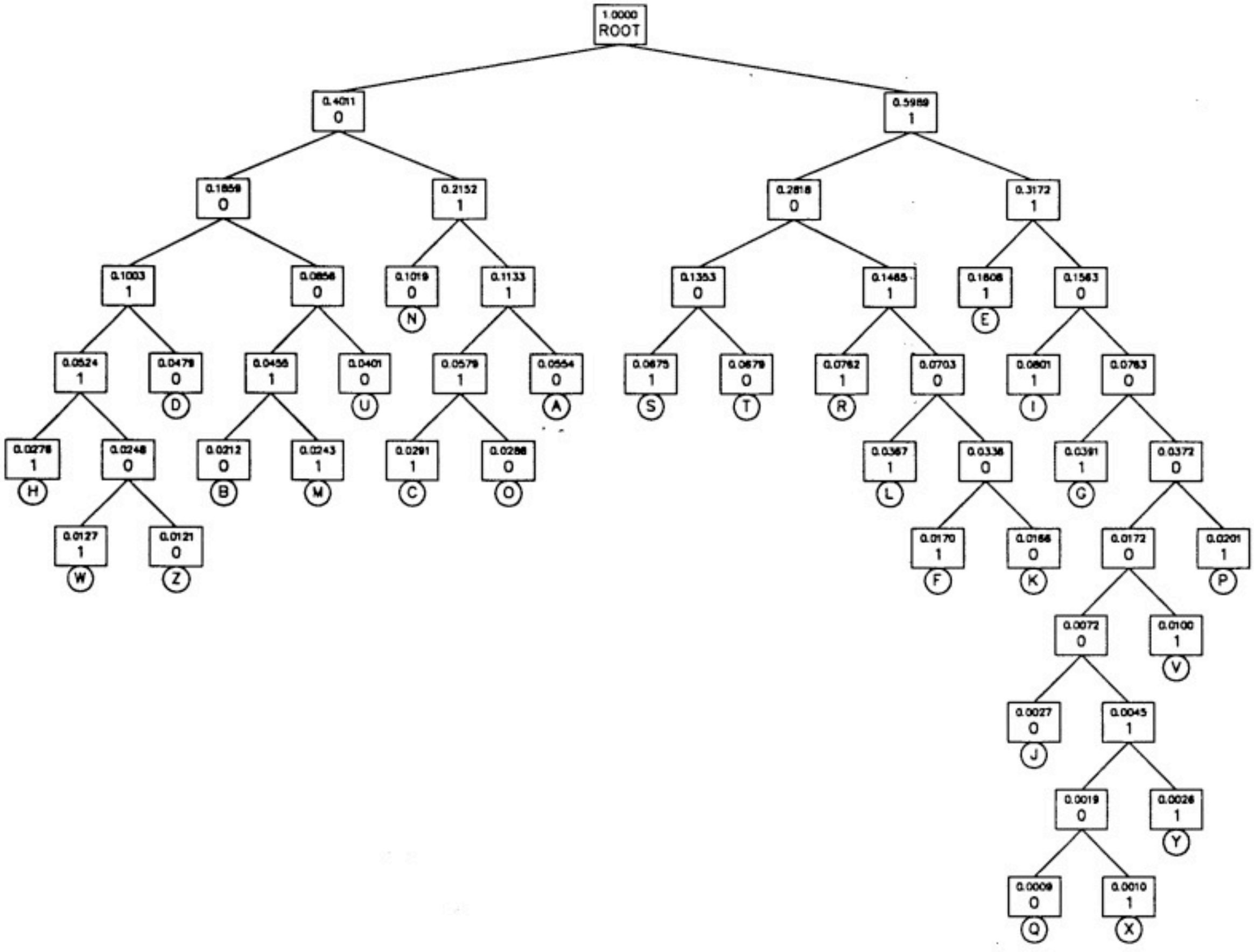
	n	p		n	p
A:	1501	0.05541	N:	2759	0.10185
B:	574	0.02119	O:	780	0.02880
C:	789	0.02913	P:	543	0.02005
D:	1298	0.04792	Q:	24	0.00089
E:	4356	0.16081	R:	2064	0.07620
F:	461	0.01702	S:	1827	0.06745
G:	1058	0.03906	T:	1838	0.06785
H:	748	0.02761	U:	1087	0.04013
I:	2169	0.08007	V:	271	0.01000
J:	73	0.00269	W:	343	0.01266
K:	450	0.01661	X:	27	0.00100
L:	993	0.03666	Y:	70	0.00258
M:	657	0.02425	Z:	328	0.01211

- Variable length encoding
  - use shorter codes for common letters

Saturday, 3 December 2011

Just as some characters are more frequent in some languages – and so different languages require different encodings to reduce the size of the encoded text – so different characters have different frequencies within a given language.

Can we use shorter codes for more frequent characters? What would such a code look like?



Saturday, 3 December 2011

This tree represents a Huffman encoding.

The 26 characters of the alphabet are at the leaves of the tree.

Each node, except the root node, is labelled, either 0 or 1.

Each non-leaf node has two children, one labelled 0, the other labelled 1.

Given a stream of bits, we can decode it as follows:
We start at the root and use successive bits from the stream to tell us which path to take through the tree, until we reach a leaf node. When we reach a leaf node, we write out the letter at that node and jump back to the root.

To encode a text, for each character, we just find the path from the root to the leaf labelled with that letter, and write out the sequence of bit-labels on that path.

The more-common letters are higher-up in the tree.

Lossless compression

- exploit statistical redundancy
- represent data concisely
- without error

- eg an html file has many occurrences of
  - <p>
- encode these with short sequences

Saturday, 3 December 2011

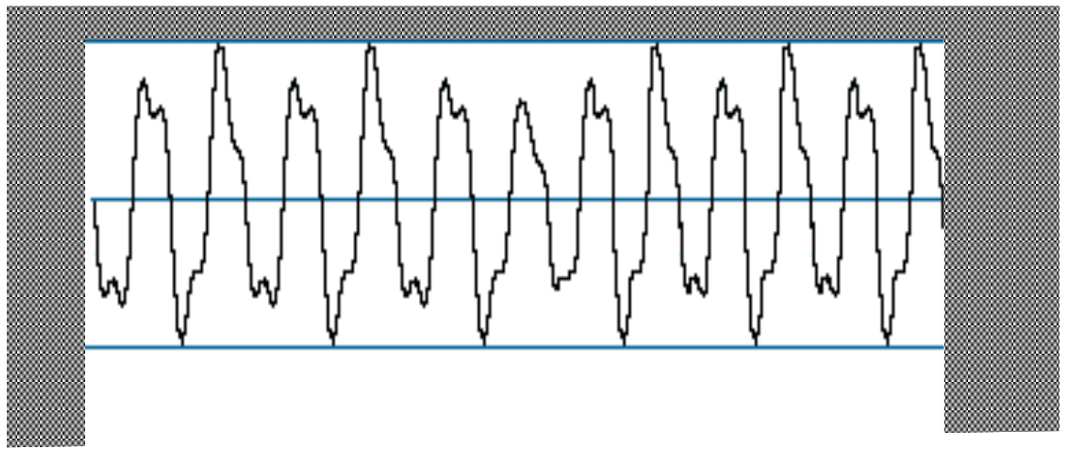
Huffman encoding is an example of lossless compression. We find a way to encode a message using fewer bits, that allows us to recreate the original message exactly.

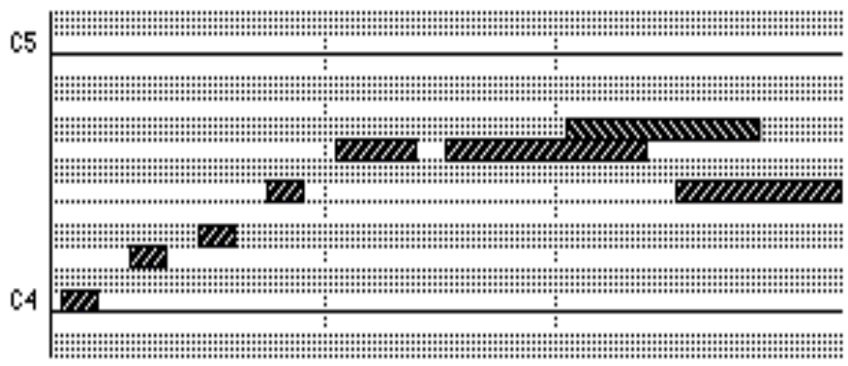
We can compute an optimal encoding for any text. Unless the text is very short, sending the encoding then the encoded text will be shorter than just sending the original.

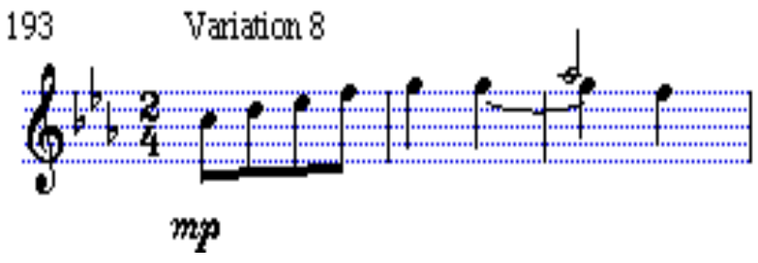
The same idea as for Huffman encoding can be used to encode common sequences of characters (eg common words in English, or particular patterns that are common in the file in question). This gives encodings such as zip and gzip used to compress files on the internet. This speeds up the web.

Representations of Music & Audio

- Audio (e.g., CD, MP3): like speech
- Time-stamped Events (e.g., MIDI file): like unformatted text
- Music Notation: like text with complex formatting



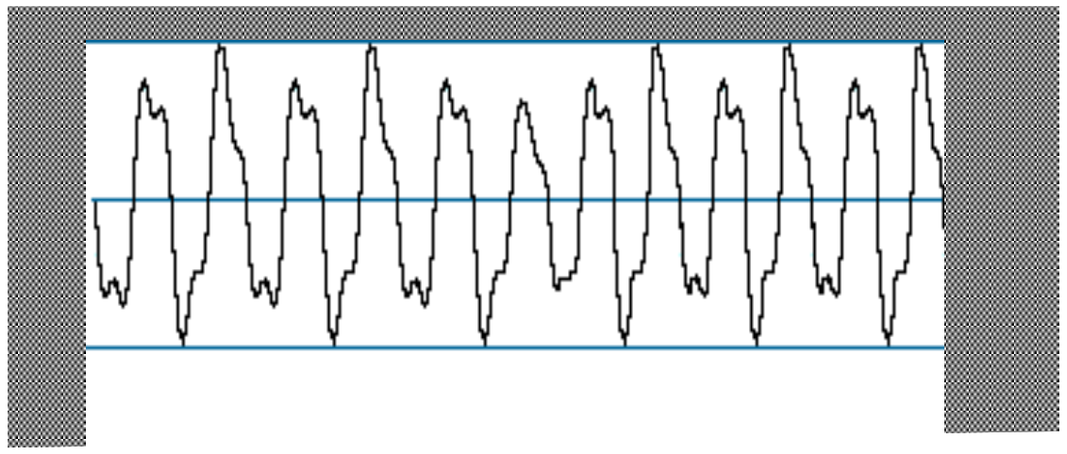


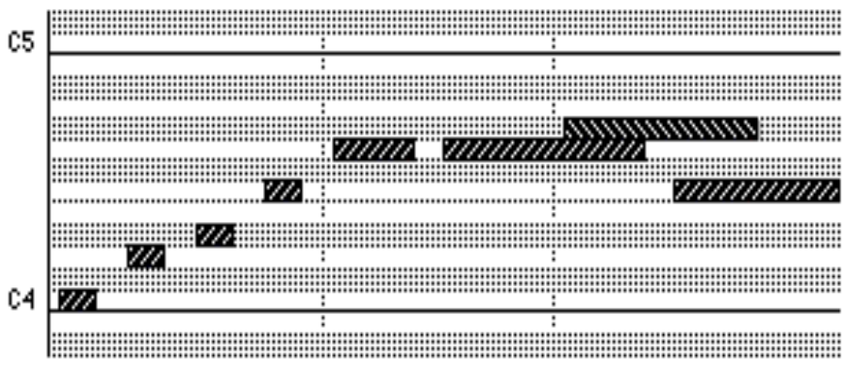


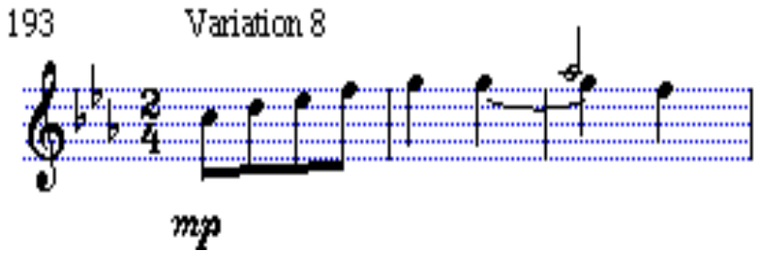
Multimedia files are often very large. They don't have the same kinds of repeated patterns that we see in text – so compression algorithms designed for text don't typically do much for music or pictures. A musician never plays the exactly the same note twice (and even if she did, random variations in the recording would introduce perhaps imperceptible differences).

MP3 up to 10:1

- perceptual audio encoding
- reconstruction *sounds like* the original
- knowledge from psychoacoustics







Saturday, 3 December 2011

On the other hand, for multimedia files, the details of the encoding may not be so important. We care what the music sounds like, or what a picture looks like. Imperceptible differences don't matter, and for some applications (eg speech) even perceptible differences don't matter provided we still get the message.

For example, telephones only transmit part of the speech signal. They are designed for communication. Listening to music down the telephone is an impoverished experience.

Even for music, there are well-researched effects that mean that some changes are imperceptible. For example, a loud sound 'masks' softer sounds at nearby frequencies. The ear can't hear whether they are there or not. So an encoding for music (such as MP3) can drop these softer sounds, imperceptibly.

Tricks such as this allow music to be compressed so it takes up less space on a memory stick and uses less bandwidth when transmitted over the internet.

Image Compression Formats

JPG or JPEG

GIF

TIF or TIFF

PNG

SVG

Saturday, 3 December 2011

There are many competing encodings for images.

Some (eg SVG) are descriptions of geometric objects, that can be rendered in many different ways.

Others are representations of the rendered form of a photograph or image.

Image Compression Formats

JPG or JPEG — Joint Photographic Expert Group

GIF

TIF or TIFF

PNG

SVG

Saturday, 3 December 2011

There are many competing encodings for images.

Some (eg SVG) are descriptions of geometric objects, that can be rendered in many different ways.

Others are representations of the rendered form of a photograph or image.

Image Compression Formats

JPG or JPEG — Joint Photographic Expert Group

GIF — Graphics Interchange Format

TIF or TIFF

PNG

SVG

Saturday, 3 December 2011

There are many competing encodings for images.

Some (eg SVG) are descriptions of geometric objects, that can be rendered in many different ways.

Others are representations of the rendered form of a photograph or image.

Image Compression Formats

JPG or JPEG — Joint Photographic Expert Group

GIF — Graphics Interchange Format

TIF or TIFF — Tagged Image File Format

PNG

SVG

There are many competing encodings for images.

Some (eg SVG) are descriptions of geometric objects, that can be rendered in many different ways.

Others are representations of the rendered form of a photograph or image.

Image Compression Formats

JPG or JPEG	Joint Photographic Expert Group
GIF	Graphics Interchange Format
TIF or TIFF	Tagged Image File Format
PNG	Portable Network Graphics
SVG	

Saturday, 3 December 2011

There are many competing encodings for images.

Some (eg SVG) are descriptions of geometric objects, that can be rendered in many different ways.

Others are representations of the rendered form of a photograph or image.

Image Compression Formats

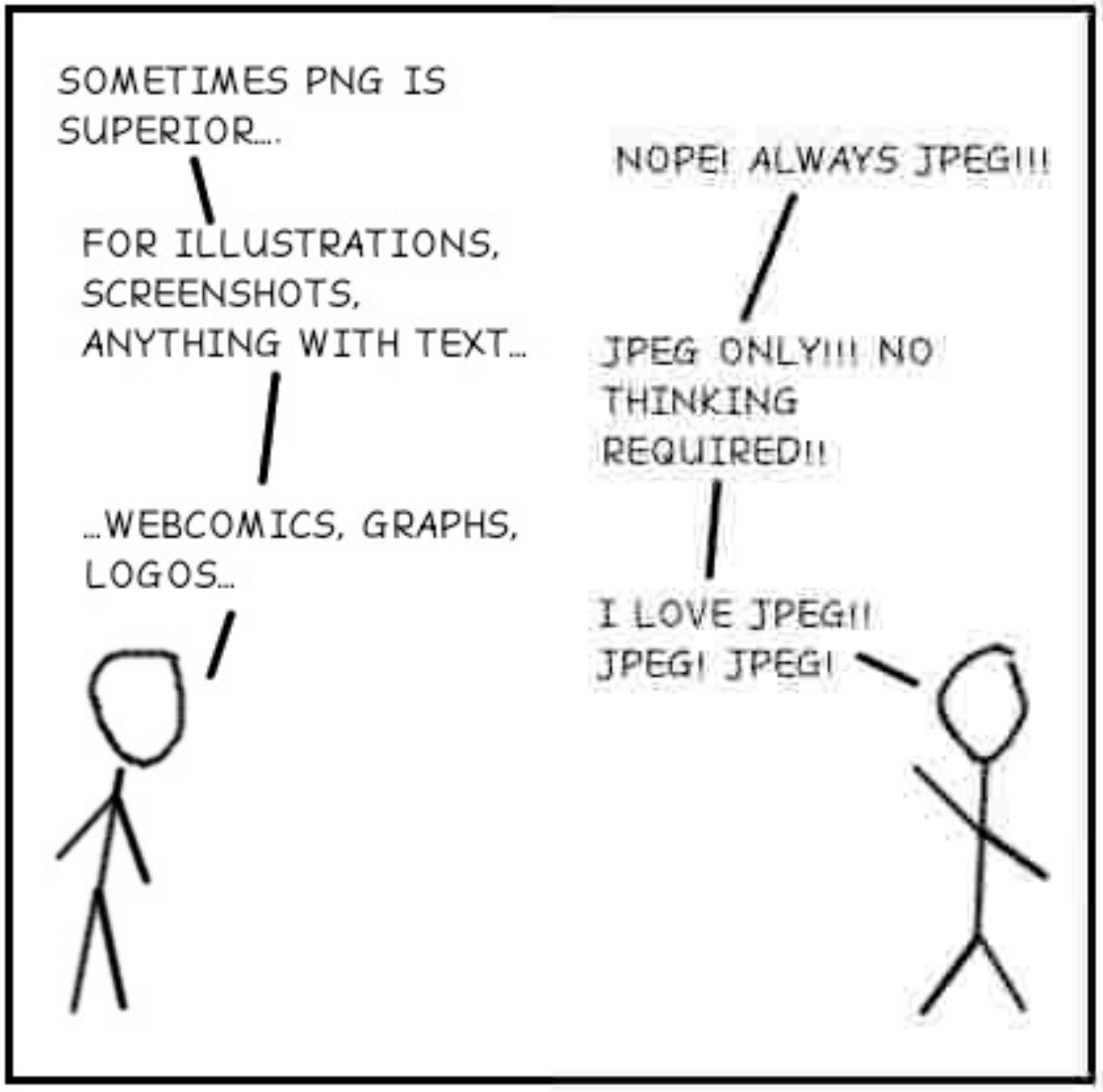
JPG or JPEG	Joint Photographic Expert Group
GIF	Graphics Interchange Format
TIF or TIFF	Tagged Image File Format
PNG	Portable Network Graphics
SVG	Scalable Vector Graphics

There are many competing encodings for images.

Some (eg SVG) are descriptions of geometric objects, that can be rendered in many different ways.

Others are representations of the rendered form of a photograph or image.

PNG vs JPEG



Saturday, 3 December 2011

JPG

RGB - 24 bits

Grayscale - 8 bits

JPG

RGB - 24 bits

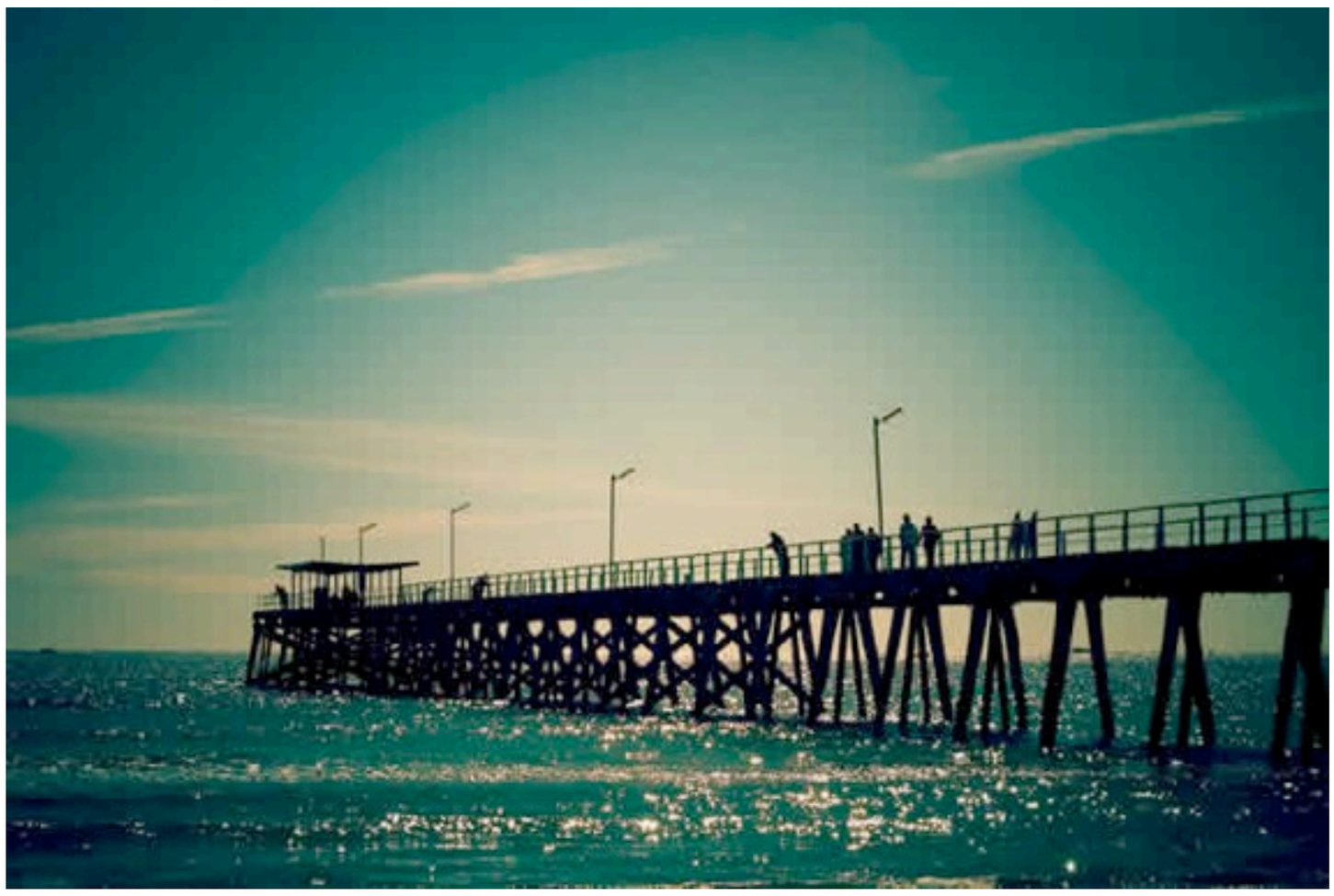
Grayscale - 8 bits

JPEG always uses **lossy** JPG compression, but the degree of compression can be chosen – for higher quality and larger files, or lower quality and smaller files.

JPG

RGB - 24 bits

Grayscale - 8 bits



JPEG always uses **lossy** JPG compression, but the degree of compression can be chosen – for higher quality and larger files, or lower quality and smaller files.

GIF

Indexed colour - 1 to 8 bits (2 to 256 colours)

GIF

Indexed colour - 1 to 8 bits (2 to 256 colours)

GIF uses **lossless** compression, effective on indexed colour. GIF files contain no dpi information for printing purposes.

GIF

Indexed colour - 1 to 8 bits (2 to 256 colours)

GIF uses **lossless** compression, effective on indexed colour. GIF files contain no dpi information for printing purposes.

GIF

Indexed colour - 1 to 8 bits (2 to 256 colours)

GIF uses **lossless** compression, effective on indexed colour. GIF files contain no dpi information for printing purposes.

TIF

RGB - 24 or 48 bits

Grayscale - 8 or 16 bits

Indexed colour - 1 to 8 bits

TIF

RGB - 24 or 48 bits

Grayscale - 8 or 16 bits

Indexed colour - 1 to 8 bits

For TIF files, most programs allow either no compression or LZW compression (lossless, but is less effective for 24 bit color images).

TIF

RGB - 24 or 48 bits

Grayscale - 8 or 16 bits

Indexed colour - 1 to 8 bits

For TIF files, most programs allow either no compression or LZW compression (lossless, but is less effective for 24 bit color images).

PNG

RGB - 24 or 48 bits

Grayscale - 8 or 16 bits

Indexed colour - 1 to 8 bits

PNG

RGB - 24 or 48 bits

Grayscale - 8 or 16 bits

Indexed colour - 1 to 8 bits

PNG uses ZIP compression which is lossless.

PNG

RGB - 24 or 48 bits

Grayscale - 8 or 16 bits

Indexed colour - 1 to 8 bits

PNG uses ZIP compression which is lossless.

PNG

RGB - 24 or 48 bits

Grayscale - 8 or 16 bits

Indexed colour - 1 to 8 bits

PNG uses ZIP compression which is lossless.

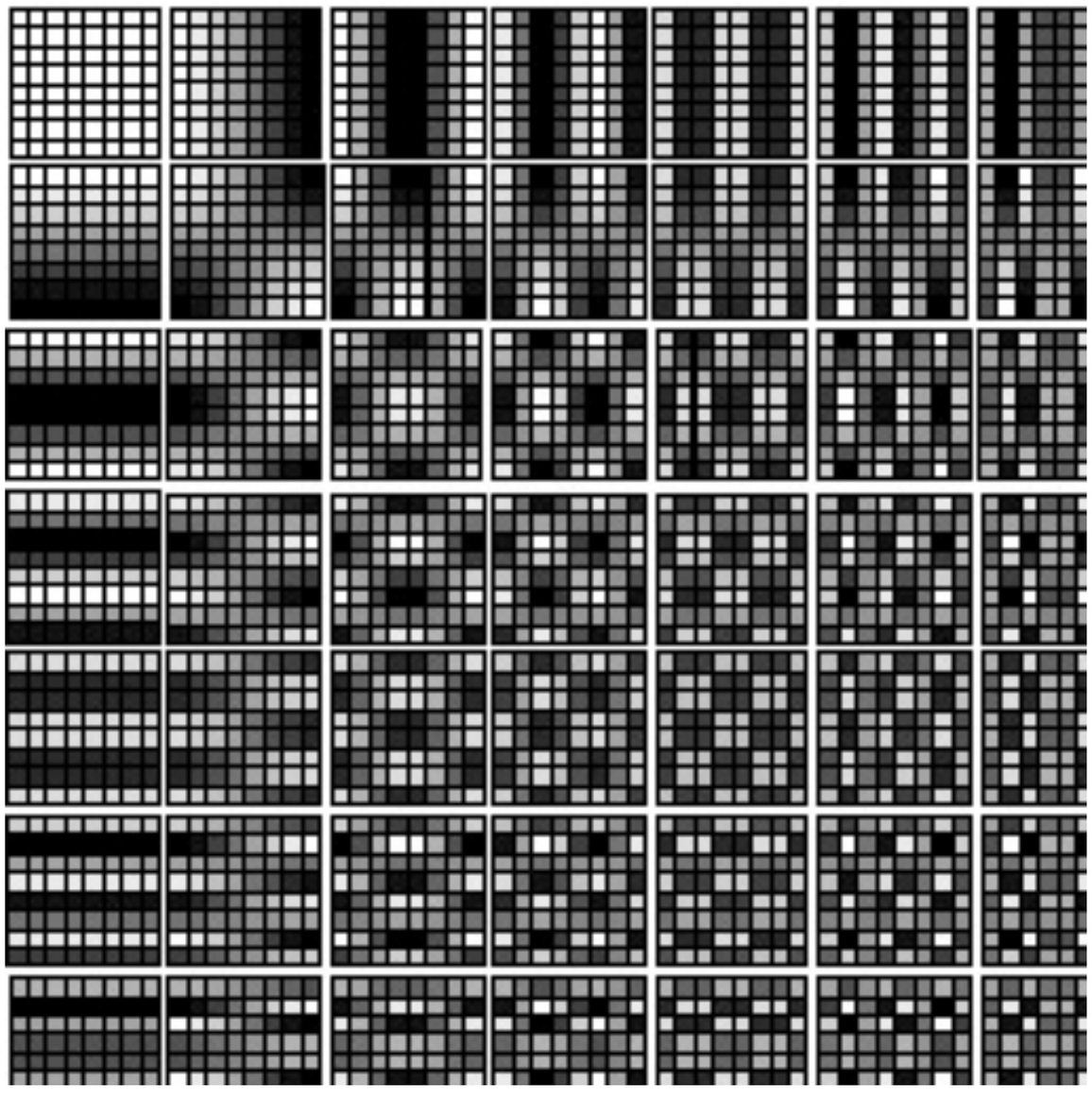
PNG was created to improve upon and replace GIF as an image-file format not requiring a patent license.

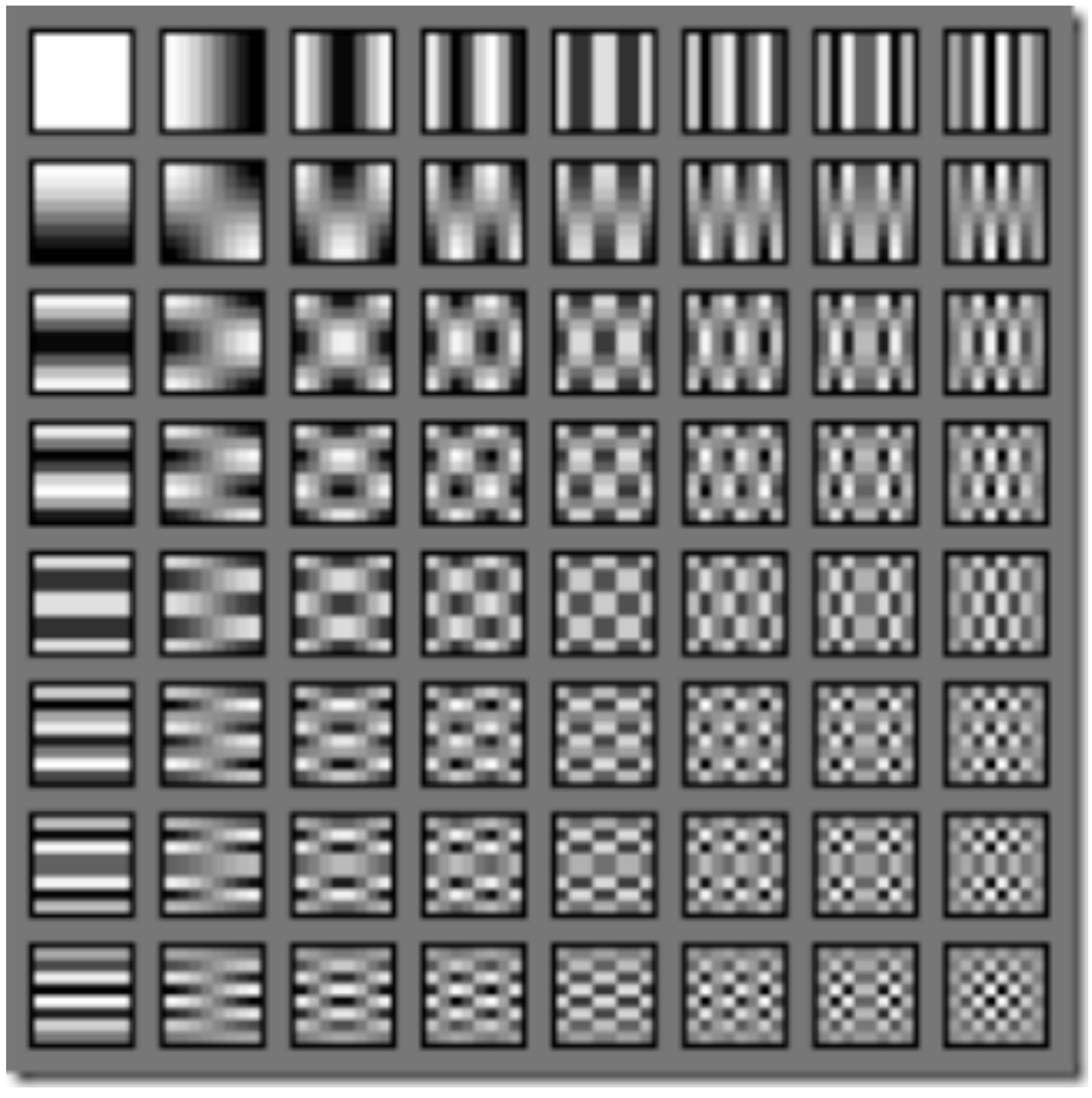
Lossy Compression

- In a lossy compression scheme, some of the original information is lost.
- It is impossible to produce an exact replica of the original signal when the audio or video is played.
- Lossy compression schemes add artefacts, small imperfections created by the loss of the actual data.

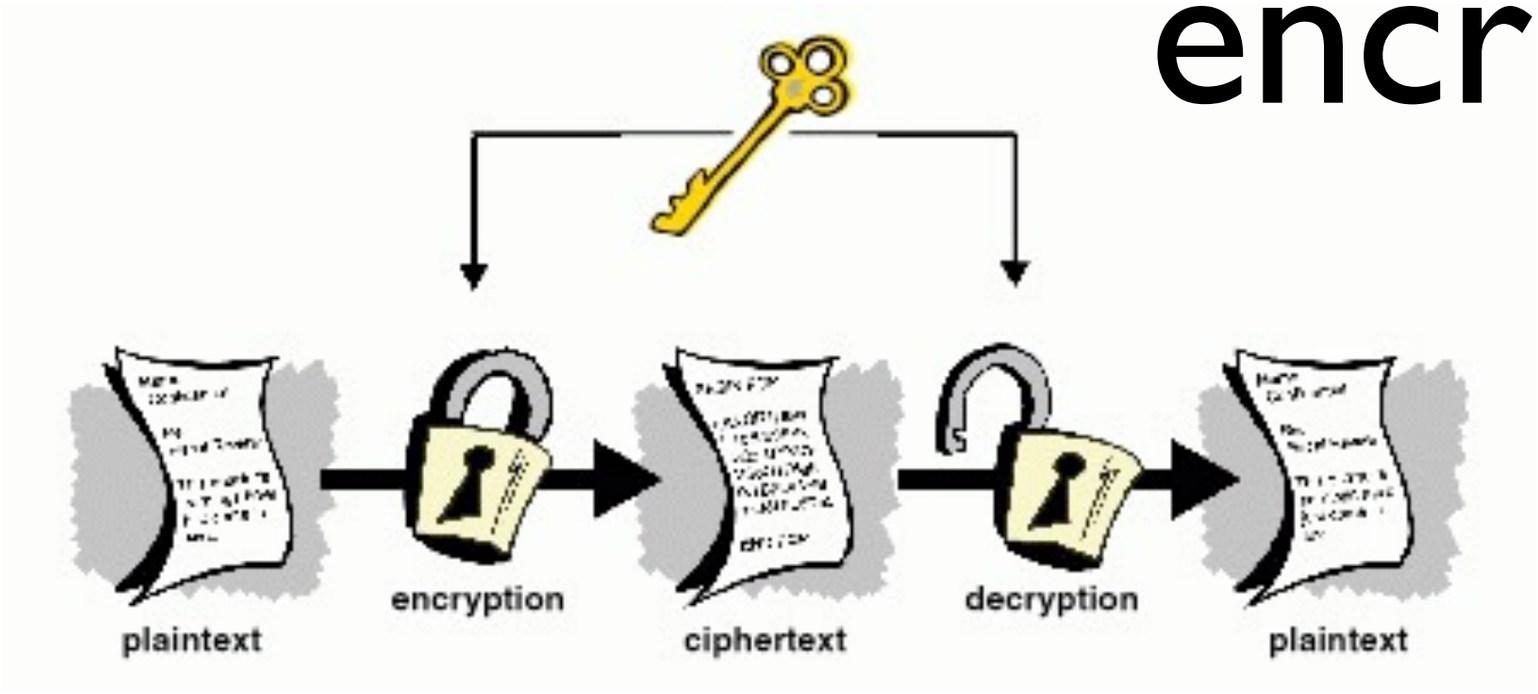
Lossy vs Lossless

Lossy





encryption



- shared key
- public key
- creating a shared secret

Keys are used to encrypt (lock) and decrypt (unlock) whatever data is being encrypted/ decrypted.

Symmetric-key algorithms use a single shared key; keeping data secret requires keeping this key secret.

Public-key algorithms use a public key and a private key. The public key is made available to anyone (often by means of a digital certificate). A sender encrypts data with the public key; only the holder of the private key can decrypt this data.

public key

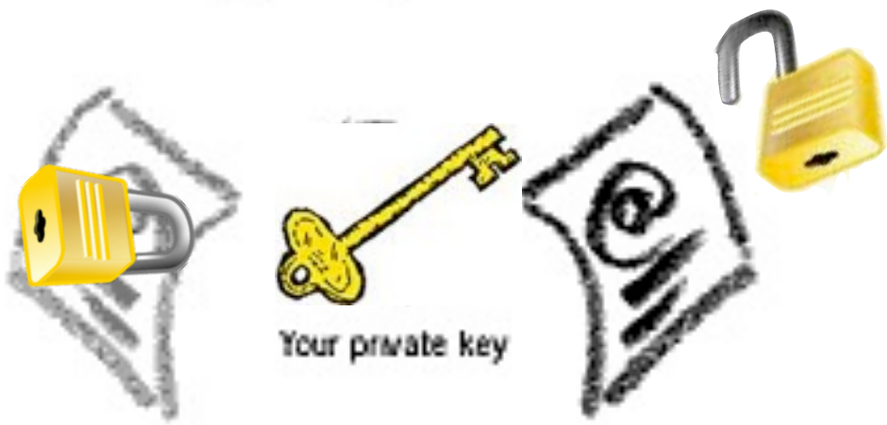
a key pair
lock (public key)
unlock (private key)

Step 1: Give your public key to the sender

Step 2: Sender uses your public key to encrypt the plaintext

Step 3: Sender gives the ciphertext to you

Step 4: Use your private key (and passphrase) to decrypt the ciphertext



making a shared secret

Alice makes up a secret: x
Bob makes up a secret: y

Alice sends Bob $A = g^x$
Bob sends Alice $B = g^y$

Bob calculates $A^y = g^{xy}$
Alice calculates $B^x = g^{xy}$

Saturday, 3 December 2011

Diffie–Hellman key exchange method allows two strangers (with no prior knowledge of each other) to jointly establish a shared secret key over an insecure communications channel

Two or more parties use a public exchange to agree on a shared secret they can use as a key without revealing the key to any eavesdropper.

The first publicly known key agreement protocol was this Diffie–Hellman exponential key exchange

Anonymous key exchange, like Diffie–Hellman, does not provide authentication of the parties, and is thus vulnerable to Man-in-the-middle attacks.

In practice the computation uses modular arithmetic to keep the sizes of numbers involved manageable.