Кодирование информации

Кодирование информации

Язык и алфавит

Язык

Язык — это система знаков, используемая для хранения, передачи и обработки информации.

Иероглифы:

Египетское письмо	
	рука
	дом
	кобра
	лев
	вода

Иероглифы (Китай)	
日	солнце
月	луна
雨	дождь
山	гора
马	лошадь

Алфавитное письмо

Алфавит — это набор знаков, который используется в языке.

Мощность алфавита — это количество знаков в алфавите.

АБВГДЕЁЖЗИЙКЛМНОПРСТУФХЦЧШЩЪЫЬЭЮЯ
0123456789 .,;?!-:…«»()

мощность 56

Слово — это последовательность символов алфавита, которая используется как самостоятельная единица и имеет определённое значение.

Сообщения

Сообщение — это любая последовательность символов некоторого алфавита.

Пример: алфавит @ # $ %.

Сообщения длины 1: @ # $ %.

всего 4

Сообщения длины 2:

@@ @# @$ @%
#@ ## #$ #%
$@ $# $$ $%
%@ %# %$ %%

всего 16

Сколько сообщений длины L ?

Количество возможных сообщений

Если алфавит языка состоит из *N символов (имеет мощность N),* количество различных сообщений длиной *L знаков равно*

$$Q = N^L$$

Сколько

- возможных 5-буквеных слов в русском языке? 33^5
- возможных 3-буквеных слов в английском языке? 26^3

Какие бывают языки?

•русский •английский •китайский •шведский •суахили •…	$y = 3\sin x + 1$ $2H_2 + O_2 = 2H_2O$  1. **e2-e4 e7-e5…**

Формальный язык – это язык, в котором однозначно определяется значение каждого слова, а также правила построения предложений и придания им смысла.

Естественные и формальные языки

Естественные	*Формальные*
•результат развития общества	• созданы людьми
•для общения в быту	• в специальных областях знаний
•значения слов зависят от контекста	• значения слов не зависит от контекста
•есть синонимы	• синонимов нет
•есть омонимы	• омонимов нет
•нет строгих правил образования предложений	• правила образования предложений строго определены
•есть исключения	• нет исключений

Кодирование информации

Кодирование

Что такое кодирование?

Кодирование — это представление информации в форме, удобной для её хранения, передачи и обработки. Правило такого преобразования называется **кодом**.

Текст:

- в России: ***Привет, Вася!***
- передача за рубеж (*транслит*): ***Privet, Vasya!***
- *Windows-1251*: *CFF0F8F2E52C20C2E0F1FF21*
- стенография: 
- шифрование: ***Рсйгжу-!Гбта”***

Числа:

- для вычислений: ***25***
- прописью: ***двадцать пять***
- римская система: ***XXV***

? Как зашифровано?

? Зачем?

Код Морзе

А	•–	О	–––	Э	••–••
Б	–•••	П	•––•	Ю	••––
В	•––	Р	•–•	Я	•–•–
Г	––•	С	•••		
Д	–••	Т	–	1	•————
Е	•	У	••–	2	••–––
Ж	•••–	Ф	••–•	3	•••––
З	––••	Х	••••	4	••••–
И	••	Ц	–•–•	5	•••••
Й	•–––	Ч	–––•	6	–••••
К	–•–	Ш	————	7	––•••
Л	•–••	Щ	––•–	8	–––••
М	––	Ь	–••–	9	————•
Н	–•	Ы	–•––	0	—————

Самюэль Морзе
(1791–1872)

! Код неравномерный, нужен разделитель!

•–– •– ••• •–•– **ВАСЯ**

•––•– **ВА, АК, ПТ, ЕМЕТ?**

Двоичное кодирование

Двоичное кодирование — это кодирование с помощью двух знаков.

Равномерный код:

А	Б	В	Г
00	01	10	11

АБАВГБ → 000100101101

Количество сообщений длиной **I** битов: $N = 2^I$

Пример. Нужно закодировать номер спортсмена от 1 до 200. Сколько битов потребуется?

$2^7 < 200 \le 2^8 = 256$

8 битов

Декодирование

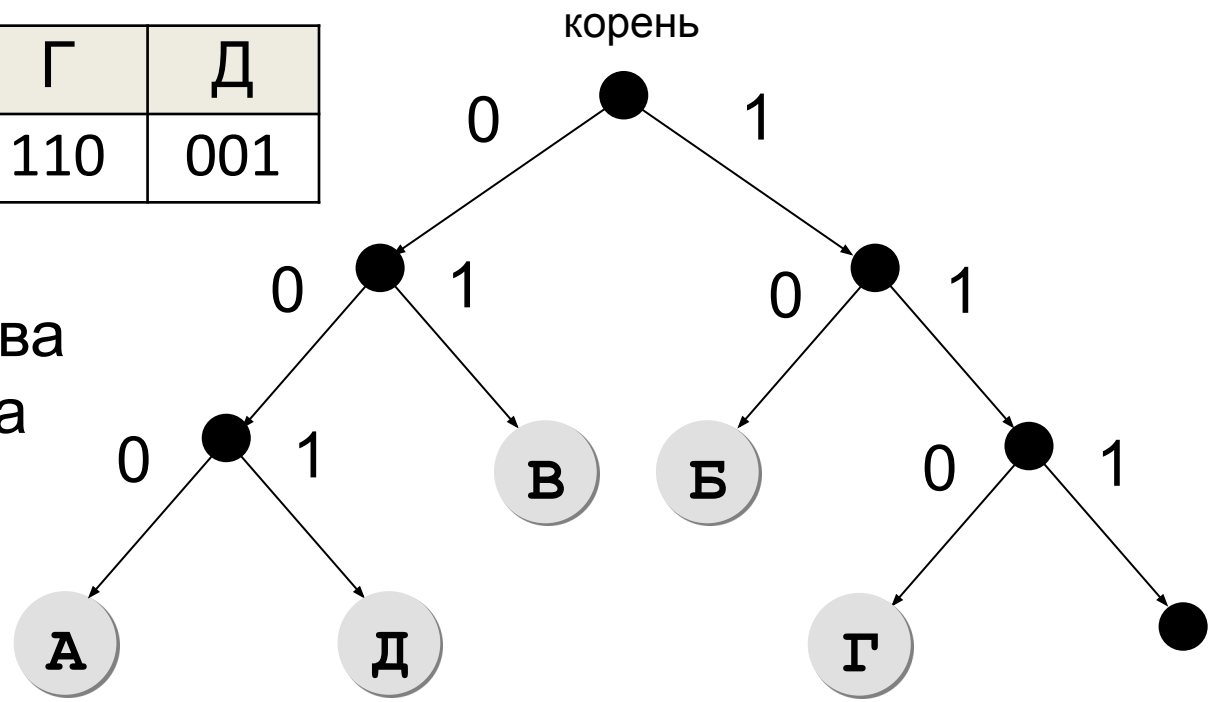
Декодирование — это восстановление сообщения из последовательности кодов.

•– – •– ••• •–•– **ВАСЯ**

? Когда разделитель не нужен?

А	Б	В	Г	Д
000	10	01	110	001

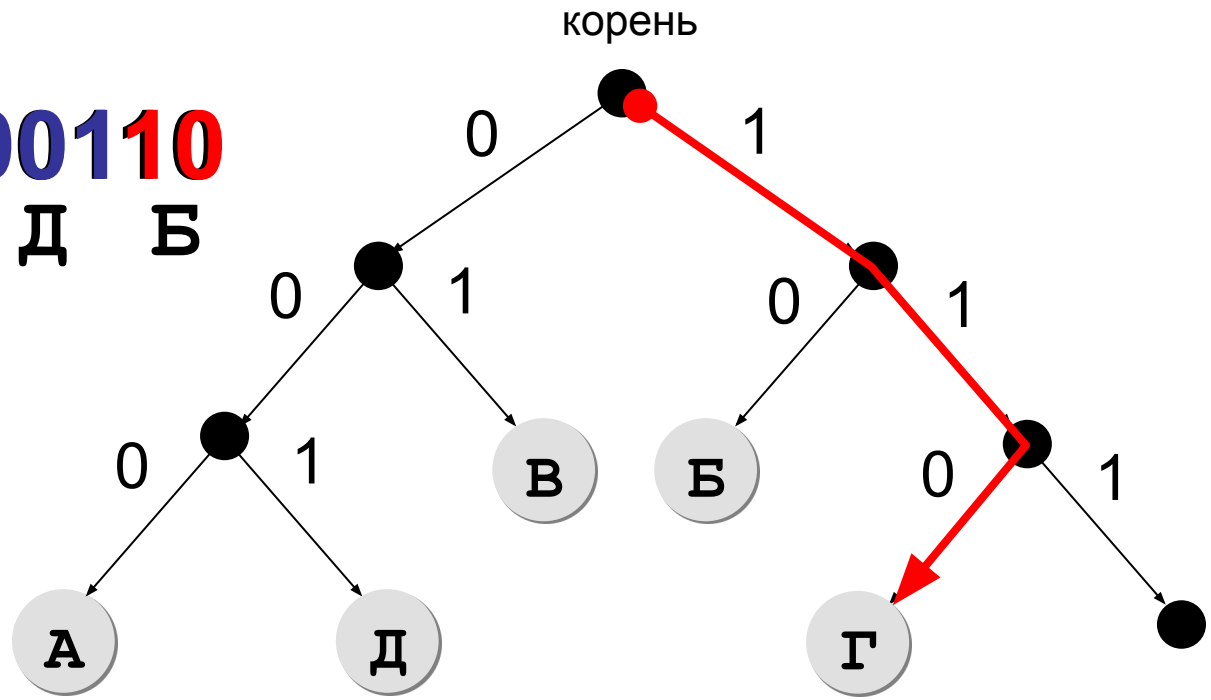
Все кодовые слова заканчиваются на листьях дерева!



Декодирование

1100000100110

Г А В Д Б



Префиксный код — это код, в котором ни одно кодовое слово не совпадает с началом другого кодового слова (*условие Фано*). Сообщения декодируются однозначно.

Постфиксные коды

Постфиксный код — это код, в котором ни одно кодовое слово не совпадает с **окончанием** другого кодового слова. Сообщения декодируются однозначно (**с конца**!).

А	Б	В	Г	Д
000	01	10	011	100

0110001101110

Б Д 1 Г Б В

Неоднозначное декодирование

А	Б	В	Г	Д
01	010	011	11	101

? Выполняются ли условия Фано?

Декодирование *может быть* неоднозначным…

010100111101 → **АБАГД**

010100111101 → **АБВГ**

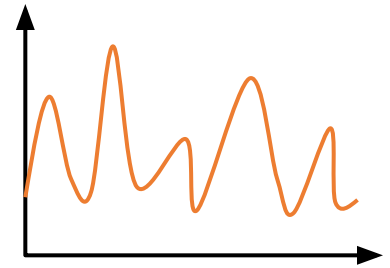
! Может быть, что условия Фано не выполнены, а декодирование однозначно (см. учебник)!

Кодирование информации

Дискретность

Аналоговые сигналы и устройства

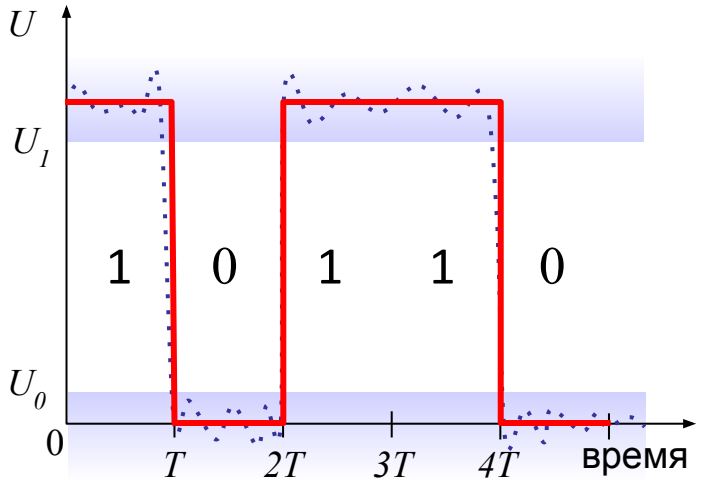
Аналоговый сигнал — это сигнал, который в любой момент времени может принимать любые значения в заданном диапазоне.



Аналоговые компьютеры

- невозможно «очистить» сигнал от помех
- при измерении сигнала вносится ошибка
- при копировании аналоговая информация искажается

Дискретные (цифровые) сигналы



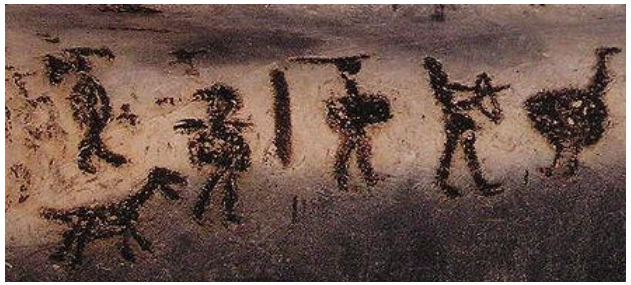
Свойства:

- сигнал изменяется только в отдельные моменты времени (*дискретность по времени*);
- принимают только несколько возможных значений (*дискретность по уровню*).

Дискретный сигнал — это последовательность значений, каждое из которых принадлежит некоторому конечному множеству.

Дискретность

Цель – максимально точно передавать сообщения при сильных помехах.



Pacta sunt servanda.

•– – •– ••• •–•–

01000011001

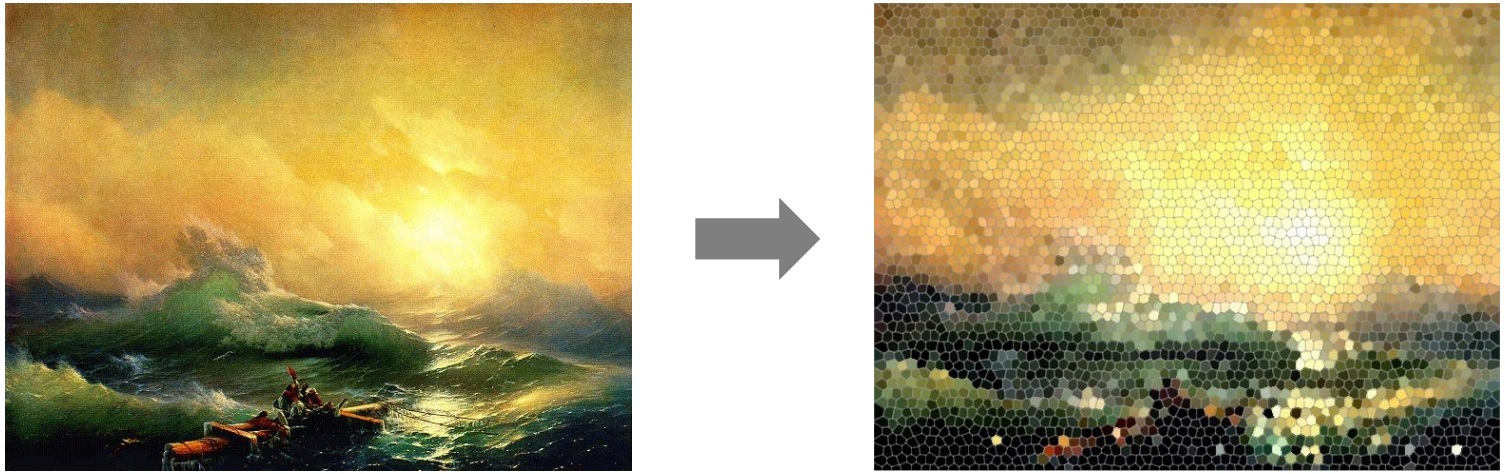
! Компьютеры могут хранить и обрабатывать только дискретную информацию!

… закодированную с помощью конечного количества знаков некоторого алфавита.

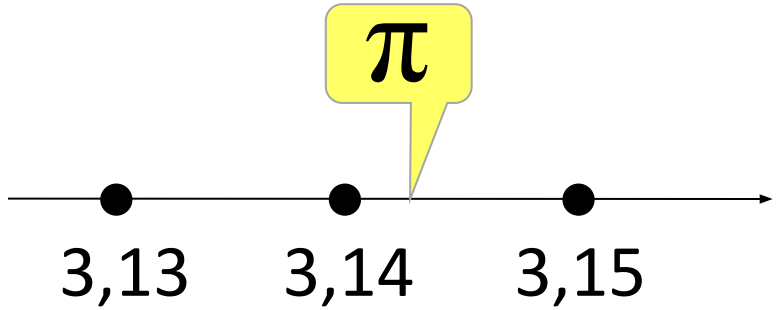
! Все виды информации нужно перевести в дискретный вид!

Дискретизация

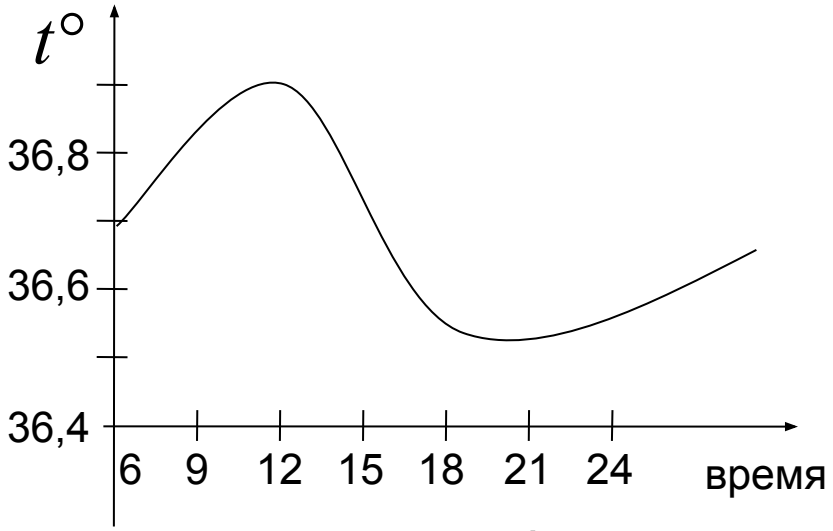
Дискретизация — это представление единого объекта в виде множества отдельных элементов.



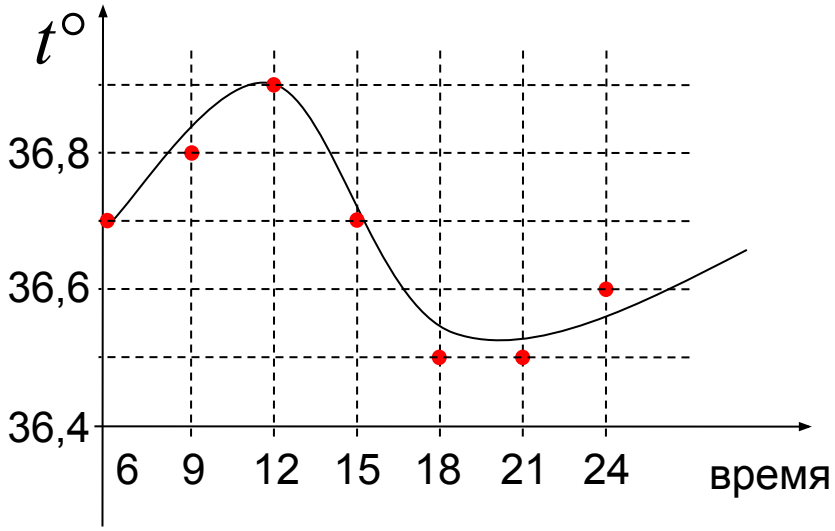
π



Дискретизация



аналоговая информация



дискретизация

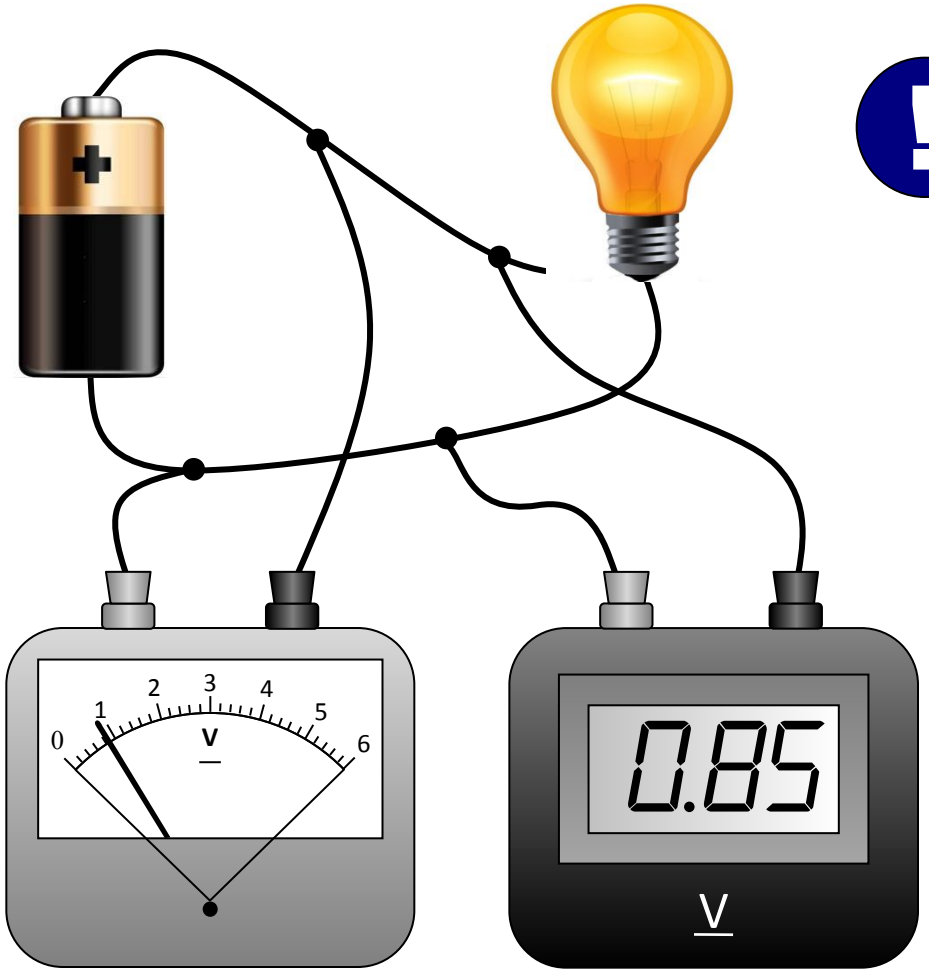
6 ч. 36,7°
9 ч. 36,8°
12 ч. 36,9°
15 ч. 36,7°
18 ч. 36,5°
21 ч. 36,5°
24 ч.

дискретная информация

! При дискретизации есть потеря информации!

? Как уменьшить потери?

Непрерывность и дискретность



аналоговые данные

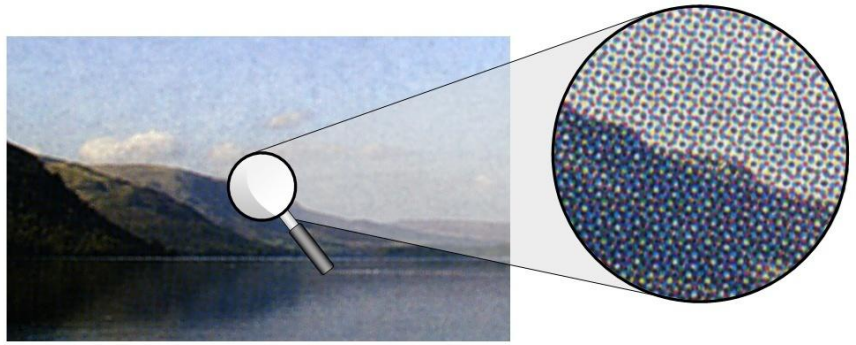
дискретные данные

Дискретность — это свойство не информации, а её представления.

Непрерывность и дискретность

! При увеличении точности дискретизации свойства аналоговой и дискретной информации практически совпадают!

$$\pi \approx 3{,}1415926$$



Кодирование информации

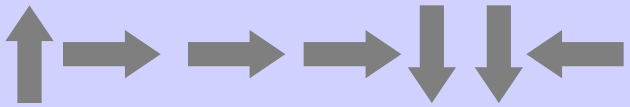
Алфавитный подход к измерению количества информации

Алфавитный подход

Количество информации в битах определяется длиной сообщения в двоичном коде.

10101100 — 8 битов

вперёд	00
назад	01
вправо	10
влево	1
	1



? Сколько битов?

00101010010111 — 14 битов

Алфавитный подход

1) определяем мощность алфавита N;
2) определяем количество битов информации i, приходящихся на один символ, — информационную ёмкость (объём) символа:

N, символов	2	4	8	16	32	64	128	256	512	1024
i, битов информации	1	2	3	4	5	6	7	8	9	10

1) количество информации в сообщении:

$$I = L \cdot i$$

где L – количество символов в сообщении.

Алфавитный подход

- каждый символ несёт одинаковое количество информации
- частота появления разных символов (и сочетаний символов) не учитывается
- количество информации определяется только длиной сообщения и мощностью алфавита
- смысл сообщения не учитывается

Кодирование символов

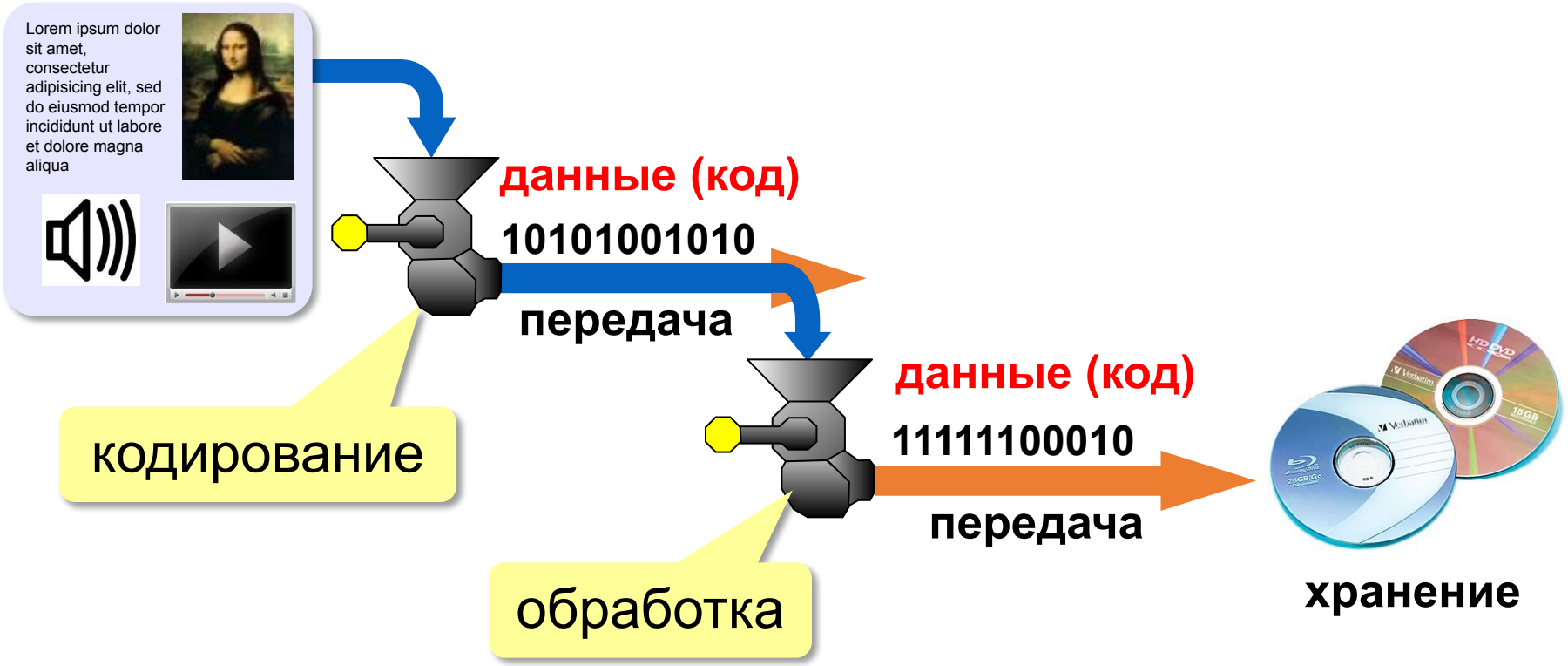
Кодирование графической информации

Кодирование звуковой и видеоинформации

Зачем кодировать информацию?

Кодирование — это представление информации в форме, удобной для её хранения, передачи и обработки.

В компьютерах используется двоичный код:

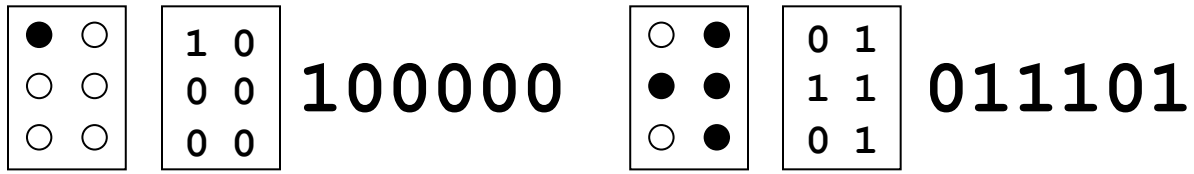


Кодирование информации

Кодирование символов

Кодирование символов

Система Брайля:



Общий подход:

- нужно использовать N символов
- выберем число битов k на символ: $2^k \geq N$
- сопоставим каждому символу код – число от 0 до $2^k - 1$
- переведем коды в двоичную систему

Откуда формула?

Кодирование символов

Текстовый файл

- на экране (символы)
- в памяти – коды

1000001_2	1000010_2	1000011_2	1000100_2
65	66	67	68

В файле хранятся не изображения символов, а их числовые коды!

Файлы со шрифтами: ***.fon**, ***.ttf**, ***.otf**

Кодировка ASCII (7-битная)

ASCII = *American Standard Code for Information Interchange*

Коды 0-127:

0-31 **управляющие символы**:
7 – звонок, 10 – новая строка,
13 – возврат каретки, 27 – Esc.

32 пробел

знаки препинания: `. , : ; ! ?`

специальные знаки: `+ - * / () {} []`

48-57 цифры **0..9**

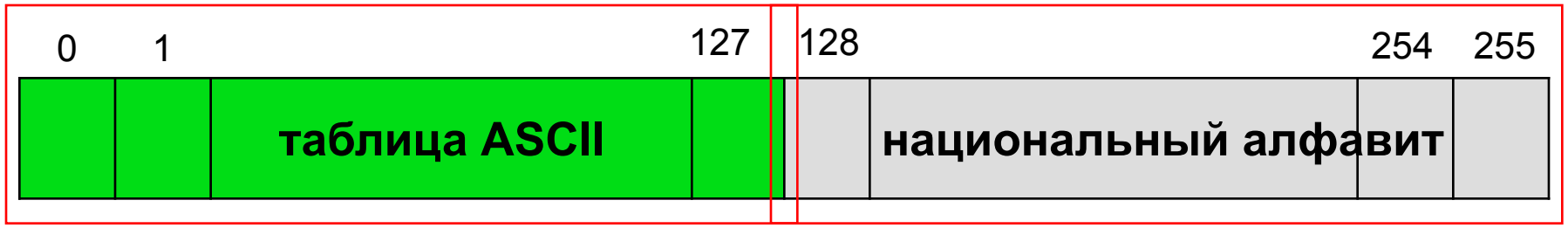
65-90 заглавные латинские буквы **A-Z**

97-122 строчные латинские буквы **a-z**

Где русские буквы?

8-битные кодировки

Кодовые страницы (расширения ASCII):



Для русского языка:

CP-866	для *MS DOS*
CP-1251	для *Windows* (Интернет)
KOI8-R	для *UNIX* (Интернет)
MacCyrillic	для компьютеров *Apple*

Проблема:

Windows-1251	*KOI8-R*
Привет, Вася!	оПХБЕР, бЮЯЪ!
рТЙЧЕФ, чБУС!	Привет, Вася!

8-битные кодировки

- 1 байт на символ – файлы небольшого размера!
- просто обрабатывать в программах

- нельзя использовать символы разных кодовых страниц одновременно (русские и французские буквы, и т.п.)
- неясно, в какой кодировке текст (перебор вариантов!)
- для каждой кодировки нужен свой шрифт (изображения символов)

Стандарт UNICODE

1 112 064 знаков, используются около **100 000**

Windows: **UTF-16**

16 битов на распространённые символы,
32 бита на редко встречающиеся

Linux: **UTF-8**

8 битов на символ для ASCII,
от 16 до 48 бита на остальные

- совместимость с ASCII
- более экономична, чем UTF-16, если много символов ASCII

2010 г. – 50% сайтов использовали UTF-8!

Кодирование графической информации

Растровое кодирование

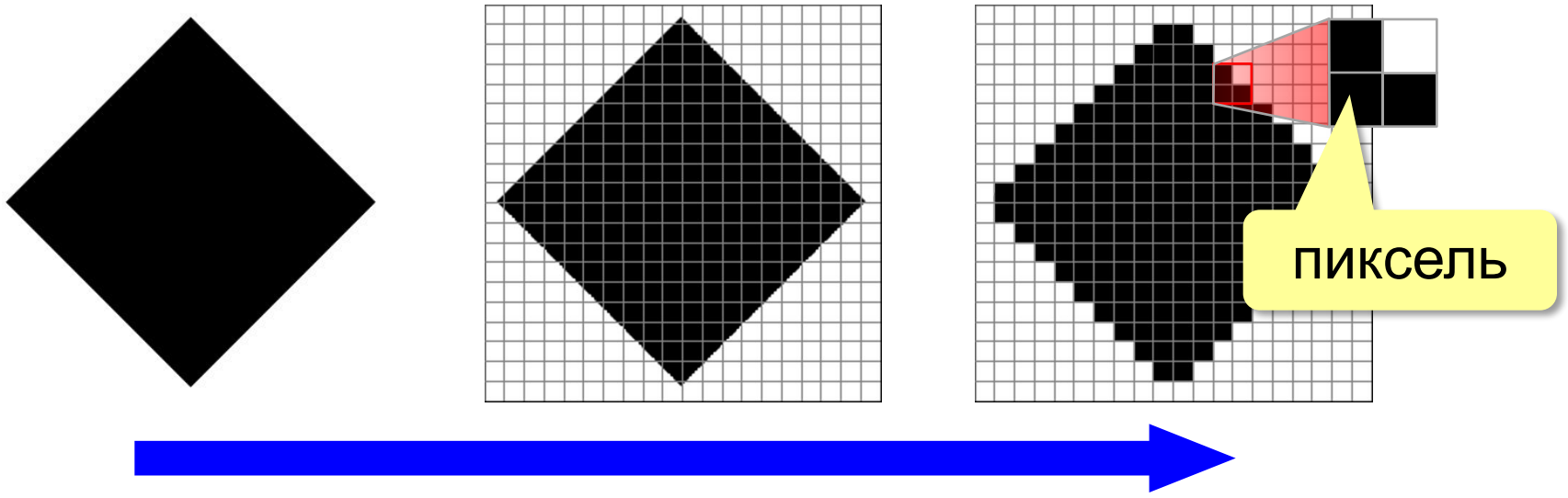
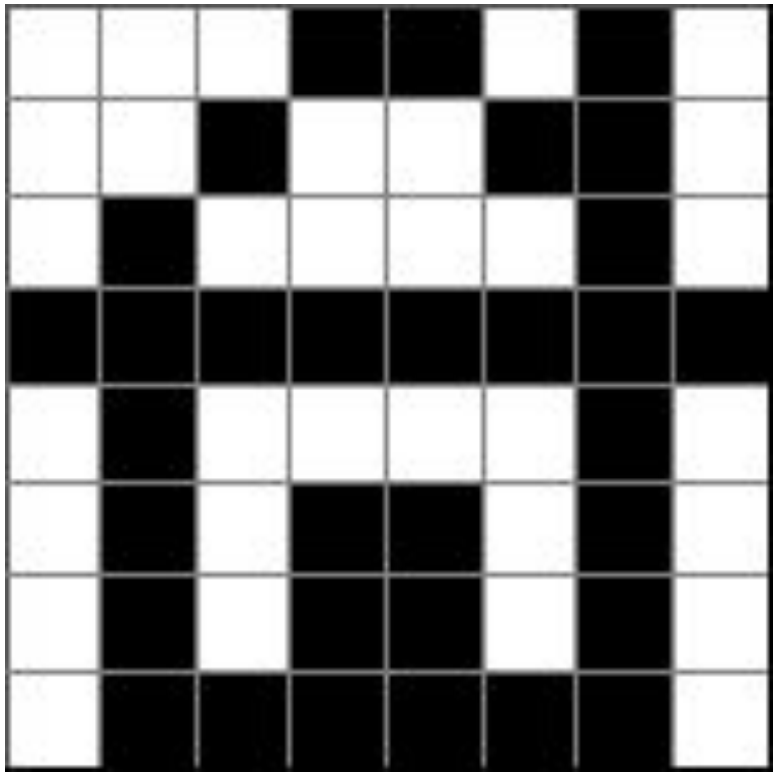


Рисунок искажается!

Пиксель – это наименьший элемент рисунка, для которого можно задать свой цвет.

Растровое изображение – это изображение, которое кодируется как множество пикселей.

Растровое кодирование



0	0	0	1	1	0	1	0	1A
0	0	1	0	0	1	1	0	26
0	1	0	0	0	0	1	0	42
1	1	1	1	1	1	1	1	FF
0	1	0	0	0	0	1	0	42
0	1	0	1	1	0	1	0	5A
0	1	0	1	1	0	1	0	5A
0	1	1	1	1	1	1	0	7E

$1A2642FF425A5A7E_{16}$

Разрешение

Разрешение – это количество пикселей, приходящихся на дюйм размера изображения.

ppi = *pixels per inch*, пикселей на дюйм

1 дюйм = 2,54 см

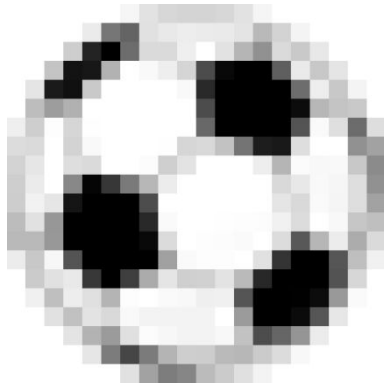
300 ppi

печать

96 ppi

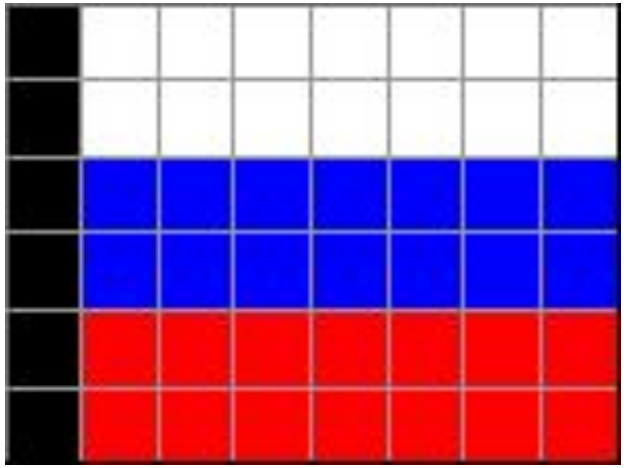
экран

48 ppi



24 ppi

Кодирование цвета

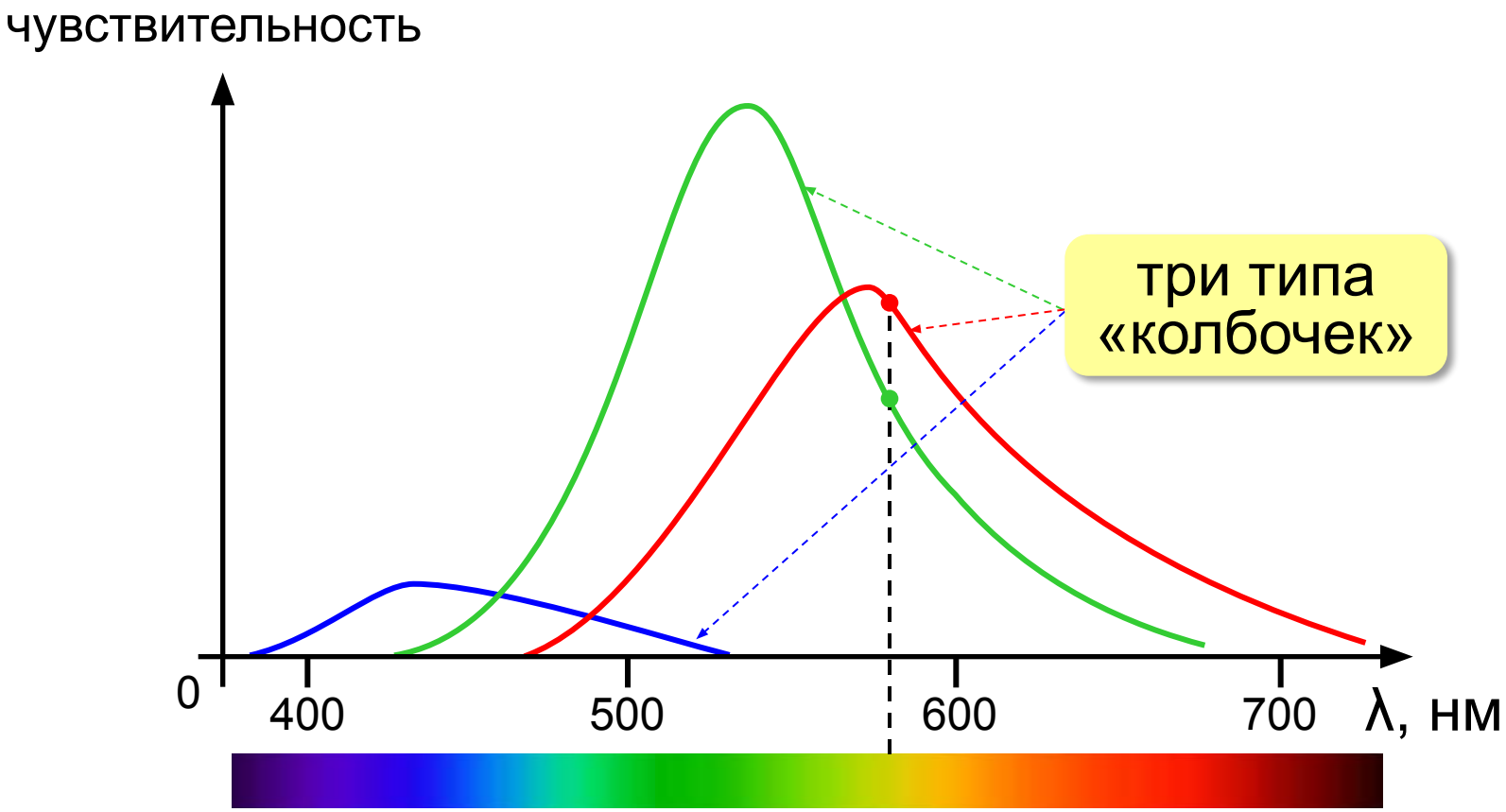


00	11	11	11	11	11	11	11
00	11	11	11	11	11	11	11
00	01	01	01	01	01	01	01
00	01	01	01	01	01	01	01
00	10	10	10	10	10	10	10
00	10	10	10	10	10	10	10

Как выводить на монитор цвет с кодом 00?

Как закодировать цвет в виде чисел?

Теория цвета Юнга-Гельмгольца

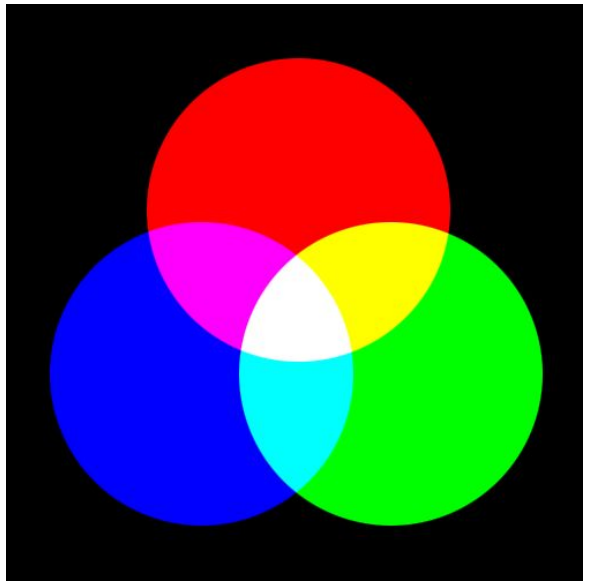


Свет любой длины волны можно заменить на красный, зелёный и синий лучи!

Цветовая модель RGB

Д. Максвелл, 1860 **цвет = (R, G, B)**

red красный 0..255 ***green*** зеленый 0..255 ***blue*** синий 0..255



(0, 0, 0)
(255, 255, 255)
(255, 0, 0)
(255, 150, 150)
(0, 255, 0)
(255, 255, 0)
(0, 0, 255)
(100, 0, 0)

? Сколько разных цветов можно кодировать?

256·256·256 = **16 777 216** (*True Color,* «истинный цвет»)

! RGB – цветовая модель для устройств, излучающих свет (мониторов)!

Цветовая модель RGB

(255, 255, 0) → #FFFF00

	RGB	Веб-страница
	(0, 0, 0)	#000000
	(255,255,255)	#FFFFFF
	(255, 0, 0)	#FF0000
	(0, 255, 0)	#00FF00
	(0, 0, 255)	#0000FF
	(255, 255, 0)	#FFFF00
	(204,204,204)	#CCCCCC

Глубина цвета

Глубина цвета — это количество битов, используемое для кодирования цвета пикселя.

Сколько памяти нужно для хранения цвета 1 пикселя в режиме *True Color*?

R (0..255) 256 = 2^8 вариантов 8 битов = 1 байт

R G B: 24 бита = 3 байта

True Color (истинный цвет)

Задача. Определите размер файла, в котором закодирован растровый рисунок размером 20×30 пикселей в режиме истинного цвета (*True Color*)?

20 · 30 · 3 байта = **1800 байт**

Кодирование с палитрой

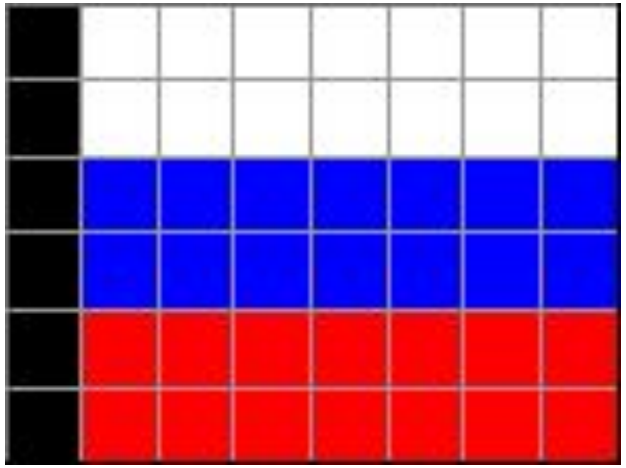
Как уменьшить размер файла?

- уменьшить разрешение
- уменьшить глубину цвета

снижается качество

Цветовая палитра – это таблица, в которой каждому цвету, заданному в виде составляющих в модели RGB, сопоставляется числовой код.

Кодирование с палитрой



00	11	11	11	11	11	11	11
00	11	11	11	11	11	11	11
00	01	01	01	01	01	01	01
00	01	01	01	01	01	01	01
00	10	10	10	10	10	10	10
00	10	10	10	10	10	10	10

Палитра:

0	0	0	0	0	255	255	0	0	255	255	255
цвет 00_2			цвет 01_2			цвет 10_2			цвет 11_2		

Какая глубина цвета?

2 бита на пиксель

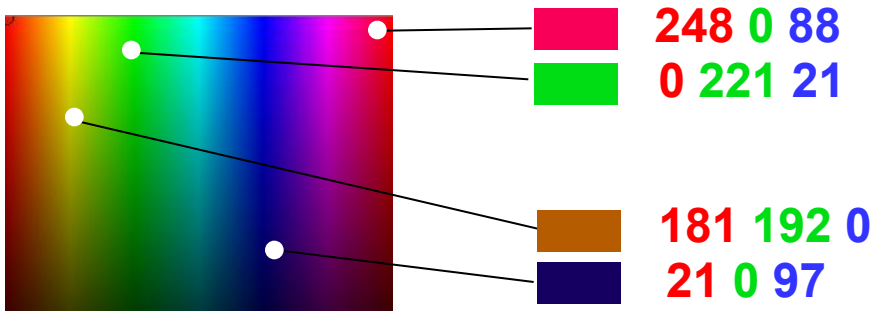
Сколько занимает палитра?

$3 \cdot 4 = 12$ байтов

Кодирование с палитрой

Шаг 1. Выбрать количество цветов: 2, 4, … **256.**

Шаг 2. Выбрать 256 цветов из палитры:



Шаг 3. Составить палитру (каждому цвету – номер 0..255)

палитра хранится в начале файла

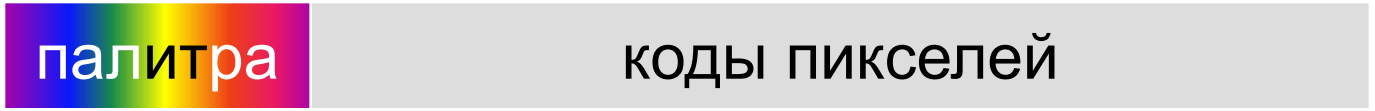
0	1		254	255
248 0 88	0 221 21	…	181 192 0	21 0 97

Шаг 4. Код пикселя = номеру его цвета в палитре

2	45	65	14	…	12	23

Кодирование с палитрой

Файл с палитрой:

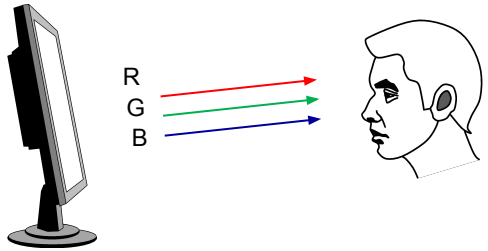


Количество цветов	Размер палитры (байтов)	Глубина цвета (битов на пиксель)
2	6	1
4	12	2
16	48	4
256	768	8

Растровые рисунки: форматы файлов

Формат	True Color	Палитра	Прозрачность	Анимация
BMP	✔	✔	⛔	⛔
JPG	✔	⛔	⛔	⛔
GIF	⛔	✔	✔	✔
PNG	✔	✔	✔	⛔

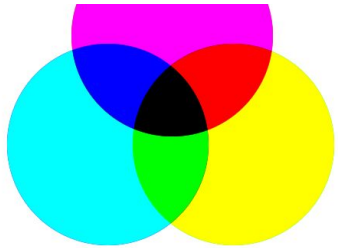
Кодирование цвета при печати (CMYK)



Белый – красный = голубой **C = Cyan**

Белый – зелёный = пурпурный **M = Magenta**

Белый – синий = желтый **Y = Yellow**



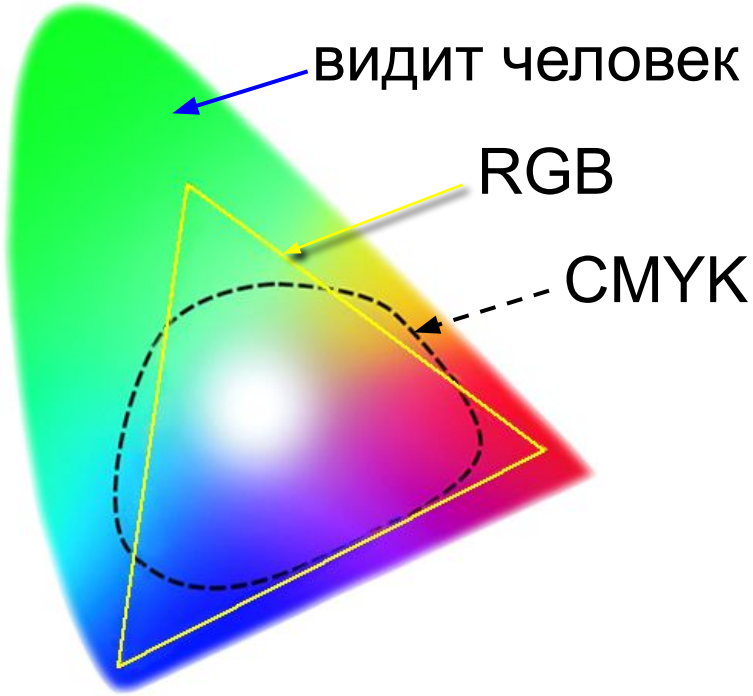
Модель CMY

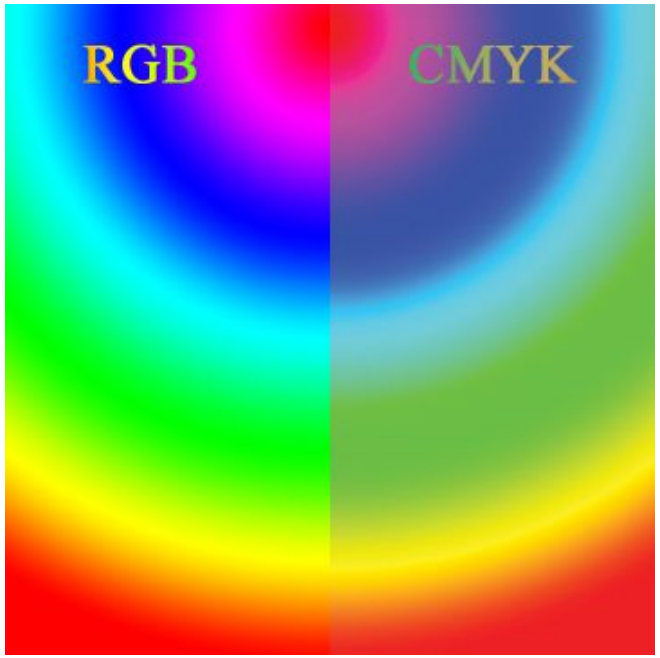
C	M	Y	
0	0	0	
255	255	0	
255	0	255	
0	255	255	
255	255	255	

Модель CMYK: + *Key color*

- меньший расход краски и лучшее качество для чёрного и серого цветов

RGB и CMYK





- не все цвета, которые показывает монитор (RGB), можно напечатать (CMYK)
- при переводе кода цвета из RGB в CMYK цвет искажается

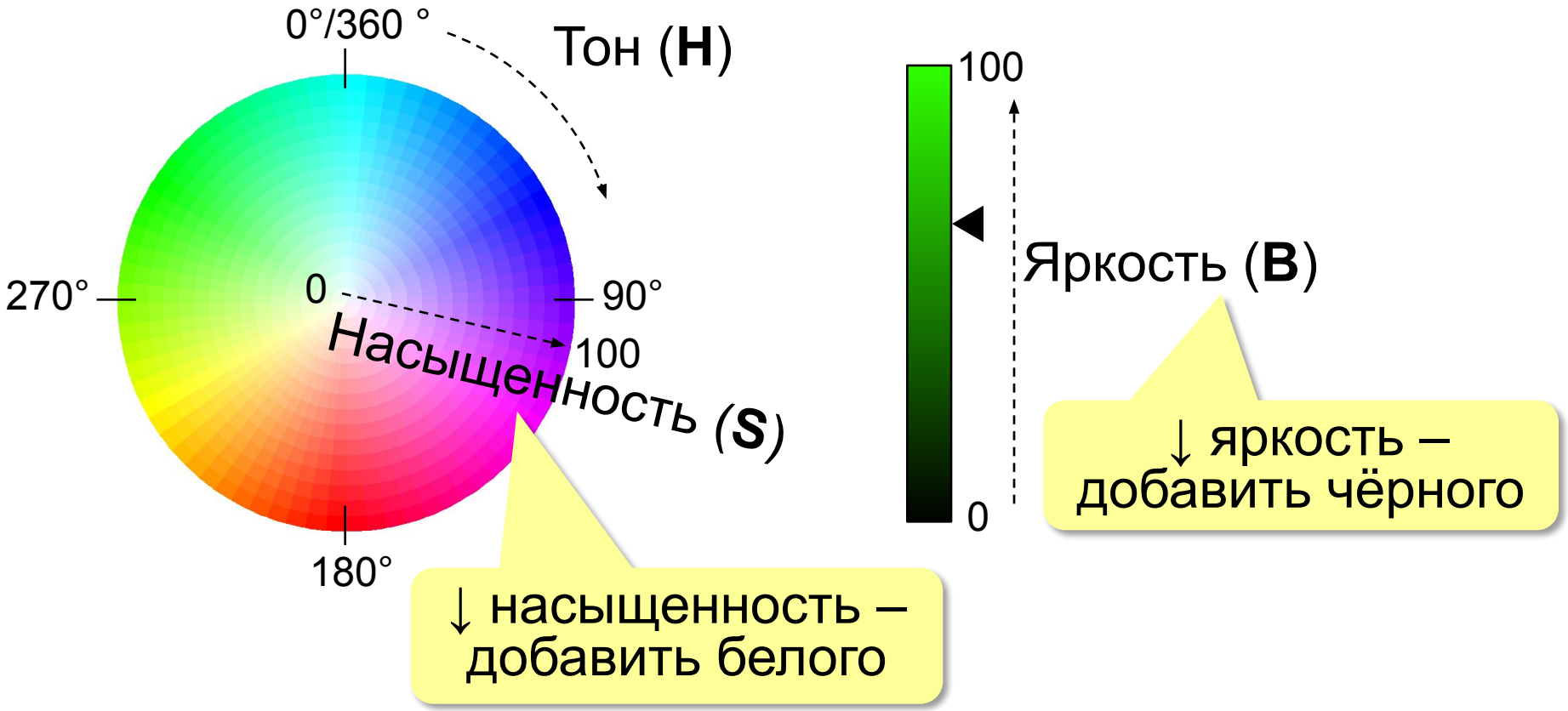
RGB(0,255,0)
→ CMYK(65,0,100,0)
→ RGB(104,175,35)

Цветовая модель HSB (HSV)

HSB = ***H****ue* (тон, оттенок)

S*aturation* (насыщенность)

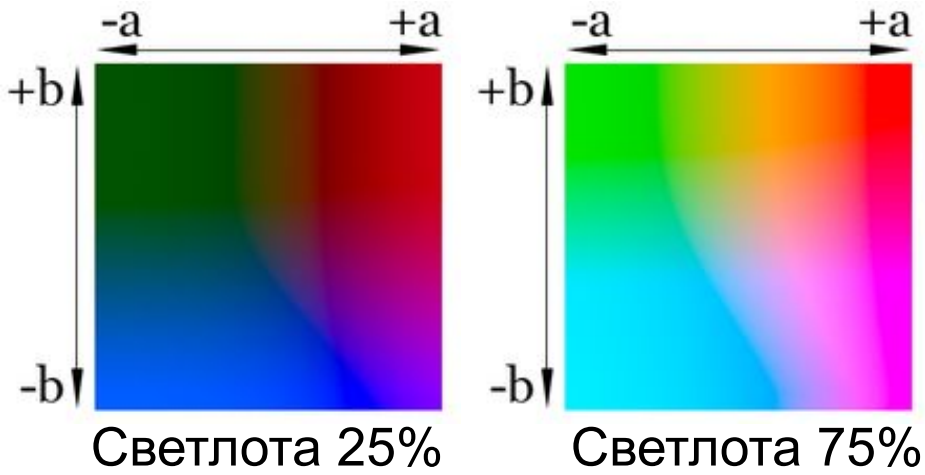
B*rightness* (яркость) или ***V****alue* (величина)



Цветовая модель Lab

Международный стандарт кодирования цвета, независимого от устройства (1976 г.)
Основана на модели восприятия цвета человеком.

Lab = *L*ightness (светлота)
a, b (задают цветовой тон)



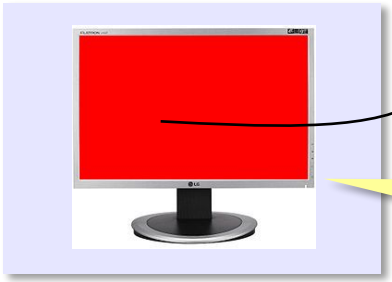
Светлота 25% Светлота 75%

- для перевода между цветовыми моделями: RGB → Lab → CMYK
- для цветокоррекции фотографий

Профили устройств

Какой цвет увидим?

RGB(255,0,0)



как λ ≈ 680нм

RGB(255,0,0) → профиль монитора

λ ≈ 680нм →  → **RGB(225,10,20)** → профиль сканера

CMYK(0,100,100,0) →  → профиль принтера

Растровое кодирование: итоги

- универсальный метод (можно закодировать любое изображение)
- единственный метод для кодирования и обработки размытых изображений, не имеющих чётких границ (фотографий)

- есть **потеря информации** (почему?)
- при изменении размеров цвет и форма объектов на рисунке **искажается**
- **размер файла** не зависит от сложности рисунка (а от чего зависит?)

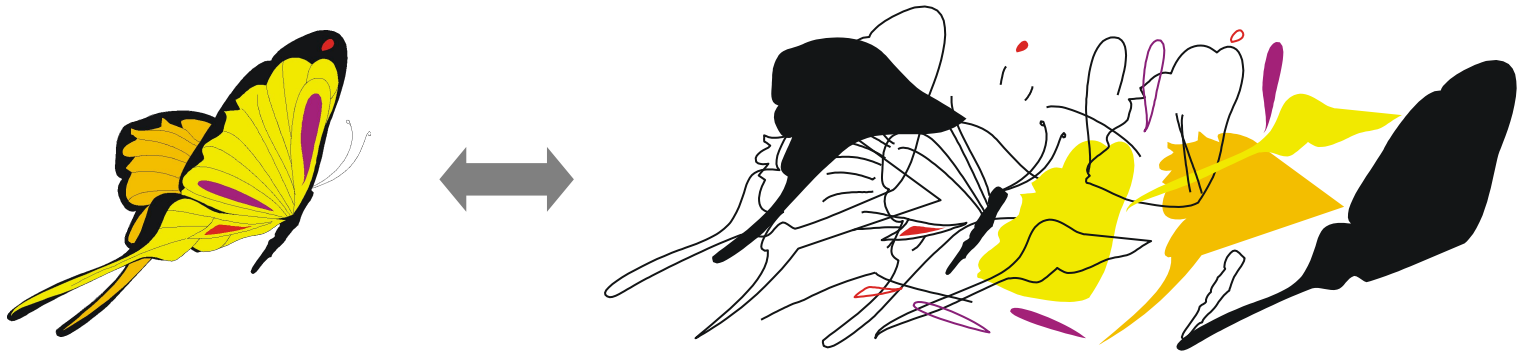
Векторное кодирование

Рисунки из геометрических фигур:

- отрезки, ломаные, прямоугольники
- окружности, эллипсы, дуги
- сглаженные линии (кривые Безье)

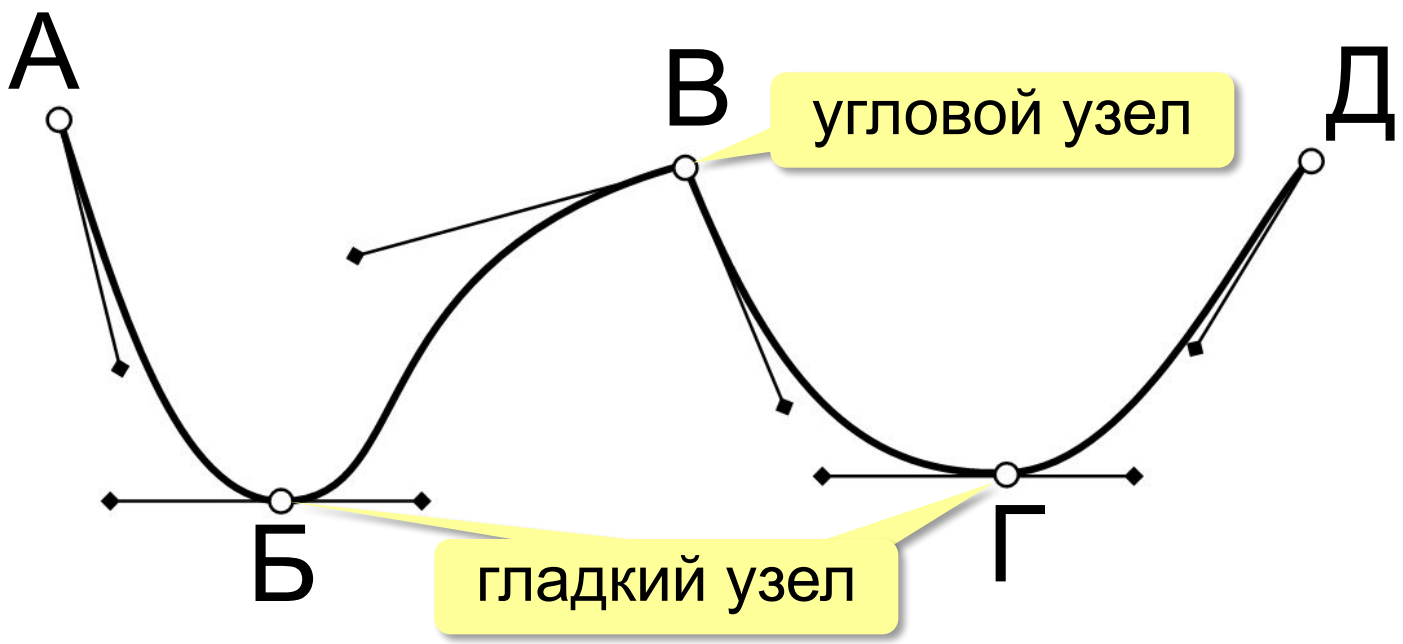
Для каждой фигуры в памяти хранятся:

- размеры и координаты на рисунке
- цвет и стиль границы
- цвет и стиль заливки (для замкнутых фигур)



Векторное кодирование

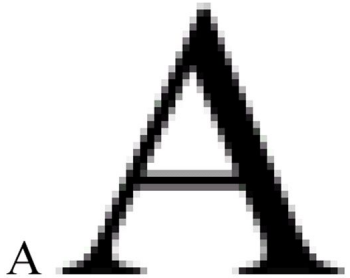
Кривые Безье:



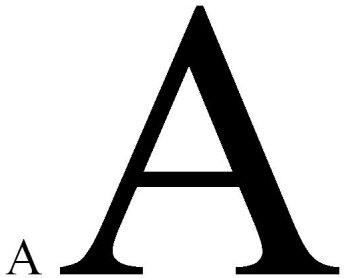
Хранятся координаты узлов и концов «рычагов»
(3 точки для каждого узла, кривые 3-го порядка).

Векторное кодирование (итоги)

- лучший способ для хранения **чертежей, схем, карт**
- при кодировании **нет потери информации**
- при изменении размера **нет искажений**



растровый рисунок



векторный рисунок

- меньше **размер файла**, зависит от сложности рисунка

- неэффективно использовать для **фотографий** и размытых изображений

Векторное кодирование: форматы файлов

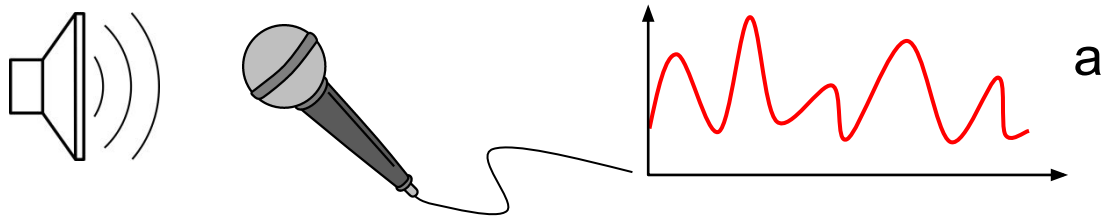
- **WMF** (*Windows Metafile*)
- **EMF** (*Windows Metafile*)
- **CDR** (программа *CorelDraw*)
- **AI** (программа *Adobe Illustrator*)
- **SVG** *(Scalable Vector Graphics,* масштабируемые векторные изображения*)*

для веб-страниц

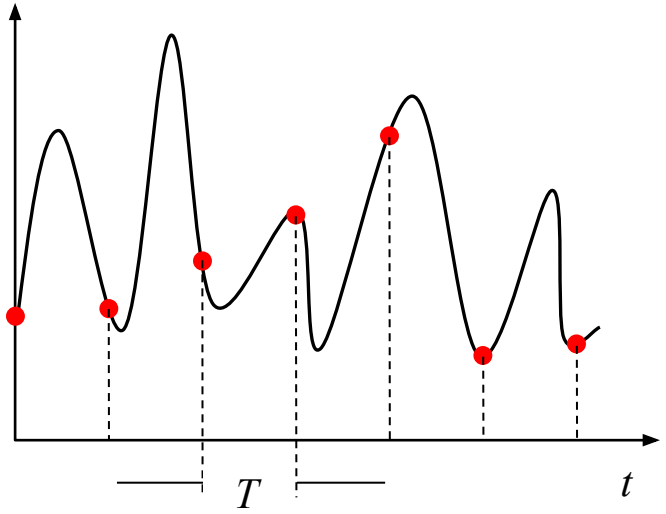
Кодирование информации

Кодирование звуковой и видеоинформации

Оцифровка звука



Оцифровка – это преобразование аналогового сигнала в цифровой код (дискретизация).



T – интервал дискретизации (с)

$f = \frac{1}{T}$ – частота дискретизации (Гц, кГц)

8 кГц – минимальная частота для распознавания речи

11 кГц, 22 кГц,

44,1 кГц – качество CD-дисков

48 кГц – фильмы на DVD

96 кГц, 192 кГц

Человек слышит

16 Гц … 20 кГц

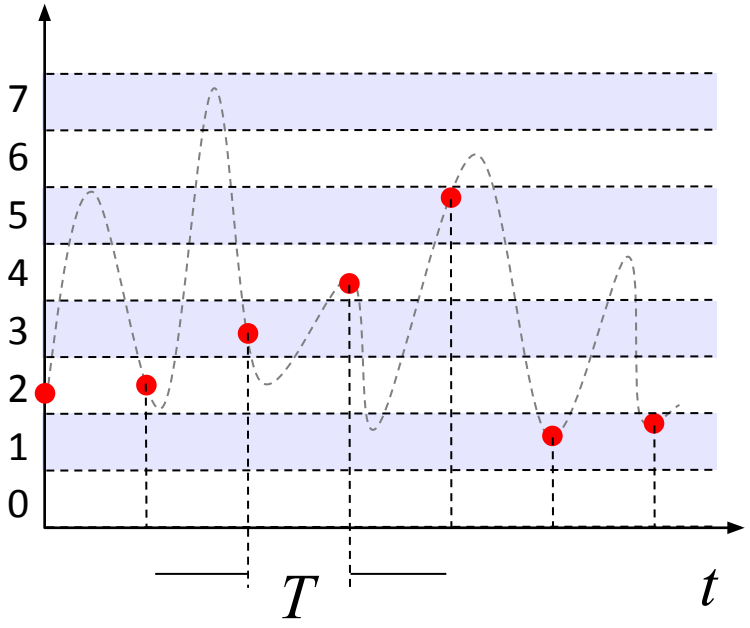
Оцифровка звука: квантование

Сколько битов нужно, чтобы записать число 0,6?

Квантование (дискретизация по уровню) – это представление числа в виде цифрового кода конечной длины.

АЦП = **А**налого-**Ц**ифровой **П**реобразователь

3-битное кодирование:



8 битов = 256 уровней

16 битов = 65536 уровней

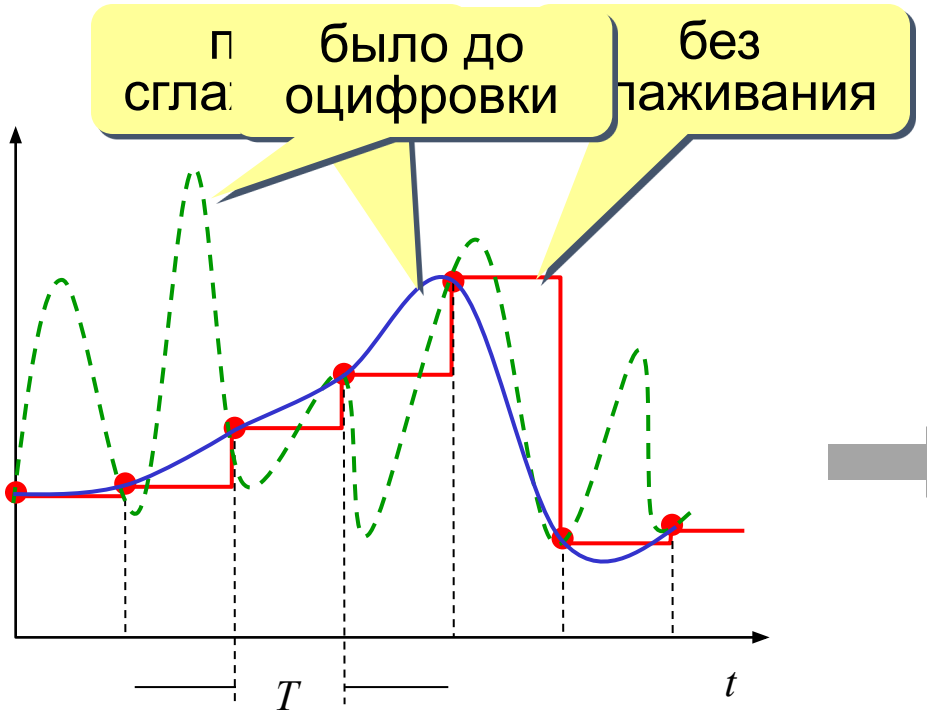
24 бита = 2^{24} уровней

Разрядность кодирования — это число битов, используемое для хранения одного отсчёта.

Оцифровка звука

Как восстановить сигнал?

ЦАП = **Ц**ифро-**А**налоговый **П**реобразователь



аналоговые устройства!

Как улучшить качество? — уменьшать T

Что при этом ухудшится? — ↑ размер файла

Оцифровка – итог

можно закодировать **любой звук** (в т.ч. голос, свист, шорох, …)

- есть **потеря информации**
- большой **объем файлов**

Какие свойства оцифрованного звука определяют качество звучания?

Форматы файлов:

WAV (*Waveform audio format),* часто без сжатия (размер!)

MP3 (*MPEG-1 Audio Layer 3*, сжатие с учётом восприятия человеком)

AAC (*Advanced Audio Coding*, 48 каналов, сжатие)

WMA (*Windows Media Audio,* потоковый звук, сжатие)

OGG (*Ogg Vorbis,* открытый формат, сжатие)

Инструментальное кодирование

MIDI (*Musical Instrument Digital Interface* — цифровой интерфейс музыкальных инструментов).

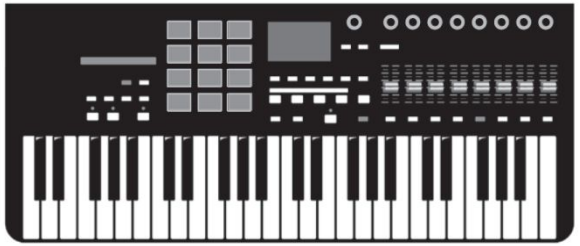
в файле **`.mid`**:
- нота (высота, длительность)
- музыкальный инструмент
- параметры звука (громкость, тембр)
- до 1024 каналов

128 мелодических и 47 ударных

программа для звуковой карты!

в памяти звуковой карты:
- образцы звуков (волновые таблицы)

MIDI-клавиатура:



- **нет потери информации** при кодировании инструментальной музыки
- небольшой **размер файлов**

невозможно закодировать нестандартный звук, голос

Трекерная музыка

В файле (модуле):

- образцы звуков (*сэмплы*)
- нотная запись, трек *(track)* – дорожка
- музыкальный инструмент
- до 32 каналов

Форматы файлов:

MOD разработан для компьютеров *Amiga*

S3M оцифрованные каналы + синтезированный звук, 99 инструментов

XM, STM, …

Использование: демосцены (важен размер файла)

Кодирование видео

Видео = изображения + звук Синхронность!

изображен
- ≥ 25 кад
- **PAL**: 768
  - за 1
  - за 1
- **HDTV**: 1
- исходны
- сжатие (
- DivX, Xvi

звук:
- 48 кГц, 1
- сжатие (
- MP3, AA

Форматы видеофайлов

AVI – *Audio Video Interleave* – чередующиеся звук и видео; контейнер – могут использоваться разные *кодеки*

MPEG – *Motion Picture Expert Group*

WMV – *Windows Media Video,* формат фирмы *Microsoft*

MP4 – *MPEG-4*, сжатое видео и звук

MOV – *Quick Time Movie*, формат фирмы *Apple*

WebM – открытый формат, поддерживается браузерами