Кодирование информации

Кодирование информации

§ 5. 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-буквенных слов в русском языке?
- возможных 3-буквенных слов в английском языке?

33^5

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

Задачи

1. Сколько различных пятизначных чисел можно записать с помощью цифр 4 и 2?
2. В языке разрешены только четырёхбуквенные слова, которые можно образовывать из букв алфавита в любых комбинациях. Словарный запас языка составляет 81 слово. Какова мощность алфавита?
3. Какое наименьшее число символов должно быть в алфавите, чтобы с помощью всевозможных трёхбуквенных слов можно было передать не менее 9 различных сообщений?

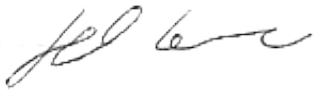
Кодирование информации

§ 6. 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*: *CFF0E8E2E52C20C2E0F1FF21*
- стенография: 
- шифрование: ***Рсйгжу-!Гбта”***

Числа:

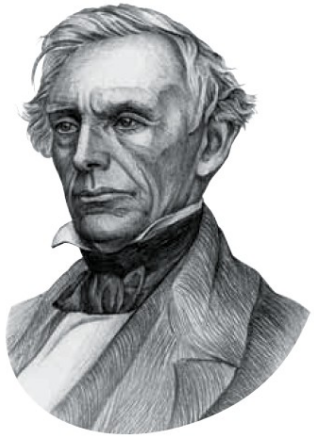
- для вычислений: ***25***
- прописью: ***двадцать пять***
- римская система: ***XXV***

? 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

Задачи

1. Сколько существует в коде Морзе различных последовательностей из точек и тире, длина которых от 4 до 6 символов?
2. Вася и Петя передают друг другу сообщения, используя синий, красный и зелёный фонарики. Это они делают, включая по одному фонарику на одинаковое короткое время в некоторой последовательности. Количество вспышек в одном сообщении — 3 или 4, между сообщениями — паузы. Сколько различных сообщений могут передавать мальчики?

Задачи

3. Шахматная доска состоит из 8 столбцов и 8 строк. Какое минимальное количество битов потребуется для кодирования координат одной шахматной фигуры?
4. Для кодирования значений температуры воздуха (целое число в интервале от –50 до 40) используется двоичный код. Какова минимальная длина двоичного кода?
5. Дорожный светофор подаёт шесть видов сигналов (непрерывные красный, жёлтый и зелёный, мигающие жёлтый и зелёный, мигающие красный и жёлтый одновременно). Подряд записано 100 сигналов светофора. Определите информационный объём этого сообщения в битах.

Задачи

6. Автомобильный номер длиной 6 символов составляется из заглавных букв (всего используется 12 букв) и десятичных цифр в любом порядке. Каждый символ кодируется одинаковым и минимально возможным количеством битов, а каждый номер — одинаковым и минимально возможным количеством байтов. Определите объём памяти, необходимый для хранения 32 автомобильных номеров.

Декодирование

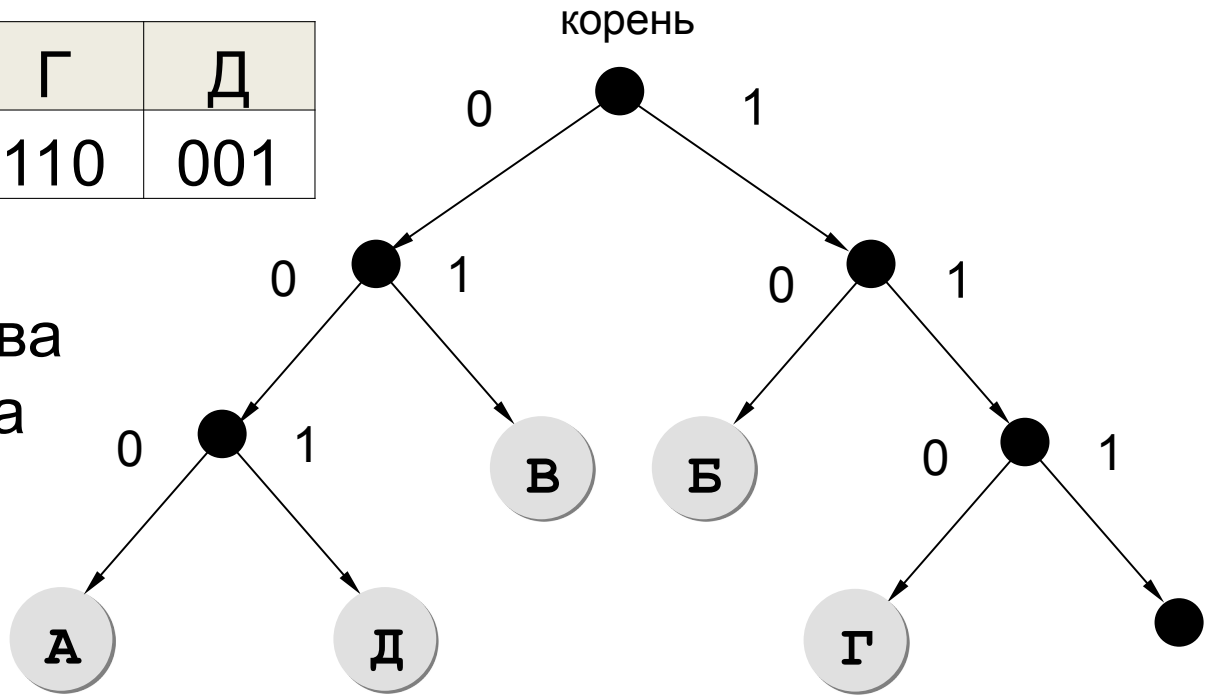
Декодирование — это восстановление сообщения из последовательности кодов.

•– – •– ••• •–•– **ВАСЯ**

? Когда разделитель не нужен?

А	Б	В	Г	Д
000	10	01	110	001

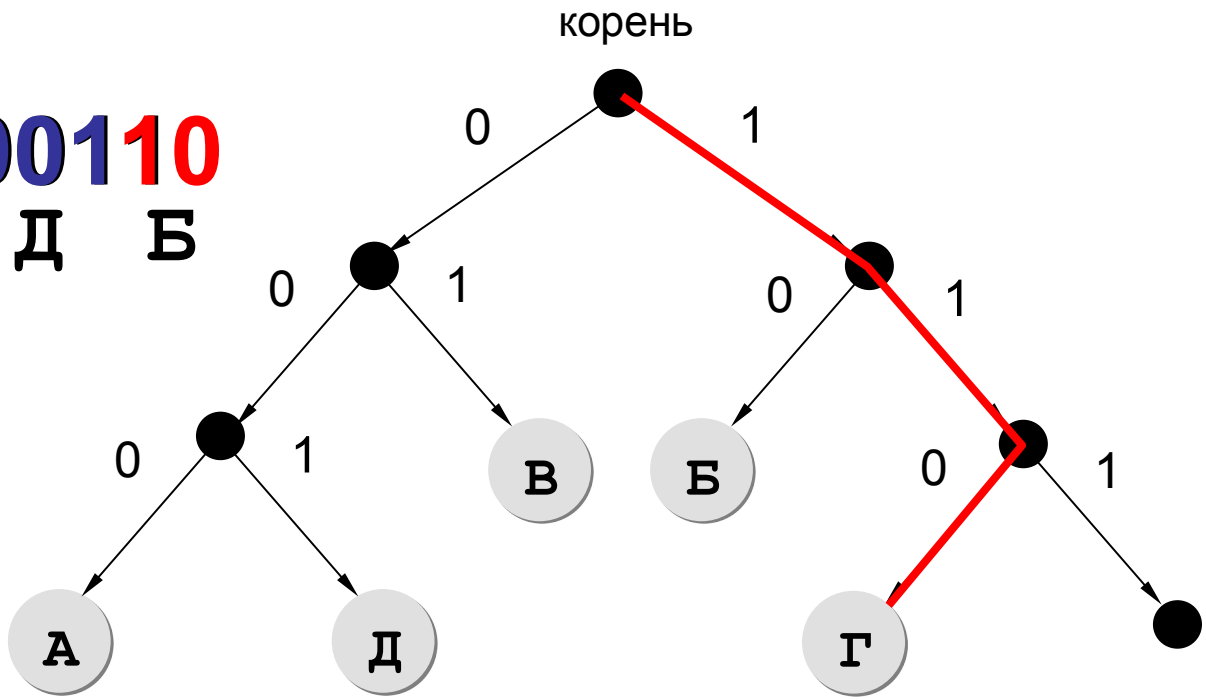
Все кодовые слова заканчиваются на листьях дерева!



Декодирование

1100000100110

0 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.

Задачи

1. Для передачи сообщения, состоящего только из букв А, Б, В, Г, решили использовать неравномерный код:

   А = 0, Б = 10, В = 110.

   Как нужно закодировать букву Г, чтобы длина кода была минимальной и допускалось однозначное декодирование?

2. Для передачи сообщения, состоящего только из букв А, Б, В, Г, решили использовать неравномерный код:

   А = 0, Б = 100, В = 101.

   Как нужно закодировать букву Г, чтобы длина кода была минимальной и допускалось однозначное декодирование?

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

011000110110

Б Д 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 → **АБВГА**

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

Задача

*Докажите, что все сообщения, закодированные этим кодом, декодируются однозначно.

А	Б	В
0	11	010

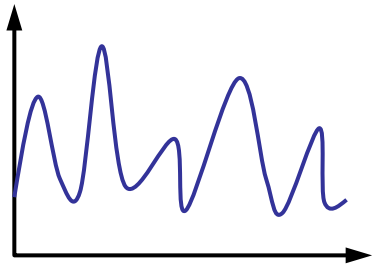
0100001100101111000100

Кодирование информации

§ 7. Дискретность

Аналоговые сигналы и устройства

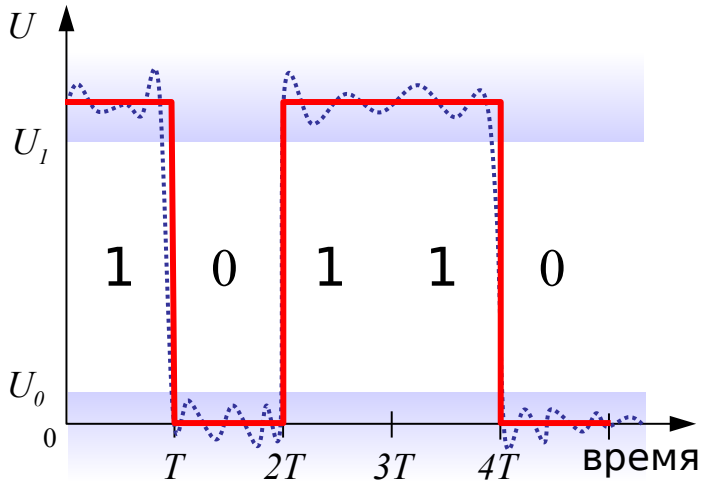
Аналоговый сигнал — это сигнал, который в любой момент времени может принимать любые значения в заданном диапазоне.



Аналоговые компьютеры

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

Дискретные (цифровые) сигналы



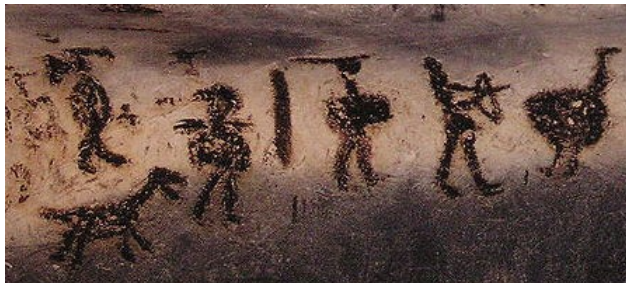
Свойства:

- 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

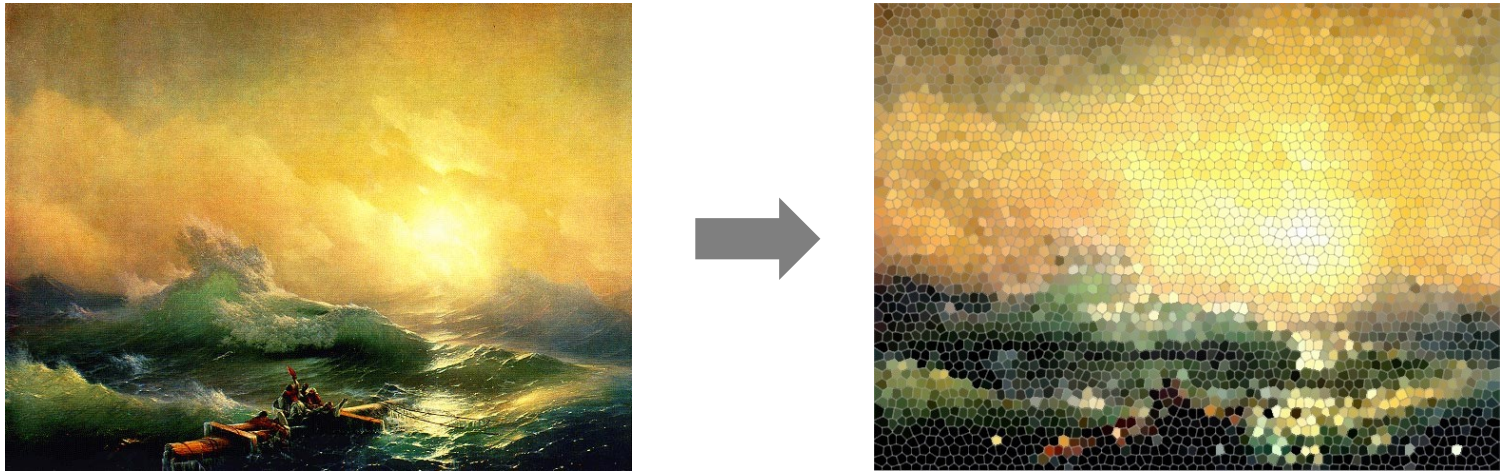
! Компьютеры могут хранить и обрабатывать только дискретную информацию!

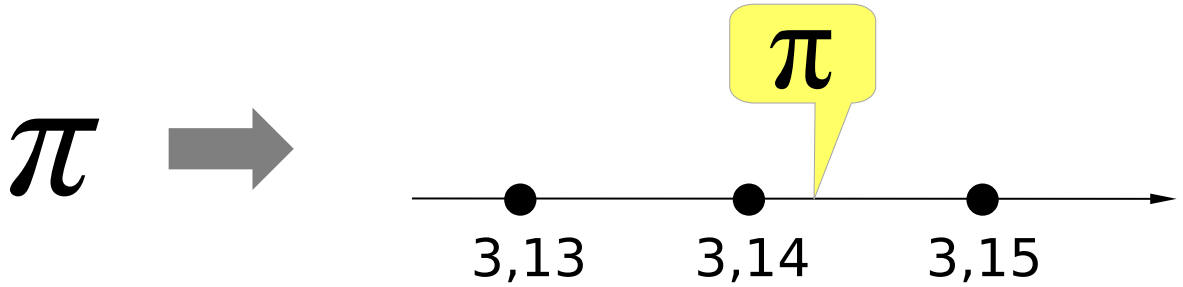
… закодированную с помощью конечного количества знаков некоторого алфавита.

! Все виды информации нужно перевести в дискретный вид!

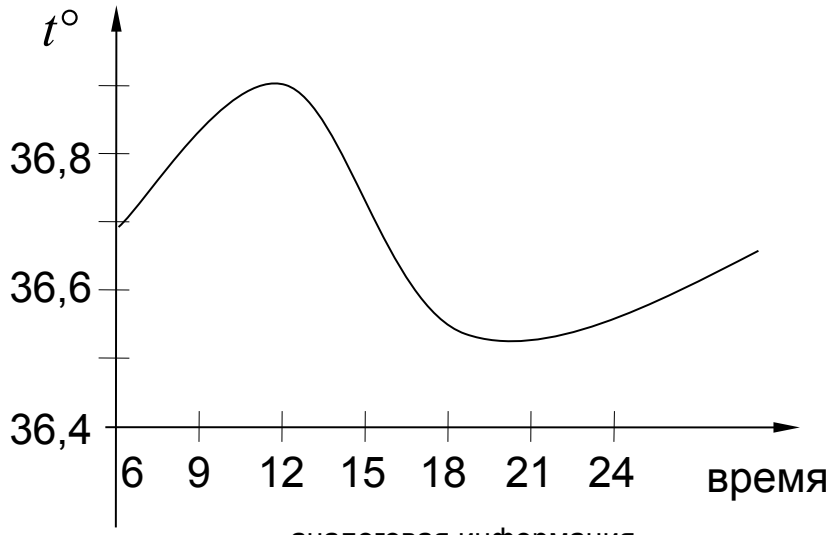
Дискретизация

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

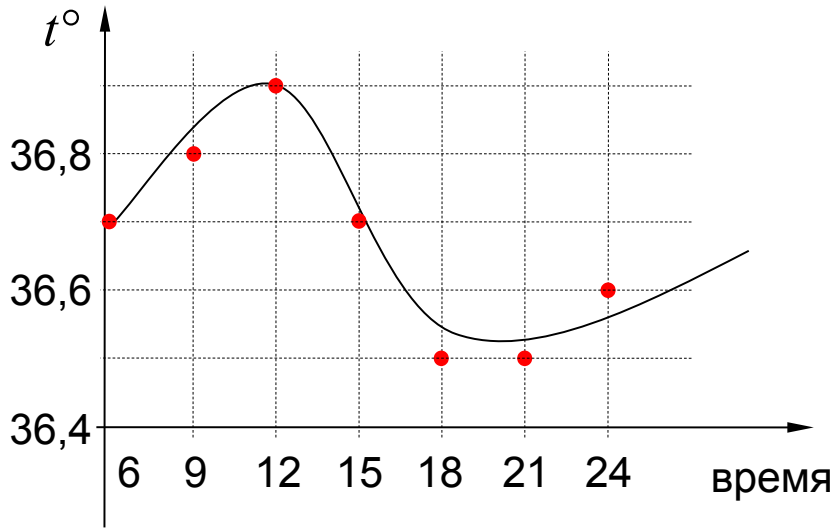




Дискретизация



аналоговая информация



дискретизация

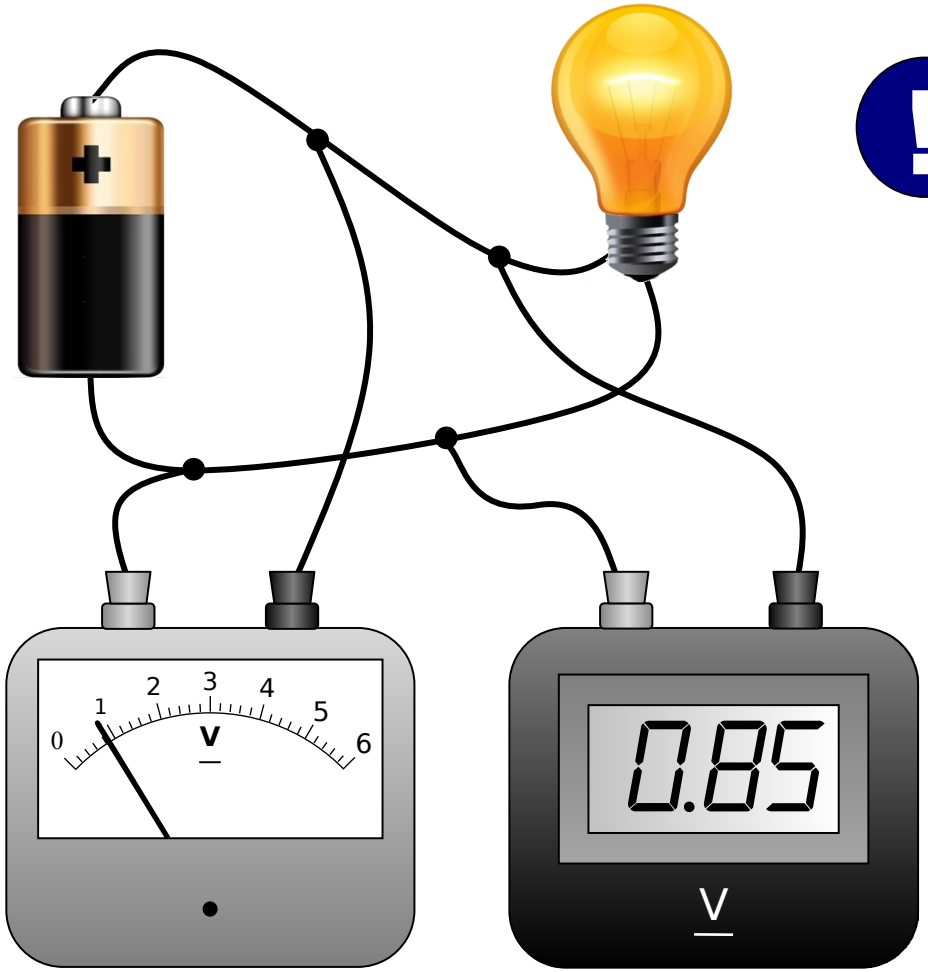
6 ч.	36,7°
9 ч.	36,8°
12 ч.	36,9°
15 ч.	36,7°
18 ч.	36,5°
21 ч.	36,5°
24 ч.	36,6°

дискретная информация

! При дискретизации есть потеря информации!

? Как уменьшить потери?

Непрерывность и дискретность



аналоговые данные

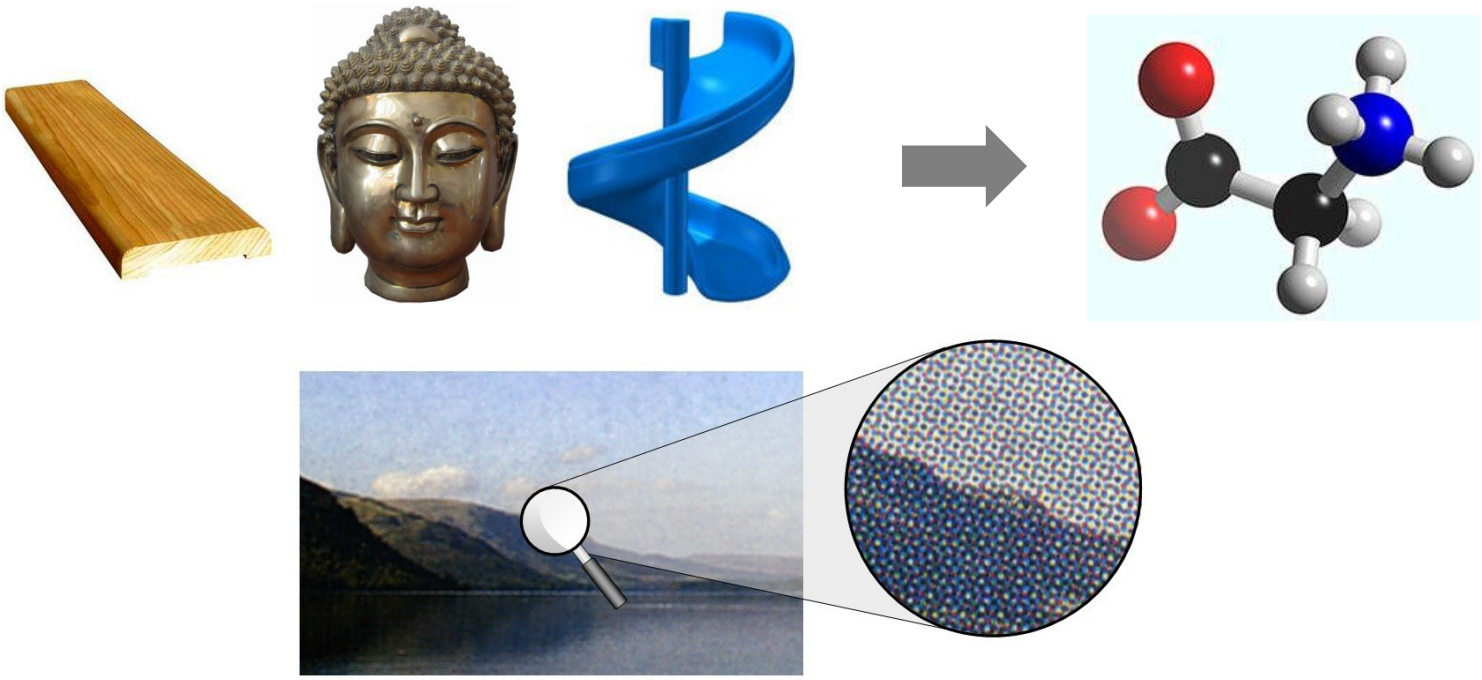
дискретные данные

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

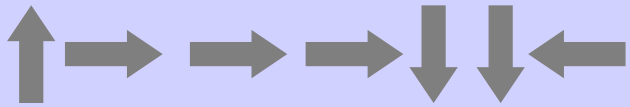
§ 8. Алфавитный подход к измерению количества информации

Алфавитный подход

Количество информации в битах определяется длиной сообщения в двоичном коде.

10101100 — 8 битов

вперёд	00
назад	01
вправо	10
влево	11



? 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

3) 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

Задача

Определить количество информации в 10 страницах текста (на каждой странице 32 строки по 64 символа) при использовании алфавита из 256 символов.

1) информационная ёмкость символа:

$$256 = 2^8 \Rightarrow i = 8 \text{ бит} = 1 \text{ байт}$$

2) количество символов на странице:

$$32 \cdot 64 = 2^5 \cdot 2^6 = 2^{11}$$

3) общее количество символов:

$$L = 10 \cdot 2^{11}$$

4) информационный объём сообщения:

$$I = L \cdot i = 10 \cdot 2^{11} \cdot 1 \text{ байтов} = 20 \text{ Кбайт}$$