

Сжатие данных

Информационные технологии.

Лекция №2



Коэффициент сжатия

сжатия

$$K = \frac{V_x}{V_y}$$

V_x — объем входного потока

V_y — объем выходного потока

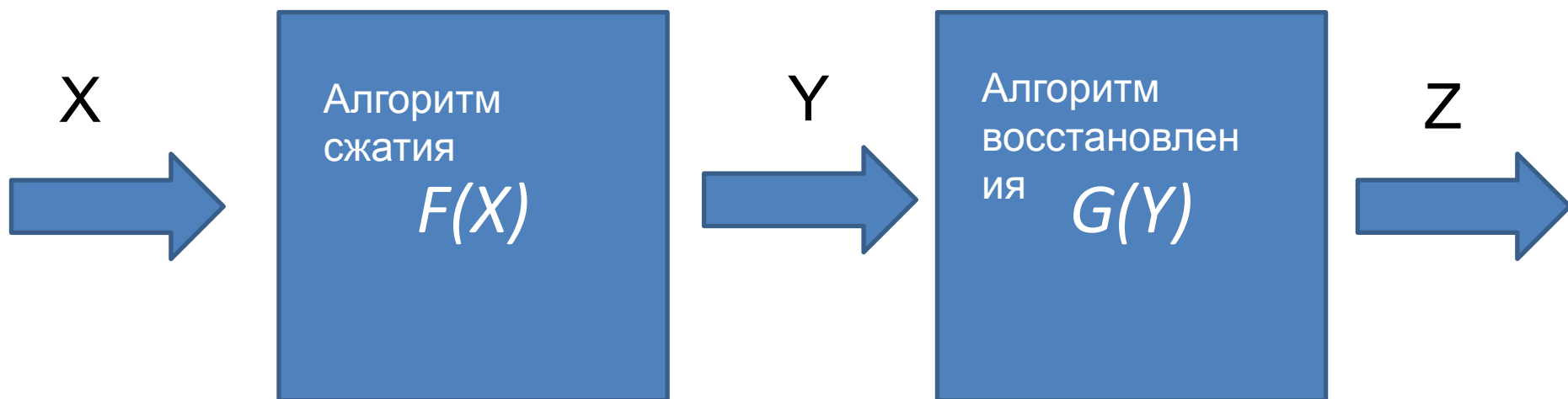
$$K \neq const$$

для разных данных могут

быть:
 $K > 1$

$$K = 1$$

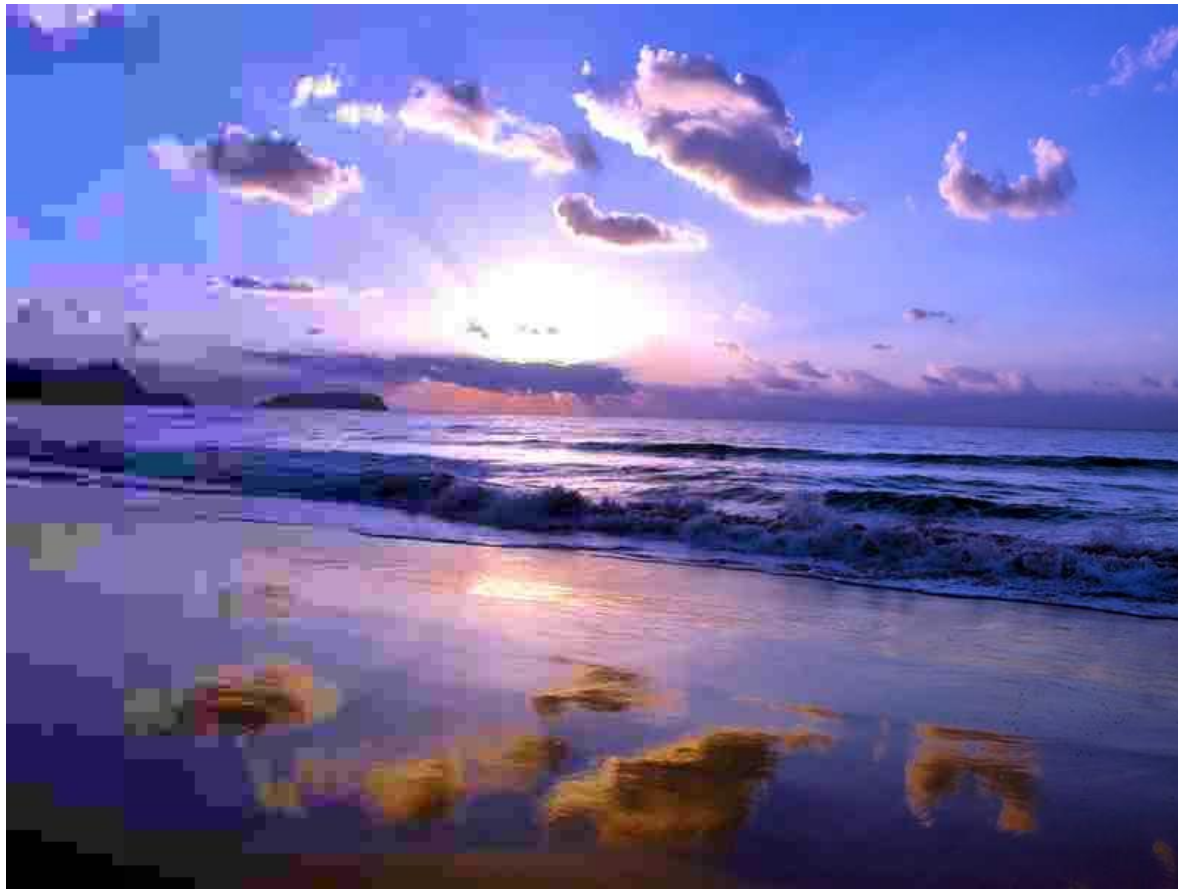
$$K < 1$$



$X = Z$ – сжатие без потерь

$X \neq Z$ – сжатие с потерями

Пример сжатия с потерями (JPEG)



Quality

Статистические алгоритмы сжатия данных

Для сжатия данных используется информация об *энтропии* входного потока.

Используется *частота/вероятность* появления символа в потоке.

Буква	Частота %	Буква	Частота %	Буква	Частота %	Буква	Частота %
О	11,08	Р	4,45	Ы	1,96	Х	0,89
Е, Ё	8,41	В	4,33	Ь	1,92	Ш	0,81
А	7,92	К	3,36	З	1,75	Ю	0,61
И	6,83	М	3,26	Г	1,74	Э	0,38
Н	6,72	Д	3,05	Б	1,71	Щ	0,37
Т	6,18	П	2,81	Ч	1,47	Ц	0,36
С	5,33	У	2,80	Й	1,12	Ф	0,19
Л	5,00	Я	2,13	Ж	1,05	Ъ	0,02

Примеры алгоритмов:

1. Алгоритм Шеннона-Фано.
2. Алгоритм Хаффмана.
3. Арифметическое кодирование.

Алгоритм Хаффмана

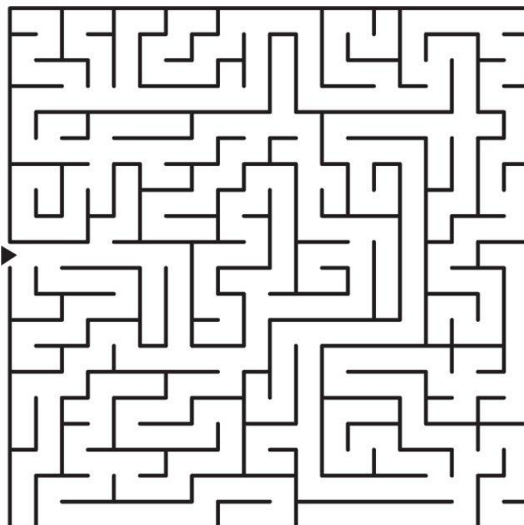
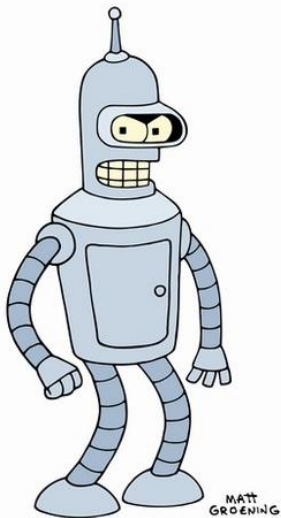
1. *Определить частотную таблицу символов входного потока данных.*
2. *Упорядочить таблицу по убыванию частоты символов.*
3. *Построить дерево Хаффмана по частотной таблице.*
4. *Определить коды символов по дереву Хаффмана.*
5. *Закодировать сообщение – сформировать выходной поток.*

Задача

Дистанционно управляемый робот должен пройти лабиринт (клеточный).
Оптимальный алгоритм прохождения лабиринта известен, но необходимо его
(алгоритм) закодировать.

Робот способен выполнять следующие команды:

- 1) Идти вперед (1 кл.)
- 2) Повернуть вправо на 90 градусов
- 3) Повернуть влево на 90 градусов
- 4) Повернуться на 180 градусов
- 5) Перепрыгнуть обрыв



В результате исследования была
найдена оптимальная
последовательность:

```
11112111311112111311131113111211111  
131121213111111111111112111311511211  
11131111211311111211113111111111112  
11131111131151113111111211113113111  
11131111211112111121111211121121131  
11111111115111311121113113113121313  
11121111211211
```

Решение

Количество шагов алгоритма (символов) - 223

Частотная
таблица

Символ	Частота
1	175
2	20
3	25
4	0
5	3

Пример 2

Количество шагов алгоритма (символов) - 223

Частотная
таблица

Символ	Частота
1	80
2	50
3	60
4	0
5	33

Арифметическое кодирование

1. *Определить вероятностную таблицу* символов входного потока данных.
2. Разделить отрезок $[0;1)$ на отрезки в соответствии с вероятностями символов.
3. Поочередно кодировать символы входного потока на получившихся отрезках.

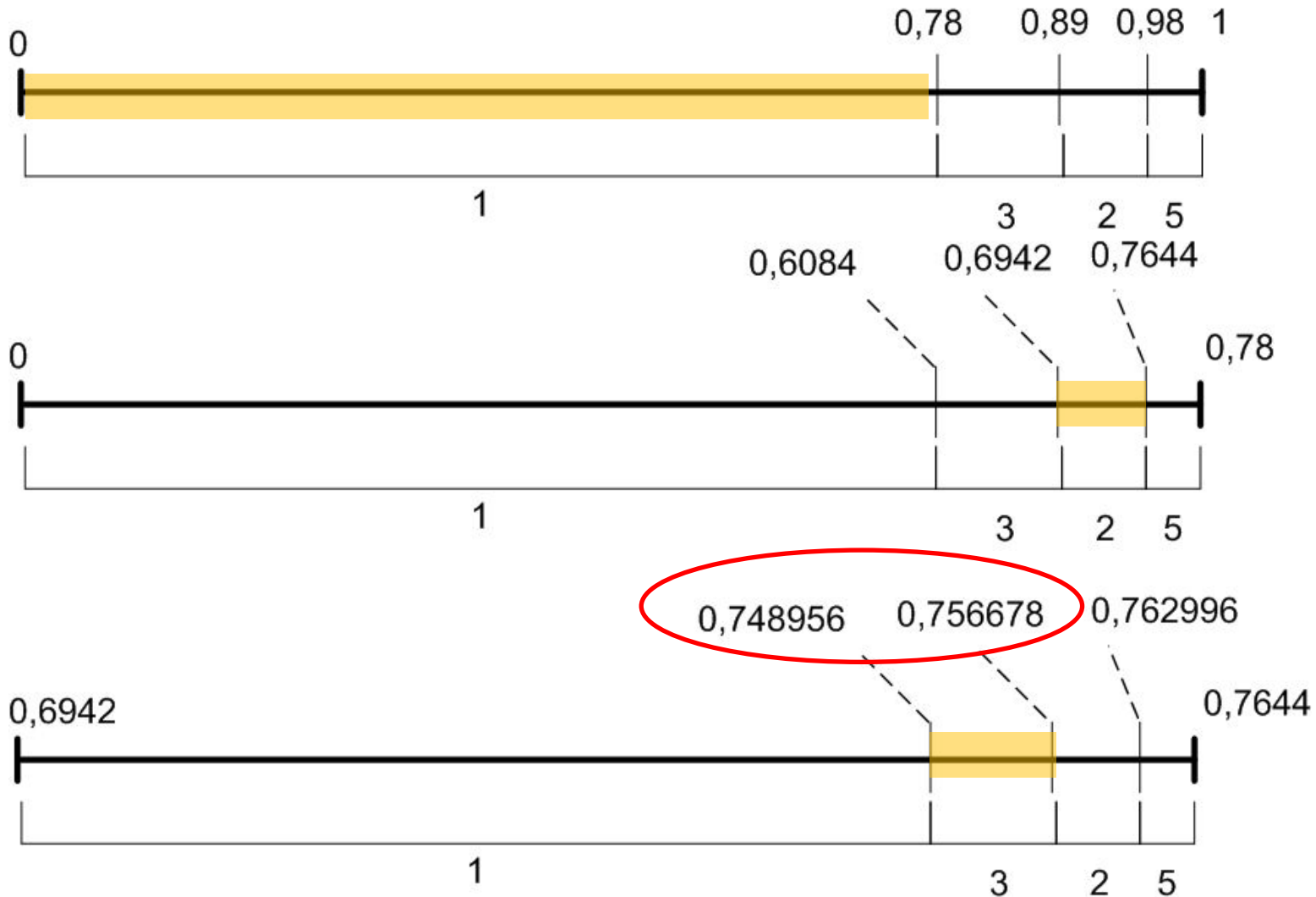
Пример арифметического кодирования

Данные возьмем из предыдущей задачи (алгоритм Хаффмана).

Вероятностная таблица

Символ	Частота	Вероятность
1	175	0,78
2	20	0,09
3	25	0,11
4	0	0,02
5	3	0

Пример арифметического кодирования



RLE

Run-length encoding – кодирование длин серий

Пример:

11112111311112111311131113111211111
131121213111111111111112111311511211
1113111121131111121111311111111112
111311113115111311111211113113111
11131111211112111121111211121121131
1111111115111311121113113113121313
1113111211311

11112111311112111311131113111211111
131121213111111111111112111311511211
1113111121131111121111311111111112
1113111131151113111111211113113111
11131111211112111121111211121121131
1111111115111311121113113113121313
1113111211311

Другие популярные алгоритмы

LZ77

LZW

BWT

DEFLATE