# *Кодирование данных*

1. Способы представления и методы передачи информации
2. Передача данных по каналам связи
3. Кодирование данных
4. Контроль передачи данных
5. Сжатие данных

# Способы физического представления информации

**Информация представляется в двоичном алфавите.**

**Физическими аналогами знаков 0 и 1 служат сигналы, способные принимать два хорошо различимых значения.**

**Такт** — временной интервал между двумя соседними моментами дискретного времени.

**Потенциальный способ**. Двум значениям переменной 1 и 0 соответствуют разные уровни напряжения — потенциальный код. Потенциальный сигнал сохраняет постоянный уровень в течение такта; его значение в переходные моменты является неопределенным.

**Импульсный способ**. Двум значениям двоичной переменной 1 и 0 соответствует наличие и отсутствие электрического импульса либо разнополярные импульсы — импульсный код.

# Методы передачи информации

**Последовательный**

Каждый временной такт предназначен для отображения одного разряда кода слова.

Все разряды слова фиксируются по очереди одним и тем же элементом.

Все разряды слова проходят через одну линию передачи информации.

**Параллельный**

- Все разряды кода слова представляются в одном временном такте.
- Все разряды слова фиксируются отдельными элементами.
- Все разряды слова проходят через отдельные линии.
- Каждая линия служит для передачи только одного разряда слова.
- Значения всех разрядов слова передаются по нескольким линиям одновременно.

# Типы каналов


**Канал связи** — совокупность средств, обеспечивающих передачу сигналов.

**Физическая среда** — пространство или материал, обеспечивающие распространение сигналов (проводная воздушная или кабельная линия, скрученная пара проводов, коаксиальный кабель, стекловолоконная линия, эфир).

# Передача информации

**Линия связи** — это среда, по которой передаются сигналы.

**Шум** — это помехи в канале связи при передаче информации.

**Кодирование** — преобразование дискретной информации одним из следующих способов: шифрование, сжатие, защита от шума.

# Основные понятия

**Кодирование данных** — процесс преобразования символов алфавита X в символы алфавита Y.

**Декодирование** — процесс, обратный кодированию.

**Символ** — наименьшая единица данных, рассматриваемая как единое целое при кодировании/декодировании.

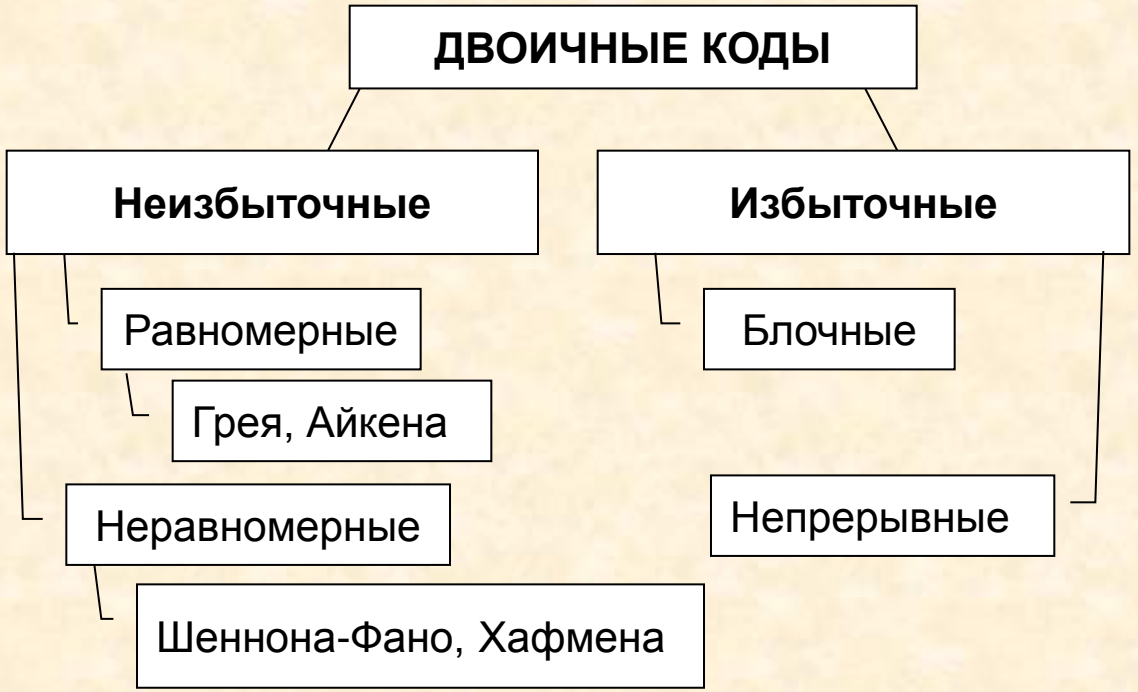
**Код** — совокупность правил, в соответствии с которыми производится кодирование

**Кодовое слово** — последовательность символов из алфавита Y, однозначно обозначающая конкретный символ алфавита X.

**Средняя длина кодового слова** — это величина, которая вычисляется как взвешенная вероятностями сумма длин всех кодовых слов

$$L = \sum_{i=1}^{N} p_i * l_i$$

# Классификация двоичных кодов




Код называется *простым*, если все разряды слова служат для представления информации.

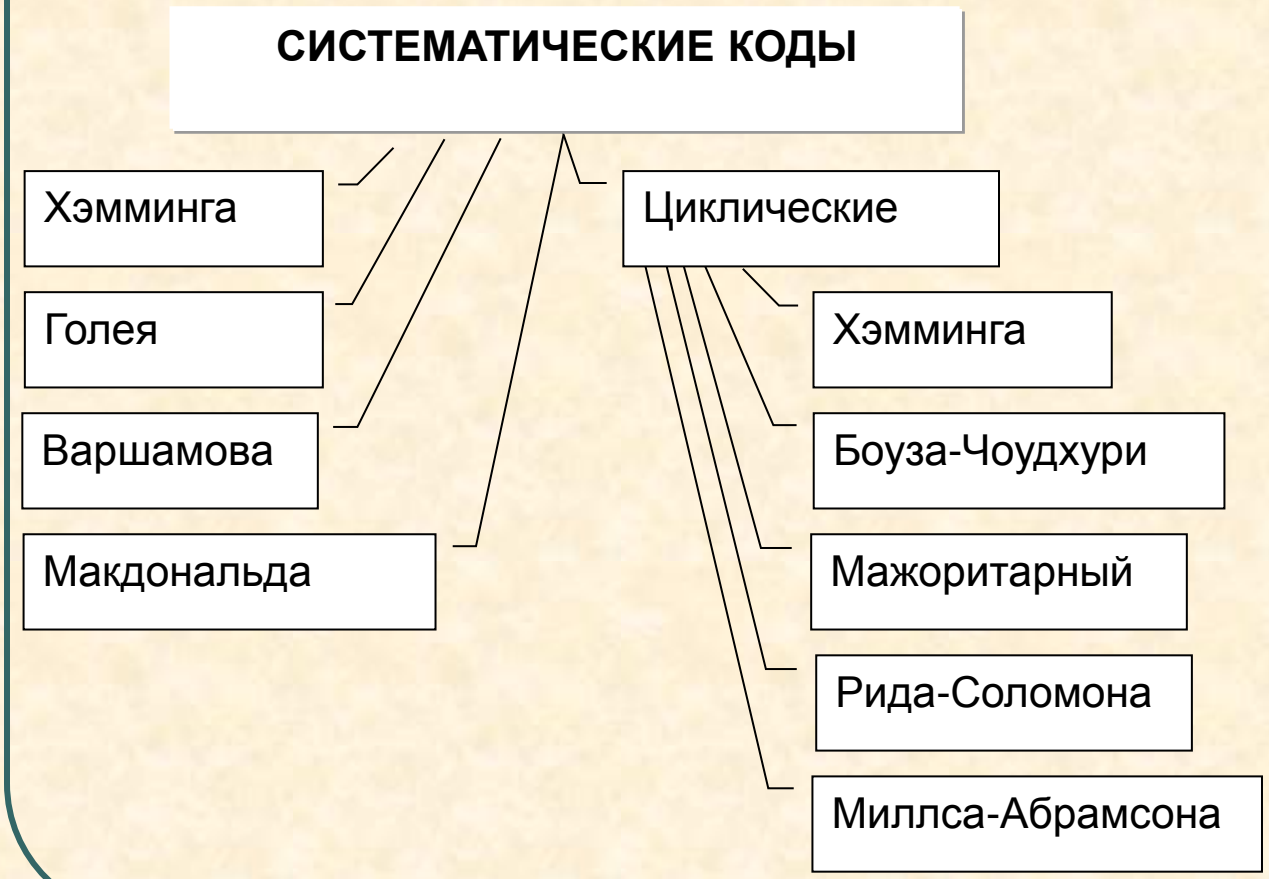
В *равномерном* коде кодовые слова имеют одинаковую длину.

В *неравномерном* коде встречаются кодовые слова разной длины.

В *блочных* кодах каждому сообщению соответствует кодовая комбинация (блок) из n символов. Блоки кодируются и декодируются отдельно друг от друга.

В *избыточных* кодах кроме информационных есть проверочные разряды.

# Классификация двоичных кодов




В *систематических* кодах каждый проверочный символ выбирается таким образом, чтобы его сумма по модулю два с определенными информационными символами была равной нулю.

# Характеристики кодов

**Длина кода** n — число разрядов, составляющих кодовую комбинацию.

**Основание кода** m — количество отличающихся друг от друга значений импульсных признаков, используемых в кодовых комбинациях. Для случая двоичных кодов m = 2. Значения импульсных признаков — цифры 0 и 1.

**Мощность кода** $N_p$ — число кодовых комбинаций, используемых для передачи сообщений.

**Вес кодовой комбинации** — количество единиц в кодовой комбинации.

**Полное число кодовых комбинаций** N — число всех возможных комбинаций длины n из m различных символов, равное $m^n$ (для двоичных кодов $N = 2^n$).

**Вероятность необнаруженной ошибки** — это вероятность события, при котором свойства данного кода не позволяют определить факт наличия ошибки в принятой комбинации.

# Характеристики кодов

**Число информационных символов** — количество символов (разрядов) кодовой комбинации, предназначенных для передачи собственно сообщения.

**Число проверочных символов** — количество символов (разрядов) кодовой комбинации, необходимых для коррекции ошибок.

**Скорость передачи кодовых комбинаций** — отношение числа информационных разрядов к длине кода.

**Оптимальность кода** — свойство кода, которое обеспечивает наименьшую вероятность не обнаружения ошибки среди всех кодов той же длины и избыточности.

Под **избыточностью кода** понимают относительную избыточность, равную отношению числа проверочных разрядов к длине кода.

# Ошибки при передаче данных

Ошибки при передаче данных происходят из-за шума в канале, а также при кодировании и декодировании.

**Достоверность передачи данных** оценивается отношением числа ошибочно принятых символов к общему числу переданных.

Теория информации изучает, в частности, способы минимизации количества таких ошибок.

Для минимизации вероятности ошибки при передаче данных используют *помехозащитные* коды. Идея состоит в добавлении к символам исходных кодов нескольких контрольных символов.

При контроле передачи информации наибольшее распространение получили *методы информационной избыточности*, использующие коды с обнаружением и коррекцией ошибок.

# Обнаружение ошибок при передаче данных

Способность кода обнаруживать или исправлять ошибки определяется *минимальным кодовым расстоянием*.

**Кодовое расстояние** — расстояние между двумя любыми словами в коде, определяемое весом суммы по модулю 2 этих кодовых комбинаций.

**Для избыточных кодов $d_{min} > 1$.**

**Пример**

Если $d_{min} = 2$, то любые два слова в данном коде отличаются не менее чем в двух разрядах. Следовательно, любая одиночная ошибка приведет к появлению запрещенного слова и может быть обнаружена.

**Условие обнаружения ошибки кратностью r**

$$d_{min} \geq r + 1.$$

Одновременная ошибка в r разрядах слова создает новое слово. Чтобы оно не совпало с другим разрешенным словом, $d_{min}$ должно быть хотя бы на 1 больше, чем r.

# Контроль четности

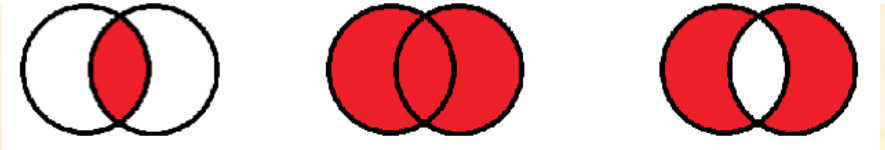
**Контроль четности** — простейший код для борьбы с шумом.

**Контрольная сумма** — некоторое число, рассчитанное путем применения определенного алгоритма к набору данных и используемое для проверки целостности этого набора при передаче или хранении.

**Бит чётности** — частный случай контрольной суммы, представляющий из себя 1 контрольный бит, используемый для проверки четности количества единичных битов в двоичном числе.

**Сумма по модулю 2** — исключающее «ИЛИ» (для двух операндов), логическое сложение или битовое сложение, разность двух/трёх множеств.

$$A \mathit{mod2} B = A \oplus B = (\neg(A \wedge B)) \wedge (A \vee B) = \neg((A \wedge B) \vee (\neg A \vee \neg B))$$



# Использование кода с проверкой четности в схемах контроля

**Код с проверкой четности** применяется для контроля передачи данных между регистрами и для контроля считываемой информации в оперативной памяти.

Код образуется добавлением к группе информационных разрядов, представляющих простой код, одного избыточного разряда.

При формировании кода слова в контрольный разряд записывается 0 или 1 так, чтобы сумма 1 в слове, включая избыточный разряд, была *четной*.

Если при передаче данных приемное устройство обнаруживает, что в принятом слове значение контрольного разряда не соответствует четности суммы слова, то это считается *признаком ошибки*.

Минимальное расстояние кода $d_{min} = 2$. Код обнаруживает все одиночные ошибки и все случаи нечетности числа ошибок.

# Код Хэмминга

**Код Хэмминга** — блочный равномерный разделимый самокорректирующийся код. Исправляет одиночные битовые ошибки, возникшие при передаче или хранении данных.

При построении кода Хэмминга к имеющимся информационным разрядам слова добавляется определенное число контрольных разрядов, после чего вся конструкция записывается в оперативную память.

При считывании слова контрольная аппаратура образует из информационных и контрольных разрядов *корректирующее число.*

Корректирующее число равно 0 при отсутствии ошибки либо указывает *номер ошибочного разряда* в слове.

Ошибочный разряд автоматически корректируется изменением его состояния на противоположное.

# Разрядность корректирующего числа в коде Хэмминга

Пусть кодовое слово длиной **n** разрядов имеет **m** информационных и **k = n - m** контрольных разрядов.

Корректирующее число длиной **k** разрядов описывает $2^k$ состояний, соответствующих отсутствию ошибки и появлению ошибки в одном разряде.

Должно соблюдаться соотношение

$$2^k = n - 1 \text{ или } 2^k - k + 1 = m$$

**Пример**

Если в оперативную память одновременно записываются или считываются 8 информационных байт (64 разряда), то при использовании кода Хэмминга потребуется 7 дополнительных контрольных разрядов.

**Синдром** — набор контрольных сумм информационных и проверочных разрядов.

# Модифицированный код Хэмминга

К контрольным разрядам кода Хэмминга добавляется еще один *разряд контроля четности* всех одновременно считываемых (записываемых) информационных и контрольных разрядов.

**Модифицированный код Хэмминга** позволяет устранять одиночные и обнаруживать двойные ошибки.

**Пример**

Пусть X — слово, записанное в оперативную память, а X' — считанное из оперативной памяти слово, в котором обнаружены две ошибки.

В неисправную ячейку оперативной памяти записывается обратный код считанного слова  и затем производится его считывание Y'.

Коды X' и Y' складываются по модулю 2.

Полученный код Z содержит 1 в разрядах, в которых имеются ошибки. Схемы управления оперативной памяти по коду Z корректируют одну ошибку. Затем схема коррекции одной ошибки исправляет вторую ошибку.

# Понятие сжатия данных

**Сжатие данных** — процесс, обеспечивающий уменьшение объёма данных путём сокращения их избыточности.

**Сжатие данных** — частный случай кодирования данных.

**Коэффициент сжатия** — отношение размера входного потока к размеру выходного потока.

**Отношение сжатия** — отношение размера выходного потока к размеру входного потока.

**Пример**
Размер входного потока равен 500 бит. Размер выходного равен 400 бит.
Коэффициент сжатия = 500 бит / 400 бит = 1,25.
Отношение сжатия = 400 бит / 500 бит = 0,8.

**Случайные данные невозможно сжать,
так как в них нет избыточности.**

# Методы сжатия данных

**Цель сжатия** — уменьшение количества бит, необходимых для хранения или передачи заданной информации.

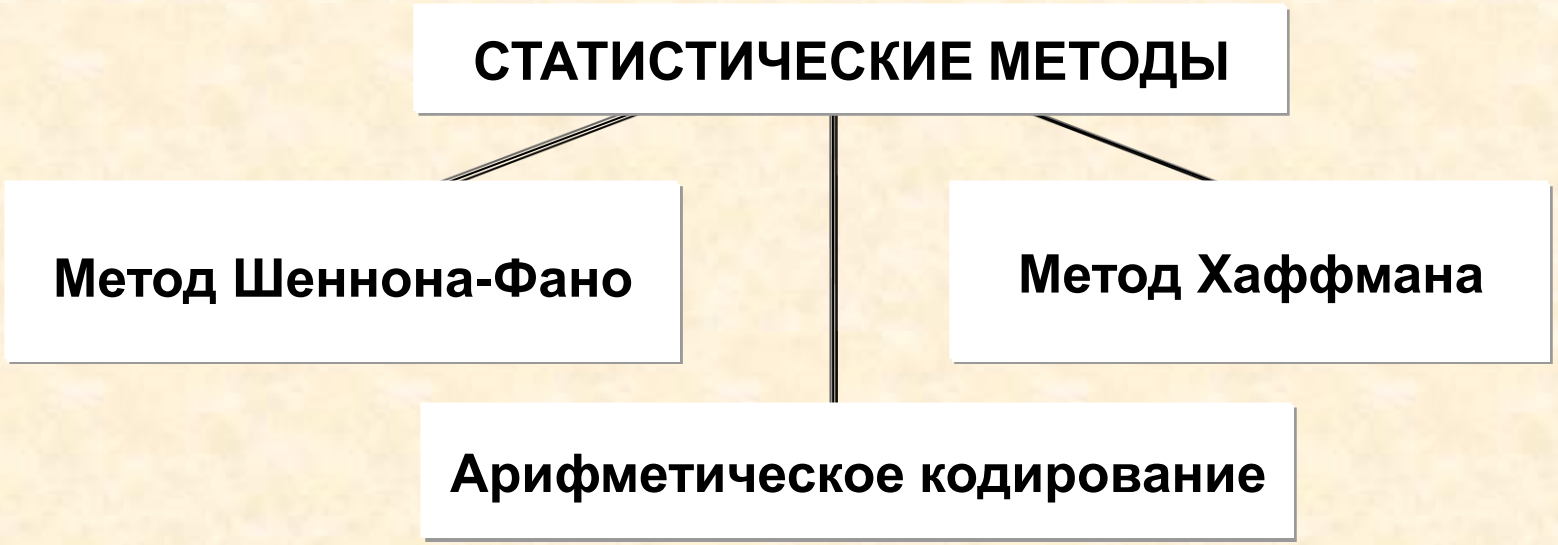
**Результат сжатия** — возможность передавать сообщения более быстро и хранить более экономно.

**Статистические методы** — кодирование с помощью усреднения вероятности появления элементов в закодированной последовательности.

**Словарные методы** — использование статистической модели данных для разбиения данных на слова с последующей заменой на их индексы в словаре.

# Сжатие без потерь

**Сжатие без потерь** (полностью обратимое) — сжатые данные после декодирования (распаковки) не отличаются от исходных.



# Сжатие с потерями

**Сжатие с потерями** (частично обратимое) — сжатые данные после декодирования (распаковки) отличаются от исходных, так как при сжатии часть исходных данных была отброшена для увеличения коэффициента сжатия.

Сжатие с потерями используется в основном для трех видов данных:
- полноцветная графика ($2^{24} \approx 16$ млн. цветов)
- звук
- видеоинформация

На первом этапе сжатия исходная информация приводится (с потерями) к виду, в котором ее можно эффективно сжимать алгоритмами второго этапа сжатия без потерь.

# Стандарты сжатия информации с потерями

Для сжатия **графической информации** установлен единый стандарт — формат **JPEG** (Joint Photographic Experts Group). В этом формате можно регулировать степень сжатия, задавая степень потери качества.

Сжатая **видеоинформация** представляет собой запись некоторых базовых кадров и последовательности изменений в них. Сжатую с потерями информацию можно сжимать далее другими методами. Наиболее распространенными являются стандарты **MPEG** (Motion Picture Experts Group).

Стандарт для сжатия **аудиоинформации** — **MPEG** без видеоданных. Стандарт **LPC** (Linear Predictive Coding) используется для сжатия речи.