И. Б. Жилина,
заслуженный учитель Российской Федерации, учитель информатики и ИКТ средней общеобразовательной школы № 814, Москва,

С. А. Жилин,
учитель информатики и ИКТ средней общеобразовательной школы № 887, Москва

ДВА МИФА О КОЛИЧЕСТВЕ ИНФОРМАЦИИ

Вопрос о количестве информации и методах его измерения обычно не считается очень трудным для изучения, особенно после знакомства с двоичной системой счисления и принципом двоичного кодирования информации. Но эта тема кажется простой только на первый взгляд. На самом деле именно здесь мы наблюдаем целый ряд серьезных заблуждений, которые часто встречаются в учебной практике. Мы будем условно называть эти заблуждения мифами. Поговорим о самых распространенных.

Миф 1. *Количество информации — это размер набора данных, который используется для записи информации в памяти компьютера или передачи по каналу связи. Соответственно, если измерить объем занятой памяти, то можно определить и количество информации.*

Проблему можно выразить в форме вопроса: если информация занимает в памяти компьютера 100 бит, то означает ли это, что мы храним в памяти именно 100 бит информации? Очень многие отвечают на этот вопрос утвердительно, и в этом заключается суть первого заблуждения. Причина ошибки в том, что два важнейших понятия информатики — «данные» и «информация» — часто и необоснованно считаются совершенно тождественными. Иногда это действительно не может привести к недоразумению. Но бывают ситуации, когда от четкого разделения этих понятий зависит не только понимание смысла задачи, но и сама возможность решения.

Для примера мы обычно предлагаем ученикам задачу, которая составлена на основе одного события из истории Древнего мира. В 490 г. до н. э. около небольшого греческого селения Марафон войска Афин победили армию персидского царя Дария. До наших дней дошла красивая легенда, связанная с этим событием. Полководец греков Мильтиад приказал воину бежать в Афины и сообщить о победе. Афинский воин пробежал по горным тропам 42 километра. У городских ворот он успел сказать только две фразы. «Радуйтесь, афиняне! Мы победили!» После этих слов он упал замертво.

Нам интересно посмотреть на эту ситуацию с точки зрения информатики. Сколько информации принес афинский воин в своем сообщении?

Если не различать информацию и данные, то у этой задачи вообще не может быть единственного решения. Данные — это конкретная форма представления информации, которая используется для ее записи в памяти технического устройства или передачи по каналу связи. Для вычисления количества данных в любом сообщении надо определить форму представления и кодирования информации. Таких форм у любой информации может быть очень много.

Теперь немного фантазии. У нас есть машина времени, и мы с ее помощью перемещаемся к воротам Афин за несколько минут до появления воина из Марафона. Конечно, не одни, а со всей необходимой техникой и программными средствами. Какие у нас есть способы записи информации, которую через несколько минут принесет воин? Первый метод. Мы записываем речь воина с помощью микрофона и программы звукозаписи. На диске получается запись сообщения в формате звукового файла. На этом этапе уже обозначилась первая трудность данного подхода к измерению информации. Звук можно записать в разных цифровых форматах, и для каждого из них мы будем получать разную длину кода. О единственном решении задачи уже можно забыть. Второй метод. Сообщение воина записываем в форме текста. Для полной точности это можно сделать на его родном греческом языке.

Запись текста — это тоже другой объем занятой памяти, т. е. другое количество данных. На этапе кодирования символов опять возникает альтернатива: какой набор символов выбрать? При использовании набора ASCII каждый символ будет представлен восьмиразрядным двоичным числом и займет в памяти 8 бит. Но если выбрать набор символов международной системы Unicode, то для каждой буквы греческого алфавита требуется уже 16 бит памяти.

Таким образом, если определять количество информации по размеру занятой памяти, то поставленную задачу решить невозможно. Точнее, невозможно решить единственным способом. Если хорошо подумать, то ситуация выглядит немного странно. Содержание сообщения не меняется, а количество информации получается разное. И зависит это количество не от предмета сообщения, а от формы записи. По смыслу это очень похоже на совершенно невозможную ситуацию, когда масса пакета картошки зависит не от массы картошки, а от формы пакета, в котором она продается.

Можем мы все-таки найти точное и единственное решение задачи или нет? Можем. Путь к решению содержится в самом условии задачи. Нас интересует не размер данных, а количество информации. Поэтому искать надо не любую форму кода, а наиболее простую и короткую из всех возможных, которые позволяют передать информацию без потери ее смыслового значения. Информация в сообщении афинского воина — это информация о победе или поражении, т. е. результат выбора одной из двух возможностей. Для передачи полной информации о таком событии можно использовать один символ из двоичного набора. Афиняне могли договориться о двоичном кодировании информации с помощью двух обыкновенных камней. Если победили греки, то воин должен принести и бросить у городских ворот камень белого цвета, в случае поражения камень должен быть черного цвета. Правила кодирования информации можно выразить следующим образом:

Цвет камня	**Код сообщения**	**Значение сообщения**
Белый	1	Радуйтесь, мы победили.
Черный	0	Горе, мы проиграли.

После получения белого камня жители Афин получают полную информацию о результате битвы. Более короткую форму сообщения для передачи информации найти невозможно. Даже если разделить камень на две половины, то для передачи информации все равно будет достаточно только одного камня. Для полного решения задачи необходимо допустить еще одну простую и очевидную вещь: шансы греков и персов на победу следует считать примерно равными. Это дополнительное условие не очень противоречит исторической реальности: обычно перед сражением очень трудно предсказать, чем оно закончится. Количество информации в любом сообщении, которое выбирается с одинаковой вероятностью из двух возможных, — это основная единица количества информации, бит (bit). Таким образом, решение задачи получено: афиняне получили от воина один бит информации.

Разницу между количеством данных и количеством информации можно понять еще лучше, если представить, что воин положил в мешок 16 белых и 16 черных камней. У городской стены он поворачивает, бежит вдоль стены и бросает на землю белые камни. Все 16 штук. Чтобы больше жителей могли увидеть, чем закончилась битва с персами. Это сообщение можно записать в форме шестнадцатиразрядного двоичного числа: «1111111111111111». При этом возникает вполне закономерный вопрос: «Сообщил ли воин таким образом еще какую-то дополнительную информацию, кроме информации о результате сражения?» Конечно нет. У нас нет совершенно никаких оснований считать, что произошло увеличение количества информации. Это на самом деле так. Дублирование единиц не увеличило и не могло увеличить количество информации в сообщении. Количество данных действительно стало больше, но количество информации осталось прежним. При выборе одного из двух равновероятных сообщений мы можем получить только 1 бит информации. Независимо от длины сообщений и метода кодирования.

От чего же зависит количество информации? Прежде всего от количества возможных сообщений. Математики в таких случаях говорят: зависит от мощности множества. Чем больше количество сообщений, тем больше количество информации. Это легко увидеть с помощью небольшого усложнения нашего учебного примера.

Пусть сражение с персами имеет не два, а три возможных исхода. Кроме победы и поражения может быть третий вариант: армии несут примерно равные потери и уходят на прежние позиции, по всем признакам это ничья. Метод кодирования информации в новых условиях может иметь следующий вид:

Цвет камня	Код сообщения	Значение сообщения
Белый	1	Радуйтесь, мы победили.
Черный	0	Горе, мы проиграли.
Белый и черный	10	Никто не победил, ничья.

У этой схемы есть один недостаток. При передаче данных о нескольких сражениях невозможно выделить из потока двоичных чисел отдельные сообщения одним-единственным способом, невозможно однозначно декодировать информацию. Например, по значению набора данных вида «1010» трудно понять, какие сообщения переданы. Это две победы и два поражения или два ничейных исхода подряд? Для однозначного выделения сообщений надо совсем немного изменить метод кодирования. Например, так:

Цвет камня	Код сообщения	Значение сообщения
Белый	1	Радуйтесь, мы победили.
Два черных	00	Горе, мы проиграли.
Черный и белый	01	Никто не победил, ничья.

Теперь из любого двоичного слова можно выделить все элементарные сообщения одним-единственным способом. Например, последовательность вида «0010001111» может означать только один набор сообщений: проиграли, победили, проиграли, ничья, победили, победили, победили.

Таким образом, для кодирования трех сообщений одной двоичной цифры уже мало. Это вполне очевидное свойство любой информации. Чем больше сообщений мы получаем от источника, тем больше символов придется использовать для кодирования отдельных сообщений. При этом одни сообщения могут быть более короткими, другие — более длинными. Но средняя длина кодовых слов обязательно будет расти. Эту среднюю величину называют ценой кодирования. Цена кодирования — это очень полезный показатель. Но это не количество информации. Средняя длина сообщения зависит от правил кодирования и может быть больше или меньше при одинаковом смысловом содержании информации.

Количество информации может зависеть только от самой природы источника информации и ни от чего другого. Такую числовую меру нашли американские ученые К. Шеннон и У. Уивер. Для определения среднего количества информации в каждом сообщении, которое поступает от источника, они предложили использовать специальную функцию состояния, которой дали название «энтропия». Энтропия — это универсальная мера разнообразия системы и неопределенности наших знаний о ее возможном состоянии. Чем больше различных сообщений мы можем получить от источника информации и чем труднее угадывать заранее содержание этих сообщений, тем больше будет значение энтропии.

Эту функцию и название для нее придумал и использовал для описания тепловых явлений немецкий физик Р. Клаузиус. Было это в середине XIX в. Австрийский физик Л. Больцман нашел связь между значением энтропии и вероятностью состояния физической системы. Шеннон и Уивер применили энтропию для описания статистических свойств каналов передачи информации. Итоги своих исследований они представили в 1949 г. в работе «Математическая теория связи» [3]. Ма-

териал этой книги стал основой новой науки — теории информации. Основные выводы ученых можно выразить в следующих утверждениях.

Информация и энтропия любой системы связаны между собой и количественно эквивалентны. Получение информации сопровождается равносильным уменьшением энтропии. Обратное тоже верно: рост энтропии связан с потерей информации. Естественно, энтропию и информацию можно измерять в одних единицах.

Выражение энтропии у Шеннона и Уивера было дано в следующей форме:

$$H = K\sum_{i=1}^{n} p_i \log p_i$$

Здесь:

H — энтропия источника информации (набора сообщений);

n — количество всех возможных сообщений;

p_i — вероятность появления сообщения с номером i;

log — функция логарифма;

K — коэффициент пропорциональности, который зависит только от выбора единиц измерения и основания логарифма.

Энтропия Шеннона—Уивера выражает среднее количество информации в одном сообщении. Чему в таком случае равно количество информации в одном конкретном сообщении? Оно равно отрицательному значению логарифма вероятности данного сообщения. Обращение знака логарифма требуется только для того, чтобы исключить отрицательные значения (логарифм вероятности всегда будет не больше нуля).

Из формулы хорошо видно, что величина энтропии не зависит от особенностей кодирования. Она зависит только от количества возможных сообщений и от вероятности появления каждого сообщения в наборе данных. При этом количество и вероятности сообщений зависят только от природы тех явлений, которые образуют источник информации. От чего зависят шансы греков победить в сражении? От военного мастерства, качества вооружения, моральной стойкости воинов. От кодовых таблиц и форматов данных эти шансы не зависят. Соответственно, не зависит и количество информации. Шеннон и Уивер не только заметили это важное свойство энтропии, но и доказали теорему о связи между ценой кодирования и средним количеством информации в одном сообщении.

Теорема. При любом способе кодирования данных среднее количество символов в одном сообщении не может быть меньше, чем среднее количество информации (энтропия источника информации).

Эта теорема позволяет провести четкую границу между количеством данных и количеством информации. Количество информации в любом наборе данных — это тот минимальный предел, к которому стремится количество данных при наиболее эффективном способе кодирования. В данном случае эффективность означает, что данные должны иметь наименьшую длину. Когда пользователь компьютера выполняет архивацию файла, то он сталкивается именно с этим явлением. Программа архивации может уменьшить размер файла не больше, чем на величину разницы между количеством данных и количеством информации. После сжатия размер файла уменьшается. Куда при этом исчезает часть информации? Никуда. Всё дело в том, что этого количества информации в файле никогда не было. Это был просто избыток данных, который образовался при кодировании.

Определим количество информации в сообщении афинского воина с помощью энтропии Шеннона—Уивера. Количество возможных сообщений равно двум. Шансы греков победить или проиграть будем считать равными. На языке теории вероятностей это означает, что вероятность победы армии составляет одну вторую и равна вероятности поражения. В качестве основания логарифма используем число 2, а для коэффициента выбираем значение −1. Это позволяет сделать значение энтропии неотрицательным числом и выразить его в двоичных единицах — битах.

$$H = -0{,}5 \log_2 0{,}5 - 0{,}5 \log_2 0{,}5 = -0{,}5 \cdot (-1) - 0{,}5 \cdot (-1) = 0{,}5 + 0{,}5 = 1 \text{ (бит)}.$$

Энтропия марафонской битвы как источника информации составляет 1 бит. После получения сообщения от воина неопределенность знаний греков о возможном исходе сражения полностью исчезает, «снимается». Исход сражения становится полностью определенным, а значение энтропии — нулевым ($\log_2 1 = 0$). Таким образом, после получения сообщения энтропия стала меньше на 1 бит. Это и есть то количество информации, которое получили афиняне. Методы теории информации полностью подтвердили наши выводы, полученные ранее с помощью эффективного кодирования.

Один бит — это очень маленькое количество информации. Энтропия одной буквы в нашем алфавите составляет примерно 4,35 бит. Правила системы Unicode разрешают использовать для кодирования одного символа от 8 до 32 бит. Восемь бит информации называют байтом (byte). Но и байт не очень удобная единица для измерения большого количества информации. Часто приходится использовать более крупные, кратные единицы измерения. К сожалению, в этой области тоже накопилось много проблем и ошибок. Об этом — миф номер два.

Миф 2. *Обычные правила для получения кратных единиц измерения величин не действуют в информационных технологиях. При использовании приставок надо всегда умножать основную величину не на степень числа 10, а на степень числа 2, которая наиболее близка к соответствующей десятичной степени. Например, многие уверены, что обозначения «килобит» и «Кбит» означают одну и ту же величину, которая составляет 1024 бит (2^{10}).*

Разумеется, это не так. Точнее — не совсем так. Давайте разберемся, в чем тут дело. Сначала о приставках. Международная система единиц измерения величин СИ (SI) устанавливает специальные приставки для получения кратных и дольных единиц измерения во всех областях науки и техники. Эти приставки имеют полные наименования и сокращенные обозначения и позволяют умножать значение основной единицы на определенную степень числа 10. Для удобства мы будем называть эти приставки десятичными. Приведем наиболее важные десятичные приставки.

Название приставки	Сокращение приставки	Значение приставки
кило	к	$10^3 = 1000$
мега	М	$10^6 = 1\,000\,000$
гига	Г	$10^9 = 1\,000\,000\,000$
тера	Т	$10^{12} = 1\,000\,000\,000\,000$
пета	П	$10^{15} = 1\,000\,000\,000\,000\,000$

Ни в одной области науки и техники эти приставки не могут иметь другие значения. И компьютерные технологии тоже не могут отменить правила системы СИ. Убедиться в этом не очень трудно. Обратимся к официальному документу. На территории Российской Федерации таким документом является «Государственная система обеспечения единства измерений. Единицы величин». Этот государственный стандарт действует с 1 сентября 2003 г. и имеет номер 8.417-2002. В приложении А данного стандарта определяются две основные единицы для измерения количества информации: «бит» и «байт». Про кратные единицы ничего не говорится, но имеется важное примечание. Приведем текст этого примечания полностью: «В соответствии с международным стандартом МЭК 60027-2 единицы «бит» и «байт» применяют с приставками СИ». Смысл примечания понятен — все приставки СИ, которые мы употребляем при измерении количества информации, должны соответствовать правилам этой системы единиц. Это значит, что в одном килобите может быть только одна тысяча бит, а в одном килобайте — только одна тысяча байт. Разумеется, то же самое относится к мегабитам, мегабайтам и т. д.

Проблема с приставками возникла из-за того, что в ряде областей информатики удобнее применять приставки не с десятичным значением, а с двоичным. Например, при измерении объема оперативной памяти компьютера. Производители микросхем оперативной памяти обычно указывают емкость схемы в Мбитах. Марки-

ровка вида 64Mx8 означает, что емкость составляет 512 Мбит (64 · 8 = 512). Буква М означает здесь вовсе не миллион, а два в двадцатой степени, т. е. степень двух, наиболее близкую к шестой степени десяти. Эта величина больше, чем миллион, и составляет 1 048 576. Соответственно, емкость в битах надо вычислять следующим образом:

512 · 1 048 576 = 536 870 912 (бит).

Емкость устройства или одного модуля оперативной памяти обычно выражают в более крупных единицах: Мбайтах или Гбайтах. И здесь символы М и Г тоже надо использовать с двоичным значением. Причина использования двоичных приставок заключается в том, что адрес байта в электронных запоминающих устройствах задается в форме целого двоичного числа. Поэтому количество байтов в микросхемах и модулях памяти удобно делать таким, чтобы значение этого количества было степенью двух.

На первом этапе развития компьютерных технологий никаких особых проблем с двоичными приставками не было. Для обозначения умножения на 1024 (два в десятой степени) была выбрана большая буква К. Например, емкость памяти 640К байт означала 655 360 байт (640 · 1024). Никаких противоречий это не вызывало, так как в СИ для умножения на тысячу используется не большая, а маленькая буква. Например, в сокращении «кг» (килограмм).

Проблемы с двоичными приставками начались позже. Первая причина — это быстрый рост размеров памяти компьютера. Возникла потребность в более крупных приставках. Появились двоичные приставки «М», «Г», «Т», и эти обозначения были выбраны неудачно — они полностью совпали с десятичными приставками СИ. Но была еще вторая причина, более серьезная. Кто-то решил для удобства называть двоичные кратные приставки так же, как это принято для десятичных. Таким образом, «Кбайты» стали «килобайтами», «Мбайты» — «мегабайтами» и т. д. Эта «вредная» привычка очень быстро распространилась среди пользователей.

Но это еще не всё. На ситуацию с приставками очень негативно влияет еще одна вещь. В информационных технологиях есть по меньшей мере две области, в которых правила системы СИ выполняются совершенно точно. Это производство жестких магнитных дисков и системы передачи данных (телекоммуникации).

При измерении емкости жестких дисков мегабайты и гигабайты являются десятичными единицами. Емкость диска на 200 гигабайт означает, что на этом диске можно записать примерно двести миллиардов байт. Двести обычных «десятичных миллиардов», а не тридцатых степеней двух. Причина, по которой производители жёстких дисков предпочитают правила СИ, не только техническая, но и коммерческая. Количество десятичных гигабайтов всегда будет больше, чем двоичных. В системах передачи данных основной единицей количества информации является бит. При этом двоичные множители не имеют никаких объективных преимуществ перед десятичными. Поэтому в этой области правила системы СИ тоже не отменяются. При передаче данных по каналу связи один «килобит» (кб) должен означать одну тысячу бит, а один «мегабит» (Мб) — один миллион бит.

Таким образом, в информационных технологиях образовалась большая путаница с приставками. При вычислении мегабитов в системах передачи данных надо умножать количество битов на миллион, а при вычислении мегабитов в оперативной памяти — на два в двадцатой степени. Двести гигабайт на жестком диске — это двести миллиардов байт, а два ГБ в оперативной памяти — это уже больше, чем два миллиарда. Представить себе что-нибудь подобное в других областях науки и техники совершенно невозможно. Разве может один километр содержать тысячу метров при измерении длины реки, но 1024 метра при измерении высоты горы?

Международная электротехническая комиссия (МЭК) сделала попытку покончить с этим «ужасным» положением. В ноябре 2000 г. были приняты поправки к международному стандарту МЭК 60027-2 («Телекоммуникация и электроника»). Суть их состоит в следующем. Приставки системы СИ для образования кратных единиц разрешается использовать только с десятичным значением. То есть в одном килобите может быть только тысяча бит, а в одном мегабайте — только миллион

байт. Для приставок с двоичным значением МЭК предлагает использовать следующее решение проблемы. Названия всех двоичных приставок меняются. От приставки СИ берутся только две первые буквы. К ним добавляется слог «би» (bi) — от английского «binary» («двоичный»). В результате образуется название новой приставки с двоичным значением.

Для наиболее распространенных двоичных приставок это должно выглядеть следующим образом:

Название приставки	Сокращение приставки	Значение приставки
киби	Ки	$2^{10} = 1024$
меби	Ми	$2^{20} = 1\,048\,576$
гиби	Ги	$2^{30} = 1\,073\,741\,824$
теби	Ти	$2^{40} = 1\,099\,511\,627\,776$
пеби	Пи	$2^{50} = 1\,125\,899\,906\,842\,624$

Теперь основы для противоречий больше нет. Двоичные приставки получили свои собственные названия и обозначения. Один кибибайт данных (КиБ) содержит 1024 байт. Два гибибайта оперативной памяти (2ГиБ) — это 2 147 483 648 байт. Но все-таки говорить об успешном решении проблемы еще рано. У новых правил МЭК есть один, но очень большой недостаток: никто не спешит их выполнять. Мешают многолетняя привычка и традиции языка. Уж очень неудобно и непривычно для многих пользователей звучат эти «кибибиты» и «мебибайты».

На своих учебных занятиях мы нашли выход из этого трудного положения. Там, где двоичное значение приставки не ясно из контекста задачи, мы даем специальный комментарий или добавляем прилагательное «двоичный». При этом часть условия может звучать примерно так: «объем данных составляет тридцать два двоичных килобайта». С одной стороны, ученики сразу понимают, какую приставку использовать при вычислении. С другой стороны, здесь нет слишком грубого нарушения правил системы СИ.

Теперь последнее, но очень важное замечание. В стране идет реформа образования, переход к активному использованию письменных форм массовой проверки знаний, таких, как единые государственные экзамены (ЕГЭ). Любая письменная форма аттестации предъявляет особо жесткие требования к точности и ясности во всех понятиях и определениях. Поэтому составители контрольно-измерительных заданий по информатике и авторы учебных пособий обязаны с предельным вниманием относиться ко всем вопросам, где причиной ошибки может быть неточность в условии задачи. Наиболее очевидные меры для наведения порядка в терминологии могут быть следующими. Первое. Пора полностью прекратить использовать десятичные приставки системы СИ с неправильным двоичным значением и предусмотреть необходимые изменения во всех учебных курсах, которые были созданы до появления нового стандарта МЭК. Второе. Понятие «количество информации» лучше использовать в том смысле, который определяется в рамках классической теории информации, и, по возможности, не смешивать его с понятием «количество данных». Особенно в тех задачах, где эти величины не являются эквивалентными.

Литературные и интернет-источники

1. *Винер Н.* Кибернетика и общество. М.: Изд-во иностр. лит-ры, 1958.
2. *Панин В. В.* Основы теории информации. М.: БИНОМ. Лаборатория знаний, 2007.
3. *Shannon C. E., Weaver W.* The Mathematical Theory of Communication. Urbana: University of Illinois Press, 1949.
4. http://nolik.ru/systems/gost.htm
5. http://www.gost.ru