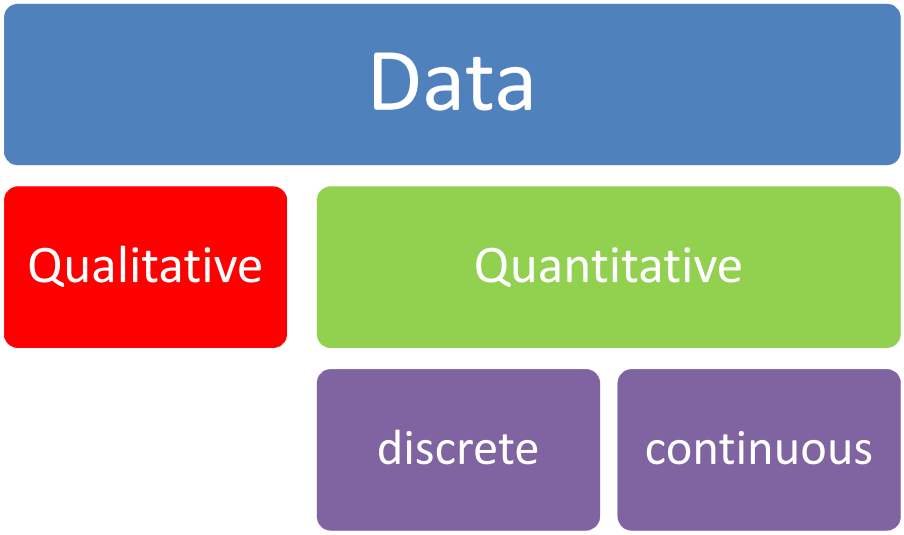
Data



Qualitative data – not given numerically (favourite colour, the most hated food)

Quantitave data – in the numeric form

1. **Discrete –** only certain values (integers, rational numbers). Examples:
    a. number of customers or empolyees
    b. parcel count
    c. number of routers
2. **Continuous** - representation of all values within an interval. There are no limits to the gaps between the measurements. Examples:
    a. height
    b. area
    c. voltage

Types of data

- Numerical, textual , audio, visual, mixed

Computers store all data in 0s and 1s – there must be a rule, how to represent letters, numbers, colours, sound etc. in the binary form – ***code***.

Computer Code = a rule defining the binary representation of a certain type of data.

Example:

0011 0101 0111 0001 1101 0110 (53, 113, 214 in decimal) can be:

- symbols **5**, ***q***, **Ö** *in extended ASCII (version ISO 8859-2)*
- *a pixel of the colour #3571D6*
- *3 samples of the sound recorded in 8-bit resolution.*

It all depends on the code applied to the binary values.

Alphanumerical/Textual Codes

A binary digit or bit can represent only two symbols as it has only two states '0' or '1'. But this is not enough for communication between two computers because there we need many more symbols for communication. These symbols are required to represent 26 alphabets with capital and small letters, numbers from 0 to 9, punctuation marks and other symbols.

The alphanumeric codes are the codes that represent numbers and alphabetic characters. Mostly such codes also represent other characters such as symbol and various instructions necessary for conveying information. An alphanumeric code should at least represent 10 digits and 26 letters of alphabet i.e. total 36 items. The following alphanumeric codes are very commonly used for the data representation:

- American Standard Code for Information Interchange (**ASCII**) – 7 bits per symbol.
- Extended ASCII – 8 bits per symbol.

Code point – a numerical value, which can represent a character, symbol, or a formatting rule.

ASCII Code

- *American Standard Code for Information Interchange*
- Original ASCII: 7 bits – defined by telecommunication industry (1st version: 1963; the latest one in 1986) – **128 code points.**
- First 33 symbols: control (non-printable) characters – many of them are obsolete; the rest – printable characters.

Example: in Windows the ENTER is stored as a combination of 2 ASCII symbols: symbol no. 13 (CR – carriage return) and symbol no. 10 (LF – line feed) – originally instructions for automatic typewriters: carriage return – go back to the beginning of the line; line feed – scroll the drum for a new line.

Modern computers store data in bytes[1] – 1B = 8 bits – the range is doubled → **extended ASCII** with **256 code points** – a table, where the first half is the original ASCII; the second half had a lot of versions, e.g. in Slovakia.

- Latin2 (ISO 8859-2)
- Windows Latin2 (Windows 1250)
- Kamenicky Code (895)

Those variants brought compatibility problems – files, which used symbols from the upper half, where displayed/printed differently.

Example:

*Code point $BE_{16} = 190_{10}$ in Latin 2 was the letter **ž**, while in Windows 1250 it was **ľ**. And in Latin 1 (which was used for the Western Europe) it was the symbol **¾**.*

Solution → UNICODE.

[1] Early systems used different lengths for data units equivalent to bytes, e.g. 7, 18 or 36 bits

UNICODE

- One code system for all symbols (*currently cca 100 000 symbols, it is ready for more than 1 million symbols*) - e.g. Latin, Cyrillic, Greek, Hebrew, Arabic, Hindu, Thai, Tamil, mathematical symbols …
- UNICODE encodes **graphemes** – characters instead of glyphs

Figure 1: Various glyphs of the grapheme A (source: http://scripts.sil.org/cms/scripts/page.php?site_id=nrsi&id=IWS-Chapter02)

A **grapheme** is a semantic unit representing an indivisible unit of text in memory.

A **glyph** is the visual representation of a character or sequence of characters.

A **font** is a collection of glyphs.

- Each grapheme has its UNICODE code point written in hexadecimal notation, e.g.:
  - A is U+0041
  - ž is U+017E
  - Æ is U+1D01
  - ∞ is U+221E
- There are multiple implementations of the UNICODE (how the code points are stored)
  - **UTF-32** – each grapheme has 32-bit long representation
  - **UTF-16** – graphemes have 1x or 2x 16-bit representation
  - **UTF-8** – graphemes may have representation by 1 up to 6 bytes.

UTF-8

- The most frequent version of UNICODE
- Structure

Bits of code point	First code point	Last code point	Bytes	Byte 1	Byte 2	Byte 3	Byte 4	Byte 5	Byte 6
7	U+0000	U+007F	1	0xxxxxxx					
11	U+0080	U+07FF	2	110xxxxx	10xxxxxx				
16	U+0800	U+FFFF	3	1110xxxx	10xxxxxx	10xxxxxx			
21	U+10000	U+1FFFFF	4	11110xxx	10xxxxxx	10xxxxxx	10xxxxxx		
26	U+200000	U+3FFFFFF	5	111110xx	10xxxxxx	10xxxxxx	10xxxxxx	10xxxxxx	
31	U+4000000	U+7FFFFFFF	6	1111110x	10xxxxxx	10xxxxxx	10xxxxxx	10xxxxxx	10xxxxxx

- Advantages:
  - Backward compatibility with ASCII
  - The most common symbols require only 1 or 2 Bytes
- Current standard for encoding of XML, HTML

Other Codes

Binary Coded Decimal (BCD) Code

This code is used to represent numeric values. Each decimal digit is represented by a 4-bit binary number. BCD is a way to express each of the decimal digits with a binary code. In the BCD, with four bits we can represent sixteen numbers (0000 to 1111). But in BCD code only first ten of these are used (0000 to 1001). The remaining six code combinations i.e. 1010 to 1111 are invalid in BCD.

Decimal Digit	0	1	2	3	4	5	6	7	8	9
BCD Code	0000	0001	0010	0011	0100	0101	0110	0111	1000	1001

Example:

4921 → 0100 1001 0010 0001

Pros:

- Similar to decimal system
- Easy to understand

Cons:

- Arithmetic for BCD has rules, which differ from the simple binary arithmetic
- BCD arithmetic is more complicated
- The representation of numbers is less efficient (4 bits for any digit) → e.g. 4921 in BCD has 16 bits, while its binary version takes 14 bits (1001100111001). It becomes more obvious with the increased length of the numbers.

Gray Code

A Gray Code represents numbers using a binary encoding scheme that groups a sequence of bits so that **only one bit in the group changes from the number before and after**. This is referred to as **Hamming distance of 1** between adjacent codes.

Example of the 3-bit Gray code:

Value	Gray Code
0	000
1	001
2	011
3	010
4	110
5	111
6	101
7	100

The length of the code depends on its use.

The Gray code is used in some systems, where mechanical position is translated into a digital value (e.g. altimeter (*výškomer*) in aircraft). Another use is in digital systems operating over satellites.

How to make Gray code of an arbitrary length[2]

Starting with bits in the first column:

0
1

Take the reflection, as if a mirror were held up to the column:

0
1
_____mirror
1
0

This results in a column with 4 entries, but the first and last are the same, as are the middle ones, so another column and alternate bits are added:

00 (added bits are in bold)
01
11
10

Then reflect that:

00
01
11
10
_____mirror
10
11
01
00

and add another column with alternate bits:

000
001
011
010
110
111
101
100

and it continues.

.

[2] https://www.allaboutcircuits.com/technical-articles/gray-code-basics/

Exercises

1. How many symbols can be encoded
    a. in the original ASCII?
    b. in the extended version?
    c. In the 3rd group of UTF-8 symbols?
2. You have got two messages in 8-bit ASCII. Try to decipher them!
    a. 84 104 97 116 115 32 111 110 101 32 115 109 97 108 108 32 115 116 101 112 32 102 111 114 32 97 32 109 97 110 32 111 110 101 32 103 105 97 110 116 32 108 101 97 112 32 102 111 114 32 109 97 110 107 105 110 100
    b. 01001001011001100010000001111001011011110111010100100000011000110110000010110111000100000001110010011001010110000010110010000100000011010010111010000101100001000000011110010110111011101010010000001100100011001010101110011011001010111001001110110011001010010000001101111011011100110010
3. Try to decipher the following UTF-8 messages:
    a. Ak chceU+0161 byU+0165 U+0161 U+0165astnU+00FD, vedz, U+017Ee to zU+00E1leU+017E U+00ED len od teba.
    b. Na%C5%A1e%20vyn%C3%A1lezy%20s%C3%BA%20zvy%C4%8Dajne%20iba%20pekn%C3%A9%20hra%C4%8Dky%2C%20ktor%C3%A9%20odvr%C3%A1tili%20na%C5%A1u%20pozornos%C5%A5%20od%20vec%C3%AD%20v%C3%A1%C5%BEnych. (*this type of encoding is used for URI/URL*)

Wordstock

character	znak
printable	tlačiteľný
grapheme	graféma
glyph	glyf