# Data Representation

Networks and Embedded Software

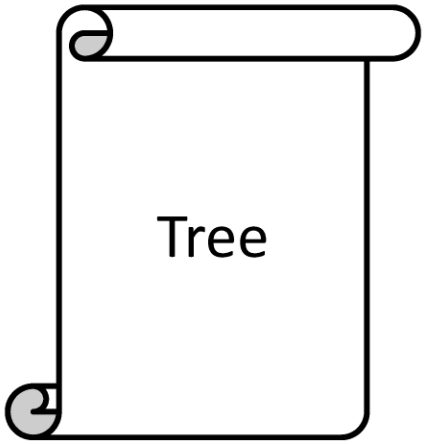
Module 5.4.1

by Wolfgang Neff

# Coding (1)

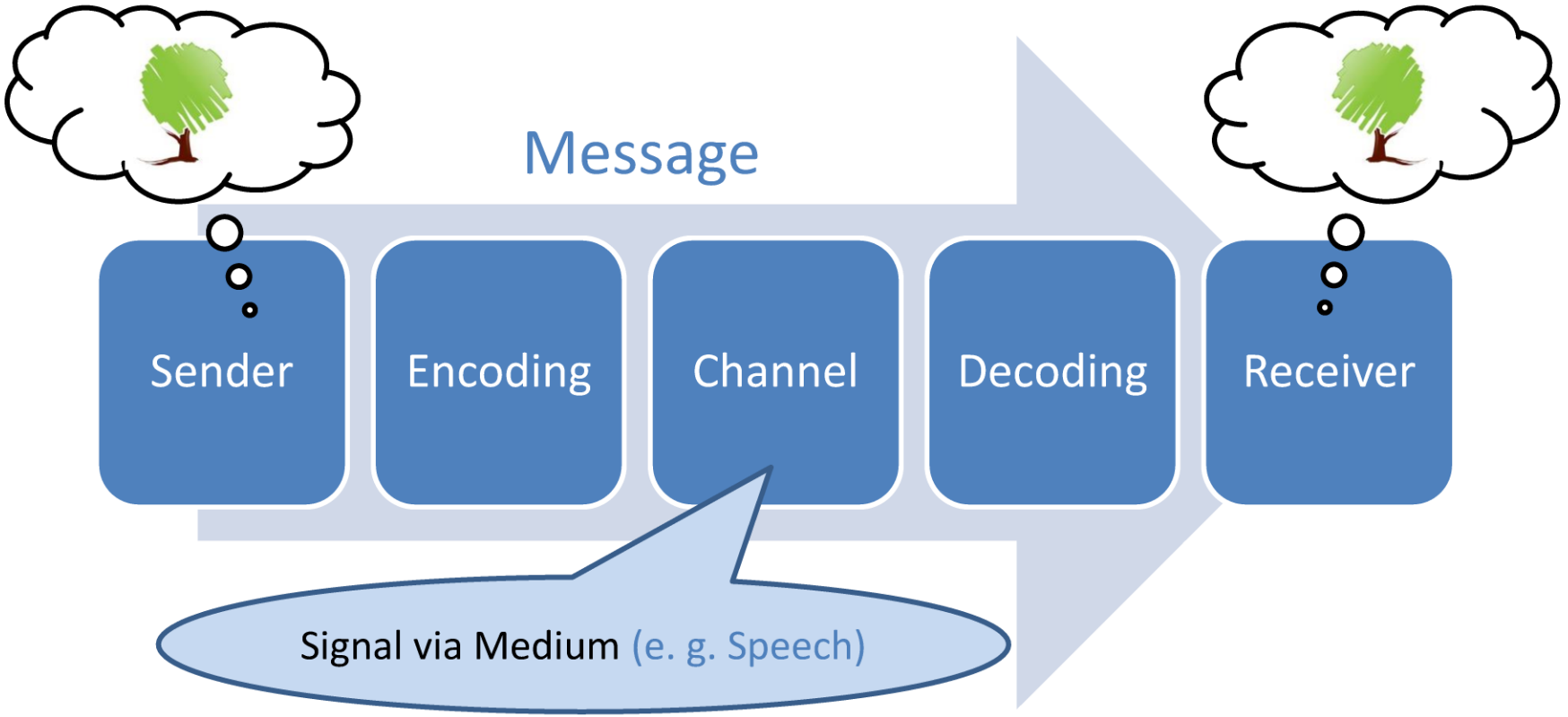
- Basic concepts
  - Information
    - Knowledge about something
    - Abstract concept (just in mind, can not be touched)

  - Data
    - Representation of information
    - Signals on a medium (characters on paper)
  - Coding
    - Write down information



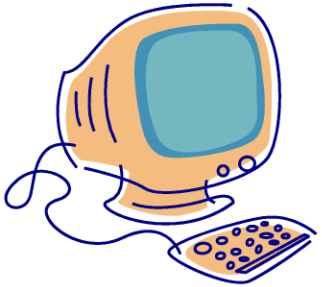
# Coding (2)

- Coding and communication are related

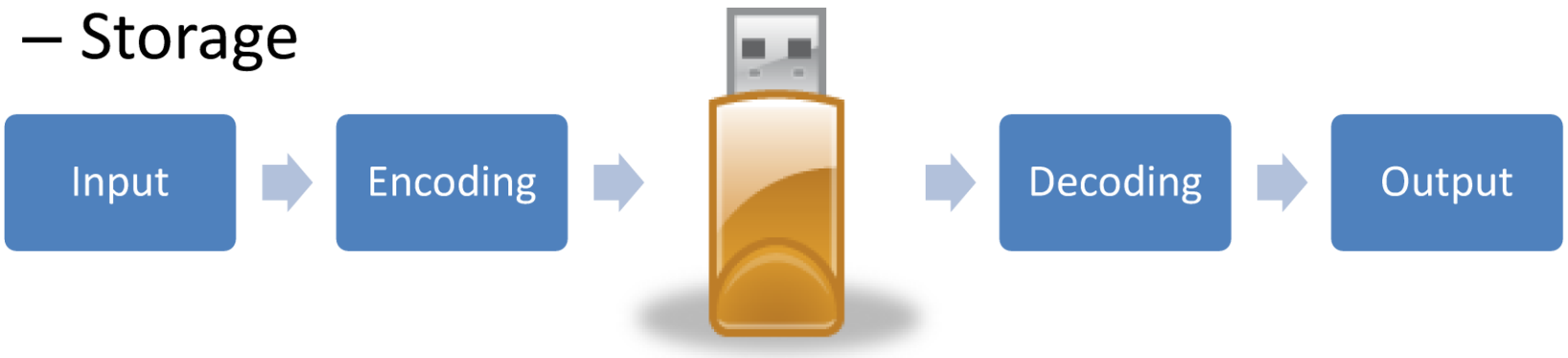


# Coding (3)

- Coding and computer science are related
  - IPO Model

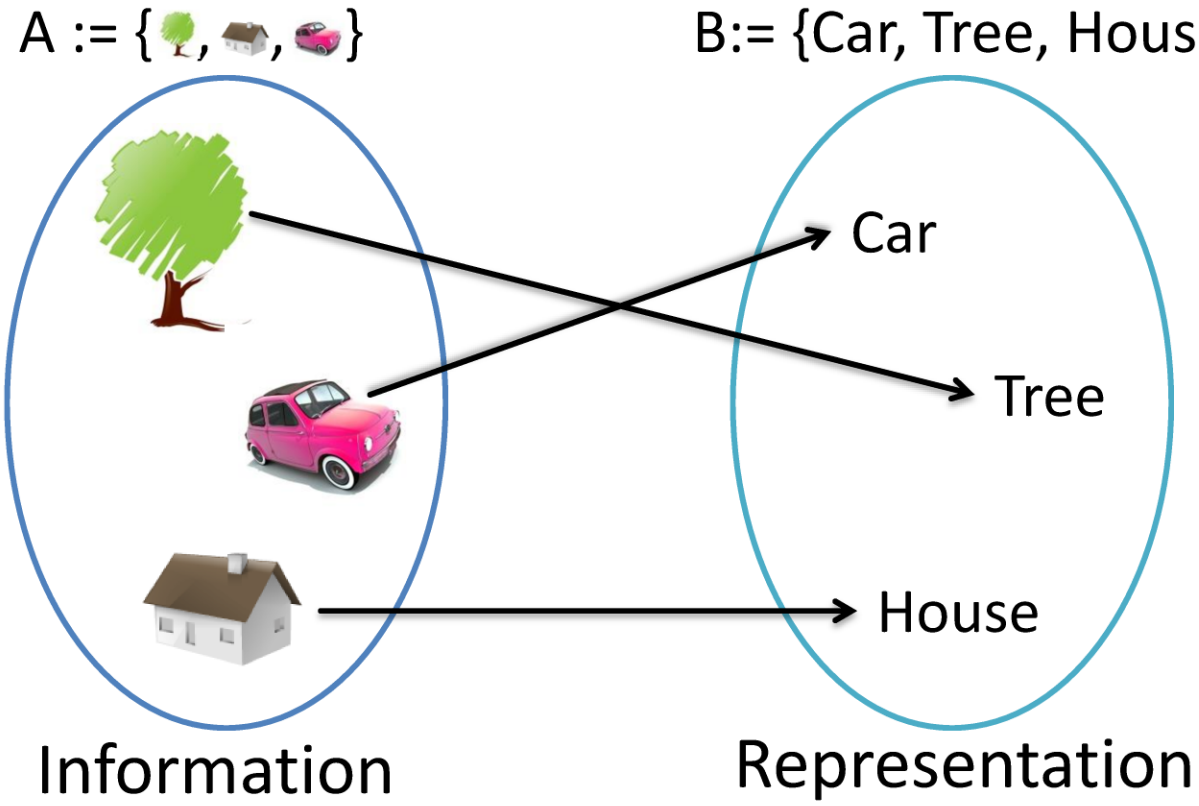



  - Storage



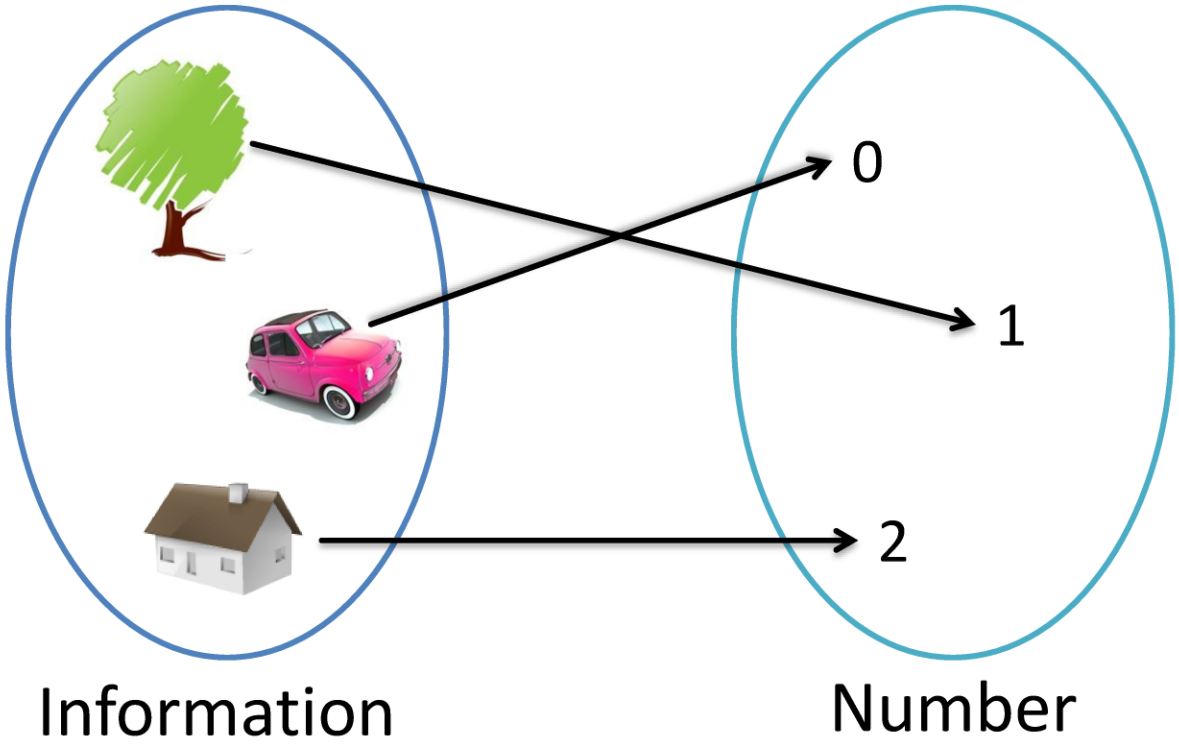
# Coding (4)

- Coding maps the elements of two sets



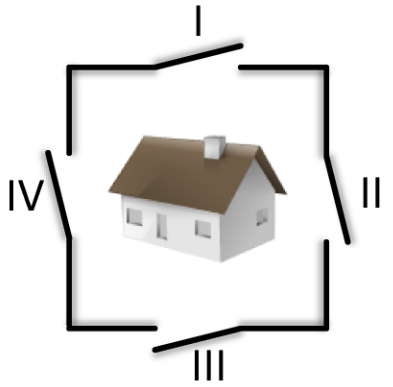
# Coding (5)

- Computers represent information by numbers



# Coding (6)

- Example
  - A door has the three states open, closed, locked
  - The house has four doors
  - We will code the state of the doors of the house
    - Binary coding of the states of a door
    - Hot one coding of the state of a door
    - Binary coding of the state of the house



# Coding (7)

- Binary coding of the states of a door
  - We need two bits to encode the state of the door

| State | First bit | Second Bit |
|---|---|---|
| Open | 0 | 0 |
| Close | 0 | 1 |
| Locked | 1 | 0 |
| Unused | 1 | 1 |

  - We need eight bits for the state of the house

| Door 1 | | Door 2 | | Door 3 | | Door 4 | |
|---|---|---|---|---|---|---|---|
| 0 | 0 | 0 | 1 | 1 | 0 | 1 | 0 |

# Coding (8)

- Hot one coding of the state of a door
  - We use one bit per state of the door

| State | First bit | Second Bit | Third Bit |
|---|---|---|---|
| Open | 0 | 0 | 1 |
| Close | 0 | 1 | 0 |
| Locked | 1 | 0 | 0 |

  - We need twelve bits for the state of the house

| Door 1 | | | Door 2 | | | Door 3 | | | Door 4 | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 0 | 0 | 1 | 0 | 1 | 0 | 1 | 0 | 0 | 1 | 0 | 0 |

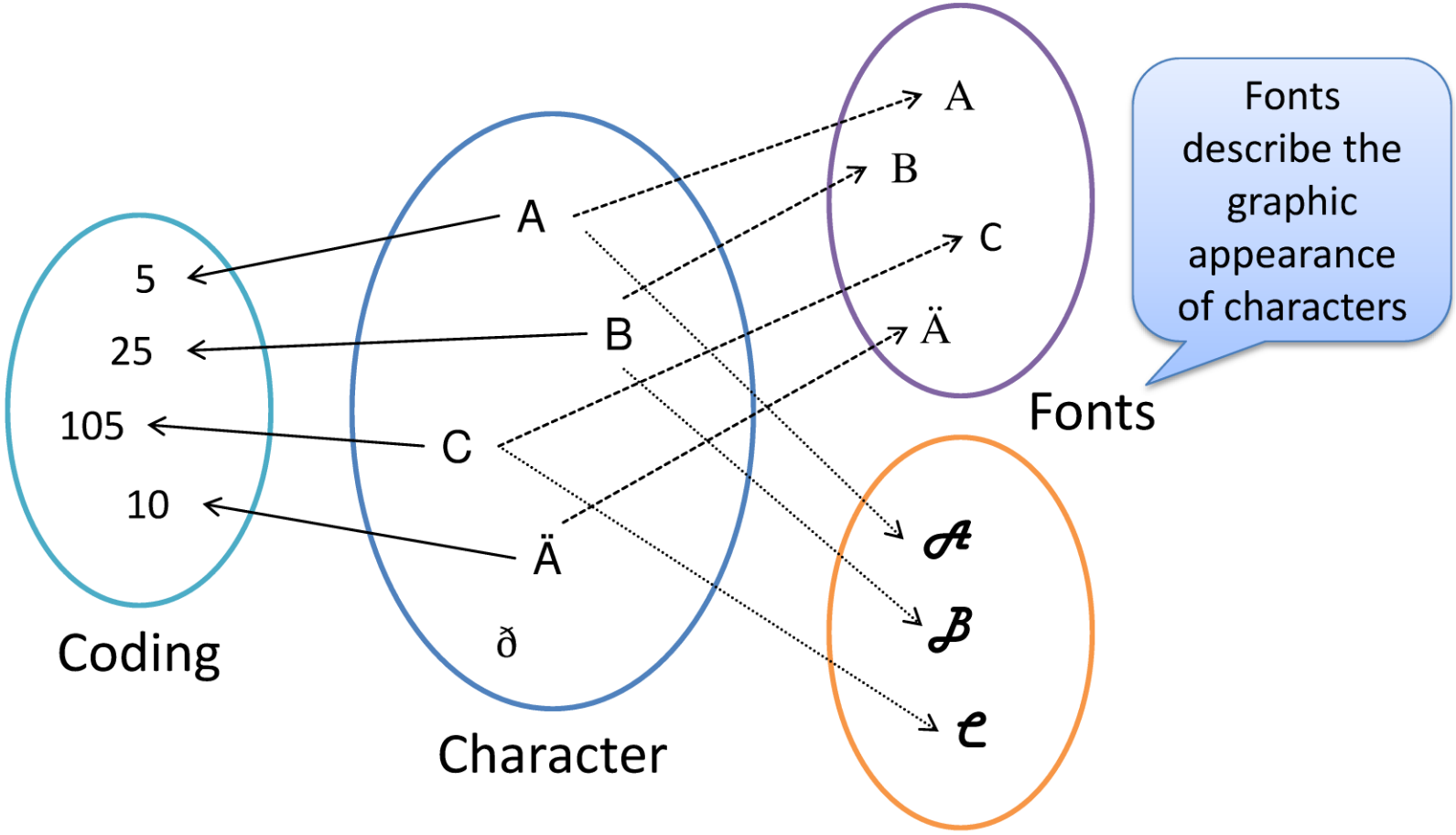
# Coding (9)

- Direct coding
  - A door has three states
    - O: open
    - C: closed
    - L: locked
  - Our house has $3^4 = 81$ states
  - We need seven bits to encode all states $2^7 > 81$
  - There are many don't-care terms

| n | Door I | Door II | Door III | Door IV |
|---|---|---|---|---|
| 0 | O | O | O | O |
| 1 | O | O | O | C |
| 2 | O | O | O | L |
| 3 | O | O | C | O |
| 4 | O | O | C | C |
| 5 | O | O | C | L |
| 6 | O | O | L | O |
| 7 | O | O | L | C |
| 8 | O | O | L | L |
| 9 | O | C | O | O |
| … | … | … | … | … |

# Characters (1)

- Characters are mapped to codes and fonts.



# Characters (2)

- Characters are encoded by numbers.
- Characters are displayed by fonts.
- The mapping need not be total.
  - There can be characters without code
  - There can be characters not present in a font
- There are many different fonts.
  - Times New Roman
  - *Harlow Solid Italic*

# Characters (3)

- ASCII: the ancestor
  - American Standard Code for Information Interchange
  - Published by ASA in 1963
  - Uses 7 bits per character
  - Contains
    - Control characters
    - Letters and Numbers
    - Punctuation and special characters

# Characters (4)

- The ASCII code table

| ASCII | | Lower Hex Digit | | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | 0 | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | A | B | C | D | E | F |
| Higher Hex Digit | 0 | NUL | SOH | STX | ETX | EOF | ENQ | ACK | BEL | BS | HT | LF | VT | FF | CR | SO | SI |
| | 1 | DLE | DC1 | DC2 | DC3 | DC4 | NAK | SYN | ETB | CAN | EM | SUB | ESC | FS | GS | RS | US |
| | 2 | | ! | " | # | $ | % | & | ' | ( | ) | * | + | , | - | . | / |
| | 3 | 0 | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | : | ; | < | = | > | ? |
| | 4 | @ | A | B | C | D | E | F | G | H | I | J | K | L | M | N | O |
| | 5 | P | Q | R | S | T | U | V | W | X | Y | Z | [ | \ | ] | ^ | _ |
| | 6 | ` | a | b | c | d | e | f | g | h | i | j | k | l | m | n | o |
| | 7 | p | q | r | s | t | u | v | w | x | y | z | { | \| | } | ~ | DEL |

# Characters (5)

- Latin-1: One nation – one character set
  - Deficiencies of ASCII
    - Suitable for English, only
    - Other languages use additional symbols
    - Only 7 bits of a byte used
  - ISO 8859
    - Uses all 8 bits of a byte for coding
    - 128 additional symbols
    - One standard per region

-1: Western European
-2: Central European
…
-7: Latin/Greek
…
-11: Latin/Thai
…

# Characters (6)

- The ISO 8859-1 code table: Latin 1

| Latin | | Lower Hex Digit | | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | 0 | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | A | B | C | D | E | F |
| Higher Hex Digit | … | … same as ASCII … | | | | | | | | | | | | | | | |
| | 8 | … unused … | | | | | | | | | | | | | | | |
| | 9 | | | | | | | | | | | | | | | | |
| | A | NBSP | ¡ | ¢ | £ | ¤ | ¥ | ¦ | § | ¨ | © | ª | « | ¬ | | ® | ¯ |
| | B | ° | ± | ² | ³ | ´ | µ | ¶ | · | ¸ | ¹ | º | » | ¼ | ½ | ¾ | ¿ |
| | C | À | Á | Â | Ã | Ä | Å | Æ | Ç | È | É | Ê | Ë | Ì | Í | Î | Ï |
| | D | Ð | Ñ | Ò | Ó | Ô | Õ | Ö | × | Ø | Ù | Ú | Û | Ü | Ý | Þ | ß |
| | E | à | á | â | ã | ä | å | æ | ç | è | é | ê | ë | ì | í | î | ï |
| | F | ð | ñ | ò | ó | ô | õ | ö | ÷ | ø | ù | ú | û | ü | ý | þ | ÿ |

# Characters (7)

- Unicode: One code for all nations
  - Consistent encoding of all writing system
  - Comprises modern and historic scripts
  - Consists of more than 110 000 characters
  - Uses 32 bits for coding
  - Standardized as ISO/IEC 10646
  - Special formats for storage (UTF)

# Characters (8)

- Unicode is or organized in planes.
  - Diacritics and Ligatures
    - à, å, ä, æ
  - Non-Latin writing systems
    - Arabic: العربية
  - Logographic writing system
    - Chinese: 汉语
  - Special purpose characters
    - Mathematics: ·, ×, ÷, ≈, ±, ∫, ∞

# Characters (9)

- Example: Greek and Coptic (U+0370-U+03FF)

| Code | 0 | 1 | 2 | 3 | 4 | 5 | 6 | 7 |
|---|---|---|---|---|---|---|---|---|
| 039 | ΐ | Α | Β | Γ | Δ | Ε | Ζ | Η |

  - U+0390: ΐ Greek small litter iota with …
  - U+0391: Α Greek capital letter alpha
  - U+0392: Β Greek capital letter beta
  - U+0393: Γ Greek capital letter gamma
  - U+0394: Δ Greek capital letter delta

# Characters (10)

- Unicode Transformation Format
  - Used to store and process Unicode characters
  - UTF-16
    - Used in Windows, OS X, Java, .Net
  - UTF-8
    - Used to store Unicode characters in files
    - Aims to reduce the size of Unicode files
    - Variable length 8 bit code
    - ASCII files are valid UTF-8 files

# Characters (11)

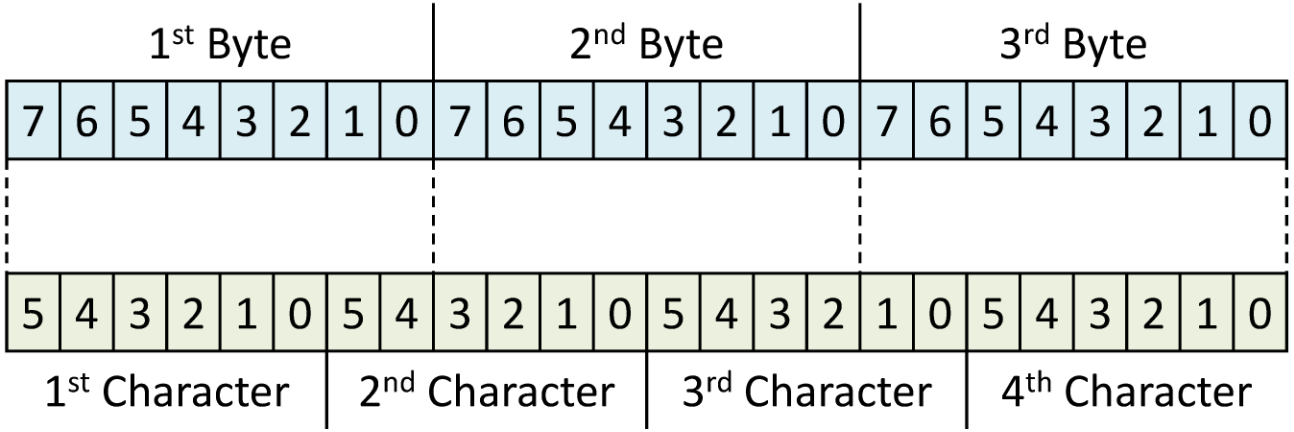
- UTF-8 Coding procedure
  - Unicode Character U+ xxxx xyyy yzzz zzzz
  - If each x and each y is 0
    - The code is **0**zzz zzzz (n. b. this is the ASCI code)
  - If each x is 0
    - The code is **11**0y yyyz **10**zz zzzz
  - Otherwise
    - The code is **1110** xxxx **10**xy yyyz **10**zz zzzz

# Characters (12)

- UTF-8 Coding examples
  - y → U+0079 → 00000 0000 1111001
    - Each x and each y is 0
    - y → 01111001
  - ä → U+00E4 → 00000 0001 1100100
    - Each x is 0
    - ä → 11000011 10100100
  - € → U+20AC → 00100 0001 0101100
    - € → 11100010 10000010 10101100

# Base64 (1)

- Base64 coding uses 64 digits
- Used in internet communication
- Three bytes are encoded as four ASCII chars



# Base64 (2)

- Coding table

| 00 | **A** | 08 | **I** | 10 | **Q** | 18 | **Y** | 20 | **g** | 28 | **o** | 30 | **w** | 38 | **4** |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 01 | **B** | 09 | **J** | 11 | **R** | 19 | **Z** | 21 | **h** | 29 | **p** | 31 | **x** | 39 | **5** |
| 02 | **C** | 0A | **K** | 12 | **S** | 1A | **a** | 22 | **i** | 2A | **q** | 32 | **y** | 3A | **6** |
| 03 | **D** | 0B | **L** | 13 | **T** | 1B | **b** | 23 | **j** | 2B | **r** | 33 | **z** | 3B | **7** |
| 04 | **E** | 0C | **M** | 14 | **U** | 1C | **c** | 24 | **k** | 2C | **s** | 34 | **0** | 3C | **8** |
| 05 | **F** | 0D | **N** | 15 | **V** | 1D | **d** | 25 | **l** | 2D | **t** | 35 | **1** | 3D | **9** |
| 06 | **G** | 0E | **O** | 16 | **W** | 1E | **e** | 26 | **m** | 2E | **u** | 36 | **2** | 3E | **+** |
| 07 | **H** | 0F | **P** | 17 | **X** | 1F | **f** | 27 | **n** | 2F | **v** | 37 | **3** | 3F | **/** |

# Base64 (3)

- Example
  - Hello!
    - 48, 65, 6C, 6C, 6F, 41
    - 01001000, 01100101, 01101100 01101100, 01101111, 01000001
    - 010010, 000110, 010101, 101100 011011, 000110, 111101, 000001
    - 12, 06, 15, 2C, 1B, 06, 3D, 01
  - SGVsbG9B