JÖNKÖPING UNIVERSITY

*School of Engineering*

# CHARACTER ENCODINGS

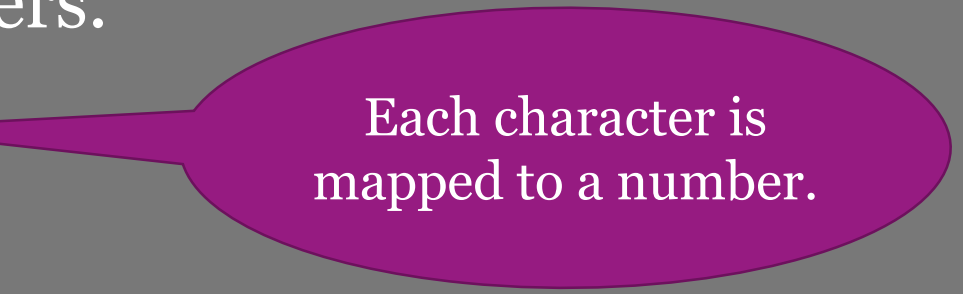
**Peter Larsson-Green**
Jönköping University
Spring 2018

# CHARACTER ENCODINGS

"Computers only understand numbers."

```
char string[] = "Hello";
```

A string in C.




Which character is mapped to which number?

- Is described by the used character encoding.

There exists many character encodings! 😢

De facto standard encoding: ASCII

# ASCII

*American Standard Code for Information Interchange*

- Each character represented by 7 bits.
  - $2^7 = 128$.
- https://www.ascii-code.com
- Does not contain å, ä, ö, Å, Ä or Ö ☹

| Number | Character |
|---|---|
| 0 | NUL |
| … | … |
| 65 | A |
| 66 | B |
| … | … |
| 97 | a |
| 98 | b |
| … | … |
| 127 | DEL |

Computers usually work with 8 bits.

- Encodings extending ASCII has been created.
- 128 additional characters!

# EXTENDED ASCII

There exists many of them!

ISO 8859-1 / ISO Latin-1

| Number | Character |
|---|---|
| 0 | NUL |
| … | … |
| 127 | DEL |
| 128 | PAD |
| … | … |
| 196 | Ä |
| 197 | Å |
| … | … |
| 214 | Ö |
| … | … |
| 255 | ÿ |

Windows-1252 / ANSI

| Number | Character |
|---|---|
| 0 | NUL |
| … | … |
| 127 | DEL |
| 128 | € |
| … | … |
| 196 | Ä |
| 197 | Å |
| … | … |
| 214 | Ö |
| … | … |
| 255 | ÿ |

# UNICODE

The solution to all encoding problems!

- Can store ~1.000.000 characters.
- Some of the encodings:
  - UTF-32: each character represented by 32 bits.
  - UTF-8: each character represented by 8, 16, 24 or 32 bits.