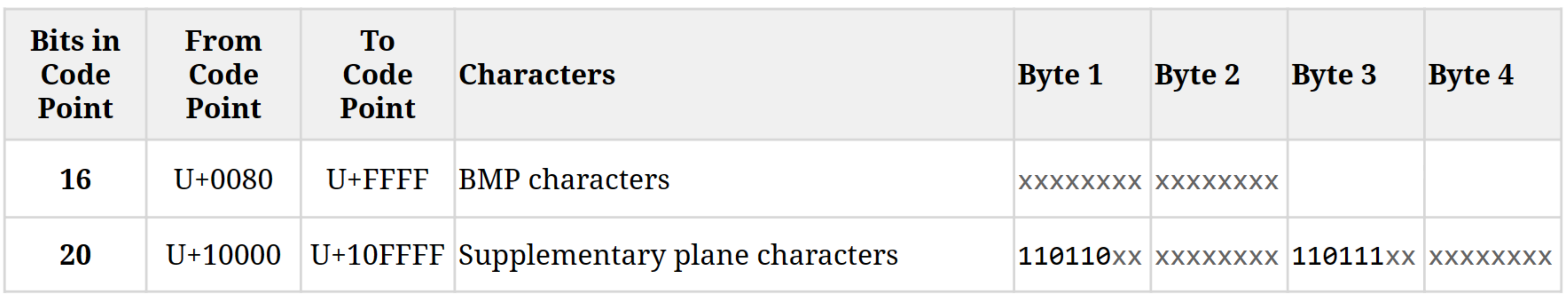
What a Character

Encoding - Details

Sualeh Fatehi

UTF-16 Encoding



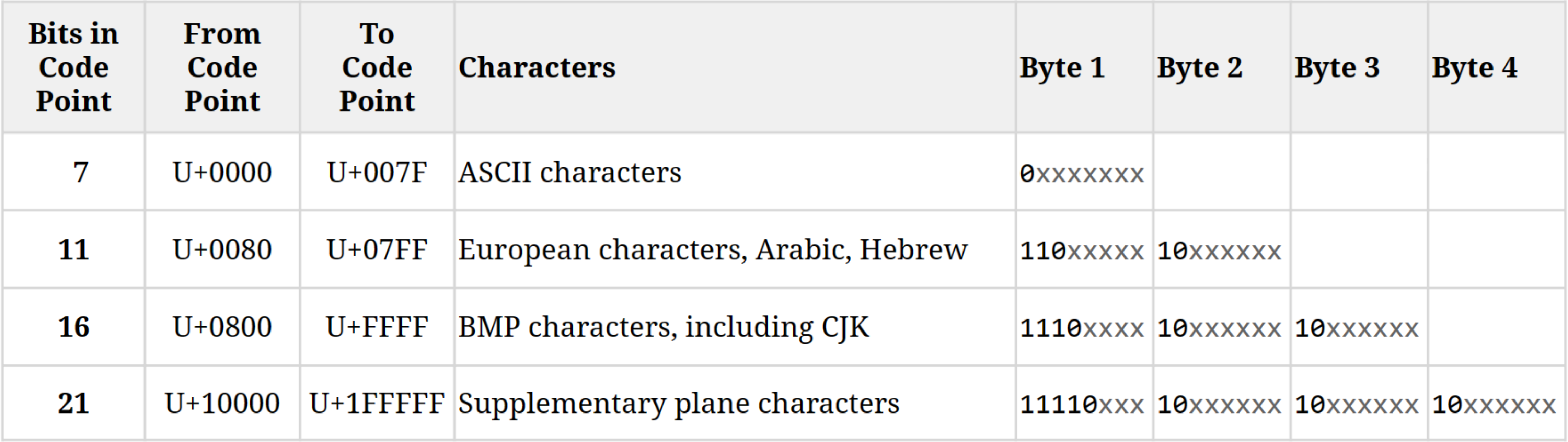
Bits in Code Point	From Code Point	To Code Point	Characters	Byte 1	Byte 2	Byte 3	Byte 4
16	U+0080	U+FFFF	BMP characters	xxxxxxxx	xxxxxxxx		
20	U+10000	U+10FFFF	Supplementary plane characters	110110xx	xxxxxxxx	110111xx	xxxxxxxx

UTF-16 Encoding

- **High surrogates** (two bytes) start with bits **110110** (0xD800)
- **Low surrogates** (two bytes) start with bits **110111** (0xDC00)
- Other bits encode the the supplementary plane and code point

(See previous slide)

UTF-8 Encoding



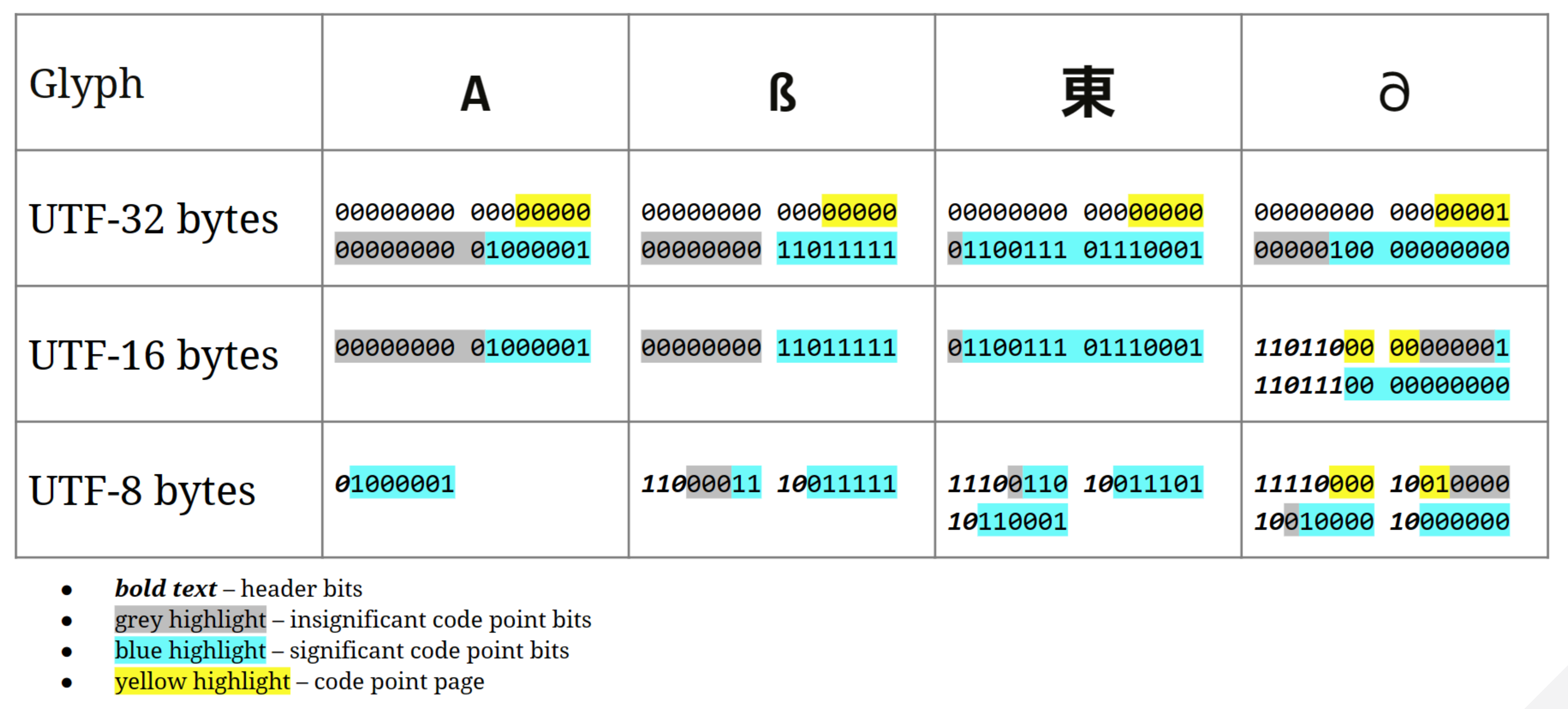
Bits in Code Point	From Code Point	To Code Point	Characters	Byte 1	Byte 2	Byte 3	Byte 4
7	U+0000	U+007F	ASCII characters	0xxxxxxx			
11	U+0080	U+07FF	European characters, Arabic, Hebrew	110xxxxx	10xxxxxx		
16	U+0800	U+FFFF	BMP characters, including CJK	1110xxxx	10xxxxxx	10xxxxxx	
21	U+10000	U+1FFFFF	Supplementary plane characters	11110xxx	10xxxxxx	10xxxxxx	10xxxxxx

UTF-8 Encoding

- **0** first bit signifies 7-bit ASCII character
- **110** leading bits signify 1 continuation byte
- **1110** leading bits signify 2 continuation bytes
- **11110** leading bits signify 3 continuation bytes
- **10** leading bits signify the continuation byte

(See previous slide)

Encoding Details

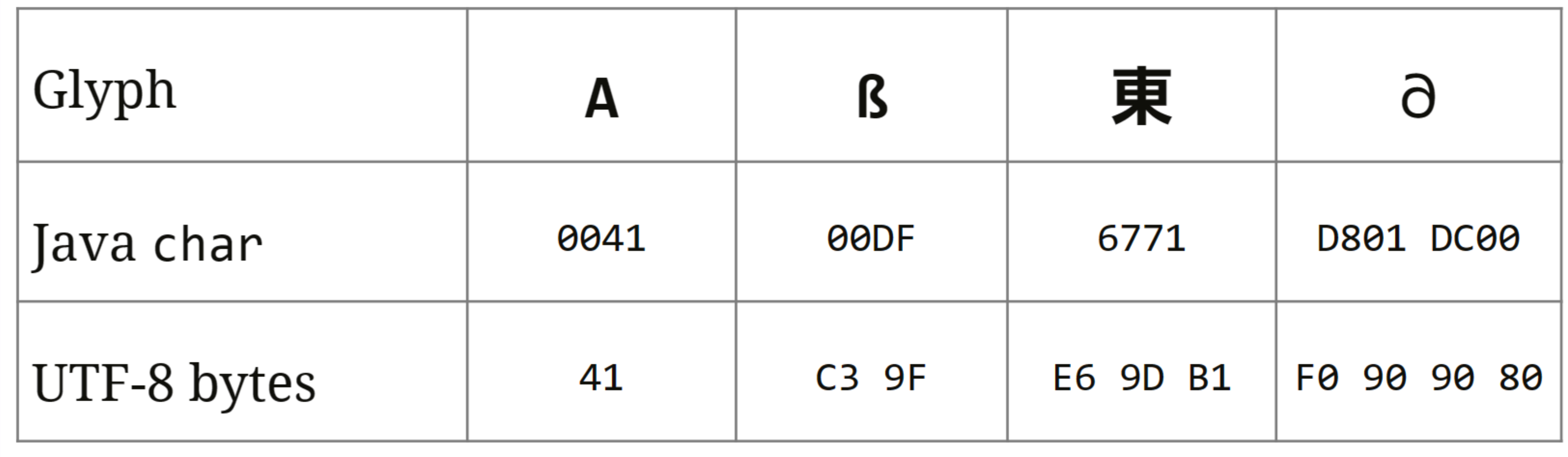


Glyph	A	ß	東	∂
UTF-32 bytes	00000000 00000000 00000000 01000001	00000000 00000000 00000000 11011111	00000000 00000000 01100111 01110001	00000000 00000001 00000100 00000000
UTF-16 bytes	00000000 01000001	00000000 11011111	01100111 01110001	***110110***00 00000001 ***110111***00 00000000
UTF-8 bytes	01000001	***110***00011 ***10***011111	***1110***0110 ***10***011101 ***10***110001	***11110***000 ***10***010000 ***10***010000 ***10***000000

- ***bold text*** – header bits
- grey highlight – insignificant code point bits
- blue highlight – significant code point bits
- yellow highlight – code point page

Where Do You Truncate?

How and where do you truncate string “Aß東𐐀”?



Glyph	A	ß	東	𐐀
Java char	0041	00DF	6771	D801 DC00
UTF-8 bytes	41	C3 9F	E6 9D B1	F0 90 90 80

TIP: There is no easy answer. Use a library to truncate strings.

Code Examples

Slides and all code examples are on GitHub

https://github.com/**sualeh/What-a-Character**