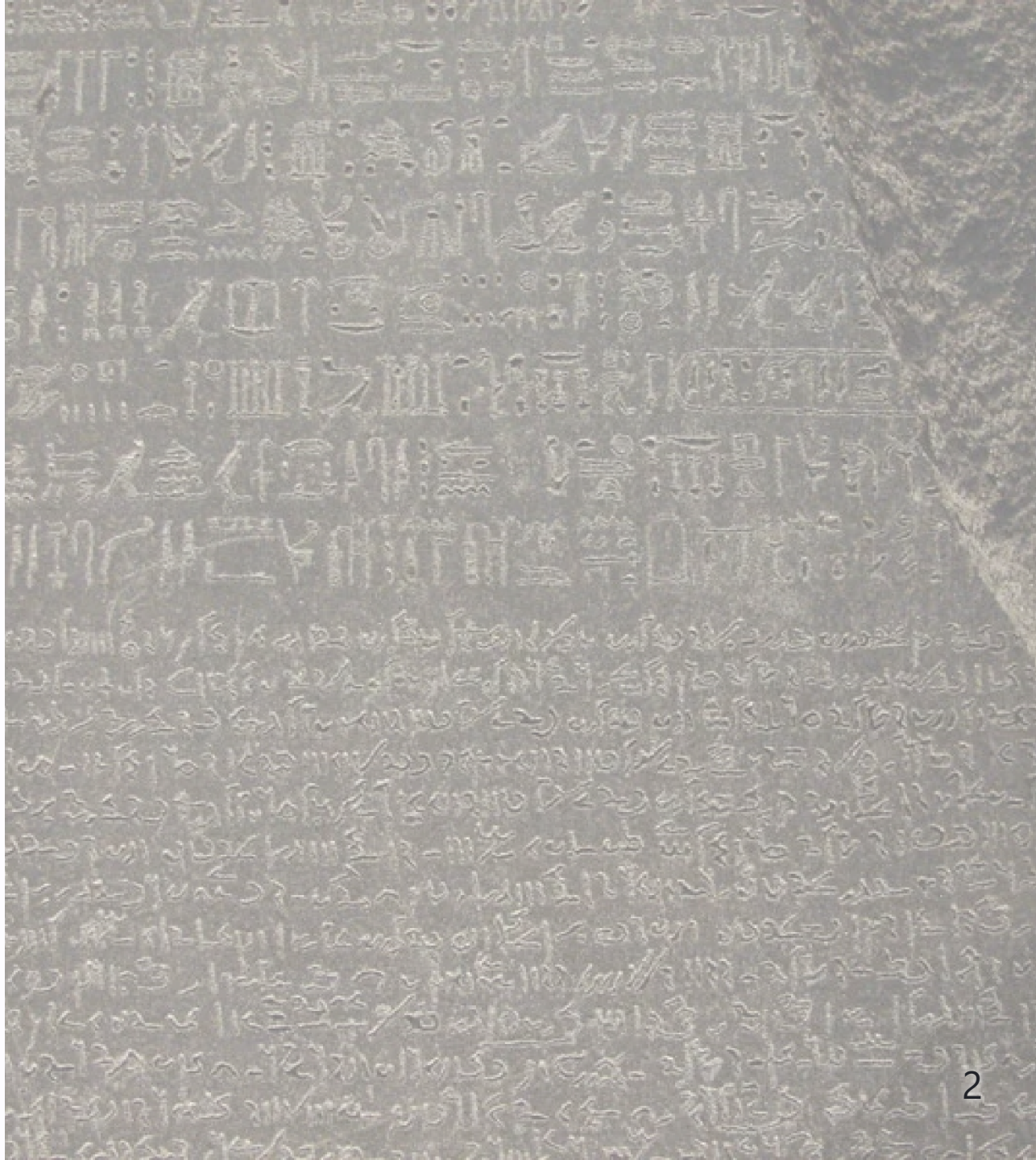
What a Character

Encoding - Concepts

Sualeh Fatehi

Encoding

Encoding specifies conversion of characters to bytes



Encoding

- **Encoding** is the process of converting code points to a byte representation
- **Decoding** is the process of converting a stream of bytes to code points
- Encoding or decoding may not always be successful

Data in Bytes

Careful encoding and decoding is required when data is serialized to bytes.

- File I/O operations
- Network communication
- Database persistence
- RAM

Common Encodings

Unicode Code Point	U+0041	U+00DF	U+6771	U+010400
Glyph	A	ß	東	𐐀
UTF-32 bytes	00 00 00 41	00 00 00 DF	00 00 67 71	00 01 04 00
UTF-16 bytes	00 41	00 DF	67 71	D8 01 DC 00
UTF-8 bytes	41	C3 9F	E6 9D B1	F0 90 90 80

UTF-32

- Four bytes for all characters
- Does not need surrogate pairs

UTF-16

- Two bytes for BMP characters
- Four bytes for supplementary plane characters
- Uses surrogate pairs

UTF-8

- Most common encoding today
- One byte for ASCII characters
- Two or three bytes for BMP characters
- Four bytes for supplementary plane characters
- Does not need surrogate pairs

Base64 Encoding

- Base64 is a binary-to-text encoding scheme
- Base64 does not encode characters or code points
- Encoded text is in "printable" ASCII characters

On the other hand,

- Unicode encoding schemes encode text-to-binary

Explicitly Specify Encoding

Always explicitly specify encoding to avoid cross-platform surprises.

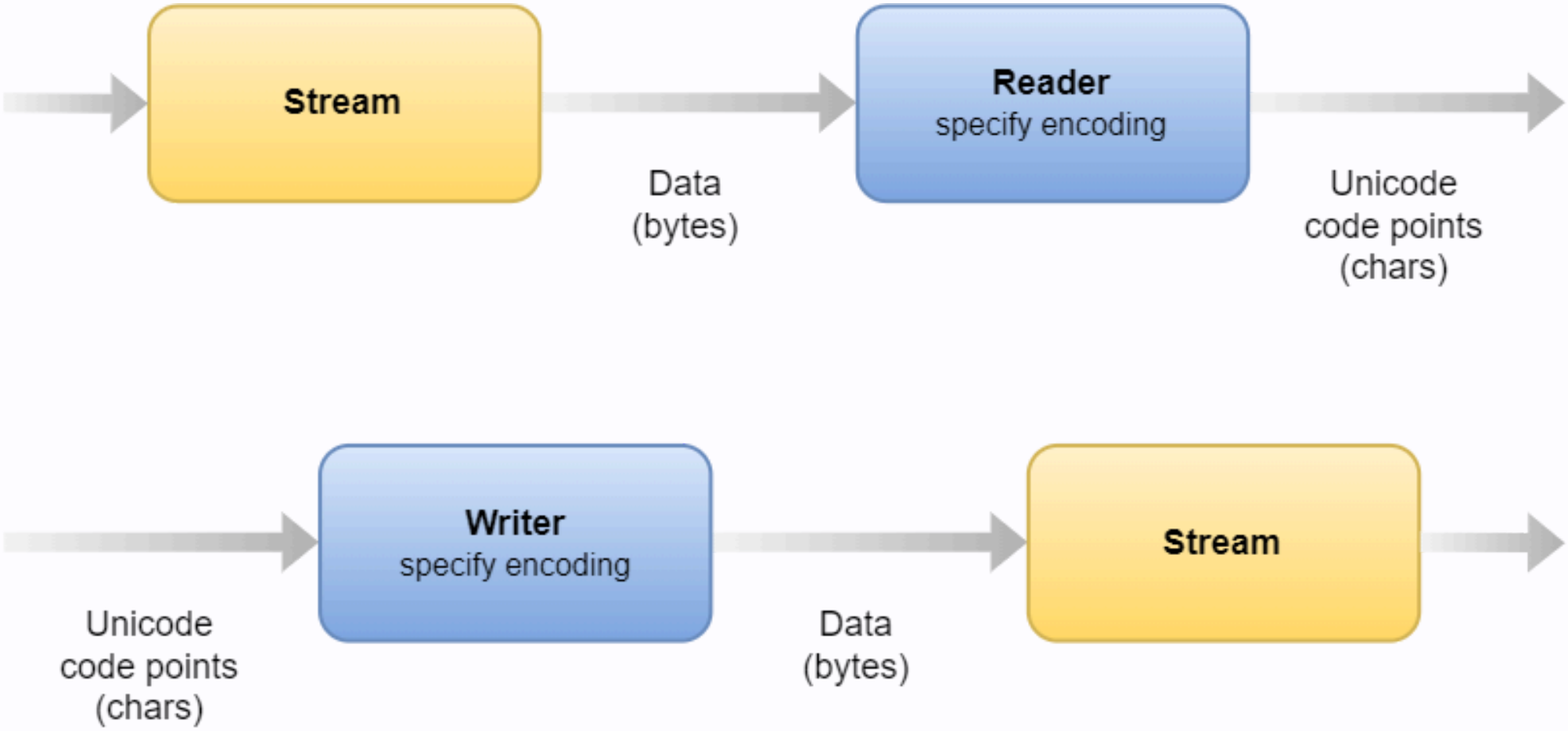
You may not always get errors - just garbled data.

Streams and Readers

In Java and C#,

- **Streams** read and write bytes
- **Readers** read characters from a byte stream
- **Writers** write characters to a byte stream
- Always specify encoding to avoid cross-platform surprises

Readers and Writers



Text Data in Databases

- `VARCHAR` and `CHAR` specify lengths in bytes, by default
- `NVARCHAR` and `NCHAR` specify lengths in characters, but use a certain multiplier for bytes

You may run out of space if you do not calculate right.

Code Examples

Slides and all code examples are on GitHub

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