JavaScript and the farmer emoji 

it's me, evan hahn

evanhahn.com/chicagojs2023

```
"hi".length
// => ???
```

```
"hi".length
// => 2
```

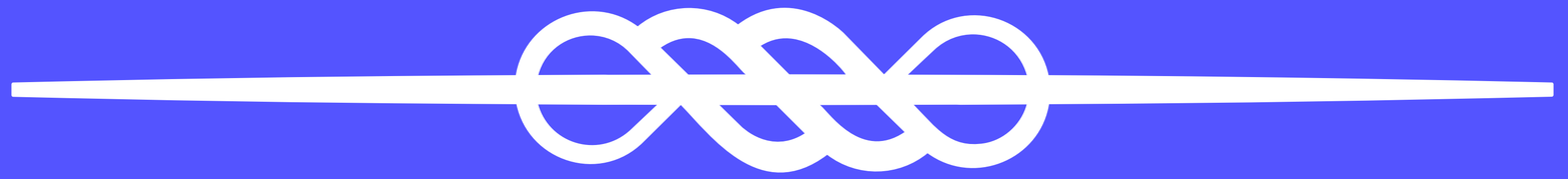
```
"👩‍🌾".length
// => ???
```

```
"👩‍🌾".length
// => 7
```

what???

what does `length` do?

Glossary of Unicode Terms



extended grapheme cluster

Extended Grapheme Cluster.

The text between extended grapheme cluster boundaries as specified by Unicode Standard Annex #29, "Unicode Text Segmentation." Abbreviated as EGC.

Extended Grapheme Cluster.

The text between extended grapheme cluster boundaries as specified by Unicode Standard Annex #29, "Unicode Text Segmentation." Abbreviated as EGC.

what???

%

س

⚖

◍

ყ

$

€

ç

♬

廣

ಐ

Д

y

E

ñ

ß

日

한

香

8

Φ

what most people would call a "character"

what does `length` count?

what does `length` count?

grapheme clusters

what does `length` count?

~~grapheme clusters~~

questions?

U+0069

U+042

U+1F341

U+D55C

scalar values

U+6068

U+00F1

U+6669

U+6068

Unicode Scalar Value.

Any Unicode code point except high-surrogate and low-surrogate code points. In other words, the ranges of integers 0 to D7FF16 and E00016 to 10FFFF16 inclusive.

Unicode Scalar Value.

Any Unicode code point except high-surrogate and low-surrogate code points. In other words, the ranges of integers 0 to D7FF16 and E00016 to 10FFFF16 inclusive.

what???

integer between
0 and ~1.1 million
(many unused)

most extended grapheme clusters contain one scalar

J

74

9836

127800

scalars are usually written with U+

9836

AKA

U+266C

some extended grapheme clusters contain many scalars

128105
127998
8205
127806

what does `length` count?

~~grapheme clusters~~

what does `length` count?

~~grapheme clusters~~

scalars

what does `length` count?

~~grapheme clusters~~

~~scalars~~

questions?

UTF-16

how do you store these scalars?

javascript stores scalars with UTF-16

UTF-16 Encoding Form.

The Unicode encoding form that assigns each Unicode scalar value in the ranges U+0000..U+D7FF and U+E000..U+FFFF to a single unsigned 16-bit code unit with the same numeric value as the Unicode scalar value, and that assigns each Unicode scalar value in the range U+10000..U+10FFFF to a surrogate pair, according to Table 3-5, “UTF-16 Bit Distribution.”

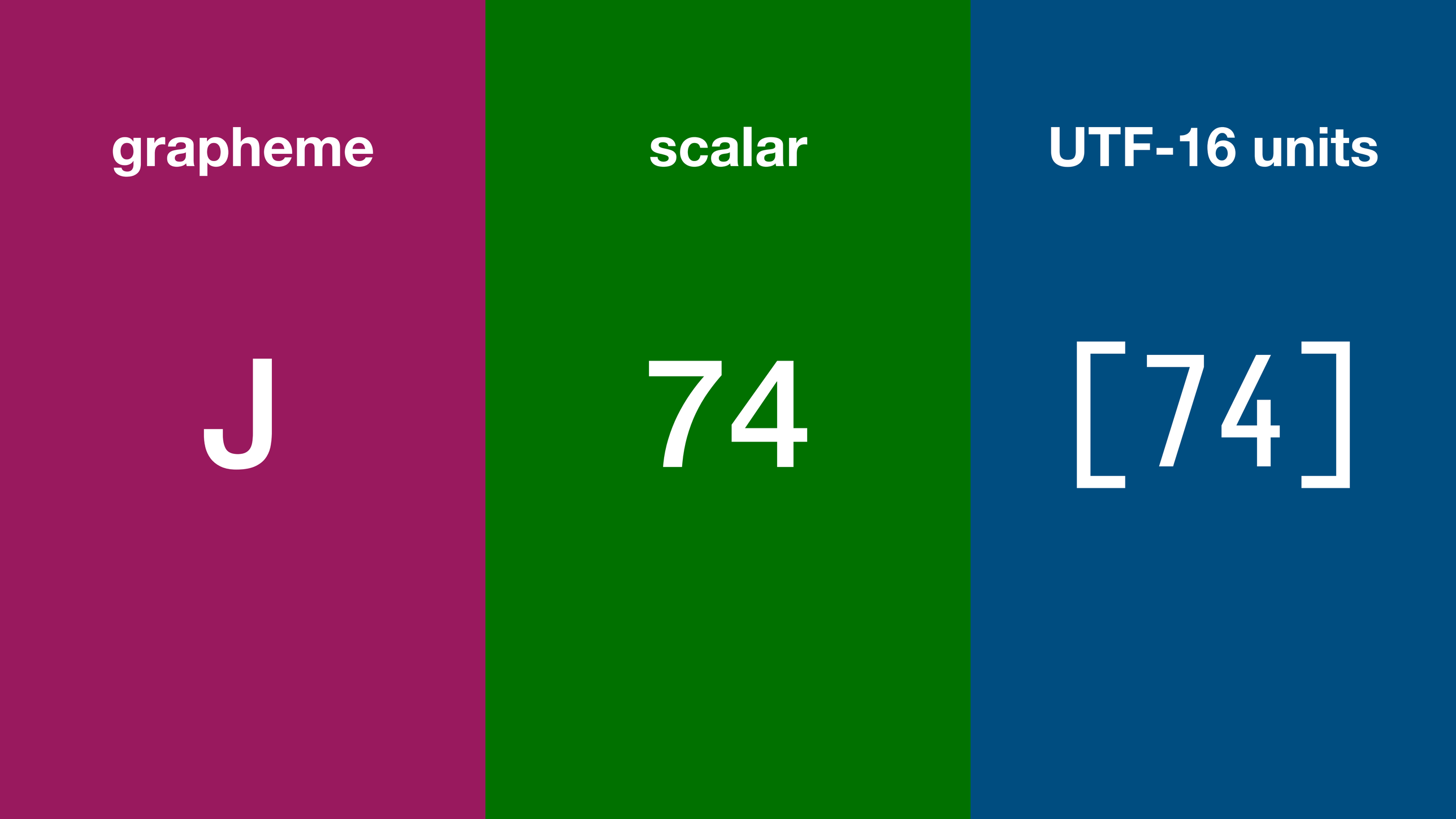
UTF-16 Encoding Form.

The Unicode encoding form that assigns each Unicode scalar value in the ranges U+0000..U+D7FF and U+E000..U+FFFF to a single unsigned 16-bit code unit with the same numeric value as the Unicode scalar value, and that assigns each Unicode scalar value in the range U+10000..U+10FFFF to a surrogate pair, according to Table 3-5, “UTF-16 Bit Distribution.”

what???

units of
16-bit integers
(0 – 65,535)

many scalars fit in a 16-bit unit


grapheme
scalar
UTF-16 units
J
74
[74]

grapheme	scalar	UTF-16 units
	9836	[9836]

some scalars are too big and get split in two

"surrogate pair"

grapheme	scalar	UTF-16 units
	127800	[55356, 57144]

grapheme	scalars	UTF-16 units
	128105 127998 8205 127806	[55357, 56425, 55356, 57342, 8205, 55356, 57150]

what does `length` count?

what does `length` count?

~~grapheme clusters~~

what does `length` count?

~~grapheme clusters~~

~~scalars~~

what does `length` count?

~~grapheme clusters~~

~~scalars~~

UTF-16 units

```
"👩‍🌾".length
// => 7
```

✨ *advice time* ✨

here be bugs

character limits

0/100

1/100

He

2/100

Hel

3/100

Hell

4/100

Hello

5/100

Hello

6/100

Hello 

13/100

syncing with a backend

```
let isValid = (s) => (
  (s.length > 0) &&
  (s.length < 20)
);
```

frontend (JS)

```
def is_valid(s):
    return 0 < len(s) < 20
```

backend (Python)

```
"♬"[0]
// ⇒ "♬"
```

```
"🌸"[0]
// ⇒ "□"
```

most of javascript uses UTF-16 code units

.length

.at()

.charAt()

.charCodeAt()

.codePointAt() *(kinda)*

.includes()

.indexOf()

.lastIndexOf()

.match()

.matchAll()

.padEnd()

.padStart()

.replace()

.replaceAll()

.search()

.slice()

uses UTF-16

.split()

.endsWith()

.startsWith()

.substring()

String.
fromCharCode()

iterating uses scalars

```
[..."👩‍🌾"]
// ⇒ ["👩", "🟫", "◻", "🌾"]
```

`Intl.Segmenter()` *(some browsers) & graphemer (library) use extended grapheme clusters*

not just a javascript problem!

summary

extended grapheme clusters

U+0069

U+042

U+1F341

U+D55C

scalars

U+6068

U+00F1

U+6669

U+6068

UTF-16

`length` **counts**

UTF-16 units

watch out for bugs

thank you!!

evanhahn.com/chicagojs2023