String Algorithms and Data Structures

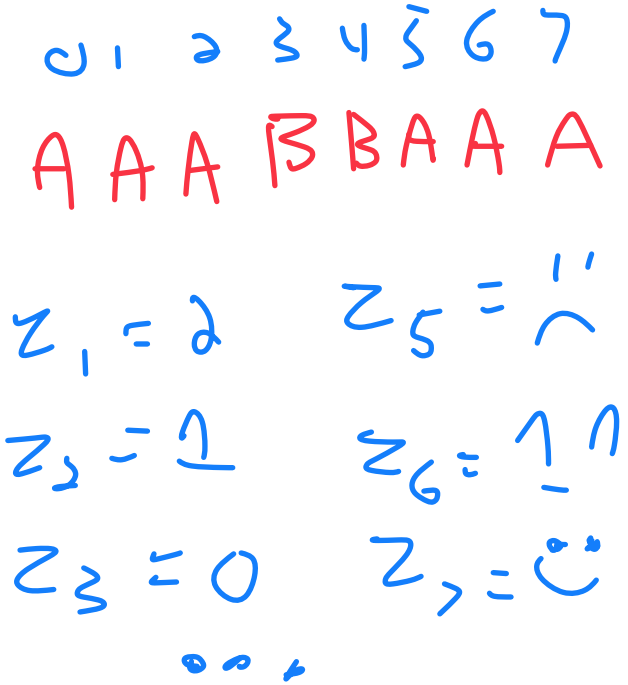
The Z-algorithm

CS 199-225
Brad Solomon

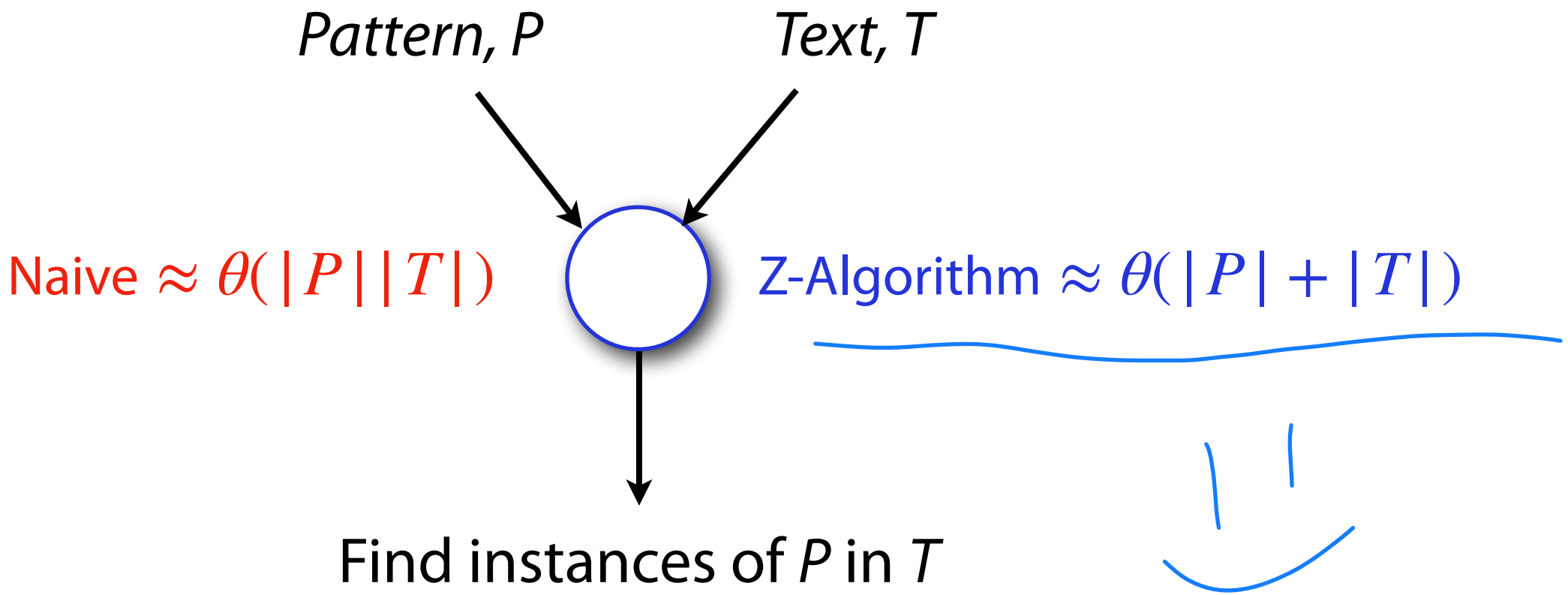
February 13, 2023

UNIVERSITY OF ILLINOIS URBANA-CHAMPAIGN

Department of Computer Science



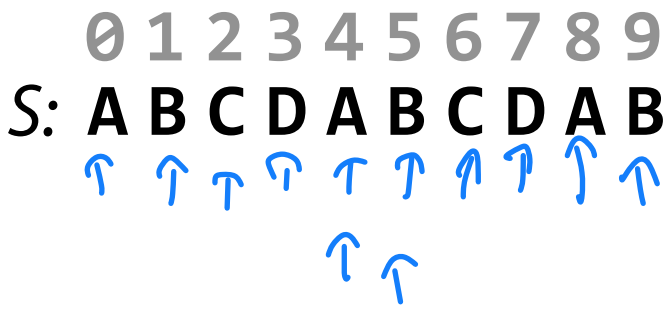
Exact Pattern Matching w/ Z-algorithm



'instances': An exact, full length copy

The Z-value [$Z_i(S)$]

Given a string S, $Z_i(S)$ is the length of the longest substring in S, starting at position $i > 0$, that matches a prefix of S.



$Z_4(S) = 6$

S: C G C G A ? ? ? ? ?

C G C X

Not G

$Z_5(S) = 3$

S: A ? ? ? ? ? ? ? ? ?

A A A A A A A X

. . . . .

Not A

$Z_1(S) = 7$

The Z-Algorithm

P $ T

```
S: 1 0 1 $ 1 0 1 0 1 1
   0 1 $ 1 0 1 0 1 1
   1 $ 1 0 1 0 1 1
   $ 1 0 1 0 1 1
   1 0 1 0 1 1
   0 1 0 1 1
   1 0 1 1
   0 1 1
   1 1
   1
```

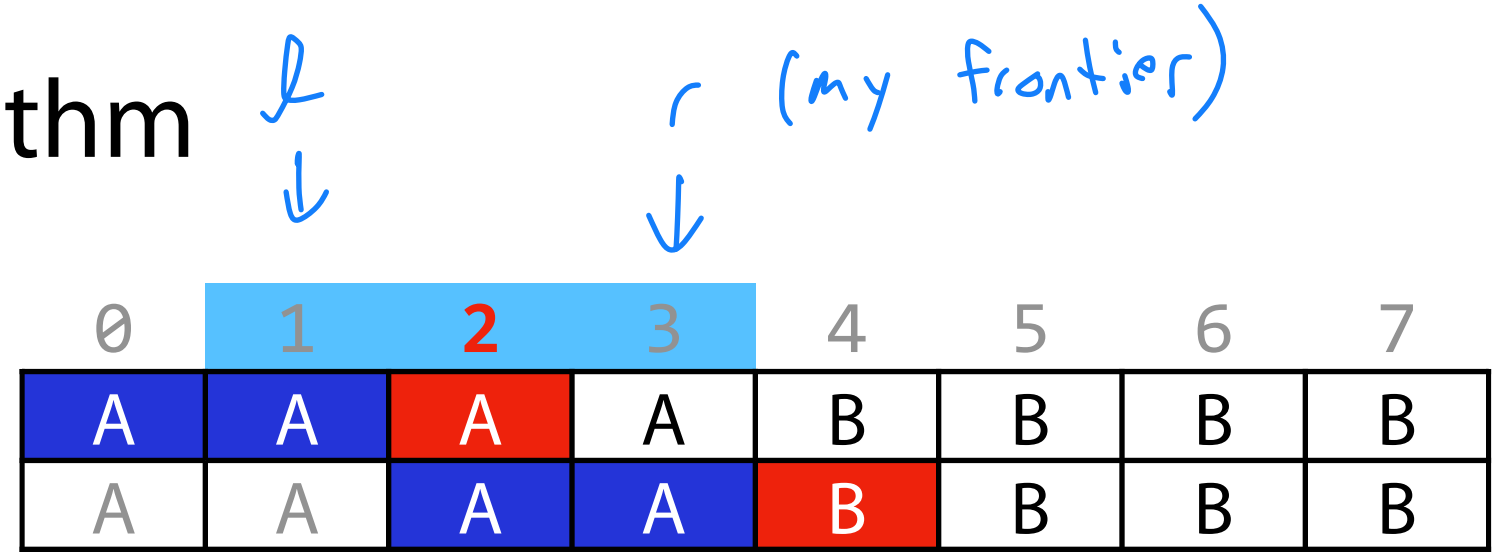
16 total char comps

↳ More work than naive!

The Z-Algorithm

$Z_1 = 3$

$Z_2 =$



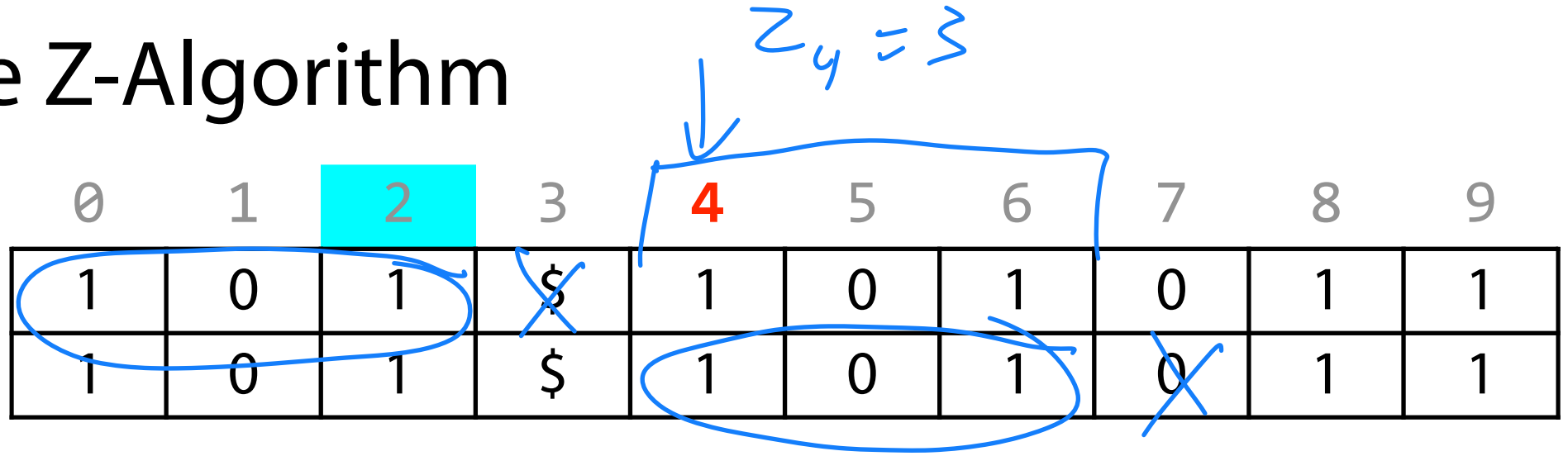
We track our current knowledge of *S* using three values: i, r, l

i gets updated every iteration (as we compute Z_i)

r gets updated when $Z_i > 0$ *AND* $r_{new} > r_{old}$

l gets updated whenever r is updated (it stores the index of r's Z-value)

The Z-Algorithm



	Start	End
i, the current index =	4	5
r, the furthest match char =	2	6
l, the furthest reaching Z-value =	2	4

The Z-Algorithm

0	1	2	3	4	5	6	7	8	9
1	0	1	$	1	0	1	0	1	1
1	0	1	$	1	0	1	0	1	1

$Z_5 = 0$, w/o doing any work

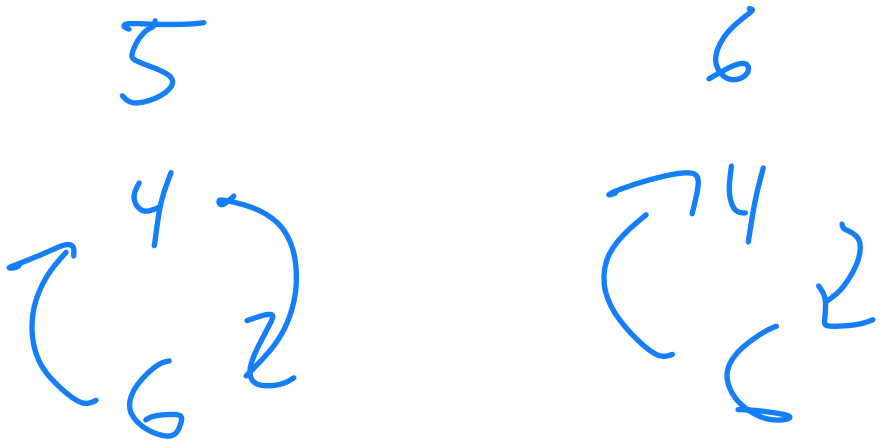
i, the current index =

r, the furthest match char =

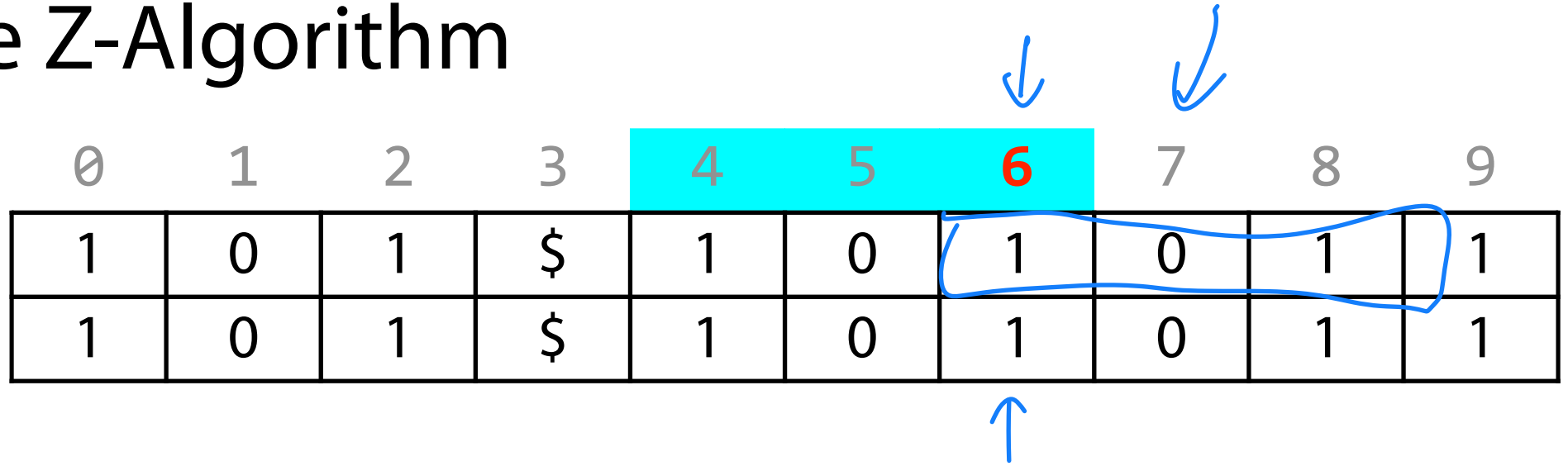
l, the furthest reaching Z-value =

Start

End

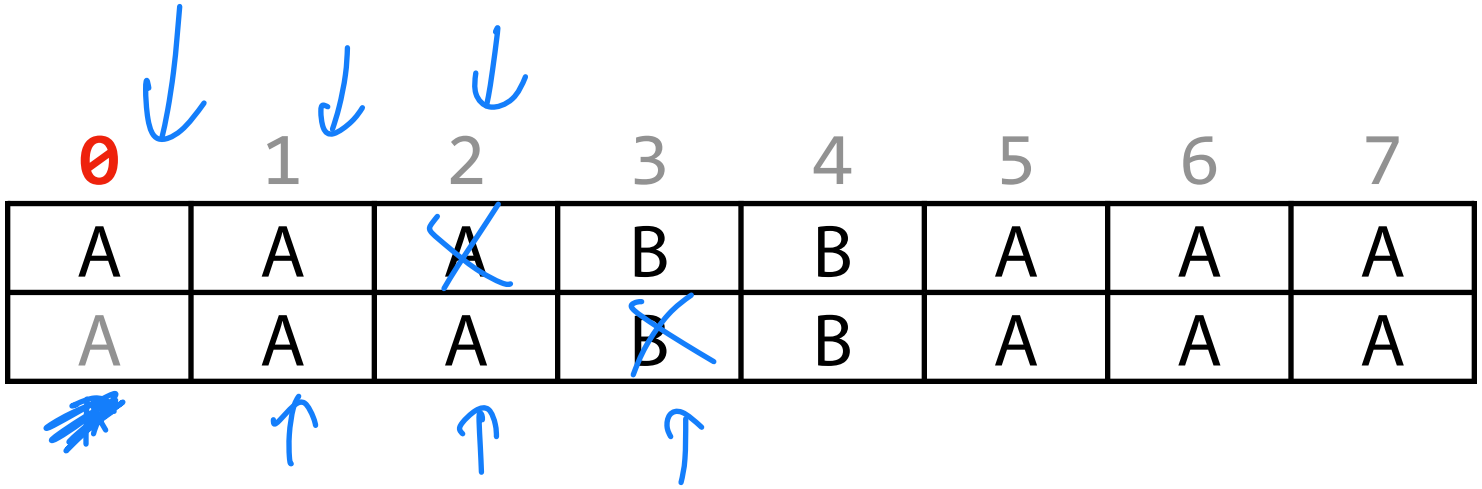


The Z-Algorithm



	Start	End
i, the current index =	6	7
r, the furthest match char =	6	8
l, the furthest reaching Z-value =	4	6

The Z-Algorithm



The values of i, r, l tell us how much work we need to do to compute Z_i

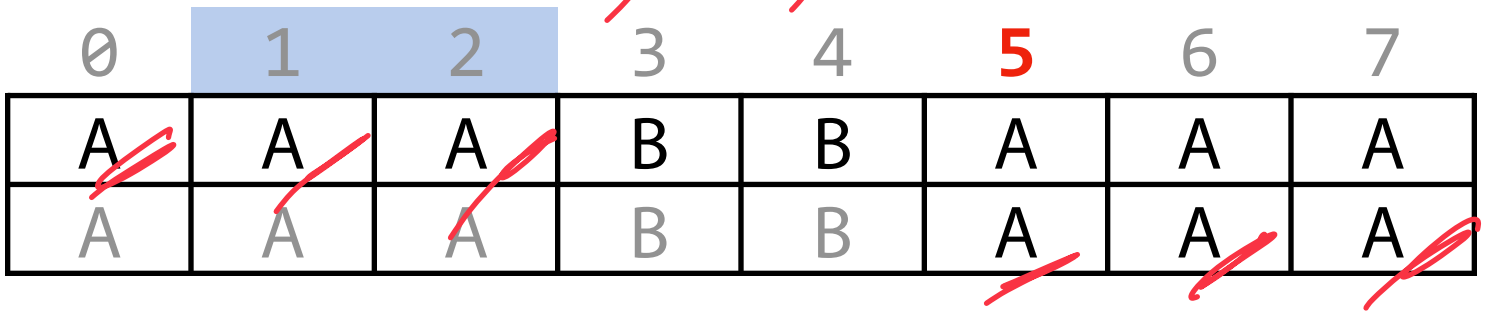
Case 1: $i > r$ default to 0

Ex: $i = 1, r = 0, l = 0$

$Z_1 = 2$

We must compute Z_i explicitly!

The Z-Algorithm



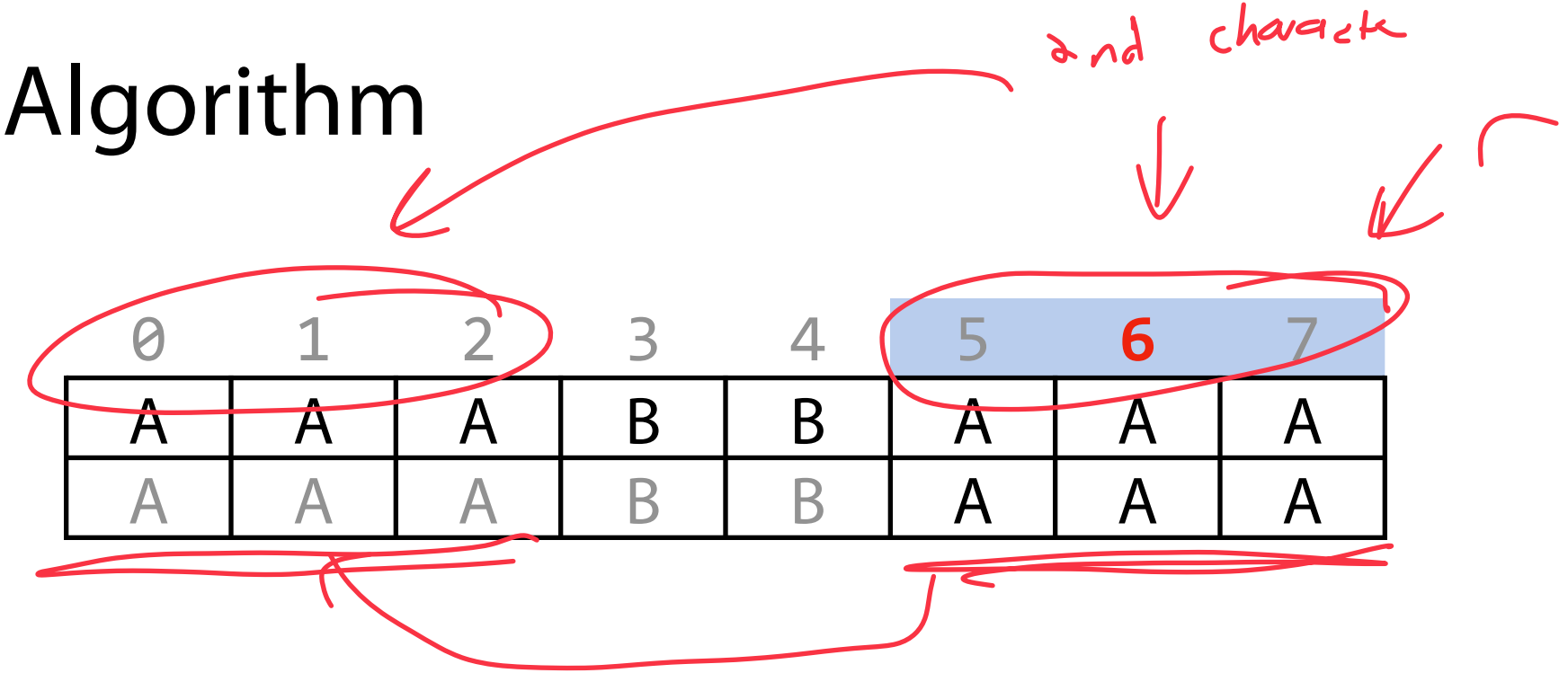
The values of i, r, l tell us how much work we need to do to compute Z_i

Case 1: $i > r$

Ex: $i = 5, r = 2, l = 1$

We must compute Z_i explicitly!

The Z-Algorithm



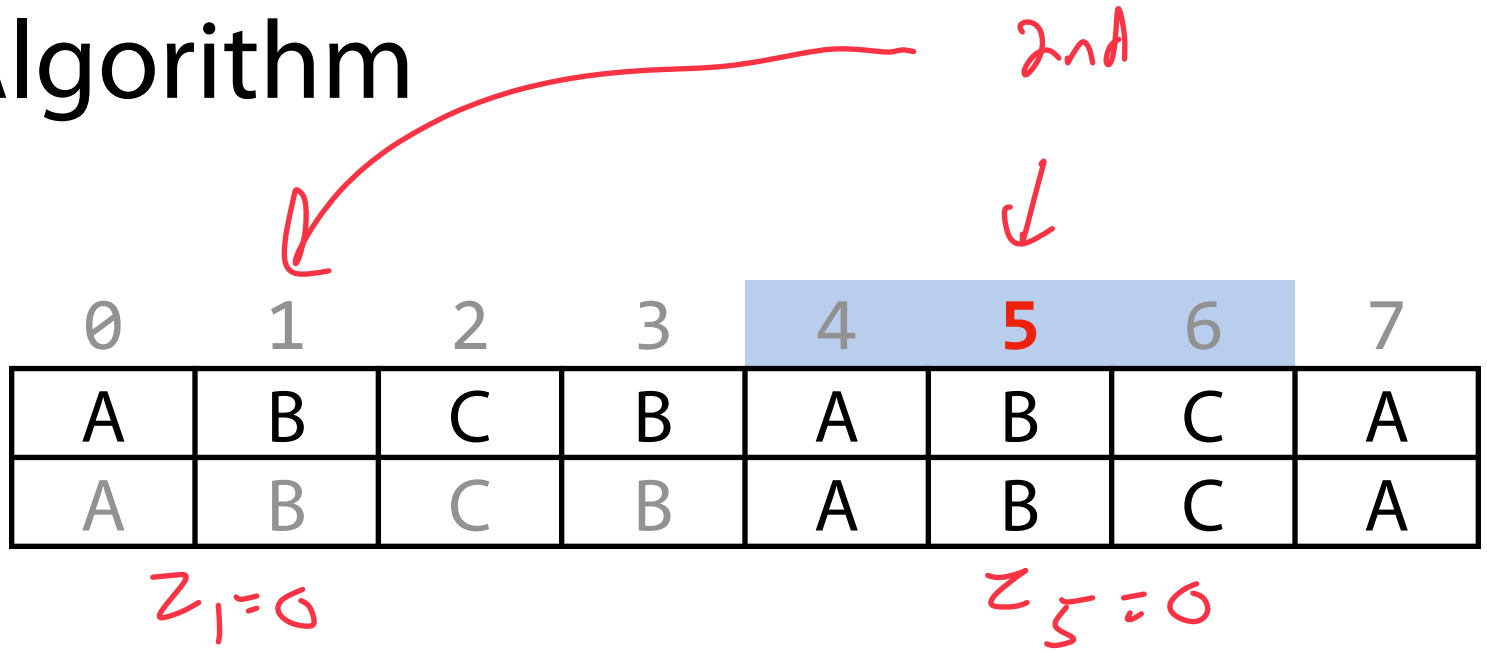
The values of i, r, l tell us how much work we need to do to compute Z_i

Case 2: $i \leq r$

Ex: $i = 6, r = 7, l = 5$

To find Z_6, we can save time by looking up the value _____

The Z-Algorithm



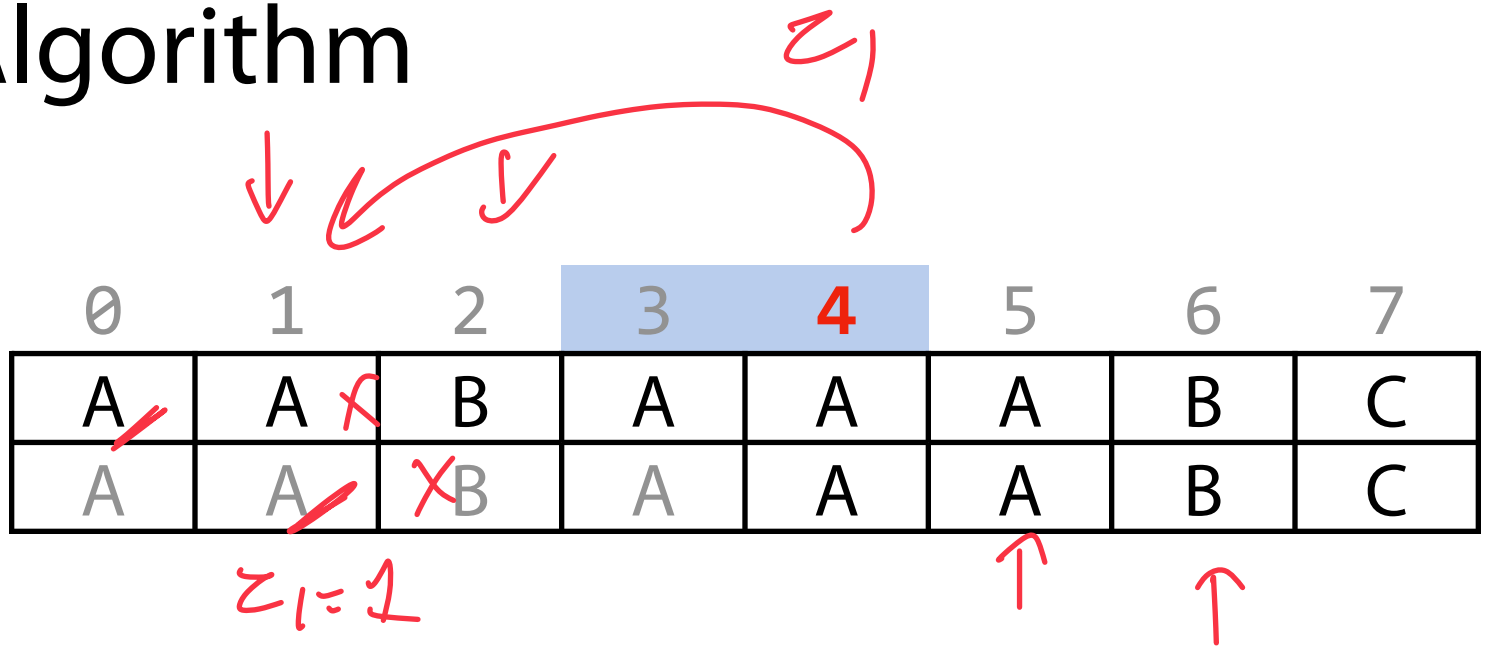
The values of i, r, l tell us how much work we need to do to compute Z_i

Case 2: $i \leq r$

Ex: $i = 5, r = 6, l = 4$

To find Z_5, we can save time by looking up the value Z_1

The Z-Algorithm



The values of i, r, l tell us how much work we need to do to compute Z_i

Case 2: $i \leq r$

Ex: $i = 4, r = 4, l = 3$

To find Z_4, we can save time by looking up the value _____

The Z-Algorithm

Let $l = 0, r = 0$, for $i = [1, \ldots, |S| - 1]$:

Compute Z_i using irl:

Case 1 ($i > r$): Compute explicitly; update irl (only if match)

Case 2 ($i \leq r$):

Use previous Z-values to avoid work

Explicitly compute only *'new'* characters

How can we tell the difference between cases?

The Z-Algorithm

$i = 6, r = 7, l = 5$

0	1	2	3	4	5	6	7	8
A	**A**	**A**	**A**	**C**	**A**	**A**	**A**	**B**
A	A	A	A	C	A	A	A	B
A	A	A	A	C	A	A	A	B

The amount of work required depends on two pieces of information

1. # of characters at or after i that we have seen before

↳ these are our Z-value calc characters

2. The Z-value that matches part or all of the string starting at i

↳ what Z-value do I look up?

The Z-Algorithm

$i = 6, r = 7, l = 5$

0	1	2	3	4	5	6	7	8
A	**A**	**A**	**A**	**C**	**A**	**A**	**A**	**B**
A	A	A	A	C	A	A	A	B
A	A	A	A	C	A	A	A	B

The amount of work required depends on two pieces of information

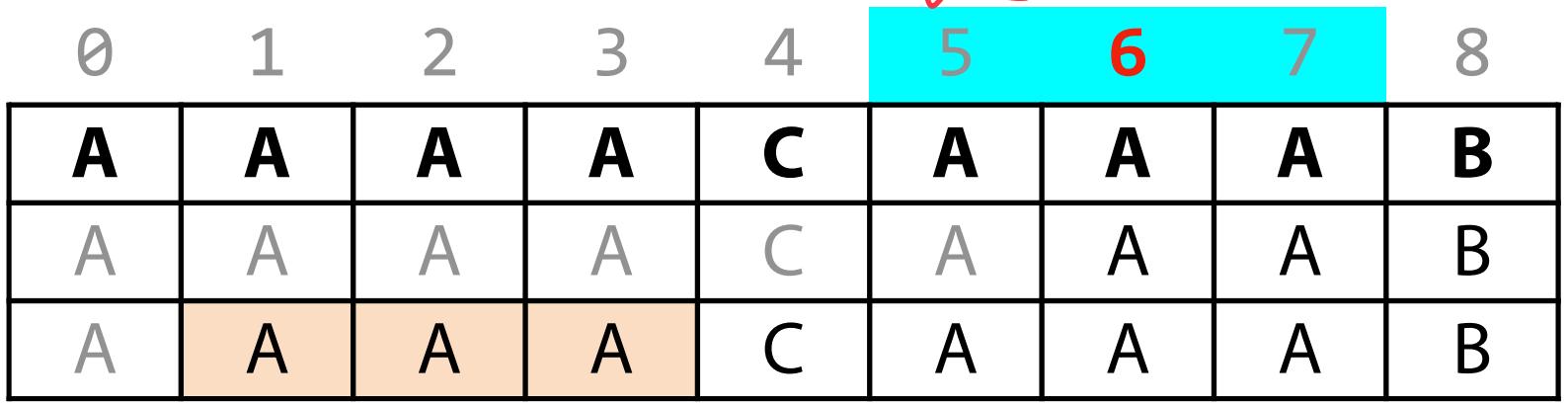
1. # of characters at or after i that we have seen before

Call this value $|\beta|$. What is $|\beta|$ in terms of i, r, l?

$\beta = r - i + 1$

The Z-Algorithm

$i = 6, r = 7, l = 5$



The amount of work required depends on two pieces of information

2. The Z-value that matches part or all of the string starting at i

Call this value Z_k. What is k in terms of i, r, l ?

$k = i - l$

The Z-Algorithm

$Z_5 = 3$

$l = 5$

$i = 6, r = 7, l = 5$

	0	1	2	3	4	5	6	7	8
	A	**A**	**A**	**A**	**C**	**A**	**A**	**A**	**B**
	A	A	A	A	C	A	A	A	B
$Z_k = Z_1 = 3$	A	A	A	A	C	A	A	A	B

The amount of work required depends on two pieces of information

1. # of characters at or after i that we have seen before

$$|\beta| = 7 - 6 + 1 = 2$$

2. The Z-value that matches part or all of the string starting at i

$$k = 6 - 5 = 1$$

The Z-Algorithm

$i = 5, r = 7, l = 4$

0	1	2	3	4	5	6	7
A	A	A	B	A	A	A	B
A	A	A	B	A	A	A	B

Case 2a: $i \leq r, Z_k < |\beta|$

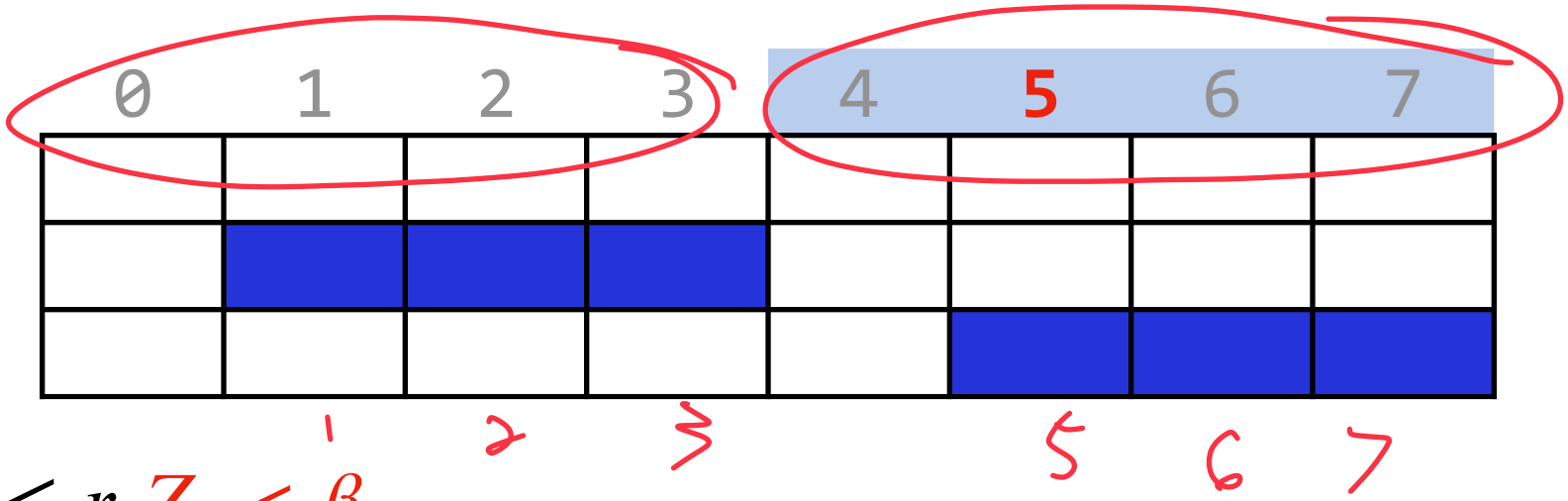
$|\beta| =$ 7 - 5 + 1 = 3 (c - i + 1), $k =$ 1, $Z_k =$ 2 (Z_1)

$Z_i =$ 2

$Z_i = Z_k$

The Z-Algorithm

$i = 5, r = 7, l = 4$

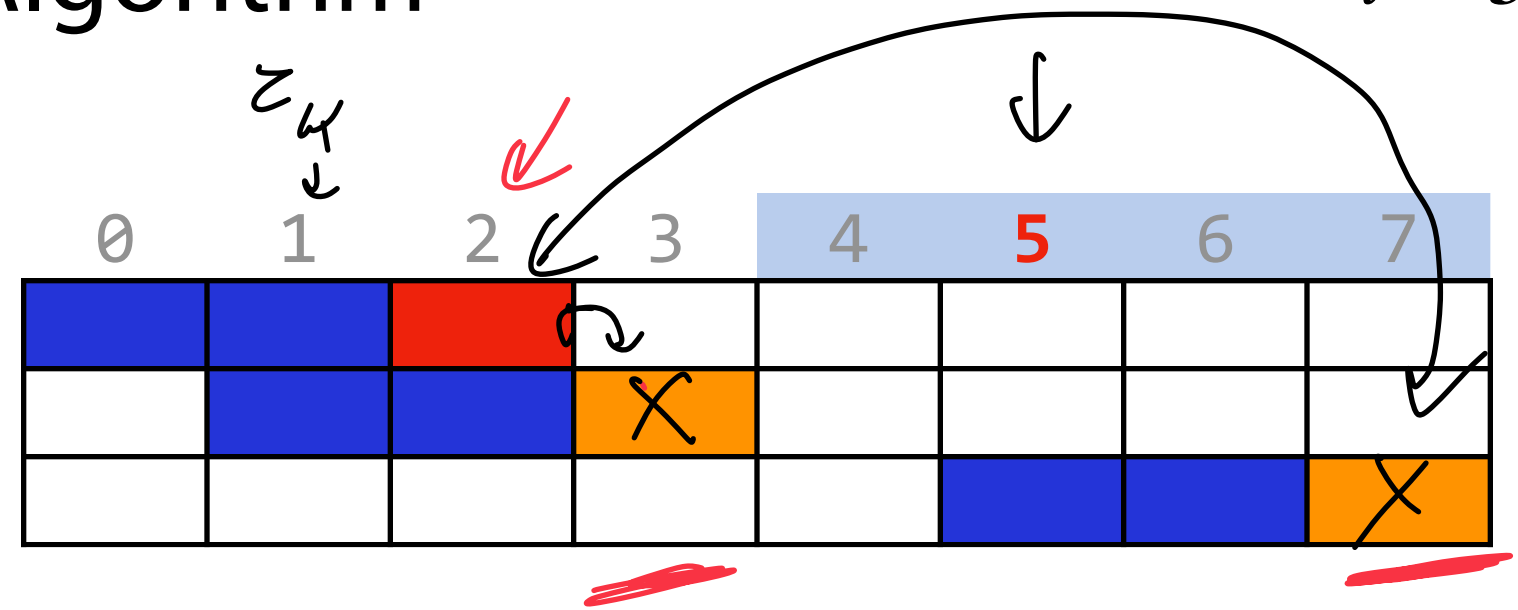


Case 2a: $i \leq r, Z_k < \beta$

Z_l (defined by r, l) tells us that β matches earlier.

The Z-Algorithm

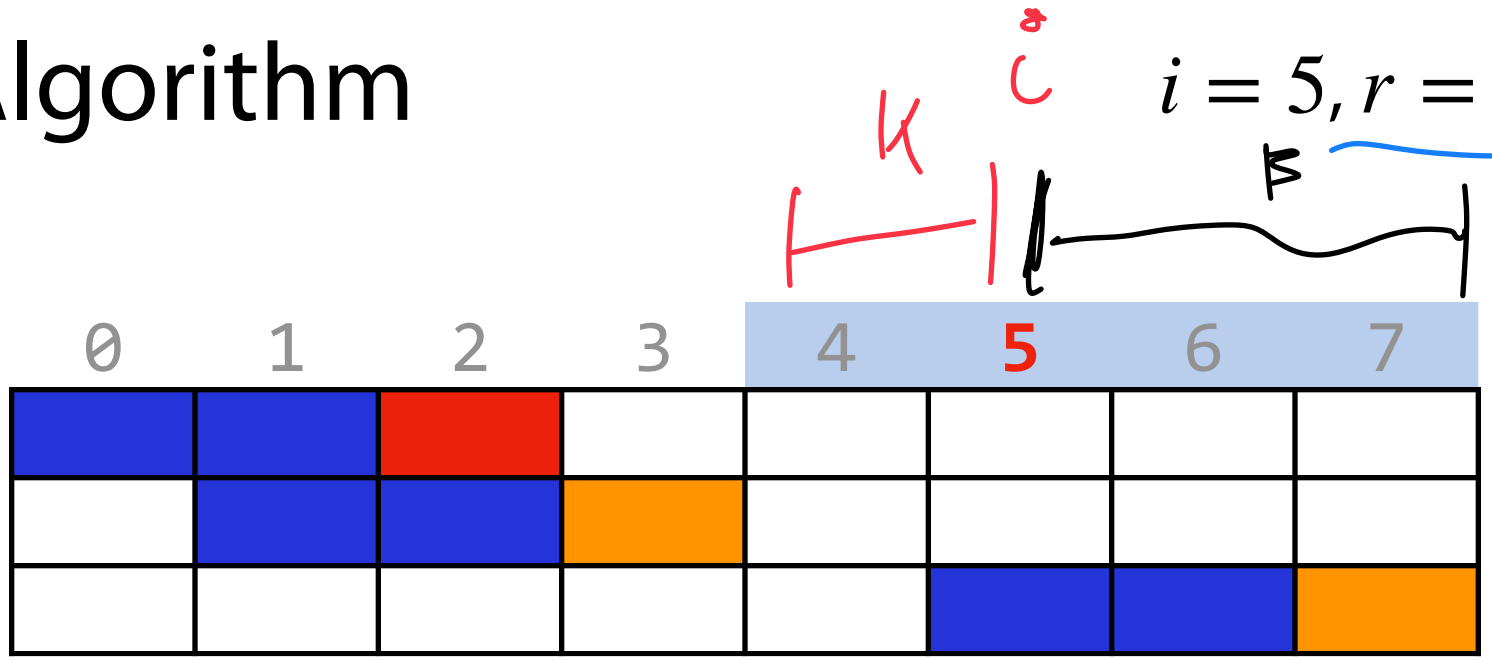
$i = 5, r = 7, l = 4$



Case 2a: $i \leq r, Z_k < |\beta|$

Z_l tells us that β matches earlier. Z_k tells us how much matches the prefix.

The Z-Algorithm



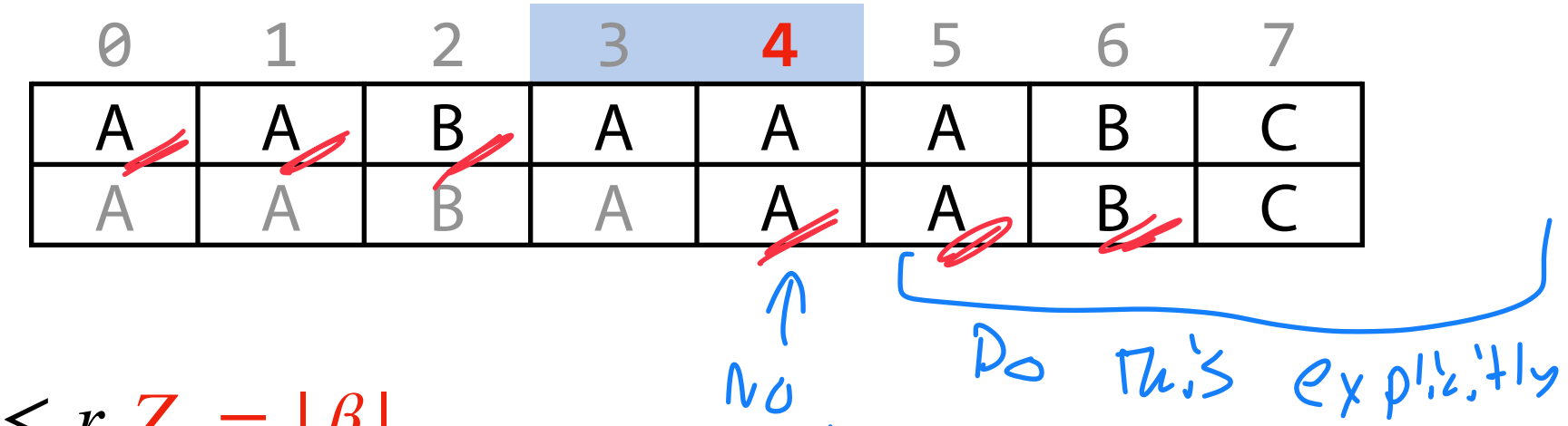
Case 2a: $i \leq r, Z_k < |\beta|$

Z_l tells us that β matches earlier. Z_k tells us how much matches the prefix.

Because $Z_k < |\beta|, Z_i =$ Z_K

The Z-Algorithm

$i = 4, r = 4, l = 3$



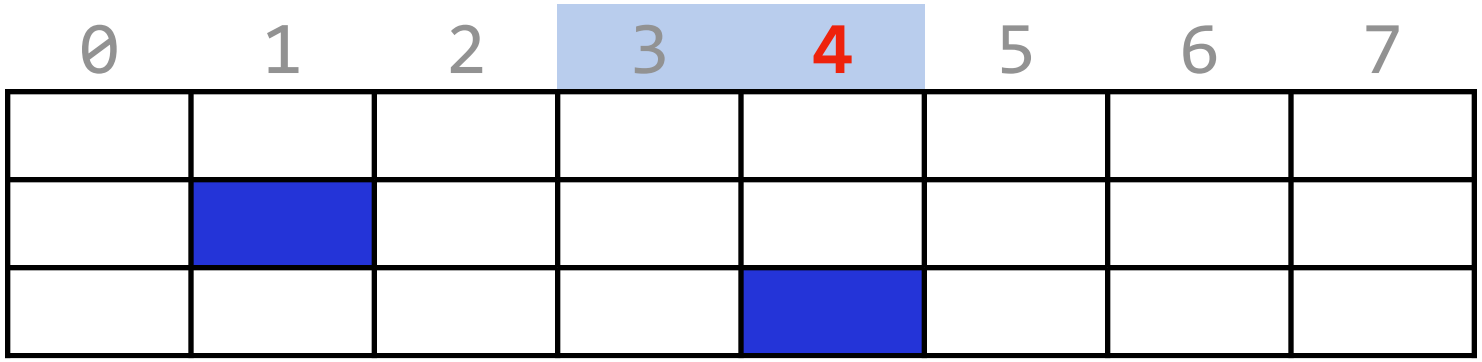
Case 2b: $i \leq r, Z_k = |\beta|$

$|\beta| =$ 4 - 4 + 1 = 1 $, k =$ 1 $, Z_k =$ 1

$Z_i =$ 3

The Z-Algorithm

$i = 4, r = 4, l = 3$

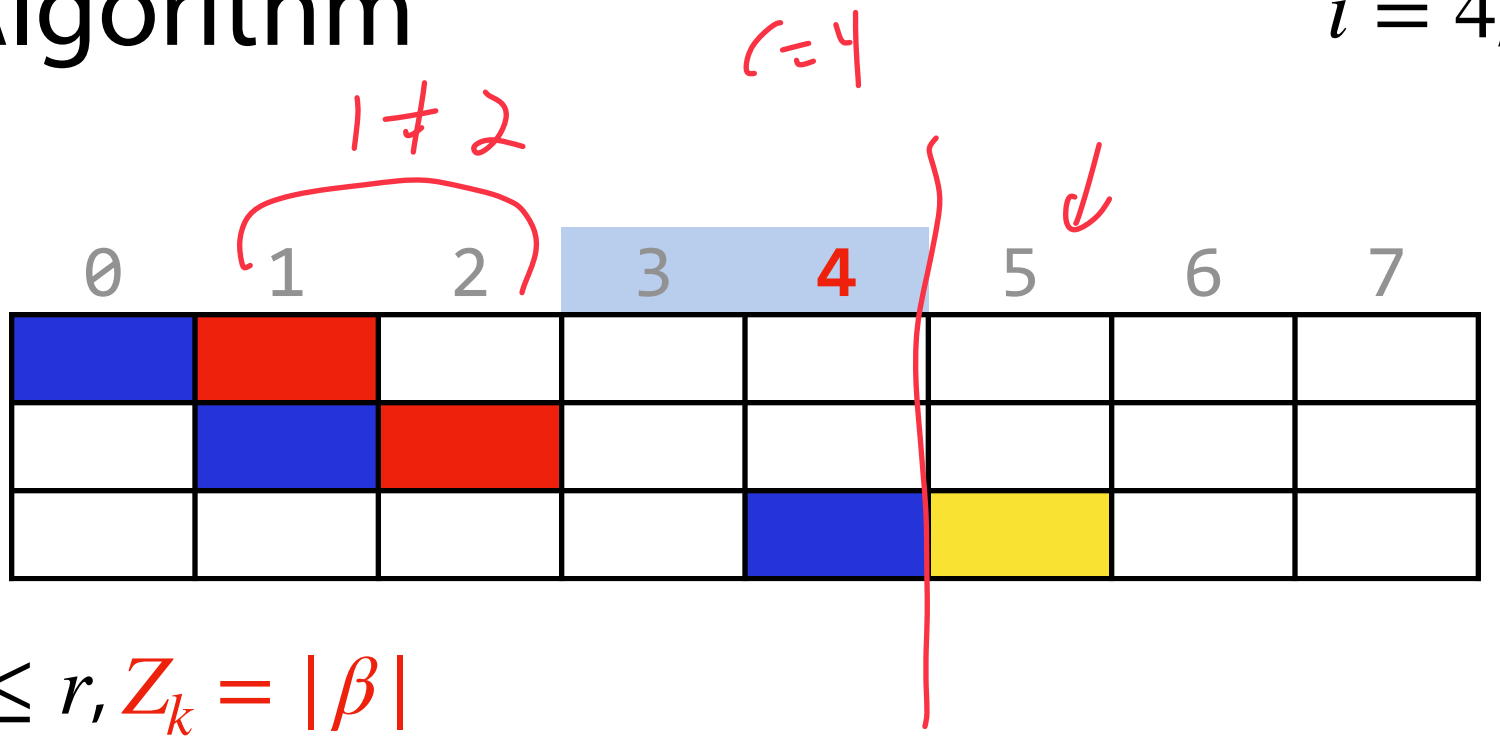


Case 2b: $i \leq r, Z_k = |\beta|$

Z_l (defined by r, l) tells us that β matches earlier.

The Z-Algorithm

$i = 4, r = 4, l = 3$



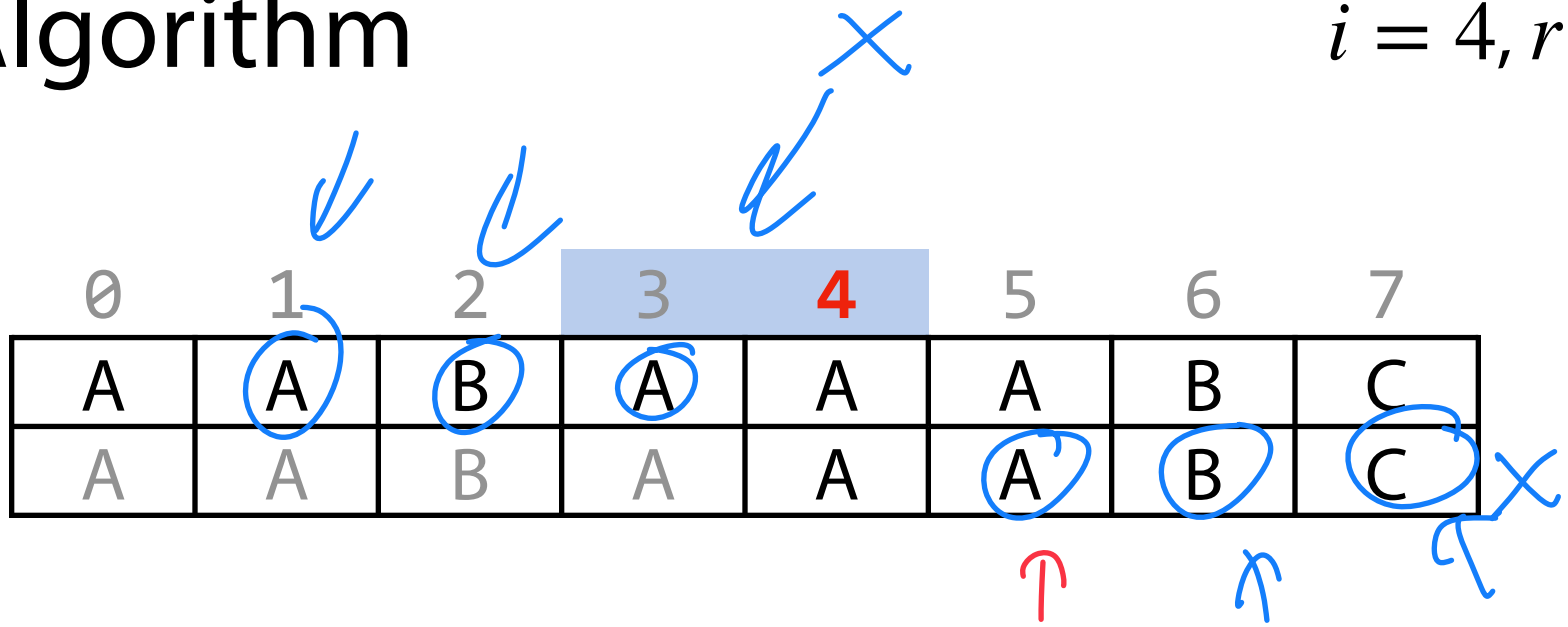
Case 2b: $i \leq r, Z_k = |\beta|$

Z_l (defined by r, l) tells us that β matches earlier.

Z_k tells us how much matches the prefix… but not everything!

The Z-Algorithm

$i = 4, r = 4, l = 3$



0	1	2	3	4	5	6	7
A	A	B	A	A	A	B	C
A	A	B	A	A	A	B	C

Case 2b: $i \leq r, Z_k = |\beta|$

$|\beta| = 1, k = 1, Z_k = 1$

$Z_i = Z\_k +$ explicit calc (r+1 vs Z_k / β)

The Z-Algorithm

$i = 3, r = 5, l = 1$

0	1	2	3	4	5	6	7
A	A	A	A	A	A	B	C
A	A	A	A	A	A	B	C

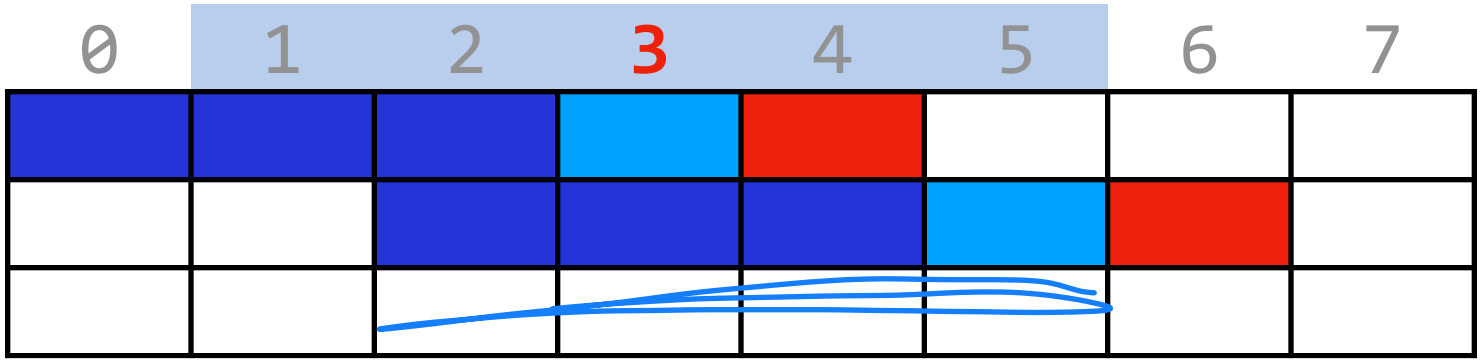
Case 2c: $i \leq r, Z_k > |\beta|$

$|\beta| =$ 3 , $k =$ 2 , $Z_k =$ 4

$Z_i =$ $|B| = 3$

The Z-Algorithm

$i = 3, r = 5, l = 1$



Case 2c: $i \leq r, Z_k > |\beta|$

$3 = 5$

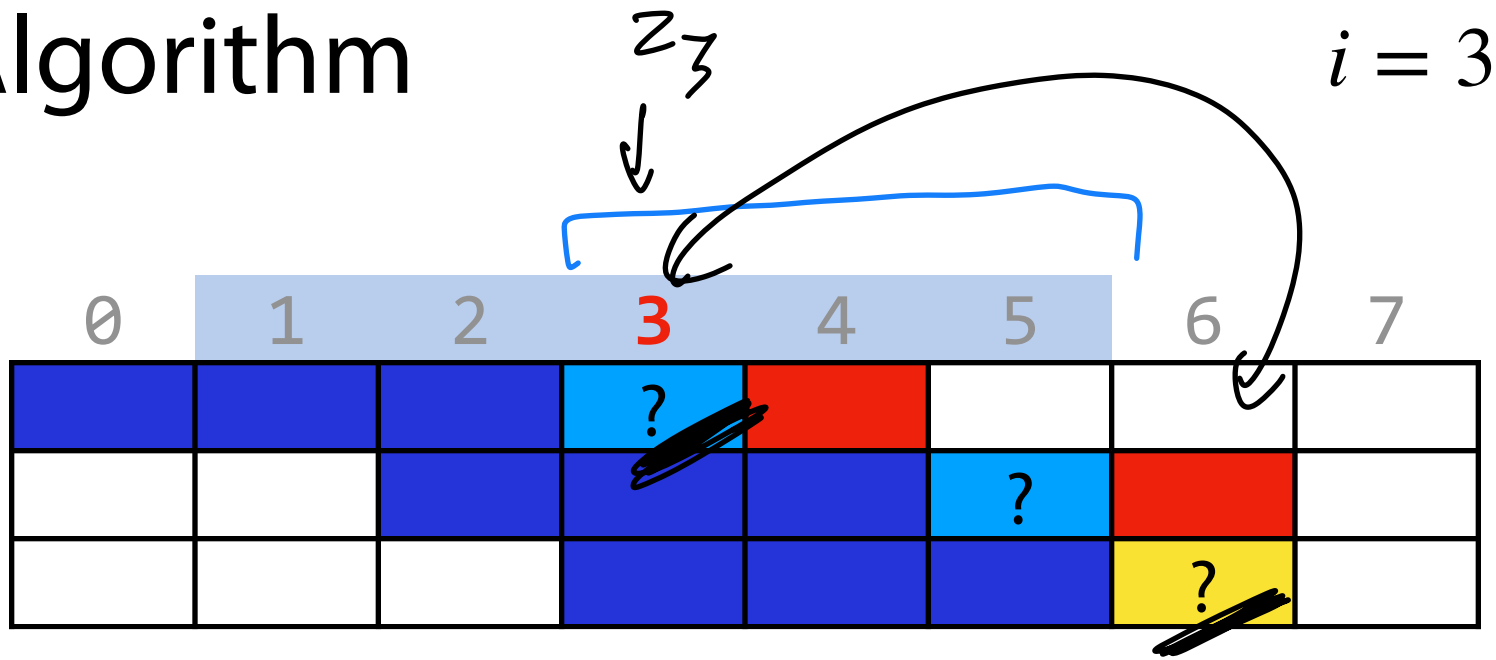
Z_k tells us how much matches the prefix.

$Z_2 = 4$

0123 == 2345

The Z-Algorithm

$i = 3, r = 5, l = 1$



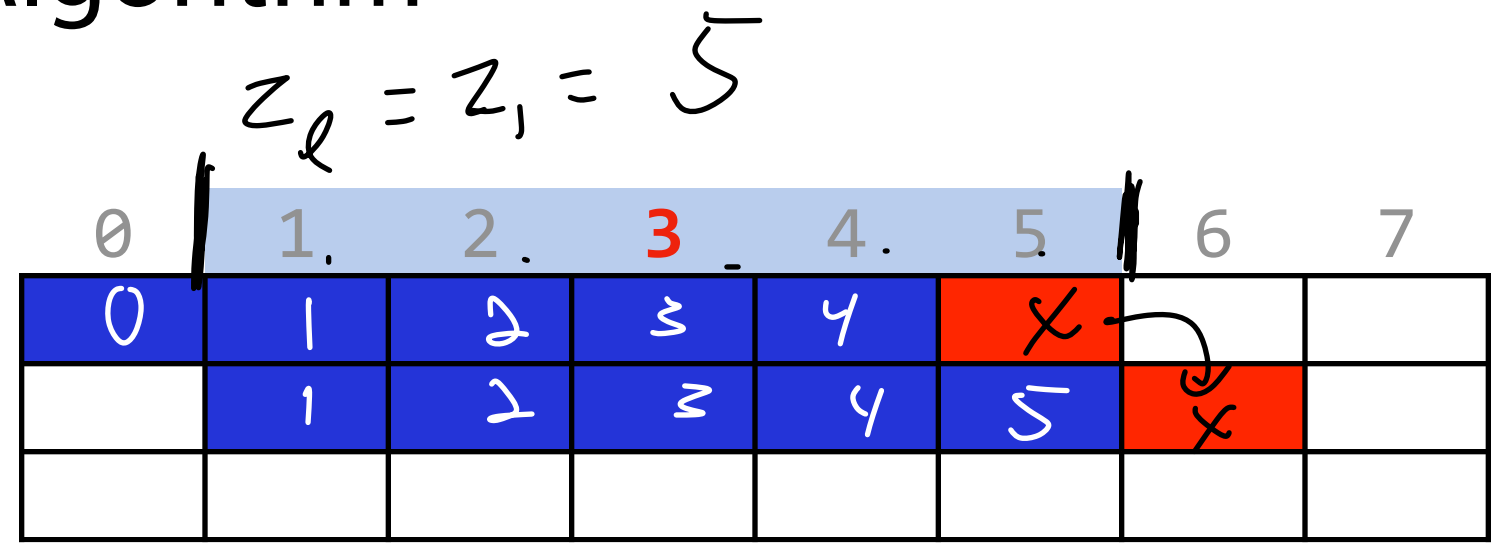
Case 2c: $i \leq r, Z_k > |\beta|$

Z_l tells us that β matches earlier. Z_k tells us how much matches the prefix.

What do we know about yellow?

The Z-Algorithm

$i = 3, r = 5, l = 1$



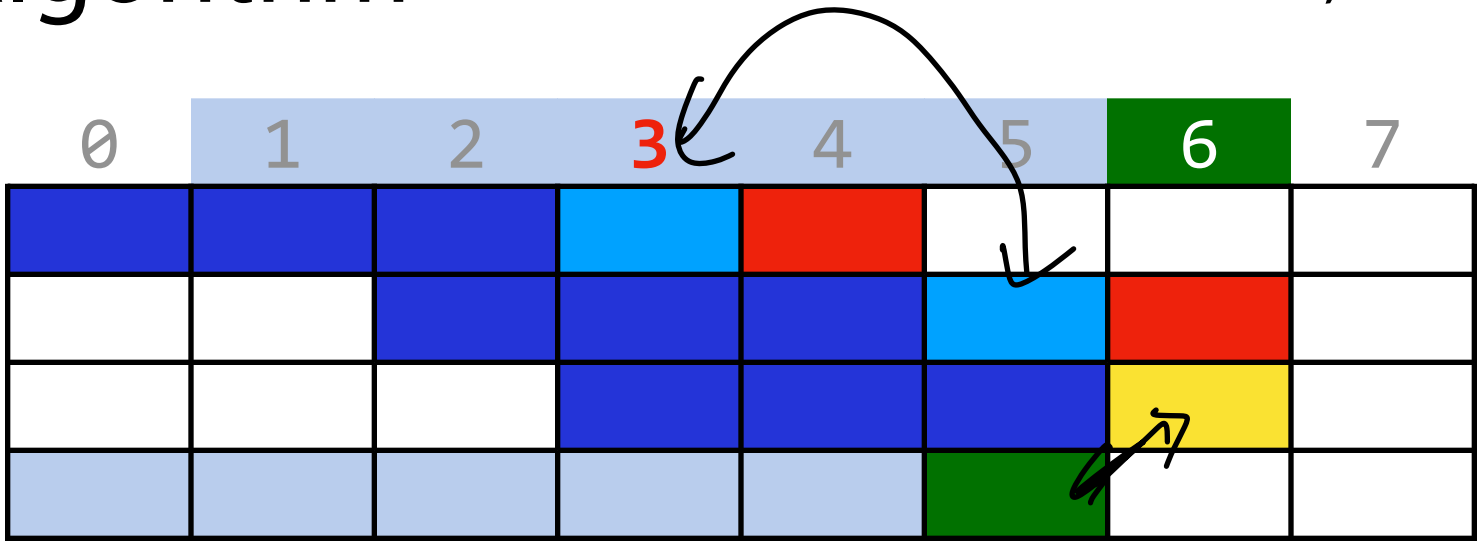
Case 2c: $i \leq r, Z_k > |\beta|$

Z_l tells us that our entire range (β included) matches earlier

… and that it failed to match the next character.

The Z-Algorithm

$i = 3, r = 5, l = 1$



Case 2c: $i \leq r, Z_k > |\beta|$

$3 = 5 \neq 6$

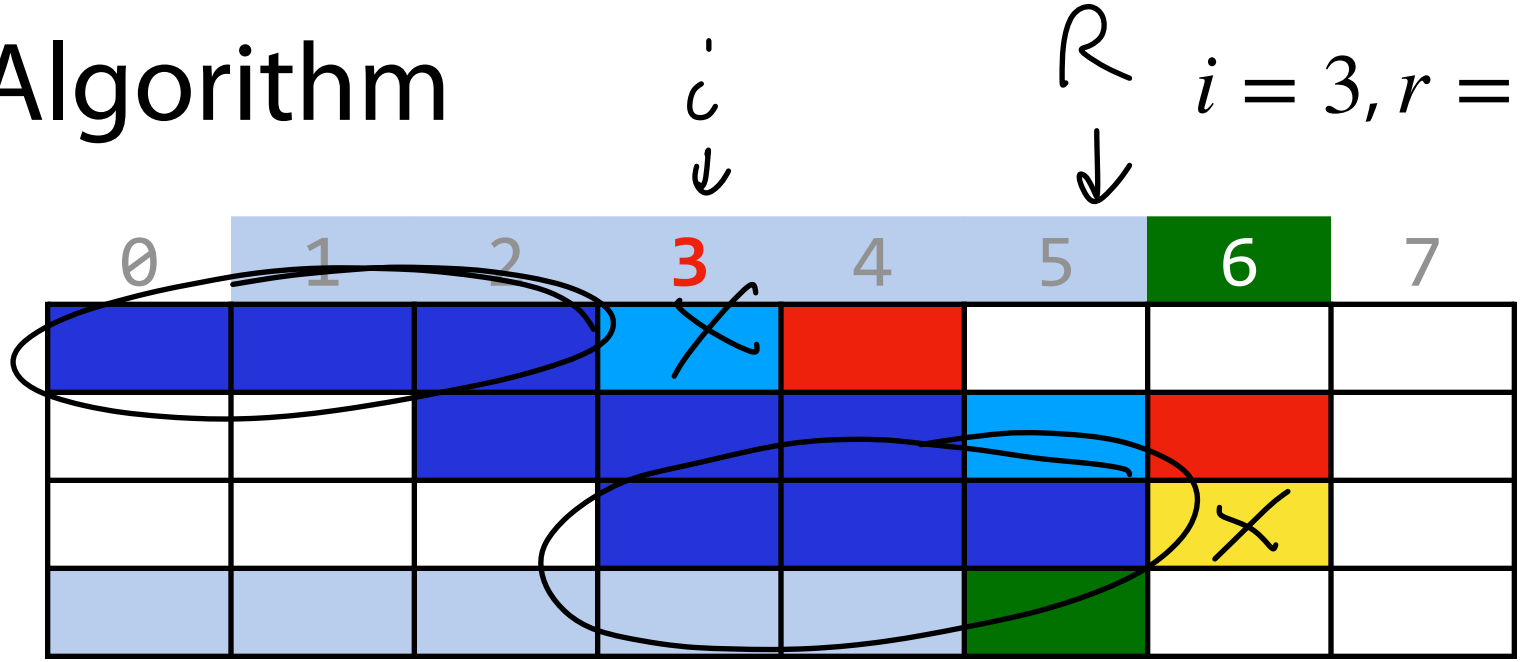
$3 \neq 6$

Z_l tells us that β matches earlier. Z_k tells us how much matches the prefix.

Z_l also tells us that yellow and green can't be equal!

The Z-Algorithm

$i = 3, r = 5, l = 1$

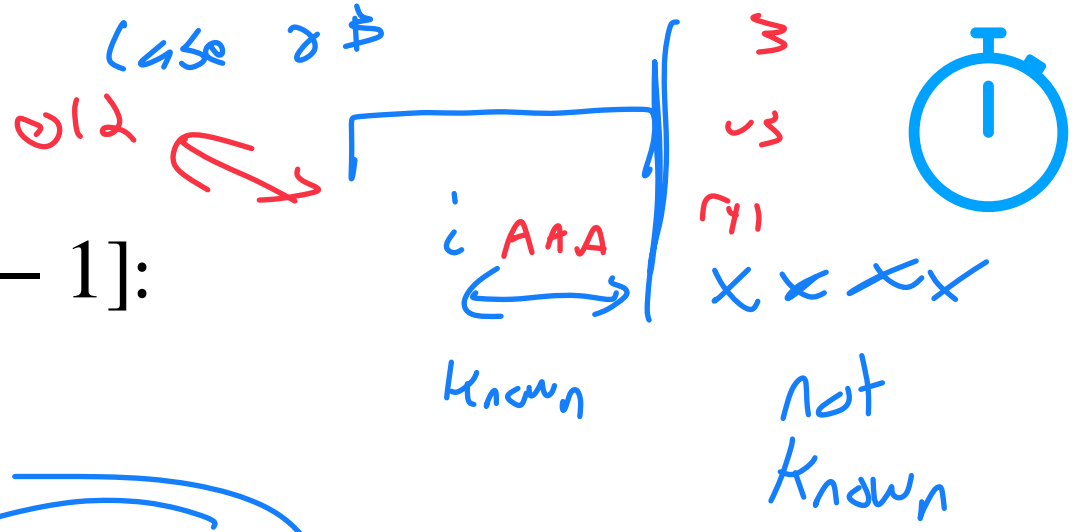


Case 2c: $i \leq r, Z_k > |\beta|$

Z_l tells us that β is our prefix. Z_k is also a previously computed prefix.

Because $Z_k > |\beta|, Z_i =$ $|\beta|$

The Z-Algorithm



Let $l = 0, r = 0$, for $i = [1, \ldots, |S| - 1]$:

Compute Z_i using irl:

Case 1 ($i > r$): Compute explicitly; update irl

Case 2 ($i \leq r$):

2a: ($Z_k < |\beta|$): $Z_i = Z_k$

2b: ($Z_k = |\beta|$): $Z_i = Z_k$ + *explicit*(indices $r + 1$ vs Z_k); update irl

2c: ($Z_k > |\beta|$): $Z_i = |\beta|$

Assignment 3: a_zalg

Learning Objective:

Construct the full Z-algorithm and measure its efficiency

160 pts

Demonstrate use of Z-algorithm in pattern matching

0 pts!

Consider: Our goal is $\theta(|P| + |T|)$. Does Z-alg search match this?

Next week:

If I gave you the pattern I was interested in ahead of time, what could you pre-compute to speed up search?

Ex: I'm going to try to look up the word **'arrays'** — but you don't know what text I'm going to search through.