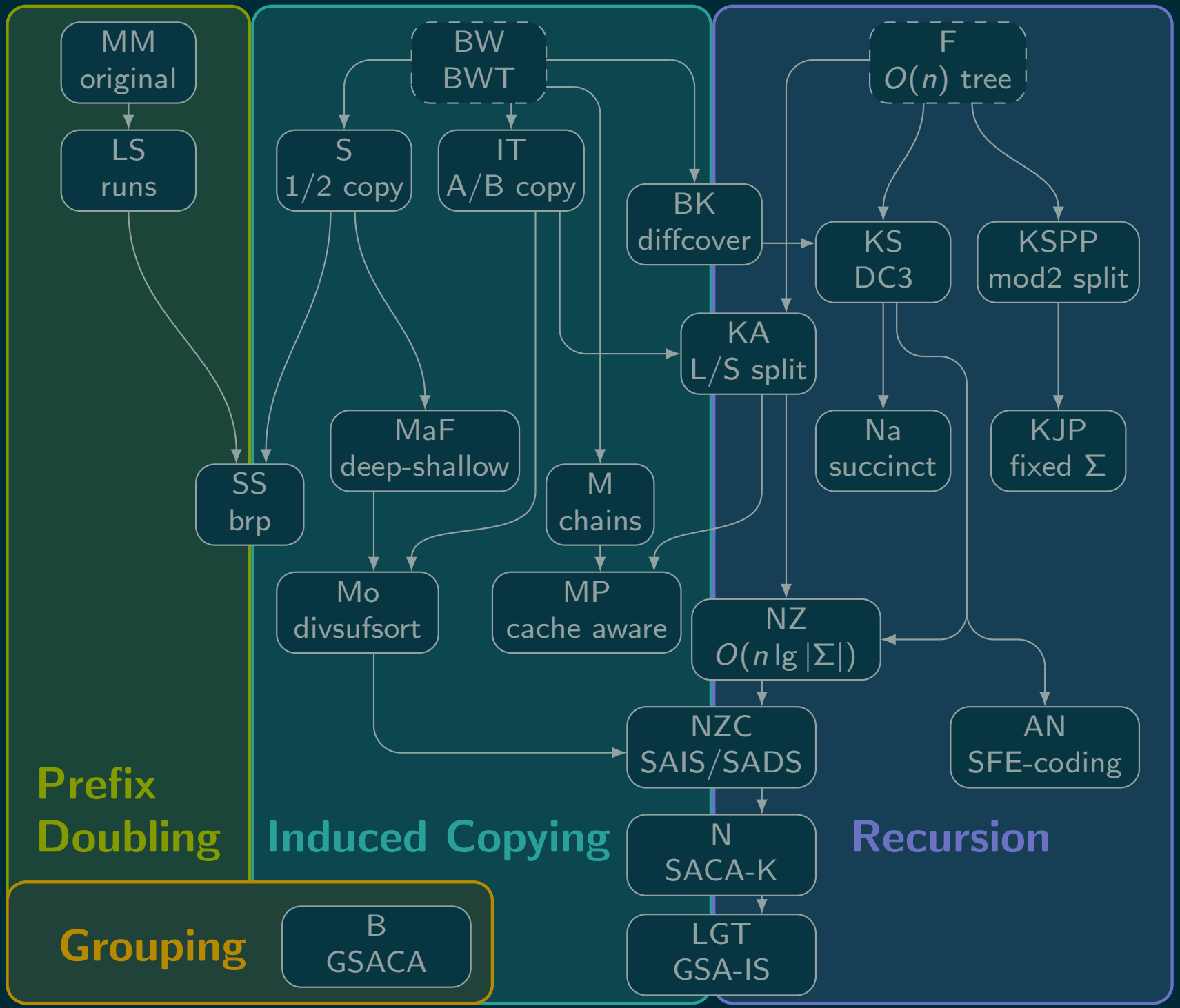
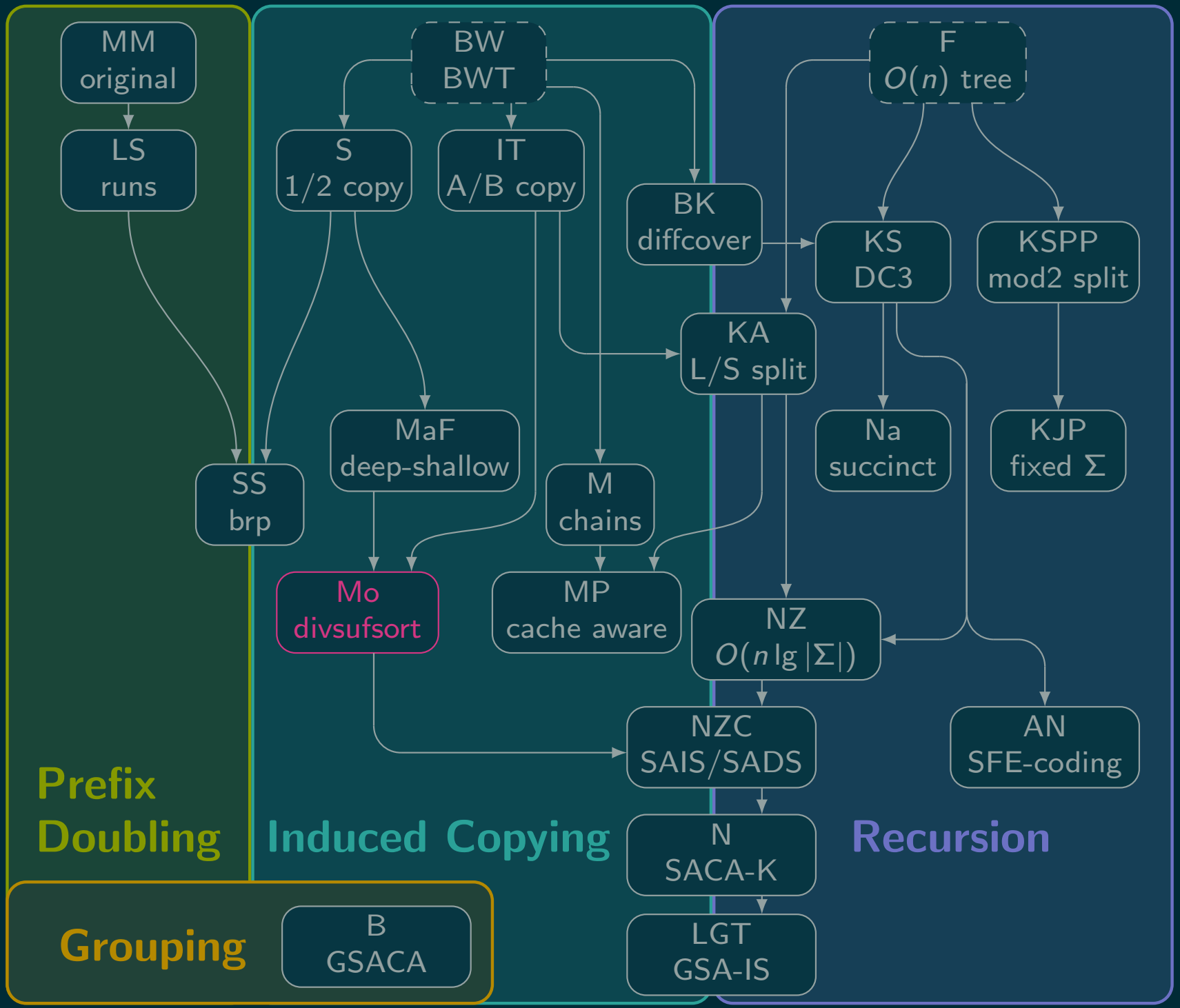
dismantling DivSufSort

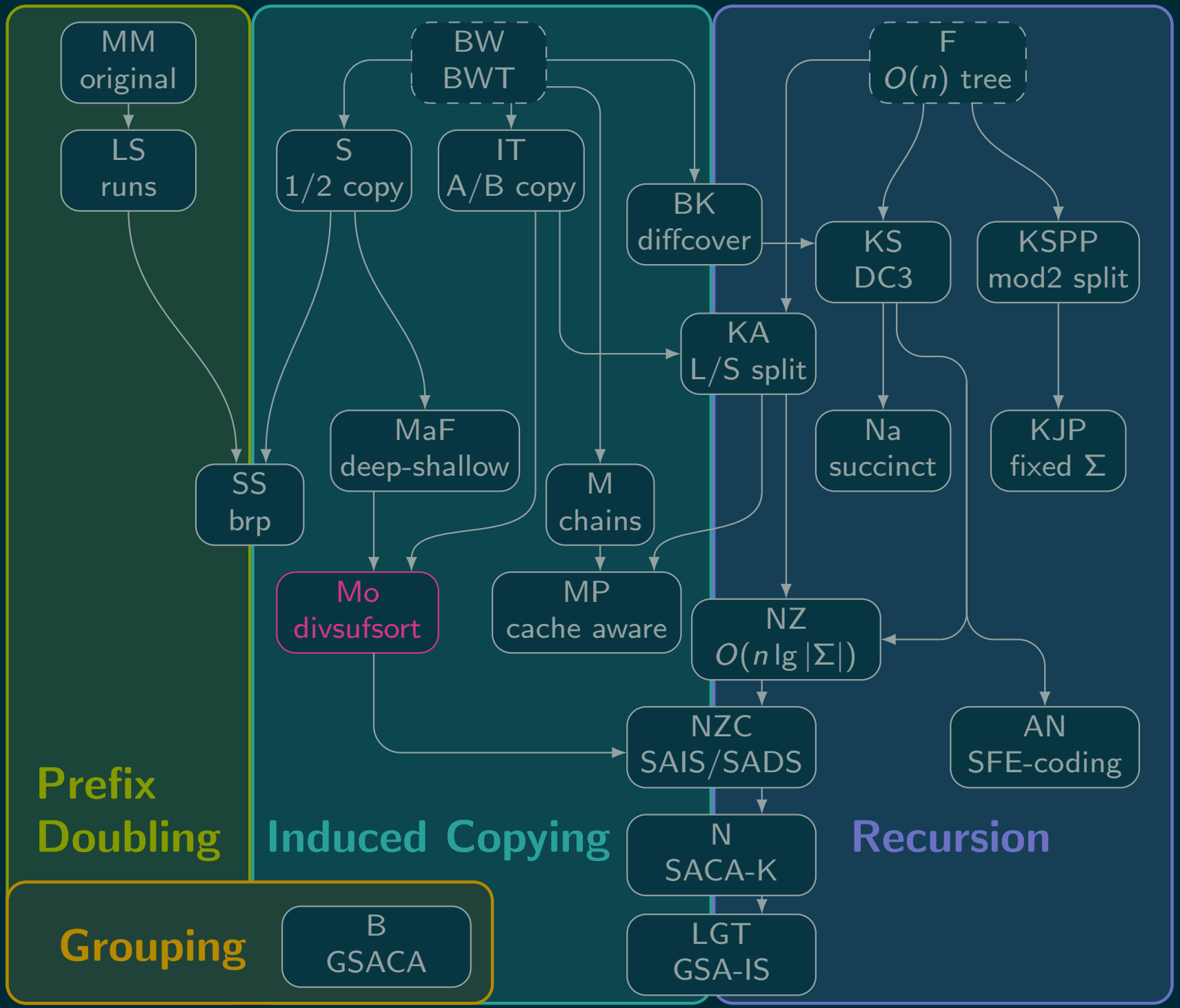
Practical Suffix and LCP Array Construction

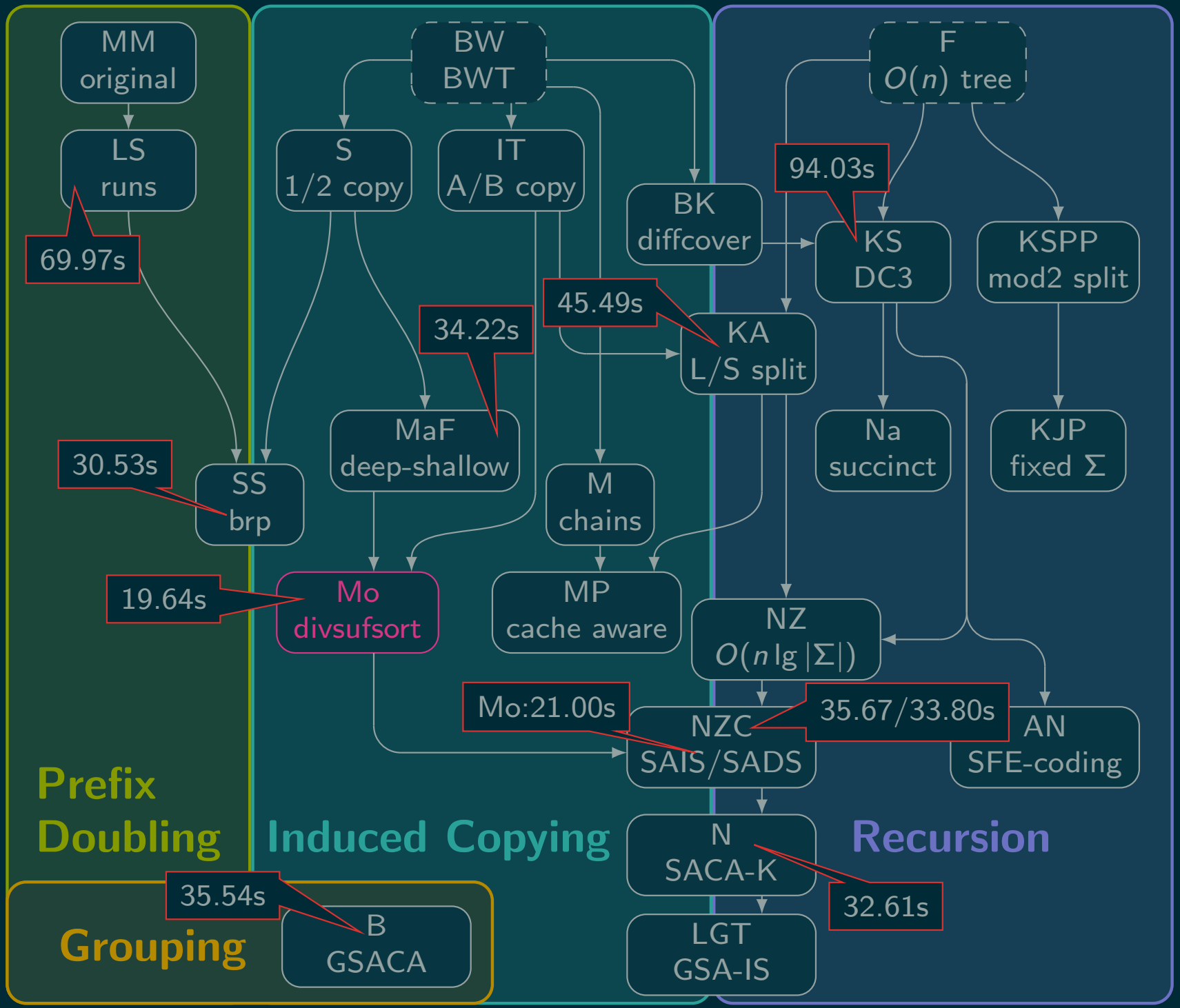
Johannes Fischer *Florian Kurpicz*

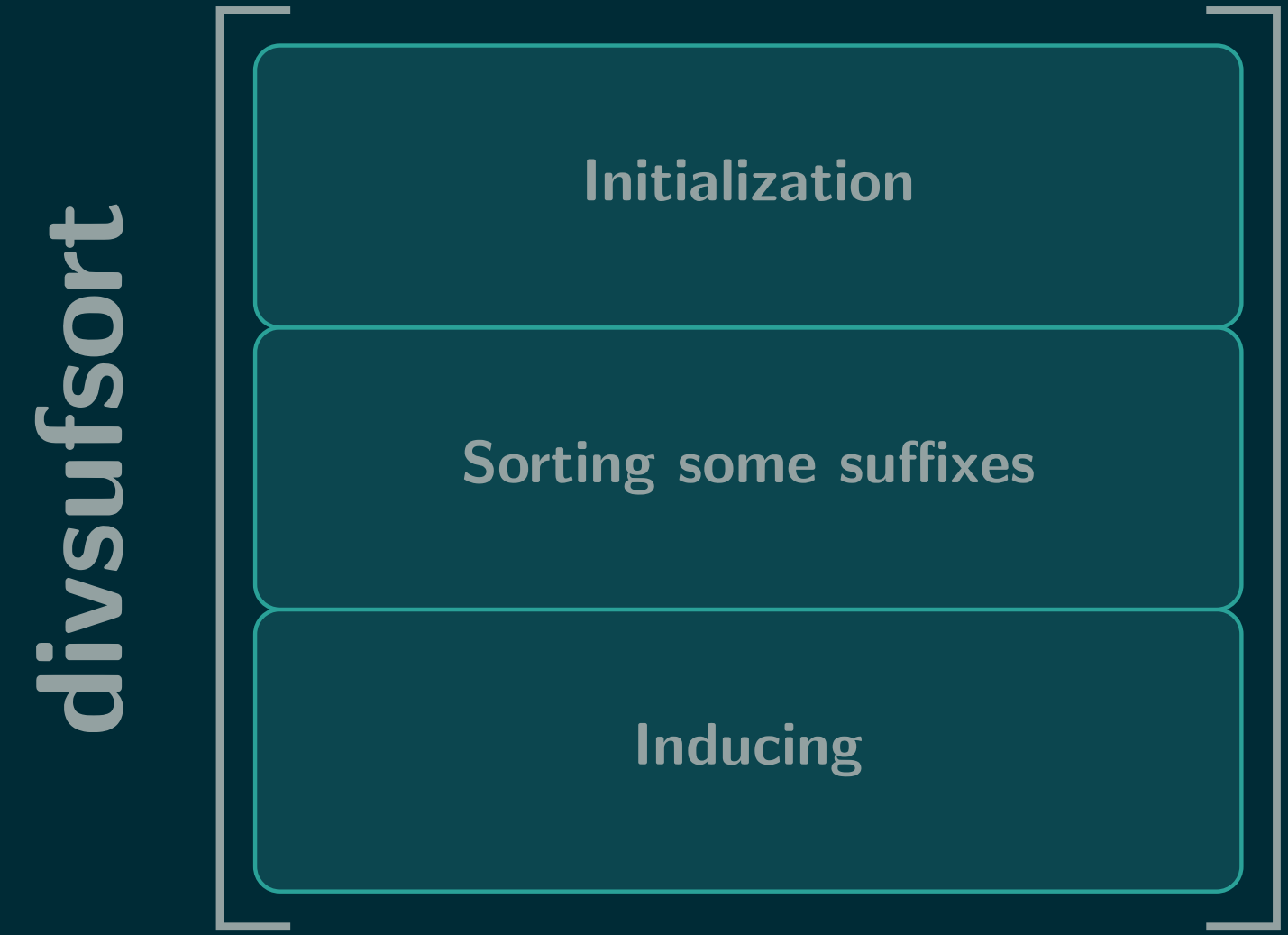

MM
original
LS
runs
BW
BWT
S
1/2 copy
IT
A/B copy
BK
diffcover
F
$O(n)$ tree
KS
DC3
KSPP
mod2 split
KA
L/S split
MaF
deep-shallow
Na
succinct
KJP
fixed Σ
SS
brp
M
chains
Mo
divsufsort
MP
cache aware
NZ
$O(n \lg |\Sigma|)$
NZC
SAIS/SADS
AN
SFE-coding
N
SACA-K
LGT
GSA-IS
B
GSACA
Prefix Doubling
Induced Copying
Recursion
Grouping
[PST07]
[BFO13]
updated
1999
2009
2011
2016

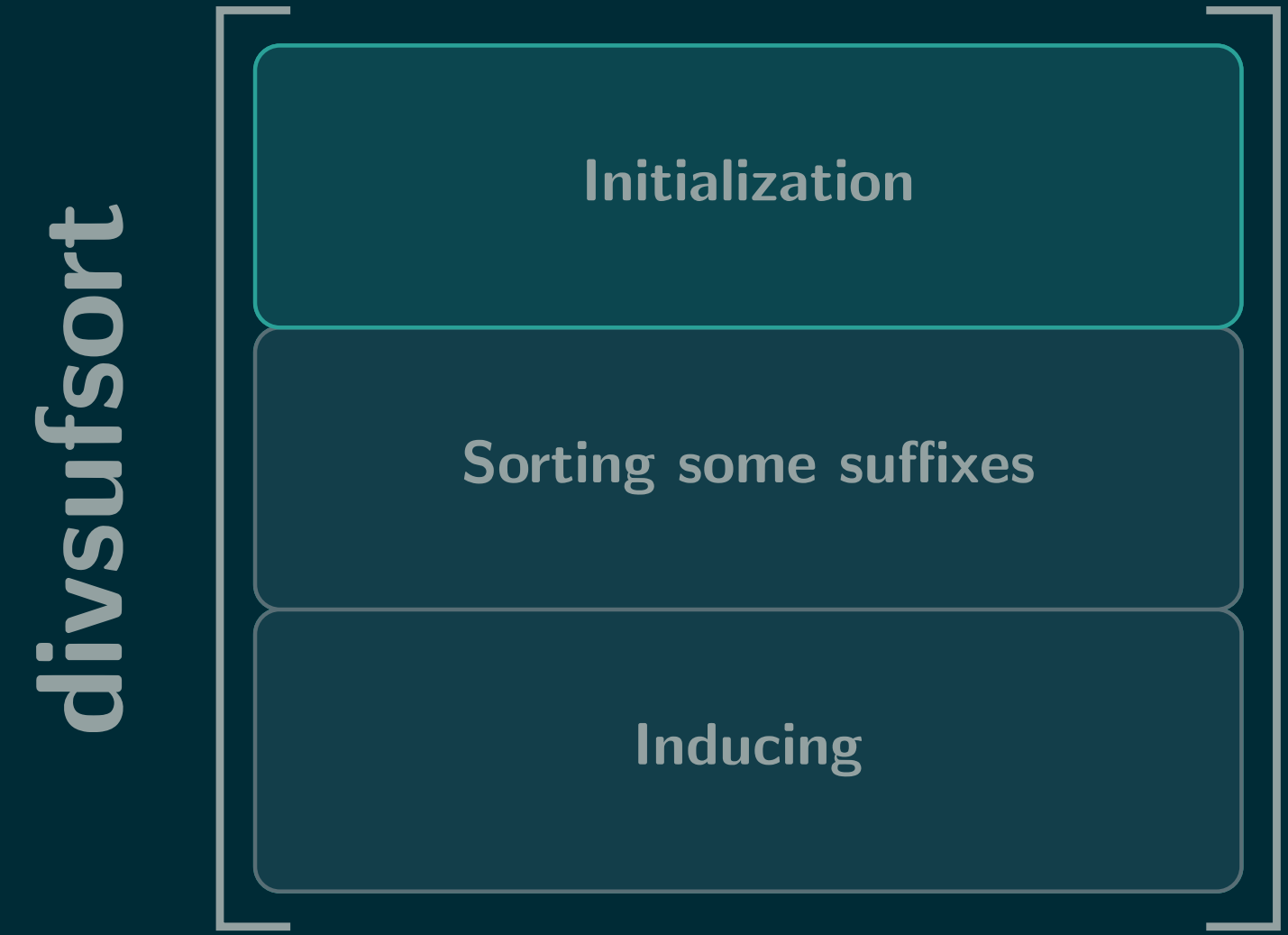

MM
original
LS
runs
BW
BWT
S
1/2 copy
IT
A/B copy
BK
diffcover
F
$O(n)$ tree
KS
DC3
KSPP
mod2 split
KA
L/S split
MaF
deep-shallow
Na
succinct
KJP
fixed Σ
SS
brp
M
chains
Mo
divsufsort
MP
cache aware
NZ
$O(n \lg |\Sigma|)$
NZC
SAIS/SADS
AN
SFE-coding
N
SACA-K
LGT
GSA-IS
B
GSACA
Prefix Doubling
Induced Copying
Recursion
Grouping
[PST07]
[BFO13]
updated
1999
2009
2011
2016

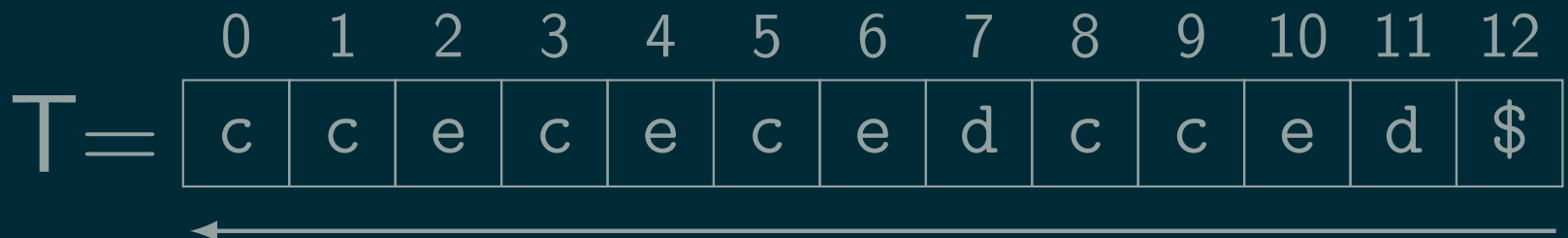
It's fast

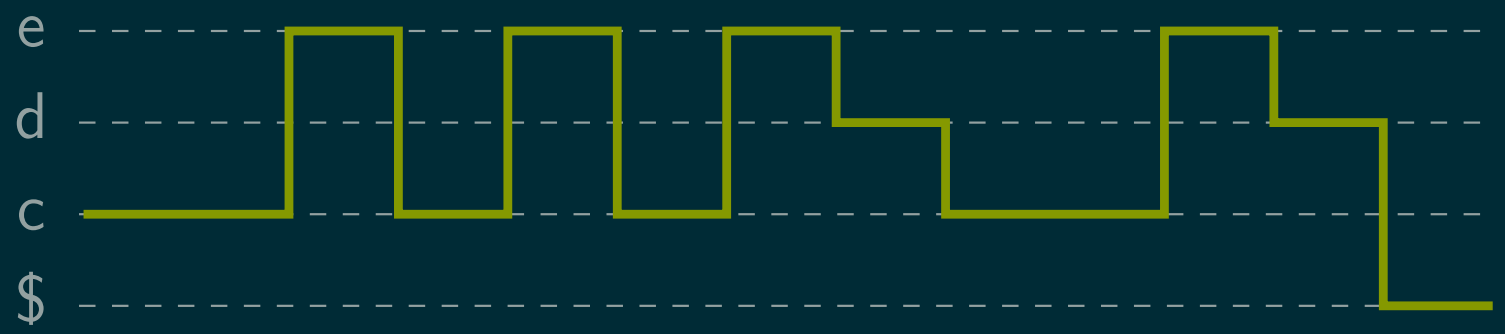

MM
original
LS
runs
BW
BWT
S
1/2 copy
IT
A/B copy
BK
diffcover
F
$O(n)$ tree
KS
DC3
KSPP
mod2 split
KA
L/S split
MaF
deep-shallow
Na
succinct
KJP
fixed Σ
SS
brp
M
chains
Mo
divsufsort
MP
cache aware
NZ
$O(n \lg |\Sigma|)$
NZC
SAIS/SADS
AN
SFE-coding
N
SACA-K
LGT
GSA-IS
B
GSACA
Prefix Doubling
Induced Copying
Recursion
Grouping
[PST07]
[BFO13]
updated
1999
2009
2011
2016


MM
original
LS
runs
69.97s
BW
BWT
S
1/2 copy
IT
A/B copy
BK
diffcover
F
$O(n)$ tree
94.03s
KS
DC3
KSPP
mod2 split
45.49s
34.22s
KA
L/S split
MaF
deep-shallow
Na
succinct
KJP
fixed Σ
30.53s
SS
brp
M
chains
19.64s
Mo
divsufsort
MP
cache aware
NZ
$O(n \lg |\Sigma|)$
Mo:21.00s
NZC
SAIS/SADS
35.67/33.80s
AN
SFE-coding
Prefix
Doubling
Induced Copying
Recursion
N
SACA-K
32.61s
35.54s
B
GSACA
Grouping
LGT
GSA-IS
[PST07]
[BFO13]
updated
1999
2009
2011
2016


divsufsort
Initialization
Sorting some suffixes
Inducing


divsufsort
Initialization
Sorting some suffixes
Inducing





A: $T[i] > T[i+1]$ and B: $T[i] < T[i+1]$

$T[i] = T[i+1]$: same as before [IT1999]

A: $T[i] > T[i+1]$ and B: $T[i] < T[i+1]$

$T[i] = T[i+1]$: same as before [IT1999]

A: $T[i] > T[i+1]$ and B: $T[i] < T[i+1]$

$T[i] = T[i+1]$: same as before [IT1999]

A: $T[i] > T[i+1]$ and B: $T[i] < T[i+1]$

$T[i] = T[i+1]$: same as before [IT1999]

A: $T[i] > T[i+1]$ and B: $T[i] < T[i+1]$

$T[i] = T[i+1]$: same as before [IT1999]

A: $T[i] > T[i+1]$ and B: $T[i] < T[i+1]$

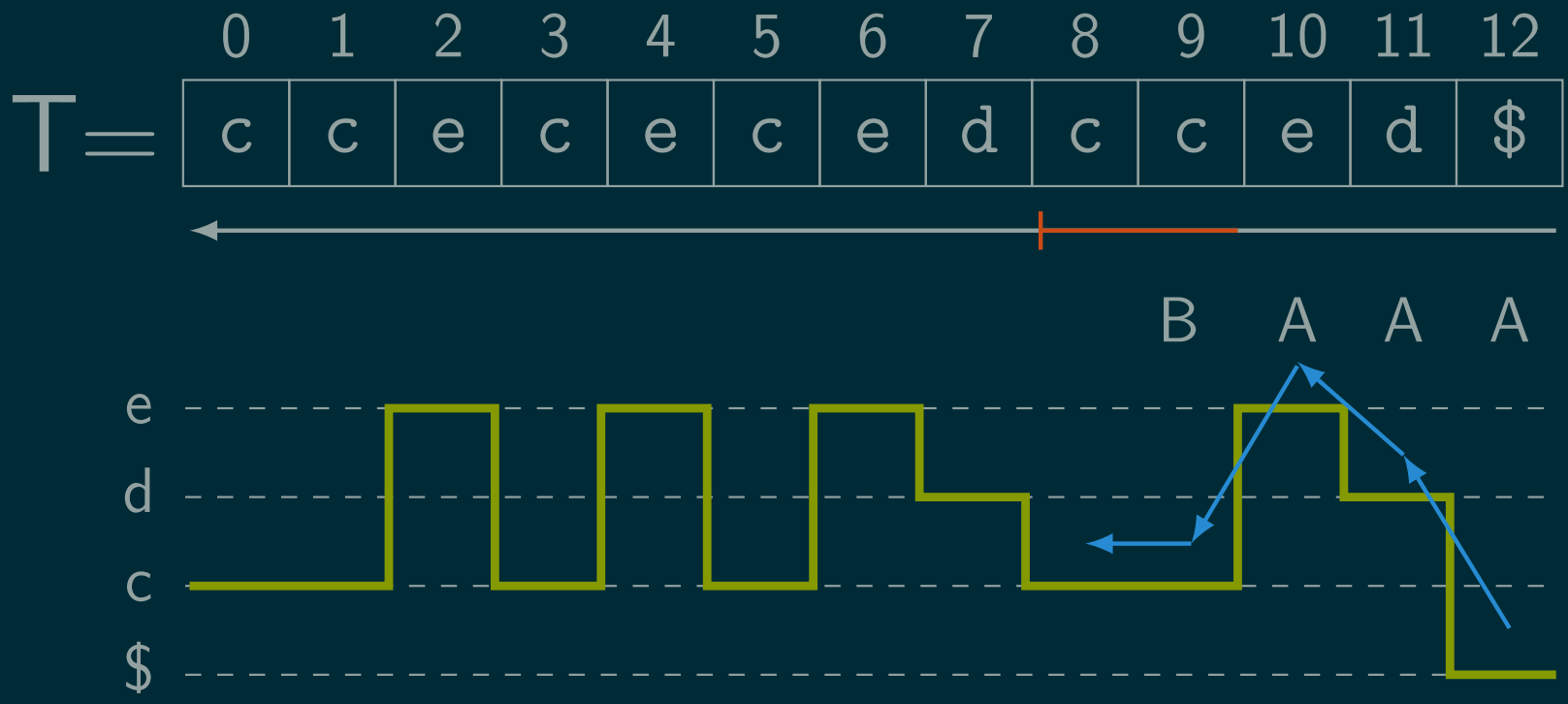
$T[i] = T[i+1]$: same as before [IT1999]

A: $T[i] > T[i+1]$ and B: $T[i] < T[i+1]$

$T[i] = T[i+1]$: same as before [IT1999]

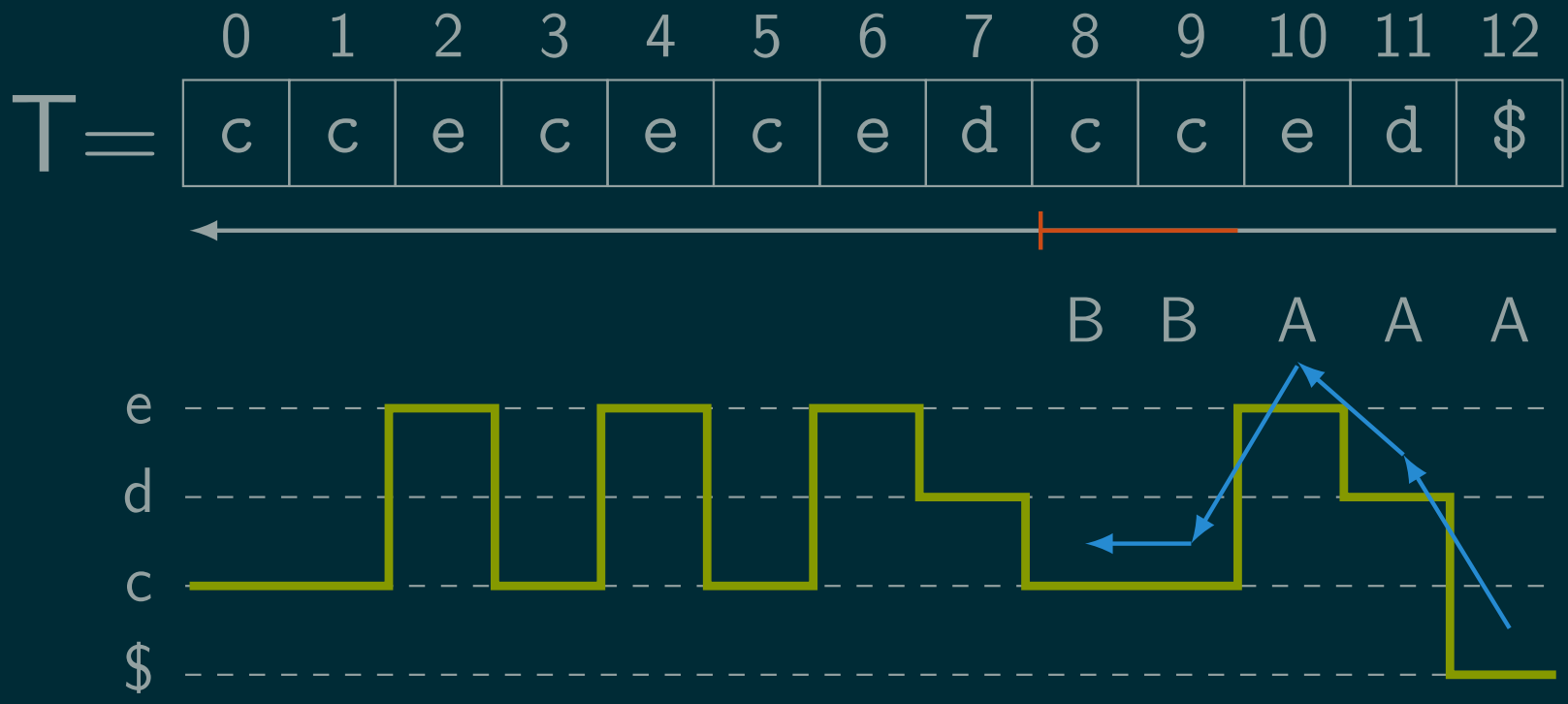
A: $T[i] > T[i+1]$ and B: $T[i] < T[i+1]$

$T[i] = T[i+1]$: same as before [IT1999]



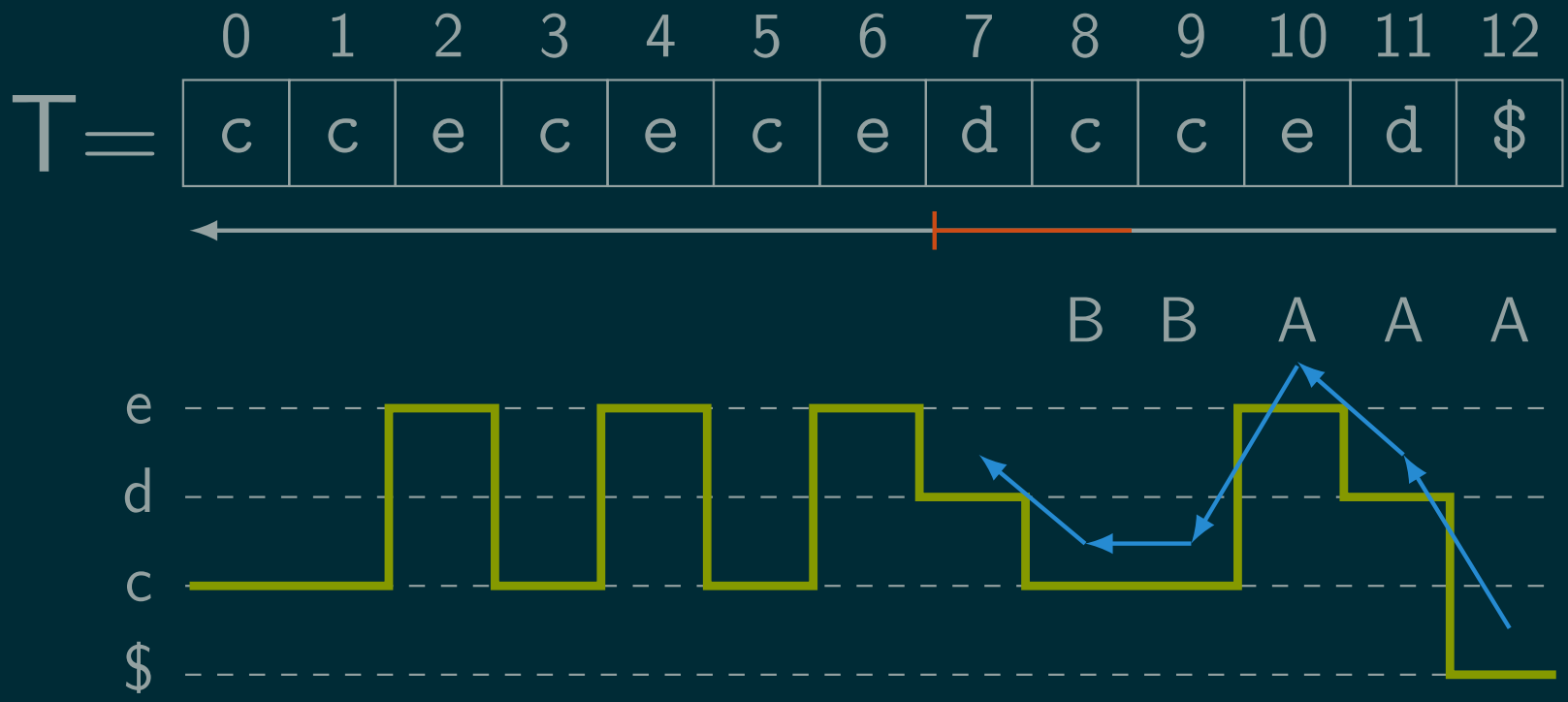
A: $T[i] > T[i+1]$ and B: $T[i] < T[i+1]$

$T[i] = T[i+1]$: same as before [IT1999]



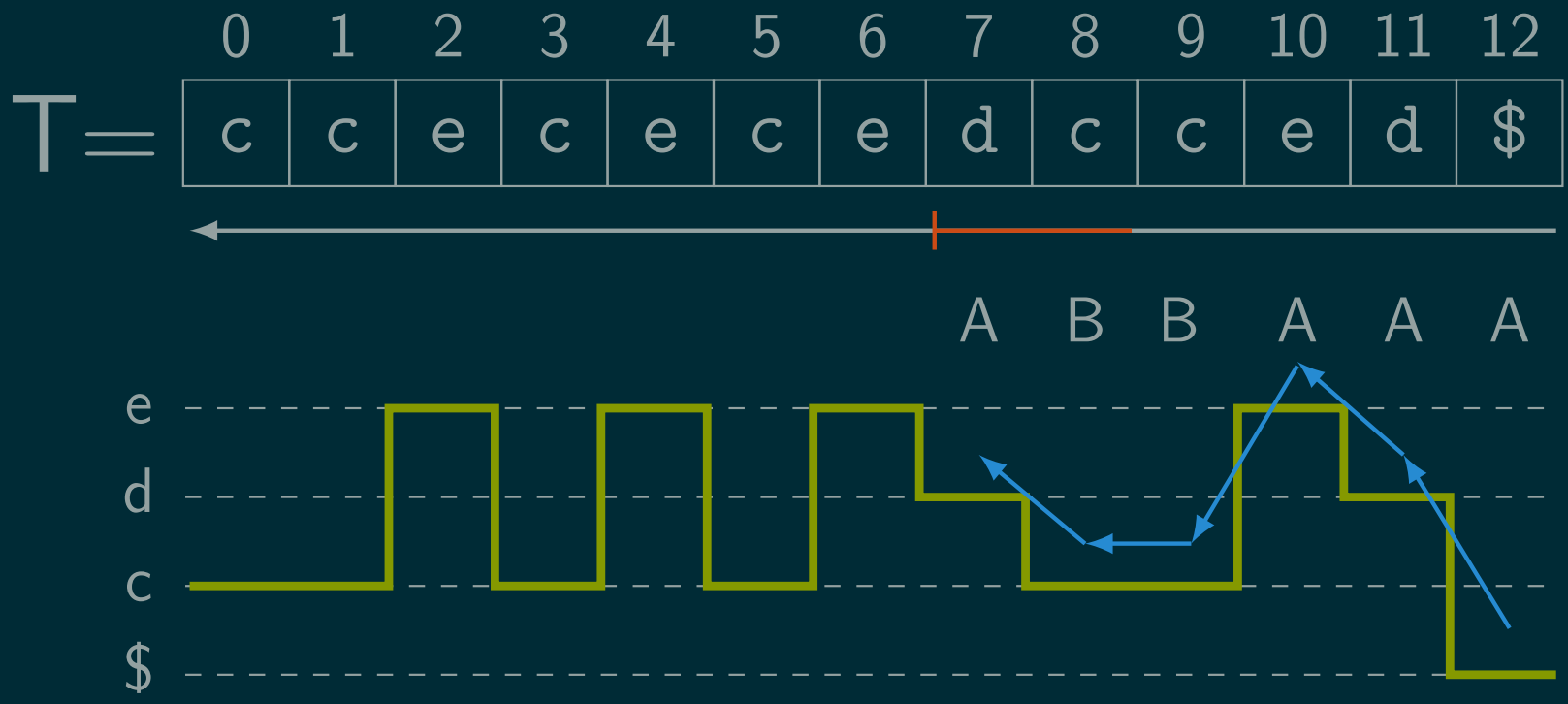
A: $T[i] > T[i+1]$ and B: $T[i] < T[i+1]$

$T[i] = T[i+1]$: same as before [IT1999]



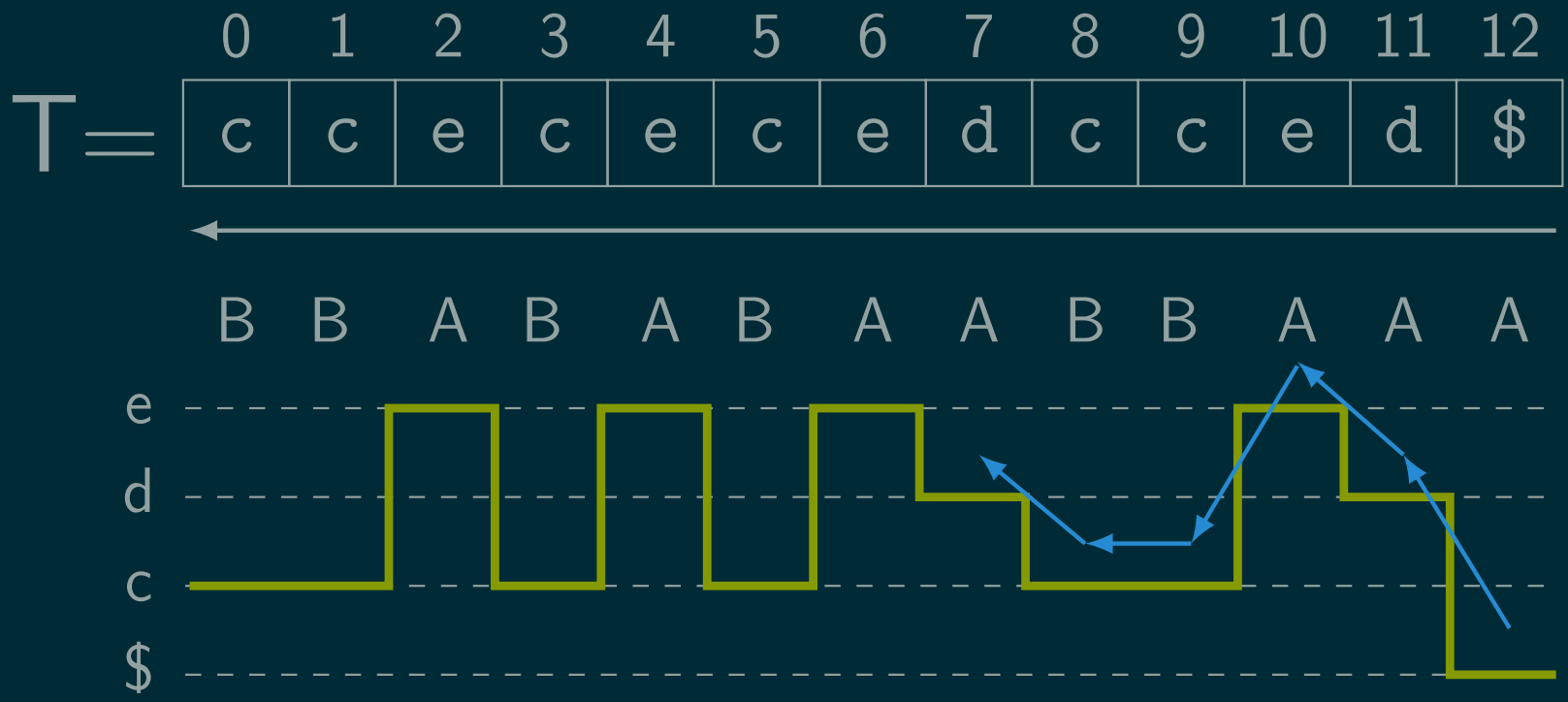
A: $T[i] > T[i+1]$ and B: $T[i] < T[i+1]$

$T[i] = T[i+1]$: same as before [IT1999]



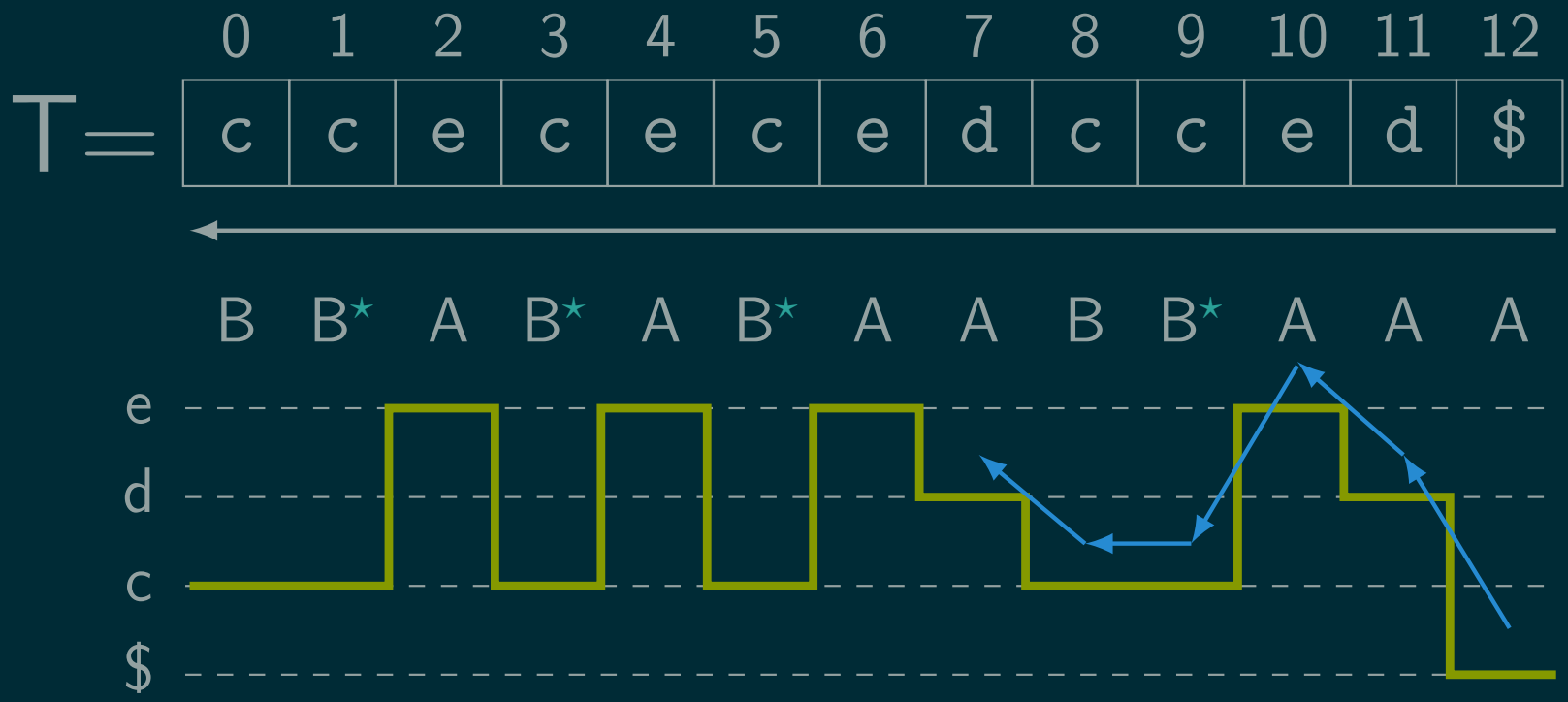
A: $T[i] > T[i+1]$ and B: $T[i] < T[i+1]$

$T[i] = T[i+1]$: same as before [IT1999]



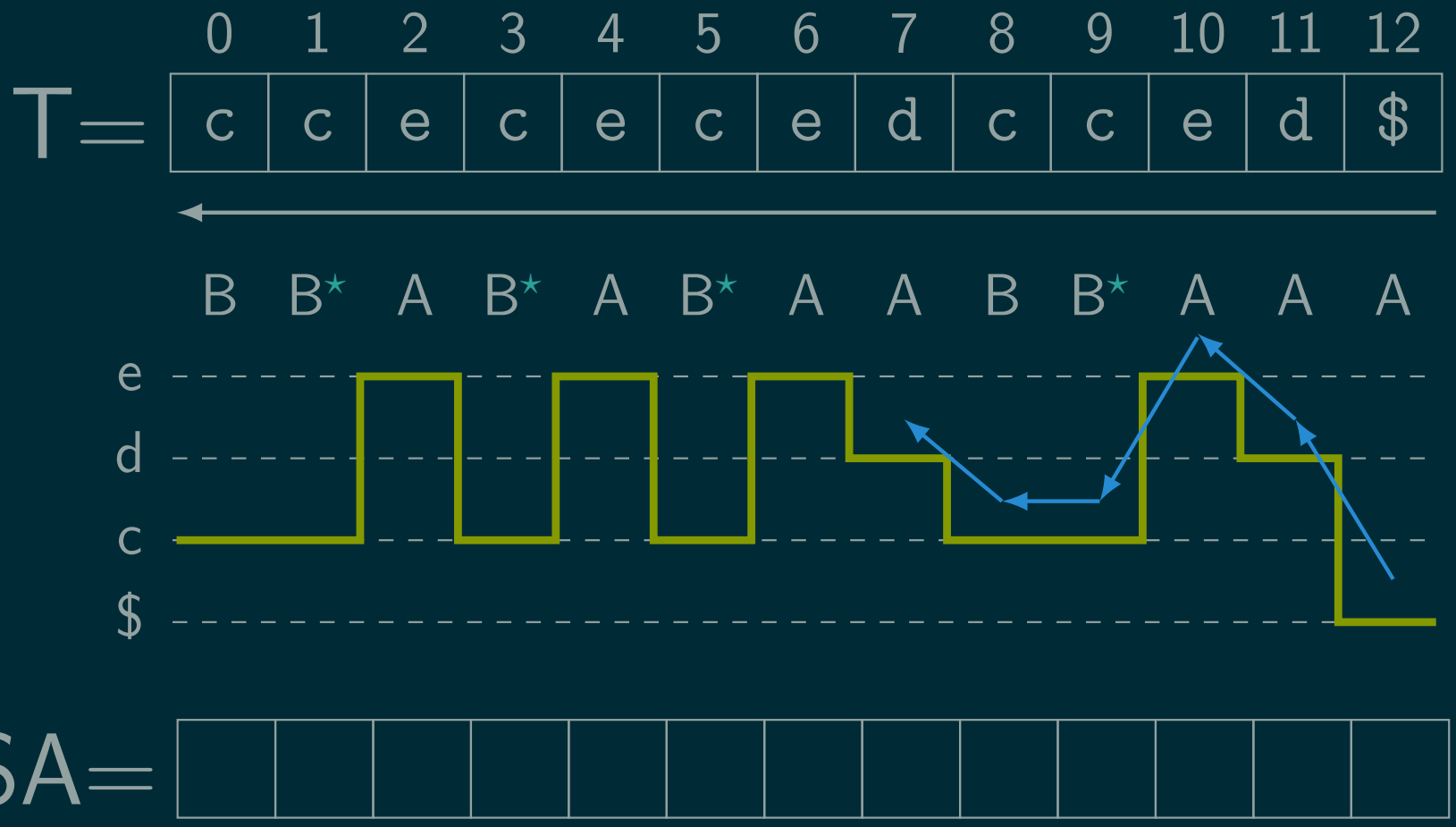
A: $T[i] > T[i+1]$ and B: $T[i] < T[i+1]$

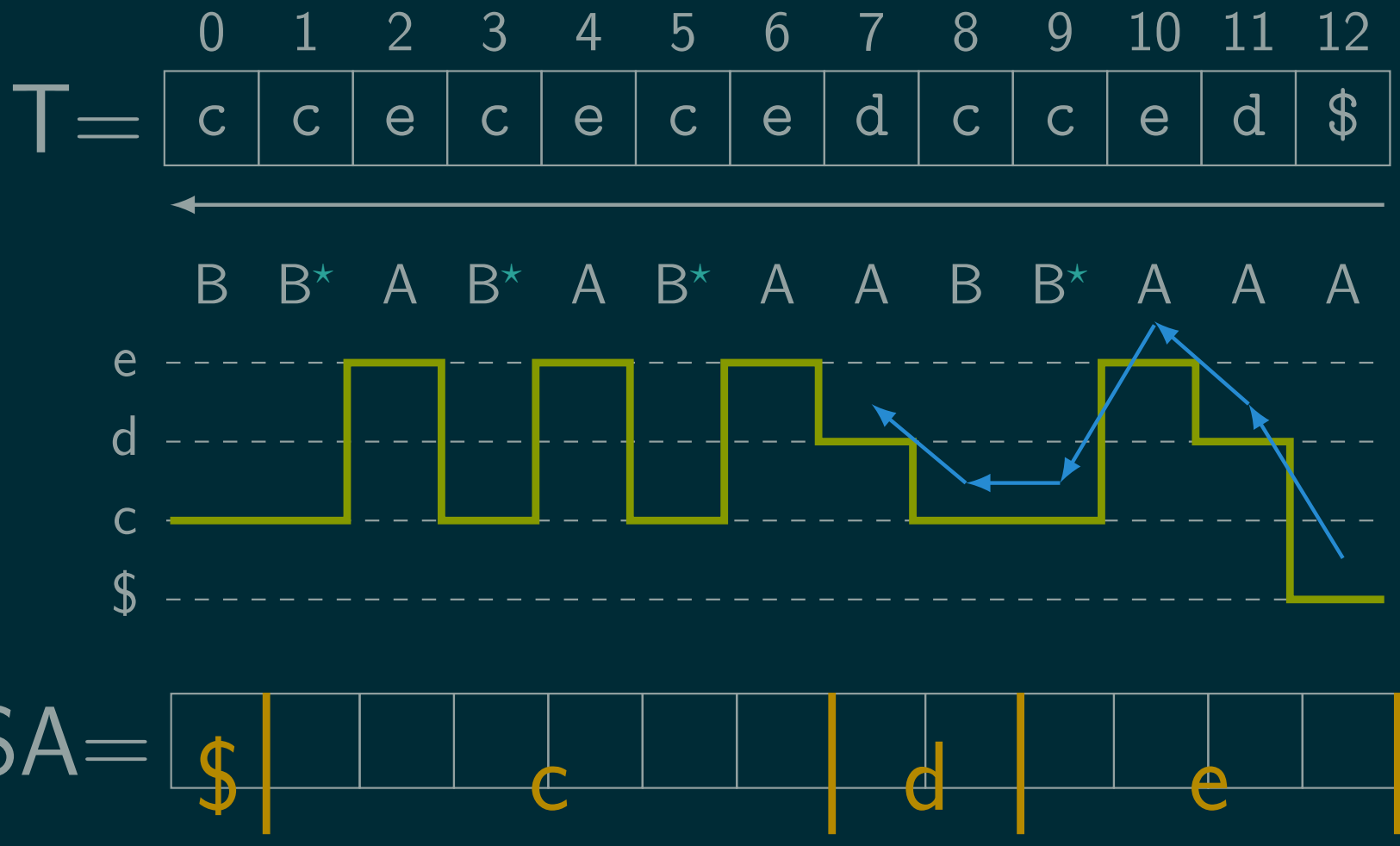
$T[i] = T[i+1]$: same as before [IT1999]

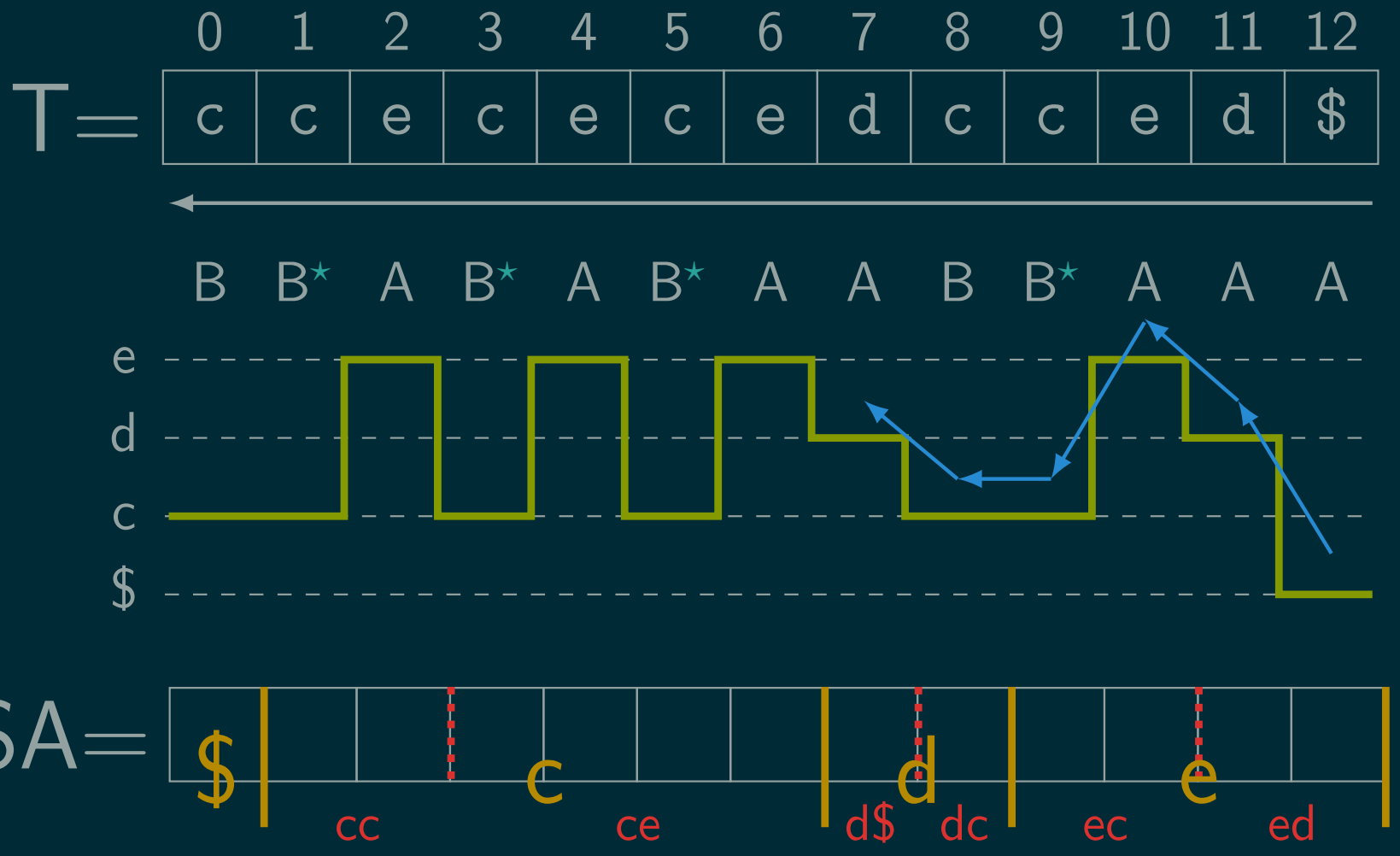


A: $T[i] > T[i+1]$ and B: $T[i] < T[i+1]$

$T[i] = T[i+1]$: same as before [IT1999]

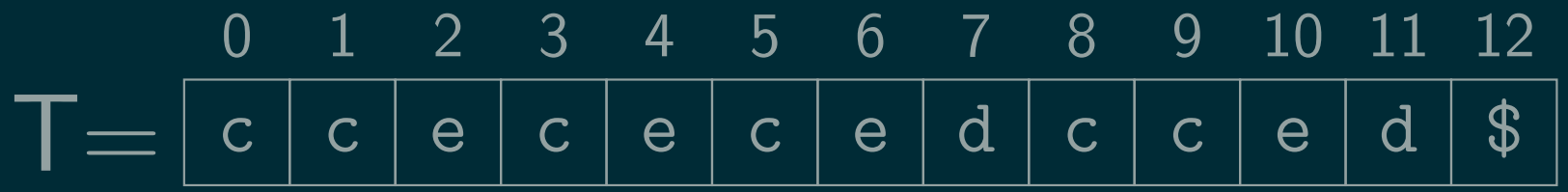

0 1 2 3 4 5 6 7 8 9 10 11 12
T=
c c e c e c e d c c e d $
B B* A B* A B* A A B B* A A A
e
d
c
$
SA=


0 1 2 3 4 5 6 7 8 9 10 11 12
T=
c c e c e c e d c c e d $
B B* A B* A B* A A B B* A A A
e
d
c
$
SA=
$
c
d
e


0 1 2 3 4 5 6 7 8 9 10 11 12
T=
c c e c e c e d c c e d $
B B* A B* A B* A A B B* A A A
e
d
c
$
SA=
$
c
d
e
cc
ce
d$
dc
ec
ed

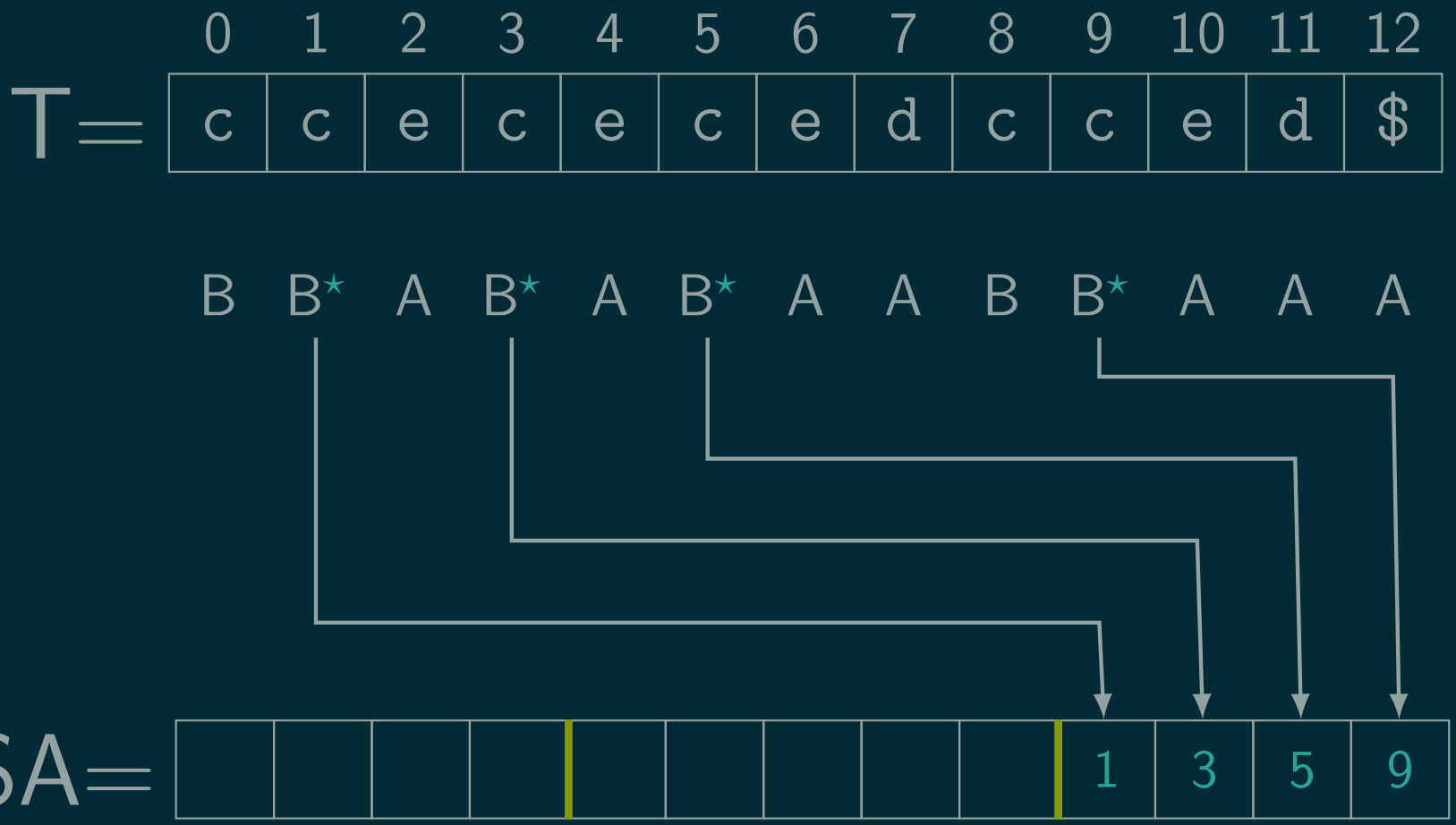
	0	1	2	3	4	5	6	7	8	9	10	11	12
T=	c	c	e	c	e	c	e	d	c	c	e	d	$
	B	B*	A	B*	A	B*	A	A	B	B*	A	A	A

SA=

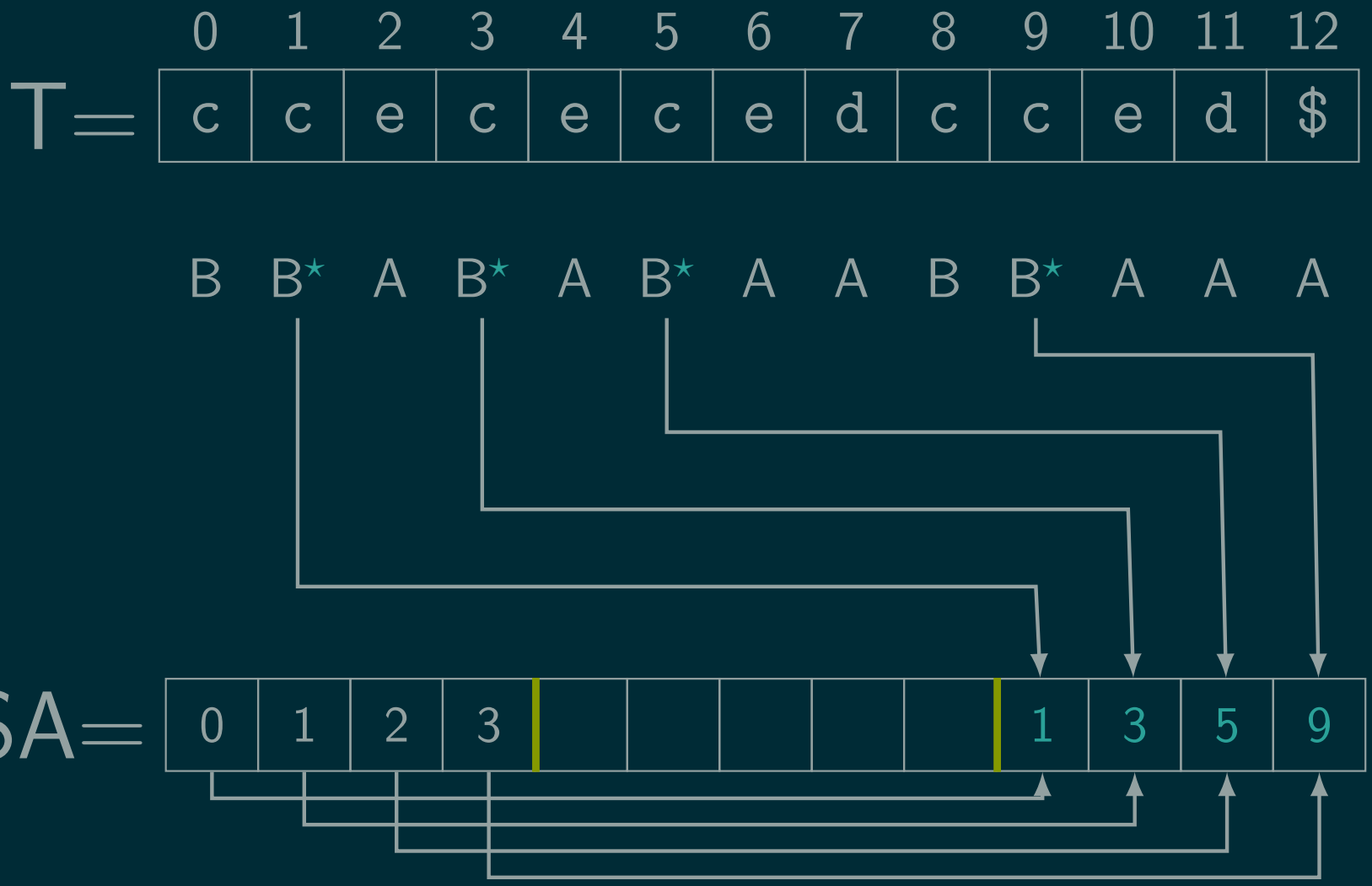


B B⋆ A B⋆ A B⋆ A A B B⋆ A A A

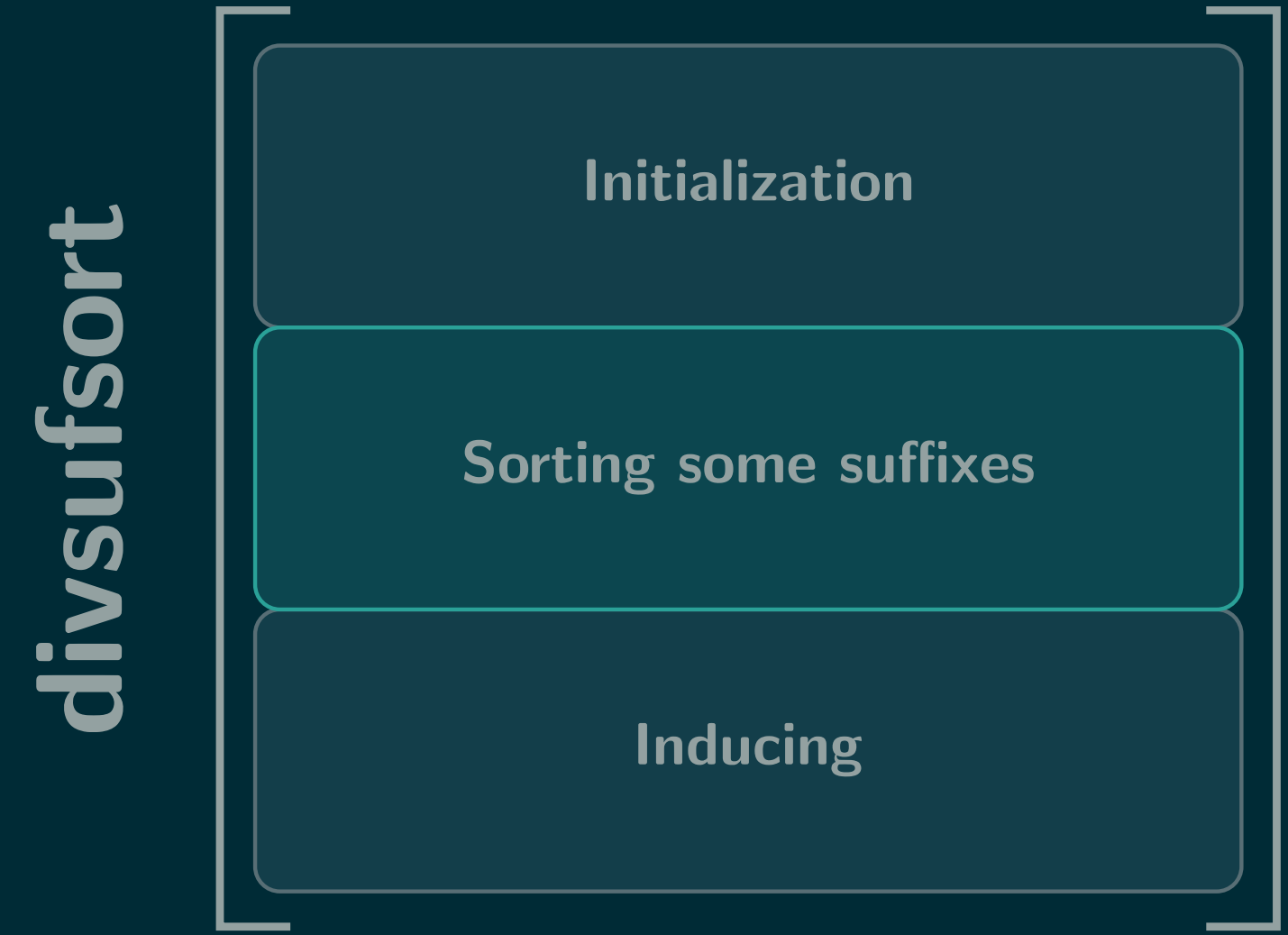
$$\#\mathrm{B}^\star \leq n/2$$

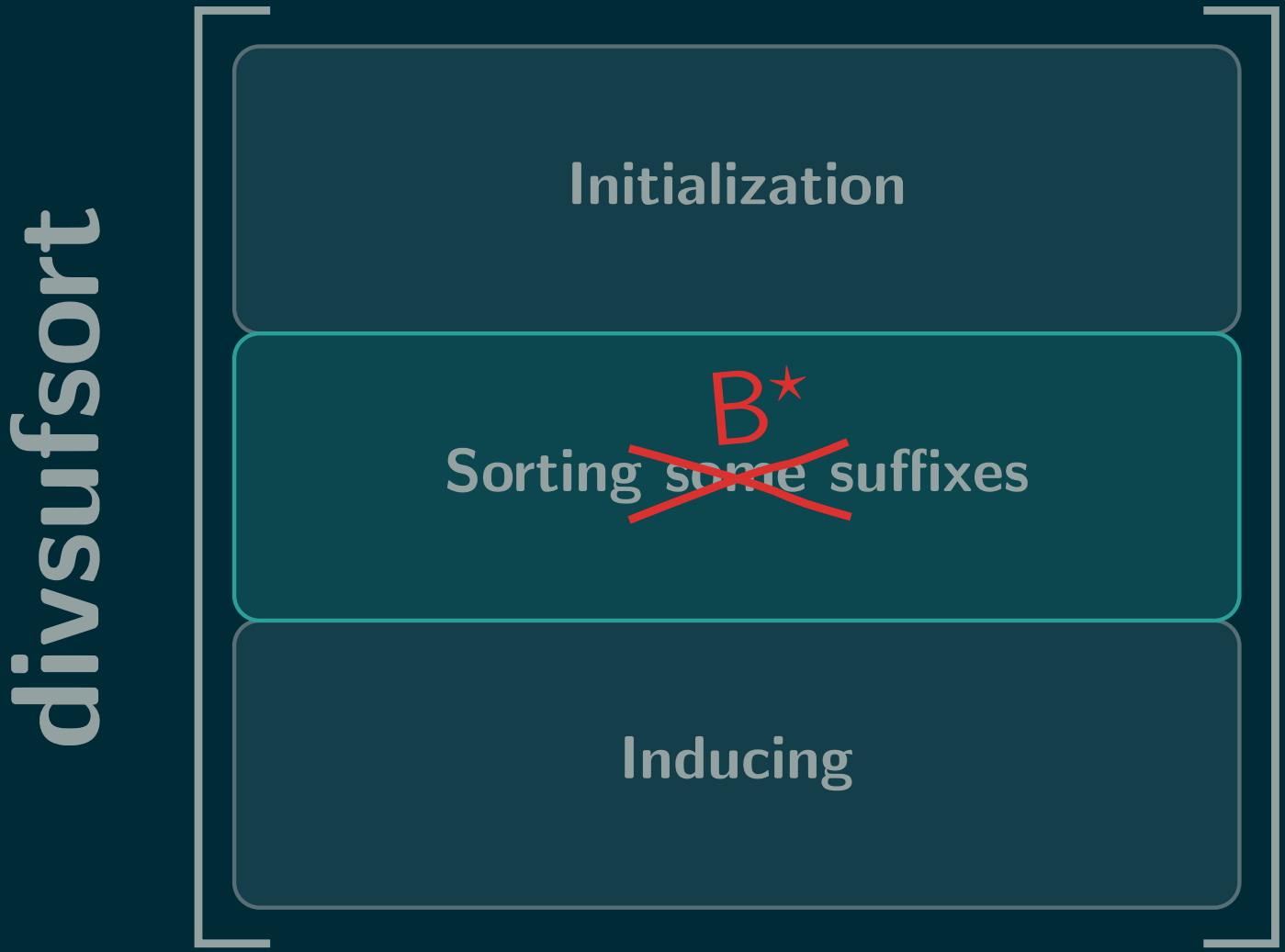


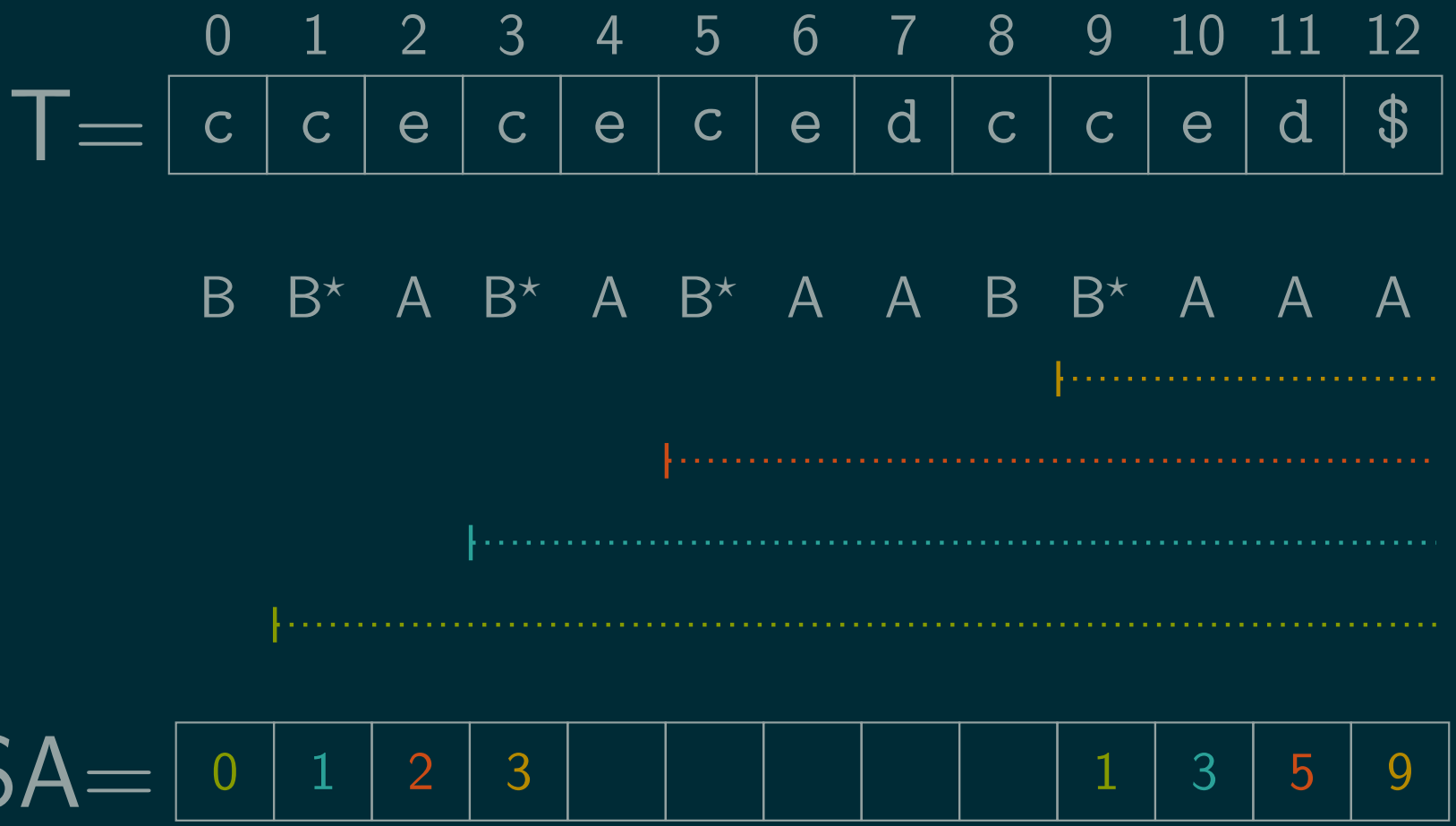
$$\#\mathrm{B}^\star \leq n/2$$

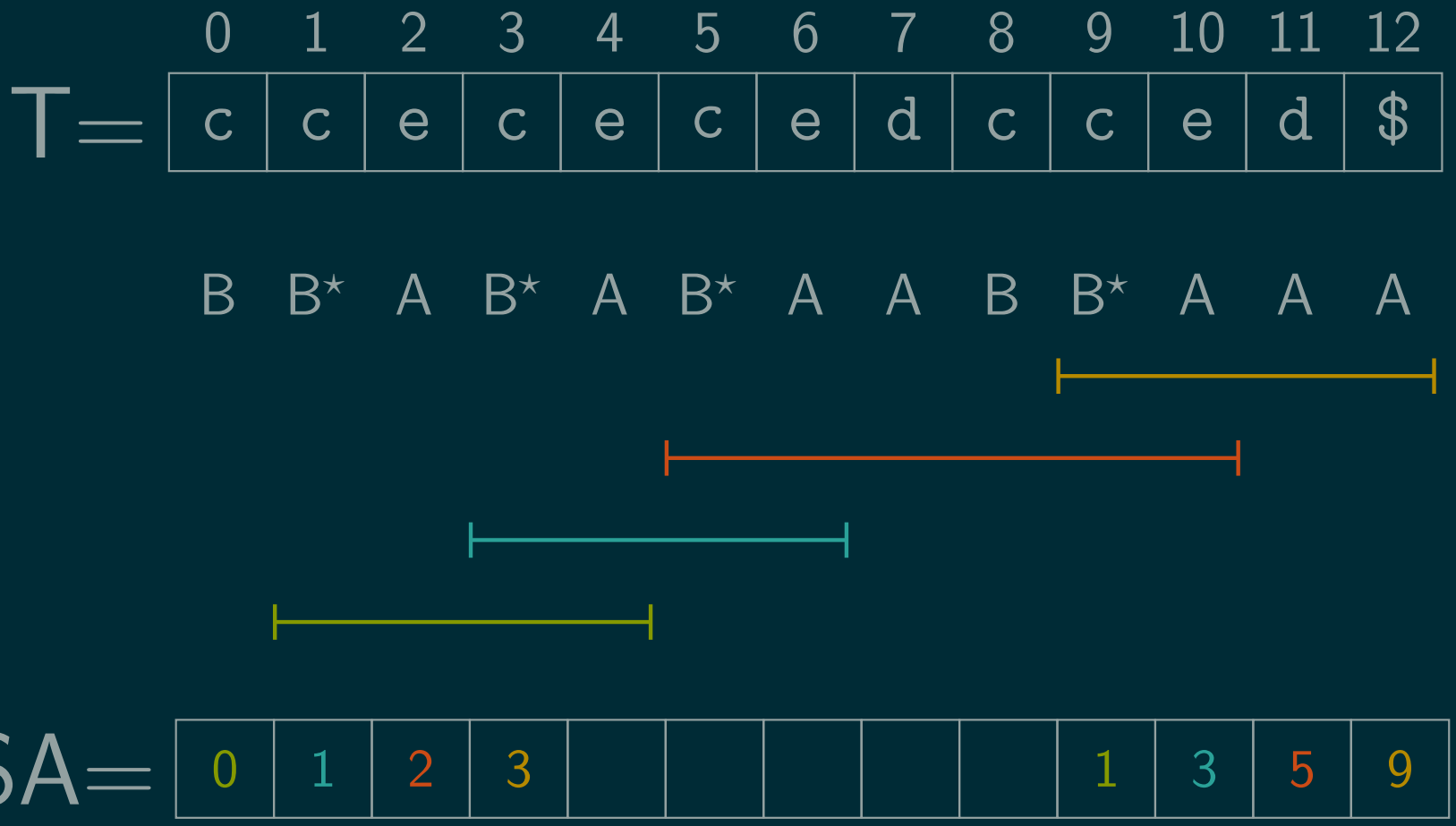


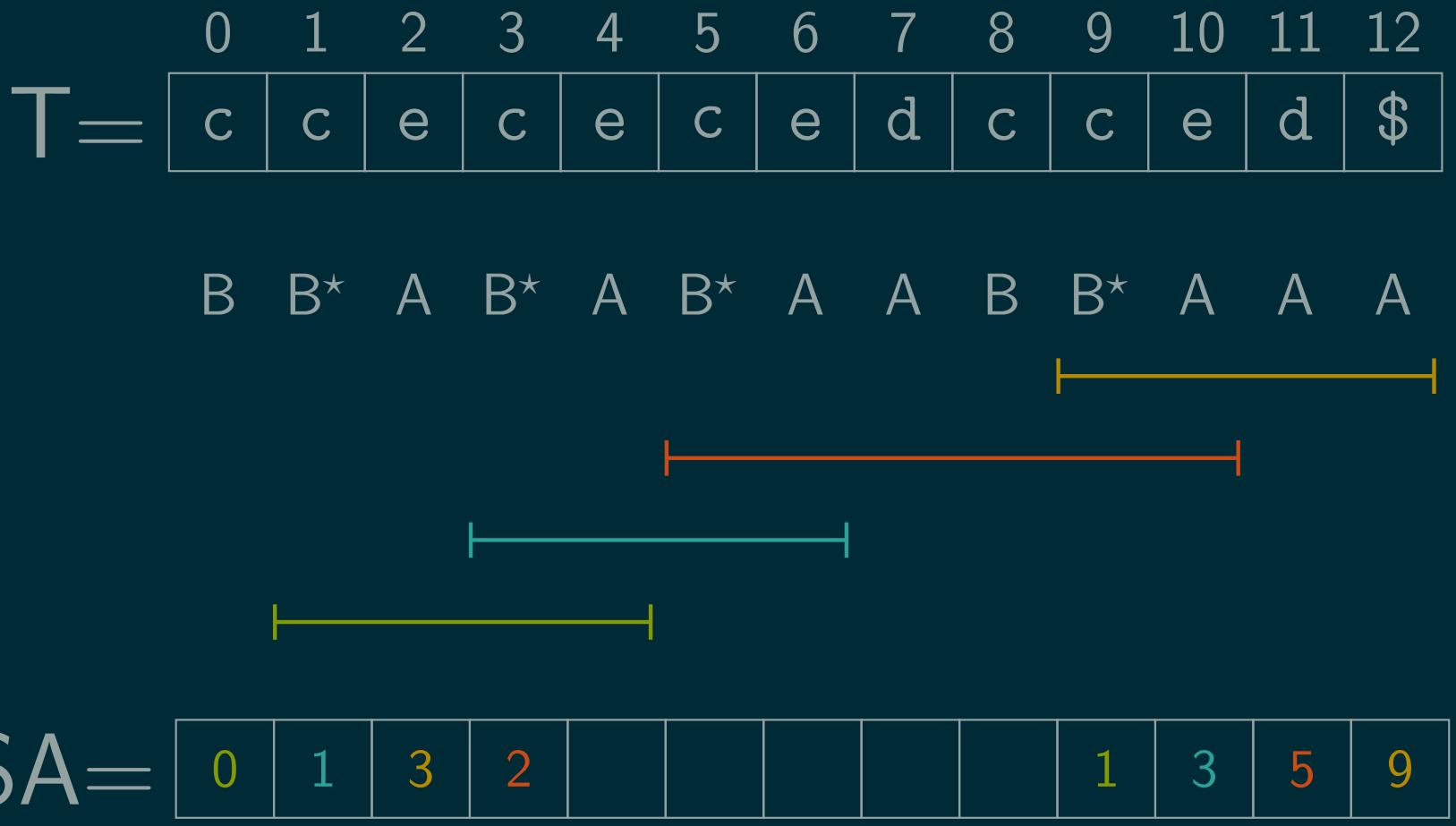
$$\#\mathrm{B}^\star \leq n/2$$

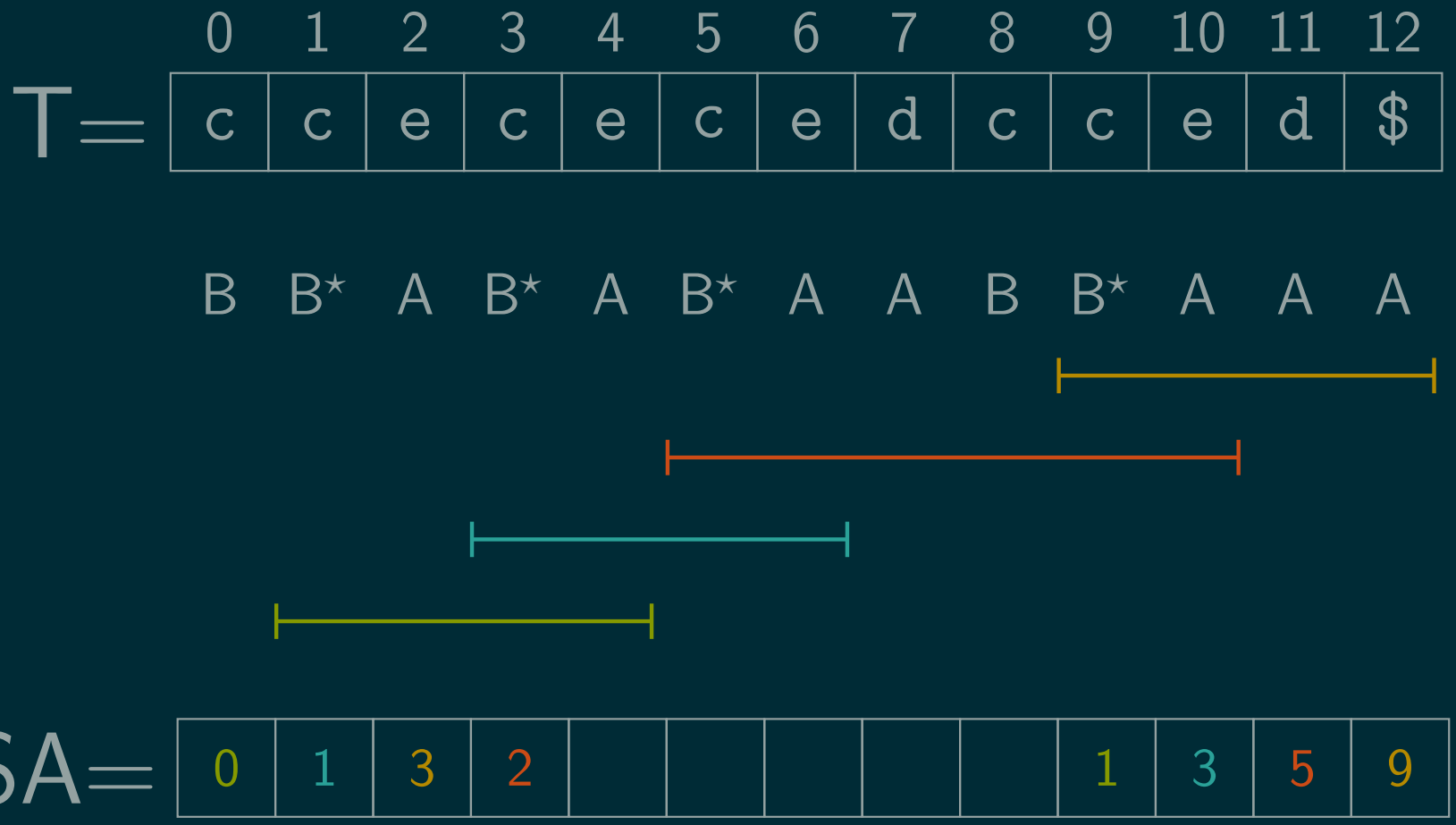

divsufsort
Initialization
Sorting some suffixes
Inducing


divsufsort
Initialization
Sorting B*
suffixes
Inducing


0 1 2 3 4 5 6 7 8 9 10 11 12
T=
c c e c e c e d c c e d $
B B* A B* A B* A A B B* A A A
SA=
0 1 2 3 1 3 5 9

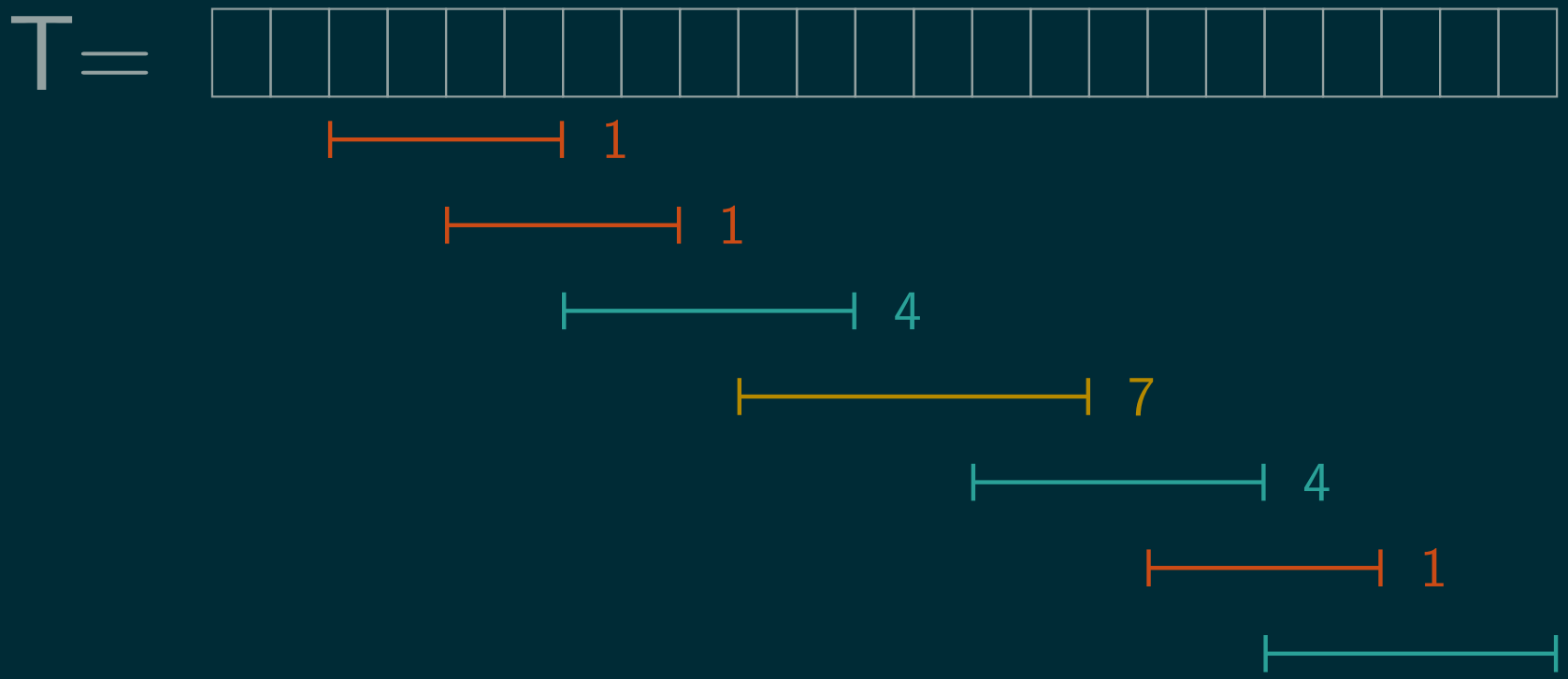

0 1 2 3 4 5 6 7 8 9 10 11 12
T=
c c e c e c e d c c e d $
B B* A B* A B* A A B B* A A A
SA=
0 1 2 3 1 3 5 9


0 1 2 3 4 5 6 7 8 9 10 11 12
T=
c c e c e c e d c c e d $
B B* A B* A B* A A B B* A A A
SA=
0 1 3 2 1 3 5 9

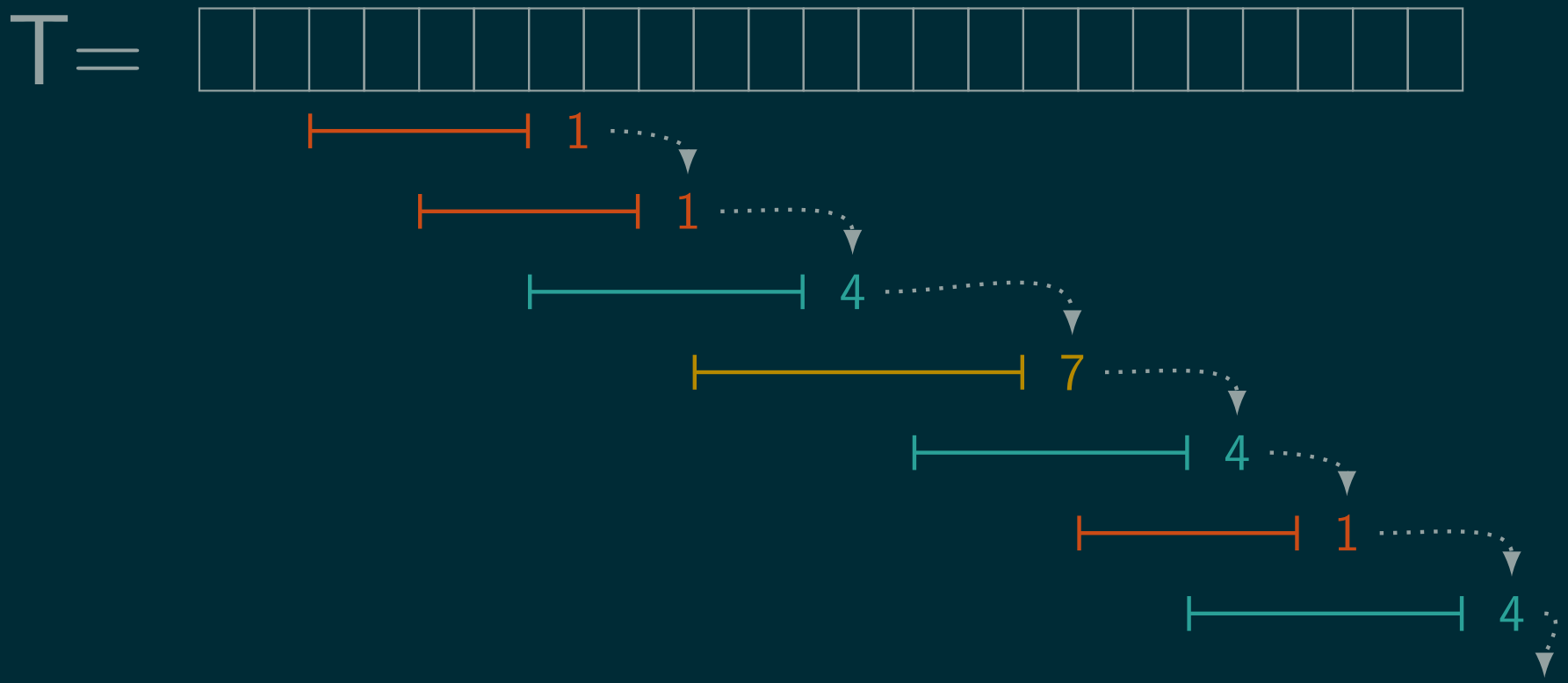


$\longrightarrow$ Introsort, Heapsort and Insertionsort

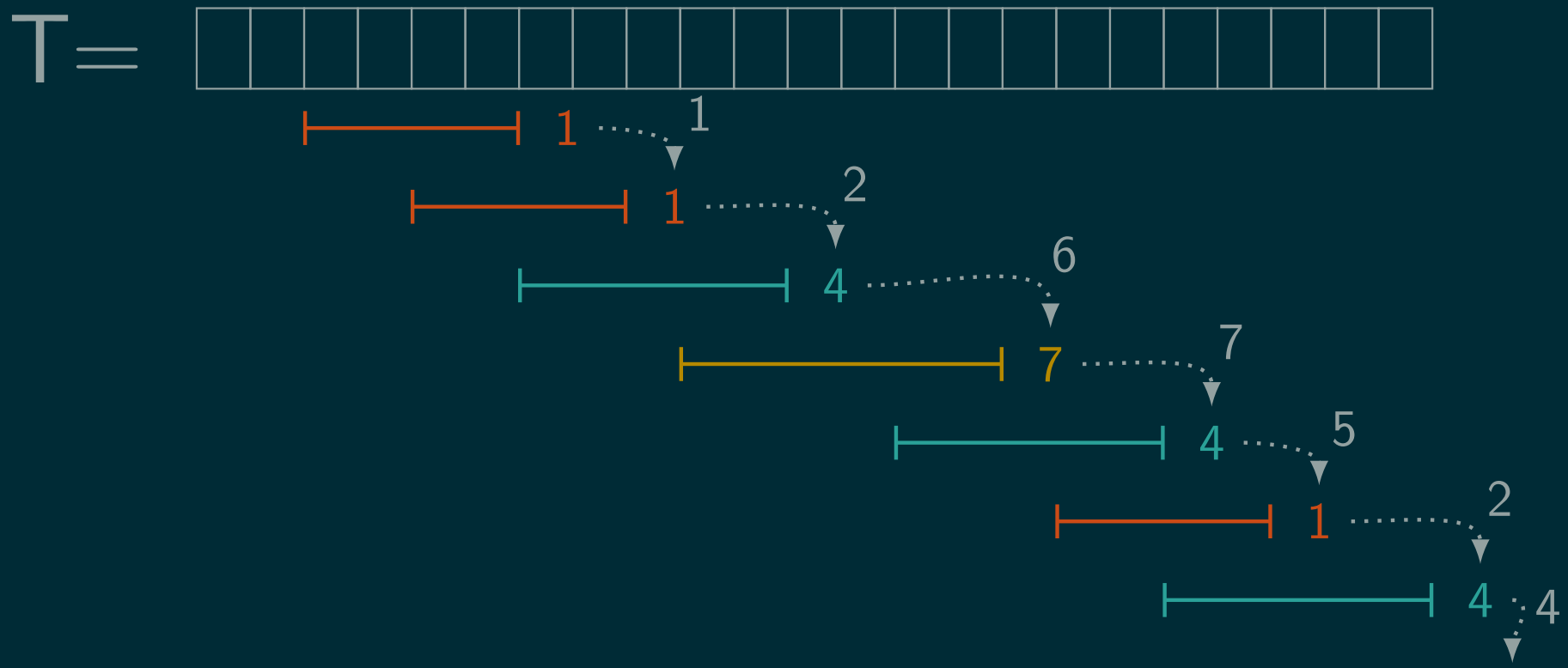
Substring Doubling



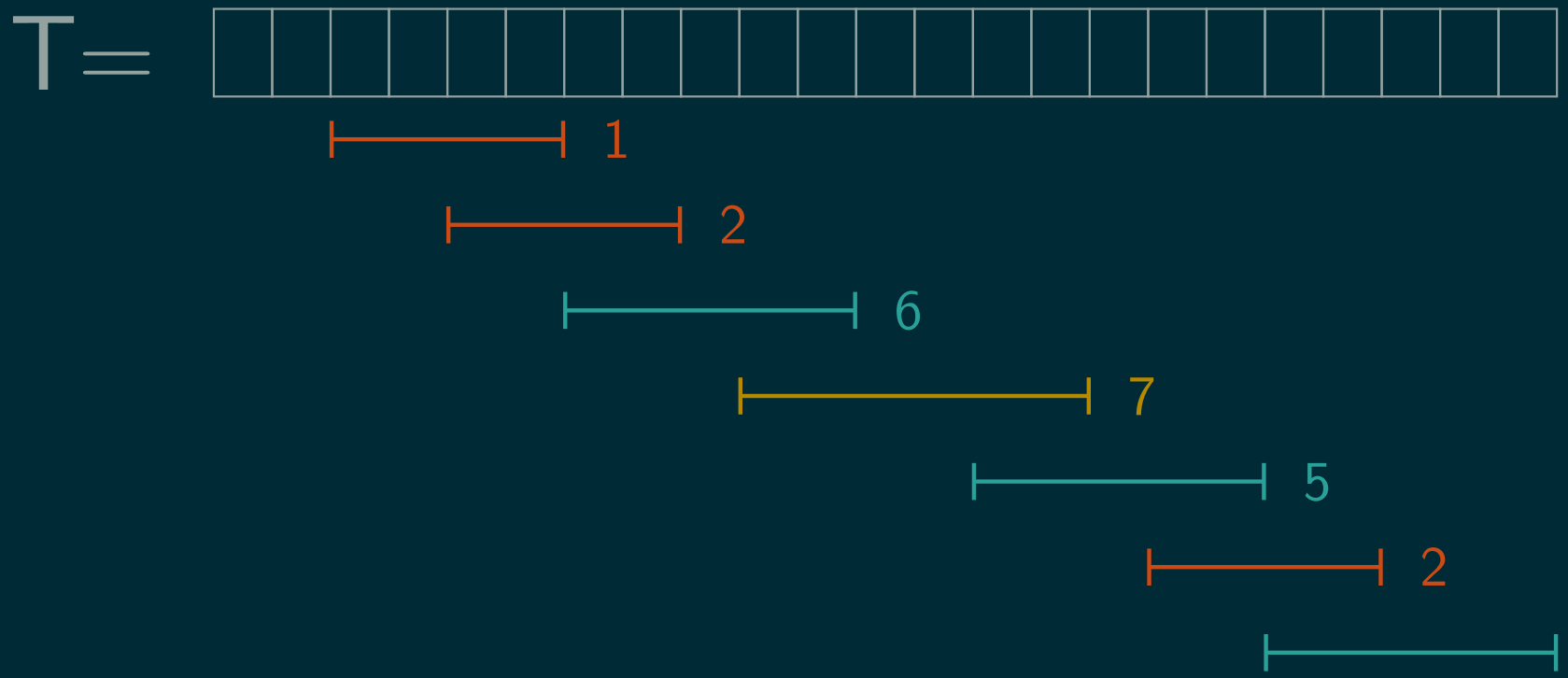
Substring Doubling



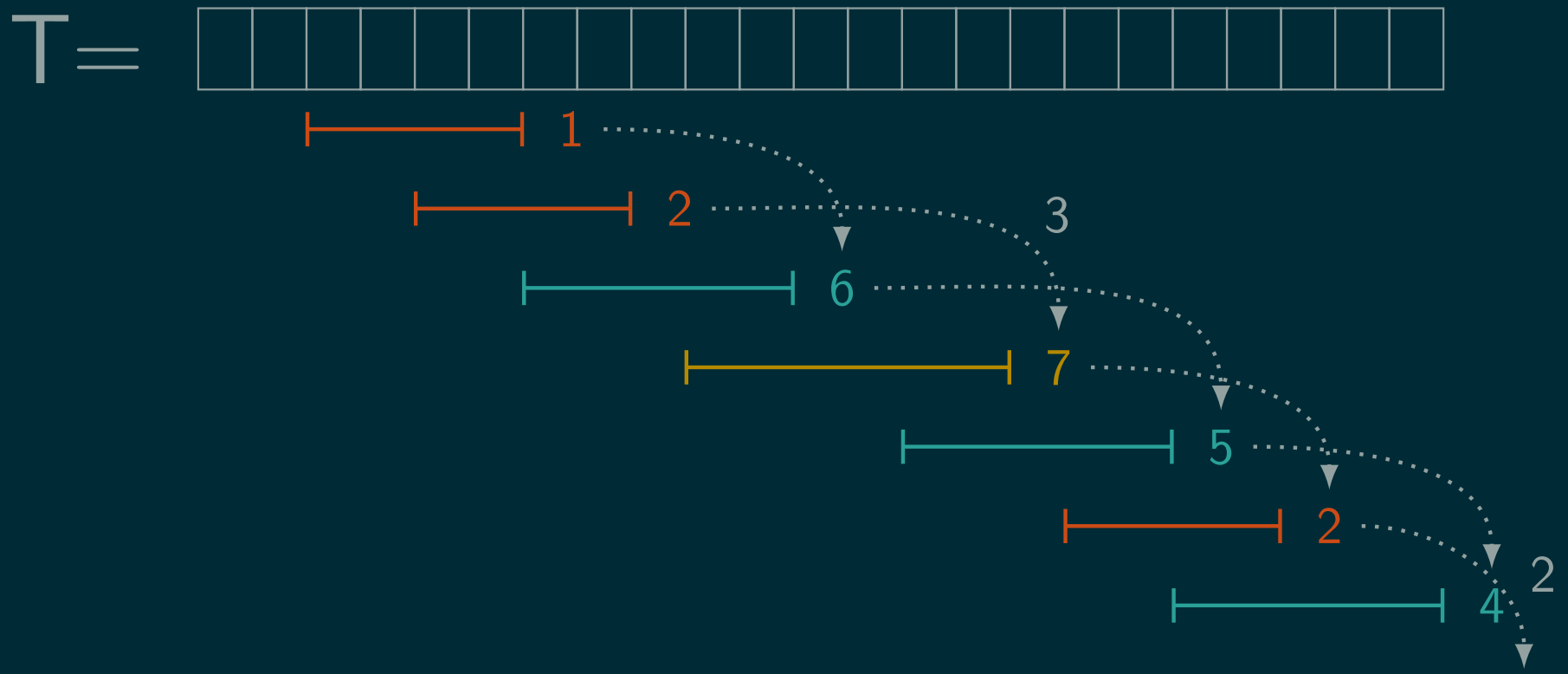
Substring Doubling



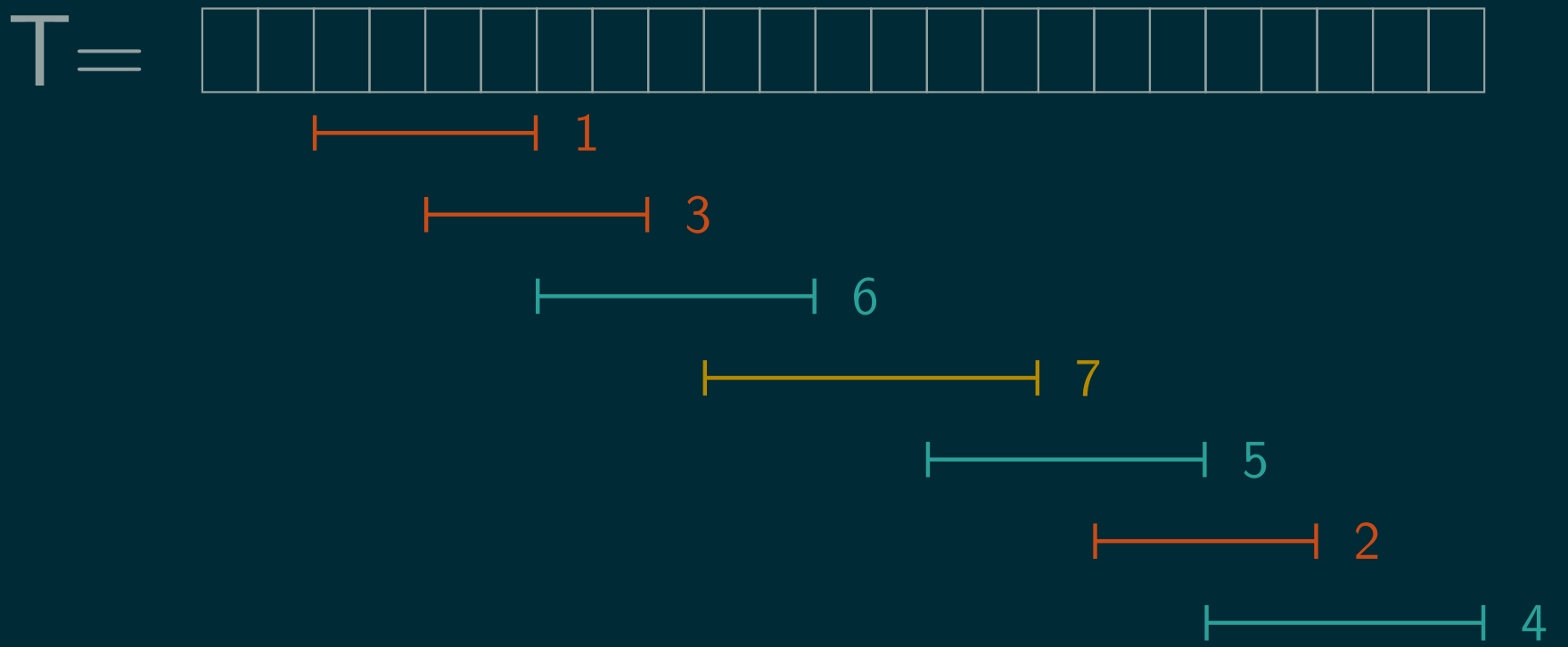
Substring Doubling



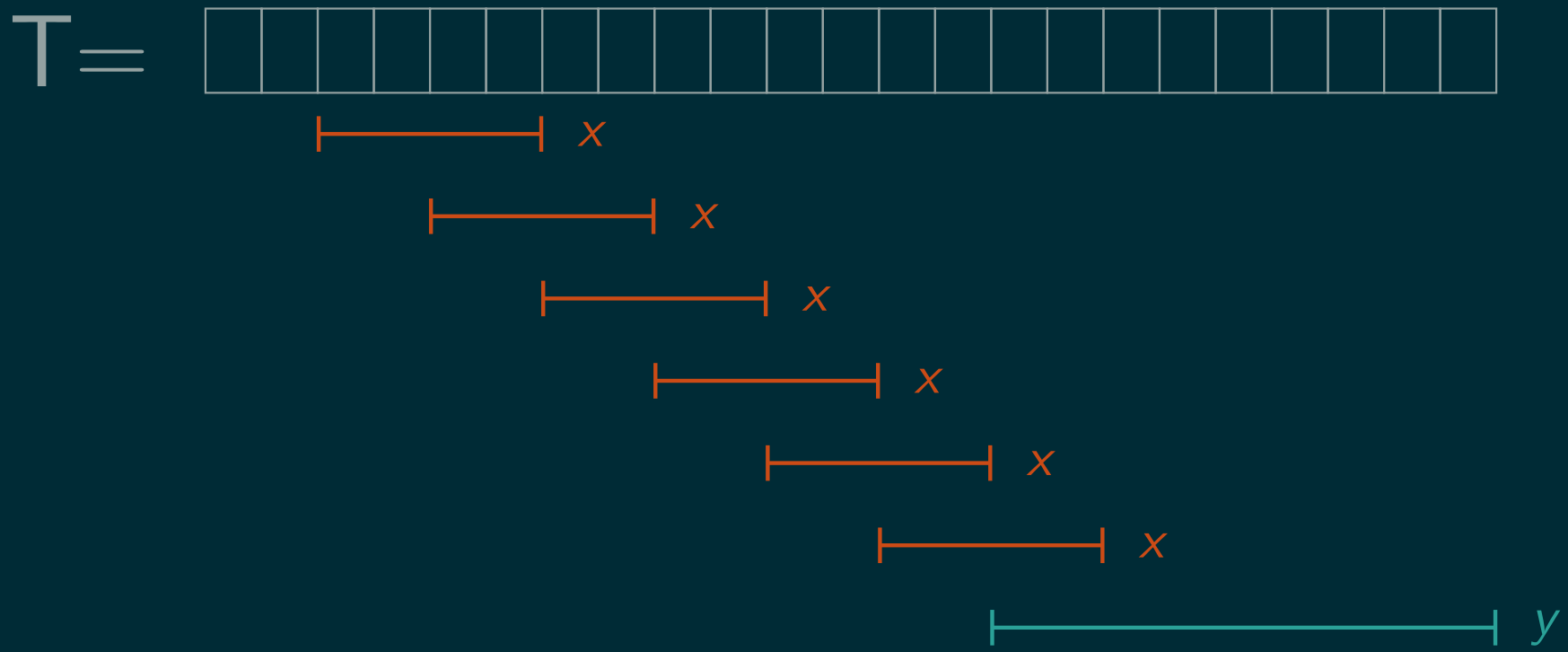
Substring Doubling



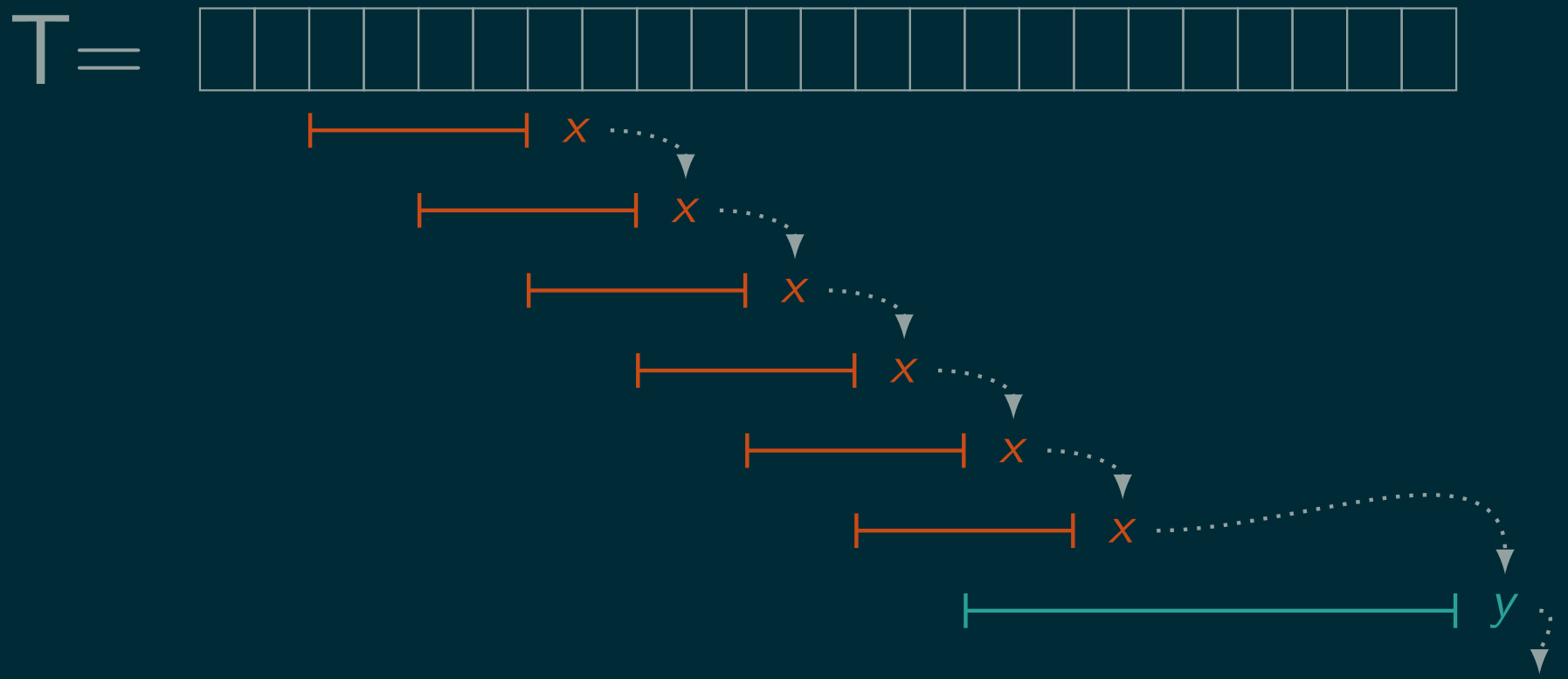
Substring Doubling



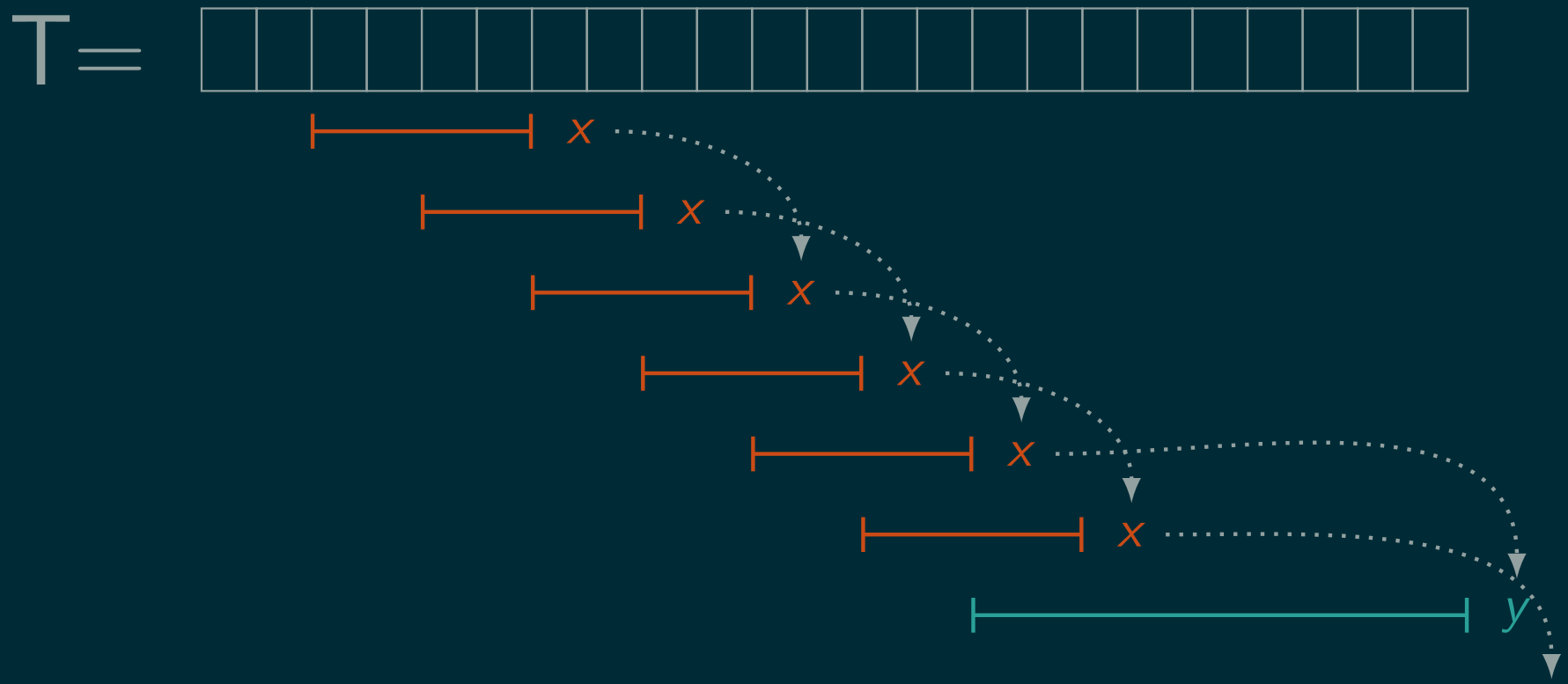
Repetition Detection [MP2007]



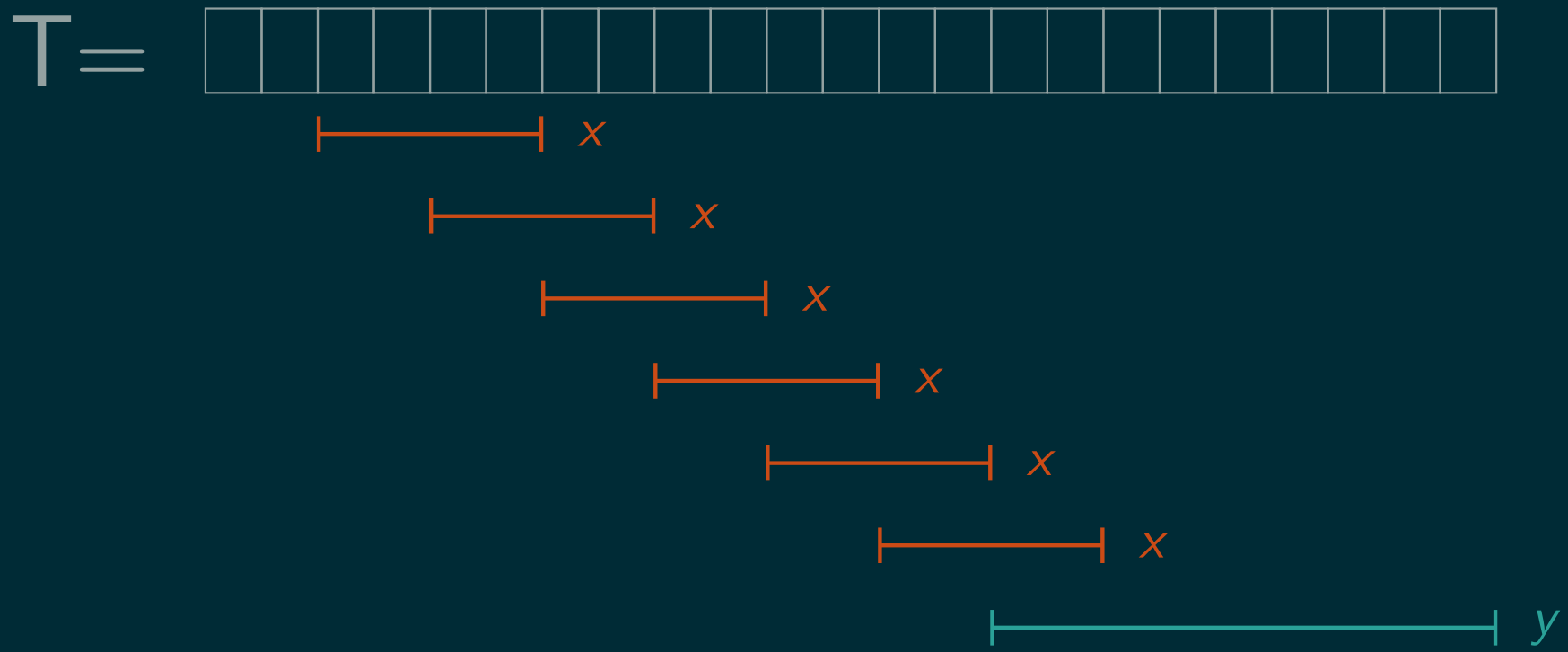
Repetition Detection [MP2007]

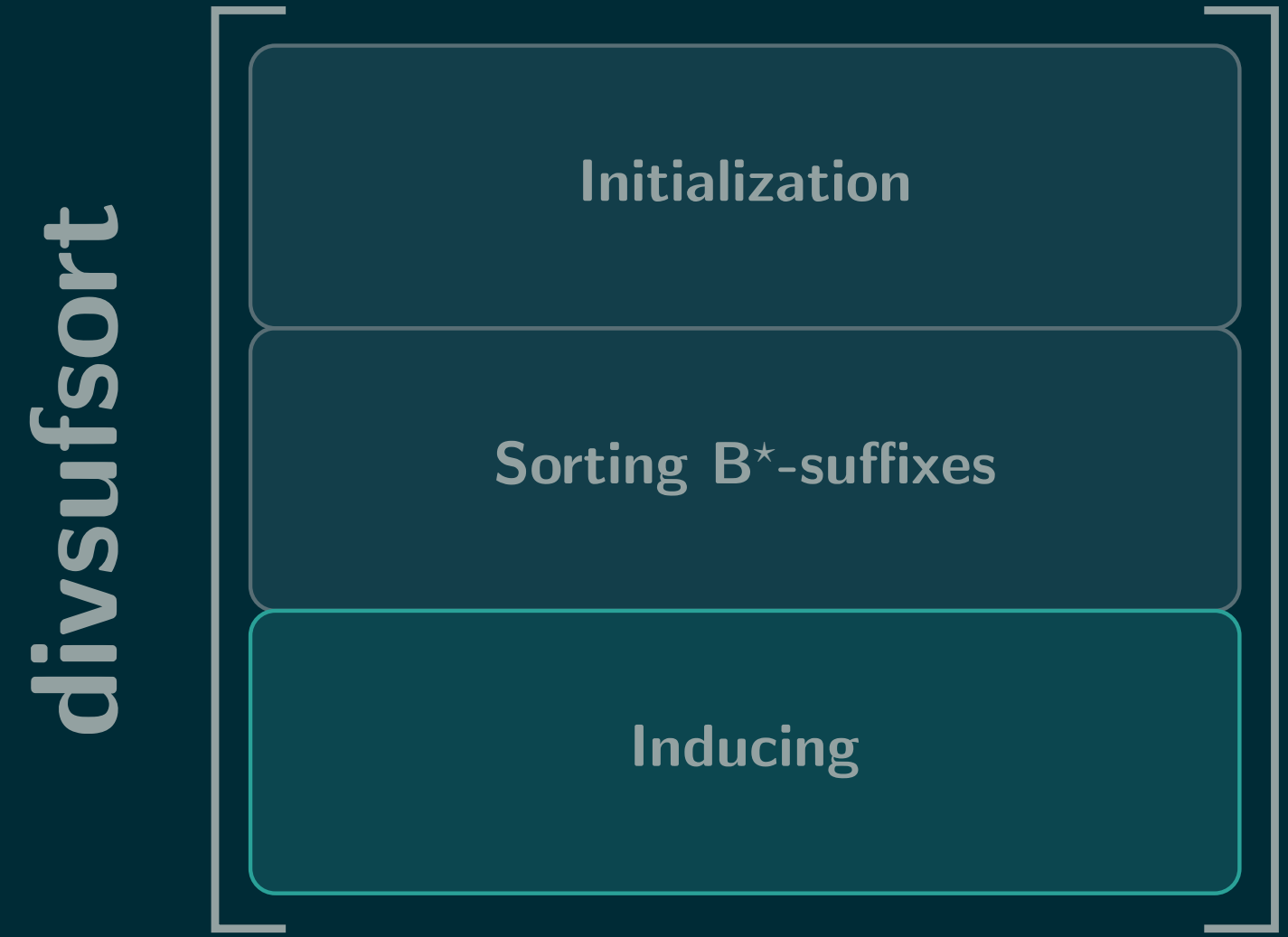


Repetition Detection [MP2007]

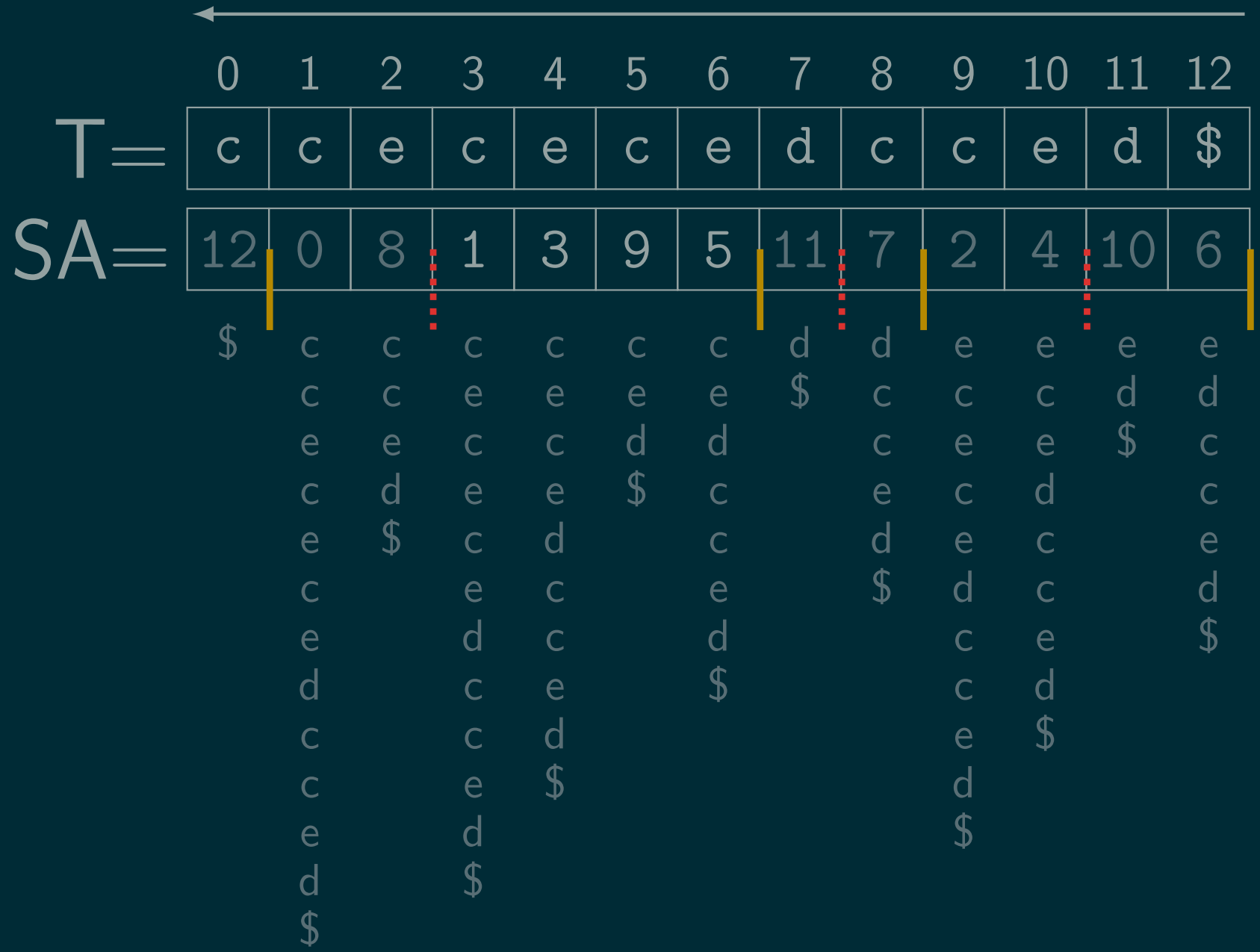


Repetition Detection [MP2007]

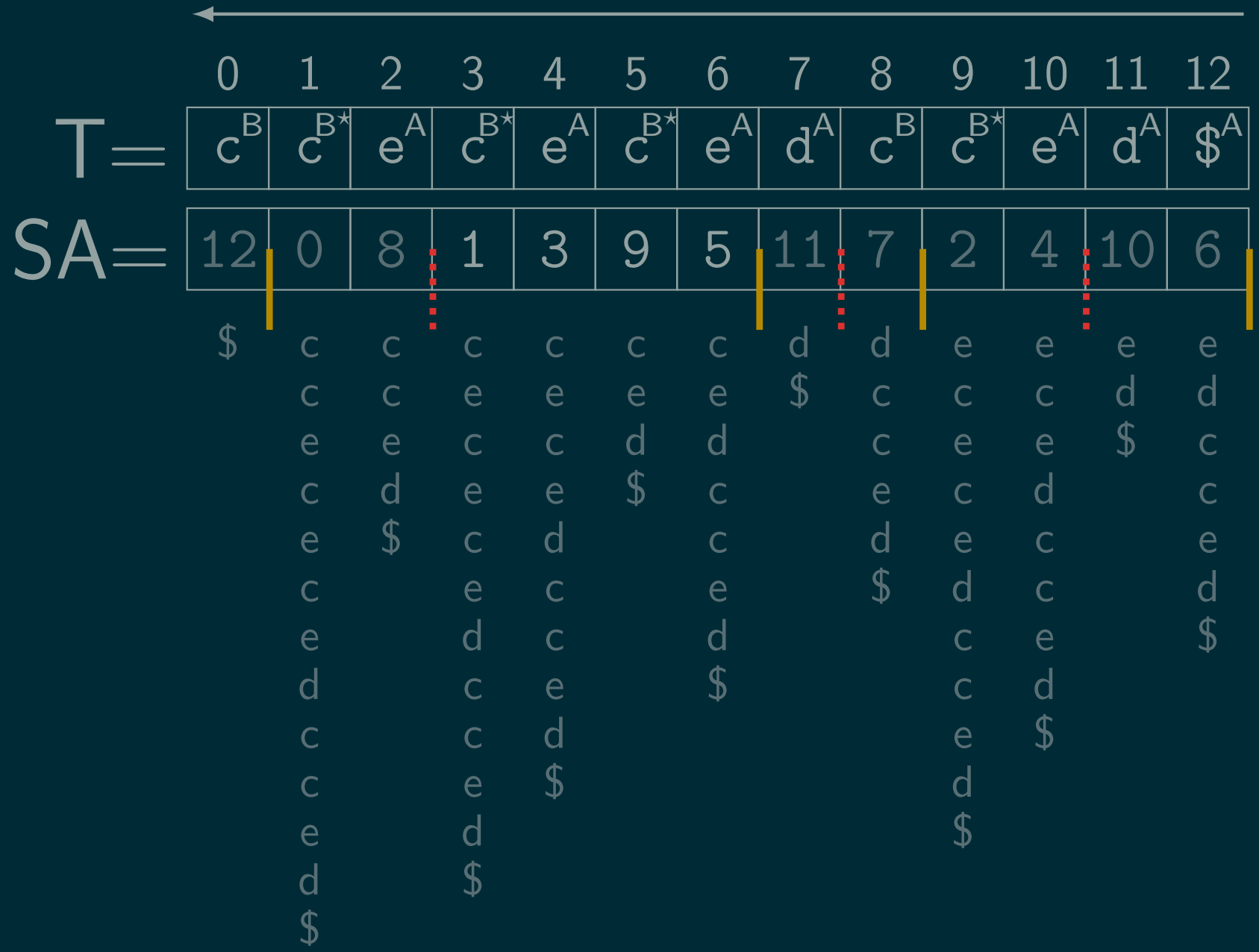



divsufsort
Initialization
Sorting B*-suffixes
Inducing

Inducing B


0 1 2 3 4 5 6 7 8 9 10 11 12
T= c c e c e c e d c c e d $
SA= 12 0 8 1 3 9 5 11 7 2 4 10 6
$
cccedcced$
ccced$
ceccedcced$
cceccedcced$
ceed$
cedccced$
d$
dccced$
eccedcced$
ecedcced$
ed$
edccced$

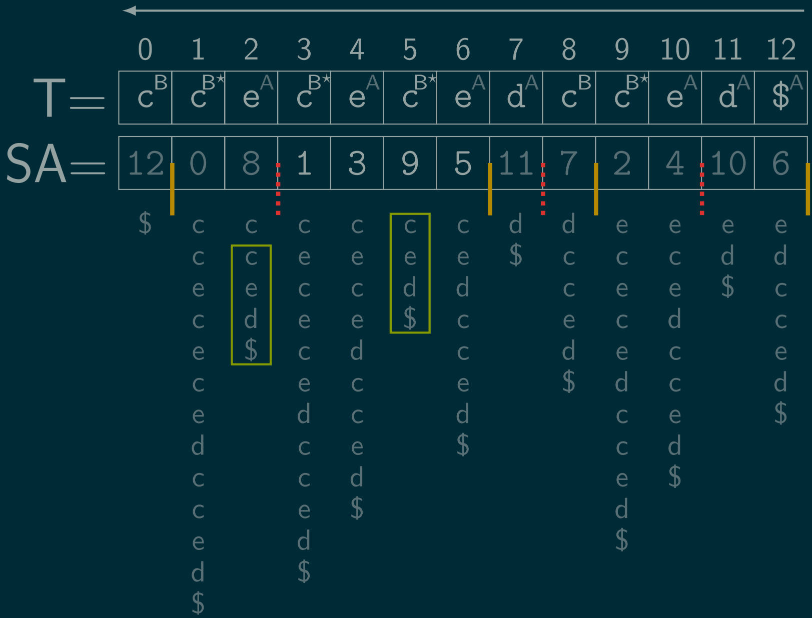
Inducing B


0 1 2 3 4 5 6 7 8 9 10 11 12
T=
cB cB* eA cB* eA cB* eA dA cB cB* eA dA $A
SA=
12 0 8 1 3 9 5 11 7 2 4 10 6
$
ccecedcced$
ced$
cecedcced$
ceedcced$
cedcced$
d$
dcced$
ecedcced$
ecedcced$
ed$
edcced$

Inducing B

	0	1	2	3	4	5	6	7	8	9	10	11	12
T=	c^{B}	$c^{B\star}$	e^{A}	$c^{B\star}$	e^{A}	$c^{B\star}$	e^{A}	d^{A}	c^{B}	$c^{B\star}$	e^{A}	d^{A}	$\$^{A}$
SA=	12	0	8	1	3	9	5	11	7	2	4	10	6
	$	ccececedcced$	cced$	cececedcced$	cecedcced$	ced$	cedcced$	d$	dcced$	ececedcced$	ecedcced$	ed$	edcced$

Inducing B


0 1 2 3 4 5 6 7 8 9 10 11 12
T= c^B c^B* e^A c^B* e^A c^B* e^A d^A c^B c^B* e^A d^A $^A
SA= 12 0 8 1 3 9 5 11 7 2 4 10 6
$
ccececedcced$
cced$
cececedcced$
cecedcced$
ced$
cedcced$
d$
dcced$
ececedcced$
ecedcced$
ed$
edcced$

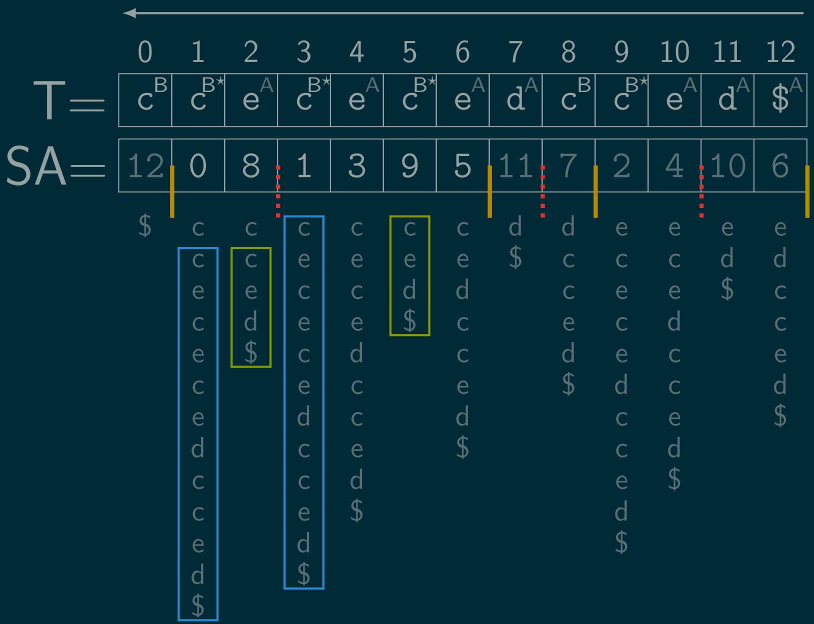
Inducing B

	0	1	2	3	4	5	6	7	8	9	10	11	12
T=	c^{B}	$c^{B\star}$	e^{A}	$c^{B\star}$	e^{A}	$c^{B\star}$	e^{A}	d^{A}	c^{B}	$c^{B\star}$	e^{A}	d^{A}	$\$^{A}$
SA=	12	0	8	1	3	9	5	11	7	2	4	10	6
	\$	ccececedcced\$	cced\$	cececedcced\$	cecedcced\$	ced\$	cedcced\$	d\$	dcced\$	ececedcced\$	ecedcced\$	ed\$	edcced\$

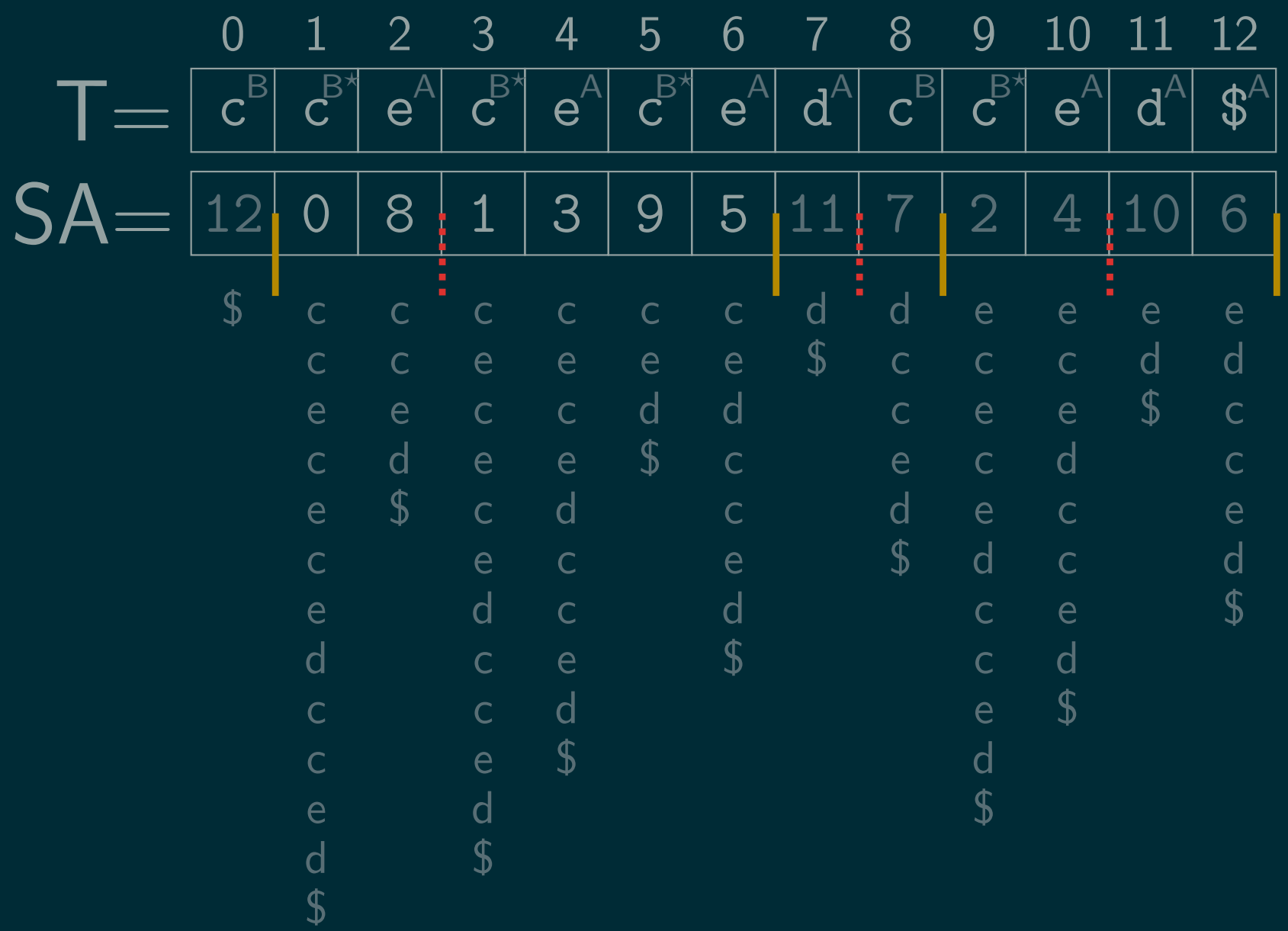
Inducing B

	0	1	2	3	4	5	6	7	8	9	10	11	12
T=	c^B	$c^{B\star}$	e^A	$c^{B\star}$	e^A	$c^{B\star}$	e^A	d^A	c^B	$c^{B\star}$	e^A	d^A	$\$^A$
SA=	12	0	8	1	3	9	5	11	7	2	4	10	6
	$	ccececedcced$	cced$	cececedcced$	cecedcced$	ced$	cedcced$	d$	dcced$	ececedcced$	ecedcced$	ed$	edcced$

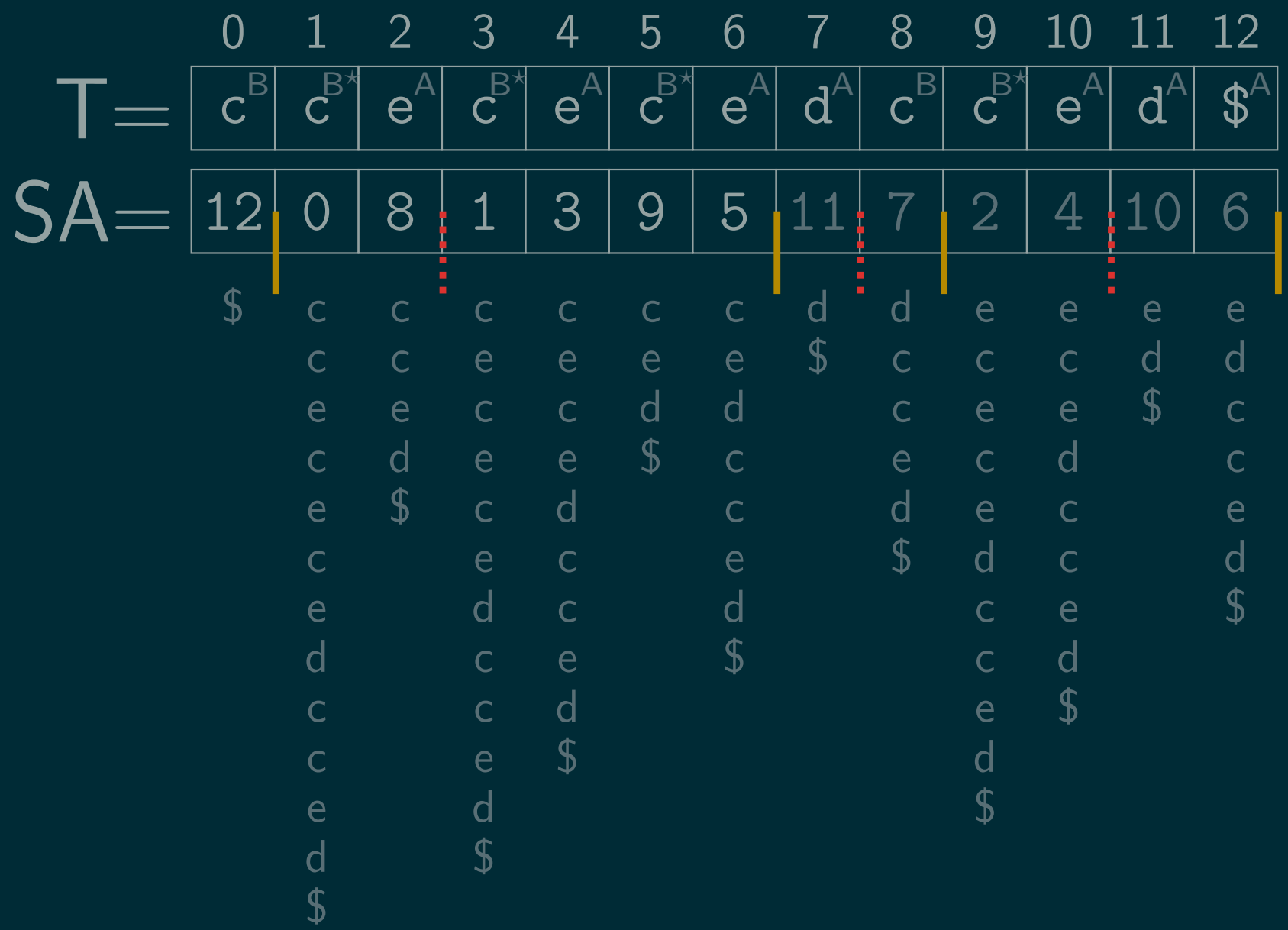
Inducing B


0 1 2 3 4 5 6 7 8 9 10 11 12
T=
c^B c^B* e^A c^B* e^A c^B* e^A d^A c^B c^B* e^A d^A $^A
SA=
12 0 8 1 3 9 5 11 7 2 4 10 6
$
ccecececedccced$
cced$
cecececedccced$
cecedccced$
ced$
cedccced$
d$
dcced$
ececedccced$
ecedccced$
ed$
edcced$

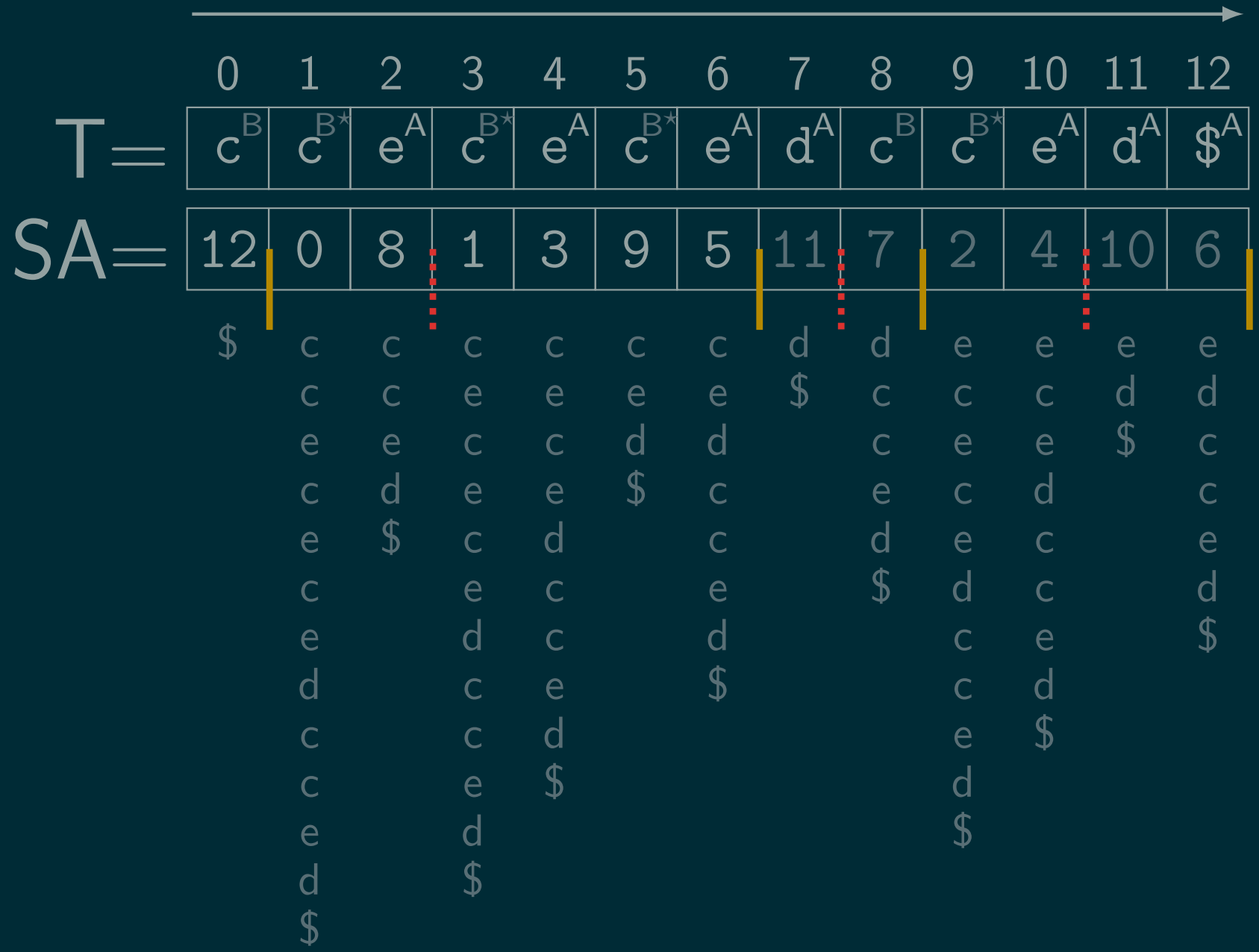
In between


0 1 2 3 4 5 6 7 8 9 10 11 12
T=
cB cB* eA cB* eA cB* eA dA cB cB* eA dA $A
SA=
12 0 8 1 3 9 5 11 7 2 4 10 6
$
ccecedcceed$
cceed$
ceceecedcced$
ceecedcced$
ceed$
cedcced$
d$
dcced$
ecedcced$
ecedcced$
ed$
edcced$

In between


0 1 2 3 4 5 6 7 8 9 10 11 12
T=
c^B c^B* e^A c^B* e^A c^B* e^A d^A c^B c^B* e^A d^A $^A
SA=
12 0 8 1 3 9 5 11 7 2 4 10 6
$
cccedcceed$
cced$
ceccedcced$
ceedccedd$
cedcced$
d$
dccced$
eccedcced$
ecdcced$
ed$
edcced$

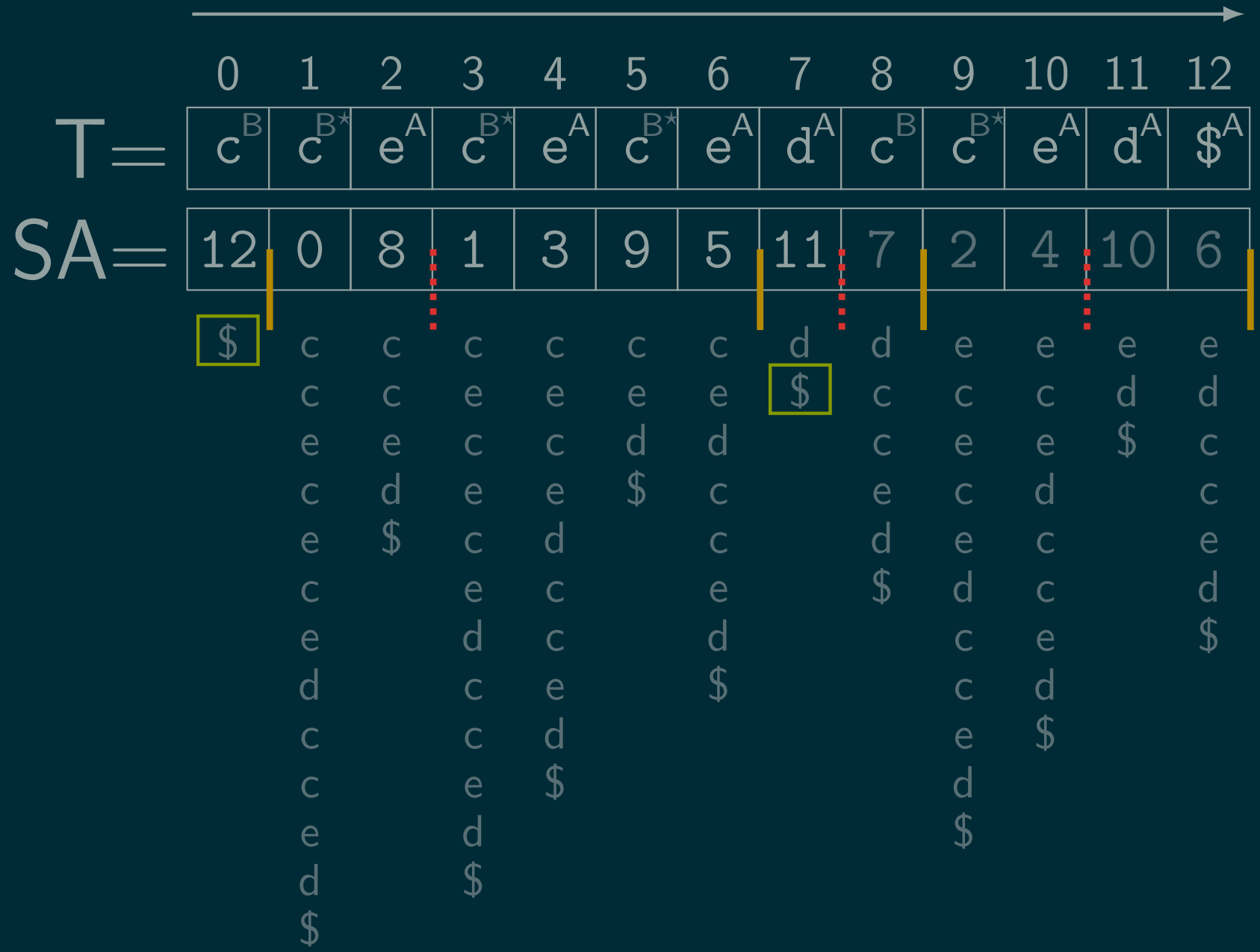
Inducing A


0 1 2 3 4 5 6 7 8 9 10 11 12
T=
c^B c^B* e^A c^B* e^A c^B* e^A d^A c^B c^B* e^A d^A $^A
SA=
12 0 8 1 3 9 5 11 7 2 4 10 6
$
cccecedcced$
ccecd$
ceccedcced$
ceced$
cedcced$
d$
dcced$
ecedcced$
ecedcced$
ed$
edcced$

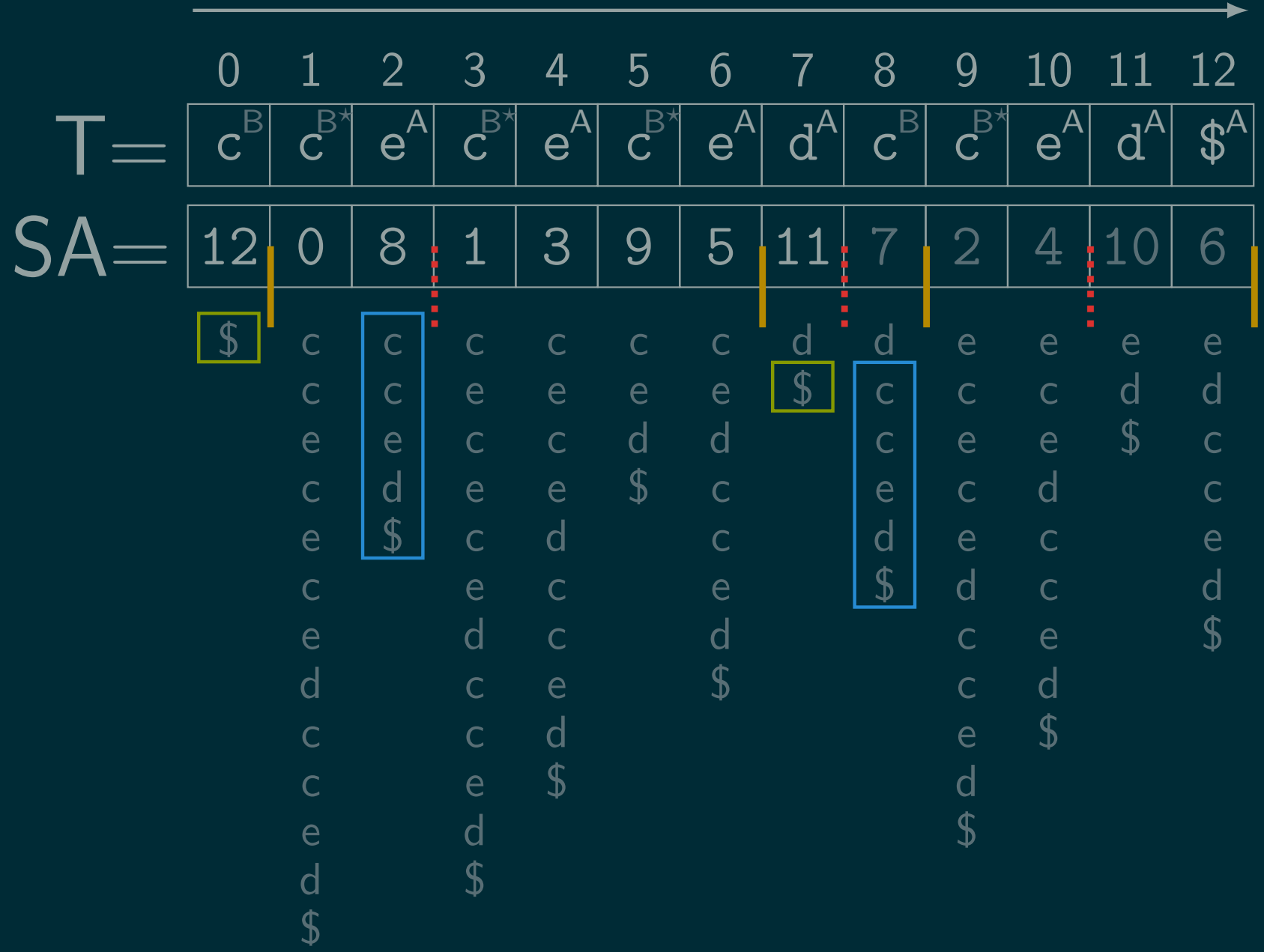
Inducing A

	0	1	2	3	4	5	6	7	8	9	10	11	12
T=	c^B	$c^{B\star}$	e^A	$c^{B\star}$	e^A	$c^{B\star}$	e^A	d^A	c^B	$c^{B\star}$	e^A	d^A	$\$^A$
SA=	12	0	8	1	3	9	5	11	7	2	4	10	6
	$	ccecececedcced$	cced$	cecececedcced$	cecedcced$	ced$	cedcced$	d$	dcced$	ececedcced$	ecedcced$	ed$	edcced$

Inducing A


0 1 2 3 4 5 6 7 8 9 10 11 12
T=
c^B c^B* e^A c^B* e^A c^B* e^A d^A c^B c^B* e^A d^A $^A
SA=
12 0 8 1 3 9 5 11 7 2 4 10 6
$
c c e c e d c c e d $
c c e d $
c e c e d c c e d $
c e c e d c e d $
c e d $
c e d c c e d $
d $
d c e d c e d $
e c e d c e d $
e d $
e d c c e d $

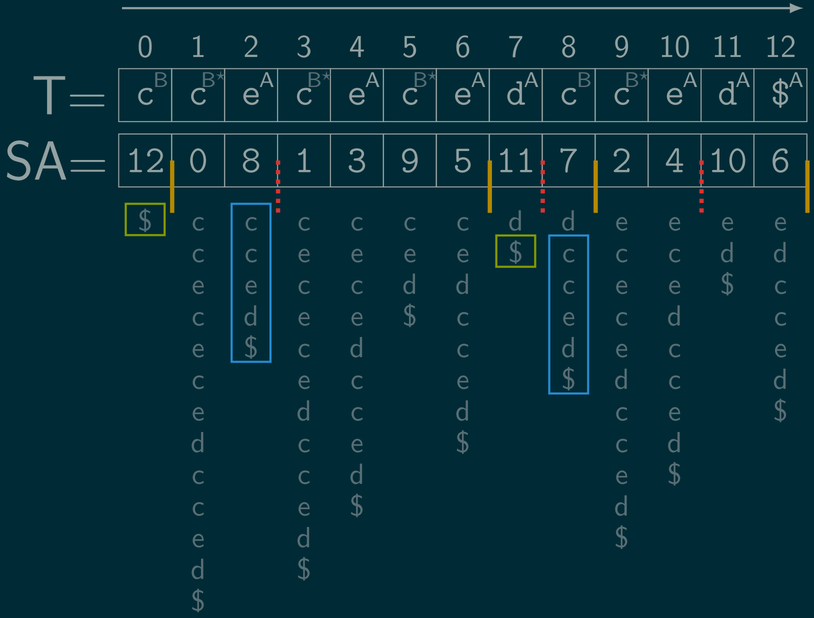
Inducing A

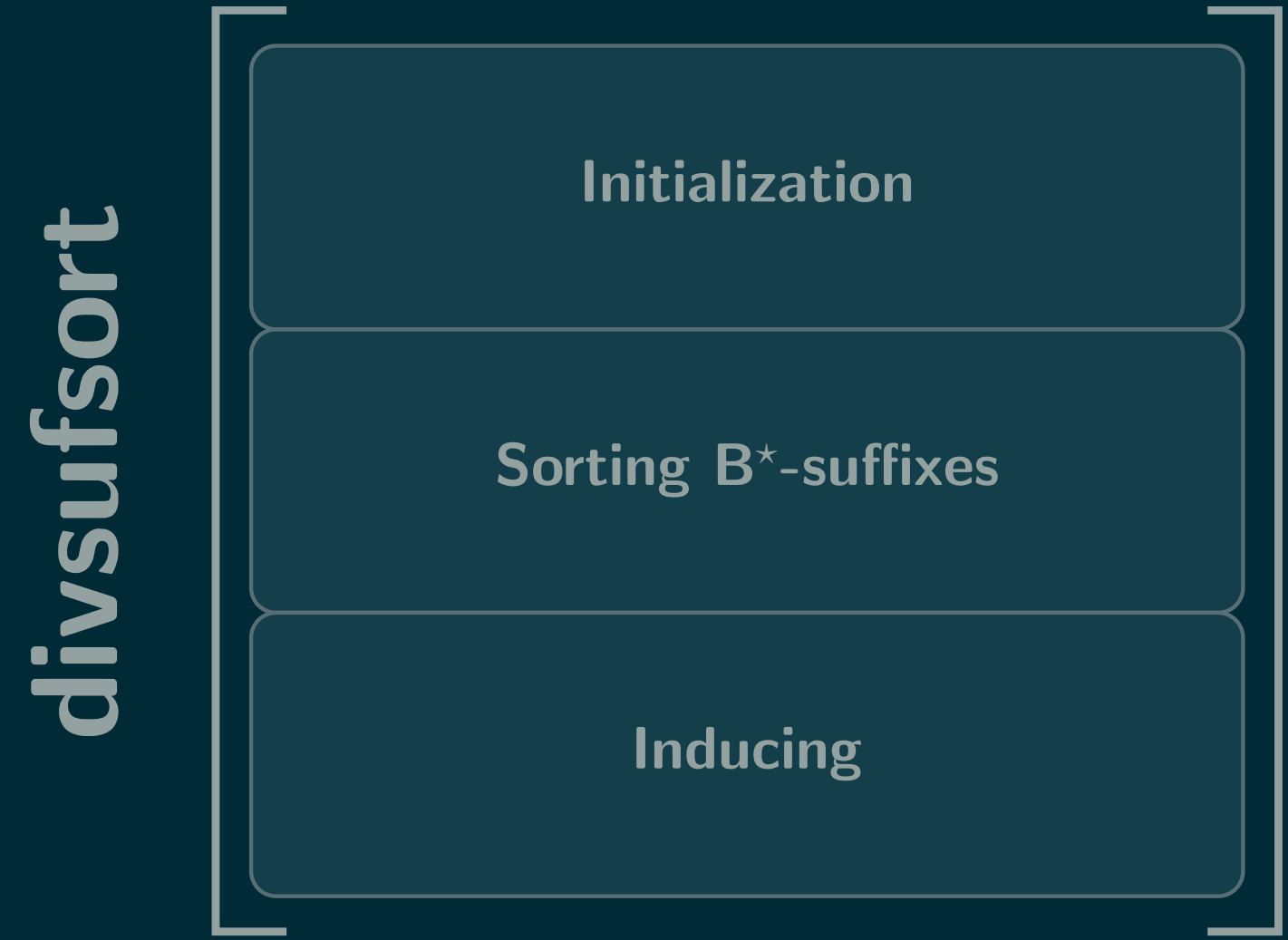

0 1 2 3 4 5 6 7 8 9 10 11 12
T=
c^B c^B* e^A c^B* e^A c^B* e^A d^A c^B c^B* e^A d^A $^A
SA=
12 0 8 1 3 9 5 11 7 2 4 10 6
$
c c e c e d $
c e d $
c e c e d $
c e d $
c e d $
d $
d c e d $
e c e d c e d $
e c d $
e d $
e d c e d $

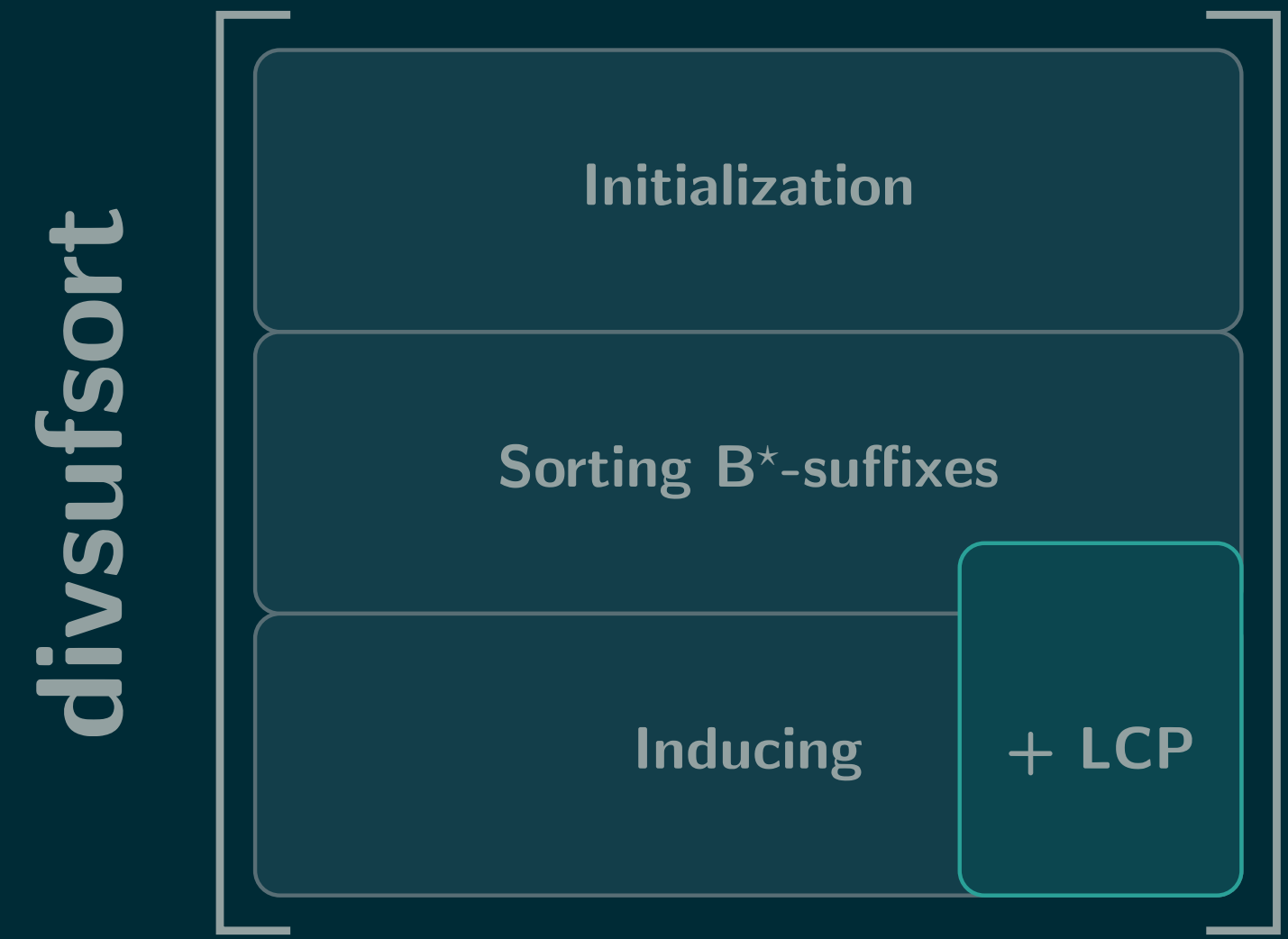
Inducing A

	0	1	2	3	4	5	6	7	8	9	10	11	12
T =	c^B	$c^{B\star}$	e^A	$c^{B\star}$	e^A	$c^{B\star}$	e^A	d^A	c^B	$c^{B\star}$	e^A	d^A	$\$^A$
SA =	12	0	8	1	3	9	5	11	7	2	4	10	6
	$	ccececedcced$	cced$	cececedcced$	cecedcced$	ced$	cedcced$	d$	dcced$	ececedcced$	ecedcced$	ed$	edcced$

Inducing A


0 1 2 3 4 5 6 7 8 9 10 11 12
T=
c B, c B⋆, e A, c B⋆, e A, c B⋆, e A, d A, c B, c B⋆, e A, d A, $ A
SA=
12 0 8 1 3 9 5 11 7 2 4 10 6

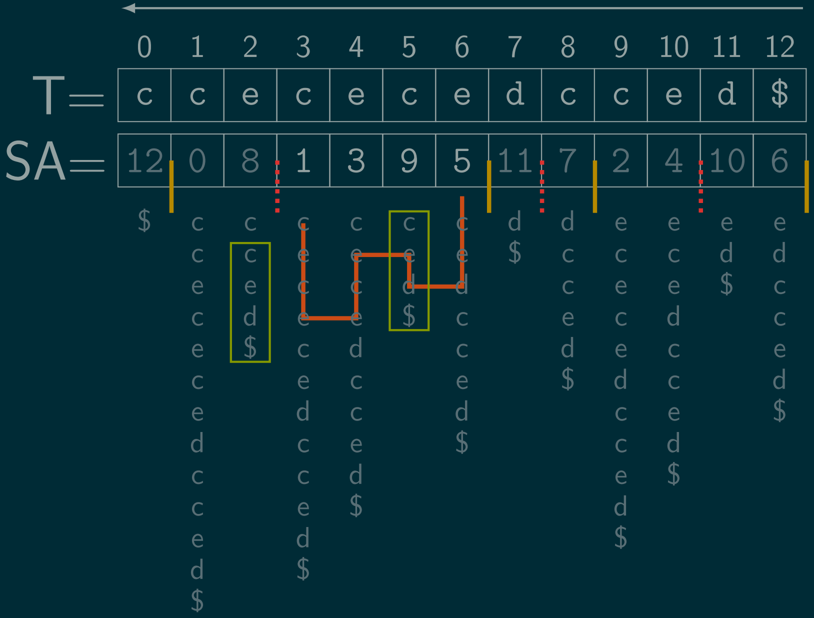

divsufsort
Initialization
Sorting B*-suffixes
Inducing


divsufsort
Initialization
Sorting B*-suffixes
Inducing
+ LCP

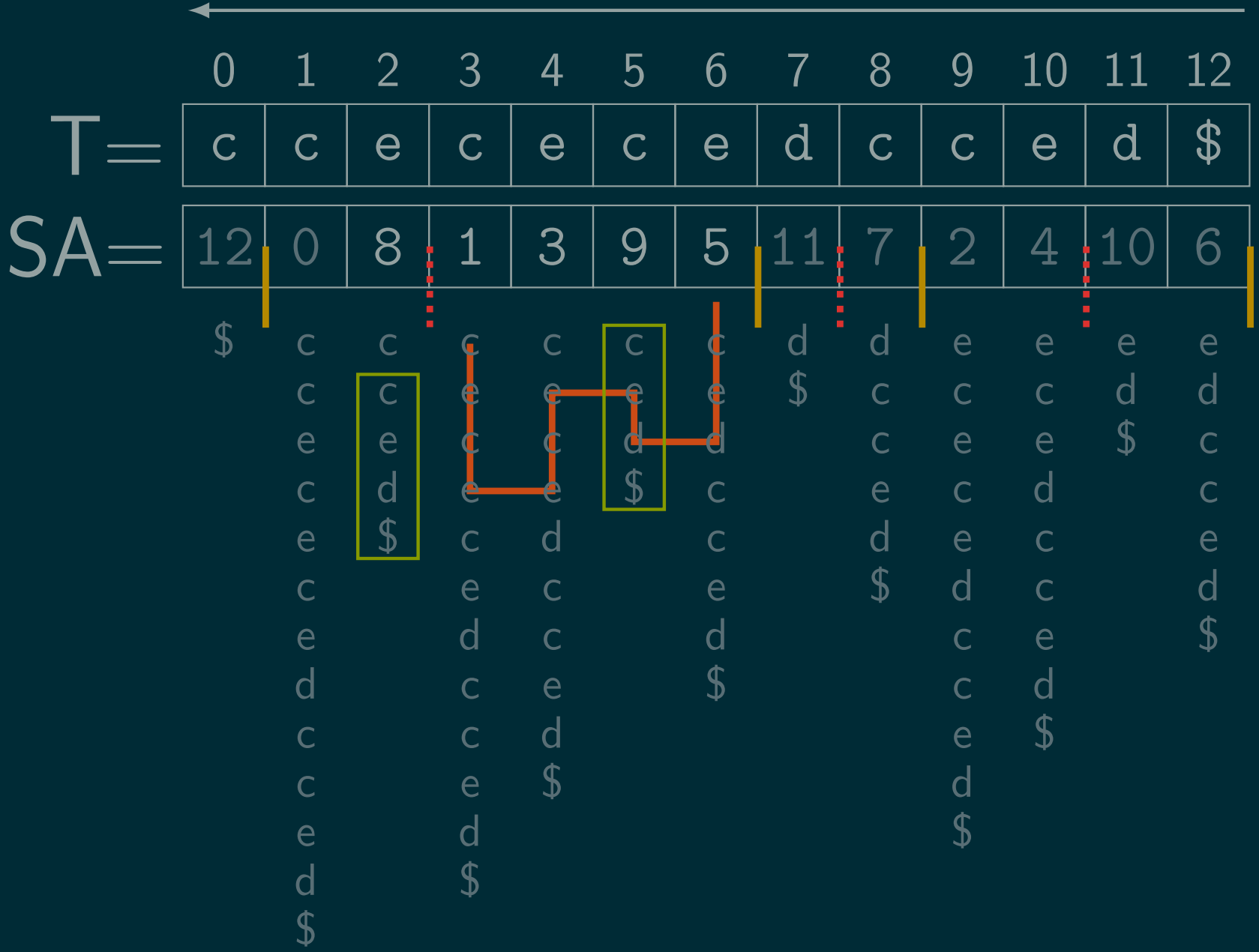
Inducing B


0 1 2 3 4 5 6 7 8 9 10 11 12
T= c c e c e c e d c c e d $
SA= 12 0 8 1 3 9 5 11 7 2 4 10 6
$
ccecedcceed$
cedcced$
ceccedcced$
ceccedcced$
cedcced$
d$
dccedcced$
ecceedccedcced$
eccedcced$
edcced$
edcced$

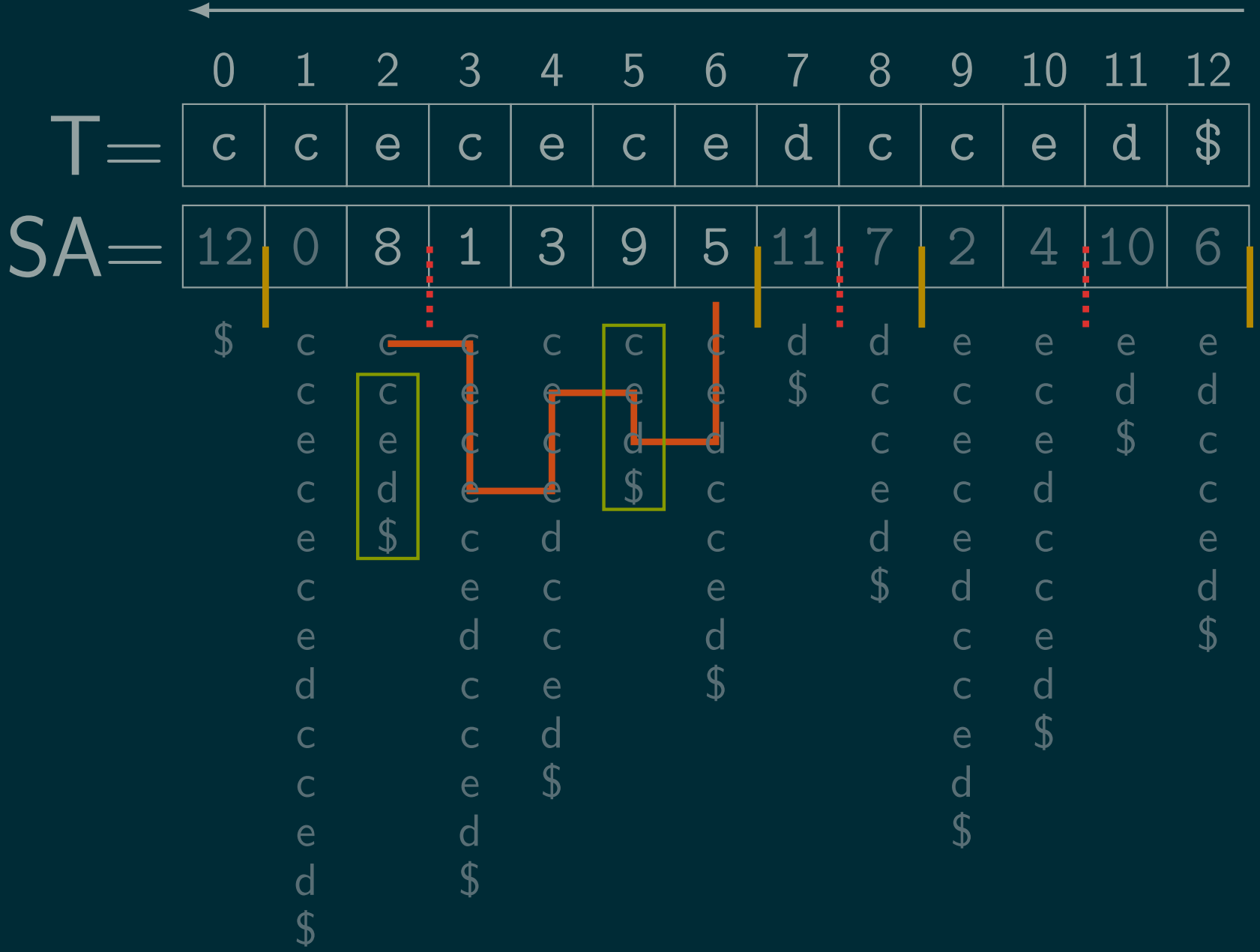
Inducing B


0 1 2 3 4 5 6 7 8 9 10 11 12
T= c c e c e c e d c c e d $
SA= 12 0 8 1 3 9 5 11 7 2 4 10 6
$
ccececedccecd$
cced$
cececcecced$
cecedccced$
ced$
cedccced$
d$
dcced$
ececedcced$
ecedccecd$
ed$
edcced$

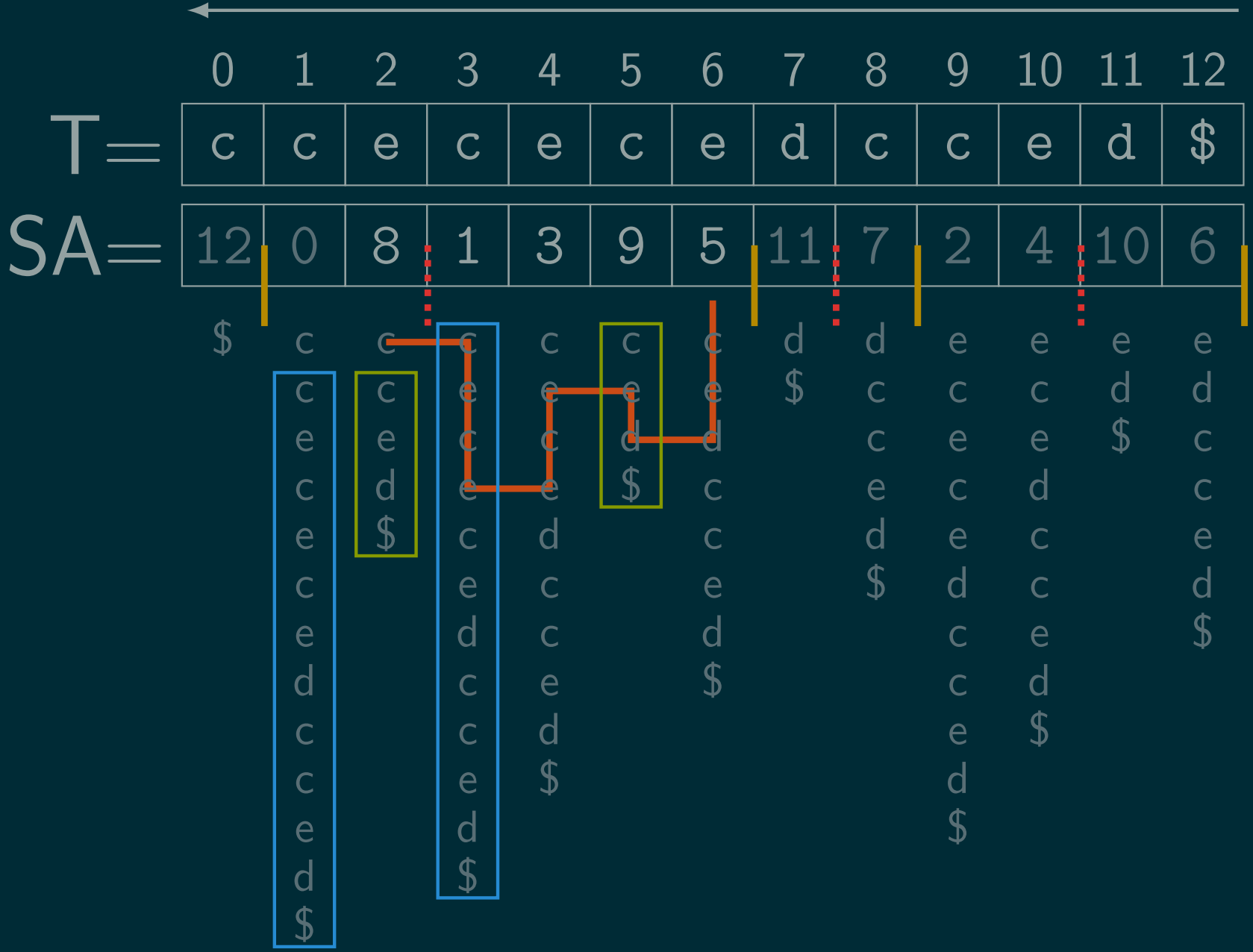
Inducing B


0 1 2 3 4 5 6 7 8 9 10 11 12
T= c c e c e c e d c c e d $
SA= 12 0 8 1 3 9 5 11 7 2 4 10 6
$
ccecedcceed$
ceed$
cecedccedd$
cecedccedd$
cedccedd$
d$
dccedd$
eccedccedd$
ecedccedd$
ed$
edccedd$

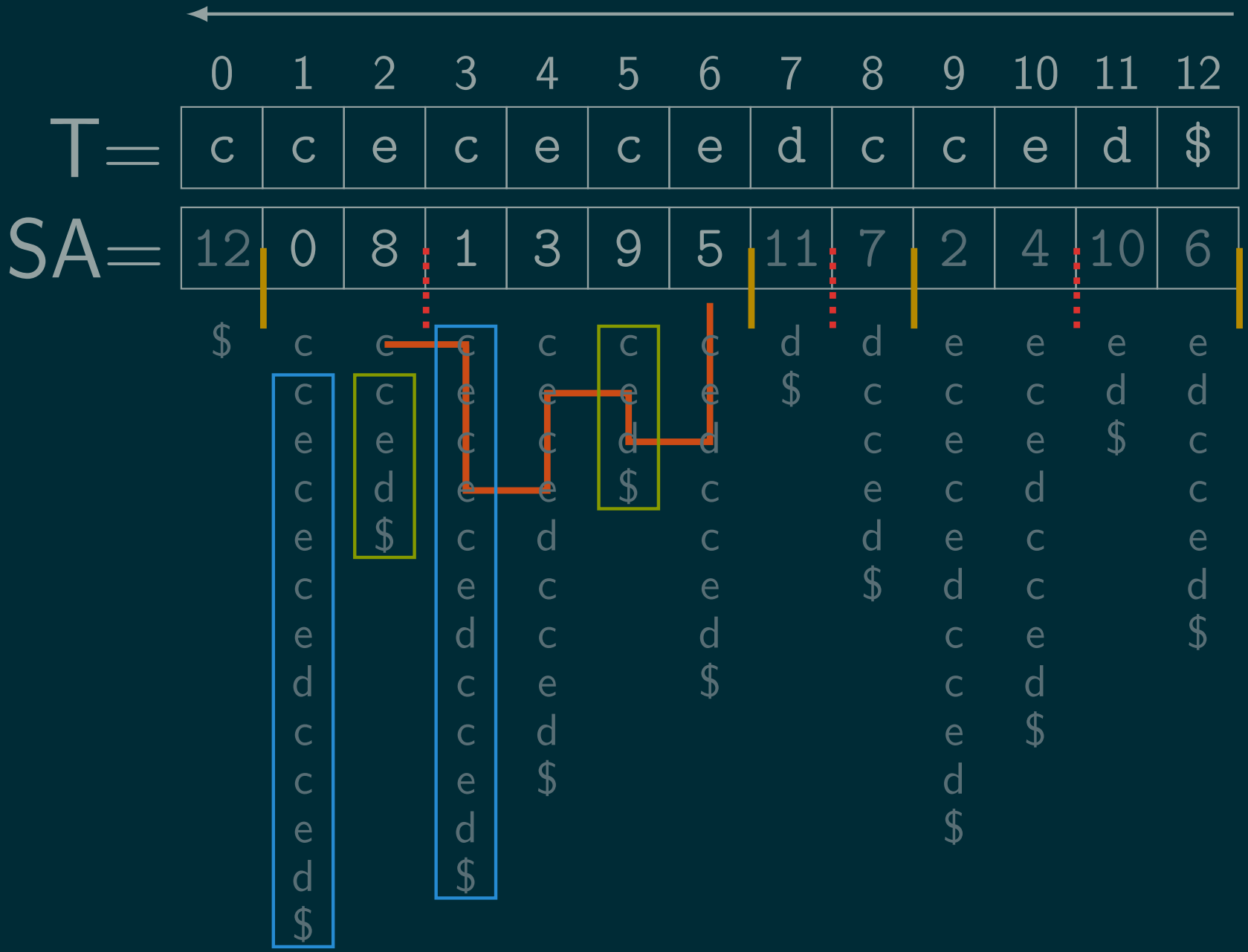
Inducing B


0 1 2 3 4 5 6 7 8 9 10 11 12
T= c c e c e c e d c c e d $
SA= 12 0 8 1 3 9 5 11 7 2 4 10 6
$
ccecedcced$
cecedcced$
ceccedced$
ceccedced$
cedcced$
d$
dcced$
ecceedcced$
ecedcced$
ed$
ecedcced$

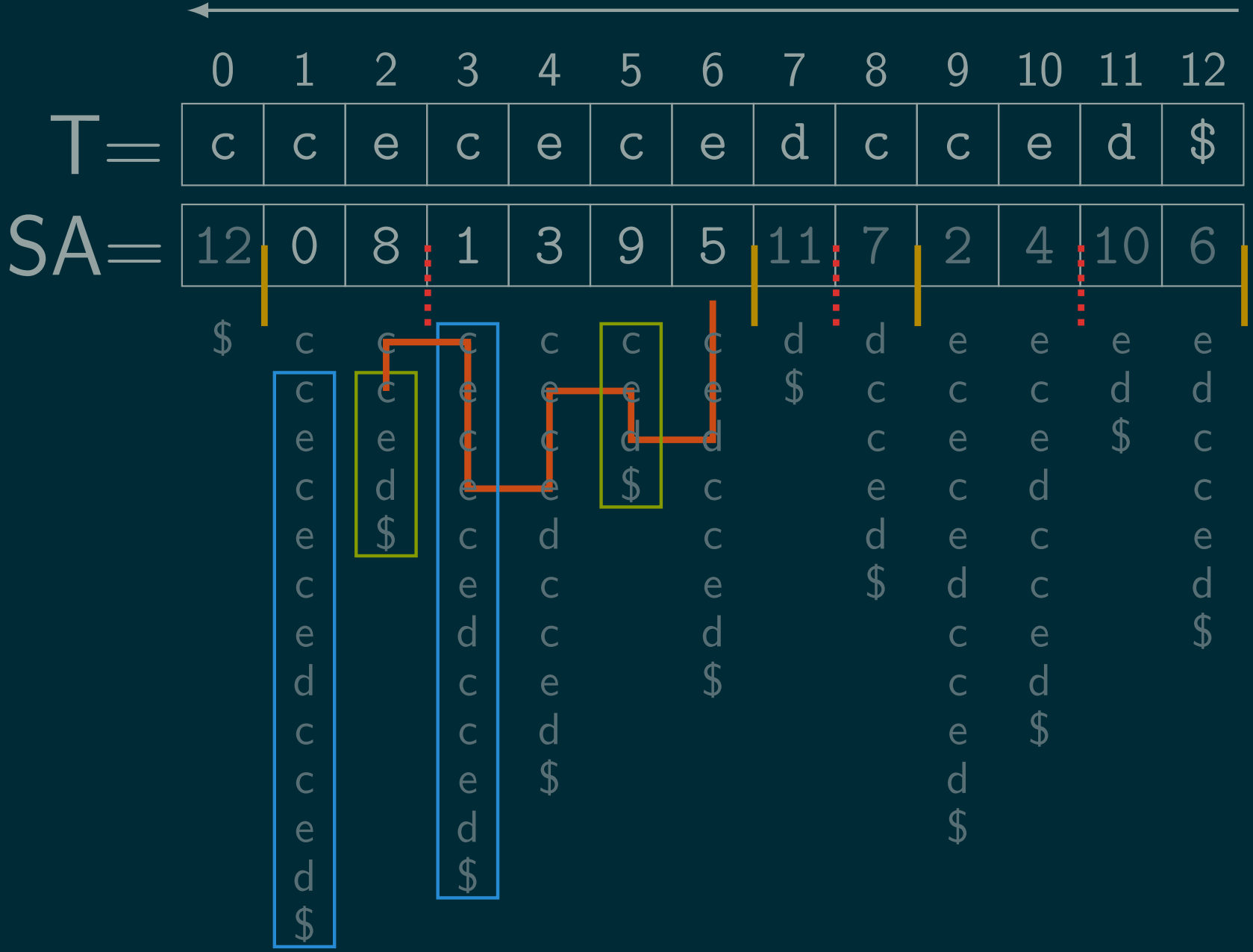
Inducing B


0 1 2 3 4 5 6 7 8 9 10 11 12
T=
c c e c e c e d c c e d $
SA=
12 0 8 1 3 9 5 11 7 2 4 10 6
$
c c e c e c e d c c e d $
c e c e c e d $
c e c e d $
c e c e d $
c e c e d $
c e d $
d $
d c c e d $
e c c e d $
e c e d $
e d $
e d c c e d $

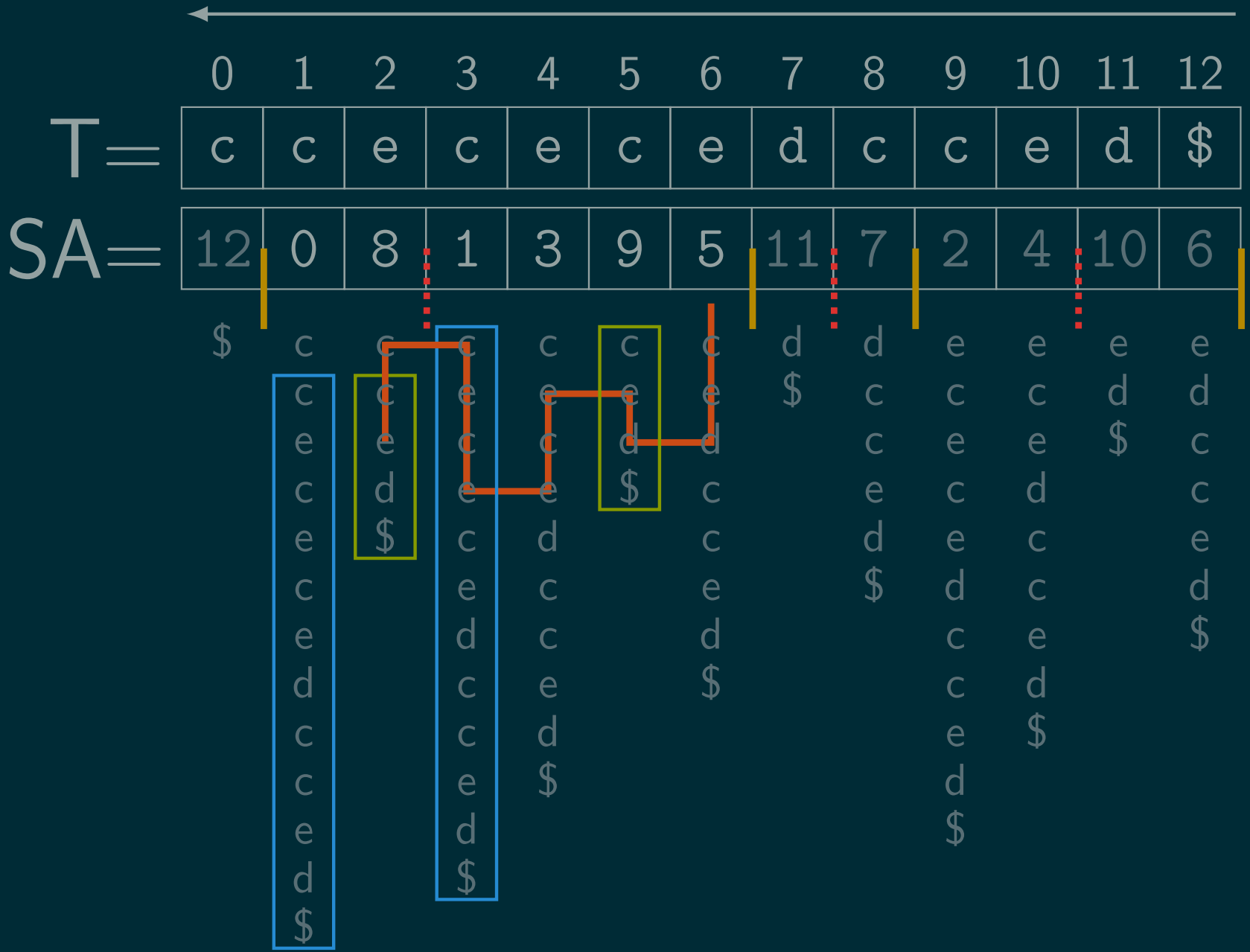
Inducing B


0 1 2 3 4 5 6 7 8 9 10 11 12
T= c c e c e c e d c c e d $
SA= 12 0 8 1 3 9 5 11 7 2 4 10 6
$
c e c e c e d $
c e c e d $
c e c e d $
c e c e d $
c e d $
d $
d c c e d $
e c c e d $
e c e d $
e d c c e d $

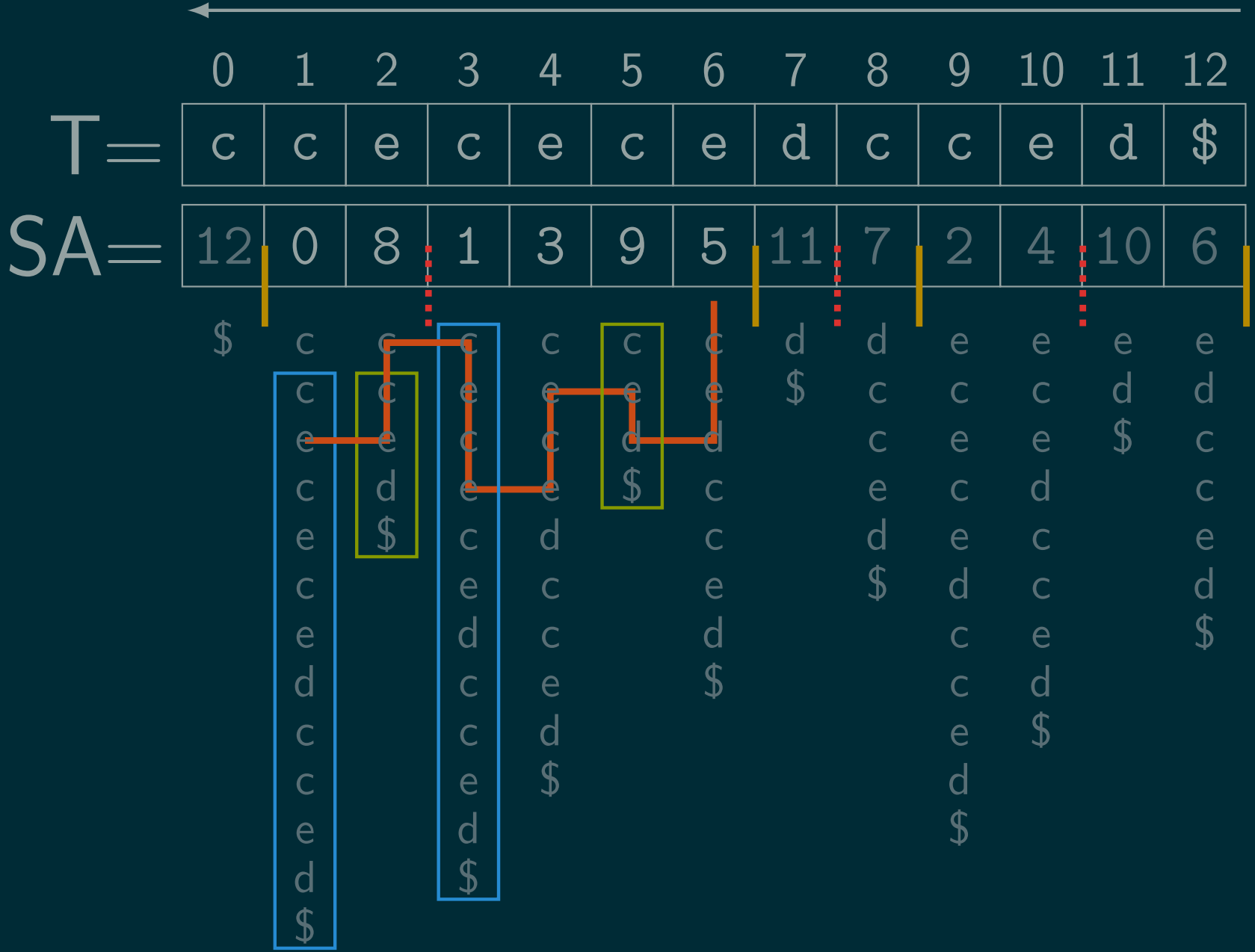
Inducing B



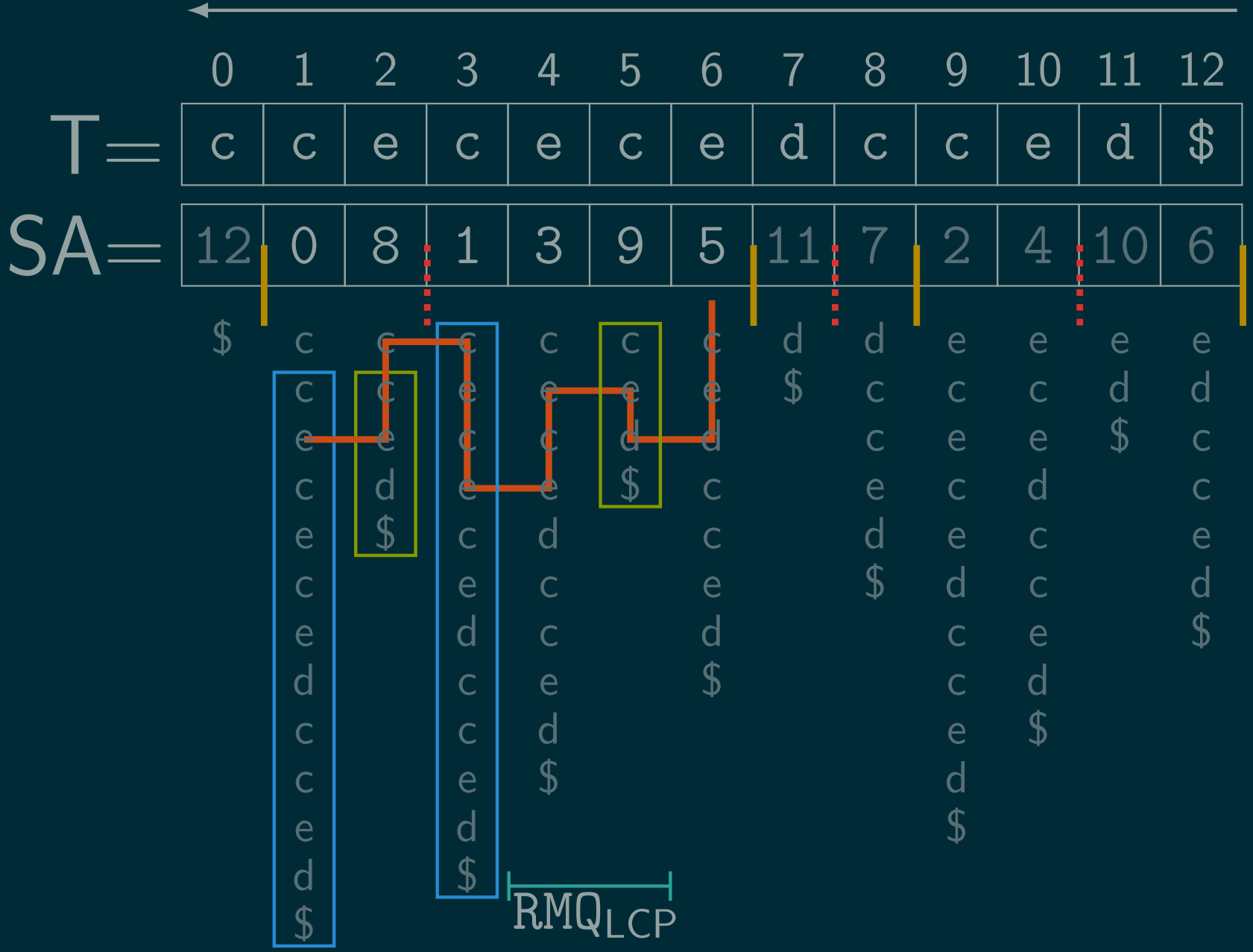
Inducing B


0 1 2 3 4 5 6 7 8 9 10 11 12
T= c c e c e c e d c c e d $
SA= 12 0 8 1 3 9 5 11 7 2 4 10 6
$
c e c e c d $
c e c e d $
c e c e d c e d $
c e c e d $
c e d c c e d $
d $
d c c e d $
e c c e d c c e d $
e c e d $
e d c c e d $

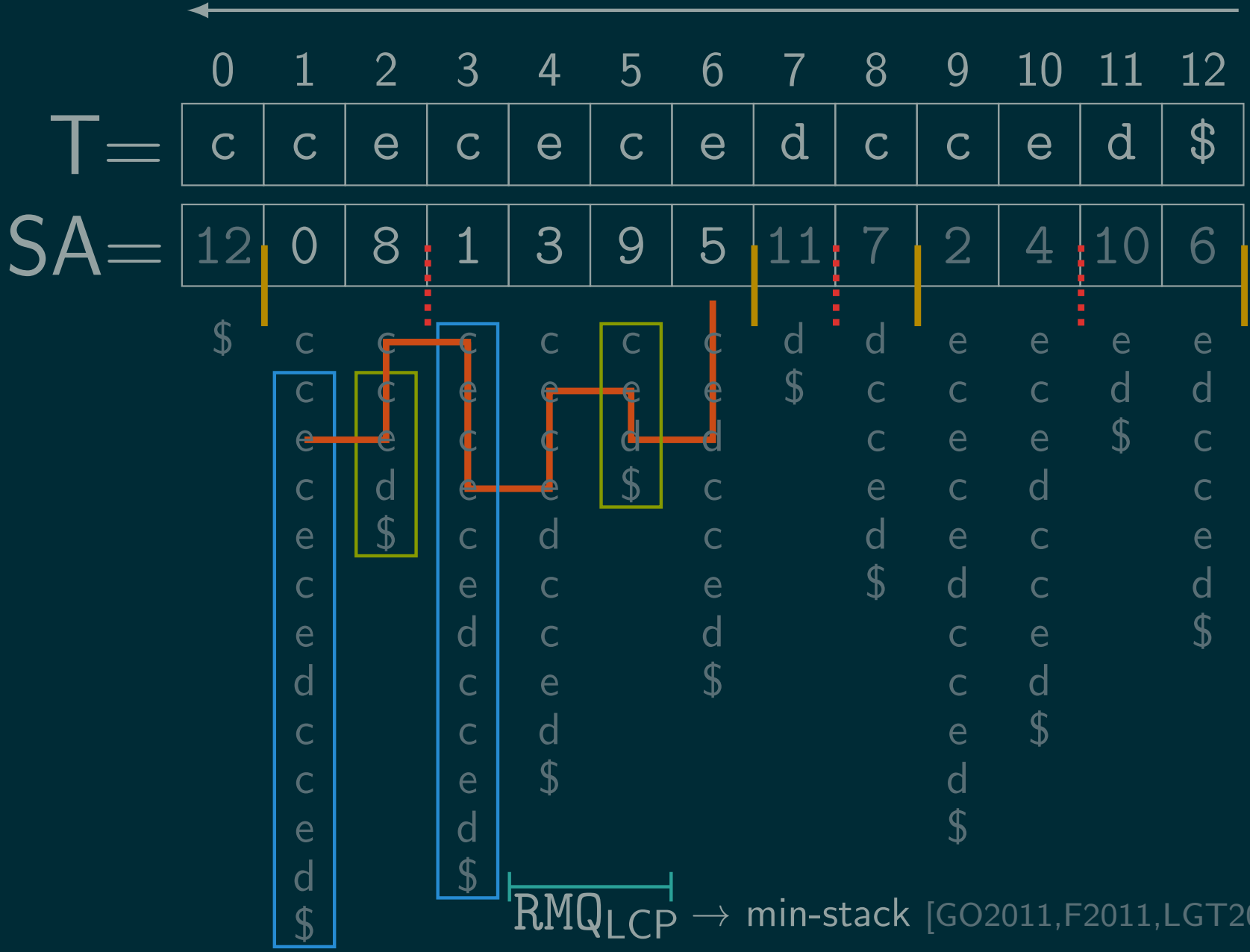
Inducing B


0 1 2 3 4 5 6 7 8 9 10 11 12
T= c c e c e c e d c c e d $
SA= 12 0 8 1 3 9 5 11 7 2 4 10 6
$
c e c e c e d $
c e d $
c e c e d $
c e c e d $
c e d $
c e d $
d $
d c c e d $
e c c e d c c e d $
e c e d c c e d $
e d $
e d c c e d $

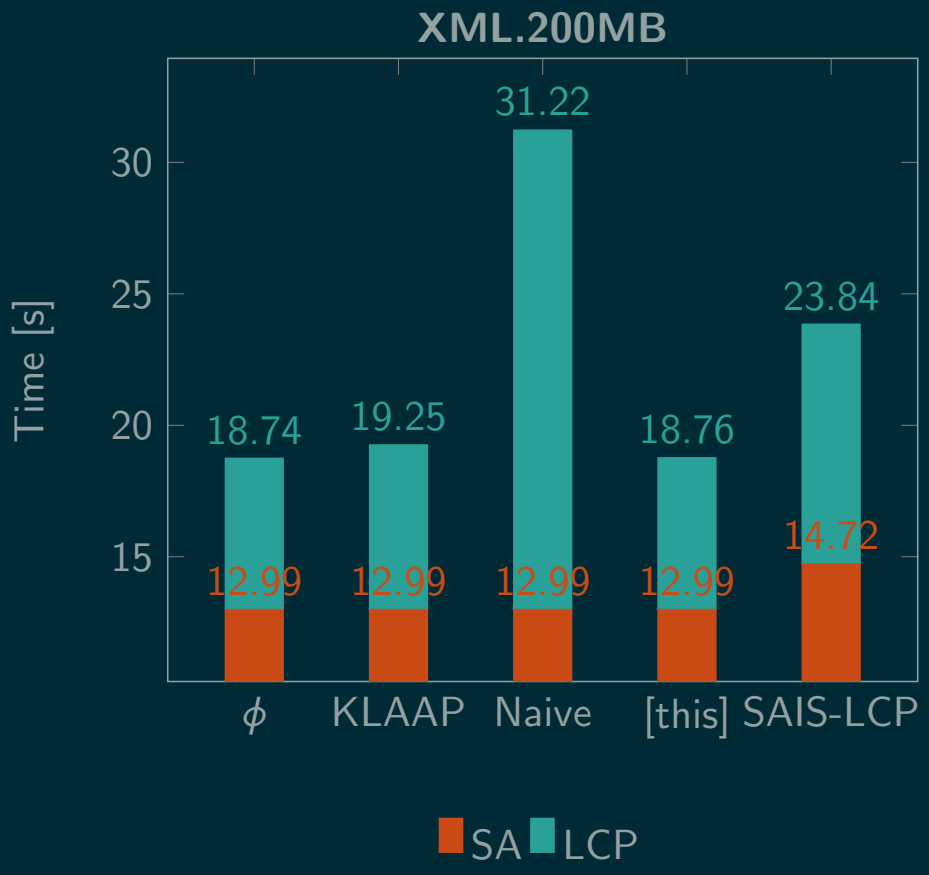
Inducing B

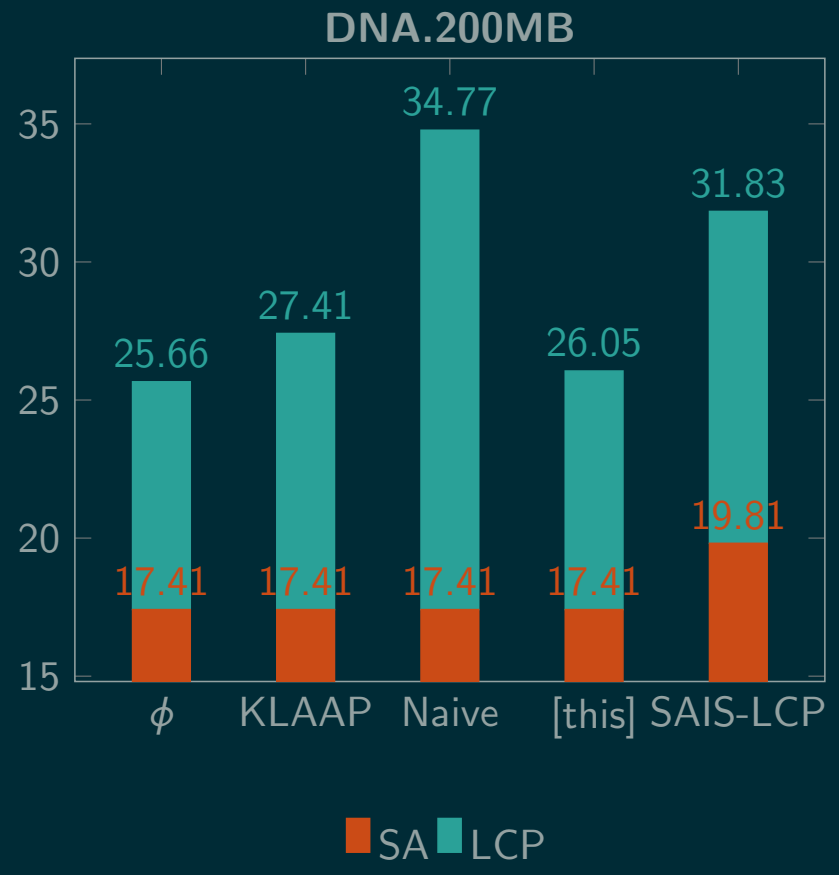

0 1 2 3 4 5 6 7 8 9 10 11 12
T=
c c e c e c e d c c e d $
SA=
12 0 8 1 3 9 5 11 7 2 4 10 6
RMQLCP

Inducing B


0 1 2 3 4 5 6 7 8 9 10 11 12
T=
c c e c e c e d c c e d $
SA=
12 0 8 1 3 9 5 11 7 2 4 10 6
RMQLCP → min-stack [GO2011,F2011,LGT2017]

Experimental Results





Conclusion & Outlook

- Detailed description of DivSufSort
- Fast LCP-construction Algorithm
- (Currently) working on faster LCP-computation during sorting

Conclusion & Outlook

- Detailed description of DivSufSort
- Fast LCP-construction Algorithm
- (Currently) working on faster LCP-computation during sorting

Thank You!

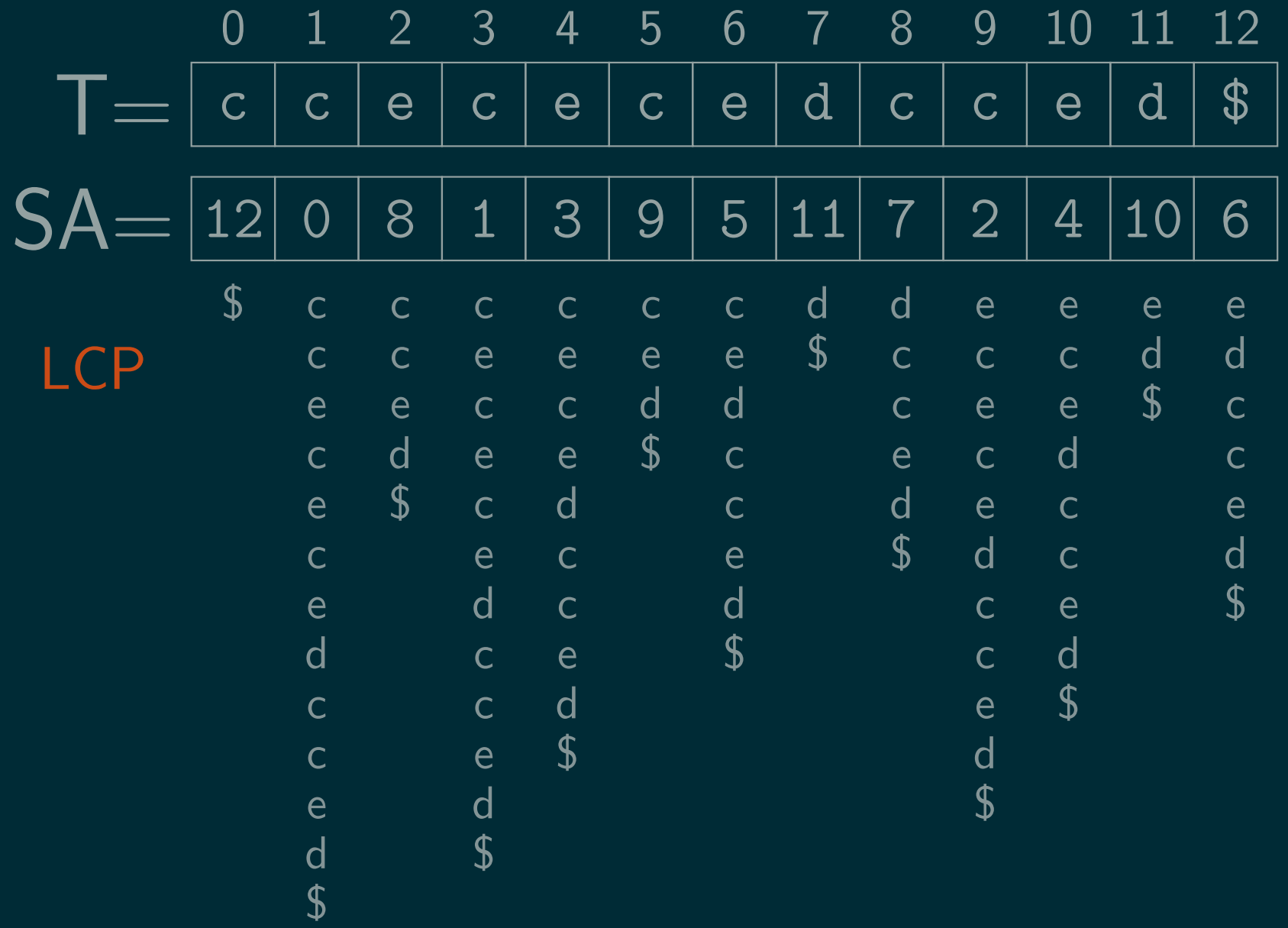
	0	1	2	3	4	5	6	7	8	9	10	11	12
T=	c	c	e	c	e	c	e	d	c	c	e	d	$

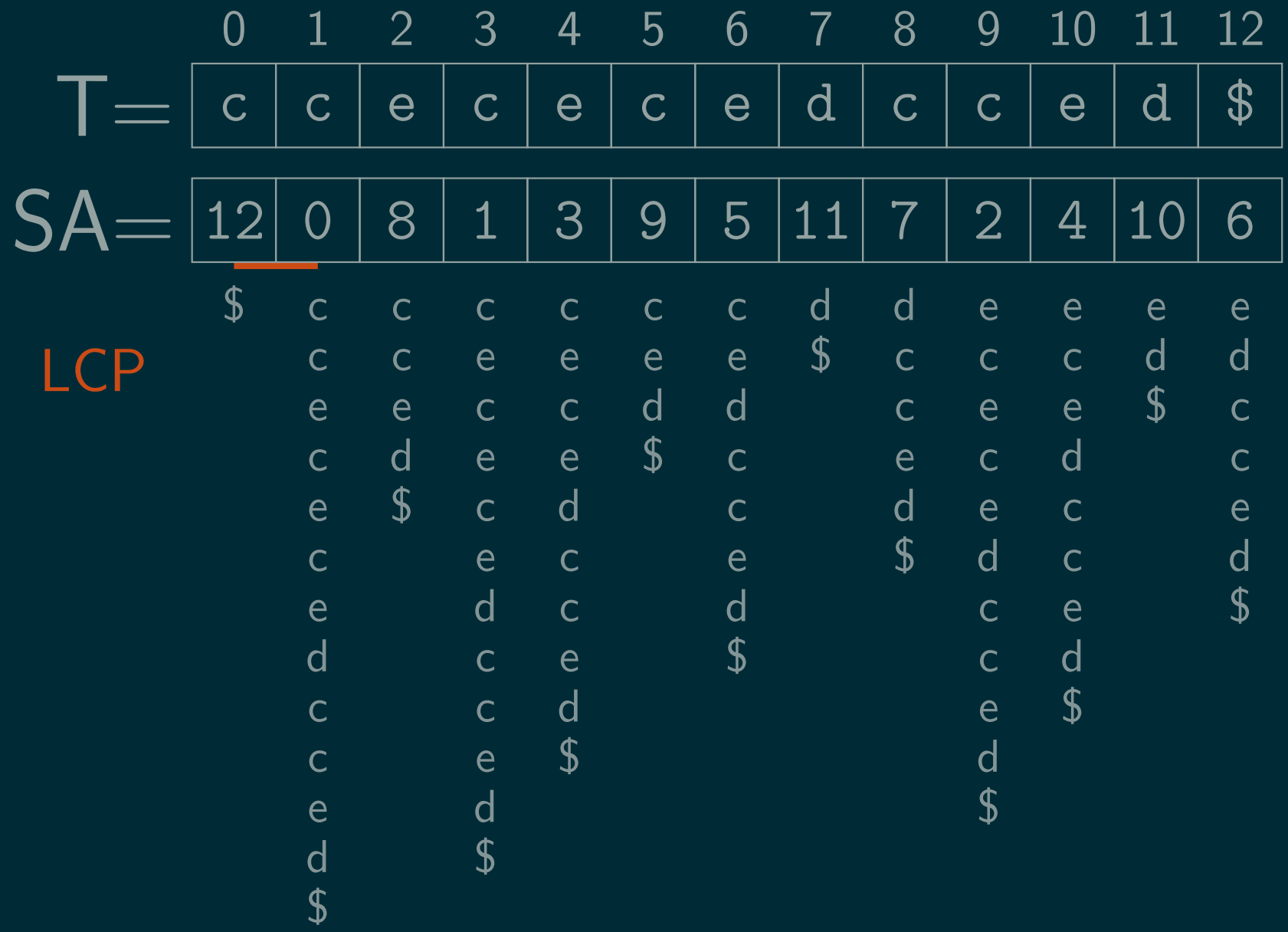
	0	1	2	3	4	5	6	7	8	9	10	11	12
SA=	0	1	2	3	4	5	6	7	8	9	10	11	12

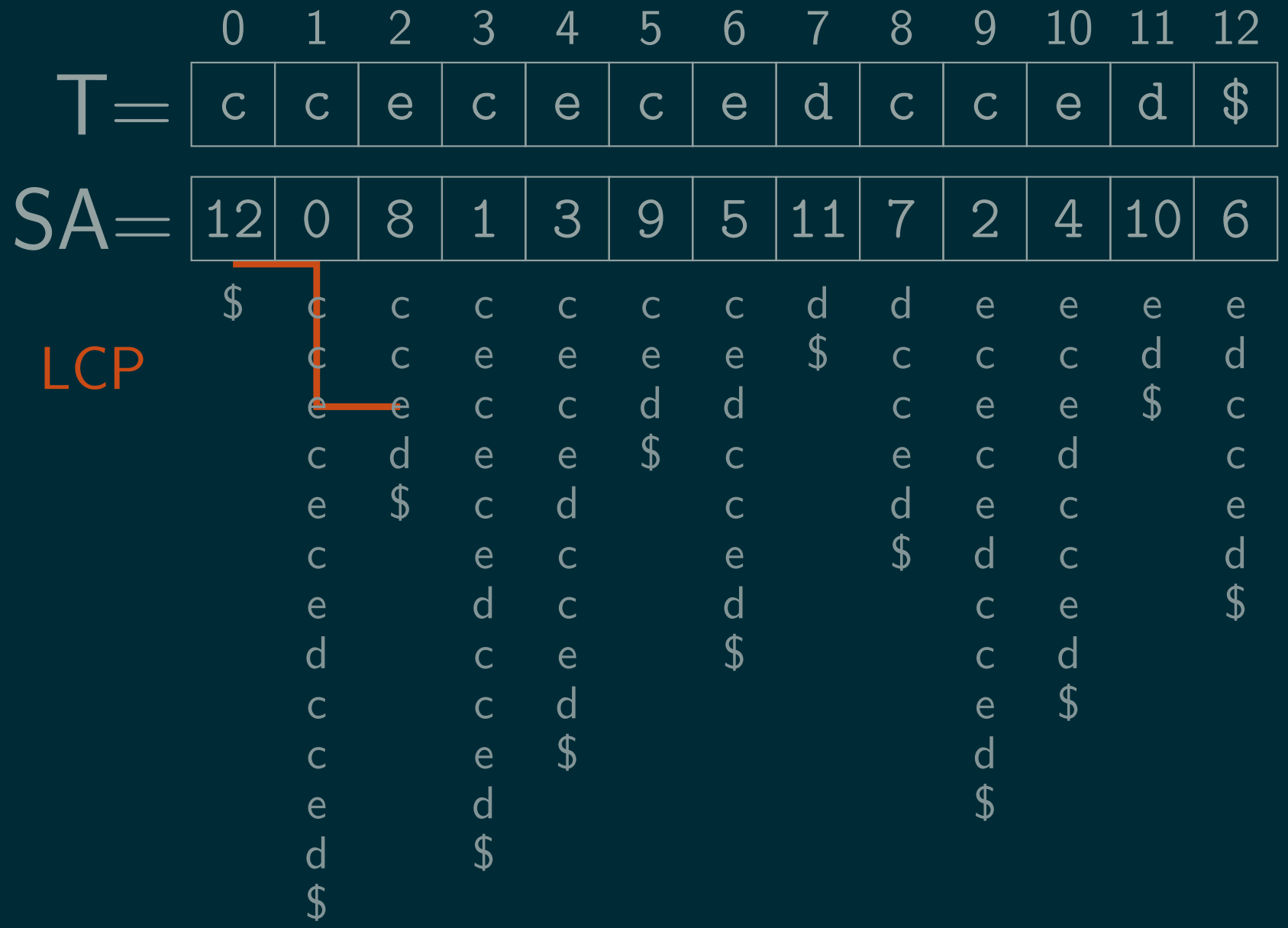
c	c	e	c	e	c	e	d	c	c	e	d	$
c	e	c	e	c	e	d	c	c	e	d	$	
e	c	e	c	e	d	c	c	e	d	$		
c	e	c	e	d	c	c	e	d	$			
e	c	e	d	c	c	e	d	$				
c	e	d	c	c	e	d	$					
e	d	c	c	e	d	$						
d	c	c	e	d	$							
c	c	e	d	$								
e	e	d	$									
d	d	$										
$	$											

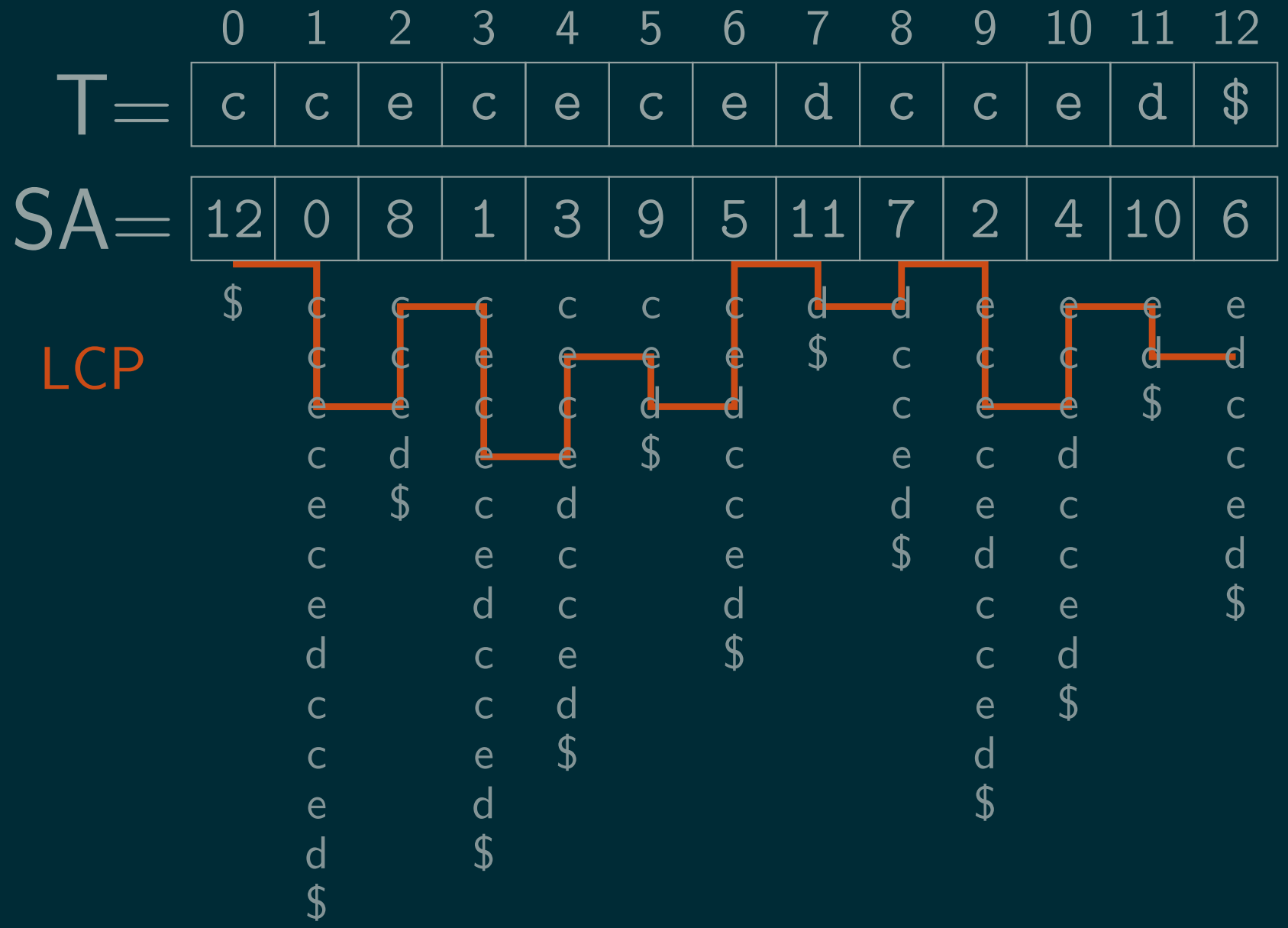
	0	1	2	3	4	5	6	7	8	9	10	11	12
T=	c	c	e	c	e	c	e	d	c	c	e	d	$
SA=	12	0	8	1	3	9	5	11	7	2	4	10	6

$	c	c	c	c	c	c	d	d	e	e	e	e	
	c	c	e	e	e	e	$	c	c	c	c	d	
	e	e	c	c	d	d		c	e	e	d	$	
	c	d	e	e	c	c		e	c	c	c		
	e	$	c	d	c	e		d	e	e	e		
	c		e	c	e	d		$	d	d	d		
	e		d	c	d	$			c	c	$		
	d		c	e	$				e	d			
	c		e	d					d	$			
	e		d	$					$				
	d		$										
	$												


0 1 2 3 4 5 6 7 8 9 10 11 12
T= c c e c e c e d c c e d $
SA= 12 0 8 1 3 9 5 11 7 2 4 10 6
LCP
$
c c e c e d c c e d $
c c e d $
c e c e d c c e d $
c e c e d c c e d $
c e d $
c e d c c e d $
d $
d c c e d $
e c e c e d c c e d $
e c e d c c e d $
e d $
e d c c e d $


0 1 2 3 4 5 6 7 8 9 10 11 12
T= c c e c e c e d c c e d $
SA= 12 0 8 1 3 9 5 11 7 2 4 10 6
LCP
$
c c e c e d $
c e d $
c e c e d $
c e c e d $
c e d $
c e d c c e d $
d $
d c c e d $
e c e c e d $
e c e d $
e d $
e d c c e d $


0 1 2 3 4 5 6 7 8 9 10 11 12
T= c c e c e c e d c c e d $
SA= 12 0 8 1 3 9 5 11 7 2 4 10 6
LCP
$
c c e c e d $
c e d $
c e c e d $
c e d c e d $
c e d $
c e d $
d $
d c e d $
e c e d $
e c d $
e d $
e d c e d $


0 1 2 3 4 5 6 7 8 9 10 11 12
T= c c e c e c e d c c e d $
SA= 12 0 8 1 3 9 5 11 7 2 4 10 6
LCP
$
c c e c e d $
c e d $
c e c e d $
c e c d $
c e d $
c e d $
d $
d c e d $
e c e d $
e d $
e d $