

---

# Algorithmen und Datenstrukturen

Prof. Dr. Ralf Möller

**Universität zu Lübeck**

**Institut für Informationssysteme**

Tanya Braun (Übungen)

sowie viele Tutoren



# Danksagung

---

- Das folgende Präsentationsmaterial wurde von Sven Groppe für das Modul Algorithmen und Datenstrukturen erstellt und mit Änderungen hier übernommen (z.B. werden Algorithmen im Pseudocode präsentiert)



# Motivation Zeichenkettenabgleich

---

- Gegeben eine Folge von Zeichen (Text), in der eine Zeichenkette (Muster) gefunden werden soll
- Varianten
  - Alle Vorkommen des Musters im Text
  - Ein beliebiges Vorkommen im Text
  - Erstes Vorkommen im Text
- Anwendungen
  - Suchen von Mustern in DNA-Sequenzen (begrenzttes Alphabet: A, C, G, T)

# Teilzeichenkette, Präfix, Suffix

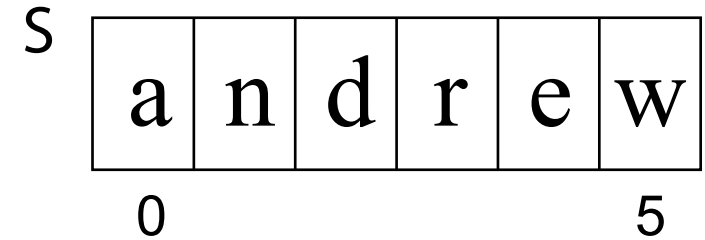
---

- $S$  sei eine **Zeichenkette** der Länge  $m$
- $S[i..j]$  ist dann eine **Teilzeichenkette** von  $S$  zwischen den Indizes  $i$  und  $j$  ( $0 \leq i \leq j \leq m-1$ )
- Ein **Präfix** ist eine Teilzeichenkette  $S[0..i]$  ( $0 \leq i \leq m-1$ )
- Eine **Suffix** ist eine Teilzeichenkette  $S[i..m-1]$  ( $0 \leq i \leq m-1$ )

# Beispiele

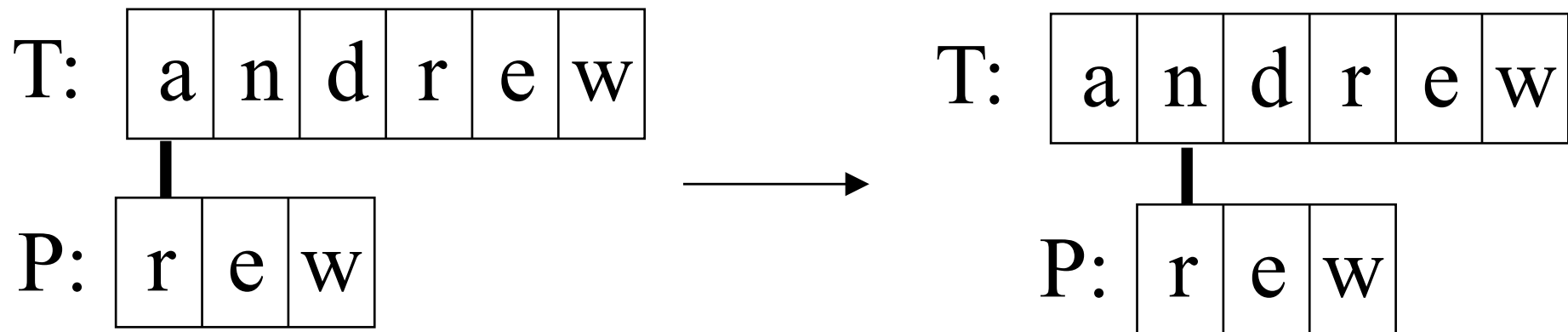
---

- Teilzeichenkette  $S[1..3] = \text{"ndr"}$
- Alle möglichen Präfixe von S:
  - "andrew", "andre", "andr", "and", "an", "a"
- Alle möglichen Suffixe von S:
  - "andrew", "ndrew", "drew", "rew", "ew", "w"



# Der Brute-Force-Algorithmus

- **Problem:** Bestimme Position des ersten Vorkommens von Muster **P** in Text **T** oder liefere **-1**, falls **P** nicht in **T** vorkommt
- **Idee:** Überprüfe jede Position im Text, ob das Muster dort startet



P bewegt sich jedes Mal um 1 Zeichen durch T

...

# Brute-Force-Suche

---

```
function BFsearch(text, pattern: String): Integer
  n := length(text); m := length(pattern)
  for i from 0 to (n-m) do
    j := 0
    while j < m and text[i+j] = pattern[j] do // passende Teilkette
      j := j + 1
    if j = m then
      return i // erfolgreiche Suche
  return -1 // erfolglose Suche
```

- Datentyp `String` entspricht `Array [0..length-1]` of `Character`
- `Character` umfasse hier 128 Zeichen (ASCII)

# Analyse der Komplexität für Suche

---

- Schlechtester Fall für erfolglose Suche
  - Beispiel
    - Text: „aaaaaaaaaaaaaaaaaaaaaaaa“
    - Muster: „aaaah“
  - Das Muster wird an jeder Position im Text durchlaufen:  $O(n \cdot m)$
- Bester Fall für erfolglose Suche
  - Beispiel
    - Text: „aaaaaaaaaaaaaaaaaaaaaaaa“
    - Muster: „bbbbbb“
  - Das Muster kann an jeder Position im Text bereits am ersten Zeichen des Musters falsifiziert werden:  $O(n)$
- Komplexität erfolgreiche Suche im Durchschnitt
  - Meist kann das Muster bereits an der ersten Stelle des Musters falsifiziert werden und in der Mitte des Textes wird das Muster gefunden:  $O(n+m)$

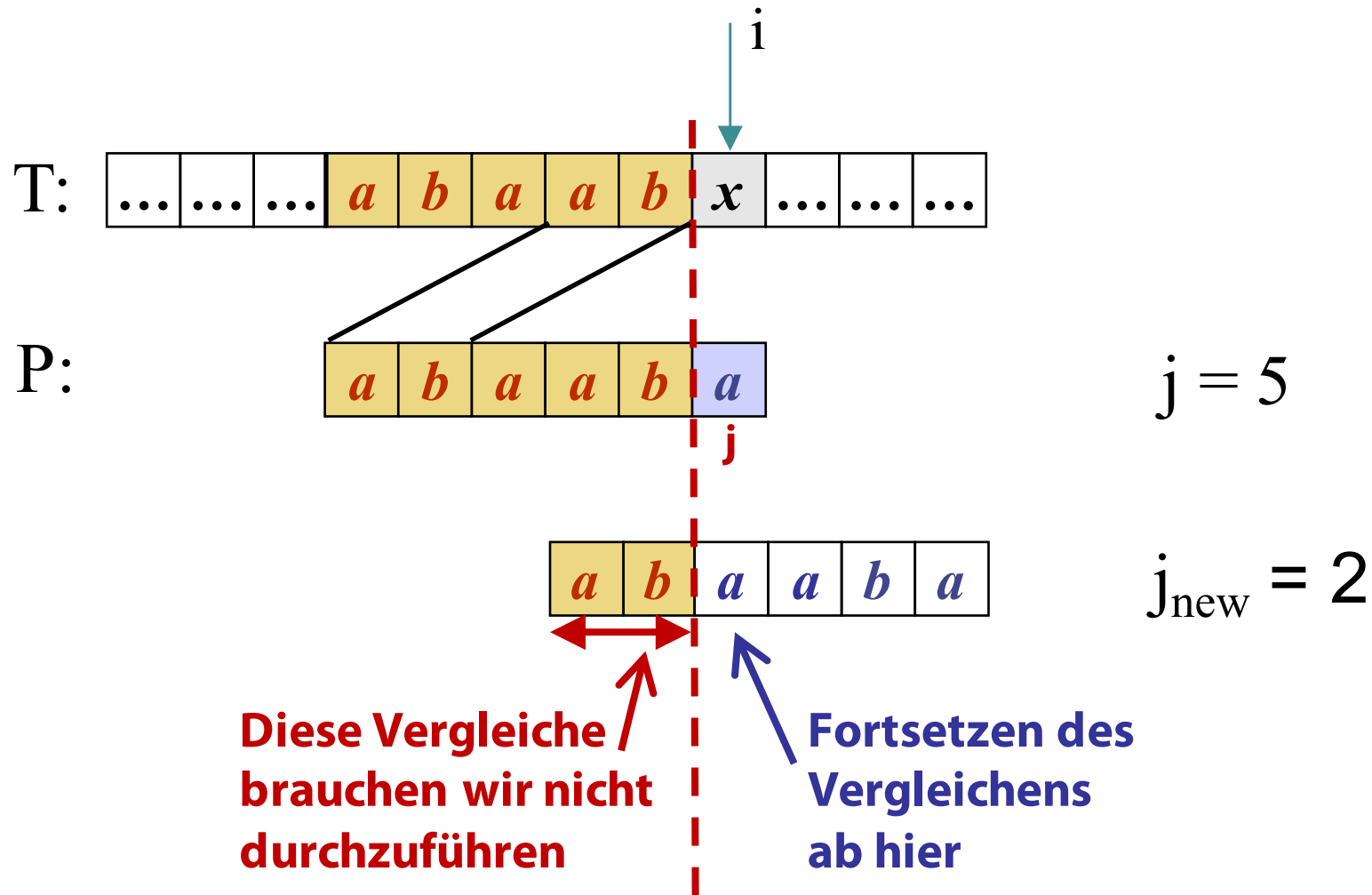


# Weitere Analyse

---

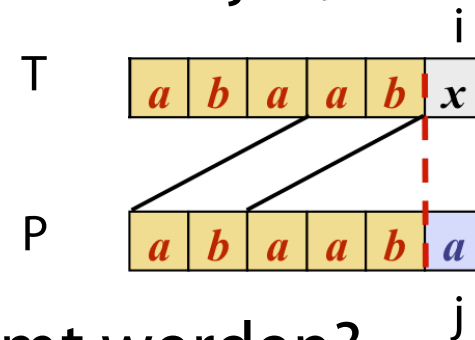
- Brute-Force-Algorithmus ist um so schneller, je größer das Alphabet ist
  - Größere Häufigkeit, dass das Muster bereits in den ersten Zeichen falsifiziert werden kann
- Für kleine Alphabete (z.B. binär 0,1) ungeeigneter
- Bessere Verschiebung des Musters zum Text als bei Brute-Force möglich?

# Beispiel

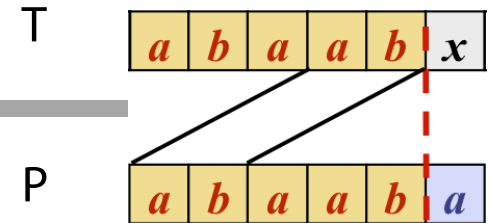


# Knuth-Morris-Pratt-Algorithmus (KMP)

- Vergleich des Musters im Text von links nach rechts
  - Wie der Brute-Force-Ansatz
- Bessere Verschiebung des Musters zum Text als bei Brute-Force
  - **Frage:** Falls das Muster an der Stelle  $j$  falsifiziert wird, was ist die größtmögliche Verschiebung, um unnötige Vergleiche zu sparen?
  - **Antwort:** um den längsten Präfix von  $P[0..j-1]$ , der ein Suffix von  $T[i-j..i-1]$  ist
- Wie können solche Präfixe mit vertretbarem Aufwand bestimmt werden?

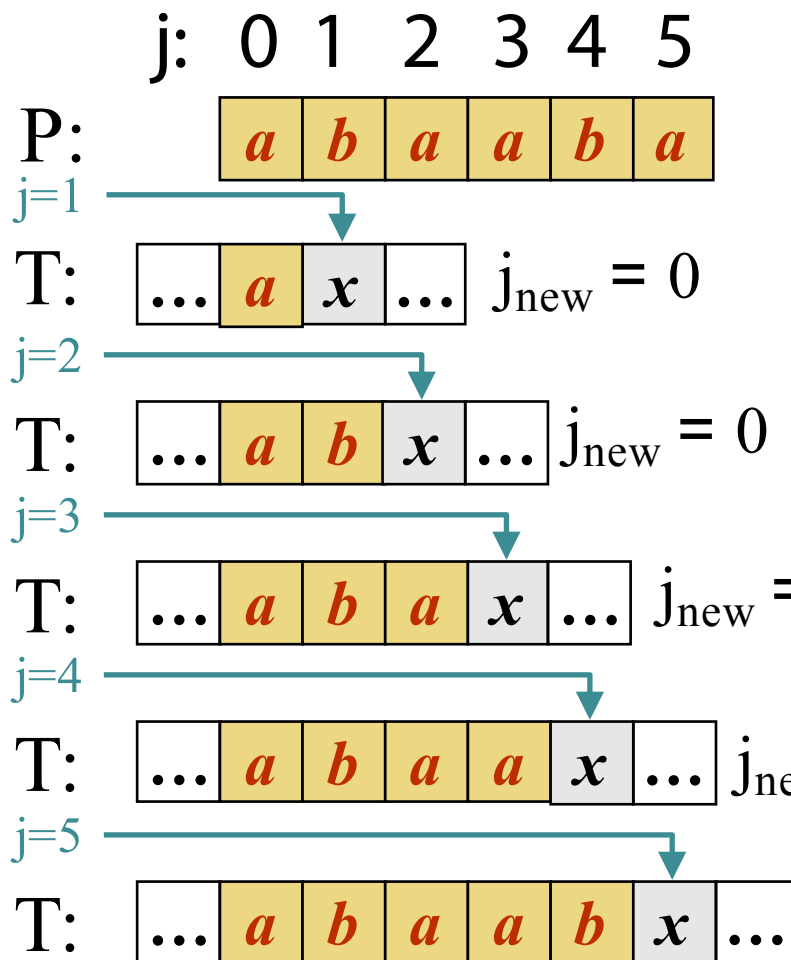


# KMP Fehlerfunktion



- KMP verarbeitet das Muster vor, um Übereinstimmungen zwischen den Präfixen des Musters mit sich selbst zu finden
- $j$  = Position der Ungleichheit in P
- $k$  = Position vor der Ungleichheit ( $k = j-1$ ).
- Die sog. Fehlerfunktion  $F(k)$  ist definiert als die Länge des **längsten Präfixes von  $P[0..k]$** , welcher auch ein Suffix  $T[i-j..i-1]$  von T ist
- Oder welcher **auch ein Suffix von  $P[0..k]$**  ist!
  - Wenn das nicht so wäre, käme man nicht an die Pos.  $j$

# Beispiel Fehlerfunktion



<b>j</b>	1	2	3	4	5
<b>k = j-1</b>	0	1	2	3	4
<b>F(k)</b>	0	0	1	1	2

- F(k) ist die Länge des längsten Präfix ( $j_{\text{new}}$  für  $j=k+1$ )
- Im Code wird F(k) als ein Feld gespeichert

Das erste a in P braucht bei fortgesetzter Suche nicht überprüft zu werden!

a und b in P braucht bei fortgesetzter Suche nicht überprüft zu werden!



# Beispiel

T: 

<i>a</i>	<i>b</i>	<i>a</i>	<i>c</i>	<i>a</i>	<i>a</i>	<i>b</i>	<i>a</i>	<i>c</i>	<i>c</i>	<i>a</i>	<i>b</i>	<i>a</i>	<i>c</i>	<i>a</i>	<i>b</i>	<i>a</i>	<i>a</i>	<i>b</i>	<i>b</i>
0	1	2	3	4	5														

P: 

<i>a</i>	<i>b</i>	<i>a</i>	<i>c</i>	<i>a</i>	<i>b</i>
0	1	2	3	4	5

 $F(4)=1$

<i>a</i>	<i>b</i>	<i>a</i>	<i>c</i>	<i>a</i>	<i>b</i>
0	1	2	3	4	5

 $F(0)=0$

<i>a</i>	<i>b</i>	<i>a</i>	<i>c</i>	<i>a</i>	<i>b</i>
0	1	2	3	4	5

 $F(3)=0$

<i>a</i>	<i>b</i>	<i>a</i>	<i>c</i>	<i>a</i>	<i>b</i>
0	1	2	3	4	5

 Kein übereinstimmendes Zeichen  
 $i:=i+1$

<i>a</i>	<i>b</i>	<i>a</i>	<i>c</i>	<i>a</i>	<i>b</i>
0	1	2	3	4	5

<i>k</i>	0	1	2	3	4
$F(k)$	0	0	1	0	1

# Unterschiede im Code zum Brute-Force-Algorithmus

---

- Knuth-Morris-Pratt Ansatz modifiziert den Brute-Force-Algorithmus
- Falls keine Übereinstimmung bei  $P[j]$  (d.h.  $P[j] \neq T[i]$ ), dann
  - $k := j-1;$
  - $j := F(k);$  // berechne neues  $j$

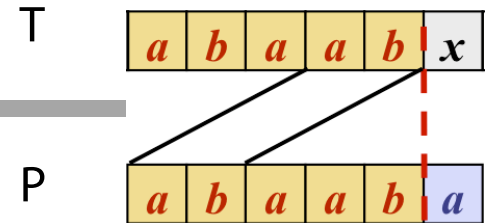
# Verfahren Knuth-Morris-Pratt

---

```
Procedure KMPsearch(text, pattern: String): Integer
  n := length(text); m := length(pattern)
  F[] := computeF(pattern)
  i := 0; j := 0
  while i < n do // passende Teilkette
    if pattern[j] = text[i] then
      if j = m - 1 then
        return i - m + 1 // erfolgreiche Suche
      i := i + 1
      j := j + 1
    else if j > 0 then
      j := F[j - 1]
    else i := i + 1
  return -1 // erfolglose Suche
```



# Fehlerfunktion computeF



```
function computeF(pattern: String): Array[] of IN
  F := <0,...,0>: Array [0..length(pattern)-1 - 1] of IN
```

```
F[0] := 0
```

```
m := length(F)
```

```
i := 1; j := 0
```

```
while i < m do
```

```
  if pattern[j] = pattern[i] then
```

```
    // j+1 Zeichen stimmen überein
```

```
    F[i] := j + 1
```

```
    i := i + 1; j := j + 1
```

```
  else if j > 0 then
```

```
    // j folgt dem übereinstim-
```

```
    // menden Präfix
```

```
    j := F[j - 1]
```

```
  else // keine Übereinstimmung
```

```
    F[i] := 0
```

```
    i := i + 1
```

```
return F
```

										Länge
<b>Position:</b>	0	1	2	3	4	5	6	7	8	
<b>Muster:</b>	a	b	a	b	c	a	b	a	b	9
<b>Präfix (0..0):</b>										0
<b>Präfix (0..1):</b>										0
<b>Präfix (0..2):</b>	a		a							1
<b>Präfix (0..3):</b>	a	b	a	b						2
<b>Präfix (0..4):</b>										0
<b>Präfix (0..5):</b>	a					a				1
<b>Präfix (0..6):</b>	a	b				a	b			2
<b>Präfix (0..7):</b>	a	b	a			a	b	a		3
<b>Präfix (0..8):</b>	a	b	a	b		a	b	a	b	4

# Analyse des Knuth-Morris-Pratt-Algorithmus

---

- Der Algorithmus springt niemals zurück im Text
  - Geeignet für Datenströme
- Komplexität  $O(m+n)$
- Algorithmus wird i.a. langsamer, wenn das Alphabet größer ist
  - Werte der Fehlerfunktion werden tendenziell kleiner

# Erweiterungen von Knuth-Morris-Pratt

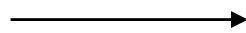
- Der Original-Algorithmus berücksichtigt bei der Verschiebung nicht das Zeichen auf Grund dessen keine Übereinstimmung gefunden wurde

T: 

	a	b	a	a	b	x	
--	---	---	---	---	---	---	--

P: 

a	b	a	a	b	a
---	---	---	---	---	---

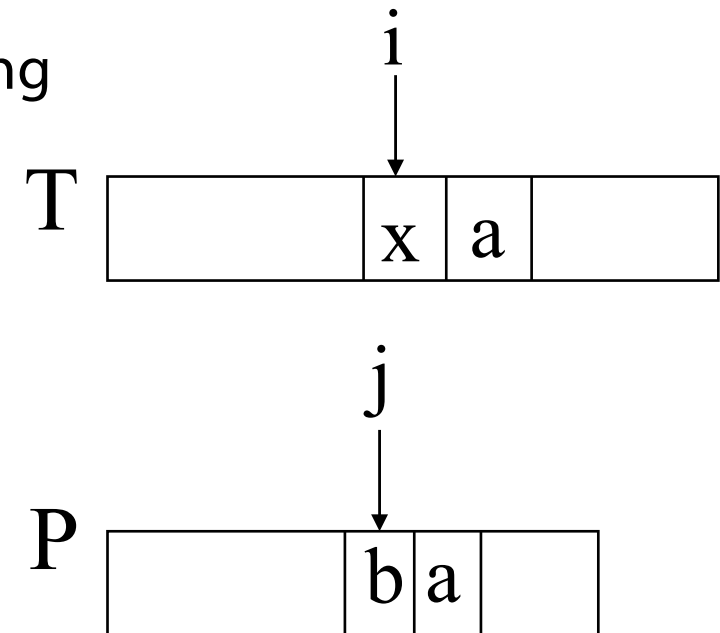


a	b	a	a	b	a
---	---	---	---	---	---

Original-KMP  
„macht“ dieses **nicht!**

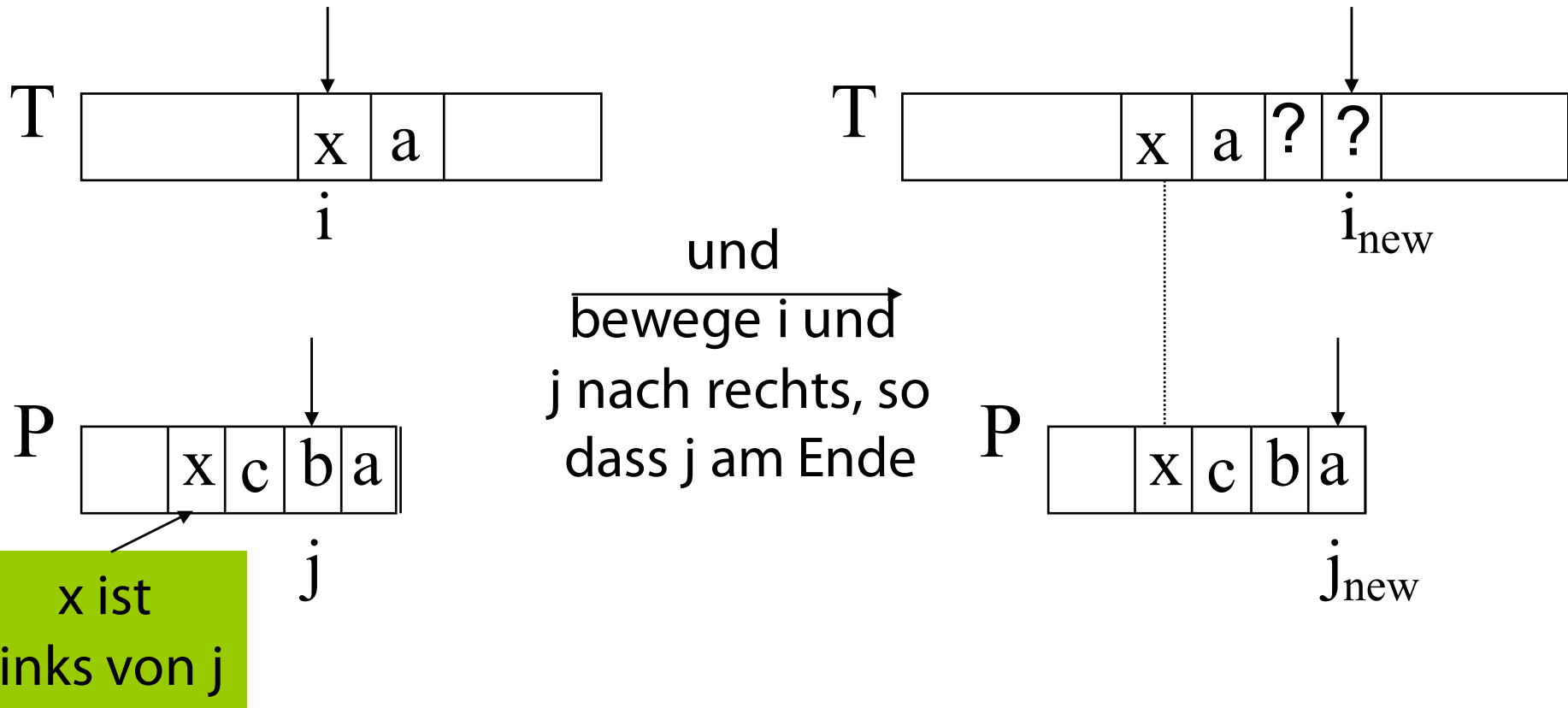
# Boyer-Moore-Algorithmus

- Basiert auf 2 Techniken
  - Spiegeltechnik
    - Finde  $P$  in  $T$  durch Rückwärtslaufen durch  $P$ , am Ende beginnend
  - Zeichensprungtechnik
    - Im Falle von Nichtübereinstimmung des Textes an der  $i$ -ten Position ( $T[i]=x$ ) und des Musters an der  $j$ -ten Position ( $P[j] \neq T[i]$ )
    - 3 Fälle...



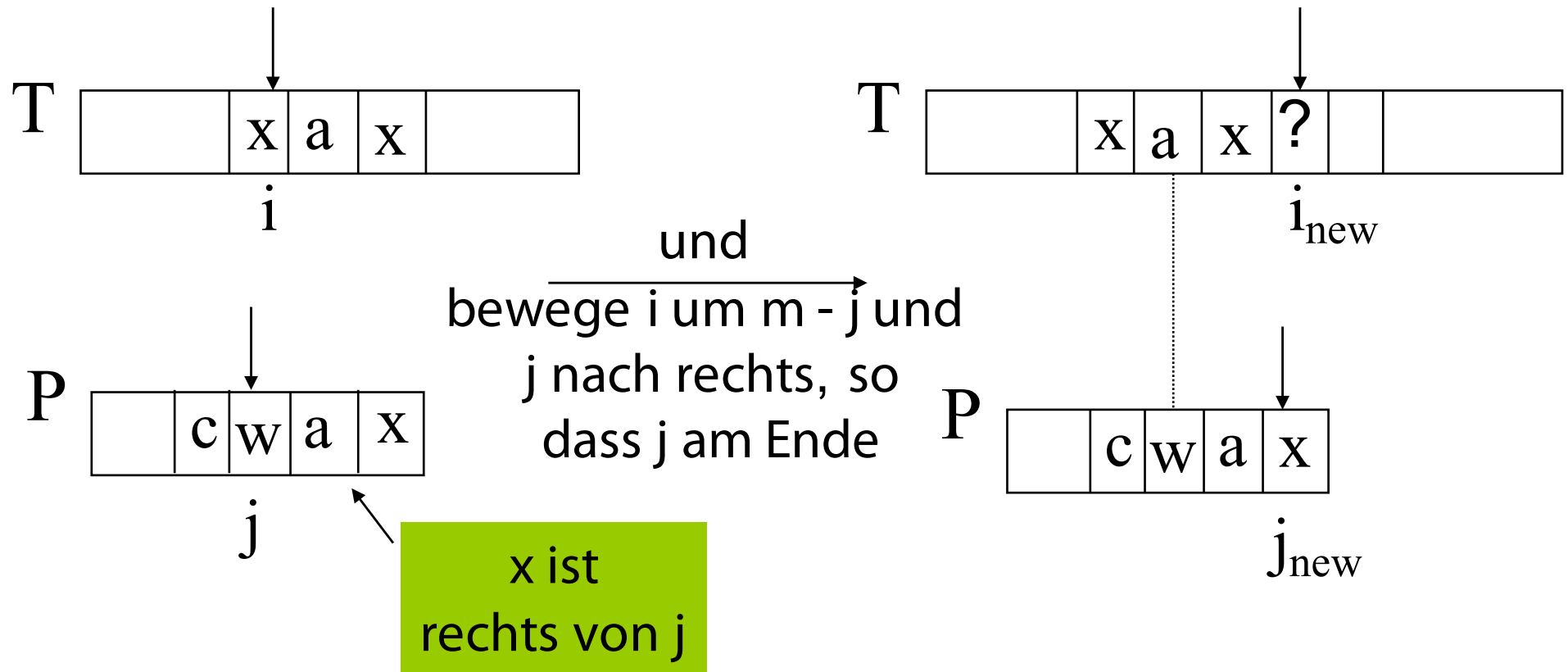
# Fall 1: P enthält x nur links von j

- Bewege P nach rechts, um das letzte Vorkommen von x in P mit T[i] abzugleichen



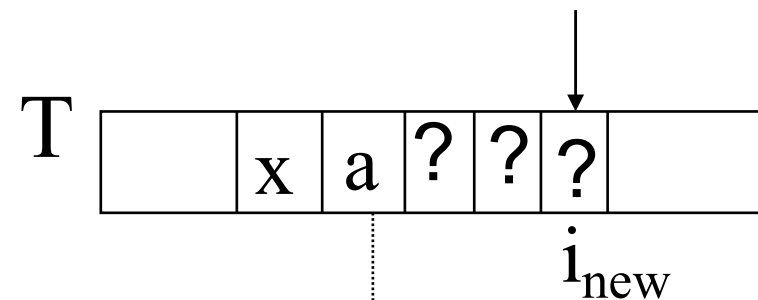
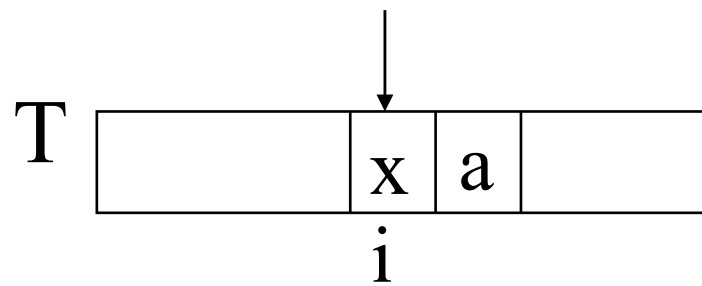
# Fall 2: P enthält x rechts von j

- Bewege P um 1 Zeichen nach T[i+1]

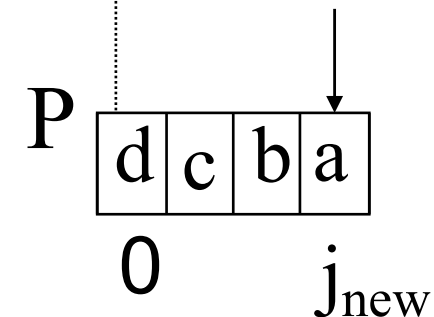
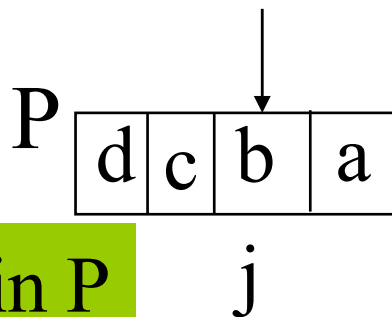


# Fall 3: Falls Fall 1 und 2 nicht anzuwenden sind ( $x$ ist *nicht* in $P$ enthalten)

- Bewege  $P$  nach rechts, um  $P[0]$  und  $T[i+1]$  abzugleichen

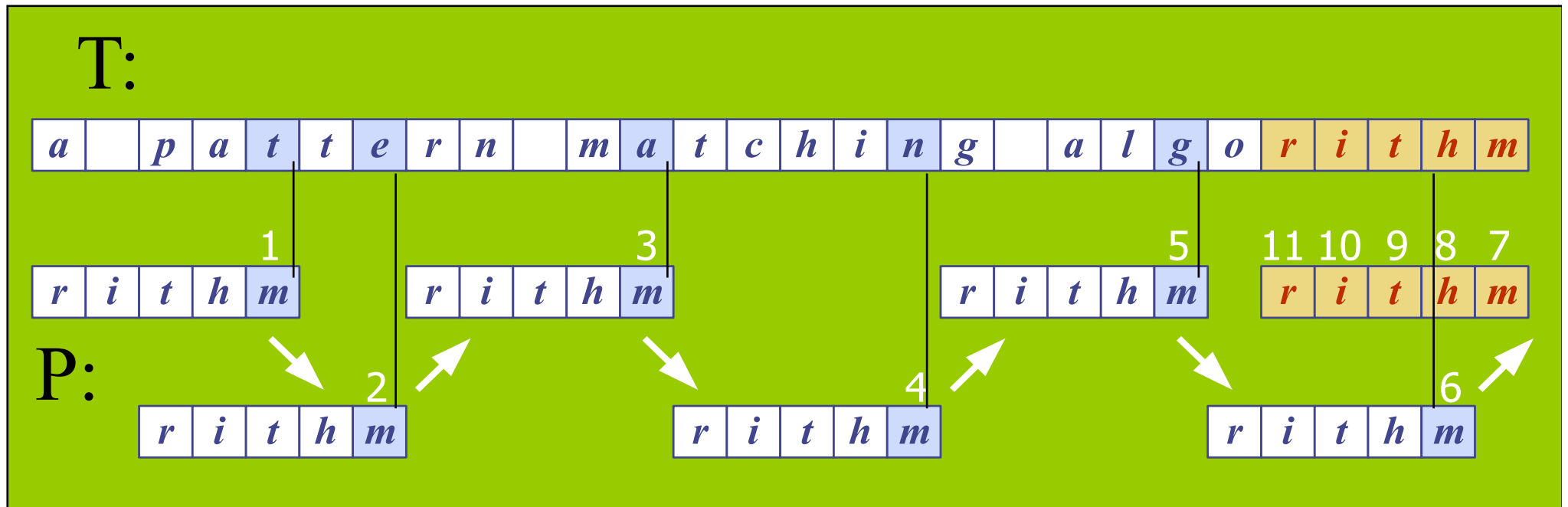


und  
bewege  $i$  und  
 $j$  nach rechts, so  
dass  $j$  am Ende



Kein  $x$  in  $P$

# Beispiel (Boyer-Moore)





# Funktion des letzten Vorkommens

---

- Der Boyer-Moore Algorithmus verarbeitet das Muster  $P$  und das Alphabet  $A$  vor, so dass eine Funktion  $L$  des letzten Vorkommens berechnet wird
  - $L$  bildet alle Zeichen des Alphabets auf ganzzahlige Werte ab
- $L(x)$  (mit  $x$  ist Zeichen aus  $A$ ) ist definiert als
  - den größten Index  $i$ , so dass  $P[i]=x$ , oder
  - $-1$ , falls solch ein Index nicht existiert
- Implementierungen
  - $L$  ist meist in einem Feld der Größe  $|A|$  gespeichert

# Beispiel L

$A = \{a, b, c, d\}$   
 $P = \text{„abacab“}$

P

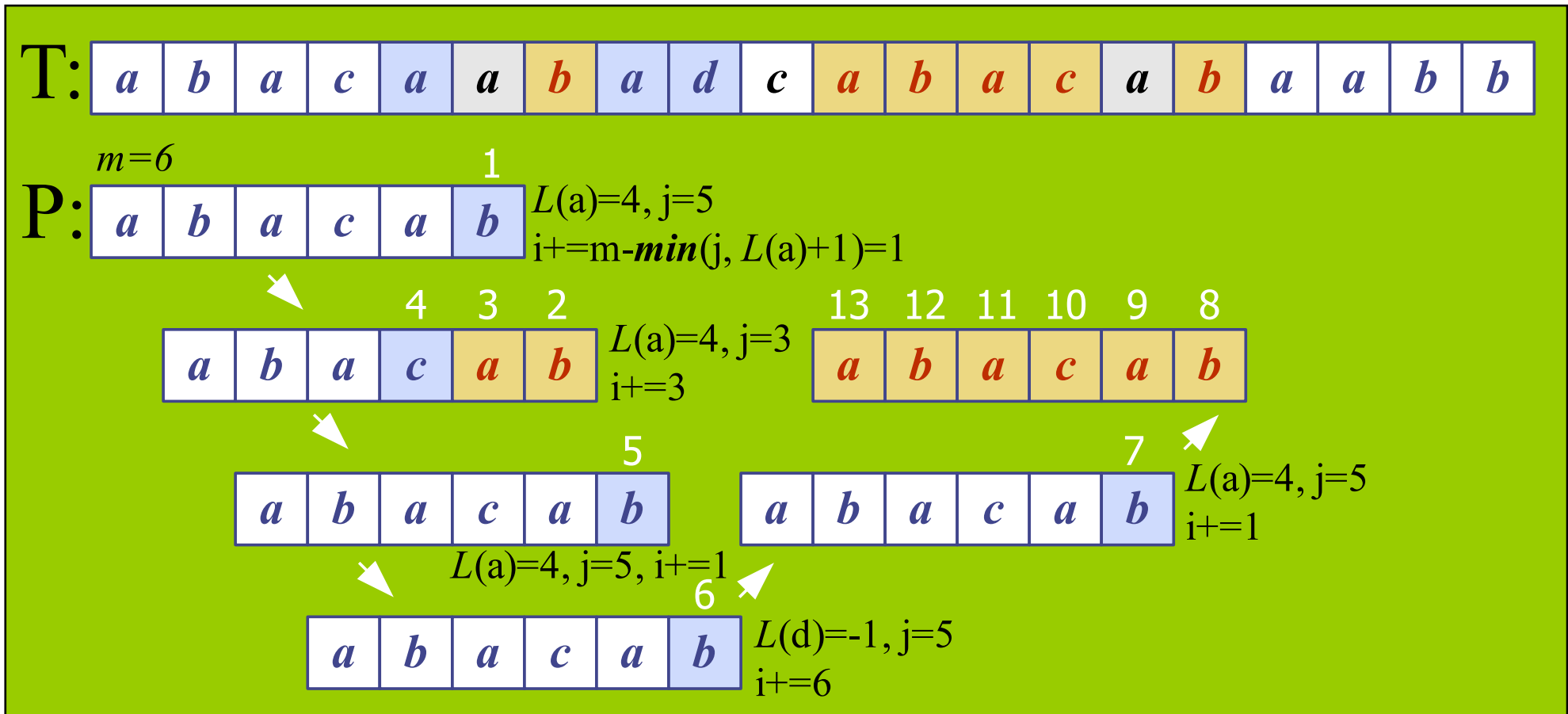
a	b	a	c	a	b
0	1	2	3	4	5



<b><i>x</i></b>	<i>a</i>	<i>b</i>	<i>c</i>	<i>d</i>
<b><i>L(x)</i></b>	4	5	3	-1

L speichert Indexe von P

# Zweites Boyer-Moore-Beispiel



$x$	$a$	$b$	$c$	$d$
$L(x)$	4	5	3	-1

# Funktion des letzten Vorkommens

---

```
function buildL(pattern: String): Array[0..127] of Integer
  L := <-1,...,-1>: Array [0..127] of Integer // nur ASCII unterstützt
  for i from 0 to length(pattern)-1 do
    L[pattern[i]] := i
  return L
```

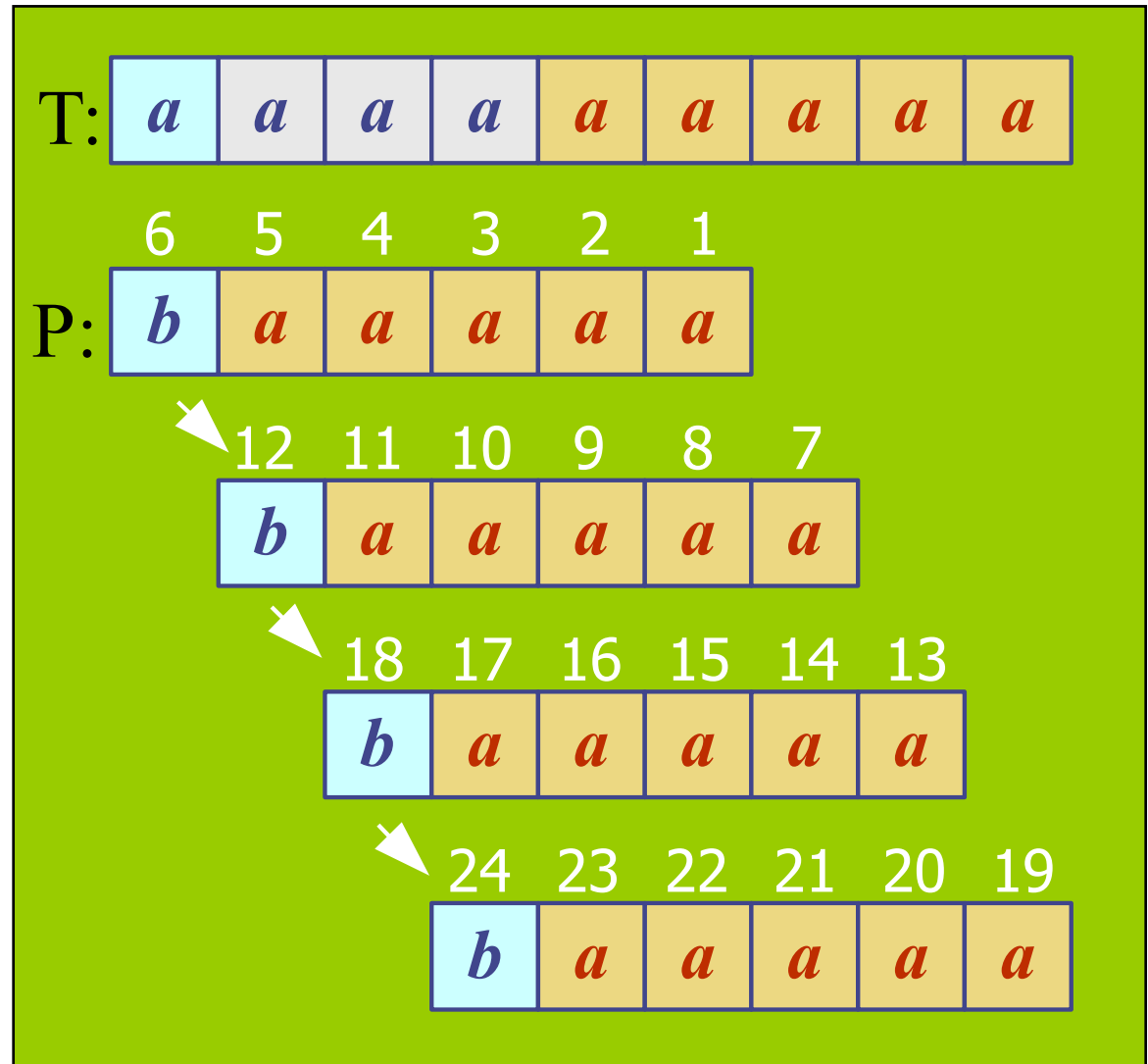
# Boyer-Moore-Verfahren

---

```
procedure BMsearch(text, pattern: String): Integer
  l := buildL(pattern)
  n := length(text); m := length(pattern)
  i := m - 1
  if i > n - 1 then
    return -1 // Muster länger als Text, Suche erfolglos
  j := m - 1
  repeat
    if pattern[j] = text[i] then
      if j = 0 then
        return i // Suche erfolgreich
      else // Spiegeltechnik
        i := i - 1; j := j - 1
      else // Zeichensprungtechnik
        lo := l[text[i]]; i := i + m - min(j, lo + 1); j := m - 1
  until j ≤ n - 1
  return -1 // Suche erfolglos
```

# Analyse der Komplexität

- Schlechtester Fall
  - Beispiel
    - T: „aaaaaaaaa...a“
    - P: „baa...a“
  - $O(m \cdot n + |A|)$
  - Wahrscheinlichkeit hoch für schlechtesten Fall bei kleinem Alphabet



# Analyse der Komplexität

---

- Bester Fall
  - Immer Fall 3, d.h. P wird jedes Mal um m nach rechts bewegt
  - Beispiel
    - T: „aaaaaaaa...a“
    - P: „bbb...b“
  - $O(n/m + |A|)$
  - Wahrscheinlichkeit für besten Fall höher bei großem Alphabet
- Durchschnittlicher Fall
  - nahe am besten Fall:  $O(n/m + |A|)$

# Rabin-Karp-Algorithmus

---

- Idee
  - Ermittle eine Hash-Signatur des Musters
  - Gehe durch den zu suchenden Text durch und vergleiche die jeweilige Hash-Signatur mit der des Musters
    - Mit geeigneten Hash-Funktionen ist es möglich, dass die Hash-Signatur iterativ mit konstantem Aufwand pro zu durchsuchenden Zeichen berechnet wird
    - Falls die Hash-Signaturen übereinstimmen, dann überprüfe noch einmal die Teilzeichenketten



# Hash-Funktion für Rabin-Karp-Algorithmus

---

- Für ein Zeichen
  - $h(k) = k \cdot q$  mit  $k$  z.B. ASCII-Code des betrachteten Zeichens und  $q$  eine Primzahl
- Für eine Zeichenkette
  - $h'(k_1..k_m) = h(k_1) + \dots + h(k_m)$
- Beispiel
  - $q = 5$  (in der Praxis sollte allerdings eine möglichst große Primzahl gewählt werden)
  - $A = \{1, 2, 3, 4\}$
  - der Einfachheit halber sei der Code des Zeichens  $i$  wiederum  $i$
- Berechnung der Hash-Signatur des Musters 1234:
  - $h'(1234) = 1 \cdot 5 + 2 \cdot 5 + 3 \cdot 5 + 4 \cdot 5 = 50$

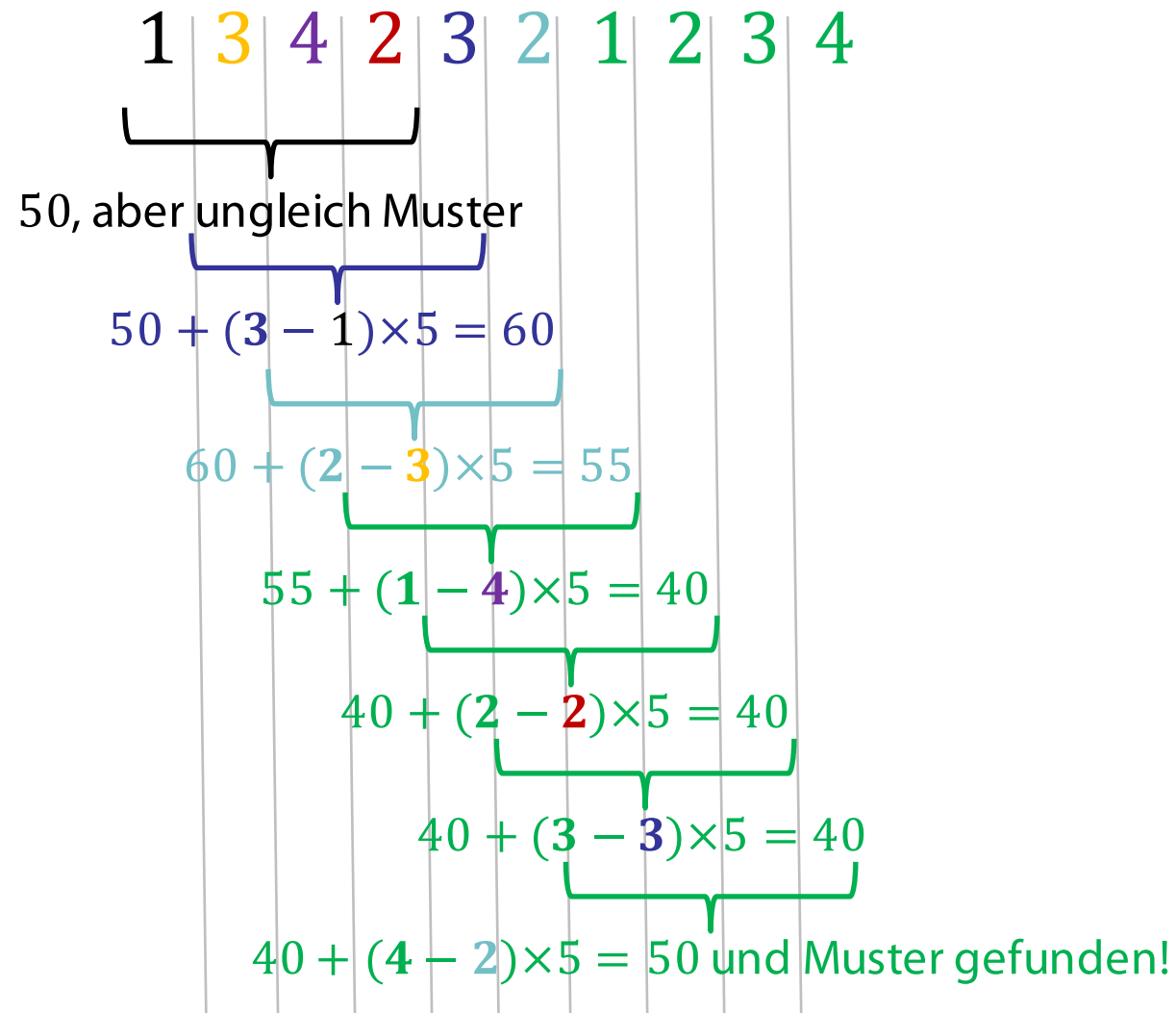
# Suche nach der Hash-Signatur

---

- Einmaliges Durchlaufen des zu durchsuchenden Textes und Vergleich der aktualisierten Hash-Signatur mit Hash-Signatur des Musters
- Hash-Signatur kann iterativ gebildet werden:  
$$h'(k_2..k_m k_{m+1}) = h'(k_1 k_2 \dots k_m) - k_1 \cdot q + k_{m+1} \cdot q$$
$$= h'(k_1 k_2 \dots k_m) + (k_{m+1} - k_1) \cdot q$$
- **Man beachte:** Konstanter Aufwand pro Zeichen

# Beispiel: Suche nach der Hash-Signatur

Musters 1234



# Komplexitätsanalyse Rabin-Karp-Algorithmus

---

- Ermittle die Hash-Signatur des Musters:  $O(m)$
- Gehe durch den zu suchenden Text durch und vergleiche die jeweilige Hash-Signatur mit der des Musters
  - Best (und Average) Case:  $O(n)$ 
    - Hash-Signaturen stimmen nur bei einem Treffer überein
  - Worst Case:  $O(n \cdot m)$ 
    - Hash-Signaturen stimmen immer überein, auch bei keinem Treffer
- Insgesamt  $O(n + m)$  im Best/Average Case und  $O(n \cdot m)$  im schlimmsten Fall

# Variante des Rabin-Karp-Algorithmus für Suche nach vielen Mustern

---

- Alle Hash-Signaturen (für Muster bis zu einer gegebenen maximalen Länge) können mit einem Durchlauf durch den Text in eine geeignete Datenstruktur z.B. eine Hashtabelle abgelegt werden
- Anschließende Suche nach vielen Mustern schnell möglich
  - im durchschnittlichen Fall in konstanter Zeit pro Muster
- **Aber:** Großer Speicherbedarf!

# Zusammenfassung

---

- Textsuche
  - Brute-Force
  - Knuth-Morris-Pratt
  - Boyer-Moore
  - Rabin-Karp



# Acknowledgements

---

- Präsentationen im nachfolgenden Teil sind entnommen aus dem Material zur Vorlesung Indexierung und Suchen von Tobias Scheffer, Univ. Potsdam

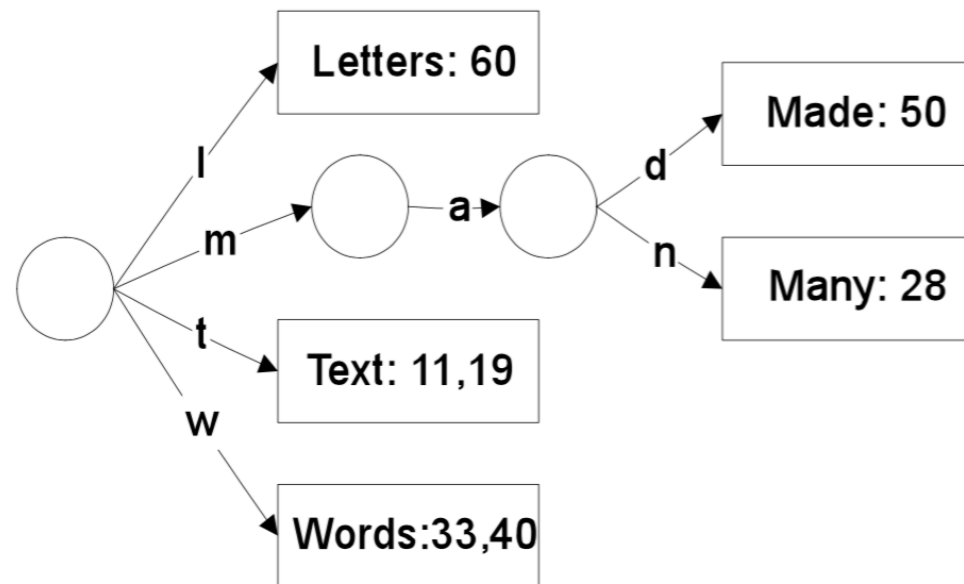


# Indexstrukturen für eindimensionale Daten

## Wiederholung: Tries

1 6 9 11 17 19 24 28 33 40 46 50 55 60

This is a text. A text has many words. Words are made from letters.





# Indexstrukturen für eindimensionale Daten

## Invertierter Index

1    6 9 11    17 19 24    28    33    40    46 50    55    60

This is a text. A text has many words. Words are made from letters.

Terme	Vorkommen
Letters	60
Made	50
Many	28
Text	11, 19
words	33, 40

Realisierung des Index:

Hashtabelle

# Invertierter Index mit Blockadressierung

1	2	3	4
This is a text.	A text has many	words. Words are	made from letters.

Terme	Vorkommen
Letters	4
Made	4
Many	2
Text	1, 2
words	3

Innerhalb eines Blocks  
kann dann ein  
Zeichenkettenabgleich  
realisiert werden  
(KMP, BM, RK, ...)

# Aufbau eines invertierten Index mit Tries

---

- Iteriere über alle Texte, iteriere über alle Positionen des Textes, an denen ein neues Wort beginnt.
  - Füge (Wort, Position) sortiert in einen Trie ein.
  - Wenn Speicher voll ist, dann speichere Trie, lade einen Unterbaum in den Speicher, berücksichtige nur Wörter, die in diesen Unterbaum gehören. Dann abspeichern, mit nächstem Unterbaum weitermachen, dazu Text neu iterieren.
- Traversiere den Trie und schreibe die Wörter in den invertierten Index.

# Suche mit invertiertem Index

---

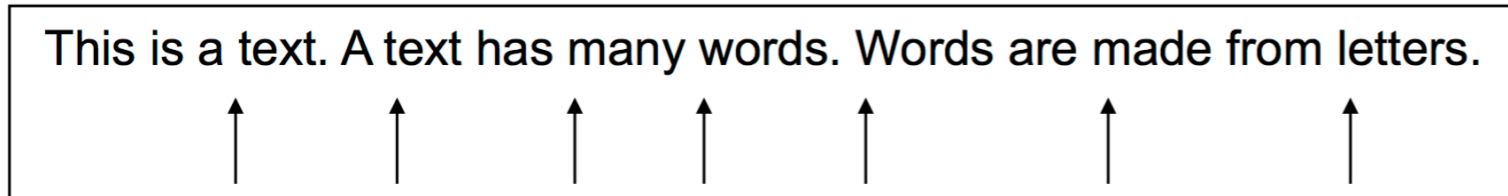
- Gegeben: Suchanfrage  $q$ 
  - Ausdruckstärke der Anfragesprache
- Schlage einzelne Terme der Suchanfrage im Index nach (Hashzugriff)
- Greife auf die gefundenen Dokumente zu
- Löse Phrasen-, Nachbarschafts- und Boolesche Anfrage anhand der Dokumente auf
- (Bei Blockadressierung, Suche in den Blöcken)
- Array der Terme (Hashtabelle) passt in der Regel in den Hauptspeicher, Listen der Vorkommen aller Terme häufig nicht

# Suffix-Bäume

---

- Indexpunkte können Wortanfänge oder alle Zeichenkettenpositionen sein.
- Text ab Position: Suffix.

This is a text. A text has many words. Words are made from letters.

A rectangular box containing the text "This is a text. A text has many words. Words are made from letters." Below the text, seven upward-pointing arrows are positioned at the start of the words "This", "is", "a", "text.", "A", "text", and "has".

Suffizes:

text. A text has many words. Words are made from letters.

text has many words. Words are made from letters.

many words. Words are made from letters.

words. Words are made from letters.

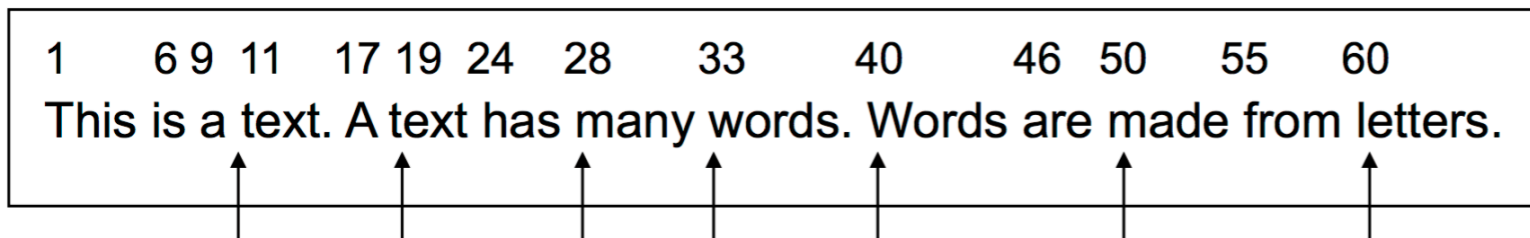
Words are made from letters.

made from letters.

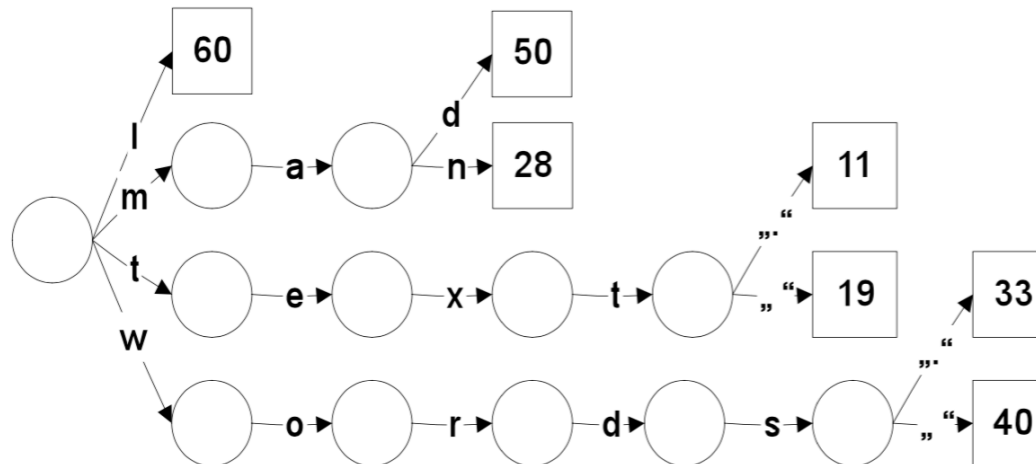
letters.

# Suffix-Tries

- Aufbau eines Suffix-Tries:
- Für alle Indexpunkte:
  - Füge Suffix ab Indexpunkt in den Trie ein.



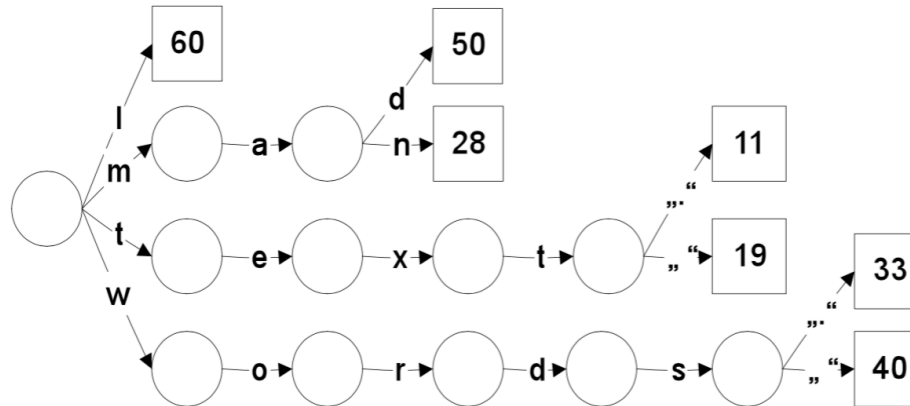
Suffix-Trie:



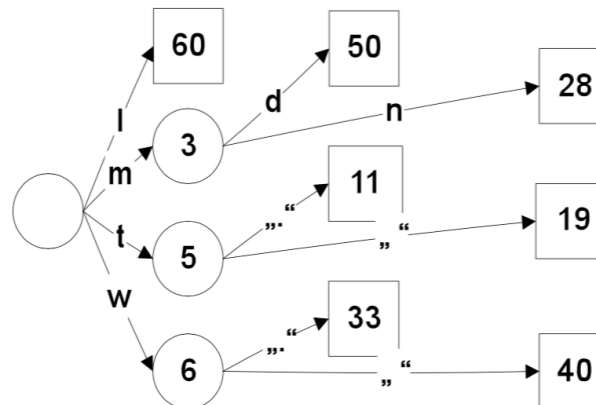
# Patricia-Tries

- Ersetze interne Knoten mit nur einer ausgehenden Kante durch „Überleseknoten“, beschrifte sie mit der nächsten zu beachtenden Textposition.

Suffix-Trie:



Patricia-Tree:



# Suche im Suffix-Baum

---

Eingabe: Suchzeichenkette, Wurzelknoten.

Wiederhole

1. Wenn Terminalknoten, liefere Position zurück, überprüfe, ob Suchzeichenkette an dieser Position steht.
2. Wenn „Überleseknoten“, spring bis zur angegebenen Textposition weiter.
3. Folge der Kante, die den Buchstaben an der aktuellen Position akzeptiert.



# Suffix-Bäume

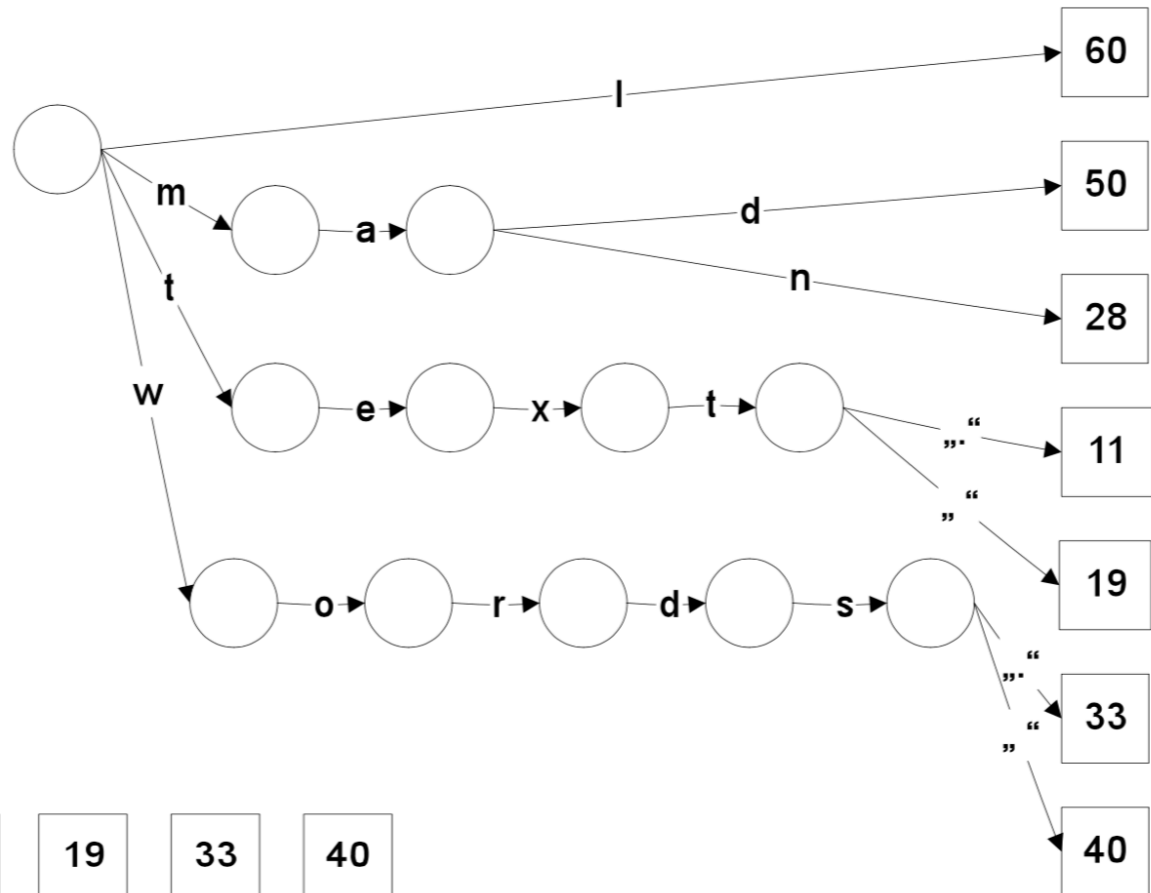
---

- Konstruktion:  $O(\text{Länge des Textes})$ .
- Algorithmus funktioniert schlecht, wenn die Struktur nicht in den Hauptspeicher passt.
- Problem: Speicherstruktur wird ziemlich groß, ca. 120-240% der Textsammlung, selbst wenn nur Wortanfänge (Längenbegrenzung) indexiert werden.
- Suffix-Felder (Arrays): kompaktere Speicherung.

# Suffix-Felder (Aufbau naiv)

- Suffix-Trie in lexikographische Reihenfolge bringen.
- Suffix-Feld= Folge der Indexpositionen.

Suffix-Trie,  
Lexikographisch  
sortiert:



Suffix-Array:



# Suche in Suffix-Feldern: Binärsuche

---

```
procedure find(p, A): Integer
  // Suchzeichenkette p, Suffix-Feld A
  // liefert Position von p in A oder -1, falls p nicht in A
  s := 1
  e := length(A)
  while s < e do
    m :=  $\lceil (e + s) / 2 \rceil$  // nach oben runden
    if p = Text(A[m]) then
      return m
    if p < Text(A[m])
    then e := m
    else s := m
  return -1
```

# Beispiel

Suffixsortierung → Binäre Suche

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

length(SA) = 11

Beispiel: Suche „sip“ in „mississippi“

SA	L	7	4	1	0	M	8	6	3	5	R
----	---	---	---	---	---	---	---	---	---	---	---

i

i

i

i

m

p

p

s

s

s

s

p

s

s

i

i

p

i

i

s

s

p

s

s

s

i

p

s

i

i

i

i

i

s

p

s

p

s

p

s

i

i

i

p

s

p

s

s

p

i

i

i

i

s

p

p

sip < ssi

p

i

i

p

sip > i

p

p

i

sip > pi

i

p

i

Suche: sip

SA	10	7	4	1	0	L	8	6	M	5	R
----	----	---	---	---	---	---	---	---	---	---	---

i	i	i	i	m	p	p	s	s	s	s
	p	s	s	i	i	p	i	i	s	s
	p	s	s	s		i	p	s	i	i
	i	i	i	s			p	s	p	s
		p	s	i			i	i	p	s
		p	s	s				p	i	i
		i	i	s				p		p
			p	i				i		p
			p	p						i
			i	p						
				i						

sip < sis

SA	10	7	4	1	0	L	8	M	R	5	2
----	----	---	---	---	---	---	---	---	---	---	---

i	i	i	i	m	p	p	s	s	s	s
	p	s	s	i	i	p	i	i	s	s
	p	s	s	s		i	p	s	i	i
	i	i	i	s			p	s	p	s
		p	s	i			i	i	p	s
		p	s	s				p	i	i
		i	i	s				p		p
			p	i				i		p
			p	p						i
			i	p						
				i						



1 Vorkommen

Mehrere Vorkommen?

Vorkommen benachbart!

sip = sip

SA	10	7	4	1	0	9	8	6	L	M	R
----	----	---	---	---	---	---	---	---	---	---	---

i	i	i	i	m	p	p	s	s	s	s
	p	s	s	i	i	p	i	i	s	s
	p	s	s	s		i	p	s	i	i
	i	i	i	s			p	s	p	s
		p	s	i			i	i	p	s
		p	s	s				p	i	i
		i	i	s				p		p
			p	i				i		p
			p	p						i
			i	p						
				i						

Suche nach ssi

2 Vorkommen



# Suche: Analyse der Aufwände

---

- Suffix-Bäume:  $O(\text{Länge der Suchzeichenkette})$
- Suffix-Felder:  $O(\log(\text{Länge der Textsammlung}))$

# Aufbau von Suffix-Feldern

---

- Ukkonens Algorithmus (1995, Hauptspeicher)
- Später: Datenbank-basierte Verfahren (ab 2001)  
z.B. für Bioinformatik-Anwendungen

Weiner, Peter (1973). Linear pattern matching algorithms, 14th Annual Symposium on Switching and Automata Theory, pp. 1–11, **1973**

McCreight, Edward Meyers, A Space-Economical Suffix Tree Construction Algorithm". Journal of the ACM 23 (2): 262–272, **1976**

Ukkonen, E., On-line construction of suffix trees, Algorithmica 14 (3): 249–260, **1995**

Hunt, E., Atkinson, M. and Irving, R. W. "A Database Index to Large Biological Sequences". VLDB **2001**

Tata, Hankins, Patel: „Practical Suffix Tree Construction“, VLDB **2004**



# Weitere Abgleichsalgorithmen

---

- Eigenes, umfangreiches Thema. → Bioinformatik.
- Aho-Corasick-Trie, Idee: mehrere Suchzeichenketten
  - Suchzeichenketten werden in Automaten umgebaut,
  - Kanten akzeptieren Buchstaben,
  - Wenn keine passende Kante mehr existiert, Sprung über „failure transition“ in Zustand, der der größten Übereinstimmung zwischen Text und einem Prefix eines Suchzeichenketten entspricht.
- Suffix-Automaten

# Approximativer Zeichenkettenabgleich

- Suche in  $y$  Unterzeichenketten, die Edit-Abstand von höchstens  $k$  von  $x$  haben.
- Dynamische Programmierung

Berechnung von  $C[0..m, 0..n]$ ;  $C[i,j]$  = minimale # Fehler beim Abgleich von  $x[1..i]$  mit  $y[1..j]$

$$C[0, j] = j$$

$$C[i, 0] = i$$

$$C[i, j] = \begin{cases} C[i-1, j-1] , & \text{wenn } x[i] = y[j] \\ 1 + \min \{C[i-1, j], C[i, j-1], C[i-1, j-1]\} , & \text{sonst} \end{cases}$$

$$O(nm)$$







# Beispiel

---





- $C[0, j] = j$
- $C[i, 0] = i$
- $C[i, j] = \begin{cases} C[i-1, j-1], & \text{wenn } x[i] = y[j] \\ 1 + \min \{C[i-1, j], C[i, j-1], C[i-1, j-1]\}, & \text{sonst} \end{cases}$

		<b>a</b>	<b>b</b>	<b>c</b>	<b>a</b>	<b>b</b>	<b>b</b>	<b>b</b>	<b>a</b>	<b>a</b>
	<b>0</b>	<b>1</b>	<b>2</b>	<b>3</b>	<b>4</b>	<b>5</b>	<b>6</b>	<b>7</b>	<b>8</b>	<b>9</b>
<b>c</b>	<b>1</b>	<b>1</b>	<b>2</b>	<b>2</b>	<b>3</b>	<b>4</b>	<b>5</b>	<b>6</b>	<b>7</b>	<b>8</b>
<b>b</b>	<b>2</b>	<b>2</b>	<b>1</b>	<b>2</b>	<b>3</b>	<b>3</b>	<b>4</b>	<b>5</b>	<b>6</b>	<b>7</b>
<b>a</b>	<b>3</b>	<b>2</b>	<b>2</b>	<b>2</b>	<b>2</b>	<b>3</b>	<b>4</b>	<b>5</b>	<b>5</b>	<b>6</b>
<b>b</b>	<b>4</b>	<b>3</b>	<b>2</b>	<b>3</b>	<b>3</b>	<b>2</b>	<b>3</b>	<b>4</b>	<b>5</b>	<b>6</b>
<b>a</b>	<b>5</b>	<b>4</b>	<b>3</b>	<b>3</b>	<b>3</b>	<b>3</b>	<b>3</b>	<b>4</b>	<b>4</b>	<b>5</b>
<b>c</b>	<b>6</b>	<b>5</b>	<b>4</b>	<b>3</b>	<b>4</b>	<b>4</b>	<b>4</b>	<b>4</b>	<b>5</b>	<b>5</b>



# Beispiel

		a	b	c	a	b	b	b	a	a
	0	1	2	3	4	5	6	7	8	9
<b>c</b>	1	1	2	2	3	4	5	6	7	8
<b>b</b>	2	2	1 =	2	3	3	4	5	6	7
<b>a</b>	3	2	2	2	2 =	3	4	5	5	6
<b>b</b>	4	3	2	3	3	2	3 =	4	5	6
<b>a</b>	5	4	3	3	3	3	3	4	4 =	5
<b>c</b>	6	5	4	3	4	4	4	4	5	5

Löschen

Einfügen

Substitution

Keine Änderung

cbabac -> ababac -> abcabac -> abcabbac ->  
 abcabbac -> abcabbbaa

# Suffix-Bäume und Suffix-Felder

---

- Zeichenkettensuche mit Suffix-Bäumen und –Feldern kein Problem, aber jede Zeichenkettenposition muss indexiert werden. Speicherbedarf empirisch 1200-2400% der Textsammlung.
- Aktuelle Themen in der Forschung: Suche in komprimiertem Text, Suche mit komprimierten Indexdateien