http://algs4.cs.princeton.edu

5.1 STRING SORTS

- *strings in Java*
- *key-indexed counting*
- *LSD radix sort*
- *MSD radix sort*
- *3-way radix quicksort*
- *suffix arrays*

Last updated on 4/25/16 2:27 PM

http://algs4.cs.princeton.edu

5.1 STRING SORTS

- *strings in Java*
- *key-indexed counting*
- *LSD radix sort*
- *MSD radix sort*
- *3-way radix quicksort*
- *suffix arrays*

String processing

String. Sequence of characters.

Important fundamental abstraction.

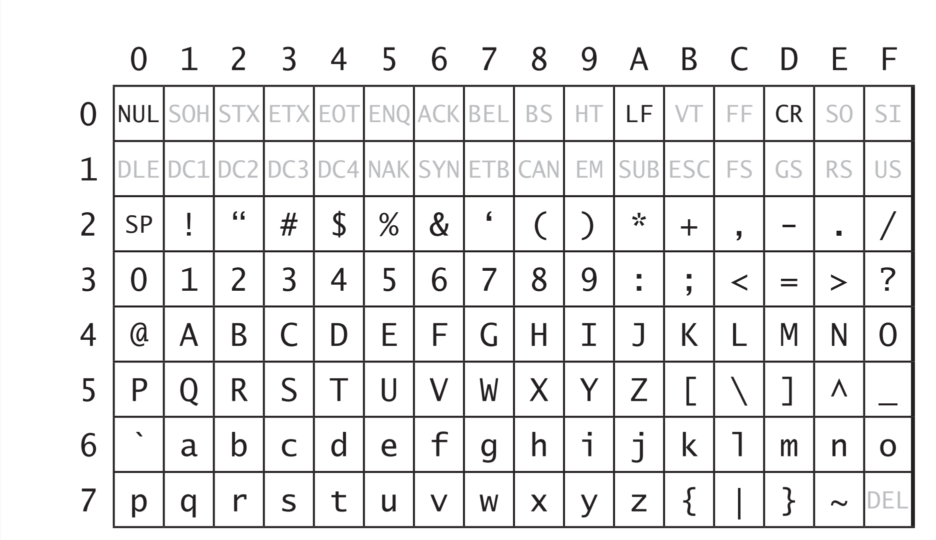
- Programming systems (e.g., Java programs).
- Communication systems (e.g., email).
- Information processing.
- Genomic sequences.
- ...

> *" The digital information that underlies biochemistry, cell biology, and development can be represented by a simple string of G's, A's, T's and C's. This string is the root data structure of an organism's biology. "* — *M. V. Olson*

The char data type

C char data type. Typically an 8-bit integer.

- Supports 7-bit ASCII.
- Can represent at most 256 characters.



	0	1	2	3	4	5	6	7	8	9	A	B	C	D	E	F
0	NUL	SOH	STX	ETX	EOT	ENQ	ACK	BEL	BS	HT	LF	VT	FF	CR	SO	SI
1	DLE	DC1	DC2	DC3	DC4	NAK	SYN	ETB	CAN	EM	SUB	ESC	FS	GS	RS	US
2	SP	!	“	#	$	%	&	‘	(	)	*	+	,	-	.	/
3	0	1	2	3	4	5	6	7	8	9	:	;	<	=	>	?
4	@	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O
5	P	Q	R	S	T	U	V	W	X	Y	Z	[	\	]	^	_
6	`	a	b	c	d	e	f	g	h	i	j	k	l	m	n	o
7	p	q	r	s	t	u	v	w	x	y	z	{	|	}	~	DEL

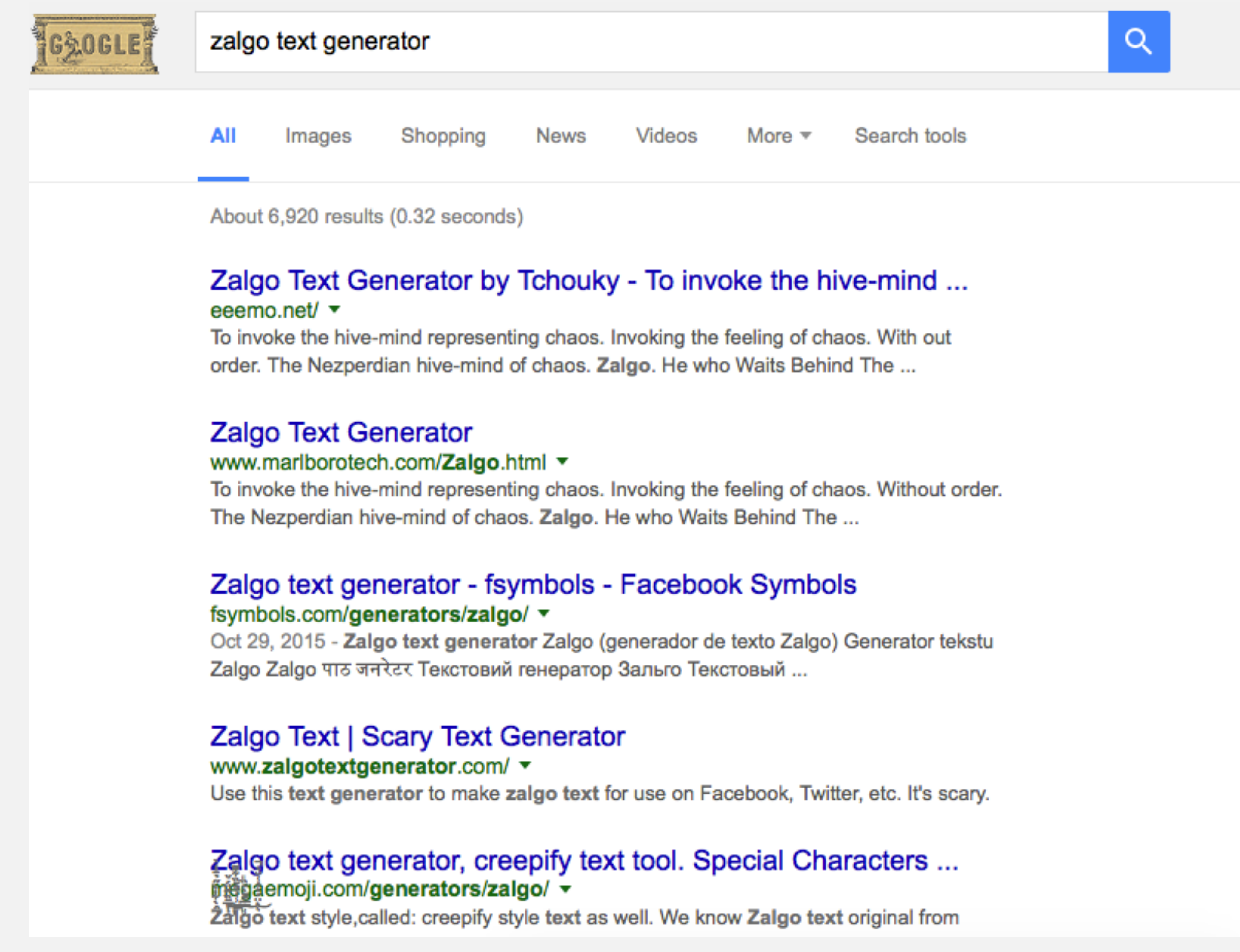
Hexadecimal to ASCII conversion table

A	á	∂	𝔊
U+0041	U+00E1	U+2202	U+1D50A

some Unicode characters

Java char data type. A 16-bit unsigned integer.

- Supports original 16-bit Unicode.
- Supports 21-bit Unicode 3.0 (awkwardly).



About 6,920 results (0.32 seconds)

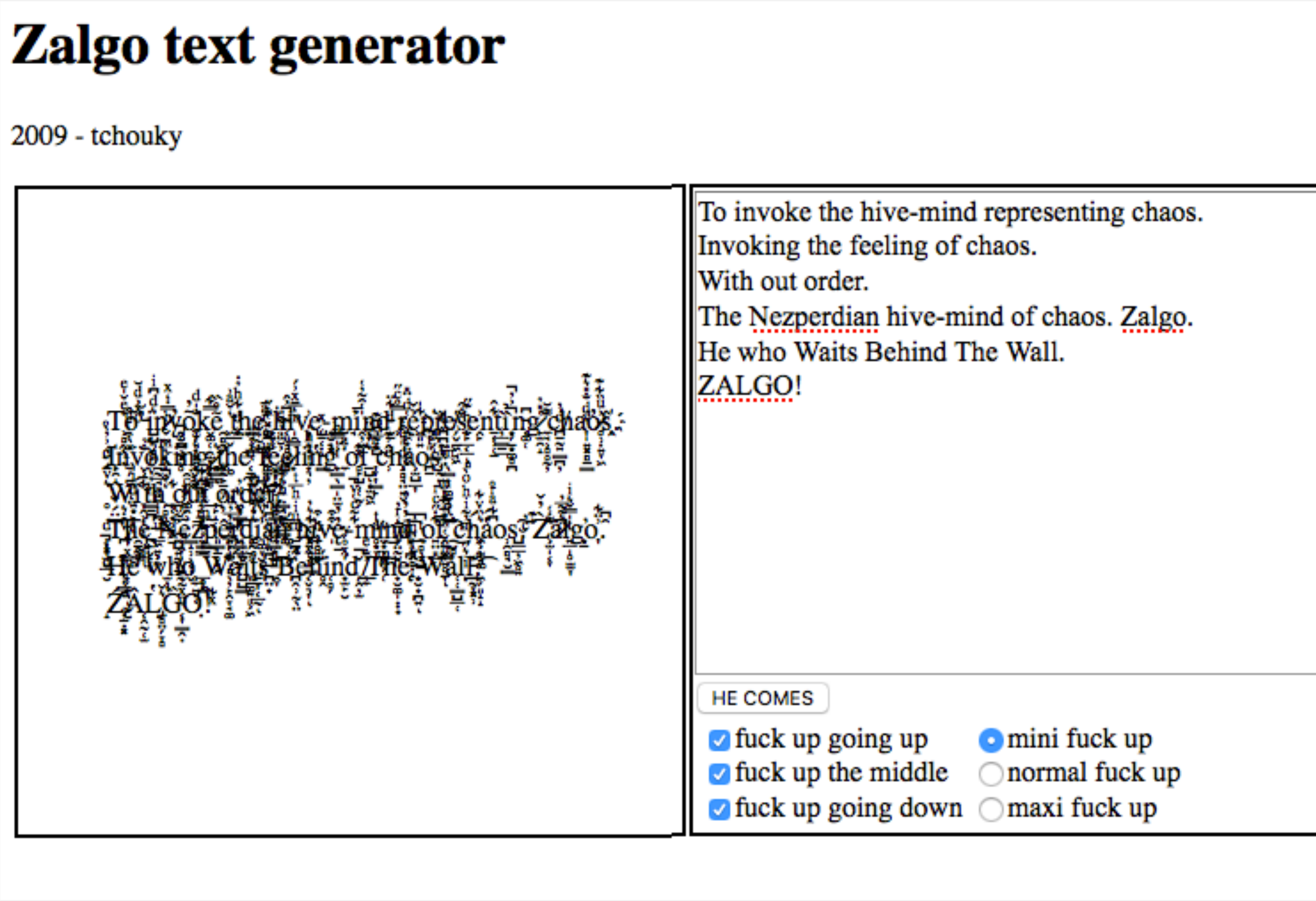
Zalgo Text Generator by Tchouky - To invoke the hive-mind ...
eeemo.net/
To invoke the hive-mind representing chaos. Invoking the feeling of chaos. With out order. The Nezperdian hive-mind of chaos. **Zalgo**. He who Waits Behind The ...

Zalgo Text Generator
www.marlborotech.com/**Zalgo**.html
To invoke the hive-mind representing chaos. Invoking the feeling of chaos. Without order. The Nezperdian hive-mind of chaos. **Zalgo**. He who Waits Behind The ...

Zalgo text generator - fsymbols - Facebook Symbols
fsymbols.com/**generators/zalgo**/
Oct 29, 2015 - **Zalgo text generator** Zalgo (generador de texto Zalgo) Generator tekstu Zalgo Zalgo पाठ जनरेटर Текстовий генератор Зальго Текстовый ...

Zalgo Text | Scary Text Generator
www.**zalgotextgenerator**.com/
Use this **text generator** to make **zalgo text** for use on Facebook, Twitter, etc. It's scary.

Zalgo text generator, creepify text tool. Special Characters ...
megaemoji.com/**generators/zalgo**/
Zalgo text style,called: creepify style **text** as well. We know **Zalgo text** original from



Zalgo text generator

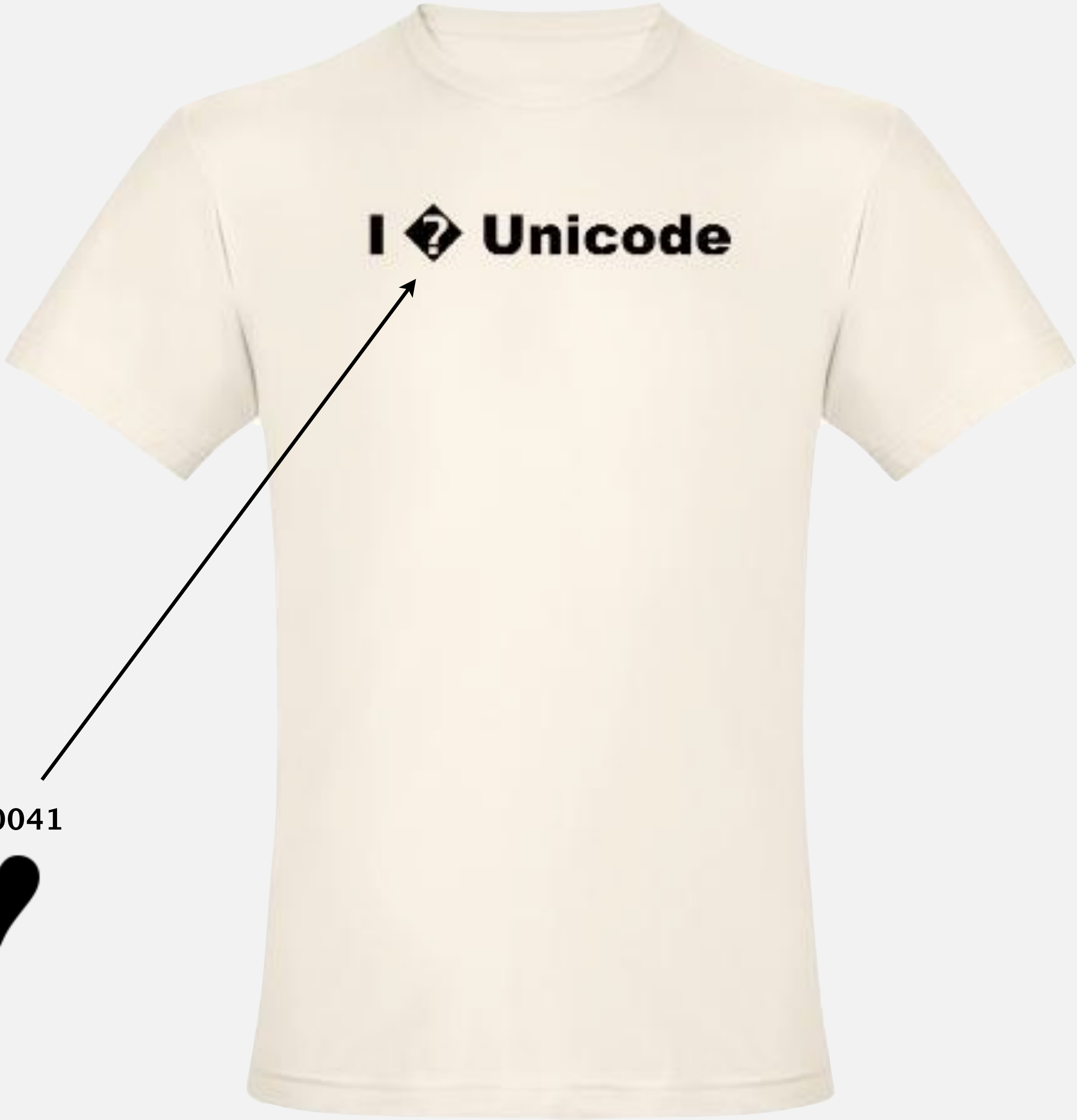
2009 - tchouky

To invoke the hive-mind representing chaos.
Invoking the feeling of chaos.
With out order.
The Nezperdian hive-mind of chaos. Zalgo.
He who Waits Behind The Wall.
ZALGO!

☑ fuck up going up
☑ fuck up the middle
☑ fuck up going down

● mini fuck up
○ normal fuck up
○ maxi fuck up

I ♥ Unicode



U+0041

♥

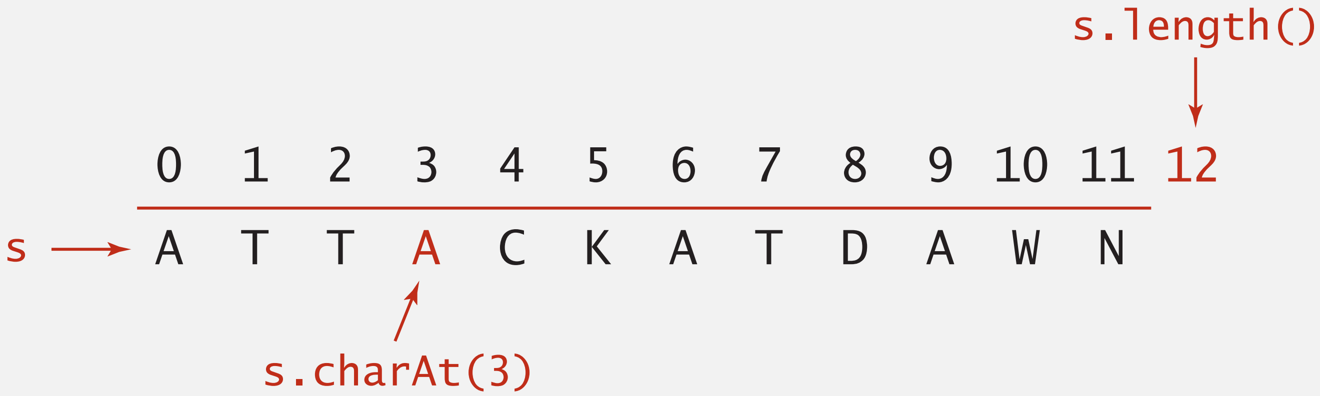
The String data type

String data type in Java. Immutable sequence of characters.

Length. Number of characters.

Indexing. Get the i^{th} character.

Concatenation. Concatenate one string to the end of another.



THE STRING DATA TYPE: IMMUTABILITY

Q. Why are Java strings immutable?

A. All the usual reasons.

- Provides security.
- Ensures consistent state.
- Can use as keys in symbol table.
- Removes need to defensively copy.
- Supports concurrency / thread safety.
- Simplifies tracing and debugging code.
- Enables compiler to perform certain optimizations.
- ...

Immutable strings. Java, C#, Python, Scala, ...

Mutable strings. C, C++, Matlab, Ruby, ...

The String data type: representation

Representation (Java 7). Immutable `char[]` array + cache of hash.

operation	Java	running time
length	`s.length()`	1
indexing	`s.charAt(i)`	1
concatenation	`s + t`	$M + N$
⋮	⋮	⋮

Q. Could concatenation be O(1)?

A. Yes, but charAt would no longer be.

String performance trap

Q. How to build a long string, one character at a time?

```
public static String reverse(String s)
{
   String reverse = "";
   for (int i = s.length() - 1; i >= 0; i--)
      rev += s.charAt(i);
   return reverse;
}
```

quadratic time
(1 + 2 + 3 + … + N)

A. Use `StringBuilder` data type (mutable `char[]` resizing array).

```
public static String reverse(String s)
{
   StringBuilder reverse = new StringBuilder();
   for (int i = s.length() - 1; i >= 0; i--)
      reverse.append(s.charAt(i));
   return reverse.toString();
}
```

linear time

Comparing two strings

Q. How many character compares to compare two strings, each of length W ?

s.compareTo(t)

s	p	r	e	f	e	t	c	h
	0	1	2	3	4	5	6	7
t	p	r	e	f	i	x	e	s

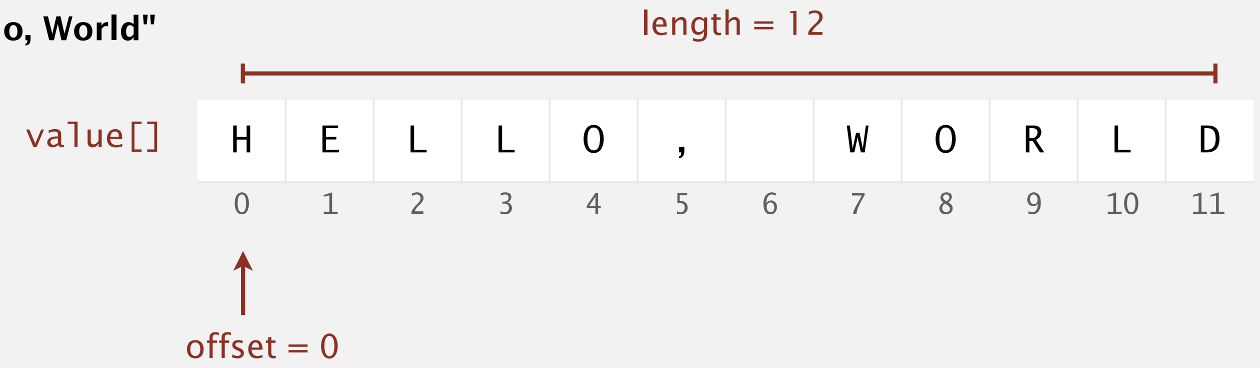
Running time. Proportional to length of longest common prefix.

- Proportional to W in the worst case.
- But, often sublinear in W.

The String data type: Java 7u5 implementation

```
public final class String implements Comparable<String>
{
   private char[] value;  // characters
   private int offset;    // index of first char in array
   private int length;    // length of string
   private int hash;      // cache of hashCode()
   …
```

String s = "Hello, World"



length = 12

value[]	H	E	L	L	O	,		W	O	R	L	D
	0	1	2	3	4	5	6	7	8	9	10	11

offset = 0

String t = s.substring(7, 12);

length = 5

value[]	H	E	L	L	O	,		W	O	R	L	D
	0	1	2	3	4	5	6	7	8	9	10	11

offset = 7

The String data type: Java 7u6 implementation

```
public final class String implements Comparable<String>
{
   private char[] value;   // characters
   private int hash;       // cache of hashCode()
   …
```

String s = "Hello, World"

value[]	H	E	L	L	O	,		W	O	R	L	D
	0	1	2	3	4	5	6	7	8	9	10	11

String t = s.substring(7, 12);

value[]	W	O	R	L	D
	0	1	2	3	4

The String data type: performance

String data type (in Java). Sequence of characters (immutable).

Java 7u5. Immutable `char[]` array, offset, length, hash cache.

Java 7u6. Immutable `char[]` array, hash cache.

operation	Java 7u5	Java 7u6
length	1	1
indexing	1	1
substring extraction	1	N
concatenation	$M + N$	$M + N$
immutable?	✔	✔
memory	$64 + 2N$	$56 + 2N$

A Reddit exchange

> I'm the author of the substring() change. As has been suggested in the analysis here there were two motivations for the change
>
> - Reduce the size of String instances. Strings are typically 20-40% of common apps footprint.
> - Avoid memory leakage caused by retained substrings holding the entire character array.

bondolo

> Changing this function, in a bugfix release no less, was totally irresponsible. It broke backwards compatibility for numerous applications with errors that didn't even produce a message, just freezing and timeouts... All pain, no gain. Your work was not just vain, it was thoroughly destructive, even beyond its immediate effect.

cypherpunks

http://www.reddit.com/r/programming/comments/1qw73v/til_oracle_changed_the_internal_string

Alphabets

Digital key. Sequence of digits over fixed alphabet.

Radix. Number of digits R in alphabet.

name	R()	lgR()	characters
BINARY	2	1	01
OCTAL	8	3	01234567
DECIMAL	10	4	0123456789
HEXADECIMAL	16	4	0123456789ABCDEF
DNA	4	2	ACTG
LOWERCASE	26	5	abcdefghijklmnopqrstuvwxyz
UPPERCASE	26	5	ABCDEFGHIJKLMNOPQRSTUVWXYZ
PROTEIN	20	5	ACDEFGHIKLMNPQRSTVWY
BASE64	64	6	ABCDEFGHIJKLMNOPQRSTUVWXYZabcdef ghijklmnopqrstuvwxyz0123456789+/
ASCII	128	7	*ASCII characters*
EXTENDED_ASCII	256	8	*extended ASCII characters*
UNICODE16	65536	16	*Unicode characters*


Algorithms

ROBERT SEDGEWICK | KEVIN WAYNE

http://algs4.cs.princeton.edu

5.1 STRING SORTS

- strings in Java
- **key-indexed counting**
- LSD radix sort
- MSD radix sort
- 3-way radix quicksort
- suffix arrays

Program optimization

" The first rule of program optimization: don't do it.
The second rule of program optimization (for experts only!):
don't do it yet. " – *Michael A. Jackson*

Review: summary of the performance of sorting algorithms

Frequency of operations.

algorithm	guarantee	random	extra space	stable?	operations on keys
insertion sort	½ N^2	¼ N^2	1	✔	`compareTo()`
mergesort	$N \lg N$	$N \lg N$	N	✔	`compareTo()`
quicksort	$1.39\, N \lg N$ *	$1.39\, N \lg N$	$c \lg N$ *		`compareTo()`
heapsort	$2\, N \lg N$	$2\, N \lg N$	1		`compareTo()`

* probabilistic

Lower bound. $\sim N \lg N$ compares required by any compare-based algorithm.

Q. Can we do better (despite the lower bound)?

A. Yes, if we don't depend on key compares. ←

use array accesses
to make R-way decisions
(instead of binary decisions)

Key-indexed counting: assumptions about keys

Assumption. Keys are integers between 0 and $R - 1$.

Implication. Can use key as an array index.

Applications.

- Sort string by first letter.
- Sort class roster by section.
- Sort phone numbers by area code.
- Subroutine in a sorting algorithm. [stay tuned]

Remark. Keys may have associated data ⇒ can't just count up number of keys of each value.

input

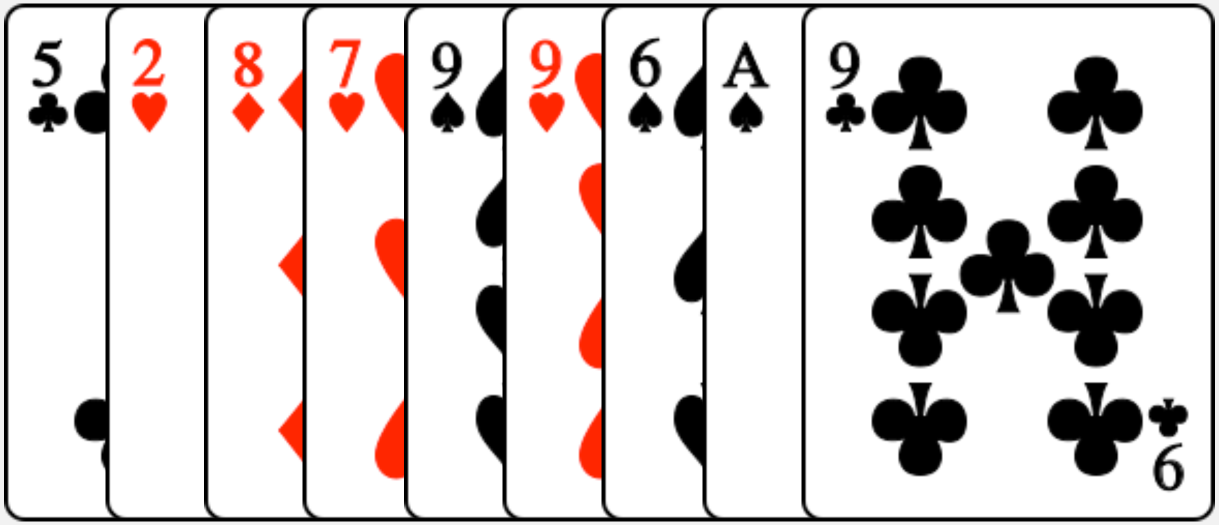
name	*section*
Anderson	2
Brown	3
Davis	3
Garcia	4
Harris	1
Jackson	3
Johnson	4
Jones	3
Martin	1
Martinez	2
Miller	2
Moore	1
Robinson	2
Smith	4
Taylor	3
Thomas	4
Thompson	4
White	2
Williams	3
Wilson	4

↑ *keys are small integers*

sorted result *(by section)*

Harris	1
Martin	1
Moore	1
Anderson	2
Martinez	2
Miller	2
Robinson	2
White	2
Brown	3
Davis	3
Jackson	3
Jones	3
Taylor	3
Williams	3
Garcia	4
Johnson	4
Smith	4
Thomas	4
Thompson	4
Wilson	4

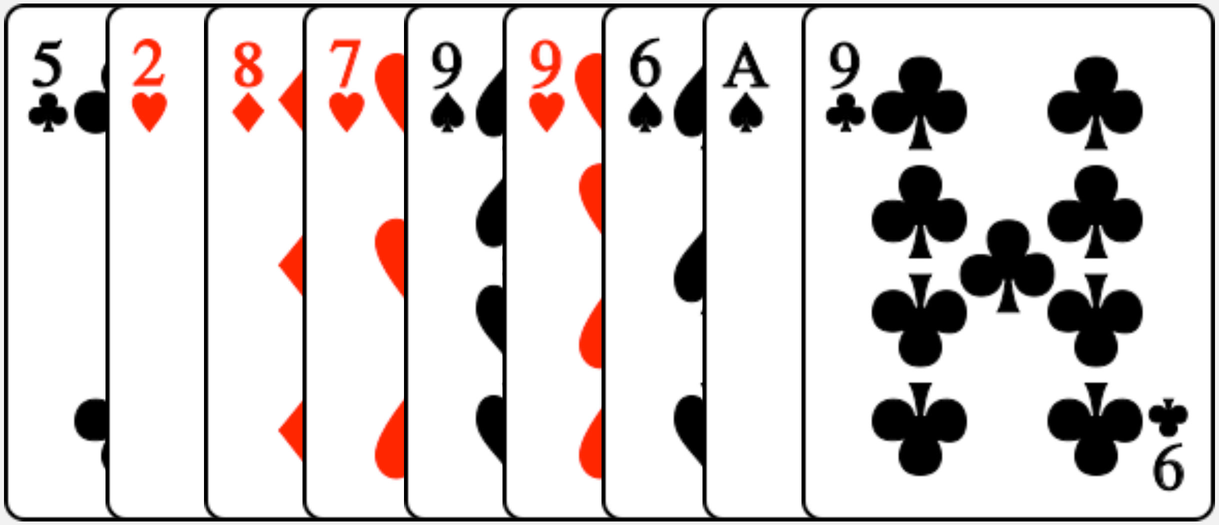
Stably sorting a set of cards by suit



We want clubs, then diamonds, hearts, spades

Stability: within each suit, same order as original

Stably sorting a set of cards by suit

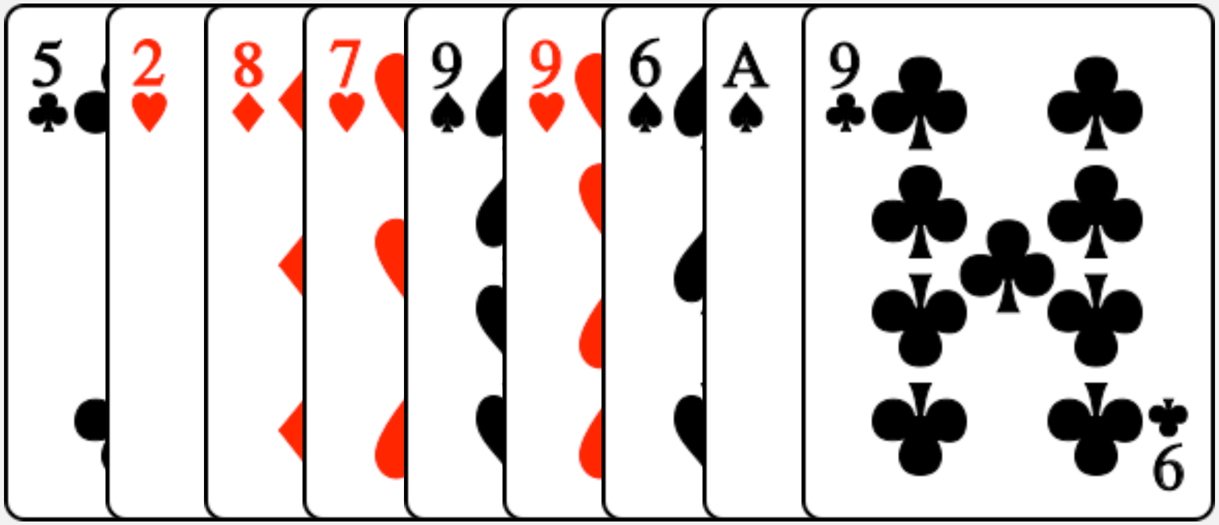


Suit	Count	Start pos
Clubs	2	0
Diamonds	1	2
Hearts	3	3
Spades	3	6

Count the number of cards in each suit

Use this to calculate starting position of each suit in sorted order

Stably sorting a set of cards by suit

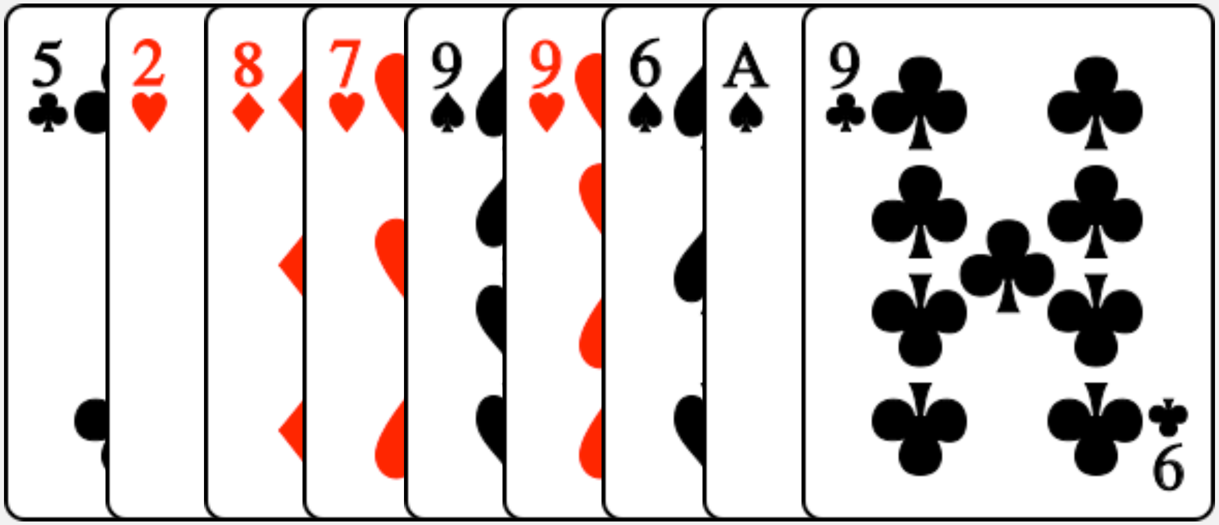


Suit	Count	Start pos
Clubs	2	0
Diamonds	1	2
Hearts	3	3
Spades	3	6

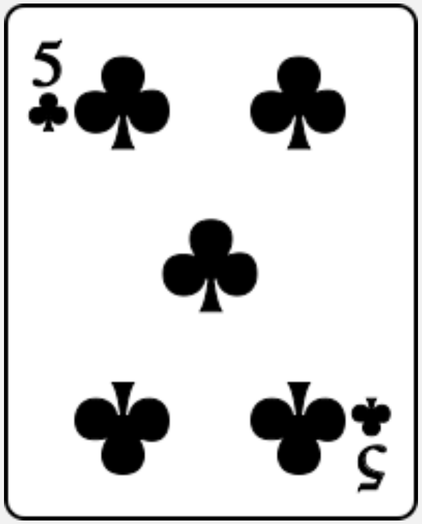
Move cards into final position one by one

Suit	Next pos
Clubs	0
Diamonds	2
Hearts	3
Spades	6

Stably sorting a set of cards by suit

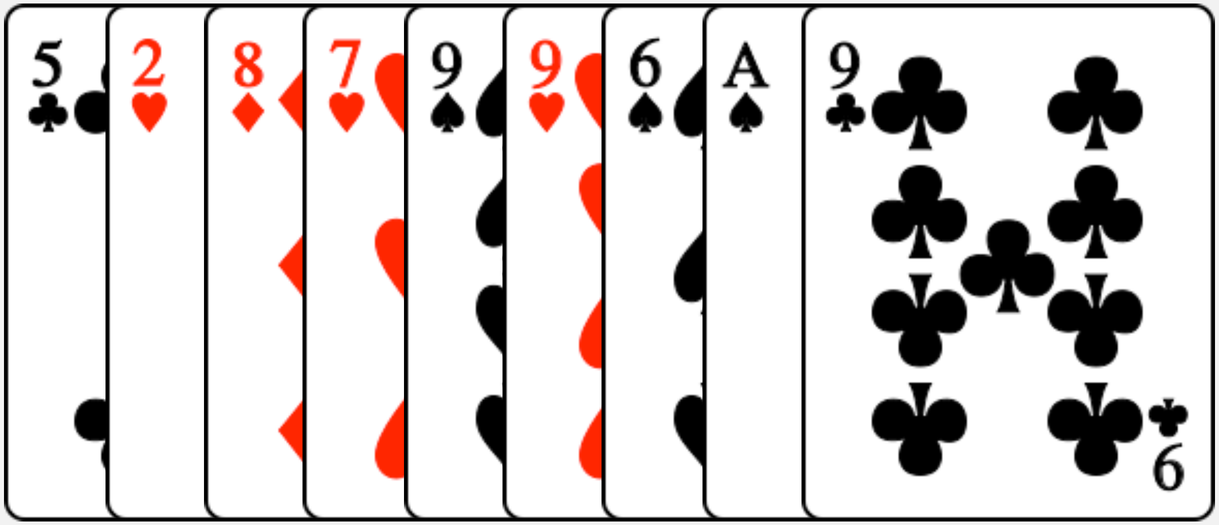


Suit	Count	Start pos
Clubs	2	0
Diamonds	1	2
Hearts	3	3
Spades	3	6

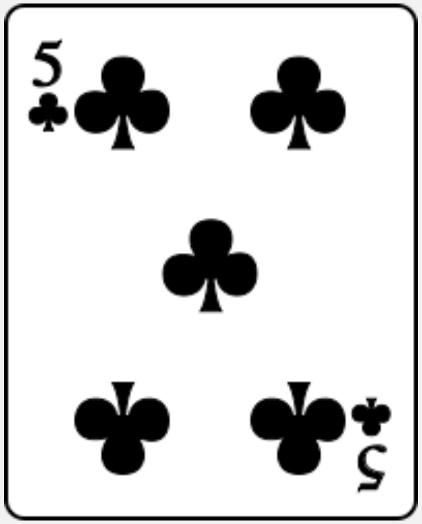


Suit	Next pos
Clubs	0
Diamonds	2
Hearts	3
Spades	6

Stably sorting a set of cards by suit

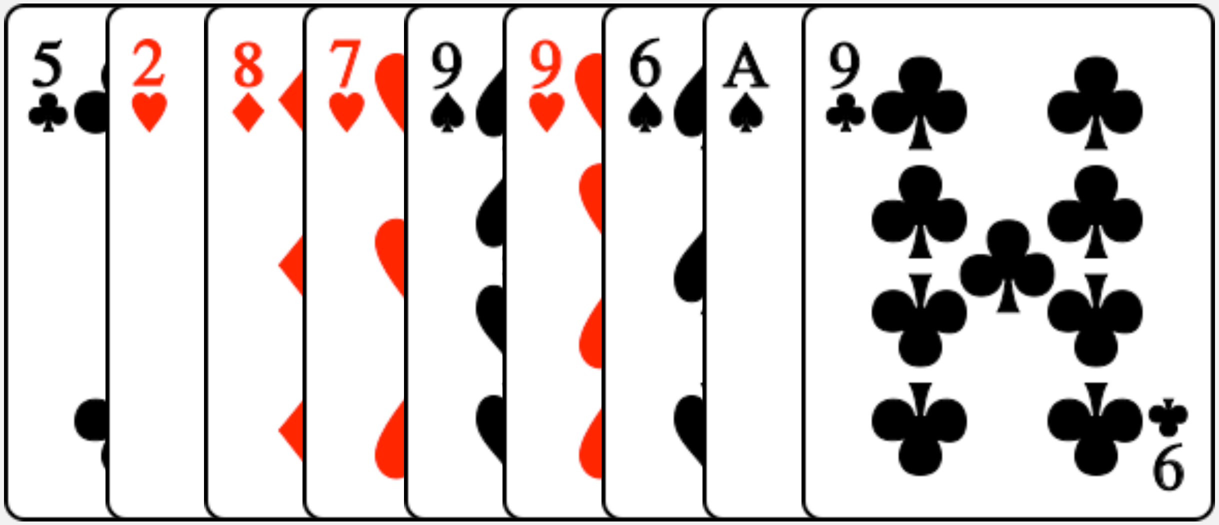


Suit	Count	Start pos
Clubs	2	0
Diamonds	1	2
Hearts	3	3
Spades	3	6

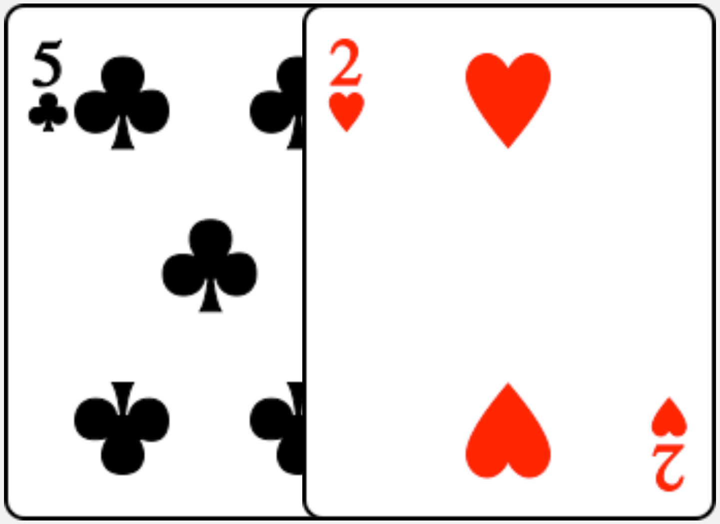


Suit	Next pos
Clubs	1
Diamonds	2
Hearts	3
Spades	6

Stably sorting a set of cards by suit

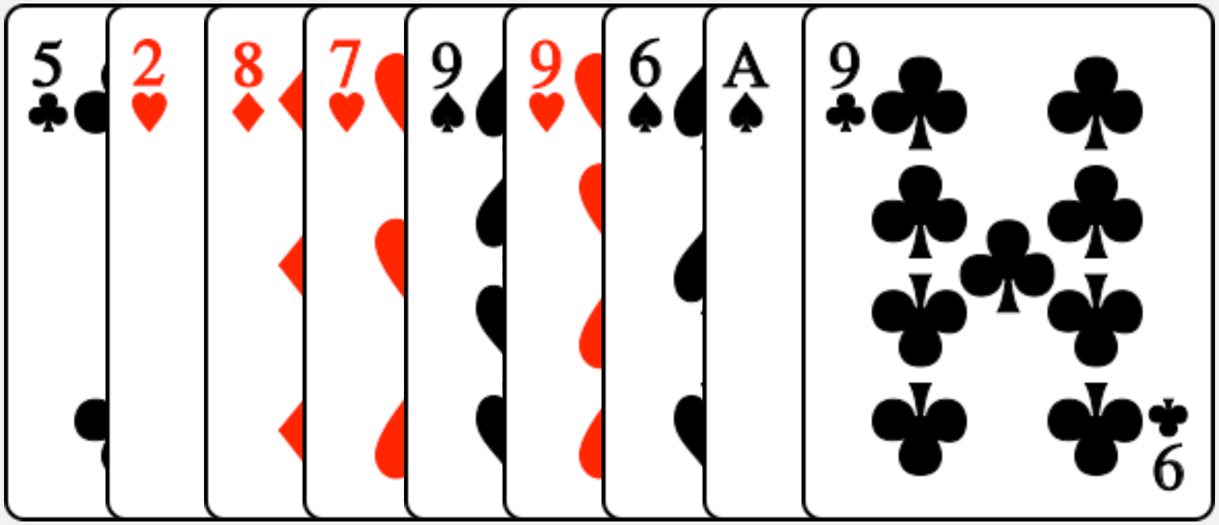


Suit	Count	Start pos
Clubs	2	0
Diamonds	1	2
Hearts	3	3
Spades	3	6

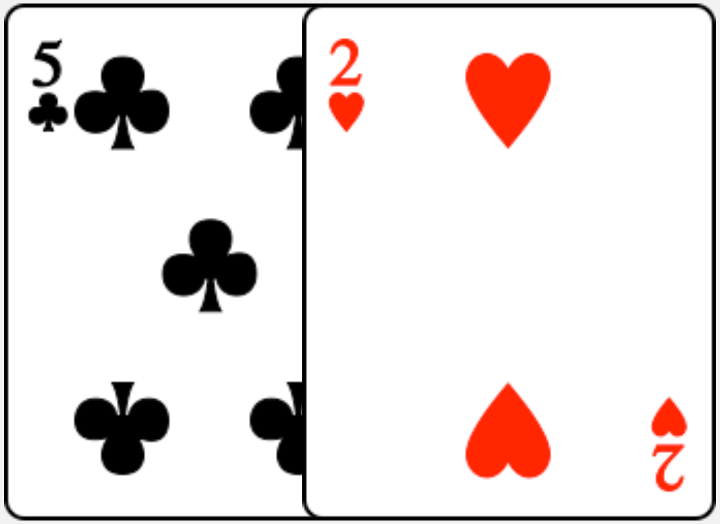


Suit	Next pos
Clubs	1
Diamonds	2
Hearts	3
Spades	6

Stably sorting a set of cards by suit

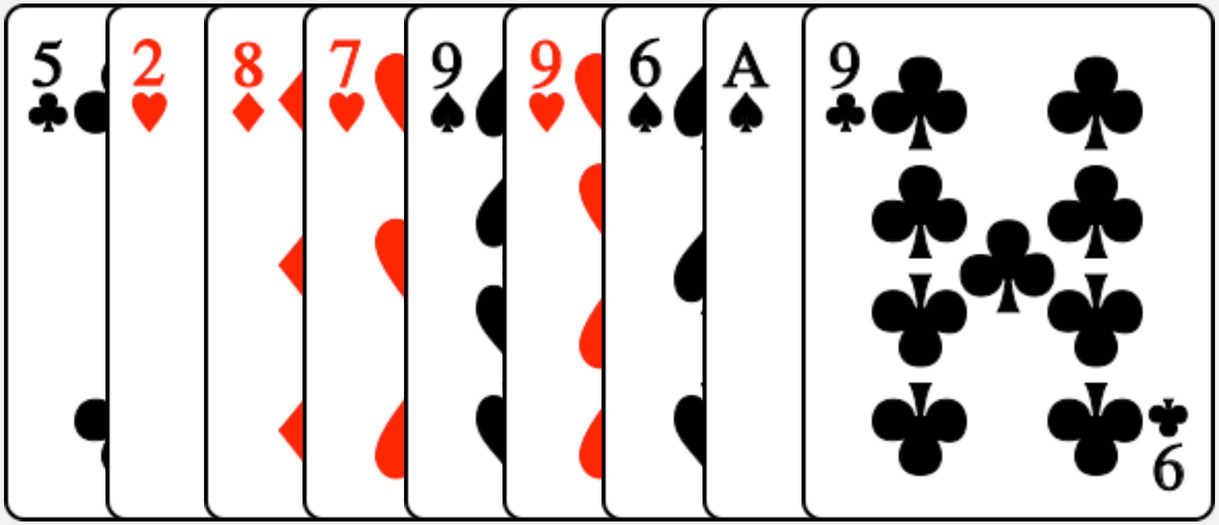


Suit	Count	Start pos
Clubs	2	0
Diamonds	1	2
Hearts	3	3
Spades	3	6

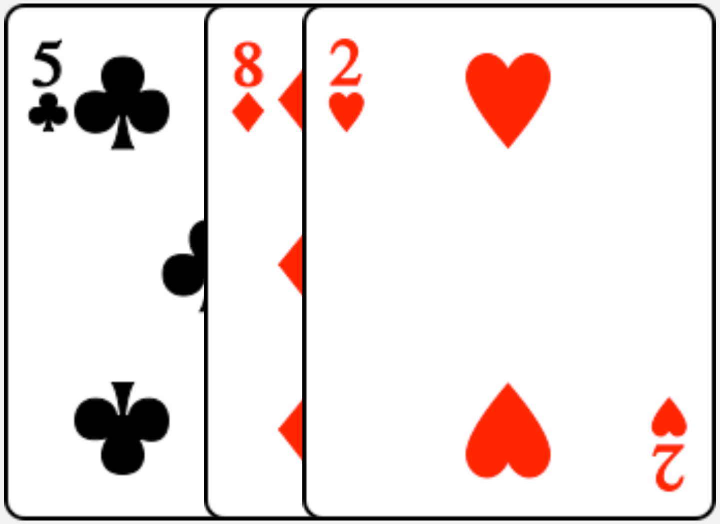


Suit	Next pos
Clubs	1
Diamonds	2
Hearts	4
Spades	6

Stably sorting a set of cards by suit

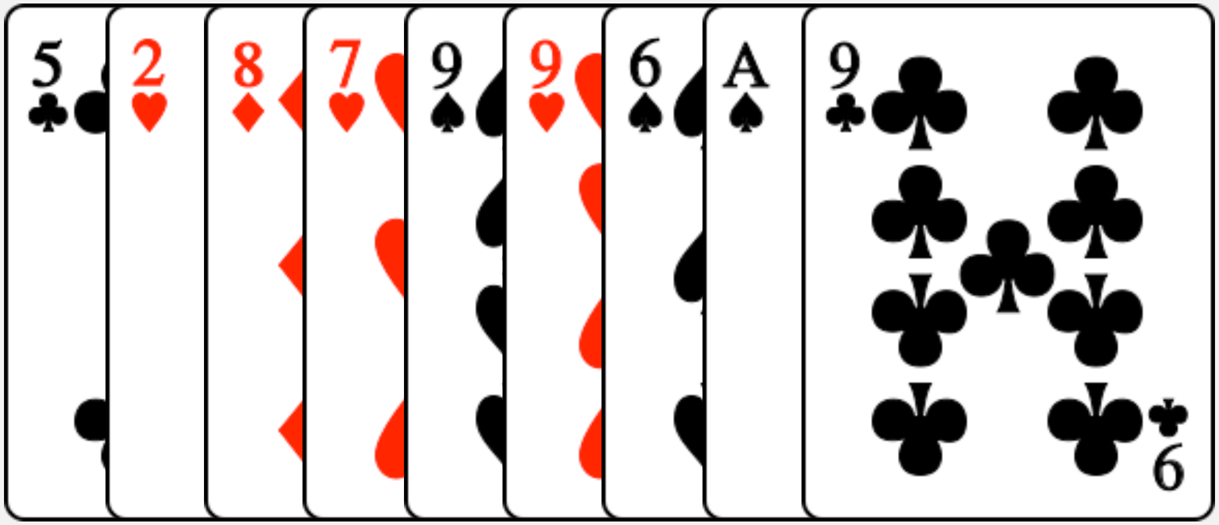


Suit	Count	Start pos
Clubs	2	0
Diamonds	1	2
Hearts	3	3
Spades	3	6

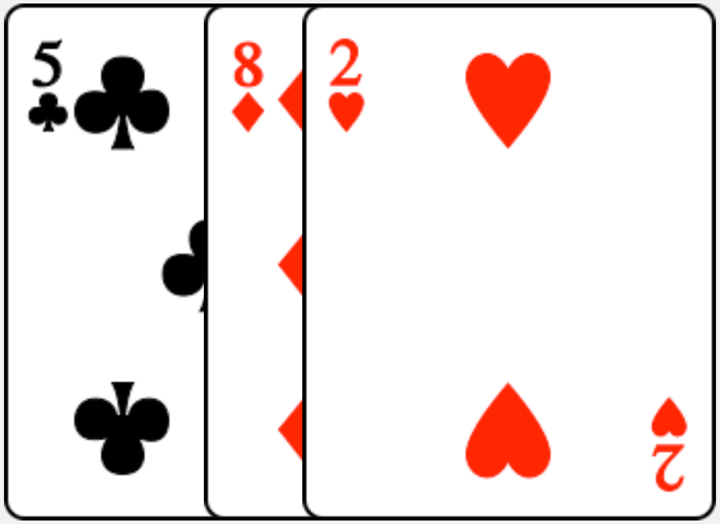


Suit	Next pos
Clubs	1
Diamonds	2
Hearts	4
Spades	6

Stably sorting a set of cards by suit

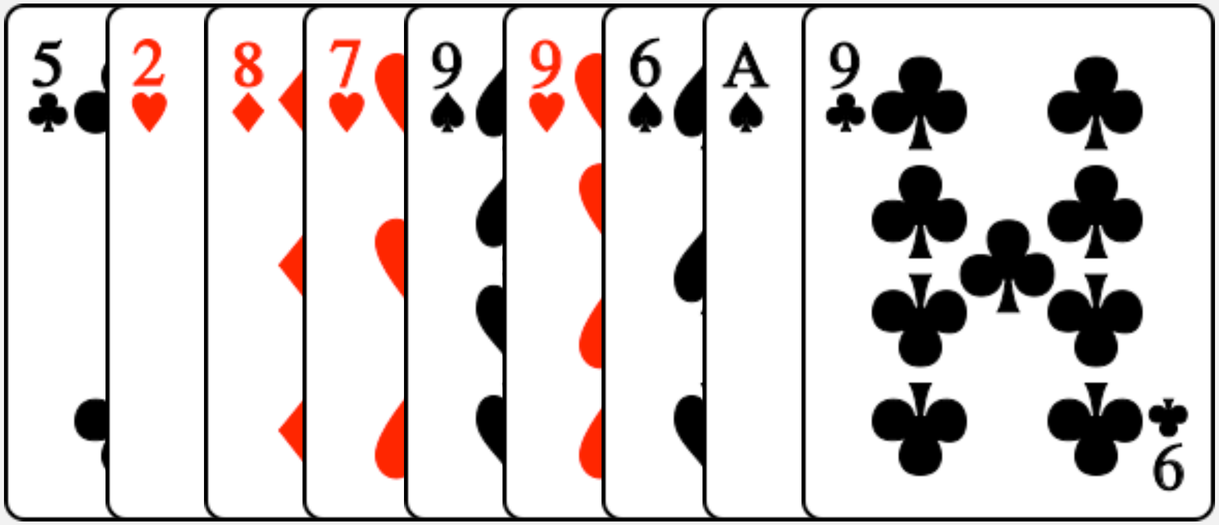


Suit	Count	Start pos
Clubs	2	0
Diamonds	1	2
Hearts	3	3
Spades	3	6

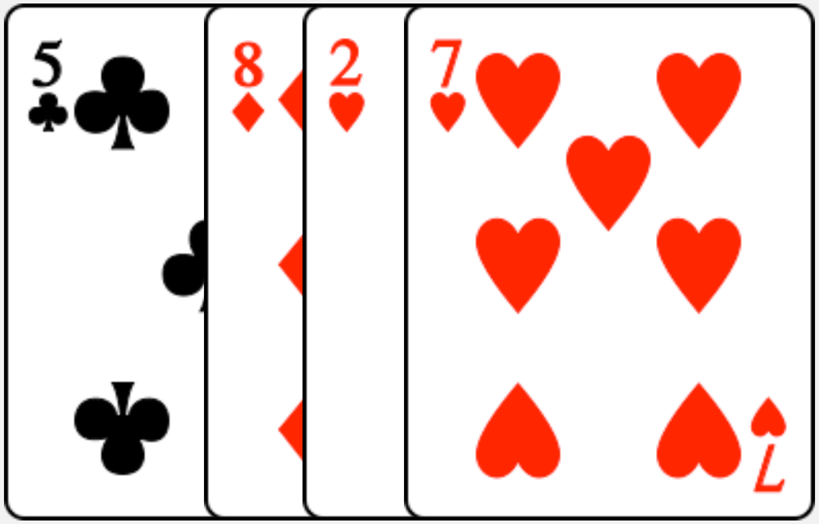


Suit	Next pos
Clubs	1
Diamonds	3
Hearts	4
Spades	6

Stably sorting a set of cards by suit

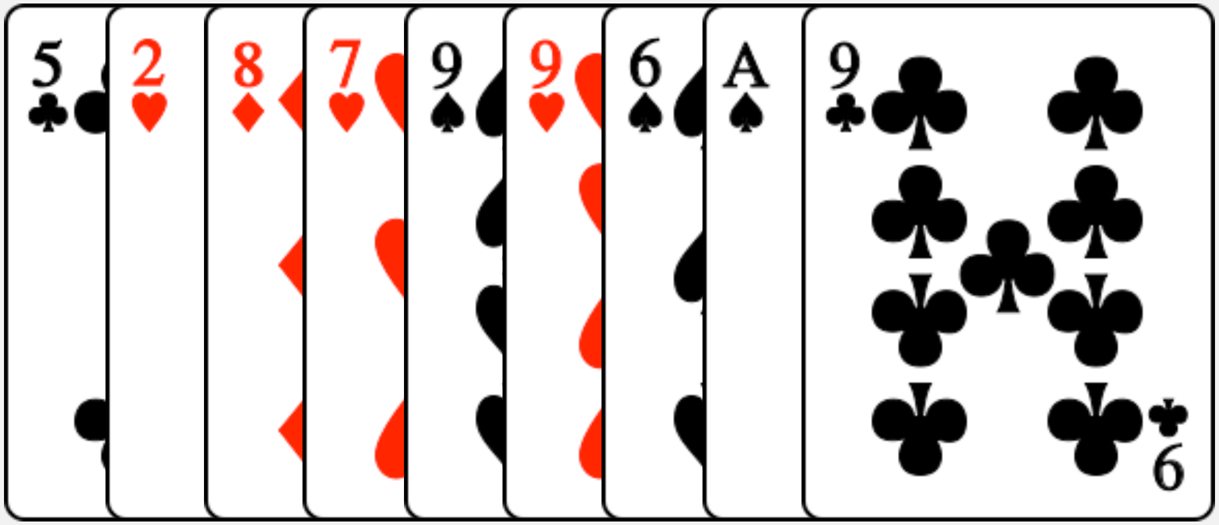


Suit	Count	Start pos
Clubs	2	0
Diamonds	1	2
Hearts	3	3
Spades	3	6

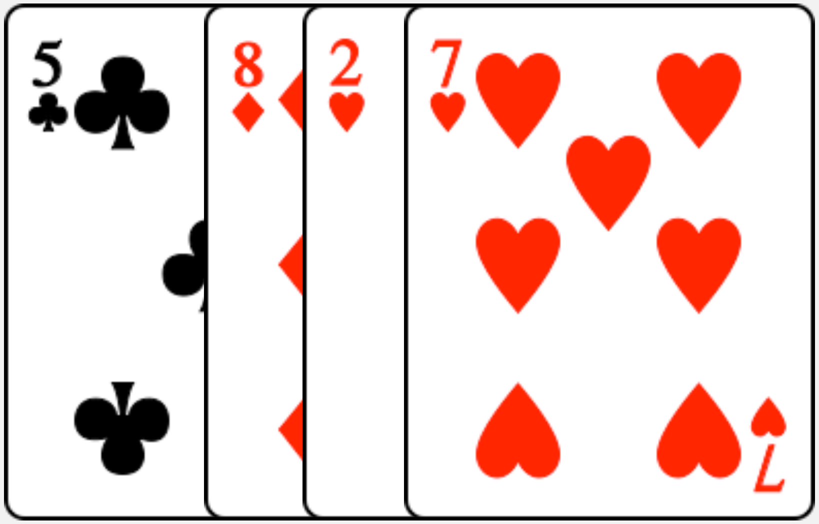


Suit	Next pos
Clubs	1
Diamonds	3
Hearts	4
Spades	6

Stably sorting a set of cards by suit

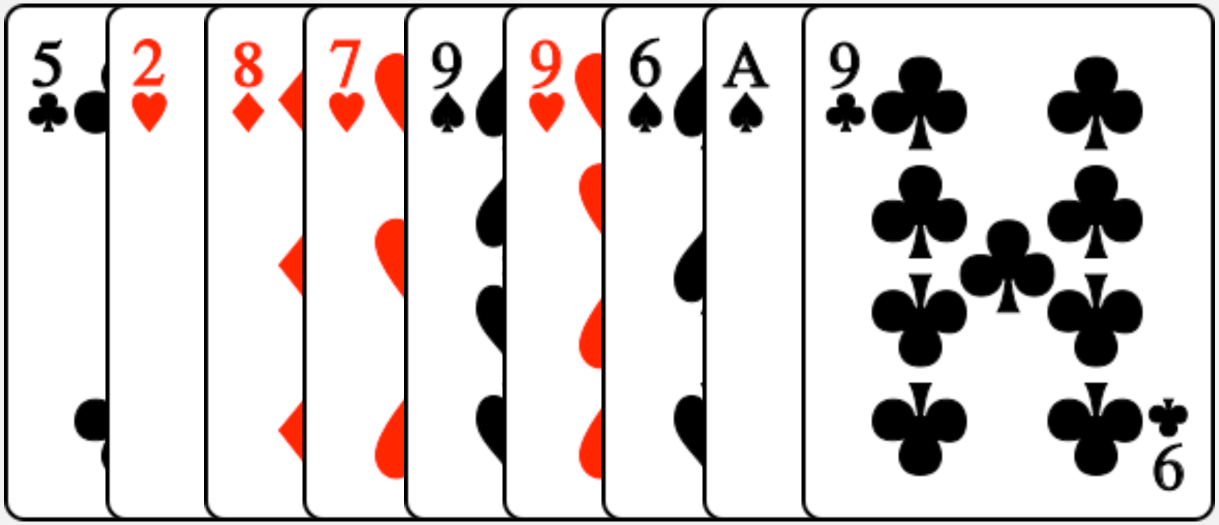


Suit	Count	Start pos
Clubs	2	0
Diamonds	1	2
Hearts	3	3
Spades	3	6

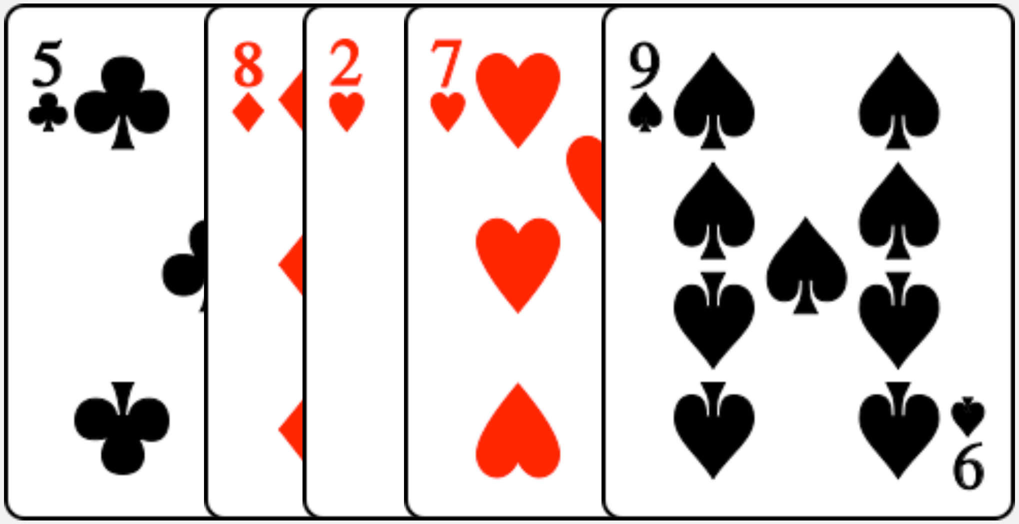


Suit	Next pos
Clubs	1
Diamonds	3
Hearts	5
Spades	6

Stably sorting a set of cards by suit

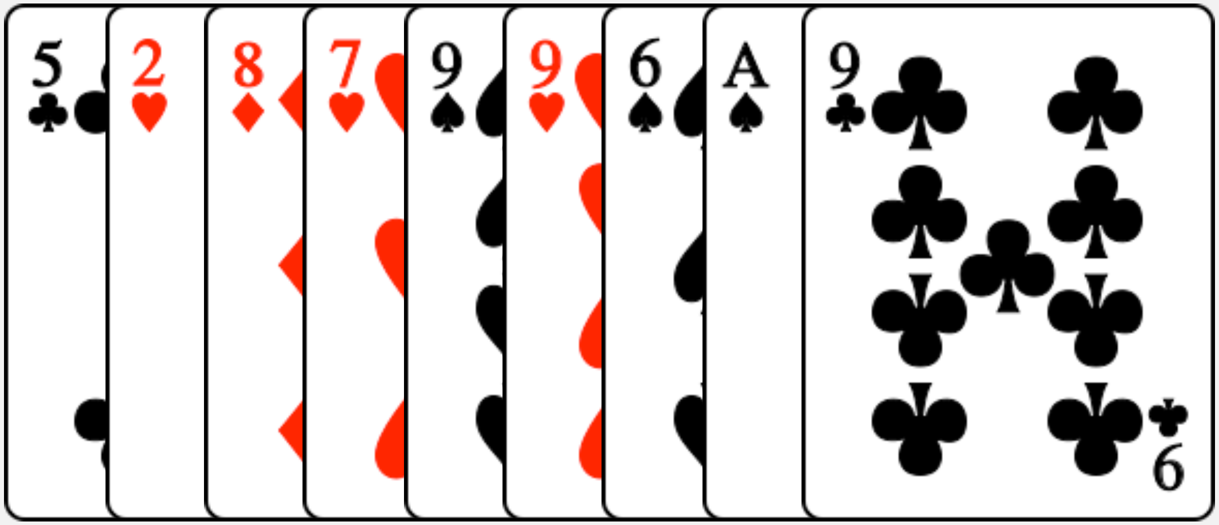


Suit	Count	Start pos
Clubs	2	0
Diamonds	1	2
Hearts	3	3
Spades	3	6

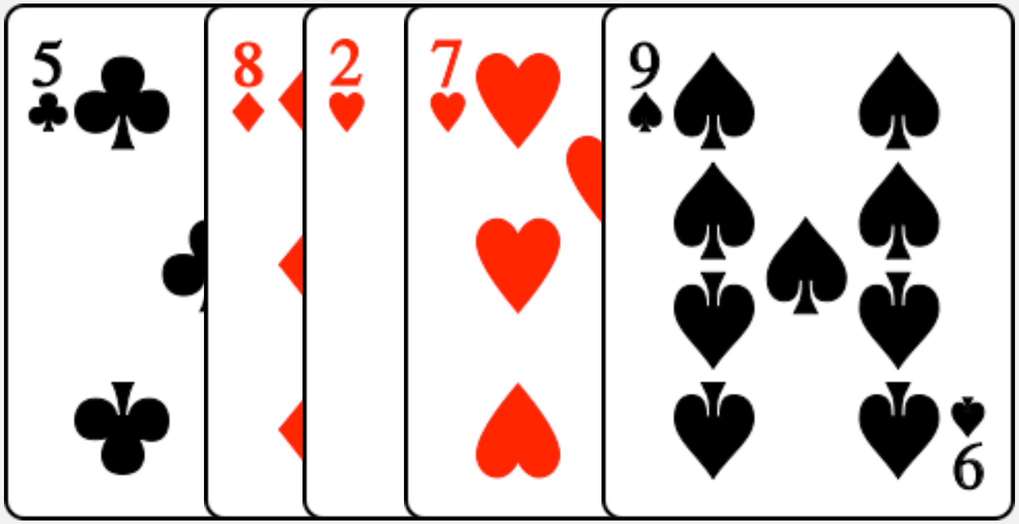


Suit	Next pos
Clubs	1
Diamonds	3
Hearts	5
Spades	6

Stably sorting a set of cards by suit

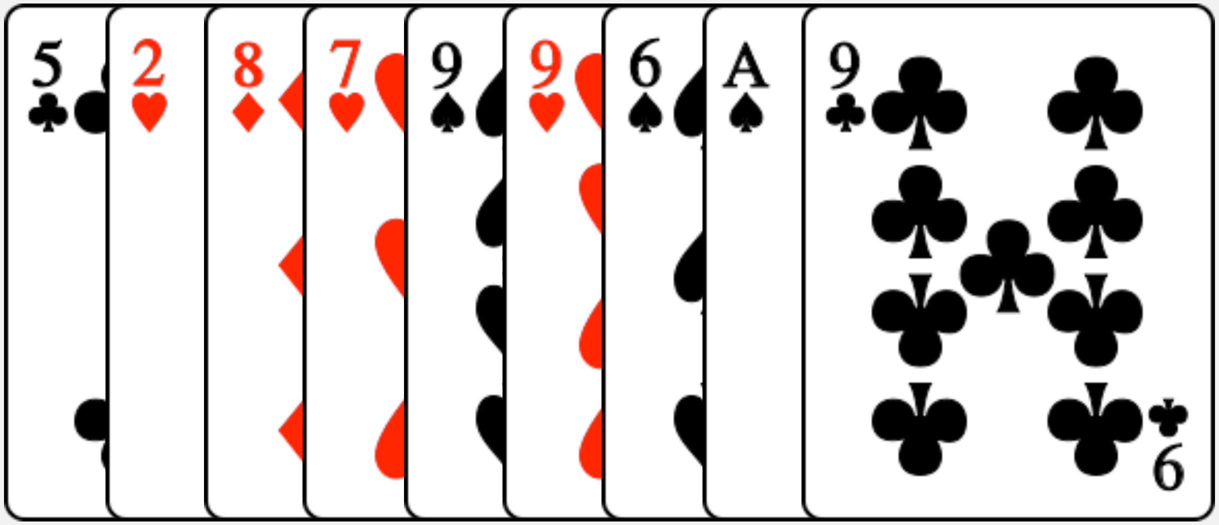


Suit	Count	Start pos
Clubs	2	0
Diamonds	1	2
Hearts	3	3
Spades	3	6

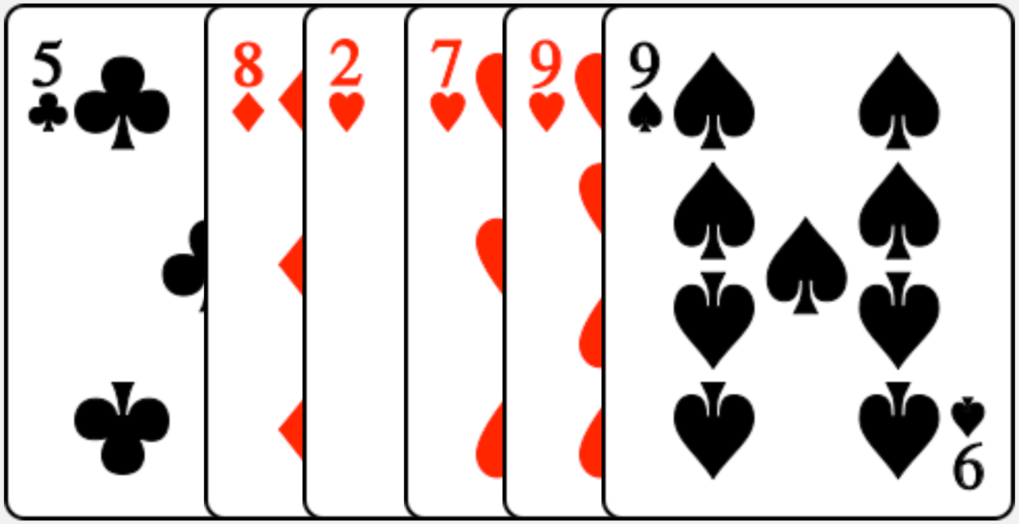


Suit	Next pos
Clubs	1
Diamonds	3
Hearts	5
Spades	7

Stably sorting a set of cards by suit

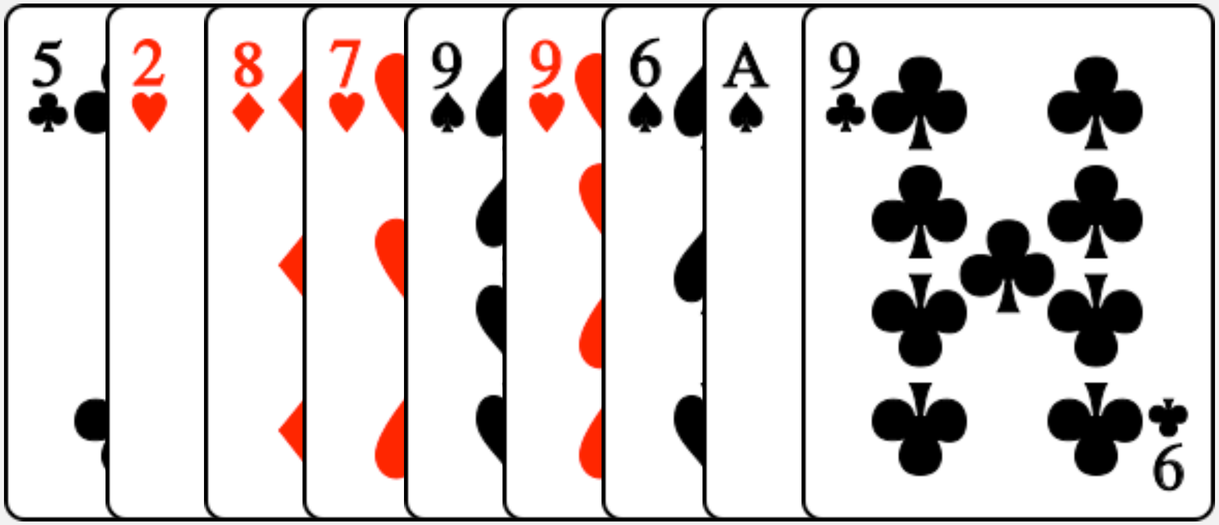


Suit	Count	Start pos
Clubs	2	0
Diamonds	1	2
Hearts	3	3
Spades	3	6

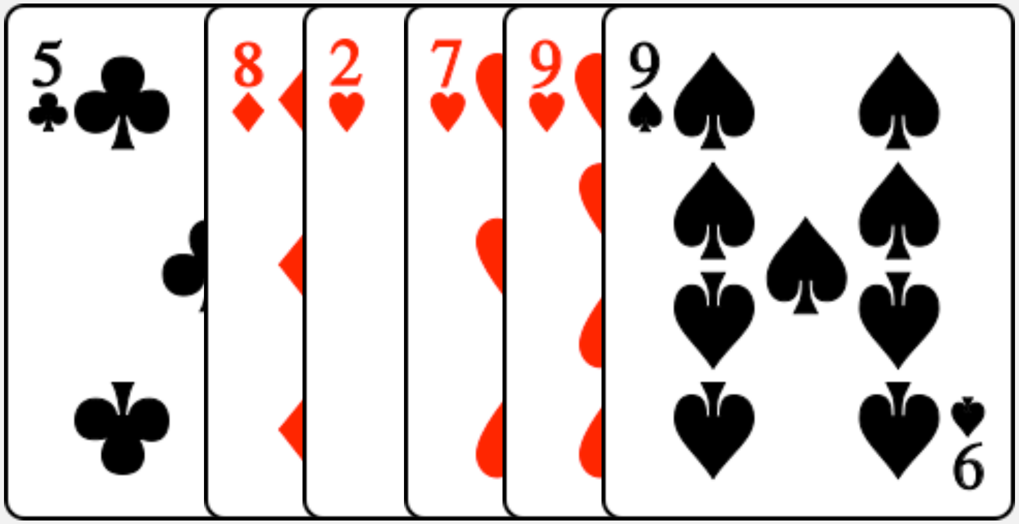


Suit	Next pos
Clubs	1
Diamonds	3
Hearts	5
Spades	7

Stably sorting a set of cards by suit

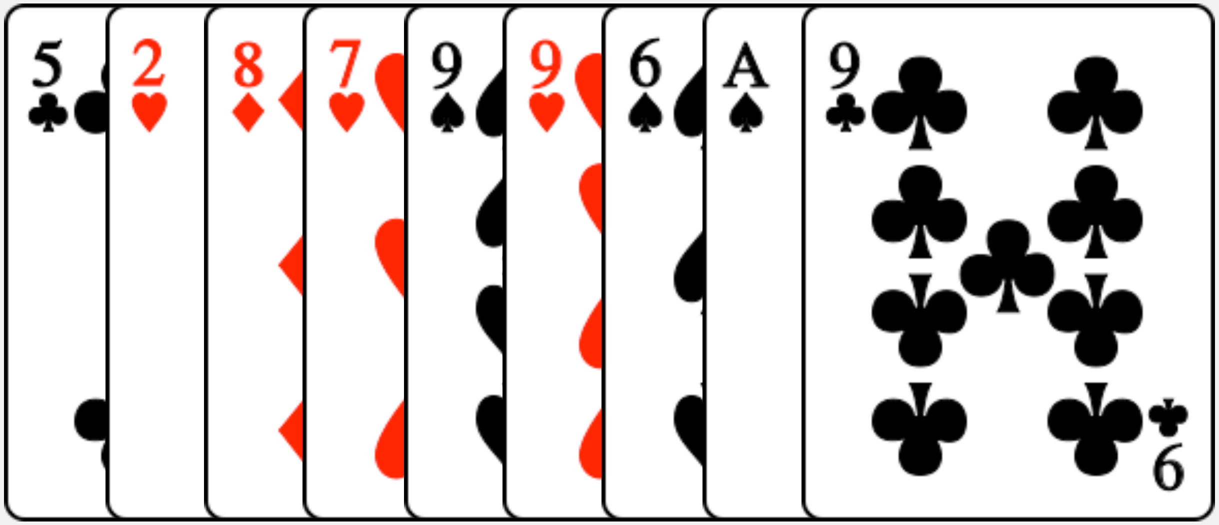


Suit	Count	Start pos
Clubs	2	0
Diamonds	1	2
Hearts	3	3
Spades	3	6

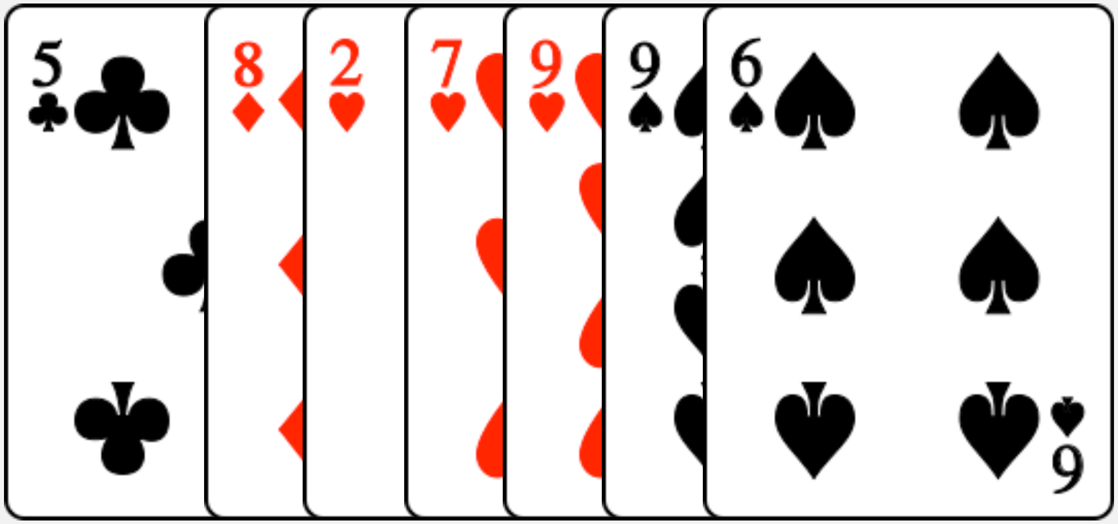


Suit	Next pos
Clubs	1
Diamonds	3
Hearts	6
Spades	7

Stably sorting a set of cards by suit

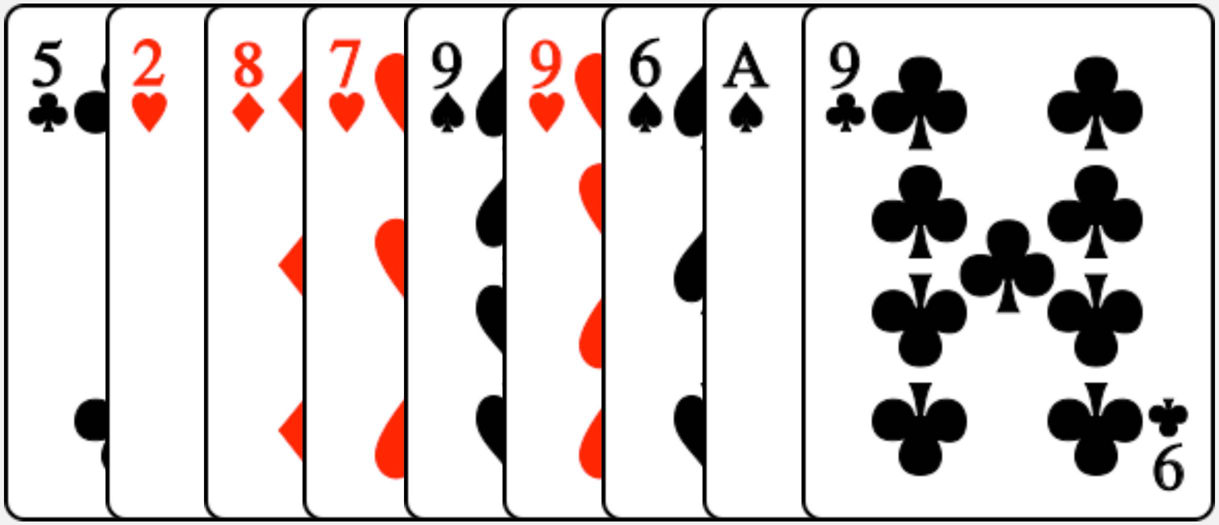


Suit	Count	Start pos
Clubs	2	0
Diamonds	1	2
Hearts	3	3
Spades	3	6

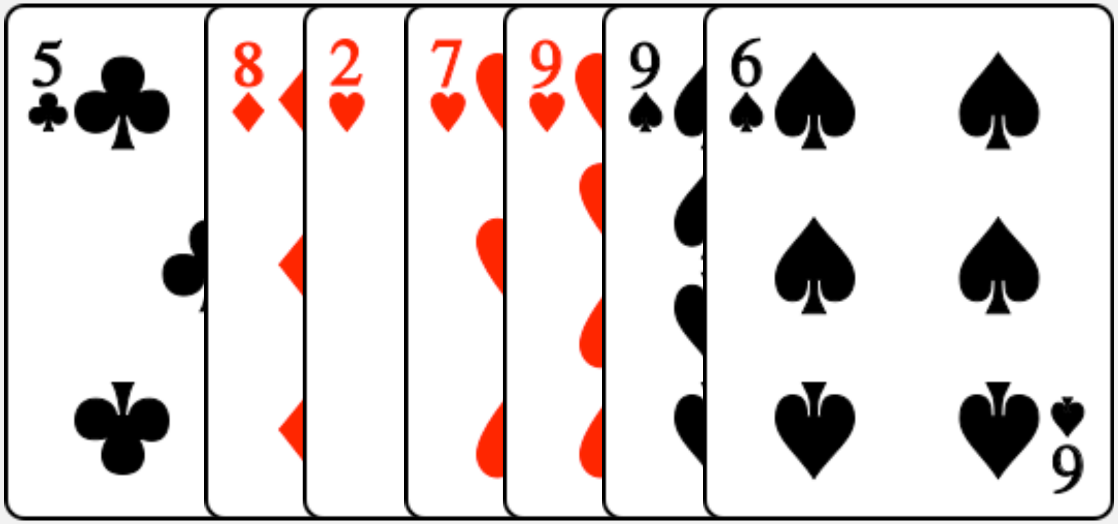


Suit	Next pos
Clubs	1
Diamonds	3
Hearts	6
Spades	7

Stably sorting a set of cards by suit

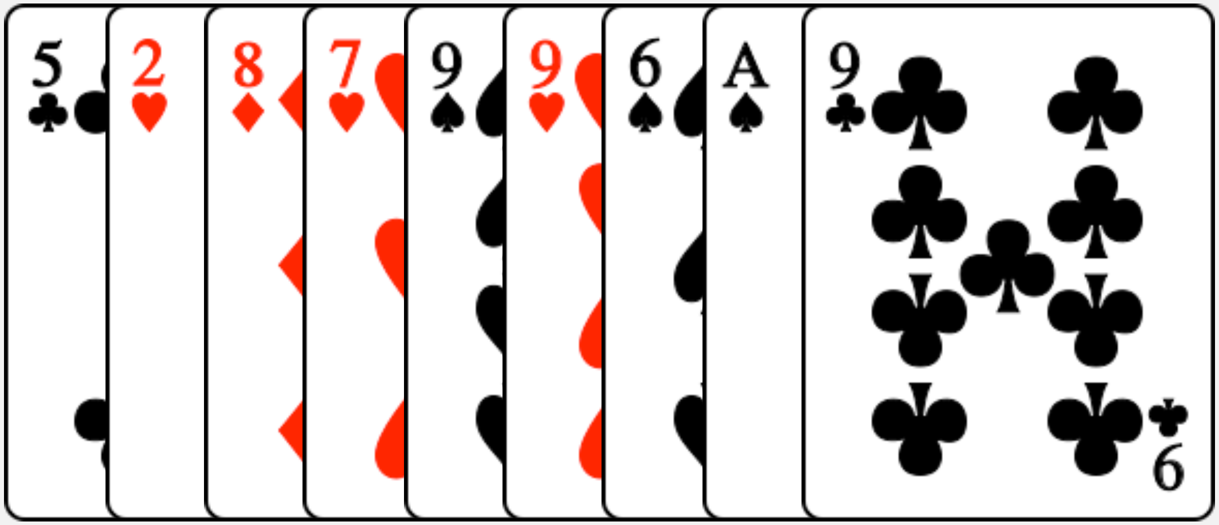


Suit	Count	Start pos
Clubs	2	0
Diamonds	1	2
Hearts	3	3
Spades	3	6

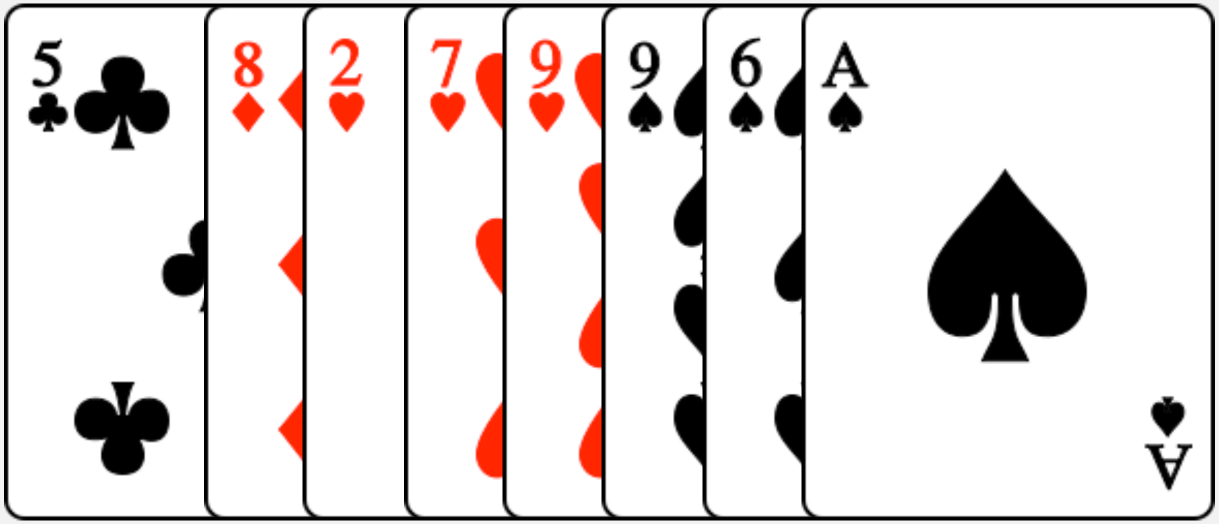


Suit	Next pos
Clubs	1
Diamonds	3
Hearts	6
Spades	8

Stably sorting a set of cards by suit

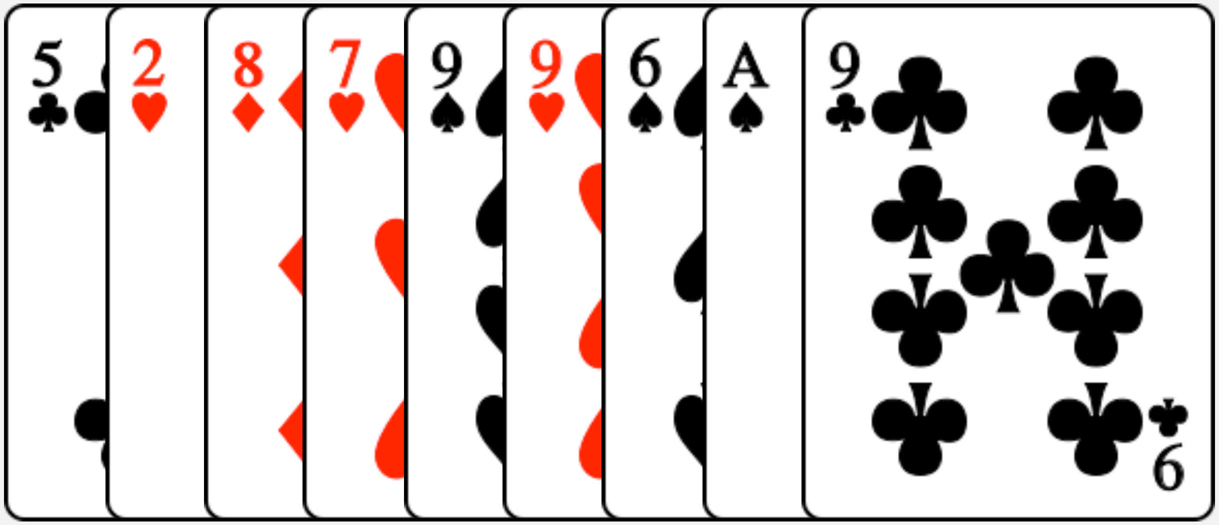


Suit	Count	Start pos
Clubs	2	0
Diamonds	1	2
Hearts	3	3
Spades	3	6

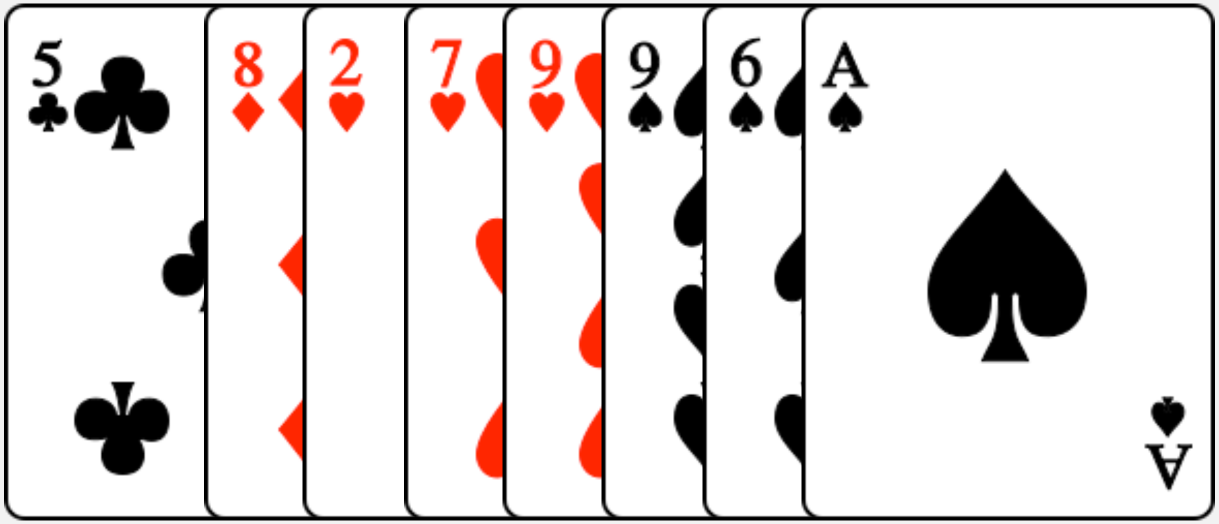


Suit	Next pos
Clubs	1
Diamonds	3
Hearts	6
Spades	8

Stably sorting a set of cards by suit

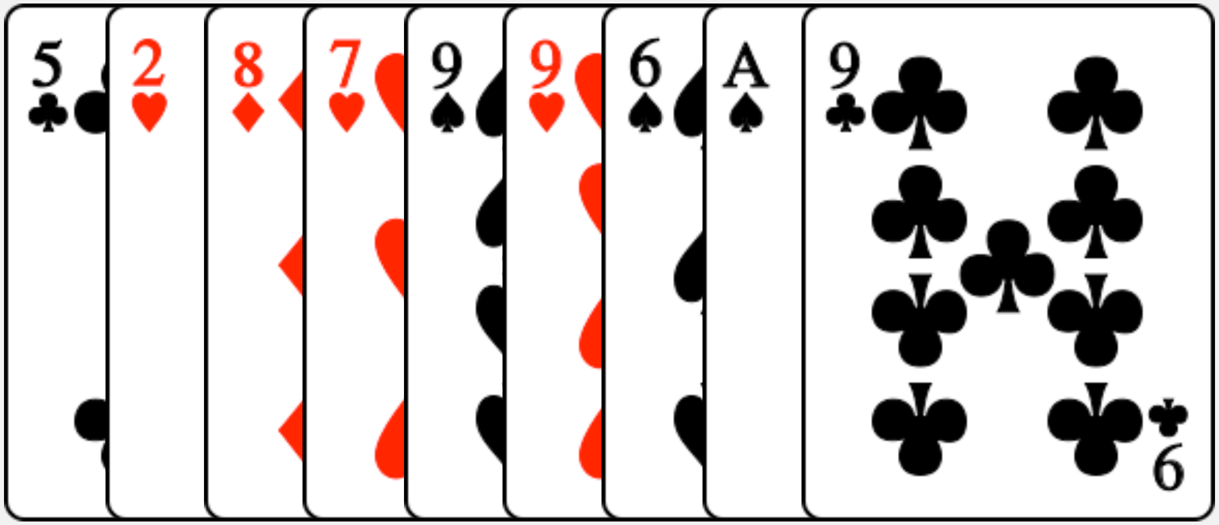


Suit	Count	Start pos
Clubs	2	0
Diamonds	1	2
Hearts	3	3
Spades	3	6

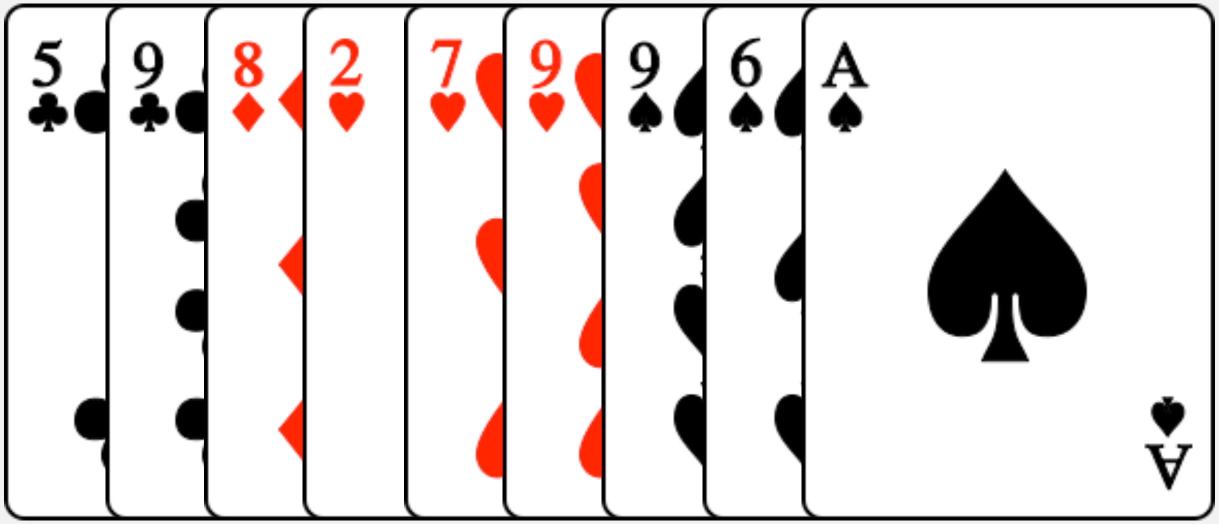


Suit	Next pos
Clubs	1
Diamonds	3
Hearts	6
Spades	9

Stably sorting a set of cards by suit



Suit	Count	Start pos
Clubs	2	0
Diamonds	1	2
Hearts	3	3
Spades	3	6



Suit	Next pos
Clubs	1
Diamonds	3
Hearts	6
Spades	9

Key-indexed counting

Goal. Sort an array `a[]` of N integers between 0 and $R - 1$.

- Count frequencies of each letter using key as index.
- Compute frequency cumulates which specify destinations.
- Access cumulates using key as index to move items.
- Copy back into original array.

```
int N = a.length;
int[] count = new int[R];
int[] pos = new int[R];

for (int i = 0; i < N; i++)
   count[a[i]]++;

for (int r = 1; r < R; r++)
   pos[r] = count[r-1] + pos[r-1];

for (int i = 0; i < N; i++)
   aux[pos[a[i]]++] = a[i];

for (int i = 0; i < N; i++)
   a[i] = aux[i];
```

Key-indexed counting

Goal. Sort an array `a[]` of N integers between 0 and $R - 1$.

- Count frequencies of each letter using key as index.
- Compute frequency cumulates which specify destinations.
- Access cumulates using key as index to move items.
- Copy back into original array.

```
int N = a.length;
int[] count = new int[R];
int[] pos = new int[R];

for (int i = 0; i < N; i++)
   count[a[i]]++;

for (int r = 1; r < R; r++)
   pos[r] = count[r-1] + pos[r-1];

for (int i = 0; i < N; i++)
   aux[pos[a[i]]++] = a[i];

for (int i = 0; i < N; i++)
   a[i] = aux[i];
```

Q. Modify this code to sort an array `a[]` of `Objects`, assuming a `key()` method that returns `int` between 0 and $R-1$.

Key-indexed counting

Goal. Sort an array `a[]` of N integers between 0 and $R - 1$.

- Count frequencies of each letter using key as index.
- Compute frequency cumulates which specify destinations.
- Access cumulates using key as index to move items.
- Copy back into original array.

```
int N = a.length;
int[] count = new int[R];
int[] pos = new int[R];

for (int i = 0; i < N; i++)
   count[a[i].key()+1]++;

for (int r = 1; r < R; r++)
   pos[r] = count[r-1] + pos[r-1];

for (int i = 0; i < N; i++)
   aux[pos[a[i].key()]++] = a[i];

for (int i = 0; i < N; i++)
   a[i] = aux[i];
```

Q. Modify this code to sort an array `a[]` of `Objects`, assuming a key() method that returns `int` between 0 and $R-1$.

Radix sorting: quiz 1

Which of the following are properties of key-indexed counting?

A. Running time proportional to $N + R$.

B. Extra space proportional to $N + R$.

C. Stable.

D. All of the above.

E. *I don't know.*

Algorithms

ROBERT SEDGEWICK | KEVIN WAYNE

http://algs4.cs.princeton.edu

5.1 STRING SORTS

- *strings in Java*
- *key-indexed counting*
- *LSD radix sort*
- *MSD radix sort*
- *3-way radix quicksort*
- *suffix arrays*

Recall from midterm

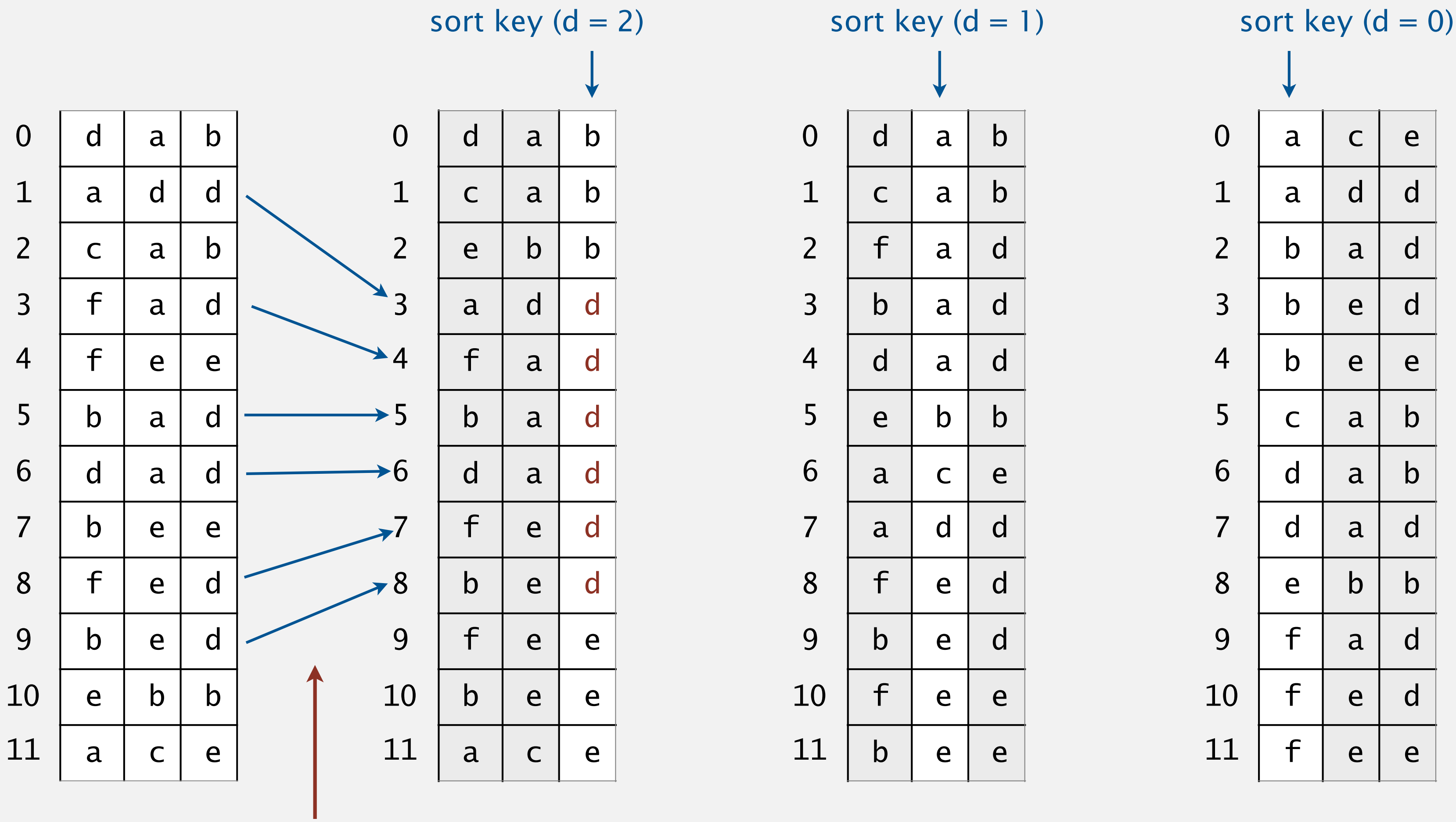
We can sort an array by date by first sorting it by day, then by month, then by year, but only if we use a stable sorting algorithm.

yyyy	mm	dd
yyyy	mm	dd
yyyy	mm	dd
yyyy	mm	dd
yyyy	mm	dd

Least-significant-digit-first string sort

LSD string (radix) sort.

- Consider characters from right to left.
- Stably sort using d^{th} character as the key (using key-indexed counting).



0	d	a	b
1	a	d	d
2	c	a	b
3	f	a	d
4	f	e	e
5	b	a	d
6	d	a	d
7	b	e	e
8	f	e	d
9	b	e	d
10	e	b	b
11	a	c	e

sort key (d = 2)

0	d	a	b
1	c	a	b
2	e	b	b
3	a	d	d
4	f	a	d
5	b	a	d
6	d	a	d
7	f	e	d
8	b	e	d
9	f	e	e
10	b	e	e
11	a	c	e

sort key (d = 1)

0	d	a	b
1	c	a	b
2	f	a	d
3	b	a	d
4	d	a	d
5	e	b	b
6	a	c	e
7	a	d	d
8	f	e	d
9	b	e	d
10	f	e	e
11	b	e	e

sort key (d = 0)

0	a	c	e
1	a	d	d
2	b	a	d
3	b	e	d
4	b	e	e
5	c	a	b
6	d	a	b
7	d	a	d
8	e	b	b
9	f	a	d
10	f	e	d
11	f	e	e

sort is stable
(arrows do not cross)

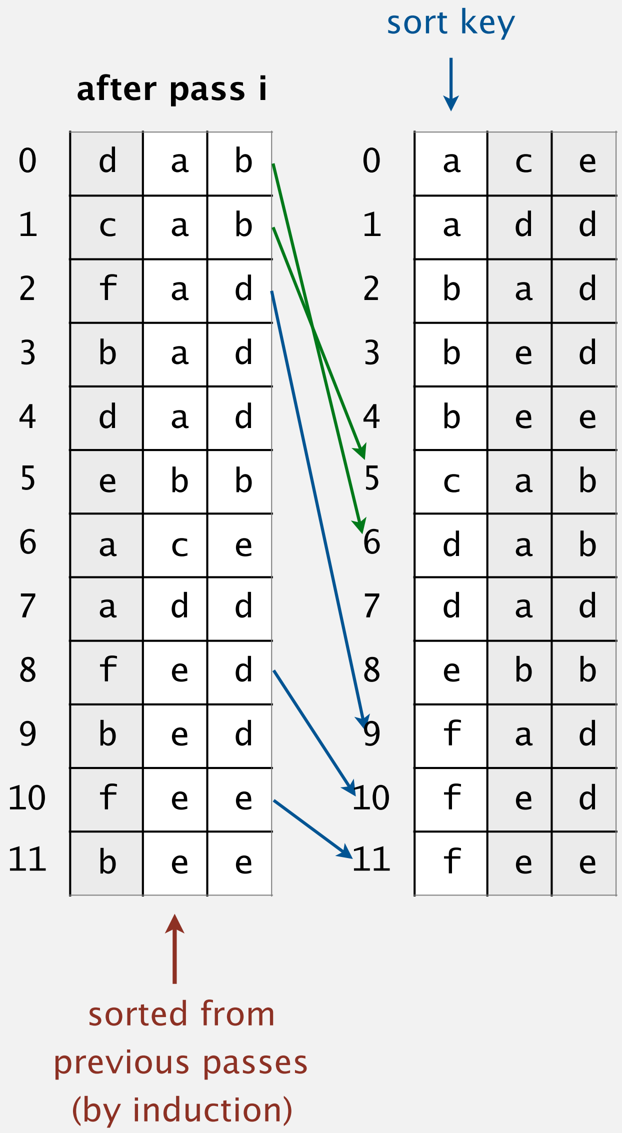
LSD string sort: correctness proof

Proposition. LSD sorts fixed-length strings in ascending order.

Pf. [by induction on i]

After pass i, strings are sorted by last i characters.

- If two strings differ on sort key, key-indexed sort puts them in proper relative order.
- If two strings agree on sort key, stability keeps them in proper relative order.



sort key

after pass i

0	d	a	b
1	c	a	b
2	f	a	d
3	b	a	d
4	d	a	d
5	e	b	b
6	a	c	e
7	a	d	d
8	f	e	d
9	b	e	d
10	f	e	e
11	b	e	e

0	a	c	e
1	a	d	d
2	b	a	d
3	b	e	d
4	b	e	e
5	c	a	b
6	d	a	b
7	d	a	d
8	e	b	b
9	f	a	d
10	f	e	d
11	f	e	e

sorted from previous passes (by induction)

Proposition. LSD sort is stable.

Pf. Key-indexed counting is stable.

LSD string sort: Java implementation

```
public class LSD
{
   public static void sort(String[] a, int W)
   {
      int R = 256;
      int N = a.length;
      String[] aux = new String[N];

      for (int d = W-1; d >= 0; d--)
      {
         int[] count = new int[R+1];
         for (int i = 0; i < N; i++)
            count[a[i].charAt(d) + 1]++;
         for (int r = 0; r < R; r++)
            count[r+1] += count[r];
         for (int i = 0; i < N; i++)
            aux[count[a[i].charAt(d)]++] = a[i];
         for (int i = 0; i < N; i++)
            a[i] = aux[i];
      }
   }
}
```

fixed-length W strings

radix R

do key-indexed counting
for each digit from right to left

key-indexed counting

Summary of the performance of sorting algorithms

Frequency of operations.

algorithm	guarantee	random	extra space	stable?	operations on keys
insertion sort	$\frac{1}{2} N^2$	$\frac{1}{4} N^2$	1	✔	`compareTo()`
mergesort	$N \lg N$	$N \lg N$	N	✔	`compareTo()`
quicksort	$1.39 N \lg N$ *	$1.39 N \lg N$	$c \lg N$		`compareTo()`
heapsort	$2 N \lg N$	$2 N \lg N$	1		`compareTo()`
LSD sort †	$2 W (N + R)$	$2 W (N + R)$	$N + R$	✔	`charAt()`

* probabilistic

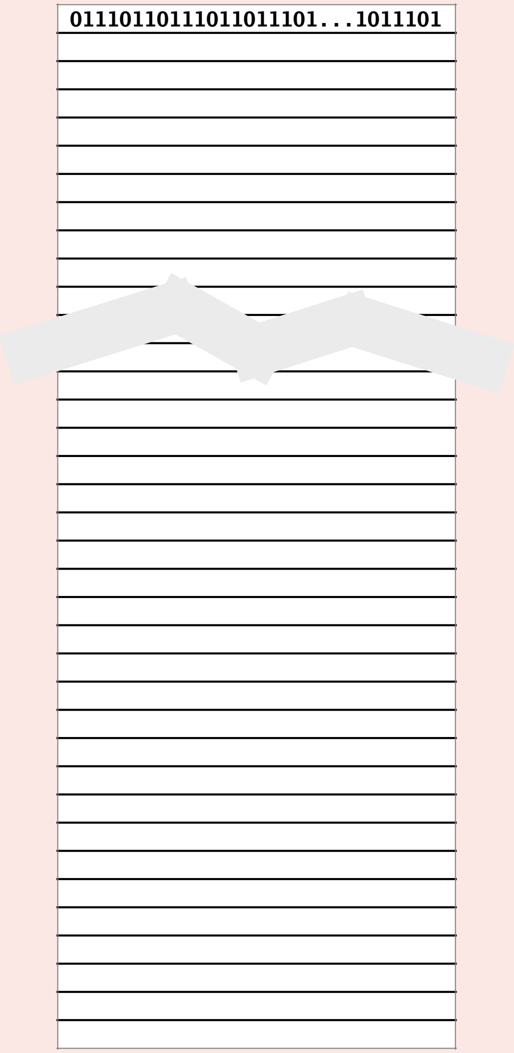
† fixed-length W keys

Q. What if strings are not all of same length?

Radix sorting: quiz 2

Which sorting method to use to sort 1 million 32-bit integers?

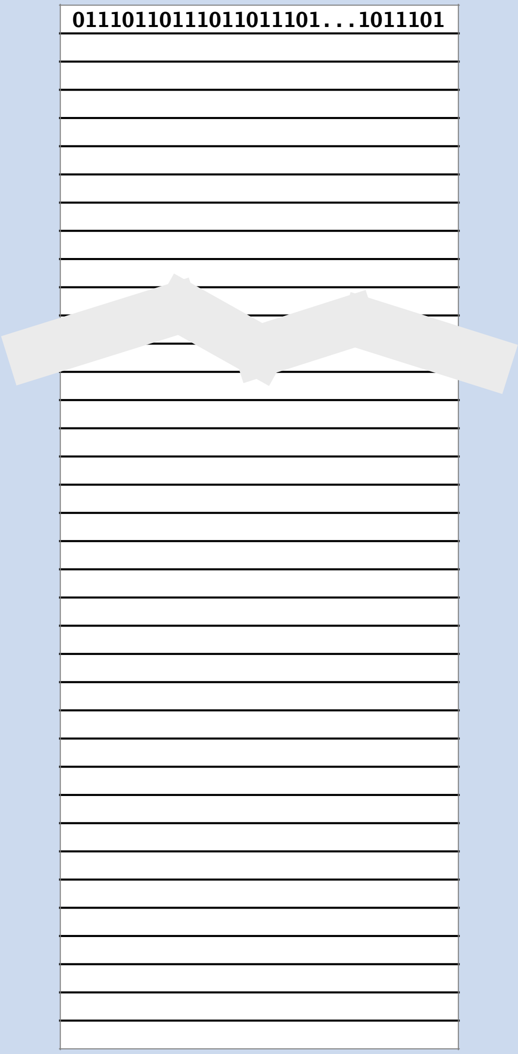
- **A.** Insertion sort.
- **B.** Mergesort.
- **C.** Quicksort.
- **D.** LSD radix sort.
- **E.** I don't know.



SORT ARRAY OF 128-BIT NUMBERS

Problem. Sort huge array of random 128-bit numbers.

Ex. Supercomputer sort, internet router.



Which sorting method to use?

- Insertion sort.
- Mergesort.
- Quicksort.
- Heapsort.
- LSD string sort.

SORT ARRAY OF 128-BIT NUMBERS

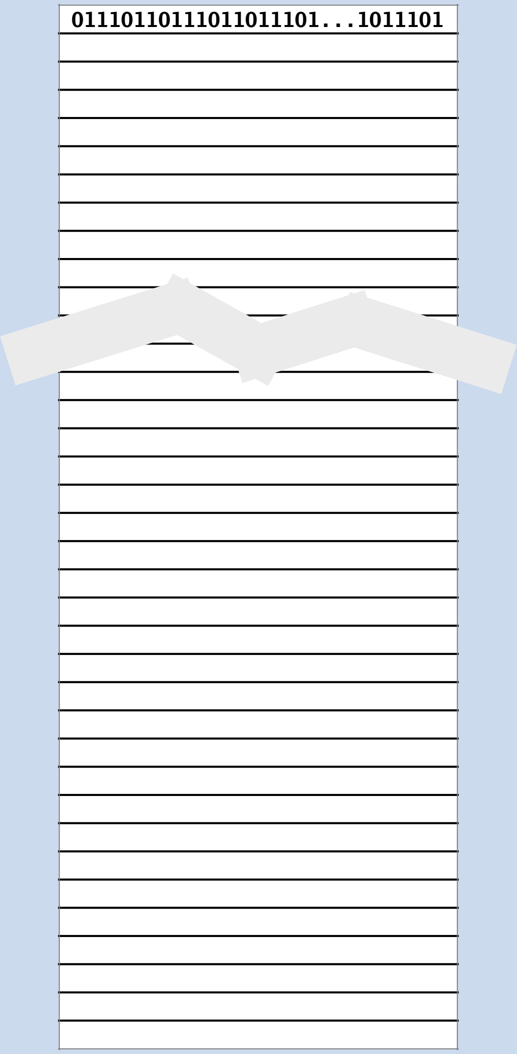
Problem. Sort huge array of random 128-bit numbers.

Ex. Supercomputer sort, internet router.

Which sorting method to use?

- Insertion sort.
- Mergesort.
- Quicksort.
- Heapsort.
- ✓ LSD string sort.

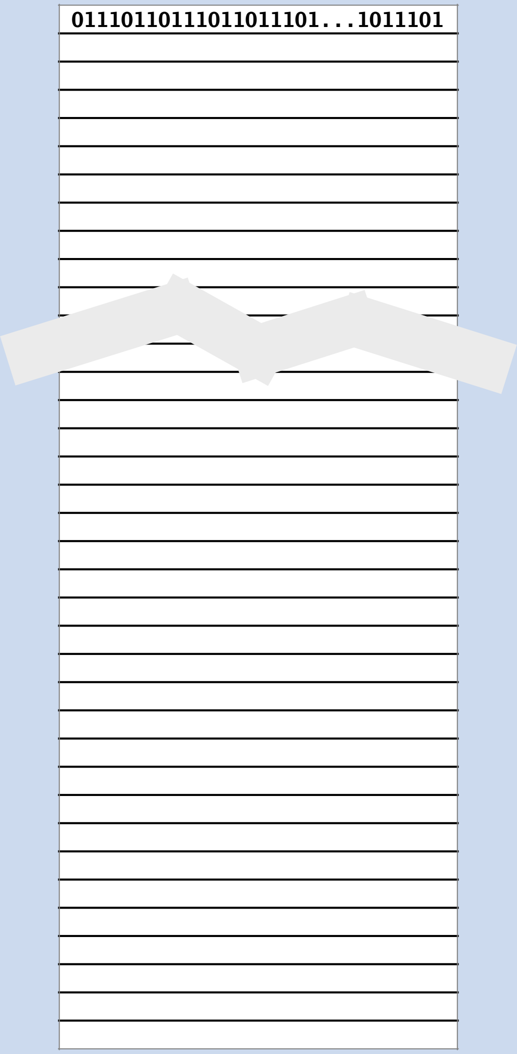
Divide each word into eight 16-bit "chars"
$2^{16} = 65,536$ counters.
Sort in 8 passes.



SORT ARRAY OF 128-BIT NUMBERS

Problem. Sort huge array of random 128-bit numbers.

Ex. Supercomputer sort, internet router.



Which sorting method to use?

- ✓ Insertion sort.
- Mergesort.
- Quicksort.
- Heapsort.
- ✓ LSD string sort.

Divide each word into eight 16-bit “chars”
2^{16} = 65,536 counters
LSD sort on leading 32 bits in 2 passes
Finish with insertion sort
Examines only ~25% of the data

How to take a census in 1900s?

1880 Census. Took 1500 people 7 years to manually process data.

Herman Hollerith. Developed a tabulating and sorting machine.

- Use punch cards to record data (e.g., sex, age).
- Machine sorts one column at a time (into one of 12 bins).
- Typical question: how many women of age 20 to 30?

Hollerith tabulating machine and sorter

punch card (12 holes per column)

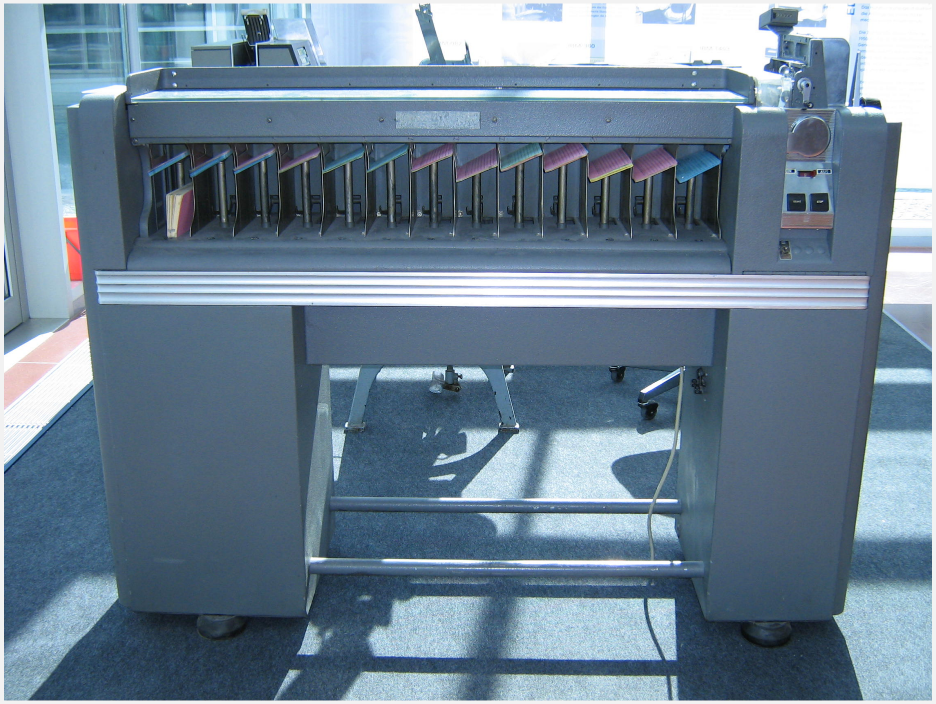
1890 Census. Finished in 1 year (and under budget)!

How to get rich sorting in 1900s?

Punch cards. [1900s to 1950s]

- Also useful for accounting, inventory, and business processes.
- Primary medium for data entry, storage, and processing.

Hollerith's company later merged with 3 others to form Computing Tabulating Recording Corporation (CTRC); company renamed in 1924.



IBM 80 Series Card Sorter (650 cards per minute)

IBM®

LSD string sort: a moment in history (1960s)

card punch

punched cards

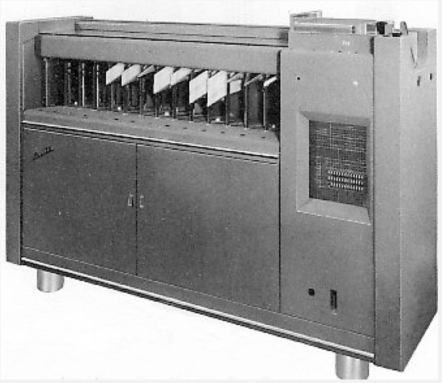
card reader

mainframe

line printer

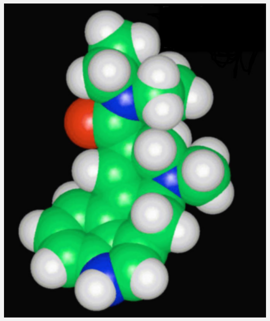
To sort a card deck

- start on right column
- put cards into hopper
- machine distributes into bins
- pick up cards (stable)
- move left one column
- continue until sorted



card sorter

not directly related
to sorting



Lysergic Acid Diethylamide

(Lucy in the Sky with Diamonds)

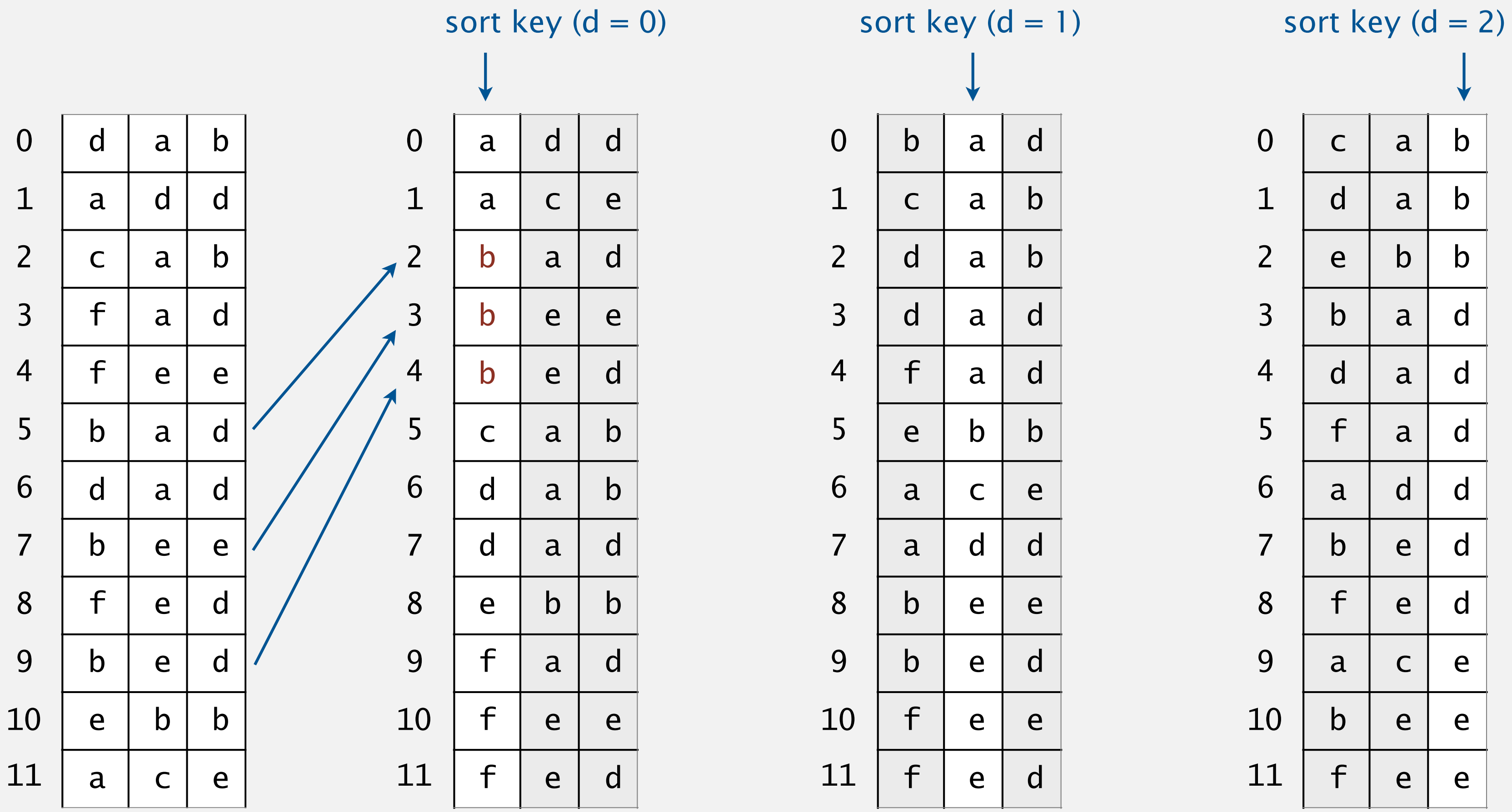
http://algs4.cs.princeton.edu

5.1 STRING SORTS

- *strings in Java*
- *key-indexed counting*
- *LSD radix sort*
- *MSD radix sort*
- *3-way radix quicksort*
- *suffix arrays*

Reverse LSD

- Consider characters from left to right.
- Stably sort using d^{th} character as the key (using key-indexed counting).



0	d	a	b
1	a	d	d
2	c	a	b
3	f	a	d
4	f	e	e
5	b	a	d
6	d	a	d
7	b	e	e
8	f	e	d
9	b	e	d
10	e	b	b
11	a	c	e

sort key (d = 0)

0	a	d	d
1	a	c	e
2	b	a	d
3	b	e	e
4	b	e	d
5	c	a	b
6	d	a	b
7	d	a	d
8	e	b	b
9	f	a	d
10	f	e	e
11	f	e	d

sort key (d = 1)

0	b	a	d
1	c	a	b
2	d	a	b
3	d	a	d
4	f	a	d
5	e	b	b
6	a	c	e
7	a	d	d
8	b	e	e
9	b	e	d
10	f	e	e
11	f	e	d

sort key (d = 2)

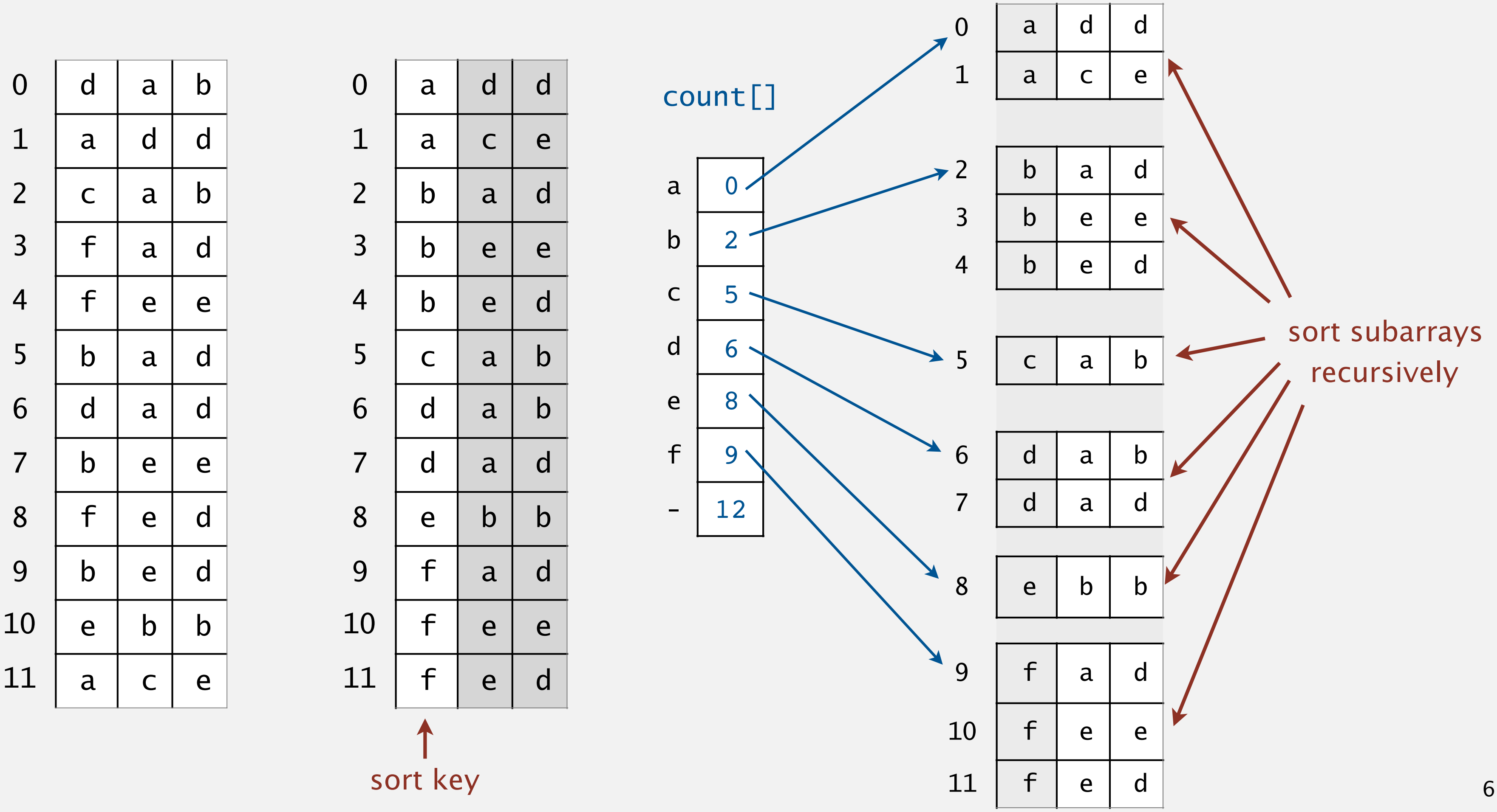
0	c	a	b
1	d	a	b
2	e	b	b
3	b	a	d
4	d	a	d
5	f	a	d
6	a	d	d
7	b	e	d
8	f	e	d
9	a	c	e
10	b	e	e
11	f	e	e

not sorted!

Most-significant-digit-first string sort

MSD string (radix) sort.

- Partition array into R pieces according to first character (use key-indexed counting).
- Recursively sort all strings that start with each character (key-indexed counts delineate subarrays to sort).



0	d	a	b
1	a	d	d
2	c	a	b
3	f	a	d
4	f	e	e
5	b	a	d
6	d	a	d
7	b	e	e
8	f	e	d
9	b	e	d
10	e	b	b
11	a	c	e

0	a	d	d
1	a	c	e
2	b	a	d
3	b	e	e
4	b	e	d
5	c	a	b
6	d	a	b
7	d	a	d
8	e	b	b
9	f	a	d
10	f	e	e
11	f	e	d

sort key

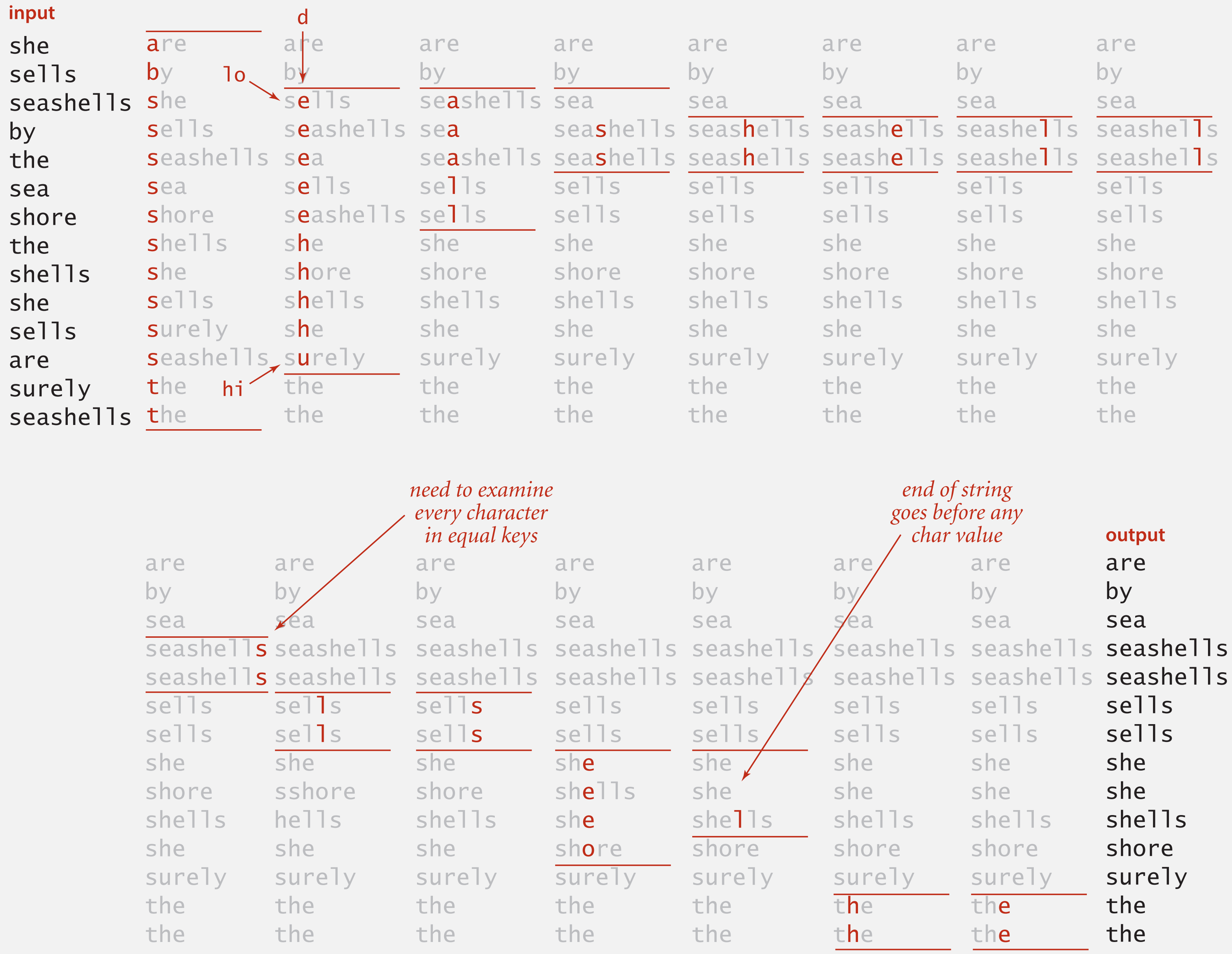
count[]

a	0
b	2
c	5
d	6
e	8
f	9
-	12

0	a	d	d
1	a	c	e
2	b	a	d
3	b	e	e
4	b	e	d
5	c	a	b
6	d	a	b
7	d	a	d
8	e	b	b
9	f	a	d
10	f	e	e
11	f	e	d

sort subarrays recursively

MSD string sort: example



she	are	are	are	are	are	are	are	are
sells	by	by	by	by	by	by	by	by
seashells	she	sells	seashells	sea	sea	sea	sea	sea
by	sells	seashells	sea	seashells	seashells	seashells	seashells	seashells
the	seashells	sea	seashells	seashells	seashells	seashells	seashells	seashells
sea	sea	sells	sells	sells	sells	sells	sells	sells
shore	shore	seashells	sells	sells	sells	sells	sells	sells
the	shells	she	she	she	she	she	she	she
shells	she	shore	shore	shore	shore	shore	shore	shore
she	sells	shells	shells	shells	shells	shells	shells	shells
sells	surely	she	she	she	she	she	she	she
are	seashells	surely	surely	surely	surely	surely	surely	surely
surely	the	the	the	the	the	the	the	the
seashells	the	the	the	the	the	the	the	the

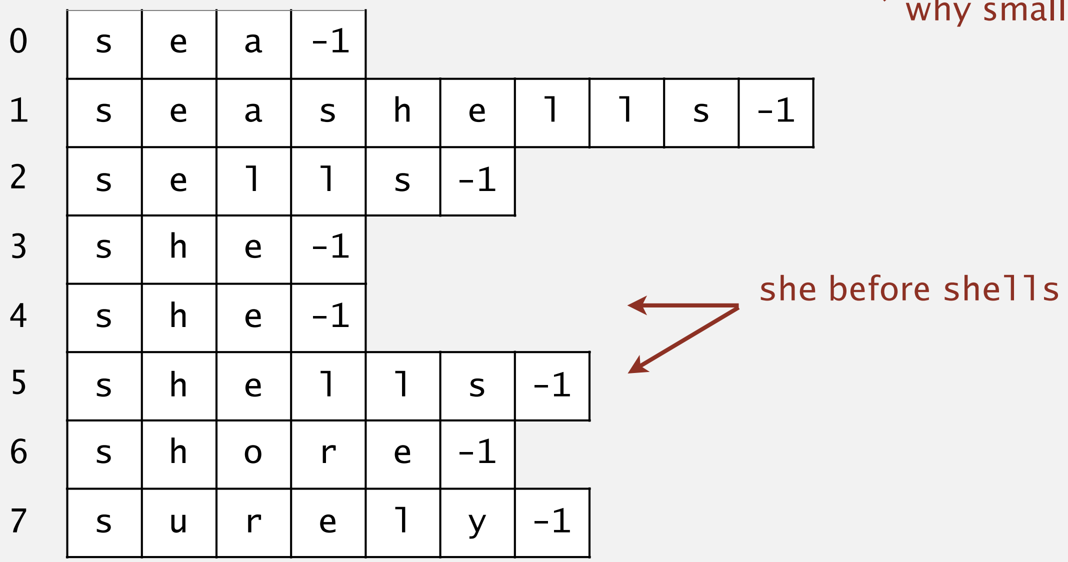
are	are	are	are	are	are	are	are
by	by	by	by	by	by	by	by
sea	sea	sea	sea	sea	sea	sea	sea
seashells	seashells	seashells	seashells	seashells	seashells	seashells	seashells
seashells	seashells	seashells	seashells	seashells	seashells	seashells	seashells
sells	sells	sells	sells	sells	sells	sells	sells
sells	sells	sells	sells	sells	sells	sells	sells
she	she	she	she	she	she	she	she
shore	sshore	shore	shells	she	she	she	she
shells	hells	shells	she	shells	shells	shells	shells
she	she	she	shore	shore	shore	shore	shore
surely	surely	surely	surely	surely	surely	surely	surely
the	the	the	the	the	the	the	the
the	the	the	the	the	the	the	the

Trace of recursive calls for MSD string sort (no cutoff for small subarrays, subarrays of size 0 and 1 omitted)

Variable-length strings

Treat strings as if they had an extra char at end (smaller than any char).

why smaller?



0	s	e	a	-1						
1	s	e	a	s	h	e	l	l	s	-1
2	s	e	l	l	s	-1				
3	s	h	e	-1						
4	s	h	e	-1						
5	s	h	e	l	l	s	-1			
6	s	h	o	r	e	-1				
7	s	u	r	e	l	y	-1			

she before shells

```
private static int charAt(String s, int d)
{
   if (d < s.length()) return s.charAt(d);
   else return -1;
}
```

C strings. Have extra char `'\0'` at end ⇒ no extra work needed.

MSD string sort: Java implementation

```
public static void sort(String[] a)
{
   aux = new String[a.length];
   sort(a, aux, 0, a.length - 1, 0);
}

private static void sort(String[] a, String[] aux, int lo, int hi, int d)
{
   if (hi <= lo) return;
   int[] count = new int[R+2];
   for (int i = lo; i <= hi; i++)
      count[charAt(a[i], d) + 2]++;
   for (int r = 0; r < R+1; r++)
      count[r+1] += count[r];
   for (int i = lo; i <= hi; i++)
      aux[count[charAt(a[i], d) + 1]++] = a[i];
   for (int i = lo; i <= hi; i++)
      a[i] = aux[i - lo];

   for (int r = 0; r < R; r++)
      sort(a, aux, lo + count[r], lo + count[r+1] - 1, d+1);
}
```

recycles aux[] array
but not count[] array

key-indexed counting

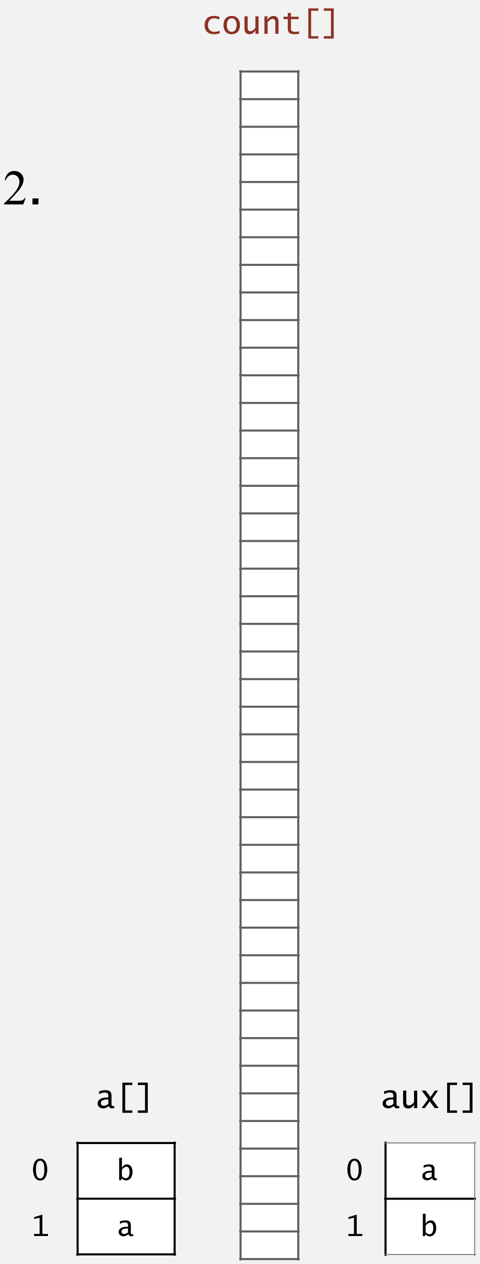
sort R subarrays recursively

MSD string sort: potential for disastrous performance

Observation 1. Much too slow for small subarrays.

- Each function call needs its own `count[]` array.
- ASCII (256 counts): 100x slower than copy pass for $N = 2$.
- Unicode (65,536 counts): 32,000x slower for $N = 2$.

Observation 2. Huge number of small subarrays because of recursion.



Cutoff to insertion sort

Solution. Cutoff to insertion sort for small subarrays.

- Insertion sort, but start at d^{th} character.

```
private static void sort(String[] a, int lo, int hi, int d)
{
   for (int i = lo; i <= hi; i++)
      for (int j = i; j > lo && less(a[j], a[j-1], d); j--)
         exch(a, j, j-1);
}
```

- Implement `less()` so that it compares starting at d^{th} character.

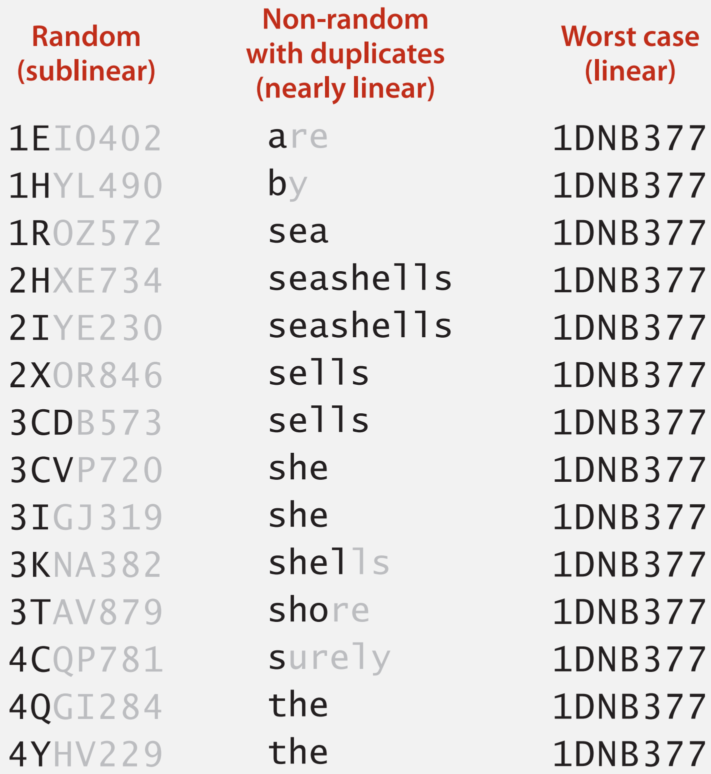
```
private static boolean less(String v, String w, int d)
 {
   for (int i = d; i < Math.min(v.length(), w.length()); i++)
   {
      if (v.charAt(i) < w.charAt(i)) return true;
      if (v.charAt(i) > w.charAt(i)) return false;
   }
   return v.length() < w.length();
 }
```

MSD string sort: performance

Number of characters examined.

- MSD examines just enough characters to sort the keys.
- Number of characters examined depends on keys.
- Can be sublinear in input size!

`compareTo()` based sorts can also be sublinear!



Random (sublinear)	Non-random with duplicates (nearly linear)	Worst case (linear)
1EIO402	are	1DNB377
1HYL490	by	1DNB377
1ROZ572	sea	1DNB377
2HXE734	seashells	1DNB377
2IYE230	seashells	1DNB377
2XOR846	sells	1DNB377
3CDB573	sells	1DNB377
3CVP720	she	1DNB377
3IGJ319	she	1DNB377
3KNA382	shells	1DNB377
3TAV879	shore	1DNB377
4CQP781	surely	1DNB377
4QGI284	the	1DNB377
4YHV229	the	1DNB377

Characters examined by MSD string sort

Summary of the performance of sorting algorithms

Frequency of operations.

algorithm	guarantee	random	extra space	stable?	operations on keys
insertion sort	$\frac{1}{2} N^2$	$\frac{1}{4} N^2$	1	✔	`compareTo()`
mergesort	$N \lg N$	$N \lg N$	N	✔	`compareTo()`
quicksort	$1.39 N \lg N$ *	$1.39 N \lg N$	$c \lg N$ *		`compareTo()`
heapsort	$2 N \lg N$	$2 N \lg N$	1		`compareTo()`
LSD sort †	$2 W (N + R)$	$2 W (N + R)$	$N + R$	✔	`charAt()`
MSD sort ‡	$2 W (N + R)$	$N \log_R N$	$N + D R$	✔	`charAt()`

D = function-call stack depth
(length of longest prefix match)

* probabilistic
† fixed-length W keys
‡ average-length W keys

MSD string sort vs. quicksort for strings

Disadvantages of MSD string sort.

- Extra space for `aux[]`.
- Extra space for `count[]`.
- Inner loop has a lot of instructions.
- Accesses memory "randomly" (cache inefficient).

Disadvantage of quicksort.

- Linearithmic number of string compares (not linear).
- Has to rescan many characters in keys with long prefix matches.

doesn't rescan characters

tight inner loop, cache friendly

Goal. Combine advantages of MSD and quicksort.

Engineering a radix sort (American flag sort)

Optimization 0. Cutoff to insertion sort.

Optimization 1. Replace recursion with explicit stack.

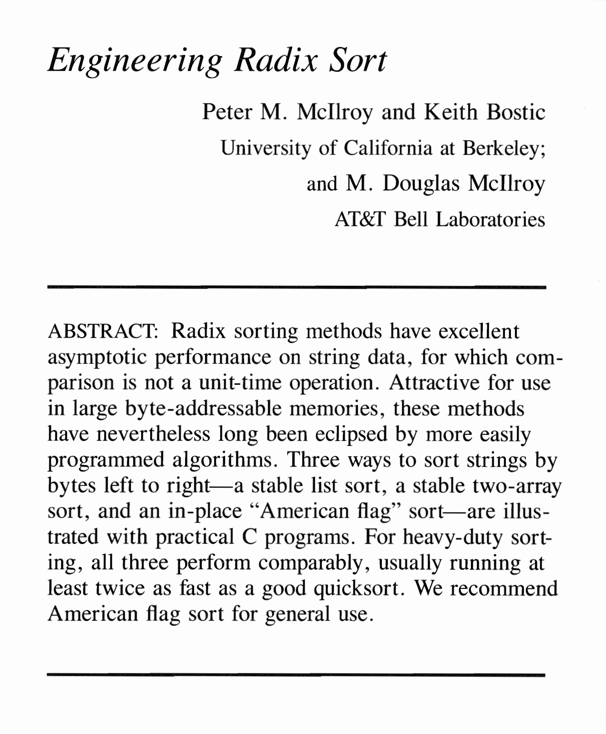
- Push subarrays to be sorted onto stack.
- One `count[]` array now suffices.

Optimization 2. Do *R*-way partitioning in place.

- Eliminates `aux[]` array.
- Sacrifices stability.

American national flag problem

Dutch national flag problem



Engineering Radix Sort

Peter M. McIlroy and Keith Bostic
University of California at Berkeley;
and M. Douglas McIlroy
AT&T Bell Laboratories

ABSTRACT: Radix sorting methods have excellent asymptotic performance on string data, for which comparison is not a unit-time operation. Attractive for use in large byte-addressable memories, these methods have nevertheless long been eclipsed by more easily programmed algorithms. Three ways to sort strings by bytes left to right—a stable list sort, a stable two-array sort, and an in-place "American flag" sort—are illustrated with practical C programs. For heavy-duty sorting, all three perform comparably, usually running at least twice as fast as a good quicksort. We recommend American flag sort for general use.

Algorithms

Robert Sedgewick | Kevin Wayne

http://algs4.cs.princeton.edu

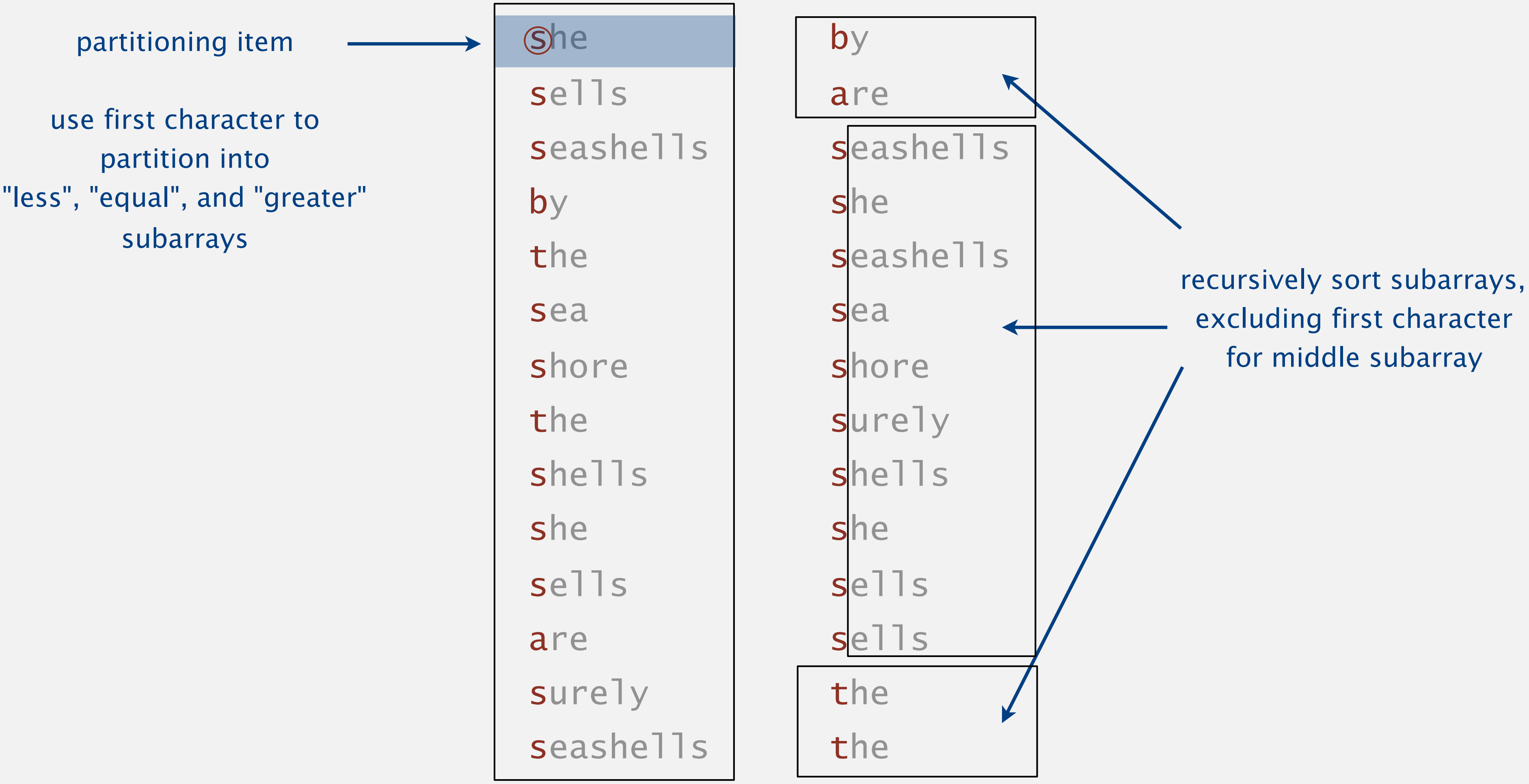
5.1 STRING SORTS

- *strings in Java*
- *key-indexed counting*
- *LSD radix sort*
- *MSD radix sort*
- *3-way radix quicksort*
- *suffix arrays*

3-way string quicksort (Bentley and Sedgewick, 1997)

Overview. Do 3-way partitioning on the d^{th} character.

- Less overhead than R-way partitioning in MSD radix sort.
- Does not re-examine characters equal to the partitioning char.
  (but does re-examine characters not equal to the partitioning char)



partitioning item

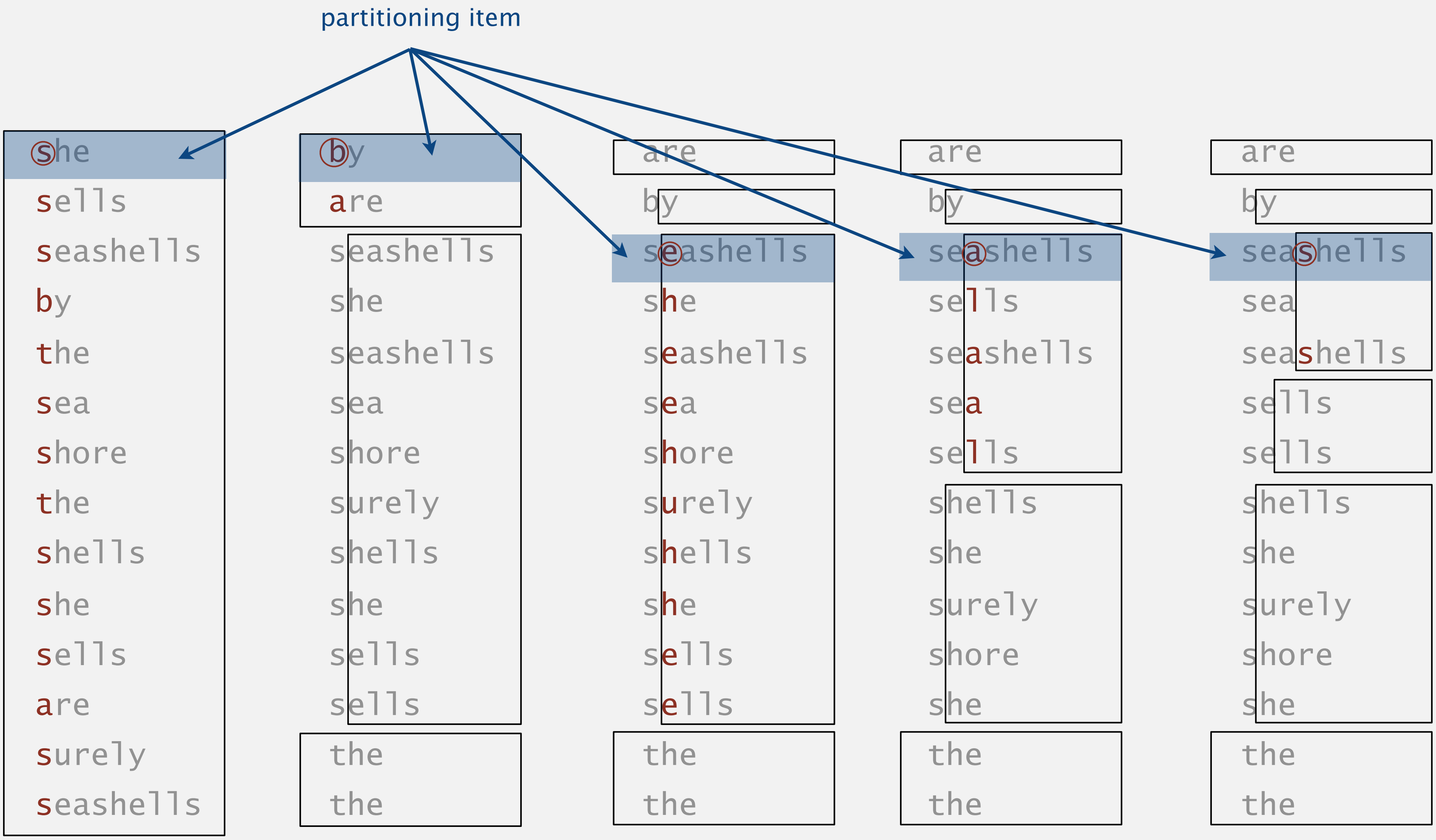
use first character to partition into "less", "equal", and "greater" subarrays

recursively sort subarrays, excluding first character for middle subarray

she
sells
seashells
by
the
sea
shore
the
shells
she
sells
are
surely
seashells

by
are
seashells
she
seashells
sea
shore
surely
shells
she
sells
sells
the
the

3-way string quicksort: trace of recursive calls



she	by	are	are	are
sells	are	by	by	by
seashells	seashells	seashells	seashells	seashells
by	she	she	sells	sea
the	seashells	seashells	seashells	seashells
sea	sea	sea	sea	sells
shore	shore	shore	sells	sells
the	surely	surely	shells	shells
shells	shells	shells	she	she
she	she	she	surely	surely
sells	sells	sells	shore	shore
are	sells	sells	she	she
surely	the	the	the	the
seashells	the	the	the	the

Trace of first few recursive calls for 3-way string quicksort (subarrays of size 1 not shown)

3-way string quicksort: Java implementation

```
private static void sort(String[] a)
{  sort(a, 0, a.length - 1, 0);  }

private static void sort(String[] a, int lo, int hi, int d)
{
   if (hi <= lo) return;
   int lt = lo, gt = hi;
   int v = charAt(a[lo], d);
   int i = lo + 1;
   while (i <= gt)
   {
      int t = charAt(a[i], d);
      if      (t < v) exch(a, lt++, i++);
      else if (t > v) exch(a, i, gt--);
      else            i++;
   }

   sort(a, lo, lt-1, d);
   if (v >= 0) sort(a, lt, gt, d+1);
   sort(a, gt+1, hi, d);
}
```

3-way partitioning
(using dth character)

to handle variable-length strings

sort 3 subarrays recursively

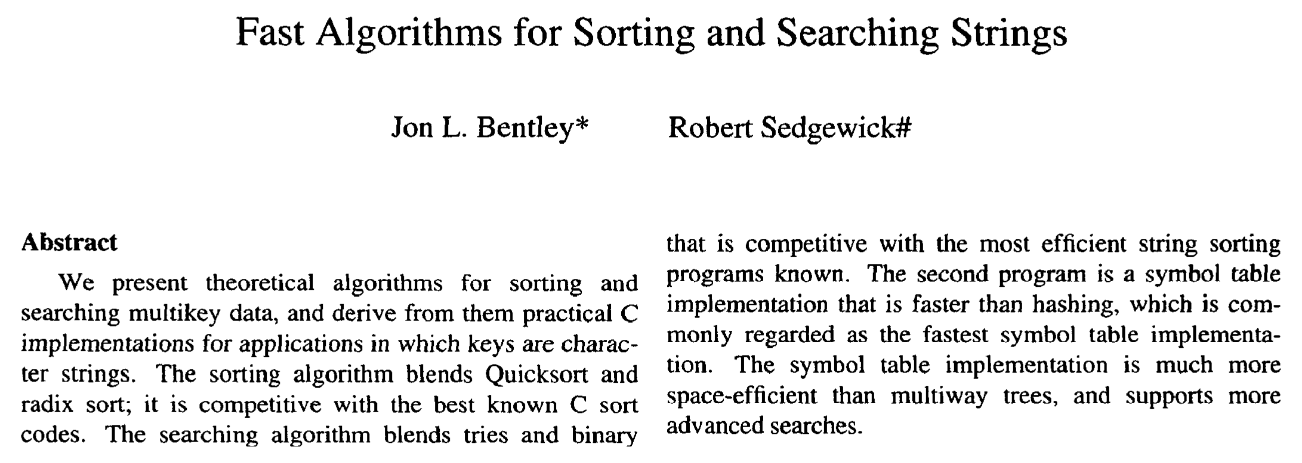
3-way string quicksort vs. standard quicksort

Standard quicksort.

- Uses $\sim 2 N \ln N$ string compares on average.
- Costly for keys with long common prefixes (and this is a common case!)

3-way string (radix) quicksort.

- Uses $\sim 2 N \ln N$ character compares on average for random strings.
- Avoids re-comparing long common prefixes.



Fast Algorithms for Sorting and Searching Strings

Jon L. Bentley*　　Robert Sedgewick#

Abstract

We present theoretical algorithms for sorting and searching multikey data, and derive from them practical C implementations for applications in which keys are character strings. The sorting algorithm blends Quicksort and radix sort; it is competitive with the best known C sort codes. The searching algorithm blends tries and binary search trees that is competitive with the most efficient string sorting programs known. The second program is a symbol table implementation that is faster than hashing, which is commonly regarded as the fastest symbol table implementation. The symbol table implementation is much more space-efficient than multiway trees, and supports more advanced searches.

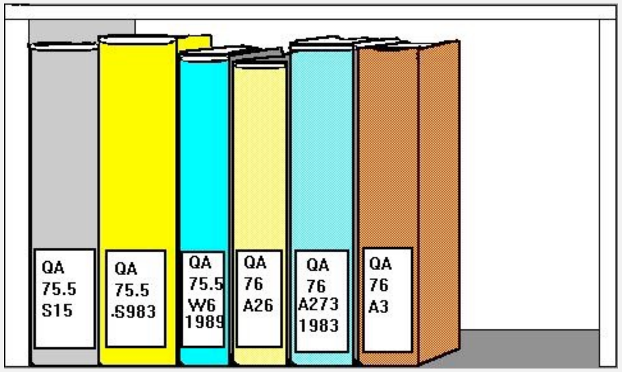
3-way string quicksort vs. MSD string sort

MSD string sort.

- Is cache-inefficient.
- Too much memory storing `count[]`.
- Too much overhead reinitializing `count[]` and `aux[]`.

3-way string quicksort.

- Is in-place.
- Is cache-friendly.
- Has a short inner loop.
- But not stable.



library of Congress call numbers

Bottom line. 3-way string quicksort is method of choice for sorting strings.

Summary of the performance of sorting algorithms

Frequency of operations.

algorithm	guarantee	random	extra space	stable?	operations on keys
insertion sort	$\frac{1}{2} N^2$	$\frac{1}{4} N^2$	1	✔	`compareTo()`
mergesort	$N \lg N$	$N \lg N$	N	✔	`compareTo()`
quicksort	$1.39 N \lg N$ *	$1.39 N \lg N$	$c \lg N$ *		`compareTo()`
heapsort	$2 N \lg N$	$2 N \lg N$	1		`compareTo()`
LSD sort †	$2 W (N + R)$	$2 W (N + R)$	$N + R$	✔	`charAt()`
MSD sort ‡	$2 W (N + R)$	$N \log_R N$	$N + D R$	✔	`charAt()`
3-way string quicksort	$1.39 W N \lg R$ *	$1.39 N \lg N$	$\log N + W$ *		`charAt()`

* probabilistic
† fixed-length W keys
‡ average-length W keys

http://algs4.cs.princeton.edu

5.1 STRING SORTS

- *strings in Java*
- *key-indexed counting*
- *LSD radix sort*
- *MSD radix sort*
- *3-way radix quicksort*
- *suffix arrays*

Keyword-in-context search

Given a text of N characters, preprocess it to enable fast substring search (find all occurrences of query string context).

```
% more tale.txt
it was the best of times
it was the worst of times
it was the age of wisdom
it was the age of foolishness
it was the epoch of belief
it was the epoch of incredulity
it was the season of light
it was the season of darkness
it was the spring of hope
it was the winter of despair
    ⋮
```

Keyword-in-context search

Given a text of N characters, preprocess it to enable fast substring search (find all occurrences of query string context).

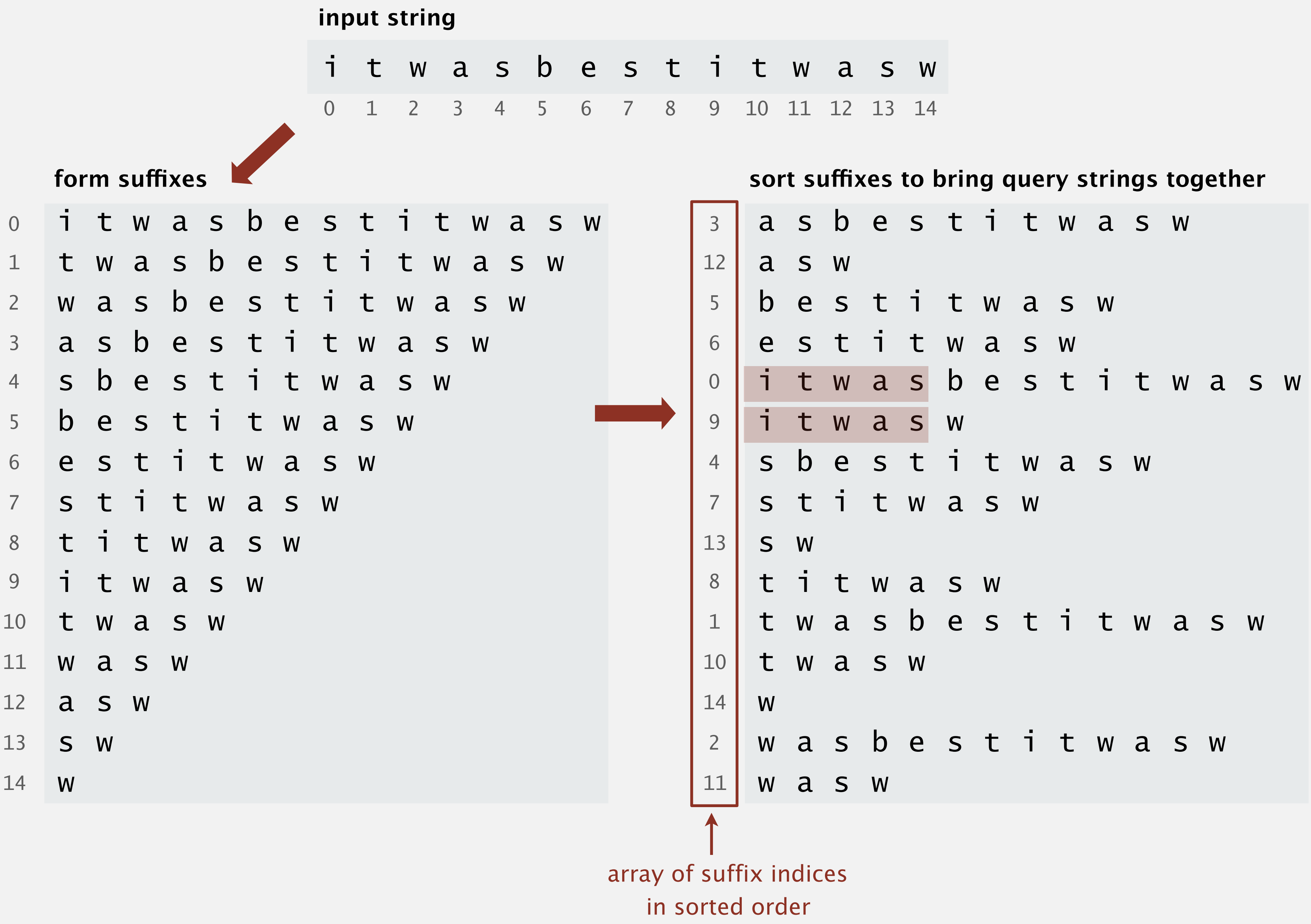
```
% java KWIC tale.txt 15
search
o st giless to search for contraband
her unavailing search for your fathe
le and gone in search of her husband
t provinces in search of impoverishe
 dispersing in search of other carri
n that bed and search the straw hold

better thing
t is a far far better thing that i do than
 some sense of better things else forgotte
was capable of better things mr carton ent
```

← characters of surrounding context

Applications. Linguistics, databases, web search, word processing, ….

Suffix sort



input string

i	t	w	a	s	b	e	s	t	i	t	w	a	s	w
0	1	2	3	4	5	6	7	8	9	10	11	12	13	14

form suffixes

0	i t w a s b e s t i t w a s w
1	t w a s b e s t i t w a s w
2	w a s b e s t i t w a s w
3	a s b e s t i t w a s w
4	s b e s t i t w a s w
5	b e s t i t w a s w
6	e s t i t w a s w
7	s t i t w a s w
8	t i t w a s w
9	i t w a s w
10	t w a s w
11	w a s w
12	a s w
13	s w
14	w

sort suffixes to bring query strings together

3	a s b e s t i t w a s w
12	a s w
5	b e s t i t w a s w
6	e s t i t w a s w
0	i t w a s b e s t i t w a s w
9	i t w a s w
4	s b e s t i t w a s w
7	s t i t w a s w
13	s w
8	t i t w a s w
1	t w a s b e s t i t w a s w
10	t w a s w
14	w
2	w a s b e s t i t w a s w
11	w a s w

array of suffix indices
in sorted order

Keyword-in-context search: suffix-sorting solution

- Preprocess: suffix sort the text.
- Query: binary search for query; scan until mismatch.

KWIC search for "search" in Tale of Two Cities

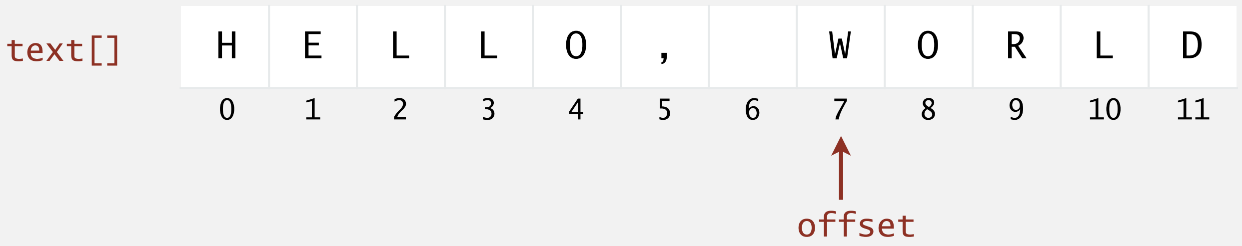
	⋮
632698	sealed_my_letter_and_…
713727	seamstress_is_lifted_…
660598	seamstress_of_twenty_…
67610	seamstress_who_was_wi…
4430	search_for_contraband…
42705	search_for_your_fathe…
499797	search_of_her_husband…
182045	search_of_impoverishe…
143399	search_of_other_carri…
411801	search_the_straw_hold…
158410	seared_marking_about_…
691536	seas_and_madame_defar…
536569	sease_a_terrible_pass…
484763	sease_that_had_brough…
	⋮

Suffix sort

Q. How to efficiently form (and sort) suffixes in Java 7u6?

A. Define `Suffix` class ala Java 7u5 `String`.

```
public class Suffix implements Comparable<Suffix>
{
   private final String text;
   private final int offset;
   public Suffix(String text, int offset)
   {
      this.text = text;
      this.offset = offset;
   }
   public int length()                { return text.length() - offset;   }
   public char charAt(int i)          { return text.charAt(offset + i);  }
   public int compareTo(Suffix that) { /* see textbook */                }
}
```



text[]	H	E	L	L	O	,		W	O	R	L	D
	0	1	2	3	4	5	6	7	8	9	10	11

Radix sorting: quiz 3

What is worst-case running time of our suffix array algorithm?

A. Quadratic.

B. Linearithmic.

C. Linear.

D. None of the above.

E. *I don't know.*

Hint: this is a worst-case input

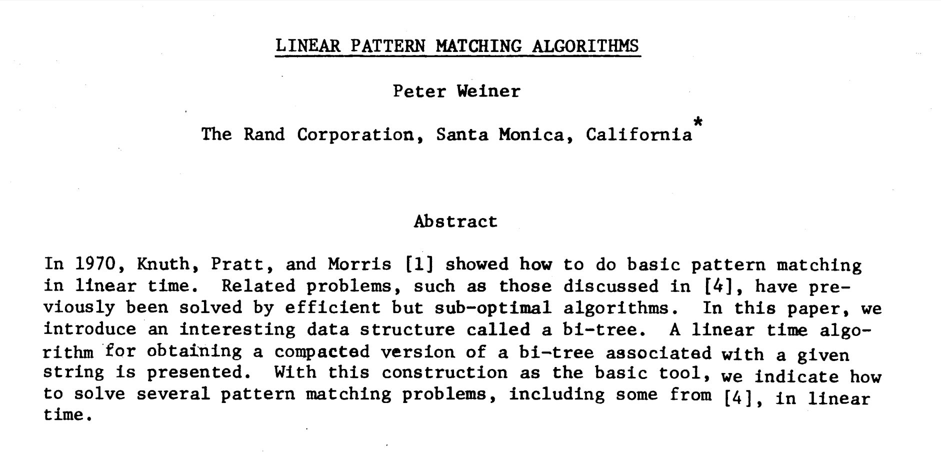
0	a a a a a a a a a a
1	a a a a a a a a a
2	a a a a a a a a
3	a a a a a a a
4	a a a a a a
5	a a a a a
6	a a a a
7	a a a
8	a a
9	a

Suffix arrays: theory

Conjecture (Knuth 1970). No linear-time algorithm.

Proposition. Linear-time algorithms (suffix trees).

> “ has no practical virtue… but a historic monument in the area of string processing. ”



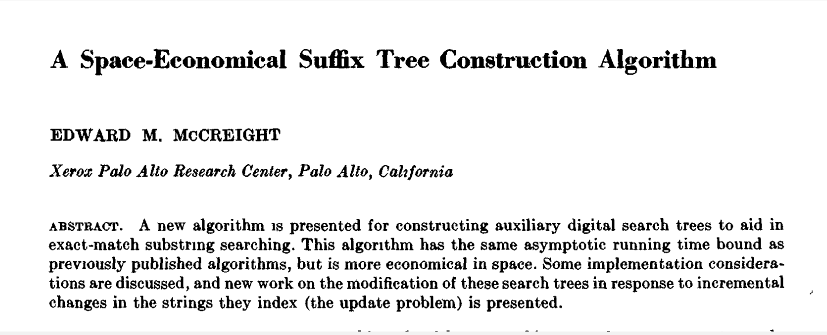
LINEAR PATTERN MATCHING ALGORITHMS

Peter Weiner

The Rand Corporation, Santa Monica, California*

Abstract

In 1970, Knuth, Pratt, and Morris [1] showed how to do basic pattern matching in linear time. Related problems, such as those discussed in [4], have previously been solved by efficient but sub-optimal algorithms. In this paper, we introduce an interesting data structure called a bi-tree. A linear time algorithm for obtaining a compacted version of a bi-tree associated with a given string is presented. With this construction as the basic tool, we indicate how to solve several pattern matching problems, including some from [4], in linear time.



A Space-Economical Suffix Tree Construction Algorithm

EDWARD M. McCREIGHT

Xerox Palo Alto Research Center, Palo Alto, California

ABSTRACT. A new algorithm is presented for constructing auxiliary digital search trees to aid in exact-match substring searching. This algorithm has the same asymptotic running time bound as previously published algorithms, but is more economical in space. Some implementation considerations are discussed, and new work on the modification of these search trees in response to incremental changes in the strings they index (the update problem) is presented.

On–line construction of suffix trees [1]

Esko Ukkonen

Department of Computer Science, University of Helsinki,
P. O. Box 26 (Teollisuuskatu 23), FIN–00014 University of Helsinki, Finland
Tel.: +358-0-7084172, fax: +358-0-7084441
Email: `ukkonen@cs.Helsinki.FI`

Suffix arrays: practice

Applications. Bioinformatics, information retrieval, data compression, ...

Many ingenious algorithms.

- Constants and memory footprint very important.
- State-of-the art still changing.

year	algorithm	worst case	memory
1991	**Manber–Myers**	$N \log N$	$8\,N$
1999	**Larsson–Sadakane**	$N \log N$	$8\,N$
2003	**Kärkkäinen–Sanders**	N	$13\,N$
2003	**Ko–Aluru**	N	$10\,N$
2008	**divsufsort2**	$N \log N$	$5\,N$
2010	**sais**	N	$6\,N$

see lecture videos (1991, Manber–Myers)

about 10× faster than Manber–Myers (1999, Larsson–Sadakane)

good choices (libdivsufsort) (divsufsort2, sais)

String sorting summary

We can develop linear-time sorts.

- Key compares not necessary for string keys.
- Use characters as index in an array.

We can develop sublinear-time sorts.

- Input size is amount of data in keys (not number of keys).
- Not all of the data has to be examined.

3-way string quicksort is asymptotically optimal.

- $1.39\, N \lg N$ chars for random data.

Long strings are rarely random in practice.

- Goal is often to learn the structure!
- May need specialized algorithms.