

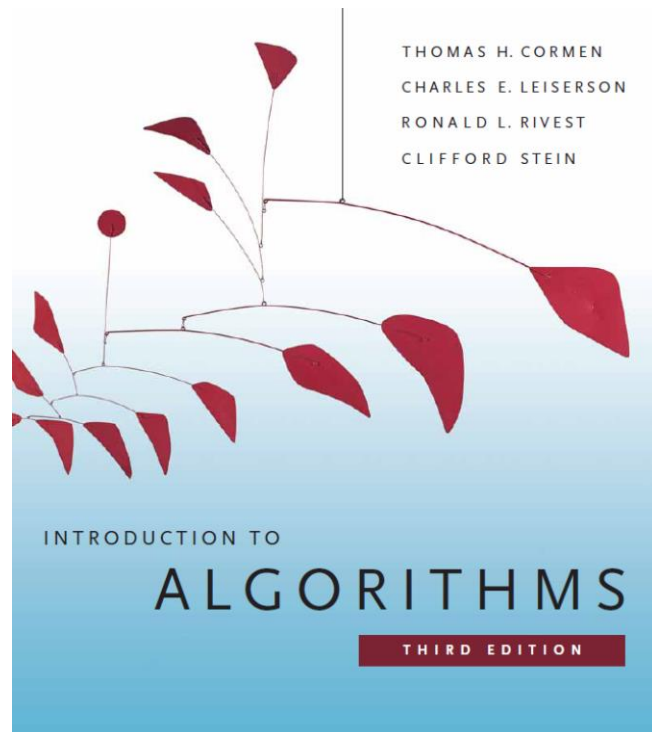
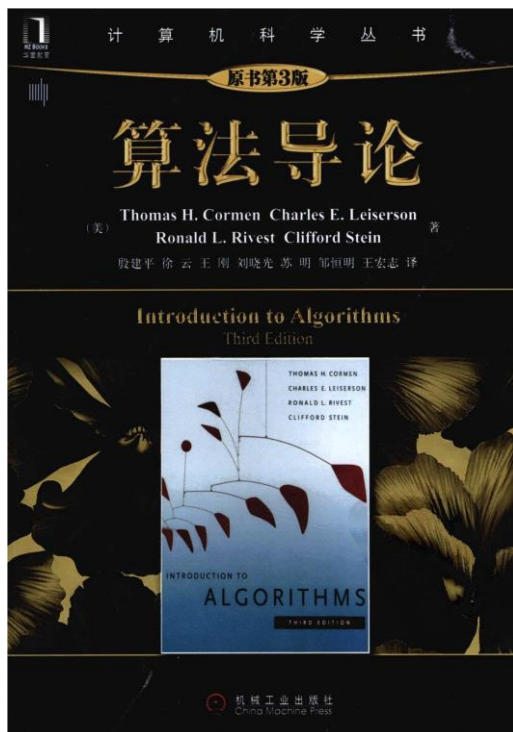
算法设计与分析

苏州大学 计算机科学与技术学院

汪笑宇

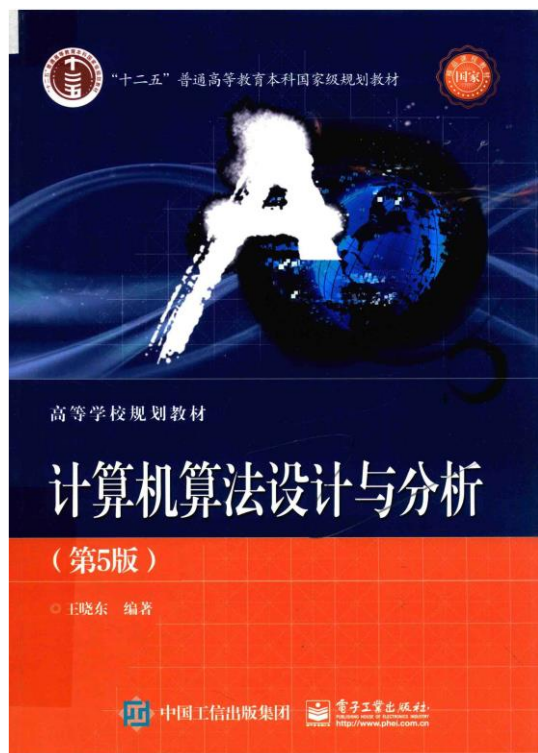
Email: xywang21@suda.edu.cn

教材



《算法导论》（第3版）

参考书目



王晓东 编著
《计算机算法设计与分析》(第5版)
电子工业出版社

课程总览

■ 本学期课程涉及内容：

- Chapter 1：算法在计算中的作用
- Chapter 2：算法基础
- Chapter 3：函数的渐近增长
- Chapter 34：NP完全性理论
- Chapter 4：分治策略
- Chapter 7：快速排序
- Chapter 9：中位数和顺序统计量
- Chapter 15：动态规划
- Chapter 16：贪心算法
- Chapter 35：近似算法
- 回溯法（参考书目第5章）
- 分支限界法（参考书目第6章）
- Chapter 5：随机算法

评分要求

■平时10%

■期中20%

■实验20%

■期末50%

作业及实验报告严禁抄袭！！！！

引入1

■ 假设有100个砝码外观完全相同，已知其中有一个是劣质的质量稍轻，而其他99个质量相同，如何用一台天平快速得出哪一个是劣质的砝码？

➤ 想法一：随机挑一个砝码作为基准，放在天平一侧，剩下的砝码依次放在天平另一侧进行称重

- 最坏情况：前98次称重天平都平衡
- 最好情况：第1次称重就天平不平衡

引入1 (续)

想法二：

- 第1次：天平一边50个，重的那一侧排除
- 第2次：天平一边25个，重的那一侧排除
- 第3次：天平一边12个，剩下1个，一样重则剩下的一个劣质；不一样重则重的一侧和剩下的一个排除
- 第4次：天平一边6个，重的那一侧排除
- 第5次：天平一边3个，重的那一侧排除
- 第6次：天平一边1个，剩下1个，一样重则剩下的一个劣质；否则哪边轻哪边劣质

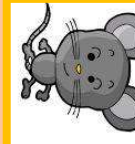
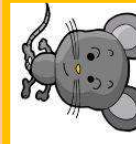
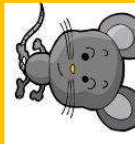
$$\text{二分法} \\ \lfloor \log_2 100 \rfloor = 6$$

引入2

■有100只一模一样的苹果，编号1-100。其中99个是正常的，一个被注射了毒药。只要老鼠咬一口毒苹果，**一天后则死亡**。现在，你有**7只**老鼠和**一天**的时间，如何检验出哪个号码的苹果被注射了毒药？



编号	binary	1	2	3	4	5	6	7
1	0000001	0	0	0	0	0	0	1
2	0000010	0	0	0	0	0	1	0
3	0000011	0	0	0	0	0	1	1
4	0000100	0	0	0	0	1	0	0
...								
...								
100	1100100	1	1	0	0	1	0	0



0010110
22号苹果有毒

第1章 算法基础

苏州大学 计算机科学与技术学院

汪笑宇

Email: xywang21@suda.edu.cn

本章内容

- 算法定义及基本概念（教材Chapter 1）
- 算法描述（教材Chapter 2）
- 函数增长及渐近记号表示（教材Chapter 3）
- 标准记号与常用函数（教材Chapter 3）
- NP完全性理论（区分并理解P/ NP/ NPC/ NP-Hard 几类问题）（教材Chapter 34）

本章内容

- 算法定义及基本概念（教材Chapter 1）
- 算法描述（教材Chapter 2）
- 函数增长及渐近记号表示（教材Chapter 3）
- 标准记号与常用函数（教材Chapter 3）
- NP完全性理论（区分并理解P/ NP/ NPC/ NP-Hard 几类问题）（教材Chapter 34）

基本概念——算法

■非形式定义

- 一个**算法**是任何一个**良定义(well-defined)**的计算过程，它接收某个值或值的集合作为输入，产生某个值或值的集合作为输出。因此，一个算法是一个**计算步骤的序列**，这些步骤将输入转化为输出。

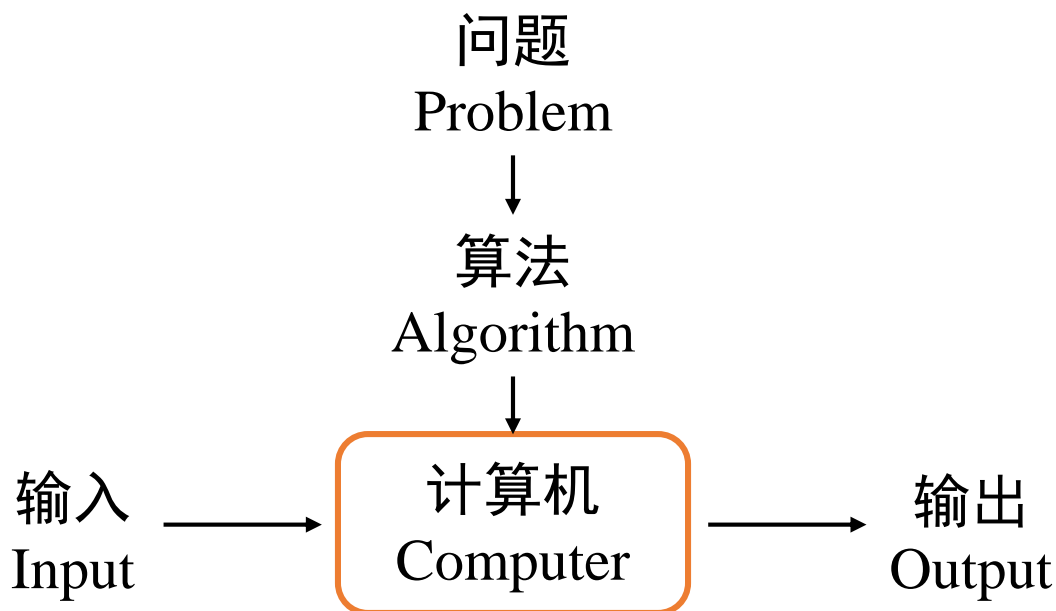


- 或者说，算法所描述的计算过程就是**怎样达到所期望的I/O关系**

基本概念——算法 (Algorithm)

■ 另一种定义

- An **algorithm** is a sequence of **unambiguous** instructions for solving a problem, i.e., for obtaining a required output for any **legitimate** input in a **finite** amount of time.



例：排序问题

■问题描述

- 输入：具有 n 个数的序列 $\langle a_1, a_2, \dots, a_n \rangle$
- 输出：输入序列的一个排列 $\langle a_1', a_2', \dots, a_n' \rangle$ ，满足 $a_1' \leq a_2' \leq \dots \leq a_n'$

■计算步骤

- 如何达到上述关系

基本概念

- **实例 (Instance)**: 一个问题的**实例**由计算该问题的一个解所需要的**所有输入**所组成
- **正确性**: 若对**每个**输入实例, 算法都以**正确的输出停机**, 则称该算法是**正确的**
 - 不正确的算法: 对某些输入实例不停机
以不正确的输出停机
 - **注**: 不正确的算法只要其错误率可控有时可能是有用的 (例如Chapter 31 大素数算法)
- **算法的描述**: 必须精确描述计算过程

算法 vs. 程序

■ 算法Algorithm \neq 程序Program

- Algorithm (webster.com): A procedure for solving a mathematical problem (as of finding the greatest common divisor) in a finite number of steps that frequently involves repetition of an operation.
- Broadly: a step-step procedure for solving a problem or accomplishing some end especially by a computer.
- Issues: correctness, efficiency (amount of work done and space used), storage (simplicity, clarity), optimality, etc.

算法 vs. 程序 (续)

■ **算法**是指解决问题的一种方法或一个过程，是若干**指令**的**有穷序列**，满足性质：

- **输入**：有外部提供的量作为算法的输入
- **输出**：算法产生至少一个量作为输出
- **确定性**：组成算法的每条指令是清晰、无歧义的
- **有限性**：算法中每条指令的执行次数是有限的，执行每条指令的时间也是有限的

注：正确性是前提

算法 vs. 程序 (续)

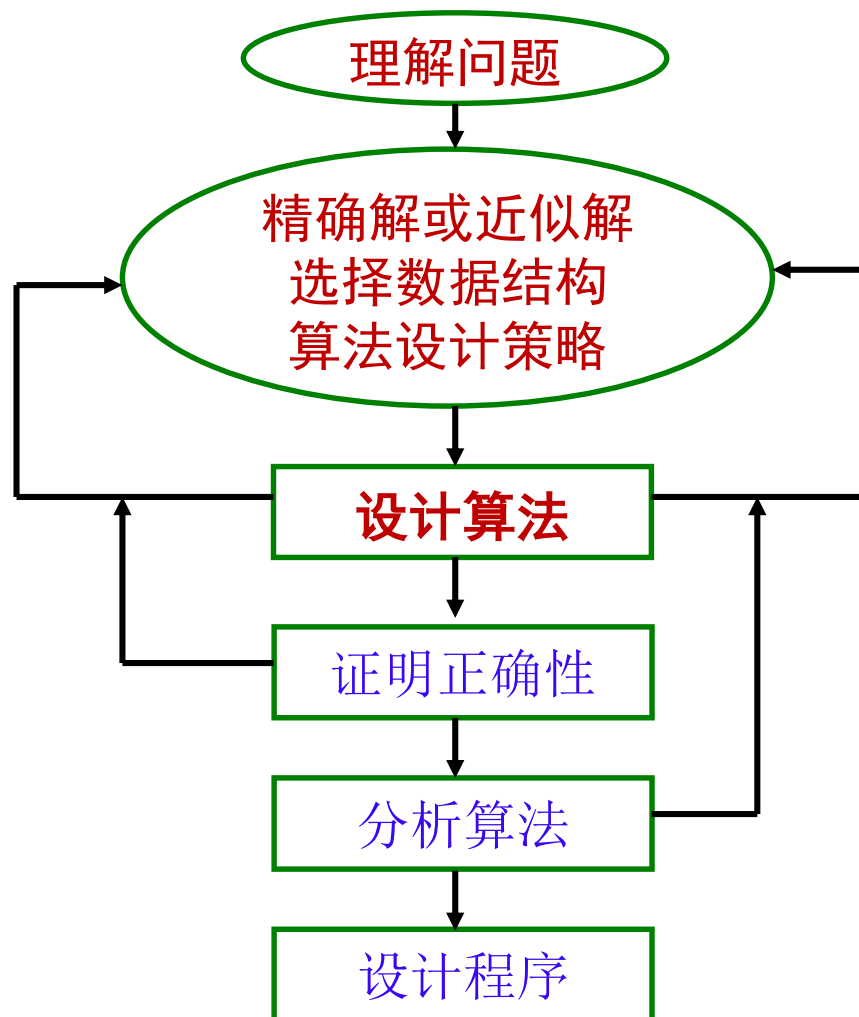
■程序

- 程序是算法用某种程序设计语言的具体实现
- 程序可以不满足有限性

例如：操作系统是一个在无限循环中执行的程序，因而不是一个算法。

操作系统的各种任务可看成是单独的问题，每一个问题由操作系统中的一个子程序通过特定的算法来实现。该子程序得到输出结果后便终止。

算法 -> 问题求解



算法解决哪些问题

■ 人类基因工程

- 识别10万个基因，确定30亿个化学碱基对序列

■ 互联网

- 为数据传输寻找好的路由
- 使用搜索引擎快速找到特定信息所在网页

■ 电子商务

- 公钥密码和数字签名

■ 制造业

- 按最有益的方式分配稀有资源

例：求最大公约数

■问题：求 m 和 n 的最大公约数 $\gcd(m, n)$ ，其中， m 和 n 为非负整数且不同时为0

➤例： $\gcd(60, 24) = 12$, $\gcd(8, 0) = 8$

■欧几里得算法 (Euclid's algorithm) 基于以下性质：

$$\gcd(m, n) = \gcd(n, m \bmod n)$$

直到第二个参数为0则停止

➤例： $\gcd(60, 24) = \gcd(24, 12) = \gcd(12, 0) = 12$

欧几里得算法的两种描述

步骤1：若 $n = 0$ ，则返回 m 并终止；
否则执行步骤2

步骤2： m 除以 n ，并将 r 赋值为余数

步骤3：将 n 的值赋给 m 、 r 的值赋给 n ，执行步骤1

Step 1: If $n = 0$, return m and stop;
otherwise go to Step 2.

Step 2: Divide m by n and assign the
value for the remainder to r .

Step 3: Assign the value of n to m and
the value of r to n . Go to Step 1.

```
Euclid( $m, n$ )  
  while  $n \neq 0$  do  
     $r \leftarrow m \bmod n$   
     $m \leftarrow n$   
     $n \leftarrow r$   
  return  $m$ 
```

求 $\gcd(m, n)$ 的其他算法

■例1：连续整数测试

Step 1: Assign the value of $\min\{m, n\}$ to t

Step 2: Divide m by t . If the remainder is 0, go to Step 3; otherwise, go to Step 4

Step 3: Divide n by t . If the remainder is 0, return t and stop; otherwise, go to Step 4

Step 4: Decrease t by 1 and go to Step 2



正确性?

求gcd(m, n)的其他算法 (续)

■例2：分解质因数

Step 1: Find the prime factorization of m

Step 2: Find the prime factorization of n

Step 3: Find all the common prime factors

Step 4: Compute the product of all the common prime factors and return it as gcd(m, n)



这是一个算法吗?

埃拉托色尼筛选法 (Sieve of Eratosthenes)

Input: Integer $n \geq 2$

Output: List of primes less than or equal to n

for $p \leftarrow 2$ to n do $A[p] \leftarrow p$

for $p \leftarrow 2$ to $\lfloor \sqrt{n} \rfloor$ do

 if $A[p] \neq 0$

$j \leftarrow p * p$

 while $j \leq n$ do

$A[j] \leftarrow 0$

$j \leftarrow j + p$

	2	3	4	5	6	7	8	9	10	Prime numbers
11	12	13	14	15	16	17	18	19	20	
21	22	23	24	25	26	27	28	29	30	
31	32	33	34	35	36	37	38	39	40	
41	42	43	44	45	46	47	48	49	50	
51	52	53	54	55	56	57	58	59	60	
61	62	63	64	65	66	67	68	69	70	
71	72	73	74	75	76	77	78	79	80	
81	82	83	84	85	86	87	88	89	90	
91	92	93	94	95	96	97	98	99	100	
101	102	103	104	105	106	107	108	109	110	
111	112	113	114	115	116	117	118	119	120	

算法的重要性

■问题：对 $n=10,000,000$ 个整数排序

➤计算机A：每秒执行**百亿**(10^{10})条指令，算法需要 $2n^2$ 条指令

$$\frac{2 \cdot (10^7)^2 \text{ 条指令}}{10^{10} \text{ 条指令/秒}} = 20000 \text{ 秒 (多于5.5小时)}$$

➤计算机B：每秒执行**千万**(10^7)条指令，算法需要 $50n \lg n$ 条指令

$$\frac{50 \cdot 10^7 \lg 10^7 \text{ 条指令}}{10^7 \text{ 条指令/秒}} \approx 1163 \text{ 秒 (少于20分钟)}$$

算法的重要性 (续)

Run time (nanoseconds)		$1.3 N^3$	$10 N^2$	$47 N \log_2 N$	$48 N$
Time to solve a problem of size	1000	1.3 seconds	10 msec	0.4 msec	0.048 msec
	10,000	22 minutes	1 second	6 msec	0.48 msec
	100,000	15 days	1.7 minutes	78 msec	4.8 msec
	million	41 years	2.8 hours	0.94 seconds	48 msec
	10 million	41 millennia	1.7 weeks	11 seconds	0.48 seconds
Max size problem solved in one	second	920	10,000	1 million	21 million
	minute	3,600	77,000	49 million	1.3 billion
	hour	14,000	600,000	2.4 billion	76 billion
	day	41,000	2.9 million	50 billion	1,800 billion
N multiplied by 10, time multiplied by		1,000	100	10+	10

本章内容

- 算法定义及基本概念（教材Chapter 1）
- **算法描述（教材Chapter 2）**
- 函数增长及渐近记号表示（教材Chapter 3）
- 标准记号与常用函数（教材Chapter 3）
- NP完全性理论（区分并理解P/ NP/ NPC/ NP-Hard 几类问题）（教材Chapter 34）

伪代码 (Pseudo code)

- 伪代码描述：更简洁地表达算法的本质，忽略细节
- 伪代码中的一些约定（书本p11）

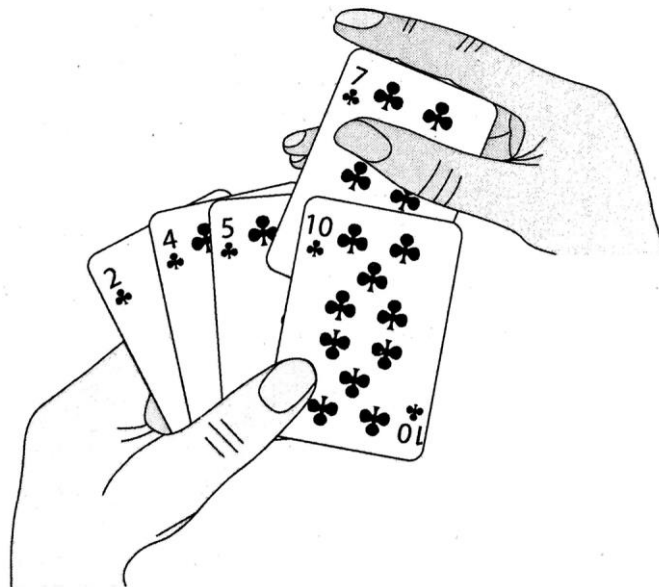
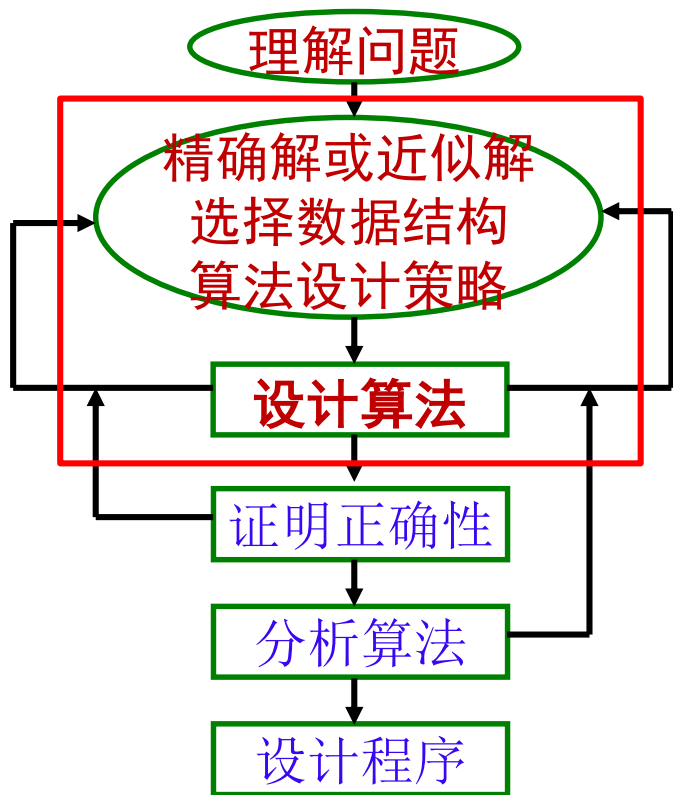
```
INSERTION_SORT(A)
1  for  $j \leftarrow 2$  to  $A.length$  do
2     $key \leftarrow A[j]$ 
3    // Insert  $A[j]$  into the sorted sequence  $A[1..j-1]$ 
4     $i \leftarrow j - 1$ 
5    while  $i > 0$  and  $A[i] > key$  do
6       $A[i+1] \leftarrow A[i]$ 
7       $i \leftarrow i - 1$ 
8     $A[i+1] \leftarrow key$ 
```

排序问题

- 输入：具有 n 个数的序列 $\langle a_1, a_2, \dots, a_n \rangle$
- 输出：输入序列的一个排列 $\langle a_1', a_2', \dots, a_n' \rangle$,
满足 $a_1' \leq a_2' \leq \dots \leq a_n'$
- 例：
 - 输入： $\langle 5, 2, 4, 6, 1, 3 \rangle$
 - 输出： $\langle 1, 2, 3, 4, 5, 6 \rangle$

排序问题求解：以插入排序为例

■精确解、线性表顺序存储结构、插入排序方法



插入排序执行过程

INSERTION_SORT(*A*)

1 **for** $j \leftarrow 2$ **to** $A.length$ **do**

2 $key \leftarrow A[j]$

3 // Insert $A[j]$ into the sorted sequence $A[1..j-1]$

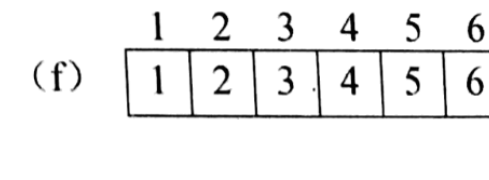
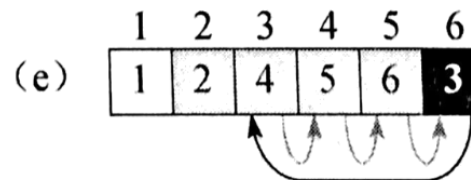
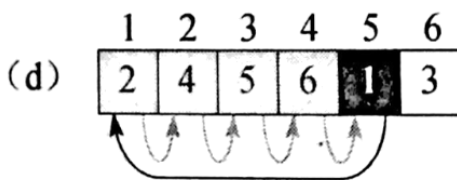
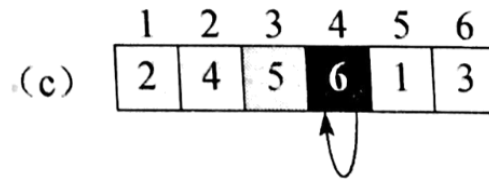
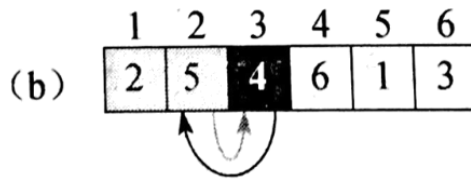
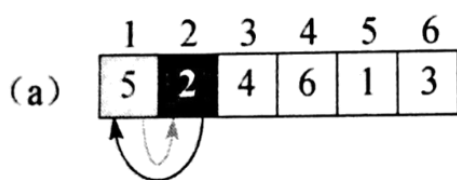
4 $i \leftarrow j - 1$

5 **while** $i > 0$ and $A[i] > key$ **do**

6 $A[i+1] \leftarrow A[i]$

7 $i \leftarrow i - 1$

8 $A[i+1] \leftarrow key$



■循环不变式：循环体每次执行前后均为真的谓词

➤作用：主要用来帮助我们理解算法的正确性。

■循环不变式性质：

➤初始化：循环的第一次迭代之前，它为真

➤保持：如果循环的某次迭代之前它为真，那么下次迭代之前它仍为真

➤终止：在循环终止时，不变式为我们提供一个有用的性质，该性质有助于证明算法是正确的

循环不变式 (续)

■前两条性质与我们熟知的数学归纳法很类似

■数学归纳法：

- 1、证明当 $n = 1$ 时命题成立。（基本情况）
- 2、假设 $n = m$ 时命题成立，那么可以推导出在 $n = m + 1$ 时命题也成立。（ m 代表任意正整数）（归纳步）

➤循环不变式“初始化”对应于“基本情况”

➤循环不变式“保持”对应于“归纳步”

➤区别在于：数学归纳法中的归纳步无限的使用，而循环不变式中当循环终止时，停止“归纳”

插入排序的正确性证明

INSERTION_SORT(A)

1 **for** $j \leftarrow 2$ **to** $A.length$ **do**

2 $key \leftarrow A[j]$

3 // Insert $A[j]$ into the sorted sequence $A[1..j-1]$

4 $i \leftarrow j - 1$

5 **while** $i > 0$ and $A[i] > key$ **do**

6 $A[i+1] \leftarrow A[i]$

7 $i \leftarrow i - 1$

8 $A[i+1] \leftarrow key$

循环不变式：
子数组 $A[1..j-1]$
有序

■ **初始化：** 首先证明在第一次循环迭代之前 (当 $j=2$ 时), 循环不变式成立。子数组 $A[1..j-1]$ 仅由单个元素 $A[1]$ 组成, 实际上就是 $A[1]$ 中原来的元素。而且该子数组是排序好的 (当然很平凡)。这表明第一次循环迭代之前循环不变式成立。

插入排序的正确性证明 (续)

```
INSERTION_SORT(A)
```

```
1  for  $j \leftarrow 2$  to  $A.length$  do
```

```
2     $key \leftarrow A[j]$ 
```

```
3    // Insert  $A[j]$  into the sorted sequence  $A[1..j-1]$ 
```

```
4     $i \leftarrow j - 1$ 
```

```
5    while  $i > 0$  and  $A[i] > key$  do
```

```
6         $A[i+1] \leftarrow A[i]$ 
```

```
7         $i \leftarrow i - 1$ 
```

```
8     $A[i+1] \leftarrow key$ 
```

■ **保持：** 非形式化地，for循环体的第4~7行将 $A[j-1]$ 、 $A[j-2]$ 、 $A[j-3]$ 等向右移动一个位置，直到找到 $A[j]$ 的适当位置，第8行将 $A[j]$ 的值插入该位置。这时子数组 $A[1..j]$ 由原来在 $A[1..j]$ 中的元素组成，但已按序排列。那么对for循环的下一迭代增加 j 将保持循环不变式。

插入排序的正确性证明 (续)

```
INSERTION_SORT(A)
```

```
1  for  $j \leftarrow 2$  to  $A.length$  do
```

```
2     $key \leftarrow A[j]$ 
```

```
3    // Insert  $A[j]$  into the sorted sequence  $A[1..j-1]$ 
```

```
4     $i \leftarrow j - 1$ 
```

```
5    while  $i > 0$  and  $A[i] > key$  do
```

```
6       $A[i+1] \leftarrow A[i]$ 
```

```
7       $i \leftarrow i - 1$ 
```

```
8     $A[i+1] \leftarrow key$ 
```

■ **终止：** 导致for循环终止的条件是 $j > A.length = n$ 。因为每次循环迭代 j 增加1，那么必有 $j = n + 1$ 。在循环不变式的表述中将 j 用 $n + 1$ 代替，我们有：子数组 $A[1..n]$ 由原来在 $A[1..n]$ 中的元素组成，但已按序排列。注意到，子数组 $A[1..n]$ 就是整个数组，我们推断出整个数组已排序。因此算法正确。

插入排序的正确性证明 (续)

```
INSERTION_SORT(A)
1  for  $j \leftarrow 2$  to  $A.length$  do
2     $key \leftarrow A[j]$ 
3    // Insert  $A[j]$  into the sorted sequence  $A[1..j-1]$ 
4     $i \leftarrow j - 1$ 
5    while  $i > 0$  and  $A[i] > key$  do
6       $A[i+1] \leftarrow A[i]$ 
7       $i \leftarrow i - 1$ 
8     $A[i+1] \leftarrow key$ 
```

- 注：第二条性质的一种更形式化的处理要求我们对第5~7行的while循环给出并证明一个循环不变式
- 非形式化的分析来证明第二条性质对外层循环成立
- 思考：第5~7行while循环的循环不变式是什么？

- 目的：分析算法就是**估计**算法所需**资源** (时间，空间，通信带宽等)
- 计算模型：单处理机，RAM (Random Access Machine, 随机存取机)模型，其中指令是顺序执行的，无并发操作
- 涉及的知识基础
 - 离散组合数学、概率论、代数等(分析)
 - 程序设计、数据结构(算法设计)

Analyzing algorithms

- We shall assume a generic one-processor, random-access machine (RAM) model of computation.
 - Instructions are executed one after another, with no concurrent operations.
 - Each time, an instruction of a program is executed as an atom operation. An instruction includes arithmetic operations, logical operations, data movement and control operations.
 - Each such instruction takes a constant amount of time.
 - RAM capacity is large enough.

算法分析 (续)

■时间分析：算法耗费时间与输入规模和实例的构成有关

例：插入排序如同手牌整理，整理速度与纸牌数量和牌本身是否有序相关

➤输入规模：通常用整数表示，取决于被研究的问题

- 例1：数组排序：项数
- 例2：两数相乘：最好的度量是总的位数
- 有时需用两个或多个整数表示输入规模，如图的顶点和边

算法分析 (续)

■时间分析

➤运行时间

- 用基本操作的数目(执行步数)来度量；(好处是算法分析独立于机器，即任何基本操作看作是单位时间)
- 用更接近实际的计算机上实现的时间来度量；(如RAM模型，不同的指令具有不同的执行时间)

两者相差一个常数因子

➤最坏情况运行时间

- 输入规模为 n 时，任何输入的最长运行时间

算法分析 (续)

■时间分析

➤为何要分析算法的最坏运行时间？(p15)

- 它是算法对于任何输入的运行时间的上界
- 对于某些算法，最坏情况常常发生，如在DB中搜索一个并不存在的记录
- 平均运行时间往往和最坏运行时间相当，仅常数因子不同 (存在反例！)

➤平均运行时间

- 常常假定一个给定输入规模的所有输入是等概率的
- 这种可能并不一定成立，但可以用随机化算法强迫它成立

➤有时平均时间和最坏时间不是同数量级，算法选择依据： 最好、最坏的概率较小时，尽量选择平均时间较小的算法

算法分析 (续)

■最坏情况 (Worst case, 通常)

➤ $T(n)$ = **maximum time** of algorithm on any input of size n

■平均情况 (Average case, 有时)

➤ $T(n)$ = **expected time** of algorithm over all inputs of size n

➤ Need assumption of statistics distribution of inputs

■最好情况 (Best case, 假象)

➤ Cheat with a slow algorithm that works fast on some input

插入排序算法分析

	代价	次数
INSERTION_SORT(A)		
1 for $j \leftarrow 2$ to $A.length$ do	c_1	n
2 $key \leftarrow A[j]$	c_2	$n-1$
3 // Insert $A[j]$ into the sorted sequence $A[1..j-1]$	0	$n-1$
4 $i \leftarrow j - 1$	c_4	$n-1$
5 while $i > 0$ and $A[i] > key$ do	c_5	$\sum_{j=2}^n t_j$
6 $A[i+1] \leftarrow A[i]$	c_6	$\sum_{j=2}^n (t_j - 1)$
7 $i \leftarrow i - 1$	c_7	$\sum_{j=2}^n (t_j - 1)$
8 $A[i+1] \leftarrow key$	c_8	$n-1$

t_j : 对于值 j , 第5行执行while循环测试的次数

插入排序算法分析 (续)

■ 算法执行时间

$$\begin{aligned} T(n) = & c_1 n + c_2(n-1) + c_4(n-1) \\ & + c_5 \sum_{j=2}^n t_j + c_6 \sum_{j=2}^n (t_j - 1) + c_7 \sum_{j=2}^n (t_j - 1) + c_8(n-1) \end{aligned}$$

➤ **最好情况：** 顺序排列—— $t_j = 1$

$$\begin{aligned} T(n) &= c_1 n + c_2(n-1) + c_4(n-1) + c_5(n-1) + c_8(n-1) \\ &= (c_1 + c_2 + c_4 + c_5 + c_8)n - (c_2 + c_4 + c_5 + c_8) \\ &= \Theta(n) \end{aligned}$$

- 执行时间是 n 的线性函数

插入排序算法分析 (续)

■ 算法执行时间

➤ 最坏情况：逆序排列

$$\sum_{j=2}^n t_j = \sum_{j=2}^n j = \frac{n(n+1)}{2} - 1$$

$$\sum_{j=2}^n (t_j - 1) = \sum_{j=2}^n (j - 1) = \frac{n(n-1)}{2}$$

$$T(n) = c_1 n + c_2(n-1) + c_4(n-1)$$

$$+ c_5\left(\frac{n(n+1)}{2} - 1\right) + c_6\left(\frac{n(n-1)}{2}\right) + c_7\left(\frac{n(n-1)}{2}\right) + c_8(n-1)$$

$$= \left(\frac{c_5}{2} + \frac{c_6}{2} + \frac{c_7}{2}\right)n^2 + \left(c_1 + c_2 + c_4 + \frac{c_5}{2} - \frac{c_6}{2} - \frac{c_7}{2} + c_8\right)n - (c_2 + c_4 + c_5 + c_8)$$

$$= \Theta(n^2)$$

- 执行时间是 n 的二次函数

插入排序算法分析 (续)

■ 算法执行时间

➤ **平均情况**：在 $A[1..j-1]$ 中一半大于 $A[j]$ 、一半小于 $A[j]$

➤ t_j 约为 $j/2$

➤ $T(n) = \Theta(n^2)$

- 执行时间是 n 的二次函数