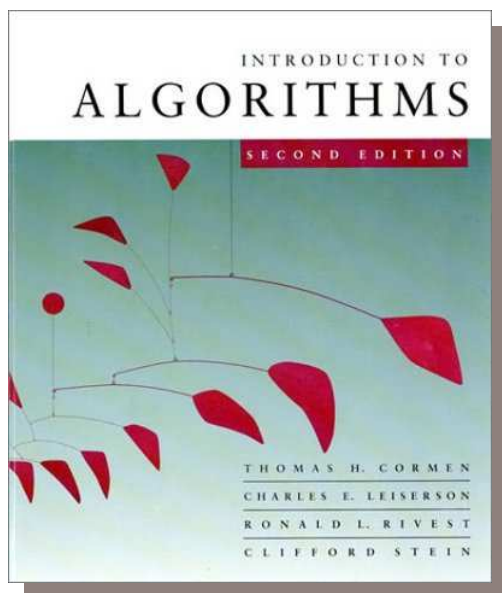


# *Data Structures and Algorithms*

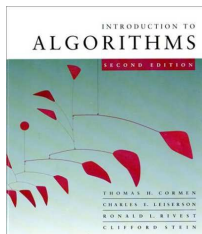
## *CSE 465*



### **LECTURE 11**

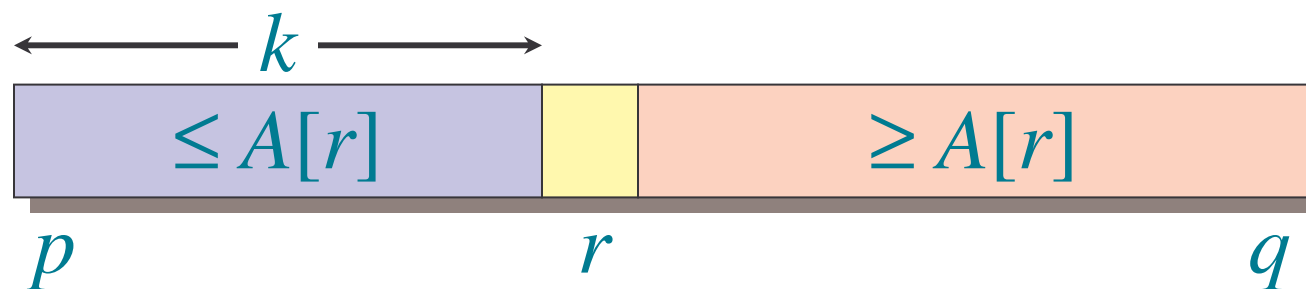
- **Order Statistics**
- **Randomized Selection**
- **Analysis of expected running time**

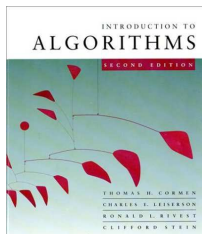
**Sofya Raskhodnikova and Adam Smith**



# Recall: randomized selection

**RAND-SELECT**( $A, p, q, i$ )  $\triangleright$   $i$ th smallest of  $A[p..q]$   
**if**  $p = q$  **then return**  $A[p]$   
 $r \leftarrow$  **RAND-PARTITION**( $A, p, q$ )  
 $k \leftarrow r - p + 1$   $\triangleright k = \text{rank}(A[r])$   
**if**  $i = k$  **then return**  $A[r]$   
**if**  $i < k$   
**then return** **RAND-SELECT**( $A, p, r - 1, i$ )  
**else return** **RAND-SELECT**( $A, r + 1, q, i - k$ )





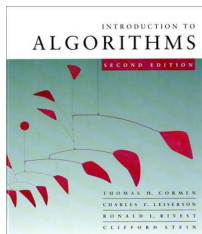
# Analysis of expected time

The analysis follows that of randomized quicksort, but it's a little different.

Throughout the analysis we assume that all array elements are **distinct**.

Let  $T(n)$  = the random variable for the running time of RAND-SELECT on an input of size  $n$ , assuming random numbers are independent.

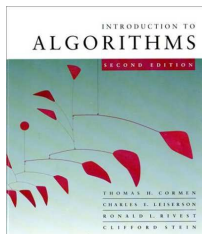
***Our goal: find the expectation  $E[T(n)]$ .***



# Analysis (continued)

To obtain an upper bound, assume that the  $i$ th element always falls in the larger side of the partition:

$$T(n) \leq \begin{cases} T(\max\{0, n-1\}) + \Theta(n) & \text{if } 0 : n-1 \text{ split,} \\ T(\max\{1, n-2\}) + \Theta(n) & \text{if } 1 : n-2 \text{ split,} \\ \vdots & \\ T(\max\{n-1, 0\}) + \Theta(n) & \text{if } n-1 : 0 \text{ split,} \end{cases}$$
$$= \sum_{k=0}^{n-1} X_k \cdot T(\max\{k, n-k-1\}) + \Theta(n)$$



# Review question

As in the analysis of Quicksort,

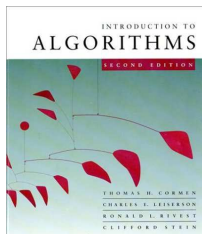
for  $k = 0, 1, \dots, n-1$ ,

define the *indicator random variable*

$$X_k = \begin{cases} 1 & \text{if PARTITION generates a } k : n-k-1 \text{ split,} \\ 0 & \text{otherwise.} \end{cases}$$

What is the expectation of  $X_k$ ?

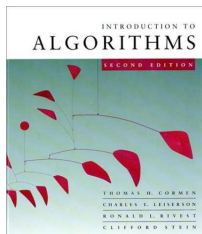
**Answer:**  $E[X_k] = 1/n$ .



# Calculating expectation

$$E[T(n)] = E \left[ \sum_{k=0}^{n-1} X_k (T(\max\{k, n-k-1\}) + \Theta(n)) \right]$$

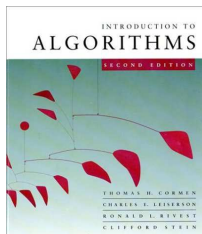
Take expectations of both sides.



# Calculating expectation

$$\begin{aligned} E[T(n)] &= E\left[\sum_{k=0}^{n-1} X_k (T(\max\{k, n-k-1\}) + \Theta(n))\right] \\ &= \sum_{k=0}^{n-1} E[X_k (T(\max\{k, n-k-1\}) + \Theta(n))] \end{aligned}$$

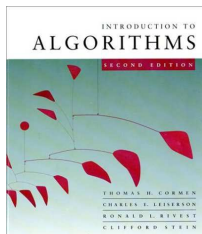
Linearity of expectation.



# Calculating expectation

$$\begin{aligned} E[T(n)] &= E\left[\sum_{k=0}^{n-1} X_k (T(\max\{k, n-k-1\}) + \Theta(n))\right] \\ &= \sum_{k=0}^{n-1} E[X_k (T(\max\{k, n-k-1\}) + \Theta(n))] \\ &= \sum_{k=0}^{n-1} E[X_k] \cdot E[T(\max\{k, n-k-1\}) + \Theta(n)] \end{aligned}$$

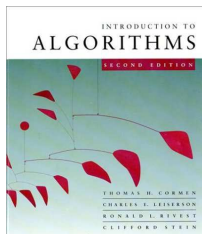
Independence of  $X_k$  from other random choices.



# Calculating expectation

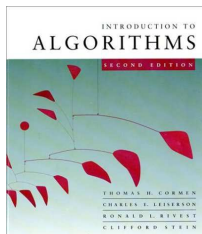
$$\begin{aligned} E[T(n)] &= E\left[\sum_{k=0}^{n-1} X_k (T(\max\{k, n-k-1\}) + \Theta(n))\right] \\ &= \sum_{k=0}^{n-1} E[X_k (T(\max\{k, n-k-1\}) + \Theta(n))] \\ &= \sum_{k=0}^{n-1} E[X_k] \cdot E[T(\max\{k, n-k-1\}) + \Theta(n)] \\ &= \frac{1}{n} \sum_{k=0}^{n-1} E[T(\max\{k, n-k-1\})] + \frac{1}{n} \sum_{k=0}^{n-1} \Theta(n) \end{aligned}$$

Linearity of expectation;  $E[X_k] = 1/n$ .



# Calculating expectation

$$\begin{aligned} E[T(n)] &= E\left[\sum_{k=0}^{n-1} X_k (T(\max\{k, n-k-1\}) + \Theta(n))\right] \\ &= \sum_{k=0}^{n-1} E[X_k (T(\max\{k, n-k-1\}) + \Theta(n))] \\ &= \sum_{k=0}^{n-1} E[X_k] \cdot E[T(\max\{k, n-k-1\}) + \Theta(n)] \\ &= \frac{1}{n} \sum_{k=0}^{n-1} E[T(\max\{k, n-k-1\})] + \frac{1}{n} \sum_{k=0}^{n-1} \Theta(n) \\ &\leq \frac{1}{n} \sum_{k=0}^{n-1} E[T(\max\{k, n-k-1\})] + \Theta(n) \end{aligned}$$



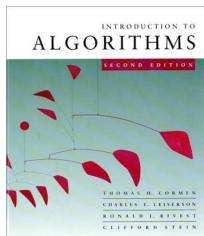
# Simplifying the expression

We will use **monotonicity** of  $T$ : the fact that  $E[T(n_1)] \leq E[T(n_2)]$  when  $n_1 < n_2$ .

Consider two cases:

- when the split is OK ( $n/4 \cdot k \cdot 3n/4$ )
- when the split is not OK (the rest)

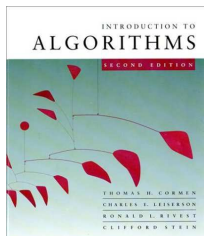
$$\max\{k, n-k-1\} \leq \begin{cases} 3n/4 & \text{if } n/4 \leq k \leq 3n/4, \\ n & \text{otherwise.} \end{cases}$$



# Simplifying the expression for $E[T(n)]$

$$\begin{aligned} E[T(n)] &\leq \frac{1}{n} \sum_{k=n/4}^{3n/4} E\left[T\left(\frac{3n}{4}\right)\right] + \frac{1}{n} \sum_{k:0 \leq k \leq n/4 \text{ or } 3n/4 \leq k \leq n-1} E[T(n)] + \Theta(n) \\ &= \frac{1}{2} E\left[T\left(\frac{3n}{4}\right)\right] + \frac{1}{2} E[T(n)] + \Theta(n) \end{aligned}$$

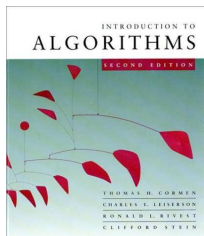
Break the sum into two sums: the sum over OK splits and the sum over the rest.  
Substitute the bound on max.



# Simplifying the expression for $E[T(n)]$

$$\begin{aligned} E[T(n)] &\leq \frac{1}{n} \sum_{k=n/4}^{3n/4} E\left[T\left(\frac{3n}{4}\right)\right] + \frac{1}{n} \sum_{k:0 \leq k \leq n/4 \text{ or } 3n/4 \leq k \leq n-1} E[T(n)] + \Theta(n) \\ &= \frac{1}{2} E\left[T\left(\frac{3n}{4}\right)\right] + \frac{1}{2} E[T(n)] + \Theta(n) \\ 2E[T(n)] &\leq E\left[T\left(\frac{3n}{4}\right)\right] + E[T(n)] + \Theta(n) \end{aligned}$$

Multiply both sides by 2.



# Simplifying the expression for $E[T(n)]$

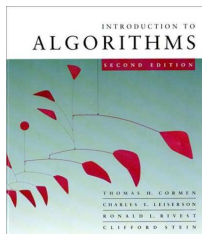
$$E[T(n)] \leq \frac{1}{n} \sum_{k=n/4}^{3n/4} E\left[T\left(\frac{3n}{4}\right)\right] + \frac{1}{n} \sum_{k:0 \leq k \leq n/4 \text{ or } 3n/4 \leq k \leq n-1} E[T(n)] + \Theta(n)$$

$$= \frac{1}{2} E\left[T\left(\frac{3n}{4}\right)\right] + \frac{1}{2} E[T(n)] + \Theta(n)$$

$$2E[T(n)] \leq E\left[T\left(\frac{3n}{4}\right)\right] + E[T(n)] + \Theta(n)$$

$$E[T(n)] \leq E\left[T\left(\frac{3n}{4}\right)\right] + \Theta(n)$$

Simplify.



# Use Master Theorem

- Let  $F(n)$  denote  $E[T(n)]$ .

- We get

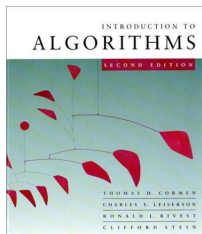
$$F(n) \leq F\left(\frac{3n}{4}\right) + \Theta(n)$$

- By Master Theorem, case 2,

$$F(n) = O(n)$$

- Since RAND-PARTITION takes time  $\Omega(n)$ ,  
RAND-SELECT also takes  $\Omega(n)$  time.

$$E[T(n)] = \Theta(n)$$



# Summary of randomized order-statistic selection

- Works fast: linear expected time.
- Excellent algorithm in practice.
- But, the worst case is *very* bad:  $\Theta(n^2)$ .

**Q.** Is there an algorithm that runs in linear time in the worst case?

**A.** Yes, due to Blum, Floyd, Pratt, Rivest, and Tarjan [1973].