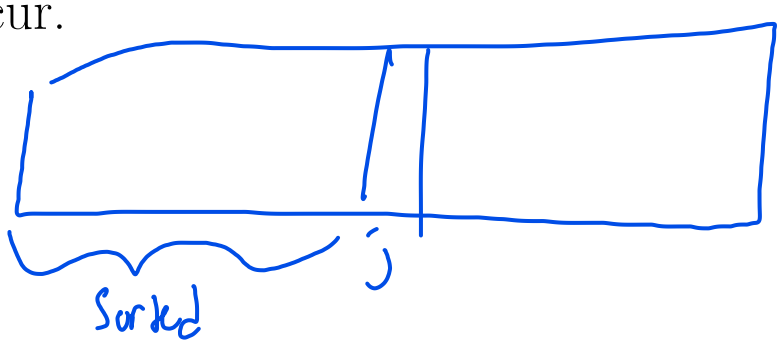
Last time we considered the average case running time of a deterministic algorithm.

That is, we assumed that the input was uniformly distributed (any permutation of the input is equally likely to occur in the case of the hiring problem), and we analyzed the running time of the algorithm of a randomly chosen input.

Consider the average case of insertion sort when any permutation of the numbers in the array is equally likely to occur.



On average, we will make $j/2$ swaps.

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

arithmetic series

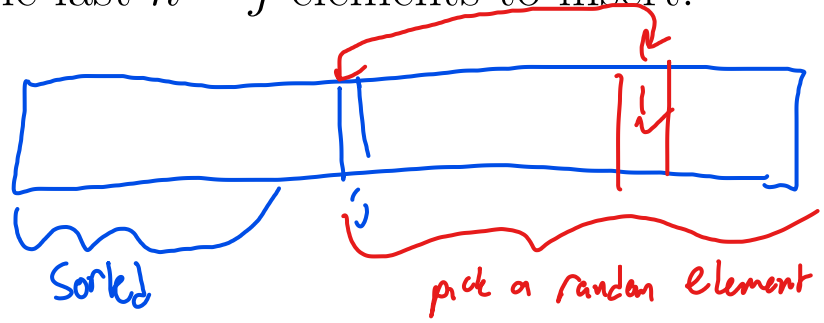
There are a few drawbacks to analyzing algorithms in this way.

- The distribution of inputs are heavily reliant on the application. Therefore the analysis may be misleading for certain applications.
- For many applications, it is difficult to determine if the inputs come from some "well-behaved" distribution, and if we are wrong about the distribution then our analysis of the running time could be significantly off base.

One approach to deal with this would be to "shuffle" the input before running the algorithm. Essentially we can force the input to come from a known distribution. Of course the downside to this is we pay a price in the running time of the algorithm.

A better approach to deal with this is to use a **randomized algorithm**. That is, an algorithm which uses a random number generator to determine some of the steps of the algorithm.

- For example in insertion sort, instead of inserting the element in position j, randomly choose a one of the last $n - j$ elements to insert.



The running time of a randomized algorithm is a random variable. We are interested in determining the **expected running time** of the algorithm. That is, let X be the random variable that measures the running time of a randomized algorithm. We are interested in bounding $E[X]$.

Now we do not need to make any assumptions about the input distribution.

No specific input will force the algorithm's worst-case behavior (although no input will force the algorithm's best-case behavior either). The worst-case (and best-case) are determined only by the output of a random number generator.

For these reasons, it is generally considered better to analyze the expected running time of a randomized algorithm rather than the average case running time of a deterministic algorithm.

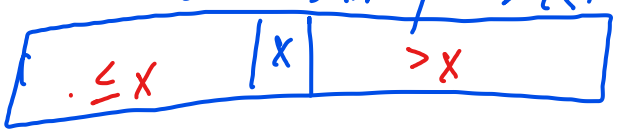
Quicksort: A divide-and-conquer sorting algorithm (like mergesort).

Quicksort is an "in place" algorithm (limited extra space is needed, unlike mergesort).

We are going to first analyze the deterministic quicksort algorithm, and then perform an expected running time analysis on the randomized version of quicksort.

Quicksort

1) Divide: Partition array into two subarrays around a pivot x such that elements in first subarray $\leq x$ and elements in second subarray $> x$.

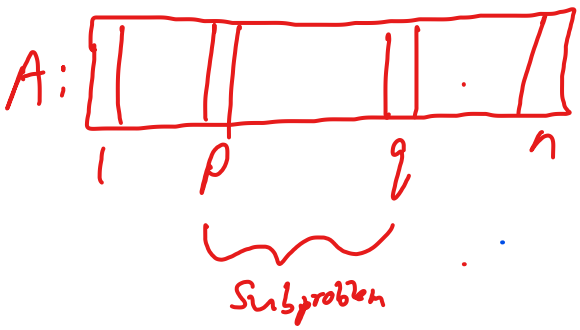


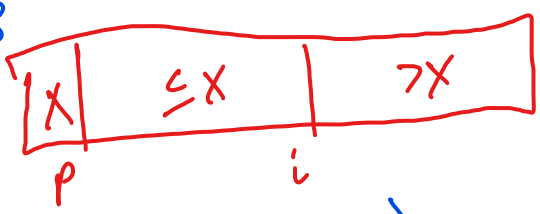
$O(n)$ time.

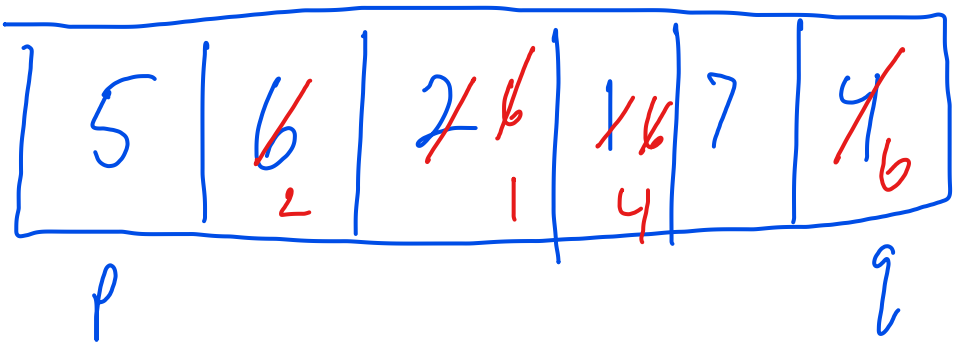
2) Conquer: Recursively sort two subarrays.

3) Combine: trivial

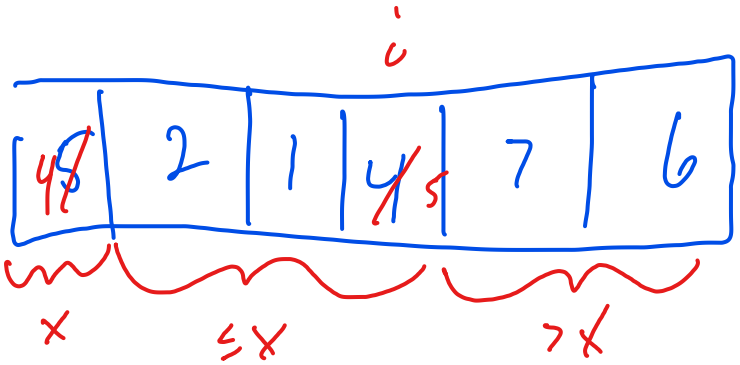
```
Partition(A, p, q){
    x = A[p]
    i = p
    for(j = p+1 to q)
    {
        if(A[j] ≤ x)
        {
            i++
            swap(A[i], A[j])
        }
    }
    swap(A[p], A[i])
    return i
}
```



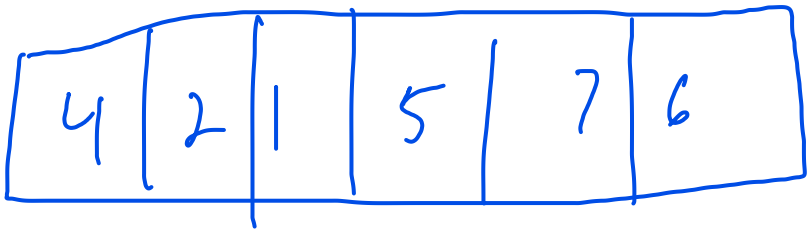




x	i	j
5	~~p~~	~~p+1~~
	~~p+1~~	~~p+2~~
	~~p+2~~	~~p+3~~
	p+3	~~p+4~~
		p+5



Final:



```
Quicksort(A, p, r)
{
    if(p<r)
    {
        q = Partition(A, p, r)
        Quicksort(A, p, q-1)
        Quicksort(A, q+1, r)
    }
}
```

Call in main(): Quicksort(A, 1, n)

Worst Case Running time

Pivot smallest or largest every time.

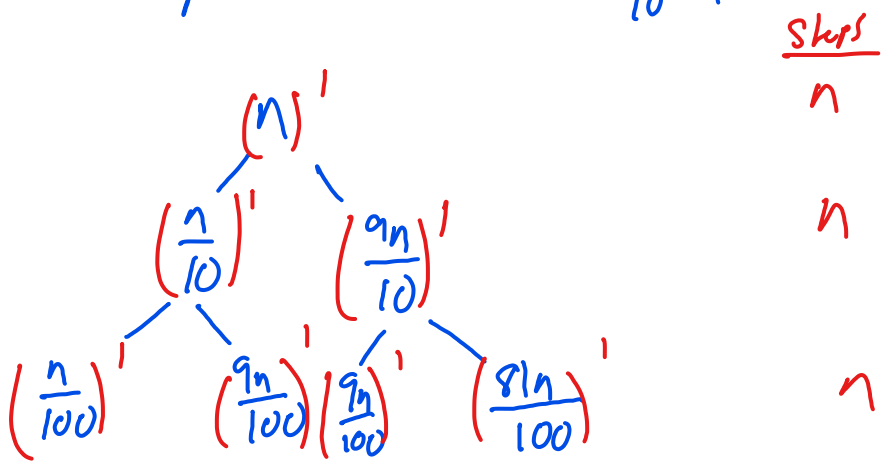
$$T(n) = T(0) + T(n-1) + \Theta(n)$$

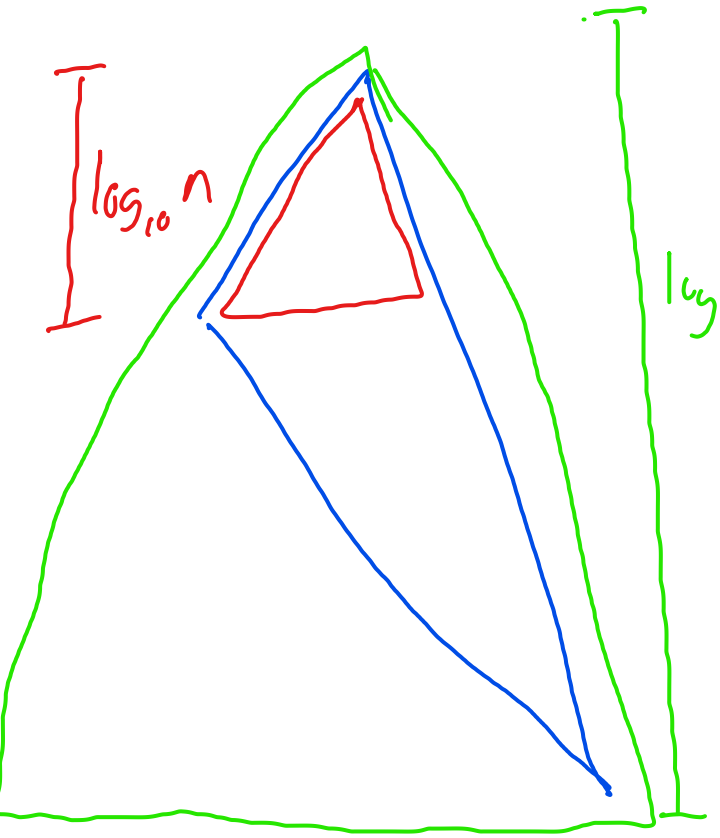
$$\sum_{i=1}^{n} i = \Theta(n^2)$$ not good!

Best Case: pivot is median every time

$$T(n) = 2T\left(\frac{n}{2}\right) + \Theta(n) \Rightarrow \Theta(n \log n)$$

What if we always split into $\frac{n}{10}$ + $\frac{9n}{10}$?





red ≤ blue ≤ green

red: $\Theta(n \log_{10} n)$

green $\Theta(n \log_{\frac{10}{9}} n)$

Randomly pick pivot to randomize QS.

Choice of pivot affects running time $\Rightarrow$ running time is a RV.

What is EV of the running time?

Let X denote the RV of the running time of the alg. We want to compute $E[X]$.

Let X_{ij} be an indicator RV denoting whether we compared the ith smallest value with the jth smallest value.

$$X_{ij} = \begin{cases} 1 & \text{if so} \\ 0 & \text{if not} \end{cases}$$

We never compare two numbers twice, so $X = \sum_{i=1}^{n} \sum_{j=i+1}^{n} X_{ij}$.

$$E[X] = E\left[\sum\sum X_{ij}\right] = \sum\sum E[X_{ij}]$$

↑ linearity of expectation

$$E[X_{ij}] = P(X_{ij} = 1)$$

$$P(X_{1n} = 1) = \frac{2}{n}$$

$$P(X_{1,2} = 1) = 1 = \frac{2}{2}$$

$$P(X_{1,3} = 1) = \frac{2}{3}$$

$$P(X_{ij} = 1) = \frac{2}{j-i+1}$$

$$\sum_{i=1}^{n} \sum_{j=i+1}^{n} \frac{2}{j-i+1} = 2 \sum_{i=1}^{n} \sum_{j=i+1}^{n} \frac{1}{j-i+1}$$

$$< 2 \sum_{i=1}^{n} \left(\frac{1}{2} + \frac{1}{3} + \cdots + \frac{1}{n} \right)$$

Harmonic Series: $O(\log n)$

$$= \Theta(n \log n)$$