20. Searching a List

Topics:

- Linear Search
- Binary Search
- Measuring Execution Time
- The Divide and Conquer Framework
Search

Examples:

Is this song in that playlist?
Is this number in that phone book?
Is this name in that phone book?
Is this fingerprint in that archive of fingerprints?
Is this photo in that yearbook?
More on Using Phone Books

The Manhattan phone book has 1,000,000+ entries.

How is it possible to locate a name by examining just a tiny, tiny fraction of those entries?

There must be a great search algorithm behind the scenes.
Linear Search
def LinSearch(x,a):
    """ Returns an int k with the property that a[k]==x is True. If no such k exists, then k==−1.
    """

    PreC: a is a nonempty list of ints and x is an int.
    """
def LinSearch(x, a):
    for k in range(len(a)):
        if x == a[k]:
            return k
    return -1

Walk down the list looking for a match
Walk down the list looking for a match

def LinSearch(x, a):
    for k in range(len(a)):
        if x == a[k]:
            return k
    return -1
Linear Search

def LinSearch(x, a):
    for k in range(len(a)):
        if x == a[k]:
            return k
        return -1

Walk down the list looking for a match
Linear Search

```
def LinSearch(x,a):
    for k in range(len(a)):
        if x == a[k]:
            return k
    return -1
```

Walk down the list looking for a match
Linear Search

```
def LinSearch(x, a):
    for k in range(len(a)):
        if x == a[k]:
            return k
    return -1
```

Walk down the list looking for a match
Linear Search

Walk down the list looking for a match

def LinSearch(x,a):
    for k in range(len(a)):
        if x == a[k]:
            return k
    return -1
Linear Search

```
def LinSearch(x,a):
    for k in range(len(a)):
        if x == a[k]:
            return k
    return -1
```

Walk down the list looking for a match
def LinSearch(x, a):
    for k in range(len(a)):
        if x == a[k]:
            return k
    return -1

Walk down the list looking for a match
Linear Search: No Match Case

```
def LinSearch(x, a):
    for k in range(len(a)):
        if x == a[k]:
            return k
    return -1
```

Return -1 if no match
Linear Search: While Implementation

def LinSearchW(x,a):
    k=0
    while k<len(a) and a[k]!=x:
        k+=1
    if k==len(a):
        return -1
    else:
        return k
Binary Search

Now we assume that the list to be searched is sorted from little to big.

\[ a = [10, 20, 40, 60, 90] \]

\[ a = [\text{brown}, \text{dog}, \text{fox}, \text{lazy}, \text{quick}, \text{the}] \]
The Ithaca phone book has 10,000+ entries.

The Manhattan phone book has 1,000,000+ entries. But it does not take 100 x longer to look something up. Why?
To find Derek Jeter’s number...

\[
\begin{align*}
B &= \text{phone book} \\
\text{while } (B \text{ is longer than 1 page}): \\
&\quad 1. \ P = \text{middle page of } B \\
&\quad 2. \ \text{Let } Q \text{ be the first name on } P \\
&\quad 3. \ \text{if } \text{‘Jeter” comes before } Q:\ \\
&\quad \quad \text{Rip away the 2}^{\text{nd}} \text{ half of } B \\
&\quad \text{else:\ } \\
&\quad \quad \text{Rip away the 1}^{\text{st}} \text{ half of } B. \\
\end{align*}
\]

Scan remaining page \( P \) line-by-line for ‘Jeter’
What Happens to Phone Book Length?

Original: 3000 pages
After 1 rip: 1500 pages
After 2 rips: 750 pages
After 3 rips: 375 pages
After 4 rips: 188 pages
After 5 rips: 94 pages

After 12 rips: 1 page
Binary Search

The idea of repeatedly halving the size of the “search space” is the main idea behind the method of binary search.

An item in a sorted array of length \( n \) can be located with approximately \( \log_2 n \) comparisons.

\[
\text{log}_2 8 = 3 \quad \text{log}_2 64 = 7 \quad \text{log}_2 2^{**k} = k
\]
What is $\log_2(n)$?

<table>
<thead>
<tr>
<th>n</th>
<th>ceil((\log_2(n)))</th>
</tr>
</thead>
<tbody>
<tr>
<td>10</td>
<td>4</td>
</tr>
<tr>
<td>100</td>
<td>7</td>
</tr>
<tr>
<td>1000</td>
<td>10</td>
</tr>
<tr>
<td>10000</td>
<td>14</td>
</tr>
<tr>
<td>100000</td>
<td>17</td>
</tr>
<tr>
<td>1000000</td>
<td>20</td>
</tr>
</tbody>
</table>
def BinSearch(x,a):
    ''' Returns an int k with the property that a[k]==x is True. If no such k exists, then k===-1. 

    PreC: a is a nonempty list of ints that is sorted from smallest to largest. x is an int. 
    '''
Example:
Does this List have an Element With Value Equal to 70?
Let's Look For x in a

Mid = (L+R)/2

L: 0
Mid: 5
R: 11

a[Mid] <= x ????

x: 70

Mid = (L+R)/2
The Midpoint Computations

<table>
<thead>
<tr>
<th>L</th>
<th>R</th>
<th>(L+R)/2</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>11</td>
<td>5</td>
</tr>
<tr>
<td>2</td>
<td>6</td>
<td>4</td>
</tr>
<tr>
<td>1</td>
<td>100</td>
<td>50</td>
</tr>
</tbody>
</table>
Let's Look For $x$ in $a$

$a \rightarrow \{ 12, 15, 33, 35, 42, 45, 51, 62, 73, 75, 86, 98 \}$

$L$: 0

$Mid$: 5

$R$: 11

$x$: 70

$a[Mid] \leq x$ ????
Let's Look For x in a

a->

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

L: 0

Mid: 5

R: 11

x: 70

a[Mid] <= x

Yes!

So throw away The “left half”
Let's Look For x in a

\[
\begin{array}{cccccccccccc}
0 & 1 & 2 & 3 & 4 & 5 & 6 & 7 & 8 & 9 & 10 & 11 \\
\hline
12 & 15 & 33 & 35 & 42 & 45 & 51 & 62 & 73 & 75 & 86 & 98 \\
\end{array}
\]

a \rightarrow 12 15 33 35 42 45 51 62 73 75 86 98

\[L: \quad 0\]

\[\text{Mid: } \quad 5\]

\[R: \quad 11\]

x: 70

\[a[\text{Mid}] \leq x\]

Yes!

So throw away The “left half”
Let's Look For x in a

L: 0
Mid: 5
R: 11

x: 70

a[Mid] <= x
Revise L and Mid

a-→ 12 15 33 35 42 45 51 62 73 75 86 98
Let's Look For $x$ in $a$

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

$$a\rightarrow$$

```
12 15 33 35 42 45 51 62 73 75 86 98
```

$L$: 5

$Mid$: 8

$R$: 11

$x$: 70

$a[Mid] \leq x$ ???
Let's Look For $x$ in $a$

a $\rightarrow$ 12 15 33 35 42 45 51 62 73 75 86 98

L: 5
Mid: 8
R: 11

$x$: 70

$a[\text{Mid}] \leq x$

No

So throw away the “right half”
Let's Look For $x = 70$

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

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

```
12 15 33 35 42 45 51 62 73 75 86 98
```

- **L**: 5
  - a[Mid] $\leq x$
  - Revise $R$ and $Mid$

- **Mid**: 8

- **R**: 11

- $x$: 70
Let's Look For $x = 70$

L: 5  
Mid: 6  
R: 8  

Revise R and Mid

$a[\text{Mid}] \leq x$
Let's Look For $x$ in $a$

$a \rightarrow 12 \ 15 \ 33 \ 35 \ 42 \ 45 \ 51 \ 62 \ 73 \ 75 \ 86 \ 98$

$L: \quad 5$

$Mid: \quad 6$

$R: \quad 8$

$x: \quad 70$

$a[Mid] \leq x$ ????
Let's Look For $x$ in $a$

$L$: 5  
$Mid$: 6  
$R$: 8  

$x$: 70

$a[Mid] \leq x$

Yes

Throw away the Left half
Let's Look For x in a

L: 6
Mid: 7
R: 8

x: 70

a[Mid] <= x
Yes
Let's Look For x in a

L: 6
Mid: 7
R: 8
x: 70

a[Mid] ≤ x
Throw away the left half
Let’s Look For $x$ in $a$

![Array diagram]

- **L:** 6
- **Mid:** 7
- **R:** 8
- **x:** 70

Done! At this point we just compare $x$ with $a[L]$ and $a[L+1]$.
What We Just Did

L = 0
R = len(a) - 1
while R - L > 1:
    # a[L] <= x <= a[R]
    Mid = (L + R) / 2
    if x <= a[mid]:
        R = Mid
    else:
        L = Mid

Note that a[L] <= x <= a[R] remains True throughout the loop
What We Just Did

```
L = 0
R = len(a) - 1
while R - L > 1:
    # a[L] <= x <= a[R]
    Mid = (L + R) / 2
    if x <= a[Mid]:
        R = Mid
    else:
        L = Mid
```

What is the situation when the loop terminates?
What We Just Did

\[ L = 0 \]
\[ R = \text{len}(a) - 1 \]

while \( R-L > 1 \):

# \( a[L] \leq x \leq a[R] \)

\[ \text{Mid} = (L+R)/2 \]

if \( x \leq a[\text{mid}] \):

\[ R = \text{Mid} \]

else:

\[ L = \text{Mid} \]

**What We Just Did**

\[ R-L \leq 1 \quad \text{implies} \quad R = L+1 \]
After the Loop Ends

This is True: $a[L] \leq x \leq a[L+1]$
After the Loop Ends

if x==a[L]:
    return L
elif x==a[L+1]:
    return L+1
else:
    return -1
Measuring Execution Time

We now have two ways to search a list:

\[ \text{LinSearch}(x, a) \]
\[ \text{BinSearch}(x, a) \]

Intuition: BinSearch much faster.

Can we quantify this with a “stop watch”?
The `timeit` Module

This module can be used to time how long it takes to execute a chunk of code.

Typical chunk = some function of interest.

This is called benchmarking.
Let's benchmark LinSearch(x,a) and BinSearch(x,a).

Compare how long it takes when len(a) equals 1000, 10000, 100000, and 1000000.

Our intuition tells us that as len(a) increases, BinSearch will be dramatically faster.
## BinSearch vs LinSearch

<table>
<thead>
<tr>
<th>n</th>
<th>tBin</th>
<th>tLin</th>
<th>tLinW</th>
</tr>
</thead>
<tbody>
<tr>
<td>1000</td>
<td>0.0007</td>
<td>0.0064</td>
<td>0.0119</td>
</tr>
<tr>
<td>10000</td>
<td>0.0009</td>
<td>0.0668</td>
<td>0.1203</td>
</tr>
<tr>
<td>100000</td>
<td>0.0011</td>
<td>0.8296</td>
<td>1.2082</td>
</tr>
<tr>
<td>1000000</td>
<td>0.0015</td>
<td>17.7388</td>
<td>13.9341</td>
</tr>
</tbody>
</table>

- **tBin** = time for BinSearch
- **tLin** = time for LinSearch (for loop version)
- **tLinW** = time for LinSearch (while-loop version)
BinSearch vs LinSearch

<table>
<thead>
<tr>
<th>n</th>
<th>tLin/tBin</th>
</tr>
</thead>
<tbody>
<tr>
<td>1000</td>
<td>9</td>
</tr>
<tr>
<td>10000</td>
<td>74</td>
</tr>
<tr>
<td>100000</td>
<td>754</td>
</tr>
<tr>
<td>1000000</td>
<td>7095</td>
</tr>
</tbody>
</table>

Reporting ratios is more illuminating since we do not really care about the time units in this informal comparison.
Using the `timeit` Module

We show how this module was use to get the results on the previous slides.

Our LinSearch vs BinSearch example is very typical: is one function faster than another?
from timeit import *
S = ""

Set-up code

B = ""

Code to Benchmark

p = 10; m = 100
t = min(Timer(B, setup=S).repeat(p, m))

Yes, these are doc strings.
The Set-Up and Bench Codes

```python
from random import randint as randi
from ShowSearch import BinSearch
n = 10000
s = [randi(0,10*n) for i in range(n)]
s.sort()
x = s[n//2]
```

The set-up code is run once.
It is not timed.
It just sets up the code to be timed.

```python
k=BinSearch(x,s)
```
A Benchmarking Framework

```python
from timeit import *

S = ""

B = ""

p = 10; m = 100

t = min(Timer(B, setup=S).repeat(p, m))
```

An “experiment” consists of running the blue code m times.
The stopwatch will time how long it takes to do one experiment.

Larger values necessary if the blue code executes very quickly
from timeit import *

S = ""

B = ""

Set-up code

p = 10; m = 100

t = min(Timer(B, setup=S).repeat(p, m))

This helps control for other stuff that may be running on your computer.
from timeit import *

S = ""

B = ""

p = 10; m = 100

t = min(Timer(B,setup=S).repeat(p,m))

In general, it is best to take the minimum as the most reliable. The benchmark time is assigned to t.

This helps control for other stuff that may be running on your computer.
Why Benchmarking is Important

Confirms/refutes what our intuition might say about efficiency.

Makes us sensitive to the various issues that affect efficiency.

Steers us away from simplistic comparisons of different methods that can be used on the same problem.