CS 1114:
Sorting and selection (part three)

Prof. Graeme Bailey

http://cs1114.cs.cornell.edu

(notes modified from Noah Snavely, Spring 2009)

Cornell University
Computer Science

Back to the selection problem

- Can solve with sorting
- Is there a better way?
- Rev. Charles L. Dodgson's problem
  - Based on how to run a tennis tournament
  - Specifically, how to award 2nd prize fairly

Problems, algorithms, programs

- A central distinction in CS
- Problem: what you want to compute
  - "Find the median"
  - Sometimes called a specification
- Algorithm: how to do it, in general
  - "Repeated find biggest"
  - "Quicksort", "Mergesort", "Quickselect" (later)
- Program: how to do it, in a particular programming language

```
function [med] = find_median[A]
...
```


- How many teams were in the tournament?
- How many games were played?
- Which is the second-best team?

Finding the second best team

- Could use quicksort to sort the teams

- Step 1: Choose one team as a pivot (say, Arizona)
- Step 2: Arizona plays every team
- Step 3: Put all teams worse than Arizona in Group 1, all teams better than Arizona in Group 2 (no ties allowed)
- Step 4: Recurse on Groups 1 and 2
- … eventually will rank all the teams …

Quicksort Tournament

Quicksort Tournament
Step 1: Choose one team (say, Arizona)
Step 2: Arizona plays every team
Step 3: Put all teams worse than Arizona in Group 1, all teams better than Arizona in Group 2 (no ties allowed)
Step 4: Recurse on groups 1 and 2
… eventually will rank all the teams …

- (Note this is a bit silly – AZ plays 63 games)
- This gives us a ranking of all teams
  - What if we just care about finding the 2nd-best team?

Modifying quicksort to select

- Suppose Arizona beats 36 teams, and loses to 27 teams

$\underbrace{\text{36 teams}}_{\text{Group 1}}$ < A < $\underbrace{\text{27 teams}}_{\text{Group 2}}$

- If we just want to know the 2nd-best team, how can we save time?

Modifying quicksort to select – Finding the 2nd best team

$\underbrace{\text{36 teams}}_{\text{Group 1}}$ < A < $\underbrace{\text{27 teams}}_{\text{Group 2}}$

$\underbrace{\text{16 teams}}_{\text{Group 2.1}}$ < Cornell < $\underbrace{\text{10 teams}}_{\text{Group 2.2}}$

7 teams < NC < 2 teams

Modifying quicksort to select – Finding the 32nd best team

36 teams < < 27 teams

Group 1 Group 2

20 teams < < 15 teams

Group 1.1 Group 1.2

- Q: Which group do we visit next?
- The 32nd best team overall is the 4th best team in Group 1

Find kth largest element in A (< than k-1 others)

A = [6.0 5.4 5.5 6.2 5.3 5.0 5.9]

MODIFIED QUICKSORT(A, k):

- Pick an element in A as the pivot, call it x
- Divide A into A1 (<x), A2 (=x), A3 (>x)
- If k < length(A3)
  - MODIFIED QUICKSORT (A3, k)
- If k > length(A2) + length(A3)
  - Let j = k – [length(A2) + length(A3)]
  - MODIFIED QUICKSORT (A1, j)
- Otherwise, return x

Modified quicksort

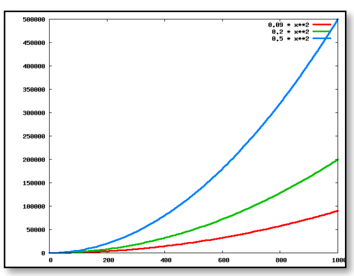
MODIFIED QUICKSORT(A, k):

- Pick an element in A as the pivot, call it x
- Divide A into A1 (<x), A2 (=x), A3 (>x)
- If k < length(A3)
  - Find the element < k others in A3
- If k > length(A2) + length(A3)
  - Let j = k – [length(A2) + length(A3)]
  - Find the element < j others in A1
- Otherwise, return x

- We'll call this *quickselect*
- Let's consider the running time…

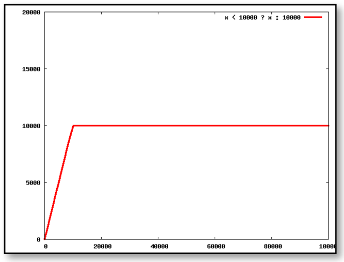
Big-O notation

- "Constant of proportionality" doesn't matter

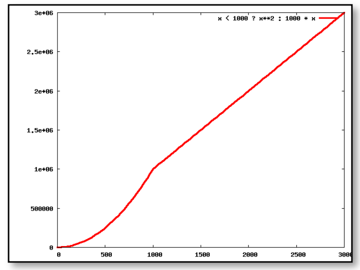


$O(n^2)$

- We only care about how the function grows for "large" values of n

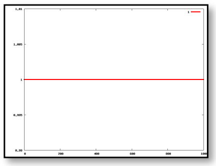


O(1)

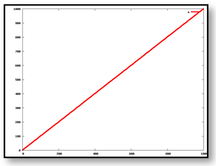


O(n)

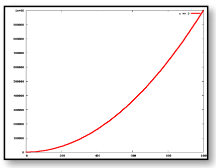
What is the running time of:



- Finding the 1st element?
  - O(1) (effort doesn't depend on input)



- Finding the biggest element?
  - O(n) (constant work per input element)



- Finding the median by repeatedly finding and removing the biggest element?
  - $O(n^2)$ (linear work per input element)

- Finding the median using quickselect?
  - Worst case? O(n^2)
  - Best case? O(n) …… we'll show that now ….

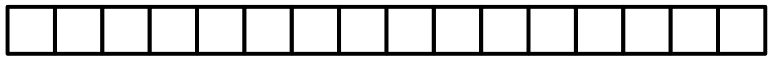
Quickselect – "medium" case

- Suppose we split the array in half each time (i.e., happen to choose the median as the pivot)

- How many comparisons will there be?

How many comparisons? (“medium” case)

- Suppose `length(A) == n`



- Round 1: Compare *n* elements to the pivot

  … now break the array in half, quickselect one half …

- Round 2: For remaining half, compare *n* / 2 elements to the pivot (total # comparisons = n / 2)

  … now break the half in half …

- Round 3: For remaining quarter, compare *n* / 4 elements to the pivot (total # comparisons = n / 4)

How many comparisons? (“medium” case)

Number of comparisons =

$$n + n / 2 + n / 4 + n / 8 + \ldots + 1$$

$$= ?$$

→ The “medium” case is O(n)!

Quickselect

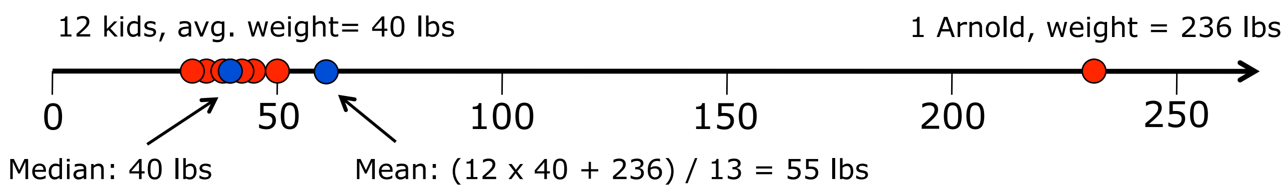
- For random input this method actually runs in linear time (beyond the scope of this class)
- The worst case is still bad
- Quickselect gives us a way to find the k^{th} element without actually sorting the array!
- It's possible to select in *guaranteed* linear time (1973)
  - But the code is a little messy
    - And the analysis is messier
    http://en.wikipedia.org/wiki/Selection_algorithm
  - Beyond the scope of this course

Back to the lightstick

- By using quickselect we can find the 5% largest (or smallest) element
  - This allows us to compute the trimmed mean efficiently

What about the median?

- Another way to avoid our bad data points:
  - Use the median instead of the mean



Median vector for 2D data

- Mean, like median, was defined in 1D
  - For a 2D mean we used the centroid
  - Mean of x coordinates and y coordinates separately
    - Call this the “mean vector”
  - Does this work for the median also?

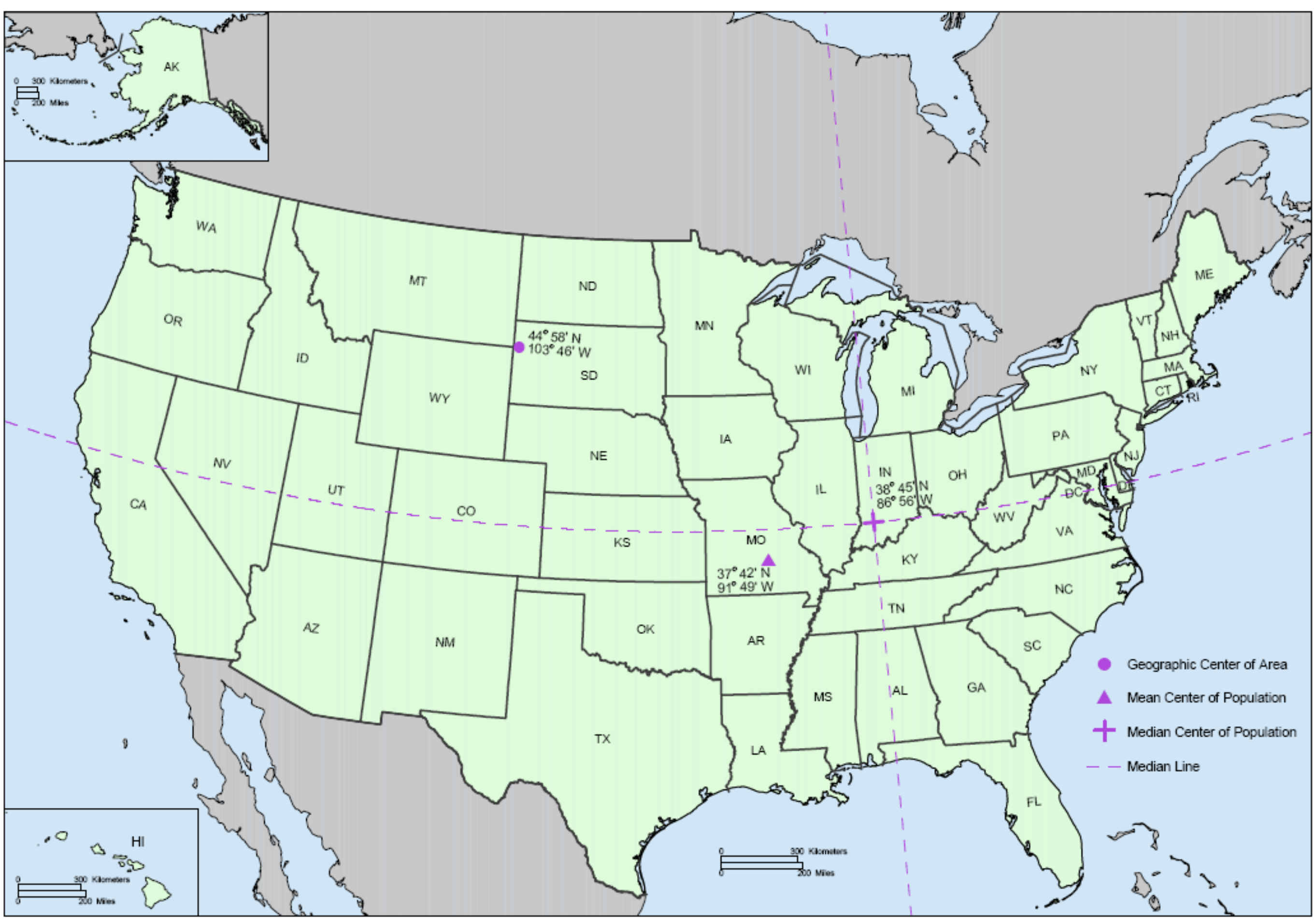
What is the median vector?

- In 1900, statisticians wanted to find the "geographical center of the population" to quantify westward shift

- Why not the centroid?
  - Someone being born in San Francisco changes the centroid much more than someone being born in Indiana

- What about the "median vector"?
  - Take the median of the x coordinates and the median of the y coordinates separately
  - Would this be different if done in polar coordinates?

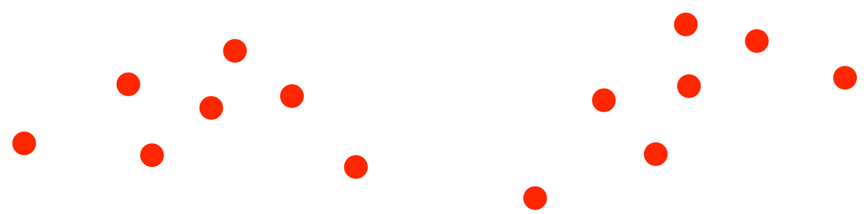
Position of the Geographic Center of Area, Mean and Median Centers of Population: 2000



U.S. Department of Commerce Economics and Statistics Administration U.S. Census Bureau
Prepared by the Geography Division

Median vector

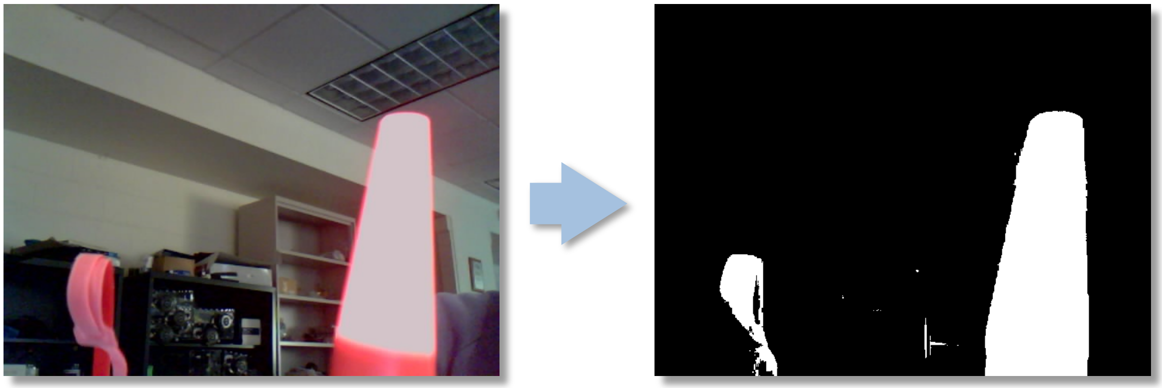
- A little thought will show you that this doesn't really make a lot of sense
  - Nonetheless, it's a common approach, and we will implement it for CS1114
  - In situations like ours it works pretty well
- It's almost never an actual datapoint



Can we do even better?

- None of what we described works that well if we have widely scattered red pixels
  - And we can't figure out lightstick orientation
- Is it possible to do even better?
  - Yes!
- We will focus on:
  - Finding "blobs" (connected red pixels)
  - Summarizing the shape of a blob
  - Computing orientation from this
- We'll need brand new tricks!

Back to the lightstick



- The lightstick forms a large "blob" in the thresholded image (among other blobs)

What is a blob?

1	0	0	0	0	0	0	0	1	0
0	0	0	0	0	0	0	0	1	0
0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0
0	0	0	0	1	0	0	0	0	0
0	0	0	1	1	1	0	0	0	0
0	0	0	1	1	1	0	0	0	0
0	0	0	1	1	1	0	0	0	0
0	0	0	1	1	1	0	0	0	0
0	0	0	1	0	0	0	0	0	0

Finding blobs

1. Pick a 1 to start with, where you don't know which blob it is in
   – When there aren't any, you're done
2. Give it a new blob color
3. Assign the same blob color to each pixel that is part of the same blob

Finding blobs

1	0	0	0	0	0	0	0	1	0
0	0	0	0	0	0	0	0	1	0
0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0
0	0	0	0	1	0	0	0	0	0
0	0	0	1	1	1	0	0	0	0
0	0	0	1	1	1	0	0	0	0
0	0	0	1	1	1	0	0	0	0
0	0	0	1	1	1	0	0	0	0
0	0	1	0	0	0	0	0	0	0

Finding blobs

1	0	0	0	0	0	0	0	1	0
0	0	0	0	0	0	0	0	1	0
0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0
0	0	0	0	1	0	0	0	0	0
0	0	0	1	1	1	0	0	0	0
0	0	0	1	1	1	0	0	0	0
0	0	0	1	1	1	0	0	0	0
0	0	0	1	1	1	0	0	0	0
0	0	1	0	0	0	0	0	0	0

Finding blobs

1	0	0	0	0	0	0	0	1	0
0	0	0	0	0	0	0	0	1	0
0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0
0	0	0	0	1	0	0	0	0	0
0	0	0	1	1	1	0	0	0	0
0	0	0	1	1	1	0	0	0	0
0	0	0	1	1	1	0	0	0	0
0	0	0	1	1	1	0	0	0	0
0	0	1	0	0	0	0	0	0	0

Finding blobs

1	0	0	0	0	0	0	0	1	0
0	0	0	0	0	0	0	0	1	0
0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0
0	0	0	0	1	0	0	0	0	0
0	0	0	1	1	1	0	0	0	0
0	0	0	1	1	1	0	0	0	0
0	0	0	1	1	1	0	0	0	0
0	0	0	1	1	1	0	0	0	0
0	0	1	0	0	0	0	0	0	0

Finding blobs

1	0	0	0	0	0	0	0	1	0
0	0	0	0	0	0	0	0	1	0
0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0
0	0	0	0	1	0	0	0	0	0
0	0	0	1	1	1	0	0	0	0
0	0	0	1	1	1	0	0	0	0
0	0	0	1	1	1	0	0	0	0
0	0	0	1	1	1	0	0	0	0
0	0	1	0	0	0	0	0	0	0

Finding blobs

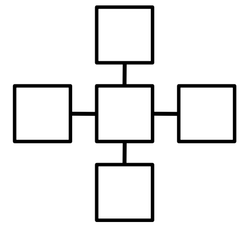
1	0	0	0	0	0	0	0	1	0
0	0	0	0	0	0	0	0	1	0
0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0
0	0	0	0	1	0	0	0	0	0
0	0	0	1	1	1	0	0	0	0
0	0	0	1	1	1	0	0	0	0
0	0	0	1	1	1	0	0	0	0
0	0	0	1	1	1	0	0	0	0
0	0	1	0	0	0	0	0	0	0

Finding blobs

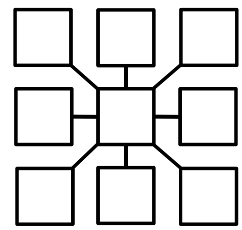
1. Pick a 1 to start with, where you don't know which blob it is in
   - When there aren't any, you're done
2. Give it a new blob color
3. Assign the same blob color to each pixel that is part of the same blob
   - How do we figure this out?
   - You are part of the blob if you are next to someone who is part of the blob
   - But what does "next to" mean?

What is a neighbor?

- We need a notion of neighborhood
  - Sometimes called a neighborhood system
- Standard system: use vertical and horizontal neighbors
  - Called “NEWS”: north, east, west, south
  - 4-connected, since you have 4 neighbors



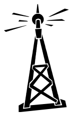
- Another possibility includes diagonals
  - 8-connected neighborhood system



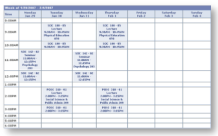
The long winding road to blobs

- We actually need to cover a surprising amount of material to get to blob finding
  - Some of which is not obviously relevant
  - But (trust me) it will all hang together!

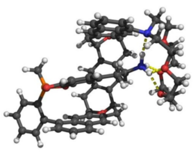
A single idea can be used to think about:



– Assigning frequencies to radio stations



– Scheduling your classes so they don't conflict



– Figuring out if a chemical is already known

– Finding groups in Facebook

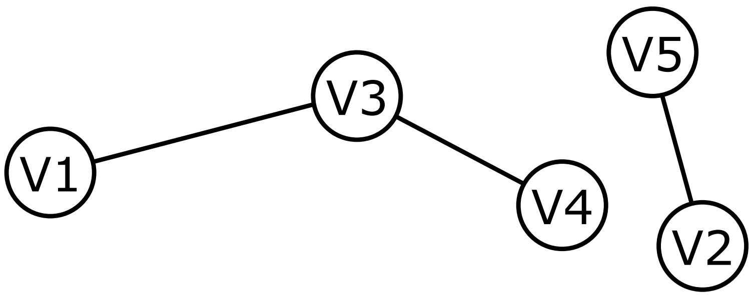
– Ranking web search results

Graphs: always the answer

- We are going to look at an incredibly important concept called a graph
  – Note: not the same as a plot
- Most problems can be thought of in terms of graphs
  – But it may not be obvious, as with blobs

What is a graph?

- Loosely speaking, a set of things that are paired up in some way
- Precisely, a set of vertices **V** and edges **E**
  - Vertices sometimes called nodes
  - An edge (or link) connects a pair of vertices

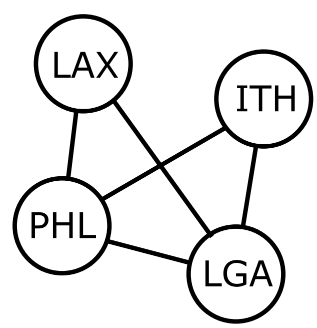


V = { V1, V2, V3, V4, V5 }

E = { (V1,V3), (V2,V5), (V3,V4) }

Notes on graphs

- What can a graph represent?
  - Cities and direct flights
  - People and friendships
  - Web pages and hyperlinks
  - Rooms and doorways
  - IMAGES!!!



Notes on graphs

- A graph isn’t changed by:
  - Drawing the edges differently
    - While preserving endpoints
  - Renaming the vertices