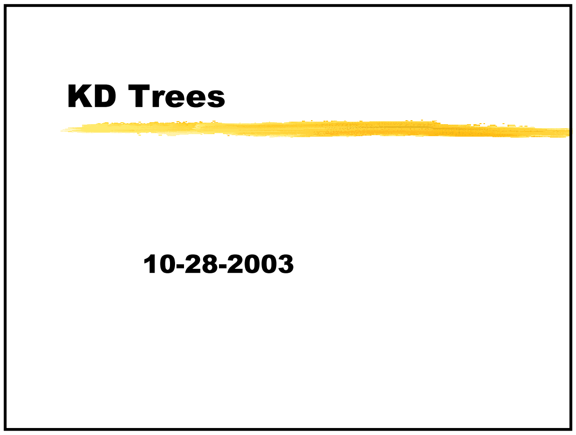
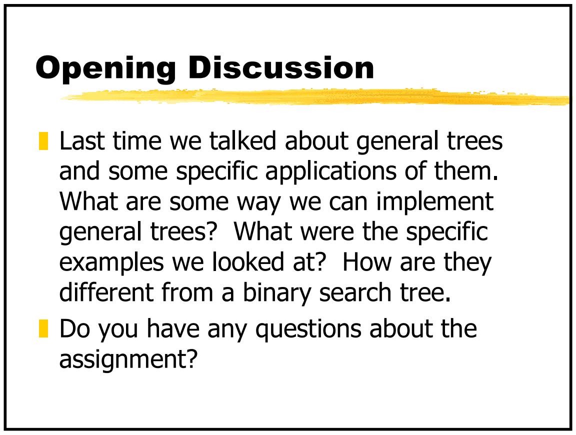


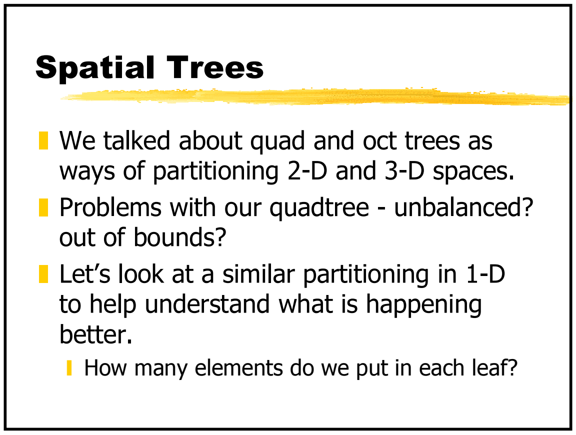
KD Trees

10-28-2003



Opening Discussion

- Last time we talked about general trees and some specific applications of them. What are some way we can implement general trees? What were the specific examples we looked at? How are they different from a binary search tree.
- Do you have any questions about the assignment?



Spatial Trees

- We talked about quad and oct trees as ways of partitioning 2-D and 3-D spaces.
- Problems with our quadtree - unbalanced? out of bounds?
- Let's look at a similar partitioning in 1-D to help understand what is happening better.
  - How many elements do we put in each leaf?

Higher Dimensions

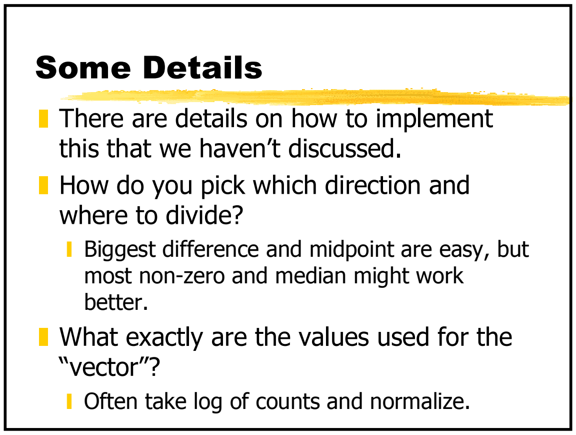
- The approach of a quad/oct tree was to divide a space evenly into 2^n children. This works well for low dimensions. We could also choose to divide unevenly if we had some other information about the distribution.
- We can also chose to make a tree binary, but the problem is picking how to divide it. A BSP is a tree typically used in 3-D graphics that divides with a random plane.

KD-tree

- Defining general "planes" in high dimensions isn't exactly easy and has diminishing returns. Instead, for high dimensionality systems we typically divide using a "plane" that is perpendicular to one axis. We can pick a different axis and location at each division.
- We can draw this out in 2-D to see what it looks like.

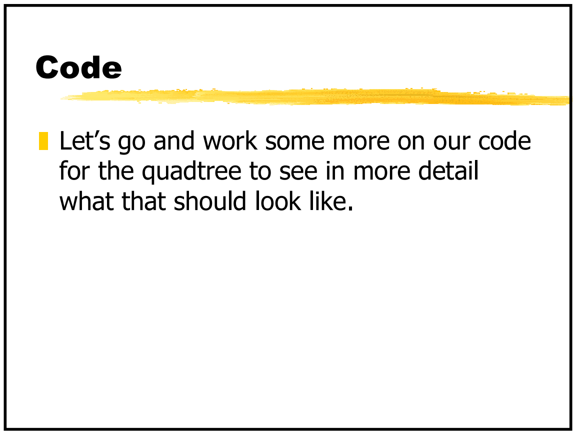
Web Pages as High Dimension Points

- In assignment #5 you will write a KD-tree as the basis for a "search engine". This is based on the common practice that the word counts for a page can be used to describe points in a very high dimension space.
- Points that are close together in this space should be for pages with similar contents. The tree allows you to search that quickly.



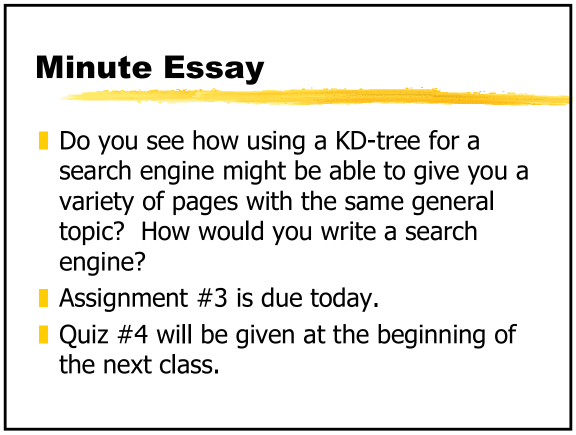
Some Details

- There are details on how to implement this that we haven’t discussed.
- How do you pick which direction and where to divide?
  - Biggest difference and midpoint are easy, but most non-zero and median might work better.
- What exactly are the values used for the “vector”?
  - Often take log of counts and normalize.



Code

- Let’s go and work some more on our code for the quadtree to see in more detail what that should look like.



Minute Essay

- Do you see how using a KD-tree for a search engine might be able to give you a variety of pages with the same general topic? How would you write a search engine?
- Assignment #3 is due today.
- Quiz #4 will be given at the beginning of the next class.