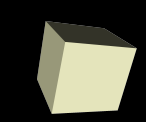
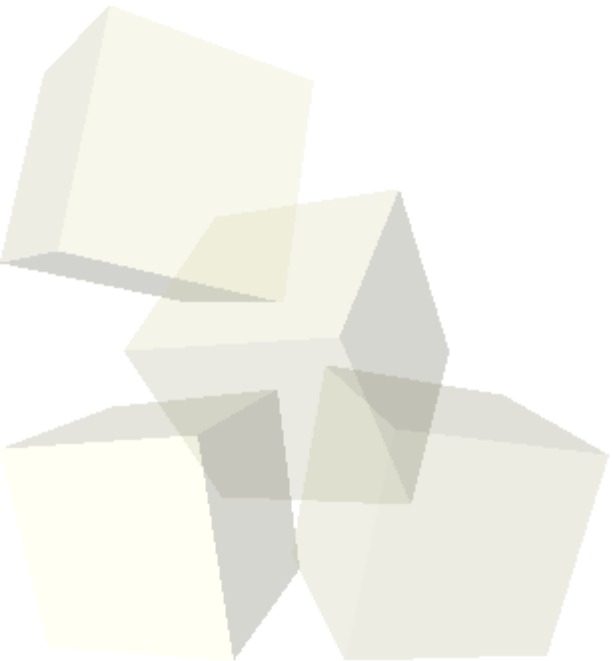


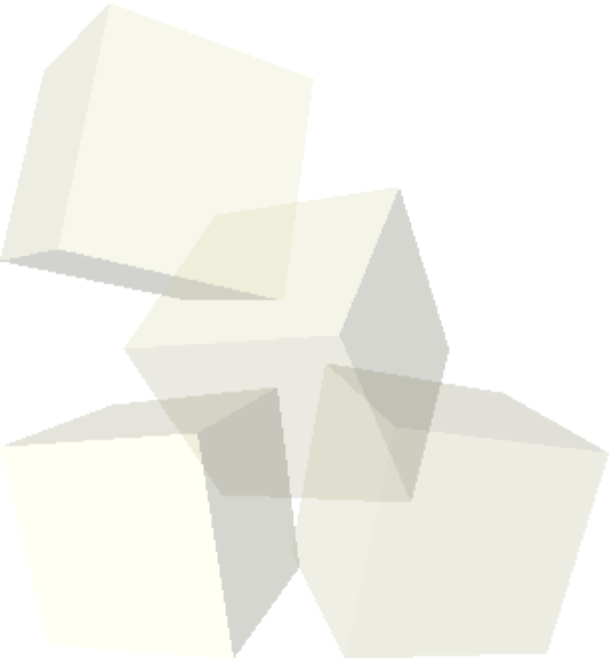
Data Processing and Mining

10-19-2006



Opening Discussion

- How are things going with the project? You should try to get the writeup of that done soon, hopefully today.
- How much do you want to go through Perl? How do you want to do it? What should we focus on?



Beyond Numerics

- So far what we have explored is mainly the use of computers to do numerics. The reason is obvious, computers can do calculations much faster than you or I can by hand.
- The ability to do numerics was one of the key reasons that computers were developed at all.
- More recently scientists have begun using computers for different reasons, to process vast amounts of data.
- Early computers didn't have that much memory so this wasn't an option. Now you can easily get a 1TB drive which can hold more information than every book you will ever touch in your life.

Applications

- Many different areas of science now have a need to process large quantities of data.
  - Biology has the areas of bioinformatics and genomics. Basically, the ability to look at the genome has produced large data sets that have to be compared.
  - Astronomy has large surveys of the sky that keep track of all types of information on a large number of stars.
  - Geology and atmospheric physics have more monitoring stations and in the past and they return more data than in the past.
  - The biggest data sets currently come from particle physics. Collider events can produce many GB from a single smash and colliders will produce TBs from an experiment.

Flat Files

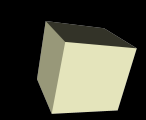
- We will begin talking about handling data in standard text files and tools looking through those.
- These types of files are easy to use and have the advantage that you don't really require special tools.
- They are less than ideal when the data sets get really big.
- We will use Perl for this type of processing.

Databases

- When the amount of data gets beyond many MB and into the GB and TB range, using flat files becomes much less efficient and it becomes useful to put the data into databases.
- Databases give you the ability to quickly find information based on certain indexes and are typically optimized for efficiency and to operate in high volume.
- The drawback is that databases aren't as simple to use. That's part of why we teach an entire course on the topic.

Languages

- For flat files we could use pretty much any language that we wanted.
- Just like Matlab was designed to make it easy to do numeric processing, Perl is designed for doing text manipulation.
- We could do the text processing with Java, C++, or even Matlab, but Perl has features supporting text processing built into the language. In Java or C++ you would find those features in the libraries instead of being part of the language itself.



Reminders

- Do some readings on Perl.