Stat196k Schedule

1 minute read

Here is a tentative outline of topics for STAT 196K. The goal of the course is to achieve the following learning outcomes.

Learning Outcomes

Students will be able to:

1. Develop complete statistical computer programs based on high level directions, using standard software packages. Their programs will be complete in the sense that they start with processing raw data, and finish by producing final summaries and results necessary for reports.
2. Apply standard statistical techniques suitable for data larger than memory, for example, the split-apply-combine strategy for grouped data, memory efficient streaming statistics, discretization, and dimension reduction through principal components analysis.
3. Identify, extract, and summarize elements of interest from complex data sets, including tabular, hierarchical, streaming, image, and text data.
4. Summarize their approach and conclusions for a data analysis problem through technical written reports with appropriate graphics.
5. Perform data analysis using remote machines, which may include databases, remote compute clusters, and cloud services.
6. Accelerate and scale data analysis programs by identifying and fixing performance bottlenecks.

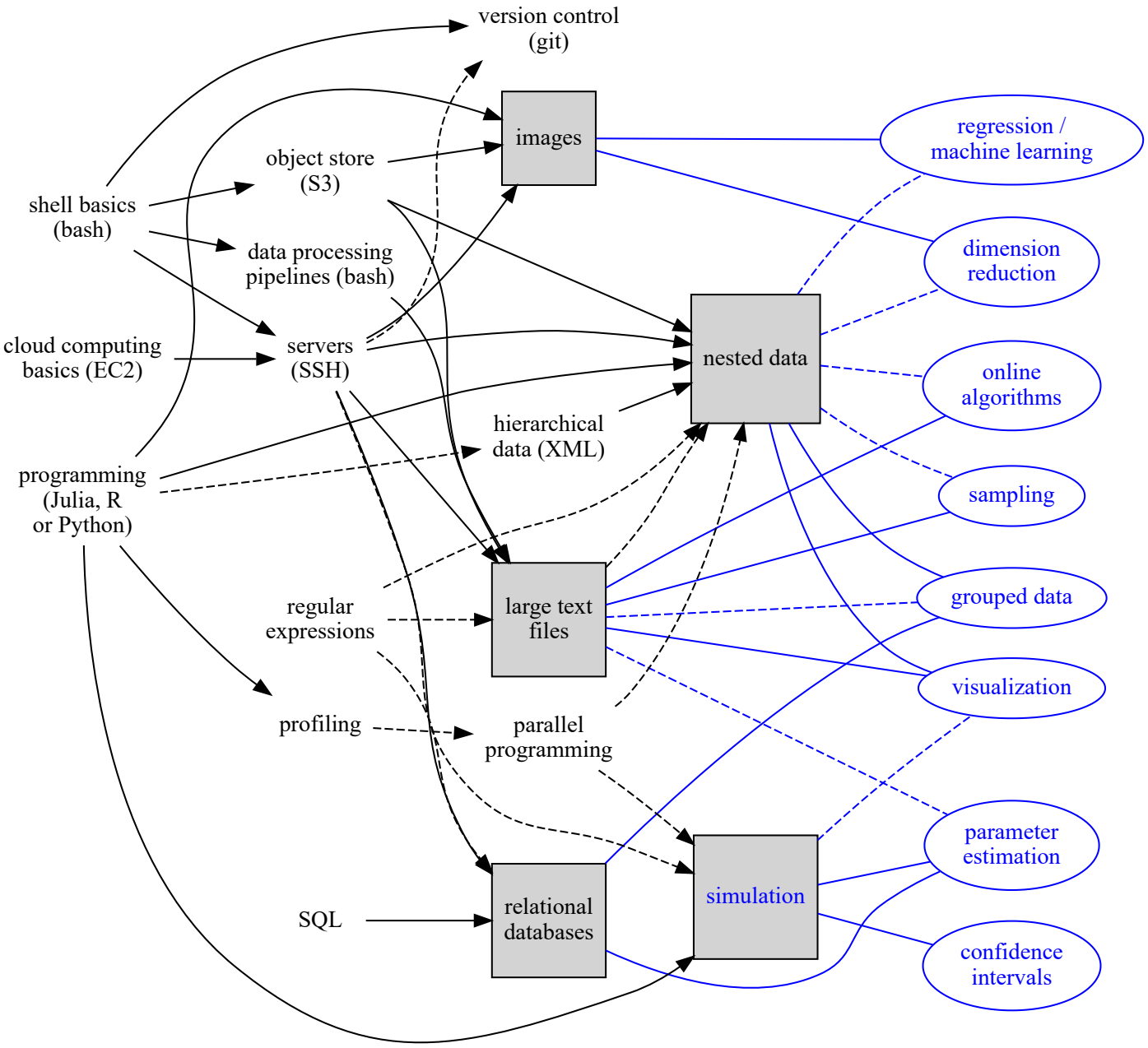
Assignments

We'll have around 5 large assignments in the class, each focused on a different kind of data or problem. Each assignment covers most of the above learning outcomes, with varying levels of emphasis. For example, we will run the programs for all the assignments on AWS, so all the assignments support the learning outcome "Perform data analysis using remote machines".

1. **Large Text Data** is one or more text files larger than memory. Example: CSV file with many rows.
2. **Nested Data** has a hierarchical, nested structure in one or many files, possibly larger than memory. Example: XML, JSON.
3. **Relational Databases** contain many related tables to join together. Example: Remote database server accessed via a SQL client.
4. **Images** are collections of pictures, one per file. We're going to use off the shelf high level image processing software in our language of choice rather than implement the details ourselves. Example: PNG, JPG.
5. **Simulation** could be based on any kind of data or random process. A pandemic simulation based on actual physical population densities could be very interesting.

These assignments require specific computing skills to complete as illustrated in the image below. The edges represent prerequisites. For example, the edge from "shell basics" to "version control" indicates that students should learn "shell basics" before "version control". Shaded boxes represent assignments, and the blue ovals on the right represent statistical concepts.

Relationship Between Concepts and Assignments



Updated: January 23, 2021