Computing has always been an essential ingredient of statistical practice. While probability theory provides us with a mathematical foundation for describing data and studying statistical inference, computing technologies act as a medium through which analyses are actually realized. Our ability to manipulate data and to audition new methodologies depends on and is limited by our familiarity with computing technologies. To some extent, even our notion of what constitutes “data” is a product of our background in computing.

Through a series of group projects, we will study tools for “exploratory computing.” We will emphasize programming and scripting languages over point-and-click interfaces. We hope to instill a problem solving ability so that you will learn languages on your own, cull online documentation or tutorials, find books and manuals.

Organizational details

The course will be structured around a series of group projects. Groups will be formed during the first lecture so that each consists of students with different computing skills. While we will be covering several programming languages and data technologies, our approach will be motivated by the demands of the projects. Time permitting, we will cover the following topics, roughly in the order given below:

- Operating systems and Unix
- Unix tools, pipes, job control, editors
- Awk and Perl
- Regular expressions, manipulating text
- R, data types, basic computations, objects and methods, vectorized operations
- Code distribution, R and packages
- Databases, SQL, R interface to MySQL, XML
- Extending R, C basics and calling C from R
- Statistical computation and realtime systems

Through class discussions of the projects, we will evaluate the strengths of each language and compare different approaches. By the end of this quarter, we expect students to have developed a
kind of aesthetic when it comes to computing with data. More importantly, we hope to initiate a
culture in which students routinely discuss computation, sharing experiences or reporting on new
computing platforms and emerging technologies.

There is no computing, probability or statistics prerequisite for this course.

Grading

Grades will be based on class participation (20%), and a series of group projects (80%). Some of
these projects will culminate in oral presentations. You are expected to be an active, contributing
member of your group. In some cases, tasks will be divided among group members, in others,
everyone will perform the same exercises and the group will act as a forum to discuss various
solutions.

Textbooks

There is no required textbook for this class. Instead, readings and other materials will be made
available throughout the quarter on the course Web site. A few useful online references are listed
below.

- The documents An Introduction to R, R Language Definition and R Data Import/Export, all
  available on the R project site (www.cran.r-project.org/manuals)
- Perl resources (www.cpan.org, www.perldoc.com)
- Unix tutorials (www.sims.berkeley.edu/~kevin/unix-tutorial)