

STA250

Data Technologies

(Graduate)

Computing technology is changing rapidly. Both software and data are now available to us in many different forms from dynamically updated sources. This course aims to introduce many of the important and emerging tools from information technology in the context of their use in statistical practice and research. Students will learn to access complex data in “raw” forms such as documents, citations, log files, DNA sequences, or annotations. We will discuss software development for the data analysis process, including text processing, accessing and creating relational databases, and working with web servers.

This is intended to be a graduate level course. It may also be useful for some undergraduates, as well as graduate students from other programs.

The course builds on knowledge of a high-level computing language (e.g. *R* or *Matlab*).

- Week 1. Introduction to Information Technology for Statistics
- Week 2. Text Processing
Regular Expressions
- Week 3. *Perl* Programming
- Week 4. Relational Databases
- Week 5. *SQL* Programming
- Week 6. Structured Self-Describing Data
- Week 7. *XML* Programming
- Week 8. Integrating Software Components
- Week 9. Client-Server Model
- Week 10. Parallel/Distributed Computing