# Syllabus

**Course Name:** ST599 Statistical Computing and Big Data
**Credits:** 3
**Term:** Spring 2014

**Prerequisites:** ST 412/512 or ST 552 or equivalent

**Lectures**: MWF 900-950 OWEN 103
**Instructors**:
Charlotte Wickham, 76 Kidder charlotte.wickham@stat.oregonstate.edu
Alix Gitelman, 48 Kidder gitelman@stat.oregonstate.edu

**Office hours**:
Wickham: 1-2pm WF in 76 Kidder
Gitelman: 2-3pm M in 48 Kidder

## Course Objectives

We'll start with data that is just a little bigger and a little messier than you are used to and focus on strategies for making big data small. We'll progress to tools for extending our skills to datasets that exceed the capabilities of our local computer. Once we are equipped to physically handle big data, we'll explore some of the statistical issues that arise. The class revolves around three big data analysis projects that you will work on in groups. By the end of the course you will confidently approach massive data sets and be aware of the possible statistical pitfalls.

R will be our lingua franca, but expect to pick up a little of some other languages along the way.

### Learning Outcomes

- Be able to access big data sets from remote locations for analysis
- Articulate statistical issues specific to big data
- Suggest approaches to answering questions of interest sensitive to the concerns of data size and computational complexity
- Understand the appropriate use of data mining tools and their limitations
- Use R in combination with other tools to access and process big datasets.

## Course policies

### Class Time

Class time will consist of discussion, group exercises and problem solving, group project work and some lectures.

## Course website

Course materials, lecture notes, handouts and readings will be posted on the class website stat599.cwick.co.nz. Blackboard will be used for submitting questions, project reports and team evaluations, and to record grades.

## Learning Resources

There is no textbook for this class. Readings will be assigned for each topic. You will also be expected to research some concepts and computing tools on your own.

## Tentative Schedule

| Week | Topic |
|---|---|
| 1 | Introduction |
| 2-3 | Getting started with big data |
| 4 | Project 1 presentations |
| 5-6 | Statistical issues with big data |
| 7 | Project 2 presentations |
| 8-9 | Data mining |
| 10 | Project 3 presentations |

## Assessment

- **90% projects**: 3 projects x 30% each. Your grade on each project will be your group's project score adjusted by an individual group citizenship score. Your group citizenship score will be based on self and peer evaluations of your ability to work in a team.

- **10% participation:** You must submit one independent question per week by Monday 5pm on Blackboard (under Assignments). These questions can be about lecture materials, computing tools, data or something you are struggling with in your group. Some weeks we will focus the scope of the questions. Common questions will be discussed in class.

## Group guidelines

You will be assigned groups for the first project at the start of the second week. At the completion of that project groups will be rearranged, unless there is unanimous agreement from all group members to stay together. Group policies and guidelines will be discussed at the start of the second week.

### Project deliverables

For each project, your group will be responsible for:

- an in class presentation in which all members will participate
- a two page written summary
- a git repository documenting your work

For each project, you will also individually submit:

- a self and peer evaluation form

## University and Department policies

### Disability statement

Accommodations are collaborative efforts between students, faculty and Disability Access Services (DAS). Students with accommodations approved through DAS are responsible for contacting me prior to or during the first week of the term to discuss accommodations. Students who believe they are eligible for accommodations but who have not yet obtained approval through DAS should contact DAS immediately at (541) 737-4098.

### Academic integrity

Academic dishonesty is a serious offense and will be addressed following the guidelines set out in the Academic Regulations of OSU (go to http://catalog.oregonstate.edu, click on Registration Information then Academic Regulations, and read AR 15).

The Student Conduct Code defines Academic dishonesty as

> ... an act of deception in which a Student seeks to claim credit for the work or effort of another person, or uses unauthorized materials or fabricated information in any academic work or research, either through the Student's own efforts or the efforts of another.

Examples include, but are not limited to, the following:

- verbatim copying of another student's homework assignment
- copying off another student's exam
- using prohibited materials (e.g., cell phone, cheat sheet) during an exam
- communicating with another student during an exam
- changing answers on an exam after the exam has been graded
- unattributed use of material copied from an article, textbook, or web site
- continuing to write on an exam after the instructor or TA has asked for the exams to be handed in.

You are responsible for knowing what academic dishonesty is, and for avoiding it. Ignorance of these rules does not absolve you from responsibility.