

STAT 663 and CSI 773
Statistical Graphics and Data Exploration
*Data Mining Certificate Program Class

Instructor: Professor Daniel B. Carr
Office: Engineering, Room 1711
Phone: (703) 993-1671
Hours: Tues 6:00 - 7:00 PM and by appointment,
Email: dcarr@gmu.edu

Prerequisites: A 300 level statistics course or higher
Stat 554 strongly recommended

Class Web Site: <http://classweb.gmu.edu/dcarr/eda/>
This contains additional information such as links to this syllabus,
related resources and the class schedule

Class Schedule: <http://classweb.gmu.edu/dcarr/eda/schedule.htm>
This contains lecture topics, reading schedule, links to notes and
assignments. Bookmarking this is a good idea.

Student Web Sites: All students are expected to create and/or maintain a web set and
post assignments on the web site.

Grades: Available via blackboard. Grade component weights:
Homework 20%, Quizzes 5%, Midterm 20%, Redesign 20%, Project 35%
The instructor reserves the right to discontinue quizzes and reallocate its
percent.

Texts: Required: *Visualizing Data*, W. S. Cleveland, Hobart Press, 1994

R Texts: **R documentation and class notes usually suffice.**
There are many helpful texts and I may strongly recommend one in class.

The <http://www.r-project.org/> book link shows at least 85 texts.

Graphics I will discuss these in class,
ggplot2: Elegant Graphics for Data Analysis. H. Wickham - Springer 2009
Lattice, Multivariate Data Visualization with R, D. Sarkar, Springer 2008
R Graphics, P. Murrel, Chapman & Hall/CRC 2006

Data and Stat. I will discuss options. Two examples used in previous classes:
A Handbook of Statistical Analyses Using R. Everitt and Hothorn,
Chapman & Hall/CRC 2006 (new version soon to be out)
Modern Applied Statistics with S-Plus, Venables and Ripley, Springer
Version 4 (2002) or later:

Software: R is basic software for this class. R is free software is available under a GNU license. See class assignments for download directions for R, other packages used in assignments. Some software is windows specific so a few students may need to get short term access to a windows system.

Extended Class Description

This class presents **statistical graphics methodology for exploring data**. The methodology includes visualization strategies related to data description and to the study of relationships within the data. Early descriptive strategies often employ graphic templates such as dot plots, box plots, qqplots, rplots, splots, time series plots, and scatterplot matrices. The class introduces these templates and related data transformations. Many strategies involve interactive and dynamic graphics. The class will introduce examples in presentations and assignments.

Statistical graphics are intimately connected with **statistical analysis methodology**. Statistical analysis methods often fall under the two headings of density estimation and model building. Class assignments include density estimation, smoothing, clustering, regression, and other modeling/data mining methodology along with graphics. A single semester class can only touch on these topics. The instructor is open to providing individualized help to those pushing beyond the assignment examples for their final project. Final projects presentations are often instructive to the class in terms of illustrating more advanced use of methodology previously introduced.

This class devotes substantial attention to **maps**. I have a long research history in integrating statistical graphics with maps and software development, as indicated by my new book with Linda Pickle. It is entitled, *Visualizing Pattern in Data With Micromap* and scheduled to appear in February 2010. The National Cancer Institute uses my linked micromap design to communicate with health planners across the nation. See statecancerprofiles.cancer.gov/micromaps.

Designing effective graphics depends upon knowledge about human beings. The class addresses issues of human **perception and cognition** that are important in graphics design. The design guidance is organized around four tasks: encourage accurate comparisons, provide context for appropriate interpretation, simplify appearance, and engage the reader or, better yet, the analyst.

Data exploration can be a lot of fun when we are curious. My curiosity has led me to **wide range of applications** including genomic, proteomics, land cover, biodiversity, labor statistics, agricultural statistics, network packet header data, transportation statistics, even global atmospheric data. Over my career the challenge of visualizing massive data sets has lured me in many directions.

Data exploration can be very **meaningful** when we are passionate about the topic being addressed. Like millions of fathers, I care about the well being of my children and by extension people of the world. My concern for people has led me to produce many graphics about the environment and human health. If you have a strong interest, this class gives you the chance to pursuit it.

We data explorers face many hazards and have many limitations that we bring with us. The

hazards, of course, include **data problems** such as the lack of data, poor data, purposely deceptive data, and unrepresentative data. The quality of quantitative graphics or models is limited by the quality of the data. I share some of my learning experiences related to “data quality” and some of conjectures about reasons data is not collected.

Noise in the data complicates finding real patterns, Often we need models to help us see through the noise. We are so good at seeing patterns that we often need models to warn us that the patterns we see are very likely based on noise.

As we become aware of all the **perceptual and cognitive limitations** we bring with us, we can remember to smile at ourselves before we move forward. We are visually plagued by change blindness and inattention blindness. Yes, people do drive into railroad trains while attending to something else. In terms of mental transforms of data we are saddled with tiny working memories. We have to use computing tools to help us see and to refine what we vaguely envision in our minds. We suffer from many well-documented reasoning biases such as anchoring. The result from an irrelevant spinner influences our starting guess at the number of beans in a jar. Our first graph can influence our thinking well past the time when we realized it was wrong. In our eagerness to solve problems we often solve the wrong problem. In statistics this is known as the TYPE III error. In this class you are encouraged to face such limitations and to appreciate your amazing abilities.

Communicating with ourselves can be hard enough, but when we have found something we still need to **communicate with other people**. We need language to augment the graphics, by directing attention and telling stories. Positive unambiguous language and consistent metaphors help make communication pleasant and clear. However in statistics we are often saddled with negative language such as penalty functions and error rates and language, such as “false positives” and “false negatives” that often confuse people. This class doesn’t attempt to solve the many statistical language problems. However communication can get better with practice so the class includes both oral presentations and written papers. For some students this may be their first presentation in English. All students are treated with respect and presentations are followed by applause.

Students are expected to apply the guidance and advocated encoding methods in a **midterm graph redesign project**. The redesign project involves finding a bad or improvable graph (or table) and improving it. Students are also expected to apply the class guidance in the final project. More details about the redesign and final projects are available on class web site. Ideally students leave this class seeing quantitative graphics and communication in a new and brighter light and use vision this in the rest of their careers.

I endeavor to make students aware of **recent developments related to quantitative graphics**. This includes new web sites, books, graphics templates, hardware, and software. I am quite interested in citizen science projects on the web such as galaxy zoo (<http://www.galaxyzoo.org/>). Your many young eyes are way more powerful than my two old eyes. I appreciate students enlightening me about developments in quantitative graphics and in related areas such as computer science and cognitive science.