Statistics 695V: Data Visualization, Spring 2005

Data Visualization

All areas of learning from data — statistics, machine learning, and data mining — can benefit immensely from data visualization. Visualization provides a front line of attack in the analysis of data, revealing intricate structure that cannot be absorbed in any other way. We discover unimagined effects, and we challenge imagined ones.

Many approaches to learning from data today involve the use of complex tools that extent tailor themselves to the patterns of the data, a form of automated learning. But human guidance from data visualizations added to the algorithms can vastly increase their performance. Other approaches involve the building of complex statistical models that are used to describe relationships among variables and to carry out probabilistic inductive inference. Data visualization enables the building of models that follow the patterns of the data, resulting in valid inferences.

Tools matter. There are exceptionally powerful visualization tools, and there are others, some well known, that rarely outperform the best ones. The analyst learning from data needs to be hard-boiled in evaluating the efficacy of a visualization tool. It is easy to be dazzled by a display of data, especially if it is rendered with color or depth. Our tendency is to be misled into thinking we are absorbing relevant information when we see a lot. But the success of a visualization tool should be based solely on the amount we learn through the data about the phenomenon under study. Some tools in the course are new and some are old, but all have a proven record of success in the analysis of common types of statistical data that arise in science and technology.

Statistics 695V is a Research Course

While the prerequisites do not require previous knowledge of data visualization, the course nevertheless has a research orientation. The objectives are to provide students with the opportunity to

- understand basic frameworks for data visualization
- understand in depth selected tools for data visualization
- experience using data visualization to display data
- experience surveying, evaluating, and in some cases improving research ideas in a specific area of data visualization
- experience giving a research talk
- experience writing a research paper.

Prerequisites, Class Limit, Permissions, and Questions

Prerequisites are

- probability: basic
- statistics: basic, including least-squares fitting of parametric functions to data
- mathematics: calculus and linear algebra
- data visualization: no previous knowledge needed

The level of the course will be that of the book *Visualizing Data* by the instructor.

Permission of the instructor is required. The class will be limited to 20 students. Send questions about the course or requests for permission to attend to the instructor at wsc@stat.purdue.edu.

Student Responsibilities

Students will form groups of size one up to a maximum size that will be determined by the class size. Each group will select either (1) a methodological area of data visualization or (2) a set of data. If (1), a group will

- read papers in the area
- give a talk on the area
- prepare a paper on the area that reviews and evaluates it, and if desired, try out a new idea and discuss in the paper.

If (2), a group will

- use a collection of visualization tools to study the data set
- evaluate the use of the tools to show important characteristics of the data and lead to conclusions about the subject matter.
- give a talk on the conclusions
- prepare a paper on the conclusions.

The TA will help facilitate the formation of groups, which should be completed by Jan. 21. Each group should discuss the selection of the area or data set with the instructor. Groups should be formed and topics selected by Feb. 9.

The word or latex template from the 2005 IEEE InfoVis Conference should be used for the paper. The 2004 templates may be found at 142.58.111.29/∼vis/Tasks/camera.html but it is expected that the 2005 templates will be available in Jan. The paper should be sent to the TA electronically as pdf by May 3.

Talks should be given from and either the classroom computer or a student computer using the classroom projector. At the end of each presentation, there will be an active student question and discussion session.

The course depends heavily on teamwork, joint planning, and feedback. It is important that students attend classes to foster this.

For all projects students should use S or R for the data visualization. The reason is that there is a particularly rich set of tools in S and R, and the course methodology is closely linked with S and R. Part of the course lectures will be devoted to design issues for the S and R graphics software.

Computing Environments

A research computing environment will be provided for students to carry out their work. This new instructional endeavor to provide such an environment is being carried out by ITaP and the Statistics Department. The operating system will be linux running on a cluster of computers with the S language (S-PLUS and R implementations).

Proposed Instructor Lecture Topics

Lectures will take material in part from the book *Visualizing Data*. In addition there will be lectures about the trellis display framework for visualizing multidimensional databases, visualization of massive databases, and the design principles of the trellis display visualization framework in the S language. Written material, papers and book chapters, will be provided for the additional lecture topics.

Course Instructor

William S. Cleveland has been a Professor of Statistics and Computer Science at Purdue University since January 2004. Previous to this he was a Distinguished Member of Technical Staff in the Statistics and Data Mining Research Department at Bell Labs.

His areas of research include machine learning, data mining, data visualization, statistical methods and models, and computer networking.

Cleveland has introduced tools for local learning, as well as many visualization tools, that are widely used in engineering, science, medicine, and business. He has participated in the design and implementation of software for these tools that is now a part of many commercial systems. He has been involved in many projects applying learning and visualization tools to data from several fields including environmental science, customer opinion polling, visual perception, and computer networking.

His books *The Elements of Graphing Data* and *Visu-*

alizing Data have been reviewed in many journals from a wide variety of disciplines, and *Elements* was selected for the Library of Science. More information about them, including reviews, is available at amazon.com and www.stat.purdue.edu/∼wsc/.