Data Visualization (CSE 578)

Note: Below outline is subject to modifications and updates.

About this Course
Visual representations generated by statistical models help us to make sense of large, complex datasets through interactive exploration, thereby enabling big data to realize its potential for informing decisions. This course covers techniques and algorithms for creating effective visualizations based on principles from graphic design, visual art, perceptual psychology, and cognitive science to enhance the understanding of complex data.

Specific topics covered include
- data transformations
- exploratory querying
- statistical graphics
- time series analysis
- exploratory spatial data analysis

Learning Outcomes
Learners completing this course will be able to
- Develop exploratory data analysis and visualization tools using Python and Jupyter notebooks
- Apply design principles for a variety of statistical graphics and visualizations including scatterplots, line charts, histograms, and choropleth maps
- Combine exploratory queries, graphics, and interaction to develop functional tools for exploratory data analysis and visualization

Projects:
Project 1: Analyzing Theme Park Patronage
Project 2: Analyzing Wait Times and Dynamics of Theme Park Patronage
Project 3: Exploring and Clustering Trajectories in a Theme Park
Project 4: Finding Commonalities Between Theme Park Patrons
Project 5: Design a Visual Analytics System for Exploring Theme Park Dynamics

Course Content

<table>
<thead>
<tr>
<th>Instruction</th>
<th>Assessments</th>
</tr>
</thead>
<tbody>
<tr>
<td>Video Lectures</td>
<td>Practice activities and quizzes (auto-graded)</td>
</tr>
<tr>
<td>Other Videos</td>
<td>Practice assignments (instructor- or peer-reviewed)</td>
</tr>
<tr>
<td>Readings</td>
<td></td>
</tr>
</tbody>
</table>

Data Visualization
Lead: Ross Maciejewski, Ph.D.
Updated 12/28/2017
Estimated Workload/Time Commitment Per Week
Approximately 9 hours per week

Required Prior Knowledge and Skills
- Basic statistics and computer science knowledge including computer organization and architecture, discrete mathematics, data structures, and algorithms
- Knowledge of high-level programming languages (e.g., C++, Java) and scripting language (e.g., Python)

Technology Requirements

Hardware
Standard with major OS

Software and Other
To complete course projects, the following application will be required: Jupyter Notebooks.

Course Outline

Unit 1: Introduction to Data Visualization

Learning Objectives
1.1 Define visualization.
1.2 Describe key purposes of visualization.
1.3 Identify the data types of elements within a dataset’s data.
1.4 Demonstrate the process of loading datasets for analysis into Python.
1.5 Explain data warehouses and exploratory querying.
1.6 Define the properties of Bertin’s visual variables.
1.7 Identify appropriate color schemes for different data types.
1.8 Explain how combinations of visual variables impact the usefulness of a graphic.

- Unit Introduction
- Core Concepts
Unit 2: Introduction to Statistical Graphics

Learning Objectives
2.1 Describe exploratory data analysis
2.2 Define design principles for pie charts and donut charts
2.3 Define design principles for bar charts and line charts
2.4 Define design principles for histograms and the impact of parameter choices on the visualization
2.5 Understand how to create and use box-plots and Q-Q Plots

- Unit Introduction
- Exploratory Data Analysis
  - What is Exploratory Data Analysis?
- Design principles for Pie and Donut Charts
  - Introduction to Pie Charts
- Design Principles for Bar Charts and Line Charts
  - Bar and Line Charts
  - Design Considerations for Non-Data Components of Graphs
- Design principles for Histograms
  - Creating Histograms
Unit 3: Multivariate Analysis

Learning Objectives
3.1 Describe attributes of multivariate data visualization
3.2 Apply methods of visualizing discrete data values along two axes
3.3 Identify issues associated with parallel coordinate plots
3.4 Compare correlation and covariance

Define supervised learning and describe supervised learning methods
Describe methods for evaluating supervised learning
Define unsupervised learning and describe unsupervised learning methods
Describe methods for evaluating unsupervised learning

- Unit Introduction
- Multivariate Analysis
  - Introduction to Scatterplots
- Mosaic Plots and Pixel Based Displays
  - Mosaic Plots
  - Pixel Based Displays
- Parallel Coordinate Plots
- Text Visualization
- Jupyter Notebook: Advanced Graphics in Python
- Supervised and Unsupervised Learning
  - Data Visualization and Machine Learning Connection
  - Supervised Learning
  - Nearest Neighbor Classifier
  - Regression
  - Evaluation
- Unsupervised Learning
  - Introduction to Unsupervised Learning
  - Evaluation
- Jupyter Notebook: Machine Learning
- Assignment 1: Dino Fun World Analysis
Unit 4: Temporal Analysis

Learning Objectives
4.1 Learners completing this course will be able to:
4.2 Apply methods of temporal analysis
4.3 Perform a multivariate visual analysis

- Unit Introduction
- Module 1: Aigner classification
- Module 2: Time series modeling
- Module 3: Time series motifs

Unit 5: Geographical Data Analysis

Learning Objectives
5.1 Describe tools and techniques that are designed to support analyses that focus on datasets with a geographic component
5.2 Differentiate types of geographic visualizations
5.3 Explain spatial statistics
5.4 Apply methods of spatial analysis
5.5 Develop a geographic visualization

- Unit Introduction
- Module 1: Introduction to Geographic Analysis and Visualization
  - Introduction to Geographical Analysis
  - Thematic Maps
  - Coordinate System
  - Map Design: Map projections
  - Map Design: Map Elements and Typography
- Module 2: Choropleth Maps
  - Introduction to Choropleth Maps and Color Schemes
  - Data Classifications
- Module 3: Common Geographic Visualizations
  - Common Geographic Visualizations
- Module 4: Spatial Statistics
  - Spatial Statistics
- Module 5: Spatial Autocorrelation
  - Spatial Autocorrelation
- Module 6: Spatial Scan Statistics
Unit 6: Hierarchical Data Analysis

**Learning Objectives**
6.1 Explain hierarchical representation schemes
6.2 Apply methods of hierarchical data analysis
6.3 Appraise a cluster visualization

- Unit Introduction
- Module 1: Hierarchical Clustering
  - Single Link
  - Centroid
  - Ward’s Distance
- Module 2: Hierarchical Data Analysis
  - Tree Maps
  - Dendrogram

<LOs and outline not yet confirmed>

Unit 7: Additional Tools Used for Data Visualization

*In development*

**Creators**
Established in Tempe in 1885, Arizona State University (ASU) has developed a new model for the American Research University, creating an institution that is committed to access, excellence and impact.

As the prototype for a New American University, ASU pursues research that contributes to the public good, and ASU assumes major responsibility for the economic, social and cultural vitality of the communities that surround it. Recognizing the university’s groundbreaking initiatives, partnerships, programs and research, *U.S. News and World Report* has named ASU as the most innovative university all three years it has had the category.

The innovation ranking is due at least in part to a more than 80 percent improvement in ASU’s graduation rate in the past 15 years, the fact that ASU is the fastest-growing research university in the country and the emphasis on inclusion and student success that has led to more than 50 percent of the school’s in-state freshman coming from minority backgrounds.
Ross Maciejewski

Ross Maciejewski (Dr. Ross) is an Associate Professor at Arizona State University in the School of Computing, Informatics & Decision Systems Engineering and Director of the Center for Accelerating Operational Efficiency, a Department of Homeland Security Center of Excellence. His primary research interests are in the areas of geographical visualization and visual analytics focusing on public health, dietary analysis, social media, criminal incident reports, and the food-energy-water nexus.

Huan Liu

Professor Huan Liu joined ASU in 2000 after conducting research in Telecom (Telstra) Australia Research labs and teaching at the National University of Singapore. He has extensive experience in research and development. Liu’s research and teaching focuses on machine learning, data mining, and real world applications.

Selcuk Candan

K. Selcuk Candan is a professor of computer science and engineering at Arizona State
University and the director of ASU’s Center for Assured and Scalable Data Engineering (CASCADE). His primary research interest is in the area of management and analysis of non-traditional, heterogeneous, and imprecise (such as multimedia, web, and scientific) data.