**Midterm Review CS4460**

**Intro**

What is Information Visualization?

* **Engineers Def**: Binding data values and structure to graphical elements on a display
* **Prof basic Def**: Preventing data via interactive charts, graphs, maps so that users can understand the data, answer questions about the data and gain insights from the data (interaction is key)
* Why?
  + Help people **think about their data**
  + Help people answer **questions** (and come up with new ones)
  + Data **usability**
* **Def**: use of computer cupported, interactive visual representations of data to **amplify cognition**
* Symbolic into the geometric
* Depiction of information using spatial or graphical representations
* Making use of visual system

Applications:

* Financial/ business data – “business intelligence”
* Internet information – traffic, topology
* Software development
* Intelligence analysis
* Social networking analysis
* Weather
* Sports
* Politics

What is “information”?

* Items, entities, things which do not necessarily have a direct physical correspondence
* Often collected from the “real world”, and through sensors, made into “information” (or “data”)
* Abstractness of the entities is important too.

Challenge of info vis

* Transform data into knowledge, making it useful
  + Understand data
  + Make decisions based on understanding
* Visual analysis
  + Integrate ML and DM

Anscombe’s Quartet – tables or graph?

* Stats same for each (x-y table (quartet))
  + Mean=9, variance = 9, correlation = 0.816, linear regression: y=3 + 0.5x
* What’s different? => graph the data => see different understanding of the data

SciVis: scientific data visualization && MedVis: medical data visualization

* Data generally associated with physical positions in a 2d or 3d space – has a geometry

**Overview:**

Main Idea

* Visuals help us think (provide a frame of reference, a temporary storage area)
* Cognition=> perception
* Pattern matching
* External cognition aid
* Can yield conclusions to questions not even originally asked, there can be surprises

Analysis (purpose 1)

* Understand your data better and act upon that understanding
* Given all the data, then: understand, compare, decide, judge, evaluate, access, determine, etc
* Solving problems, answering questions, and performing tasks, accomplishing goals

Presentation (purpose 2)

* Communicate and inform others more effectively
* Visuals can serve as evidence or support

When to Use Data vis?

* Exploratory data analysis
* Let yourself discover, but also explain and answer specific questions

Examples

* Business
  + Did a recent marketing campaign work as intended?
  + How do sales reflect this? By region? By demographic? Compared to norms?
  + Did their product improve in quality, or just perceived quality? What defines quality?
  + How does their product compare to competitors?
* Airlines
  + What are the key factors causing flight delays in the US?
  + Are delays worse in the summer or winter?
  + Is the seasonal effect influenced by geographic location?
  + How does competition at an airport affect flight delays?

Components of info vis

* Representation
* Interaction

**Multivariate Data, Basic Charts:**

Basic Data Types

* Nominal (categorical)
  + Equal or not equal to other values (ex gender)
* Ordinal
  + Obeys< relations, ordered set
  + Ex: freshman, sophomore, junior, senior
* Quantitative
  + Can do math, equal intervals
  + Ex: distance, weight, temp, population count, age

Data Marks:

* Visual primitives in 2d or 3d space
  + Points, lines, areas, volumes
* Graphical properties of data marks encode data valuse
  + Size, shape, color (HSV), orientation, texture, border, thickness
* Info vis make up of data marks placed inside of visual metaphors, and mapped to data values
* Slide (3 and 4 for data marks examples)

Terminology

* Data items: commonly called data cases
* Data attributes: also often referred to as data features, data dimensions, and data variables
* Data marks: visual encodings, visual marks, graphical encodings

Number of variables:

* Univariate data (1)
  + Think of data items as being shown along one dimension and the value in another
* Bivariate data (2)
  + Scatterplot is common
  + 2 variables, want to see relationship
  + Is there a linear, curved or random pattern?
* Tri-variate data (3)
  + 3d representation of data is not great
  + Still use 2d but have visual mark property represent third variable (size)
  + Represent each in its own explicit way
* Hyper-variate data (>3)

Few’s selection & design process:

* Determine your message and identify your data
* Determine if a table, or graph, or both is needed to communicate your message
* Determine the best means to encode the values
* Determine where to display each variable
* Determine the best design for the remaining objects
  + Determine the range of the quantitative scale
  + If a legend is required, determine where to place it
  + Determine the best location for the quantitative scale
  + Determine if grid lines are required
  + Determine what descriptive text is needed
* Determine if particular data should be featured and how

Points

* Useful in scatter plots for 2 valuse
* Can apply shapes, color for additional variables to encode

Lines

* Connect values in a series
* Show changes, tends, patterns
* Not for a set of nominal or ordinal values

Bars

* Emphasizes individual values
* Good for comparing individual values (when along common axis)
* Horizontal bars are good if long labels or many items
* Multiple grouped bars can be used to encode another variable

Boxes

* Show distribution of values

Hyper-variate data examples:

* Parallel coordinates
* Dimensional reordering: changing order of dimensions creates new visual patterns
* Chernoff faces
* Star plots

**Perception and Gestalt**

Colin Ware, simple model

* Two stage process
  + 1: Parallel extraction of low-level properties of scene (color, texture, shape, spatial attributes)
    - Automatically
    - Rapid
    - Bottom up data driven model of processing
    - “pre-attentive” processing
  + 2: Sequential goal directed processing (of object identification (using memory) and spatial layout action)
    - Slow serial processing
    - Involves working long term memory
    - Top down processing

How does human visual system analyze images?

* Without need for focused attention
* Less than 200-250 msecs
* Done in parallel by low-level vision system
* “mental shortcut”

Gestalt Principles

* Proximity
  + Gorup together
* Closure
  + Creating a larger, more complete object
* Similarity
  + Group similar objects together
* Symmetry
* Continuity
  + Separate overlapping objects to give them a smooth interpretation
* Common fate
  + Moving in the same direction
* Figure/ground
* Connectedness
  + Connected visual elements are seen as a whole
* German: Gestalt – “essence or shape of an entity’s complete form”

**Multivariate Visual Rep**

View Slides:

**Tufte Few:**

Graphical Excellence

* Well-designed presentation of interesting data --- a matter of substance, of statistics, and of design
* Clarity, precision, and efficiency
* Telling the truth about the data

Graphical integrity:

* Graphic should tell truth about your data
* Baseline?
* Scale?
* Context?

Measuring misrepresentation

* Visual attribute value should be directly proportional to data attribute value
* Lie factor = (size of effect shown in graphic) / (size of effect in data)

Design Aesthetics

* Set of principles to help guide designers
* Maximize data ink ration
  + Data ink ratio = (data ink) / (total ink used in graphic)
* Show the data
* Erase non-data ink
* Erase redundant data ink
* Revise and edit

Design principle

* Utilize multifunctioning graphical elements (macro/micro readings)
  + Graphical elements that convey data information and design function
* Small multiples
  + Repeat visually similar graphical elements nearby rather than spreading far apart
* Utilize narratives of space and time
  + Tell a story of position and chronology through visual elements
* Content is king
  + Quality, relevance and integrity of the content is fundamental

Graph and chart tips

* Avoid separate legends and keys – just have that info in the graphic
* Make grids, labeling, etc. very faint so that they recede in the back ground

**HTML CSS SVG**

HTML: Hyper Text Markup Language

* Standard for world wide web
* Syntax: <tagname>Content</tagname>
* <!DOCTYPE html>
* <html>
* <head> <title>tittle text</title> </head> //title, links to stylesheets, etc.
* <body> <h1>text</h1> </body> //visible content
* </html>

CSS: Cascading Style Sheets

|  |  |  |
| --- | --- | --- |
| syntax | example | What it does |
| .class | .key-names | Select all element with class=”key-names” |
| #id | #tufte | Select the element with id=”tufte” |
| Element | h1 | Select all h1 elements |
| :hover | a:hover | Select links on mouse over |

SVG: Scalable Vector Graphics

* Renders graphics with shapes, coordinates, etc.
* <svg width="100" height="100" style="border: 1px solid black;">
  + <circle cx="50" cy="50" r="40" stroke="red" stroke-width="4" fill="#ffc0cb" />
* </svg>

**Tasks and Analysis**

Why do we need task taxonomies?

* To build better visualizations, we need to understand what people might use them for
  + What tasks do they want to accomplish?
  + What models/diagrams for those tasks?
  + What do we know about the tasks that people perform?
  + How do people want to make sense of their data?
* Task Taxonomies:
  + Structural decompositions of concepts to help us understand categories and their relationships
  + Important to understand what process they focus on
    - Creating understanding
    - Creating a final vis that presents something
  + Low and high level taxonomies exist

User tasks

* Low level, domain independent taxonomy of user tasks in visualization environments
* 11 basic action:
* Identify, locate, distinguish, categorize, cluster, distribution, rank, compare within relations, compare between relations, associate, correlate
* Task x data type taxonomy to understand what people do with visualization
* Data case: an entity in the data set (data item)
* Attribute: a value measure for all data cases (dimension, variable)
* Aggregation function: a function that create a numeric representation for a set of data cases ( average, count, sum)

1. Retrieve value
   * Given a set of specific cases, find attributes of those cases (ie. mpg of audi?, how long movie)
2. Filter
   * Given some concrete conditions on attribute values, find data casese satisfying those conditions (and hide others) (ie. cereals have high fiber?, comedies have won awards?)
3. Compute derived value
   * Given set of data cases, compute an aggregate numeric representation of those data cases (how many manufacturers of cars are there? average calorie content of post cereals?)
4. Find extremum
   * Find data cases possessing an extreme value of an attribute over its range within the data set (highest mpg car? Director/film most awards?)
5. Sort
   * Rank (rank cereals by calories)
6. Determine range
   * Find span of values within the set (range of film lengths? Range of car horsepowers?)
7. Characterize distribution
   * Characterize the distribution of that attribute’s values over the set (distribution of carbs in cereals? Age distribution of shoppers?)
8. Find anomalies
   * Identify any anomalies (outliers in protein? Exceptions to relationship btwn horsepower and acceleration?)
9. Cluster
   * Find clusters of data items with similar attribute values (cluster of typical film lengths?)
10. Correlate
    * Determine useful relationships between the values of those attributes (correlation btwn carbs and fats? Country of origin and MPG?)

**Time Series, geo, storytelling**