Exam Date:

1. Thursday, February 21

Exam Format:

1. Multiple choices
2. Listing
3. Sketching
4. Short answer

Online Resources:

PRACTICE QUESTIONS (Staskos class from a while ago. Not all are relevant): https://gist.github.com/ldong/151820f6adb968c94844

(question 57 is really good) Answer:

- 

- **But it's the contrast effect tho**

Mosaic Plots explained:
https://ncss-wpengine.netdna-ssl.com/wp-content/themes/ncss/pdf/Procedures/NCSS/Mosaic_Plots.pdf
Graphical Integrity:
http://jcsites.juniata.edu/faculty/rhodes/ida/graphicalIntRedes.html
FLINA:
https://pure.tue.nl/ws/portalfiles/portal/3493590/687784949356560.pdf

Key: Blue = definitions Yellow = purpose Green = possible short answer questions
9
Possible questions/lists we'll be tested on at the bottom. If you have answers, add them!
Sample Test Questions •

- Given this data, use an appropriate visual representation to depict it. •
- Name this visual representation and what data it is best/ideally used for. •
- Given a task, what visualization technique would you choose and why. •
- Given a visualization, identify pros/cons and discuss how you could improve it

Topics:

A. InfoVis overview

a. Visualization → "use of computer supported, interactive visual representation of data to **amplify cognition**.
   i. Process of making a graphic or an image → "show me data/visualize it!"
   ii. A cognitive process → creates a mental image + understanding.
   iii. Purpose of vis: gaining insight (decision making, an explanation)
   iv. Visuals help us think, pattern matching, external cognition aid

b. **Purpose of info vis: analysis and presentation**
   i. Analysis: The process of understanding your data better and act upon the understanding (cognitive process)
      1. Understand the limits of data, compare the data, make decisions, judge and assess the quality of the data, evaluate a certain topic.
      2. Ultimately about solving problems, answering questions, performing tasks and solving goals.
         a. Especially useful when questions are vague and you want to explore the dataset (aka **exploratory data analysis)**
      3. Not always useful:
         a. Does not provide explicit results like that of a specific query or statistical analysis (e.g. average score: 80.56)
         b. When you already know what questions to ask or what patterns you are looking for, maybe less effective
         c. Can use other techniques for data analysis such as statistics, data mining, machine learning, DB querying/searching → Good for when you know what questions to ask and there is a well-defined stop rule.
      4. Analytics tasks:
         a. Search vs Browsing
            i. Browsing is better: Look over or inspect something in a more casual manner, seek interesting information
         b. Comparison-difference
         c. Outliers, extremes
         d. patterns

   ii. Presentation: Communicate and inform others the data more effectively.
      1. Use visualization to communicate ideas, influence, explain and persuade.
      2. Visuals can serve as evidence or support.
         a. Must be careful: can potentially **misrepresent** the data.
      3. Visuals can often make things much easier and replace words. But sometimes words are needed such as captions, axis, labels

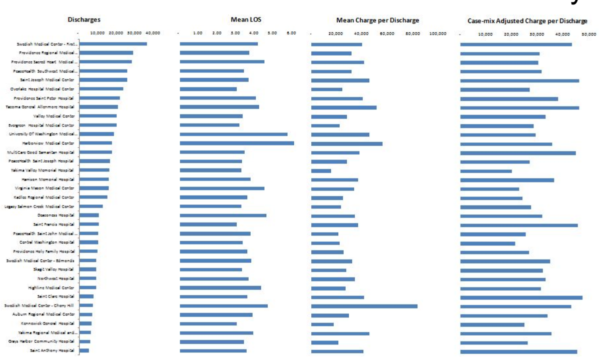
c. Goal: transform data into understandable information so that it is useful for people.

B. Multivariate Data & Basic Charts

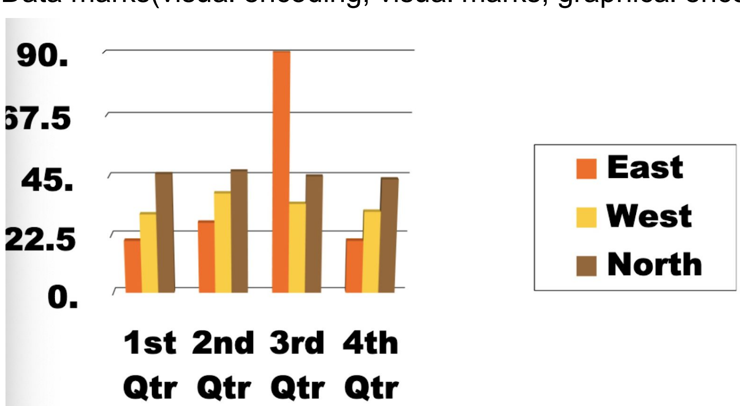
a. Basic data types
   i. **Nominal** (Categorical -> Color, Type, Gender) → equal or not equal to other values
   ii. **Ordinal** (Obeys an ordered set → FR, SP, JR, SR)
   iii. **Quantitative** (Distance, Weight) → can do math, equal intervals

b. Data Marks: Visual primitives in 2D or 3D space like points, lines, etc.

        i. Placed in visual metaphors and mapped to values
        ii. Can vary in effectiveness
    c. Basics of visual representation:
        i. **Binding** data values to visual glyphs
        ii. **Emphasize** salient features of data for perceptual system
        iii. **Augment** cognitive processes of people to reason about data
    d. Basic charts
        i. Scatterplot matrix = small multiples, a scatterplot for each pair of variables
        ii. Chernoff faces = encode variables' values in the facial characteristics
        iii. Star plots = each spoke represents variable, can connect the lines forming a star shape to represent a multivariable data point
        iv. Parallel Coordinates = vertical line for each variable, can have multiple data points across multiple variables
    e. TableLens
        i. Spreadsheet = hypervariate data presentation
            1. Each variable is positioned into a column
            2. Data cases in rows
            3. It is a projection (mapping)
            4. TableLens makes the text more visual and symbolic



    f. Data items(data cases)
    Data attributes(data features, dimensions, variables)
    Data marks(visual encoding, visual marks, graphical encodings)



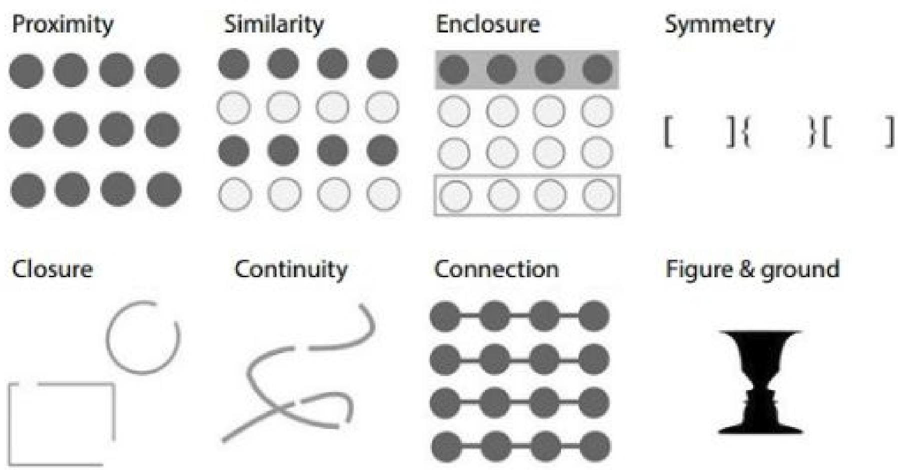
    Q: How many cases, items? A: 12
    Q: How many variables? 3 trivariate
    Q: What are variables type? A: Region- Nominal Sales-Quantitative Quarter-Ordinal
    g. Color → choose wisely,

ordering of color should be used when data has order.
Separation of color = separation of data
Limits to # of colors. Less is better! :)

C. **Perception & Gestalt**
   a. **Perceptual Processing** → Seek to better understand visual perception and visual information processing
      i. Simple Model (two stages)
         1. Stage 1 - low-level, Parallel
            a. **Rapid**
            b. Occurs **"automatically"**
            c. Often called **"pre-attentive" processing** (bottom-up, data driven)
            d. Parallel detection of color, texture, shape, spatial attributes
            e. Pre-attentive processing
               i. w/o need for focused attention
               ii. Done in parallel by low-level vision system
               iii. "Mental shortcut"
               iv. good for: target detection(sth there), boundary detection(element group) and counting(#elements of a type)! Eg: red 3's in a group of numbers (Easier to find)

         2. Stage 2 - Sequential, Goal-Directed
            a. Splits into subsystems for object recognition and for interacting with the environment
            b. Increasing evidence supports the independence of systems for symbolic object manipulation and for locomotion & action
            c. Spatial layout
            d. **Slow serial processing**
            e. Involves working and long-term memory
            f. More emphasis on arbitrary aspects of symbols
            g. Top-down processing
      ii. Preattentive Features
         1. Hue → color! (finding red circle among blue circles, "distractor") (Pre-attentive!)
         2. Shape (Pre-attentive!)
         3. Hue and Shape (NOT pre-attentive, feature conjunction)
            a. Can cancel each other out (Red square vs Blue circle: sequential search needed)
         4. Fill and Shape(Boundary in the display, white circle+square vs black circle+square)
         5. Emergent Features (visual form == preattentive)
            a. Things that stick out when a Gestalt principle is applied (like a square in circles or a blue dot in red)

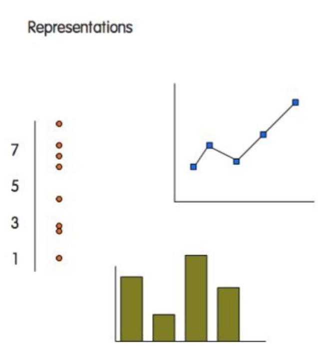
   b. Gestalt Principles - "essence or shape of an entity's complete form"



i.
ii. Proximity → group together proximate objects
    1. Stronger than similarity
iii. Closure → try to see collections of objects as creating a larger, more complete
iv. Similarity → group similar objects together
v. Symmetry
vi. Continuity −(aka Good Continuation) → separate overlapping objects to give them a “smooth” interpretation.
vii. Common Fate → group together objects to seen to be moving in the same direction.
viii. Connectedness → connected visual elements are seen as a whole.
ix. Figure/Ground

D. <u>Multivariate Data Representation</u>

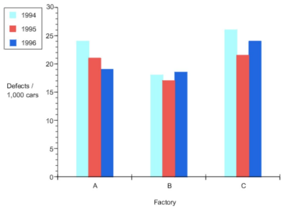
a. Univariate data - 1 variable
    i. bar charts and line graphs to show changes.
    ii. Data items shown along one dimension, and the value in another.
    iii.

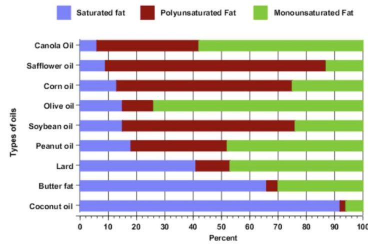


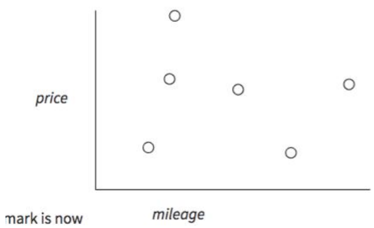
Label	Variable1
A	4
B	8
C	3
D	1

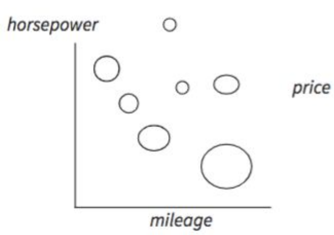
Clustered Bar Chart

Stacked Bar Chart

iv. 



b. Bivariate data - 2 variables
   i. scatter-plot to show the correlation.
   ii. Want to see the relationship(linear, curve, random pattern)
   iii. 

c. Trivariate data - 3 variables
   i. often use encoding such as size to represent 3rd variable on a 2d plane
   ii. 3D scatter plot possible but 3D isn't a great representation so the following are good examples:
   iii. 

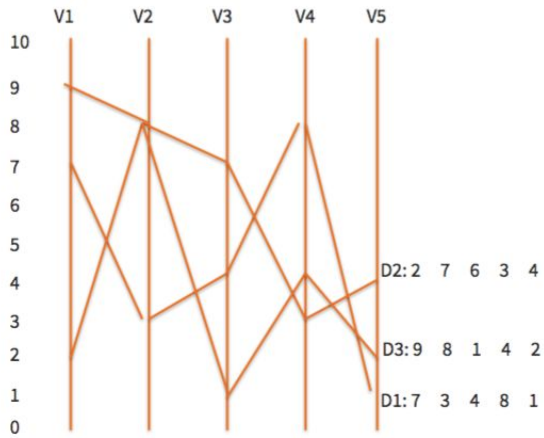
(2D, visual mark property ⇒ 3rd variable(size))
   iv. 

(each variable in explicit way, can't perceive the relations)
   v.

d. Hyper-variate data - > 3 variables
   i. Can use additional data mark properties to encode additional data variables.
   ii. Spreadsheets and instantiation
   iii. Multiple graphs view
   iv. Scatterplot matrix → represent each possible pair of variables in their own 2-D scatterplot
   v. star plots
   vi. Star coordinates (same idea as star plots)
      1. Data cases with similar values will lead to clusters of points
      2. The naive approach to projection or multidimensional scaling

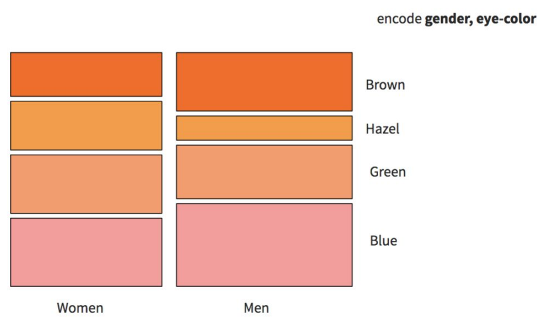
a. Axes can cancel each other

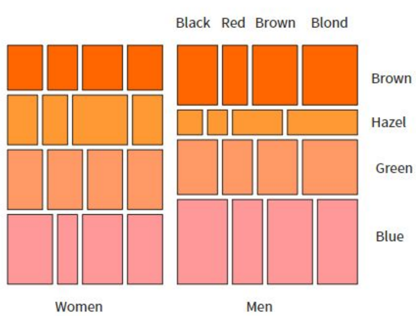
3. Parallel coordinates



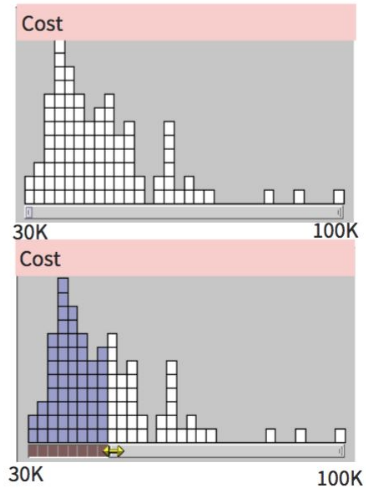
vii. Parallel Sets
1. Visualization method adopting parallel coordinates layout but uses frequency-based representation
2. User-driven!
3. Color used for different categories
4. No details on demand, at least on a case by case basis

viii. Mosaic plot

1. 

2. 

ix. Attribute Explorer
1. Multiple histogram views, one per attribute
2. Each data item represented by a square
3. Square is positioned relative to that case's value on that attribute



4. Selecting the **case in one view lights it up in others**

x. FLINA View
   1. https://pure.tue.nl/ws/portalfiles/portal/3493590/687784949356560.pdf
   2. It is a way of visualizing multivariate data that can combine multiple types of visualization that are linked together and can be rearranged. The second paragraph of the intro does a good job of describing it concisely:
      a. "In this article we propose yet another approach: Flexible Linked Axes. The method is based on a simple idea. In all methods so far the structure of the visualization is more or less fixed, and the user can only change some properties of the representations provided. We propose to enable users to define and position coordinate axes freely, and specify suitable visualizations by linking these axes. This approach enables users to define scatterplots, PCPs, and radar charts, but also to develop highly customized visual representations."

xi. Dust & Magnet
   1. Higher values will move more quickly to the correlated “attribute magnet” (if the magnet was income, the higher income values would move more quickly towards the magnet)
   2. Data cases represented as small bits of iron dust.
   3. Moving the magnet makes all the dust move
   4. Iron bits (Data) are drawn towards magnets(attributes) proportional to that data’s elements value in that attribute
   5. All magnets present on display affect position of all dust
   6. Individual power of magnets can be changed
   7. Dust color and size are connected to attributes.

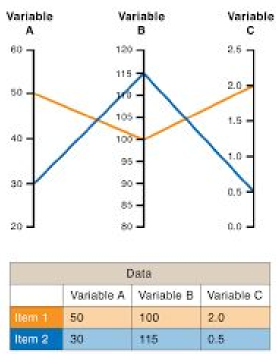
xii. Set Operations → venn diagram, bubble sets

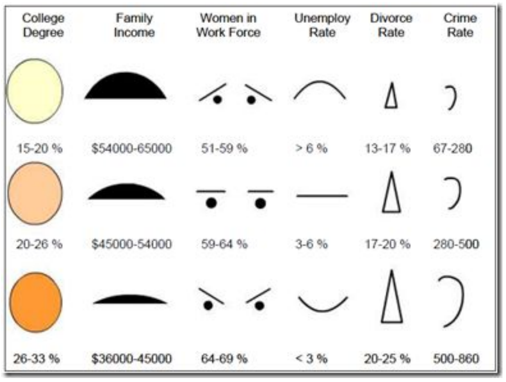
xiii. Pixel Based Systems → smallest visual representation
   1. Can scale up well at a distance
   2. Allows for details on demand
   3. Some data marks aggregate well, some don’t
   4. One pixel per variable/data cases

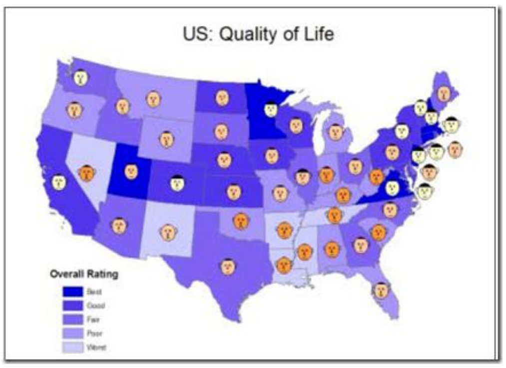
5. Color the pixel ⇒ value of one of data point's variable

xiv. Points, Lines, Bars, Boxes
1. Points: useful in scatterplots for 2-values; Can apply shapes, color for additional variable to encode.
2. Lines: connect values in a series; show changes,trends, patterns; not for a set of nominal or ordinal values
3. Bars: emphasizes individual values; good for comparing individual values (when along common axis)
4. Boxes: shows a distribution of values.

xv. Why are horizontal bars good?
1. They are good for long labels or many items.

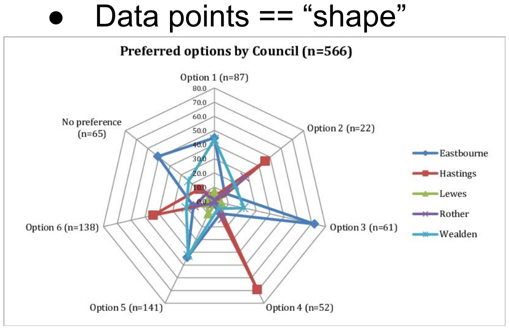
Chart Type	Pros	Cons
Parallel Coordinates 	● Representation of high dimensional as a 2-dimensional visualization ● Easy to perceive trend shown by data entries ● Axes order can be changed to check out the trend between 2 different axes (if possible). Creates new visual patterns	● Overlaying of data lines for common data values among data entries
Scatterplot Matrix ● Small multiples: each pair of variables in scatterplot	● Shows relationship between two variables ● Best method to show NON-LINEAR PATTERN ● Easily determine max, min, range of data flow ● Straightforward observation ● Relatively simple	● Unable to give exact extent of correlation ● Doesn't show quantitative measure of the relationship btw the variable(only shows quantitative expression of the quantitative change) ● Does not show the relationship for > 2 vars
Chernoff Faces: encode different variables' values in characteristics of human face	● Can be a quick way to qualitatively assess the performance/detect and comprehend important phenomena ● Remember major conclusions ● Ease of monitoring the sensitivity of variables to each other, fast identify key differentiating dimensions	● To make all faces to equal size, the width and length of each face need to be normalized(almost cancel out the effect of variables assigned to two features) ● Extreme values of certain parameter compress the range of other parameters(artificial dependencies/misrepresent)

<table>
<tr>
<td>




</td>
<td></td>
<td></td>
</tr>
<tr>
<td>
Star plots:

- Space out the n variables at equal angles around a circle
- Each encodes a variable's value
- Data points == "shape"



</td>
<td>

- Can help identify outliers (e.g. student perform well will stand out/particularly strong in one area)
- Can easily observe: "Balloon out" pattern, "Cave in" pattern
- Highlighting obvious deficiencies
- Discover subtle issues often hidden under the weight of avg.ratings
- Compare individual performance against a standard/group performance(not applicable for many axes)

</td>
<td>

- Difficult to read for most people?
- Unusual chart type may confuse audience
- Confusing when too many variables/axes
- Compare values across differ axes
- Unclear mapping(e.g. Car properties: speed is measured in miles/hour, but 'comfort' are not explicitly defined) leads to unclear share scale maps to the number
- Can't deal with nominal data

</td>
</tr>
<tr>
<td>Mosaic Plots</td>
<td>

- Useful for displaying contingency table data
- Easy to view visual attributes independently(e.g.colors, area of objects, larger or smaller objects)
- Some other attributes are easy to perceive together as a whole(integral visual dimensions)
- Good for general overview of

</td>
<td>

- Hard to make accurate comparison of the objects that are not arranged along to the baseline one next to one

</td>
</tr>
</table>

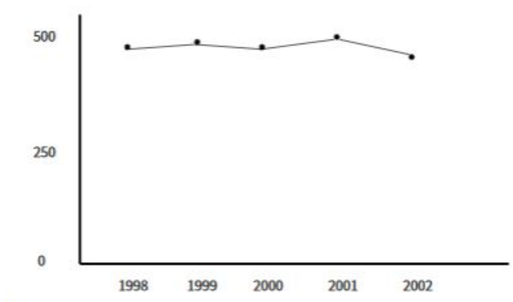
	dataset	
Parallel Sets		

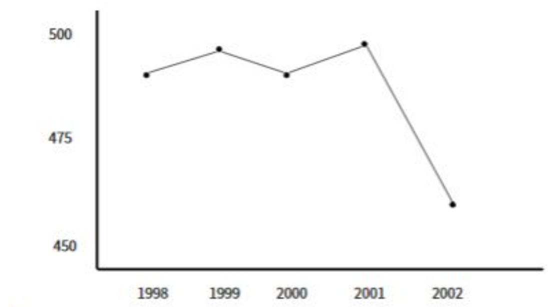
E. Design Principles

a. Tufte

i. **Graphical Excellence** -> promotion of precision and clarity in data presentation. Utilize human capabilities to present information effectively. Tell the truth!

1. **Tell the truth about the data, no misrepresentation(Graphical Integrity)**
    a. Is the data set across the relevant baseline?





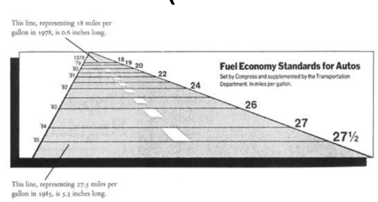
(Notice have the change in baseline/scale may change how we interpret the data)

b. Does the data use effective scaling? (show the entire scale)

c. Does the data provide enough context for interpretation?

d. Are the size encodings appropriate (area, size, volume)?

i. **Lie factor = (size of effect in graphics)/(size of effect in data)**



$$\text{Lie Factor} = \frac{\frac{5.3 - 0.6}{0.6}}{\frac{27.5 - 18}{18}} = 14.8$$

The standard required an increase in mileage from 18 to 27.5, an increase of 53%. The magnitude of increase shown in the graph is 783%, which results in a lie factor of 14.8

the second bar is much taller than the first one, even though both claim to represent the value 100

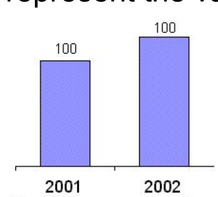


Chart 2: False representation of data

$$\text{lie factor} = \frac{\text{size of effect shown in graphic}}{\text{actual effect in data}}$$

$$= \frac{(\text{size of right bar} / \text{size of left bar})}{(\text{value of right bar} / \text{value of left bar})}$$

$$= \frac{(3{,}58\ cm / 2{,}92\ cm)}{(100\ /100)} \approx 1{,}23$$

2. **Present effectively -> aesthetics**

a. Maximize data-ink ratio → (data-ink) / (total ink used in viz)
    i. Most of the graphic should be devoted to a non-redundant display of data
    ii. Display data first
b. Avoid Chart Junk
    i. Remove extraneous visual elements (unnecessary gridlines/labels)
    ii. Note: sometimes chart junk can increase memorability of the visualization, consider the audience
c. Utilize multifunctional graphical elements (micro and macro analyzation)
    i. Ex: Stem and leaf plot -> displays overall distribution and individual data points
d. Use Small Multiples
    i. Keep similar elements near each other
    ii. Useful for static representations
e. Utilize Narratives of Time and Space
    i. Tell a story of position (space) and chronology (time)
f. Content is King
    i. Maintain quality, relevance, and integrity of data

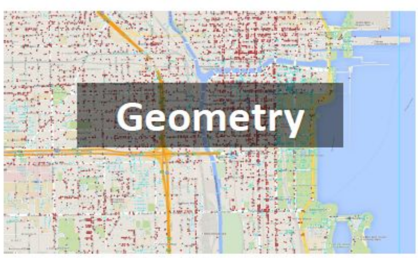
b. Few's Selection and Design Process
    i. Determine your message and identify your data
    ii. Determine if a table, graph or both is needed to communicate the message
    iii. Determine the best means to encode the values
    iv. Determine where to display each variable
    v. Determine the best design for the remaining objects
        1. Determine the range of the quantitative scale
        2. If a legend is required, determine where to place it
        3. Determine the best location for the quantitative scale
        4. Determine if grid lines are required
        5. Determine what description text is needed.
    vi. Determine if particular data should be featured and how.
c. Few: table vs charts
d.
e. Gestalt Principles
    i. proximity - group together to show similarity, membership, or as an encoding for a data variable being visualized
    ii. closure - try to see collections of objects as creating a larger, more complete object (i.e. box, WWG logo)
    iii. similarity - group similar objects together (color, shape)
    iv. symmetry
    v. continuity - separate (parse) overlapping objects to give them a smooth interpretation, even if something ain't continuous we see it as continuous
    vi. common fate - group together objects seen to be moving in the same direction
    vii. figure/ground - perception so that part of a stimulus appears to stand out as an object against a less prominent background
    viii. connectedness - connected visual elements are seen as a whole

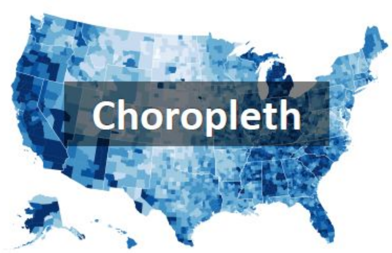
F. HTML, CSS, SVG

G. Tasks & Analysis
   a. Start with high-level tasks of searching and browsing → understanding what and how people may use our visualizations
      i. Searching and Browsing are generic → break it down
      ii. How do users create new, better questions about data?
      iii. How do users know what to search for or get the overview of the data?
   b. Task Taxonomies →
      i. Structural decomposition of user tasks to help us understand their process in using a vis
      ii. Many different types (both low and high-level):
         1. User Tasks
            a. Low-level, domain independent taxonomy of user tasks
            b. 11 basic actions: identify, locate, distinguish, categorize, cluster, distribution, rank, compare within/between relations, associate, correlate
         2. Shneiderman
            a. Mantra: "Overview first, zoom and filter, then details on demand"
            b. Task x data taxonomy to understand what people do with vis
               i. Map specific data types to specific tasks (unlike user tasks taxonomy)
         3. Amar, Eagan, & Stasko (Primary)
            a. Derived by pooling together tasks mentioned in variety of other publications and grouping them together on whiteboard
            b. Terms:
               i. Data case: entity in the data set
               ii. Attribute: value measured for all data cases
               iii. Aggregate Function: creates numeric representation of set of data cases
            c. Tasks
               i. Retrieve Value -> given set of specific cases, find attributes of those cases
               ii. Filter -> find data cases satisfying some concrete conditions of attribute values
               iii. Compute Derived Value -> compute aggregate numeric representation of set of data cases
               iv. Find Extremum -> data cases with extreme attribute values
               v. Sort -> rank data cases according to some ordinal metric
               vi. Determine range -> find span of values of some attribute
               vii. Characterize distribution -> normalize? Skewed? Ect., density of the values across the range
               viii. Find anomalies -> Weird data points/Statistical outliers, could even be bad data
               ix. Cluster -> find cluster of data items sharing similar attribute values
               x. Correlate -> determine the relationship of pairs of attributes among set of data cases
            d. Tasks can be compounded

        e. Tasks do not include:
            i. Basic Math
            ii. Uncertain criteria
            iii. High-level tasks
            iv. More qualitative comparisons
            v. How these tasks should be completed
        f. Provides set of grounded, low-level analysis tasks and language for comparing and evaluating infovis systems
    c. Interaction
        i. Tasks carried out through interaction with visualization
        ii. Users are active participants in the analysis
            1. Move viz from one state to another

H. Timeseries, Geospatial, and Storytelling
    a. Timeseries - a time taxonomy
        i. Continuous
            1. EventFlow
            2. Project management (Gantt Chart)
        ii. Discrete
    b. Geospatial
        i. Important new factors that we have to visualize
            1. **Where** is the data located?
            2. **When** did the data happen (or get captured)?
            3. **What** is it we're visualizing?
        ii. Geometry, choropleth, cartograms, fields/lines





    c. Storytelling
        i. Tell a narrative to your audience using data, visualizations, and story
    d. Summary
        i. Time is a special attribute of data that we need to pay particular attention to when visualizing.
        ii. Geo data has many alternatives that we need to consider depending on the user task
        iii. Storytelling involves an author narrative in addition to visualizing data

Possible lists we could be tested on

InfoVis Overview
Multivariate Data & Basic Charts
Perception and Gestalt

Colin Ware Simple Model

1. Stage 1 (low-level, parallel, automatically, rapid, pre-attentive processing, color, shape, etc.)
2. State 2 (slow serial processing, involves working and long-term memory)

Pre attentive processing works well for

- Target detection
- Boundary detection
- Counting

Gestalt Principles

- Proximity (group together)
- Closure (create a larger, more complete object)
- Similarity (group similar objects together)
- Symmetry (can spot symmetry and group together)
- Continuity (separate overlapping objects for a smooth interpretation)(fill in the blanks)
- Common Fate (moving in the same direction)
- Figure/Ground (looking at negative vs positive of an object)
- Connectedness (seen as a whole)

Basic Data Types

- Nominal
- Ordinal
- Quantitative

Chart Integrity Principles

- Where's the baseline
- What's the scale
- What's the context

Basic User Actions

- Identify
- Locate
- Distinguish
- Categorize
- Cluster
- Distribution
- Rank
- Compare within relations
- Compare between relations

- Associate
- Correlate

Primary Task Taxonomy (Amar, Eagan, Stasko)

1. Retrieve Value (find attributes of those cases)
2. Filter (find data cases satisfying those conditions)
3. Compute Derived Value (compute an aggregate numeric representation of those data cases)
4. Find Extremum (find data cases possessing an extreme value)
5. Sort (rank)
6. Determine Range (find the span of values)
7. Characterize Distribution (characterize the distribution of that attribute values over the set)
8. Find Anomalies (identify any anomalies)
9. Cluster (find clusters of data items with similar attribute values)
10. Correlate (determine useful relationships between the values of those attributes)

Survey of Techniques

1. Detail only (single window with x & y panning, works with zoom factor is relatively small)
2. Single window with zoom and replace (selectable zoom area, context switch can be disorienting, variations can let users pan and adjust zoomed area/level of magnification)
3. Single coordinated pair (combined display of overview and local magnified view, issue of how big are views and if they overlap, some have bigger overview, some have bigger detail)
4. Tiled multilevel browser (combined global/intermediate/detailed view, good if panning in one affects panning in others, views don't overlap)
5. Free zoom/multiple overlap (overview presented first and users selects areas to zoom, flexible layout but users must perform manual window management)
6. Bifocal magnified (magnifying glass zoomed image float over overview, neighboring objects obscured by zoomed window)
7. Fish-eye view (magnified image is distorted so that focus is at high magnification, all in one view, distortion can be disorienting)

Forms of degree of interest function

- Continuous (smooth interpolation away from focus)
- Filtering (past a certain point objects disappear)
- Step (levels or regions dictating rendering)
- Semantic changes (objects change rendering at different levels)

Sensemaking Loop

Bottom-up

- Search and filter
- Read and extract
- Schematize
- Build case
- Tell story

Top-down

- Re-evaluate
- Search for support
- Search for evidence
- Search for relations
- Search for information

Information Foraging Tasks

- Exploring (like recall in IR, increasing span of new information)
- Enriching (narrowing to a smaller set of items)
- Exploiting (generate inferences, notice pattern)

Data-Frame Model

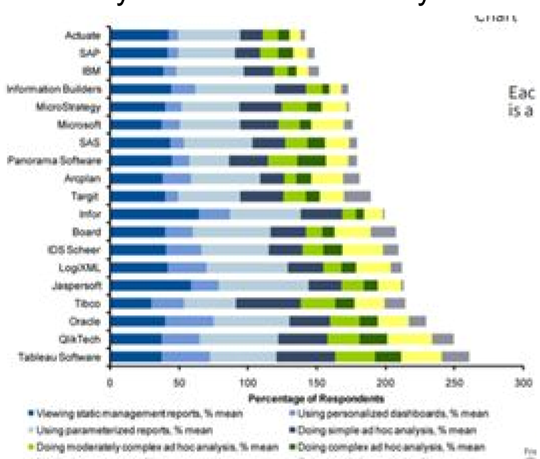
- Seeking a frame (establishing new anchors or relationships that allow you to explain data)
- Elaborate the frame (seek data to fill it out/confirm it)
- Questioning the frame (inconsistency with what you find)
- Preserving the frame (explain the anomaly in the data that made you question the frame)
- Comparing frames (sharpening and refining the distinction between frames)

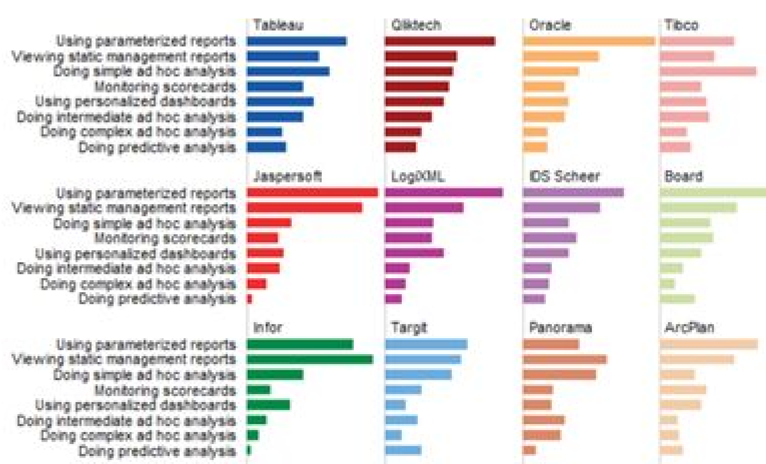
Hyper-variate data charts

- Spreadsheets and instantiation
- Multiple graphs view
- Scatterplot matrix
- Star plots
- Star coordinates
- Parallel coordinates
- Parallel Sets
- Mosaic plot
- Attribute Explorer
- FLINA View
- Dust & Magnet
- Set Operations
- Pixel Based Systems

<u>Possible (Probable) Questions</u>

- How many variables? How many cases?





-
- Why do we visualize data? (and when should we not)
    - **Why we should use info vis:** exploratory data analysis, don't have set-in-stone questions
    - **Why we shouldn't:** if we need explicit results, already know what kinds of questions to ask
- What are the 2 high-level purposes of InfoVis?

- **Analysis and presentation**

- What are the 2 components of InfoVis?
    - **Representation and interaction**
- What is Shneiderman's mantra?
    - **Overview first, zoom and filter, and details on demand**

- **Visual encodings in order of effectiveness (Ordered Attributes):**
    - 1) Position along common scale (Best)
    - 2) Position along unaligned scale
    - 3) Length
    - 4) Tile/Angle
    - 5) Area
    - 6) Depth
    - 7) Color
    - 8) Curvature
    - 9) Volume (Worst)

- Visual encodings for categorical variables in order of effectiveness:
    - 1) Spatial region
    - 2) Color hue
    - 3) Motion
    - 4) Shape
    - Note that his slide has them all listed as pretty effective