Foundations of Data Science Lecture: Intro to Visualization

Rumi Chunara, PhD

CS3943/9223

First Half..

- What is Data Science?
- Intro to R
- Data cleaning, sampling, processing
- Intro to ML – what is it
- Two Basic Algorithms
  - kNN
  - Linear Regression
- Time-series Analyses
  - Regression and lagged data in R
- Supervised Learning
  - Support Vector Machines
- Unsupervised Learning
  - PCA

Second Half So far..

- Language Models
- Networks
  - Important centrality measures
  - Domain expert: networks in biomedicine
- Visualization

Visualization

- What is Visualization?
- Visualization Principles
- Visualization Properties and Information Types
- Lying with Visuals
- Problem-solving with Visuals

Introduction

- Goals of Information Visualization
- Case Study: The Journey of the TreeMap
- Key Questions

What is Information Visualization?

"Transformation of the symbolic into the geometric"
(McCormick et al., 1987)

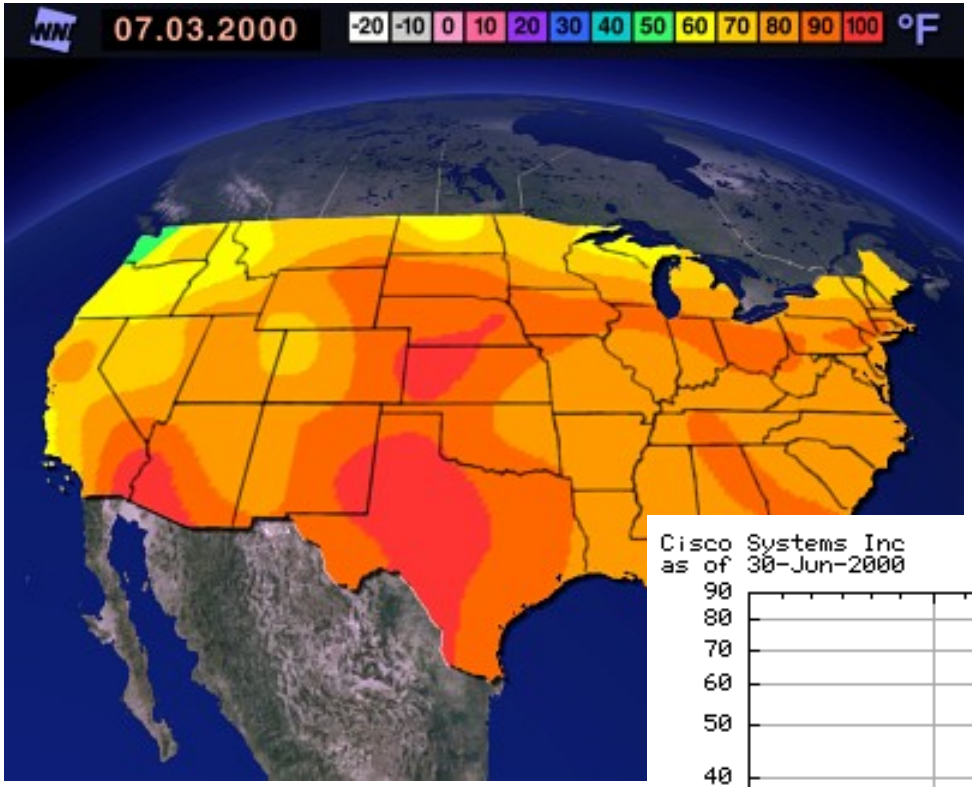
"... finding the artificial memory that best
supports our natural means of perception.''
(Bertin, 1983)

The depiction of information using spatial or graphical representations, to facilitate comparison, pattern recognition, change detection, and other cognitive skills by making use of the visual system.

Information Visualization

- Problem:
  - Large data, nuanced results, How to understand them?
- Solution
  - Take better advantage of human perceptual system
  - Convert information into a graphical representation.
- Issues
  - How to convert abstract information into graphical form?
  - Do visualizations do a better job than other methods?

Visualization Success Stories



Images from yahoo.com

The Power of Visualization

1. Start out going Southwest on ELLSWORTH AVE
   Towards BROADWAY by turning right.
2: Turn RIGHT onto BROADWAY.
3. Turn RIGHT onto QUINCY ST.
4. Turn LEFT onto CAMBRIDGE ST.
5. Turn SLIGHT RIGHT onto MASSACHUSETTS AVE.
6. Turn RIGHT onto RUSSELL ST.

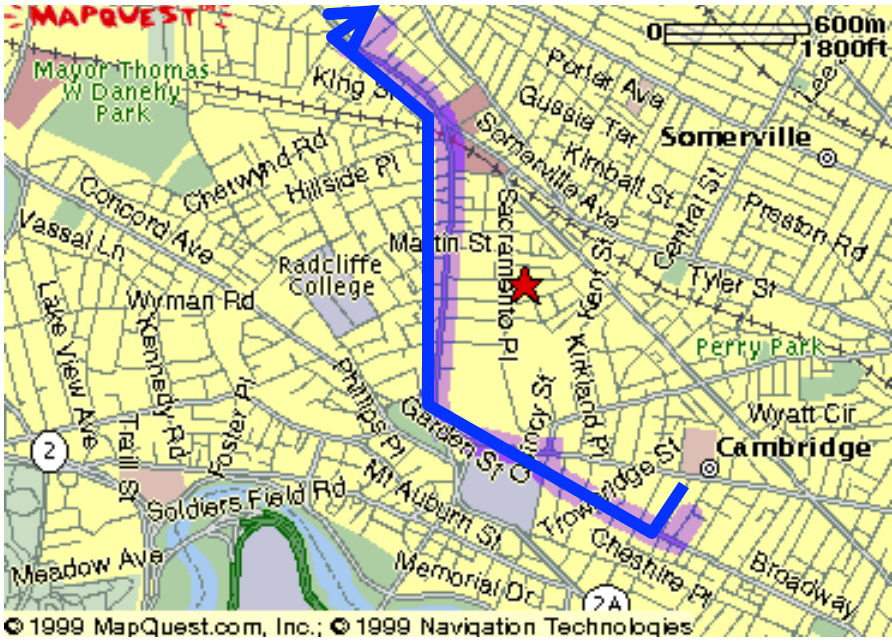


Image from mapquest.com

Visualization Success Story

Mystery: what is causing a cholera epidemic in London in 1854?

Visualization Success Story

Illustration of John Snow's deduction that a cholera epidemic was caused by a bad water pump, circa 1854.

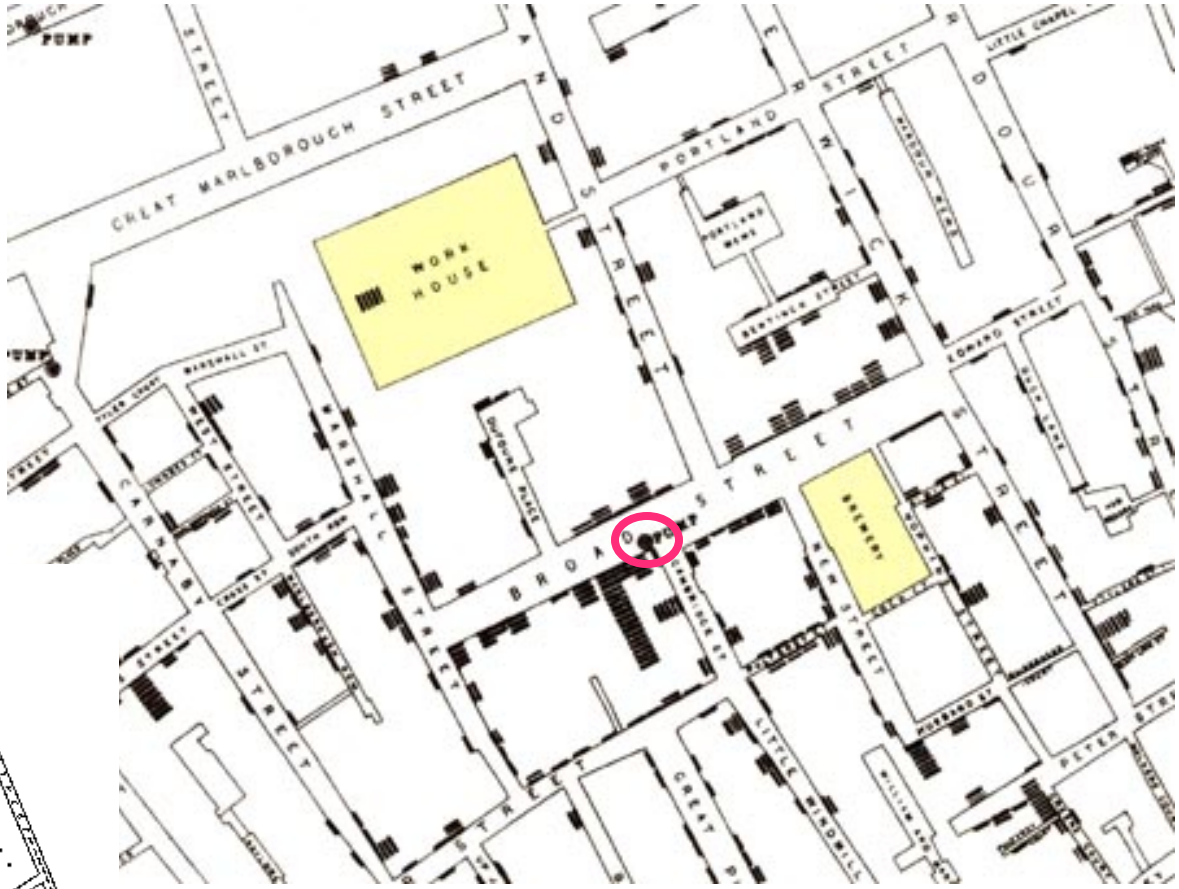
Horizontal lines indicate location of deaths.

From Visual Explanations by Edward Tufte, Graphics Press, 1997

Visualization Success Story

Illustration of John Snow's deduction that a cholera epidemic was caused by a bad water pump, circa 1854.

Horizontal lines indicate location of deaths.



From Visual Explanations by Edward Tufte, Graphics Press, 1997

Purposes of Information Visualization

To help:

- Explore
- Calculate
- Communicate
- Decorate

Two Different Primary Goals: Two Different Types of Viz

Explore/Calculate

Analyze

Reason about Information

Communicate

Explain

Make Decisions

Reason about Information

Goals of Information Visualization

More specifically, visualization should:

- Make large datasets coherent
  (Present huge amounts of information compactly)
- Present information from various viewpoints
- Present information at several levels of detail
  (from overviews to fine structure)
- Support visual comparisons
- Tell stories about the data

Why Visualization?

Use the eye for pattern recognition; people are good at

- scanning
- recognizing
- remembering images

Graphical elements facilitate comparisons via

- length
- shape
- orientation
- texture

Animation shows changes across time

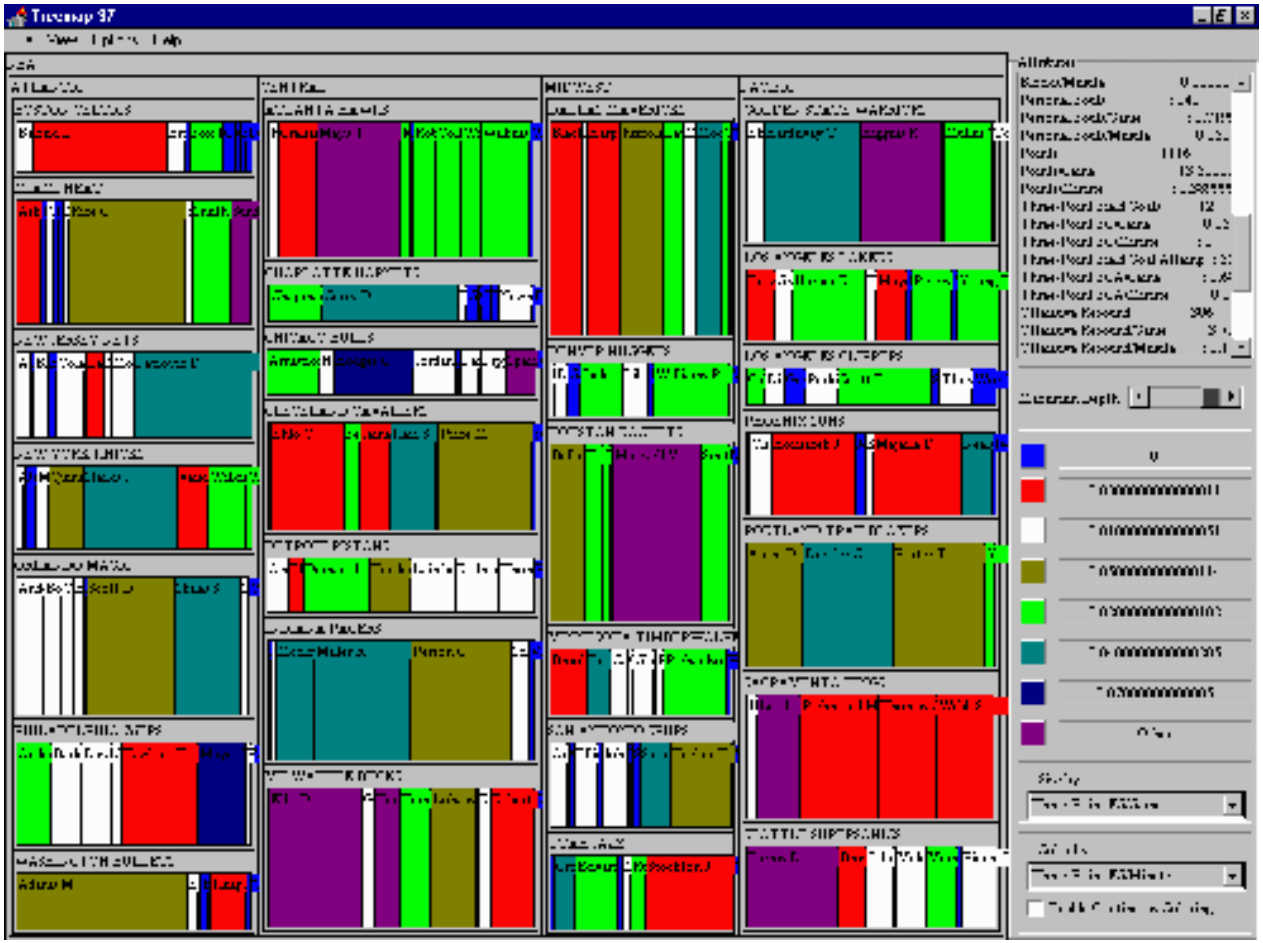
Color helps make distinctions

Aesthetics make the process appealing

Case Study:
The Journey of the TreeMap

- The TreeMap **(Johnson & Shneiderman '91)**
- Idea:
  - Show a hierarchy as a 2D layout
  - Fill up the space with rectangles representing objects
  - Size on screen indicates relative size of underlying objects.

Early Treemap Applied to File System



Treemap Problems

- Too disorderly
  - What does adjacency mean?
  - Aspect ratios uncontrolled leads to lots of skinny boxes that clutter
- Color not used appropriately
  - In fact, is meaningless here
- Wrong application
  - Don't need all this to just see the largest files in the OS

Successful Application of Treemaps

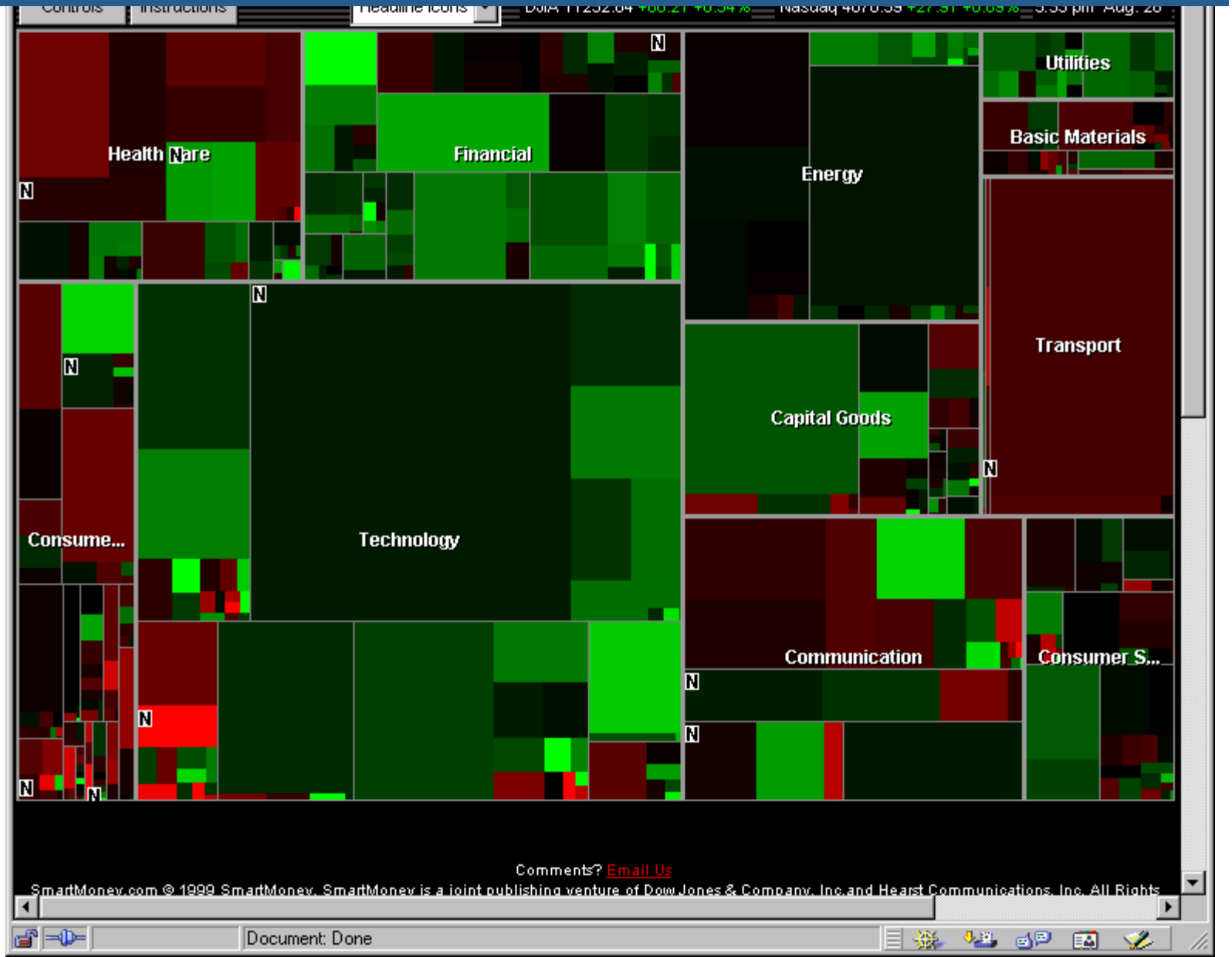
- Think more about the use
  - Break into meaningful groups
  - Fix these into a useful aspect ratio
- Use visual properties properly
  - Use color to distinguish meaningfully
    - Use only two colors:
      - Can then distinguish one thing from another
    - When exact numbers aren't very important
- Provide excellent interactivity
  - Access to the real data
  - Makes it into a useful tool

TreeMaps in Action

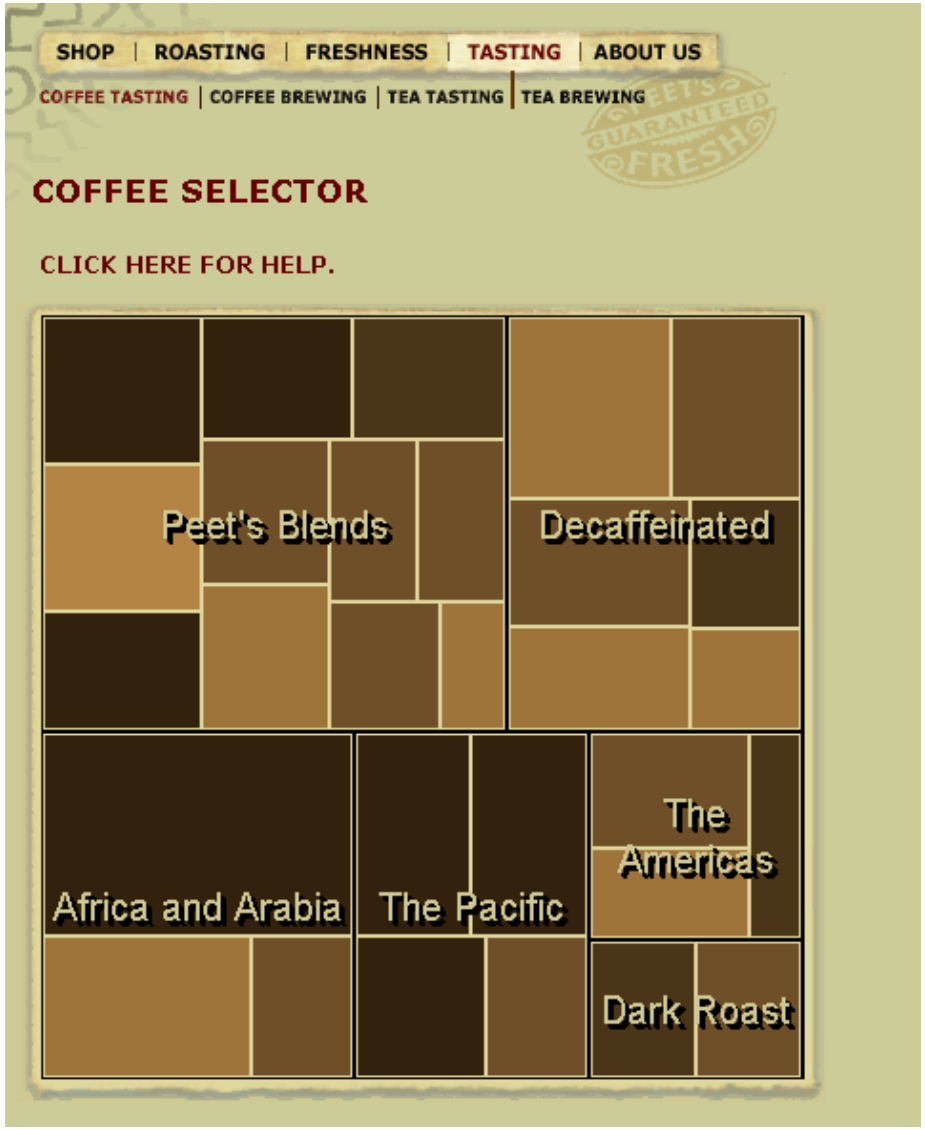
http://www.smartmoney.com/maps

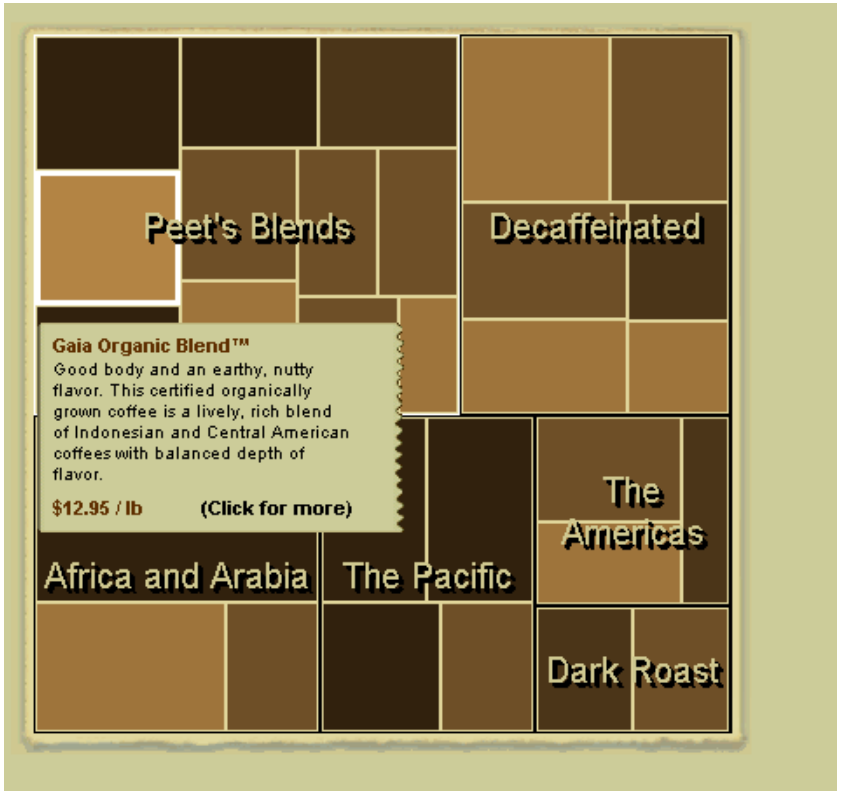
http://www.peets.com/tast/11/coffee_selector.asp

A Good Use of TreeMaps and Interactivity



Treemaps in Peets site





Analysis vs. Communication

- MarketMaps use of TreeMaps allows for sophisticated analysis
- Peets' use of TreeMaps is more for presentation and communication
- This is a key contrast

Open Issues

- Does visualization “sell” or help?
  - This is an active area of research
  - Supplemental for text collections
    - A correlation with spatial ability
    - Learning effects: with practice ability on visual display begins to equal that of text
    - Lots of use cases for conveying information (e.g. medicine labeling)

Key Questions to Ask about a Viz

1. What does it teach/show/elucidate?
2. Could it have been done more simply?
3. How is usability tested or evaluated?

Visualization

- What is Visualization?
- Visualization Principles
- Visualization Properties and Information Types
- Lying with Visuals
- Problem-solving with Visuals

Visual Principles

Visual Principles

- Types of Graphs
- Pre-attentive Properties
- Relative Expressiveness of Visual Cues
- Visual Illusions
- Tufte's notions
  - Graphical Excellence
  - Data-Ink Ratio Maximization
  - How to Lie with Visualization

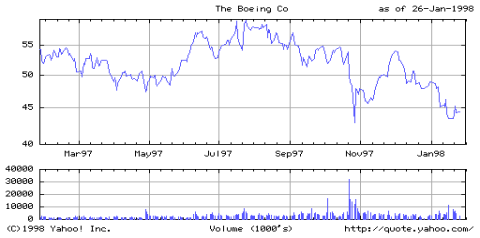
A Graph is: (Kosslyn)

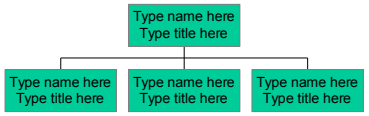
- A visual display that illustrates one or more relationships among entities
- A shorthand way to present information
- Allows a trend, pattern, or comparison to be easily apprehended

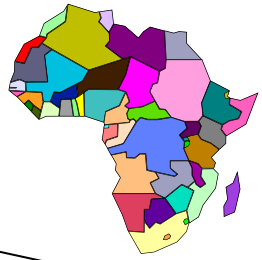
Types of Symbolic Displays

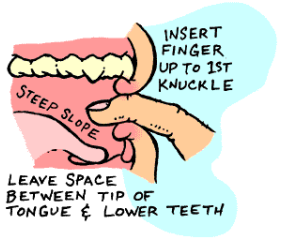
(Kosslyn 89)

- Graphs
- Charts
- Maps
- Diagrams



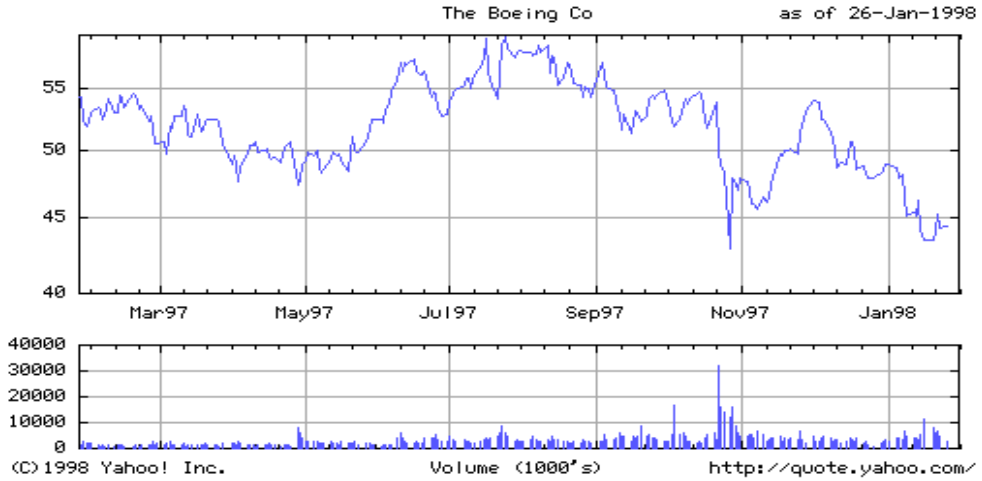






Types of Symbolic Displays

- Graphs
  - at least two scales required
  - values associated by a symmetric “paired with” relation
    - Examples: scatter-plot, bar-chart, correlation-plot



Types of Symbolic Displays

Charts

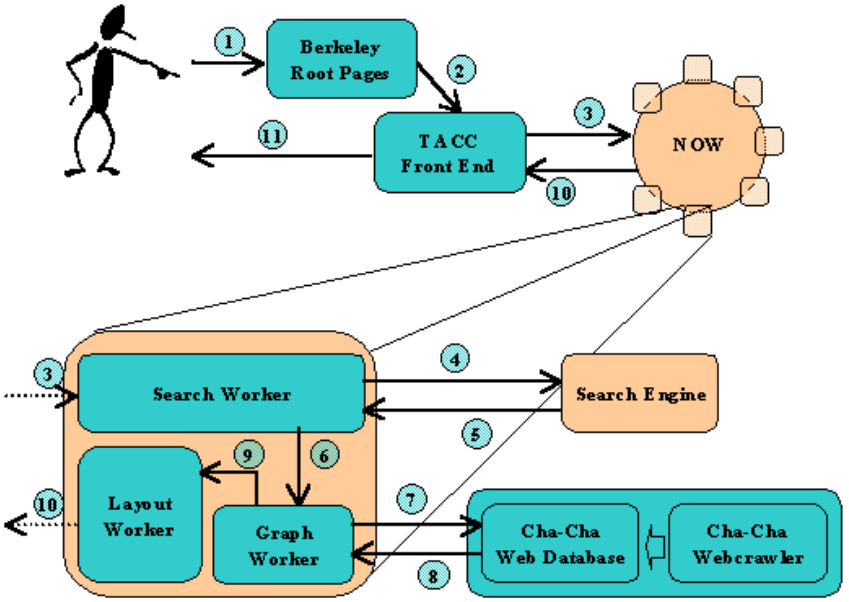
- discrete relations among discrete entities
- structure relates entities to one another
- lines and relative position serve as links

Examples:

family tree

flow chart

network diagram



Types of Symbolic Displays

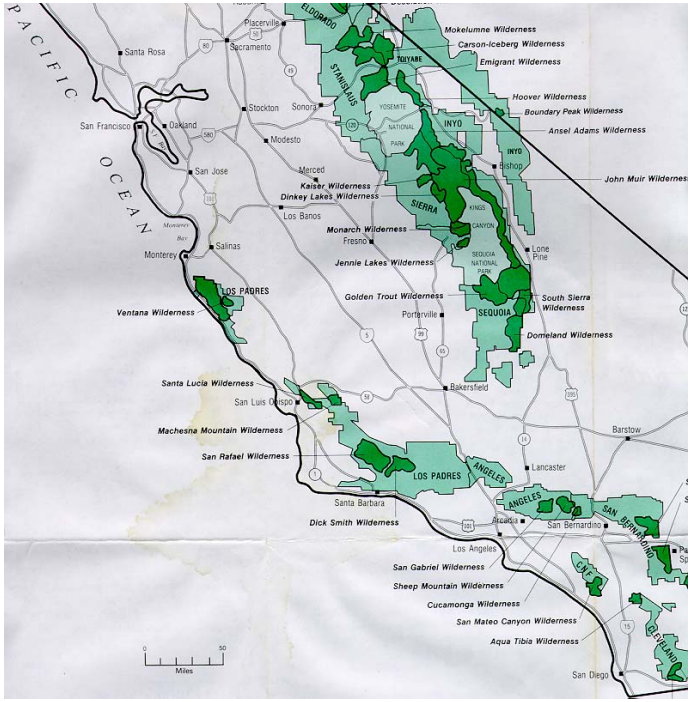
- Maps
  - internal relations determined (in part) by the spatial relations of what is pictured
  - labels paired with locations

Examples:

map of census data

topographic maps

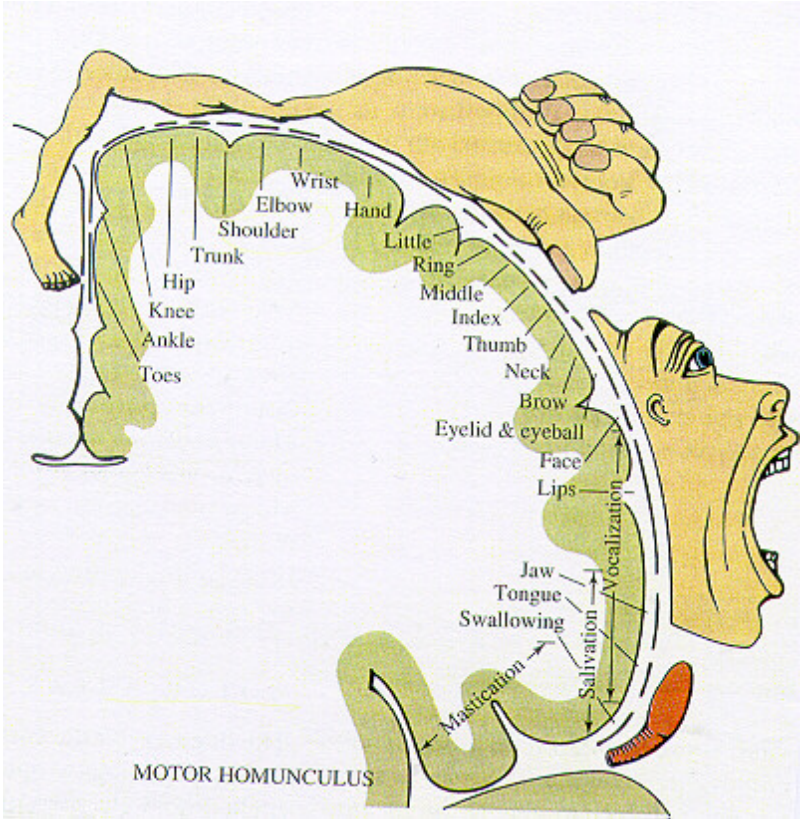
From www.thehighsierra.com



Types of Symbolic Displays

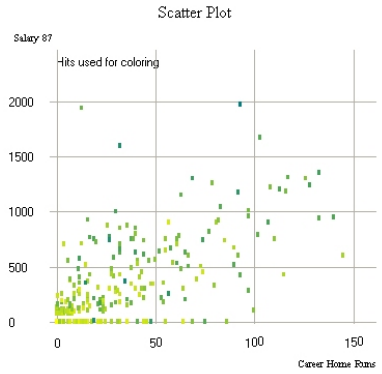
Diagrams

- schematic pictures of objects or entities
- parts are symbolic (unlike photographs)
  - how-to illustrations
  - figures in a manual



From Glietman, Henry. Psychology.
W.W. Norton and Company, Inc.
New York, 1995

Scatter Plots



- Qualitatively determine if variables
  - are highly correlated
    - linear mapping between horizontal & vertical axes
  - have low correlation
    - spherical, rectangular, or irregular distributions
  - have a nonlinear relationship
    - a curvature in the pattern of plotted points
- Place points of interest in context
  - color representing special entities

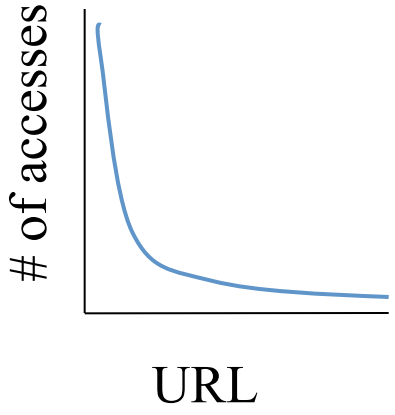
Anatomy of a Graph (Kosslyn 89)

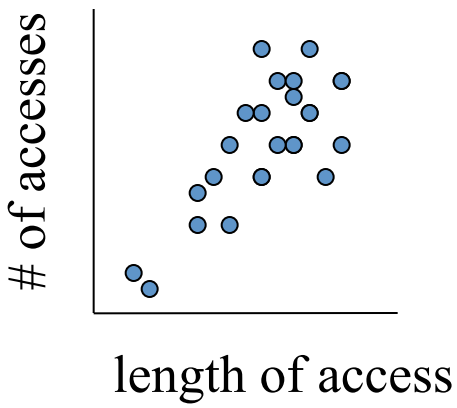
- Framework
  - sets the stage
  - kinds of measurements, scale, ...
- Content
  - marks
  - point symbols, lines, areas, bars, …
- Labels
  - title, axes, tic marks, ...

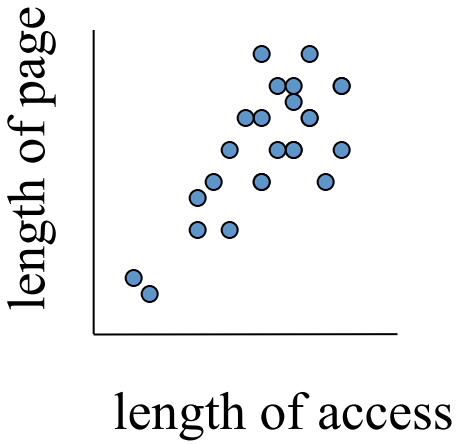
Basic Types of Data

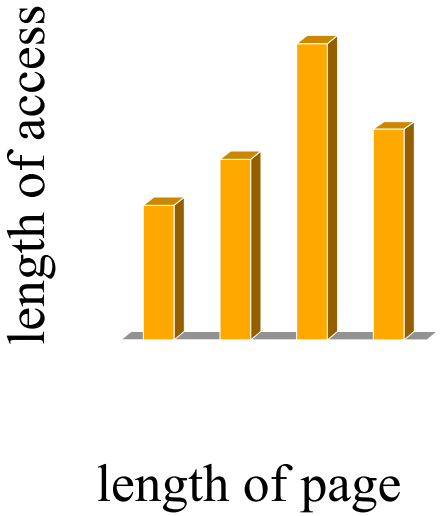
- Nominal (qualitative)
  - (no inherent order)
  - city names, types of diseases, ...
- Ordinal (qualitative)
  - (ordered, but not at measurable intervals)
  - first, second, third, …
  - cold, warm, hot
- Interval (quantitative)
  - list of integers or reals

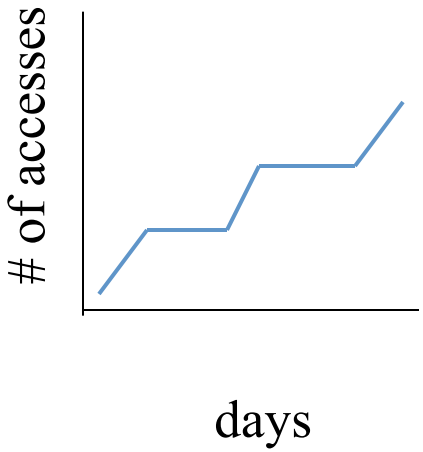
Common Graph Types

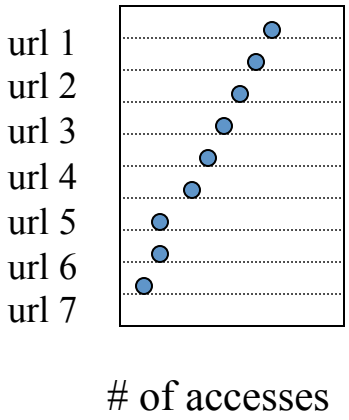












Combining Data Types in Graphs

Nominal	**Nominal**
Nominal	**Ordinal**
Nominal	**Interval**
Ordinal	**Ordinal**
Ordinal	**Interval**
Interval	**Interval**

When to use which type?

- Line graph
  - x-axis requires quantitative variable
  - Variables have contiguous values
  - familiar/conventional ordering among ordinals
- Bar graph
  - comparison of relative point values
- Scatter plot
  - convey overall impression of relationship between two variables
- Pie Chart?
  - Emphasizing differences in proportion among a few numbers

Practice in Selecting Tables and Graphs

- Should the message be presented in the form of a table or a graph?
- If a table, which kind of relationship?
- If a graph, which kind of relationship does it need to display?
  - Nominal comparison
  - Time-series
  - Part-to-whole
  - Ranking
  - Deviation
  - Frequency distribution
  - Correlation
- If a graph, which object or combination of objects for encoding the quantitative values would work best?
  - Points
  - Lines
  - Points and lines
  - bars

Summary of Value-Encoding Objects

	Points	Lines	Points & Lines	Bars
Nominal comparison	When there is a need to narrow the quantitative scale, and in doing so, remove zero from its base	Avoid	Avoid	Either horizontal or vertical bars
Time series	Avoid	Categorical subdivisions on X axis, quantitative values on Y axis, emphasis on overall pattern	Categorical subdivisions (X), quantitative values on Y, emphasis on overall pattern AND individual values	Categorical subdivisions on X axis, quantitative values on Y axis, emphasis on individual values
Ranking	When there is a need to narrow the quantitative scale, and in doing so, remove zero from its base	Avoid	Avoid	Horizontal bars are preferable, with values sorted in descending order
Part-to-whole	Avoid	Avoid	Avoid	Either horizontal or vertical bars
Distribution	Avoid	*Frequency polygon*: emphasis on overall pattern	Avoid	*Histogram*: emphasis primarily on individual values
Correlation	*Scatter Plot*	Avoid	Line is a trend line, not a line that connects the points	Either horizontal or vertical bars

Scenario

- Six months ago you developed and began teaching a new course entitled *Ethical Management*. When you initially proposed the idea for the class, your director was a little apprehensive about how well it would be received, but your past successes encouraged him to give you a shot. Now that you've been teaching it for a while and have worked the bugs out, it's time to give your director some evidence that he made the right decision.
- You've taught the course 4 times during the past month to a total of 100 students. Each student filled out an evaluation form at the end of the class, and you've tabulated the results. On a rating scale of 1 to 5 (1 represents *worthless*, and 5 *excellent*), the median rating for the course is 4, and the mean is 4.3. Not only is the average rating high, the range of ratings is tightly grouped aroudn the ratings of 3, 4, 5 with very few 2 and none 1. When you compare these ratings to those that you received for another class, their averages were about the same, but the spread of ratings for this other class were more broadly distributed, indicating that it doesn't work for all students as well as the new course.
- You want to give this information to your director in a form that he will grasp with little difficulty. Once before, when you tried to communicate differences in the range of ratings between classes using standard deviations, you could tell that the director didn't' really understand how to interpret them but was too embarassed to admit it. This time, what form with your presentation take?

- Table or graph?
- If a table, which kind?
- If a graph, what kind of relationship?
- If a graph, which graphical objects for quantitative encoding?
- Anything else?

Scenario

- You have been promoted from Director of Customer Service to Vice President of Services. Before you were a le to move full time into your new position, you had to recruit someone to replace you as director.
- Your company spreads the word of customer service across four different customer service centers, one in each of our major geographical regions. Customers are able to rate their experiences with the service centers by responding to surveys distributed via email. Because you want the new director to focus on improving the centers that are scoring lowest in customers' ratings, you need to provide her with the mean rating of service for each service center during the most recent quarter. In what form will you present these summarized ratings?

- Table or graph?
- If a table, which kind?
- If a graph, what kind of relationship?
- If a graph, which graphical objects for quantitative encoding?
- Anything else?

Today and Upcoming

- Visualization
  - What is Visualization?
  - Visualization Principles
  - Visualization Properties and Information Types (Scenarios)

  - Lying with Visuals/Negative Examples
  - Problem-solving with Visuals

- Data Science Industry Experts
- Your Presentations
- Ethics and Data

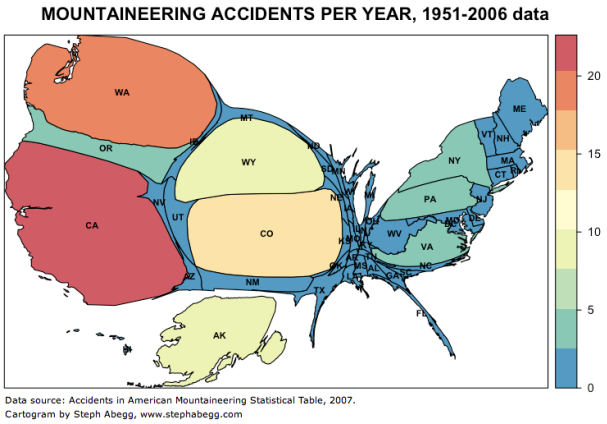
Experimentally Motivated Classification (Lohse et al. 94)

- Graphs
- Tables (numerical)
- Tables (graphical)
- Charts (time)
- Charts (network)
- Diagrams (structure)
- Diagrams (network)
- Maps
- Cartograms
- Icons
- Pictures

Interesting Findings

Lohse et al. 94

- Photorealistic images were least informative
  - Echos results in icon studies – better to use less complex, more schematic images
- Graphs and tables are the most self-similar categories
  - Results in the literature comparing these are inconclusive
- Cartograms were hard to understand
  - Echos other results – better to put points into a framed rectangle to aid spatial perception
- Temporal data more difficult to show than cyclic data
  - Recommend using animation for temporal data



Visual Properties

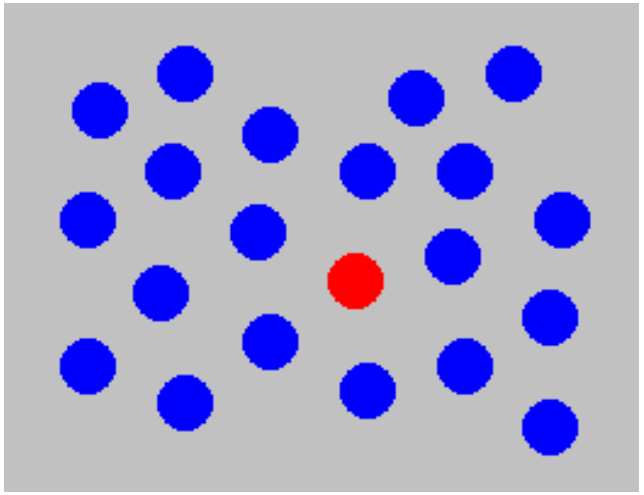
- Preattentive Processing
- Accuracy of Interpretation of Visual Properties
- Illusions and the Relation to Graphical Integrity

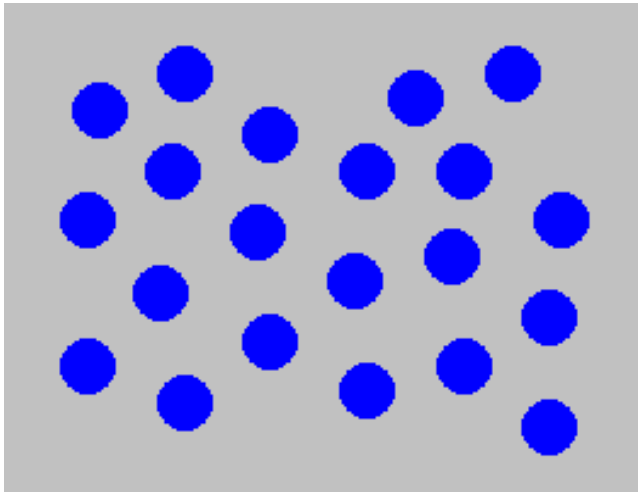
All Preattentive Processing figures from Healey 97
http://www.csc.ncsu.edu/faculty/healey/PP/PP.html

Preattentive Processing

- A limited set of visual properties are processed preattentively
  - (without need for focusing attention).
- This is important for design of visualizations
  - what can be perceived immediately
  - what properties are good discriminators
  - what can mislead viewers

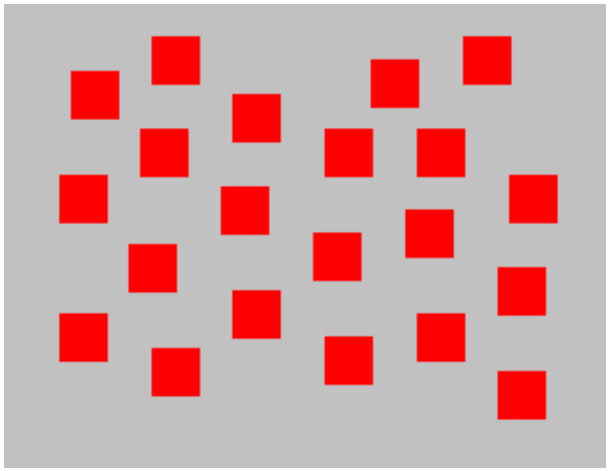
Example: Color Selection

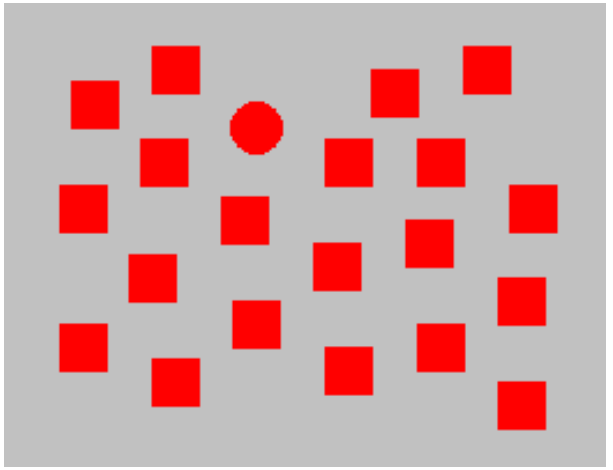




Viewer can rapidly and accurately determine whether the target (red circle) is present or absent. Difference detected in color.

Example: Shape Selection



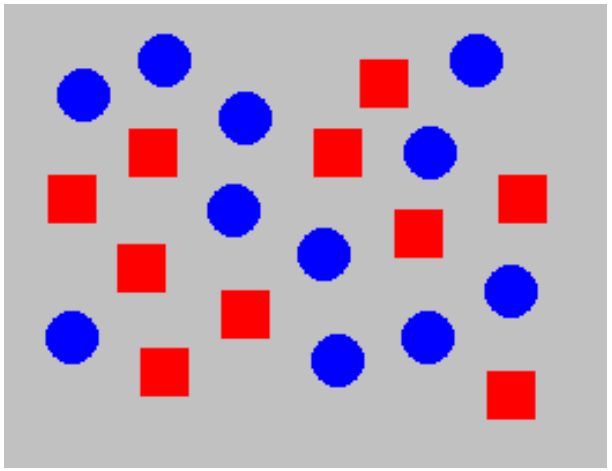


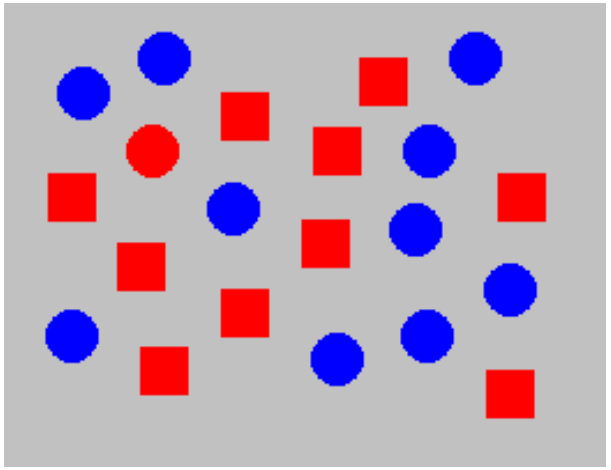
Viewer can rapidly and accurately determine whether the target (red circle) is present or absent. Difference detected in form (curvature)

Pre-attentive Processing

- < 200 - 250ms qualifies as pre-attentive
  - eye movements take at least 200ms
  - yet certain processing can be done very quickly, implying low-level processing in parallel
- If a decision takes a fixed amount of time regardless of the number of distractors, it is considered to be preattentive.

Example: Conjunction of Features





Viewer *cannot* rapidly and accurately determine whether the target (red circle) is present or absent when target has two or more features, each of which are present in the distractors. Viewer must search sequentially.

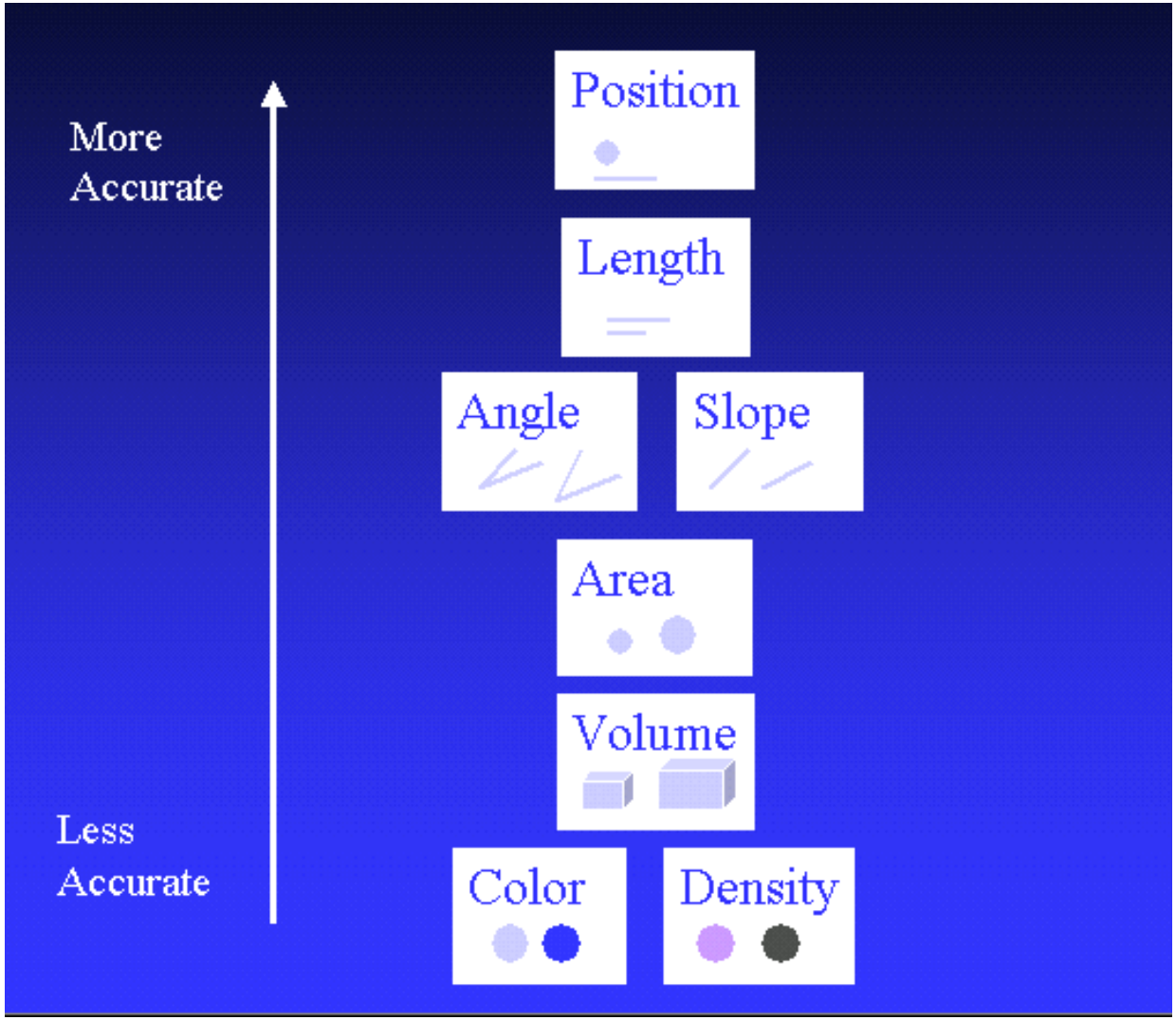
All Preattentive Processing figures from Healey 97
http://www.csc.ncsu.edu/faculty/healey/PP/PP.html

Visualization

- What is Visualization?
- Visualization Principles
- Visualization Properties and Information Types
- Lying with Visuals
- Problem-solving with Visuals

Which Properties are Appropriate for Which Information Types?

Accuracy Ranking of Quantitative Perceptual Tasks
Estimated; only pairwise comparisons have been validated
(Mackinlay 88 from Cleveland & McGill)



Interpretations of Visual Properties

Some properties can be discriminated more accurately but don't have intrinsic meaning

(Senay & Ingatious 97, Kosslyn, others)

- Density (Greyscale)
  - Darker -> More
- Size / Length / Area
  - Larger -> More
- Position
  - Leftmost -> first, Topmost -> first
- Hue
  - ??? no intrinsic meaning
- Slope
  - ??? no intrinsic meaning

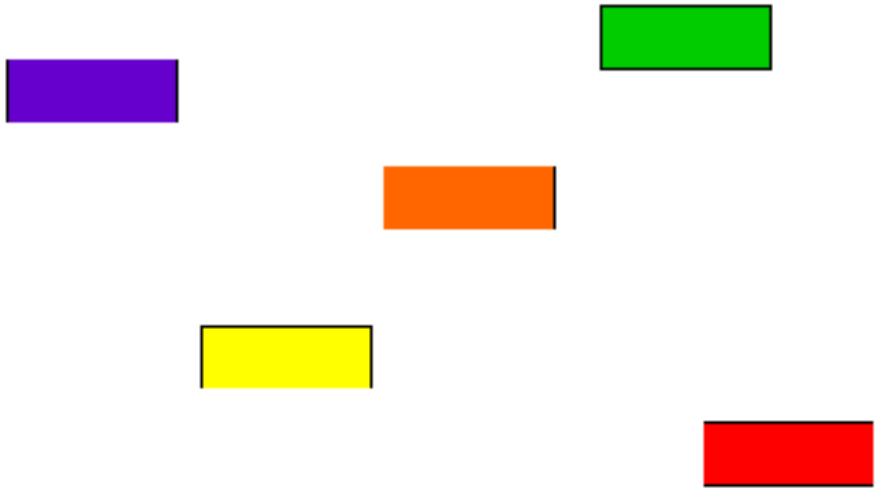
Ranking of Applicability of Properties for Different Data Types

(Mackinlay 88, Not Empirically Verified)

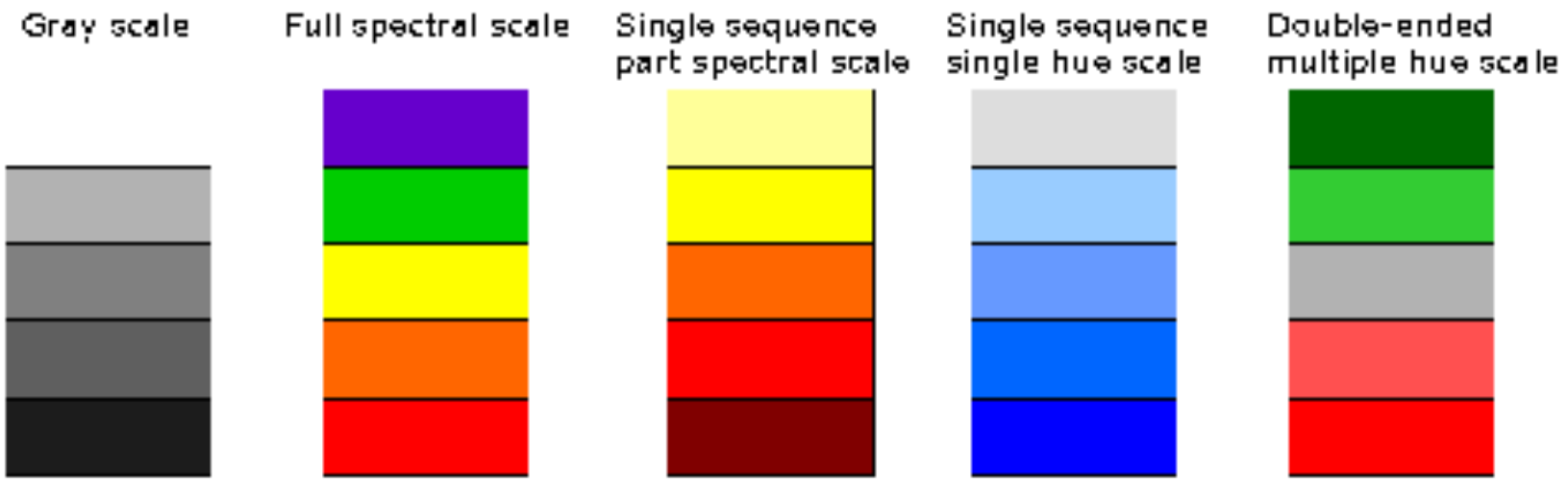
QUANTITATIVE	ORDINAL	NOMINAL
Position	Position	Position
Length	Density	Color Hue
Angle	Color Saturation	Texture
Slope	Color Hue	Connection
Area	Texture	Containment
Volume	Connection	Density
Density	Containment	Color Saturation
Color Saturation	Length	Shape
Color Hue	Angle	Length

Color Schemes

Order these (low->hi)



Color Schemes



Color Purposes

- Call attention to specific items
- Distinguish between classes of items
  - Increases the number of dimensions for encoding
- Increase the appeal of the visualization

Using Color

- Proceed with caution
  - Less is more
  - Representing magnitude is tricky
- Examples
  - Red-orange-yellow-white
    - Works for costs
    - Maybe because people are very experienced at reasoning shrewdly according to cost
  - Green-light green-light brown-dark brown-grey-white works for atlases
  - Grayscale is unambiguous but has limited range

What are good guidelines for Infoviz?

- Use graphics appropriately
  - Don't use images gratuitously
  - Don't lie with graphics!
    - Link to original data
  - Don't conflate area with other information
    - E.g., use area in map to imply amount
- Make it interactive (feedback)
  - Brushing and linking
  - Multiple views
  - Overview + details
- Match mental models

Tufte

- Principles of Graphical Excellence
  - Graphical excellence is
    - the well-designed presentation of interesting data – a matter of substance, of statistics, and of design
    - consists of complex ideas communicated with clarity, precision and efficiency
    - is that which gives to the viewer the greates number of ideas in the shortest time with the least ink in the smallest space
    - requires telling the truth about the data.

Tufte's Notion of Data Ink Maximization

- What is the main idea?
  - draw viewers attention to the substance of the graphic
  - the role of redundancy
  - principles of editing and redesign
- What's wrong with this? What is he really getting at?

Tufte Principle

Maximize the data-ink ratio:

$$\text{Data-ink ratio} = \frac{\text{data ink}}{\text{total ink used in graphic}}$$

Avoid "chart junk"

Tufte Principles

- Use multifunctioning graphical elements
- Use small multiples
- Show mechanism, process, dynamics, and causality
- High data density
  - Number of items/area of graphic
  - This is controversial
    - White space thought to contribute to good visual design
    - Tufte's book itself has lots of white space

Tufte's Graphical Integrity

- Some lapses intentional, some not
- Lie Factor = $\frac{\text{size of effect in graph}}{\text{size of effect in data}}$
- Misleading uses of area
- Misleading uses of perspective
- Leaving out important context
- Lack of taste and aesthetics

Visualization

- What is Visualization?
- Visualization Principles
- Visualization Properties and Information Types
- Lying with Visuals
- Problem-solving with Visuals

From Tim Craven's LIS 504 course
http://instruct.uwo.ca/fim-lis/504/504gra.htm#data-ink_ratio

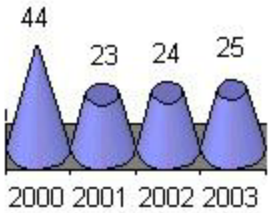
LIS 504 - Graphic displays of data - Netscape

File Edit View Go Communicator Help

A common example of a high lie factor occurs when both dimensions of a two-dimensional figure are made proportional to the same data, so that the size of the figure is proportional to the square of the data; for instance,

Year	Books circulated
2001	100 
2002	141 
2003	200 

An example of a **low** lie factor can be seen in the "Cones" custom chart format in Microsoft Excel.

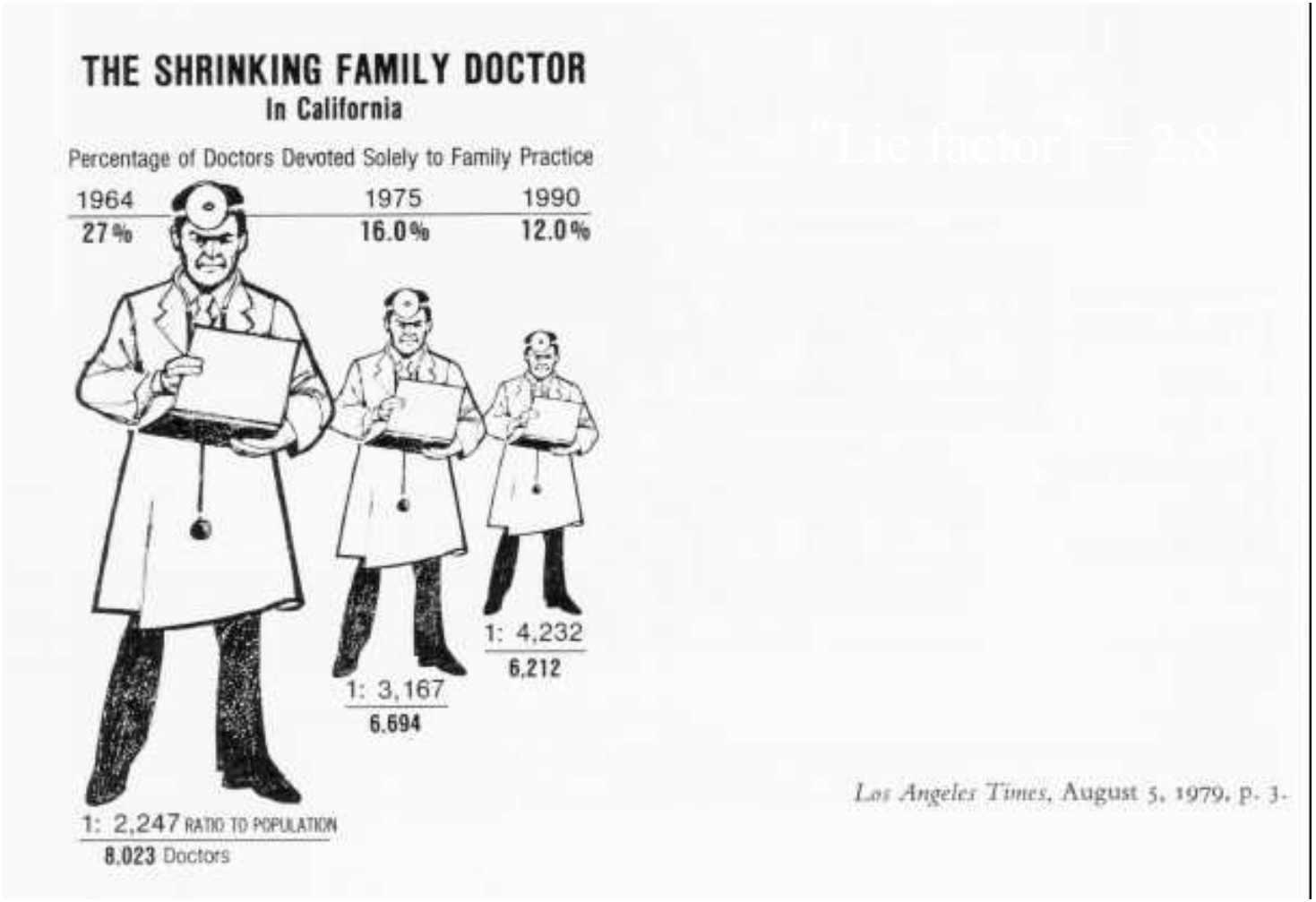


The heights of the (truncated) cones are proportional to the data, but their areas on the screen and their apparent volumes make the larger data values seem relatively small.

Document: Done

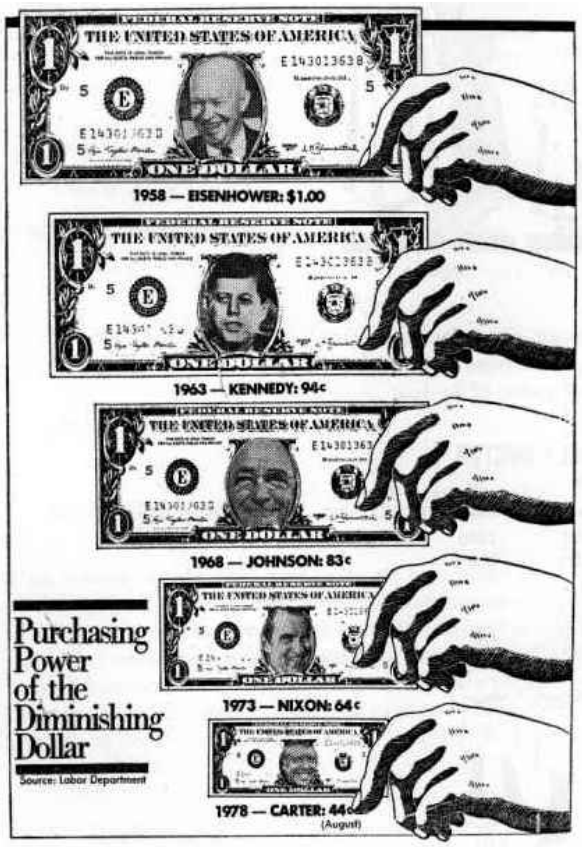
How to Exaggerate with Graphs

from Tufte ’83



How to Exaggerate with Graphs

from Tufte ’83



Washington Post, October 25, 1978, p. 1.

Visualization

- What is Visualization?
- Visualization Principles
- Visualization Properties and Information Types
- Lying with Visuals
- Problem-solving with Visuals

Multidimensional Detective

A. Inselberg, Multidimensional Detective, Proceedings of IEEE Symposium on Information Visualization (InfoVis '97), 1997.

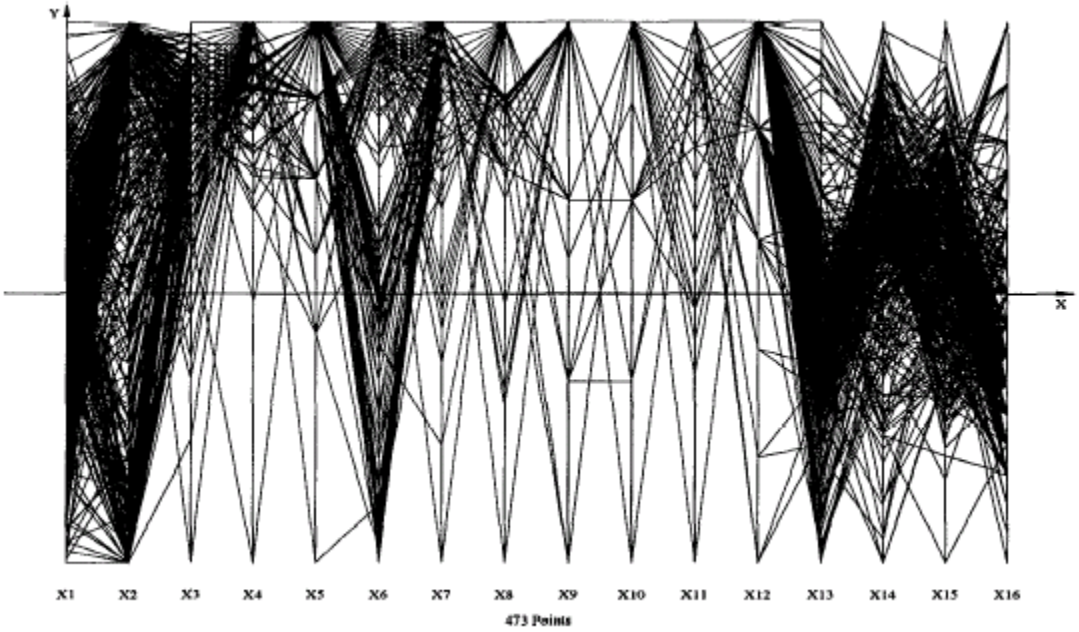


Figure 1: The full dataset consisting of 473 batches

A Detective Story

A. Inselberg, Multidimensional Detective, Proceedings of IEEE Symposium on Information Visualization (InfoVis '97), 1997

Inselberg's Principles for analysis using visualizations:

1. Do not let the picture scare you
2. Understand your objectives
   - Use them to obtain visual cues
3. Carefully scrutinize the picture
4. Test your assumptions, especially the "I am really sure of's"
5. You can't be unlucky all the time!

A Detective Story

A. Inselberg, Multidimensional Detective, Proceedings of IEEE Symposium on Information Visualization (InfoVis '97), 1997

- The Dataset:
  - Production data for 473 batches of a VLSI chip
  - 16 process parameters
  - The yield: % of produced chips that are useful
    - X1
  - The quality of the produced chips (speed)
    - X2
  - 10 types of defects (zero defects shown at top)
    - X3 … X12
  - 4 physical parameters
    - X13 … X16
- The Objective:
  - Raise the yield (X1) and maintain high quality (X2)

Multidimensional Detective

A. Inselberg, Multidimensional Detective, Proceedings of IEEE Symposium on Information Visualization (InfoVis '97), 1997.

Do Not Let the Picture Scare You!!

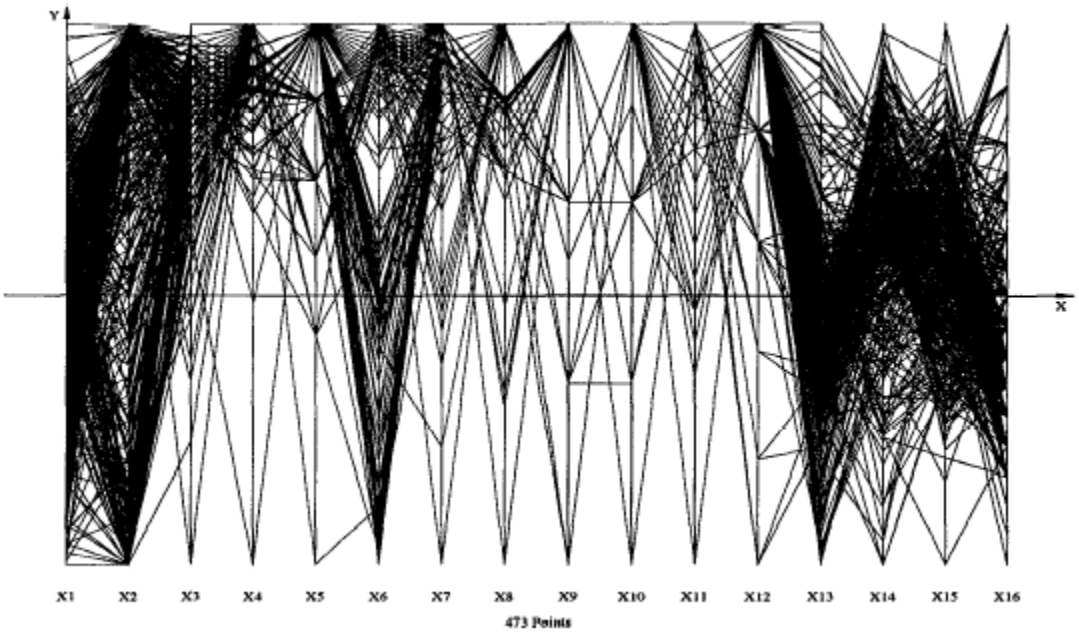


Figure 1: The full dataset consisting of 473 batches

Multidimensional Detective

- Each line represents the values for one batch of chips
- This figure shows what happens when only those batches with both high X1 and high X2 are chosen
- Notice the separation in values at X15
- Also, some batches with few X3 defects are not in this high-yield/high-quality group.

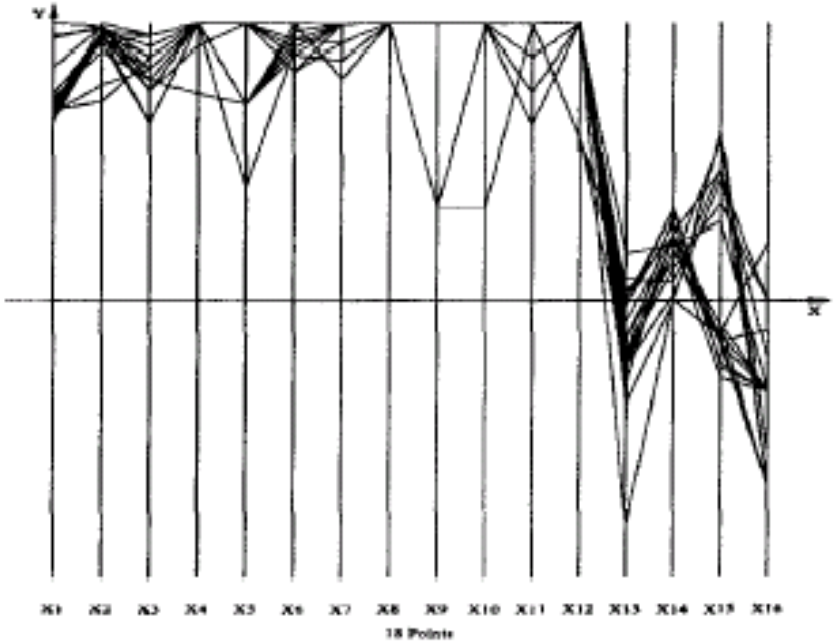


Figure 2: The batches high in Yield, $X1$, and Quality, $X2$.

Multidimensional Detective

- Now look for batches which have *nearly* zero defects.
  - For 9 out of 10 defect categories
- Most of these have low yields
- Surprising because we know from first diagram that some defects are ok.
- Go back to first diagram, looking at defect categories
- Notice that X6 behaves differently than the rest
- Allow two defects, where one defect in X6
- This results in the very best batch appearing

Multidimensional Detective

- Fig 5 and 6 show that high yield batches don't have non-zero values for defects of type X3 and X6
  - Don't believe your assumptions …
- Looking now at X15 we see the separation is important
  - Lower values of this property end up in the better yield batches

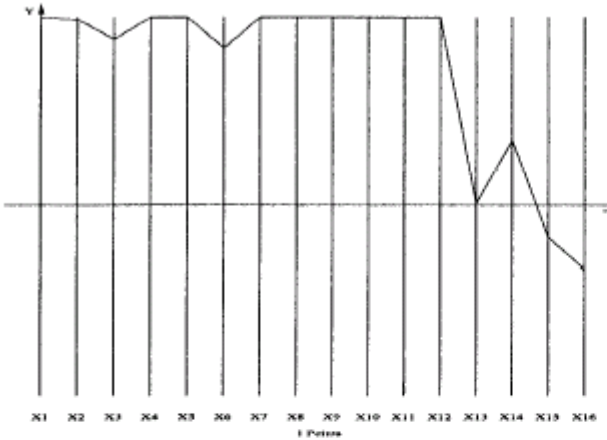


Figure 5: The best batch. Highest in Yield, $X1$, and very high in Quality, $X2$.

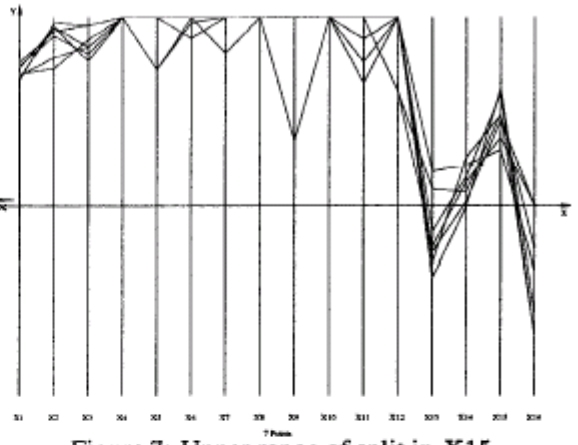
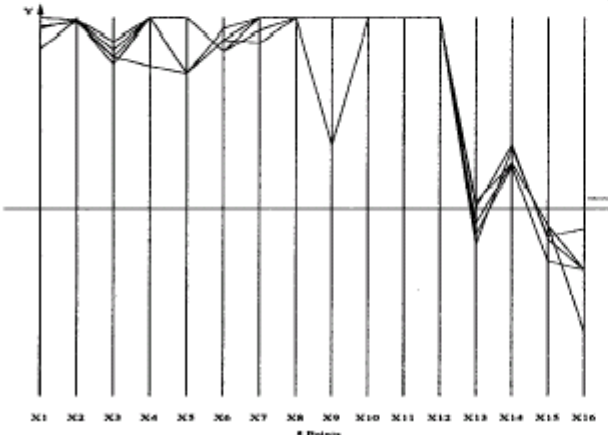
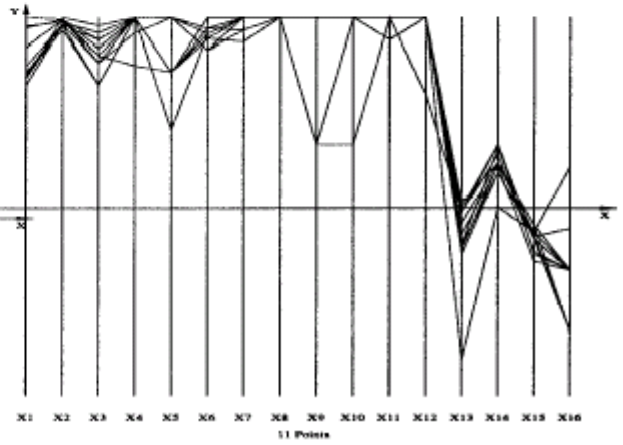


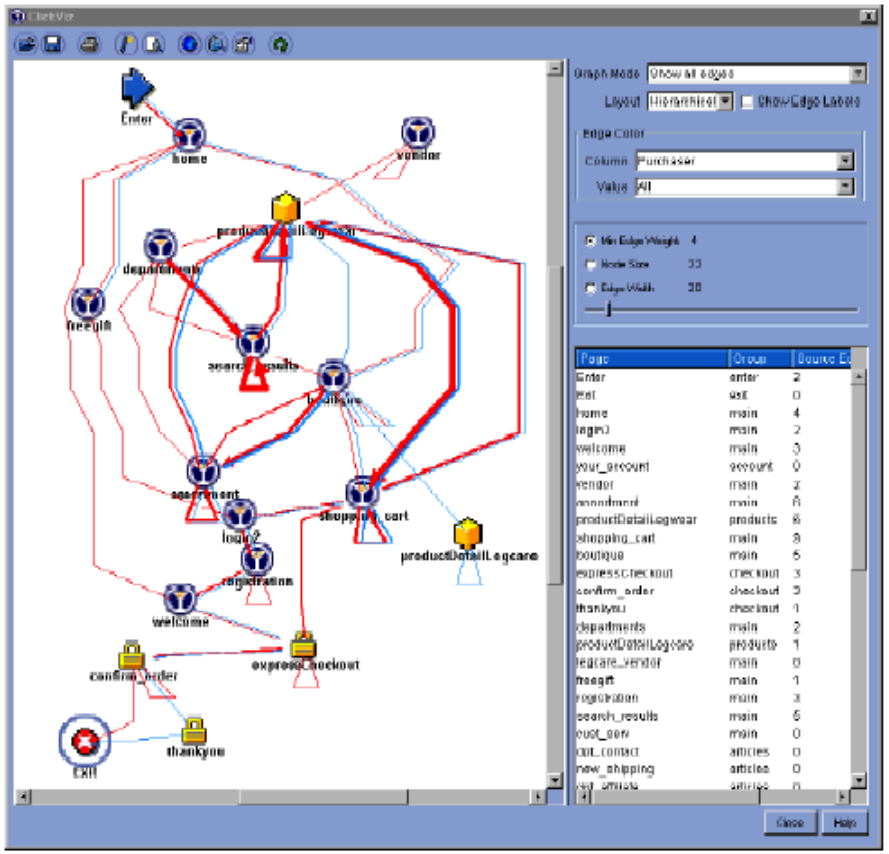
Figure 7: Upper range of split in $X15$





Case Study: E-Commerce Clickstream Visualization

- Brainerd & Becker, IEEE Infovis 2001
- Aggregate nodes using an icon (e.g. all the checkout pages)
- Edges represent transitions
  - Wider means more transitions



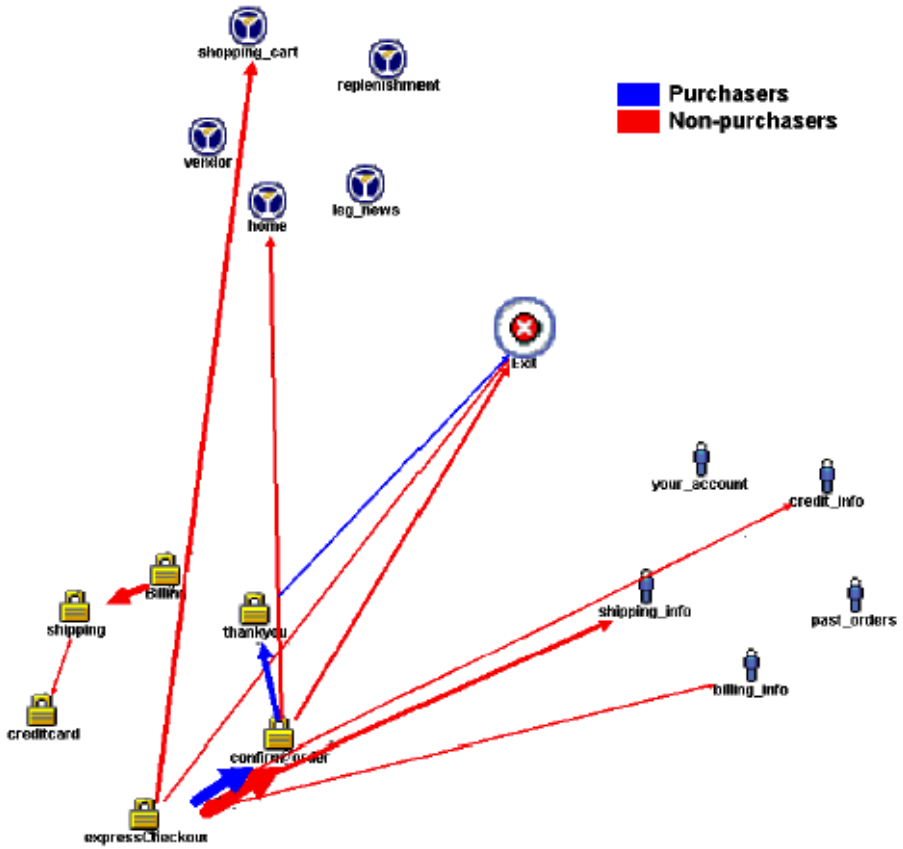
Slide by Wayne Kao

Customer Segments

- Collect
  - Clickstream
  - Purchase history
  - Demographic data
- Associates customer data with their clickstream
- Different color for each customer segment

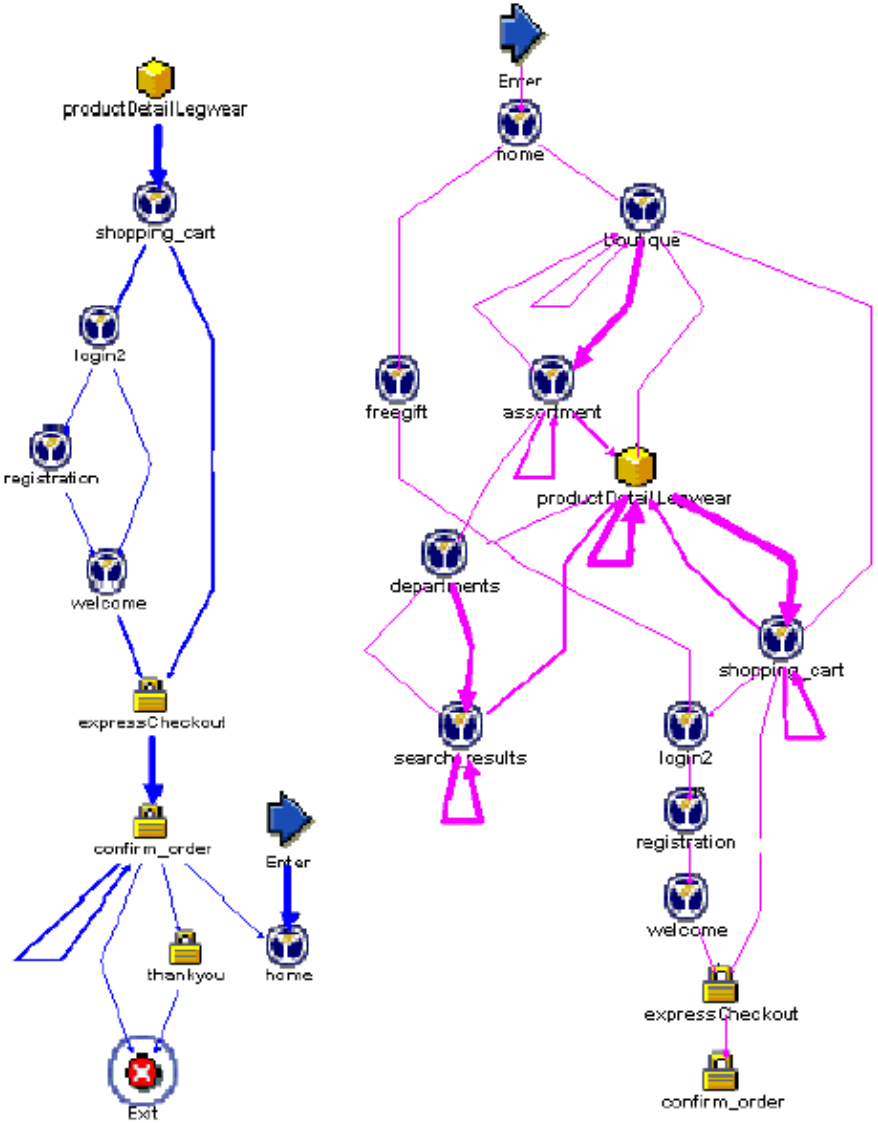
Layout

- Aggregation based on file system path



Initial Findings

- Gender shopping differences



(a) Males (b) Females

Visualization for Analysis

- Carlis & Konstan, UIST 1998

- Problem: data that is both periodic and serial
  - Time students spend on different activities
  - Tree growth patterns
    - Time: which year
    - Period: yearly
  - Multi-day races such as the Tour de France
  - Calendars arbitrarily wrap around at end of month
  - Octaves in music
- How to find patterns along both dimensions?

Analyzing Complex Periodic Data

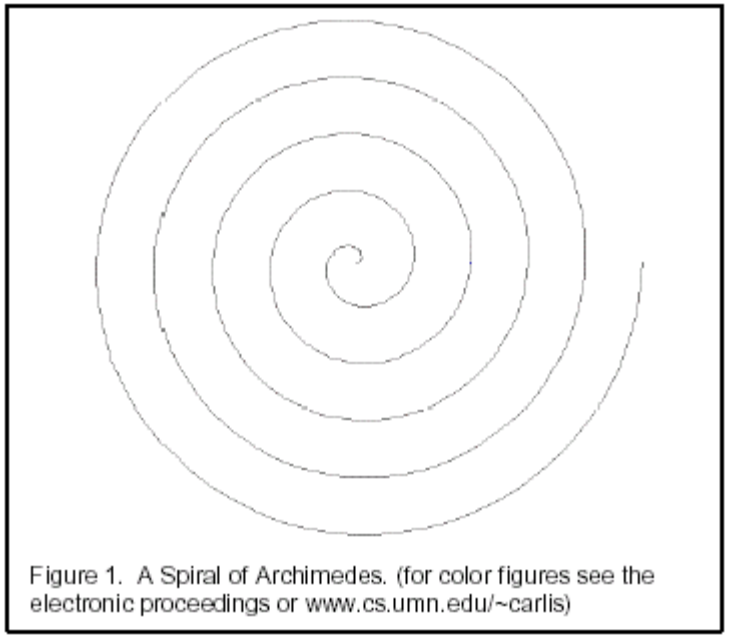


Figure 1. A Spiral of Archimedes. (for color figures see the electronic proceedings or www.cs.umn.edu/~carlis)

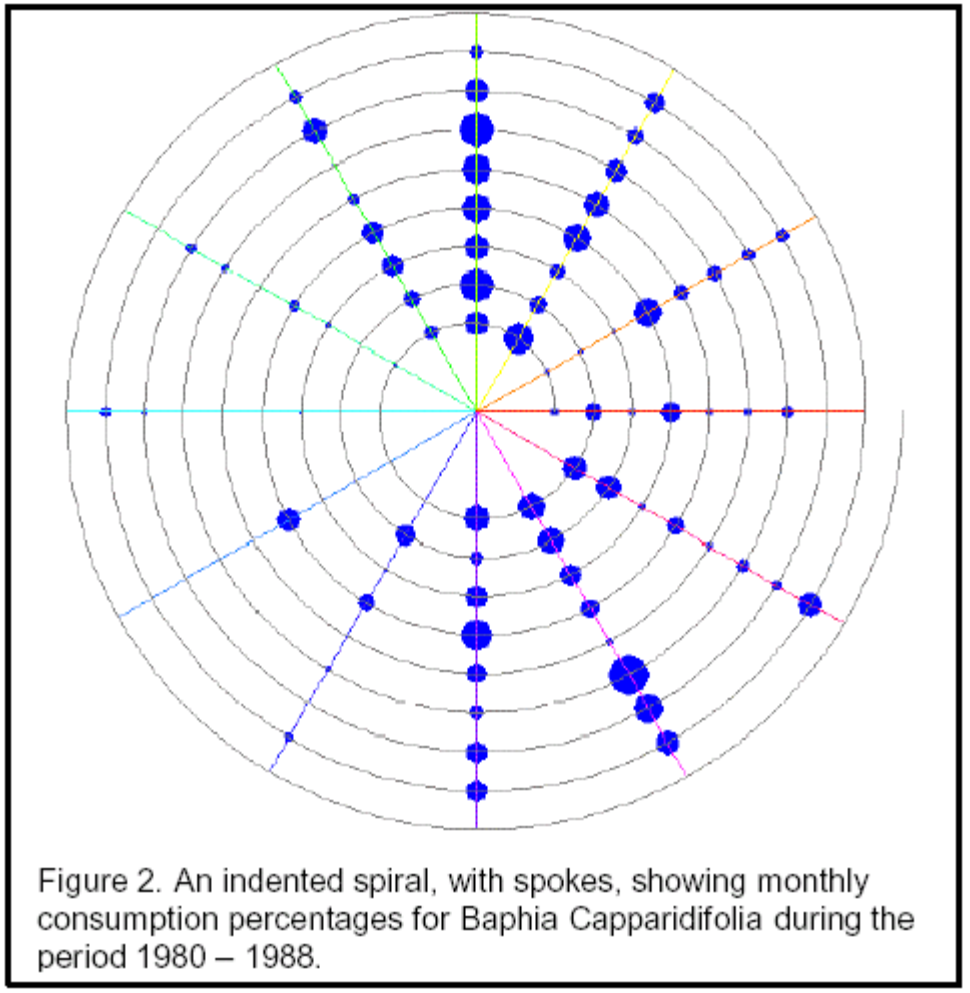


Figure 2. An indented spiral, with spokes, showing monthly consumption percentages for Baphia Capparidifolia during the period 1980 – 1988.

Carlis & Konstan, UIST 1998.

Analyzing Complex Periodic Data

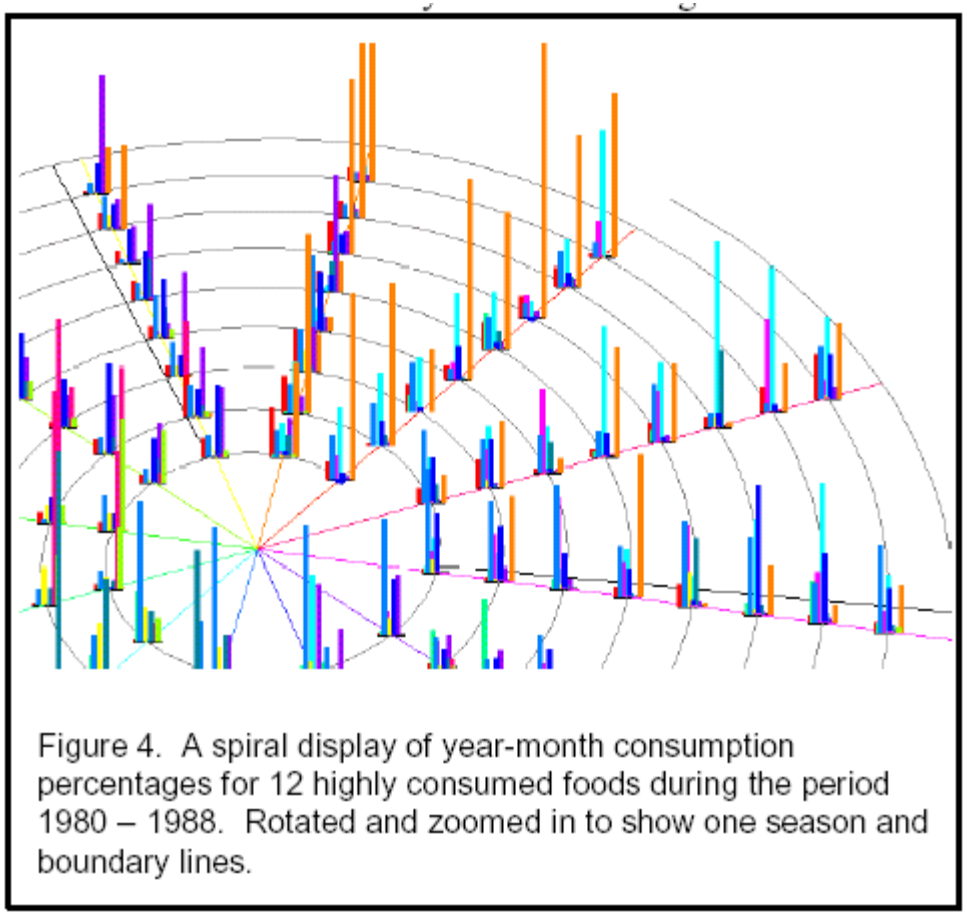


Figure 4. A spiral display of year-month consumption percentages for 12 highly consumed foods during the period 1980 – 1988. Rotated and zoomed in to show one season and boundary lines.

- Consumption values for each month appear as spikes
- Each food has its own color
- Boundary line (in black) shows when season begins/ ends

Carlis & Konstan, UIST 1998.

Key Questions to Ask about a Viz

1. What does it teach/show/elucidate?
2. Could it have been done more simply?
3. How is usability tested or evaluated?

Today and Upcoming

- Visualization
  - What is Visualization?
  - Visualization Principles
  - Visualization Properties and Information Types (Scenarios)

  - Lying with Visuals/Negative Examples
  - Problem-solving with Visuals

- Data Science Industry Experts
- Your Presentations
- Ethics and Data