Foundation of Data Science
Lecture 10, Module 1
Fall 2022

Rumi Chunara, PhD

Introduction

- Why visualization?
- Goals of Information Visualization
- Case Study: The Journey of the TreeMap
- Key Questions
- Lying with Visuals
- Problem Solving with Visuals

Why Visualization?

Anscombe's quartet

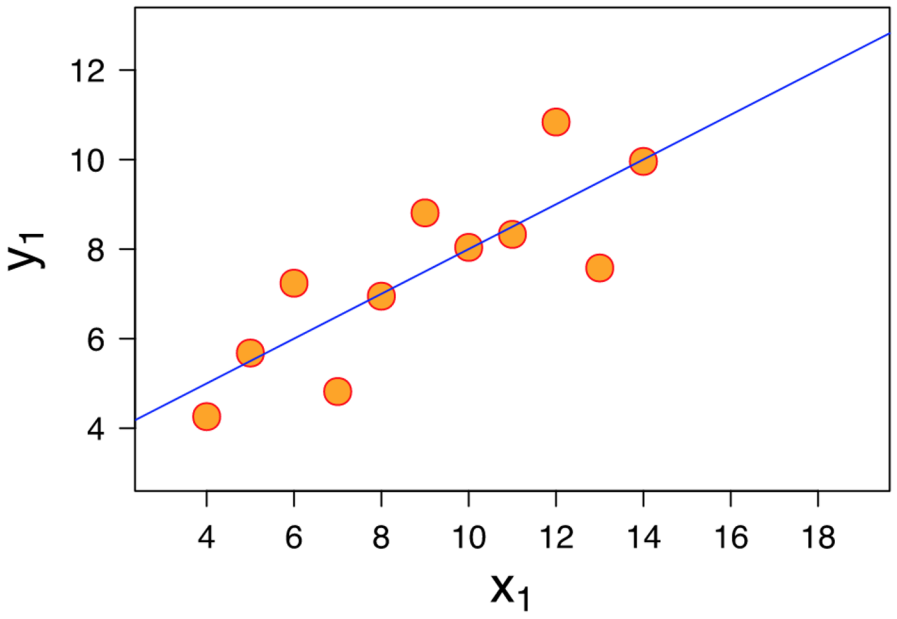
I		II		III		IV	
x	y	x	y	x	y	x	y
10.0	8.04	10.0	9.14	10.0	7.46	8.0	6.58
8.0	6.95	8.0	8.14	8.0	6.77	8.0	5.76
13.0	7.58	13.0	8.74	13.0	12.74	8.0	7.71
9.0	8.81	9.0	8.77	9.0	7.11	8.0	8.84
11.0	8.33	11.0	9.26	11.0	7.81	8.0	8.47
14.0	9.96	14.0	8.10	14.0	8.84	8.0	7.04
6.0	7.24	6.0	6.13	6.0	6.08	8.0	5.25
4.0	4.26	4.0	3.10	4.0	5.39	19.0	12.50
12.0	10.84	12.0	9.13	12.0	8.15	8.0	5.56
7.0	4.82	7.0	7.26	7.0	6.42	8.0	7.91
5.0	5.68	5.0	4.74	5.0	5.73	8.0	6.89

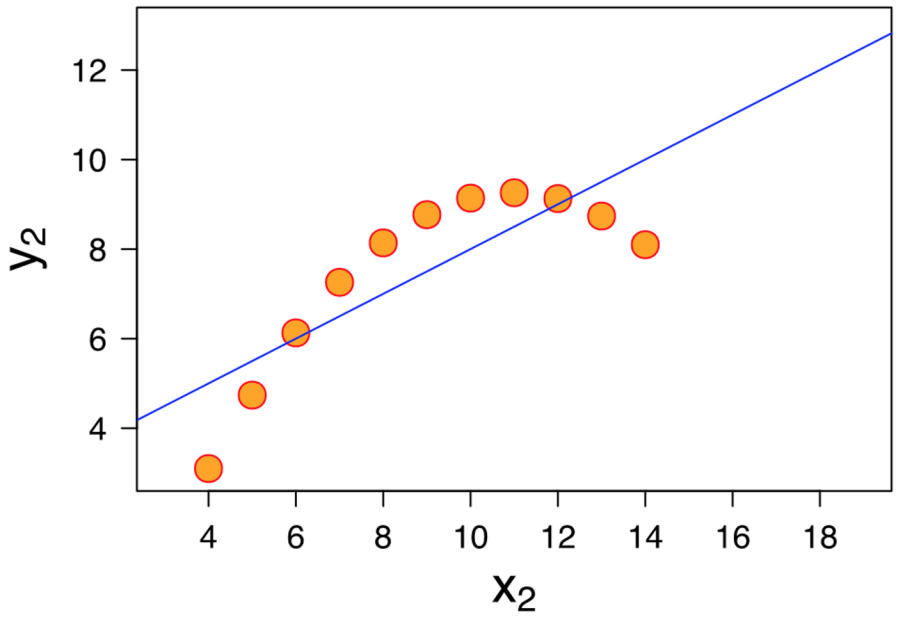
Anscombe's quartet

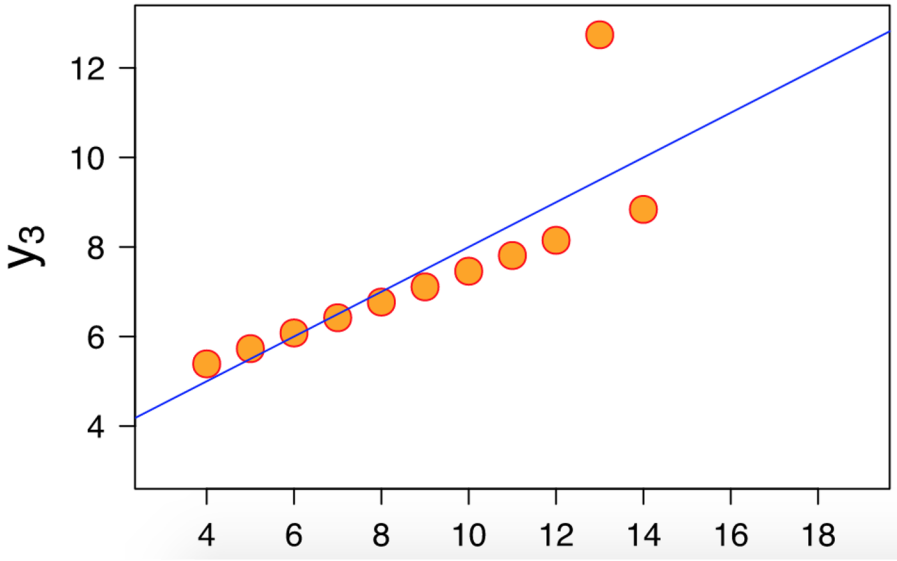
I		II		III		IV	
x	y	x	y	x	y	x	y
10.0	8.04	10.0	9.14	10.0	7.46	8.0	6.58
8.0	6.95	8.0	8.14	8.0	6.77	8.0	5.76
13.0	7.58	13.0	8.74	13.0	12.74	8.0	7.71
9.0	8.81	9.0	8.77	9.0	7.11	8.0	8.84
11.0	8.33	11.0	9.26	11.0	7.81	8.0	8.47
14.0	9.96	14.0	8.10	14.0	8.84	8.0	7.04
6.0	7.24	6.0	6.13	6.0	6.08	8.0	5.25

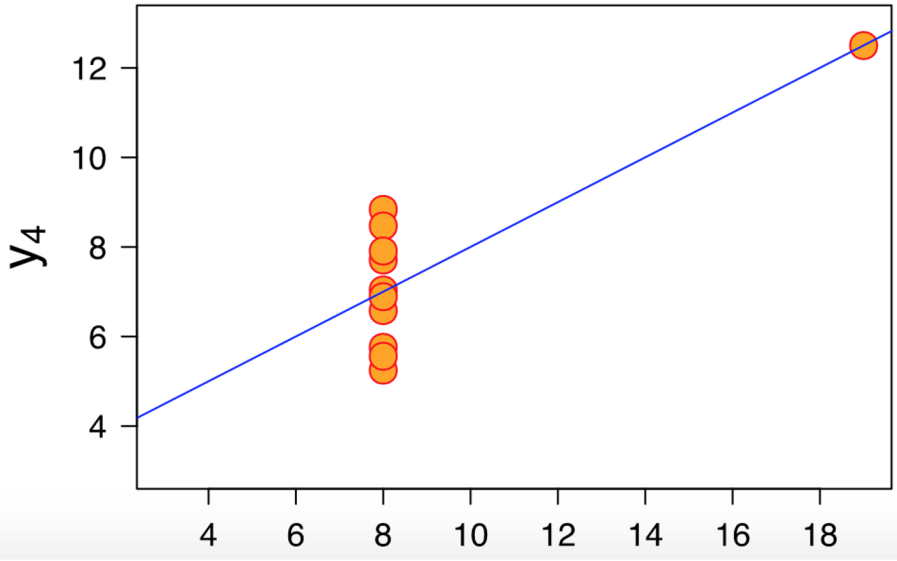
All of them have the same values for the statistics below!

Property	Value	Accuracy
Mean of *x*	9	exact
Sample variance of *x*	11	exact
Mean of *y*	7.50	to 2 decimal places
Sample variance of *y*	4.125	plus/minus 0.003
Correlation between *x* and *y*	0.816	to 3 decimal places
Linear regression line	$y = 3.00 + 0.500x$	to 2 and 3 decimal places, respectively
Coefficient of determination of the linear regression	0.67	to 2 decimal places


y1
x1
4
6
8
10
12
14
16
18
4
6
8
10
12


y2
x2
4
6
8
10
12
14
16
18
4
6
8
10
12


y3
4
6
8
10
12
14
16
18
4
6
8
10
12


y4
4
6
8
10
12
14
16
18
4
6
8
10
12

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

Two Different Types of Viz

- **Explore/Calculate** (Goal 1)
  - Analyze
  - Reason about Information

- **Communicate** (Goal 2)
  - Explain
  - Make Decisions
  - Reason about Information *(again!)*

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.

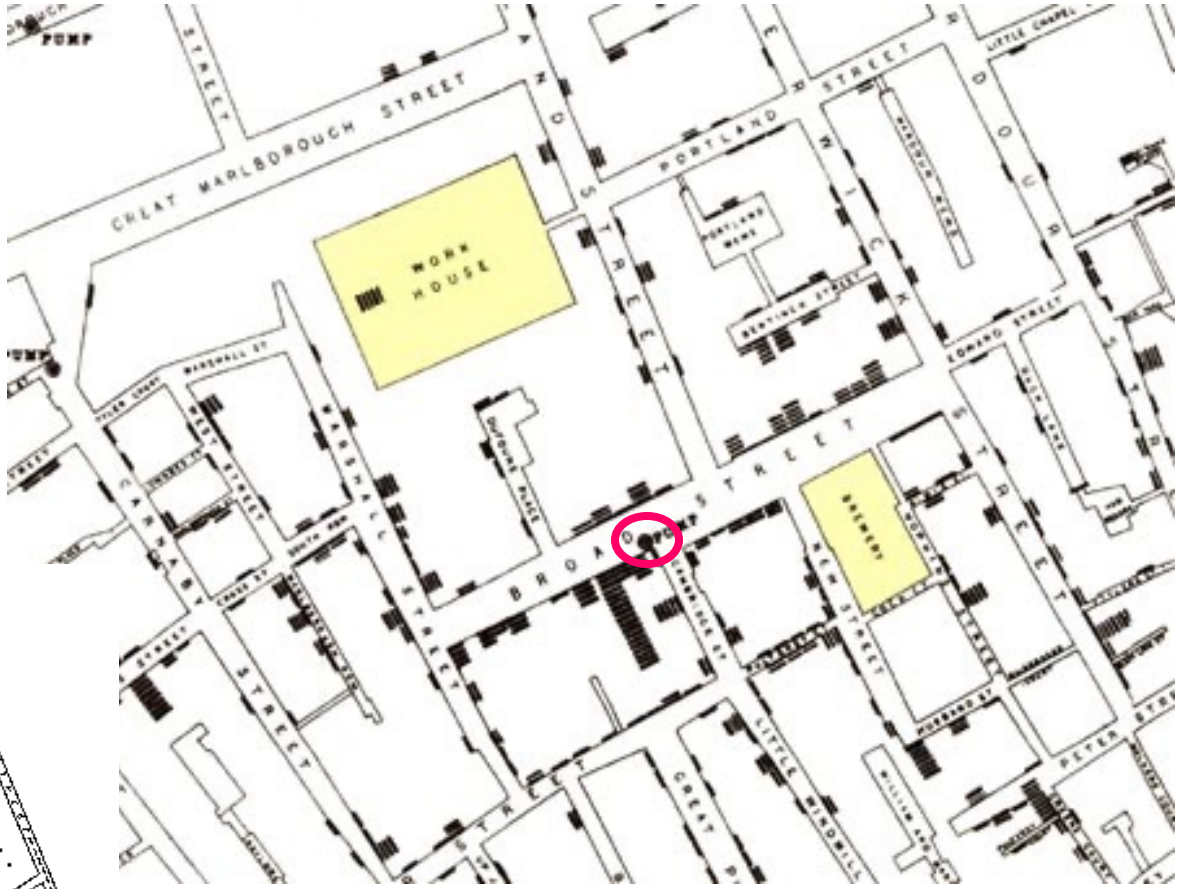
Horizontal lines indicate location of deaths.

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



From Visual Explanations by Edward Tufte, Graphics Press, 1997

Why Visualization?

- Use the eye for pattern recognition. Humans 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

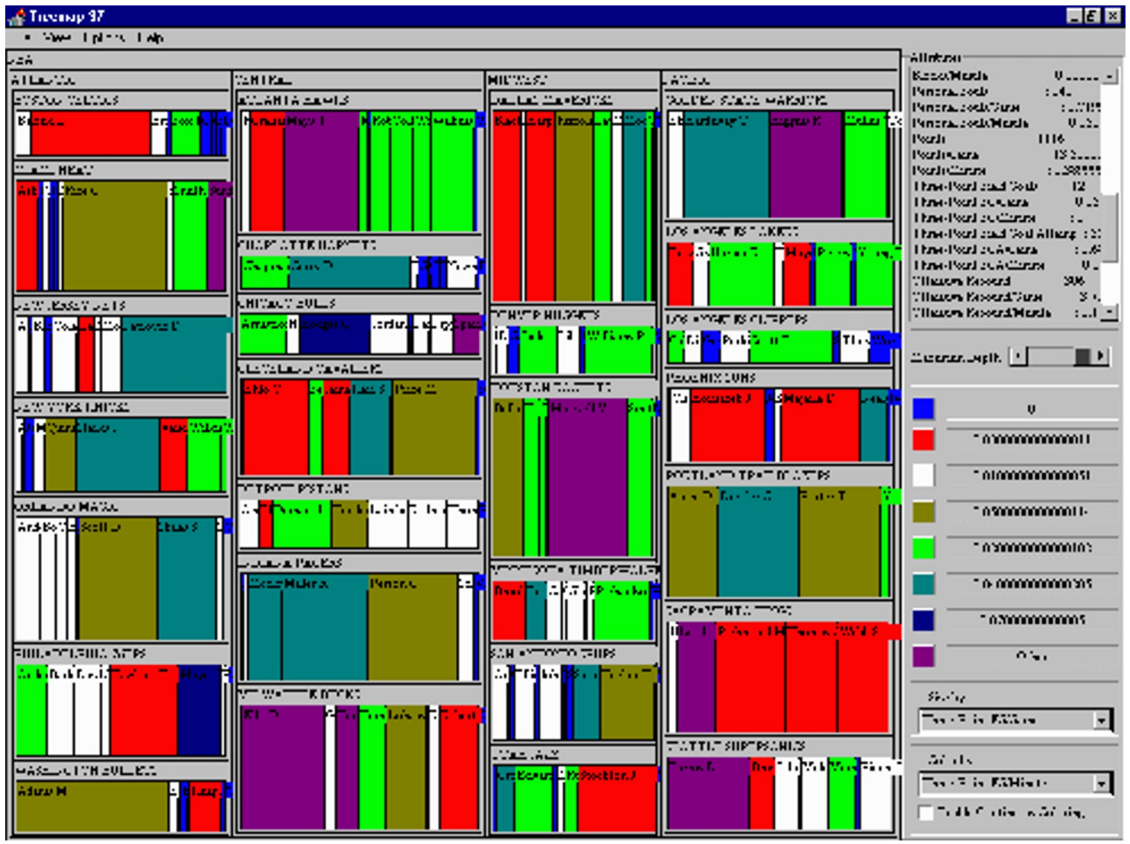
Case Study: The Treemap

- Technique proposed 30 years ago (Johnson & Shneiderman ‘91)

- Idea of Treemap:
  - Show a hierarchy as a 2D layout
  - Fill up the space with rectangles representing objects
  - Size on space indicates relative size of underlying objects

Early Treemap Applied to File Systems

- Each file is a colored rectangle with area proportional to the file's size
- Rectangles are arranged such that directories again make up rectangles with an area that is proportional to the size of the subtrees
- The color of a rectangle indicates the type of the file (extension)



Treemap Problems

- Too disorderly
  - What does adjacency stand for?
  - Uncontrolled aspect ratios leads to lots of cluttered skinny boxes
- Color not used appropriately
  - In fact, color is almost meaningless here
- Wrong application
  - Nobody needs all of this just to see the largest files in the OS

Successful Application of Treemaps

- Think more about why you are using a treemap
  - Break the space into meaningful groups
  - Fix these into a useful aspect ratio
- Use visual properties properly
  - Use color for meaningful distinctions
    - The fewer colors the better
- If possible, provide excellent interactivity
  - It has a navigable tree structure, after all!
  - Access to the real data
  - Makes it into a useful tool

TreeMaps in Action

Finviz: S&P 500 Map

A Good Use of TreeMaps and Interactivity

Key Questions to Ask about a Viz

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

Visual Principles

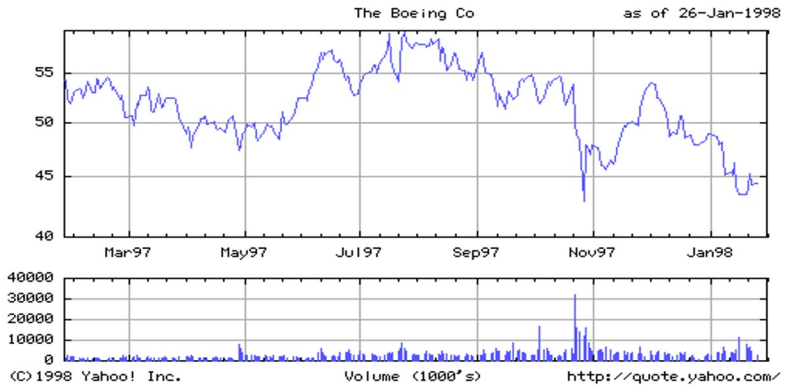
- Types of Graphs
- Pre-attentive Properties
- Expressiveness of Visual Properties

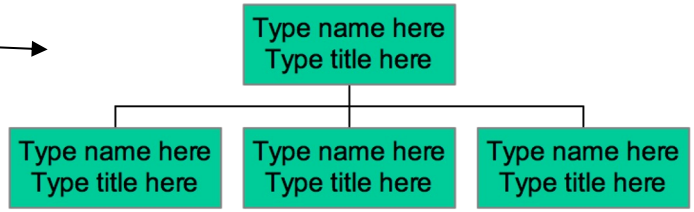
A Graph is: (Kosslyn)

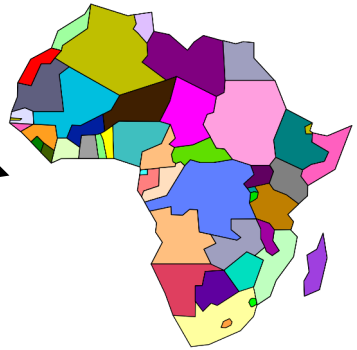
- A visual display that illustrates one or more relationships among entities
- A shorthand way to present information
- Allows a trend, pattern, or comparison to be easily apprehended

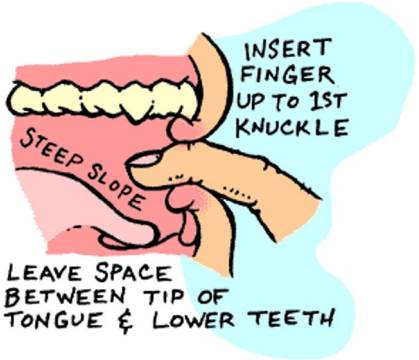
Types of Symbolic Displays (Kosslyn 89)

- Graphs
- Charts
- Maps
- Diagrams







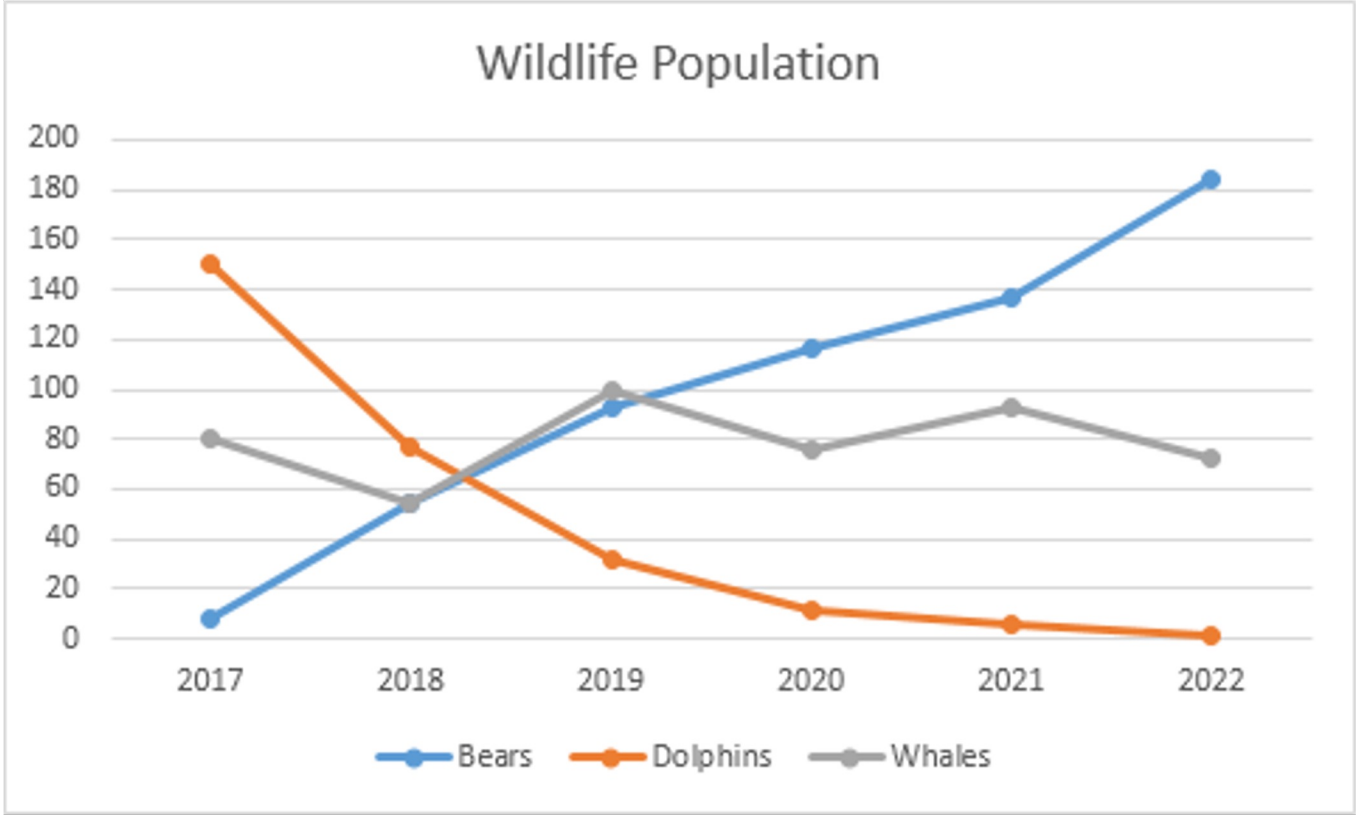


Anatomy of a Graph (Kosslyn 89)

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

When to use which type of graph?

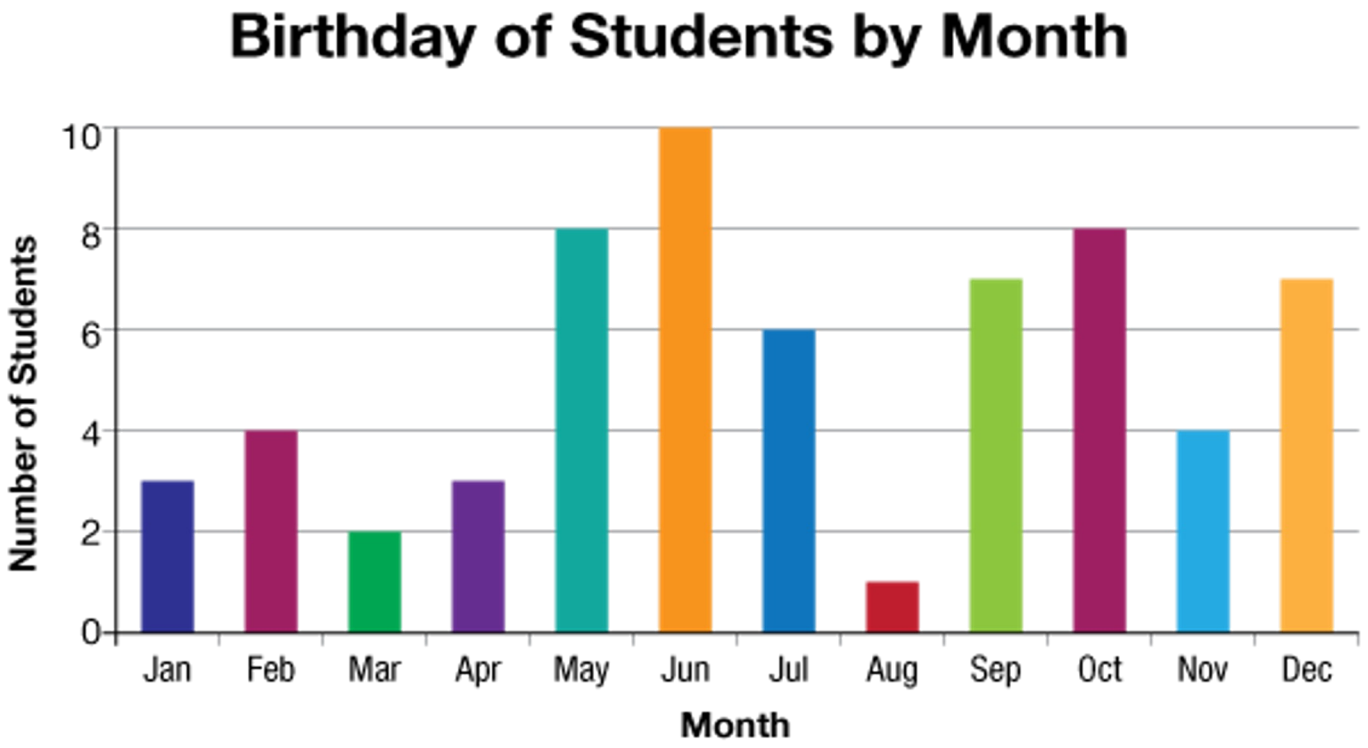
- Line graph
  - X-axis requires quantitative variable
  - Variables can have contiguous values
    - Good for time series and trends



https://medium.com/@patrickbfuller/line-plot-7b4068a3a9fc

When to use which type of graph?

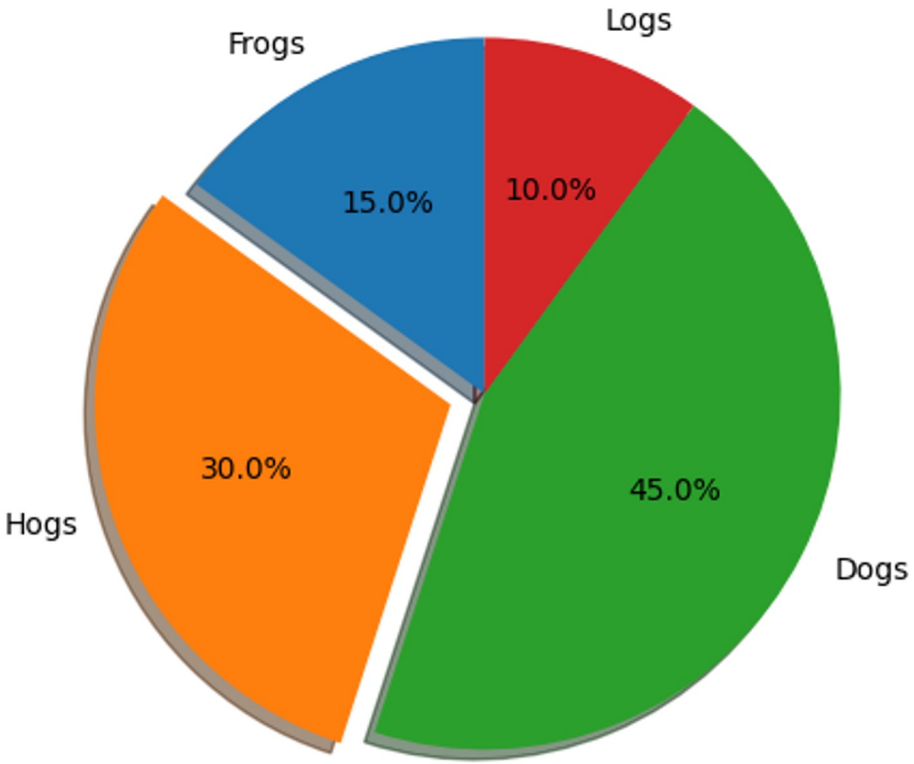
- Bar graph
  - Comparison of relative point values for different groups
  - If there is a natural order in the groups, respect it
    - e.g., each bar corresponds to a month



https://gfchart.com/2021/01/bar-graph-vs-line-graph-which-is-right-for-your-form/

When to use which type of graph?

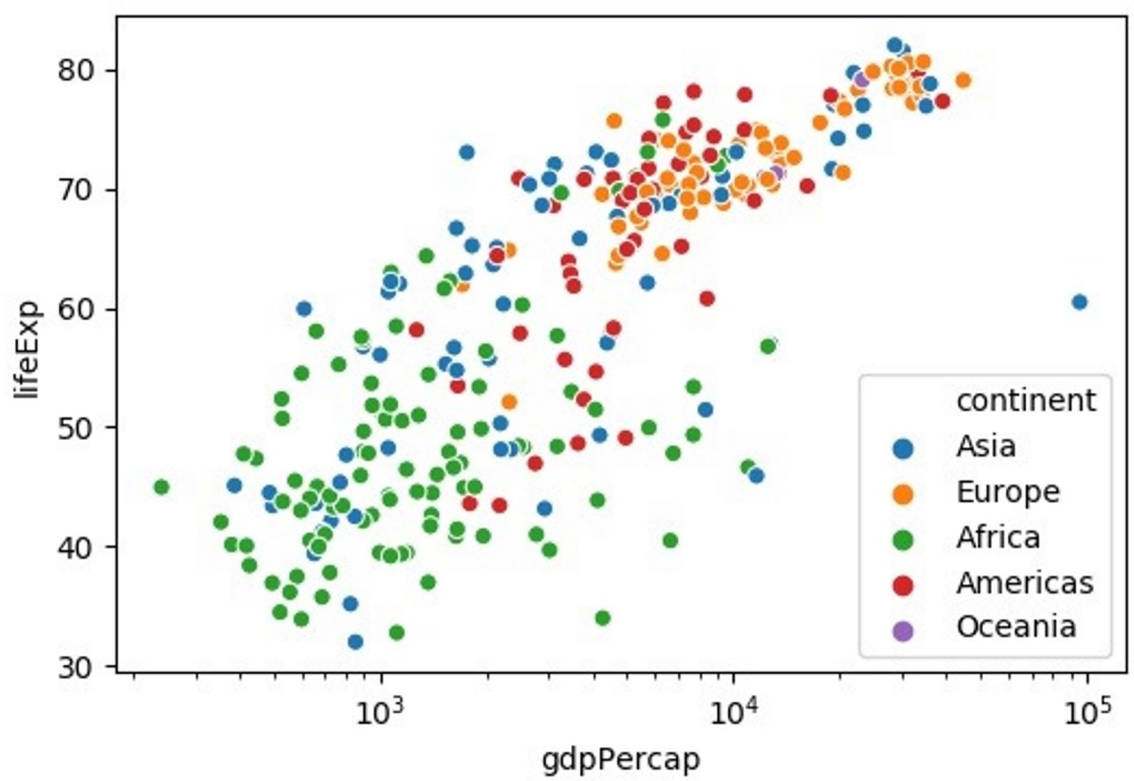
- Pie chart
  - Emphasizes differences in proportion for a few groups



https://matplotlib.org/stable/gallery/pie_and_polar_charts/pie_features.html

When to use which type of graph?

- Scatter plot
  - Illustrates the relationship between two variables
  - Useful for bivariate visual exploration
    - e.g., during exploratory data analysis



https://cmdlinetips.com/2019/04/how-to-specify-colors-to-scatter-plots-in-python/

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 around 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 embarrassed 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 allowed 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?

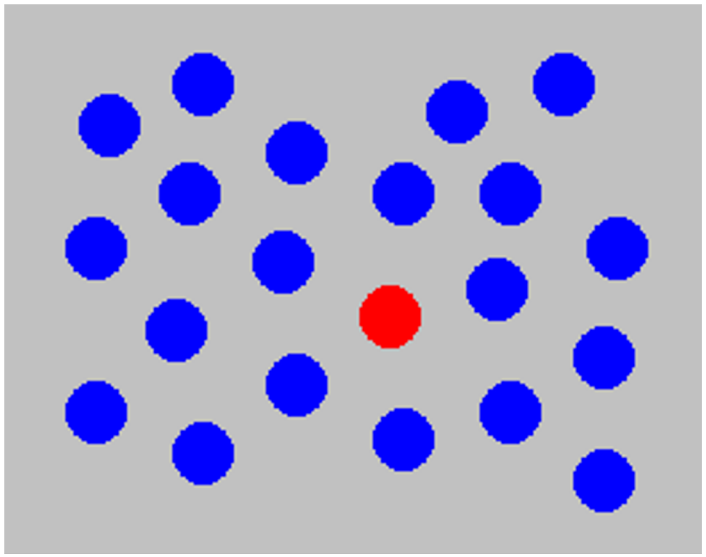
- Table or graph?
  - If a table, which kind?
  - If a graph, what kind of relationship?
  - If a graph, which graphical objects for quantitative encoding?
  - Anything else?

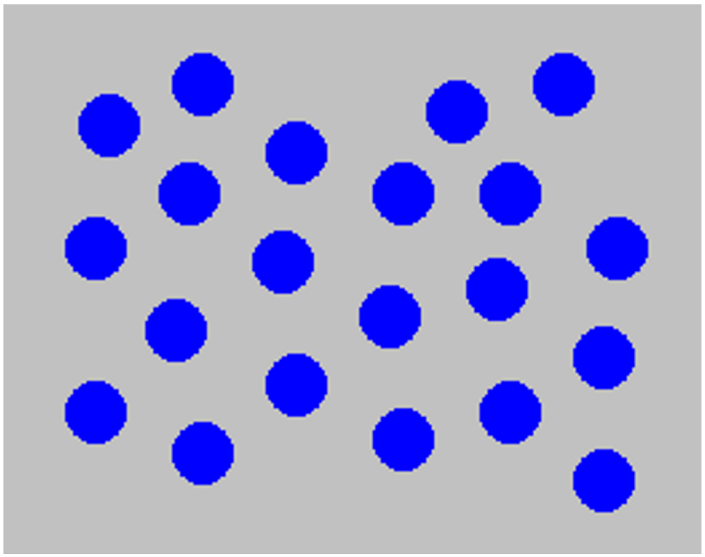
Pre-attentive Processing

- A limited set of visual properties are processed pre-attentively, i.e., without need for focused 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

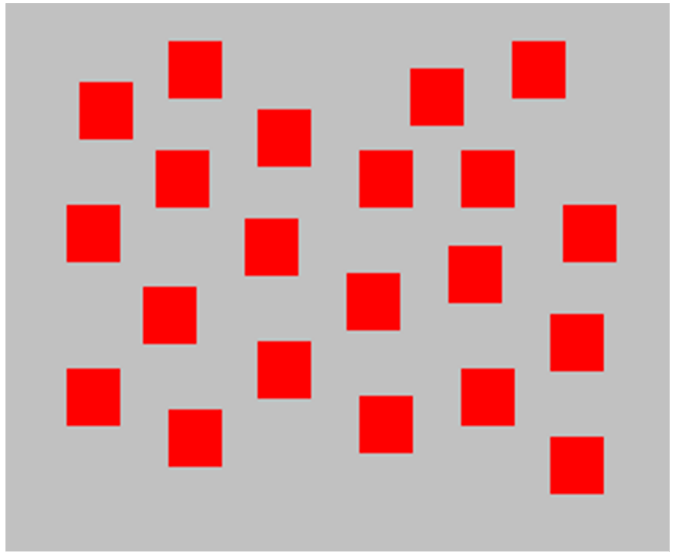


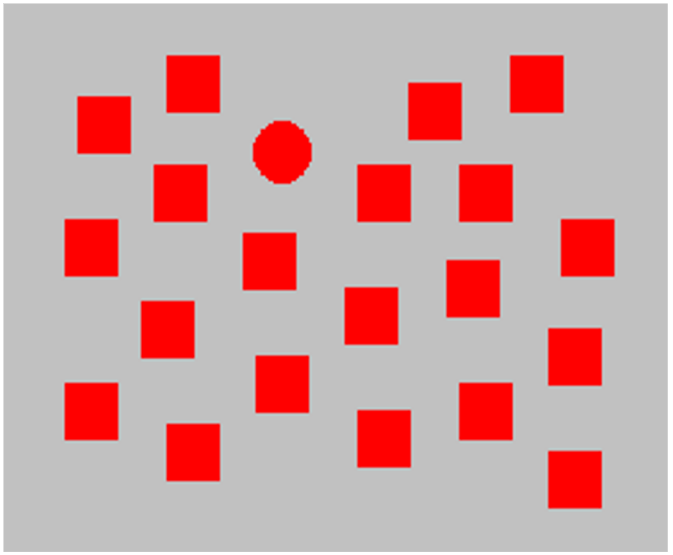


Viewer can rapidly and accurately determine whether the target (red circle) is present or absent

Difference detected in color

Example: Shape Selection





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

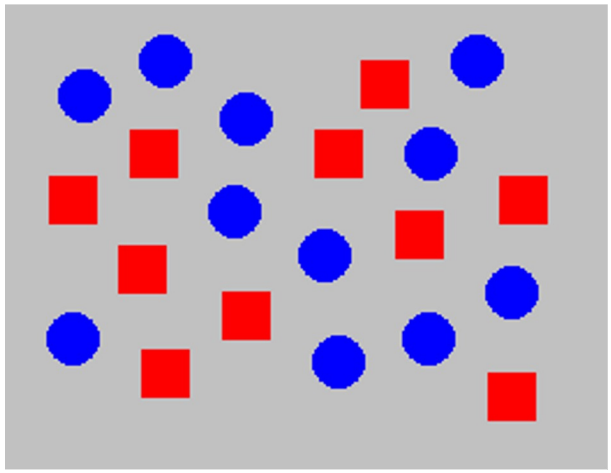
Difference detected in form (curvature)

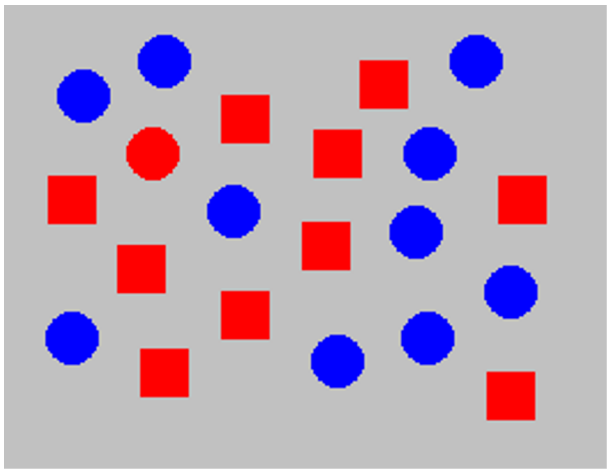
Pre-attentive Processing

- < 200ms - 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 (and very small) amount of time regardless of the number of distractors, it is considered to be pre-attentive

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

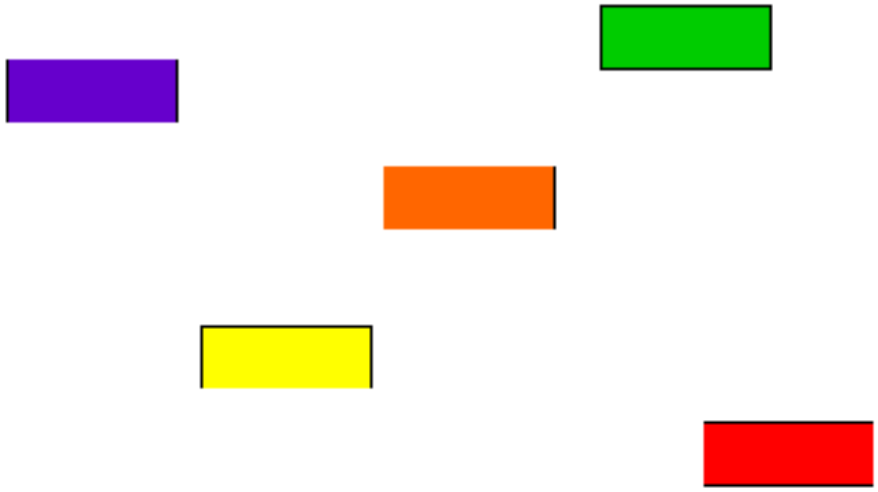
Which Visualization Properties are Appropriate for Which Information Types?

Interpretations of Visual Properties

- Some properties can be discriminated more accurately, even if they do not have intrinsic meaning (Senay & Ingatious 97):
  - Density (Grayscale)
    - Darker ➡ More
  - Size / Length / Area
    - Larger ➡ More
  - Position
    - Leftmost ➡ First
    - Topmost ➡ First
  - Hue
    - No intrinsic meaning

Color Schemes

Order these (low->hi)



Color: Different Purposes

- Call attention to specific items
- Distinguish between classes of items
- Increase the appeal of the visualization

Using Color

- Proceed with caution
  - Less is more
  - Representing magnitude with color is tricky
- Examples
  - Green-light green-light brown-dark b grey-white:
    - works for atlases and maps
  - Grayscale:
    - unambiguous but has limited range
- Be inclusive
  - Consider working with scales that are colorblind-friendly

topography vs. vegetation?
green == lush?
yellow == desert?

What are good guidelines for Infoviz?

- Use graphics appropriately
  - Do not use images gratuitously
  - Do not lie with graphics
    - Add links to original data when possible
- Make it interactive when possible (feedback)
  - Multiple views
  - Overview + details
- Match mental models

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

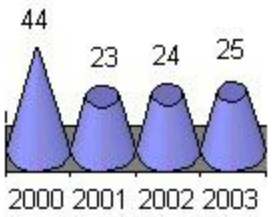
LIS 504 - Graphic displays of data - Netscape

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.

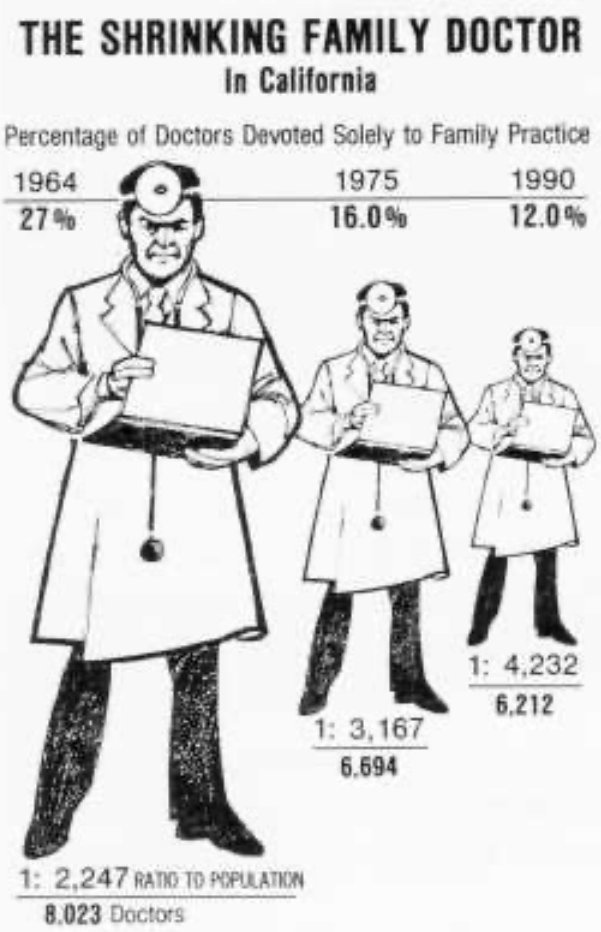


The heights of the (truncated) cones are proportional to the data, but their areas on the screen and their apparent volumes make the larger data values seem relatively small.

Document: Done

How to Exaggerate with Graphs

from Tufte ’83

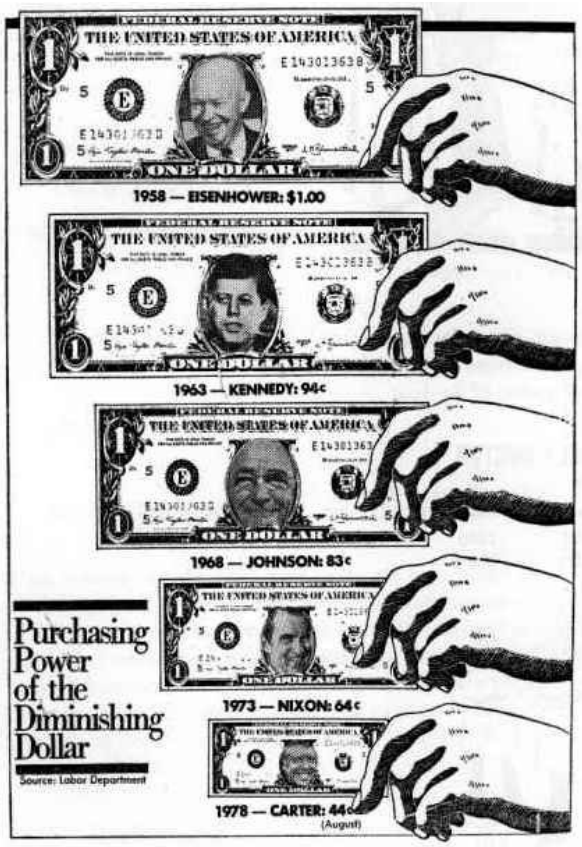


"Lie factor" = 2.8

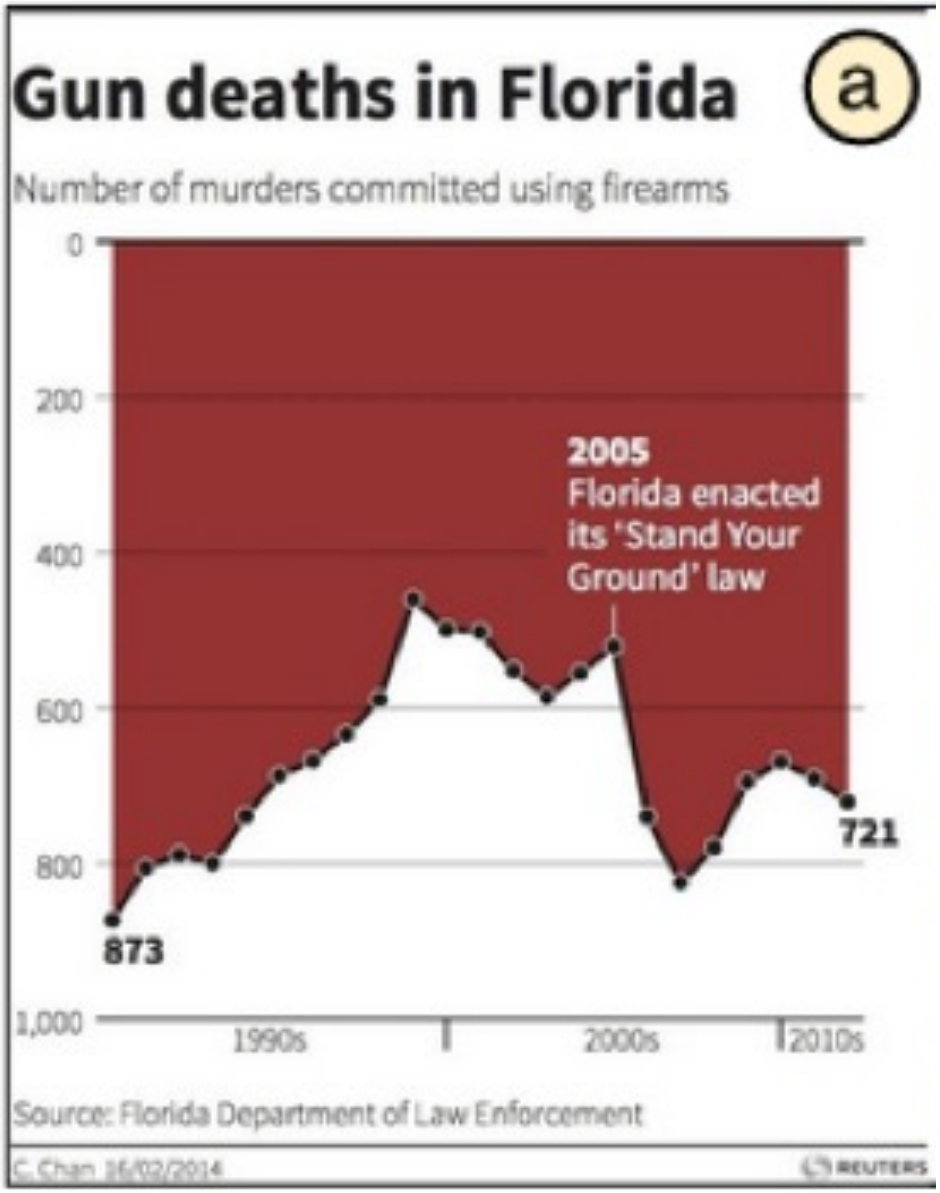
Los Angeles Times, August 5, 1979, p. 3.

How to Exaggerate with Graphs

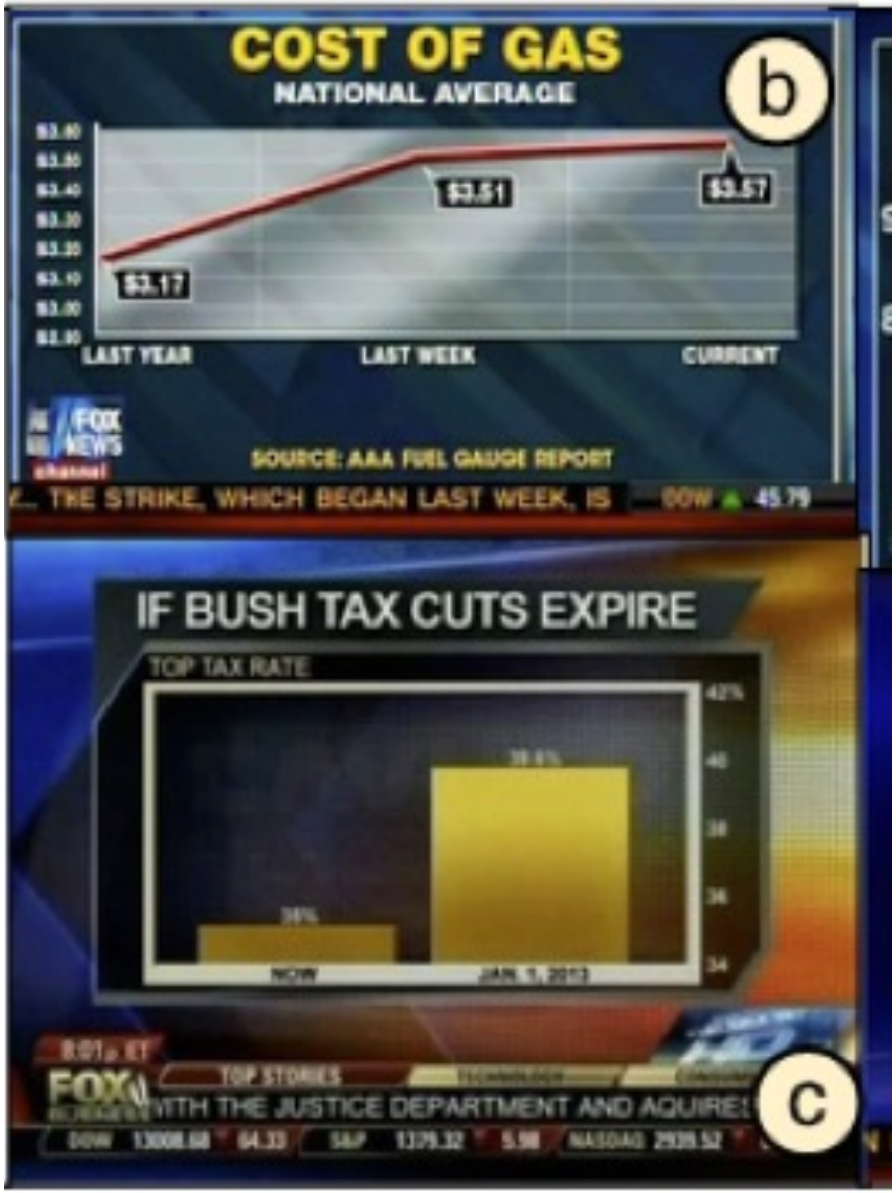
from Tufte ’83



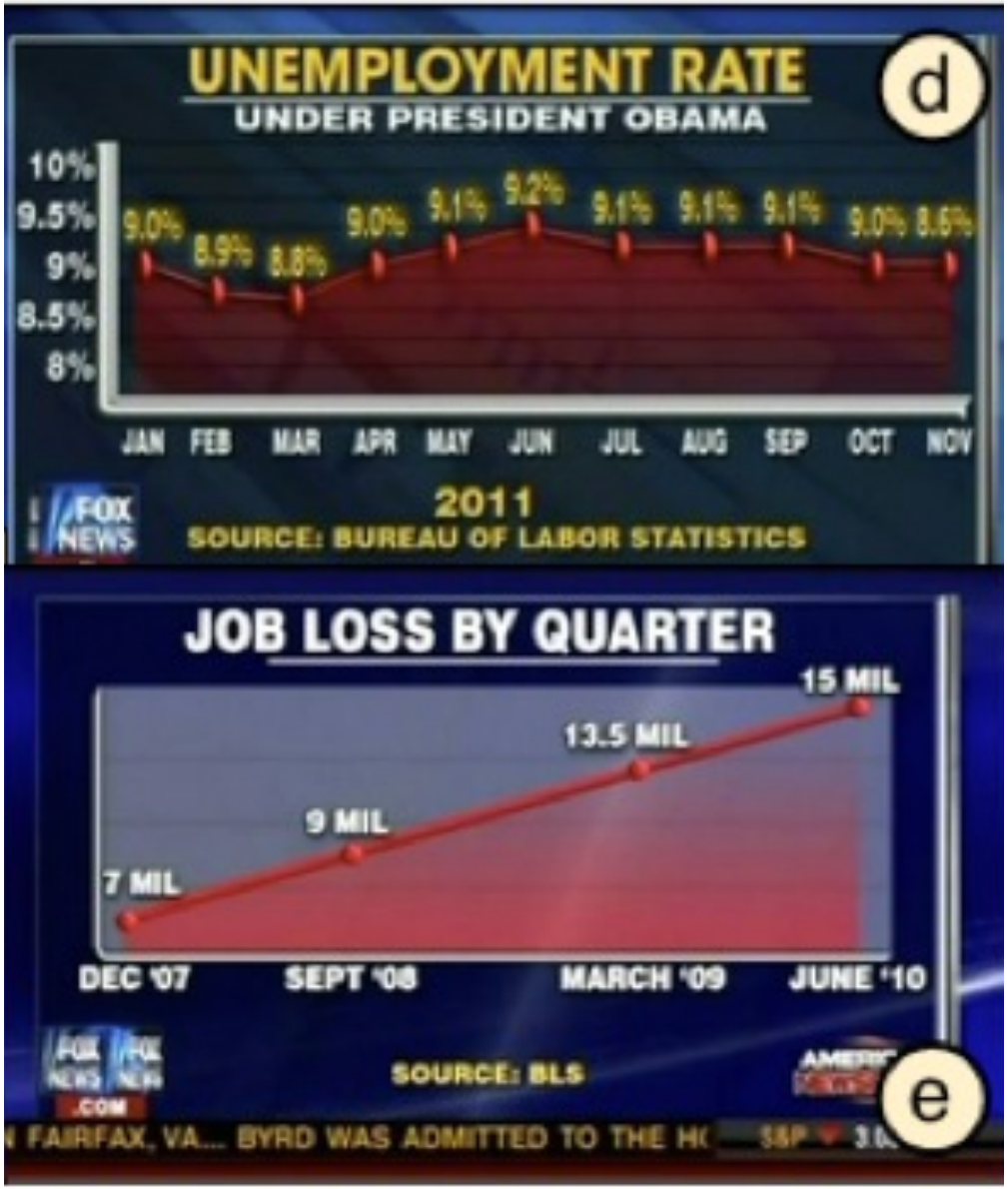
Washington Post, October 25, 1978, p. 1.



taken from: How Deceptive Are Deceptive Visualizations: An Empirical Analysis of Common Distortion Techniques. Anshul Vikram Pandey, Katharina Rall, Margaret L. Satterthwaite, Oded Nov, Enrico Bertini; ACM CHI 2015

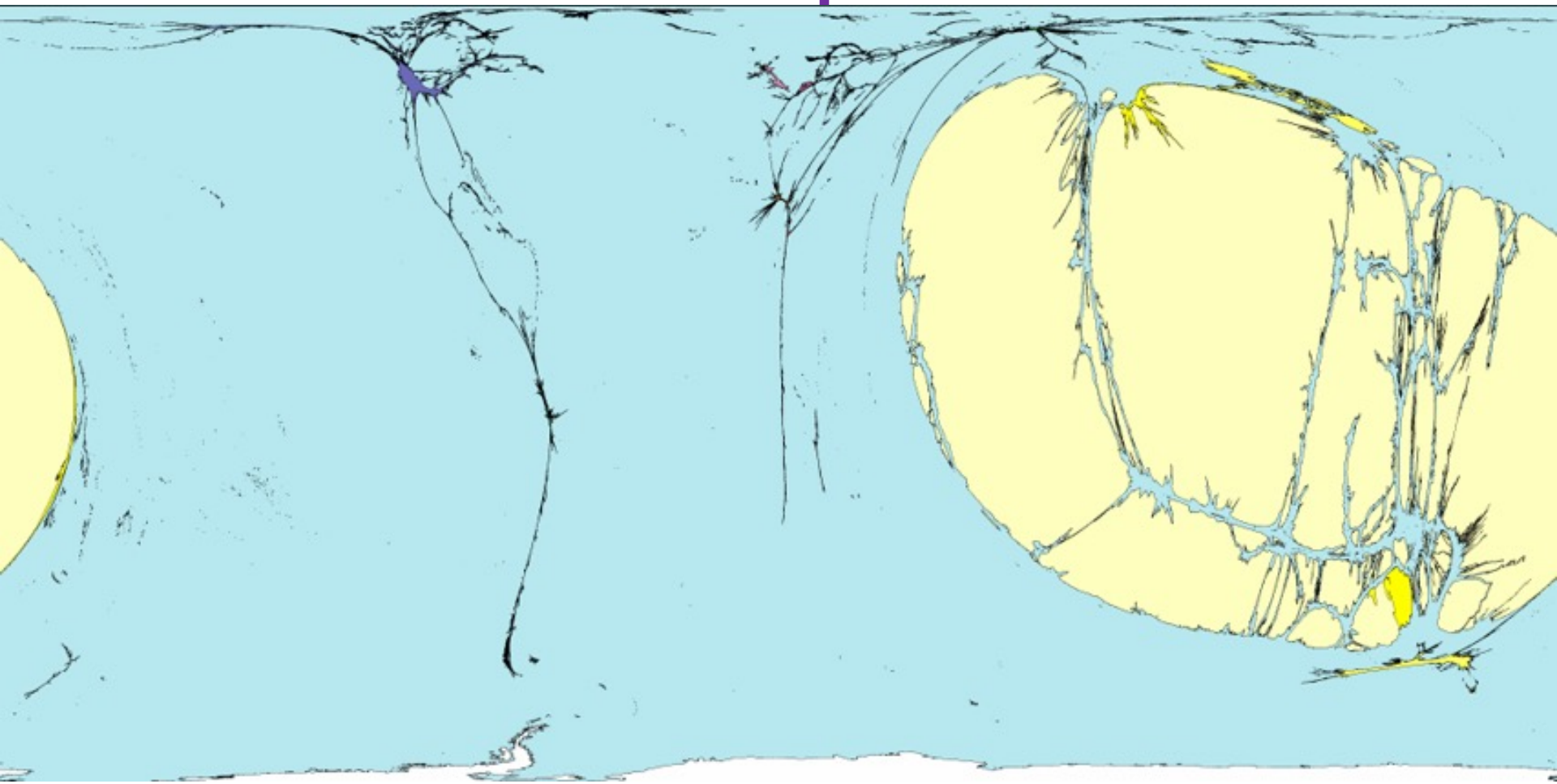


taken from: How Deceptive Are Deceptive Visualizations: An Empirical Analysis of Common Distortion Techniques. Anshul Vikram Pandey, Katharina Rall, Margaret L. Satterthwaite, Oded Nov, Enrico Bertini; ACM CHI 2015



taken from: How Deceptive Are Deceptive Visualizations: An Empirical Analysis of Common Distortion Techniques. Anshul Vikram Pandey, Katharina Rall, Margaret L. Satterthwaite, Oded Nov, Enrico Bertini; ACM CHI 2015

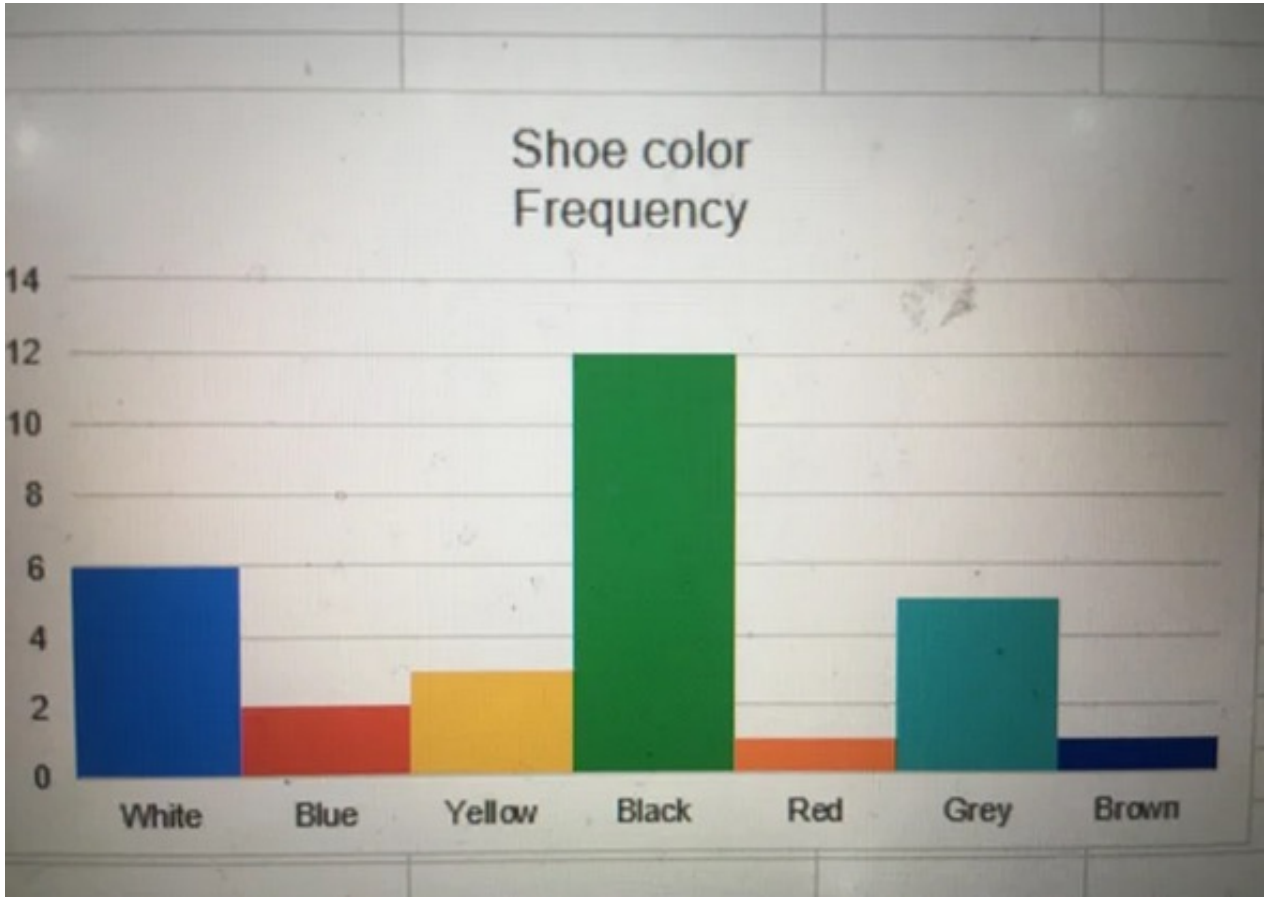
Worst Visualization Examples



"Indonesian Speakers of the WorldMicroscope Slide"

Examples from: Jared Hutchins

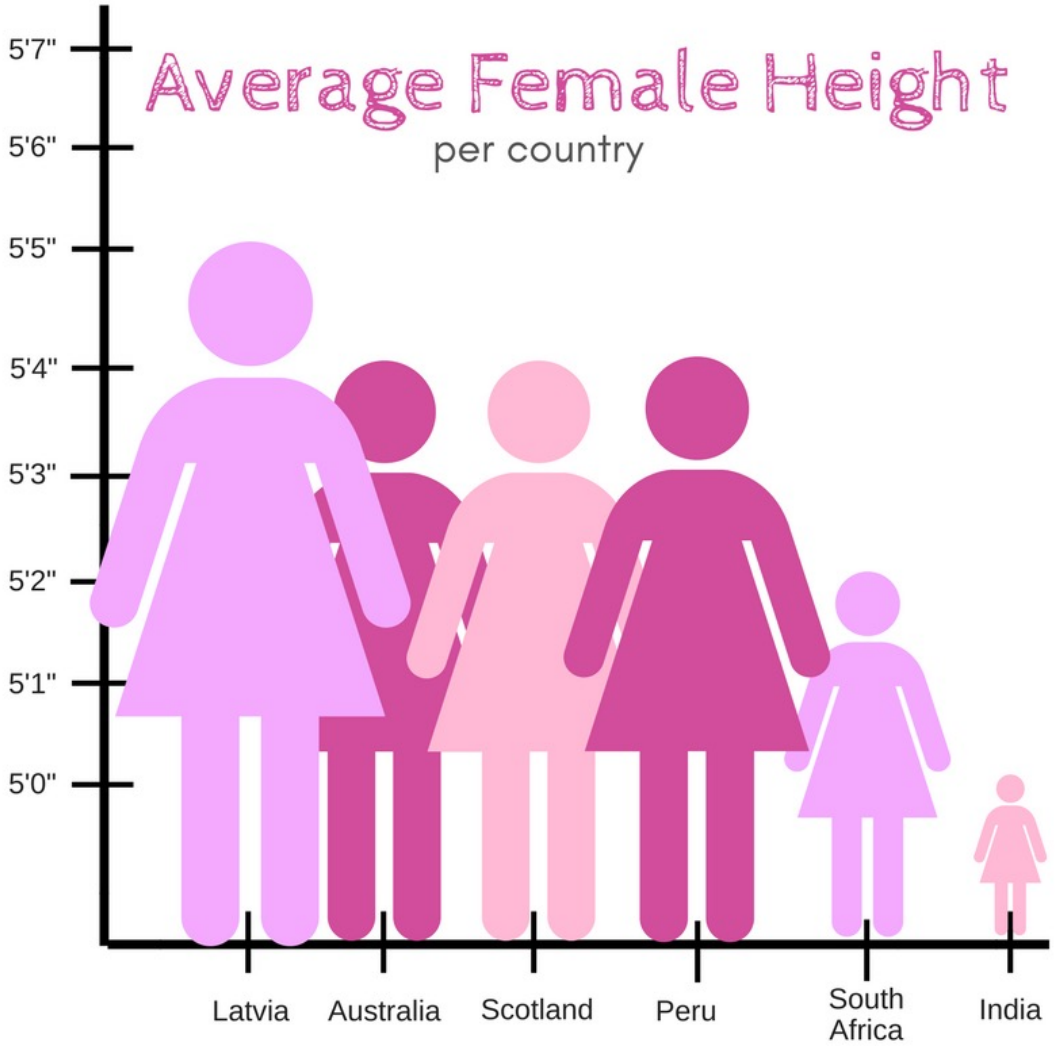
Worst Visualization Examples



"Excel Defaults"

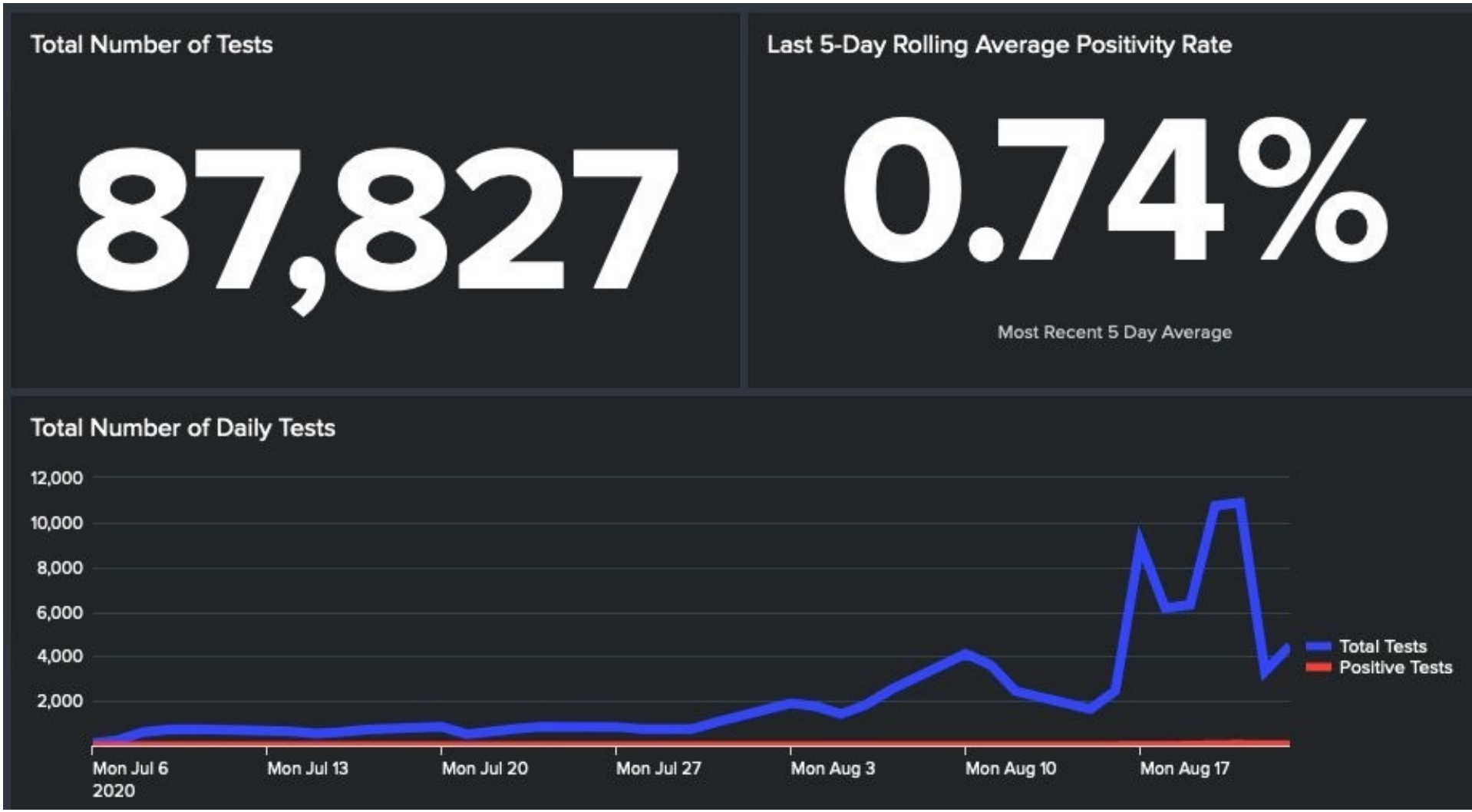
Examples from: Jared Hutchins

Worst Visualization Examples



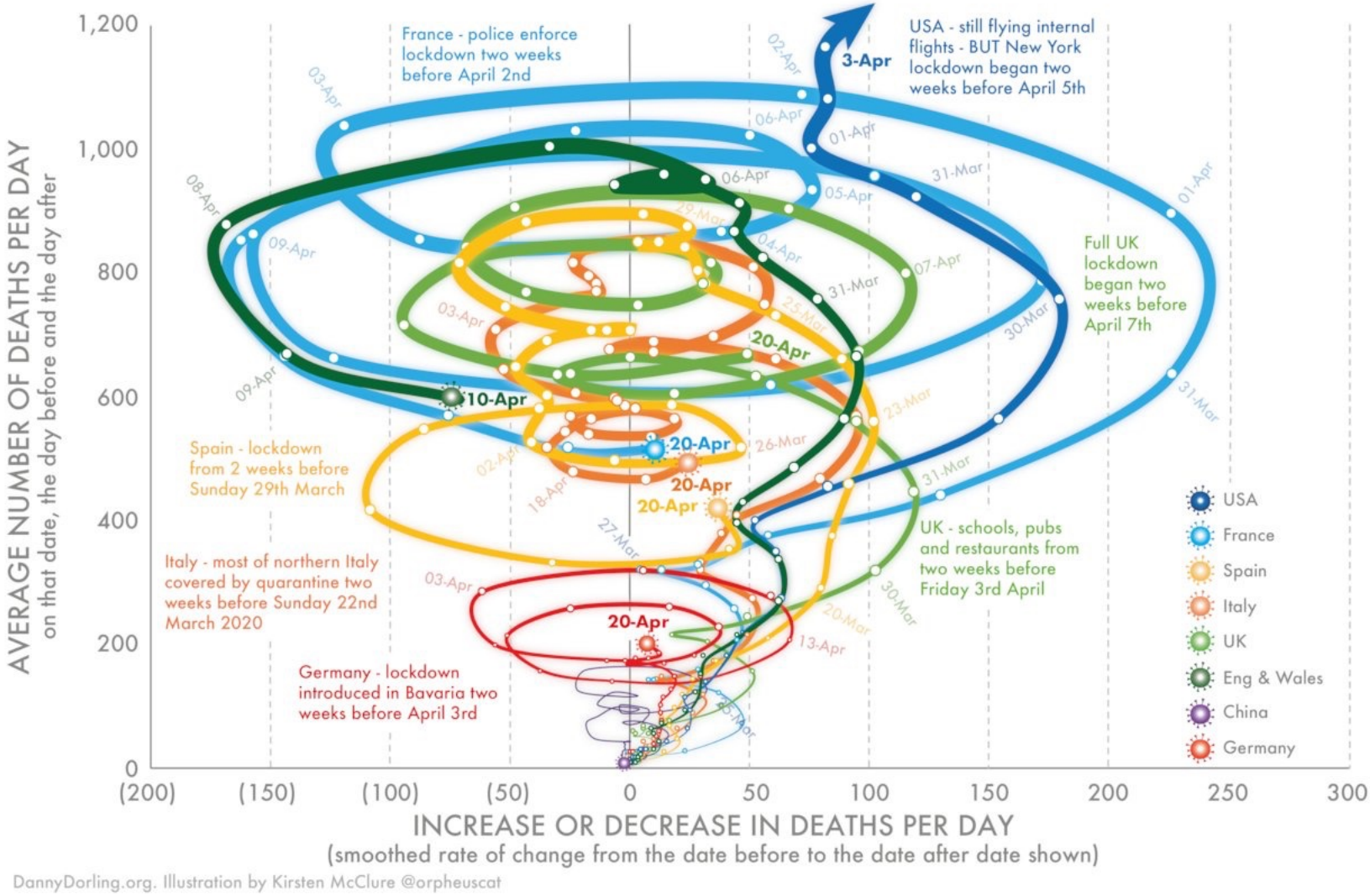
Worst Visualization Examples

Worst Visualization Examples



"What positive tests?"

Worst Visualization Examples



COVID Death Tornado

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.

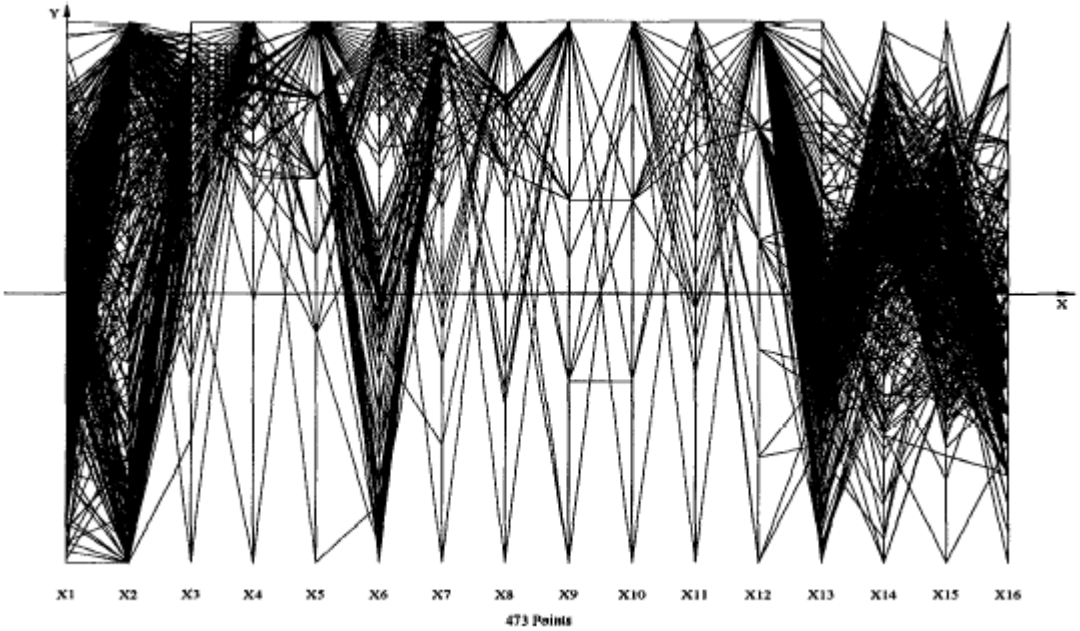


Figure 1: The full dataset consisting of 473 batches

A Detective Story

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

- 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

- 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.

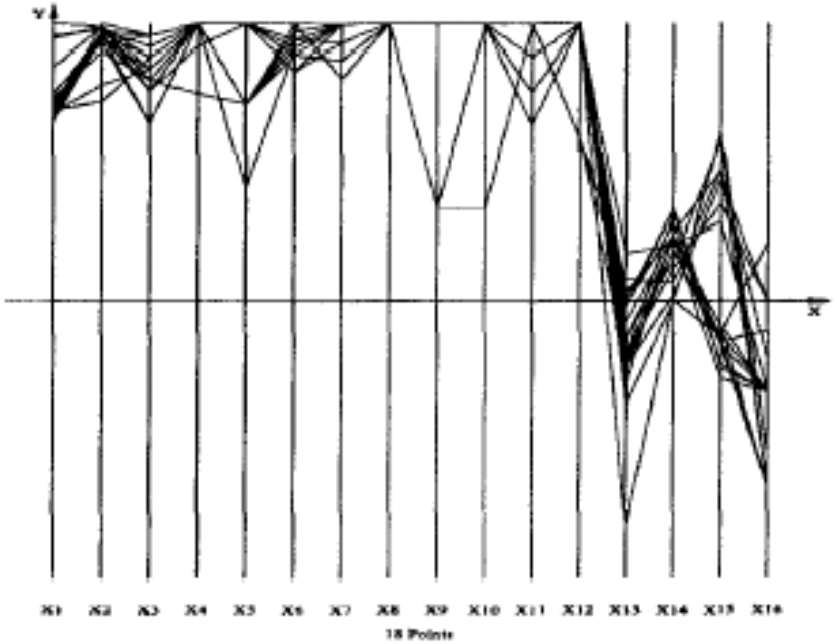


Figure 2: The batches high in Yield, $X1$, and Quality, $X2$.

Multidimensional Detective

- 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

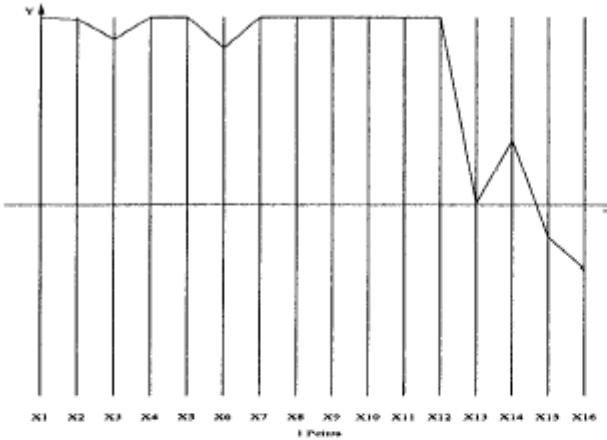


Figure 5: The best batch. Highest in Yield, $X1$, and very high in Quality, $X2$.

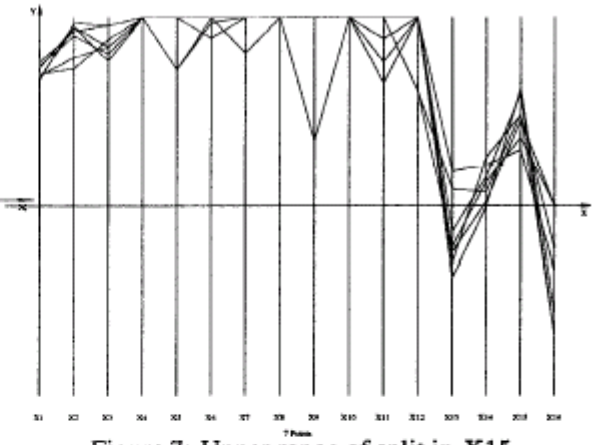
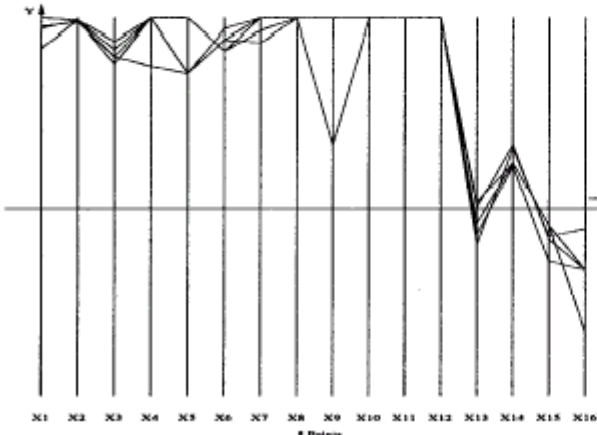
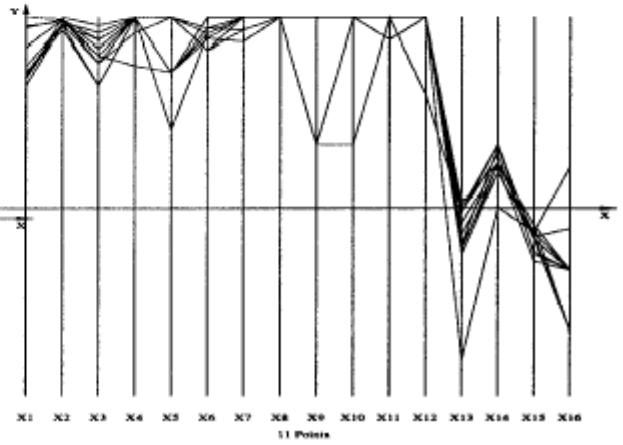


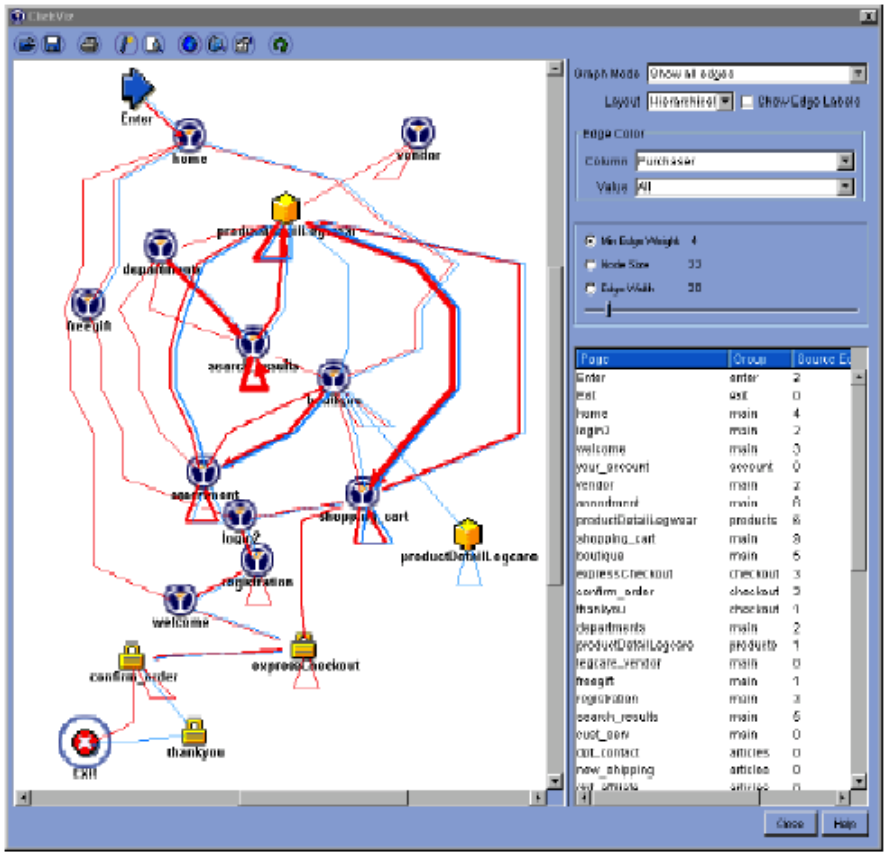
Figure 7: Upper range of split in $X15$





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



Customer Segments

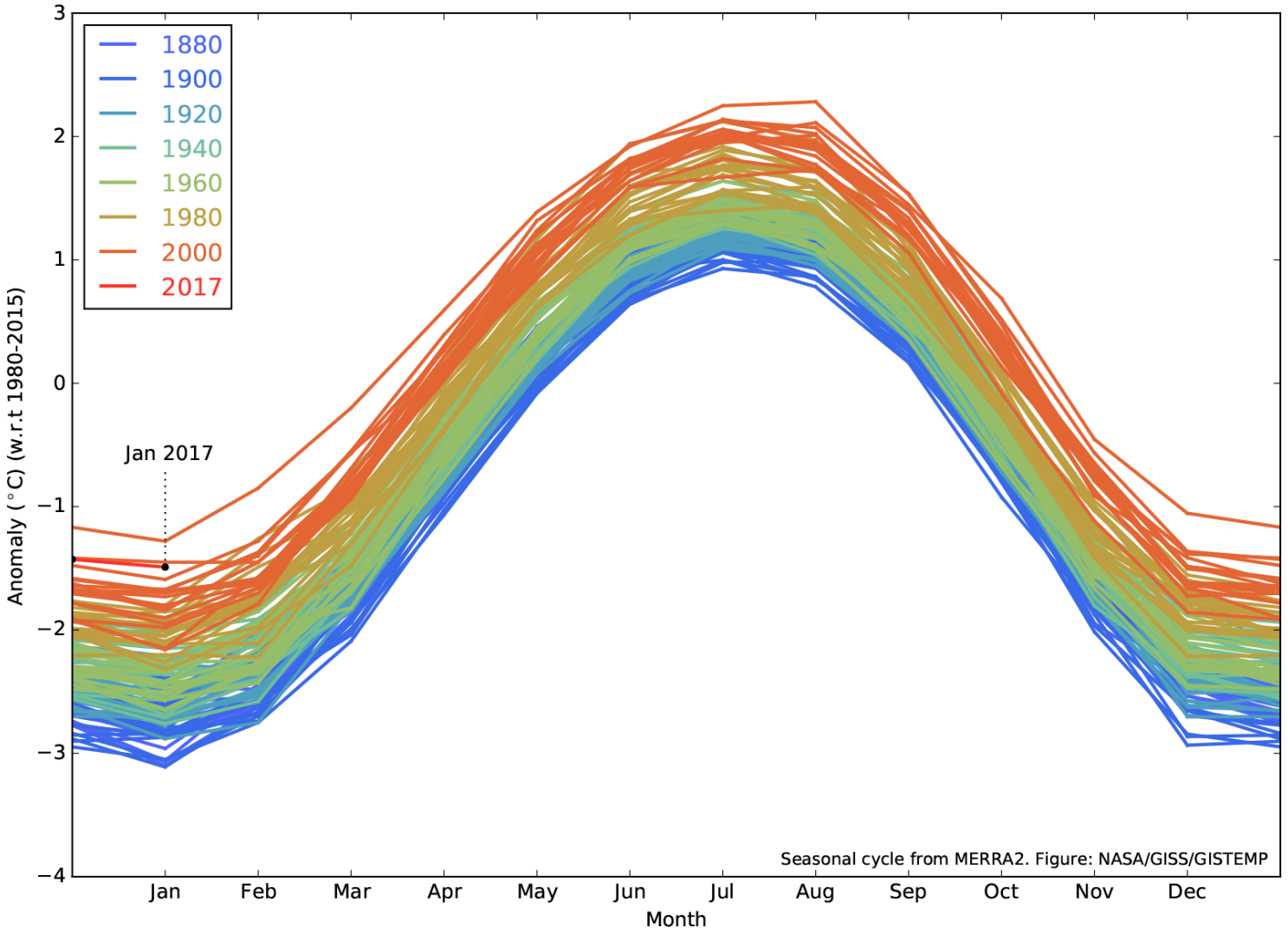
- Collect
  - Clickstream
  - Purchase history
  - Demographic data
- Associates customer data with their clickstream
- Different color for each customer segment

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?

Cycle Plot



The GISTEMP monthly temperature anomalies superimposed on a 1980-2015 mean seasonal cycle.

Credit: NASA/GISS.

Analyzing Complex Periodic Data

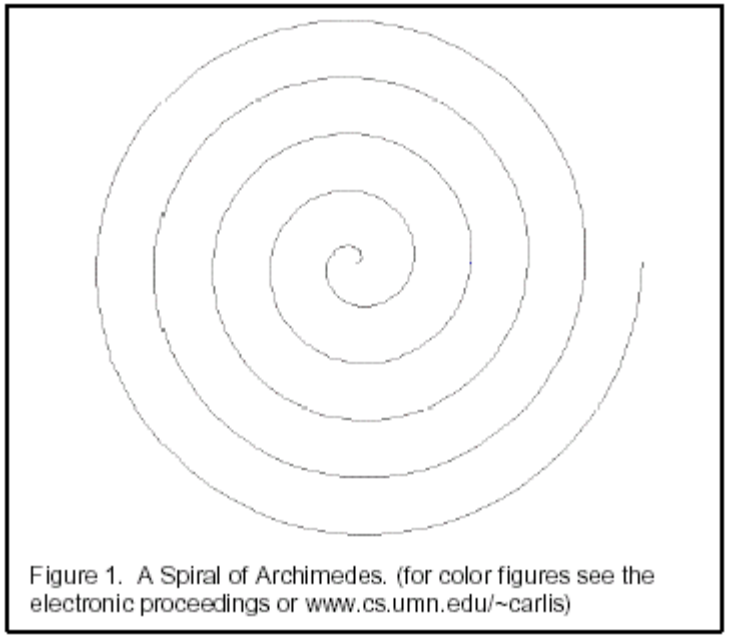


Figure 1. A Spiral of Archimedes. (for color figures see the electronic proceedings or www.cs.umn.edu/~carlis)

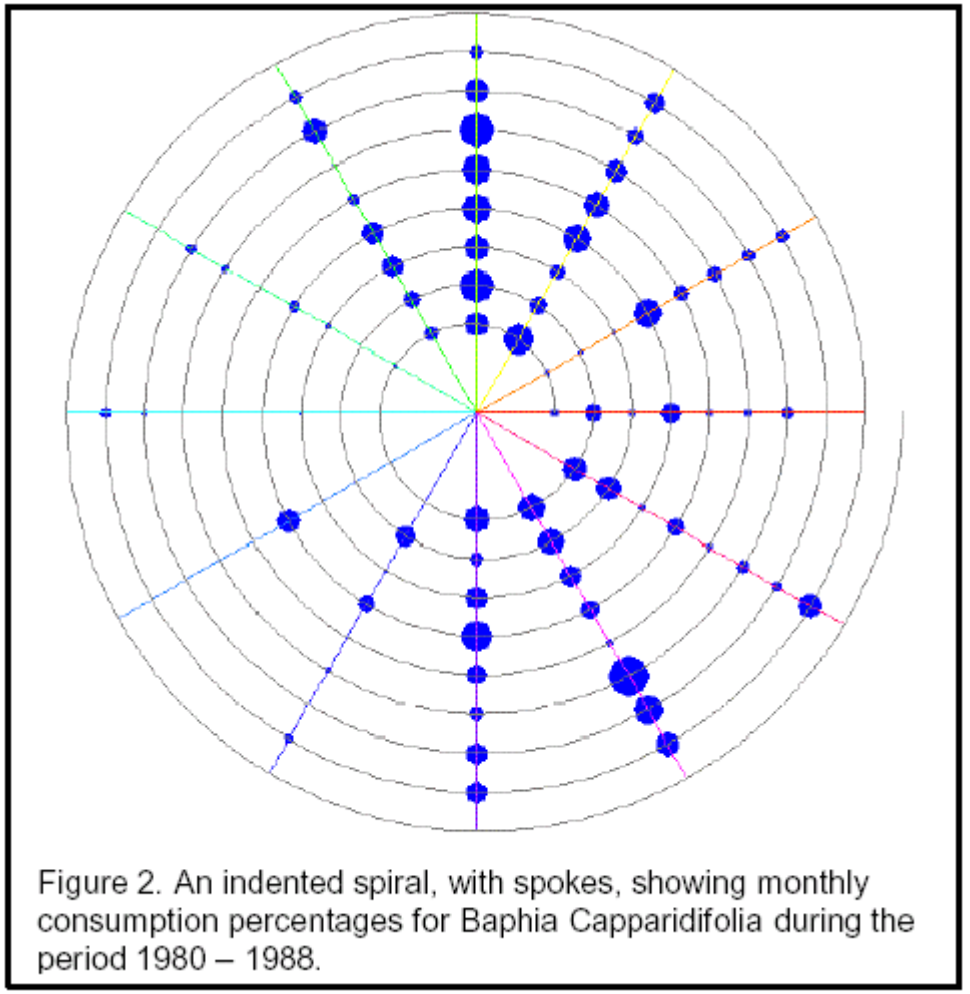


Figure 2. An indented spiral, with spokes, showing monthly consumption percentages for Baphia Capparidifolia during the period 1980 – 1988.

Carlis & Konstan, UIST 1998.

Analyzing Complex Periodic Data

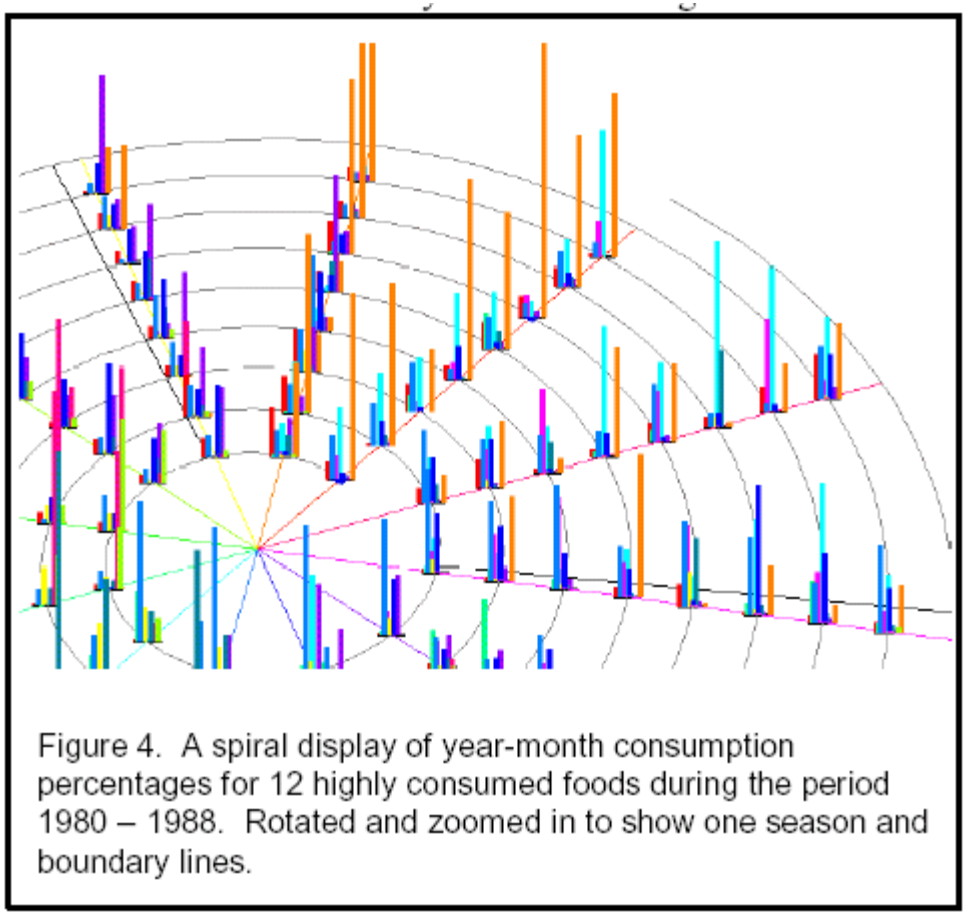


Figure 4. A spiral display of year-month consumption percentages for 12 highly consumed foods during the period 1980 – 1988. Rotated and zoomed in to show one season and boundary lines.

- 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?

Learning more about Info Viz

- Online visualization courses/certificates
- matplotlib: library for python
- d3: javascript library (out of the box)
- react: javascript library

Visualization and ML

- Improving interpretability
- Human-in-the-loop models
- Automated testing of visualization performance

New Course: CS-GY 9223 - I (24680)
Visualization: Connections with Machine Learning