Introduction to Data Visualization

Jason Leigh

Laboratory for Advanced Visualization & Applications
University of Hawai‘i at Mānoa

We Only Have 2 Hours

- Visualization is like literacy – it takes practice to do it well
- Being aware of and developing an allergic reaction to ugly
- Feeding your brain with beautiful things
- Some tips that are hopefully useful on your journey
  - Having a Process
  - Picking the right visual encodings
  - Picking the right colors
  - Picking the right chart
  - Knowing the tools
  - Wielding the tools

Why Visualize?

	I		II		III		IV	
	x	y	x	y	x	y	x	y
	10	8,04	10	9,14	10	7,46	8	6,58
	8	6,95	8	8,14	8	6,77	8	5,76
	13	7,58	13	8,74	13	12,74	8	7,71
	9	8,81	9	8,77	9	7,11	8	8,84
	11	8,33	11	9,26	11	7,81	8	8,47
	14	9,96	14	8,1	14	8,84	8	7,04
	6	7,24	6	6,13	6	6,08	8	5,25
	4	4,26	4	3,1	4	5,39	19	12,5
	12	10,84	12	9,13	12	8,15	8	5,56
	7	4,82	7	7,26	7	6,42	8	7,91
	5	5,68	5	4,74	5	5,73	8	6,89
SUM	99,00	82,51	99,00	82,51	99,00	82,50	99,00	82,51
AVG	9,00	7,50	9,00	7,50	9,00	7,50	9,00	7,50
STDEV	3,32	2,03	3,32	2,03	3,32	2,03	3,32	2,03

- Anscombe's Quartet
- 4 Data Sets
- Same statistical properties

- Averages are the same
- Variance of X is 11, Variance of Y is 4.13
- Correlation coefficient (how strong a relationship is between 2 variables) between X & Y is 0.816 for each data set

y1
x1
y2
x2
y3
x3
y4
x4
4 6 8 10 12 14 16 18
4 6 8 10 12

More info enters the brain via sight than any other sense, both sub-consciously and consciously

Role of Visualization

- Allows you to find problems with the data itself & hence rectify data gathering early before more data is collected.
- Allows you to comprehend large amounts of data.
- Helps you see both large-scale and small-scale features to understand the forest for the trees.
- Allows the perception of unanticipated emergent features.
- Helps you answer questions.
- Helps you ask new questions (hypothesis formation).
- Facilitates the explanation of findings / storytelling.

Napoleon’s March to Moscow

Charles Joseph Minard's 1861 graphic showing Napoleon's losses during his 1812 march to and from Moscow - possibly the best statistical graph ever drawn ... why?

- The chart show 5 variables
- size of the army
- location on a 2D surface
- direction of the army's movement
- temperature
- location of major river crossings
- Like many analytical visualizations this gives us a way to see relationships between different types of related data.
- Today we have the ability to make dynamic visualizations that encourage active exploration beyond just looking.
- How would you enhance this visualization if it was software-based?

90 Years of Measles
Before and After Vaccines
University of Pittsburg School of Public Health

The Process

1. Understand your audience
   - Who is the visualization for?
   - What is their level of understanding of a subject?
   - What are they trying to verify/confirm?
   - What are they trying to discover?
   - What questions / hypotheses are they trying to answer?
   - What story are they trying to tell?
2. For each of your questions, identify which attributes in the data need to be plotted to answer them.
   Each grouping of attributes could be considered 1 chart that needs to be plotted.
3. Draw empty boxes and sketch out how attributes might be visualized. Consider also how a user might interact with the visualization- e.g. turn on/off data layers, brush layers.
4. Now lay out the empty boxes as a “dashboard” and also consider which boxes could potentially be merged because they share data attributes.
5. Make a prototype. As a prototype emerges, more questions will emerge & new charts will need to be created and/or merged until you reach the final product.
6. Get feedback from your customer. Refine based on feedback.

COVID Example

https://miro.com/app/board/o9J_koVDLcI=/

You can also sketch on note cards

5 You Might End Up With Something Like This

https://hotspotmap-coronavirus.github.io

Recognizing Ugly

Rank	Top 10 PC Game Titles: 10/08	Share	Avg. Mins Per Week	TMP%
1	World of Warcraft	12.509	546	62.280
2	Warcraft III	1.890	328	4.751
3	Half-Life 2	1.917	288	4.301
4	Spore	2.234	268	4.220
5	Bookworm	3.074	133	2.716
6	Cake Mania 3	4.260	138	2.709
7	Virtual Villagers: The Secret City	1.567	204	2.456
8	Chessmaster Challenge	8.856	47	2.123
9	Halo: Combat Evolved	2.005	165	2.105
10	Build-a-Lot 2: Town of the Year	1.661	183	1.668

Source: Nielsen Games

Share is the percentage of the gaming audience measured that played the title

Overall TMP% represents the percent of total minutes played as compared to all other PC games measured

Rank	Top 10 PC Game Titles: 10/08	Share	Avg. Mins Per Week	TMP%
1	World of Warcraft	12.509	546	62.280
2	Warcraft III	1.890	328	4.751
3	Half-Life 2	1.917	288	4.301
4	Spore	2.234	268	4.220
5	Bookworm	3.074	133	2.716
6	Cake Mania 3	4.260	138	2.709
7	Virtual Villagers: The Secret City	1.567	204	2.456
8	Chessmaster Challenge	8.856	47	2.123
9	Halo: Combat Evolved	2.005	165	2.105
10	Build-a-Lot 2: Town of the Year	1.661	183	1.668

Rank	Top 10 PC Game Titles: 10/08	Share	Avg. Mins Per Week	TMP%
1	World of Warcraft	12.51	546	62.28
2	Warcraft III	1.89	328	4.75
3	Half-Life 2	1.92	288	4.30
4	Spore	2.23	268	4.22
5	Bookworm	3.07	133	2.72
6	Cake Mania 3	4.26	138	2.71
7	Virtual Villagers: The Secret City	1.57	204	2.46
8	Chessmaster Challenge	8.86	47	2.12
9	Halo: Combat Evolved	2.01	165	2.11
10	Build-a-Lot 2: Town of the Year	1.66	183	1.67

- The results should not show greater accuracy than the original measurements. Don't let Excel etc add on extra digits.
- The numbers should only show enough precision for the audience to see the trend – more digits can obscure that trend.
- Space padding for improved readability
- Right justify numbers
- Left justify text
- Align decimal points
- Do not show more decimal points than original accuracy of data
- Show only enough decimal points for viewer to see a trend

Tips on Font Selection

More On Text

- sans-serif (e.g. Helvetica, Arial, Verdana, Tahoma) good for on-screen text
- serif (e.g. Times, Georgia) - good for printed text – not great for screens.
- monospace - (e.g. courier) good for certain occasions when you need exact alignment of the text (e.g. code)
- Fantasy / cute / brush strokes / cursive / dripping blood - save these fonts for party invitations.
- lower case words are read faster than words in upper case
- individual letters and nonsense words like UA1416 are read faster in upper case.

Messages That Fonts Convey

Graphs

- Make sure graph lines and glyphs are easily visible from back of the room- both in terms of size and color.
- Avoid fully saturated colors (e.g. Red = 255,0,0)
- Microsoft's default graph coloring ISN'T always the best.
- Make sure there are units.
- Make sure your numbers don't have extraneous significant digits.

Graphs

- Avoid 3D charts & graphs – they look pretty but make it harder to compare data & are often used to deceive viewers.
- 29% looks bigger than 35% !!!!

Could also simplify this and put (millions)

- Former graph suggests US is experiencing aggressive population growth.
- But is that really true?
- Vertical axis should start from zero if possible. Otherwise it can also deceive the viewer.

When Presenting Graphs in a Slide Presentation

- Give viewers time to interpret the axes of the graph and then interpret the data in the graph.
- So take the time to explain the graph e.g. explain the axes.
- Put a conclusion at the bottom of each graph.

You achieve the best lap time with cars that are lighter and more powerful.

Pie charts are ONLY good for comparing 2-3 different data points whose values are very different.

Pie charts are POOR for comparing between themselves

E.g. Data comparing 3 polling stations for votes for 5 candidates

More informative and easier to compare the % with bar charts

A pie chart is poor for comparing between ITSELF

This pie chart only tells you that the 2 largest parties are EUL-NGL and S&D

EUL-NGL > S&D

A simple bar chart is more informative and you don‘t need to waste a whole lot of colors.

McDonald's
9.1 mg / fl. oz
Seattle's Best
10.4 mg / fl. oz
Biggby Coffee
12.5 mg / fl. oz
Dunkin' Donuts
12.7 mg / fl. oz
Dutch Bros. Coffee
12.8 mg / fl. oz
Caribou Coffee
15 mg / fl. oz
Peet's Brewed Coffee
16.7 mg / fl. oz
Starbucks
20.6 mg / fl. oz
Deathwish Coffee
54.2 mg / fl. oz

How We See

Alberto Cairo's The Functional Art

Iconic Memory

- Short-term buffer & processor to maintain a coherent picture of the world at all times. Also perceives basic visual attributes like shape, edges, relative size, patches of color.
- These visual attributes are also referred to as **Pre-Attentive** attributes.
- It means you don’t have to think hard to do it.
- If you know what the brain pre-attentively processes, you can use that to make important things in your visualizations pop out.

Visual Working Memory

- Iconic memory's info is passed to visual working memory
- Visual working memory is also a short-term storage (**stores about 5 +/- 2 things at a time**).
- Long-term memory kicks in to associate things in short-term memory to enable comprehension of what you are seeing.

The Goal In Producing a Good Visualization is to…

- take advantage of pre-attentive visual attributes to make important features in the data pop out.
- not require visual working memory to hold more than 5 +/- 2 things at a time. I.e. you don’t want the user to have to refer to more than 5 encodings to make sense of something.
- associate it with something that the user might be familiar with in the past. E.g. seeing your home relative to the destination on a map.

Pre-Attentively Processed Visual Encodings

- Pre-attentive - visual properties that are processed in sensory memory without our conscious thought.
- It takes our brain less than half a second to process a pre-attentive property of an image.
- 4 basic visual properties that can be defined as pre-attentive:
    - Form
    - Movement
    - Spatial Positioning
    - Color

- www.csc2.ncsu.edu/faculty/healey/PP/

Form

Orientation and collinearity
Curvature
Length
Width
Added marks
Numerosity
Shapes
Size
Spatial grouping

Movement

Flicker

Velocity

Direction

Spatial Positioning

2D positioning – plotting something on a 2D map.

Stereoscopic depth - seeing depth by virtue of the fact that we have 2 eyes.
This is the most rapidly detected preattentive stimulus.

Concave and convex positioning – being able to see whether something is concave or convex.

Relative accuracy of comparison using different basic visual features
(Cleveland and McGill)

Extensive testing concludes that blue-orange divergent provides the most resolving power

blue-orange divergent

BOD - Blue/Orange
YGB-Yellow/Green/Blue
RA-Rainbow
CW-Cool/Warm

VI-Virdis
RG-Red/Green
GP-Grayscale
SAT-Saturation

sciviscolor.org

CW
BOD
VI
RA
YGB

The Ubiquitous Rainbow Color Map

Used for a long time by the scientific community even though it is not great. But people still use it so they can compare against old visualizations.

http://www.kennethmoreland.com/color-advice/

- If data is ordered (numerical), Natural ordering to colors is important. Rainbow is not a natural ordering.
- i.e. Why is green > blue? Why is Yellow > Green?

- Perceptual changes between values should be uniform - yellow region changes faster than green.

- 5% of population have deficiencies in distinguishing colors (e.g. red and green)

Basically if you use only color you are limited to shades of blue, gray and yellow

sciviscolor.org/tools

Colormoves

What it is: An interactive online tool equipped with a selection of pre-made, designer-crafted colormaps, a viewing window for testing colormaps on your own data, and an adjustment window for simple and fast colormap building. The adjustment window allows a user to see where the density of their data lies, thereby providing for easy drag-and-drop colormap allocation, which shows the results on data (in the viewing window) in real-time. Once a user creates a new map, they can export that map in the relevant format for any major vis program.

When to use: Use when you are looking for a quick, easy, and effective way to explore your data interactively.

CCC Tool

What it is: CCC Tool is an online colormap crafting tool for users with significant experience who are looking for advanced control over every aspect of their map. The tool provides a starting set of maps to work off of in addition to several extremely fine-tune metrics that enable a user to see each hue point along a map in three different colorspaces at once. The maps can be exported in any relevant format for vis programs. A user can also import their own starting map and adjust it within CCCTool.

When to use: Use only if you need extremely detailed control over a colormap. If you are interested in adding minute control points and adjusting each point until satisfied, this is the tool for you.

David Johnstone

What it is: A color picker and gradient creation tool for a variety of color spaces that interpolates between discrete hues to create perceptually uniform continuous colormaps.

When to use: Good for interpolating the distance between two specific hues in a perceptually uniform space.

Color Brewer

What it is: ColorBrewer is an online colormapping tool for categorical data with one permanent data set showing counties in the Eastern United States. The tool allows a user to define parameters such as number of data classes, nature of their data, distribution considerations, and context. Rather than uploading their own data, a user can request a colormap based on predetermined needs and test it before applying it to their own. It is also

Colorgorical

What it is: Colorgorical is an online tool used to automatically generate color palettes for categorical scientific visualization based on a set of pre-defined preferences. A user can define parameters such as: perceptual distance in colorspace, name difference, pair preference (how well the colors go together), name uniqueness, and lightness range. The user can also identify specific sections of the colorwheel from which the program should compose a palette, as well as add starting, or "seed" colors for the palette.

Color Loom

What it is: Colorloom is a tool developed by the Sculpting Vis Collaborative and inspired to mimic palette creation in the arts. The tool extracts a selection of hues from images and enables users to create continuous colormaps by dragging these extracted hues into a desired order, all within the same interface. These colormaps can be exported in a variety of formats for use in major visualization software.

When to use: Best for generating non-traditional colormaps semantically tied to specific imagery. For example, a continuous colormap for ocean eddy temperature drawn from images of the ocean.

Chart Chooser Guides

www.data-to-viz.com/#explore
www.data-to-viz.com/poster.html

chart.guide/charts/chart-choosing/

Chart Suggestions—A Thought-Starter

Modified with permision -Doug Hull
blogs.mathworks.com/videos
hull@mathworks.com 2009

www.ExtremePresentation.com
© 2009 A. Abela — a.v.abela@gmail.com

Compare among items: Gas, Electricity, etc.

2 Variables are: Cost and % share

Compare among items:
France, Germany, etc

1 Variable per Item:
Hours

Categories are: Paid work & study, etc.

Compare among items: graphics cards

1 Variable per Item: performance metric

Many items: Many different graphics cards

Sorting can also be useful

Compare among items:
Dolphin & Whales

1 Variable per Item:
Population

Few items: Years 2009 to 2014

Few items so there is space for labels

Compare over time:
January - December

Many Periods:
January - December

Cyclical Data: Avg temperature is cyclic

Compare over time:
January - December

Many Periods:
January - December

Non-Cyclical Data:
Hotel price

Compare over time:
1890 - 1920

Few Periods:
1890 - 1920

Single or Few Categories:
Total and Women

Compare over time:
1992-2012

Few Periods:
1992-2012

Many Categories:
Countries

Relationship between:
Lap time and Power/Weight ratio for different cars

Two Variables:
Time and Power/Weight ratio

Relationship between:
Rent, Income, Population
for different cities

Three Variables:
Rent, Income, Population

Distribution of a single variable:
Grade in class

Few Bins:
5

Distribution of two variables:
Weight and Height

Notice the use of transparency

Distribution of four variables:
Latitude, Longitude,
Earthquake Intensity (height),
Population density (color)

2011 Christchurch earthquake
(New Zealand)

Composition changing over time:
Jan-Dec

Few Periods where only relative differences matter

Composition changing over time: Jan-Dec

Few Periods where relative & absolute differences matter

Composition changing over time:
1850-2000

Many Periods (16) where only relative differences matter

Stacked time series to visualize 150 years of changes in the United States Labor Force

A Composition of a Static Simple Share of Total:

Pie chart of most popular spoken languages

A Composition of a Static Accumulation or Subtraction to a Total:

Inventory audit of t-shirts showing numbers damaged and refurbished.

A Composition of a Static Components of Components:

Essentially a hierarchical stacked column chart.

Tree maps

There’s Much More…

- Information or Statistical Visualization
- Geospatial Visualization
- Network Visualization
- Scientific Visualization
- Medical Visualization
- Visualization Infrastructure

ICS 484

Feeding Your Brain With Beautiful Things

informationisbeautiful.net