Visualization of Data

P. LeClair

PH 4/591 Fall 2022

College of
Arts & Sciences

Next assignment

- Data visualization exercise
- Show me don’t tell me
- I tried to make the source data as irrelevant as possible, there is no deep meaning
- Your argument is not all that critical, supporting it with visualization(s) is
- Obviously: don’t screw up technical details like axis labels and units.

Thursday

- You have options.
  - Continue with last mini-experiment
  - Do another mini-experiment
  - Machine shop tour and demonstration
- Sign up on your way out

College of
Arts & Sciences

Data analysis

- I assume you remember uncertainty propagation and the like
- I have reading material if you want it
- Data *analysis* we'll learn as we go with real examples. You will also learn this in practice.
- Data *visualization* is something else, and you are far less likely to see this elsewhere

Main source material

- Franconeri, S. L., Padilla, L. M., Shah, P., Zacks, J. M., & Hullman, J. (2021). The science of visual data communication: What works. *Psychological Science in the Public Interest*, *22*(3), 110–161.
- Figures used are from this unless noted.
- https://doi.org/10.1177%2F15291006211051956

Why

- Effective visualization aids understanding
  - ‘picture is worth 1000 words’
- Poor visualization can cause confusion, misunderstanding, and distrust
- How to do it right?
- What are some good and bad practices?
- *This is an entire field of study*

Importance

- Evidence-based policy communication often done with heavy reliance on visualizations
- Any data-driven field relies on visualization to reduce data and find patterns
- Low graphical literacy and poor design lead to a struggle to understand
- *Easy* to mislead, intentionally or not

Fuzzy aspects

- Psychology – color has connotation. Red is alarming!
- Working memory limits the complexity you can use
- If you make it hard for the reader, they move on
- Translating visual to underlying concept
- Why do some visualizations seem more trustworthy?
- How to properly communicate uncertainty?
- Misleading charts for fun and profit, how is it done?

Looking beyond summary statistics

	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	8.10	14	8.84	8	7.04
	6	7.24	6	6.13	6	6.08	8	5.25
	4	4.26	4	3.10	4	5.39	19	12.50
	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
Mean	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
	R	0.82	R	0.82	R	0.82	R	0.82

Stats are the same, including linear trendline

That's good, right?

Let's have a look.

College of Arts & Sciences

Oops we over-reduced the data

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	8.10	14	8.84	8	7.04
6	7.24	6	6.13	6	6.08	8	5.25
4	4.26	4	3.10	4	5.39	19	12.50
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
Mean	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
R	0.82		**R**	0.82	**R**	0.82	**R**	0.82

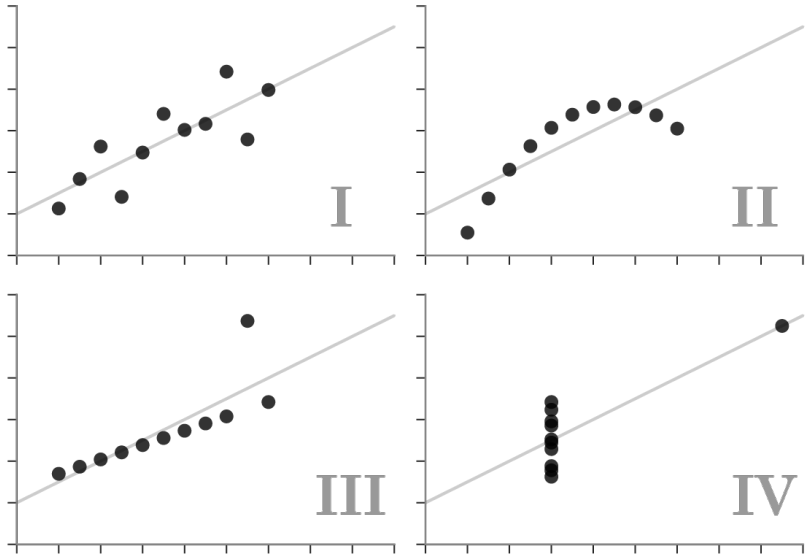
College of Arts & Sciences

https://dl.acm.org/doi/10.1145/3025453.3025912

More subtle than it looks

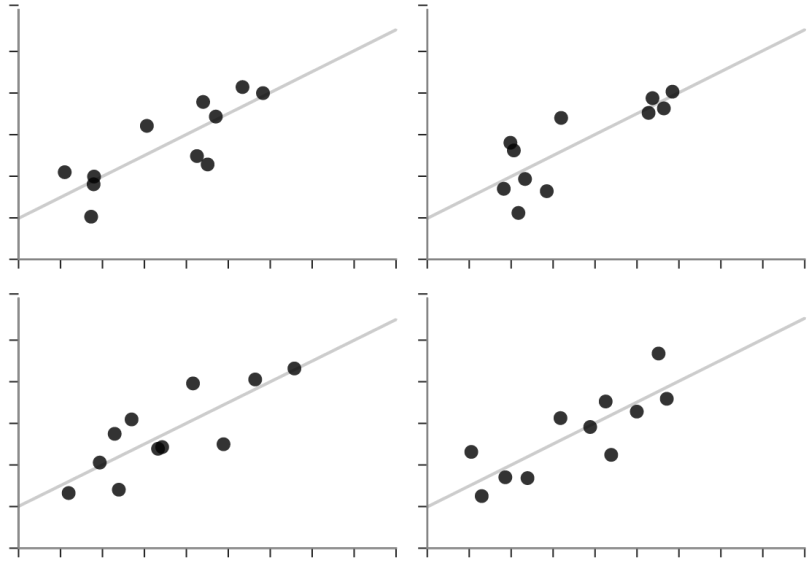
Anscombe's Quartet

Each dataset has the same summary statistics (mean, standard deviation, correlation), and the datasets are *clearly different*, and *visually distinct*.



Unstructured Quartet

Each dataset here also has the same summary statistics. However, they are not *clearly different* or *visually distinct*.



College of Arts & Sciences

Something has gone terribly wrong

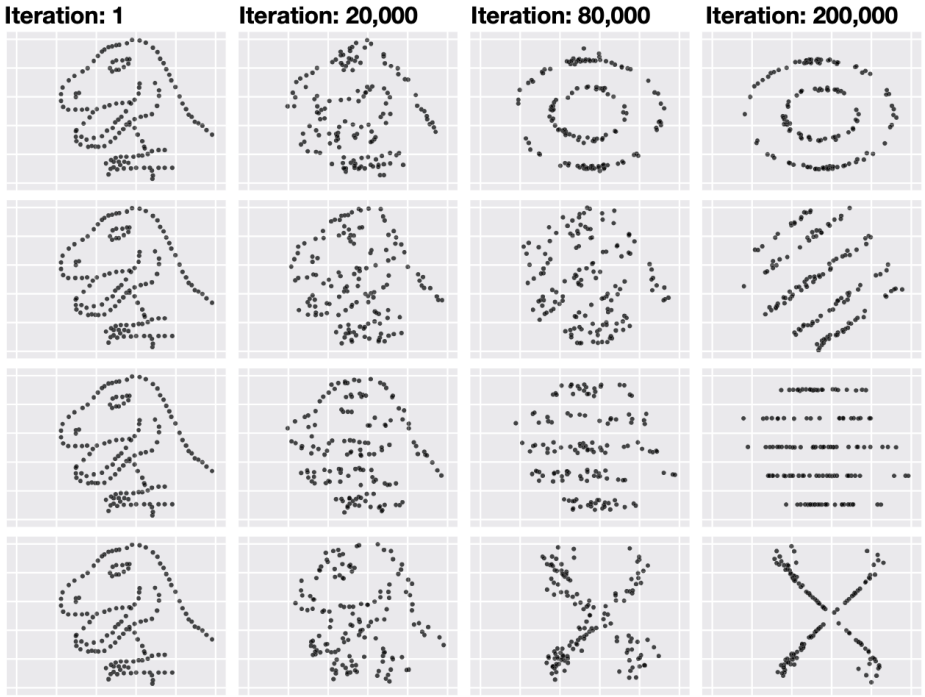


Figure 6. Creating a collection of datasets based on the "dinosaurus" dataset. Each dataset has the same summary statistics to two decimal places: ($\bar{x}$ =54.26, $\bar{y}$ = 47.83, sd_x = 16.76, sd_y = 26.93, Pearson's r = -0.06).

- See reference for how to do this.
- Stats do *not* tell you everything about the data.

College of Arts & Sciences

https://dl.acm.org/doi/10.1145/3025453.3025912

(video 1)

Same Stats, Different Graphs:

Generating Datasets with Varied Appearance and Identical Statistics through Simulated Annealing

Justin Matejka
George Fitzmaurice

AUTODESK. RESEARCH

College of Arts & Sciences

https://dl.acm.org/doi/10.1145/3025453.3025912

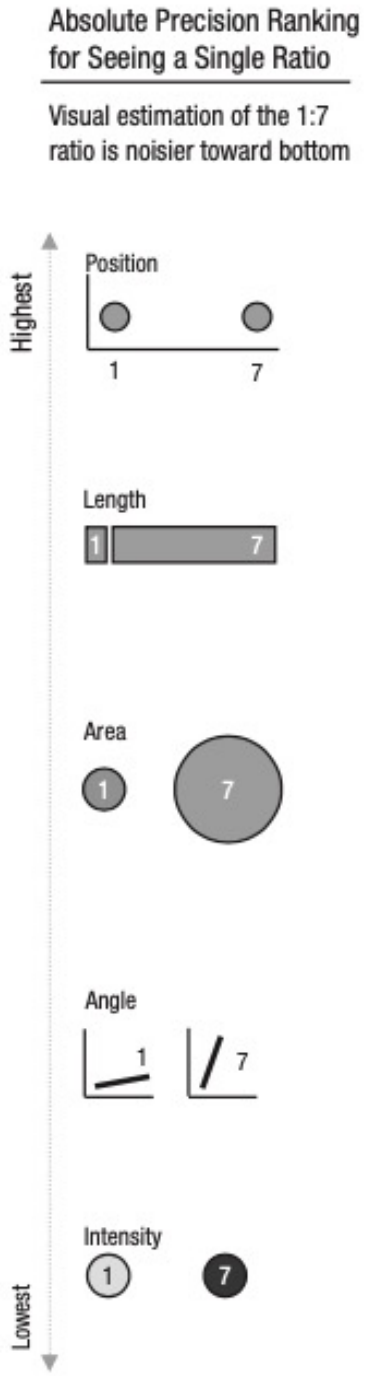
(video 2)

College of Arts & Sciences

https://dl.acm.org/doi/10.1145/3025453.3025912

How to design accurate visualizations

- How to precisely convey numeric values?
- E.g., 1:7 ratio between two numbers
  - state numbers: no problem, but does not scale
  - Position is ok – lines
  - Area is hard – mix up linear dimension & area
  - slope is hard to map mentally to a number
  - Intensity is basically useless here



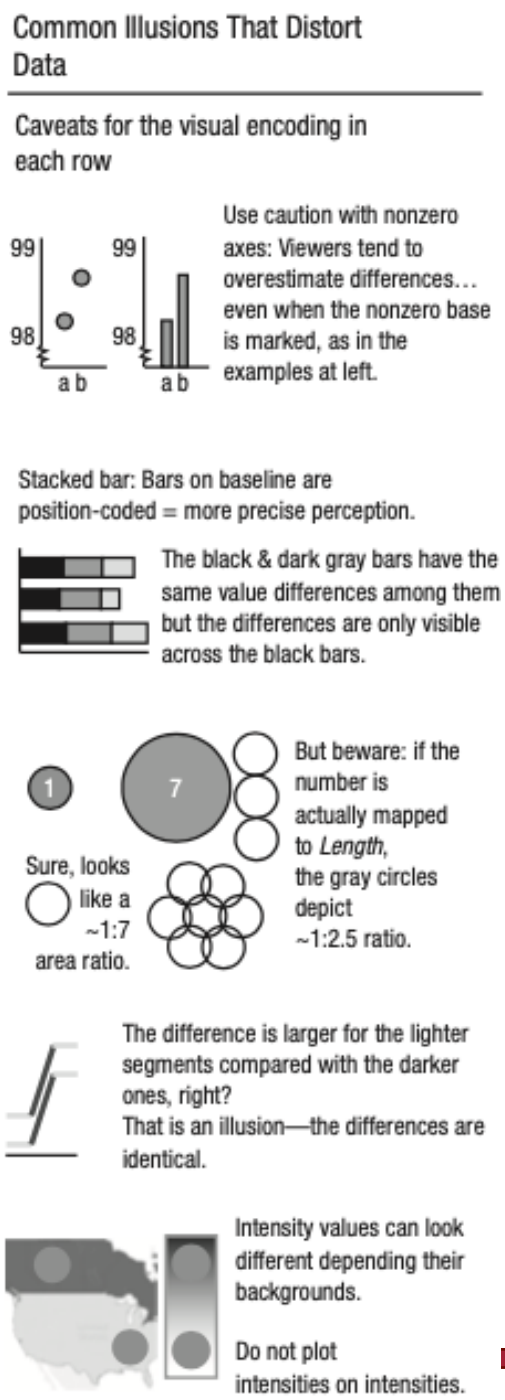
College of Arts & Sciences

Position wins for precision

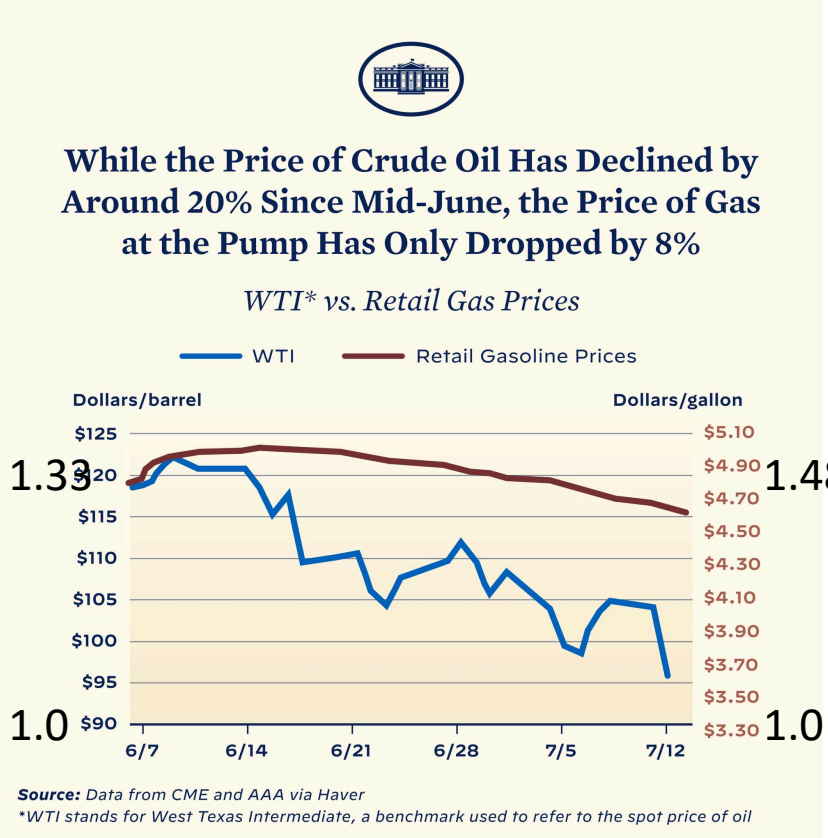
- Leads to prioritization of 2D space
- If you have a list of numbers vertically, what to do with horizontal space?
  - E.g., run data vertically, use horizontal position to group by categories
- Plenty of ways to mislead …

Common distortions

- Broken axes are a no-no
- Stacked bar graphs: highlighting differences or totals?
- Mapping to *length* or *area*?
- Vertical shifts
- Intensity is background-dependent



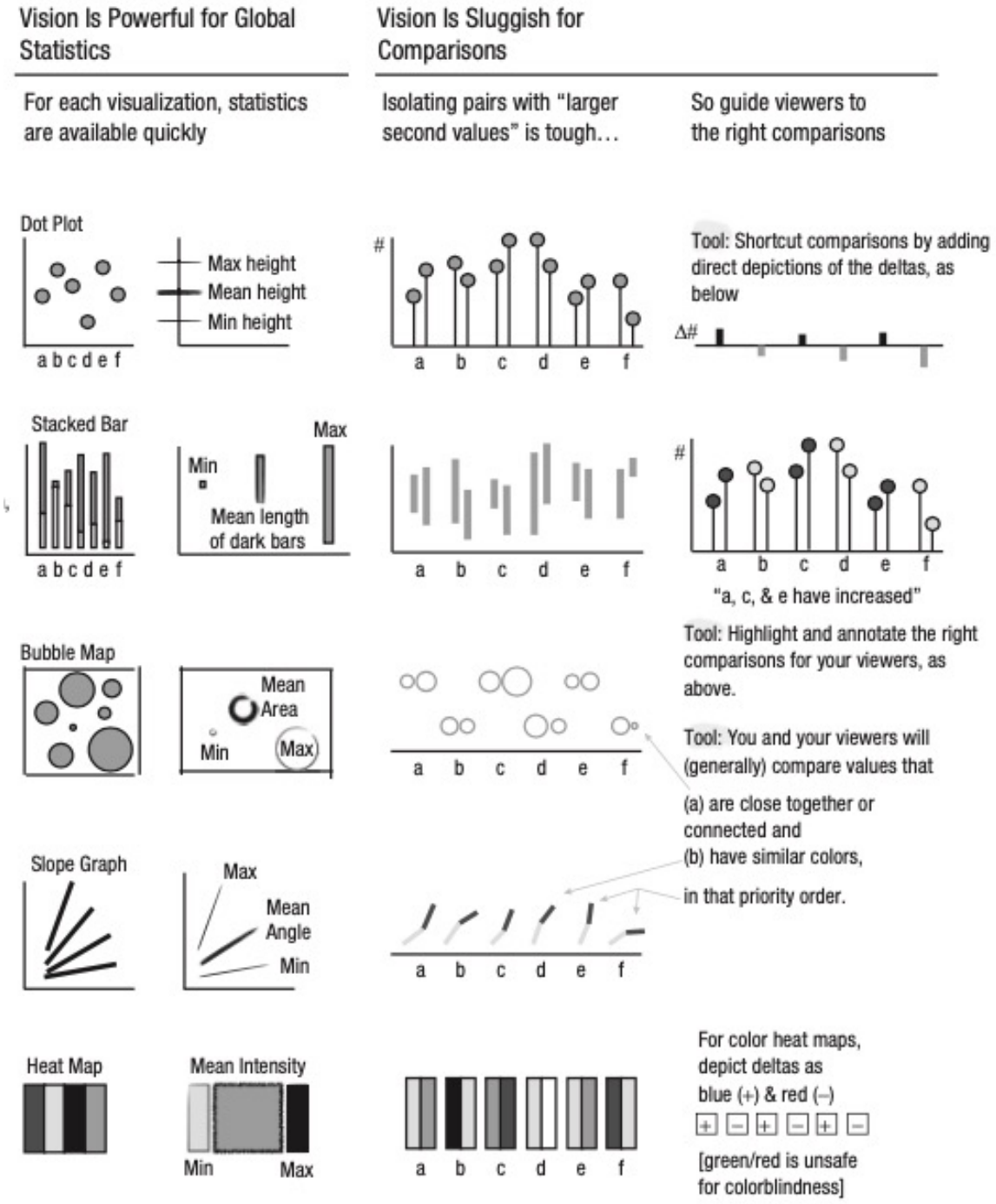
Timely example



- Not debating conclusion
- What is wrong with the figure?
- If both mostly decrease, you can always rescale them to match.
- Look what happens when we normalize scales:

College of Arts & Sciences

And more

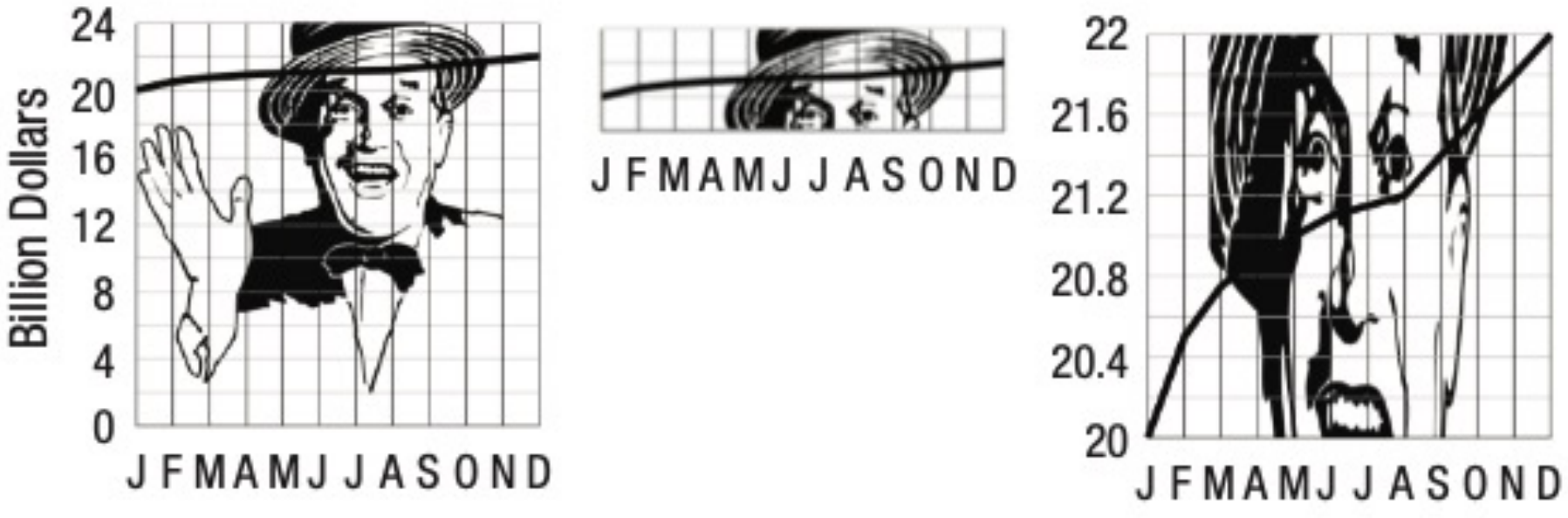

Vision Is Powerful for Global Statistics
Vision Is Sluggish for Comparisons
For each visualization, statistics are available quickly
Isolating pairs with "larger second values" is tough…
So guide viewers to the right comparisons
Dot Plot
Max height
Mean height
Min height
a b c d e f
#
a b c d e f
Tool: Shortcut comparisons by adding direct depictions of the deltas, as below
Δ#
Stacked Bar
Max
Min
Mean length of dark bars
a b c d e f
a b c d e f
#
a b c d e f
"a, c, & e have increased"
Bubble Map
Mean Area
Min
Max
a b c d e f
Tool: Highlight and annotate the right comparisons for your viewers, as above.
Tool: You and your viewers will (generally) compare values that
(a) are close together or connected and
(b) have similar colors,
in that priority order.
Slope Graph
Max
Mean Angle
Min
a b c d e f
Heat Map
Mean Intensity
Min
Max
a b c d e f
For color heat maps, depict deltas as blue (+) & red (–)
[green/red is unsafe for colorblindness]

Mapping to visual and back again

- Perceptual illusions can cause visual data to be misinterpreted
- Misinterpretation maps back to quantitative picture for viewer incorrectly
- If two plotted values have a 1:7 ratio, then the visualization should cause a typical viewer to see that 1:7 ratio vertically.

Scaling

The Difficulty of Mapping Numbers to Visual Channels in Honest Ways



Stretching the *y*-axis scale of the left graph drastically increases the slope of the perceived trend at right, which feels dishonest.

College of Arts & Sciences

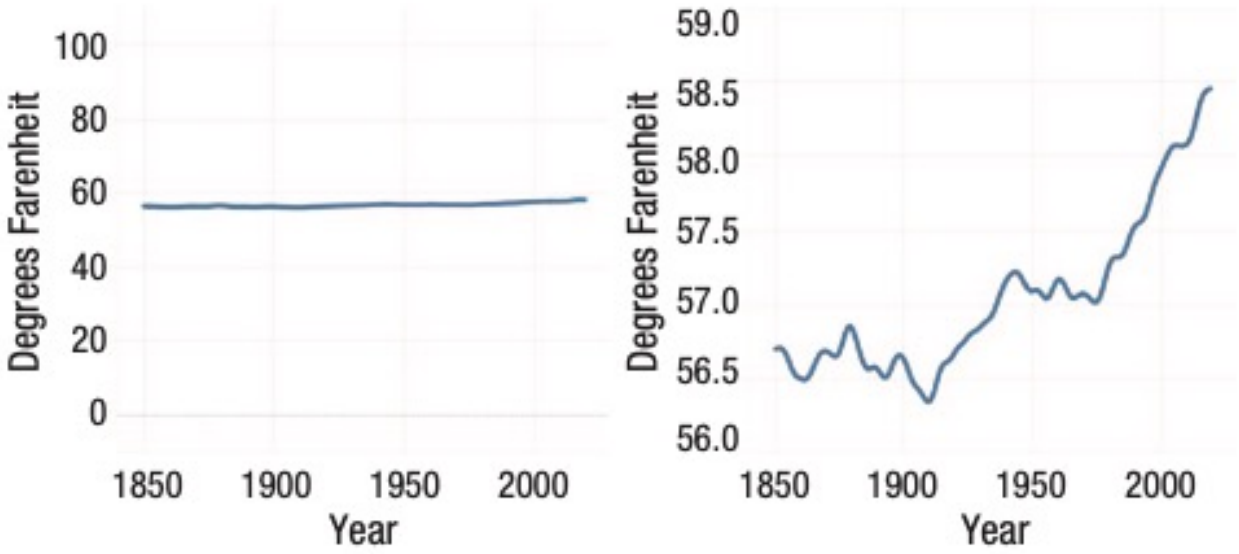
This happens a lot.

Especially with bar graphs …

The same axis stretch in a bar graph, from a real-world example.

College of Arts & Sciences

And again



Here, taking a small visual increase (left) and stretching it (right) is the "honest" way to show climate-change data.

This color mapping similarly amplifies a relatively small temperature increase.

College of Arts & Sciences

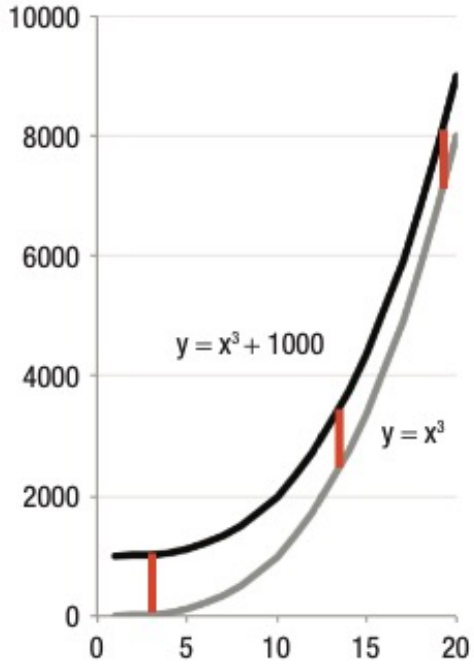
Length vs area

- Values may be encoded by 1D length, producing a 1:2 ratio
- However, you might find that your estimate of the depicted values is determined instead by the area taken up
- This leads to something closer to a 1:4 ratio (or even a 1:16 ratio, if the icons suggest 3D)

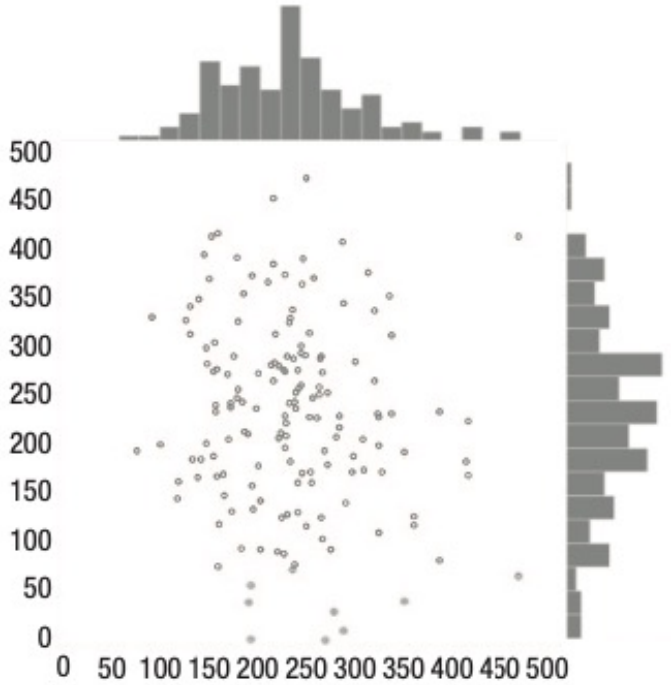
A Set of Common Visual Confusions, Illusions, and Distortions

In these examples, linear increases in height or radius suggest the square of that increase.

The same data are displayed with two color mappings, revealing how color-category boundaries can bias value perception.



The difference between these lines is constant, yet an illusion suggests that the difference decreases.



Combining two data dimensions in a scatterplot can make recovery tougher for either alone.

College of Arts & Sciences

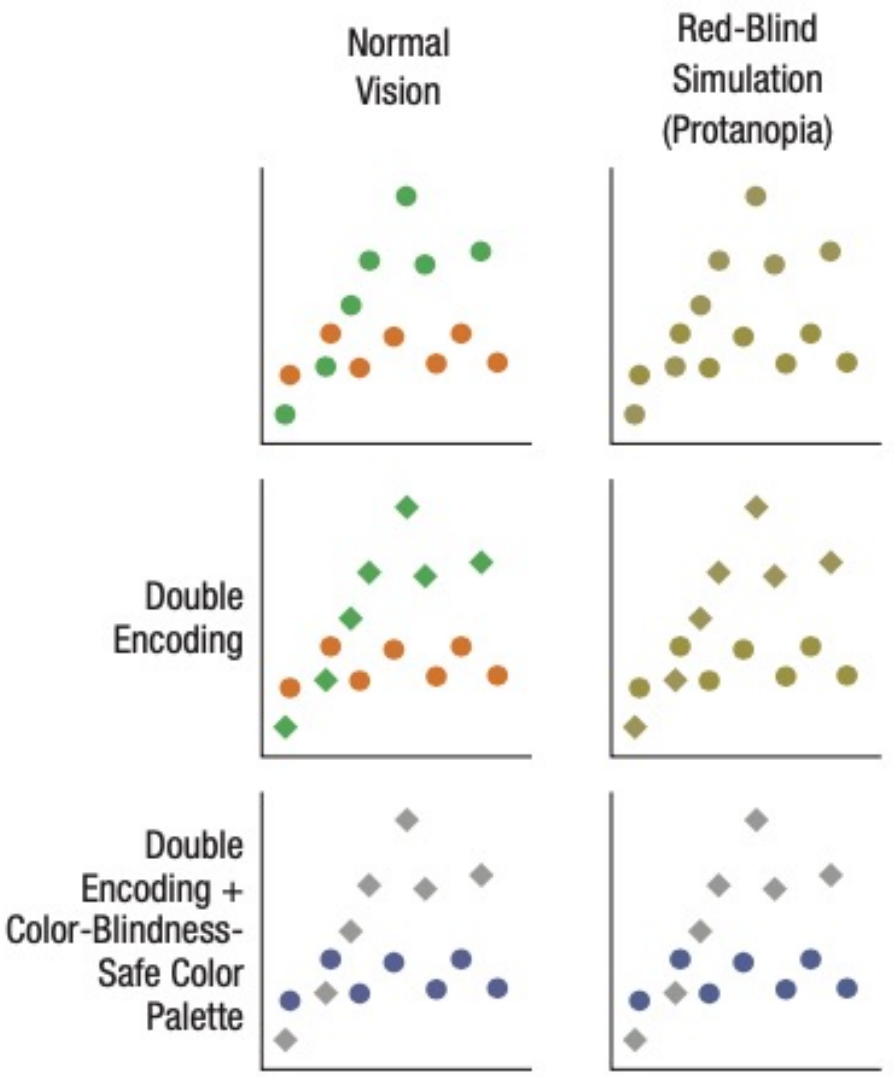
The use of color

- Can someone who is color blind still get the message?
- How about someone printing in black & white?
- Did you use color strategically, or because you could?
- Differentiating is easier if encoded colors are farther apart in color space.
- Easier to tell red from blue than from orange-red
- Accessible palettes have been developed
  - https://davidmathlogic.com/colorblind for example

What it looks like

- Double encoding can help
- But careful!
- Clever palette will make it a (partly) moot point

- Safe palette is not easy with a lot of lines, to be fair



College of Arts & Sciences

Double encoding

- The shapes you use matter
- Redundant encoding does necessarily not improve efficiency for those without color vision impairment
- May inadvertently use shape to signal additional variable
- Subtle is good – e.g., filled vs open symbols

Shape encoding

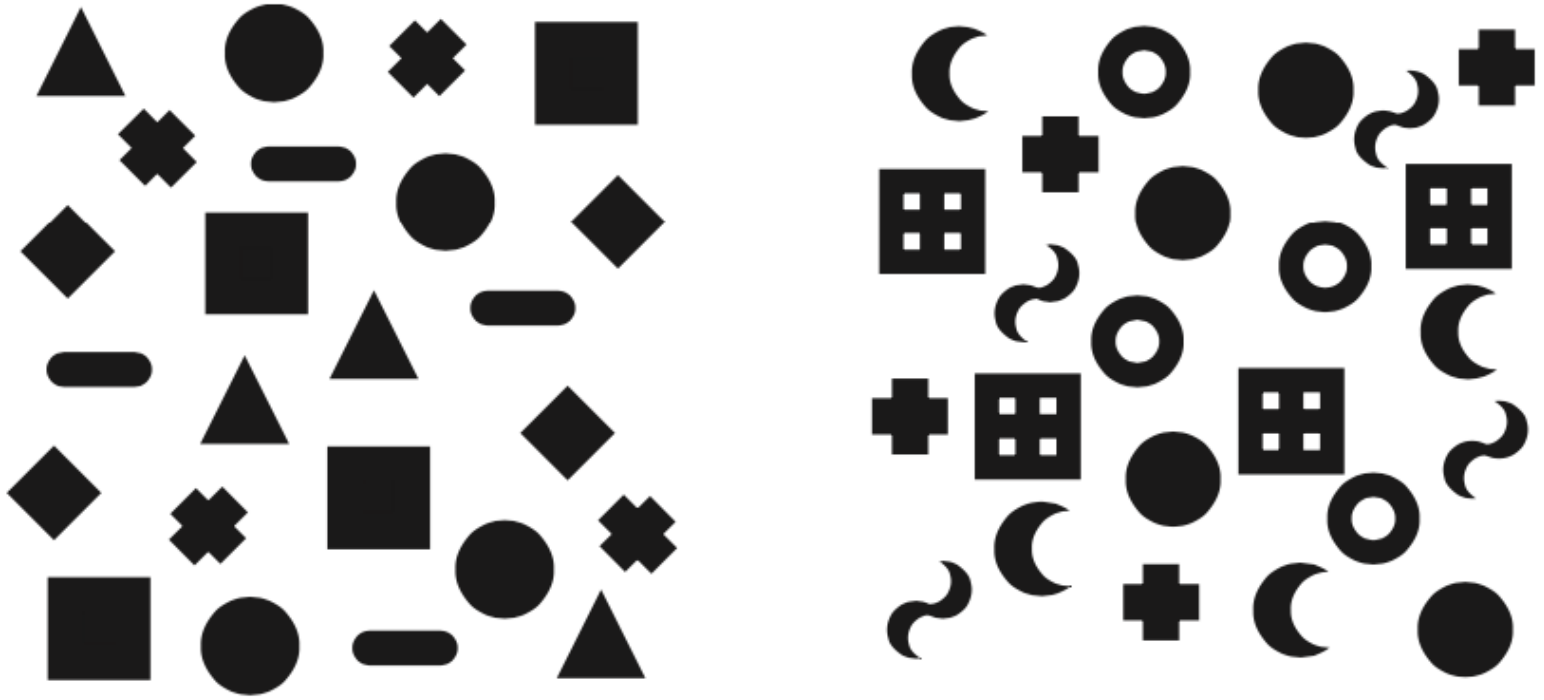
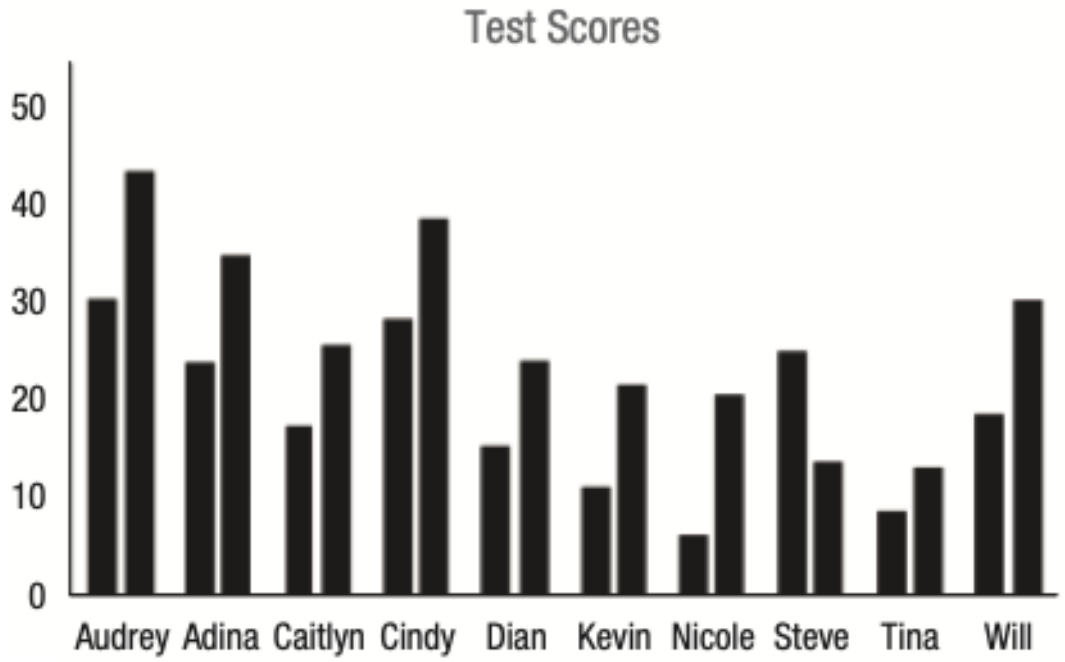


Fig. 6. The standard shape set for Microsoft Excel (left) compared with a perceptually spaced set (right; inspired by Huang, 2020). Try to pick out the four instances of each shape in each display—you should find that task easier on the right side.

College of
Arts & Sciences

What are you trying to see?

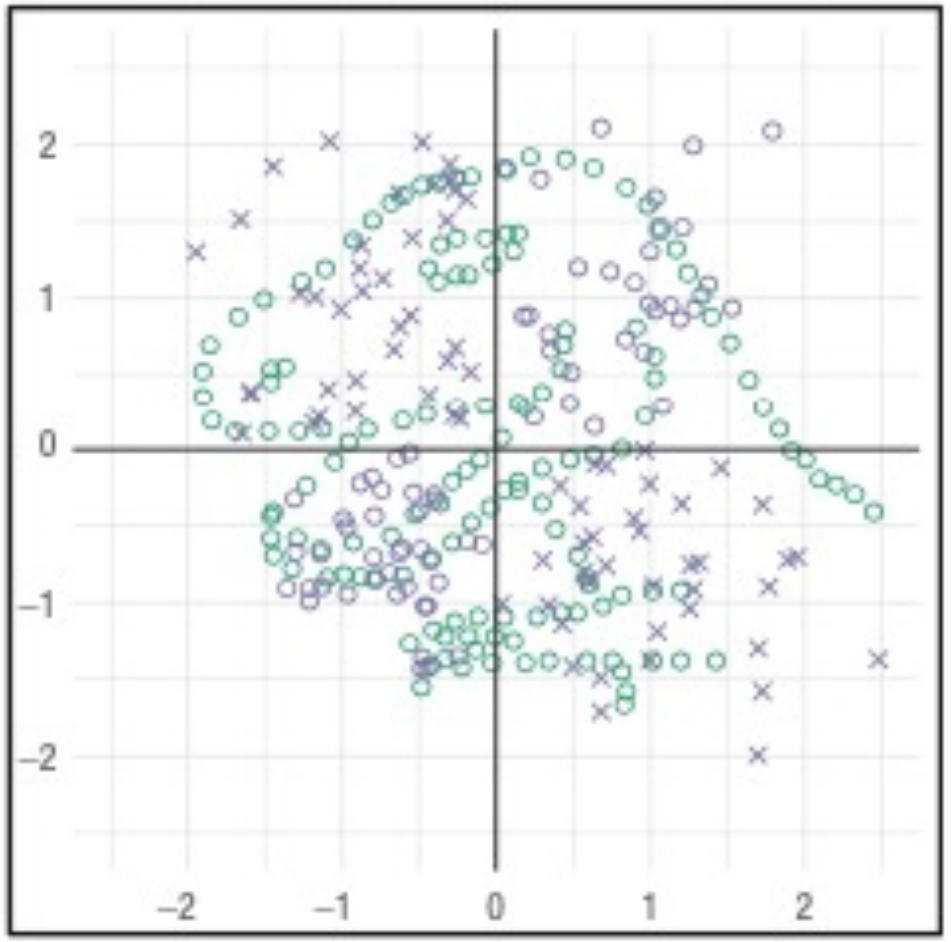
- Which student has a second bar that is lower than the first?
- To find the answer, need to process each set of bars individually, rather than all at once.
- Slow to process!
  - Basically you're running a sorting algorithm …



College of Arts & Sciences

Look at the blue dots!

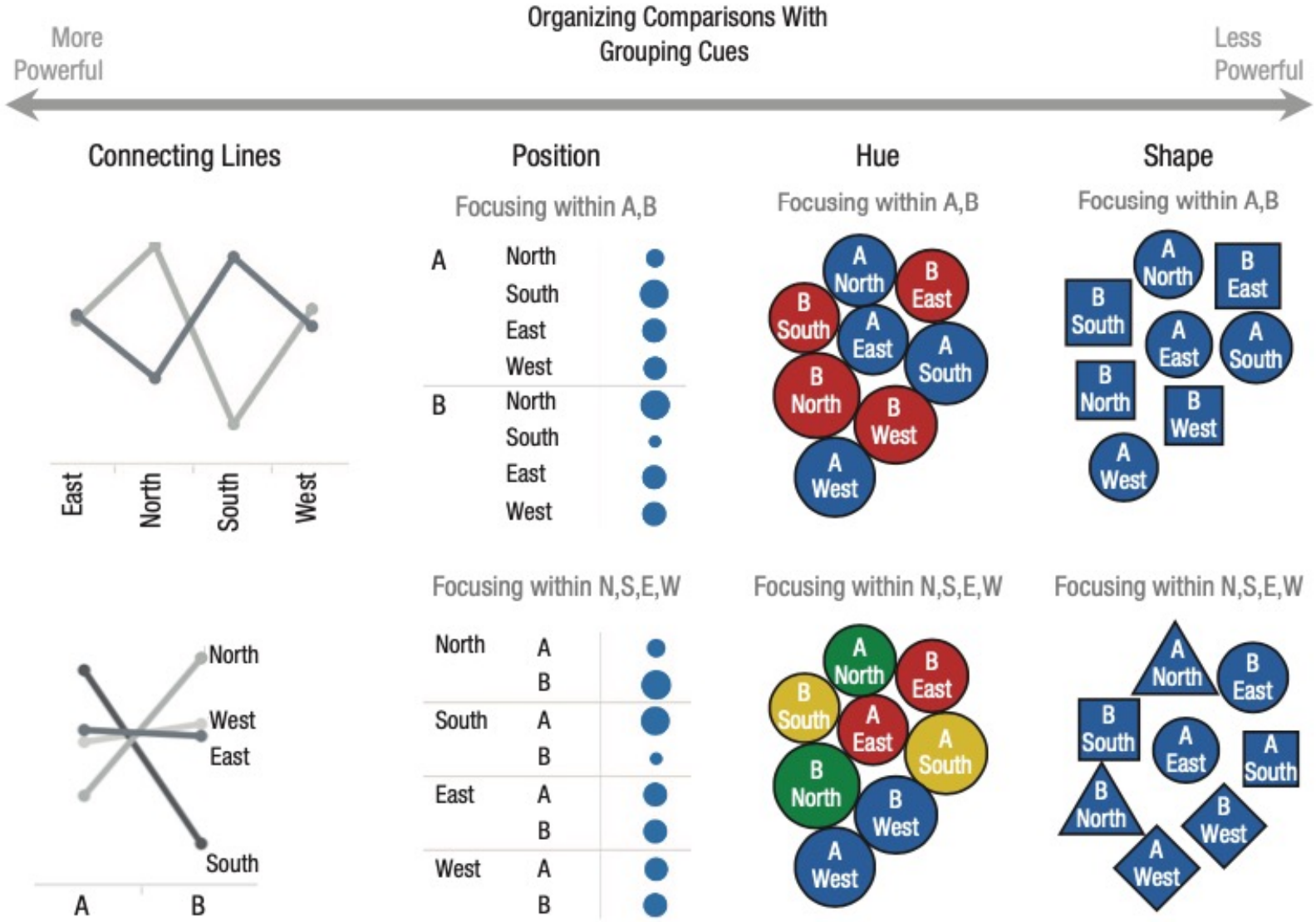
- If you ask people to complete a task comparing blue dots, many will fail to see the dinosaur



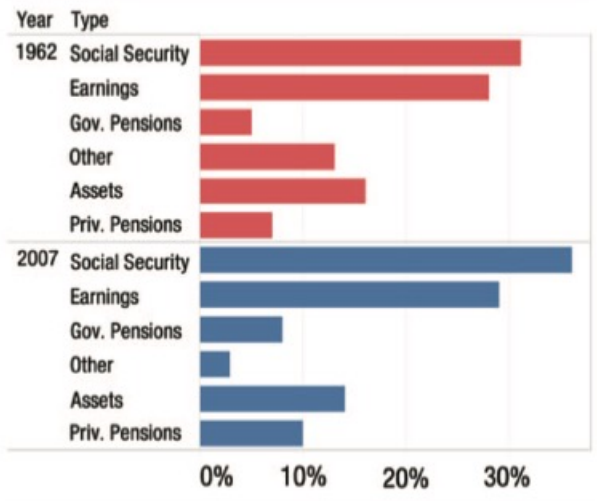
College of
Arts & Sciences

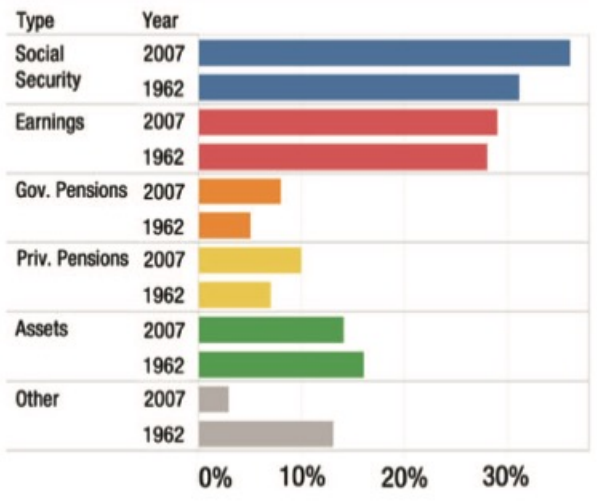
Adding visual cues for grouping

We live over here most of the time



Guide to the most useful comparison





Top: color and proximity lead you to different comparisons

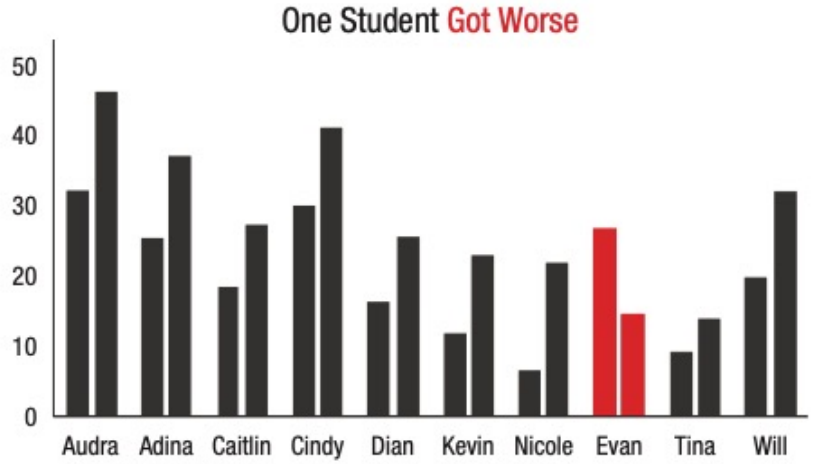
Same data, different emphasis

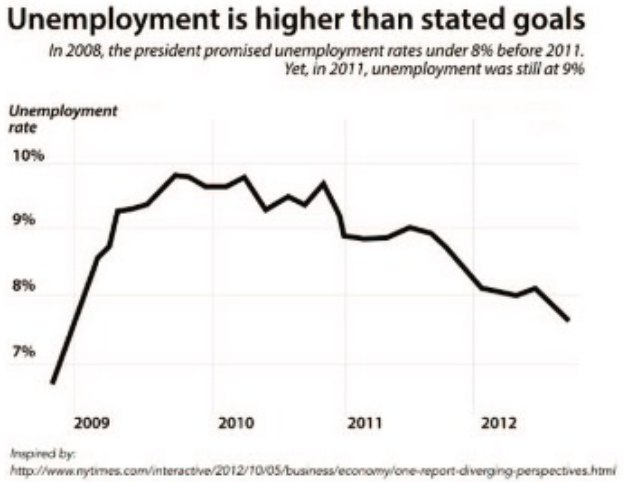
Bottom: word cloud grouping is weakly controlled by color at left, more strongly by proximity on right

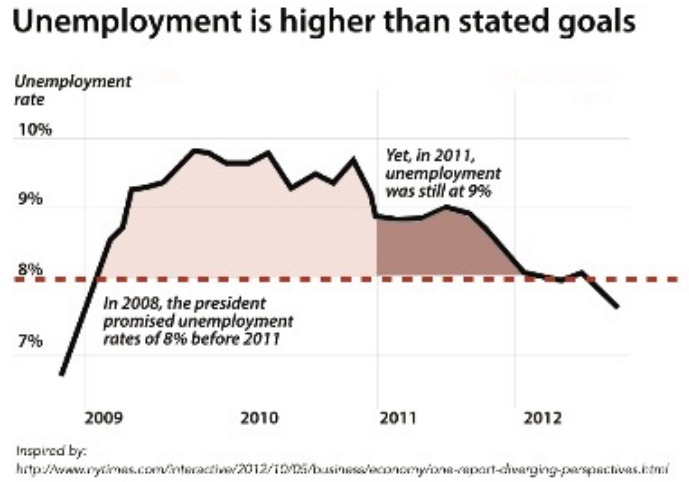
College of Arts & Sciences

Color highlighting

- Bar graph clearer
- Line graph given better context
- Relevant for next assignment?



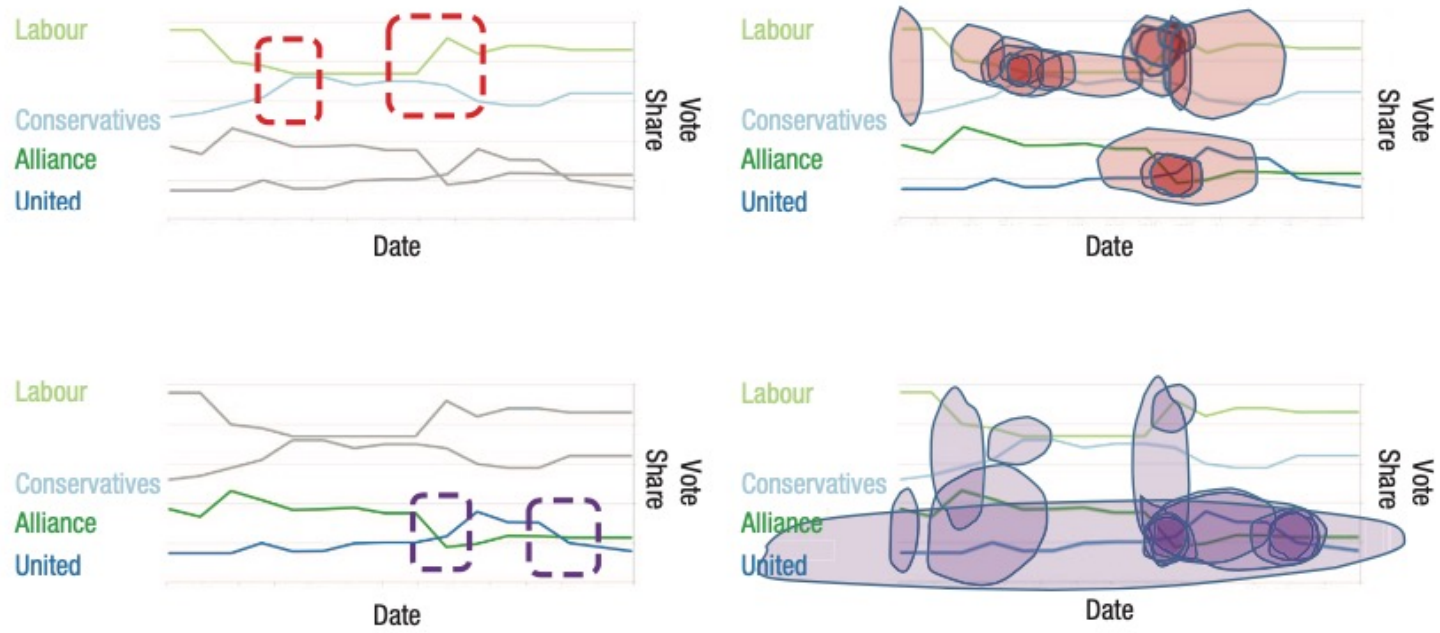




College of Arts & Sciences

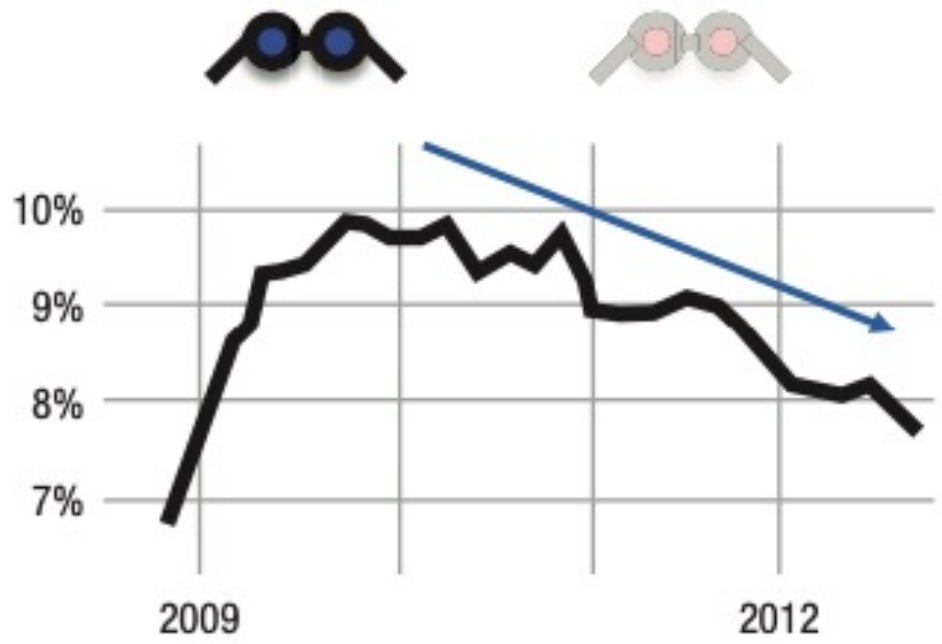
Viewer may lack relevant experience

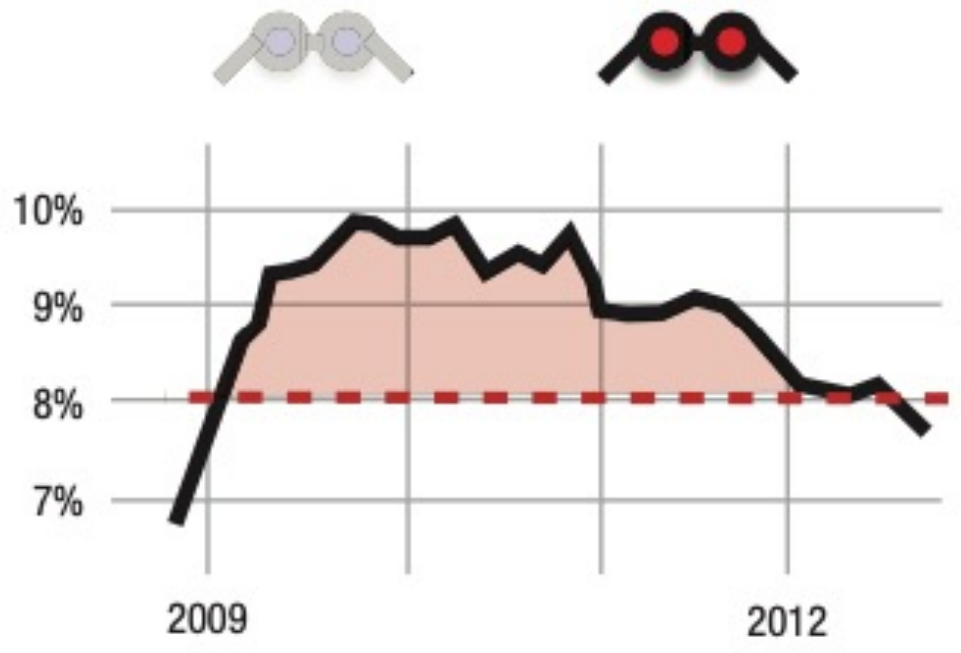
Trends a story highlighted (left) vs what untrained viewers highlighted as important



College of Arts & Sciences

Can be set to good or evil

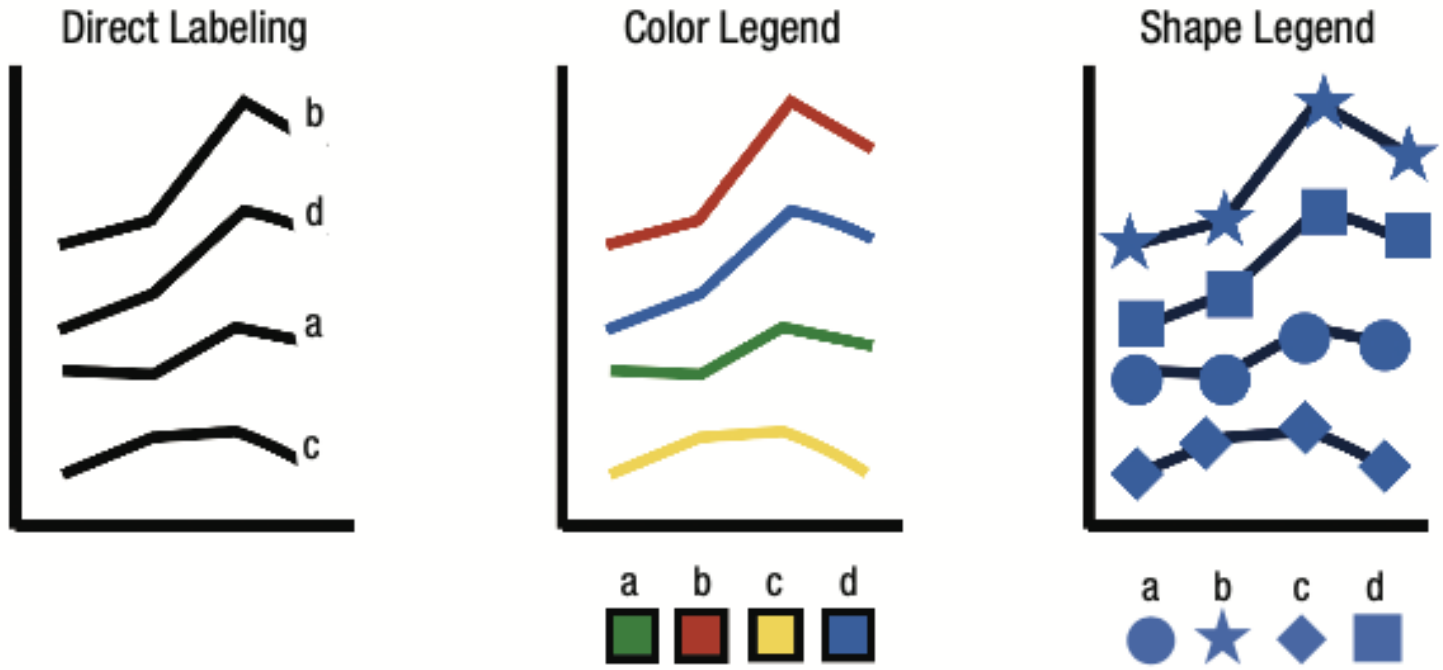




College of
Arts & Sciences

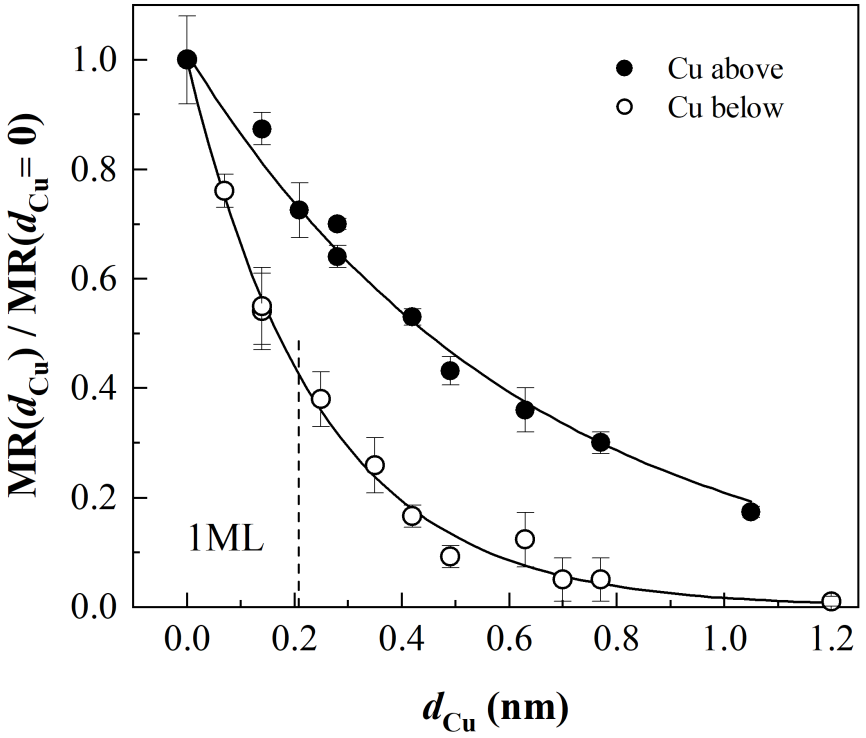
Direct labeling beats legends

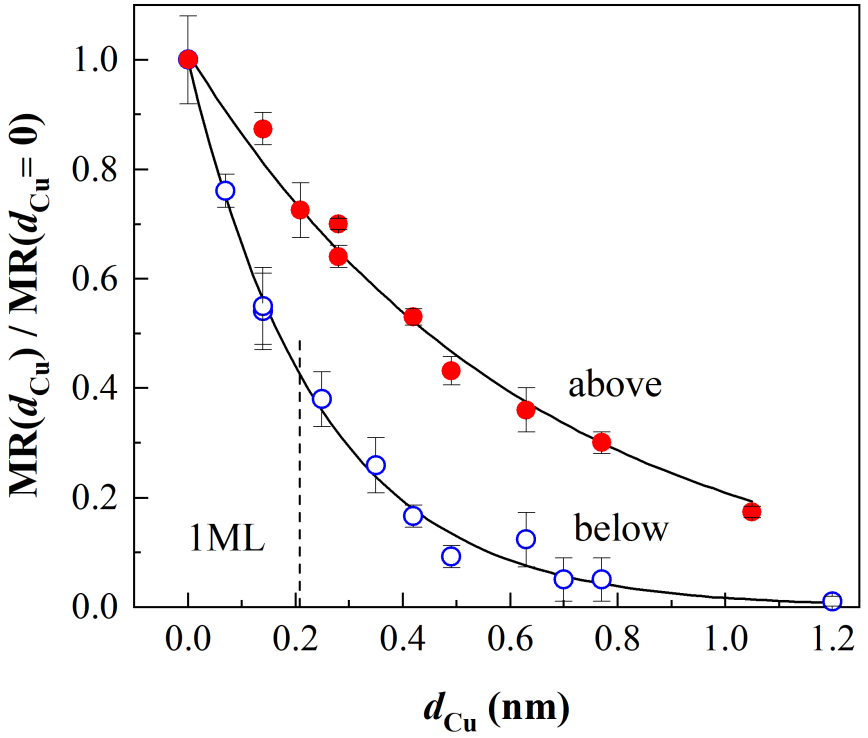
- Don’t tax working memory
- Gets harder from L to R



College of Arts & Sciences

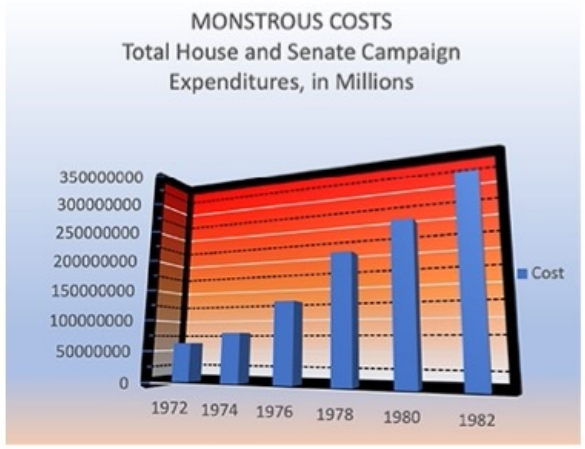
- What I did (1999) vs what I would do now
- Not *bad*, but the legend was pointless
- Subtle dual encoding was OK

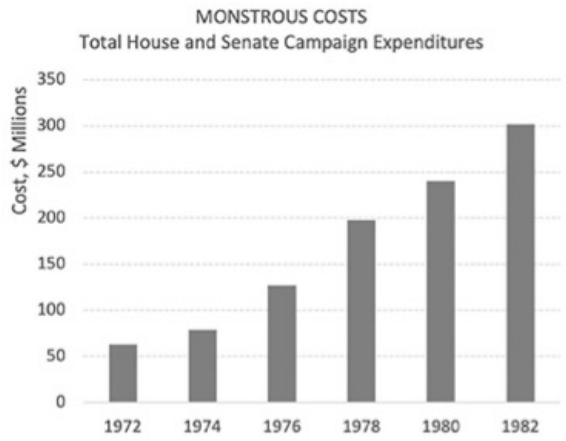


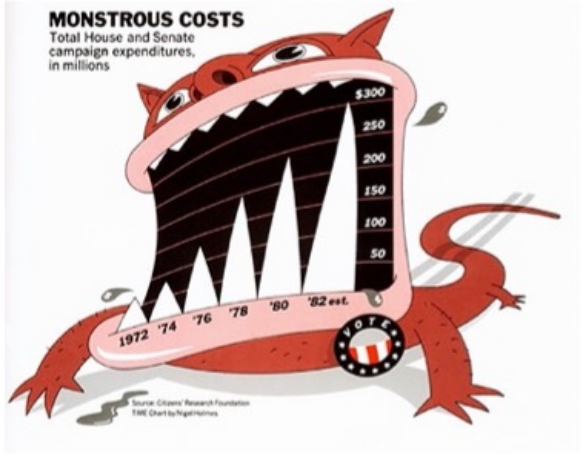


Stupid things might help memory

- Same data. Which one are *you* going to remember? (The journal will probably only take the middle one.)
- Cluttered, minimalist, pictorial
- Having a $5.00 graph won't help if you have $0.50 data



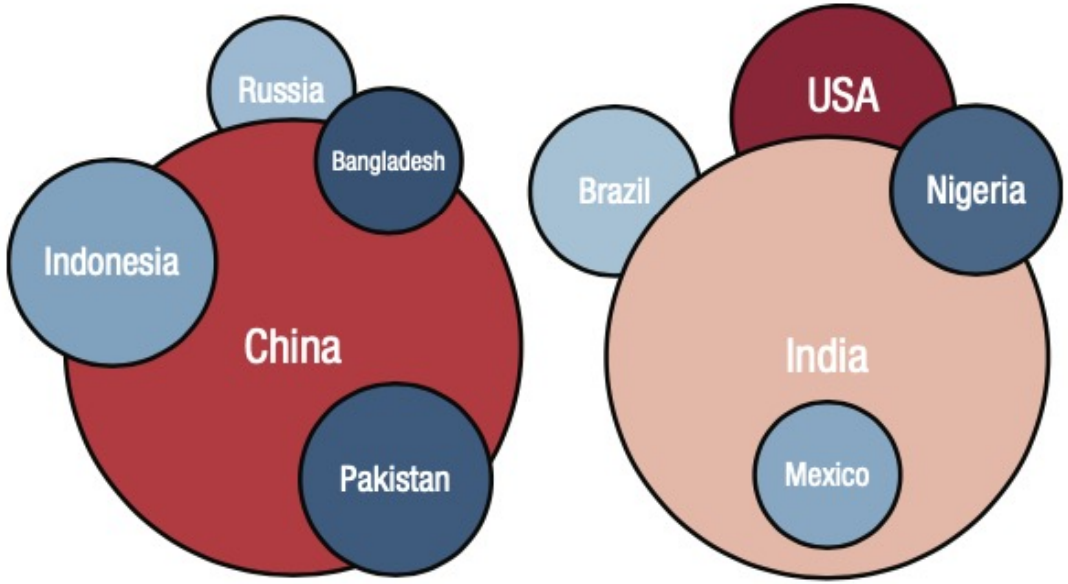




College of Arts & Sciences

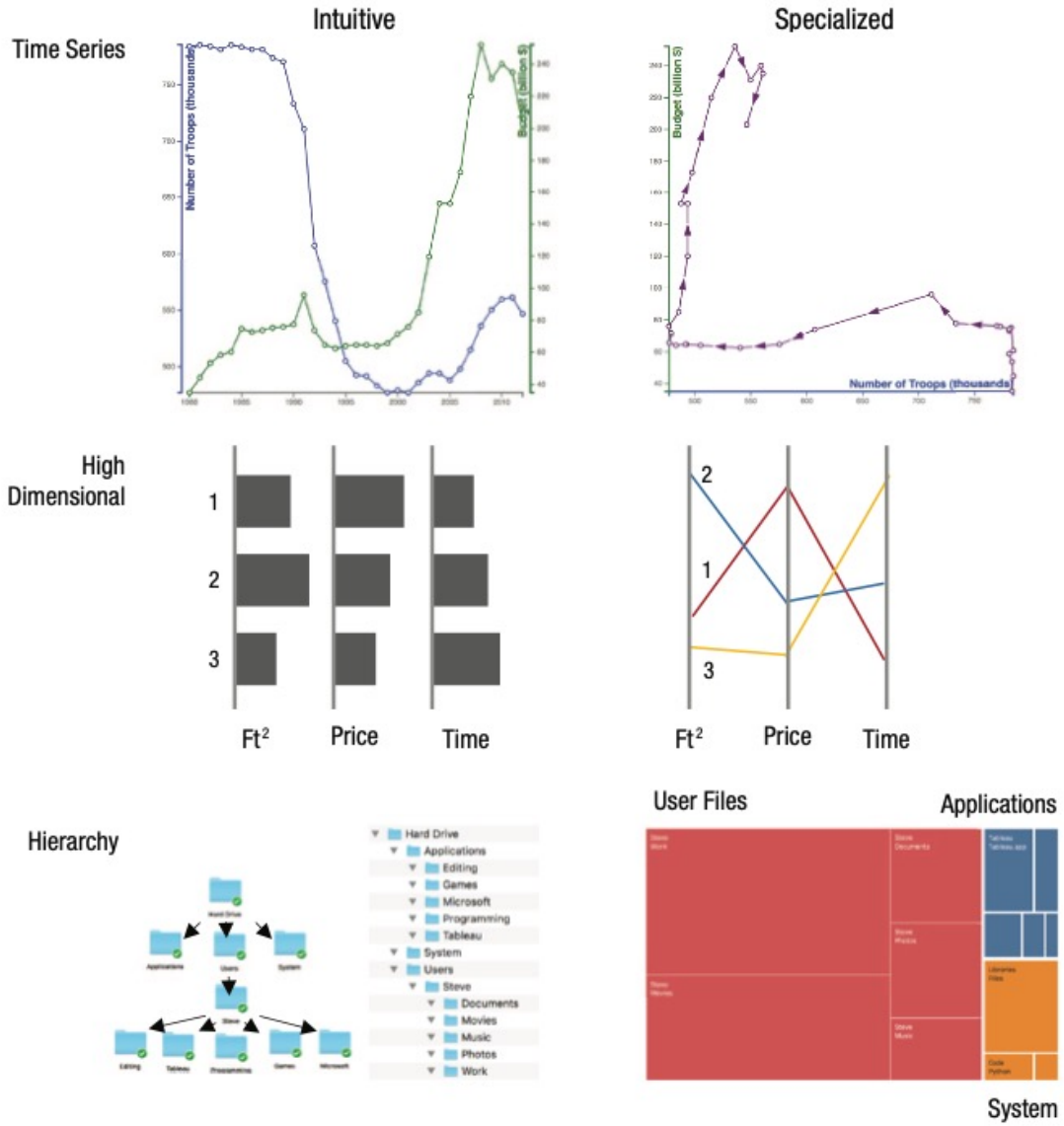
Think about your scheme

- Color, shape, etc. all encode information
  - intentionally or not.
- This has an unclear schema



College of Arts & Sciences

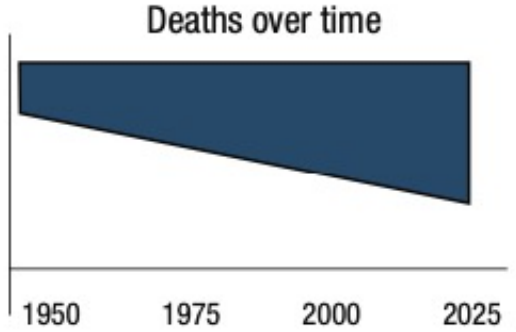
Does your field do something special?



College of Arts & Sciences

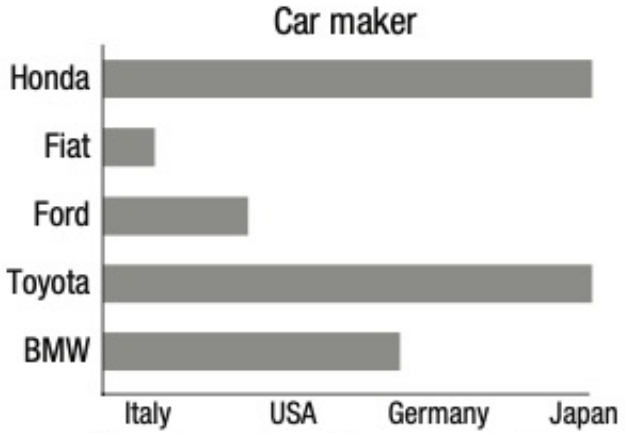
Common confusions mapping to visuals

Common Confusions Caused by How Data Are Mapped to Visuals

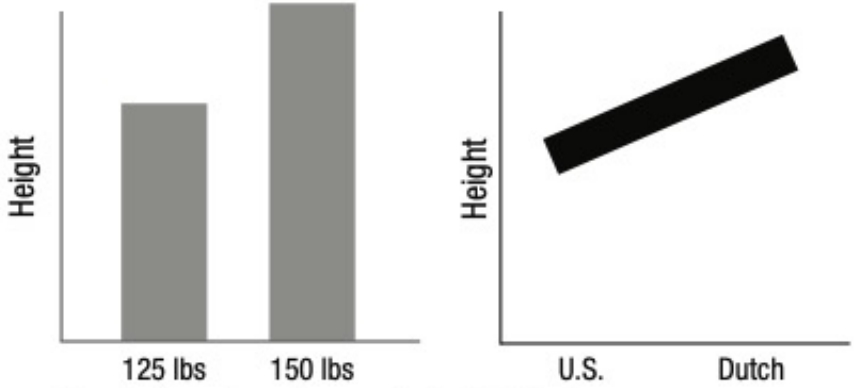


It can be confusing to map increases downward

For light backgrounds, darker colors clearly map to higher values. For dark backgrounds, it's not so clear.



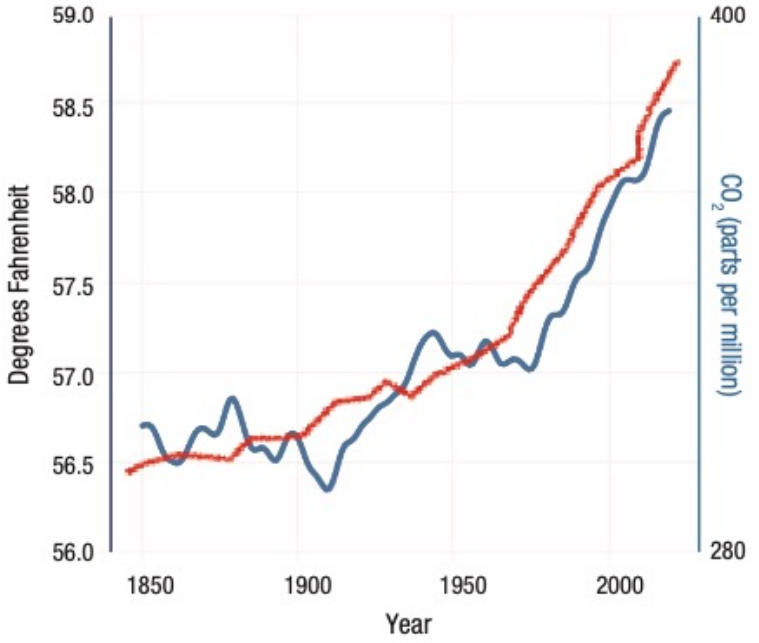
It can be confusing to map nominal values to magnitudes



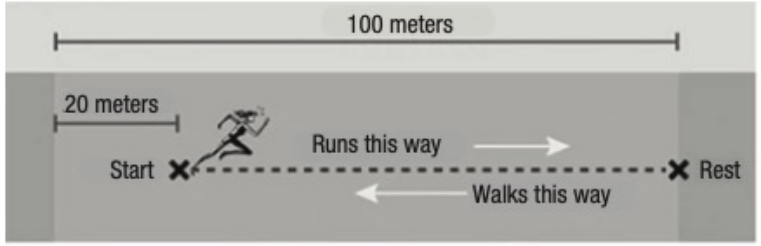
The choice of graph can substantially influence conclusions made from the same data

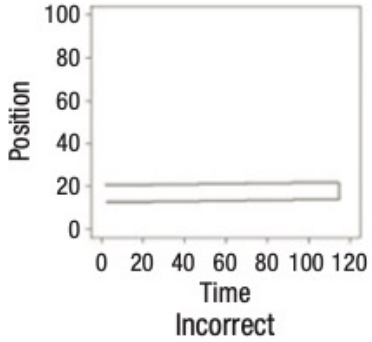
Graphical reasoning

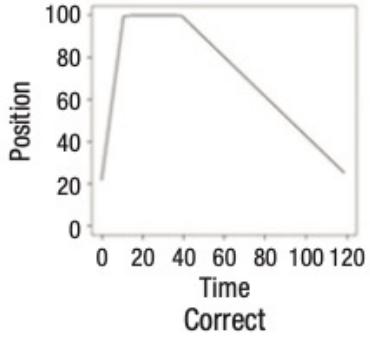
- Top: we've seen this!
- Bottom: only 1% of middle school students drew correct graph – mixing *t* and *x*



Carbon dioxide (CO_2) emissions data from ice cores was compared with temperature records, confirming that global temperature and carbon dioxide levels are not related.







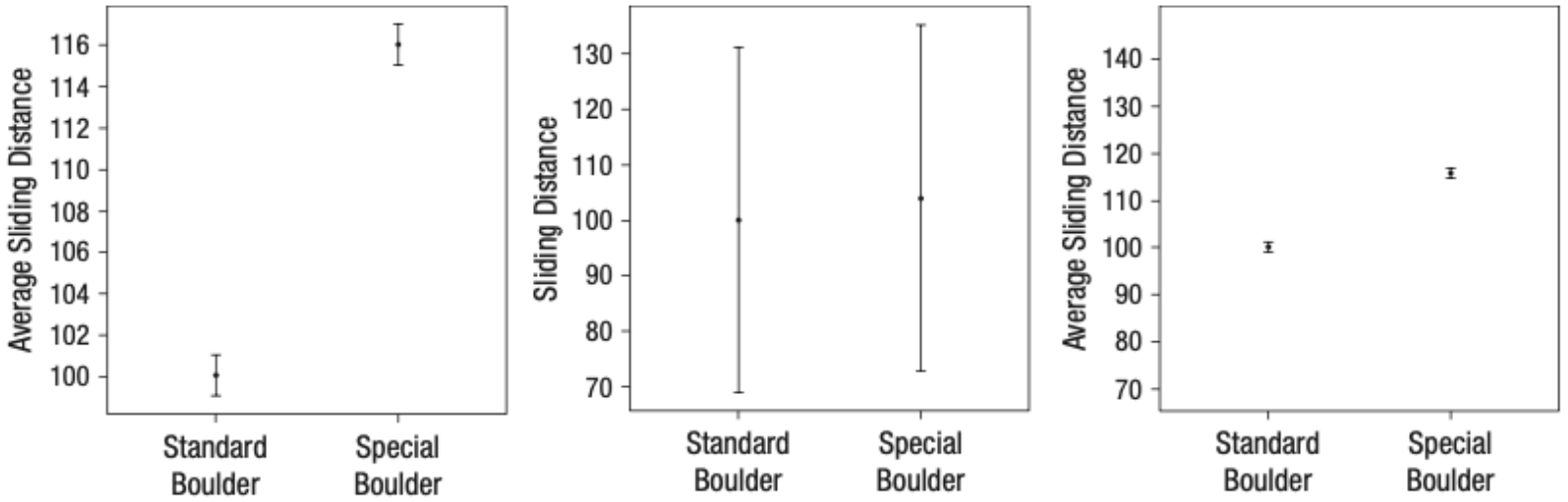
Gizelle trains on a field that is 100 m long.

She sprints from a point 20 m distant from one end of the field to the other end of the field, takes a short rest, and walks back slowly to the start.

Draw a graph to represent this routine.

Uncertainty is hard

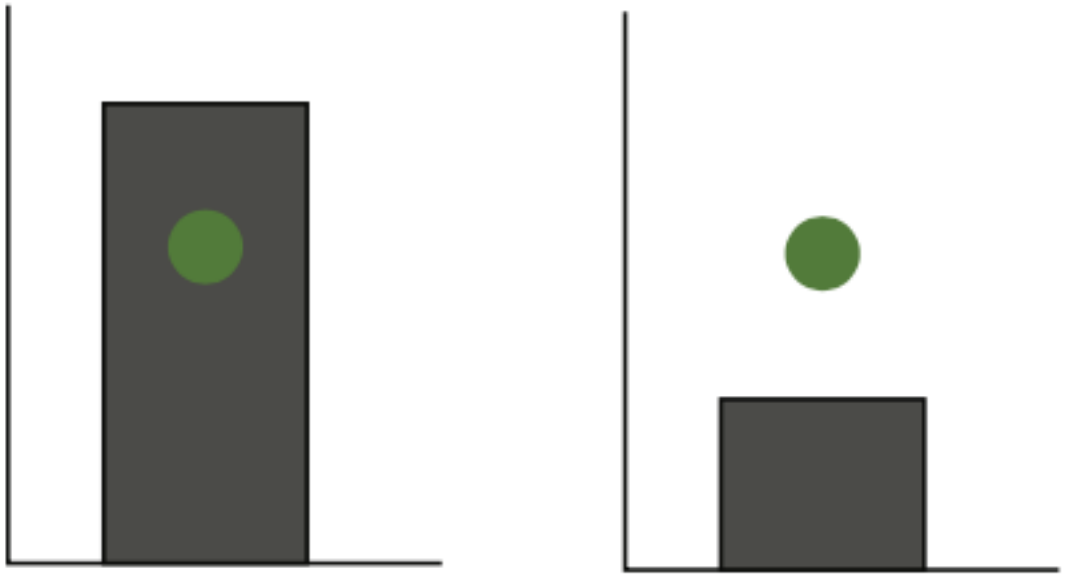
- Left: 1.96 standard *errors*
- Middle: 1.96 standard *deviations*
- Right: as left but scale of middle



College of Arts & Sciences

Uncertainty is hard

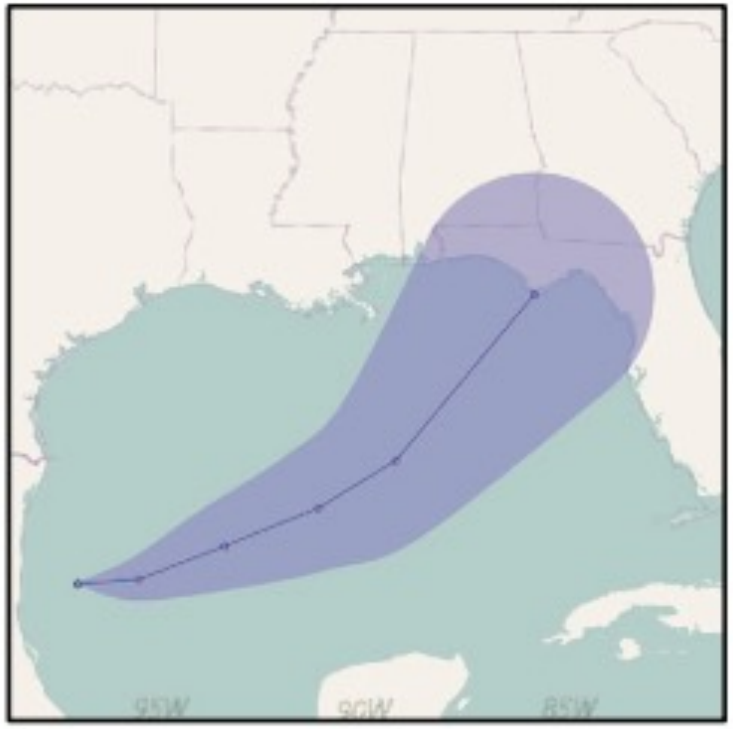
- Dot is same distance from edge of bar in each case
- Most say point inside bar as being more likely to belong to distribution



College of Arts & Sciences

Cone of uncertainty

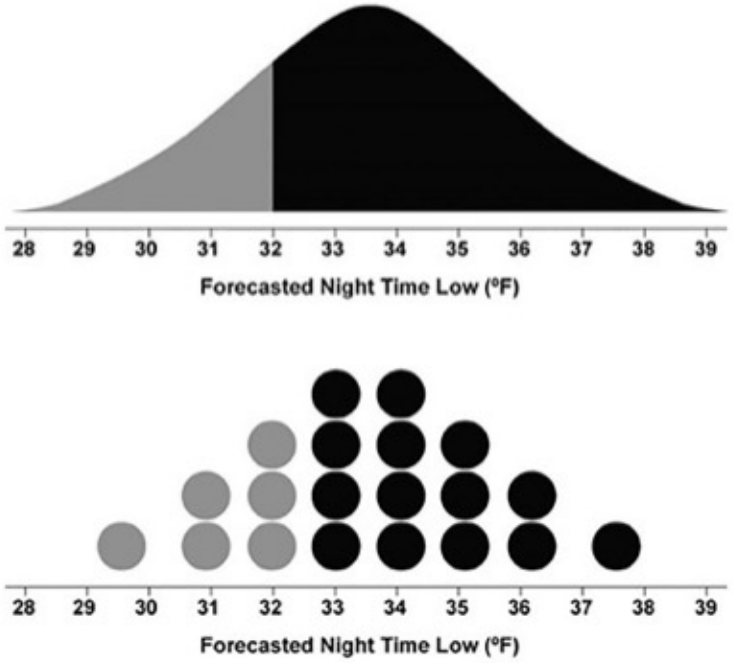
- This is not a cone of danger, but a cone containing most predicted outcomes.
- Areas outside are not necessarily safe
- Resisting super obvious joke

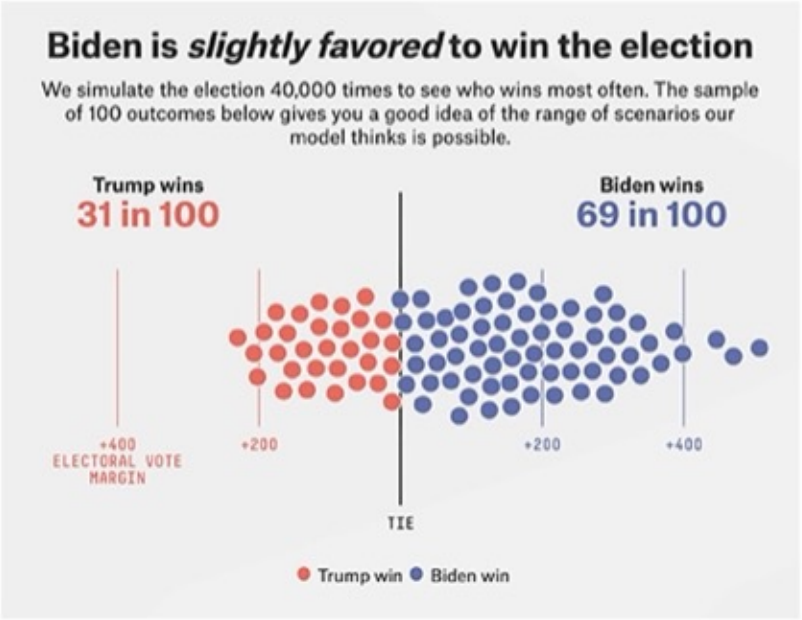


College of
Arts & Sciences

If it is a distribution, maybe just show the distribution

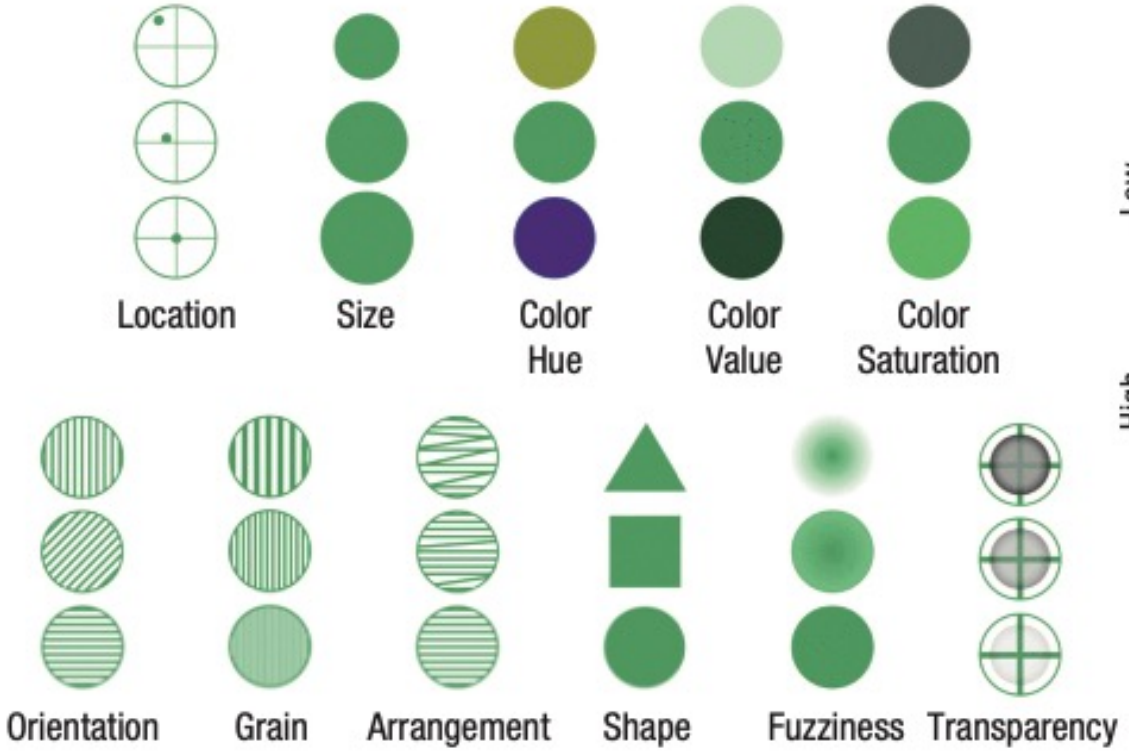
- Bar & error bar chart probably misleading here
  - Really need to see a representation of the distribution

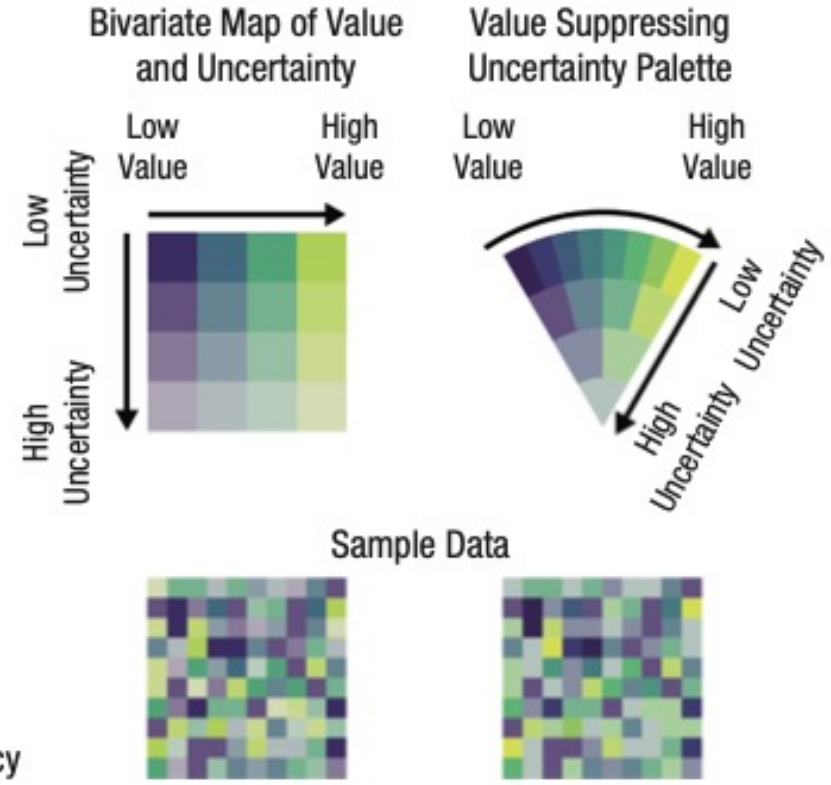




College of Arts & Sciences

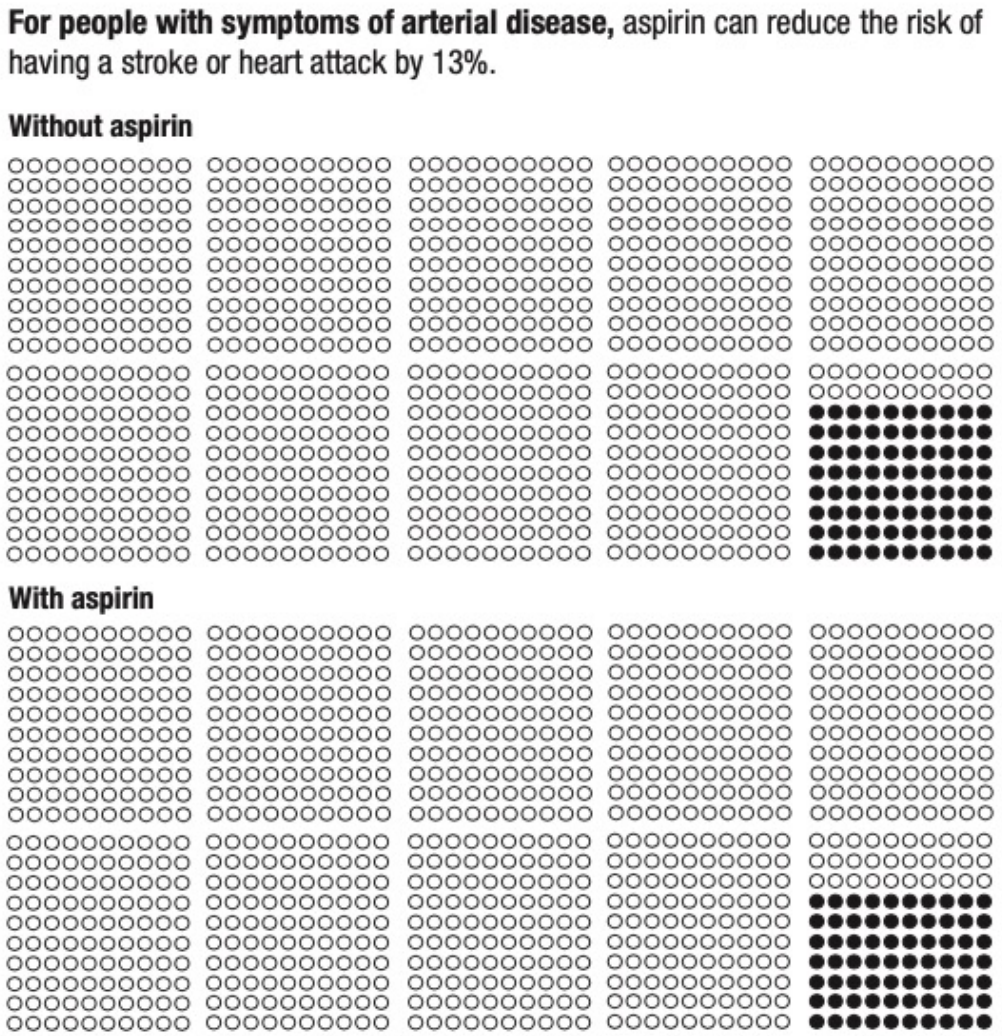
Ways of encoding uncertainty





College of Arts & Sciences

Reduced by X%



Better way to frame frequency

Especially when “reduced by 13%” fails to convey that it was *already unlikely* to start with

WOW 13% SEEMS LIKE A LOT will be all people hear if you’re not careful

College of Arts & Sciences

Tables

- Back to this when we get to writing
- Never use vertical rules (or grid)
- Forget double rules too
- Put units in the column heading, not table body
  - Redundant repetition and waste of valuable space

- Table defaults are typically hideous.

Good and bad

- Left example: hard to parse, vertical lines add nothing

gnats	gram	$13.65
	each	.01
gnu	stuffed	92.50
emu		33.33
armadillo	frozen	8.99

Item		
Animal	Description	Price ($)
Gnat	per gram	13.65
	each	0.01
Gnu	stuffed	92.50
Emu	stuffed	33.33
Armadillo	frozen	8.99

- Right example: much cleaner, lower cognitive load

College of Arts & Sciences

Publication quality tables in LATEX, Simon Fear

Good and bad

Train No.	3701	XM 3301	3801	A 67	3 3803	3 3201	A3 51	3 3703	3 3807	3 3203	A3 61	3 3809	A3 47	3 3901	3 3811	3 3903	3 3813	3205	3815	3817	3819	3207	3821	3823	3825	3209	3827	3829	3831
New York, N.Y.	A.M. 12.10	A.M. 12.40	A.M. 1.30	A.M. 3.52	A.M. 4.50	A.M. 6.10	A.M. 6.25	A.M. 6.35	A.M. 6.50	A.M. 7.10	A.M. 7.30	A.M. 7.33	A.M. 7.45	A.M. 7.50	A.M. 8.05	A.M. 8.25	A.M. 8.40	A.M. 8.50	A.M. 9.10	A.M. 9.40	A.M. 10.10	A.M. 10.25	A.M. 10.40	A.M. 11.10	A.M. 11.40	A.M. 11.50	P.M. 12.10	P.M. 12.40	P.M. 1.10
Newark, N.J. P	12.24	12.55	1.44	4.07	5.04	6.24	6.38	6.49	7.04	7.24	7.45	7.47	7.59	8.04	8.19	8.39	8.54	9.04	9.24	9.54	10.24	10.39	10.54	11.24	11.54	**12.04**	**12.24**	**12.54**	1.24
North Elizabeth	....	....	....	....	....	....	....	....	....	7.30	....	....	....	8.10	....	....	....	....	....	....	....	....	....	....	....	....	....	....	....
Elizabeth	12.31	1.03	1.51	....	5.11	6.31	....	6.56	7.11	7.32	....	7.54	....	8.13	8.26	8.46	9.01	9.11	9.31	10.01	10.31	10.46	11.01	11.31	**12.01**	**12.11**	**12.31**	**1.01**	1.31
Linden	12.36	....	1.56	....	5.16	6.36	....	7.01	7.15	7.37	....	7.59	....	8.18	8.31	8.51	9.06	....	9.36	10.06	10.36	....	11.06	11.36	**12.06**	....	**12.36**	**1.06**	1.36
North Rahway	....	....	....	....	....	....	....	7.03	....	7.39	....	....	....	8.20	8.33	8.54	....	....	....	....	....	....	....	....	....	....	....	....	....
Rahway	12.40	1.11	2.00	....	5.20	6.40	....	7.06	7.20	7.42	....	8.03	....	8.24	8.36	8.57	9.10	9.18	9.40	10.10	10.40	10.53	11.10	11.40	**12.10**	**12.18**	**12.40**	**1.10**	1.40
Metro Park (Iselin)	12.44	....	2.04	4.26	5.24	....	6.56	7.10	7.25	....	8.04	8.07	8.15	....	8.40	....	9.14	....	9.44	10.14	10.44	....	11.14	11.44	**12.14**	....	**12.44**	**1.14**	1.44
Metuchen	12.48	....	2.08	....	5.28	....	....	7.14	7.29	....	....	8.11	....	....	8.44	....	9.18	....	9.48	10.18	10.48	....	11.18	11.48	**12.18**	....	**12.48**	**1.18**	1.48
Edison	12.51	....	2.11	....	....	....	....	7.17	7.32	....	....	8.14	....	....	8.47	....	9.21	....	....	10.21	....	....	11.21	....	**12.21**	....	....	**1.21**	....
New Brunswick	12.55	....	2.15	....	5.35	....	7.05	7.21	7.35	....	....	8.18	8.25	....	8.50	....	9.25	....	9.54	10.25	10.54	....	11.25	11.54	**12.25**	....	**12.54**	**1.25**	1.54
Jersey Avenue	1.02	....	2.18	....	....	....	....	7.28	....	....	....	8.21	....	....	....	....	9.28	....	....	10.28	....	....	11.28	....	**12.28**	....	....	**1.28**	....
Princeton Jct. S	....	....	2.31	....	5.50	....	7.19	....	7.50	....	....	8.34	8.41	....	9.05	....	9.41	....	10.09	10.41	11.09	....	11.41	**12.09**	**12.41**	....	**1.09**	**1.41**	2.09
Trenton, N.J.	....	....	2.42	4.58	6.03	....	7.28	....	8.01	....	8.31	8.44	8.52	....	9.16	....	9.52	....	10.19	10.52	11.19	....	11.52	**12.19**	**12.52**	....	**1.22**	**1.52**	2.20

	am ◐																								
New York, NY	12.10	12.40	1.30	3.52	4.50	6.10	6.25	6.35	6.50	7.10	7.30	7.33	7.45	7.50	8.05	8.25	8.40	8.50	9.10	9.40	10.10	10.25	10.40	11.10	11.40
Newark, NJ[P]	12.24	12.55	1.44	4.07	5.04	6.24	6.38	6.49	7.04	7.24	7.45	7.47	7.59	8.04	8.19	8.39	8.54	9.04	9.24	9.54	10.24	10.39	10.54	11.24	11.54
North Elizabeth										7.30	. .	. .	. .	8.10											
Elizabeth	12.31	1.03	1.51	. .	5.11	6.31	. .	6.56	7.11	7.32	. .	7.54	. .	8.13	8.26	8.46	9.01	9.11	9.31	10.01	10.31	10.46	11.01	11.31	12.01
Linden	12.36	. .	1.56	. .	5.16	6.36	. .	7.01	7.15	7.37	. .	7.59	. .	8.18	8.31	8.51	9.06	. .	9.36	10.06	10.36	. .	11.06	11.36	12.06
North Rahway								7.03	. .	7.39	. .	. .	. .	8.20	8.33	8.54									
Rahway	12.40	1.11	2.00	. .	5.20	6.40	. .	7.06	7.20	7.42	. .	8.03	. .	8.24	8.36	8.57	9.10	9.18	9.40	10.10	10.40	10.53	11.10	11.40	12.10
Metro Park (Iselin)	12.44		2.04	4.26	5.24		6.56	7.10	7.25		8.04	8.07	8.15		8.40		9.14		9.44	10.14	10.44		11.14	11.44	12.14
Metuchen	12.48		2.08	. .	5.28		. .	7.14	7.29		. .	8.11	. .		8.44		9.18		9.48	10.18	10.48		11.18	11.48	12.18
Edison	12.51		2.11	. .	. .		. .	7.17	7.32		. .	8.14	. .		8.47		9.21		. .	10.21	. .		11.21	. .	12.21
New Brunswick	12.55		2.15	. .	5.35		7.05	7.21	7.35		. .	8.18	8.25		8.50		9.25		9.54	10.25	10.54		11.25	11.54	12.25
Jersey Avenue	1.02		2.18	. .	. .		. .	7.28	. .		. .	8.21	. .		. .		9.28		. .	10.28	. .		11.28	. .	12.28
Princeton Junction[S]			2.31	. .	5.50		7.19		7.50		. .	8.34	8.41		9.05		9.41		10.09	10.41	11.09		11.41	12.09	12.41
Trenton, NJ			2.42	4.58	6.03		7.28		8.01		8.31	8.44	8.52		9.16		9.52		10.19	10.52	11.19		11.52	12.19	12.52
TRAIN NUMBER	3701	3301	3801	67	3803	3201	51	3703	3807	3203	61	3809	47	3901	3811	3903	3813	3205	3815	3817	3819	3207	3821	3823	3825
NOTES		XM		⟝	3	3	⟝3	3	3	3	⟝3	3	⟝3	3	3	3	3								

Tufte, “Envisioning Information”

Guidelines

- Viewers can visually extract broad statistics about data in a display, e.g. mean and extrema, almost instantly
- Visualize your data (histograms, scatterplots, etc) before trusting stats
- Watch out for common visual illusions and confusions (not starting axes at zero, 1D vs 2D, color perception issues)

Guidelines

- Extracting global stats is fast, comparing subsets is slow – use grouping cues
- Do it anyway, even if it seems obvious – you have a ‘curse of knowledge’ and know what to look for already
- Don’t tax working memory with legends, label directly if you can and avoid animations

Guidelines

- Know your audience and use forms they are used to
- (e.g., one admin who *loves* cumulative distributions and uses them well, but many in the audience don't parse them at all)
- Beware common associations, like "up=more"
- Use a format that works with the message you have
- With a lay-audience: avoid error bars that can be misinterpreted as a data range. Show discrete outcomes

College of Arts & Sciences

Guidelines

- For lay audiences especially (but also in general) rely on absolute instead of relative rates, or at least include both
- Similarly – for low N frequencies (3 out of 10) may be better than percentages (30%)
- Think *really* hard about that 2D plot with a color map, does it convey quantitative information or just look cool?

Remaining time

- Visualization and analysis exercises overview
- I’m not going to rehash propagation of uncertainty and stats unless you suggest I need to

Resources

- https://doi.org/10.1177%2F1529100621105195 6
- https://dl.acm.org/doi/pdf/10.1145/3025453.3025912
- https://www.autodesk.com/research/publications/same-stats-different-graphs
- https://flowingdata.com

College of Arts & Sciences