Data Visualization

CMPSC 110, ERSP

Ziad Matni, Fall 2023

...a, CRs only, and ...proposed
... under fou... ...ssing mechanisms

	Bias[a]		Variance[b]		95% CI[c]	
	$(\hat{\beta}_W)$	$(\hat{\beta}_X)$	$(\hat{\beta}_W)$	$(\hat{\beta}_X)$	$(\hat{\beta}_W)$	$(\hat{\beta}_X)$
(M.1) $P(R=1)=0.66$						
	0.01346	0.02229	0.04008	0.03685	0.955	0.950
	0.03062	−0.003561	0.1149	0.06732	0.960	0.955
...pu	0.01431	0.021	0.04088	0.05169	0.980	0.975
(M.2) logit $P(R=1)=2Y$						
Full	0.007908	−0.02116	0.03838	0.03624	0.975	0.925
Comp	0.01945	0.07096	0.107	0.06581	0.960	0.950
Impu	0.006966	0.01597	0.04227	0.05226	0.975	0.985
(M.3) logit $P(R=1)=2X$						
Full	0.007908	−0.02116	0.03838	0.03624	0.975	0.925
Comp	0.01225	0.0589	0.08856	0.06818	0.980	0.975
...pu	0.009563	−0.04699	0.03865	0.04923	0.985	0.970
(M.4) logit $P(R=1)=X+Y$						
Full	0.01346	0.02229	0.04008	0.03685	0.955	0.950
Comp	0.02404	1.613	0.1102	0.08202	0.955	0.580
Impu	0.01814	0.08289	0.0578	0.06075	0.955	0.970

[a]Bias = $(\hat{\beta}-\beta_0)/\beta_0$.
...imulation variance.
...e interval using jackknife standard error.

Table 5

...using full dat... ...CRs only, and proposed
...hanisms

Simulation results for u...
method under ...

	Bias[a]		
Method	$(\hat{\beta}_W)$	$(\hat{\beta}_X)$	$(\hat{\beta}$
(M.1) $P(R=1)$			
Full	0.01346	0.02229	0.04
Comp	0.03062	−0.003561	0.11
Impu	0.01431	0.021	0.0
(M.2) logit $P(R=$			
Full	0.007908	−0.02116	0.038
Comp	0.01945	0.07096	0.107
Impu	0.006966	0.01597	0.04227 0.052
(M.3) logit $P(R=1)=2X$			
Full	0.007908	−0.02116	0.03838 0.03624 0.97
Comp	0.01225	0.0589	0.08856 0.06818 0.9
Impu	0.009563	−0.04699	0.03865 0.04
(M.4) logit $P(R=$			
Full	0.01346	0.02229	
Comp	0.02404	1.61	
Impu	0.01814	0.0	

[a]Bias = $(\hat{\beta}-\beta_0)/\beta_0$.
[b]Simulation variance
[c]Confidence interv

Tell us a story, Daddy!!!

Tables That Hurt

Table 5

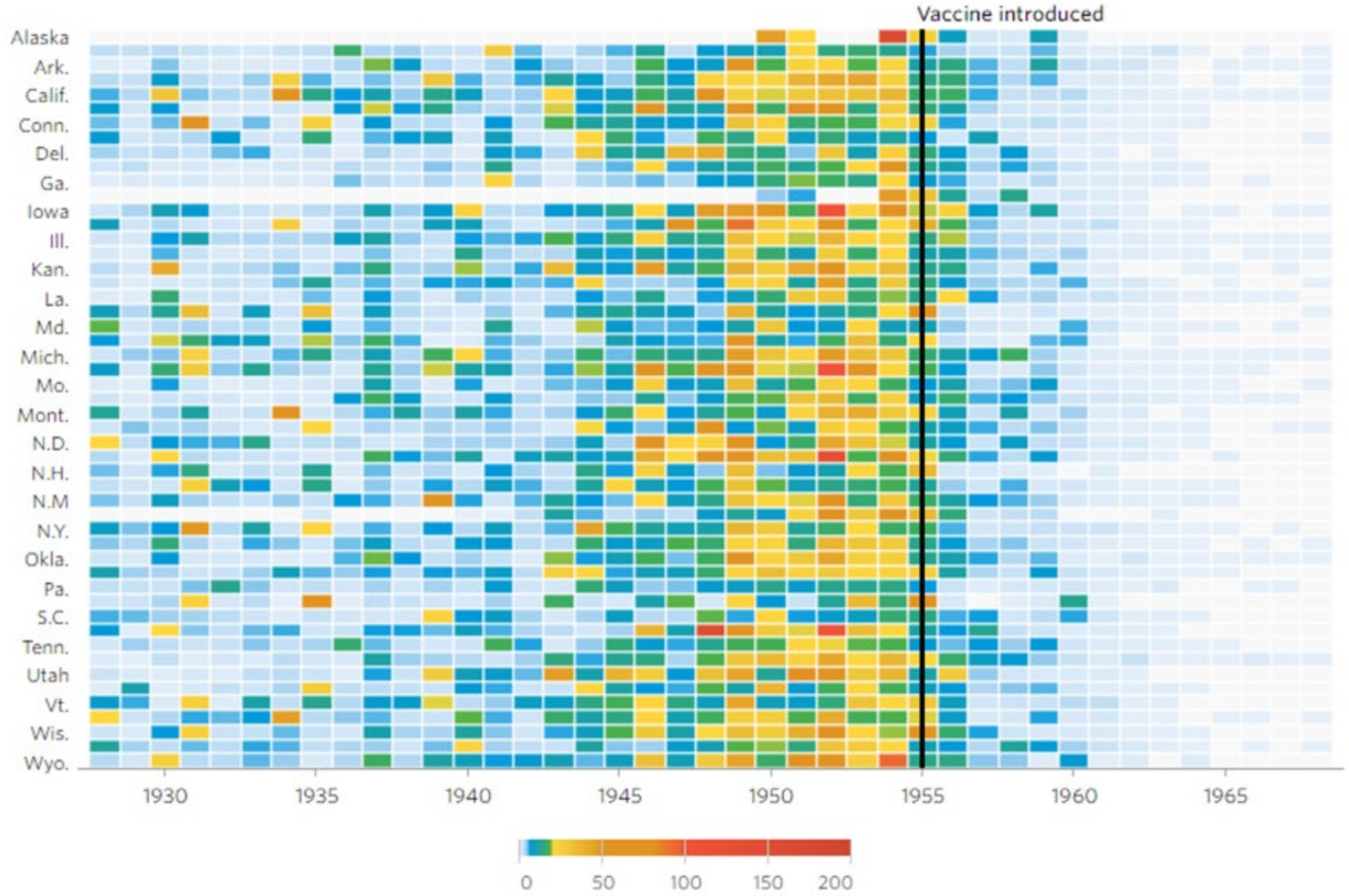
Simulation results for using full data, CRs only, and proposed method under four missing mechanisms

	Bias[a]		Variance[b]		95% CI[c]	
Method	$(\hat{\beta}_W)$	$(\hat{\beta}_X)$	$(\hat{\beta}_W)$	$(\hat{\beta}_X)$	$(\hat{\beta}_W)$	$(\hat{\beta}_X)$
			(M.1) $P(R=1)=0.66$			
Full	0.01346	0.02229	0.04008	0.03685	0.955	0.950
Comp	0.03062	−0.003561	0.1149	0.06732	0.960	0.955
Impu	0.01431	0.021	0.04088	0.05169	0.980	0.975
			(M.2) logit $P(R=1)=2Y$			
Full	0.007908	−0.02116	0.03838	0.03624	0.975	0.925
Comp	0.01945	0.07096	0.107	0.06581	0.960	0.950
Impu	0.006966	0.01597	0.04227	0.05226	0.975	0.985
			(M.3) logit $P(R=1)=2X$			
Full	0.007908	−0.02116	0.03838	0.03624	0.975	0.925
Comp	0.01225	0.0589	0.08856	0.06818	0.980	0.975
Impu	0.009563	−0.04699	0.03865	0.04923	0.985	0.970
			(M.4) logit $P(R=1)=X+Y$			
Full	0.01346	0.02229	0.04008	0.03685	0.955	0.950
Comp	0.02404	1.613	0.1102	0.08202	0.955	0.580
Impu	0.01814	0.08289	0.0578	0.06075	0.955	0.970

[a]Bias $= (\hat{\beta}-\beta_0)/\beta_0$.
[b]Simulation variance.
[c]Confidence interval using jackknife standard error.

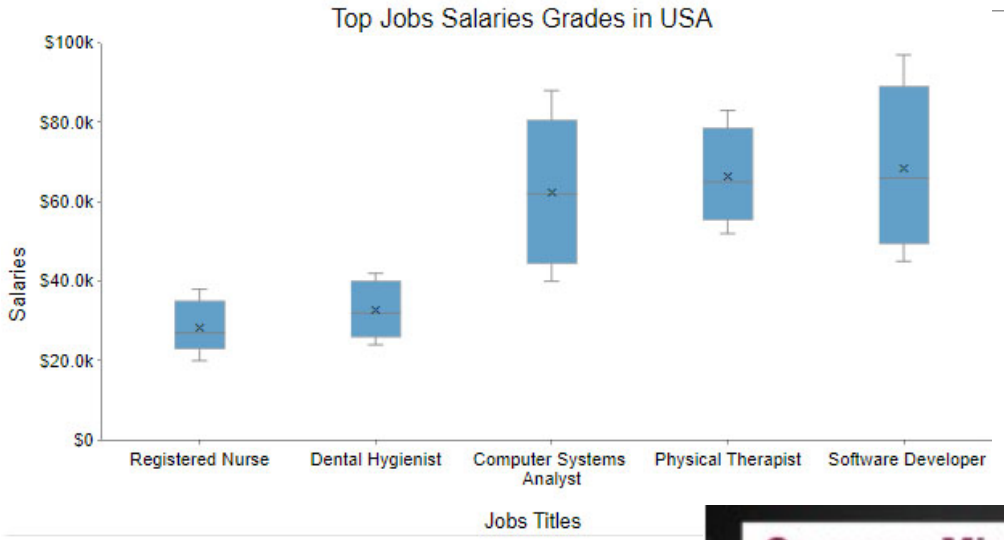
I		II		III		IV	
x	*y*	*x*	*y*	*x*	*y*	*x*	*y*
10.0	8.04	10.0	9.14	10.0	7.46	8.0	6.58
8.0	6.95	8.0	8.14	8.0	6.77	8.0	5.75
13.0	7.58	13.0	8.74	13.0	12.74	8.0	7.71
9.0	8.81	9.0	8.77	9.0	7.11	8.0	8.84
11.0	8.33	11.0	9.26	11.0	7.81	8.0	8.47
14.0	9.96	14.0	8.10	14.0	8.84	8.0	7.04
6.0	7.24	6.0	6.13	6.0	6.08	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

Polio



Battling Infectious Diseases in the 20th Century: The Impact of Vaccines, Wall Street Journal

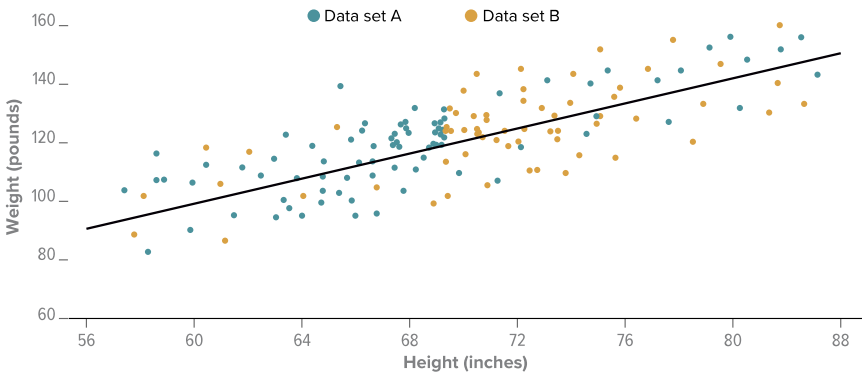
Visuals >> Tables



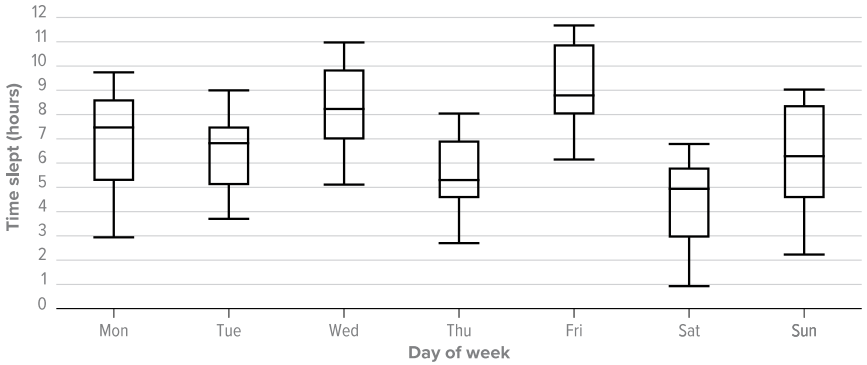
Choosing the right chart for the data

Continuous data are often better displayed in scatterplots, box plots and histograms than in simple bar charts.

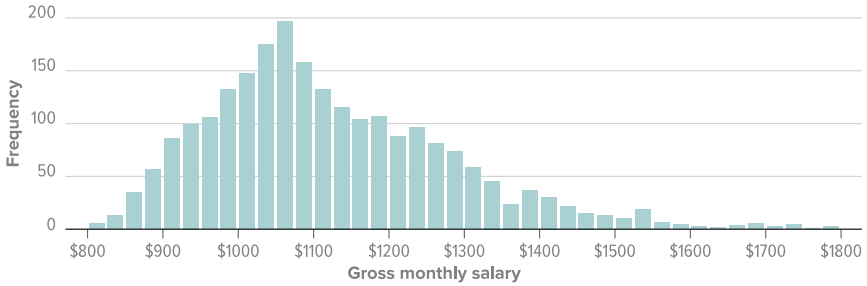
SCATTERPLOT



BOX PLOT



HISTOGRAM



5W INFOGRAPHIC / KNOWABLE

Data Representation

Looking at data without context is difficult!

Goal of data visualization = facilitating understanding

- Help tell a story of what the data is saying!

The 3 Cognitive Stages of Understanding:

- Perceiving — *what does it show?*
- Interpreting — *what does it mean?*
- Comprehending — *what does it mean to me?*

Ask Yourself: "What do I Want to Show?"

"What visual is best used for my data?"

- Bar Graph?
- Pie Chart?
- Line Graph?
- Scatter Plot?
- Histogram?
- Other?

- Dynamic or Static?

"What am I telling people about?"

- A Comparison?
- A Relationship?
- Some Distribution?
- A Description of data composition?

Chart Suggestions—A Thought-Starter

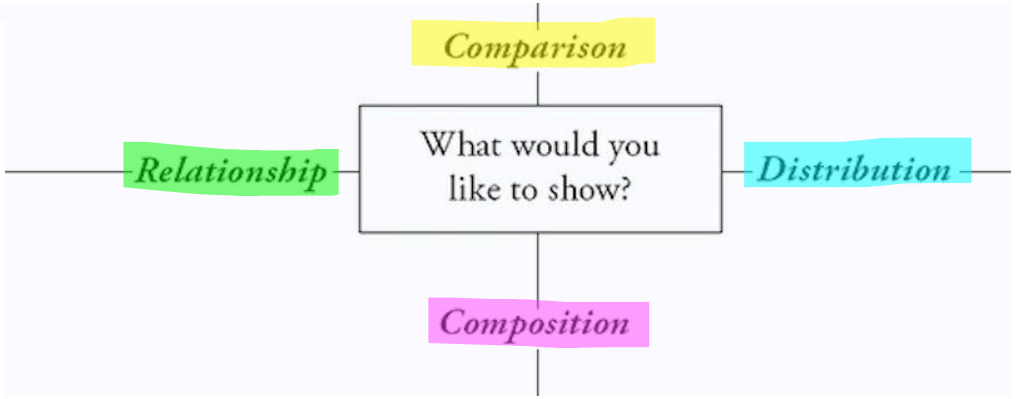
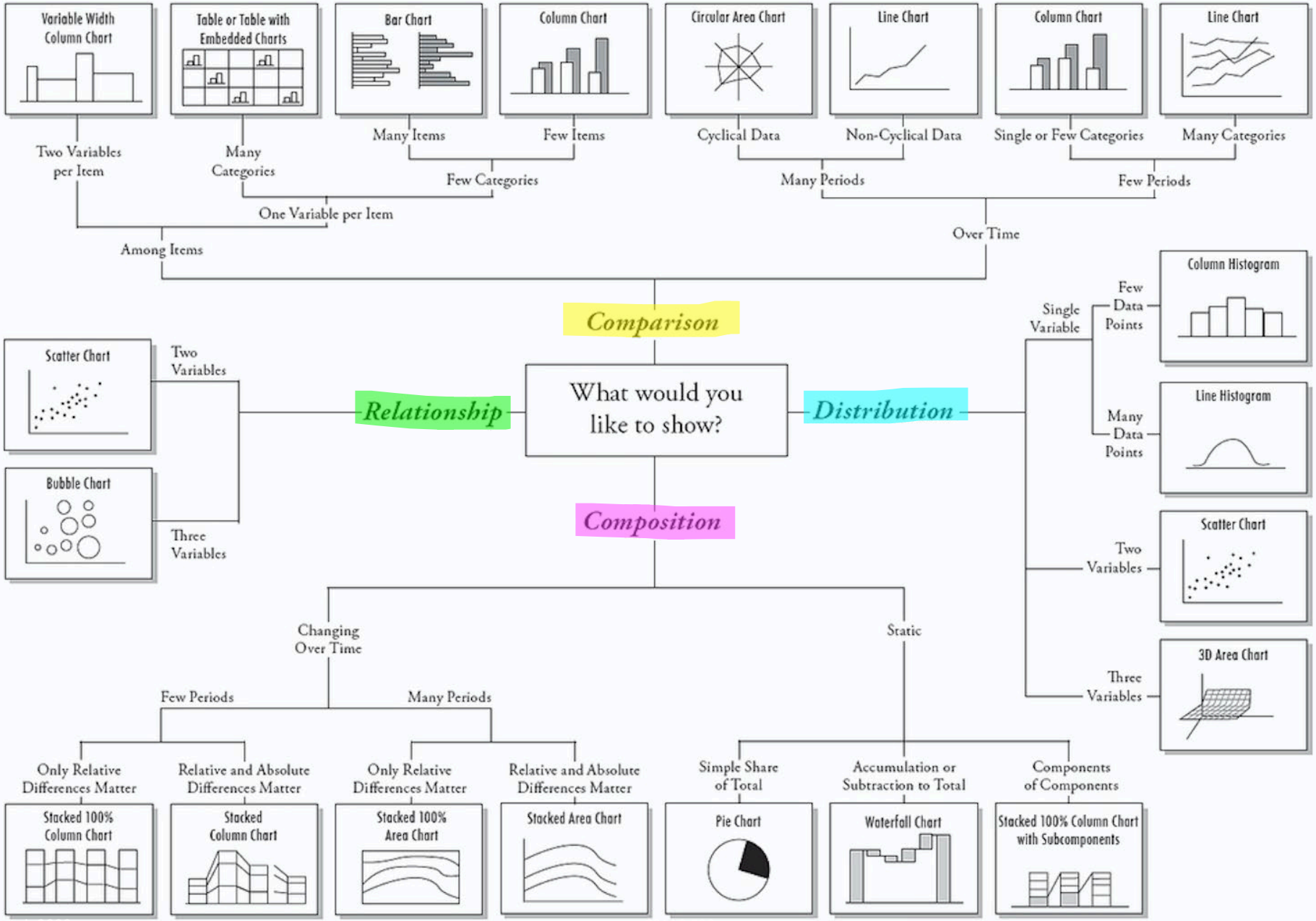


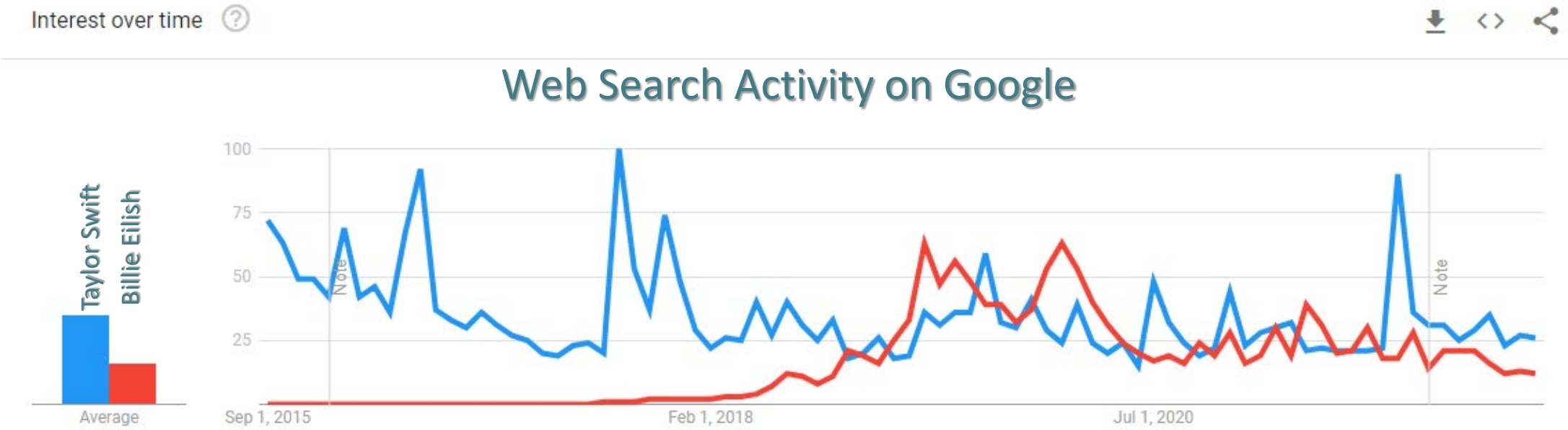
Chart Suggestions—A Thought-Starter



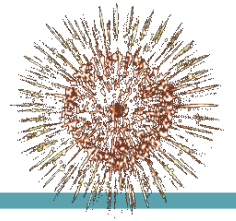
What Makes a Data Presentation *Effective*?

A Picture Is Worth 1000 Words

- **Pictorial superiority effect -** Our brains are led by our eyes
- Large parts of our brain activity are devoted to just visual processing



What Makes a Data Presentation *Effective*?



Making use of "**Early Attention**"

- Boosts the audience's ability to recall the information
- Use of: Color, alignment, symmetry, motion, orientation, size

Engaging with "**Working Memory**"

- Helps the audience comprehend what is being presented
- Use of: Good organization of the data, emphasis of important info
- Use of: Simplification (without over-simplifying)
- All this, in turn, helps the audience engage with their longer-term memory (*why is this important???*)

Presenting Data Effectively by S. Evergreen (2011)

Good Data Representation is *Good Design*

There's a matter of **aesthetics** and taste

- A lot of that comes from experience – it is a *skill*
- So, you can get better at it with practice

But! Good Design of <anything> is also about making the design…

1. …**useful**
2. …**understandable**
3. …**unobtrusive**
4. …**honest**
5. …**thorough**
6. …as **minimal** (uncluttered) as you can

Inspired by Dietar Ram's "Rules of Good Design" (per Kirk, 2016)

Kirk's 3 Principles of **Good** Data Visualization

Good Data Visualization is Good Design

1. Good Data Visualization is **Trustworthy**
2. Good Data Visualization is **Accessible**
3. Good Data Visualization is **Elegant**

Data Visualization by A. Kirk (2011)

Trustworthiness of Data Viz

Good data viz should **not be misleading**

A part of this is about the ***source*** of the data…

- **Professional researchers**, Professional news agency, *etc…*

  vs. gossip-oriented news, my Aunt Karen, TikTok, *etc…*

- Hidden biases and intents can present a risk in our interpretation of data

…Another part is ***how*** the data is presented aesthetically

- Flashy colors (generally seen as "amateurish") vs low key colors (better)
- Avoid "gimicky" fonts and busy backgrounds
- Make use of good x-y axes in graphs (*more on this in a bit…*)

Accessibility of Data Viz

Remember: ***It's all about getting the information across AND it's about the audience!***

To make your data viz accessible, you need to think of:

- Subject matter appeal (if audience not interested, then…)
- Subject matter knowledge (what might they know about this stuff?)
- Always come back to: **What do they need to know?**
- Their time, (make it quick,
  Your format, make it easy to "get",
  Their attitude/emotion make it relevant to them)
- The art should be more "utilitarian art" than "fine art"
  (balancing cool artistic themes/flair w/ goal of being practical/effective)

Elegance of Data Viz

Elegance is important, but it can be an elusive pursuit…

Hard to design for (without stylish confidence)– i.e. it's a skill

More appreciated when it's *not* there

What can you do to ensure that your design is "elegant"?

- Make it less cumbersome, more consistent in style
- Eliminate the arbitrary/unnecessary (this needs good editing skills)
- Be thorough (no short cuts, don't assume, respect your audience)
- Don't go for "*style over substance*"
- Practice minimization, but realize "too little" is just as bad as "too much" (so find a balance)

Bad Data Viz Example 1

inelegant
misleading

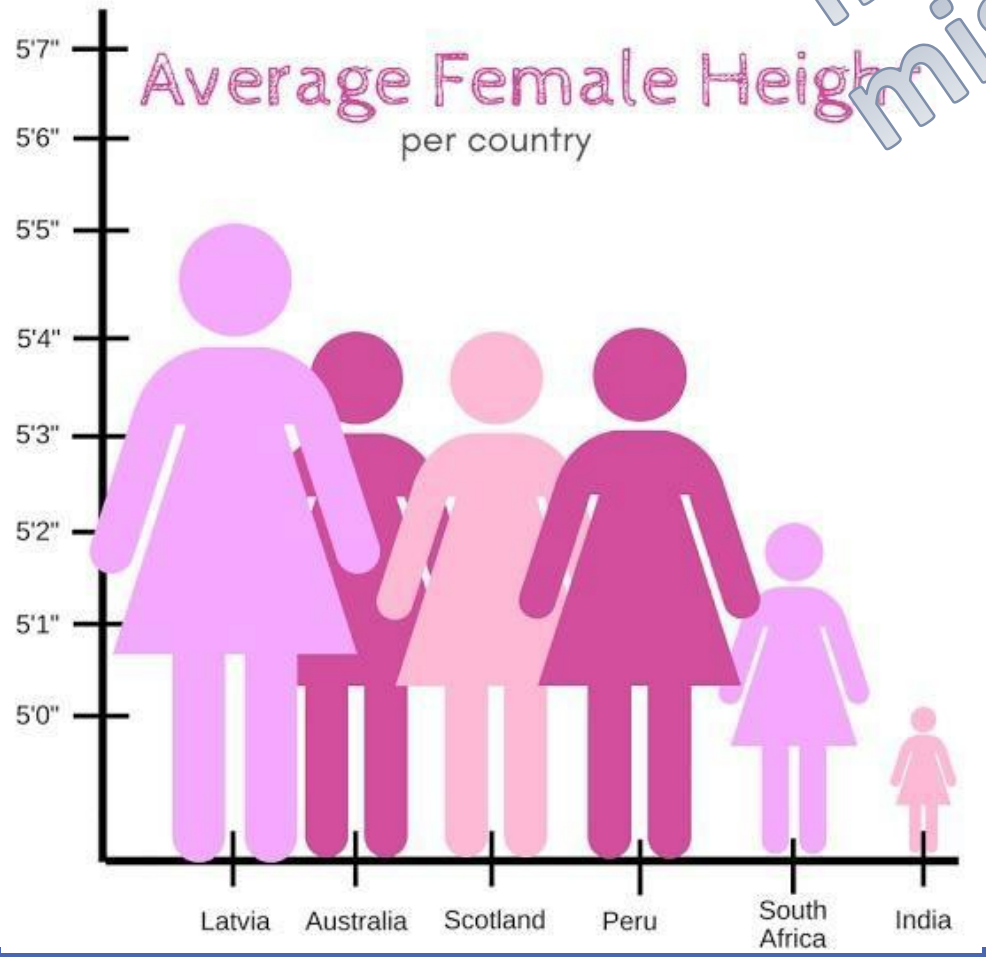


Image from: https://badvisualisations.tumblr.com/

- The y-axis should start at zero
  - Are Indian women really 1/5th the size of Latvian women??!!
  - Truncating the y-axis can give a dramatically different (i.e. ***false***) impression of the data

Also: why not use bars instead of figures?

- Cleaner, less-cluttered

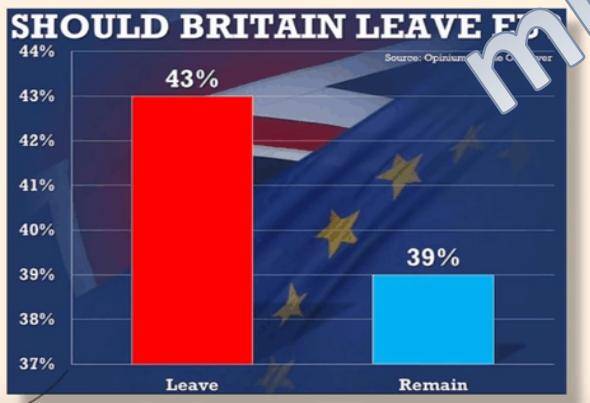
1. Good Data Visualization is **Trustworthy**
2. Good Data Visualization is **Accessible**
3. Good Data Visualization is **Elegant**

Bad Data Viz Example 2

misleading

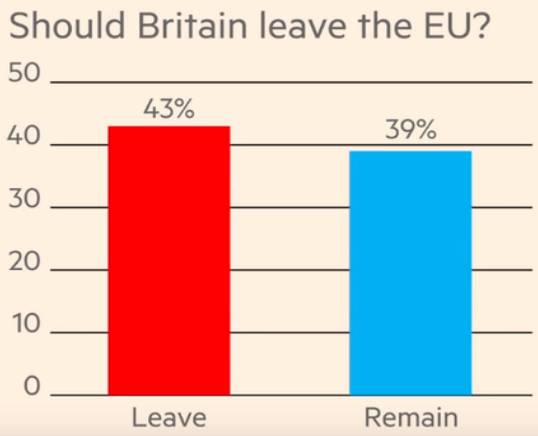
Graphics that are accurate but misleading

Baseline should start at zero, not 37



Graphics that are accurate but misleading

A better chart of the same data



Similar to the previous example:

Baseline not at zero for y-axis can bias the interpretation of the data!

1. Good Data Visualization is **Trustworthy**
2. Good Data Visualization is **Accessible**
3. Good Data Visualization is **Elegant**

Bad Data Viz Example 3

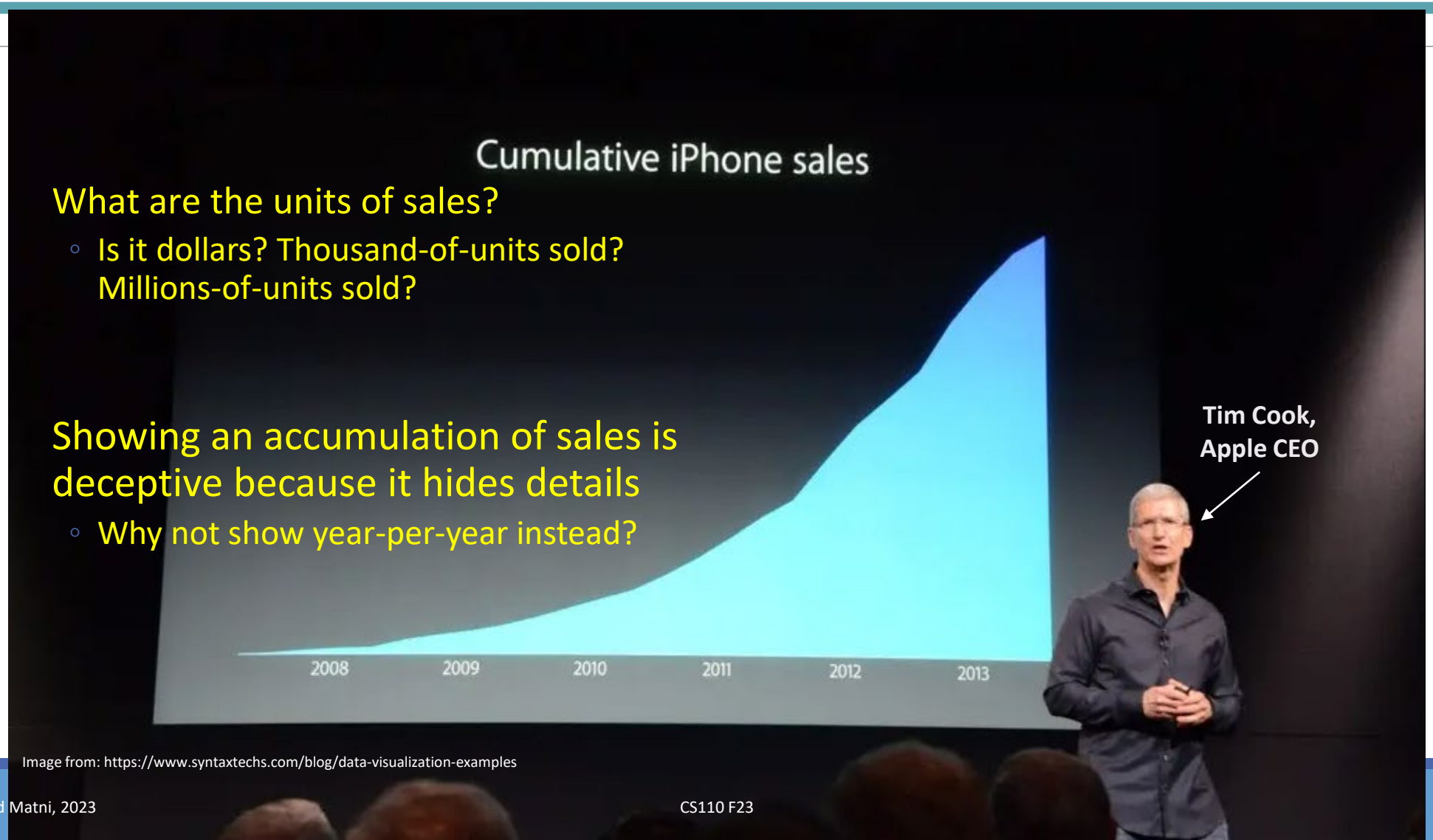


Image from: https://www.syntaxtechs.com/blog/data-visualization-examples

Bad Data Viz Example 4

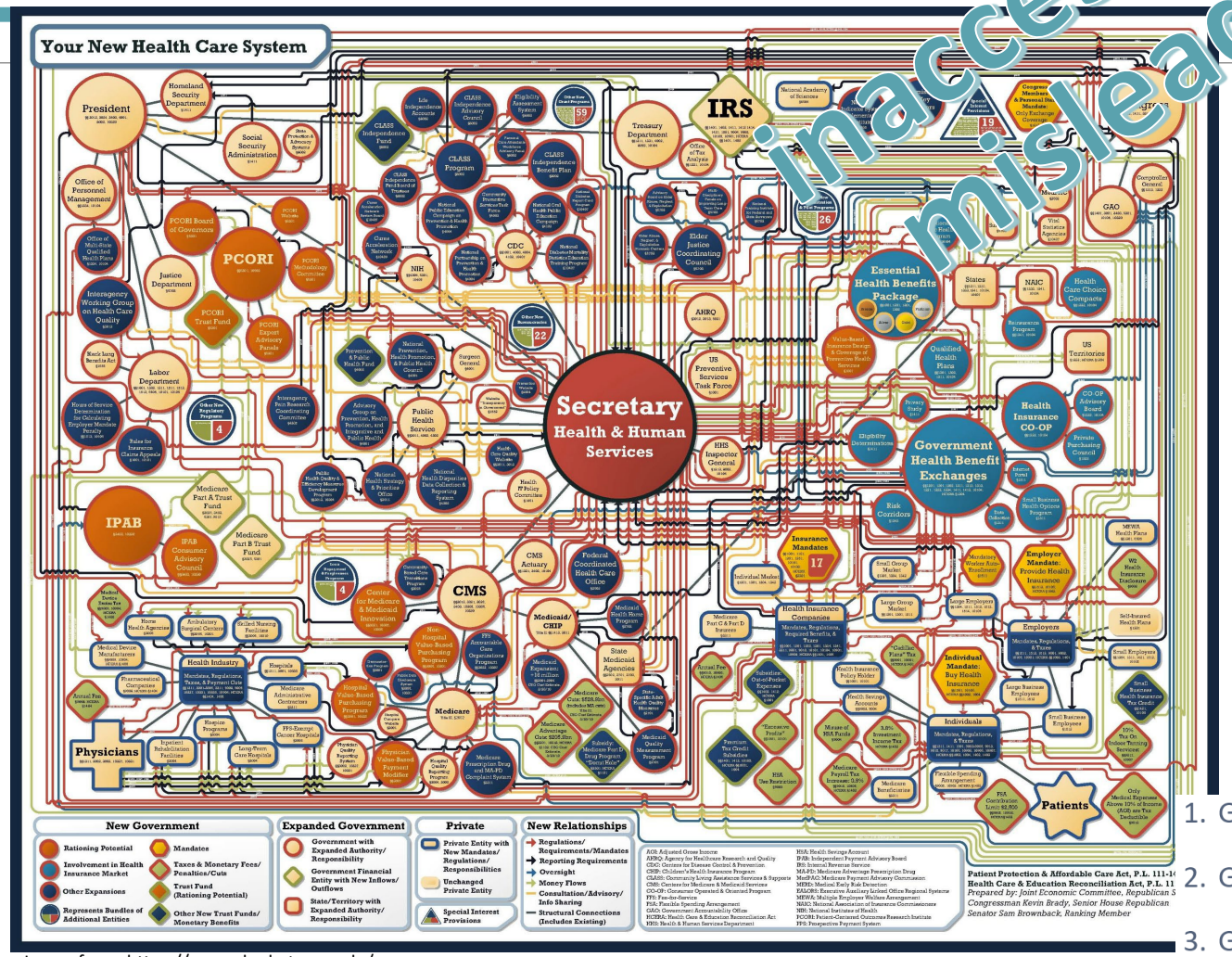


Image from: https://www.dwrl.utexas.edu/

inaccessible

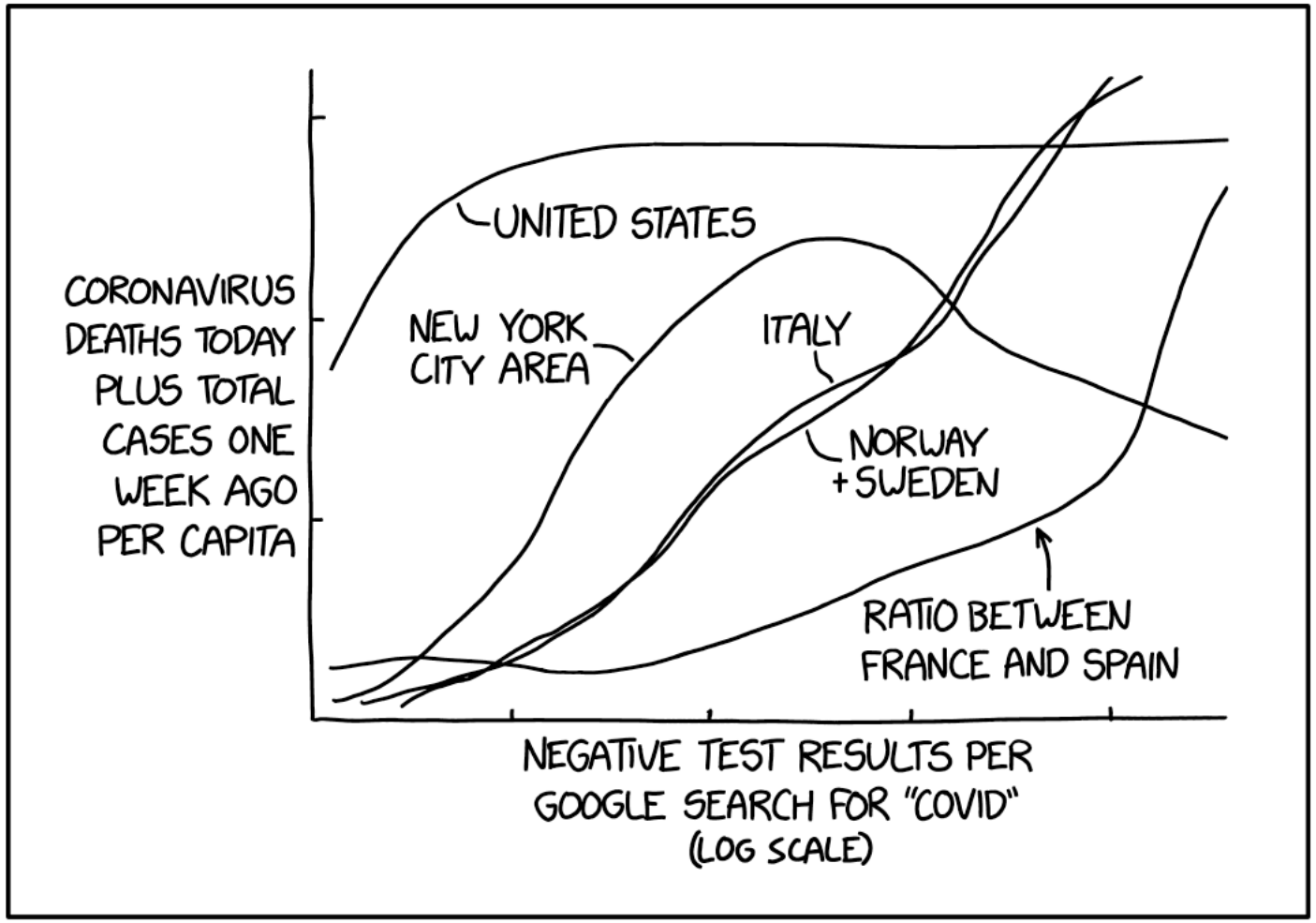
misleading

Overly complicated way to explain “Obama Care”

Can bias people against this health care initiative

The chart itself seeks to dissuade the viewer from investigating further!

1. Good Data Visualization is **Trustworthy**
2. Good Data Visualization is **Accessible**
3. Good Data Visualization is **Elegant**



I'M A HUGE FAN OF WEIRD GRAPHS, BUT EVEN I ADMIT SOME OF THESE CORONAVIRUS CHARTS ARE LESS THAN HELPFUL.

https://xkcd.com/2294/

Well Done Data Visualizations Examples 1

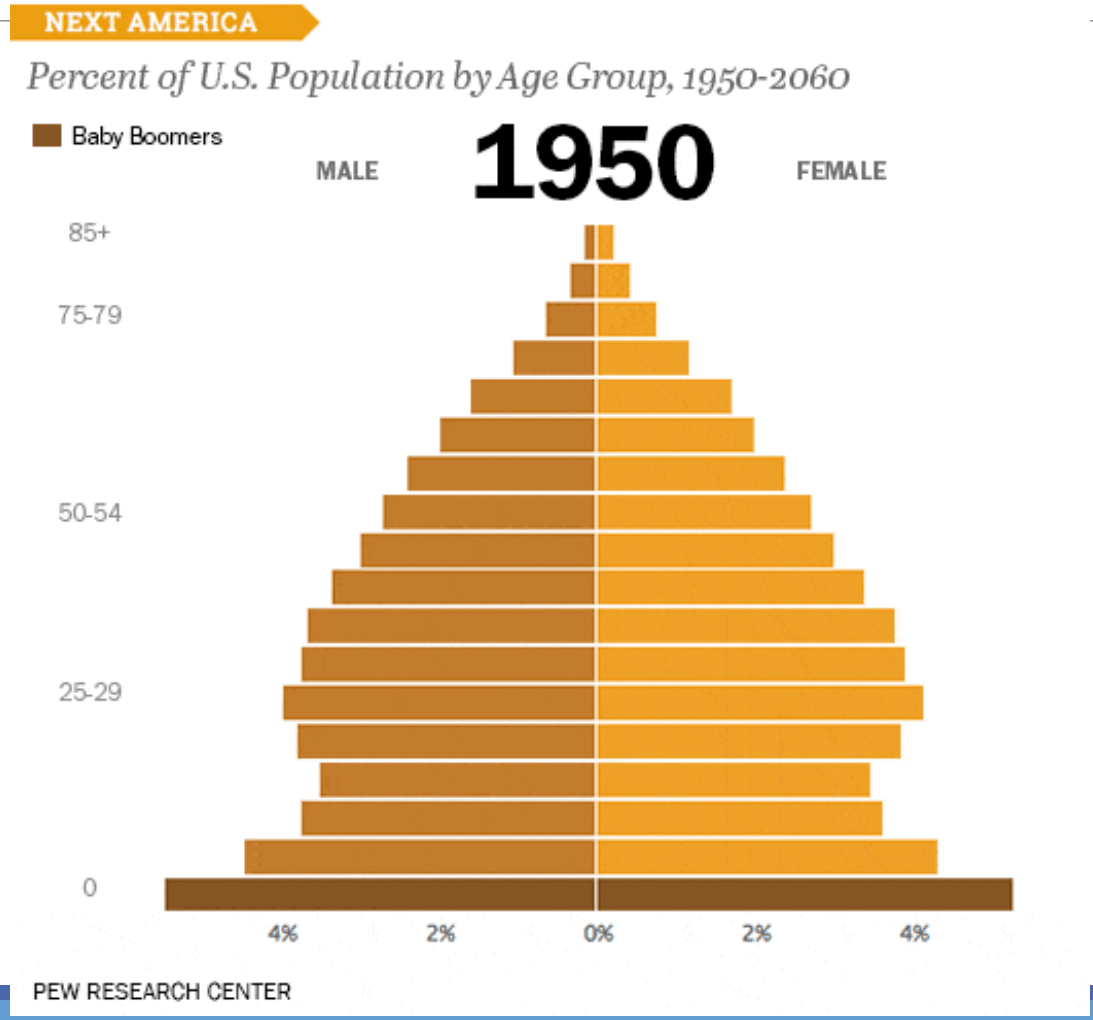
US 2013 Budget Proposal Exploration

https://archive.nytimes.com/www.nytimes.com/interactive/2012/02/13/us/politics/2013-budget-proposal-graphic.html

US Electricity Generation

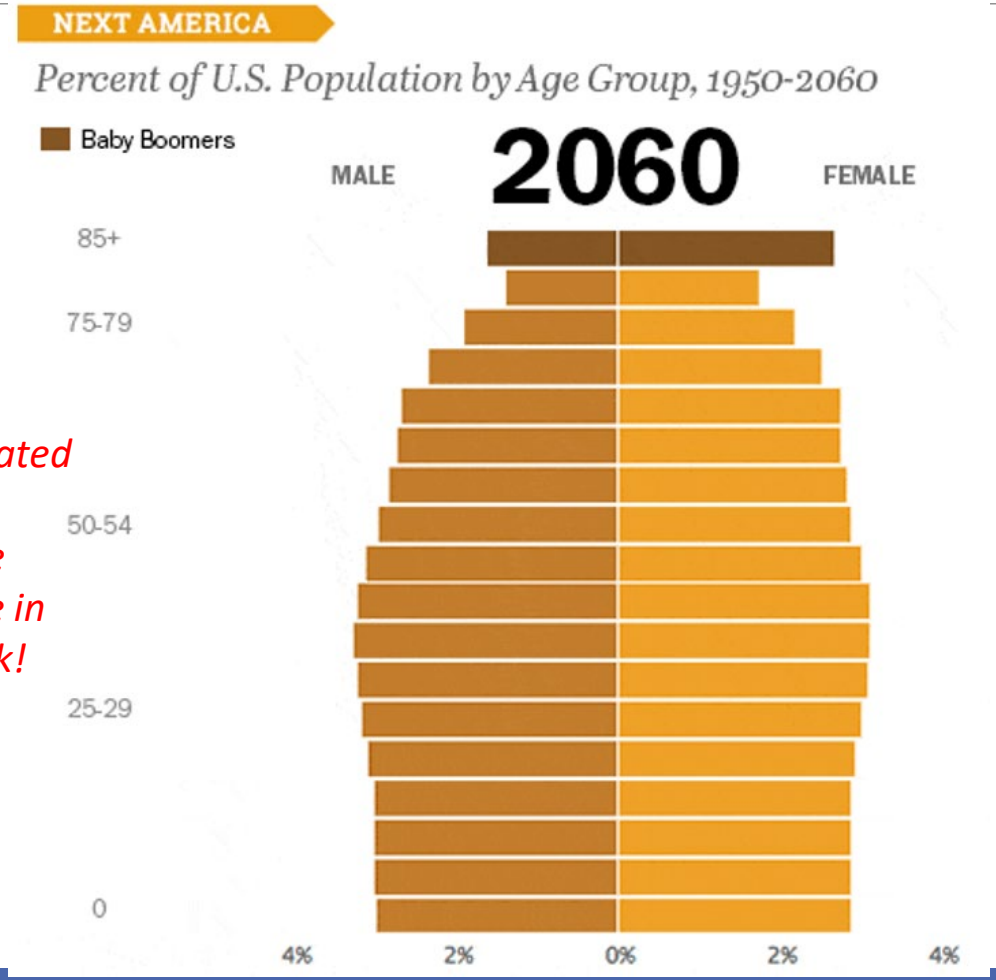
https://www.carbonbrief.org/mapped-how-the-us-generates-electricity/

Well Done Data Visualizations Examples 2



Illustrates how the US Population is aging in relevant animation style

Well Done Data Visualizations Examples 2



This was an animated picture, so I'm showing what the end of it looks like in this PDF slide deck!

Illustrates how the US Population is aging

Some Example S/W for Producing Data Visualizations

Excel / Google Sheets

- Great for basic plots and histograms

Python

- https://mathdatasimplified.com/top-6-python-libraries-for-visualization-which-one-to-use/
- https://rklopotek.blog.uksw.edu.pl/files/2017/09/data-visualization-2.1.pdf

R

- If you're inclined to do some more advanced statistical analysis

Interested in Learning More About Data Viz?

Extended topics can be about:

Designing for interactivity

Designing for Color, Composition

Designing for specific contexts (research, marketing, management, etc…)

Remember: it's a skill, so you get better with practice, practice, practice

Good textbooks:

Data Visualisation	by Andy Kirk
The Visual Display of Quantitative Information	by Edward R. Tufte
Storytelling With Data	by Cole Nussbaumer Knaflic

</LECTURE>