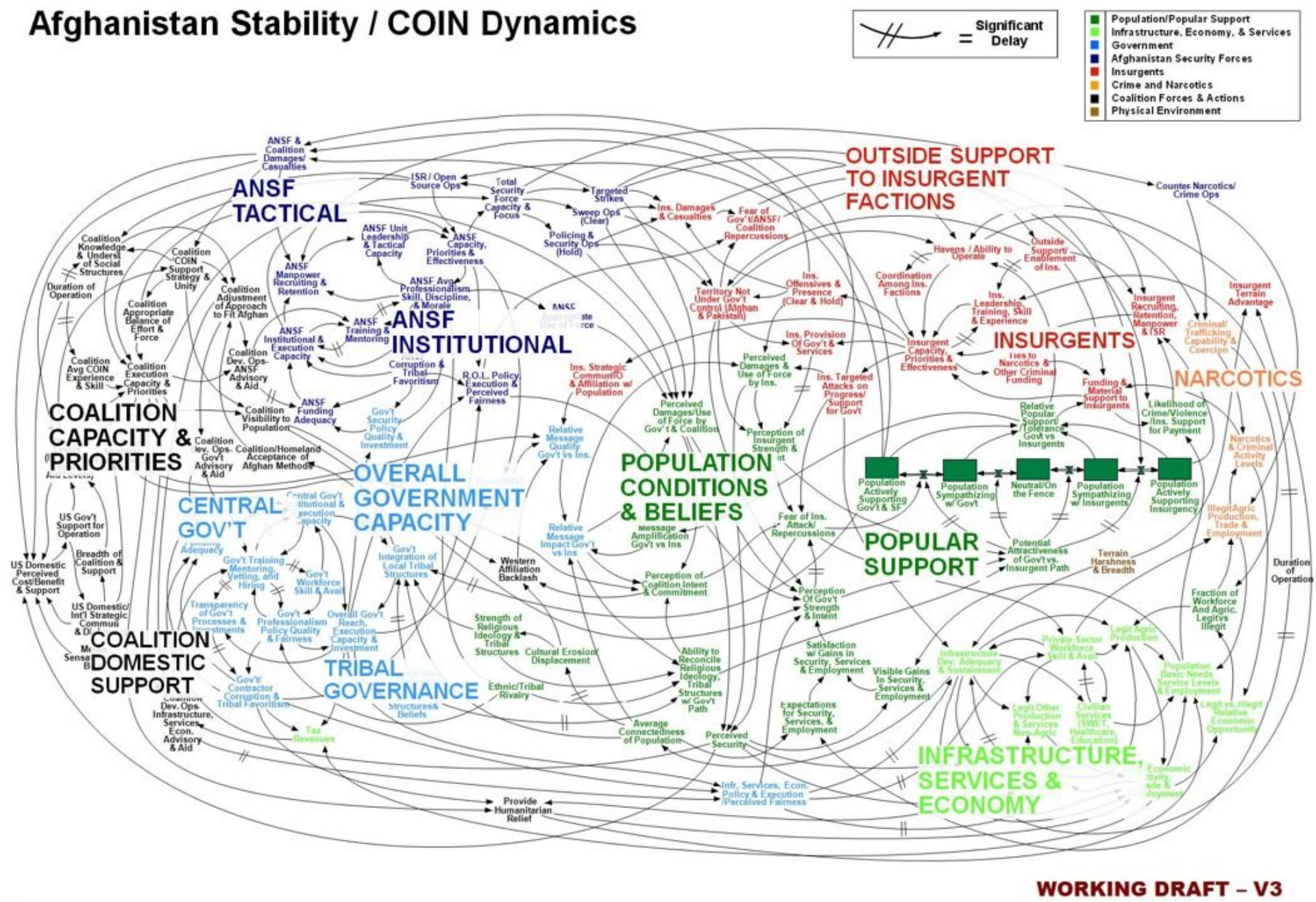
Data Visualization Principles

Michael Beckstrand, Ph.D.
LATIS Research
mjbeckst@umn.edu

Basic Graphical Principles

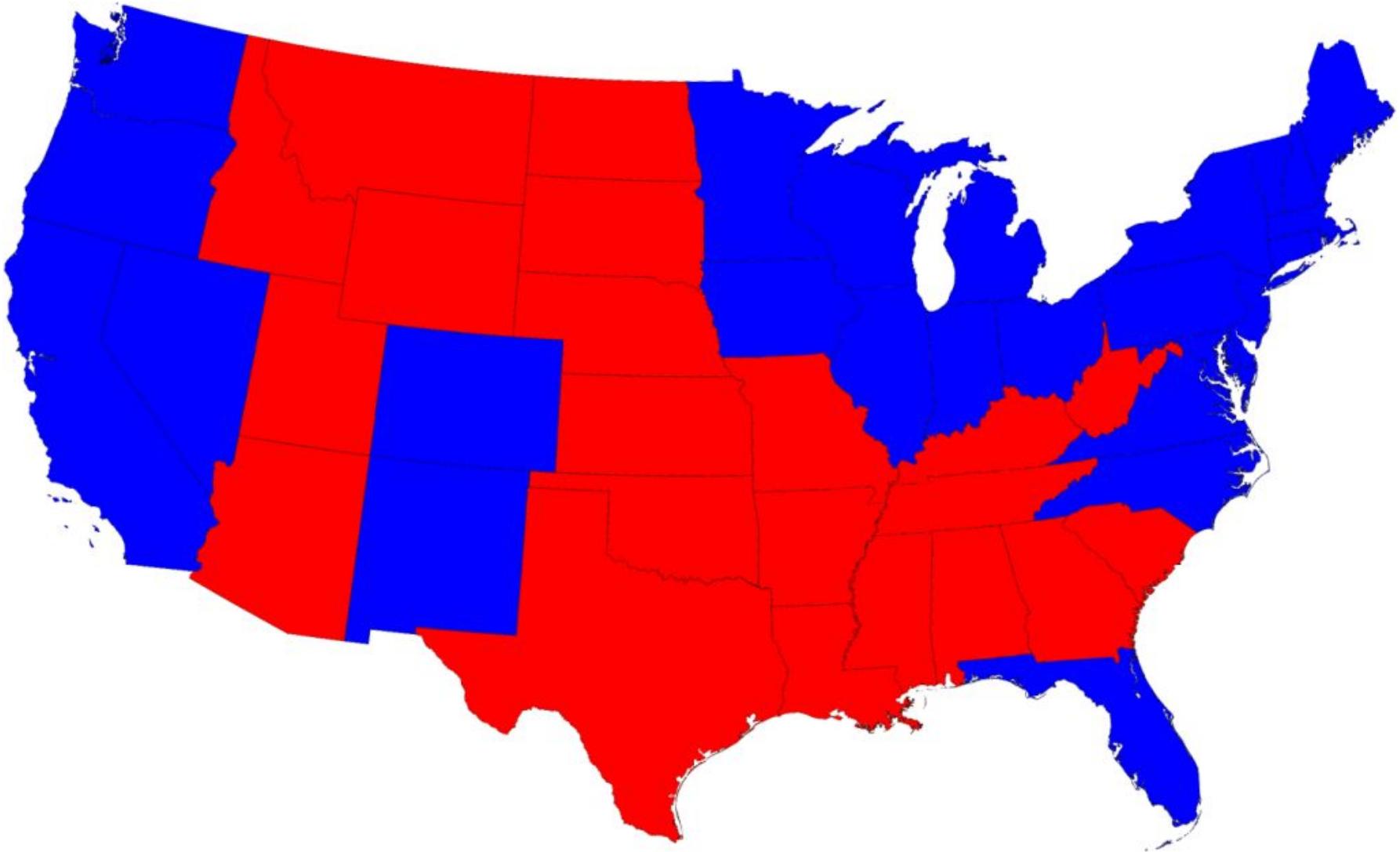
- Graphics should
  - Be substantive
  - Avoid distorting data
  - Present complex data more simply
- Which graph to use when?
  - Univariate: histogram or pie chart
  - Multivariate: bar charts, line graphs, scatter plots
- Be aware of “data-ink” ratio

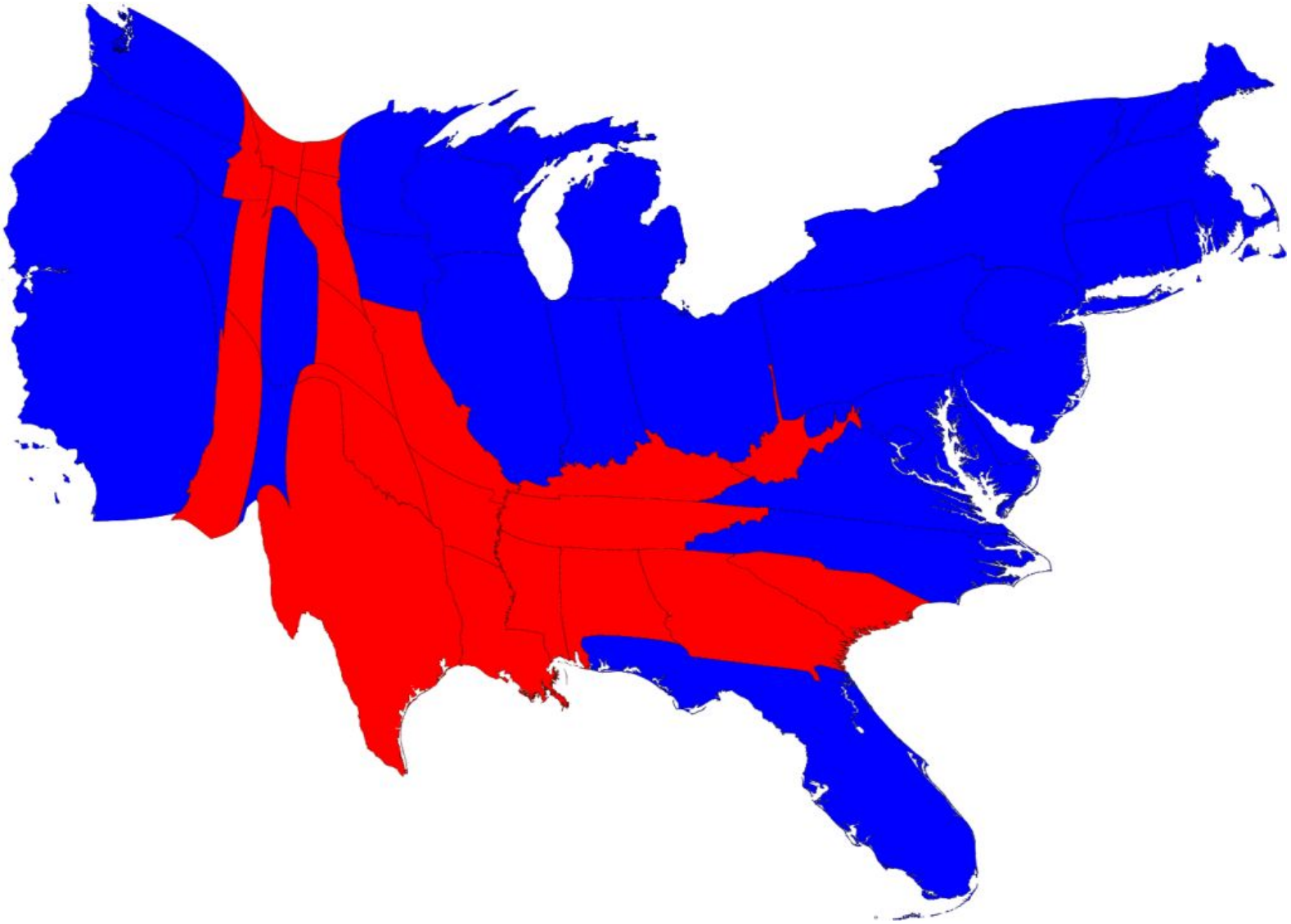
Afghanistan Stability / COIN Dynamics

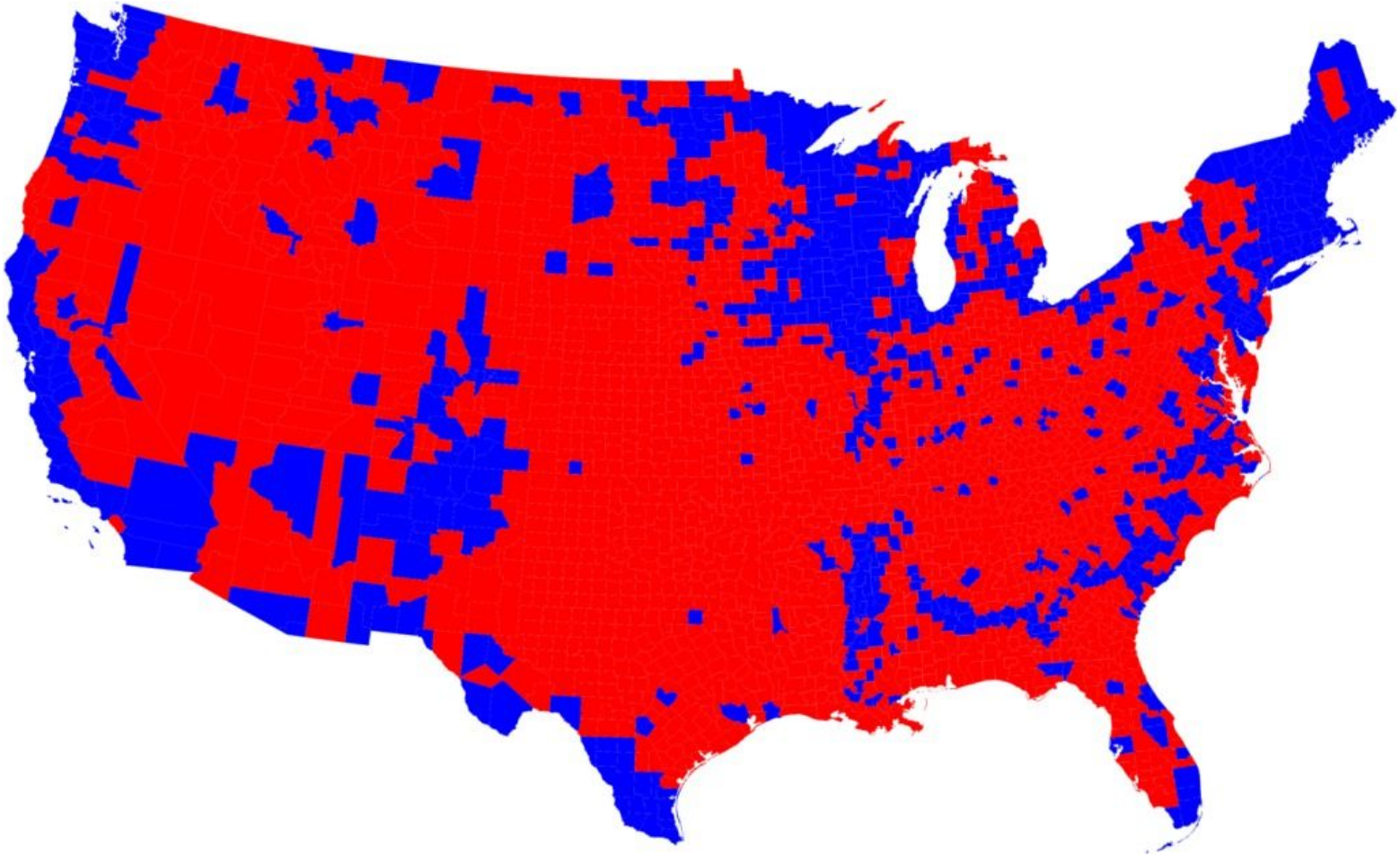


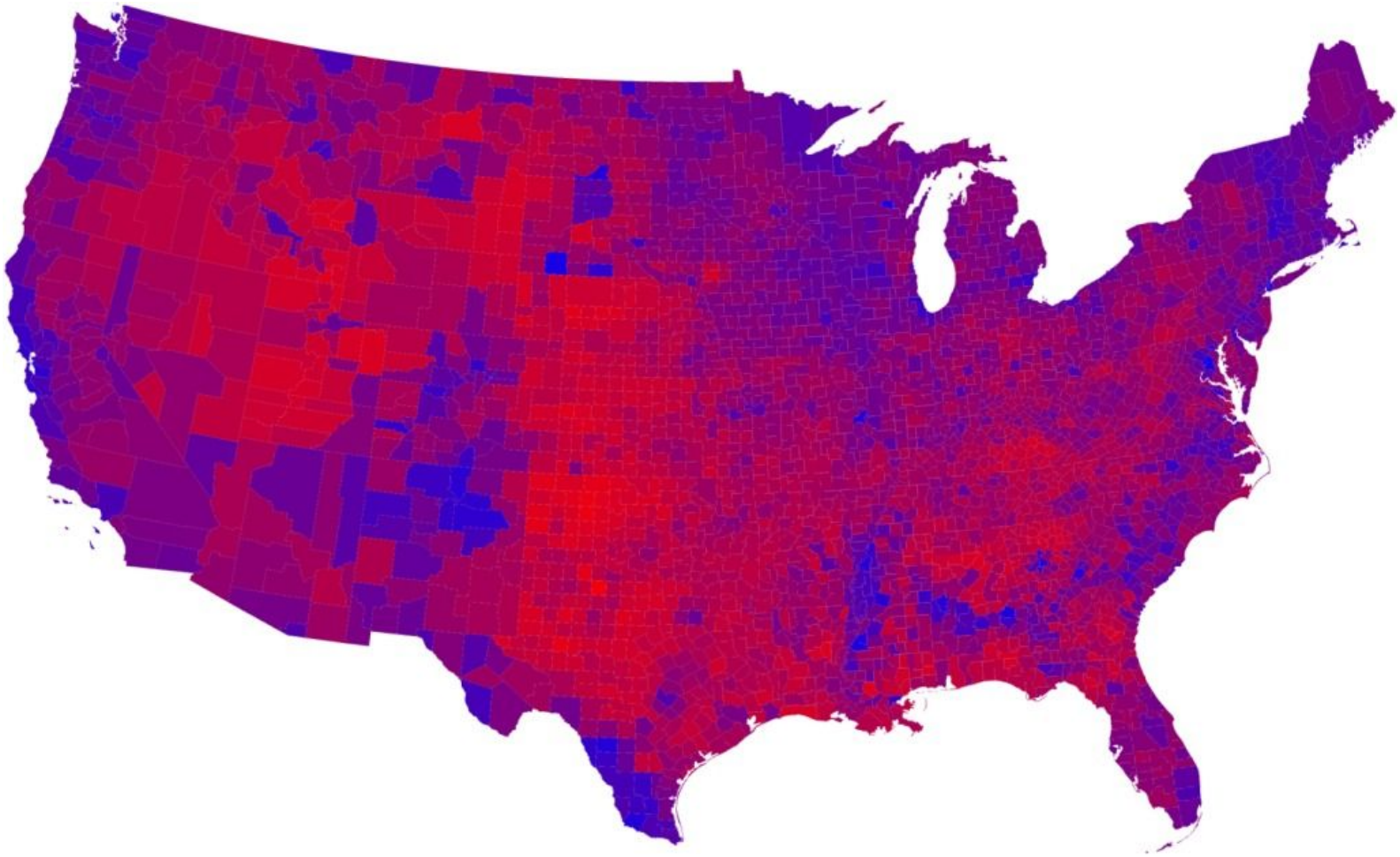
WORKING DRAFT – V3

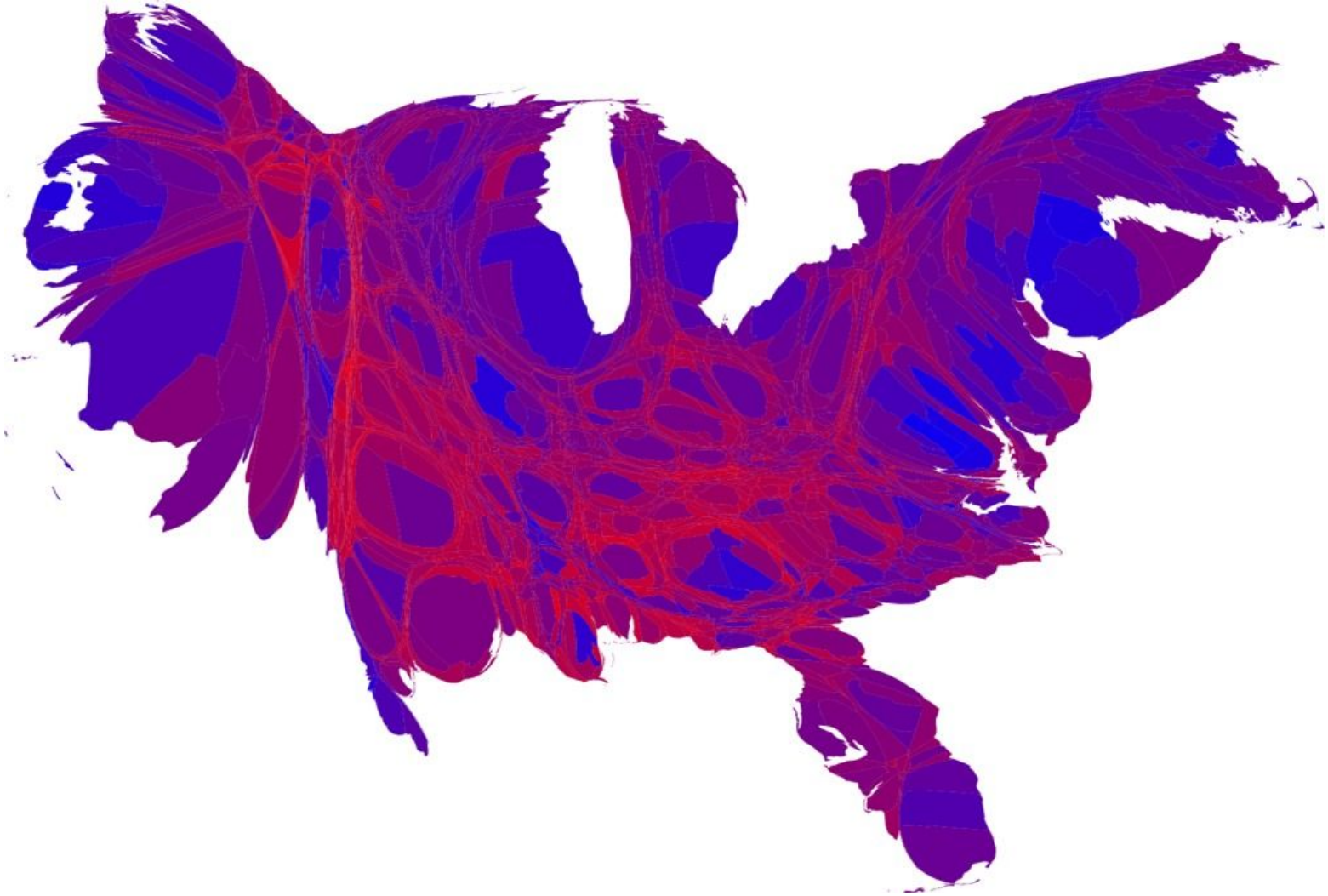
Electoral Map (2008)











Keep it Simple

- Avoid unnecessary lines
- Report only necessary numbers
- Avoid unnecessary ‘3D’ effects
- Avoid other fancy effects that distort data

Messy starting point

% Voters who Report Contact from Campaign

State	Obama	McCain	Gap
Nevada	50	29	21
Colorado	51	34	17
Indiana	37	22	15
Virginia	50	38	12
Pennsylvania	50	39	11
Iowa	41	30	11
Florida	29	20	9
North Carolina	34	26	8
Missouri	44	37	7
Ohio	43	36	7
Wisconsin	42	39	3
West Virginia	29	31	-2

Minus unnecessary lines

% Voters who Report Contact from Campaign

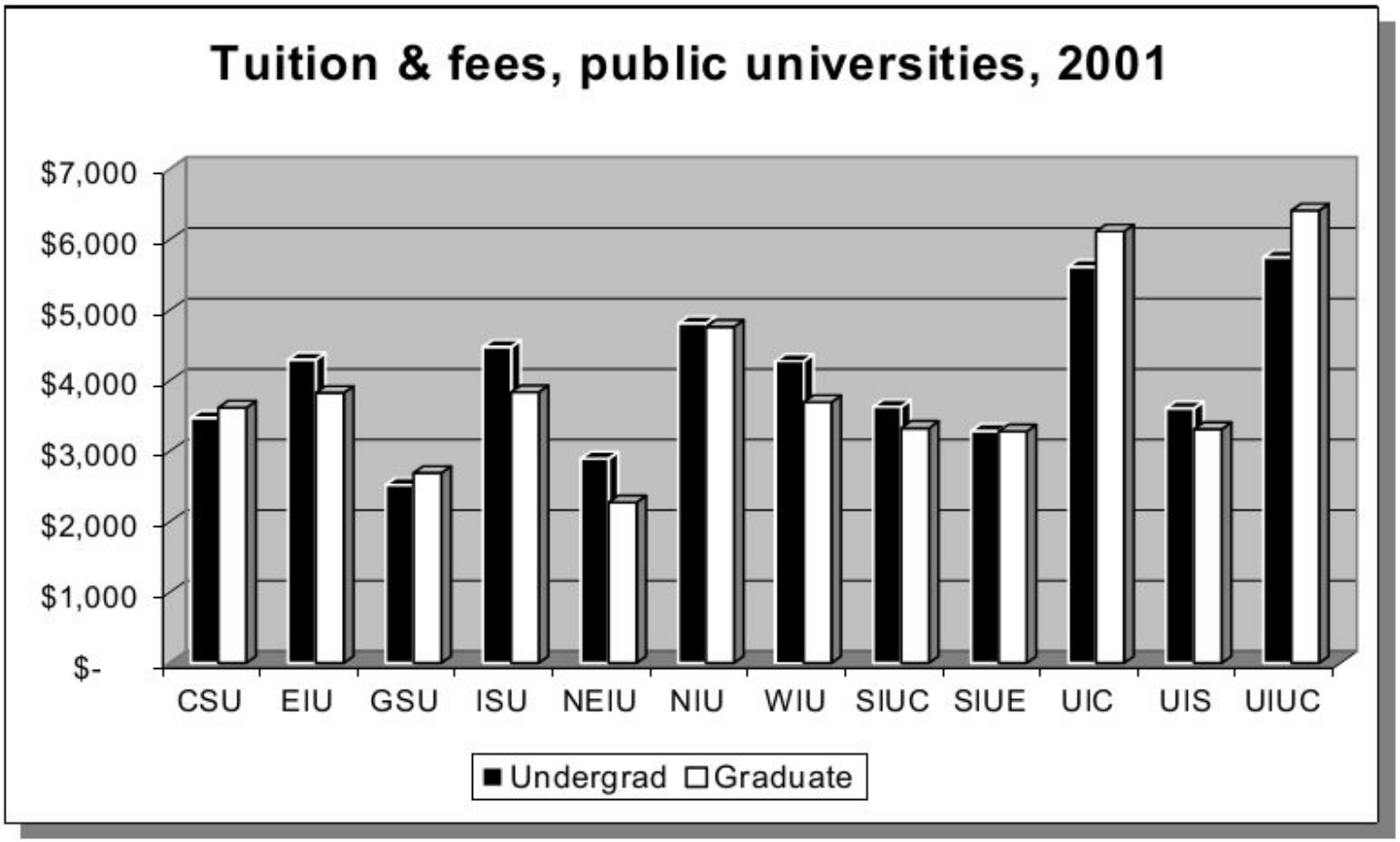
State	Obama	McCain	Gap
Nevada	50	29	21
Colorado	51	34	17
Indiana	37	22	15
Virginia	50	38	12
Pennsylvania	50	39	11
Iowa	41	30	11
Florida	29	20	9
North Carolina	34	26	8
Missouri	44	37	7
Ohio	43	36	7
Wisconsin	42	39	3
West Virginia	29	31	-2

With only necessary #s

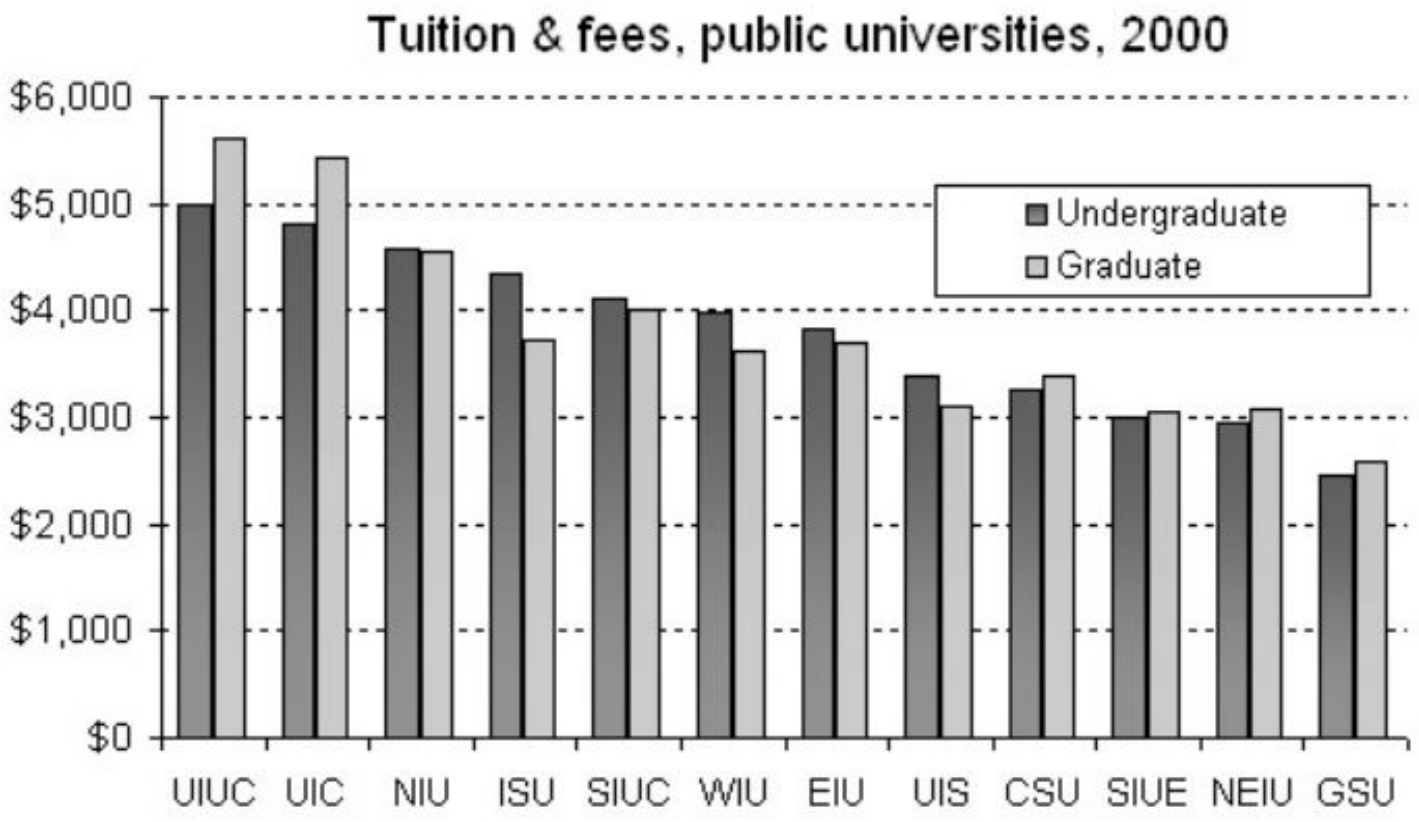
Voters Contacted: Obama's lead over McCain

State	Gap (%)
Nevada	21
Colorado	17
Indiana	15
Virginia	12
Pennsylvania	11
Iowa	11
Florida	9
North Carolina	8
Missouri	7
Ohio	7
Wisconsin	3
West Virginia	-2

Don’t: 3D

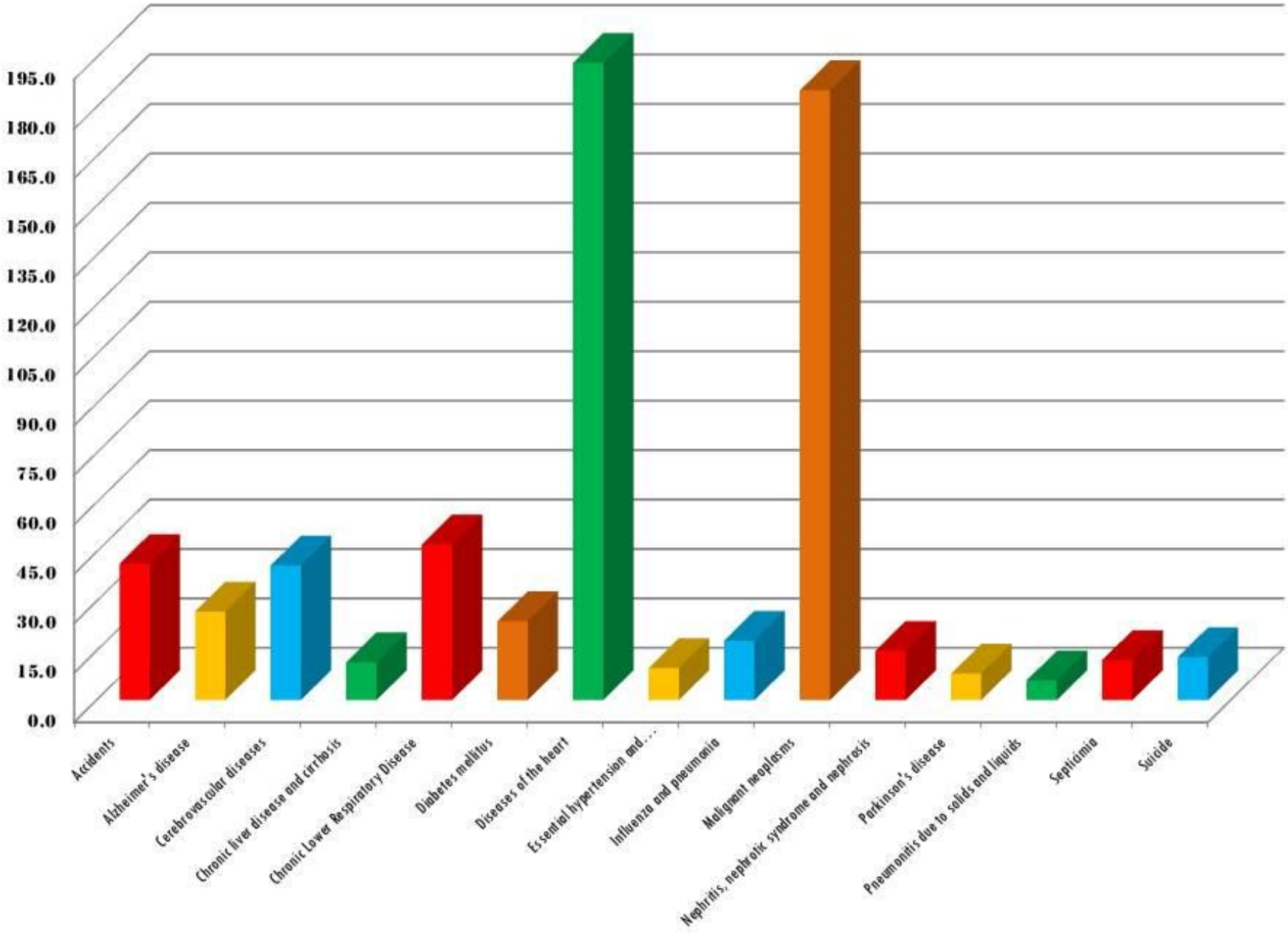

Tuition & fees, public universities, 2001
$7,000
$6,000
$5,000
$4,000
$3,000
$2,000
$1,000
$-
CSU
EIU
GSU
ISU
NEIU
NIU
WIU
SIUC
SIUE
UIC
UIS
UIUC
Undergrad
Graduate

Clearer Without

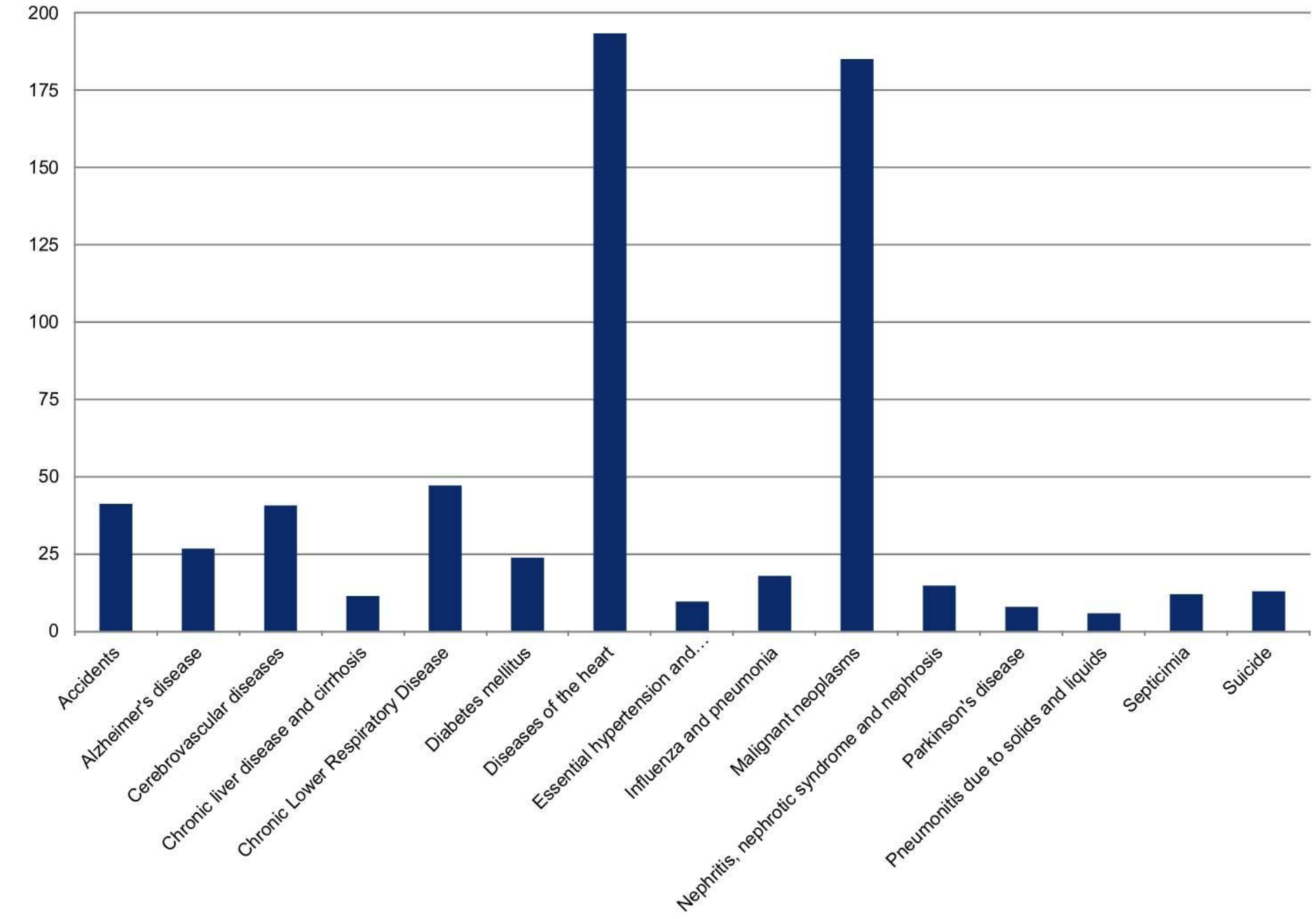


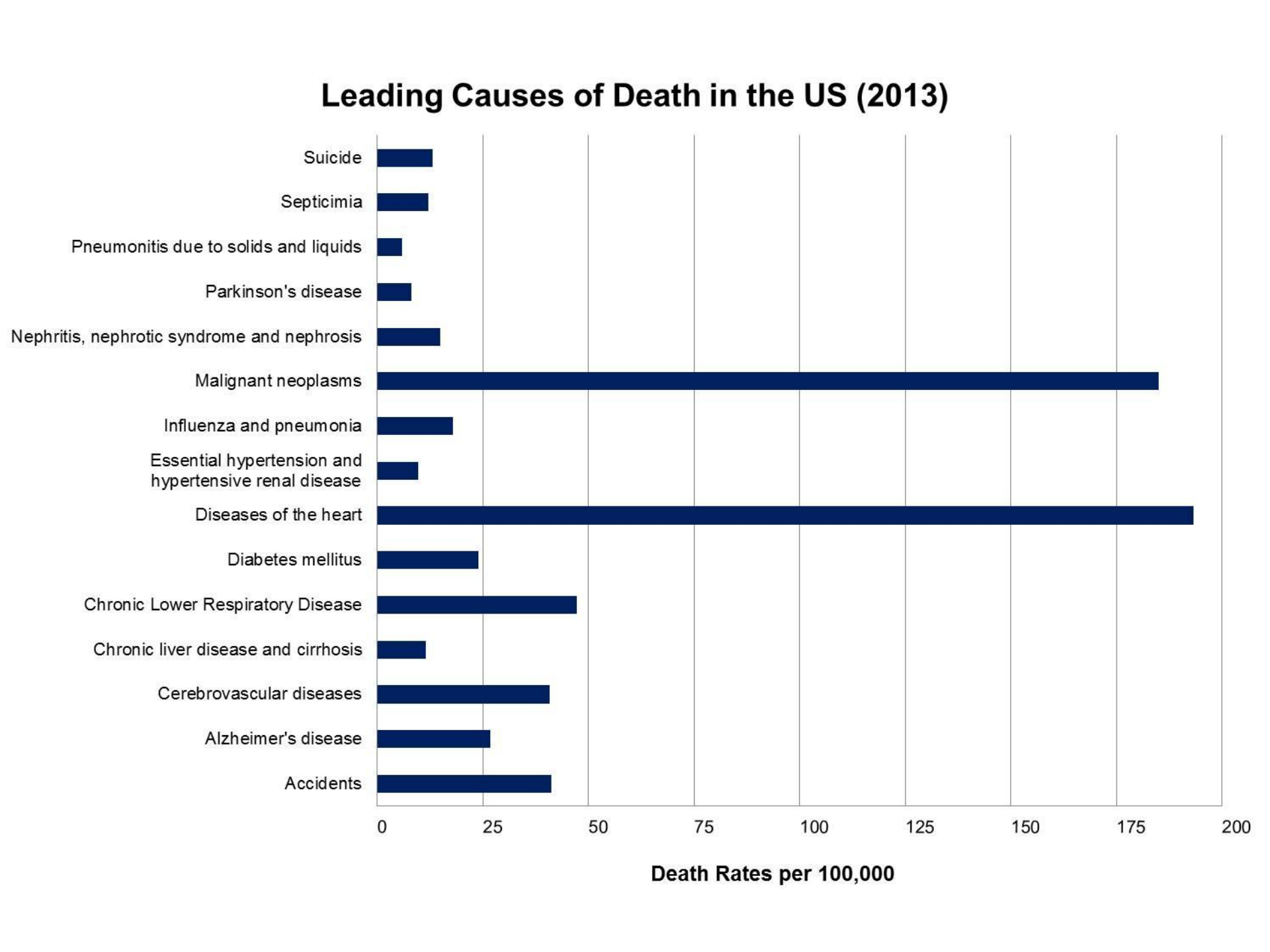
Making a terrible graphic not so terrible

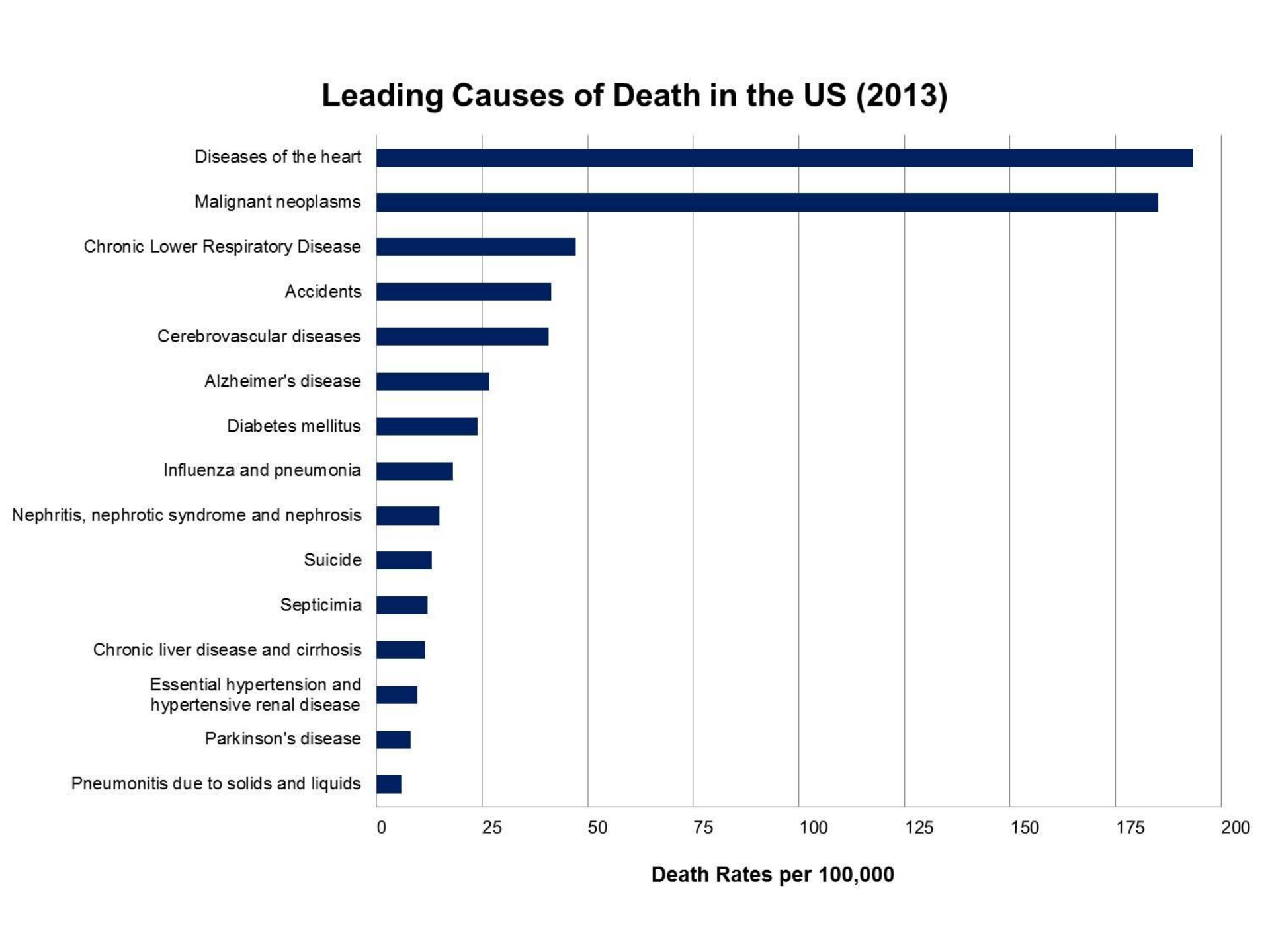
Leading Causes of Death in the US (2013)



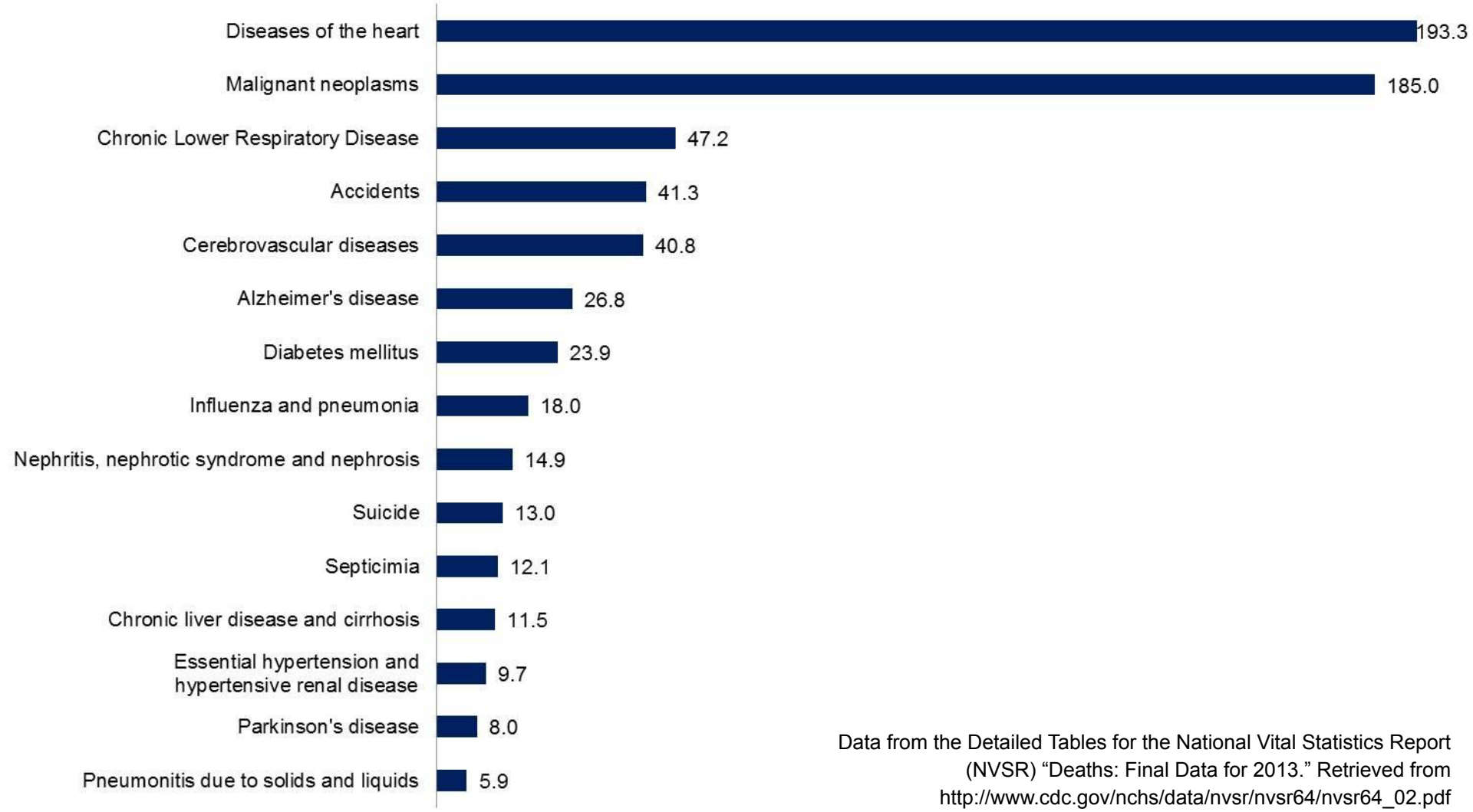
Leading Causes of Death in the US (2013)




Leading Causes of Death in the US (2013)
Suicide
Septicimia
Pneumonitis due to solids and liquids
Parkinson's disease
Nephritis, nephrotic syndrome and nephrosis
Malignant neoplasms
Influenza and pneumonia
Essential hypertension and hypertensive renal disease
Diseases of the heart
Diabetes mellitus
Chronic Lower Respiratory Disease
Chronic liver disease and cirrhosis
Cerebrovascular diseases
Alzheimer's disease
Accidents
0
25
50
75
100
125
150
175
200
Death Rates per 100,000


Leading Causes of Death in the US (2013)
Diseases of the heart
Malignant neoplasms
Chronic Lower Respiratory Disease
Accidents
Cerebrovascular diseases
Alzheimer's disease
Diabetes mellitus
Influenza and pneumonia
Nephritis, nephrotic syndrome and nephrosis
Suicide
Septicimia
Chronic liver disease and cirrhosis
Essential hypertension and hypertensive renal disease
Parkinson's disease
Pneumonitis due to solids and liquids
0
25
50
75
100
125
150
175
200
Death Rates per 100,000

Leading Causes of Death per 100,000 in the US (2013)



Data from the Detailed Tables for the National Vital Statistics Report (NVSR) “Deaths: Final Data for 2013.” Retrieved from http://www.cdc.gov/nchs/data/nvsr/nvsr64/nvsr64_02.pdf

Selecting Quotations

- Length: ideally one line - a third of a page long.
- Choose quotations that are typical of the points that you are making.
- Ideal quotations are well expressed and striking – they provide the *feeling of listening to participants*

Presenting Quotes

"To invent, you need a good imagination and a pile of junk." — Thomas A. Edison

- For short quotes/visual displays: place line breaks at naturally occurring points
- Condensed font so the text is less spread out
- Add color on selected key words in the quote
- Photo of person being quoted or representative image

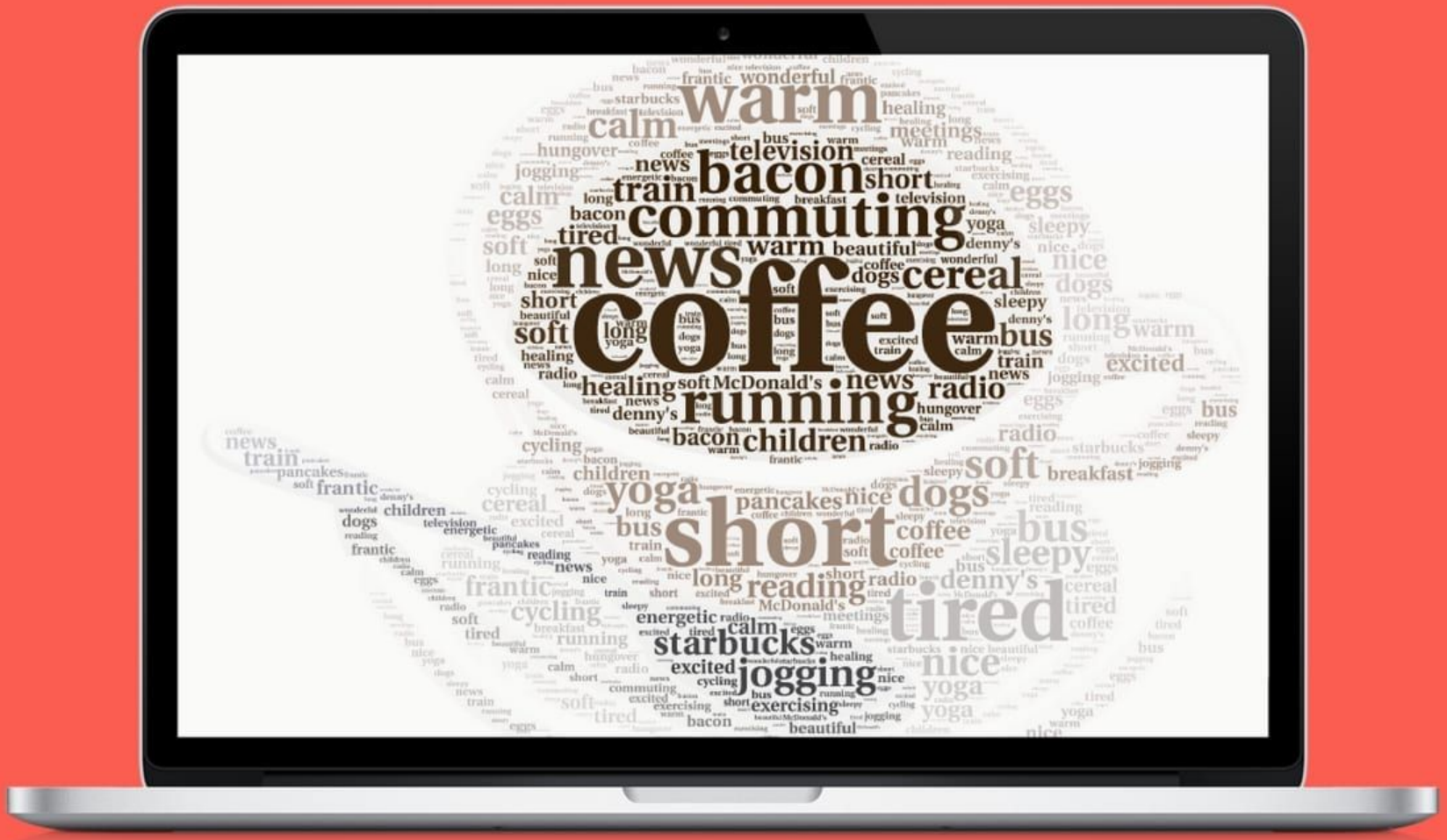
Visualizing Textual Themes

- Word clouds present words in a set (a document, corpus, a single interview, etc) scaled by their frequency of use

- Useful in giving the gist of a text and drawing out key commonalities

- Can quickly get confusing or muddle the point. Use strategically

More art than effective



But shape can add meaning

ColorBrewer2.org

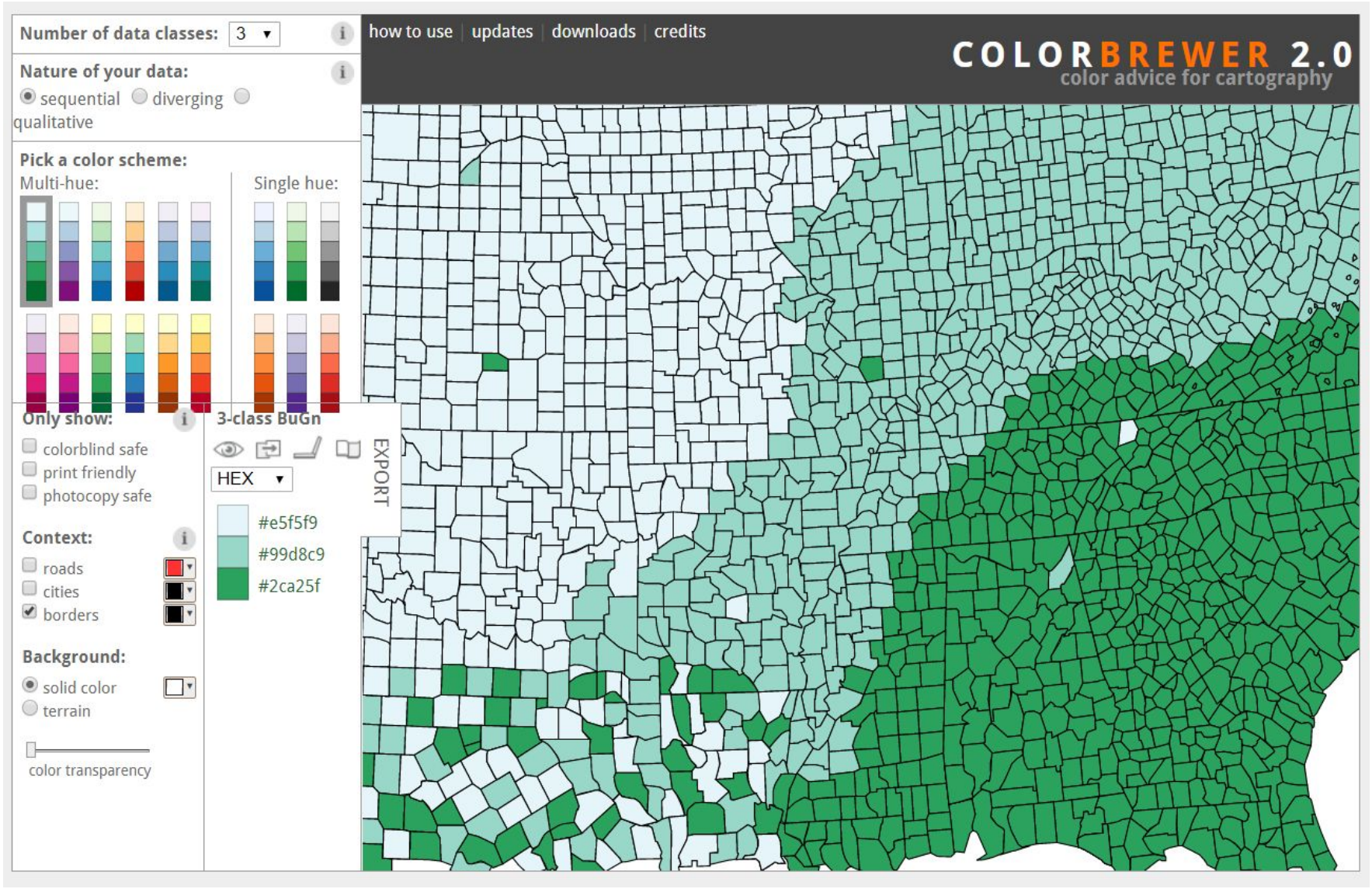


Chart Selection

Questions to ask about your data:

- What research question am I trying to visualize?
- What story am I trying to tell?
- How, and to whom, am I presenting my data?
- How many variables do I need to display?
- Is my data best conveyed as a visual?

Common Chart Types

Distribution: histogram, box, scatter, kernel density, Q-Q

Relationship: scatter, heatmap, radar, column/bar

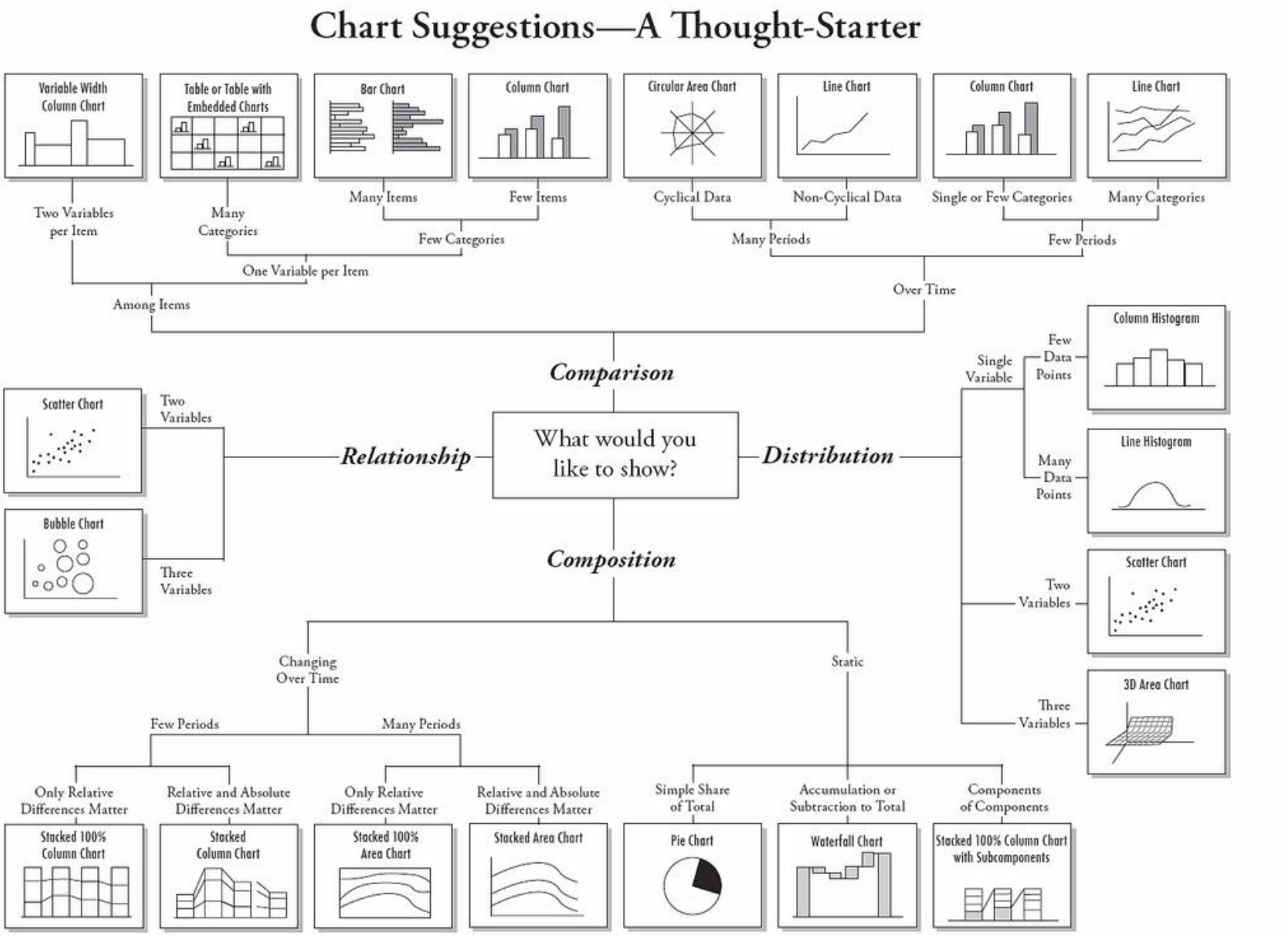
Temporal: line, column/bar, radar

Hierarchical: ring, tree

Network: node-link, alluvial

Spatial: choropleth, dot density

Chart Selection



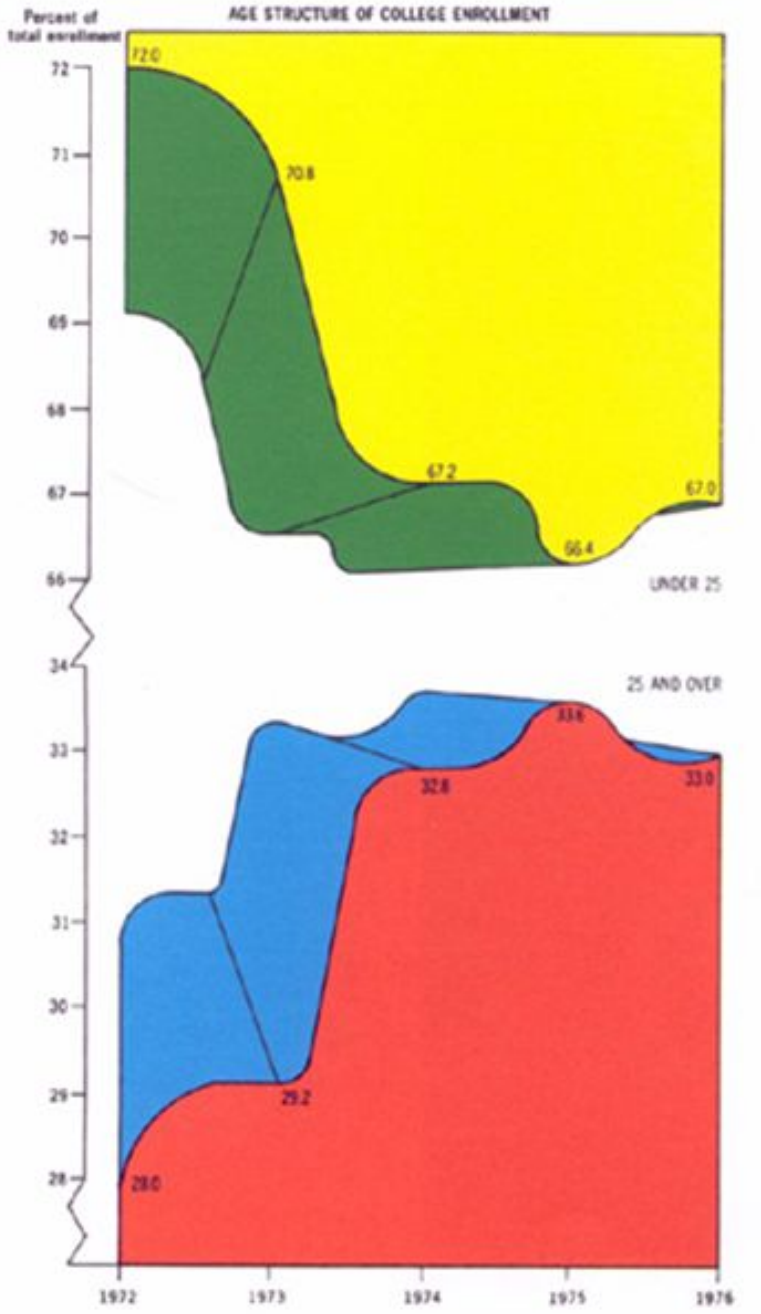
Visualizing Qualitative Data

Table 2

Types of Visual Displays and Purposes

Source: Verdinelli and Scagnoli, 2013

Visual display	Purpose
Boxed display	To highlight a specific narrative considered important and frame it in a box
Decision tree modeling	To describe options, decisions, and actions
Flow chart	To illustrate directional flow and show pathways of different groups
Ladder	To represent the dimensions of the progression of certain phenomenon through time or to show levels or stages
Matrix	To cross two or more dimensions, variables, or concepts of relevance to the topic of interest
Metaphorical visual display	To depict in a metaphorical way the topics or themes found
Modified Venn diagram	To indicate shared or overlapping aspects of a concept, a category, or a process
Network	To depict relationships between themes and subthemes or categories and subcategories
Taxonomy	To classify or organize information



The Data-Ink Ratio

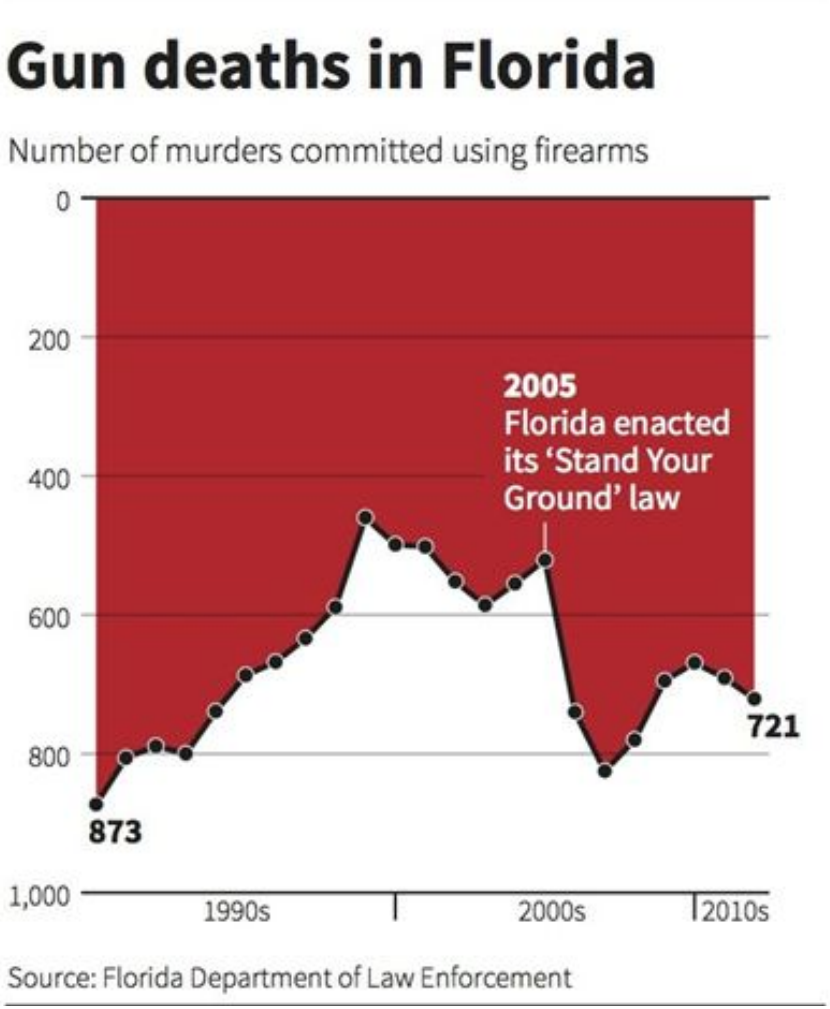
Ink used to represent the data

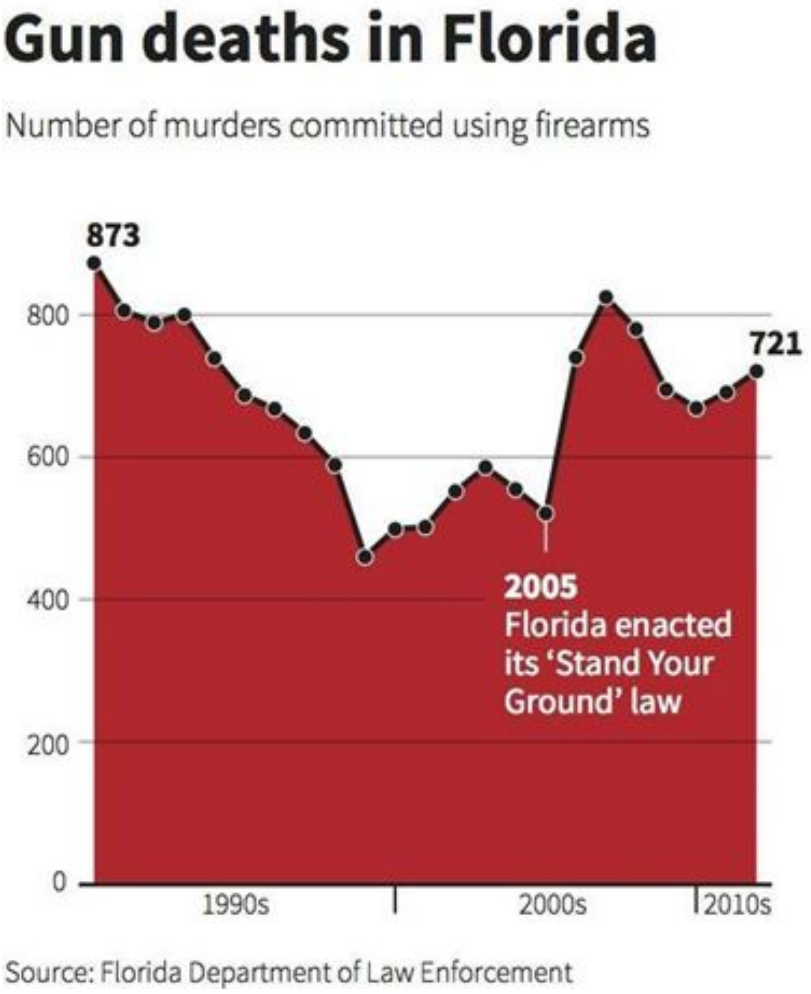
Ink used to print the graphic

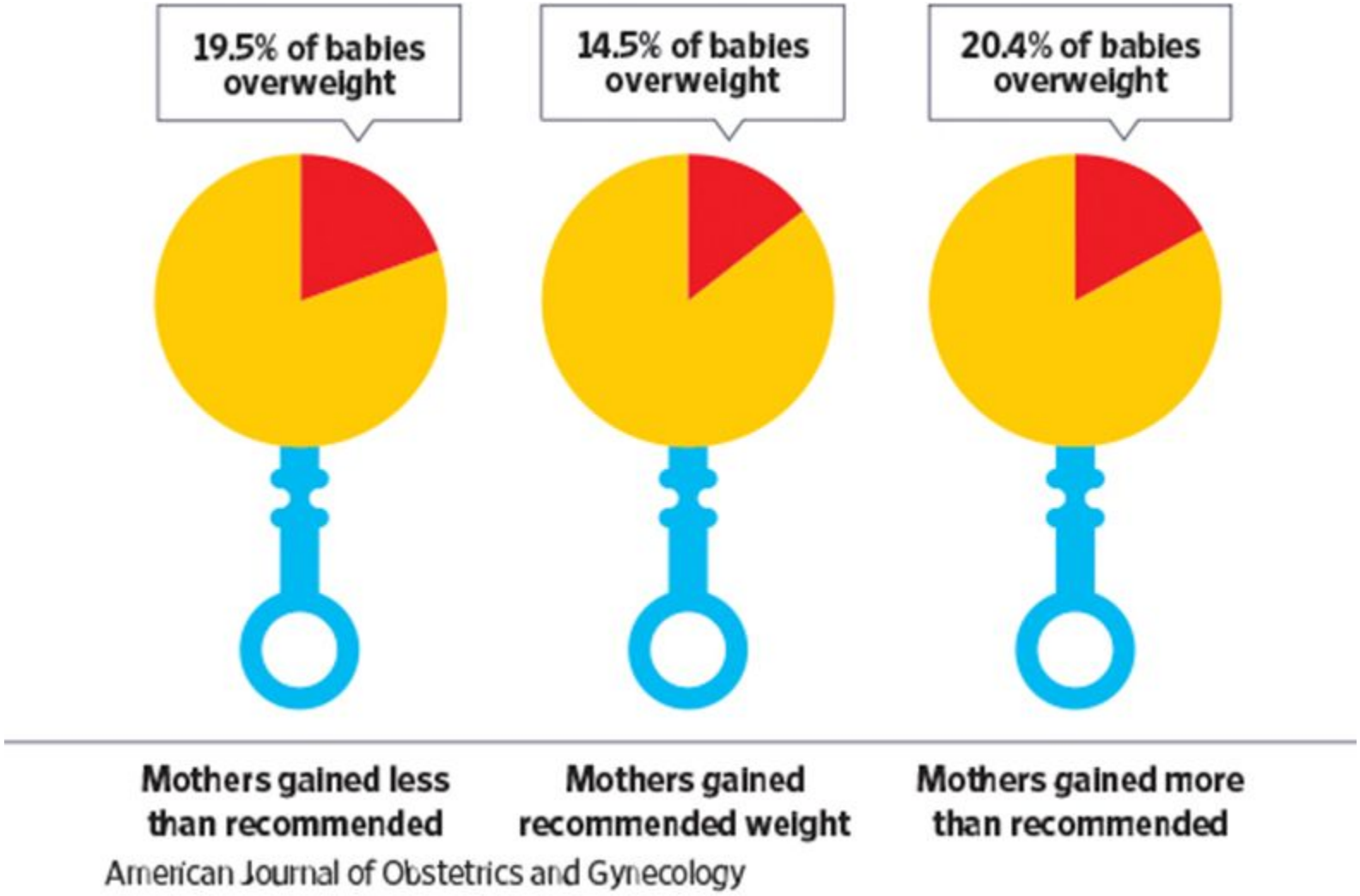
Chart Junk: “any element… that does not contribute to clarifying the intended message”

Tufte, E.R. (2001). The visual display of quantitative information. 2nd edition. Cheshire, CT: Graphics Press; Kosara, R. (2013). A better definition of chart junk. EagerEyes. Retrieved from https://eagereyes.org/blog/2013/definition-chart-junk

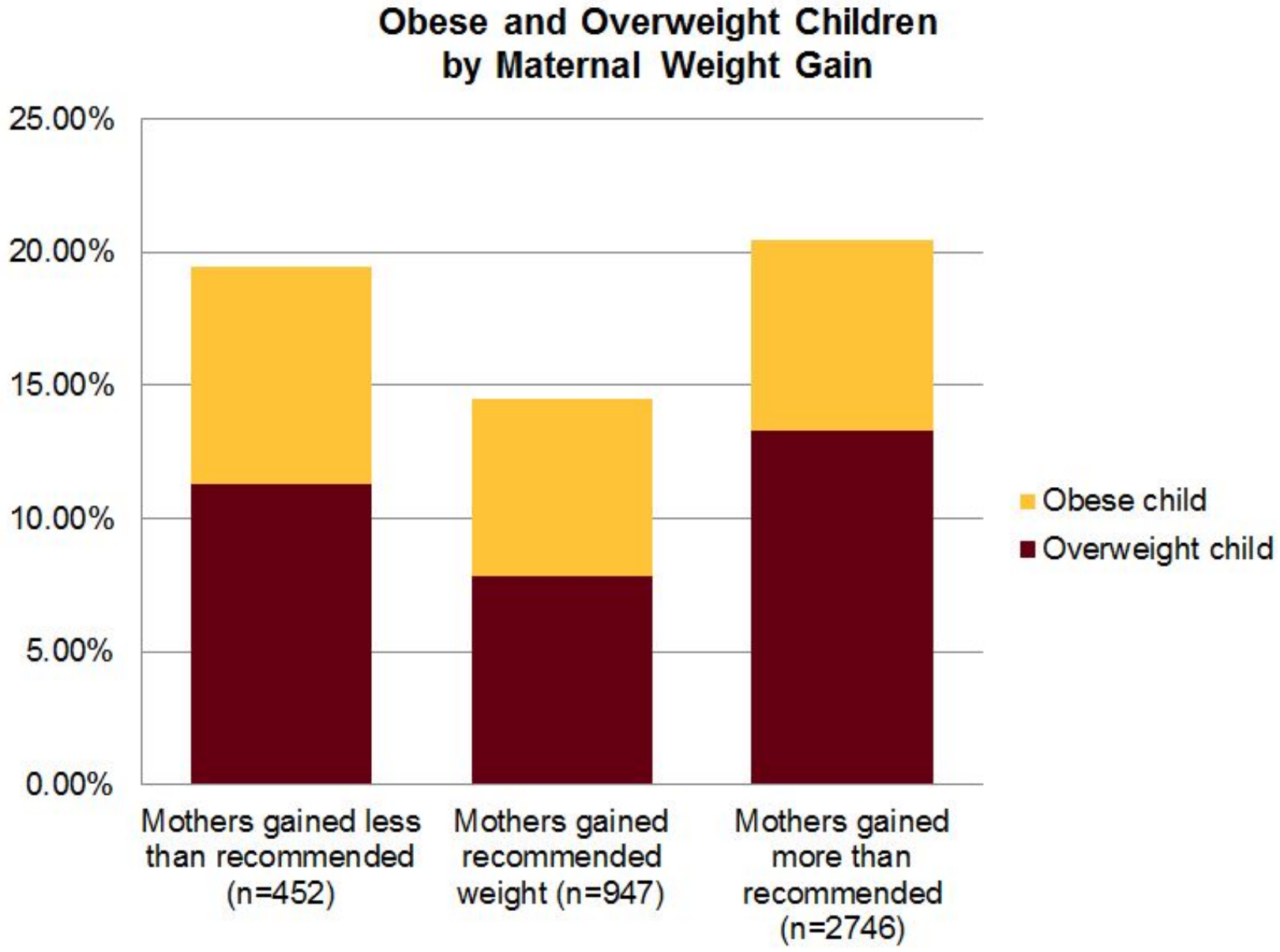
What Not To Do







Goldenberg, D. (2014). Statshot: CitiBike, baby weight, NBA playoffs. The Wall Street Journal. Retrieved from http://blogs.wsj.com/numbers/statshot-citibike-baby-weight-nba-playoffs-1335/



Data from Sridhar, S.B., Darbinian, J., Ehrlich, S.F., Markman, M.A., Gunderson, E.P., Ferrara, A., & Hedderson, M.M. (2014). Maternal gestational weight gain and offspring risk for childhood overweight or obesity. American Journal of Obstetrics and Gynecology, 211(3):259.e1-259.e8. doi:10.1016/j.ajog.2014.02.030.

Vizualization Tool Recommendations

Ggplot GUI (Online)

site.shinyserver.dck.gmw.rug.nl/ggplotgui

Tableau (free for students for 1 year, or public is always free)

tableau.com/academic/students

Wordle Word Clouds

wordle.net

Resources

Websites:

- matplotlib examples: http://matplotlib.org/examples/index.html
- ggplot2 docs: http://docs.ggplot2.org/current/
- https://www.tapclicks.com/the-ultimate-guide-to-data-visualization/

Books:

- Mcgreggor, Duncan M. *Mastering matplotlib.* Packt, 2015.
- Wickham, Hadley. *ggplot2: Elegant Graphics for Data Analysis*, Springer, 2016.
- Tufte, Edward. *The Visual Display of Quantitative Information, 2nd ed*. Graphics Press, 2001.

Video:

- Olson, Randy. *Data Visualization Basics with Python.* O'Reilly, 2016.

On Campus:

- LATIS: latisresearch.umn.edu
- Libraries: https://www.lib.umn.edu/datamanagement