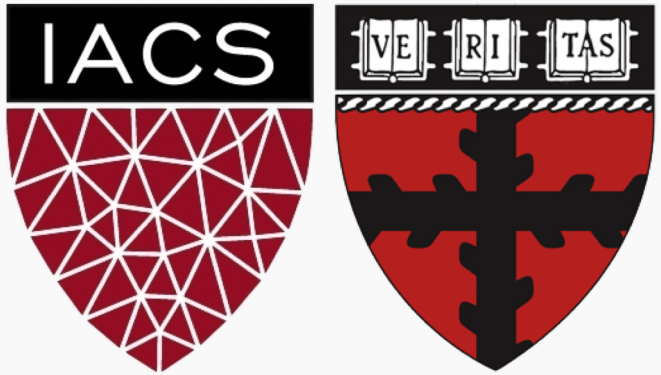
Lecture 9: Visualization for EDA and Communication

CS109A Introduction to Data Science

Pavlos Protopapas, Kevin Rader and Chris Tanner



As the matplotlib thickens …

ANNOUNCEMENTS

Projects have been posted!

- Check the website for details and read the guidelines file for due dates

Added more Office Hours!

- Check the weekly schedule on our website

Lecture Outline

- EDA Refresher
- Effective Visualization
  - Graphical Integrity
  - Scope
  - Displays
  - Sensible Design
- Communication
  - Motivation
  - Key Considerations

Lecture Outline

- EDA Refresher
- Effective Visualization
  - Graphical Integrity
  - Scope
  - Displays
  - Sensible Design
- Communication
  - Motivation
  - Key Considerations

EDA Refresher

Assume you know a given dataset is credible, complete with the info you want, and has no missing values. Why do further EDA?

EDA Refresher

Purposes of EDA:

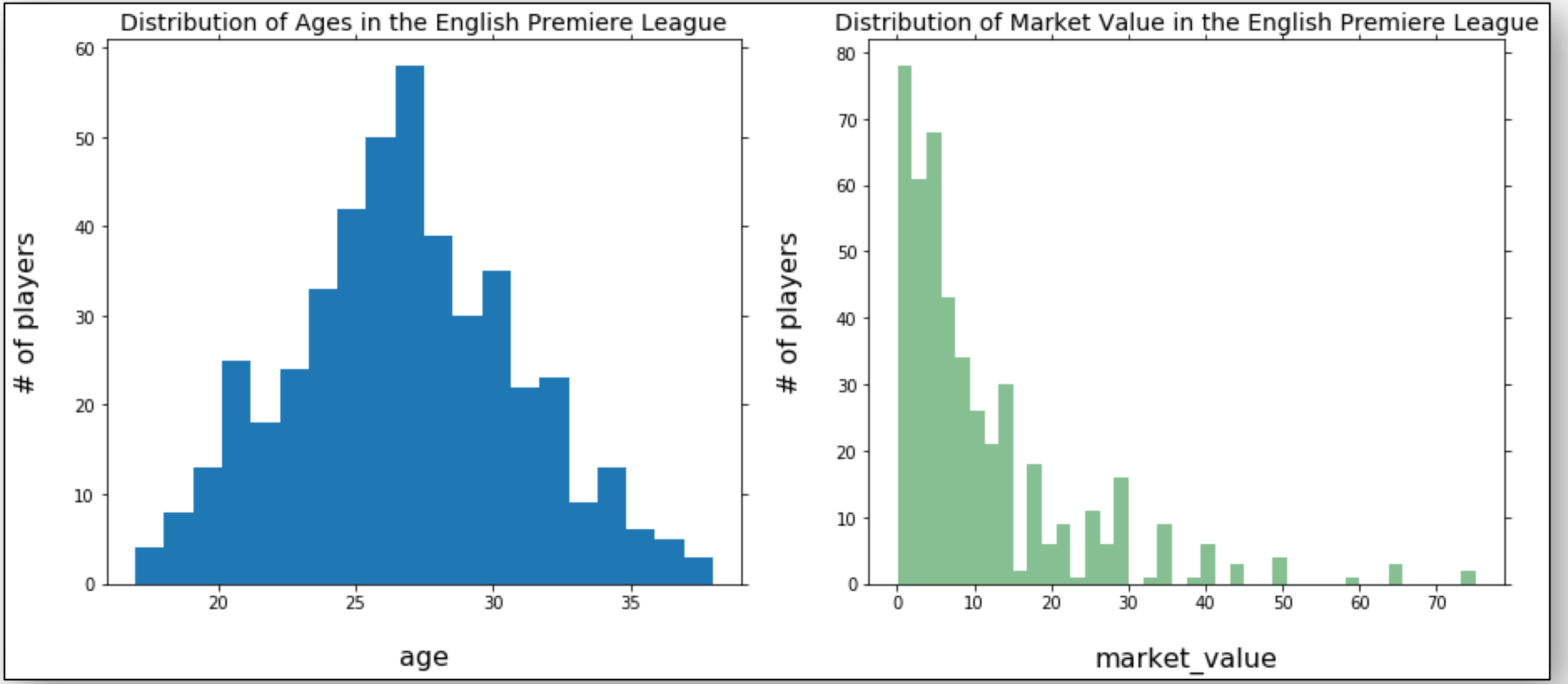
- Maximize insight into a dataset
- Uncover underlying structure
- Detect outliers
- Test underlying assumptions
- Develop parsimonious models

EDA Refresher: English Premier League

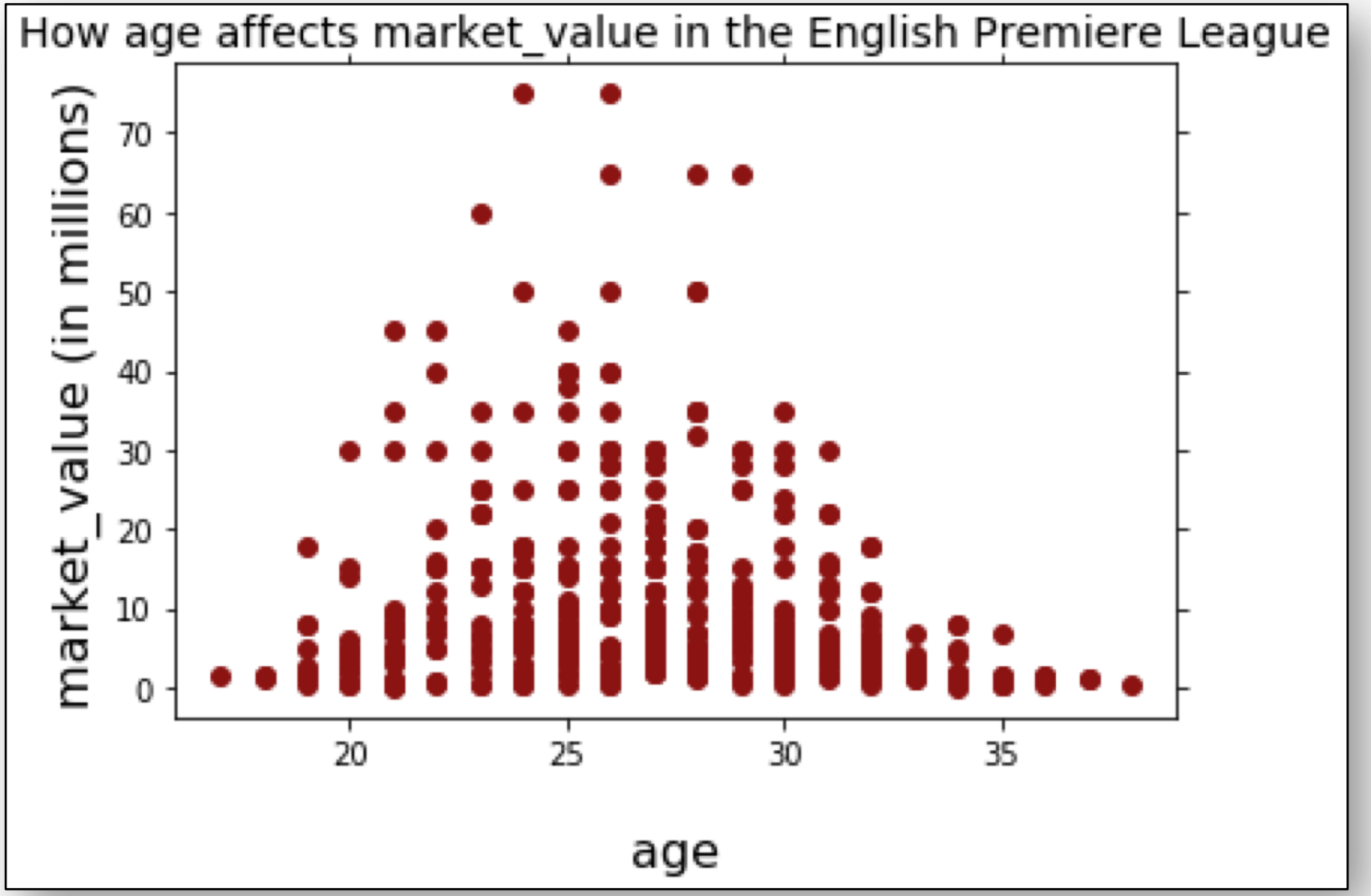
name	club	age	position	market value
Alexis Sanchez	Arsenal	28	LW	65
Mesut Ozil	Arsenal	28	AM	50
Petr Cech	Arsenal	35	GK	7
Theo Walcott	Arsenal	28	RW	20
Laurent Koscielny	Arsenal	31	CB	22

from www.transfermarkt.us

EDA Refresher: English Premier League

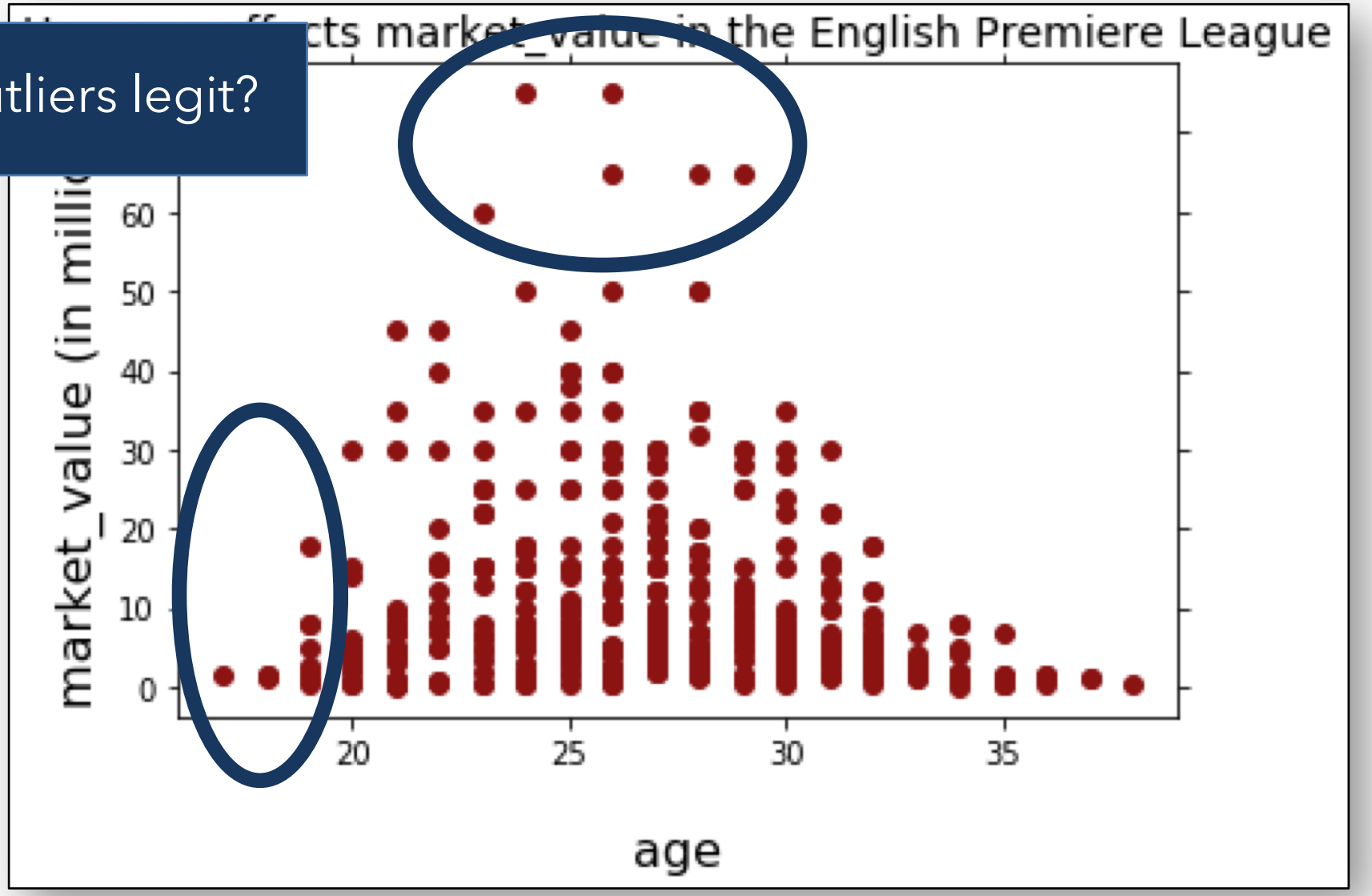


EDA Refresher: English Premier League



EDA Refresher: English Premier League

Are the outliers legit?



EDA Refresher: English Premier League

```
league_df.loc[league_df['age']<20][['name', 'club', 'age', 'position', \
    'market_value']].sort_values(by="age")
```

	name	club	age	position	market_value
233	Ben Woodburn	Liverpool	17	LW	1.50
231	Trent Alexander-Arnold	Liverpool	18	RB	1.50
350	Josh Tymon	Stoke+City	18	LB	1.00
435	Jonathan Leko	West+Brom	18	RW	1.50
147	Tom Davies	Everton	19	CM	8.00
155	Ademola Lookman	Everton	19	LW	5.00
239	Dominic Solanke	Liverpool	19	CF	2.00
270	Marcus Rashford	Manchester+United	19	CF	18.00
281	Axel Tuanzebe	Manchester+United	19	CB	1.00
282	Timothy Fosu-Mensah	Manchester+United	19	DM	2.50
375	Tammy Abraham	Swansea	19	CF	8.00
436	Sam Field	West+Brom	19	CM	0.25

EDA Refresher: English Premier League

```
league_df.loc[league_df['market_value']>=60][['name', 'club', \
    'age', 'position', 'market_value']].sort_values(by="age")
```

	name	club	age	position	market_value
377	Harry Kane	Tottenham	23	CF	60.0
263	Paul Pogba	Manchester+United	24	CM	75.0
92	Eden Hazard	Chelsea	26	LW	75.0
240	Kevin De Bruyne	Manchester+City	26	AM	65.0
0	Alexis Sanchez	Arsenal	28	LW	65.0
241	Sergio Aguero	Manchester+City	29	CF	65.0

EDA Refresher: English Premier League

```
league_df.loc[league_df['market_value']>=60][['name', 'club', \
    'age', 'position', 'market_value']].sort_values(by="age")
```

	name	club	age	position	market_value
377	Harry Kane	Tottenham	23	CF	60.0
263	Paul Pogba	Manchester+United	24	CM	75.0
92	Eden Hazard	Chelsea	26	LW	75.0
240	Kevin De Bruyne	Manchester+City	26	AM	65.0
0	Alexis Sanchez	Arsenal	28	LW	65.0
241	Sergio Aguero	Manchester+City	29	CF	65.0

Lecture Outline

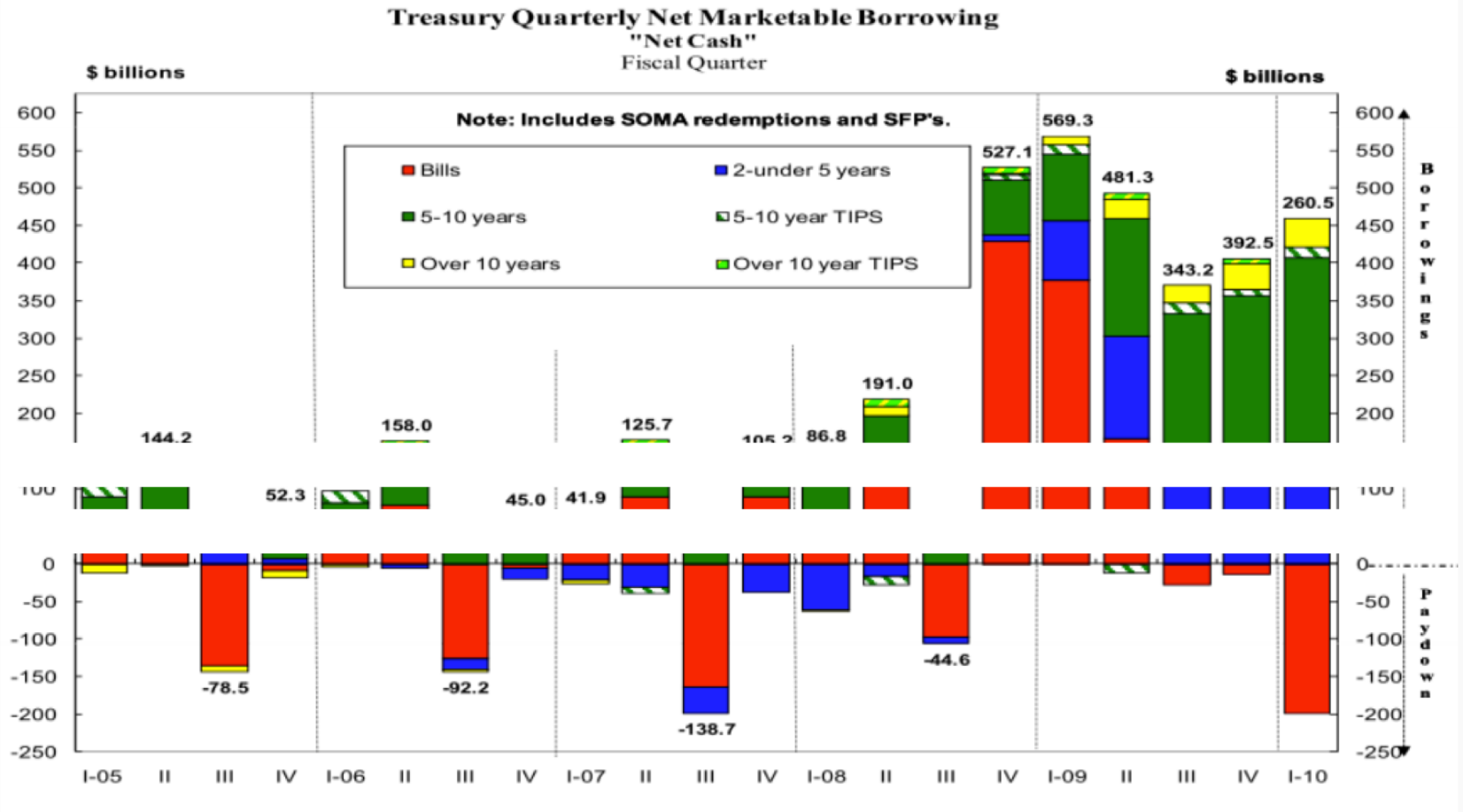
- EDA Refresher
- Effective Visualization
  - Graphical Integrity
  - Scope
  - Displays
  - Sensible Design
- Communication
  - Motivation
  - Key Considerations

Lecture Outline

- EDA Refresher
- Effective Visualization
  - Graphical Integrity
  - Scope
  - Displays
  - Sensible Design
- Communication
  - Motivation
  - Key Considerations

Effective Visualization

Not effective:



Effective Visualization

Not effective:

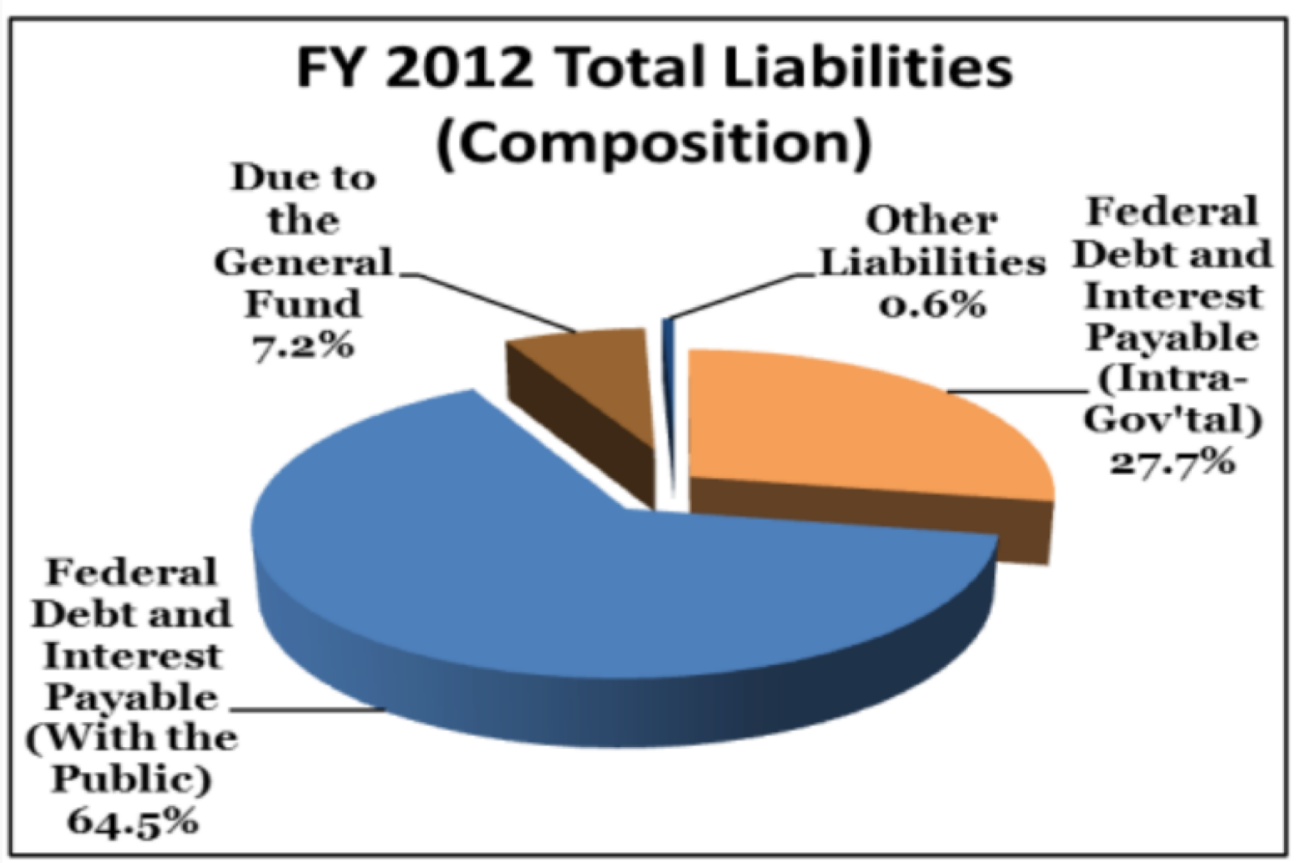
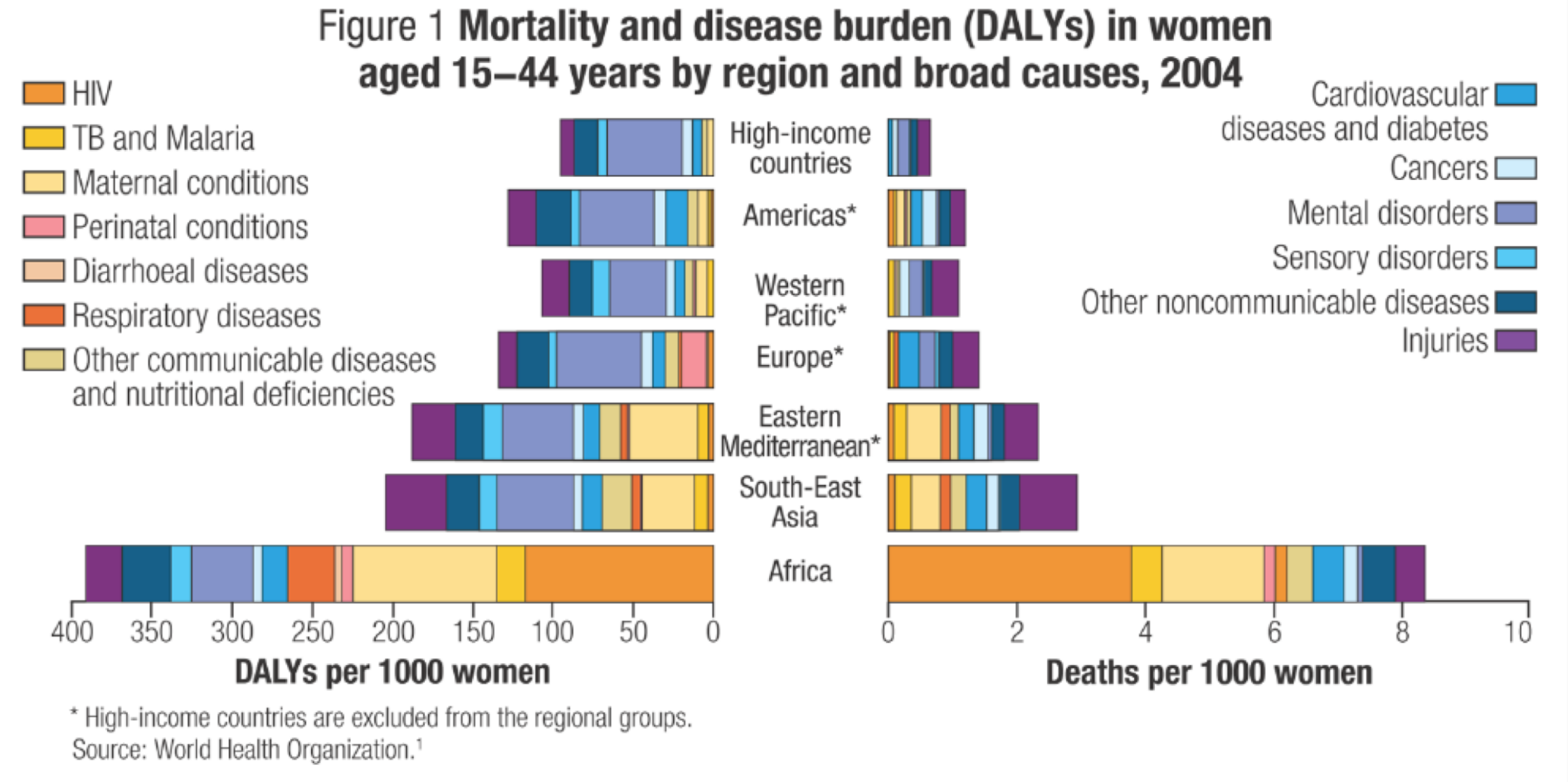


Figure 10

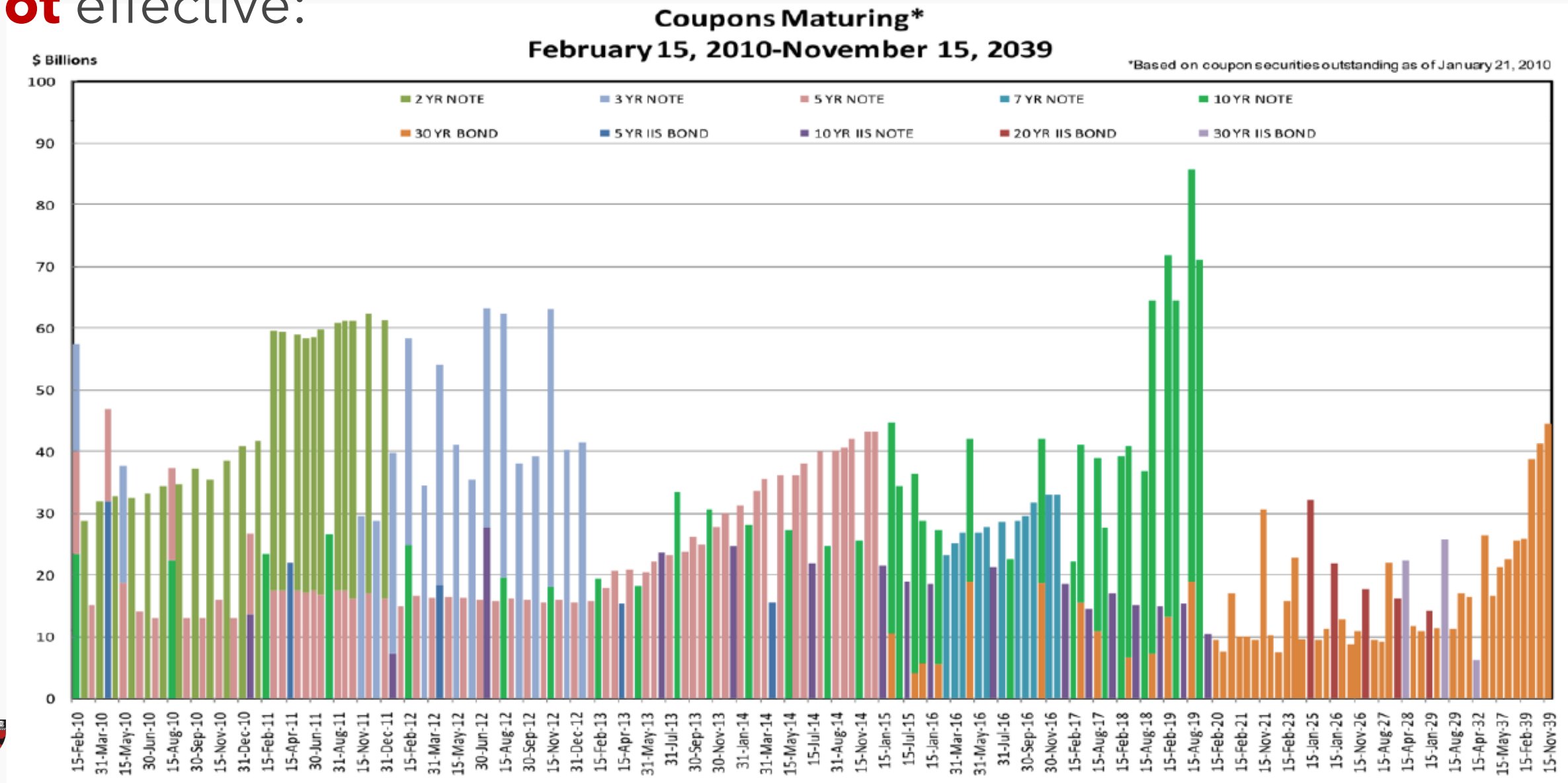
Effective Visualization

Not effective:



Effective Visualization

Not effective:



Effective Visualization

Not effective:

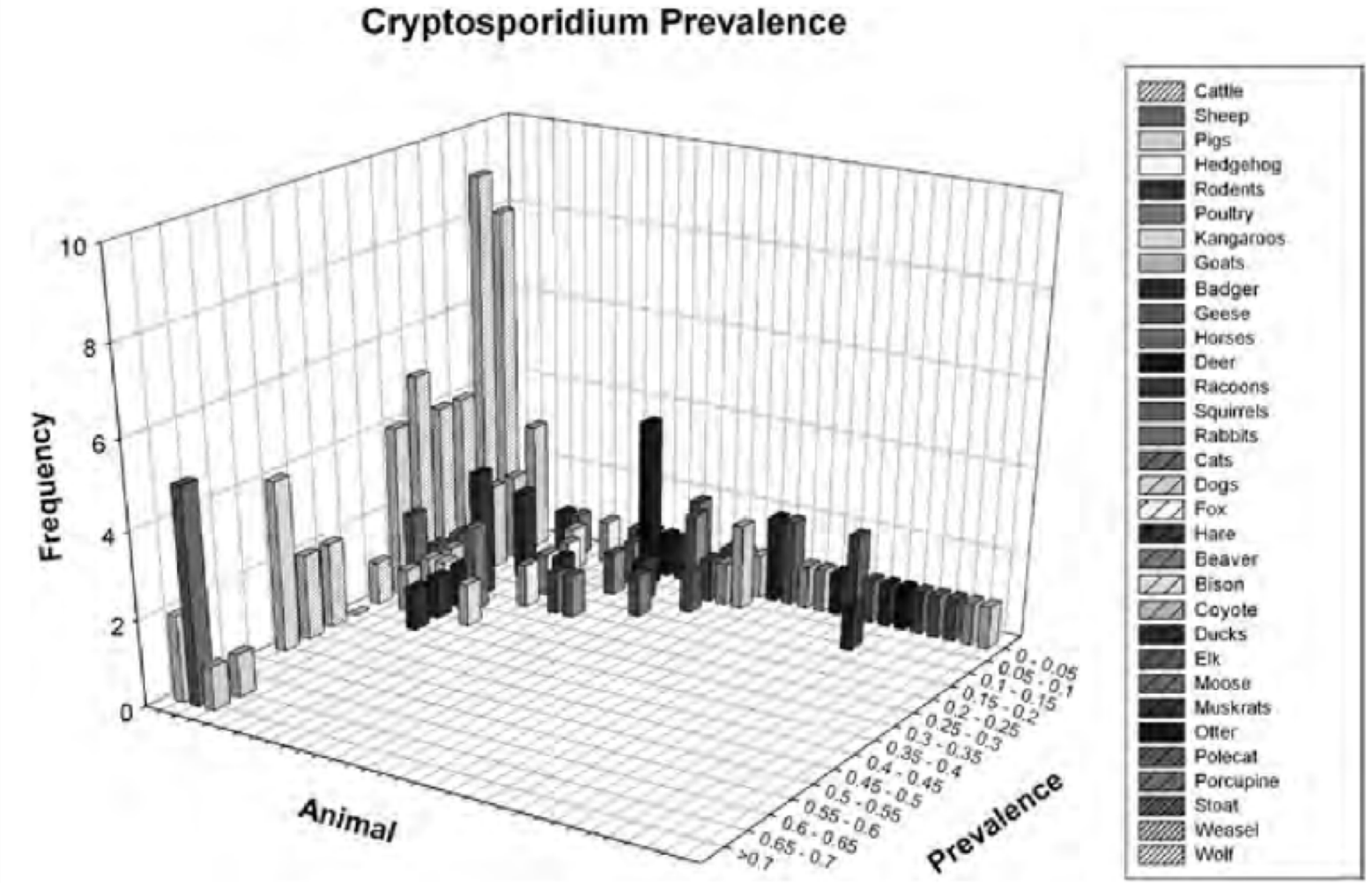


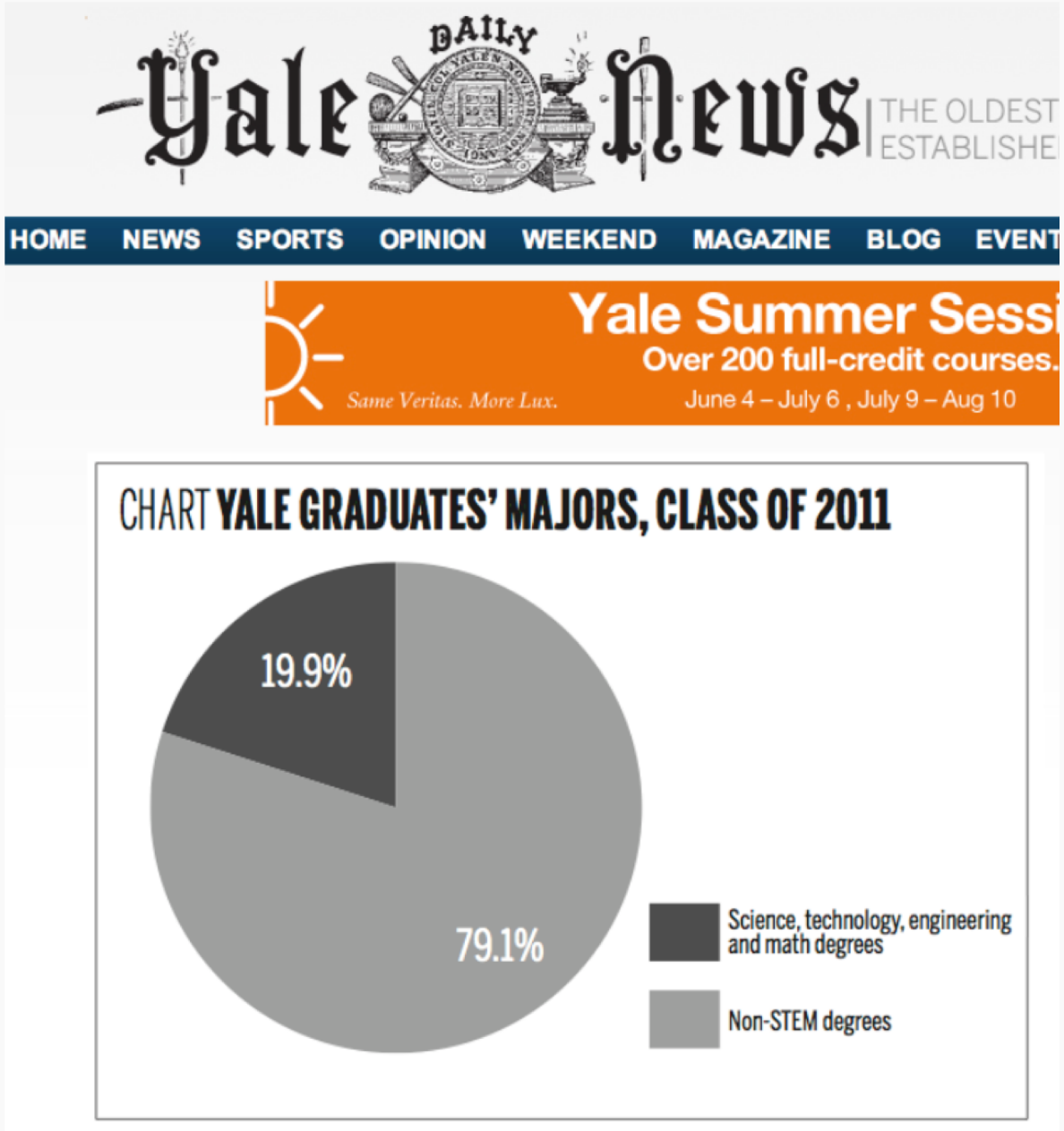
Figure 5.2 Mean prevalence rates of *Cryptosporidium* oocysts by animal species.

Lecture Outline

- EDA Refresher
- Effective Visualization
  - Graphical Integrity
  - Scope
  - Displays
  - Sensible Design
- Communication
  - Motivation
  - Key Considerations

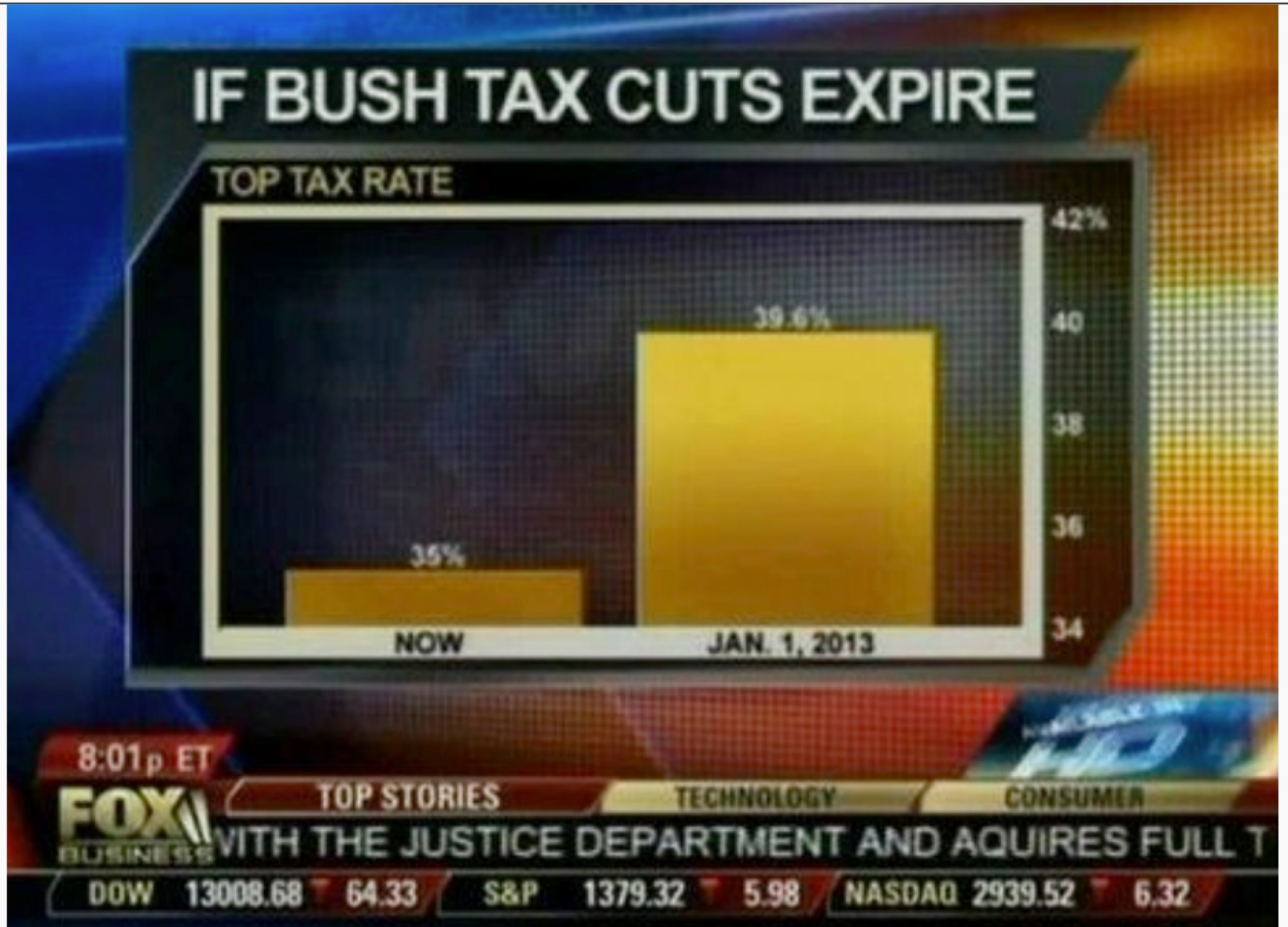
Graphical Integrity

What's wrong?

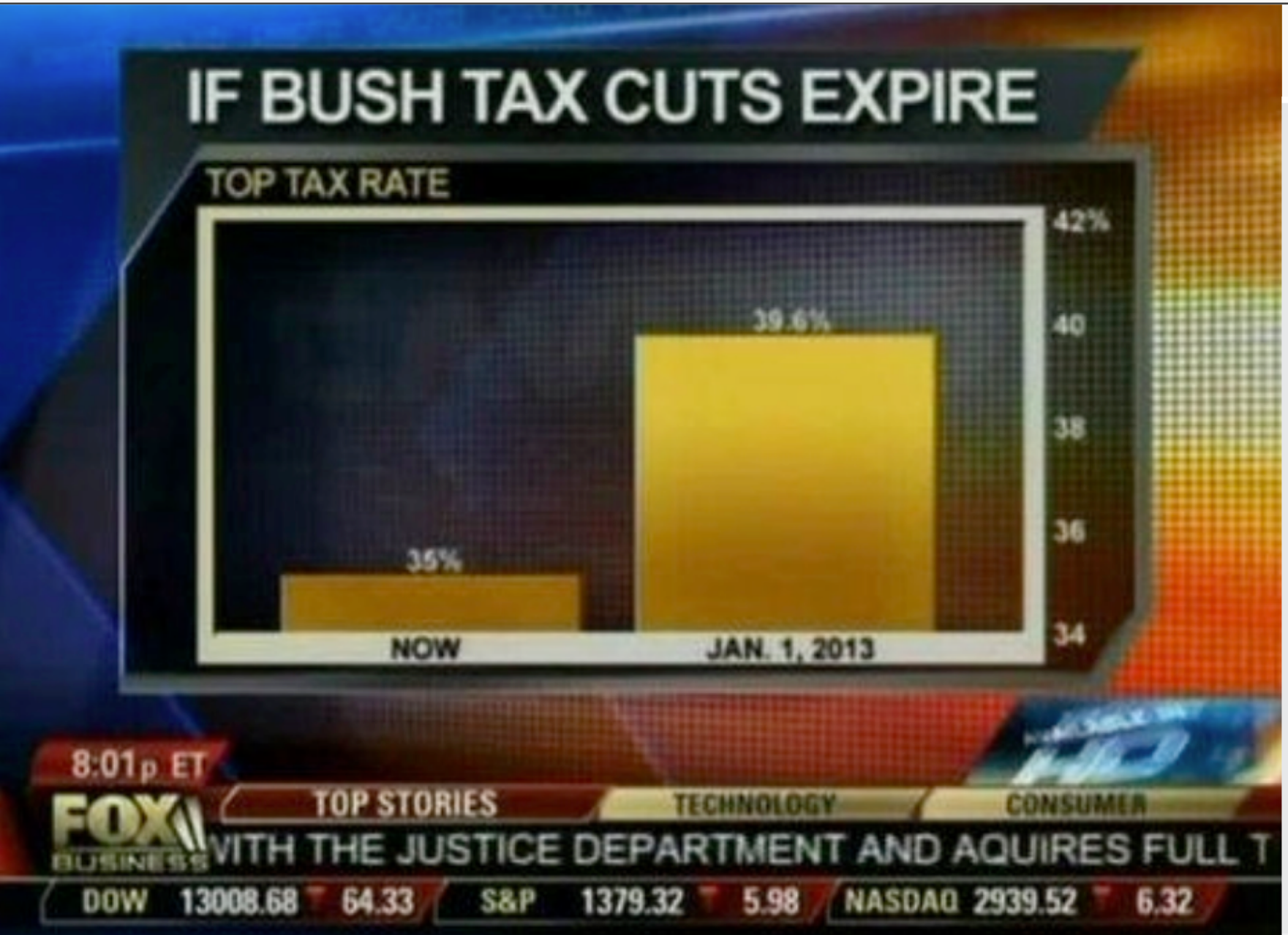


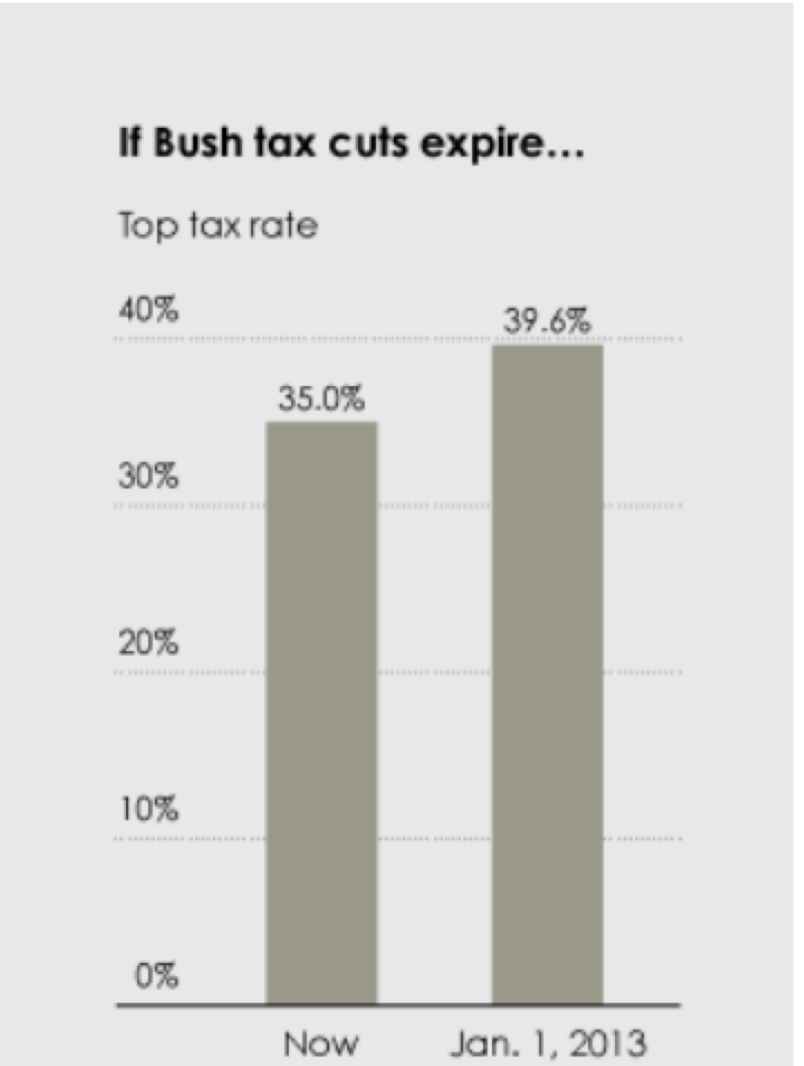
Graphical Integrity

What's wrong?



Graphical Integrity

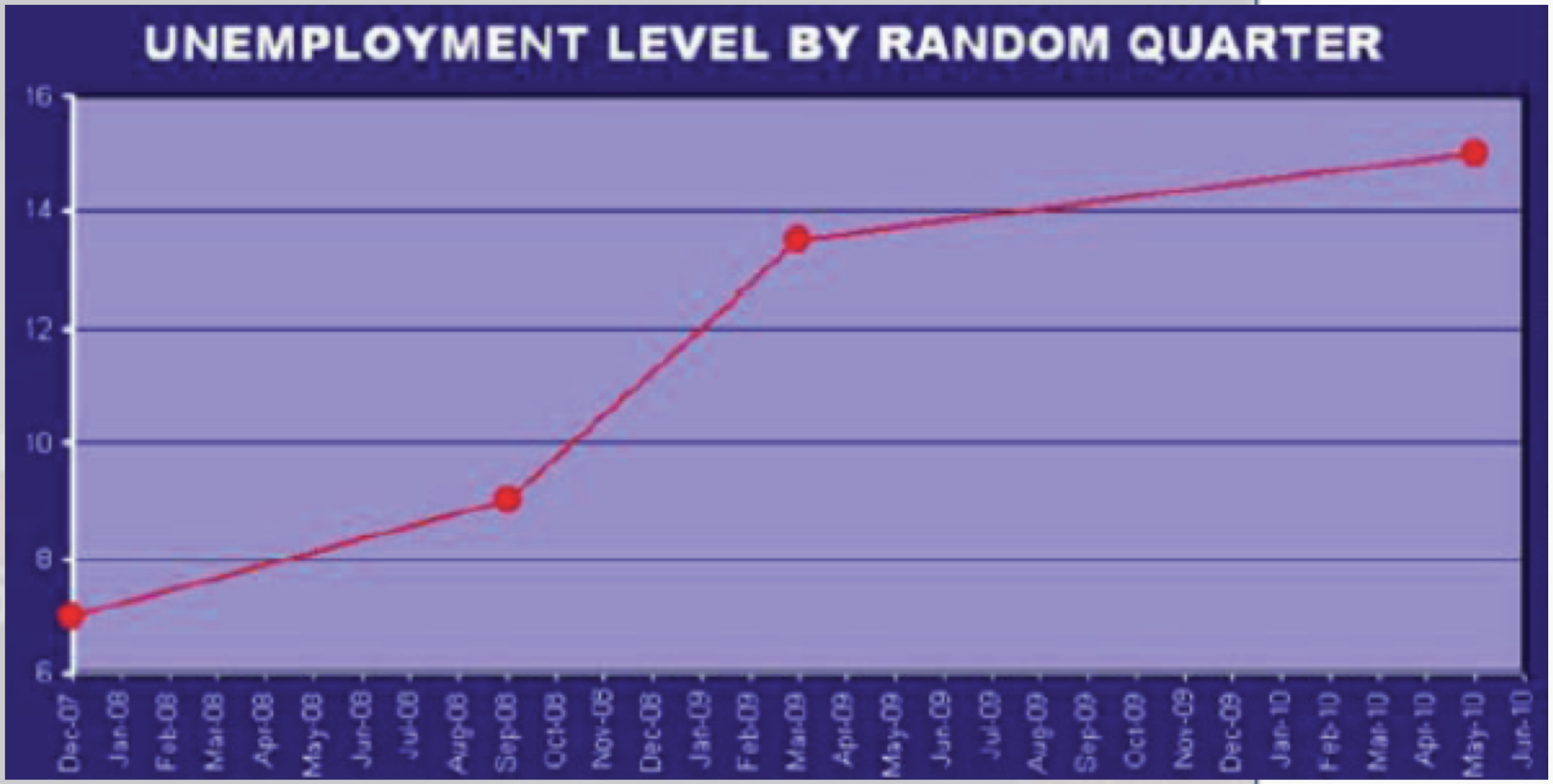




Graphical Integrity

What's wrong?

Graphical Integrity



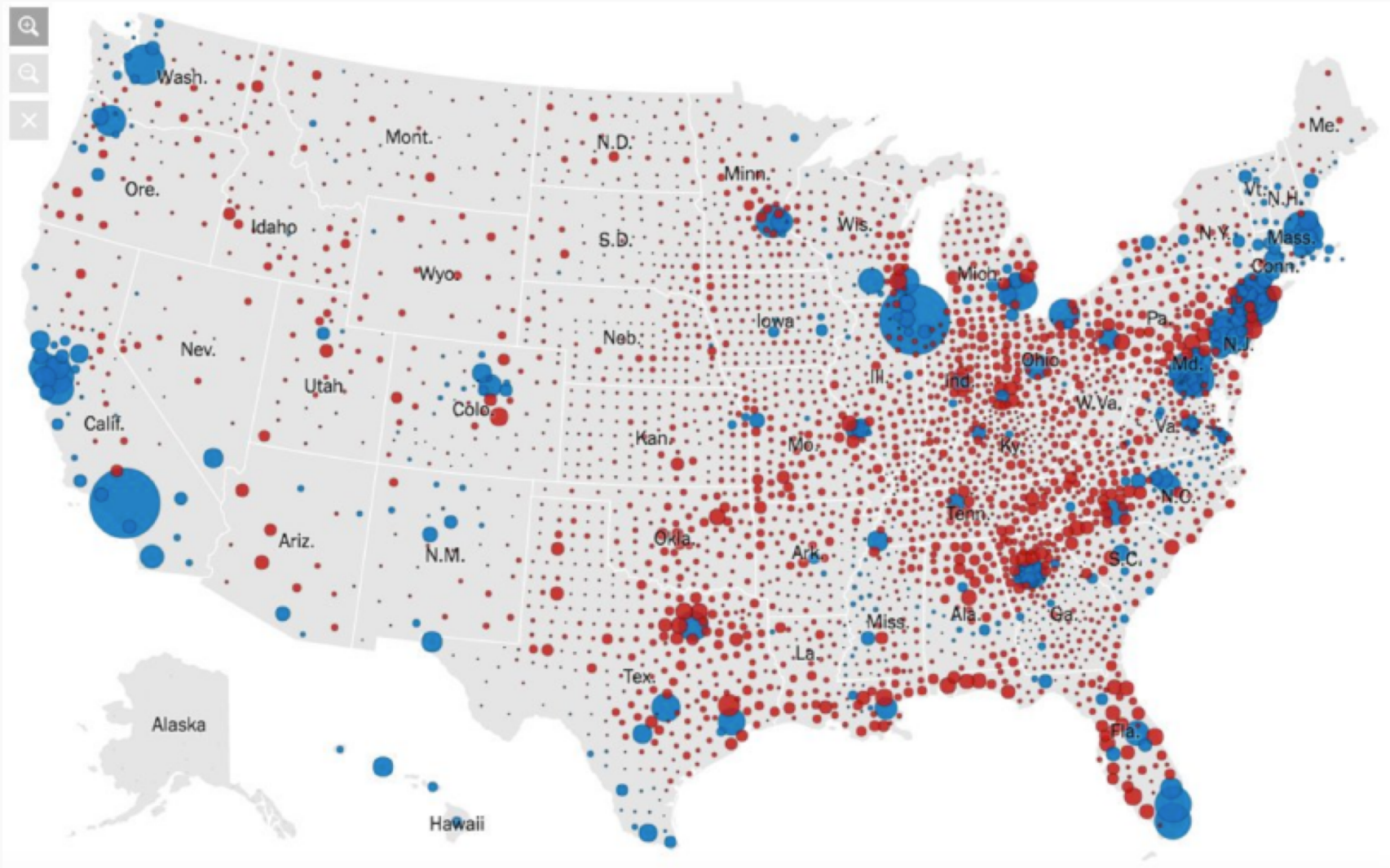
Graphical Integrity

What's wrong?

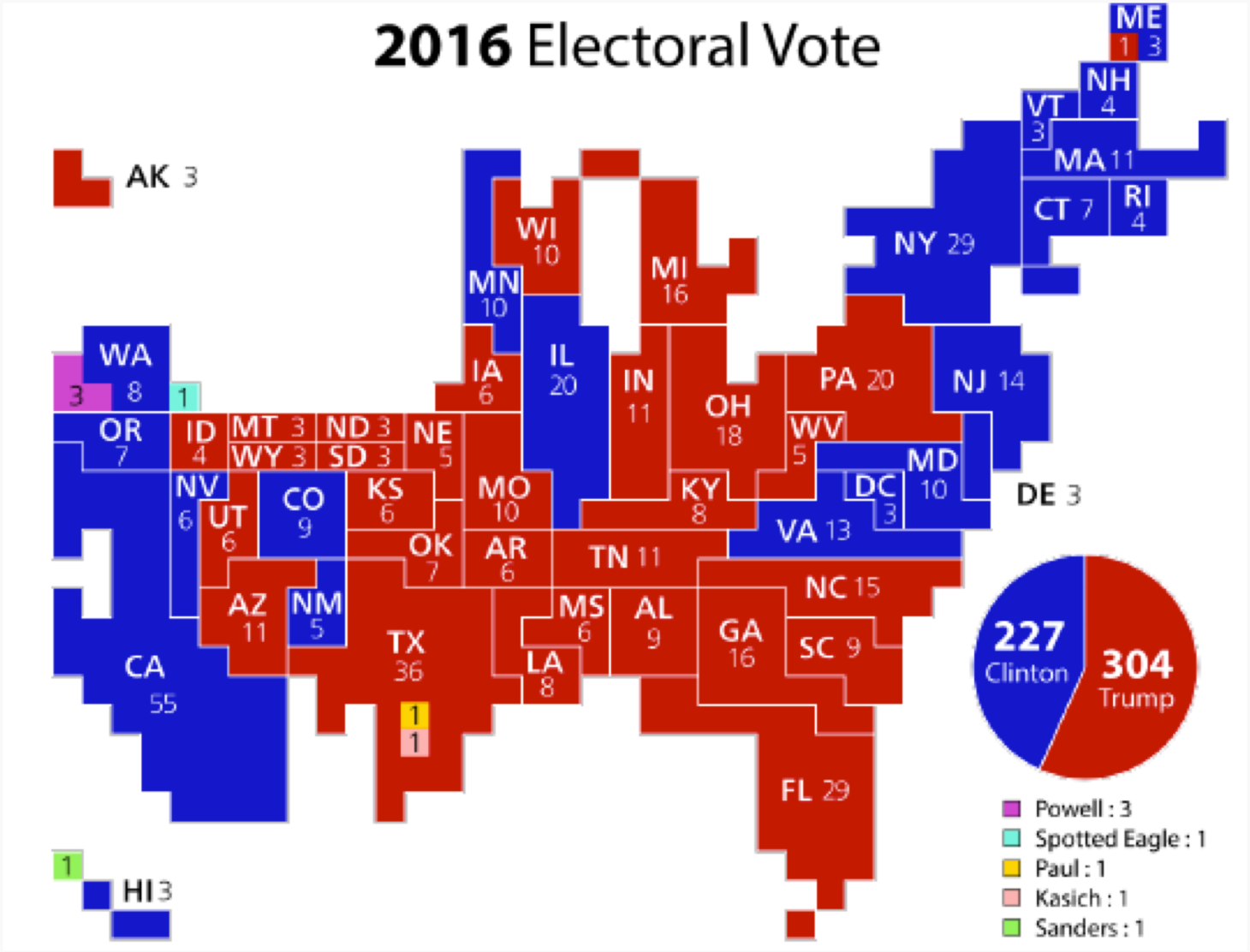
October 1, 2019

Graphical Integrity

Graphical Integrity



Graphical Integrity

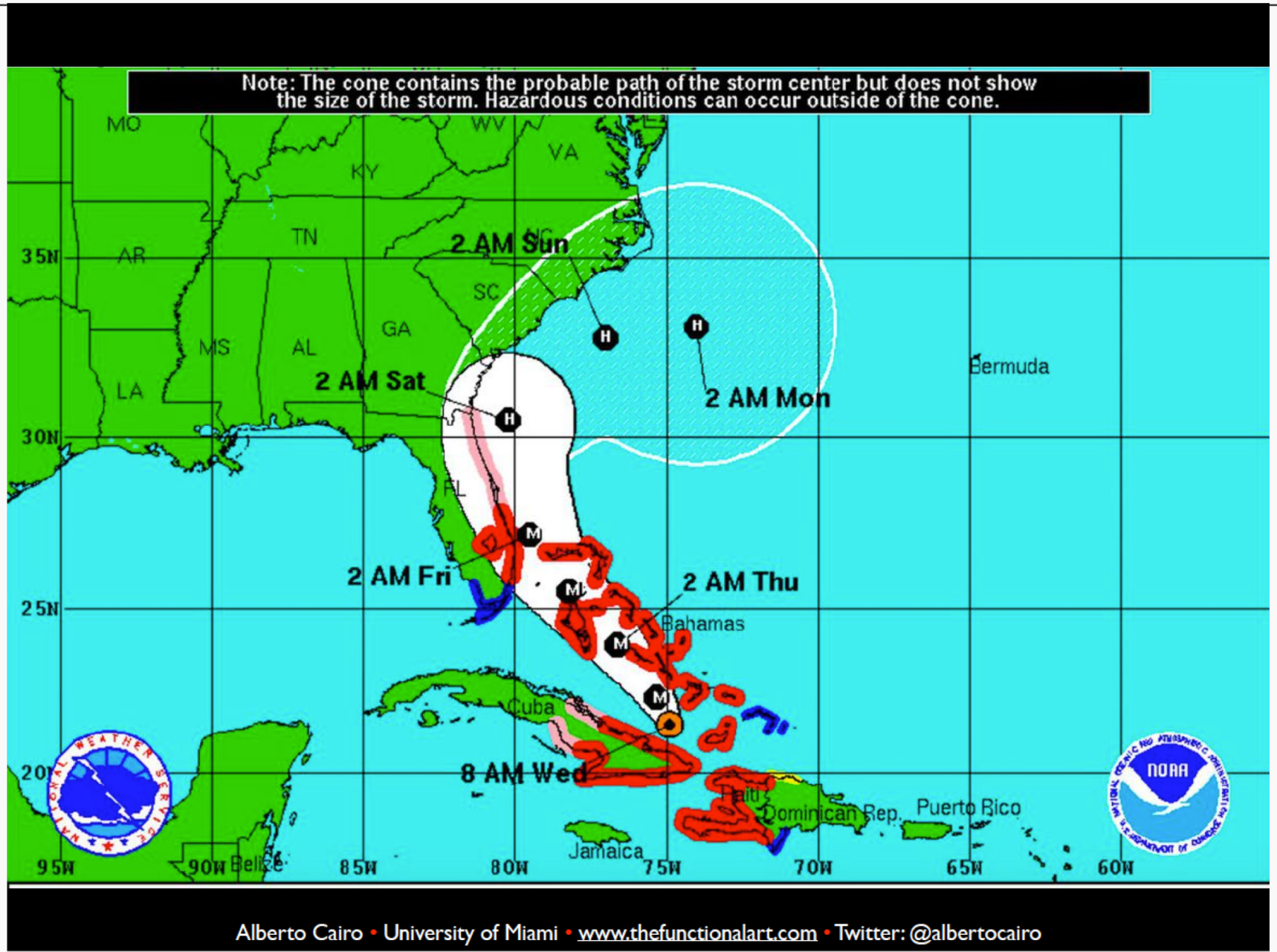


Graphical Integrity

Lesson: be proportional

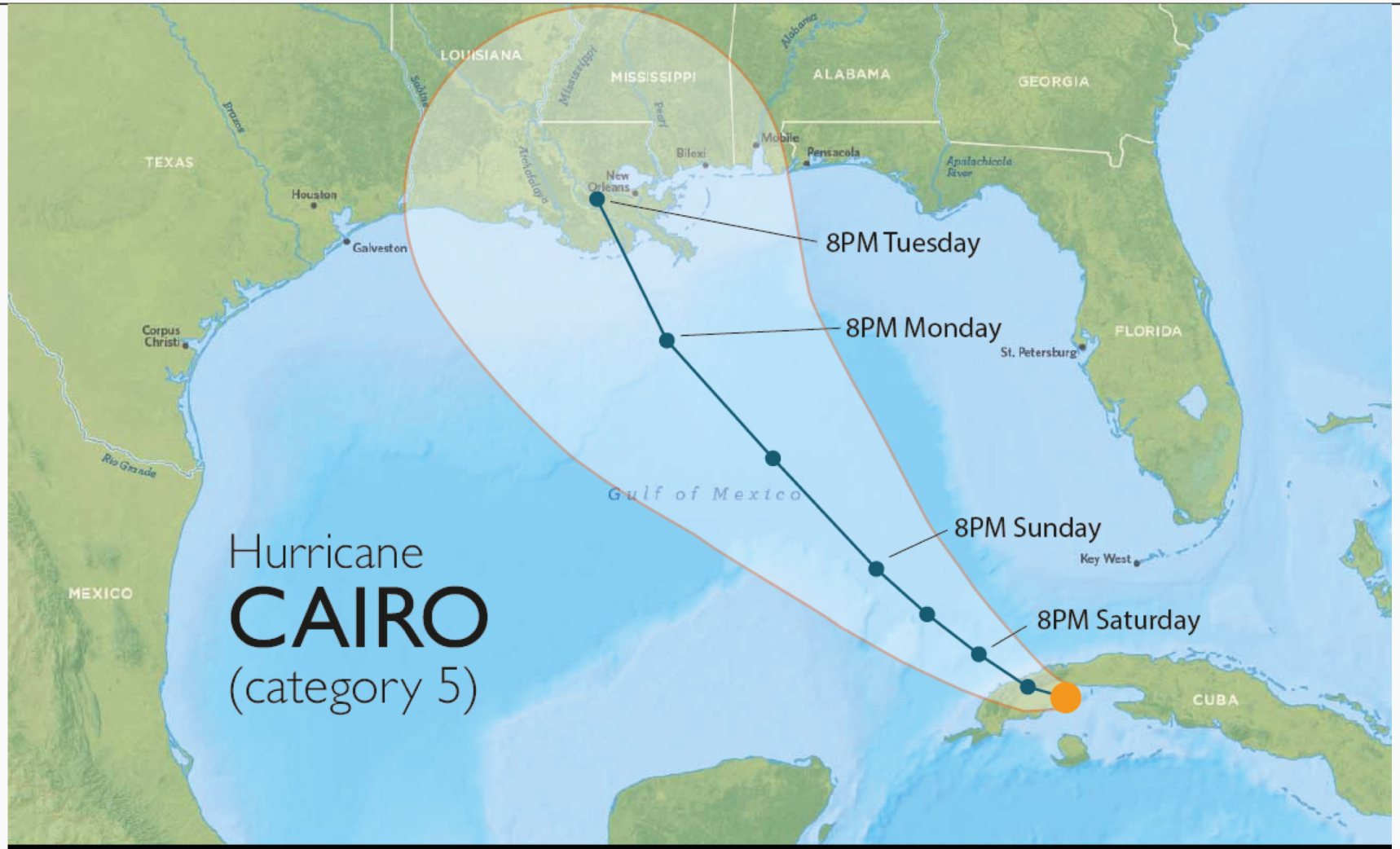
Graphical Integrity

What's wrong?



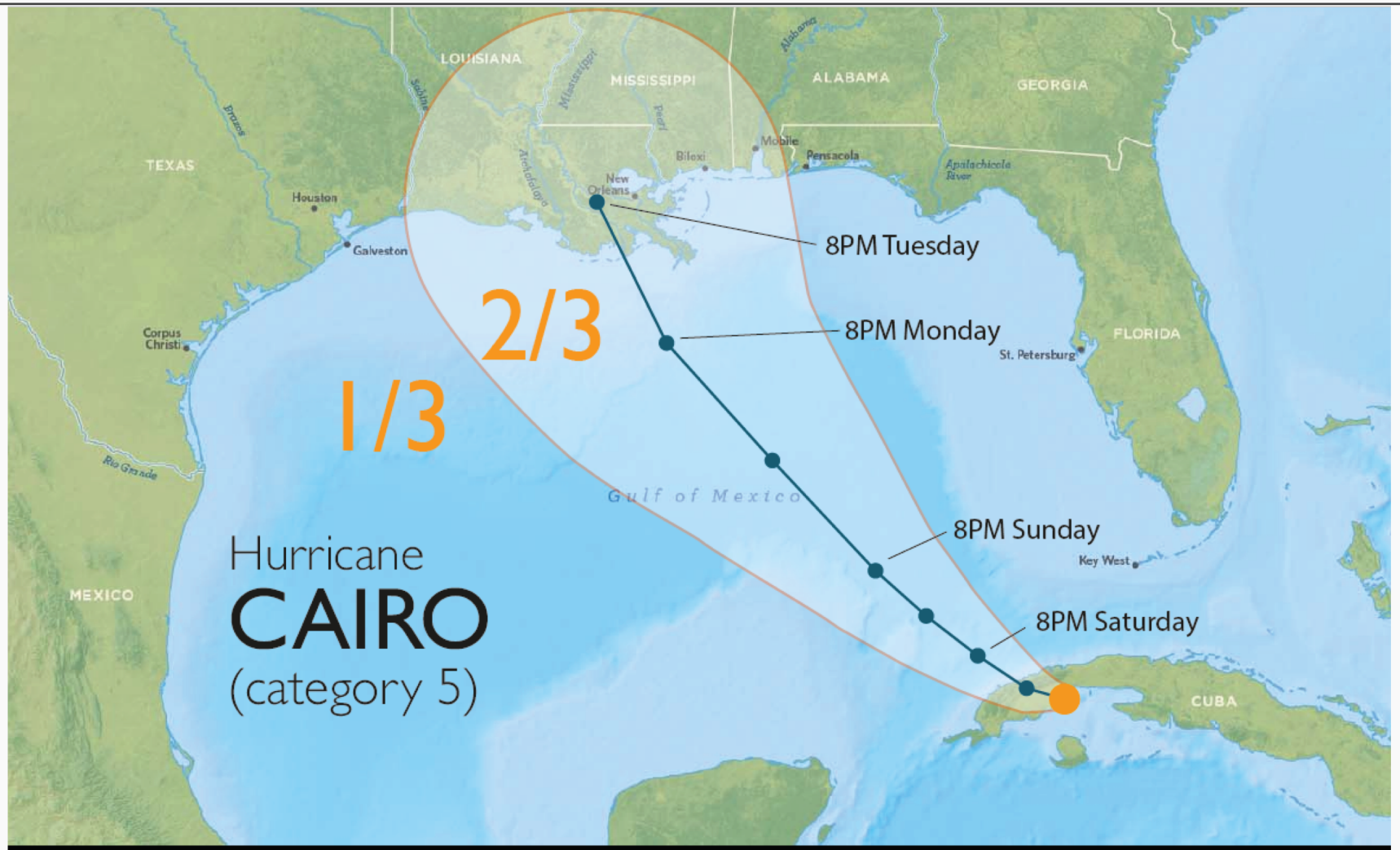
Graphical Integrity

What's wrong?



Graphical Integrity

What's wrong?



Graphical Integrity

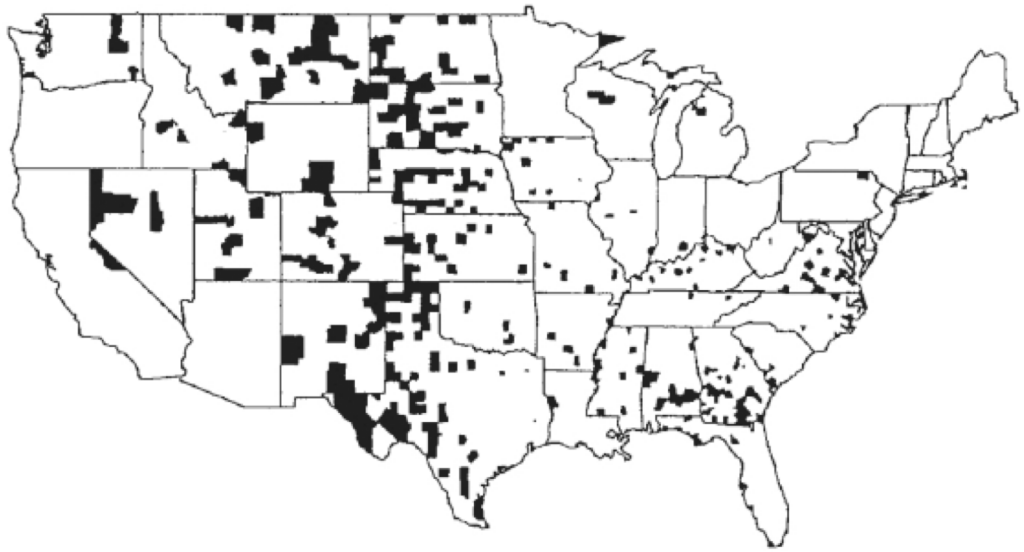
What's wrong?

Graphical Integrity

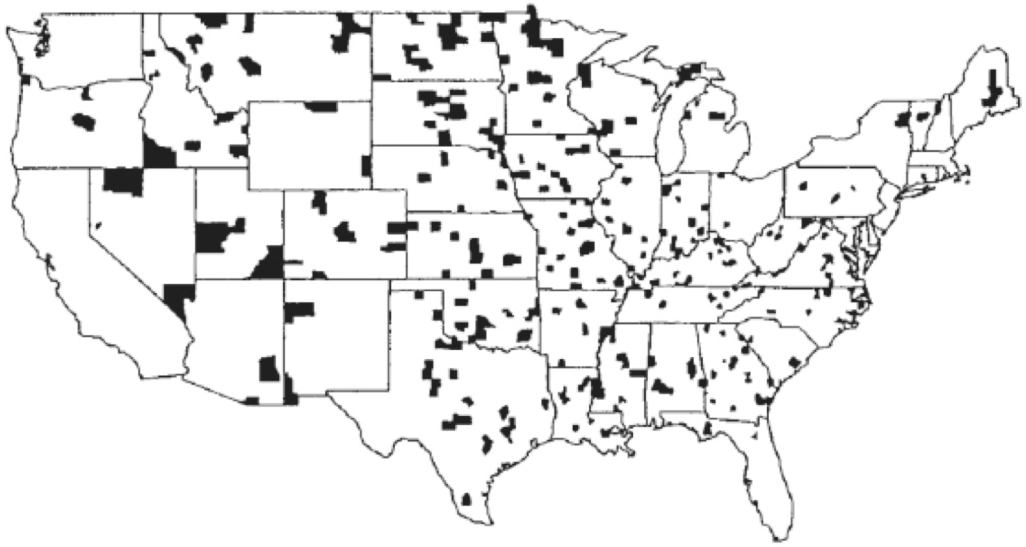
Lesson: include uncertainty

Graphical Integrity

What's wrong?

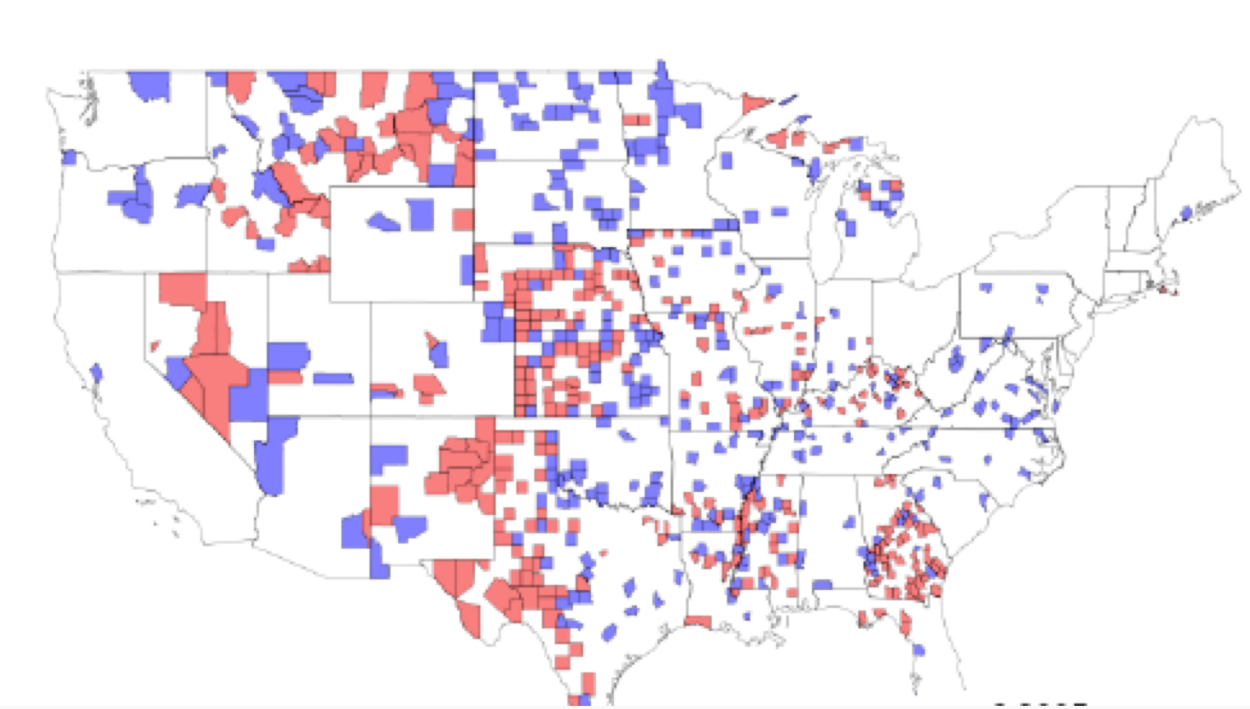


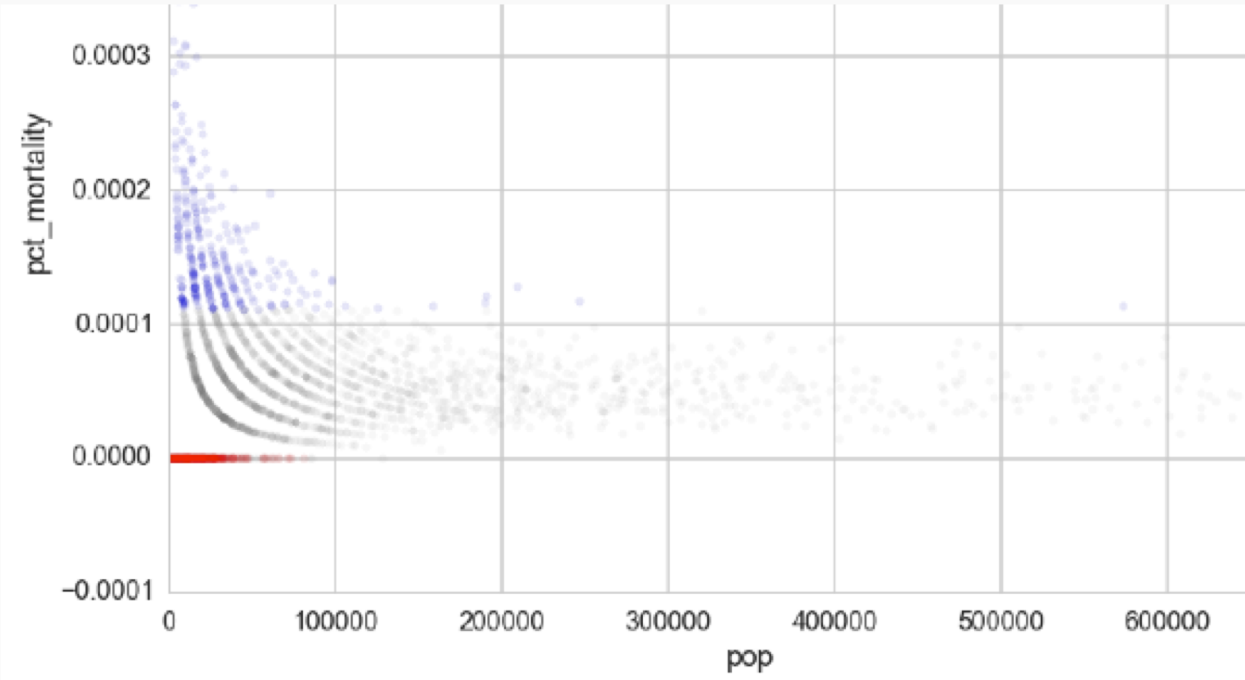
Counties with the LOWEST kidney cancer death rates (1980-1989)



Counties with the HIGHEST kidney cancer death rates (1980-1989)

Graphical Integrity





Graphical Integrity

Lesson: plot all the data

Lecture Outline

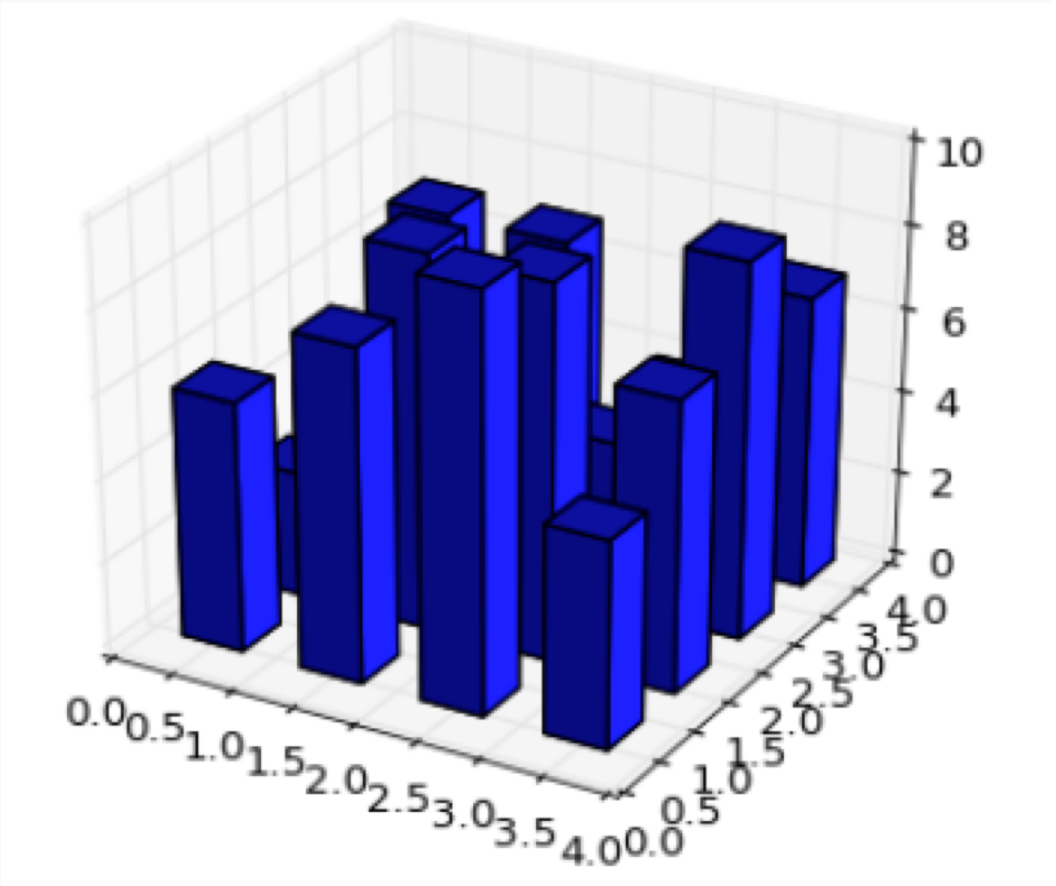
- EDA Refresher
- Effective Visualization
  - Graphical Integrity
  - Scope
  - Displays
  - Sensible Design
- Communication
  - Motivation
  - Key Considerations

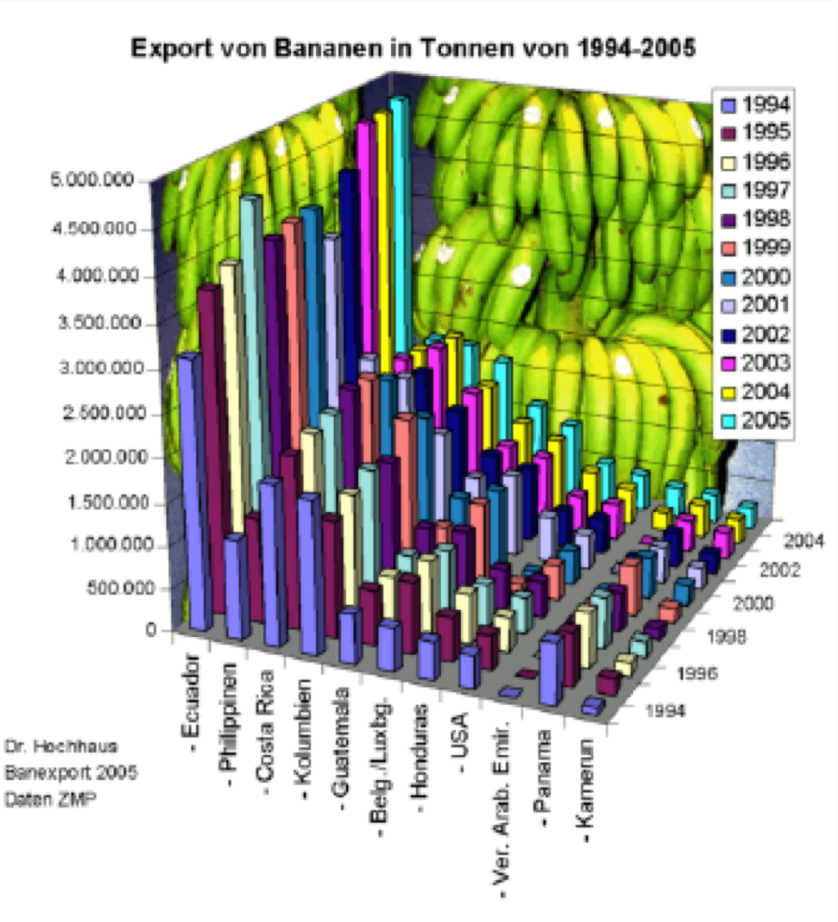
Lecture Outline

- EDA Refresher
- Effective Visualization
  - Graphical Integrity
  - Scope
  - Displays
  - Sensible Design
- Communication
  - Motivation
  - Key Considerations

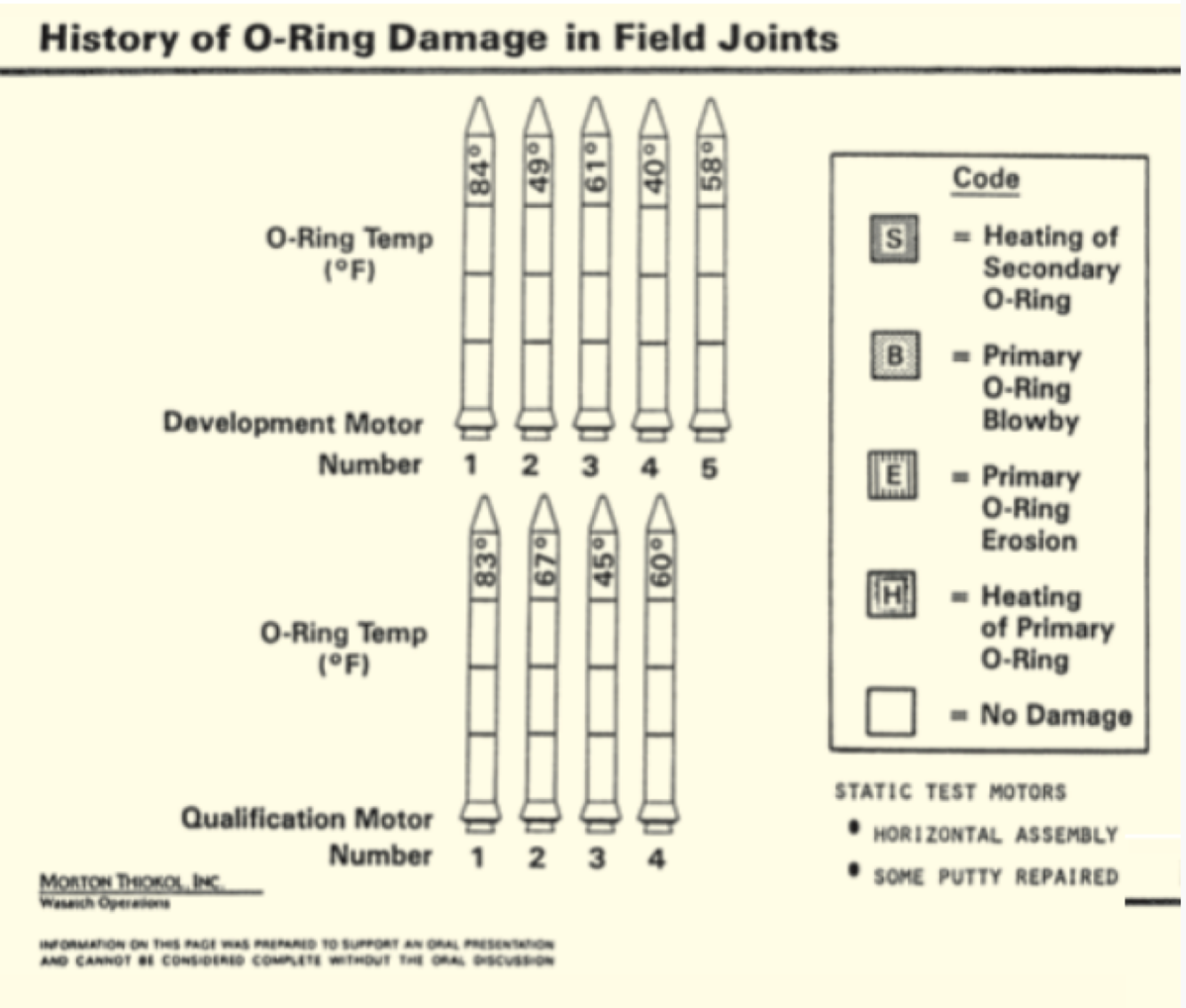
Scope

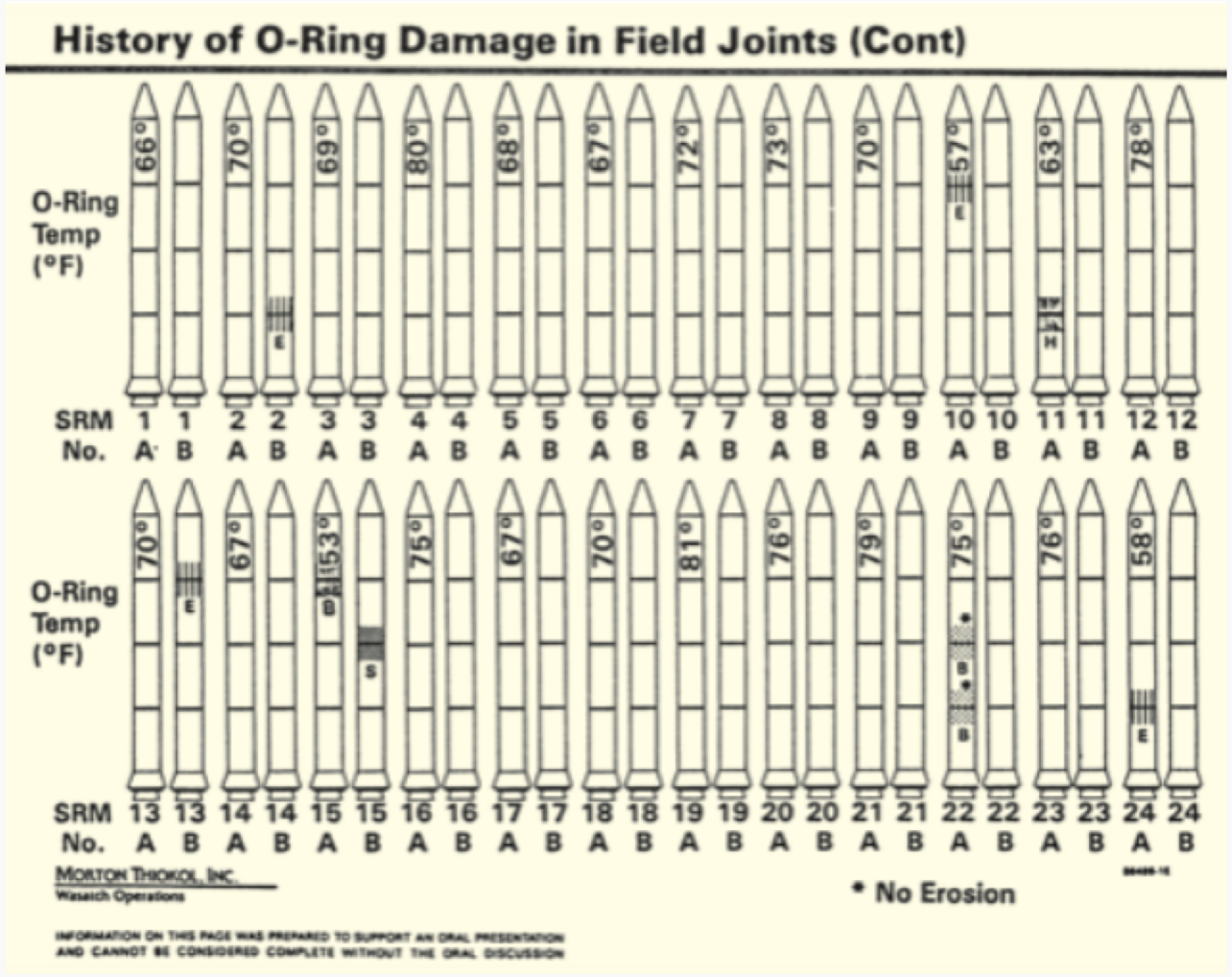
What's wrong?





Scope


History of O-Ring Damage in Field Joints
O-Ring Temp (°F)
84° 49° 61° 40° 58°
Development Motor Number
1 2 3 4 5
O-Ring Temp (°F)
83° 67° 45° 60°
Qualification Motor Number
1 2 3 4
Code
S = Heating of Secondary O-Ring
B = Primary O-Ring Blowby
E = Primary O-Ring Erosion
H = Heating of Primary O-Ring
= No Damage
STATIC TEST MOTORS
• HORIZONTAL ASSEMBLY
• SOME PUTTY REPAIRED
MORTON THIOKOL, INC.
Wasatch Operations
INFORMATION ON THIS PAGE WAS PREPARED TO SUPPORT AN ORAL PRESENTATION AND CANNOT BE CONSIDERED COMPLETE WITHOUT THE ORAL DISCUSSION


History of O-Ring Damage in Field Joints (Cont)
O-Ring Temp (°F)
66° 70° 69° 80° 68° 67° 72° 73° 70° 57° 63° 78°
SRM No.
1 A 1 B 2 A 2 B 3 A 3 B 4 A 4 B 5 A 5 B 6 A 6 B 7 A 7 B 8 A 8 B 9 A 9 B 10 A 10 B 11 A 11 B 12 A 12 B
O-Ring Temp (°F)
70° 67° 53° 75° 67° 70° 81° 76° 79° 75° 76° 58°
SRM No.
13 A 13 B 14 A 14 B 15 A 15 B 16 A 16 B 17 A 17 B 18 A 18 B 19 A 19 B 20 A 20 B 21 A 21 B 22 A 22 B 23 A 23 B 24 A 24 B
MORTON THIOKOL, INC.
Wasatch Operations
• No Erosion
INFORMATION ON THIS PAGE WAS PREPARED TO SUPPORT AN ORAL PRESENTATION AND CANNOT BE CONSIDERED COMPLETE WITHOUT THE ORAL DISCUSSION

Scope

Lesson: keep it simple… enough

Scope

"You should have stayed with the soup question. The object of a question is to obtain information that matters only to us"

-- Sean Connery in Finding Forrester (movie)

Scope

- Making plots is effectively providing an answer to an implicit question
- You get to pick the answer
- Ensure the answer doesn't leave the viewer with uncertainty as to what it's answering or the completeness of the answer
- A good plot should invoke and inspire new questions

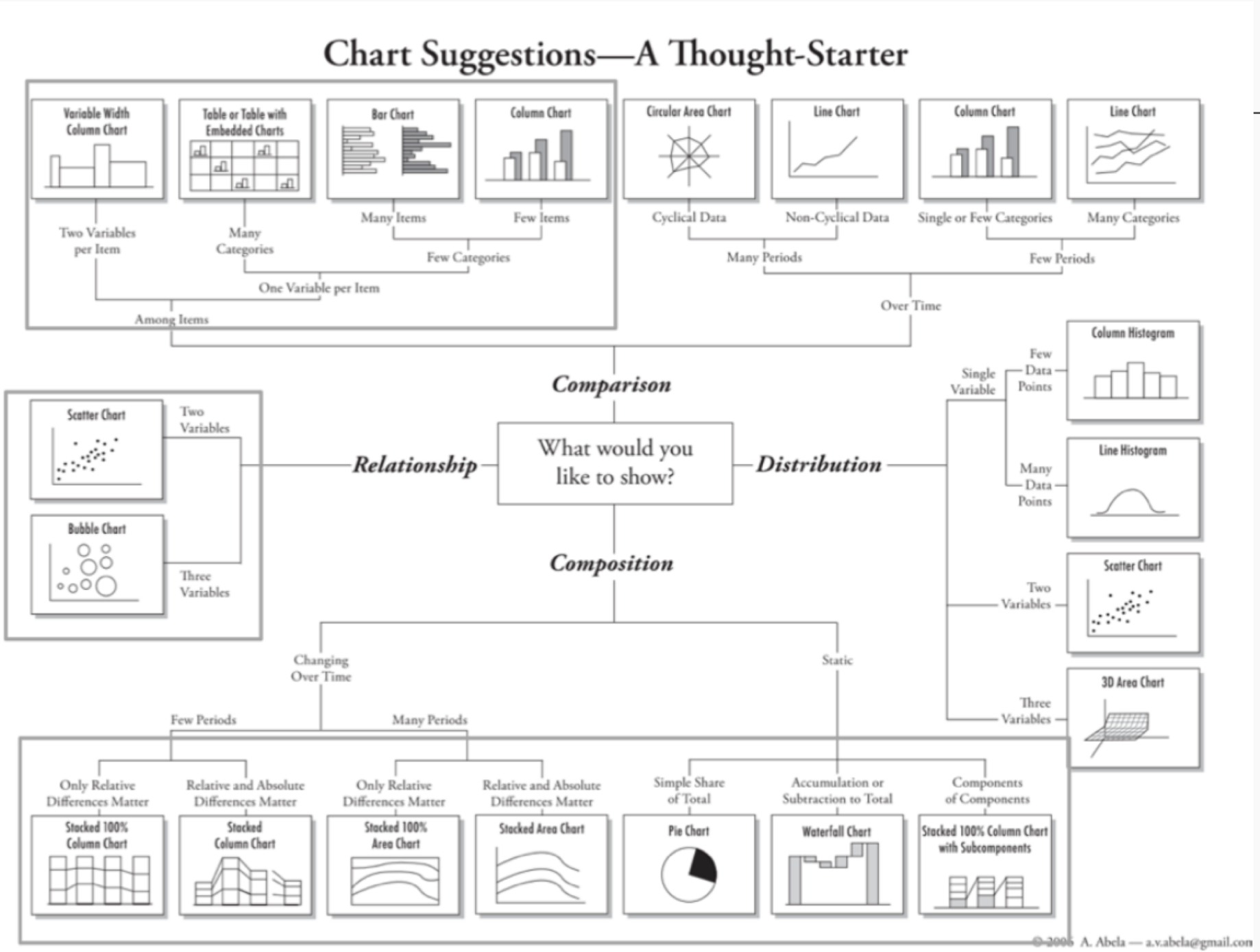
Lecture Outline

- EDA Refresher
- Effective Visualization
  - Graphical Integrity
  - Scope
  - Displays
  - Sensible Design
- Communication
  - Motivation
  - Key Considerations

Lecture Outline

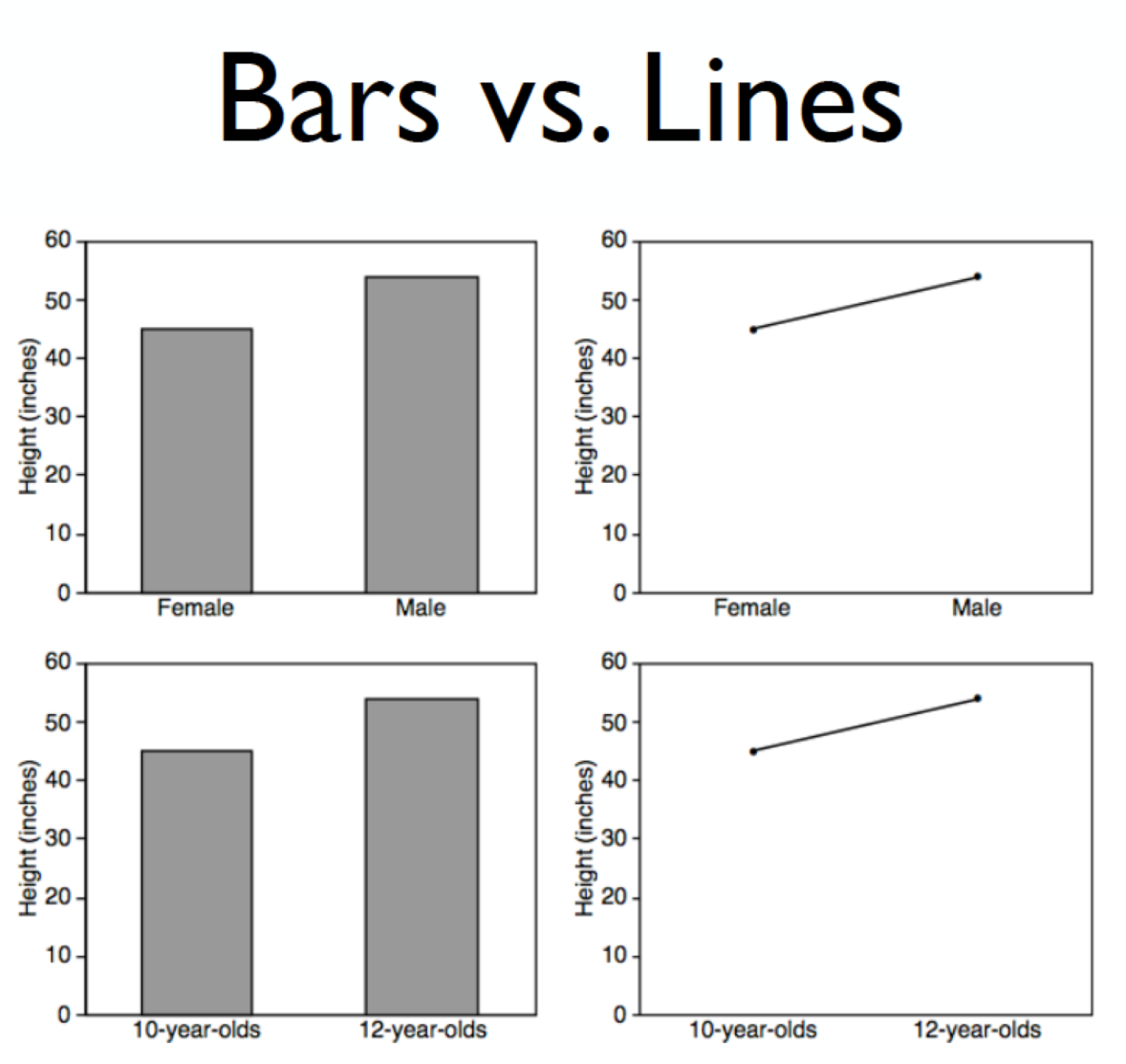
- EDA Refresher
- Effective Visualization
  - Graphical Integrity
  - Scope
  - Displays
  - Sensible Design
- Communication
  - Motivation
  - Key Considerations

Displays



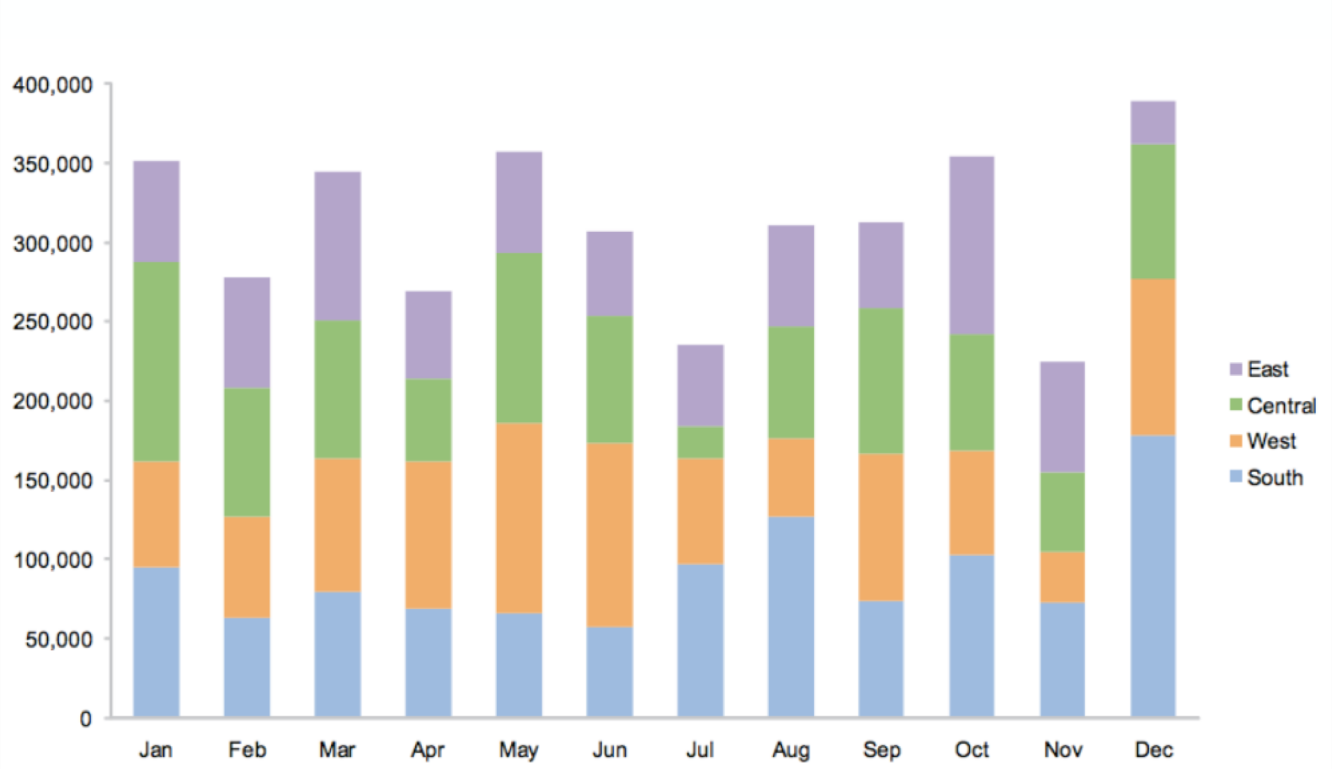
http://extremepresentation.typepad.com/blog/files/choosing_a_good_chart.pdf

Displays: comparisons

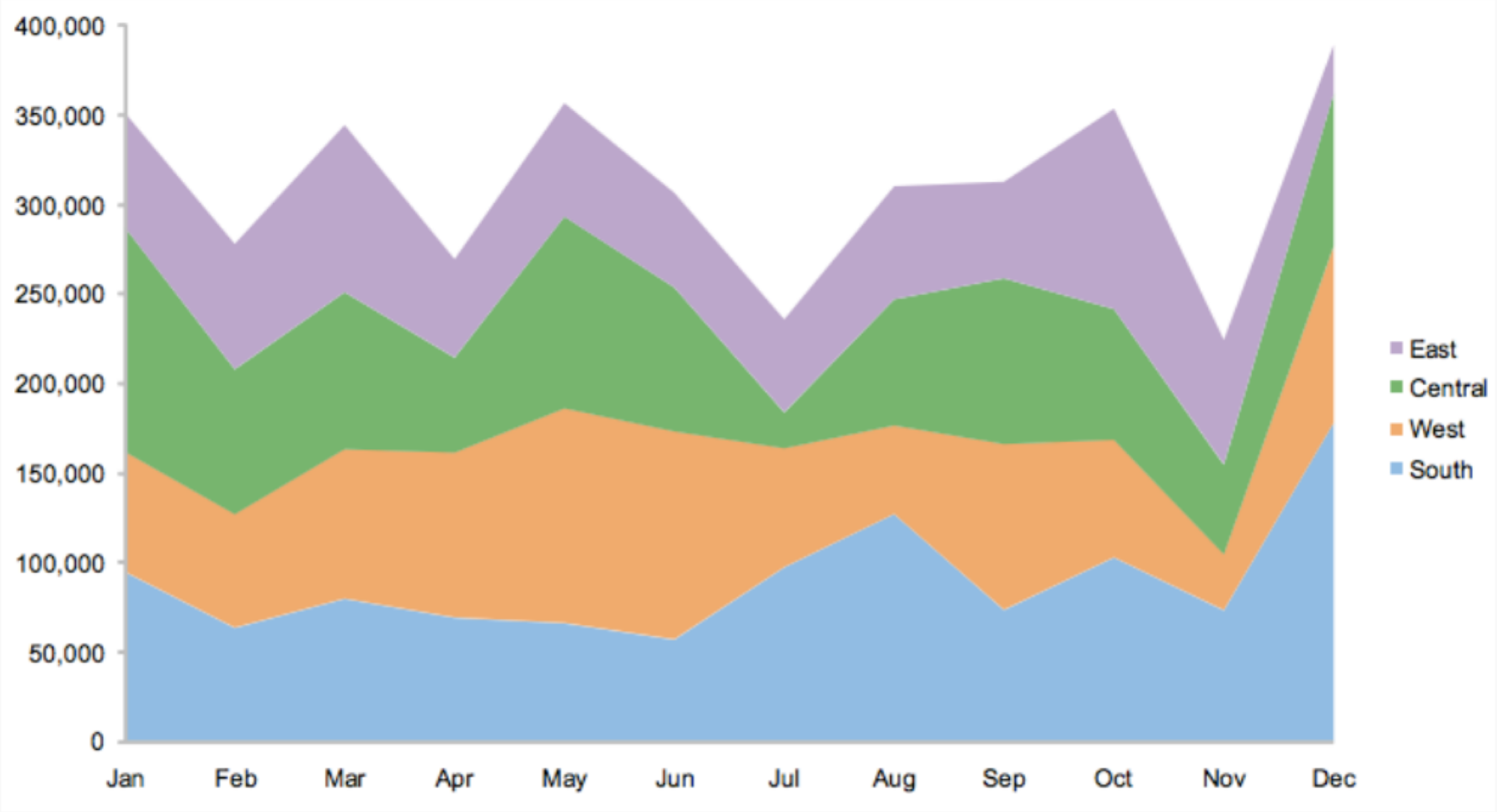


Displays: proportions

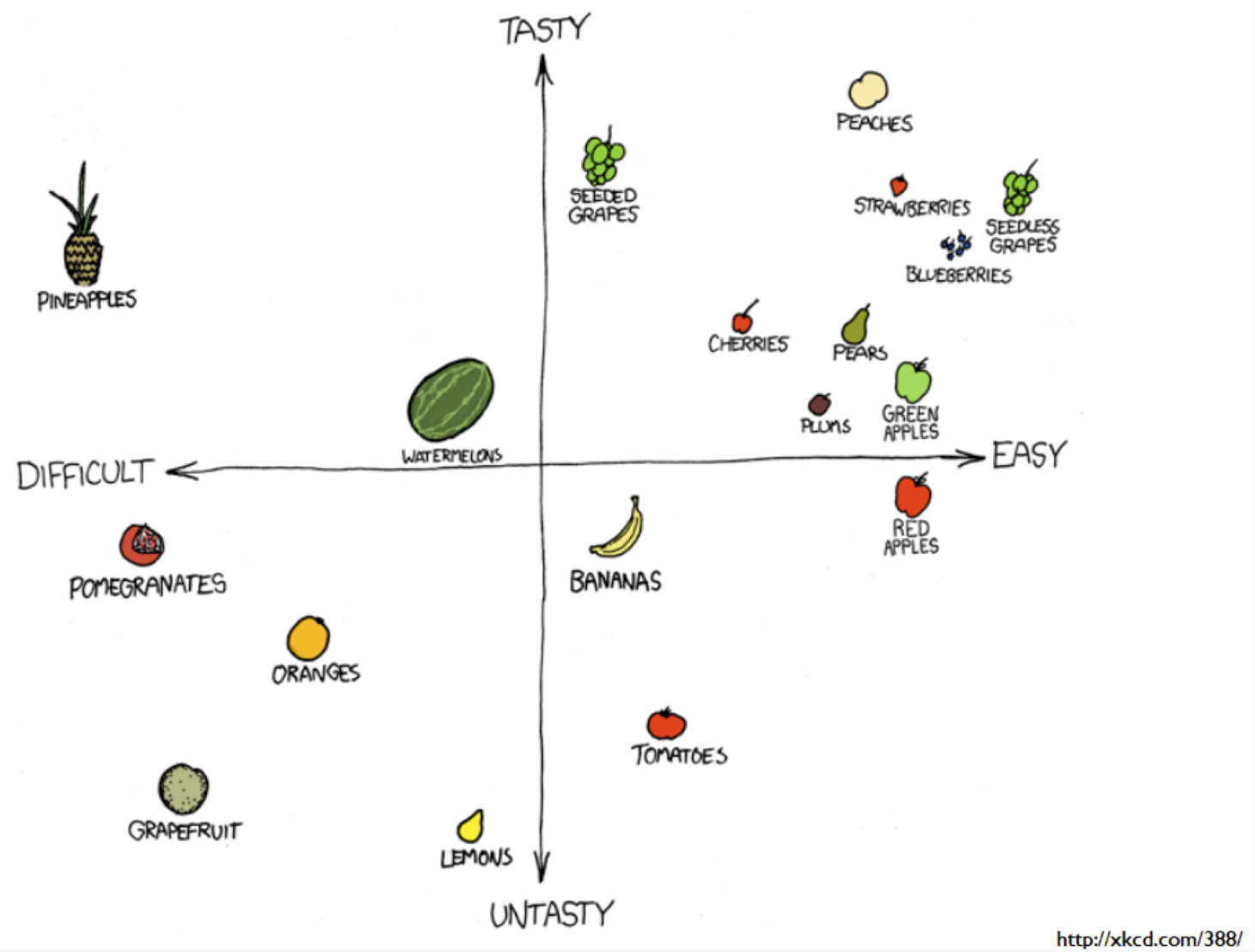
Displays: proportions



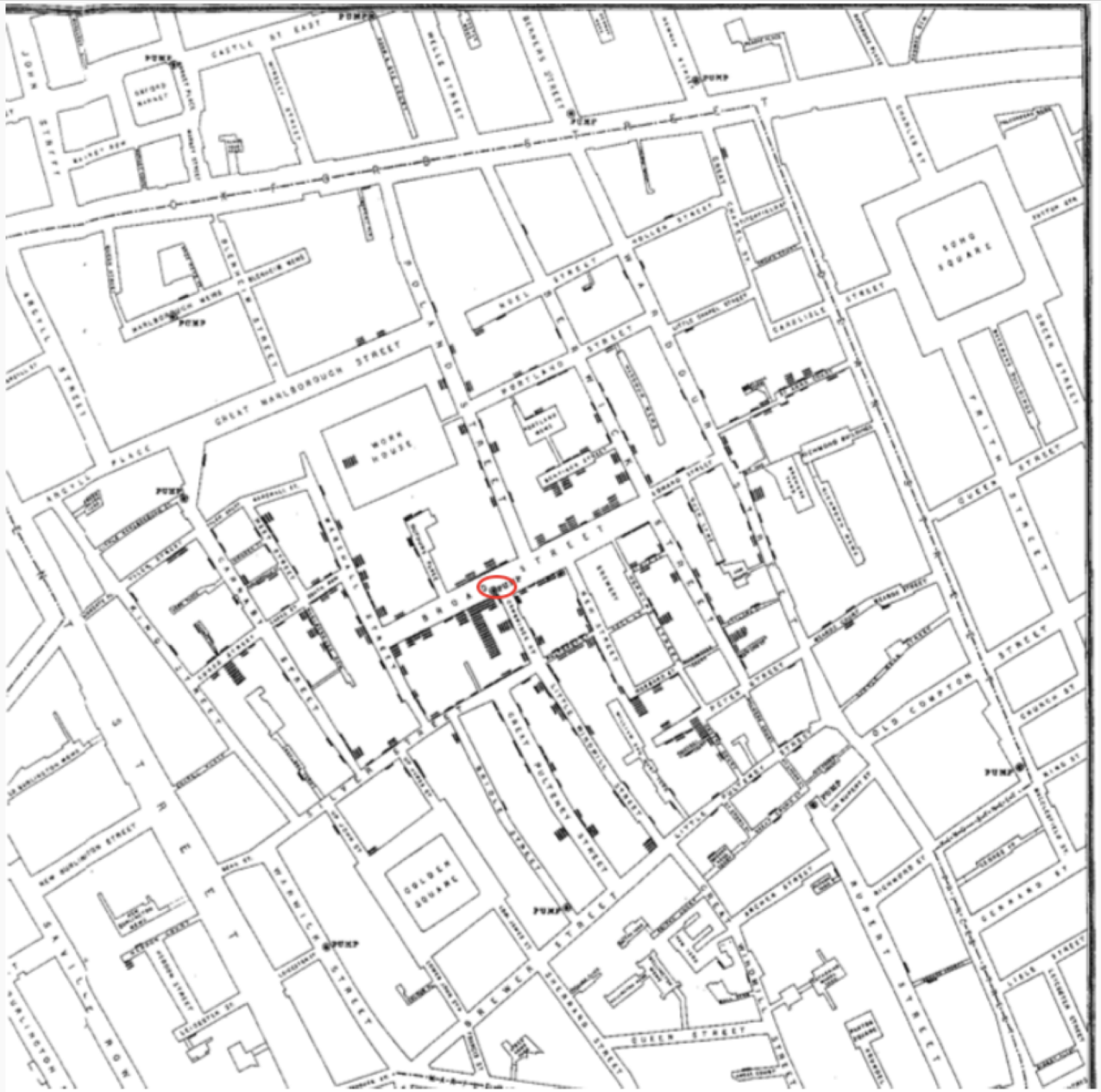
Displays: proportions



Displays: proportions



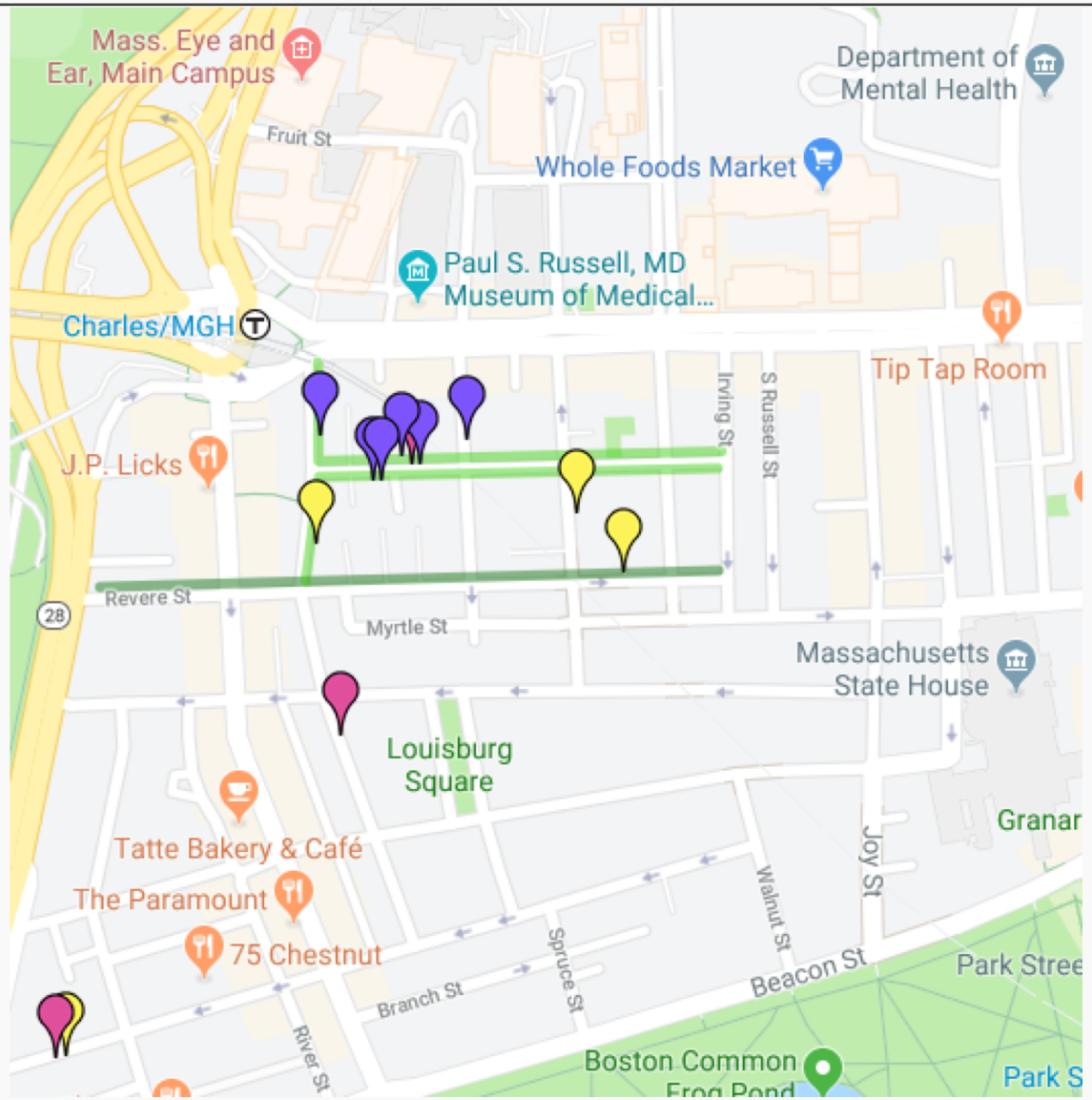
Displays: proportions



London Cholera Epidemic

-- Edward Tufte,
Visual and Statistical
Thinking

Displays: proportions



friday @ 10:30am
sunday @ 5pm
saturday @ 2pm
sunday @ 2pm
friday @ 11:30am
1st and 3rd wed
2nd and 4th wed
2nd and 4th wed
saturday @ 3pm
sunday @ 4:30pm
friday @ 7:30pm
saturday @ 8pm
friday @ 8pm
saturday @ 6pm
saturday @ 7
saturday @ 6