

# **STA2453**

# **Data Visualization**

Prof. Nathan Taback

# Overview

- What Is Data Visualization?
- Visualization Components
- Telling Stories With Data
- Data background
- Who is your audience?
- Data narrative

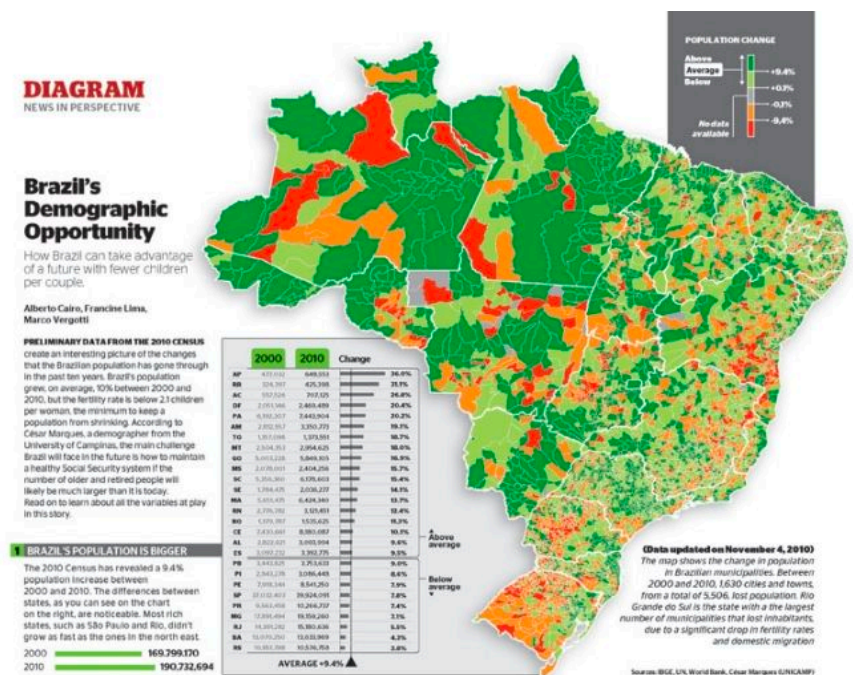
# What is Data Visualization?

Visualization is a way to represent data, an abstraction of the real world, in the same way that the written word can be used to tell different kinds of stories.

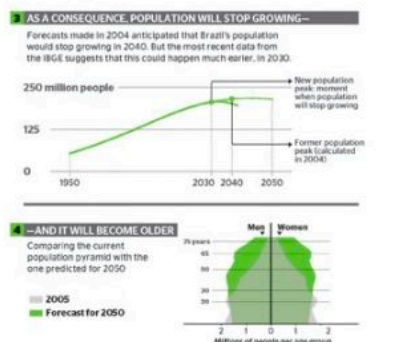
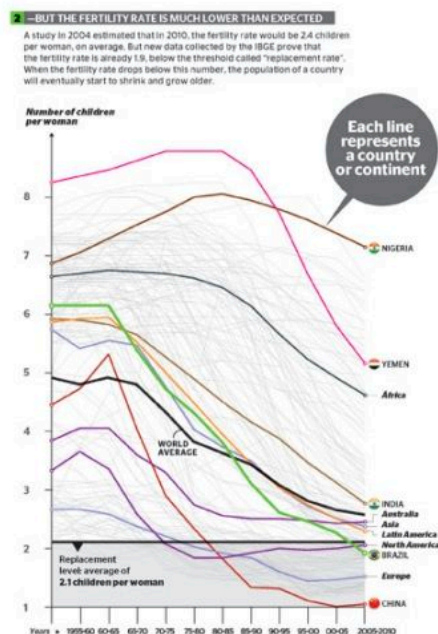
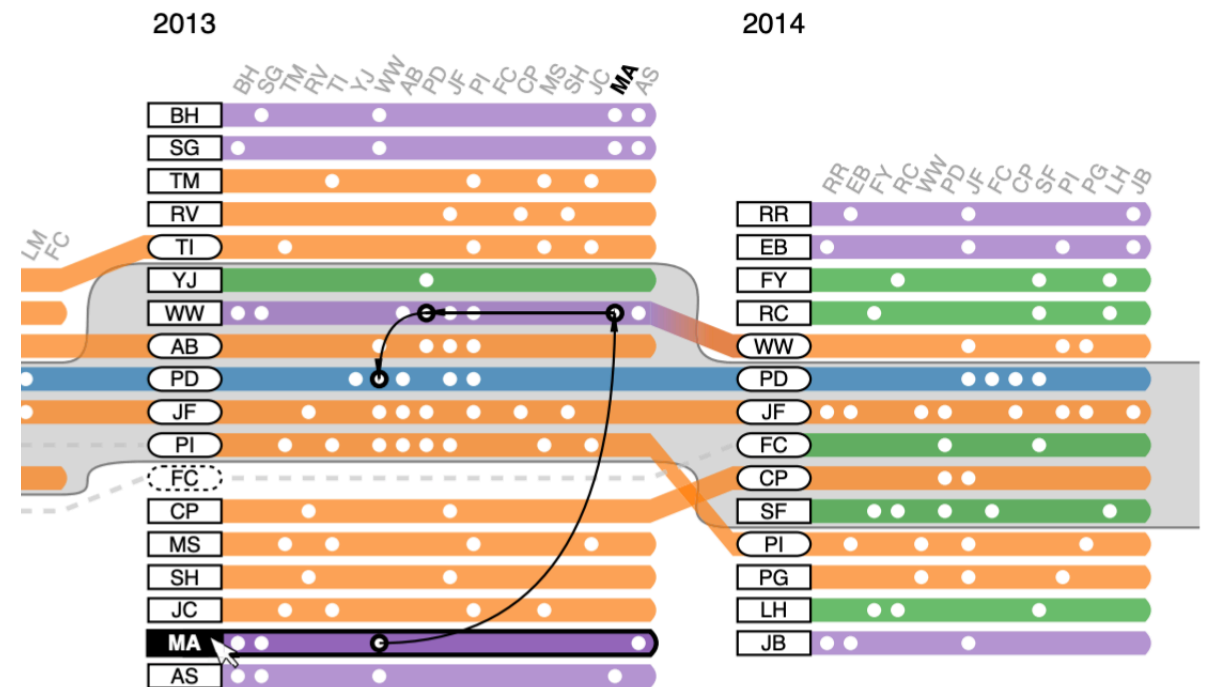
*Yau, N. Data Points*

# MAIN USES OF DATA VISUALIZATION

## EXPLANATORY Storytelling



## EXPLORATORY Acquiring insights



**How Brazil can transform the population challenge into an opportunity**

As the population ages, the proportion of people of working age increases. The country will therefore have more people producing wealth (of the labor market can absorb them) and fewer children to consume investments. It is a window of opportunity, because in some cases the number of people of working age to fall back when older people are leaving the market.

The population under 15 years of age is falling today. A smaller number of student in public schools will facilitate the quality of teaching, if the amount invested in education stays the same.

Educational policy focused on low-income youth favors the formation of more skilled workforce and greater social mobility.

In the future, Brazil will reach the stage of Europe and Japan, which struggle to support their elders. This is why it's so important to prepare a more balanced retirement system, which will include retirement at a later age.



# Visualization Components

- Visualization maps data to geometry and colour.
- It works because your brain is wired to find patterns, and you can switch back and forth between the visual and the numbers it represents.
- **Important:** You must make sure that the essence of the data isn't lost in that back and forth between visual and the value it represents because if you can't map back to the data, the visualization is just a bunch of shapes.

**37 75**

# 37

# 75

1. Grab a set of post-it notes or paper and a pen; gather up in pairs



2. Try to come up with as many possible representations/encodings for the "data" above as you can, in the paper segments.

**Feel free to be creative!**

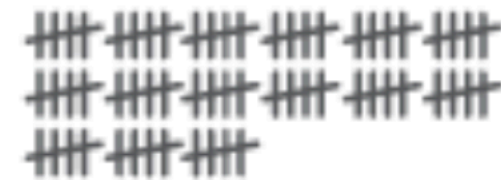
37

75

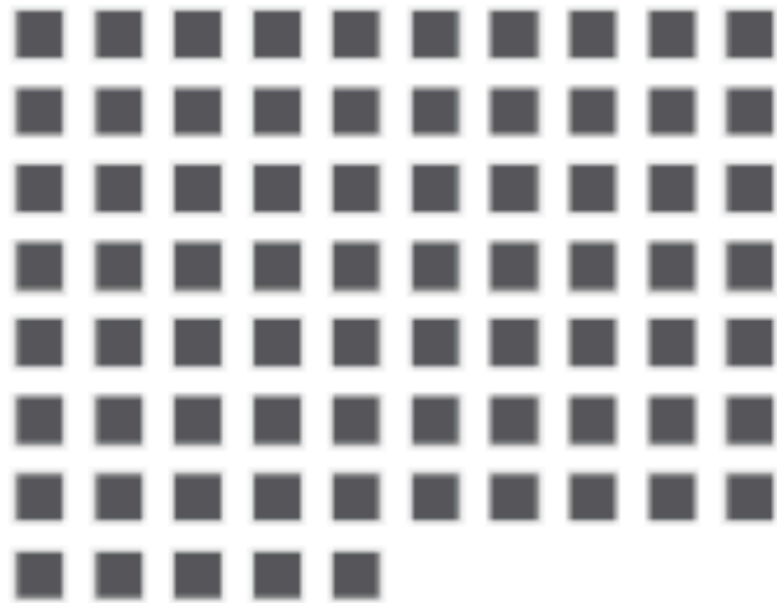
Thirty-seven  
Seventy-five

XXXVII

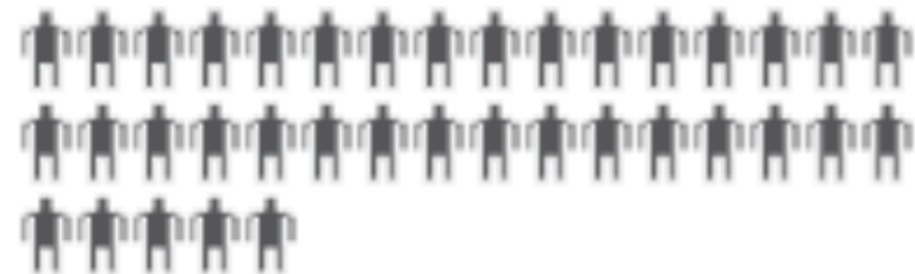
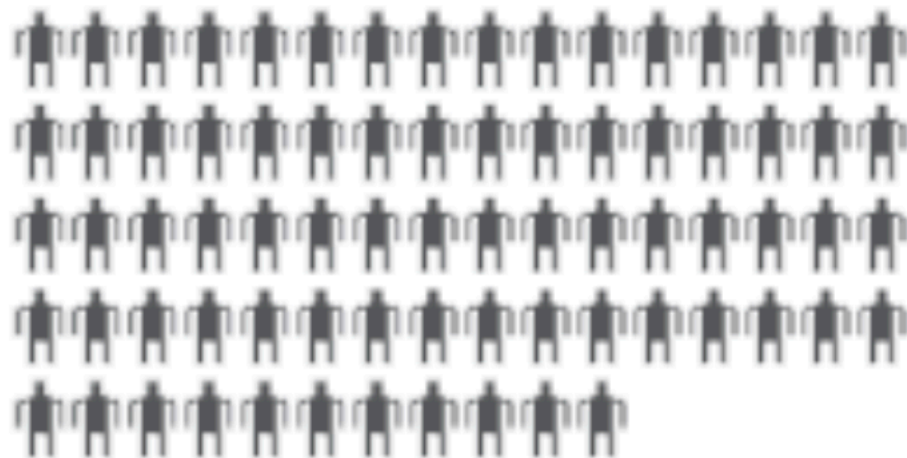
LXXV



# Squares



# Isotypes



75, 37

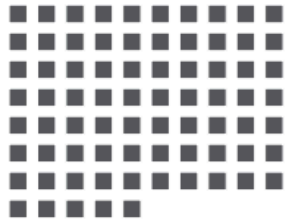
*a*



*b*



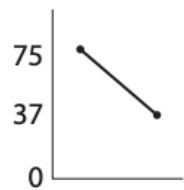
*c*



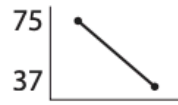
*b*



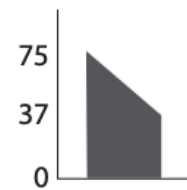
*c*



*a*



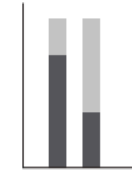
*b*



*c*



*a*



*b*



*a*



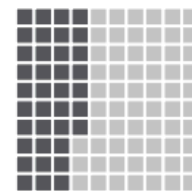
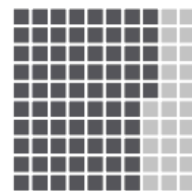
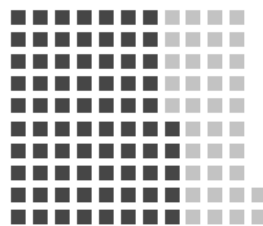
*b*



*a*



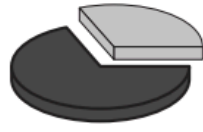
*b*



*a*



*b*



*c*



# Visualization Components

## Visual cues

When you visualize data, you encode values to shapes, sizes, and colors.

### Position

Where in space the data is



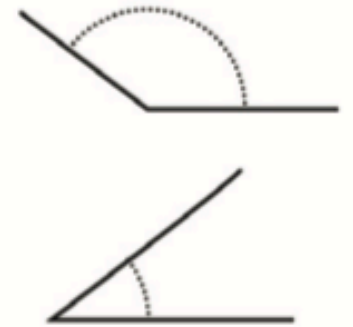
### Length

How long the shapes are



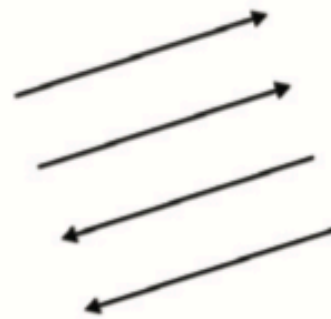
### Angle

Rotation between vectors



### Direction

Slope of a vector in space



### Shapes

Symbols as categories



### Area

How much 2-D space



### Volume

How much 3-D space



### Color saturation

Intensity of a color hue



### Color hue

Usually referred to as color

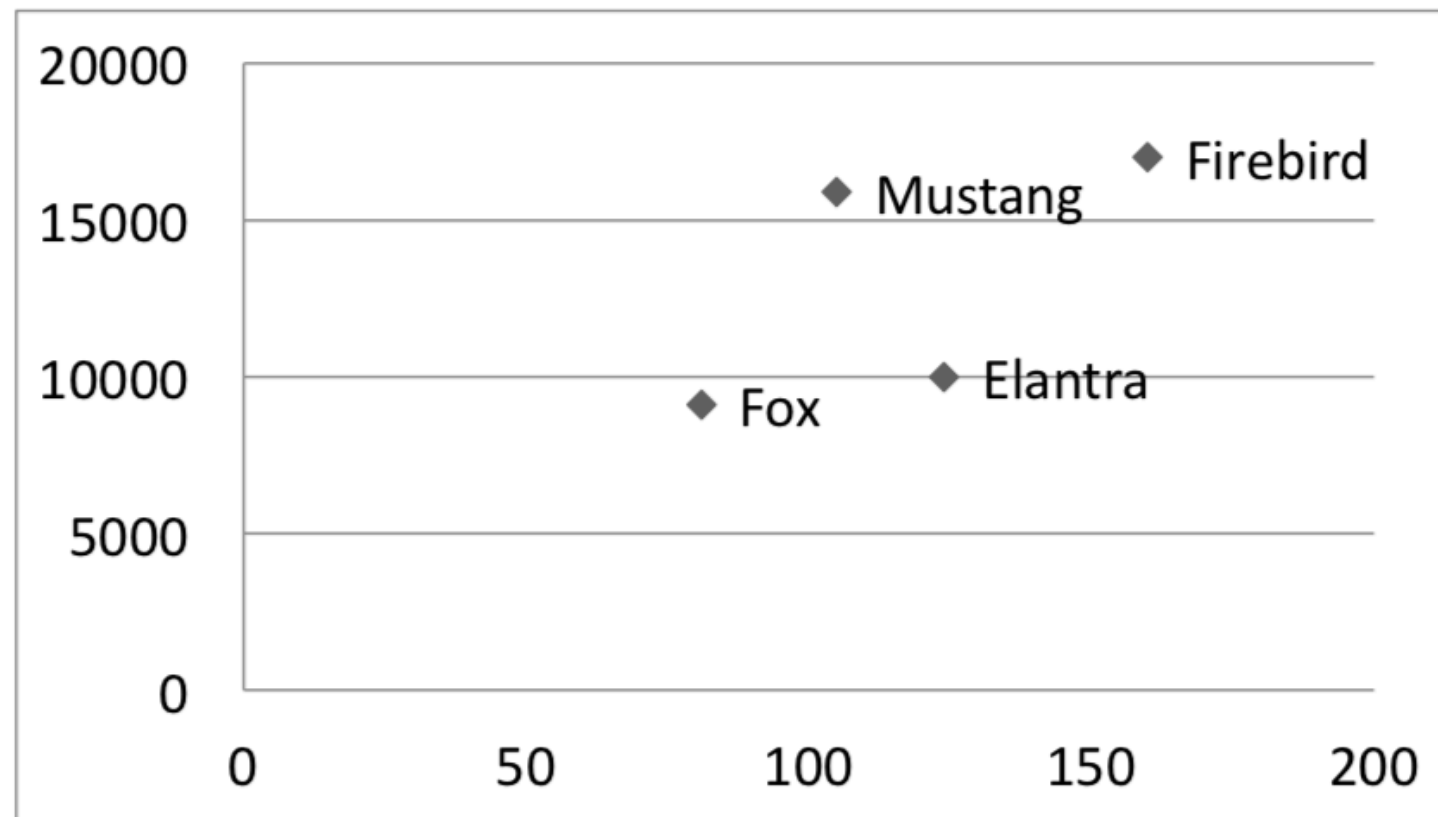


Yau, N. Data Points

FIGURE 3-3 Visual cues

# MAPPING DATA TO VISUAL VARIABLES

Model	Type	HP	Price	Number Airbags
Volkswagen Fox	Compact	81	9100	0 (no airbag)
Hyundai Elantra	Compact	124	10000	0 (no airbag)
Pontiac Firebird	Sport	160	17000	2 (driv. & pass.)
Ford Mustang	Sport	105	15900	1 (driver)



Position on X-axis

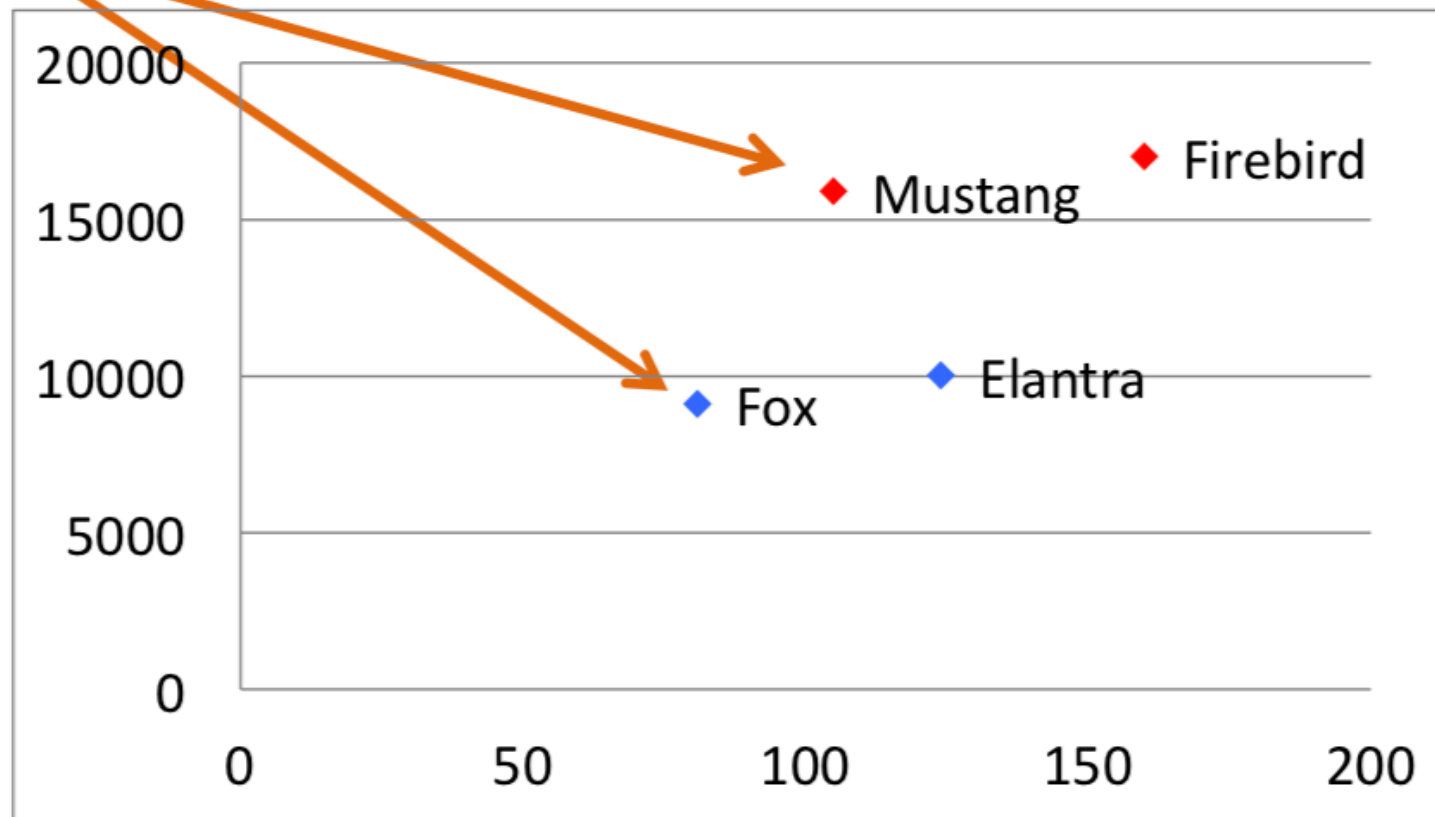
Position on Y-axis



# MAPPING DATA TO VISUAL VARIABLES

Model	Type	HP	Price	Number Airbags
Volkswagen Fox	Compact	81	9100	0 (no airbag)
Hyundai Elantra	Compact	124	10000	0 (no airbag)
Pontiac Firebird	Sport	160	17000	2 (driv. & pass.)
Ford Mustang	Sport	105	15900	1 (driver)

Color  
(Hue)



Position on Y-axis

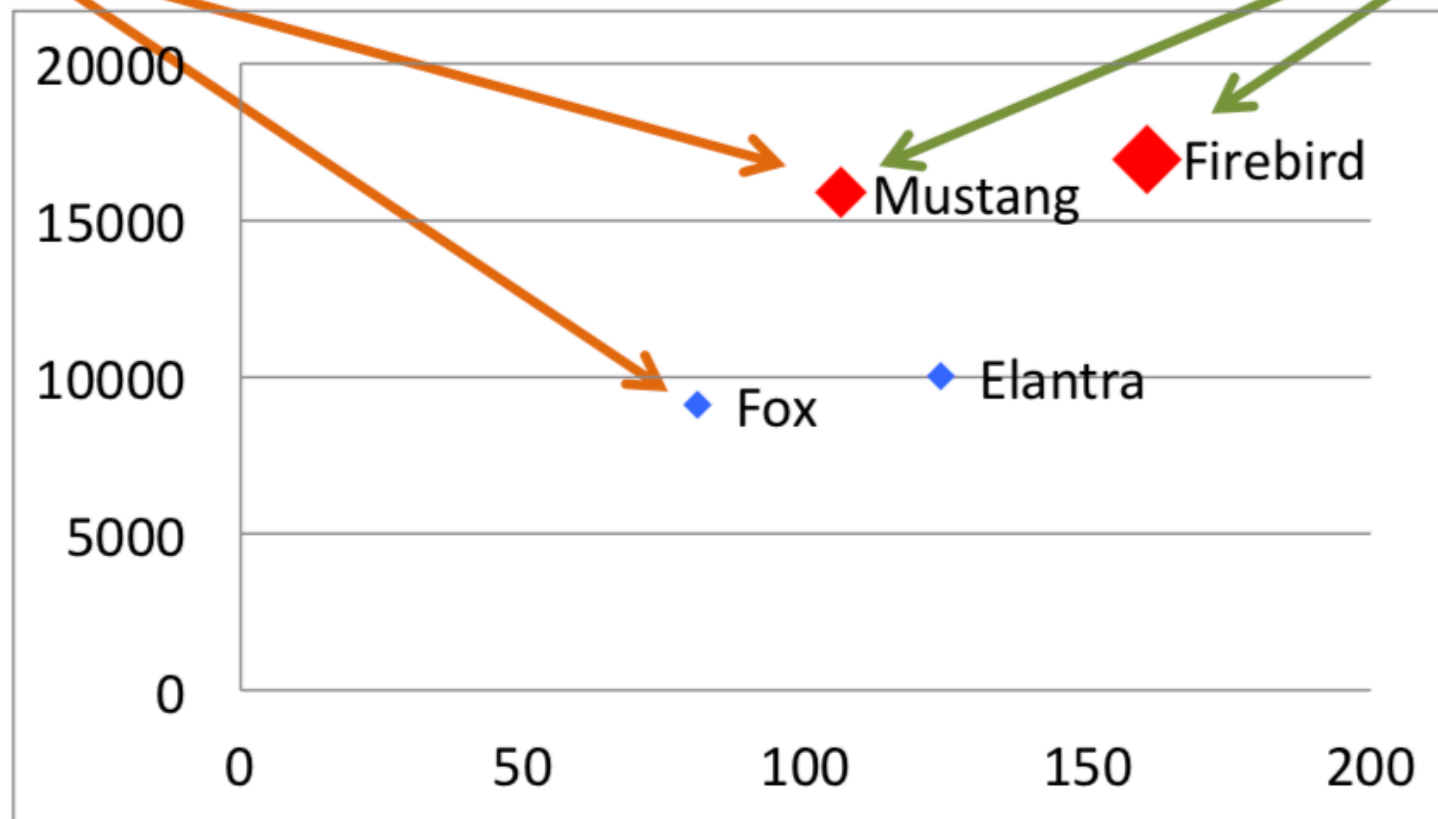
Position on X-axis

# MAPPING DATA TO VISUAL VARIABLES

Model	Type	HP	Price	Number Airbags
Volkswagen Fox	Compact	81	9100	0 (no airbag)
Hyundai Elantra	Compact	124	10000	0 (no airbag)
Pontiac Firebird	Sport	160	17000	2 (driv. & pass.)
Ford Mustang	Sport	105	15900	1 (driver)

Color  
(Hue)

Size



Position on Y-axis

Position on X-axis

**Which representations can I use to encode data?**

**Expressiveness principle:  
use adequate/suitable data representations**

Encodings should convey all, and only, the information of associated attributes.

e.g. Ordinal data representation should convey “order”; similarly, “categorical data” should not be shown in a way that implies order.

Which representations are more suitable to ensure I'm conveying the right message?

## **Effectiveness principle: choosing the best representation to your data**

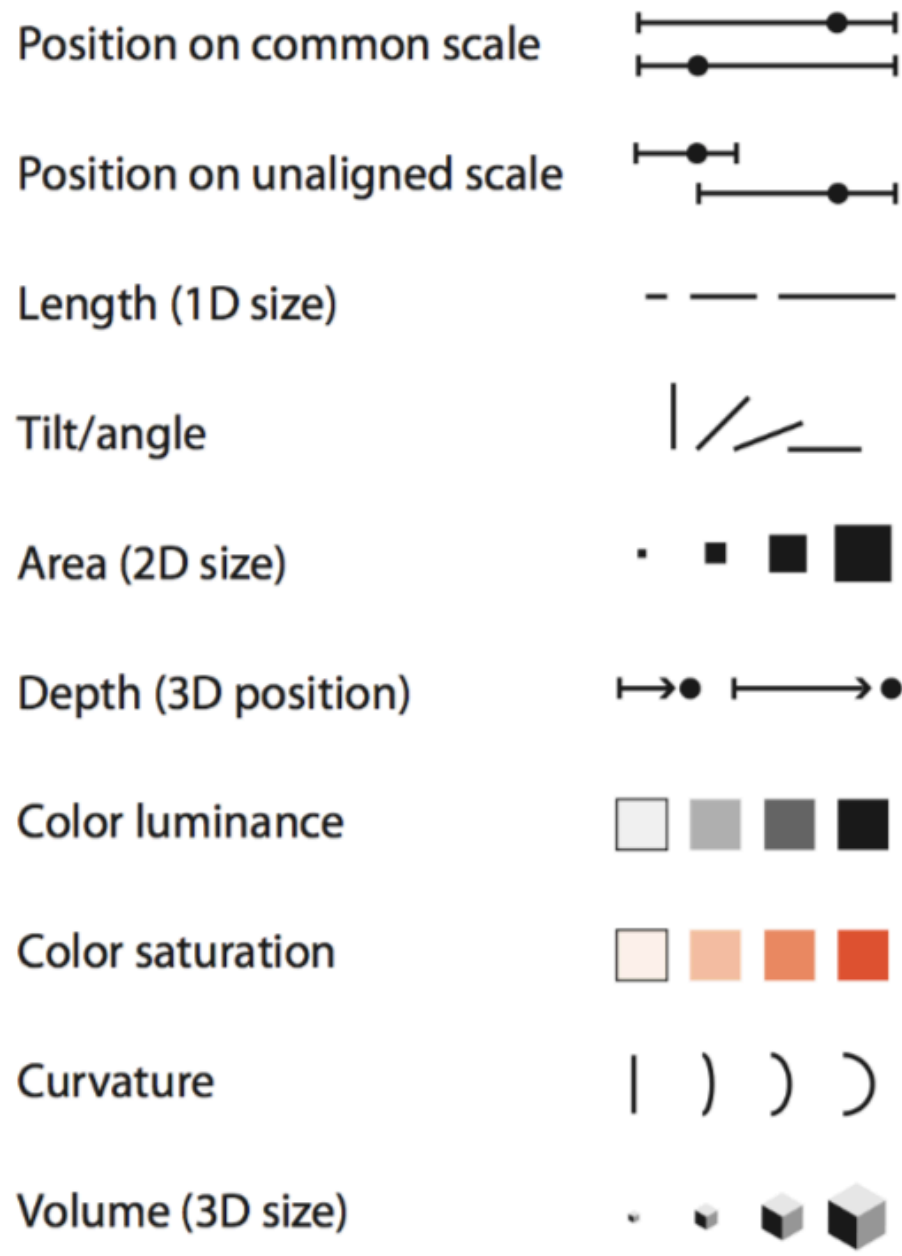
Importance of attributes should match the "saliency" of the channel;

Most important attributes should be encoded using the most effective and noticeable channels.

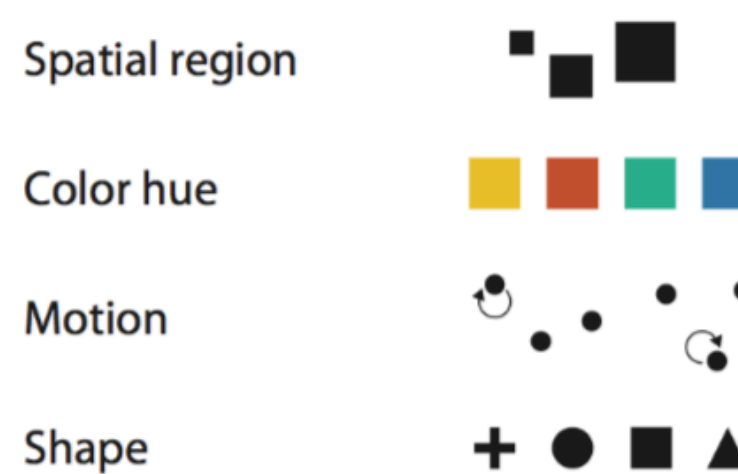
# EFFECTIVENESS PRINCIPLE

Some variables are perceptually better than others

## ➔ Magnitude Channels: Ordered Attributes

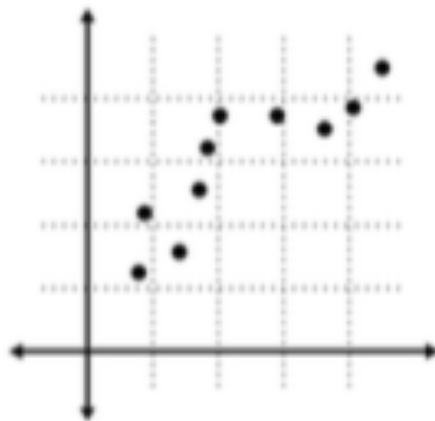


## ➔ Identity Channels: Categorical Attributes

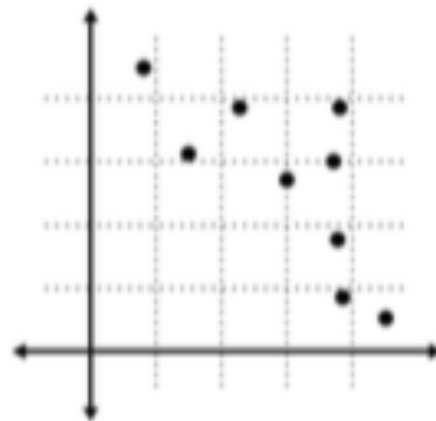


# Position

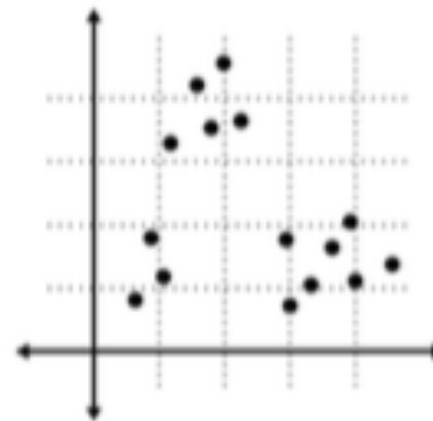
Upward trend



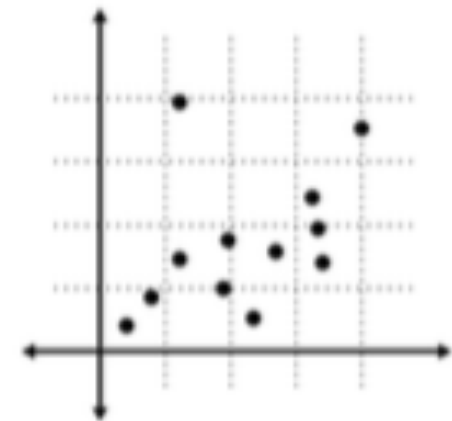
Downward trend



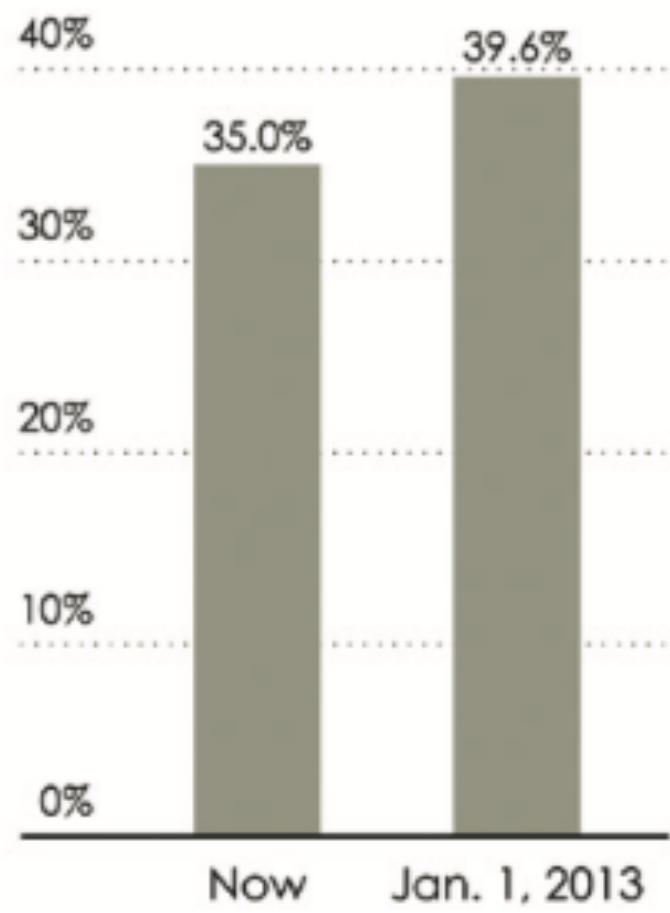
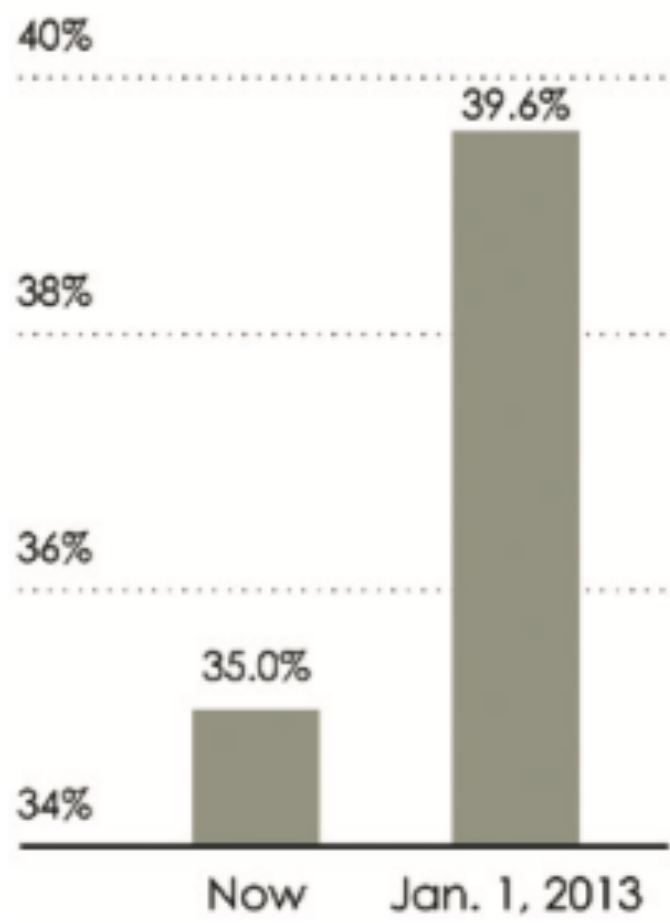
Clustering



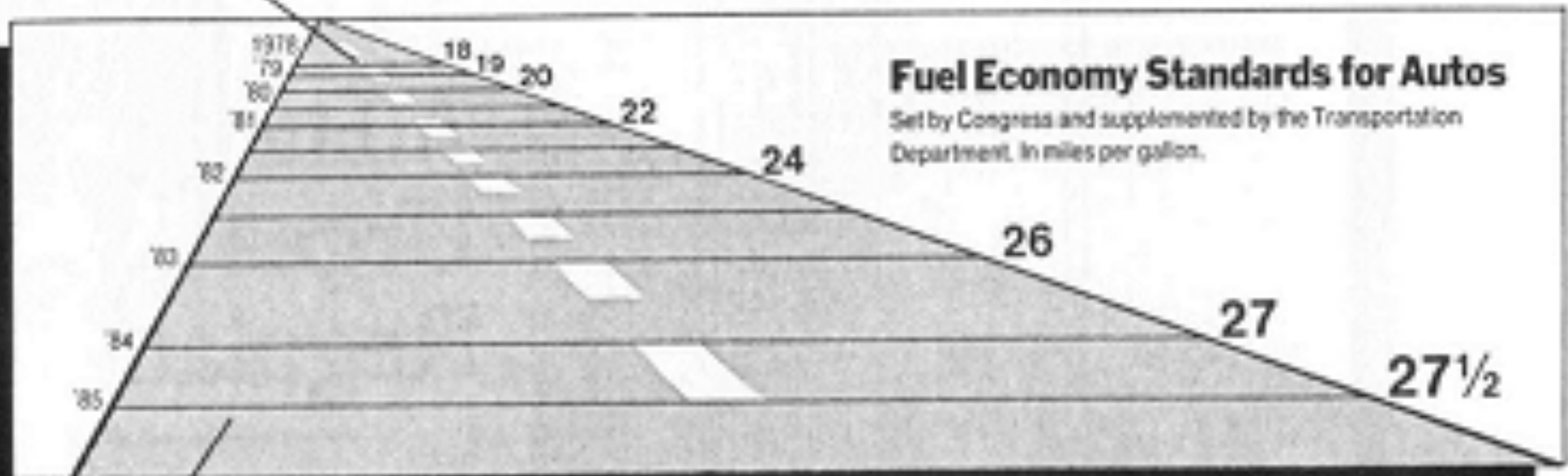
Outlier



# Length



This line, representing 18 miles per gallon in 1978, is 0.6 inches long.



This line, representing 27.5 miles per gallon in 1985, is 5.3 inches long.

New York Times, August 9, 1978, p. D-2.

**Tufte, E. Visual Display of Quantitative Information.**

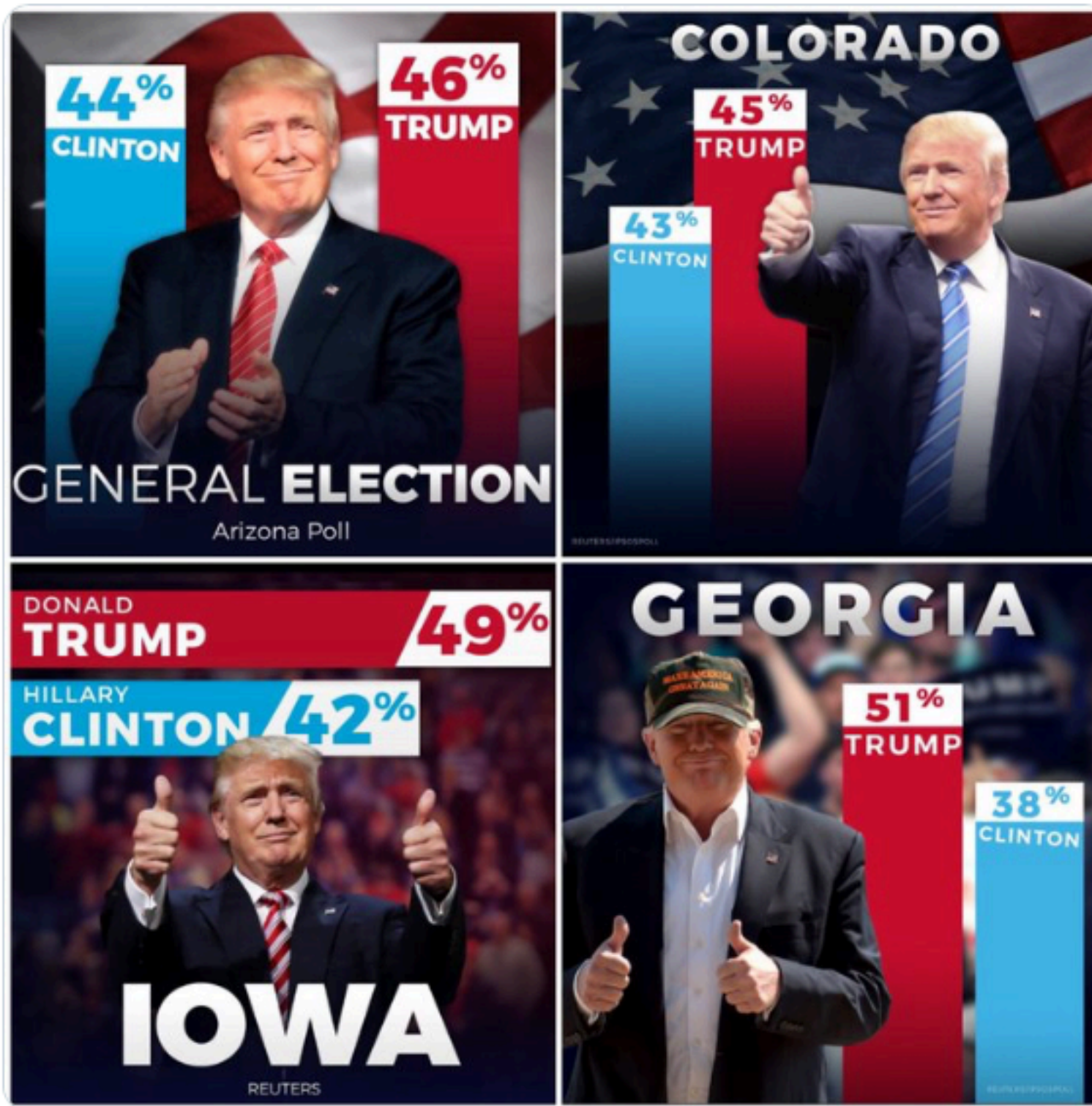


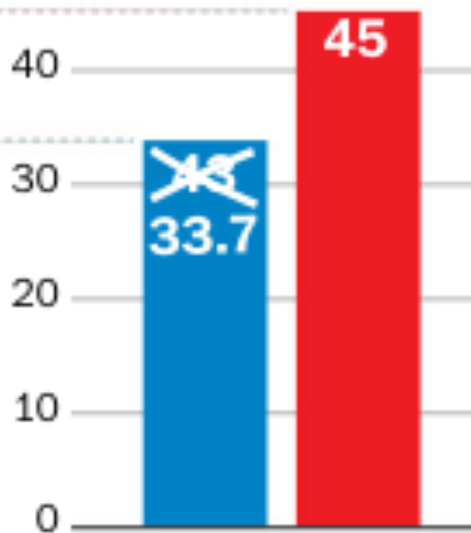


Donald J. Trump   
@realDonaldTrump

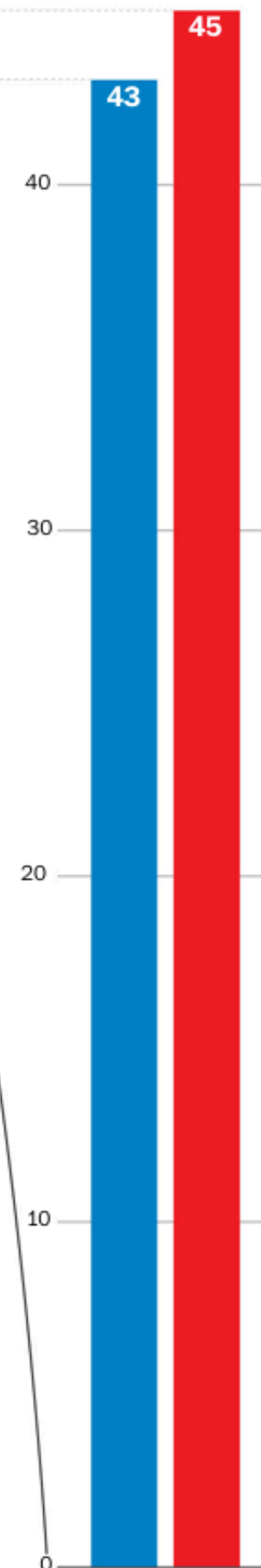


Reuters polling just out- thank you!  
[#MakeAmericaGreatAgain](#)

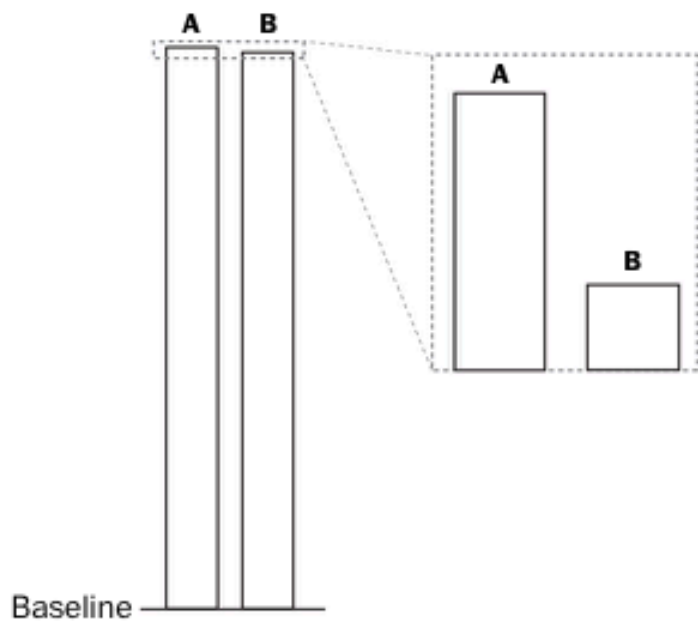




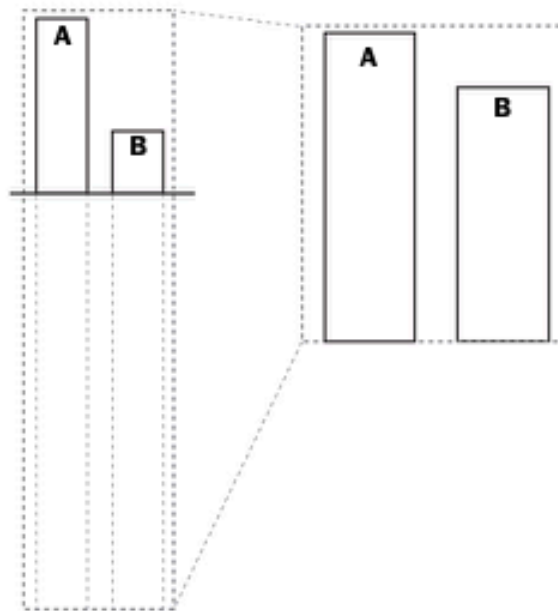
The real baseline is way down there



**EXAGGERATING A LEAD**



**DIMINISHING A LEAD**







**Donald J. Trump** ✓

@realDonaldTrump



♥ 214K 6:05 AM - Oct 1, 2019



💬 110K people are talking about this



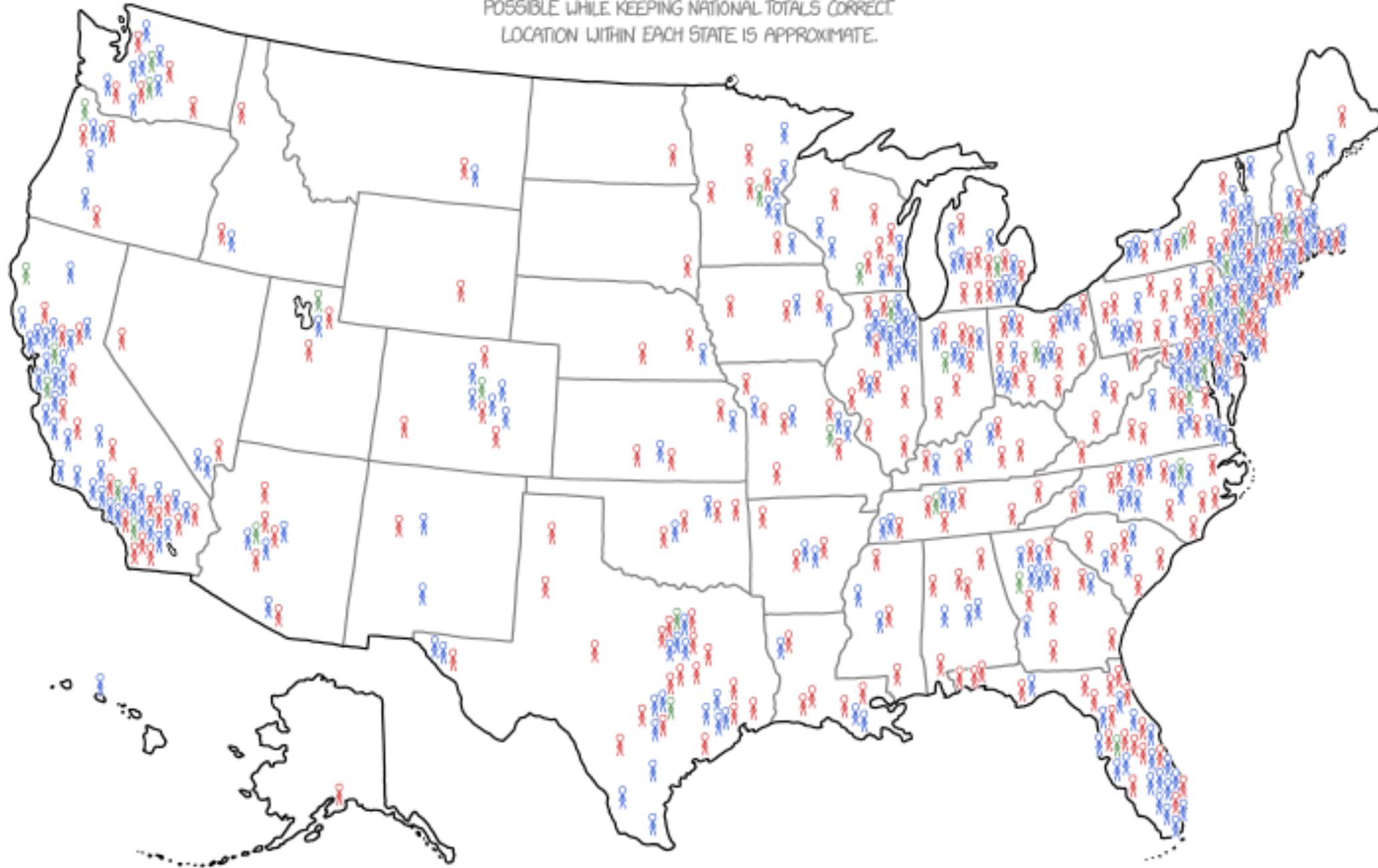
# 2016 ELECTION MAP

EACH FIGURE REPRESENTS 250,000 VOTES

TRUMP CLINTON OTHER

VOTES ARE DISTRIBUTED BY STATE AS ACCURATELY AS POSSIBLE WHILE KEEPING NATIONAL TOTALS CORRECT

LOCATION WITHIN EACH STATE IS APPROXIMATE.





### United States presidential election, 2008



2004 ← **November 4, 2008** → 2012

All **538 electoral votes** of the **Electoral College**  
270 electoral votes needed to win

Turnout 58.2%<sup>[1]</sup> ▲ 1.5%




Nominee	<b>Barack Obama</b>	John McCain
Party	Democratic	Republican
Home state	Illinois	Arizona
Running mate	<b>Joe Biden</b>	Sarah Palin
Electoral vote	365	173
States carried	<b>28 + DC + NE-02</b>	22
Popular vote	<b>69,498,516</b>	59,948,323
Percentage	<b>52.9%</b>	45.7%

### United States presidential election, 2012



2008 ← **November 6, 2012** → 2016


All **538 electoral votes** of the **Electoral College**  
270 electoral votes needed to win

Turnout 54.9%<sup>[1]</sup> ▼ 3.3%






Nominee	<b>Barack Obama</b>	Mitt Romney
Party	Democratic	Republican
Home state	Illinois	Massachusetts
Running mate	<b>Joe Biden</b>	Paul Ryan
Electoral vote	<b>332</b>	206
States carried	<b>26 + DC</b>	24
Popular vote	<b>65,915,795</b>	60,933,504
Percentage	<b>51.1%</b>	47.2%

### United States presidential election, 2016

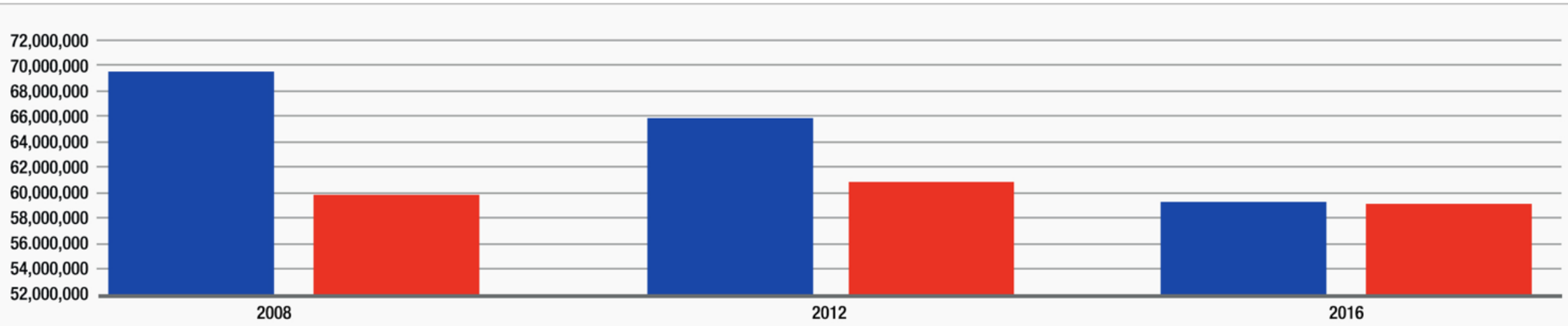


2012 ← **November 8, 2016** → 2020

538 members of the **Electoral College**  
270 electoral votes needed to win

Nominee	<b>Hillary Clinton</b>	Donald Trump
Party	Democratic	Republican
Home state	New York	New York
Running mate	Tim Kaine	<b>Mike Pence</b>
Projected electoral vote	232 <sup>[1][2][3]</sup>	<b>306<sup>[1][2][3]</sup></b>
States carried	20 + DC	<b>30 + ME-02</b>
Popular vote	<b>59,861,516<sup>[4]</sup></b>	59,639,462 <sup>[4]</sup>
Percentage	<b>47.7%</b>	47.5%



source: [https://en.wikipedia.org/wiki/United\\_States\\_presidential\\_election,\\_2008](https://en.wikipedia.org/wiki/United_States_presidential_election,_2008)  
[https://en.wikipedia.org/wiki/United\\_States\\_presidential\\_election,\\_2012](https://en.wikipedia.org/wiki/United_States_presidential_election,_2012)  
[https://en.wikipedia.org/wiki/United\\_States\\_presidential\\_election,\\_2016](https://en.wikipedia.org/wiki/United_States_presidential_election,_2016)

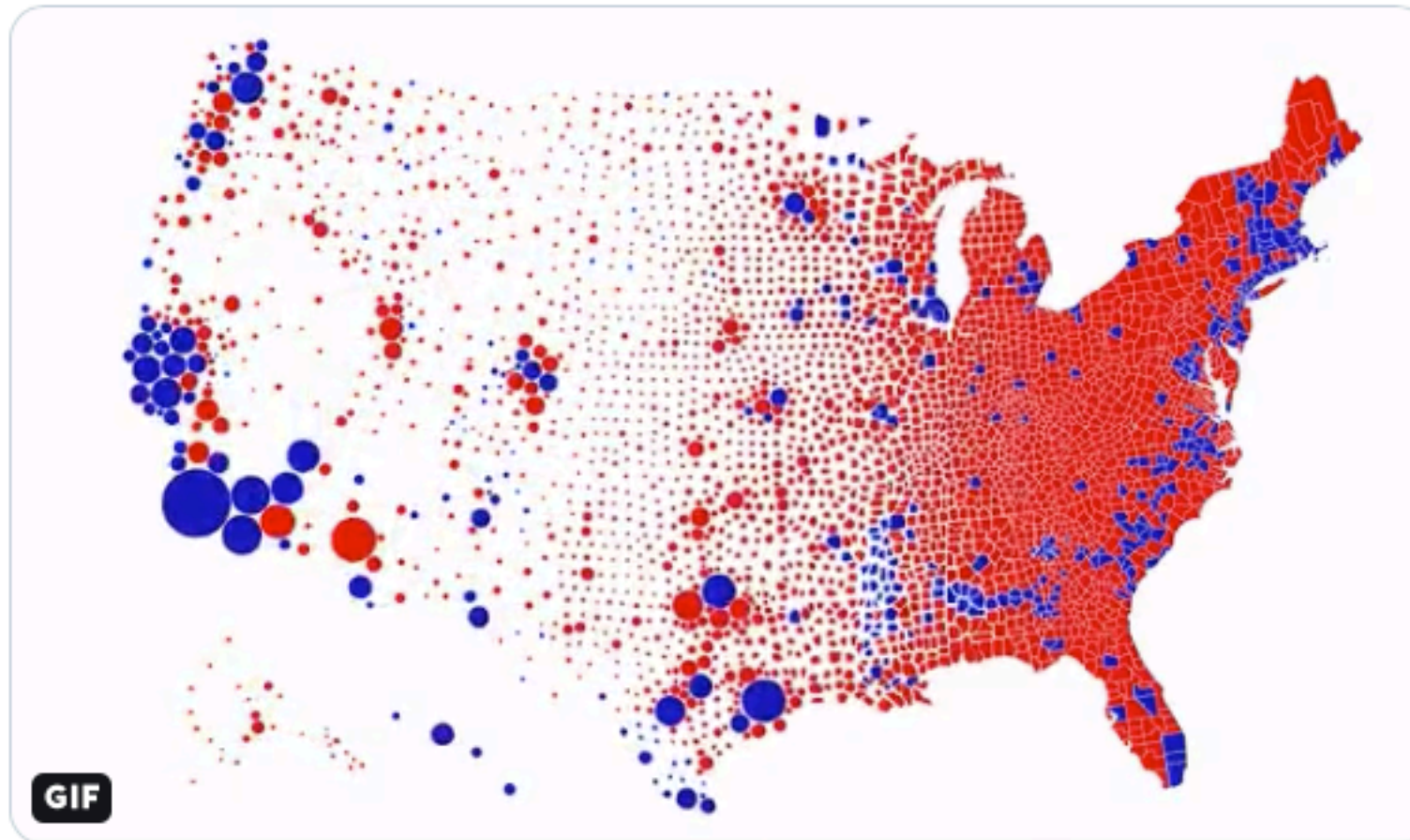


**Karim Douieb**  
@karim\_douieb



Challenge accepted! Here is a transition between surface area of US counties and their associated population. This arguably provides a much more accurate reading of the situation. [@observablehq](#) notebook:

[observablehq.com/@karimdouieb/t...](#) #HowChartsLie  
#DataViz #d3js



[https://twitter.com/karim\\_douieb/status/1181695687005745153](https://twitter.com/karim_douieb/status/1181695687005745153)



# Color

Colour can be used to encode data.

The screenshot displays the ColorBrewer 2.0 web application interface. The main map shows the United States with a 3-class sequential color scheme (BuGn) applied to county-level data. The interface includes several control panels:

- Number of data classes:** Set to 3.
- Nature of your data:** Sequential (selected), Diverging, Qualitative.
- Pick a color scheme:** Multi-hue and Single hue options with various color palette thumbnails.
- Only show:** Colorblind safe, print friendly, photocopy safe (all unchecked).
- Context:** Roads, cities, borders (checked).
- Background:** Solid color (selected), terrain.
- 3-class BuGn:** Color palette with three classes: #e5f5f9, #99d8c9, and #2ca25f.
- EXPORT:** HEX color format selected.

© Cynthia Brewer, Mark Harrower and The Pennsylvania State University  
[Source code and feedback](#)  
[Back to Flash version](#)  
[Back to ColorBrewer 1.0](#)

axismaps

<http://colorbrewer2.org/#type=sequential&scheme=BuGn&n=3>

# Perception of Visual Cues

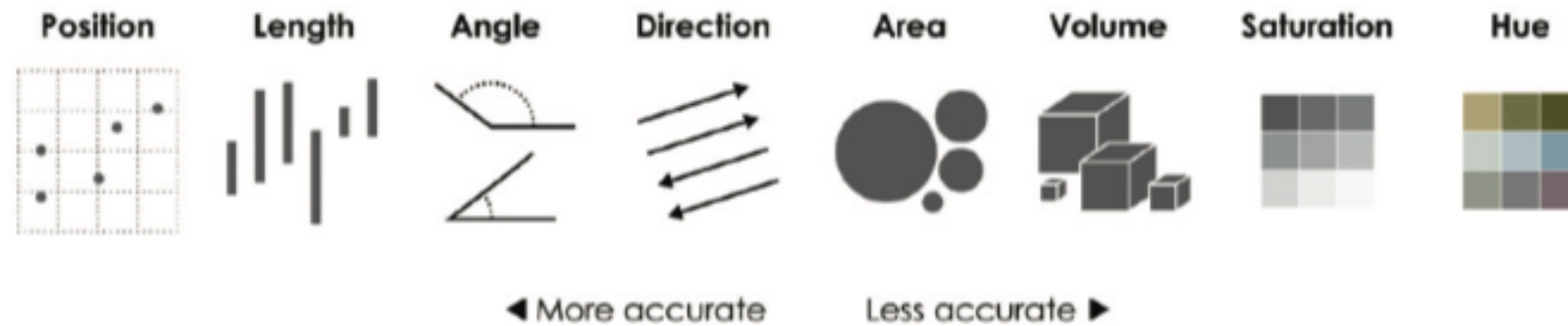


FIGURE 3-12 Visual cues ranked by Cleveland and McGill



# Telling Stories with Data

- Data are represented by numbers and words.
- Data is a representation of something in real life.
- Statistics and visualization can help tell a story.
- It's up to the statistician, data scientist, to decide how to tell that story.

- This graphic presents data in a clear and concise manner.
- Important points, areas are annotated, symbols and colours explained, and it's easy to see the story in the data.
- This is a simple line chart, but design elements help tell a better story.
- Line width and colour direct your eyes to what's important.

## graph<sup>1</sup> | gra:f, graf |

noun

a diagram showing the relation between variable quantities, typically of two variables, each measured along one of a pair of axes at right angles.

- *Mathematics* a collection of points whose coordinates satisfy a given relation.

verb [*with object*]

plot or trace on a graph.

## graphic | 'grafɪk |

adjective

- 1 relating to visual art, especially involving drawing, engraving, or lettering: *his mature graphic work.*
  - *Computing* relating to or denoting a visual image: *graphic information such as charts and diagrams.*
- 2 giving clear and vividly explicit details: *a graphic account of the riots.*
- 3 of or in the form of a graph.
- 4 [*attributive*] *Geology* of or denoting rocks having a surface texture resembling cuneiform writing.

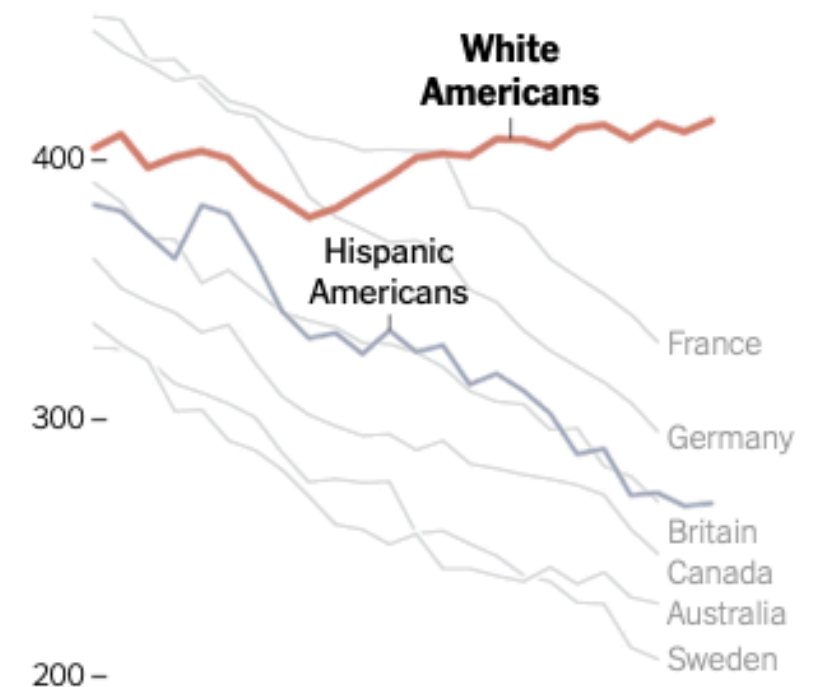
noun *Computing*

a graphical item displayed on a screen or stored as data.

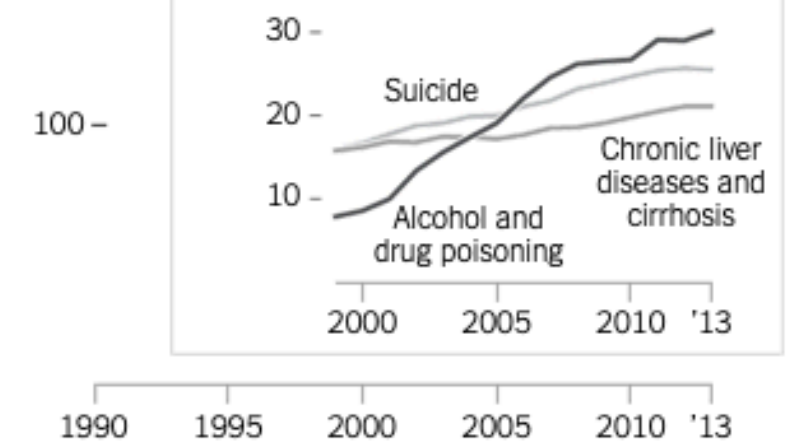
## Dying in Middle Age

Death rates are rising for middle-aged white Americans, while declining in other wealthy countries and among other races and ethnicities. The rise appears to be driven by suicide, drugs and alcohol abuse.

**DEATHS** per 100,000 people aged 45–54



### INCREASING CAUSES OF DEATHS Per 100,000 white Americans, 45–54



Sources: Anne Case and Angus Deaton; PNAS

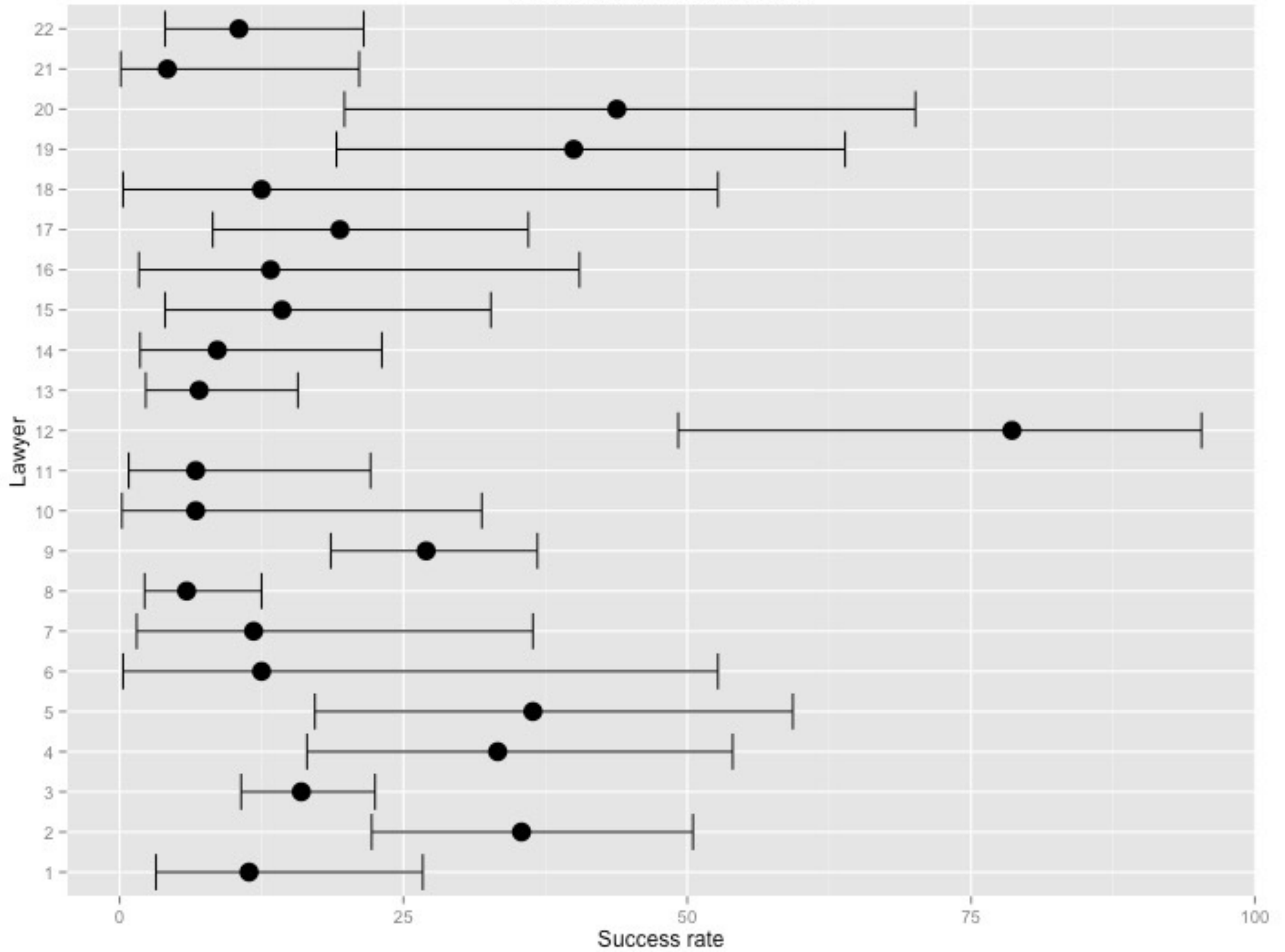
By The New York Times

Counsel	Abandoned/ Withdrawn	Negative	Positive	Total	Recognition Rate	95% Lower Limit	95% Upper Limit
MANZARARU, LEONARD	22	3	11	36	78.6	49.2	95.3
VALOIS, STEPHANIE	24	9	7	40	43.8	19.8	70.1
Vallieres, Alain	20	12	8	40	40	19.1	63.9
GOLDMAN, JEFFREY	9	14	8	31	36.4	17.2	59.3
BHATTI, ROGER	66	31	17	114	35.4	22.2	50.5
FINE, DANIEL	70	18	9	97	33.3	16.5	54
Ivanyi, Peter	149	73	27	249	27	18.6	36.8
SILCOFF, MAUREEN	16	29	7	52	19.4	8.2	36
FARKAS, JOSEPH	223	137	26	386	16	10.7	22.5
Rodrigues, Roger	24	24	4	52	14.3	4	32.7
SARKOZI, JOZEF	10	13	2	25	13.3	1.7	40.5
GRICE, JOHN	26	7	1	34	12.5	0.3	52.7
TAHERI, DJAWID	65	7	1	73	12.5	0.3	52.7
HEGYI, ILDIKO	14	15	2	31	11.8	1.5	36.4
	297	31	4	332	11.4	3.2	26.7
YOUNES, DIANA	18	51	6	75	10.5	4	21.5
NO COUNSEL, IDENTIFIED	67	32	3	102	8.6	1.8	23.1
NO COUNSEL,	66	66	5	137	7	2.3	15.7
JASZI, ELIZABETH	80	14	1	95	6.7	0.2	31.9
KORMAN, MICHAEL	25	28	2	55	6.7	0.8	22.1
HOHOTS, VIKTOR	403	95	6	504	5.9	2.2	12.5
Wang, Yaqian	9	23	1	33	4.2	0.1	21.1

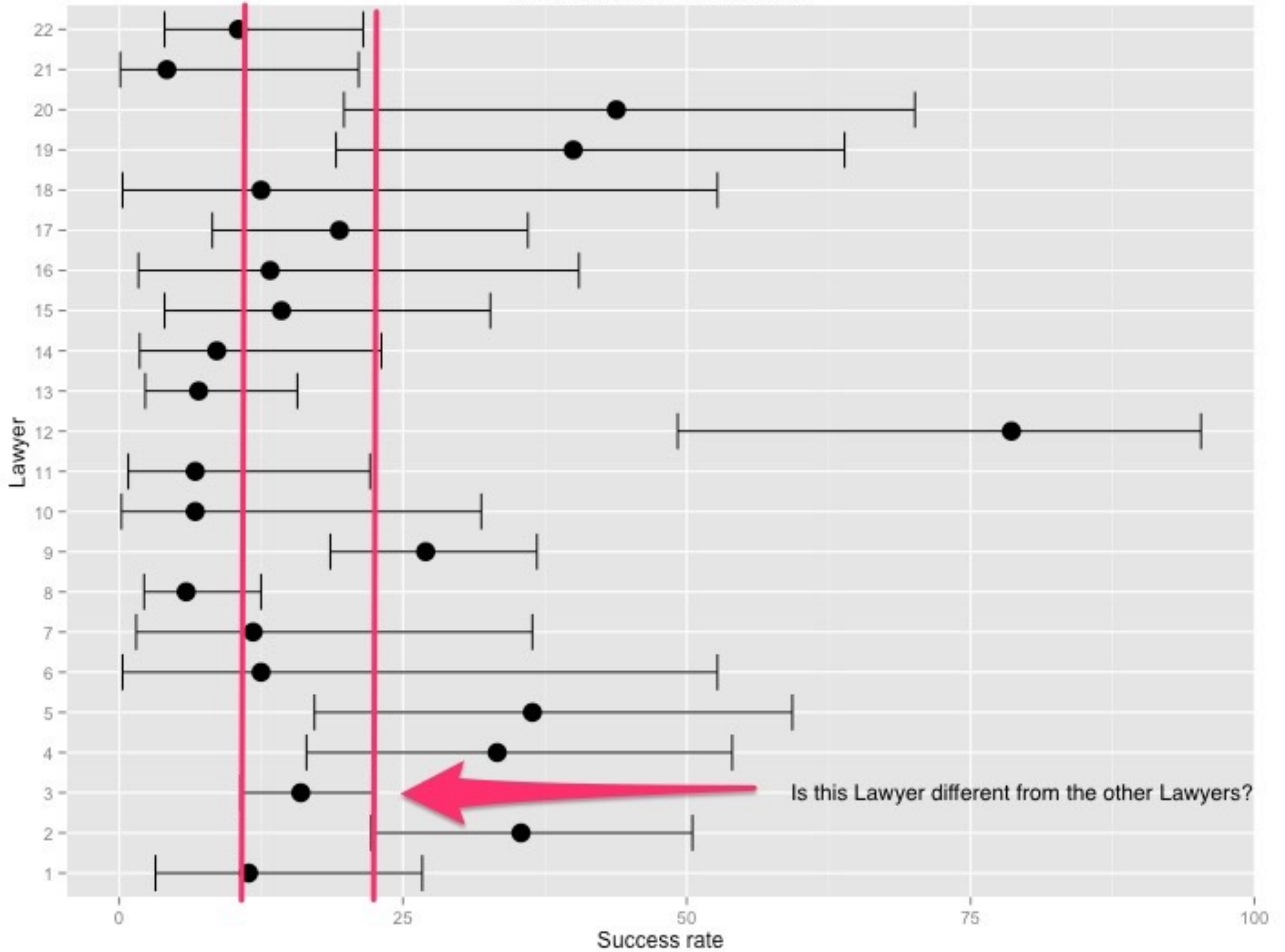
**Table 2: 2008-2012 recognition rates for high volume counsel (25+ decisions) with 95% Confidence**

**Intervals - ranked by recognition rate**

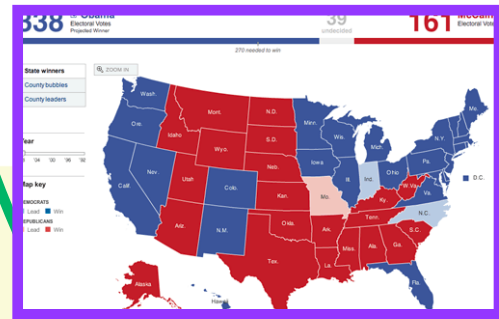
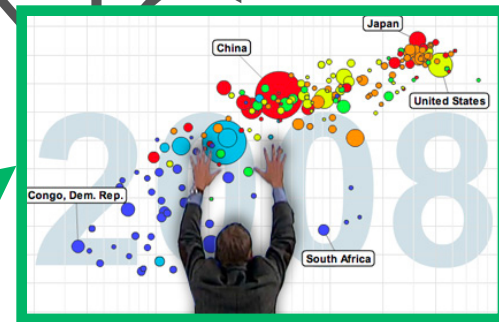
95% Confidence Intervals



# 95% Confidence Intervals



# VISUALIZATION ENCODING PIPELINE



## Visualization Encoding Pipeline

Structured Data map to Visual Attributes draw as Marks plot on a Layout

Nominal

Apple, Banana, Pear

Ordered

Mint, Good, Fair, Poor

Quantitative

0, 3, 4.2, -31.2,  $6.6 \times 10^6$

Position



Size



Hue



Shape



Brightness



Etc



Point



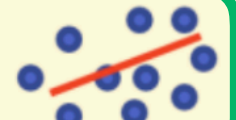
Line



Area



Scatter



Graph



Line



Tree map



Etc





# What Do Data Visualizations Show?

- Patterns
- Relationships

# State of the World

Consider three estimates about the state of the world:

- A. Life expectancy at birth is 70 years
- B. The literacy rate of youth females ages 15 to 24 is 87 percent,
- C. The gross domestic product is approximately \$70 trillion.

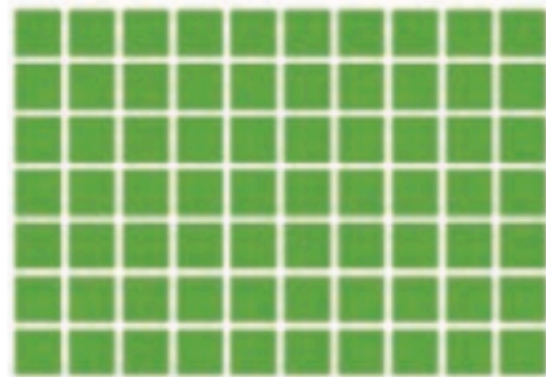
Should you visualize this data?



# Random numbers about the world

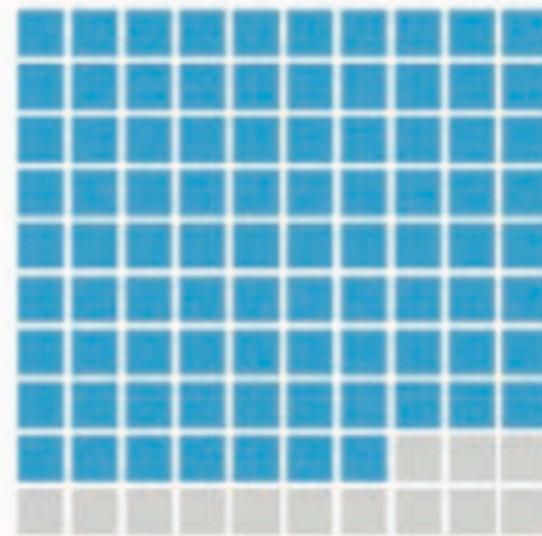
Life expectancy

70 years



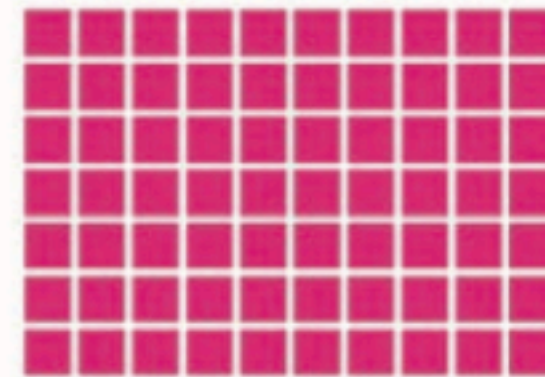
Literacy rate of youth females

87%



Gross domestic product

\$70 trillion



# Random numbers about the world

LIFE  
EXPECTANCY

70 years

LITERACY RATE OF  
YOUTH FEMALES

87%

GROSS DOMESTIC  
PRODUCT

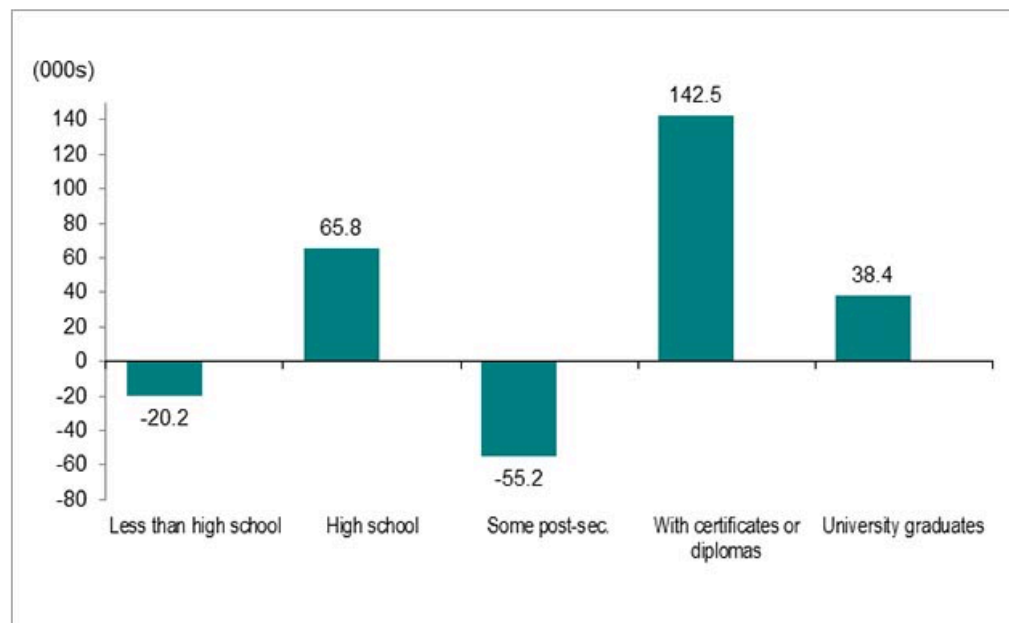
\$70 trillion

# Patterns

- Changes over time.
- Data can be split in different ways to reveal different patterns.

## Employment increase and decrease by education level

Chart 4 shows Ontario employment change by highest level of education attained, aged 25 and older, January 2018 to January 2019.

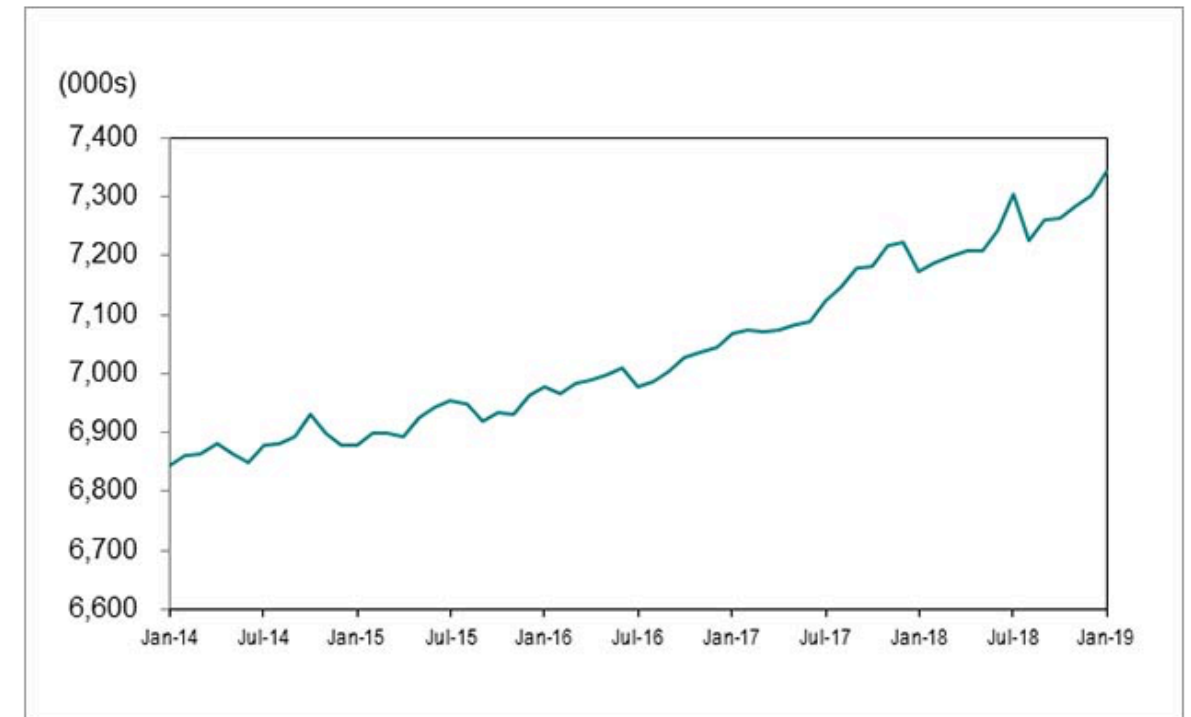


Source: Statistics Canada, Labour Force Survey, Table 14-10-0019-01, unadjusted data

## Employment increased in January

Employment in Ontario increased in January (41,400), after rising by 16,100 jobs in December. January's job gain was the largest increase since July 2018.

Chart 1 shows employment in Ontario from January 2014 to January 2019.



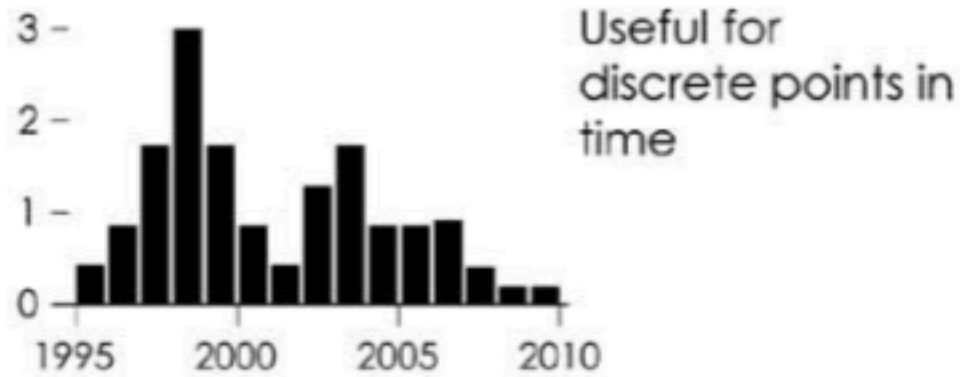
Source: Statistics Canada, Labour Force Survey, Table 14-10-0019-01, (seasonally adjusted data).

<https://www.ontario.ca/page/labour-market-report-january-2019>

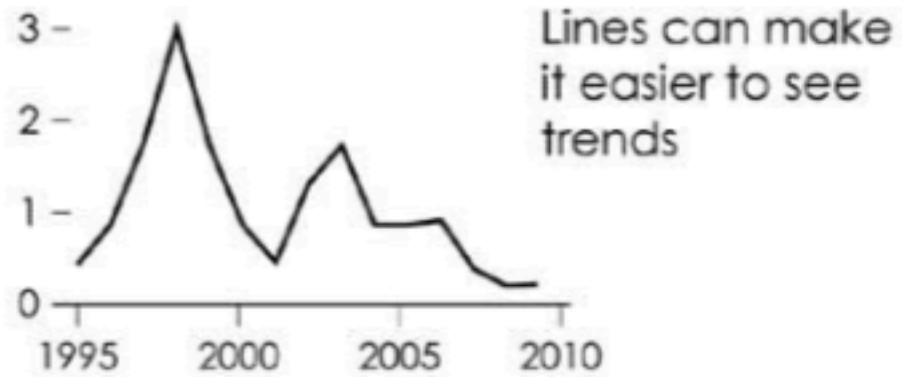
## Time series

There are a variety of ways to see patterns over time, using cues such as length, direction, and position.

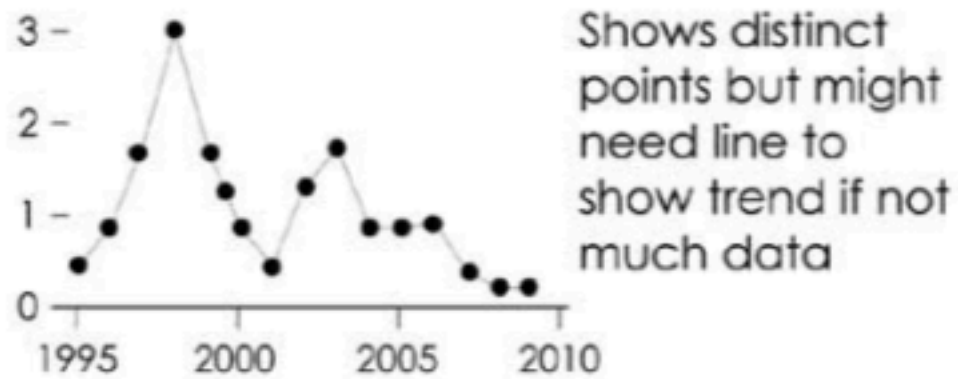
### Bar graph



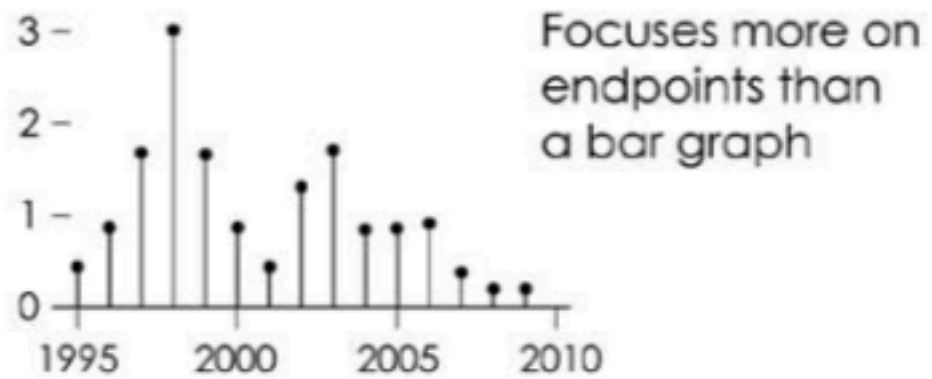
### Line chart



### Dot plot



### Dot-bar graph

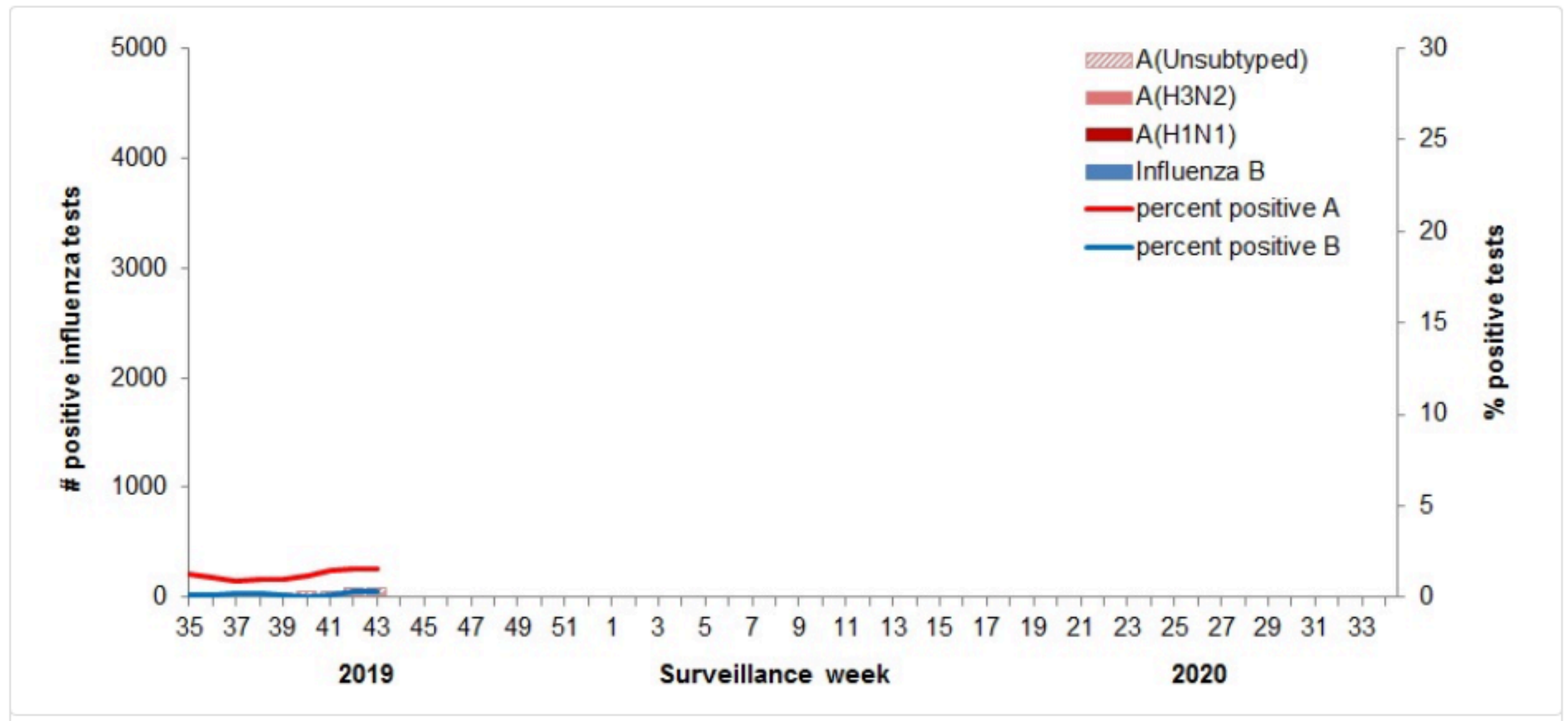


# Answer These Questions Before Presenting A Visualization ...

- What is your message?
- Who is your audience?
- What does your audience need to know?

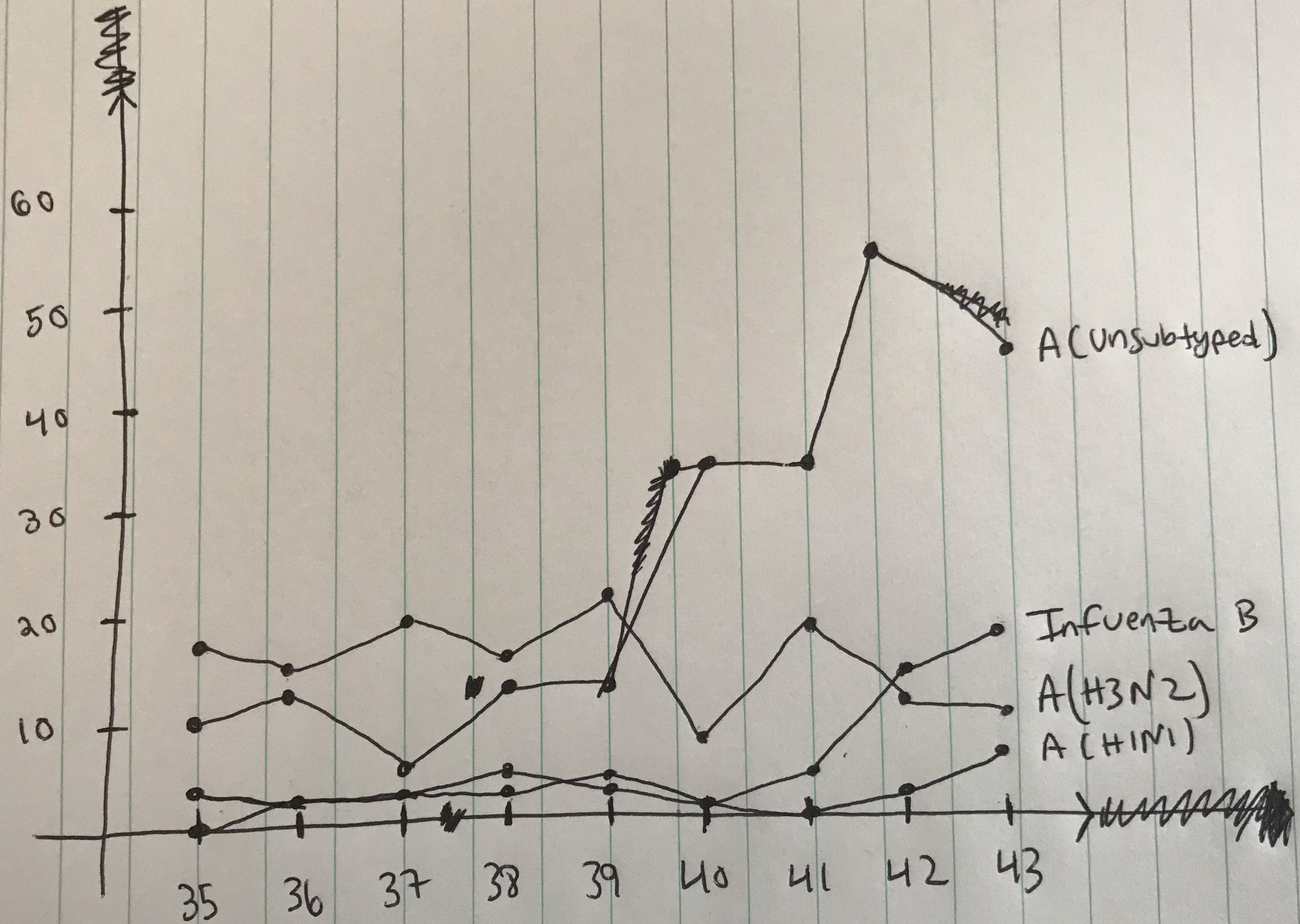
**Figure 2 - Number of positive influenza tests and percentage of tests positive, by type, subtype and report week, Canada, week 2019-43**

Number of Laboratories Reporting in Week 43: 33 out of 34



<https://www.canada.ca/en/public-health/services/publications/diseases-conditions/fluwatch/2019-2020/week-43-october-20-26-2019.html>





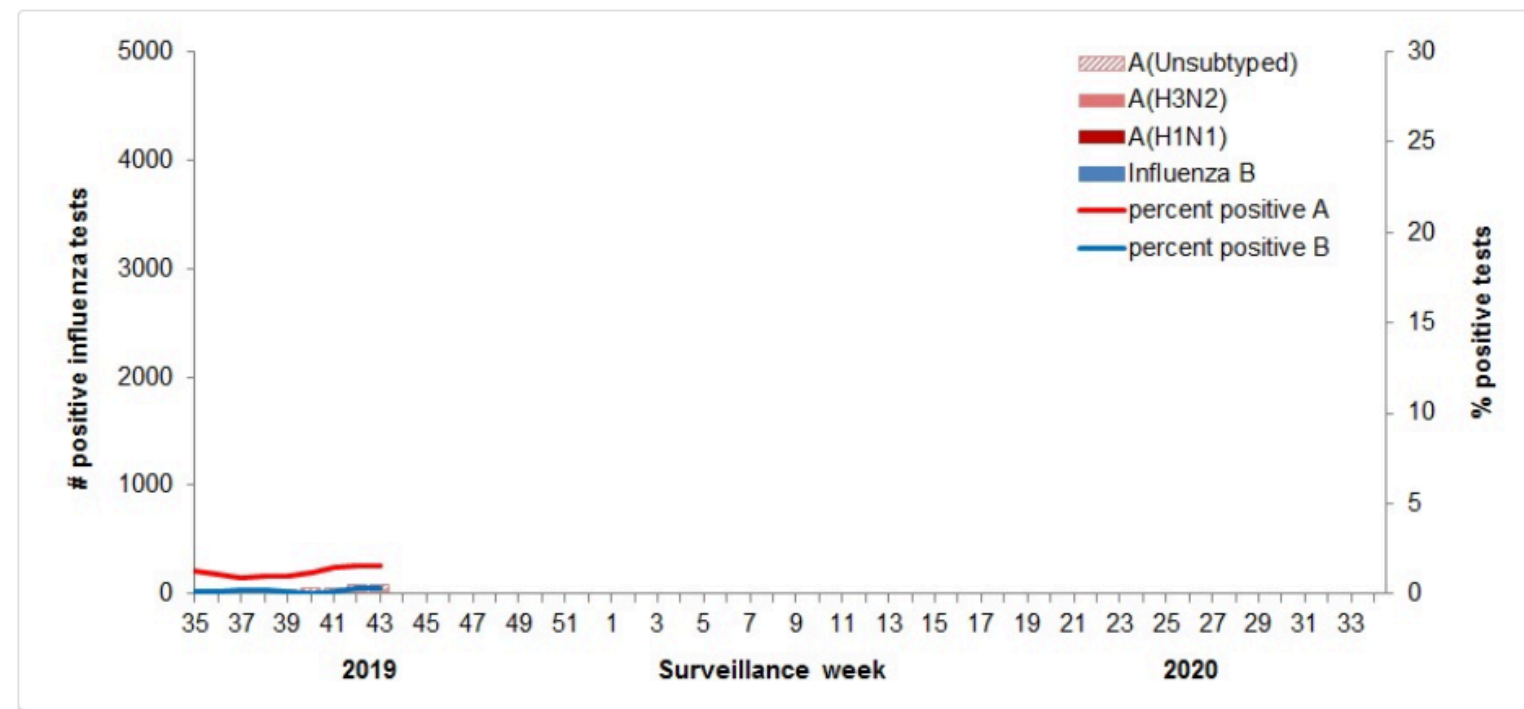


▼ Figure 2 - Text equivalent

Surveillance Week	A(Unsubtyped)	A(H3N2)	A(H1N1)pdm09	Influenza B	Percent Positive A	Percent Positive B
35	10	16	0	2	1.3	0.1
36	11	13	2	2	1.1	0.1
37	5	17	2	5	0.9	0.2
38	11	15	3	6	1.0	0.2
39	11	21	2	3	1.0	0.1
40	34	9	1	2	1.2	0.1
41	34	18	0	5	1.4	0.1
42	54	12	1	14	1.6	0.3
43	45	12	6	17	1.6	0.3

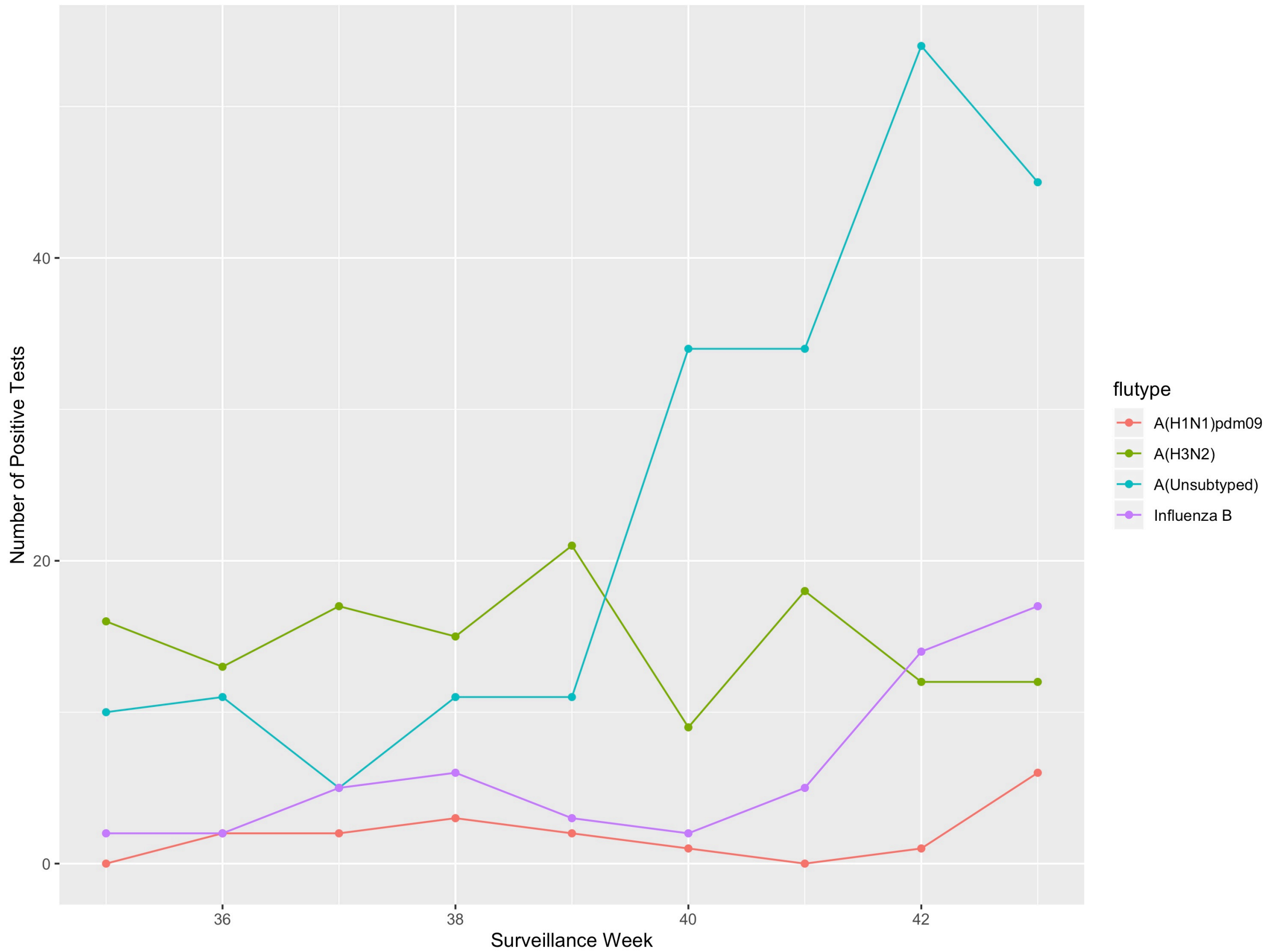
**Figure 2 - Number of positive influenza tests and percentage of tests positive, by type, subtype and report week, Canada, week 2019-43**

Number of Laboratories Reporting in Week 43: 33 out of 34



# Lab Confirmed Influenza in Week 43

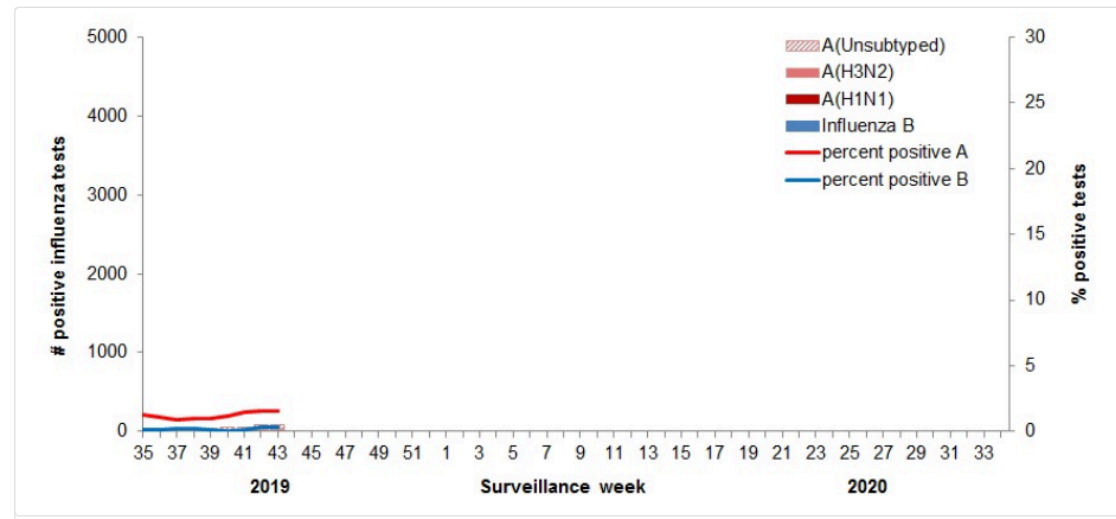
Source: Government of Canada FluWatch





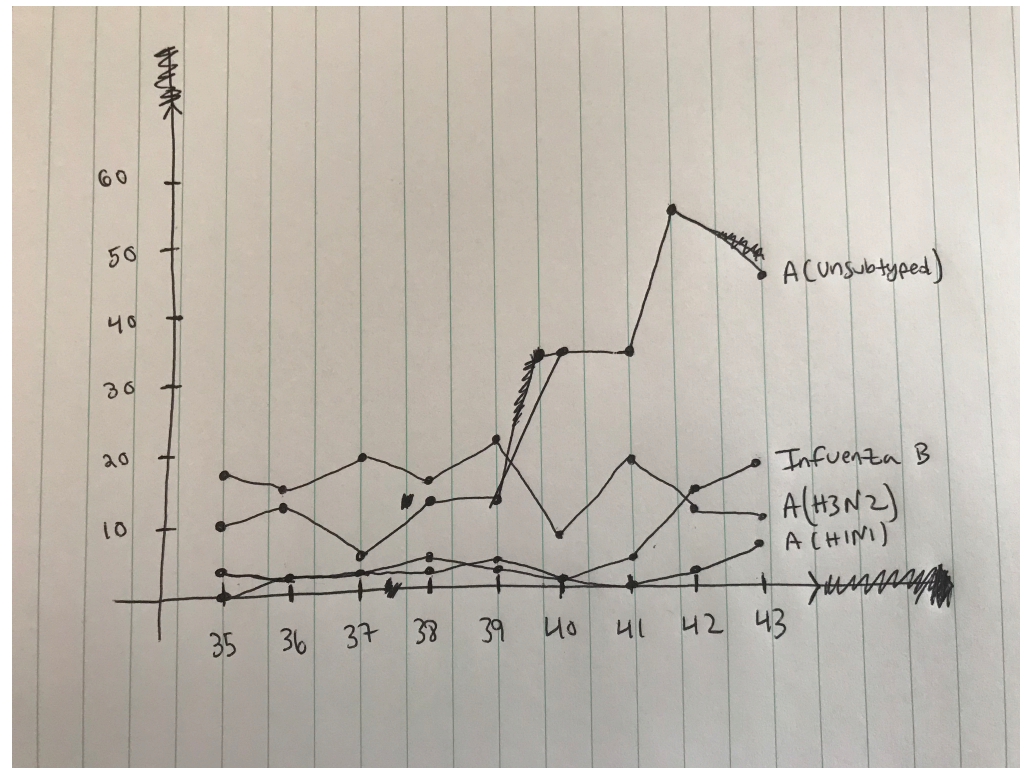
**Figure 2 - Number of positive influenza tests and percentage of tests positive, by type, subtype and report week, Canada, week 2019-43**

Number of Laboratories Reporting in Week 43: 33 out of 34



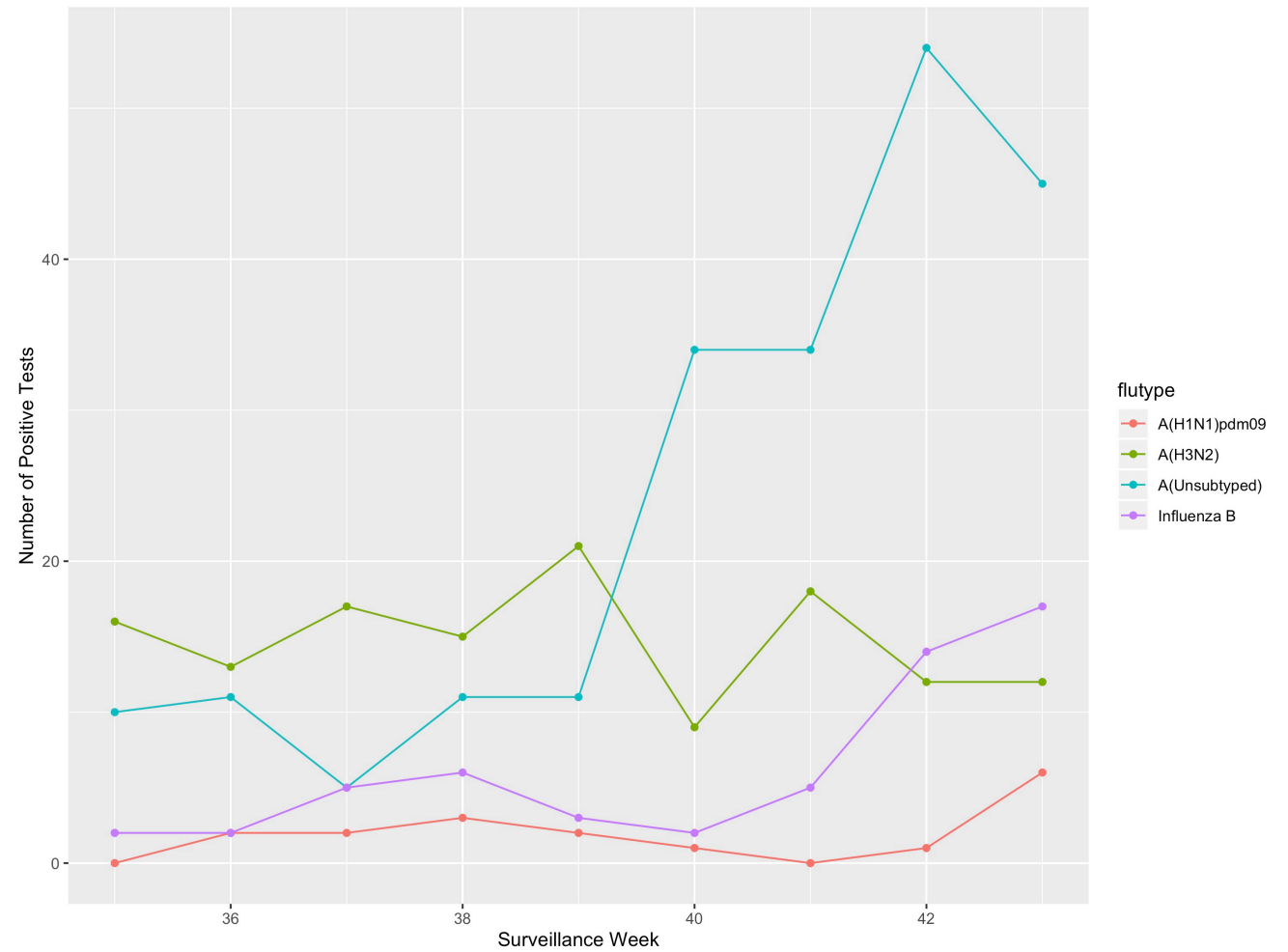
▼ Figure 2 - Text equivalent

Surveillance Week	A(Unsubtyped)	A(H3N2)	A(H1N1)pdm09	Influenza B	Percent Positive A	Percent Positive B
35	10	16	0	2	1.3	0.1
36	11	13	2	2	1.1	0.1
37	5	17	2	5	0.9	0.2
38	11	15	3	6	1.0	0.2
39	11	21	2	3	1.0	0.1
40	34	9	1	2	1.2	0.1
41	34	18	0	5	1.4	0.1
42	54	12	1	14	1.6	0.3
43	45	12	6	17	1.6	0.3



Lab Confirmed Influenza in Week 43

Source: Government of Canada FluWatch



# Presenting Data to People

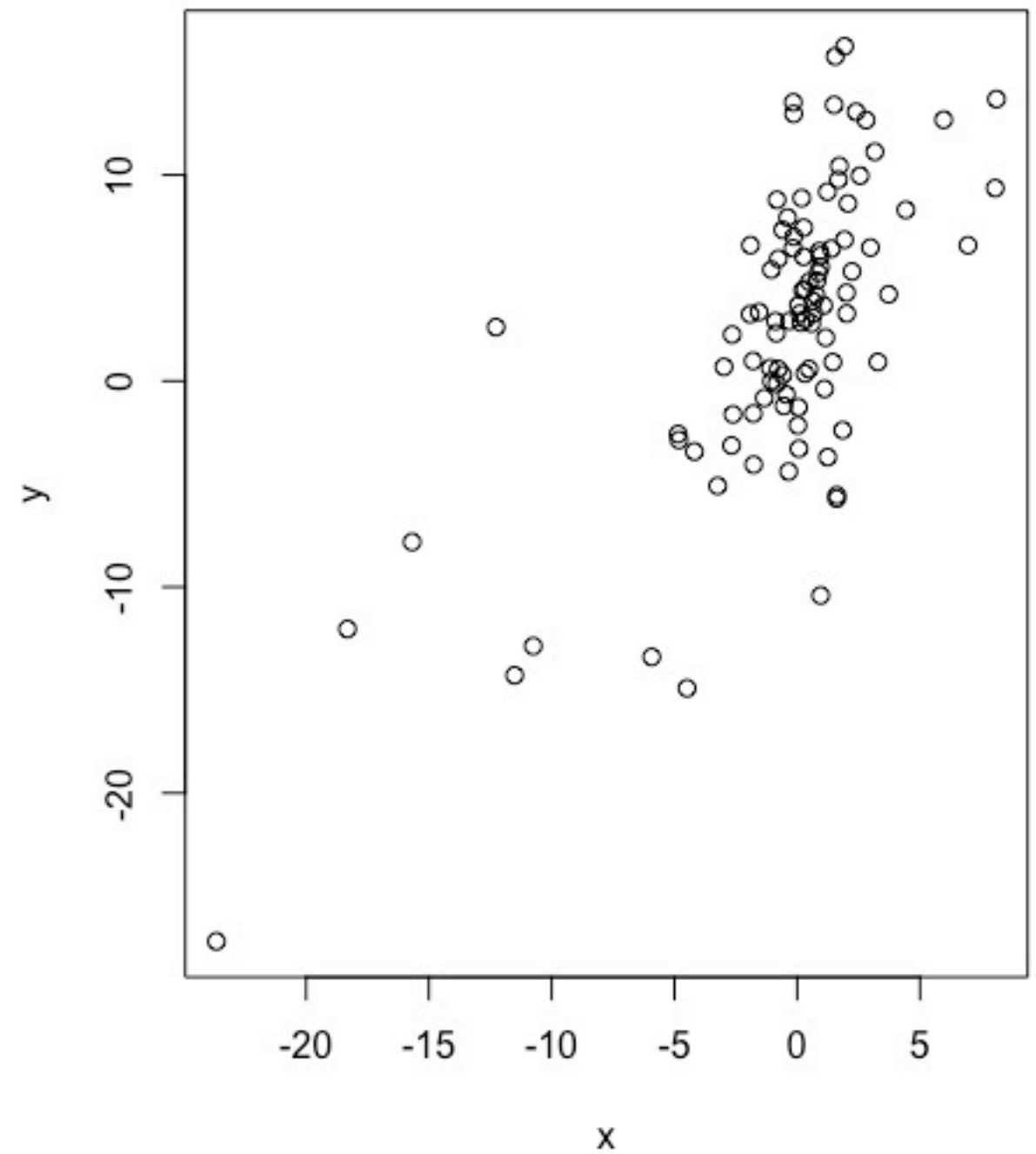
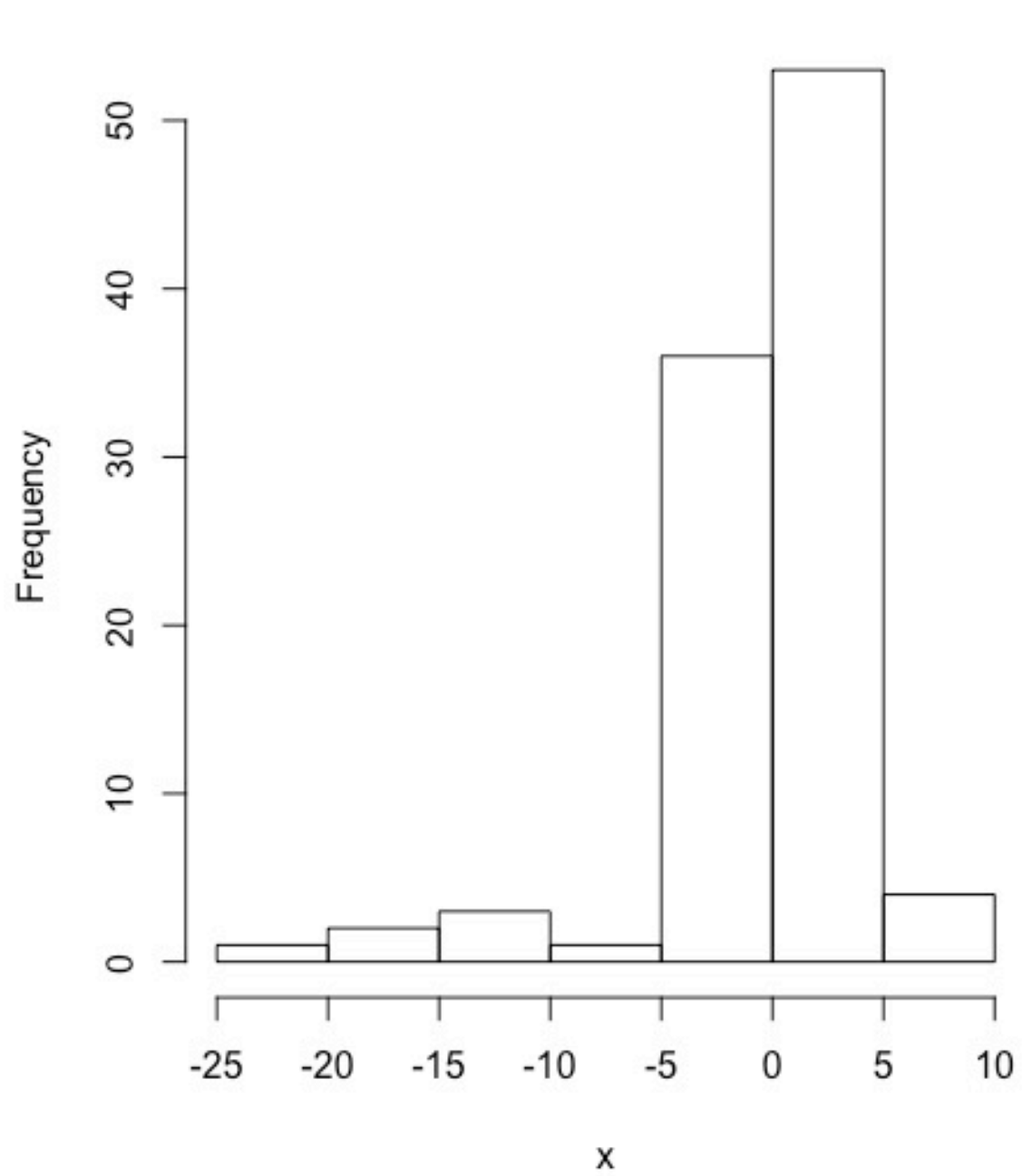
- Me, myself, and I
- A specific audience
- A wider audience

# Presenting Data to People

- How much control does the audience have over the presentation?
- How much detail can they get?

# Me, myself, and I

Histogram of x

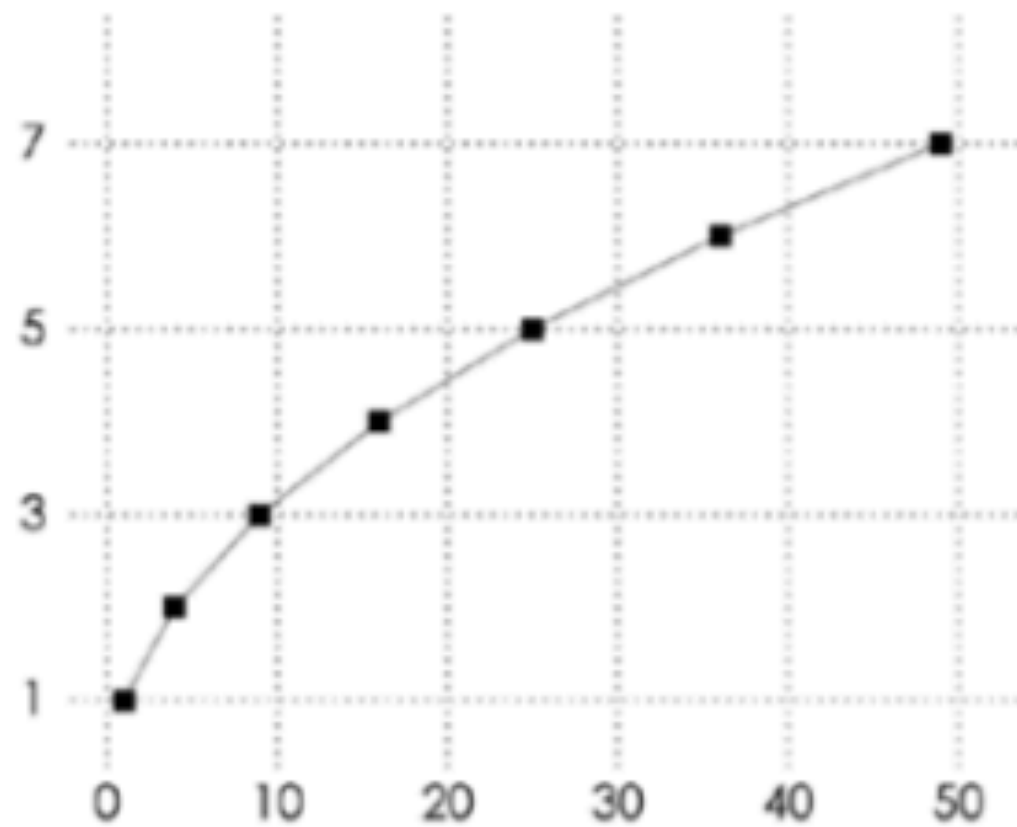


# A Specific Audience

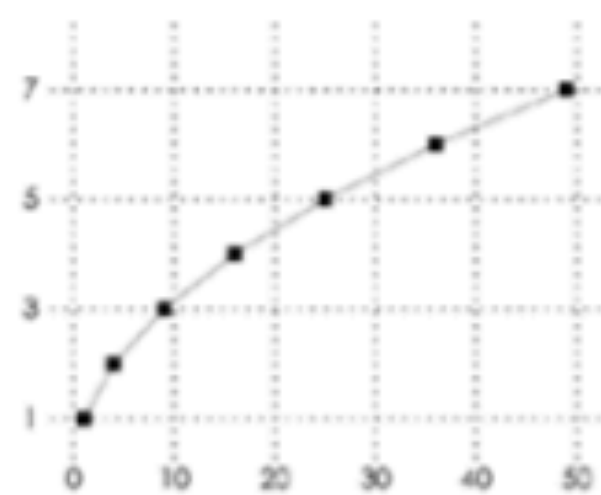
- Your audience should be able to decode your encodings so that they can understand the data.
- If your audience is already familiar with the background behind your data or has perhaps even worked with it, the barriers are lower, but still exist.
- Consider how your audience will examine your work.

# Visualization In A Presentation

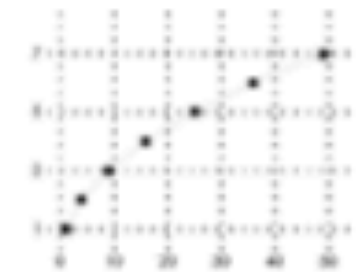
You can see this okay.



This too, if you squint.



Um, what?



# Designing For A Wider Audience

- As your audience grows so do the challenges, such as the range of data literacy, and familiarity with your data's context.
- Avoid jargon and be sure you explain complex concepts in a way so that people can relate.

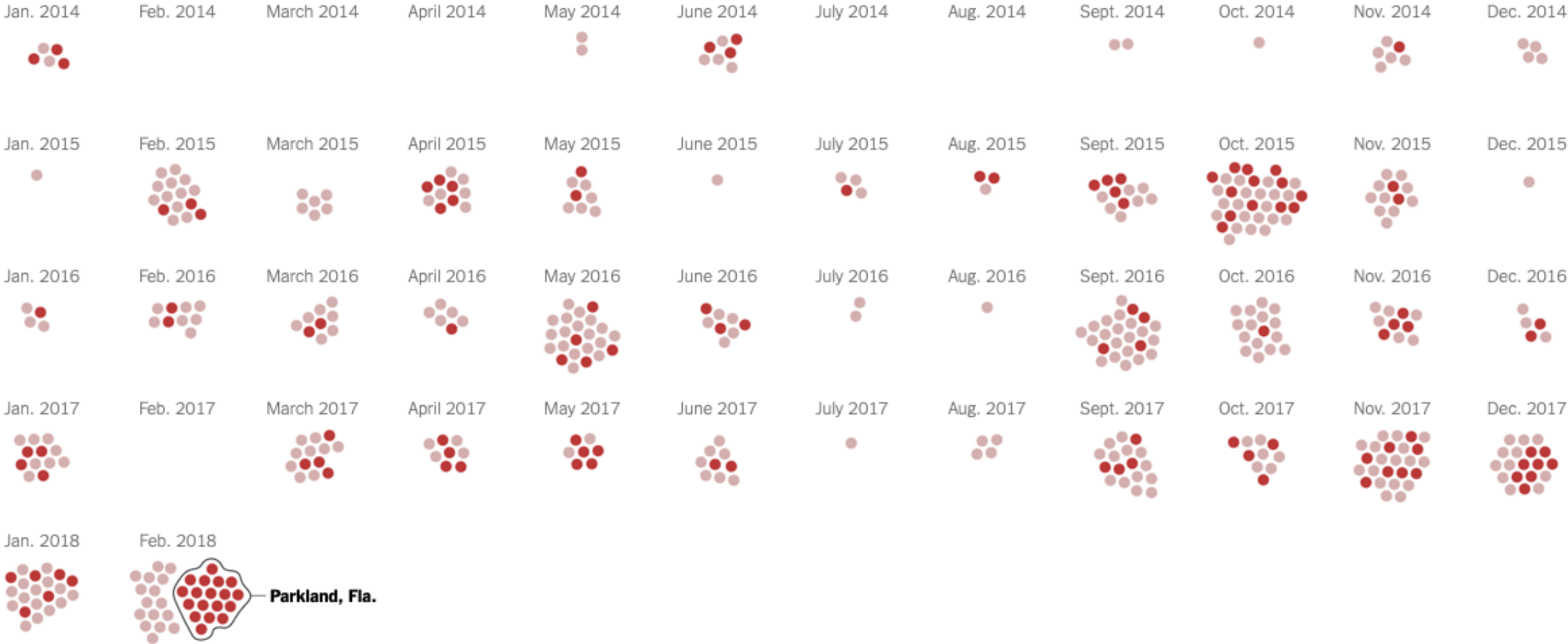
# After Sandy Hook, More Than 400 People Have Been Shot in Over 200 School Shootings

By JUGAL K. PATEL FEB. 15, 2018

## Gunshot Victims in School Shootings

● Killed ● Injured

### Sandy Hook



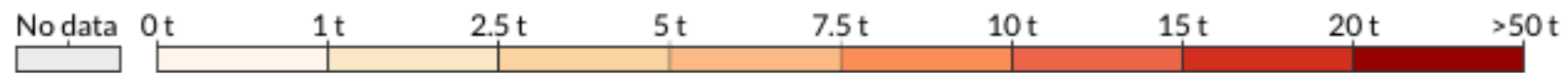
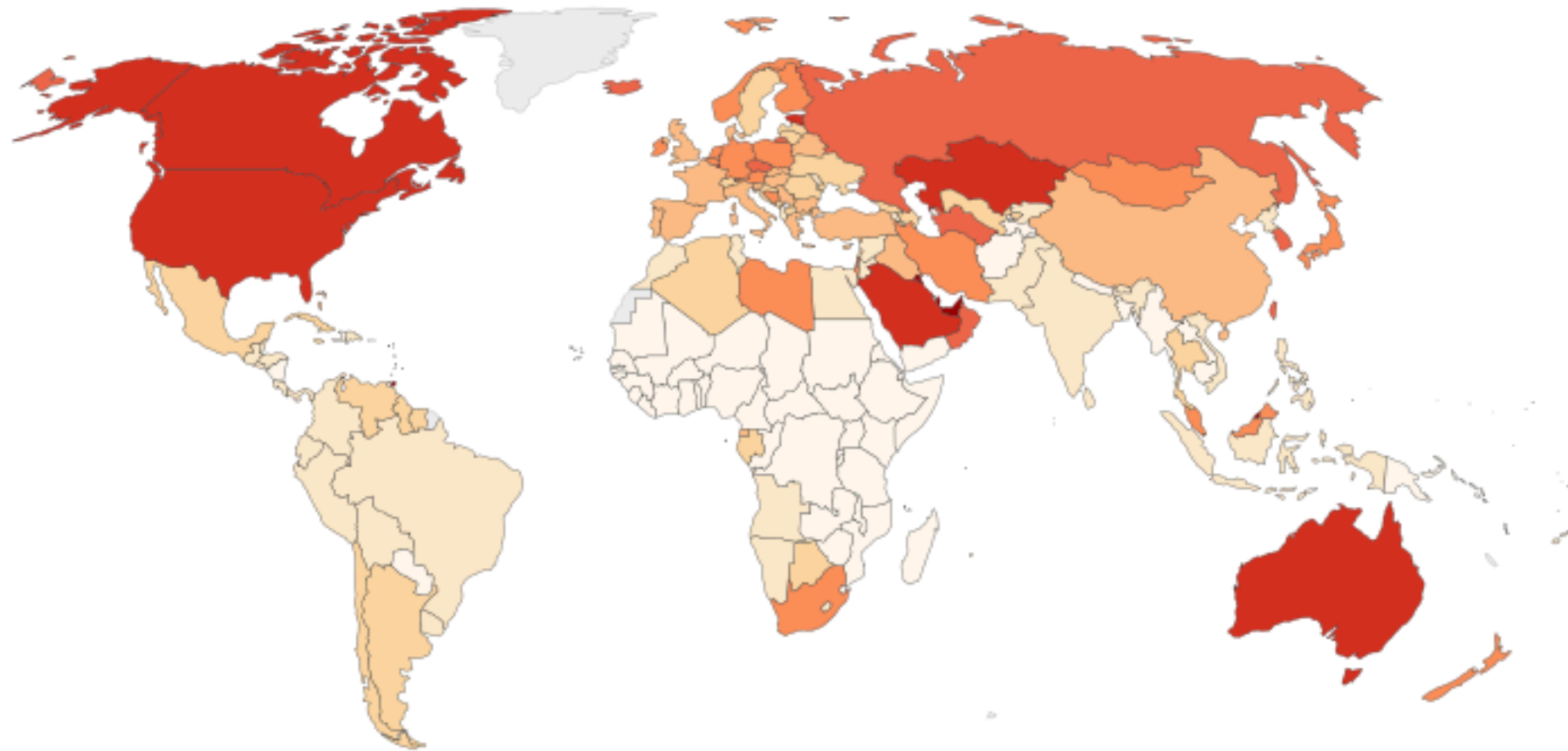
Source: Gun Violence Archive

Note: Shootings in 2013 are not included because complete data was not available in that year. Months with blanks indicate no shootings archived.



# CO<sub>2</sub> emissions per capita, 2017

Average carbon dioxide (CO<sub>2</sub>) emissions per capita measured in tonnes per year.



Source: OWID based on CDIAC; Global Carbon Project; Gapminder & UN

CC BY

▶ 1800  2017 CHART MAP DATA SOURCES

<https://ourworldindata.org/per-capita-co2>

# Things to Consider

- Imagine you are a tourist in a new place.
- What do you want a tour guide to tell you?
- It's your job to point out the direction of interest, provide background, and make sure you don't confuse people.

# Data Provenance

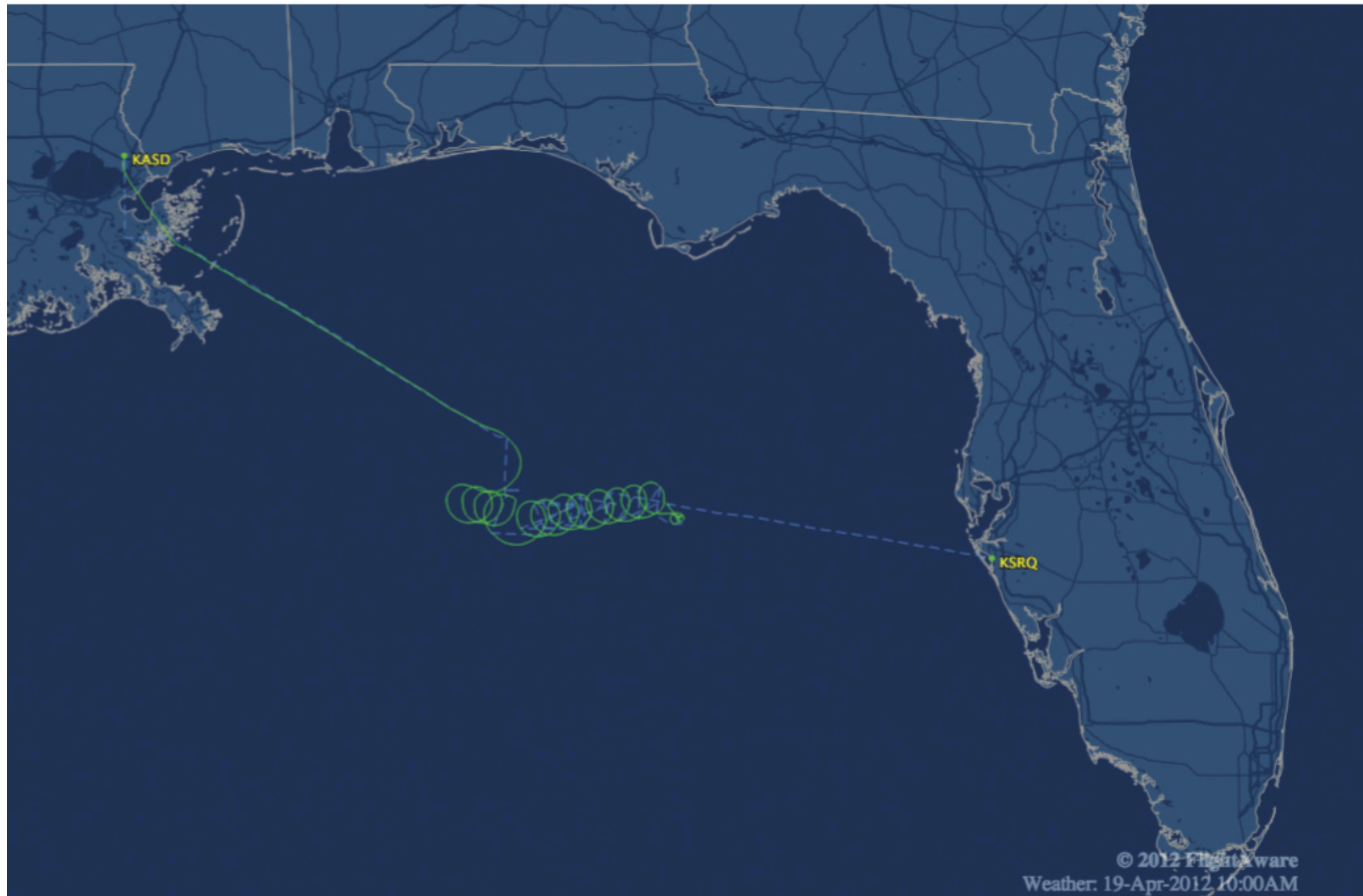

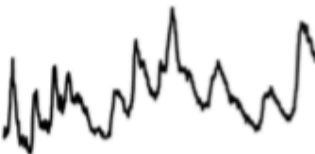

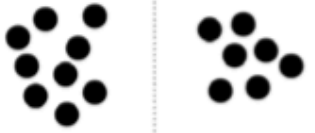
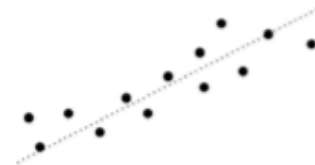


FIGURE 6-17 A flight from Slidell, Louisiana to Sarasota, Florida, according to FlightAware, <http://flightaware.com/live/flight/N48DL>

# Data Narrative

- Ask a question about the data and then try to answer that through the visualization.
- How do you want your audience to read the read? How will your audience read your graph?

Possible questions Fill in the blanks	Statistical concepts	Possible visuals
What _____ is the best and worst?	Maximums and minimums	
How has _____ changed over time?	Temporal patterns	
What _____ stands out from the rest?	Outliers	
What makes _____ different from _____?	Clustering	
How are _____ and _____ related to each other?	Correlation	
What's the breakdown for _____?	Distributions	