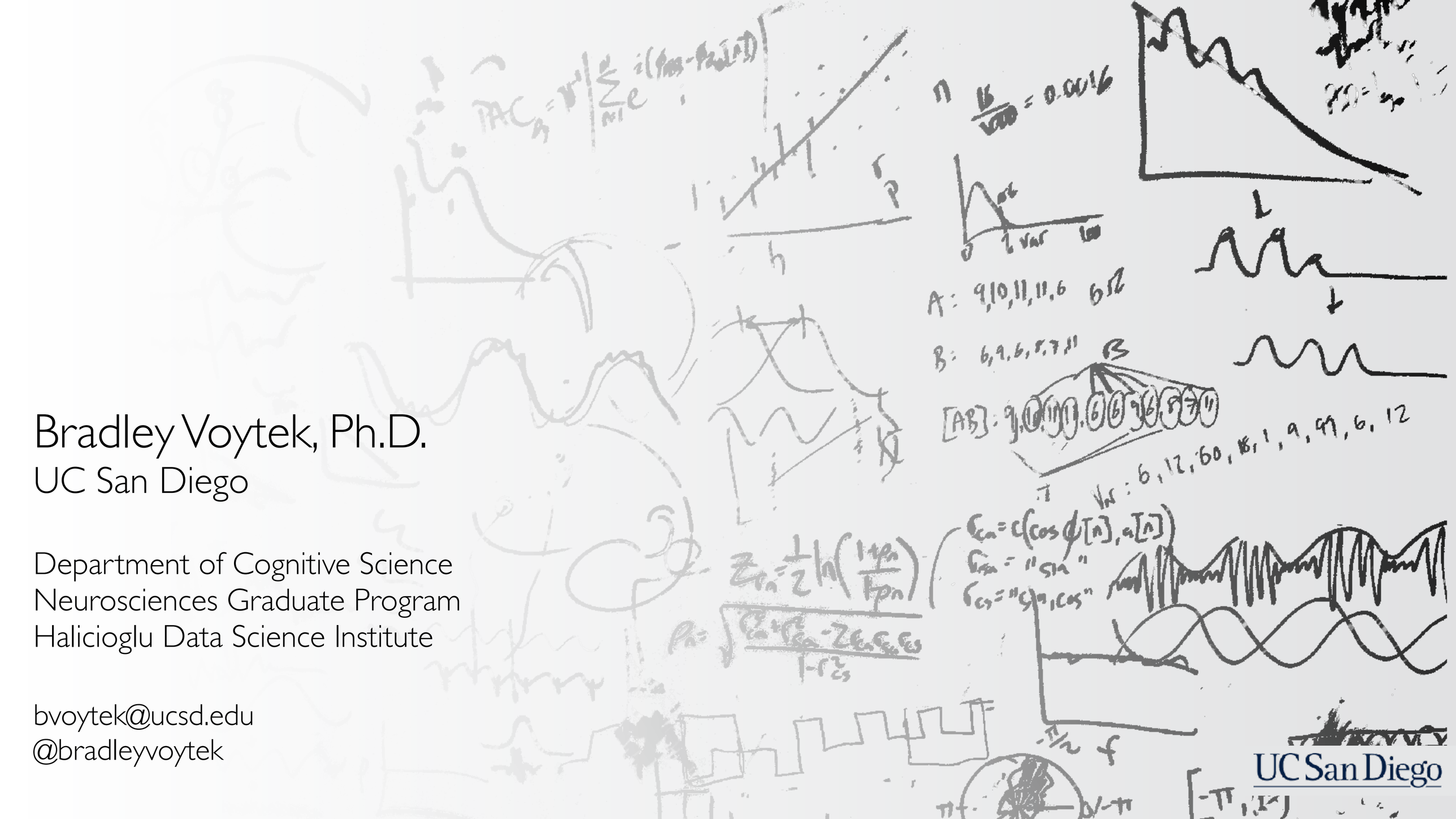


Bradley Voytek, Ph.D.

UC San Diego

Department of Cognitive Science
Neurosciences Graduate Program
Halicioglu Data Science Institute

bvoytek@ucsd.edu
@bradleyvoytek

UC San Diego

Administrative stuff

- **REQUIRED LECTURE ANNOUNCEMENTS!**
  - İlkay Altıntaş: Chief Data Science Officer, San Diego Supercomputer Center
  - 2018 February 13 (Tuesday)

  - Special guest!
  - 2018 February 20 (Tuesday)

Survey!

- **http://bit.ly/quickfbsurvey**

COGS 108
Data Science in Practice

Data Visualization

UC San Diego

Data visualization

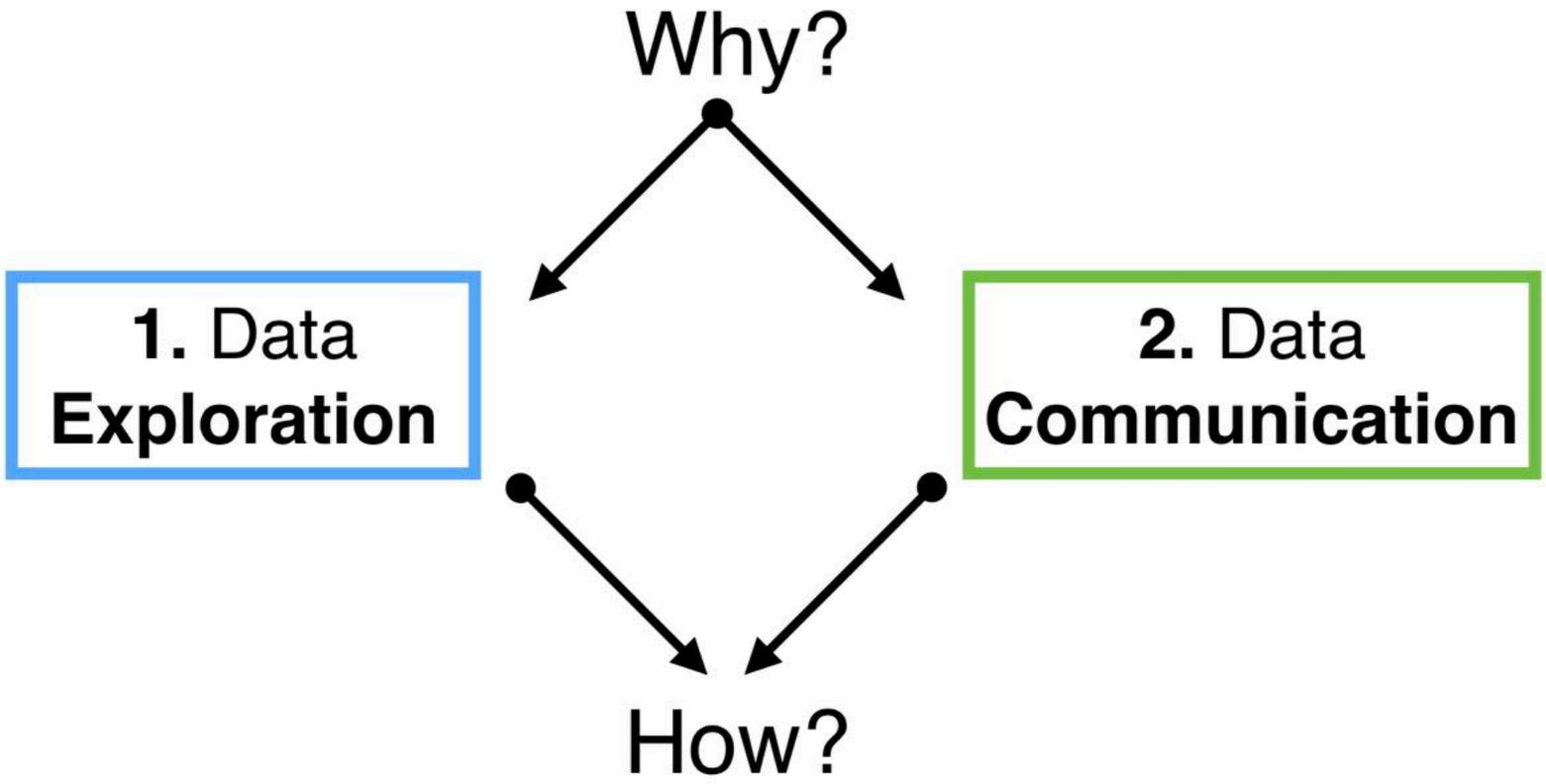
- Visualizing your data should be your very first action.

Data visualization

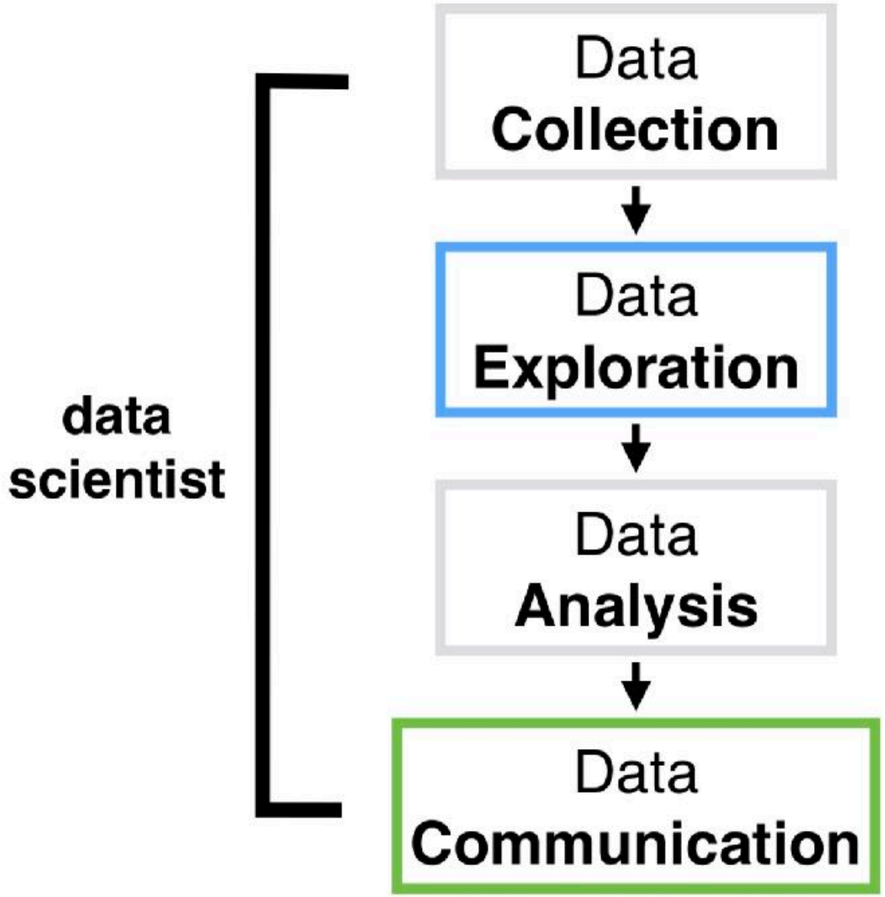
Visualizing your data should be
your very first action.

This act lays the groundwork for
everything else that comes after.

Data visualization



Data exploration



Data exploration

Goal: get fast intuition about your data and avoid Garbage-In-Garbage-Out

Things to look out for: data quality (noise), distributions, outliers

Python tools for basic exploration

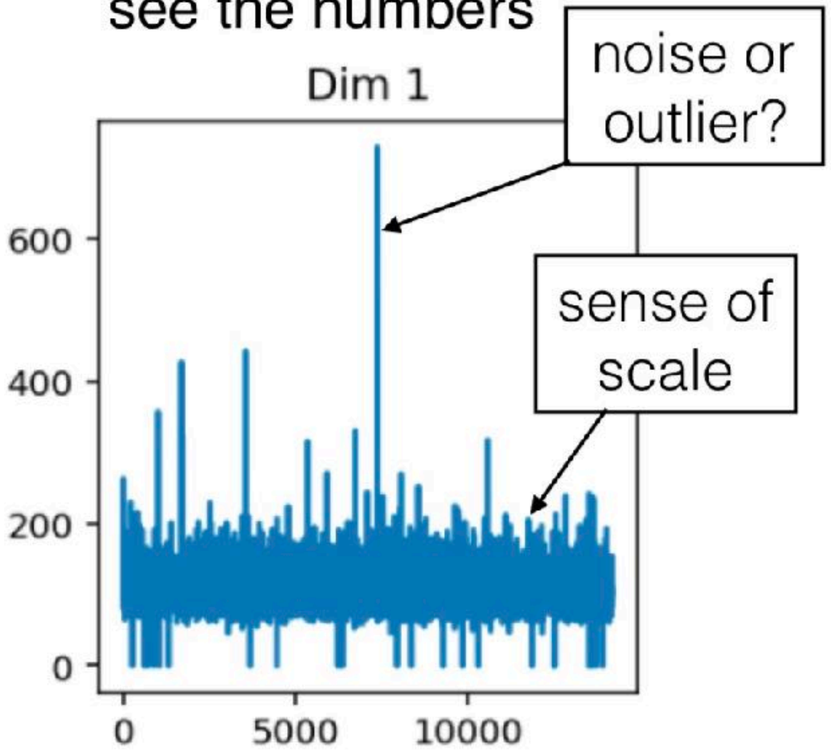
Data exploration

1956	118	1588	1588	6.2
1982	120	31589	31589	6.9
1946	103	2640	2640	7
2005	95	48291624	85309	6.7
1981	131	560	560	6.7
1986	93	13431806	11274	6.4
1949	263	1258	1258	6.2
1982	82	620	620	2.2
2010	80	2231024	29346	7.5
1966	110	3235	3235	6.5
2014	90	1428647	11292	7
1957	85	297	297	6.3
2001	120	50173190	154520	7.4
1978	96	1474	1474	5.7
1957	125	1497	1497	6.6
1963	88	415	415	5.2
1982	103	4870	4870	7.9
1984	87	1867	1867	3.5
1976	100	168	168	4.1
1990	101	37963281	12247	6.7

- What questions are we trying to answer?
- What types of analyses do we want to apply?
- What unexpected findings are there?
- Staring at the numbers is **NOT** the answer.

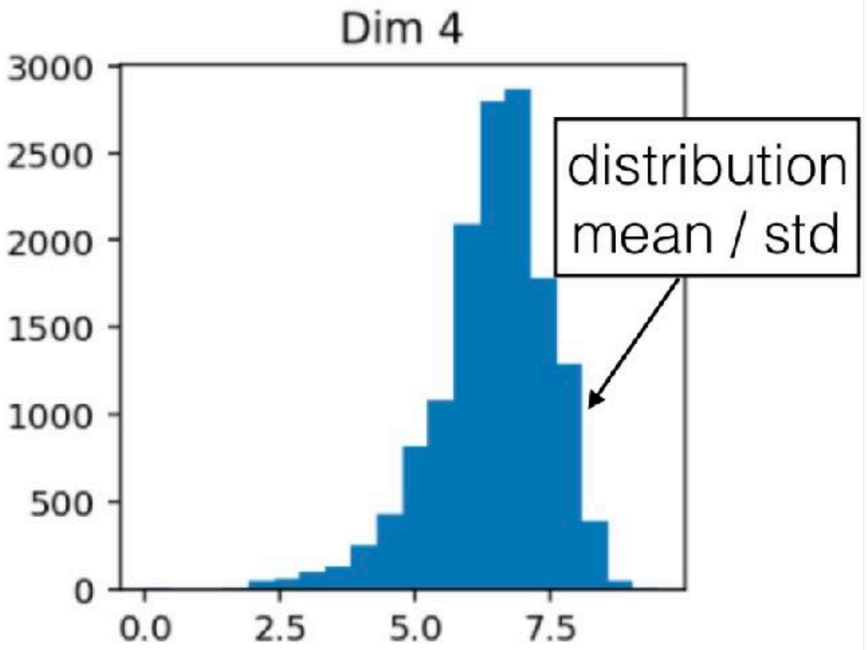
Basic data visualization

"Plot" plot
see the numbers



`plt.plot()`

Histogram
summary stats

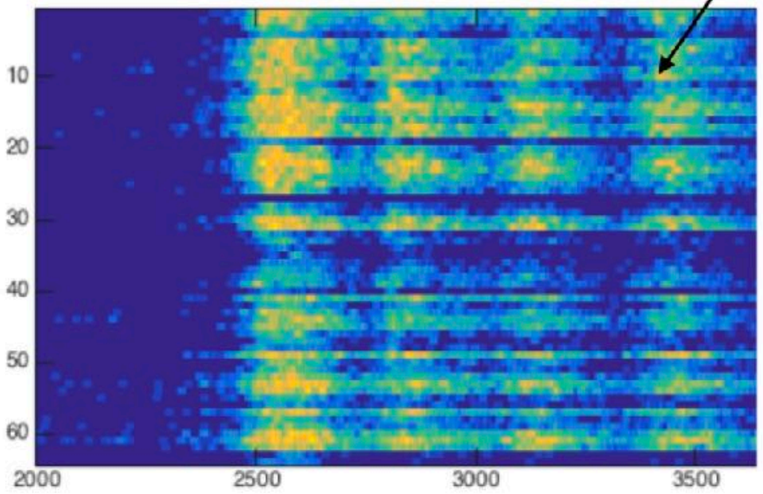


`plt.hist()`

Basic data visualization

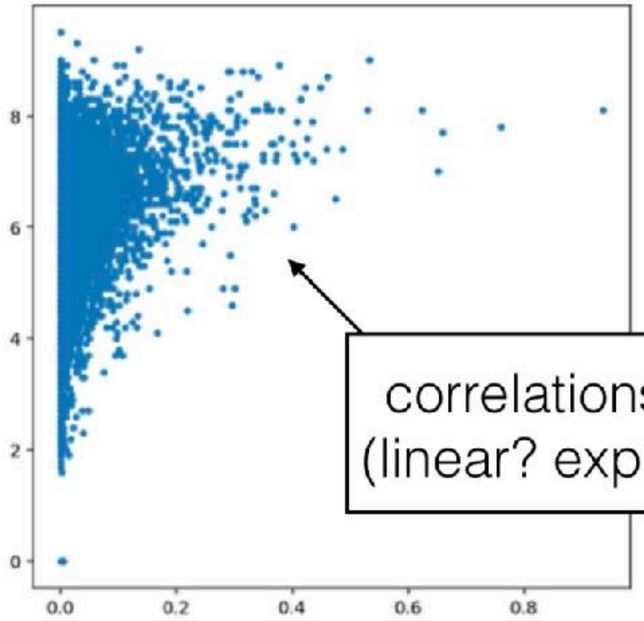
Heatmap
multi-dimensional

global patterns



`plt.imshow()`

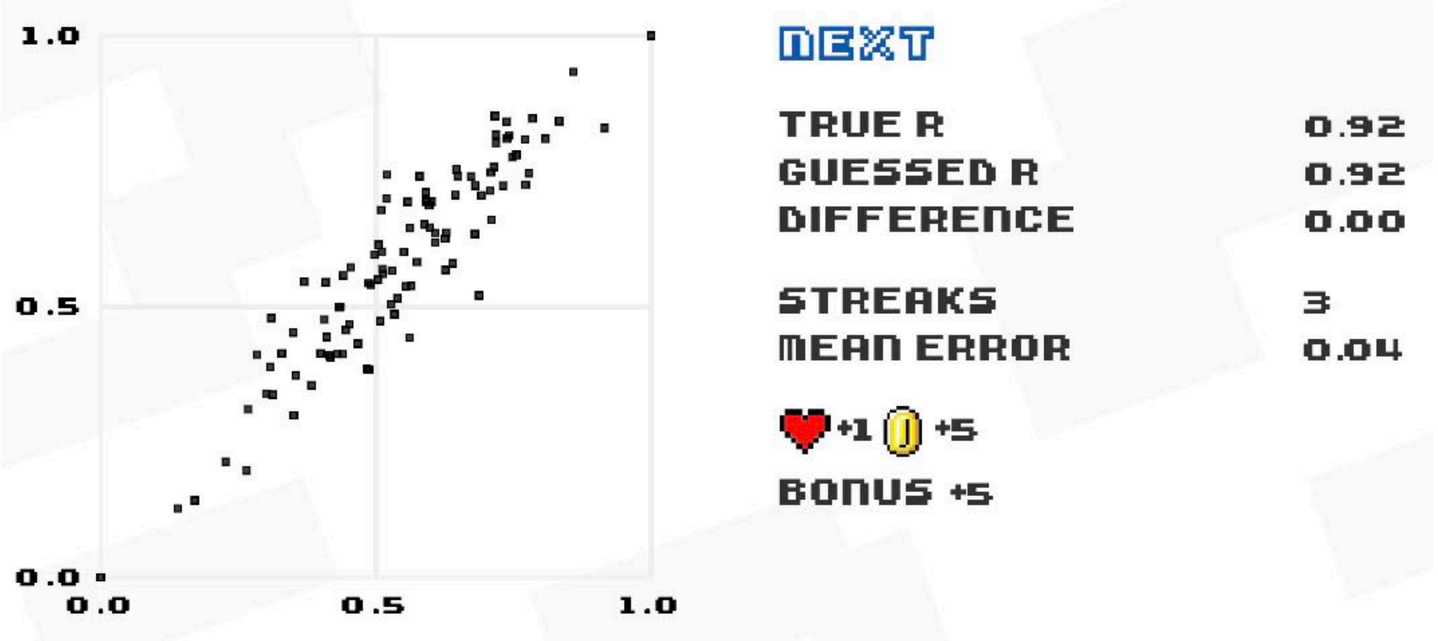
Scatter plot
X-Y relationships



`plt.scatter()`

COGS108Wi18-07 notebook

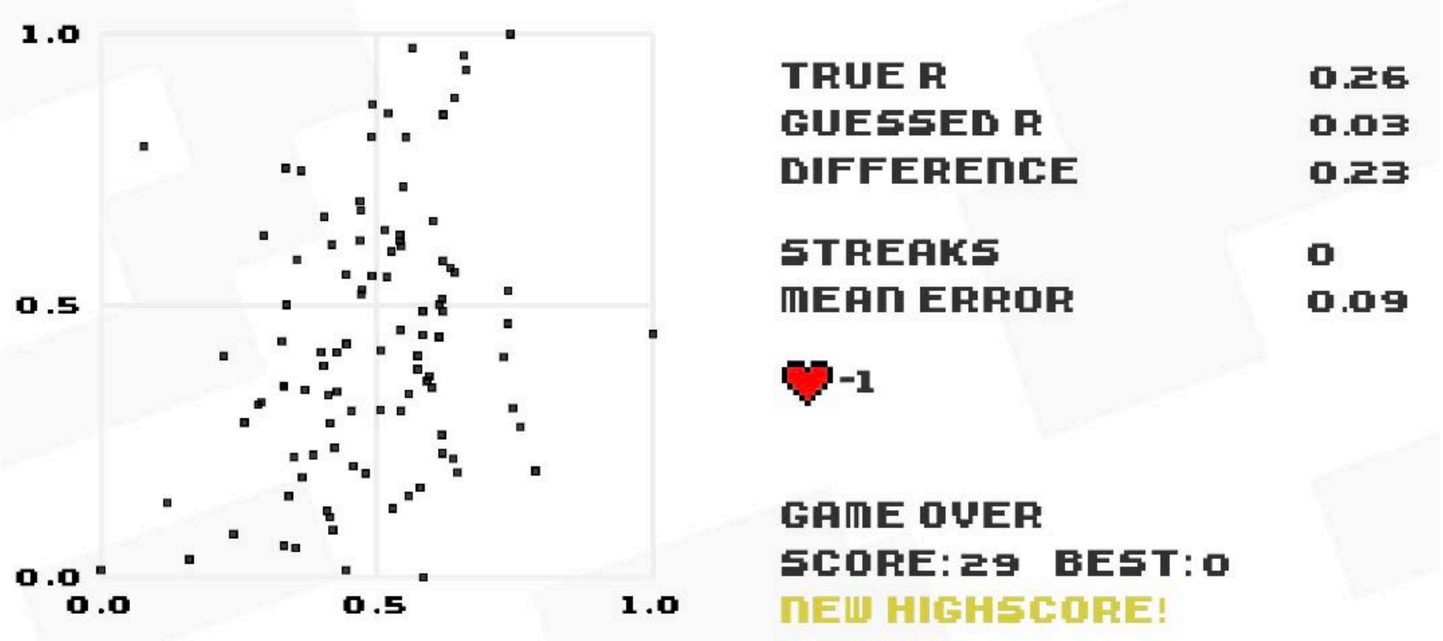
Data visualization CAUTION



- Even if you ARE awesome…

Source: http://guessthecorrelation.com/

Data visualization CAUTION

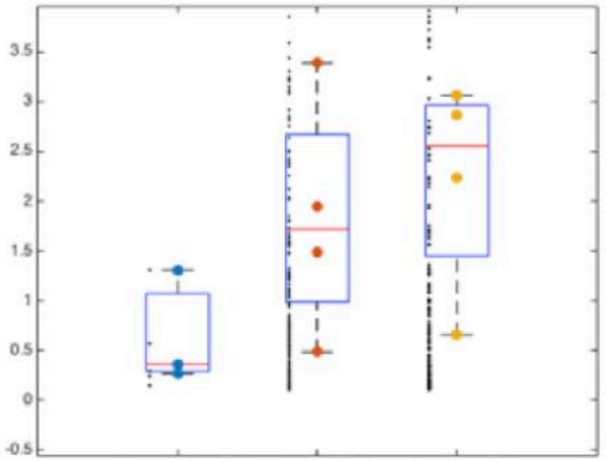


- Even if you ARE awesome…
- Your eyes ***will*** deceive you.
- So quantify when possible!

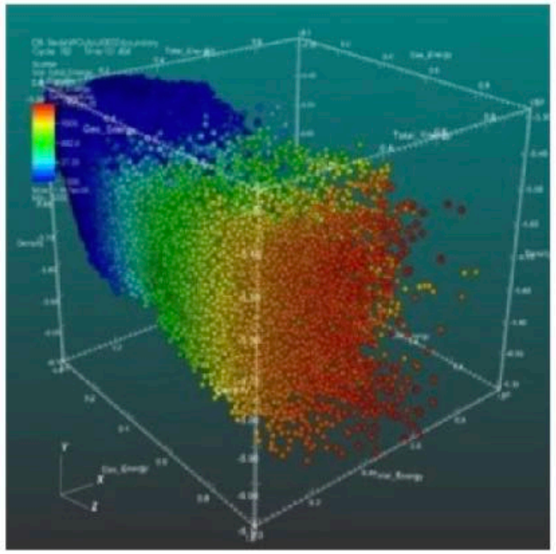
Source: http://guessthecorrelation.com/

Basic data visualization

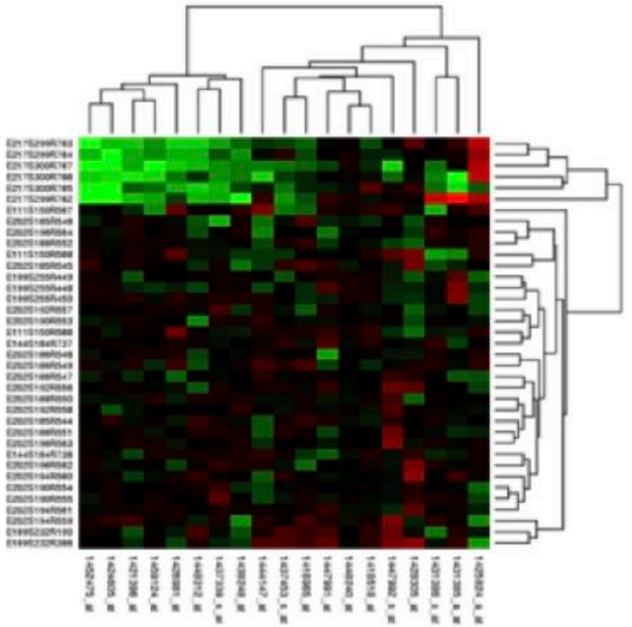
Boxplot



3D scatter

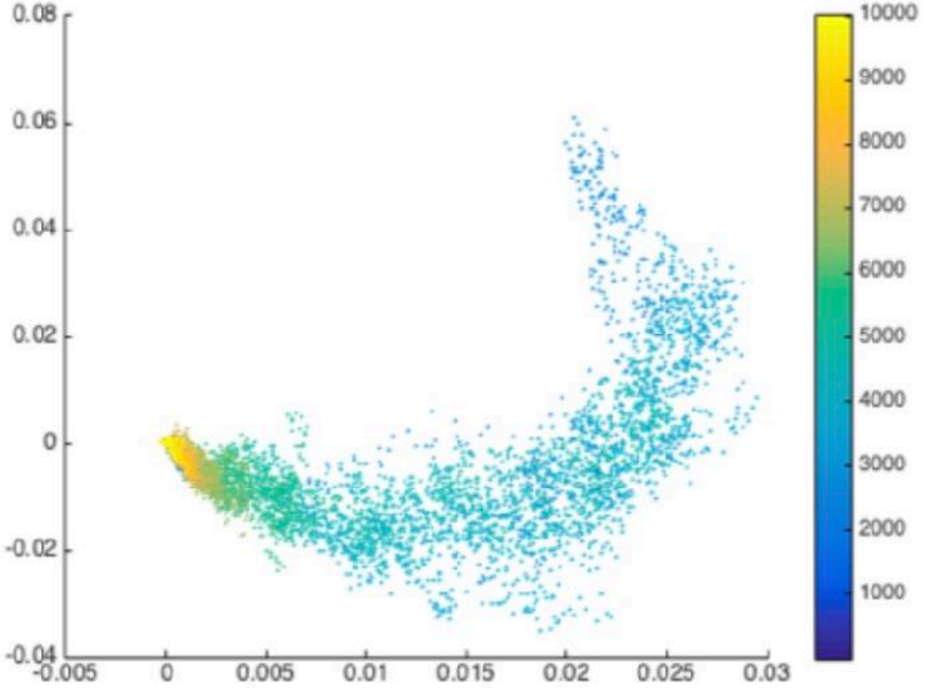


Domain specific

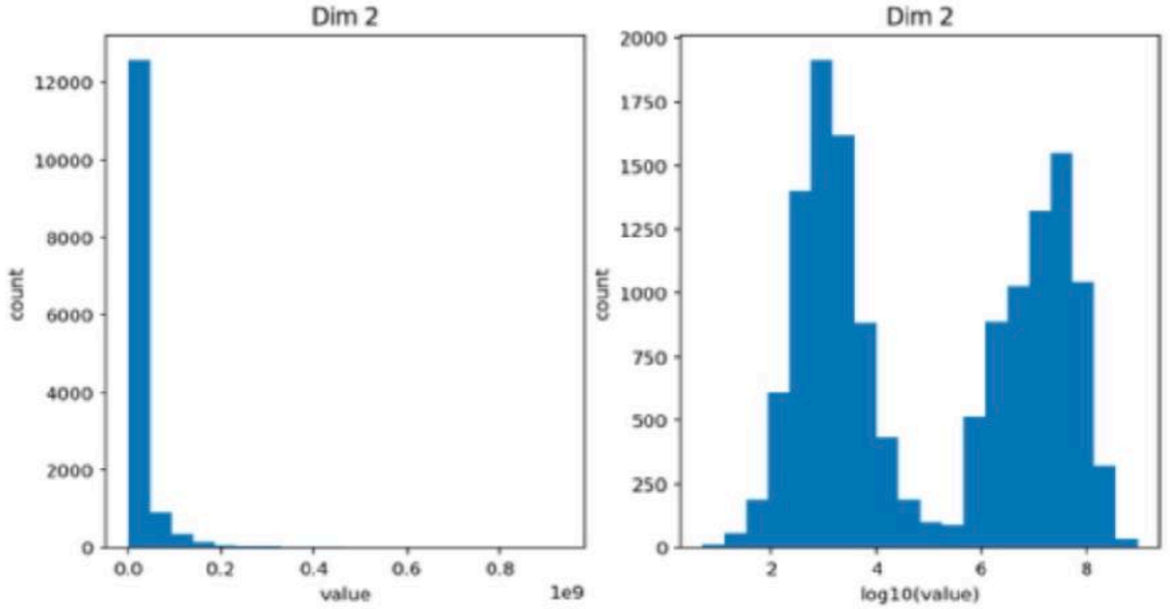


Basic data visualization tips

Use color to represent more dimensions (like time)



Rescale values to better separate data



Data visualization in Python

- **Matplotlib:** comprehensive functionality
  - **Seaborn:** beautify matplotlib

- **Bokeh:** interactive, pretty
  - **Plotly:** similar, but needs sign-up

Data visualization takeaways

Visually manipulate data to distill information

High dimensional to low dimensional

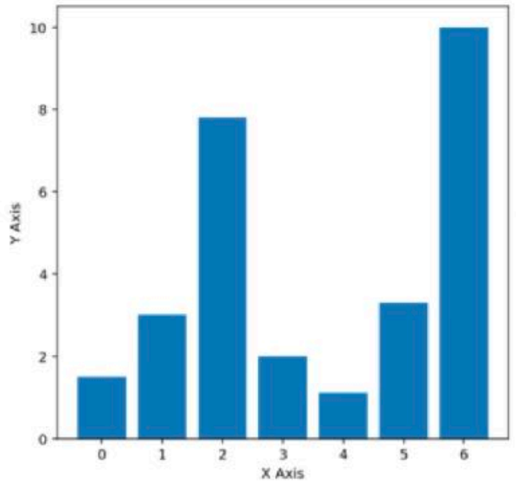
Get multiple perspectives

Choosing your plot

Pie

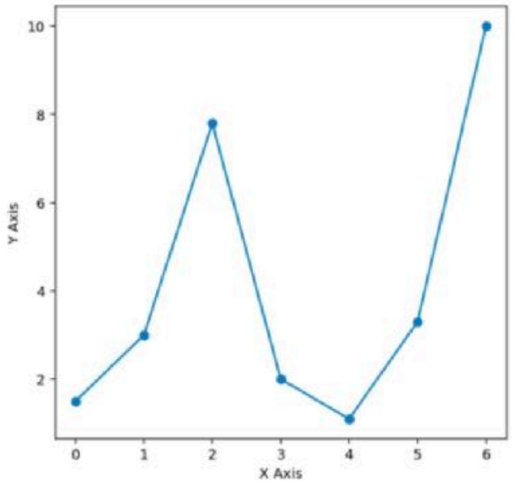
Almost never

Bar



Usually categorical

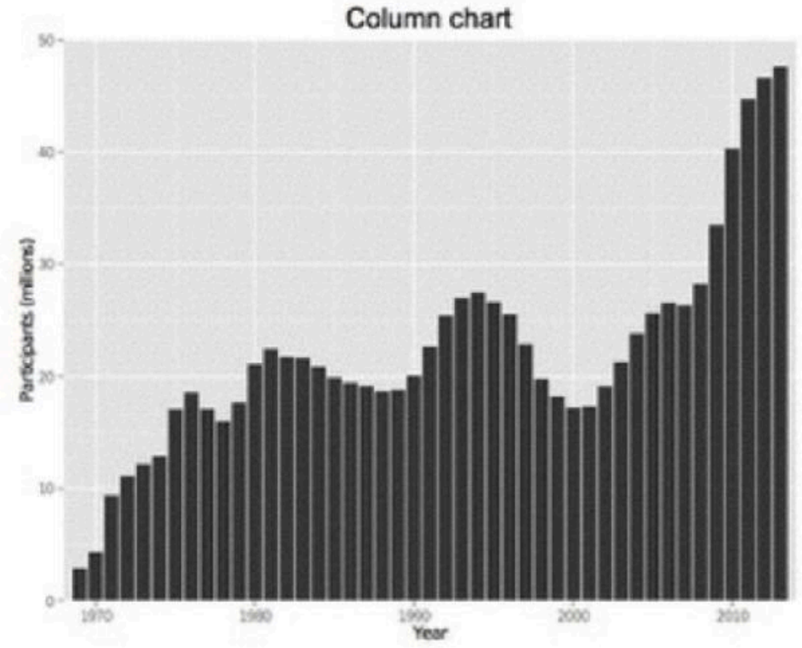
Line



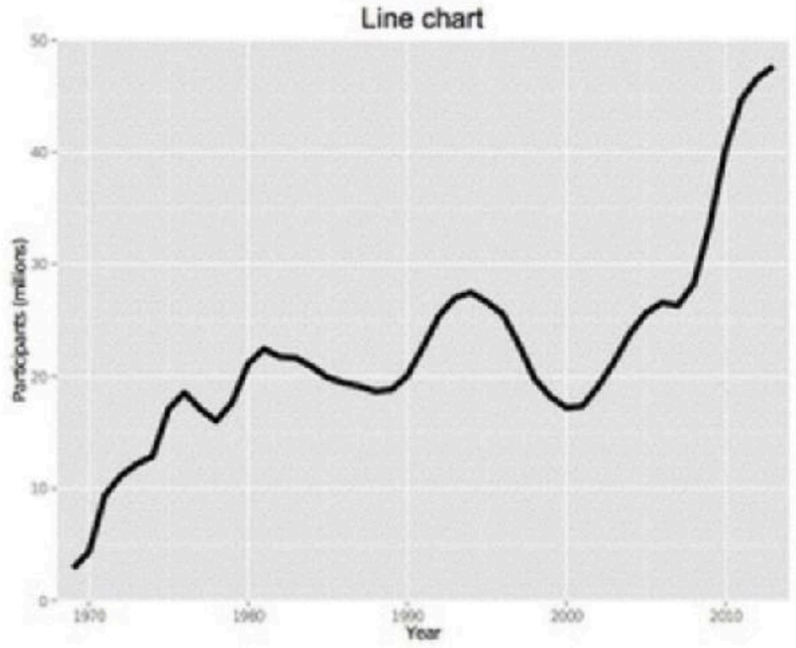
Change over time

“What am I trying to show?”

Choosing your plot

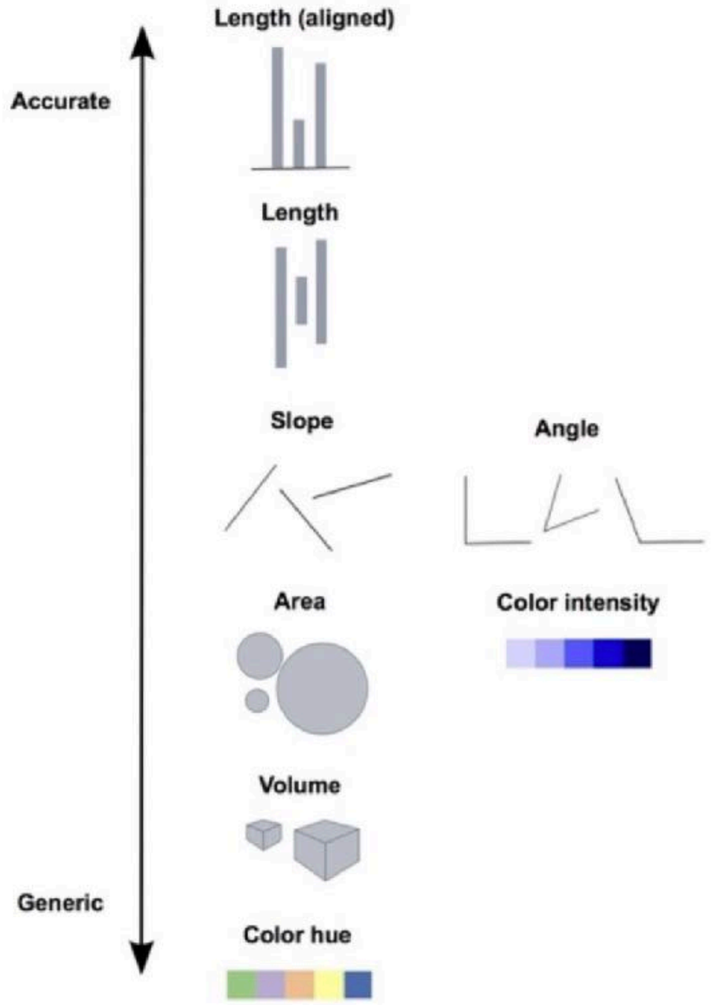


Time is discrete



Time is continuous

Visual principles

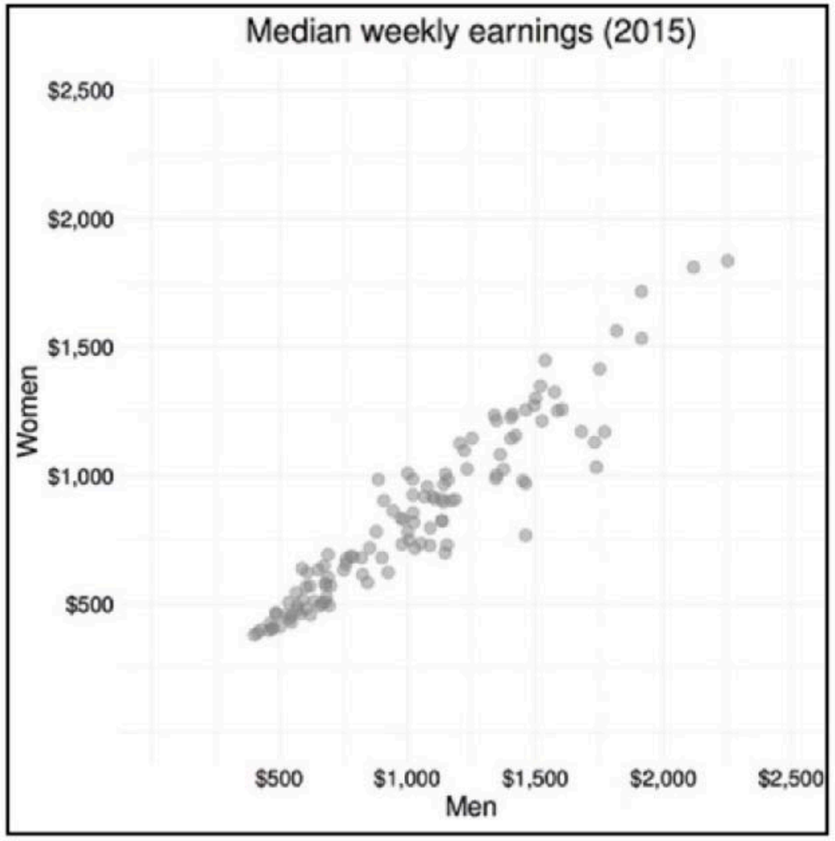


Pre-attentive Processing
Patterns and anomalies automatically grab our attention.

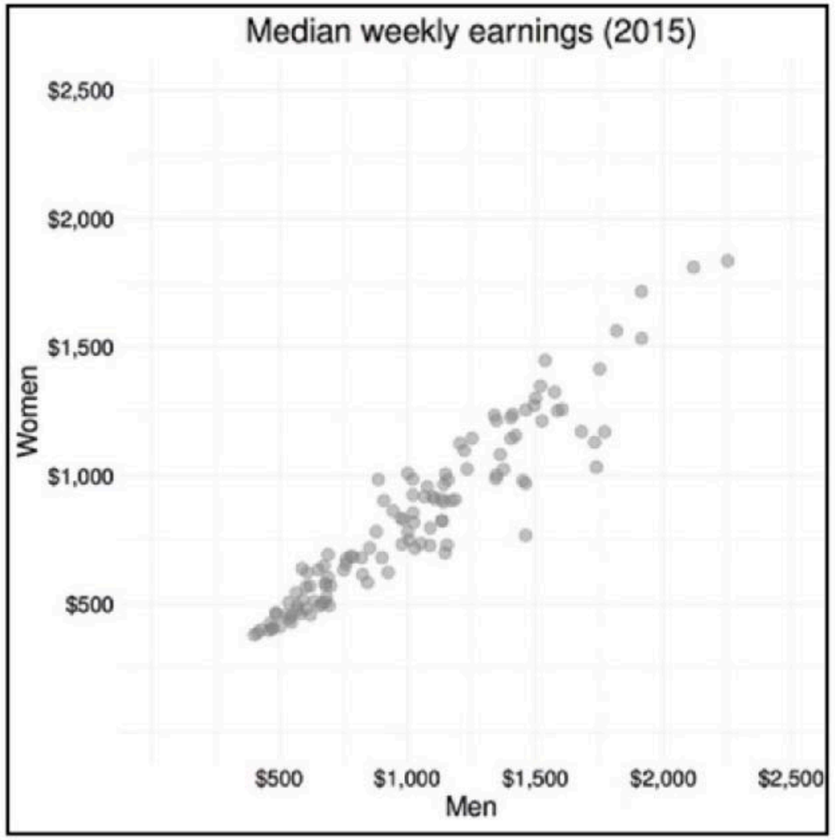
Quantitative Judgement
More accurate is **often**, but not **always**, better.

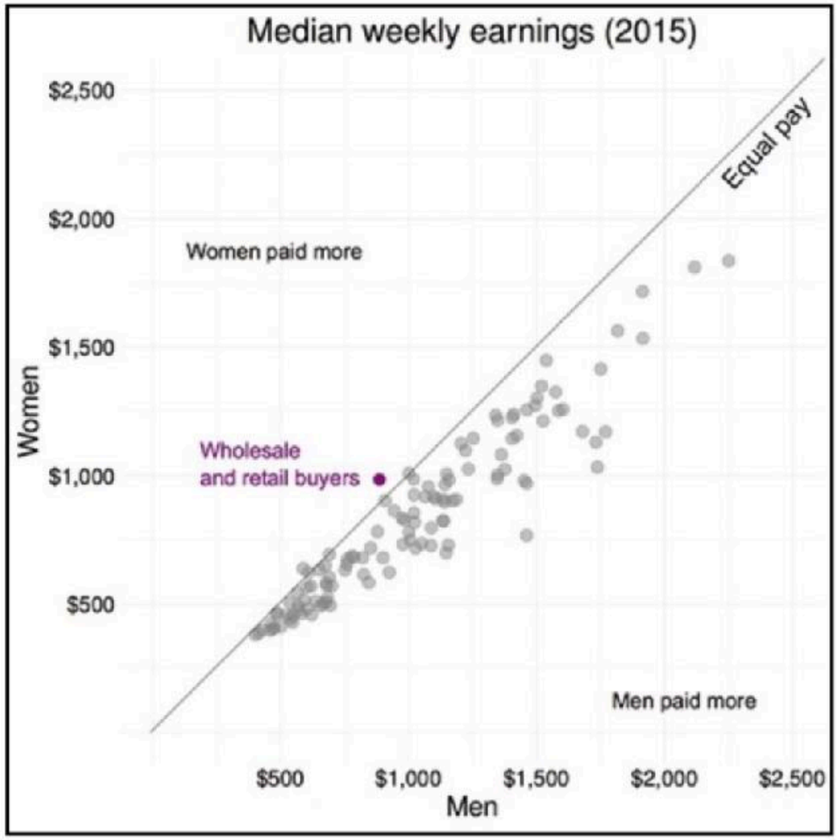
Source: http://paldhous.github.io/ucb/2016/dataviz/index.html

Visual principles



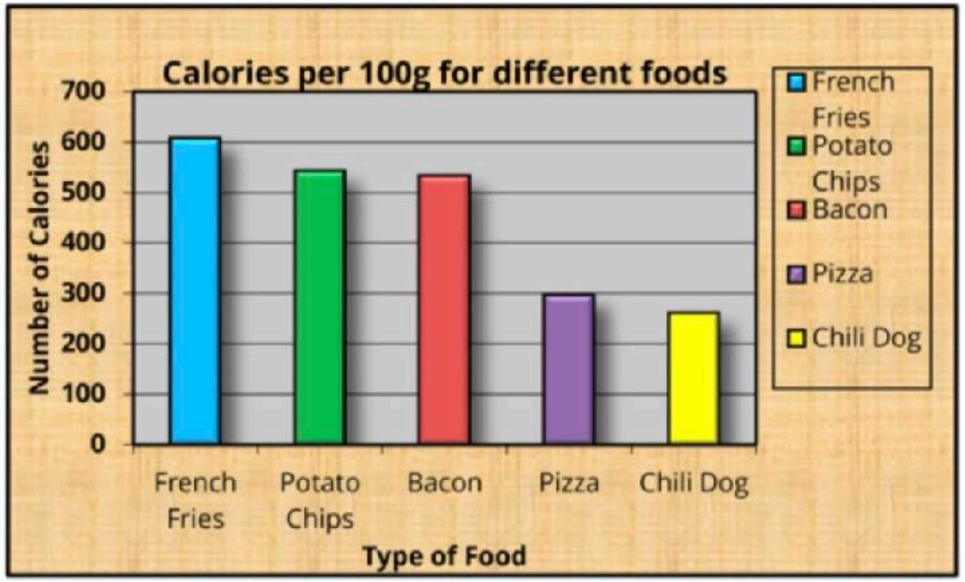
Visual principles

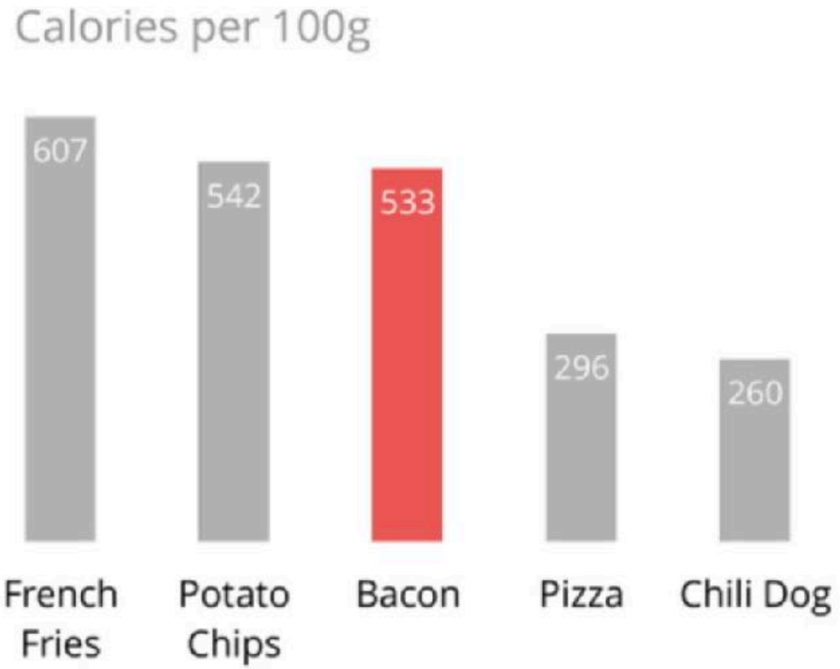




Emphasize your message.

Visual principles



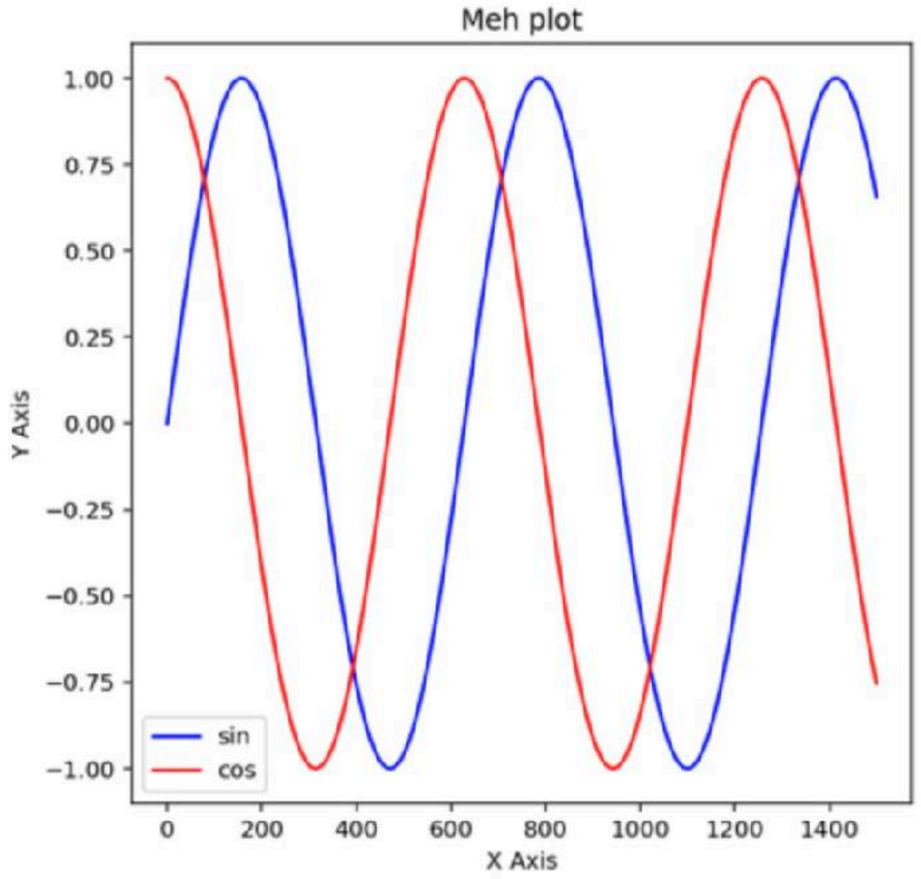


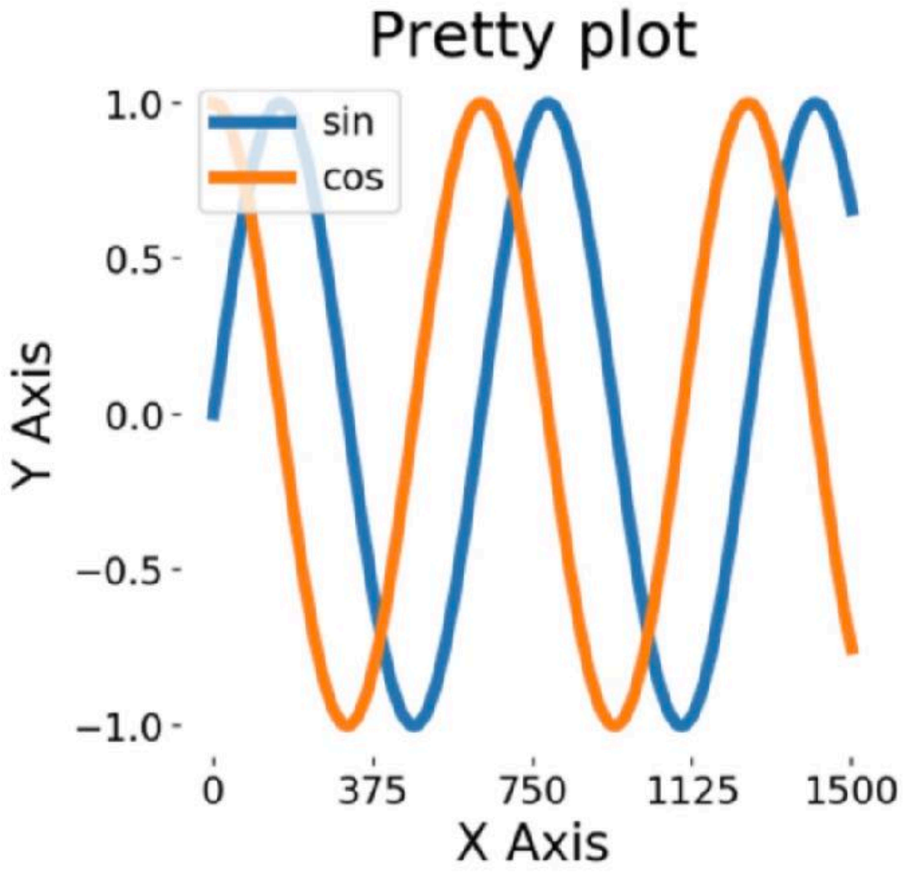
De-emphasize everything else.

Source: Darkhorse Analytics

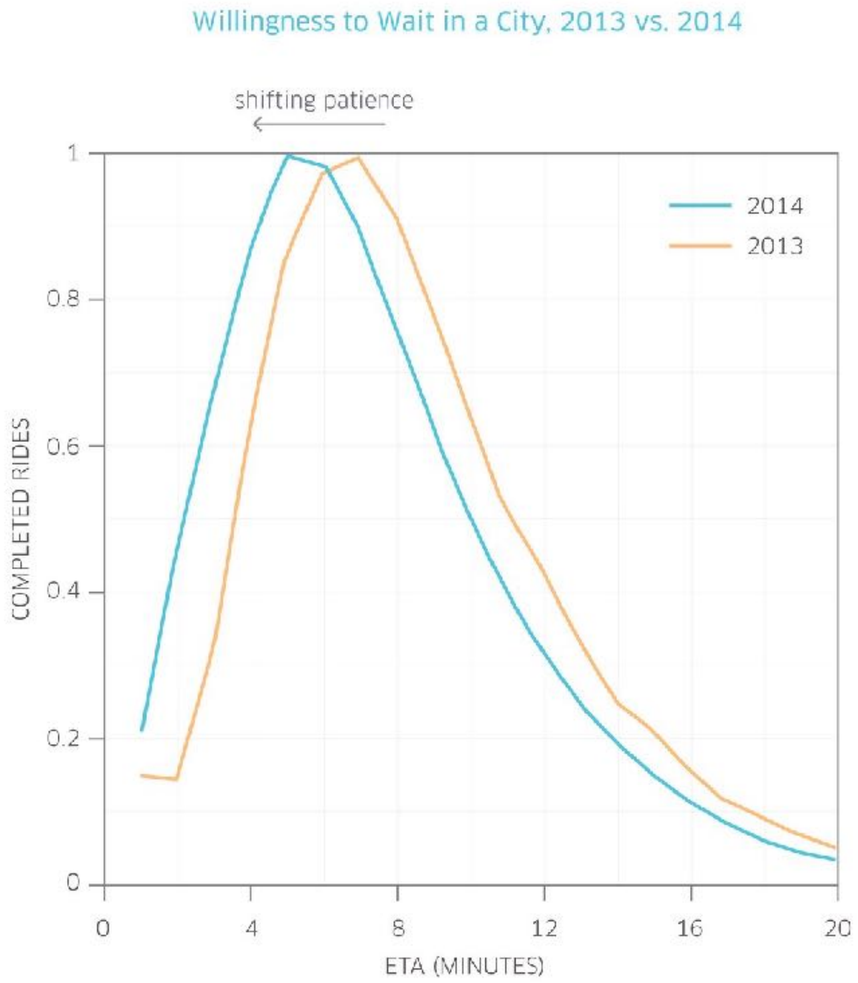
Visual principles

The Devil is in the Detail



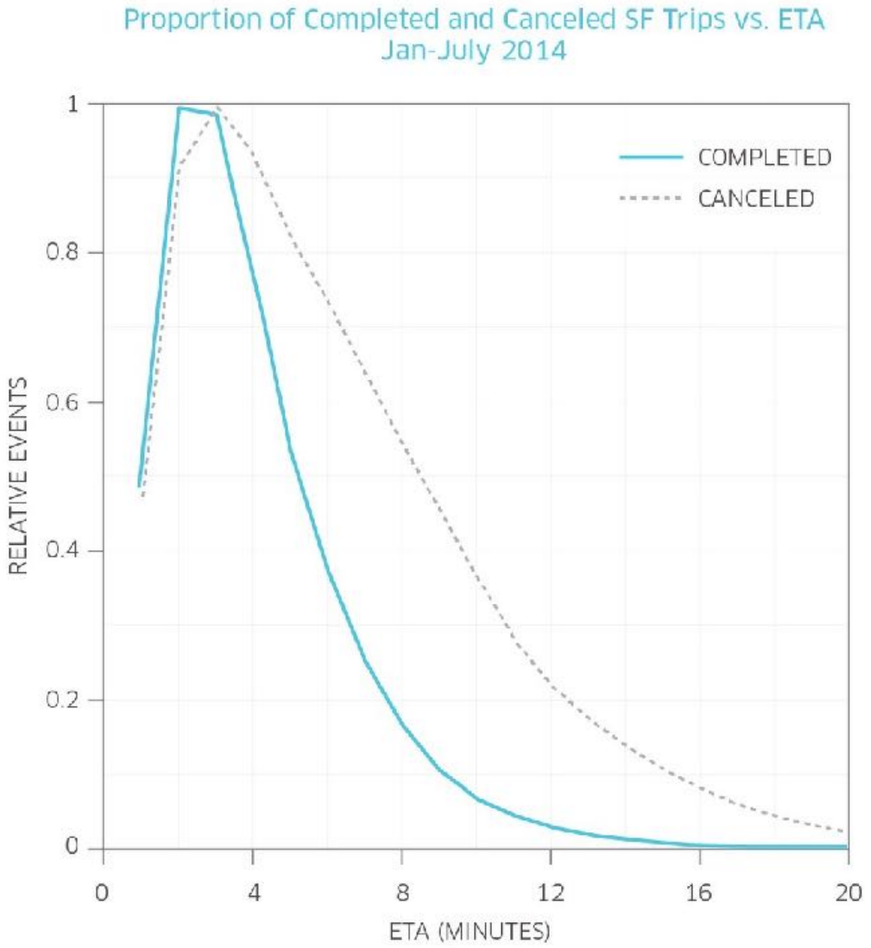


Personal example: Quantifying patience



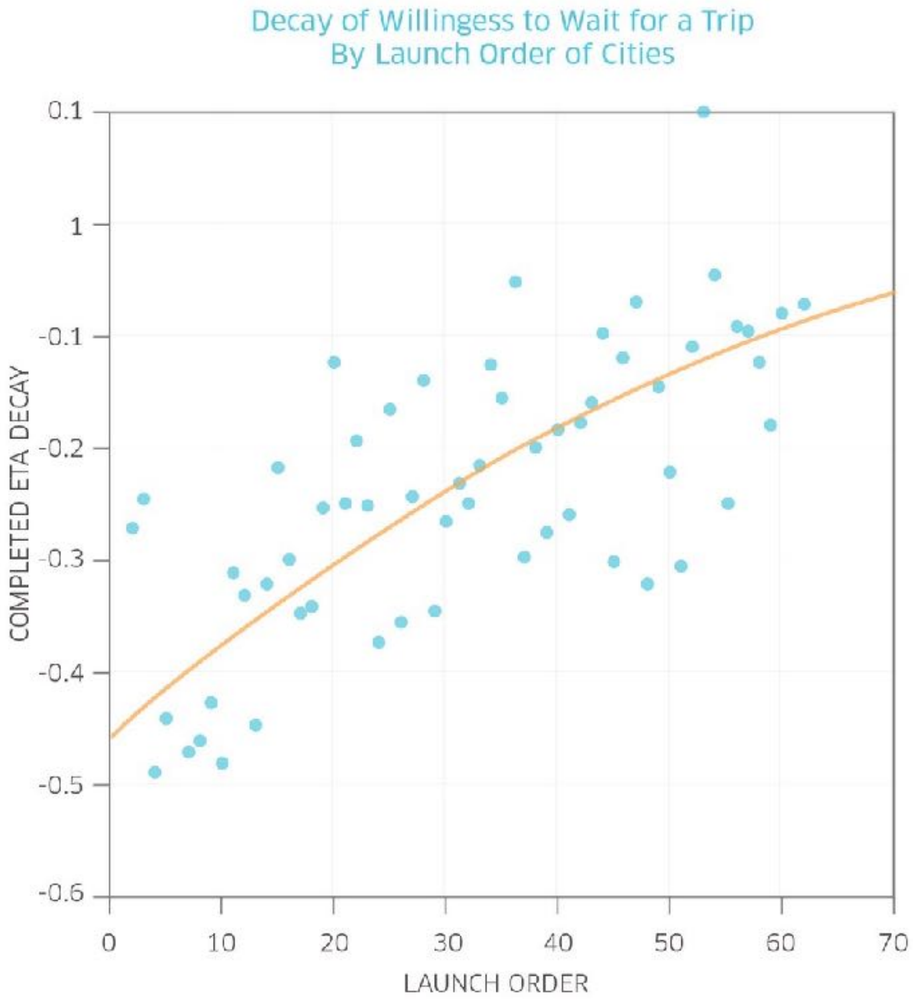
Source: Voytek, *Uber, Inc.*

Personal example: Quantifying patience



Source: Voytek, *Uber, Inc.*

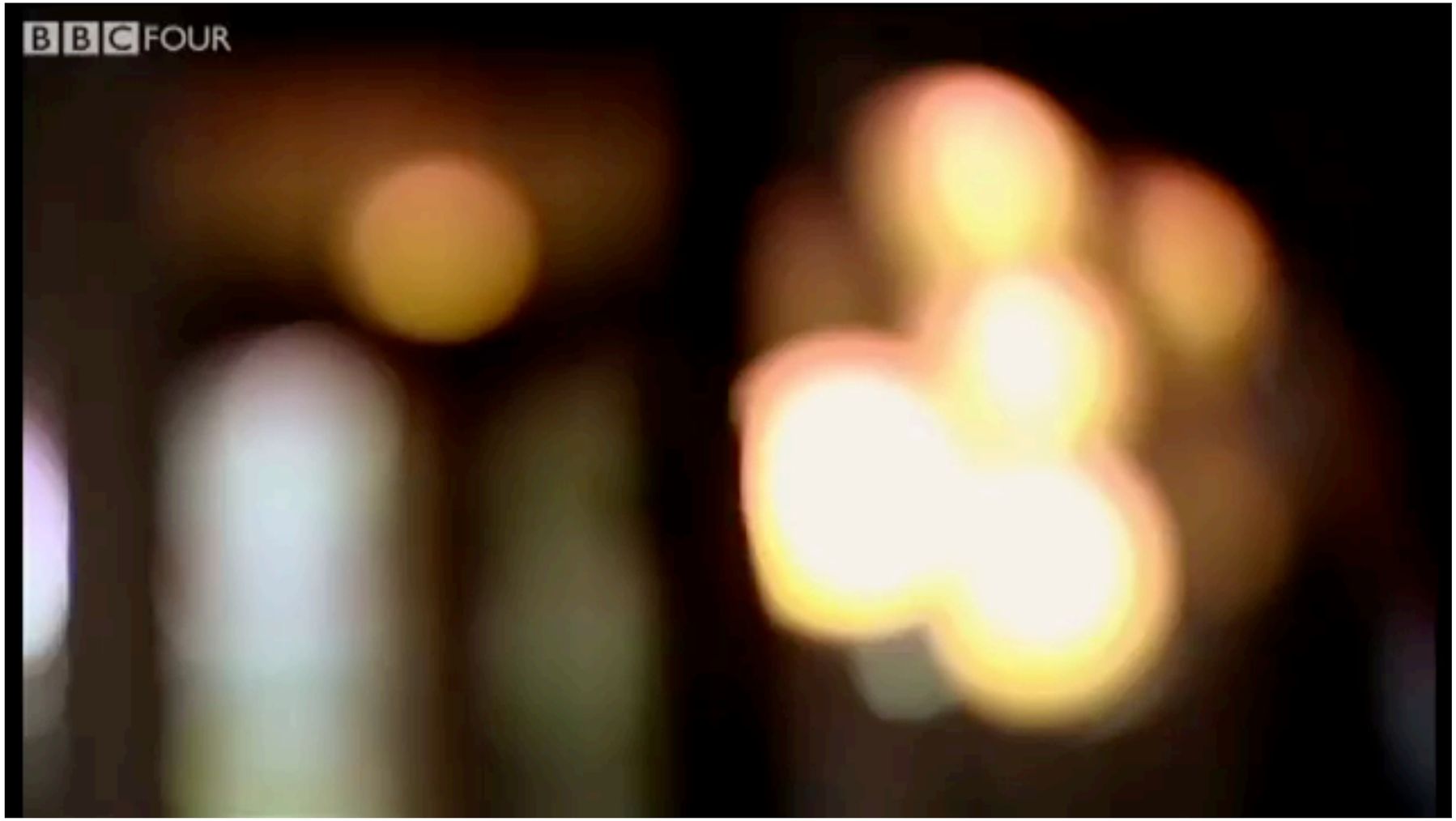
Personal example: Quantifying patience



Source: Voytek, *Uber, Inc.*

The slide deck of ideas!

Health and wealth



Source: http://www.gapminder.org/world/

Names

Looking at the absolute chance in percentages is interesting, but would not tell the full story. A change of, say 15% to 14% would be quite different and less drastic than a change from 2% to 1%, but the absolute change in percentage would measure those two things equally. Thus, I need a measure of the relative change in the percentages — that is, the percent change in percentages (confusing, I know). Fortunately the public health field has dealt with this problem for a long time, and has a measurement called the relative risk, where "risk" refers to the proportion of babies given a certain name. For example, let's say the percentage of babies named "Jane" is 1% of the population in 1990, and 1.2% of the population in 1991. The relative risk of being named "Jane" in 1991 versus 1990 is 1.2 (that is, it's (1.2/1)=1.2 times as probable, or (1.2-1)*100=20% more likely). In this case, however, I'm interested in instances where the percentage of children with a certain name decreases. The way to make the most sensible statistics in this case is to calculate the relative risk again, but in this case think of it as a decrease. That is, if "Jane" was at 1.5% in 1990 and 1.3% in 1991, then the relative risk of being named "Jane" in 1991 compared to 1990 is (1.3/1.5)=0.87. That is, it is (1-0.87)*100=13% less likely that a baby will be named "Jane" in 1991 compared to 1990.

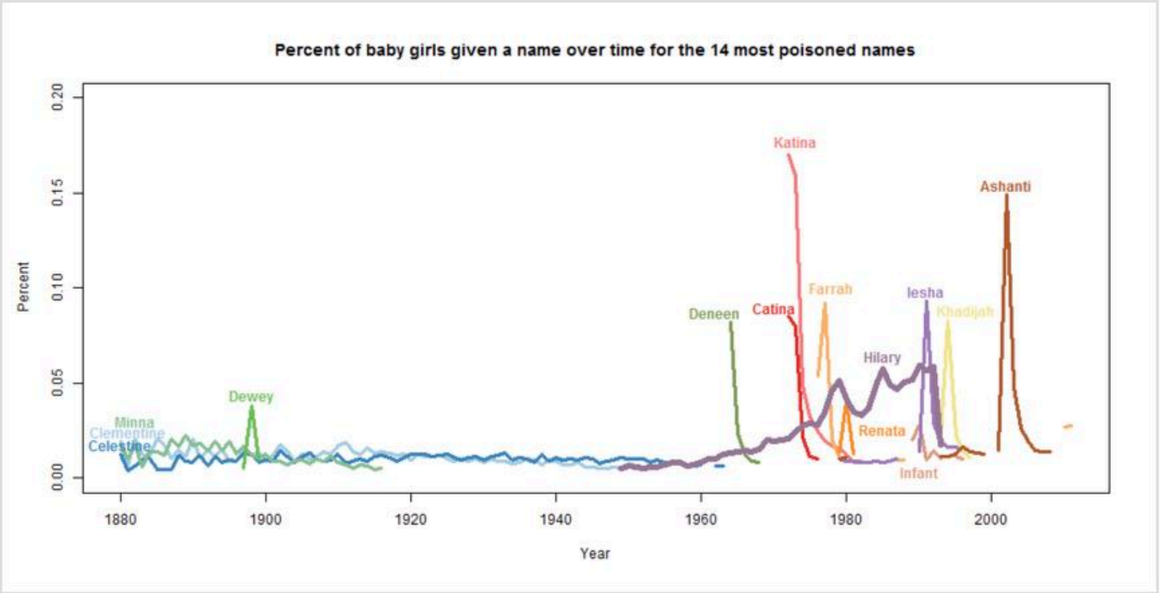
(Note that I'm not doing any model fitting here because I'm not interested in any parameter estimates — I have my entire population! I'm just summarizing the data in a way that makes sense.)

Source: Hilary Parker, https://hilaryparker.com/2013/01/30/hilary-the-most-poisoned-baby-name-in-us-history/

Names

Looking at the absolute chance in percentages is interesting, but would not tell the full story. A change of, say 15% to 14% would be quite different and less drastic than a change from 2% to 1%, but the absolute change in percentage would measure those two things equally. Thus, I need a measure of the relative change in the percentages — that is, the percent change in percentages (confusing, I know). Fortunately the public health field has dealt with this problem for a long time, and has a measurement called the relative risk, where "risk" refers to the proportion of babies given a certain name. For example, let's say the percentage of babies named "Jane" is 1% of the population in 1990, and 1.2% of the population in 1991. The relative risk of being named "Jane" in 1991 versus 1990 is 1.2 (that is, it's (1.2/1)=1.2 times as probable, or (1.2-1)*100=20% more likely). In this case, however, I'm interested in instances where the percentage of children with a certain name decreases. The way to make the most sensible statistics in this case is to calculate the relative risk again, but in this case think of it as a decrease. That is, if "Jane" was at 1.5% in 1990 and 1.3% in 1991, then the relative risk of being named "Jane" in 1991 compared to 1990 is (1.3/1.5)=0.87. That is, it is (1-0.87)*100=13% less likely that a baby will be named "Jane" in 1991 compared to 1990.

(Note that I'm not doing any model fitting here because I'm not interested in any parameter estimates — I have my entire population! I'm just summarizing the data in a way that makes sense.)



These plots looked quite curious to me. While the names had very steep drop-offs, they also had very steep drop-ins as well.

This is where this project got deliriously fun. For each of the names that "dropped in" I did a little research on the name and the year. "Dewey" popped up in 1898 because of the Spanish-American War — people named their daughters after George Dewey. "Deneen" was one name of a duo with a one-hit wonder in 1968. "Katina" and "Catina" were wildly popular because in 1972 in the soap opera *Where the Heart Is* a character is born named

Source: Hilary Parker, https://hilaryparker.com/2013/01/30/hilary-the-most-poisoned-baby-name-in-us-history/

Names

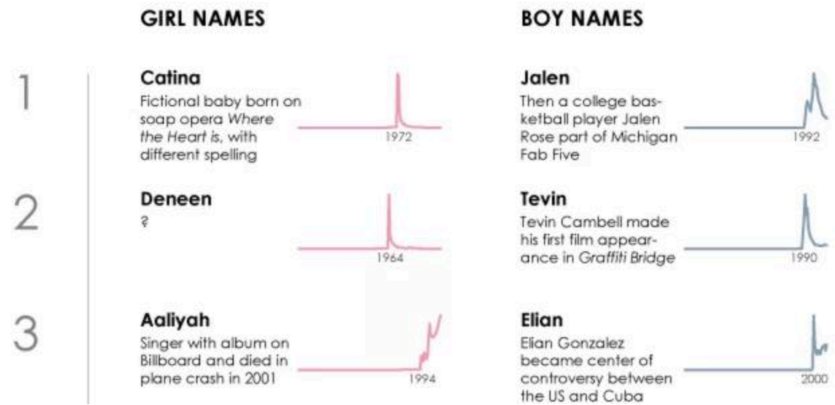
The Most Trendy Names in US History

POSTED TO DATA UNDERLOAD | TAGS: NAMES | NATHAN YAU

Names are incredibly personal things. It's how we identity ourselves. We associate others, places, and points in our past with names. Maybe you recall a family member, a celebrity, or a significant other.

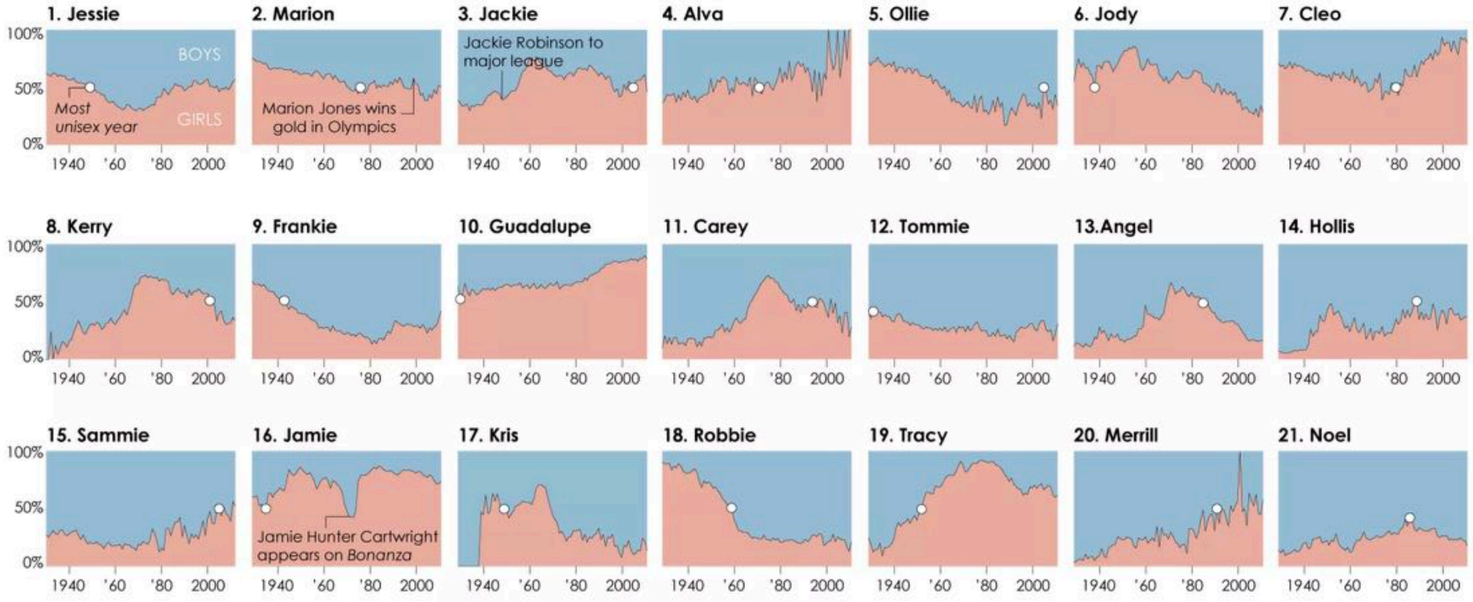
At the same time, it's not uncommon for two people with the same name to run into each other, and it's why gift shops can sell and profit from those mini license plates. Parents decide what they want to call their kid at some point. So as you walk through history, you end up with names that surge, some that die off, and some that come back again.

Hilary Parker already looked at the most poisoned name in US history (her own). Here we look at names from the other direction. The most trendy:



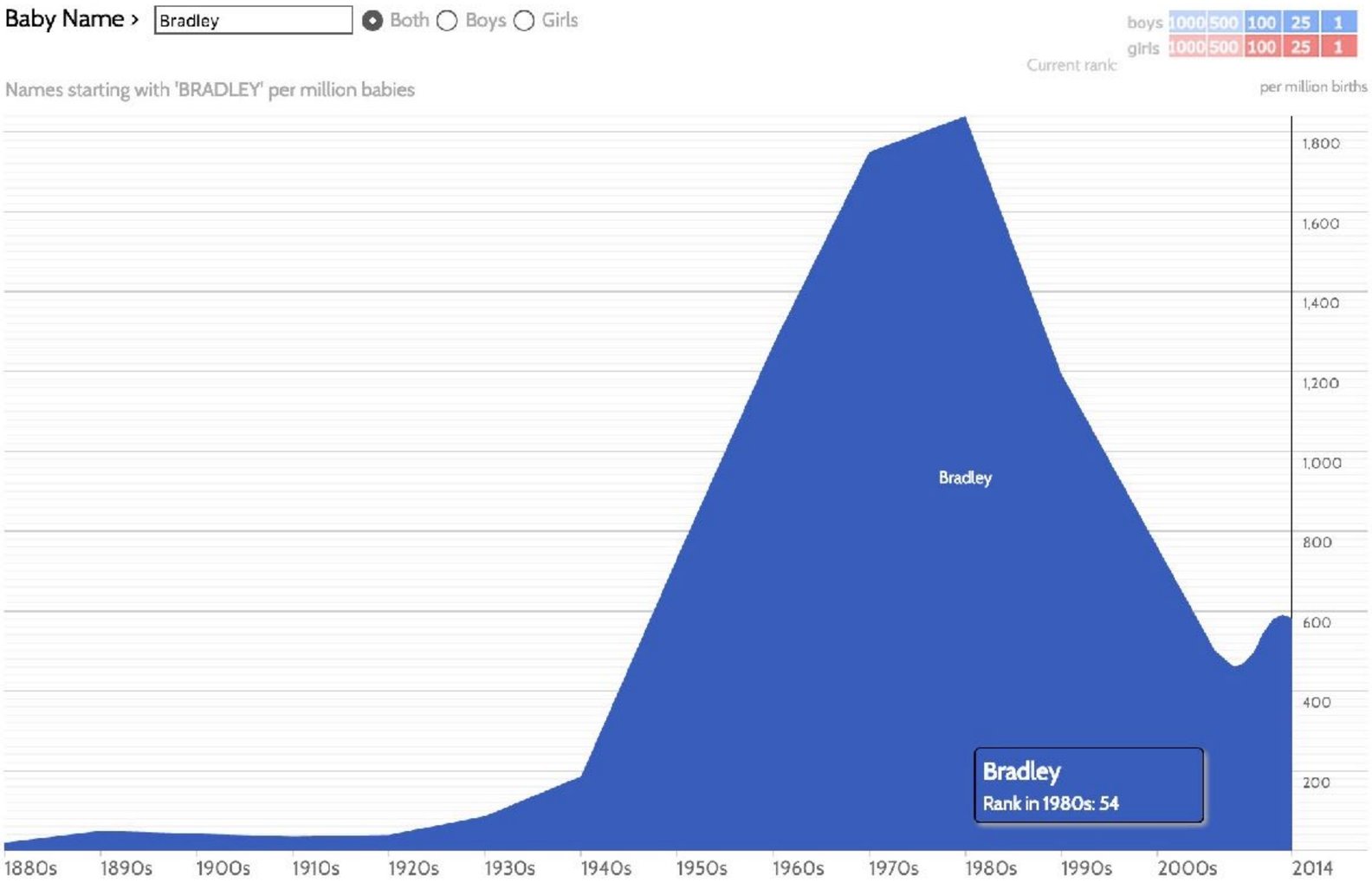
The Most Unisex Names in US History

BY NATHAN YAU / POSTED TO DATA UNDERLOAD / TAGS: NAMES



Source: Nathan Yau, http://flowingdata.com/2013/07/29/the-most-trendy-names-in-us-history/; /2013/09/25/the-most-unisex-names-in-us-history/

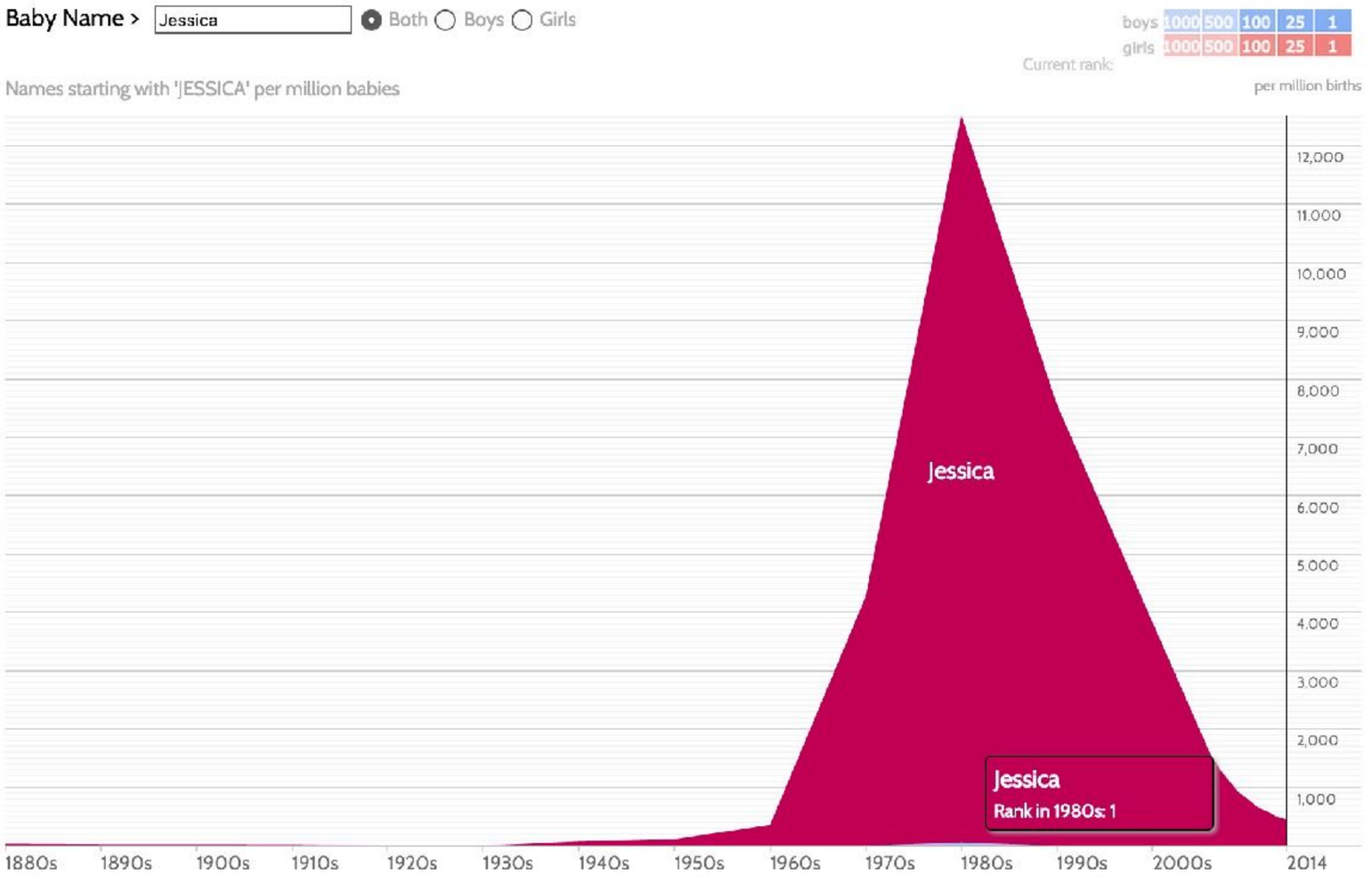
Names over time



Source: babynamewizard.com

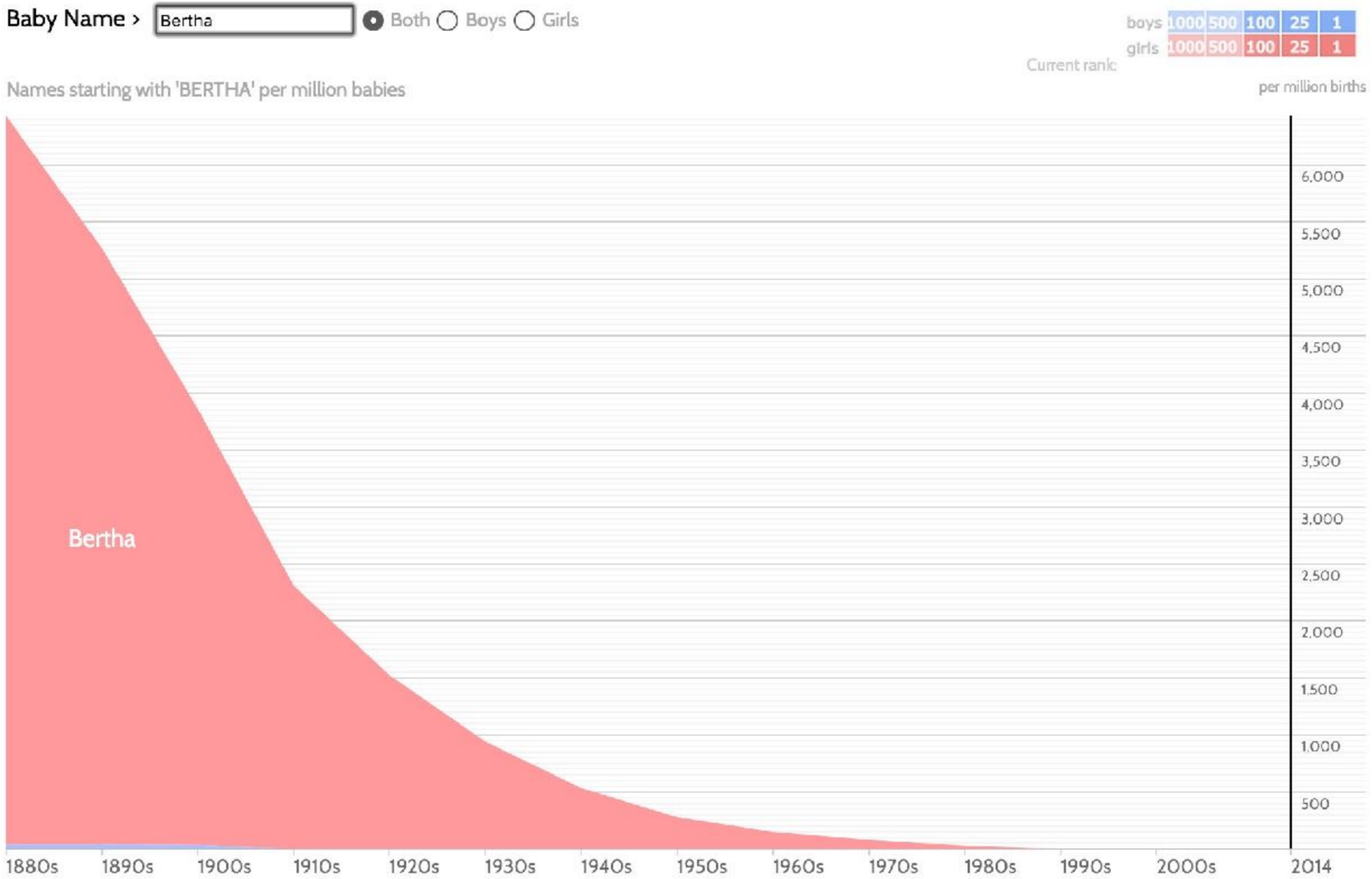
UC San Diego

Names over time



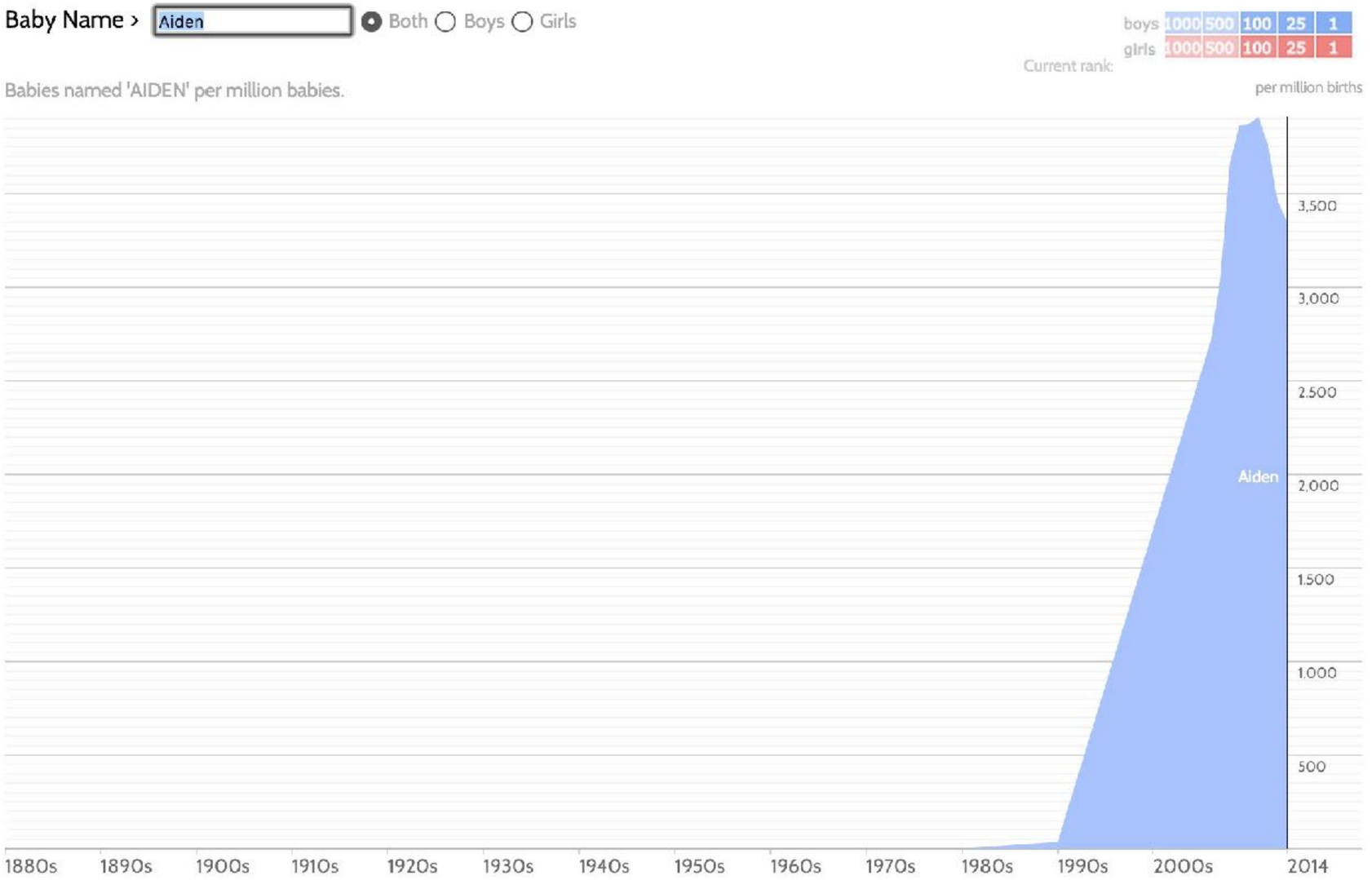
Source: babynamewizard.com

Names over time



Source: babynamewizard.com

Names over time



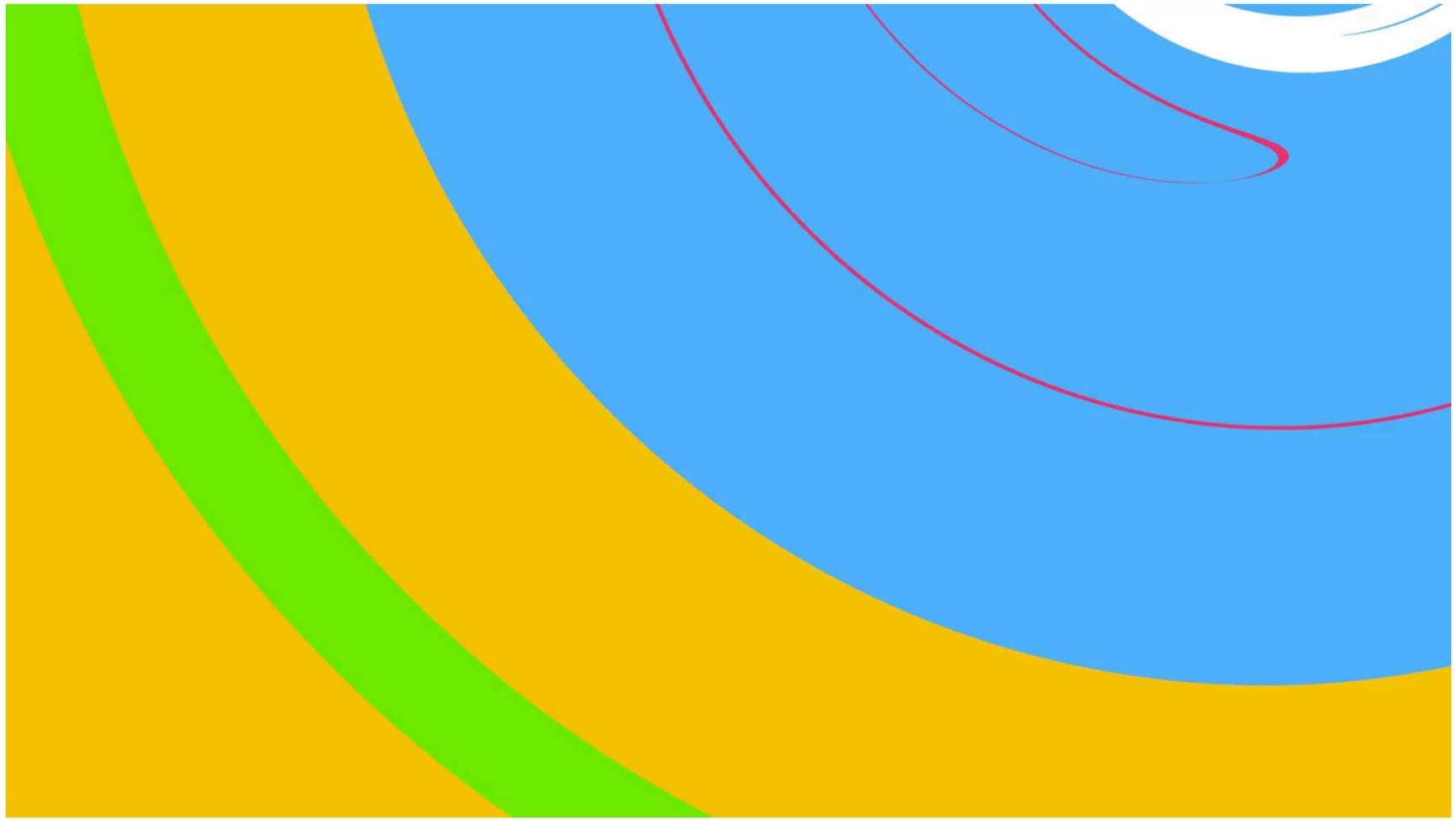
Source: babynamewizard.com

UC San Diego

Names over time

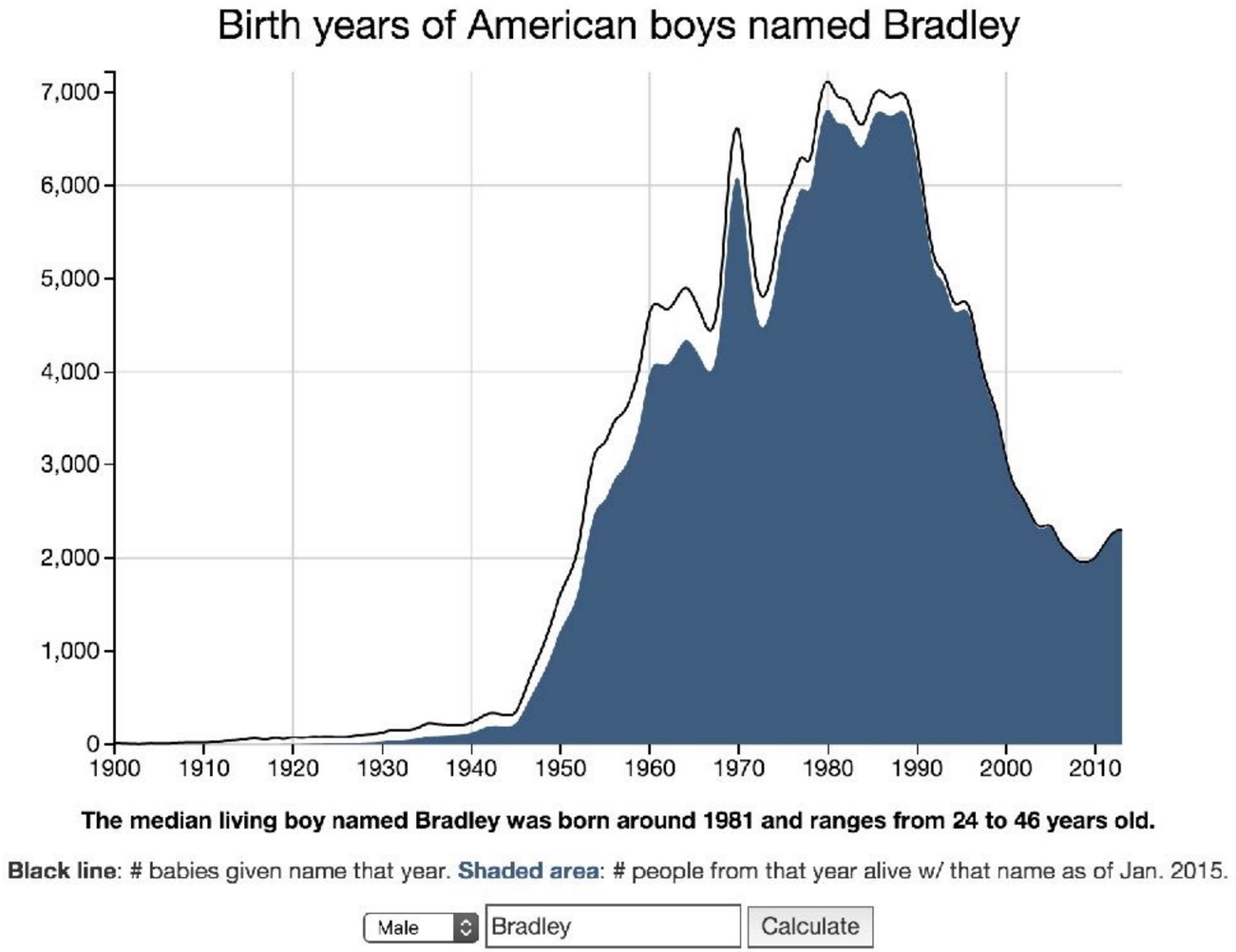
Source: *Abacaba*

Names over time



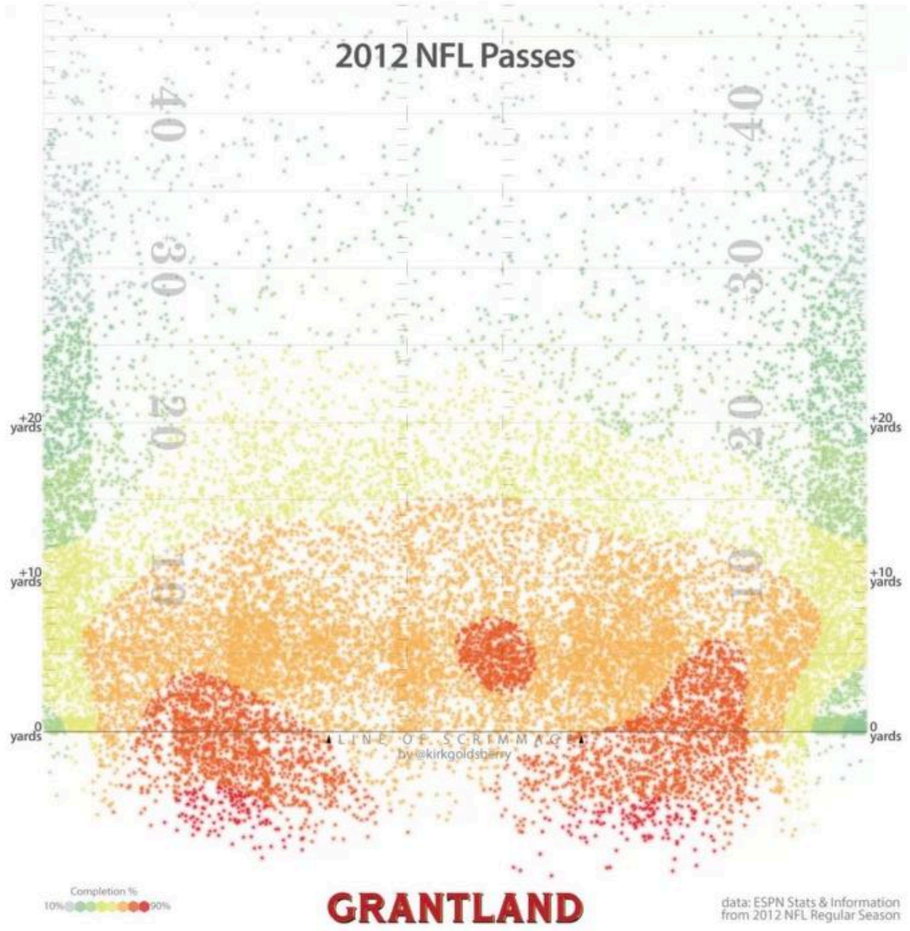
Source: *Abacaba*

Names over time



Source: http://rhiever.github.io/

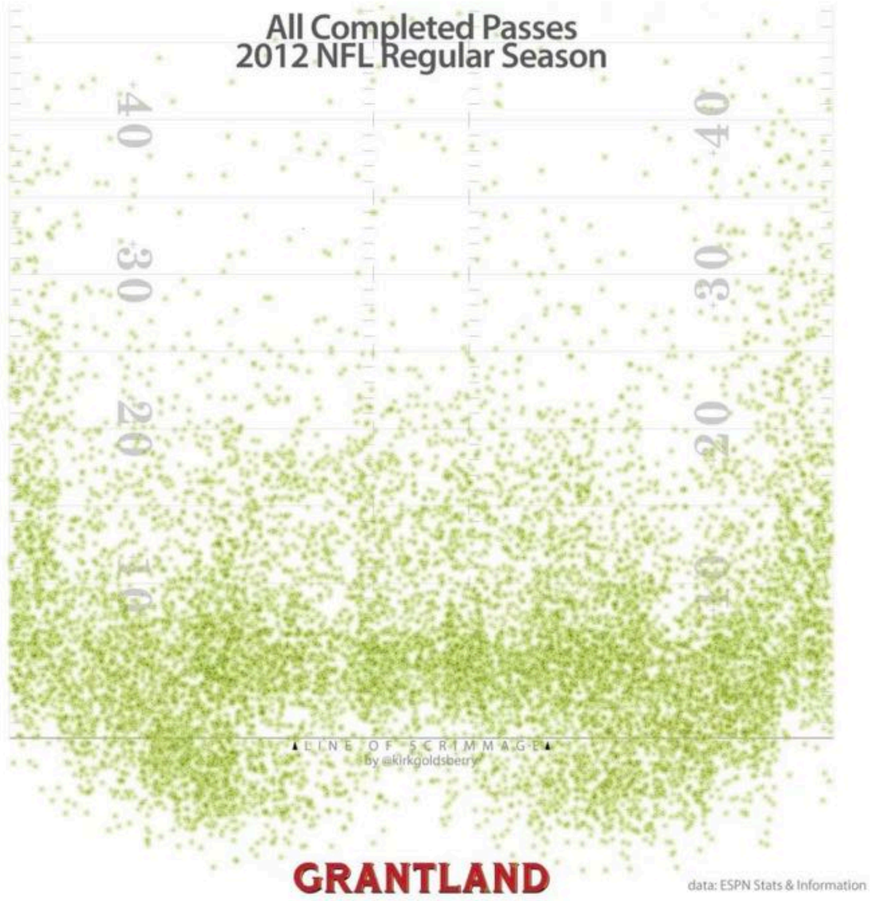
NFL passes



During the 2012 regular season, NFL quarterbacks attempted more than 19,000 passes and completed more than 11,500; the chart above shows the location of nearly every pass. Dots are placed in spots nearby the receiver or the intended receiver. In the cases of out-of-bounds throwaways, those dots are placed at the sideline near where the ball went out of bounds. The color of the dot corresponds to the league's overall completion percentage in that zone.

Source: http://grantland.com/the-triangle/pass-atlas-a-map-of-where-nfl-quarterbacks-throw-the-ball/

NFL passes

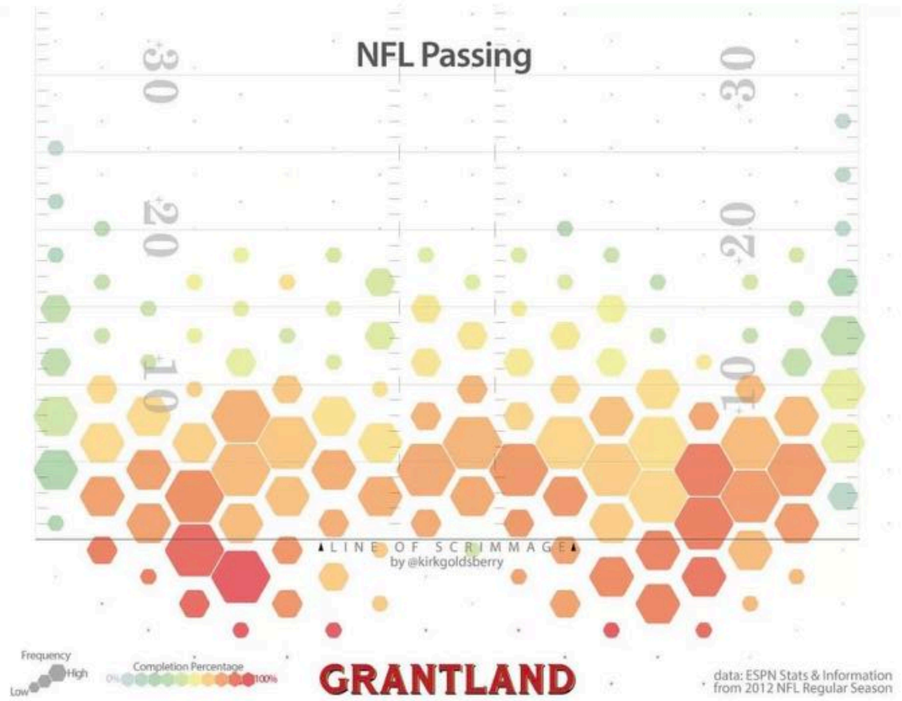


If we look only at completed passes, we see similar patterns. Long "bomb" completions are rare and a vast majority of NFL completions occur within 10 yards of the line of scrimmage. In fact, only about 31 percent of all passes are thrown beyond 10 yards of the line of scrimmage. We hear a lot that the NFL is a passing league, but it's more accurate to say that it's a short-passing league.

The chart also immediately reveals the importance of the sideline, particularly downfield. Of passes thrown more than 20 yards, 69 percent are directed between the numbers and the sideline, while only 9 percent target the area between the hashes.

Source: http://grantland.com/the-triangle/pass-atlas-a-map-of-where-nfl-quarterbacks-throw-the-ball/

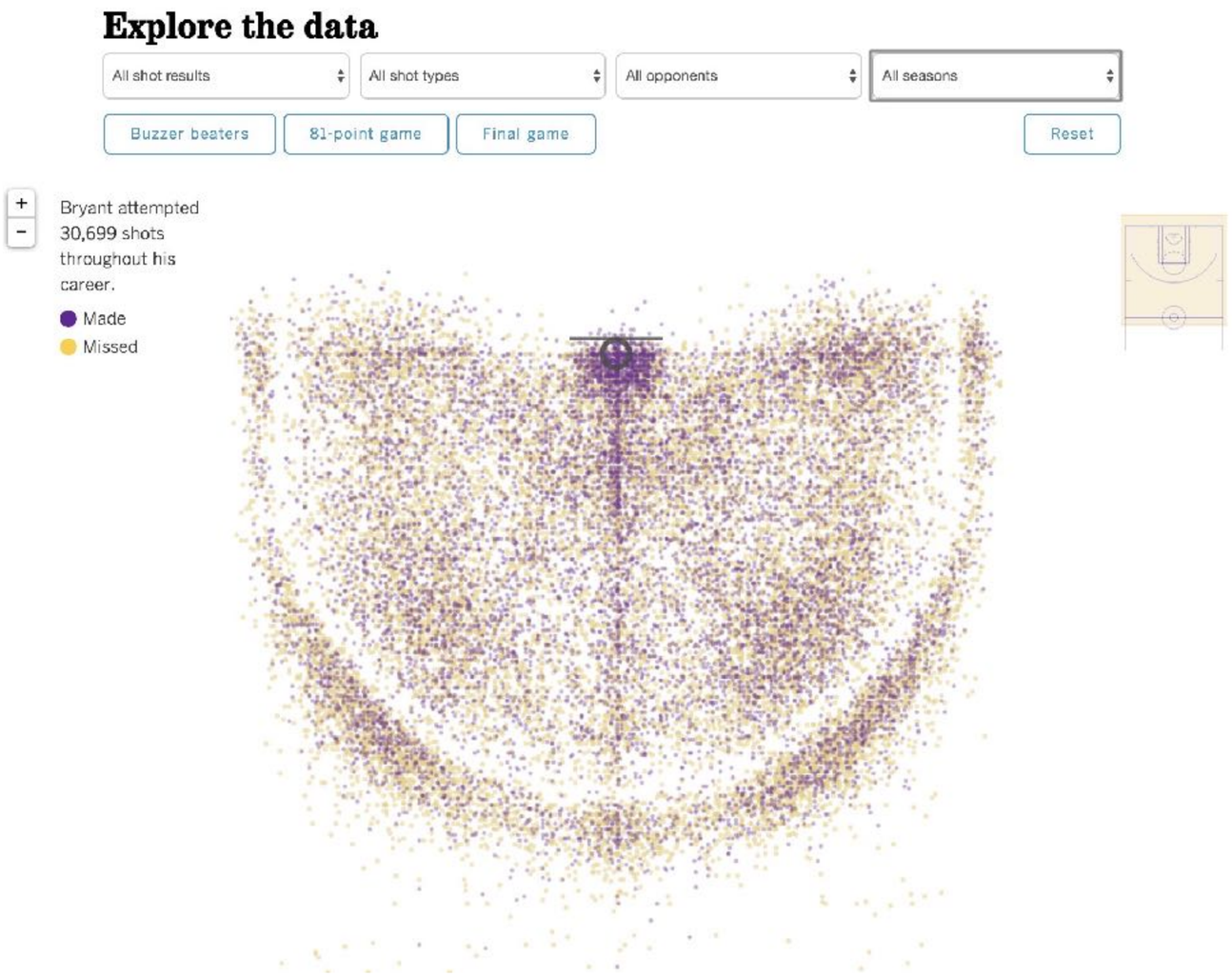
NFL passes



Although it's fun to look at thousands of dots, it's not really that informative. The chart below aggregates the 19,000-plus pass attempts into hexagonal zones, reinforcing the idea that short passes are king and revealing the league's most common passing targets. Although at first glance the pattern may appear mostly symmetric, NFL quarterbacks target receivers on the right side (46 percent) of the field more than the left side (41 percent).

Source: http://grantland.com/the-triangle/pass-atlas-a-map-of-where-nfl-quarterbacks-throw-the-ball/

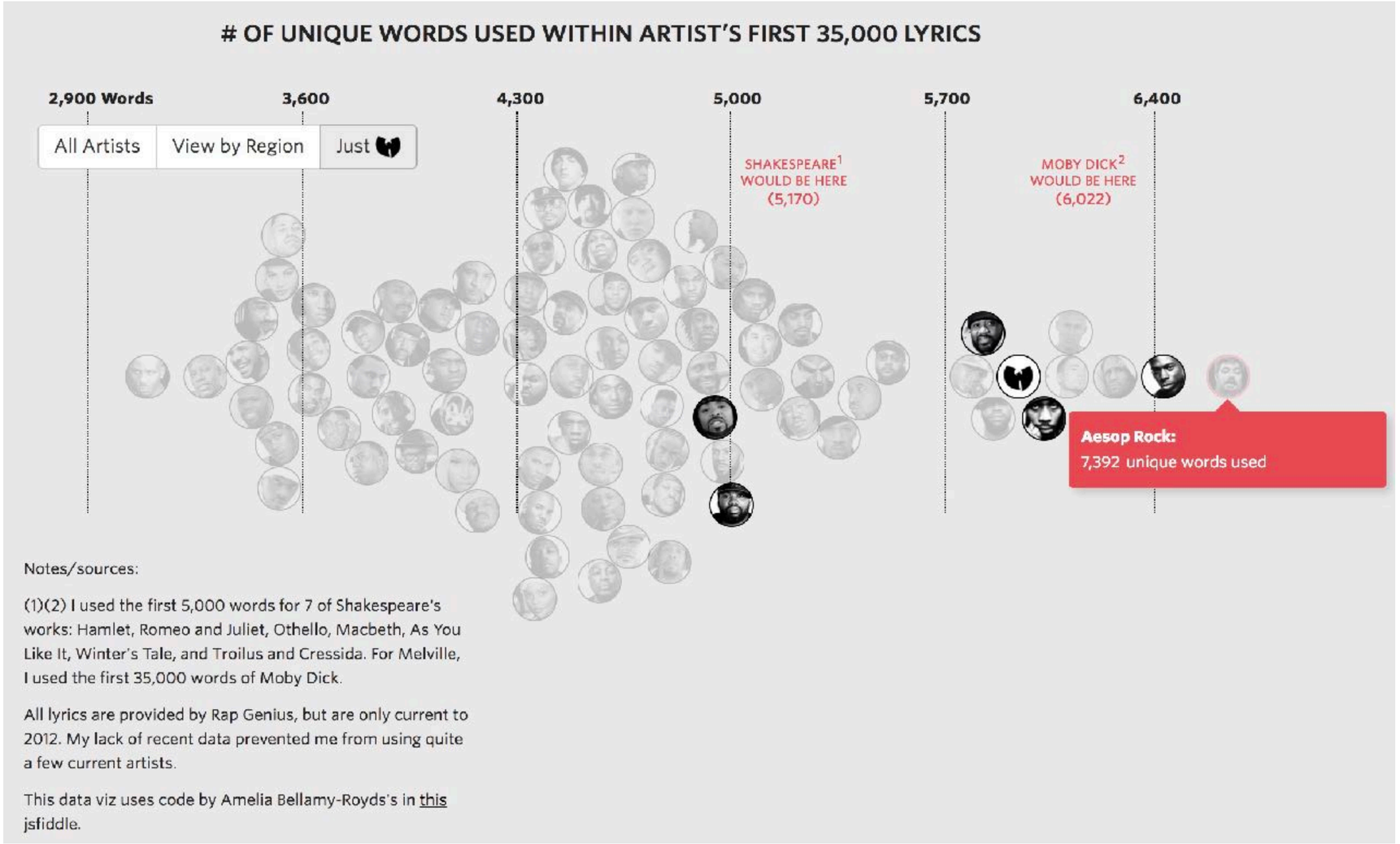
Kobe



source: http://graphics.latimes.com/kobe-every-shot-ever/

UC San Diego

Hip hop vocabulary



Source: Matt Daniels, https://pudding.cool/2017/02/vocabulary/index.html

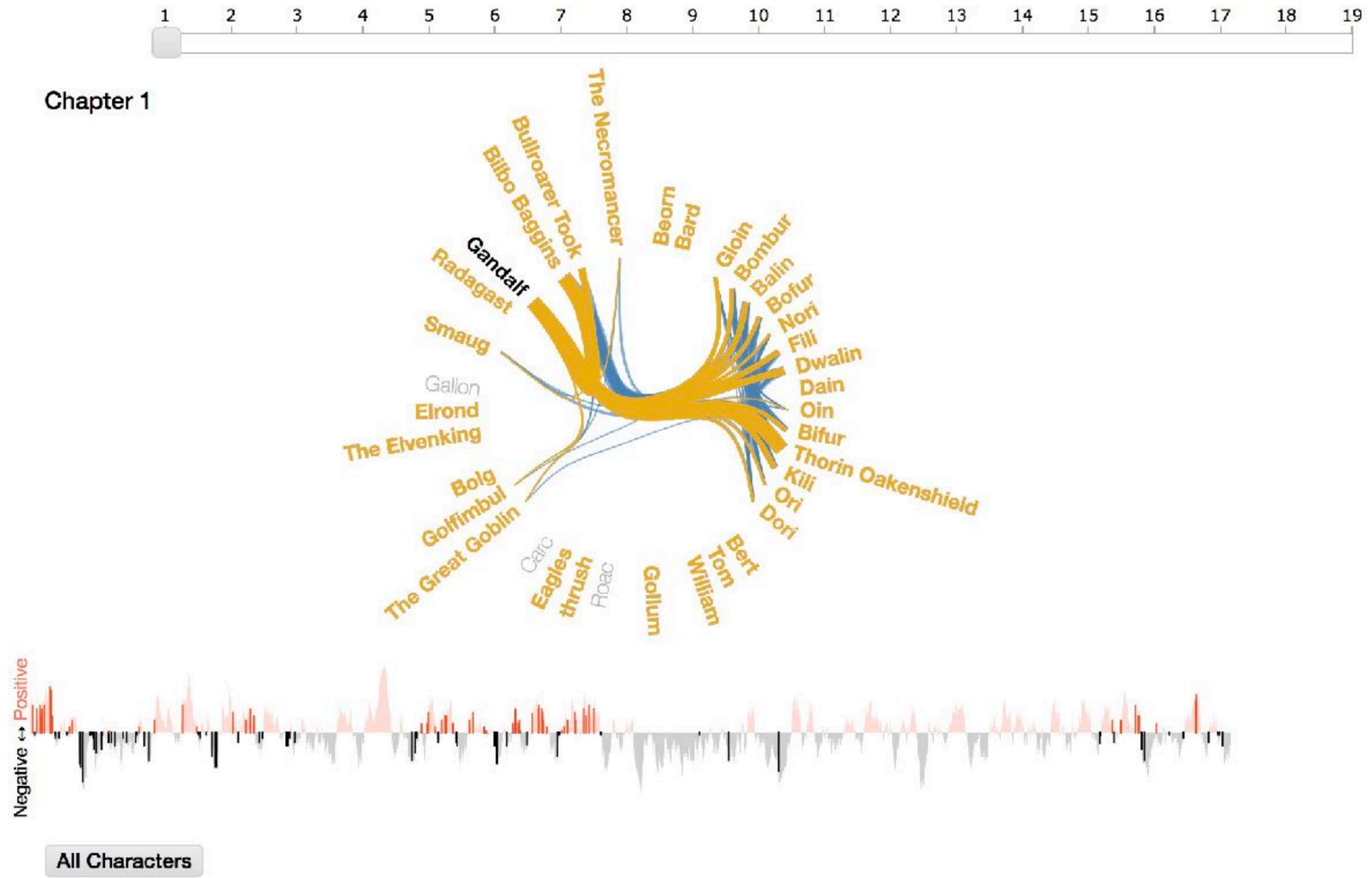
Hip hop vocabulary

Literary elites love to rep Shakespeare's vocabulary: across his entire corpus, he uses 28,829 words, suggesting he knew over 100,000 words and arguably had the largest vocabulary, ever.

I decided to compare this data point against the most famous artists in hip hop. I used each artist's first 35,000 lyrics. That way, prolific artists, such as Jay-Z, could be compared to newer artists, such as Drake.

Source: Matt Daniels, https://pudding.cool/2017/02/vocabulary/index.html

Narratives



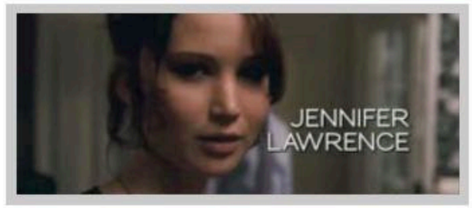
Source: Natalia Bilenko and Asako Miyakawa, http://nbilenko.com/projects/hobbit.html

UC San Diego

Film trailer narratives

Silver Linings Playbook

"Silver Linings Playbook" follows the standard model for trailers, according to Bill Woolery, a trailer specialist in Los Angeles who once worked on trailers for movies like "The Usual Suspects" and "E.T. the Extra-Terrestrial." While introducing the movie's story and its characters, the trailer largely follows the order of the film itself.

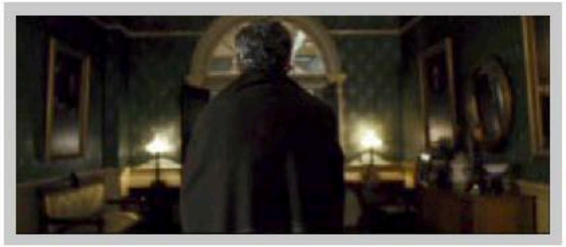


Source: Shan Carter, Amanda Cox, Mike Bostock, http://www.nytimes.com/interactive/2013/02/19/movies/awardsseason/oscar-trailers.html

Film trailer narratives

Lincoln

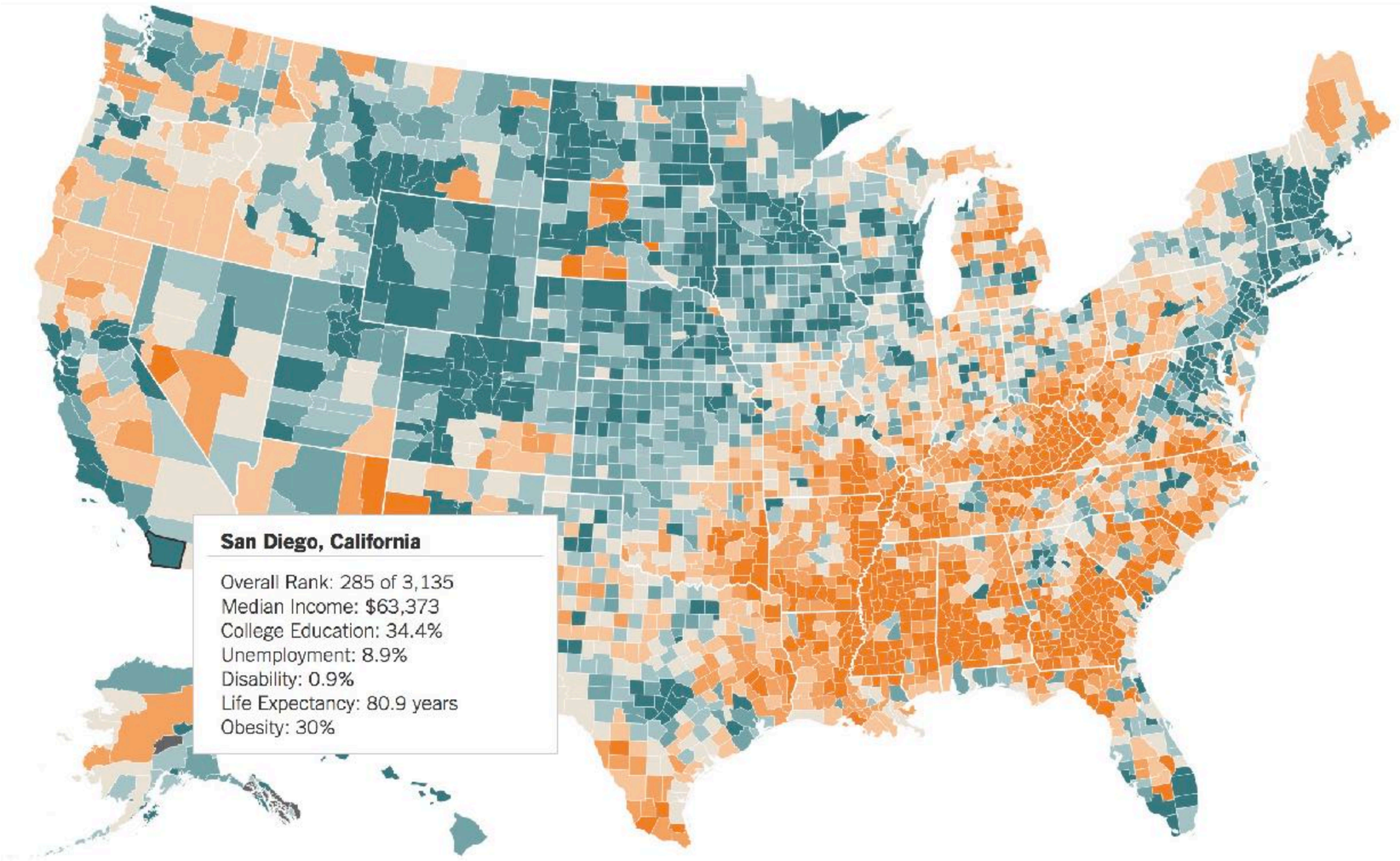
The "Lincoln" trailer is more like the typical teaser than a trailer, according to Stephen Garrett, who owns Jump Cut, a trailer house that specializes in foreign, independent and documentary films. While trailers often focus on plot or character descriptions, teasers establish the mood and tone of a film. Teasers "don't have to be chronological," Mr. Garrett said.



UC San Diego

Source: Shan Carter, Amanda Cox, Mike Bostock, http://www.nytimes.com/interactive/2013/02/19/movies/awardsseason/oscar-trailers.html

Geography



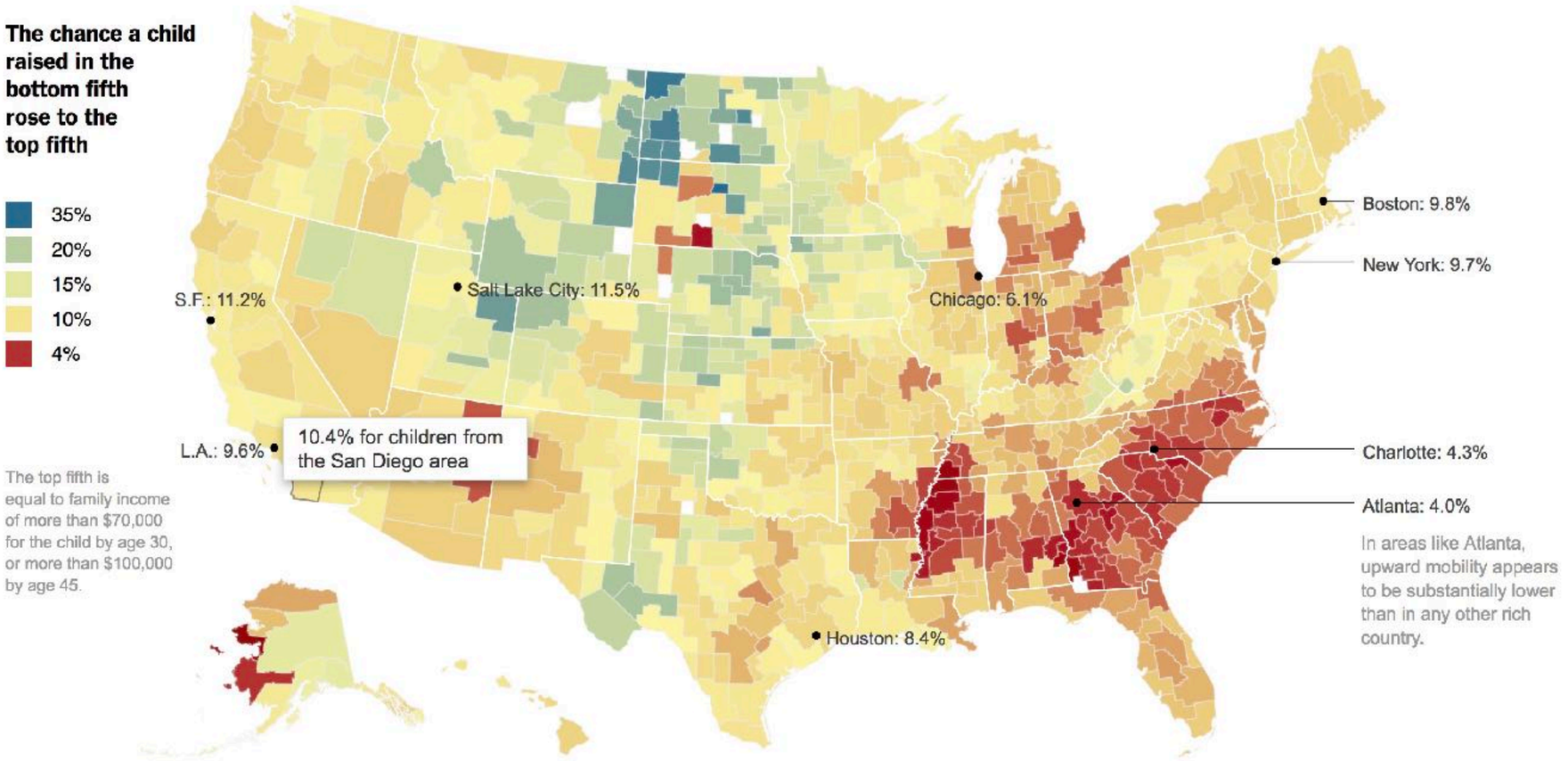
A composite ranking of where Americans are healthy and wealthy, or struggling.

UC San Diego

Source: Alan Flippen, https://www.nytimes.com/2014/06/26/upshot/where-are-the-hardest-places-to-live-in-the-us.html

Geography

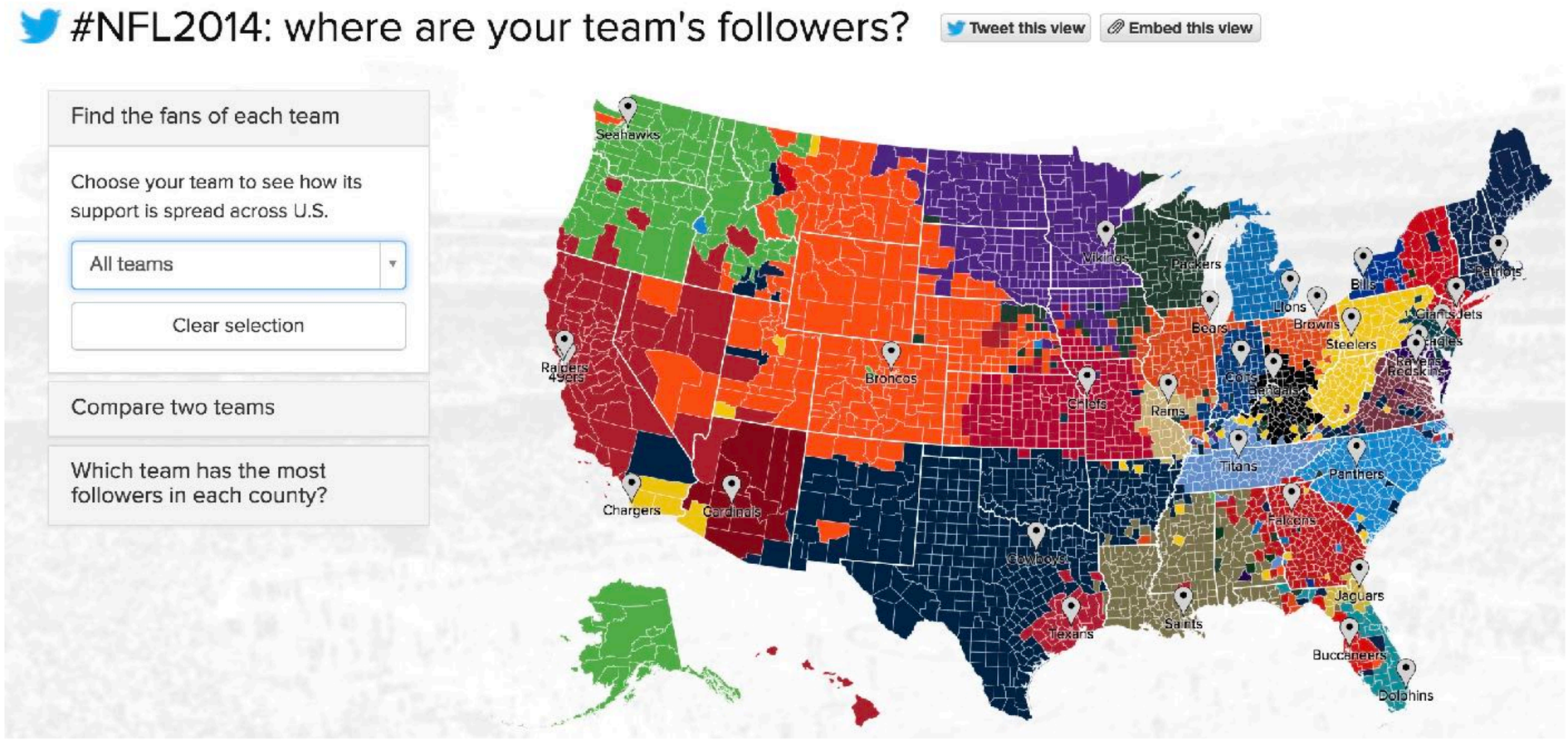
A study finds the odds of rising to another income level are notably low in certain cities, like Atlanta and Charlotte, and much higher in New York and Boston.



UC San Diego

Source: David Leonhardt, http://www.nytimes.com/2013/07/22/business/in-climbing-income-ladder-location-matters.html

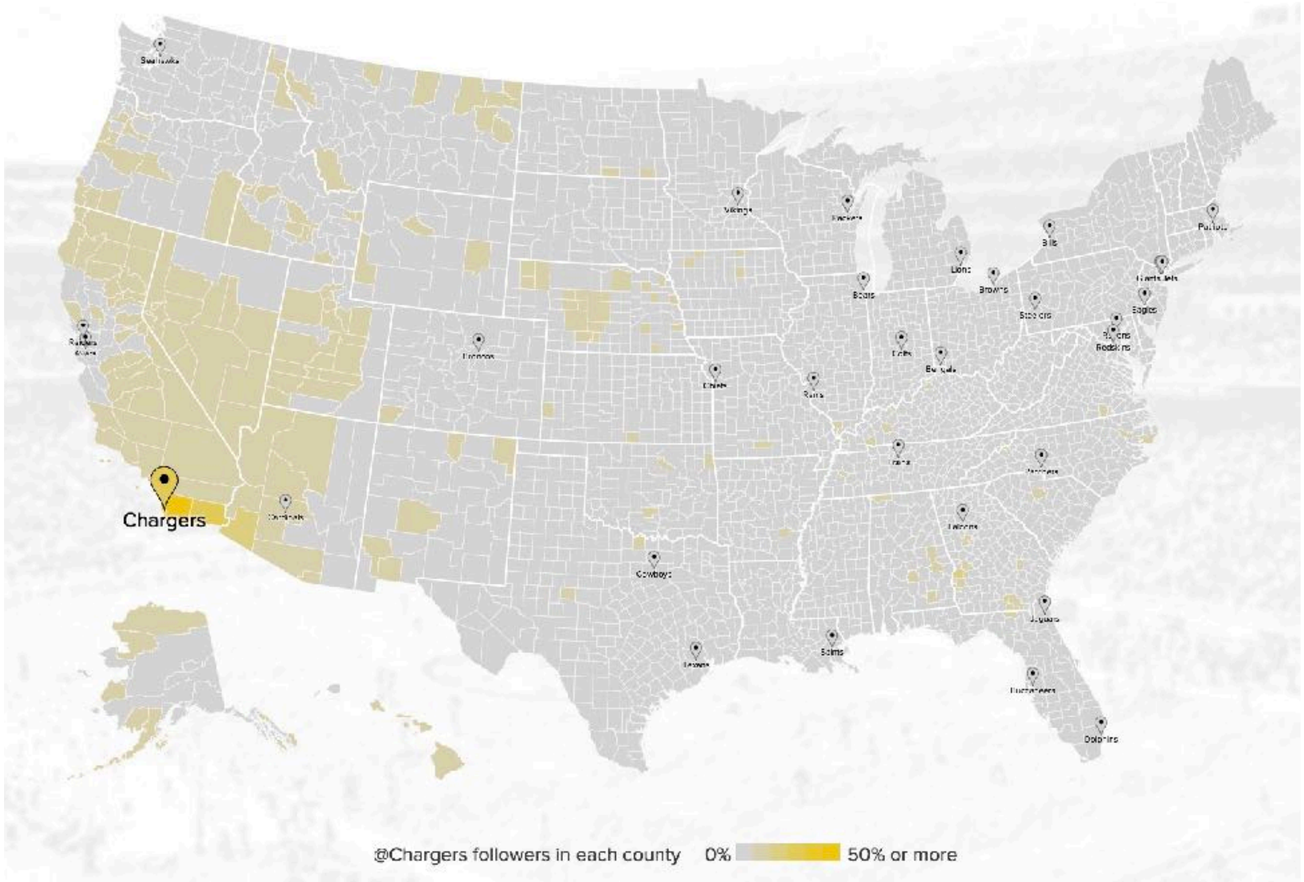
More NFL



Source: https://interactive.twitter.com/nfl_followers2014

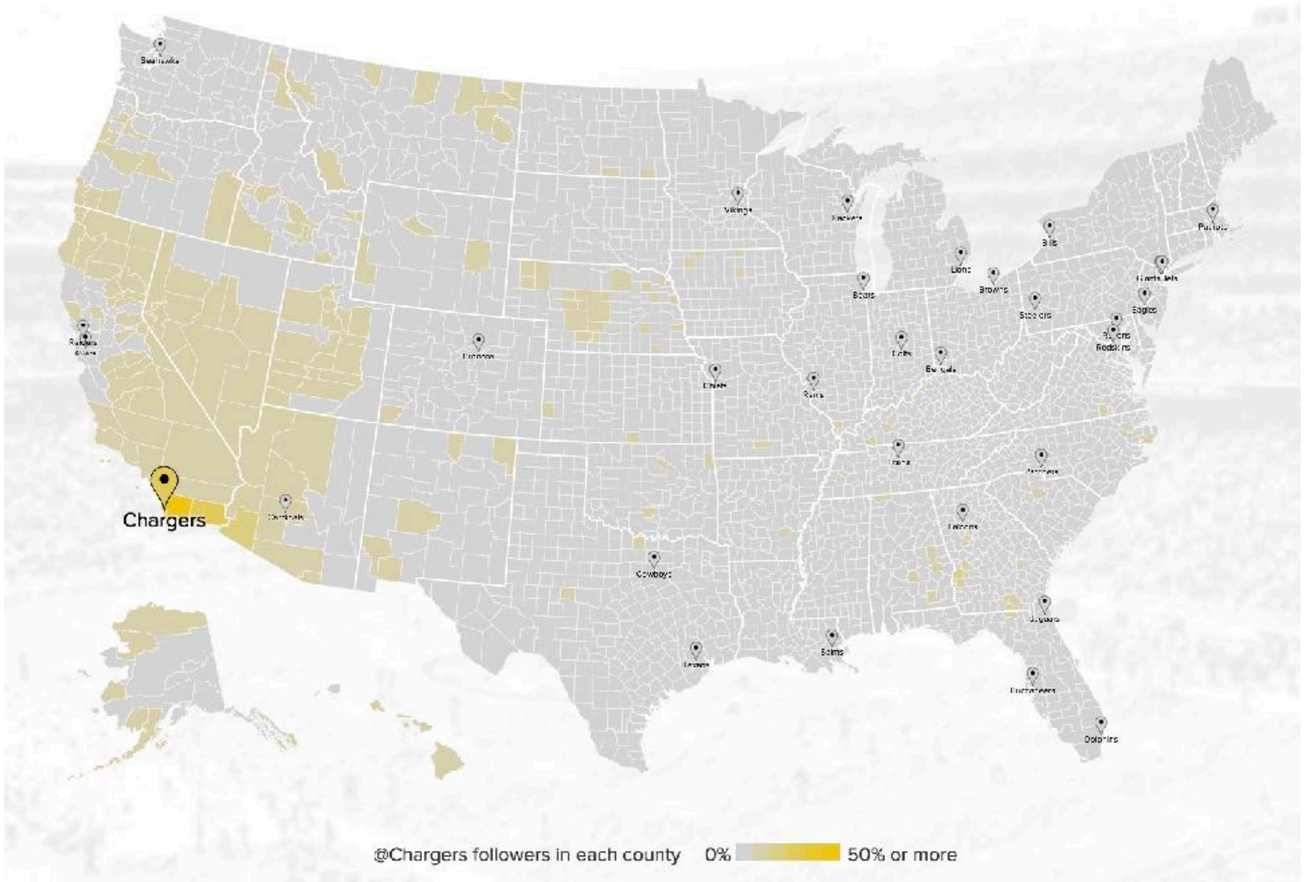
UC San Diego

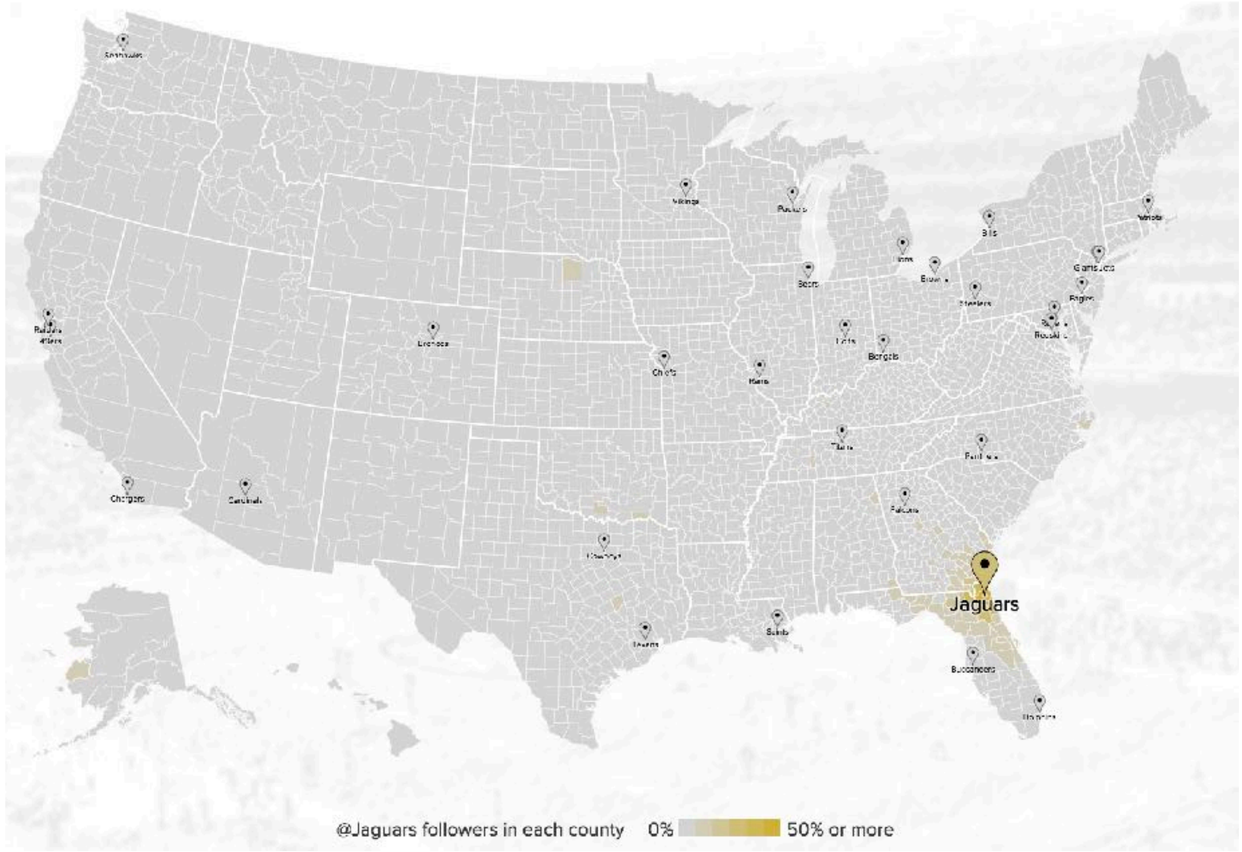
More NFL



Source: https://interactive.twitter.com/nfl_followers2014

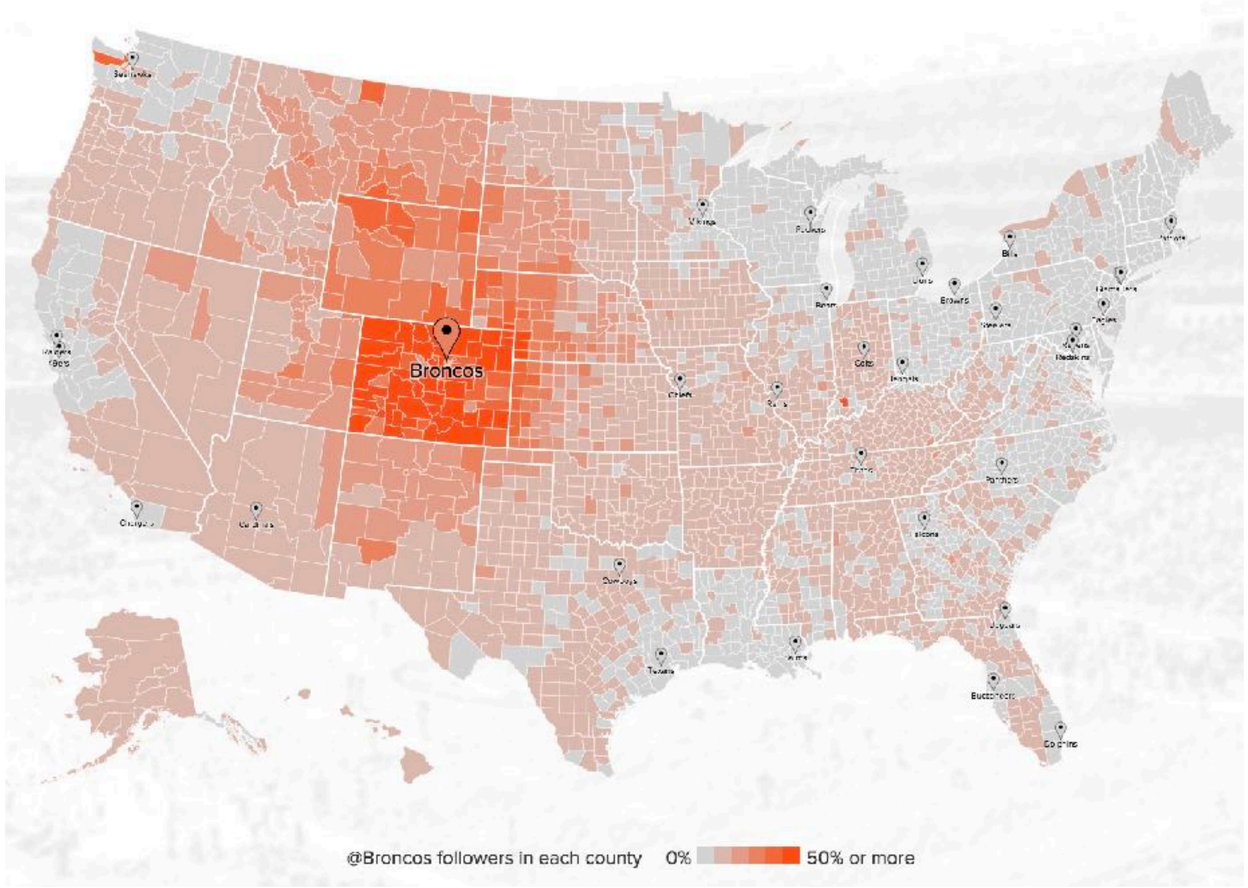
More NFL

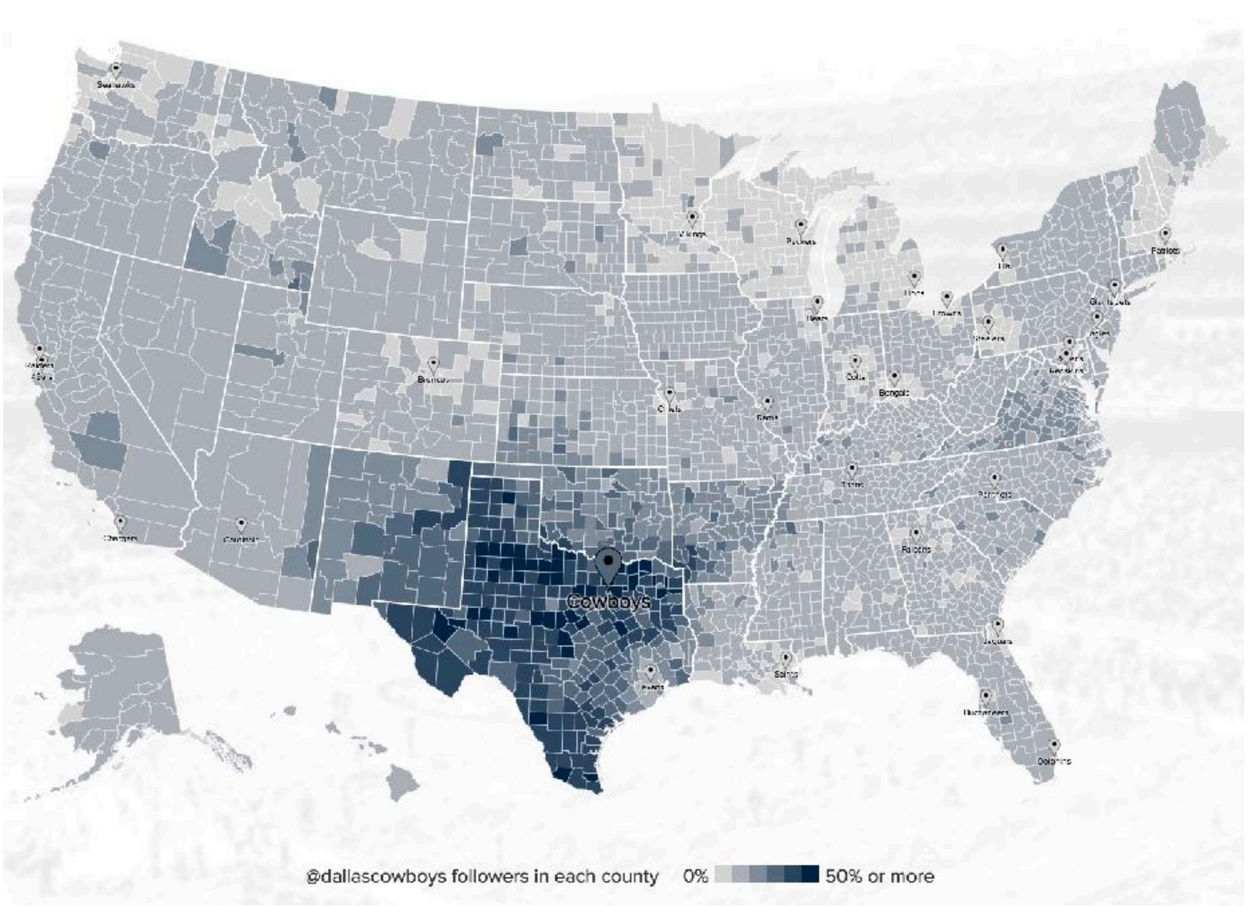




Source: https://interactive.twitter.com/nfl_followers2014

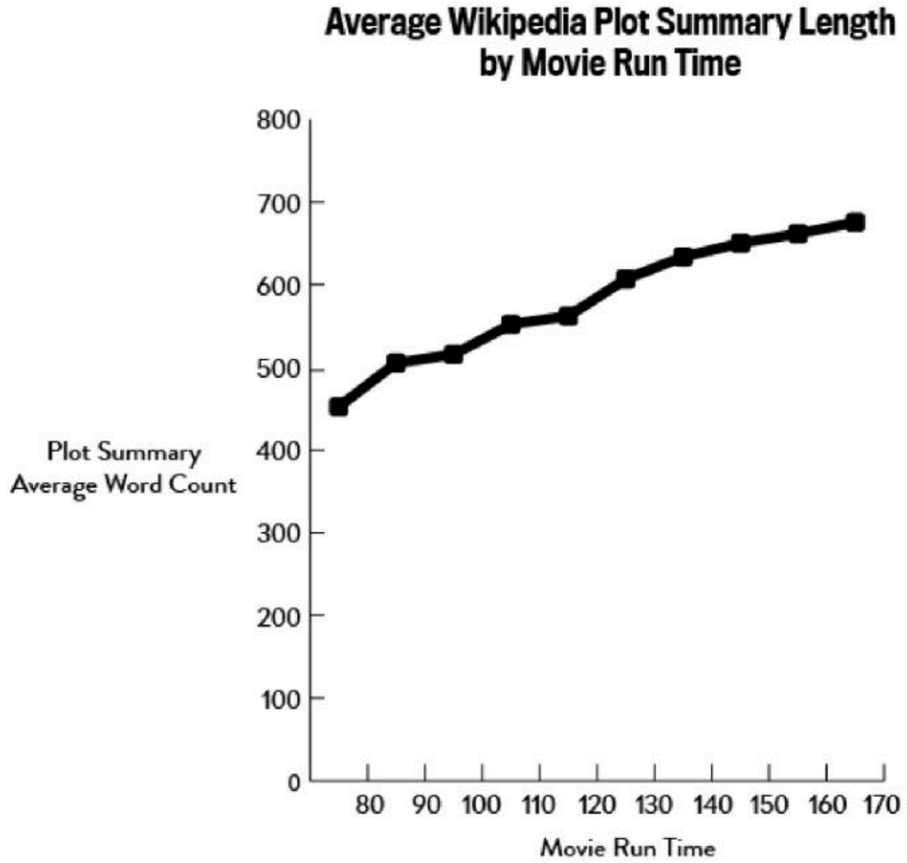
More NFL





Source: https://interactive.twitter.com/nfl_followers2014

Wikipedia



If we consider only films that enjoyed a theatrical release in the United States, the top 10 list looks like this:

1. ***Band of the Hand:*** 3,530 words
2. ***V/H/S/2***: 3,360 words
3. ***Premonition (2007)***: 3,130 words
4. ***Underground (1995)***: 2,824 words
5. ***The Insider***: 2,395 words
6. ***For Colored Girls***: 2,361 words
7. ***2046***: 2,346 words
8. ***Problem Child 2***: 2,282 words
9. ***K-19: The Widowmaker***: 2,046 words
10. ***The Good Girl***: 2,042 words

Many of these articles appear to have been stuffed with detail by superfans.

UC San Diego

Source: Ben Blatt, http://www.slate.com/articles/arts/culturebox/2014/06/what_is_the_most_detailed_wikipedia_movie_plot_summary_of_all_time.html

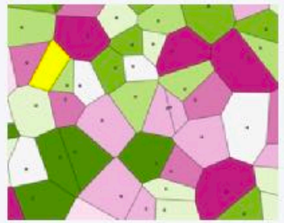
d3.js

Sunburst

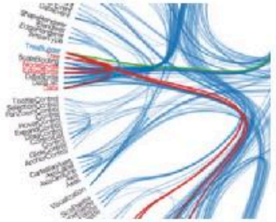
Node-Link Tree

Treemap

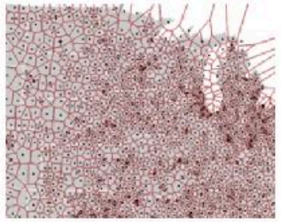
Voronoi Diagram



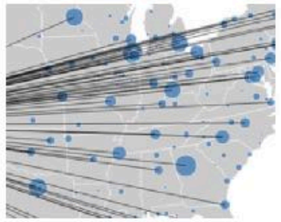
Hierarchical Edge Bundling



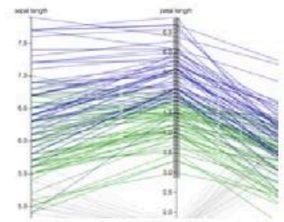
Voronoi Diagram



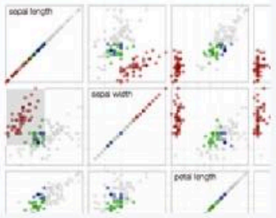
Symbol Map



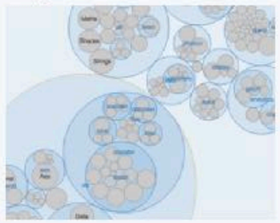
Parallel Coordinates



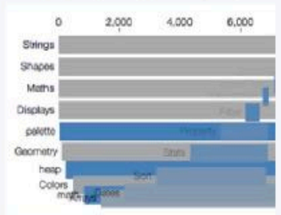
Scatterplot Matrix



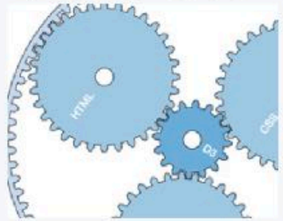
Zoomable Pack Layout



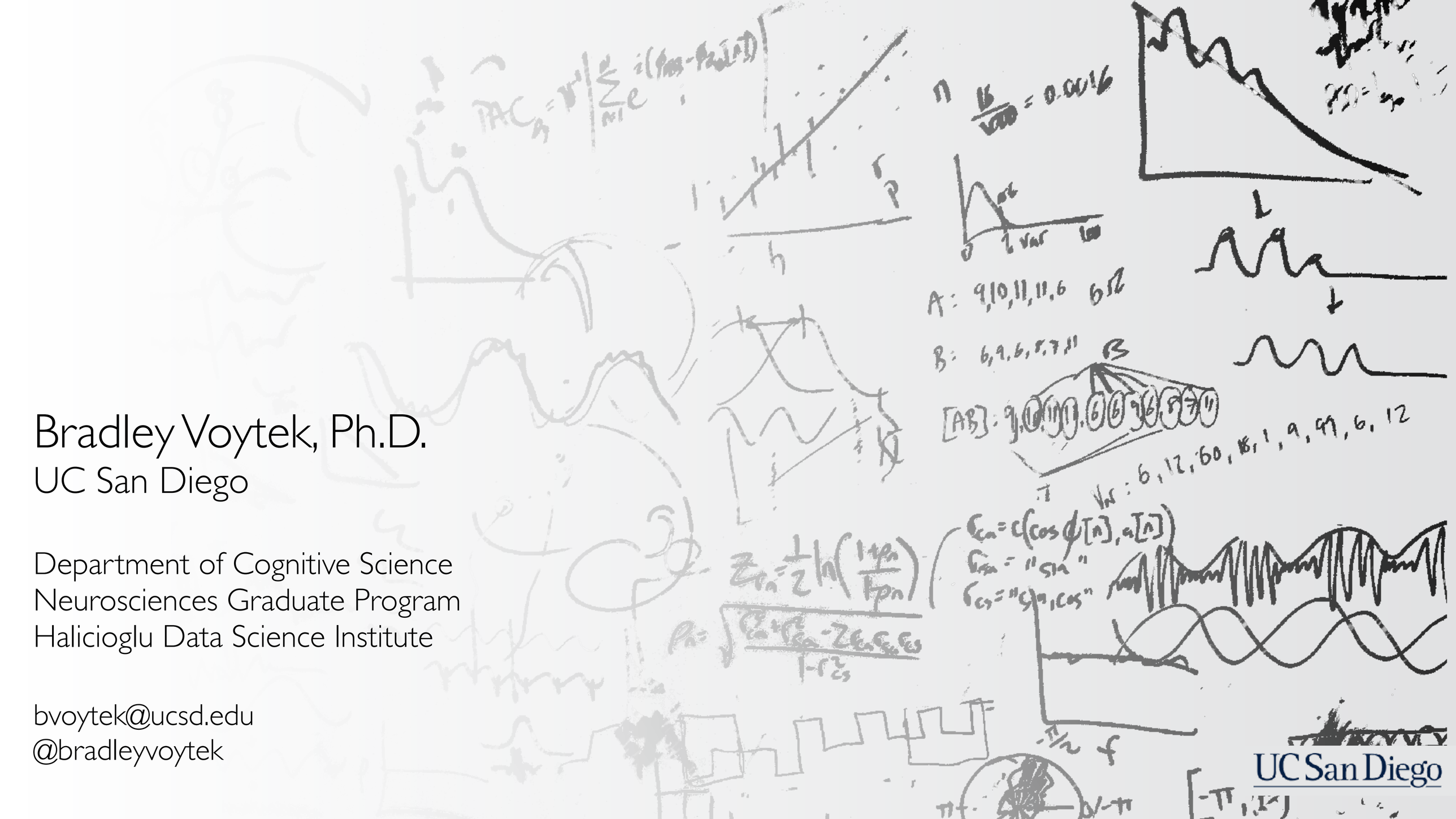
Hierarchical Bars



Epicyclical Gears



Source: https://d3js.org/



Bradley Voytek, Ph.D.

UC San Diego

Department of Cognitive Science
Neurosciences Graduate Program
Halicioglu Data Science Institute

bvoytek@ucsd.edu
@bradleyvoytek

UC San Diego