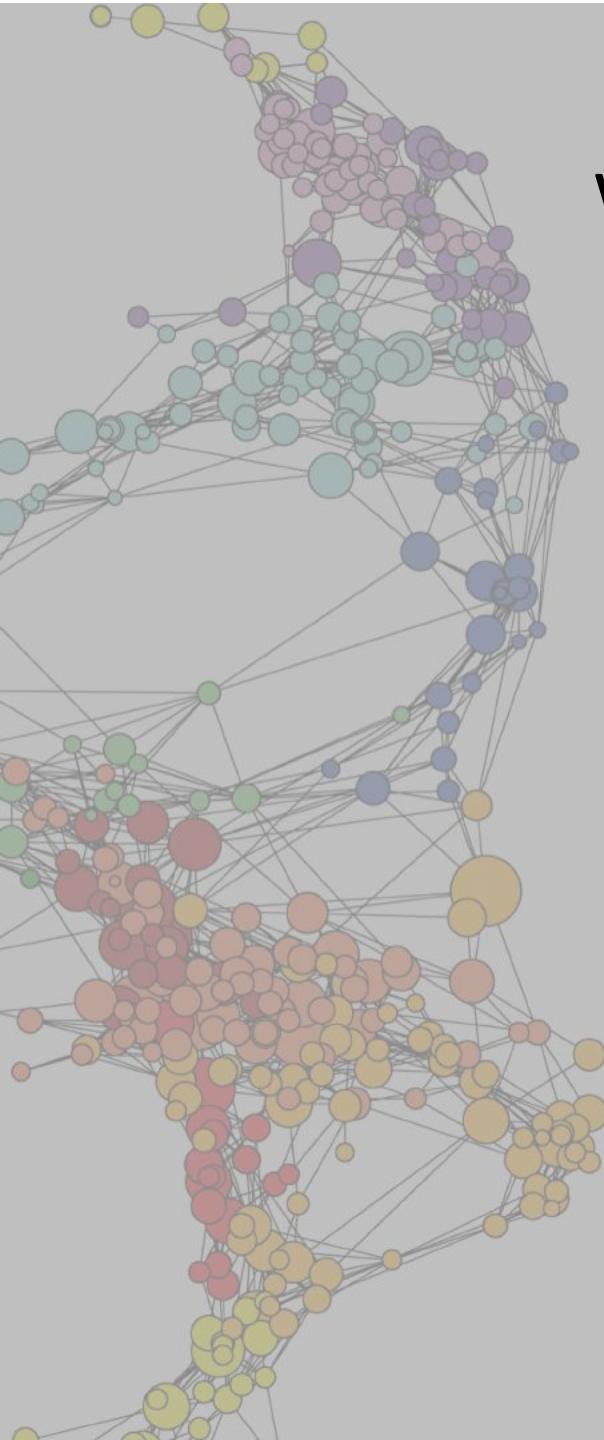


# Tips for Effective Data Visualization

Eric E Monson, PhD  
Duke Data and Visualization Services  
STA199 · Spring 2019

Slides: <https://bit.ly/STA199visSpring2019>



# What is data visualization?

**Anything that converts data sources into a visual representation**

*charts, graphs, maps, even just tables*

<http://guides.library.duke.edu/datavis>

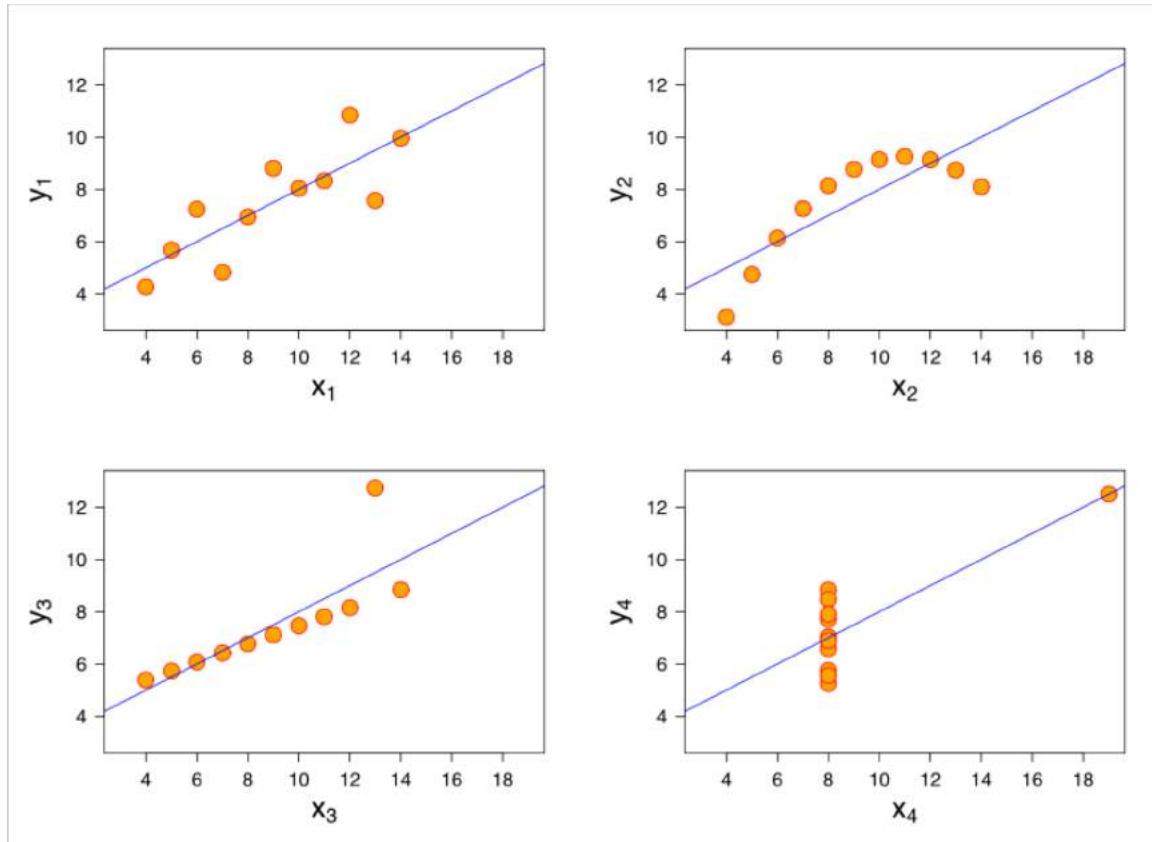
# Why do we visualize?

1		2		3		4	
x	y	x	y	x	y	x	y
10.0	8.04	10.0	9.14	10.0	7.46	8.0	6.58
8.0	6.95	8.0	8.14	8.0	6.77	8.0	5.76
13.0	7.58	13.0	8.74	13.0	12.74	8.0	7.71
9.0	8.81	9.0	8.77	9.0	7.11	8.0	8.84
11.0	8.33	11.0	9.26	11.0	7.81	8.0	8.47
14.0	9.96	14.0	8.10	14.0	8.84	8.0	7.04
6.0	7.24	6.0	6.13	6.0	6.08	8.0	5.25
4.0	4.26	4.0	3.10	4.0	5.39	19.0	12.50
12.0	10.84	12.0	9.13	12.0	8.15	8.0	5.56
7.0	4.82	7.0	7.26	7.0	6.42	8.0	7.91
5.0	5.68	5.0	4.74	5.0	5.73	8.0	6.89

*Almost identical summary statistics:*  
x & y mean  
x & y variance  
x-y correlation  
x-y linear regression

[https://en.wikipedia.org/wiki/Anscombe%27s\\_quartet](https://en.wikipedia.org/wiki/Anscombe%27s_quartet)

# We visualize to see patterns

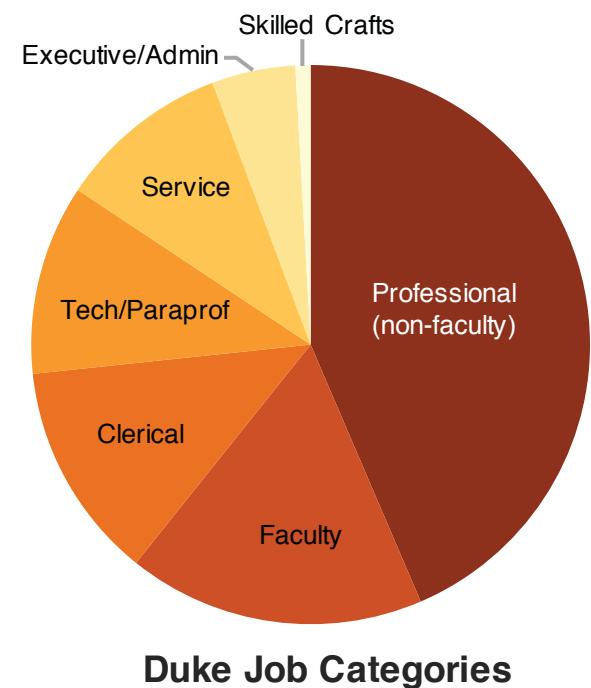
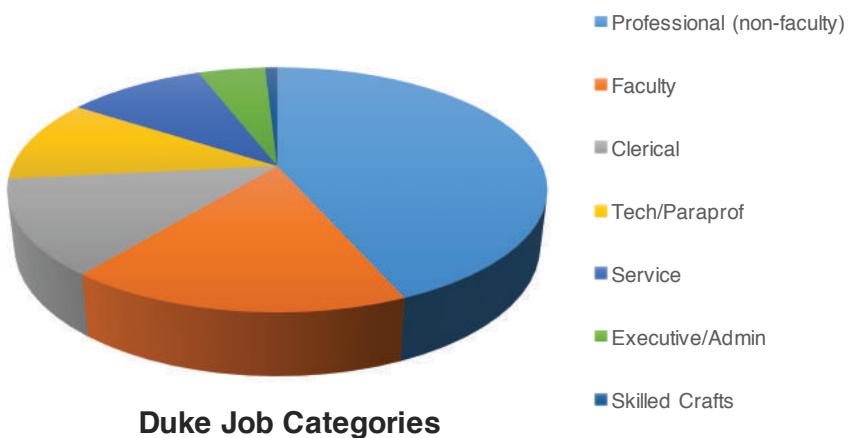


**Anscombe's Quartet**

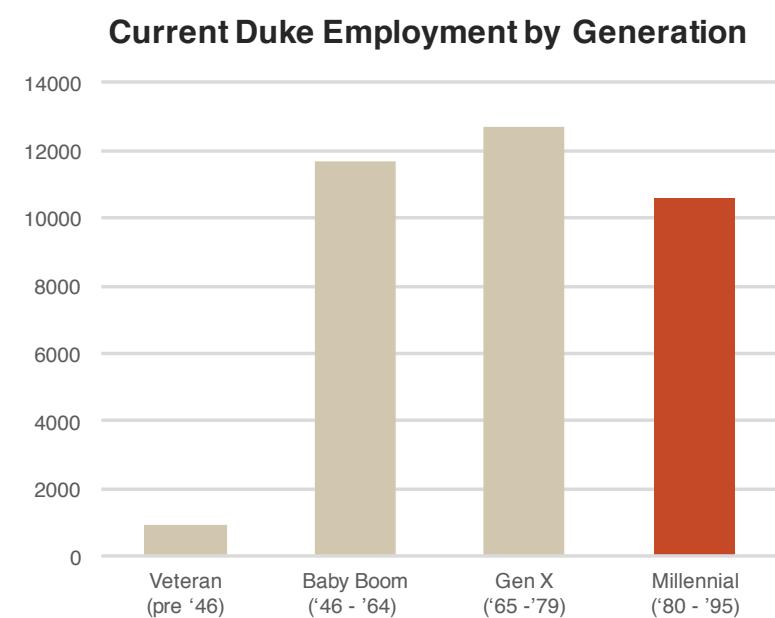
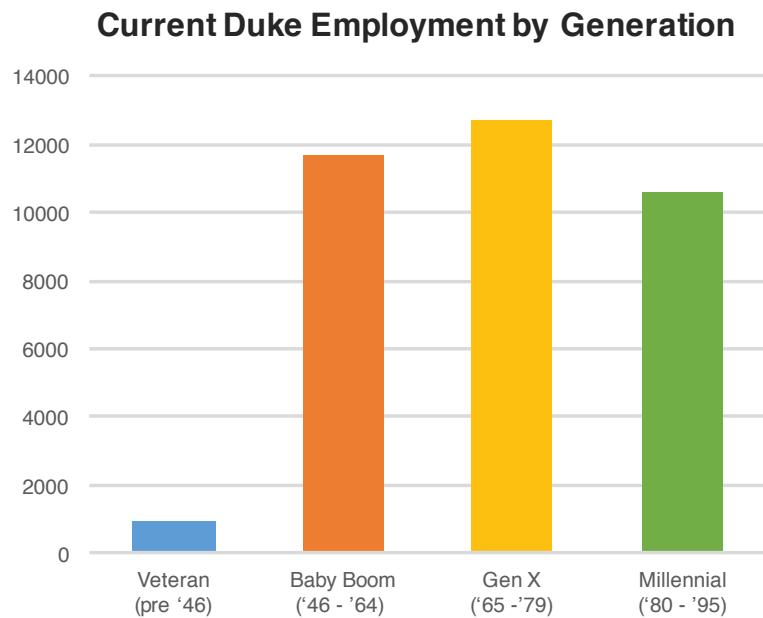
[http://en.wikipedia.org/wiki/Anscombe%27s\\_quartet](http://en.wikipedia.org/wiki/Anscombe%27s_quartet)

# Designing effective visualizations

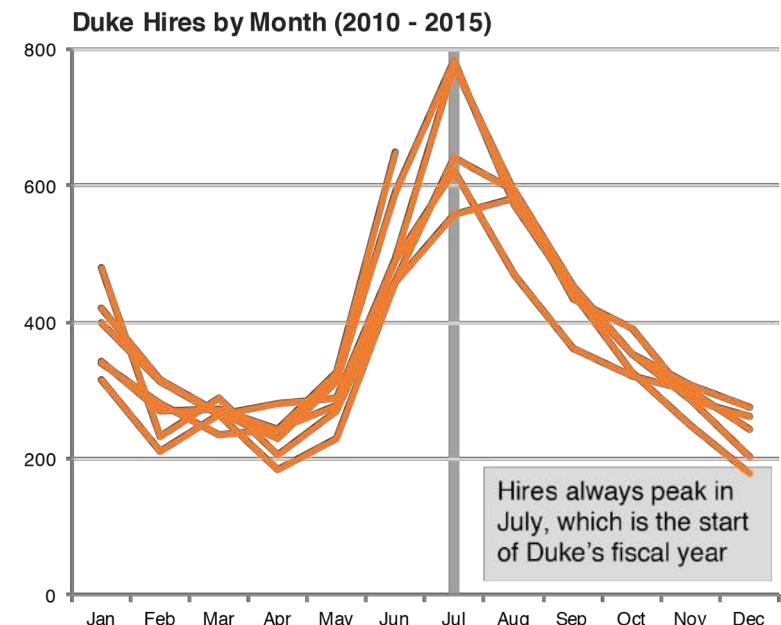
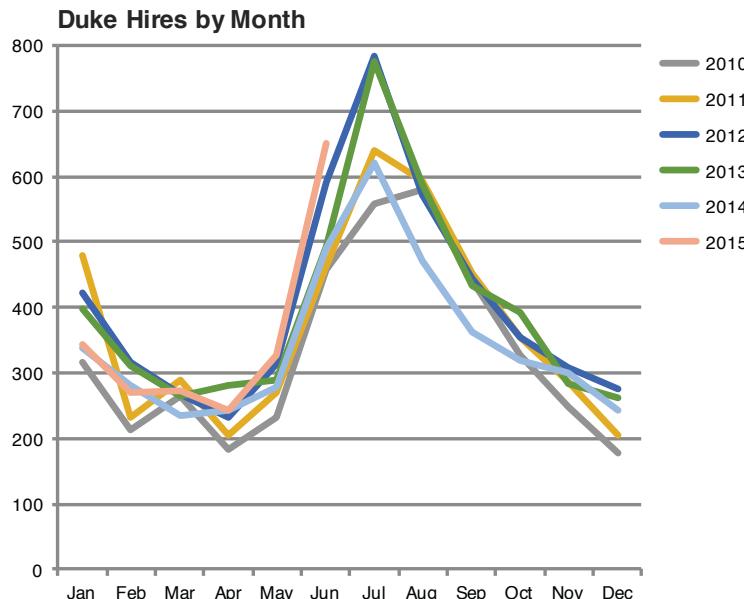
# Keep it simple



# Use color to draw attention

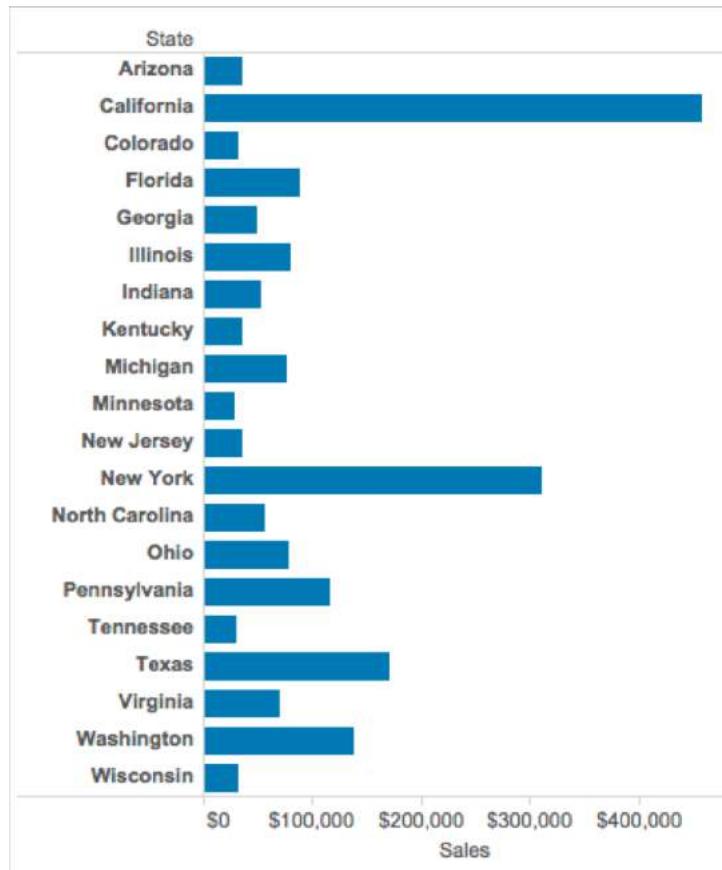


# Every figure should have a purpose: tell a story or make an argument

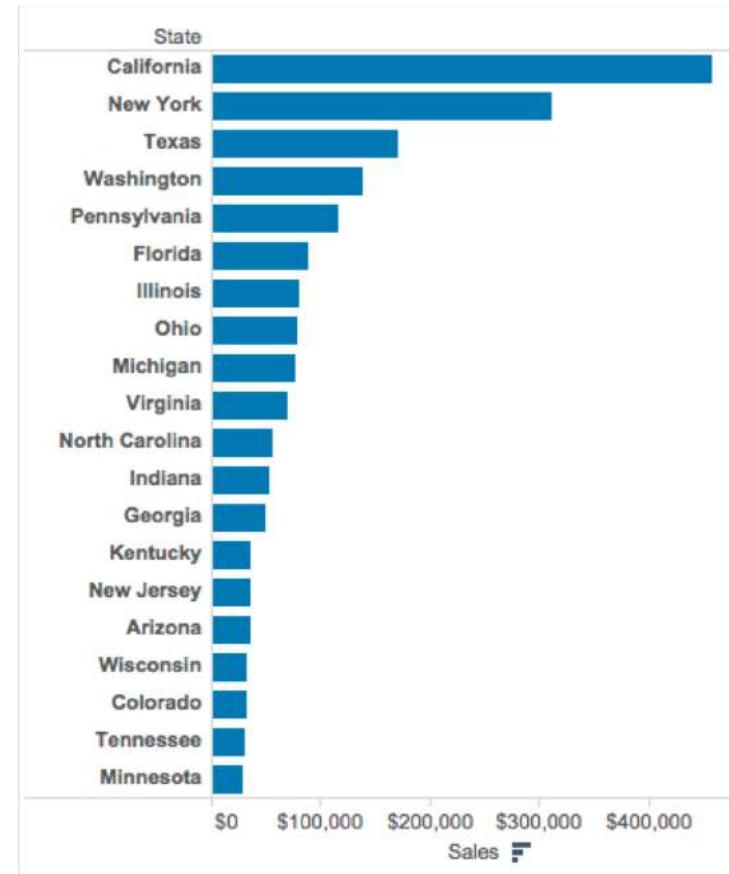
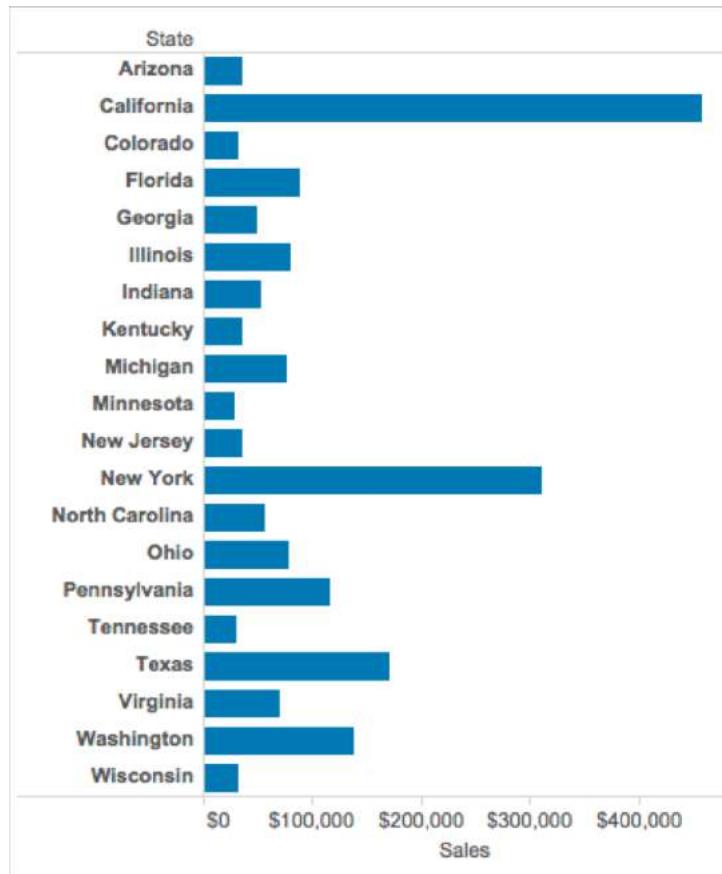


# Common missteps

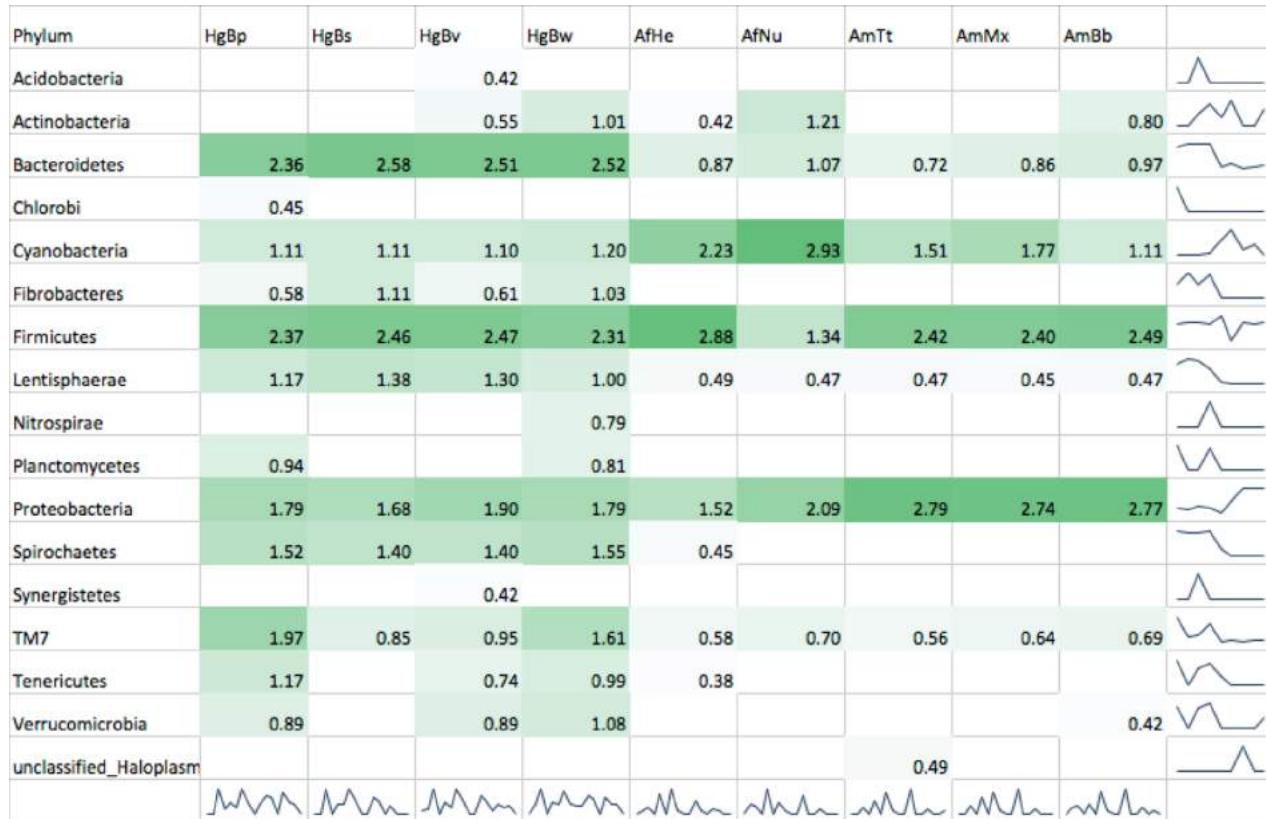
# Default ordering hides patterns



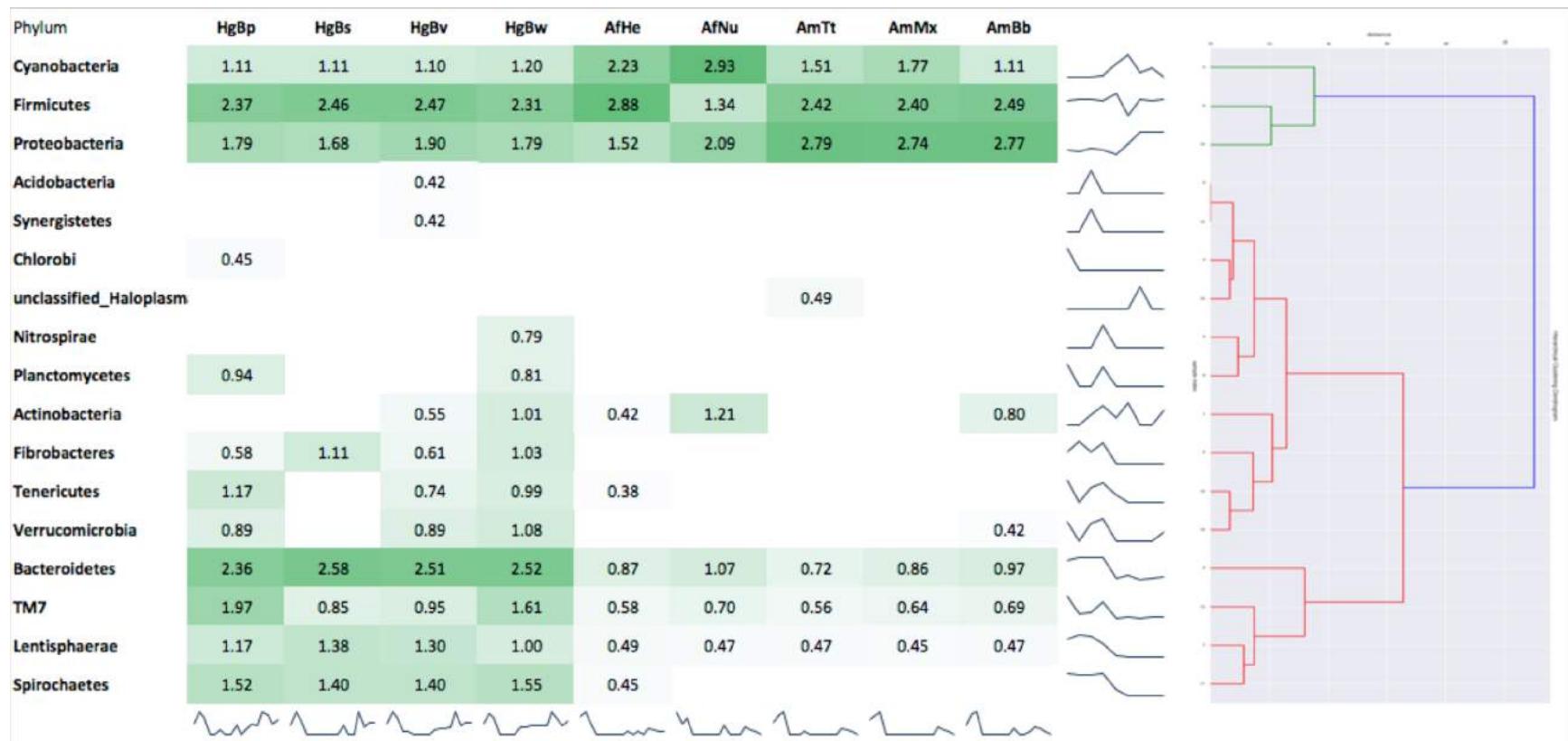
# Sorting reveals patterns



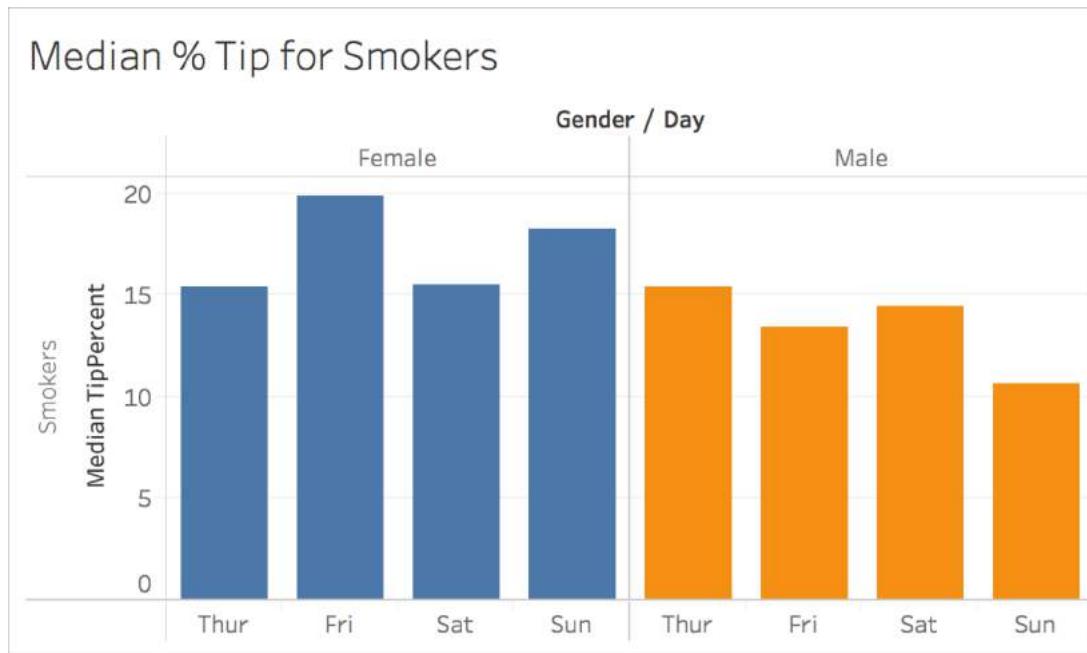
# Alphabetical again hides patterns



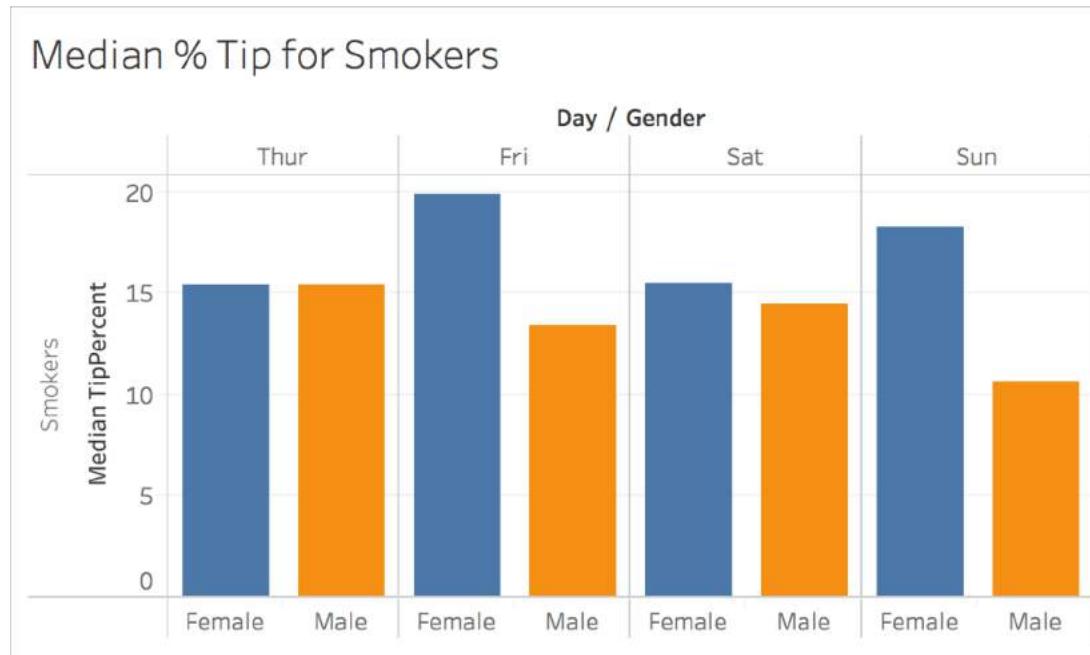
# Clustering to see response groups



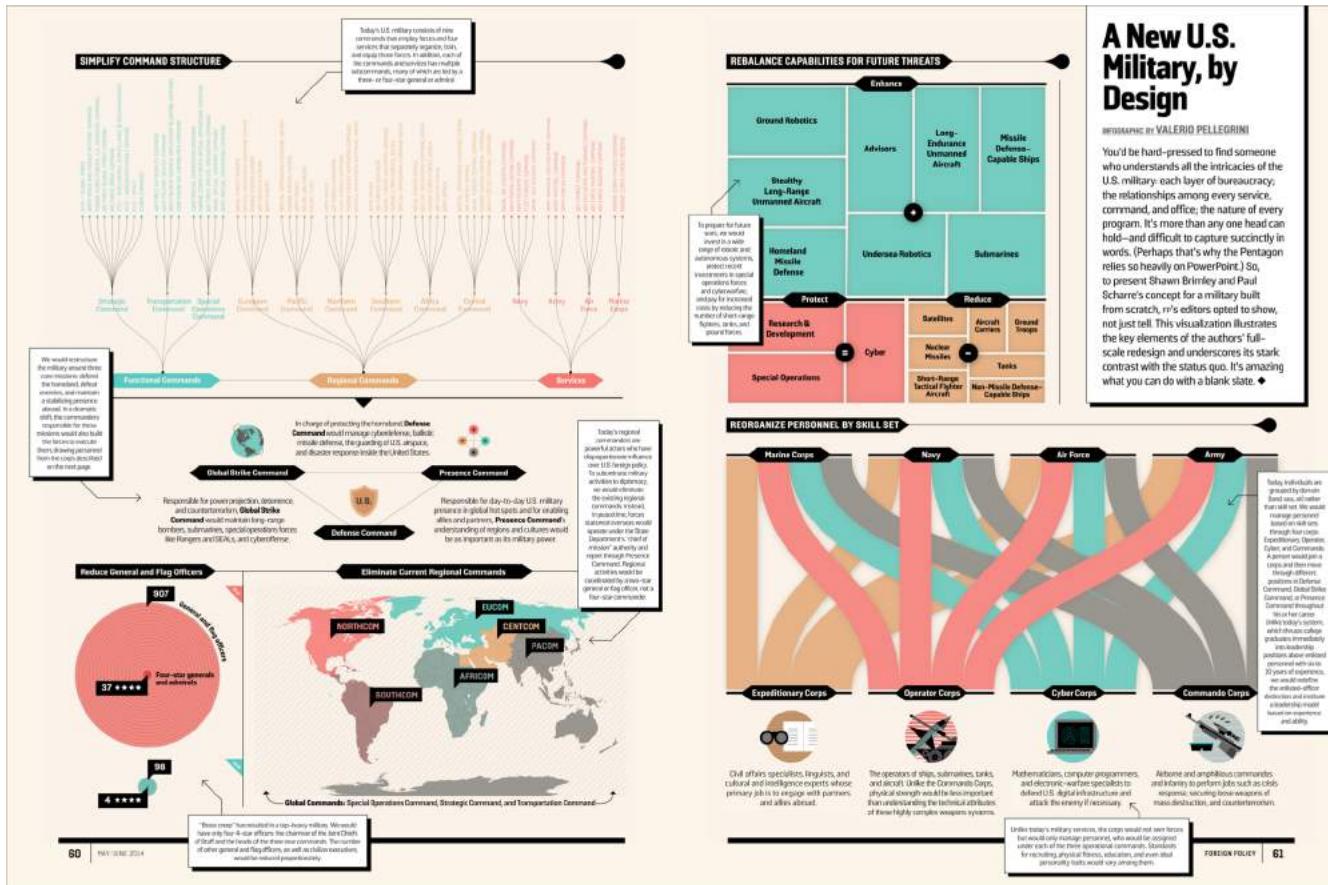
# Distant comparisons are harder



# Put important comparisons nearby



# Colors shouldn't switch meanings



# Colors shouldn't switch meanings

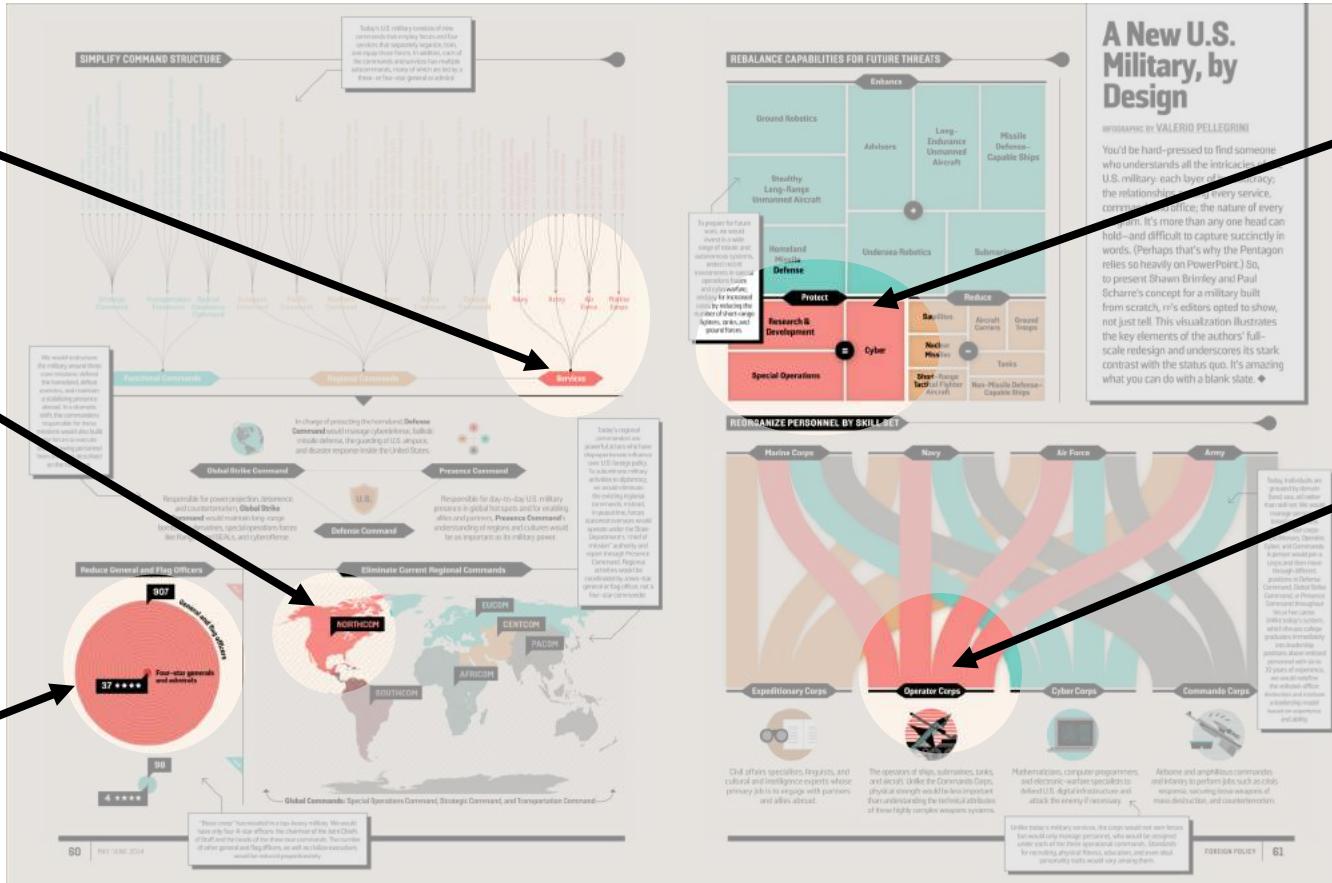
Military services

North America

Old officers

Protected priorities

Operator corps



# Tables easily hide patterns

The study group (to compare with the control group)	Cancer and age group	Type of analysis	White females	AA females
Females living in communities located in zip codes with kind of bad contions	Uterine cancer, all ages	Univariate	1.09, p=0.1027	
		Multivariate	1.24, p=0.0002	
	Uterine cancer, age 65+	Univariate	1.05, p=0.5702	1.05, p=0.6438
		Multivariate	1.18, p=0.0723	1.18, p=0.1789
	Cancer of corpus uteri, age 65+	Univariate	1.13, p=0.2861	1.07, p=0.6547
		Multivariate	1.27, p=0.0513	1.07, p=0.6763
	Uterine cancer, all ages	Univariate	1.29, p=0.0665	1.14, p=0.3630
		Multivariate	1.57, p=0.0019	1.28, p=0.1440
	Uterine cancer, age 65+	Univariate	1.55, p=0.0061	1.16, p=0.3800
		Multivariate	1.94, p=7.3x10 <sup>-5</sup>	1.27, p=0.2300
	Cancer of corpus uteri, all ages	Univariate	1.68, p=0.0022	1.19, p=0.3820
		Multivariate	1.98, p=0.0001	1.09, p=0.7170
	Cancer of corpus uteri, age 65+	Univariate	1.91, p=0.0011	1.20, p=0.4330
		Multivariate	2.41, p=3.01x10 <sup>-5</sup>	1.07, p=0.8120

# Highlight patterns for the reader

**The study group  
(to compare with  
the control group)**

	Cancer	Age group	Analysis type	White females	p-value	AA females	p-value
<b>Females living in communities located in zip codes with kind of bad conditions</b>	Uterine	all ages	Univariate	1.09	0.1027	1.09	0.1027
			Multivariate	<b>1.24</b>		<b>0.0002</b>	<b>0.0002</b>
	Uterine	65+	Univariate	1.05	0.5702	1.05	0.6438
			Multivariate	1.18		1.18	
	Corpus uteri	65+	Univariate	1.13	0.2861	1.07	0.6547
			Multivariate	1.27		1.07	
	Uterine	all ages	Univariate	1.29	0.0665	1.14	0.3630
			Multivariate	<b>1.57</b>		<b>0.0019</b>	<b>0.1440</b>
	Uterine	65+	Univariate	<b>1.55</b>	<b>7.30E-05</b>	1.16	0.3800
			Multivariate	<b>1.94</b>		1.27	
	Corpus uteri	all ages	Univariate	<b>1.68</b>	<b>0.0022</b>	1.19	0.3820
			Multivariate	<b>1.98</b>		<b>0.0001</b>	<b>0.7170</b>
	Corpus uteri	65+	Univariate	<b>1.91</b>	<b>0.0011</b>	1.2	0.4330
			Multivariate	<b>2.41</b>		<b>3.01E-05</b>	<b>0.8120</b>

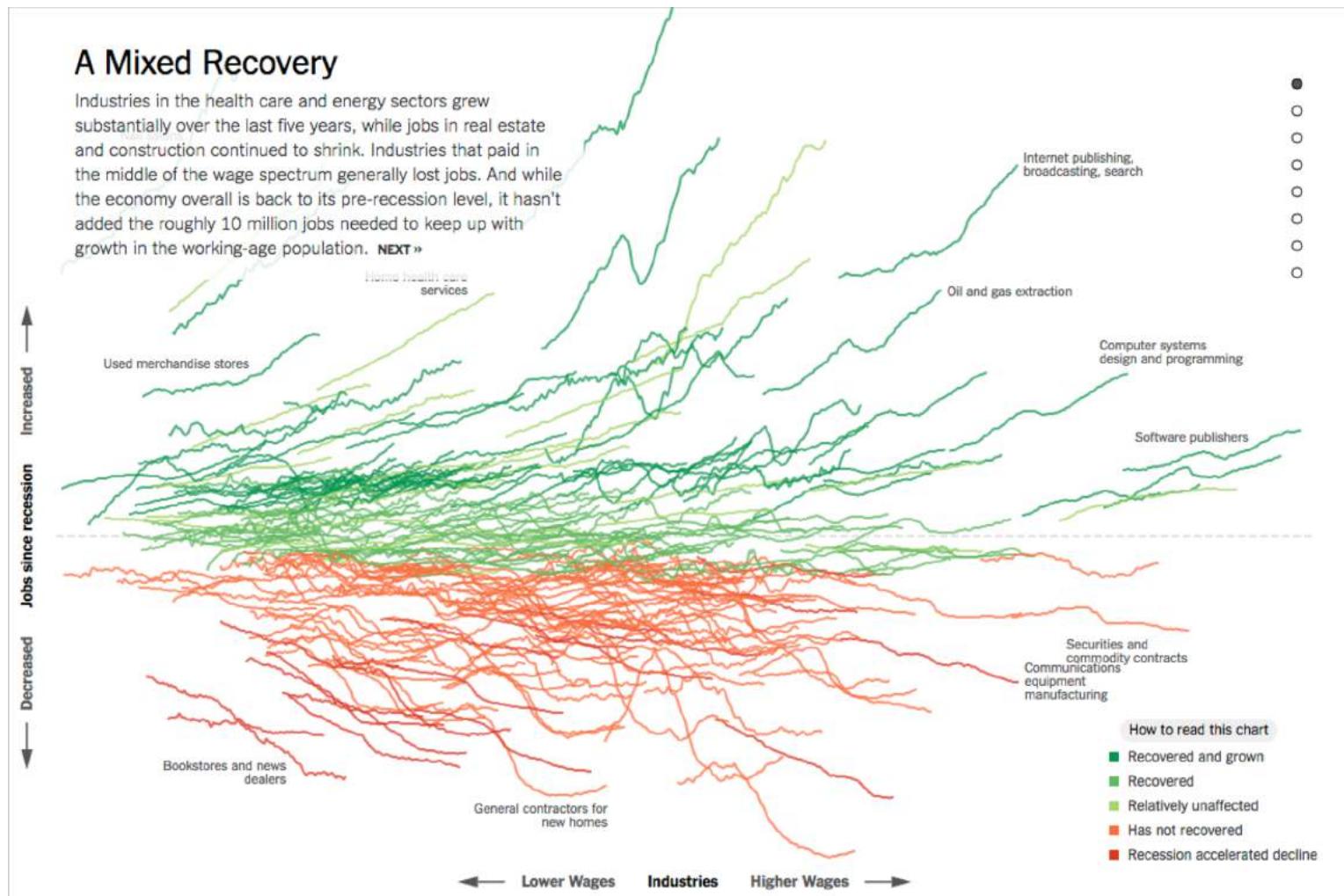
# Highlight patterns for the reader

**The study group  
(to compare with  
the control group)**

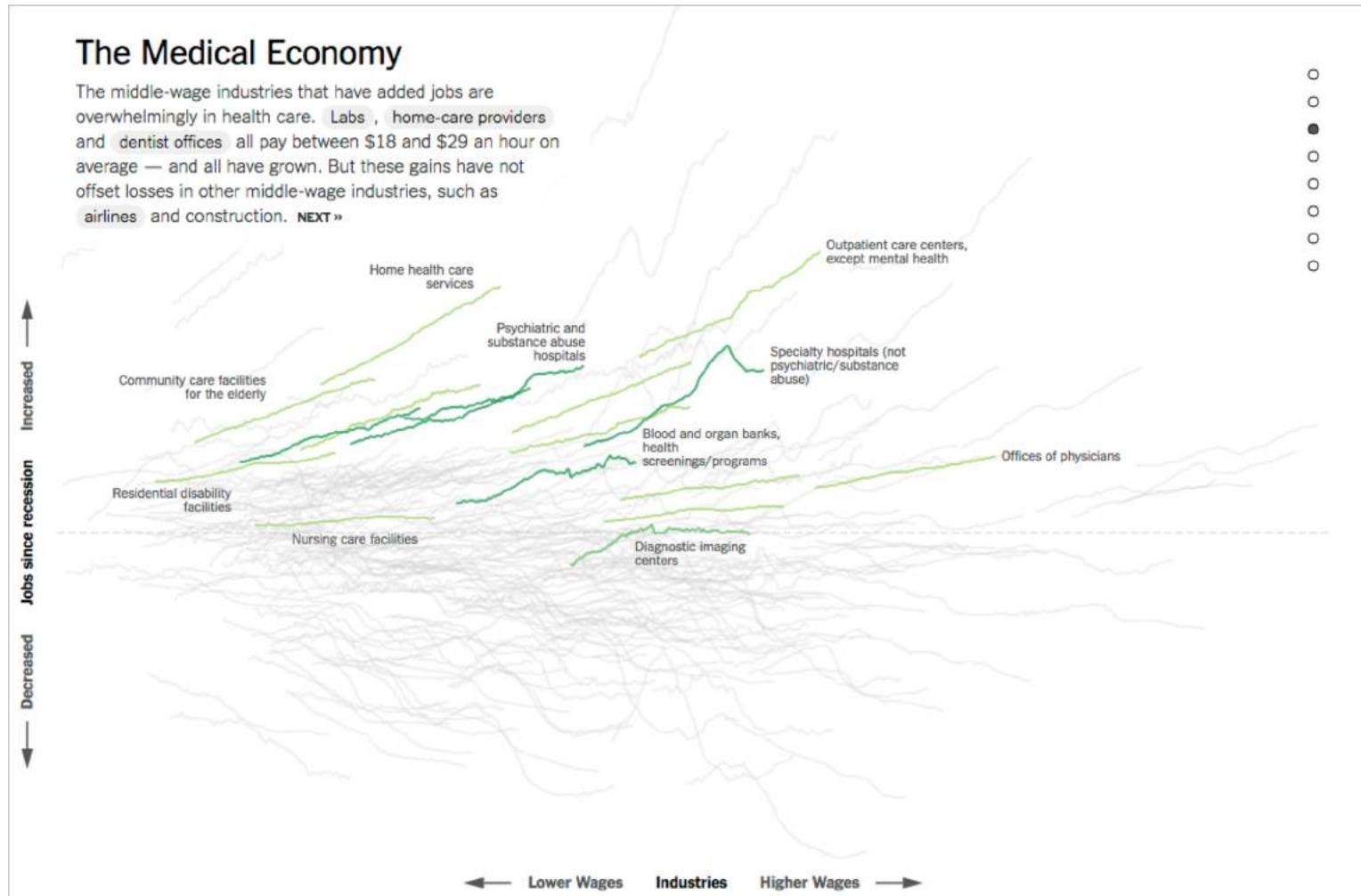
	Cancer	Age group	Analysis type	White females	p-value	AA females	p-value
Females living in communities located in zip codes <b>with kind of bad conditions</b>	Uterine	all ages	Univariate	1.09	0.1027	1.09	0.1027
			Multivariate	<b>1.24</b>	<b>0.0002</b>	<b>1.24</b>	<b>0.0002</b>
	Uterine	65+	Univariate	1.05	0.5702	1.05	0.6438
			Multivariate	1.18	0.0723	1.18	0.1789
	Corpus uteri	65+	Univariate	1.13	0.2861	1.07	0.6547
			Multivariate	1.27	0.0513	1.07	0.6763
Females living in communities with <b>lots of really bad stuff</b>	Uterine	all ages	Univariate	1.29	0.0665	1.14	0.3630
			Multivariate	<b>1.57</b>	<b>0.0019</b>	1.28	0.1440
	Uterine	65+	Univariate	<b>1.55</b>	<b>0.0061</b>	1.16	0.3800
			Multivariate	<b>1.94</b>	<b>7.30E-05</b>	1.27	0.2300
	Corpus uteri	all ages	Univariate	<b>1.68</b>	<b>0.0022</b>	1.19	0.3820
			Multivariate	<b>1.98</b>	<b>0.0001</b>	1.09	0.7170
	Corpus uteri	65+	Univariate	<b>1.91</b>	<b>0.0011</b>	1.2	0.4330
			Multivariate	<b>2.41</b>	<b>3.01E-05</b>	1.07	0.8120

Significantly impacted residents

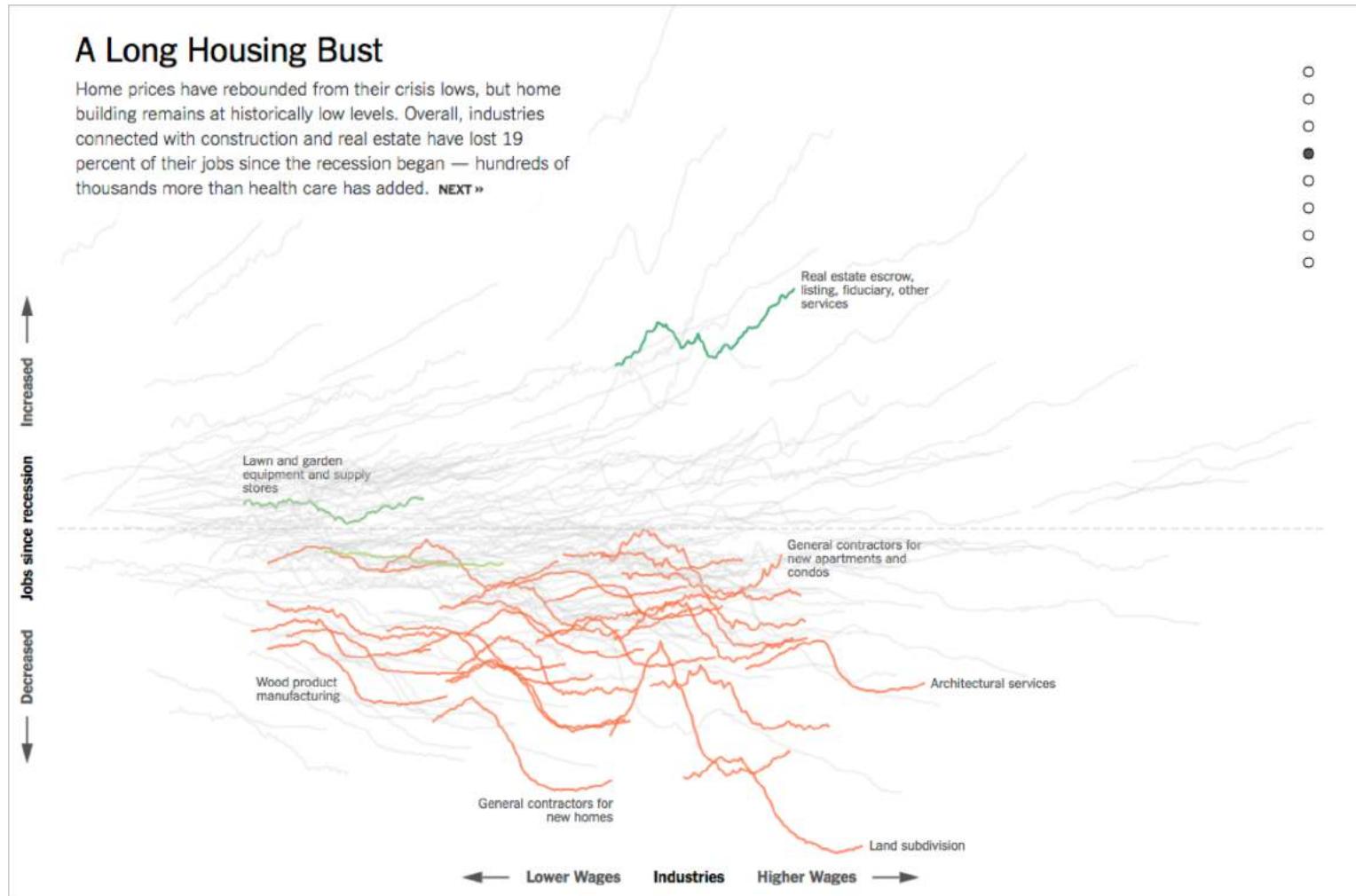
# All the data doesn't tell a story



# All the data doesn't tell a story



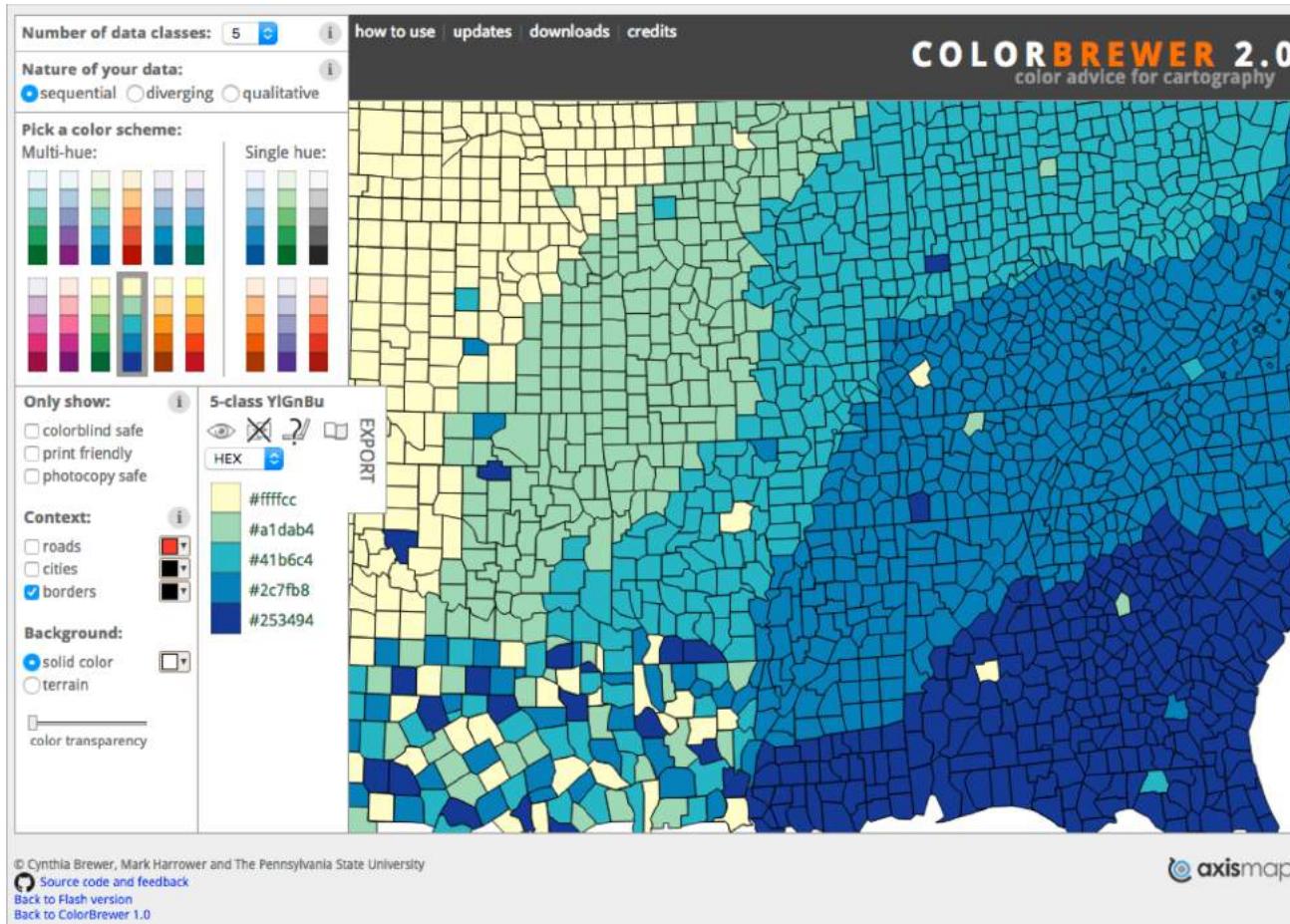
# All the data doesn't tell a story



# Color

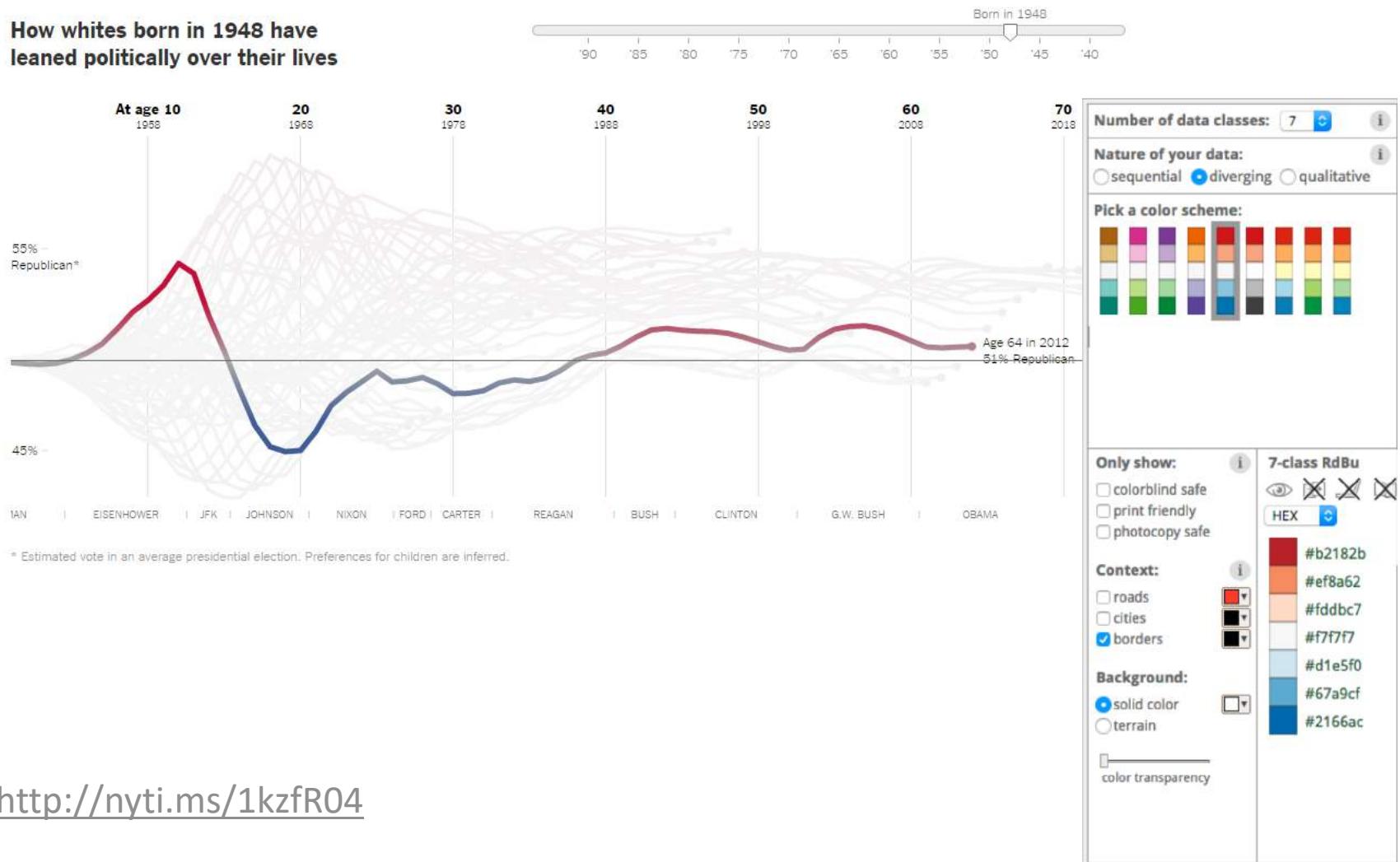
# Luminance ramp for good colormap

<http://colorbrewer2.org/>



# Diverging if there's a natural center

How whites born in 1948 have leaned politically over their lives



<http://nyti.ms/1kzfR04>

# Keep colors perceptually similar

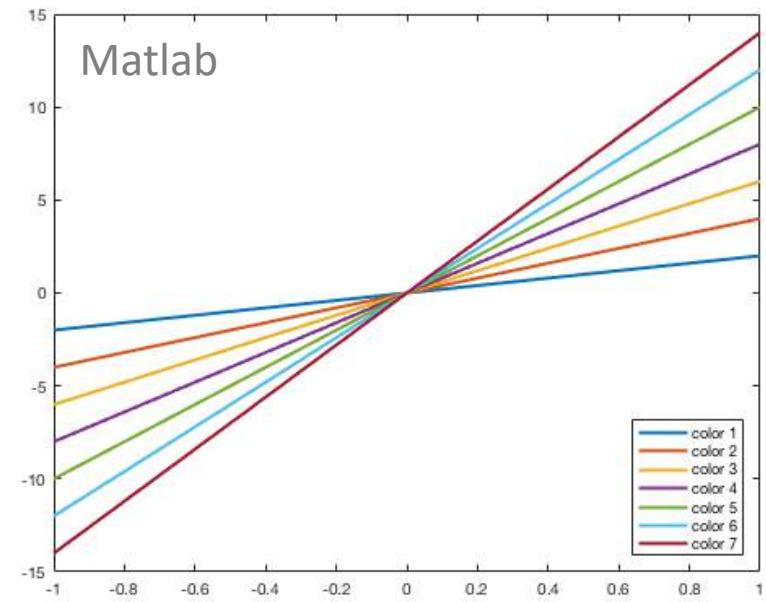
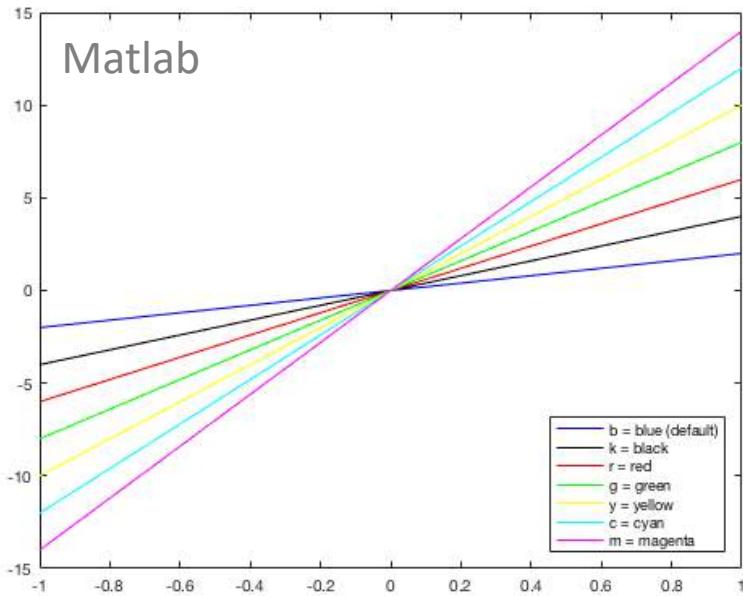
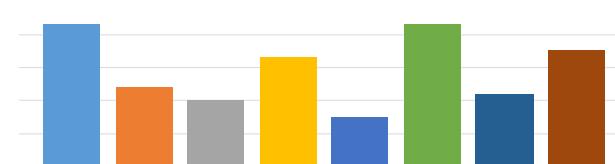


Chart Fills	16	17	18	19	20	21	22	23
Chart Lines	24	25	26	27	28	29	30	31

Excel



Excel

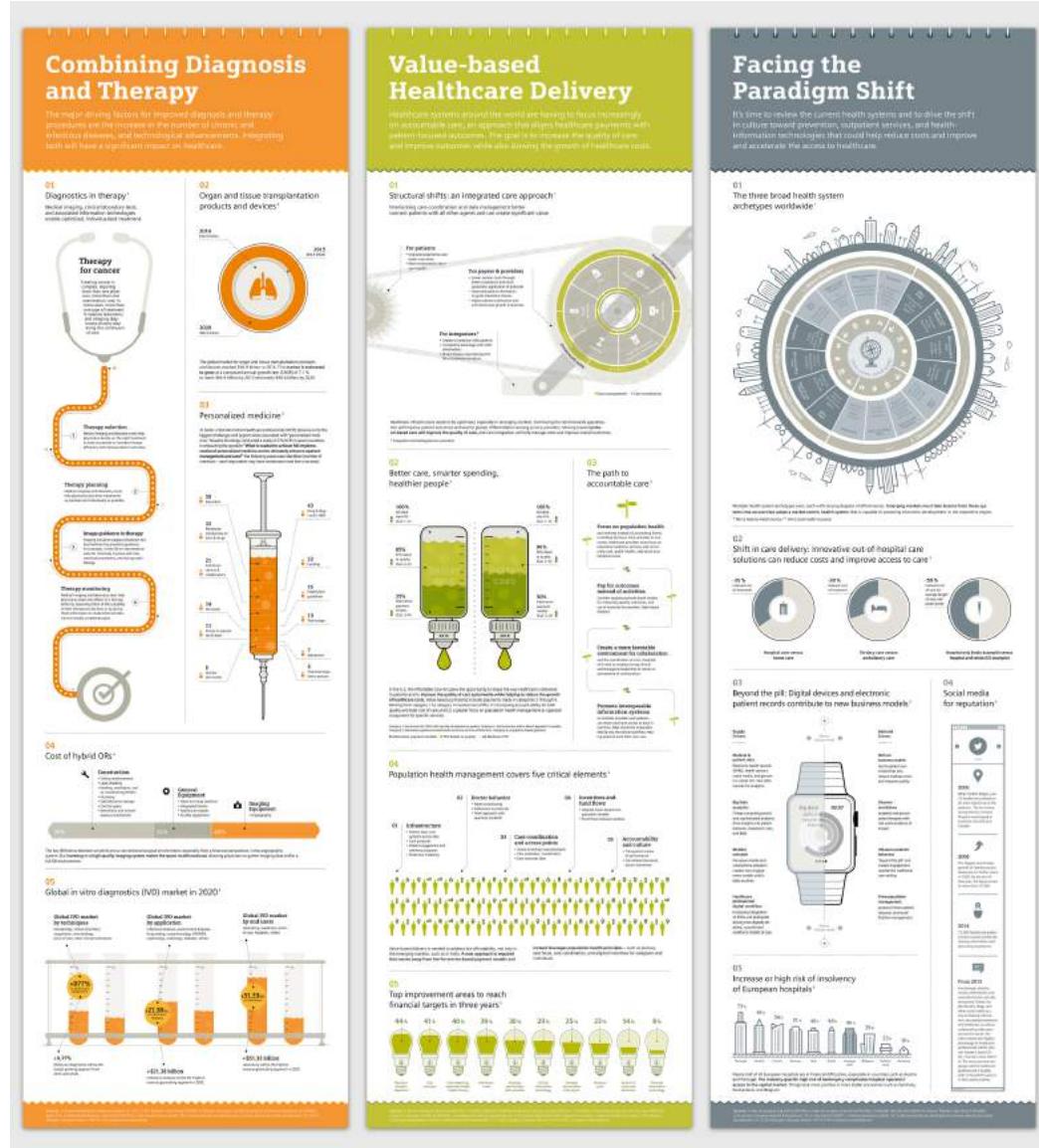


Tableau

# Color

*Professional designs often have more limited, muted colors*

*Use to create hierarchy and unity*

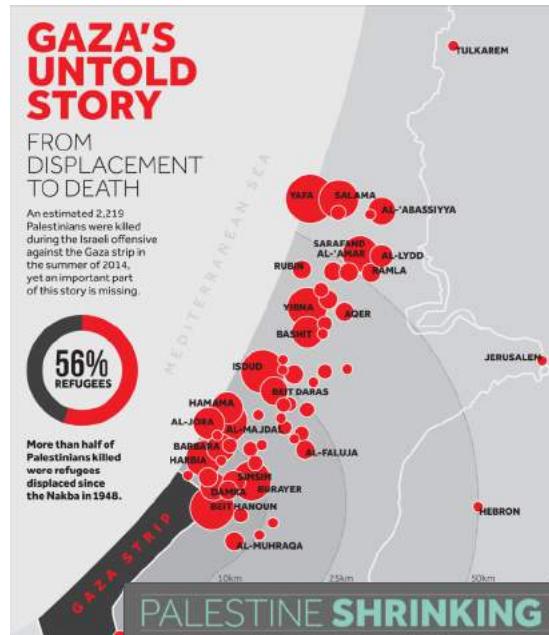


<https://www.informationisbeautifulawards.com/showcase/2370-global-trends-in-healthcare-5-part-infographic-for-siemens-healthcare>

# Color

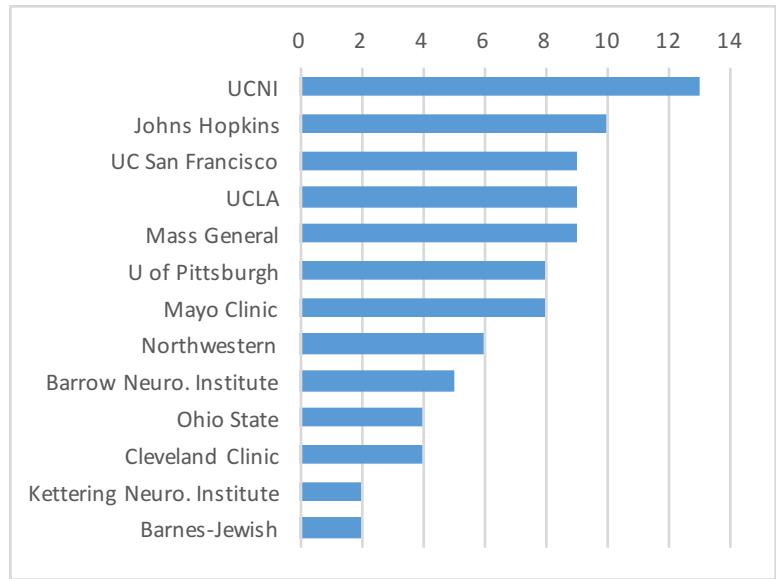
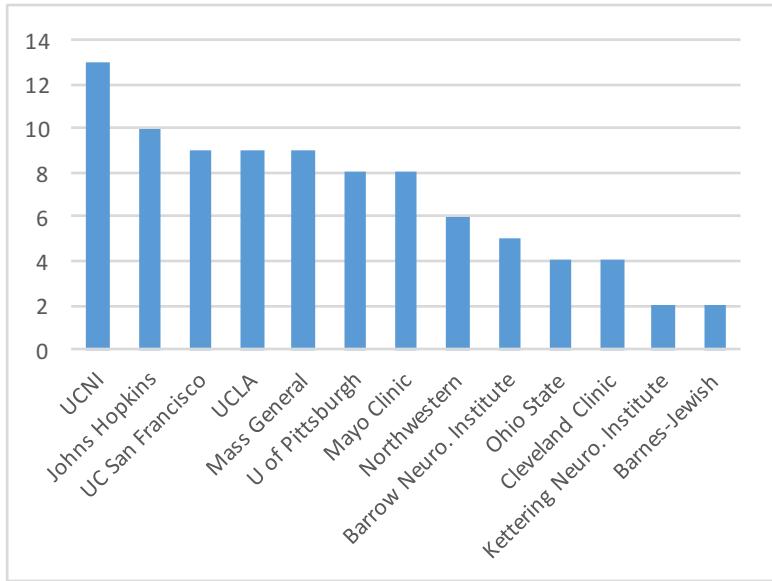
*Professional designs often have more limited, muted colors*

*Use to create hierarchy and unity*



Text to clarify

# Keep text horizontal

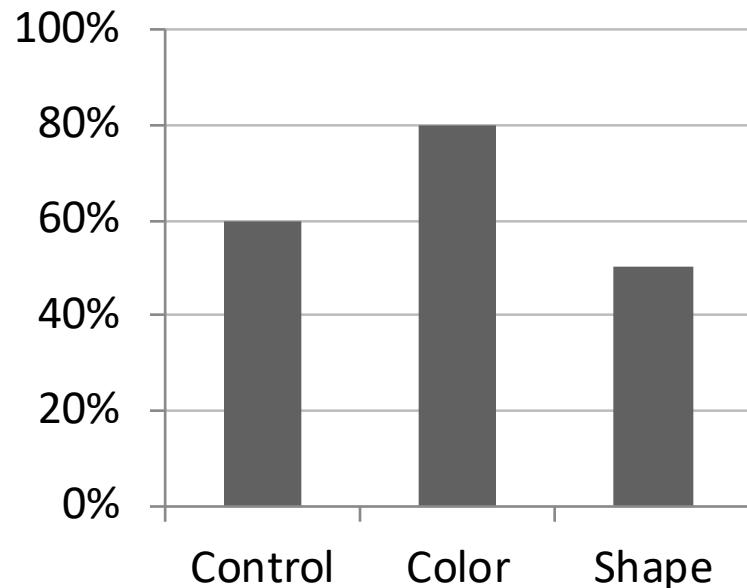


<http://www.storytellingwithdata.com/2012/09/some-finer-points-of-data-visualization.html>

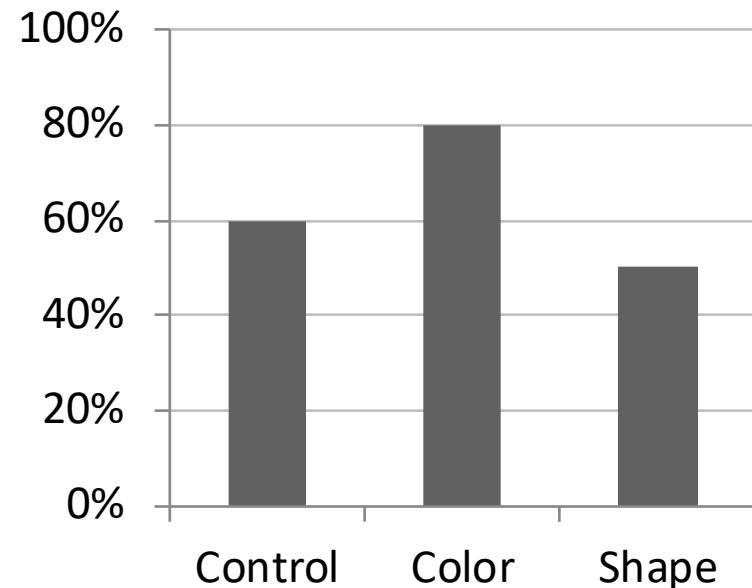
# Use descriptive titles

*Active titles summarize trends in the figure  
and reinforce your message.*

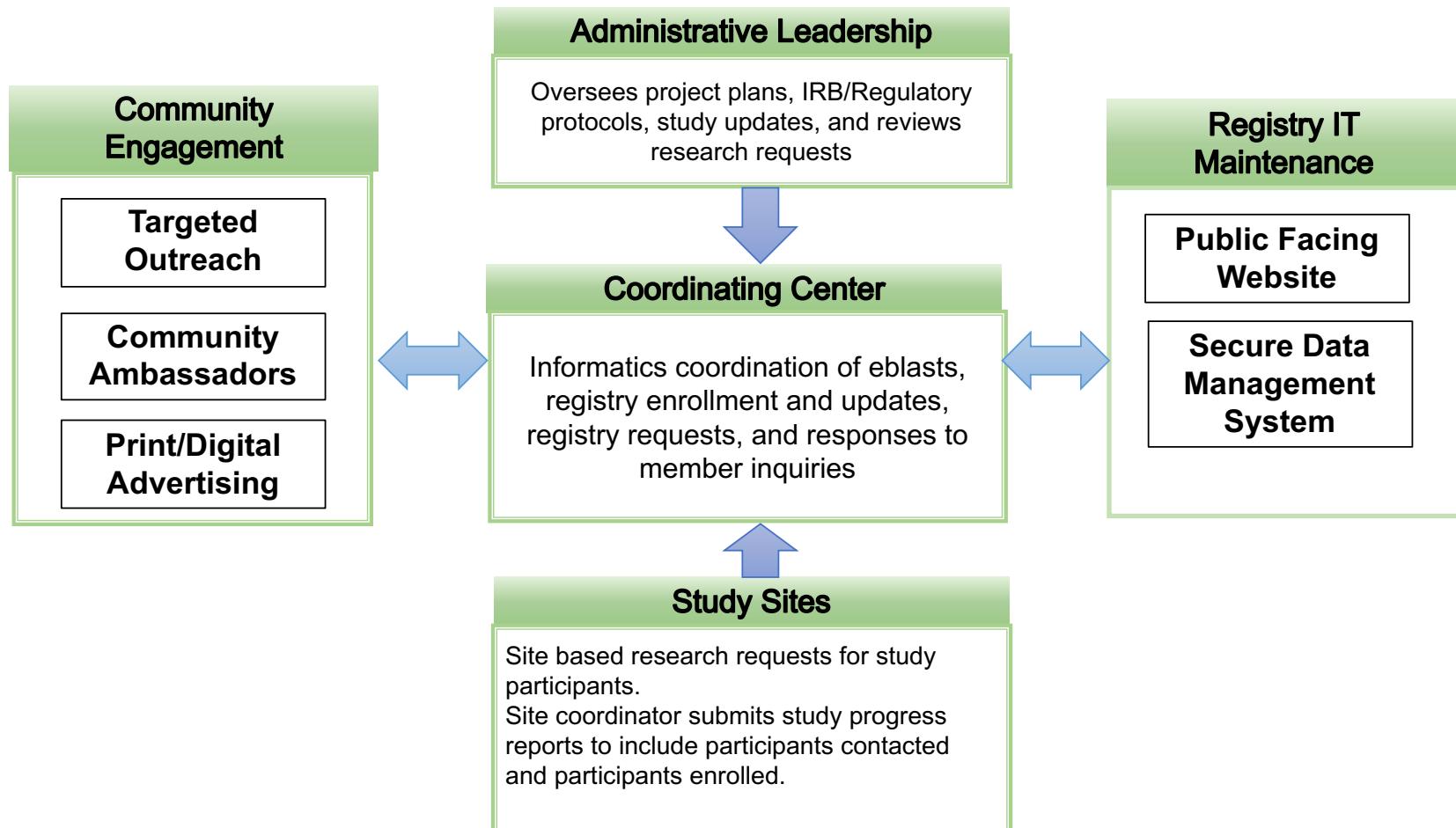
## Accuracy versus Color and Shape



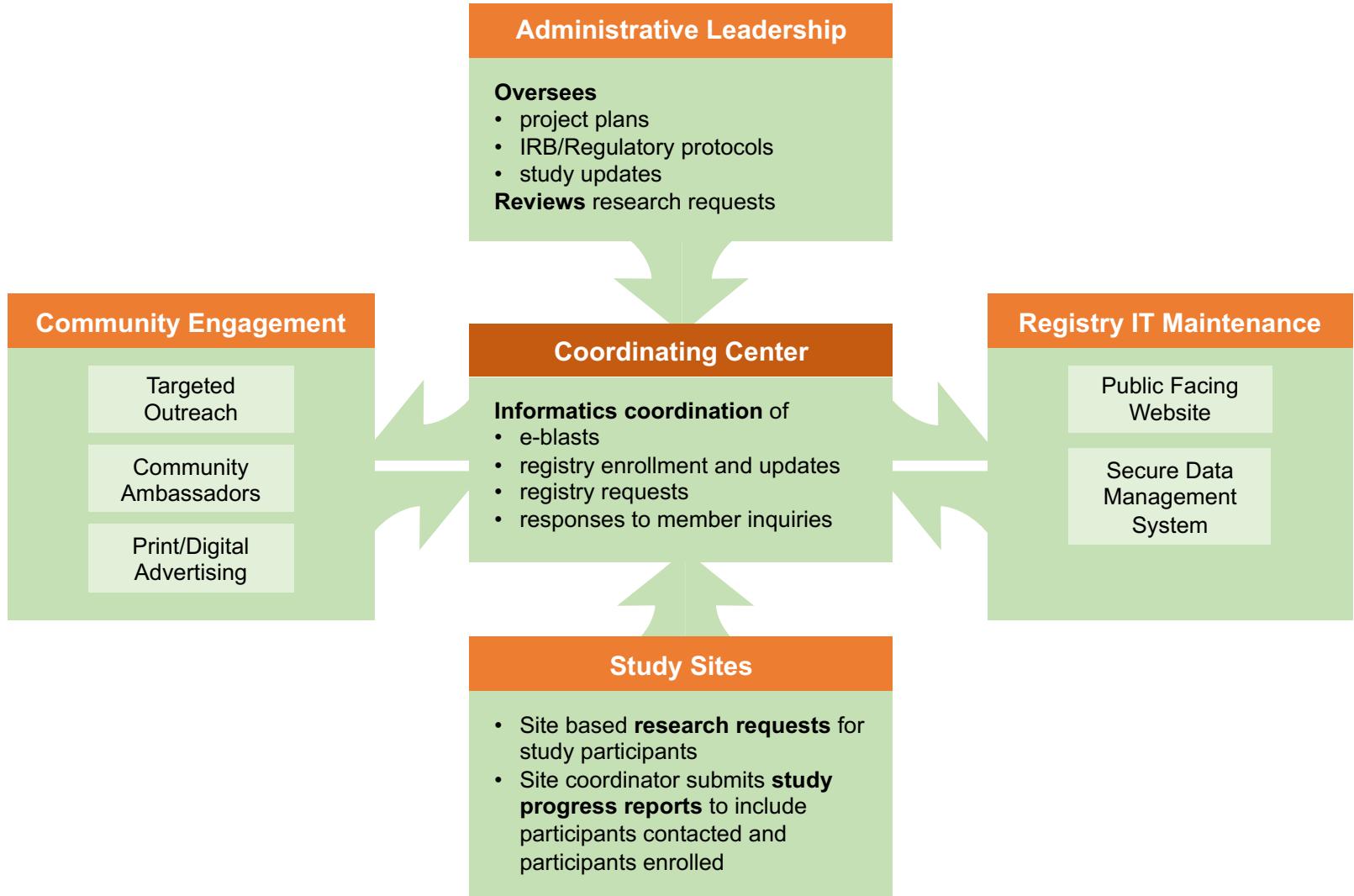
## Accuracy Improved by Color, not Shape



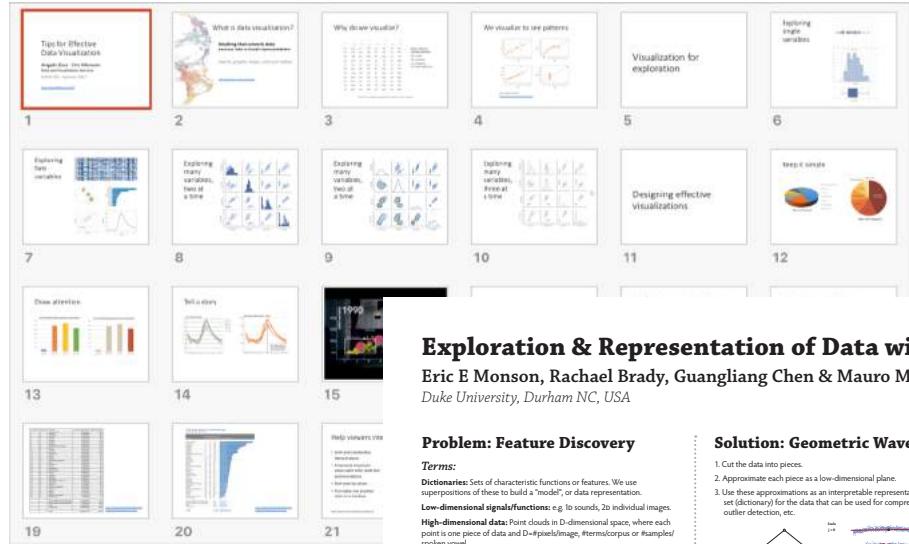
# Remove distractors & customize



# Remove distractors & customize



# With posters, presentations, and infographics, tell your story more like a comic book than an academic paper



## Exploration & Representation of Data with Geometric Wavelets

Eric E Monson, Rachael Brady, Guangliang Chen & Mauro Maggioni  
Duke University, Durham NC, USA

### Problem: Feature Discovery

#### Terms:

**Dictionaries:** Sets of characteristic functions or features. We use superpositions of them to build a "model", or data representation.

**Low-dimensional signals/functions:** e.g. 10 sounds, 20 individual images.

**High-dimensional data:** Point clouds in D-dimensional space, where each point is one piece of data and D={pixelsImage, Thermocouple or Sample/ spoken vowel}.

#### Challenge:

For low-dimensional signals we have many different "general-purpose" dictionaries (Fourier basis, Wavelets, Curvelets) to model our data. For high-dimensional data, de-noising, sharpening, etc.

For high-dimensional data, we would like to do the same things, but we do not have good, tractable models, so we must find features from the data itself.

Given the data ( $X$ ) we want to learn a set of features ( $\Theta$ ) and coefficients ( $\alpha$ ) to build a representation such that

$$X \approx \Theta \cdot \alpha$$

• Is it "good" if  $\alpha$  is sparse (lots of zeros or very small values).

Most methods are "black boxes" which have no guarantees, are costly to compute and don't yield interpretable data features.

#### Our Approach:

We do not try to solve this problem "in general", but exploit the fact that often real data has lower-dimensional geometric structure, such as lying near a manifold ( $M$ ) or being a "skew

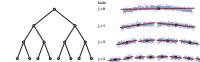
Geometric wavelets allow construction which discovers features in high-dimensional data under these geometric assumptions.

It is explicit, which leads to **interpretable features**, and it comes with guarantees (as a function of approximation error parameter) on computational cost, number of elements in the dictionary and sparsity of representation.

The approach is non-linear, but piecewise linear, or, in a fact, but can adapt to arbitrary non-linear manifolds.

### Solution: Geometric Wavelets

1. Cut the data into pieces.
2. Approximate each piece as a low-dimensional plane.
3. Use these approximations as an interpretable representation or feature set (dictionary) for the data that can be used for compression, filtering, outlier detection, etc.



#### More Details:

- Let's use a simple measure to create a graph from the data points. Construct a set of multi-scale partitions of  $M$  by using recursive spectral partitioning (METIS uses Laplacian eigenvalues). In the graph – we are also implementing Cover Tree. This step often involves scale-specific methods:

2. Compute the SVD of the data covariance for each piece. This gives Scaling Functions &  $M_{\text{sf}}$  – a manifold approximation at scale  $j$  for piece  $k$  – a projection onto that local approximate tangent space.

3. Higher-scale pieces are obtained by merging the nodes of the parent. Efficiently encode the differences between  $M_{\text{sf}} \rightarrow M$  by constructing Wavelet spanning spaces & "detail" operators analogous to Wavelet Theory.



This results in a multi-scale nonlinear transform mapping data to a family of pieces of planes which approximates the original data to any given precision.



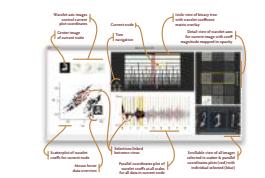
### Experience: Interactive GUI

- Quickly see much more data so mathematicians can evaluate methods during development.

- Begin developing a platform onto which we can easily add new applications for new tasks and data types.

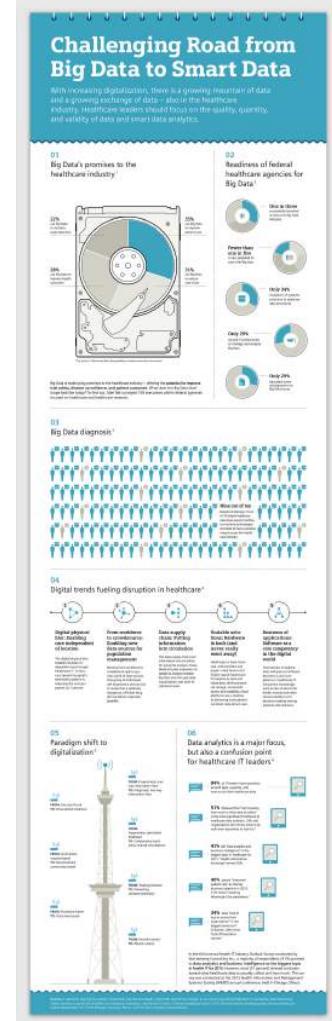
- Help explain Geometric Wavelets and gain intuition about the representation.

Implemented in Python, using PyQt to glue together custom VTK views. Currently the representation is not computed in the GUI – Matlab output is loaded from files.



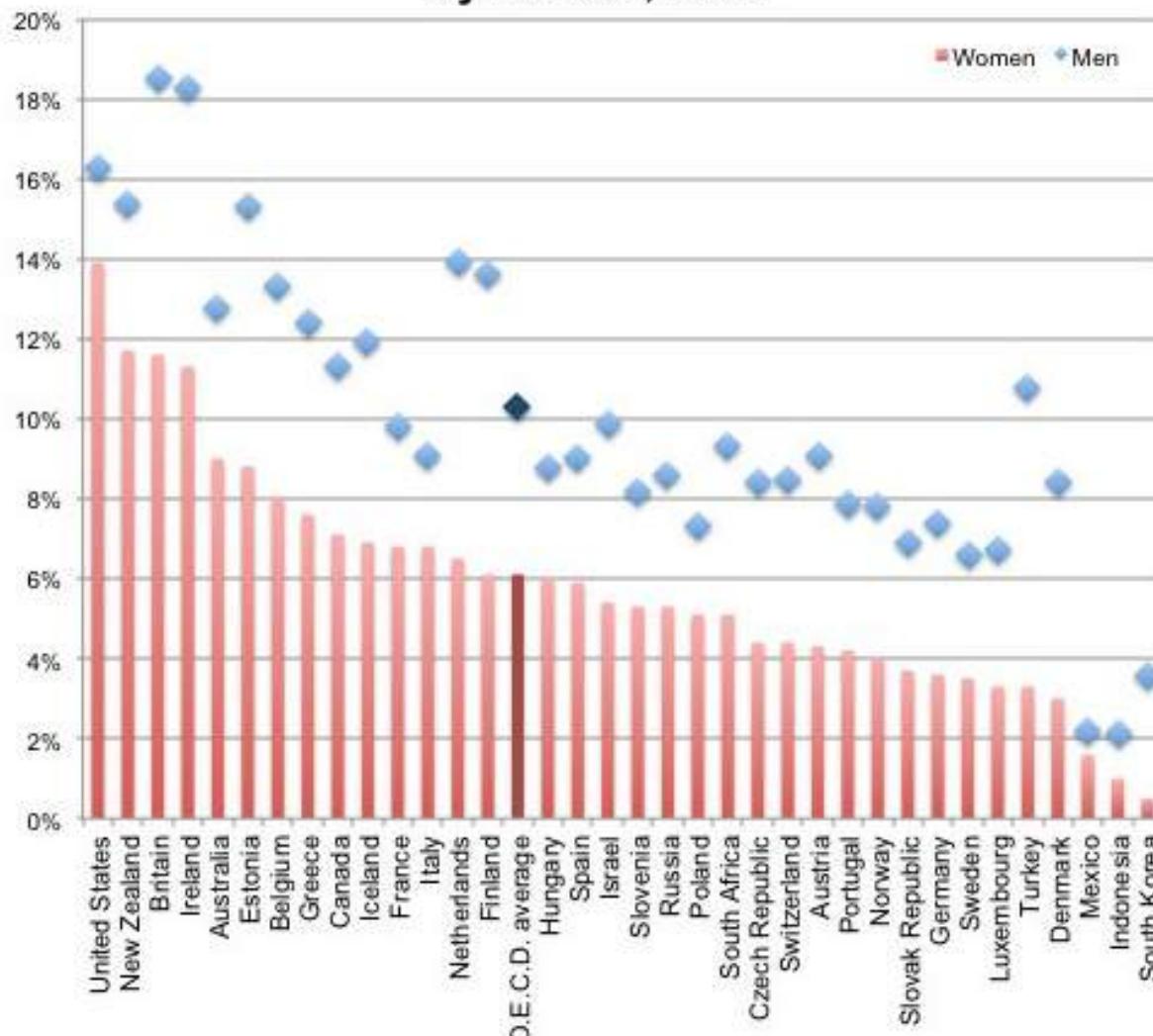
### Observations & Future Directions:

- Coarser scales contain general applications of the data, such as interpretable node centers and wavelet directions.
- Finer scales reveal anomalous data through extreme wavelet coefficients and wavelet boundaries.
- Coarser wavelets contain information which could be good for classification tasks, but finer-scale wavelets encode more specific features which cluster and characterize individuals.
- Developers have changed their ideas about data encoding after viewing their results in the GUI.
- We are already working on variable dimensionality, group definition & labeling for classification and outlier detection, views for new data types.



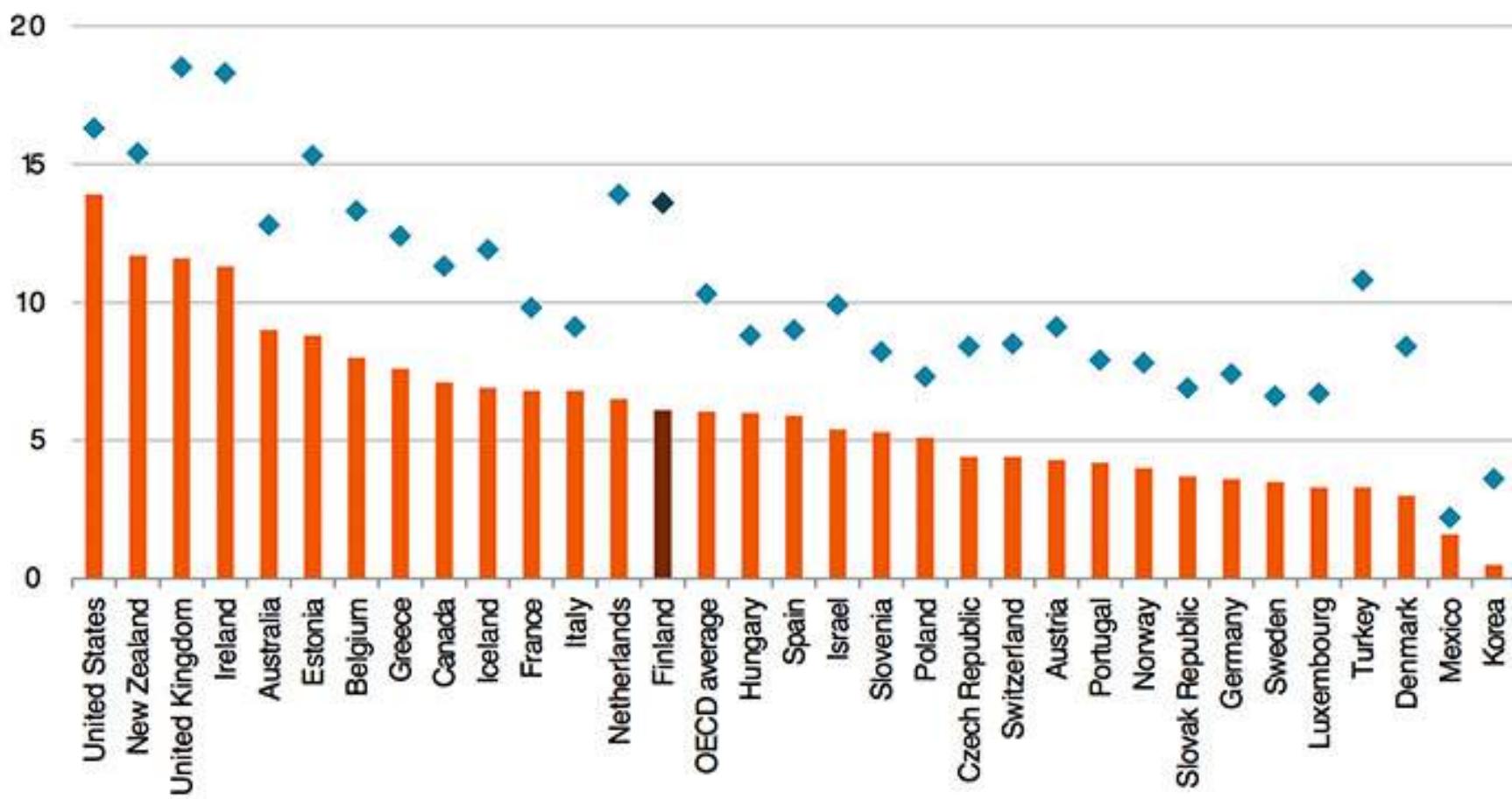
# Figure critique & reworks

## Percentage of Employed Who Are Senior Managers, by Gender, 2008



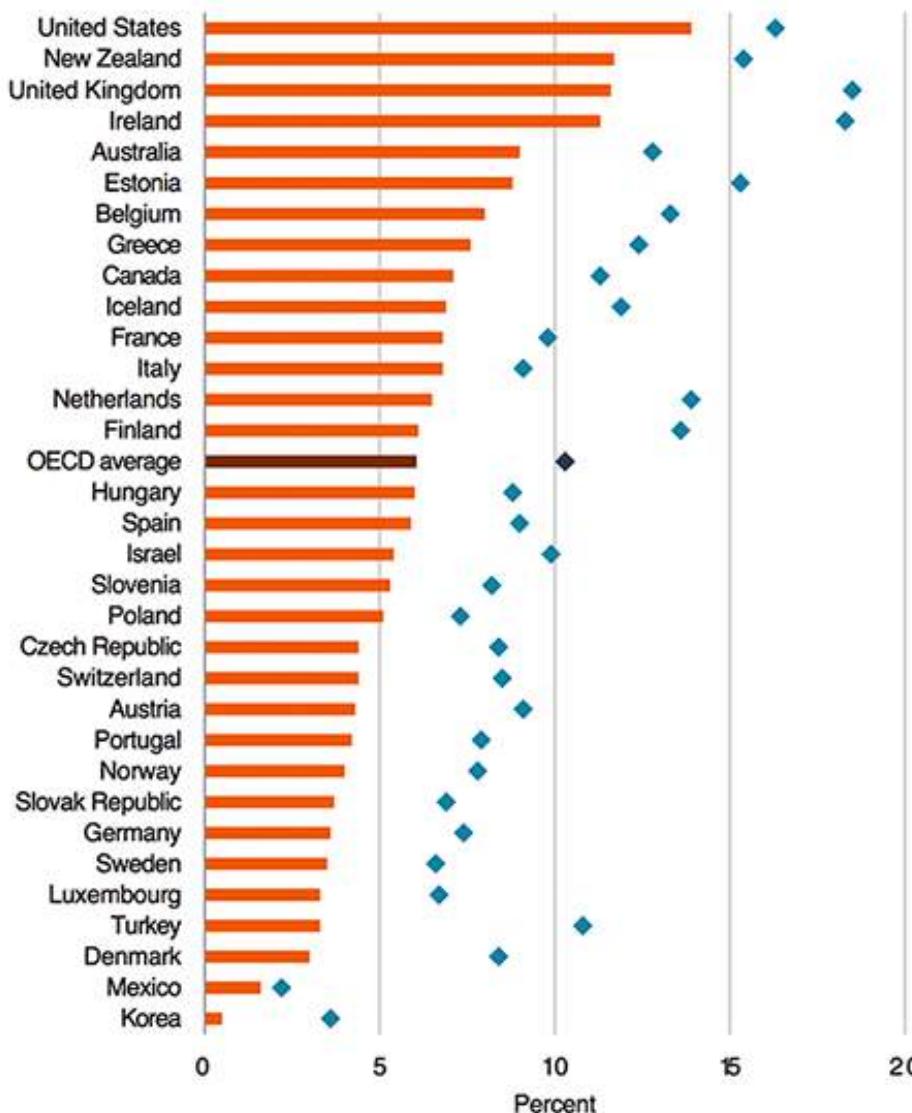
## Percentage of Employed Who are Senior Managers, by Gender, 2008

(Percent)    □ Women    □ Men

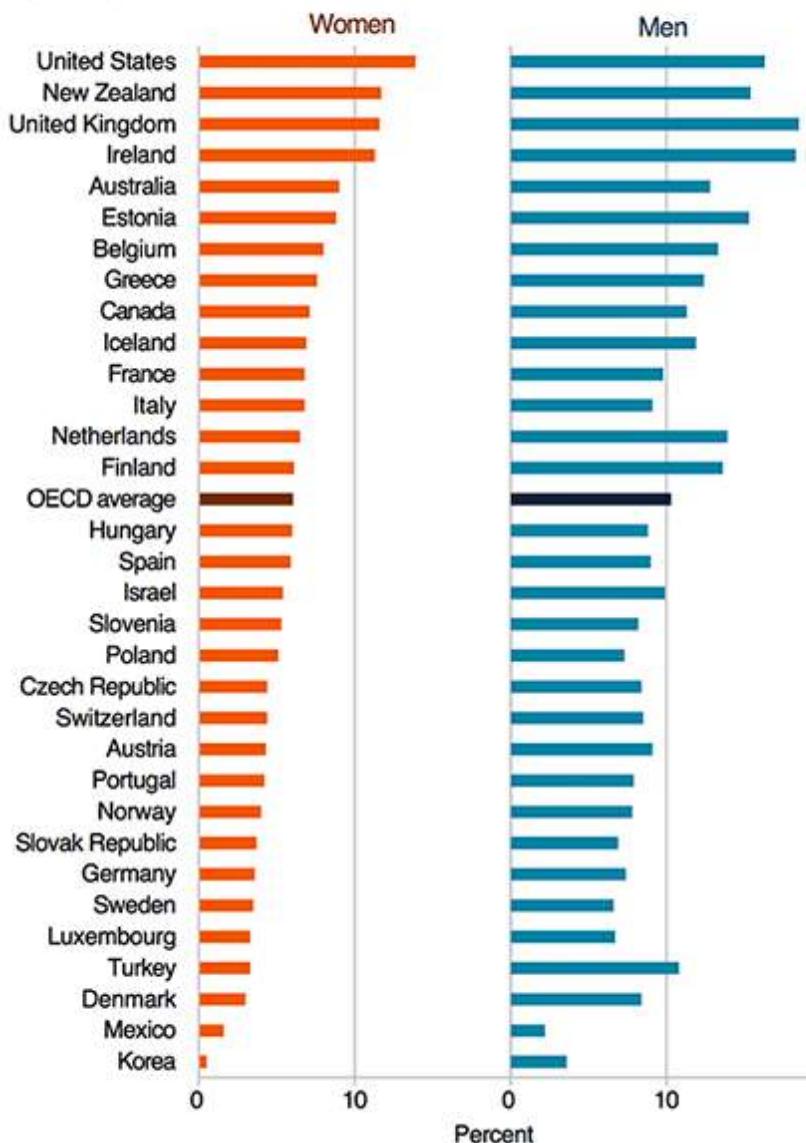


Percentage of Employed Who are Senior Managers,  
by Gender, 2008

(Percent)    ■ Women    ♦ Men

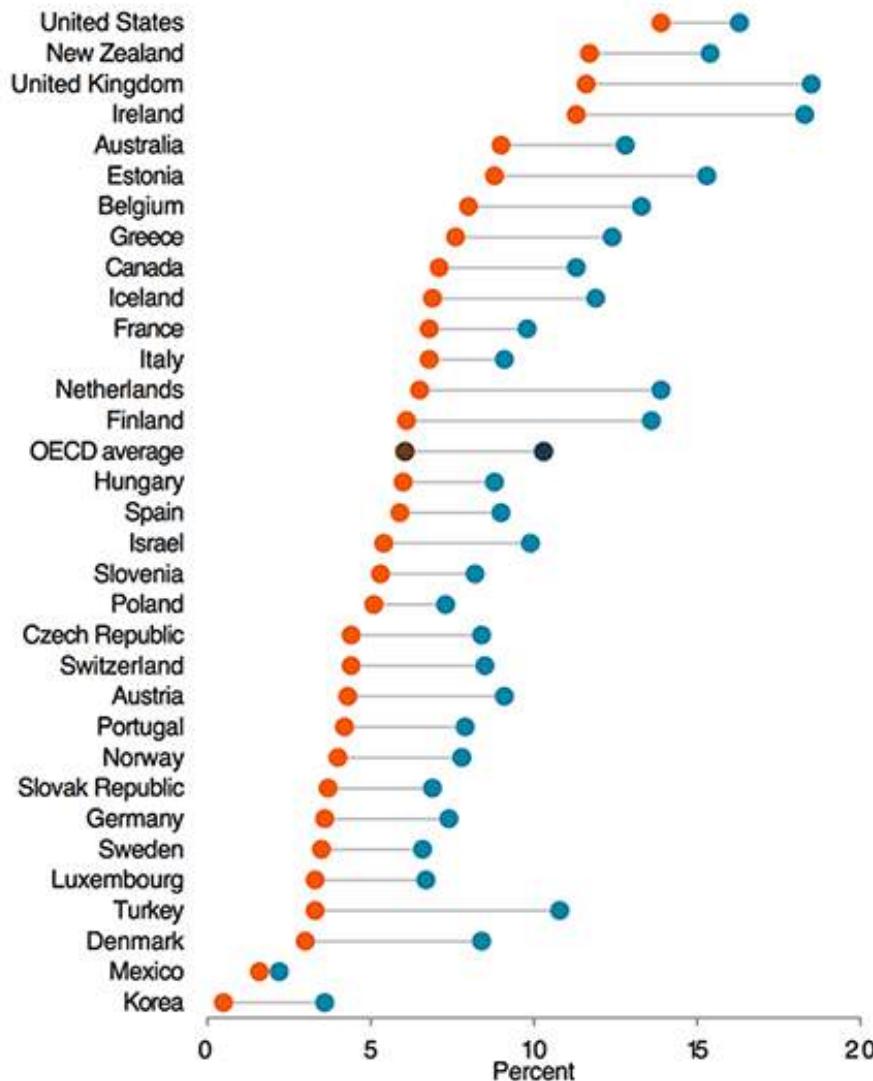


Percentage of Employed Who are Senior Managers,  
by Gender, 2008  
(Percent)

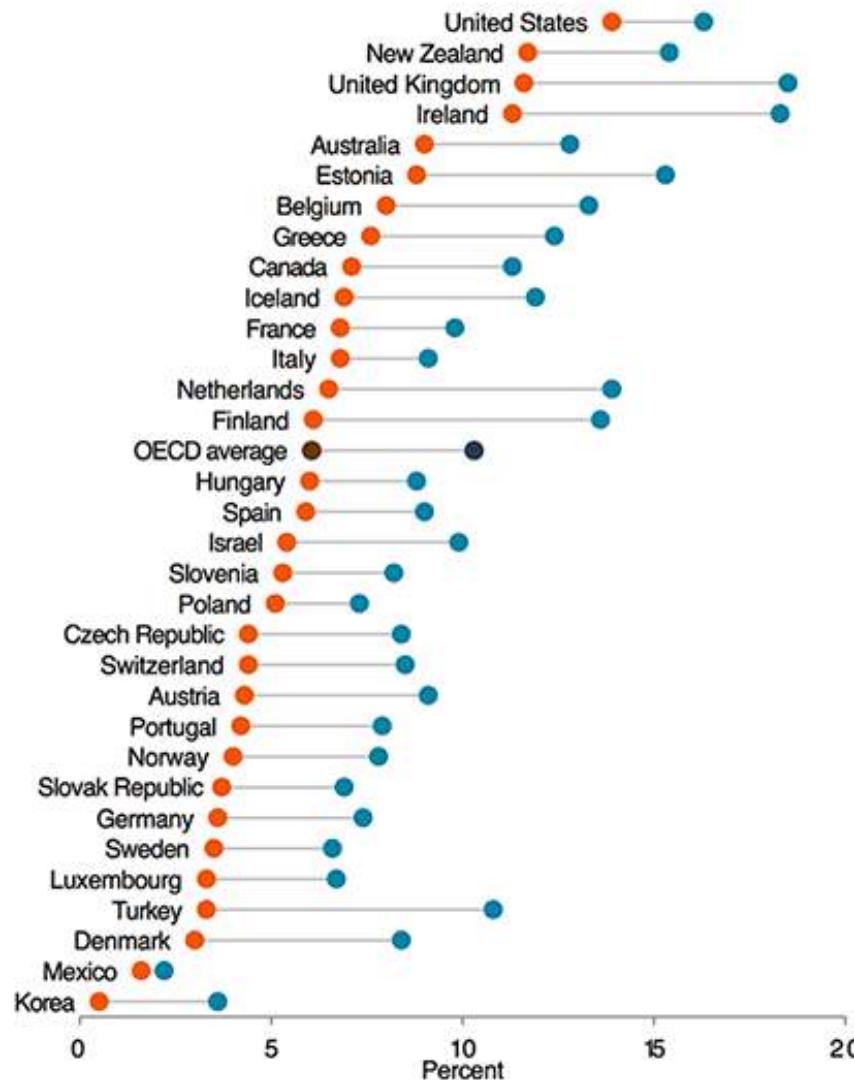


Jon Schwabish: <http://thewhyaxis.info/gap-remake/>

Percentage of Employed Who are Senior Managers,  
by Gender, 2008  
(Percent)    ● Women    ● Men



Percentage of Employed Who are Senior Managers,  
by Gender, 2008  
(Percent)    ● Women    ● Men



# Data and Visualization Services



<http://library.duke.edu/data>  
askdata@duke.edu

# Support Areas



Data Sources



Data Transformation



Mapping and GIS



Data Management



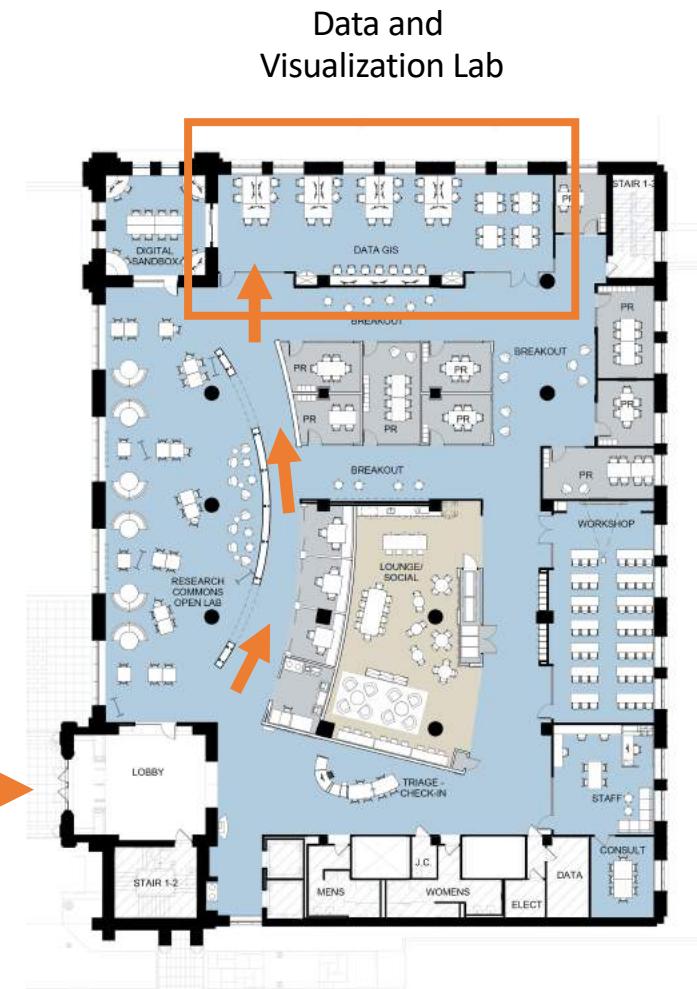
Data Visualization

# Brandaleone Family Lab for Data and Visualization Services

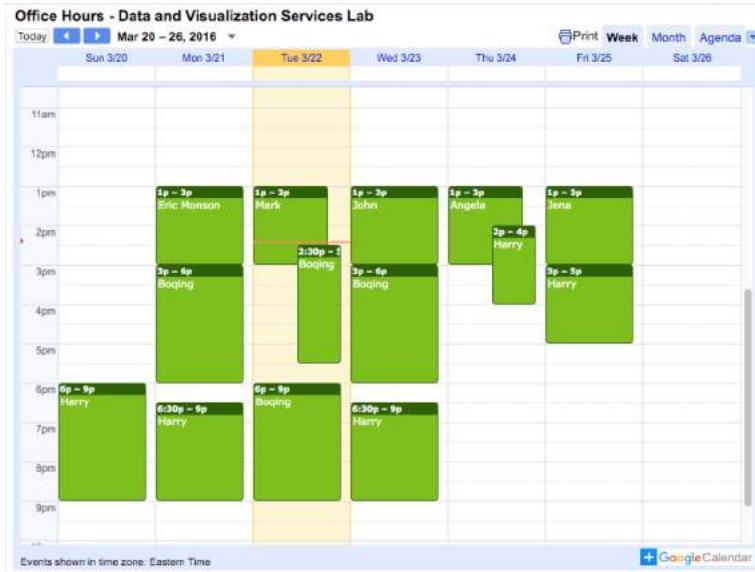
<http://library.duke.edu/data/about/lab>

- **The Edge** (*1st floor of Bostock*)
- Open whenever the library is open
- 12 high-powered Dell workstations
- 3 Bloomberg financial workstations
- Software for data analysis, GIS, and visualization

Entrance to  
Bostock



# Consulting

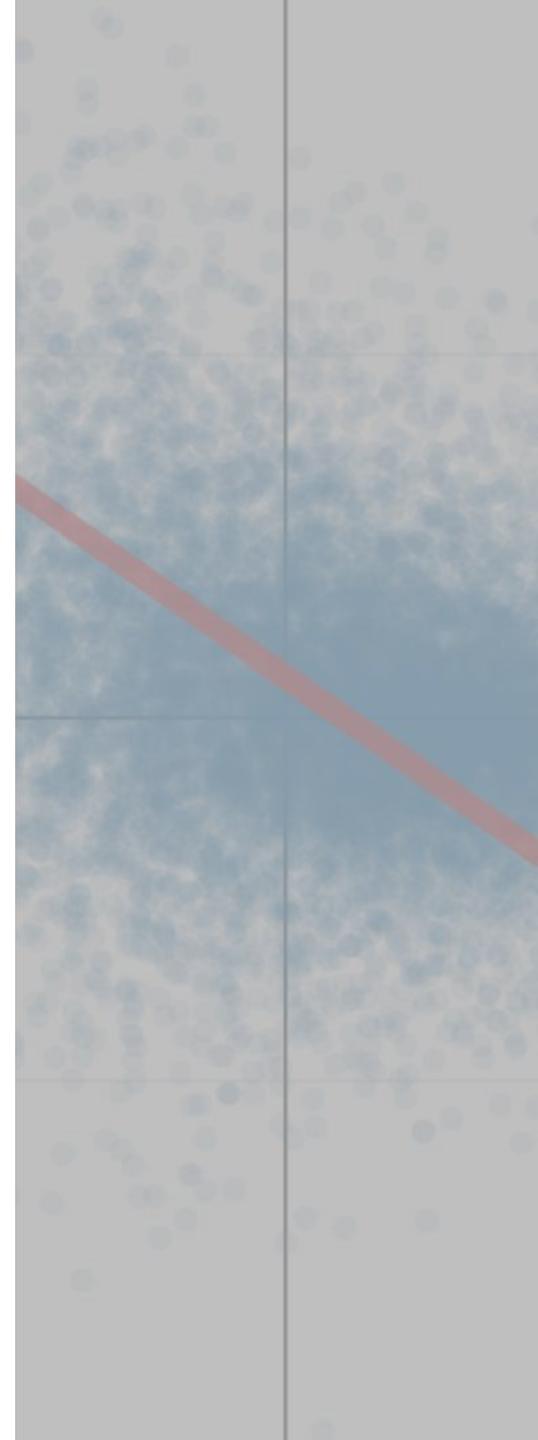


<http://library.duke.edu/data/about/schedule>

...or email  
[askdata@duke.edu](mailto:askdata@duke.edu)  
for an appointment

# Types of visualization consulting

- Look at data and brainstorm about the best visualization
- Recommend appropriate tools
- Troubleshoot software problems
- Help with cleaning and structuring data
- Offer graphic design advice for figures, diagrams, slides and posters



# Workshops

<http://library.duke.edu/data/workshops>

- Typically toward the beginning of the semester
- Covering: visualization, data processing, GIS/mapping
- 1-2 hours, often hands-on



For announcements, sign up for our listserv:  
<https://lists.duke.edu/sympa/subscribe/dvs-announce>

# Spring 2019 visualization workshops

- Intro to ArcGIS Pro · Jan 24
- Rfun: Visualization in R using ggplot2 · Jan 25
- Rfun: Mapping with R · Jan 29
- Web Mapping · Jan 31
- Intro to QGIS · Feb 6
- Intro to Tableau – Easy Charts and Maps · Feb 7
- Introduction to GIS · Feb 12
- Story Maps · Feb 13
- Visualization in Python with Altair · Feb 19
- Map Design · Feb 21
- Rfun: Interactive Dashboards Visualizations with R · Feb 26
- Adobe Illustrator for Modifying Charts and Graphs · Mar 7

# Videos of past workshops

<http://bit.ly/DVSvideos>

The screenshot shows a Panopto video player interface. At the top, it displays the title "Panopto Figures and Posters" and the date "March 4, 2016 in DVS Training". On the left, there's a thumbnail image of two people standing in front of a whiteboard in a classroom setting. Below the thumbnail is a search bar labeled "Search this recording" and a "Discussion" button. The main content area features the title "Designing Academic Figures and Posters" in large bold letters, followed by the date "March 4, 2016" and a link to "Slides: <http://duke.box.com/PostersSpring2016>". Below the title, two speakers are introduced: "Angela Zoss" (Data Visualization Coordinator, Data and Visualization Services) and "Eric Monson" (Data Visualization Analyst, Data and Visualization Services). At the bottom, there's a navigation bar with a play button, a timestamp of "0:03", a volume icon, and controls for "Speed" (set to 1x), "Quality", and "Hide". There are also four small thumbnail images at the bottom: "Good Posters", "Causal Diagrams", "A poster", and "Purpose of a poster".

# Visualization Friday Forum

<https://vis.duke.edu/FridayForum>

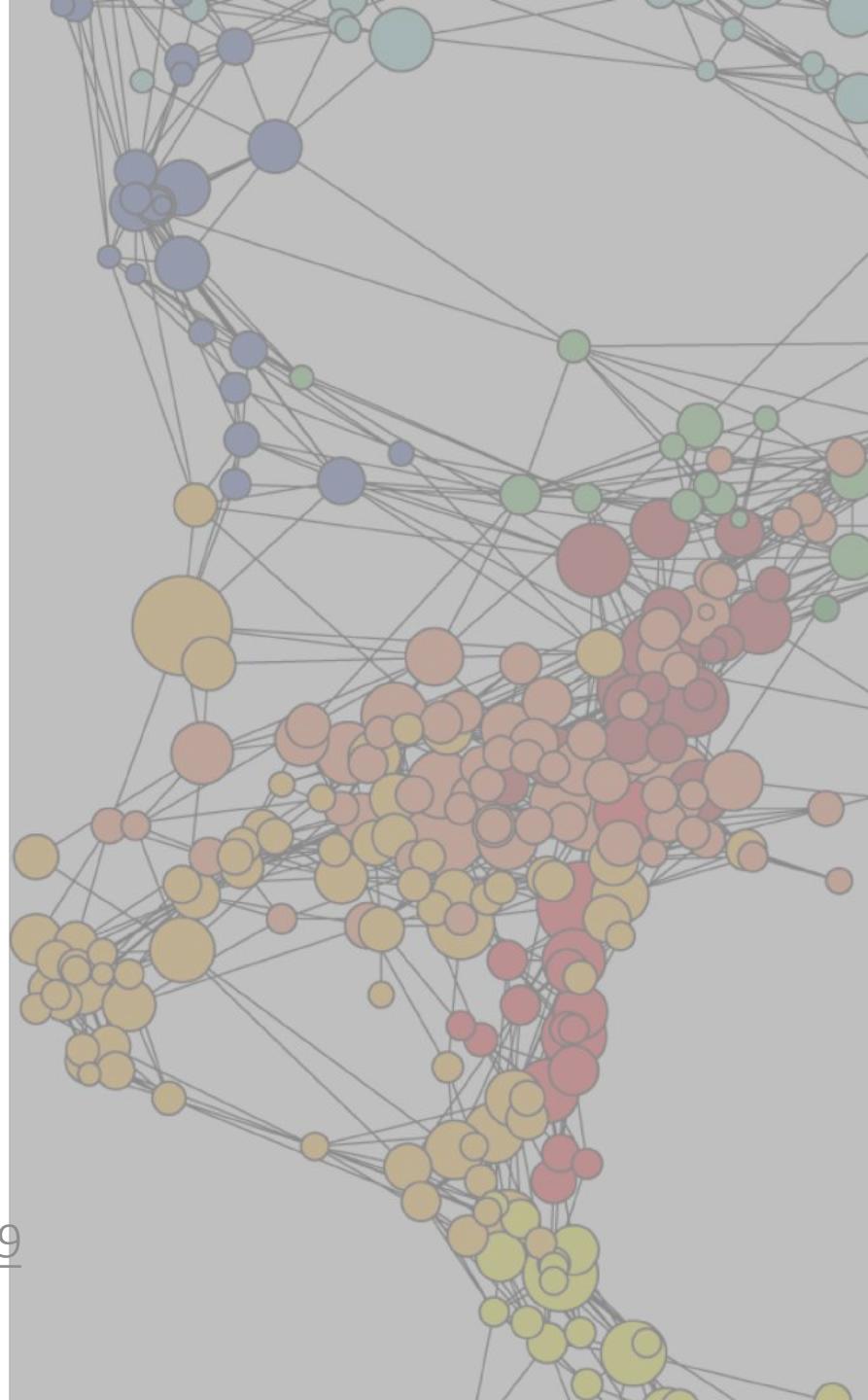
- Every Friday during the semester
- Noon in LSRC D106
- Free lunch (sssh!)
- Live streamed and recorded
- Email me to get on the mailing list



# Questions

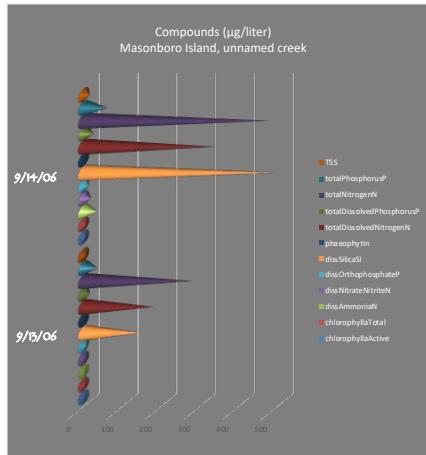
[askdata@duke.edu](mailto:askdata@duke.edu)

Slides: <https://bit.ly/STA199visSpring2019>

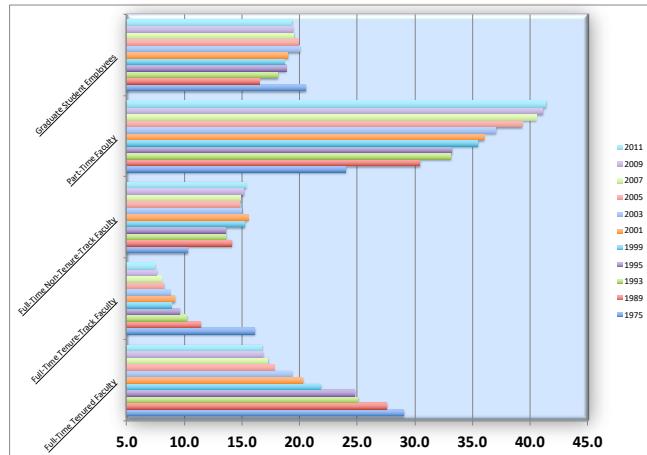


# Visualization Reworks Activity

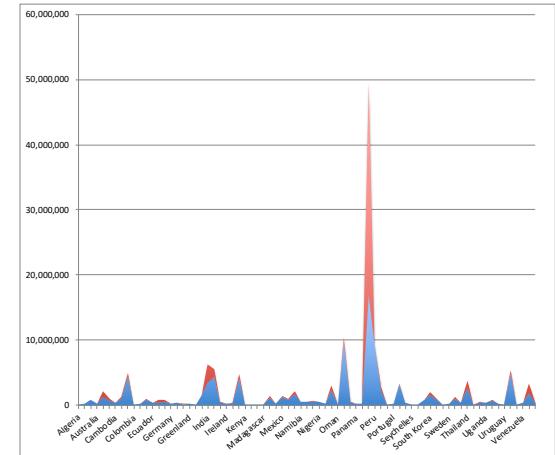
NERRS Nutrients



Instructional Staff Employment

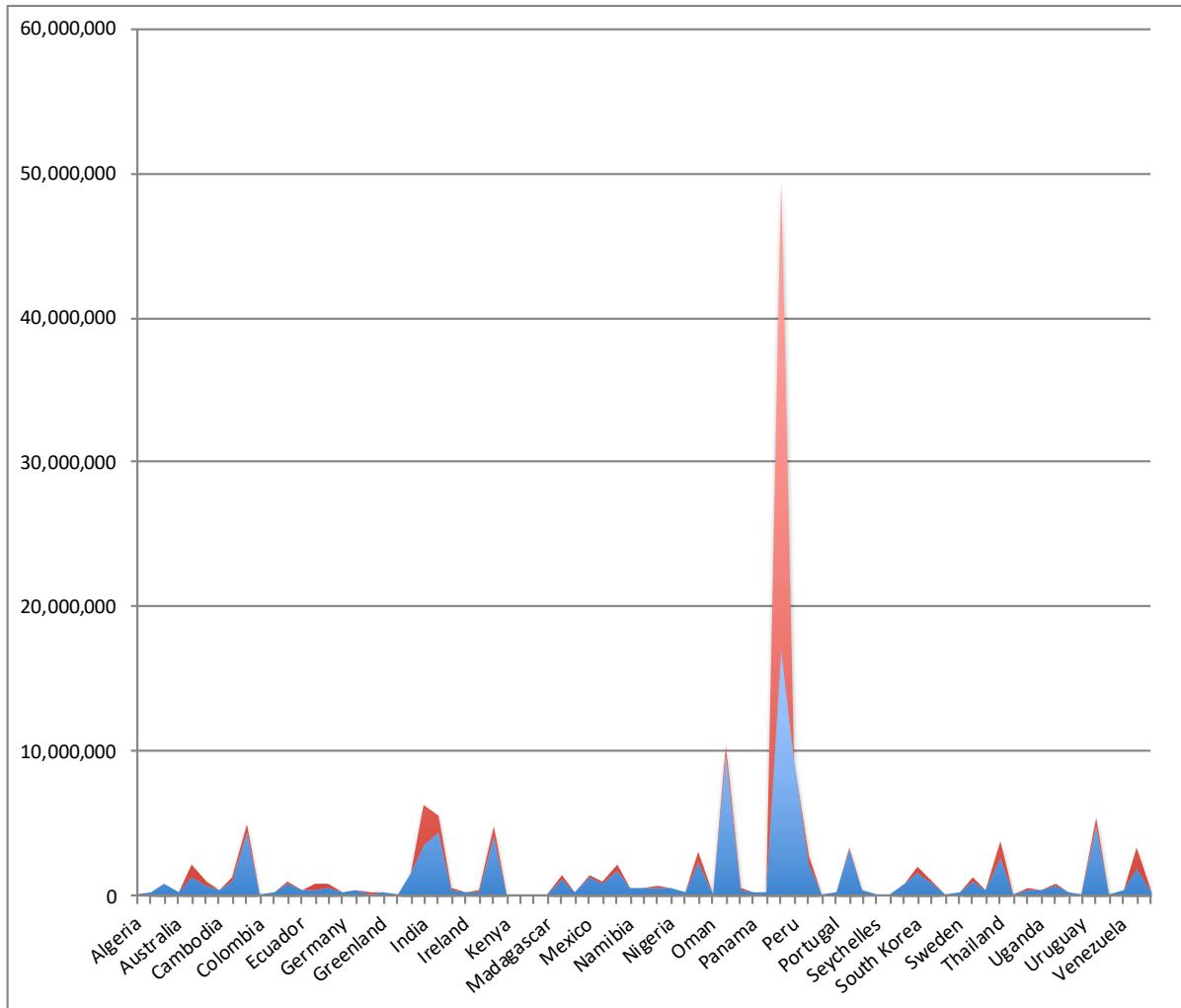


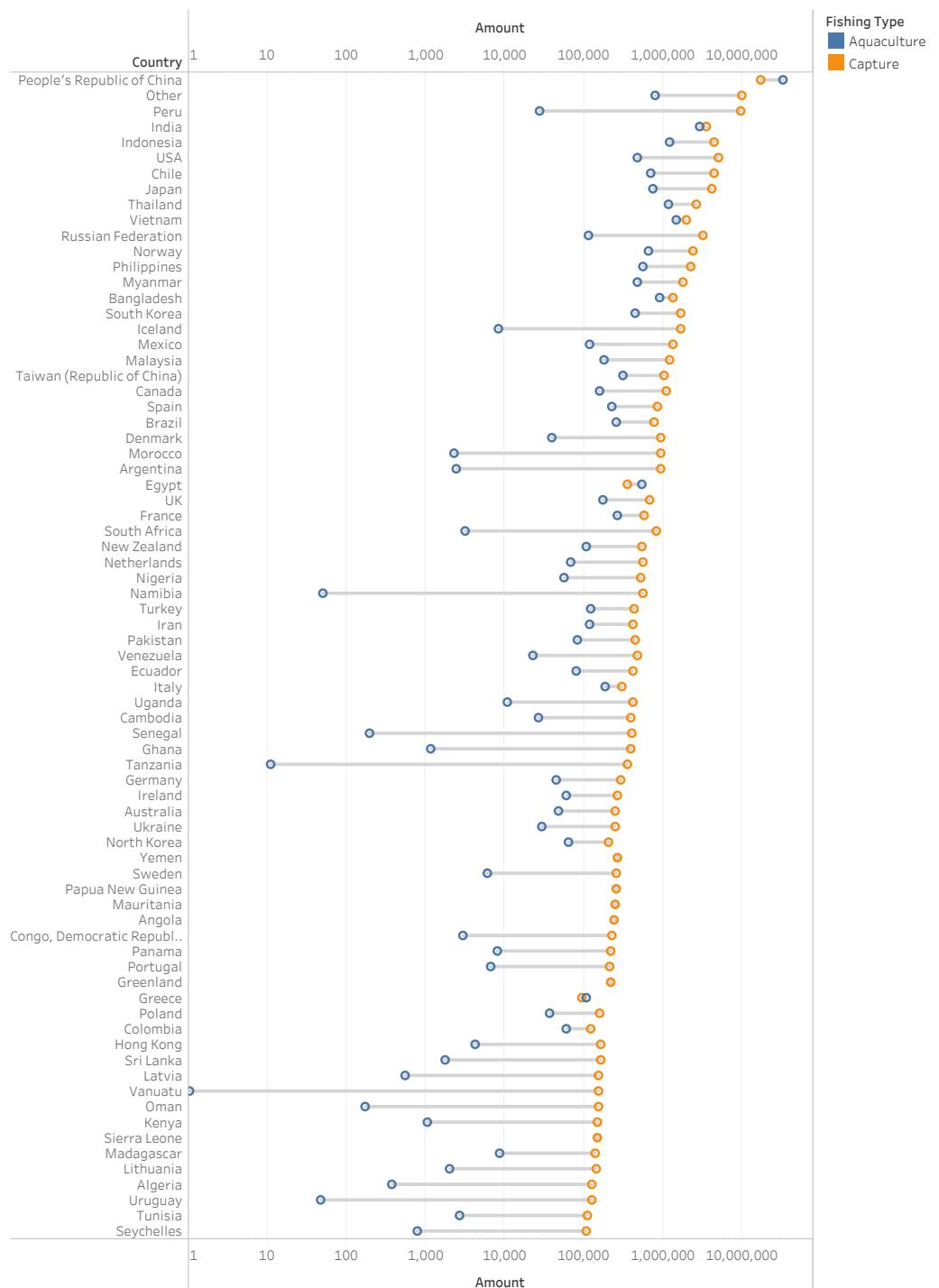
Fishing Industry

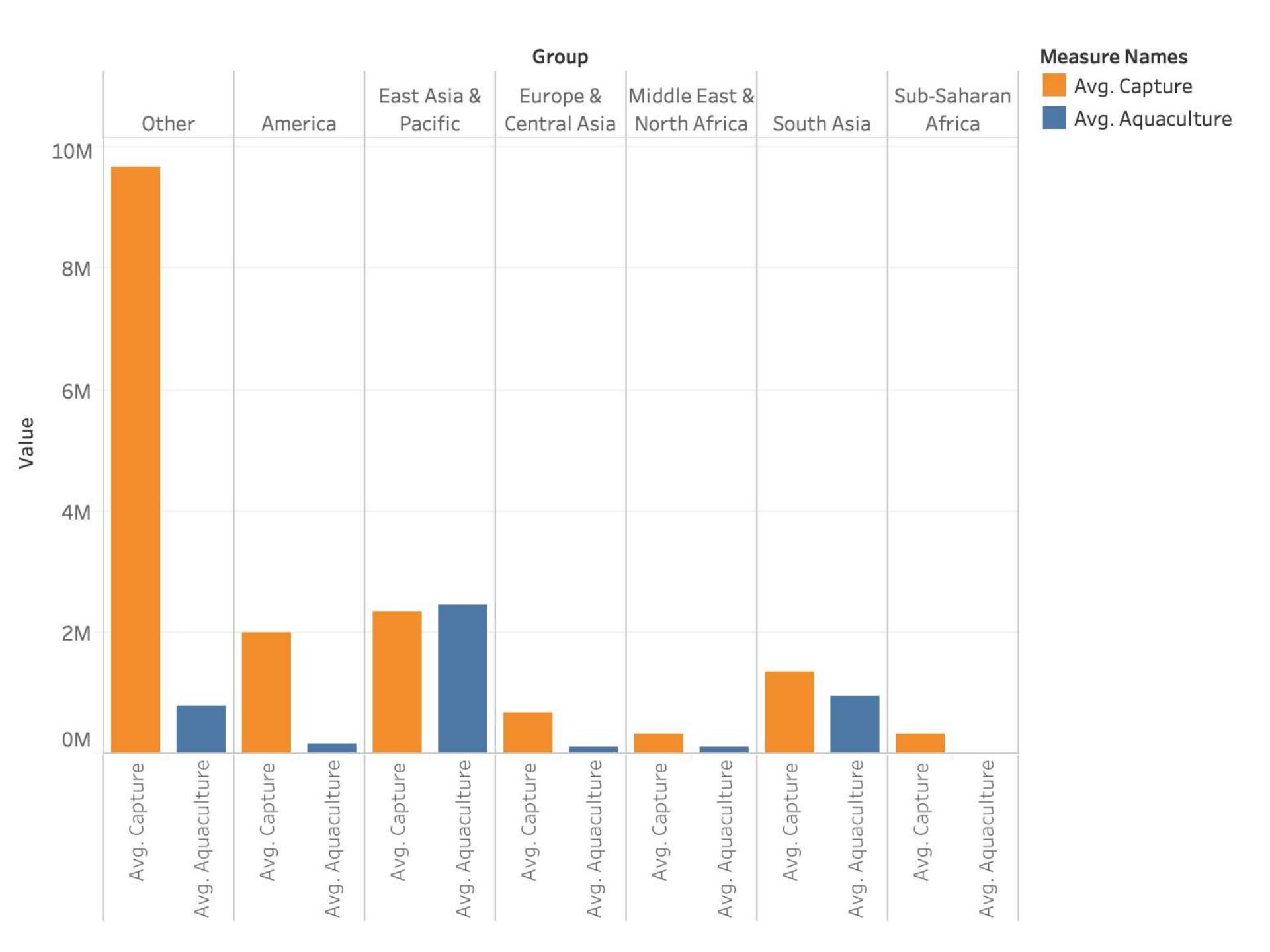


# Rework activity

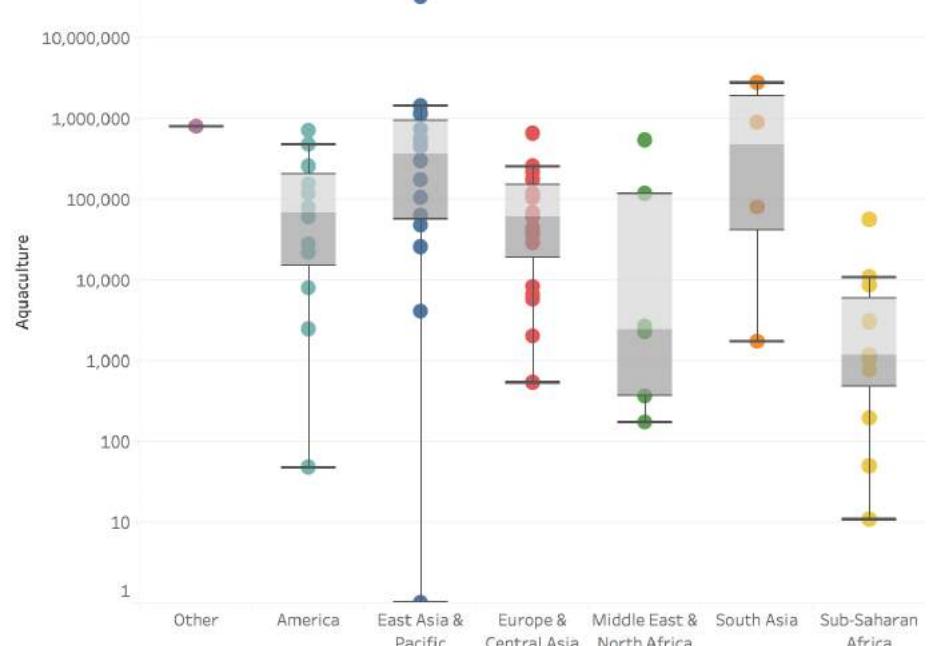
# Fishing industry by country



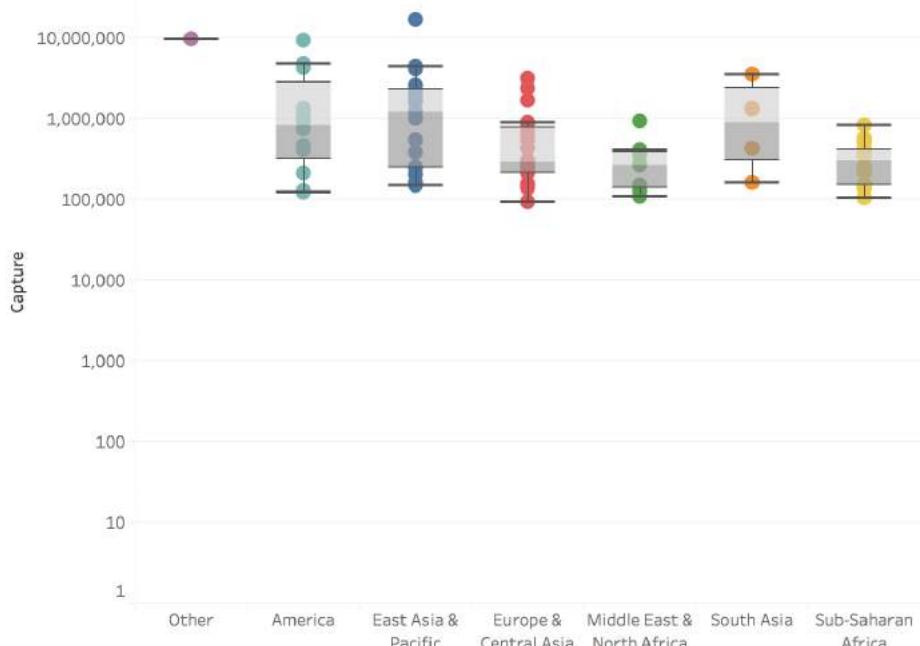


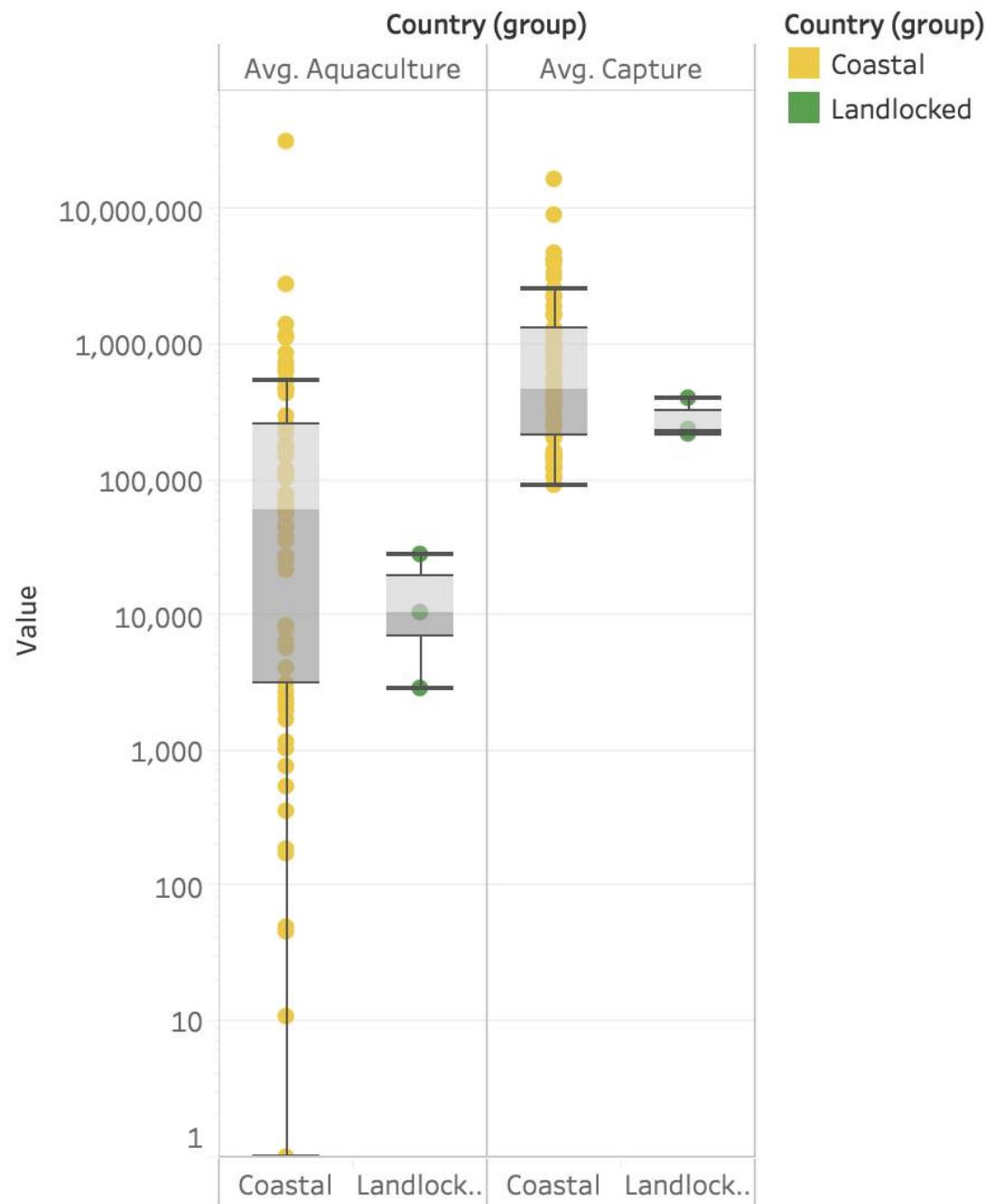


## Aquaculture



## Capture

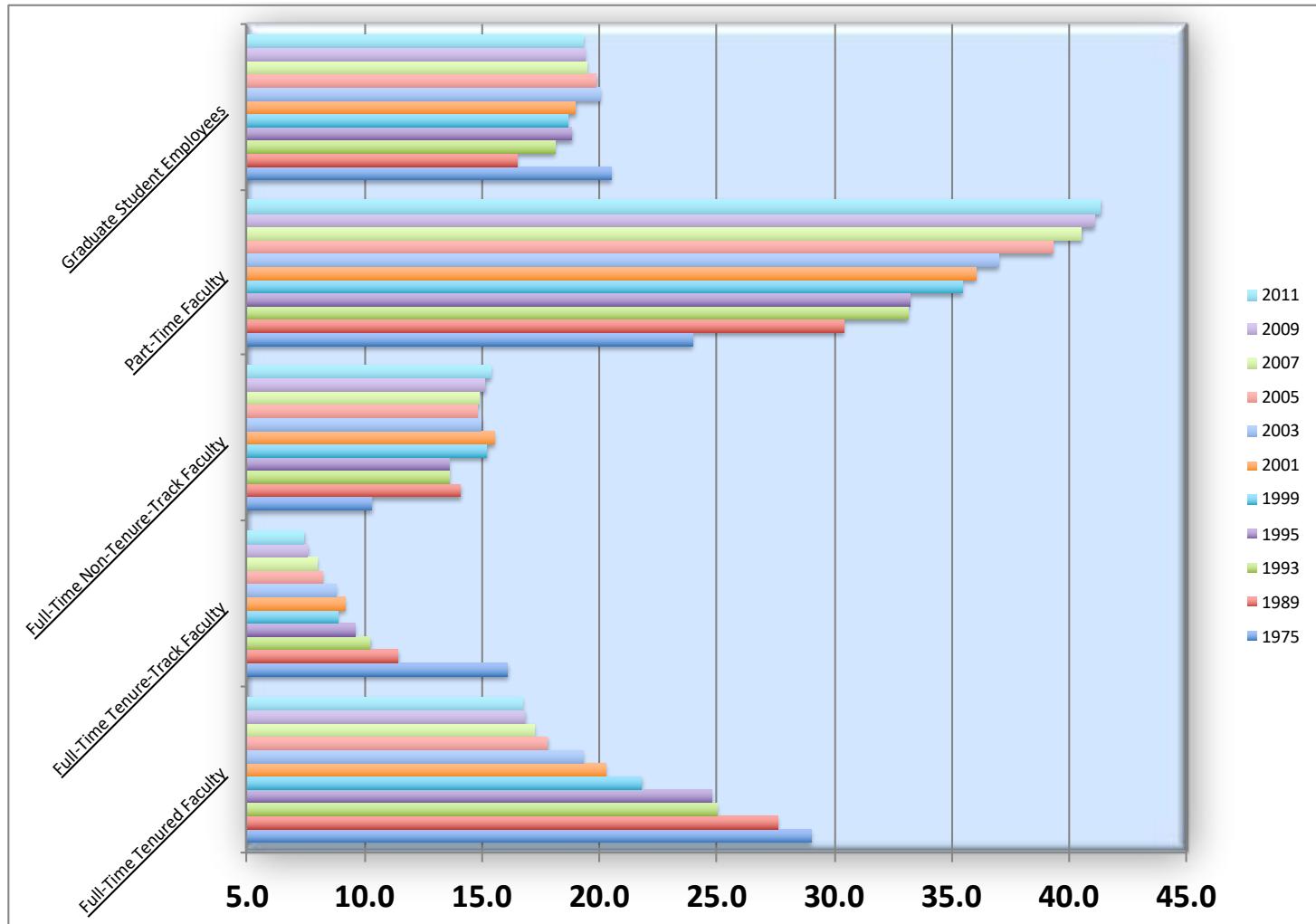


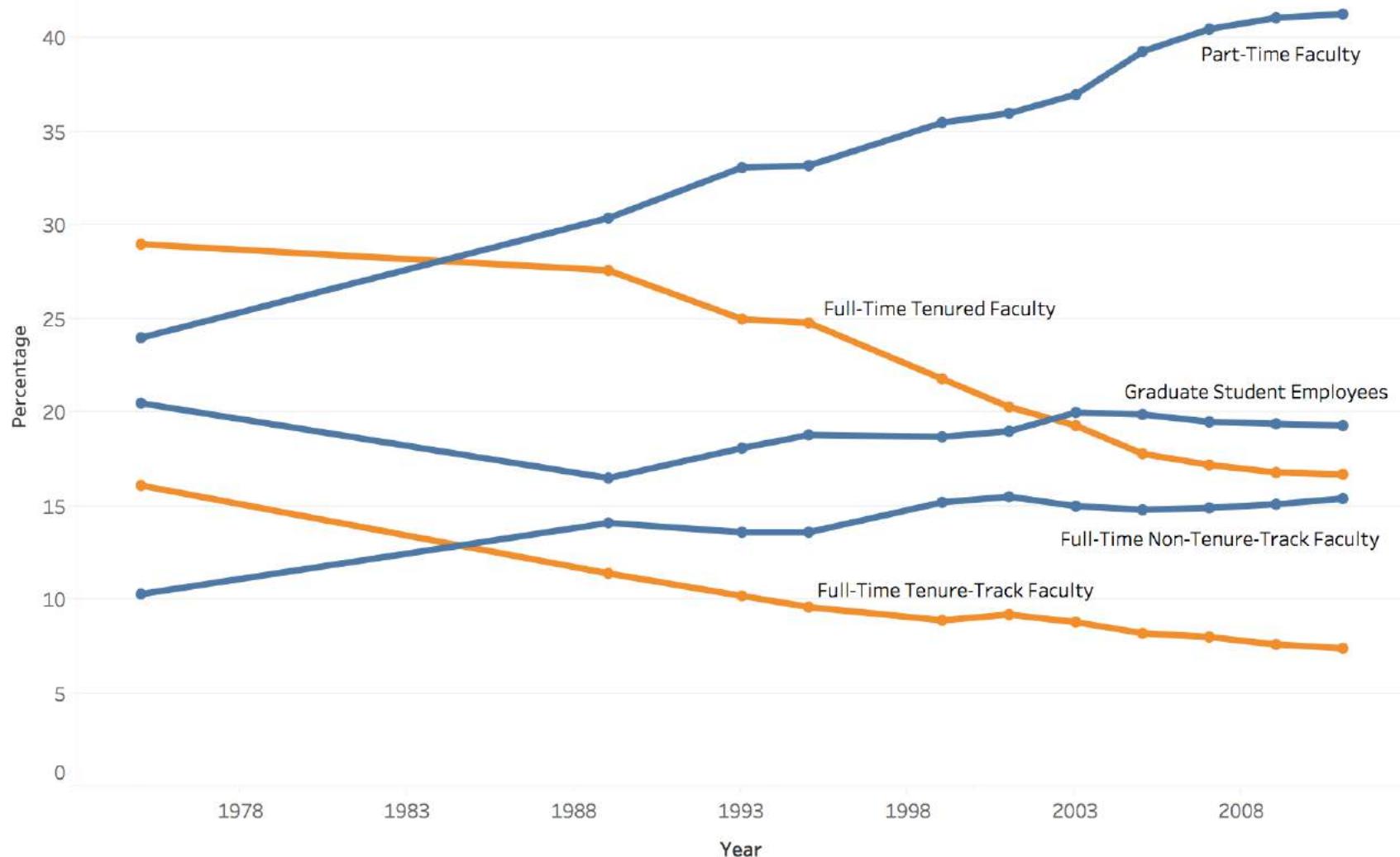


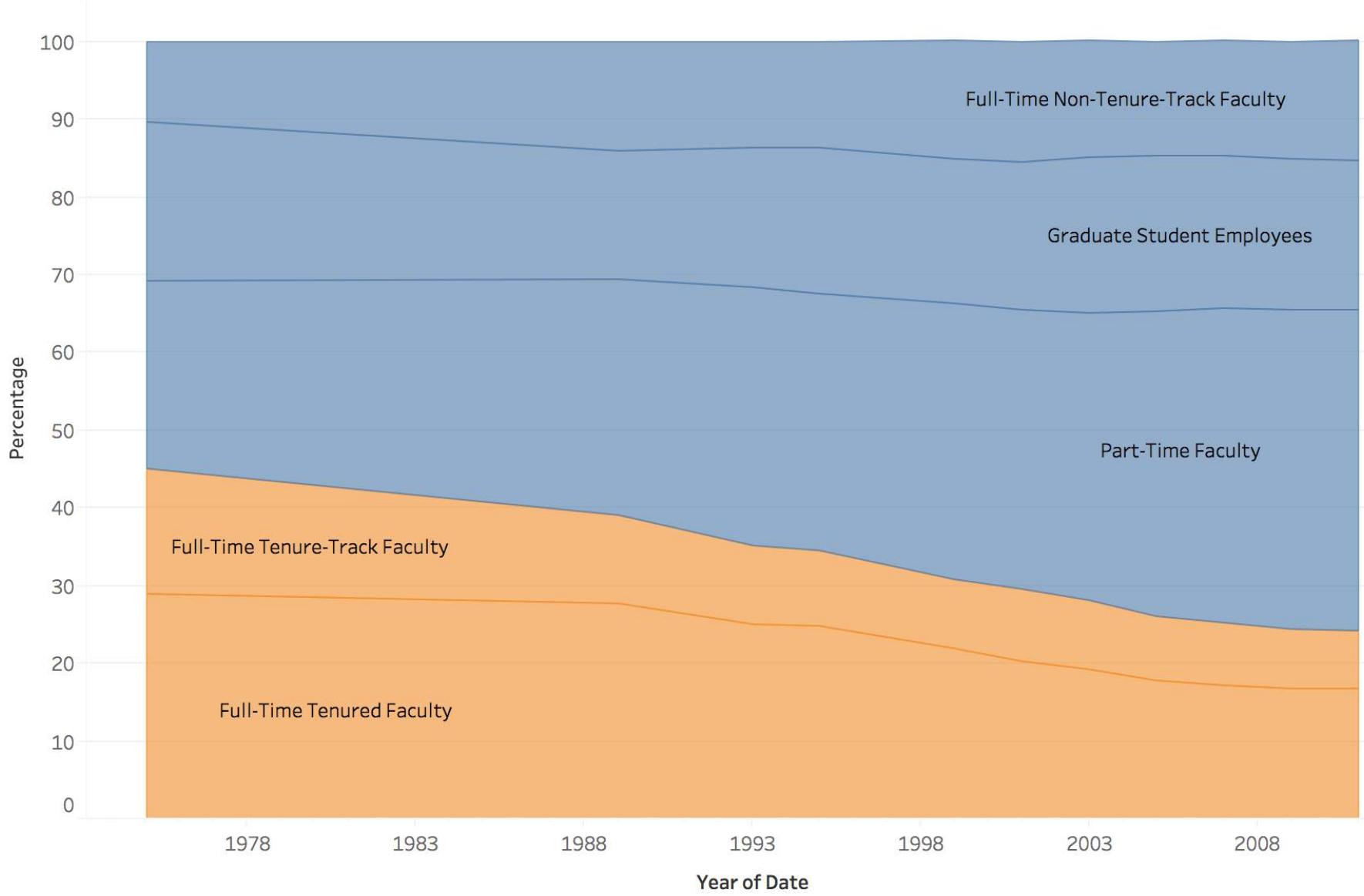
Size: Total fishing – Color: Fraction capture



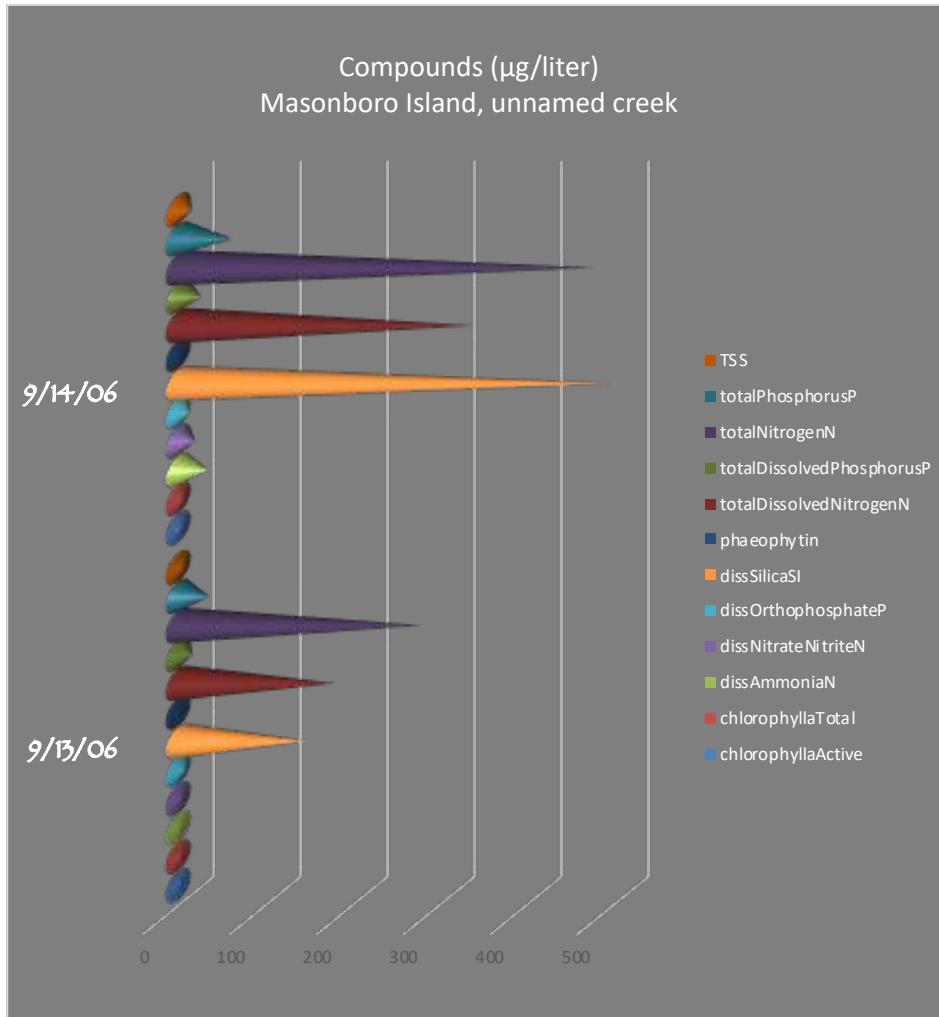
# Instructional staff trends







# NERRS nutrients



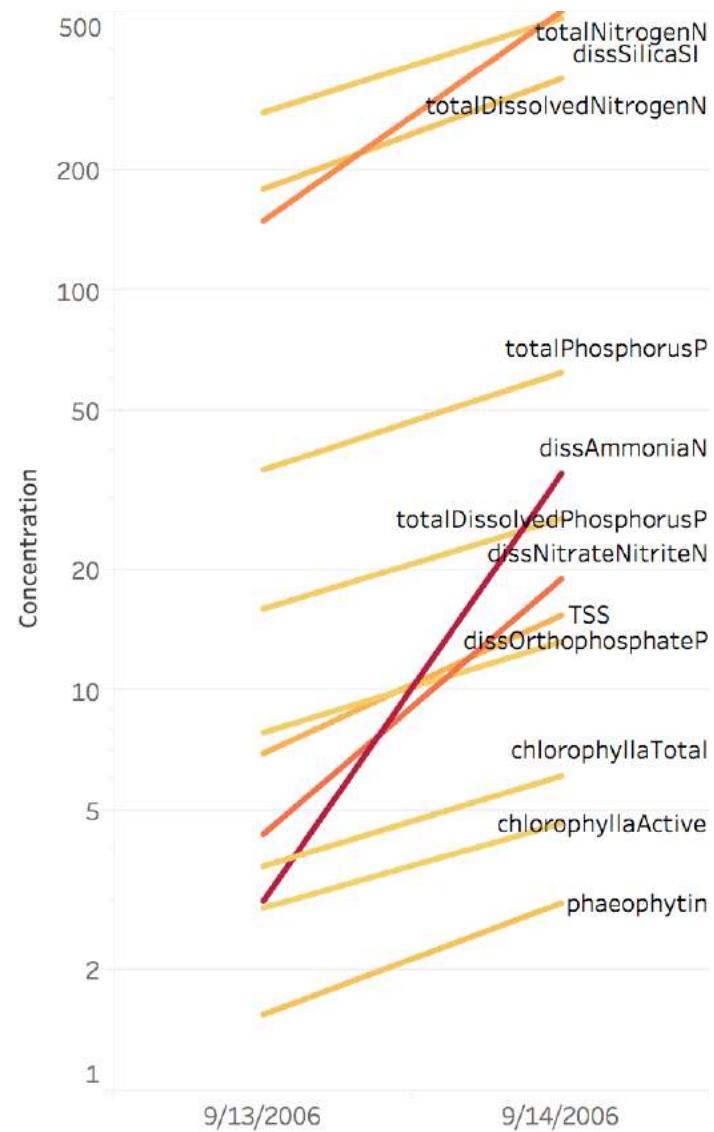
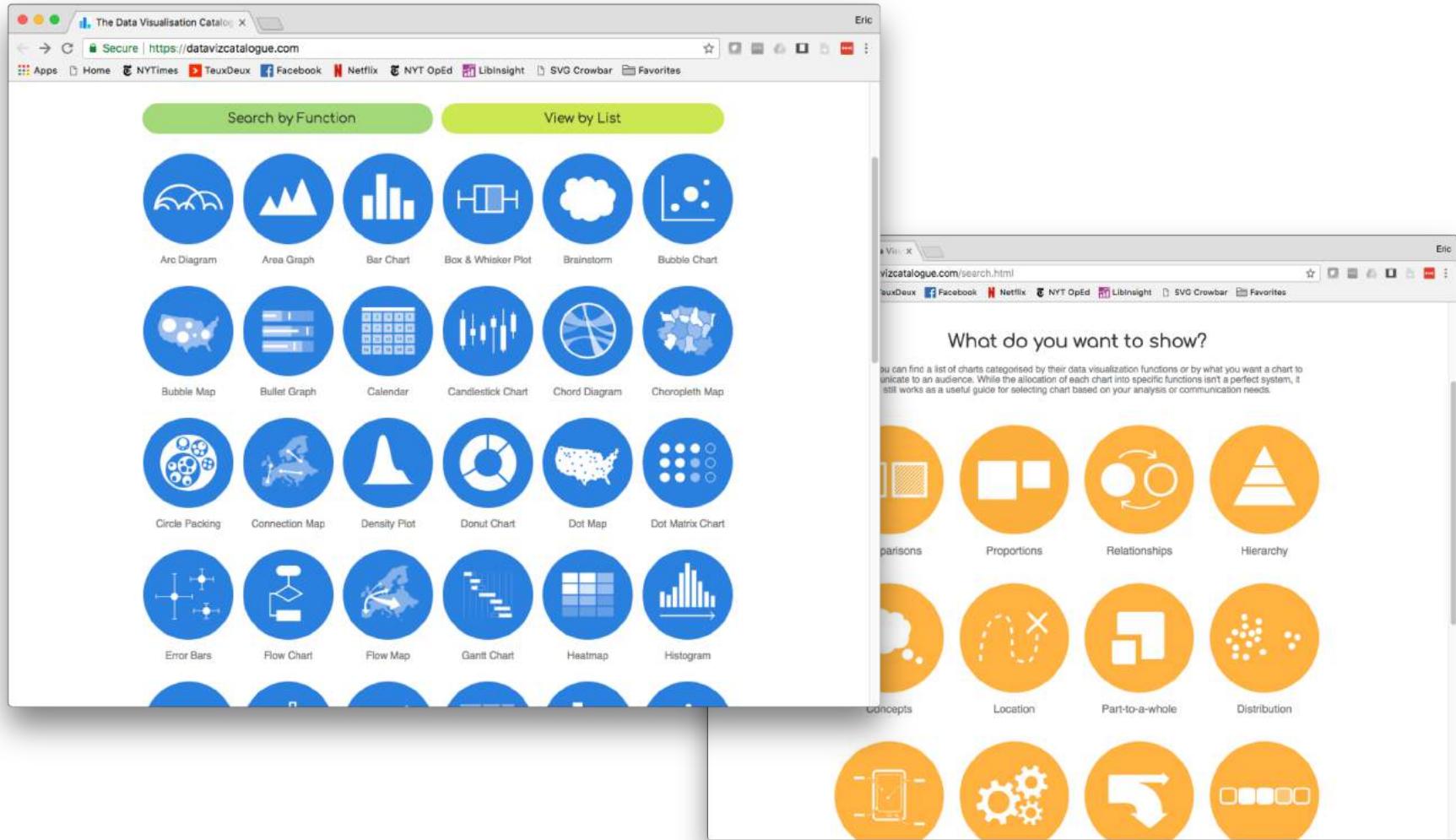


Chart chooser helper  
sites

# Chart choosing helper sites

<https://datavizcatalogue.com/>



# Chart choosing helper sites

<http://datavizproject.com/>

The screenshot displays the Data Viz Project website, a collection of data visualization tools. The interface includes a top navigation bar with links for Apps, Home, NYTimes, TeuxDeux, Facebook, Netflix, NYT OpEd, LibInsight, SVG Crowbar, and Favorites. A search bar and filter buttons for ALL, FAMILY, INPUT, FUNCTION, and SHAPE are also present.

The main content area features a grid of chart examples and a section for input matrices:

- Sankey Diagram:** A flow diagram with nodes A, B, C, D, E, F.
- Alluvial Diagram:** A flow diagram with nodes A, B, C.
- Donut Chart:** A donut chart divided into three segments labeled A, B, and C.
- Radial Bar Chart:** A radial chart with segments A, B, and C.
- Radial Histogram:** A sunburst-style radial histogram.
- Sorted Stream Graph:** A stream graph with nodes A, B, C, D.
- Fishbone Diagram:** A cause-and-effect diagram with categories like PEOPLE, PROCESSES, EQUIPMENT, METHODS, and MATERIALS.
- Matrix Diagram:** A simple matrix with columns 1, 2, 3.
- Matrix Diagram (Roof Shaped):** A roof-shaped matrix diagram.
- Arc Diagram:** A circular arc diagram.
- Hexagonal Binning:** A scatter plot using hexagonal binning.
- Radial Line Graph:** A radial line graph.
- Line Graph:** A standard line graph.
- Cluster Analysis:** Two scatter plots showing cluster analysis results.
- Area Chart:** A line area chart.
- Bagplot:** A boxplot-like chart.
- Radial Area Chart:** A radial area chart.
- Spline Graph:** A line graph with a spline fit.

Below the charts, there is a section for input matrices:

X Y	X Y <sub>z</sub>	A 32%	A 14	X Y
1 30	0 - 2 30 45	B 40%	B 16	1 14
2 34	2 - 4 34 80	C 28%	C 12	2 16
3 38	4 - 6 38 60	2 10 2 15	4 100 4 85	3 12 20

# Chart choosing helper sites

<https://github.com/ft-interactive/chart-doctor/tree/master/visual-vocabulary>

The screenshot shows a GitHub repository page for 'chart-doctor/visual-vocabulary'. The main view displays a grid of visualizations categorized into ten groups: Dimension, Correlation, Ranking, Distribution, Change over Time, Magnitude, Part-to-whole, Spatial, and Filter. Each category has a small icon and a brief description. Below this grid, there's a section titled 'Visual vocabulary' with a note about its purpose and a link to 'ft.com/vocabulary'. A separate window is open showing a detailed poster for 'Ranking', 'Distribution', and 'Change over Time', each with examples and descriptions of their use.

**Visual vocabulary**

There are as many ways to visualize data - how do we know which one to use? This guide helps you decide the most effective way to show your relationship is most important in your story, then look at the different types of chart, graphs and maps available and what they can and cannot do best. This bar is not meant to be exhaustive, nor a textbook, but is a starting point for evaluating, discussing and deciding what's best for your data.

[ft.com/vocabulary](http://ft.com/vocabulary)

**Financial Times Visual Vocabulary**

A poster (available in English, Japanese, traditional Chinese and simplified Chinese) and we journalists to select the optimal symbology for data visualisations, by the Financial Times V

**Ranking**

Use where an item's position in an ordered list is more important than its absolute or relative value. Don't be afraid to highlight the points of interest.

**Example FT uses**  
Wealth, deprivation, league tables, constituency election results

**Distribution**

Show values in a dataset and how often they occur. The shape (or 'skew') of a distribution can be a memorable way of highlighting the lack of uniformity or equality in the data.

**Example FT uses**  
Income distribution, population (age/sex) distribution

**Change over Time**

Give emphasis to changing trends. These can be short (intra-day) movements or extended series traversing decades or centuries. Choosing the correct time period is important to provide suitable context for the reader.

**Example FT uses**  
Share price movements, economic time series

**Line**

The standard way to show a changing time series. If data are irregular, consider markers to represent data points.

**Column**

Columns work well for showing change over time - but usually best with only one series of data at a time.

**Column + line timeline**

A good way of showing the relationship over time

**Ordered bar**  
Standard bar charts display the ranks of values much more easily when sorted into order.

**Histogram**  
The standard way to show a statistical distribution. Use the gaps between columns small to highlight the 'shape' of the data.

**Ordered column**  
See above.

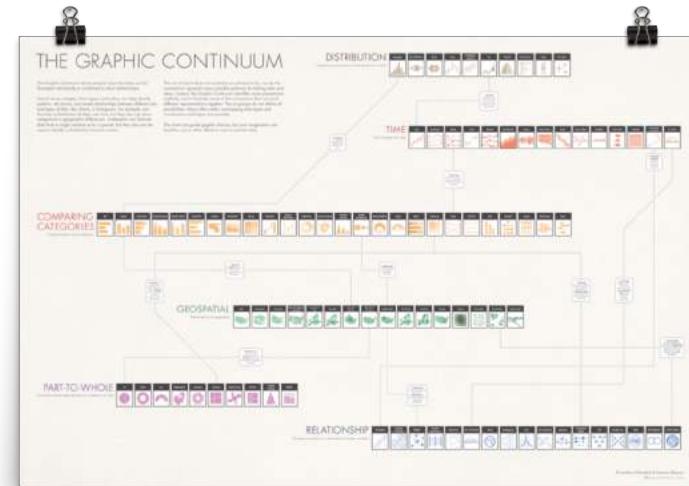
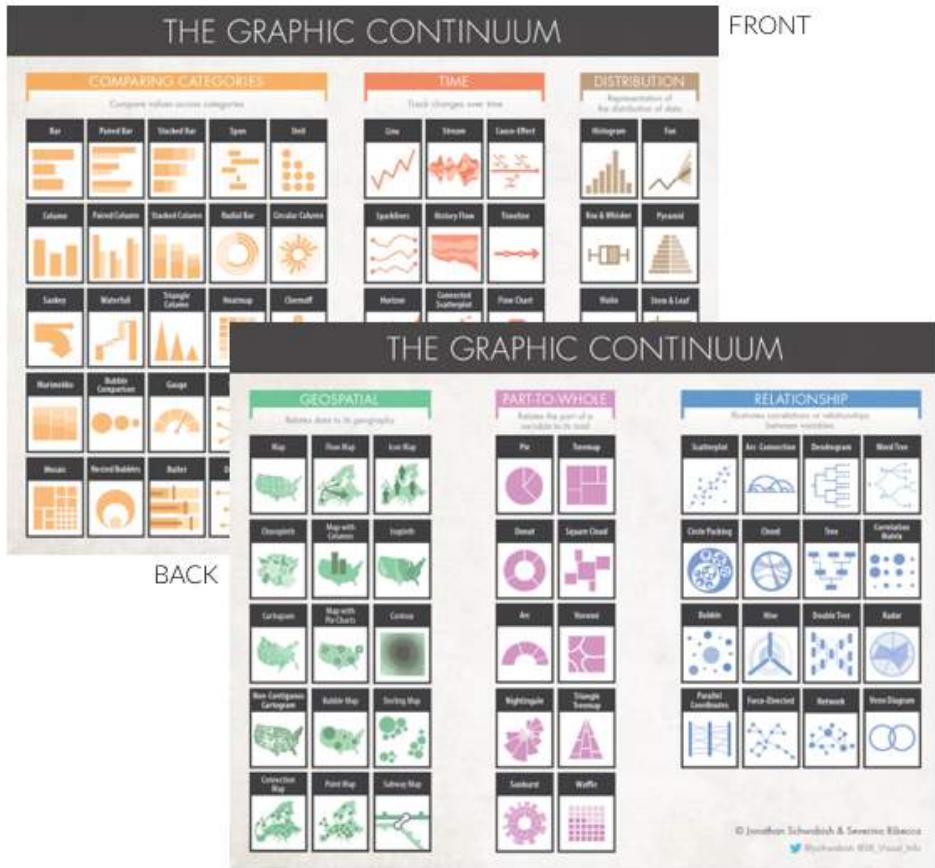
**Dot plot**  
A simple way of showing the change or range (minimum) of data across multiple categories.

**Ordered proportional symbol**  
Use when there are big variations between values and/or seeing

**Dot strip plot**  
Good for showing individual values in a distribution, can be a

# Chart choosing helper sites

<https://policyviz.com/shop/>



# Chart choosing helper sites

<http://chartmaker.visualisingdata.com/>

The screenshot shows a web browser window titled "The Chartmaker Directory". The URL in the address bar is "chartmaker.visualisingdata.com". The page features a header with a search bar, navigation links (ABOUT, V), and a sidebar with social media icons (Twitter, Facebook, Google+, LinkedIn). The main content is a grid-based chart compatibility matrix. The columns represent different charting tools: Amazon QuickSight, ArcGIS, ChartJS, D3.js, Data Illustrator, Datawrapper, Flourish, FusionCharts, Gephi, Google Charts, and Highcharts. The rows list various chart types: Bar chart, Clustered bar chart, Bullet chart, Connected dot plot, Pictogram, Bubble chart, Word cloud, and Radar chart. Each cell in the grid contains a black dot icon, where the position of the dots indicates compatibility: top-left dot means the tool supports the chart type, bottom-right dot means it's a solution for that chart type, and other positions indicate specific reference types like Categorical, Hierarchical, Relational, Temporal, or Spatial.

	Amazon QuickSight	ArcGIS	ChartJS	D3.js	Data Illustrator	Datawrapper	Flourish	FusionCharts	Gephi	Google Charts	Highcharts
Bar chart	●			●	○	●	○	○		●	●
Clustered bar chart	●				●	●	○	○		●	
Bullet chart				●		●			○		
Connected dot plot				●		●					
Pictogram					○						
Bubble chart				●	○		○○	○		●	
Word cloud					●						
Radar chart				○	●			○		○	

# Chart choosing helper sites

<https://xeno.graphics/>

