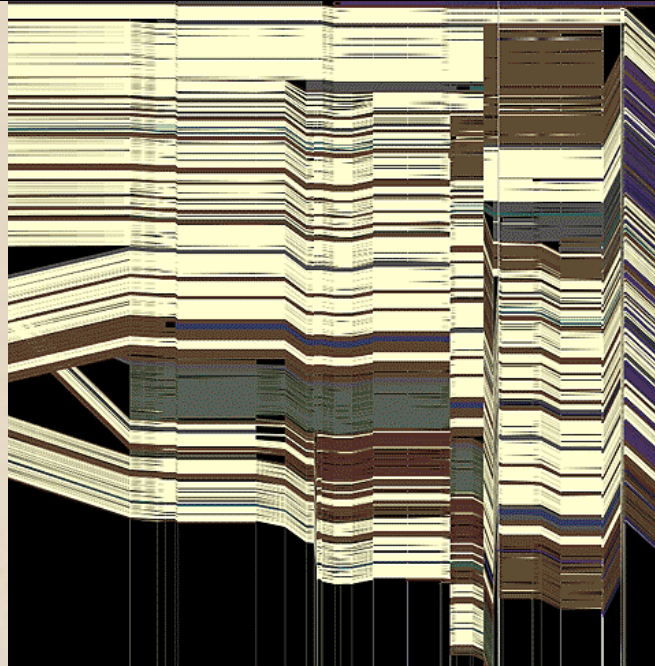
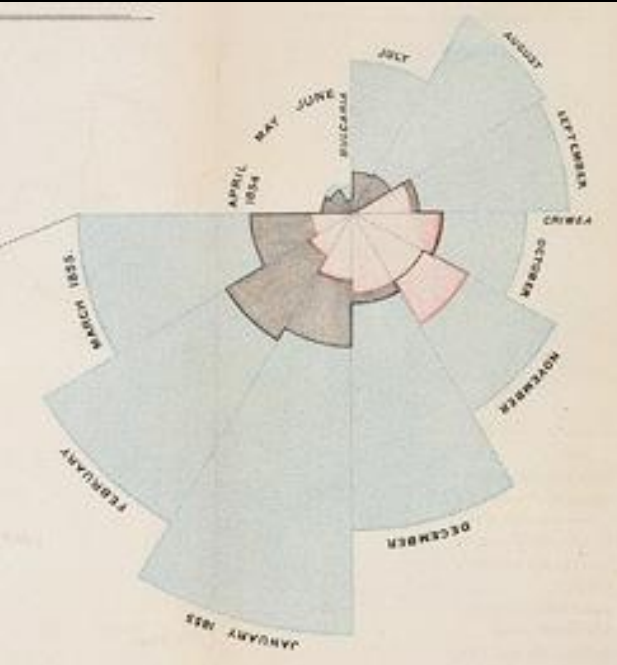


CSE 442 - Data Visualization

# Visual Encoding Design



Leilani Battle University of Washington

# Lecture Overview

## Learning Goals

How do we apply existing encoding principles to univariate, bivariate, and multivariate data?

## Topics

Exploring the Visual Design Space

Encoding Effectiveness

Scales & Axes

Multidimensional Data

# A Design Space of Visual Encodings

# Mapping Data to Visual Variables

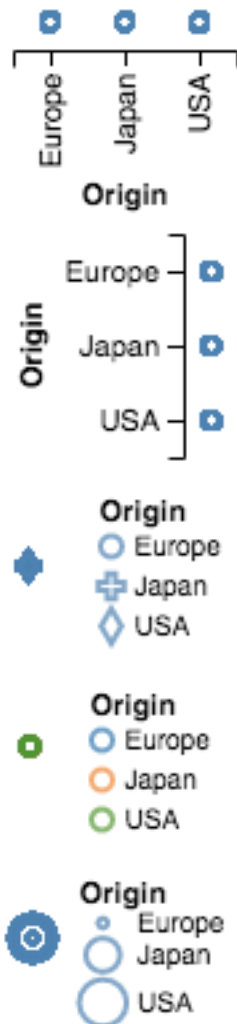
Assign **data fields** (e.g., with *N*, *O*, *Q* types) to **visual channels** (*x*, *y*, *color*, *shape*, *size*, ...) for a chosen **graphical mark type** (*point*, *bar*, *line*, ...).

Additional concerns include choosing appropriate **encoding parameters** (*log scale*, *sorting*, ...) and **data transformations** (*bin*, *group*, *aggregate*, ...).

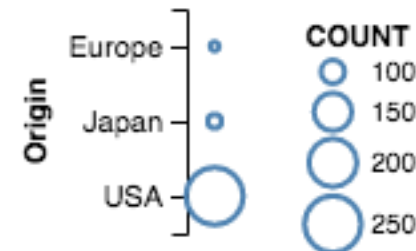
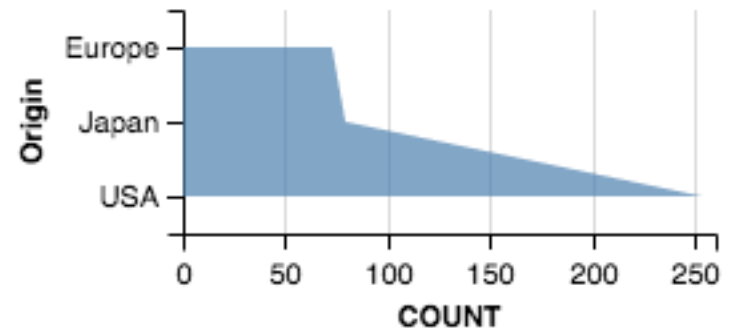
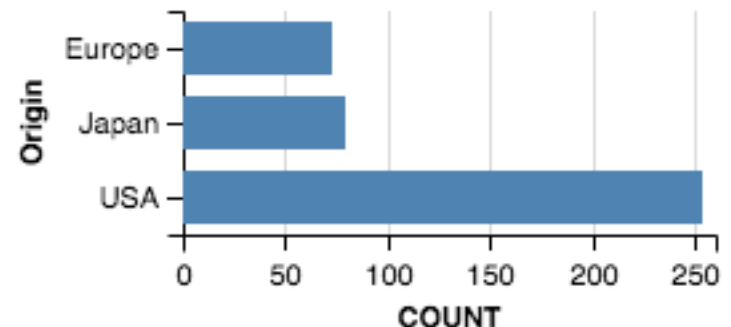
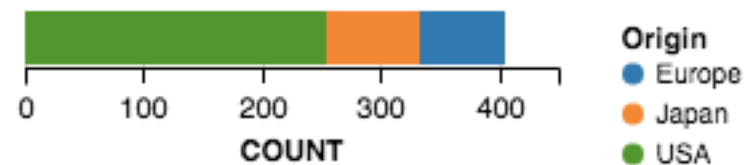
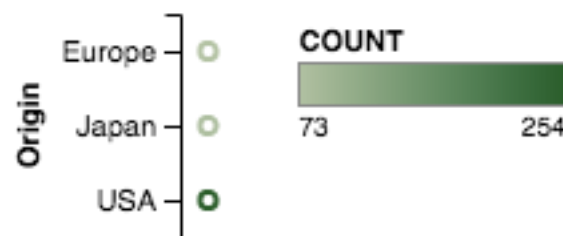
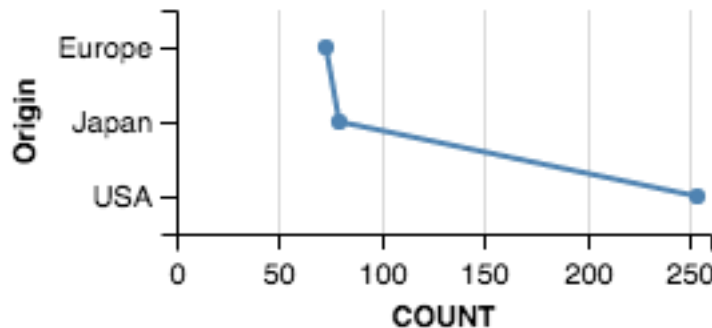
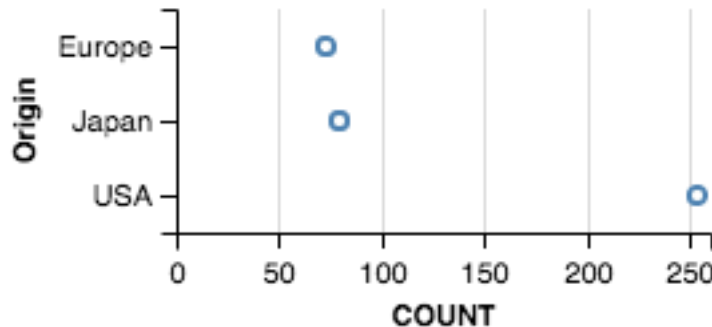
These options define a large combinatorial space, containing both useful and questionable charts!

# 1D: Nominal

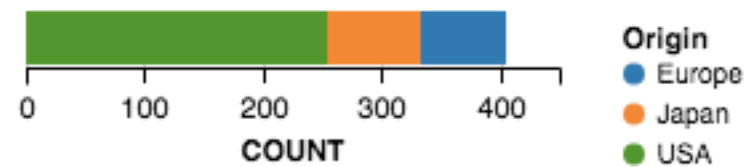
Raw



Aggregate (Count)

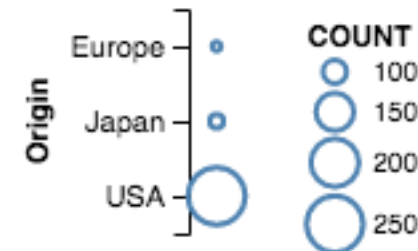
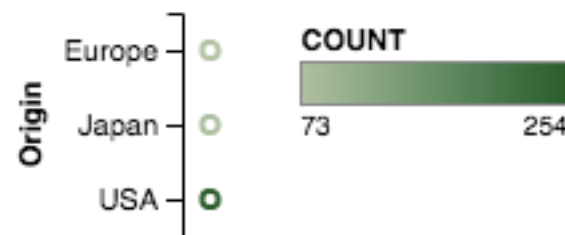
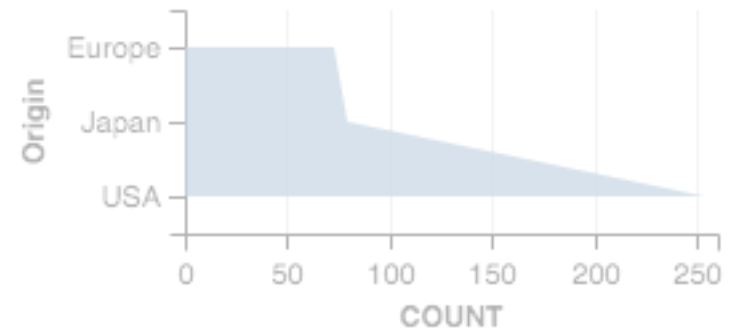
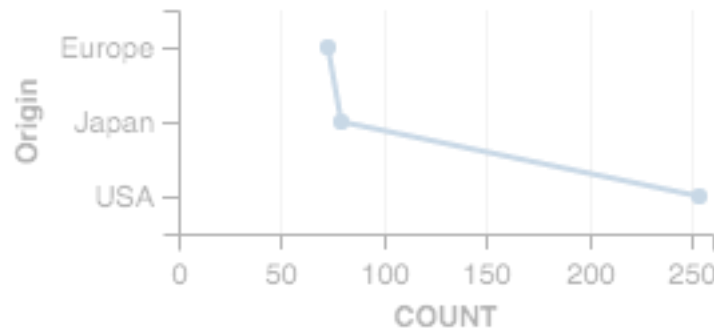
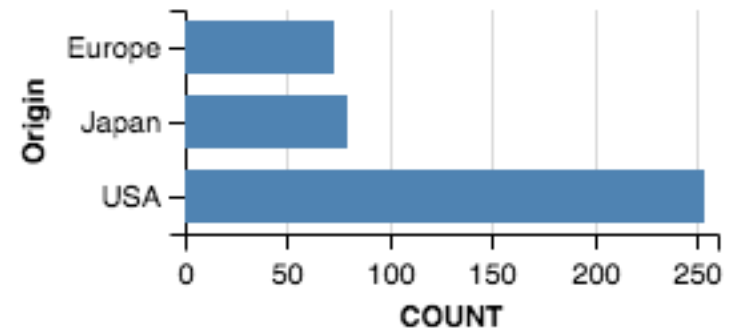
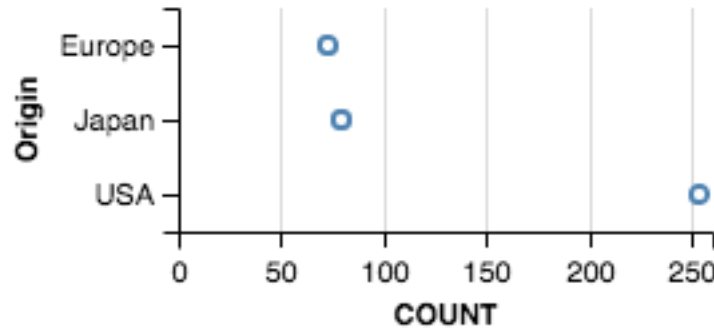


# Expressive?



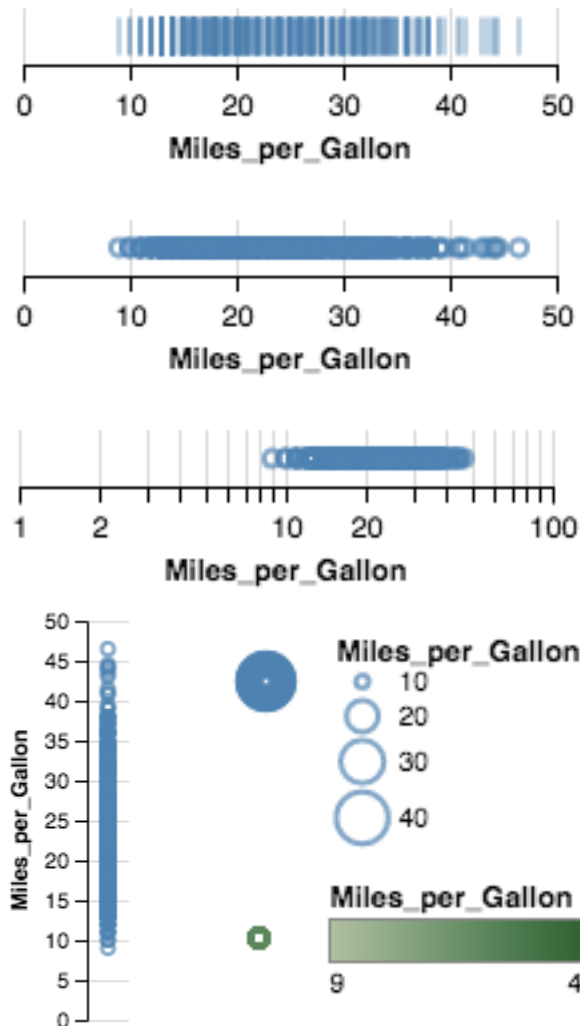
Raw

Aggregate (Count)

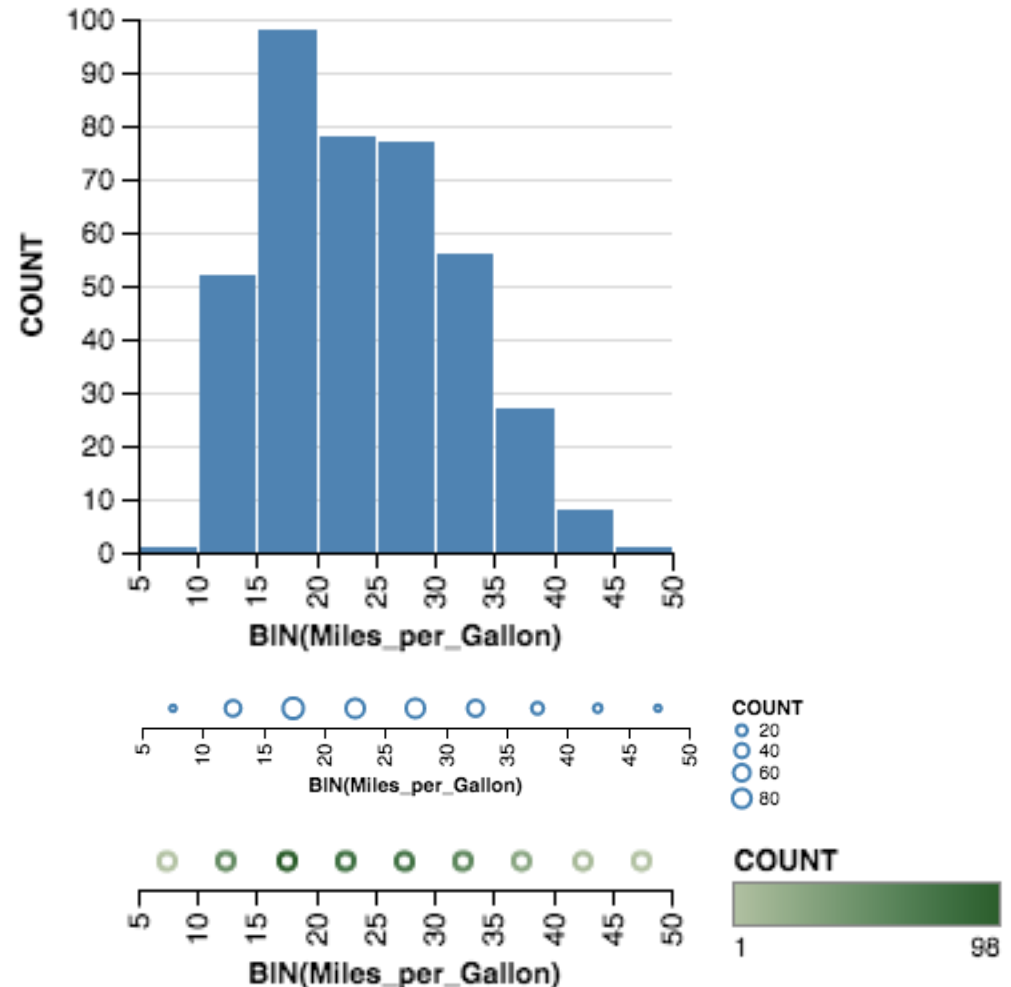


# 1D: Quantitative

Raw

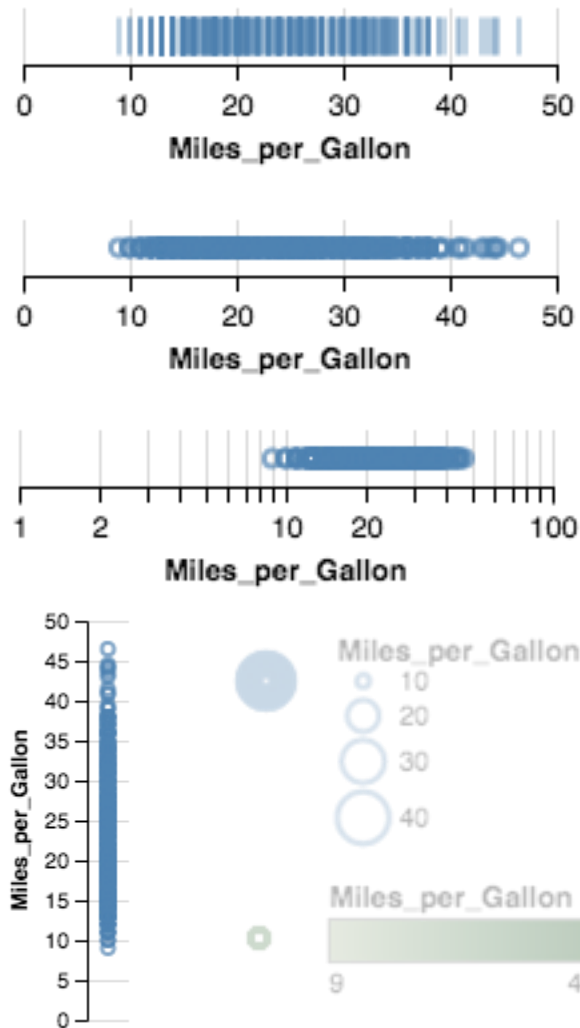


Aggregate (Count)

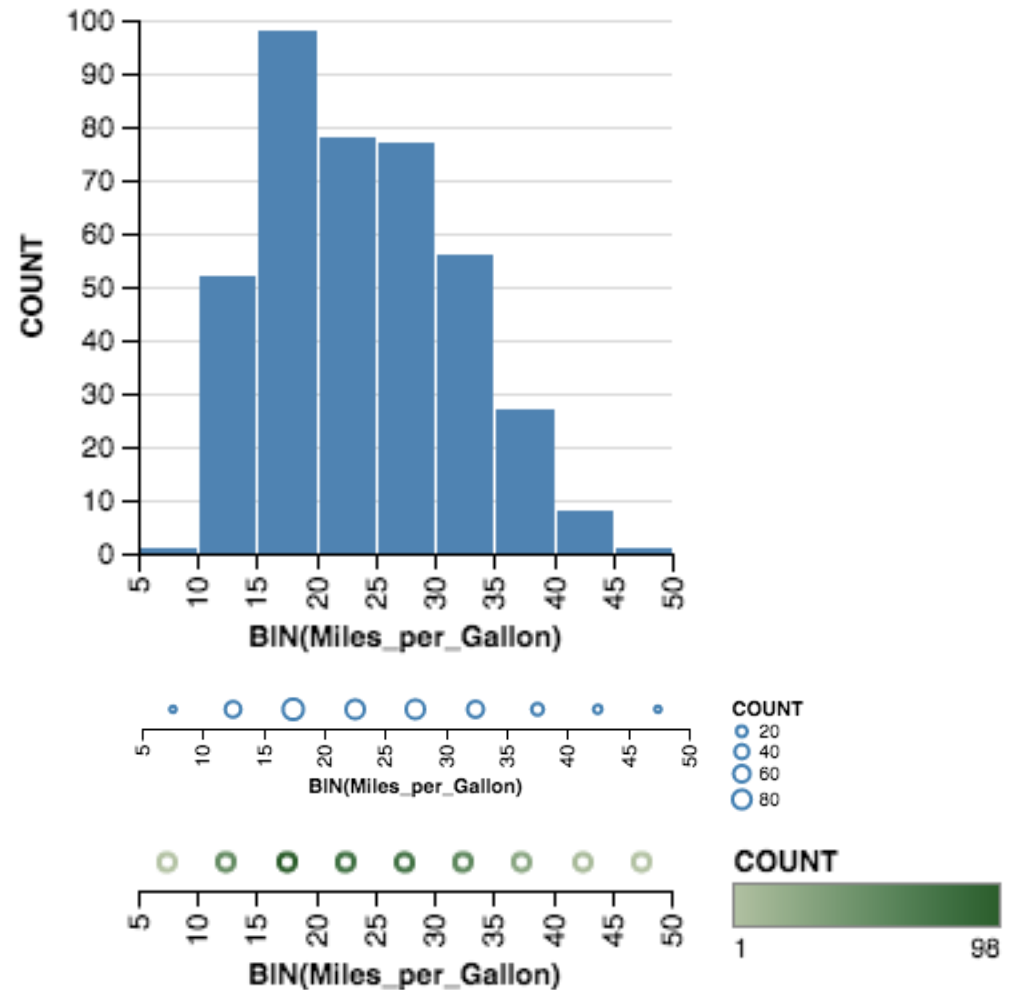


# Expressive?

Raw



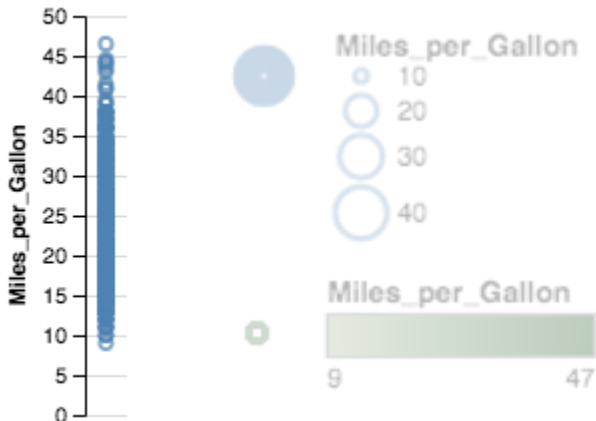
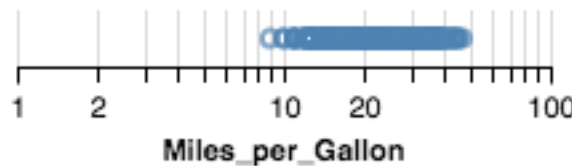
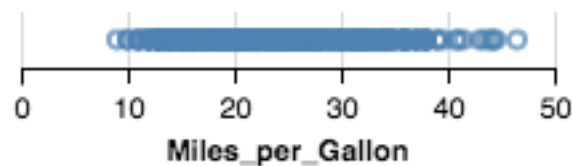
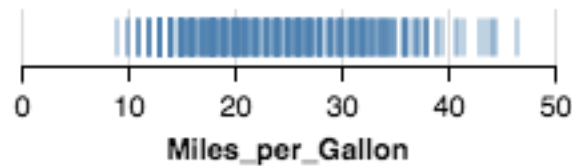
Aggregate (Count)



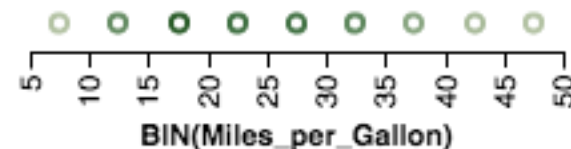
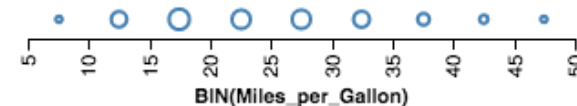
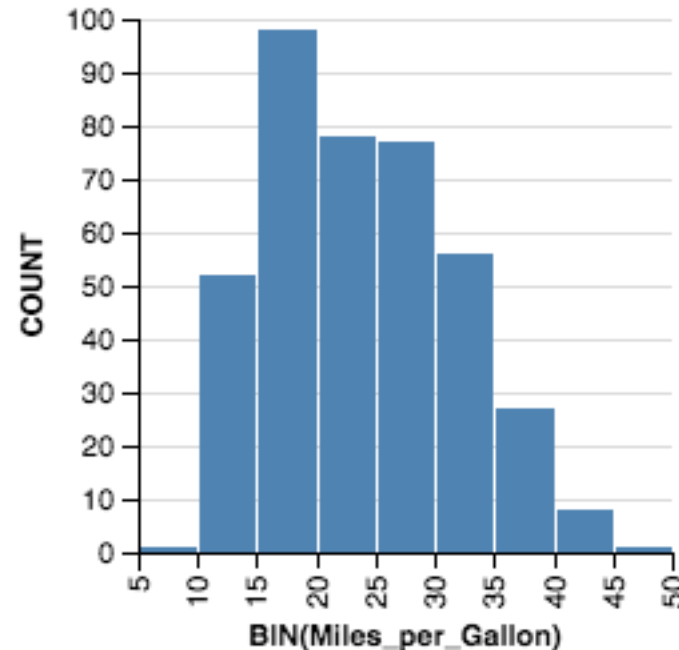


# Effective?

Raw

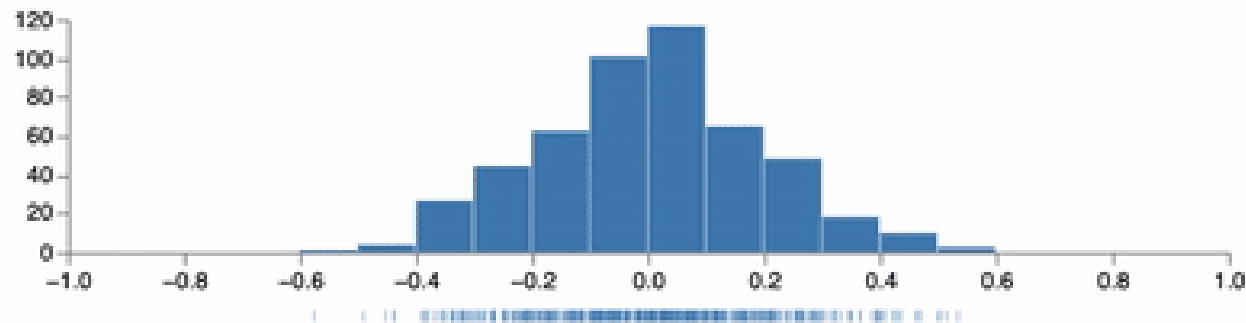


Aggregate (Count)



# Histogram Example

A [histogram](#) subdivides a numerical range into bins, and counts the number of data points within each segment. This chart provides a discrete estimate of the probability density function.



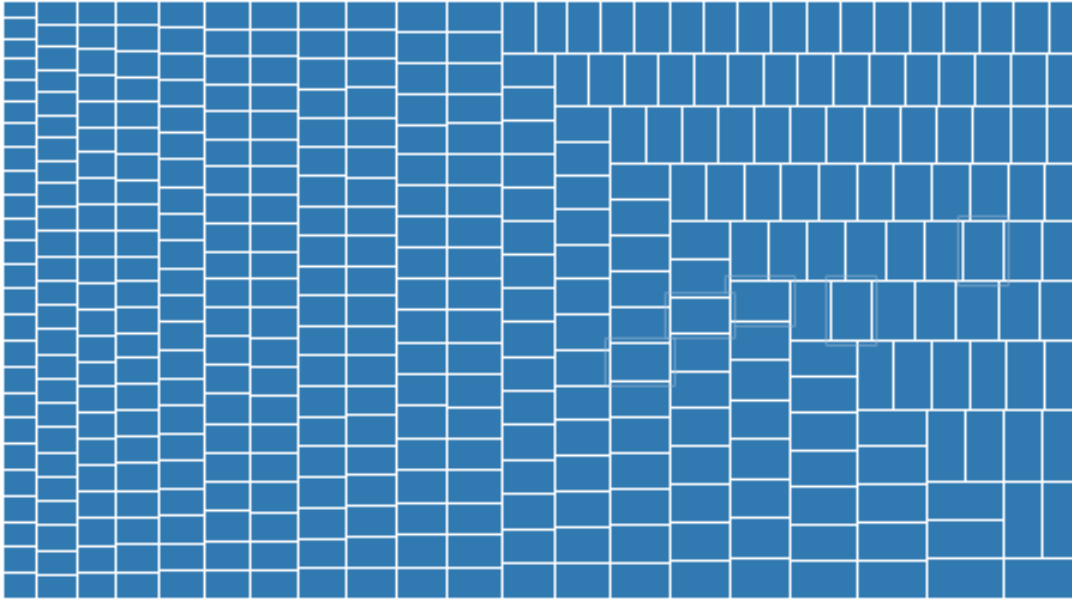
binOffset

 0

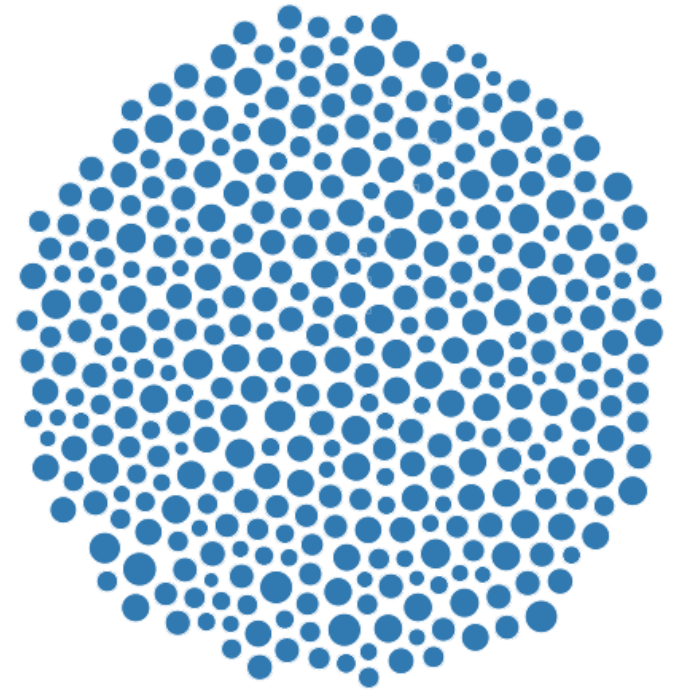
binStep

 0.1[View in Online Vega Editor](#)

## Raw (with Layout Algorithm)

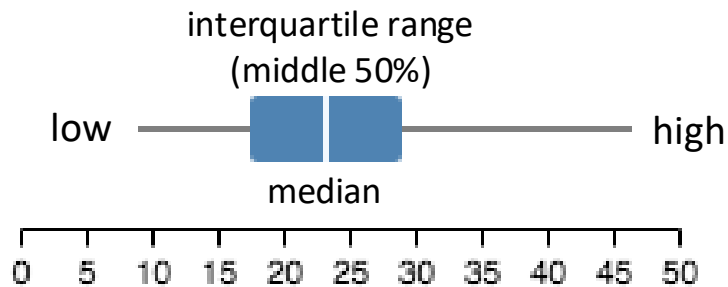


Treemap

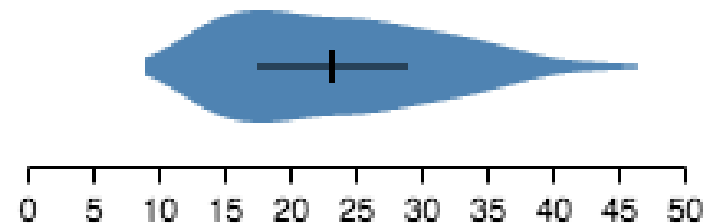


Bubble Chart

## Aggregate (Distributions)



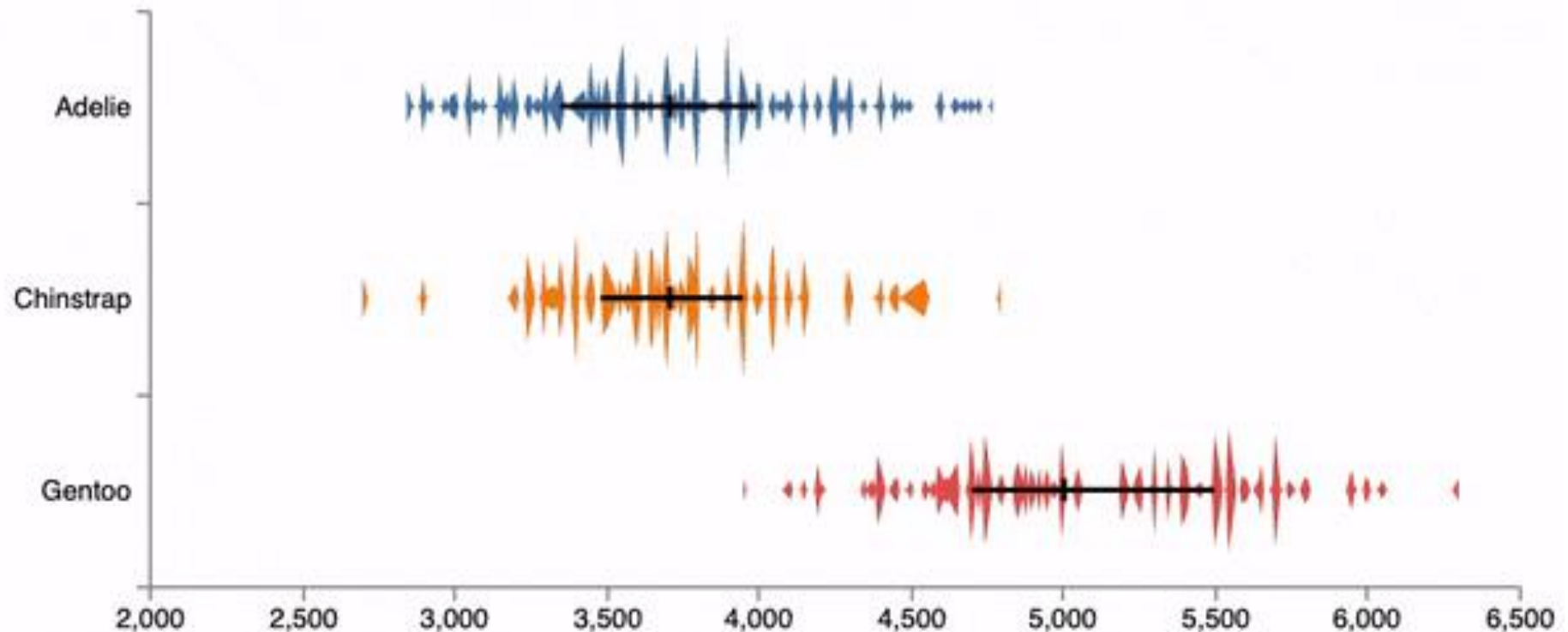
Box Plot



Violin Plot

# Violin Plot Example

A **violin plot** visualizes a distribution of quantitative values as a continuous approximation of the probability density function, computed using **kernel density estimation (KDE)**. The densities are additionally annotated with the **median** value and **interquartile range**, shown as black lines. Violin plots can be more informative than classical **box plots**.



trim



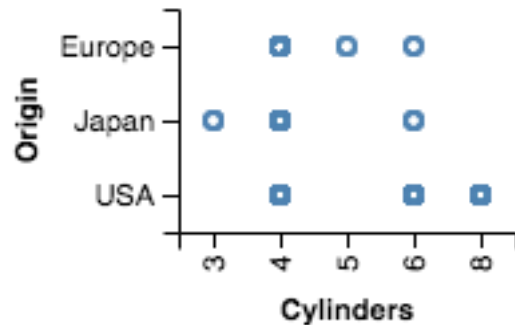
bandwidth



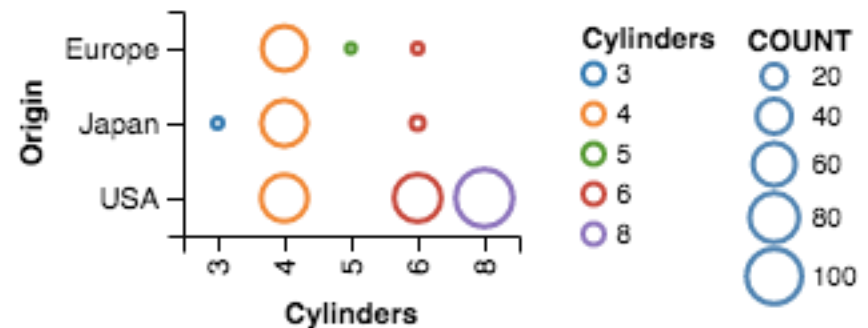
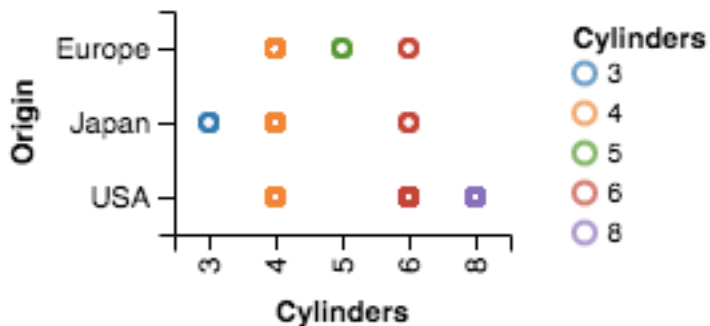
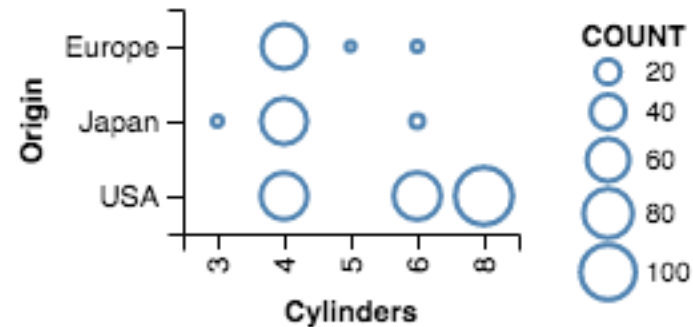
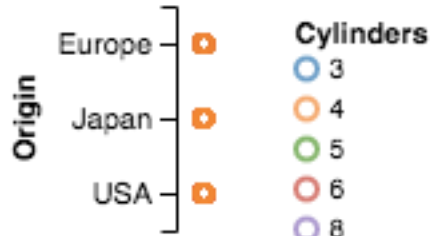
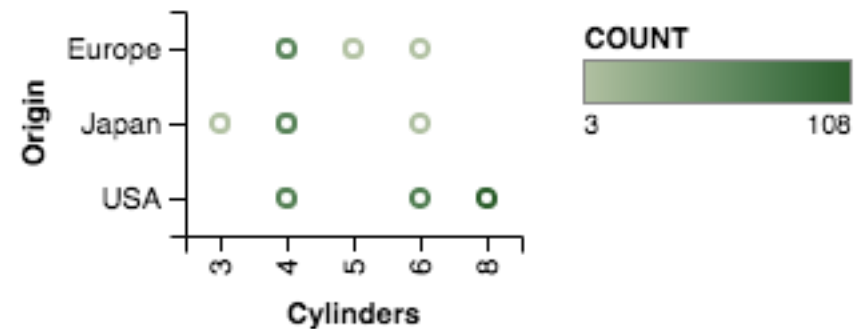
5

# 2D: Nominal x Nominal

Raw

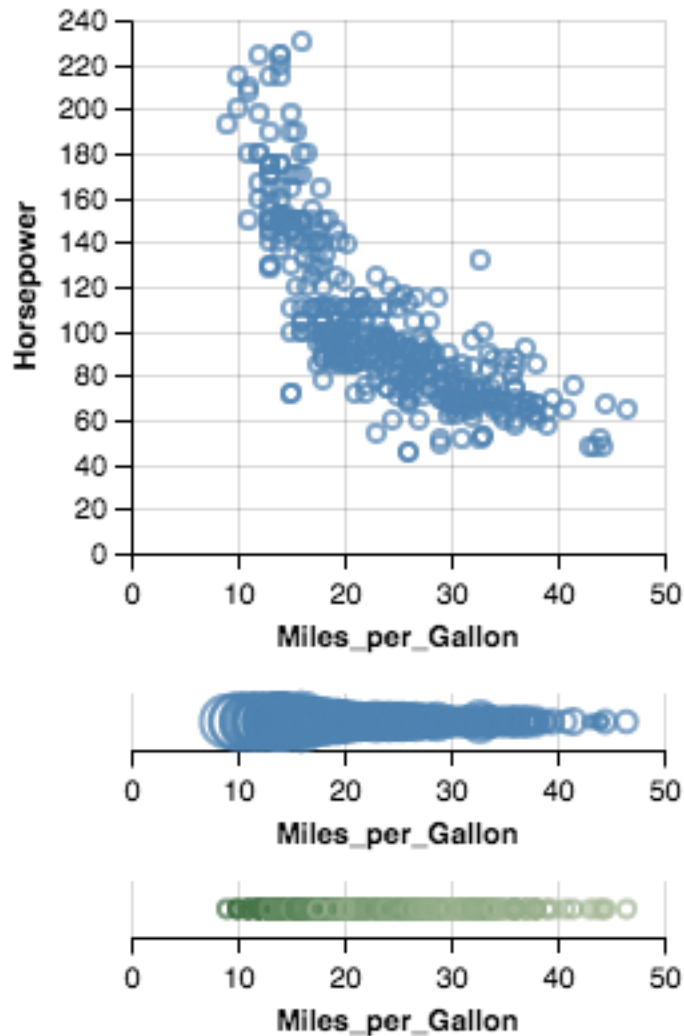


Aggregate (Count)

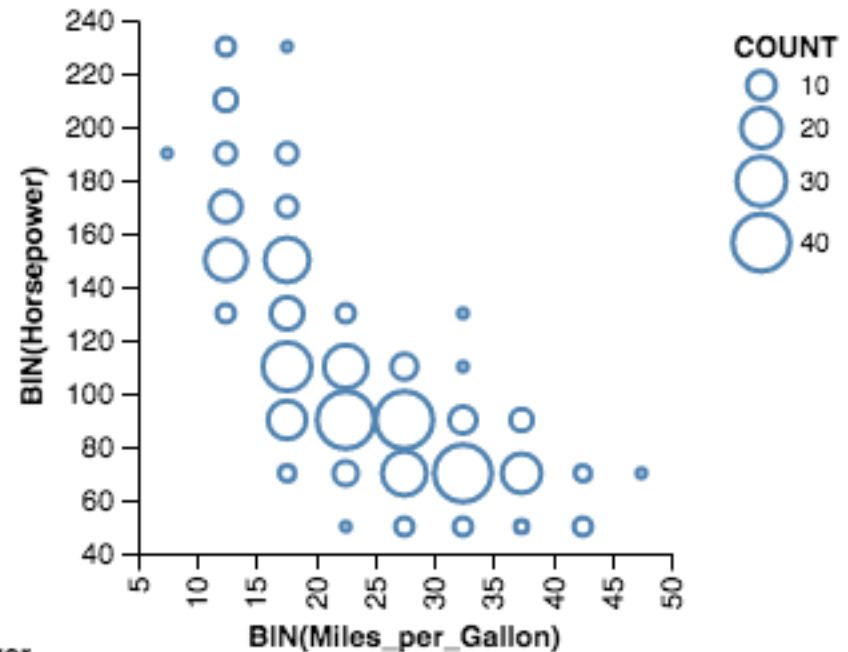


# 2D: Quantitative x Quantitative

Raw

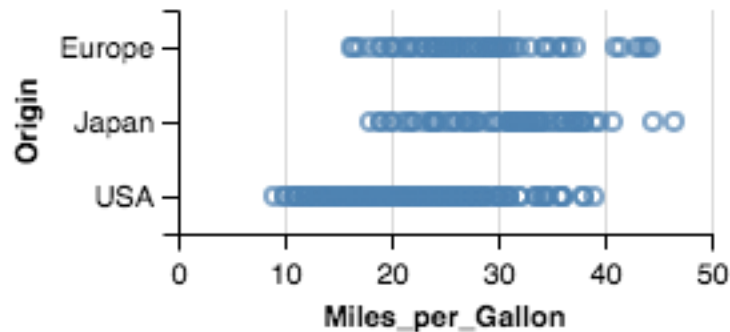


Aggregate (Count)

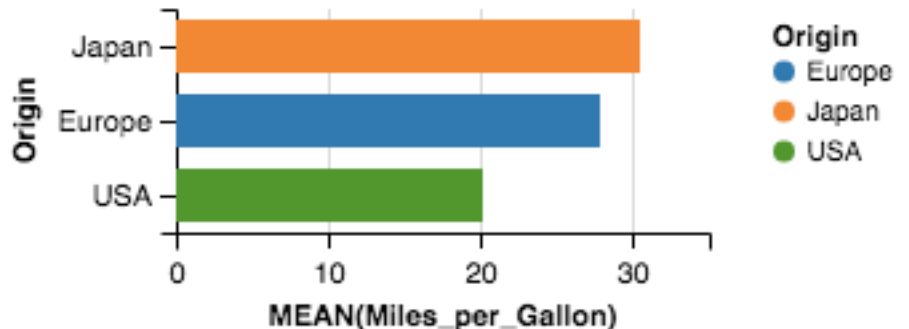
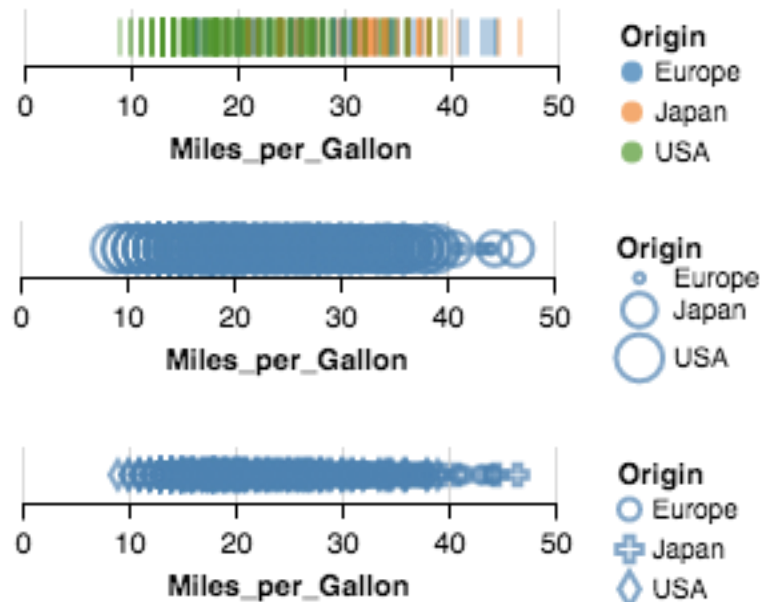
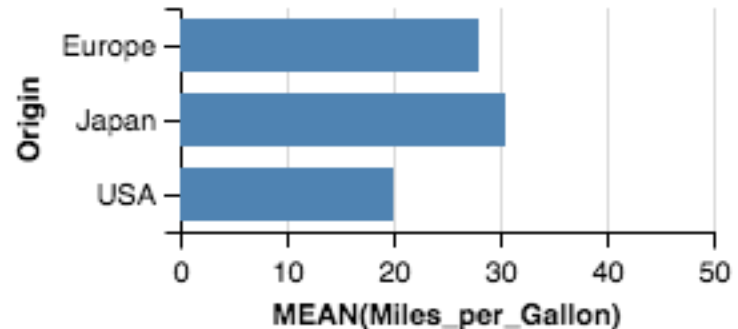


# 2D: Nominal x Quantitative

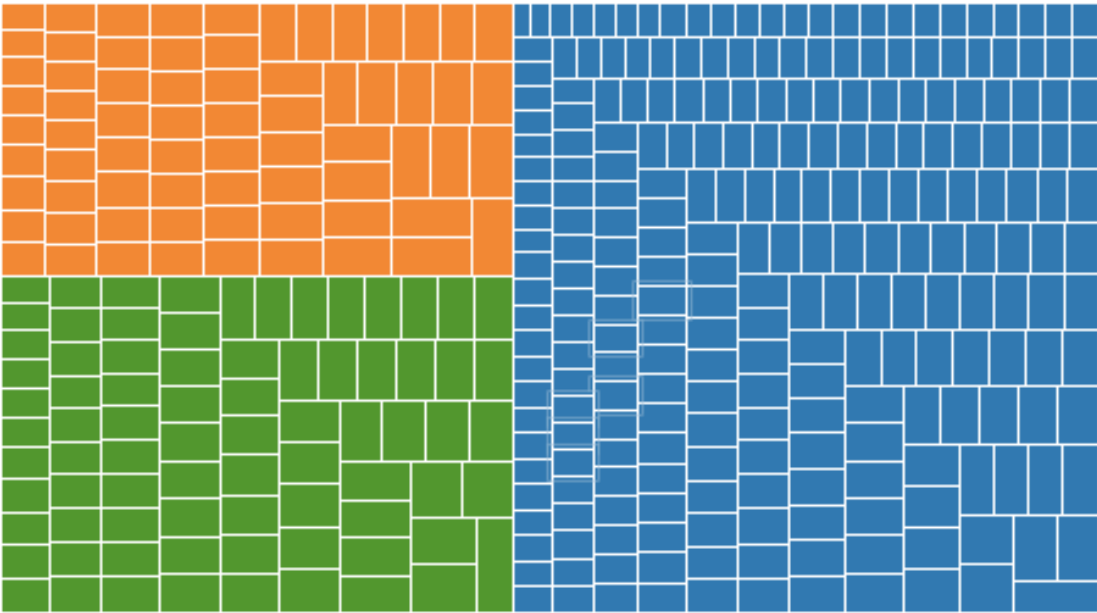
Raw



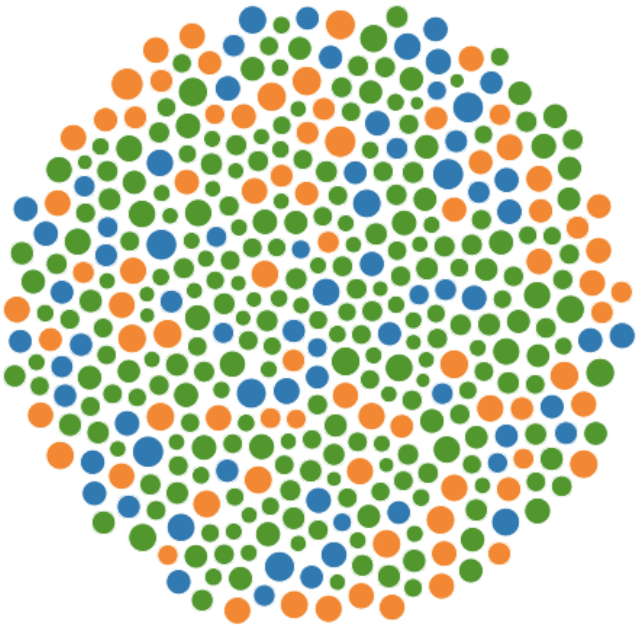
Aggregate (Mean)



Raw (with Layout Algorithm)

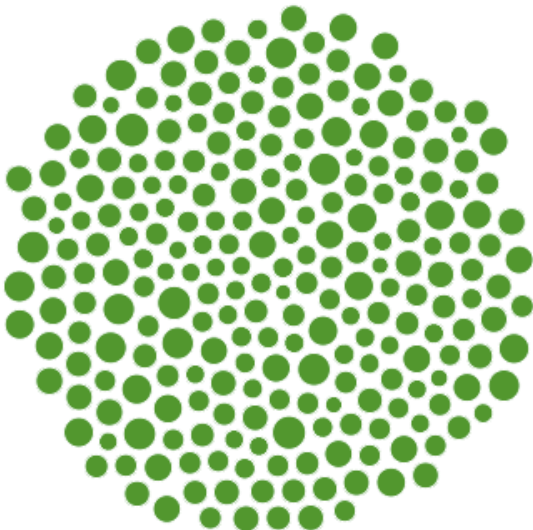
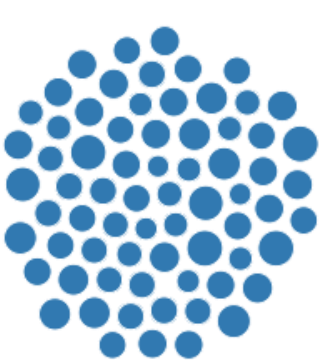


Treemap



Bubble Chart

Origin  
● Europe  
● Japan  
● USA



Beeswarm Plot



# 3D and Higher

**Two variables** [x,y]

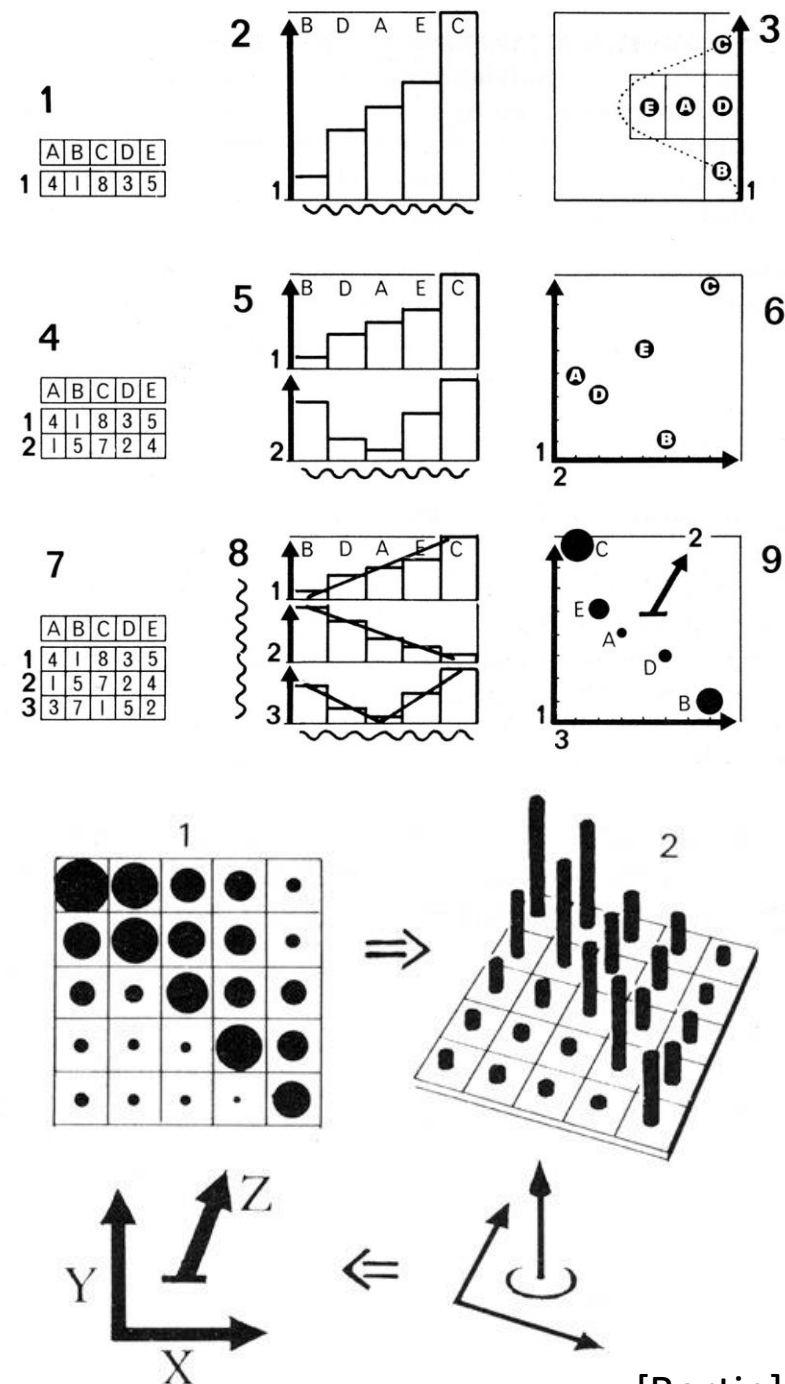
Can map to 2D points.

Scatterplots, maps, ...

**Third variable** [z]

Often use one of size, color, etc. Or, one can further partition space.

*What about 3D rendering?*



# Other Visual Encoding Channels?

## wind map

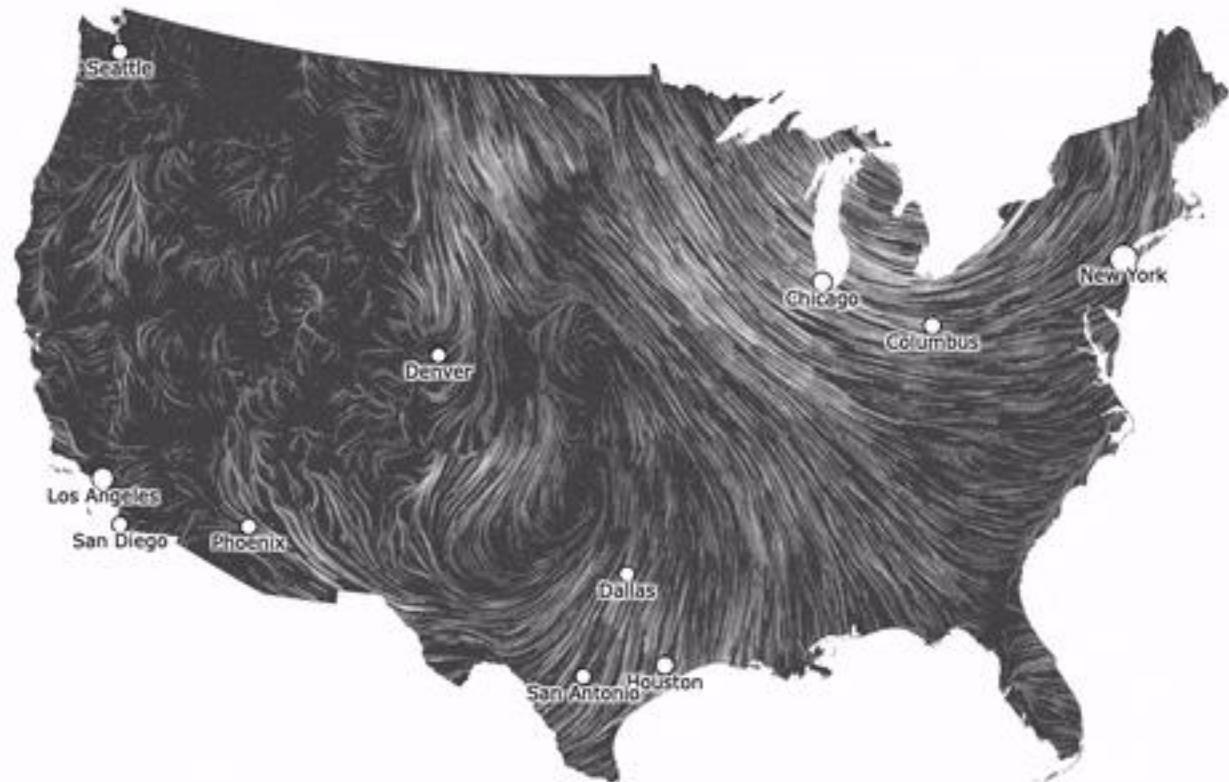
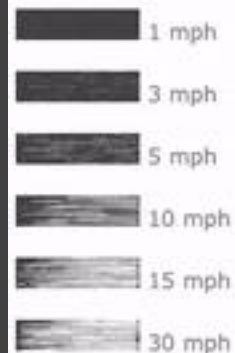
**Sept. 26, 2022**

5:27 pm EST

(time of forecast download)

top speed: 26.0 mph

average: 8.7 mph



# Design Space Summary

Visualization design **principles** help us navigate a large combinatorial space of chart designs.

**Expressive** charts show all and only the data we aim to display.

**Effective** charts maximize the readability of the data.

Data **transformations** such as aggregations can make large data easier to read.

Encoding **parameters** such as bin widths should be selected to suit the input dataset.

# Encoding Effectiveness

# Effectiveness Rankings [Mackinlay 86]

## QUANTITATIVE ORDINAL

Position	Position
Length	Density (Value) Color Hue
Angle	Color Sat
Slope	Color Hue
Area (Size)	Texture
Volume	Connection
Density (Value)	Containment
Color Sat	Length
Color Hue	Angle
Texture	Slope
Connection	Area (Size)
Containment	Volume
Shape	Shape

## NOMINAL

Position
Texture
Connection
Containment
Density (Value)
Color Sat
Shape
Length
Angle
Slope
Area
Volume

# Effectiveness Rankings [Mackinlay 86]

## QUANTITATIVE ORDINAL

## NOMINAL

<b>Position</b> .....		<b>Position</b> .....		<b>Position</b>
Length	Density (Value)	Color Hue		
Angle	Color Sat		Texture	
Slope	Color Hue		Connection	
Area (Size)	Texture		Containment	
Volume	Connection		Density (Value)	
Density (Value)	Containment		Color Sat	
Color Sat	Length		Shape	
Color Hue	Angle		Length	
Texture	Slope		Angle	
Connection	Area (Size)		Slope	
Containment	Volume		Area	
Shape	Shape		Volume	

# Effectiveness Rankings [Mackinlay 86]

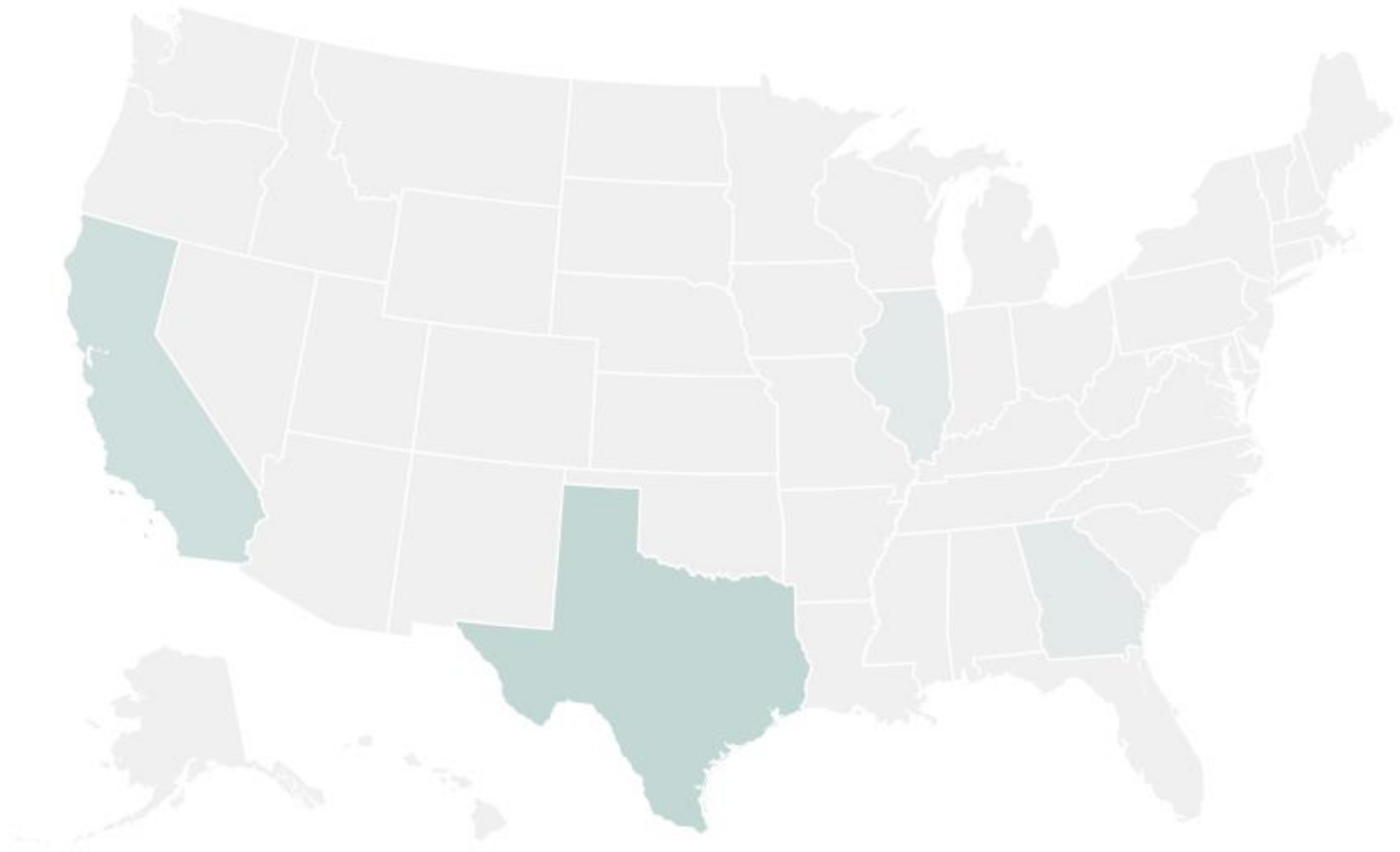
## QUANTITATIVE ORDINAL

Position  
Length  
Angle  
Slope  
Area (Size)  
Volume  
Density (Value)  
Containment  
Color Sat  
**Color Hue**  
Texture  
Connection  
Containment  
Shape

Position  
Density (Value)  
**Color Hue**  
Color Sat  
**Color Hue**  
Texture  
Connection  
Containment  
Length  
Angle  
Slope  
Area (Size)  
Volume  
Shape

## NOMINAL

Position  
Texture  
Connection  
Containment  
Density (Value)  
Color Sat  
Shape  
Length  
Angle  
Slope  
Area  
Volume



Color Encoding (Choropleth Map)



# Effectiveness Rankings

## QUANTITATIVE ORDINAL

Position

Length

Angle

Slope

★ **Area (Size)**

Volume

~~**Density (Value)**~~

Color Sat

Color Hue

Texture

Connection

Containment

Shape

Position  
Density (Value) Color Hue

Color Sat

Color Hue

Texture

Connection

Containment

Length

Angle

Slope

Area (Size)

Volume

Shape

## NOMINAL

Position

Texture

Connection

Containment

Density (Value)

Color Sat

Shape

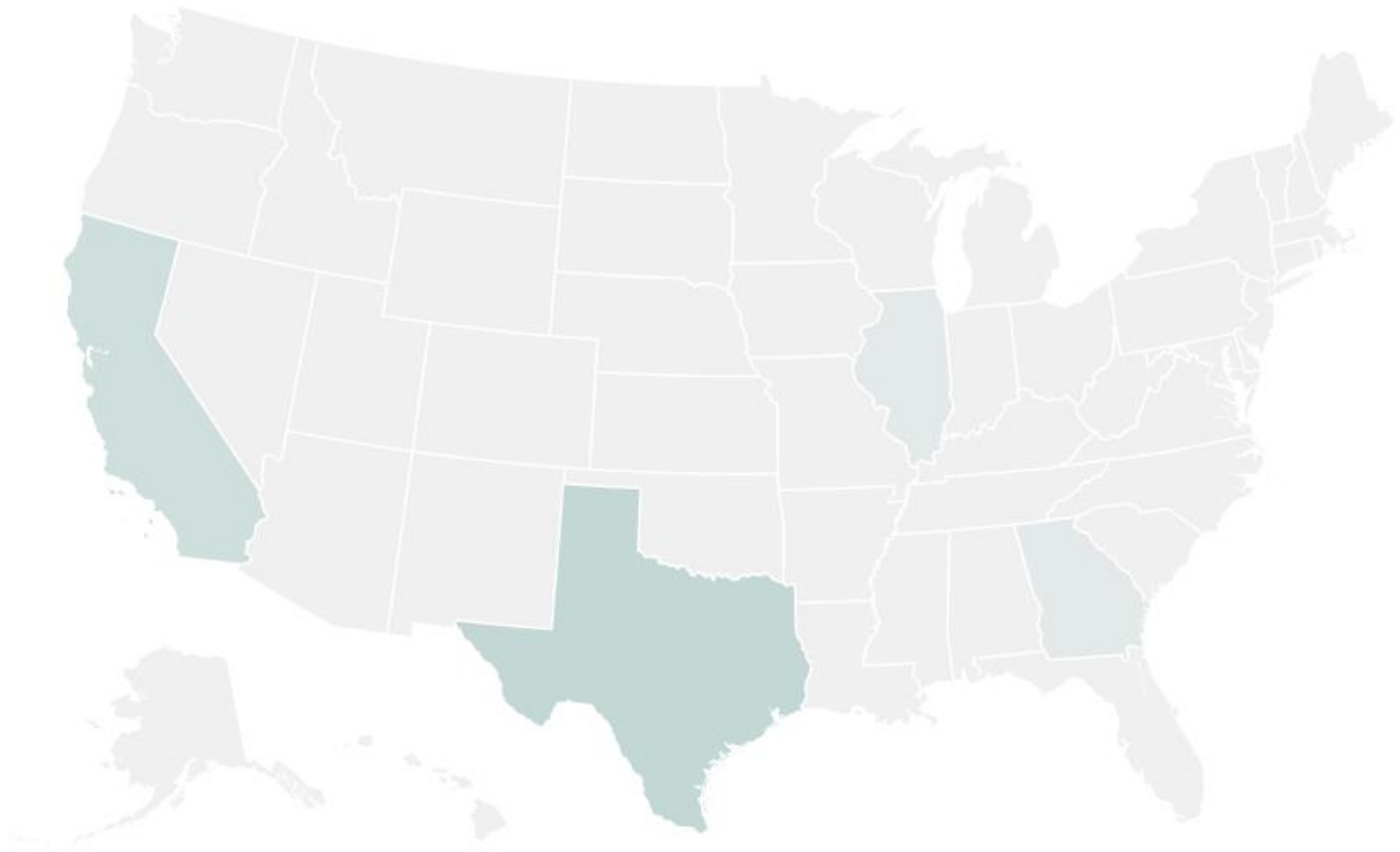
Length

Angle

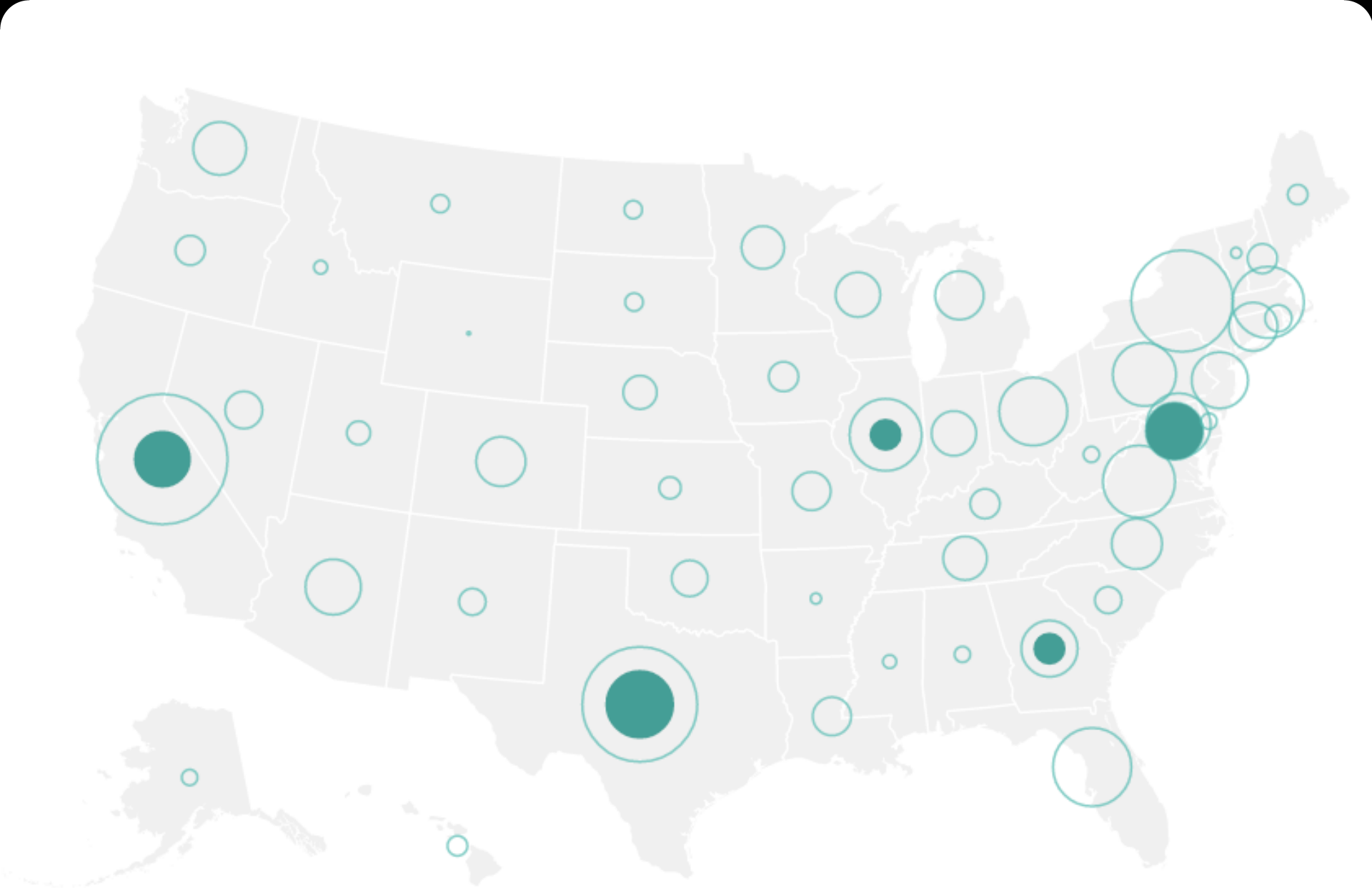
Slope

Area

Volume



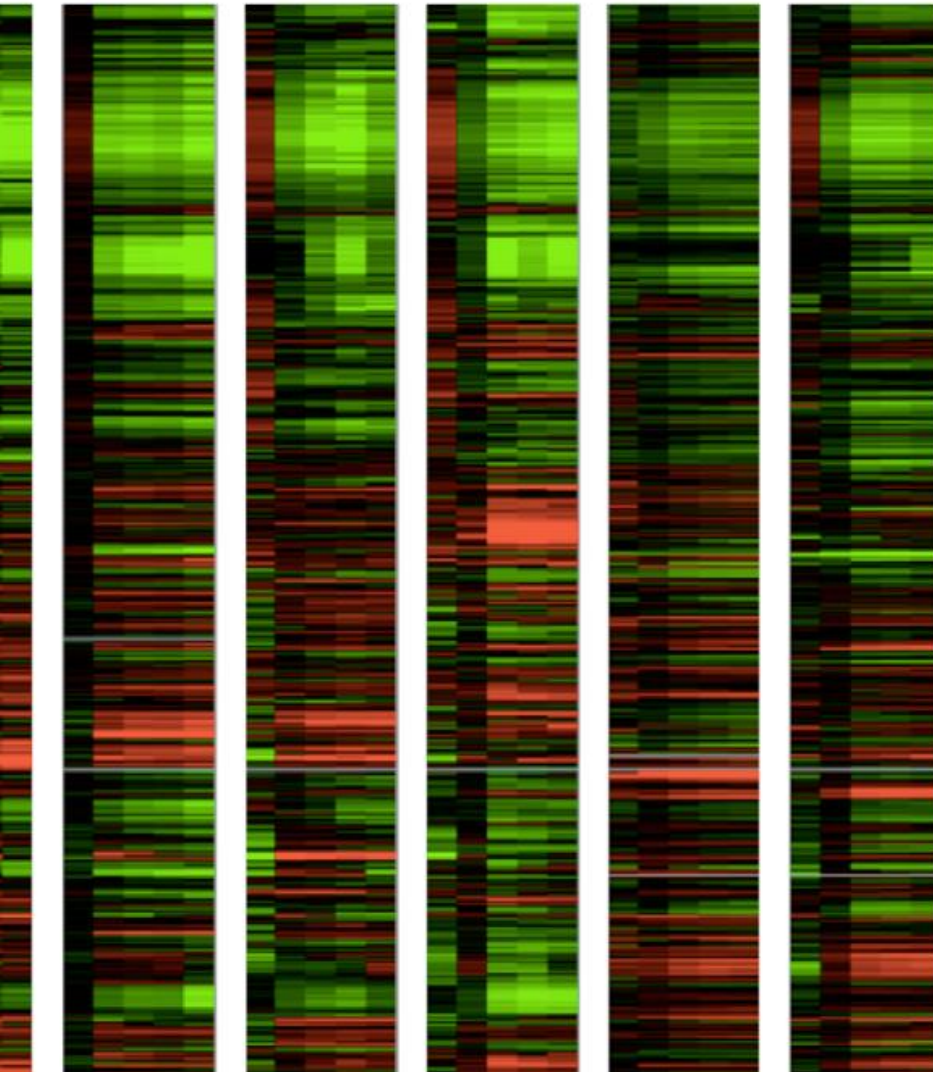
Color Encoding (Choropleth Map)



Area Encoding (Symbol Map)

# Gene Expression Time-Series [Meyer et al '11]

Color Encoding



# Effectiveness Rankings

## QUANTITATIVE ORDINAL

## NOMINAL

### ★ Position

Length

Angle

Slope

Area (Size)

Volume

~~Density (Value)~~

Color Sat

~~Color Hue~~

Texture

Connection

Containment

Shape

Position  
Density (Value) Color Hue

Color Sat

Color Hue

Texture

Connection

Containment

Length

Angle

Slope

Area (Size)

Volume

Shape

Position

Texture

Connection

Containment

Density (Value)

Color Sat

Shape

Length

Angle

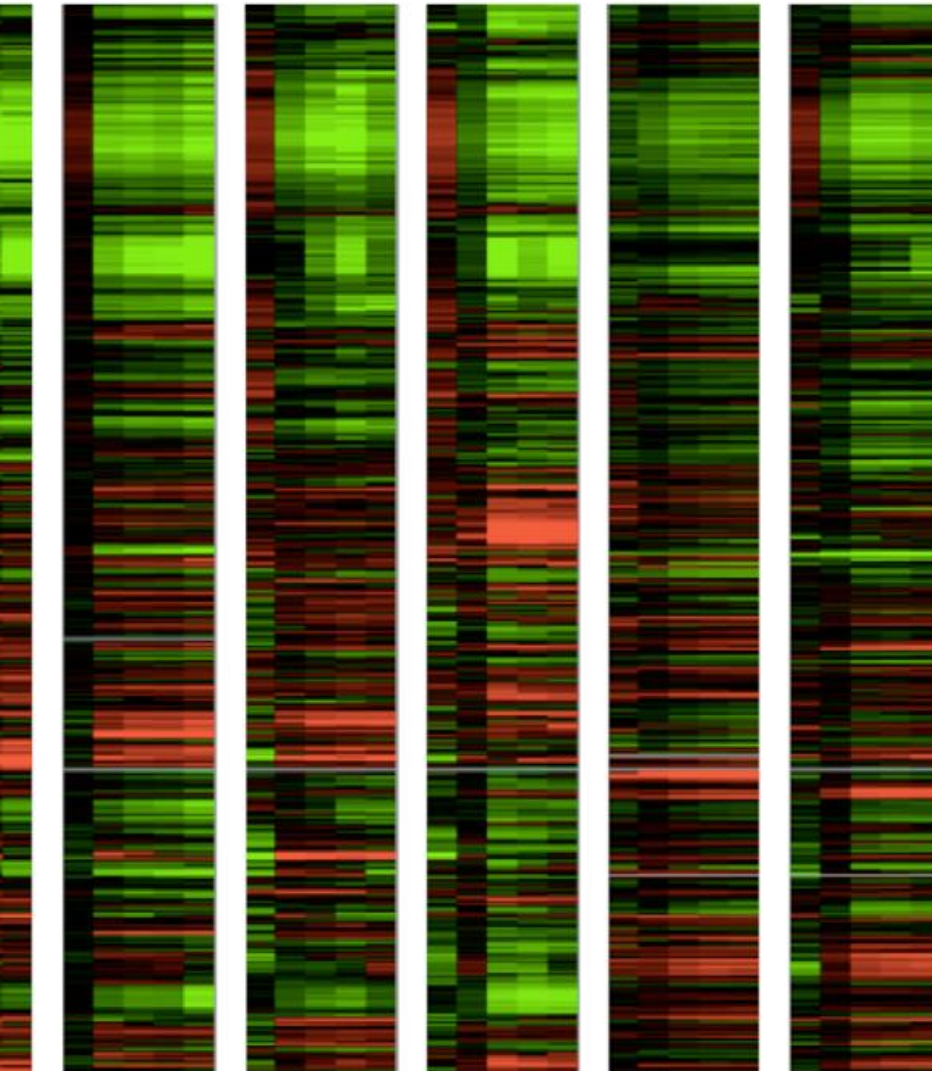
Slope

Area

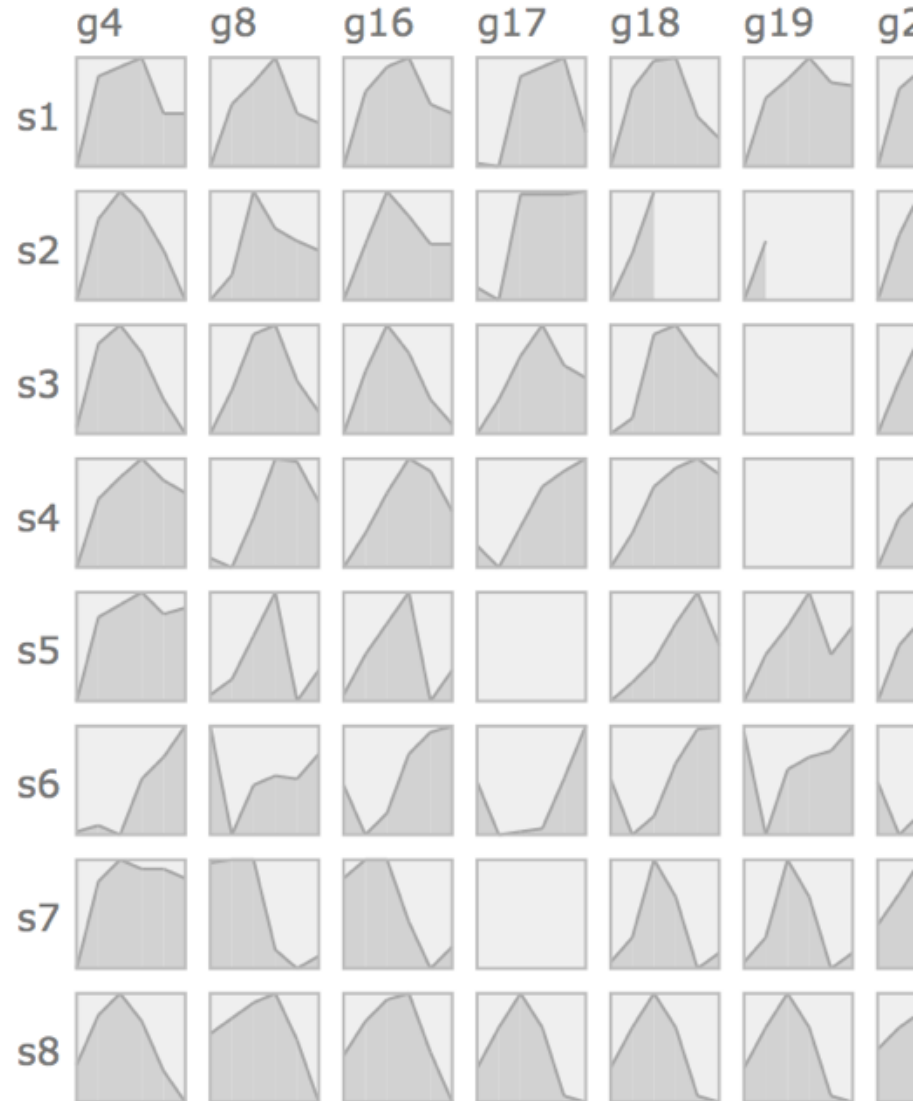
Volume

# Gene Expression Time-Series [Meyer et al '11]

Color Encoding



Position Encoding



# Artery Visualization [Borkin et al '11]

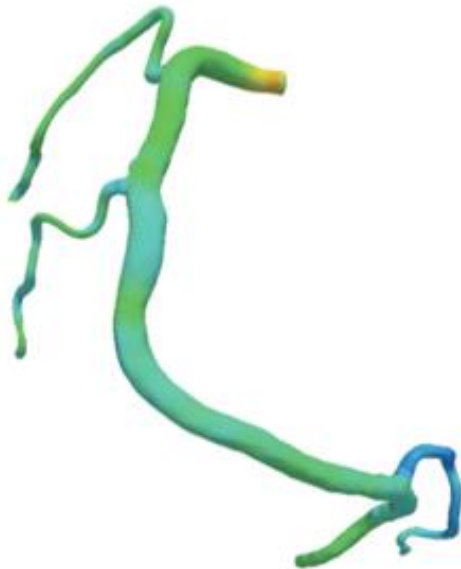
Rainbow Palette

Diverging Palette

2D



3D





# Artery Visualization [Borkin et al '11]

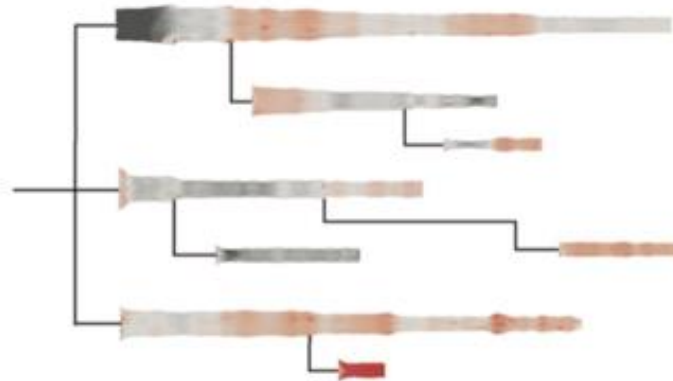
Rainbow Palette

Diverging Palette

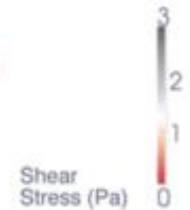
2D



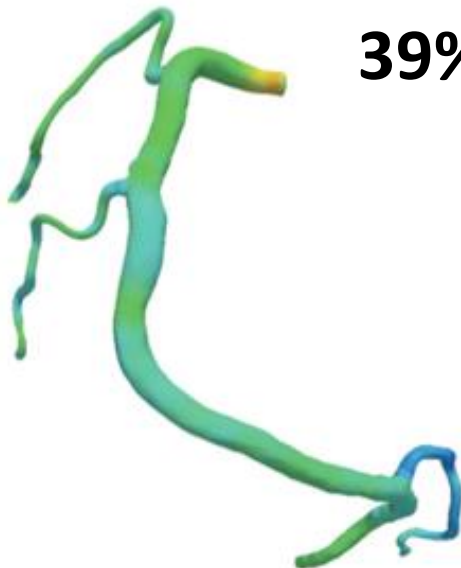
**62%**



**92%**



3D



**39%**



**71%**



# Effectiveness Rankings

## QUANTITATIVE ORDINAL



**Position**

Length

Angle

Slope

Area (Size)

Volume

**Density (Value)**

★ **Color Sat**

~~Color Hue~~

Texture

Connection

Containment

Shape

Position  
Density (Value) Color Hue

Color Sat

Color Hue

Texture

**Connection**

Containment

Length

Angle

Slope

Area (Size)

Volume

Shape

## NOMINAL

Position

Texture

Connection

Containment

Density (Value)

Color Sat

Shape

Length

Angle

Slope

Area

Volume

# Encoding Effectiveness Summary

**Rank** your target data **attributes** by importance.

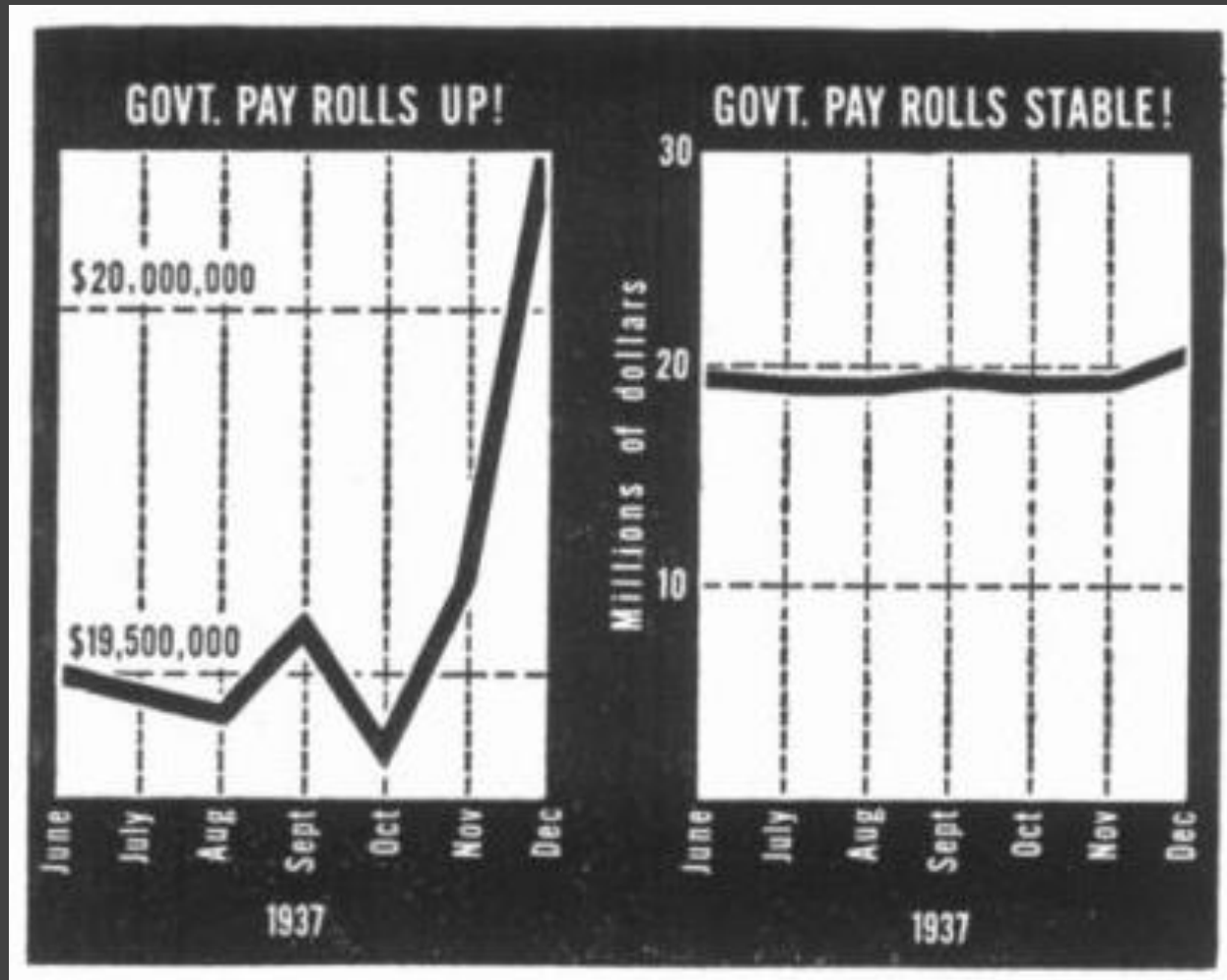
The **most important attributes** should map to their **most effective encoding channels**.

**Effective** charts maximize the readability of the data.

We can use **Mackinlay's rankings** to select better encodings for a given data attribute.

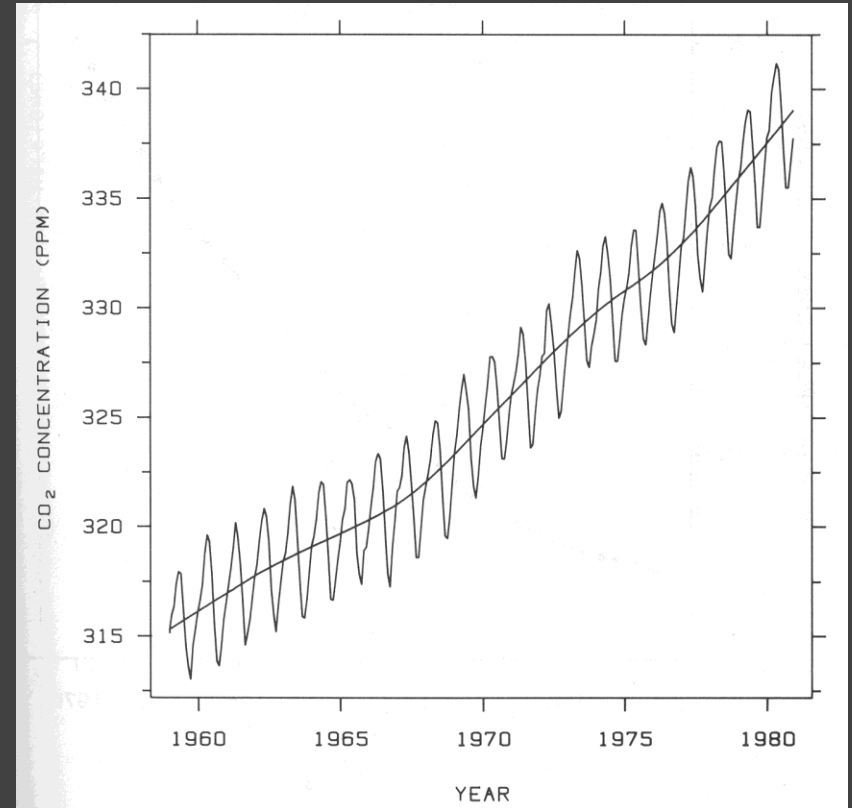
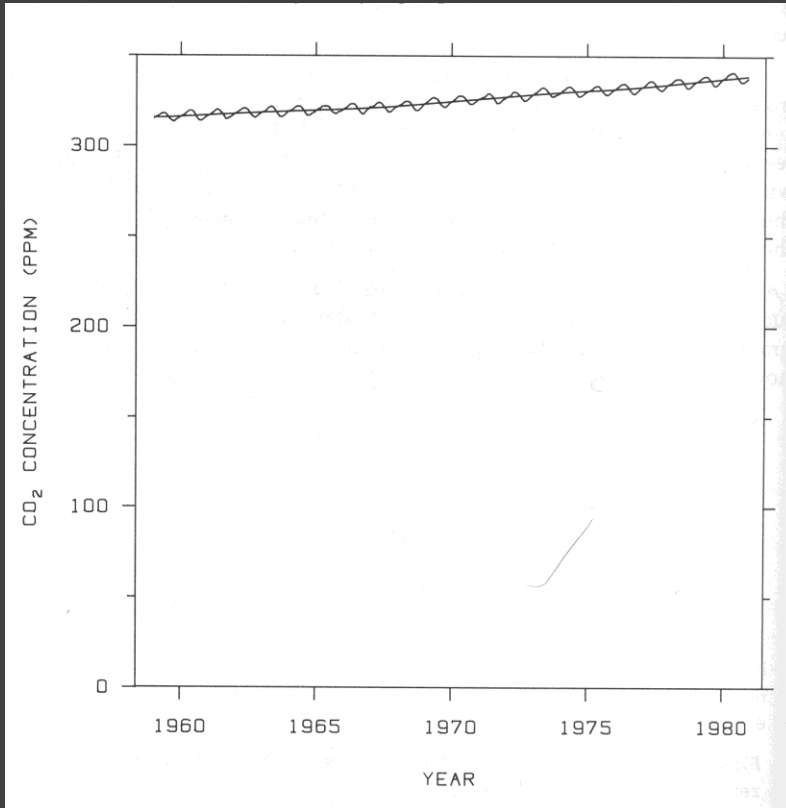
# Scales & Axes

# Include Zero in Axis Scale?



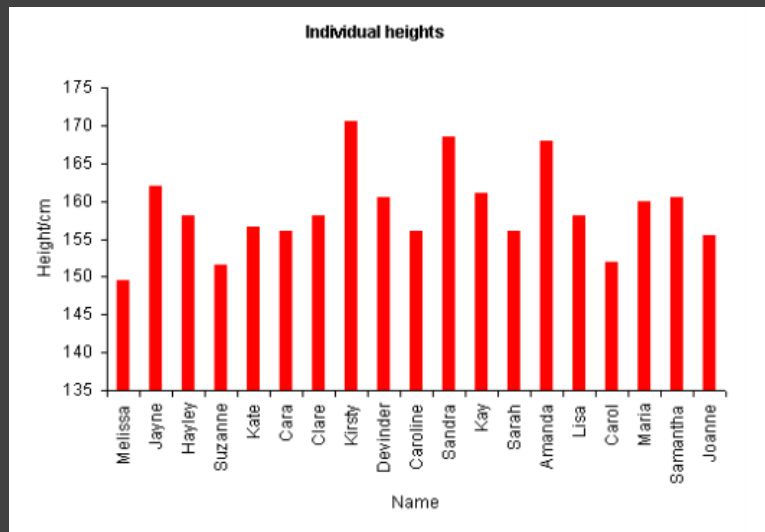
Government payrolls in 1937 [How To Lie With Statistics. Huff]

# Include Zero in Axis Scale?



Yearly CO<sub>2</sub> concentrations [Cleveland 85]

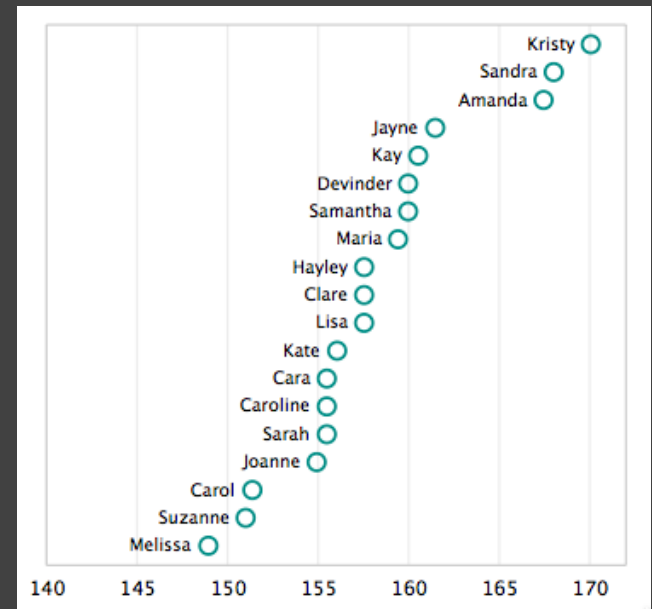
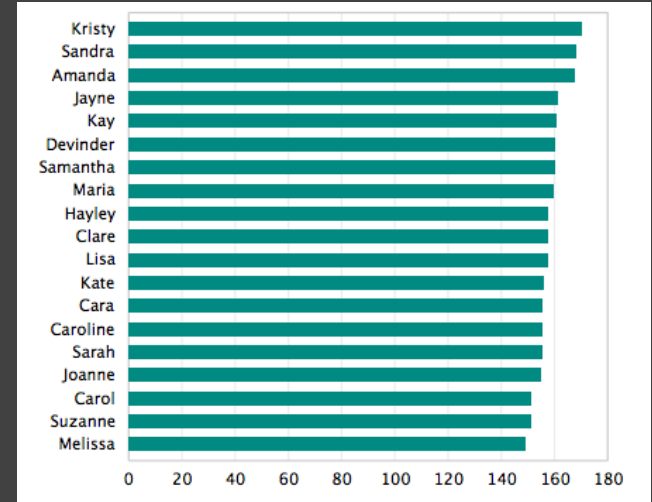
# Include Zero in Axis Scale?



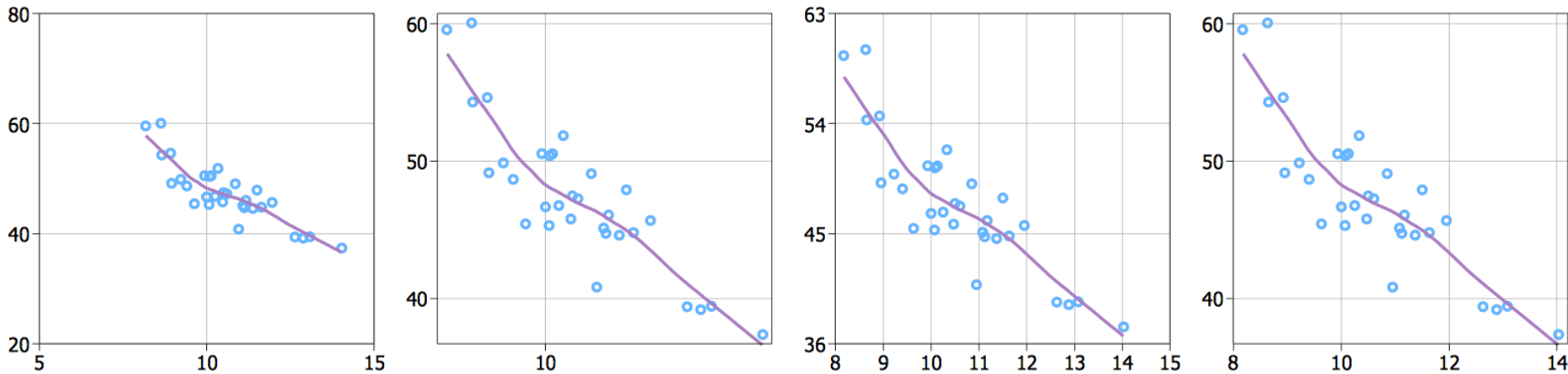
Violates Expressiveness Principle!

Compare Proportions (Q-Ratio)

Compare Relative Position (Q-Interval)

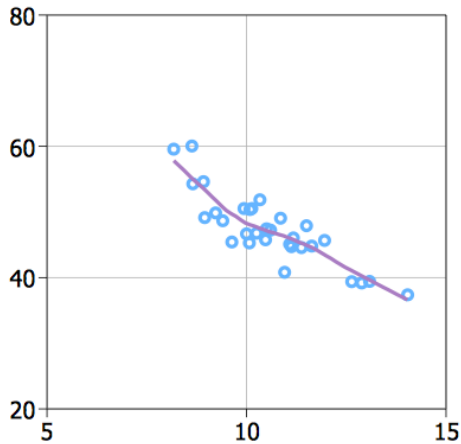


# Axis Tick Mark Selection

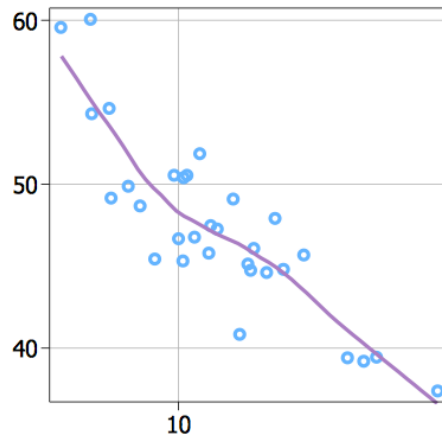


What are some properties of “good” tick marks?

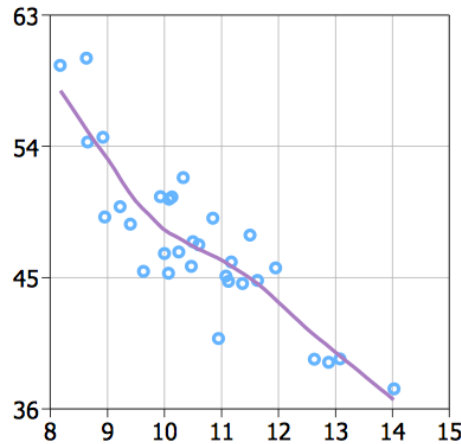
# Axis Tick Mark Selection



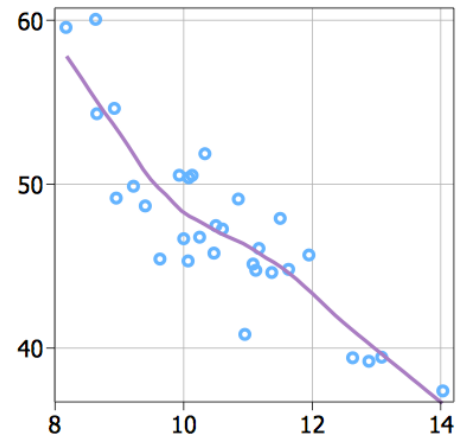
(a) Heckbert



(b) R's pretty



(c) Wilkinson



(d) Extended

**Simplicity** - numbers are multiples of 10, 5, 2

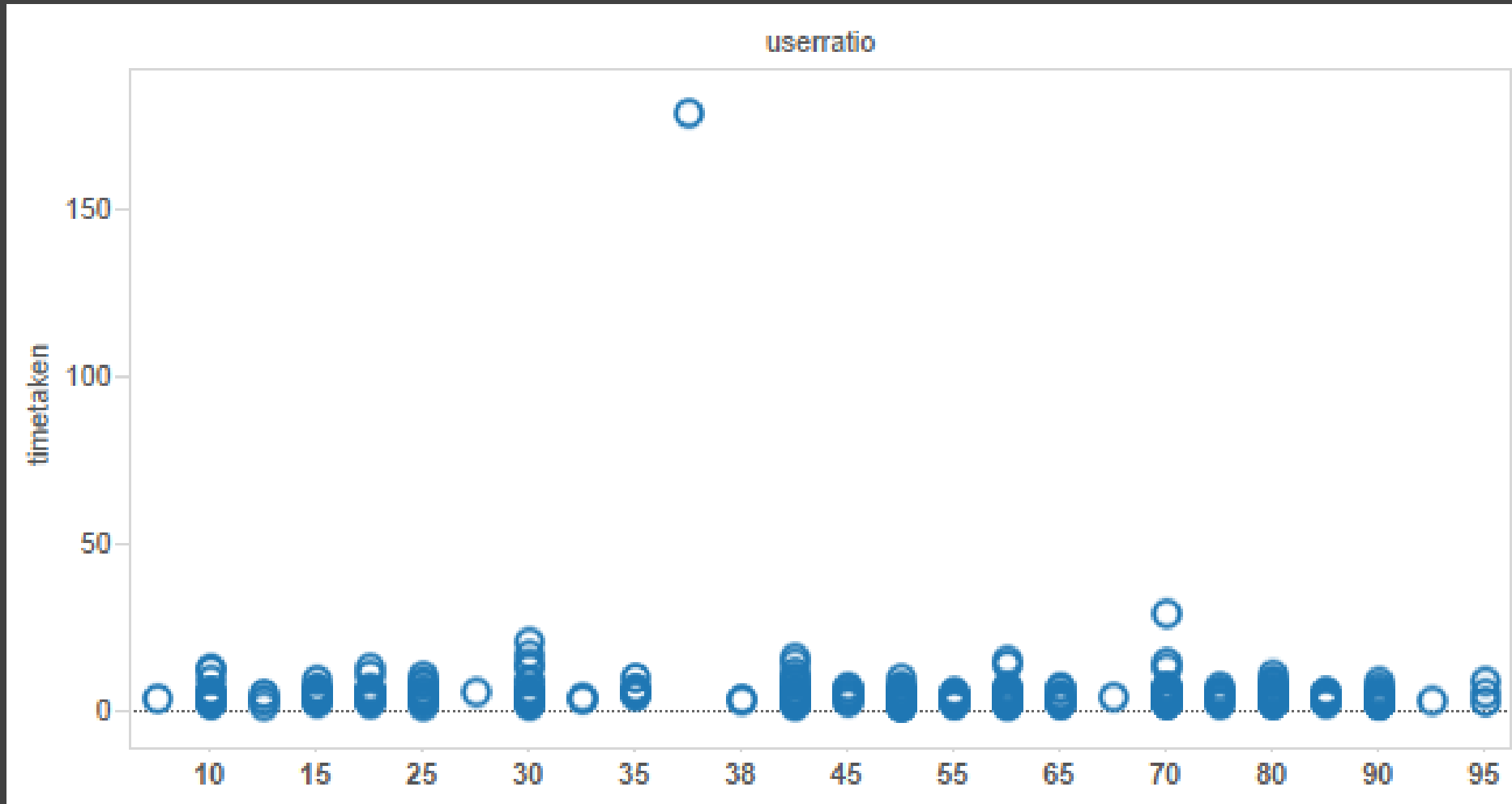
**Coverage** - ticks near the ends of the data

**Density** - not too many, nor too few

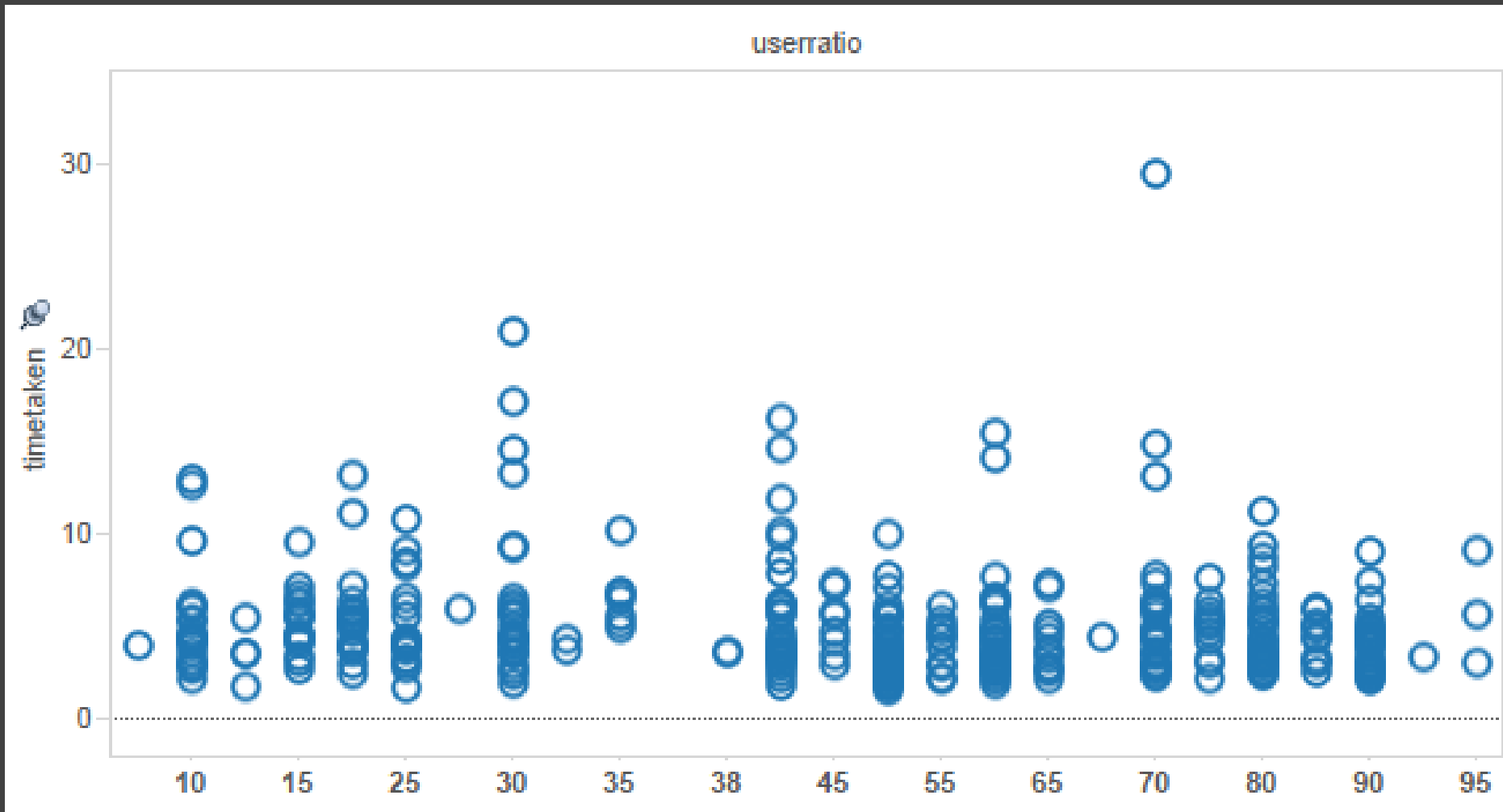
**Legibility** - whitespace, horizontal text, size



# How to Scale the Axis?

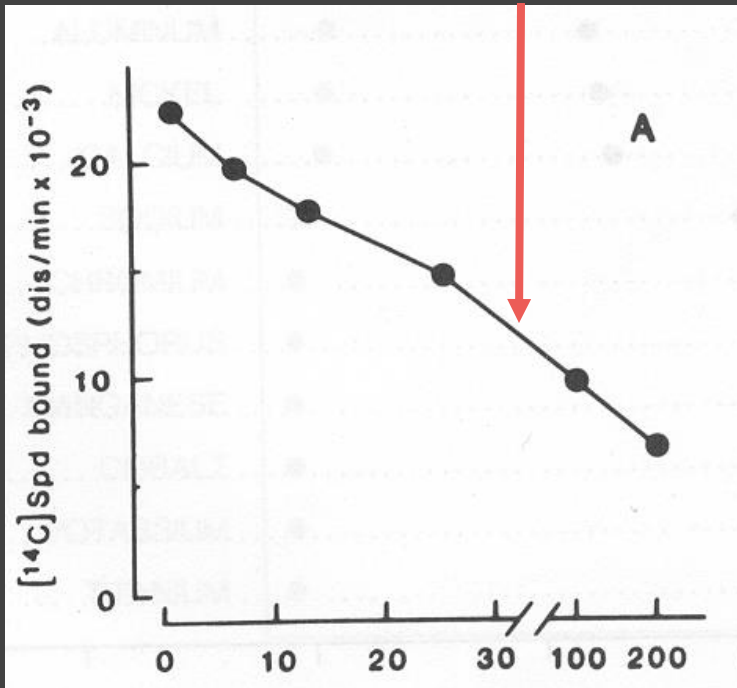


# One Option: Clip Outliers

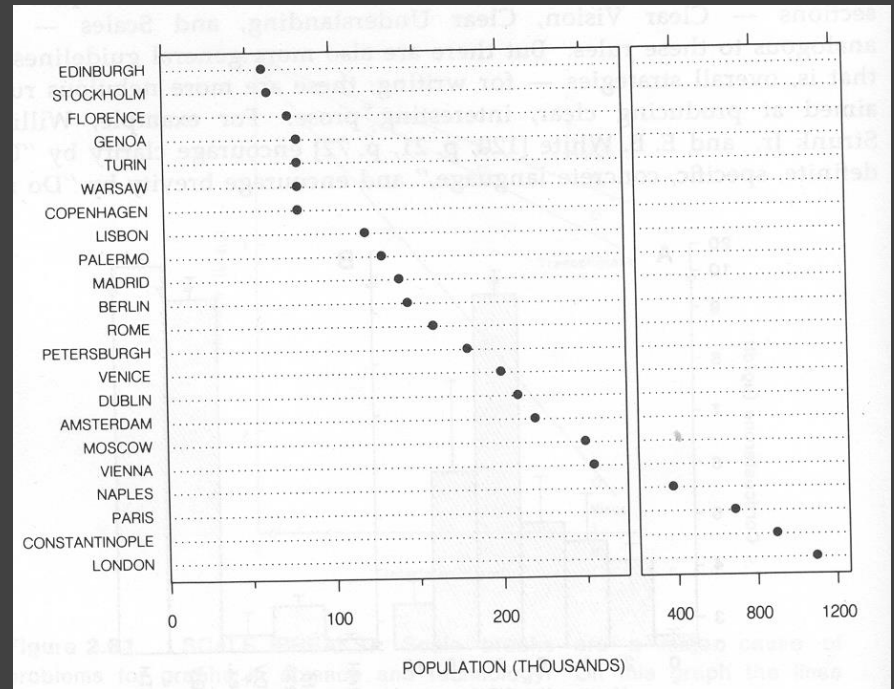


# Clearly Mark Scale Breaks

Violates Expressiveness Principle!

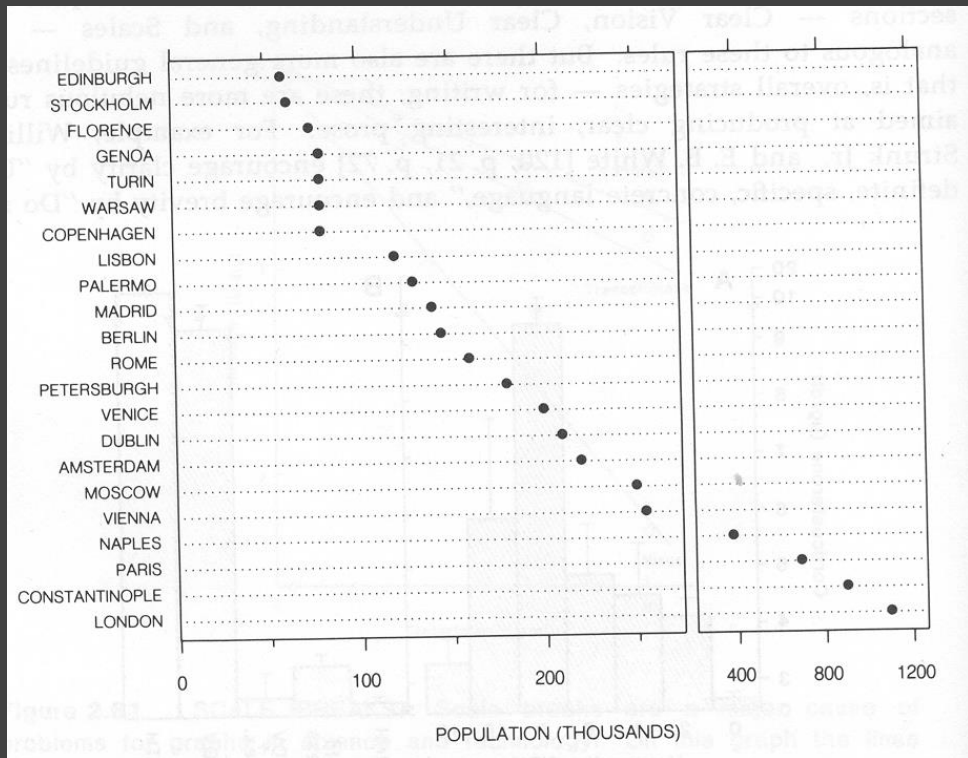


Poor scale break [Cleveland 85]

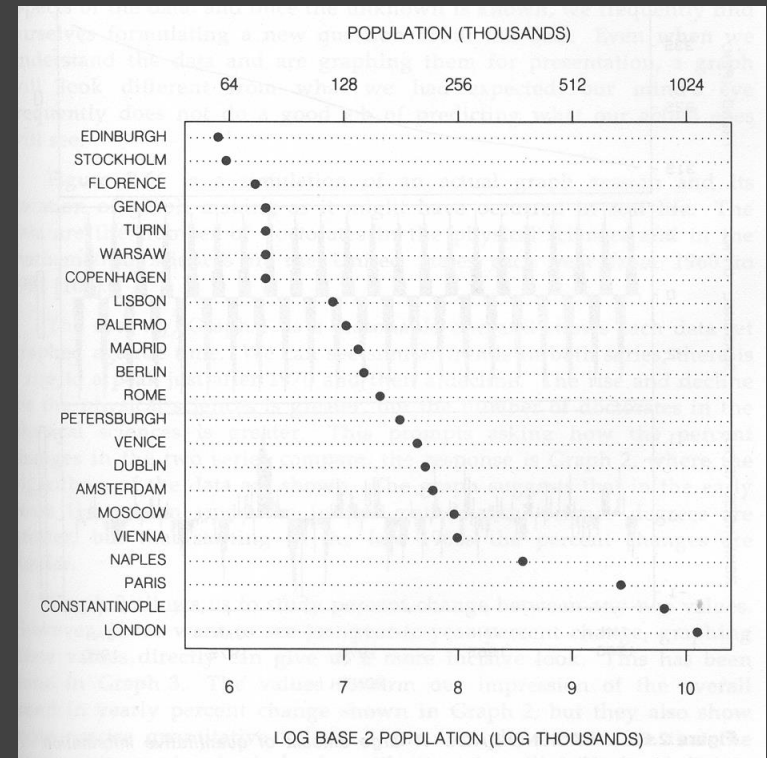


Well-marked scale break [Cleveland 85]

# Scale Break vs. Log Scale

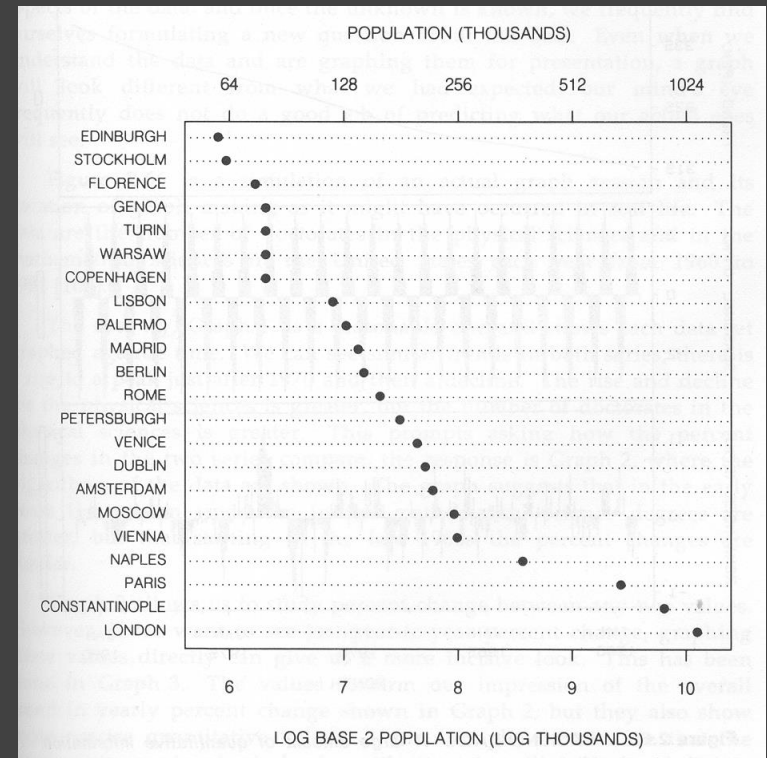
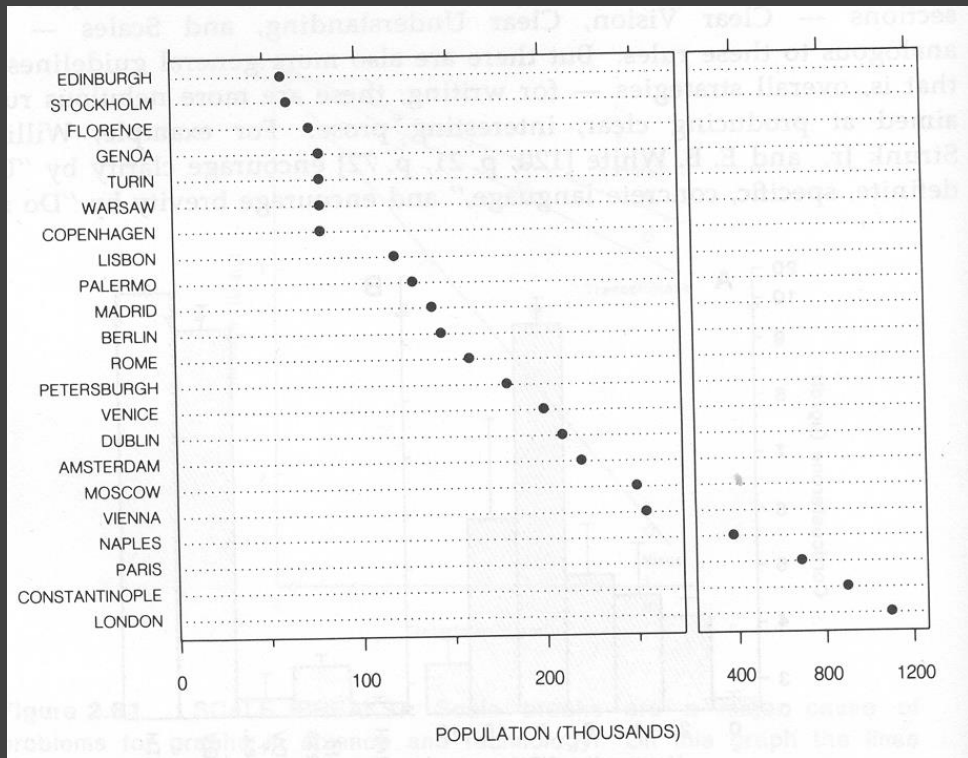


Scale Break



Log Scale

# Scale Break vs. Log Scale



**Both increase visual resolution**

Scale break: difficult to compare (*cognitive* — not *perceptual* — work)

Log scale: direct comparison of all data

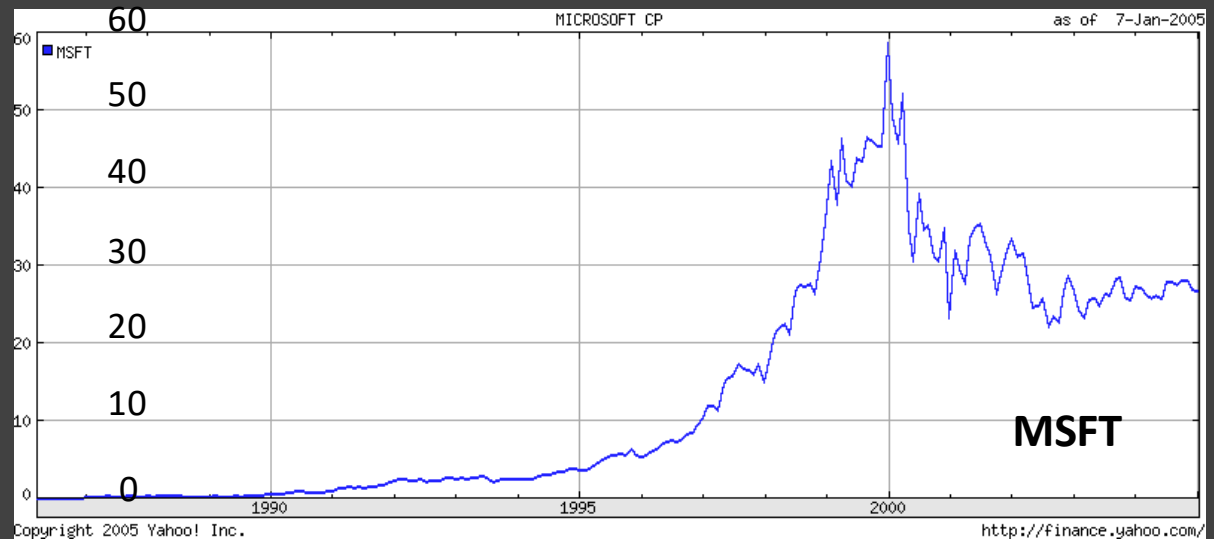
Logarithms turn *multiplication* into *addition*.

$$\log(x \cdot y) = \log(x) + \log(y)$$

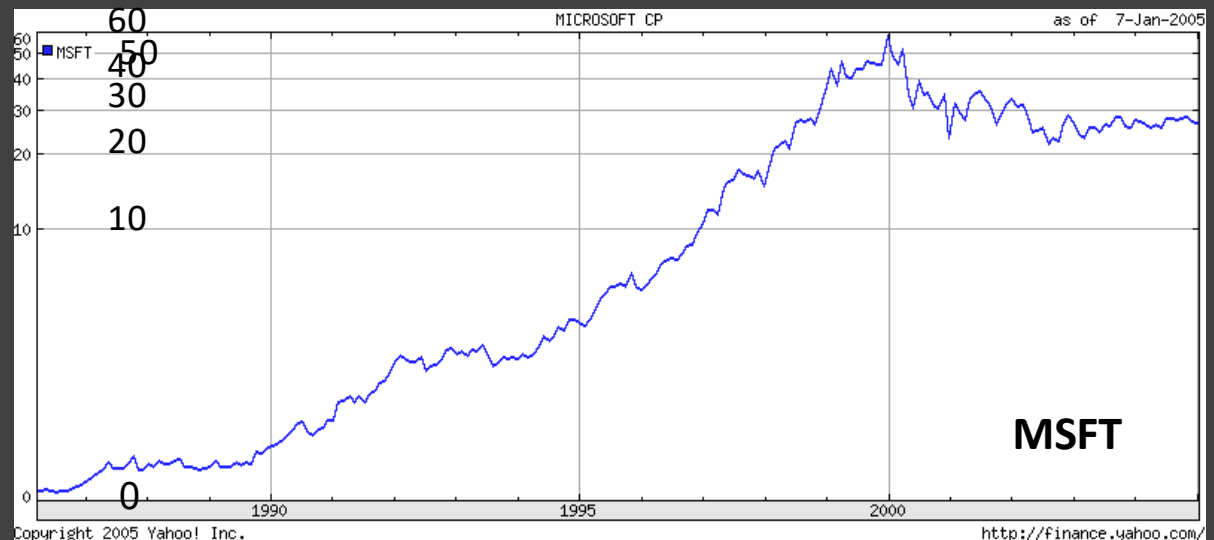
Equal steps on a log scale  
correspond to equal changes to a  
multiplicative scale factor.

# Linear Scale vs. Log Scale

## Linear Scale



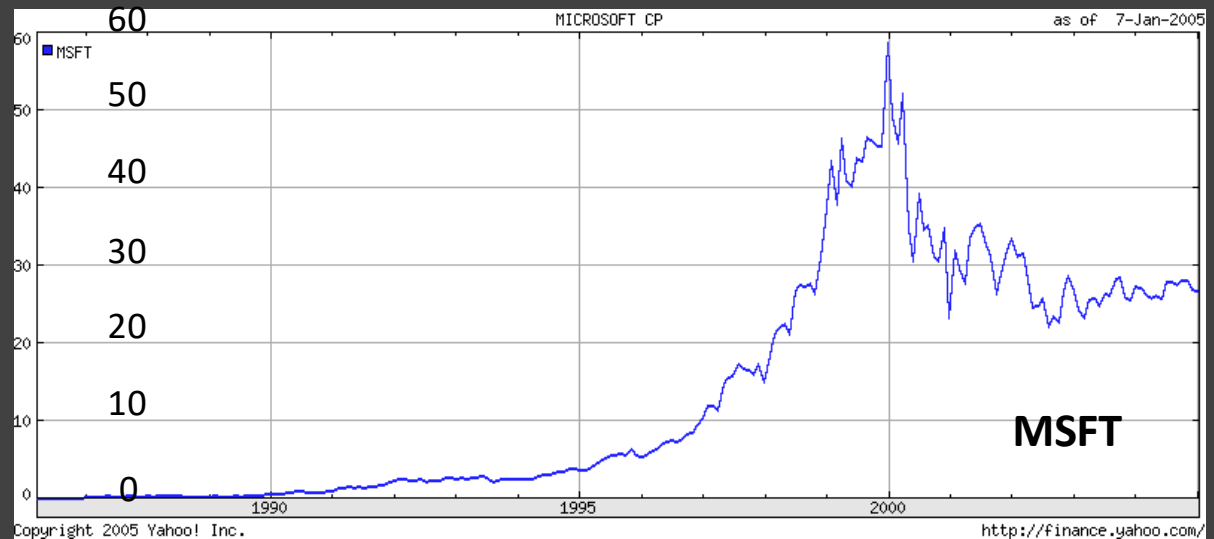
## Log Scale



# Linear Scale vs. Log Scale

## Linear Scale

Absolute change

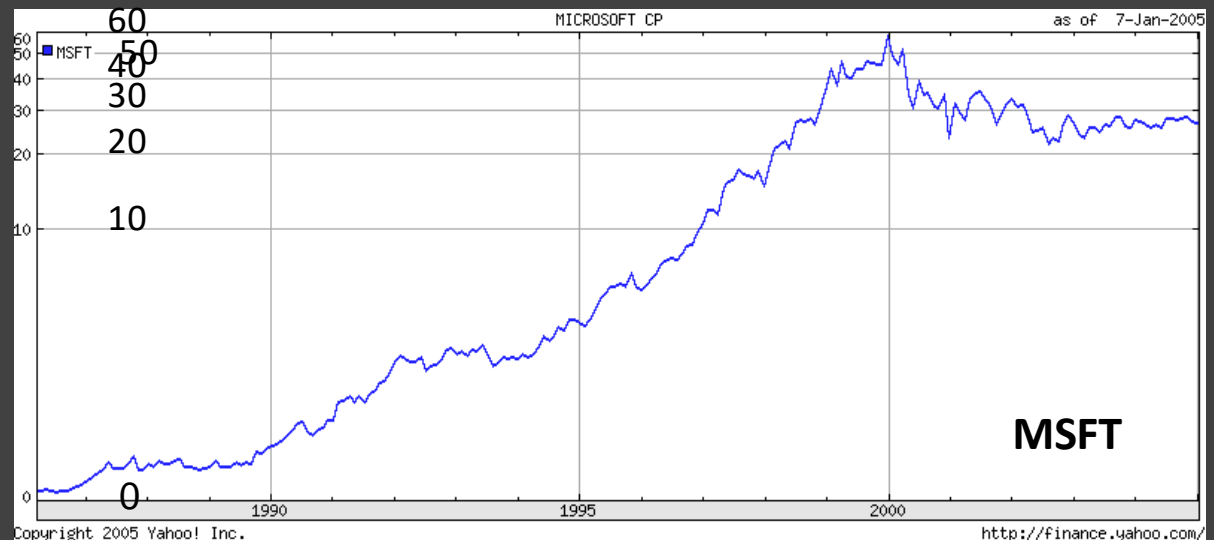


## Log Scale

Small fluctuations

Percent change

$$d(10,30) > d(30,60)$$

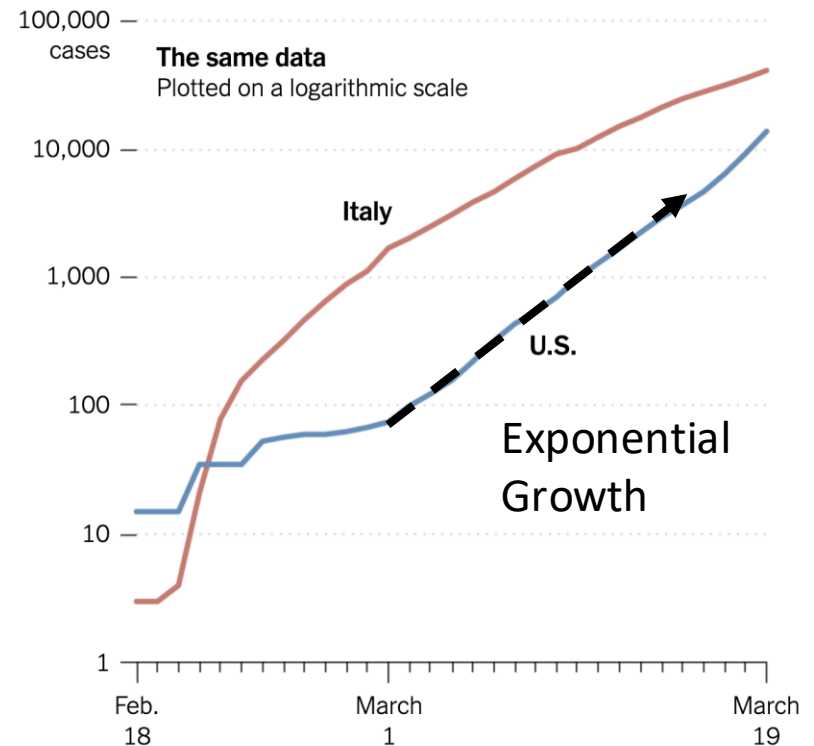
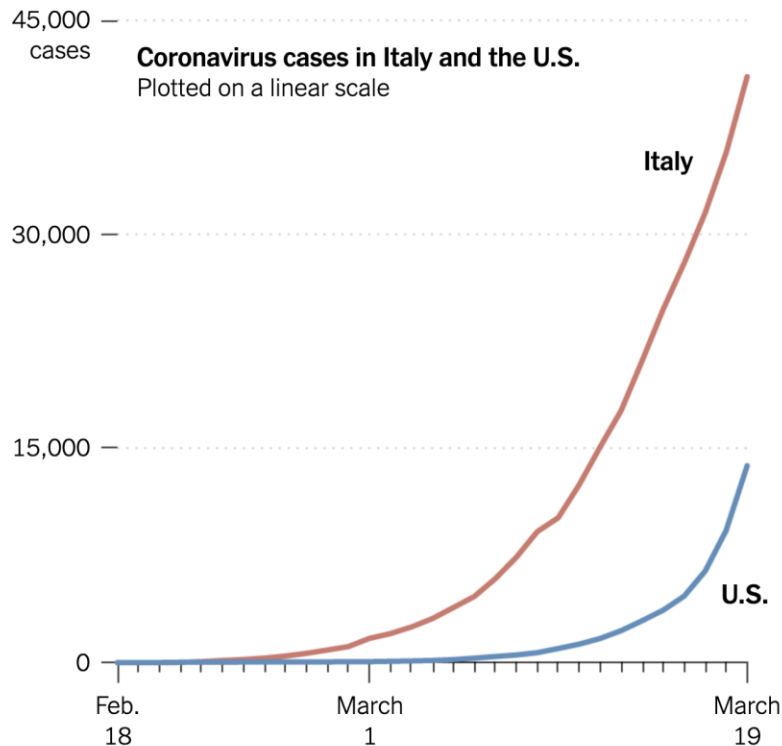




# Linear Scale vs. Log Scale

## Bending the Curve

Logarithmic scales can emphasize the rate of change in a way that linear scales do not. Italy seems to be slowing the coronavirus infection rate, while the number of cases in the United States continues to double every few days.



# When To Apply a Log Scale?

Address data skew (e.g., long tails, outliers)

Enables comparison within and across multiple orders of magnitude.

Focus on multiplicative factors (not additive)

Recall that the logarithm transforms  $\times$  to  $+$ !

Percentage change, not linear difference.

Constraint: positive, non-zero values

Constraint: audience familiarity?

# Scales & Axes Summary

**Zero-based scales** help viewers interpret individual **Q-Ratio** data values.

**Grid lines** help viewers compare multiple data values.

**Data outliers** and **dataset skew** can be displayed more easily using **logarithmic scales**.

However, **logarithmic scales** can be **hard to read**, especially for **non-expert** audiences.

Break Time!

# Multidimensional Data

# Visual Encoding Variables

Position (X)

Position (Y)

Area

Value

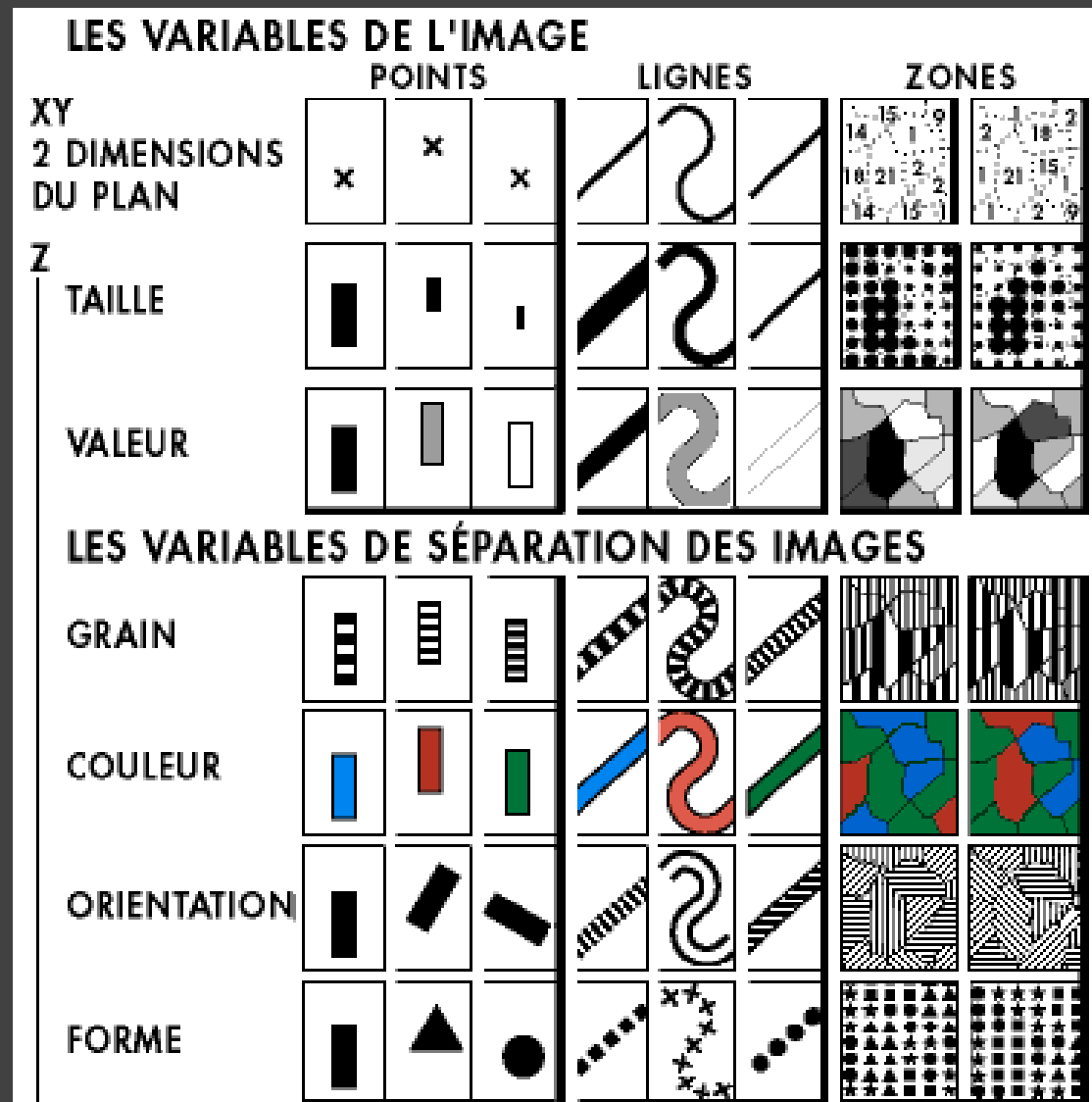
Texture

Color

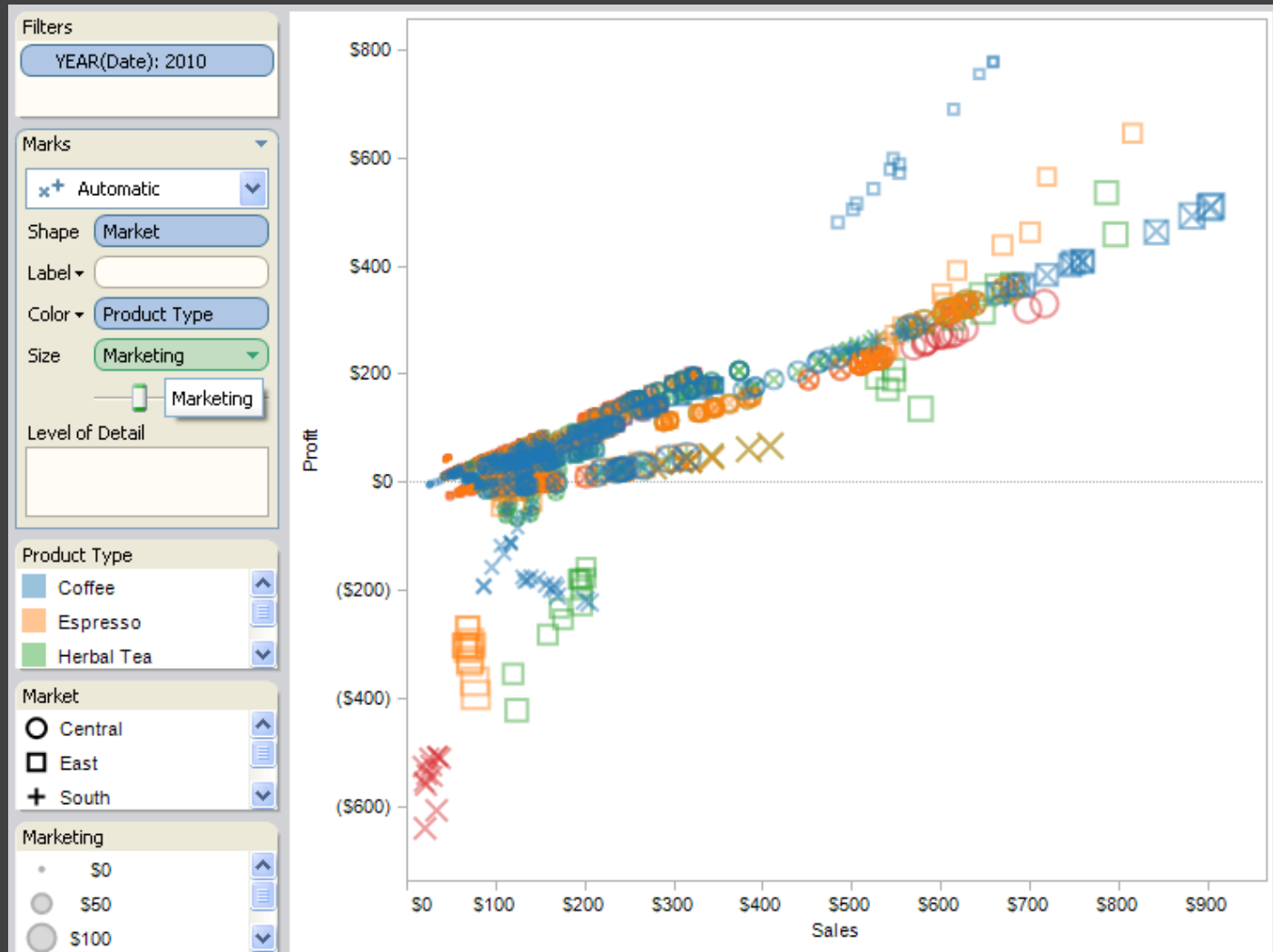
Orientation

Shape

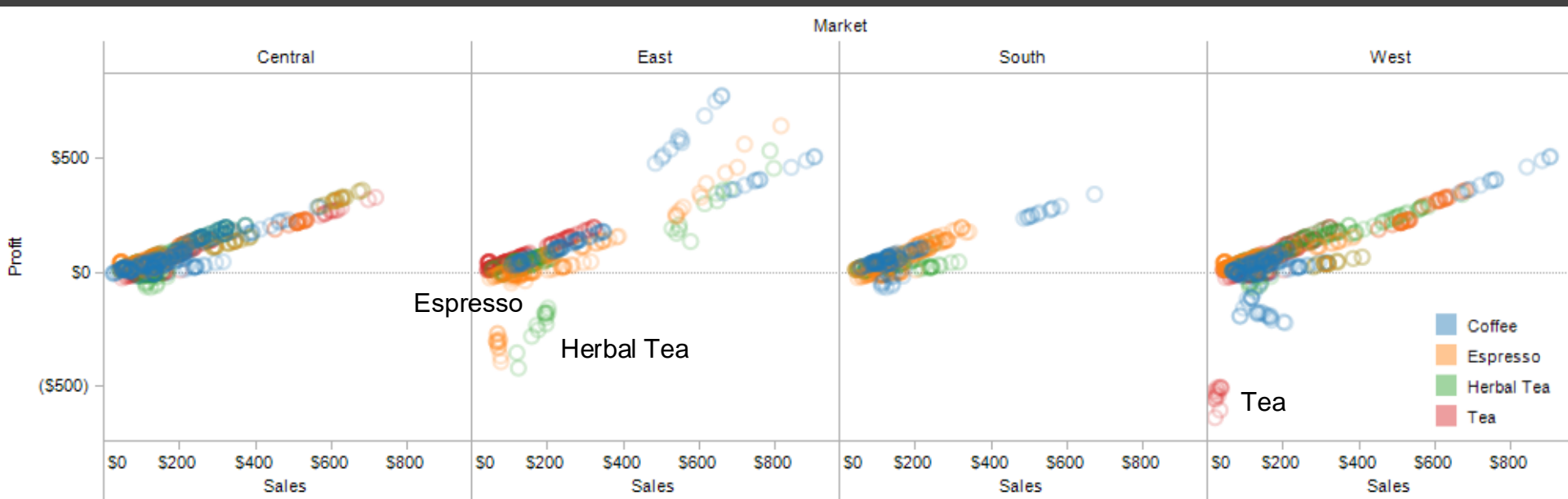
~8 dimensions?



# Too Many Encodings Cause Clutter



# Trellis Plots Subdivide the Data



A *trellis plot* subdivides space to enable comparison across multiple plots.

Typically nominal or ordinal variables are used as dimensions for subdivision.



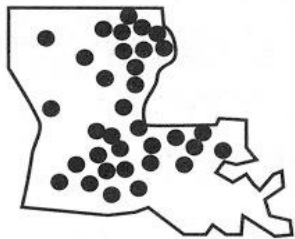
# Small Multiples Generalize Trellis Plots



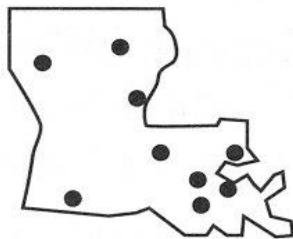
[MacEachren '95, Figure 2.11, p. 38]

# Small Multiples Generalize Trellis Plots

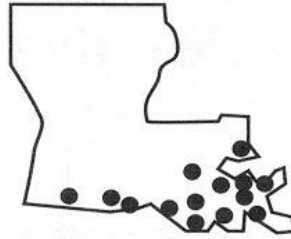
alfisol



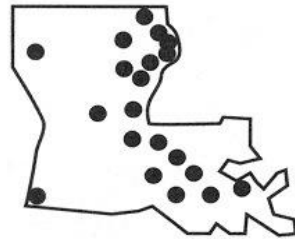
entisol



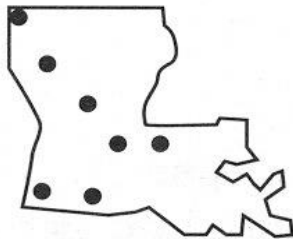
histosol



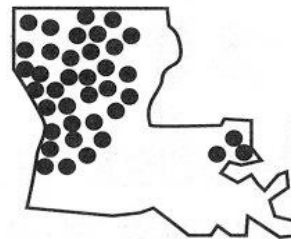
inceptisol



mollisol

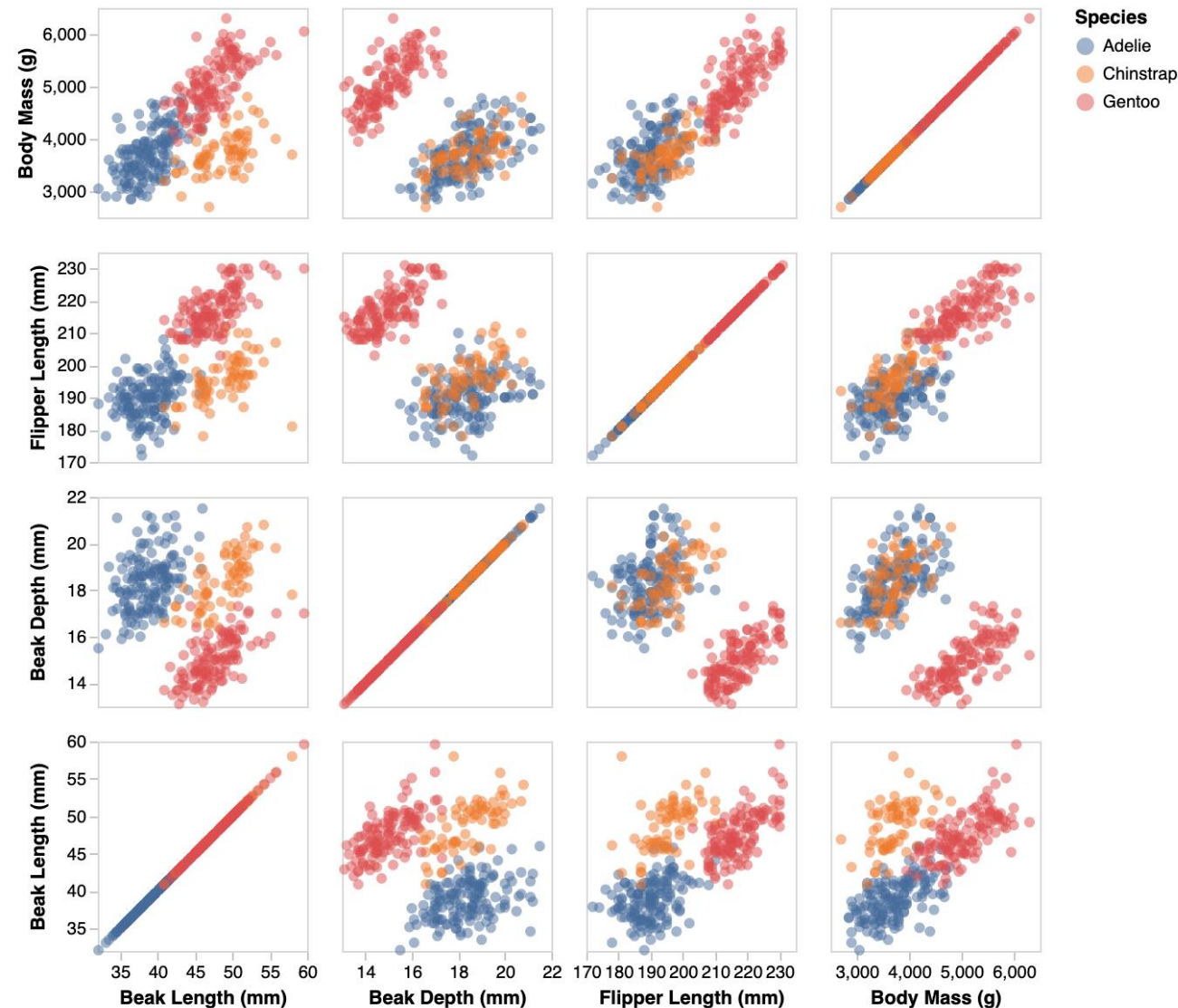


ultisol



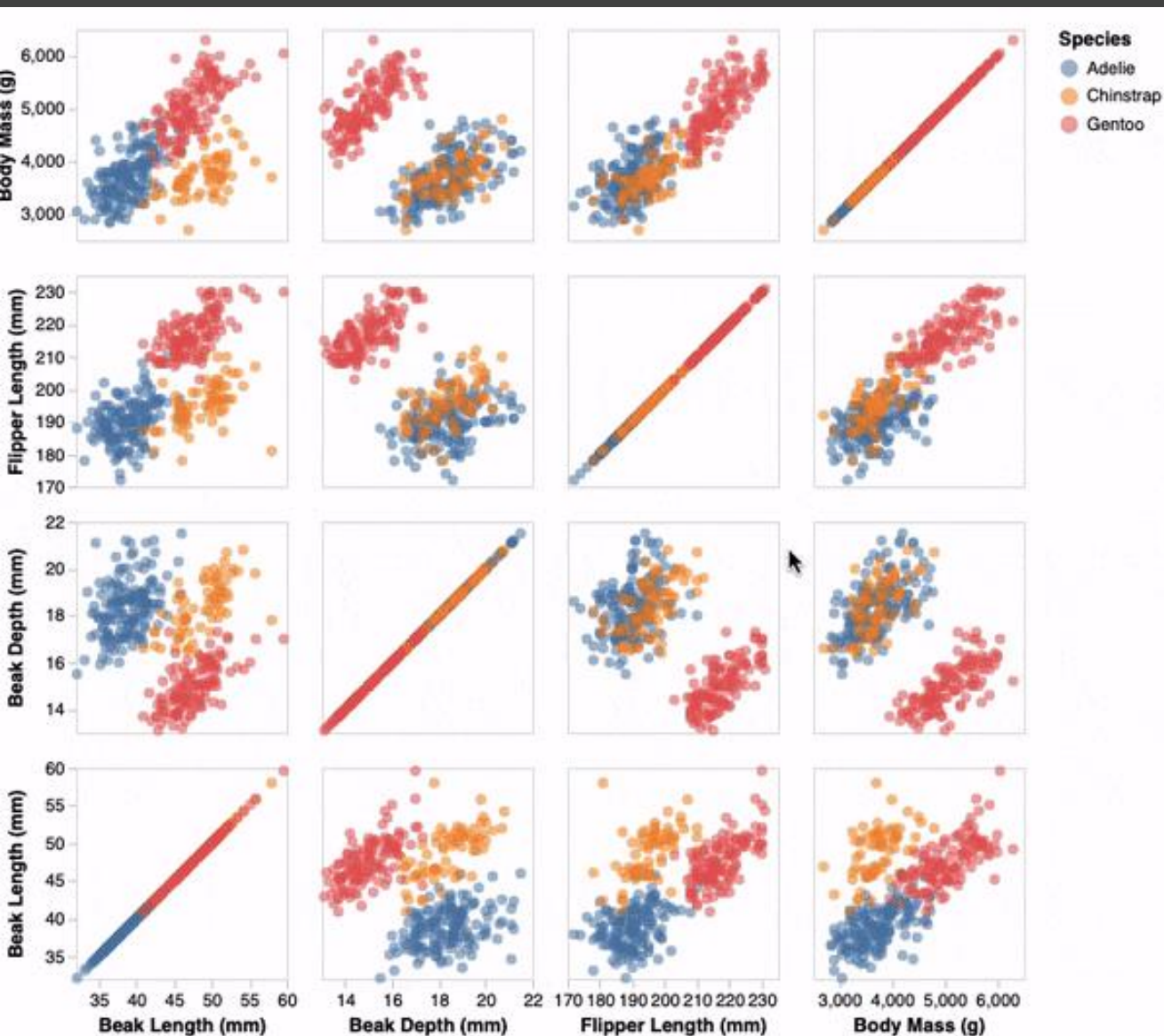
[MacEachren '95, Figure 2.11, p. 38]

# Scatterplot Matrix (SPLOM)



Scatter plots for pairwise comparison of each data dimension.

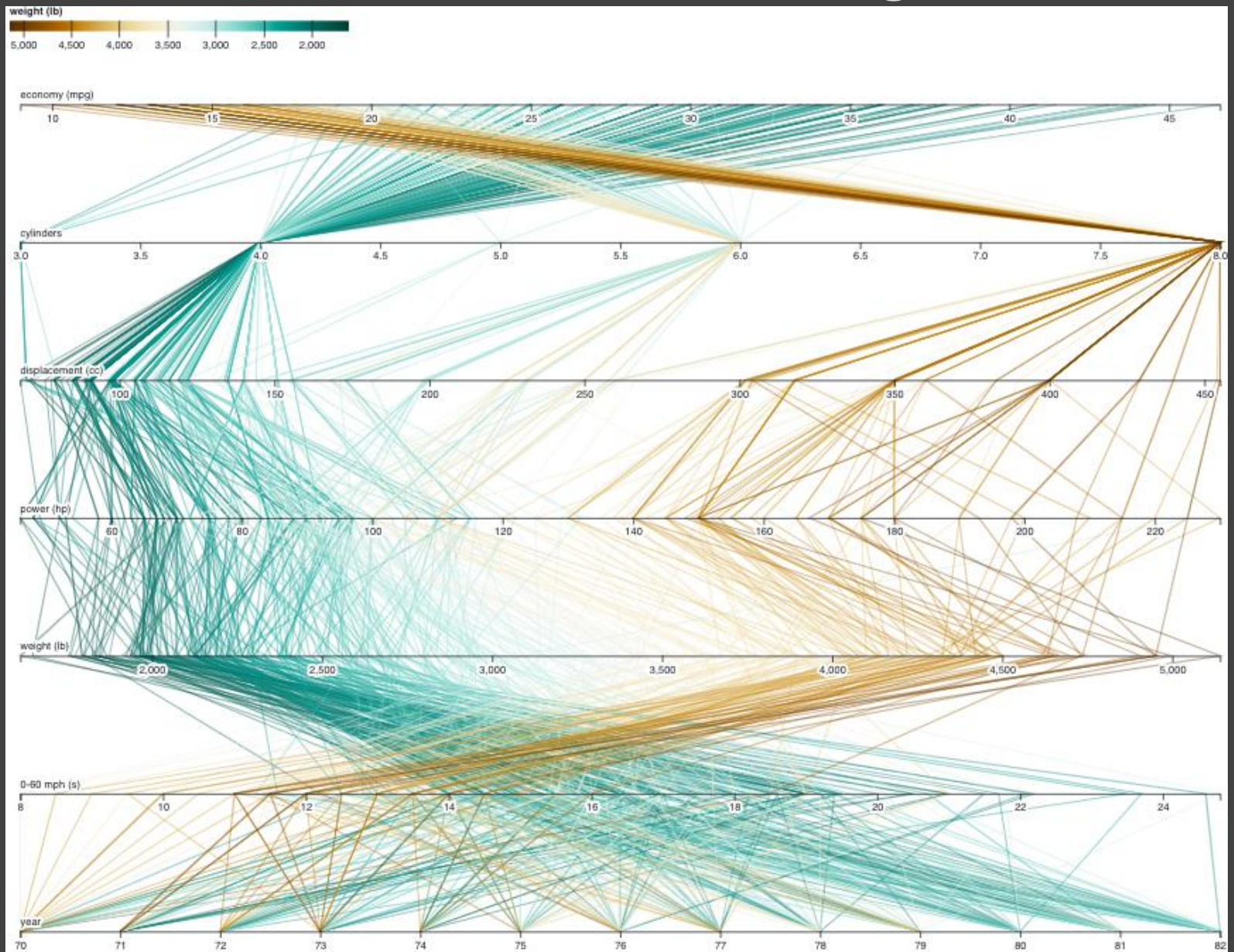
# Scatterplot Matrix (SPLOM)



Scatter plots for pairwise comparison of each data dimension.



# Parallel Coordinates [Inselberg]



# Parallel Coordinates [Inselberg]

Visualize up to ~two dozen dimensions at once

1. Draw parallel axes for each variable
2. For each tuple, connect points on each axis

Between adjacent axes: line crossings imply neg. correlation, shared slopes imply pos. correlation.

Full plot can be cluttered. **Interactive selection** can be used to assess multivariate relationships.

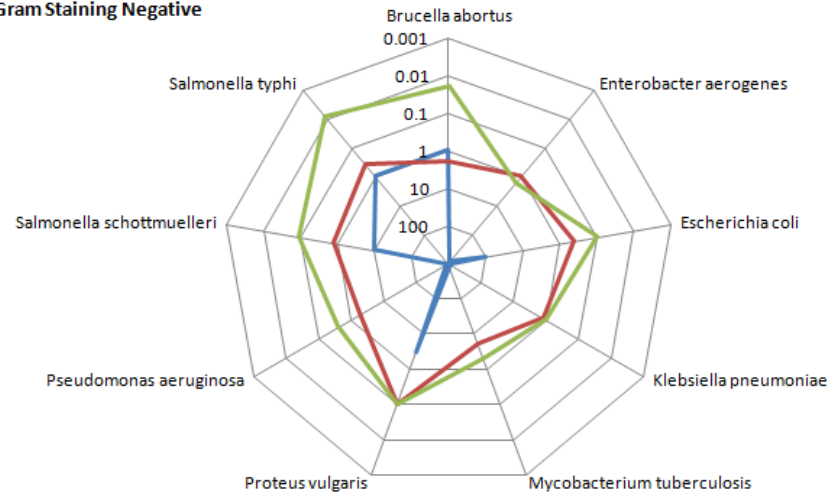
Highly sensitive to axis **scale** and **ordering**.

Expertise required to use effectively!

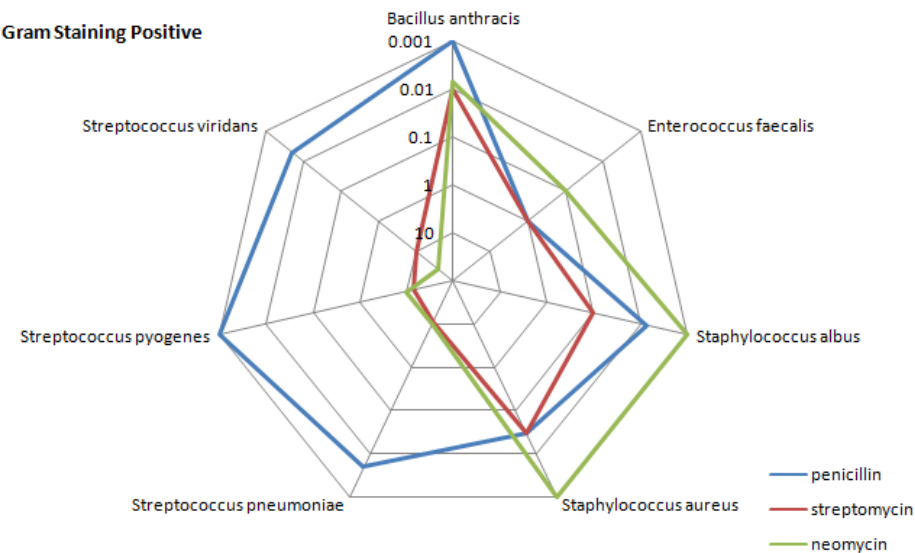
# Radar Plot / Star Graph

Antibiotics MIC Concentrations

Gram Staining Negative



Gram Staining Positive



“Parallel” dimensions in polar coordinate space  
Best if same units apply to each axis

# Multidimensional Data Summary

Rendering **more than three encodings** tends to produce a **cluttered visualization**.

**Small multiples** subdivide a **visualization** into more easily readable partitions.

Trellis plots and scatterplot matrices (SPLOMS) are examples of small multiples.

**Parallel coordinates** plots can render many attributes in one visualization but **require expertise** to interpret.



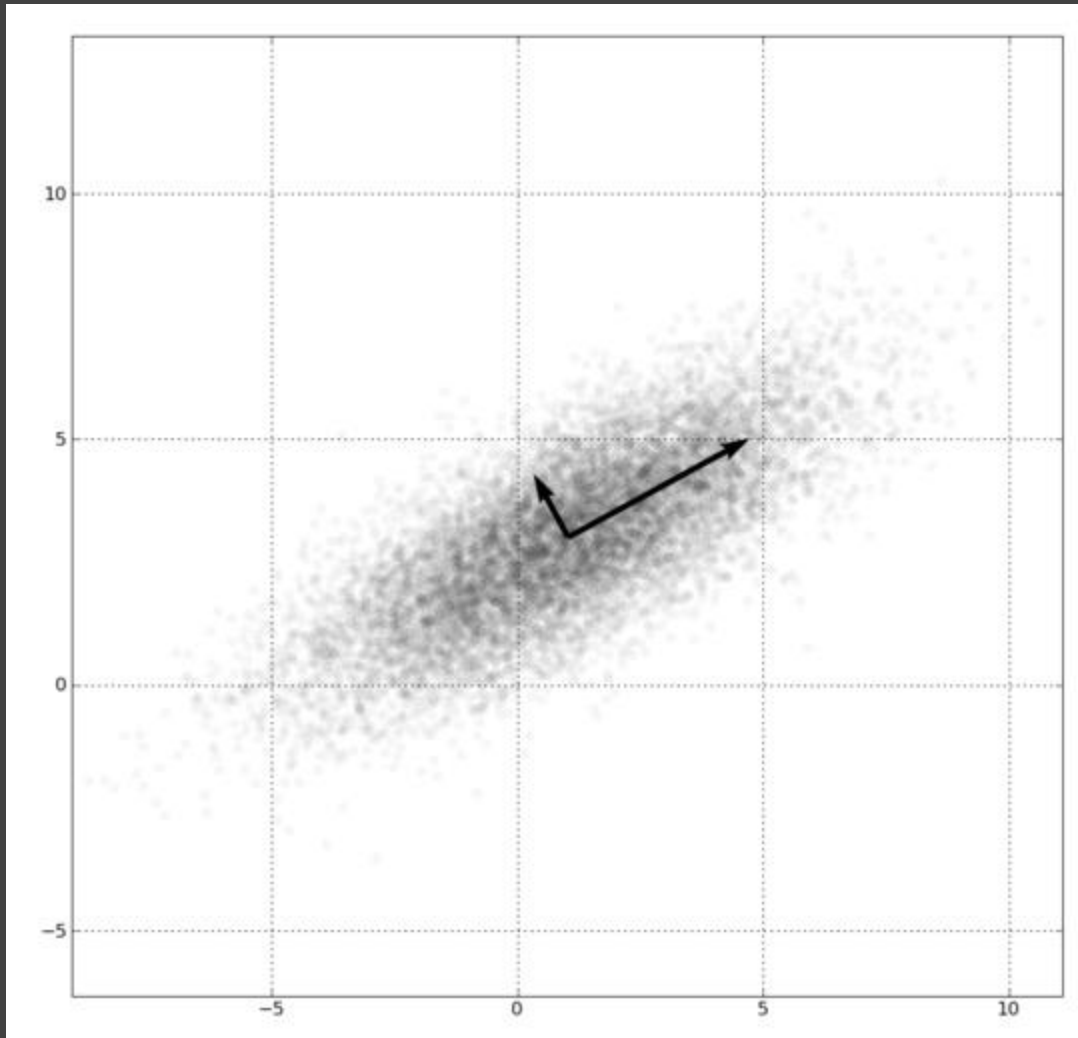
# Dimensionality Reduction

# Dimensionality Reduction (DR)

Project  $n$ D data to 2D or 3D for viewing. Often used to interpret and sanity check high-dimensional representations fit by machine learning methods.

Different DR methods make different trade-offs: for example to preserve global structure (e.g., PCA) or emphasize local structure (e.g., nearest-neighbor approaches, including t-SNE and UMAP).

# Principal Components Analysis

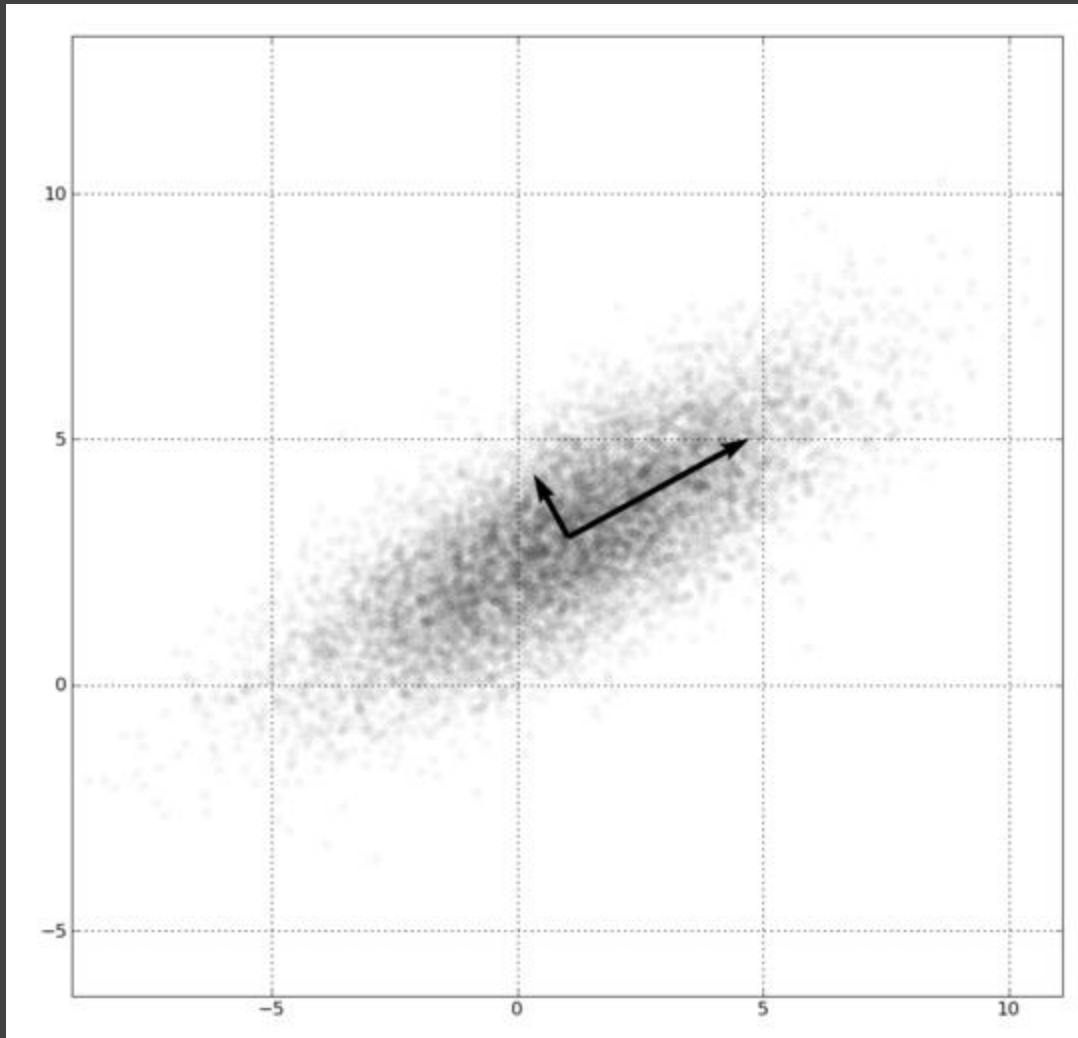


1. Mean-center the data.

2. Find  $\perp$  basis vectors that maximize the data variance.

3. Plot the data using the top vectors.

# Principal Components Analysis

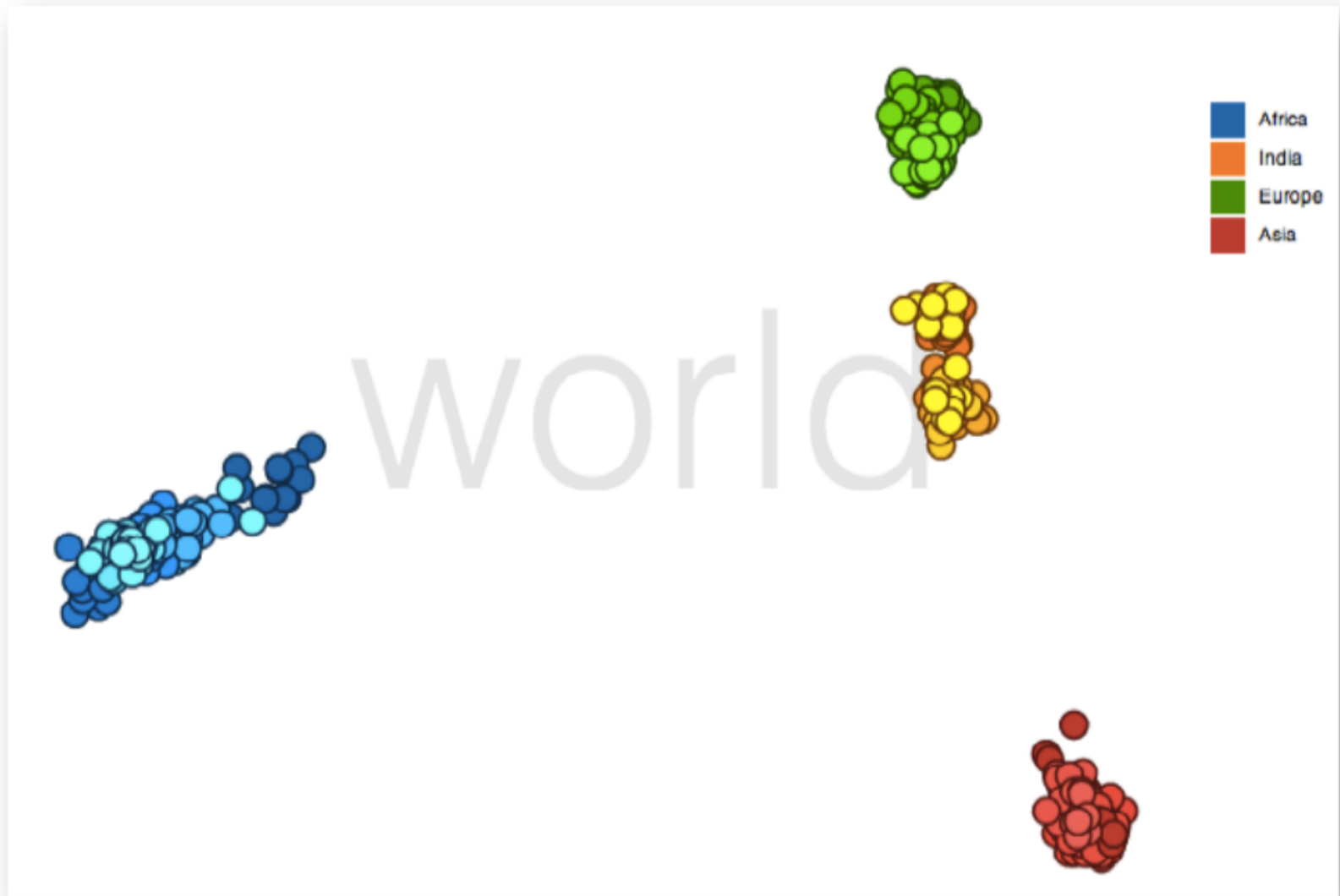


Linear transform:  
scale and rotate  
original space.

Lines (vectors)  
project to lines.

Preserves global  
distances.

# PCA of Genomes [Demiralp et al. '13]



# Reduction Techniques

## **LINEAR - PRESERVE GLOBAL STRUCTURE**

Principal Components Analysis (PCA)

Linear transformation of basis vectors, ordered by amount of data variance they explain.

## **NON-LINEAR - PRESERVE LOCAL TOPOLOGY**

t-Dist. Stochastic Neighbor Embedding (t-SNE)

Probabilistically model distance, optimize positions.

Uniform Manifold Approx. & Projection (UMAP)

Identify local manifolds, then stitch them together.

# Non-Linear Techniques

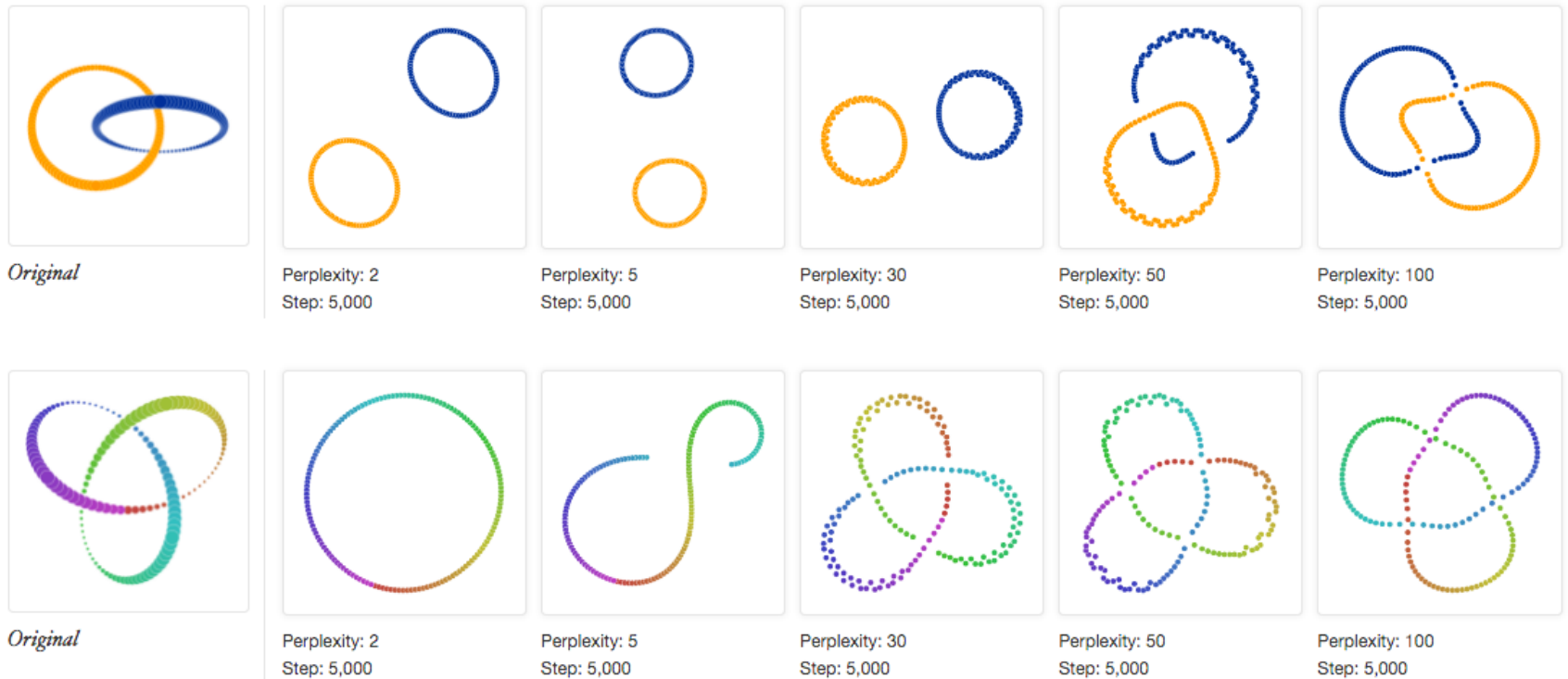
Distort the space, trade-off preservation of global structure to emphasize local neighborhoods. Use topological (nearest neighbor) analysis.

Two popular contemporary methods:

t-SNE - probabilistic interpretation of distance

UMAP - tries to balance local/global trade-off

# Visualizing t-SNE [Wattenberg et al. '16]

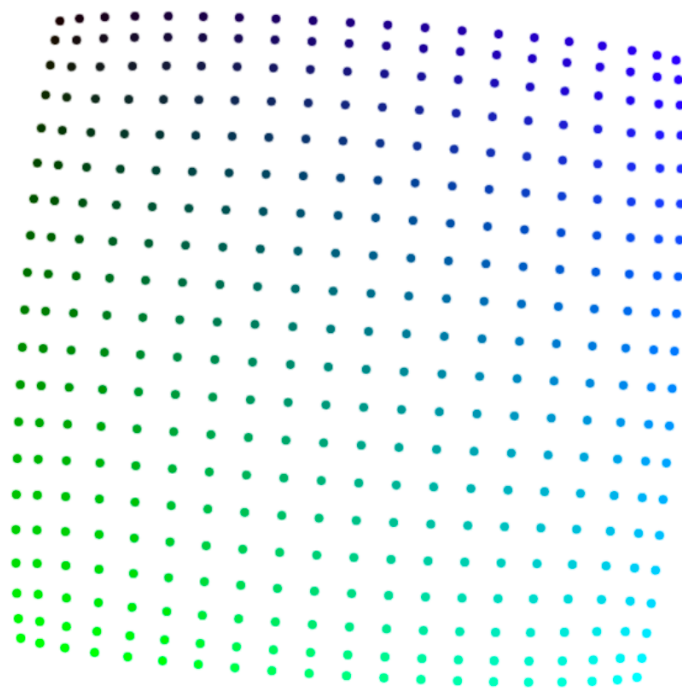


Results can be highly sensitive to the algorithm parameters!



# How to Use t-SNE Effectively

Although extremely useful for visualizing high-dimensional data, t-SNE plots can sometimes be mysterious or misleading. By exploring how it behaves in simple cases, we can learn to use it more effectively.





 Step  
1,910

Points Per Side 20

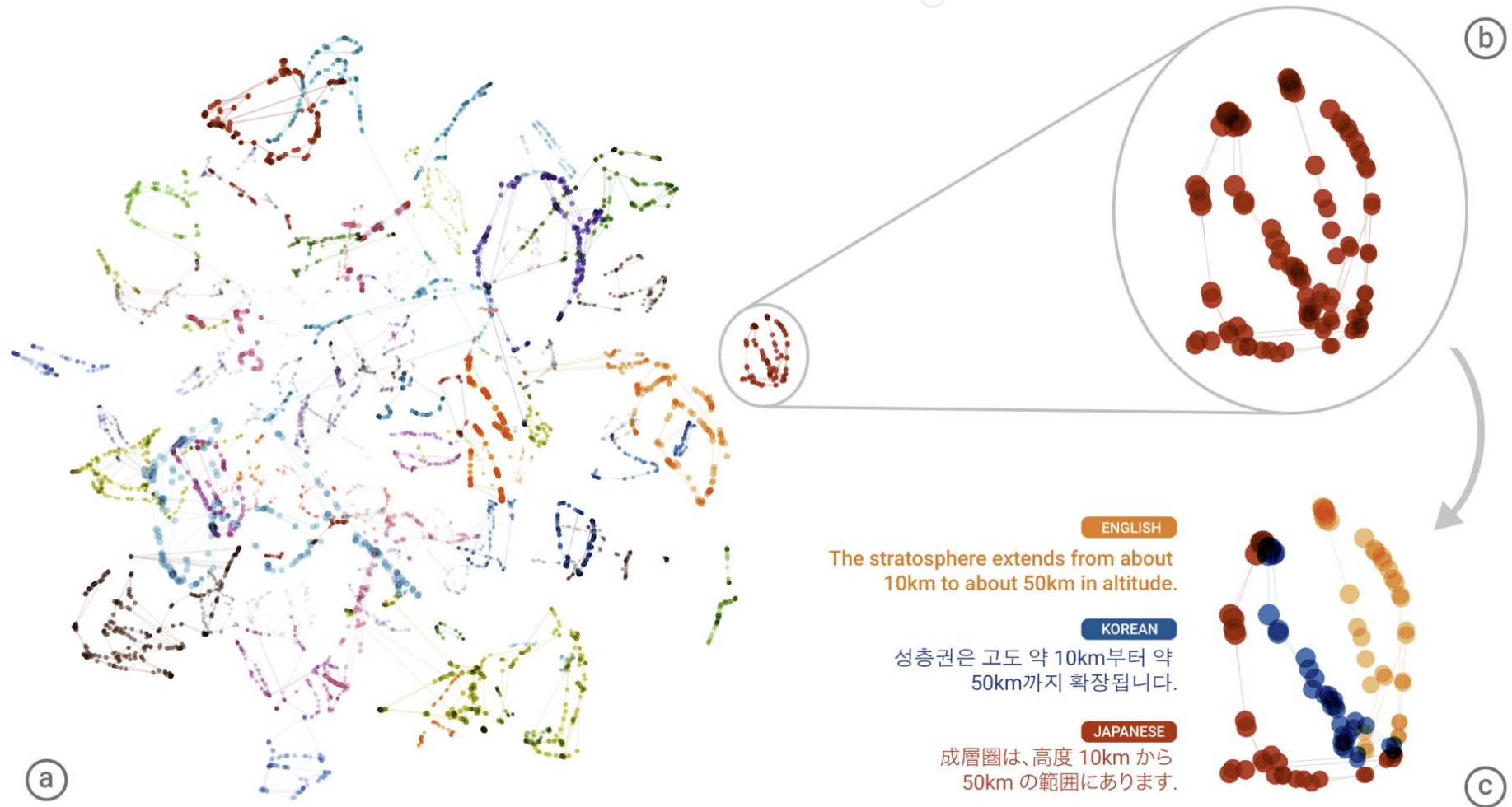
Perplexity 10

Epsilon 5

A square grid with equal spacing between points. Try convergence at different sizes.

distill.pub

# MT Embedding [Johnson et al. 2018]



t-SNE projection of latent space of language translation model.

# Lecture Summary

Use **expressive** and **effective** encodings

**Reduce** the problem space

Avoid **over-encoding**

Use **space** and **small multiples** intelligently

Use **interaction** to generate *relevant* views

Rarely does a single visualization answer all questions. Instead, the ability to generate appropriate visualizations quickly is critical!

# About the design process...

Visualization draws upon both science and art!

Principles like expressiveness & effectiveness are not hard-and-fast rules, but can assist us to guide the process and articulate alternatives.

They can lead us to think more deeply about our design rationale and prompt us to reflect.

It helps to know “the rules” in order to wisely bend (*or break*) them at the right times!

Administrivia

# A1: Expository Visualization

Pick a **guiding question**, use it to title your vis.

Design a **static visualization** for that question.

You are free to **use any tools** (inc. pen & paper).

**Deliverables** (submit on Gradescope; see A1 page)

Image of your visualization (PNG or JPG format)

Short description + design rationale ( $\leq 4$  paragraphs)

Due by **11:59 pm, Fri Jan 17.**

# Tableau Tutorial (Optional)

Friday Jan 17, 4-5:30pm over Zoom

Led by Han

Zoom link will be available on Canvas

Session will be recorded

# I Like... / I Wish... / What If?

## I LIKE...

Praise for design ideas and/or well-executed implementation details.

*Example: "I like the navigation through time via the slider; the patterns observed as one moves forward are compelling!"*

## I WISH...

Constructive statements on how the design might be improved or further refined. *Example: "I wish moving the slider caused the visualization to update immediately, rather than the current lag."*

## WHAT IF?

Suggest alternative design directions, or even wacky half-baked ideas.

*Example: "What if we got rid of the slider and enabled direct manipulation navigation by dragging data points directly?"*