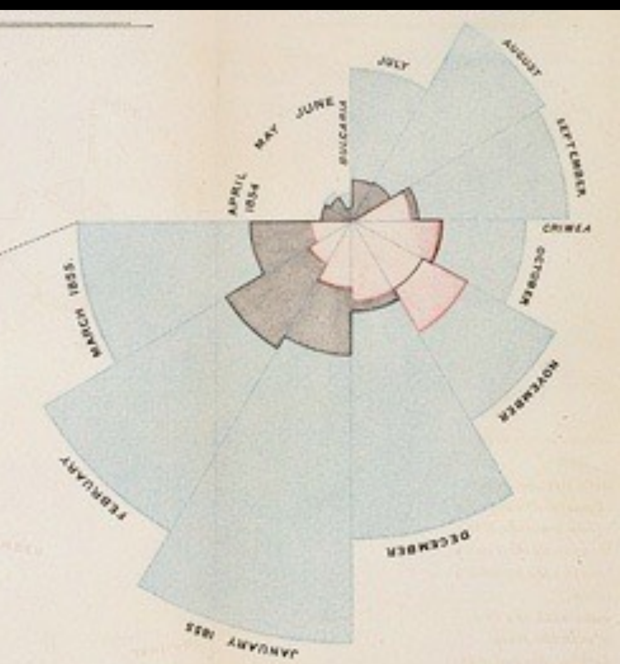
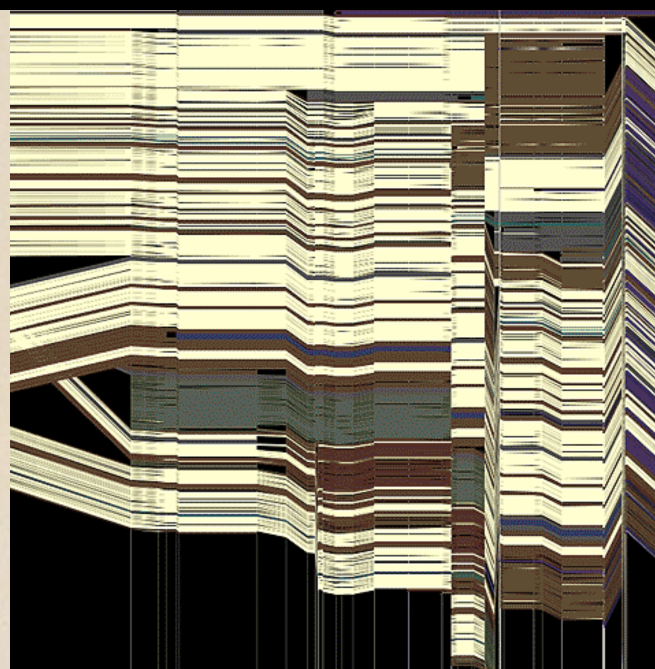
CSE 512 - Data Visualization

Visual Encoding Design





Jeffrey Heer University of Washington

A Design Space of Visual Encodings

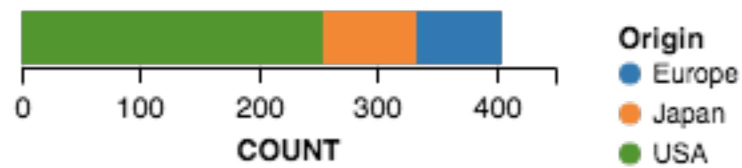
Mapping Data to Visual Variables

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, …*).

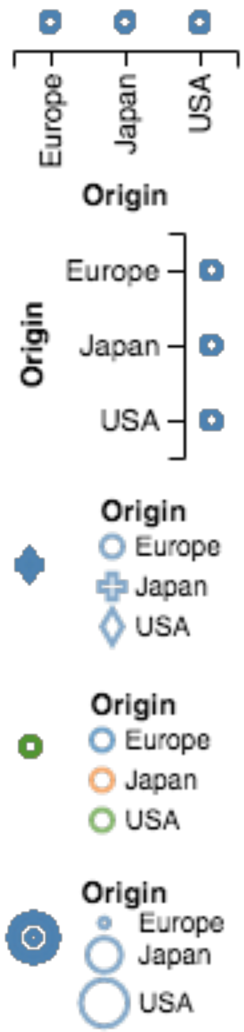
Additional concerns include choosing appropriate **encoding parameters** (*log scale, sorting, …*) and **data transformations** (*bin*, *group, aggregate, …*).

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

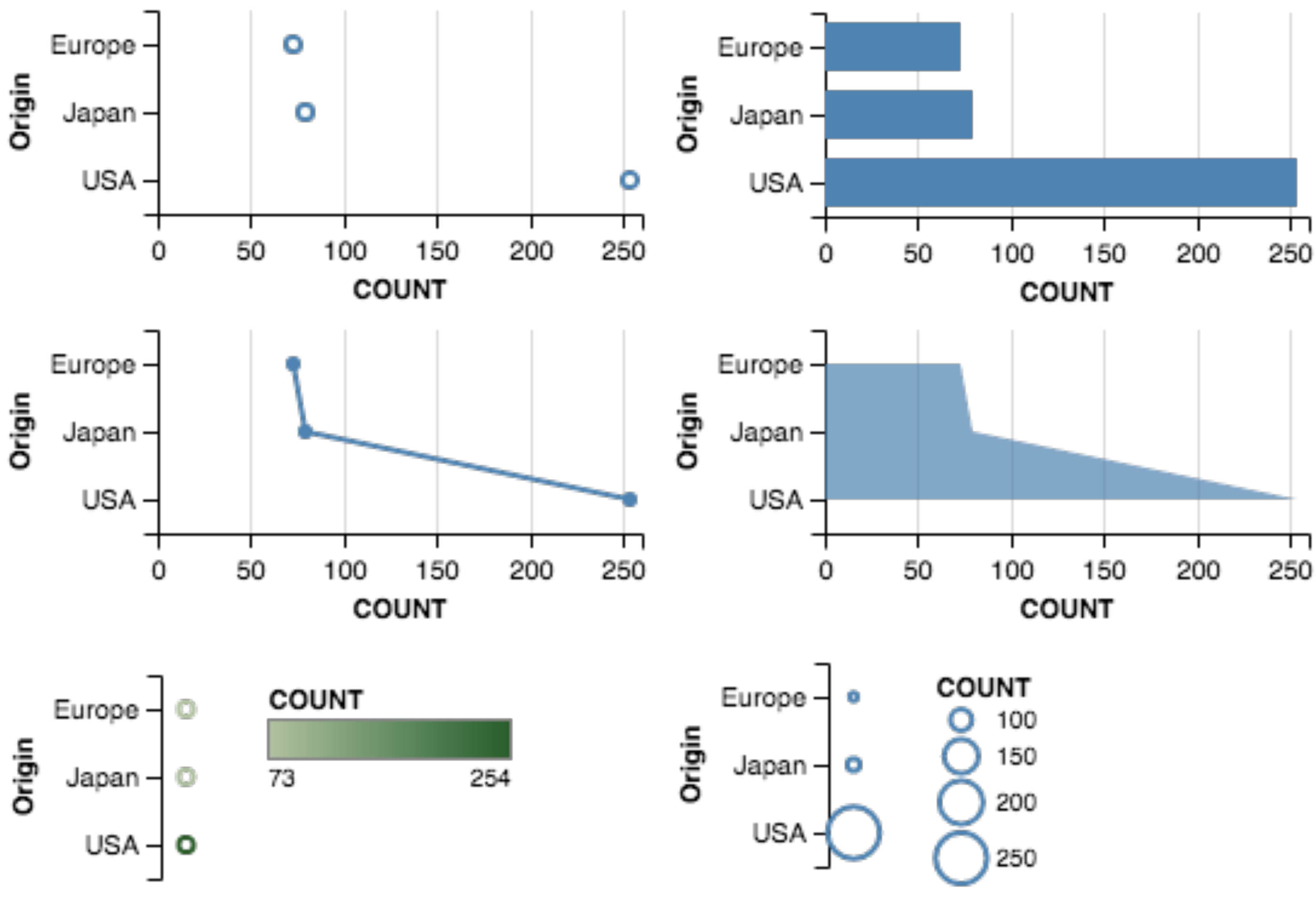
1D: Nominal



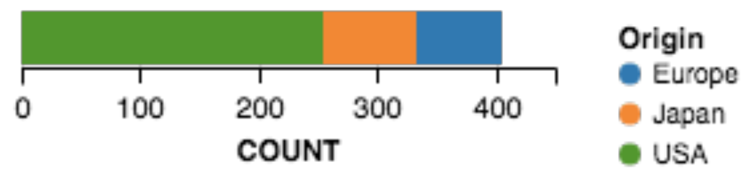
Raw



Aggregate (Count)

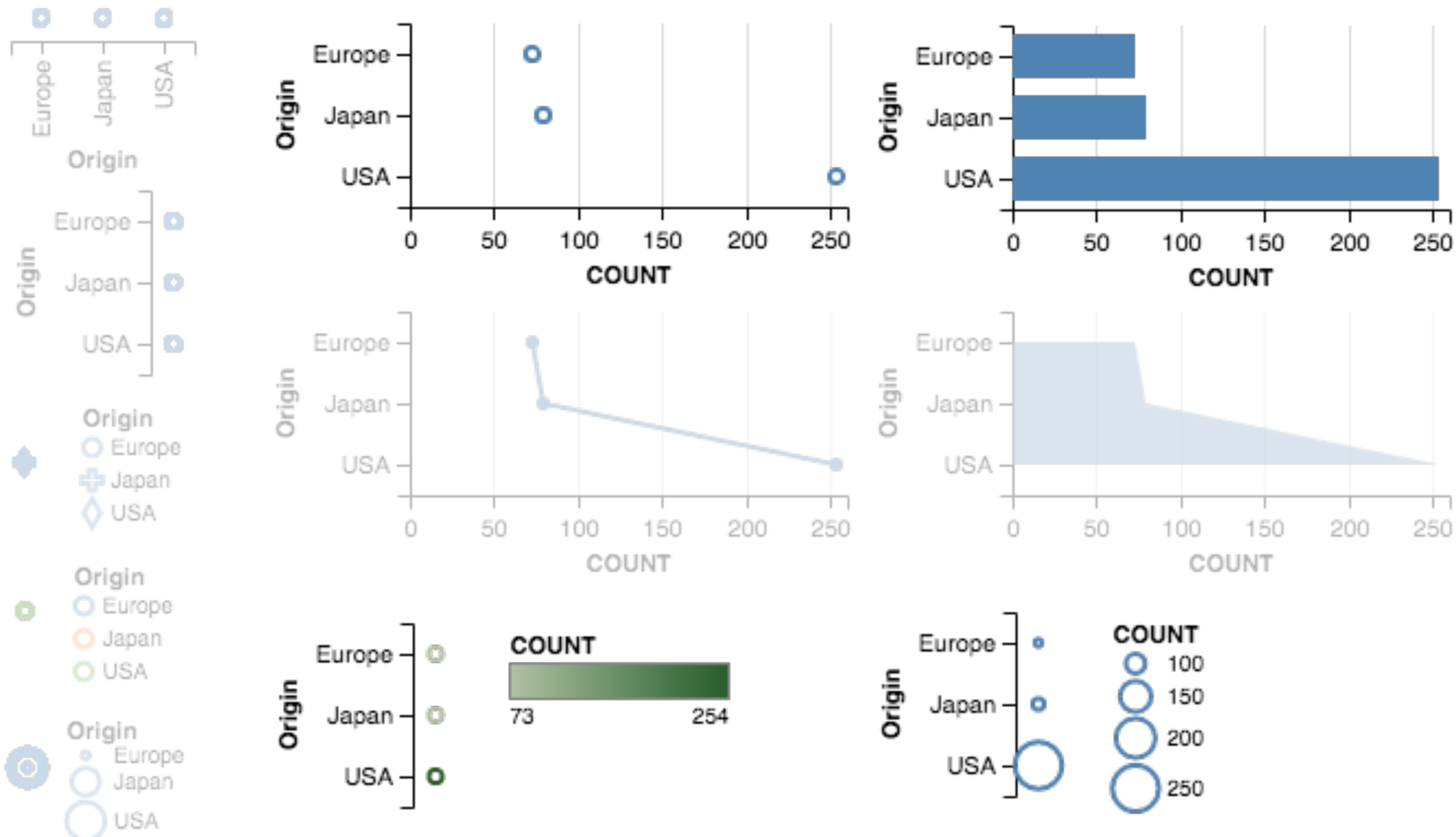


Expressive?



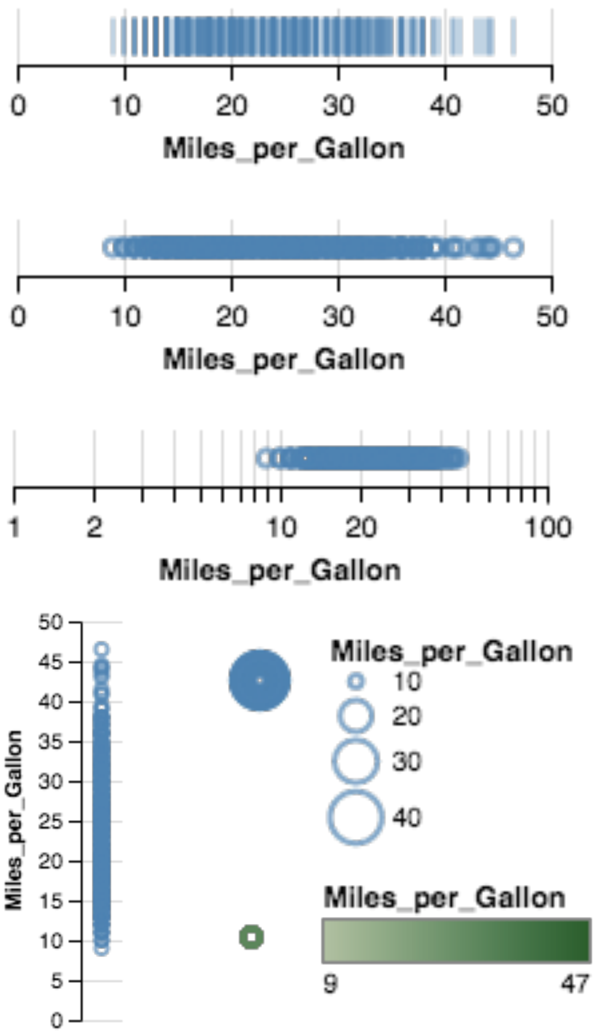
Raw

Aggregate (Count)

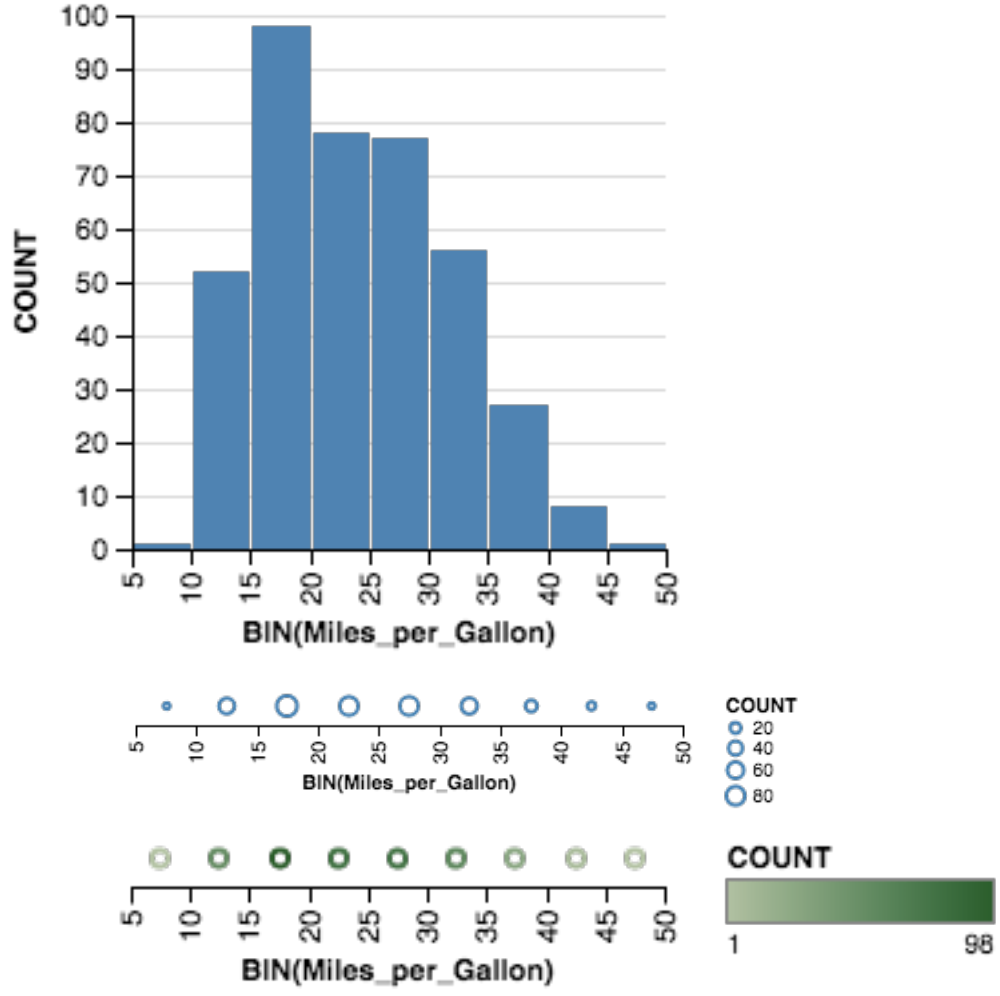


1D: Quantitative

Raw

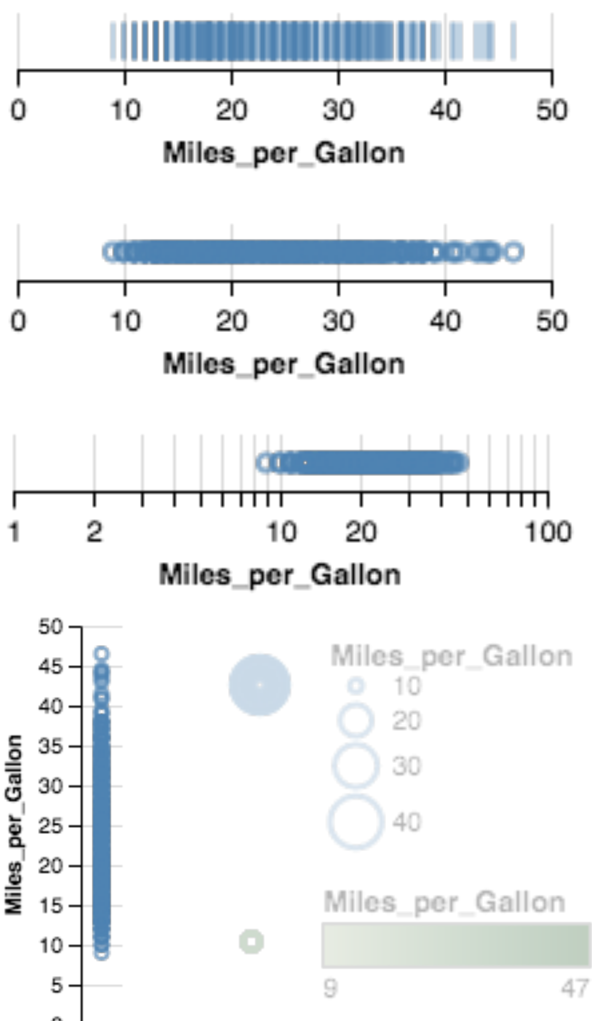


Aggregate (Count)

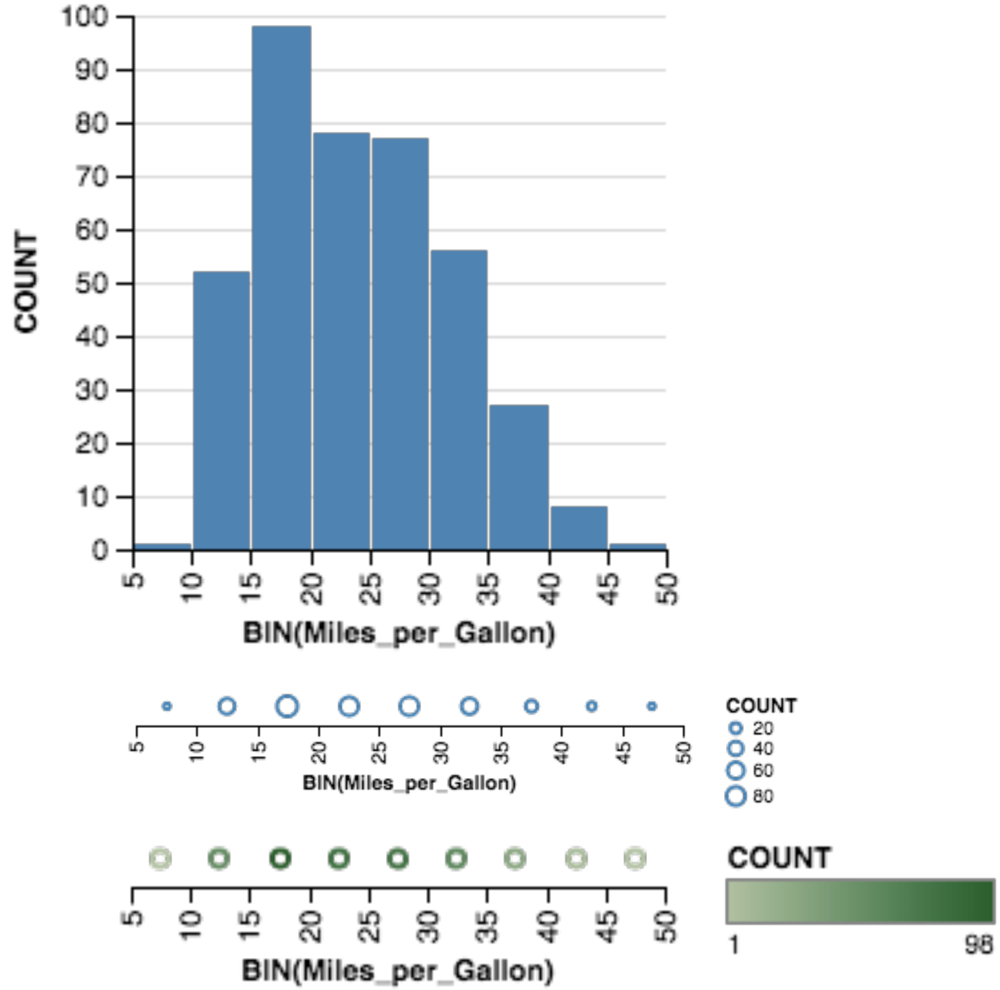


Expressive?

Raw

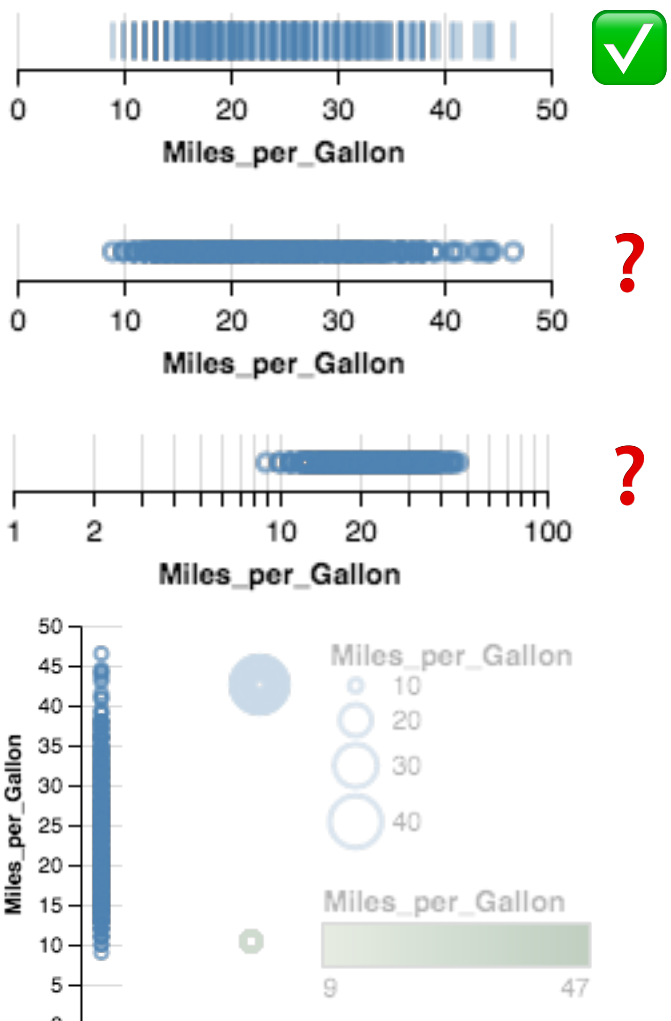


Aggregate (Count)

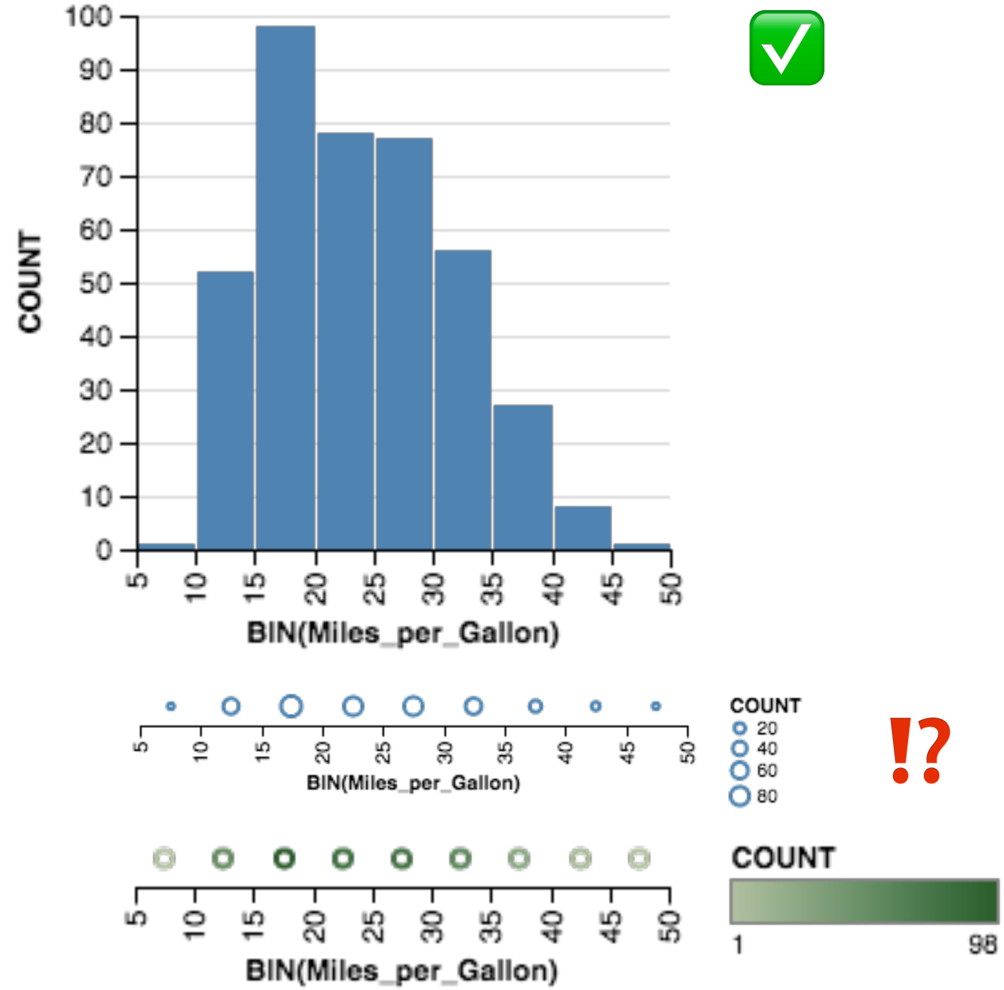


Effective?

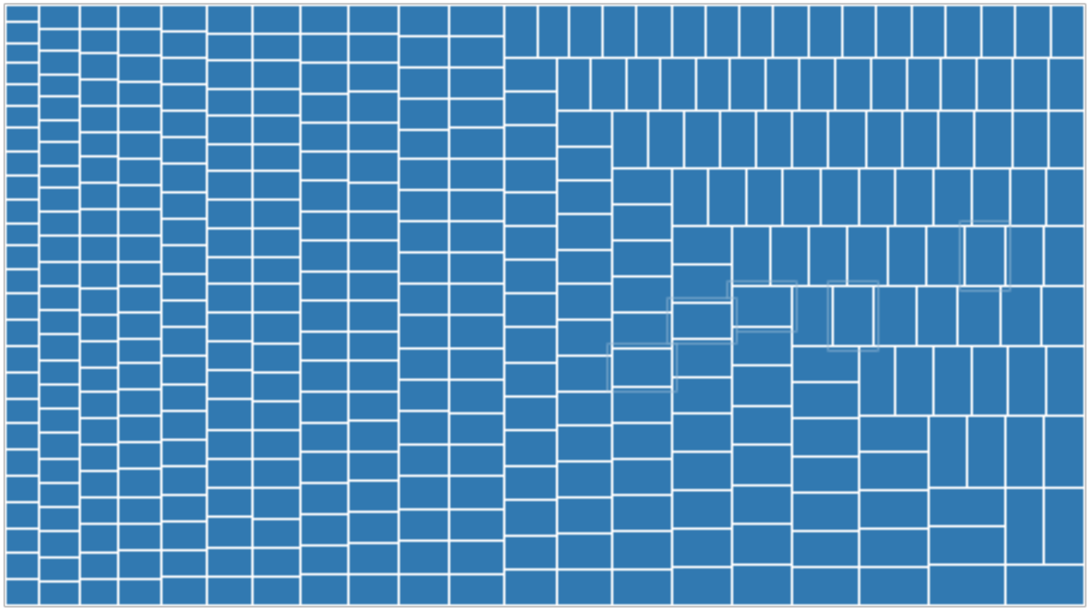
Raw



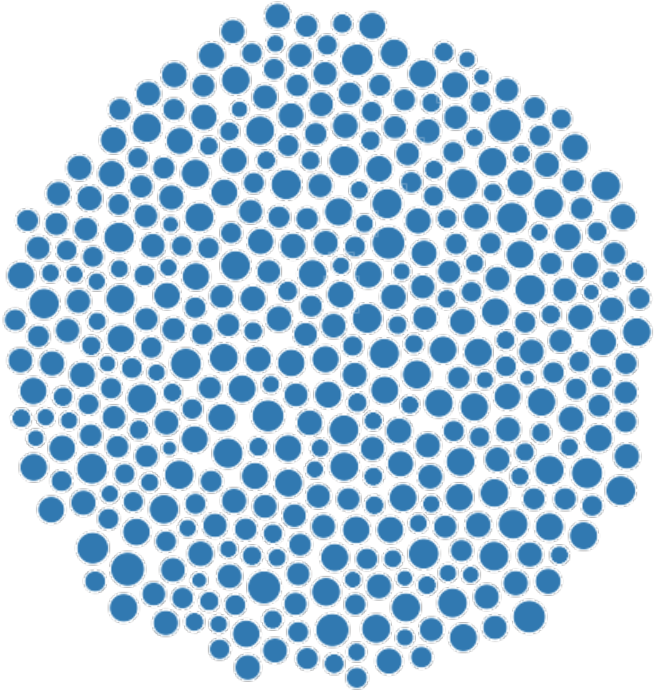
Aggregate (Count)



Raw (with Layout Algorithm)

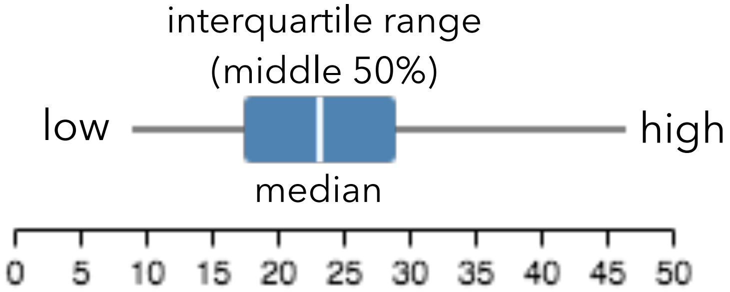


Treemap

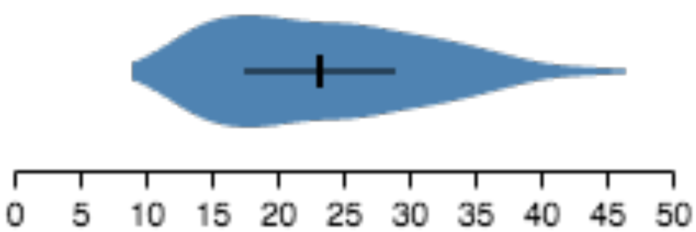


Bubble Chart

Aggregate (Distributions)



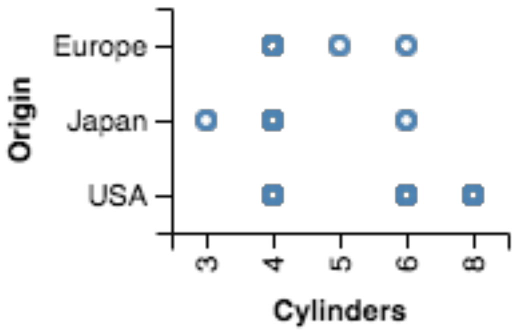
Box Plot

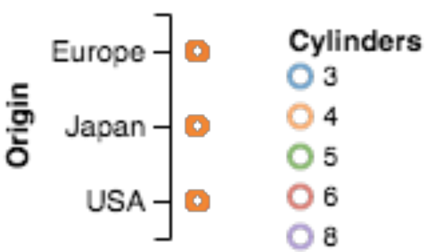


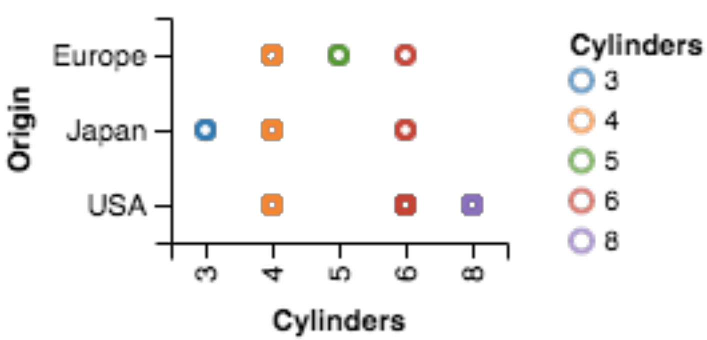
Violin Plot

2D: Nominal x Nominal

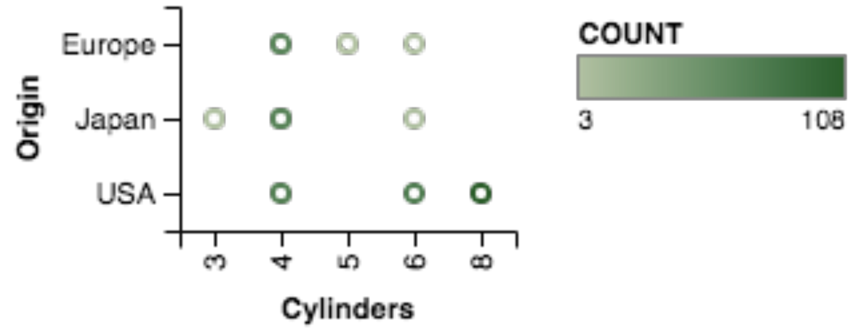
Raw

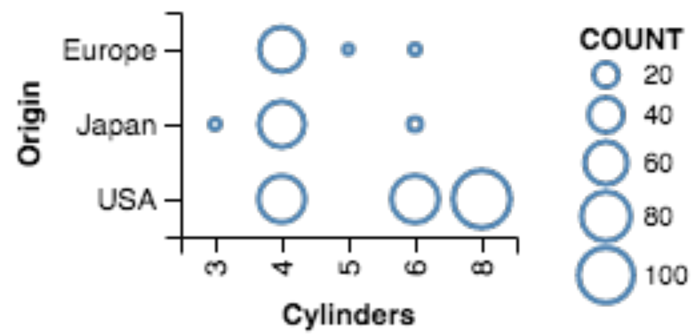


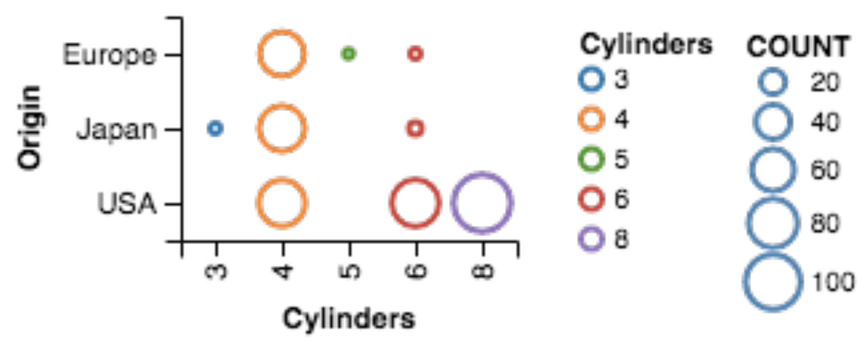




Aggregate (Count)

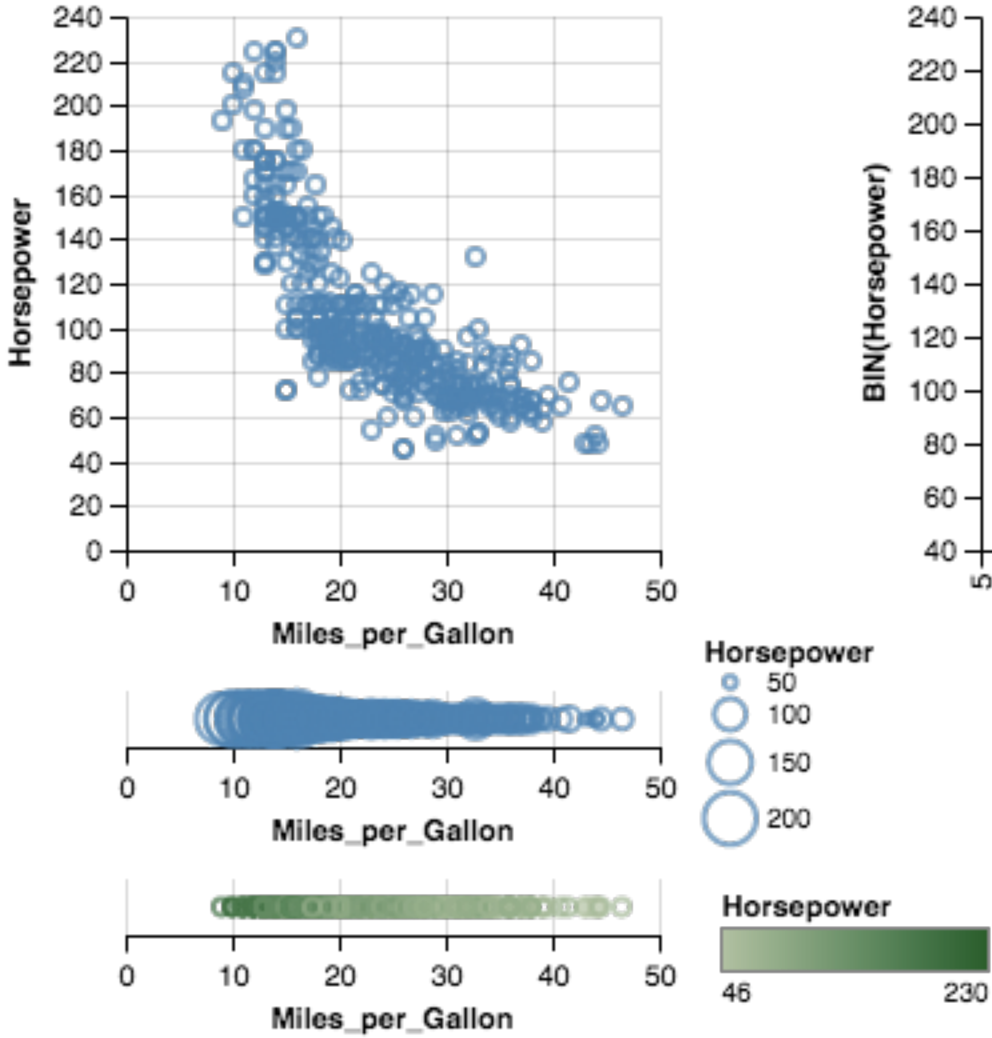




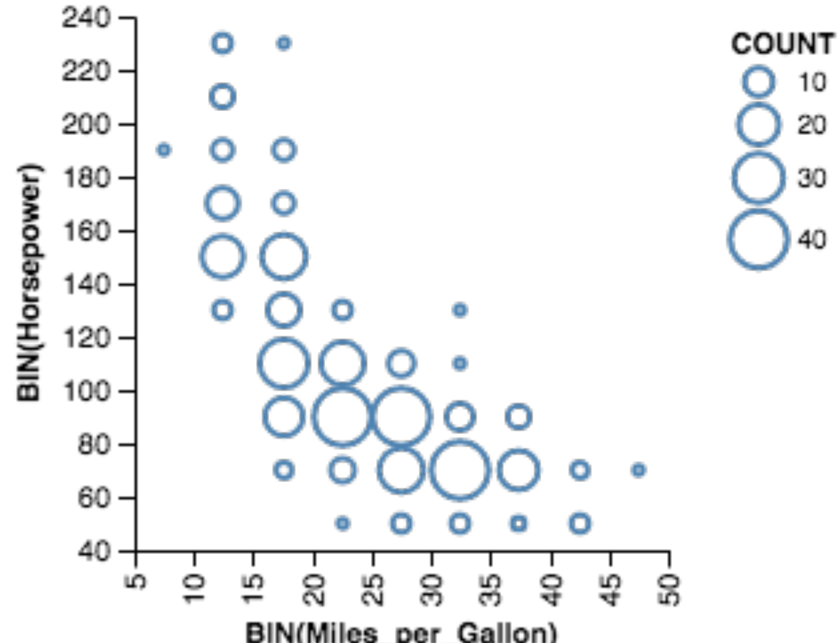


2D: Quantitative x Quantitative

Raw

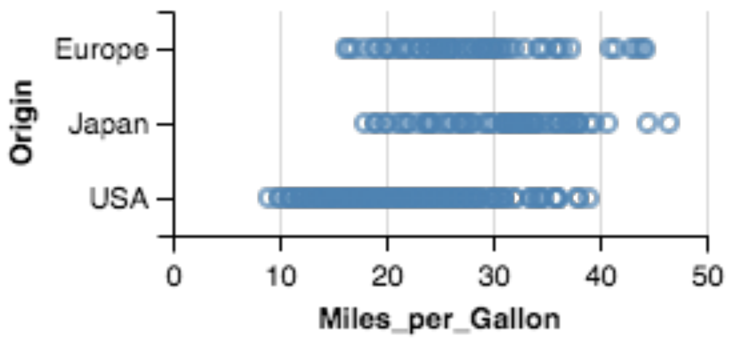


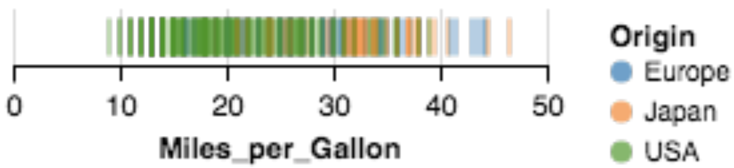
Aggregate (Count)

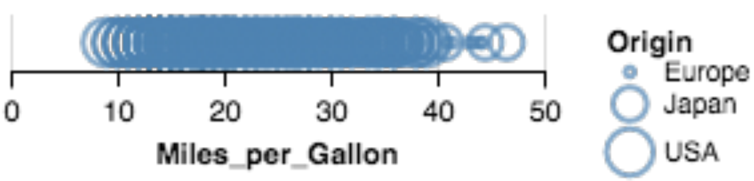


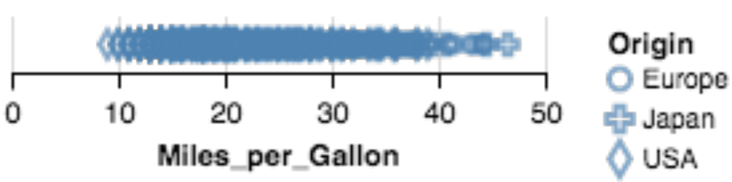
2D: Nominal x Quantitative

Raw

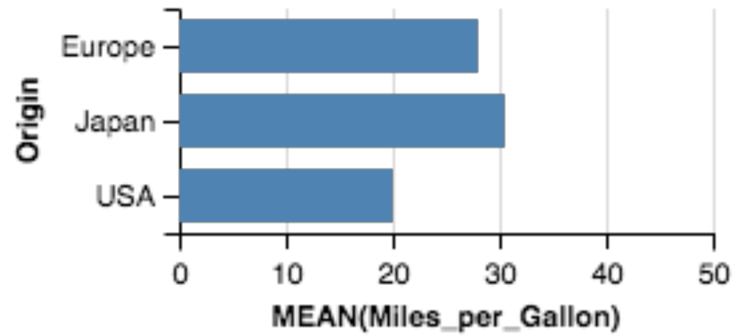


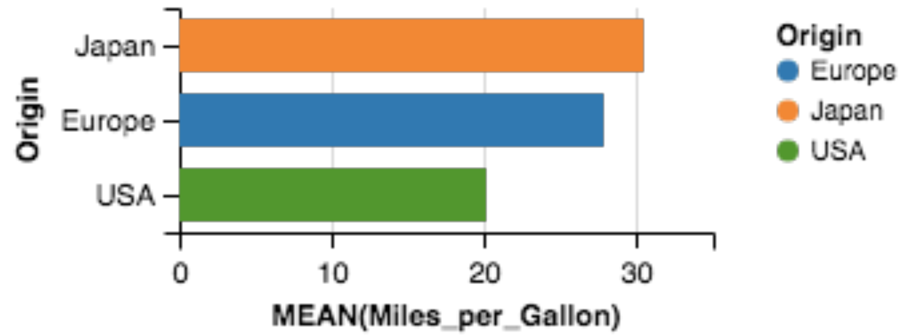




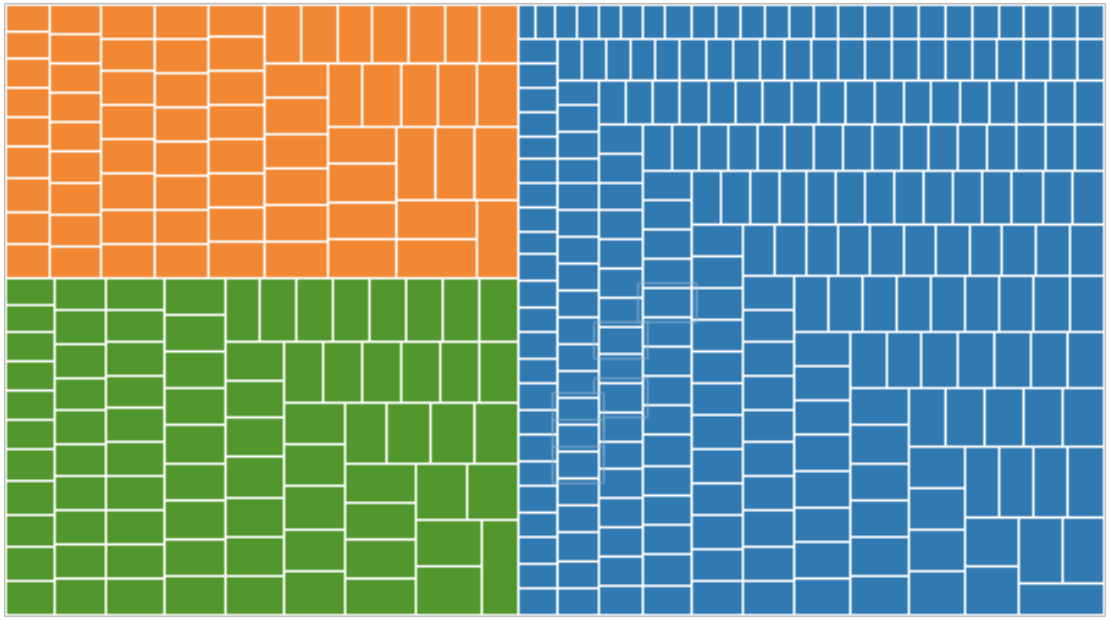


Aggregate (Mean)

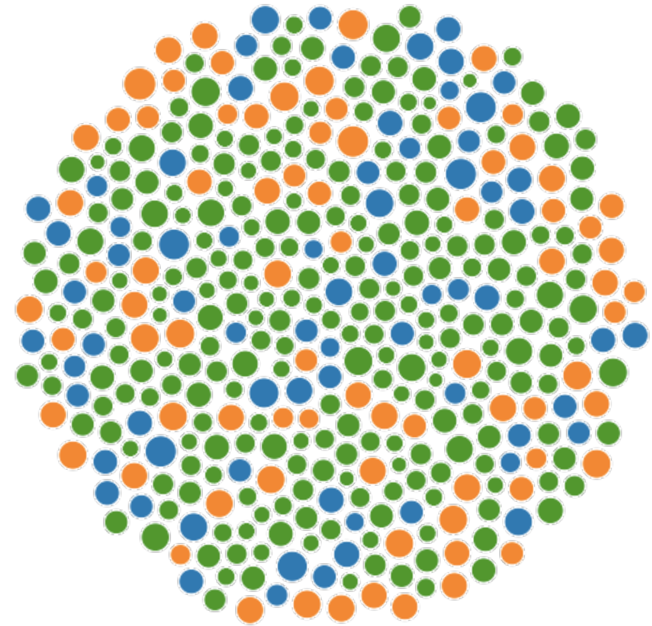




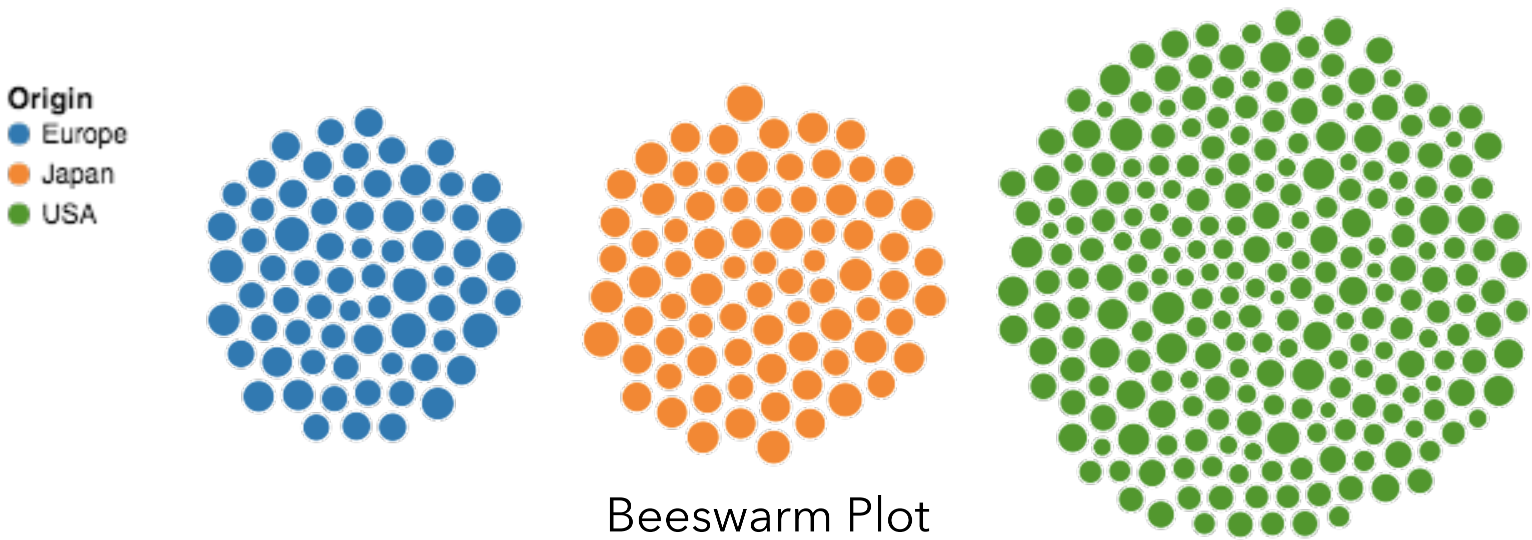
Raw (with Layout Algorithm)



Treemap



Bubble Chart



Beeswarm Plot

3D and Higher

Two variables [x,y]

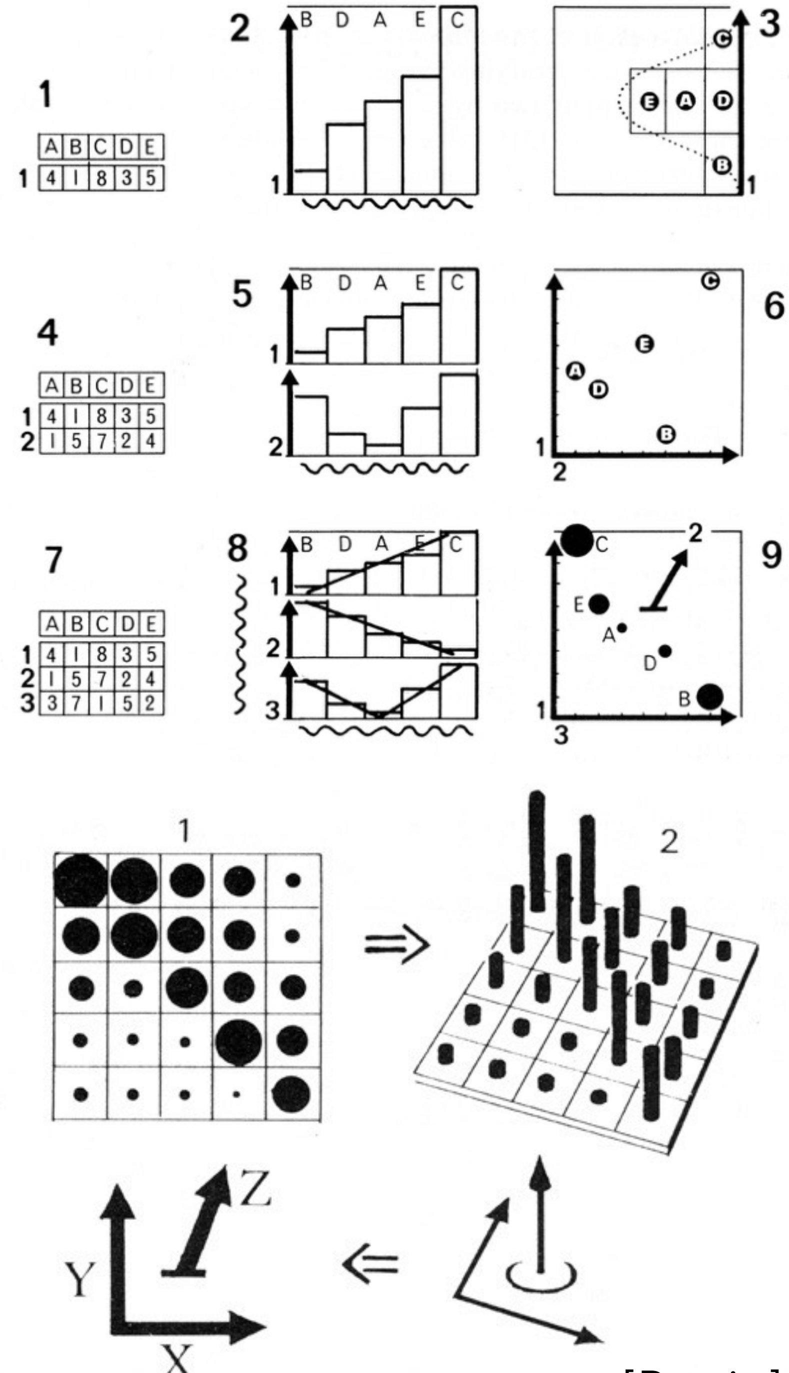
Can map to 2D points.

Scatterplots, maps, …

Third variable [z]

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

What about 3D rendering?



[Bertin]

Other Visual Encoding Channels?

Encoding Effectiveness

Effectiveness Rankings [Mackinlay 86]

QUANTITATIVE	ORDINAL	NOMINAL
Position	Position	Position
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	NOMINAL
Position	**Position**	**Position**
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]

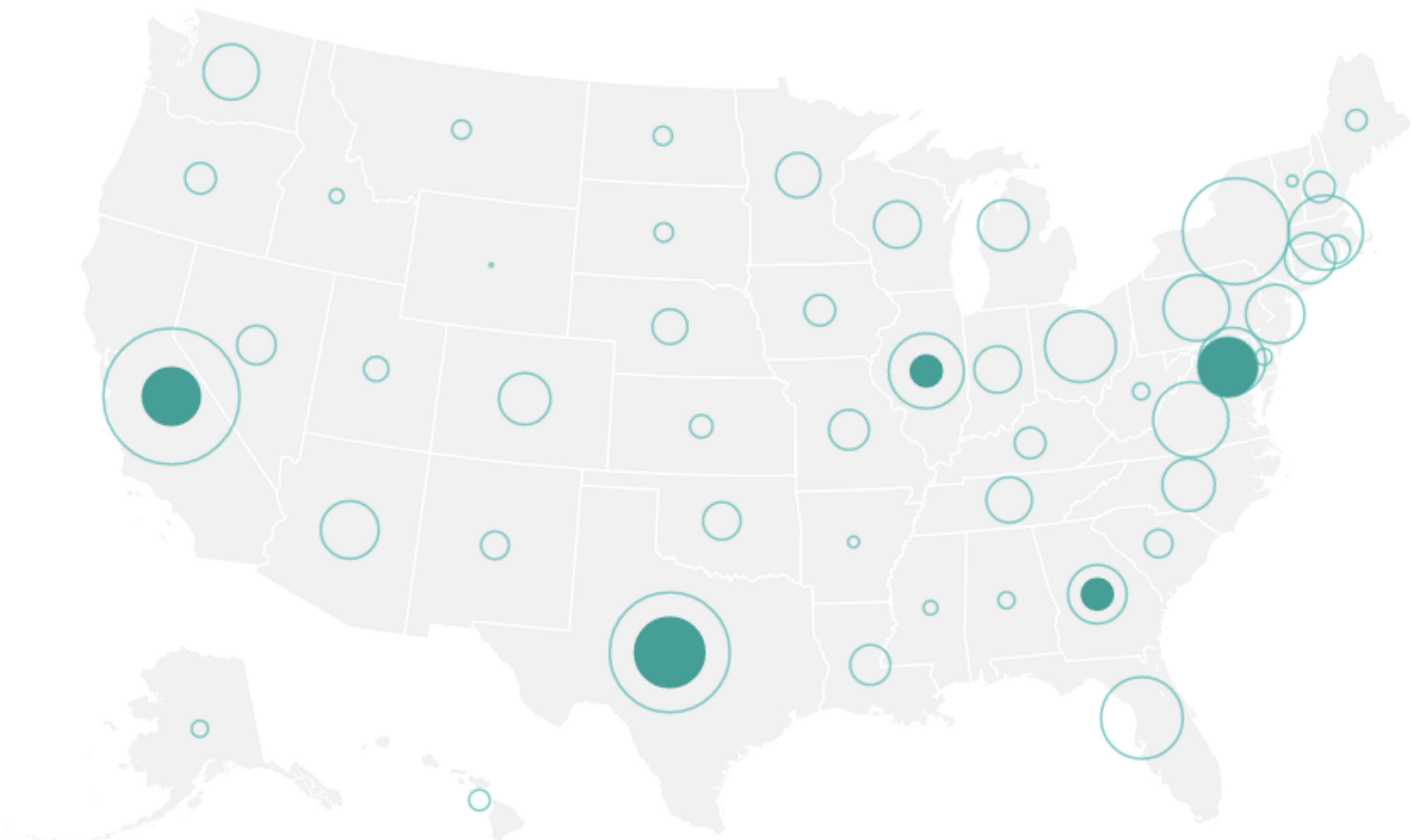
QUANTITATIVE	ORDINAL	NOMINAL
Position	Position	Position
Length	Density (Value)	**Color Hue**
Angle	Color Sat	Texture
Slope	**Color Hue**	Connection
Area (Size)	Texture	Containment
Volume	Connection	Density (Value)
Density (Value)	Containment	Color Sat
Color Sat	Length	Shape
Color Hue	Angle	Length
Texture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

Color Encoding (Choropleth Map)

Effectiveness Rankings

QUANTITATIVE	ORDINAL	NOMINAL
Position	Position	Position
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

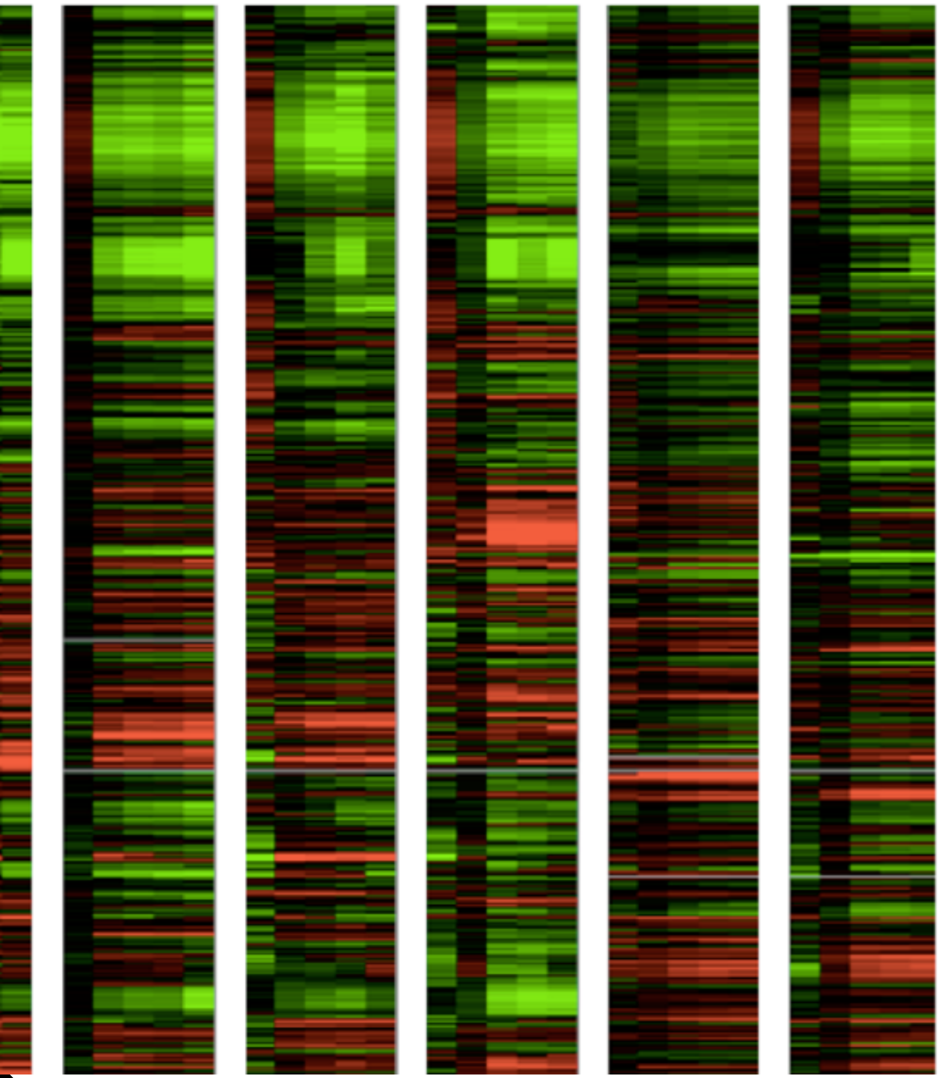
Color Encoding (Choropleth Map)



Area Encoding (Symbol Map)

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

Color Encoding

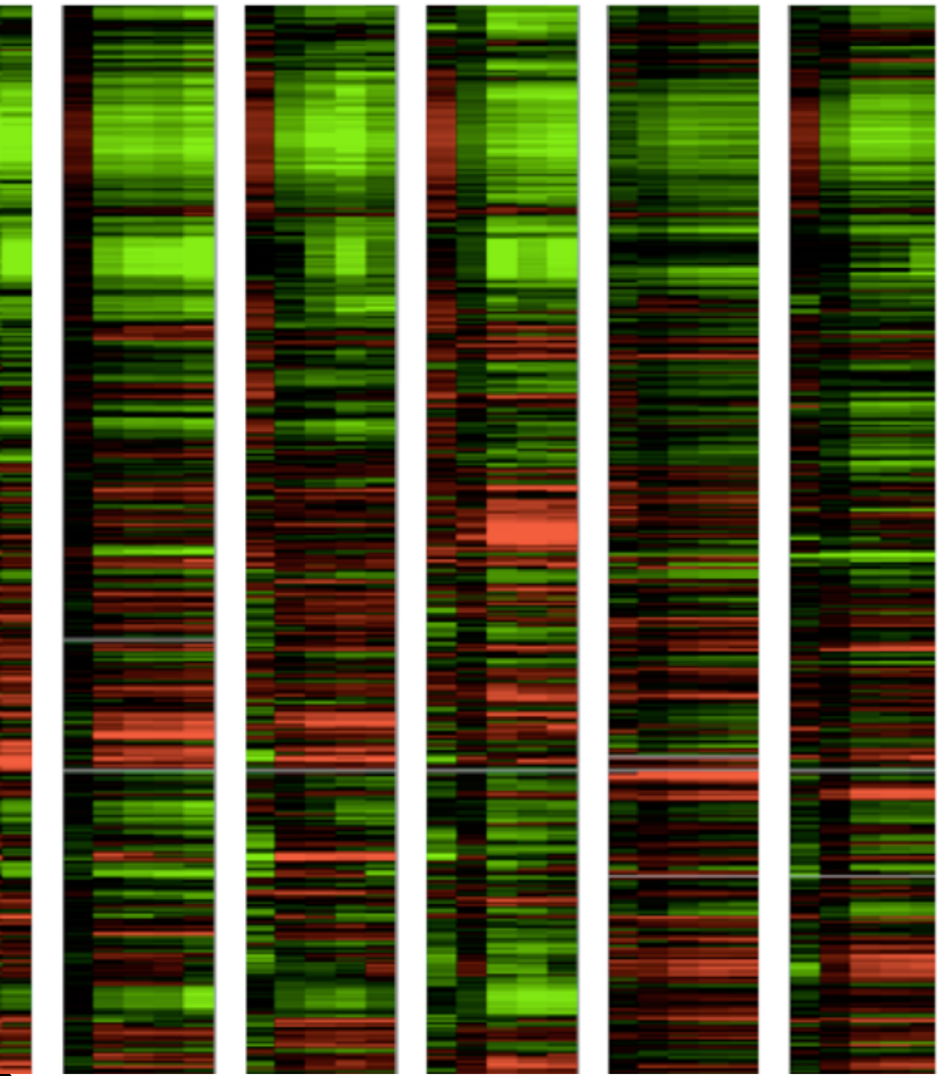


Effectiveness Rankings

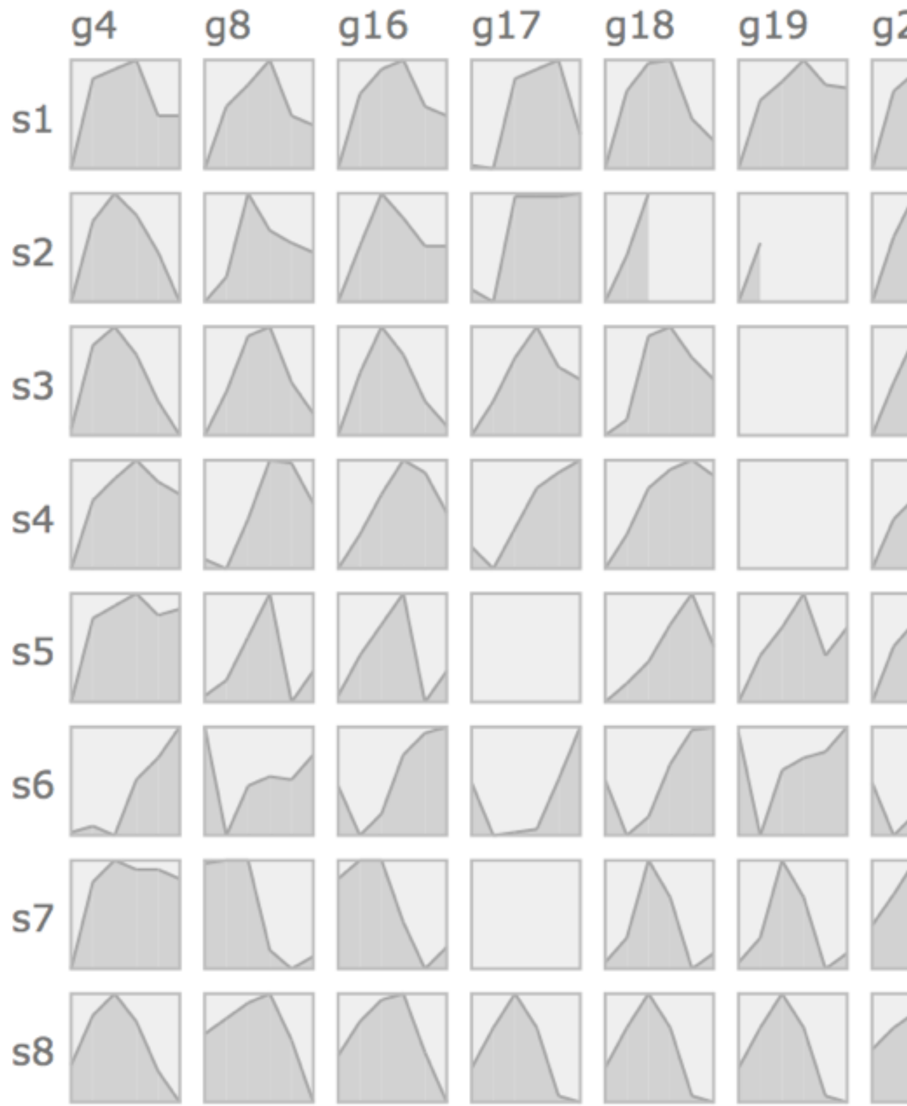
QUANTITATIVE	ORDINAL	NOMINAL
Position	Position	Position
Length	Density (Value)	Color Hue
Angle	Color Sat	Texture
Slope	Color Hue	Connection
Area (Size)	Texture	Containment
Volume	Connection	Density (Value)
~~**Density (Value)**~~	Containment	Color Sat
Color Sat	Length	Shape
~~**Color Hue**~~	Angle	Length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

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

Color Encoding

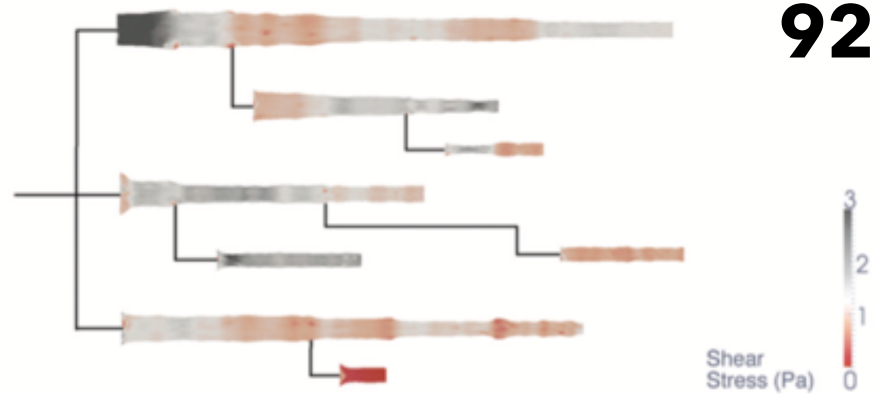


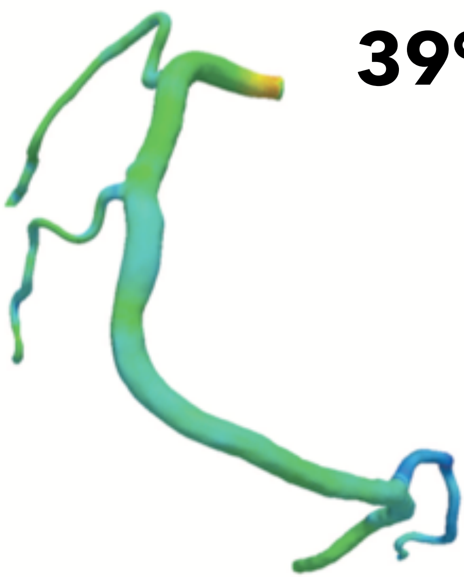
Position Encoding



Artery Visualization [Borkin et al '11]

	Rainbow Palette	Diverging Palette
2D	62%	92%
3D	39%	71%



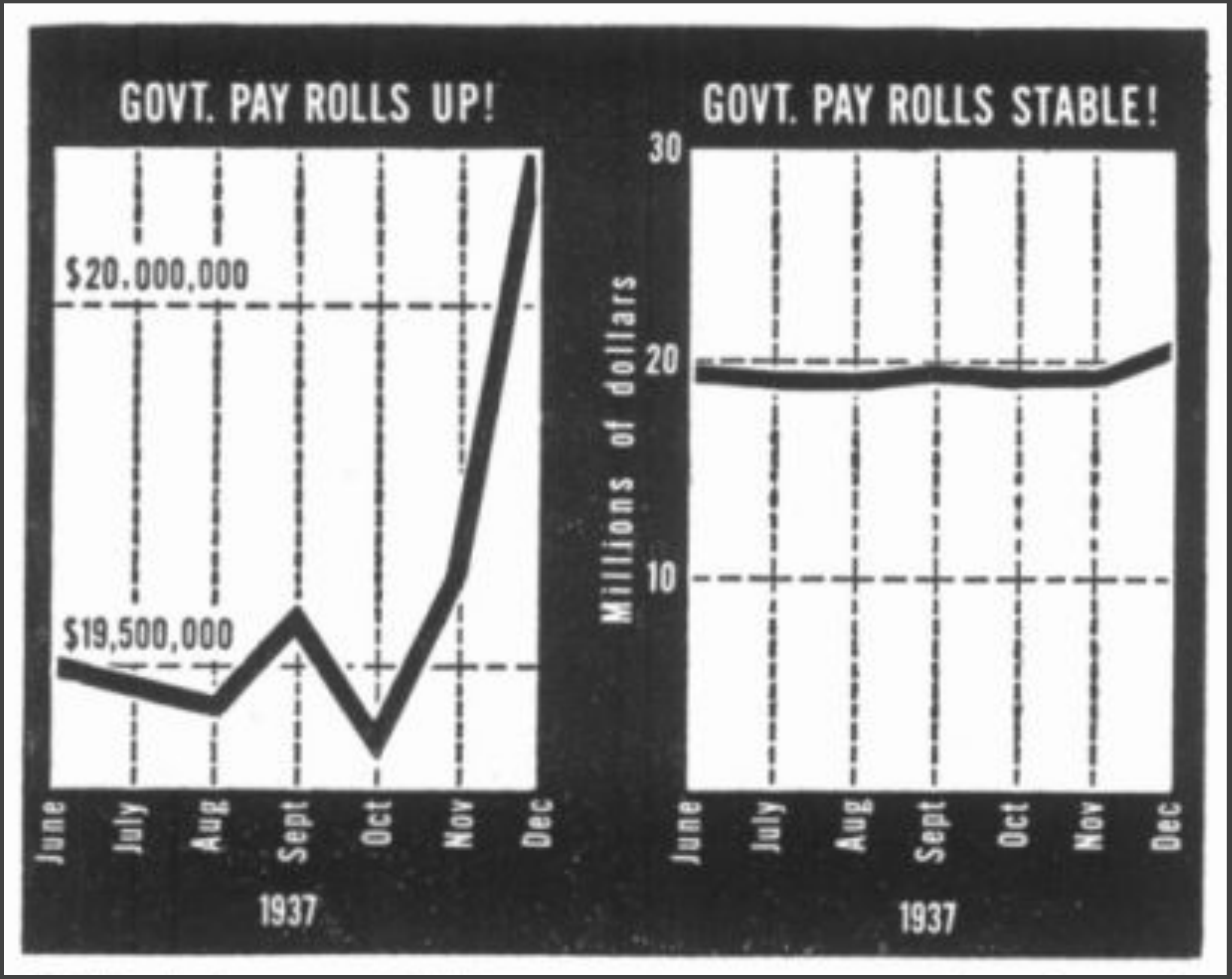


Effectiveness Rankings

QUANTITATIVE	ORDINAL	NOMINAL
Position	Position	Position
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

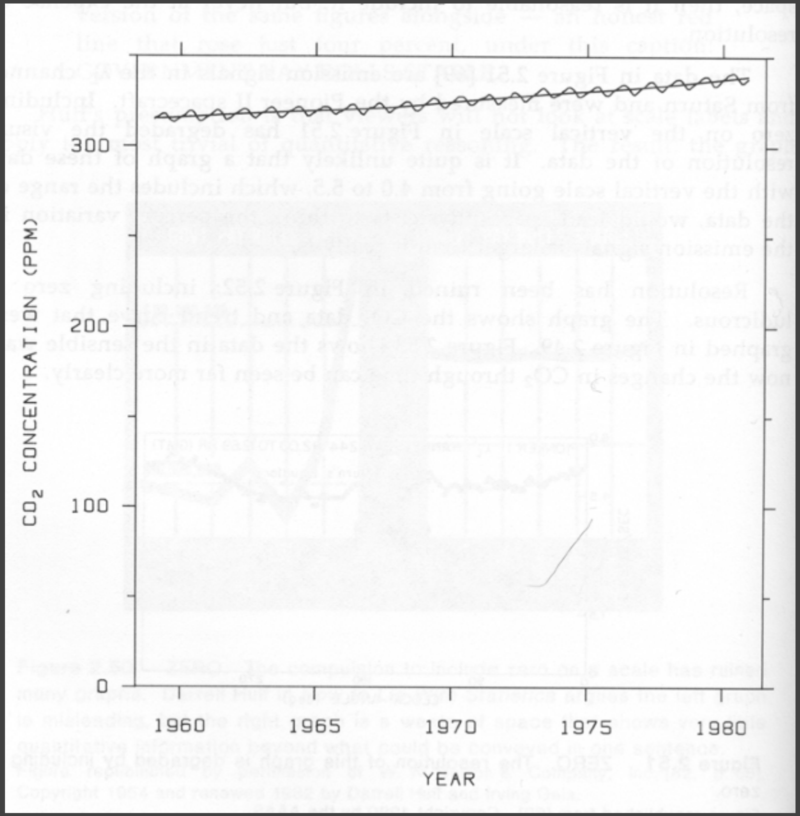
Scales & Axes

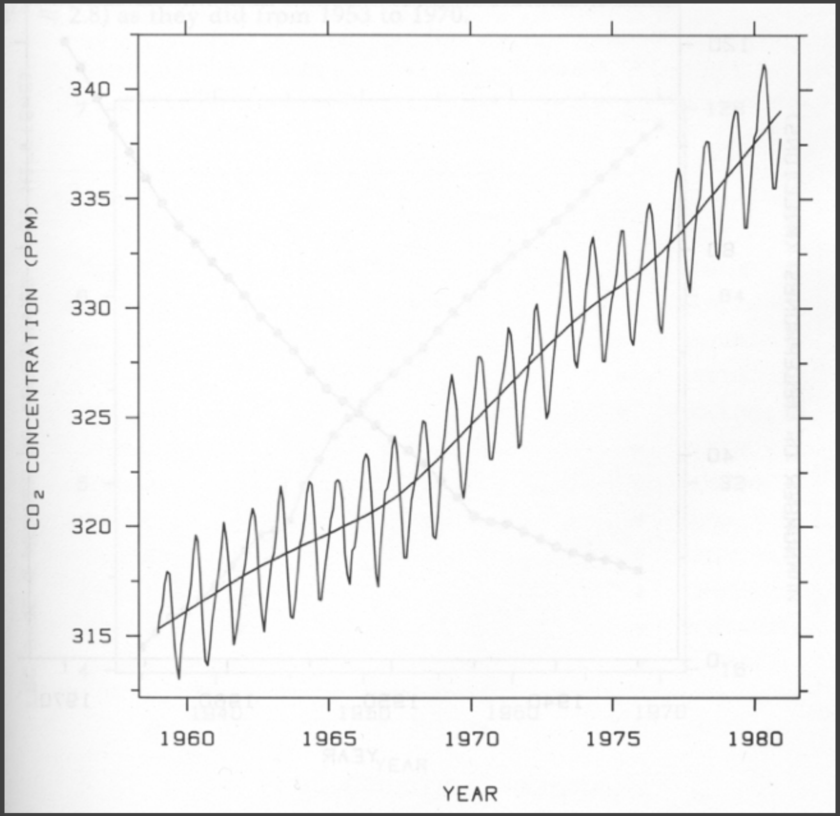
Include Zero in Axis Scale?



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

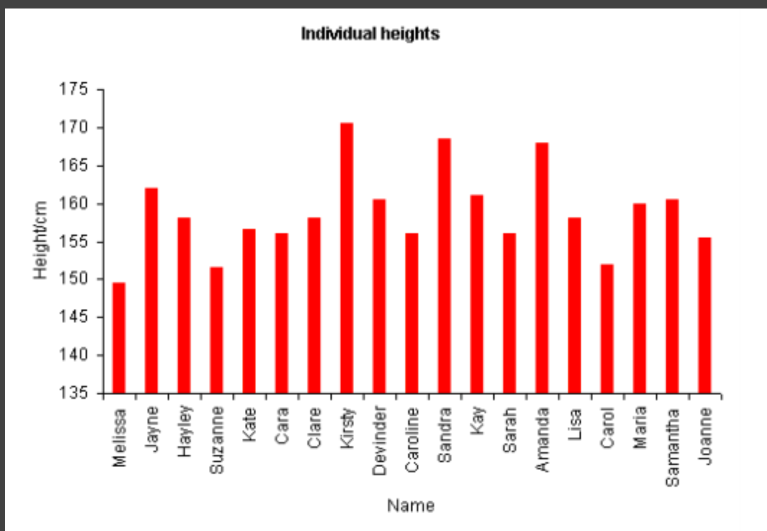
Include Zero in Axis Scale?





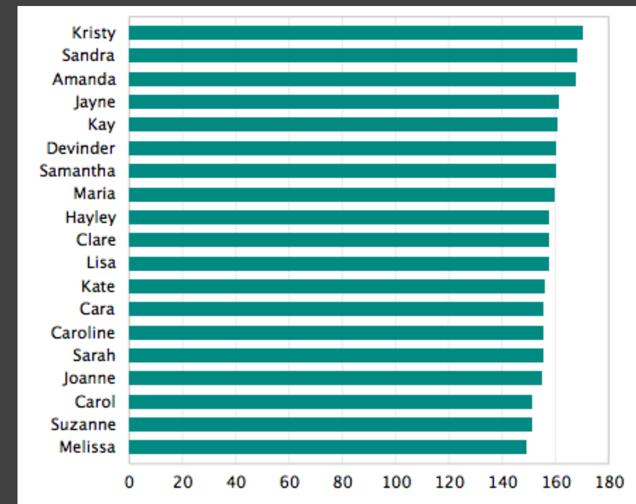
Yearly CO_2 concentrations [Cleveland 85]

Include Zero in Axis Scale?

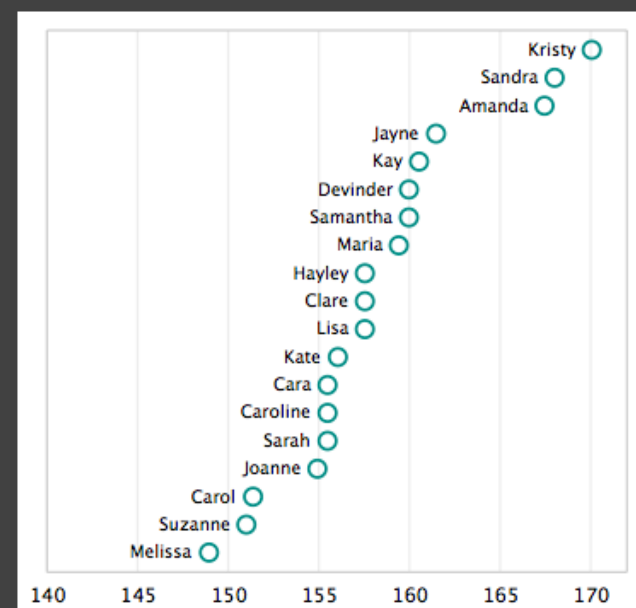


Violates Expressiveness Principle!

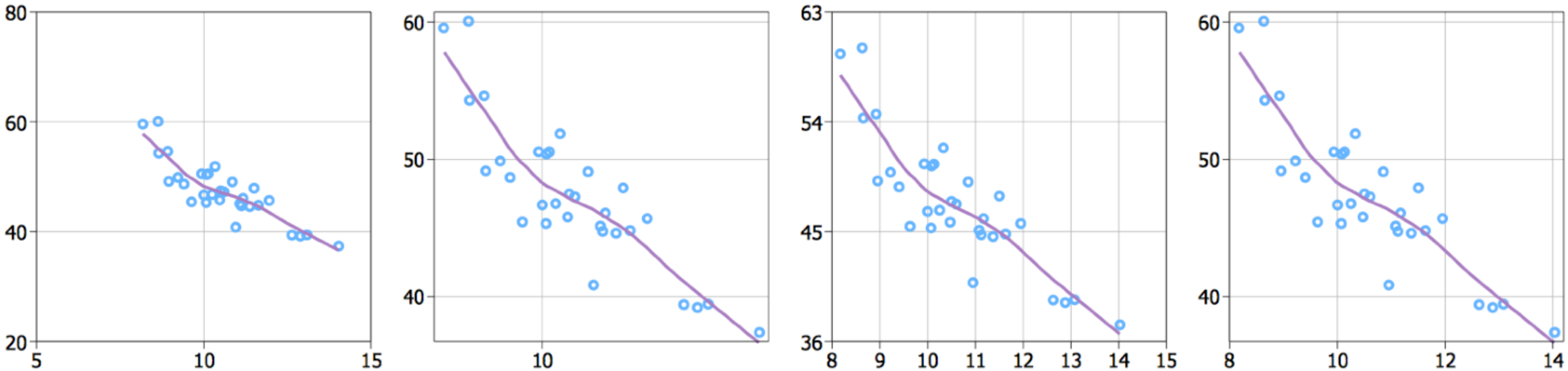
Compare Proportions (Q-Ratio)



Compare Relative Position (Q-Interval)

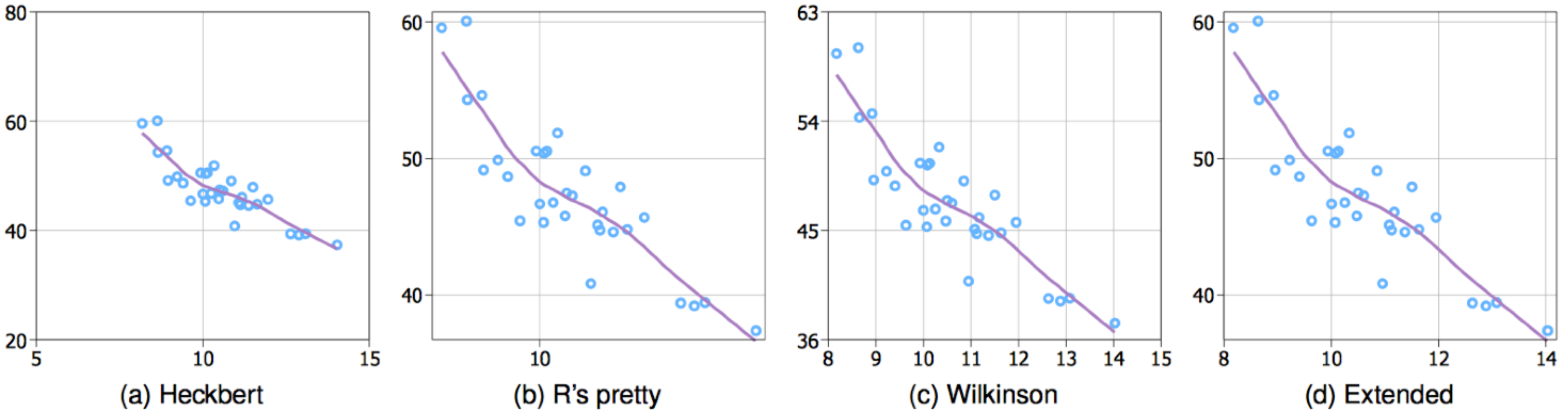


Axis Tick Mark Selection



What are some properties of “good” tick marks?

Axis Tick Mark Selection



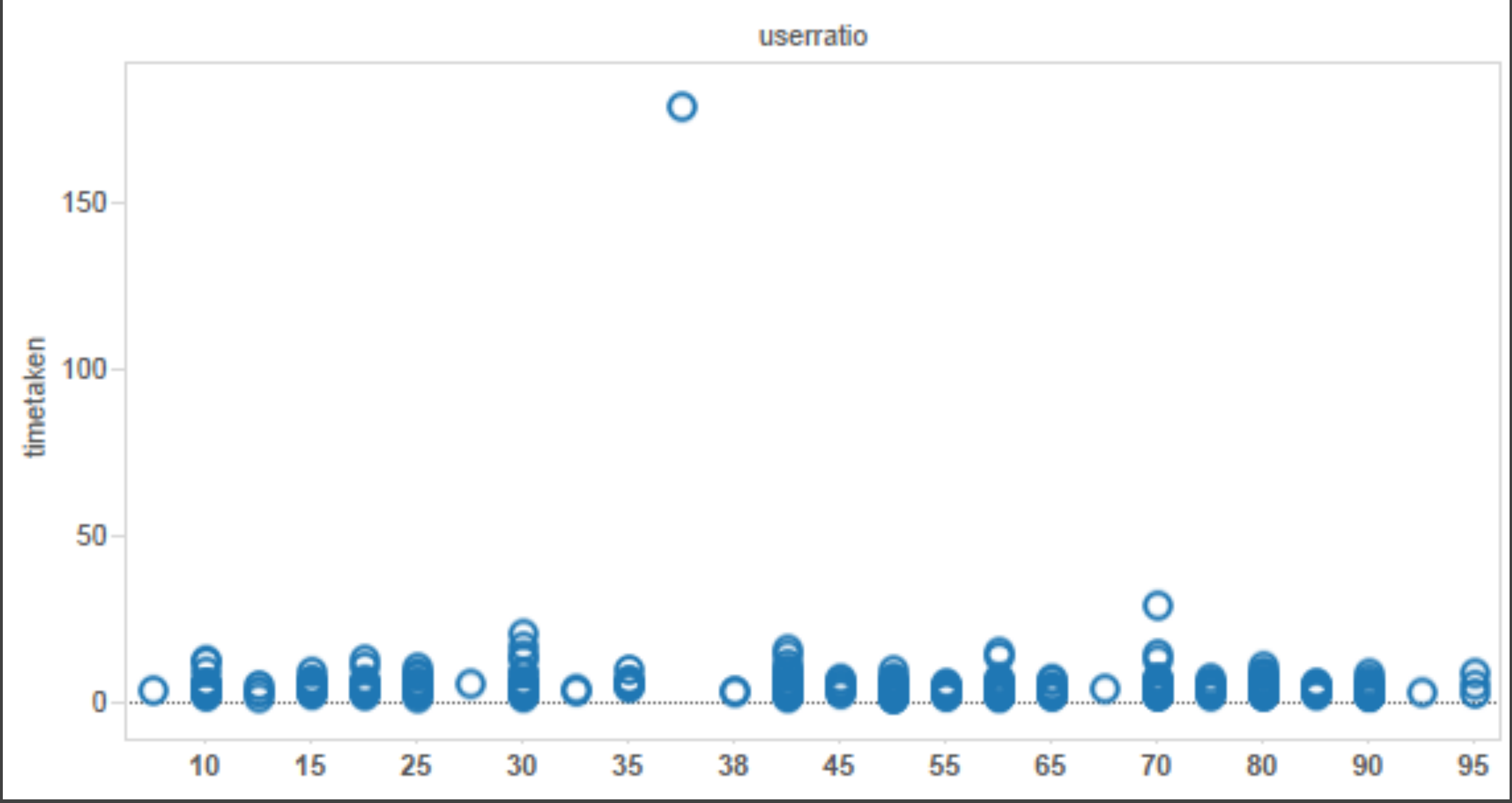
Simplicity - numbers are multiples of 10, 5, 2

Coverage - ticks near the ends of the data

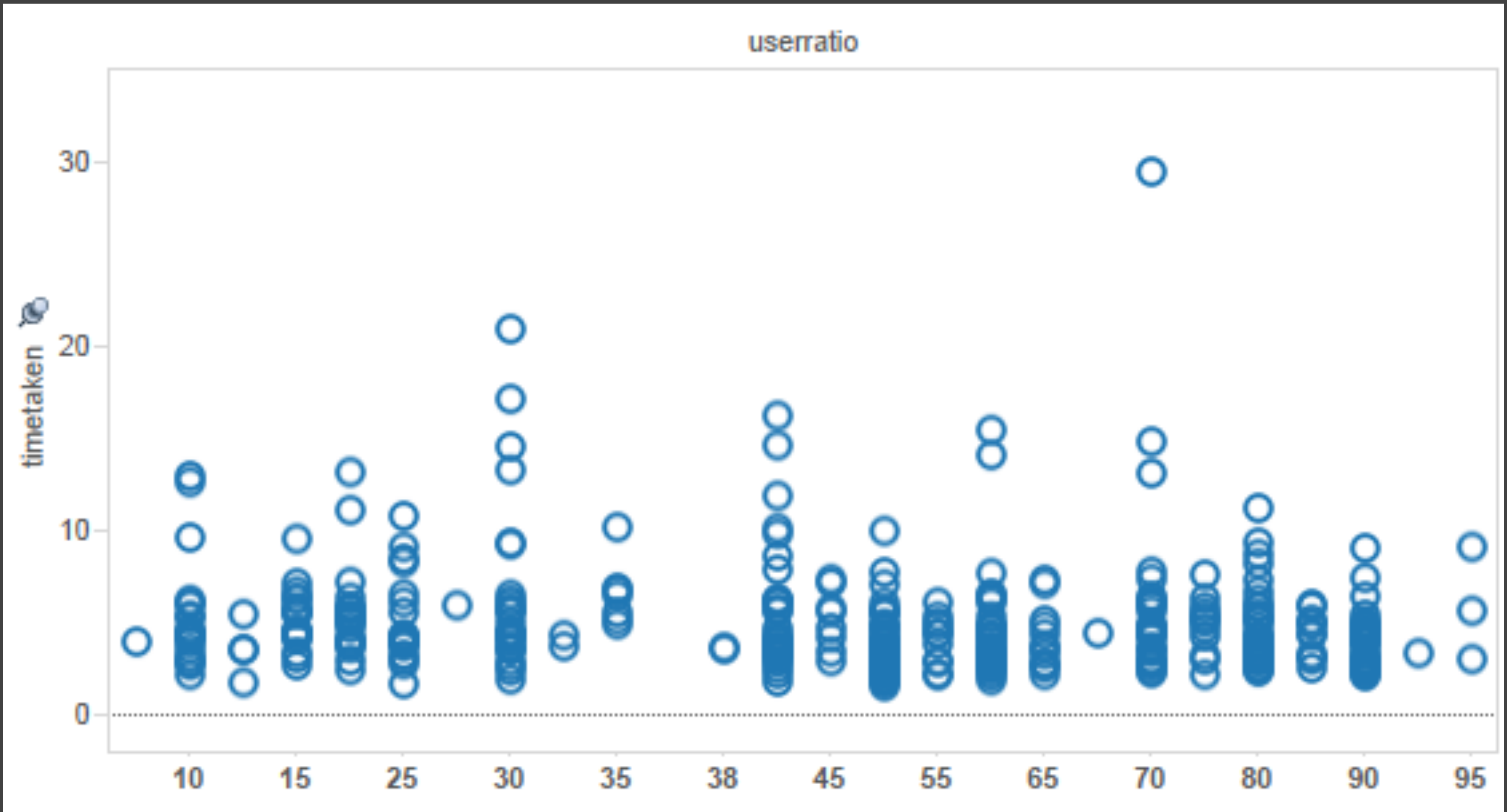
Density - not too many, nor too few

Legibility - whitespace, horizontal text, size

How to Scale the Axis?

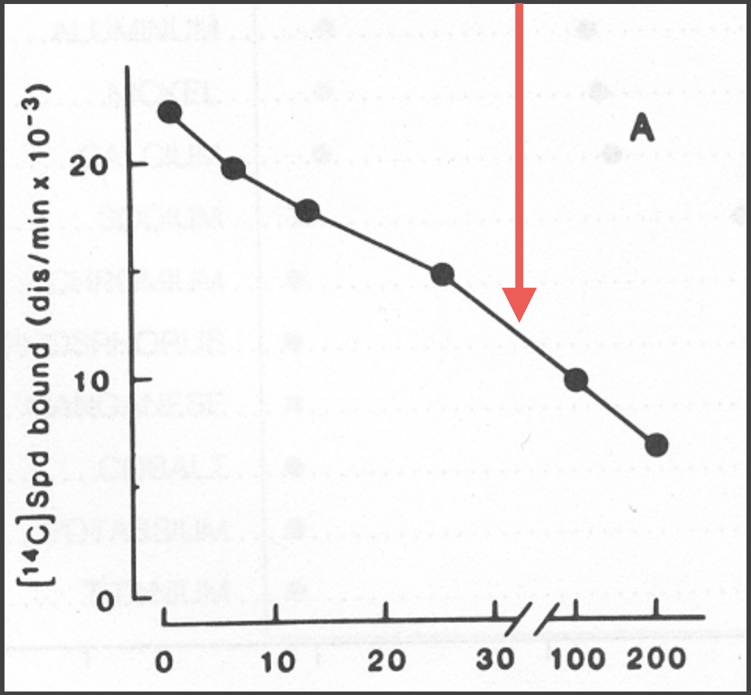


One Option: Clip Outliers

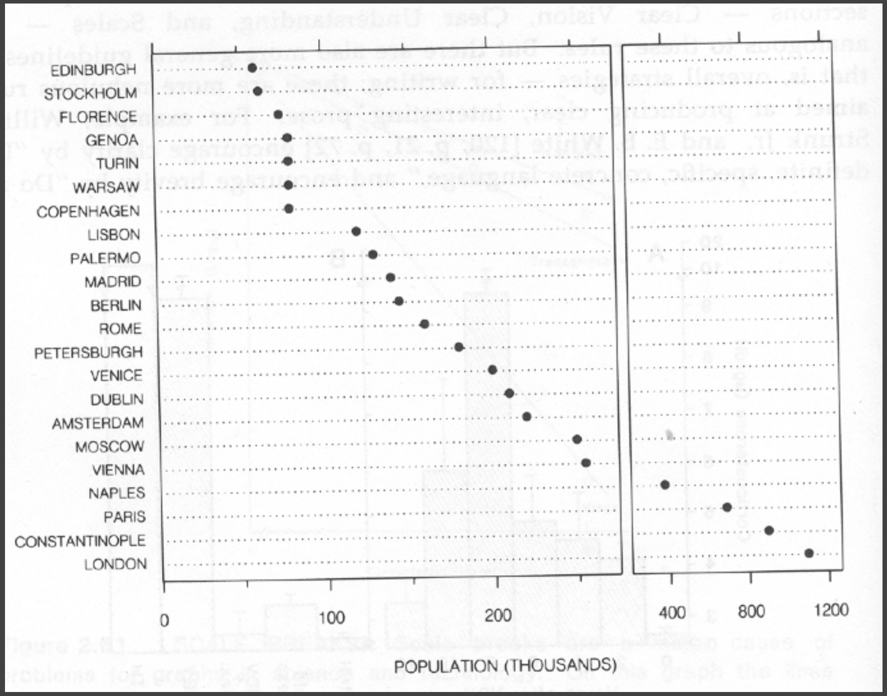


Clearly Mark Scale Breaks

Violates Expressiveness Principle!

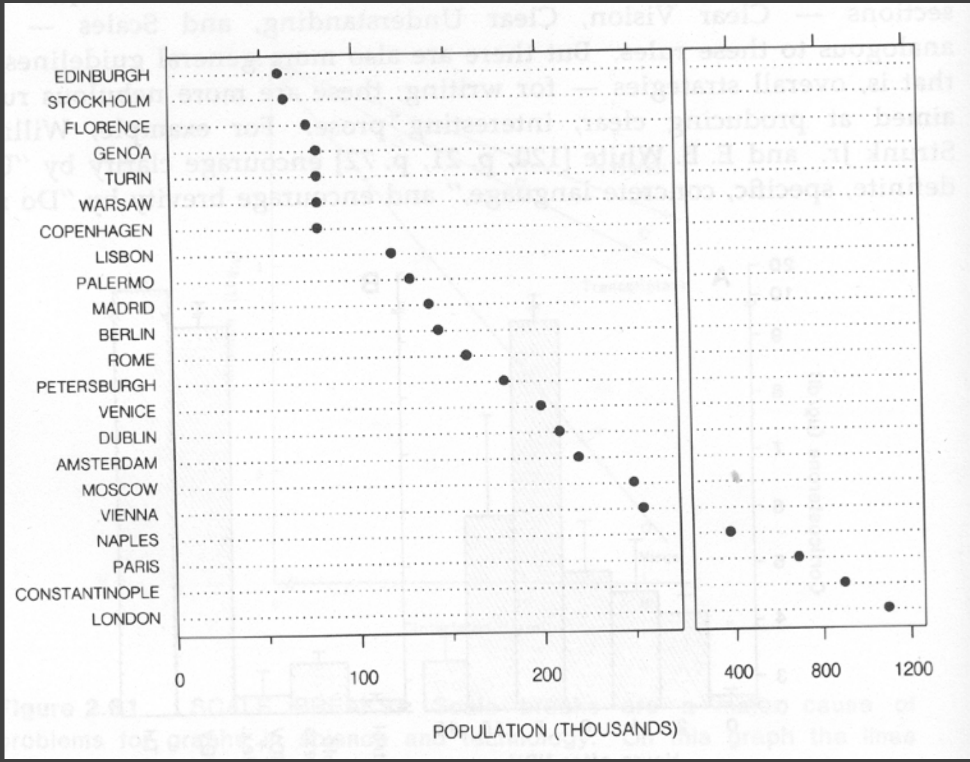


Poor scale break [Cleveland 85]

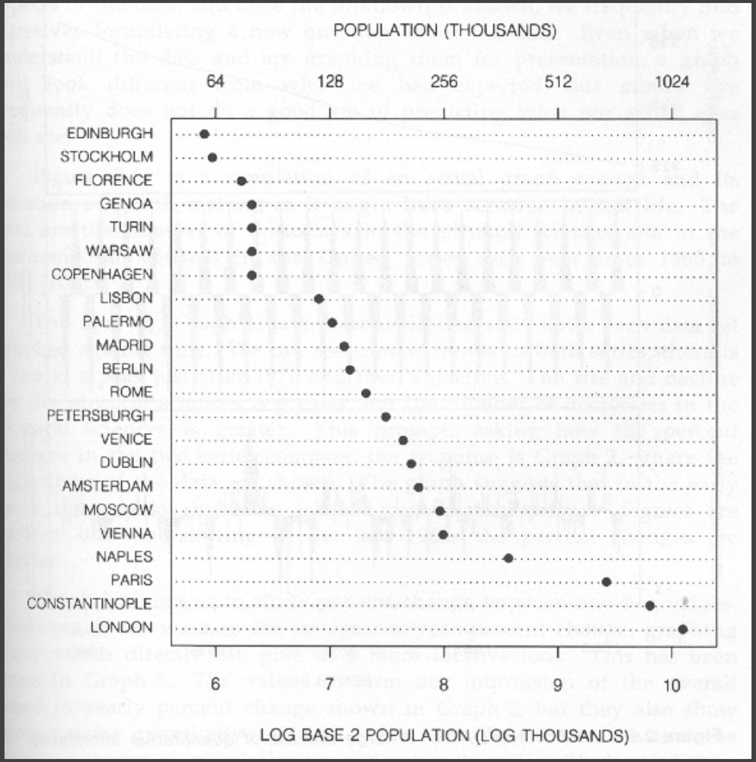


Well-marked scale break [Cleveland 85]

Scale Break vs. Log Scale



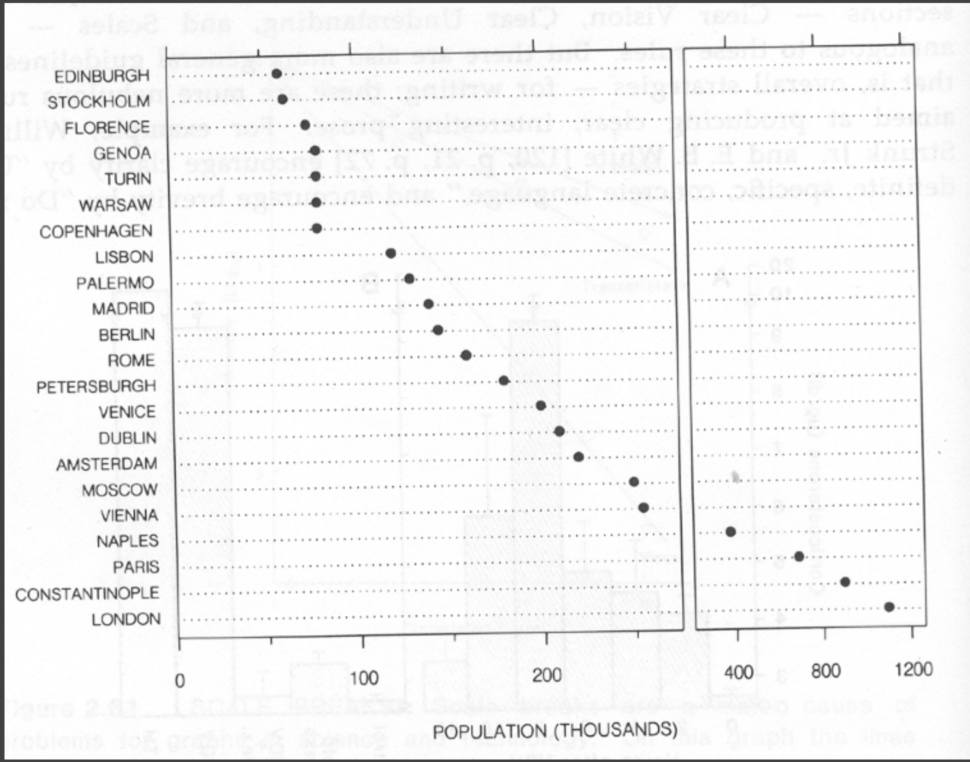
Scale Break

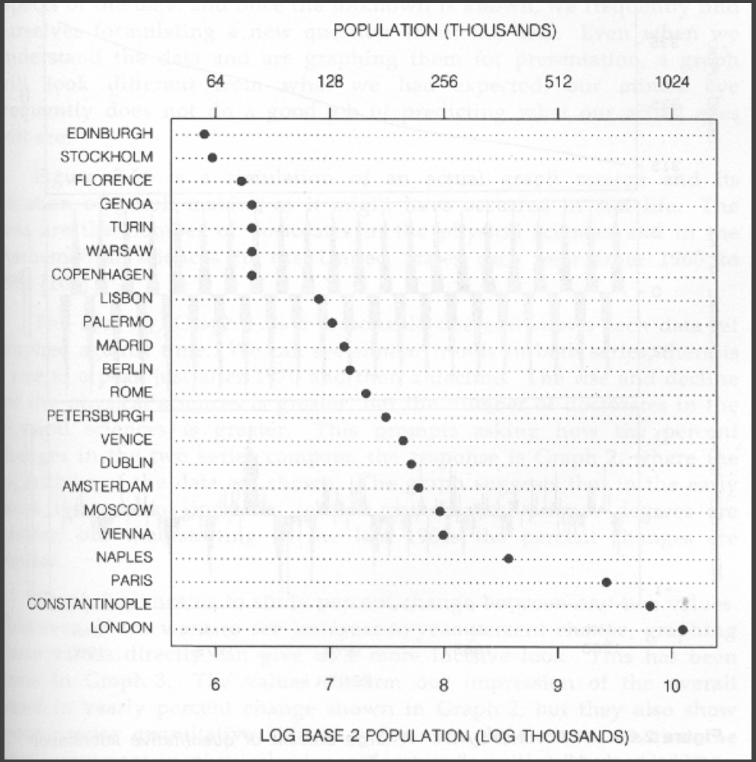


Log Scale

[Cleveland 85]

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

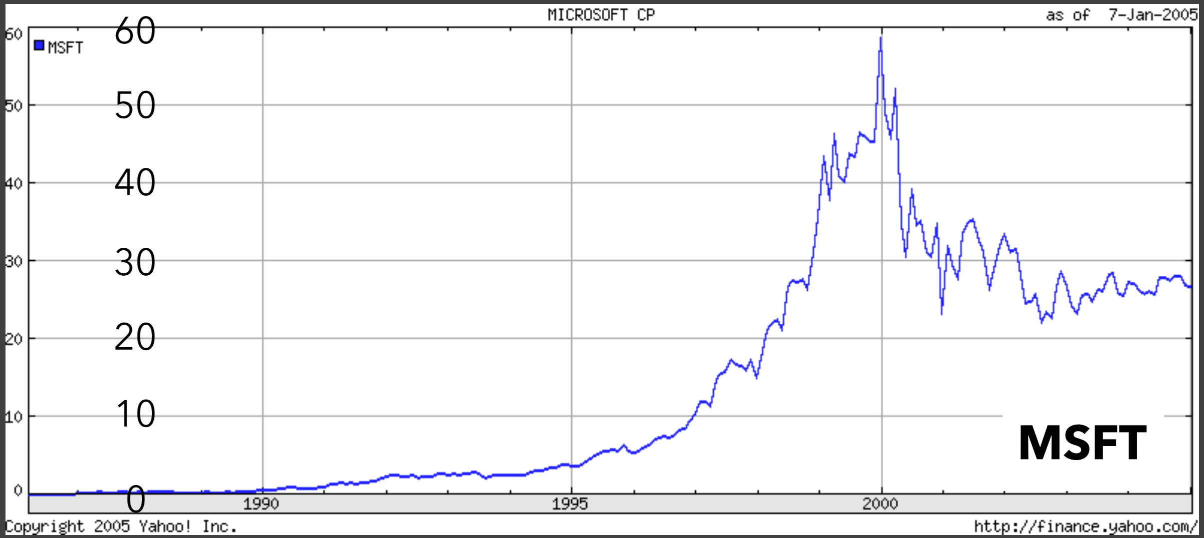
Logarithms turn *multiplication* into *addition*.

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

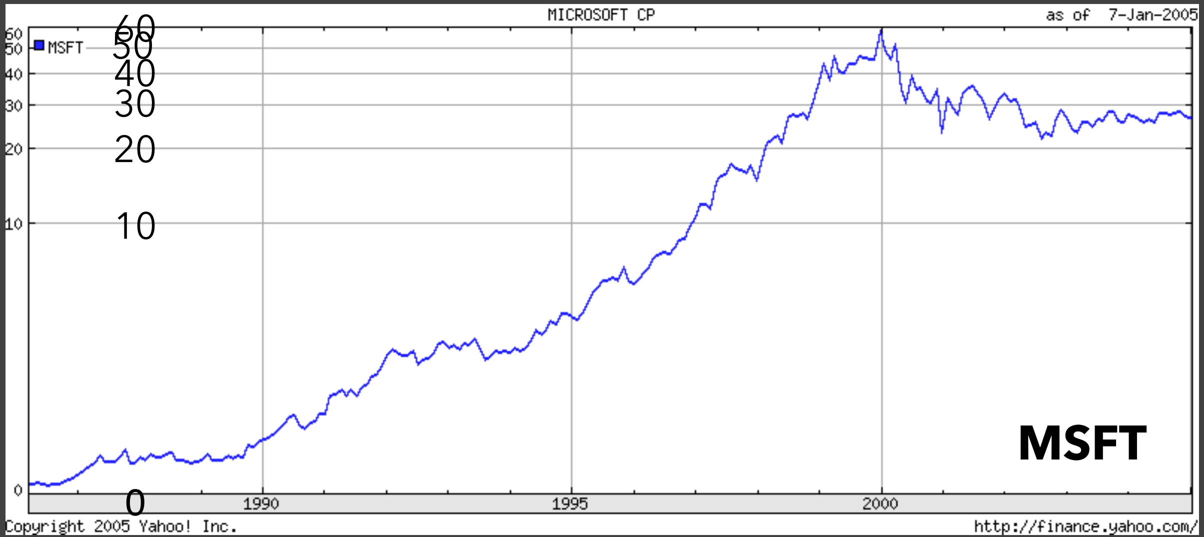
Equal steps on a log scale correspond to equal changes to a multiplicative scale factor.

Linear Scale vs. Log Scale

Linear Scale



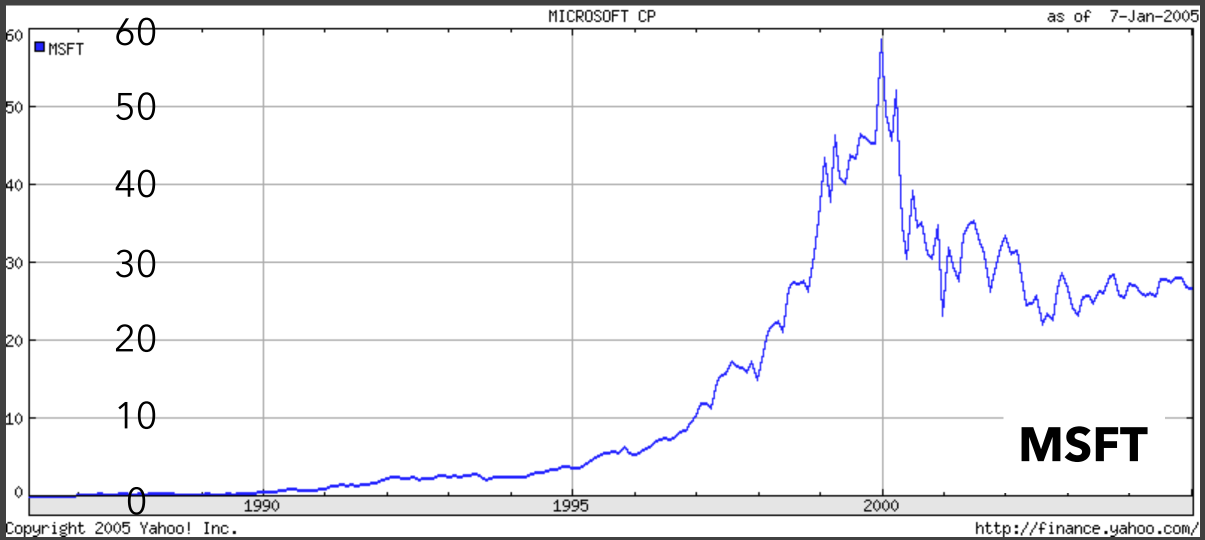
Log Scale



Linear Scale vs. Log Scale

Linear Scale

Absolute change

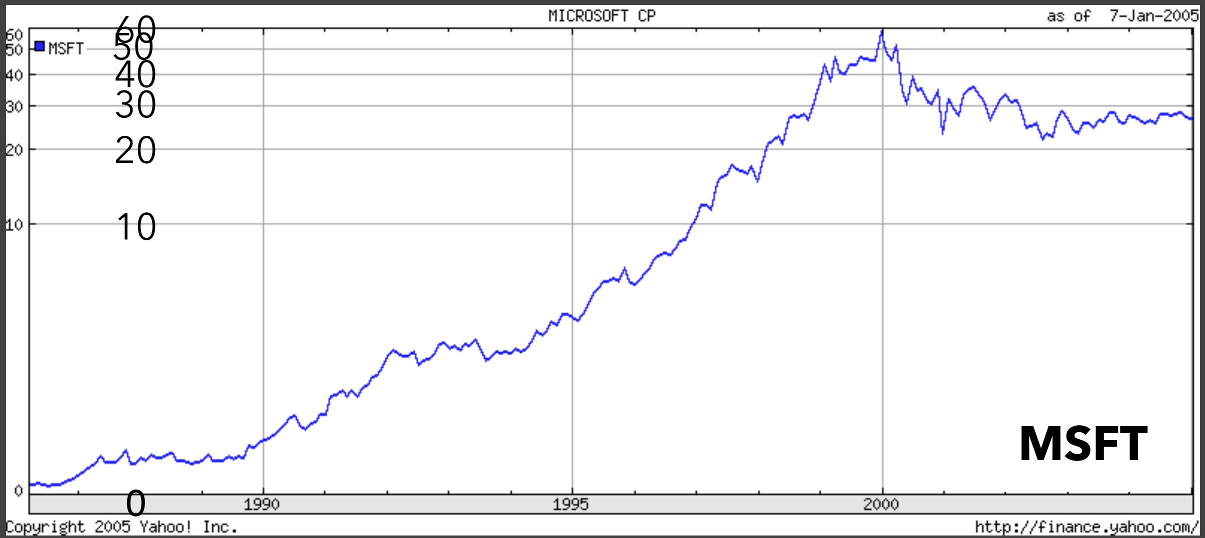


Log Scale

Small fluctuations

Percent change

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



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 **×** to **+**!

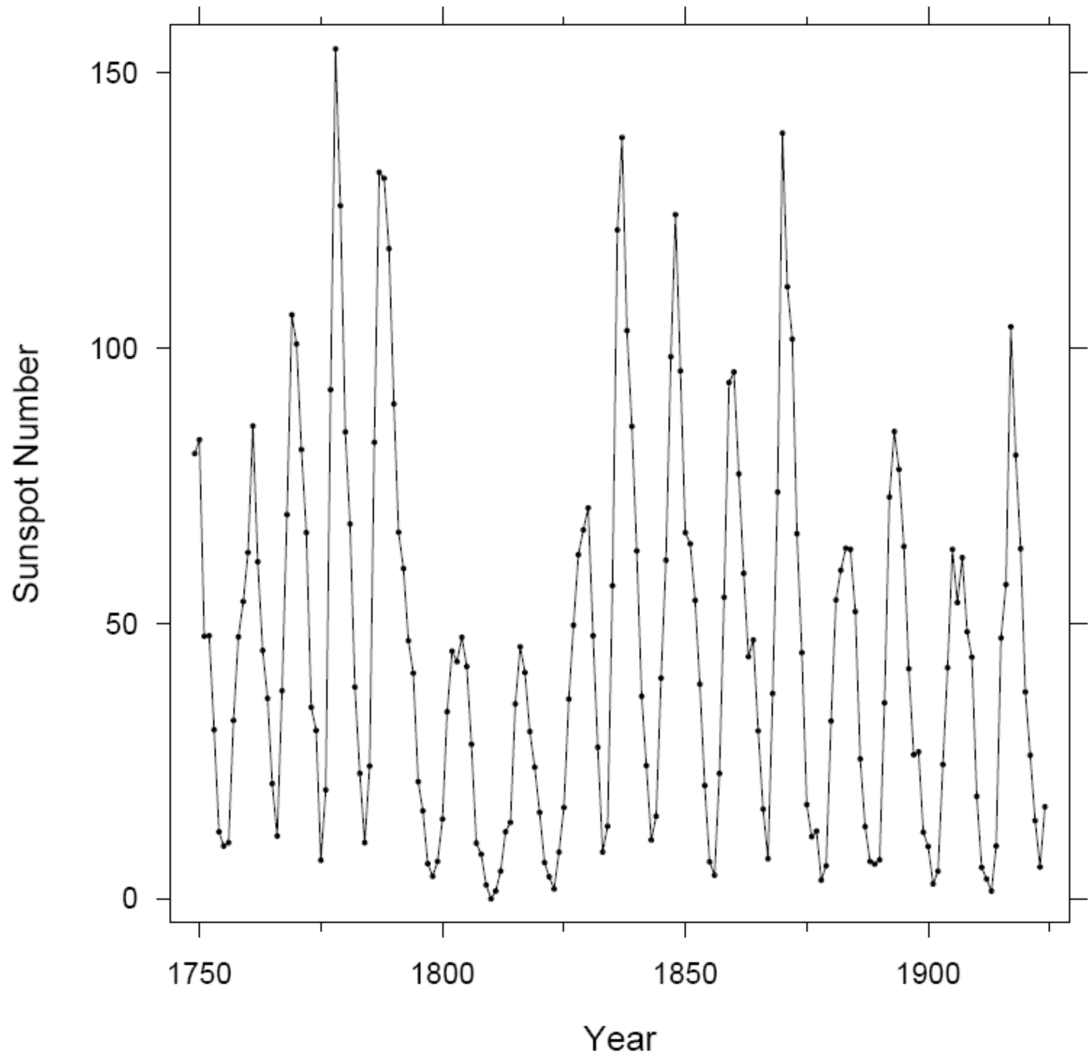
Percentage change, not linear difference.

Constraint: **positive, non-zero values**

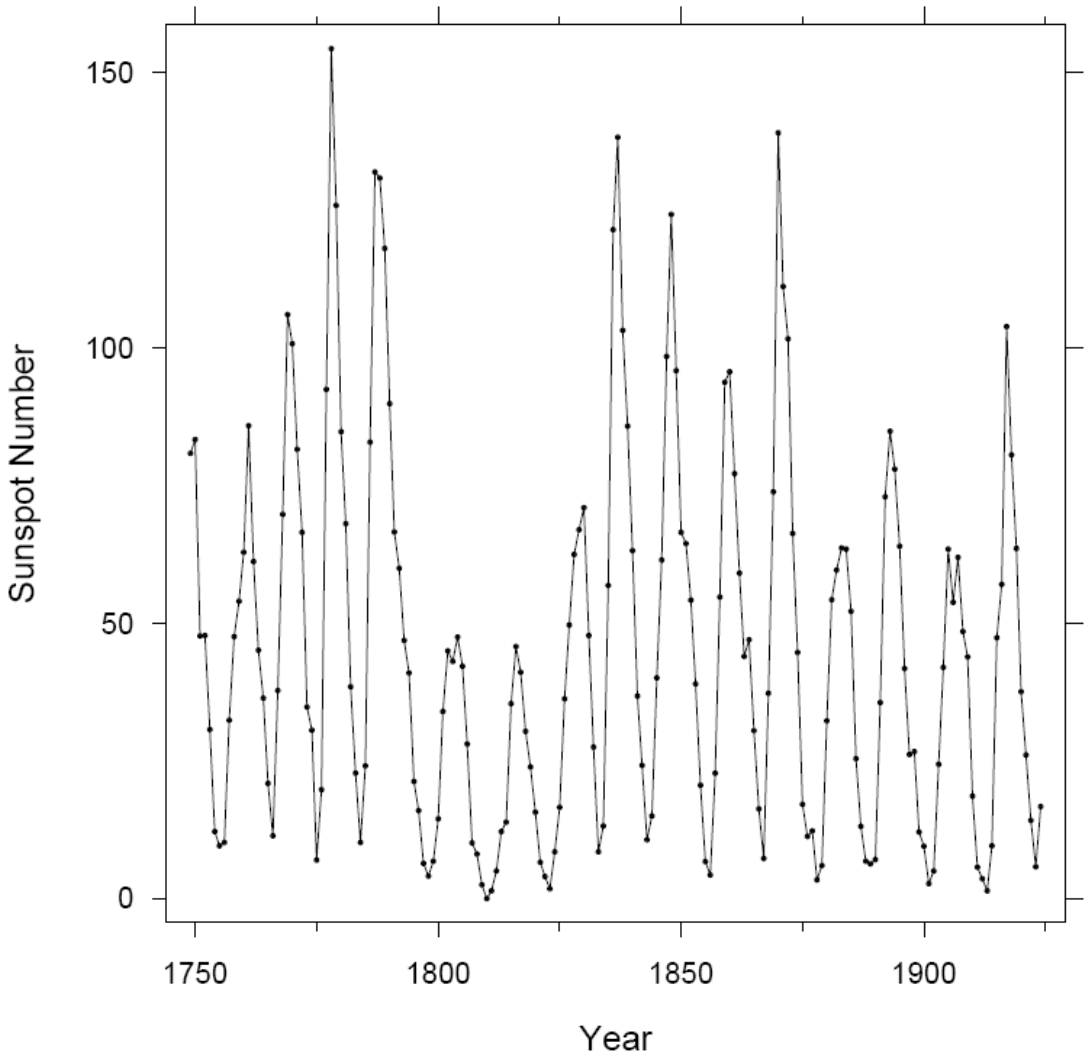
Constraint: **audience familiarity?**

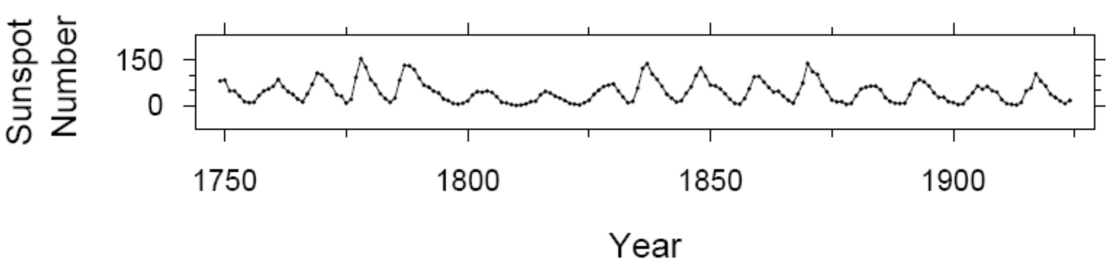
Aspect Ratio

(width : height)



William S. Cleveland
The Elements of Graphing Data

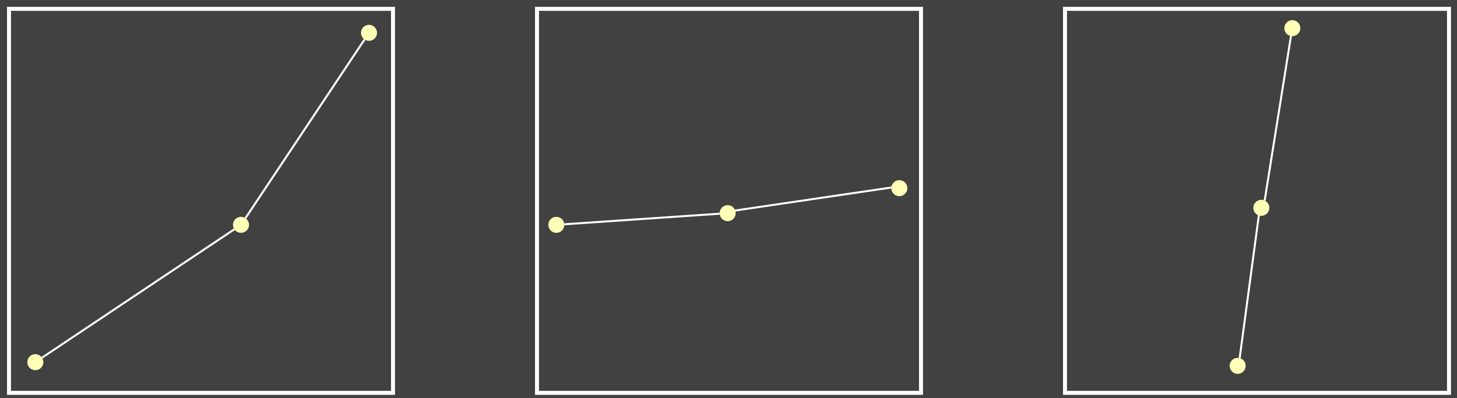




William S. Cleveland
The Elements of Graphing Data

Banking to 45° [Cleveland]

To facilitate perception of trends, maximize the discriminability of line segment orientations

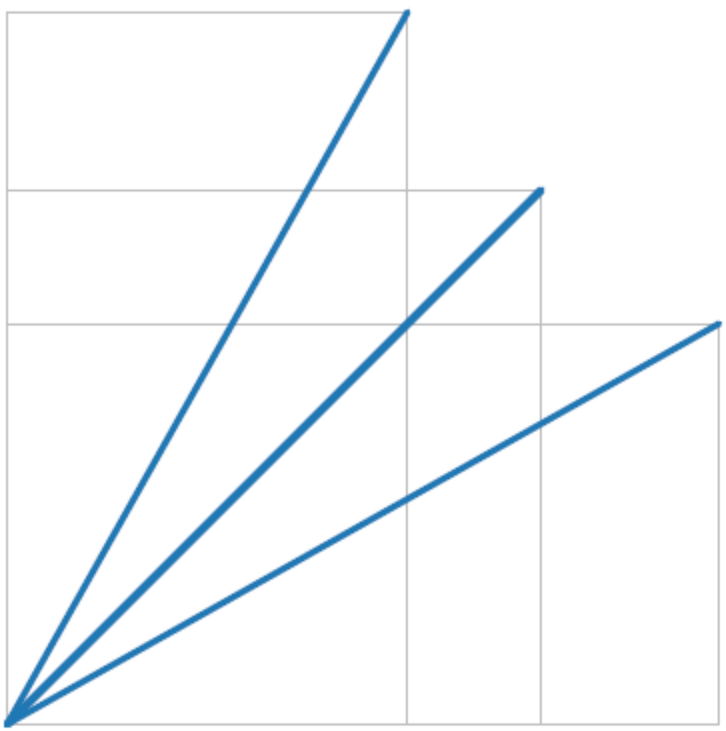


Two line segments are maximally discriminable when their average absolute angle is 45°

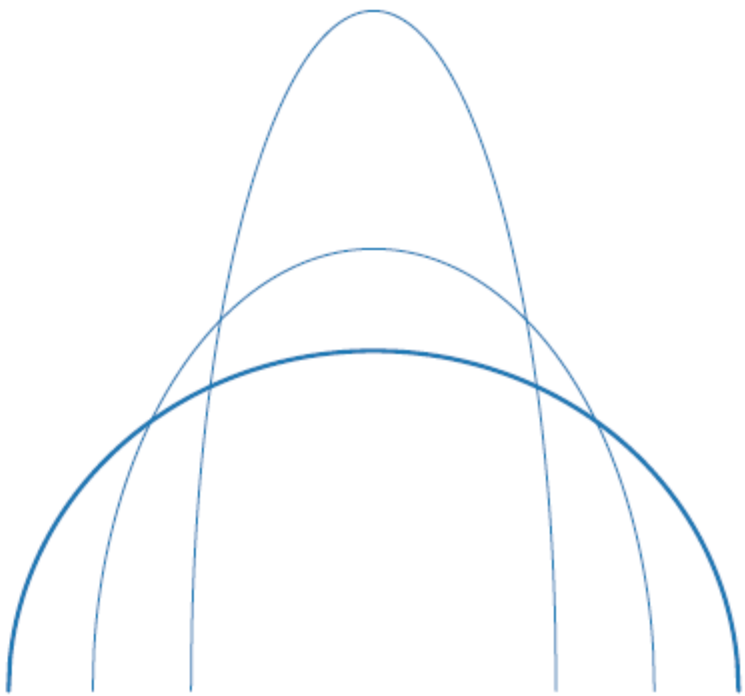
Method: optimize the aspect ratio such that the average absolute angle of all segments is 45°

Alternative: Minimize Arc Length

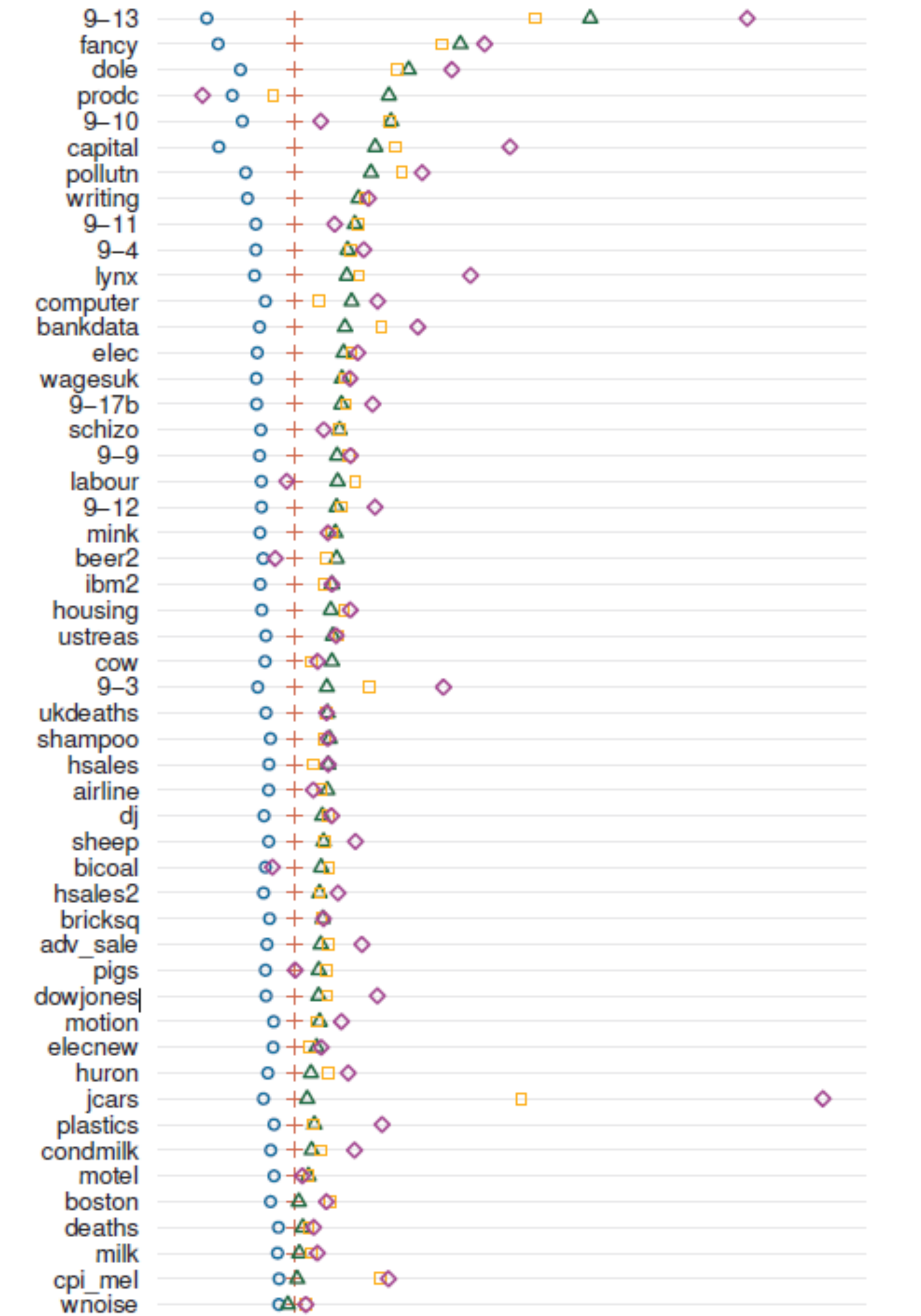
while holding area constant [Talbot et al. 2011]



Straight line -> 45°



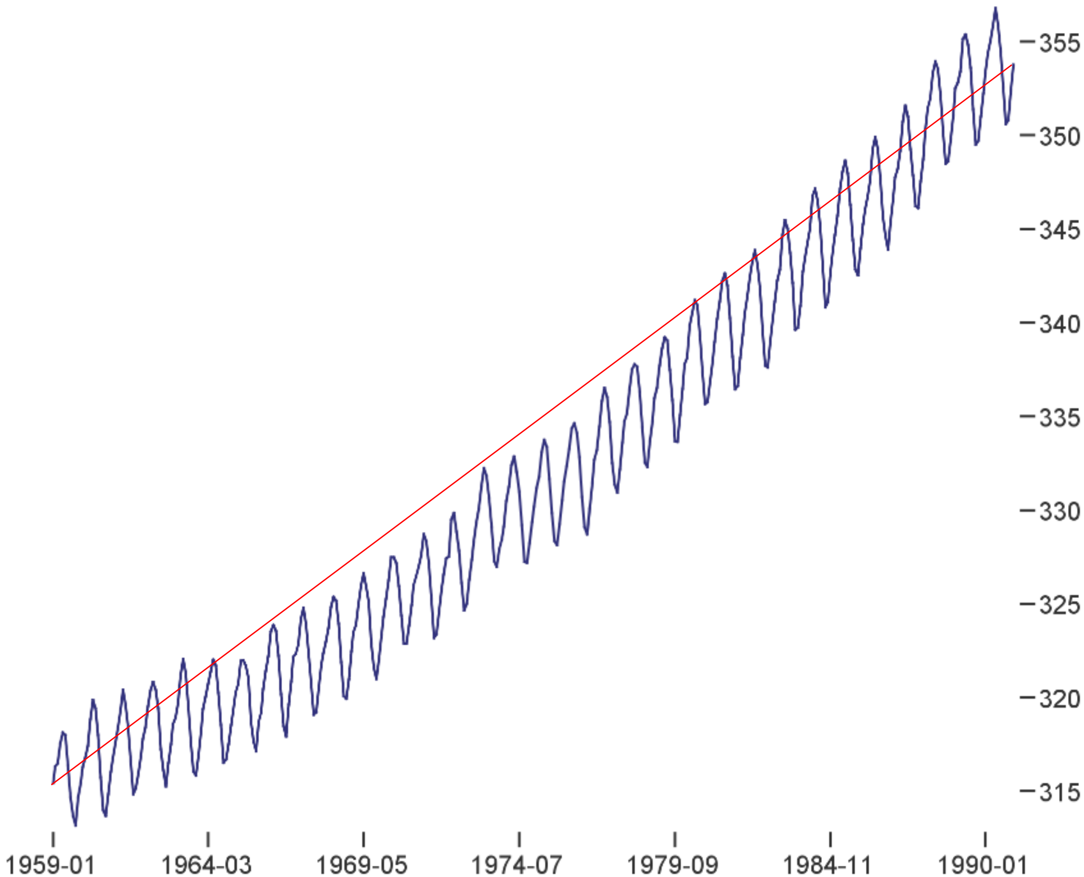
Ellipse -> Circle



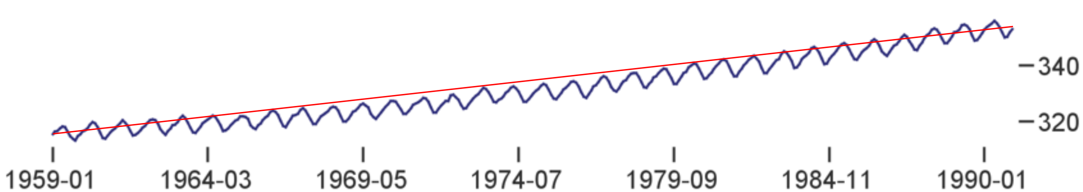
A Good Compromise

Arc-length banking produces aspect ratios in-between those produced by other methods.

[Talbot et al. 2011]



Aspect Ratio = 1.17



Aspect Ratio = 7.87

Trends may occur at different scales!

Apply banking to the original data *or* to fitted trend lines.

[Heer & Agrawala '06]

CO_2 Measurements

William S. Cleveland
Visualizing Data

Administrivia

A2: Exploratory Data Analysis

Use visualization software to form & answer questions

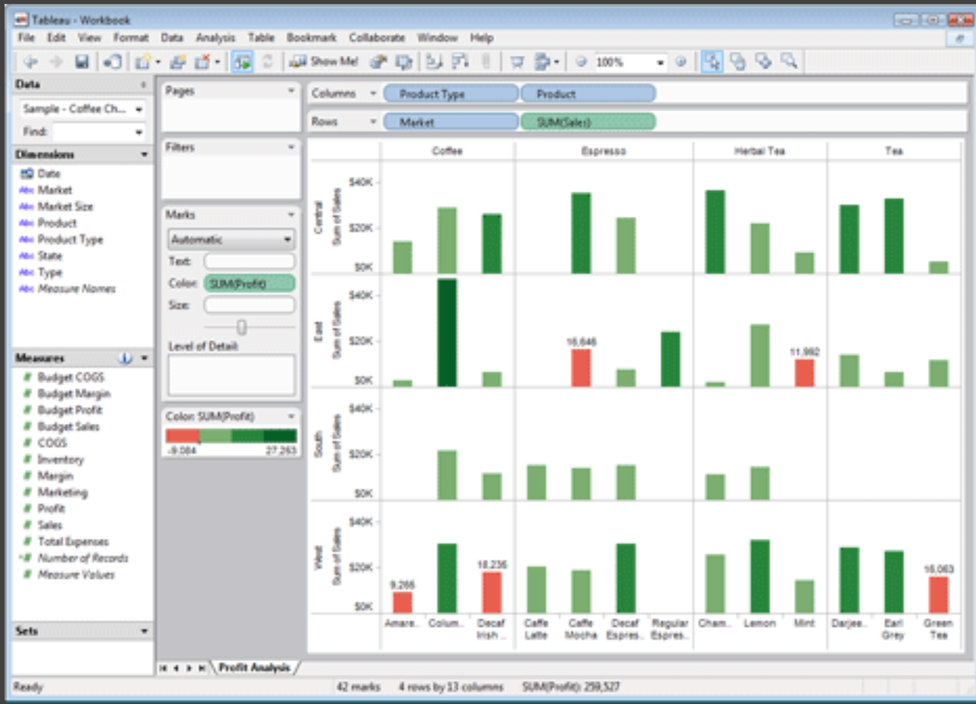
First steps:

Step 1: Pick domain & data

Step 2: Pose questions

Step 3: Profile the data

Iterate as needed



Create visualizations

Interact with data

Refine your questions

Author a report

Screenshots of most insightful views *(8+)*

Include titles and captions for each view

Due by 11:59pm

Friday, Apr 23

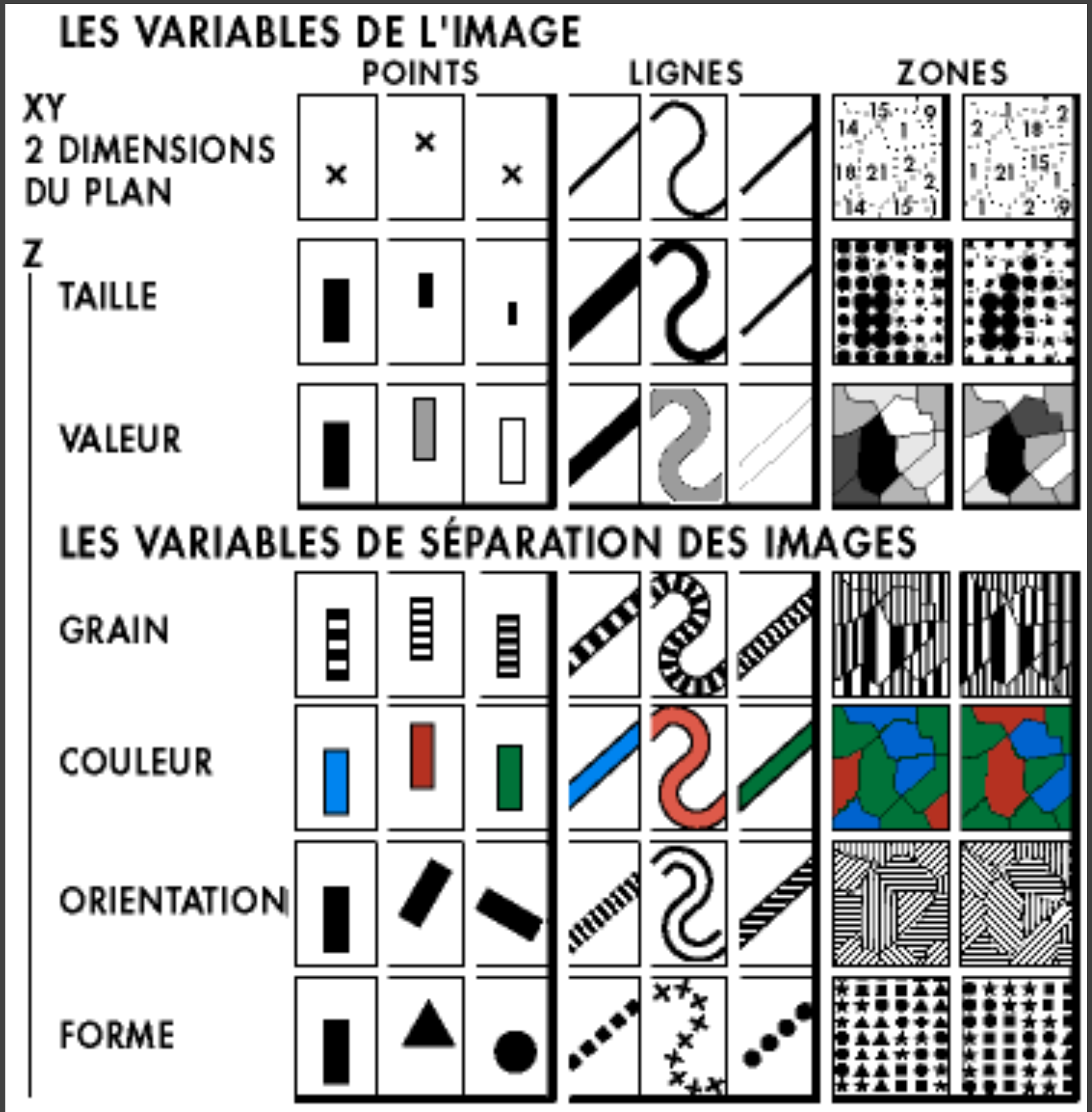
Break Time!

Multidimensional Data

Visual Encoding Variables

Position (X)
Position (Y)
Area
Value
Texture
Color
Orientation
Shape

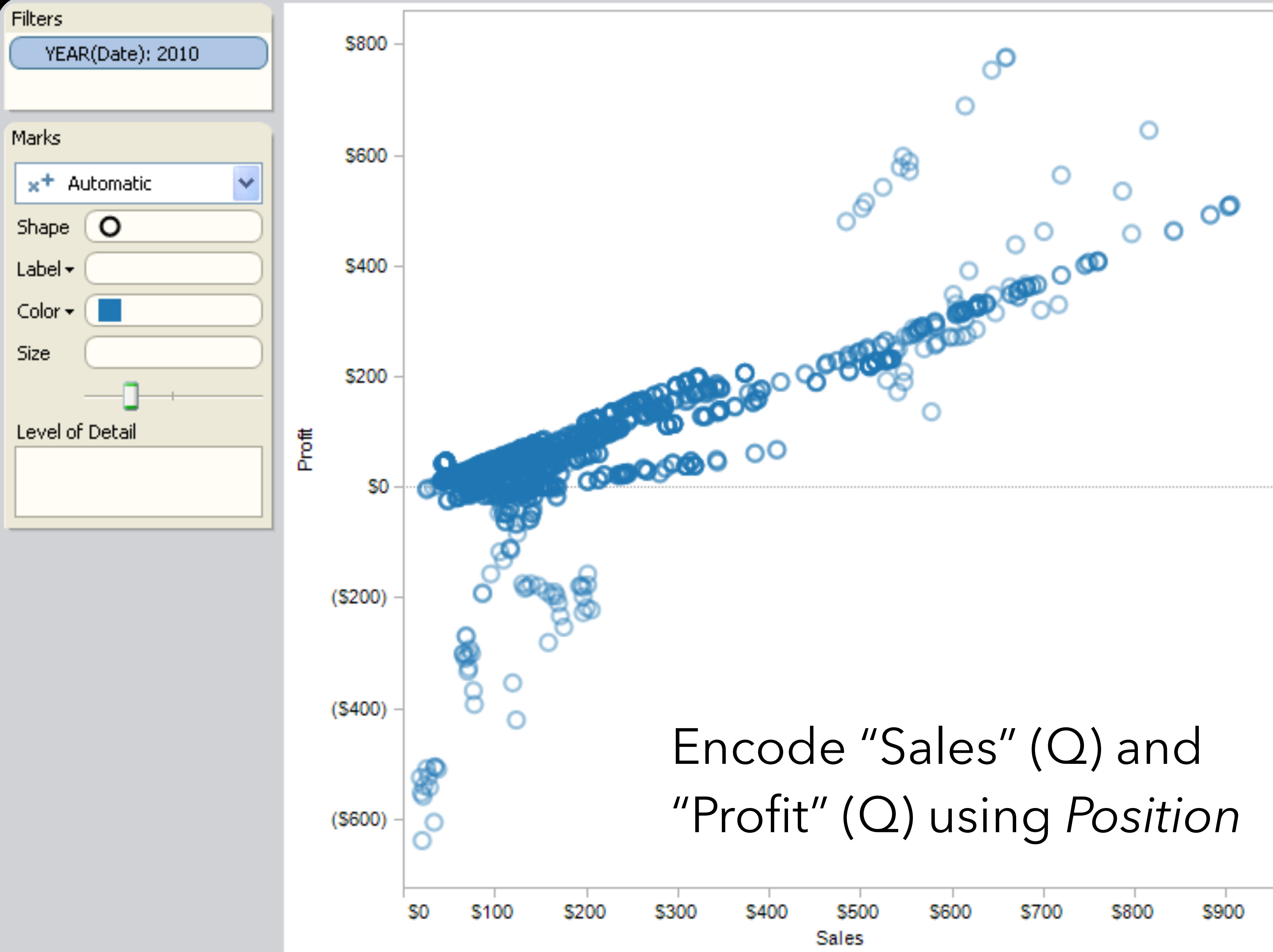
~8 dimensions?



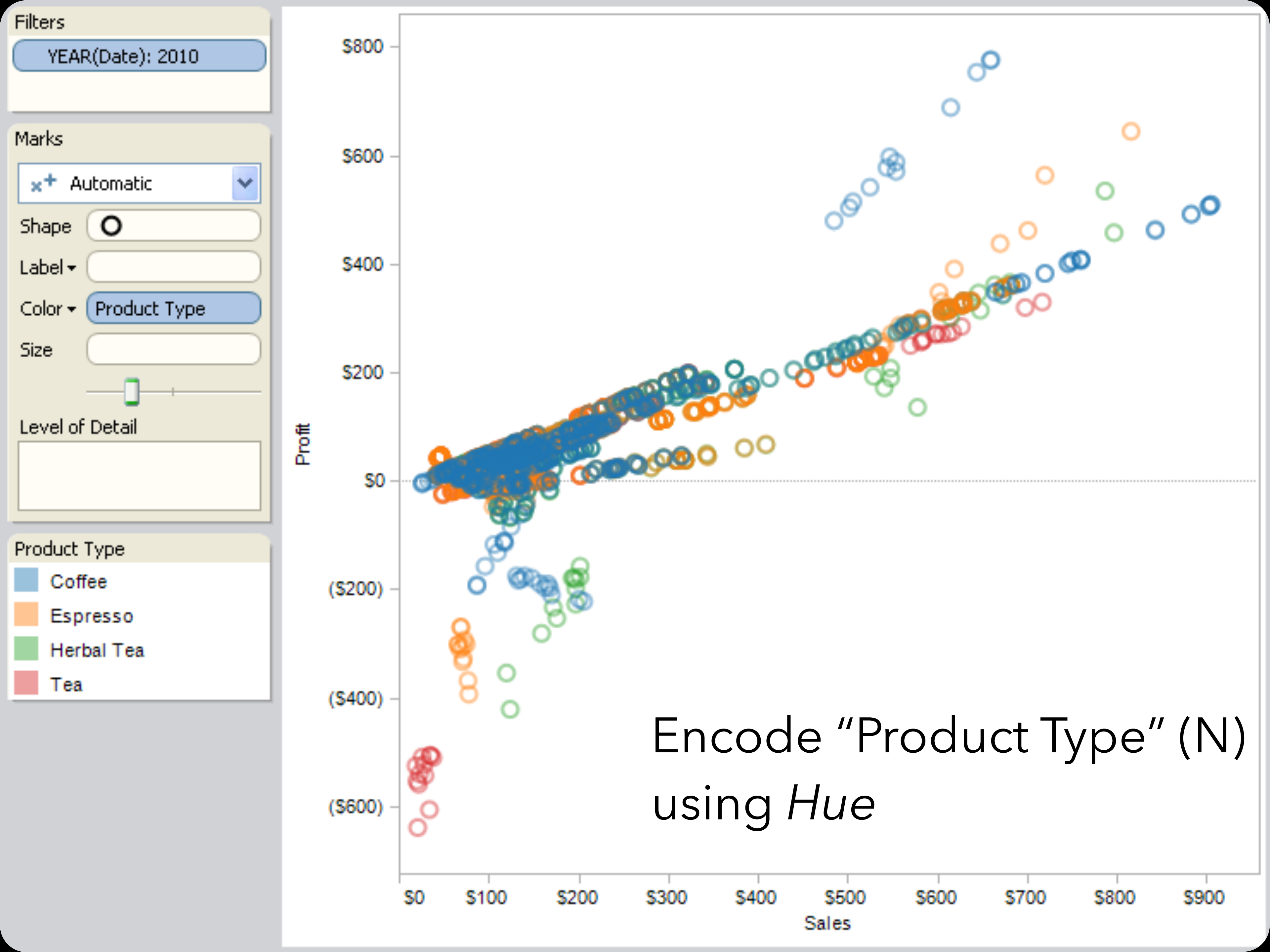
Example: Coffee Sales

Sales figures for a fictional coffee chain

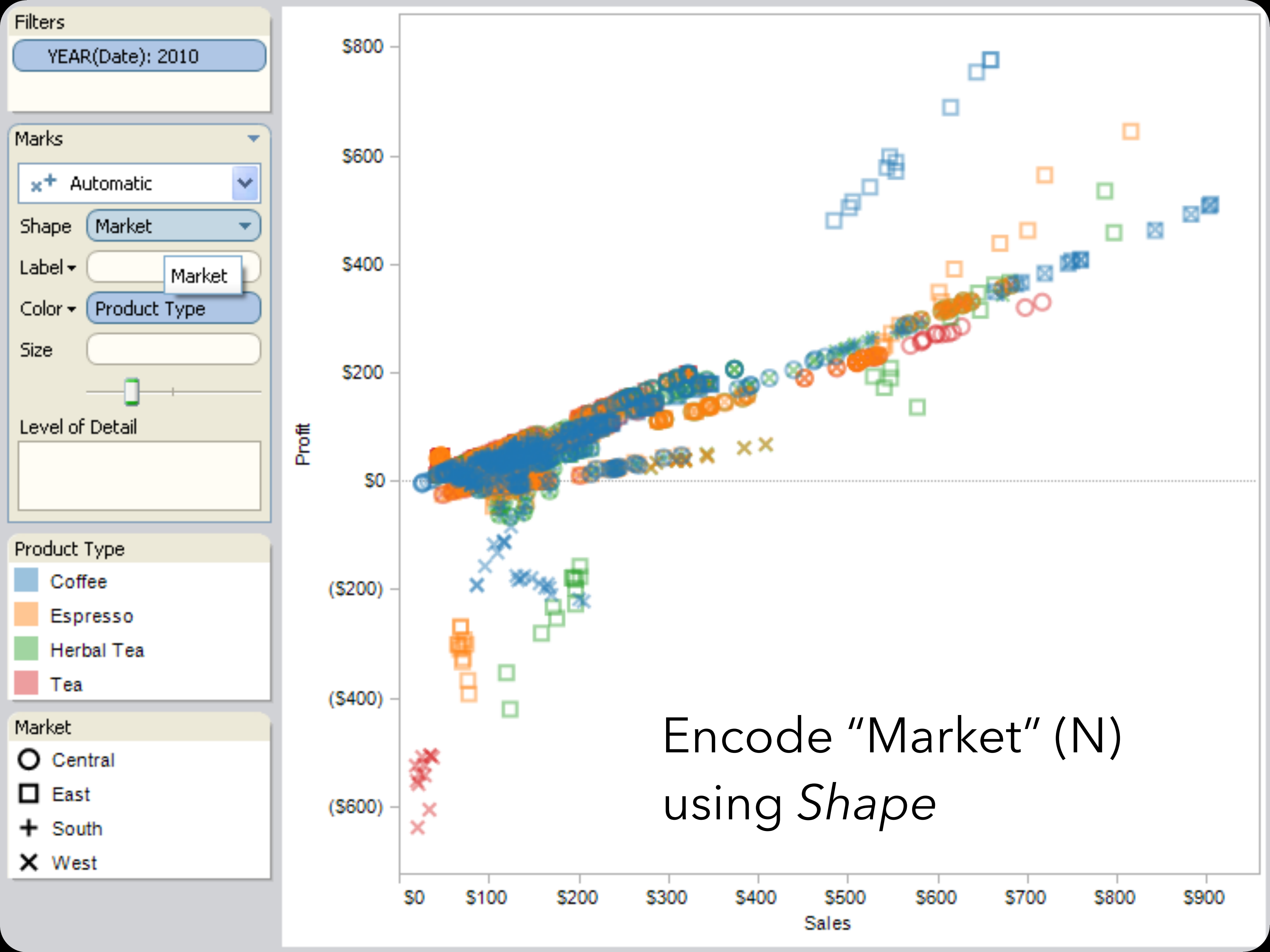
Sales	Q-Ratio
Profit	Q-Ratio
Marketing	Q-Ratio
Product Type	N {Coffee, Espresso, Herbal Tea, Tea}
Market	N {Central, East, South, West}



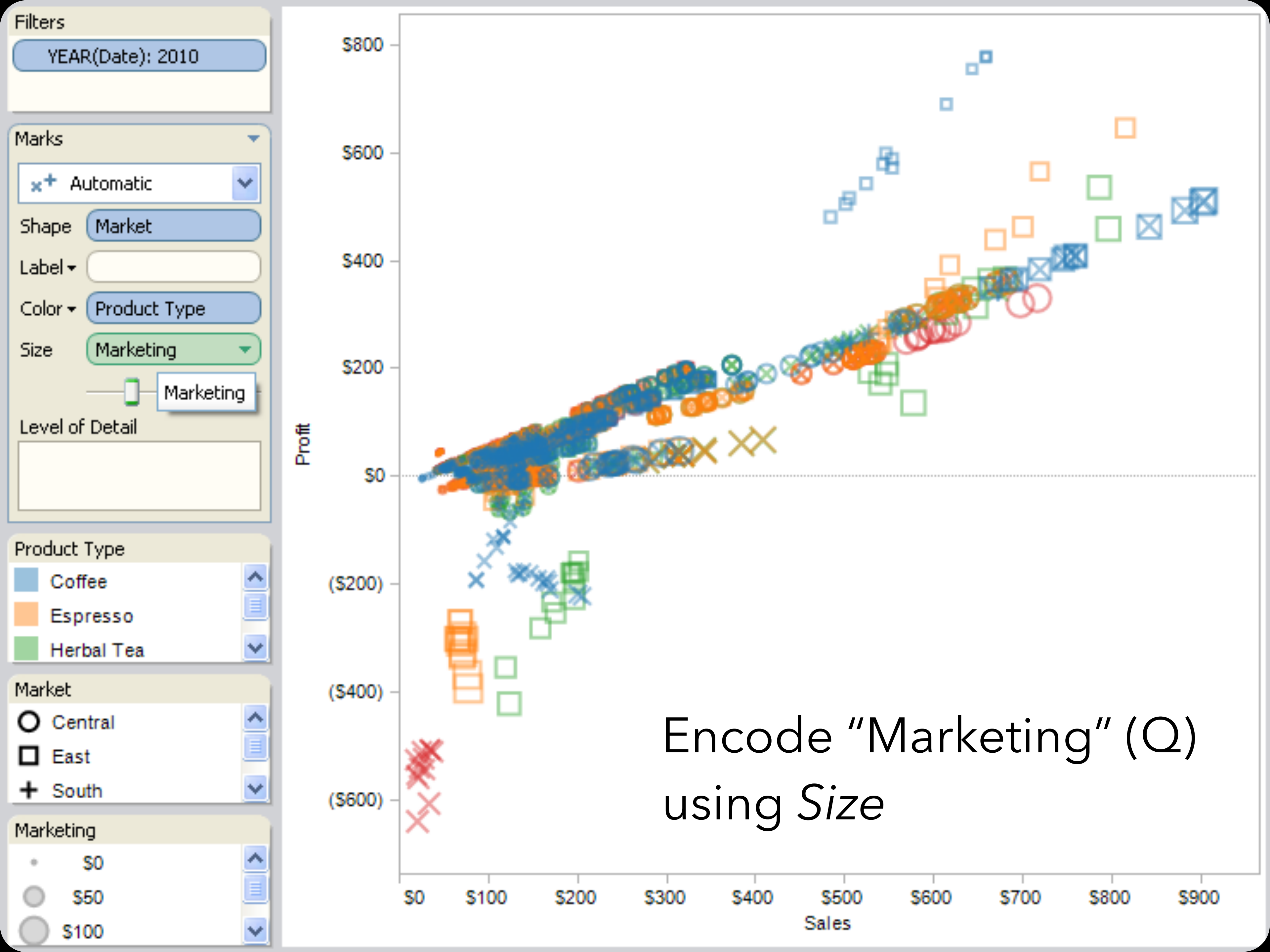
Encode "Sales" (Q) and "Profit" (Q) using *Position*



Encode "Product Type" (N) using *Hue*

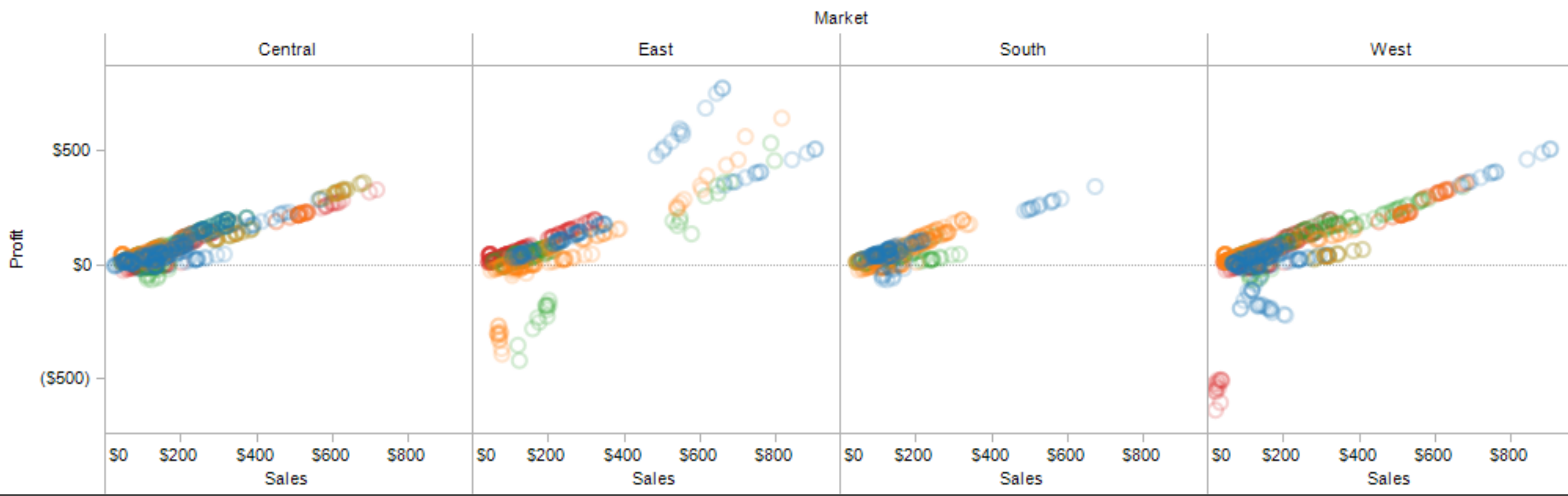


Encode "Market" (N) using *Shape*



Encode "Marketing" (Q) using *Size*

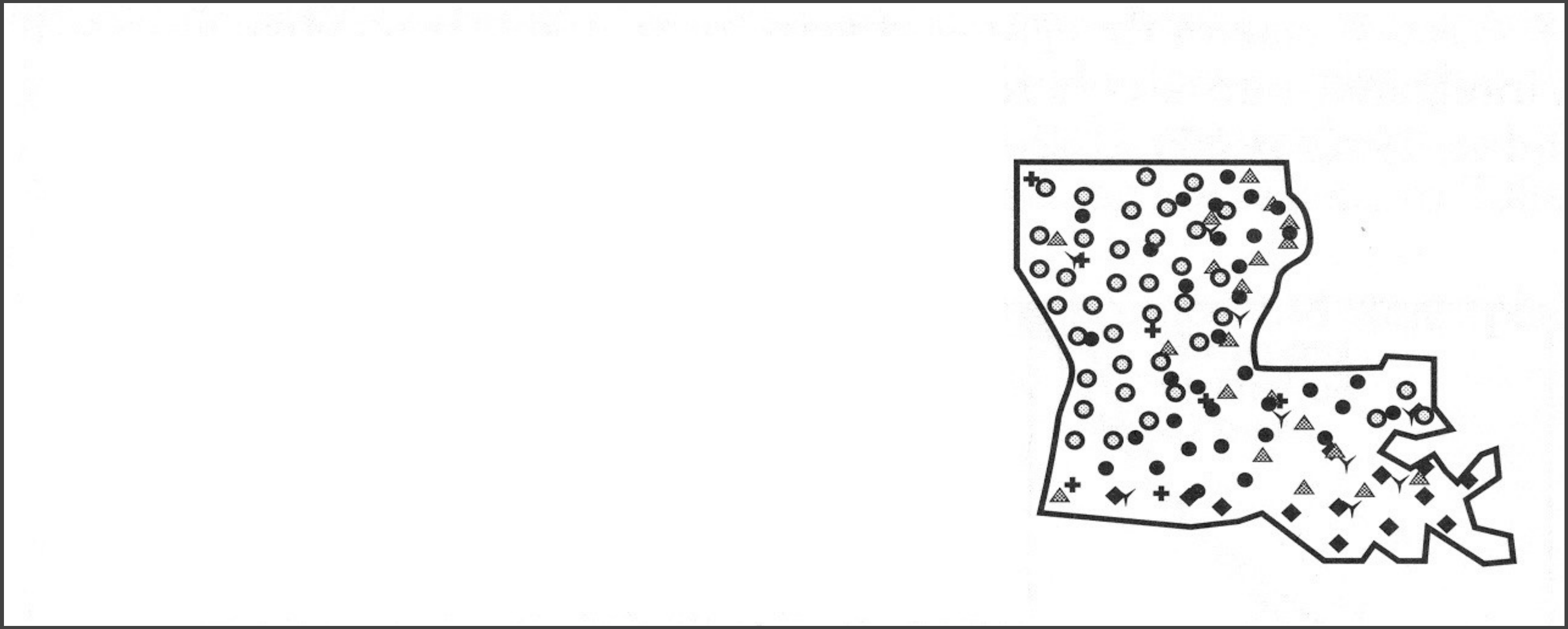
Trellis Plots



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

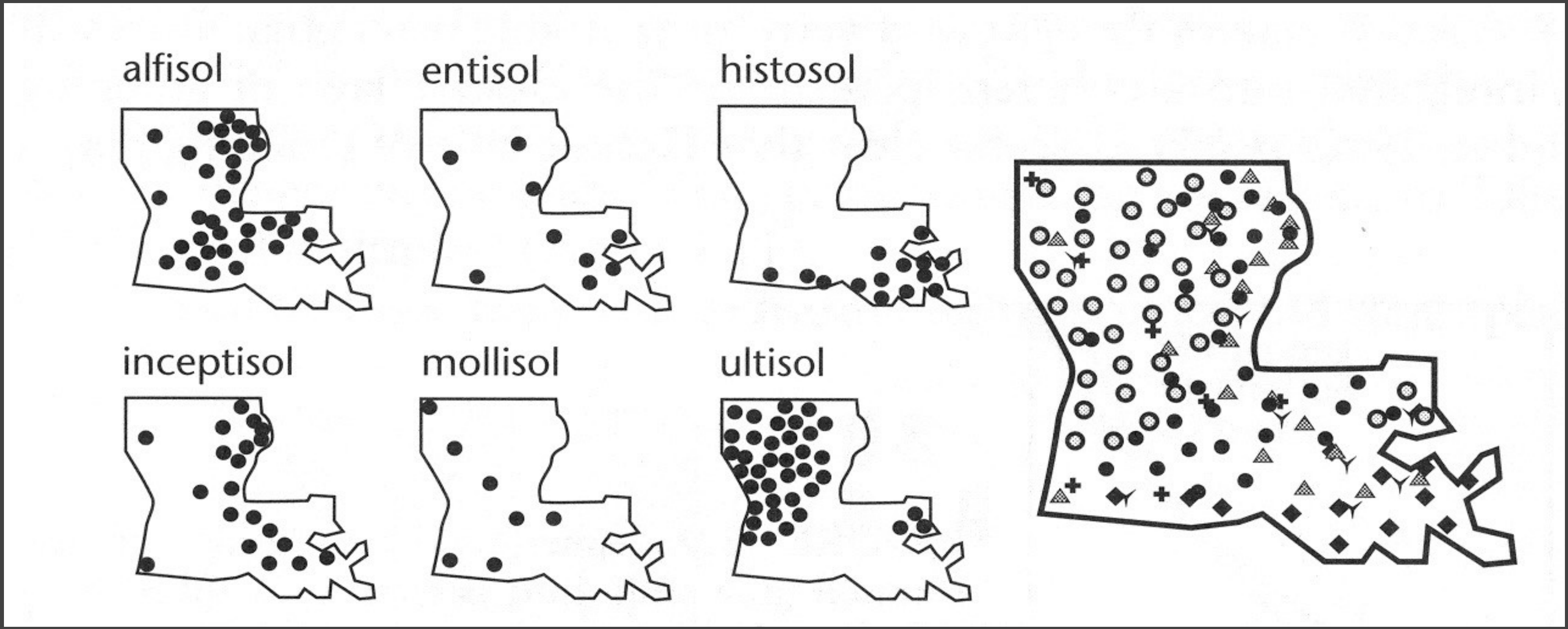
Typically nominal or ordinal variables are used as dimensions for subdivision.

Small Multiples



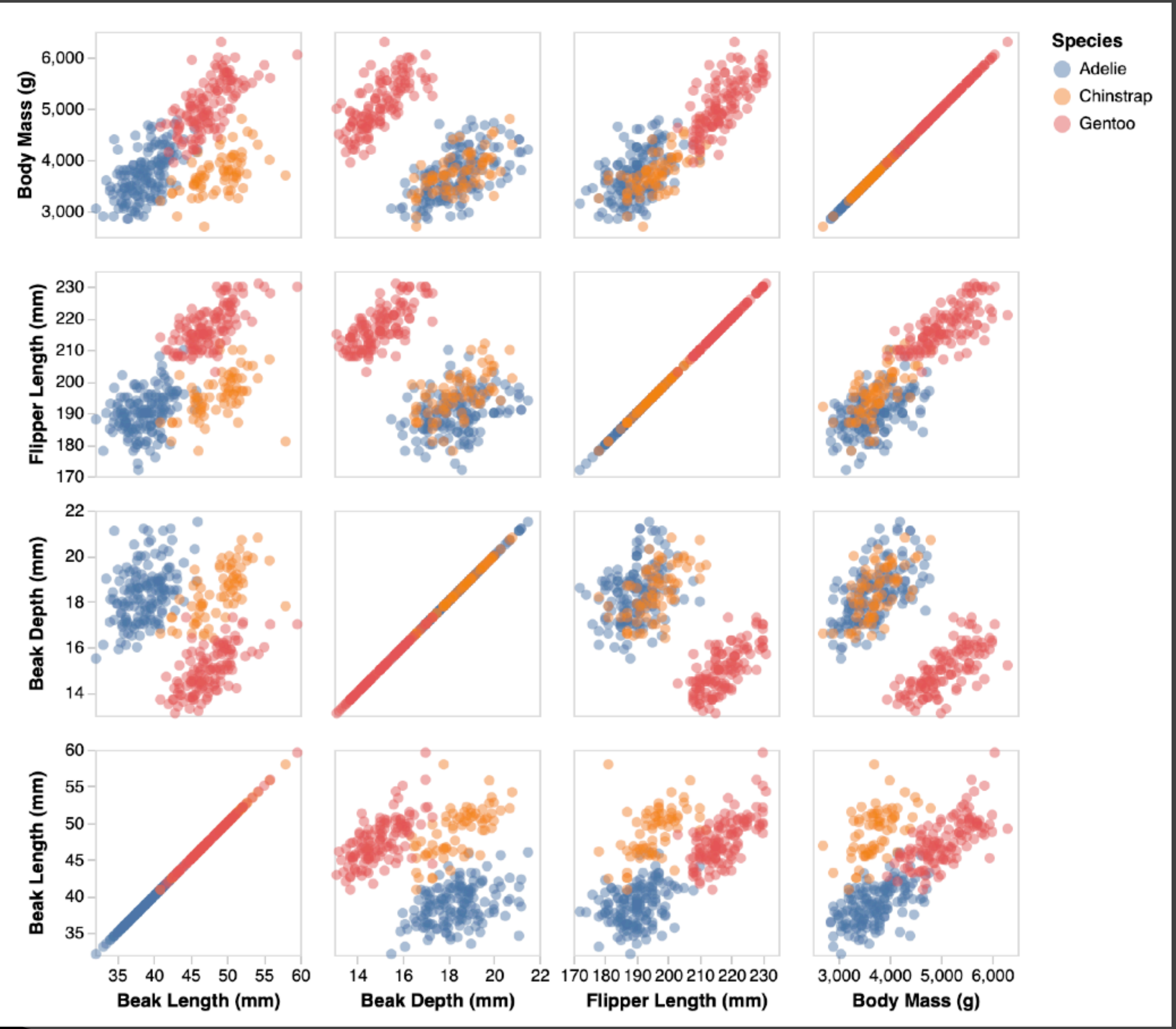
[MacEachren '95, Figure 2.11, p. 38]

Small Multiples



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

Scatterplot Matrix (SPLOM)



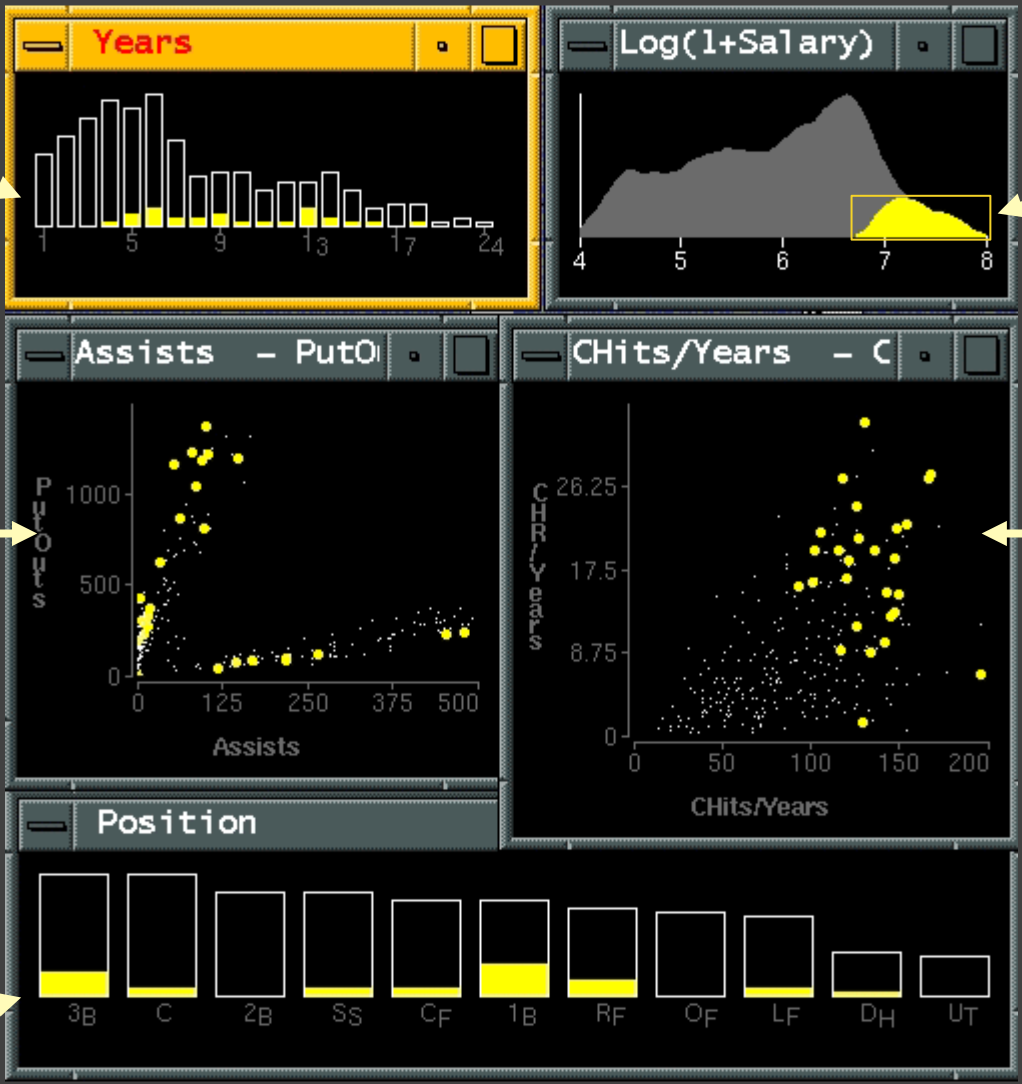
Scatter plots for pairwise comparison of each data dimension.

Multiple Coordinated Views

how long in majors

avg assists vs avg putouts (fielding ability)

distribution of positions played

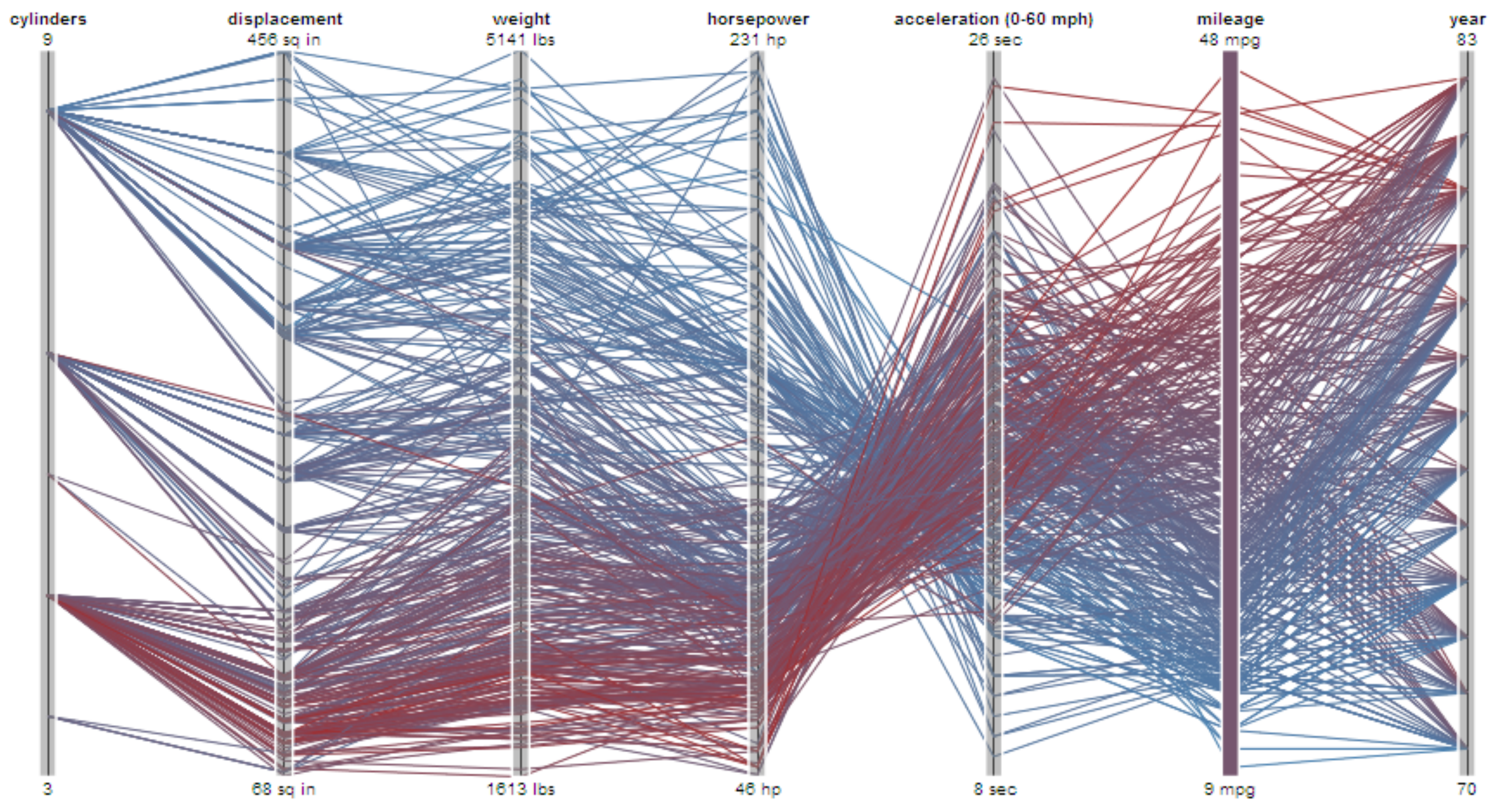


select high salaries

avg career HRs vs avg career hits (batting ability)

Parallel Coordinates

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

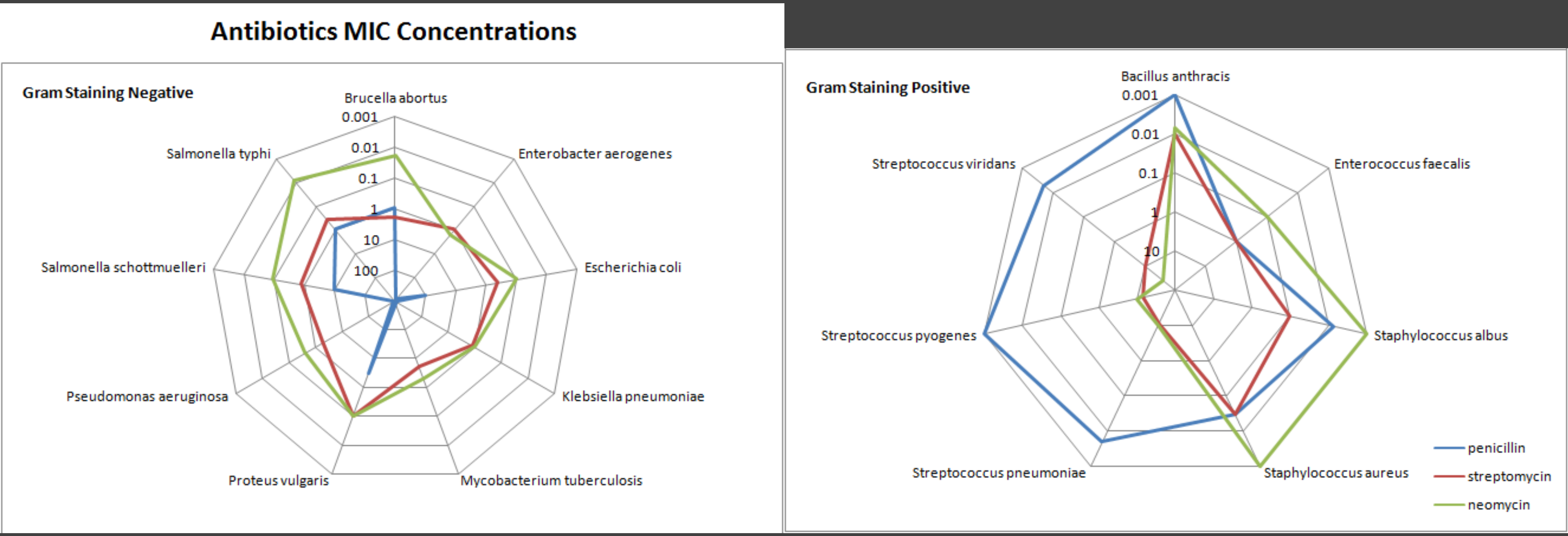
Between adjacent axes: line crossings imply neg. correlation, shared slopes imply pos. correlation.

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



“Parallel” dimensions in polar coordinate space

Best if same units apply to each axis

[Loren Yu, CS448B 2009]

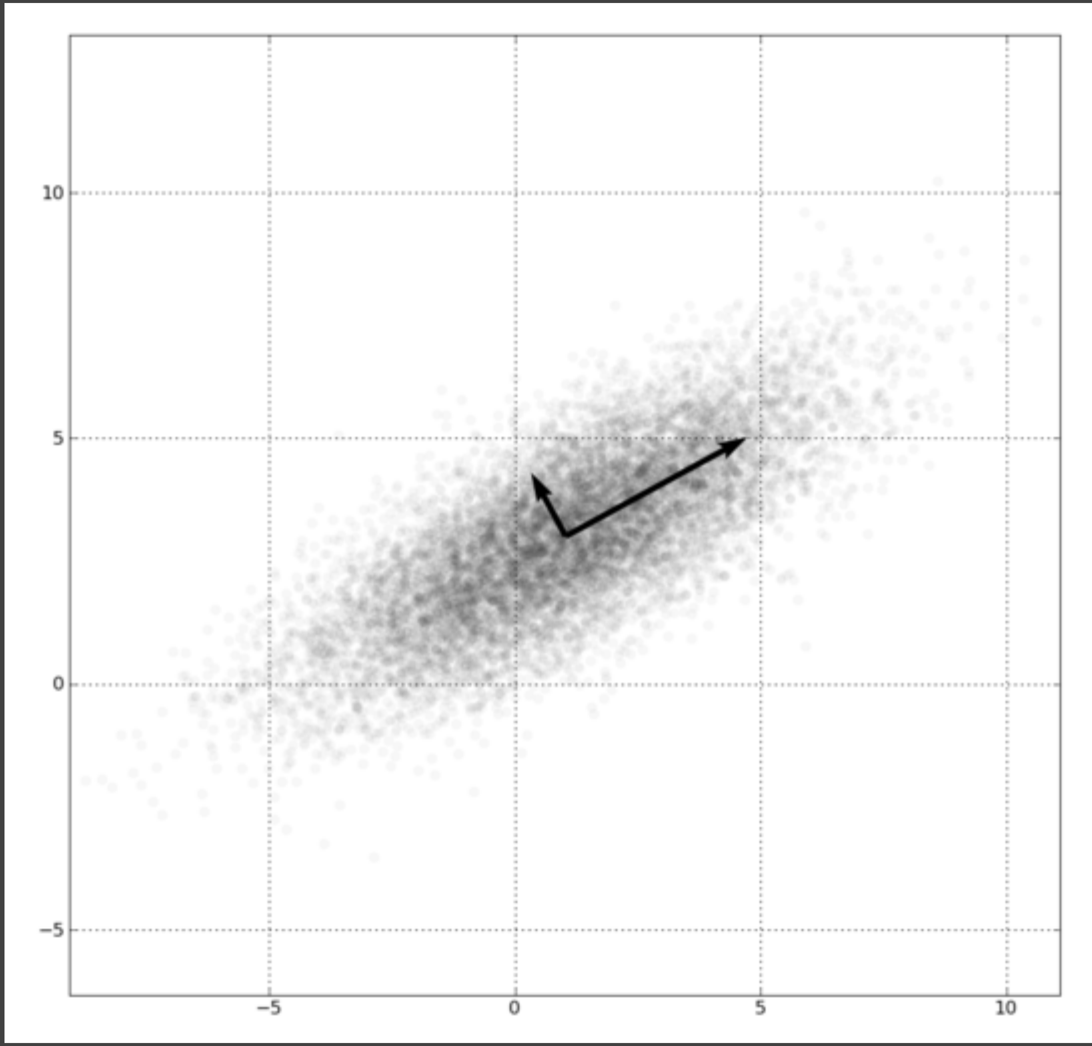
Dimensionality Reduction

Dimensionality Reduction (DR)

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

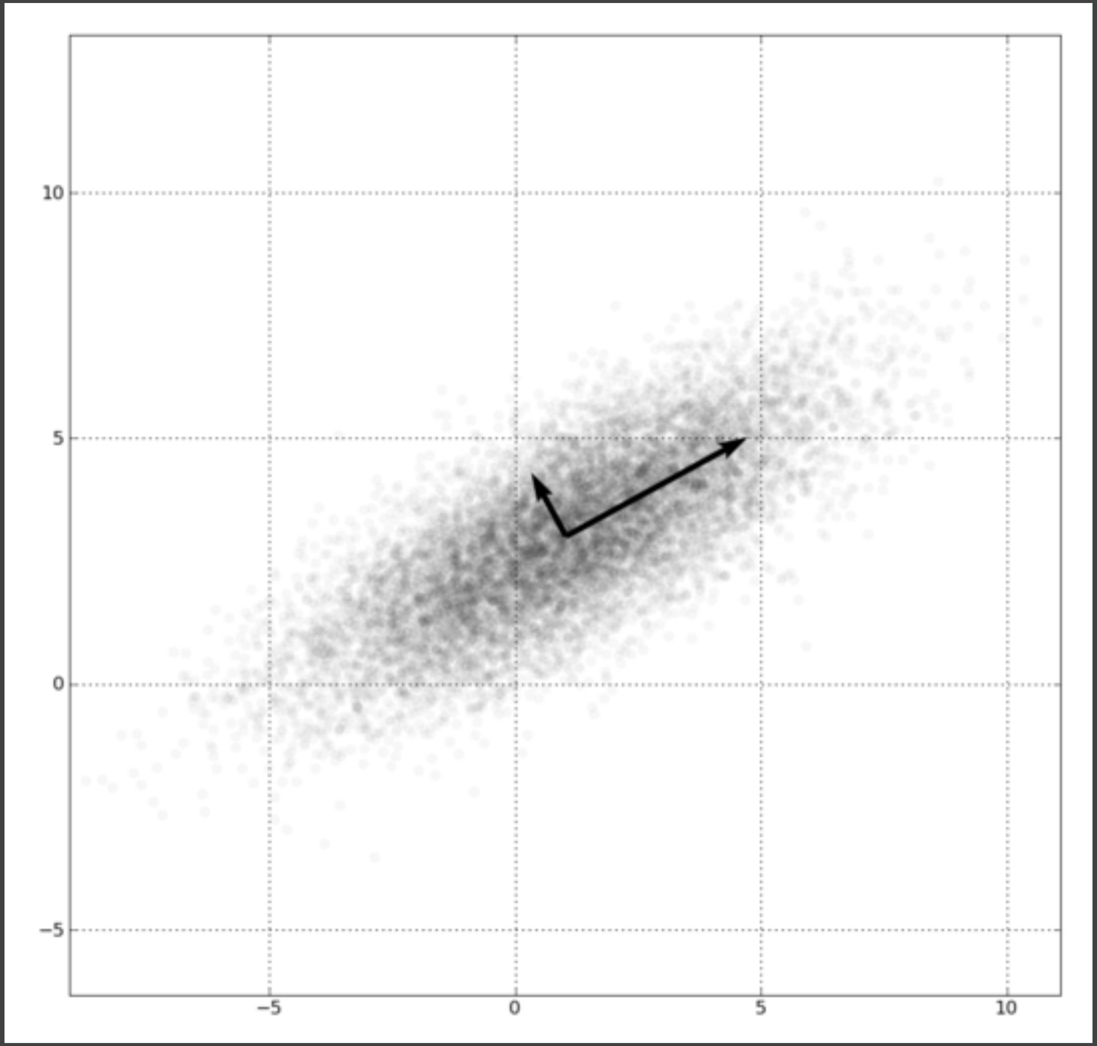
Different DR methods make different trade-offs: for example to **preserve global structure** (e.g., PCA) or **emphasize local structure** (e.g., nearest-neighbor approaches, including t-SNE and UMAP).

Principal Components Analysis



1. Mean-center the data.
2. Find ⊥ basis vectors that maximize the data variance.
3. Plot the data using the top vectors.

Principal Components Analysis

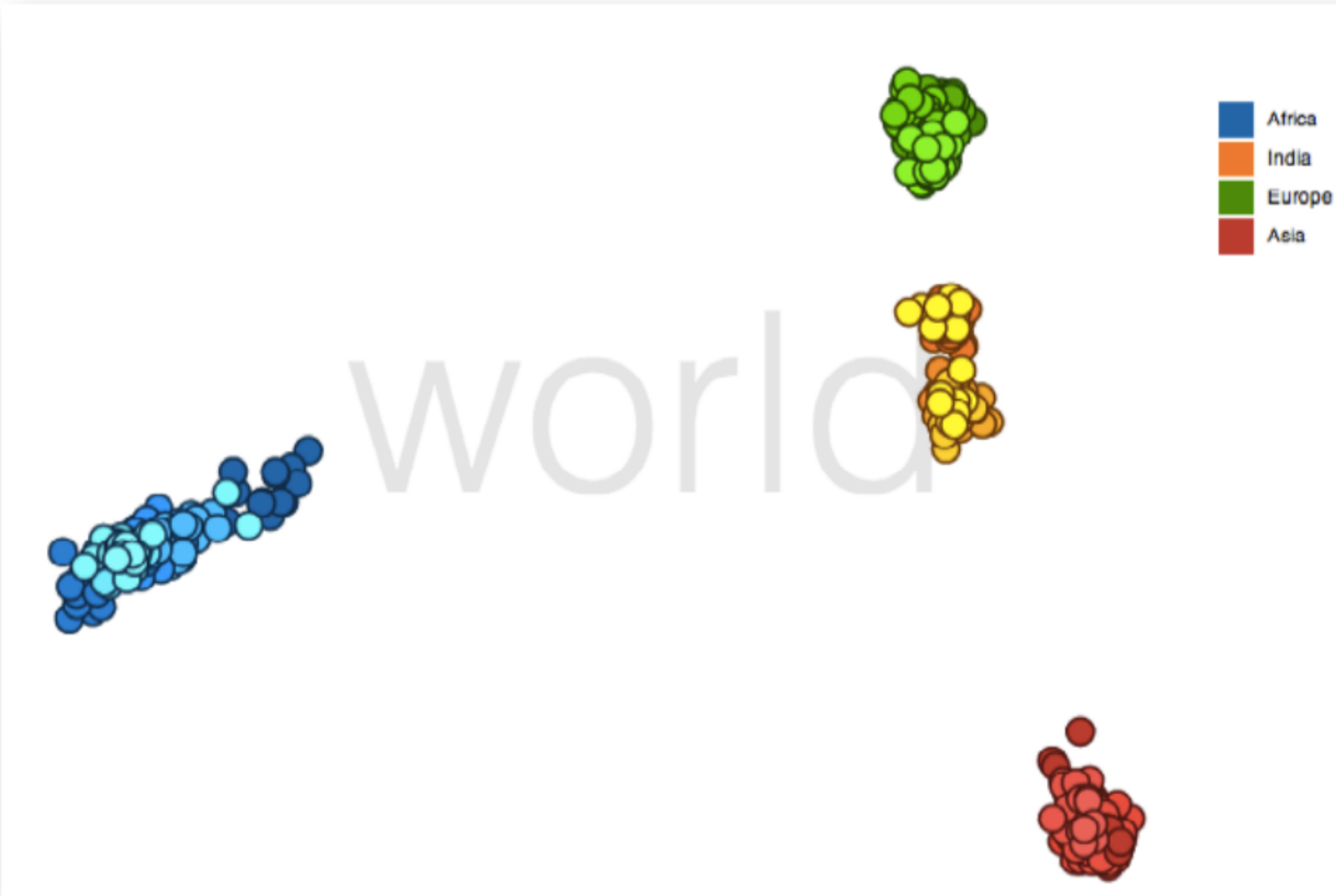


Linear transform: scale and rotate original space.

Lines (vectors) project to lines.

Preserves global distances.

PCA of Genomes [Demiralp et al. '13]



Non-Linear Techniques

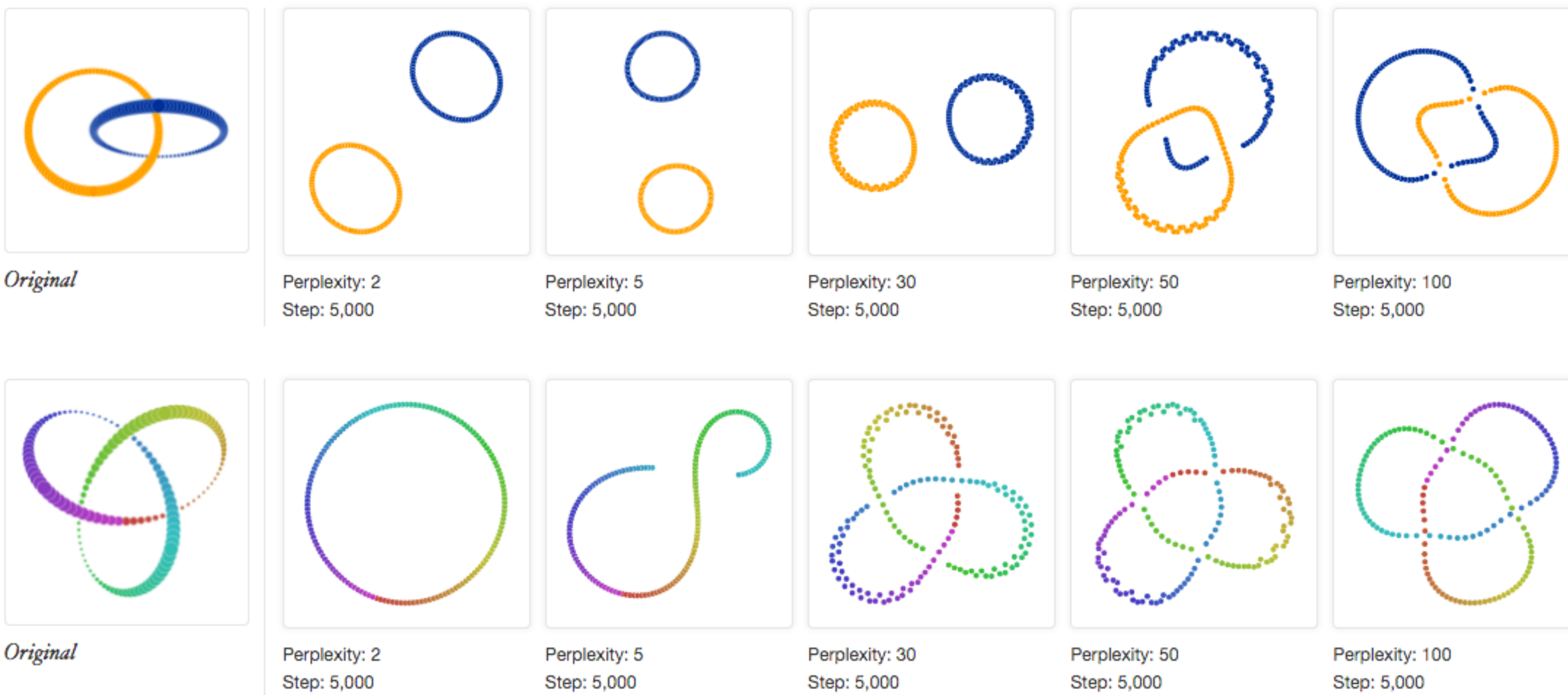
Distort the space, trade-off preservation of global structure to emphasize local neighborhoods. Use topological (nearest neighbor) analysis.

Two popular contemporary methods:

t-SNE - probabilistic interpretation of distance

UMAP - tries to balance local/global trade-off

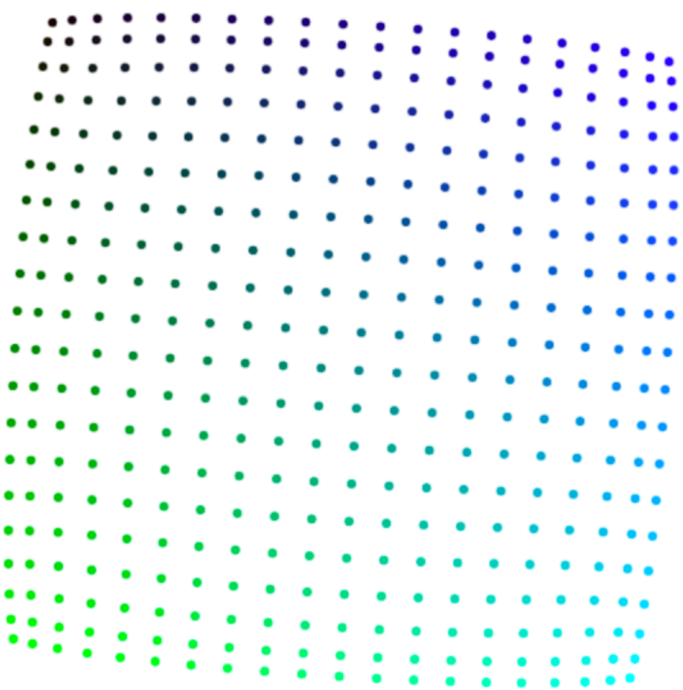
Visualizing t-SNE [Wattenberg et al. '16]



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

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

Points Per Side 20

Perplexity 10

Epsilon 5

MARTIN WATTENBERG
Google Brain

FERNANDA VIÉGAS
Google Brain

IAN JOHNSON
Google Cloud

Oct. 13
2016

Citation:
Wattenberg, et al., 2016

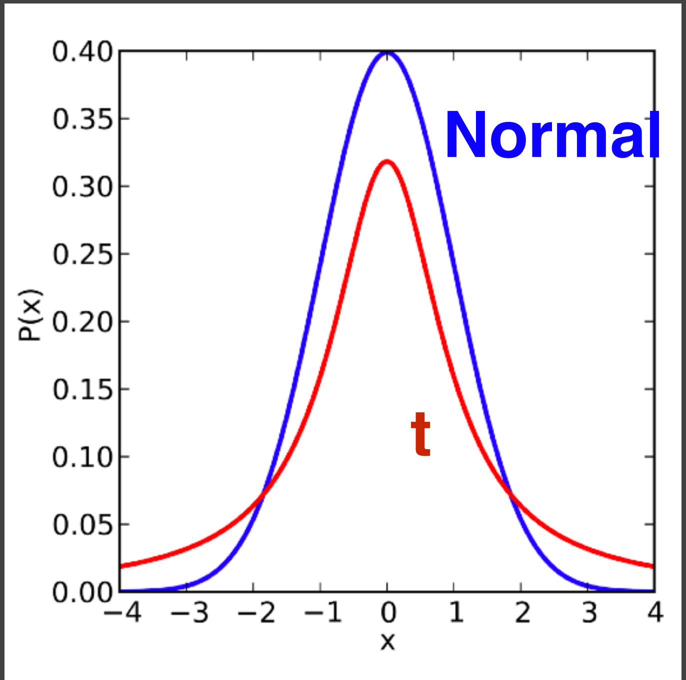
distill.pub

t-SNE [Maaten & Hinton 2008]

1. Model probability **P** of one point "choosing" another as its neighbor in the original space, using a Gaussian distribution defined using the distance between points. Nearer points have higher probability than distant ones.

t-SNE [Maaten & Hinton 2008]

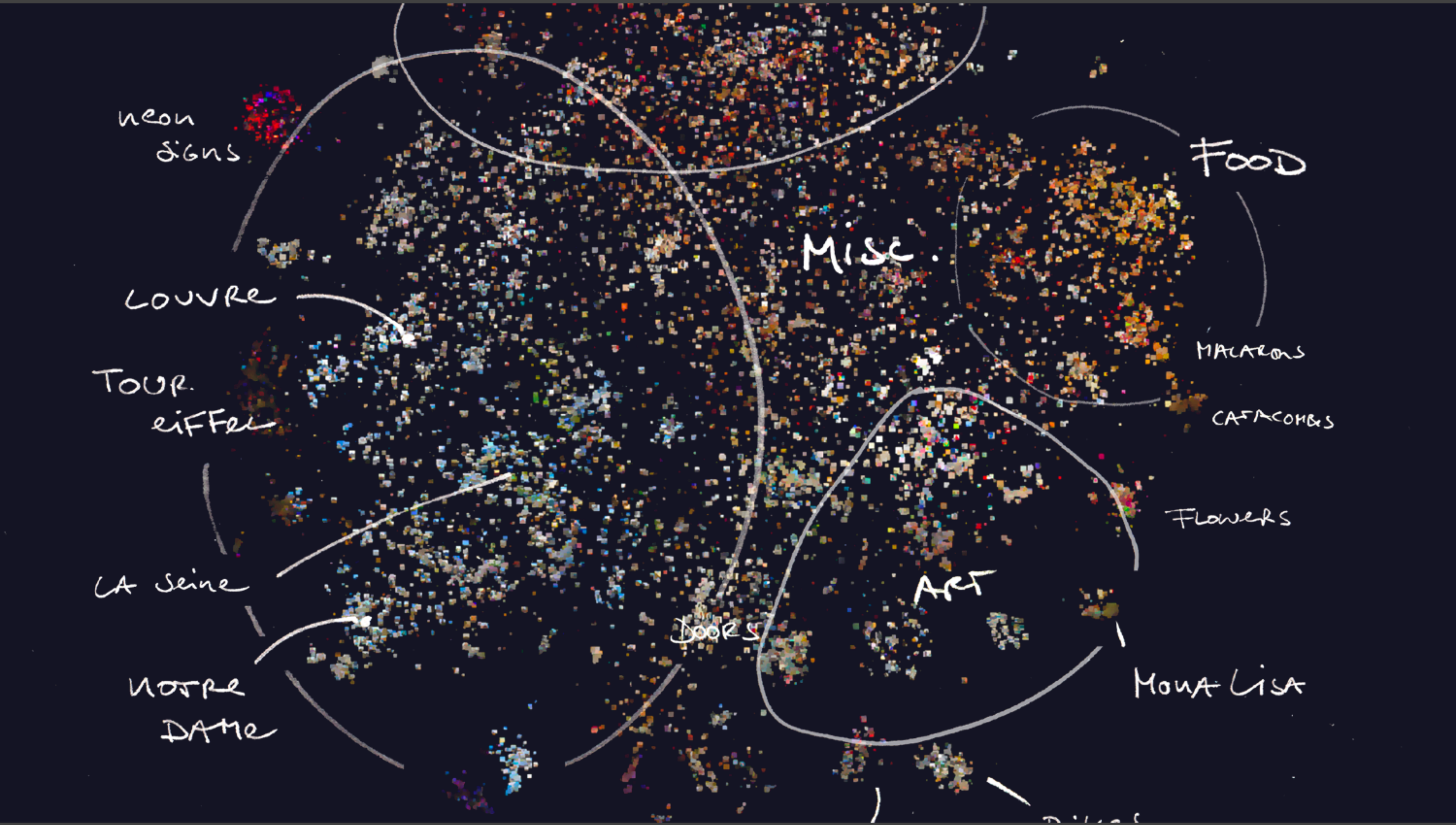
2. Define a similar probability **Q** in the low-dimensional (2D or 3D) embedding space, using a Student's *t* distribution *(hence the "t-" in "t-SNE"!)*. The *t*-distribution is heavy-tailed, allowing distant points to be even further apart.



t-SNE [Maaten & Hinton 2008]

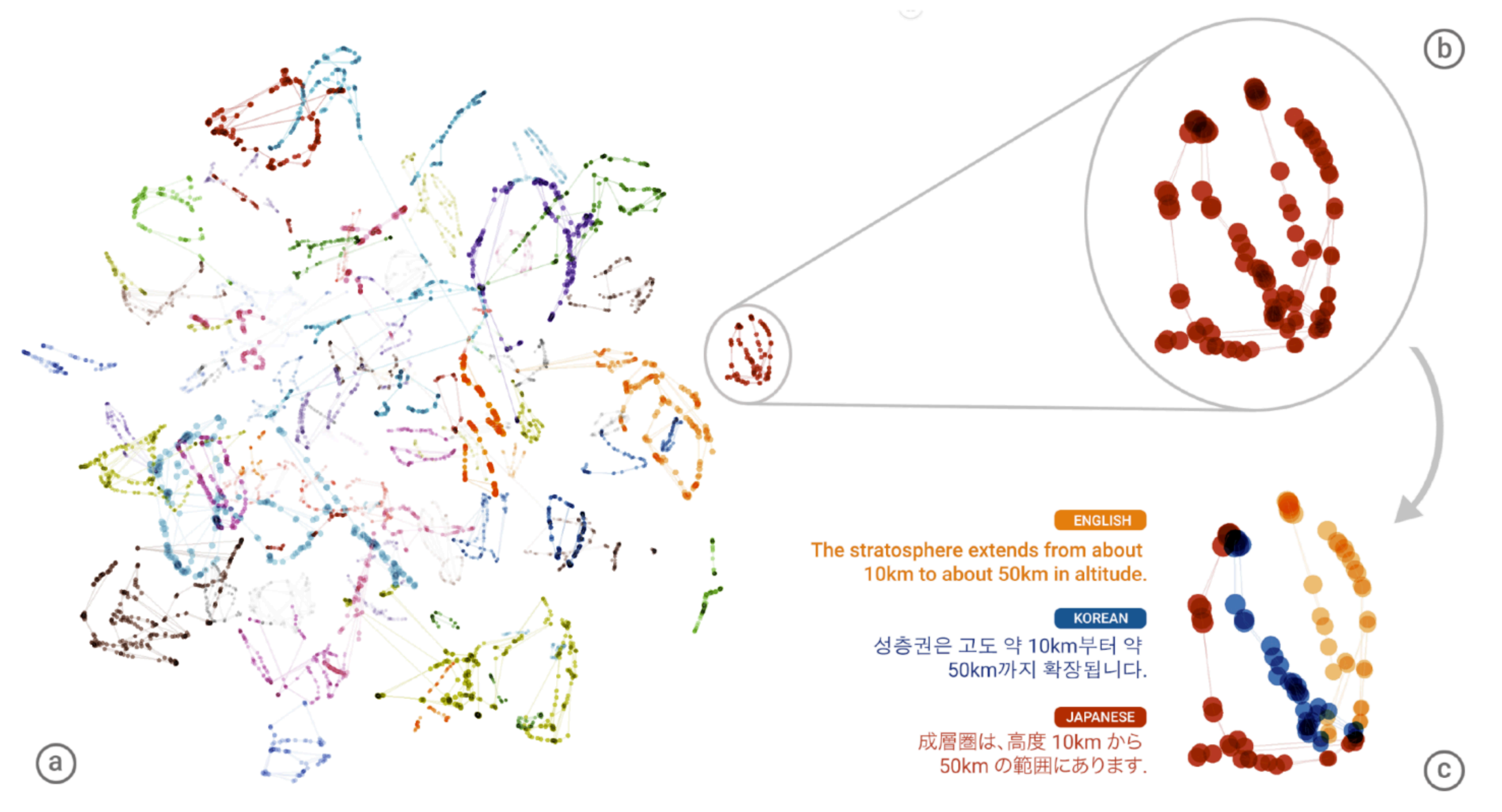
1. Model probability **P** of one point "choosing" another as its neighbor in the original space, using a Gaussian distribution defined using the distance between points. Nearer points have higher probability than distant ones.
2. Define a similar probability **Q** in the low-dimensional (2D or 3D) embedding space, using a Student's *t* distribution *(hence the "t-" in "t-SNE"!)*. The *t*-distribution is heavy-tailed, allowing distant points to be even further apart.
3. Optimize to find the positions in the embedding space that minimize the Kullback-Leibler divergence between the **P** and **Q** distributions: *KL(P || Q)*

Multiplicity [Stefaner 2018]



t-SNE projection of photos taken in Paris, France

MT Embedding [Johnson et al. 2018]



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

UMAP [McInnes et al. 2018]

Form weighted nearest neighbor graph, then layout the graph in a manner that balances embedding of local and global structure.

"Our algorithm is competitive with t-SNE for visualization quality and arguably preserves more of the global structure with superior run time performance." - McInnes et al. 2018

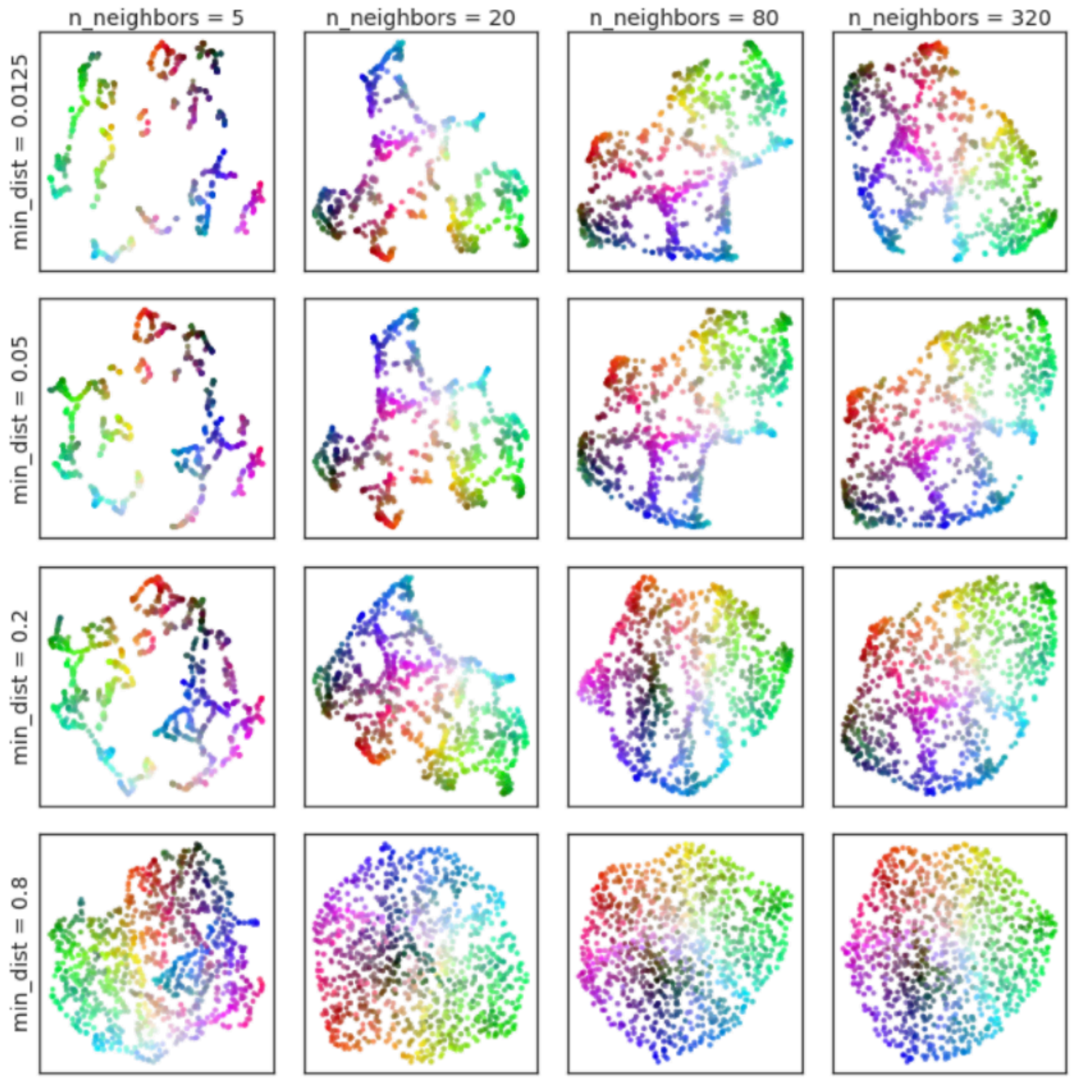
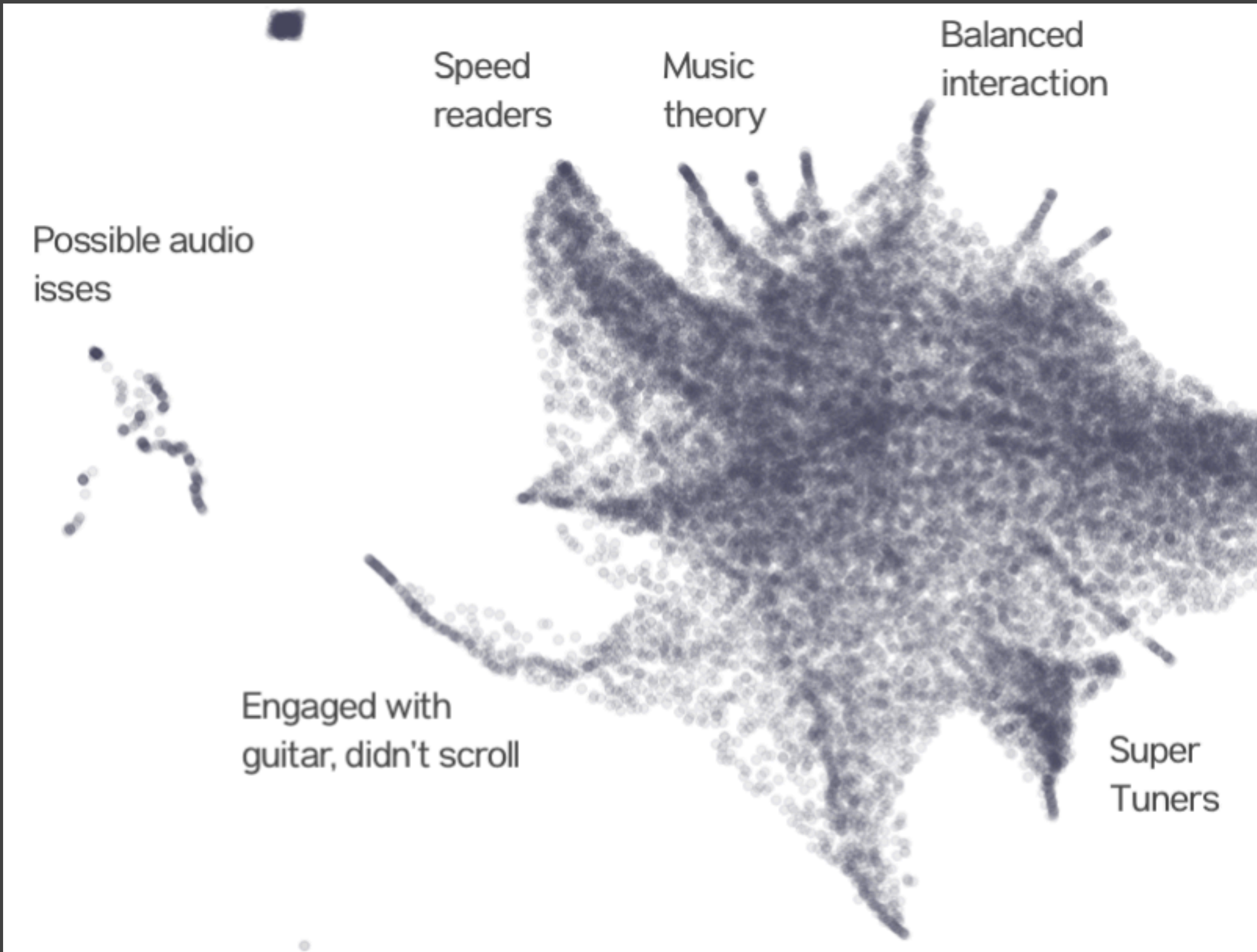


Figure 1: Variation of UMAP hyperparameters n and min-dist result in different embeddings. The data is uniform random samples from a 3-dimensional color-cube, allowing for easy visualization of the original 3-dimensional coordinates in the embedding space by using the corresponding RGB colour. Low values of n spuriously interpret structure from the random sampling noise – see Section 6 for further discussion of this phenomena.

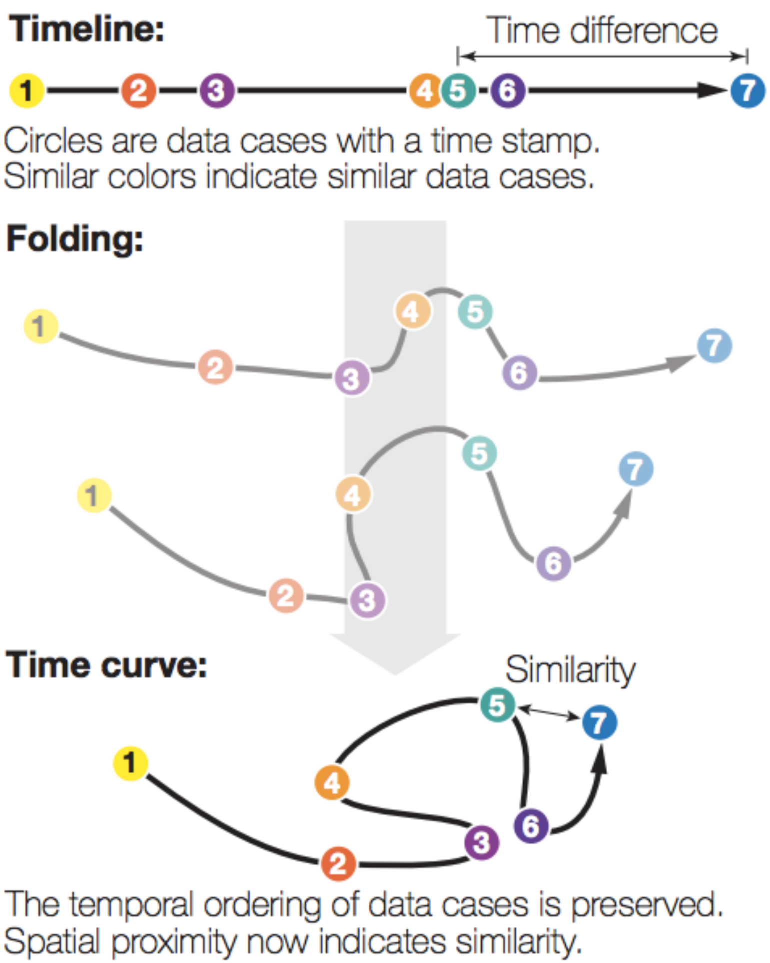
Reader Behavior [Conlen et al. 2019]



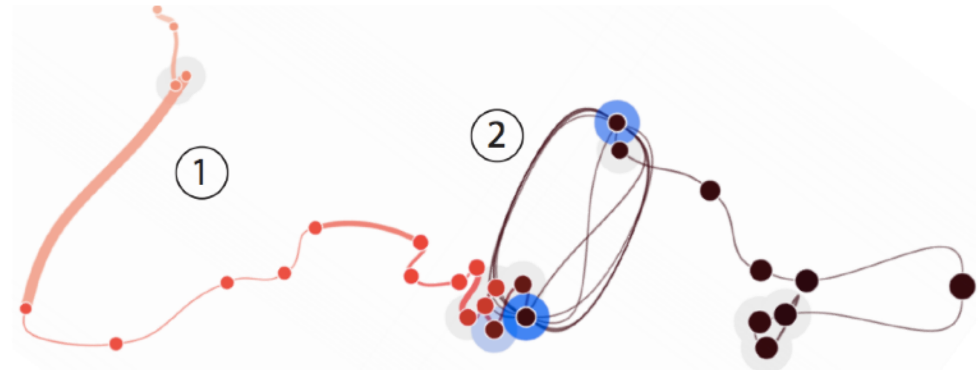
"Tentacles" map to activity archetypes, "blob" body maps to sessions that blend behaviors.

UMAP projection of reader activity for an interactive article.

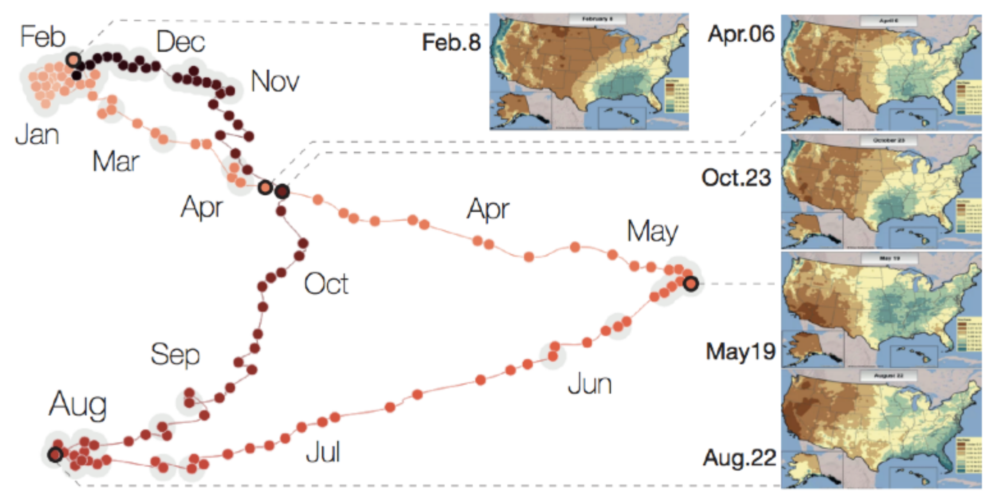
Time Curves [Bach et al. '16]



(a) Folding time



Wikipedia "Chocolate" Article



U.S. Precipitation over 1 Year

Visual Encoding Design

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!