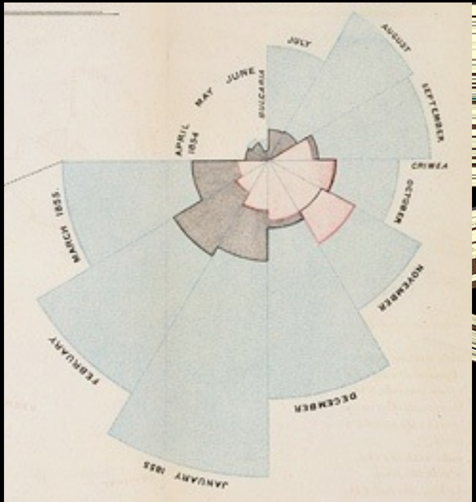
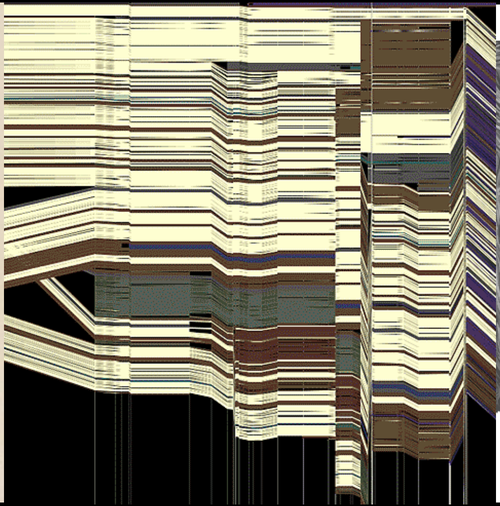
CSE 512 - Data Visualization

Visual Encoding





Leilani Battle University of Washington

The Big Picture

task
questions, goals
assumptions

data
physical data type
conceptual data type

domain
metadata
semantics
conventions

processing algorithms

→

mapping
visual encoding

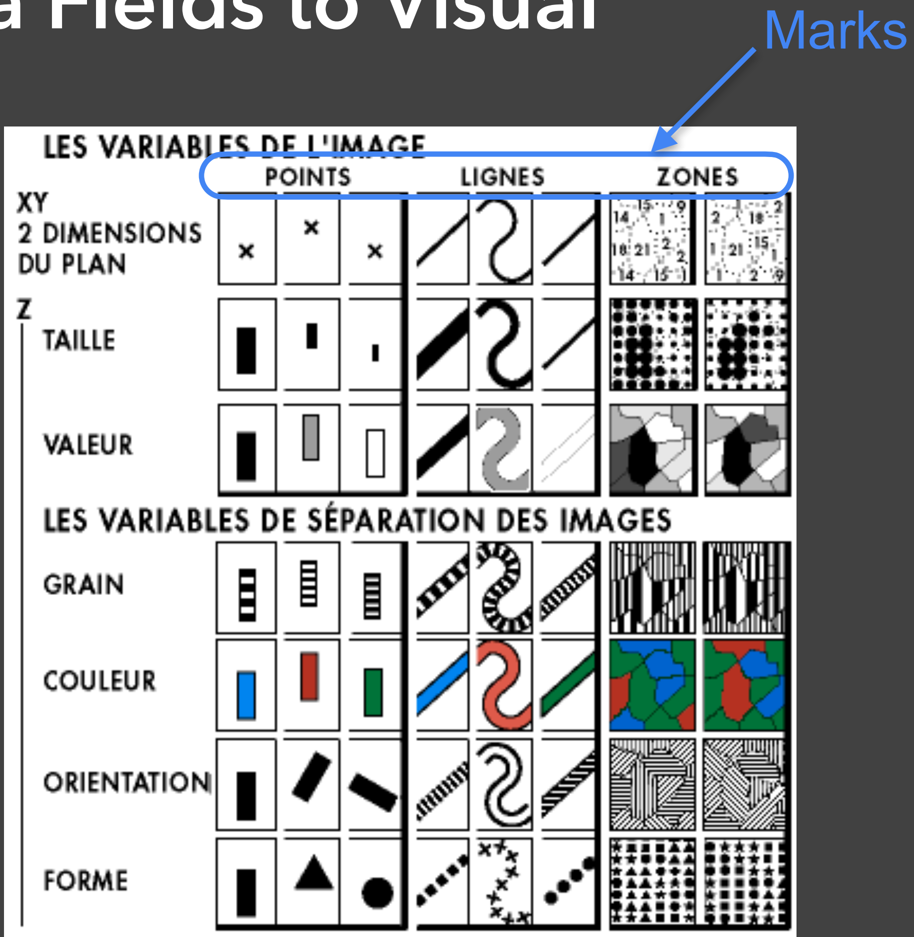
image
visual channel
graphical marks

Formalizing Design

Encodings Map Data Fields to Visual Properties of Marks

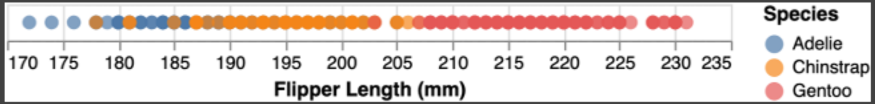
Encoding Channels:

Position (x 2)
Size
Value
Texture
Color
Orientation
Shape

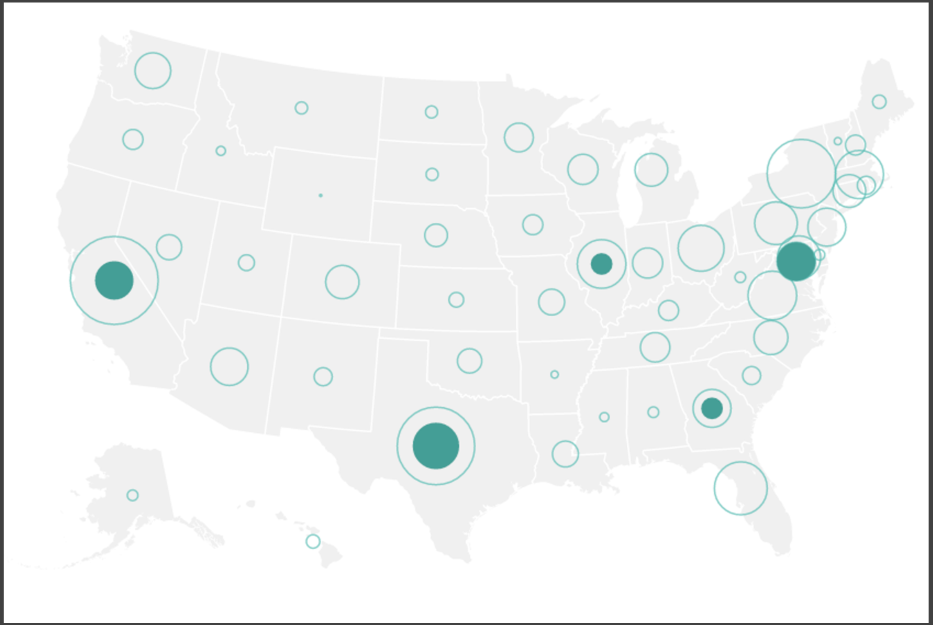


What Encoding Channels are Being Used?

Dot Plot



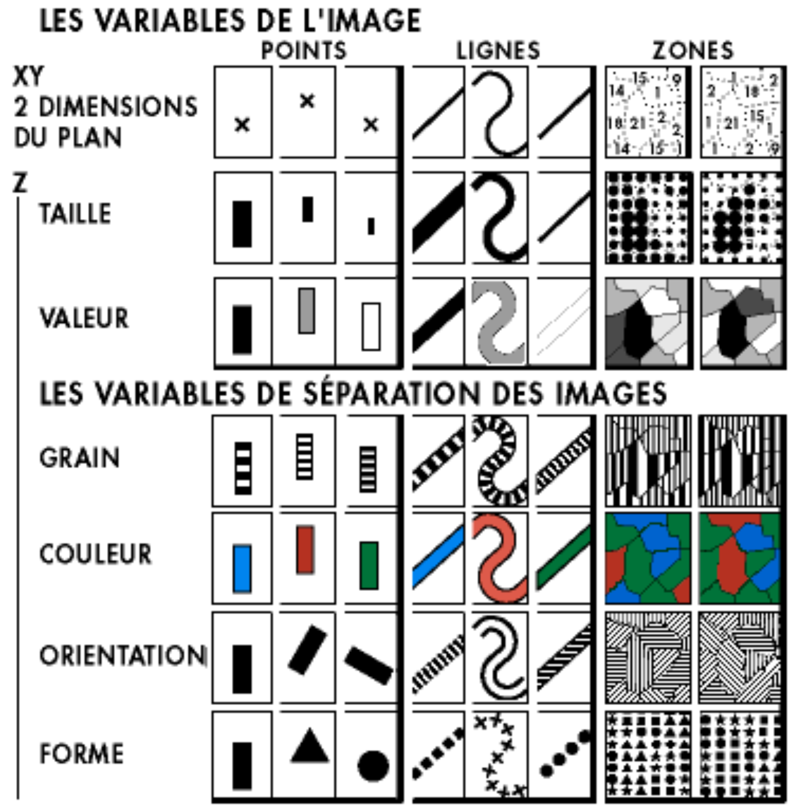
Choropleth Map



How do We Know Which Channels to Use?

Encoding Channels:

Position (x 2)
Size
Value
Texture
Color
Orientation
Shape



Choosing Visual Encodings

Assume *k* visual encodings and *n* data attributes.
We would like to pick the "best" encoding among a combinatorial set of possibilities of size $(n+1)^k$

Principle of Consistency
The properties of the image (visual variables) should match the properties of the data.

Principle of Importance Ordering
Encode the most important information in the most effective way.

Some Useful Data Types

N - Nominal (labels or categories)

- Fruits: apples, oranges, …

O - Ordered

- Letter Grades: A+, A , A-, B+, …

Q - Quantitative (intervals and ratios)

- Dates: Jan, 19, 2006
- Location: (LAT 33.98, LONG -118.45)
- Physical measurements: Length, Mass, Time duration, …

Bertin’s Levels of Organization

Position	N	O	Q
Size	N	O	Q
Value	N	O	Q
Texture	N	O	
Color	N		
Orientation	N		
Shape	N		

Nominal

Ordinal

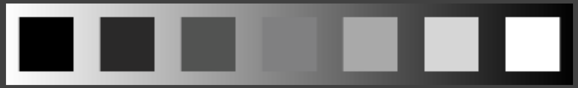
Quantitative

Note: **Q** ⊂ **O** ⊂ **N**

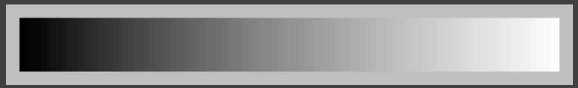
Information in Hue and Lightness

Lightness (“value”) is perceived as ordered

∴ Encode ordinal variables (O)



∴ Encode continuous variables (Q) [not as well]



Hue is normally perceived as unordered

∴ Encode nominal variables (N) using color

Design Criteria [Mackinlay 86]

Expressiveness

A set of facts is *expressible* in a visual language if the sentences (i.e. the visualizations) in the language express all the facts in the set of data, and only the facts in the data.

Effectiveness

A visualization is more *effective* than another visualization if the information conveyed by one visualization is more readily perceived than the information in the other visualization.

Design Criteria [Mackinlay 86]

Expressiveness

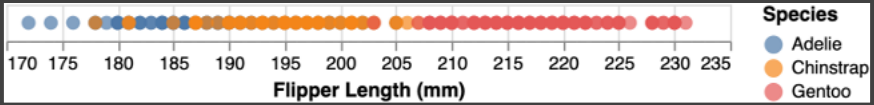
A set of facts is *expressible* in a visual language if the sentences (i.e. the visualizations) in the language express all the facts in the set of data, and only the facts in the data.

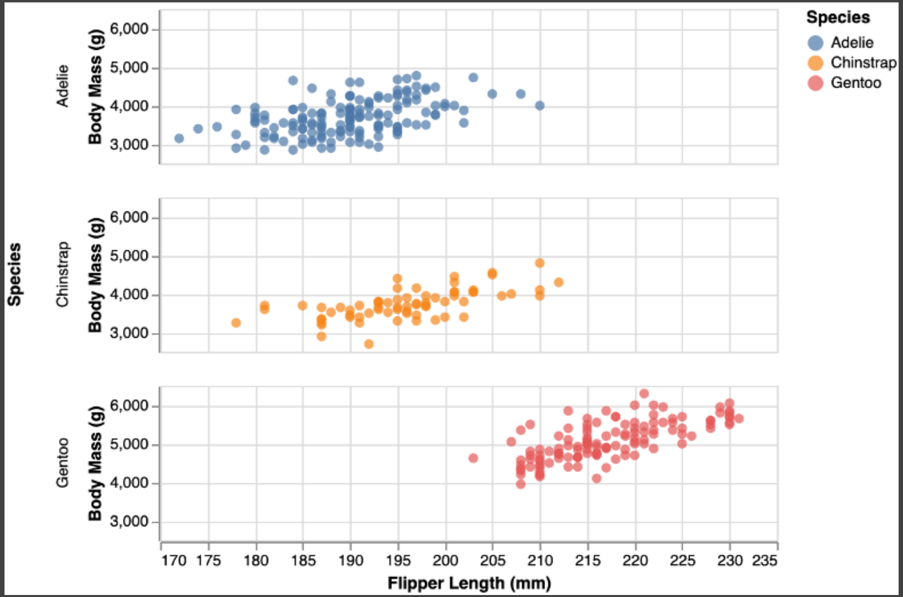
Effectiveness

A visualization is more *effective* than another visualization if the information conveyed by one visualization is more readily perceived than the information in the other visualization.

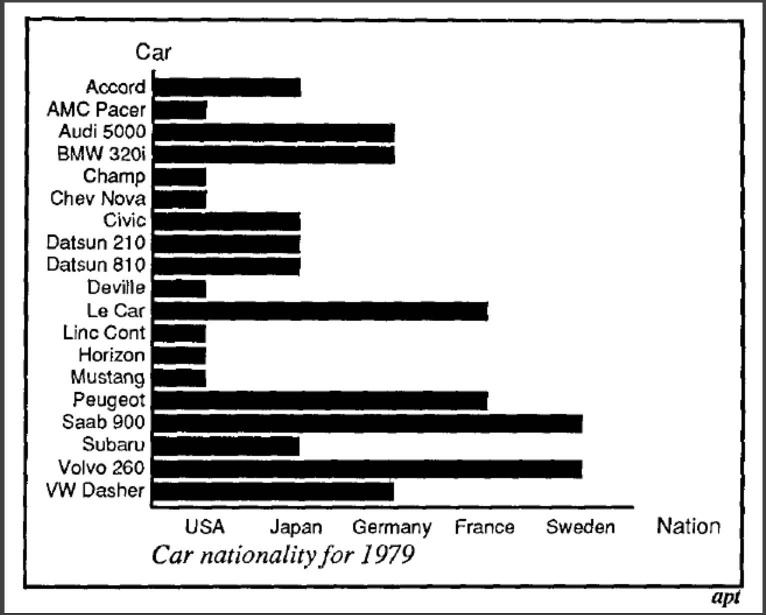
Can not express the facts

A multivariate relation may be *inexpressive* in a single horizontal dot plot because multiple records are mapped to the same position.





Expresses facts not in the data



A length is interpreted as a quantitative value.

Design Criteria [Mackinlay 86]

Expressiveness

A set of facts is *expressible* in a visual language if the sentences (i.e. the visualizations) in the language express all the facts in the set of data, and only the facts in the data.

Effectiveness

A visualization is more *effective* than another visualization if the information conveyed by one visualization is more readily perceived than the information in the other visualization.

Design Criteria [Mackinlay 86]

Expressiveness

A set of facts is *expressible* in a visual language if the sentences (i.e. the visualizations) in the language express all the facts in the set of data, and only the facts in the data.

Effectiveness

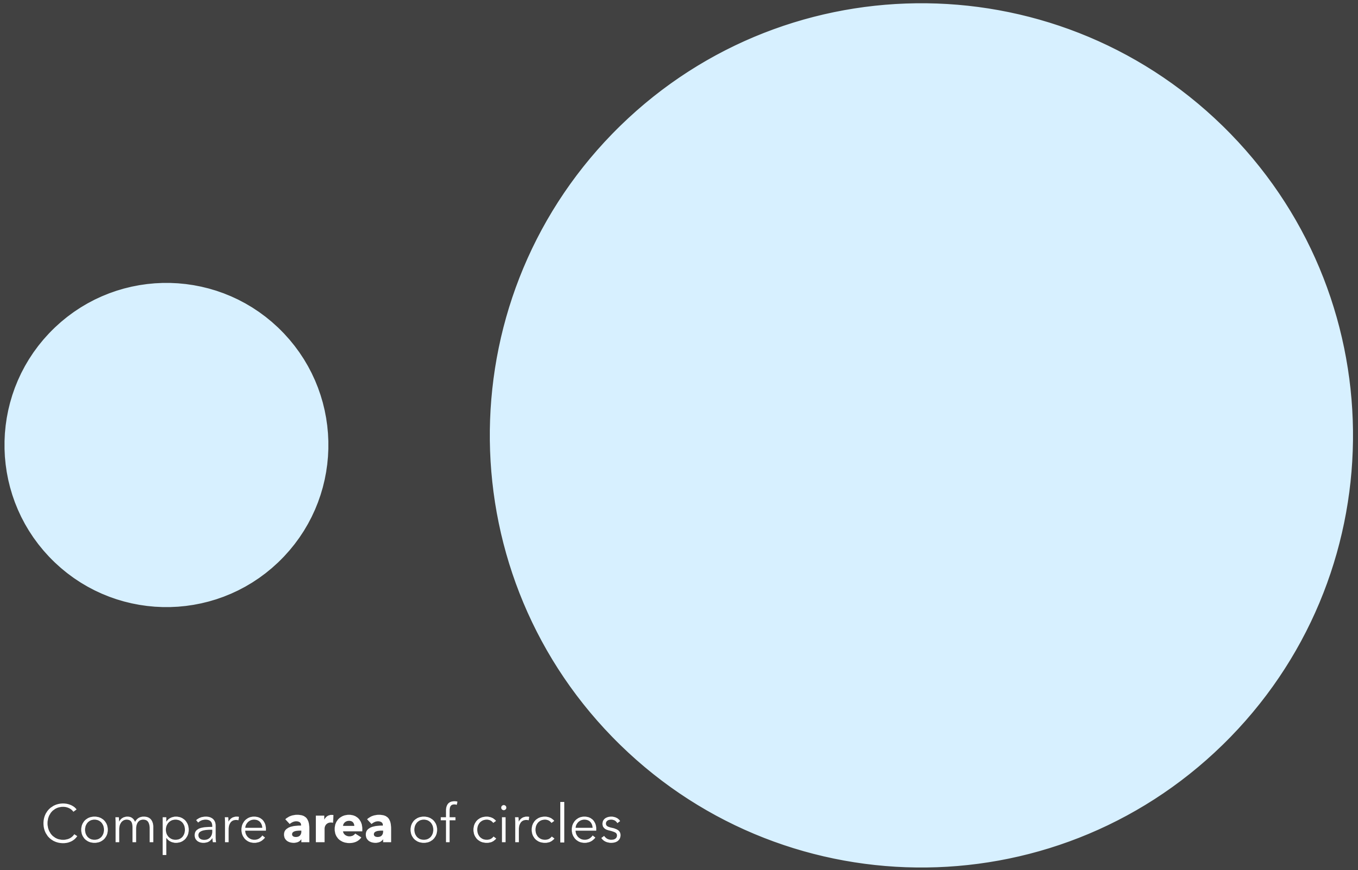
A visualization is more *effective* than another visualization if the information conveyed by one visualization is more readily perceived than the information in the other visualization.

Design Criteria *Translated*

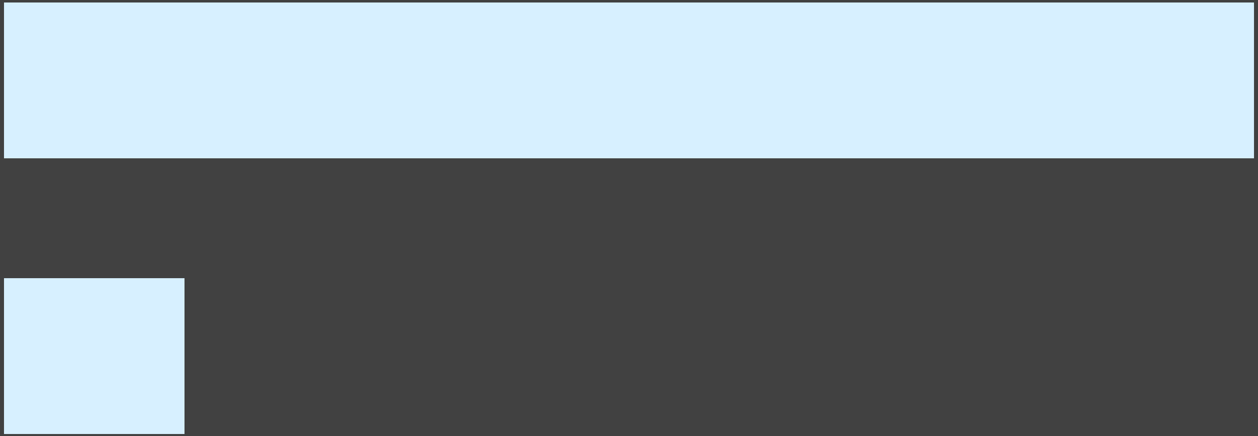
Tell the truth and nothing but the truth
(don't lie, and don't lie by omission)

Use encodings that people decode better
(where better = faster and/or more accurate)

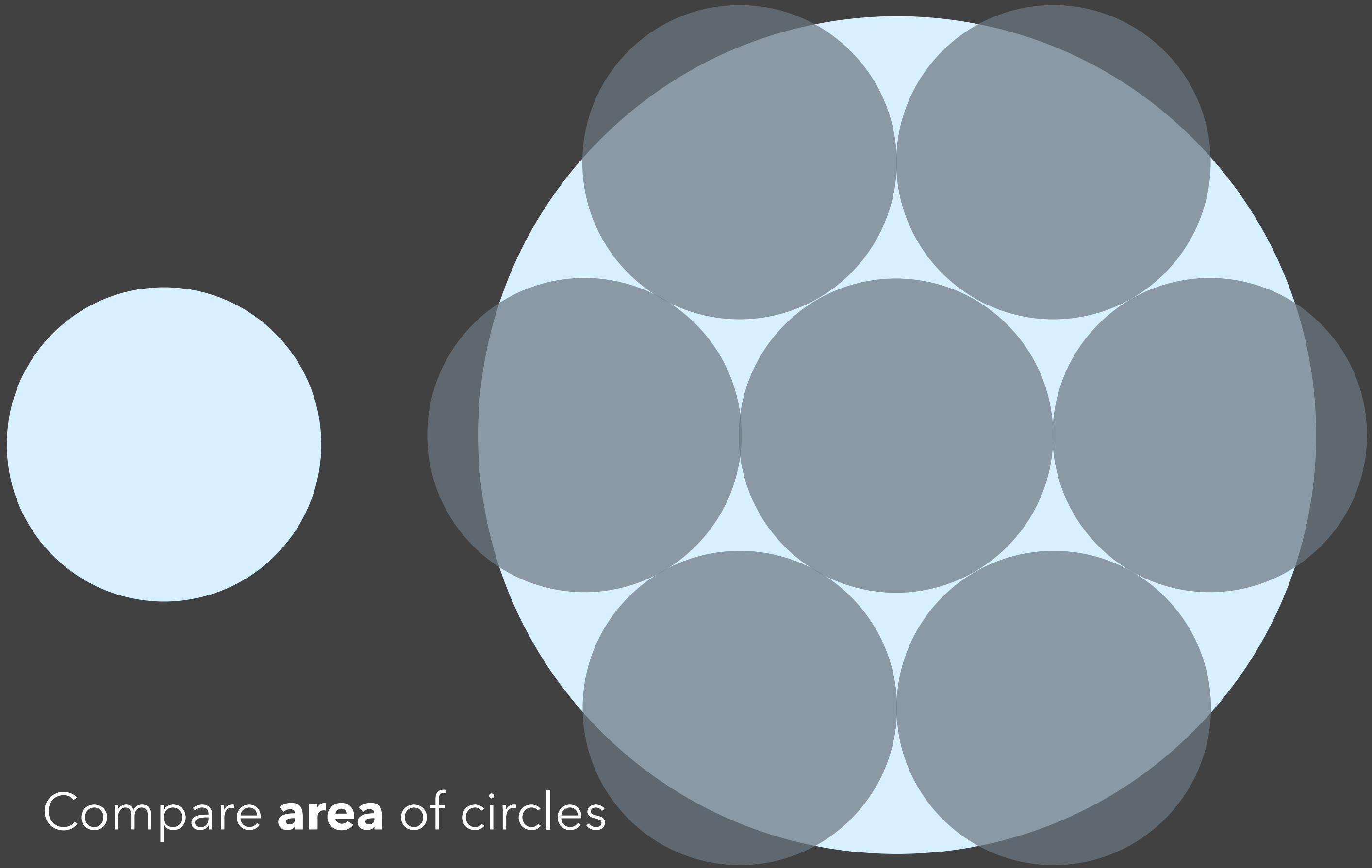
A Quick Experiment…



Compare **area** of circles



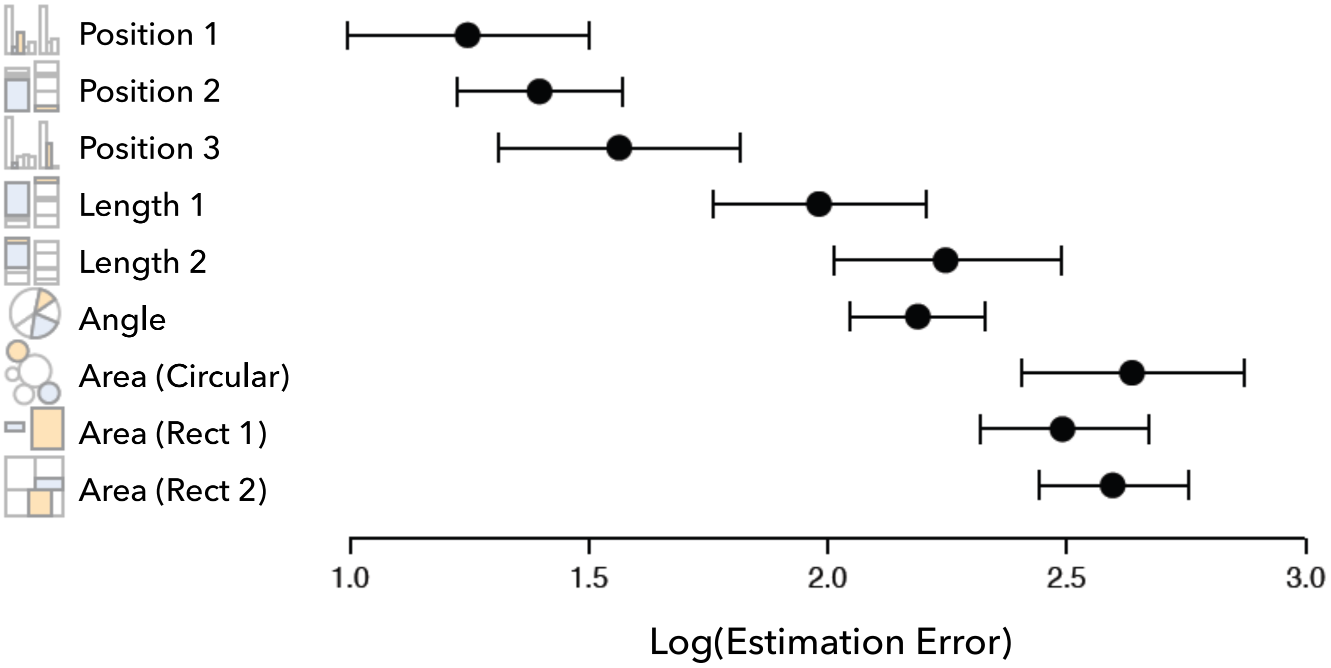
Compare **length** of bars



Compare **area** of circles

Compare **length** of bars

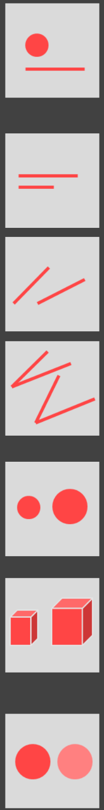
Accuracy of Visual Decoding



Ranking Visual Encodings

Comparing
Proportions

Most accurate



- Position (common) scale
- Position (non-aligned) scale
- Length
- Slope
- Angle
- Area
- Volume
- Color hue-saturation-density

Least accurate

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

Conjectured *effectiveness* of encodings by data type

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

Conjectured *effectiveness* of encodings by data type

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

Conjectured *effectiveness* of encodings by data type

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

Conjectured *effectiveness* of encodings by data type

Mackinlay's Design Algorithm

APT - "A Presentation Tool", 1986

User formally specifies data model and type
Input: ordered list of data variables to show

APT searches over design space
Test expressiveness of each visual encoding
Generate encodings that pass test
Rank by perceptual effectiveness criteria

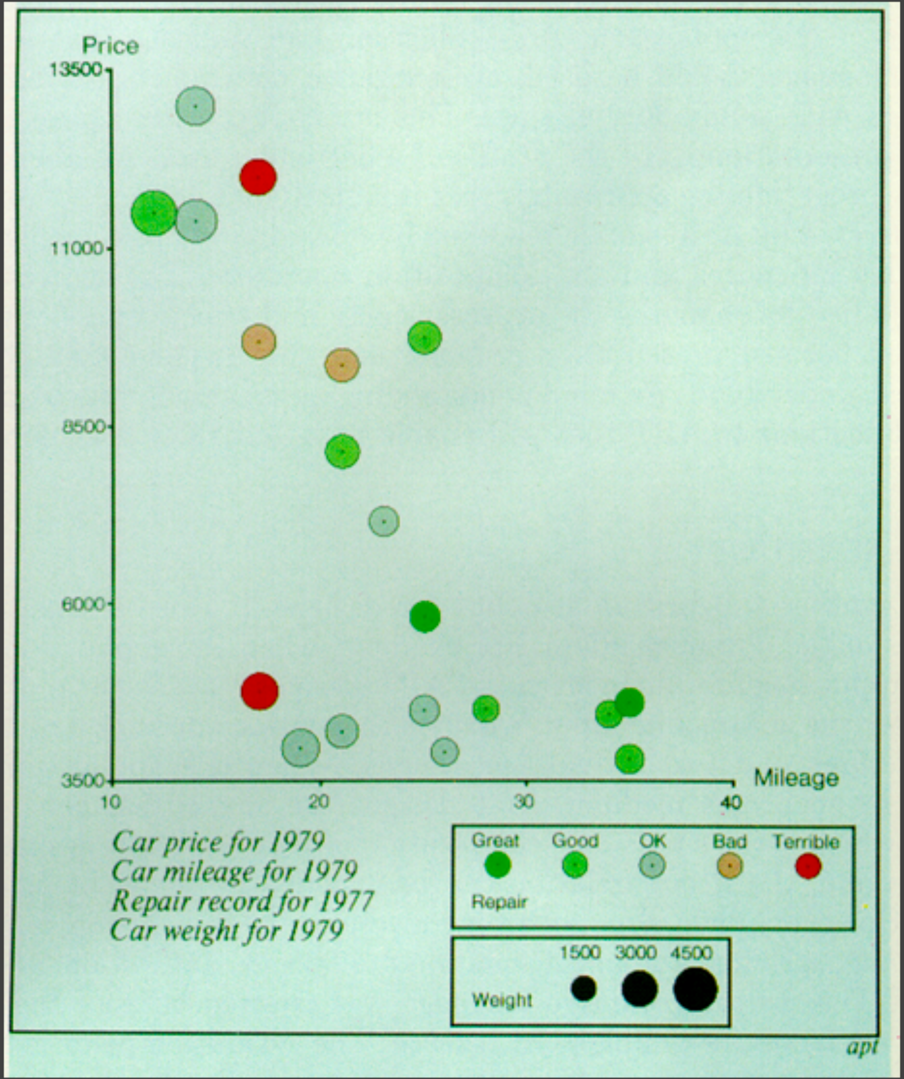
Output the "most effective" visualization

APT

Automatically generate chart for car data

Input variables:

1. Price
2. Mileage
3. Repair
4. Weight



Limitations of APT

Does not cover many visualization techniques
Networks, hierarchies, maps, diagrams
Also: 3D structure, animation, illustration, …

Does not consider interaction

Does not consider semantics / conventions

Assumes single visualization as output

Still an active area of research, e.g., the
Draco visualization design knowledge base

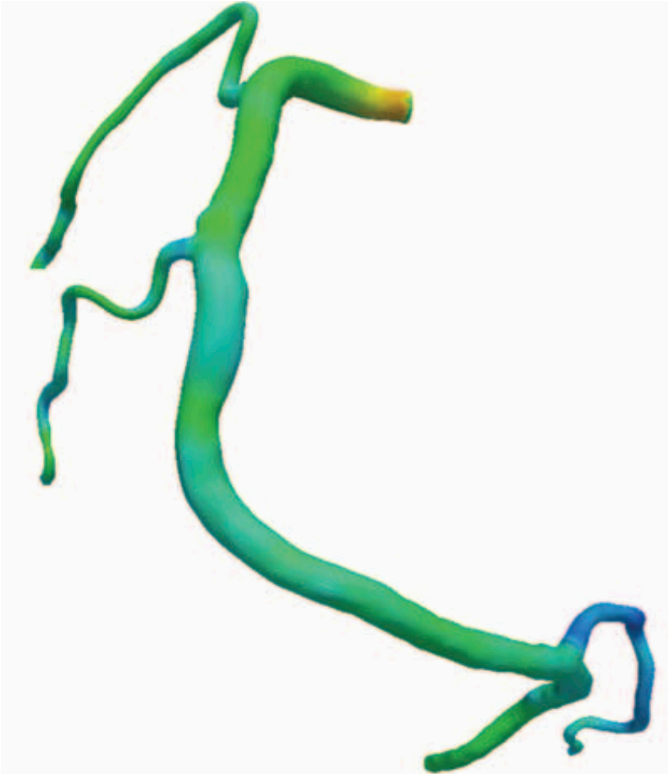
Design Examples

Artery Visualization [Borkin et al. 2011]

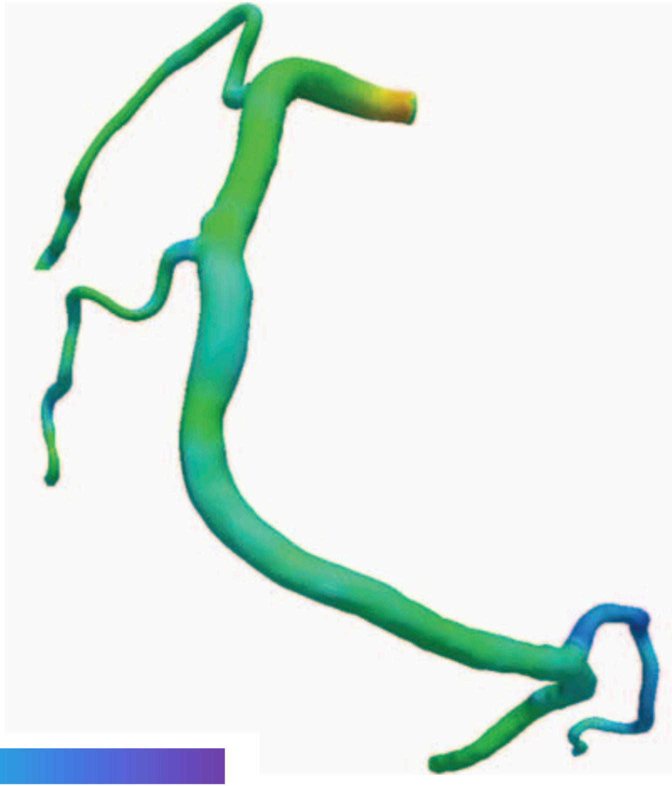
What encoding channels are used?

Which choices are effective?

What could be improved?



Artery Visualization [Borkin et al. 2011]



What encoding channels are used?

- X position
- Y position
- Z position
- Color hue (rainbow color map)

Which choices are effective?

- X position, y position

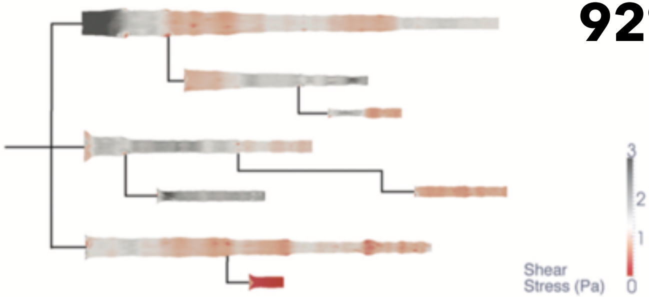
What could be improved?

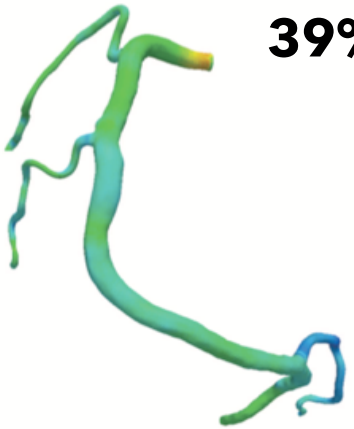
- Z position, rainbow color map

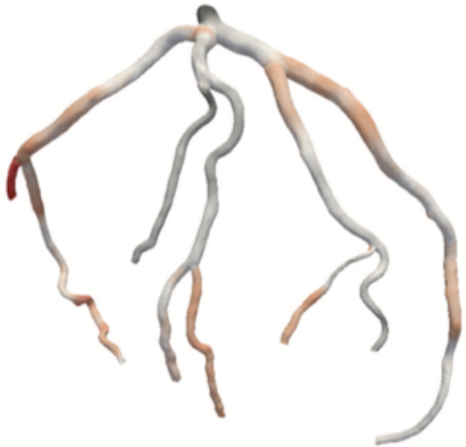
https://observablehq.com/@d3/color-schemes

Artery Visualization [Borkin et al. 2011]

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





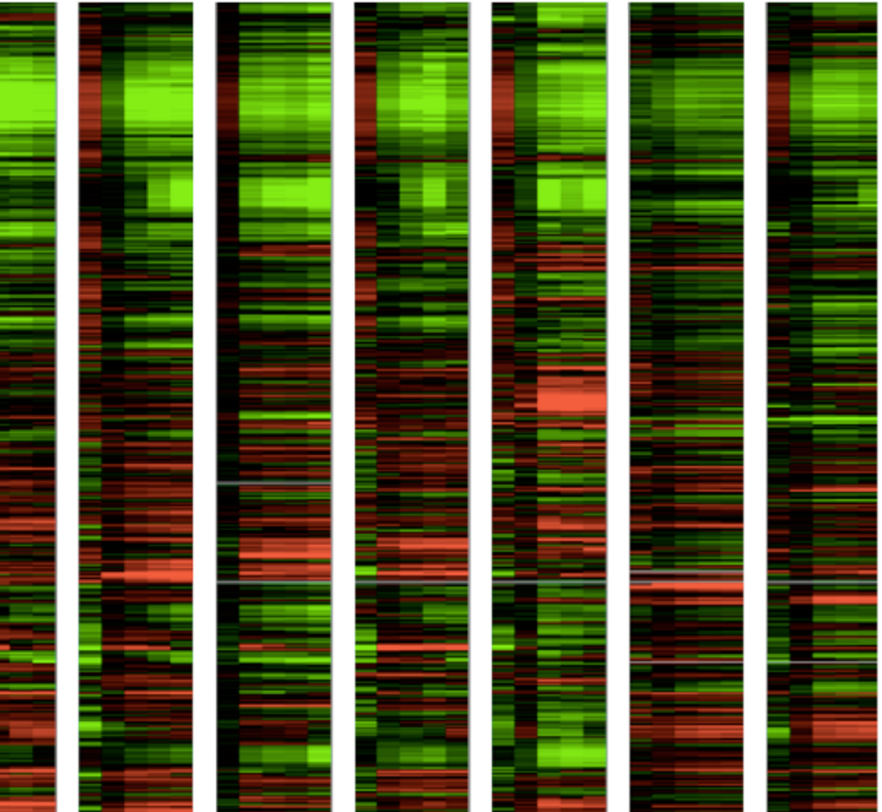


Effectiveness Rankings

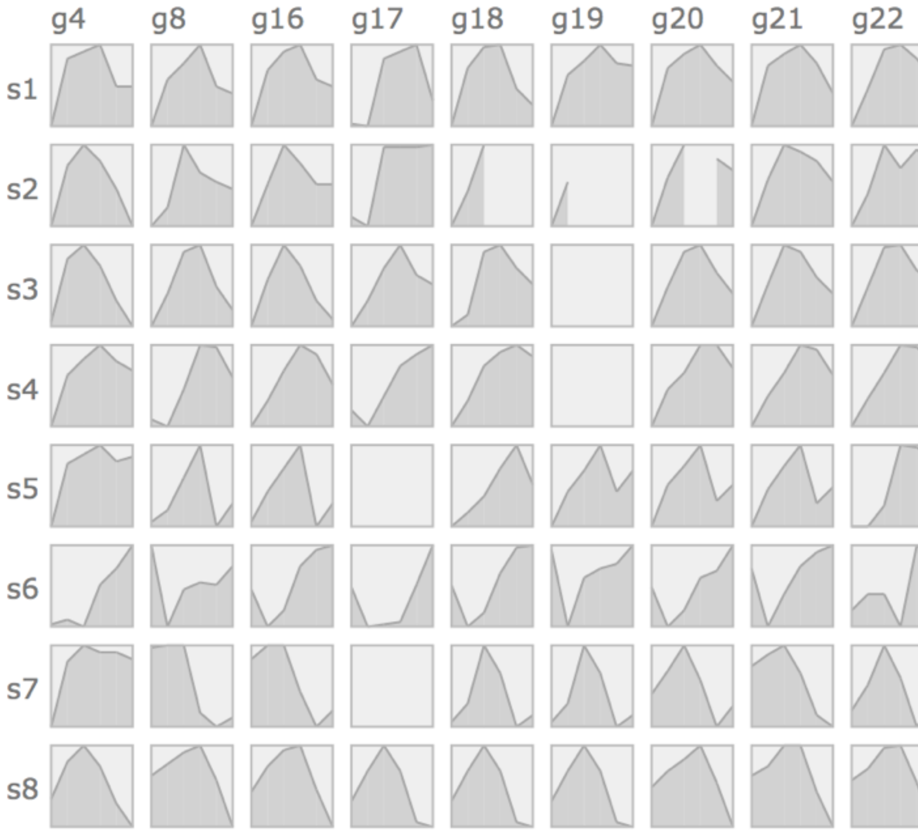
QUANTITATIVE	ORDINAL	NOMINAL
Position	Position	Position
Length	Density (Value)	Color Hue
Angle	Color Sat	Texture
Slope	Color Hue	Connection
Area (Size)	Texture	Containment
Volume	**Connection**	Density (Value)
Density (Value)	Containment	Color Sat
Color Sat	Length	Shape
~~**Color Hue**~~	Angle	Length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

Gene Expression Time-Series [Meyer et al. 2011]

Color Encoding



Position 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

A Design Space
of Visual Encodings

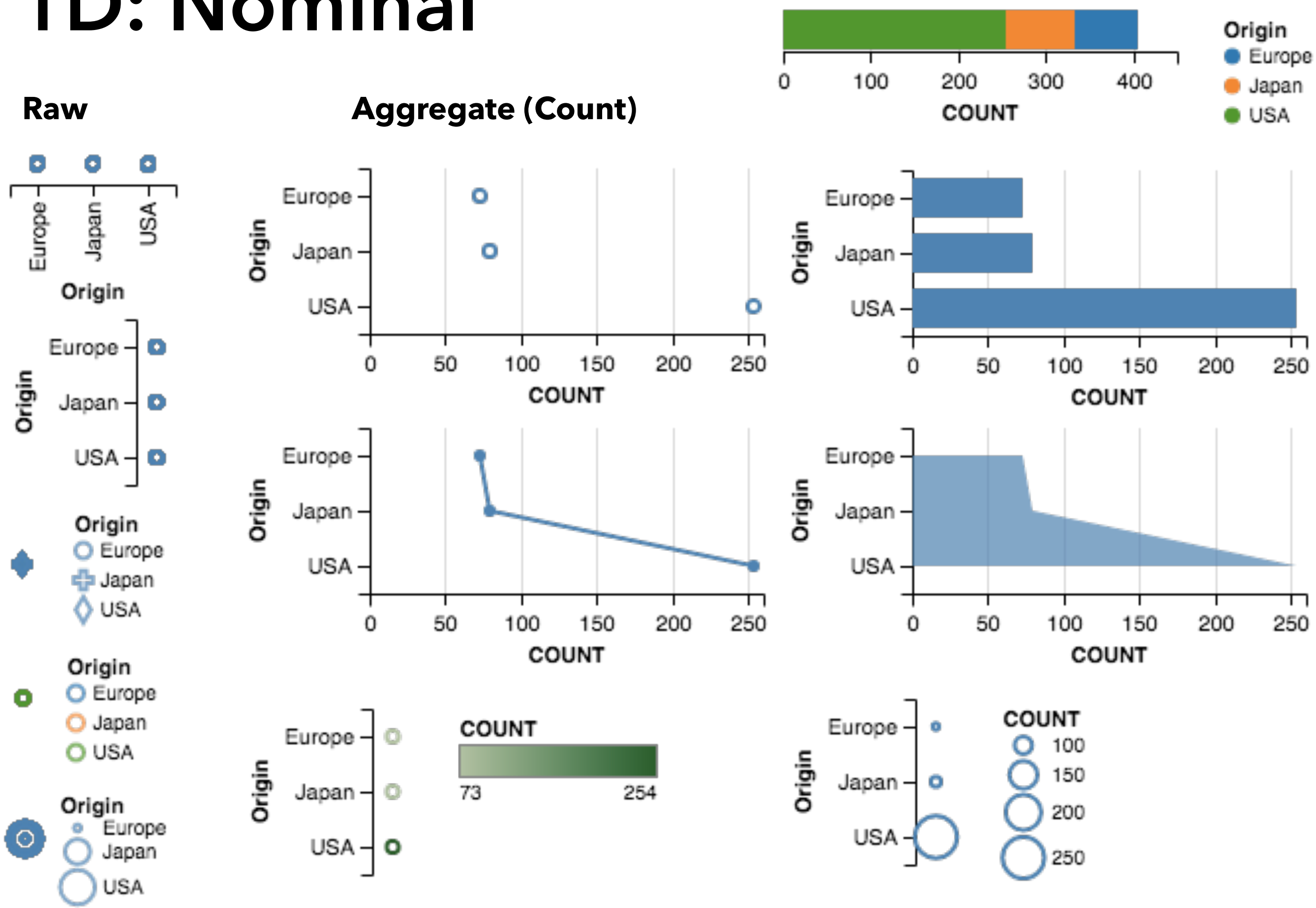
Mapping Data to Visual Variables

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, …*).

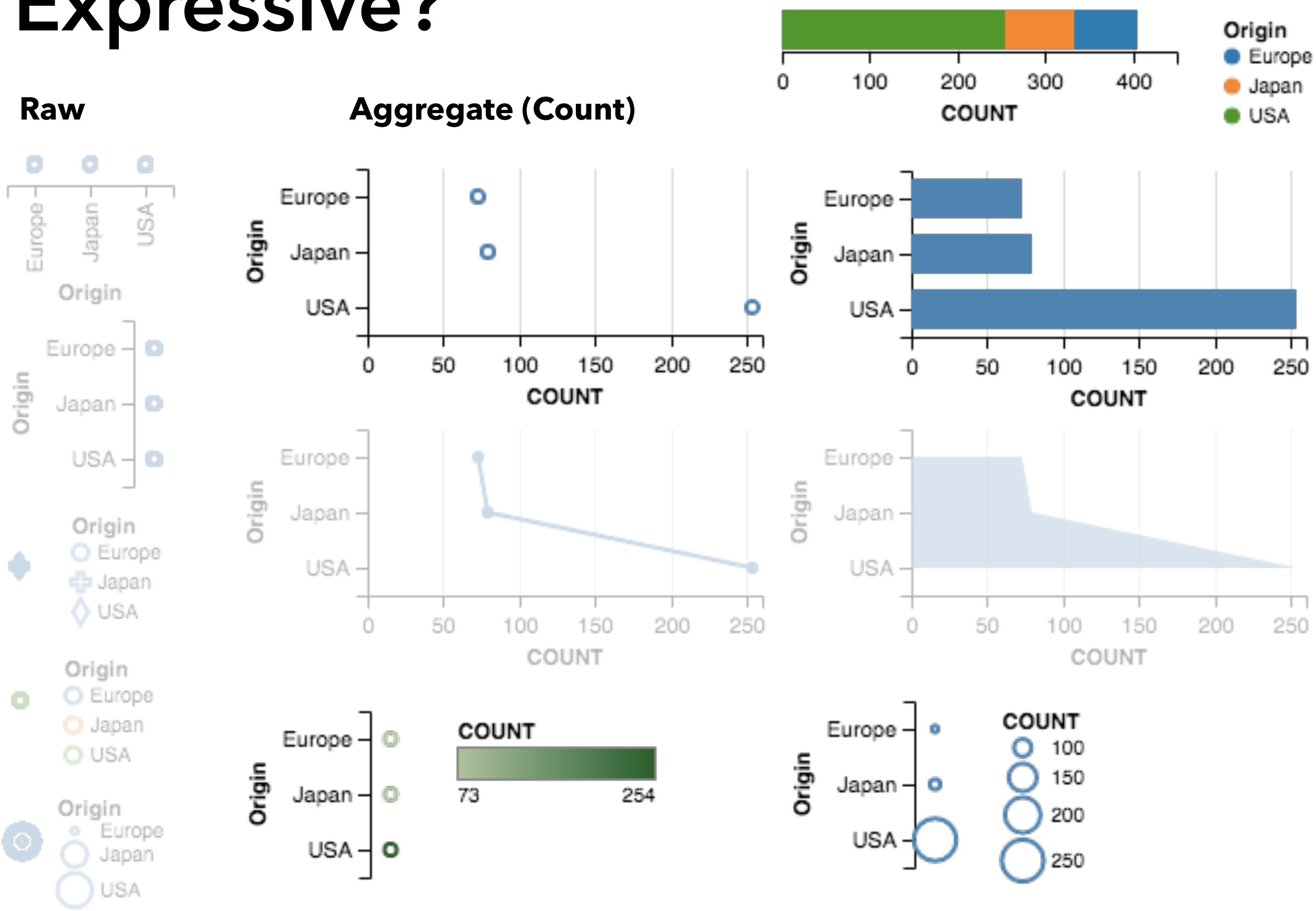
These options define a large combinatorial space, containing both useful and questionable charts!

1D: Nominal



Raw
Aggregate (Count)
Origin
Europe
Japan
USA
COUNT
0
50
100
150
200
250
300
400
73
254

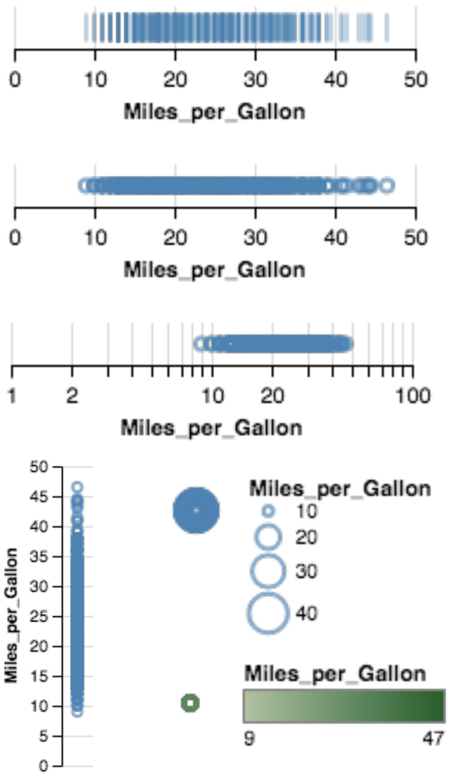
Expressive?



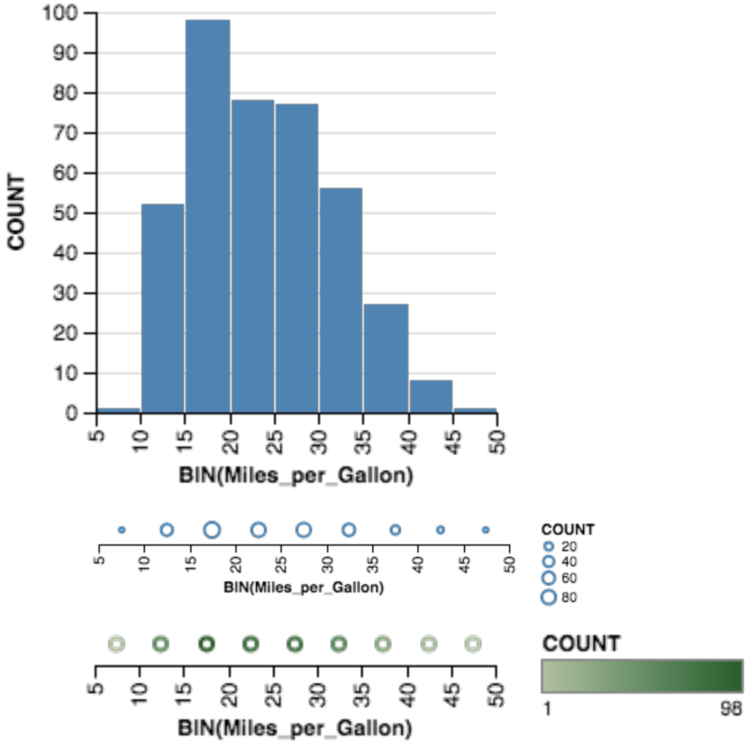
Raw
Aggregate (Count)
0
100
200
300
400
COUNT
Origin
Europe
Japan
USA
Europe
Japan
USA
Origin
0
50
100
150
200
250
COUNT
COUNT
73
254
COUNT
100
150
200
250

1D: Quantitative

Raw

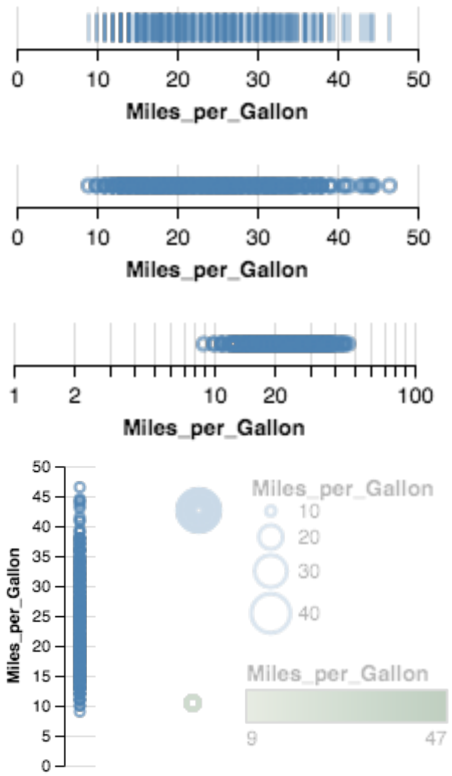


Aggregate (Count)

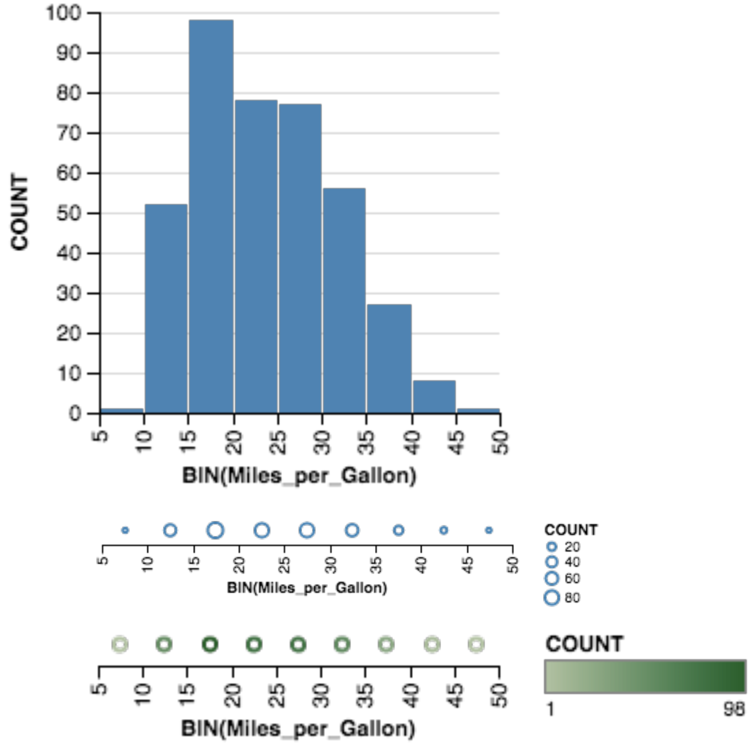


Expressive?

Raw

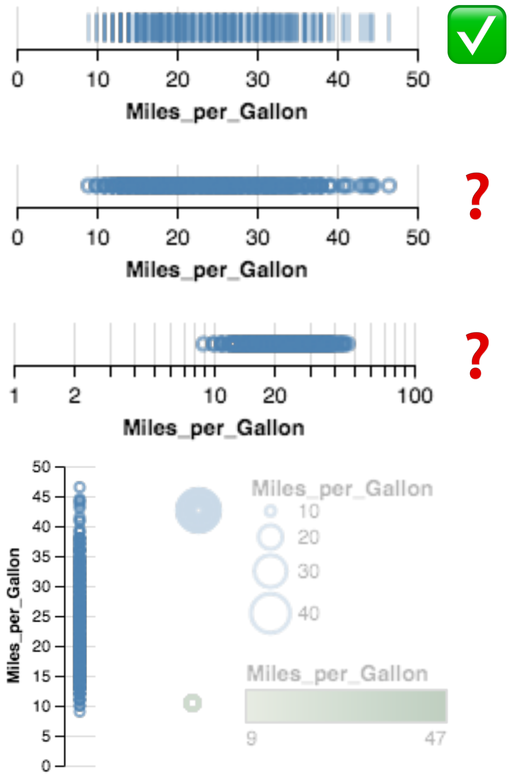


Aggregate (Count)

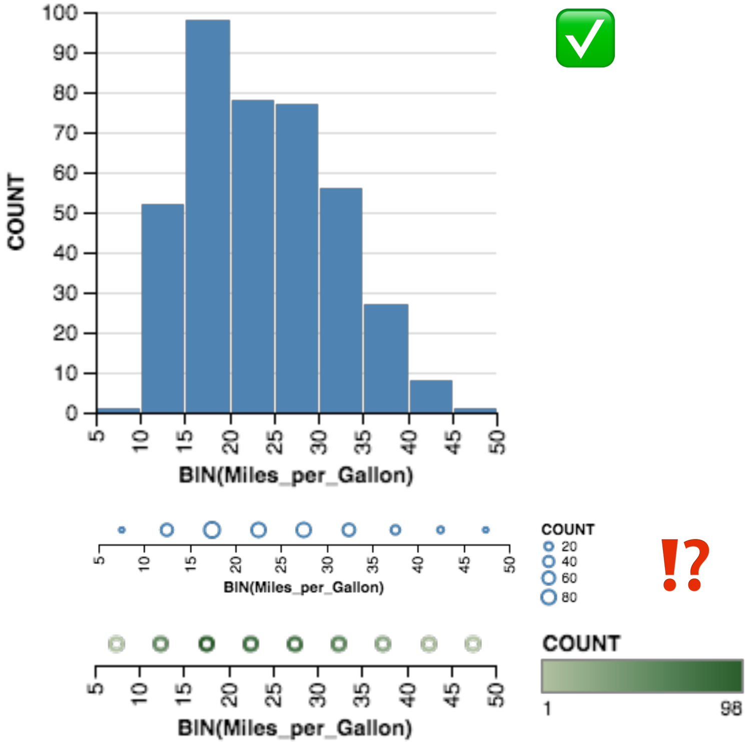


Effective?

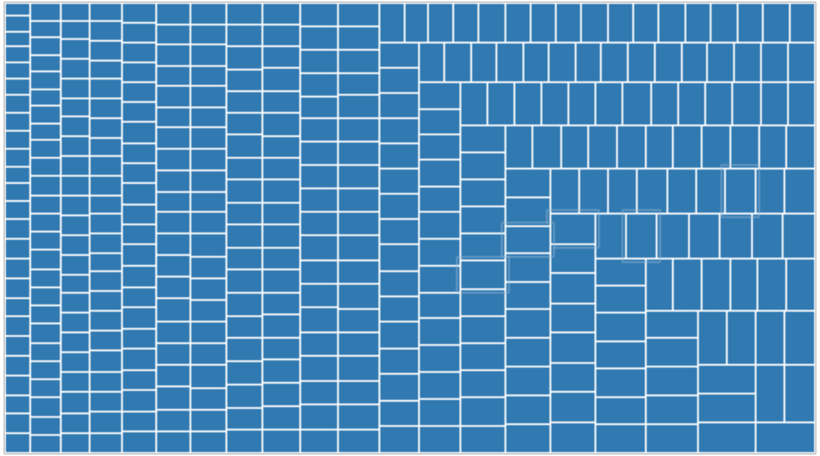
Raw



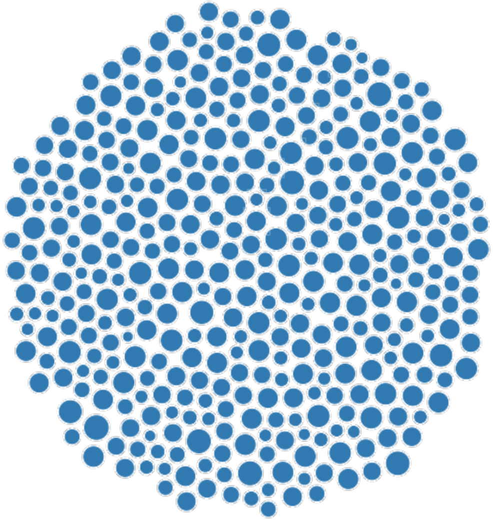
Aggregate (Count)



Raw (with Layout Algorithm)

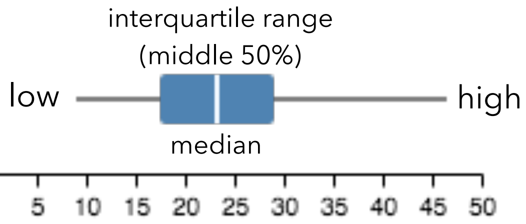


Treemap

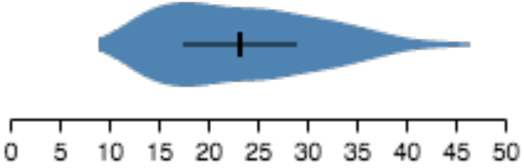


Bubble Chart

Aggregate (Distributions)



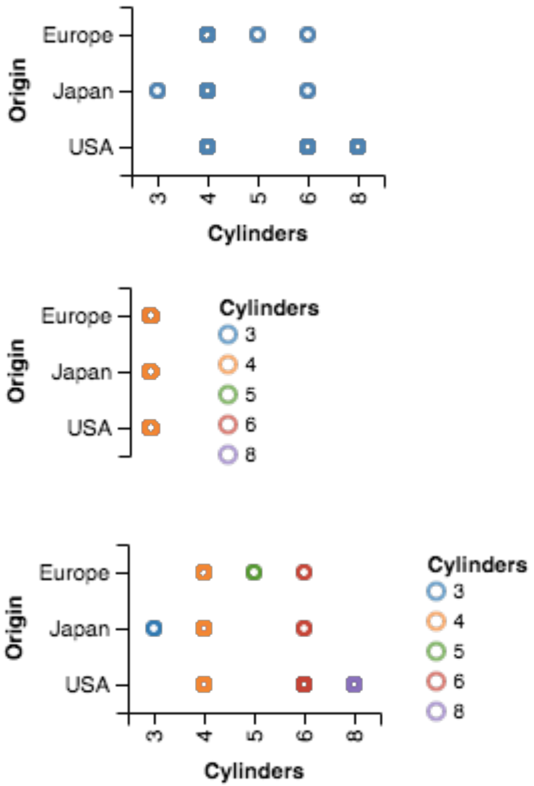
Box Plot



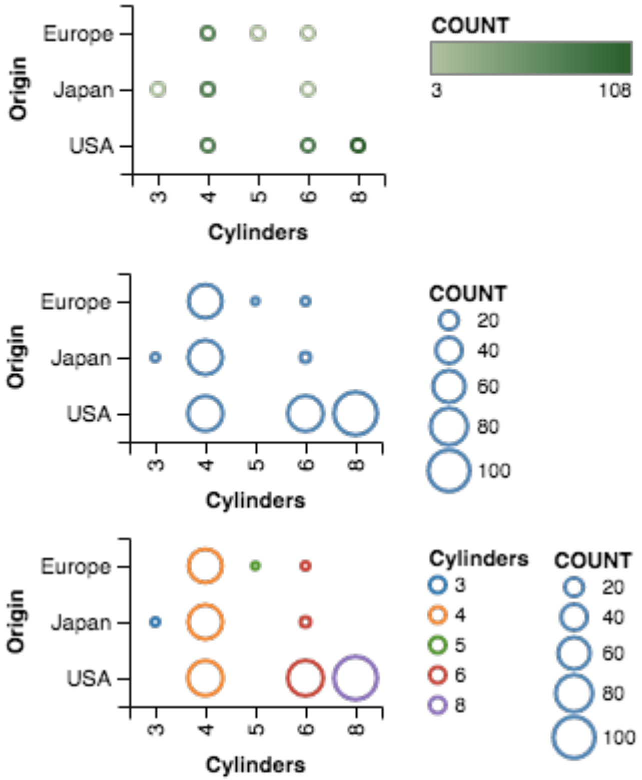
Violin Plot

2D: Nominal x Nominal

Raw

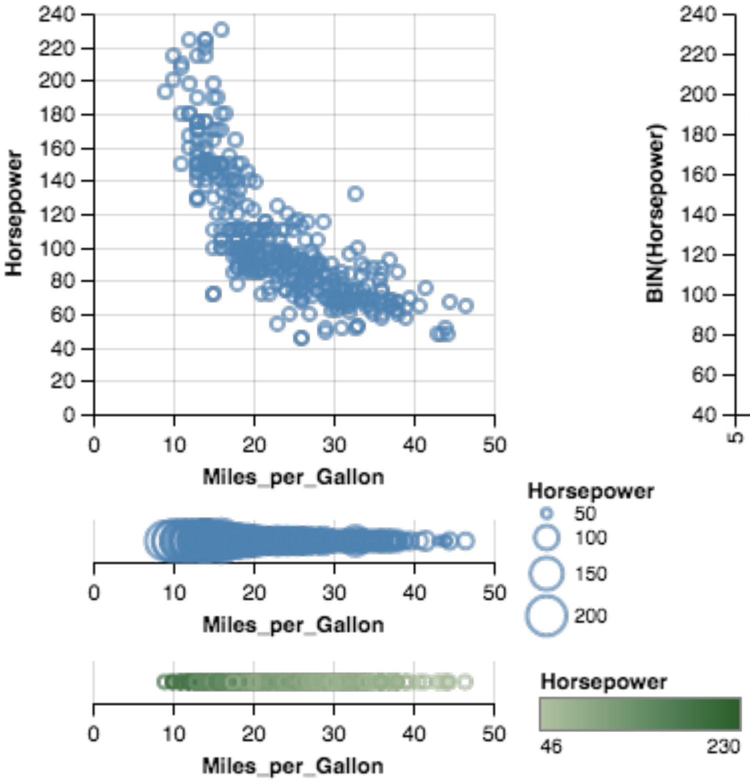


Aggregate (Count)

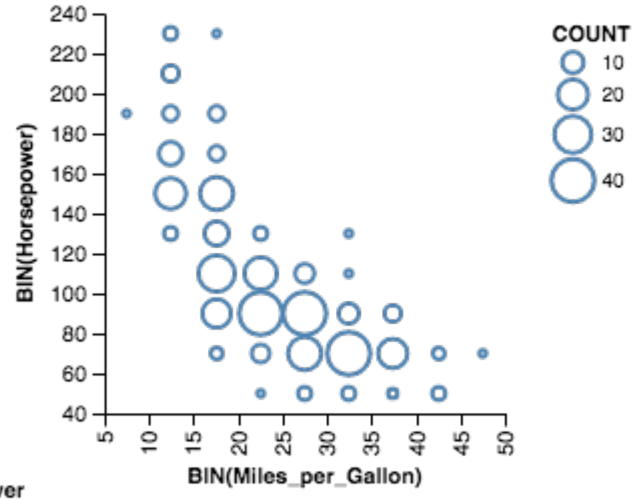


2D: Quantitative x Quantitative

Raw

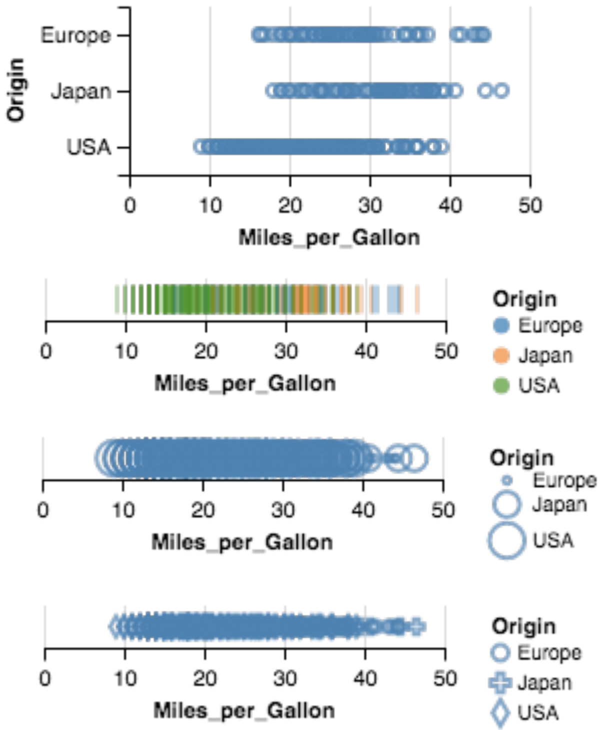


Aggregate (Count)

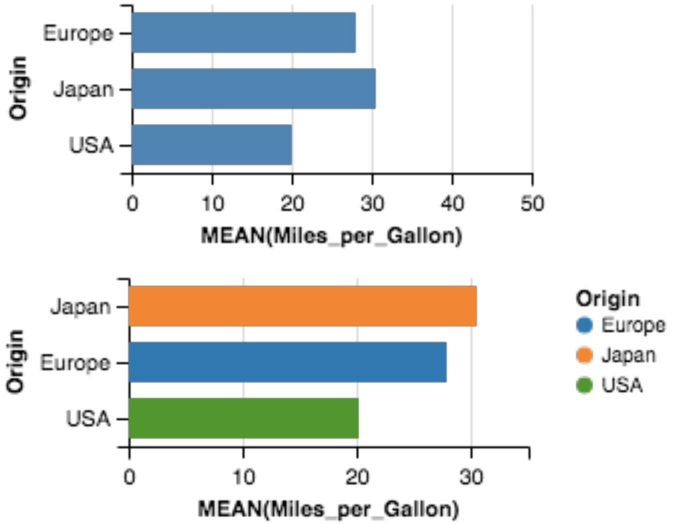


2D: Nominal x Quantitative

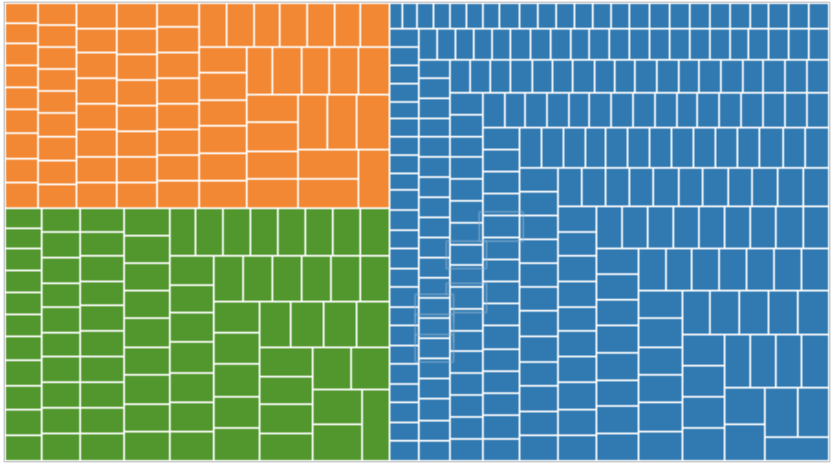
Raw



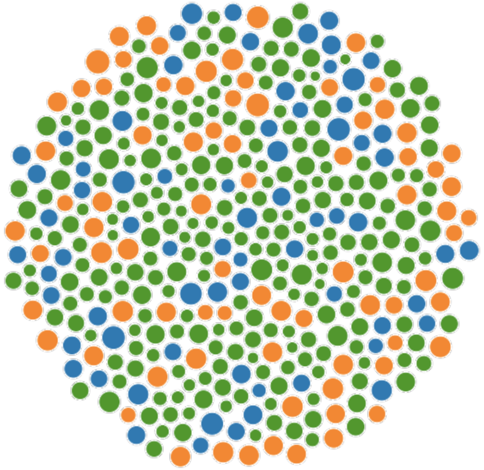
Aggregate (Mean)



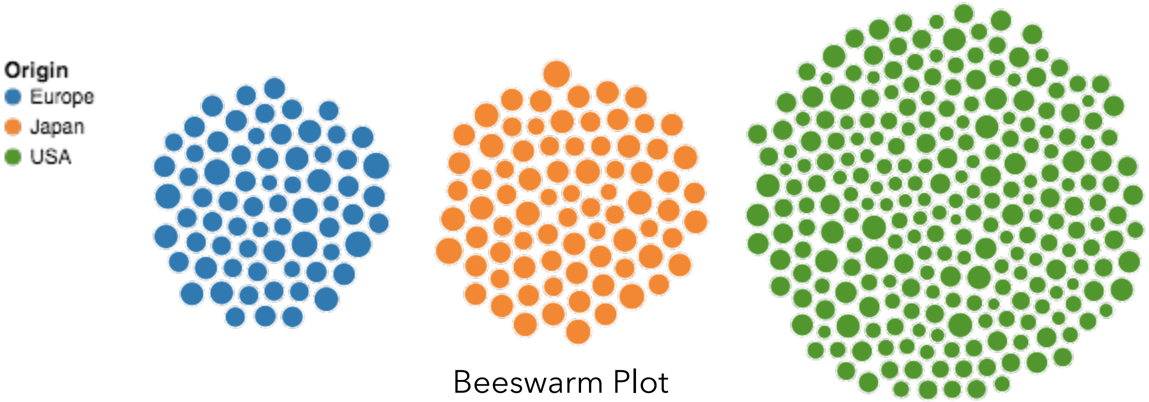
Raw (with Layout Algorithm)



Treemap



Bubble Chart



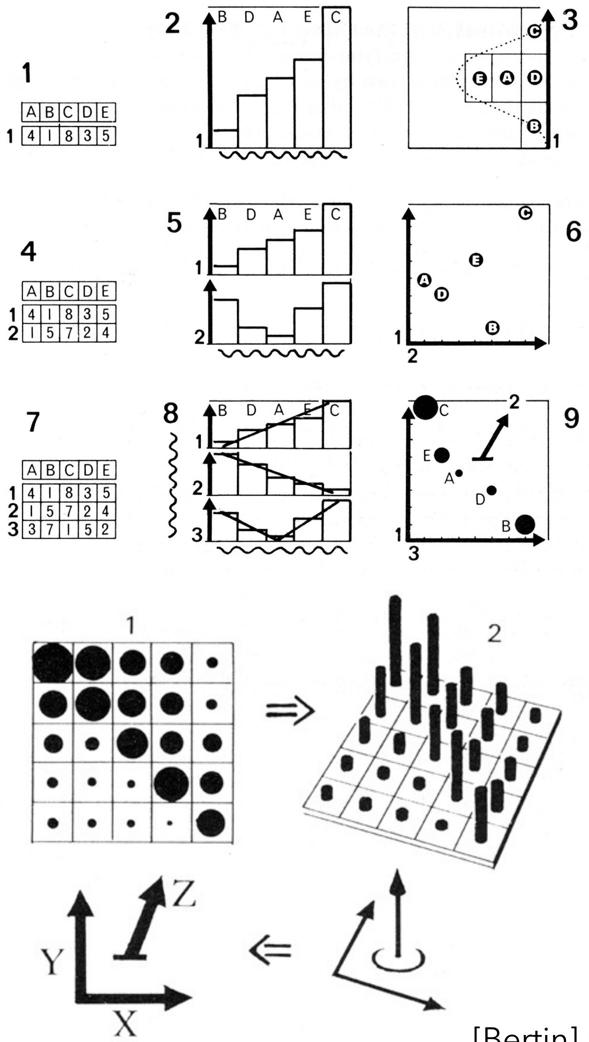
Beeswarm Plot

3D and Higher

Two variables [x, y]
Can map to 2D points.
Scatterplots, maps, …

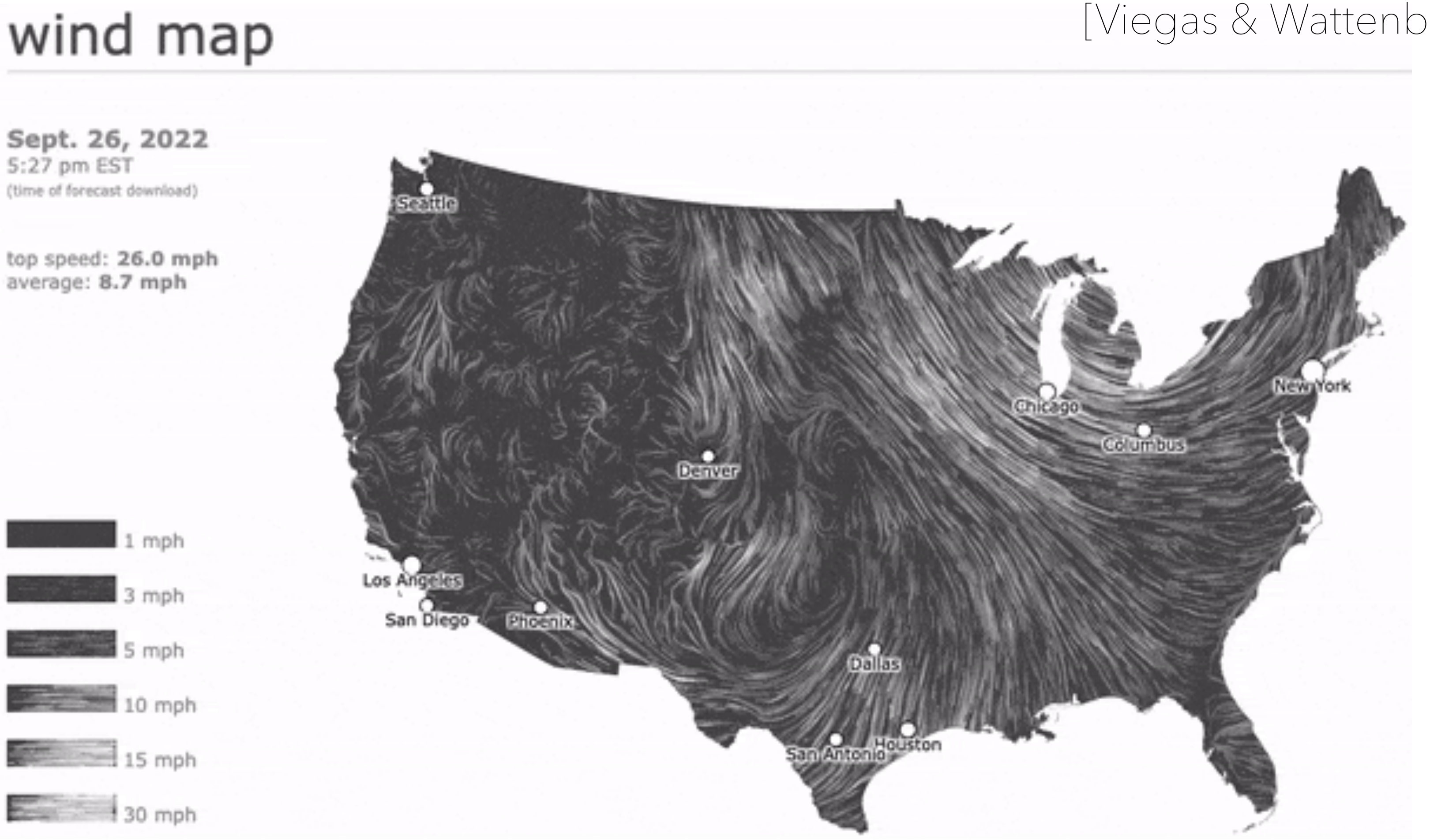
Third variable [z]
Often use one of size, color, opacity, shape, *etc*. Or, one can further partition space.

What about 3D rendering?



[Bertin]

[Viegas & Wattenberg]



Administrivia

A1: Expository Visualization

Using the given climate data set…

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

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

You are free to **use any visualization tool**.

Deliverables via Gradescope

Image of your visualization (PNG or JPG format)

Short description + design rationale (≤ 4 paragraphs)

Due by **EOD, Tue April 7 (today!)**.

Tableau Tutorial

Tableau is a graphical tool for creating visualizations. It can be valuable for rapid exploration and prototyping, even if you ultimately plan to code your own visualization.

Friday, April 11, 4pm

On **Zoom** (see Ed and Canvas later this week for links)

Recording will be posted afterward

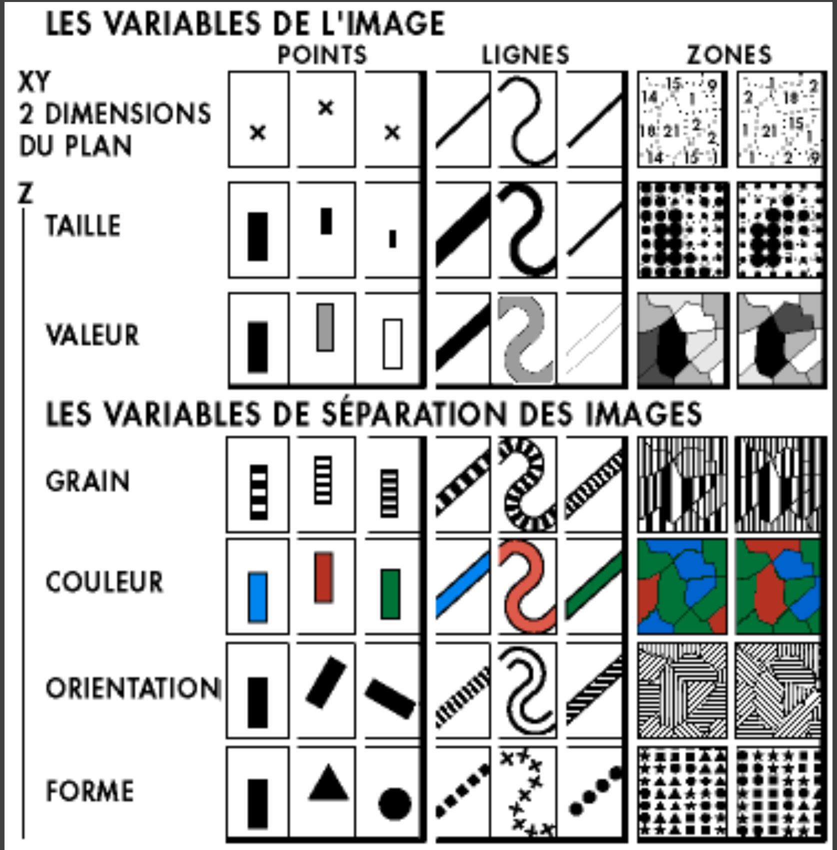
Led by **Katherine**

Multidimensional Data

Visual Encoding Channels

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

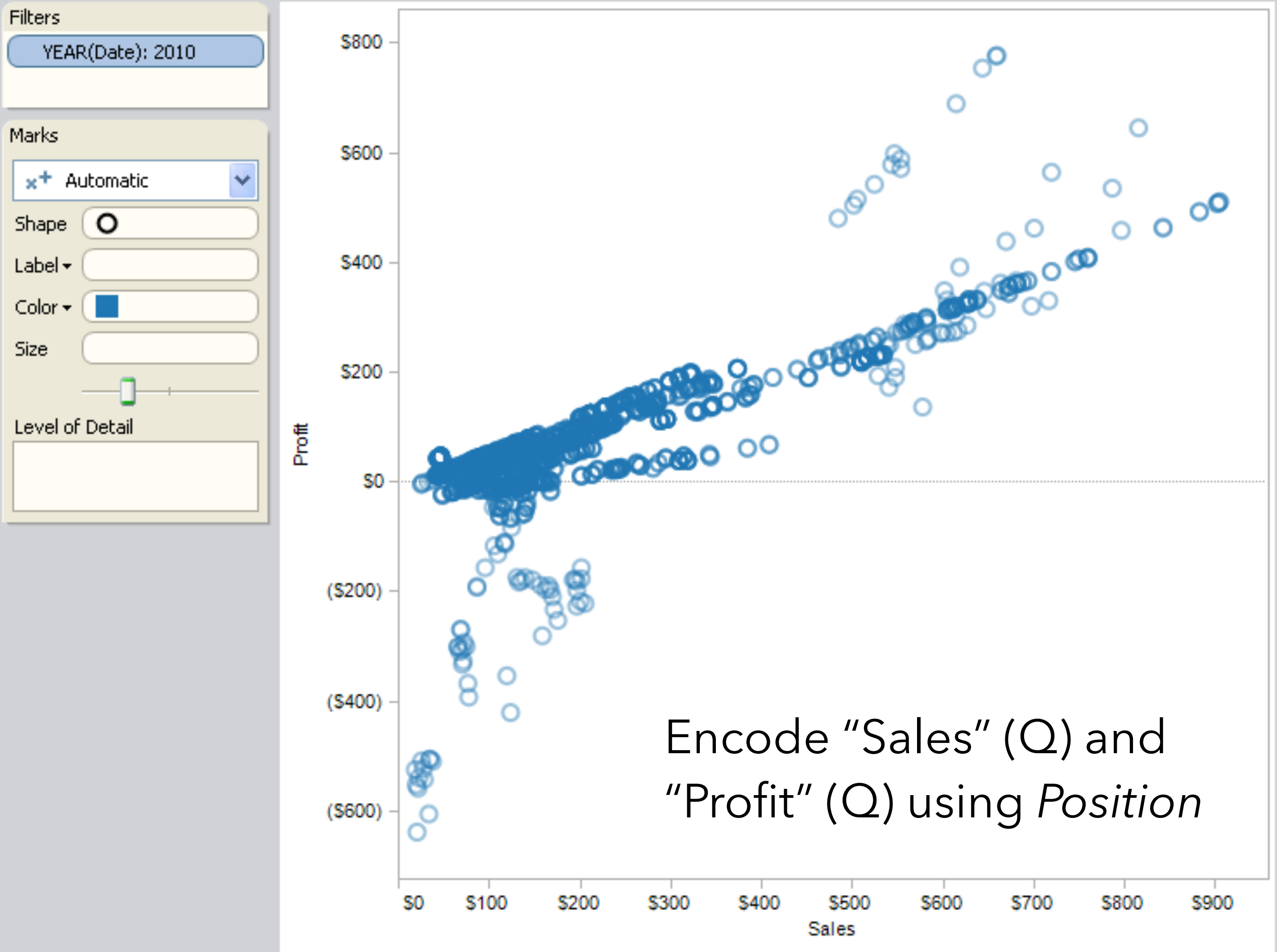
~8 dimensions?



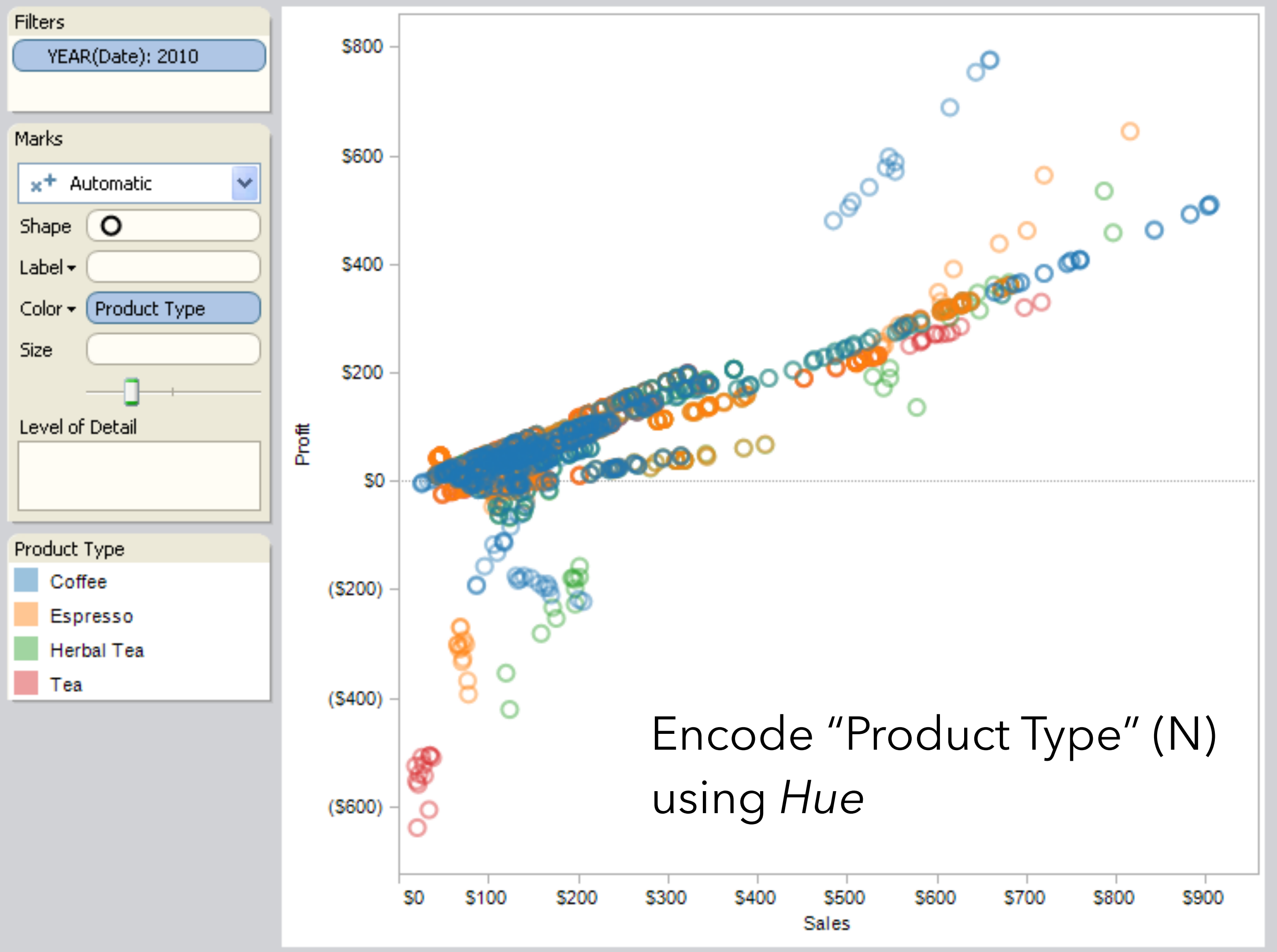
Example: Coffee Sales

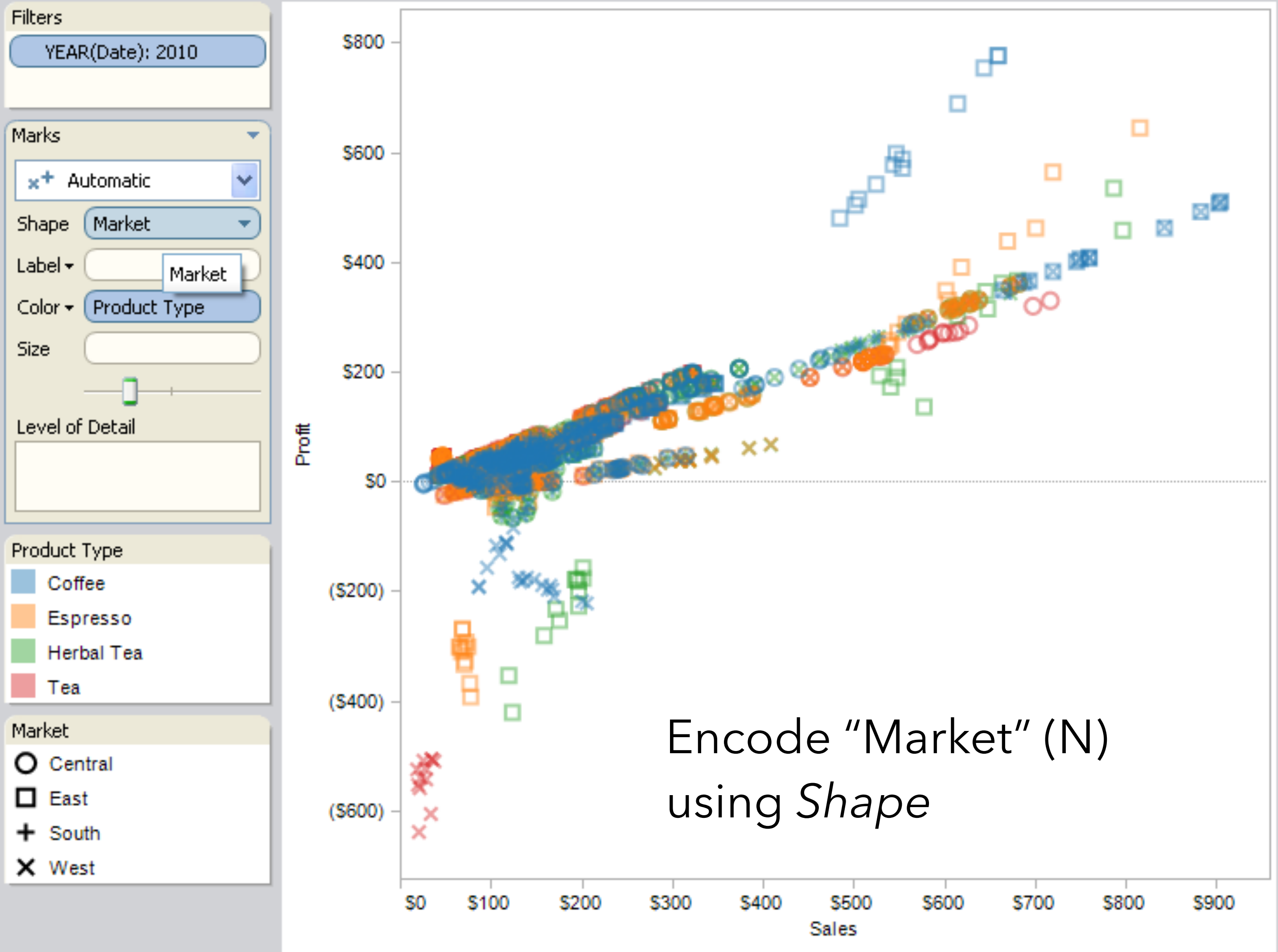
Sales figures for a fictional coffee chain

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

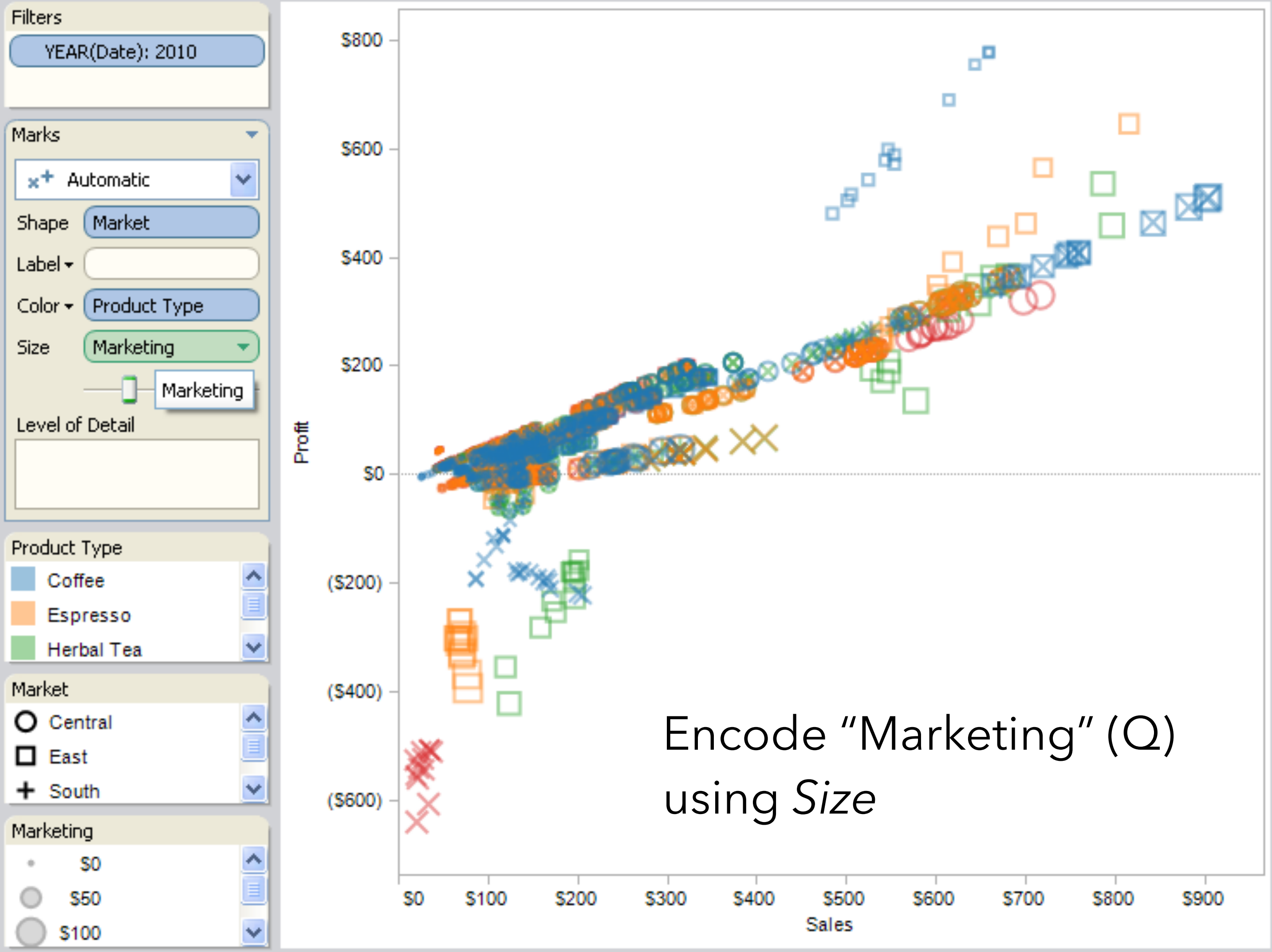


Encode “Sales” (Q) and “Profit” (Q) using *Position*


Filters
YEAR(Date): 2010
Marks
Automatic
Shape
Label
Color
Product Type
Size
Level of Detail
Product Type
Coffee
Espresso
Herbal Tea
Tea
$800
$600
$400
$200
$0
($200)
($400)
($600)
Profit
$0
$100
$200
$300
$400
$500
$600
$700
$800
$900
Sales
Encode "Product Type" (N) using *Hue*

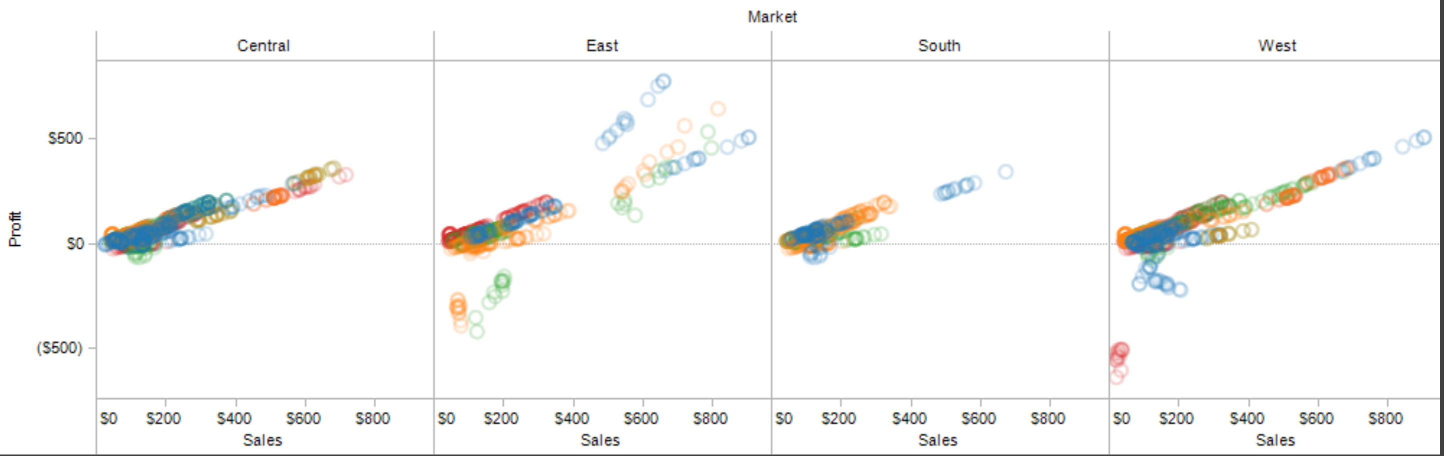


Encode "Market" (N) using *Shape*



Encode "Marketing" (Q) using *Size*

Trellis Plots



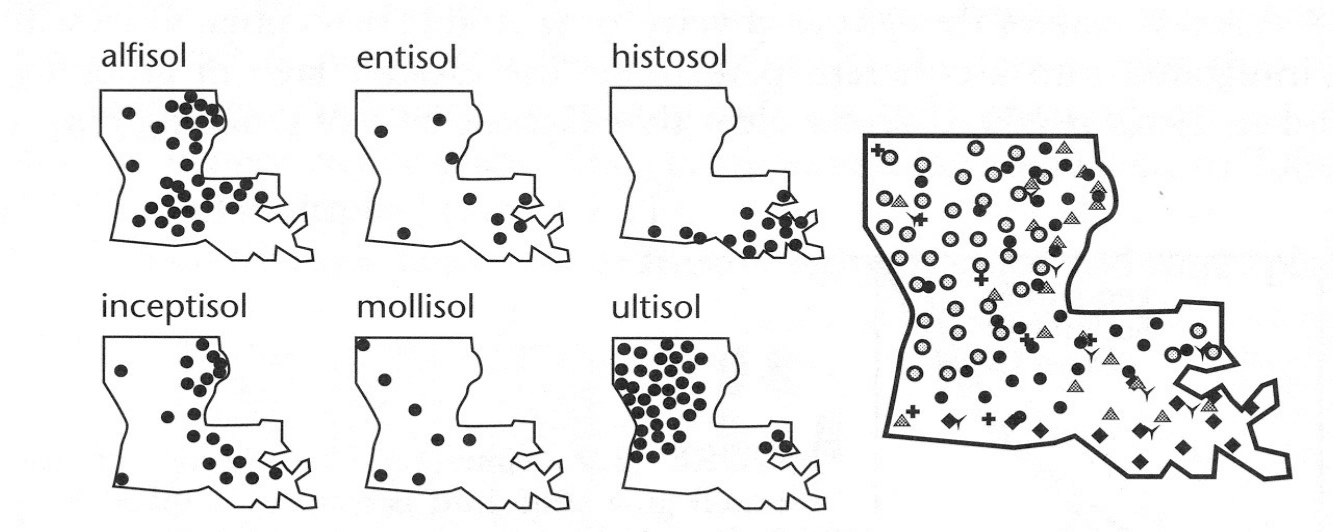
A *trellis plot* subdivides space to enable comparison across multiple plots.

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

Small Multiples

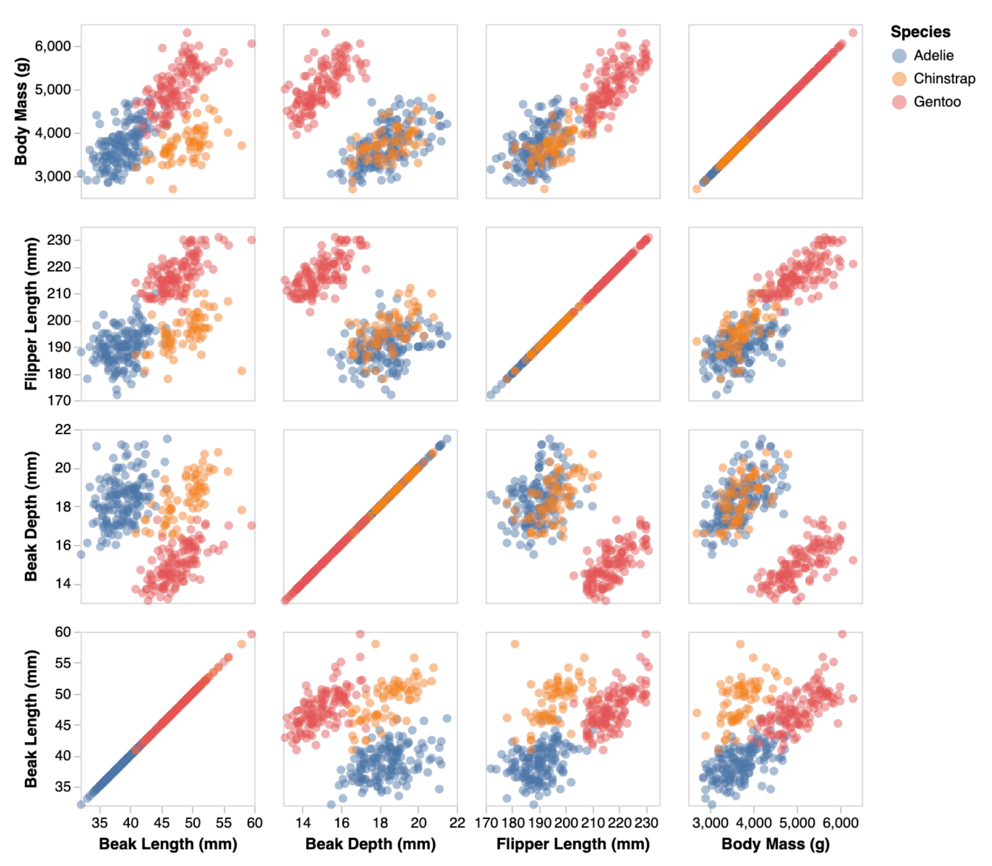
[MacEachren '95, Figure 2.11, p. 38]

Small Multiples



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

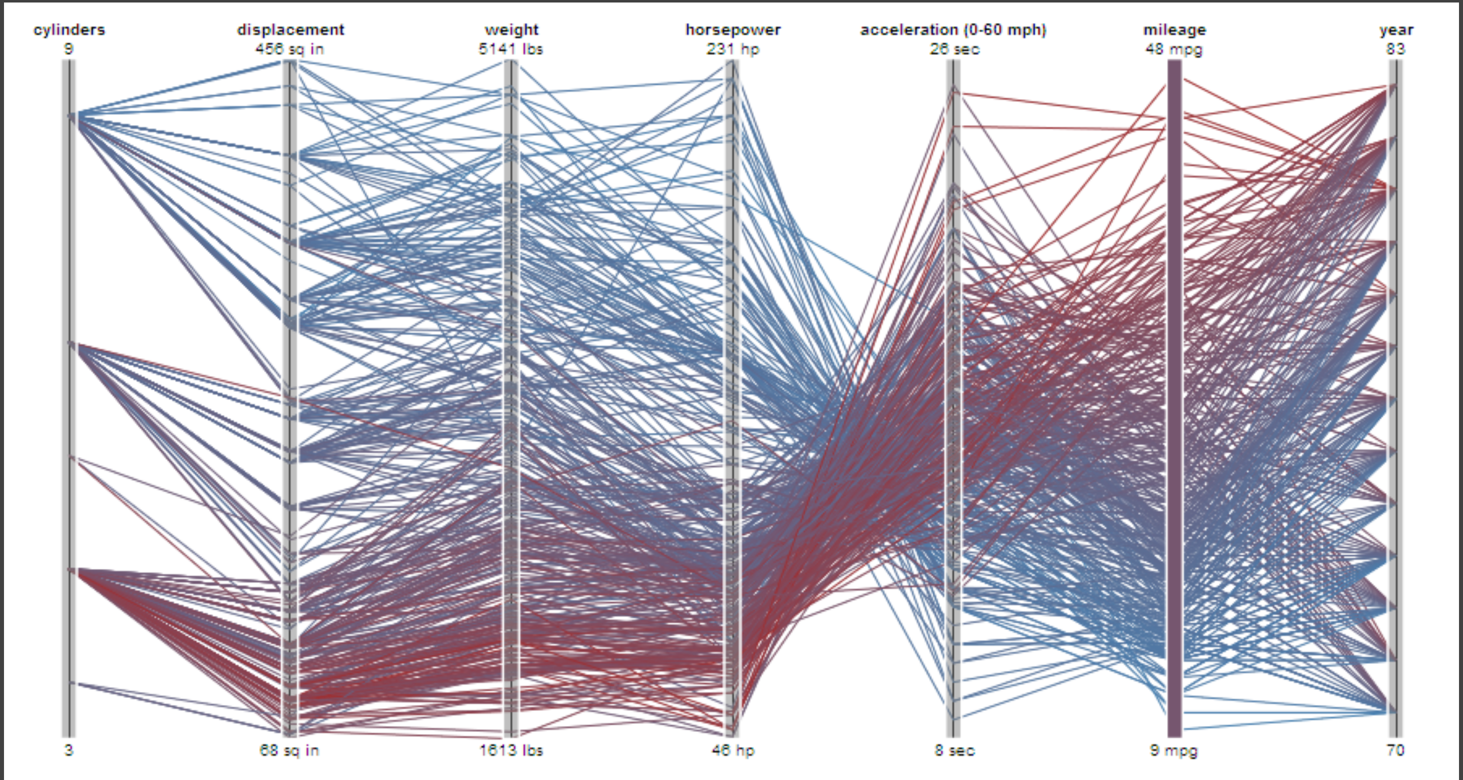
Scatterplot Matrix (SPLOM)



Scatter plots for pairwise comparison of each data dimension.

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!

Scales & Axes

Scale Transforms

$$f\colon D \rightarrow R$$

A **scale** is a function that maps a domain D of data values to a range R of visual values.

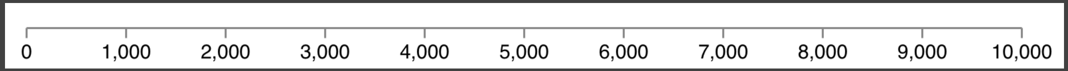
Example ranges: x-position, color, size, angle

Scales are the workhorses of visual encoding!

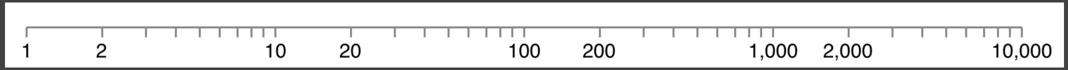
We can modify domains, ranges, transforms (log, etc.), padding, and more…

Positional Scales $R = pixels$

Continuous / Quantitative



linear

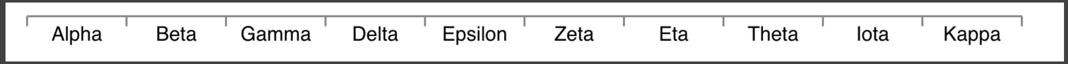


log

sqrt

Discrete / Ordinal

point



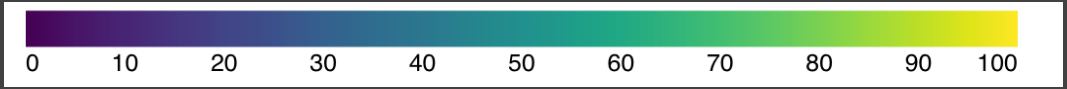
band

Color Scales $R = colors$

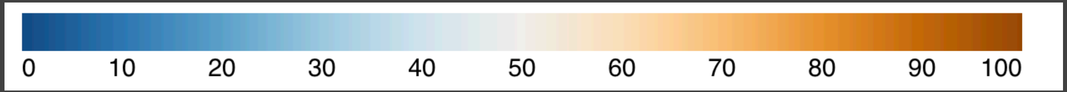
Discrete / Categorical

ordinal

Continuous / Quantitative

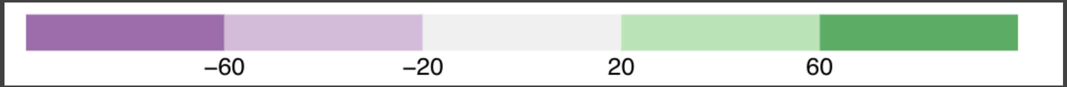


sequential



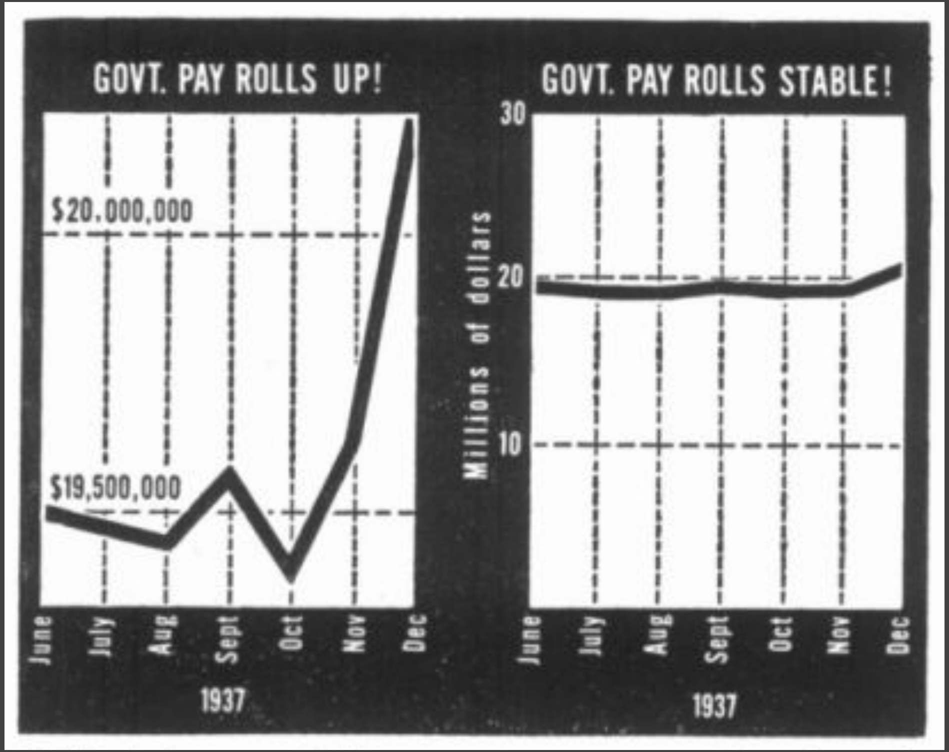
diverging

Discretized / Binned Quantitative



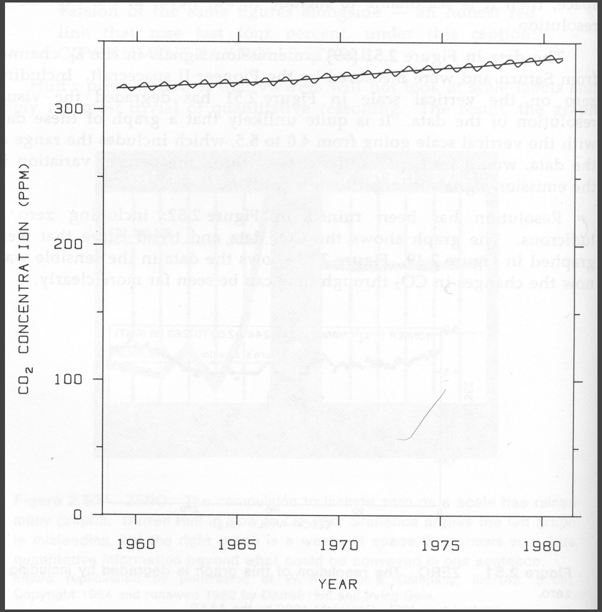
quantize

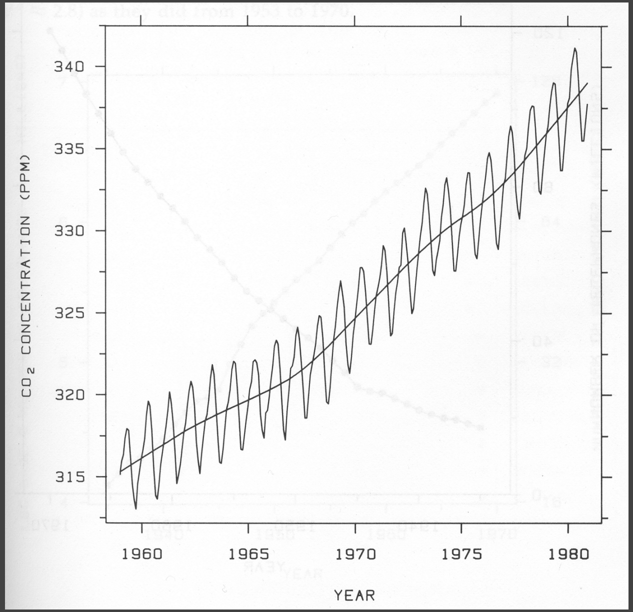
Include Zero in Axis Scale?



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

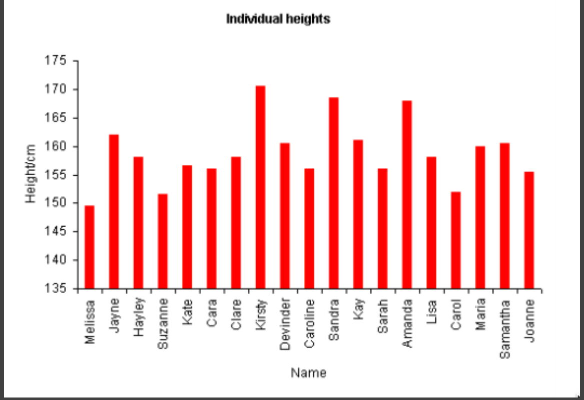
Include Zero in Axis Scale?





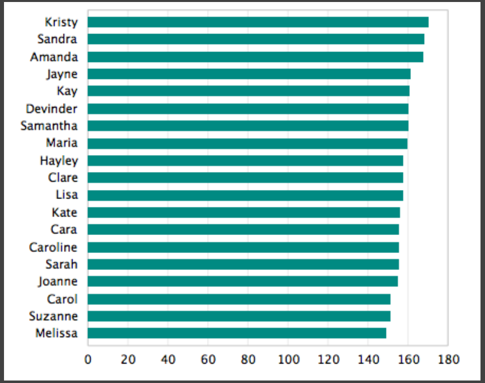
Yearly CO_2 concentrations [Cleveland 85]

Include Zero in Axis Scale?

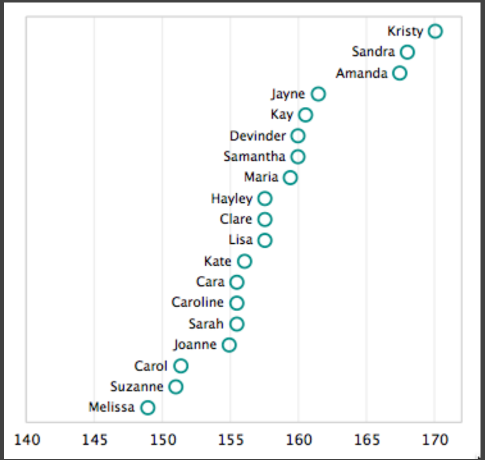


Violates Expressiveness Principle!

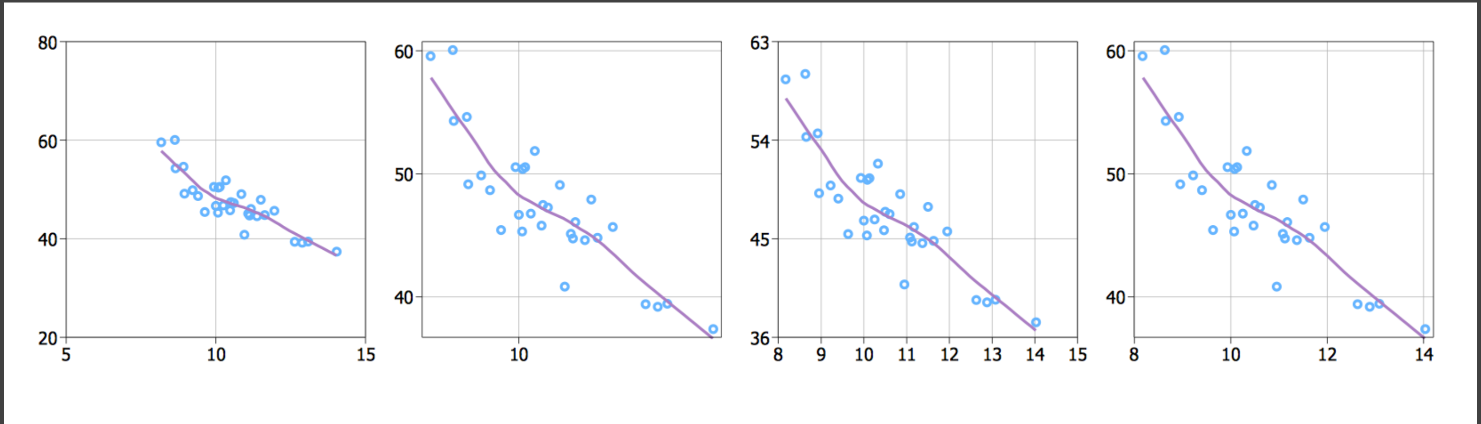
Compare Proportions (Q-Ratio)



Compare Relative Position (Q-Interval)

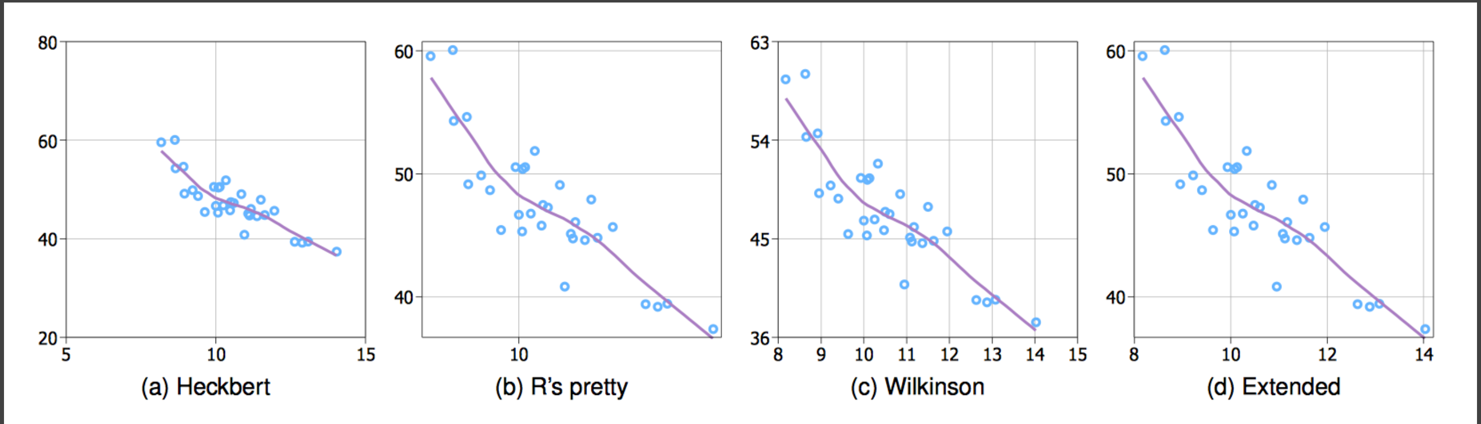


Axis Tick Mark Selection



What are some properties of "good" tick marks?

Axis Tick Mark Selection



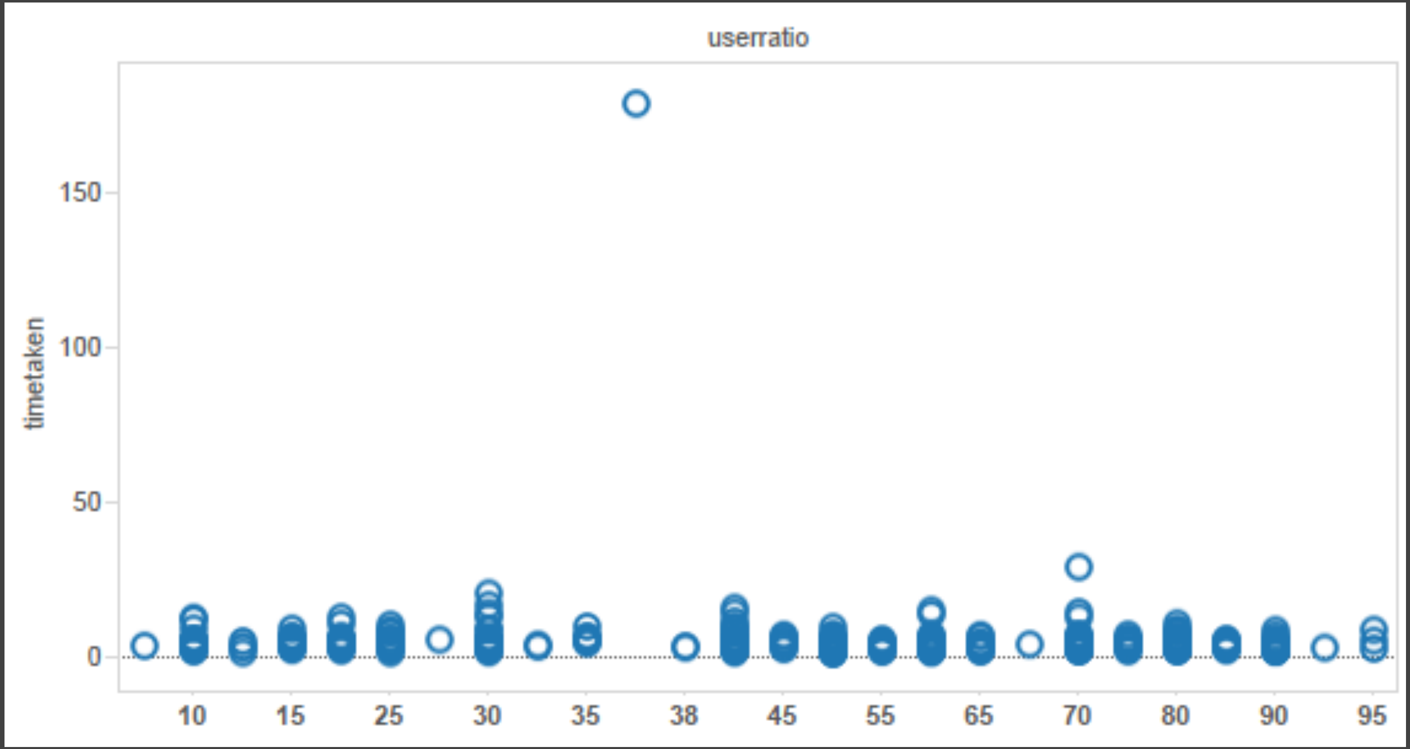
Simplicity - numbers are multiples of 10, 5, 2

Coverage - ticks near the ends of the data

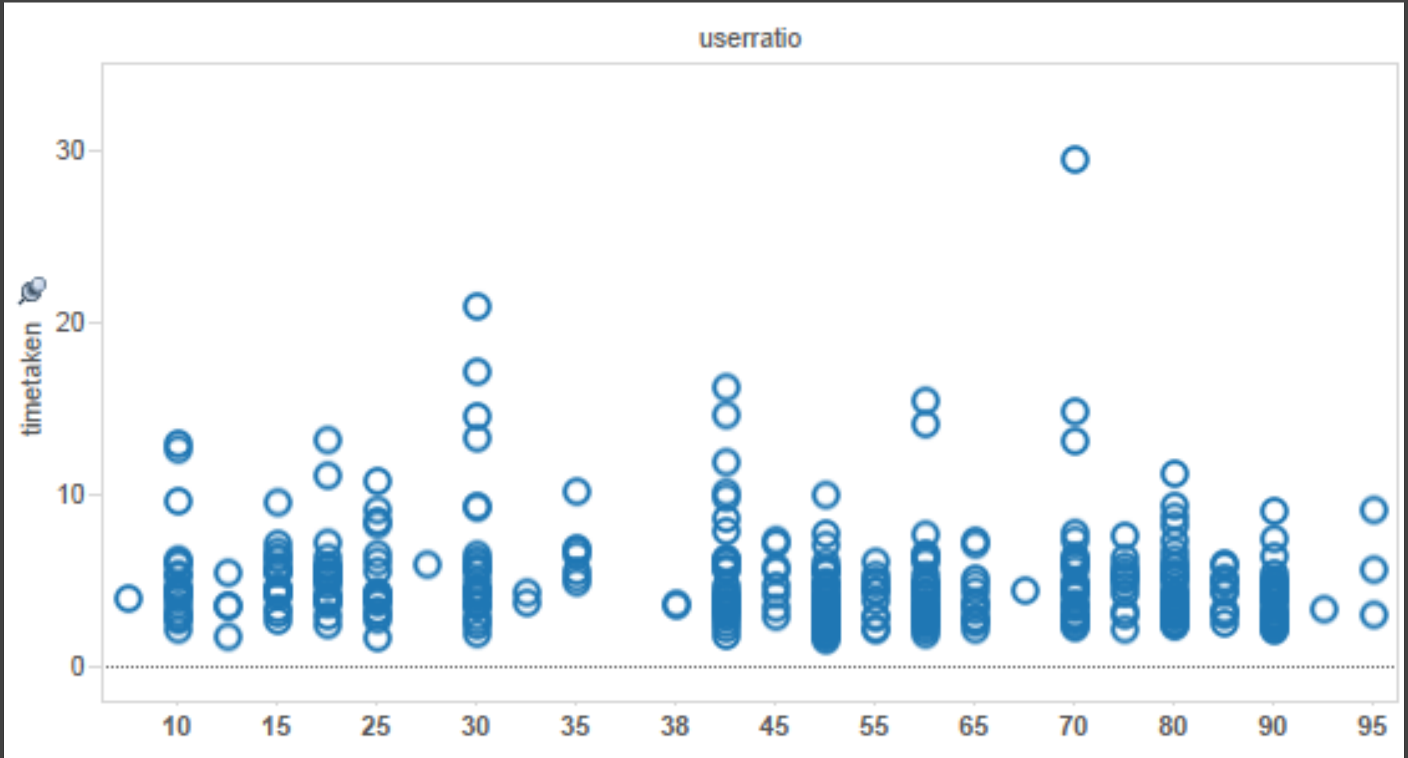
Density - not too many, nor too few

Legibility - whitespace, horizontal text, size

How to Scale the Axis?

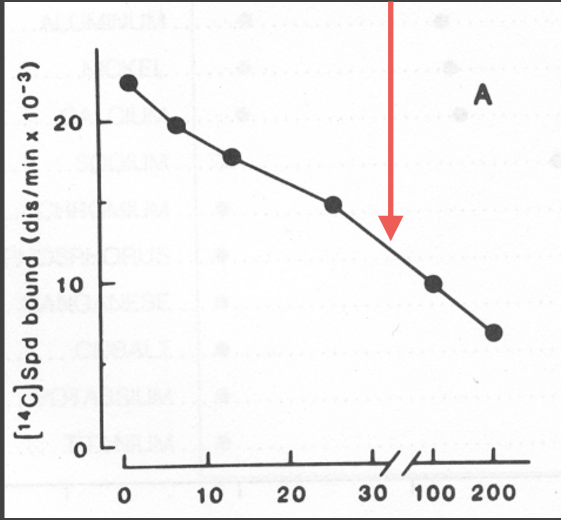


One Option: Clip Outliers

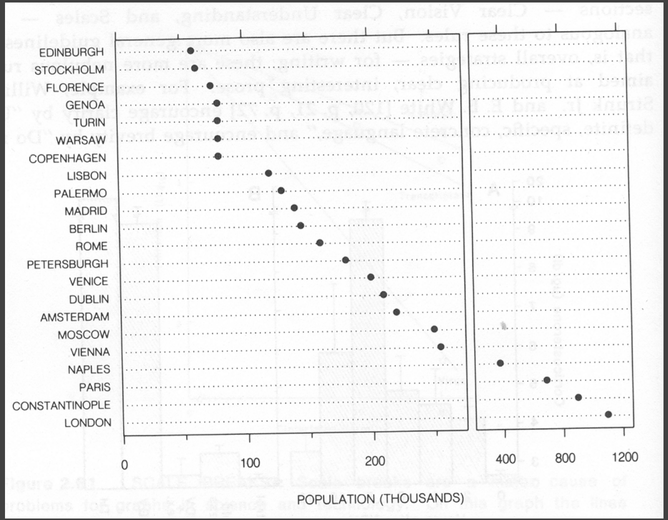


Clearly Mark Scale Breaks

Violates Expressiveness Principle!

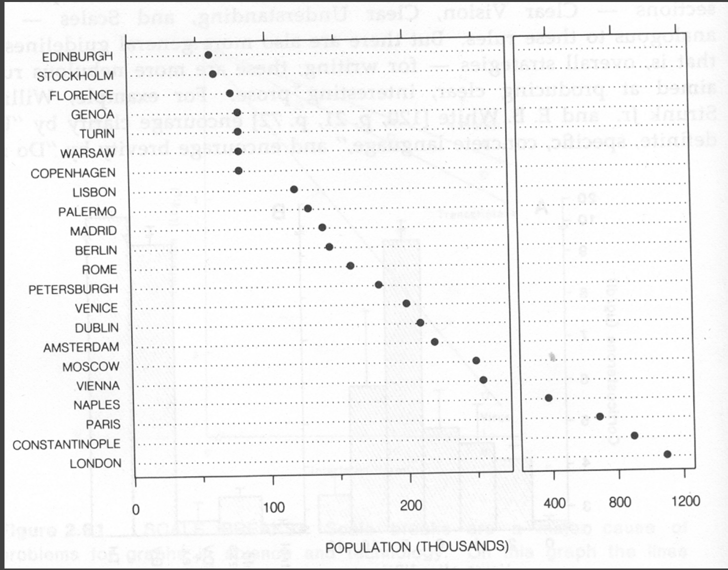


Poor scale break [Cleveland 85]

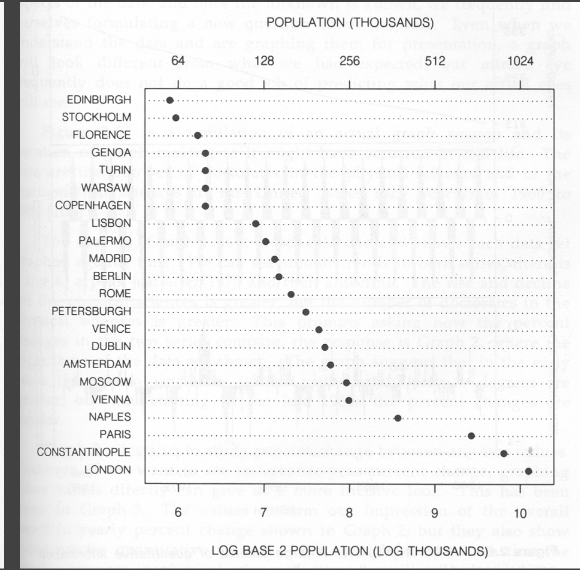


Well-marked scale break [Cleveland 85]

Scale Break vs. Log Scale



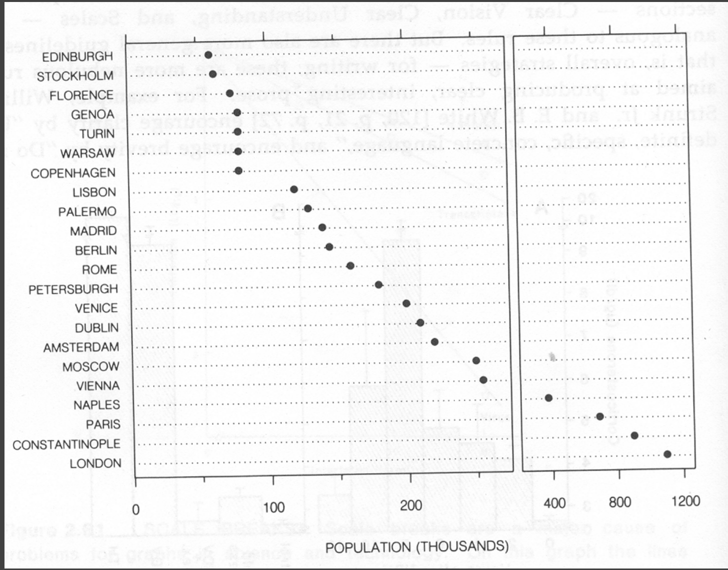
Scale Break

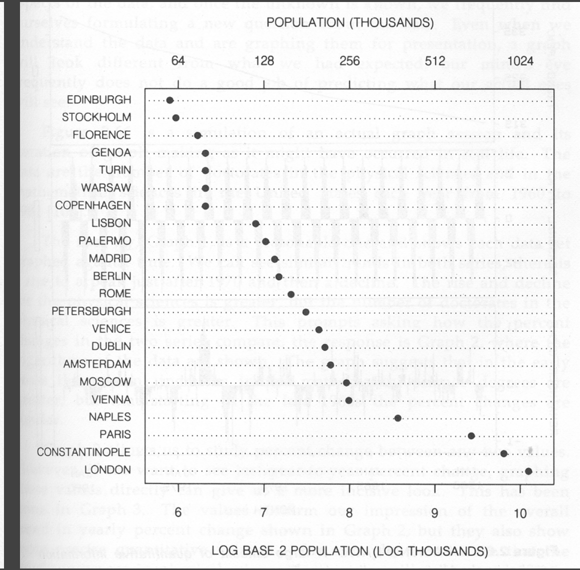


Log Scale

[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

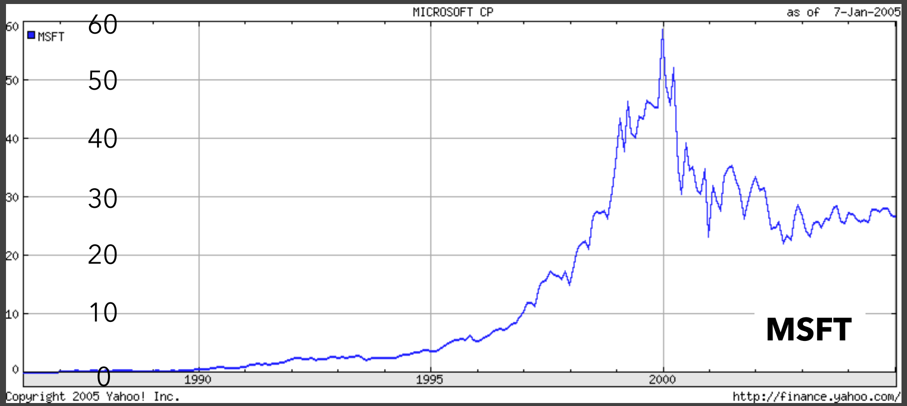
Logarithms turn *multiplication* into *addition*.

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

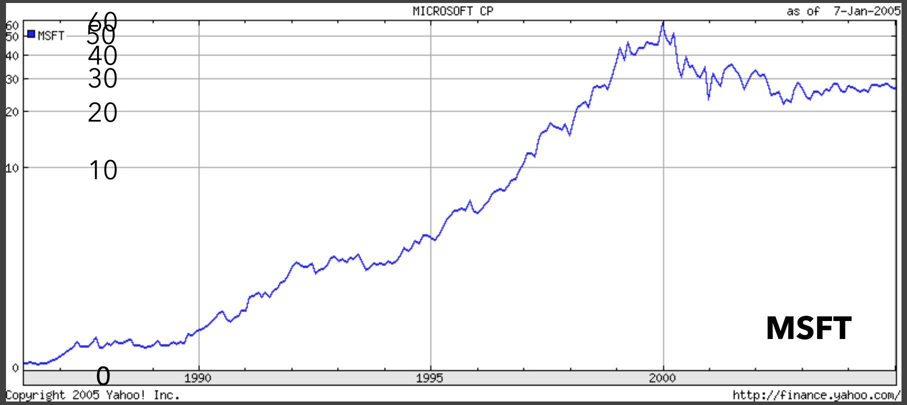
Equal steps on a log scale correspond to equal changes to a multiplicative scale factor.

Linear Scale vs. Log Scale

Linear Scale



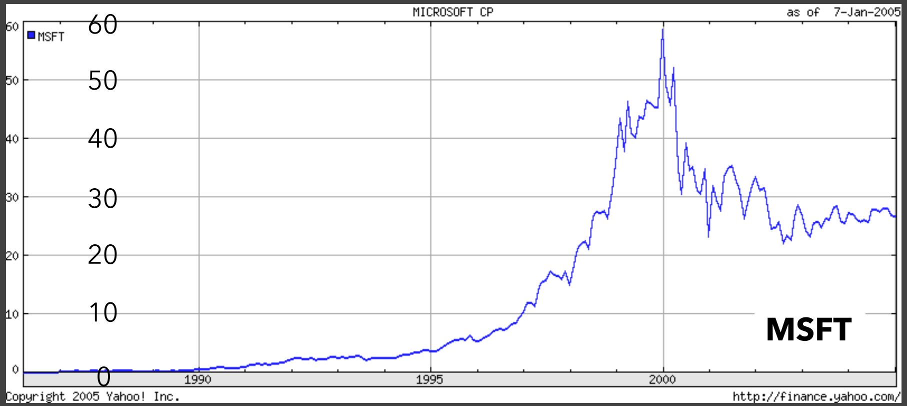
Log Scale



Linear Scale vs. Log Scale

Linear Scale

Absolute change

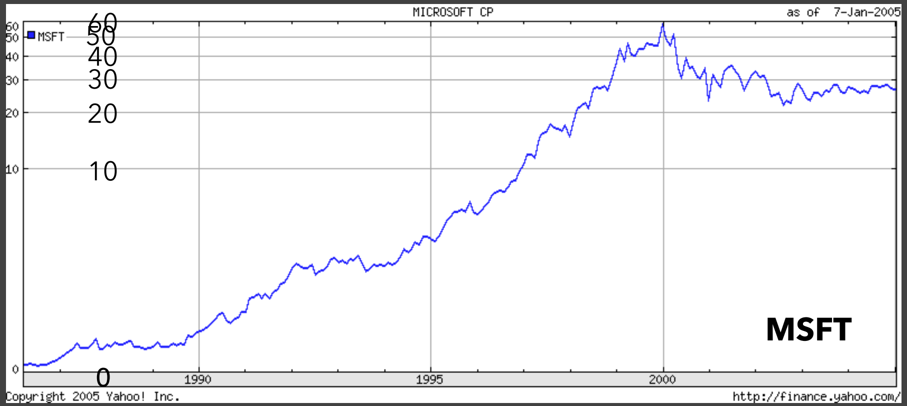


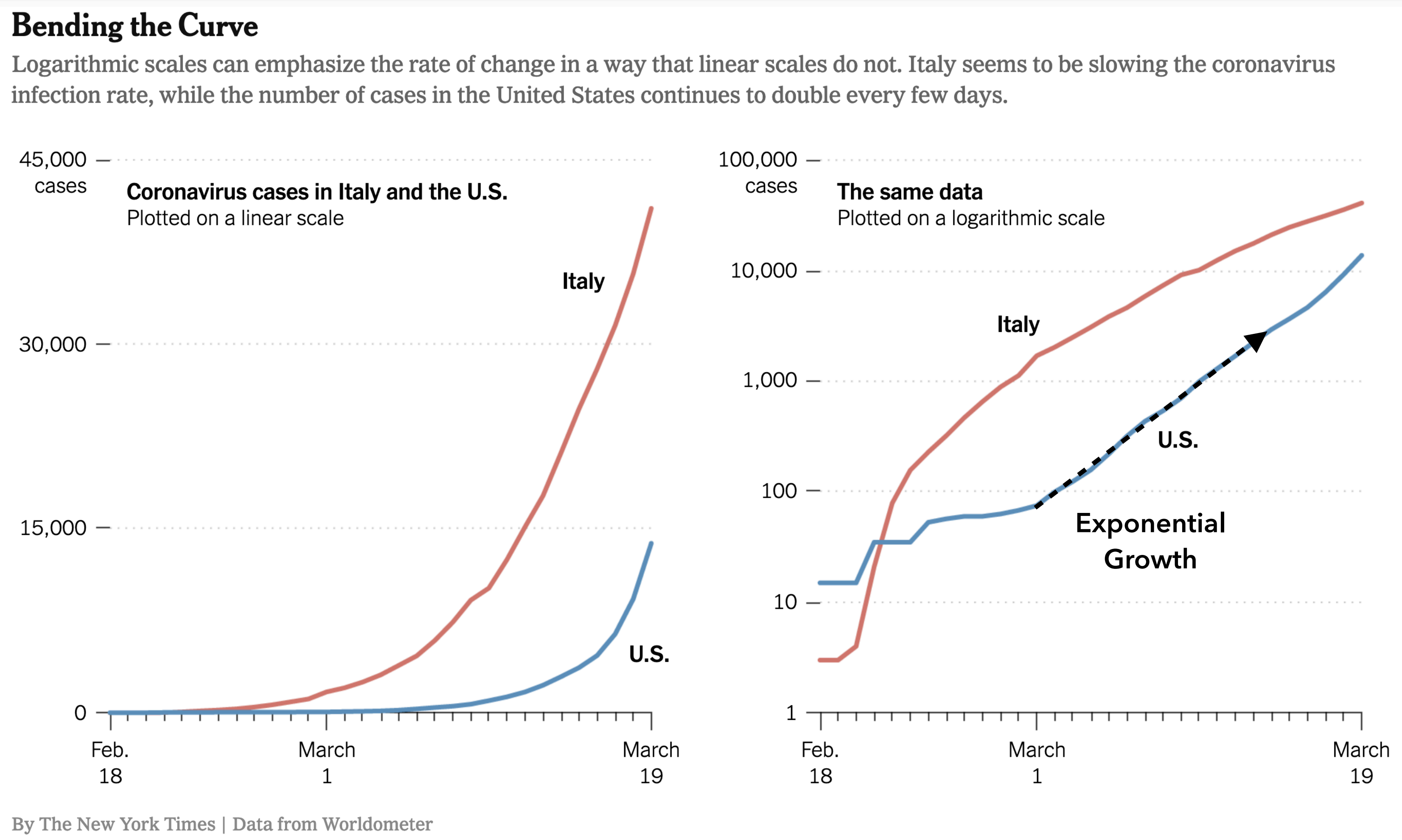
Log Scale

Small fluctuations

Percent change

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




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.
45,000 cases
30,000
15,000
0
Coronavirus cases in Italy and the U.S.
Plotted on a linear scale
Italy
U.S.
Feb. 18
March 1
March 19
100,000 cases
10,000
1,000
100
10
1
The same data
Plotted on a logarithmic scale
Italy
U.S.
Exponential Growth
Feb. 18
March 1
March 19
By The New York Times | Data from Worldometer

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?**

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!