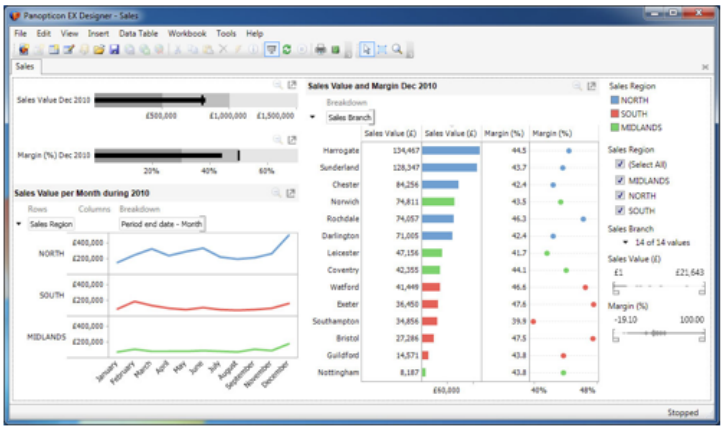
IAT 355 Visual Analytics

Tabular Data Visualization Techniques

Lyn Bartram

SFU

Visualization techniques serve different needs



Tabular data

ID	Sepal Length	Sepal Width	Petal Length	Petal Width	Species
14	4.3	3	1.1	0.1	setosa
39	4.4	3	1.3	0.2	setosa
43	4.4	3.2	1.3	0.2	setosa
9	4.4	2.9	1.4	0.2	setosa
42	4.5	2.3	1.3	0.3	setosa
23	4.6	3.6	1	0.2	setosa
48	4.6	3.2	1.4	0.2	setosa
4	4.6	3.1	1.5	0.2	setosa
7	4.6	3.4	1.4	0.3	setosa
3	4.7	3.2	1.3	0.2	setosa
30	4.7	3.2	1.6	0.2	setosa
13	4.8	3	1.4	0.1	setosa
12	4.8	3.4	1.6	0.2	setosa
31	4.8	3.1	1.6	0.2	setosa
25	4.8	3.4	1.9	0.2	setosa
46	4.8	3	1.4	0.3	setosa
38	4.9	3.6	1.4	0.1	setosa
10	4.9	3.1	1.5	0.1	setosa
2	4.9	3	1.4	0.2	setosa
35	4.9	3.1	1.5	0.2	setosa
58	4.9	2.4	3.3	1	versicolor
107	4.9	2.5	4.5	1.7	virginica
36	5	3.2	1.2	0.2	setosa
5	5	3.6	1.4	0.2	setosa
50	5	3.3	1.4	0.2	setosa

Q

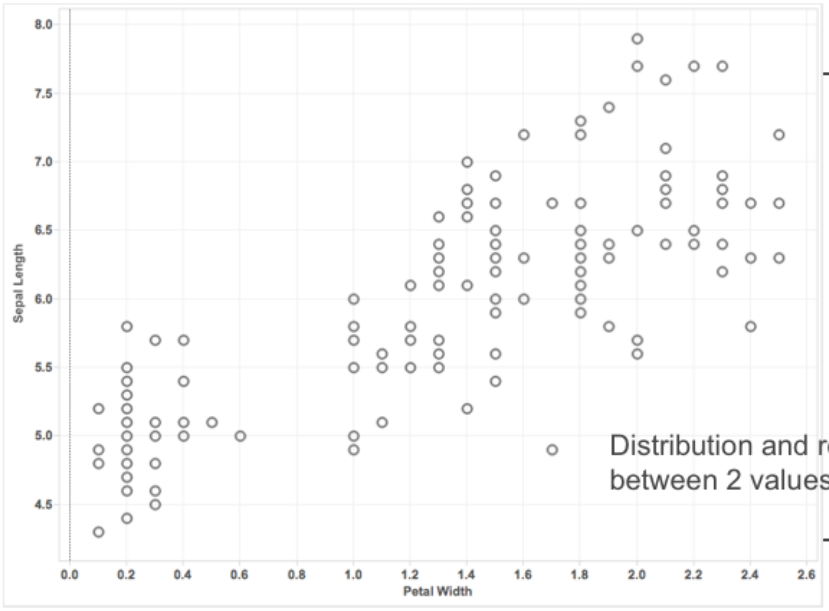
O

N

Sepia and petal length for three species of iris [Fisher 1936]

To date we have been talking about TABULAR DATA – data that we can express in a matrix, or table, indexed by rows and columns

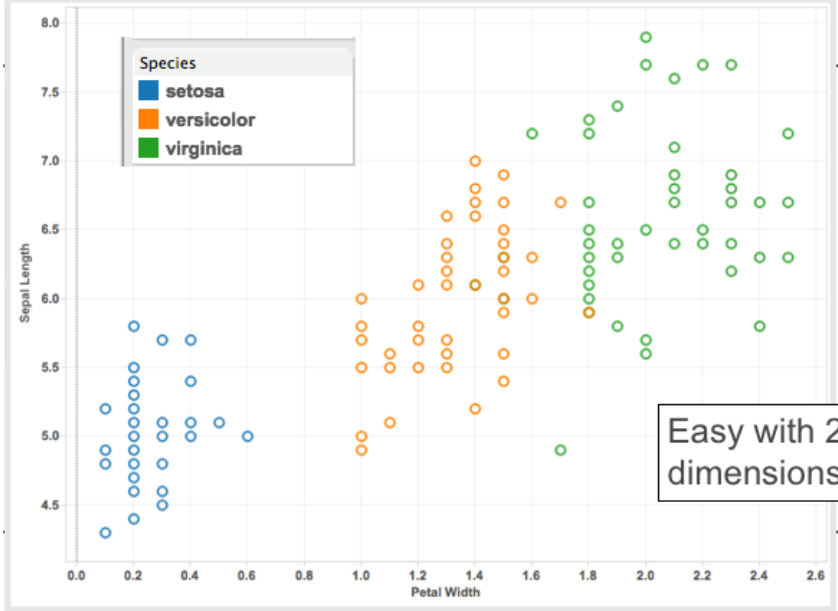
Possible views – scatter plot – why?



Distribution and relations between 2 values

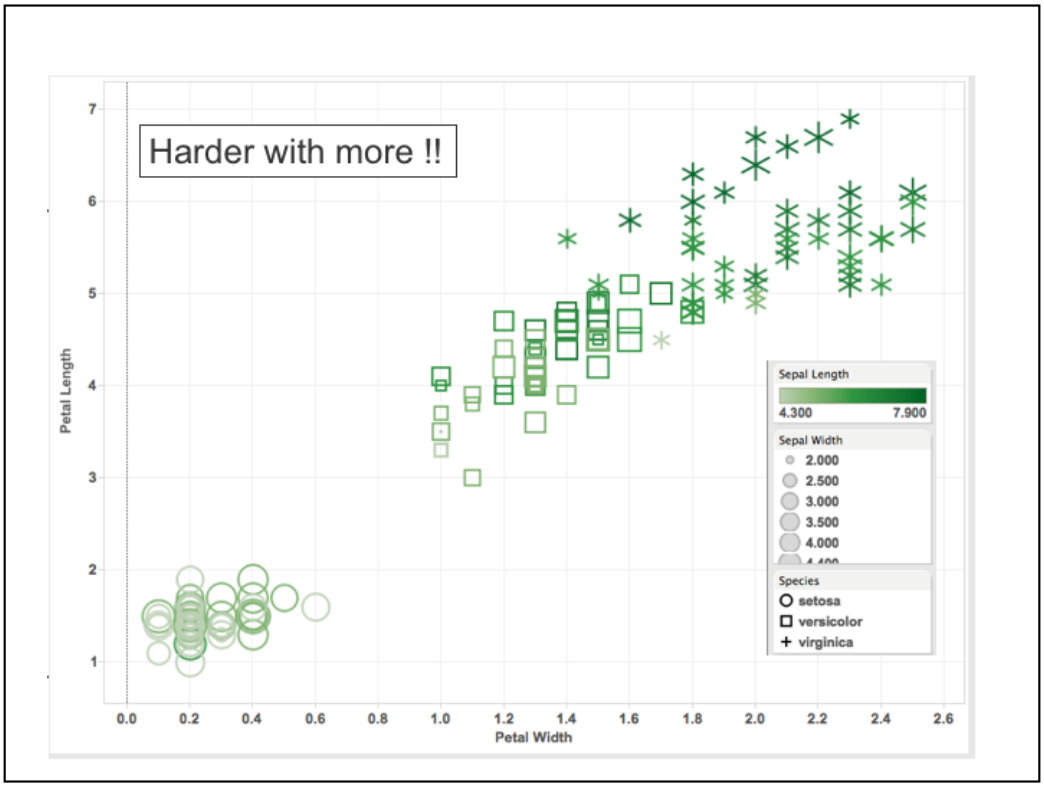
4

Possible views – scatter plot – why?

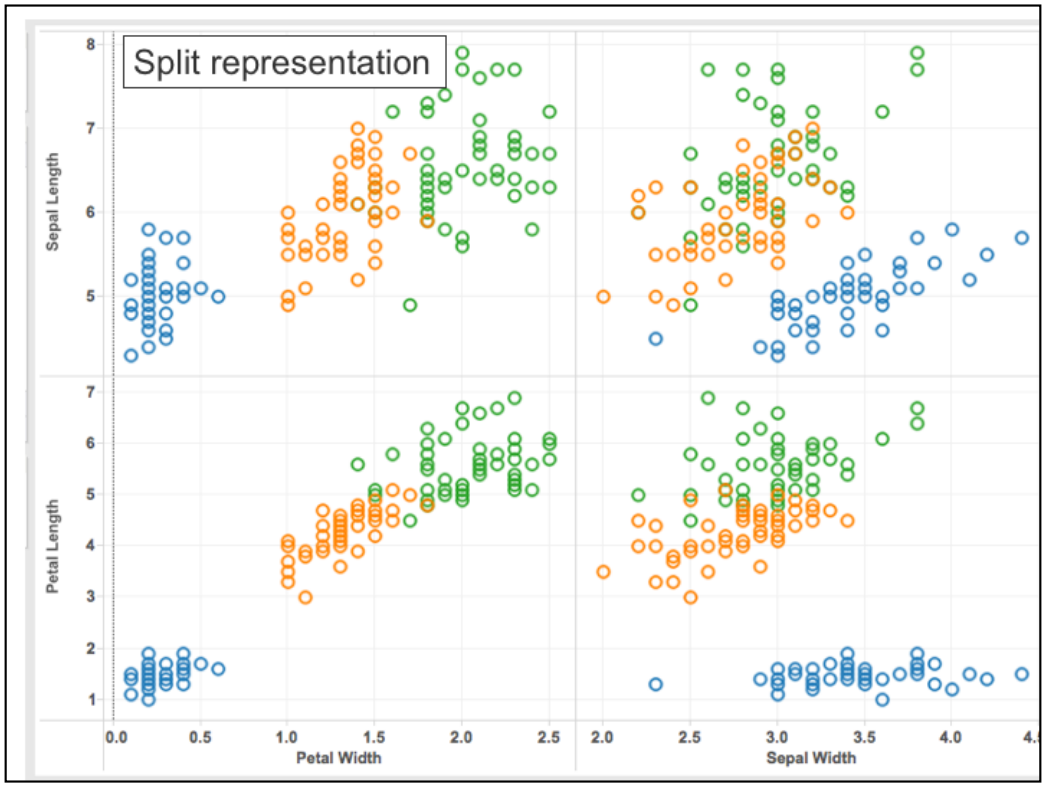


Easy with 2/3 dimensions

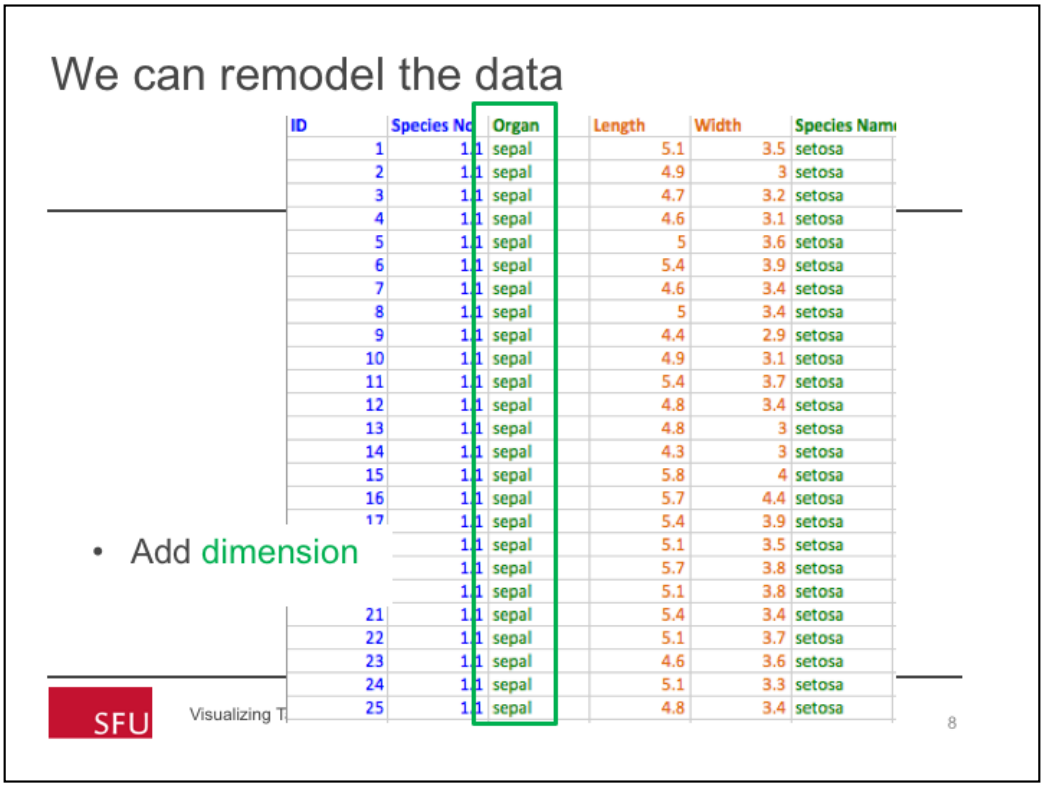
5



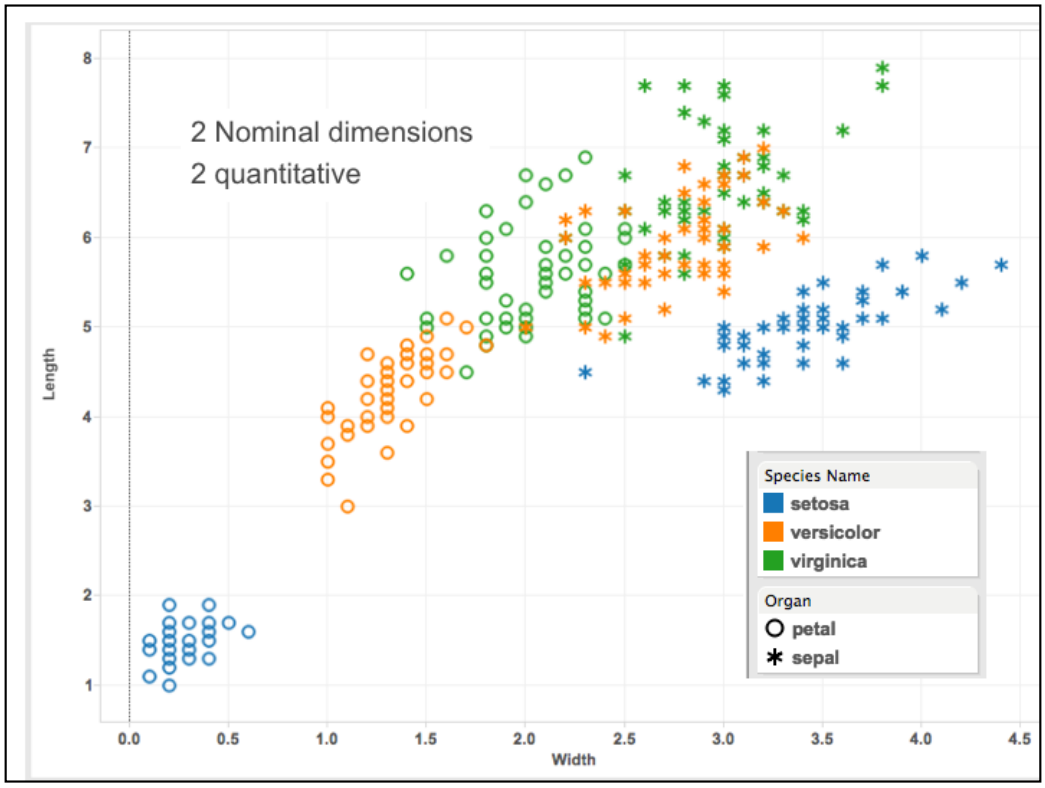
But when I add a 3rd dimension things get gnarly. I have now used colour to show the quantity of the third dimension, moved my categiry coding to shape, and what I have now lost is the easy way to extract the patterns from the plot



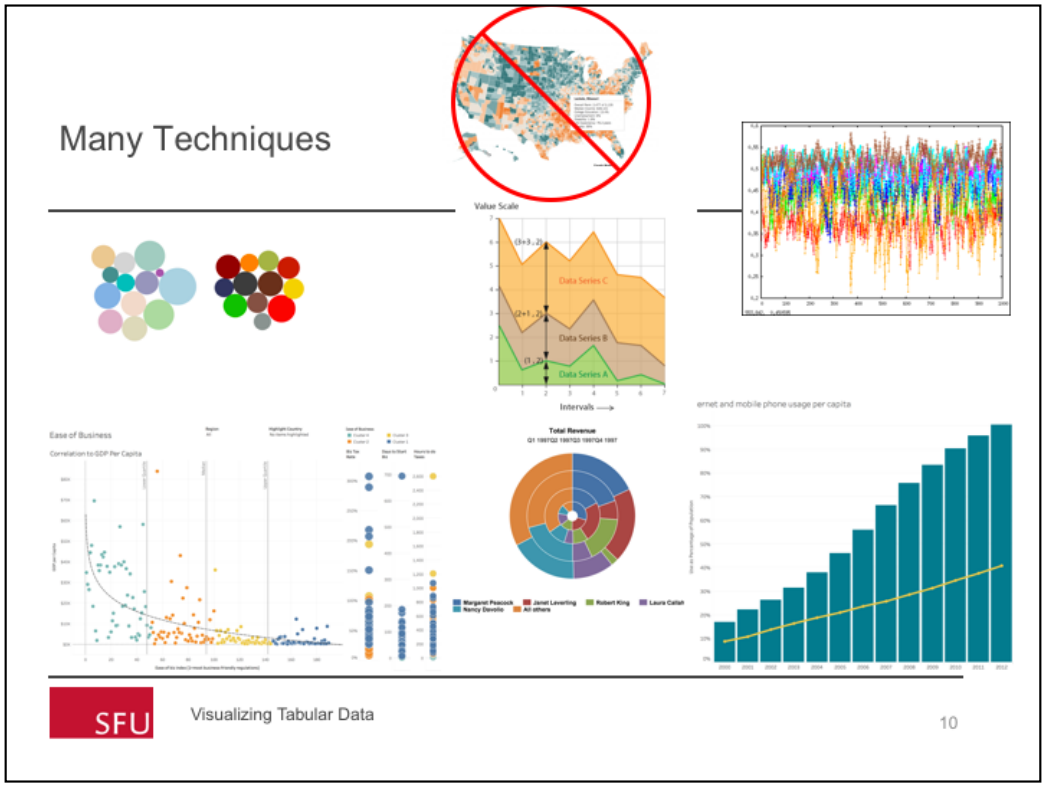
Thre are options: I can splti the views into a 2x2 matrix where I plot all combinations of the dimensions and this sort of works but I miss the overall patterns



I can make my life easier by restructuring the data to move the classification of organ into its own dimension with two levels, giving me another categorical dimension and leaving me two quanititative values.



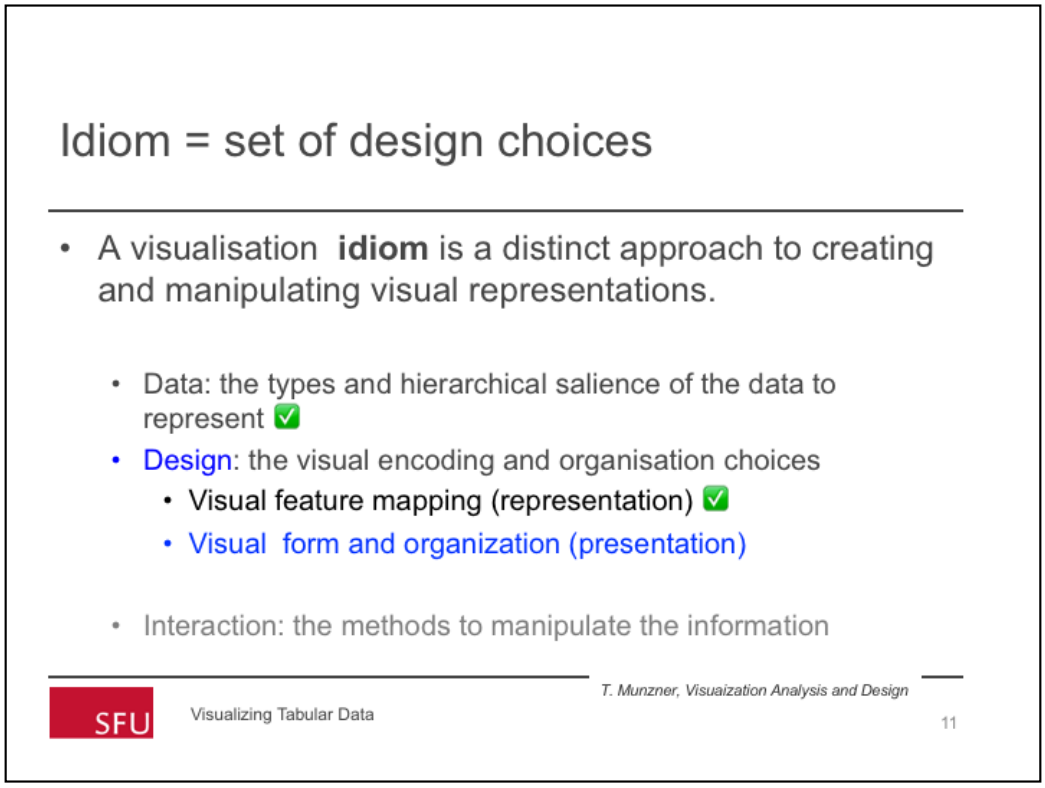
And now I can start to see patterns. What kinds of questions do these data afford? What can I ask?



What alternatives are there?

More generally, what kinds of presentation techniques are best for what kinds of problems?

How do we make design choices?



There are many possible visualization techniques, but many are based on a smaller set of organizational principles that guide the choice of design idiom. Formally, we say a visualisation **idiom** is a set of design choices that make up a distinct approach to a visual representation (form).

. Today we are talking about the choice of presentation idioms, visual forms that structure and organize the data in ways that support different questions and insights.

Recall: Data Abstraction

Row/column = **key**

Tables

- Data item (row)
- attributes (columns)

- Data queries are **Key-value** selections

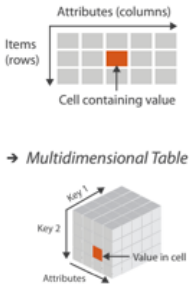
Cell = **value**

ID	Species No	Organ	Length	Width	Species Nam
1	1.1	sepal	5.1	3.5	setosa
2	1.1	sepal	4.9	3	setosa
3	1.1	sepal	4.7	3.2	setosa
4	1.1	sepal	4.6	3.1	setosa
5	1.1	sepal	5	3.6	setosa
6	1.1	sepal	5.4	3.9	setosa
7	1.1	sepal	4.6	3.4	setosa
8	1.1	sepal	5	3.4	setosa
9	1.1	sepal	4.4	2.9	setosa
10	1.1	sepal	4.9	3.1	setosa
11	1.1	sepal	5.4	3.7	setosa
12	1.1	sepal		3.4	setosa
13	1.1	sepal		3	setosa
14	1.1	sepal		3	setosa
15	1.1	sepal	5.8	4	setosa
16	1.1	sepal	5.7	4.4	setosa
17	1.1	sepal	5.4	3.9	setosa
18	1.1	sepal	5.1	3.5	setosa
19	1.1	sepal	5.7	3.8	setosa
20	1.1	sepal	5.1	3.8	setosa
21	1.1	sepal	5.4	3.4	setosa
22	1.1	sepal	5.1	3.7	setosa
23	1.1	sepal	4.6	3.6	setosa
24	1.1	sepal	5.1	3.3	setosa
25	1.1	sepal	4.8	3.4	setosa

Keys and values

Tabular data are Key→Value vectors.

- Key/Attribute: property of the data that can be used to index into (sort by/look up) the set. (independent variable)
  - N, O
- Value : the actual value of an individual item
  - N,O,Q



Key

independent attribute

used as unique index to look up items

simple tables: 1 key

multidimensional tables: multiple keys

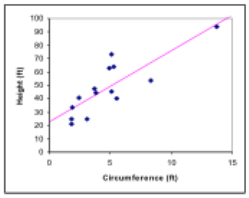
value

dependent attribute, value of cell

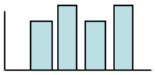
(But how might we represent N and O as values?)

1. Key-value question suggests **idiom**

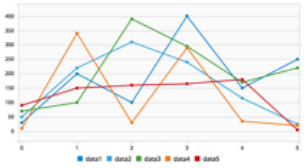
- 2 values

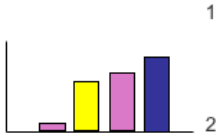


- 1 Key and 1 value



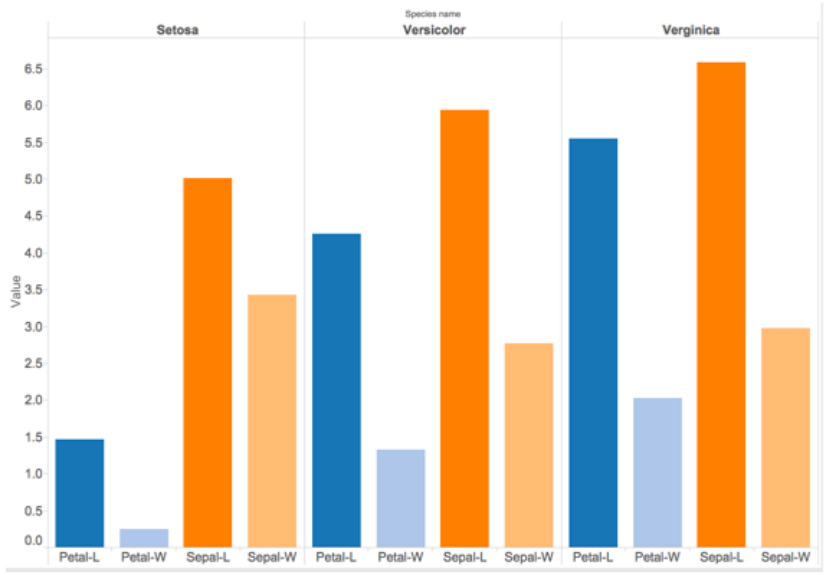
- 2 Keys and 1 value



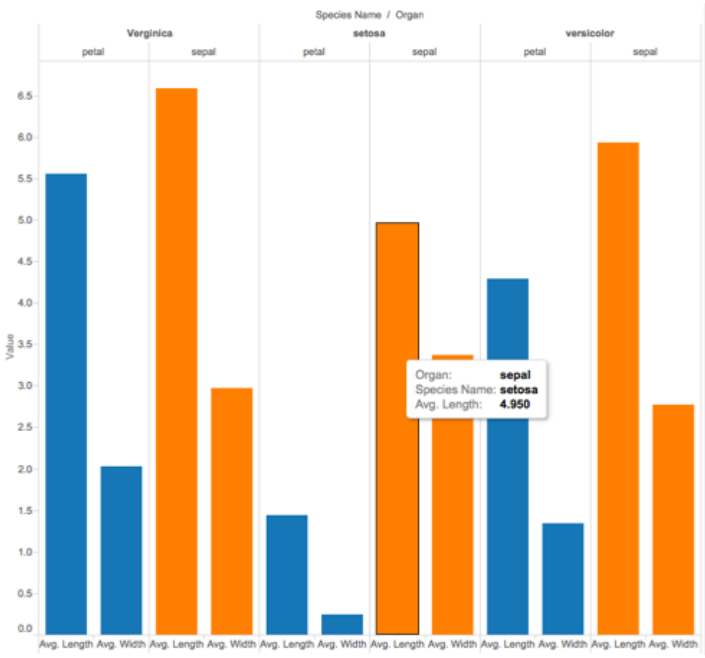


A visualisation **idiom** is a set of design choices that make up a distinct approach to a visual representation (form).

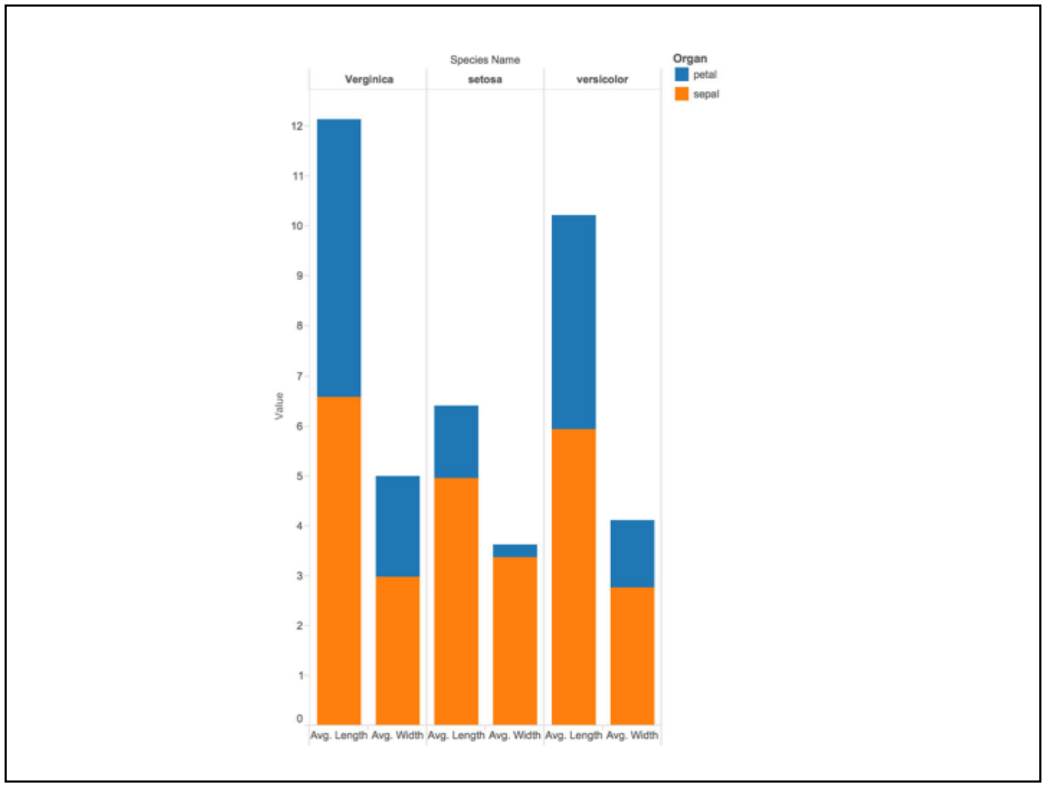
Keys, values and structure

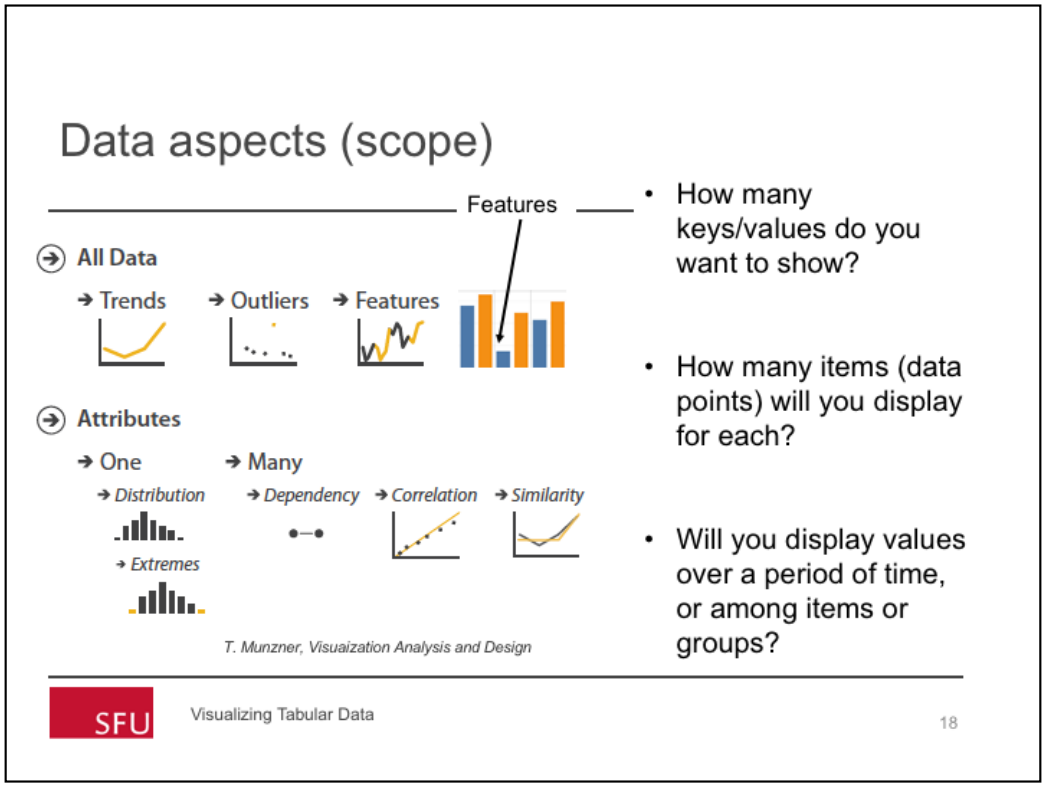


What is this vis good for? Not so good for?


Species Name / Organ
Verginica
setosa
versicolor
petal
sepal
petal
sepal
petal
sepal
Value
6.5
6.0
5.5
5.0
4.5
4.0
3.5
3.0
2.5
2.0
1.5
1.0
0.5
0.0
Organ: sepal
Species Name: setosa
Avg. Length: 4.950
Avg. Length
Avg. Width
Avg. Length
Avg. Width
Avg. Length
Avg. Width
Avg. Length
Avg. Width
Avg. Length
Avg. Width
Avg. Length
Avg. Width

16


Species Name
Verginica
setosa
versicolor
Organ
petal
sepal
Value
0
1
2
3
4
5
6
7
8
9
10
11
12
Avg. Length
Avg. Width
Avg. Length
Avg. Width
Avg. Length
Avg. Width



How many variables do you want to show in a single chart? One, two, three, many?

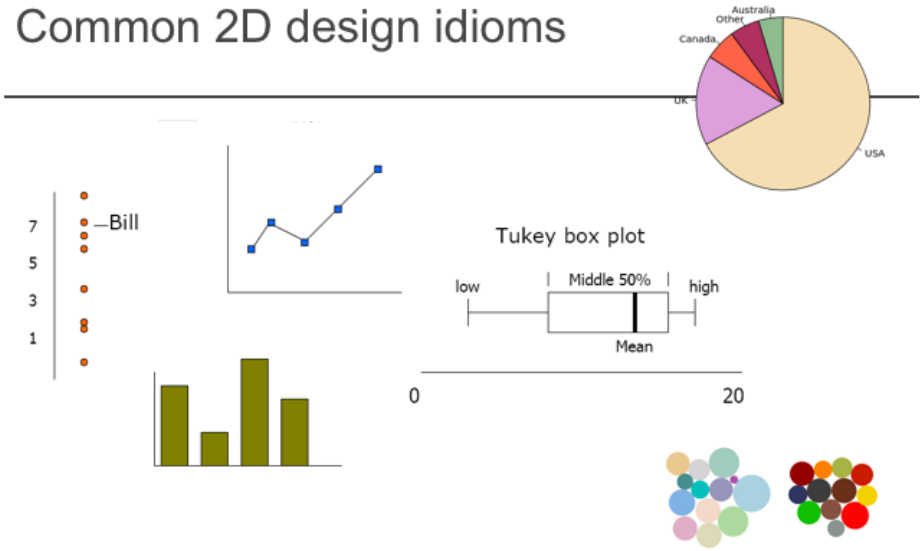
How many items (data points) will you display for each variable? Only a few or many?

Will you display values over a period of time, or among items or groups?

If all you want is a single precise value ….

Year	Jan	Feb	Mar	Apr	May	Jun	Jul	Aug	Sep	Oct	Nov	Dec	Annual
1990	127.4	128.0	128.7	128.9	129.2	129.9	130.4	131.6	132.7	133.5	133.8	133.8	130.7
1991	134.6	134.8	135.0	135.2	135.6	136.0	136.2	136.6	137.2	137.4	137.8	137.9	136.2
1992	138.1	138.6	139.3	139.5	139.7	140.2	140.5	140.9	141.3	141.8	142.0	141.9	140.3
1993	142.6	143.1	143.6	144.0	144.2	144.4	144.4	144.8	145.1	145.7	145.8	145.8	144.5
1994	146.2	146.7	147.2	147.4	147.5	148.0	148.4	149.0	149.4	149.5	149.7	149.7	148.2
1995	150.3	150.9	151.4	151.9	152.2	152.5	152.5	152.9	153.2	153.7	153.6	153.5	152.4
1996	154.4	154.9	155.7	156.3	156.6	156.7	157.0	157.3	157.8	158.3	158.6	158.6	156.9
1997	159.1	159.6	160.0	160.2	160.1	160.3	160.5	160.8	161.2	161.6	161.5	161.3	160.5
1998	161.6	161.9	162.2	162.5	162.8	163.0	163.2	163.4	163.6	164.0	164.0	163.9	163.0
1999	164.3	164.5	165.0	166.2	166.2	166.2	166.7	167.1	167.9	168.2	168.3	168.3	166.6
2000	168.8	169.8	171.2	171.3	171.5	172.4	172.8	172.8	173.7	174.0	174.1	174.0	172.2
2001	175.1	175.8	176.2	176.9	177.7	178.0	177.5	177.5	178.3	177.7	177.4	176.7	177.1
2002	177.1	177.8	178.8	179.8	179.8	179.9	180.1	180.7	181.0	181.3	181.3	180.9	179.9

Common 2D design idioms



2. Task emphasizes idiom choice

- Comparison
- Composition
- Distribution
- Relationship

1. Time Series
   - Quantitative across equal intervals (of time)
2. Ranking
   - Sequenced by attribute value
3. Relative value
   - Differ from reference (baseline or other element)

Adapted from S. Few, Effectively Communicating Numbers:Selecting the Best Means and Manner of Display

2. Task emphasizes idiom choice

- Comparison
- Composition
- Distribution
- Relationship

4. Part-Whole
   - Ratio
   - Components of components
5. Find value
   - Simple descriptive stats of categories

- Static or dynamic

Adapted from S. Few, Effectively Communicating Numbers:Selecting the Best Means and Manner of Display

2. Task emphasizes idiom choice

- Comparison
- Composition
- Distribution
- Relationship

6. Distribution:
   - Clusters
   - Outliers
   - Deviations
   - Modes
7. Deviation
   - Differ from centrality reference (mean)

Adapted from S. Few, Effectively Communicating Numbers:Selecting the Best Means and Manner of Display

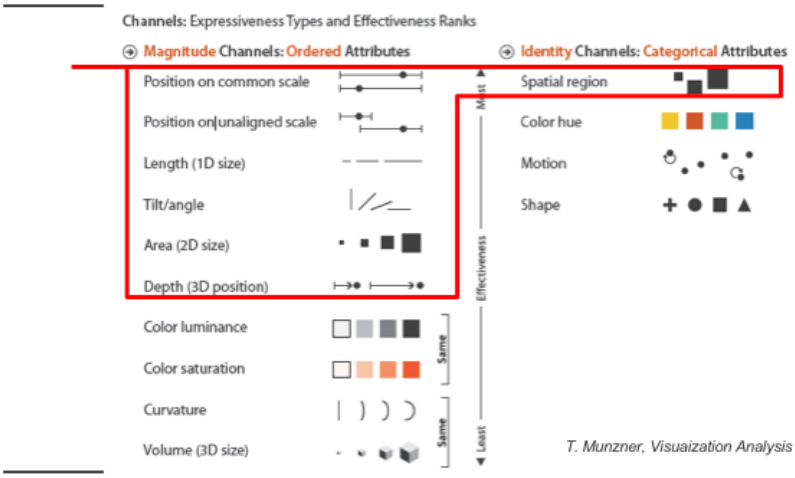
2. Task emphasizes idiom choice

- Comparison
- Composition
- Distribution
- **Relationship**

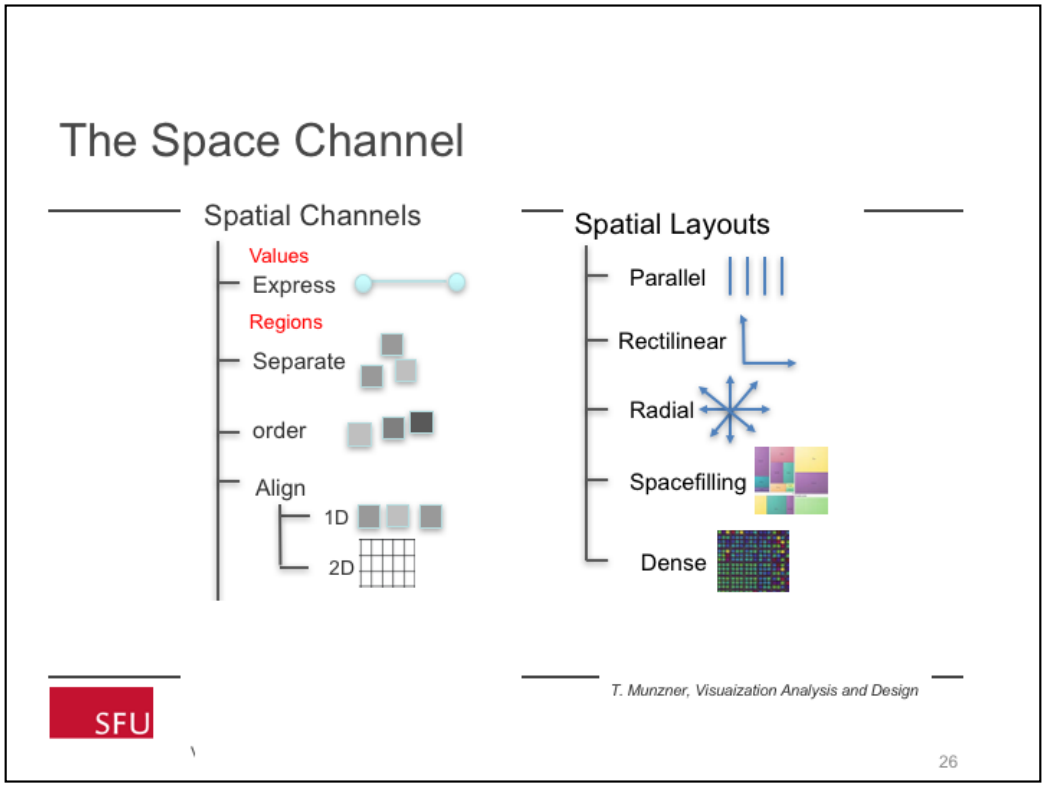
8. Trends:
   - Time series
9. Correlations
   - How one value affects another
10. Dependency and causality

Adapted from S. Few, Effectively Communicating Numbers:Selecting the Best Means and Manner of Display

Spatial organization is the basis of all vis idioms



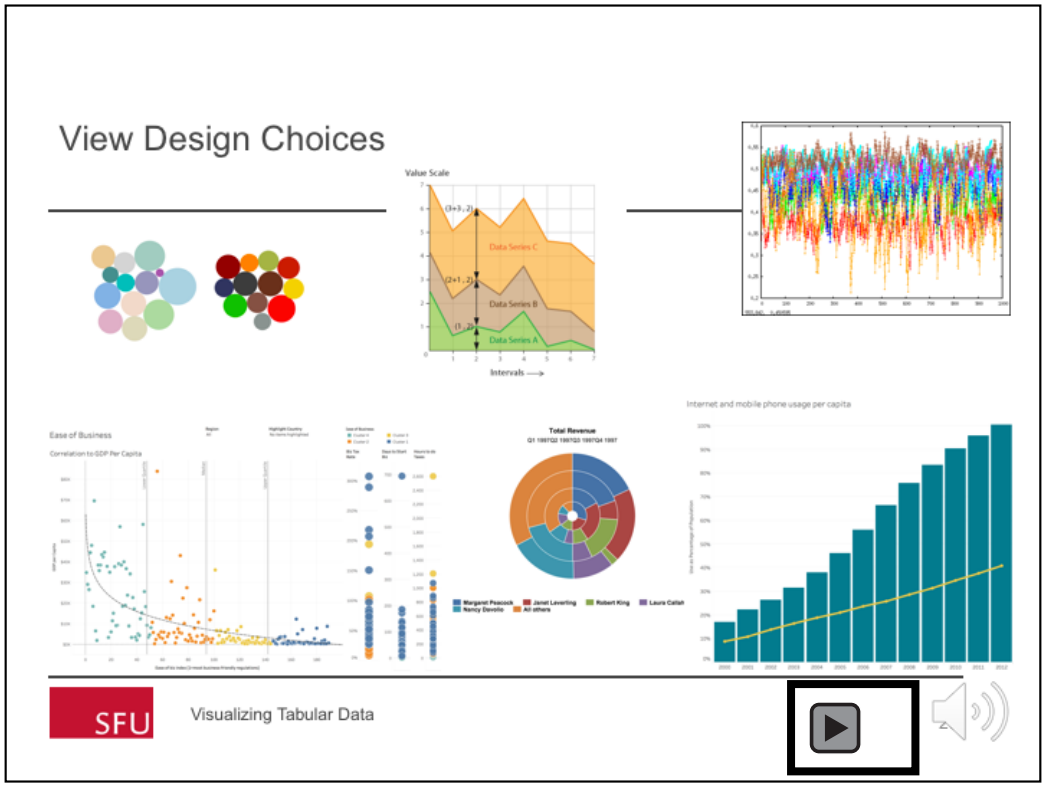
Integrality and seprability



Express (show) values

Arrange data groupings

- Separate /distinguish regions by categorical key
- Order groups by ordinal key
- Align for visual comparison along reference value



What alternatives are there?

More generally, what kinds of techniques are best for what kinds of problems?

All information integrated in one view

basic visual encodings

- spatial position
- color
- other channels
- pixel-oriented techniques

visual layering

- global compositing
- item-level stacking

glyphs

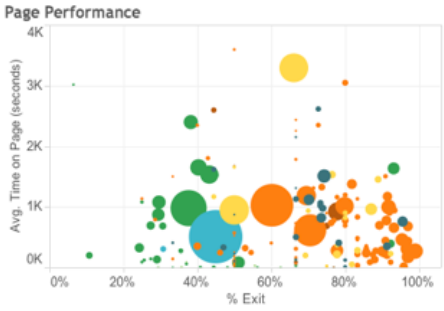
Rectilinear

Scatter plot

- Axes encode values
- Emphasizes how 2 values relate to each other

- Point marks (hundreds)
- Color for key/category
- Additional value in size, shape ..

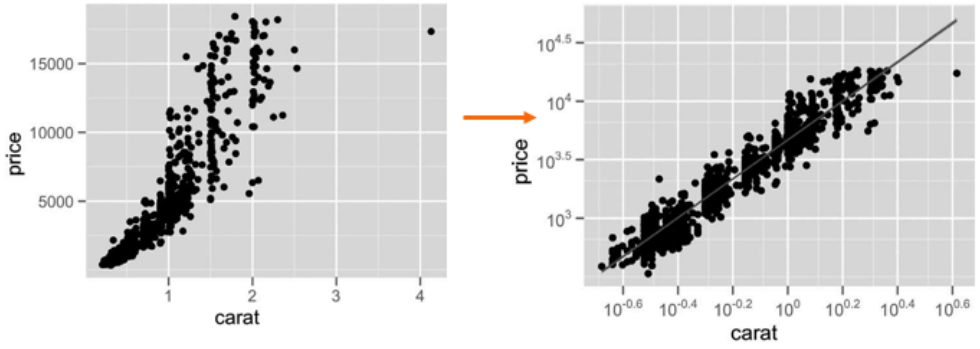
- Relation
- distribution



Scatter plots express the relationship of two continuous values

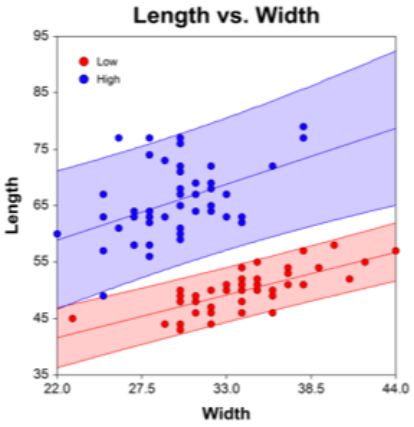
Rectilinear

Data transformations can enhance clarity, reveal patterns



Log transformations show strong correlation between size and price

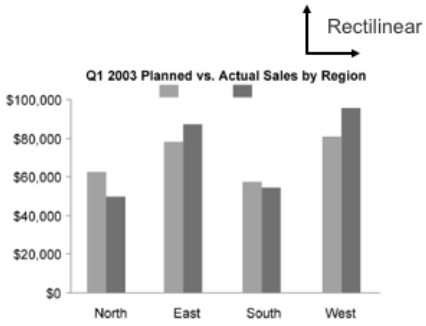
Why (not) use a scatter plot ?

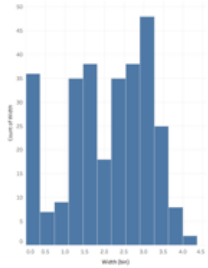


- Excellent for **clusters** and **relations**
- See all the data
- can be difficult to understand
- (Can only) model relations between 2 quantitative variables
- occlusion

Bar (column) chart

- List alignment of keys
- Can be ordered along the list axis or by value
- Nominal and ordinal (if axis ordered)
- Emphasizes individual values and relative differences
  - **Comparison**
  - Composition
  - Distribution (1 attribute histogram)



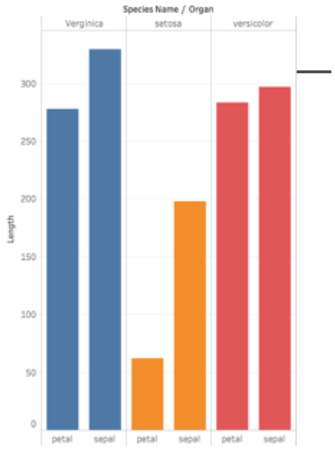


Rectilinear

Bar chart idiom

- Categorical attributes match well with spatial **regions**
- **Separate, order, align**
- Can be hard to find patterns in the data “shape”
- Dozens to hundred items

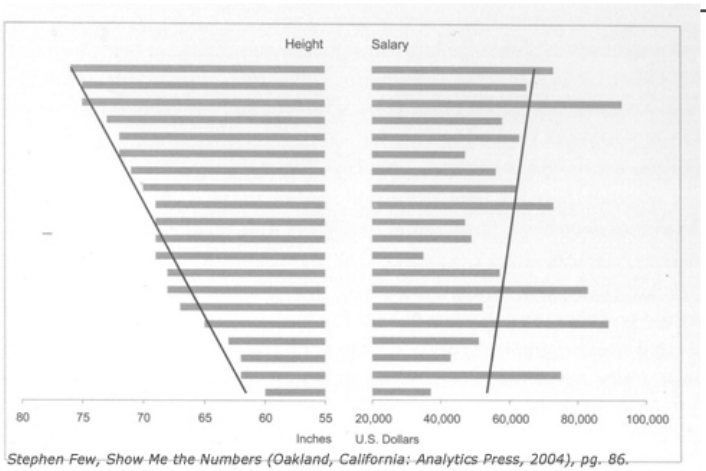
- Usually used with aggregates



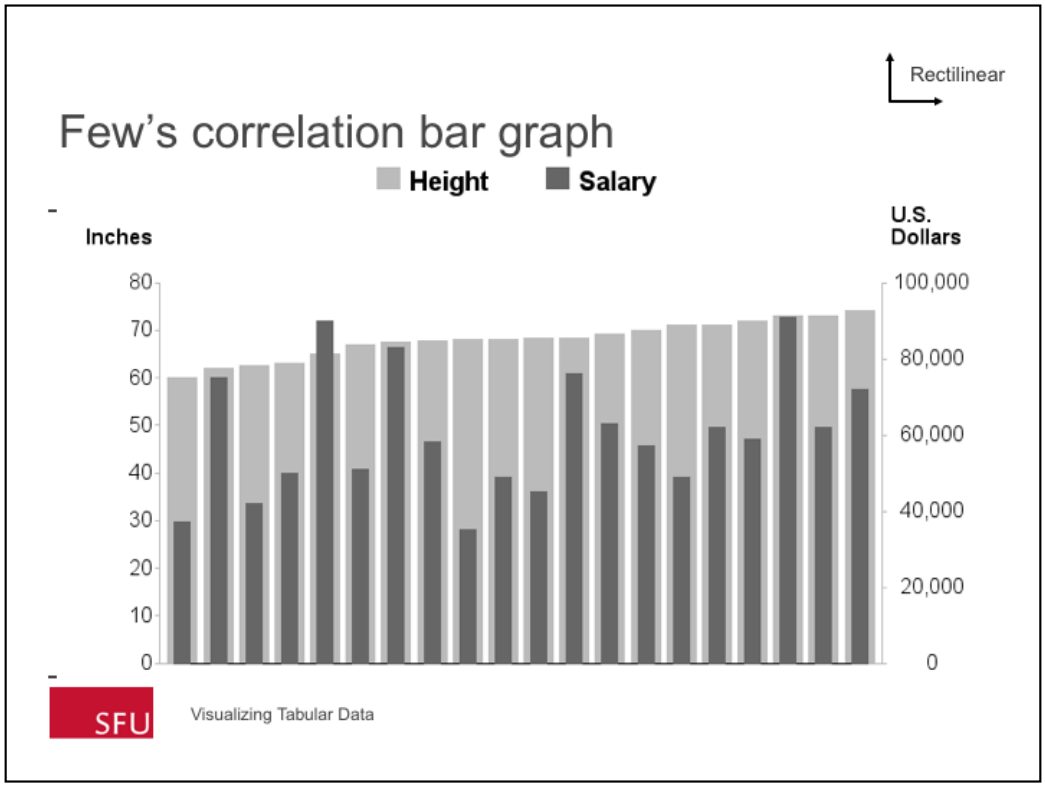
Credit: T. Munzner, 2014

Rectilinear

Paired Bar graph with trend lines (Few)



Stephen Few, Show Me the Numbers (Oakland, California: Analytics Press, 2004), pg. 86.



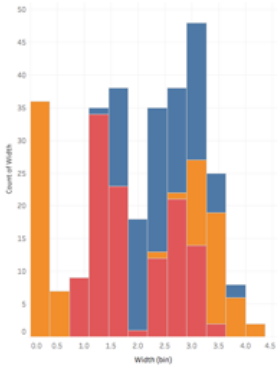
This is a form of bullet graph, in which one value is displayed within another for context.

Because these are two sets of values, this can require two concurrent value axes. This can take a while to interpret, and can easily become confusing.

Rectilinear

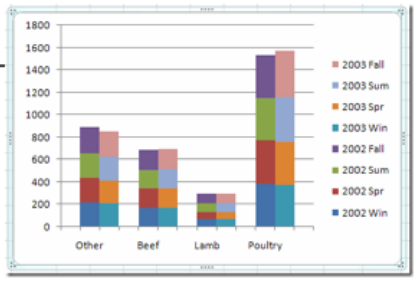
Stacked bar chart

- Barchart with 2 keys/attributes, 1 value
  - Typically use colour or texture
  - <= 10 attributes for 2nd key
- Emphasize **part-whole** level of detail
  - Most important key gets axis alignment
  - Second key gets non-alignment
- Composition
- Comparison
  - Hard to calculate/compare sub-categorical values



Rectilinear

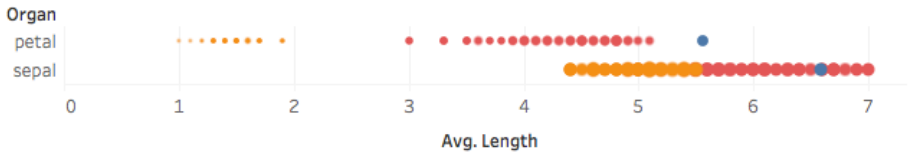
Stacked bars grouped



- Barchart with 3 keys/attributes, 1 value
  - Typically use colour or texture
  - <= 10 attributes for 2 key
  - Spatial grouping used for 3 key
- mark: vertical stack of line marks
  - ***glyph***: composite object, internal structure from multiple marks
  - spatial regions: one per glyph
    - aligned: full glyph, lowest bar component
    - unaligned: other bar

Rectilinear

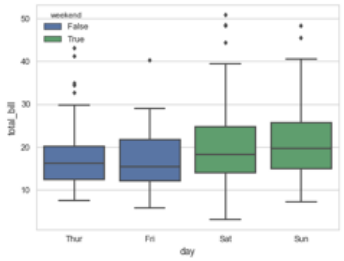
dot plots



- 1 value, 1 key
- Can use 2nd nominal key (color)
- Position to express value according to key 1
- Shows detail of points
- Composition, comparison

Box plot

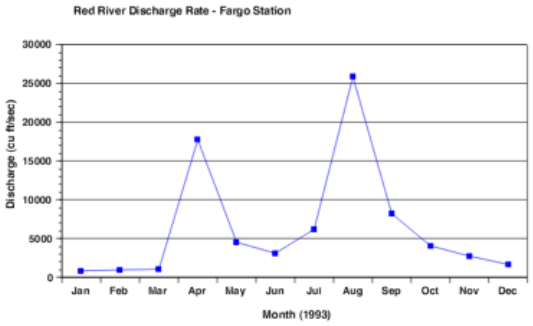
- 1 value as position (length)
- 1 key as list alignment along axis
  - multiple boxes
- Optional 2nd key (color)
- Shows spread of all data
  - “infinite values” but not individual items
  - > few dozen boxes (key)
- Distribution



Line charts

→ Rectilinear

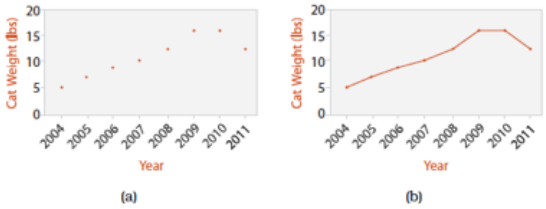
- Position to express value according to key
- Line charts use angle/shape to show trends
  - Frequently time



Rectilinear

Line Chart

- Line charts, dot chart
- Good for ordered data
- Hundreds of points

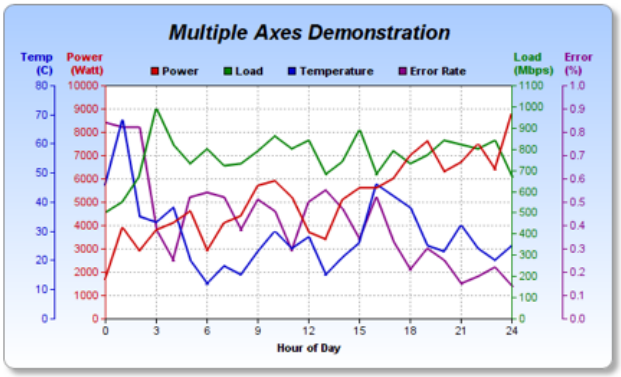


- Key must be ordered and interval

- Composition (find trend)
- Comparison

Rectilinear

Line charts: multiple axes

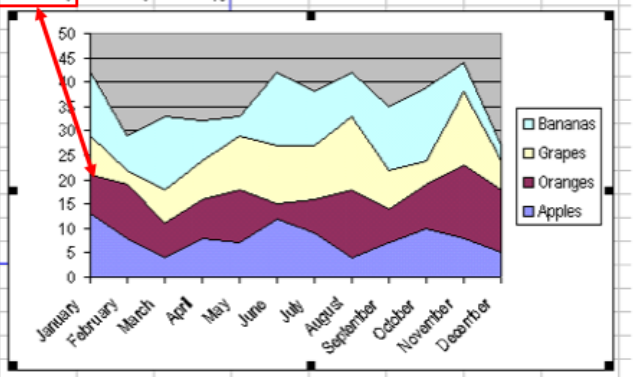


- 2nd key : axis colour
- Add value: second axis
- additional superimposed chart

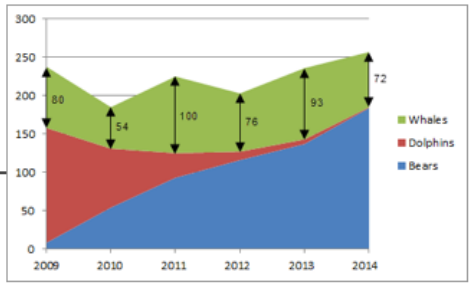
Rectilinear

Stacked area charts

	Apples	Oranges	Grapes	Bananas
January	13	8	8	13
February	8			
March	4			
April	8			
May	7			
June	12			
July	9			
August	4			
September	7			
October	10			
November	8			
December	5			



Stacked area charts



- 1 value (length)
- 1 ordered key (spatial region aligned along axis)
- 2 or more categorical key (color)

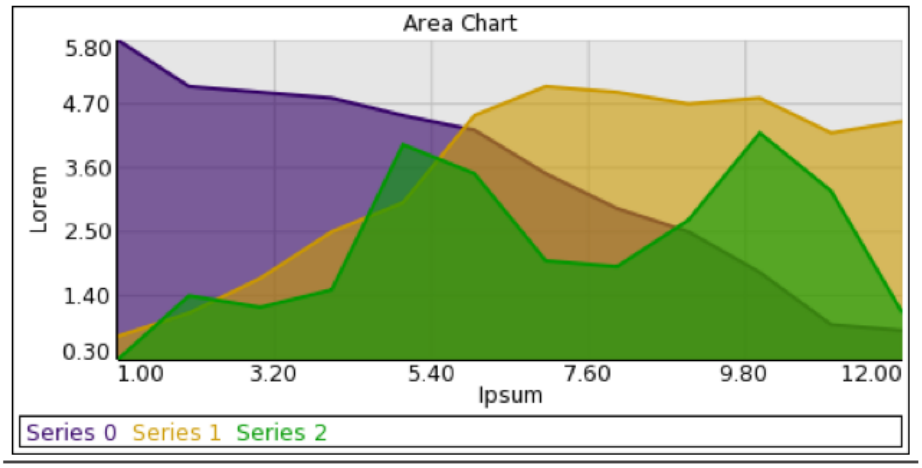
- Scale
  - Ordered key: dozens to hundreds
  - Stacked axis: several

- Composition:
  - part-whole, find value,

- Relationship
  - see trends

Rectilinear

Overlaid area charts



SFU Visualizing Tabular Data 44

Whats different here?

Streamgraphs

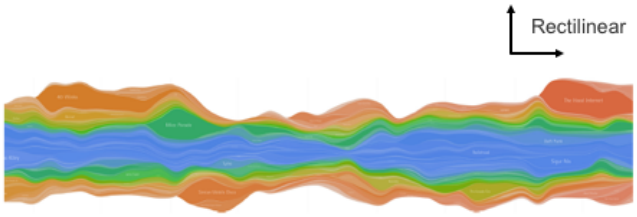
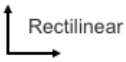


Figure 7.6. Streamgraph of music listening history. From [Byron and Wattenberg 08, Figure 0].

- Stacked time series
- 2 nominal keys, 1 value
- Dozen-hundreds of items along main axis
- Few ~(1-12) stacked glyph levels
-
- Composition (find trends, find value, part-whole)
  - De-emphasizes individual values

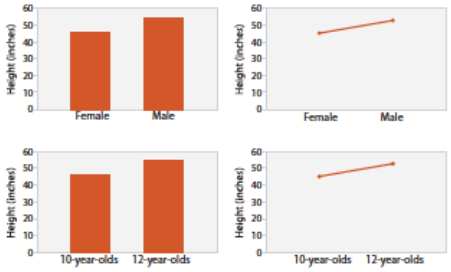


Partial summary

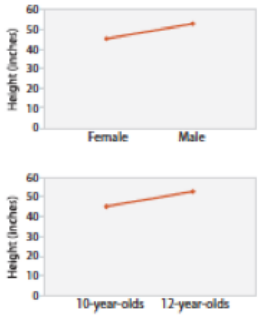
- Bar charts, line charts and dotplots all encode a quantitative value against a key attribute in a rectilinear layout.
  - Often use additional encoding for other keys
- Lines also use connection marks to show inter-item relations
  - Only use for ordered data!

Which to use when

- Bars and bubbles emphasize comparison and association of individual values
- Lines (explicit and implied) emphasize trends



Lines and bars

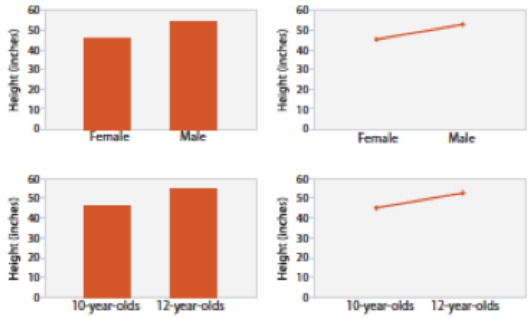


Lines imply connections
Use when there is some ordered progression between the levels on the x-axis

- "12 year olds are taller than 10 year olds"
- NOT
- "the more male someone is the taller he is"

Je Zacks and Barbara Tversky. Bars and Lines:
A Study of Graphic Communication." Memory and Cognition 27:6(1999), 1073{1079

Lines and bars



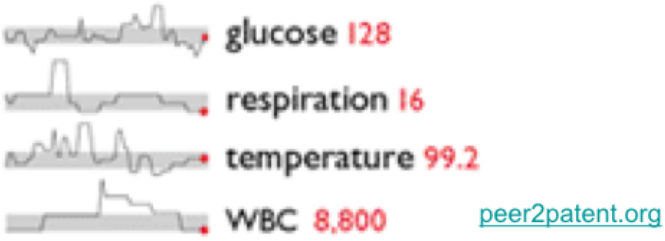
Bars enable comparison between categories

- Males are in average taller than females

Je Zacks and Barbara Tversky. Bars and Lines:
A Study of Graphic Communication." Memory and Cognition 27:6(1999), 1073{1079

Tufte's Sparklines

- Give a hint of the trend, but don't show the actual axes and scales.



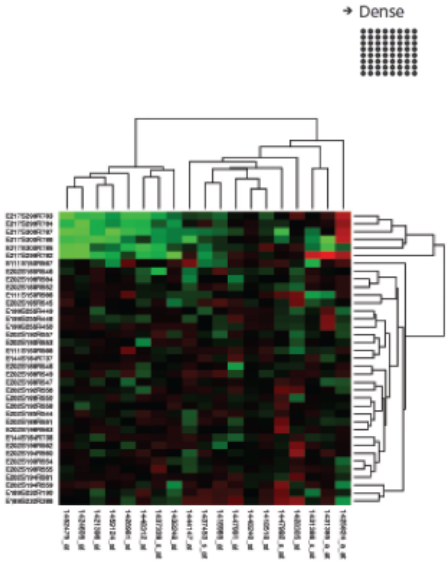
peer2patent.org

- Good for dashboards and small spaces

Pixel-dense displays

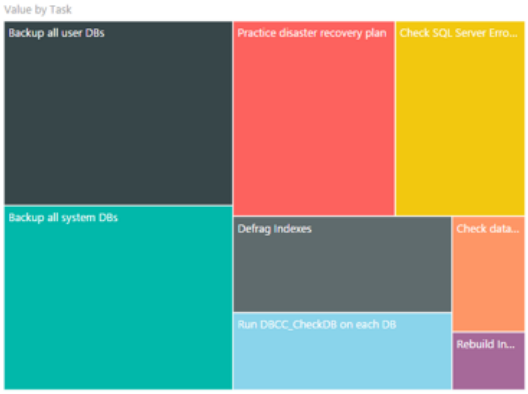
- Matrix Heatmap
- 2 keys, 1 value
  - Good for dense encoding
  - Re-ordering for clusters
- Hard to find ind. values

- Distribution
- Composition



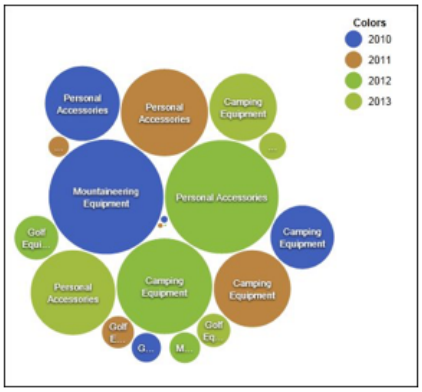
Space filling displays

- Treemap
- 1 key, 1 value
  - Supports hierarchy
- Hard to compare ind. values
- Distribution
- Composition



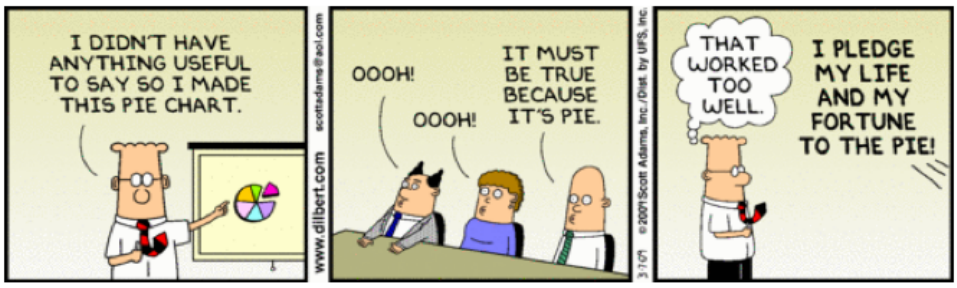
Packed bubbles

- 1 key, 1 value
  - Size = value
  - Colour = key
- Spatial location does **not** hold meaning
- Dense screen packing
- Comparison
- Composition

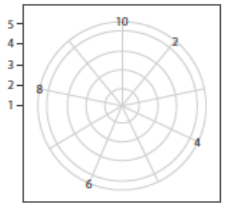


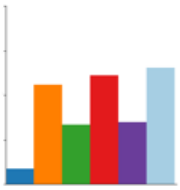
What about Pies?

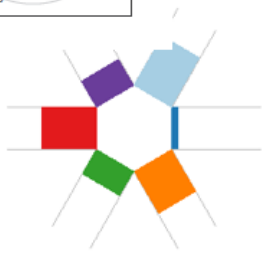
→ Radial



Radial layouts







Use polar coordinates

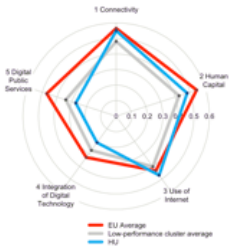
- 1 categorical key, 1 quantitative value

Star charts

→ Radial

- Star plot
  - 1 Key = point on circumference
  - 1 Value = length of mark spoke
  - Can use colour , line thickness for 2nd key

- Radar plot
  - 2 keys, point on circumference, colour
  - 1 value, length of polygon edge



- Composition (find value)
- Comparison (some shape with radar)

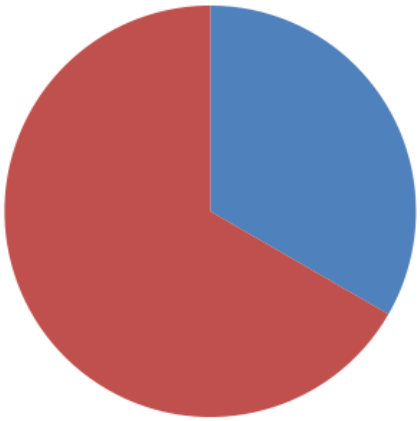
Pie charts

→ Radial

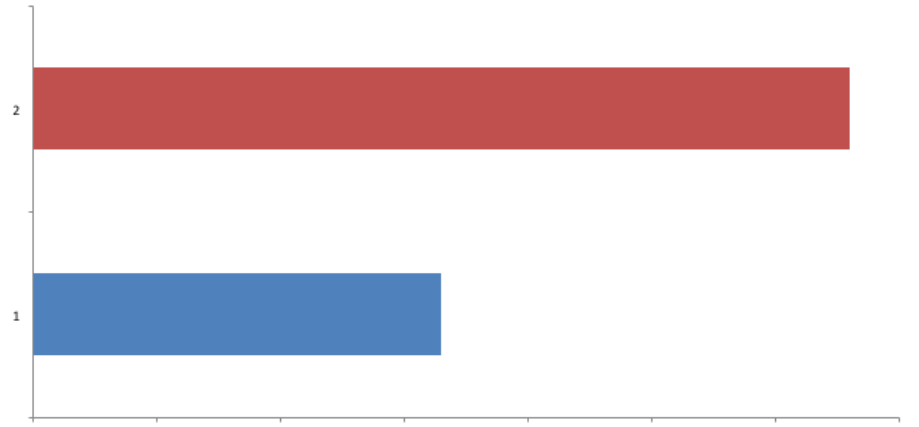
- One of most common forms
- 1 Key = colour
  - 1-7 levels of key (distinct areas)
- 1 Value = angle, area

- Composition – part whole, ratios
- Comparison: partial
  - We misjudge area

Percent Blue relative to Red?

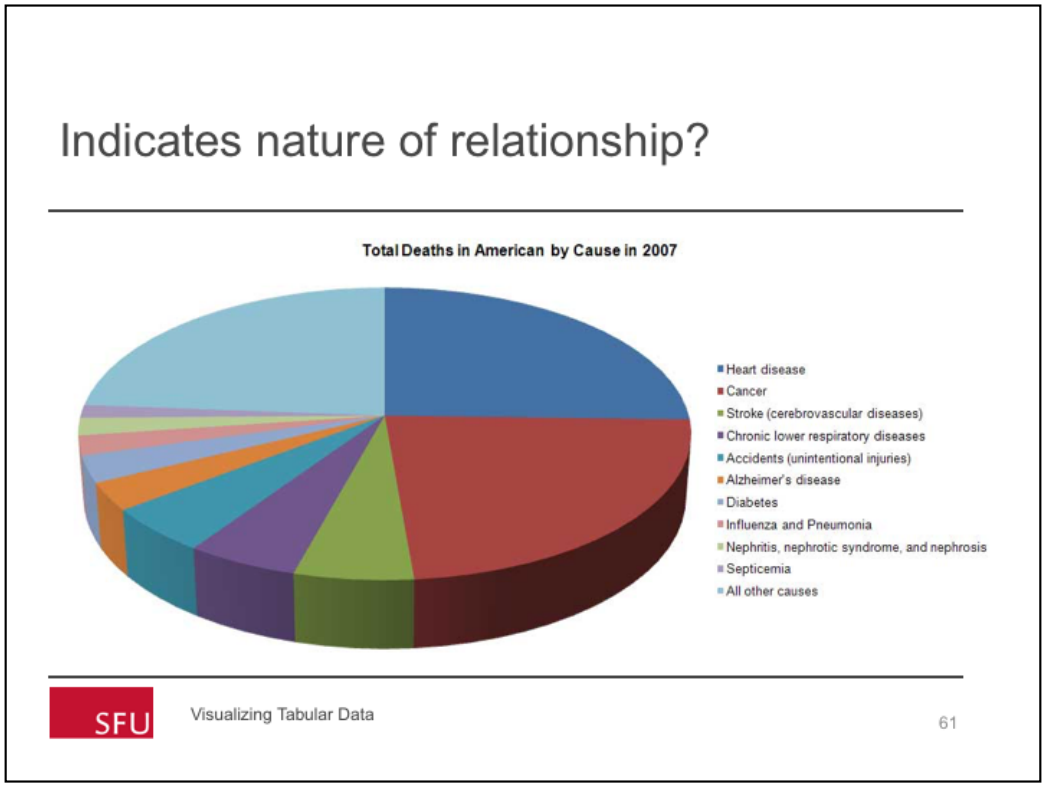


Percent Blue relative to Red?



Few's criteria for an effective visualization

- Clearly indicate the nature of the relationship
- Represent the quantities accurately
- Makes it easy to compare the quantities
- Makes it easy to see the ranked order of values
- Makes obvious how people should use the information



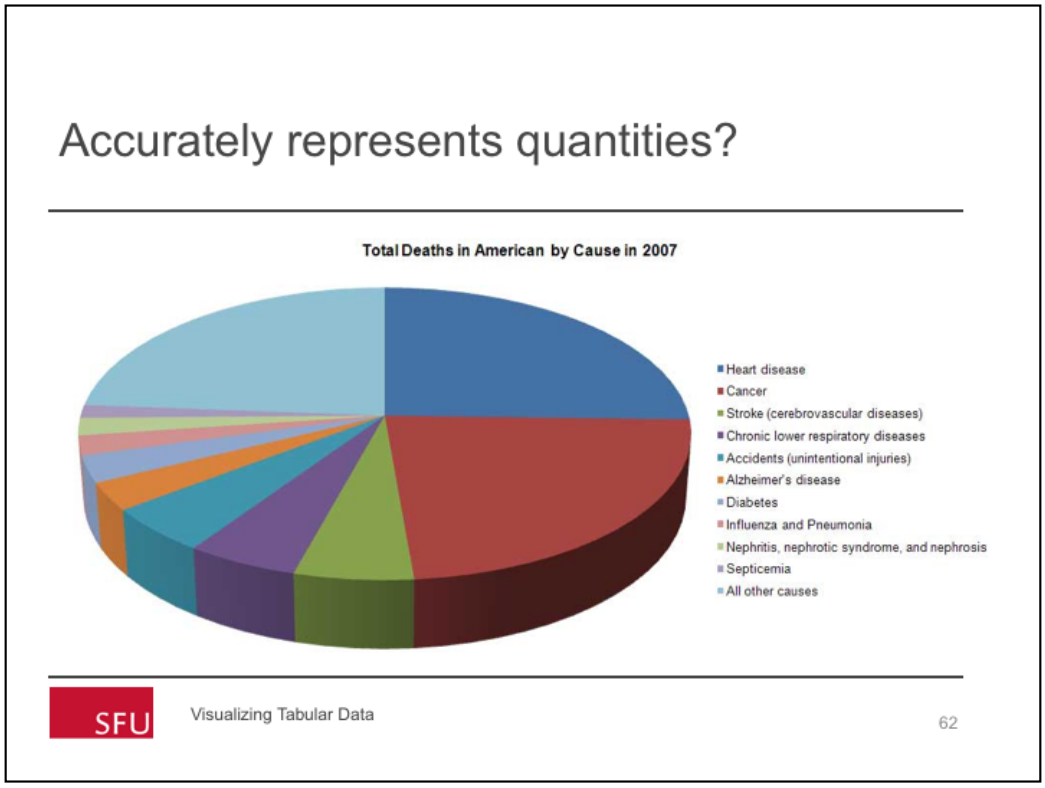
Clearly indicates the nature of the relationship? Yes. The primary strength of a pie chart is the fact that it clearly indicates a part-to-whole relationship between the values.

Represents the quantities accurately? No. Pie charts encode values redundantly through the use of three visual attributes: the area of each slice, the angle formed by each slice at the center of the pie, and the length of the each slice along the pie's perimeter. Even when the area, angle, and perimeter of each slice is calculated properly, it fails in that we cannot perceive any one of these attributes accurately. Visual perception in humans has not evolved to support accurate decoding of areas, angles, or distance along a curve.

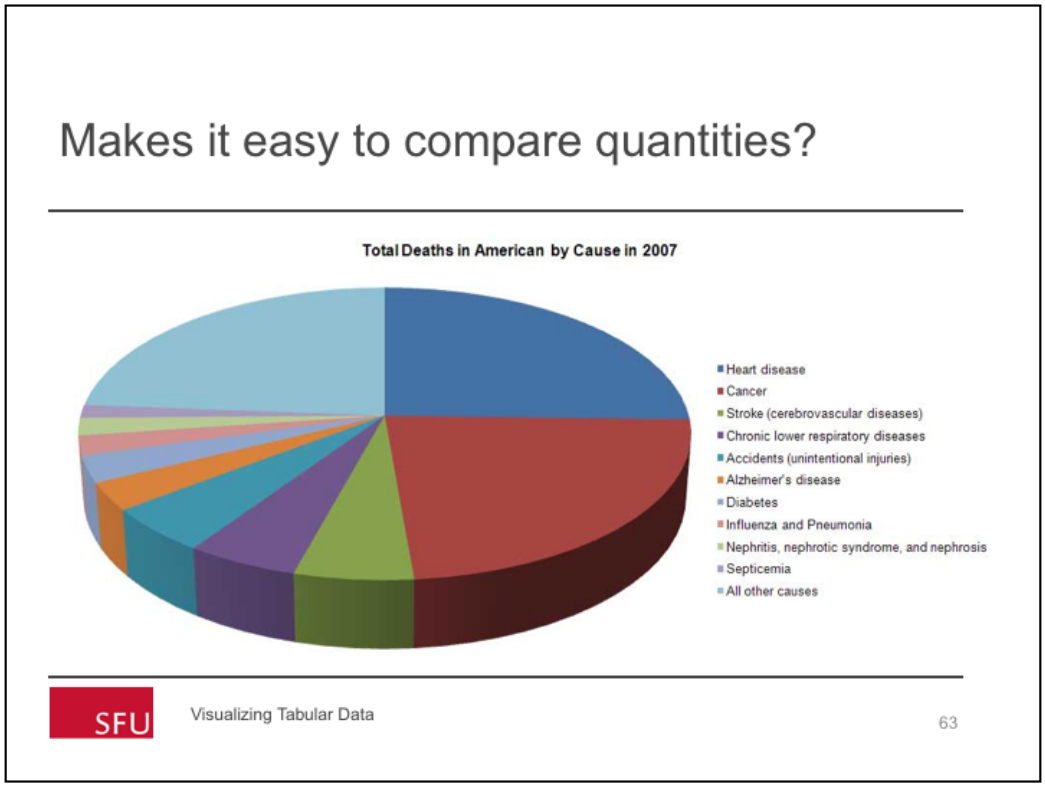
Makes it easy to compare the quantities? No. Because we cannot perceive the values accurately, we also cannot compare them easily or accurately. Furthermore, in this particular pie chart, because a legend has been used to label the slices, we are forced over and over to look up the meaning of the slices we wish to compare by finding the right color, which is often difficult to discriminate. The fact that this pie chart has been rendered in 3-D also complicates the simple act of comparison because the perspective skews the relative size and shape of the slices, making slices on the bottom appear larger and more salient than similarly sized slices on the top.

Makes it easy to see the ranked order of values? No. Even though the slices are displayed in ranked order from the highest value (heart disease) at the top and continuing clockwise to the smallest, excluding the final "All other causes" slice, this ranking isn't obvious, because it's difficult to compare the slices. For example, the red cancer slice appears to be larger than the blue heart disease slice due to the 3-D effect, which has given it more visual weight. Effects such as the 3-D rendering of this pie chart are sometimes used to intentionally mislead.

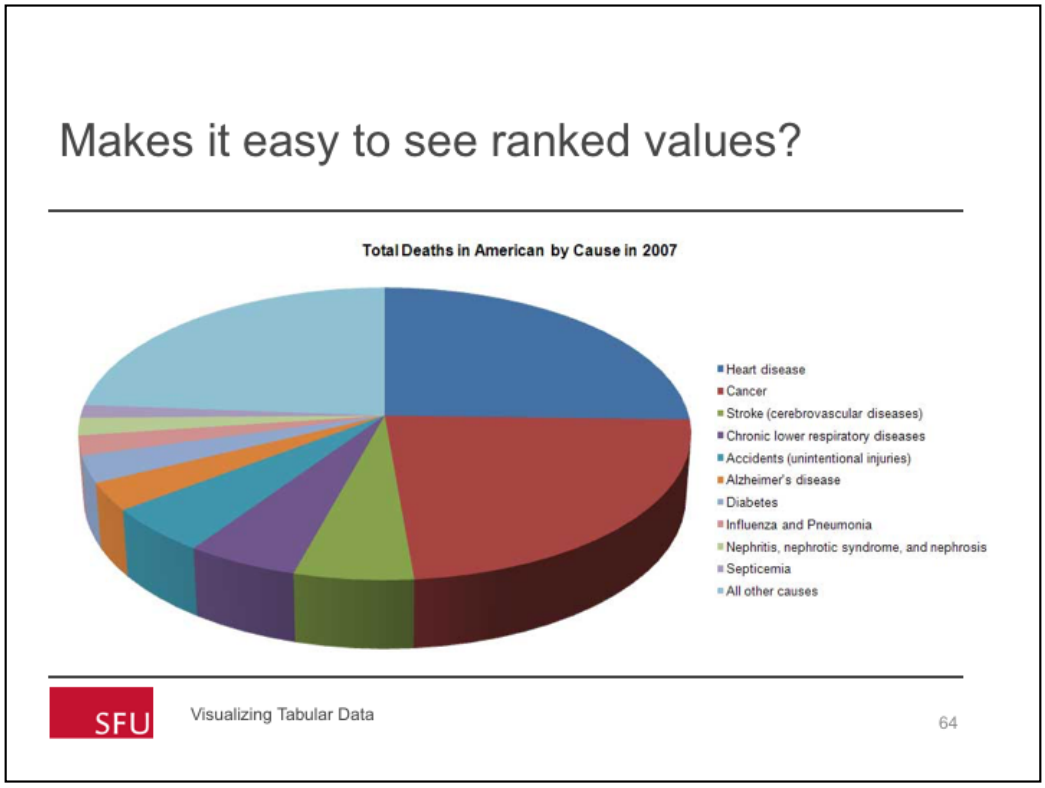
Makes obvious how people should use the information? Partially. Although the pie chart succeeds in encouraging people to compare the slices to understand the relative contributions of each part to the whole, it fails to support this operation effectively.



Represents the quantities accurately? No. Pie charts encode values redundantly through the use of three visual attributes: the area of each slice, the angle formed by each slice at the center of the pie, and the length of the each slice along the pie's perimeter. Even when the area, angle, and perimeter of each slice is calculated properly, it fails in that we cannot perceive any one of these attributes accurately. Visual perception in humans has not evolved to support accurate decoding of areas, angles, or distance along a curve.



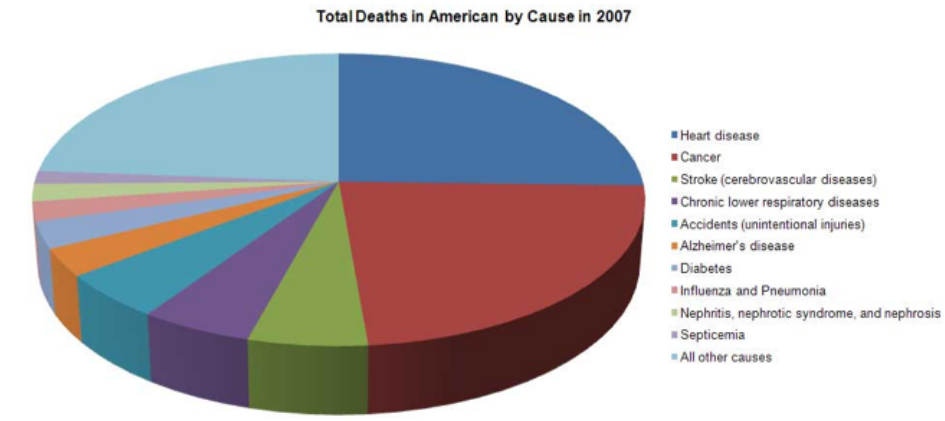
Makes it easy to compare the quantities? No. Because we cannot perceive the values accurately, we also cannot compare them easily or accurately. Furthermore, in this particular pie chart, because a legend has been used to label the slices, we are forced over and over to look up the meaning of the slices we wish to compare by finding the right color, which is often difficult to discriminate. The fact that this pie chart has been rendered in 3-D also complicates the simple act of comparison because the perspective skews the relative size and shape of the slices, making slices on the bottom appear larger and more salient than similarly sized slices on the top.



Makes it easy to see the ranked order of values? No. Even though the slices are displayed in ranked order from the highest value (heart disease) at the top and continuing clockwise to the smallest, excluding the final "All other causes" slice, this ranking isn't obvious, because it's difficult to compare the slices. For example, the red cancer slice appears to be larger than the blue heart disease slice due to the 3-D effect, which has given it more visual weight. Effects such as the 3-D rendering of this pie chart are sometimes used to intentionally mislead.

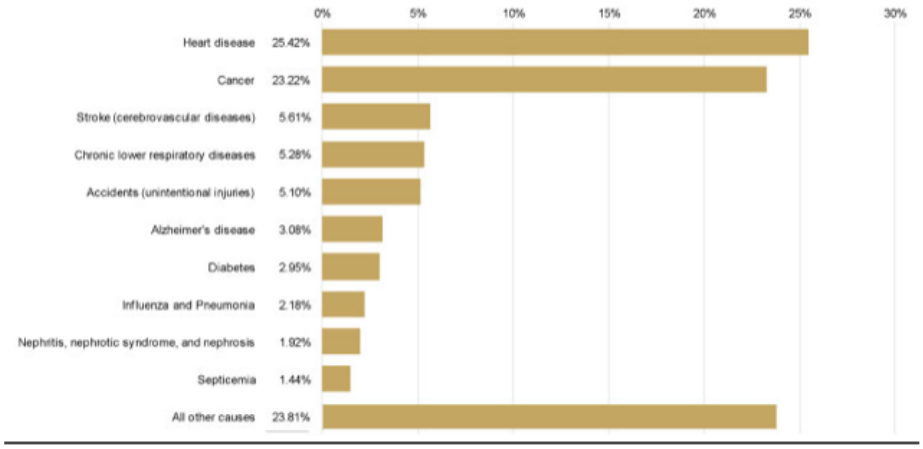
Makes obvious how people should use the information? Partially. Although the pie chart succeeds in encouraging people to compare the slices to understand the relative contributions of each part to the whole, it fails to support this operation effectively.

Makes it easy to see how people should use information?



Makes obvious how people should use the information? Partially. Although the pie chart succeeds in encouraging people to compare the slices to understand the relative contributions of each part to the whole, it fails to support this operation effectively.

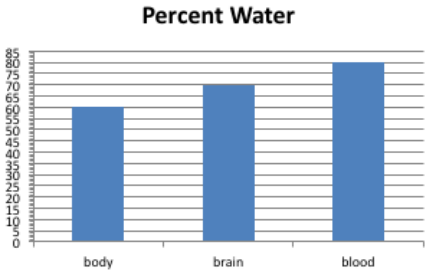
A better way

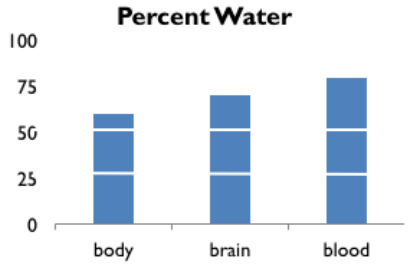


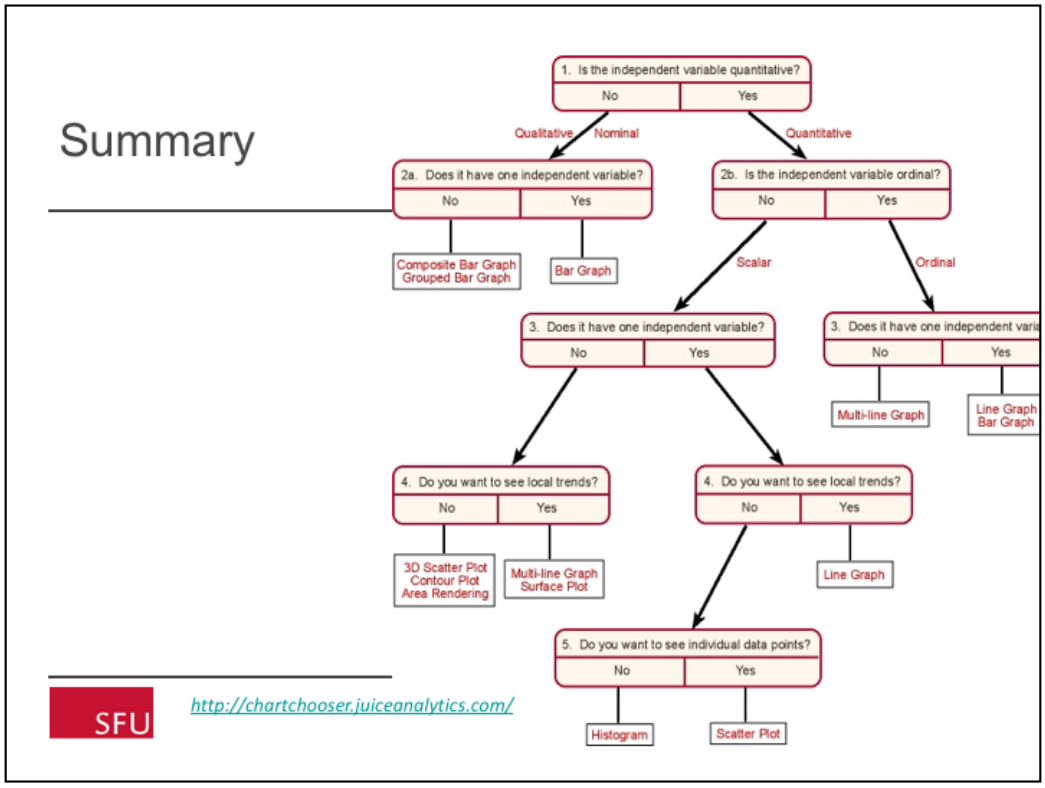
The bar chart is often a better way to show the data. Howeverm if WHAT WE WANT is the part whole percept, this DOES NOT support that insight.

Reference structures [Bartram et al., 2012]

- Visual structure Grids and lines

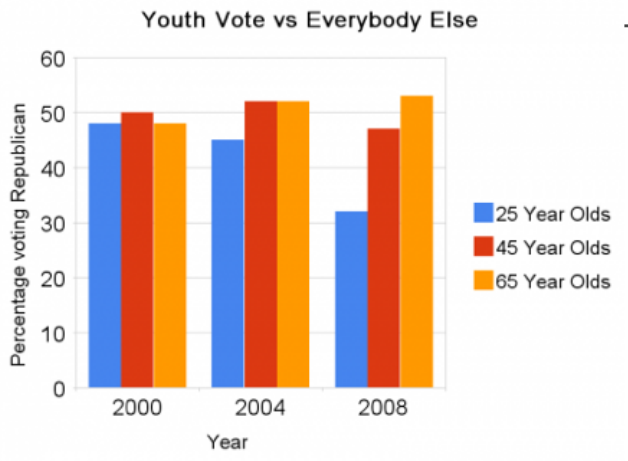


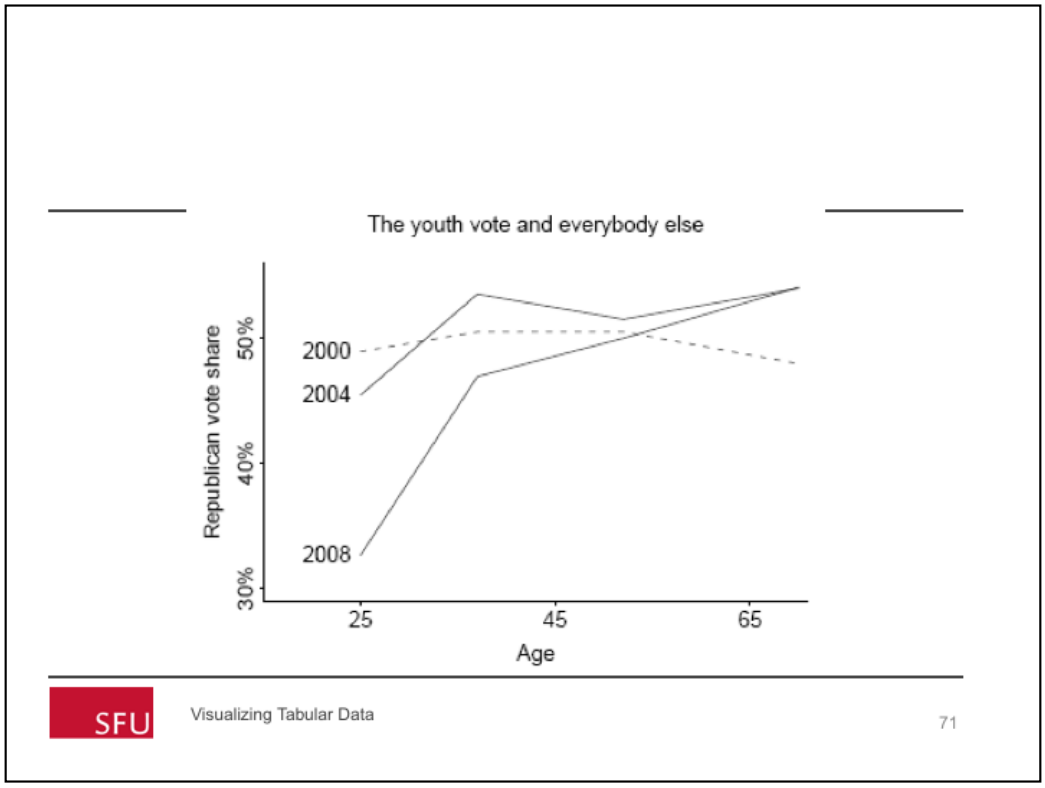




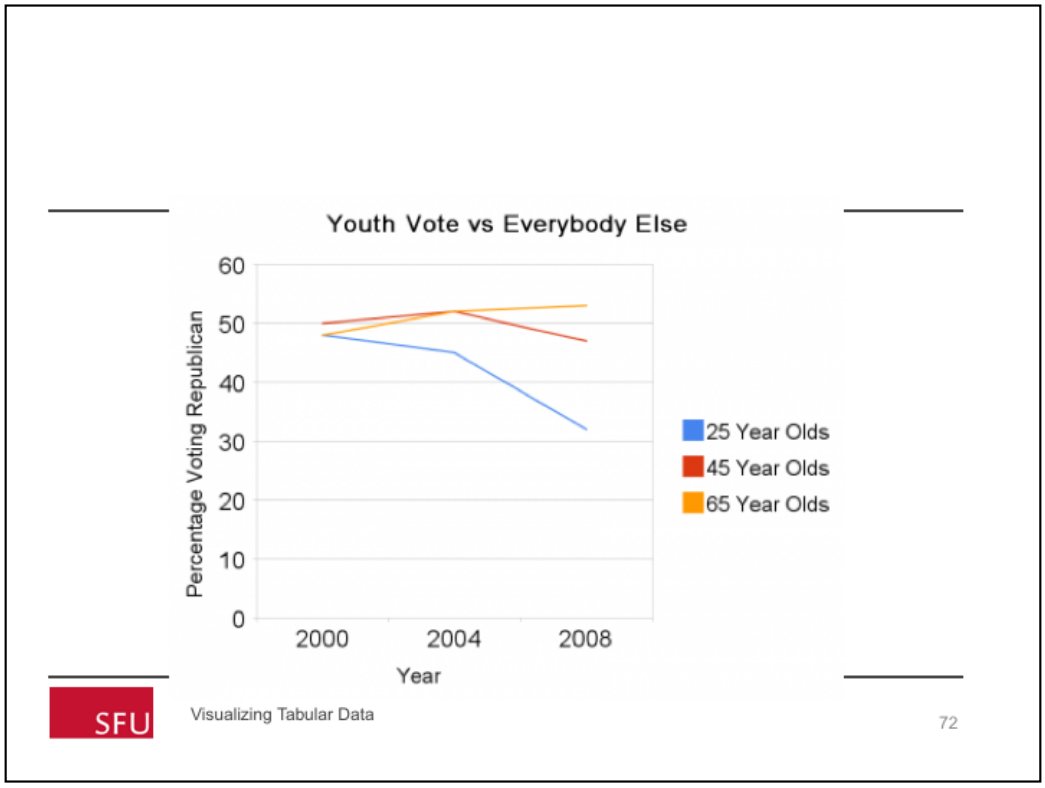
http://chartchooser.juiceanalytics.com/

- Let's look at some examples
- What types of questions do these visualizations best support?

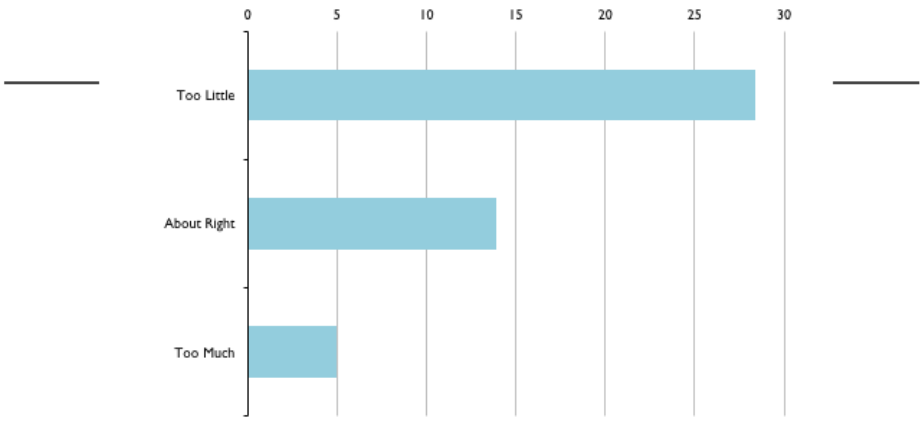

Youth Vote vs Everybody Else
Percentage voting Republican
60
50
40
30
20
10
0
2000
2004
2008
Year
25 Year Olds
45 Year Olds
65 Year Olds



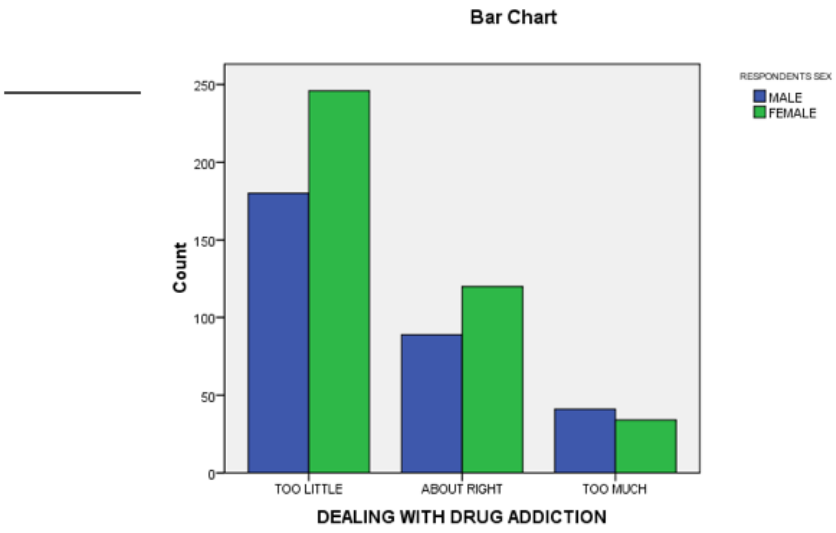
From http://hobershort.wordpress.com/2008/11/06/do-the-numbers/



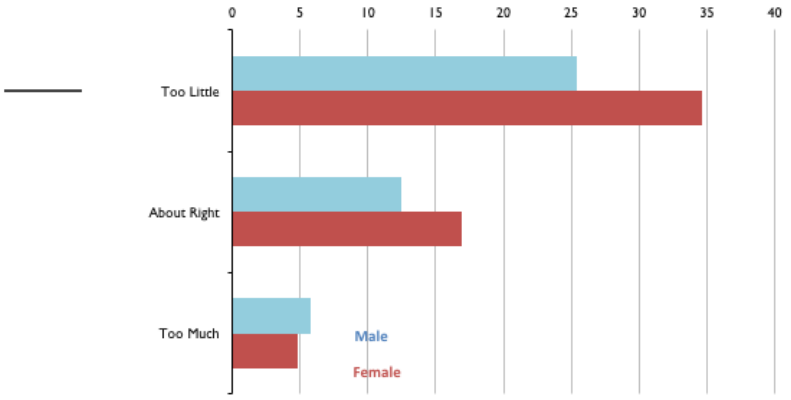
The problem remaining is starting Y axis at 0 compresses the differences. This is good and bad. Its bad because there is too much useless whitespace. Its good because it doesn't distort the data. The other problem is it connects data points across time when in fact there are 4 years intervening and the composition of the groups are different of those time periods as some people move groups, but this is minor.



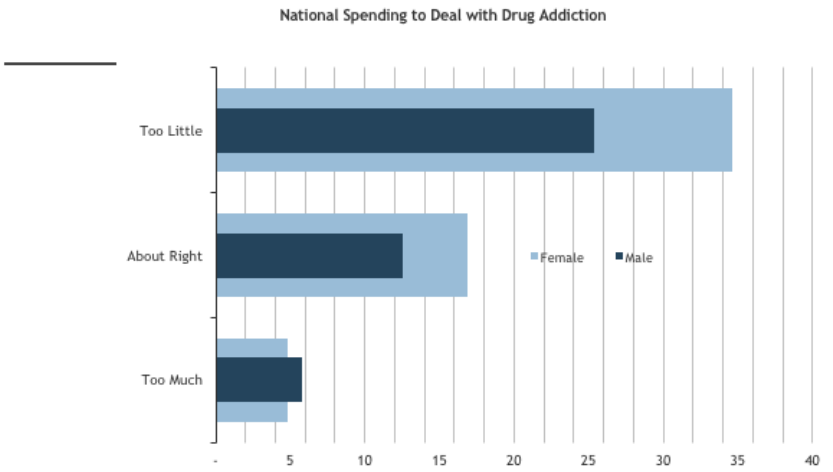
National Spending to Deal with Drug Addiction



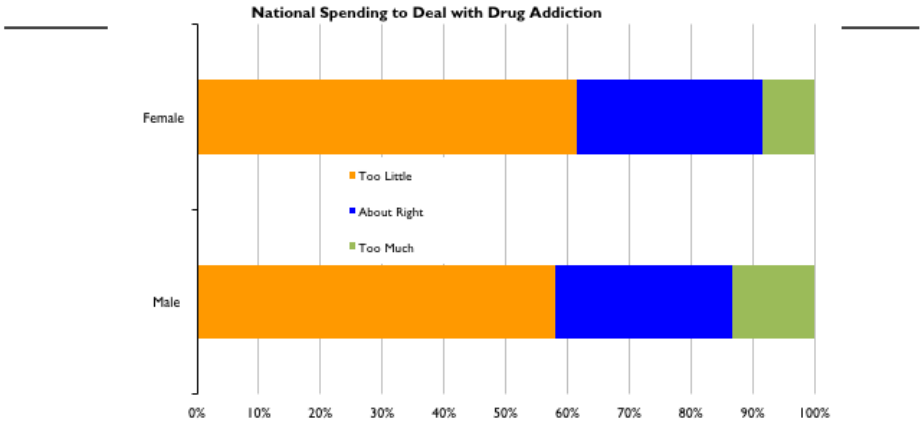
SFU Visualizing Tabular Data



National Spending to Deal with Drug Addiction



SFU Visualizing Tabular Data



SFU Visualizing Tabular Data