Detecting outliers

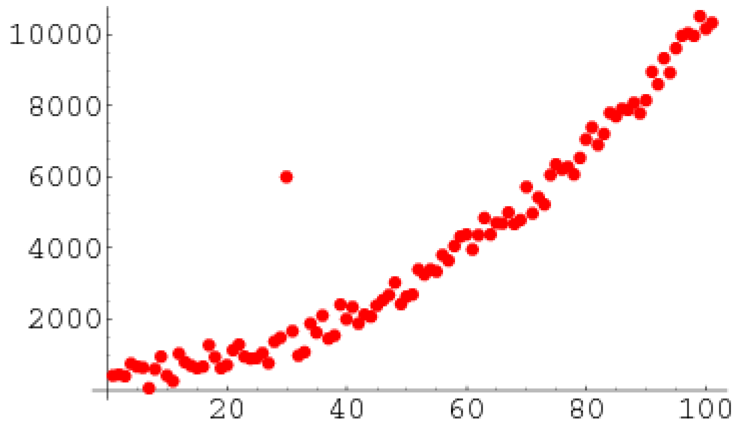
Effectiveness and expressiveness principle

Visual exaggeration the lie factor Tufte

Detecting outliers

- Ways to Detect Outliers/Anomalies

What is Anomaly/Outlier?

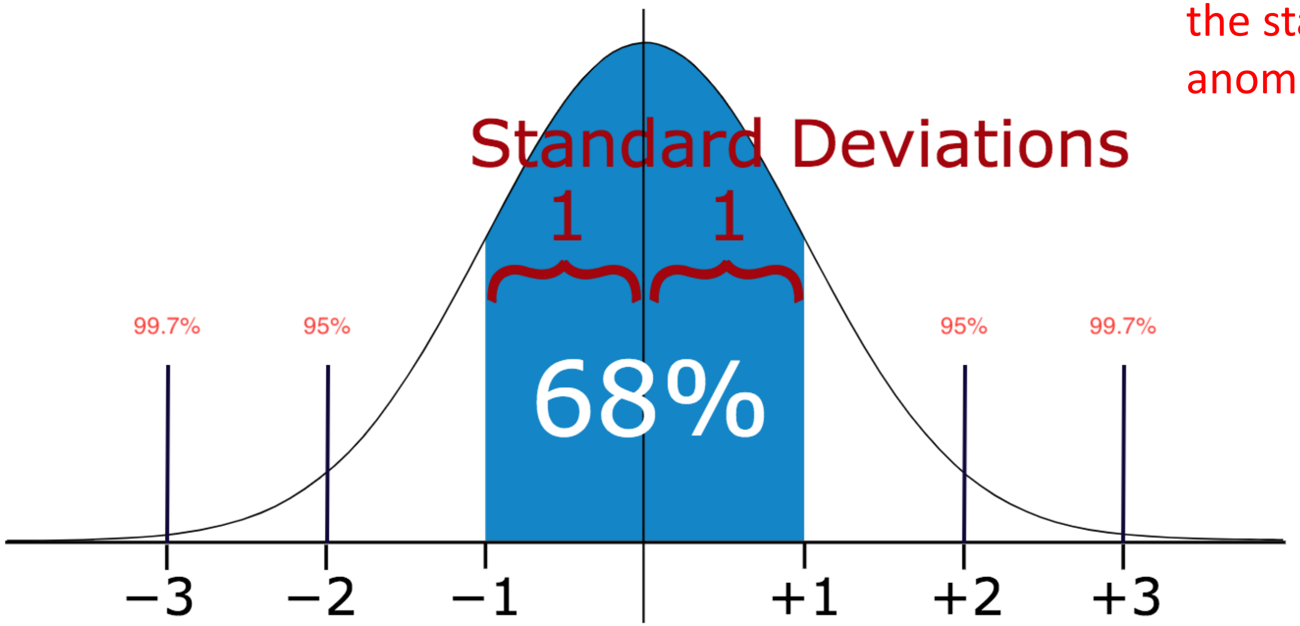


- **In statistics**, outliers are data points that don’t belong to a certain population.
- It is an abnormal observation that lies far away from other values.
- An outlier is an observation that diverges from otherwise well-structured data.

For Example, you can clearly see the outlier in this list:
[20,24,22,19,29,18,**4300**,30,18]

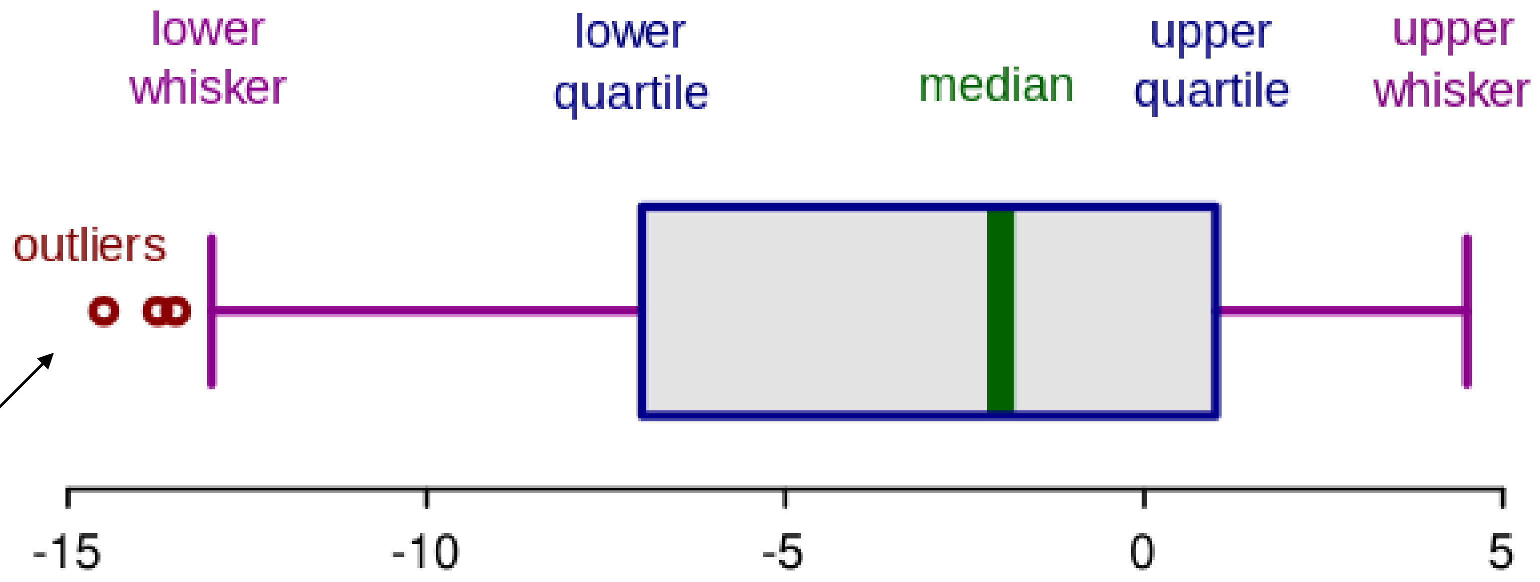
Method 1 — Standard Deviation:

In statistics, If a data distribution is approximately normal then about 68% of the data values lie within one standard deviation of the mean and about 95% are within two standard deviations, and **about 99.7%** lie within three standard deviations.



Therefore, if you have any data point that is more than 3 times the standard deviation, then those points are very likely to be anomalous or outliers.

Method 2 — Boxplots

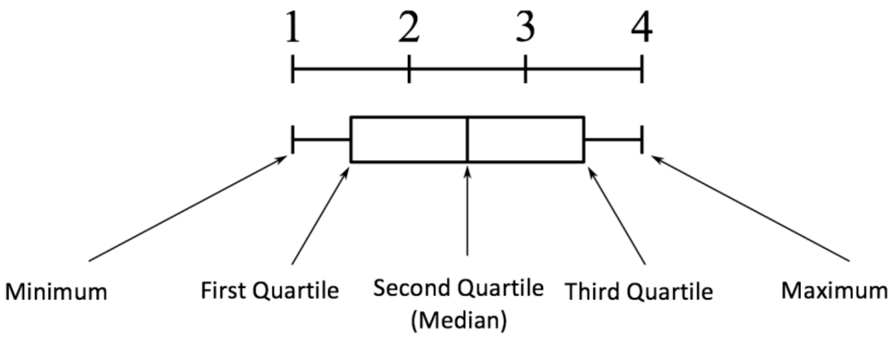


outliers

An outlier is defined here as a data point that is located outside the whiskers of the box plot.

Method 3- Tukey's fences

<u>Boxplot Anatomy:</u>



- Interquartile Range (IQR) is important because it is used to define the outliers.
- It is the difference between the third quartile and the first quartile (IQR = Q3 -Q1).
- Outliers in this case are defined as the observations that are below (Q1 − 1.5x IQR) or *boxplot lower whisker* or above (Q3 + 1.5x IQR) or *boxplot upper whisker*.

<u>Boxplot Anatomy:</u>

- The concept of the I**nterquartile Range** (**IQR**) is used to build the boxplot graphs. IQR is a concept in statistics that is used to measure the statistical dispersion and data variability by dividing the dataset into quartiles.

<u>Find Quartiles: Examples</u>

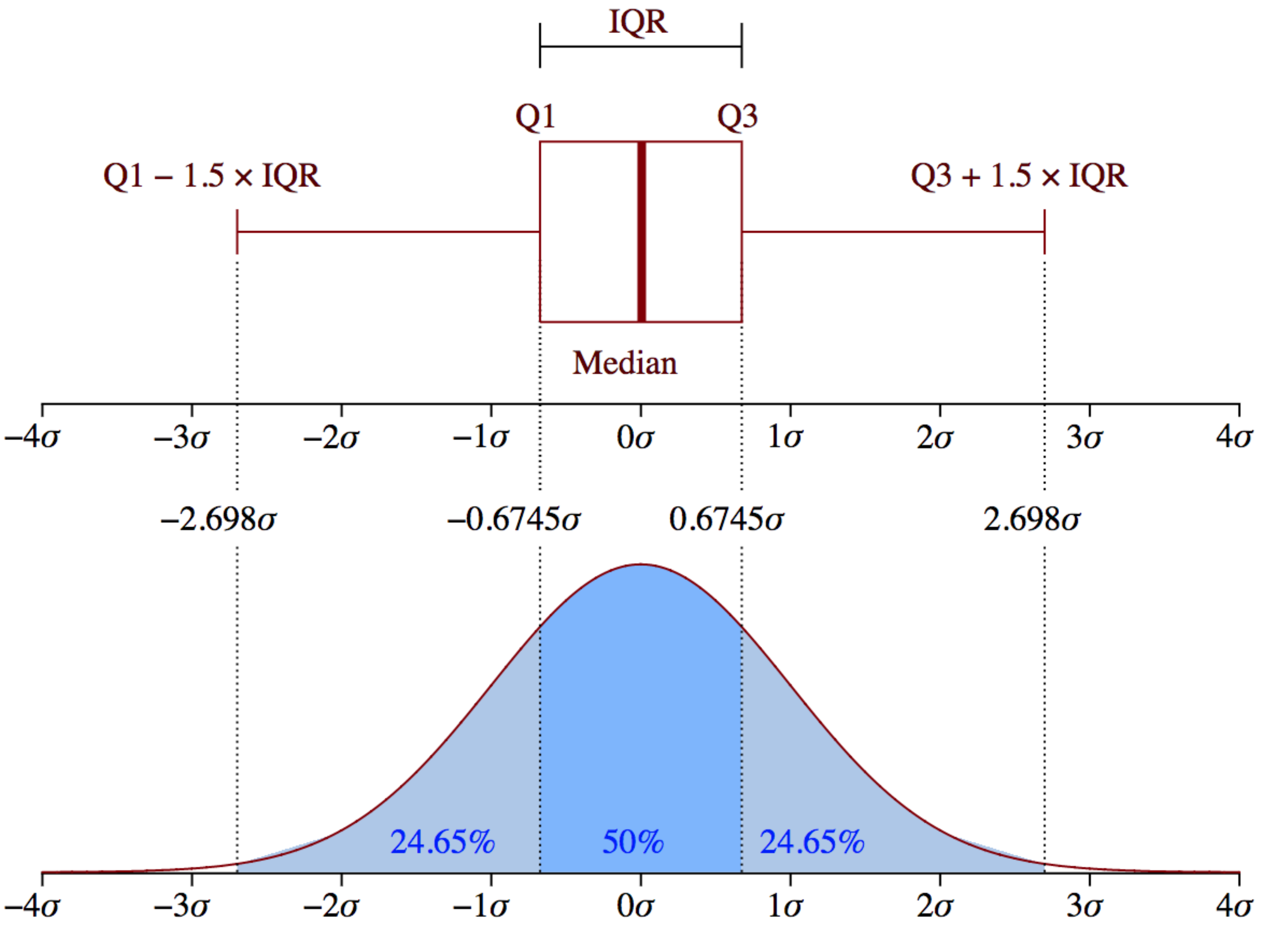
Example: Divide the following data set into quartiles: 2, 5, 6, 7, 10, 22, 13, 14, 16, 65, 45, 12.

<u>Step 1:</u> Put the numbers in order: 2, 5, 6, 7, 10, 12 13, 14, 16, 22, 45, 65.

<u>Step 2:</u> Count how many numbers there are in your set and then divide by 4 to cut the list of numbers into quarters. There are 12 numbers in this set, so you would have 3 numbers in each quartile.

2, 5, 6, | 7, 10, 12 | 13, 14, 16, | 22, 45, 65

Superimposing when data is normal distributed



Method 4- MAD distance

In statistics, the **median absolute deviation** (**MAD**) is a robust measure of the variability of a univariate sample of quantitative data. It can also refer to the population parameter that is estimated by the MAD calculated from a sample. For a univariate data set X_1, X_2, ..., X_n, the MAD is defined as the median of the absolute deviations from the data's median.

Find the MAD of the following set of numbers:
3, 8, 8, 8, 8, 9, 9, 9, 9.

median=8

Subtract the median from each x-value using the formula $|y_i - \text{median}|$.
|3-8|, |8-8|, etc…

Find the **median of the absolute differences**. The median of the differences (0,0,0,0,**1**,1,1,1,5) is 1.

The median absolute deviation formula is:
MAD = median(X_i - m), where

- m is the median of a dataset; and
- X_i is the dataset in question.

Method 5- Mahalanobis Distance

- The most common use for the **Mahalanobis distance** is to find multivariate outliers, which indicates unusual combinations of two or more variables. For **example**, it's fairly common to find a 6′ tall woman weighing 185 lbs, but it's rare to find a 4′ tall woman who weighs that much.

- The Mahalanobis distance (MD) is **the distance between two points in multivariate space**. In a regular Euclidean space, variables (e.g. x, y, z) are represented by axes drawn at right angles to each other; The distance between any two points can be measured with a ruler.

- For uncorrelated variables, the Euclidean distance equals the MD. However, if two or more variables are correlated, the axes are no longer at right angles, and the measurements become impossible with a ruler. In addition, if you have more than three variables, you can’t plot them in regular 3D space at all. The MD solves this measurement problem, as it measures distances between points, even correlated points for multiple variables.

Mahalanobis Distance Definition

- MD is primarily used in classification and clustering problems where there is a need to establish correlation between different groups/clusters of data. Another application of MD is discriminant analysis and pattern analysis, which are based on classification.

- The Mahalanobis distance of an observation $x = (x_1, x_2, x_3....x_N)^T$ from a set of observations with mean $\mu = (\mu_1, \mu_2, \mu_3....\mu_N)^T$ and covariance matrix S is defined as:

  $MD(x) = \sqrt{(x - \mu)^T S^{-1} (x - \mu)}$

- The covariance matrix provides the covariance associated with the variables (the reason covariance is followed is to establish the effect of two or more variables together).

Evaluating Expressiveness and Effectiveness of Informative Charts

- The main objective of data visualization is to pass the messages to readers expressively and effectively.

- The priority of the beautifulness of the chart is put lower than the expressiveness and effectiveness of the chart.

Evaluating Expressiveness and Effectiveness of Informative Charts

How do we measure expressiveness in the context of data visualization?

- The chart is considered **expressive** if it can express the information in the dataset attributes.

- The dataset attributes normally consist of 3 main types, which are nominal, ordinal and quantitative. To summarize, both nominal and ordinal attributes are qualitative attributes.

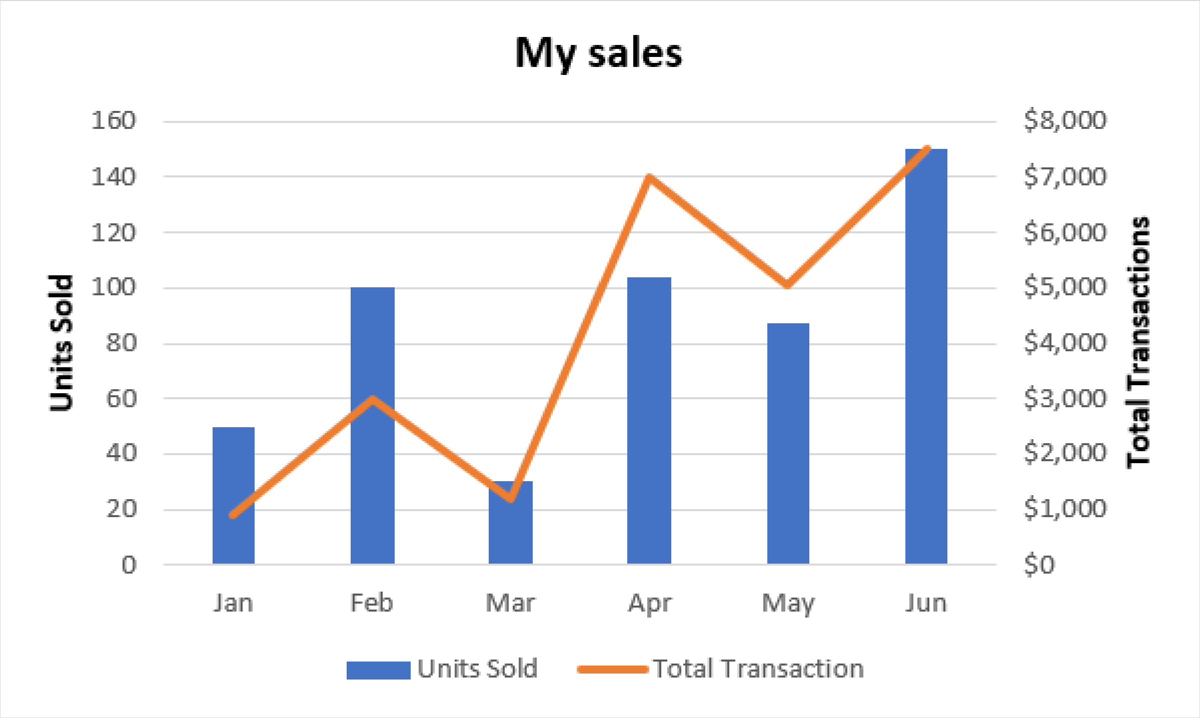
- However, nominal attributes have no implicit ordering whereas ordinal attributes are the vice versa. Both ordinal and quantitative attributes have implicit ordering but there is no meaning in applying arithmetic operations on ordinal attributes only.

- The examples of nominal, ordinal and quantitative attributes include:

• Nominal — Types of car, Gender
• Ordinal — Date, Size of T-shirt
• Quantitative — Revenue, Profit

Evaluating Expressiveness and Effectiveness of Informative Charts

- If the attribute is ordered, it should appear in the chart as ordered attributes.
- Conversely, unordered data should not be shown in a way that perceptually implies an ordering that does not exist.
- For an instance, if x-axis is a nominal attribute and y-axis is a quantitative attribute, a bar chart should be drawn instead of a line chart. If x-axis is an ordinal attribute and y-axis is a quantitative attribute, a line chart should be drawn instead of a bar chart.



A mistake of combining these plots!!

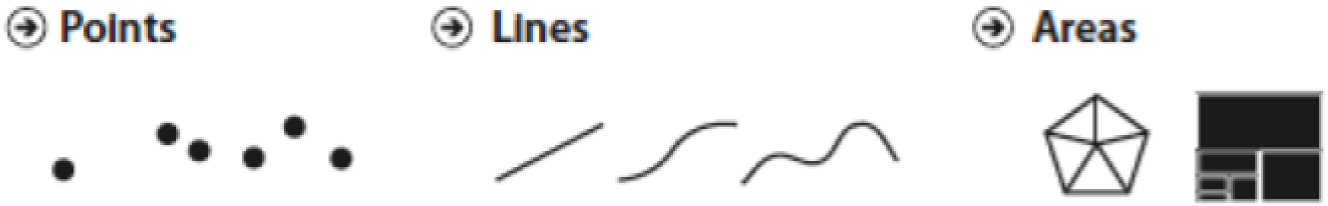
Marks and Channels
(Chapter 5 in Visualization Analysis and Design)

Why Marks and Channels

- Learning the reason about marks and channels gives you the building blocks for analyzing visual encoding.
- The core of the design space of visual encodings can be described as an orthogonal combinations of two aspects: Graphical elements called marks, and visual channels to control their appearance. Even complex visual encodings can be broken down into components that can be analyzed in terms of their marks and channel structure.

Defining Marks and Channels

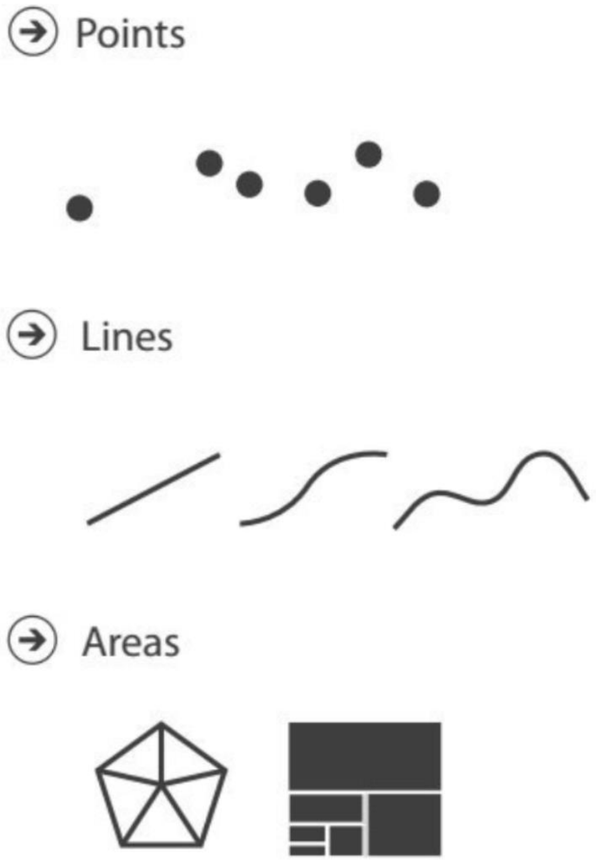
- Marks: basic geometric elements that depict items and links



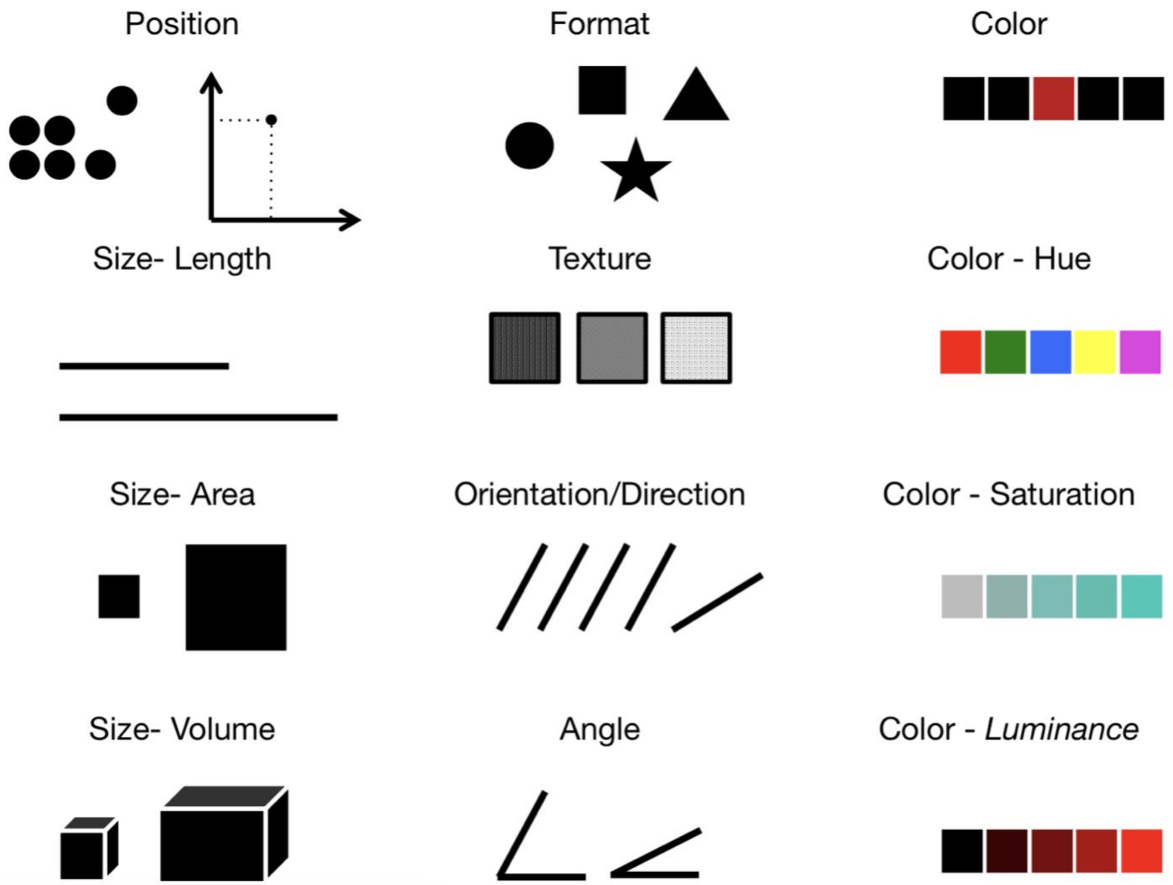
- Channels: control the marks' appearance
  - Magnitude for ordered data
  - Indentify for categorical data
- Marks and channels are building blocks for visual encoding

Defining Marks and Channels

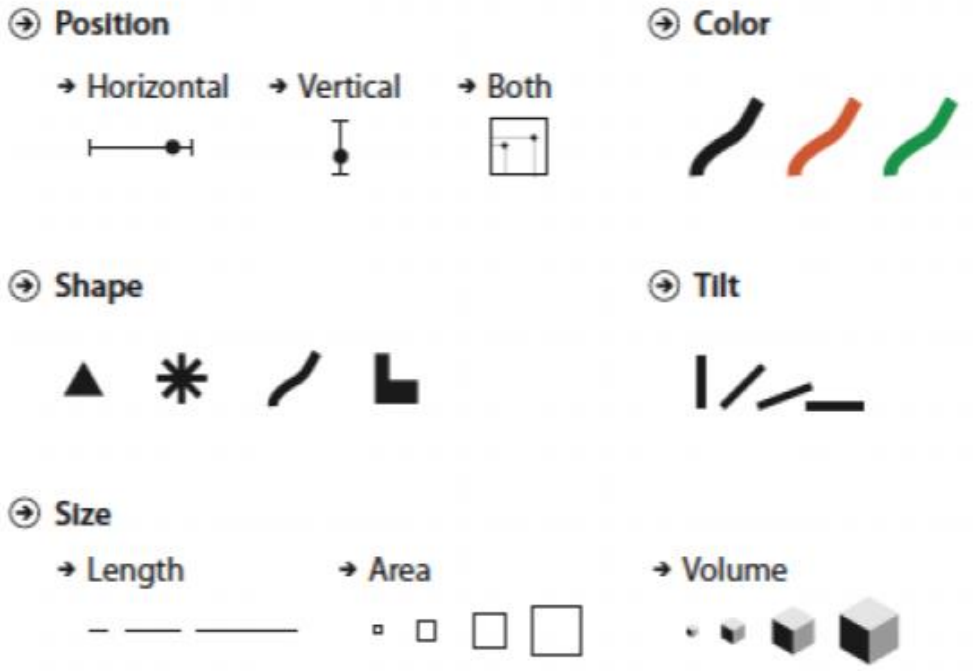
MARKS



CHANNELS

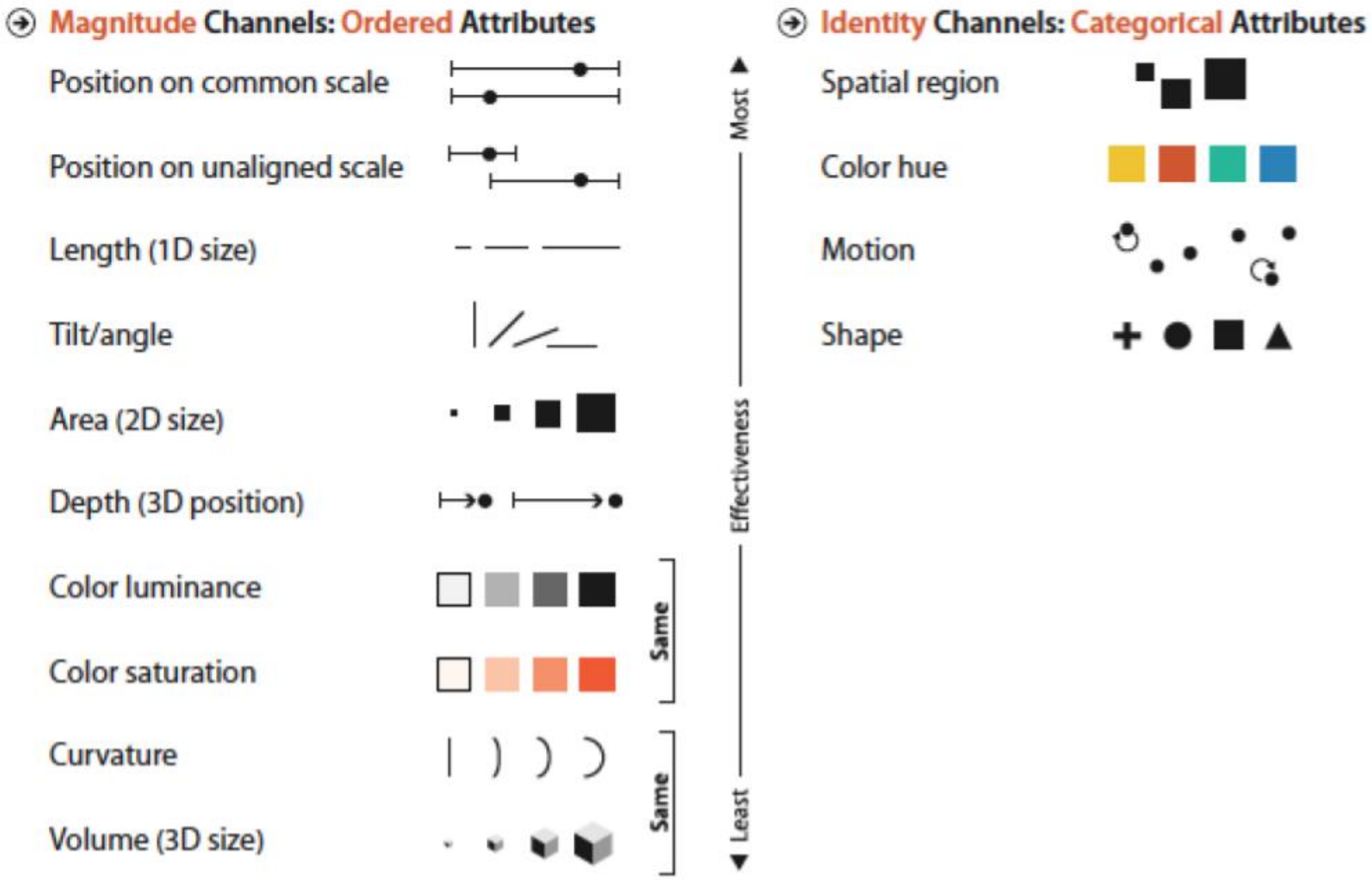


- Marks can be classified according to their spatial dimensions
  - 0 D: points; 1D: lines; 2D: areas, etc.
- Channels encode properties of a mark



Channels: Expressiveness Types and Effectiveness Ranks

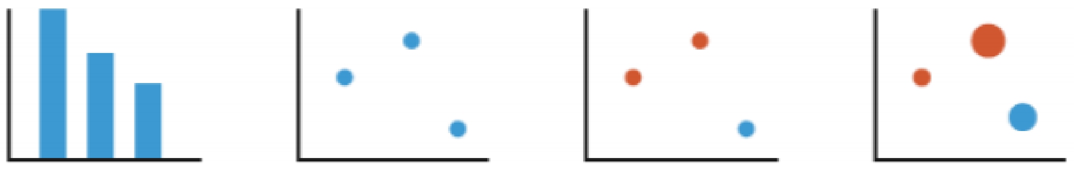
The effectiveness of the channels that modify the appearance of marks depends on matching the expressiveness of channels with the attributes being encoded.



- As shown, the channels for ordered attributes are known as magnitude channels whereas the channels for categorical attributes are known as identity channels.
- Higher the positions of the channels, higher the ranking of effectiveness.
- According to Tamara Munzer (2014), the most important attributes should be encoded with the most effective channels to be most noticeable, and then decreasingly important attributes can be matched with less effective channels.

PLOTS with marks and channels

- Bar charts:
  - Marks: Lines
  - Channels: Vertical lengths and horizontal positions
- Scatterplots:
  - Marks: Points
  - Channels: Vertical and Horizontal positions + colors (optional)

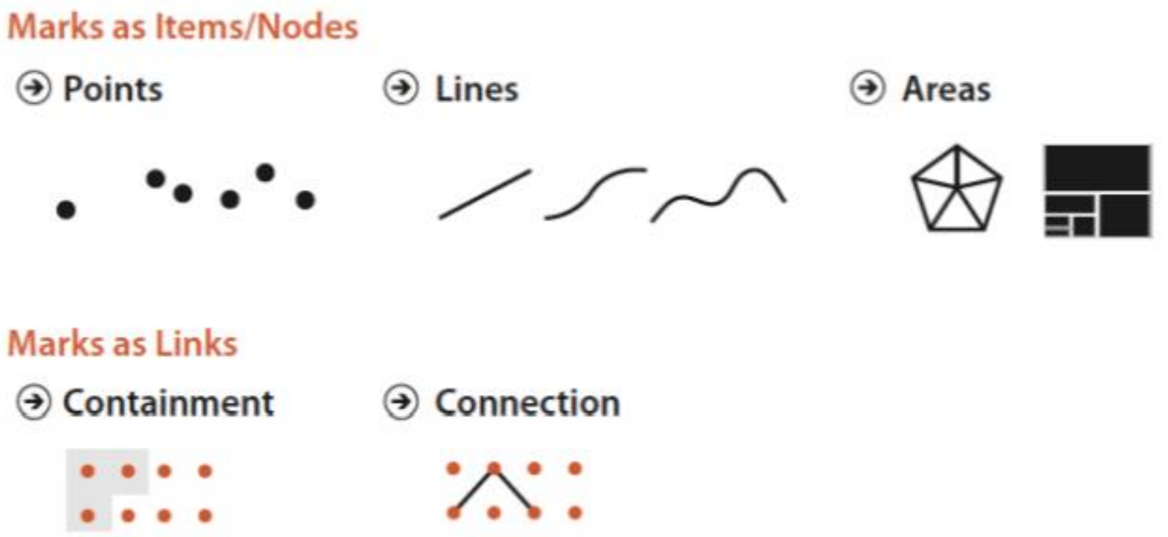


(a) (b) (c) (d)

Channel and Mark Types

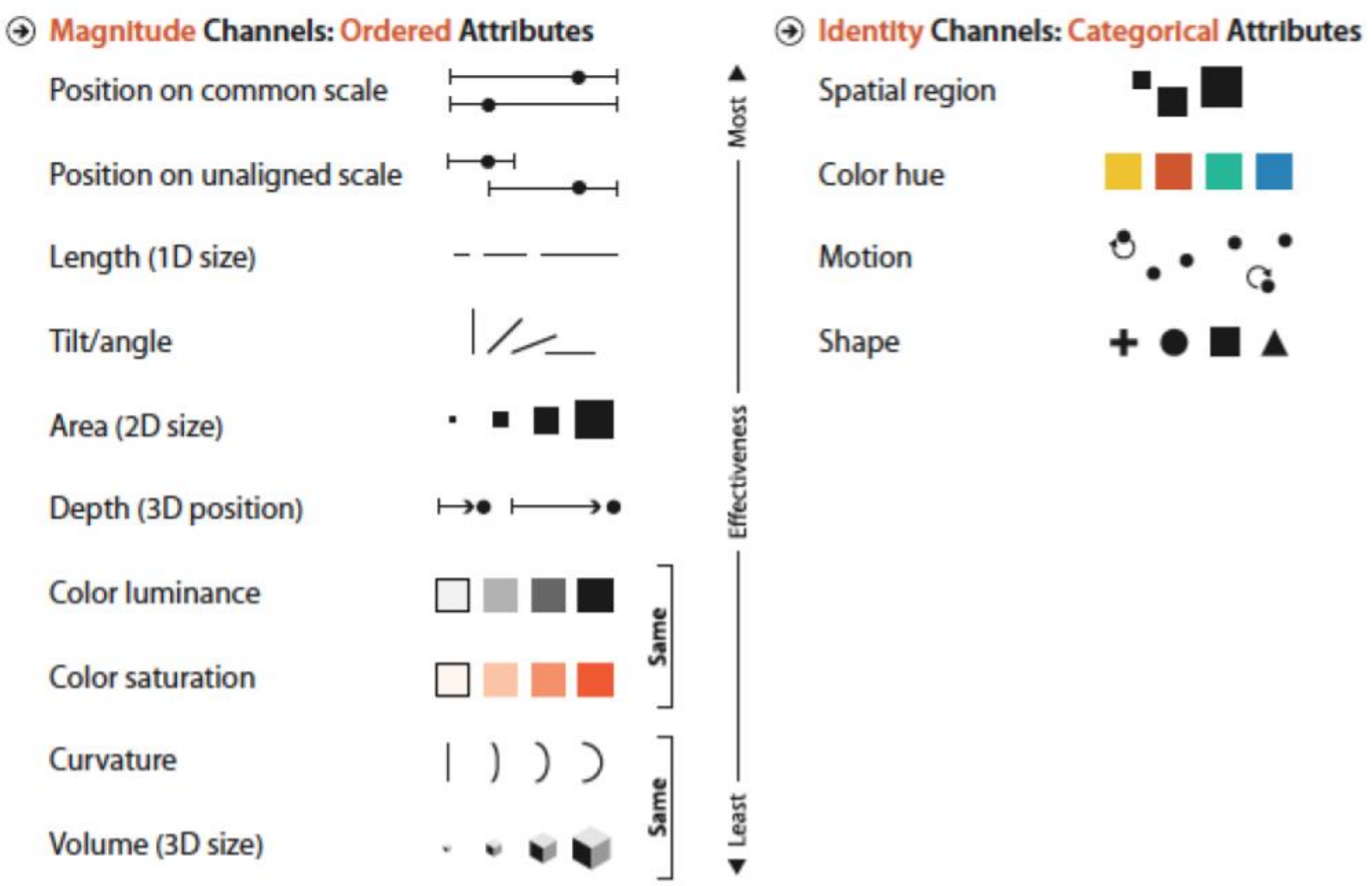
- Channel Types:
  - Identify channels: what something is and where it is (circle, triangle, cross, etc.)
  - Magnitude channels: how much something there is (length, luminance, etc.)
- Mark Types:
  - Item marks
  - Link marks: show relationship between items
    - Connection marks: show pair wise relationship
    - Containment marks: show hierarchical relationship

- Mark Types:
  - Item marks
  - Link marks: show relationship between items
    - Connection marks: show pair wise relationship
    - Containment marks: show hierarchical relationship



Choice of Marks and Channels

- Expressiveness
  - The visual encoding should express all of the information in the data set
    - For example, ordered data are seen as orders (and vice versa)
- Effectiveness
  - The importance of the attribute should match the salience of the channel
    - For example, important items are made the most noticeable


Magnitude Channels: Ordered Attributes
Position on common scale
Position on unaligned scale
Length (1D size)
Tilt/angle
Area (2D size)
Depth (3D position)
Color luminance
Color saturation
Curvature
Volume (3D size)
Same
Same
Most
Effectiveness
Least
Identity Channels: Categorical Attributes
Spatial region
Color hue
Motion
Shape

Channel Effectiveness

- How do we determine the ranking above?
  - Accuracy
  - Discriminability
  - Separability
  - Popout
  - Grouping

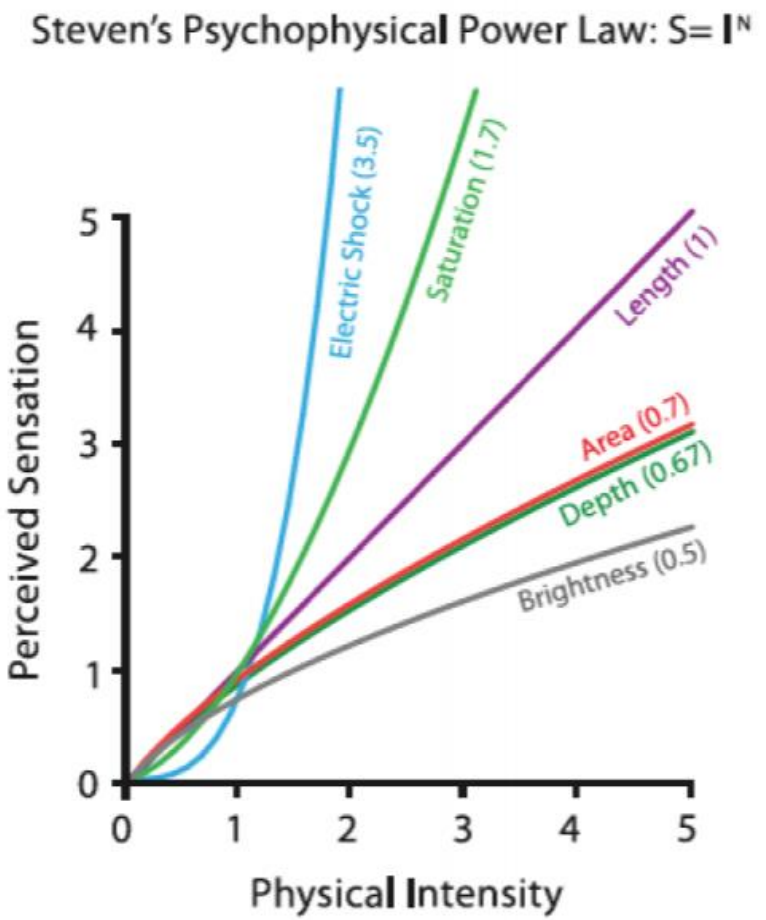
Channel Accuracy

- How close is human perceptual judgment to some objective measurement of the stimulus?
- Our responses to the sensory experience of magnitude follow power laws

$$S = I^n$$

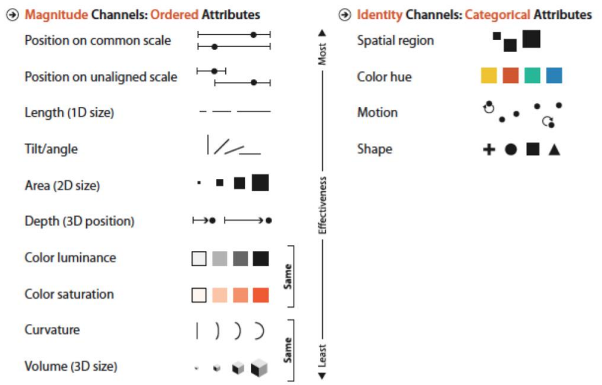
  - S: perceived sensation
  - I: physical intensity

Steven's Pyschophysical Law



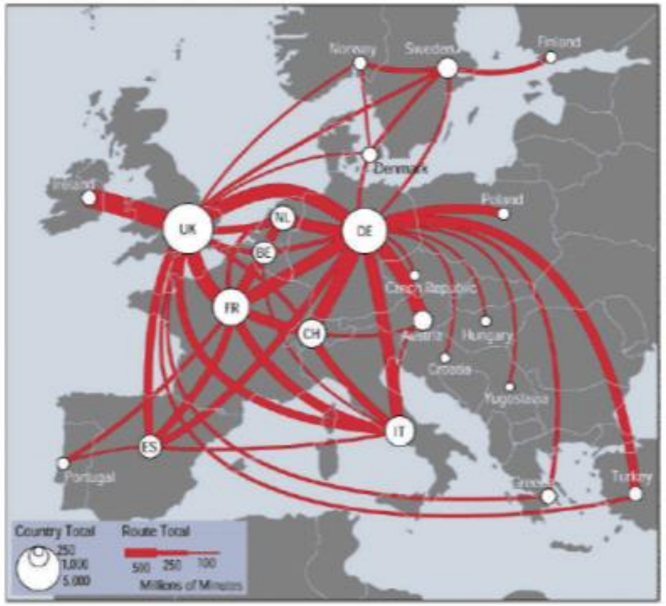
Channel Accuracy

- Cleveland and McGill’s experiment on magnitude channel
- Aligned position > unaligned position > length > angle > area > volume > curvature > luminance > hue



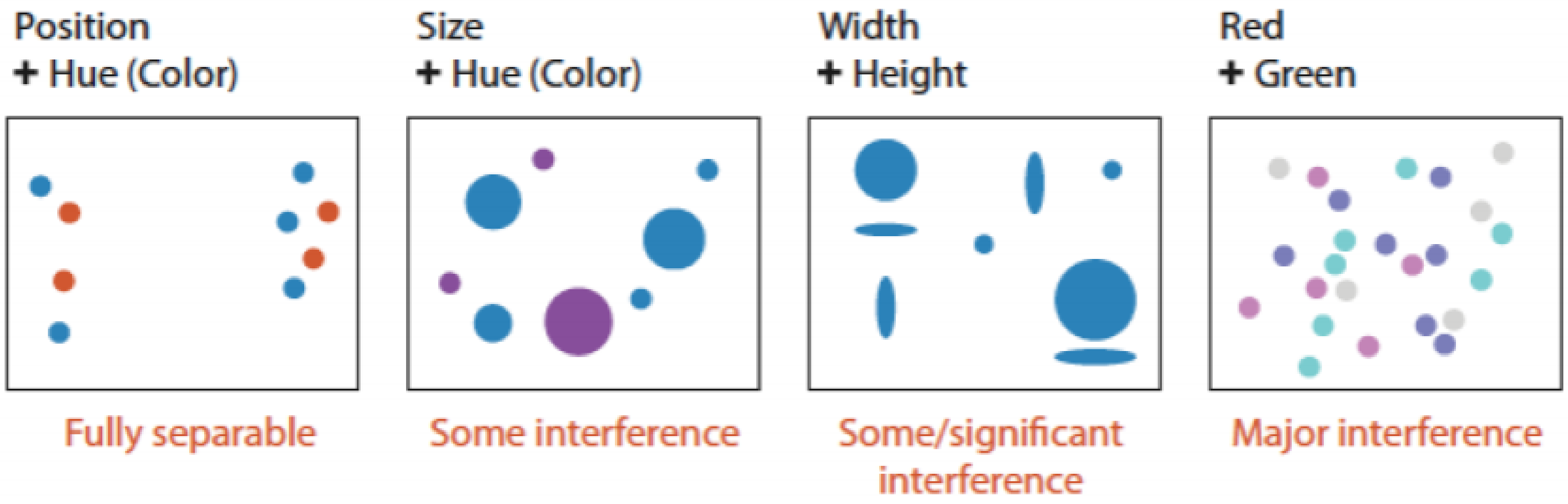
Discriminability

- How many distinguishable levels (bins) in the channel?
  - Linewidth works well for 3 or 4 levels



Separability

- Not all channels are independent



Popout

- Distinct items stand out

Harder

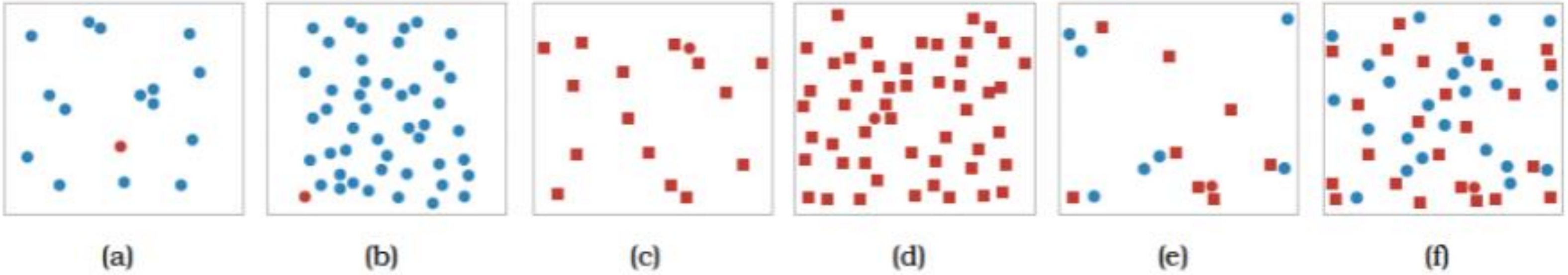
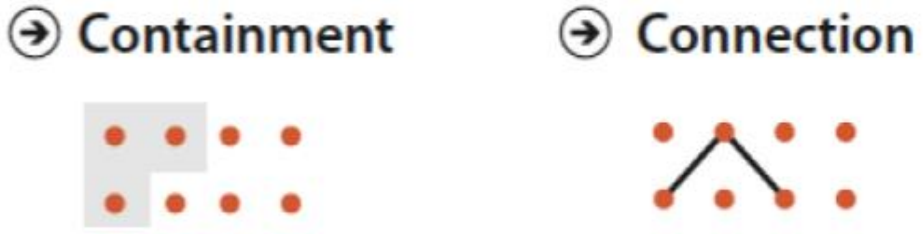


Figure 5.11. Visual popout. (a) The red circle pops out from a small set of blue circles. (b) The red circle pops out from a large set of blue circles just as quickly. (c) The red circle also pops out from a small set of square shapes, although a bit slower than with color. (d) The red circle also pops out of a large set of red squares. (e) The red circle does not take long to find from a small set of mixed shapes and colors. (f) The red circle does not pop out from a large set of red squares and blue circles, and it can only be found by searching one by one through all the objects. After http://www.csc.ncsu.edu/faculty/healey/PP by Christopher G. Healey.

More difficult to pop out with multiple channels combined together

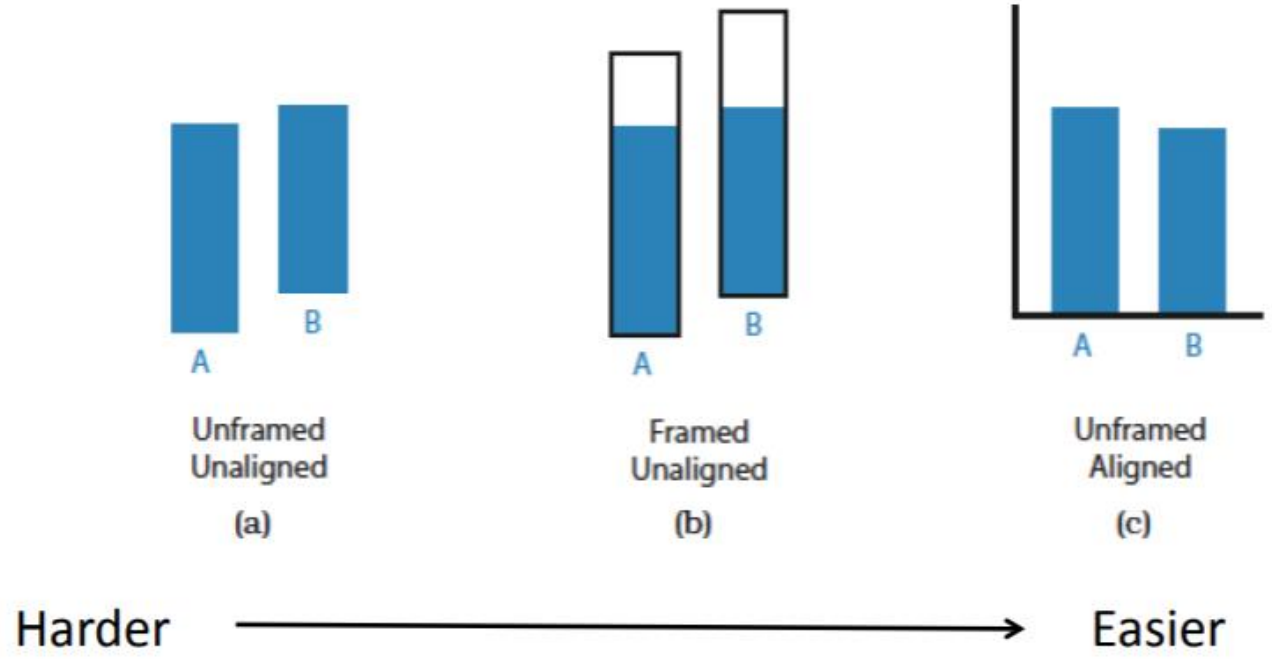
Grouping

- Select proper channels that allow visual grouping or visual clustering
  - Use link marks with area of containment
  - Use identify channel to encode categorical data
  - Proximity: placing similar items nearby
  - Similarity: hue, motion, etc

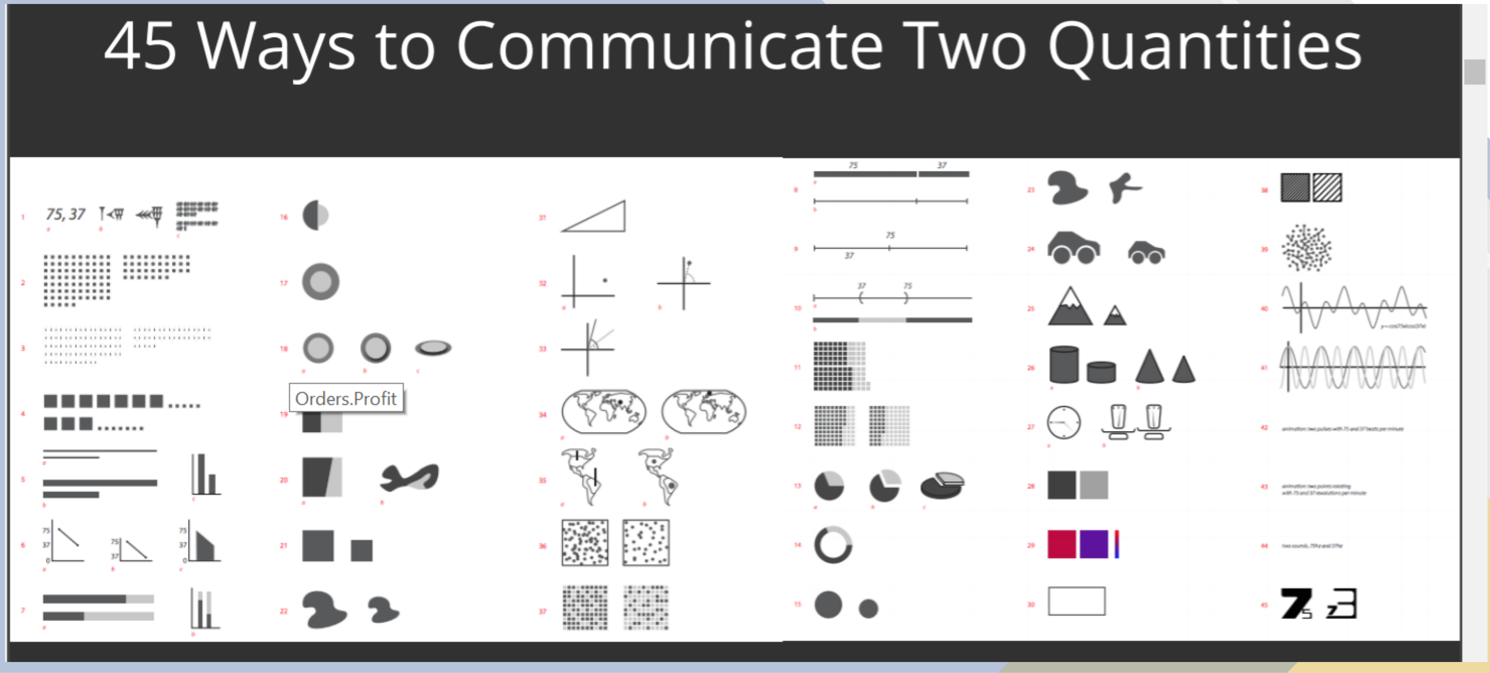


Relative vs. Absolute Judgment

- Human perception is based on relative judgment, not absolute - Weber's law



Improve Your Visualization Skills Using Tufte's Principles of Graphical Design



Which one is the best and why?

Is there an ideal way to visualize a data set?

It depends on

- Data types e.g., table, network, spatial, temporal
- Context of the data
- Tasks to perform e.g., identify trends, compare values
- Questions to answer
- Messages to deliver

But, is there at least a guide for visualization design?

Edward Tufte’s Design Principles

Tell the Truth!

The representation of numbers … should be directly proportional to the numerical quantities measured. — Edward Tufte 1983

Lie Factor

$$\text{Lie Factor} = \frac{\text{Size of effect in graphic}}{\text{Size of effect in data}}$$

Lie factor greater than 1.05 or less than 0.95 indicate substantial distortion, far beyond minor accuracies in plotting.

$$\text{Lie Factor} = \frac{\text{Size of effect in graphic}}{\text{Size of effect in data}}$$

$$\text{Size of effect} = \text{Percentage change}$$

$$= \frac{|\,V_1 - V_2\,|}{|\,V_1\,|}$$

Example

- The idea of the lie factor is to express in numbers, how much a graphic deviates from the actual data it should represent.

Table 1: Example data for lie factor graphics

year	*earnings*
2001	100
2002	100

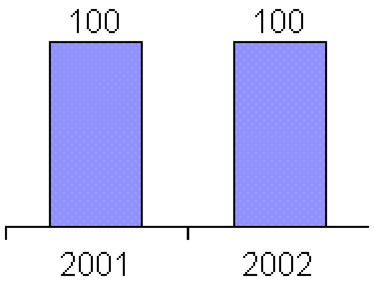


Chart 1: Truthfull representation of data

- Since this is an accurate representation of the data, the lie factor equals one. On the other hand, the artist in charge of making charts might for some weird reason decide to "emphasize" the second bar in the chart:

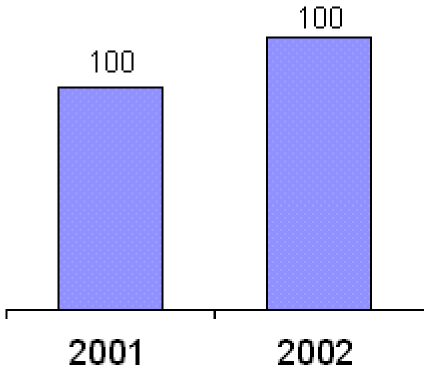


Chart 2: False representation of data

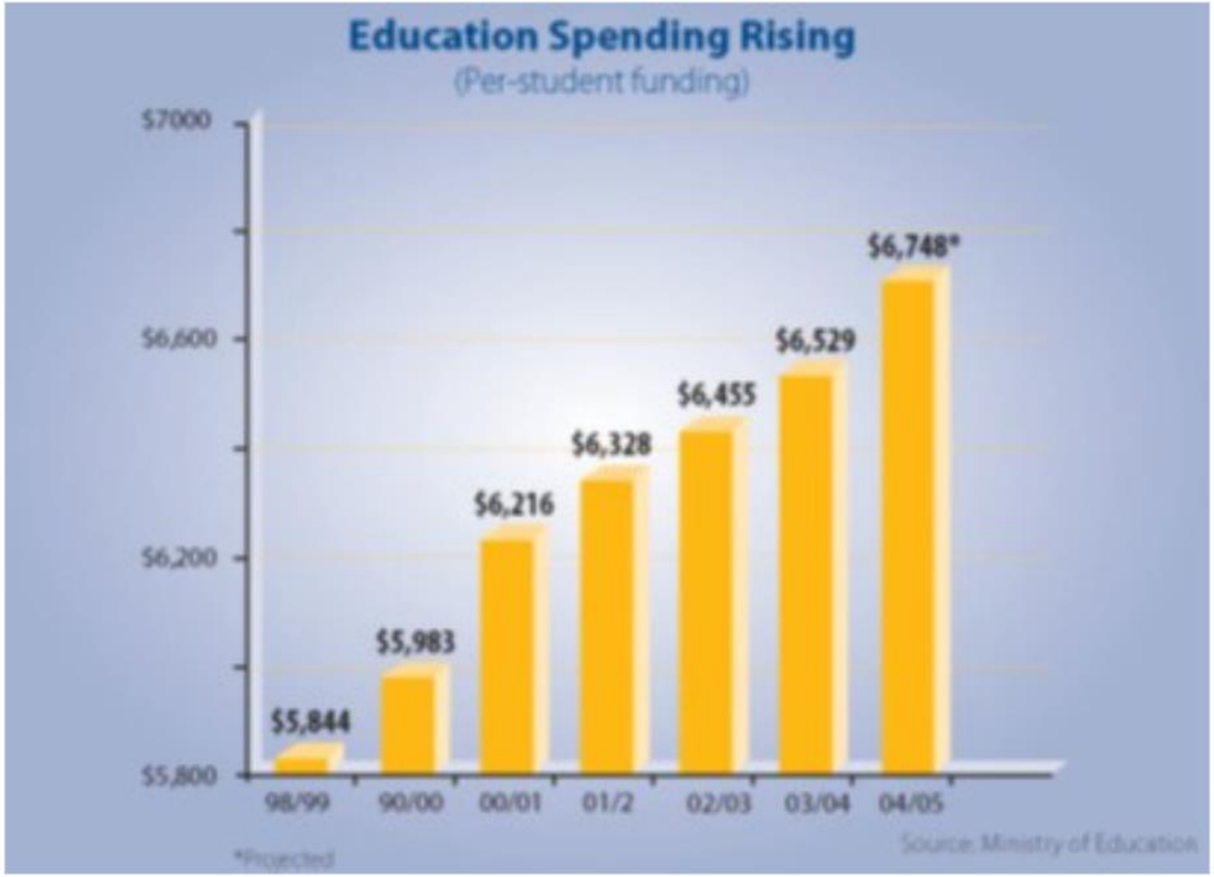
$$\text{lie factor} = \frac{\text{size of effect shown in graphic}}{\text{actual effect in data}}$$

$$= \frac{(\text{size of right bar}/\text{size of left bar})}{(\text{value of right bar}/\text{value of left bar})}$$

$$= \frac{(3{,}58\ cm/2{,}92\ cm)}{(100\ /100)} \approx 1{,}23$$

- Truthful charts always have a lie factor of one, whereas any other lie factor indicates a misrepresentation

Example

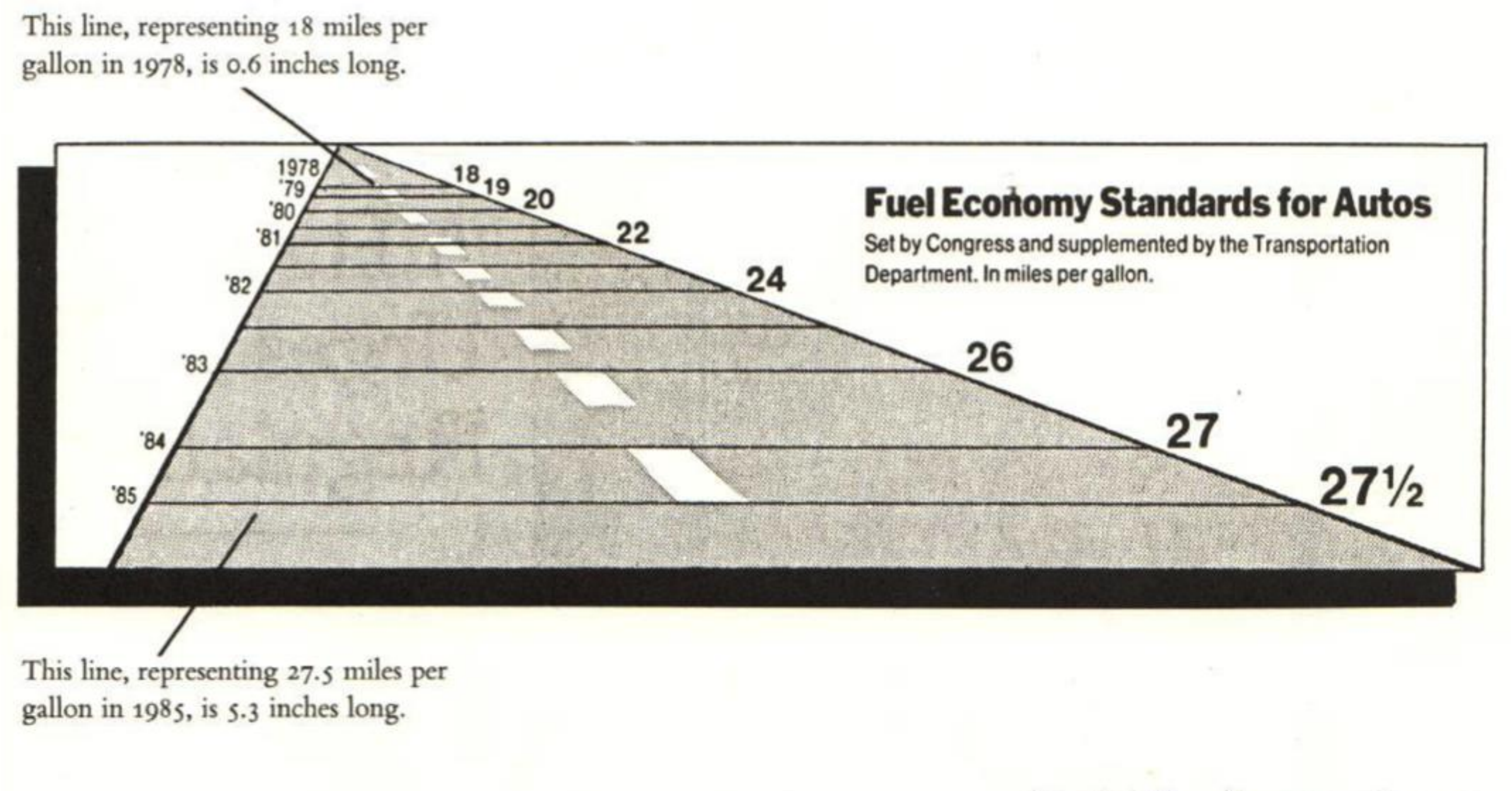


Effect in graphic: 2.33/0.08
= 29.1

Effect in data: 6748/5844
= 1.15

Lie factor = 29.1 / 1.15
= 25.3

Example

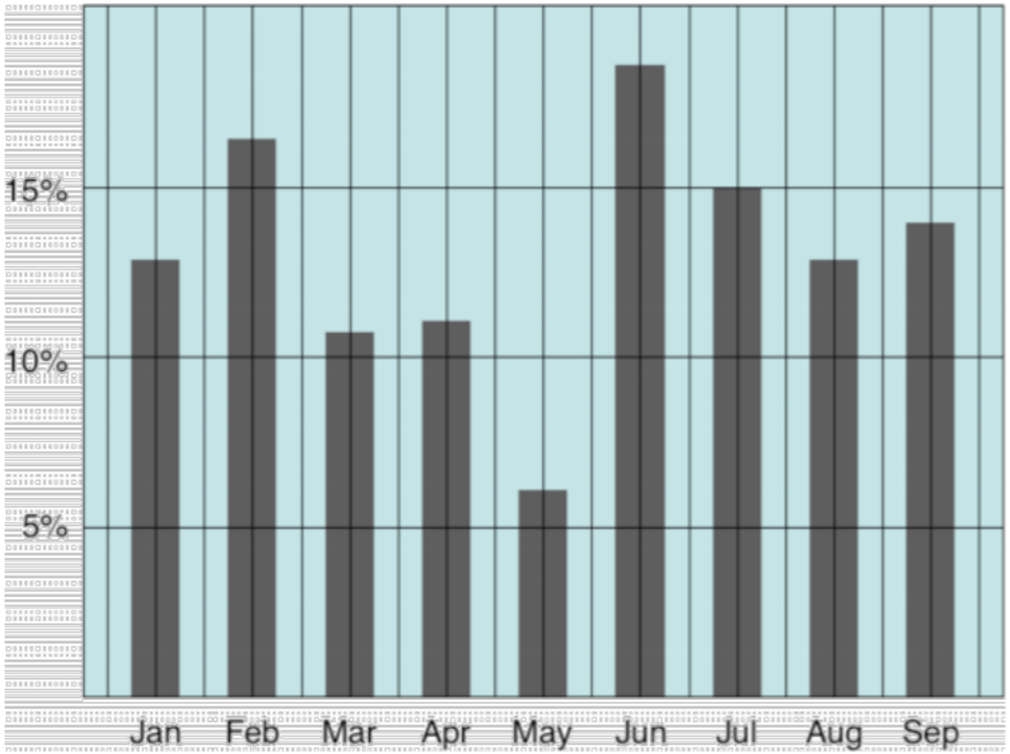


New York Times, August 9, 1978, p. D-2.

- Using this metric, we can see that although the true change in fuel economy standards from ’78 to ’85 was a (27.5−18)/18=53% change, the chart shows it as a (5.3−.6)/.6=783% , or a Lie factor of 14.814.8, where we’d hope for a number around 1 (note this example is just a re-presentation of the comments from Tufte).

http://benjaminleroy.github.io/pages/blog/public/post/2018/05/16/review-of-tufte-s-the-visual-display-of-quantitative-information/

<u>Chart Junks</u> = Unnecessary visual elements in charts that distracts the viewer from the information



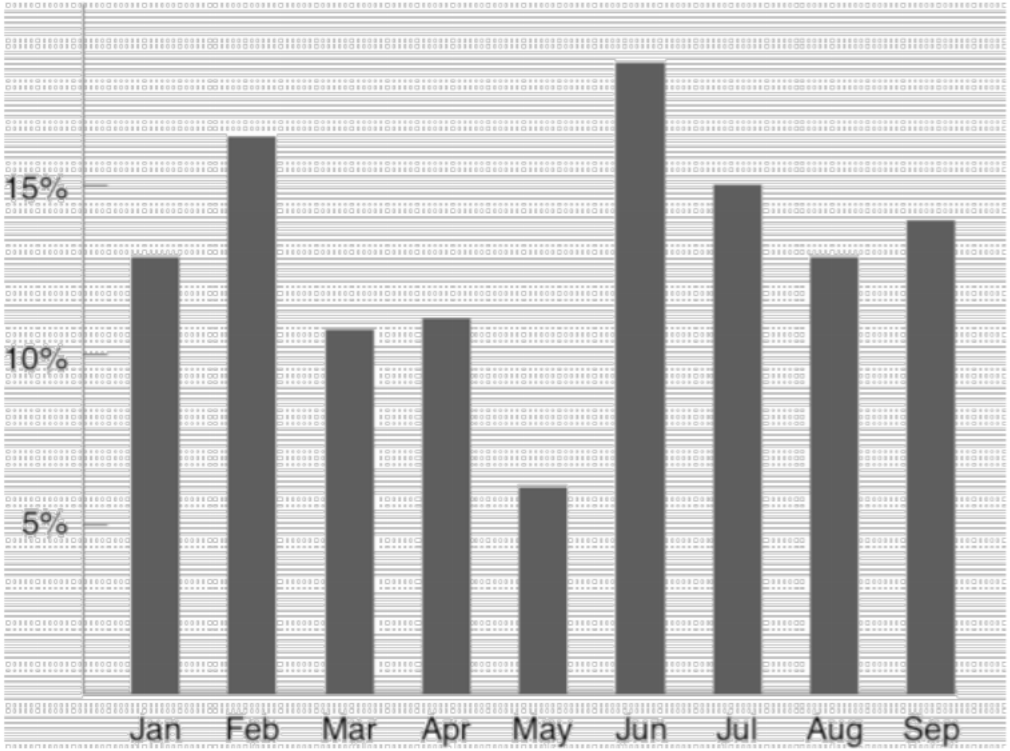
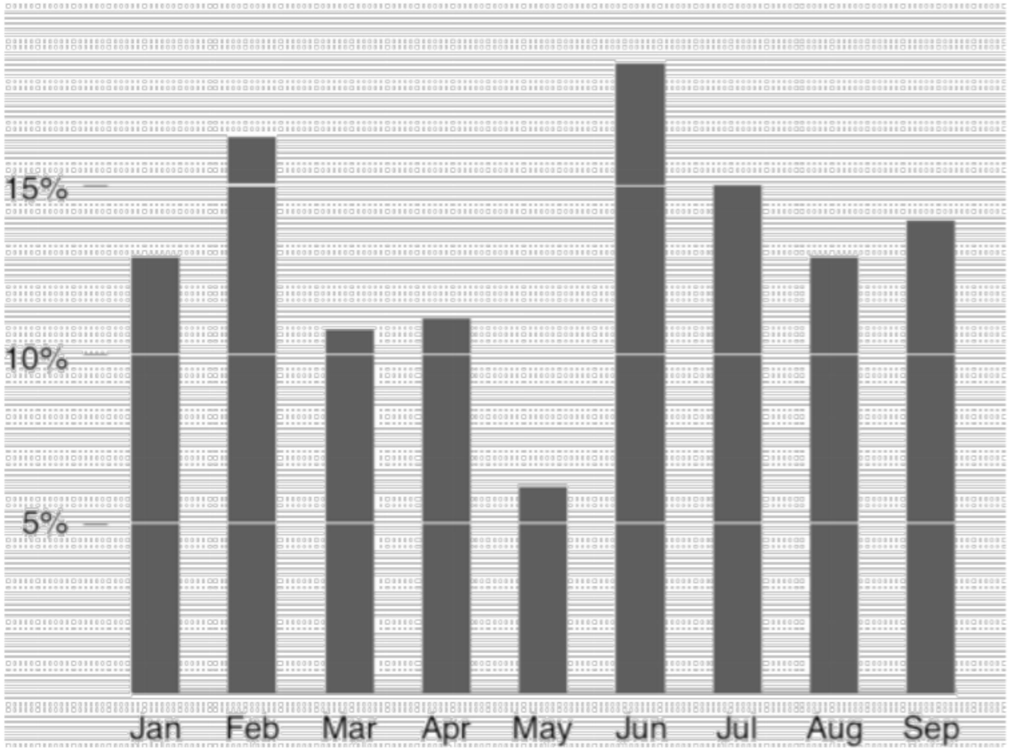
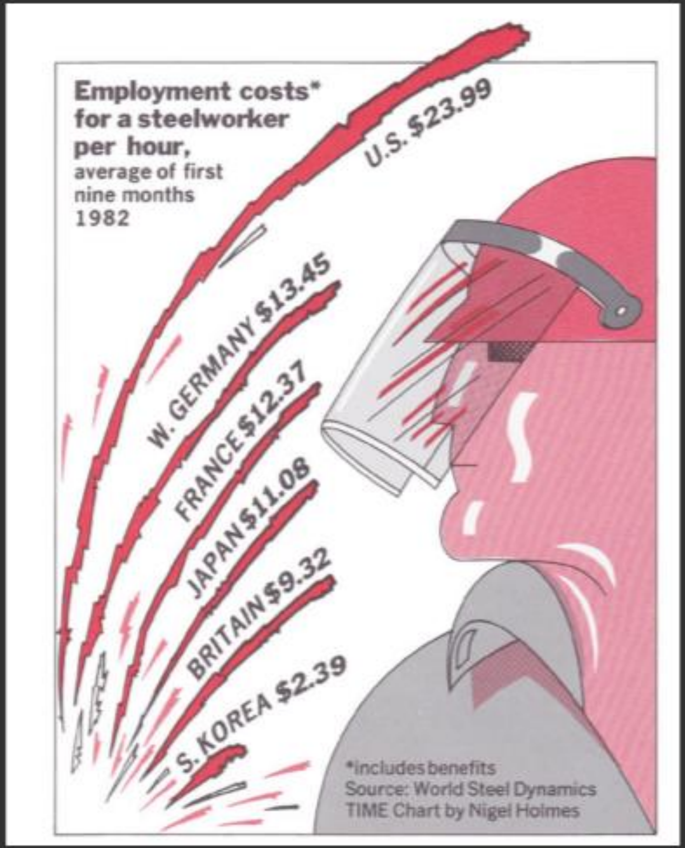


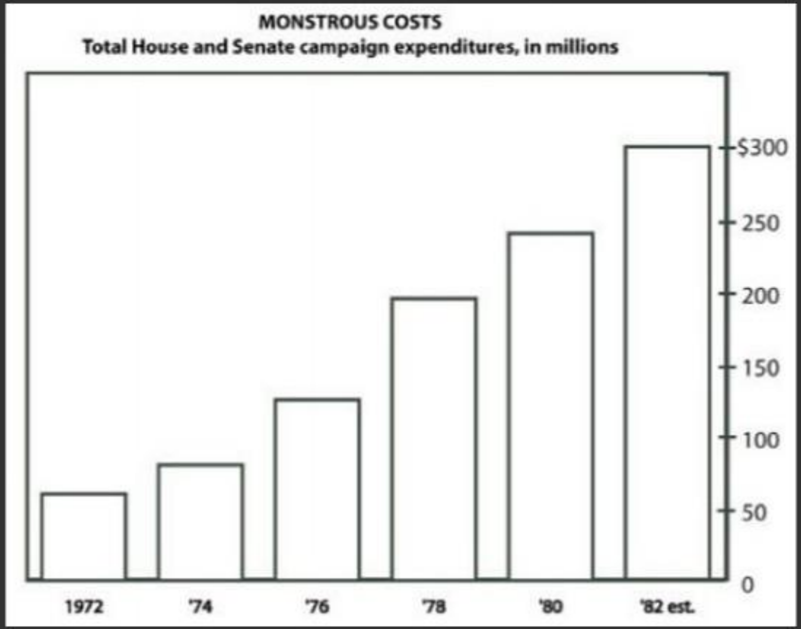
Chart Junks = Unnecessary visual elements in charts that distracts the viewer from the information



Are these chart junks?

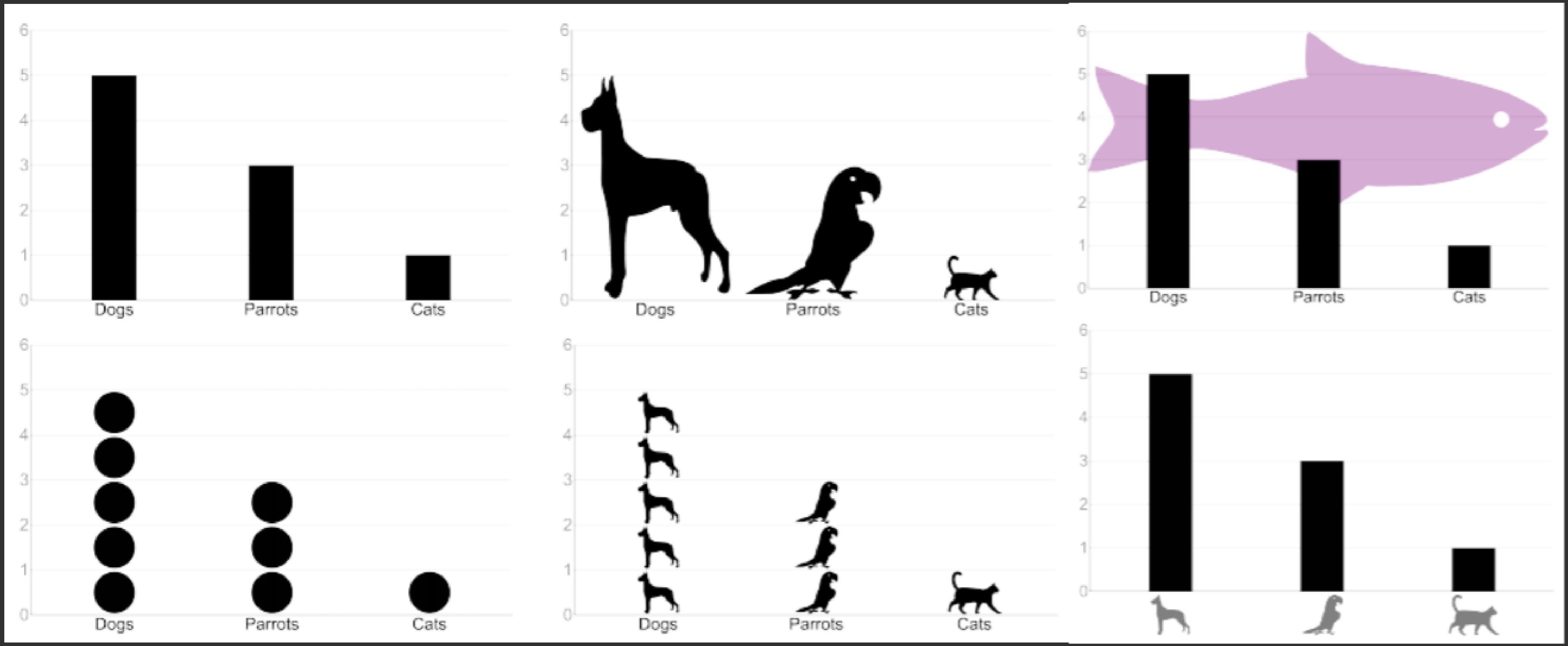


Useful chart junks?



Source: Useful Junk? The Effects of Visual Embellishment on Comprehension and Memorability of Charts, CHI'10.

Contextual representation can be helpful



Haroz et al. CHI'15

Summary on graphical representation

- Graphical Integrity
- Visual representations of data must tell the truth.
- Tufte shows a whole range of graphs that either over or underrepresent the effects in the data.
- He does this by calculating a graph's Lie Factor which can be calculated by dividing the size of the effect shown in the graphic by the size of the effect in the data.

- <u>If the Lie Factor is greater than 1 the graph overstates the effect</u>.

- Tufte goes on to list the following 6 principles of graphical integrity:
- The representation of numbers, as physically measured on the surface of the graph itself, should be directly proportional to the numerical quantities represented
- Clear, detailed and thorough labeling should be used to defeat graphical distortion and ambiguity. Write out explanations of the data on the graph itself. Label important events in the data.
- Show data variation, not design variation.
- In time-series displays of money, deflated and standardized units of monetary measurement are nearly always better than nominal units.
- The number of information carrying (variable) dimensions depicted should not exceed the number of dimensions in the data.
- Graphics must not quote data out of context.