Best Practices in Graphical Data Presentation

Ray Lyons
Independent Consultant
Cleveland OH, USA

Library Assessment Conference,
October 25 – 27, 2010 Baltimore MD, USA

Workshop Objectives

To gain an understanding of the basics of graphical data presentation theory and practice in order to create high quality quantitative graphs

To appreciate the usefulness of graphical methods in investigating data

To develop skills in recognizing clarity and objectivity in graphical presentation

Expert Sources

Cleveland, William S.

The Elements of Graphing Data, 1994, revised ed., Murray Hill, NJ: AT&T Bell Laboratories.

Tufte, Edward R.

The Visual Display of Quantitative Information, 1983, Cheshire, CT: Graphics Press.

Visual Explanations: Images and Quantities, Evidence and Narrative, 1997, Cheshire, CT: Graphics Press.

Expert Sources

Wainer, Howard

Visual Revelations: Graphical Tales of Fate and Deception from Napolean Bonaparte to Ross Perot, 1997, New York: Copernicus.

Graphic Discovery: A Trout in the Milk and Other Visual Adventures, 2005, Princeton, NJ: Princeton University Press.

Expert Sources

Wong, Dona M.

The Wall Street Journal Guide to Information Graphics: The Dos and Don'ts of Presenting Data, Facts, and Figures, 2010, New York: W.W. Norton and Co.

Few, Stephen

Now You See It: Simple Visualization Techniques for Quantitative Analysis, 2009, Oakland, CA: Analytics Press.

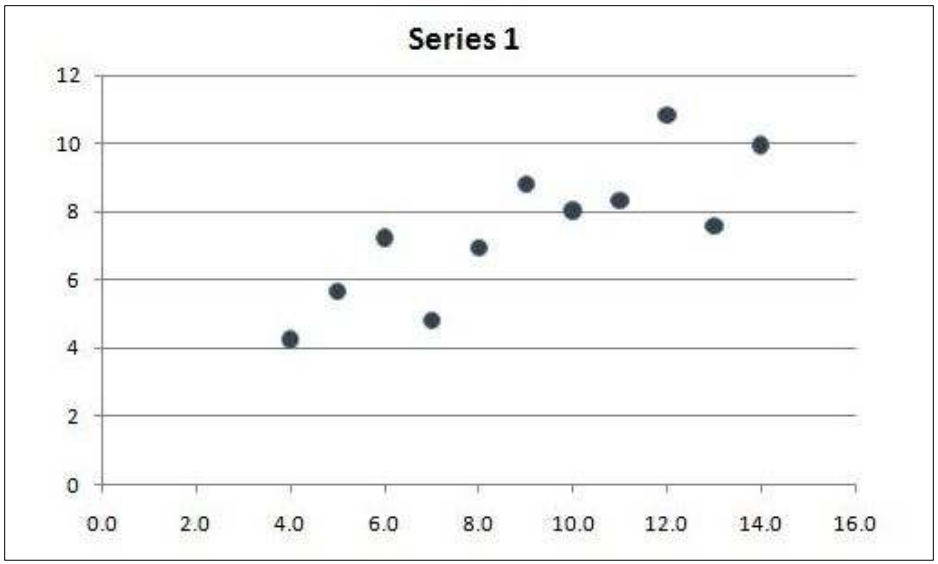
SERIES 1		SERIES 2		SERIES 3		SERIES 4	
X	Y	X	Y	X	Y	X	Y
10.0	8.04	10.0	9.14	10.0	7.46	8.0	6.58
8.0	6.95	8.0	8.14	8.0	6.77	8.0	5.76
13.0	7.58	13.0	8.74	13.0	12.74	8.0	7.71
9.0	8.81	9.0	8.77	9.0	7.11	8.0	8.84
11.0	8.33	11.0	9.26	11.0	7.81	8.0	8.47
14.0	9.96	14.0	8.10	14.0	8.84	8.0	7.04
6.0	7.24	6.0	6.13	6.0	6.08	8.0	5.25
4.0	4.26	4.0	3.10	4.0	5.39	19.0	12.5
12.0	10.84	12.0	9.13	12.0	8.15	8.0	5.56
7.0	4.82	7.0	7.26	7.0	6.42	8.0	7.91
5.0	5.68	5.0	4.74	5.0	5.73	8.0	6.89

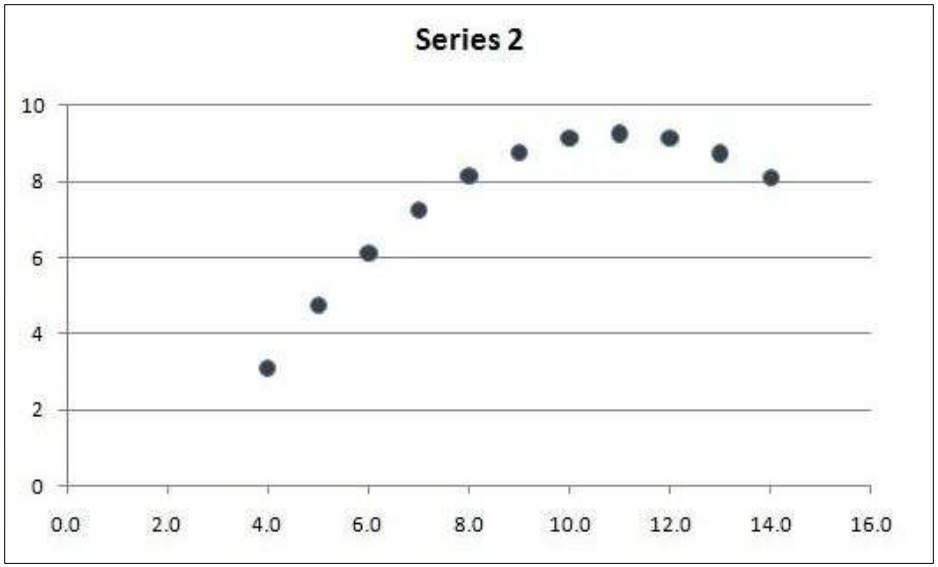
SERIES 1		SERIES 2		SERIES 3		SERIES 4	
X	Y	X	Y	X	Y	X	Y
10.0	8.04	10.0	9.14	10.0	7.46	8.0	6.58
8.0	6.95	8.0	8.14	8.0	6.77	8.0	5.76
13.0	7.58	13.0	8.74	13.0	12.74	8.0	7.71
9.0	8.81	9.0	8.77	9.0	7.11	8.0	8.84
11.0	8.33	11.0	9.26	11.0	7.81	8.0	8.47
14.0	9.96	14.0	8.10	14.0	8.84	8.0	7.04
6.0	7.24	6.0	6.13	6.0	6.08	8.0	5.25
4.0	4.26	4.0	3.10	4.0	5.39	19.0	12.5
12.0	10.84	12.0	9.13	12.0	8.15	8.0	5.56
7.0	4.82	7.0	7.26	7.0	6.42	8.0	7.91
5.0	5.68	5.0	4.74	5.0	5.73	8.0	6.89

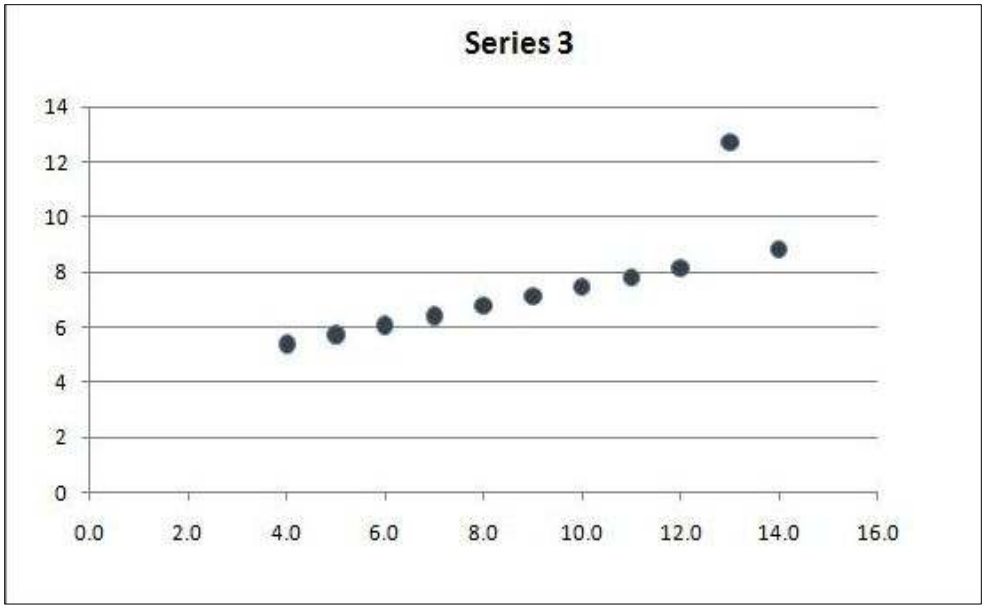
N = 11
Mean of X's= 9.0
Mean of Y's = 7.5

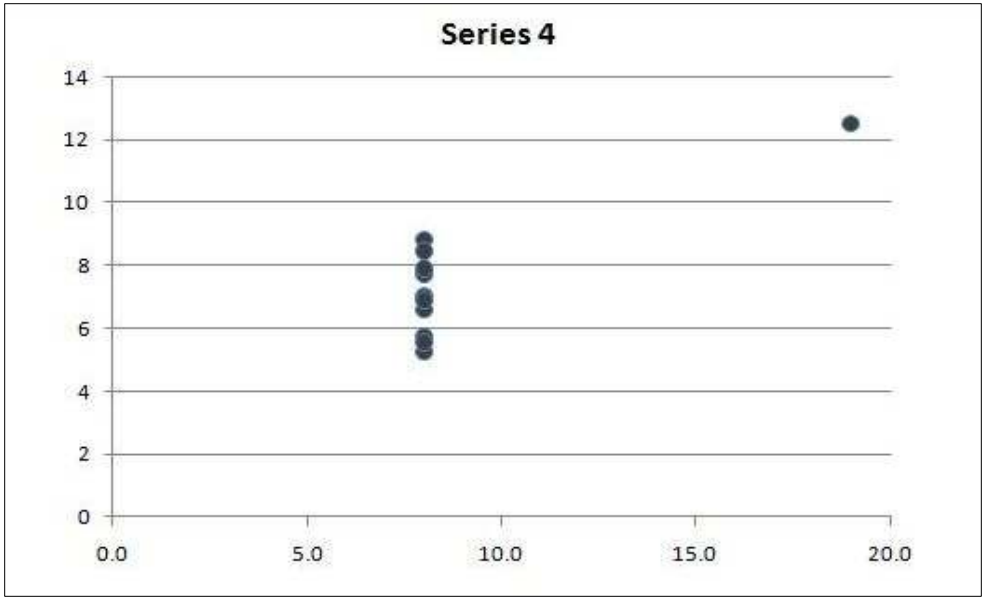
regression line: Y = 3 + 0.5X
sum of squares: 110.0
regression sum of squares = 27.5

correlation coeffecient = .82
R-square = .67


Series 1
12
10
8
6
4
2
0
0.0
2.0
4.0
6.0
8.0
10.0
12.0
14.0
16.0


Series 2
10
8
6
4
2
0
0.0
2.0
4.0
6.0
8.0
10.0
12.0
14.0
16.0


Series 3
14
12
10
8
6
4
2
0
0.0
2.0
4.0
6.0
8.0
10.0
12.0
14.0
16.0


Series 4
14
12
10
8
6
4
2
0
0.0
5.0
10.0
15.0
20.0

Why Graphical Data?

Visual perception more immediate than sequential scan of numbers and letters

Takes us from the specific and literal to the general and abstract (which is what we are most often interested in)

Visual arrangement of data can tell a story – in both senses of the word!

Why Graphical Data?

Portraying data graphically reveals patterns in the data that are difficult to detect otherwise

Visual depictions of data are almost universally understood without requiring knowledge of language

Graphical Excellence a la Edward Tufte *

"... is the well-designed presentation of interesting data—a matter of substance, of statistics, and of design.

... consists of complex ideas communicated with clarity, precision, and efficiency."

* Edward R. Tufte, 1983. *The Visual Display of Quantitative Information.*

More Graphical Excellence a la Edward Tufte

"... gives the viewer the greatest number of ideas in the shortest time with the least ink in the smallest space.

... is almost always multivariate.

... requires telling the truth about the data."

Tufte's Principles of Graphical Integrity

"Show data variation, not design variation."

Tufte calls superfluous graphical elements
"chartjunk"

No need to entertain or distract the reader

William Cleveland 's Principles of Data Graphics *

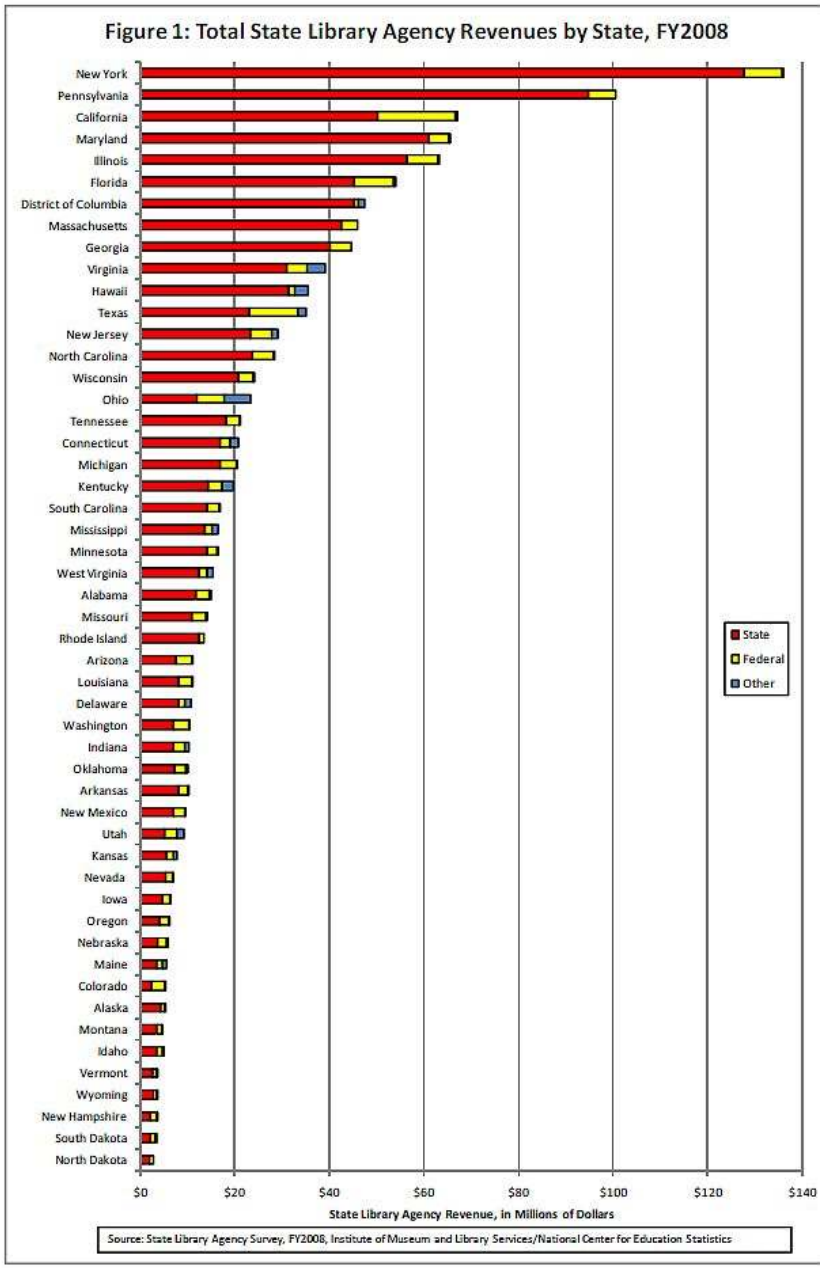
"Make the data stand out. Avoid superfluity."

"Use visually prominent graphical elements to show the data."

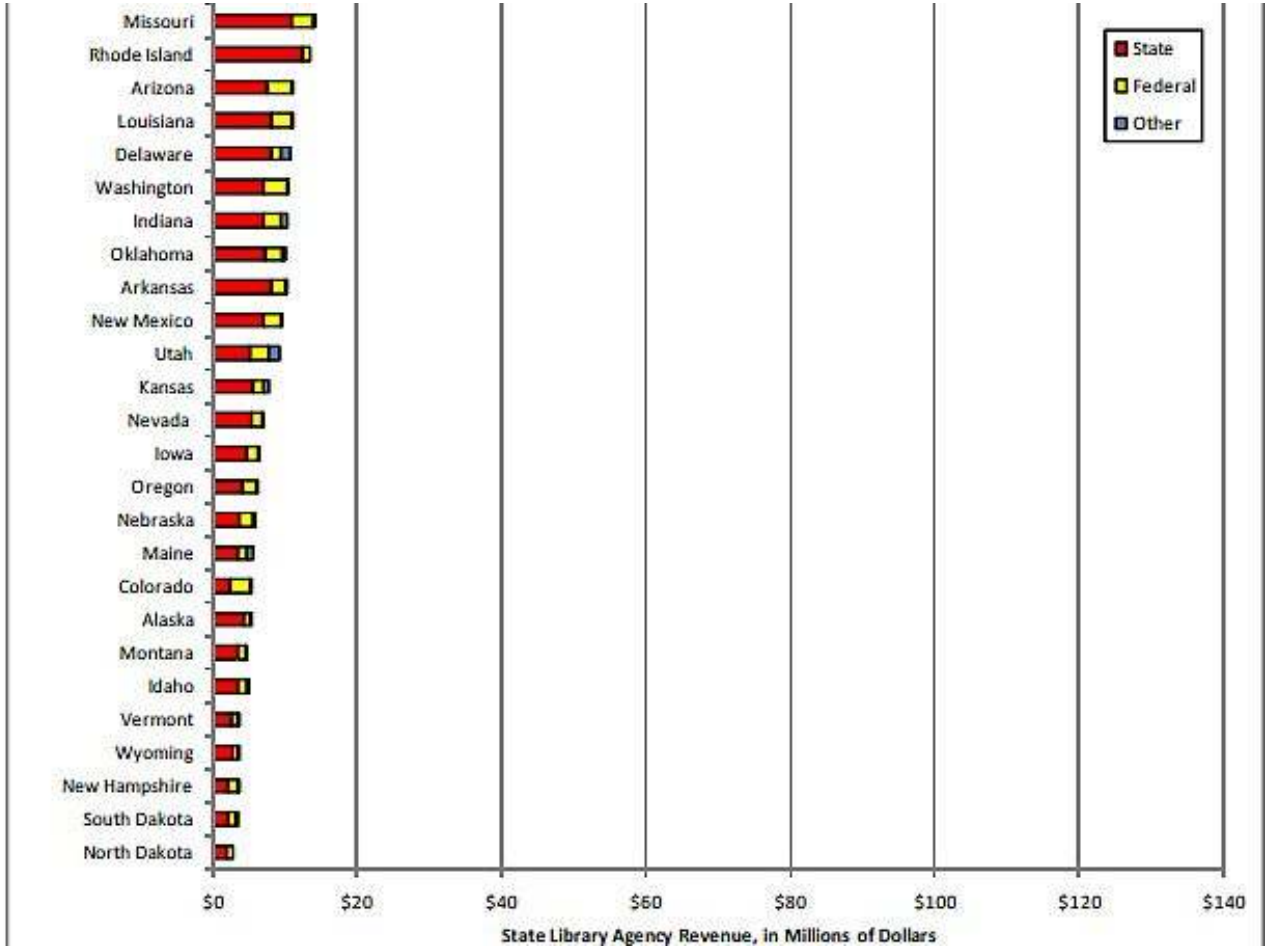
i.e. Do not obscure data by using ambiguous or overlapping symbols, inadequate spacing, poor labeling, and so on

* William S. Cleveland, 1994, *The Elements of Graphing Data.*

The purpose of graphical data presentation is to communicate information clearly and accurately.



Source: Institute of Museum and Library Services, 2010, *State Library Agency Survey Fiscal Year 2008*



Source: Institute of Museum and Library Services, 2010.

Bar segments in this chart excerpt are too small to decipher.

William Cleveland Principles of Data Graphics

"Strive for clarity."

"Proofread graphs."

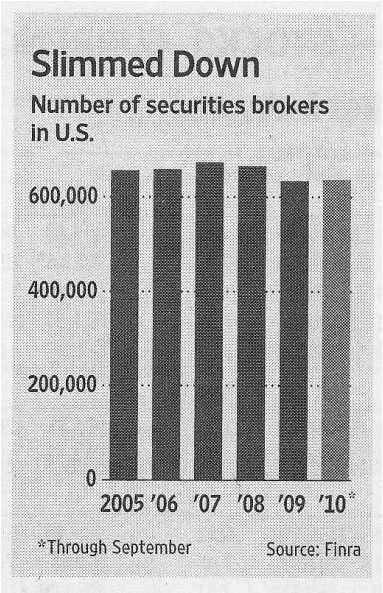
"Visual clarity must be preserved under reduction and reproduction. "

William Cleveland Principles of Data Graphics

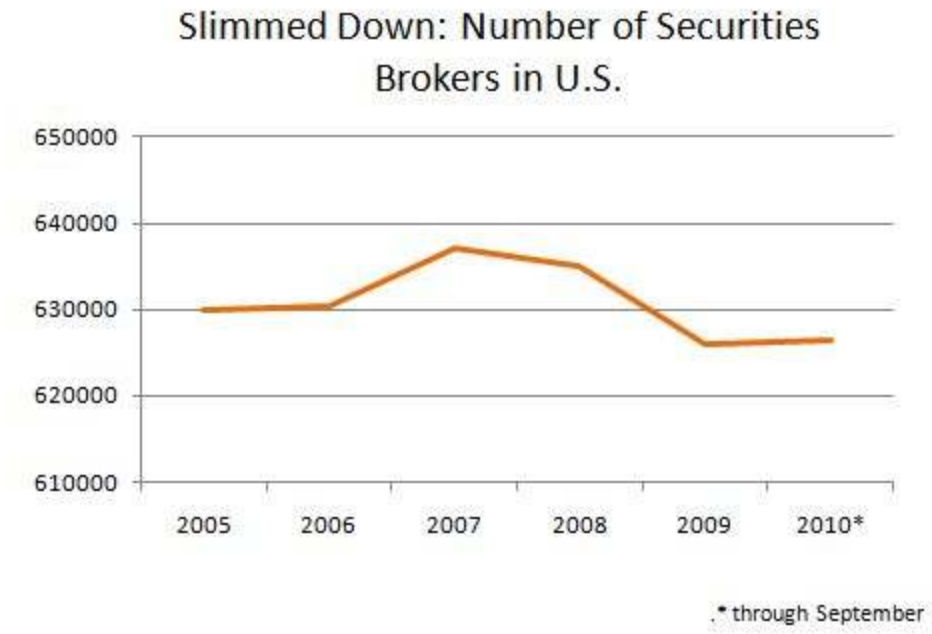
"Make captions comprehensive and informative."

"Draw attention to the important features of the data."

Source: Wall Street Journal, Oct. 2010



The excessive span of the vertical scale masks the trend. (And graphs in the Wall Street Journal are generally tiny, already!)



A line chart makes the trend more discernable.

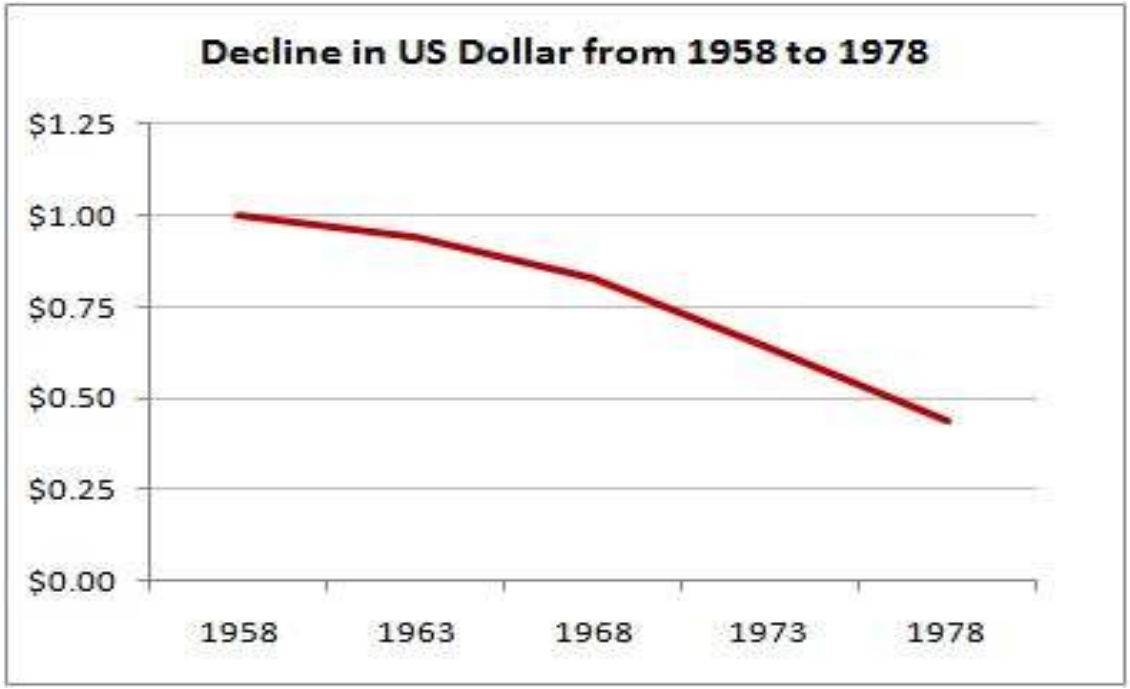
More of Tufte's Principles of Graphical Integrity

"Clear, detailed, and thorough labeling should be used to defeat graphical distortion and ambiguity."

"The representation of numbers, as physically measured on the surface of the graphic itself, should be directly proportional to the numerical quantities represented."

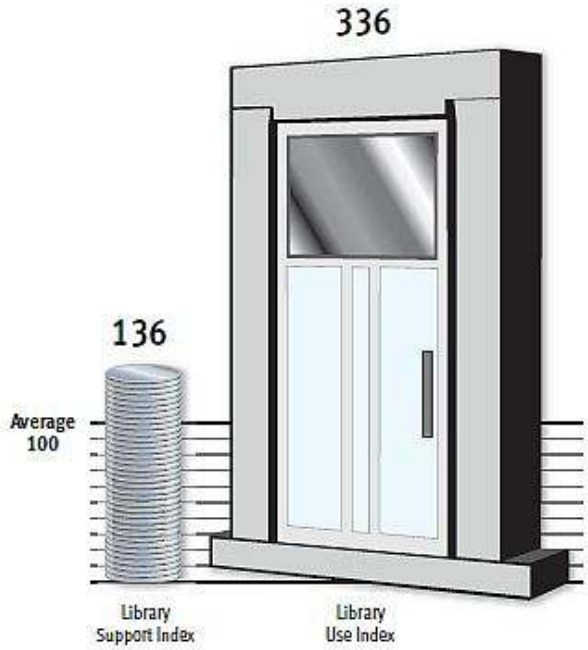
More of Tufte's Principles of Graphical Integrity

"The number of information-carrying (variable) dimensions depicted should not exceed the number of dimensions in the data."


Decline in US Dollar from 1958 to 1978
$1.25
$1.00
$0.75
$0.50
$0.25
$0.00
1958
1963
1968
1973
1978

The 1978 dollar should be twice as big as shown.

Source: Graphic reprinted in Edward Tufte, 1983.

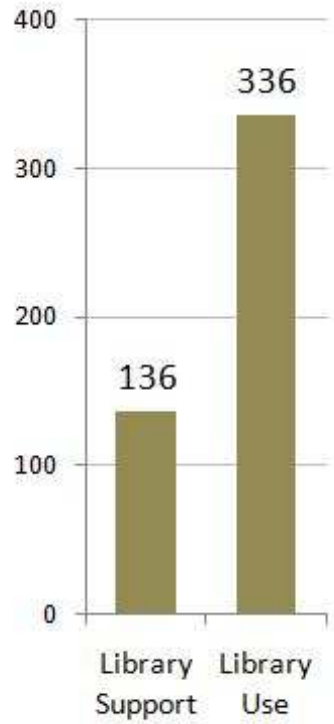


$$\text{Library Support Index} = \frac{\text{\% Definite Library Supporters}}{\text{\% population ages 18–69 in communities < 200,000}} \times 100$$

$$\text{Library Use Index} = \frac{\text{\% Library Visits}}{\text{\% population ages 18–69 in communities < 200,000}} \times 100$$

An index of 100 is average
Source: *From Awareness to Funding*, OCLC, 2008

From Awareness to Funding: A study of library support in America


400
300
200
100
0
336
136
Library Support
Library Use

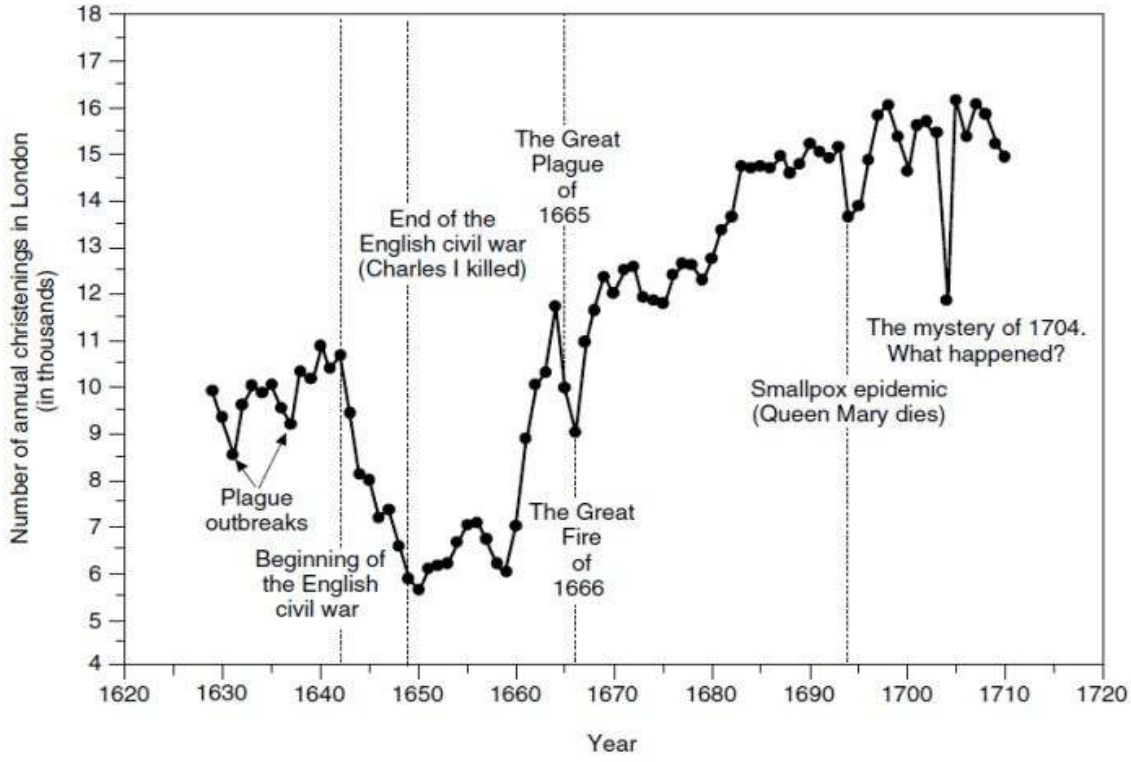
More of Tufte's Principles of Graphical Integrity

"Write out explanations of the data on the graphic itself."

"Label important events in the data."

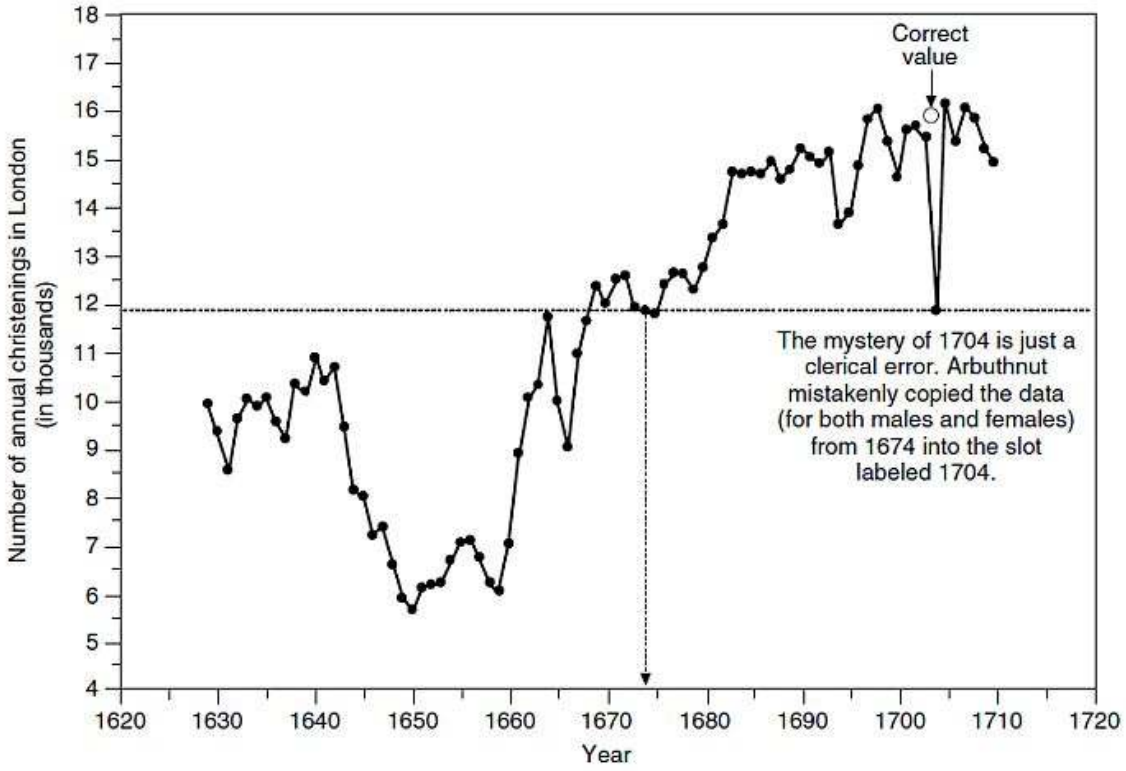
"Graphics must not quote data out of context."

Historical Example



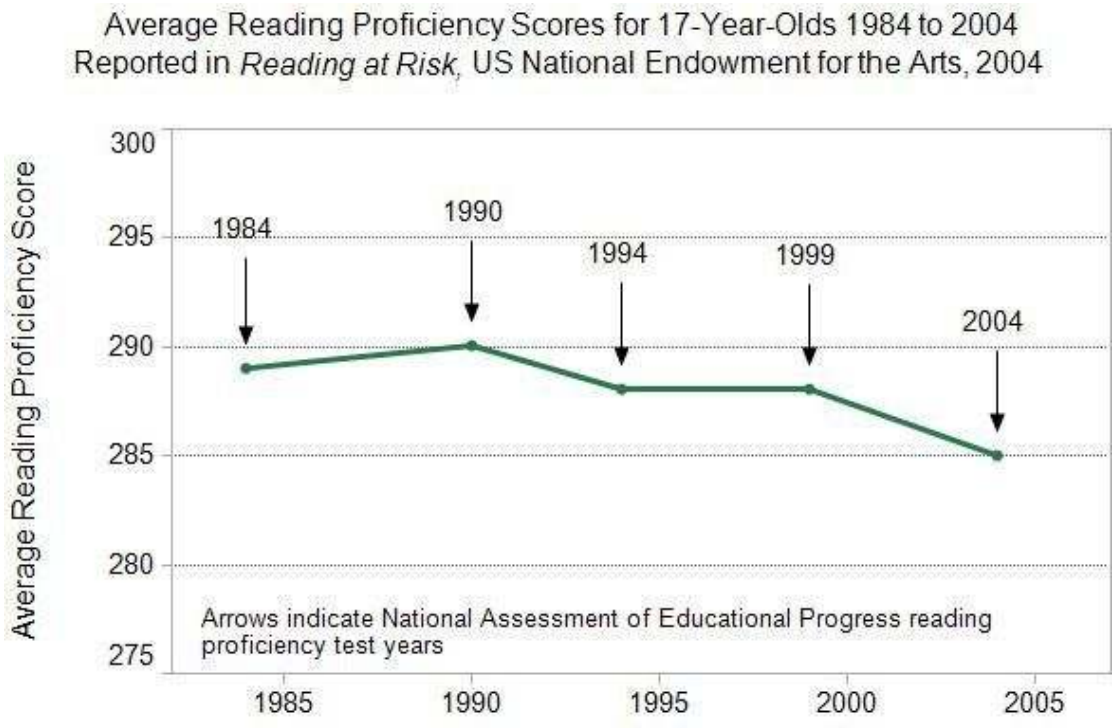
Source: Howard Wainer, 2005.

John Arbuthnot's 1710 analysis of London *Bills of Mortality* did not utilize graphical methods. Wainer's plot of the the data is a revelation.



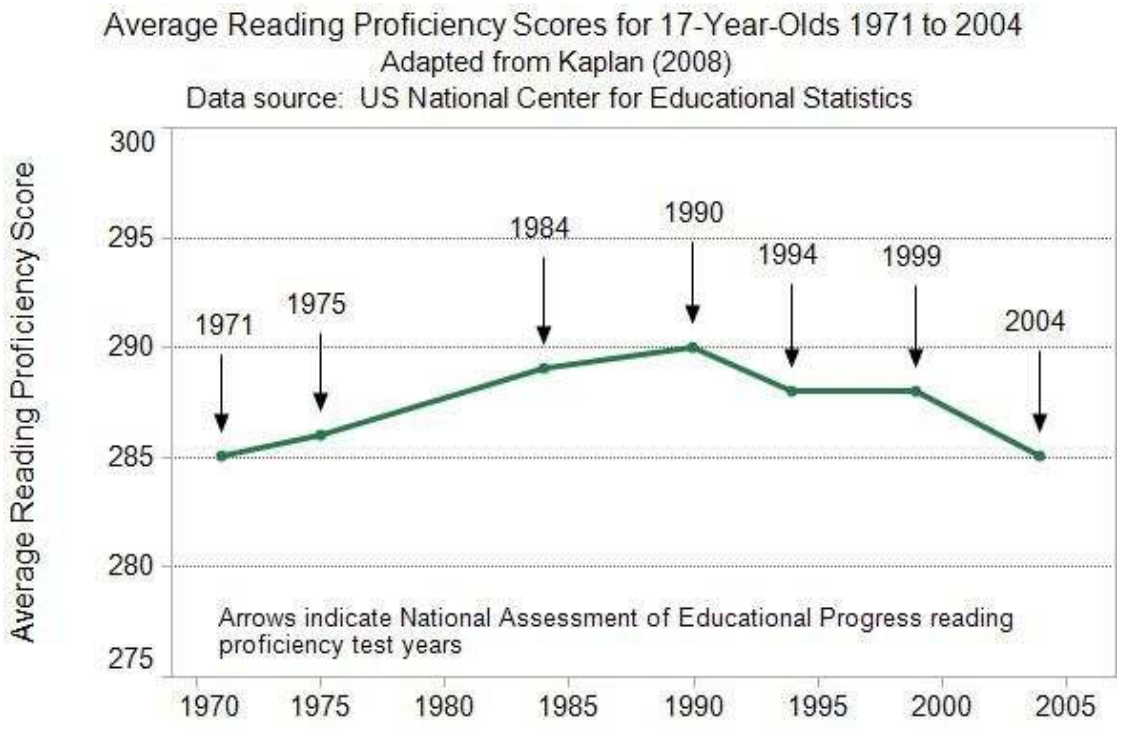
Source: Howard Wainer, 2009.

Rendering data graphically is a powerful investigative tool.



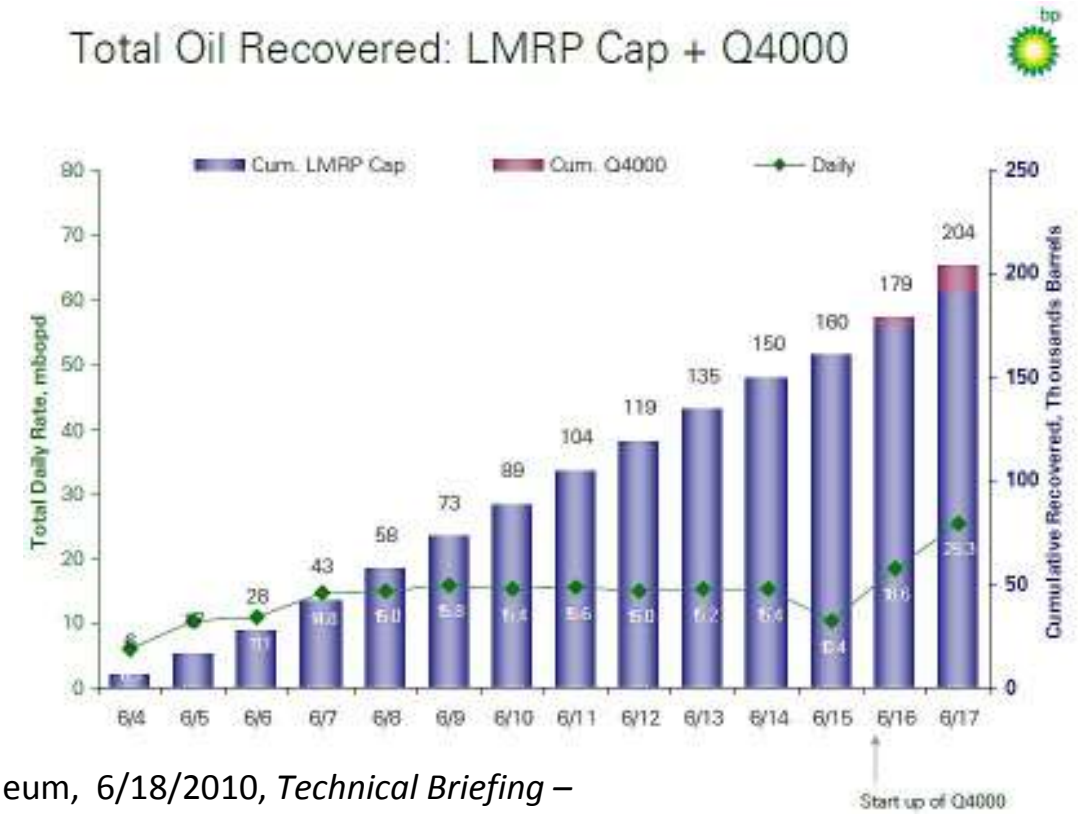
Data source: US National Center for Education Statistics

US National Endowment for the Arts report described this trend in read proficiency.



The story is quite different when the larger range of data is viewed.

See Nancy Kaplan, "To Read, Responsibly," *Public Library Quarterly*, 27, no. 3 (2008): 193-201.



Source: British Petroleum, 6/18/2010, *Technical Briefing – Kent Wells.*

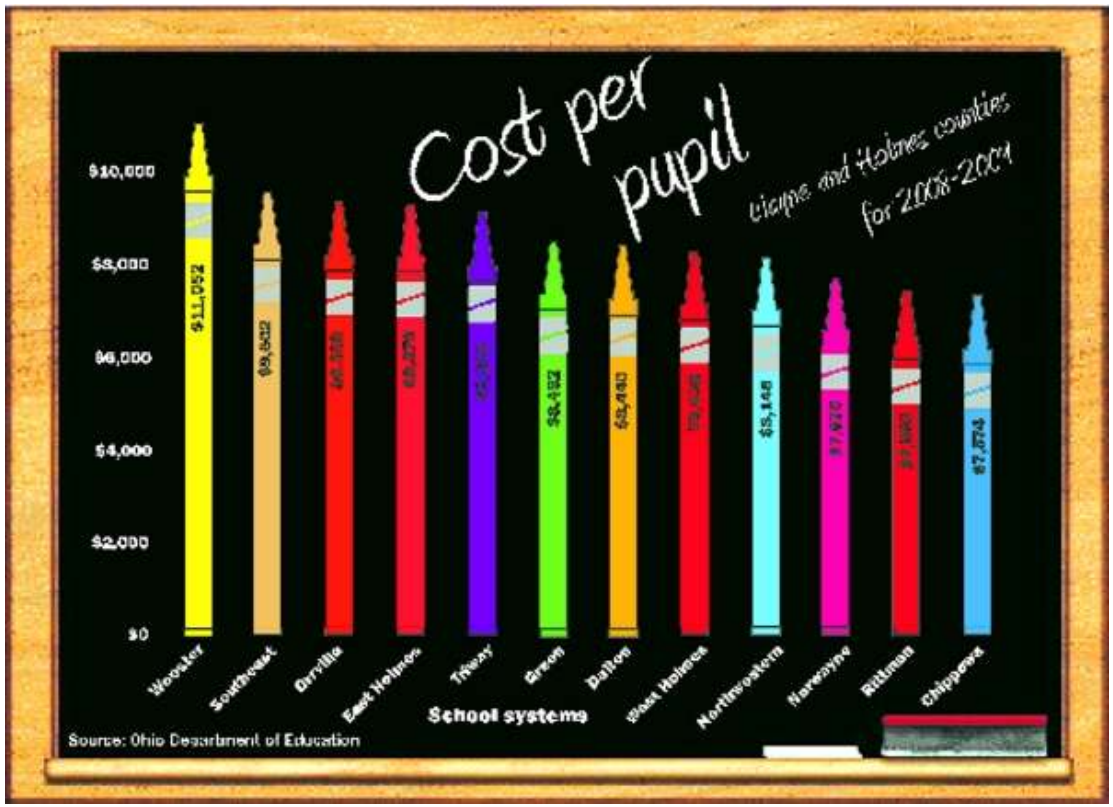
What is the context of these data? What magnitude of progress is shown? Compared to what?

Show the Data

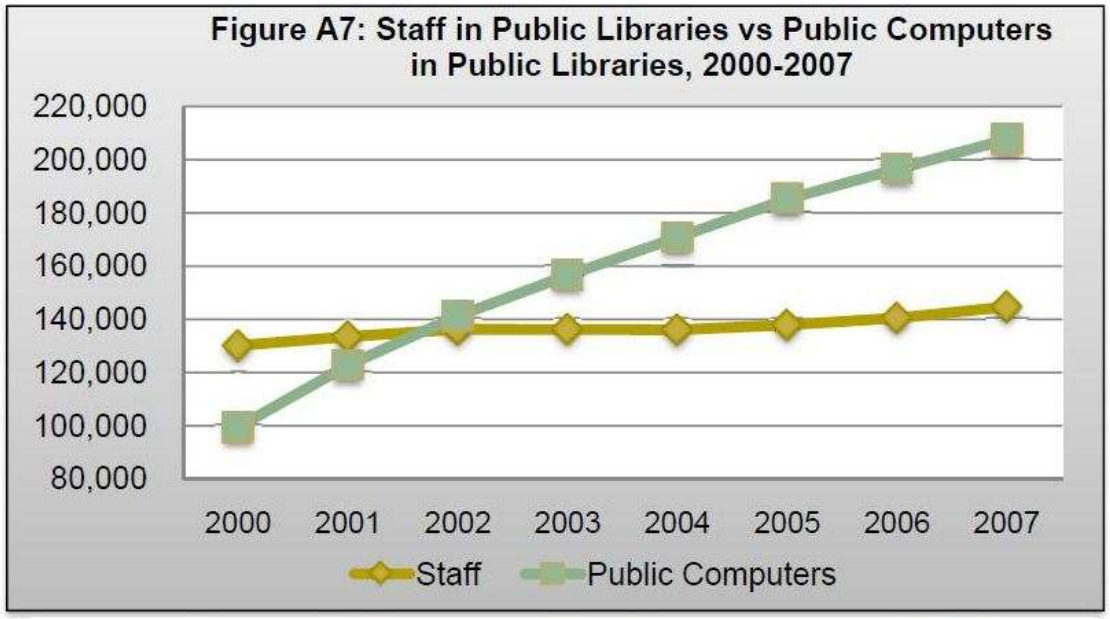
Each graphical element should aid communication of information

Avoid distractions that hinder visibility and ‘decoding’ (interpretation) of data

Source: Graphic reprinted in Edward Tufte, 1983.

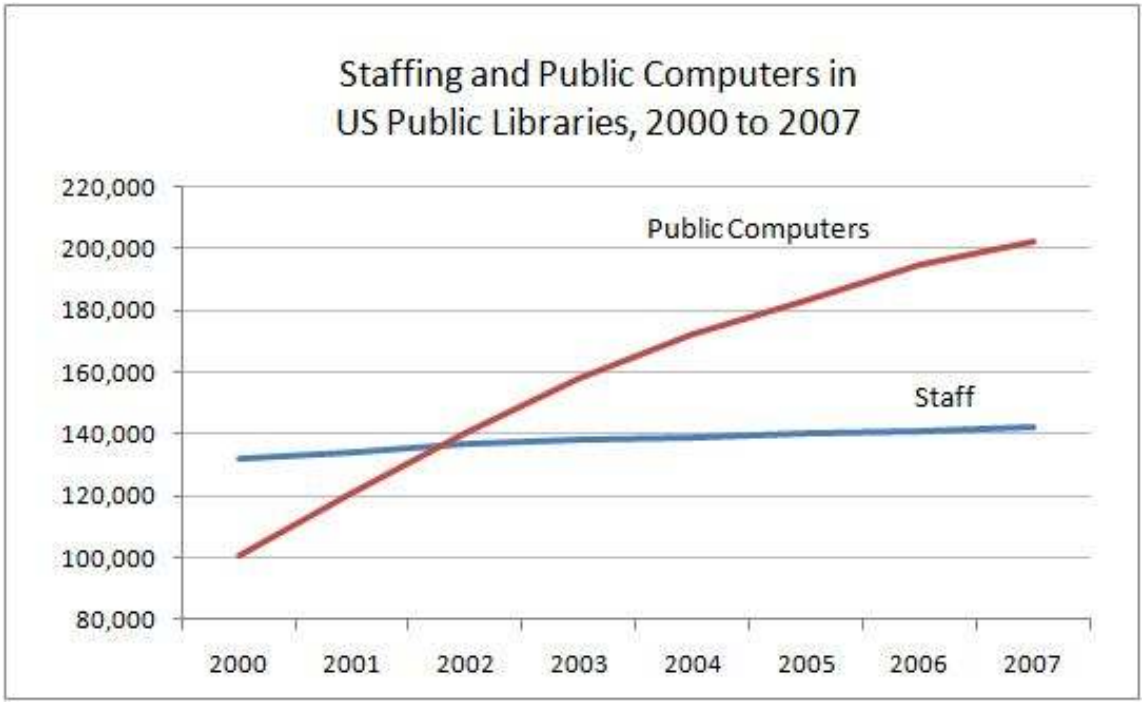


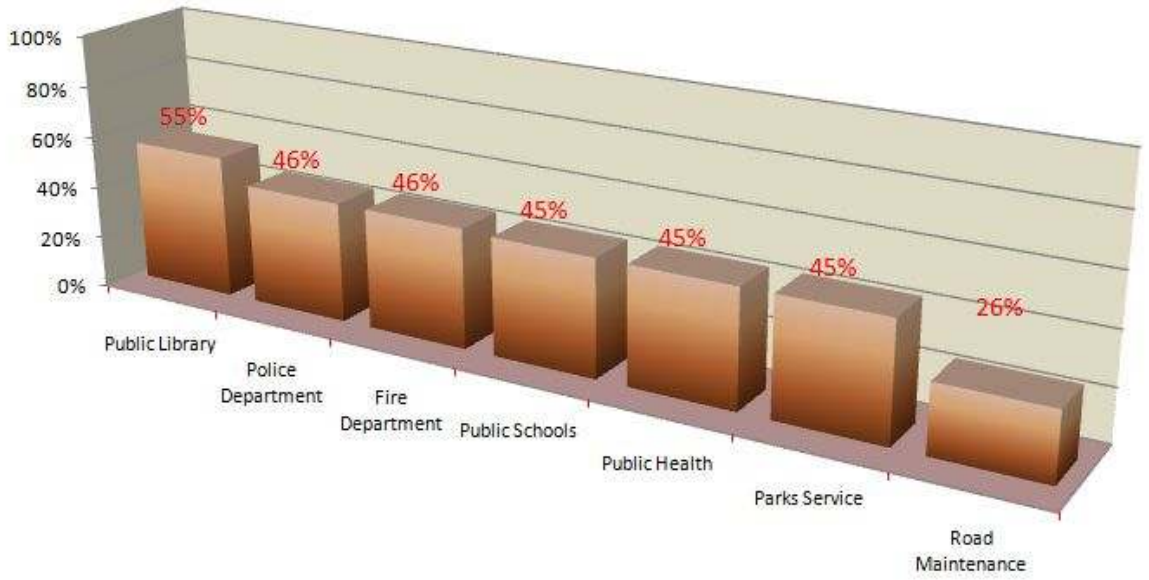
Source: An unnamed Ohio newspaper, 2010.



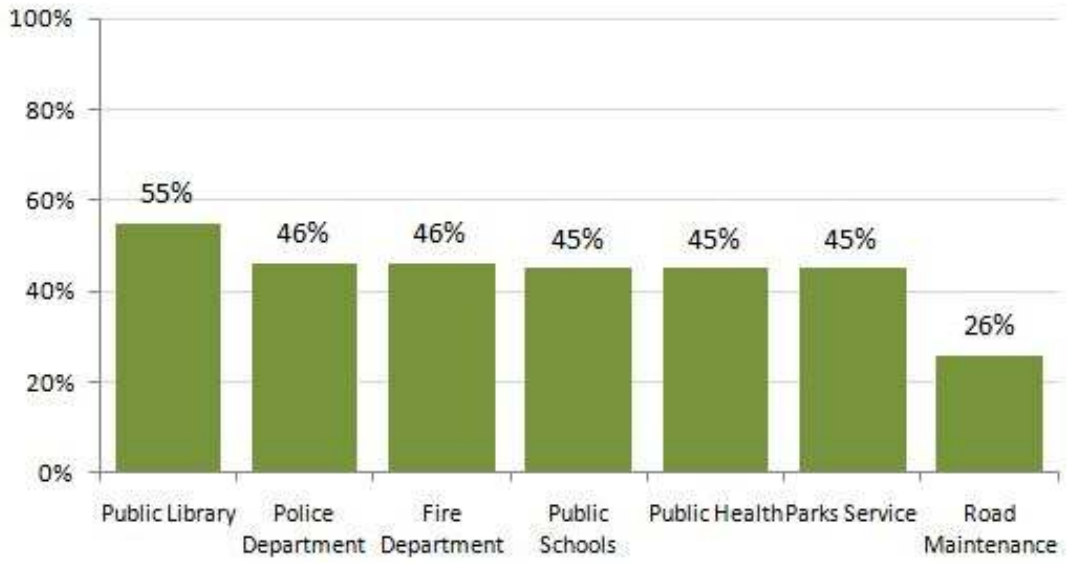
Source: American Library Association, 2010, *Perfect Storm.*

Line markers in this chart interrupt the visual flow of the trend lines. Gray squares are about 13,000 units tall.


Staffing and Public Computers in
US Public Libraries, 2000 to 2007
220,000
200,000
180,000
160,000
140,000
120,000
100,000
80,000
Public Computers
Staff
2000
2001
2002
2003
2004
2005
2006
2007



Never use 3-D effects. Never.



Clarity, clarity, clarity.

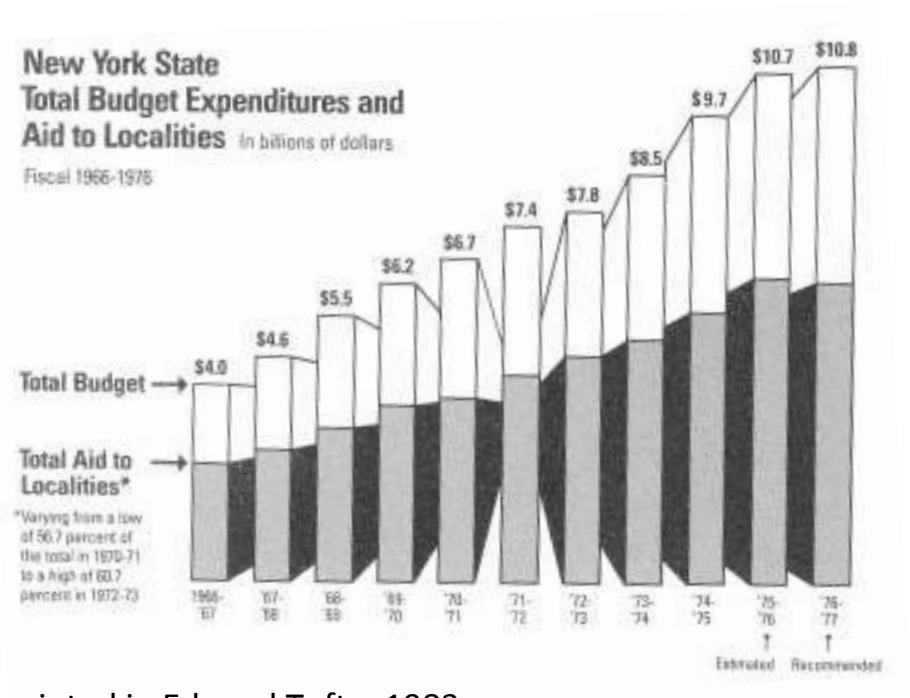
Edward Tufte's
Graphical Efficiency Measures

Data-Ink Ratio =

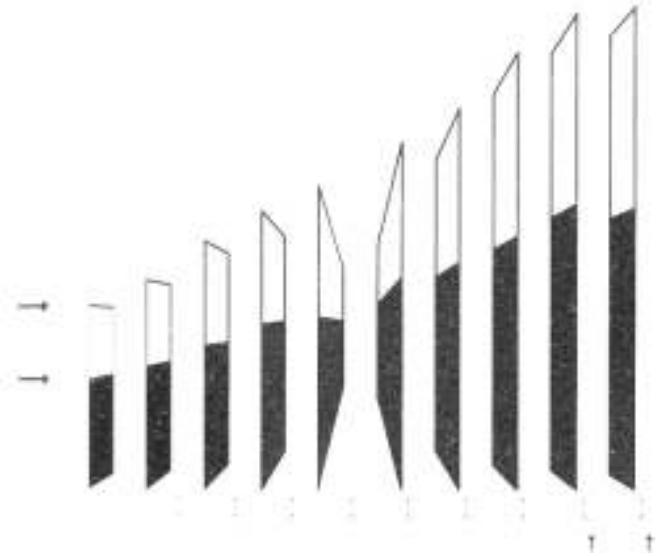
$$\frac{\text{Ink used portraying data}}{\text{Total ink used}}$$

= proportion of a graphic's ink devoted to the non-redundant display of information

= 1.0 – proportion of a graphic that can be erased without loss of information

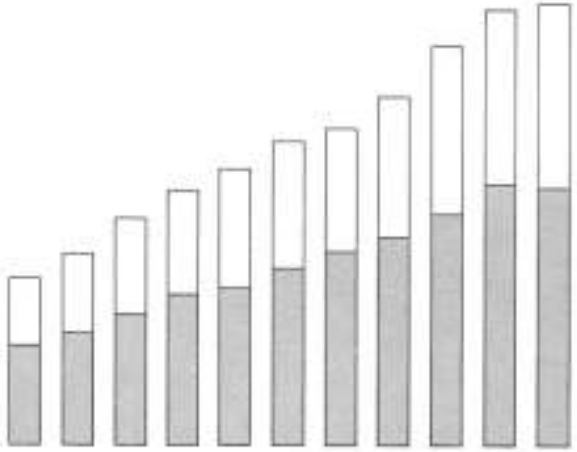


Source: Graphic reprinted in Edward Tufte, 1983.

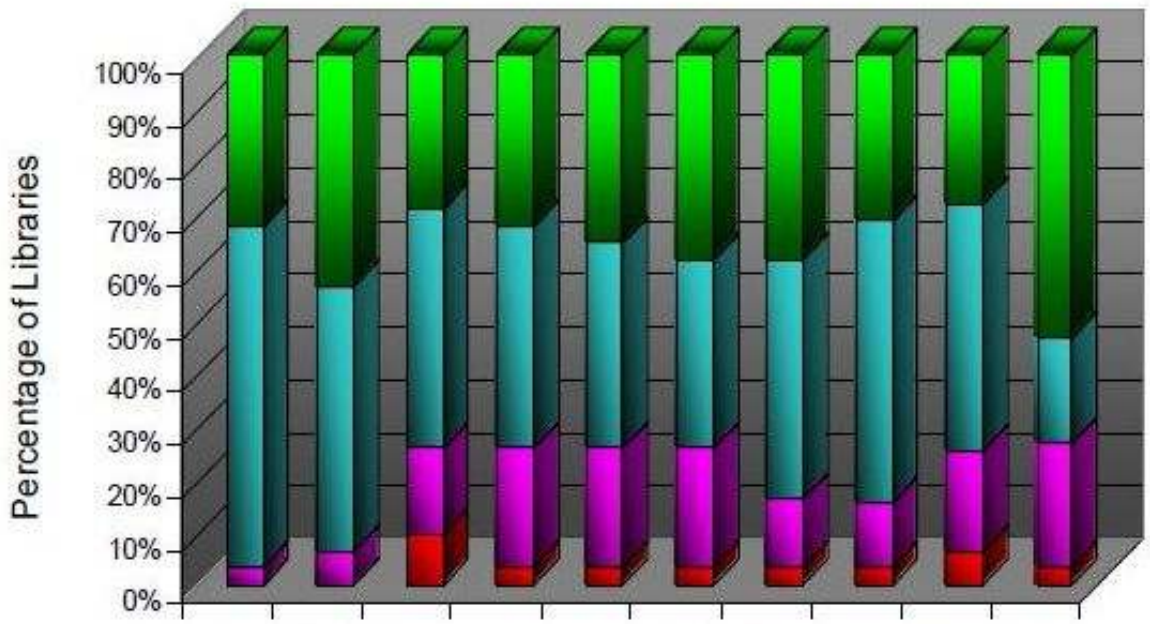


Redundant or superfluous

Source: Edward Tufte, 1983.



Non-redundant information



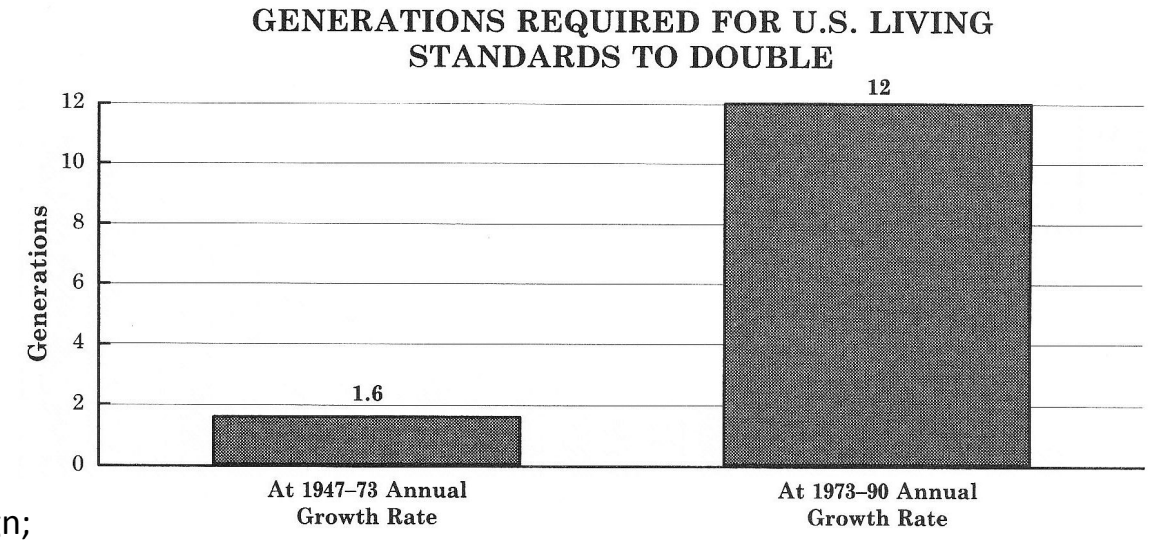
With two-dimensional stacked bars, segment heights are difficult to evaluate. Three-dimensional bars are worse. The luminescent shading in this chart is useless.*

*I can say so because I designed this graph for a 2008 survey report.

Edward Tufte's Two Graphical Efficiency Measures

Data Density =

$$\frac{\text{Number of data values displayed}}{\text{Total area of graph}}$$



Source: Ross Perot's presidential campaign; reprinted in Howard Wainer, 1997.

Growth Rate	Generations Required to Double US Standard of Living
1947 – 1973	1.6
1973 – 1990	12

Tufte's Advice

"Above all else show the data.

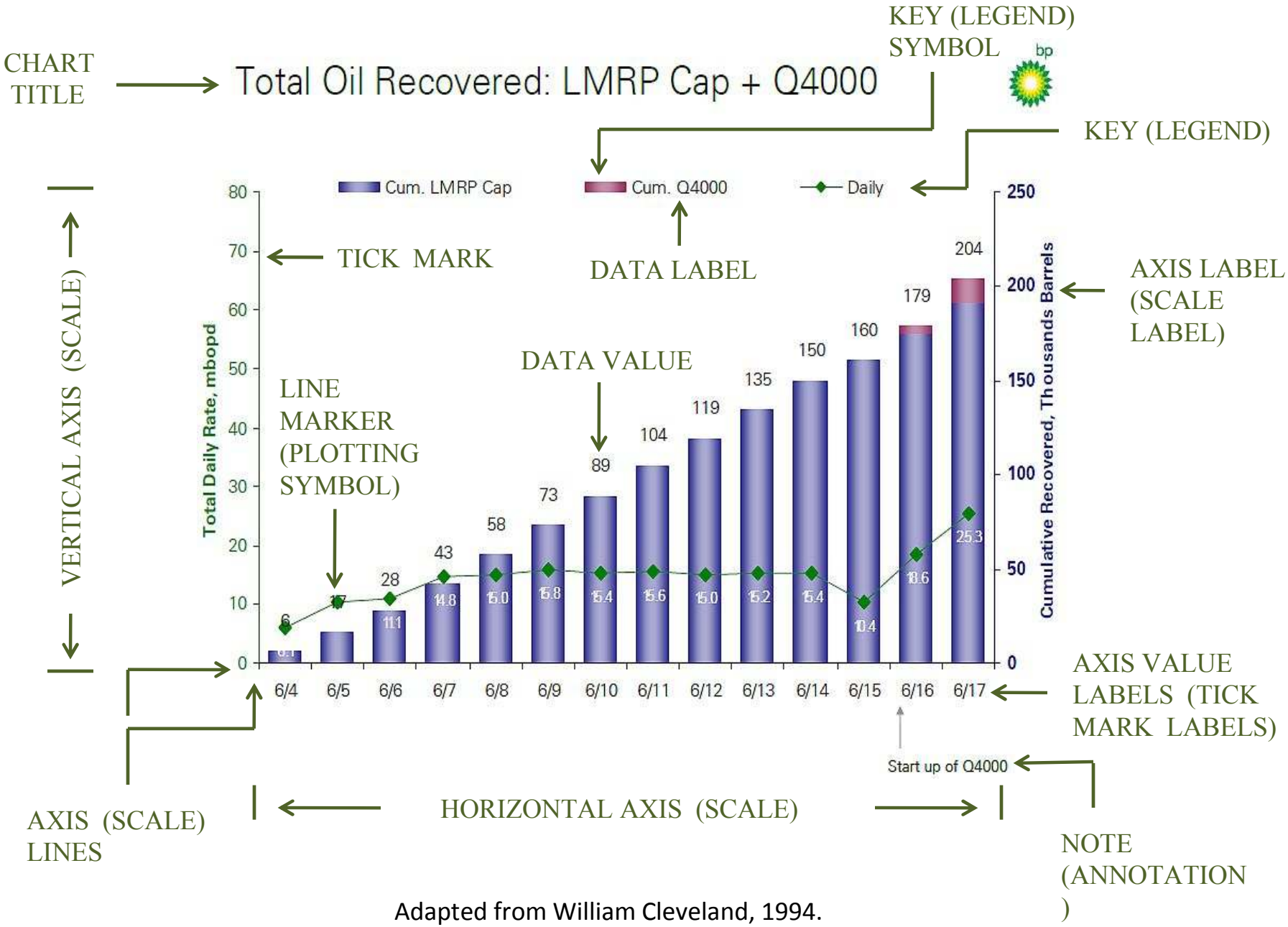
Maximize the data-ink ratio.

Erase non-data ink.

Erase redundant data-ink.

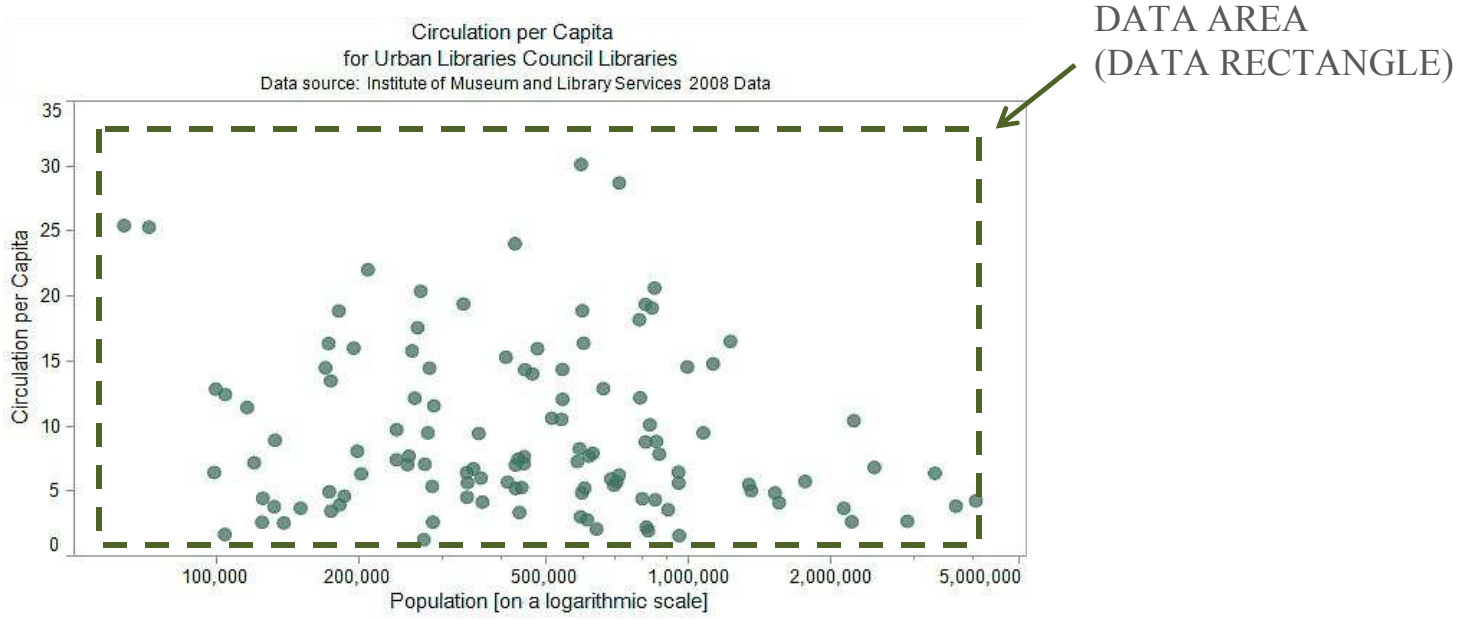
Revise and edit."

Components of a Chart



Adapted from William Cleveland, 1994.

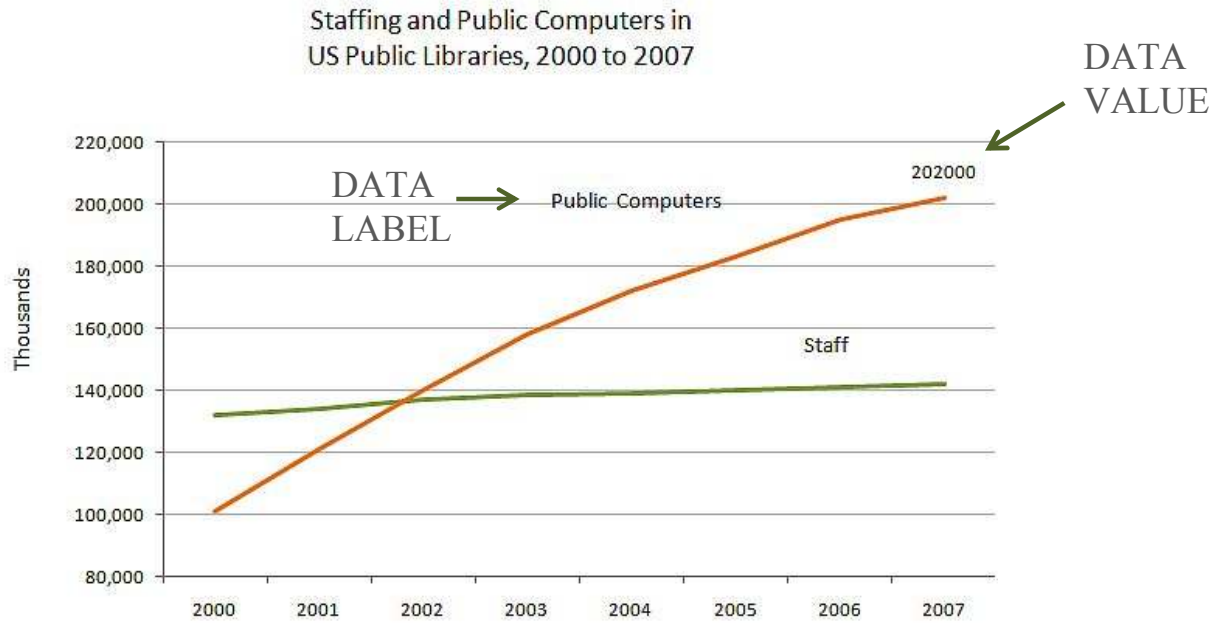
Components of a Chart



The *data area* or *data rectangle* is the area inside the boundary of the axes where the data are charted.*

*MS Excel calls this the *plot area;* Excel calls the area outlined by the axes the *chart area.*

Components of a Chart



A *data label* identifies the name of the variable or series plotted. A *data value* is a number indicating a specific value in the data.

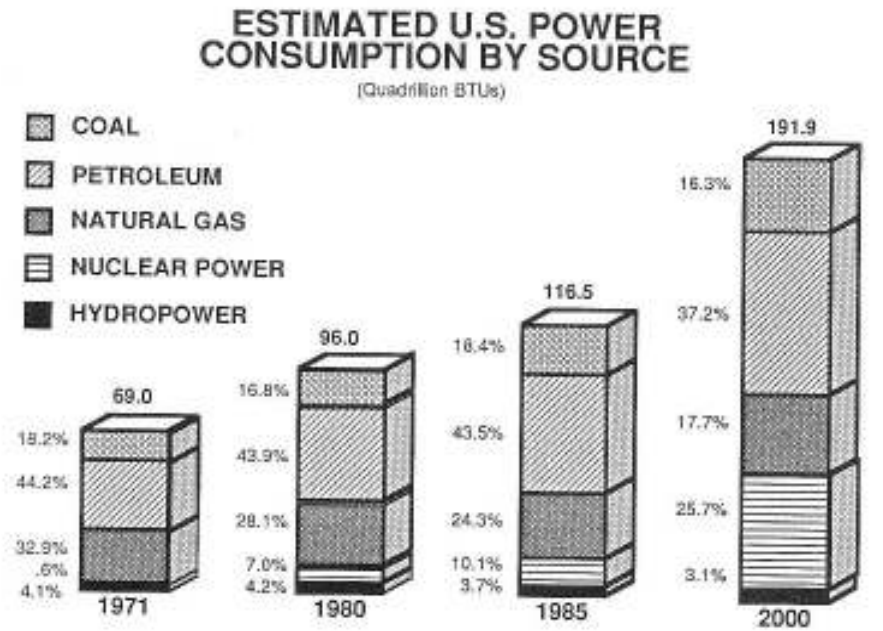
Components of a Chart

"Scale" has two meanings in graphical construction:

1. The line and associated markings representing the magnitude of the data (e.g. tick marks arranged along the length of a thermometer)
2. The range (span or extent) of the values depicted by the axes

Exercise: Deciphering Charts

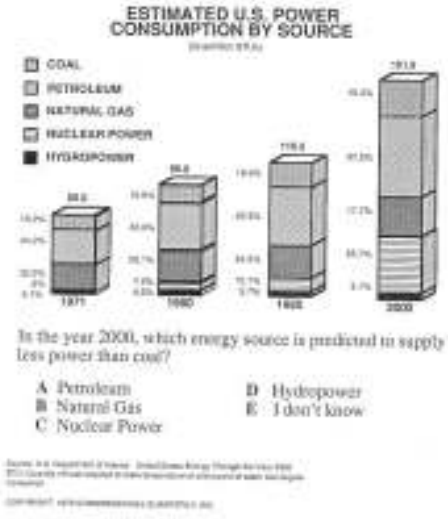
Source: Test question reprinted in Howard Wainer, 1997.



In the year 2000, which energy source is predicted to supply less power than coal?

A Petroleum
B Natural Gas
C Nuclear Power
D Hydropower
E I don't know

Source: U.S. Department of Interior United States Energy Through the Year 2000
BTU: Quantity of heat required to raise temperature of one pound of water one degree Fahrenheit

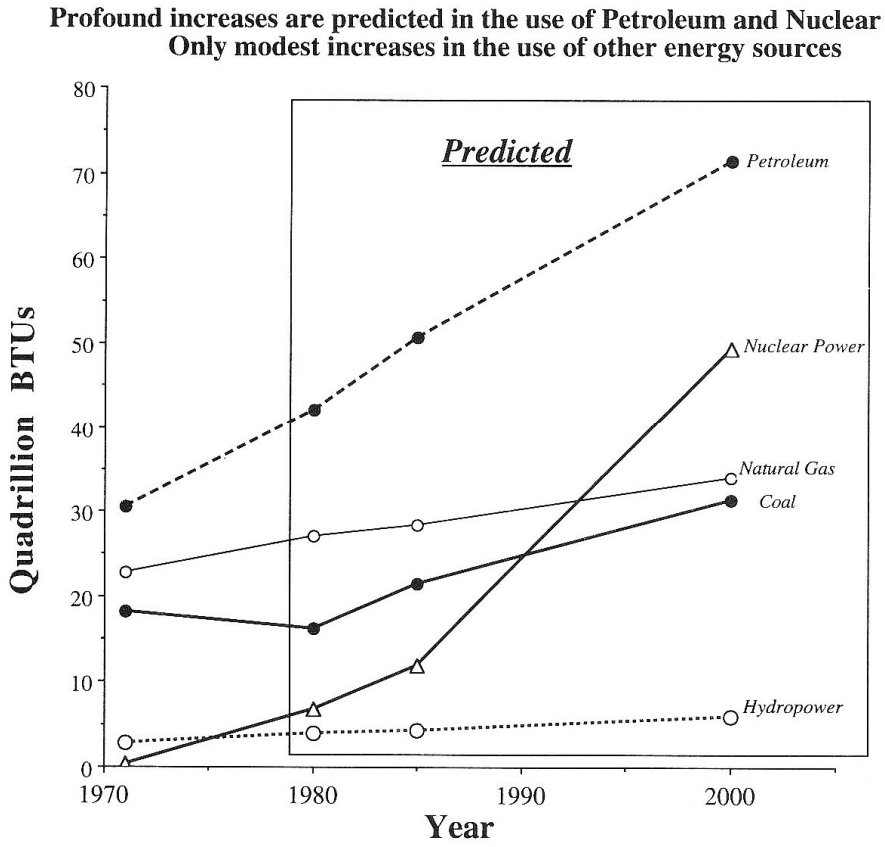


1. Answer the question appearing below the 3-D stacked bars in the chart to the left. Make notes of each step you follow: List each item of information you seek in the order you seek it, and where you find it—if you do find it. Also note whether each item of information you seek could or could not be found in the chart.

2. Answer the following question, keeping the same notes you did for item #1:

   How does hydropower use in 1971 compare to its predicted use in 2000?

3. Did the bar images help or hinder your locating the information needed? In what ways?



Source: Howard Wainer, 1997.

Wainer's re-draw of US energy consumption chart makes trends obvious and comparisons straightforward.

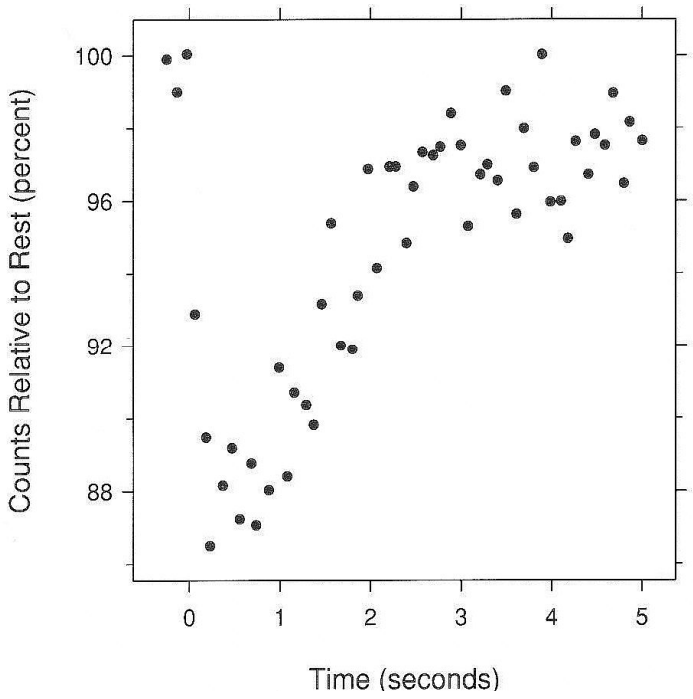
Axes (Scales)

Usually two axes (scales) are sufficient, though William Cleveland often uses axes on all four sides of his charts

When using two scale lines use visible but non-intrusive gridlines

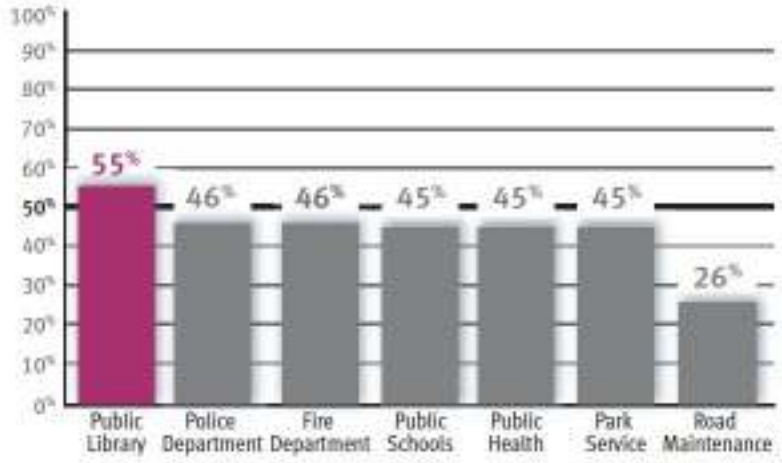
Make the data area (rectangle) slightly smaller than the chart rectangle.

Source: William Cleveland, 1994.



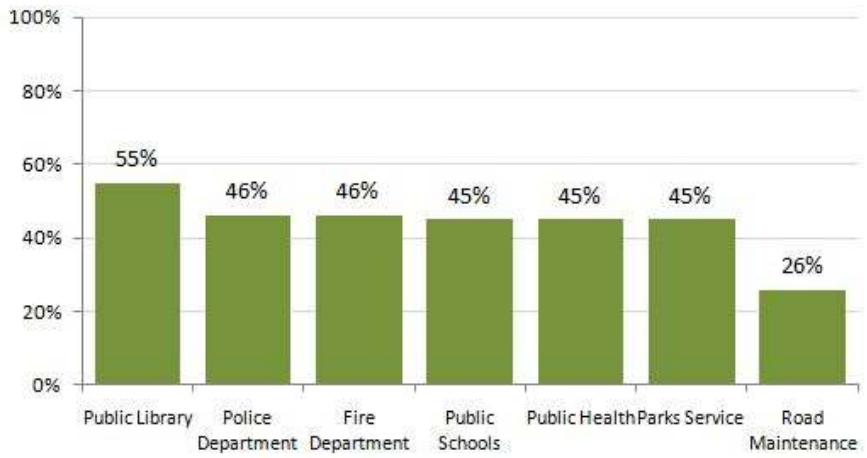
2.13 SCALE LINES AND THE DATA RECTANGLE. The four scale lines provide a clearly defined region for our eyes to look for data. Now, none of the data from Figure 2.12 are in danger of being overlooked.

William Cleveland suggests four scale lines. His captions are precise and thorough.

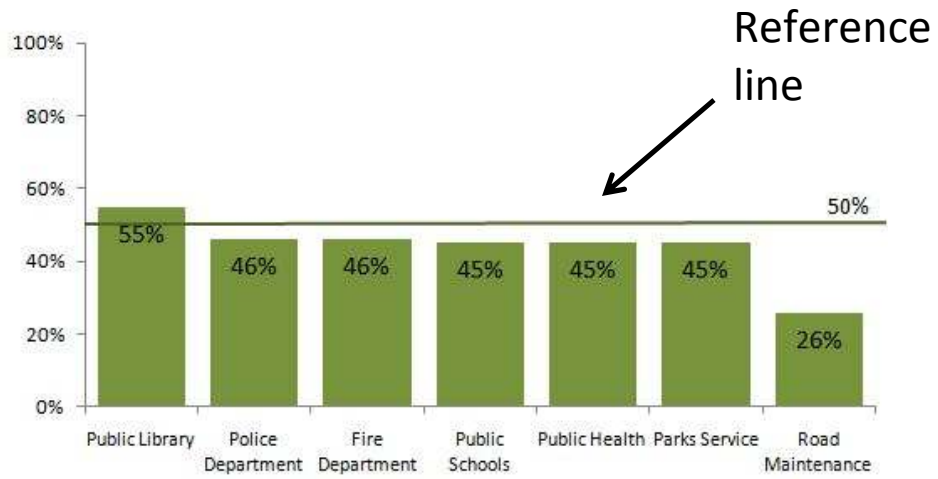


Percentage of voting respondents with an agreement rating of 8, 9 or 10
Source: *From Awareness to Funding*, OCLC, 2008

Gridlines are useful, but distracting when too heavy.



Use fewer and lighter gridlines, or no gridlines



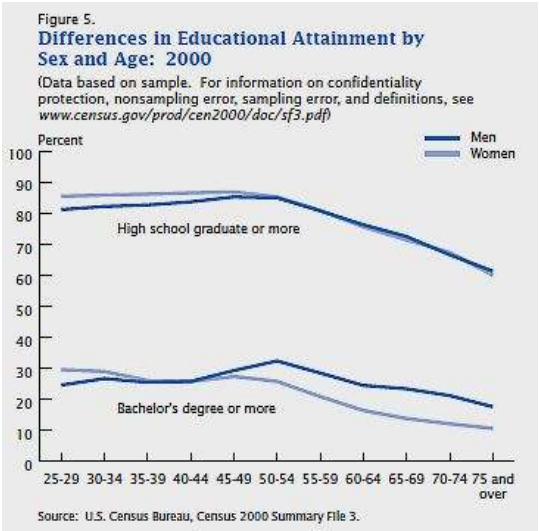
When using a *reference line** use light or no gridlines.

* A *reference line* is a line used to highlight a particular value or region of the chart that is notable, such as a benchmark value.

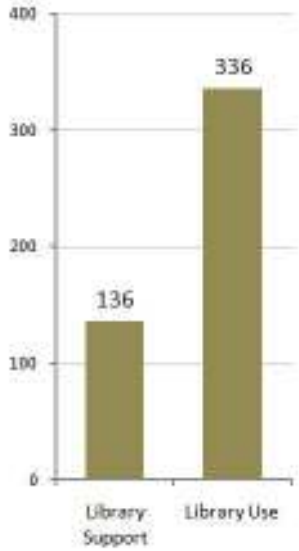
Axes (Scales)

Point tick marks outward, away from data rectangle

Do not overdo the number of tick marks

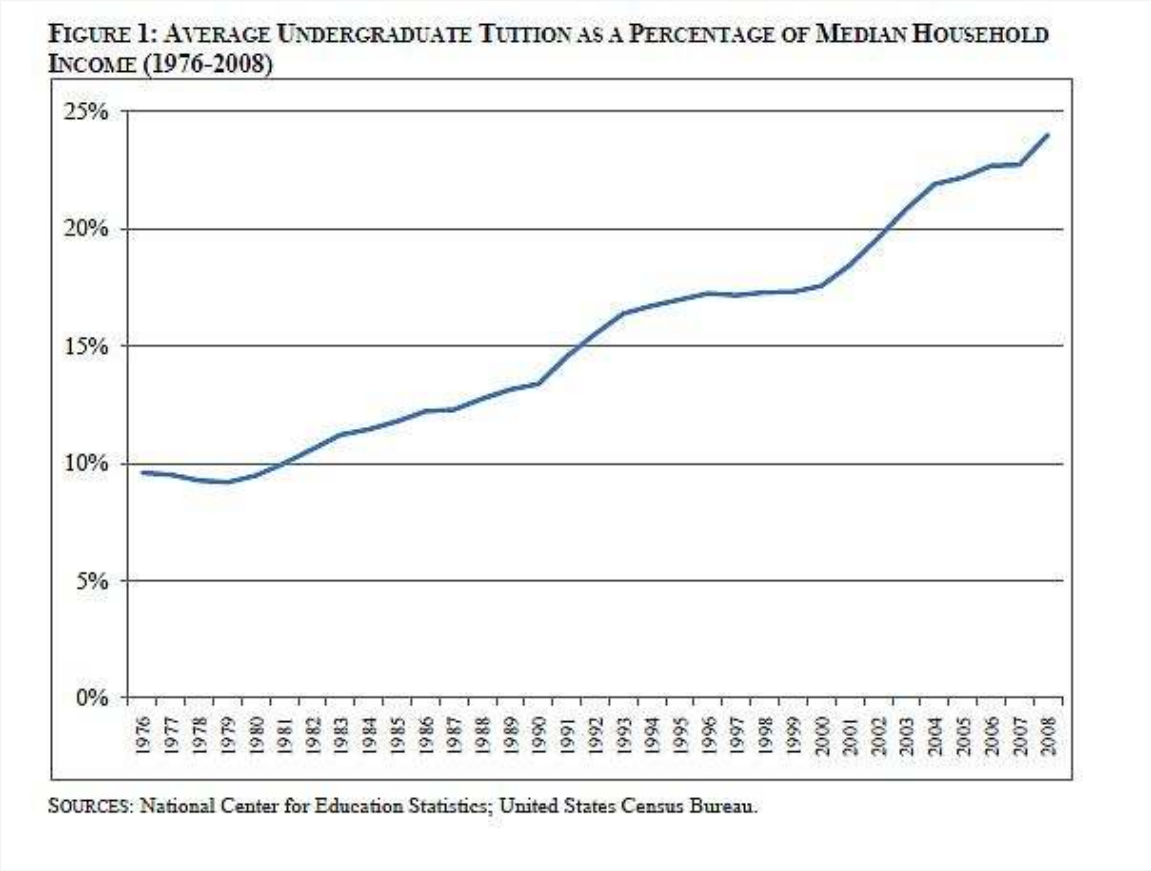


Axis tick marks inside the data area can interfere with plotted data.



Axis tick marks should point outward, away from the data area.

Source: Richard Vedder, 2010, *The Coming Revolution in Higher Education.*



Crowded and repetitious tick mark values are nearly unreadable.

Axes (Scales)

User familiar numbering increments:

0, 1, 2, 3, 4, 5 . . .
0, 2, 4, 6, 8, 10 . . .
0, 5, 10, 15 . . .
0, 10, 20, 30 . . .
0, 25, 50, 75 . . .
0, 0.2, 0.4, 0.6 . . .
0, 0.25, 0.50, 0.75 . . .

Axes (Scales)

Use highest reasonable units:

Use	Instead of:
$3 million or 3M	$3,000 thousand
2.6 billion or 2.6B	2,582,000,000
1K, 2K, 3K	1000, 2000, 3000

Don’t make readers do the math

Axes (Scales)

Always include zero on bar charts*

Zero optional on line, scatter plots, and box plots (box-and-whiskers plots)

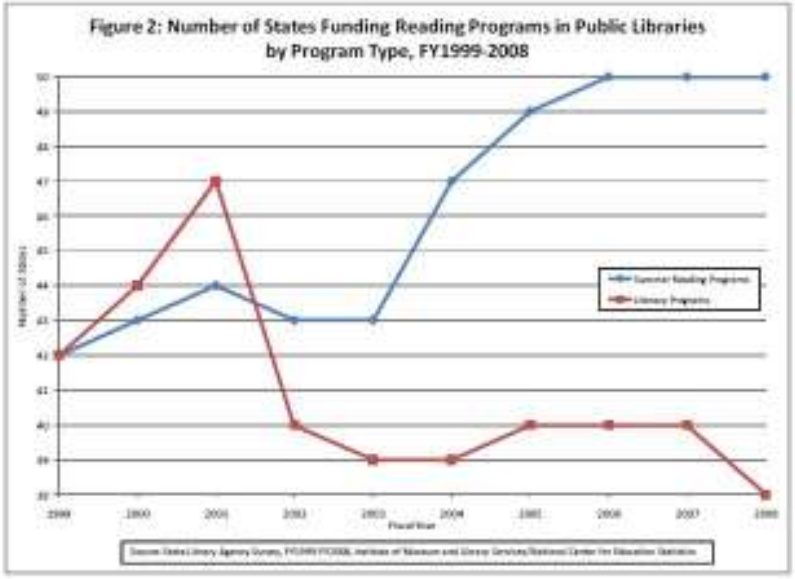
* Unless there is a compelling reason ~~not~~ to exclude zero. In this case, add a jagged line to bottom of each bar indicating it has been truncated.

Determining Scaling

Scale span must encompass all of the data (except outliers justifiably excluded)

Begin the scale at an even number, or multiple of 2, 5, or 10

If data values are close to zero (depending on data range), include zero on the axes



Source: Institute of Museum and Library Services, 2010, *State Library Agency Survey Fiscal Year 2008*

Make sure span of axis values exceeds the data values. Do not plot data on the chart axis lines.

Chart Orientation

To Edward Tufte, horizontal displays are preferable since:

> "Our eye is naturally practiced in detecting deviations from the horizon…"

Ease of labeling. More space for labeling

Emphasis on causal influence

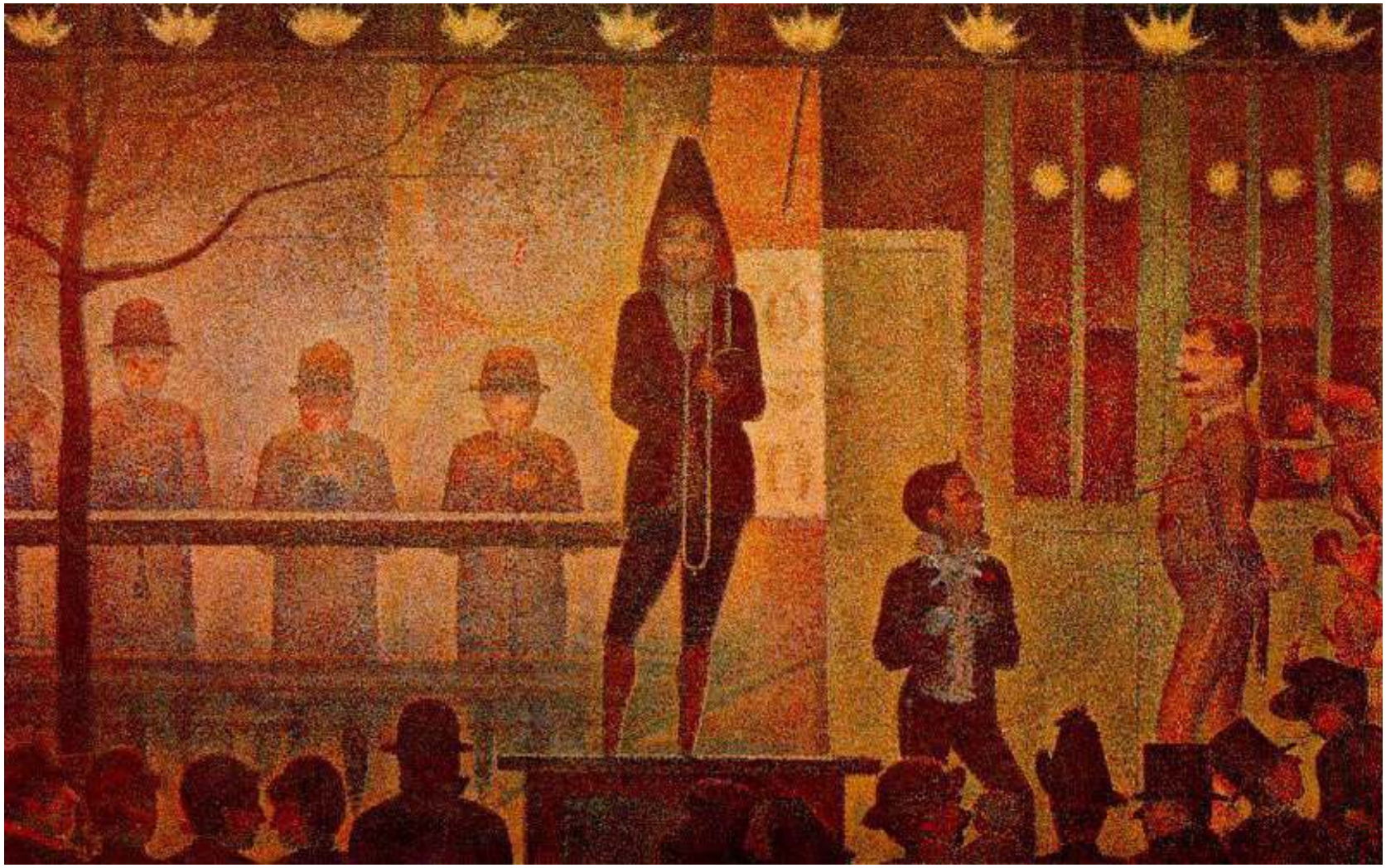
Chart Orientation and Proportions

If the data suggest a shape for the graphic, use that shape (portrait vs. landscape)

Otherwise, use landscape (rectangular) orientation about 50% wider than tall

Aim for the *Golden Rectangle*:

Ratio = 1.0 to 1.618...



The Parade, Seurat

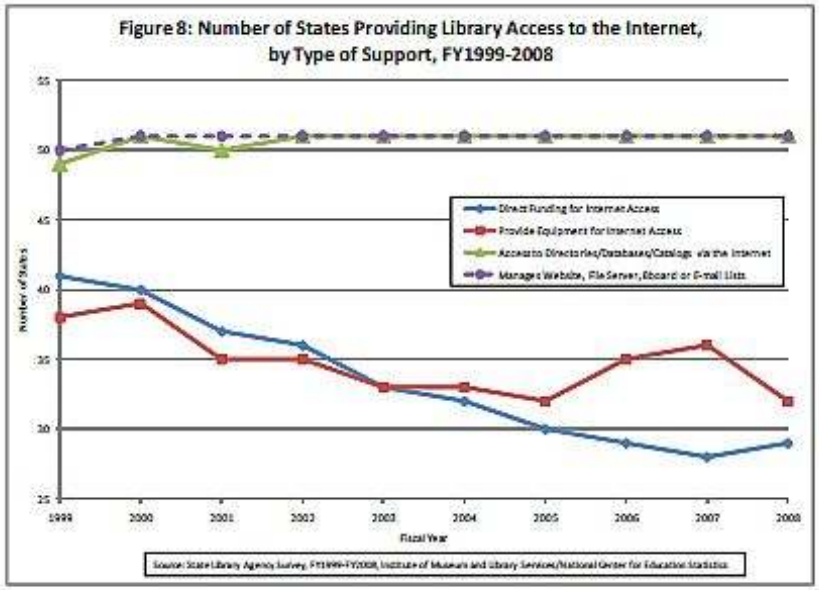
Legends (Keys)

Place legend at or near top of chart

Do not place legend in chart's data area

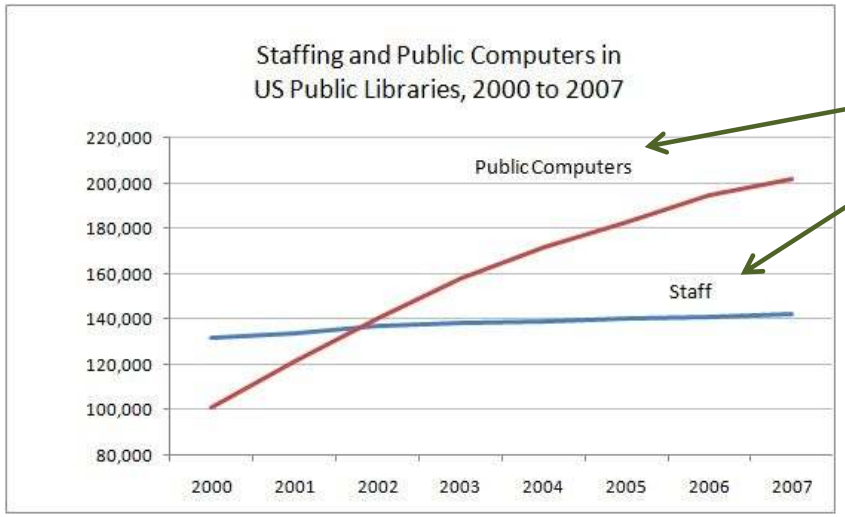
In line charts, annotate lines with data labels instead (no legend needed)

Order data labels to match the order of the data

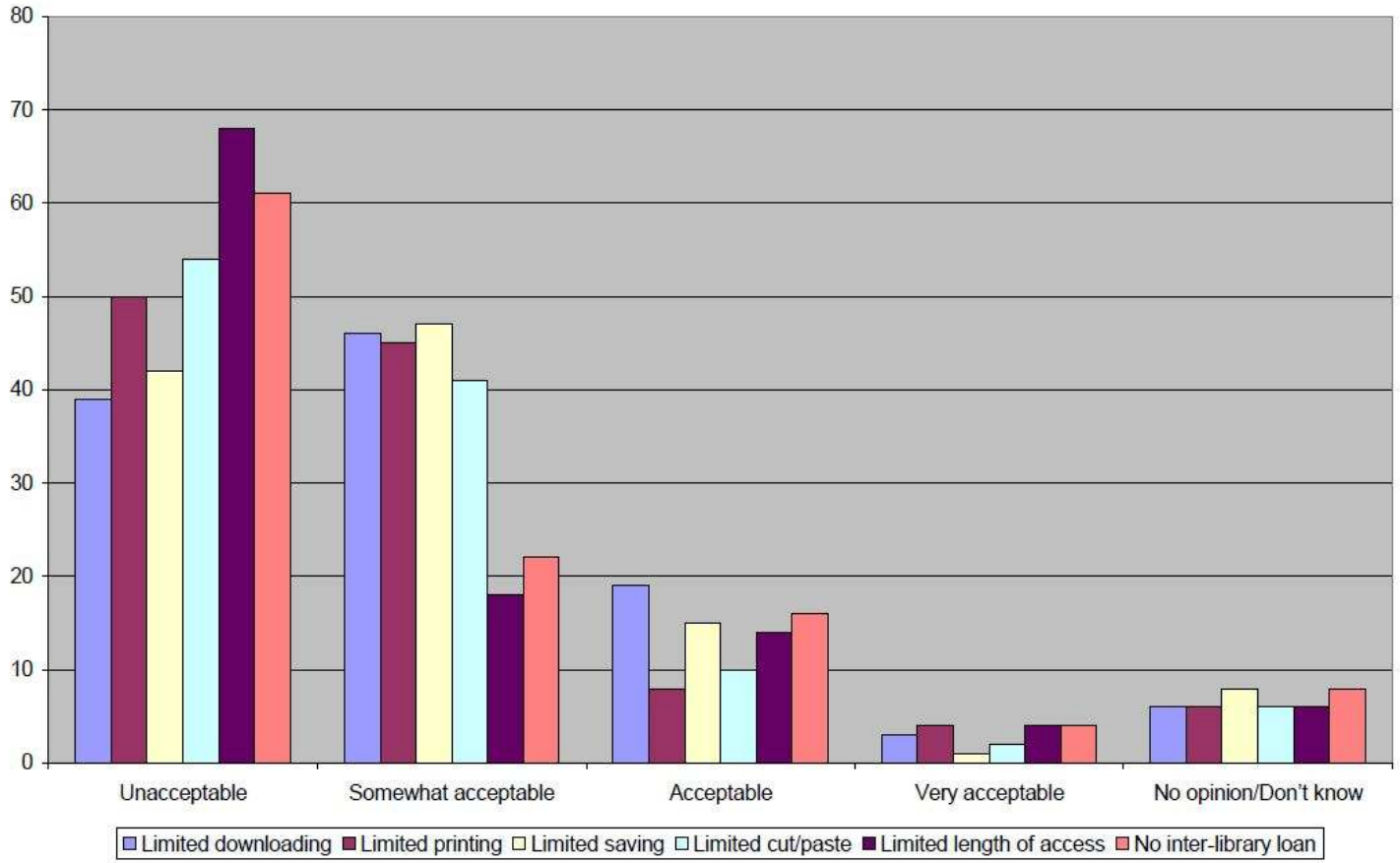


Source: Institute of Museum and Library Services, 2008.

Avoid placing a legend inside the data rectangle.



With *data labels* the reader does not have to translate legend colors, patterns, or labels.



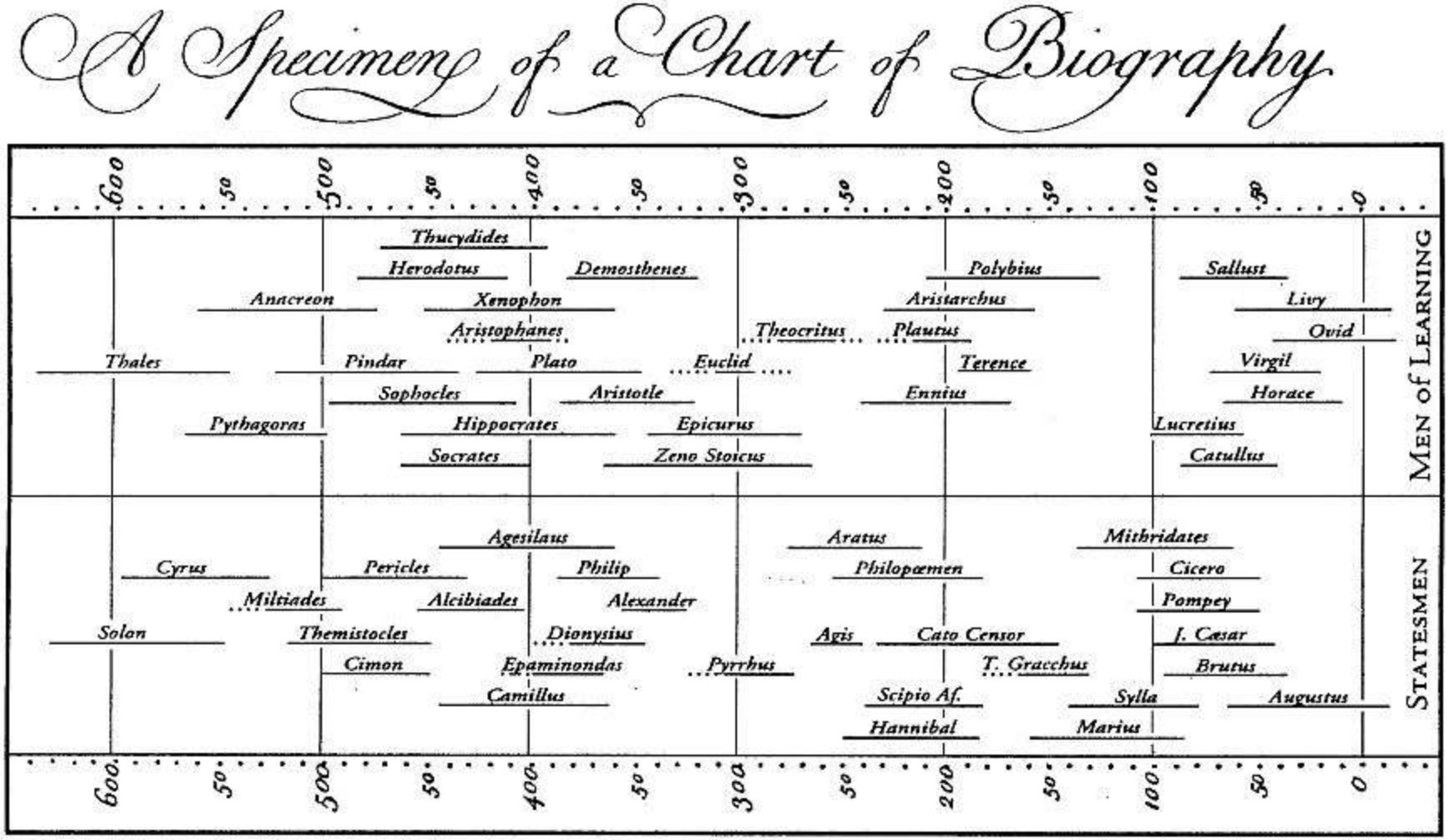
Source: Highwire Press, *2009 Librarian eBook Survey* , Stanford University, 2010.

This chart's legend matches the order of the data.

Line Charts

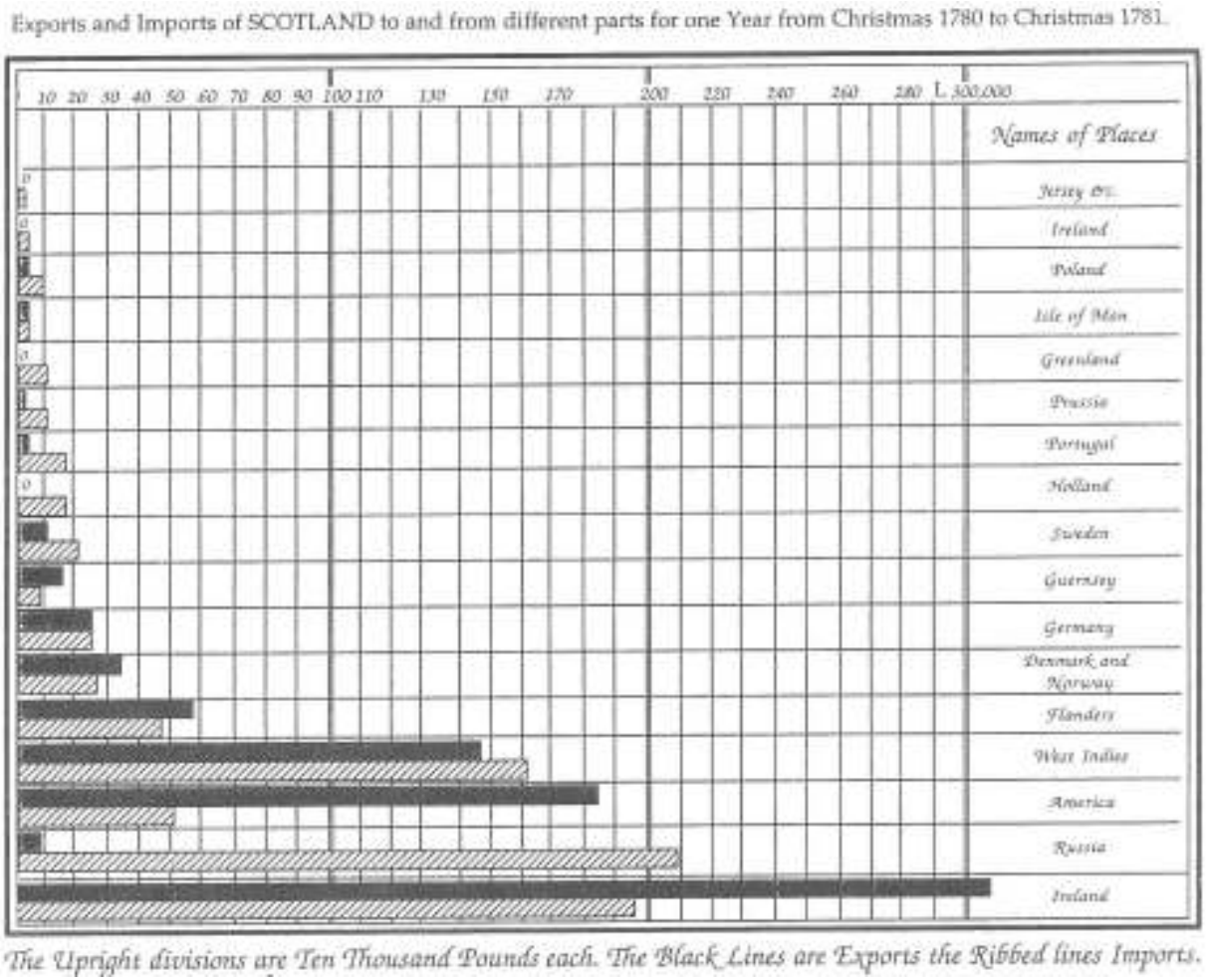
Joseph Priestley invented the line chart in 1765

Perhaps the first use of a line chart to portray quantitative data was the one by William Playfair in 1786



Source: Howard Wainer, 2005, Graphic Discovery

Joseph Priestley's line chart (1765) depicted biographical history.



Source: Howard Wainer, 2005.

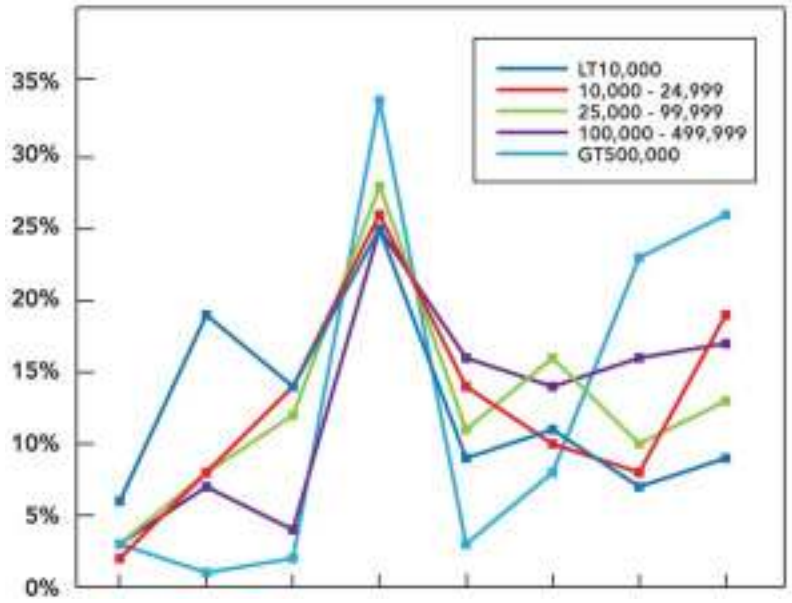
William Playfair's data line chart (1786)

Line Charts

Use four or fewer lines. For more, use separate panels

No spaghetti or spidery line charts!

Never use 3-D. Never.



Excerpted from a chart in *American Libraries*, Nov. 2009.

If the data are intertwined, consider producing separate, smaller charts for each measure.

Source: Stephen Few, 2009, *Now You See It.**

Attack of the 3D rainbow pasta!

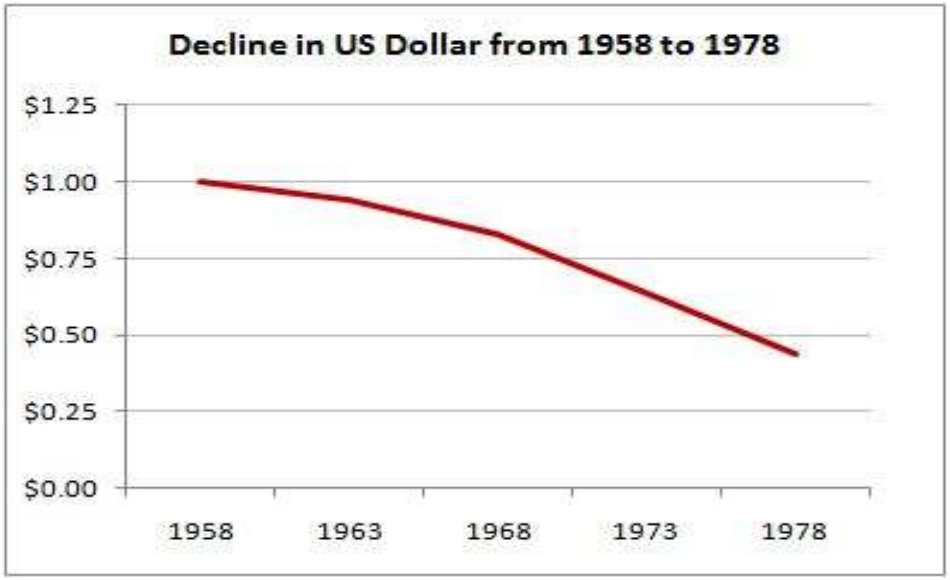
*Stephen Few presents this chart as a good *BAD* example.

Line Charts

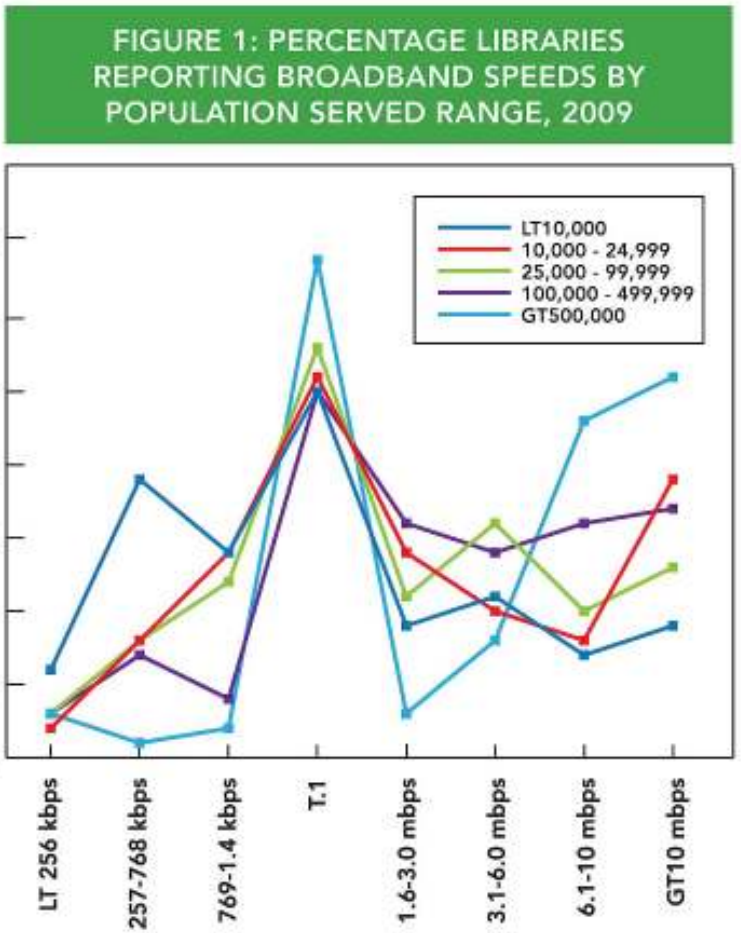
Lines should be thick enough to see without masking data peaks and valleys

Use line charts to plot time-series data

Do not use line charts for categorical data



Line charts are the preferred format for time-series data.



Source: *American Libraries*, Nov. 2009

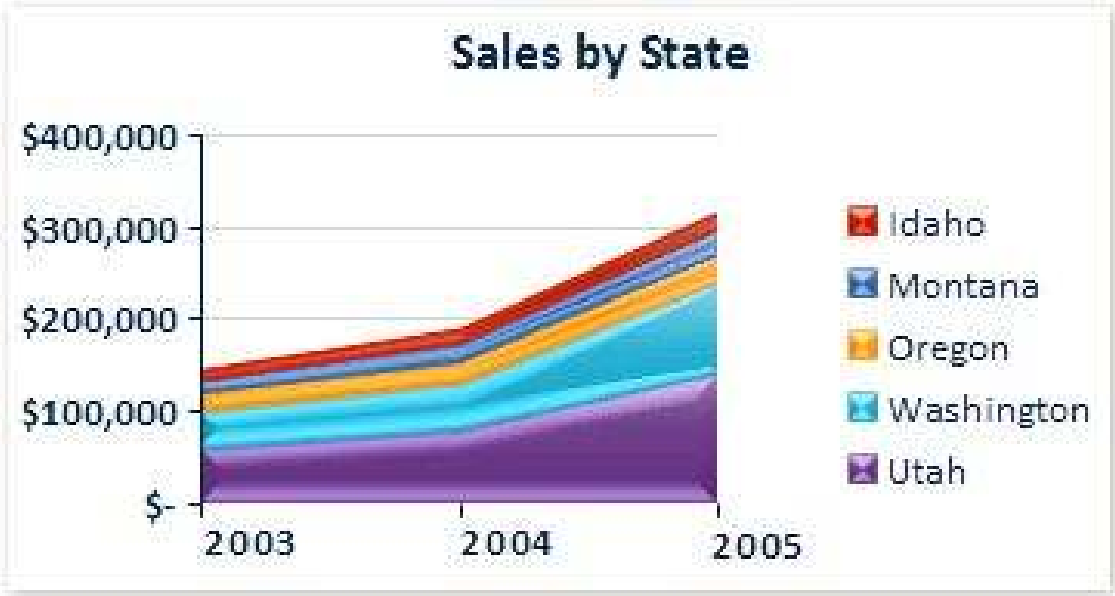
Avoid using line charts to depict categorical data

Line Charts

Never shade below a line unless chart has a zero baseline

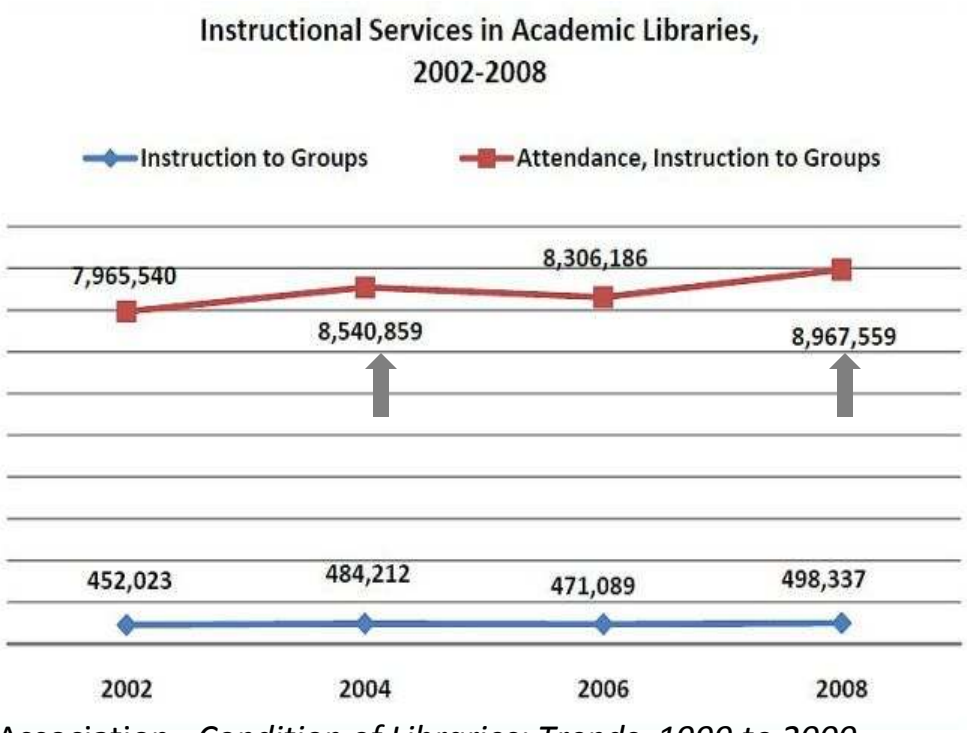
Place data values above lines whenever feasible

Do not vary placement of data values; either all above or all below the line



Source: www.microsoft.com, *Office 2007 Online Help.*

Oops. For line charts with shading below the line(s), the vertical scale should begin at zero, with zero clearly marked.



Gray arrows are my annotation

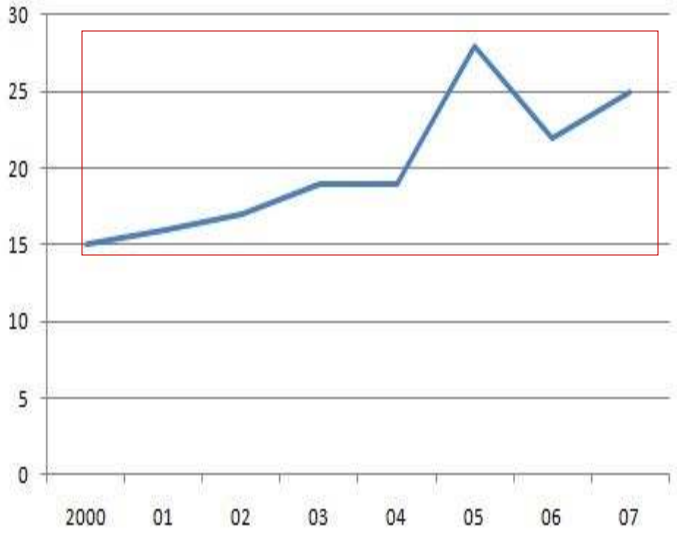
Source: American Library Association, *Condition of Libraries: Trends, 1999 to 2009*

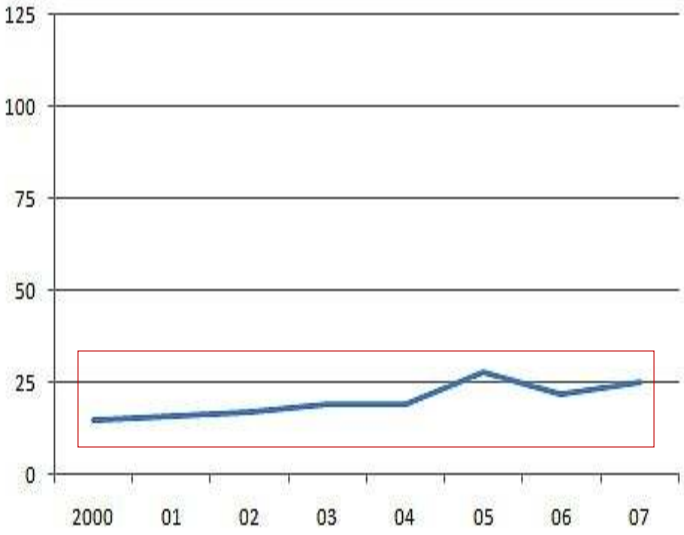
Positions of two of the numbers (arrows) suggest they are smaller than other numbers in the series.

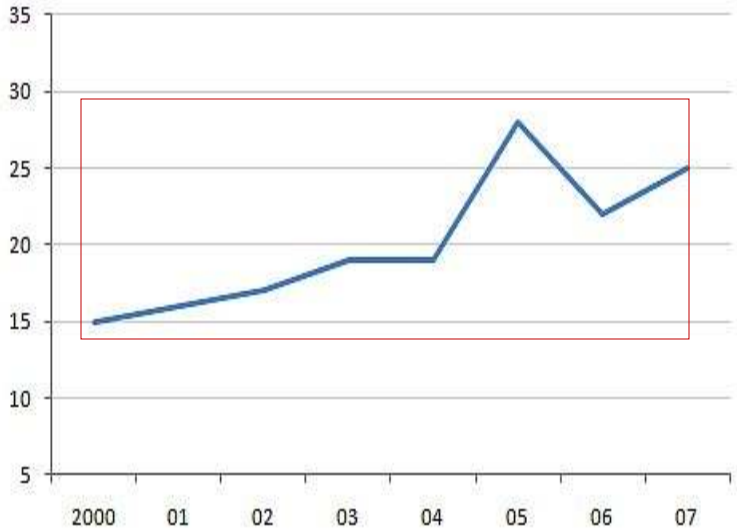
Scaling for Line Charts

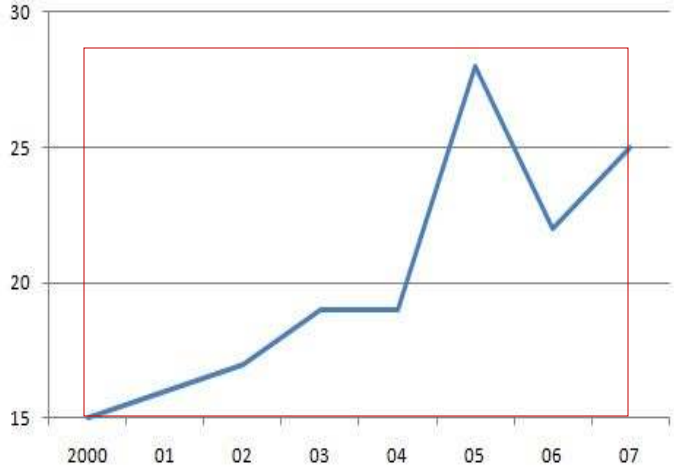
Dona Wong recommends choosing vertical axis scaling so that a rectangle surrounding the data line(s) covers about 2/3 of data area

William Cleveland recommends sizing scales so that data cover the majority of the data area (data rectangle’







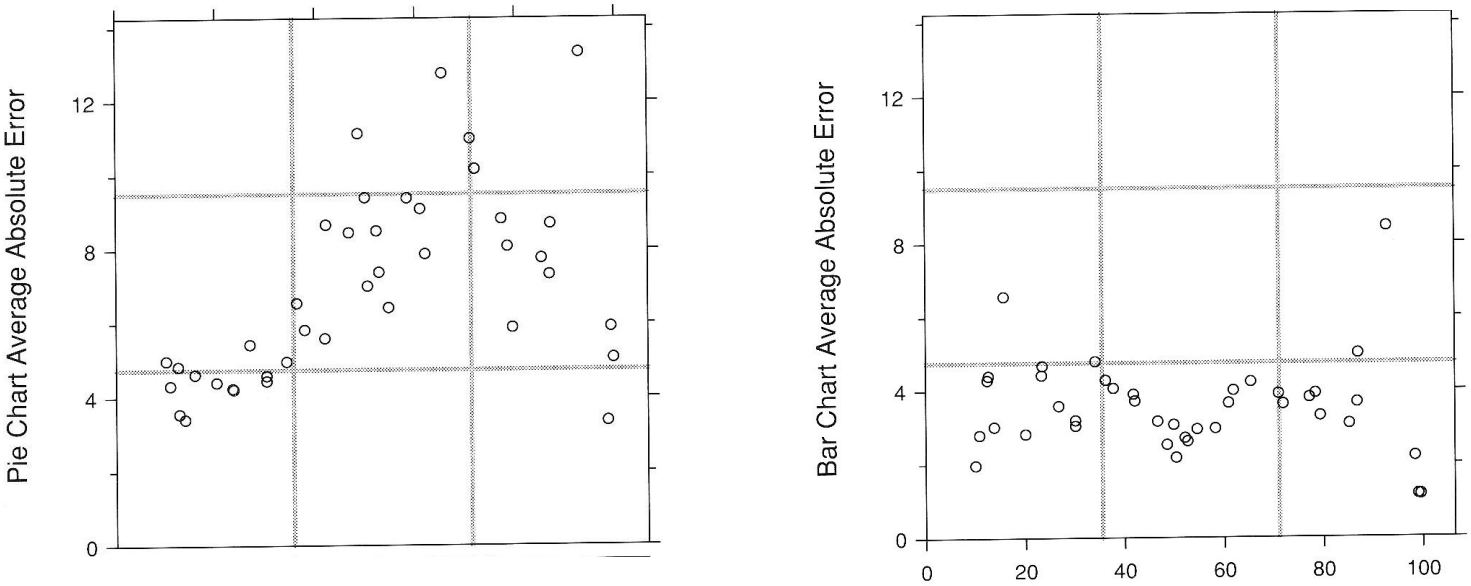


Bottom two charts utilize 2/3 or more of data area.

Scaling for Line Charts

Try to size the chart so that the data cover 2/3 or more of the data rectangle.

Don’t be dogmatic: A compelling reason—like the need for equivalent scales when comparing two sets of data—overrules this recommendation.

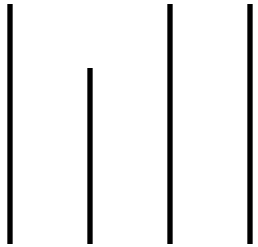


Source: William Cleveland, 1994.

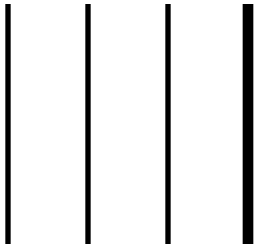
The importance of equivalent scaling for this comparison overrules the idea of filling 2/3 or more of the data area.

Pre-attentive Attributes

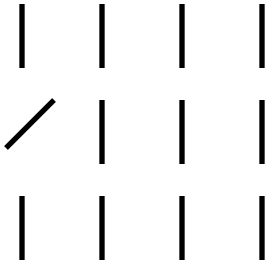
Length



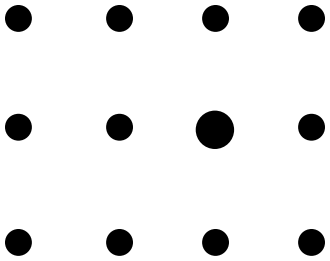
Thickness



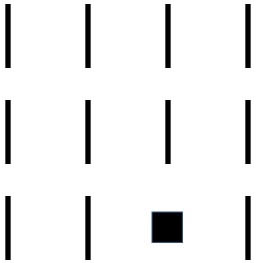
Orientation



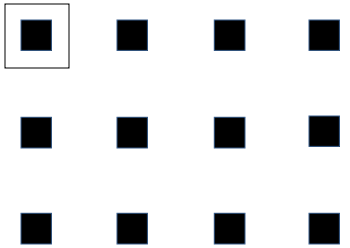
Size



Relative Shape



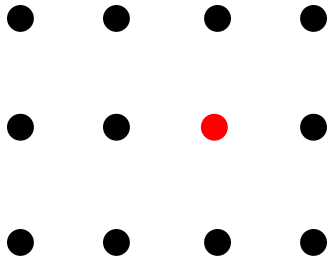
Enclosure



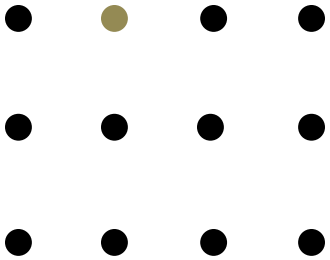
Adapted from Stephen Few, 2009.

Pre-attentive Attributes

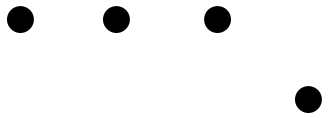
Hue



Intensity / Brightness



Spatial Position

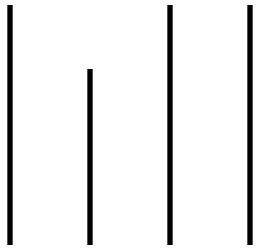


Adapted from Stephen Few, 2009.

Pre-attentive Attributes

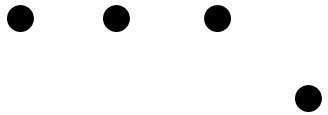
Quantitative perception is very precise with …

Length



Longer = greater quantity

Spatial Position



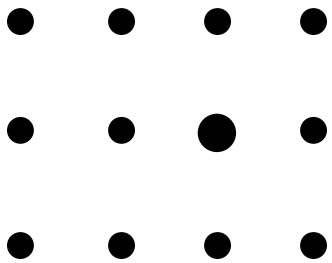
Higher or further right = greater quantity

Adapted from Stephen Few, 2009.

Pre-attentive Attributes

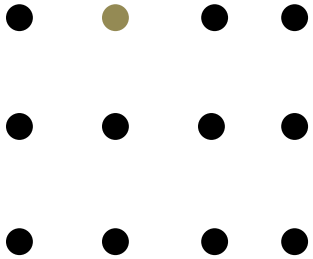
Quantitative perception is less precise with …

Size



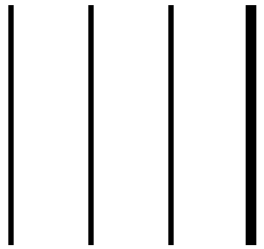
Bigger = greater

Intensity / Brightness



Darker = greater

Thickness



Thicker = greater

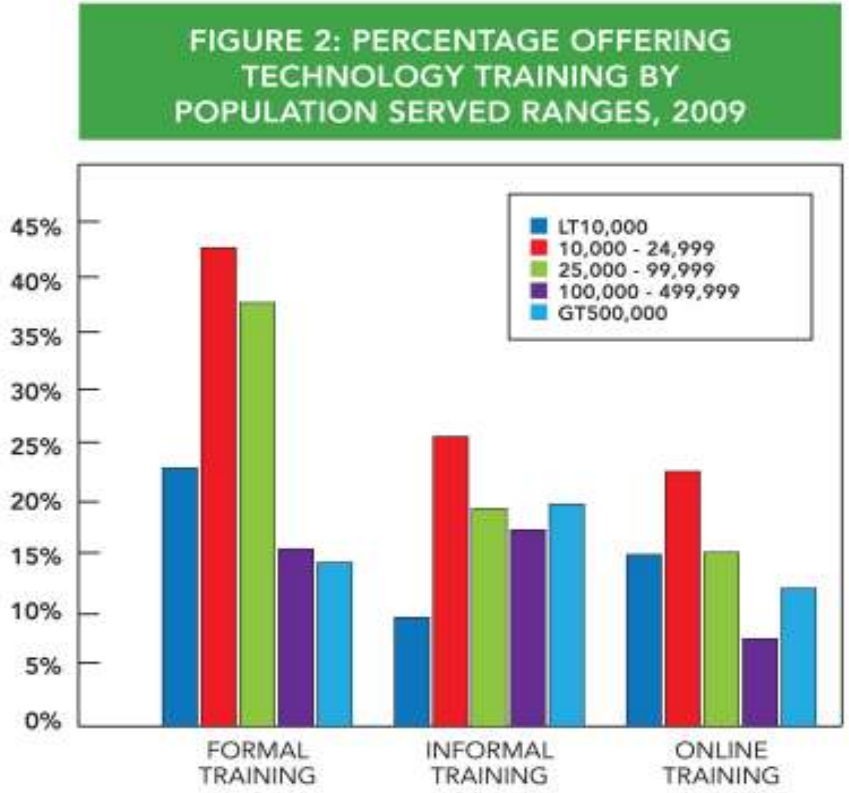
* Adapted from Stephen Few, 2009.

Using Color

"Admit colors into charts gracefully, as you would receive in-laws into your home." - Dona Wong, 2010.

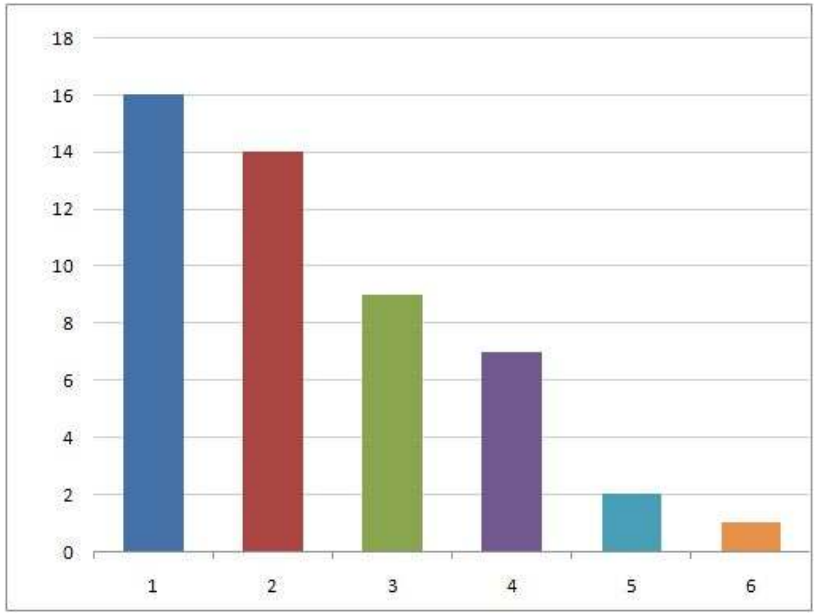
Color either to distinguish or emphasize data

No decorating!

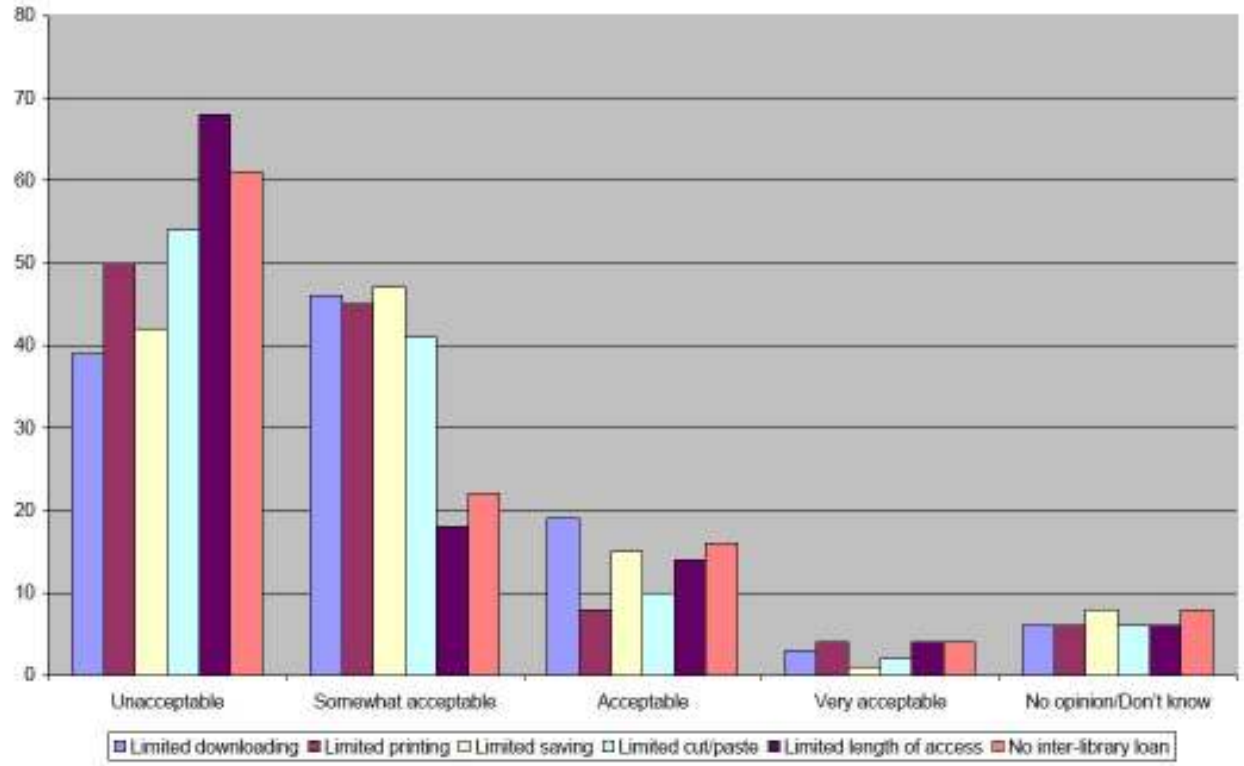


Source: *American Libraries,* Nov 2009

Multi-colored graphs make the eyes and brain labor to make sense of a myriad of cues.

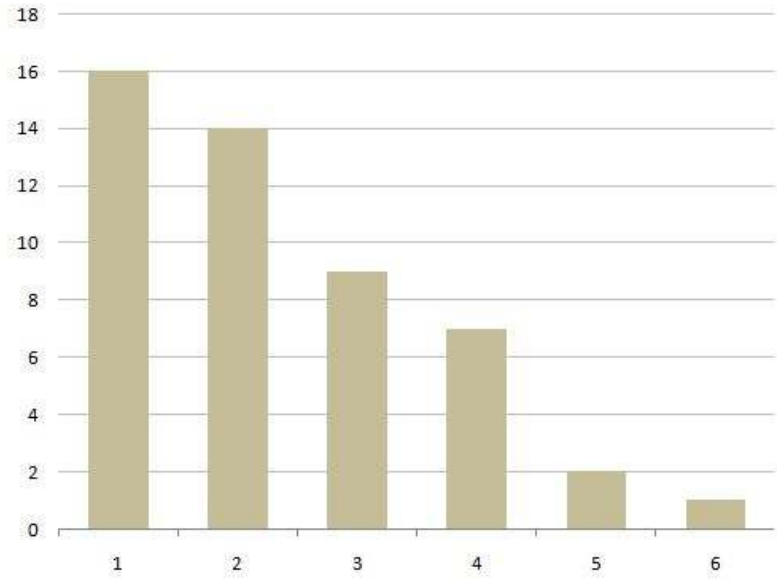


MS Excel 2007 default chart color scheme: circus stripes!



Source: Highwire Press, 2010

Stylish coloring does not make circus striping less distracting.



In well-designed charts, graphic elements stay out of the way, allowing the data to be easily perceived.

Warm and Cool Colors

Warm

Cool

Warm and Cool Colors

Warm colors: red, orange, yellow

Cool colors: blue, violet, neutral gray

Objects rendered in warm colors appear larger than cool colors

Do not use colors from opposite sides of the color wheel together

Warm and Cool Colors

Warm colors: red, orange, yellow

Cool colors: blue, violet, neutral gray

Objects rendered in warm colors appear larger than cool colors

Do not use colors from opposite sides of the color wheel together

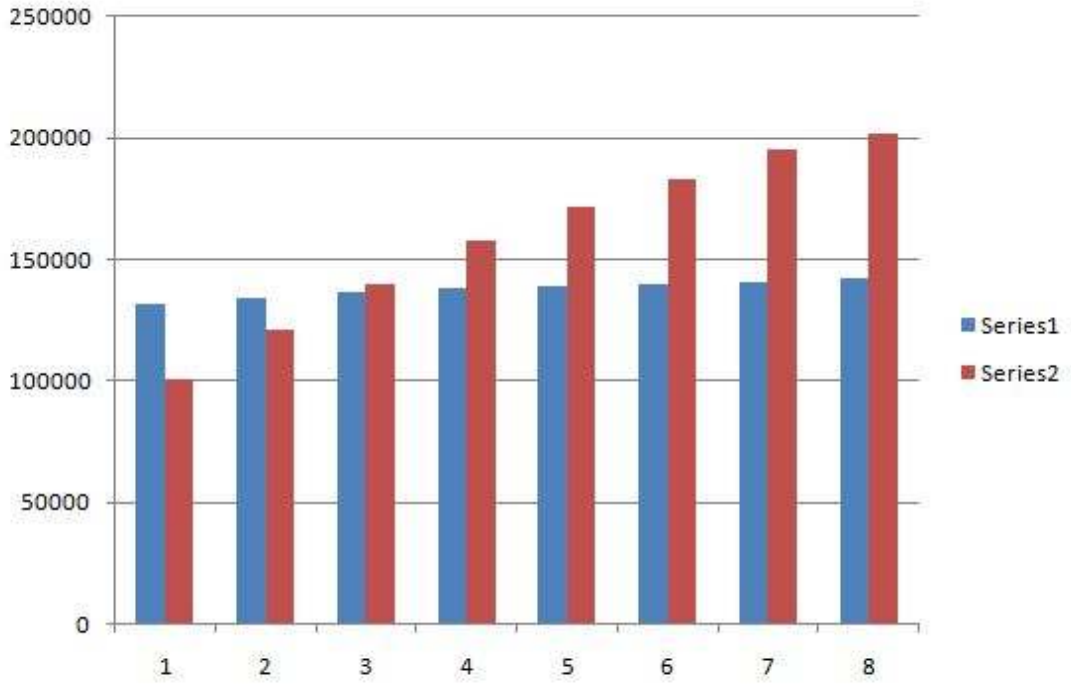
Warm and Cool Colors

Warm colors: red, orange, yellow

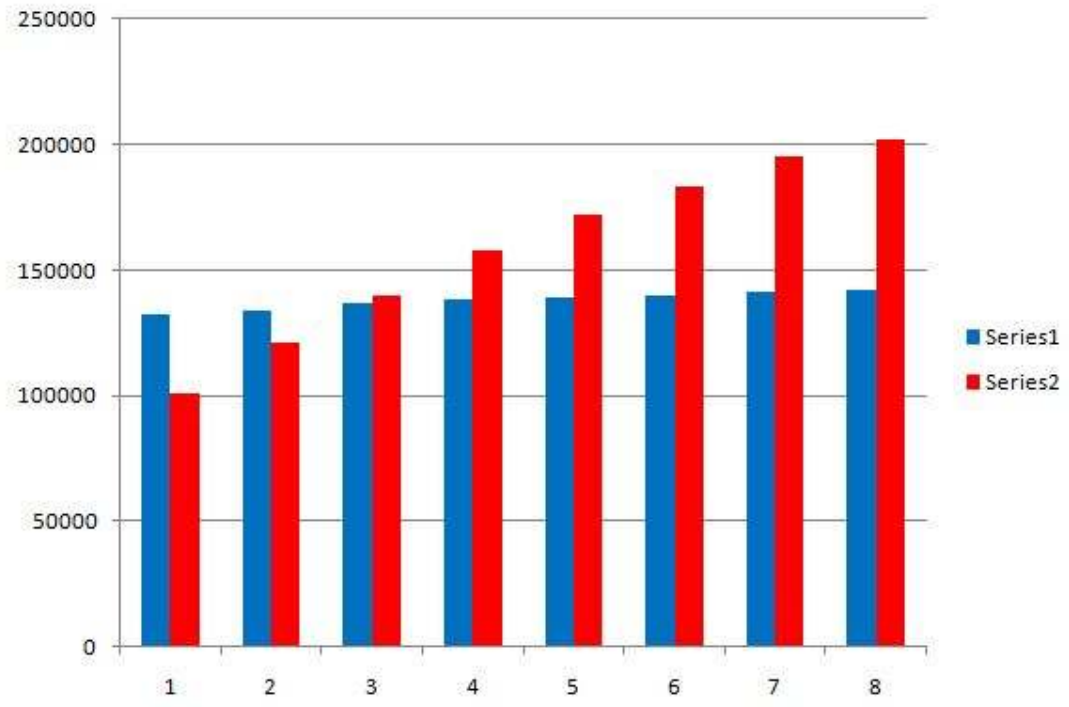
Cool colors: blue, violet, neutral gray

Objects rendered in warm colors appear larger than cool colors

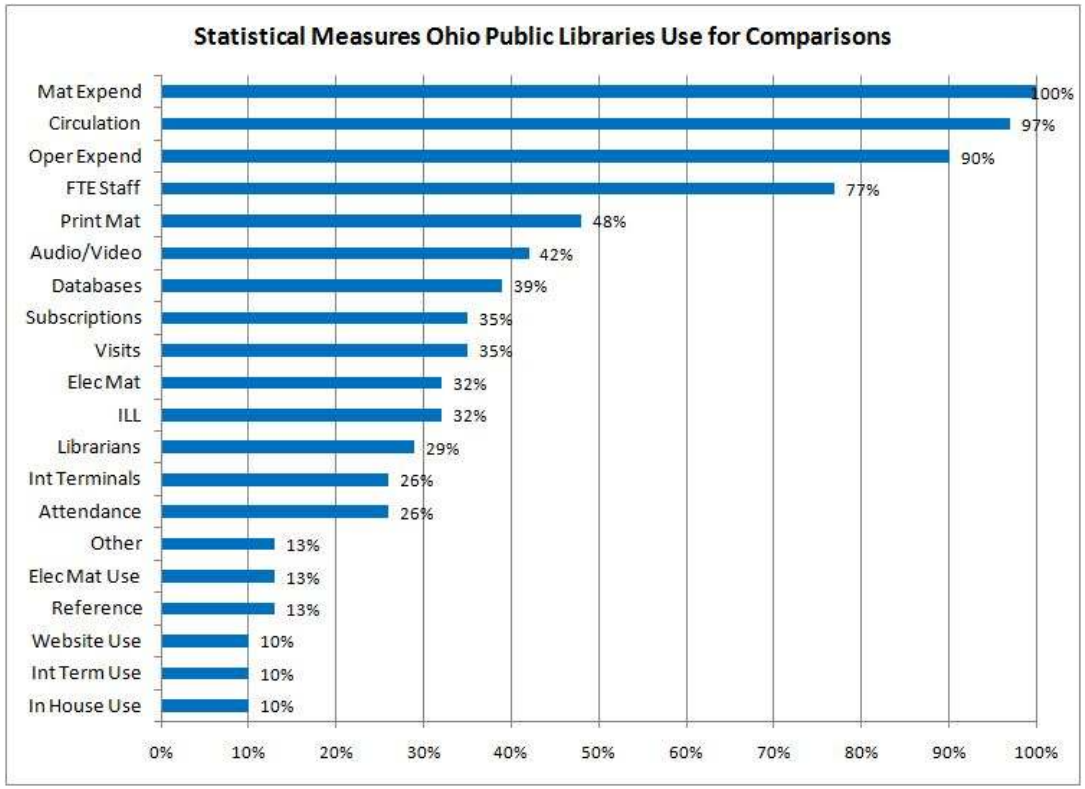
Do not use colors from opposite sides of the color wheel together



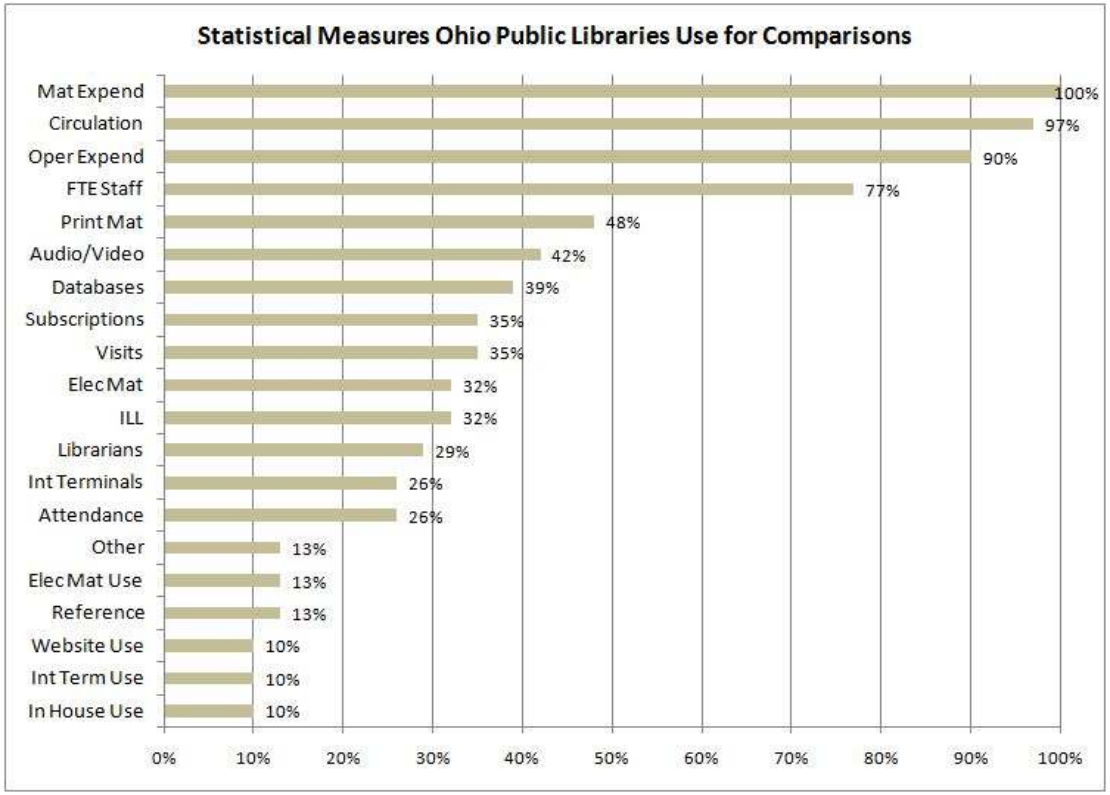
MS Excel 2007 default coloring for 2-bar chart



MS Excel coloring for 2-bar chart enhanced!



High contrast (color or black & white) produces moire.



Lower contrast, dimmer hues avoid moire.

Using Color

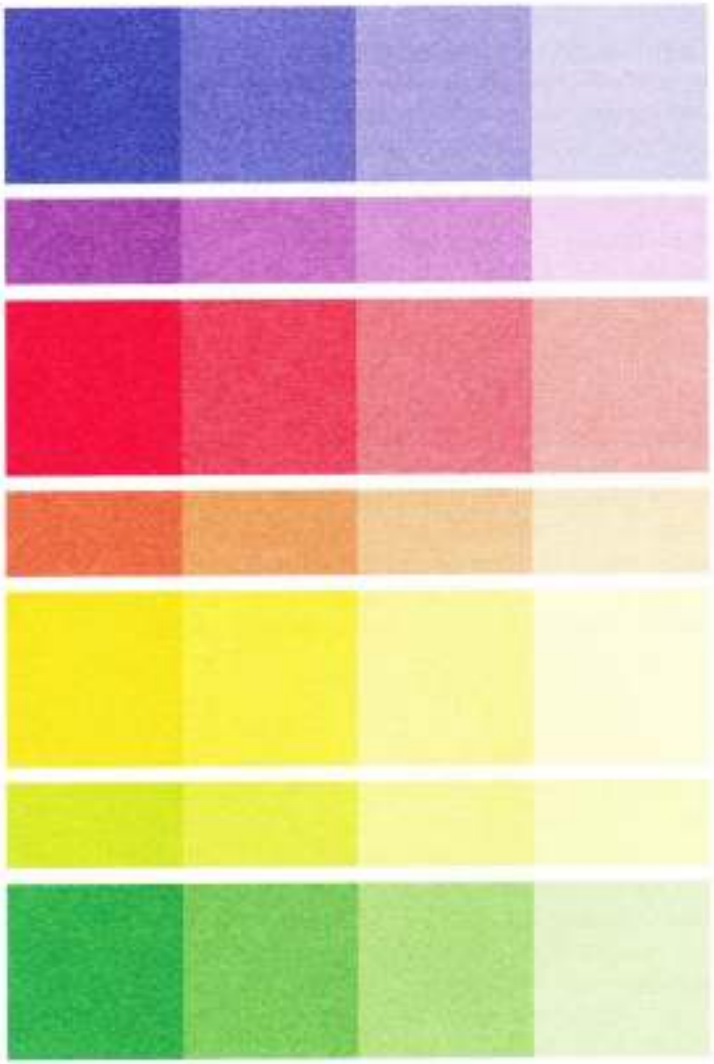
Choose a single color palette (hues and shades hues) for your entire set of graphs

Do not vary these for decorative purposes

Use shades of each hue*

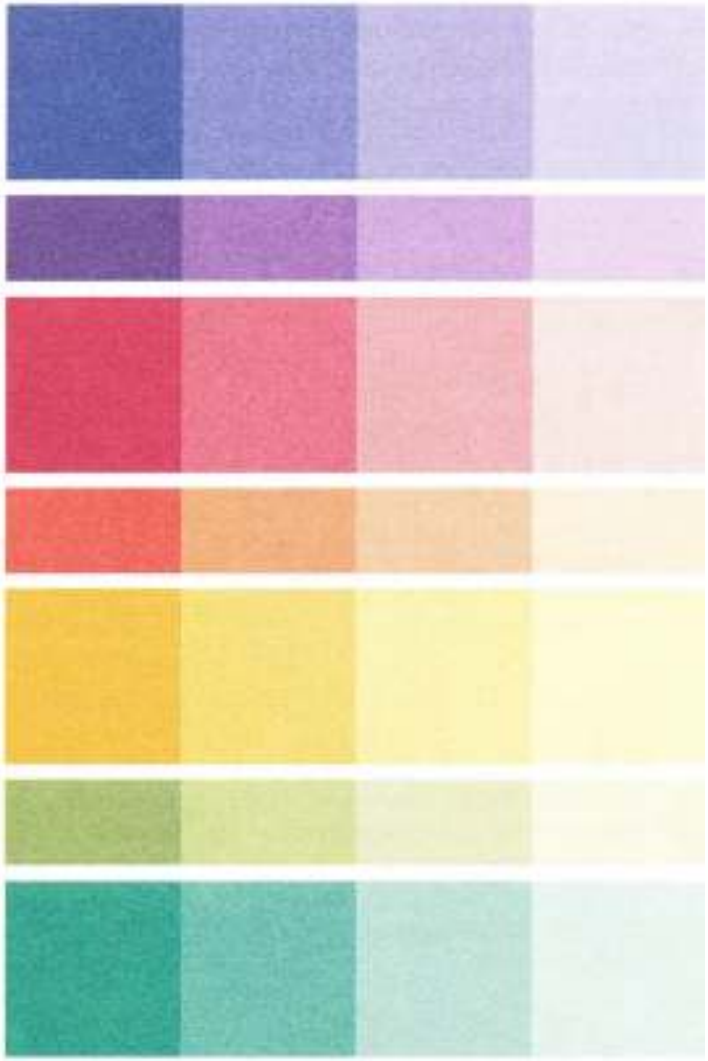
* Depends on capability of graphing software used.

Color palette - bright

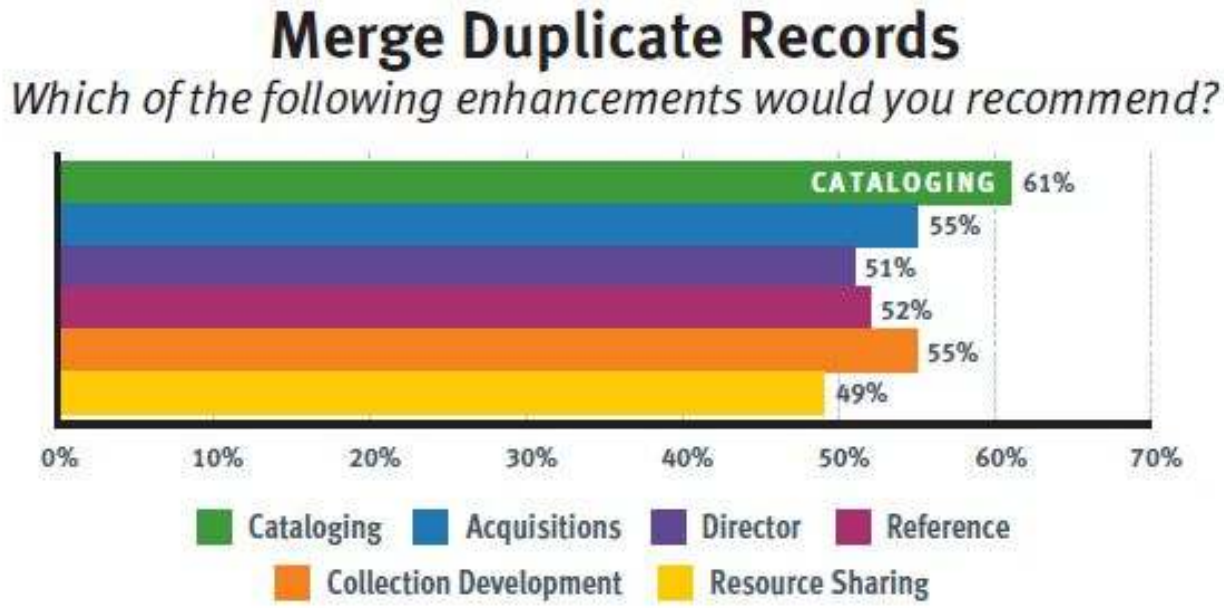


Source: Wong, Dona, 2010, *The Wall Street Journal Guide to Information Graphics.*

Color palette - muted



Source: Wong, Dona, 2010.

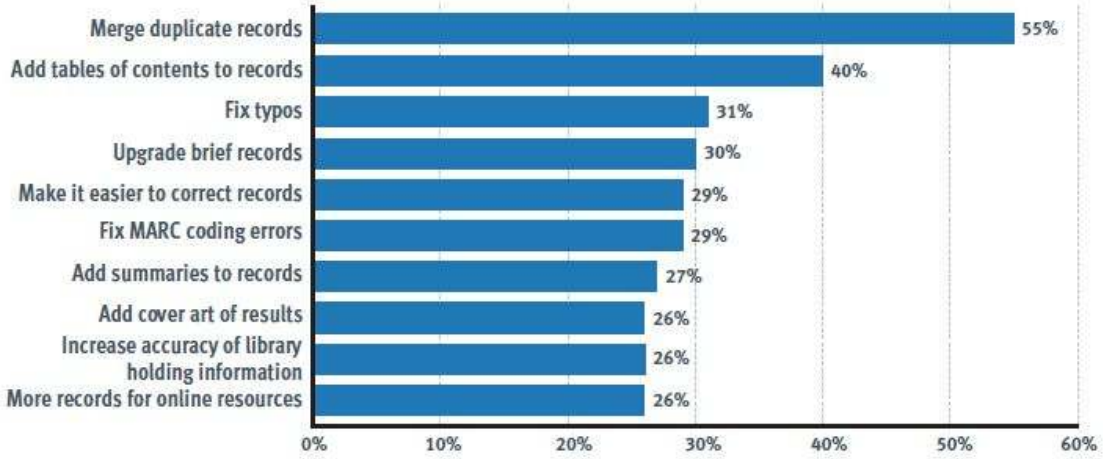


Source: *Online Catalogs: What Users and Librarians Want*, OCLC, 2009 (Library survey)

Color schemes are best when they have a rhyme and reason (as we will see this one does).

Acquisitions—Top Ten Enhancements

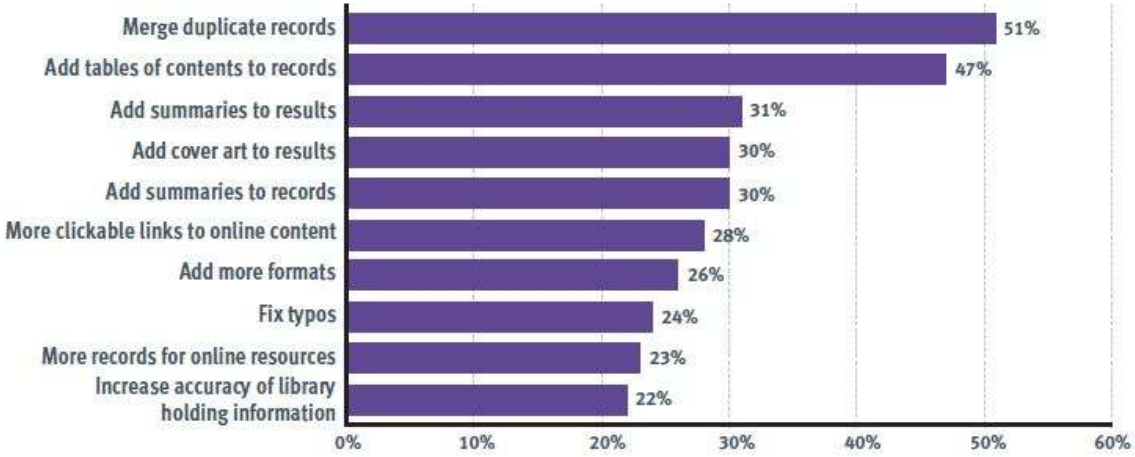
Which of the following enhancements would you recommend?



Source: *Online Catalogs: What Users and Librarians Want*, OCLC, 2009 (Library survey)

Library Directors—Top Ten Enhancements

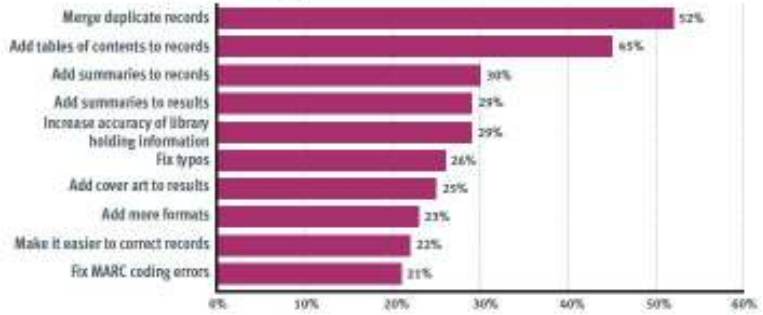
Which of the following enhancements would you recommend?



Source: *Online Catalogs: What Users and Librarians Want*, OCLC, 2009 (Library survey)

Reference—Top Ten Enhancements

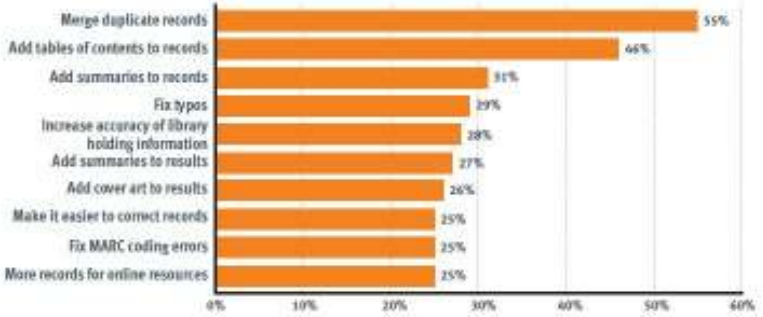
Which of the following enhancements would you recommend?



Source: *Online Catalogs: What Users and Librarians Want*, OCLC, 2009 (Library survey)

Collection Development—Top Ten Enhancements

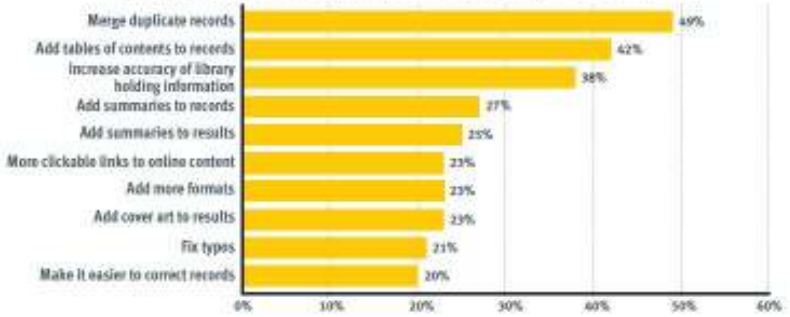
Which of the following enhancements would you recommend?



Source: *Online Catalogs: What Users and Librarians Want*, OCLC, 2009 (Library survey)

Resource Sharing—Top Ten Enhancements

Which of the following enhancements would you recommend?



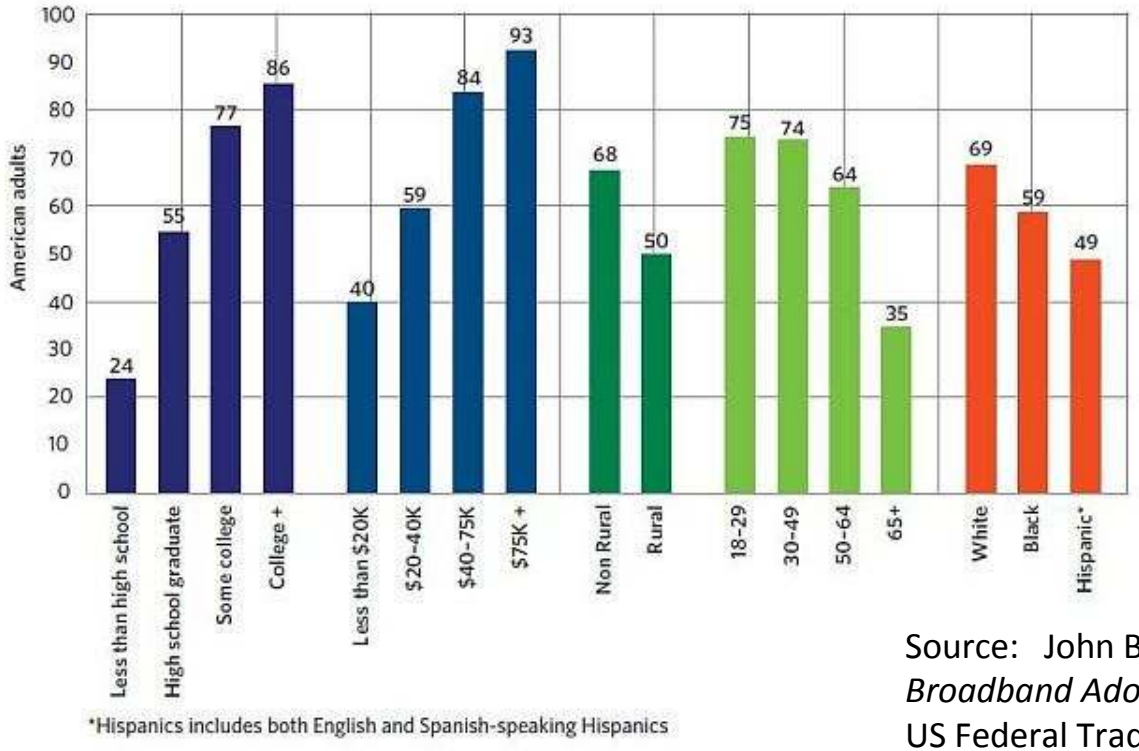
Source: *Online Catalogs: What Users and Librarians Want*, OCLC, 2009 (Library survey)

Using Color

Use a single color for each data type, including pie charts

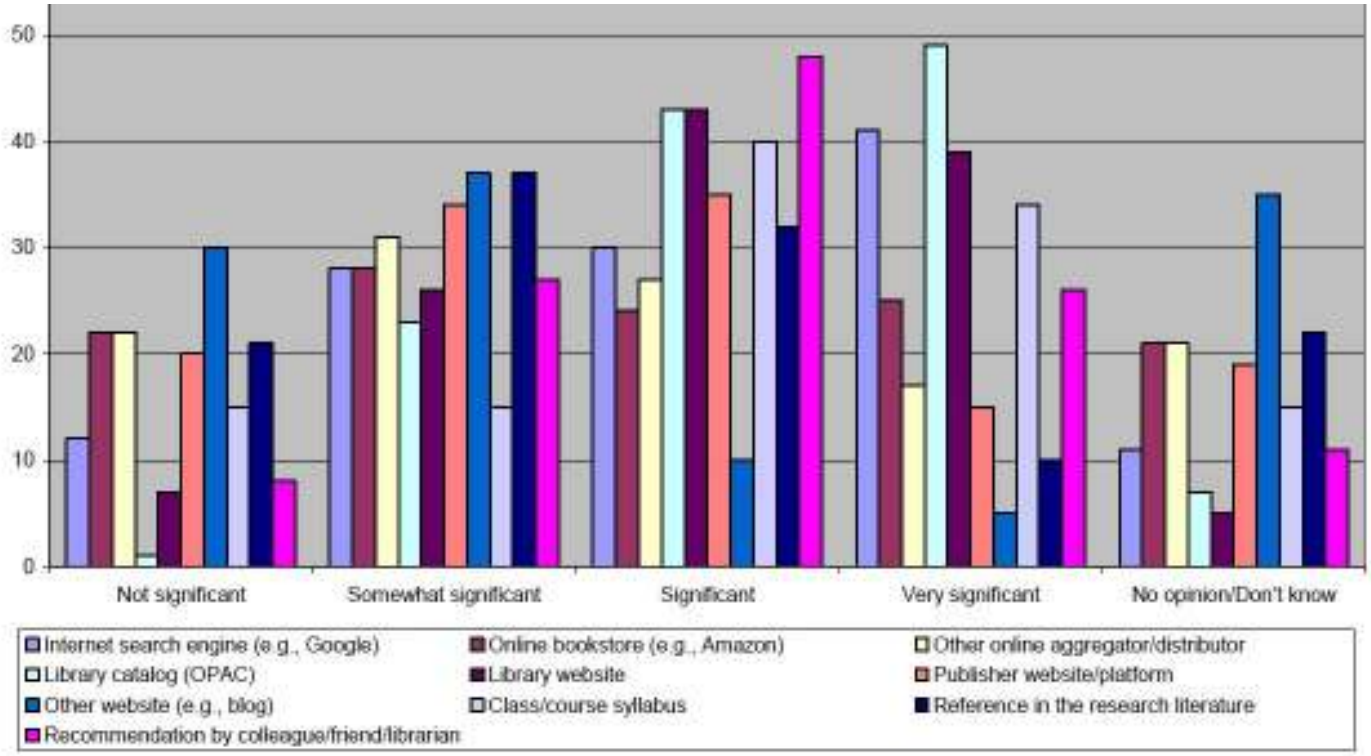
To repeat: Do not use multiple colors to represent the same type of data (as MS Excel's default setting for color bar charts does!)

Use a different shade, or occasionally a different color to highlight data



Source: John B. Horrigan, 2010, *Broadband Adoption and Use in America,* US Federal Trade Commission

This chart uses one color for each variable. Hues are not bright or inharmonious.

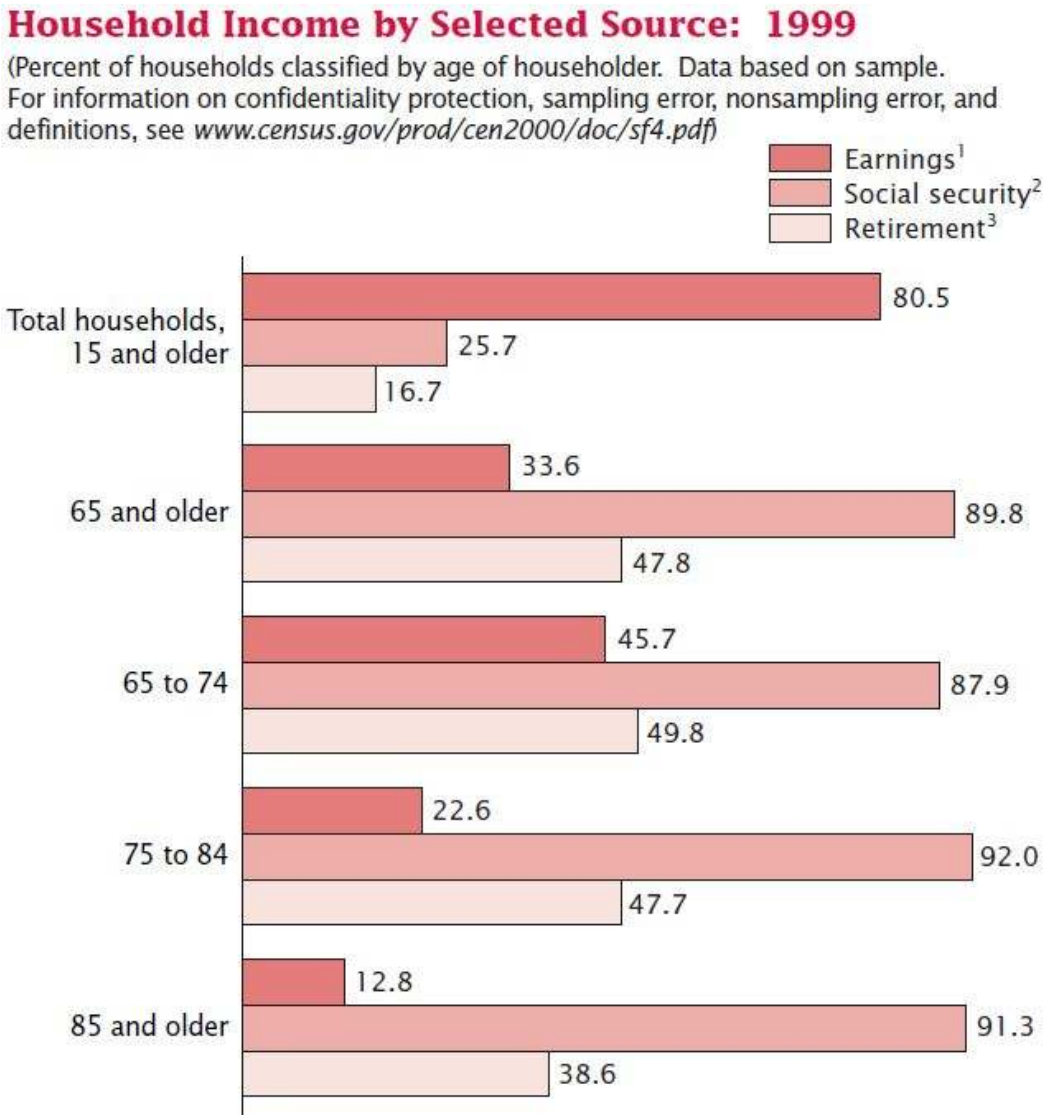


Source: Highwire Press, *2009 Librarian eBook Survey* , Stanford University, 2010.

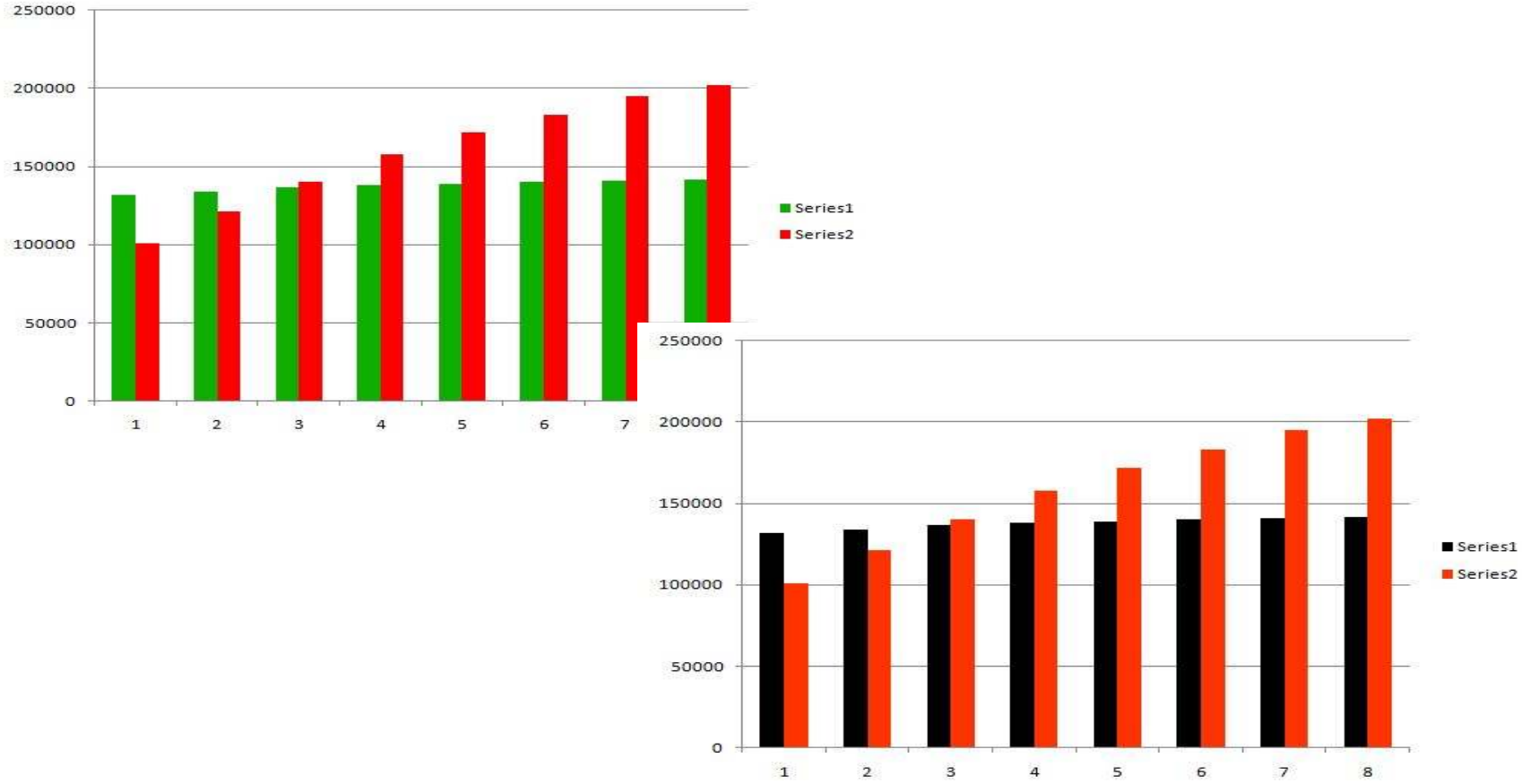
Follows the rule, use a single color for each data type. But the result is disturbing to the eye.

Using Color

In a single, related set of measures, use graduating shades of one color or colors on the same side of the color wheel in a multiple-bar chart



Source: U.S. Bureau of the Census 2004, *We the People: Aging in the U.S.*



Avoid thematic representation of colors, such as holiday colors

Using Color

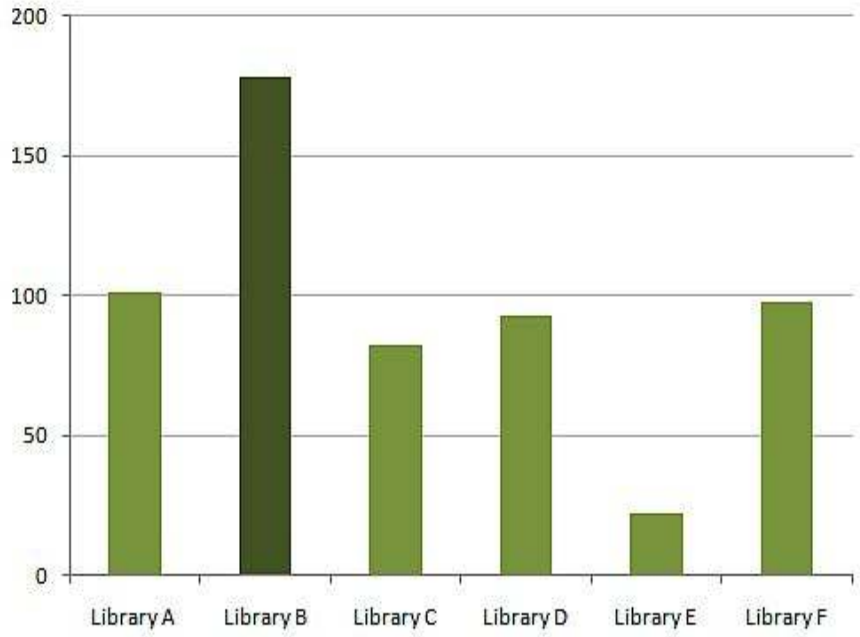
Highlight most important data with bright color (e.g. red); render others in a single, less prominent color

With financial data, avoid use of red to indicate positive values

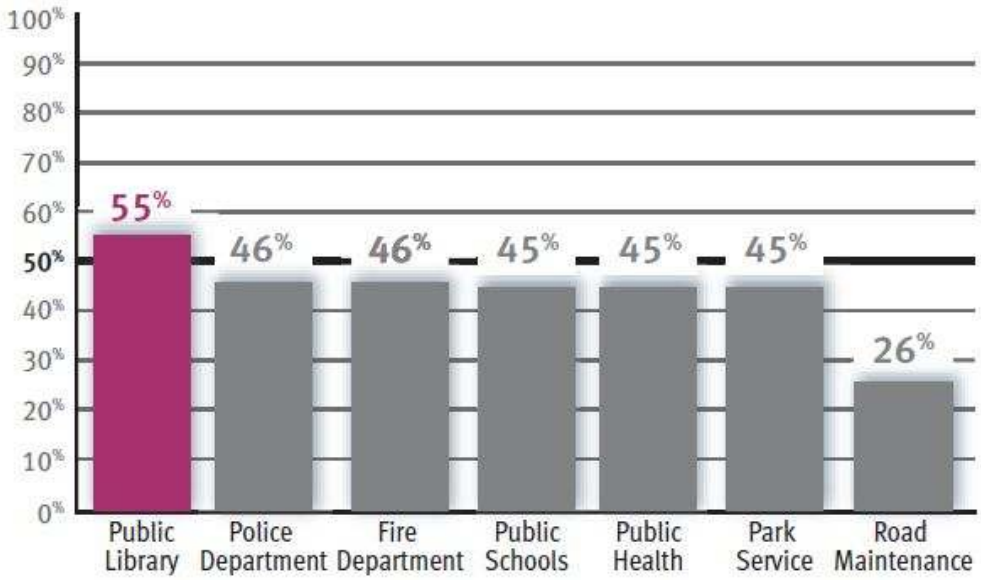
Using Color

When using a color for emphasis, use a distinctly lighter or darker shade than the color used for the other data

Highlight most important data with bright primary color; render others in a single, less prominent color



Data can be emphasized by darker or brighter shades of a single hue.



Percentage of voting respondents with an agreement rating of 8, 9 or 10
Source: *From Awareness to Funding*, OCLC, 2008

When portraying data about a single variable* use alternate color for emphasis only

* Responses to one questionnaire item, in this example.

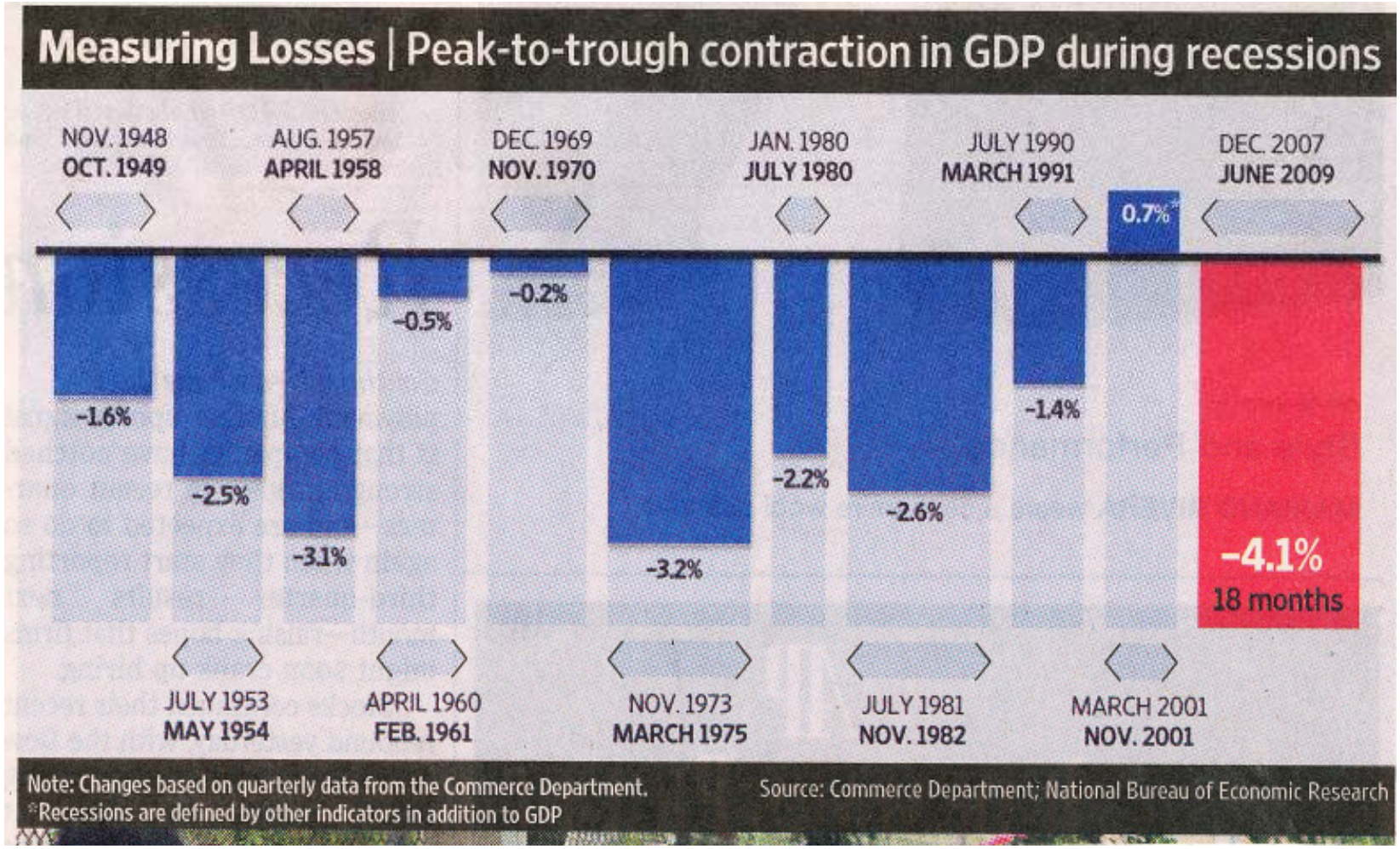
Using Color

In black and white charts, emphasize important data series with **dark black,** and the others in grayer shades

As a test, convert color charts to gray scale to evaluate shading

If using a dark background for design reasons, render text in white only (not yellow, beige, etc.)

Exercise 2



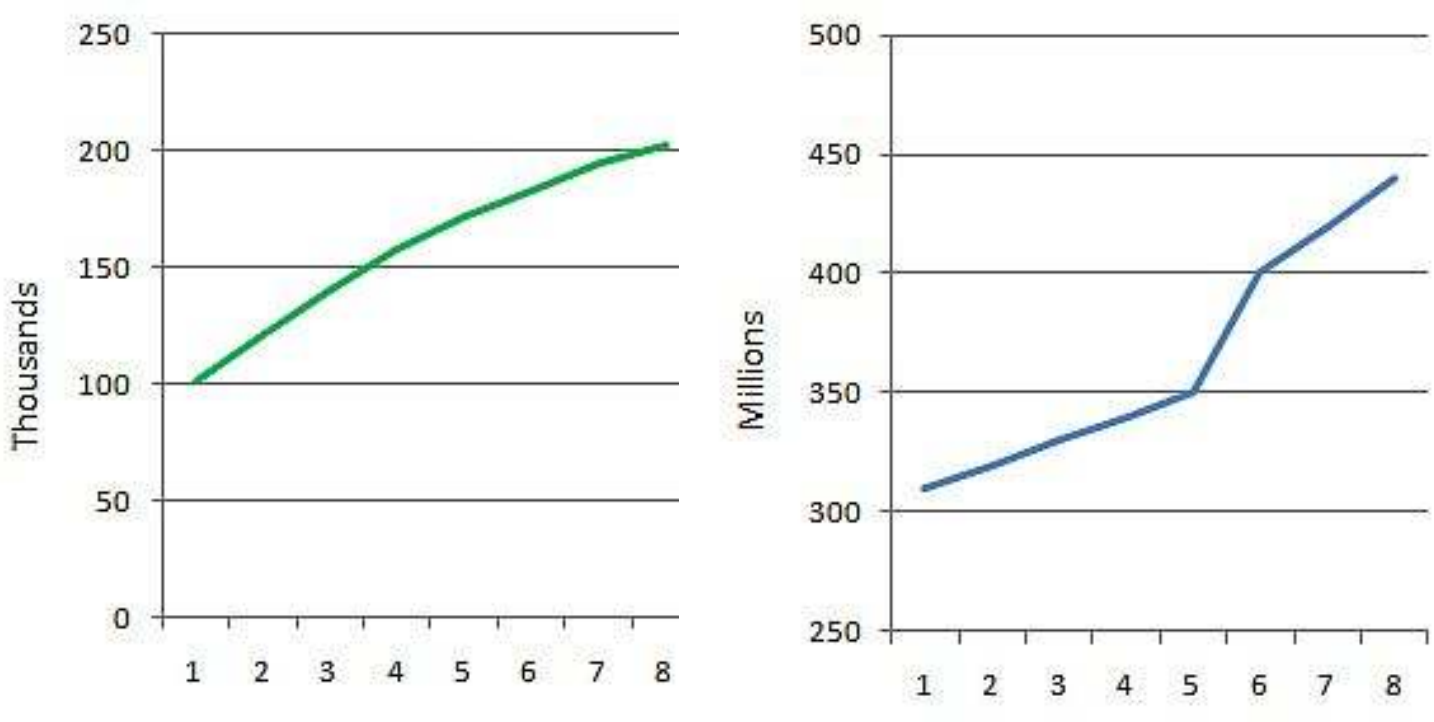
Source: *Wall Street Journal,* Oct. 2010

Comparisons on Two Different Vertical Scales

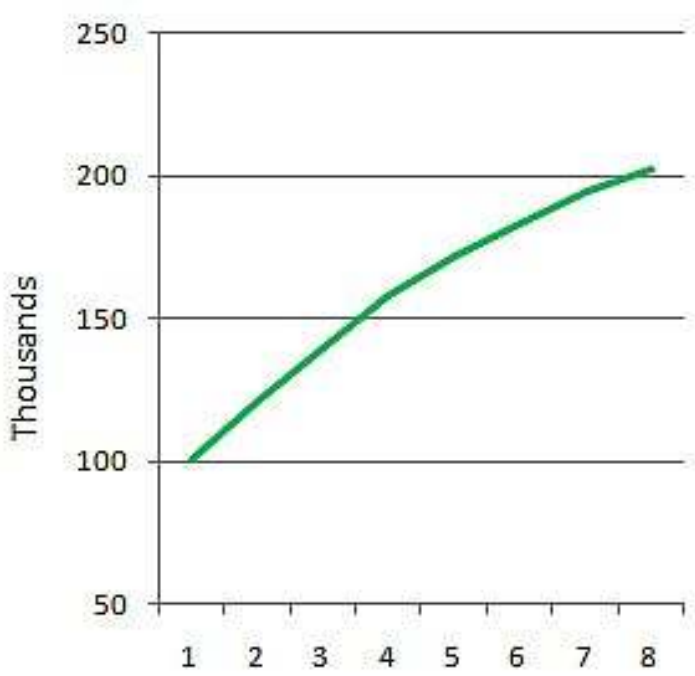
Use two scales to demonstrate how two related variables trend

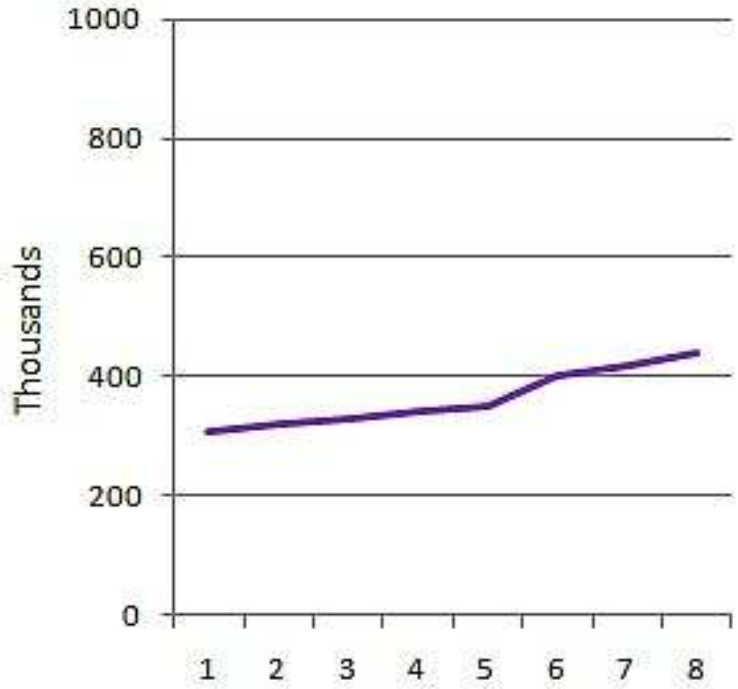
Don’t plot unrelated data

Choice of scale/scaling changes comparison

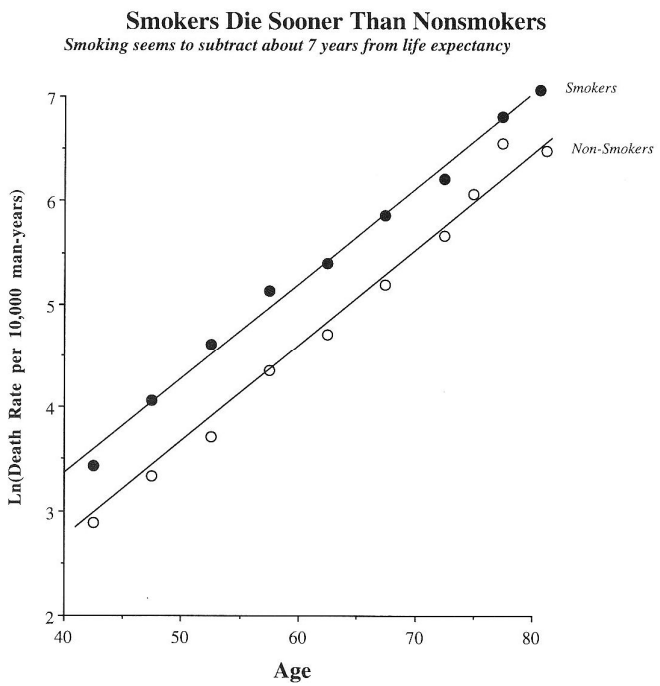


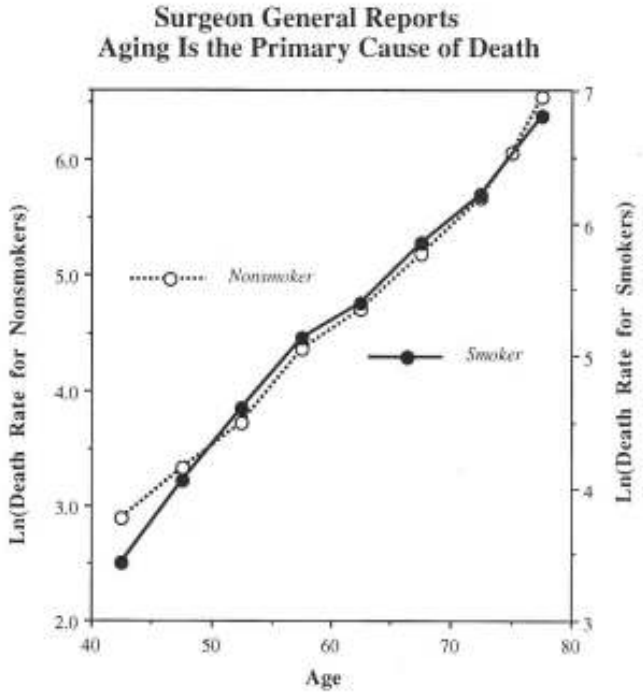
Scale sizes (spans) on vertical axes suggest that gain of 100,000 units is equivalent to a gain of 150 million.





Though the scale units are similar, the scales chosen exaggerate the left trend.



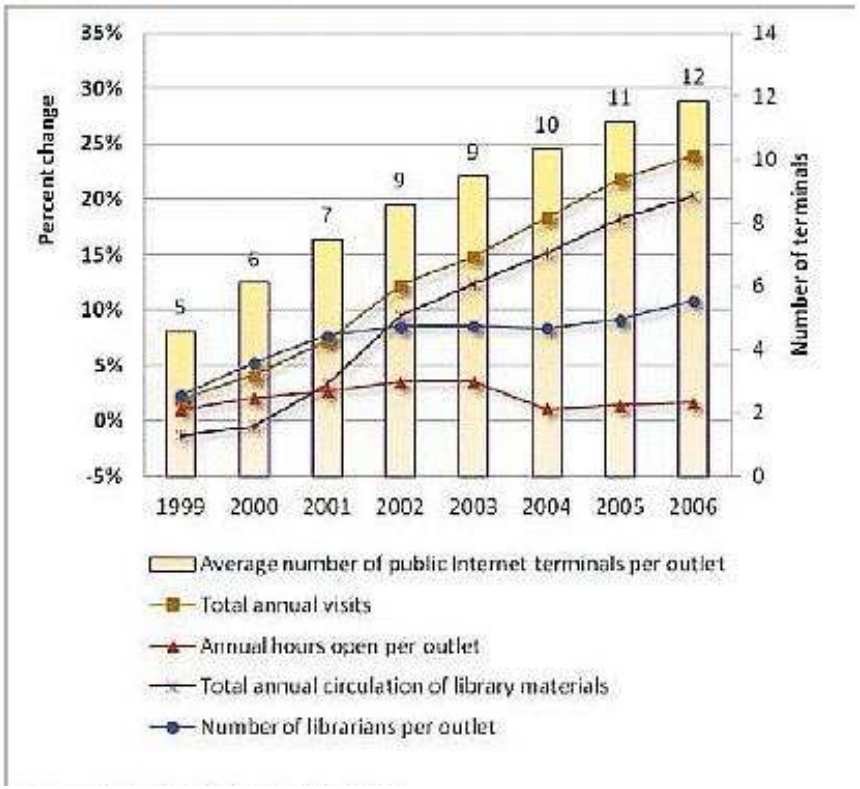


Source: Howard Wainer, 1997.

The chart at the use a 'double Y axis' (two vertical axes*) to put a definite *slant* on the data.

*In this case the axes are calibrated unequally to produce the effect the tobacco industry wanted.

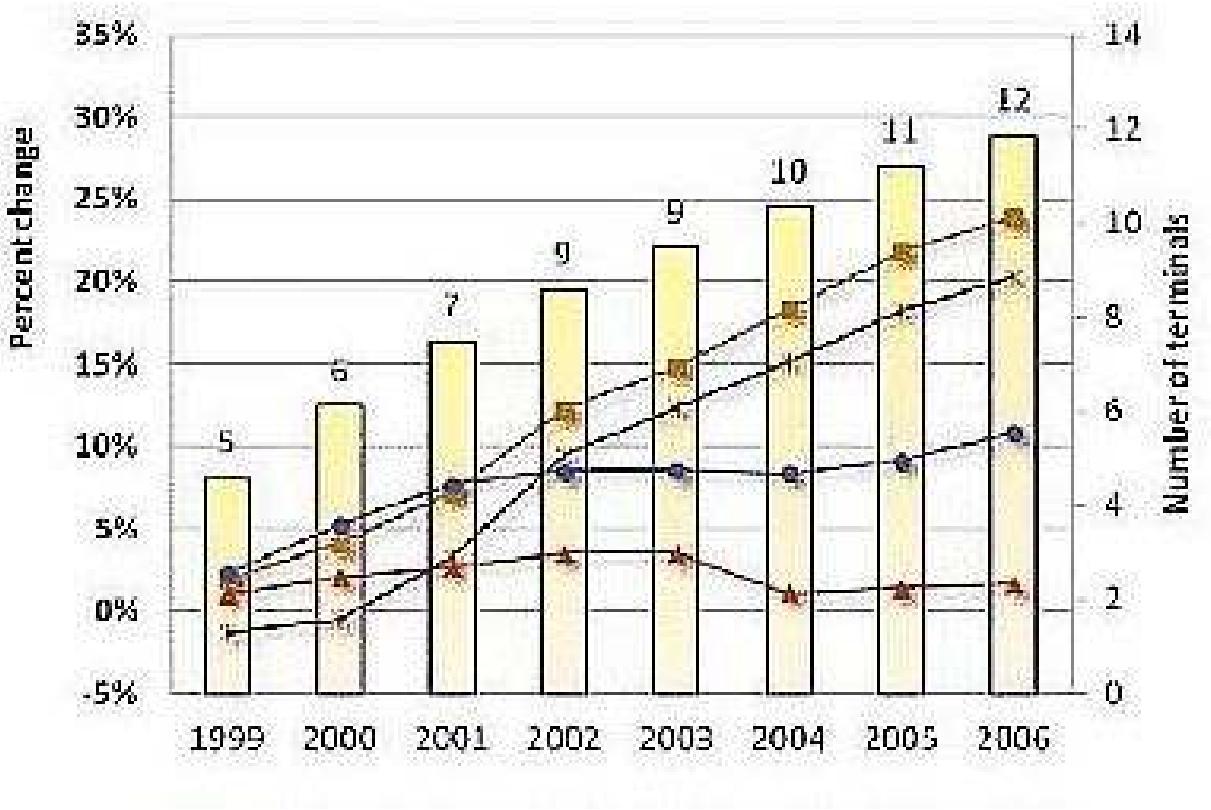
Figure 2: Change in library use and resources since 1998



Source: Samantha Becker et al., 2010, *Opportunity for All: How the American Public Benefits from Internet Access at U.S. Libraries,* Institute of Museum and Library Services.

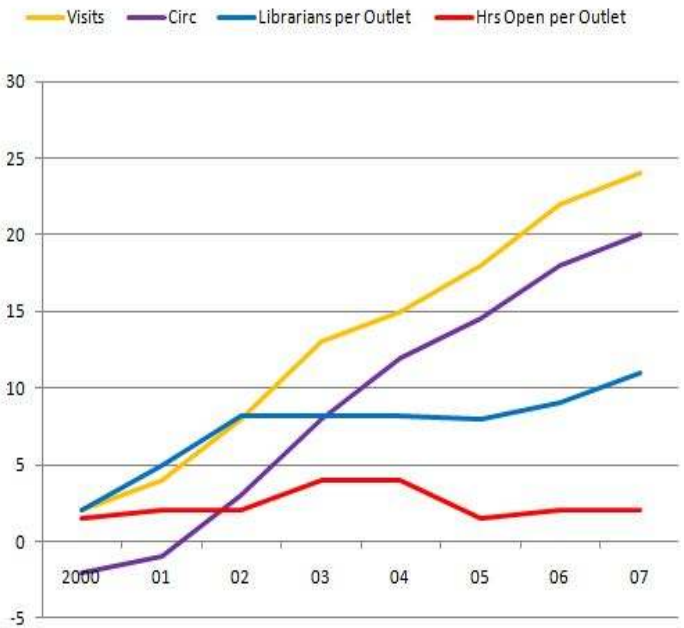
This chart uses a ‘double Y-axis’ that exaggerates terminal installations in US public libraries.

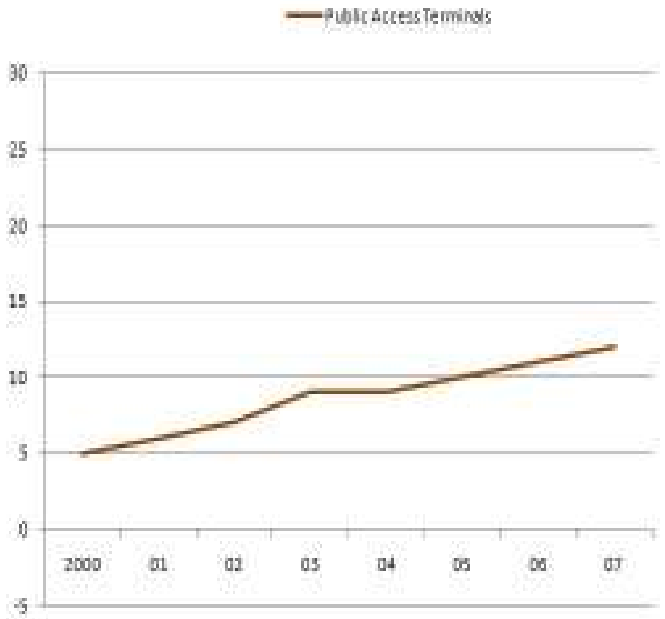
(Scale label would be better as “Cumulative Percent Change”)



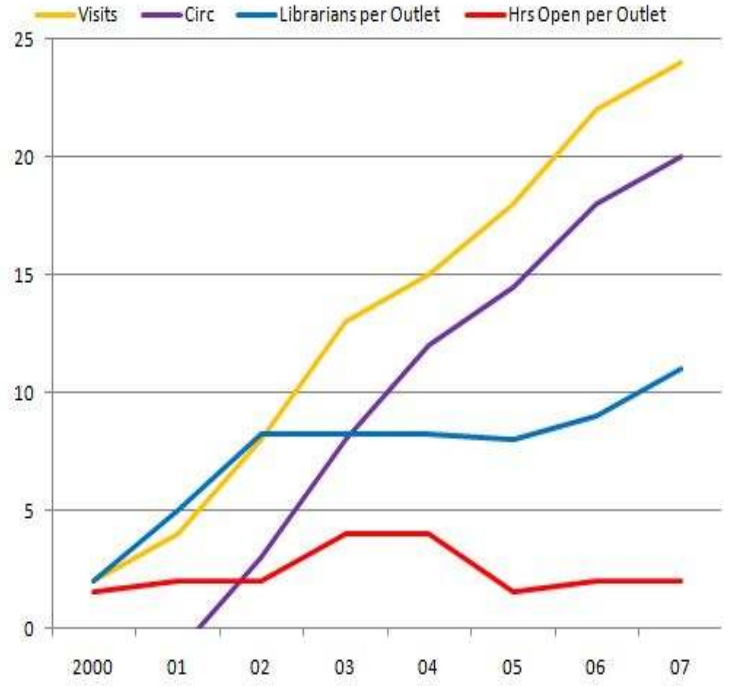
Connected lines = cumulative percent change in 4 public library measures

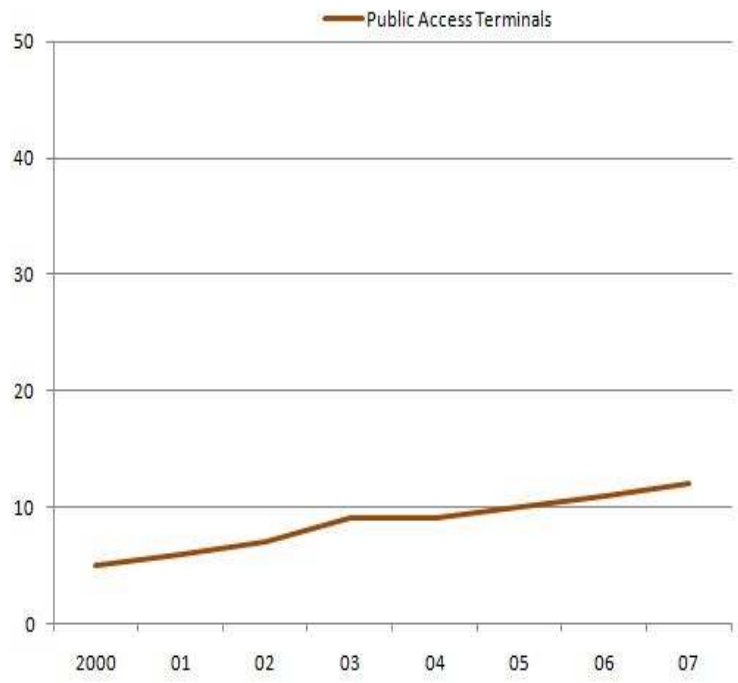
Bars = number of public access terminals per outlet



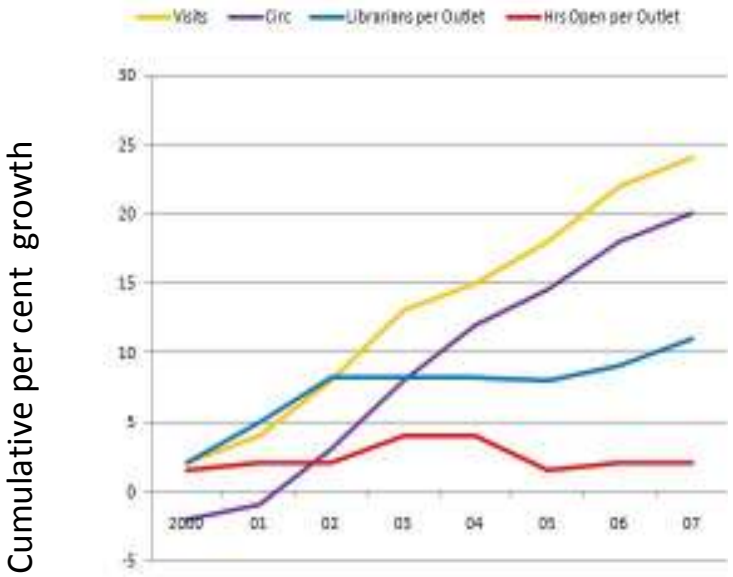


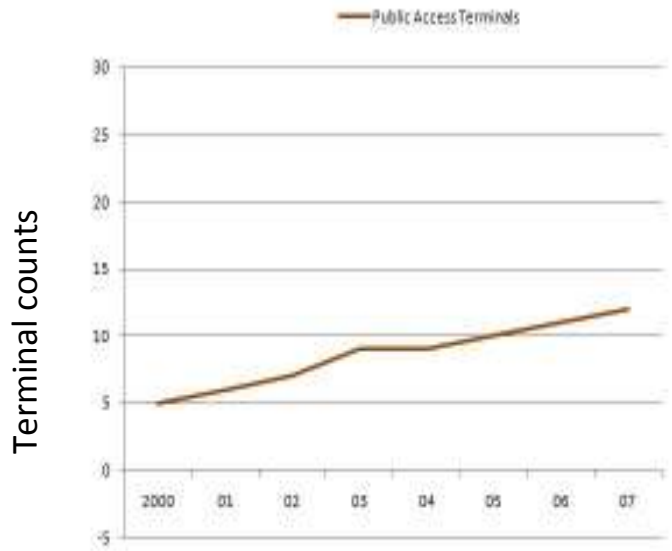
Using equivalent vertical scales (axes) produces an accurate display of the data.





Non-equivalent scaling can make the library statistics trends (left chart) even more impressive.





TWO BIGGER PROBLEMS WITH THIS COMPARISON, whether depicted in the Becker et al. chart or redrawn charts:

1.A comparison of per cent growth to actual counts is specious.

2.Comparisons of per cent growth in data of different magnitudes should be done with caution.

Comparisons on Two Different Vertical Scales

Scaling can easily exaggerate one trend over another

For this reason, comparisons of data having different scales should be done with forethought and care (and trepidation, really)

Comparing Disparate Data

Scaling two measures having very different magnitudes:

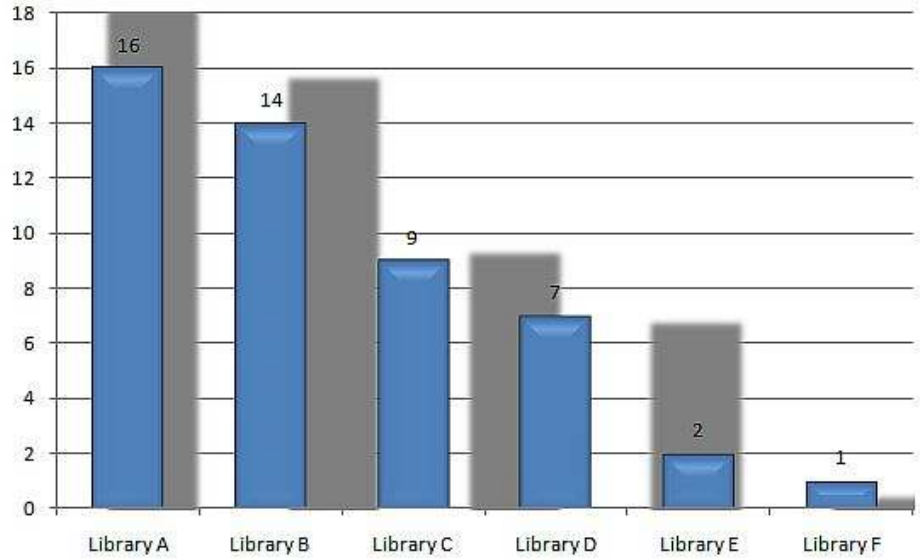
- Set horizontal axes equal to same proportions based on each chart's baseline

Bar Charts

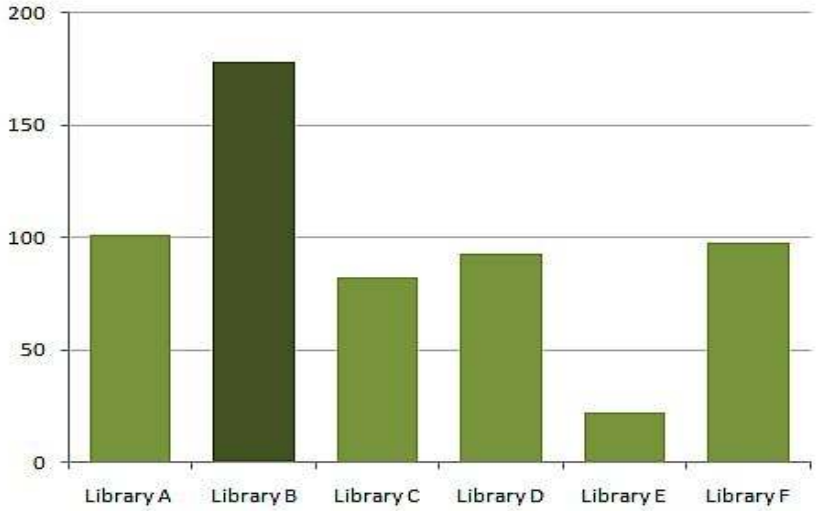
Don't use shading or shadows

Make width of bar about one to two times space between bars

Make projections/estimates paler shade

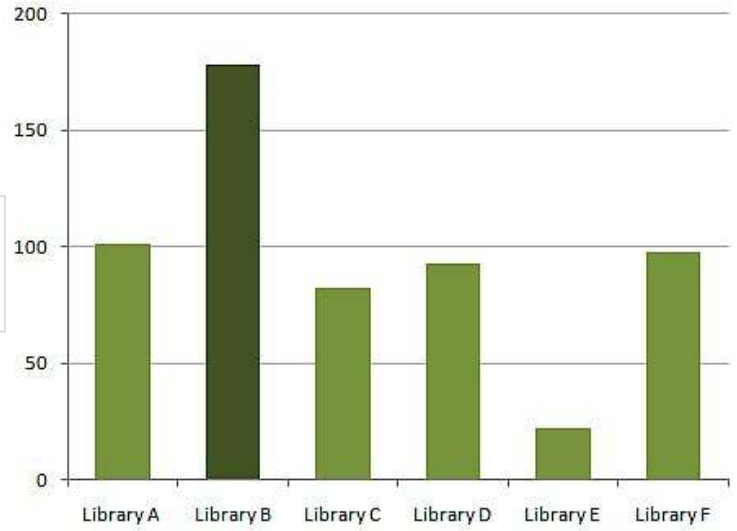


MS Excel designers must believe bar charts are like sundials. They are not. We have no need to see where shadows might or might not fall.



Wall Street Journal standard: gap = ½ bar width

Gap = bar width seems okay also

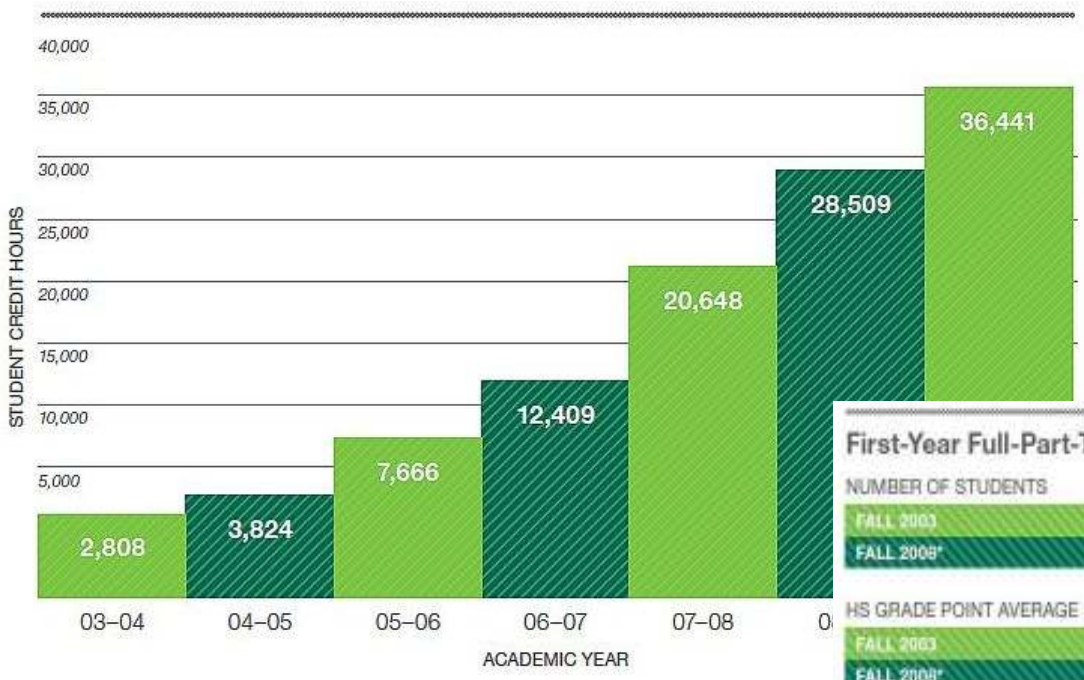


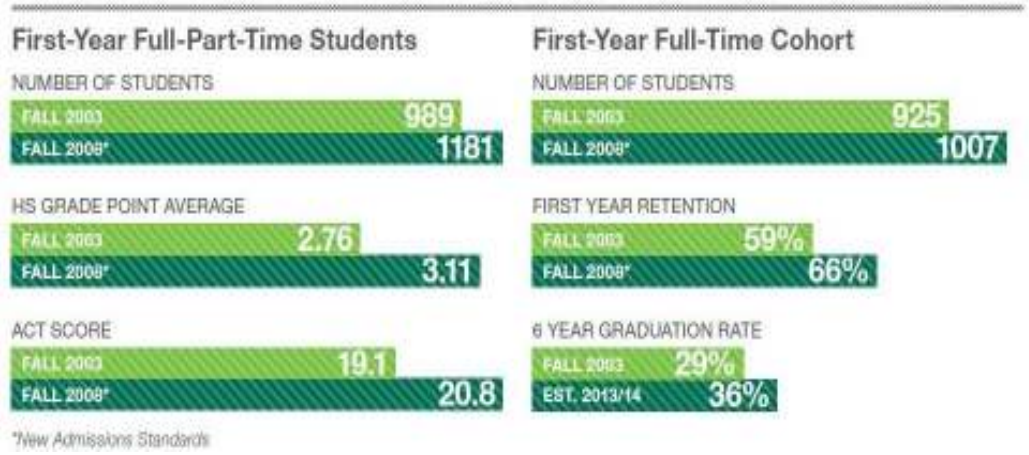
Bar Charts

Don't mix colors or hash-patterns

Never use 3-D. On bar charts the values are impossible to interpret.

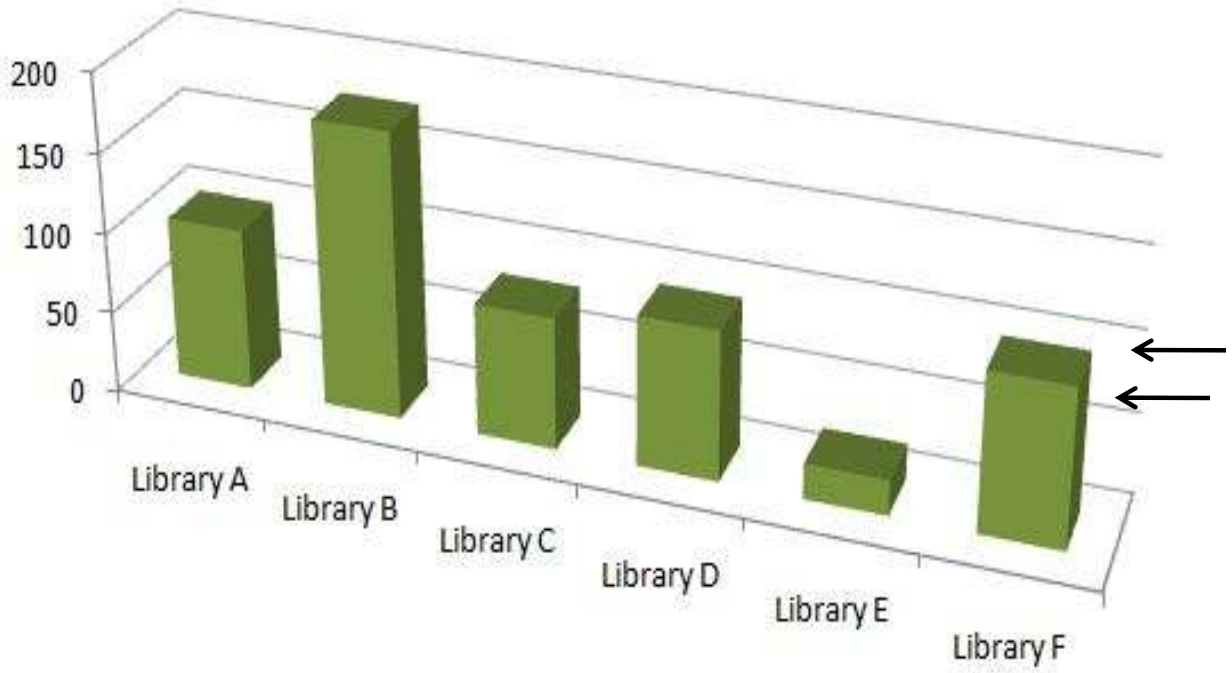
Use gray background to separate negative from positive zone of chart, if desired





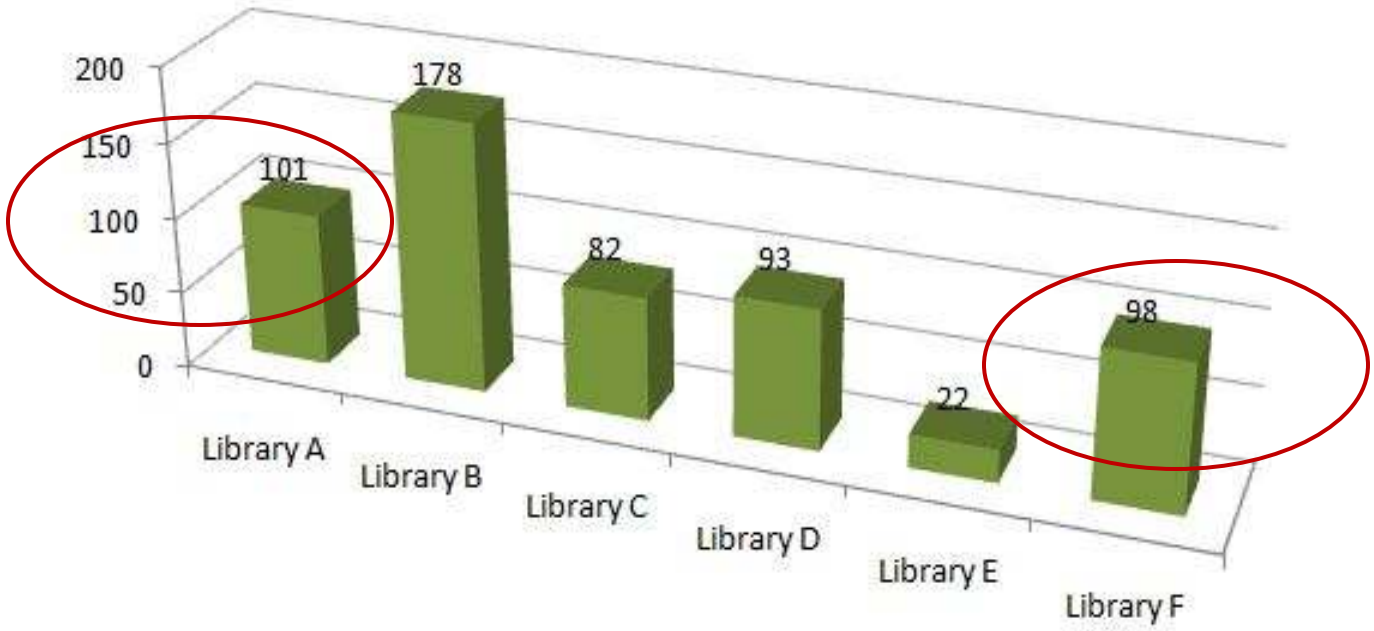
Source : Cleveland State University, *2010, Self Study.*

Circus striping and hash patterns distract the eye from the data.



Which indicates the bar's height? Front edge or back edge?

Never use 3-D. Never.



MS Excel 2007 depicts bar heights inaccurately.

Bar Charts

Plot bars from a zero baseline

Non-zero baselines permissible for special purposes

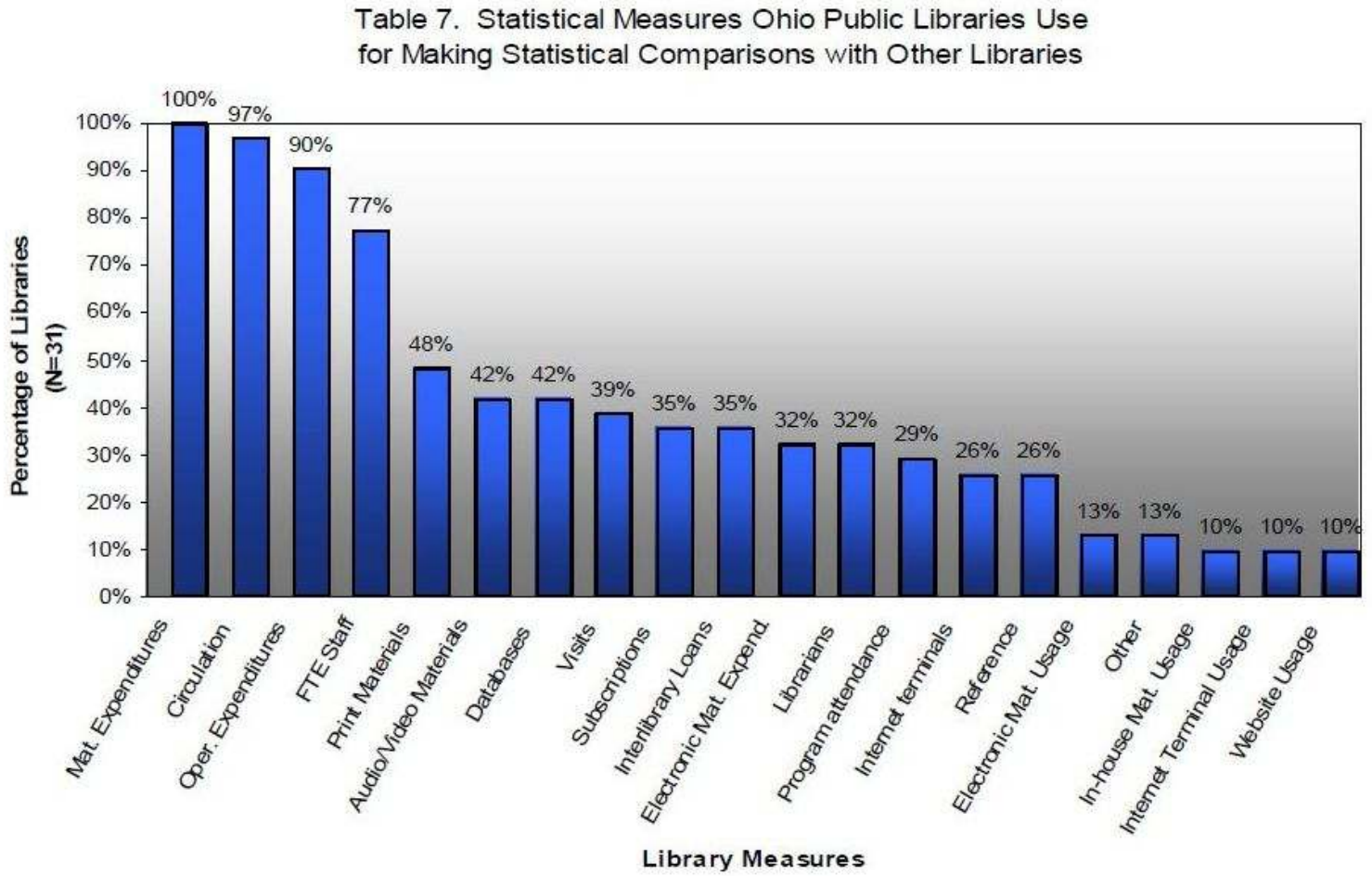
When bars are similar in height so as to be indistinguishable, consider plotting the differences between the values.

Bar Charts

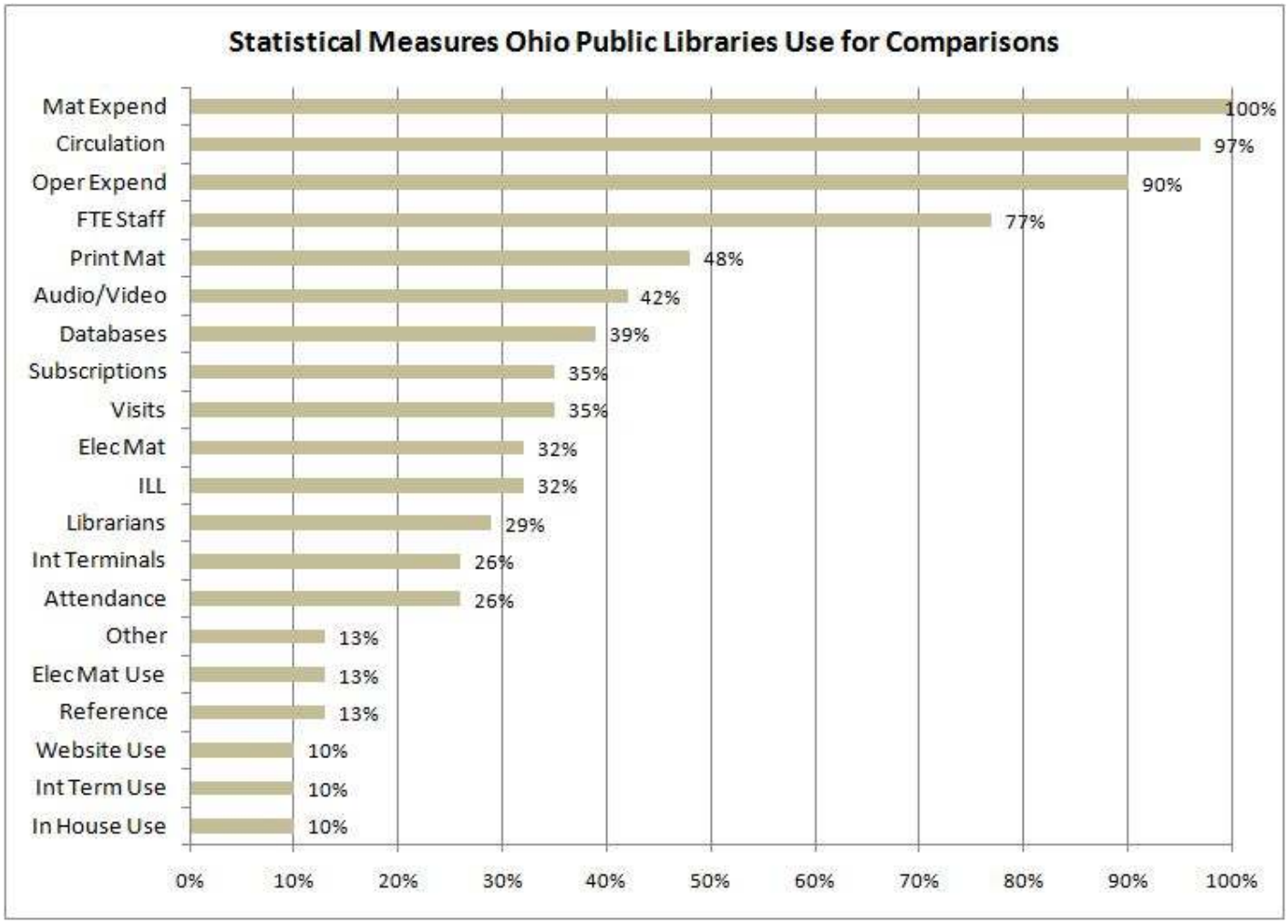
When a bar is so small it is close to zero, label it

Do not use angled tick mark labels; redo as horizontal bar chart instead

When using bars shaded in a single hue, go left to right from lightest to darkest



No angled text



Horizontal bar charts avoid angled text

Bar Charts

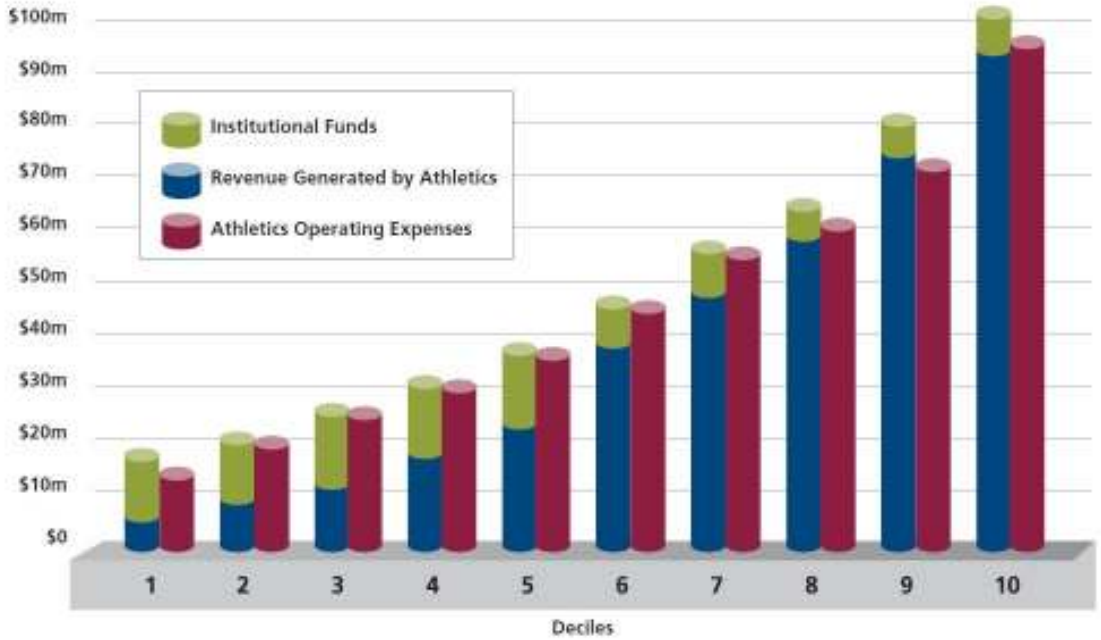
For multiple category charts, maximum number of categories is four

Limiting to three categories is preferable since it is difficult to decipher more

Colors don't help; they make it worse

Bar Charts

To repeat: No ‘circus‘ or ‘zebra’ stripes, hashed, polka-dot, 3-D, or other cutesy patterns



Source: Knight Commission on Intercollegiate Athletics, 2009, *Restoring the Balance.*

Leave eye-catching shapes and colors for *Cirque de Soleil*.

Go for simplicity and clarity.

Bar Charts

If depicting exceptionally high values (‘outliers’), put break mark (zig-zag) in bar to show the gap in that bar

Label the outlier value

Make that broken bar much taller than other bars to indicate the magnitude of the data value

Horizontal Bar Charts

Typically used to rank items by a single characteristic, e.g. rankings; rank from largest to smallest or vice versa

When listing in order by time interval, start with most recent interval first

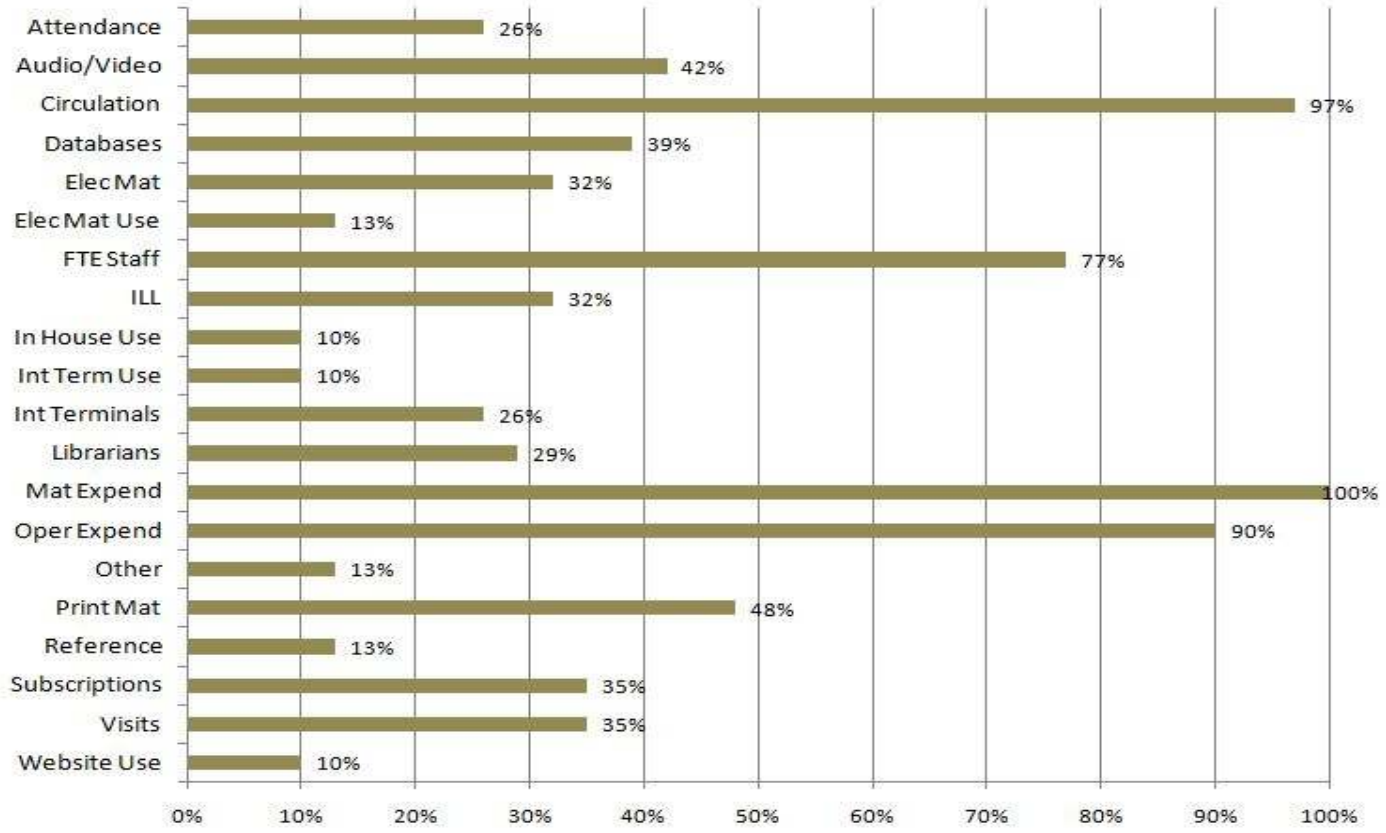
No shading or 3-D

Horizontal Bar Charts

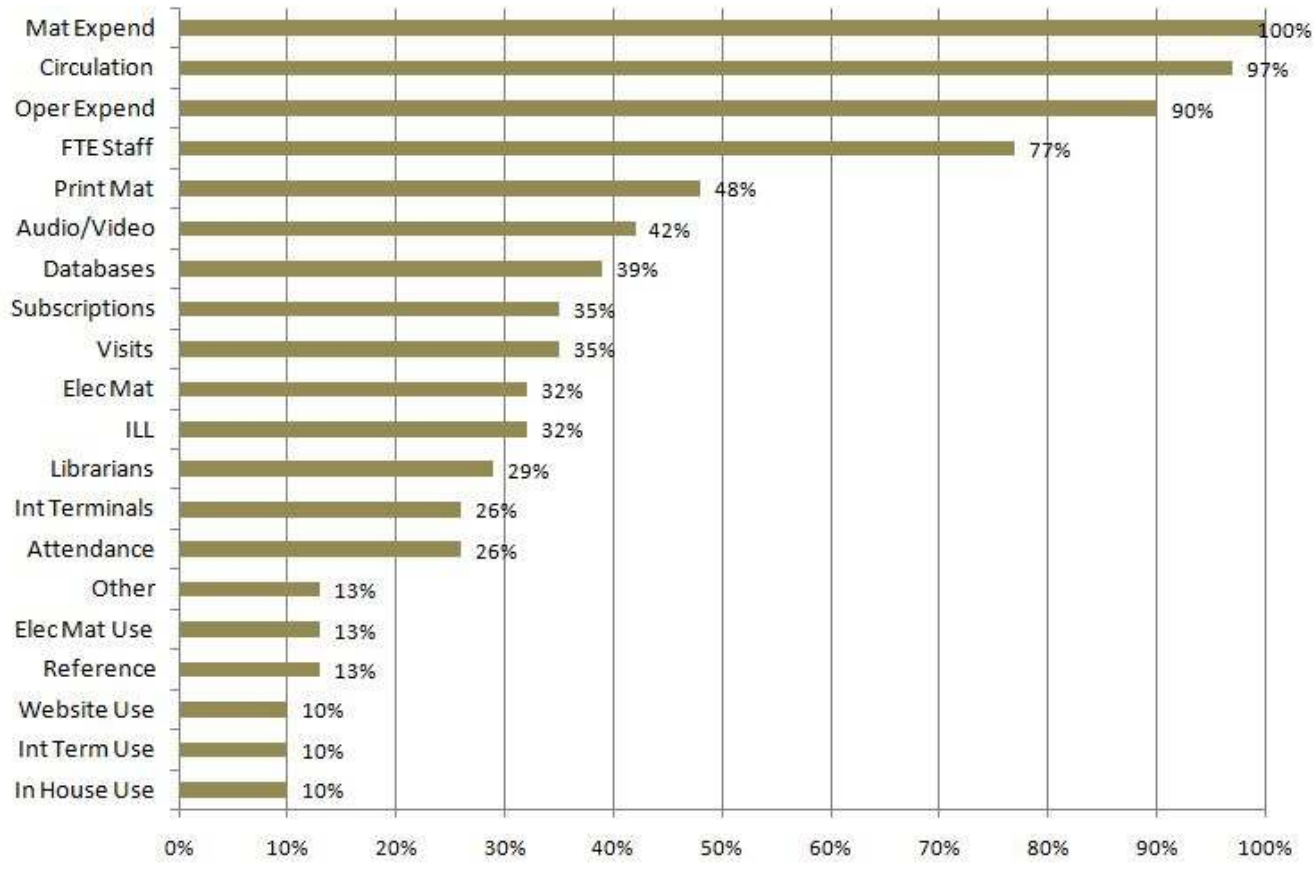
For a long list, label data points to the right, flush

Order by important magnitude, not random or alphabetical order, unless chart is meant for lookup use

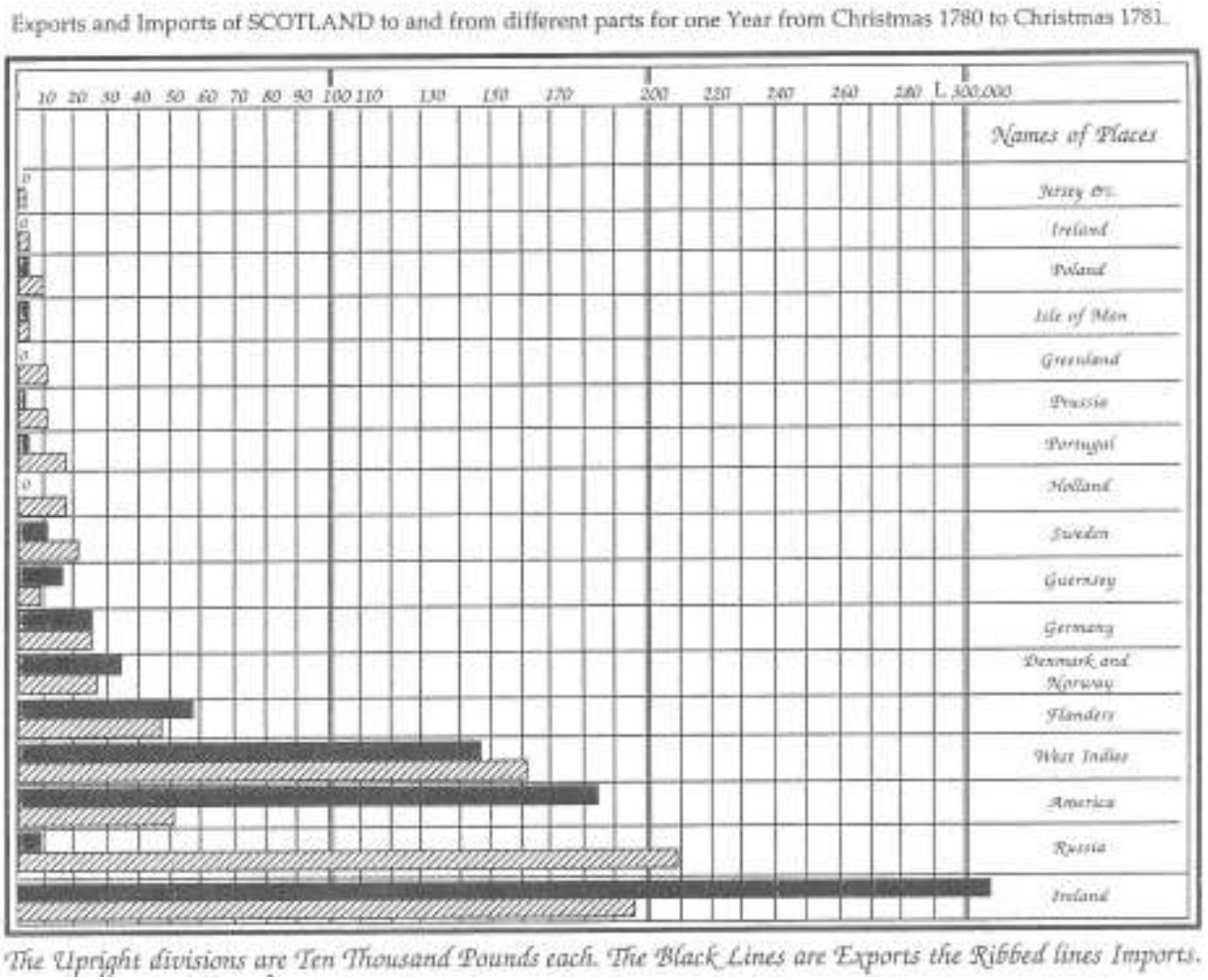
With negative numbers, place zero on vertical axis at right and show negative numbers to the left



Alphabetical ordering makes comparisons difficult.



Sort horizontal bar charts in the order of the data.



Source: Howard Wainer, 2005.

In 1786 William Playfair realized the importance of sorting the order of the data (small to large).

Pie Charts

Don’t use them. The eye cannot accurately evaluate relative sizes of the pieces, except for very simple pies.

If you cannot resist using pie charts, follow basic principles of graphical data presentation.

Pie Charts

Start at “noon” on the circle and place largest segment first (to emphasize its importance)

Places second largest slice on left of noon.

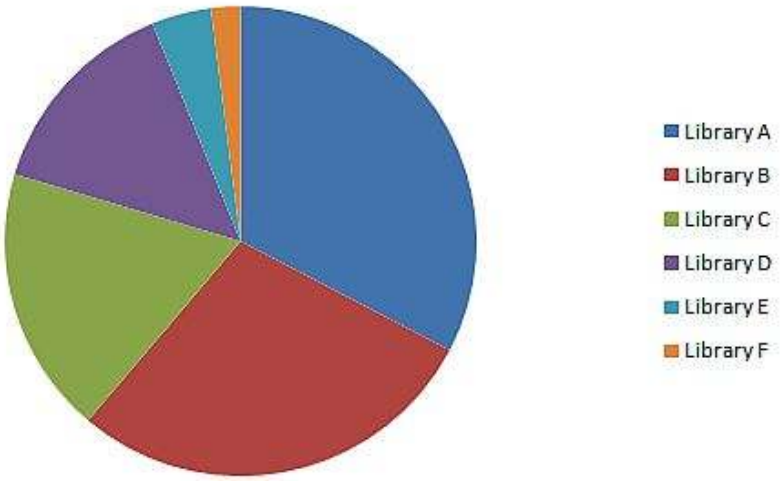
If all slices are close in size/value, order by size, clockwise.

Pie Charts

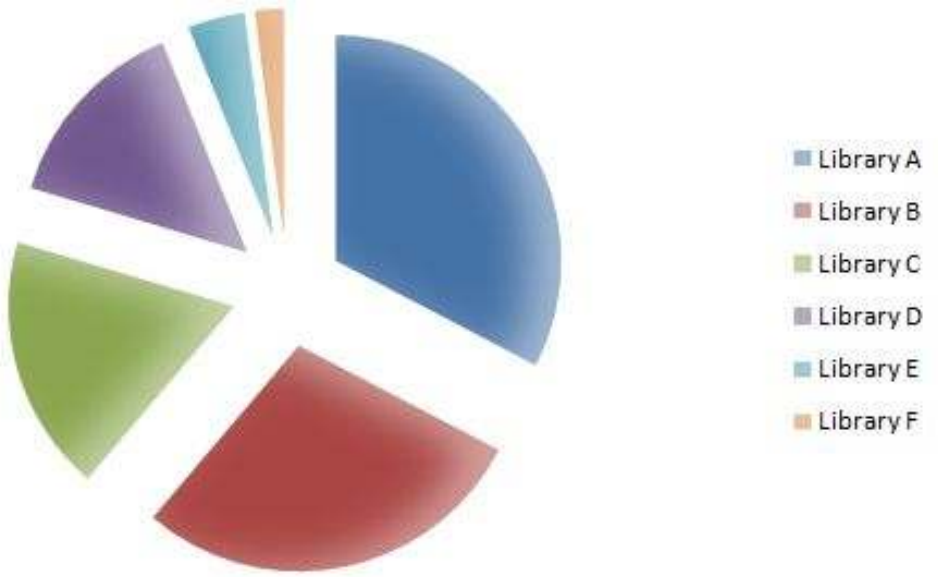
Use no more than 5 slices

No bright, contrasting colors

No special effects (shading and pull-out of one slice)



MS Excel 2007 pie chart default coloring

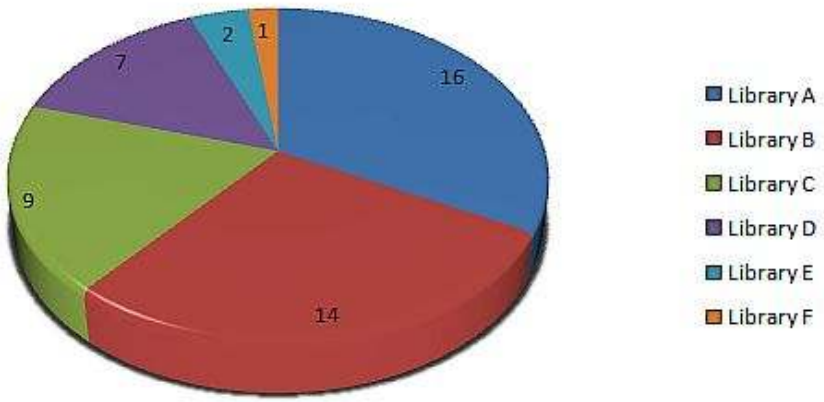


Separating slices and shading/shadowing make gauging proportions even more daunting.

Pie Charts

3-D pie charts are inaccurate since they mask real sizes of slices

Don't subdivide a pie slice and represent as another pie. Don’t make users do the math.



Which slice is largest in this 3-D pie?

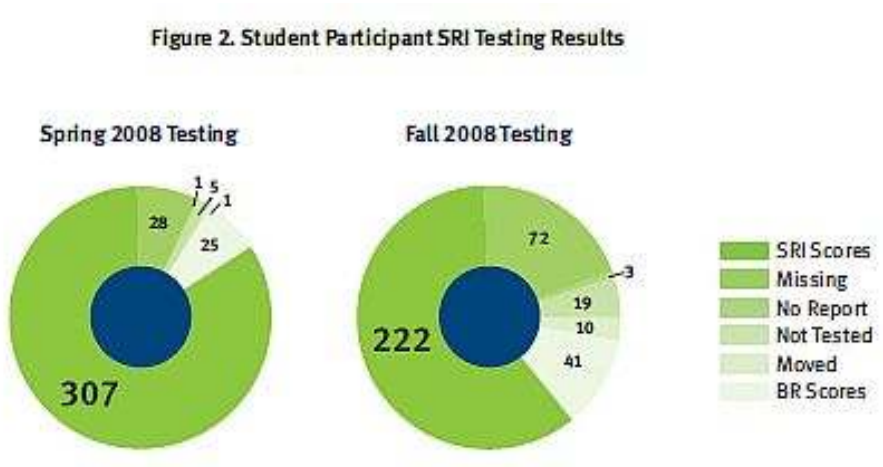
Pie Charts

Don't chart use donut-pie charts. Center circles distort the proportions

If comparing multiple pies, always display in proportion to their values.

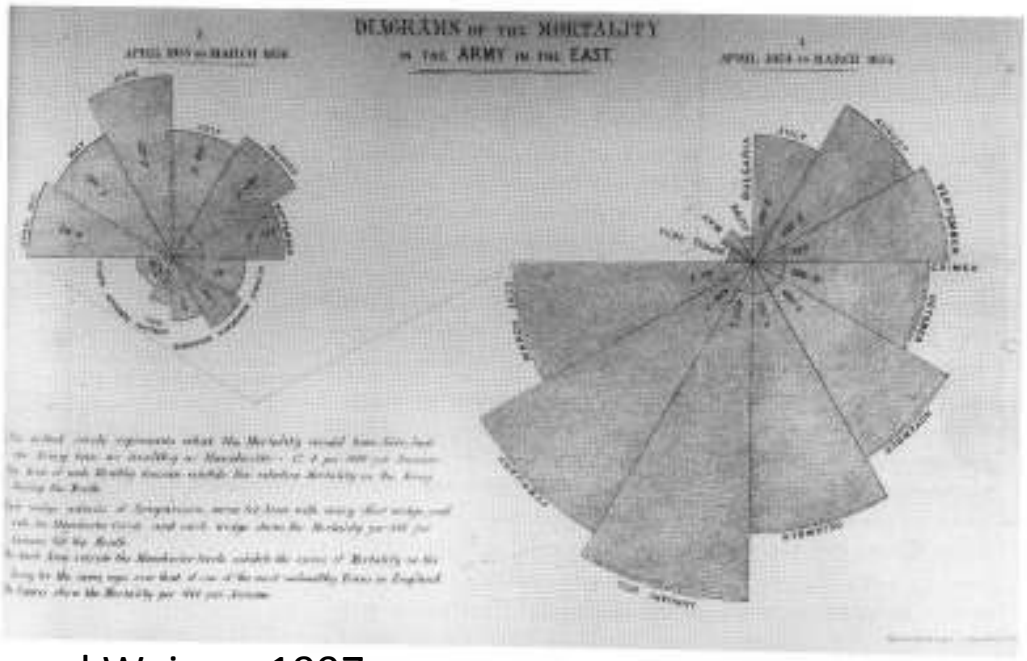
Recall the formula:

$$\text{area of a circle} = \pi r^2$$



Source: Roman et al., 2010, *Public Library Summer Reading Programs Close the Reading Gap*.

Pie centers are deceiving and distracting



Source: Howard Wainer, 1997.

Florence Nightingale's 'rose charts' used circular graphics appropriately and effectively. Data proportions are much easier to decode since length of slice faithfully reflects the data. Wainer tracked the history of rose charts back to the 15th century.

EXERCISE 3

Needs assessment

1. Identify problems, shortcomings with the graph.
2. Make a list the graph's shortcomings, based on the principles presented in the class.
3. Try to determine if there is a story in the data that needs told.

Design and production

1. Decide how the group will address artistic design and production functions: (a) consensus or (b) appoint *leaders* for each function
2. Determine improvements needed; perhaps they will be minor, perhaps major; or perhaps you're group will reach an impasse.
3. Create prototype design (rough sketch) - guided by the *artistic leader*, if you choose to appoint one
4. Create final graph - under the direction of *production leader*, if you chose to appoint one

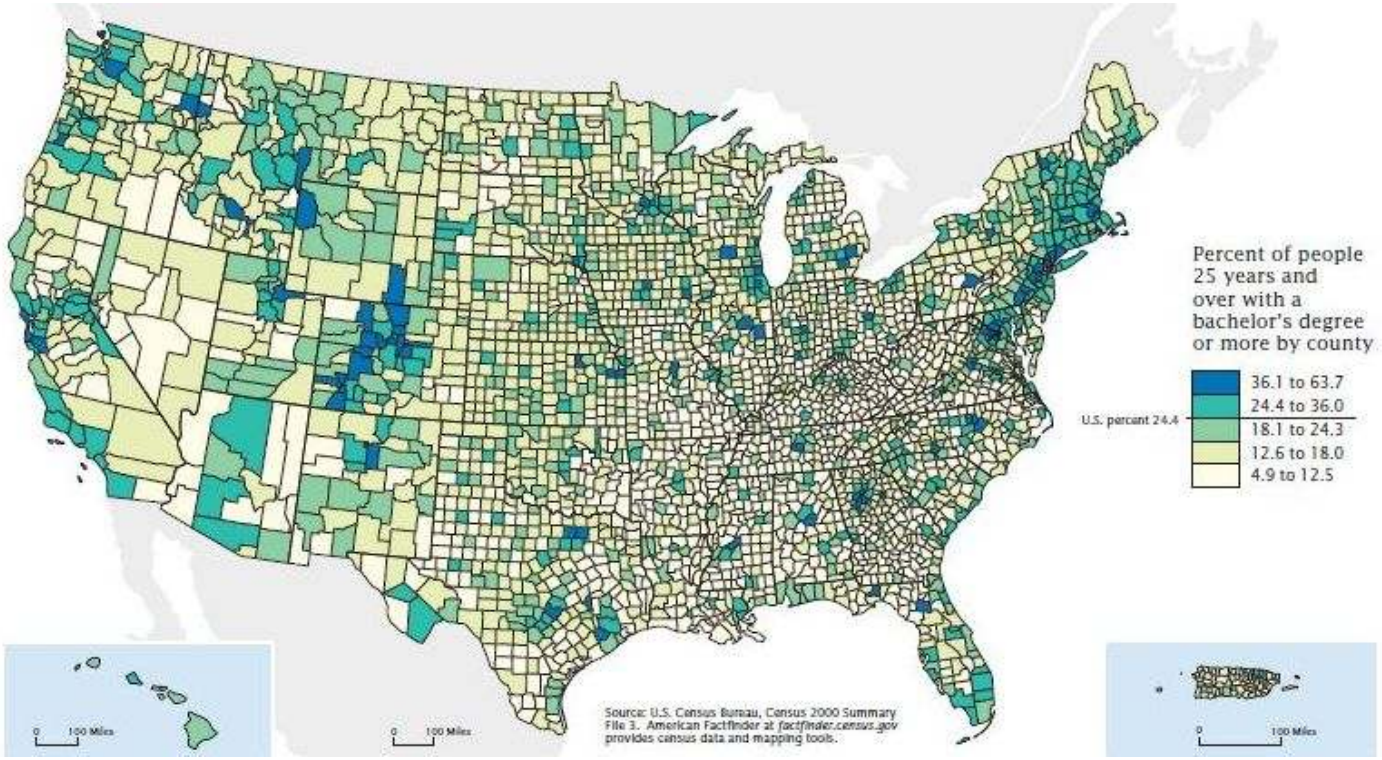
* Don't worry about re-producing scale and data values precisely. Create and draw the general idea. Make up numbers, scaling as needed.

Geographic Maps

Trendy, but this graphical form typically does not clarify or illuminate data

Maps do not accurately represent magnitude due to difference in state/province sizes

Use only when spatial distributions are central to the analysis



This map obscures rather than illuminates the data. Large US counties in the west and southwest are visually prominent. Small counties in the northeast with highest density of graduates are nearly invisible.

Source: Bertot et al., 2007, *Public Libraries and the Internet,* Information Use Management and Policy Institute, Florida State University

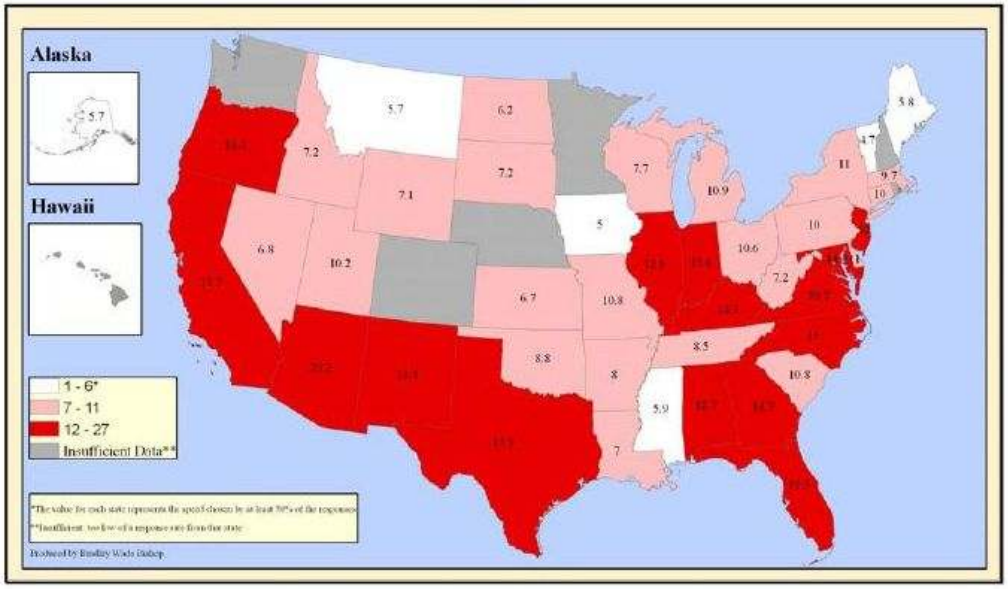


Figure GIS-1. Average Number of Public Access Workstations by State.

Maps are poor information channels for portraying comparisons of magnitude. Here states color-coded white (lowest per outlet category) stand out more than states in the middle (pink) category. Gray states are visually similar to pink states even though their information value is nil.

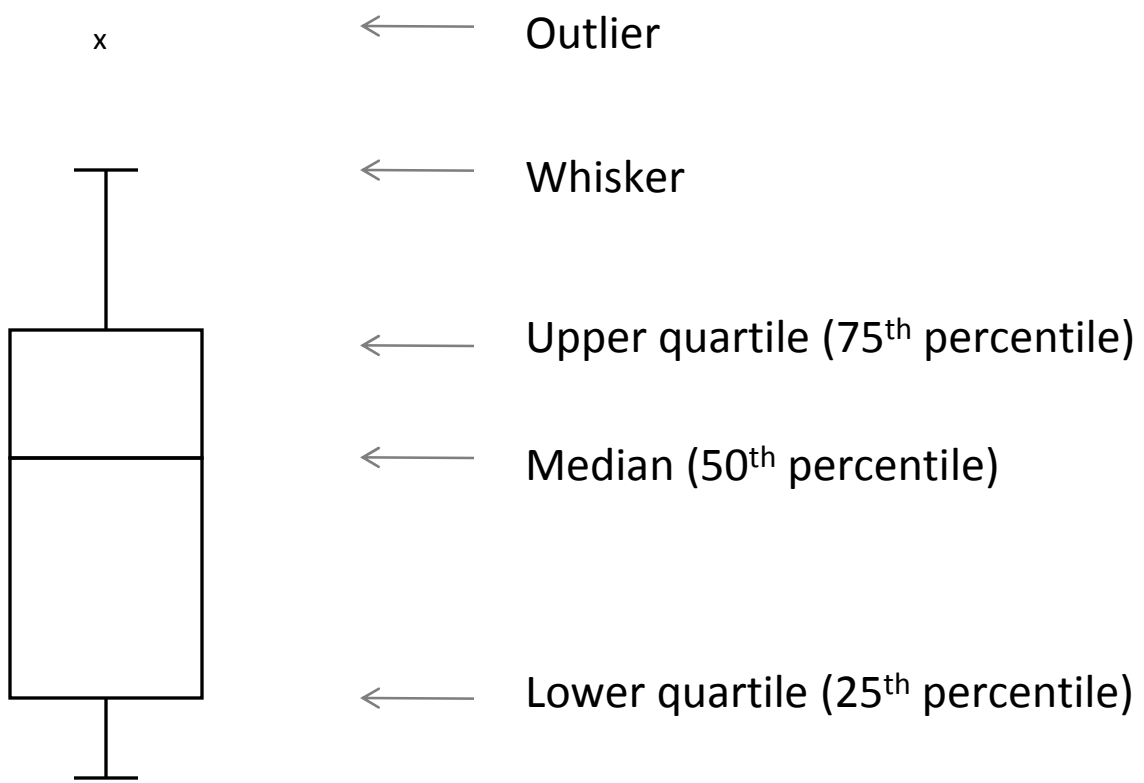
Box Plots

Also called ’Box-and-Whisker Plots’

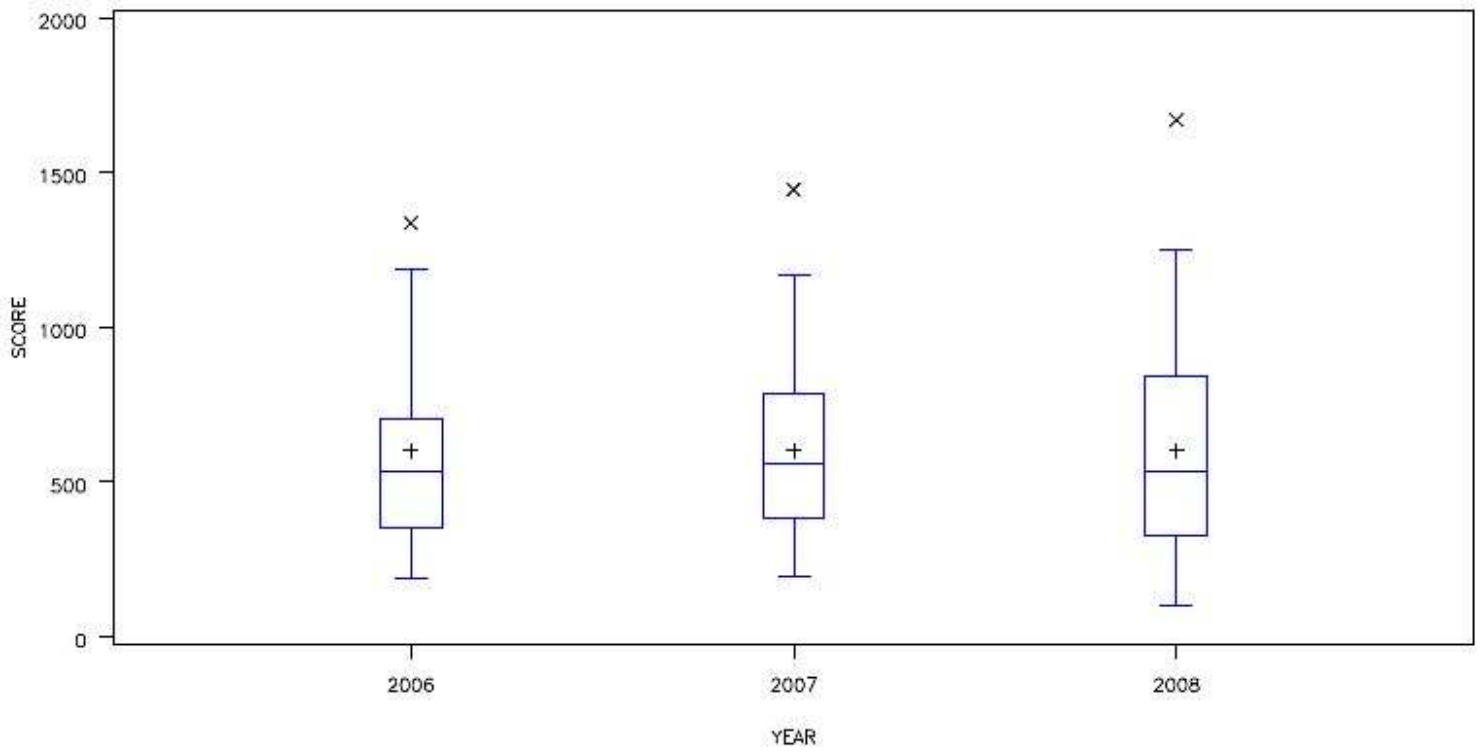
Developed by John Tukey

Describe basic aspects of distribution of a set of data (range, median, outliers, etc.)

Makes comparisons of distributions easy



Box Plot Components

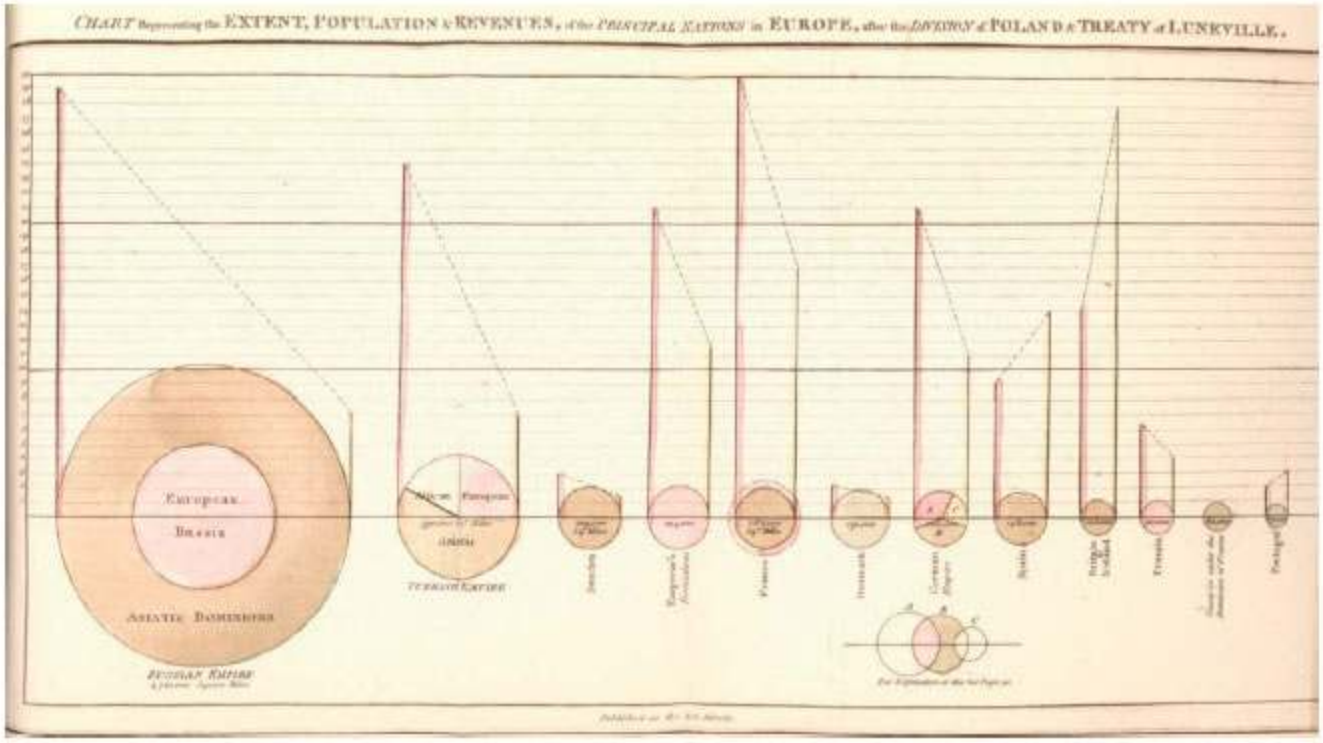


Library Journal Index scores for US public libraries with expenditures $30 million and above. ‘+’ indicates mean value.

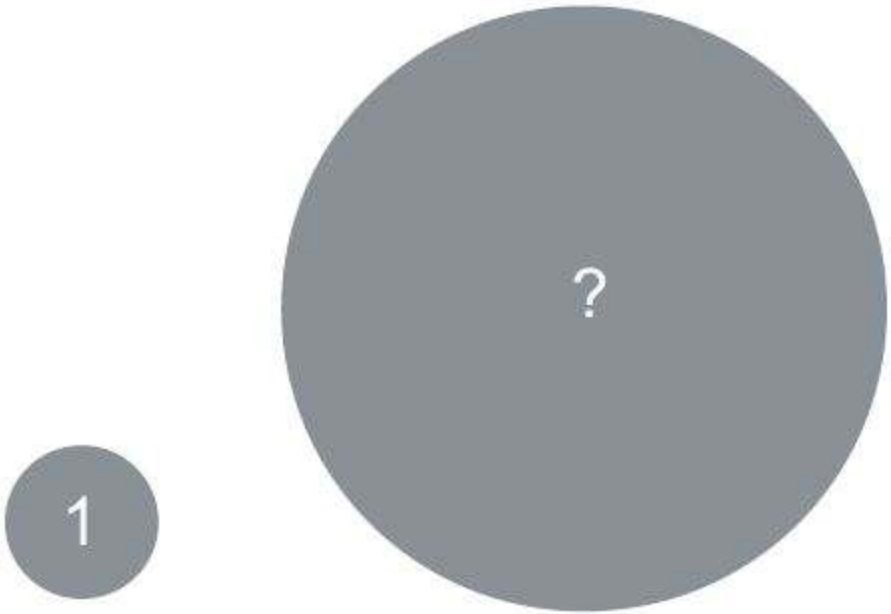
Area Plots

William Playfair was probably the first to represent data magnitude using different sized circles (the precursor to the dreaded pie chart!)

Circles are too difficult for judging comparative sizes.

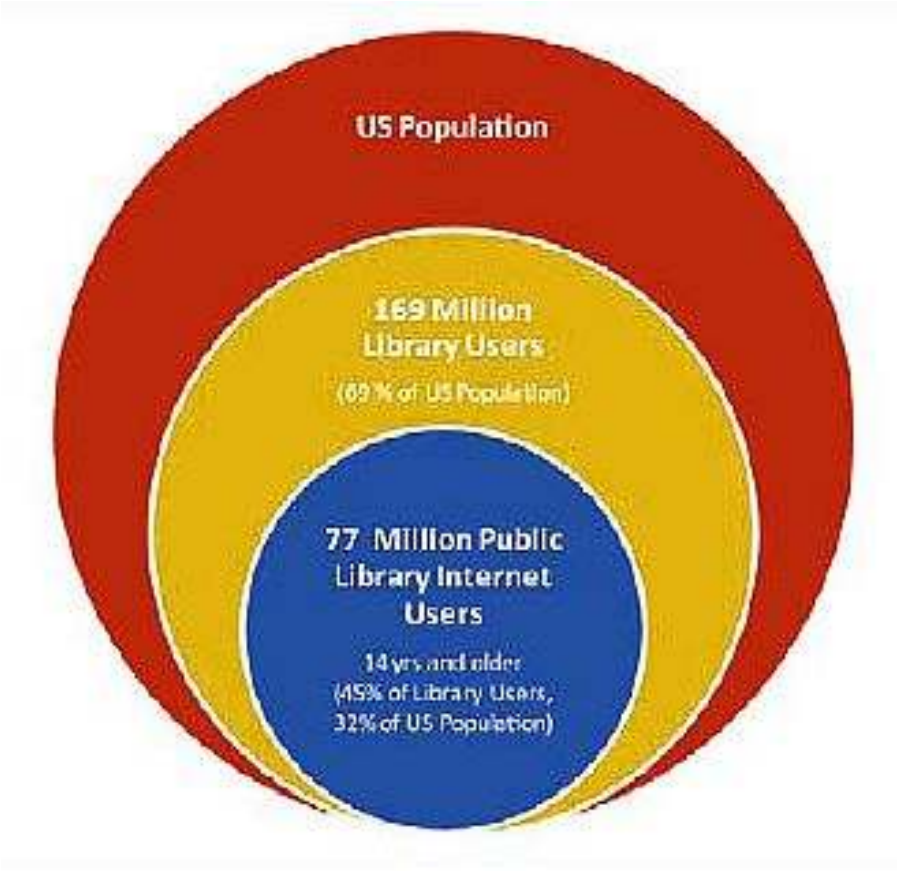


Judging comparative magnitude using circles is not intuitive. William Playfair's alignment of circles by height amounts to a square root conversion, since ½ of a circle's height (the radius) is proportional to the square root of its area.



Source: Stephen Few, 2007, *Save the Pies for Dessert,* Visual Business Intelligence Newsletter.

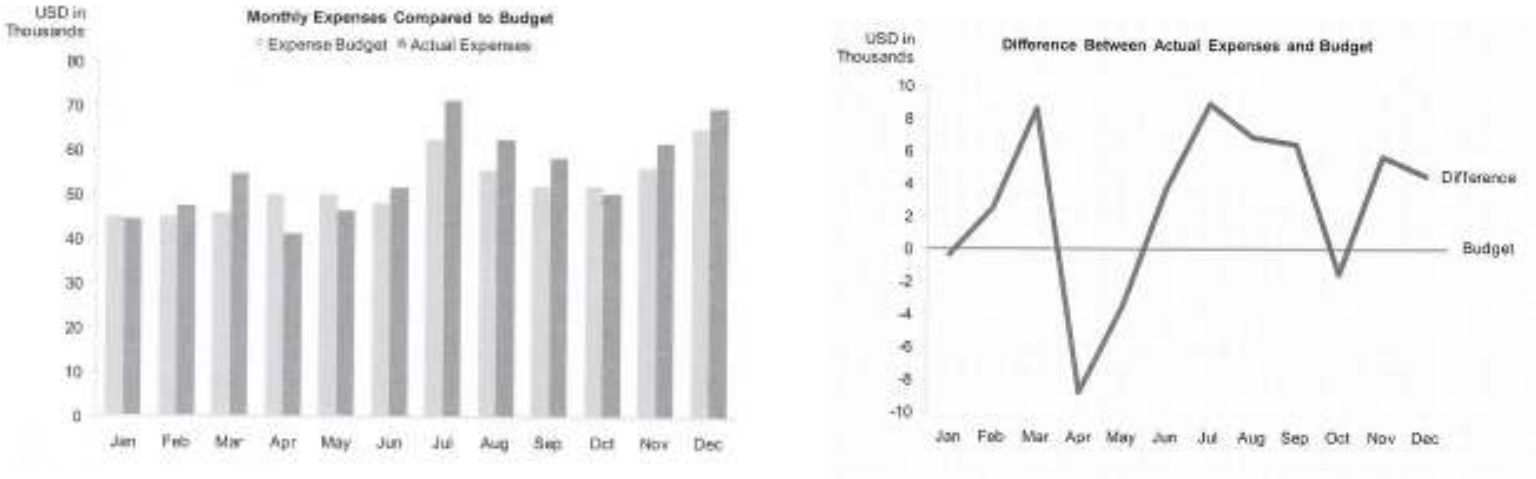
Homework Problem



Source: Samantha Becker et al., 2010.

Redraw this graphic so that the areas of the circles are proportionate to the numbers represented.

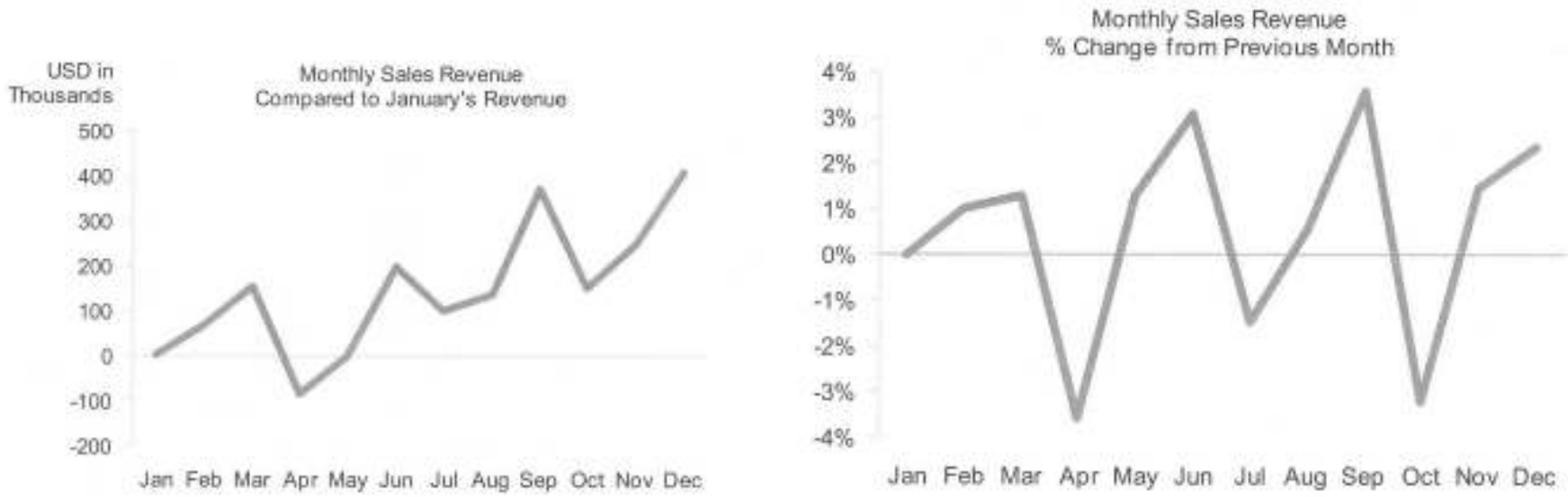
Re-Expressing Data



Source: Stephen Few, 2009.

When comparing two-variable bar chart depicting time intervals, re-charting difference for each interval enhances the comparison.

Re-Expressing Data



Source: Stephen Few, 2009.

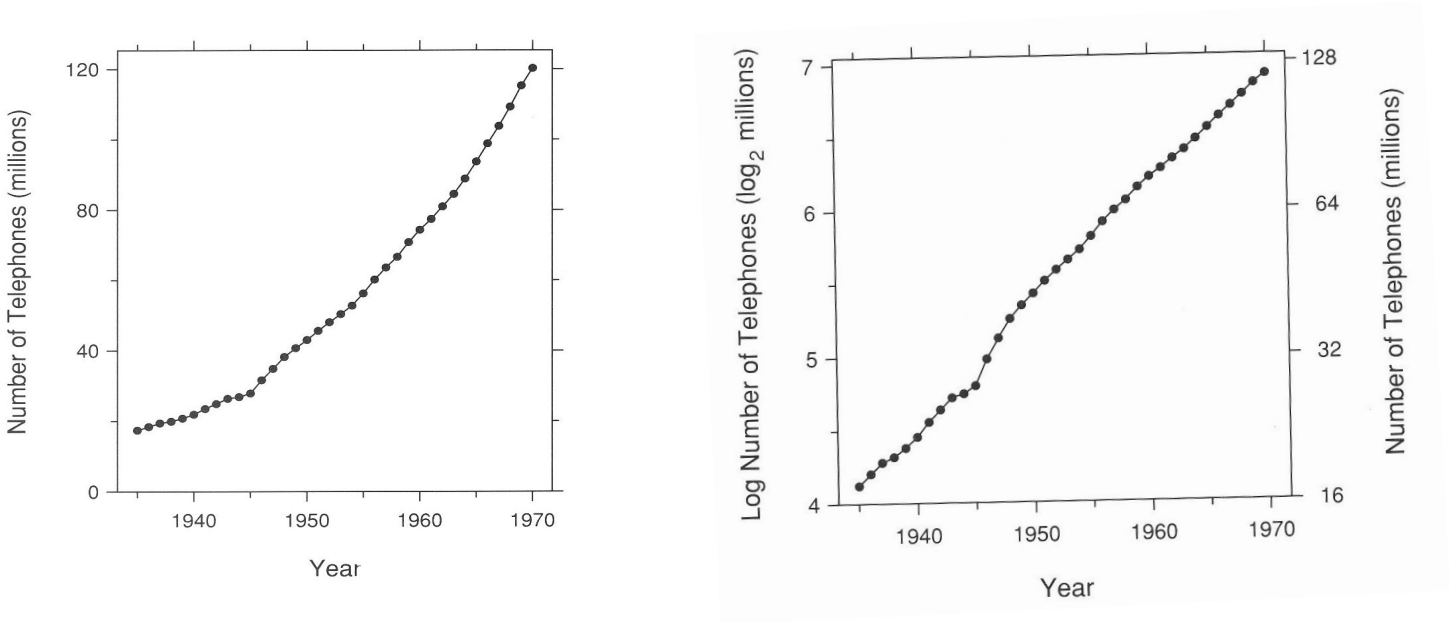
When viewing changes in a trend over time, it may be useful to look at periodic rate of change in the data also

Re-Expressing Data

John Tukey, the grandfather of *exploratory data analysis*, emphasized re-expressing ('transforming') statistical data

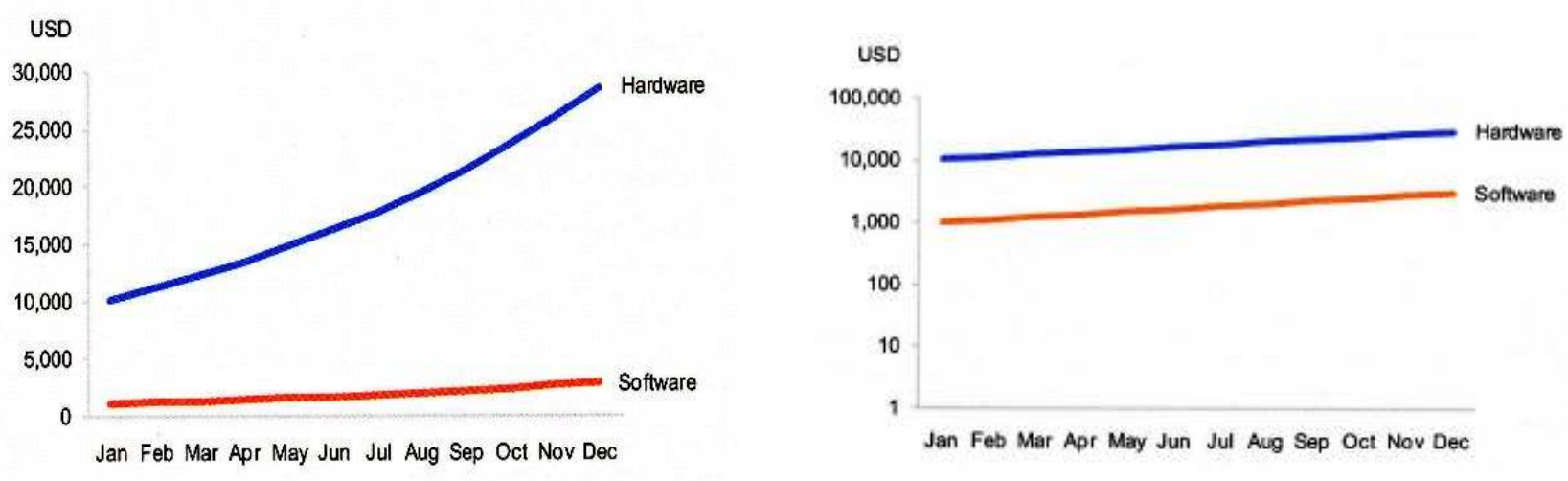
Re-expression maintains the information value of the original data

One common transformation = logarithms



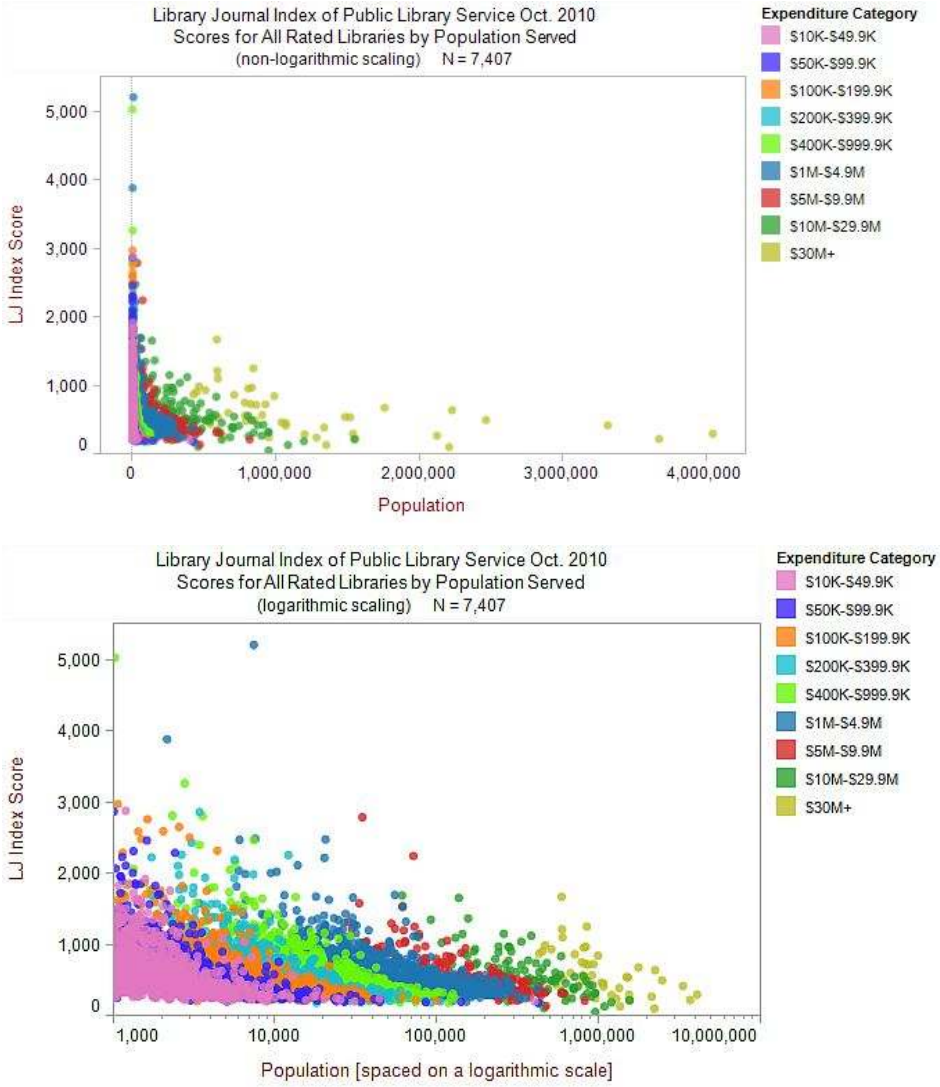
Source: William Cleveland, 1994.

The same percent of an increasing base amount yields larger numbers over time. When displayed in logarithmic form in the right chart, growth of the data is shown to be fairly constant.



Source: Stephen Few, 2009.

Converting these two data series (software and hardware sales) to logarithms enables us to see that both are growing at the same rate over time.

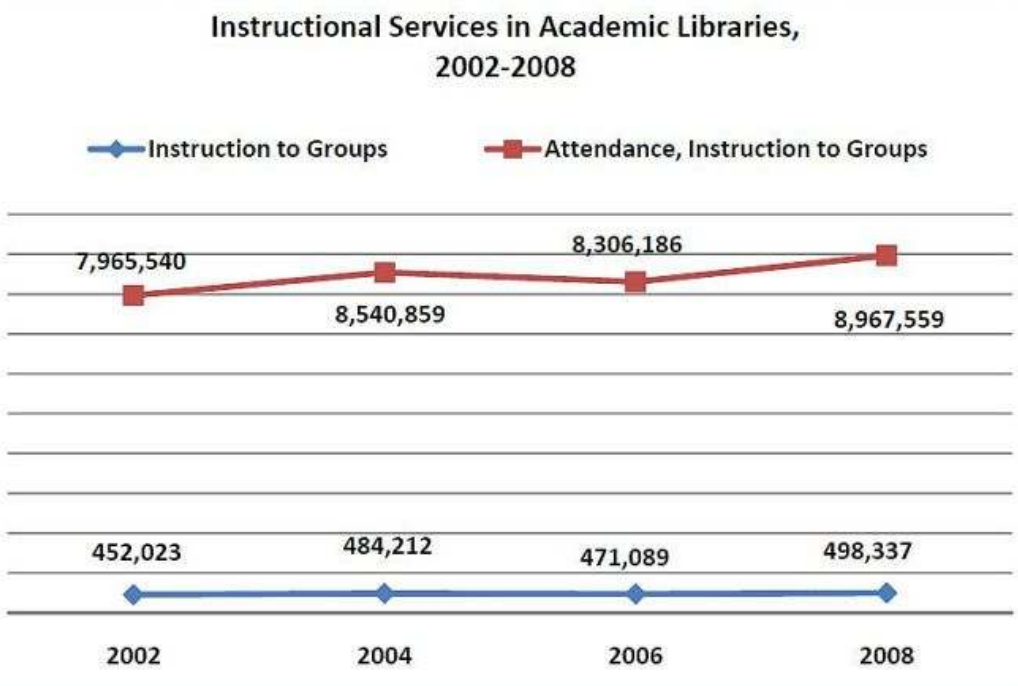


Logarithmic scaling used to extend the horizontal axis when data are clustered toward one end of a scale (skewed).

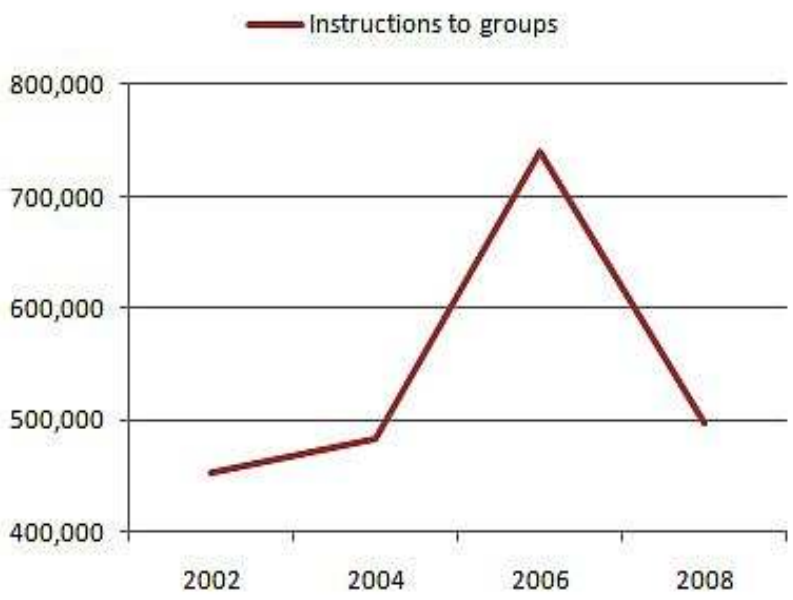
Comparing Disparate Data

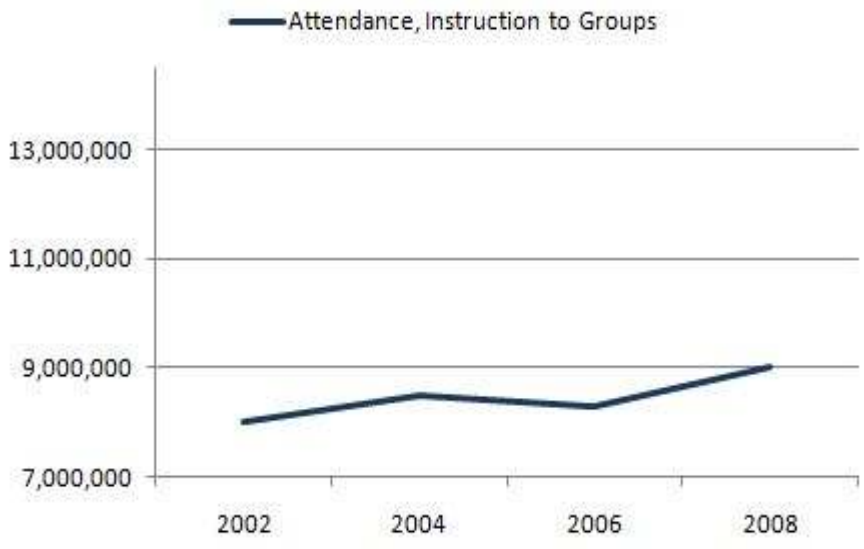
Scaling two measures having very different magnitudes:

- Set horizontal axes equal to same proportions based on each chart's baseline

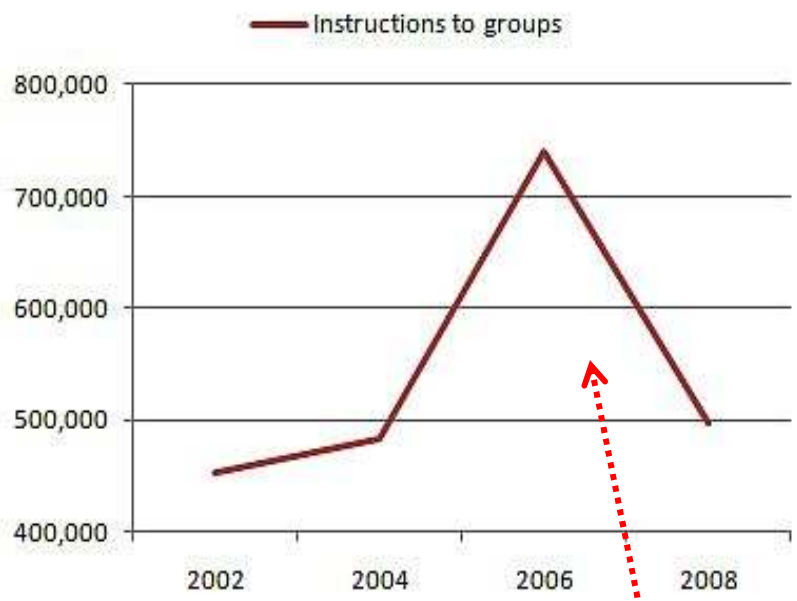


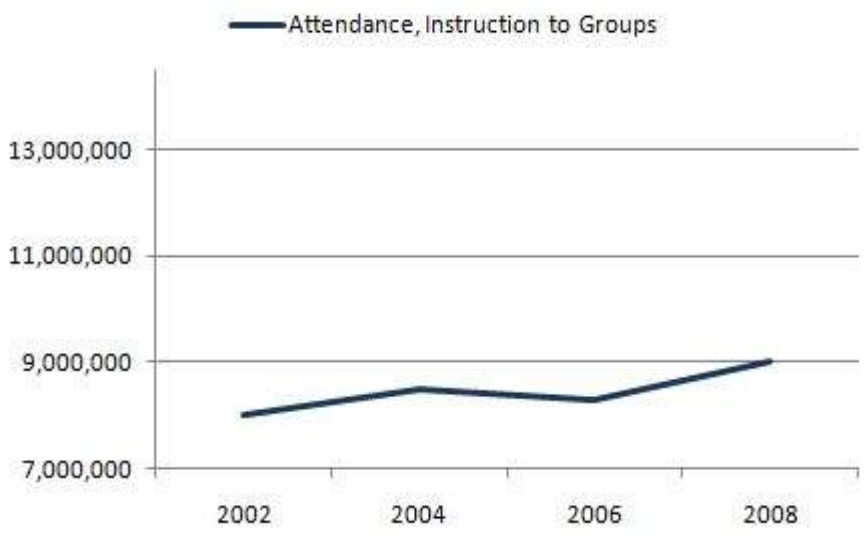
Source: American Library Association, *Condition of Libraries: Trends, 1999 to 2009*





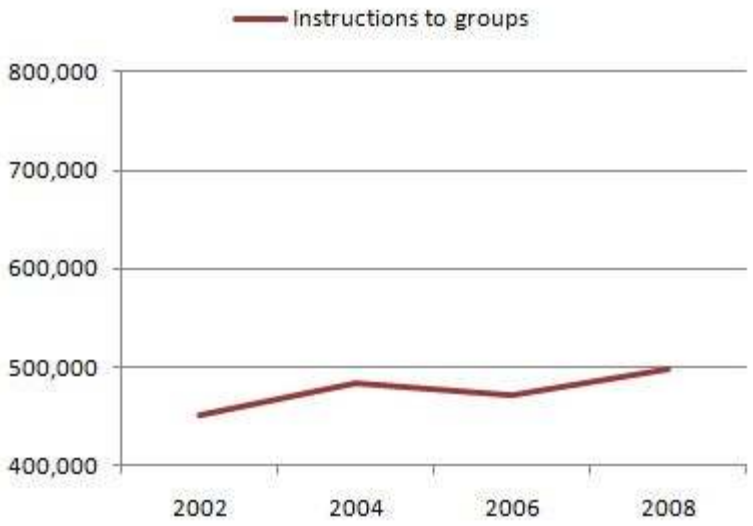
Using each series' lowest value, choose a baseline (400,000 and 7,500,000 in these charts). Set each horizontal scale to twice its baseline, or to some other equal proportion. The scales then represent an equal proportion of each measure's baseline value.

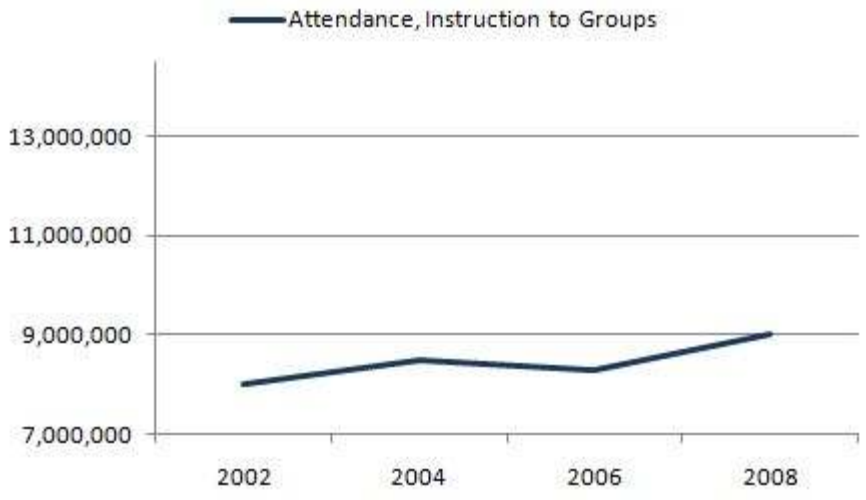




Oops. I typed 741,000 instead of 471,000.

Using ea
(400,000 and 7,500,000 in these charts). Set each horizontal scale to twice its baseline, or to some other equal proportion. The scales then represent an equal proportion of each measure's baseline value.





The left chart now has the correct data.

Remember to proofread.

Which Charts To Use

Line Charts

- To analyze trends, patterns, and exceptions

Bar Charts

- To investigate specific comparisons in time
- To compare categorical data

Scatter Plots

- To visualize how two attributes vary together

Box plots

- To view and compare distributions

Fonts and Typography

Avoid hyphenation

Can use serif and sans serif fonts together (tastefully)

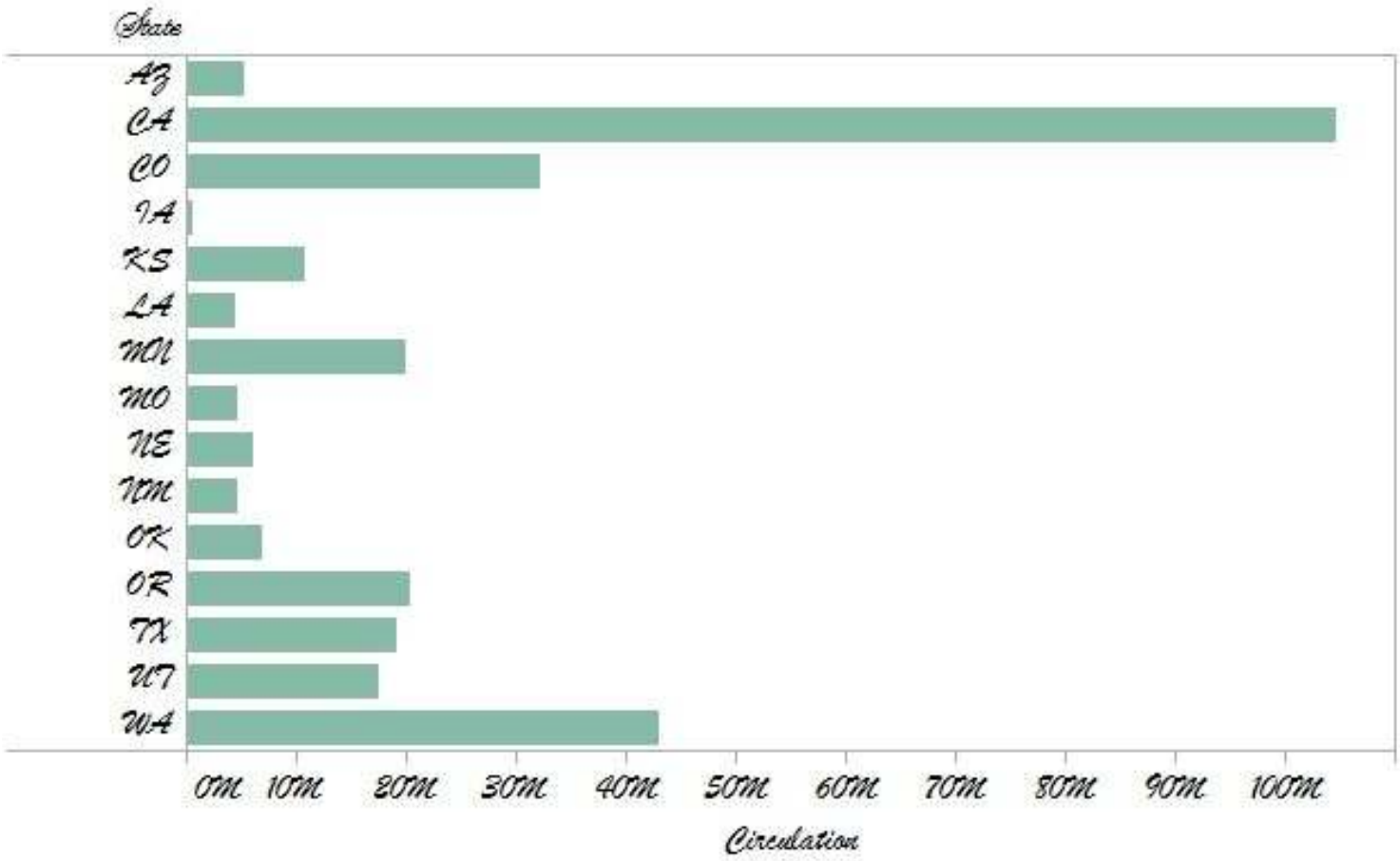
To test legibility, reduce on copy machine to a small size (say 50%) and see if text is still legible

Fonts and Typography

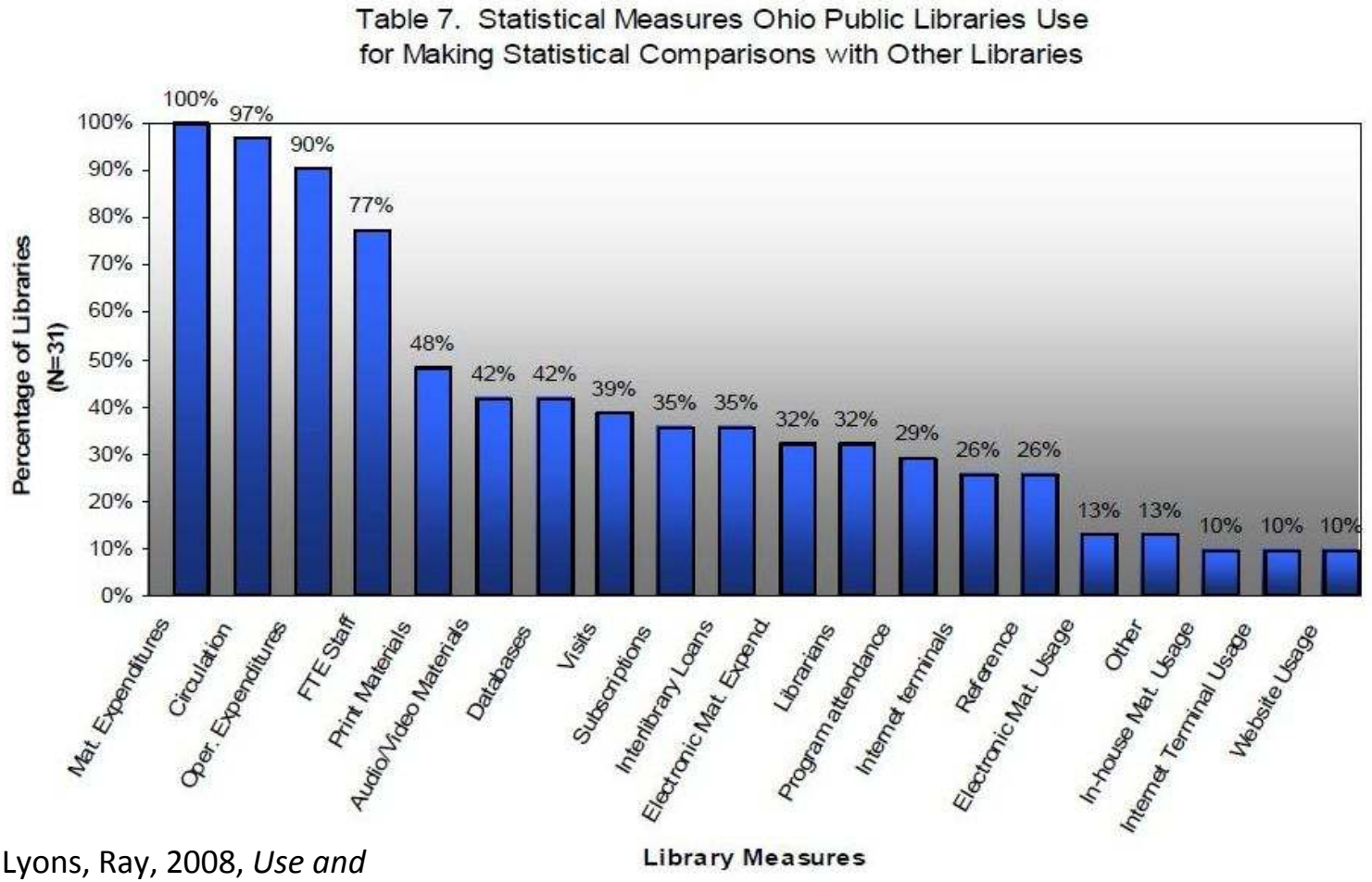
Leading (vertical distance between baseline of each row of text) should be 2 points larger than type size

Do not used condensed fonts

Keep the style simple. Use bold or italic for emphasis (but never together).

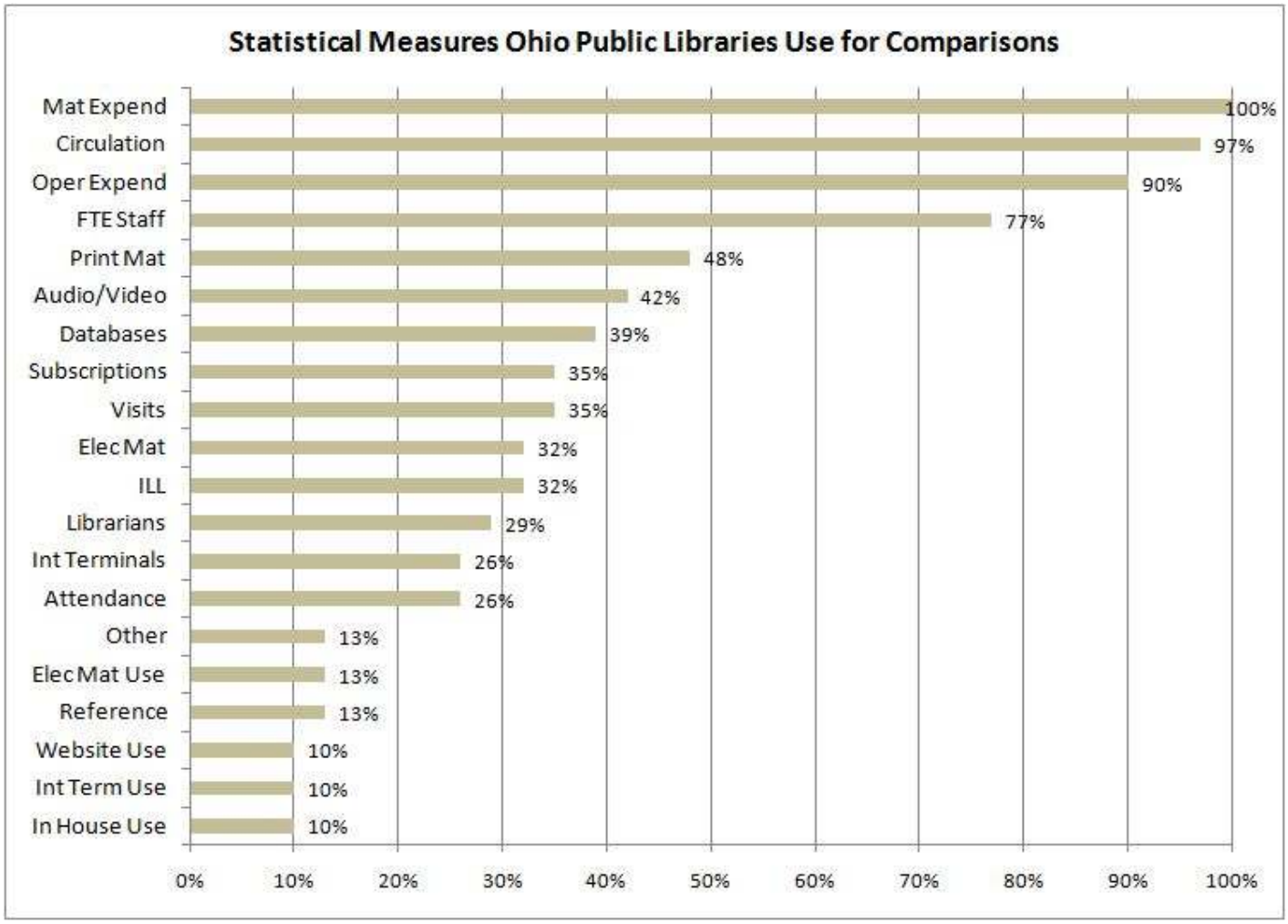


No stylized fonts

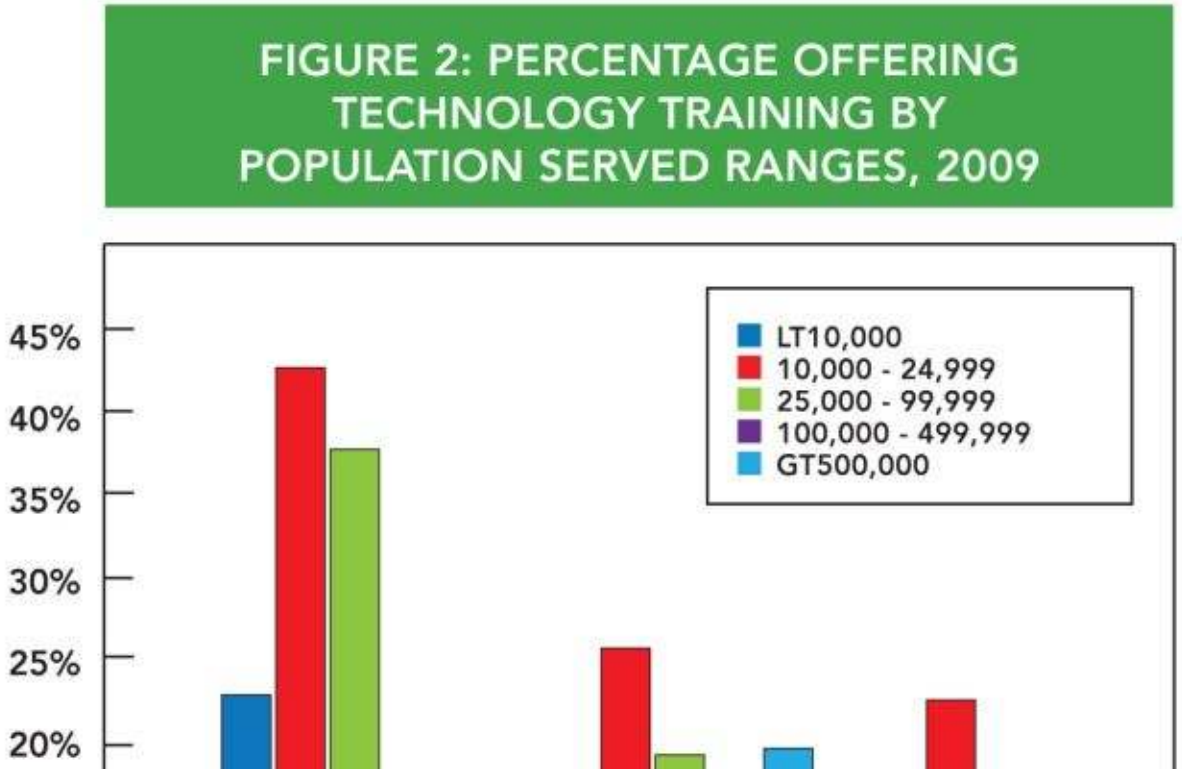


Source: Lyons, Ray, 2008, *Use and Perceptions of Comparative Statistics by Ohio Public Libraries*

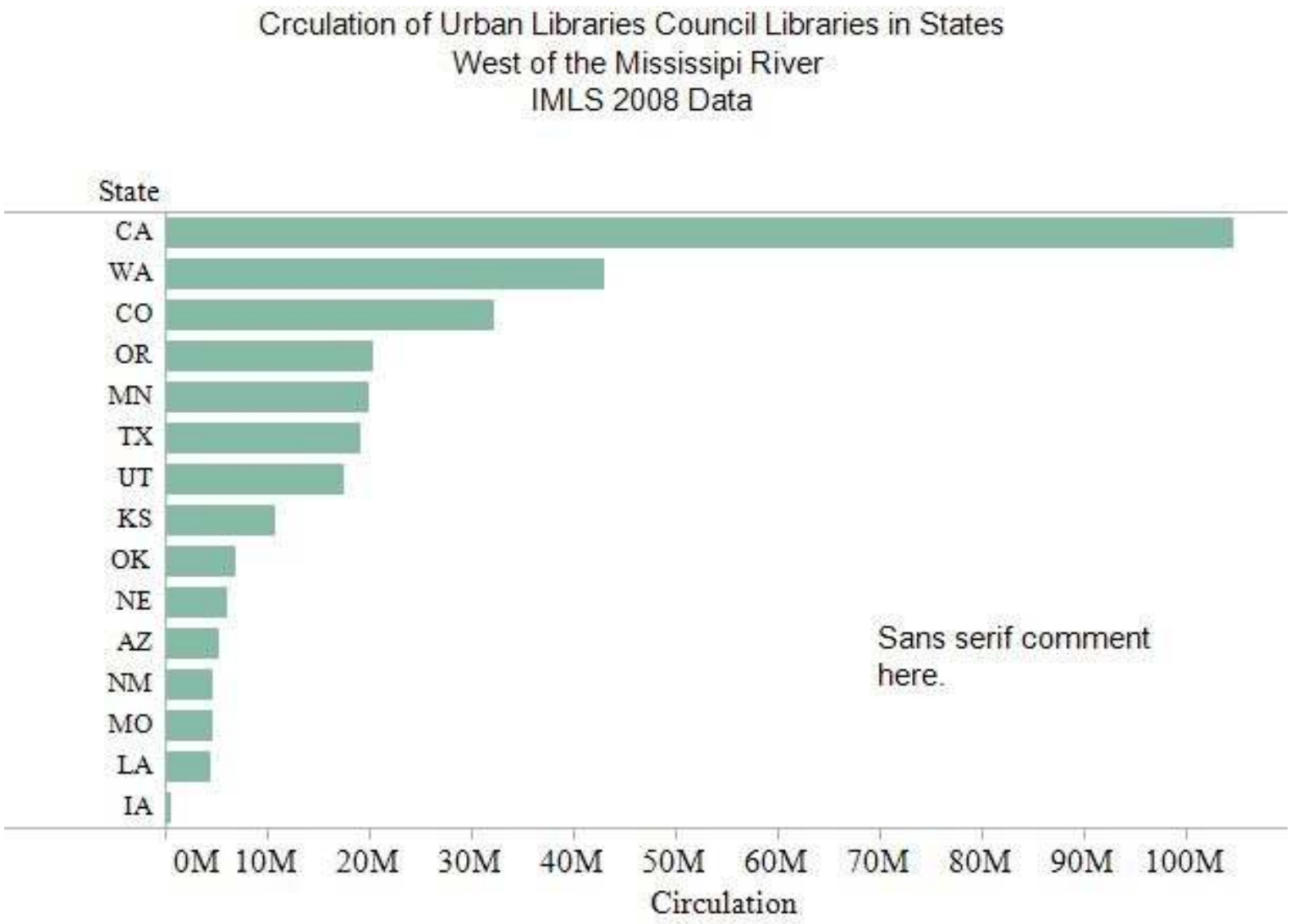
No angled text



Horizontal bar charts avoid angled text



Do not use inverse text (light text on dark background)



Serif and sans serif fonts can be used together

Name	Outlets	Staff	Computers
Library A	16	**101**	**23**
Library B	14	**178**	**41**
Library C	9	**82**	**29**
Library D	7	**93**	**21**
Library E	2	**22**	**12**
Library F	1	**98**	**16**

Don’t overdo use of emphasis

Words and Abbreviations

Spell out formal names (no IMLS, ALA, NCES)

Spell out months when feasible

On horizontal chart axis use :

Jan Feb March April May June

July Aug Sept Oct Nov Dec

Always spell out months in tables

States and Provinces

Avoid 2-character state or province abbreviations

Ala. *Ariz.* *Ark.* *Calif.* *Colo.* *Conn.* *Del.*

Fla. *Ga.* *Ill.* *Ind.* *Kan.* *Ky.* *La.*

Mass. *Md.* *Mich.* *Minn.* *Miss.* *Mo.* *Mont.*

N.C. *N.D.* *H.H.* *N.J.* *N.M.* *N.Y.* *Neb.*

Nev. *Oklah.* *Ore.* *Pa.* *R.I.* *S.C.* *S.D.*

Tenn. *Va.* *Vt.* *W.Va.* *Wash.* *Wis.* *Wyo.*

Dates

User 4-digit years when feasible

If not feasible, begin sequence with full year:

2001, 02, 03, 04, 05 … **or**

2001, ‘02, ‘03, ’04, ‘05 …

Indicate year with quarterly data:

Q1	Q2	Q3	Q4
2006			

I	II	III	IV
2006			

Dates

On horizontal chart axis use :

- Jan Feb March April May June
- July Aug Sept Oct Nov Dec

Always spell out months in tables

Graphic Icons

Use only when comparing a small series

Too-detailed symbols are distracting, hinder the message.

Use simple pictograms

Graphic Icons

To represent variables, shade a single symbol (don't use alternate symbols).

Dona Wong’s standards for good icons:

- simple
- symmetrical
- clear when reduced
- square-shaped

Graphic Icons

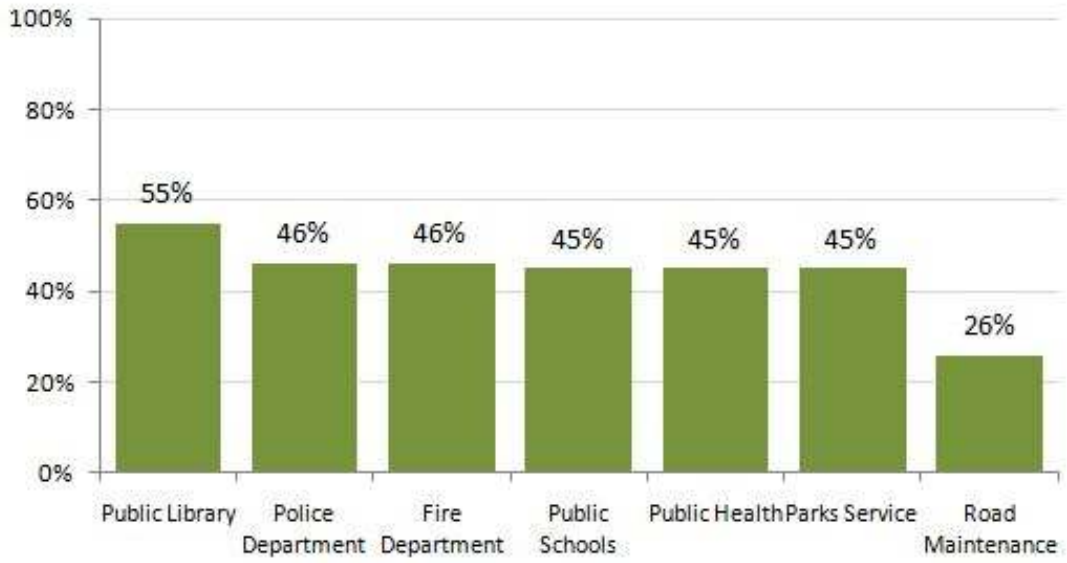
Don't us pictograms when counts are not that different; too hard to distinguish magnitude.

Don't use icons to represent relative size—shrunken for less, expanded for more.

Icons can represent multiple units; use multiples of 1, 2, 10, 50, 100, etc.

Edward Tufte suggested a *USA Today* style graphical icons to modernize John Snow's famous graphical study of cholera-related deaths in London 1854.

IN CONCLUSION…



SHOW THE DATA!