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.


Tufte, Edward R.


Expert Sources

Wainer, Howard


Expert Sources

Wong, Dona M.


Few, Stephen

<table>
<thead>
<tr>
<th>SERIES 1</th>
<th>SERIES 2</th>
<th>SERIES 3</th>
<th>SERIES 4</th>
</tr>
</thead>
<tbody>
<tr>
<td>X</td>
<td>Y</td>
<td>X</td>
<td>Y</td>
</tr>
<tr>
<td>10.0</td>
<td>8.04</td>
<td>10.0</td>
<td>9.14</td>
</tr>
<tr>
<td>8.0</td>
<td>6.95</td>
<td>8.0</td>
<td>8.14</td>
</tr>
<tr>
<td>13.0</td>
<td>7.58</td>
<td>13.0</td>
<td>8.74</td>
</tr>
<tr>
<td>9.0</td>
<td>8.81</td>
<td>9.0</td>
<td>8.77</td>
</tr>
<tr>
<td>11.0</td>
<td>8.33</td>
<td>11.0</td>
<td>9.26</td>
</tr>
<tr>
<td>14.0</td>
<td>9.96</td>
<td>14.0</td>
<td>8.10</td>
</tr>
<tr>
<td>6.0</td>
<td>7.24</td>
<td>6.0</td>
<td>6.13</td>
</tr>
<tr>
<td>4.0</td>
<td>4.26</td>
<td>4.0</td>
<td>3.10</td>
</tr>
<tr>
<td>12.0</td>
<td>10.84</td>
<td>12.0</td>
<td>9.13</td>
</tr>
<tr>
<td>7.0</td>
<td>4.82</td>
<td>7.0</td>
<td>7.26</td>
</tr>
<tr>
<td>5.0</td>
<td>5.68</td>
<td>5.0</td>
<td>4.74</td>
</tr>
<tr>
<td>SERIES 1</td>
<td>SERIES 2</td>
<td>SERIES 3</td>
<td>SERIES 4</td>
</tr>
<tr>
<td>----------</td>
<td>----------</td>
<td>----------</td>
<td>----------</td>
</tr>
<tr>
<td>X</td>
<td>Y</td>
<td>X</td>
<td>Y</td>
</tr>
<tr>
<td>10.0</td>
<td>8.04</td>
<td>10.0</td>
<td>9.14</td>
</tr>
<tr>
<td>8.0</td>
<td>6.95</td>
<td>8.0</td>
<td>8.14</td>
</tr>
<tr>
<td>13.0</td>
<td>7.58</td>
<td>13.0</td>
<td>8.74</td>
</tr>
<tr>
<td>9.0</td>
<td>8.81</td>
<td>9.0</td>
<td>8.77</td>
</tr>
<tr>
<td>11.0</td>
<td>8.33</td>
<td>11.0</td>
<td>9.26</td>
</tr>
<tr>
<td>14.0</td>
<td>9.96</td>
<td>14.0</td>
<td>8.10</td>
</tr>
<tr>
<td>6.0</td>
<td>7.24</td>
<td>6.0</td>
<td>6.13</td>
</tr>
<tr>
<td>4.0</td>
<td>4.26</td>
<td>4.0</td>
<td>3.10</td>
</tr>
<tr>
<td>12.0</td>
<td>10.84</td>
<td>12.0</td>
<td>9.13</td>
</tr>
<tr>
<td>7.0</td>
<td>4.82</td>
<td>7.0</td>
<td>7.26</td>
</tr>
<tr>
<td>5.0</td>
<td>5.68</td>
<td>5.0</td>
<td>4.74</td>
</tr>
</tbody>
</table>

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 coefficient = .82
R-square = .67
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.”

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

The purpose of graphical data presentation is to communicate information clearly and accurately.
Figure 1: Total State Library Agency Revenues by State, FY2008

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

“Strive for clarity.”

“Proofread graphs.”

“Visual clarity must be preserved under reduction and reproduction.”
“Make captions comprehensive and informative.”

“Draw attention to the important features of the data.”
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.”
“The number of information-carrying (variable) dimensions depicted should not exceed the number of dimensions in the data.”
The 1978 dollar should be twice as big as shown.

Source: Graphic reprinted in Edward Tufte, 1983.
An index of 100 is average
Source: From Awareness to Funding, OCLC, 2003

From Awareness to Funding: A study of library support in America
“Write out explanations of the data on the graphic itself.”

“Label important events in the data.”

“Graphics must not quote data out of context.”
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.
Rendering data graphically is a powerful investigative tool.
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.

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

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.

<table>
<thead>
<tr>
<th>Growth Rate</th>
<th>Generations Required to Double US Standard of Living</th>
</tr>
</thead>
<tbody>
<tr>
<td>1947 – 1973</td>
<td>1.6</td>
</tr>
<tr>
<td>1973 – 1990</td>
<td>12</td>
</tr>
</tbody>
</table>
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.
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.
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


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

BTU: Quantity of heat required to raise temperature of one pound of water one degree Fahrenheit.

Copyright 1973 Congressional Quarterly, Inc.
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?
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.
William Cleveland suggests four scale lines. His captions are precise and thorough.


**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.
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.
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:

<table>
<thead>
<tr>
<th>Use</th>
<th>Instead of:</th>
</tr>
</thead>
<tbody>
<tr>
<td>$3 \text{ million}$ or 3M</td>
<td>$3,000 \text{ thousand}$</td>
</tr>
<tr>
<td>2.6 billion or 2.6B</td>
<td>2,582,000,000</td>
</tr>
<tr>
<td>1K, 2K, 3K</td>
<td>1000, 2000, 3000</td>
</tr>
</tbody>
</table>

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
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
Avoid placing a legend inside the data rectangle.
With *data labels* the reader does not have to translate legend colors, patterns, or labels.

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
Joseph Priestley’s line chart (1765) depicted biographical history.
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.
If the data are intertwined, consider producing separate, smaller charts for each measure.

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.
Avoid using line charts to depict categorical data

Source: *American Libraries*, Nov. 2009
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
Oops. For line charts with shading below the line(s), the vertical scale should begin at zero, with zero clearly marked.

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

Pre-attentive Attributes

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

Spatial Position

Longer = greater quantity

Higher or further right = greater quantity

Adapted from Stephen Few, 2009.
Pre-attentive Attributes

Quantitative perception is less precise with ...

<table>
<thead>
<tr>
<th>Size</th>
<th>Intensity / Brightness</th>
<th>Thickness</th>
</tr>
</thead>
<tbody>
<tr>
<td>● ● ● ● ●</td>
<td>● ○ ● ● ●</td>
<td>●</td>
</tr>
<tr>
<td>● ● ● ●</td>
<td>● ● ● ● ●</td>
<td>●</td>
</tr>
<tr>
<td>● ● ● ●</td>
<td>● ● ● ● ●</td>
<td>●</td>
</tr>
<tr>
<td>● ● ● ●</td>
<td>● ● ● ● ●</td>
<td>●</td>
</tr>
</tbody>
</table>

Bigger = greater  Darker = greater  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!
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!
Stylish coloring does not make circus striping less distracting.

Source: Highwire Press, 2010
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

Color palette - muted

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?

- Merge duplicate records: 55%
- Add tables of contents to records: 40%
- Fix typos: 31%
- Upgrade brief records: 30%
- Make it easier to correct records: 29%
- Fix MARC coding errors: 29%
- Add summaries to records: 27%
- Add cover art of results: 26%
- Increase accuracy of library holding information: 26%
- More records for online resources: 26%

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?

- Merge duplicate records: 51%
- Add tables of contents to records: 47%
- Add summaries to results: 31%
- Add cover art to results: 30%
- Add summaries to records: 30%
- More clickable links to online content: 28%
- Add more formats: 26%
- Fix typos: 24%
- More records for online resources: 23%
- Increase accuracy of library holding information: 22%

Source: Online Catalogs: What Users and Librarians Want, OCLC, 2009 (Library survey)
Reference—Top Ten Enhancements
Which of the following enhancements would you recommend?

- Merge duplicate records: 52%
- Add tables of contents to records: 46%
- Add summaries to records: 29%
- Add summaries to results: 28%
- Increase accuracy of library holdings information: 26%
- Fix types: 26%
- Add cover art to results: 25%
- Add more formats: 23%
- Make it easier to correct records: 22%
- Fix MARC coding errors: 21%

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?

- Merge duplicate records: 55%
- Add tables of contents to records: 46%
- Add summaries to records: 31%
- Fix types: 29%
- Increase accuracy of library holdings information: 28%
- Add summaries to results: 27%
- Add cover art to results: 26%
- Make it easier to correct records: 25%
- Fix MARC coding errors: 23%
- More records for online resources: 22%

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?

- Merge duplicate records: 49%
- Add tables of contents to records: 42%
- Increase accuracy of library holdings information: 38%
- Add summaries to records: 27%
- Add summaries to results: 26%
- More clickable links to online content: 25%
- Add more formats: 23%
- Add cover art to results: 23%
- Fix types: 21%
- Make it easier to correct records: 20%

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


*Hispanics includes both English and Spanish-speaking Hispanics
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
Household Income by Selected Source: 1999
(Percent of households classified by age of householder. Data based on sample. For information on confidentiality protection, sampling error, nonsampling error, and definitions, see www.census.gov/prod/cen2000/doc/sf4.pdf)

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.
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

Measuring Losses | Peak-to-trough contraction in GDP during recessions

Note: Changes based on quarterly data from the Commerce Department. Recessions are defined by other indicators in addition to GDP.

Source: Commerce Department; National Bureau of Economic Research

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.
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.
This chart uses a ‘double Y-axis’ that exaggerates terminal installations in US public libraries.
Connected lines = cumulative percent change in 4 public library measures

Bars = number of public access terminals per outlet

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

Cumulative Percent Change in 4 Public Library Measures

- 1999
- 2000
- 2001
- 2002
- 2003
- 2004
- 2005
- 2006

Number of terminals
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
Circus striping and hash patterns distract the eye from the data.
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
Table 7. Statistical Measures Ohio Public Libraries Use for Making Statistical Comparisons with Other Libraries

<table>
<thead>
<tr>
<th>Library Measures</th>
<th>Percentage of Libraries (N=31)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Net Expenditures</td>
<td>100%</td>
</tr>
<tr>
<td>Circulation</td>
<td>97%</td>
</tr>
<tr>
<td>Operating Expenditures</td>
<td>90%</td>
</tr>
<tr>
<td>FTE Staff</td>
<td>77%</td>
</tr>
<tr>
<td>Print Materials</td>
<td>48%</td>
</tr>
<tr>
<td>Audio/Video Materials</td>
<td>42%</td>
</tr>
<tr>
<td>Database</td>
<td>42%</td>
</tr>
<tr>
<td>Visits</td>
<td>39%</td>
</tr>
<tr>
<td>Subscriptions</td>
<td>36%</td>
</tr>
<tr>
<td>Electronic/Net Expense</td>
<td>35%</td>
</tr>
<tr>
<td>Libraries</td>
<td>32%</td>
</tr>
<tr>
<td>Patrons/Attendance</td>
<td>29%</td>
</tr>
<tr>
<td>In-house terminals</td>
<td>26%</td>
</tr>
<tr>
<td>Reference</td>
<td>26%</td>
</tr>
<tr>
<td>Electronic/Net Usage</td>
<td>13%</td>
</tr>
<tr>
<td>In-house Net Usage</td>
<td>13%</td>
</tr>
<tr>
<td>Internet Terminal Usage</td>
<td>10%</td>
</tr>
<tr>
<td>Website Usage</td>
<td>10%</td>
</tr>
</tbody>
</table>

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

Source: Knight Commission on Intercollegiate Athletics, 2009, *Restoring the Balance*. 
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.
In 1786 William Playfair realized the importance of sorting the order of the data (small to large).

Source: Howard Wainer, 2005.
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:

\[
\text{area of a circle} = \pi r^2
\]
Pie centers are deceiving and distracting

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

- Outlier
- Whisker
- Upper quartile (75th percentile)
- Median (50th percentile)
- Lower quartile (25th percentile)
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 \( \frac{1}{2} \) of a circle’s height (the radius) is proportional to the square root of its area.
Homework Problem

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

Source: Samantha Becker et al., 2010.
Re-Expressing Data

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

Re-Expressing Data

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.
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.
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.
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.
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 use condensed fonts.

Keep the style simple. Use bold or italic for emphasis (but never together).
No stylized fonts
Table 7. Statistical Measures Ohio Public Libraries Use for Making Statistical Comparisons with Other Libraries

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
<table>
<thead>
<tr>
<th>Name</th>
<th>Outlets</th>
<th>Staff</th>
<th>Computers</th>
</tr>
</thead>
<tbody>
<tr>
<td>Library A</td>
<td>16</td>
<td>101</td>
<td>23</td>
</tr>
<tr>
<td>Library B</td>
<td>14</td>
<td>178</td>
<td>41</td>
</tr>
<tr>
<td>Library C</td>
<td>9</td>
<td>82</td>
<td>29</td>
</tr>
<tr>
<td>Library D</td>
<td>7</td>
<td>93</td>
<td>21</td>
</tr>
<tr>
<td>Library E</td>
<td>2</td>
<td>22</td>
<td>12</td>
</tr>
<tr>
<td>Library F</td>
<td>1</td>
<td>98</td>
<td>16</td>
</tr>
</tbody>
</table>

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

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:

<table>
<thead>
<tr>
<th>Q1</th>
<th>Q2</th>
<th>Q3</th>
<th>Q4</th>
</tr>
</thead>
<tbody>
<tr>
<td>2006</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>I</th>
<th>II</th>
<th>III</th>
<th>IV</th>
</tr>
</thead>
<tbody>
<tr>
<td>2006</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
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 use 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!