Seminal Books by Edward Tufte

Standard literature for every visualization enthusiast

Well recognized for his writings on information design

- a pioneer in the field of data visualization
- taught information design at Princeton University
- now a professor at Yale University

Popularized concept of “small multiples”

- aka trellis chart or panel chart
- similar charts of same scale + axes
- allows them to be easily compared
- use multiple views to show different partitions of a dataset
E. Muybridge’s Horses in Motion (1886)

- proofed for the first time that horses CAN have all 4 legs in the air
- work was also foundational to the development of the motion picture
FA Walker’s census charts (1870)

- population is broken down by state and then occupation, including a count of those attending school
- also has tree maps!
Also popularized “sparklines”

- small integrative visualizations

Sparklines inspired “word size visualizations”

- charts or graphs tightly integrated into text or even computer code

Although Tufte is said to have invented sparklines, in actuality he invented only the name and popularized it as technique. Sparklines are a condensed way to present trends and variation, associated with a measurement such as average temperature or stock market activity, often embedded directly in the text; for example: The Dow Jones index for February 7, 2006. These are often used as elements of a small multiple with several lines used together. Tufte explains the sparkline as a kind of “word” that conveys rich information without breaking the flow of a sentence or paragraph made of other “words” both visual and conventional. To date, the earliest known implementation of sparklines was done by interaction designer Peter Zelchenko and programmer Mike Medved in early 1998.
According to Tufte (pg. 51):

- Graphical excellence is the well-designed presentation of interesting data
  - a matter of *substance, statistics, and design*

- Graphical excellence consists of complex ideas communicated with:
  - *clarity, precision, and efficiency*

- Graphical excellence is that what gives the viewer:
  - the *greatest number of ideas*
  - in the *shortest time*
  - with the *least ink*
  - in the *smallest space*

- Graphical excellence is nearly always multivariate

- Graphical excellence requires telling the truth about the data

(Nevertheless, visualizations should be visually pleasing and may very well have an artistic touch)
The Need for Visualization: Anscombe Quartet

<table>
<thead>
<tr>
<th>I</th>
<th>II</th>
<th>III</th>
<th>IV</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>
</tbody>
</table>

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

equation of regression line: \( Y = 3 + 0.5X \)
standard error of estimate of slope = 0.118
t = 4.24
sum of squares X - \( \bar{X} \) = 110.0
regression sum of squares = 27.50
residual sum of squares of Y = 15.75

correlation coefficient = .92
\( r^2 = .82 \)
### The Need for Visualization: Anscombe Quartet

<table>
<thead>
<tr>
<th></th>
<th>I</th>
<th></th>
<th>II</th>
<th></th>
<th>III</th>
<th></th>
<th>IV</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>X</td>
<td>Y</td>
<td></td>
<td>X</td>
<td>Y</td>
<td>X</td>
<td>Y</td>
<td>X</td>
<td>Y</td>
</tr>
<tr>
<td>10.0</td>
<td>8.04</td>
<td></td>
<td>10.0</td>
<td>9.14</td>
<td></td>
<td>10.0</td>
<td>7.46</td>
<td></td>
</tr>
<tr>
<td>8.0</td>
<td>6.95</td>
<td></td>
<td>8.0</td>
<td>8.14</td>
<td></td>
<td>8.0</td>
<td>6.77</td>
<td></td>
</tr>
<tr>
<td>13.0</td>
<td>7.58</td>
<td></td>
<td>13.0</td>
<td>8.74</td>
<td></td>
<td>13.0</td>
<td>12.74</td>
<td></td>
</tr>
<tr>
<td>9.0</td>
<td>8.81</td>
<td></td>
<td>9.0</td>
<td>8.77</td>
<td></td>
<td>9.0</td>
<td>7.11</td>
<td></td>
</tr>
<tr>
<td>11.0</td>
<td>8.33</td>
<td></td>
<td>11.0</td>
<td>9.26</td>
<td></td>
<td>11.0</td>
<td>7.81</td>
<td></td>
</tr>
<tr>
<td>14.0</td>
<td>9.96</td>
<td></td>
<td>14.0</td>
<td>8.10</td>
<td></td>
<td>14.0</td>
<td>8.84</td>
<td></td>
</tr>
<tr>
<td>6.0</td>
<td>7.24</td>
<td></td>
<td>6.0</td>
<td>6.13</td>
<td></td>
<td>6.0</td>
<td>6.08</td>
<td></td>
</tr>
<tr>
<td>4.0</td>
<td>4.26</td>
<td></td>
<td>4.0</td>
<td>3.10</td>
<td></td>
<td>4.0</td>
<td>5.39</td>
<td></td>
</tr>
<tr>
<td>12.0</td>
<td>10.84</td>
<td></td>
<td>12.0</td>
<td>9.13</td>
<td></td>
<td>12.0</td>
<td>8.15</td>
<td></td>
</tr>
<tr>
<td>7.0</td>
<td>4.82</td>
<td></td>
<td>7.0</td>
<td>7.26</td>
<td></td>
<td>7.0</td>
<td>6.42</td>
<td></td>
</tr>
<tr>
<td>5.0</td>
<td>5.68</td>
<td></td>
<td>5.0</td>
<td>4.74</td>
<td></td>
<td>5.0</td>
<td>3.73</td>
<td></td>
</tr>
</tbody>
</table>

- **N = 11**
- **Mean of X's = 9.0**
- **Mean of Y's = 7.5**
- **Equation of regression line:** \( Y = 3 + 0.5X \)
- **Standard error of estimate of slope:** 0.118
- **t = 4.24**
- **Sum of squares \( X - \bar{X} \) = 110.0**
- **Regression sum of squares = 27.50**
- **Residual sum of squares of Y = 13.75**
- **Correlation coefficient = .82**
- **r^2 = .67**
Age-Adjusted Cancer Rates (by County)

21,000 numbers
3056 counties
7 numbers per county:
- size (4)
- location (2)
- cancer rate (1)

1950-1969
Galaxy Maps

divide sky into
1,024 x 2,222 rectangles

tone = number of galaxies per rectangle
7,000 objects > 10 cm doubles every 5 years
Train Schedule: Paris – Lyon, 1880s

Minard: Visualization of Napoleon’s Russia Campaign (1812)

plots 6 variables: army size, 2D location, direction vector, temperature, time
Rage Fear Graph: Expressive Glyphs
Chernoff Faces: Multi-Variable Display
Chernoff Faces

Eye spacing
Head eccentricity
Eyebrow slope
Eye eccentricity
Pupil size
Nose width
Mouth openness
Mouth curvature
Mouth width
Eye size
Nose length
Chernoff Faces
Graphical Display: History

- Can be more precise and revealing than numerical display
  - example: Anscombe’s quartet (pg. 13/14)
  - example: cholera map of central London, 1854, by Dr. John Snow (pg. 24)

- Can capture a large amount of information in a very small space (billions of bits on one page)
  - example: data maps for cancer incidence (pg. 17)
  - example: galaxy maps (pg. 27)
  - example: space debris (pg. 48, Tufte “Envisioning Information”)

- Can extend to time-series display
  - example: train schedule Paris-Lyon, 1880s (pg. 31)

- Can be narrative
  - example: Napoleon’s Russia campaign, 1812, plots 6 variables on a 2D graph (pg. 41)

- Can represent each datapoint by visual information (graphic, icon, image, color, pattern)
  - examples: fear-rage graph (pg. 50), Chernoff faces (pg. 97, 142)
Tufte’s views on

- visual embellishments → “chart junk”
- abuse of physically-motivated distortions → “lie factor”
Avoid Misleading Embellishments = Chart Junk
Nigel Holmes’ Famous Chart

MONSTROUS COSTS
Total House and Senate campaign expenditures, in millions

1972 '74 '76 '80 '82 est.
Avoid Misleading Scaling
Manipulation of Axis Orientation

from Panday at al. (CHI 2015)
Avoid Misleading Scaling

Commission Payments To Travel Agents
In millions of dollars

- Delta: $36, $50, $70, $79, $84, $102, $109
- Eastern: $57, $70, $79, $84, $102, $109, $109
- TWA: $50, $52, $84, $102, $109, $109, $109
- United Airlines: $50, $52, $84, $102, $109, $109, $109

Years: '77, '78, First Half '78, Second Half '78
Avoid Misleading Use of Graphics Effects

This line, representing 18 miles per gallon in 1978, is 0.6 inches long.

This line, representing 27.5 miles per gallon in 1985, is 5.3 inches long.

real effect: \( \frac{(27.5-18)}{18} = 53\% \)

graphical effect: \( \frac{(5.3''-0.6'')}{18} = 783\% \)  \( \Rightarrow \) lie factor: \( \frac{783}{53} = 14.8 \)

Fuel Economy Standards for Autos
Set by Congress and supplemented by the Transportation Department. In miles per gallon.
Tell the Truth About the Data

REQUIRED FUEL ECONOMY STANDARDS:
NEW CARS BUILT FROM 1978 TO 1985

19.1 mpg, expected average for all cars on road, 1985

13.7 mpg, average for all cars on road, 1978
If You Must Embellish…
IN THE BARREL...
Price per bbl. of light crude, leaving Saudi Arabia on Jan. 1

April 1
$14.55

$13.34
$12.70
$12.09
$11.51
$10.46
$9.95
$2.41

1979
1978
1977
1976
1975
1974
1973

Avoid Suggestive Distortions
Show the Data in Their Proper Context

[Graphs showing Connecticut Traffic Deaths, Before (1935) and After (1936) with stricter enforcement by the Police Against Cars Exceeding Speed Limit.]

A few more data points add immensely to the account:

[Graphs showing Connecticut Traffic Deaths, 1931-1939 with fluctuations over time.]

[Traffic Deaths per 100,000 Persons in Connecticut, Massachusetts, Rhode Island, and New York, 1931-1939 with data trends.]
Avoid Display of Out-of-Context Data

SOLAR RADIATION AND STOCK PRICES

Is cosmetic decoration really needed to make data more interesting (may only distract):
- example: diamond graph (adds a useless 3rd dimension)

Misleading graphical representation
- example: missing baseline in Day Mines, Inc. annual report (pg. 54)
- example: non-uniform data spans in Commission Payments graph (pg. 54)
- example: non-uniform scaling of icons in Pittsburgh Civic Commission report (pg. 55)

The Lie Factor = \frac{\text{size of effect shown in graphic}}{\text{size of effect in data}} \quad (\text{should be within } [0.95, 1.05])
- example: graph on fuel economy standards for autos (lie factor = 14.8) (pg. 57)

Visualizing data bearing some dimension by means of objects of higher dimensions:
- example: the *growing barrel* (lie factor: 9.4 (2D), 59.4 (3D)) (pg. 62)
- example: the *growing oil pump* (lie factor: 9.5) (pg. 62)
- example: the *shrinking dollar bill* (lie factor: \sim 6) (pg. 70)
- example: the *incredibly shrinking family doctor* (pg. 69)

→ the number of information carrying dimensions should not exceed the data dimensions
Graphical Integrity

- Quoting data out of context and/or too sparse (recall: graphics allows high data density)
  - example: Connecticut traffic deaths (pg. 74/75)

Principles that ensure graphical integrity:

- The representation of numbers should be directly proportional to the numerical quantities represented (see the growing barrels)
- Clear and detailed labeling should be used to defeat graphical distortion and ambiguity
- Show data variations and not design variations (see the fuel economy graph)
- In time-series displays of money, show deflated and standardized units
- The number of information carrying dimensions should not exceed the data dimensions (see the growing barrels, the shrinking doctor)
- Graphics must not quote data out of context (see the Connecticut traffic deaths)
- Convincing graphics must demonstrate cause and effect (see Challenger disaster)
Do these bare graphs engage a human audience?
  • are they memorable?

A recent (research) trend
  • will embellishment help memorability, engagement?
  • do we need what Tufte char junk
MemViz: A Tool for Creating Memorable Visualizations

Darius Coelho, Sungsoo Ha, Shenghui Cheng, Salman Mahmoud, Jisung Kim, and Klaus Mueller

Visual Analytics and Imaging Lab, Computer Science Department, Stony Brook University and SUNY Korea
Memorability Experiment by Borkin et al.

Experiment set up as a game on Amazon Mechanical Turk

- workers were presented with a sequence of images (about 120)
- presented for 1 second, with a 1.4 second gap between consecutive images
- workers had to press a key if they saw an image for the second time in the sequence (spacing 1-7 images with “filler” images in between)
Memorability Experiment by Borkin et al.

most memorable
most memorable
least memorable
after removing human recognizable cartoons

Borkin et al. IEEE TVCG 2014
What Do People Remember?

**Experiment Design**

**Labeled Visualization Database**

- 393 visualizations

Visualizations are taken from [8], and the label taxonomy described in Table 1 is applied.

**Encoding**

- 100 “target” visualizations
- 10 seconds / image
- Output: Eye-tracking fixation locations and durations.

**Recognition**

- Same 100 targets + 100 “fillers”
- 2 seconds / image
- Output: Eye-tracking fixation locations and durations, and whether visualization is recognized.

**Recall**

- Correctly recognized blurred targets
- 20 min - as many images as participant can complete
- Output: Text descriptions of what participant recalls about the visualization.

Borkin et al. IEEE TVCG 2016
Fig. 7. Examples of the most and least recognizable visualizations from [8]. **TOP:** Eye-tracking fixation heat maps (i.e., average of all participants' fixation locations) from the encoding phase of the experiment in which each visualization was presented for 10 seconds. The fixation patterns demonstrate visual exploration of the visualization. **BOTTOM:** Eye-tracking fixation heat maps from the recognition phase of the experiment in which each visualization was presented for 2 seconds or until response. The most recognizable visualizations all have a single focus in the center indicating quick recognition of the visualization, whereas the least recognizable visualizations have fixation patterns similar to the encoding fixations indicative of visual exploration (e.g., title, text, etc.) for recognition.
Practical Rules for Visualization Design

Takeaways:

- 393 visualizations and eye movements of 33 participants and 1,000s of participant-generated text descriptions of the visualizations
- titles and supporting text should convey the message of a visualization
- if used appropriately, pictograms do not interfere with understanding and can improve recognition
- redundancy helps effectively communicate the message
- visualizations that are memorable “at-a-glance” are also capable of effectively conveying the message of the visualization

→ thus, a memorable visualization is often also an effective one
Important for Memorability

Important are:
- attributes like color
- inclusion of a human recognizable object

However, link to human engagement not explicitly established
- “just” memorability

Our own studies show that embellishments can get humans interested in studying an image
- but prefer conventional charts for problem solving
Infographic

Graphic visual representations of information, data or knowledge intended to present information quickly and clearly

Evolved in recent years to be for mass communication

- designed with fewer assumptions about the readers knowledge base than other types of visualizations
- but can be misleading and express the opinion of the author

![Infographic Image]
Using Icons as Bar Graphs

Wang et al. CHI 18
Data-Driven Design Guides

Kim et al. TVCG 17