CSE 5544: Introduction to Data Visualization

Raghu Machiraju
machiraju.1@osu.edu
Many Thanks to

– Alex Lex, University of Utah
– Torsten Moeller, University of Vienna
– Tamara Munzner, Univ of British Columbia
– Hanspeter Pfister, Harvard University
Design Principles
Previously
Visualization

To convey information through visual representations
Design Excellence

“Well-designed presentations of interesting data are a matter of substance, of statistics, and of design.”

E. Tufte
Edward Tufte
every time you make a powerpoint

edward tufte kills a kitten
And ...
Stephen Few
Robin Williams
Garr Reynolds
Design Critique & Redesign
Most important issues

What do you think is the most important problem facing New Zealand today?

- Unemployment/Jobs
- Economy

2001: 15% Unemployment, 12% Economy
2002: 10% Unemployment, 8% Economy
2003: 9% Unemployment, 9% Economy
2004: 8% Unemployment, 5% Economy
2005: 11% Unemployment, 4% Economy
2006: 13% Unemployment, 4% Economy
2007: 13% Unemployment, 3% Economy
2008: 22% Unemployment, 4% Economy

2009: 33% Unemployment, 14% Economy
2010: 25% Unemployment, 13% Economy
2011: 33% Unemployment, 14% Economy

Crime/Violence
- Race relations

2001: 7% Crime, 6% Race relations
2002: 14% Crime, 6% Race relations
2003: 13% Crime, 9% Race relations
2004: 28% Crime, 6% Race relations
2005: 10% Crime, 9% Race relations
2006: 11% Crime, 5% Race relations
2007: 13% Crime, 3% Race relations
2008: 15% Crime, 1% Race relations
2009: 11% Crime, 3% Race relations
2010: 9% Crime, 4% Race relations
2011: 4% Crime, 4% Race relations

Sunday Star Times, 2012

https://goo.gl/lHWp4x
Quantity encoded by diameter, not area!
Fixing that:
But is this **visual encoding** appropriate?
The Premise
What can you distort?
When you use geometry, photometry do not distort
Distort What?

Data

– representations
– clusters, trends
– correlations
Design Excellence

“Well-designed presentations of interesting data are a matter of substance, of statistics, and of design.” - Tufte
Design Guidelines
Mantra

– Graphical Integrity
– Design Principles
– Design Elements
Graphical Integrity
If Bush tax cuts expire...

<table>
<thead>
<tr>
<th>Top tax rate</th>
<th>Now</th>
<th>Jan. 1, 2013</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>35.0%</td>
<td>39.6%</td>
</tr>
</tbody>
</table>
Dealing with Scale Distortions
Missing Scales

- Where's baseline?
- What's scale?
- What's context?

Tufte, VDQI
Incomplete Scales

How 2012 Stacks Up
The Warmest Years on Record
Contiguous U.S.

1921: 53.81
1999: 53.93
1934: 54.14
2006: 54.31
1998: 54.32
2012: 55.3

Source: NOAA's National Climatic Data Center - State of the Climate National Overview

CLIMATECENTRAL
Consider This ...
Look at this closely …
Start y-scale at 0?
Another one …

The Daily Mail, UK, Jan 2012
Putting it in CONTEXT

Temperature Anomaly -- Annual Mean (°C)

Mother Jones
Frame the Data

Temperature Anomaly -- Annual Mean (°C)

Mother Jones
A huge problem
“the latest data shows total consumer credit collapsing at an accelerating rate” (businessinsider.com)

“Per capita is better than total” (chartingtheeconomy.com)

“have to compare credit to something like disposable income.” (Rolfe Winkler, reuters blog)

“on-going credit contraction— it’s astonishing” (ritholtz.com – The Big Picture)
Other distortions
The Lie Factor

- \( \frac{\text{Size of effect in graphic}}{\text{size of effect in data}} \)
The Lie Factor

\[
\frac{5.3 - 0.6}{0.6} / \frac{27.5 - 18}{18} = 14.8
\]
The Lie Factor

From Modest to McMansion
The average square footage of a new single-family home

<table>
<thead>
<tr>
<th>Year</th>
<th>Square Footage</th>
</tr>
</thead>
<tbody>
<tr>
<td>1950</td>
<td>983 sq. ft.</td>
</tr>
<tr>
<td>1970</td>
<td>1,500 sq. ft.</td>
</tr>
<tr>
<td>1990</td>
<td>2,080 sq. ft.</td>
</tr>
<tr>
<td>2004</td>
<td>2,349 sq. ft.</td>
</tr>
</tbody>
</table>

Source: National Association of Home Builders (Housing Facts, Figures and Trends for March 2006)

http://www.theoildrum.com/node/4645
The Lie Factor

THE SHRINKING FAMILY DOCTOR
In California
Percentage of Doctors Devoted Solely to Family Practice
1964 1975 1990
27% 16.0% 12.0%
1: 3,167 6,212
1: 4,232 6,694
1: 2,247 RATIO TO POPULATION
8,023 Doctors

IN THE BARREL...
Price per bbl. of light crude, leaving Saudi Arabia on Jan. 1
April 1
$14.55
$13.24
$12.09
$11.51
$10.46
$9.96
$2.41
$1.2

Tufte, VDQI
Design Distortions

OPEC Oil Prices: After 18 Months of Stability, Prices Are Due to Rise Again

Dollars per barrel

- Jan. 1.5% increase
- April 1.5% increase
- July 1.5% increase
- Oct. 1.5% increase

$54.54

1979

Quarterly

Yearly

Inherently pathological
Death to Pie Charts

“I hate pie charts. I mean, really hate them.”

Share of coverage on TechCrunch


Cole Nussbaumer
Redesign

TechCrunch Coverage: 2005 - 2011
A slightly better pie?

Bar are best!

- General Consumer Web: 23%
- Social Networks: 12%
- Search: 10%
- Mobile: 9%
- Software: 8%
- Entertainment: 6%
- Hardware: 6%
- E-Commerce: 5%
- Advertising: 3%
- Video: 3%
- No Category: 3%
- Enterprise: 2%
- Other: 2%
- News: 2%
- Music: 1%
- Network/Hosting: 1%
- Investor: 1%
- PR: 1%
- Cleantech: 1%
The differences?
The differences?
Tufte’s Integrity

- Show data variation, not design variation
- Clear, detailed, and thorough labeling and appropriate scales
- Size of the graphic effect should be directly proportional to the numerical quantities ("lie factor")
The problem …
Visual Encoding

TechCrunch Coverage: 2005 - 2011
Bars are best!

General Consumer Web 23%
Social Networks 12%
Search 10%
Mobile 9%
Software 8%
Entertainment 6%
Hardware 6%
E-Commerce 5%
Advertising 3%
Video 3%
No Category 3%
Enterprise 2%
Other 2%
News 2%
Music 1%
Network/Hosting 1%
Investor 1%
PR 1%
CleanTech 1%

TechCrunch Coverage: 2005 - 2011
A slightly better pie?
Not all Encodings are same

From Wilkinson 99, based on Stevens 61
Design Principles
Use Decomposition

Beer sales
Hierarchical Display
Show Context

100 combat deaths per month

Iraq
Maximize Data-Ink Ratio
Maximize Data-Ink Ratio

• Data-ink = the ink used to show data
• Data-ink ratio = data-ink / total ink used
Avoid Chartjunk

Extraneous visual elements that distract from the message

ongoing, Tim Brey
Avoid Chartjunk

Extraneous visual elements that distract from the message

ongoing, Tim Brey
Avoid Chartjunk

Extraneous visual elements that distract from the message

ongoing, Tim Brey
Avoid Chartjunk

Extraneous visual elements that distract from the message

ongoing, Tim Brey
Avoid Chartjunk

Extraneous visual elements that distract from the message

ongoing, Tim Brey
Avoid Chartjunk

Extraneous visual elements that distract from the message
Which is better?

[Bateman et al. 2010]
Which is better?

[Bateman et al. 2010]

https://eagereyes.org/criticism/chart-junk-considered-useful-after-all
Useful Junk? The Effects of Visual Embellishment on Comprehension and Memorability of Charts

Scott Bateman, Regan L. Mandryk, Carl Gutwin,
Aaron Genest, David McDine, Christopher Brooks
Department of Computer Science, University of Saskatchewan, Saskatoon, Saskatchewan, Canada
scott.bateman@usask.ca, regan@cs.usask.ca, gutwin@cs.usask.ca,
aaron.genest@usask.ca, dam085@mail.usask.ca, cab938@mail.usask.ca

ABSTRACT
Guidelines for designing information charts often state that the presentation should reduce 'chart junk'—visual embellishments that are not essential to understanding the data. In contrast, some popular chart designers wrap the presented data in detailed and elaborate imagery, raising the questions of whether this imagery is really as detrimental to understanding as has been proposed, and whether the visual embellishment may have other benefits. To investigate these issues, we conducted an experiment that compared embellished charts with plain ones, and measured both interpretation accuracy and long-term recall. We found that people's accuracy in describing the embellished charts was no worse than for plain charts, and that their recall after a two-to-three-week gap was significantly better. Although we are cautious about recommending that all charts be produced in this style, our results question some of the premises of the minimalist approach to chart design.

Author Keywords
Charts, information visualization, imagery, memorability.

ACM Classification Keywords
H.5.m. Information interfaces and presentation (e.g., HCI): Miscellaneous.

General Terms
Design, Human Factors

INTRODUCTION
Many experts in the area of chart design, such as Edward Tufte, criticize the inclusion of visual embellishment in charts and graphs; their guidelines for good chart design often suggest that the addition of chart junk, decorations and other kinds of non-essential imagery, to a chart can make interpretation more difficult and can distract readers from the data [22]. This minimalist perspective advocates data-ink—or the ink in the chart used to represent data.

Despite these minimalist guidelines, many designers include a wide variety of visual embellishments in their charts, from small decorations to large images and visual backgrounds. One well-known proponent of visual embellishment in charts is the graphic artist Nigel Holmes, whose work regularly incorporates strong visual imagery into the fabric of the chart [7] (e.g., Figure 1).
EXPERIMENTAL RESULTS

1. No significant difference between plain and image charts for interactive interpretation accuracy
2. No significant difference in recall accuracy after a five-minute gap
3. Significantly better recall for plain charts of both the chart topic and the details (categories and trend) after long-term gap (2-3 weeks).
4. Participants saw value messages in plain charts significantly more often than in the plain charts.
5. Participants found plain charts more attractive, most enjoyed them, and found that they were easiest and fastest to remember.
Before

Number of bikes sold (2002-2007)

After

Over 5,000 bikes sold in 2007

Arable land in organic production

Source: SOEL-FIBL Survey '07

G. Reynolds, Presentation Zen
Before

Number of new members per year

2004  2005  2006  2007

After

Nearly 100 new members in '07

2004  2005  2006  2007

Before

Smoking rates for 2002

China  Japan  Germany  U.K.  USA

After

Smoking rates for 2002

China  Japan  Germany  U.K.  USA

G. Reynolds, Presentation Zen
Bring in the Clowns…

World Population in 2008
A better version...

World Population in 2008
Use Chart Junk? It depends!

<table>
<thead>
<tr>
<th>PROS</th>
<th>CONS</th>
</tr>
</thead>
<tbody>
<tr>
<td>persuasion</td>
<td>unbiased analysis</td>
</tr>
<tr>
<td>memorability</td>
<td>trustworthiness</td>
</tr>
<tr>
<td>engagement</td>
<td>interpretability</td>
</tr>
<tr>
<td></td>
<td>space efficiency</td>
</tr>
</tbody>
</table>
More
Another theme of Tufte’s is that good graphics have high “data density”. To calculate this quantity, note that the numbers to be plotted in a graphic can be written as an array of one or more dimensions.

**Definition 3 (Data Density)**

\[
data \text{ density} = \frac{\text{number of entries in data array}}{\text{area of data graphic}}\]

Notwithstanding this formal definition, the concept of a “high data density” graphic is perhaps best illustrated through some examples. Fig. 2.12 compares a large number of simultaneous measurements of thermal conductivity as a function of temperature. Because more than 200 different curves, each labelled, are combined into a single graph, Tufte has canonized this as a Good Example. Is it? Well, yes, but only up to a point. First, there is no really good way to present 200 data sets. A single figure is certainly much easier to grasp than a whole of lot of figures.

Escaping Flatland
HARD ROCK
- Kingsmen
- Paul Revere and the Raiders
- Association
- Johnny Rivers
- ? & the Mysterians
- Standells
- Troggs
- Cyrkle
- Hollies
- Vogues
- Left Bank
- Neil Diamond
- Dusty Springfield
- Nancy Sinatra

ACID ROCK
- Van Morrison
- Procol Harum
- Turtles
- Doors
- Vanilla Fudge
- Jimi Hendrix
- Jefferson Airplane
- Box Tops
- Bee Gees
- Happenings
- Buckinghams
- Classics IV

POP ROCK
- Cream
- Steppenwolf
- Blue Cheer
- Guess
- Creedence
- Chicago
- Blood Sweat
- Big Brother & the Holding Company
- Three Dog Night
- Spiral
- Peter Paul

Reebee Garofalo, Genealogy of Pop/Rock Music
### Sparklines

<table>
<thead>
<tr>
<th></th>
<th>1999.1.1</th>
<th>65 months</th>
<th>2004.4.28</th>
<th>low</th>
<th>high</th>
</tr>
</thead>
<tbody>
<tr>
<td>Euro foreign exchange $</td>
<td>1.1608</td>
<td>1.1907</td>
<td></td>
<td>.8252</td>
<td>1.2858</td>
</tr>
<tr>
<td>Euro foreign exchange ¥</td>
<td>121.32</td>
<td>130.17</td>
<td></td>
<td>89.30</td>
<td>140.31</td>
</tr>
<tr>
<td>Euro foreign exchange £</td>
<td>0.7111</td>
<td>0.6665</td>
<td></td>
<td>0.5711</td>
<td>0.7235</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>2003.4.28</th>
<th>12 months</th>
<th>2004.4.28</th>
<th>low</th>
<th>high</th>
</tr>
</thead>
<tbody>
<tr>
<td>$</td>
<td>1.1025</td>
<td>1.1907</td>
<td></td>
<td>1.0783</td>
<td>1.2858</td>
</tr>
<tr>
<td>¥</td>
<td>132.54</td>
<td>130.17</td>
<td></td>
<td>124.80</td>
<td>140.31</td>
</tr>
<tr>
<td>£</td>
<td>0.6914</td>
<td>0.6665</td>
<td></td>
<td>0.6556</td>
<td>0.7235</td>
</tr>
</tbody>
</table>

Tufte, 1990
Tufte’s Integrity

• Show data variation, not design variation
• Clear, detailed, and thorough labeling and appropriate scales
• Size of the graphic effect should be directly proportional to the numerical quantities ("lie factor")
Design Elements
Tufte’s Design

- Above all else show the data
- Maximize data-ink ratio
- Erase non-data ink
- Erase redundant data ink
- Revise and edit
Design Pyramid

- Functionality: effective
- Usability: efficient
- Aesthetics: affective
Subjective

• Aesthetics: Attractive things are perceived as more useful than unattractive ones
• Style: Communicates brand, process, who the designer is
• Playfulness: Encourages experimentation and exploration
• Vividness: Can make a visualization more memorable
Contrast
Repetition
Alignment
Proximity
Contrast

Peter’s cake metaphor ties in nicely with Galls Law

A complex system that works is invariably found to have evolved from a simple system that worked. The inverse proposition also appears to be true: A complex system designed from scratch never works and cannot be made to work. You have to start over, beginning with a working simple system.
Before

[Image of a scale with a chart showing obesity rates among OECD nations in 2004. The chart lists countries like Japan, Korea, Switzerland, Mexico, Italy, Austria, Denmark, France, Sweden, Netherlands, Turkey, Iceland, Poland, Belgium, Portugal, Germany, Ireland, Spain, Finland, Czech Republic, Slovak Republic, Luxembourg, Hungary, New Zealand, Australia, Greece, Canada, United Kingdom, and Mexico, with Japan having the lowest rate of 3.2% and United States having the highest rate of 32.2%.

After

[Image of a table showing a similar data set, but with each country's obesity rate highlighted and the table organized in a clear, readable format. The table includes columns for country and obesity rate, with Japan having the lowest rate of 3.2% and United States having the highest rate of 32.2%

G. Reynolds, Presentation Zen
Before

Mobile phone internet connectivity rate

- Japan: 94%
- South Korea: 89%
- U.S.A.: 34%
- France: 13%
- U.K.: 13%
- China: 5%

After

Mobile phone internet connectivity rate

- Japan: 94%
- South Korea: 89%
- U.S.A.: 34%
- France: 13%
- U.K.: 13%
- China: 5%

G. Reynolds, Presentation Zen
Faces of the Dead in Iraq, NY Times
Alignment

Before

After

S. Few, Show me the numbers
Proximity

S. Few, Show me the numbers
Small Multiples

S. Few, Show me the numbers
Small Multiples
Layering and Separation

<table>
<thead>
<tr>
<th>Train No.</th>
<th>3701</th>
<th>3301</th>
<th>3801</th>
<th>3542</th>
<th>3765</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>New York</strong></td>
<td>12:10</td>
<td>1:30</td>
<td>3:45</td>
<td>7:30</td>
<td>4:33</td>
</tr>
<tr>
<td><strong>Newark, N. J.</strong></td>
<td>1:43</td>
<td>10:30</td>
<td>5:21</td>
<td>8:50</td>
<td>11:45</td>
</tr>
<tr>
<td><strong>North Elizabeth</strong></td>
<td>....</td>
<td>......</td>
<td>......</td>
<td>......</td>
<td>6:45</td>
</tr>
<tr>
<td><strong>Elizabeth</strong></td>
<td>3:33</td>
<td>2:05</td>
<td>......</td>
<td>......</td>
<td>7:05</td>
</tr>
<tr>
<td><strong>Peekskill</strong></td>
<td>5:34</td>
<td>6:40</td>
<td>......</td>
<td>7:20</td>
<td>8:50</td>
</tr>
<tr>
<td><strong>Edison, N. J.</strong></td>
<td>4:45</td>
<td>5:20</td>
<td>4:40</td>
<td>2:10</td>
<td>11:05</td>
</tr>
<tr>
<td><strong>Princeton, N. J.</strong></td>
<td>1:30</td>
<td>......</td>
<td>......</td>
<td>3:30</td>
<td>7:30</td>
</tr>
</tbody>
</table>

Figure 2.25: Two versions of a railway timetable with the "bad" version on top. Inspired by an example from pgs. 54-55 of Tufte (1990).
Layering and Separation

<table>
<thead>
<tr>
<th>Location</th>
<th>12:10</th>
<th>1:30</th>
<th>3:45</th>
<th>7:30</th>
<th>4:33</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>New York</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Newark, N. J.</strong></td>
<td>1:43</td>
<td>10:30</td>
<td>5:21</td>
<td>8:50</td>
<td>11:45</td>
</tr>
<tr>
<td>North Elizabeth</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>6:45</td>
</tr>
<tr>
<td>Elizabeth</td>
<td>3:33</td>
<td>2:05</td>
<td></td>
<td></td>
<td>7:05</td>
</tr>
<tr>
<td>Peekskill</td>
<td>5:34</td>
<td>6:40</td>
<td></td>
<td>7:20</td>
<td>8:50</td>
</tr>
<tr>
<td><strong>Edison, N. J.</strong></td>
<td>4:45</td>
<td>5:20</td>
<td>4:40</td>
<td>2:10</td>
<td>11:05</td>
</tr>
<tr>
<td>Princeton, N. J.</td>
<td>1:30</td>
<td></td>
<td></td>
<td>3:30</td>
<td>7:30</td>
</tr>
<tr>
<td>Train No.</td>
<td>3701</td>
<td>3301</td>
<td>3801</td>
<td>3542</td>
<td>3765</td>
</tr>
</tbody>
</table>

Figure 2.25: Two versions of a railway timetable with the “bad” version on top. Inspired by an example from pgs. 54-55 of Tufte (1990).
Negative Space

CROSSWIND TAKEOFF
THE SLIP
CROSSWIND LANDINGS
SHORT FIELD TAKE OFF & LANDING
SOFT FIELD TAKE OFF & LANDING
FORCED LANDING
720 POWER TURNS
Negative Space

GIRL SCOUTS

USA network

Negative Space Logos
To Summarize
The Zen Aesthetic

- Simplicity
- Clarity
- Uncluttered
- Restraint
Tufte’s Graphical Excellence

• Interesting data
  – Complex ideas, multivariate data

• Clear, precise, concise presentation
  – Data-ink ratio

• Accurate communication
  – Lie factor
<table>
<thead>
<tr>
<th>Relationship</th>
<th>Points</th>
<th>Lines</th>
<th>Points &amp; Lines</th>
<th>Bars</th>
</tr>
</thead>
<tbody>
<tr>
<td>Nominal Comparison</td>
<td>When there is a need to narrow the quantitative scale, and in so doing, remove zero from its base</td>
<td>Avoid</td>
<td>Avoid</td>
<td>Either horizontal or vertical bars</td>
</tr>
<tr>
<td>Time Series</td>
<td>Avoid</td>
<td>Categorical subdivisions on X axis, quantitative values on Y axis; emphasis on overall pattern</td>
<td>Categorical subdivisions on X axis, quantitative values on Y axis; mutual emphasis on overall pattern and individual values</td>
<td>Categorical subdivisions on X axis, quantitative values on Y axis; emphasis on individual values</td>
</tr>
<tr>
<td>Ranking</td>
<td>When there is a need to narrow the quantitative scale, and in so doing, remove zero from its base</td>
<td>Avoid</td>
<td>Avoid</td>
<td>Horizontal bars are preferable, with values sorted in descending order.</td>
</tr>
<tr>
<td>Part-to-Whole</td>
<td>Avoid</td>
<td>Avoid</td>
<td>Avoid</td>
<td>Either horizontal or vertical bars</td>
</tr>
<tr>
<td>Deviation</td>
<td>Avoid</td>
<td>Especially useful when combined with time series</td>
<td>Useful when combined with time series and when a slight emphasis on individual values is desired</td>
<td>Either horizontal or vertical bars, except when combined with time series, which requires vertical bars</td>
</tr>
<tr>
<td>Distribution</td>
<td>Single</td>
<td>Known as a frequency polygon; emphasis on overall pattern</td>
<td>Avoid</td>
<td>Known as a histogram; emphasis primarily on individual values</td>
</tr>
<tr>
<td></td>
<td>Multiple</td>
<td>Use to mark the median in a box plot</td>
<td>Avoid</td>
<td>Use in the form of range bars in box plots</td>
</tr>
<tr>
<td>Correlation</td>
<td>Known as a scatter plot</td>
<td>Avoid</td>
<td>In this case the line is a trend line, not a line that connects the points.</td>
<td>Either horizontal or vertical bars; can be structured either as a correlation bar graph or a paired bar graph</td>
</tr>
</tbody>
</table>

S. Few, Show Me the Numbers, p. 87
Analysis Questions

• Who is the intended audience?
• What information does this visualization represent?
• How many data dimensions does it encode?
• List several tasks, comparisons or evaluations it enables
• What principles of excellence best describe why it is good / bad?
• Can you suggest any improvements?
• Why do you like / dislike this visualization?
Graphical displays

- Show the data
- Induce the viewer to think about the substance, rather than about methodology, graphic design, [or] the technology of graphic productions...
- Avoid distorting what the data have to say
- Present many numbers in a small space
- Make large data sets coherent
- Encourage the eye to compare different pieces of data
- Reveal the data at several levels of detail
- Serve a reasonably clear purpose
- Be closely integrated with the statistical and verbal descriptions