Psychology of Data Visualization: Course Overview

Michael Friendly
Psych 6135

http://euclid.psych.yorku.ca/www/psy6135/
@datavisFriendly
Introducing: me

I wear two hats, both reflected on my license plate:

Statistical graphics developer (categorical & multivariate data analysis)

mosaic plots for frequency tables

HE plots for MANOVA
Introducing: me

History of data visualization: Les Chevaliers; The Origins of Graphical Species (2018)

John Snow’s map of cholera in London, 1854

C. J. Minard: Flow maps of cotton trade

Visual explanation: What happened in the US Civil War?
Course Topics

• Varieties of information visualization
  ▪ Goals of visualization
  ▪ Survey of graphic forms

• History of information visualization

• Psychological models, theories and results
  ▪ What can people see, understand and remember from data displays?
  ▪ Perceptual aspects, cognitive aspects

• Software tools for information visualization

• Visualization in statistics: case studies
  ▪ Categorical data; High-D data; Dynamic and interactive methods

• Human factors research: how to tell what works
Your role

• Weekly readings – see the course web site for updates
• Discussion – no formal grade, but please contribute
• Discussion leader (20%)
  ▪ Each week 1-2 of you will lead a brief discussion on one of the readings or sub-topics (~ 5 min.)
• Class presentation (40%)
  ▪ In the last 2-3 weeks, each person will give a ~ 20 min presentation on a topic of research, application or software related to data visualization
• Research proposal (40%)
  ▪ Prepare a brief research proposal on a data visualization topic
Colin Ware, *Information Visualization*, 3rd Ed.
What perceptual science has to say about data visualization, from a bottom-up perspective
Course notes at: http://ccom.unh.edu/vislab/VisCourse/index.html

Alberto Cairo, *The Truthful Art*
Information graphics from a communication perspective
Blog: http://www.thefunctionalart.com/

Steven Kosslyn, *Elements of Graph Design*
A cognitive psychologist looks at graphs and presents some dos and don’ts

Hadley Wickham, *ggplot2: Elegant graphics for data analysis*, 2nd Ed.
Complete ggplot2 documentation: http://docs.ggplot2.org/current/
More books I like

Tamara Munzner (2014), *Visualization Analysis & Design*
An attractive new book combining computer science and design perspectives

Howard Wainer (2005), *Graphic discovery: a trout in the milk and other visual adventures*
A collection of essays on the history of graphics and other topics

Manuel Lima, *The Book of Trees: Visualizing branches of knowledge*
A visual delight; an entire history of tree-type diagrams

Keiran Healy, *Data Visualization: A Practical Introduction*
An accessible primer on how to create effective graphics from data using ggplot2
Online: [http://socvis.co](http://socvis.co)
Four books by Edward Tufte largely defined the landscape for data visualization and information design

Concepts introduced:
- chart junk,
- data-ink ratio,
- small multiples,
- substance takes precedence over visual design

Web site: https://www.edwardtufte.com
Blogs & Web resources

My web site, http://datavis.ca. Contains the Milestone Project on the history of data vis, Data Visualization gallery, links to books, papers and courses.

Kaiser Fung, http://junkcharts.typepad.com/. Fung discusses a variety of data displays and discusses how they can be improved.

Nathan Yau's blog, http://flowingdata.com. A large number of blog posts illustrating data visualization methods with tutorials on how to do these with R and other software.

http://visiphilia.org/. Statisticians Di Cook and Heike Hofmann from Iowa State University blog about data visualization topics, using R

Manuel Lima’s blog, http://www.visualcomplexity.com/vc/blog/, with hundreds of projects on all types of visualizations
Blogs & Web resources

DATA STORIES

http://datastori.es/. A podcast on data visualization with Enrico Bertini and Moritz Stefaner; interviews with over 100 graphic designers & developers.

Annual awards celebrate excellence and beauty in data visualizations, infographics, interactives & information art. https://www.informationisbeautifulawards.com

https://www.r-bloggers.com/. A large collection of posts on R news and tutorials by over 750 R bloggers.

Goal: Tell a credible story about some real data problem

Measles vaccination
Global warming
...

Data, pictures, models & stories
Two paths to enlightenment

Data, pictures, models & stories

Model-based

Exploratory

visualisation

story

summary

model

data
Data, pictures, models & stories

Now, tell the story!
Words, numbers and pictures

Pictures and images in a wider context

Modes of communication, as composed of words (story), numbers (symbols) and pictures (images) in different proportions.

E.g.,
Poetry ≈ 60% words + 40% images
Table ≈ 10% words + 80% numbers + 10% images
Words, numbers and pictures

Beauty: The 4th dimension

Modes of communication also vary in beauty & aesthetic appeal
Roles of graphics in communication

• Graphs (& tables) are forms of communication:
  ▪ What is the audience?
  ▪ What is the message?

**Analysis graphs:** design to see patterns, trends, aid the process of data description, interpretation

**Presentation graphs:** design to attract attention, make a point, illustrate a conclusion
Different graphs for different purposes

Presentation

Goal: the Wow! experience
Single image for a large audience
Tells a clear story!

Exploration

Goal: the Ah ha! Experience
Many images, for a narrow audience (you!), linked to analysis
Powerful graphs: Measels and vaccines

Visualizing the impact of health policy interventions

In 2015 Tynan DeBold & Dov Friedman in the Wall Street Journal show the effect of the introduction of vaccination programs in the US states on disease incidence, using color-coded heat maps for a variety of diseases.

Measles was decimated!

The message hits you between the eyes!

Powerful graphs make comparison easy

In 2014, vaccination rates declined and measles re-emerged in those areas

Effective graphs can cure ignorance, but not stupidity.

The best graphs pass the **Interocular Traumatic Test**: the message hits you between the eyes!
Data graph: Nightingale (1857)

The same, as a data graph, using time-series line plots
Many statisticians might prefer this today, but it doesn’t draw attention or interest as Flo’s original did.

![Graph of Causes of Mortality of the British Army in the East](image)
Rhetorical graph: Welfare income and Homeless deaths after the “Common Sense Revolution” 

Scott Sorli (2007)
Scatterplot of deaths vs. income
- Loess smooth + CI band
- Labels: year
- Color: party in power

The message here is interesting, but it lacks the power and eloquence of the original graph

As well, the relationship of deaths to time & party is lost
Racial profiling: Analysis graph

- Toronto Star (2002) study of police actions on a charge of simple possession of marijuana
  - release with a summons (Form 9) vs. hold for bail (Show cause)
  - Evidence for racial bias?
- First graph: mosaic display
  - area ~ frequency
  - shading: ~ residual
    - Obs > Expected in blue
    - Obs < Expected in red
Racial profiling: The process

How to communicate these results most effectively?
• What is the message? What features are directly comprehensible to the audience?

Graphic designer’s early attempts

My early attempts

Man behind the numbers
Racial profiling: Presentation graphic

Together, we created this self-explaining infographic.

- **Title** gives the main conclusion.
- **Legend** gives a layman’s description of shading levels.
- **Text description** gives details.
- **Bar width** ~ charges divided by % release.
- **Numbers shown in the cells**.

### Same charge, different treatment

Statistical analysis of single drug possession charges shows that blacks are much less likely to be released at the scene and much more likely to be held in custody for a bail hearing. Darker colours represent a stronger statistical link between skin colour and police treatment.

**Whites** are more likely to be released at the scene:
- 6,662 charges laid
- 78% released at the scene
- 14.5% released at station
- 7.5% held for bail

**Blacks** are much more likely to be held for bail hearings:
- 2,446 charges laid
- 64% released at the scene
- 20% released at station
- 16% held for bail

**Degree of likelihood**
- *Much less* likely to occur
- *Much more* likely to occur
- *More likely* to occur

*Source: Toronto police arrest records 1996-2002*
Why plot your data?

Graphs help us to see

patterns, trends, anomalies and other features

not otherwise easily apparent from numerical summaries.

Source: http://xkcd.com/523/
Why plot your data?

Three data sets with exactly the same bivariate summary statistics:

- Same correlations, linear regression lines, etc
- Indistinguishable from standard printed output
- Totally different interpretations!

Standard data  
r=0 but + 2 outliers  
Lurking variable?
Comparing groups: Analysis vs. Presentation graphs

Six different graphs for comparing groups in a one-way design
• which group means differ?
• equal variability?
• distribution shape?
• what do error bars mean?
• unusual observations?

Never use dynamite plots
Always explain what error bars mean
Consider tradeoff between summarization & exposure
Graphs of model coefficients are often clearer than tables

Source: tables2graphs.com
Effective data display

• Make the data stand out
  - Fill the data region (axes, ranges)
  - Use visually distinct symbols (shape, color) for different groups
  - Avoid chart junk, heavy grid lines that detract from the data

• Facilitate comparison
  - Emphasize the important comparisons visually
  - Side-by-side easier than in separate panels
  - “data” vs. a “standard” easier against a horizontal line
  - Show uncertainty where possible

• Effect ordering
  - For variables and unordered factors, arrange them according to the effects to be seen
**Comparisons**— Make visual comparisons easy

- **Visual grouping**— connect with lines, make key comparisons contiguous
- **Baselines**— compare *data* to *model* against a line, preferably horizontal
- **Frequencies** often better plotted on a square-root scale

![Standard histogram with fit](image1)

![Suspended rootogram](image2)

33
Make comparisons *direct*

- Use points not bars
- Connect similar by lines
- Same panel rather than different panels
Showing uncertainty

- Standard plots of observed vs. predicted lack a basis for assessment of uncertainty
- Confidence envelopes indicate extent of deviation
- Identify “noteworthy” observations to track them down

Example: Normal QQ plots used to assess normality of data
Effect ordering

• Information presentation is always ordered
  ▪ in time or sequence (a talk or written paper)
  ▪ in space (table or graph)
  ▪ Constraints of time & space are dominant—can conceal or reveal the important message

• Effect ordering for data display
  ▪ Sort the data by the effects to be seen
  ▪ Order the data to facilitate the task at hand
    • lookup – find a value
    • comparison – which is greater?
    • detection – find patterns, trends, anomalies
Effect order failure: the *Challenger* disaster

- Few events in history provide as compelling illustration of importance of appropriate ordering and display of information
  - On January 28, 1986, the space shuttle Challenger exploded on take-off.
  - The cause was later determined to be that rubber O-rings failed due to cold weather.

- Tables and charts presented to NASA by Thiokol engineers showed data from prior launches ordered by time (launch number), rather than by temperature—the crucial factor.

- The engineers’ charts were also remarkable for information obfuscation: “erosion depth” (O-ring damage), “blow-by” (soot on O-rings), ...
Visual explanation: Physics

• NASA appointed members of the Rogers Commission to investigate the cause of the disaster
• the noted physicist Richard Feynman discovered the cause: at low temperature, O-rings became brittle and were subject to failure
• in his testimony, he demonstrated the effect by plunging a rubber O-ring into a cup of ice water
Subsequent statistical analysis showed the relationship between launch temperature and O-ring failures.

As Tufte (1997) notes: the fatal flaw was in the ordering of the data.

The graph shown here is the result of a statistical model fit to the data.

- The **thick** line shows the predicted value of failure vs. temperature.
- The **red** dotted lines show uncertainty of the predicted values.
A presentation version of the previous graph alters the scales and describes the story in text annotations.
Graphic displays: Main effect ordering

- To see trends, patterns, anomalies: **Sort unordered factors by means or medians**

Data on barley yields
10 varieties x 6 sites x 2 years

3 way dot plot, sorted by main effect means

- Which site has the highest yield?
- Which variety is highest on average?
- Which site stands out in pattern over year?
Tabular displays: Main effect ordering

- Tables are often presented with rows/cols ordered alphabetically
  - good for lookup
  - bad for seeing patterns, trends, anomalies

Table 1: Average Barley Yields (rounded), Means by Site and Variety

<table>
<thead>
<tr>
<th>Variety</th>
<th>Crookston</th>
<th>Duluth</th>
<th>Grand Rapids</th>
<th>Morris</th>
<th>University Farm</th>
<th>Waseca</th>
<th>Mean</th>
</tr>
</thead>
<tbody>
<tr>
<td>Glabron</td>
<td>32</td>
<td>28</td>
<td>22</td>
<td>32</td>
<td>40</td>
<td>46</td>
<td>33.3</td>
</tr>
<tr>
<td>Manchuria</td>
<td>36</td>
<td>26</td>
<td>28</td>
<td>31</td>
<td>27</td>
<td>41</td>
<td>31.5</td>
</tr>
<tr>
<td>No. 457</td>
<td>40</td>
<td>28</td>
<td>26</td>
<td>36</td>
<td>35</td>
<td>50</td>
<td>35.8</td>
</tr>
<tr>
<td>No. 462</td>
<td>40</td>
<td>25</td>
<td>22</td>
<td>39</td>
<td>31</td>
<td>55</td>
<td>35.4</td>
</tr>
<tr>
<td>No. 475</td>
<td>38</td>
<td>30</td>
<td>17</td>
<td>33</td>
<td>27</td>
<td>44</td>
<td>31.8</td>
</tr>
<tr>
<td>Peatland</td>
<td>33</td>
<td>32</td>
<td>31</td>
<td>37</td>
<td>30</td>
<td>42</td>
<td>34.2</td>
</tr>
<tr>
<td>Svansota</td>
<td>31</td>
<td>24</td>
<td>23</td>
<td>30</td>
<td>31</td>
<td>43</td>
<td>30.4</td>
</tr>
<tr>
<td>Trebi</td>
<td>44</td>
<td>32</td>
<td>25</td>
<td>45</td>
<td>33</td>
<td>57</td>
<td>39.4</td>
</tr>
<tr>
<td>Velvet</td>
<td>37</td>
<td>24</td>
<td>28</td>
<td>32</td>
<td>33</td>
<td>44</td>
<td>33.1</td>
</tr>
<tr>
<td>Wisconsin No. 38</td>
<td>43</td>
<td>30</td>
<td>28</td>
<td>38</td>
<td>39</td>
<td>58</td>
<td>39.4</td>
</tr>
<tr>
<td>Mean</td>
<td>37.4</td>
<td>28.0</td>
<td>24.9</td>
<td>35.4</td>
<td>32.7</td>
<td>48.1</td>
<td>34.4</td>
</tr>
</tbody>
</table>
Tabular displays: Main effect ordering

- Better: sort rows/cols by means/medians
- Shade cells according to residual from additive model

Table 2: Average Barley Yields, sorted by Mean, shaded by residual from the model $Yield = Variety + Site$

<table>
<thead>
<tr>
<th>Variety</th>
<th>Grand Rapids</th>
<th>Duluth</th>
<th>University Farm</th>
<th>Morris</th>
<th>Crookston</th>
<th>Waseca</th>
<th>Mean</th>
</tr>
</thead>
<tbody>
<tr>
<td>Svansota</td>
<td>23</td>
<td>24</td>
<td>31</td>
<td>30</td>
<td>31</td>
<td>43</td>
<td>30.4</td>
</tr>
<tr>
<td>Manchuria</td>
<td>28</td>
<td>26</td>
<td>27</td>
<td>31</td>
<td>36</td>
<td>41</td>
<td>31.5</td>
</tr>
<tr>
<td>No. 475</td>
<td>17</td>
<td>30</td>
<td>27</td>
<td>33</td>
<td>38</td>
<td>44</td>
<td>31.8</td>
</tr>
<tr>
<td>Velvet</td>
<td>28</td>
<td>24</td>
<td>33</td>
<td>32</td>
<td>37</td>
<td>44</td>
<td>33.1</td>
</tr>
<tr>
<td>Glabron</td>
<td>22</td>
<td>28</td>
<td>40</td>
<td>32</td>
<td>32</td>
<td>46</td>
<td>33.3</td>
</tr>
<tr>
<td>Peatland</td>
<td>31</td>
<td>32</td>
<td>30</td>
<td>37</td>
<td>33</td>
<td>42</td>
<td>34.2</td>
</tr>
<tr>
<td>No. 462</td>
<td>22</td>
<td>25</td>
<td>31</td>
<td>39</td>
<td>40</td>
<td>55</td>
<td>35.4</td>
</tr>
<tr>
<td>No. 457</td>
<td>26</td>
<td>28</td>
<td>35</td>
<td>36</td>
<td>40</td>
<td>50</td>
<td>35.8</td>
</tr>
<tr>
<td>Wisconsin No. 38</td>
<td>28</td>
<td>30</td>
<td>39</td>
<td>38</td>
<td>43</td>
<td>58</td>
<td>39.4</td>
</tr>
<tr>
<td>Trebi</td>
<td>25</td>
<td>32</td>
<td>33</td>
<td>45</td>
<td>44</td>
<td>57</td>
<td>39.4</td>
</tr>
</tbody>
</table>

| Mean  | 24.9 | 28.0 | 32.7 | 35.4 | 37.4 | 48.1 | 34.4 |
Tabular displays: Main effect ordering

Yield difference, $\Delta y_{ij} = 1931 - 1932$ by Variety & Site

**Ordered:** by row and column means; **shaded:** by value ($|\Delta y_{ij}| > \{2,3\} \times \sigma(\Delta y_{ij})$)

What features stand out?

<table>
<thead>
<tr>
<th>Variety</th>
<th>Morris</th>
<th>Duluth</th>
<th>University Farm</th>
<th>Grand Rapids</th>
<th>Waseca</th>
<th>Crookston</th>
<th>Mean</th>
</tr>
</thead>
<tbody>
<tr>
<td>No. 475</td>
<td>-22</td>
<td>6</td>
<td>-5</td>
<td>4</td>
<td>6</td>
<td>12</td>
<td>0.1</td>
</tr>
<tr>
<td>Wisconsin No. 38</td>
<td>-18</td>
<td>2</td>
<td>1</td>
<td>14</td>
<td>1</td>
<td>14</td>
<td>2.4</td>
</tr>
<tr>
<td>Velvet</td>
<td>-13</td>
<td>4</td>
<td>13</td>
<td>-9</td>
<td>13</td>
<td>9</td>
<td>2.9</td>
</tr>
<tr>
<td>Peatland</td>
<td>-13</td>
<td>1</td>
<td>5</td>
<td>8</td>
<td>13</td>
<td>16</td>
<td>4.8</td>
</tr>
<tr>
<td>Manchuria</td>
<td>-7</td>
<td>6</td>
<td>0</td>
<td>11</td>
<td>15</td>
<td>7</td>
<td>5.5</td>
</tr>
<tr>
<td>Trebi</td>
<td>-3</td>
<td>3</td>
<td>7</td>
<td>9</td>
<td>15</td>
<td>5</td>
<td>6.1</td>
</tr>
<tr>
<td>Svansota</td>
<td>-9</td>
<td>3</td>
<td>8</td>
<td>13</td>
<td>9</td>
<td>20</td>
<td>7.3</td>
</tr>
<tr>
<td>No. 462</td>
<td>-17</td>
<td>6</td>
<td>11</td>
<td>5</td>
<td>21</td>
<td>18</td>
<td>7.4</td>
</tr>
<tr>
<td>Glabron</td>
<td>-6</td>
<td>4</td>
<td>6</td>
<td>15</td>
<td>17</td>
<td>12</td>
<td>8.0</td>
</tr>
<tr>
<td>No. 457</td>
<td>-15</td>
<td>11</td>
<td>17</td>
<td>13</td>
<td>16</td>
<td>11</td>
<td>8.8</td>
</tr>
<tr>
<td><strong>Mean</strong></td>
<td>-12.2</td>
<td>4.6</td>
<td>6.3</td>
<td>8.2</td>
<td>12.5</td>
<td>12.5</td>
<td>5.3</td>
</tr>
</tbody>
</table>
Tukey two-way plot of average barley yield

If there is no interaction,

\[ y_{ij} = \mu + \alpha_{\text{site}} + \beta_{\text{variety}} \]

Site & variety effects sorted automatically
Effects are spaced by fitted values

More variation among sites than varieties
Waseca best, by a wide margin
Multivariate data: correlation ordering

• Arrange *variables* so that:
  - Similar variables are contiguous
  - Ordered to show patterns of relations

• Arrange *observations* so that:
  - Similar variables are contiguous
  - Ordered to show patterns of relations
Correlation matrices

Baseball data: Batting, fielding and (log) Salary
Nobody wants to see all those decimals

<table>
<thead>
<tr>
<th></th>
<th>Assists</th>
<th>Atbat</th>
<th>Errors</th>
<th>Hits</th>
<th>Homer</th>
<th>logSal</th>
<th>Putouts</th>
<th>RBI</th>
<th>Runs</th>
<th>Walks</th>
<th>Years</th>
</tr>
</thead>
<tbody>
<tr>
<td>Assists</td>
<td>1.0000</td>
<td>0.3421</td>
<td>0.7035</td>
<td>0.3040</td>
<td>-0.1616</td>
<td>0.0500</td>
<td>-0.0434</td>
<td>0.0629</td>
<td>0.1790</td>
<td>0.1025</td>
<td>-0.0851</td>
</tr>
<tr>
<td>Atbat</td>
<td>0.3421</td>
<td>1.0000</td>
<td>0.3255</td>
<td>0.5551</td>
<td>0.4149</td>
<td>0.3096</td>
<td>0.7960</td>
<td>0.9000</td>
<td>0.6244</td>
<td>0.0127</td>
<td></td>
</tr>
<tr>
<td>Errors</td>
<td>0.7035</td>
<td>0.3256</td>
<td>1.0000</td>
<td>0.2799</td>
<td>-0.0097</td>
<td>-0.0208</td>
<td>0.0753</td>
<td>0.1502</td>
<td>0.1930</td>
<td>0.0819</td>
<td>-0.1565</td>
</tr>
<tr>
<td>Hits</td>
<td>0.3040</td>
<td>0.9640</td>
<td>0.2798</td>
<td>1.0000</td>
<td>0.5306</td>
<td>0.4496</td>
<td>0.2997</td>
<td>0.7885</td>
<td>0.9110</td>
<td>0.5873</td>
<td>0.0186</td>
</tr>
<tr>
<td>Homer</td>
<td>-0.1616</td>
<td>0.5551</td>
<td>-0.0097</td>
<td>0.5306</td>
<td>1.0000</td>
<td>0.3398</td>
<td>0.2509</td>
<td>0.8491</td>
<td>0.6310</td>
<td>0.4405</td>
<td>0.1135</td>
</tr>
<tr>
<td>logSal</td>
<td>0.0500</td>
<td>0.4149</td>
<td>-0.0208</td>
<td>0.4496</td>
<td>0.3398</td>
<td>1.0000</td>
<td>0.2245</td>
<td>0.4441</td>
<td>0.4260</td>
<td>0.4324</td>
<td>0.5374</td>
</tr>
<tr>
<td>Putouts</td>
<td>-0.0434</td>
<td>0.3096</td>
<td>0.0753</td>
<td>0.2997</td>
<td>0.2509</td>
<td>1.0000</td>
<td>0.3121</td>
<td>0.7790</td>
<td>0.2710</td>
<td>0.2809</td>
<td>-0.0200</td>
</tr>
<tr>
<td>RBI</td>
<td>0.0629</td>
<td>0.7960</td>
<td>0.1501</td>
<td>0.7885</td>
<td>0.8491</td>
<td>0.4441</td>
<td>1.0000</td>
<td>0.7790</td>
<td>0.5695</td>
<td>0.1297</td>
<td></td>
</tr>
<tr>
<td>Runs</td>
<td>0.1793</td>
<td>0.8998</td>
<td>0.1926</td>
<td>0.9106</td>
<td>0.6310</td>
<td>0.4256</td>
<td>0.2712</td>
<td>1.0000</td>
<td>0.6970</td>
<td>-0.0120</td>
<td></td>
</tr>
<tr>
<td>Walks</td>
<td>0.1025</td>
<td>0.6244</td>
<td>0.0819</td>
<td>0.5873</td>
<td>0.4404</td>
<td>0.4324</td>
<td>0.2809</td>
<td>0.5695</td>
<td>0.6970</td>
<td>1.0000</td>
<td>0.1348</td>
</tr>
<tr>
<td>Years</td>
<td>-0.0851</td>
<td>0.0127</td>
<td>-0.1565</td>
<td>0.0186</td>
<td>0.1134</td>
<td>0.5374</td>
<td>-0.0200</td>
<td>0.1297</td>
<td>-0.0120</td>
<td>0.1348</td>
<td>1.0000</td>
</tr>
</tbody>
</table>

If you are going to present the numbers, round a lot

<table>
<thead>
<tr>
<th></th>
<th>Assists</th>
<th>Atbat</th>
<th>Errors</th>
<th>Hits</th>
<th>Homer</th>
<th>logSal</th>
<th>Putouts</th>
<th>RBI</th>
<th>Runs</th>
<th>Walks</th>
<th>Years</th>
</tr>
</thead>
<tbody>
<tr>
<td>Assists</td>
<td>100</td>
<td>34</td>
<td>70</td>
<td>30</td>
<td>-16</td>
<td>5</td>
<td>-4</td>
<td>6</td>
<td>18</td>
<td>10</td>
<td>-9</td>
</tr>
<tr>
<td>Atbat</td>
<td>34</td>
<td>100</td>
<td>33</td>
<td>96</td>
<td>56</td>
<td>41</td>
<td>31</td>
<td>80</td>
<td>90</td>
<td>62</td>
<td>1</td>
</tr>
<tr>
<td>Errors</td>
<td>70</td>
<td>33</td>
<td>100</td>
<td>28</td>
<td>-1</td>
<td>-2</td>
<td>8</td>
<td>15</td>
<td>19</td>
<td>8</td>
<td>-16</td>
</tr>
<tr>
<td>Hits</td>
<td>30</td>
<td>96</td>
<td>28</td>
<td>100</td>
<td>53</td>
<td>45</td>
<td>30</td>
<td>79</td>
<td>91</td>
<td>59</td>
<td>2</td>
</tr>
<tr>
<td>Homer</td>
<td>-16</td>
<td>56</td>
<td>-1</td>
<td>53</td>
<td>100</td>
<td>34</td>
<td>25</td>
<td>85</td>
<td>63</td>
<td>44</td>
<td>11</td>
</tr>
<tr>
<td>logSal</td>
<td>5</td>
<td>41</td>
<td>-2</td>
<td>45</td>
<td>34</td>
<td>100</td>
<td>22</td>
<td>44</td>
<td>43</td>
<td>43</td>
<td>54</td>
</tr>
<tr>
<td>Putouts</td>
<td>-4</td>
<td>31</td>
<td>8</td>
<td>30</td>
<td>25</td>
<td>22</td>
<td>100</td>
<td>31</td>
<td>27</td>
<td>28</td>
<td>-2</td>
</tr>
<tr>
<td>RBI</td>
<td>6</td>
<td>80</td>
<td>15</td>
<td>79</td>
<td>85</td>
<td>44</td>
<td>31</td>
<td>100</td>
<td>78</td>
<td>57</td>
<td>13</td>
</tr>
<tr>
<td>Runs</td>
<td>18</td>
<td>90</td>
<td>19</td>
<td>91</td>
<td>63</td>
<td>43</td>
<td>27</td>
<td>78</td>
<td>100</td>
<td>70</td>
<td>-1</td>
</tr>
<tr>
<td>Walks</td>
<td>10</td>
<td>62</td>
<td>8</td>
<td>59</td>
<td>44</td>
<td>43</td>
<td>28</td>
<td>57</td>
<td>70</td>
<td>100</td>
<td>13</td>
</tr>
<tr>
<td>Years</td>
<td>-9</td>
<td>1</td>
<td>-16</td>
<td>2</td>
<td>11</td>
<td>54</td>
<td>-2</td>
<td>13</td>
<td>-1</td>
<td>13</td>
<td>100</td>
</tr>
</tbody>
</table>
**Rendering:** a correlation value can be displayed in different ways, for different tasks

**Correlation ordering:**
- A PCA finds weighted sums of variable to maximize variance accounted for
- Angles between vectors reflect the correlations
- → Arrange variables in the order of their angles
This is a corrgram display of the correlations among the baseball statistics, with the variables ordered alphabetically.
The same display, with the variables sorted according to the angles between vectors in the PCA

Not that dramatic, but it isolates the positive & negative correlations
Graphs: Good/Bad, Excellent/Evil

• Like good writing, good graphical displays of data communicate ideas with:
  ▪ clarity,
  ▪ precision, and
  ▪ efficiency—avoids graphic clutter
  ▪ Even better: excellent graphs make the message obvious

• Like poor writing, bad graphical displays:
  ▪ distort or obscure the data,
  ▪ make it harder to understand or compare, or
  ▪ thwart the communicative effect the graph should convey.
  ▪ Even worse: evil graphs distort, or mislead.
Bad graphs are easy in Excel

Friends don’t let friends use Excel for data visualization or statistics

How many things are wrong with this graph?
Pie charts are easy to abuse

What’s wrong with this picture?

On the other hand, pie charts are a great source of merriment for people interested in graphics.
But, can be used to great effect

This graphic uses pie charts to show the transport of different kinds of goods to the ports of Paris and the principal maritime ports:

- the size of each pie reflects total
- the sectors reflect relative %
- location places them in context

Album de Statistique Graphique, 1885, plate 17.
What was the intent of the designer of this graphic?
Which category led to the greatest total deaths?
What was the proportion of deaths due to strokes?
Did more people die from strokes vs. accidents?
Simple re-design makes it clearer

Total Deaths in America by Cause in 2007

<table>
<thead>
<tr>
<th>Cause</th>
<th>Percentage</th>
</tr>
</thead>
<tbody>
<tr>
<td>Heart disease</td>
<td>25.42%</td>
</tr>
<tr>
<td>Cancer</td>
<td>23.22%</td>
</tr>
<tr>
<td>Stroke (cerebrovascular diseases)</td>
<td>5.61%</td>
</tr>
<tr>
<td>Chronic lower respiratory diseases</td>
<td>5.28%</td>
</tr>
<tr>
<td>Accidents (unintentional injuries)</td>
<td>5.10%</td>
</tr>
<tr>
<td>Alzheimer's disease</td>
<td>3.08%</td>
</tr>
<tr>
<td>Diabetes</td>
<td>2.95%</td>
</tr>
<tr>
<td>Influenza and Pneumonia</td>
<td>2.18%</td>
</tr>
<tr>
<td>Nephritis, nephrotic syndrome, and nephrosis</td>
<td>1.92%</td>
</tr>
<tr>
<td>Septicemia</td>
<td>1.44%</td>
</tr>
<tr>
<td>All other causes</td>
<td>23.81%</td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td><strong>100.00%</strong></td>
</tr>
</tbody>
</table>
Double Y-axis: Really evil graphs

After pie charts, double Y-axis graphs have caused more trouble than almost any other

OMG, autism has been increasing directly with sales of organic food!
William Playfair invented the pie chart, line chart and bar chart. In this figure, he shows 3 parallel time series over a 250-year period, 1560--1810

- weekly wages of a good mechanic
- price of wheat
- reigning monarch

Goal: show that workers were better off most recently (1810) than in the past
Or, another graph would have been better.

A modern re-vision plots the ratio of price of wheat to wages directly.
Rep. Jason Chaffetz, R-Utah, sparred with Planned Parenthood president Cecile Richards during a high-profile hearing on Sept. 29, 2015 and presented this graph.

"In pink, that's the reduction in the breast exams, and the red is the increase in the abortions. That's what's going on in your organization."

Created by an anti-abortion group it is a deliberate attempt to mislead.

Can you see why?

See: http://www.politifact.com/truth-o-meter/statements/2015/oct/01/jason-chaffetz/chart-shown-planned-parenthood-hearing-misleading/
This graph shows the actual data from the Planned Parenthood reports used by Americans United for Life.

The number of abortions was relatively steady.

Some services like pap smears, dropped due to changing medical standards about who should be screened and how often.

What are a few improvements that could be made to this graph?
Showing a wider range of PP activities puts these data in context.

PP activities were far higher for contraception and STD testing.
A study by Abigail Friendly (2017) wanted to show the use of benefits afforded to Toronto developers for their contributions of different types over time.

Figure 9: Section 37 benefits by type (1998–2015)

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Roads, streetscapes</td>
<td>30</td>
<td>35</td>
<td>54</td>
<td>83</td>
<td>15</td>
</tr>
<tr>
<td>Culture, community, recreation</td>
<td>26</td>
<td>50</td>
<td>59</td>
<td>47</td>
<td>16</td>
</tr>
<tr>
<td>Parks</td>
<td>27</td>
<td>41</td>
<td>41</td>
<td>52</td>
<td>16</td>
</tr>
<tr>
<td>Affordable housing</td>
<td>17</td>
<td>26</td>
<td>38</td>
<td>56</td>
<td>11</td>
</tr>
<tr>
<td>Public art</td>
<td>26</td>
<td>25</td>
<td>41</td>
<td>32</td>
<td>4</td>
</tr>
<tr>
<td>Heritage</td>
<td>16</td>
<td>13</td>
<td>26</td>
<td>18</td>
<td>3</td>
</tr>
<tr>
<td>Transit</td>
<td>11</td>
<td>7</td>
<td>10</td>
<td>20</td>
<td>3</td>
</tr>
<tr>
<td>Libraries</td>
<td>6</td>
<td>2</td>
<td>5</td>
<td>11</td>
<td>1</td>
</tr>
<tr>
<td>Other</td>
<td>3</td>
<td>6</td>
<td>7</td>
<td>8</td>
<td>3</td>
</tr>
</tbody>
</table>

Color background scale from light to dark highlights the largest values.

Most frequent benefits appear at the top.

Can see overall trends and anomalies.

What happened in 2014-2016?

This graph reports the results of a survey by Sherman Kent for the CIA with the question:

*What [probability/number] would you assign to the phrase "[phrase]"*

The goal was to contribute to an understanding of how intelligence analysts use these terms

Why can this be considered a graphical failure?

Graphical excellence

This graph shows the same data, as both dotplots & boxplots

We can see a lot more:
- “about even” has very low variability
- the last 3 categories are listed out of order
- the extreme outliers stand out
- skewness is – for high probability, + for low probability

Technical notes:
- software: ggplot2
- design: faint grid lines
- color: points use transparent color & jittering; outliers also shown in black

From: https://github.com/zonination/perceptions
This graph uses “ridgeline” plots to show the same data.

Each one is a small version of a density plot showing a smoothed version of the distribution.

Stacking them in this way allows center, variability, shape and other features to be readily compared.
Chart junk or effective info vis?

What is the message?
Who is the audience?
Suzana Herculano-Houzel has a new method for determining counts of cortical neurons across different species. How to present this effectively?

Goal: compare mammal species brain size and cortical neuron count

Neuron count is shown both as numbers and bars

What do you think?

How could this be made better?

A scatterplot makes clear how humans differ from other species

- Using scaled images as point symbols also conveys brain size
- Primates are distinguished from non-primates by text color

This is arguably a more effective display.
As a scatterplot – log scale

Perhaps even better is to make the plot using log scales for both axes.

The relationship is now approx. linear.
In the movie, *An Inconvenient Truth* (2006), Al Gore used the now-famous “hockey stick” graph to show that human activities had greatly increased the degree of global warming over the recent past.

The goal was to raise public awareness and call for action to curb environmental effects: CO₂ emissions as the main agent.

Movie: [https://www.youtube.com/watch?v=8ZUoYGAi5i0](https://www.youtube.com/watch?v=8ZUoYGAi5i0); [http://www.imdb.com/title/tt0497116/](http://www.imdb.com/title/tt0497116/)
Sir John Houghton presents the original Northern Hemisphere hockey stick graph to the Intergovernmental Panel on Climate Change (IPCC) in 2005. It is based on an analysis by Mann, Bradley & Hughes (1990), with a smoothed curve and uncertainty intervals.
The MBH (1999) paper had used a wide variety of data sources. They were combined using a novel statistical technique, the first eigenvector-based climate field reconstruction (CFR).
Climate scientists understood this; the sceptics did not.

See: https://en.wikipedia.org/wiki/Hockey_stick_controversy for details
Taking a longer view, and adding a lot of extraneous historical details, climate sceptics were easily able to mount alternative explanations.
Perhaps one fault with the original graphs was trying to show noisy data, from many sources, over too wide a time span.
A politically-incorrect graphic shows very clearly the effect of global warming on panty size

Climate change: other explanations

This infographic attempts to relate global warming to the decrease in pirates

Aside from the substance, how many things are wrong about this graphic?

**Simple explanation:**
Lack of pirates causes
global warming!

**Conclusion:**
To stop global warming,
become a pirate!

Climate change: animation

This animation shows a rotating globe indicating local effects of global warming (red is warmer).

The graph below shows the global average temperature changes from historical averages.

Video: https://youtu.be/xhqEkyJDBho
Summary

• Graphs as a form of communication
  ▪ Data (numbers), words, images → Stories

• Analysis graphs vs. presentation graphs

• Some principles of effective data display
  ▪ Make the data stand out
  ▪ Facilitate comparisons
  ▪ Effect ordering