Information Visualization I
UM Big Data Summer Institute 2019

Matthew Kay
Assistant Professor
School of Information & Computer Science and Engineering
University of Michigan
A little bit about me

Master’s and Bachelor’s in CS (Fine Art minor) from the University of Waterloo

PhD in CSE from the University of Washington

My work draws upon human–computer interaction, visualization, design, and statistics
What I would like to do today

Motivate why visualization is important.

Give you grounding to help you systematically create effective visualizations:

Perceptual principles

Design principles
Why visualize?
Anscombe’s 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>10.0</td>
<td>y</td>
<td>8.04</td>
<td></td>
<td>10.0</td>
<td>y</td>
<td>9.14</td>
<td></td>
</tr>
<tr>
<td>x</td>
<td>8.0</td>
<td>y</td>
<td>6.95</td>
<td></td>
<td>8.0</td>
<td>y</td>
<td>8.14</td>
<td></td>
</tr>
<tr>
<td>x</td>
<td>13.0</td>
<td>y</td>
<td>7.58</td>
<td></td>
<td>13.0</td>
<td>y</td>
<td>8.74</td>
<td></td>
</tr>
<tr>
<td>x</td>
<td>9.0</td>
<td>y</td>
<td>8.81</td>
<td></td>
<td>9.0</td>
<td>y</td>
<td>8.77</td>
<td></td>
</tr>
<tr>
<td>x</td>
<td>11.0</td>
<td>y</td>
<td>8.33</td>
<td></td>
<td>11.0</td>
<td>y</td>
<td>9.26</td>
<td></td>
</tr>
<tr>
<td>x</td>
<td>14.0</td>
<td>y</td>
<td>9.96</td>
<td></td>
<td>14.0</td>
<td>y</td>
<td>8.10</td>
<td></td>
</tr>
<tr>
<td>x</td>
<td>6.0</td>
<td>y</td>
<td>7.24</td>
<td></td>
<td>6.0</td>
<td>y</td>
<td>6.13</td>
<td></td>
</tr>
<tr>
<td>x</td>
<td>4.0</td>
<td>y</td>
<td>4.26</td>
<td></td>
<td>4.0</td>
<td>y</td>
<td>3.10</td>
<td></td>
</tr>
<tr>
<td>x</td>
<td>12.0</td>
<td>y</td>
<td>10.84</td>
<td></td>
<td>12.0</td>
<td>y</td>
<td>9.13</td>
<td></td>
</tr>
<tr>
<td>x</td>
<td>7.0</td>
<td>y</td>
<td>4.82</td>
<td></td>
<td>7.0</td>
<td>y</td>
<td>7.26</td>
<td></td>
</tr>
<tr>
<td>x</td>
<td>5.0</td>
<td>y</td>
<td>5.68</td>
<td></td>
<td>5.0</td>
<td>y</td>
<td>4.74</td>
<td></td>
</tr>
</tbody>
</table>

4 datasets, **same** means, variances, correlation
Anscombe's quartet
O-ring failure in the Challenger

History of O-Ring Damage in Field Joints (Cont)

O-Ring Temp
°F

SRM 1 1 2 2 3 3 4 4 5 5 6 6 7 7 8 8 9 9 10 10 11 11 12 12 13 13 14 14 15 15 16 16 17 17 18 18 19 19 20 20 21 21 22 22 23 23 24 24


* No Erosion

O-ring failure, Morton-Thiokol
O-ring failure in the Challenger

O-ring failure, Tufte
O-ring failure in the Challenger

Extrapolation of damage curve to the cold
Challenger launch: 31° forecasted temperature for January 28, 1986

Dots indicate temperature and O-ring damage for 24 successful launches prior to Challenger. Curve shows increasing damage is related to cooler temperatures.

O-ring failure, Tufte
Visualize to see patterns you wouldn’t otherwise
Visualize for communication

What’s Really Warming the World?
By Eric Roston and Blake Mejia
June 24, 2015

Skeptics of manmade climate change offer various natural causes to explain why the Earth has warmed 1.4 degrees Fahrenheit since 1880. But can these account for the planet’s rising temperature? Scroll down to see how much different factors, both natural and industrial, contribute to global warming, based on findings from NASA’s Goddard Institute for Space Studies.

[https://www.bloomberg.com/graphics/2015-whats-warming-the-world/]
How do we turn data into visualizations?
<table>
<thead>
<tr>
<th>Make</th>
<th>mpg</th>
<th>cyl</th>
<th>disp</th>
<th>hp</th>
<th>drat</th>
<th>wt</th>
<th>qsec</th>
<th>vs</th>
<th>am</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mazda RX4</td>
<td>21.0</td>
<td>6</td>
<td>160.0</td>
<td>110</td>
<td>3.50</td>
<td>2.620</td>
<td>16.40</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>Mazda RX4 Wagon</td>
<td>21.0</td>
<td>6</td>
<td>160.0</td>
<td>110</td>
<td>3.90</td>
<td>2.075</td>
<td>17.02</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>Datsun 710</td>
<td>22.8</td>
<td>4</td>
<td>108.0</td>
<td>93</td>
<td>3.85</td>
<td>2.320</td>
<td>19.51</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>Hornet 4 Drive</td>
<td>21.4</td>
<td>6</td>
<td>258.0</td>
<td>110</td>
<td>3.08</td>
<td>3.215</td>
<td>19.44</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>Hornet Sportabout</td>
<td>18.7</td>
<td>8</td>
<td>360.0</td>
<td>175</td>
<td>3.15</td>
<td>3.440</td>
<td>17.02</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Valiant</td>
<td>18.1</td>
<td>6</td>
<td>225.0</td>
<td>105</td>
<td>2.76</td>
<td>3.460</td>
<td>20.22</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>Duster 360</td>
<td>14.3</td>
<td>8</td>
<td>360.0</td>
<td>245</td>
<td>3.21</td>
<td>3.570</td>
<td>15.04</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Merc 240D</td>
<td>24.4</td>
<td>4</td>
<td>146.7</td>
<td>62</td>
<td>3.69</td>
<td>3.190</td>
<td>20.00</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>Merc 230</td>
<td>22.0</td>
<td>4</td>
<td>140.8</td>
<td>95</td>
<td>3.92</td>
<td>3.150</td>
<td>22.90</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>Merc 280</td>
<td>19.2</td>
<td>6</td>
<td>167.6</td>
<td>123</td>
<td>3.92</td>
<td>3.440</td>
<td>18.30</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>Merc 280C</td>
<td>17.8</td>
<td>6</td>
<td>167.6</td>
<td>123</td>
<td>3.92</td>
<td>3.440</td>
<td>18.90</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>Merc 450SE</td>
<td>16.4</td>
<td>8</td>
<td>275.8</td>
<td>180</td>
<td>3.07</td>
<td>4.070</td>
<td>17.40</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Merc 450SL</td>
<td>17.3</td>
<td>8</td>
<td>275.8</td>
<td>180</td>
<td>3.07</td>
<td>3.750</td>
<td>17.50</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Merc 450SLC</td>
<td>15.2</td>
<td>8</td>
<td>275.0</td>
<td>100</td>
<td>3.07</td>
<td>3.700</td>
<td>16.00</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Cadillac Fleetwood</td>
<td>10.4</td>
<td>8</td>
<td>472.0</td>
<td>205</td>
<td>2.93</td>
<td>5.350</td>
<td>17.98</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Lincoln Continental</td>
<td>10.4</td>
<td>8</td>
<td>460.0</td>
<td>215</td>
<td>3.00</td>
<td>5.424</td>
<td>17.32</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Chrysler Imperial</td>
<td>14.7</td>
<td>8</td>
<td>440.0</td>
<td>230</td>
<td>3.33</td>
<td>5.345</td>
<td>17.42</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Fiat 128</td>
<td>32.4</td>
<td>4</td>
<td>78.7</td>
<td>66</td>
<td>4.08</td>
<td>2.200</td>
<td>19.47</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>Honda Civic</td>
<td>30.4</td>
<td>4</td>
<td>75.7</td>
<td>52</td>
<td>4.93</td>
<td>1.615</td>
<td>18.52</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>Toyota Corolla</td>
<td>33.9</td>
<td>4</td>
<td>71.1</td>
<td>65</td>
<td>4.22</td>
<td>1.335</td>
<td>19.90</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>Toyota Corona</td>
<td>21.5</td>
<td>4</td>
<td>120.1</td>
<td>97</td>
<td>3.70</td>
<td>2.455</td>
<td>20.01</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>Dodge Challenger</td>
<td>15.5</td>
<td>8</td>
<td>318.0</td>
<td>150</td>
<td>2.76</td>
<td>3.520</td>
<td>16.87</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>AMC Javelin</td>
<td>15.2</td>
<td>8</td>
<td>304.0</td>
<td>150</td>
<td>3.15</td>
<td>3.435</td>
<td>17.30</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Camaro Z28</td>
<td>13.3</td>
<td>8</td>
<td>350.0</td>
<td>245</td>
<td>3.73</td>
<td>3.840</td>
<td>15.41</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Pontiac Firebird</td>
<td>19.2</td>
<td>8</td>
<td>400.0</td>
<td>175</td>
<td>3.08</td>
<td>3.845</td>
<td>17.05</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Fiat X1-9</td>
<td>27.3</td>
<td>4</td>
<td>79.0</td>
<td>66</td>
<td>4.08</td>
<td>1.355</td>
<td>18.90</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td>mpg</td>
<td>cyl</td>
<td>disp</td>
<td>hp</td>
<td>drat</td>
<td>wt</td>
<td>qsec</td>
<td>vs</td>
<td>am</td>
</tr>
<tr>
<td>----------</td>
<td>------</td>
<td>-----</td>
<td>------</td>
<td>-----</td>
<td>------</td>
<td>-----</td>
<td>------</td>
<td>-----</td>
<td>-----</td>
</tr>
<tr>
<td>Mazda RX4</td>
<td>21.0</td>
<td>6</td>
<td>160.0</td>
<td>110</td>
<td>3.90</td>
<td>2.620</td>
<td>16.46</td>
<td></td>
<td>0</td>
</tr>
<tr>
<td>Mazda RX4 Wagon</td>
<td>21.0</td>
<td>6</td>
<td>160.0</td>
<td>110</td>
<td>3.90</td>
<td>2.875</td>
<td>17.02</td>
<td></td>
<td>0</td>
</tr>
<tr>
<td>Datsun 710</td>
<td>22.8</td>
<td>4</td>
<td>108.0</td>
<td>93</td>
<td>3.85</td>
<td>2.320</td>
<td>18.61</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>Hornet 4-Door</td>
<td>23.4</td>
<td>6</td>
<td>258.0</td>
<td>110</td>
<td>3.08</td>
<td>3.215</td>
<td>19.44</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>Hornet Sportabout</td>
<td>18.7</td>
<td>8</td>
<td>360.0</td>
<td>175</td>
<td>3.15</td>
<td>3.440</td>
<td>17.02</td>
<td></td>
<td>0</td>
</tr>
<tr>
<td>Valiant</td>
<td>18.1</td>
<td>6</td>
<td>225.0</td>
<td>105</td>
<td>2.76</td>
<td>3.450</td>
<td>20.22</td>
<td></td>
<td>1</td>
</tr>
<tr>
<td>Duster 360</td>
<td>14.3</td>
<td>8</td>
<td>360.0</td>
<td>245</td>
<td>3.21</td>
<td>3.570</td>
<td>15.04</td>
<td></td>
<td>0</td>
</tr>
<tr>
<td>Merc 240D</td>
<td>24.4</td>
<td>4</td>
<td>146.7</td>
<td>62</td>
<td>3.69</td>
<td>3.190</td>
<td>20.00</td>
<td></td>
<td>1</td>
</tr>
<tr>
<td>Merc 230</td>
<td>22.8</td>
<td>4</td>
<td>140.8</td>
<td>95</td>
<td>3.92</td>
<td>3.150</td>
<td>22.90</td>
<td></td>
<td>1</td>
</tr>
<tr>
<td>Merc 280</td>
<td>19.2</td>
<td>6</td>
<td>167.6</td>
<td>123</td>
<td>3.92</td>
<td>3.440</td>
<td>18.30</td>
<td></td>
<td>1</td>
</tr>
<tr>
<td>Merc 280C</td>
<td>17.8</td>
<td>6</td>
<td>167.6</td>
<td>123</td>
<td>3.92</td>
<td>3.440</td>
<td>18.90</td>
<td></td>
<td>1</td>
</tr>
<tr>
<td>Merc 450SE</td>
<td>16.4</td>
<td>8</td>
<td>275.8</td>
<td>180</td>
<td>3.07</td>
<td>4.070</td>
<td>17.40</td>
<td></td>
<td>0</td>
</tr>
<tr>
<td>Merc 450SL</td>
<td>17.3</td>
<td>8</td>
<td>275.8</td>
<td>180</td>
<td>3.07</td>
<td>3.750</td>
<td>17.60</td>
<td></td>
<td>0</td>
</tr>
<tr>
<td>Merc 450SLC</td>
<td>15.2</td>
<td>8</td>
<td>275.0</td>
<td>160</td>
<td>3.07</td>
<td>3.750</td>
<td>16.00</td>
<td></td>
<td>0</td>
</tr>
<tr>
<td>Cadillac Fleetwood</td>
<td>10.4</td>
<td>8</td>
<td>472.0</td>
<td>205</td>
<td>2.93</td>
<td>5.520</td>
<td>17.98</td>
<td></td>
<td>0</td>
</tr>
<tr>
<td>Lincoln Continental</td>
<td>10.4</td>
<td>8</td>
<td>460.0</td>
<td>215</td>
<td>3.00</td>
<td>5.420</td>
<td>17.82</td>
<td></td>
<td>0</td>
</tr>
<tr>
<td>Chrysler Imperial</td>
<td>14.7</td>
<td>8</td>
<td>440.0</td>
<td>220</td>
<td>3.23</td>
<td>5.345</td>
<td>17.42</td>
<td></td>
<td>0</td>
</tr>
<tr>
<td>Fiat 128</td>
<td>32.4</td>
<td>4</td>
<td>78.7</td>
<td>66</td>
<td>4.08</td>
<td>2.200</td>
<td>19.47</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>Honda Civic</td>
<td>30.4</td>
<td>4</td>
<td>75.7</td>
<td>52</td>
<td>4.93</td>
<td>1.615</td>
<td>18.52</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>Toyota Corolla</td>
<td>33.9</td>
<td>4</td>
<td>71.1</td>
<td>65</td>
<td>4.22</td>
<td>1.835</td>
<td>19.90</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>Toyota Corona</td>
<td>23.5</td>
<td>4</td>
<td>120.1</td>
<td>97</td>
<td>3.70</td>
<td>2.455</td>
<td>20.01</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>Dodge Challenger</td>
<td>15.5</td>
<td>8</td>
<td>318.0</td>
<td>150</td>
<td>2.76</td>
<td>3.520</td>
<td>16.67</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>AMC Javelin</td>
<td>15.2</td>
<td>8</td>
<td>304.0</td>
<td>150</td>
<td>3.15</td>
<td>3.435</td>
<td>17.30</td>
<td></td>
<td>0</td>
</tr>
<tr>
<td>Camaro Z28</td>
<td>13.3</td>
<td>8</td>
<td>350.0</td>
<td>245</td>
<td>3.73</td>
<td>3.840</td>
<td>15.41</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Pontiac Firebird</td>
<td>19.2</td>
<td>8</td>
<td>400.0</td>
<td>175</td>
<td>3.08</td>
<td>3.845</td>
<td>17.05</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Fiat X1-9</td>
<td>27.3</td>
<td>4</td>
<td>79.0</td>
<td>66</td>
<td>4.08</td>
<td>1.935</td>
<td>18.90</td>
<td>1</td>
<td>1</td>
</tr>
</tbody>
</table>
Let’s systematize “turning data into a vis”

data -> ??? -> marks on the screen (or paper)
Let’s systematize “turning data into a vis”

data -> ??? -> marks on the screen (or paper)

??? = some vis API
     = some way of thinking about vis systematically
Let’s systematize “turning data into a vis”

data -> ??? -> marks on the screen (or paper)

??? = New function for every chart type:
    scatter_plot(data, ...)
    bar_chart(data, ...)
    ...

Let’s systematize “turning data into a visual”

data -> ??? -> marks on the screen (or paper)

??? = New function for every chart type:
   scatter_plot(data, ...)
   bar_chart(data, ...)
   ...

   Every new chart is a new adventure!
   Too many specs! — Too high level!
Let’s systematize “turning data into a vis”

data -> ??? -> marks on the screen (or paper)

??? = New function for every chart type
  = Low-level drawing functions
    draw_point(...) 
    draw_rectangle(...)
Let’s systematize “turning data into a vis”

data -> ??? -> marks on the screen (or paper)

??? = New function for every chart type
    = Low-level drawing functions
      draw_point(...) 
      draw_rectangle(...) 

Too low level!
Let’s systematize “turning data into a vis”

data -> ??? -> marks on the screen (or paper)

??? = New function for every chart type
    = Low-level drawing functions
    = Grammar of graphics

Encode data with visual channels
Display encodings with marks
Visual channels
(ggplot “aesthetics”)
Visual channels -----> Marks

(ggplot “aesthetics”) (ggplot “geometries”)

Position  
Color Hue  
Texture  
Connection  
Containment  
Density  
Color Saturation  
Shape  
Length  
Angle  
Slope  
Area  
Volume  

Points  
Lines  
Bars  

etc
Grammar of graphics

Codifies data types, encodings/channels, marks

Maps data -> channels -> marks

Makes visualization specification straightforward

Undergirds ggplot, Tableau, Vega-Lite, Altair,... (terms may vary)
<table>
<thead>
<tr>
<th></th>
<th>mpg</th>
<th>cyl</th>
<th>disp</th>
<th>hp</th>
<th>drat</th>
<th>wt</th>
<th>qsec</th>
<th>vs</th>
<th>am</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mazda RX4</td>
<td>21.0</td>
<td>6</td>
<td>160.0</td>
<td>110</td>
<td>3.90</td>
<td>2.620</td>
<td>16.46</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>Mazda RX4 Wag</td>
<td>21.0</td>
<td>6</td>
<td>160.0</td>
<td>110</td>
<td>3.90</td>
<td>2.875</td>
<td>17.02</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>Datsun 710</td>
<td>22.8</td>
<td>4</td>
<td>180.0</td>
<td>93</td>
<td>3.85</td>
<td>2.320</td>
<td>18.61</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>Hornet 4 Drive</td>
<td>21.4</td>
<td>6</td>
<td>258.0</td>
<td>110</td>
<td>3.08</td>
<td>3.215</td>
<td>19.44</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>Hornet Sportabout</td>
<td>18.7</td>
<td>8</td>
<td>360.0</td>
<td>175</td>
<td>3.15</td>
<td>3.440</td>
<td>17.02</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Valiant</td>
<td>18.1</td>
<td>6</td>
<td>225.0</td>
<td>105</td>
<td>2.76</td>
<td>3.450</td>
<td>20.22</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>Duster 360</td>
<td>14.3</td>
<td>8</td>
<td>360.0</td>
<td>245</td>
<td>3.21</td>
<td>3.570</td>
<td>15.04</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Merc 240D</td>
<td>24.4</td>
<td>4</td>
<td>146.7</td>
<td>62</td>
<td>3.69</td>
<td>3.190</td>
<td>20.00</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>Merc 230</td>
<td>22.8</td>
<td>4</td>
<td>140.8</td>
<td>95</td>
<td>3.92</td>
<td>3.150</td>
<td>22.90</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>Merc 280</td>
<td>19.2</td>
<td>6</td>
<td>167.6</td>
<td>123</td>
<td>3.07</td>
<td>3.440</td>
<td>18.90</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>Merc 280C</td>
<td>17.8</td>
<td>6</td>
<td>167.6</td>
<td>123</td>
<td>3.92</td>
<td>3.440</td>
<td>18.90</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>Merc 450SE</td>
<td>16.4</td>
<td>8</td>
<td>275.8</td>
<td>180</td>
<td>3.07</td>
<td>4.070</td>
<td>17.40</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Merc 450SL</td>
<td>17.3</td>
<td>8</td>
<td>275.8</td>
<td>180</td>
<td>3.07</td>
<td>3.750</td>
<td>17.50</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Merc 450SLC</td>
<td>15.2</td>
<td>8</td>
<td>275.8</td>
<td>180</td>
<td>3.07</td>
<td>3.750</td>
<td>17.50</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Cadillac Fleetwood</td>
<td>10.4</td>
<td>8</td>
<td>472.0</td>
<td>205</td>
<td>2.93</td>
<td>5.250</td>
<td>17.98</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Lincoln Continental</td>
<td>10.4</td>
<td>8</td>
<td>460.0</td>
<td>215</td>
<td>3.00</td>
<td>5.424</td>
<td>17.82</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Chrysler Imperial</td>
<td>14.7</td>
<td>8</td>
<td>440.0</td>
<td>220</td>
<td>3.23</td>
<td>5.345</td>
<td>17.42</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Fiat 128</td>
<td>32.4</td>
<td>4</td>
<td>70.7</td>
<td>66</td>
<td>4.08</td>
<td>2.200</td>
<td>19.47</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>Honda Civic</td>
<td>21.5</td>
<td>4</td>
<td>120.1</td>
<td>97</td>
<td>3.70</td>
<td>2.455</td>
<td>20.01</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>Toyota Corolla</td>
<td>35.0</td>
<td>4</td>
<td>71.1</td>
<td>64</td>
<td>4.22</td>
<td>1.935</td>
<td>18.90</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>Toyota Corona</td>
<td>21.5</td>
<td>4</td>
<td>120.1</td>
<td>97</td>
<td>3.70</td>
<td>2.455</td>
<td>20.01</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>Dodge Challenger</td>
<td>15.5</td>
<td>8</td>
<td>200.0</td>
<td>150</td>
<td>2.76</td>
<td>3.520</td>
<td>16.07</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>AMC Javelin</td>
<td>13.2</td>
<td>8</td>
<td>304.0</td>
<td>150</td>
<td>3.15</td>
<td>3.435</td>
<td>17.30</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Camaro Z28</td>
<td>12.3</td>
<td>8</td>
<td>350.0</td>
<td>245</td>
<td>3.73</td>
<td>3.840</td>
<td>15.41</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Pontiac Firebird</td>
<td>19.2</td>
<td>8</td>
<td>400.0</td>
<td>175</td>
<td>3.08</td>
<td>3.845</td>
<td>17.05</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Fiat X1-9</td>
<td>27.3</td>
<td>4</td>
<td>79.0</td>
<td>66</td>
<td>4.08</td>
<td>1.935</td>
<td>18.90</td>
<td>1</td>
<td>1</td>
</tr>
</tbody>
</table>
Grammar of graphics
(data types, channels, marks)
Grammar of graphics
(data types, channels, marks)

mpg: numeric
wt: numeric
Grammar of graphics
(data types, channels, marks)

mpg: numeric
wt: numeric
wt -> x position
mpg -> y position
Grammar of graphics
(data types, channels, marks)

mpg: numeric
wt: numeric
wt -> x position
mpg -> y position
mark: point
Grammar of graphics
(data types, channels, marks)

mpg: numeric
wt: numeric

wt -> x position
mpg -> y position

mark: point
Grammar of graphics
(data types, channels, marks)

mpg: numeric
wt: numeric
manual: nominal
wt -> x position
mpg -> y position

mark: point
Grammar of graphics
(data types, channels, marks)
mpg: numeric
wt: numeric
manual: nominal
wt -> x position
mpg -> y position
manual -> color
mark: point
Grammar of graphics
(data types, channels, marks)
mpg: numeric
wt: numeric
manual: nominal
wt -> x position
mpg -> y position
manual -> shape
mark: point
Grammar of graphics
(data types, channels, marks)

mpg: numeric
wt: numeric
manual: nominal

wt -> x position
mpg -> y position
manual -> color
manual -> shape

mark: point
Grammar of graphics
(data types, channels, marks)

mpg: numeric
wt: numeric
manual: nominal
wt -> x position
mpg -> y position
manual -> color

mark: point
Grammar of graphics
(data types, channels, marks)

wt: numeric
manual: nominal
bin(wt) -> x position
count(wt)-> height
manual -> color

mark: bar
Why is the grammar of graphics useful?

1. Easier to specify many charts, combinations

2. Helps you design/evaluate charts systematically
1. Easier to specify many charts, combinations

mpg: numeric
wt: numeric

wt -> x position
mpg -> y position

mark: point
1. Easier to **specify** many charts, combinations

Not:

```python
some_big_function_to_make_scatterplots(
    my_data,
    a_bunch_of_options
)
```
1. Easier to **specify** many charts, combinations

Not:

```r
some_function_to_draw_grid()
some_function_to_draw_axes()
for (row in data) {
    draw_point(data[i]["x"], ...)
}
...
```
1. Easier to specify many charts, combinations
   e.g., in ggplot
   (data, channels, marks):

   ggplot(mtcars, aes(
       x = wt,
       y = mpg
   )) +
   geom_point()
2. Helps you **design** charts systematically

Data → channels → marks → viewer

viewer’s reconstruction of the data
2. Helps you **design** charts systematically

Data $\rightarrow$ channels $\rightarrow$ marks $\rightarrow$ viewer

viewer's reconstruction of the data
2. Helps you design charts systematically

Data → channels → marks → viewer

viewer’s reconstruction of the data

How well do these match, given the channel used?
2. Helps you **design** charts systematically

E.g.,

How accurately do people **perceive position**?

How accurately do people **perceive area**?

**Channels**

- Position
- Color Hue
- Texture
- Connection
- Containment
- Density
- Color Saturation
- Shape
- Length
- Angle
- Slope
- Area
- Volume
2. Helps you **design** charts systematically

E.g.,

How accurately do people perceive **position** for quantitative data?

...for **ordered** data?

...for **nominal** data?

etc.

**Channels**

- Position
- Color Hue
- Texture
- Connection
- Containment
- Density
- Color Saturation
- Shape
- Length
- Angle
- Slope
- Area
- Volume
2. Helps you **design** charts systematically

E.g.,

What **channel** is best for quantitative data?
...for **ordered** data?
...for **nominal** data?

etc.

**Channels**
- Position
- Color Hue
- Texture
- Connection
- Containment
- Density
- Color Saturation
- Shape
- Length
- Angle
- Slope
- Area
- Volume
How good is a visual channel / encoding?

Length encoding:

→

→
How good is a **visual channel / encoding**?

**Length encoding:**

→

→
How good is a **visual channel / encoding**?

**Length encoding:**

- 
- 

**Area encoding:**
How good is a **visual channel / encoding**?

**Length encoding:**

![Length encoding diagram]

**Area encoding:**

![Area encoding diagram]
Encodings help us judge chart effectiveness

What **encoding** is best for quantitative data?  
...for ordered data?  
...for nominal data?  
etc.

### Nominal

- Position
- Color Hue
- Texture
- Connection
- Containment
- Density
- Color Saturation
- Shape
- Length
- Angle
- Slope
- Area
- Volume
Encodings help us judge chart effectiveness

**Ordinal**
- Position
- Density
- Color Saturation
- Color Hue
- Texture
- Connection
- Containment
- Length
- Angle
- Slope
- Area
- Volume
- Shape

**Nominal**
- Position
- Color Hue
- Texture
- Connection
- Containment
- Density
- Color Saturation
- Shape
- Length
- Angle
- Slope
- Area
- Volume
Encodings help us judge chart effectiveness.
Pick one, cross it off...
## Pick one, cross it off...

<table>
<thead>
<tr>
<th>Quantitative</th>
<th>Nominal</th>
</tr>
</thead>
<tbody>
<tr>
<td>Position</td>
<td>Position</td>
</tr>
<tr>
<td>Length</td>
<td>Color Hue</td>
</tr>
<tr>
<td>Angle</td>
<td>Texture</td>
</tr>
<tr>
<td>Slope</td>
<td>Connection</td>
</tr>
<tr>
<td>Area</td>
<td>Containment</td>
</tr>
<tr>
<td>Volume</td>
<td>Density</td>
</tr>
<tr>
<td>Density</td>
<td>Color Saturation</td>
</tr>
<tr>
<td>Color Saturation</td>
<td>Shape</td>
</tr>
<tr>
<td>Color Hue</td>
<td>Length</td>
</tr>
<tr>
<td>Texture</td>
<td>Angle</td>
</tr>
<tr>
<td>Connection</td>
<td>Slope</td>
</tr>
<tr>
<td>Containment</td>
<td>Area</td>
</tr>
<tr>
<td>Shape</td>
<td>Volume</td>
</tr>
</tbody>
</table>
Effectiveness

This chart works because it uses accurate channels (ones with low estimation error).

This is the essence of effectiveness.
What about this?
What about this?

Quantitative

- Position
- Length
- Angle
- Slope
- Area
- Volume
- Density
- Color Saturation
- Color Hue

Other:

- Texture
- Connection
- Containment
- Shape
Other insights from perception
Reference lines can help...
Reference lines can help...
Reference lines can help...
Reference lines can help...
Reference lines can help...
Reference lines can help...
Reference lines

Induce bias...

...but can be used to decrease error
Popout and preattentiveness

https://www.csc2.ncsu.edu/faculty/healey/PP/
Popout and preattentiveness

What do people see first?
What can people see separately?

Preattentiveness
-\rightarrow popout
-\rightarrow layering
Popout and preattentiveness

Preattentiveness
-→ popout
-→ layering

What do people see first?

What can people see separately?
Popout and preattentiveness

What do people see first?

What can people see separately?

Preattentiveness
- \( \rightarrow \) popout
- \( \rightarrow \) layering
Color
Sequential / diverging data

[http://www.research.ibm.com/people/l/lloydt/color/color.HTM]
Sequential / diverging data

[http://www.research.ibm.com/people/l/lloydt/color/color.HTM]
Sequential / diverging scales

Ordered / quantitative data may be **sequential** or **diverging**

This impacts encoding choice, for example:

**Sequential color scale:**

![Sequential color scale]

**Diverging color scale:**

![Diverging color scale]
Use perceptually uniform colormaps

For continuous color maps, Viridis (and co)...
For discrete colormaps, Color Brewer...

[http://colorbrewer2.org]
For more, hclwizard / colorspace R package

HCL Color Space

The hclwizard provides tools for manipulating and assessing colors and palettes based on the underlying colorspace software (available in R and Python). It leverages the HCL color space, a color model that is based on human color perception and thus makes it easy to choose good color palettes by varying three color properties: Hue (= type of color, dominant wavelength) - Chroma (= colorfulness) - Luminance (= brightness). As shown in the color swatches below each property can be varied while keeping the other two properties fixed.

Color Palettes

For color coding data visualizations it is crucial to choose a palette that appropriately captures the underlying information. Three types...
Grammar of graphics + Perception helps us design more effective charts
Grammar of Graphics + Perception

Think in data types, channels, and marks.

Helps you specify and design charts using perceptually effective channels.

Consider sequential / diverging nature of data.

Use perceptually uniform colormaps.
Questions about the grammar of graphics?
Grammar of graphics
(data types, channels, marks)

mpg: numeric
wt: numeric
wt -> x position
mpg -> y position
mark: point
Group activity

What are the variables / types?

Channels / encodings?

Marks?

Is this effective?
Design guidelines
Some rough design guidelines*

1. Match effectiveness with importance
2. Avoid ambiguity
3. Locality is king / eyes beat memory
4. Establish viewing order
5. Layer, layer, layer
6. When in doubt, grid
7. Treat visual attributes like adjectives

* These guidelines are drawn largely from my experience + personal preferences + the literature. Design is messy, these are not perfect, others will disagree with me, etc. Caveat emptor.
1. Match effectiveness with importance
2. Avoid ambiguity

Does the 3D mean anything here?

(Hint: No)
2. Avoid ambiguity

Marks should **not** have multiple reasonable interpretations

If it **looks like** it could **come from** data, it **should** come from data
3. Locality is king / eyes beat memory

No:
Thing → Information I need to understand thing

Yes:
Thing ← Information I need to understand thing
The left panel shows the Bayesian censored log-linear model, which gives us a posterior probability distribution over the mean log(JND) for each value of r. In the center panel we rank and group visualizations based on how precise estimations of correlations are with them (lower expected JND implies higher precision). In the right panel we estimate the ratio of average JNDs between successive groups over all values of r from 0.3 to 0.8. The low precision group is between ~1.5 and 3 times more precise than the chance group. The high precision group is between ~1.5 and 2 times more precise than the medium precision group.
The left panel shows the Bayesian censored log-linear model, which gives us a posterior probability distribution over the mean log(JND) for each value of r. In the center panel we rank and group visualizations based on how precise estimations of correlations are with them (lower expected JND implies higher precision). In the right panel we estimate the ratio of average JNDs between successive groups over all values of r from 0.3 to 0.8. The low precision group is between ~1.5 and 3 times more precise than the chance group. The high precision group is between ~1.5 and 2 times more precise than the medium precision group.
1. The final Bayesian censored log-linear model gives us a posterior probability distribution over the mean log(JND) for each value of r.

2. We rank and group visualizations based on how precise people’s estimations of correlations are with them (lower expected JND implies higher precision).

3. We estimate the ratio of average JNDs between successive groups over all values of r from 0.3 to 0.8. The low precision group is between ~1.5 and 3 times more precise than the chance group. The high precision group is between ~1.5 and 2 times more precise than the medium precision group.

The left panel shows the Bayesian censored log-linear model, which gives us a posterior probability distribution over the mean log(JND) for each value of r. In the center panel, we rank and group visualizations based on how precise estimations of correlations are with them (lower expected JND implies higher precision). In the right panel, we estimate the ratio of average JNDs between successive groups over all values of r from 0.3 to 0.8. The low precision group is between ~1.5 and 3 times more precise than the chance group. The high precision group is between ~1.5 and 2 times more precise than the medium precision group.
1. The final Bayesian censored log-linear model gives us a posterior probability distribution over the mean log(JND) for each value of $r$.

2. We rank and group visualizations based on how precise people’s estimations of correlations are with them (lower expected JND implies higher precision).

3. We estimate the ratio of average JNDs between successive groups over all values of $r$ from 0.3 to 0.8.

The low precision group is between ~1.5 and 3 times more precise than the chance group.

The high precision group is between ~1.5 and 2 times more precise than the medium precision group.
1. The final Bayesian censored log-linear model gives us a posterior probability distribution over the mean log(JND) for each value of $r$.

2. We rank and group visualizations based on how precise people’s estimations of correlations are with them (lower expected JND implies higher precision).

3. We estimate the ratio of average JNDs between successive groups over all values of $r$ from 0.3 to 0.8.

The low precision group is between ~1.5 and 3 times more precise than the chance group.

The high precision group is between ~1.5 and 2 times more precise than the medium precision group.

Count lookups!
4. Establish viewing order

Know where your audience will look first, second.

Think like a movie director. Are you telling a story?

https://www.youtube.com/watch?v=v4seDVfgwOg
4. Establish viewing order
Can be as simple as some numbers...

1. The final Bayesian censored log-linear model gives us a posterior probability distribution over the mean log(JND) for each value of $r$.

2. We rank and group visualizations based on how precise people's estimations of correlations are with them (lower expected JND implies higher precision).

3. We estimate the ratio of average JNDs between successive groups over all values of $r$ from 0.3 to 0.8.

The low precision group is between ~1.5 and 3 times more precise than the chance group.

The high precision group is between ~1.5 and 2 times more precise than the medium precision group.
4. Establish viewing order

Or more complex, relying on salience, other visual cues, viewer expectations (maybe) ...
And you will read this at the end

You will read this first

And then you will read this

Then this one
4. Establish viewing order

Or more complex, relying on **salience**, other visual cues, viewer expectations (maybe) ...
4. Establish viewing order

Or more complex, relying on salience, other visual cues, viewer expectations (maybe) ...
5. Layer, layer, layer

Design for micro-macro reading

Pre-attentive attributes help

5. Layer, layer, layer

PVI Score: State presidential vote relative to nationwide vote

(small multiples)
(small multiples = double use of position)

year -> wrapped column (x position)
(small multiples = double use of position)

year -> wrapped column (x position)

~longitude -> column (x position)
~latitude  -> row (y position)
6. When in doubt, grid

And get synchronized axes as a bonus.
7. Treat visual attributes like adjectives

Don’t use three attributes (size, color, shape, ...) to create emphasis where one or two will do.

The very tall building is very extremely tall.
The final Bayesian censored log-linear model gives us a posterior probability distribution over the mean log(JND) for each value of r.

2. We rank and group visualizations based on how precise people's estimations of correlations are with them (lower expected JND implies higher precision).

3. We estimate the ratio of average JNDs between successive groups over all values of r from 0.3 to 0.8.

The high precision group is between ~1.5 and 2 times more precise than the medium precision group.

The low precision group is between ~1.5 and 3 times more precise than the chance group.

- donut – negative
- donut – positive
- line – negative
- line – positive
- ordered line – negative
- ordered line – positive
- parallel coordinates – negative
- parallel coordinates – positive
- radar – negative
- radar – positive
- scatterplot – negative
- scatterplot – positive
- stacked area – negative
- stacked area – positive
- stacked bar – negative
- stacked bar – positive
- stacked line – negative
- stacked line – positive

low precision
medium precision
high precision
1. The final Bayesian censored log-linear model gives us a posterior probability distribution over the mean log(JND) for each value of $r$.

2. We rank and group visualizations based on how precise people’s estimations of correlations are with them (lower expected JND implies higher precision).

3. We estimate the ratio of average JNDs between successive groups over all values of $r$ from 0.3 to 0.8.

- The low precision group is between ~1.5 and 3 times more precise than the chance group.
- The high precision group is between ~1.5 and 2 times more precise than the medium precision group.

The donut plot is used to distinguish negative and positive correlations. The line plot is used to distinguish positive and negative correlations. The ordered line plot is used to distinguish negative and positive correlations. The parallel coordinates plot is used to distinguish negative and positive correlations. The stacked area plot is used to distinguish negative and positive correlations. The stacked bar plot is used to distinguish negative and positive correlations. The stacked line plot is used to distinguish negative and positive correlations.
(7b. Obey the pen)

Even visual texture is pleasing

Also makes it easier to create visual hierarchy and call out something important when you need to
Some rough design guidelines*

1. Match effectiveness with importance
2. Avoid ambiguity
3. Locality is king / eyes beat memory
4. Establish viewing order
5. Layer, layer, layer
6. When in doubt, grid
7. Treat visual attributes like adjectives

* These guidelines are drawn largely from my experience + personal preferences + the literature. Design is messy, these are not perfect, others will disagree with me, etc. Caveat emptor.
In sum

Understand the **effectiveness** of different aesthetics (encodings).

Understand **where** your viewer will look and **what** they want to do (tasks).

Visualize **as a reflex** during analysis.
Questions?
Examples / exercises
Prediction and memory

Draw your line on the chart below

[https://nyti.ms/2jX8zue]
Small multiples

Group activity

What are the variables / types?

Channels / encodings?

Marks?

Is this effective?
Hack Your Way To Scientific Glory

You're a social scientist with a hunch: The U.S. economy is affected by whether Republicans or Democrats are in office. Try to show that a connection exists, using real data going back to 1948. For your results to be publishable in an academic journal, you'll need to prove that they are "statistically significant" by achieving a low enough p-value.

1. CHOOSE A POLITICAL PARTY
   - Republicans
   - Democrats

2. DEFINE TERMS
   - Which politicians do you want to include?
     - [x] Governors
     - [x] Senators
     - [x] Representatives
   - How do you want to measure economic performance?
     - [x] Employment
     - [x] Inflation
     - [x] GDP
     - [x] Stock prices
     - Other options
       - [x] Factor in power: Weight more powerful positions more heavily
       - [x] Exclude recessions: Don't include economic recessions

3. IS THERE A RELATIONSHIP?
   - Given how you've defined your terms, does the economy do better, worse, or about the same when more Republicans are in power? Each dot below represents one month of data.

4. IS YOUR RESULT SIGNIFICANT?
   - If there were no connection between the economy and politics, what is the probability that you'd get results at least as strong as yours? That probability is your p-value, and by convention, you need a p-value of 0.05 or less to get published.

Result: Unpublishable

With a p-value of 0.43, your findings are not statistically significant. Try defining your terms differently.

SPLOM: Scatter plot matrix

[https://bl.ocks.org/mbostock/4063663]
Hyperbolic trees

[https://youtu.be/fhbQy_NCwWI]
Evolution of bacteria

The evolution of bacteria on a "mega-plate" Petri dish

https://vimeo.com/180908160
Small multiples

PVI Score: State presidential vote relative to nationwide vote

Small multiples

PVI Score: State presidential vote relative to nationwide vote

More Rep.

More Dem.