Why Visualize?
## Anscombe’s Quartet

<table>
<thead>
<tr>
<th>X</th>
<th>Y</th>
<th>X</th>
<th>Y</th>
<th>X</th>
<th>Y</th>
<th>X</th>
<th>Y</th>
</tr>
</thead>
<tbody>
<tr>
<td>10.0</td>
<td>8.04</td>
<td>10.0</td>
<td>9.14</td>
<td>10.0</td>
<td>7.46</td>
<td>8.0</td>
<td>6.58</td>
</tr>
<tr>
<td>8.0</td>
<td>6.95</td>
<td>8.0</td>
<td>8.14</td>
<td>8.0</td>
<td>6.77</td>
<td>8.0</td>
<td>5.76</td>
</tr>
<tr>
<td>13.0</td>
<td>7.58</td>
<td>13.0</td>
<td>8.74</td>
<td>13.0</td>
<td>12.74</td>
<td>8.0</td>
<td>7.71</td>
</tr>
<tr>
<td>9.0</td>
<td>8.81</td>
<td>9.0</td>
<td>8.77</td>
<td>9.0</td>
<td>7.11</td>
<td>8.0</td>
<td>8.84</td>
</tr>
<tr>
<td>11.0</td>
<td>8.33</td>
<td>11.0</td>
<td>9.26</td>
<td>11.0</td>
<td>7.81</td>
<td>8.0</td>
<td>8.47</td>
</tr>
<tr>
<td>14.0</td>
<td>9.96</td>
<td>14.0</td>
<td>8.10</td>
<td>14.0</td>
<td>8.84</td>
<td>8.0</td>
<td>7.04</td>
</tr>
<tr>
<td>6.0</td>
<td>7.24</td>
<td>6.0</td>
<td>6.13</td>
<td>6.0</td>
<td>6.08</td>
<td>8.0</td>
<td>5.25</td>
</tr>
<tr>
<td>4.0</td>
<td>4.26</td>
<td>4.0</td>
<td>3.10</td>
<td>4.0</td>
<td>5.39</td>
<td>19.0</td>
<td>12.50</td>
</tr>
<tr>
<td>12.0</td>
<td>10.84</td>
<td>12.0</td>
<td>9.13</td>
<td>12.0</td>
<td>8.15</td>
<td>8.0</td>
<td>5.56</td>
</tr>
<tr>
<td>7.0</td>
<td>4.82</td>
<td>7.0</td>
<td>7.26</td>
<td>7.0</td>
<td>6.42</td>
<td>8.0</td>
<td>7.91</td>
</tr>
<tr>
<td>5.0</td>
<td>5.68</td>
<td>5.0</td>
<td>4.74</td>
<td>5.0</td>
<td>5.73</td>
<td>8.0</td>
<td>6.89</td>
</tr>
</tbody>
</table>

Same:
N
Mean X
Mean Y
Variance X
Variance Y
Regression Correlation

F.J. Anscombe, 1973
Anscombe’s Quartet

dataset = I

dataset = II

dataset = III

dataset = IV

F.J. Anscombe, 1973
Choosing a Tool
Good News/Bad News

Arbor.js
A library of force-directed layout algorithms plus abstractions for graph organization and refresh handling.

CartoDB
A web service for mapping, analyzing and building applications with data.

Chromajs
Interactive color space explorer that allows to preview a set of linear interpolated equidistant colors.

Circos
A software package for visualizing data in a circular layout.

Cola.js
A library for arranging networks using constraint-based optimization techniques.

ColorBrewer
A web tool for selecting colors for maps.

Cubism.js
A library for creating interactive time series and horizon graphics based on D3.js.

Cytoscape
An application for visualizing complex networks and integrating these with any type of attribute data.

D3.js
An small, flexible and efficient library to create and manipulate interactive documents based on data.

Dance.js
A simple data-driven visualization framework based on Data.js and Underscore.js

Transform Script
- Import Sax
- Split data repeatedly on newline into rows
- Split split repeatedly on ",
- Promote new 0 to header

http://selection.datavisualization.ch
**Goal**

Is your goal explanation or exploration?

<table>
<thead>
<tr>
<th>Explanatory:</th>
<th>Exploratory:</th>
</tr>
</thead>
<tbody>
<tr>
<td>You already know what you want to say.</td>
<td>You want to find out what the data means.</td>
</tr>
</tbody>
</table>
Speed/Flexibility Tradeoff

Do you have lots of time? Do you need customization?

Speed, Ease

Flexibility
What is compatible with your working habits?

Python

R

Analytics app

Web

Libraries

Versions
Do you need to share with others?

- Image files (PNG, TIFF, PDF, etc.)
- Interactivity (click, brush, rollover, zoom, pan, etc.)
- Code sharing (notebook interface)
- Privacy (authentication, authorization)
Specificity

Is there a tool aimed at your problem in particular?
What tool will let you show your data without distortion or chartjunk?
What are the costs?

- Purchase price
- License over time
- License constraints (e.g., attribution, constraints on derivative works)
New Tools of Note
Seaborn

- Goal: exploratory
- Speed/Flexibility: Speed
- Dependencies: python 2.7 or 3.3+, numpy, scipy, matplotlib, pandas
- Sharing: export image files
- Specific Uses: statistical graphs
- Graphical Quality: good quality graphs, ability to tweak colors, axes, etc.
- Costs: open source
Bokeh
Bokeh

- Goal: exploratory or explanatory
- Speed/Flexibility: low-, intermediate-, or high-level
- Dependencies: python
- Sharing: Bokeh server, embedding in web pages or notebooks
- Specific Uses: novel interactive visualizations in the browser
- Graphical Quality: very good quality
- Costs: open source
Plotly

Dashboards

Line and Scatter Plots

Bubble Charts

Filled Area Plots

View Tutorial

Bar Charts

Pie Charts

2D Histograms

Range Sliders and Selectors

Gauge Chart

Dot plots

Dumbbell plots

Graphing Multiple Chart Types
<table>
<thead>
<tr>
<th>Environment name</th>
<th>Organism name</th>
<th>method</th>
<th>Project id</th>
<th>name</th>
<th>chemical formula</th>
<th>neutral mass</th>
</tr>
</thead>
<tbody>
<tr>
<td>M94</td>
<td>The Environment</td>
<td>analysis</td>
<td>ENIGMA</td>
<td>tyrosine</td>
<td>C9H11N03</td>
<td>181.0738933</td>
</tr>
<tr>
<td>M94</td>
<td>The Environment</td>
<td>analysis</td>
<td>ENIGMA</td>
<td>betaine</td>
<td>C5H11N02</td>
<td>117.0789781</td>
</tr>
<tr>
<td></td>
<td>M94</td>
<td>analysis</td>
<td>ENIGMA</td>
<td>monosaccharides (fructose, mannose, glucose, galactose, inositol)</td>
<td>C6H12O6</td>
<td>180.0633881</td>
</tr>
<tr>
<td></td>
<td>M94</td>
<td>analysis</td>
<td>ENIGMA</td>
<td>disaccharide 1 (sucrose, maltose, trehalose, cellulobiose, lactose, etc)</td>
<td>C12H22O11</td>
<td>342.1162111</td>
</tr>
<tr>
<td></td>
<td>M94</td>
<td>analysis</td>
<td>ENIGMA</td>
<td>disaccharide 2 (sucrose, maltose, trehalose, cellulobiose, lactose, etc)</td>
<td>C12H22O11</td>
<td>342.1162111</td>
</tr>
<tr>
<td></td>
<td>M94</td>
<td>analysis</td>
<td>ENIGMA</td>
<td>lactate</td>
<td>C3H6O3</td>
<td>90.0316940</td>
</tr>
</tbody>
</table>

Click the column headers to choose x and y columns to graph. Use the different colors to match x columns with y columns.

OPTIONS
- Error Bars
- Asymmetric Errors
- Group By
- Text

INSERT INTO
- Make a new plot
Goal: exploratory
Speed/Flexibility: Speed
Dependencies: library for python, R, Matlab, Excel, or JS, or use web app
Sharing: Plot.ly, Plotly Server
Specific Uses: collaborative analysis
Graphical Quality: good quality
Costs: plotly libs are open source, Plotly Server is licensed
Shiny

**Multidimensional Scaling**
Similarity analysis tool.

**Interactive Reporting**
Interactive report with Shiny and R Markdown

**Geometry of Classifiers**
Comparison of machine learning algorithms.

**Nomogram Generator**
Tune analysis then export as pdf, HTML, or Word file.

**Radiant**
Extensive app for teaching business analytics.
(documentation)

**Traveling Salesman**
Optimization fun.

**File Converter**
Upload a data file, then download in various formats.

**GDELT & BigQuery**
Dashboard to GDELT database with quarter-billion records
Movie explorer

Filter
Minimum number of reviews on Rotten Tomatoes
10  50  200

Year released
1,040  1,500  2,014

Minimum number of Oscar wins (all categories)
0  4

Dollars at Box Office (millions)
0  800

Genre (a movie can have multiple genres)
All

Director name contains (e.g., Miyazaki)

Cast names contains (e.g. Tom Hanks)

Number of movies selected:
2557

Won Oscar
- Yes
- No

To Rome with Love
2012
$16,700,000
Shiny

- Goal: exploratory or explanatory
- Speed/Flexibility: Speed
- Dependencies: R or R Studio
- Sharing: shinyapps.io, Shiny Server
- Specific Uses: interactive web applications
- Graphical Quality: good quality
- Costs: open source, freemium model
Questions?

I have some for you . . .
Questions?

• What tools do you guys use (on or off HPC)?
Questions?

- What tools do you guys use (on or off HPC)?
- What would you like to see available at NERSC?
Questions?

• What tools do you guys use (on or off HPC)?
• What would you like to see available at NERSC?
• Do you currently do data vis on HPC systems?
Questions?

- What tools do you guys use (on or off HPC)?
- What would you like to see available at NERSC?
- Do you currently do data vis on HPC systems?
- If no, why not? Are there roadblocks we can remove?
Questions?

• What tools do you guys use (on or off HPC)?
• What would you like to see available at NERSC?
• Do you currently do data vis on HPC systems?
• If no, why not? Are there roadblocks we can remove?
• Are there other issues with currently available software for data vis?