CSE 564
Visualization & Visual Analytics

Data Wrangling and Preparation

Klaus Mueller

Computer Science Department
Stony Brook University
<table>
<thead>
<tr>
<th>Lecture</th>
<th>Topic</th>
<th>Projects</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Intro, schedule, and logistics</td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>Applications of visual analytics, basic tasks, data types</td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>Data sources and introduction to D3</td>
<td>Project #1 out</td>
</tr>
<tr>
<td>4</td>
<td>Data wrangling and preparation</td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>Bias in visualization</td>
<td></td>
</tr>
<tr>
<td>6</td>
<td>Data reduction and dimension reduction</td>
<td></td>
</tr>
<tr>
<td>7</td>
<td>Visual perception and cognition</td>
<td>Project #1 due</td>
</tr>
<tr>
<td>8</td>
<td>Visual design and aesthetics</td>
<td>Project #2 out</td>
</tr>
<tr>
<td>9</td>
<td>Python/Flask hands-on</td>
<td></td>
</tr>
<tr>
<td>10</td>
<td>Cluster analysis: numerical data</td>
<td></td>
</tr>
<tr>
<td>11</td>
<td>Cluster analysis: categorical data</td>
<td></td>
</tr>
<tr>
<td>12</td>
<td>Foundations of scientific and medical visualization</td>
<td></td>
</tr>
<tr>
<td>13</td>
<td>Computer graphics and volume rendering</td>
<td>Project #2 due / Project #3 out</td>
</tr>
<tr>
<td>14</td>
<td>Scientific and medical visualization</td>
<td></td>
</tr>
<tr>
<td>15</td>
<td>Illustrative rendering</td>
<td>Project #3 due</td>
</tr>
<tr>
<td>16</td>
<td>High-dimensional data, dimensionality reduction</td>
<td>Final project proposal call out</td>
</tr>
<tr>
<td>17</td>
<td>Correlation visualization</td>
<td></td>
</tr>
<tr>
<td>18</td>
<td>Principles of interaction</td>
<td></td>
</tr>
<tr>
<td>19</td>
<td>Midterm #1</td>
<td></td>
</tr>
<tr>
<td>20</td>
<td>Visual analytics and the visual sense making process</td>
<td>Final project proposal due</td>
</tr>
<tr>
<td>21</td>
<td>Evaluation and user studies</td>
<td></td>
</tr>
<tr>
<td>22</td>
<td>Visualization of time-varying and time-series data</td>
<td></td>
</tr>
<tr>
<td>23</td>
<td>Visualization of streaming data</td>
<td></td>
</tr>
<tr>
<td>24</td>
<td>Visualization of graph data</td>
<td>Final Project preliminary report due</td>
</tr>
<tr>
<td>25</td>
<td>Visualization of text data</td>
<td></td>
</tr>
<tr>
<td>26</td>
<td>Midterm #2</td>
<td></td>
</tr>
<tr>
<td>27</td>
<td>Data journalism</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Final project presentations</td>
<td>Final Project slides and final report due</td>
</tr>
</tbody>
</table>
The variables → the attributes or properties we measured

The data items → the samples (observations) we obtained from the population of all instances

<table>
<thead>
<tr>
<th>Name</th>
<th>Country</th>
<th>Miles Per Gallon</th>
<th>Acceleration</th>
<th>Horsepower</th>
<th>Weight</th>
<th>cylinders</th>
<th>year</th>
<th>price</th>
</tr>
</thead>
<tbody>
<tr>
<td>Volkswagen Rabbit DI</td>
<td>Germany</td>
<td>43,1</td>
<td>21,5</td>
<td>48</td>
<td>1985</td>
<td>4</td>
<td>78</td>
<td>2400</td>
</tr>
<tr>
<td>Ford Fiesta</td>
<td>Germany</td>
<td>36,1</td>
<td>14,4</td>
<td>66</td>
<td>1800</td>
<td>4</td>
<td>78</td>
<td>1900</td>
</tr>
<tr>
<td>Mazda GLC Deluxe</td>
<td>Japan</td>
<td>32,8</td>
<td>19,4</td>
<td>52</td>
<td>1985</td>
<td>4</td>
<td>78</td>
<td>2200</td>
</tr>
<tr>
<td>Datsun B210 GX</td>
<td>Japan</td>
<td>39,4</td>
<td>18,6</td>
<td>70</td>
<td>2070</td>
<td>4</td>
<td>78</td>
<td>2725</td>
</tr>
<tr>
<td>Honda Civic CVCC</td>
<td>Japan</td>
<td>36,1</td>
<td>16,4</td>
<td>60</td>
<td>1800</td>
<td>4</td>
<td>78</td>
<td>2250</td>
</tr>
<tr>
<td>Oldsmobile Cutlass</td>
<td>USA</td>
<td>19,9</td>
<td>15,5</td>
<td>110</td>
<td>3365</td>
<td>8</td>
<td>78</td>
<td>3300</td>
</tr>
<tr>
<td>Dodge Diplomat</td>
<td>USA</td>
<td>19,4</td>
<td>13,2</td>
<td>140</td>
<td>3735</td>
<td>8</td>
<td>78</td>
<td>3125</td>
</tr>
<tr>
<td>Mercury Monarch</td>
<td>USA</td>
<td>20,2</td>
<td>12,8</td>
<td>139</td>
<td>3570</td>
<td>8</td>
<td>78</td>
<td>2850</td>
</tr>
<tr>
<td>Pontiac Phoenix</td>
<td>USA</td>
<td>19,2</td>
<td>19,2</td>
<td>105</td>
<td>3535</td>
<td>6</td>
<td>78</td>
<td>2800</td>
</tr>
<tr>
<td>Chevrolet Malibu</td>
<td>USA</td>
<td>20,5</td>
<td>18,2</td>
<td>95</td>
<td>3155</td>
<td>6</td>
<td>78</td>
<td>3275</td>
</tr>
<tr>
<td>Ford Fairmont A</td>
<td>USA</td>
<td>20,2</td>
<td>15,8</td>
<td>85</td>
<td>2965</td>
<td>6</td>
<td>78</td>
<td>2375</td>
</tr>
<tr>
<td>Ford Fairmont M</td>
<td>USA</td>
<td>25,1</td>
<td>15,4</td>
<td>88</td>
<td>2720</td>
<td>6</td>
<td>78</td>
<td>2275</td>
</tr>
<tr>
<td>Plymouth Volare</td>
<td>USA</td>
<td>20,5</td>
<td>17,2</td>
<td>100</td>
<td>3430</td>
<td>6</td>
<td>78</td>
<td>2700</td>
</tr>
<tr>
<td>AMC Concord</td>
<td>USA</td>
<td>19,4</td>
<td>17,2</td>
<td>90</td>
<td>3210</td>
<td>6</td>
<td>78</td>
<td>2300</td>
</tr>
<tr>
<td>Buick Century</td>
<td>USA</td>
<td>20,6</td>
<td>15,8</td>
<td>105</td>
<td>3380</td>
<td>6</td>
<td>78</td>
<td>3300</td>
</tr>
<tr>
<td>Mercury Zephyr</td>
<td>USA</td>
<td>20,8</td>
<td>16,7</td>
<td>85</td>
<td>3070</td>
<td>6</td>
<td>78</td>
<td>2425</td>
</tr>
<tr>
<td>Dodge Aspen</td>
<td>USA</td>
<td>18,6</td>
<td>18,7</td>
<td>110</td>
<td>3620</td>
<td>6</td>
<td>78</td>
<td>2700</td>
</tr>
<tr>
<td>AMC Concord D1</td>
<td>USA</td>
<td>18,1</td>
<td>15,1</td>
<td>120</td>
<td>3410</td>
<td>6</td>
<td>78</td>
<td>2425</td>
</tr>
<tr>
<td>Chevrolet MonteCarlo</td>
<td>USA</td>
<td>19,2</td>
<td>13,2</td>
<td>145</td>
<td>3425</td>
<td>8</td>
<td>78</td>
<td>3900</td>
</tr>
<tr>
<td>Buick RegalTurbo</td>
<td>USA</td>
<td>17,7</td>
<td>13,4</td>
<td>165</td>
<td>3445</td>
<td>8</td>
<td>78</td>
<td>4400</td>
</tr>
<tr>
<td>Ford Futura</td>
<td>Germany</td>
<td>18,1</td>
<td>11,2</td>
<td>139</td>
<td>3205</td>
<td>8</td>
<td>78</td>
<td>2525</td>
</tr>
<tr>
<td>Dodge Magnum XE</td>
<td>USA</td>
<td>17,5</td>
<td>13,7</td>
<td>140</td>
<td>4080</td>
<td>8</td>
<td>78</td>
<td>3000</td>
</tr>
<tr>
<td>Chevrolet Chevette</td>
<td>USA</td>
<td>30</td>
<td>16,5</td>
<td>68</td>
<td>2155</td>
<td>4</td>
<td>78</td>
<td>2100</td>
</tr>
</tbody>
</table>
Rectangular Dataset

Also called the *Data Matrix*

Car performance metrics
or Survey question responses
or Patient characteristics

One data item

Car models
or Survey respondents
or Patients

....
# Project #1: Dataset Example

Multivariate - Quantitative data and Categorical data

<table>
<thead>
<tr>
<th></th>
<th>Name</th>
<th>Country</th>
<th>Miles Per Gallon</th>
<th>Acceleration</th>
<th>Horsepower</th>
<th>weight</th>
<th>cylinders</th>
<th>year</th>
<th>price</th>
</tr>
</thead>
<tbody>
<tr>
<td>2</td>
<td>Volkswagen Rabbit DL</td>
<td>Germany</td>
<td>43.1</td>
<td>21.5</td>
<td>48</td>
<td>1985</td>
<td>4</td>
<td>78</td>
<td>2400</td>
</tr>
<tr>
<td>3</td>
<td>Ford Fiesta</td>
<td>Germany</td>
<td>36.1</td>
<td>14.4</td>
<td>66</td>
<td>1800</td>
<td>4</td>
<td>78</td>
<td>1900</td>
</tr>
<tr>
<td>4</td>
<td>Mazda GLC Deluxe</td>
<td>Japan</td>
<td>32.8</td>
<td>19.4</td>
<td>52</td>
<td>1985</td>
<td>4</td>
<td>78</td>
<td>2200</td>
</tr>
<tr>
<td>5</td>
<td>Datsun B210 GX</td>
<td>Japan</td>
<td>39.4</td>
<td>18.6</td>
<td>70</td>
<td>2070</td>
<td>4</td>
<td>78</td>
<td>2725</td>
</tr>
<tr>
<td>6</td>
<td>Honda Civic CVCC</td>
<td>Japan</td>
<td>36.1</td>
<td>16.4</td>
<td>60</td>
<td>1800</td>
<td>4</td>
<td>78</td>
<td>2250</td>
</tr>
<tr>
<td>7</td>
<td>Oldsmobile Cutlass</td>
<td>USA</td>
<td>19.9</td>
<td>15.5</td>
<td>110</td>
<td>3365</td>
<td>8</td>
<td>78</td>
<td>3300</td>
</tr>
<tr>
<td>8</td>
<td>Dodge Diplomat</td>
<td>USA</td>
<td>19.4</td>
<td>13.2</td>
<td>140</td>
<td>3735</td>
<td>8</td>
<td>78</td>
<td>3125</td>
</tr>
<tr>
<td>9</td>
<td>Mercury Monarch</td>
<td>USA</td>
<td>20.2</td>
<td>12.8</td>
<td>139</td>
<td>3570</td>
<td>8</td>
<td>78</td>
<td>2850</td>
</tr>
<tr>
<td>10</td>
<td>Pontiac Phoenix</td>
<td>USA</td>
<td>19.2</td>
<td>19.2</td>
<td>105</td>
<td>3535</td>
<td>6</td>
<td>78</td>
<td>2800</td>
</tr>
<tr>
<td>11</td>
<td>Chevrolet Malibu</td>
<td>USA</td>
<td>20.5</td>
<td>18.2</td>
<td>95</td>
<td>3155</td>
<td>6</td>
<td>78</td>
<td>3275</td>
</tr>
<tr>
<td>12</td>
<td>Ford Fairmont A</td>
<td>USA</td>
<td>20.2</td>
<td>15.8</td>
<td>85</td>
<td>2965</td>
<td>6</td>
<td>78</td>
<td>2375</td>
</tr>
<tr>
<td>13</td>
<td>Ford Fairmont M</td>
<td>USA</td>
<td>25.1</td>
<td>15.4</td>
<td>88</td>
<td>2720</td>
<td>4</td>
<td>78</td>
<td>2275</td>
</tr>
<tr>
<td>14</td>
<td>Plymouth Volare</td>
<td>USA</td>
<td>20.5</td>
<td>17.2</td>
<td>100</td>
<td>3430</td>
<td>6</td>
<td>78</td>
<td>2700</td>
</tr>
<tr>
<td>15</td>
<td>AMC Concord</td>
<td>USA</td>
<td>19.4</td>
<td>17.2</td>
<td>90</td>
<td>3210</td>
<td>6</td>
<td>78</td>
<td>2300</td>
</tr>
<tr>
<td>16</td>
<td>Buick Century</td>
<td>USA</td>
<td>20.6</td>
<td>15.8</td>
<td>105</td>
<td>3380</td>
<td>6</td>
<td>78</td>
<td>3300</td>
</tr>
<tr>
<td>17</td>
<td>Mercury Zephyr</td>
<td>USA</td>
<td>20.8</td>
<td>16.7</td>
<td>85</td>
<td>3070</td>
<td>6</td>
<td>78</td>
<td>2425</td>
</tr>
<tr>
<td>18</td>
<td>AMC Concord D1</td>
<td>USA</td>
<td>19.4</td>
<td>15.1</td>
<td>110</td>
<td>3620</td>
<td>6</td>
<td>78</td>
<td>2700</td>
</tr>
<tr>
<td>19</td>
<td>Chevrolet Monte Carlo</td>
<td>USA</td>
<td>19.2</td>
<td>13.2</td>
<td>145</td>
<td>3425</td>
<td>8</td>
<td>78</td>
<td>3900</td>
</tr>
<tr>
<td>20</td>
<td>Buick RegalTurbo</td>
<td>USA</td>
<td>17.7</td>
<td>13.4</td>
<td>165</td>
<td>3445</td>
<td>6</td>
<td>78</td>
<td>4400</td>
</tr>
<tr>
<td>21</td>
<td>Ford Futura</td>
<td>USA</td>
<td>18.1</td>
<td>11.2</td>
<td>139</td>
<td>3205</td>
<td>8</td>
<td>78</td>
<td>2525</td>
</tr>
<tr>
<td>22</td>
<td>Dodge Magnum XE</td>
<td>USA</td>
<td>17.5</td>
<td>13.7</td>
<td>140</td>
<td>4080</td>
<td>6</td>
<td>78</td>
<td>3000</td>
</tr>
<tr>
<td>23</td>
<td>Chevrolet Chevette</td>
<td>USA</td>
<td>30</td>
<td>16.5</td>
<td>68</td>
<td>2155</td>
<td>4</td>
<td>78</td>
<td>2100</td>
</tr>
<tr>
<td>24</td>
<td>Toyota Corona</td>
<td>Japan</td>
<td>27.5</td>
<td>14.2</td>
<td>95</td>
<td>2560</td>
<td>4</td>
<td>78</td>
<td>2975</td>
</tr>
</tbody>
</table>
If the data are already in a rectangular table
  - try cut and paste into Excel

If the data are on one page but cut/paste is not working
  - try a web scraper like Outwit Hub

If the data are spread across multiple webpages
  - try Outwit Hub’s automators
  - use python
  - do it by hand (probably not)
Do you think data are always clean and perfect?

Think again

Real world data are dirty

Data cleaning (wrangling)
- fill in missing values
- smooth noisy data
- identify or remove outliers
- resolve inconsistencies
- standardize/normalize data
- fuse/merge disjoint data
Data is not always available

- e.g., many tuples have no recorded value for several attributes, such as customer income in sales data

Missing data may be due to

- equipment malfunction
- inconsistent with other recorded data and thus deleted
- data not entered due to misunderstanding
- certain data may not be considered important at the time of entry
- many more reasons
Assume you get these baseball fan data

<table>
<thead>
<tr>
<th>Age</th>
<th>Income</th>
<th>Team</th>
<th>Gender</th>
</tr>
</thead>
<tbody>
<tr>
<td>23</td>
<td>24,200</td>
<td>Mets</td>
<td>M</td>
</tr>
<tr>
<td>39</td>
<td>50,245</td>
<td>Yankees</td>
<td>F</td>
</tr>
<tr>
<td>45</td>
<td>45,390</td>
<td>Yankees</td>
<td>F</td>
</tr>
<tr>
<td>22</td>
<td>32,300</td>
<td>Mets</td>
<td>M</td>
</tr>
<tr>
<td>52</td>
<td></td>
<td>Yankees</td>
<td>F</td>
</tr>
<tr>
<td>27</td>
<td>28,300</td>
<td>Mets</td>
<td>F</td>
</tr>
<tr>
<td>48</td>
<td>53,100</td>
<td>Yankees</td>
<td>M</td>
</tr>
</tbody>
</table>

How would you estimate the missing value for income (imputation)?
- ignore or put in a default value (will decimate the usable data)
- manually fill in (can be tedious or infeasible for large data)
- use the available value of the nearest neighbor’s
- average over all incomes
- average over incomes of Yankee fans
- average over incomes of female Yankees fans
- use a probabilistic method (regression, Bayesian, decision tree)
- use a neural network trained on complete data
Can help reduce influence of extreme values

See our discussion last lecture
Sometimes we like to have all variables on the same scale

- **min-max normalization**
  
  \[ v' = \frac{v - \text{min}}{\text{max} - \text{min}} \]

- **standardization / z-score normalization**
  
  \[ v' = \frac{v - \bar{v}}{\sigma_v} \]

- **clipping tails and outliers**
  - set all values beyond ± 3σ to value at 3σ
  - set values <5% and >95% to value at 5/95%
Standardization

- **Left Diagram:**
  - Mean ($m = 10.0$)
  - Standard Deviation ($s = 30.0$)

- **Right Diagram:**
  - Mean ($m = 200.0$)
  - Standard Deviation ($s = 20.0$)

- **Bottom Diagrams:**
  - **Left Diagram:**
    - Mean ($m = 0.0$)
    - Standard Deviation ($s = 1.0$)
  - **Right Diagram:**
    - Mean ($m = 0.0$)
    - Standard Deviation ($s = 1.0$)

- **Arrows:**
  - Arrows connecting the top and bottom diagrams indicate the process of standardization.

- **Comparative Text:**
  - Comparable distributions ($m = 0.0$, $s = 1.0$)
Is standardization less or more sensitive to outliers?

But you need to set a reasonable cut-off point on each side

- normal distributions are infinite
Noise = Random error in a measured variable

- faulty data collection instruments
- data entry problems
- data transmission problems
- technology limitation
- inconsistency in naming convention

Other data problems which require data cleaning

- duplicate records
- incomplete data
- inconsistent data
Binning method
- discussed last lecture

Clustering
- detect and remove outliers

Semi-automated method
- combined computer and human inspection
- detect suspicious values and check manually (need visualization)

Regression
- smooth by fitting the data to a regression function
An outlier may not be noise
- it may be an anomaly that is very valuable (e.g., the Higgs particle)
Inconsistencies in naming conventions or data codes
- e.g., 2/5/2002 could be 2 May 2002 or 5 Feb 2002

Redundant data
- duplicate tuples, which were received twice should be removed
**Data Aggregation**

Raw parallel coordinate display

Aggregate parallel coordinate display

- 4 clusters shown in different colors
- means are visualized as opaque polylines
- cluster extents are mapped to semi-transparent shapes between each axis pair
- semi-transparencies determined by linear distance from cluster center to (clipped) extent

left cluster ext. for variable i

mean

right cluster ext. for variable i
Data integration/fusion

- multiple databases
- data cubes
- files
- notes

Produces new opportunities

- can gain more comprehensive insight (value > sum of parts)
- but watch out for *synonymy and polysemy*
- attributes with different labels may have the same meaning
  - “comical” and “hilarious”
- attributes with the same label may have different meaning
  - “jaguar” can be a cat or a car
Goal is to add new thematic aspects

- enable deeper and more far-fetching insights
- can open valuable opportunities for research and $$$

How to do it

- start off with a first dataset, such as a set of listings of houses
- ask, what would house buyers be interested in?
  - education for children (school quality, pre-K, ...)
  - quality of life (entertainment, socializing, fitness, clean air, ...)
  - infrastructure (shopping, airport, roads, ...)
  - what else?
- make Google your best friend
  - determine a good attribute to link to other datasets (e.g. zip code)
  - then ask “primary education by zip code” or “livability by zip code”
### Example: Fusing Different Thematic Datasets

<table>
<thead>
<tr>
<th>Address</th>
<th>Size</th>
<th>Bedrooms</th>
<th>Baths</th>
<th>Price</th>
<th>Zip Code</th>
</tr>
</thead>
<tbody>
<tr>
<td>5 Nut Str.</td>
<td>2,345 sqft</td>
<td>3</td>
<td>1</td>
<td>$564k</td>
<td>11794</td>
</tr>
</tbody>
</table>

#### House listing data

<table>
<thead>
<tr>
<th>Zip Code</th>
<th>School Name</th>
<th>Avg. SAT</th>
<th>Class Size</th>
<th>Cost</th>
</tr>
</thead>
<tbody>
<tr>
<td>11794</td>
<td>Tree Top</td>
<td>1060</td>
<td>34</td>
<td>Public</td>
</tr>
</tbody>
</table>

#### Education by zip code

<table>
<thead>
<tr>
<th>Zip Code</th>
<th>Livability Score</th>
<th>Distance to Airport</th>
<th>Air Quality Score</th>
<th>Electricity Cost</th>
</tr>
</thead>
<tbody>
<tr>
<td>11794</td>
<td>63</td>
<td>45 miles</td>
<td>89</td>
<td>$0.34/KW</td>
</tr>
</tbody>
</table>

Make sure that all data are from the same/similar year (when time matters)

- Might need different keys for linking different thematic datasets
  - for example zip code, state, county, and so on
  - find associations for each in all tables and fuse
  - duplicate information for coarse grained tables in finer-grained tables
But Data Integration Can Also Bring Ethical Problems
Can you identify a person from these medical records?

<table>
<thead>
<tr>
<th>SSN</th>
<th>Name</th>
<th>Race</th>
<th>Date Of Birth</th>
<th>Sex</th>
<th>ZIP</th>
<th>Marital Status</th>
<th>Health Problem</th>
</tr>
</thead>
<tbody>
<tr>
<td>asian</td>
<td>9/27/64</td>
<td>female</td>
<td>94139</td>
<td>divorced</td>
<td>hypertension</td>
<td></td>
<td></td>
</tr>
<tr>
<td>asian</td>
<td>9/30/64</td>
<td>female</td>
<td>94139</td>
<td>divorced</td>
<td>obesity</td>
<td></td>
<td></td>
</tr>
<tr>
<td>asian</td>
<td>4/18/64</td>
<td>male</td>
<td>94139</td>
<td>married</td>
<td>chest pain</td>
<td></td>
<td></td>
</tr>
<tr>
<td>asian</td>
<td>4/15/64</td>
<td>male</td>
<td>94139</td>
<td>married</td>
<td>obesity</td>
<td></td>
<td></td>
</tr>
<tr>
<td>black</td>
<td>3/13/63</td>
<td>male</td>
<td>94138</td>
<td>married</td>
<td>hypertension</td>
<td></td>
<td></td>
</tr>
<tr>
<td>black</td>
<td>3/18/63</td>
<td>male</td>
<td>94138</td>
<td>married</td>
<td>shortness of breath</td>
<td></td>
<td></td>
</tr>
<tr>
<td>black</td>
<td>9/13/64</td>
<td>female</td>
<td>94141</td>
<td>married</td>
<td>shortness of breath</td>
<td></td>
<td></td>
</tr>
<tr>
<td>black</td>
<td>9/7/64</td>
<td>female</td>
<td>94141</td>
<td>married</td>
<td>obesity</td>
<td></td>
<td></td>
</tr>
<tr>
<td>white</td>
<td>5/14/61</td>
<td>male</td>
<td>94138</td>
<td>single</td>
<td>chest pain</td>
<td></td>
<td></td>
</tr>
<tr>
<td>white</td>
<td>05/08 61</td>
<td>male</td>
<td>94138</td>
<td>single</td>
<td>obesity</td>
<td></td>
<td></td>
</tr>
<tr>
<td>white</td>
<td>9/15/61</td>
<td>female</td>
<td>94142</td>
<td>widow</td>
<td>shortness of breath</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
What if you had a voter list

<table>
<thead>
<tr>
<th>Name</th>
<th>Address</th>
<th>City</th>
<th>ZIP</th>
<th>DOB</th>
<th>Sex</th>
<th>Party</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sue J. Carlson</td>
<td>900 Market St.</td>
<td>San Francisco</td>
<td>94142</td>
<td>9/15/61</td>
<td>female</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>SSN</th>
<th>Name</th>
<th>Race</th>
<th>Date Of Birth</th>
<th>Sex</th>
<th>ZIP</th>
<th>Marital Status</th>
<th>Health Problem</th>
</tr>
</thead>
<tbody>
<tr>
<td>asian</td>
<td>9/27/64</td>
<td>female</td>
<td>94139</td>
<td>divorced</td>
<td>hypertension</td>
<td></td>
<td></td>
</tr>
<tr>
<td>asian</td>
<td>9/30/64</td>
<td>female</td>
<td>94139</td>
<td>divorced</td>
<td>obesity</td>
<td></td>
<td></td>
</tr>
<tr>
<td>asian</td>
<td>4/18/64</td>
<td>male</td>
<td>94139</td>
<td>married</td>
<td>chest pain</td>
<td></td>
<td></td>
</tr>
<tr>
<td>asian</td>
<td>4/15/64</td>
<td>male</td>
<td>94139</td>
<td>married</td>
<td>obesity</td>
<td></td>
<td></td>
</tr>
<tr>
<td>black</td>
<td>3/13/63</td>
<td>male</td>
<td>94138</td>
<td>married</td>
<td>hypertension</td>
<td></td>
<td></td>
</tr>
<tr>
<td>black</td>
<td>3/18/63</td>
<td>male</td>
<td>94138</td>
<td>married</td>
<td>shortness of breath</td>
<td></td>
<td></td>
</tr>
<tr>
<td>black</td>
<td>9/13/64</td>
<td>female</td>
<td>94141</td>
<td>married</td>
<td>shortness of breath</td>
<td></td>
<td></td>
</tr>
<tr>
<td>black</td>
<td>9/7/64</td>
<td>female</td>
<td>94141</td>
<td>married</td>
<td>obesity</td>
<td></td>
<td></td>
</tr>
<tr>
<td>white</td>
<td>5/14/61</td>
<td>male</td>
<td>94138</td>
<td>single</td>
<td>chest pain</td>
<td></td>
<td></td>
</tr>
<tr>
<td>white</td>
<td>05/08 61</td>
<td>male</td>
<td>94138</td>
<td>single</td>
<td>obesity</td>
<td></td>
<td></td>
</tr>
<tr>
<td>white</td>
<td>9/15/61</td>
<td>female</td>
<td>94142</td>
<td>widow</td>
<td>shortness of breath</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Data fusion can bring insight
- the purpose is not always good
- but often it is (criminal justice, market analysis, ....)

Visualization can bring insight
- the 94142 zip code would have been an outlier
- your visualization would have shown that nicely
- then you could have dug for complementary data

How to obfuscate for protection?
- k-anonymity (generalize)
- make data less specific → use binning
- age groups, zip code groups, etc...
- make blobs instead of points
Cluster records are aggregated into k-sized bins for each variable/dimension

- Dasgupta and Kosara show this for parallel coordinates [TVCG, 2011]
- see slides for data aggregation discussed before
The Need for Data Reduction

Purpose
- reduce the data to a size that can be feasibly stored
- reduce the data so a mining algorithm can be feasibly run

Alternatives
- buy more storage
- buy more computers or faster ones
- develop more efficient algorithms (look beyond O-notation)

In practice, all of this is happening at the same time
- but the growth of data and complexities is faster
- and so data reduction is important
Data Reduction

Sampling
- random
- stratified

Data summarization
- binning (already discussed)
- clustering (see a future lecture)
- dimension reduction (see next lecture)
The goal
- pick a **representative** subset of the data

Random sampling
- pick sample points at random
- will work if the points are distributed uniformly
- this is usually not the case
- outliers will likely be missed
- so the sample will not be representative
Pick the samples according to some knowledge of the data distribution

- create a binning of some sort (outliers will form bins as well)
- also called *strata* (stratified sampling)
- the size of each bin represents its percentage in the population
- it guides the number of samples – bigger bins get more samples

**Adaptive Sampling**

sampling rate ~ cluster size
Can you “hallucinate” or “invent” realistic data?

And if so, how would you go about this?
How to Hallucinate More data...
Strategy to artificially synthesize new data from existing data

- go from small data to big data
Important topic in deep learning

Common techniques are (for images)
- rotations
- translations
- zooms
- flips
- color perturbations
- crops
- add noise by *jittering*
Definition from dictionary

- act nervously
- "an anxious student who jittered at any provocation"

- small random noise about a steady signal
Generate new samples according to the data distributions

- cluster the data (outliers will form clusters as well)
- the size of each cluster represents its percentage in the population
- randomize new samples – bigger clusters get more samples
- add a small randomized value to either the mean or an existing sample
- do this for every dimension of the chosen mean or sample

augmentation rate ~ cluster size
Also called *data narratives*

- parallel coordinates are well suited for this
- next is an example from the business world

- but first let’s review parallel coordinates in detail
Each data item is an N-dimensional vector (N variables)
- recall 2D and 3D vectors in 2D and 3D space, respectively

Now we have N-D attribute space
- the data axes extend into more than 3 orthogonal directions
- hard to imagine?
- that’s why need good visualization methods
The N=7 data axes are arranged side by side
- in parallel
Hard to see the individual cars?
  - what can we do?
Grouping the cars into sub-populations

- this is called *clustering*
- can be automated or interactive (put the user in charge)
Interactive Clustering With parallel Coordinates

Interaction in Parallel Coordinate
completely abstracted away
PC WITH ILLUSTRATIVE ABSTRACTION

blended partially
PC WITH ILLUSTRATIVE ABSTRACTION

all put together – three clusters
Our example make use of *correlation*

- correlation is a statistical measure that indicates the extent to which two or more variables fluctuate together
- a **positive correlation** indicates the extent to which those variables increase or decrease in parallel
- a **negative correlation** indicates the extent to which one variable increases as the other decreases
BACKGROUND – Anatomy of a Sales Pipeline

- Lead generator
- Lead
- Qualified lead
- Opportunity
- Opportunity+
- Opportunity++
- Opportunity+++ • Customer
- Responds
- Responds++
- Requests info (RFI)
- Requests pricing (RFP)
- Shapes deal
- Signs/buys

Potential

Prospect

Lead pool
Scene:
- a meeting of sales executives of a large corporation, Vandelay Industries

Mission:
- review the strategies of their various sales teams

Evidence:
- data of three sales teams with a couple of hundred sales people in each team
Meet Kate, a sales analyst in the meeting room:

“OK...let’s see, cost/won lead is nearby and it has a positive correlation with #opportunities but also a negative correlation with #won leads”
“Let’s go and make a revealing route!”
- she uses the mouse and designs the route shown
- she starts explaining the data like a story ...
Kate notices something else:
- now looking at the red team
- there seems to be a spread in effectiveness among the team
- the team splits into three distinct groups

She recommends: “Maybe fire the least effective group or at least retrain them”
Today’s Take Aways

How to deal with
- missing data
- noisy data and outliers
- uneven and diverse data ranges

Various strategies for segmenting data to
- visualize overall trends and groups
- reduce data
- augment/enrich data
- enable techniques to ensure privacy

Enrich datasets by adding other thematic aspects
- obtained by additional attributes from other sources
- determine proper key attributes helpful for linking

Visual storytelling with multivariate data using parallel coordinates
The course has been set up with Piazza
- [http://piazza.com/stonybrook/fall2020/cse564/home](http://piazza.com/stonybrook/fall2020/cse564/home)
- please let me know if you cannot access it

Make use of this handy discussion forum
- ask questions of general interest
- give advice to peers (those who ask questions)
- give general feedback (observe etiquette)
- but obviously, don’t provide actual solutions and aid in cheating