CSE 332
INTRODUCTION TO VISUALIZATION

DATA TYPES & BASIC APPLICATIONS

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, data, and basic tasks</td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>Introduction to D3, basic vis techniques for non-spatial data</td>
<td>Project 1 out</td>
</tr>
<tr>
<td>4</td>
<td>Visual perception and cognition</td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>Data preparation and reduction</td>
<td></td>
</tr>
<tr>
<td>6</td>
<td>Visual design and aesthetics</td>
<td></td>
</tr>
<tr>
<td>7</td>
<td>Foundations of statistics</td>
<td>Project 2 out</td>
</tr>
<tr>
<td>8</td>
<td>Data types, notion of similarity and distance</td>
<td></td>
</tr>
<tr>
<td>9</td>
<td>Data mining techniques: clusters, text, patterns, classifiers</td>
<td></td>
</tr>
<tr>
<td>10</td>
<td>Data mining techniques: clusters, text, patterns, classifiers</td>
<td></td>
</tr>
<tr>
<td>11</td>
<td>High-dimensional data, dimensionality reduction</td>
<td></td>
</tr>
<tr>
<td>12</td>
<td>Computer graphics and volume rendering</td>
<td>Project 3 out</td>
</tr>
<tr>
<td>13</td>
<td>Techniques to visualize spatial (3D) data</td>
<td></td>
</tr>
<tr>
<td>14</td>
<td>Scientific and medical visualization</td>
<td></td>
</tr>
<tr>
<td>15</td>
<td>Scientific and medical visualization</td>
<td></td>
</tr>
<tr>
<td>16</td>
<td>Non-photorealistic rendering</td>
<td></td>
</tr>
<tr>
<td>17</td>
<td>Midterm</td>
<td></td>
</tr>
<tr>
<td>18</td>
<td>Principles of interaction</td>
<td>Project 4 out</td>
</tr>
<tr>
<td>19</td>
<td>Visual analytics and the visual sense making process</td>
<td></td>
</tr>
<tr>
<td>20</td>
<td>Correlation and causal modeling</td>
<td></td>
</tr>
<tr>
<td>21</td>
<td>Big data: data reduction, summarization</td>
<td></td>
</tr>
<tr>
<td>22</td>
<td>Visualization of graphs and hierarchies</td>
<td></td>
</tr>
<tr>
<td>23</td>
<td>Visualization of text data</td>
<td>Project 5 out</td>
</tr>
<tr>
<td>24</td>
<td>Visualization of time-varying and time-series data</td>
<td></td>
</tr>
<tr>
<td>25</td>
<td>Memorable visualizations, visual embellishments</td>
<td></td>
</tr>
<tr>
<td>26</td>
<td>Evaluation and user studies</td>
<td></td>
</tr>
<tr>
<td>27</td>
<td>Narrative visualization and storytelling</td>
<td></td>
</tr>
<tr>
<td>28</td>
<td>Data journalism</td>
<td></td>
</tr>
</tbody>
</table>
Data Types Every CS Person Knows

- Primitive
  - char
  - int
  - float
  - double
  - void

- Derived
  - Array
  - Pointer
  - Function

- User defined
  - enum
  - Structure
  - Union
Data Types in Visual Analytics

Numeric
Categorical
Text
Time series
Graphs and networks
Hierarchies
**Variables in Statistics**

**Numeric variables**
- measure a **quantity** as a number
- like: ‘how many' or 'how much'
- can be continuous (grey curve)
- or discrete (red steps)

**Categorical variables**
- describe a **quality** or characteristic
- like: 'what type' or 'which category'
- can be ordinal = ordered, ranked (distances need not be equal)
  - clothing size, academic grades, levels of agreement
- or nominal = not organized into a logical sequence
  - gender, business type, eye color, brand
Most often the x-axis is ‘time’

- provides an intuitive & innate ordering of the data values
- the majority of people expect the x-axis to be ‘time’

But ‘time’ is not the only option
- engineers, statisticians, etc. will be receptive to this idea
- can you think of an example?
Another plot where ‘time’ is not the x-axis

- from the engineering / physics domain
- in some sense, it tells a story
CATEGORICAL VARIABLES

Usually plotted as bar charts or pie charts

nominal

ordinal

Customer Satisfaction
But not everything is expressed in numbers

- images
- video
- text
- web logs
- ...

Need to do **feature analysis** to turn these abstract things into numbers

- then apply your analysis as usual
- but keep the reference to the original data so you can return to the native domain where the analysis problem originated
Sensor Data

Characteristics
- often large scale
- time series

Feature Analysis
- example: Motif discovery
- encode into 5D data vector

[0.12, 0.3, 0.41, 0.12, 0.02]
Characteristics

- array of pixels

Feature Analysis

- example: value histograms
- encode into a 256-D vector

[0, 0, 0, ...., 10, ..., 1200, .....]
VIDEO DATA

Characteristics
- essentially a time series of images

Feature Analysis
- many of the image techniques apply but extension is non-trivial
Characteristics
  - often raw and unstructured

Feature analysis
  - first step is to remove stop words and stem the data
  - perform **named-entity recognition** to gain atomic elements
    - identify names, locations, actions, numeric quantities, relations
    - understand the structure of the sentence and complex events
  - example:
    - Jim bought 300 shares of Acme Corp. in 2006.
    - \([\text{Jim}]_{\text{Person}}\) bought \([300 \text{ shares}]_{\text{Quantity}}\) of \([\text{Acme Corp.}]_{\text{Organiz.}}\) in \([2006]_{\text{Time}}\)
  - distinguish between
    - application of grammar rules (old style, need experienced linguists)
    - statistical models (Google etc., need big data to build)
Create a term-document matrix

- turns text into a high-dimensional vector which can be compared
- use Latent Semantic Analysis (LSA) to derive a visualization
Maps the frequency of words in a corpus to size

https://www.jasondavies.com/wordcloud/
Weblogs

- typically represented as text strings in a pre-specified format
- this makes it easy to convert them into multidimensional representation of categorical and numeric attributes

Network traffic

- characteristics of the network packets are used to analyze intrusions or other interesting activity
- a variety of features may be extracted from these packets
  - the number of bytes transferred
  - the network protocol used
  - IP ports used
let’s look at some essential graphical representations

and do some advertising for D3
Stakeholder Hierarchy

- Stakeholders
  - Customers
  - Others
    - Procurers
    - Users
      - Favored User Classes
      - Disfavored User Classes
      - Ignored User Classes
      - Other User Classes
FUNCTION CALL TREE

- gprof_function_call_tree_view_get_type
  - gprof_function_call_tree_view_init
  - gprof_function_call_tree_view_create_columns
  - on_list_view_row_activated
  - gprof_view_show_symbol_in_editor
- gprof_function_call_tree_view_class_init
  - gprof_function_call_tree_view_refresh
  - gprof_function_call_tree_view_finalized
  - gprof_function_call_tree_view_add_function
- gprof_call_graph_block_get_first_block
  - gprof_call_graph_block_get_next
- gprof_call_graph_block_is_recursive
  - gprof_call_graph_block_get_first_child
- gprof_call_graph_find_block
  - gprof_call_graph_block_entry_get_next
- gprof_call_graph_block_entry_get_name
  - gprof_call_graph_block_get_primary_entry
- gprof_profile_data_get_call_graph
  - gprof_view_get_data
- gprof_view_get_data
  - gprof_profile_data_get_call_graph
  - gprof_function_call_tree_view_add_function
Hierarchies

Questions you might have

- how large is each group of stakeholders (or function)?
  - tree with quantities
- what fraction is each group with respect to the entire group?
  - partition of unity
- how is information disseminated among the stakeholders (or functions)?
  - information flow
- how close (or distant) are the individual stakeholders (functions) in terms of some metric?
  - force directed layout
More scalable tree, and natural with some randomness

http://animateddata.co.uk/lab/d3-tree/
A standard tree, but one that is scalable to large hierarchies

A tree that is scalable and has partial partition of unity

More space efficient since it’s radial, has partial partition of unity

https://www.jasondavies.com/coffee-wheel/

http://bl.ocks.org/kerryrodden/7090426
Bubble Charts

No hierarchy information, just quantities

http://bl.ocks.org/mbostock/4063269
Quantities and containment, but not partition of unity

Quantities, containment, and full partition of unity

Chord Diagram

Relationships among group fractions, not necessarily a tree

http://bl.ocks.org/mbostock/4062006
Hierarchical Edge Bundling

Relationships of individual group members, also in terms of quantitative measures such as information flow

Collapsible Force Layout

Relationships within organization members expressed as distance and proximity

Voronoi Tessellation

Shows the closest point on the plane for a given set of points... and a new point via interaction

http://bl.ocks.org/mbostock/4060366
Data Type Conversions and Transformation
Solution 1:
- divide the numeric attribute values into ϕ \textbf{equi-width} ranges
- each range/bucket has the same width
- example: customer age

- what is lost here?
- age ranges of customers could be unevenly distributed within a bin
Solution 2:

- divide the numeric attribute values into φ equi-depth ranges
- same number of samples in each bin
- (again) example: customer age:

  ![Histogram example](image)

  - what is the disadvantage here?
  - extra storage needed: must store the start/end value for each bin
Solution 3:

- what if all the bars have seemingly height
- or are dominated by one large peak

- switch to log scaling of the y-value
Other Transformations

Dang and Wilkinson, “Transforming Scagnostics to Reveal Hidden Features”, TVCG 2014

- none: \( x^* = x \) (leaves points unchanged)
- half: \( x^* = x/2 \) (squeezes all points together)
- square: \( x^* = x^2 \) (pulls points toward left of frame)
- square root: \( x^* = \sqrt{x} \) (mildly pulls points toward right of frame)
- log: \( x^* = \log(x) \) (strongly pulls points toward right of frame)
- inverse: \( x^* = 1/x \) (reverses scale and squeezes points into left of frame)
- logit: \( x^* = (\log(x/(1-x)) + 10)/20 \) (squeezes points toward middle of frame)
- sigmoid: \( x^* = 1/(1 + \exp(-20x + 10)) \) (expands points away from middle of frame)
Why discrete?

- because we can’t store continuous data
- we can only store samples of the continuous data
- how many samples do we need?
- also keep this in mind for data reduction
Why discrete?

- because we can’t store continuous data
- we can only store samples of the continuous data
- how many samples do we need?
Why discrete?

- because we can’t store continuous data
- we can only store samples of the continuous data
- how many samples do we need?
Why discrete?

- because we can’t store continuous data
- we can only store samples of the continuous data
- how many samples do we need?

We need a certain number of samples to represent a continuous phenomenon

- twice as many samples as the highest frequency in the signal
called the Nyquist frequency

- else we get aliasing
Ever tried to reduce the size of an image and you got this?

This is aliasing
But what you really wanted is this:

This is *anti-aliasing*.
**Why Is This Happening?**

The smaller image resolution cannot represent the image detail captured at the higher resolution

- skipping this small detail leads to these undesired artifacts
Procedure

- either sample at a higher rate
- or smooth the signal before sampling it
- the latter is called *filtering*
Anti-Aliasing Via Smoothing
ANTI-ALIASING VIA SMOOTHING
What is Smoothing?

Slide a window across the signal

- stop at each discrete sample point
- average the original data points that fall into the window
- store this average value at the sample point
- move the window to the next sample point
- repeat
What is the filter we just used called?

- it’s called a *box filter*

There are other filters

- for example, Gaussian filter
- yields a smoother result
- box filtering is simplest
Can you see some patterns?

It’s another form of aliasing
Filtering also eliminates noise in the data
In some ways, bar charts reduce noise and uncertainties in the data

- the bins do the smoothing

Example:

- obesity over age (group)
Of course, bar charts can also hold categorical data.
BAR CHARTS IN D3

http://bl.ocks.org/mbostock/3885304

Working with bar charts will be your job for Lab 2

- the next two slides offer some help with calculations
BAR CHART CALCULATIONS – BINNING

Determine bin size

- min(data) is optional, can also use 0 or some reasonable value
- max(data) is optional, can also use some reasonable value

\[
bin \ size = \frac{\text{max}(data) - \text{min}(data)}{\text{number of bins}}
\]

Given a data value \( \text{val} \) increment (++) the bin value

- but first initialize bin val array to 0

\[
\text{bin val array} \left[ \left\lfloor \frac{\text{val} - \text{min}(data)}{\text{bin size}} \right\rfloor \right] + +
\]
Determine bin size on the screen

\[
\text{bin size on screen} = \frac{\text{chart width}}{\text{number of bins}}
\]

Center of a bar for bin with index \textit{bin index}

\[
\text{bar center on screen} = (\text{bin index} \cdot \text{bin size on screen}) + 0.5
\]

Height of the bar for a bin with index \textit{bin index}

\[
\text{bar height(\text{bin index})} = \text{bin val array(\text{bin index})} \cdot \frac{\text{chart height}}{\text{max(bin val array)}}
\]

Do not forget that the origin of a web page is the top left corner
PROJECT #1

Find some interesting data on the web

- something that challenges and interests you
- there are many data sources on the web
- use google and some imagination

Criteria for selection

- more than 500 data points (observations)
- more than 10 attributes
- the more the better (you can always reduce it)

Deliverables

- 2-page report that describes the data and justifies your choice
- a URL to the data source

Due date

- Tuesday, September 20, 11:59pm
# 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>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>19</td>
<td>AMC Concord D1</td>
<td>USA</td>
<td>18.1</td>
<td>15.1</td>
<td>120</td>
<td>3410</td>
<td>8</td>
<td>78</td>
<td>2425</td>
</tr>
<tr>
<td>20</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>21</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>22</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>23</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>24</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>25</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>

**Data types**
- Quantitative (Numerical)
- Categorical (Ordinal)
Other data types are OK

- text, images, video, logs, etc.
- just convert them to numbers via appropriate mechanism as discussed in class
- must produce a spreadsheet of rows (data items) and attributes (columns)

Categorical data

- color, brand, country, etc.
- convert into numbers by assigning a numerical ID
The course has been set up with Piazza
- 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