CSE 564
Visualization & Visual Analytics
Applications and Basic Tasks

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>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>Visual design and aesthetics</td>
<td></td>
</tr>
<tr>
<td>6</td>
<td>Data types, notion of similarity and distance</td>
<td></td>
</tr>
<tr>
<td>7</td>
<td>Data preparation and reduction</td>
<td>Project #1 due</td>
</tr>
<tr>
<td>8</td>
<td>Introduction to R, statistics foundations</td>
<td>Project #2 out</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>Computer graphics and volume rendering</td>
<td></td>
</tr>
<tr>
<td>12</td>
<td>Techniques to visualize spatial (3D) data</td>
<td>Project #2 due</td>
</tr>
<tr>
<td>13</td>
<td>Scientific and medical visualization</td>
<td>Project #3 out</td>
</tr>
<tr>
<td>14</td>
<td>Scientific and medical visualization</td>
<td></td>
</tr>
<tr>
<td>15</td>
<td>Midterm #1</td>
<td></td>
</tr>
<tr>
<td>16</td>
<td>High-dimensional data, dimensionality reduction</td>
<td>Project #3 due</td>
</tr>
<tr>
<td>17</td>
<td>Big data: data reduction, summarization</td>
<td></td>
</tr>
<tr>
<td>18</td>
<td>Correlation and causal modeling</td>
<td></td>
</tr>
<tr>
<td>19</td>
<td>Principles of interaction</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>
Data Types Every CS Person Knows
DATA TYPES IN VISUAL ANALYTICS

Numeric
Categorical
Text
Time series
Graphs and networks
Hierarchies
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
Numbers are Good

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
Characteristics
- often large scale
- time series

Feature Analysis
- Fourier transform (FT, FFT)
- Wavelet transform (WT, FWT)
- Motif discovery
Characteristics

- array of pixels

Feature Analysis

- histograms
  - values
  - gradients
- FFT, FWT
- Scale Invariant Feature Transform (SIFT)
- Bag of Features (BoF)
- visual words

histograms
VIDEO DATA

Characteristics
- essentially a time series of images

Feature Analysis
- many of the above techniques apply albeit extension is non-trivial
BAG OF FEATURES (BOF)
1. Obtain the set of bags of features
   (i) Select a large set of images
   (ii) Extract the SIFT feature points of all the images in the set and obtain the SIFT descriptor for each feature point extracted from each image
   (iii) Cluster the set of feature descriptors for the amount of bags we defined and train the bags with clustered feature descriptors
   (iv) Obtain the visual vocabulary

2. Obtain the BoF descriptor for a given image/video frame
   (v) Extract SIFT feature points of the given image
   (vi) Obtain SIFT descriptor for each feature point
   (vii) Match the feature descriptors with the vocabulary we created in the first step
   (viii) Build the histogram
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
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.
    - [Jim]Person bought [300 shares]Quantity of [Acme Corp.]Organiz. in [2006]Time
  - distinguish between
    - application of grammar rules (old style, need experienced linguists)
    - statistical models (Google etc., need big data to build)
let's look at some essential graphical representations

and do some advertising for D3
Example: Stakeholders of a library system
Example: Stakeholders of a library system

Questions you might have

- how large is each stakeholder group?
  - tree with quantities
- what fraction is each group with respect to the entire group of stakeholders?
  - partition of unity
- how is information disseminated among the stakeholders?
  - information flow
- how close (or distant) are the individual stakeholders in terms of some metric?
  - force directed layout
Invoke Nature

More scalable tree, and natural with some randomness

http://animateddata.co.uk/lab/d3-tree/
COLLAPSIBLE 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

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

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 Types in Visual Analytics

<table>
<thead>
<tr>
<th>Source data type</th>
</tr>
</thead>
<tbody>
<tr>
<td>Numeric</td>
</tr>
<tr>
<td>Categorical</td>
</tr>
<tr>
<td>Text</td>
</tr>
<tr>
<td>Time series</td>
</tr>
<tr>
<td>Time series</td>
</tr>
<tr>
<td>Discrete sequence</td>
</tr>
<tr>
<td>Spatial</td>
</tr>
<tr>
<td>Graphs</td>
</tr>
<tr>
<td>Any type</td>
</tr>
</tbody>
</table>

which you may need to convert into using these methods
Solution 1:

- divide the numeric attribute values into $\varphi$ equi-width ranges
- each range has the same width
- example: customer age

- what is lost here?
- age ranges of customer could be unevenly distributed
Solution 2:

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

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
Other Transformations
World Bank data:
Rural vs. Urban population

Each point is a country

Transformations to emphasize outliers

An example of “how to lie with visualizations”

So rather not do this excessively or to simply make a point that isn’t really there

Dang and Wilkinson,
“Transforming Scagnostics to Reveal Hidden Features”, TVCG 2014
World Bank data: Rural vs. Urban population

Each point is a country

Transformations to visualize growth patterns

Again, be careful not to overemphasize a point that isn’t yours to make

Generally accepted is mostly log-scale or log-log scale (for 2D)

Dang and Wilkinson, “Transforming Scagnostics to Reveal Hidden Features”, TVCG 2014
If a categorical attribute has $\varphi$ different values, then $\varphi$ different binary attributes are created.

This can have disadvantages:

- ordinal data are not spaced by actual distance, whatever the metric
- nominal data have no order and spacing at all
- there are techniques based on correlation (stay tuned)
Use Latent Semantic Analysis (LSA)

- also known as Latent Semantic Indexing (LSI)
- turns text into a high-dimensional vector which can be compared
Word Cloud

Maps the frequency of words in a corpus to size

https://www.jasondavies.com/wordcloud/
Sequence Data

Time series to discrete sequence data

- window based averaging
  - eliminates noise in the data
- value-based discretization
  - collapse series into a smaller number of intervals
  - usually done after averaging

Time series to numeric data

- FFT, DFT
- also works for discrete series and spatial data
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

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

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

- but first initialize bin val array to 0

\[
\text{bin val array} \left[ \left\lfloor \frac{val - \text{min(data)}}{\text{bin size}} \right\rfloor \right] + +
\]
**Bar Chart Calculations – Plotting**

Determine bin size on the screen

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

Center of a bar for bin with index *bin index*

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

Height of the bar for a bin with index *bin index*

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

Do not forget that the origin of a web page is the top left corner
Get some CSV-based data from [here](#) or [here](#) or elsewhere

Your D3-based visual interface should be able to (all 10 pts):
1. pick a variable and bin it into a fixed range (equi-width) of your choice
2. create a bar chart of the variable you picked in 1.
3. respond to mouse clicks to cycle through all variables and update chart
4. only on mouse over display the value of the bar on top of the bar
5. on mouse over make the bar wider and higher to focus on it
6. on mouse click transform the bar chart into a pie chart (and back)
7. on mouse click create a force-directed layout using a chosen distance

An additional 10 pts for elegant implementation/function

Extra credit (10 pts):
- mouse moves left (right) should decrease (increase) bin width/size

Submission on Blackboard – a report and a URL to the page
Due Thursday, Feb. 18, 2016 (see note below)

We will provide a VM on the department web server where your code can be run public-access