<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>Basic vis techniques for non-spatial data</td>
<td>Project 1 out</td>
</tr>
<tr>
<td>4</td>
<td>Data preparation and reduction</td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>Perception and cognition, visual design and aesthetics</td>
<td></td>
</tr>
<tr>
<td>6</td>
<td>Foundations of statistics</td>
<td></td>
</tr>
<tr>
<td>7</td>
<td>Introduction to D3</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

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

**Numerical 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
Numerical Variables

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

Number of Colors in Bag of M&M Candies

Customer Satisfaction

Nominal

Ordinal
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
**Image Data**

**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
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
Train a shallow neural network (NN) on a corpus of text

- the NN weight vectors encode word similarity as a high-D vector
- use a 2D embedding technique to display
**Word Embedding Algebra**

Load up the word vectors

- **QUEEN** [0.3, 0.9]
- **KING** [0.5, 0.7]
- **WOMAN** [0.3, 0.4]
- **MAN** [0.5, 0.2]

**gender** = WOMAN – MAN

**QUEEN** = KING + gender

**QUEEN** = KING – MAN + WOMAN
Maps the frequency of words in a corpus to size

https://www.jasondavies.com/wordcloud/
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
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
Invoke Nature

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/

https://www.data-to-viz.com/graph/sunburst.html

https://observablehq.com/@kerryrodden/sequences-sunburst
Bubble Charts

No hierarchy information, just quantities

https://observablehq.com/@d3/bubble-chart
Quantities and containment, but not partition of unity

Quantities, containment, and full partition of unity

Relationships among group fractions, not necessarily a tree

https://observablehq.com/@d3/chord-diagram
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

https://observablehq.com/@mbostock/voronoi-particles
Data Type Conversions and Transformation
Solution 1:
- divide the numeric attribute values into $\varphi$ equi-width ranges
- each range/bucket has the same width
- example: customer age

- what is lost here?
Problem With Equi-Width Histogram

Age ranges of customers could be unevenly distributed within a bin
- this could be an interesting anomaly
Solution 2:

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

```
+-----+-----+-----+-----+-----+-----+
| 0-22| 22-31| 32-38| 38-44| 44-48| 48-55|
+-----+-----+-----+-----+-----+-----+
| 36   | 36   | 36   | 36   | 36   | 36   |
+-----+-----+-----+-----+-----+-----+
|      |      |      |      |      |      |
+-----+-----+-----+-----+-----+-----+
```

- 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
Data Representation
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.
What's the underlying problem?

- detail can't be refined upon zoom
- can just be replicated or blurred

The solution...

- represent detail as a function that can be mathematically refined
- replace raster graphics by vector graphics
Scalable Vector Graphics (SVG)
Vector graphics tends to have an “cartoonish” look

raster graphics

vector graphics
Photographs and Images in SVG
D3 Uses SVG

The Wealth & Health of Nations

42.0%
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 vs. Histograms

Histograms
- bars show the frequency of numerical data
- quantitative data
- elements are grouped together, so that they are considered as ranges
- bars cannot be reordered
- width of bars need not be the same

Bar charts
- uses bars to compare different categories of data
- comparison of discrete variables
- elements are taken as individual entities
- bars can be reordered
- width of bars need to be the same
How many bars are too many (in a chart)
- if individual categories are the focus? 12 is a good rule
- if the overall trend is the important factor? 50 or even more
- eventually you can switch to a line chart

- sort bars by height and use ‘other’ to aggregate the bar chart tails into a single bar
- find a grouping that can semantically aggregate bars, for example aggregate countries into continents

more information
Bar Charts in D3

https://observablehq.com/@d3/bar-chart

Working with bar charts and histograms is the topic of Lab 2
- the next two slides offer some help with calculations
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

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

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}

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

Do not forget that the origin of a web page is the top left corner.
D3 – Data Driven Documents (we will use for this course)
  • creates interactive webpages from data
  • lots of creations are here

Vega (see here)
  • higher-level visualization specification language on top of D3
  • D3 is still more “expressive” and allows for more creative freedom

Vega-Lite (see here)
  • a high-level grammar of interactive graphics
  • built on top of Vega
  • more concise & convenient form to author common visualizations
  • supports data analytics (both data and visual transformations)
  • better support for interactions
Tableau is a leading commercial visual analytics platform

- founded in 2003 by a group of Stanford University researchers (Chris Stolte, Pat Hanrahan, and Christian Chabot)
- recently acquired by Salesforce
- goal was to make data more accessible through visualization
- key tech was VizQL – visualizes data by translating drag-and-drop actions into data queries through an intuitive interface
EXAMPLE TABLEAU DASHBOARDS

Account tracking
Quarterly results
Top accounts

Opportunity overview
Opportunity tracking
Marketing leads
Essentially, **Tableau** is great for expediently-developed in-house use. **D3** is better for external use, real-time interactive web, and embedding into a product.