CSE 412 - Intro to Data Visualization

Course Summary

Jane Hoffswell  University of Washington
Final Project Deliverables

Demonstration Video (<= 3.5 min)
Due on YouTube & Canvas by midnight Tue 6/1. 
*Be sure to submit the video on time!*

Final Project Showcase
We will show demo videos in class, Wed & Fri.

Interactive Web Page & GitHub Repo
All materials online by midnight Mon 6/7.

Read assignment description for more!
Final Project Video

First frame of video should include: project name, team members' names, link to your website

Communicate topics and project goals
Do: Show what viewers can learn from your page
Don't: Enumerate every feature of the page

Think about overall production style/quality
Include music to set the tone; check the sound quality on the recording; focus on the narrative!

For other tips, see the video production guide!
Final Project Considerations

Overall narrative and flow: does the page present an interesting and coherent story with the text and visualizations for a general audience?

Breadth and depth: does the page include a compelling and thorough exploration of the topic?

Visualization design: are all visualization designs and details expressive and effective?

Interaction and animation: are the interactions interesting and effective, or do they unnecessarily hide information and complicate the view?
Course Summary
Value of Visualization

Anscombe's Quartet [Anscombe 73]
### Summary Statistics

**Set A**

<table>
<thead>
<tr>
<th>X</th>
<th>Y</th>
</tr>
</thead>
<tbody>
<tr>
<td>10</td>
<td>8.04</td>
</tr>
<tr>
<td>8</td>
<td>6.95</td>
</tr>
<tr>
<td>13</td>
<td>7.58</td>
</tr>
<tr>
<td>9</td>
<td>8.81</td>
</tr>
<tr>
<td>11</td>
<td>8.33</td>
</tr>
<tr>
<td>14</td>
<td>9.96</td>
</tr>
<tr>
<td>6</td>
<td>7.24</td>
</tr>
<tr>
<td>4</td>
<td>4.26</td>
</tr>
<tr>
<td>12</td>
<td>10.84</td>
</tr>
<tr>
<td>7</td>
<td>4.82</td>
</tr>
<tr>
<td>5</td>
<td>5.68</td>
</tr>
</tbody>
</table>

**Set B**

<table>
<thead>
<tr>
<th>X</th>
<th>Y</th>
</tr>
</thead>
<tbody>
<tr>
<td>10</td>
<td>9.14</td>
</tr>
<tr>
<td>8</td>
<td>8.14</td>
</tr>
<tr>
<td>13</td>
<td>8.74</td>
</tr>
<tr>
<td>9</td>
<td>8.77</td>
</tr>
<tr>
<td>11</td>
<td>9.26</td>
</tr>
<tr>
<td>14</td>
<td>8.1</td>
</tr>
<tr>
<td>6</td>
<td>6.13</td>
</tr>
<tr>
<td>4</td>
<td>3.1</td>
</tr>
<tr>
<td>12</td>
<td>9.11</td>
</tr>
<tr>
<td>7</td>
<td>7.26</td>
</tr>
<tr>
<td>5</td>
<td>4.74</td>
</tr>
</tbody>
</table>

**Set C**

<table>
<thead>
<tr>
<th>X</th>
<th>Y</th>
</tr>
</thead>
<tbody>
<tr>
<td>10</td>
<td>7.46</td>
</tr>
<tr>
<td>8</td>
<td>6.77</td>
</tr>
<tr>
<td>13</td>
<td>12.74</td>
</tr>
<tr>
<td>9</td>
<td>7.11</td>
</tr>
<tr>
<td>11</td>
<td>7.81</td>
</tr>
<tr>
<td>14</td>
<td>8.84</td>
</tr>
<tr>
<td>6</td>
<td>6.08</td>
</tr>
<tr>
<td>4</td>
<td>5.39</td>
</tr>
<tr>
<td>12</td>
<td>8.15</td>
</tr>
<tr>
<td>7</td>
<td>6.42</td>
</tr>
<tr>
<td>5</td>
<td>5.73</td>
</tr>
</tbody>
</table>

**Set D**

<table>
<thead>
<tr>
<th>X</th>
<th>Y</th>
</tr>
</thead>
<tbody>
<tr>
<td>8</td>
<td>6.58</td>
</tr>
<tr>
<td>8</td>
<td>5.76</td>
</tr>
<tr>
<td>8</td>
<td>7.71</td>
</tr>
<tr>
<td>8</td>
<td>8.84</td>
</tr>
<tr>
<td>8</td>
<td>8.47</td>
</tr>
<tr>
<td>8</td>
<td>7.04</td>
</tr>
<tr>
<td>8</td>
<td>5.25</td>
</tr>
<tr>
<td>19</td>
<td>12.5</td>
</tr>
<tr>
<td>8</td>
<td>5.56</td>
</tr>
<tr>
<td>8</td>
<td>7.91</td>
</tr>
<tr>
<td>8</td>
<td>6.89</td>
</tr>
</tbody>
</table>

### Linear Regression

**Set A**

- $u_X = 9.0$
- $\sigma_X = 3.317$
- $u_Y = 7.5$
- $\sigma_Y = 2.03$

**Set B**

- $Y = 3 + 0.5X$
- $R^2 = 0.67$

[Anscombe 73]
The Value of Visualization

**Record** information
- Blueprints, photographs, seismographs, ...

**Analyze** data to support reasoning
- Develop and assess hypotheses
- Find patterns / Discover errors in data
- Expand memory

**Convey** information
- Communicate, inform, inspire
- Collaborate and revise
Data and Image Models

Sémiologie Graphique [Bertin 67]
Nominal, Ordinal & Quantitative

N - Nominal (labels or categories) $\rightarrow =, \neq$
  - Fruits: apples, oranges, ...

O - Ordered $\rightarrow =, \neq, <, >$
  - Quality of meat: Grade A, AA, AAA

Q - Interval (location of zero arbitrary) $\rightarrow =, \neq, <, >, -$
  - Dates: Jan, 19, 2006; Location: (LAT 33.98, LONG -118.45)
  - Only differences (i.e., intervals) may be compared
  - Can measure distances or spans

Q - Ratio (zero fixed) $\rightarrow =, \neq, <, >, -, \%$
  - Physical measurement: Length, Mass, Time duration, ...
  - Counts and amounts
  - Can measure ratios or proportions
Dimensions & Measures

**Dimensions** (~ independent variables)
Often discrete variables describing data (N, O)
Categories, dates, binned quantities

**Measures** (~ dependent variables)
Data values that can be aggregated (Q)
Numbers to be analyzed
Aggregate as sum, count, avg, std. dev…

Not a strict distinction. The same variable may be treated either way depending on the task.
Design Criteria [Mackinlay 86]

Expressiveness
A set of facts is *expressible* in a visual language if the sentences (i.e. the visualizations) in the language express all the facts in the set of data, and only the facts in the data.

Effectiveness
A visualization is more *effective* than another visualization if the information conveyed by one visualization is more readily perceived than the information in the other visualization.
Design Criteria *Translated*

Tell the truth and nothing but the truth (don’t lie, and don’t lie by omission)

Use encodings that people decode better (where better = faster and/or more accurate)
Conjectured effectiveness of encodings by data type
Data Quality

“The first sign that a visualization is good is that it shows you a problem in your data...

...every successful visualization that I've been involved with has had this stage where you realize, "Oh my God, this data is not what I thought it would be!" So already, you've discovered something.”

Martin Wattenberg
Exploratory Data Analysis Lessons

Check **data quality** and your **assumptions**.

Start with **univariate summaries**, then start to consider **relationships among variables**.

**Avoid premature fixation!**

**Transform data** appropriately (e.g., invert, log)

**Show data variation, not design variation** [Tufte]
Visual Encoding and Design

SlicerDicers' Sales Compared to Other Products

- AhHuts
- NervousNellies
- RingaDingies
- RoundTuits
- SlicerDicers
- SweetNuthins
- ThingamaGigs
- Whatchamacallits
- WileyWidgets

Problematic design

Redesign
Artery Visualization [Borkin et al ’11]

Rainbow Palette

2D

Diverging Palette

3D

62% 92%

39% 71%
About the design process...

Visualization draws upon both science and art! Principles like expressiveness & effectiveness are not hard-and-fast rules, but can assist us to guide the process and articulate alternatives. They can lead us to think more deeply about our design rationale and prompt us to reflect.

It helps to know “the rules” in order to wisely bend (or break) them at the right times!
Responsive Visualization

In close decisions, Kennedy voted in the majority 76 percent of the time.

Justice Kennedy has been an essential figure in forming a majority on the court. He voted on the winning side of close decisions 76 percent of the time over his career, far more often than any other justice he served with except for Justice Neil Gorsuch, who joined the court last year.

Justices Kennedy joined both the conservative and liberal blocs of justices to

Techniques for Responsive Visualization Design [Hoffswell 20]
Responsive Visualization Summary

Good visualizations are task dependent
Who is the audience and what is the task?
Pick the right interaction technique

Visualizations are not one size fits all
Context might change user goals
Recent elections have placed a heavy emphasis on “swing states” — Ohio, Florida and the other competitive states. You can see how many states shifted between the Democratic and Republican parties. A look at how the states shifted over past elections.

Obama Re-elected
The country voted about 5 percentage points more Republican in 2012 than in 2008. Obama lost North Carolina and Indiana, but won every tossup except Florida, which remains too close to call.

As Goes Ohio
Ohio, which has voted for the winner in every election since 1964, provided the decisive electoral votes in 2004, and it is the state likeliest to play that role again this year, according to the FiveThirtyEight model.
Narrative Storytelling

**narrative** (n): An account of a series of events, facts, etc., given in order and with the establishing of connections between them.

Effective storytelling “require[s] skills like those familiar to movie directors, beyond a technical expert’s knowledge of computer engineering and science.”

- Gershon & Page ’01
Gulf of Execution
The difference between the user’s intentions and the allowable actions.

Gulf of Evaluation
The amount of effort that the person must exert to interpret the state of the system and to determine how well the expectations and intentions have been met.

[Norman 1986]
Taxonomy of Interactions

Data and View Specification
Visualize, Filter, Sort, Derive

View Manipulation
Select, Navigate, Coordinate, Organize

Process and Provenance
Record, Annotate, Share, Guide
Interaction Takeaways

Most visualizations are interactive
Even passive media elicit interactions

Good visualizations are task dependent
Pick the right interaction technique
Consider the semantics of the data domain

Fundamental interaction techniques
Selection / Annotation, Sorting, Navigation,
Brushing & Linking, Dynamic Queries
Animation

Animated transitions in statistical data graphics [Heer & Robertson 07]
Tracking Multiple Targets

How many dots can we simultaneously track?
~4-6. Difficulty increases sig. at 6. [Yantis 92, Pylyshn 88, Cavanagh 05]
Animation Takeaways

Animation is a salient visual phenomenon
Attention, object constancy, causality, timing
Design with care: congruence & apprehension

For processes, static images may be preferable
For transitions, animation has demonstrated benefits, but consider task and timing
Ethical & Deceptive Visualization

Tufte's Lie Factor, original visualization from the NYT [Tufte 01]
Deceptive Visualization

**Lie Factor:** Distorting the apparent size of the effect in your data, often through choosing ambiguous or non-standard encodings.

**Scale Manipulation:** Changing with the scales of your chart to minimize, magnify, or invert the change in the data.

**Metric Manipulation:** Choosing how data are counted or normalized in order to hide or exaggerate effects in your data.
A3 Review: Ethical or Deceptive?

"You will be assigned at least one ethical and one deceptive visualization; the other two visualizations will be randomly assigned."
Graphical Perception

The psychophysics of sensory function [Stevens 61]
Graphical Perception
The ability of viewers to interpret visual (graphical) encodings of information and thereby decode information in graphs.
Pre-Attentive Features

[Information Visualization. Figure 5. 5 Ware 04]
Change Blindness
Change Blindness
Change Blindness
Change Blindness
Change Blindness

[Example from Palmer 99, originally due to Rock]
Just Noticeable Difference (JND)

JND (Weber’s Law)

Perceived Change → \( \Delta S = k \frac{\Delta I}{I} \) → Scale Factor (Empirically Determined)

Change of Intensity ◀ Physical Intensity

Ratios more important than magnitude

Most continuous variation in stimuli are perceived in discrete steps
Summary

Choosing effective visual encodings requires knowledge of visual perception.

Visual features/attributes
Individual attributes often pre-attentive
Multiple attributes may be separable or integral

Gestalt principles provide high-level guidelines

We don’t always see everything that is there!
CIE 1931 xy chromaticity diagram
showing the gamut of the sRGB and Adobe RGB color spaces including the Planckian locus, with temperatures indicated.
Wavelengths of monochromatic light are shown in blue.
Palette Design & Color Names

Minimize overlap and ambiguity of colors.

<table>
<thead>
<tr>
<th>Color Name Distance</th>
<th>Salience</th>
<th>Name</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.00</td>
<td>.20</td>
<td>blue</td>
</tr>
<tr>
<td>1.00</td>
<td>.47</td>
<td>orange</td>
</tr>
<tr>
<td>1.00</td>
<td>.67</td>
<td>green</td>
</tr>
<tr>
<td>1.00</td>
<td>.66</td>
<td>red</td>
</tr>
<tr>
<td>1.00</td>
<td>.66</td>
<td>purple</td>
</tr>
<tr>
<td>1.00</td>
<td>.47</td>
<td>brown</td>
</tr>
<tr>
<td>1.00</td>
<td>.37</td>
<td>pink</td>
</tr>
<tr>
<td>1.00</td>
<td>.58</td>
<td>grey</td>
</tr>
<tr>
<td>1.00</td>
<td>.18</td>
<td>yellow</td>
</tr>
<tr>
<td>1.00</td>
<td>.25</td>
<td>blue</td>
</tr>
</tbody>
</table>

Tableau-10

Average 0.97

http://vis.stanford.edu/color-names
Be Wary of Naïve Rainbows!

1. Hues are not naturally ordered
2. People segment colors into classes, perceptual banding
3. Naive rainbows are unfriendly to color blind viewers
4. Some colors are less effective at high spatial frequencies
Quantitative Color Encoding

Sequential color scale
Ramp in luminance, possibly also hue
Higher value -> darker color (or vice versa)

Diverging color scale
Useful when data has meaningful “midpoint”
Use neutral color (e.g., grey) for midpoint
Use saturated colors for endpoints

Limit number of steps in color to 3-9
Avoid simultaneous contrast, hold mappings in memory
Simultaneous Contrast

Inner & outer rings are the same physical purple.
Hints for the Colorist

Use only a few colors (~6 ideal)
Colors should be distinctive and named
Strive for color harmony (natural colors?)
Use cultural conventions; appreciate symbolism
Get it right in black and white
Respect the color blind
Take advantage of perceptual color spaces
Color is cultural and a matter of taste!
Visualization Software

D3: Data-Driven Documents
Chart Typologies
Excel, Many Eyes, Google Charts

Visual Analysis Grammars
VizQL, ggplot2, Vega-Lite

Visualization Grammars
Protovis, D3.js, Vega

Component Architectures
Prefuse, Flare, Improvise, VTK

Graphics APIs
Processing, OpenGL, Java2D
Chart Typologies
Excel, Many Eyes, Google Charts

Visual Analysis Grammars
VizQL, ggplot2, Vega-Lite

Visualization Grammars
Protovis, D3.js, Vega

Component Architectures
Prefuse, Flare, Improvise, VTK

Graphics APIs
Processing, OpenGL, Java2D
Interactive Data Exploration
- Tableau, Lyra, Polestar, Voyager

Visual Analysis Grammars
- VizQL, ggplot2, Vega-Lite

Visualization Grammars
- Protovis, D3.js, Vega

Component Architectures
- Prefuse, Flare, Improvise, VTK

Graphics APIs
- Processing, OpenGL, Java2D
Exploring Projections…

https://observablehq.com/@vega/vega-lite-cartographic-projections
Tissot’s Indicatrix

Circle size indicates the amount of area distortion
Symbol Map

NY Times
Visualizations : Word tree / Alberto Gonzales

Creator: Martin Wattenberg
Tags:

Data source: CQ Transcript Wire via the Washington Post

Comments (4)
Text Processing Pipeline

Tokenization
Segment text into terms.
Remove stop words?  a, an, the, of, to be
Numbers and symbols?  #huskies, @UW, OMG!!!!!
Entities?  Washington State, Seattle, U.S.A

Stemming
Group together different forms of a word.
Porter stemmer?  visualization(s), visualize(s), visually → visual
Lemmatization?  goes, went, gone → go

Ordered list of terms
Text Visualization Takeaways

High Dimensionality
Where possible use text to represent text…
… which terms are the most descriptive?

Context & Semantics
Provide relevant context to aid understanding.
Show (or provide access to) the source text.

Modeling Abstraction
Understand abstraction of your language models.
Match analysis task with appropriate tools and models.

Currently: from bag-of-words to *vector space embeddings*
Hierarchies and Networks

Degree-Of-Interest Trees [Heer & Card 04]
Trees and Graphs

Trees
Graphs with hierarchical structure
Connected graph with N-1 edges
Nodes as parents and children

Graphs
Model relations among data
Nodes and edges
Network Analysis Tasks [Pretorius '13]

**Structure-based:** relationships and connectivity
Find all of the friends of friends for Taylor.
Find all of the people who are friends with Jordan and Alex.
Six degrees of separation: shortest path between two individuals.

**Attribute-based:** specific node/link attributes
Find all "students" attending CSE412.
Find all the "friends" and "family" of Alex.

**Browsing:** understand paths in the data
Find Alex's friend Taylor, and then Taylor's friend Jordan.

**Estimation:** summarization and temporal changes
How does Jordan's friend group change over the course of the year?
Hierarchies and Networks

**Mon 5/17 - Tree Visualization**

**Wed 5/19 - Graph Layout: Node-Link Diagrams**

**Wed 5/19 - Alternative Visualizations & Techniques**
Scalability

1.7 B stars. 1.2 TB of data.

Visualizations running in-browser.
Interactive Scalability Strategies

1. Query Database
2. Client-Side Indexing / Data Cubes
3. Prefetching
4. Approximation

These strategies are not mutually exclusive! Systems can apply them in tandem.
Uncertainty
What does uncertainty mean?

Measurement Uncertainty:
"We're not exactly sure what the values in the data are."

Forecast Uncertainty:
"We're not exactly sure what will happen to the data next."

Model Uncertainty:
"We're not exactly sure how the data fits together."

Decision Uncertainty:
"We're not exactly sure what to do with the data."
Uncertainty Visualization Summary

Uncertainty can happen at all stages of the analysis process, from data collection to final decision-making.

Variables like blur and transparency can be intuitive for showing uncertainty, but hard to decode.

Consider using discrete samples to show variation and uncertainty in a model.

Consider when uncertainty is high enough that doing nothing is the right thing to do.
Evaluation vs. Microsoft File Explorer vs. Xerox PARC Hyperbolic Tree
Design and analyze visualization techniques in context of real-world use. Time/error analyses can be insightful, but they don’t provide a complete picture. Performance measures may be more suited to serious analysis than casual use?
Zoom Poll:
Top 3 Course Topics
The Future of Visualization

Where is more work required?

What emerging technologies and societal trends will impact visualization design?

What did you find most difficult in creating visualizations and designing techniques?