Administrative notes
March 6, 2018

• Reminder: In the News Call #2 group portion due tonight
• Reminder: Reading quiz due next week
• Reminder: Midterm #2 next Tuesday in class
  • Not cumulative (i.e., material from midterm #1 will not appear)
  • Covers all material from end of midterm #1 material through end of this unit
  • Remember to bring your ID and writing implements
  • We will include a table with powers of two and the first 16 binary, decimal, and hex representations
  • Seating chart will be posted on Canvas on Monday March 12
Data Visualization
Learning Goals

You should be able to:

• [CT Application] Be able to assess strengths and weaknesses of simple data visualizations

• [CT Building Block] List factors that you should consider when interpreting an infographic or statistic

• [CT Building Block] Understand the steps involved in creating data visualizations
Cholera in 1854 London

- Cholera is an infectious disease that affects the small intestine
  - Sanitation and clean water are vital for its prevention!
  - Without treatment, the death rate can be up to 50%![1]

- In 1854, a cholera outbreak swept the Soho district in London. In the first week, there were more than 150 deaths.

Group Exercise: Cholera Outbreak

• Pretend it is 1854 (remember: no computers!) and the government has told you to find the outbreak’s cause.
  • What would you do?
  • What kind of information would you want to collect?
  • What would you do with the information? Since there are no computers help organize information, how would you deal with the amount of information that you get?

• Map locations of people who were ill or died (where they spent most of their time/lived)
• Look it up in a library
• Age distribution
• Gender
Cholera in 1854 London

• Dr. John Snow plotted cholera deaths based on where the patient lived
• He eventually isolated the cause that the public water pump was infected
• The government took away the water pump's handle to stop people from using it; the outbreak died down

https://en.wikipedia.org/wiki/1854_Broad_Street_cholera_outbreak#/media/File:Snow-cholera-map-1.jpg
Why data visualization?

Visual representations of data can help us to

• Communicate known results
• Explore data to uncover patterns
• Carry out tasks more efficiently
• Make sound policy decisions
Data visualization and statistics

- Many visualizations convey statistical information
- Some statistics you may be familiar with are:
  - Mean/Average (sum of numbers, divided by the total number of numbers)
  - Median (the number in the middle of a set of numbers when the numbers are arranged in order)
  - Mode (the number that occurs most frequently)
In 2015, the state of New York had 966 traffic related deaths while the state of Wyoming had 190 traffic related deaths. Which state has safer traffic?

- **A: New York**
- B: Wyoming
- C: Both are dangerous
- D: Both are safe

New York (state) population: 19,795,791
Wyoming population: 586,107
Normalizing the data: NY: .005%   WY: .03%

http://www.census.gov/en.html
Statistics can lie!

- Numbers can be manipulated to skew a story
- You should check who is producing or promoting a statistic
  - Are they directly related?
  - Do they have an agenda they want to push?
  - **How** did the number come about?
When looking at data visualizations...

- A lot of the same rules apply!
- You want to look at
  - who the visualization is from
  - what kind of agenda they may have
  - how the numbers they are using came about
- Sometimes, visualizations will take advantage of people’s first impressions...

https://xkcd.com/1138/
Group Exercise: Are these visualizations misleading? Why or why not?

The NSW Health system is...

RECRUITING MORE NURSES


* Nursing headcount figures at June includes non-casual staff and 3rd schedule

Same data with zero shown

Series 1

Computational Thinking
www.ugrad.cs.ubc.ca/~cs100
Group Exercise: Is this visualization misleading? Why or why not?

Gun deaths in Florida

Number of murders committed using firearms

2005 Florida enacted its ‘Stand Your Ground’ law

Source: Florida Department of Law Enforcement

C. Chan 16/02/2014
Group Exercise: Are these visualizations misleading? Why or why not?

- Y-axis has 0 at the top
Even when there is no intent to mislead, data visualizations can be poorly designed…
Can you identify flaws in this visualization?

Average Number of New Automatic Guided Vehicles Retrofit Within Existing System – Year 2007

Based on new systems, existing system expansions and existing system replacements; does not include existing system retrofits/upgrades.

Source: Automatic Guided Vehicle Systems Product Section of the Material Handling Industry of America
Can you identify flaws in this visualization?
Data visualizations can be static
Data visualizations can be dynamic
Cornell Lab of Ornithology: Bird migration patterns

Data visualizations can be dynamic

“By using eBird data and other forms of migration tracking information, we’re getting a more detailed picture than ever before about where and when birds migrate. That’s the kind of information we need to make smart conservation decisions for species that live in vastly different regions during the year…”
Data visualizations can be interactive
Music Timeline

Group exercise: Play with this site. Discuss its strengths and weaknesses:
https://research.google.com/bigpicture/music/ (on lecture page on class website)
Music timeline strengths and weaknesses
https://research.google.com/bigpicture/music/

• Colourful—kept you interested
• Interactive, fun to click around
• Genres are even—Internet access makes it easier to find different genres of music

• Goth is not a category
• Hard to connect labels to a colour
• Bit difficult to understand the popularity graph (is area based or height based when it comes to popularity?)
• The way they set up the comparisons is hard to interpret but good to see
Data visualizations can be interactive

Group exercise

How long do you have to live? Experiment with this site. Discuss its strengths and weaknesses [http://flowingdata.com/2015/09/23/years-you-have-left-to-live-probably/](http://flowingdata.com/2015/09/23/years-you-have-left-to-live-probably/) (see lecture page)

- Visualization doesn’t take into account lifestyle factors or occupation
Creating a Visualization

1. Determine your domain
2. Break down your data/task
3. Select the appropriate visual encodings/interaction
4. Create an algorithm for the computer to create the visualization
Determining your domain

• A **domain** is a specific area (e.g., microbiology, chemistry, economics, etc.)
• It has its own vocabulary/terminology to describe things
• There are often existing conventions for describing or working with data
• There may also be tools associated with these tasks
• It is important to know about the domain you’re working with!
Determining your domain

- At the end of this stage, you should know what your user wants and needs from your visualization.
- In this stage, you may use methods like:
  - Interviewing the users
  - Observing what the users normally do
  - Doing some research into the users your visualization is trying to target
Example: Describe your domain

• The domain for the data provided is educational information.
• The predominant tools that people currently use to work with this kind of data are spreadsheets (e.g., Excel) or R (a statistics programming language).
Breaking down your data/task

• At this point, you have a specific task but this task is probably in very domain specific language.
• You can extrapolate this task into more general terms. This can help you determine what visual encodings/interaction (next step) will be needed!
Breaking down your task

Break your task down into three steps:

• **Analyze**: What is the purpose of your visualization? To **consume** existing data or to **produce** new data?

• **Search**: Do we know **what** we want to find? Do we know **where** to find whatever it is we are looking for?

• **Query**: How do you want to look at the data once you find it? Do you want to look at one data point, multiple data points, or all data points?
Breaking down your data/task: Search

<table>
<thead>
<tr>
<th>Location known</th>
<th>Target known</th>
<th>Target unknown</th>
</tr>
</thead>
<tbody>
<tr>
<td>Location known</td>
<td>Lookup</td>
<td>Browse</td>
</tr>
<tr>
<td>Location unknown</td>
<td>Locate</td>
<td>Explore</td>
</tr>
</tbody>
</table>
Breaking down your data/task: Query

- Query
  - Identify
    - [Diagram of points]
  - Compare
    - [Diagram of line chart]
  - Summarize
    - [Diagram of grid]
Example: Break down your task

• Our task is:
  “..to create something that will let people see how the percentage of women in different majors has changed over time”

• Although we say we are creating something, we are actually trying to help people look at data

• Since we aren’t producing any new numbers from our existing data, we are not producing numbers—we are consuming them
Example: Using not enough idioms for something can make a visualization hard to read / confusing.

- Computer Science: 13.6%
- Engineering: 0.8%
- Psychology: 44.4%
- Health Professions: 77.11%
- Social Sciences: 36.8%
Example: Using not enough idioms for something can make a visualization hard to read / confusing

- Computer Science: 13.6%
- Engineering: 0.8%
- Psychology: 44.4%
- Health Professions: 77.11%
- Social Sciences: 36.8%
Example: Certain idioms (or certain implementations of idioms) are better
Create the visualization using a computer tool

- After making all these decisions, we can start creating the visualization on a computer
- There are many nice visualization tools available now for this purpose – you can play with some in lab
In lab you’ll get to play with visualizations and do peer evaluations of each other’s
Great overview of a presentation by Tamara on the blog of Dr. Jenny Cham: http://jennycham.co.uk/2014/07/a-peek-into-the-world-of-data-visualisation-with-prof-munzner/
Learning Goals Revisited

You should be able to:

• [CT Application] Be able to assess strengths and weaknesses of simple data visualizations

• [CT Building Block] List factors that you should consider when interpreting an infographic or statistic

• [CT Building Block] Understand the steps involved in creating data visualizations
Midterm 2 testable material ends here

Material for midterm 2 will end here!

Midterm 2 is not cumulative (i.e., material from midterm #1 will not appear).