1 Data Visualization and analysis

- Data: the plural of datum. According to the Oxford English dictionary, a datum is

1. (a) Chiefly in pl. An item of (chiefly numerical) information, esp. one obtained by scientific work, a number of which are typically collected together for reference, analysis, or calculation.

   (b) In pl. Computing. The quantities, characters, or symbols on which operations are performed by a computer and which may be stored or transmitted in the form of electrical signals and held on recording media. Also (in non-technical contexts): items of information represented in digital form.

2. (a) Something given or granted; something known or assumed as fact, and made the basis of reasoning; an assumption or premise from which inferences are drawn.

   (b) Philos. Anything immediately apprehended by or presented to the mind or senses. Freq. in datum of consciousness, datum of sense.

3. Chiefly Surveying. A line, point, etc., forming a basis for measurement; a baseline, benchmark, or reference point

Data is a set of facts or items of information. It is important to remember that it can take on many forms, e.g. a number, many numbers, a
book, a collection of newspapers. Also, data are most easily analyzed if they are stored in a computer, but not all interesting data are digitized.

- Meta-data: Data about data. Meta-data tells you the meaning of the numbers or words or images. Suppose you are given a list of numbers \([37,38,43]\). It could be the ages of professors at Colby. Or it could be the heights of their children. Without that information, you can’t learn anything from the data. Yes, you can compute a statistic about it, such as the mean or standard deviation. But you wouldn’t know how to interpret it. You would learn nothing.

- Data visualization: the process of connecting data with our brains. While data visualization primarily refers to using our eyes to connect the data with our brains, data visualization is also possible via all of our other senses. The goal of data visualization is to enable the computational machinery in our brains to identify patterns, trends, or other interesting and salient characteristics of the data. While not all of the patterns we find in data are meaningful, our brains have incredible pattern identification and recognition capabilities, and data visualization enables us to make effective use of it.

- Data analysis: the process of computationally extracting information from data. Note that both visualization and analysis have the same ultimate goal: synthesizing knowledge from data. Data visualization enhances the ability of our biological computers to extract knowledge. Data analysis uses the computer to automate the process. Both have a role to play, as it can be difficult to define data analysis strategies without having some idea of the structures that exist in the data. Conversely, visualization can be difficult without some simple computational analyses to make the patterns more obvious.

2 Some examples

- From the Chronicle of Higher Education, a map showing the education level of state congress people:  
• Tim Berners-Lee (inventor of WWW), make raw data public. What happens when you put indep data together?
  
  http://www.ted.com/talks/tim_berners_lee_the_year_open_data_went_worldwide.html

• Circadian clock data

To see a list of TED talks about visualizations, check out this link:
  
  http://www.ted.com/topics/visualizations

To follow tech news, which often talks about Big Data, check out these links:

• slashdot.org Slashdot is a technology-related current affairs website. Summaries of stories and links to news articles are submitted by Slashdot’s own readers, and each story becomes the topic of a threaded discussion among users. (description stolen from Wikipedia)

• geekwire.com GeekWire is an independent technology news site and online community based in Seattle, Wash. covering the people, companies and innovations emerging from the Pacific Northwest and impacting the world.

3 Going over the syllabus

The syllabus is on the course website. No need to repeat it here.

4 Homework assignment

The usual pattern of this class will be to have homeworks on Wed that help you to prepare for a quiz on Friday. This week, we will have a discussion instead of a quiz. Search the web to find two examples of data visualization – one that is “good” and one that is “bad”. Send the links to Stephanie by 10pm on Thursday and she will compile them in order to facilitate a brief discussion on Friday. We will talk about what makes them good or bad.

To get you started, why not look at the galleries of examples on the course web page:
http://cs.colby.edu/courses/S14/cs251/goodbad.php

Also, read the articles at Lighthouse.org on making visuals accessible to people with vision problems.

http://www.lighthouse.org/accessibility/design/accessible-print-design/effective-color-contrast/

and

http://www.lighthouse.org/accessibility/design/accessible-print-design/making-text-legible/

Be prepared to answer the questions:

- What makes a visualization “good”?
- What makes a visualization “bad”?
- What makes this visualization “good” or “bad”?
- What have we learned from the visualization?
- Could the data have been visualized differently and been as informative? (or maybe more informative?)

Acknowledgement: Some of the material in these notes is taken from Bruce Maxwell’s spring 2012 CS251 notes.