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 [38,39,44]. 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 An example: GapMinder World

We use the GapMinder World software at [http://www.gapminder.org](http://www.gapminder.org).

For the data he shows in the talk, the basic visualization has these elements:

- one bubble per country.
- x - income
- y - life expectancy
- color - continent
- size - population
- animate over time (display year as background)

We use GapMinder in the browser.

- Students choose a data set. Look at it.
- We describe the features they liked.
- What was intuitive?
- Was anything difficult?

3 Data Terminolgy

- data point, data vector, or feature vector: one or more numbers representing a single measurement event.
- variable or feature: a symbol that connects a set of numbers to a meaningful description. A variable/feature usually refers to a single number within a data point or data vector.
- multi-variable data: a data set whose data points consist of more than one measurement.
• dimension: the number of variables/features/measurements in a data point.

• min: the minimum value of a feature within a data set.

• max: the maximum value of a feature within a data set.

• range: the upper and lower bounds of potential values for a feature, sometimes refers to max - min.

• independent variable: a direct measurement or value that does not depend on another value in the data point. In the example y = mx + b, the variables m, x and b are independent variables.

• dependent variable: a variable calculated from or that is a result of other variables or measurements in the data point. In the example y = mx + b, y is the dependent variable because it is completely defined by the variables on the right side of the equation.

• missing data: sometimes a data point will not contain all of the measurements that other data points in the set possess.

• meta-information or meta-data: a description of the variables in a data set, often including their source, method of measurement, valid range, or valid values for the variable.

• precision: a description of the number of significant figures in a measurement, which is based on the repeatability, or reproducibility of a measurement. Note that a repeatable measurement is not necessarily correct.

• accuracy: a description of how close a measurement is to the true value.

• scaling: multiplying data by a scale factor to change its units (e.g. from lbs to pixels).

• normalizing: Transforming data by a linear or nonlinear function so that all values are between 0 and 1.

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.

4 Going over the syllabus

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

5 Homework assignment

The usual pattern of this class will be to have homework on Wed that help you to prepare for a quiz on Friday. This week, we will have HW but not a quiz. To get a head start on future projects and to solidify your understanding of the terminology, find a data set (you can search the web, if you like), and send it to me, along with a statement indicating how many data points it has, the dimension of the data, the type of each data feature, and the meaning of each feature (e.g. the third feature is the weight in pounds of the most recent meal someone ate).

Send the data to Stephanie by 10pm on Thursday and she will respond to any questions/comments you might have.

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