1 Good and bad examples

We look at the examples and use the following questions to guide discussion:

- 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?)
2 An example: GapMinder World

We watch a movie in which Hans Rosling uses GapMinder World to show how much the world has changed in 200 years.

http://www.youtube.com/watch?v=BPt8EiTQMig

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 Terminology

- 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 variable within a data set.

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

• range: the upper and lower bounds of potential values for a variable, 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.
Precision and accuracy are important concepts in data analysis and visualization. A measurement with high precision is not necessarily accurate, and an accurate measurement is not necessarily precise. A data set with high accuracy and high precision is called valid.

In many data sets, the original precision of the data gets lost during analysis or transformation by a computer. For example, data that is originally integral, with a precision of 0.5 may be converted to a floating point representation as part of a transformation process and stored to a file using six decimal places. The number of significant figures in the data does not change, however, which means that the final representation likely contains meaningless digits. The original precision of the data will usually be lost unless it is included in the meta-data.

3.1 Data Precision

All data has a native precision. The native precision depends on the method of data capture and the precision of the measurement. Native precision is not always the same as the number of significant figures used to store the information. We could use six significant figures to store data captured with only three significant figures of precision. In that situation, the extra significant figures fool us into thinking the data is more precise than its native precision. The opposite can also be true. We could take measurements with three significant figures and store only two. In that case, the native precision is reduced by the method of storage.

Whenever data is transformed from one format to another, there is the real danger of the data being changed because of representational errors. The more transformations occur, the more likely that errors will creep into the data. For example, some numbers cannot be exactly represented in a floating point format, or require infinite series of decimals. If the values are written out to a text file with a fixed number of decimals, they get truncated. When the data is read back into the computer, the internal representation of the number is different than it was before the data was written to a file. If the process is repeated several times, the errors accumulate.

Responsible data management requires that we keep the data as close to its original form as possible and in a format that retains at least as much
precision as the original measurements. To avoid unnecessary quantization errors, we always want to start with the original source and build our processing chain in as few steps as possible. Any change to the data format introduces quantization errors.

Acknowledgement: The material in this section is taken from Bruce Maxwell’s spring 2012 CS251 notes.