**Final Exam (Take Home)**

## Data Analytics

School of Communication and Information

**Question 1:**

**Schoneberger and Cukier define “Big Data” as “*more*, *messy*, and *good enough*”. Explain each of these concepts with an example for each. [15 points]**

In the words of Schoneberger and Cukier**,** they define big data as society being able to connect information in new ways to produce useful insights or some type of value. Whereas, within this notion of big data, they proclaim that it must be “more, messy, and good enough”.

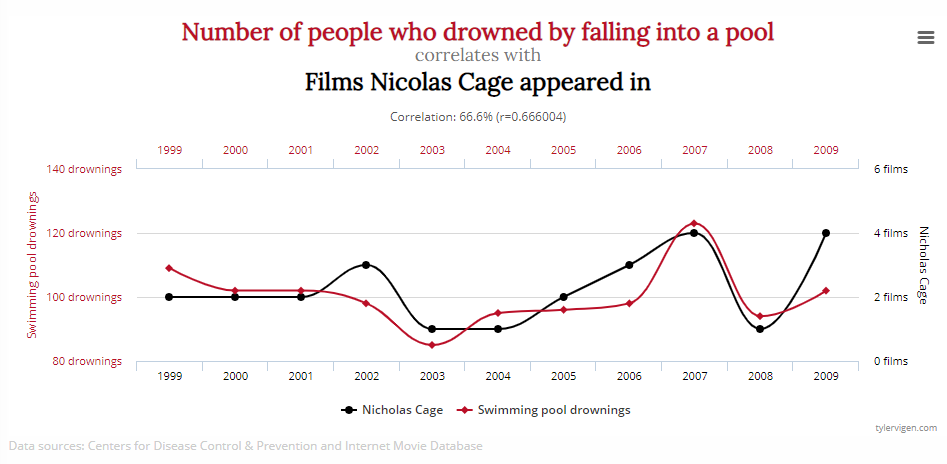
What the authors mean when they say big data is useful when there is “more”, they essentially mean that is it is important to use all of the data at hand instead of just a small portion of it. Because when we follow this practice, we are basically restricting or putting a censorship on the knowledge or information they are able to gain when working with the data. This ability to collect and work with larger sets of data is relatively newly possibly due to the technological advancements. But with these advancements came limitations as collecting large amounts of data can be complex, time-consuming, and costly. Whereas, a solution to this limitation is collecting vast amounts of data by randomizing and sampling. A real-life example that would resemble the concept of “more” is the U.S. Census Bureau. The Census Bureauconducts over two hundred economic and demographic surveys every year based on random sampling, in addition to the decennial census that tries to count everyone.

When the authors discuss that big data should be messy, they mean that none of the data should be “What we never wanted to do was consider them unavoidable and learn to live with them”

Excerpt From: Viktor Mayer-Schonberger. “Big Data: A Revolution That Will Transform How We Live, Work, and Think.” Apple Books.

**Question 2:**

**Refer to the figure below. What are the variables considered? What is the correlation score? What does correlation imply? Do you think that it is causal? Use at least one other example to explain correlation and causation. [15 points]**

****

**Question 3: What are some of the challenges/issues in applying Big Data analytics to human-facing applications (e.g. healthcare, finance)? Explain with at least three practical examples. [15 points]**

**Question 4:**

**Why is the question of *privacy* worth revisiting in the era of Big Data? In particular identify at least three aspects that have changed over the last 10 years. [15 points]**

**Answer either question 5 or 6. Please do not answer them both.**

**Question 5: Case Study**

**Look at the data set in the attached file. It contains information about average January temperature (JanTemp), latitude (Lat), and longitude (Long) in 56 different cities.**

**Undertake the following analysis.**

1. **Create a scatter plot between temperature and latitude [10 points]**
2. **Find the correlation coefficient between temperature and latitude. Is it significant? Positive? Strong? [10 points]**
3. **Build a regression model with temperature as an output (dependent) variable and latitude and longitude as the input (independent) variables. Include a screenshot of the analysis results and explain what percentage of variance in temperature is captured by the model? [10 points]**
4. **Interpret the results and also point out the potential limitations of this analysis. [10 points]**

**Question 6: Analytics Concepts**

1. **What is a “model” in terms of data analysis? Give an example. [10 points]**
2. **How will you explain the concept of hypothesis testing to a layperson? [10 points]**
3. **What is the difference between Mean and Median? If you could pick only one number to explain the salaries of NBA players, which one will you pick and why? [10 points]**
4. **“With big data we can escape the straightjacket of group identities and replace them with much more granular predictions for each individual.   
   The promise of big data is that we do what we’ve been doing all along – profiling – but make it better, less discriminatory and more individualized.” Argue a case for or against this new type of behavioral profiling. [10 points]**

**<End of Exam>**

**All the Best!**