**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”.

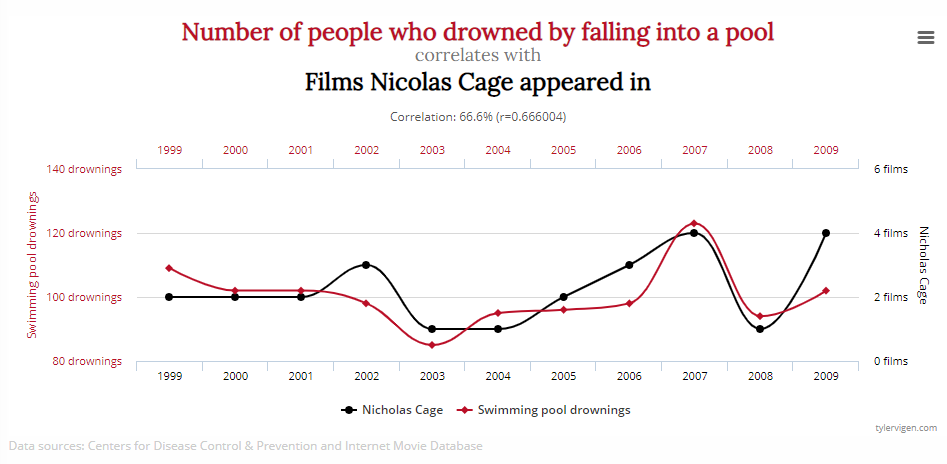
What the authors mean when they say big data is useful when there is “more”, they essentially mean that is it is important to use all of the data at hand instead of just a small portion of it. Because when we follow this practice, we are basically restricting or putting a censorship on the knowledge 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 considered unavoidable and learnt to be accepted when it is full of erroneous figures and corrupted bits from working with big data. In other words, “messy” occurs when one adds more data points, thus increasing the likelihood of errors. Another way “messy” data can come about is when one combine’s different types of information from different sources that don’t associate well. Whereas, when we combine or increase the data sets, we are able to trump the inaccuracy or errors with new information or connections that otherwise would have been missed. One example of messy data is improving the grammar checker that is part of the company’s Word program by feeding more data into the algorithm. As more data went in, the performance of the algorithms improved, and its accuracy went from 75 percent to 95 percent. Whereas this vast amount of data may not have been perfect, but the frequency of these connections allowed the algorithm to pick up on certain patterns and gain more accuracy it otherwise wouldn’t have.

Lastly, the idea that big data should also be “good enough” refers to the notion of moving away from the idea that in order to find solutions to things within data, there has to be a cause. Instead, we can look at data without a particular motive and find correlations and patterns. Whereas this brings certain knowledge or information we otherwise wouldn’t have seen to fruition because we were looking for something else. In other words, the most valuable information retrieved from big data came about from letting data speak for itself, rather than looking for a cause to a phenomenon. One example of data that is good enough is looking at medical data and noticing that a particular combination of aspirin and orange juice see their disease go into remission. This is considered good enough because although we do not know the exact reason for this, the fact that we are able to see this connection and use this information to help is good enough.

**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]**

****

Within this particular data graph, it is a depiction of a linear graph depicting the correlation (or lack thereof) two variables, the number of people who drowned by falling into a pool and the number of films Nicolas cage appeared in. Whereas this graph depicts number of people who drowned by falling into a pool is the dependent variable and the number of films Nicolas cage appeared in is the independent variable. The correlation score is approximately 0.66. Whereas this is a moderately positive relationship. This means that there a relatively strong connection between the number of people who drowned by falling into a pool and the number of films Nicolas cage appeared in. Whereas, this correlation implies that the connection is not entirely connected with each other, there is still a moderate connection between these two variables. Although given such, I do not think that this is casual given that correlation does not mean causation. In other words, just because these two variables are moderately correlated, that does not mean one is the cause and effect of the other. There needs to be more data and research to prove this correlation. Once example to explain correlation and causation is a correlation to shark attacks and ice cream. One may want to conclude they are casual. Even though it looks like that, there can be other factors to explain this phenomena such as an increase in numbers due to warmer or summer seasons.

**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]**

Some challenges and issues within the realm of big data that industries with human-facing applications such as healthcare and finance face consist of security and integrity of data.

One challenge we face with big data is the security of consumer data in terms of healthcare. With the rise of implementation of big data and the knowledge that can be retrieved from it, there have been many open data initiatives with the government to release data for analysts to do research and potentially use this data to find connections in data. Although the release of this data does not necessarily contain sensitive information that gives away personal data, it can still be harmful to communities. In the healthcare standpoint, one example where big data plays a role with security can be with open data that contains general information on the population of individuals with HIV. This can be a security issue to LGBTQ+ individuals because with the release of this information, it can further perpetuate the negative societal stereotypes of people in that community being the reason of this virus.

Challenges we face with big data and data integrity is holding the creators of these predictive algorithms curated with big data accountable. In past circumstances, it was easy to hold people accountable and uphold the law if they committed a crime or something unlawful. With big data, it opened up a new era of laws from holding people accountable for intentionally creating algorithms that hurt society to protecting companies from being sued based off the decisions they make from big data they collected. In addition, this leads to the debate of whether or not people should be based off their actions in the context of big data. One instance of this is in finance, these companies make decisions with big data to offer a mortgage or to even deny us a credit card. Companies have to watch out because although these machine learning algorithms can be extremely effective, they can also make the wrong decision. Whereas, wrong decision can potentially be seen as unlawful to the person effected.

**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]**

The question of privacy is worth revisiting in the era of Big Data because with the use of technology, we have been able to use big data in ways that we did not expect. One of these ways is big data’s ability to make accurate predictions about the world. With these predictions, uplifts information and an obstruction of privacy that we did not foresee or expect. In particular, three aspects of privacy within big data that has changed over the past ten years was the change in scale of personal information, the misuse big data, and the ability to make predictions before they happen.

In regard to the scale of big data, this can pose a serious privacy issue. With the increasing use of big data for its powerful insights it can bring also makes the size and scale of data collections increase greatly. This poses a threat to privacy because this data is getting reused, shared, and spread in a scale that is pretty much impossible to contain. Another aspect where privacy and big data play a role is in the possibility of misuse and dictatorship of big data. Big data is a powerful tool and if misused, it can lead to a form of repression. The ability to make predictions before they happen also is an important aspect within privacy to touch upon. With the accuracy in predictive models, it is possible to be used in unethical ways where it can judge and punish people just based off a prediction. Whereas, this leads to a clear use of unfairness and obstruction of justice as well as free will.

**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]**

In terms of data analysis, a model is meant to represent reality in a way that is constructed by organizing data and homogenizes how the data elements relate to one another. One example of a data model can be a predictive model that determines weather or not a plane ticket will increase or decrease in value overtime. Whereas the data collected would be particular plane tickets and routes, their prices, and the date bought perspective to the date of the flight.

1. **How will you explain the concept of hypothesis testing to a layperson? [10 points]**

I would explain the concept of hypothesis testing to a layperson by first describing what a hypothesis testing is in. Whereas hypothesis testing is the act of pursuing an experiment or observation on an educated guess about something in the world. Next, I then explain what a null and alternative hypothesis is. From there, I would draw out a chart that explains when to reject a null and alternative hypothesis. For instance, a null hypothesis is an observation that is true unless proven to be not true during observations. With the alternative hypothesis, we suspect our observation is not true unless we see information from the observation that proves otherwise.

1. **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]**

The difference between mean and median is that mean is the average of a particular set of values, the median is the middle most value within its set that is ordered from lowest to highest. If I could pick one number to explain the salaries of the salaries of NBA players, I would use median because the salaries of NBA players is usually disproportionate to the more “popular” players, whereas if I computed the average it wouldn’t really resemble most of the players salaries because of the outliers of high salary players.

1. **“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]**

I believe that this new type of behavioral profiling is beneficial because it allows for predictions, information, and knowledge to be found that wouldn’t be found if we followed the “straightjacket” of group identities. This new type of behavioral profiling is more representative of datasets and therefore allows us to make more proposed decisions. Its aspect of individualization also leads to more validity of propositions within the data. For instance, behavioral profiling can allow companies and other organizations to understand their consumers or target market in a more behavioral pattern rather than just standardized representation. The use of behavioral patterns will be much more useful in making decisions rather than just on quantifiable statistics.

**<End of Exam>**

**All the Best!**