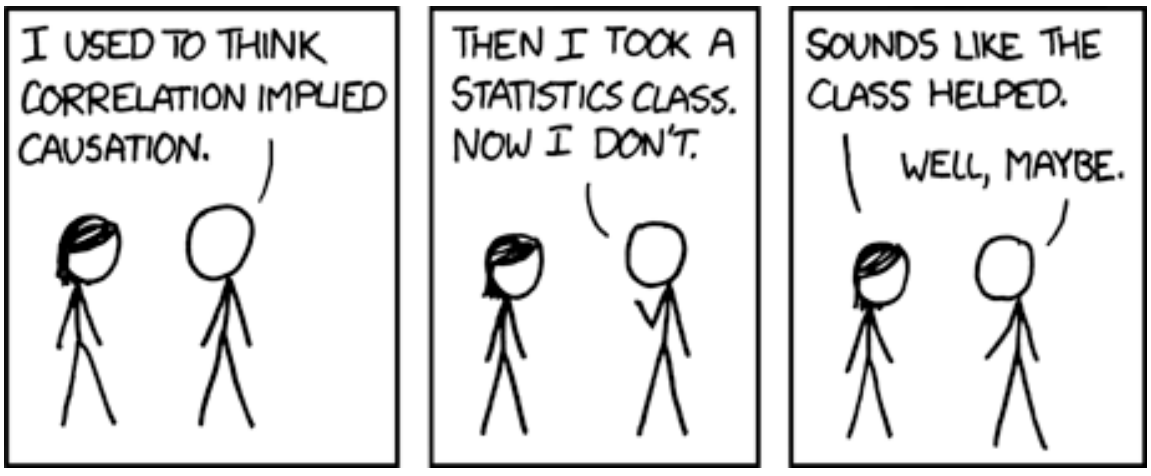
Towards a New Paradigm in Big Data:

Improving Decision Making in College Applications

(Humanities/Quantitative)

by Andrew Wong



http://xkcd.com/552/

Guiding Questions:

1. What is big data?
2. What are predictive analytics?
3. What are some current applications of predictive analytics in business and education?
4. How do educational institutions currently gather and store student data?
5. What makes a certain high school student compatible with a certain college or university?
6. How can high school extracurricular activities predict students' participation in college activities?
7. What are the analytical and computing approaches to model and analyze this data?
8. How can high school students use various data types and analytical tools to make better college application decisions?

I am a high school senior currently going through the final stage of the college admissions process. Since last year, I have been preparing for standardized tests, striving to excel in my schoolwork, working vigorously on my extracurricular activities, and practicing on essay writing. When I looked at the introspective questions colleges ask their perspective students such as “Describe where you come from and how your beliefs have been shaped” or “Write about a trait that you would like to share, that you feel your application would be incomplete without it,” I speculated about the purpose behind these topics. Having gone through a 4-week long business decision science course over the summer at Stanford University, I realize that I can in fact apply the decision-making concepts to the college application and admission processes. Decisions are made constantly in personal and professional situations. From the Stanford course, I learned that decisions are made with a combination of expected outcome and a good foundation.[1] During this course, I was introduced to the concept how probability and individual decision-making are correlated with external factors. At the same time, I was an intern at GeoSure Global, a mobile-app company that predicts travel safety based on travellers’ characteristics such as gender, age, and experience. Through the work at GeoSure, I learned that real-time user input could improve the quality of information for individual to make better travel decisions. From my own college application experience, the application of theories learned from the Stanford decision-making

[1] "Decision Education Foundation." Decision Education Foundation. DEF, n.d. Web. <http://www.decisioneducation.org/#!skills/cm8d>.

course, my internship at GeoSure, and my interaction with college counselors, I come to ask this question: with the pre-existing conditions in a college application, how can I predict the best school that fits me the best and most likely to admit me? In college admissions, I have been repeatedly informed and believe that the acceptance decisions do not solely come from a set of quantitative data points, such as GPA and SAT or ACT standard test scores. Rather, many colleges, especially those selective colleges utilize holistic review of the applications. This led me to think about the current technology trend in big data and its role in the college admission decision-making process from a broad perspective. Coincidentally, the hype around big data reveals that institutions prefer to look at situations with a broader scope, taking into account the qualitative aspects and not just the numbers.[2] Big data is currently used to spot trends in massive sets of data, so the question is can we use the same *methods* for college admissions and if so, how?

Big data can be traced back to the beginning of written history. Since then, humans have been using data (or information, in a broader sense) to account for their observations. About 7,000 years ago, Sumerian merchants started to record dates on clay beads. Later, the ancient Egyptians and Chinese conducted censuses to keep track and maintain stability of their populations.[3] In the Middle Ages, the *Domesday Book of 1086* was a census that delineated the ownership of land among

---

[2] Greene, Ilana. "On Trend: The Evolution of Big Data." The Huffington Post, 25 Sept. 2013. Web. <http://www.huffingtonpost.com/ilana-greene/on-trend-the-evolution-of_b_3981638.html>.

[3] Viktor Mayer-Schonberger and Kenneth Cukier. *Big Data: A Revolution That Will Transform How We Live, Work, and Think*. New York: Houghton, 2013. Print., (19-21)

over 13,000 people and is one of the most notable precedents to big data today.[4] The Internet has created the data revolution by generating immense amounts of data on a daily basis. In 2013, Cisco announced that the amount of data flowing across the internet reached a high of 667 exabytes (that's 667 x 10^{18} bytes!).[5] The extent of this data is now called "big data," large sets of information that can no longer be analyzed by conventional techniques. Simply put, big data is a vast amount of data and comes in many types and forms.[6] For example, consider old books. These books offer written text that are not digitized, but hold great value. We needed a way to preserve these books in case of an emergency, but digitizing all individual books is a costly and difficult ordeal. A computer scientist named Luis von Ahn, famous for creating the familiar anti-spambot test called CAPTCHA, came to aid. CAPTCHA, originally created to tell apart humans and bots, served another useful purpose; scanning old books and having humans decipher the distorted words allowed books to be translated into big data.[7] Citing reason for today's overwhelming accumulation of data, Kenneth Cukier, the data editor of famed magazine *The Economist*, believes that "one reason why we have so much data ...

---

[4] "Domesday Book." *The National Archives*. <http://www.nationalarchives.gov.uk/education/resources/domesday-book/>.

[5] "Data, Data Everywhere." The Economist. 2010. Web. <http://www.emc.com/collateral/analyst-reports/ar-the-economist-data-data-everywhere.pdf>.

[6] Mayer-Schonberger and Cukier, 78.

[7] Mayer-Schonberger and Cukier, 98.

is we're taking things that have always been informational but have never been rendered into a data format and we are putting it into data."[8]

To analyze such magnitude of data, three main levels of real-time practices called analytics exist: descriptive, prescriptive, and predictive. Descriptive analytics extracts current data and finds correlation among that data. Prescriptive analytics determines what humans can do to alter outcomes. Finally, predictive analytics uses historical data to predict the future.[9] These methods analyze the data as we receive it, and thus can expedite processes and give an answer almost instantaneously. Suppose you enter a Google search for the word "piza," but meant to search for "pizza." Since Google handles over three billion search queries every day and others have likely made a similar mistake, Google will store this seemingly incorrect data. Later, Google recognizes that the misspelling "piza" is highly correlated with the word "pizza," and automatically returns similar results for each word.[10] This is a prescriptive analytics example, which demonstrates that big data needs mainly sheer quantity of information, and even studies with no correlation can reveal hidden insights. Take another example: researchers testing Microsoft Word's grammar check. The researchers weren't sure about how to improve the script, so they tested accuracy based on quantity of words. They tested each grammar check algorithm at three different orders of magnitude: 10 million words, 100 million, and 1000 million (1

[8] Cukier, K. (2014, June). Kenneth Cukier: Big Data is Better Data. Retrieved from <http://www.ted.com/talks/kenneth_cukier_big_data_is_better_data?language=en>

[9] "Analytics Examples." Master of Science in Analytics. Northwestern University, Web. <http://www.analytics.northwestern.edu/program-overview/analytics-examples.html>.

[10] Mayer-Schonberger and Cukier, 112.

billion). To their expectations, the simplest algorithm was the least accurate, at a 75% success rate for 10 million words. However, as the amount of words increased, so did the efficiency, from 75% to 86% to 95% correct for the increasing number of words fed, respectively.[11] This descriptive analytics study proves a point that often times more data can overtake complex algorithms. Next, we look to predictive analytics for solutions.

Predictive analytics are commonly used to gain useful information by tracking and analyzing current and historical data. As the Internet usage increases, more data will be shared both in the private and the public sectors. High-tech companies like Google and Amazon, known as pioneers of web and data services, already use relevance in their search methods to detect consumers' purchasing tendencies and make recommendations for them. Amazon can even predict the items their customers will buy, and ship the items in advance to a nearby warehouse.[12] Organizations in the public sector are utilizing predictive analytics, too. Take a look at the Obama Administration: although it looks to increase restrictions of public data, it seeks to improve quality of life by recording body data for babies born prematurely to avoid infections, providing performance sensors in vehicles to optimize fuel efficiency, and targeting previous instances of tax fraud to prevent future occurrences.[13] In the education sector, post-secondary educational institutions such

---

[11] Mayer-Schonberger and Cukier, 36.

[12] Loukides, Mike. *What Is Data Science?* Rep. O'Reilly Media, Web. (3-4).

[13] "FACT SHEET: Big Data and Privacy Working Group Review." The White House. Web. <http://www.whitehouse.gov/the-press-office/2014/05/01/fact-sheet-big-data-and-privacy-working-group-review>.

as colleges and universities are using big data predictive analytics to predict graduation rate. In a study conducted by Education Advisory Board (EAB), a Washington-based educational research company, college freshman were assessed by a variety of parameters such as first year GPA, income level, ethnicity, and access to academic resources (tutoring, mentorship). This project acknowledged all factors but focused on first year GPA, providing a correlation between academic performance and resources available to the student. The research projected three groups of students: high-risk students, average students, and high-performing students (See Figure 1).

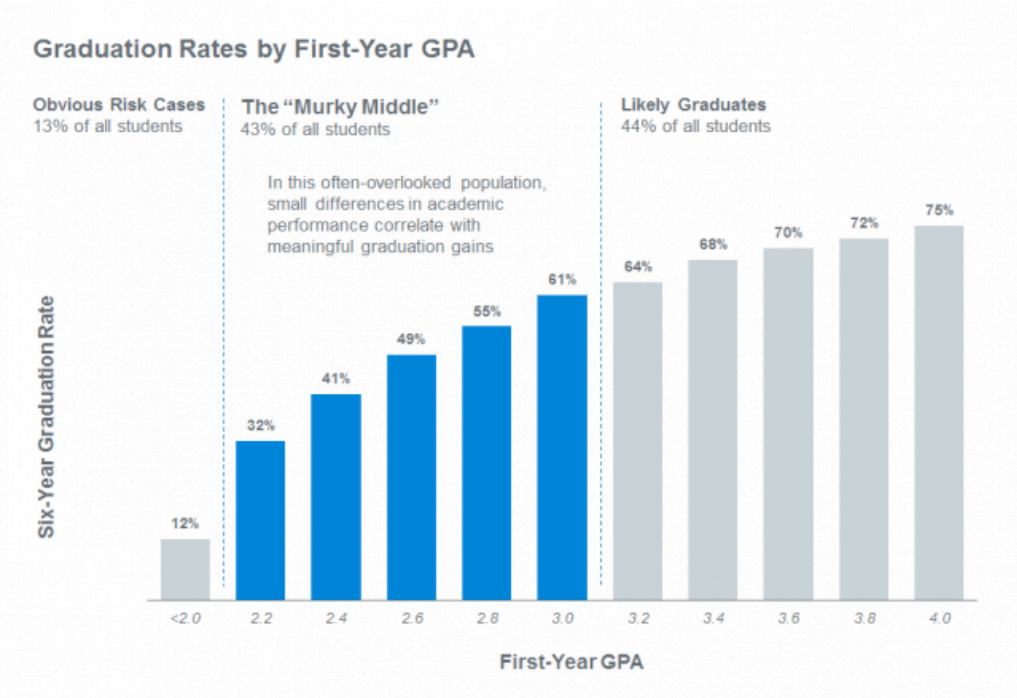


Figure 1. Graduation Rate by First-Year GPA[14]

[14]Tyson, Charlie. "Graduation Rates by First Year GPA." Inside Higher Ed, n.d. Web. <https://www.insidehighered.com/news/2014/09/10/maximize-graduation-rates-colleges-should-focus-middle-range-students-research-shows> Web. (Figure 1)

The data presented by the EAB shows that 45% of drop-outs occurred within the 2.0 to 3.0 first-year GPA range. The data also suggests that some of these drop-outs may have a fluctuating A to F grade range, rather than a consistent B to C range. With more information, the EAB further concluded that with more academic resources, these fluctuating students might be able to complete the required coursework to stay in college.[15] Based on the above results, it is not difficult to infer that colleges are likely to admit high school students whose profiles are similar to their high performing freshman students. These examples in predictive analytics demonstrate the possibility of using such methods to assist college application and decision process.

The objective of this project is to identify advanced methods currently used in analyzing big data, and to apply predictive analytics to the college application process. Like any information, big data can be erratic, unstructured, and complex. High schools and colleges are storing more student data than ever before.[16] There is a potential benefit for interpretation of such data as institutions are quite likely to utilize certain information extracted from the data to identify qualifying applicants, in order to improve the quality of education. High schools typically would have less data because of the limited number of students. Colleges, on the other hand, usually have significant more number of students; some hold up to 20,000 students at any

---

[15]https://www.insidehighered.com/news/2014/09/10/maximize-graduation-rates-colleges-should-focus-middle-range-students-research-shows.

[16] Davis, Michelle R., and Sean Cavanagh. "Cloud Computing in K-12 Expands." *Education Week*. Education Week, Web. <http://www.edweek.org/ew/articles/2014/01/08/15cloud_ep.h33.html>.

given time, thus have much more data. One possible issue that may arise when interpreting student data is the privacy issue. There is an ongoing debate to determine whether certain data are deemed public or private, and if institutions have the right to collect personal information. For example, the ACLU believes that data-tracking harms one's privacy and is a way of profiling through discrimination.[17] On the other hand, many proponents of technology argue that the prevalence of data would justify the right to analyze them.[18] For college admissions data, students no longer hold the right to their own data after submission of the same. Conversely, students may utilize whatever data that is available to aid their decisions.

College admissions become more selective each year, which means acceptance rates may decrease for certain colleges, especially for more selective ones. From the students' perspective, college applications become harder than ever. It may partially be because of services like the Common Application that allow one application to be sent out to multiple schools that increases the number of applications received by certain colleges. An article in the New York Times states, "If you work hard and get good grades and test scores, there is very likely a place in the best schools for you."[19] As a result of this common thinking pattern, many students use services like "Parchment.com" to compare their scores with those of

---

[17] "Civil Rights Principles for the Era of Big Data." The Leadership Conference on Civil and Human Rights. Web. <http://www.civilrights.org/press/2014/civil-rights-principles-big-data.html>.

[18] "Data, Data Everywhere." The Economist, 2010.

[19] Carey, Kevin. "For Accomplished Students, Reaching a Good College Isn't as Hard as It Seems." The New York Times. The New York Times, 30 Nov. 2014. Web. <http://www.nytimes.com/2014/11/30/upshot/for-accomplished-students-reaching-a-top-college-isnt-actually-that-hard.html>.

similar students.[20] The flaw in this mindset is that some students may think that colleges only draw upon numeric data such as GPA, SAT scores, and ACT scores, and neglect other aspects such as personal qualities, accomplishments, extracurricular activities, and teacher recommendations.

In addition, individual decisions should take into consideration the institutional criteria for admissions. For example, unlike many institutions, the University of California (UC) system disregards pluses (+) and minuses (-) of the high school grades.[21] Because of this rule in the application, a student with more A-'s now would look exactly the same as another students with many A+'s. As a result, assuming everything else is the same, the two students are now equally appealing as far as the grading system in the application of University of California schools is concerned. Furthermore, the UC application system only account for students' sophomore and junior transcripts and does not require any recommendations, which may benefit certain group of students.

Moreover, I believe that the possibility of a student being accepted to certain college can be modeled by conditional probability. From the Stanford summer program on decision science and its applications, I learned that conditional probability measures the probability of an event if something else has already occurred. Consider a situation in which 5 similarly or better qualified students of the same high school intend to apply to a selective college. The expectation of the

---

[20] *New York Times,* Nov. 29, 2014.

[21] "University of California - Admissions." University of California. Web. <http://admission.universityofcalifornia.edu/counselors/q-and-a/calculating-gpa/#3>.

admission for students should be lowered accordingly. As such, when assessing fit and analyzing student decisions, these shifts in frame should be considered as well. This example illustrates the application of big data approach because these types of information are not commonly available to the students and need to be consciously sought after.

While applying for colleges, I was faced with a similar situation. Maintaining a balanced perspective, I tried to take into account known priorities of certain institutions that I was interested in and the criteria and characteristics and academic compatibilities of the colleges I wanted to apply. After discussing with my high school as well as the college admissions counselors, I classified the main variables for admission decision to be the following 5: academic rigor, institutional priorities, qualitative aspects, financial concerns, and random error.[22] Within these five main variables are many other data points that admissions representatives take into account. First, academic rigor includes the quality and diversity of classes, as well as high school student performance outlined by GPA and standardized testing (SAT, ACT). Colleges weigh on academic rigor to predict how well a student will do in college courses. Institutional priorities are certain activities or characteristics that certain colleges seek, such as replacing graduating athletes and musicians. This also includes interest in underrepresented majors, notable organizations, and research, as well as geographic and demographic diversity. Qualitative aspects include teacher recommendations, essays, personal qualities, and students' career

[22] Fernandez-Rojo, Rosita. "College Admissions Decisions Variables." Personal interview. 17 Oct. 2014.

goals. These factors will be discussed later from student fit and overall contribution perspectives. Financial concerns address the compatibility of the financial aid expectations of institutions and students. Finally, random errors are those variables that have not been accounted for by the above four (4) categories and may include error such as overlooking a student's strength.[23]

One important aspect of this decision making process is the student's compatibility with the institution. Two main categories for this type of fit include students' academic curiosity and achievement as well as their leadership and contribution to the community. First, students' academic achievement is related to their personality traits. This is illustrated by a recent study conducted at the Higher Education Research Institute (HERI) at UCLA called "Habits of Mind," which have shown that there are close correlations between academic success and certain behaviors. HERI defines "Habits of Mind" as "a unified measure of the behaviors and traits associated with academic success [and] the foundation for lifelong learning."[24] From this study, researchers formulated "estimation weights" based on the importance of each trait demonstrated in the classroom.[25] The study found that students who generally had the willingness to learn outside of the classroom and help others in that pursuit found the most academic success. These same students

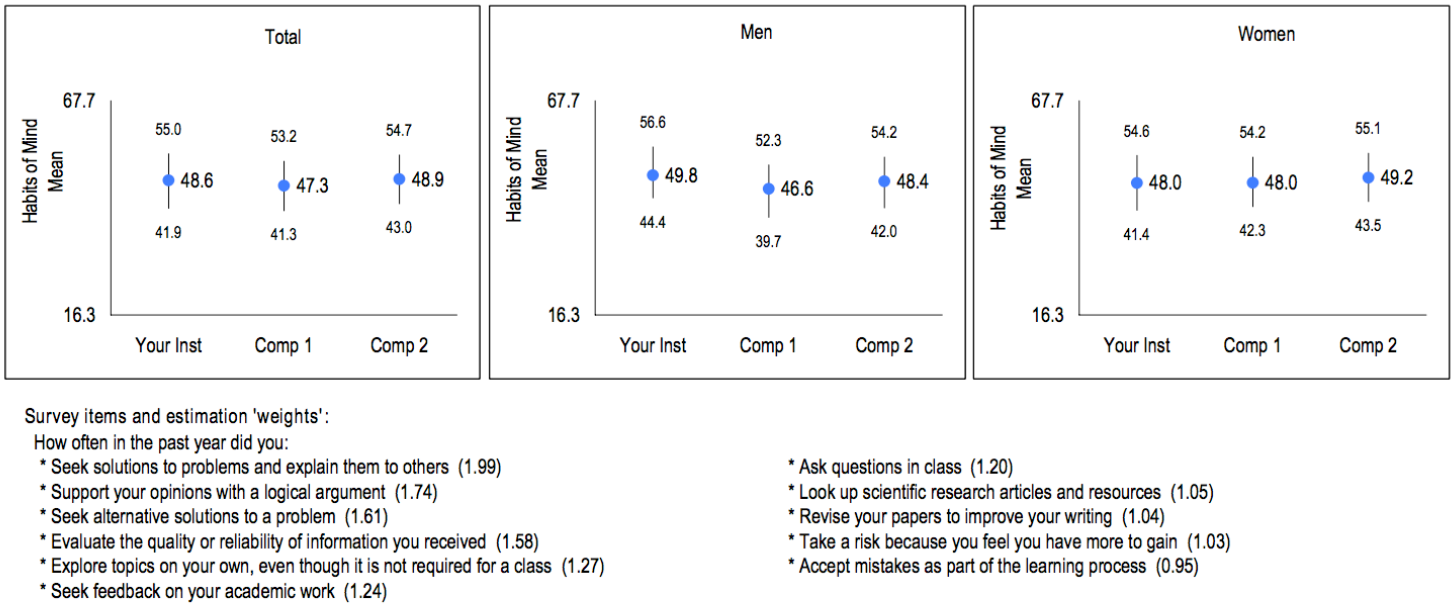
---

[23] Fernandez-Rojo, Rosita.

[24] Habits of Mind. Rep. Higher Education Research Institute at UCLA, n.d. Web. <http://www.heri.ucla.edu/PDFs/surveyAdmin/tfs/TFS_SampleReport.pdf>.

[25] Habits of Mind

are keen to embracing diversity[26], have a high tolerance of failure, and aspire to be lifelong learners, among many other traits (See Figure 2).



1A

Figure 2. Habits of Mind Study[27]

I came to the conclusion that colleges may not simply rest on GPA and testing scores, but also rely on the personality traits exhibited in essays, recommendations and extracurricular activities. As shown, these personal qualities in a student demonstrate motivation to learn and contribute to the community. Exhibiting these

[26] <http://www.heri.ucla.edu/infographics/TFS2013Infographic.pdf>. Web.

[27] <http://www.heri.ucla.edu/PDFs/surveyAdmin/tfs/TFS_SampleReport.pdf>. (Figure 2)

characteristics show positive work habits that help these individuals succeed in the long run.

In addition to academic achievement, the second category of student compatibility is based upon the contribution and leadership in activities. My current research is parallel with the analysis done by researchers at UCLA to address students' contribution by showing that participating in extracurricular activities can predict future involvement in college activities. In a particular survey, the researchers asked incoming college freshmen to assess themselves on the likelihood of participating in certain groups (See Figure 3).

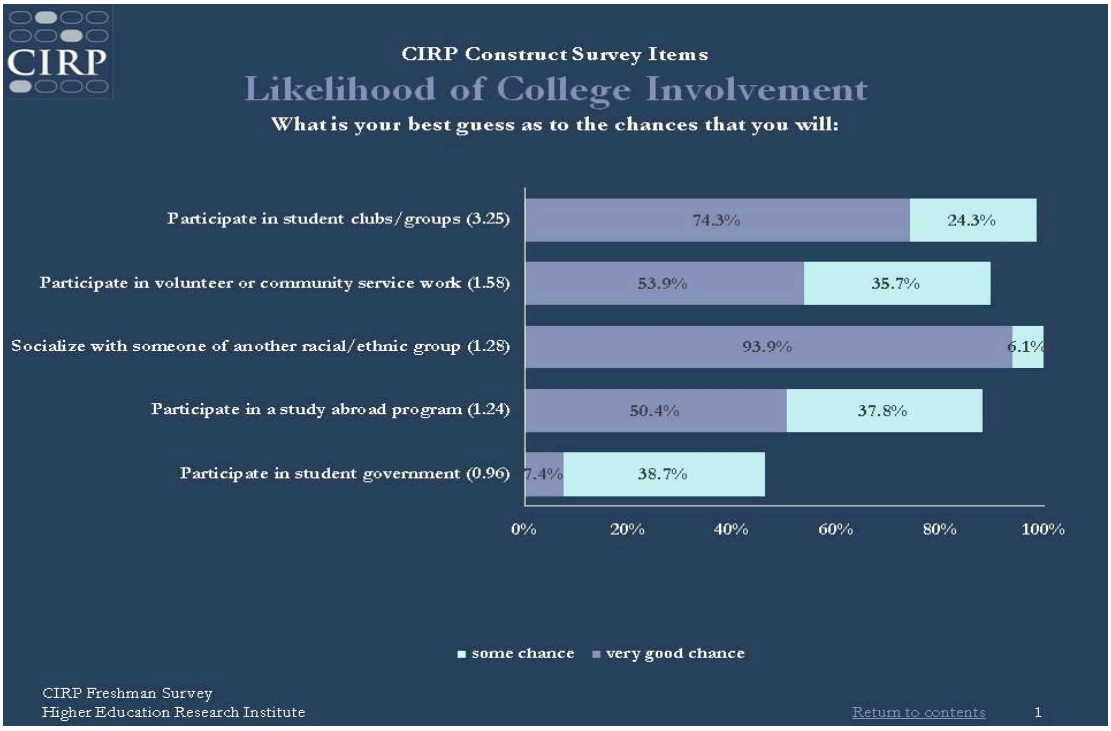


Figure 3.

Likelihood of College Involvement Survey[28]

[28] Web. <http://www.heri.ucla.edu/images/TFS.graphic.jpg>. (Figure 3)

They then compared the choices of activities selected to each student's high school participation, and found a close link between extracurricular activities and college involvement (Figure 3).

Further, the traditional data collected by these researchers can now be enhanced within big data framework by incorporating other less formal and behavioral information from channels such as social media. By collecting information from various channels, colleges can have more comprehensive view of a student to make better-informed and accurate admission decisions.

Big data, in general, enables higher education institutions to collect much richer data sets of applicants. Yet there can be challenges in collecting certain types of data and analyzing them. While academic rigor and financial concerns of the college are mostly consistent each year, institutional priorities can be unpredictable. We try to best solve this issue by delegating a network of computers to solely extract public databases of desirable students and student-related news (public information). Another challenge might be the analysis of large amount of textual information collected through social media. To overcome this challenge, the solution can be to apply natural language processing[29] and machine learning[30] methodology to extract personal traits and behavioral information. From a students' perspective, one challenge might be the prediction of the institutional assessment of their essays

[29] The process of translating text to data

[30] More commonly known as artificial intelligence, when a computer makes computations based on previous data

and teachers' recommendations. To solve that, we can perform statistical modeling of the qualitative data to produce more reliable rating through the use of signal detection theory, which in its most simple form states that an algorithm can calculate a score based on a rater's history and qualifications taking into account his/her bias.[31] This method is significant because it allows us to further recognize good candidates by using previous data. For example, if the rater of an essay is a reputable former college admissions officer, he/she will receive a high multiplier. In contrast, if the rater is an outsourced human intelligence worker (e.g. Amazon Mechanical Turk, a service that crowd-sources simple tasks), he/she will receive a much lower multiplier. Of course, those with higher multipliers will have a much higher influence on the final rating of the essays. Although more factors go into the final college decision, this idea of an algorithm that is not solely quantitative or exacting of accuracy reflects the fundamental motivations of big data.

Since we deal with big data methods, we can look to accomplishing these tasks by cloud computing, which distributes sets of data over networks of computers to expedite processes. Cloud computing proliferated when engineers at Google needed to combat the growing concern for analyzing big sets of data stored in Google's data warehouses. They developed a programming model called MapReduce, which runs on a large cluster of about 1,000 to 2,000 computers, simultaneously analyzing one large set of data.[32] This simultaneous analysis, called

[31] "Overview of Signal Detection Theory." Overview of Signal Detection Theory. Web. <http://wise.cgu.edu/sdtmod/overview.asp>.

sharding, allows for efficient investigation of data through many computers without an individual computer interfering with another. This method alleviates the problems of scaling operations and looking at unstructured data.[33] Unstructured data solves the problem of categorizing every item and looks at a set of data as a whole. It often takes paragraph form rather than a list of numbers. For example, instead of using a list like {3, 0.86, 1, 0.95}, we state: “3 students received a grade of 86% on the test, and 1 student received a grade of 95% on the test” and store the information in a database. From my research, I noticed that essays and written supplements are written in a similar form and can be analyzed in the same way. As a result, we may be finally able to assess personal qualities by comparing essays and recommendations to criteria set by colleges. Unlike a small sampling of data, which is often stored in databases with specific names for variables called relational databases, big data is often “messy” and rather unpleasant to organize.[34] Overcoming this obstacle by cloud computing, we can extract real-time educational institutional information and compare student data to produce an almost instantaneous answer. I predict in the future, a student will be able to predict accurately his/her best fit and the likelihood of each college's acceptance by pressing a few keys.

---

[32] Dean, Jeffrey, and Sanjay Ghemawat. "MapReduce: Simplified Data Processing on Large Clusters." (2004): Web. <http://static.googleusercontent.com/media/research.google.com/en/us/archive/mapreduce-osdi04.pdf>.

[33] Nist Computer Security Division. "The NIST Definition of Cloud Computing." Web. <http://csrc.nist.gov/publications/nistpubs/800-145/SP800-145.pdf>.

[34] Mayer-chonberger and Cukier, 34-35.

In conclusion, big data reflects needs that shape our society. Whether used in private sector or public sector, big data predictive analytics serves as an additional possibility of analysis that has a high potential to reveal essential insights. We learn that even a “messy” set of variables like college admissions variables may be organized after statistical analysis is performed. College admissions variables contain a great deal of variety, for each student is different. However, we learn that with big data, differences are now understood to have correlations. With modern technology and computing methods, we are able to provide these insights that we are normally unable to produce.

Looking back from where I started, I’ve learned much about big data methods and their applications to improve the quality and efficiency of college admissions decisions. I am aware that the journey will be long as I continue to acquire knowledge about the constantly evolving technology such as cloud computing. I strive to work towards the big data paradigm to improve the quality of life. Just like what big data reveals about information in general, although there is no telling *how* I will reach my answer yet, I know by this research that I *can* reach this goal.

Bibliography

"Analytics Examples." Master of Science in Analytics. Northwestern University, Web. <http://www.analytics.northwestern.edu/program-overview/analytics-examples.html>.

"Civil Rights Principles for the Era of Big Data." The Leadership Conference on Civil and Human Rights. Web. <http://www.civilrights.org/press/2014/civil-rights-principles-big-data.html>.

"Data, Data Everywhere." The Economist. 2010. Web. <http://www.emc.com/collateral/analyst-reports/ar-the-economist-data-data-everywhere.pdf>.

"Decision Education Foundation." Decision Education Foundation. DEF, n.d. Web. <http://www.decisioneducation.org/#!skills/cm8d>.

"Domesday Book." The National Archives. The National Archives, n.d. Web. 10 Oct. 2014. <http://www.nationalarchives.gov.uk/education/resources/domesday-book/>.

"FACT SHEET: Big Data and Privacy Working Group Review." The White House. The White House, Web. <http://www.whitehouse.gov/the-press-office/2014/05/01/fact-sheet-big-data-and-privacy-working-group-review>.

"Overview of Signal Detection Theory." Overview of Signal Detection Theory. Web. <http://wise.cgu.edu/sdtmod/overview.asp>.

"University of California - Admissions." University of California. Web. <http://admission.universityofcalifornia.edu/counselors/q-and-a/calculating-gpa/#3>.

Carey, Kevin. "For Accomplished Students, Reaching a Good College Isn't as Hard as It Seems." The New York Times. The New York Times, 30 Nov. 2014. Web. <http://www.nytimes.com/2014/11/30/upshot/for-accomplished-students-reaching-a-top-college-isnt-actually-that-hard.html>.

Cukier, K. (2014, June). Kenneth Cukier: Big Data is Better Data. Retrieved from <http://www.ted.com/talks/kenneth_cukier_big_data_is_better_data?language=en>

Davis, Michelle R., and Sean Cavanagh. "Cloud Computing in K-12 Expands." Education Week. Education Week, Web. <http://www.edweek.org/ew/articles/2014/01/08/15cloud_ep.h33.html>.

Dean, Jeffrey, and Sanjay Ghemawat. "MapReduce: Simplified Data Processing on Large Clusters." (2004): Web. <http://static.googleusercontent.com/media/research.google.com/en/us/archive/mapreduce-osdi04.pdf>.

Fernandez-Rojo, Rosita. "College Admissions Decisions Variables." Personal interview. 17 Oct. 2014.

Greene, Ilana. "On Trend: The Evolution of Big Data." The Huffington Post, 25 Sept. 2013. Web. <http://www.huffingtonpost.com/ilana-greene/on-trend-the-evolution-of_b_3981638.html>.

Loukides, Mike. What Is Data Science? Rep. O'Reilly Media, Web.

Mayer-Schonberger, Viktor, and Kenneth Cukier. Big Data: A Revolution That Will Transform How We Live, Work, and Think. New York: Houghton, 2013. Print.

Nist Computer Security Division. "The NIST Definition of Cloud Computing." Web. <http://csrc.nist.gov/publications/nistpubs/800-145/SP800-145.pdf>.

The American Freshman 2013, Habits of Mind. Rep. Higher Education Research Institute at UCLA, Web. <http://www.heri.ucla.edu/monographs/TheAmericanFreshman2013.pdf>.

Tyson, Charlie. "Graduation Rates by First Year GPA.” Inside Higher Ed, n.d. Web. <https://www.insidehighered.com/news/2014/09/10/maximize-graduation-rates-colleges-should-focus-middle-range-students-research-shows>.