Paper 1682-2018

BIG DATA Analytic Models: The Challenges of Philosophies as Covariates in Higher Education

Sean W. Mulvenon, Ph.D., University of Nevada, Las Vegas

ABSTRACT

Use of BIG DATA analytics in business is commonplace and is supporting the growth of industries throughout the world. A major source of BIG DATA is K-12 and higher education. Since 2004 several billion in funding has been used to create BIG DATA systems in education, but there has been limited success in the use of this data to improve educational outcomes of students in the K-12 and Higher Education systems. Why? A singular goal of business is to generate revenue and evidence from BIG DATA facilitates these efforts. Too often in education, goals/outcomes are based on philosophies, rather than scientific evidence from BIG DATA. The purpose of this session is to outline the challenges of BIG DATA in higher education and the impact of philosophies as covariates.

INTRODUCTION

The creation of large data systems in the K-12 and higher education systems has generated significant clarity to data systems. The cost of these systems is staggering and the usability to proactively support improving student outcomes is a reoccurring question. BIG DATA works in most industries, including banking and pharmaceutical fields, multiple marketing platforms, including identifying purchasing trends, in developing shopping patterns, or creation of consumer profiles to incessantly send advertisements via email messages or display in our browser windows. However, in education, the use of BIG DATA has had limited demonstrated effectiveness? Why? Are the data less accurate? Are the problems more complex? Higher education possesses the academic expertise to expand use of nanotechnology to improve food storage, cancer research to increase five-year survival rates for breast cancer patients, and design alternative fuel sources that lower CO_2 in our atmosphere. However, even with all the BIG DATA available in higher education there has been limited success in increasing college graduation rates. Why? The BIG DATA are available to grow our understanding of K-12 and higher education performance models and to transform educational outcomes of students. The goal of this paper is to outline one professor's overview of the challenges and possible solutions to improve use of BIG DATA in higher education (and K-12).

As a method to demonstrate the challenges and impact of philosophies for more effective use of BIG DATA I shall use a key question that is a goal for all higher education institutions:

> Key Question: What information do we need to improve student graduate rates?

GROUND RULES FOR UNDERSTANDING MY "PHILOSOPHIES" ON BIG DATA IN HIGHER EDUCATION

As an entry point to this paper, I think it is important to share some thoughts about my passions, theories and "philosophies" regarding BIG DATA in education. I have my own issues regarding the challenges of effectively using BIG DATA, and I believe it is important to be candid.

PERSONAL PHILOSOPHY #1: "DO NO HARM TO KIDS"

During my time as a senior advisor to the Deputy Secretary of Education, Mr. Raymond Simon, I heard him state "do no harm to kids" repeatedly. Much like an educational goal of 100% of students graduating, doing "NO harm to kids" may not be possible as some decisions invariably cause unintended consequences. However, rather than interpreting this as an actionable edict that creates paralysis in the decision-making process, I see it as an important metaphor of integrity in one's actions on behalf of students. Bad analytics using BIG DATA can have the impact of harming students.

PERSONAL PHILOSOPHY #2: PROFITEERING FROM EDUCATION

Too many companies are profiting from K-12 and higher education systems with suspect solutions/products and this inhibits our ability to identify and develop effective interventions/solutions. One of my favorite examples of this type of activity involves the per-student cost of educational data systems (Oklahoma Legislative Committee, August 2008).

> **Scenario:** "A company shares it will cost $3.00 per student to develop a comprehensive data system and you indicate there are one million students so they quote you the price of $3 million for your system. However, you based the number one million on students who are in the accountability system, yet you have another one million who need to be included using the same data platform. The company says 'no problem' and says this system will be $6 million" (Mulvenon, 2008).

What's wrong with this situation? Aside from the need to include one additional variable (1, 0) in the accountability system, the data platform is the same and results in an additional $3 million expense or 100% profit. Another example occurred when a Title I coordinator called in tears because they had forgotten to include in this variable on the coding sheets for a large school district and the testing company quoted $56,000 to include at this late date. I asked "do you have this information with the student ID number?" She replied "Yes" and if she could send it as an Excel spreadsheet? Thirty minutes later the Title I data had been merged with the state testing data and saved this testing coordinator $56,000. Both examples demonstrate a process of unnecessary profiteering on educational systems.

PERSONAL PHILOSOPHY #3: MEANINGFUL ANALYTICS

I once attended a SAS data workshop and the instructor would often say "Know thy data." Analytics is more than generating results, and should be focused on meaningful analytics. Too often BIG DATA are used to generate spurious analytics. Can anyone identify the challenge with the following statement?:

> "GRE exam scores don't predict graduate school success of students in the College of Education!"

Answer: This is a restriction of range issue with no variability due to average graduate school GPA's of 3.8 for the college. In case you are wondering, the same issue happens if you attempt to predict graduate school GPA with undergraduate GPA's. I Googled ACT predicting college success and received 193 million "hits." This type of analytic is not meaningful or helpful; and we can do better by knowing our data and studying the systems.

PERSONAL PHILOSOPHY #4: PHILOSOPHIES AS COVARIATES

The use of demographic and sociocultural variables as covariates based on social philosophies may contribute challenges with interpretation and development of meaningful results. An area of need where BIG DATA should be incredibly powerful is to expand the opportunities of underrepresented groups and women in Science, Technology, Engineering and Mathematics (STEM) fields. However, despite countless projects to expand participation of these groups in STEM fields there has been limited success. To be an engineer you must pass the 2nd semester of calculus. It is just one of those unwritten rules in engineering. However, how many STEM programs emphasize the underlying academic requirements and implementation of programs to facilitate the success in mathematics to expand inclusion of underrepresented groups in engineering?

We all have philosophies and theories we ascribe to in our day to day worlds. We all have philosophies and theories around educational models in both the K-12 and higher education systems. BIG DATA provides an academic plutonium that may be used to understand how to positively "intervene" on behalf of students and improve their educational outcomes. How can BIG DATA be used to improve the more global goal of improving graduation rates for all students in higher education?

KEY CHALLENGES OF USING BIG DATA IN HIGHER EDUCATION

What are the obstacles that prevent or inhibit more effective use of BIG DATA in higher education? In my experience there has been a series of consistently replicated errors in use of BIG DATA.

"KITCHEN SINK" DATA ARCHITECTURE

What are the elements that are really needed in a higher education data architecture? It is intriguing to hear academics, policy experts, and other educational stakeholders discuss the data required for various projects. Typically the models will include a listing of a plethora of variables. For example, in one meeting on college completion or graduation rates here are a list of variables discussed:

1) Gender (2)	2) Race (7)	3) Poverty Status (2)
4) Ethnicity (8)	5) College (7)	6) Degree (36)
7) Type of Student (3)	8) First generation (2)	9) ACT Level (4)
10) HSGPA (4)		

Note: The values in parentheses represents the number of levels

The total number of possible reports or combinations of reports from these 10 variables is 5,419,008. BIG DATA are available, but a lack of clarity on the underlying challenges to completing college convolute the ability to develop solutions in higher education. I believe a by-product of this type of model evolves from the inclusion and development of "kitchen sink" models of data architecture. As an example, the open source archival data sets provided by National Center for Educational Statistics (NCES) will have countless variables. Too often the result of research from use of these data sources, which are a great resource, is to include a plethora of variables in convoluted analytical models. As a note, is this example representative of analytics, data mining or bad research? Do the variables help answer the question of how to improve student graduate rates in higher education? As an additional comment, the number one reason students do poorly on an exam is the do not know the material. The psychometrics and science of test development since the mid-70's has virtually eliminated bias due to race, gender, and culture.

BUILDING VERSUS BUYING SOLUTIONS

Either the University of Arkansas (U of A) or the University of Nevada, Las Vegas (UNLV) have the resources to design and incorporate use of BIG DATA for analytics within their campuses. This is a situation that exists for most graduate degree granting institutions with the critical intellectual capital available on campus. However, higher education has a tradition of outsourcing the large scale data structures to companies such as ORACLE, IBM, etc. The actual storage of the BIG DATA systems require use of external agencies, but the interoperability of the "shells" or "software" placed over these systems can be prohibitive. For example, Figure 1 is a picture of the data extraction process required to obtain specific data from an Oracle system at the U of A.

The data architecture becomes a manifestation of the external company and their "solution" with the associated software for access and use. This can be prohibitive for use within higher education as the interaction requires countless additional facets and the ability to fastidiously develop your own data structures becomes prohibitive. My "fantasyland model" which requires a Single Source Solution (SSS) should provide the ability to develop subsets of data specifically for the purpose of targeted analytics projects. Figure 1 outlines the efforts to extract data as a step in the process for linking student transcript data with various demographic information to create an analytics data set. This particular figure only provides ½ of the actual steps and processes required to develop a data system, which still is incomplete for purposes of analytics. Further, data extracted via this protracted model limits the ability to write a simpler code and I am always concerned about the data quality and integrity as it is difficult to write data checks within this type of system.

Use of BIG DATA via large company based structures can be prohibitive. I requested a simple listing of variables, a process to select those for this project, and a mechanism to have them provided as a single or series of data sets where I could write SAS© code to merge and combine. SSS models facilitate two facets that are essential for improving use of BIG DATA in higher education: (1) Knowledge of data architecture (which includes understanding of data formats, clarity, etc.), and (2) a researcher interacting with data for their research. Too often data is almost mythical and never seen while the analytics are all point-and-click solutions with no relevance (Mulvenon, 2017).

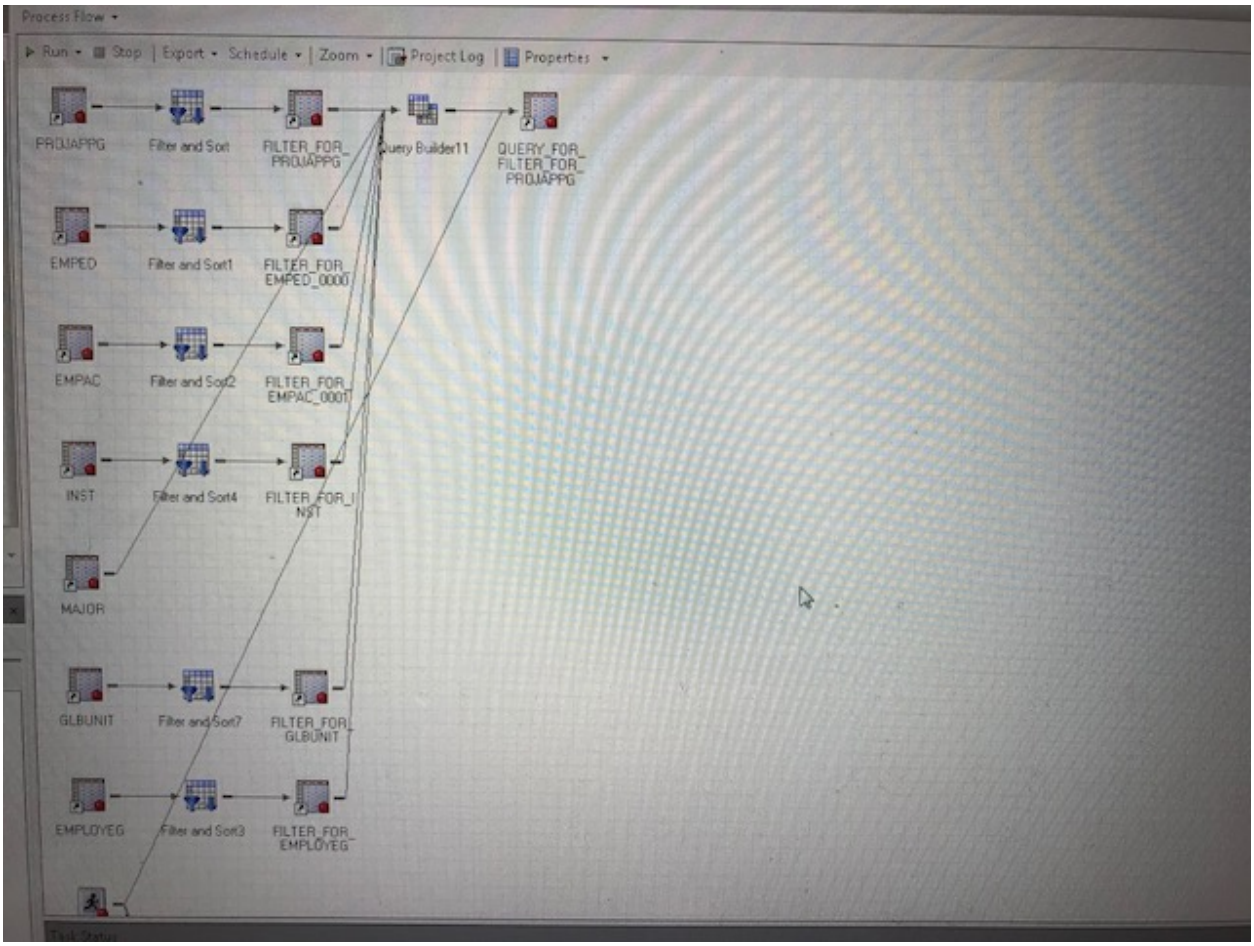


Figure 1. A mapping of data extraction process from an ORACLE System SAS Enterprise Guide®

LARGE DATA ARCHITECTURES VERSUS RESPONSE BASED DATA STRUCTURES

The ability to include information does not make BIG DATA valuable or representative of the solution to all matters in education. Are these large data architectures more valuable than specialized response based data structures designed to answer specific questions? And what do you really need to be effective? For example, what are those key data elements required to predict student success in college?

The data developed in Figure 1 is a “Reporting” data set and what is required is an “Analytics” data set. Figure 2 provides a simplistic example of a necessary transformation of data from “Reporting” to “Analytics” format required to complete advanced analytics. This transformation process can be protracted and if you rely on the external companies it may also be costly and result in significant time delays.

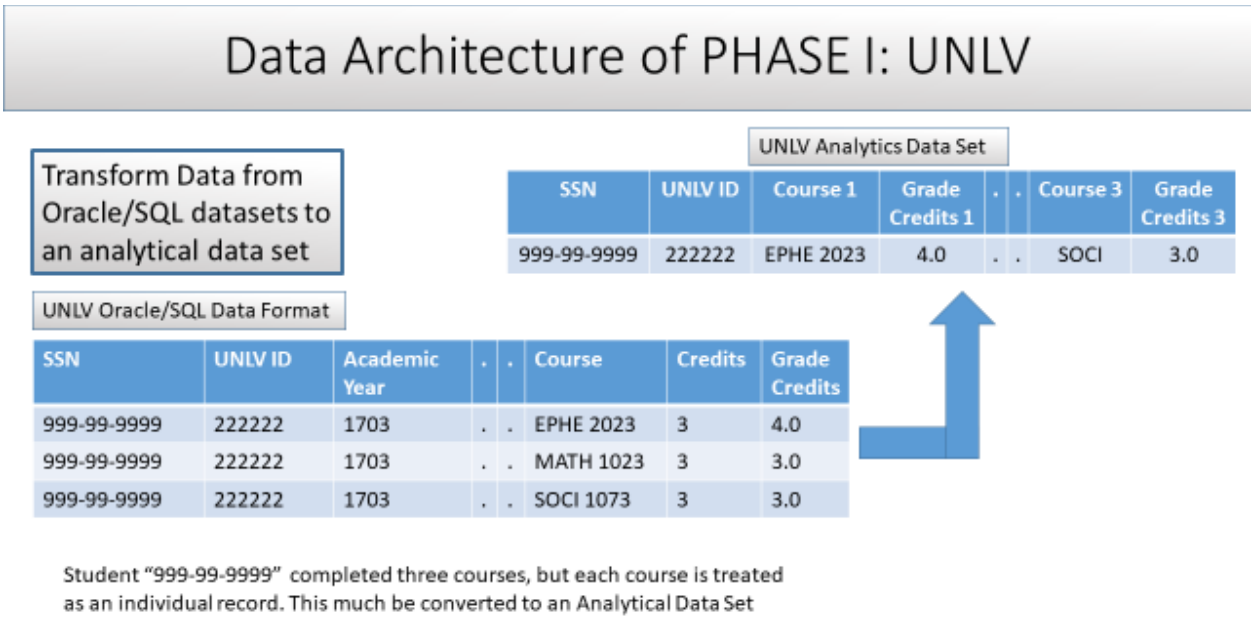


Figure 2. Data transformation from reporting to analytics data set

DIFFERENTIATING BETWEEN REPORTING AND RESEARCH

What is the difference in generating reports and conducting research? I recently had a discussion where I was attempting to explain the data architecture and tools within most IT structures versus the steps and process required for Hierarchical Linear Models (HLM). Figure 2 is a by-product of that discussion to explain the formats and the IT Toolbox of programmer options are extremely limited in their use for more conventional and advanced statistical procedures. But this discussion also extends to the difference between Data Mining and Scientific modeling. At times, I'm not sure if people are aware of the distinction and relevance of the two methods. Do graduate schools explain the difference between random events and consistent events? Does winning the lottery with numbers of 26, 27, 28, 29, 30 and 31 indicate to win the lottery just use these numbers? Does it suggest that winning is based on sequential numbers? Or is this an independent random event that has no meaning? How much of data mining is based on the former versus the latter explanation?

DATA MINING VERSUS SCIENTIFIC MODELS

A commercial I enjoy is where a golfer is using BIG DATA analytics to improve his golf game. The voice over in the commercial continues to discuss data mining and the ability to find information. What is the difference in data mining versus scientific models? One of my favorite courses in graduate school was Exploratory Data Analysis (EDA) using the seminal text by Dr. John Tukey (Tukey, 1977). The representation of most commercials on BIG DATA is that they are finding answers with the premise that "data mining" will find a needle in the haystack. The EDA course was transformative in how I became an analyst because I was "pushed" to creatively think about data patterns, underlying mechanisms in the systems, and move beyond the analytics. The data developed from studying data using EDA has always improved my analytics and the interpretation of outcomes.

I prefer the term EDA or exploratory data analysis to data mining because it reminds me that I am genuinely "exploring" data attempting to identify underlying mechanisms that may be incongruent. Identification of these aberrations provides the basis for developing scientific models to replicate and validate the consistency of the underlying mechanism. For the golfing example, if Mr. DeChambeau (the golfer in the commercial) identifies that 75% of the time he misses a fairway on the right, is this an example of effective analytics? YES, if he has the population of scores and he in fact misses the fairway

75% of the time to the right. If he concludes this is why he has not been playing well, is this accurate? NO! What is the average score relative to par for holes where he misses the fairway on the right versus left? Analytics is more than identifying a point estimate of an event (percent of fairways missed on the right) and confusing data mining/exploratory with more developed scientific models is problematic in higher education and our use of BIG DATA.

PHILOSOPHIES AS COVARIATES IN HIGHER EDUCATION

The seminal goal of research in higher education (or K-12) should be to improve/maximize the educational outcomes of all students. BIG DATA is often championed as the solution with the corresponding analytics that are possible with "point-and-click" technology of software packages. Figure 3 – 5 represent a few BIG DATA examples of common analytics mistakes made in higher education and the K-12 systems. Each figure has a significant problem in regard to the direct interpretation of the information, but can you identify the issue?

Educational Policy or Analytics Issue?

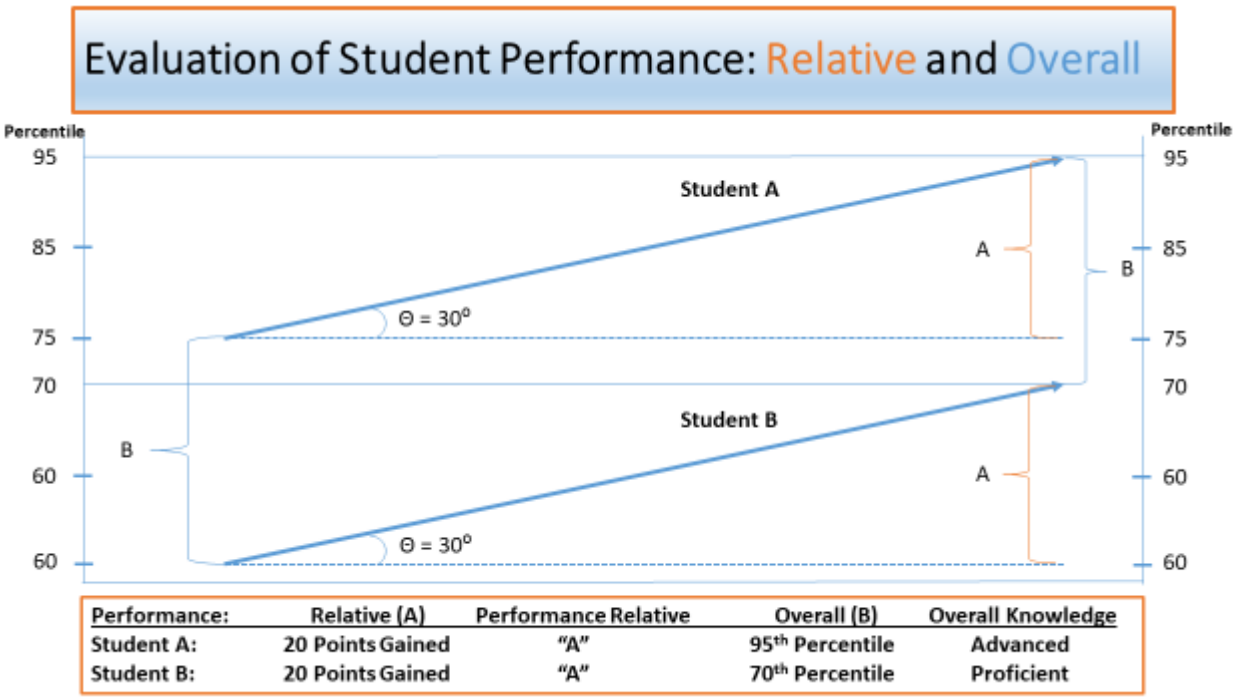
High School(s)	Literacy Proficient	Math Proficient	TAGG Literacy Proficient	TAGG Math Proficient	Graduation Rate	* ACT English	* ACT Math
School A (Needs Improvement School)	88.9%	92.8%	74.6	84.3%	85.2%	24.0	22.9
School B (Needs Improvement Focus School)	82.3%	83.5%	73.6%	73.6%	86.8%	24.3	23.4
Achieving High Schools (N = 15)	75.4%	72.0%	71.3%	69.3%	88.8%	20.6	19.7

Figure 3. Accountability data of three high schools in Arkansas in 2016

The issue in Figure 3 is not analytics, but one of validity. As you review Figure 3, notice the school classifications on the left and concurrently compare the Literacy and Math Proficiency scores. This is an example of accurate analytics via BIG DATA, but the policy applications of the scoring models create significant interpretation challenges and questions about Face Validity of the metrics. Too often in higher education, analytics are applied to poor policy models. I share this model because as people question results in Figure 3, they don't question the educational policy, they challenge the analytics. The graduation rates raise more global questions of the fidelity of the metric graduation consistent with the Key Question of this paper. Additionally, this type of example is consistent with the challenges of "point-and-click" analytics with limited thought given to the interpretability and meaningfulness of data (Mulvenon, 2017).

Figure 4 represents another example of a common interpretational problem with analytics. Can you identify the issue? If you were to conclude that both Students A and B improved by the same amount on the exam would you be correct? Yes and No! Students A and B both improve by 20 points on the exam, but the relative gain based on the distribution of scores produces two very difference conclusions. Analytical models have additional value due to variability in the systems, not the point estimates of gains or means. For example, is it harder to decrease your time to run a mile from 15 to 14 minutes or 8 to 7

minutes? Both represent one minute gains, but we can all understand the improvement from 8 to 7 minutes represents a much more significant reduction in time. Figure 4 provide an example where the gain of 20 points is much greater for Student A than Student B.



Performance:	Relative (A)	Performance Relative	Overall (B)	Overall Knowledge
Student A:	20 Points Gained	"A"	95th Percentile	Advanced
Student B:	20 Points Gained	"A"	70th Percentile	Proficient

Figure 4. A representation of challenges of interpreting relative versus overall performance

Figure 5 represents another example of the challenge of BIG DATA analytics in higher education and specifically the use of covariates in predicating student success. Too often we examine the issue of predicting success of students with the inclusion of poverty or race as a covariate in the HLM model.

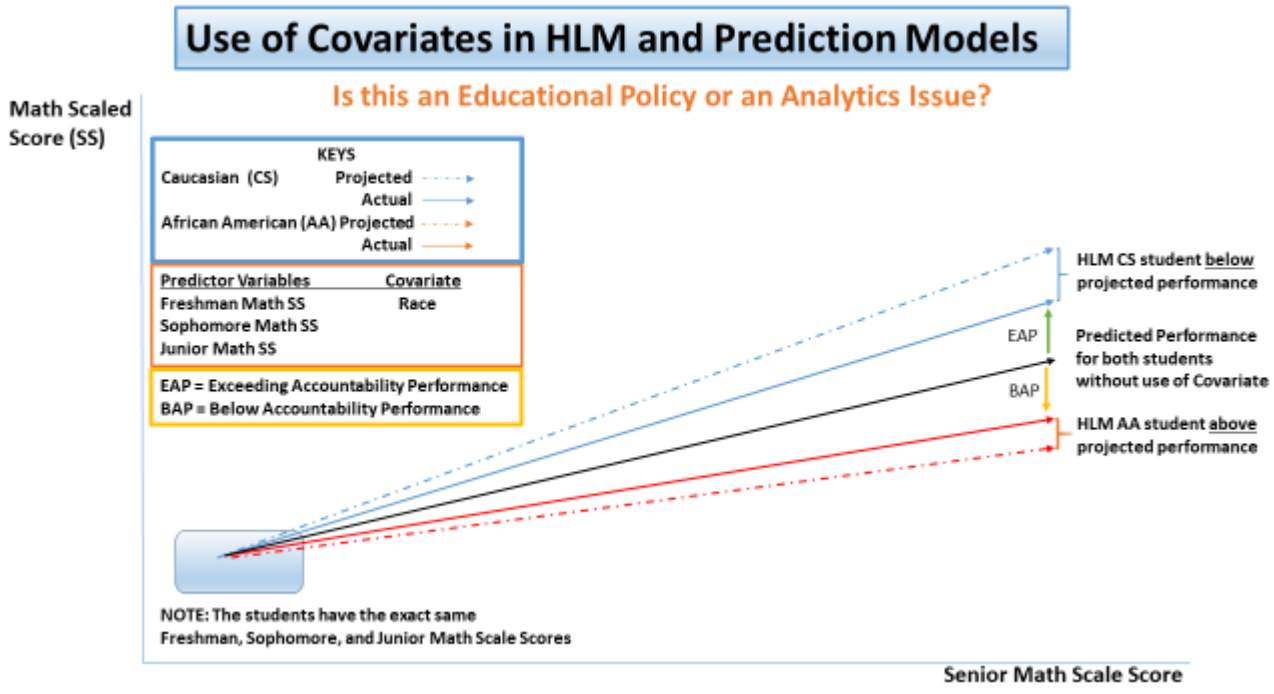


Figure 5. Example of HLM model with race as a covariate and policy challenges

The inclusion of race in Figure 5 will apply the historical performance of a specific race to each individual. In the present example, if Student A and B have the same prior performance it is anticipated they will concurrently have the same trajectory for future performance. However, due to the application of race, the historical academic performance is modified and the HLM model may not be accurate in their predicted performance (Mulvenon, 2010). Is this a policy or an analytics problem? I believe it is a combination of both, with the challenge of understanding the implications when including race (or poverty) as a policy issue; and the interpretation of the results as an analytics problem. A long understood anomaly of three years of academic performance data is it attenuates measurement error associated with race or poverty in prediction models. The use of this type of misinformation to develop policies or programs contributes to the more global examples of limited success of increasing graduation rates in higher education.

RECOMMENDATIONS AND POSSIBLE NEXT STEPS

SAS AS A SINGLE SOURCE SOLUTION

What SAS does provide is an analytical resources that extends well beyond what other companies may provide? A comprehensive data management structure, tools to transform data (proc transform), SQL interfaces, sub setting of data structures, and a comprehensive package of analytics procedures to answer your questions. I am an advocate of the single source system and have been an unrepentant use of SAS since 1988.

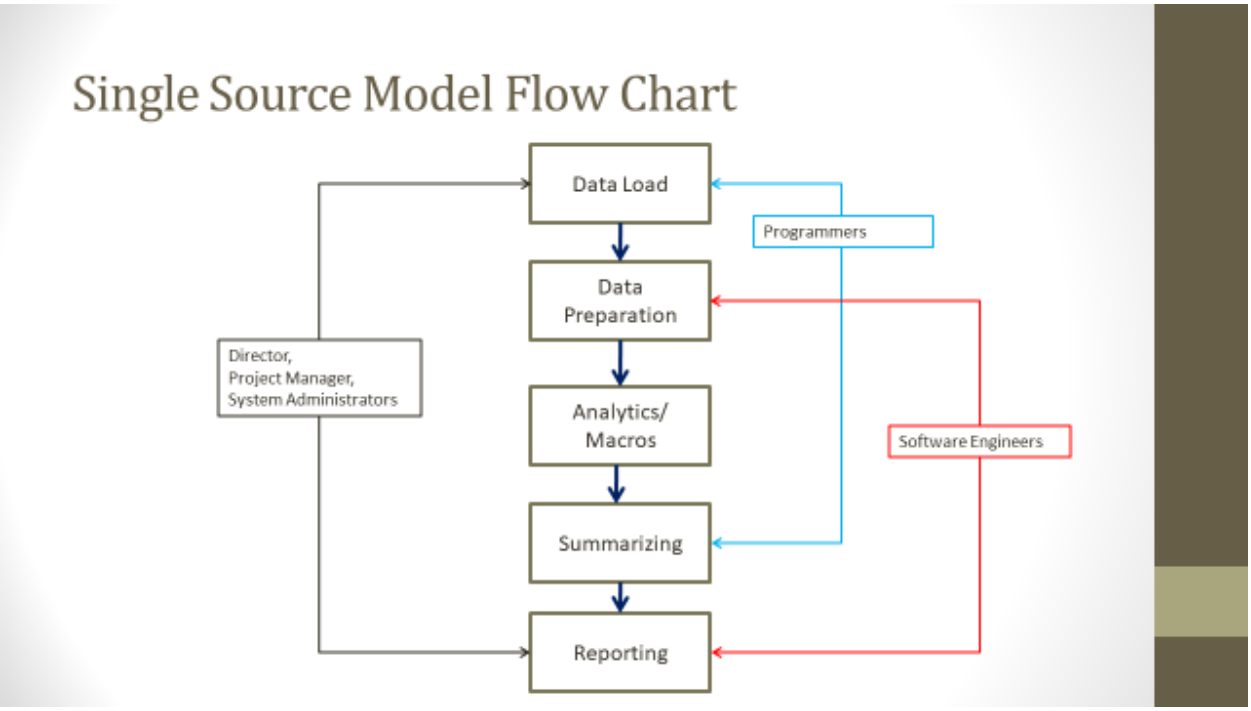


Figure 6. A single source solution analytics model

The blue line represents an element where programmers are required, red lines a role where software engineers (web programmers, etc.) may be invaluable, but the overall single source platform is the key. SSS models where data storage, procedures/transformation, analytics, and reporting are self-contained within a single software system are invaluable and reduces options to use of SAS©. I can laud the value of SAS© as an SSS model, and I have repeatedly, and those discussions consistently involve the comment "you can do all that in SPSS" or some other software. The reality is SAS© is the true SSS platform. My

personal theory as to why it is not purchased and used in this fashion in higher education is due to the fact IT Departments make software purchases. And IT programs enjoy their large dysfunctional data systems as a mechanism for job security in lieu of as a resource to facilitate use of BIG DATA.

IMPROVED KNOWLEDGE OF ANALYTICS: INSTITUTIONAL EFFECTIVENESS

The ability to complete various analytics via a software platform or access to BIG DATA are helpful in the more global goal to improve graduation rates. However, improving the effective use of BIG DATA and analytics requires a corresponding increase in an understanding of data architecture, statistics, and implications associated with policy. Figures 1 – 2 outline challenges of data development with BIG DATA platforms while Figures 3 – 5 demonstrate obstacles associated with interpretation of analytics. Until there is a commitment to invest the academic energy to grow understanding of these elements, use of BIG DATA will have a limited effect in higher education.

One approach I have advocated is the expansion of Institutional Effectiveness programs within higher education and K-12 leadership programs. The goal of these programs would be to provide a graduate certificate that emphasizes analytics, including data development and modeling. A seminal design philosophy of these program should be inclusion of data from the respective educational systems of students within the program.

CONCLUSION

BIG DATA has a **BIG** role to play in improving the educational outcomes for all students. A clear path to improving the use and success of BIG DATA in higher education is to focus on the value of this data and a recommitment to understanding the role it can play in finding answers. BIG DATA by itself does not provide solutions or represent the essential data to identify the answers. Improving outcomes requires developing researchers who understand how to use data and have a willingness to study the challenges in their efforts to identify solutions. Revisiting the Key Question “What information do we need to improve student graduation rates?” What BIG DATA do we need? I would suggest the BIG DATA are available, but the real challenges are the need to improve the scientific methods, our depth of understanding on the policy implications, and to limit the inclusion of philosophies as covariates.

REFERENCES

Mulvenon, S. (2010). Assessing Performance of School Systems: The Measurement and Assessment Challenges of NCLB, in Walford, Tucker, and Viswanathan (eds) *SAGE Handbook of Measurement*. London: SAGE Publications.

Mulvenon, S. & Bowman, S. (2015). An Evaluation of how the “Policies of K-12 Testing” Impact the Effectiveness of Global Testing Programs. Smith, W. (Editor) in *The Global Testing Culture: Shaping Education Policy, Perceptions and Practice.* Oxford Press: Oxford, UK.

Mulvenon, S. (2017). Improving the Evaluation of Higher Education: Understanding the Myths, Methods, and Metrics. Manuscript published in the proceedings of the SAS Global Forum 2017, Orlando, FL.

Tukey, J. (1977). Exploratory Data Analysis. Addison Wesley: Boston, MA.

RECOMMENDED READING

- *Base SAS® Procedures Guide*
- *The Little SAS® Enterprise Guide® Book*
- *SAS® For Dummies®*
- *Doing HLM by SAS® Proc Mixed*

CONTACT INFORMATION

Your comments and questions are valued and encouraged. Contact the author at:

Sean W. Mulvenon, Ph.D.
Associate Dean for Research and Sponsored Projects
University of Nevada, Las Vegas
702-895-4647
Sean.mulvenon@unlv.edu