Paper 3050-2019

The Transformation of Educational Data Using SAS®: Introducing Intervention Analytics

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

ABSTRACT

A goal of improved student outcomes, requires intervention analytics. The last few decades have seen a tremendous growth in the depth and clarity of the role of big data in education, including a greater emphasis on using analytics to improve student outcomes. Yet despite the genesis and transition from data-driven decision-making to actionable data to visual analytics, has there really been a systemic transformation in the effective use of analytics in education? This paper introduces the use of intervention analytics to improve the educational outcomes for all students. Key to the process is reliance on the single-source SAS® model that transforms data, integrates an innovative analytics model, identifies unique educational interventions, and transitions information to web-based platforms.

INTRODUCTION

A consistent challenge in higher education is to identify methods and techniques to improve persistence and graduation rates for ALL college students. “Persistence” is a frequented term in the educational context used to combine the concepts of continuous enrollment and progress toward the completion of a degree. “Degree Completion” and “Graduation Rates” are both terms used to refer to a similar event, with degree completion often a dichotomous event (1, 0) used to develop the numerator when computing the graduation rate(s). The basis for sharing and providing a definition here is to outline two consistent metrics used as proxy measures when evaluating the effectiveness of higher education institutions.

Higher education administrators must consistently address issues of persistence and graduation rates with legislators, policymakers, educational advocates, special interest groups, etc. Further, these simple metrics are readily reported in newspapers and magazines, as well as referenced in publications as indicators of performance for the general public to instill further interest in the performance of higher education institutions. This consistent public discussion is a healthy method of transparency, but the natural next question of “what are you doing to increase persistence in graduation rates?” is much more complex. This leads to the internal response question by higher education administrators, “what are we doing?” and extends across other levels of the academic institutions, e.g., Student Support Services, College Leadership, or Faculty. Too often, responses to these questions are focused on producing statistics of various percentages, entry scores, financial aid, or social covariates of poverty, access, and academic preparation of underrepresented students. Although all of these elements may contribute to lower student persistence and graduation rates, none are proactive in addressing the academic “roadblocks” that impact many students.

CURRENT EARLY DETECTION SYSTEMS. An all too common remedy for improved outcomes is for higher education systems to invest in “early detection systems” that identify academically at-risk students. These systems are designed with great intentions, but considerations must be made for the underlying academic metrics used to identify a student as “at-risk”. What even are they? Across several of the early detection systems, the key indicator for determining an “at-risk” student is simply a mid-term grade of C or lower in a course. This type of simplicity in a metric creates two challenges: (1) myopia in evaluating

student persistence, and (2) mistrust in metrics by those required to use this information. Attempting to overly simplify complex multivariate problems with univariate solutions fails to appreciate the consequences of these metrics. This metric may be simple to explain and may impress those in education who prescribe to the "I have to understand it to use" claim, but in reality it marginalizes the field of analytics. This ultimately results in a mistrust of analytics and false projection that analytics "do not work." For example, does a midterm grade correlate with the end-of-course grade? Yes, very highly. Thus, when a student receives an "F" as a mid-term grade and the "Early Detection" system provides an alert, it is most likely too late to provide interventions.

INTERVENTION ANALYTICS. The goal of this paper is to outline a single source SAS model for development of "Intervention Analytics!" Intervention Analytics are designed to provide information that is proactive in identifying students who are at-risk *a priori* to academic difficulties. The program aims to provide intervention models to support academic progress, and utilizes historical student pathways for developing probability models of success based on the academic major. Presently, the concept of Intervention Analytics is still evolving, but this paper will outline the developmental steps for a SAS single source solution and present several prototypes of models currently in various stages of development.

UNDERSTANDING THE NEED OF HIGHER EDUCATION

The "NEED" in higher education is predicated on the specific audience requesting information or data, including: (1) administration, (2) recruitment, (3) faculty, or (4) student support services. Each of these audiences requires specific data, but also all seek to improve persistence and graduation rates. Many educational dashboards include simple data with various levels of disaggregation (Race, Gender, First Generation, etc.), including:

- Enrollment
- Retention
- Graduation Rates
- Degrees Awarded
- Student Retention

Each of these variables provides important information, but they all have a few things in common: (a) they are descriptive, (b) they employ post-hoc analytics, and (c) they provide no information on a formal action that may improve student outcomes. The purpose of Intervention Analytics is twofold: (1) to serve as a method to proactively identify students who may be at-risk academically based on their majors and (2) to identify effective intervention models to concurrently employ as students are progressing through their degree programs.

COLLEGE READY. The term college ready is one of the most ubiquitous terms used in the K-12 and higher education systems. What is "college ready?" The American College of Testing (ACT) provided an interesting overview of College Ready in their ACT Research Report Series 2014 (5) *Broadening the Definition of College and Career Readiness: A Holistic Approach*. A challenge to this definition is failure to contextualize college ready in terms of a specific degree program. Is college ready the same for an education major as it is for a physics major? Could someone be ready for one degree program, but unprepared for success in another degree program? ACT and many other organizations, including the U.S. Department of Education, provide omnibus definitions of college ready typically focused on more global definitions of first-year success or likelihood of a grade in specific freshman courses. An element of this study will be to outline a more formal academic process for identifying likelihood of success and "college ready" status relative to specific degrees.

KEY PHASES OF THIS STUDY

For the purposes of SAS and the research surrounding this conference, this study will be presented across three phases.

PHASE I: DATA ARCHITECTURE AND DEVELOPMENT. The Office of Information Technology (OIT) at UNLV has been instrumental in helping to design and articulate the use of SAS as a single source solution. Figure 1 represents the dynamics of developing a single source solution within a university research environment.

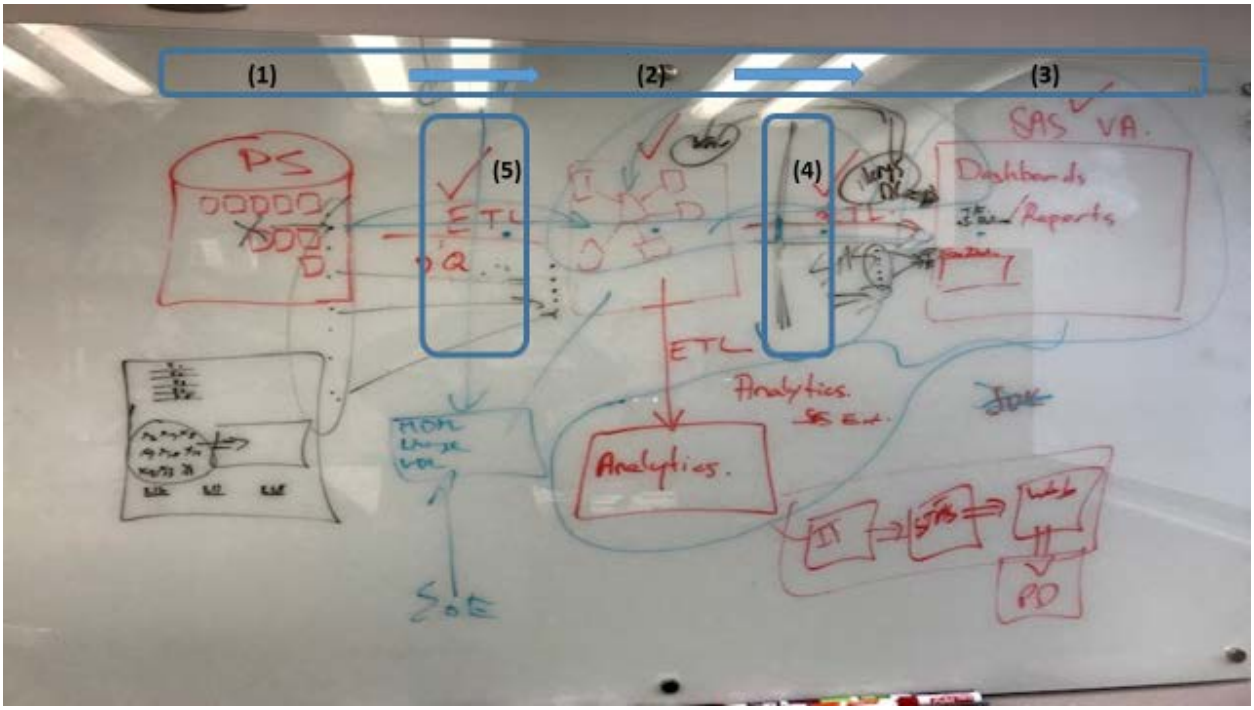

Figure 1. A Prototype of Data Architecture at UNLV.

ANALYTICS IN SINGLE SOURCE MODELS FOR INTERVENTION ANALYTICS

In Figure 1, data flows from (1) to (3), with (1) PeopleSoft (PS) or a "data lake" where all data resides, (2) a data quality phase and analytics within OIT, and (3) a SAS VA or Visual Analytics. Figure 1 represents three entities at UNLV: Data Warehousing (Mr. Brett Burnes), Enterprise Solutions, (Mr. Kivanc Oner), and Research Faculty (Dr. Sean Mulvenon) all designing a solution that embraces collaboration for developing data access and use in a University.

Item (4) and (5) are essential in understanding the progress and development of a single source SAS solution within UNLV. This is the part of the process engineering where I advocate the value of SAS as a solution. The identification of (4) is to represent the current place in the data flow were individualized data development occurs at UNLV. As part of this study, data has flowed from (1) to (2) and within (2) I have represented (4) where I am actively developing the analytic datasets for Intervention Analytics and use in SAS Visual Analytics. The line next to (5) represents a spot where I advocate the introduction of SAS as a solution in data architecture including the ability to complete additional data cleaning,

preparation, modification and, as Mr. Burnes would say, act as “a single source of truth” for the UNLV data structure. Figure 2 provides a 2nd representation of this relationship.

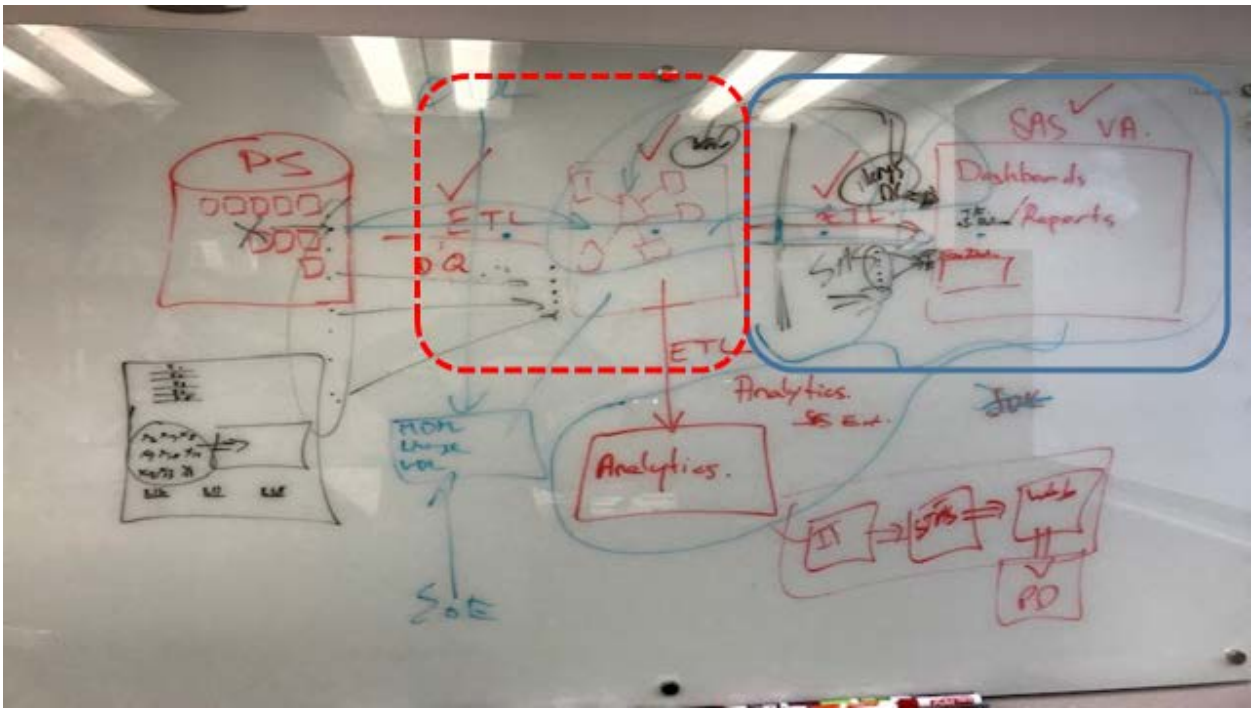

Figure 2. Outlining Current Use of SAS versus Possible Single Source Model.

The outlined blue box represents current use of SAS at UNLV with the dotted red box representing the possibilities for future single source solution design. The data development toolbox of SAS provides so many more options and possibilities for creating systematic data architecture to support data needs and campus wide enterprise solutions. Element (4) in Figure 1 still exists in Figure 2, as academic faculty will always require the ability to modify data for specific research efforts (i.e., Intervention Analytics). A successful model requires the support of Mr. Burnes and Mr. Oner in designing and developing these types of integrated single source solutions. In a perfect world (or at least my vision of a perfect world), a comprehensive data center would be represented by (2) in Figure 1.

PHASE II: DEVELOPING STATISTICAL MODELS FOR INTERVENTION ANALYTICS

Raw transcript data provided via OIT and Enterprise Solutions at UNLV represents the foundation for the development of Intervention Analytics for education. I commonly refer to data in two forms: (1) raw and (2) processed. Raw data is the form data that are provided to you and represent actual values as reported, collected, or submitted. In short, raw data is data as it has been stored and untouched. Processed is data that has been cleaned and may have additional metrics included that have been computed for your convenience for use in other calculations (e.g., statistical means or totals). Many times you will receive data that is a combination of both raw and processed. For example, if you compute the semester GPA or cumulative GPA of student performance, this represents processed data that is computed from static raw data from course performance. Figure 3 provides an example of select UNLV student transcript data that is static.

Variables in Creation Order

#	Variable	Type	Len	Label
1	college	Char	30	Student College
2	birthdate	Char	10	Student Birthdate
3	sex	Char	1	Student Gender
4	ethnicity	Char	20	Student Ethnicity
5	country_of_origin	Char	30	Student Country
6	state_of_origin	Char	30	Student State
7	last_high_sch_name	Char	20	
8	highest_actm	Num	8	Highest ACT Math
9	highest_acte	Num	8	Highest ACT English
10	is_full_time	Char	1	Student Enrollment Status
11	is_pell_eligible	Char	1	Proxy of Financial Status
12	acad_plan	Char	12	Student Academic Plan/Major
13	degree_granted	Char	20	Degree Earned
14	grad_date	Char	20	Date of Graduation

Static Data Elements that are "tagged" to every record of data provided on courses attempted/completed by the student. These data are "STATIC" in they will not change for the individual student during their period in college

Figure 3. Static Data Elements in Transcript Data at UNLV.

These data elements represent some of the important demographic data employed in the development of Intervention Analytics. For example, "is_pell_eligible" helps identify if a student is from poverty, "last_high_school_name" is part of the data on the past performance of the student in the K-12 system, and "highest ACTM" is academic data that is invaluable in providing an objective measure of the overall academic preparation of the student.

Dynamic Data Elements because they will change for each course completed by a student

Variables in Creation Order

#	Variable	Type	Len	Label
1	strm	Num	8	Academic Year and Semester
2	subject	Char	20	Course Identifier by Subject
3	catalog_nbr	Char	20	Course Catalog number
4	class_section	Num	8	Course Section Number
5	instructor	Char	20	Course Instructor
6	official_grade	Char	3	Course Grade
7	semester_gpa	Num	8	Semester GPA for Student *
8	cum_gpa	Num	8	Cumulative GPA for Student *

Key Dynamic Variables for student grades

* A Data Check Element

Figure 4. Dynamic Data Elements of Transcript Data at UNLV.

Figure 4 provides examples of the type of transcript data associated with student grades earned while attending UNLV. More importantly, this information will provide essential information for the identification of the "pathways to success" in Intervention Analytics. Additionally, variables such as "semester_gpa" provide important information to cross-validate the transformation of data from an oracle/sql format to what is referred to here as

a “processed” data set, but called an analytical data set. Figure 5 provides an example of this data transformation.

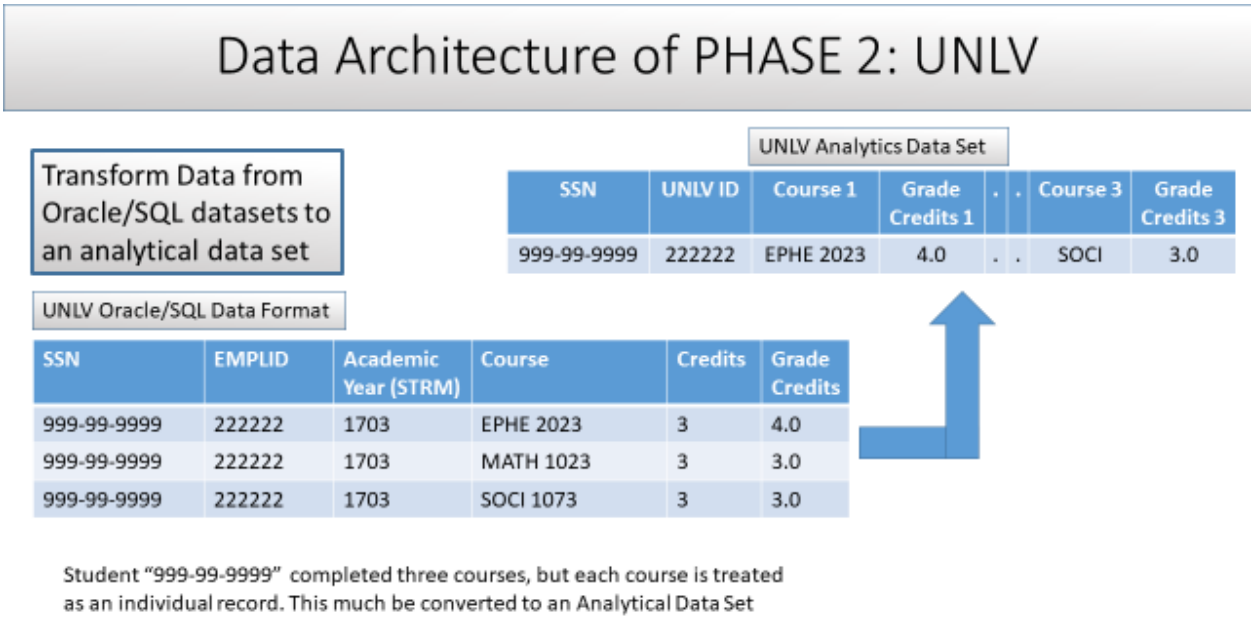

Figure 5. Transposing SQL data to an Analytical Data Set.

Transposing of transcript data allows for more advanced statistical models to be developed and completed in Intervention Analytics. Some of the possible analysis include:

- Proportional Hazards Regression (PROC PHREG)
- Hierarchical Linear Models (PROC MIXED)
- Structural Equation Models (PROC CALIS)
- Doubly Multivariate Repeated Measures (PROC GLM)

These statistical procedures represent just a few of the advanced statistical procedures that are possible and may be employed in Intervention Analytics.

PHASE III: INTERVENTION ANALYTICS

Intervention Analytics expands past the traditional metrics used in reports for higher education. For example, Figure 6 provides a listing of standard reports available at UNLV.

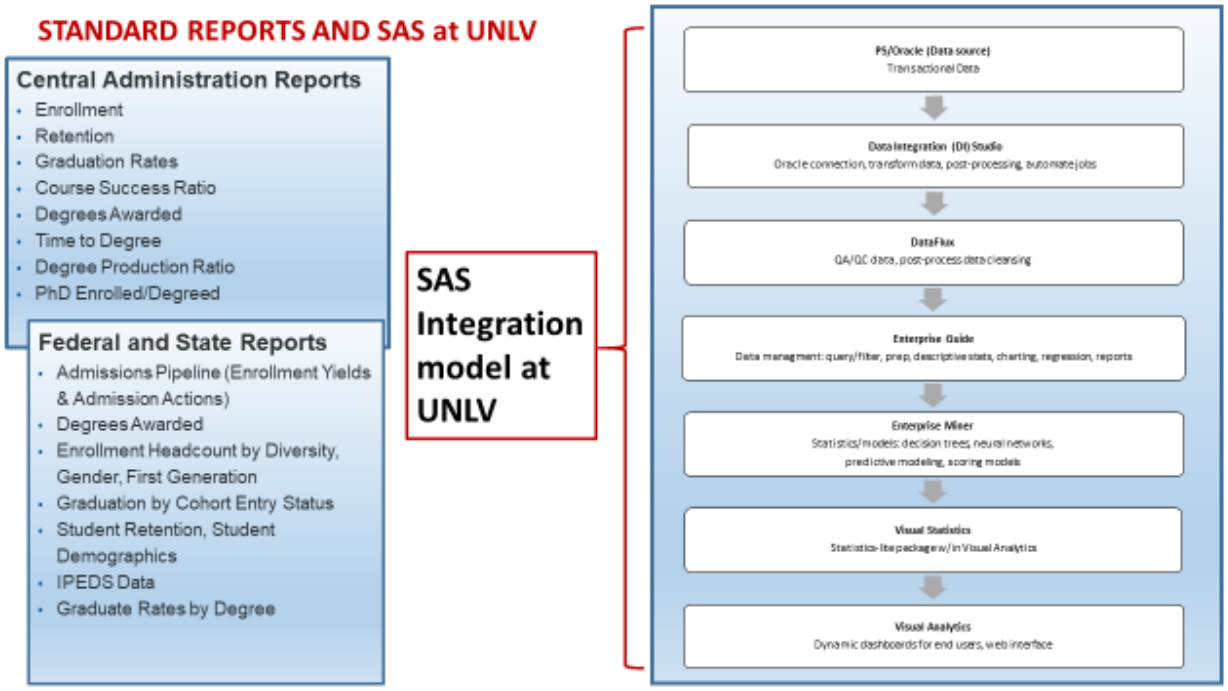

Figure 6. Standard Reports and SAS Data Flow Model at UNLV.

Many of the standard reports identified in Figure 6 are created via a SAS Integration model at UNLV. The transition from PS/Oracle to Visual Analytics is also represented, albeit simplistically, in Figures 1 and 2. A single source solution makes this data "flow" and the produces standard report models that are incredibly efficient. More importantly, the ability to update these reports is also automated via this system of "flow" of data and reporting. However, standard data storage formats which are convenient for reports is not effective for more advanced analytics and requires the transforming of data to analytic data sets for Intervention Analytics. Further, though many may interpret such reports as helping improve student persistence and graduation, in reality these reports are static and are just that … reports.

DEVELOPING A PROCESS FOR INTERVENTION ANALYTICS

The overarching goals of Intervention Analytics are to improve persistence and graduation rates, but what is the process required to attain these goals? It is not development of more standard reports and rather requires a better understanding of student pathways to success/failure in college. The first step is to better understand potential academic "road blocks". Consider this model from engineering as an example:

Engineering has a graduation rate of less than 40% which is low and needs to be improved, but what are the specific academic challenges associated with academic success in this college? For example, what are the dynamics around a student passing/failing in their effort to attain an engineering degree? The academic transcripts and performance pathways of those successful/unsuccessful students provided the foundation for developing Intervention Analytics. Using a "Decade of Data", an examination of engineering students from fall 2009 through spring 2019 was completed.

Step 1: Creating Academic Cohorts

Three cohorts of students were created: (1) Exploratory (fall 2009 – spring 2012), (2) Confirmatory (fall 2012 – spring 2016), and (3) Intervention (fall 2016 – current). The basis for these cohorts was simple. Students in the Exploratory cohort, effectively having completed their academic careers at UNLV, were studied to understand successful/unsuccessful pathways. The Confirmatory cohort represented students who may still be active at UNLV, but can be used to cross-validate trends and patterns of successful/unsuccessful students in engineering at UNLV. Finally, the Intervention cohort represented those students where proactive academic support may be "targeted" based on Intervention Analytical models.

Step 2: Preliminary Analytical Models at UNLV

Approximately 1,956 students have been enrolled in computer systems engineering since the Fall 2009 semester, with 422, 202 and 1,332 in the Confirmatory, Exploratory, and Intervention Cohorts respectively. Approximately 31% and 37% of students graduated in the Confirmatory and Exploratory Cohorts. The actual gain of 6% may sound minor, but this represents an increase of 19.4% (37/31= 1.194), which is excellent.

Survival analysis revealed that most students were active in pursuing their engineering degree for three to four semesters, until hitting a "road block" course identified in this analysis as Calculus. A "road block" course represents a course where students have lower grade point averages (GPA) which subsequently lowers the persistence rate in engineering (i.e., failing this course leads to either withdrawing from college or transferring to an alternative degree). If the persistence rate declines then concurrently the possible graduation rate would decline for the 6-year period used to measure this metric. Anecdotally, this "road-block" makes sense. I had a colleague at the University of Arkansas who was the Director of the Freshman Engineering Program and who would always congratulate students on becoming an engineer when they passed Calculus II (the 2nd semester of calculus). As Engineering seeks to improve their persistence and graduation rates it is paramount this "road block" be examined and additional academic remedies, resources, or interventions identified.

Intervention Analytics used advanced statistical method to identify the academic paths/road blocks of former and current successful/unsuccessful students and to proactively provide assistance and remedies. This simplistic model represents a seminal goal of Intervention Analytics, which is to use these more advanced methods to improve educational outcomes for all students. This simple Survival Analysis revealed that likelihood of success increased dramatically if a student remained academically persistent into their junior year, i.e., if we can keep students enrolled through the beginning of their junior year their likelihood of completion (graduation) increases dramatically.

Step 3: Implementing Interventions

Using the results from above, what are the remedies that can be provided to improve student success in Math 182 or Calculus II? At UNLV, a proactive program is being developed to remediate mathematics preparation, which includes an evaluation of the student success in Math 181 or Calculus I. A follow-up analysis revealed that in one section (Section XX1) of this course, the cumulative GPA for over six hundred students was approximately 1.2 or for every five grades assigned, there were 4 D's and 1 C. In contrast, in another section (Section XX2) of the same course the cumulative GPA was 2.4, approximately double, for the almost 1,000 students who completed it. Further examination

revealed that only 11% of students in Section XX1 completed Math 182 or Calculus II with a GPA of 2.38, but 17.4% of students in Section XX2 completed Calculus II with a GPA of 2.47. The sections of Math 181 and 182 mattered for the persistence of students in engineering. Why? This is something that is being currently discussed, as this is evidentially not random across over 600 and 1100 students in Sections XX1 and XX2, respectively.

Additionally, a Learning Management System (LMS) has been implement for Math 181 (Calculus I) to improve the preparation and analytics. The LMS is part of a Learning Analytics Initiative supported and funded by Provost Chase at UNLV. The goal of this system is to monitor student access with various resources, including tutors, course materials, syllabus, practice exercises, and course study guides. All of this information is collected electronically and is being used to develop interventions models to support student persistence and success across multiple courses at UNLV. This too is a form of Intervention Analytics, but with counts and monitoring access of students to instructional resources. In the next few months this will be integrated with more advanced methods of statistical modeling.

CONCLUSION

Intervention Analytics represents an academic effort to improve methods/approaches for increasing student persistence and graduation rates in higher education. Several additional models are currently being developed and cross-validated for use at UNLV. Additionally, efforts are under way to collaborate with Clark County School District to create a data pipeline in which High School Transcript data will be merged with UNLV transcript data. Figure 7 represents this process and the goals of this CCSD/UNLV data initiative.

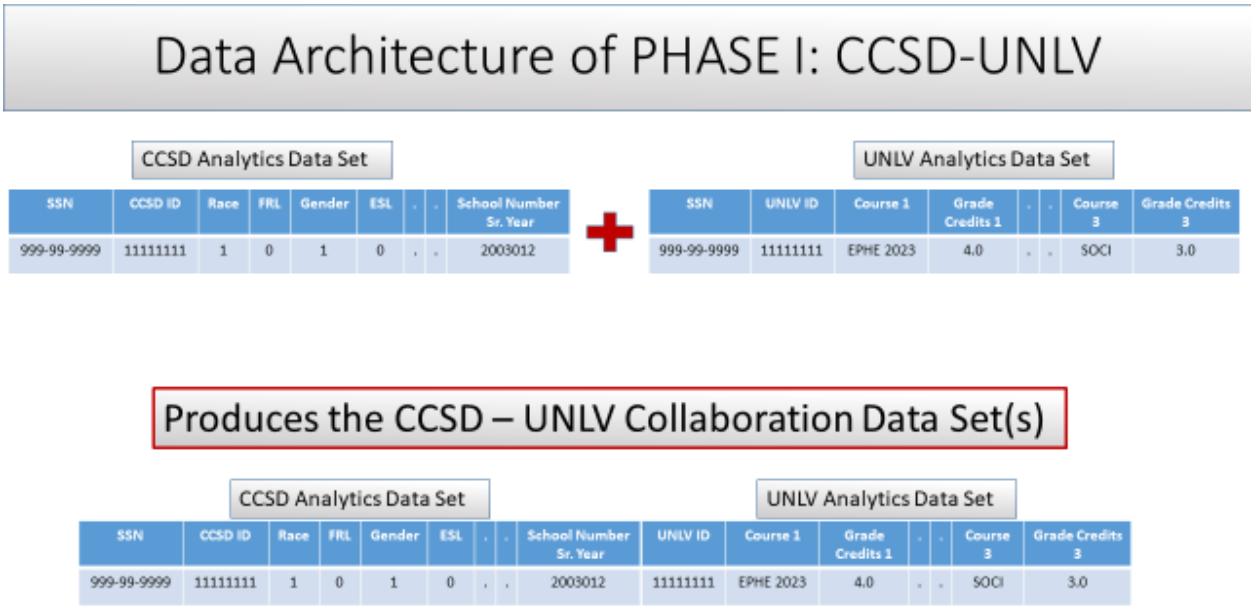

Figure 7. The CCSD/UNLV data initiative.

An amazing facet of this collaborative will be the ability to expand and identify the critical academic intersections associated with success in college. Further, the linking of this

information will provide an opportunity to more formally define “College Ready” relative to numerous academic degrees available on campus. The recent Executive Order on Higher Education by President Trump on March 21, 2019 will require this information relative to income expectations and a return on investment for students attending college.

Future Research. This study was intended to introduce Intervention Analytics at UNLV. Much more research and innovation is currently being investigated with the data collaboration with CCSD only increasing the types of Intervention Analytics to be developed. It is anticipated this will be integrated into a more formal web application for parents to better understand higher education, costs, return on investment, and important academic benchmarks associated with success in various degree programs. Finally, this information will be linked to more formal academic interventions and strategies for success.

Closing Comments

The success of this project is driven by developing an effective collaboration within UNLV and requires the efforts of Mr. Oner and Mr. Burnes to provide the essential data elements. Further, the use of SAS and the single source solution is beneficial in driving this model to make this type of research more efficient and valuable to the academic success of all students.

REFERENCES

ACT (2014). Broadening the Definition of College and Career Readiness: A Holistic Approach. Accessed March 8, 2019. http://www.act.org/content/dam/act/unsecured/documents/ACT_RR2014-5.pdf

RECOMMENDED READING

- *Base SAS® Procedures Guide*
- *SAS® System for Mixed Models®*
- *Modeling Longitudinal and Multilevel Data: Authors Little, Schnabel, & Baumert.*

CONTACT INFORMATION <HEADING 1>

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

Sean W. Mulvenon, Ph.D.
University of Nevada, Las Vegas
702-895-4647
Sean.mulvenon@unlv.edu