ANALYTICS THAT INFORM THE UNIVERSITY: USING DATA YOU ALREADY HAVE

Charles Dziuban
Patsy Moskal
Thomas Cavanagh
Andre Watts
Center for Distributed Learning
University of Central Florida

ABSTRACT

The authors describe the University of Central Florida's top-down / bottom-up action analytics approach to using data to inform decision-making at the University of Central Florida. The top-down approach utilizes information about programs, modalities, and college implementation of Web initiatives. The bottom-up approach continuously monitors outcomes attributable to distributed learning, including student ratings and student success. Combined, this top-down/bottom up approach becomes a powerful means for using large extant university datasets to provide significant insights that can be instrumental in strategic planning.

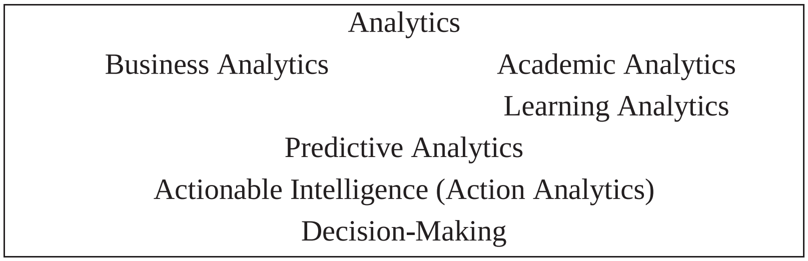
KEYWORDS

action analytics, big data, top-down/bottom-up, online courses, impact evaluation, actionable research

INTRODUCTION

Literally, the term analytics refers to the *science of logical analysis* [1] and is not a new concept. The use of analytics in business has developed into a common practice, driven in part by advances in technology, data storage, and data analysis techniques, including predictive modeling, that allow for complex computations with very large data sets. Companies such as Amazon.com, iTunes, and Netflix store members' clicks, views, and orders and "mine" these data to extract meaningful information, used to influence customers with recommended choices, additional options, and advertisements. The more informed a company is about a consumer's purchases, the better it can motivate them about the possibility of further choices and options that they might not have found otherwise. Intuitively, this makes sense in today's electronic world as the notion of "shopping" online takes on a challenge of logarithmic proportions without guidance and direction. In an effort to influence sales, shrewd businesses prefer to guide and direct their customers toward more of their own products.

While analytics is widely used in business, the use of analytics in higher education is still in its infancy. In fact, the field is so new and varied that van Barneveld, Arnold, and Campbell [2] reviewed the literature in an effort to determine a common language in the flood of applications and articles currently using the term "analytics." They found many variations in the terms and definitions, but proposed their own conceptual framework in an attempt to position learning analytics within a business and academic domain (Figure 1).



Analytics

Business Analytics   Academic Analytics

Learning Analytics

Predictive Analytics

Actionable Intelligence (Action Analytics)

Decision-Making

Figure 1. Conceptual Framework of Analytics

Siemens et al [3] differentiate between *learning analytics*—which focuses on data related to learners primarily to improve student success—and *academic analytics*—which is aimed at improving organizational effectiveness through learner, academic, and institutional data. They propose an integrated learning analytics platform that provides an open infrastructure for researchers, educators, and learners to develop new technologies and methods.

Many of the current applications of analytics in higher education are focused on Learning Analytics or Academic Analytics. The emphasis is on using very large data sets to inform faculty, students, and administrators when students are at risk and, in some cases, suggest possibilities for improving their performance within a course. Approaches vary widely both in terms of data used to develop models and application of data to inform students, faculty, and the institution. However, a number of researchers have identified models and/or applications that have shown promise on their campuses.

One of the most familiar systems was developed by Campbell [4] and his colleagues at Purdue. They have had success with their use of analytics in identifying students at risk and employing alerts to make them aware of the status within a course, utilizing student information, learning management system data, and student grades to form a model of course success. *Signals* notifies students using a traffic light to identify if they are doing well (green light), in danger (yellow light), or at risk (red light) for failing in a course. Campbell found, however, that identifying a student as at risk was not sufficient because those students who need the most help also are those that ignore the signals and do not take advantage of resources that might help them improve. Recently, the *Signals* application was acquired by Sungard and is now being marketed to campuses as a means to potentially help improve student course success.

University of Maryland, Baltimore County (UMBC) [5] found that students who have a grade of D or F used the course management system on average 39% less than higher performing students. Fritz [5] and his colleagues created a "Check My Activity" (CMA) tool to allow students to monitor their progress in Blackboard compared to their classmates. Initially, few students used the system, but when the campus developed a marketing campaign to advertise CMA and made sure the tool was easier for students to both find and use, students did increase their usage. In addition, students' behavior with the course management system also changed and they became more active participants in interacting with course materials through the system.

Goldstein and Katz's [6] survey on the use of analytics in higher education resulted in a framework of five stages: data extraction, performance analysis, what-if decision support, predictive modeling, and automatic process triggers (such as alerts). Further, they found that three factors contributed to an institution's successful use of analytics: effective institutional training, staff skilled in understanding and applying analytics, and leaders committed to evidence-based decision-making. They found most universities using analytics for admission prospects or to identify at-risk students.

Campbell and Oblinger [7] suggested considering analytics as an engine that guides the decision making process in five steps: capture (data), report (trends), predict (with a model), act (intervene), and refine (the model and process). Also, they stressed the importance of organizational readiness in terms of the support required to successfully implement learning analytics into the culture of the institution. The possibility of using analytics to oversimplify what is a complex system of student variables that create a successful course, program, or degree experience is a concern with this approach, which further typifies why clear goals, objectives, and support are critical [8].

Much has been written about the potential of learning analytics at the course level [5, 9, 10]. Certainly, there is demonstrable value in being able to identify "at risk" students and proactively intervene to get them back on track. Likewise, mining through the usage of instructional tools to understand effective technology-based teaching strategies can yield important trends that can inform future course development. However, the same potential exists to leverage data analytics strategically at the institutional level. Being able to examine macro data across departments, colleges, and the larger university can reveal institutional opportunities that might have otherwise remained hidden.

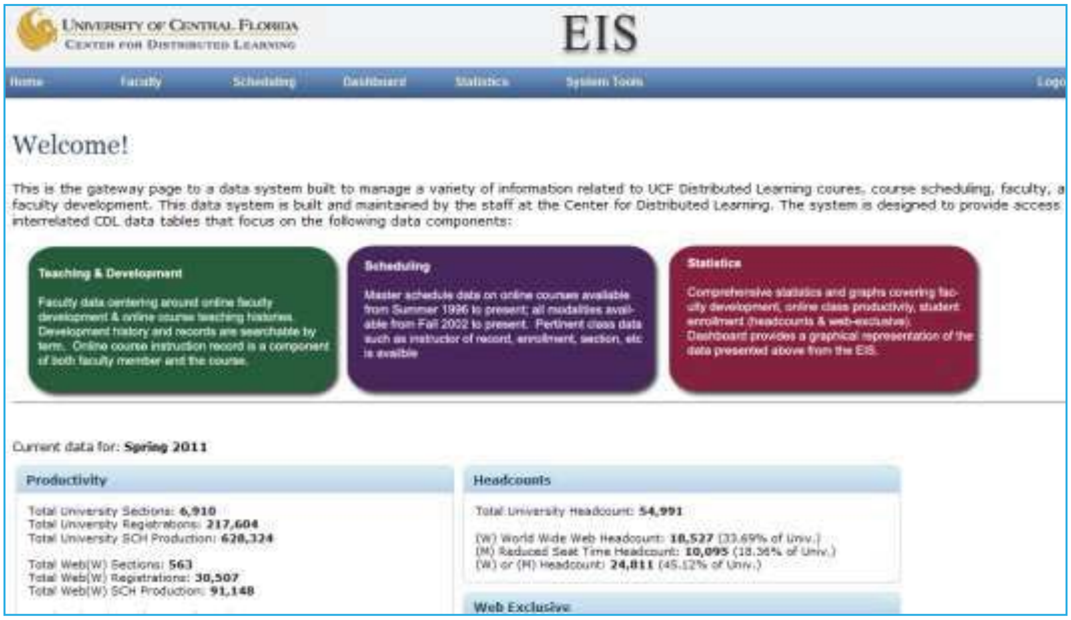
ANALYTICS AT UCF

At the University of Central Florida (UCF), the Center for Distributed Learning (CDL) is responsible for overseeing this institutional lookout of what is a combination of what van Barneveld, Arnold, and Campbell [2] have called "Business Analytics" and "Action Analytics." To do this, we maintain simultaneous "top-down" and "bottom-up" views of what is happening across the university related to distributed learning (completely online, blended, and lecture-capture courses and programs).

TOP-DOWN PERSPECTIVE

From a top-down perspective, CDL has developed a proprietary data mining platform called the Executive Information System (EIS). The EIS (Figure 2) began as a skunkworks project to better automate CDL's ability to answer various questions from senior administration. Over time it has grown into an indispensable tool in the management of a high-growth online learning initiative at the second-largest university in the nation. Among the diverse set of functions the EIS offers are:

- manages faculty development scheduling and credentialing to teach online.
- maintains historical faculty teaching records across all modalities, as well as master course schedule data.
- tracks productivity data (e.g., registrations, sections, student credit hours, etc.) by campus, college, and modality.
- permits program tracking for regional accreditation and state governing board reporting.
- monitors student demographics.


University of Central Florida
Center for Distributed Learning
EIS
Home
Faculty
Scheduling
Dashboard
Statistics
System Tools
Logo

Welcome!

This is the gateway page to a data system built to manage a variety of information related to UCF Distributed Learning coures, course scheduling, faculty, a faculty development. This data system is built and maintained by the staff at the Center for Distributed Learning. The system is designed to provide access interrelated CDL data tables that focus on the following data components:

Teaching & Development

Faculty data centering around online faculty development & online course teaching histories. Development history and records are searchable by term. Online course instruction record is a component of both faculty member and the course.

Scheduling

Master schedule data on online courses available from Summer 1996 to present; all modalities available from Fall 2002 to present. Pertinent class data such as instructor of record, enrollment, section, etc is availble

Statistics

Comprehensive statistics and graphs covering faculty development, online class productivity, student enrollment (headcounts & web-exclusive). Dashboard provides a graphical representation of the data presented above from the EIS.

Current data for: **Spring 2011**

Productivity

Total University Sections: **6,910**
Total University Registrations: **217,604**
Total University SCH Production: **628,324**

Total Web(W) Sections: **563**
Total Web(W) Registrations: **30,507**
Total Web(W) SCH Production: **91,148**

Headcounts

Total University Headcount: **54,991**

(W) World Wide Web Headcount: **18,527** (33.69% of Univ.)
(M) Reduced Seat Time Headcount: **10,095** (18.36% of Univ.)
(W) or (M) Headcount: **24,811** (45.12% of Univ.)

Web Exclusive

Figure 2. Home Page of the Executive Information System (EIS)

A. How the EIS Works

The EIS is a classic web based application that utilizes a relational database as its primary data source. It is a split system where the web server sits separate from the database server as opposed to both being on the same computing environment. This allows for increased system performance and scalability over time. At just a little over 500MB, the amount of data within the database is actually small when compared to reporting data systems of similar characteristics. Unlike larger data warehouses that cover the whole organization, the EIS is a much more focused and tailored solution. The smaller focus allows for easier adaptability, development and maintenance over time as needs and request patterns change. It also allows for lower server and storage costs as the database does not consume vast quantities of space and the overall system does not suffer from performance degradation.

The EIS is mainly driven by open source applications. The three main open source applications that power the core functionality of the EIS are:

- MySQL – Popular and widely used open source relational database system;
- PHP – Widely available scripting language primarily used for web development;
- Apache HTTP Server – Widely used HTTP server.

All of these applications have proven to be highly reliable and scalable for this particular application. Regardless of applications or technologies that power a system such as this, it is its internal architecture that becomes critical to its success.

The internal architecture of the EIS centers around four main processes: data input, data preparation, data storage, and data display. Figure 3 below outlines the overall architecture of the EIS including some examples of what is contained in each process outlined above. The system in general does not deviate from the spirit of the traditional extract, transform, and load (ETL) methodology present in modern data warehousing applications. As with most ETL processes, those of the EIS are highly specific to how the system stores and ultimately reports on the data.

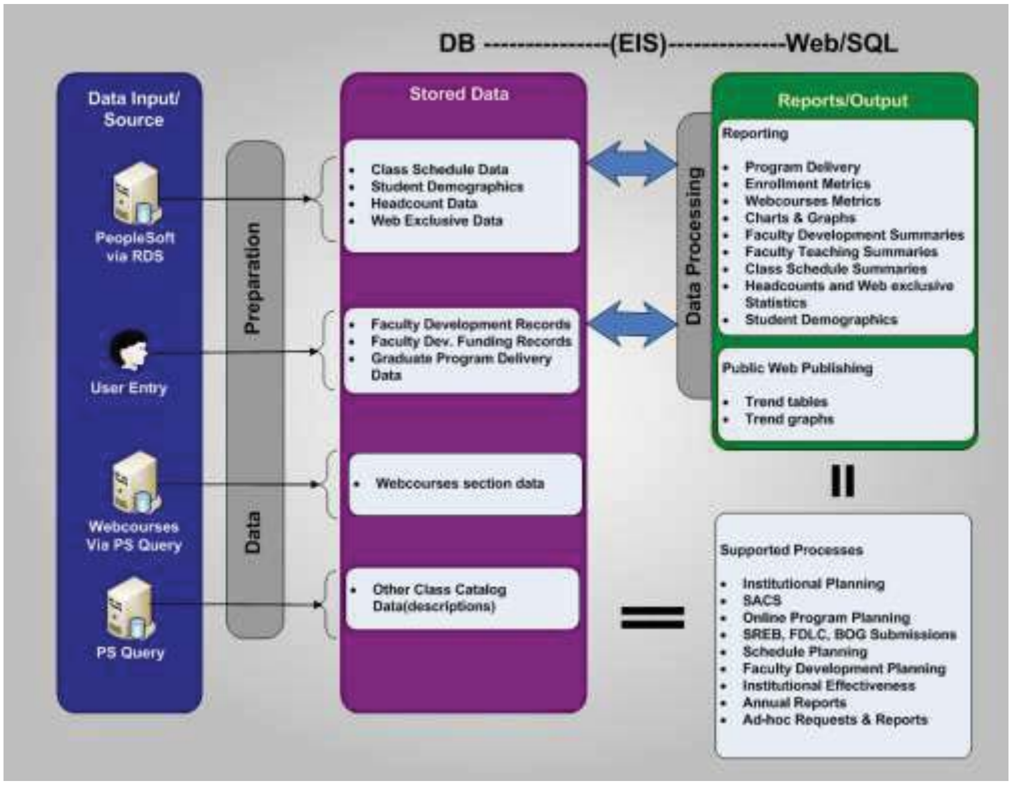

DB ---------------(EIS)--------------Web/SQL
Data Input/ Source
PeopleSoft via RDS
User Entry
Webcourses Via PS Query
PS Query
Preparation
Data
Stored Data
• Class Schedule Data
• Student Demographics
• Headcount Data
• Web Exclusive Data
• Faculty Development Records
• Faculty Dev. Funding Records
• Graduate Program Delivery Data
• Webcourses section data
• Other Class Catalog Data(descriptions)
Data Processing
Reports/Output
Reporting
• Program Delivery
• Enrollment Metrics
• Webcourses Metrics
• Charts & Graphs
• Faculty Development Summaries
• Faculty Teaching Summaries
• Class Schedule Summaries
• Headcounts and Web exclusive Statistics
• Student Demographics
Public Web Publishing
• Trend tables
• Trend graphs
Supported Processes
• Institutional Planning
• SACS
• Online Program Planning
• SREB, FDLC, BOG Submissions
• Schedule Planning
• Faculty Development Planning
• Institutional Effectiveness
• Annual Reports
• Ad-hoc Requests & Reports

Figure 3. EIS Architecture

The primary data source that feeds the EIS comes from UCF's Enterprise Resource Planning (ERP) system, PeopleSoft. These data consist primarily of class schedule data, LMS data, and student performance and demographic information. The ERP data is extracted, prepared, and entered automatically into the EIS database on a nightly basis. Once in, the class schedule data becomes the heart of the system as most all other data points are directly related to this information. The secondary sources of data come from primarily manual entry. These data consist of faculty development information and academic program data from the university graduate and undergraduate catalogs.

While most of the data in the EIS are inter-related, some numerical population data like headcounts are calculated during the input process and stored as a "fact" table. The primary purpose of these fact tables is to save system processing time and improve the user experience on the web interface. It would be quite expensive in terms of data processing time to both the system and the user if information that was based off these fact tables had to be derived ad-hoc as opposed to simply being retrieved. This delicate balance of how data is derived is critical to the overall system and its efficacy as a reporting mechanism to high level constituents. Special care is taken to ensure that all generated reports are finished and presented to users in an acceptable time frame.

The analytic reports that the EIS generates all fit into the high level report categories that are shown in Figure 3. The specific outputs and reports within these categories directly support the processes also defined in the graphic. Most reports and statistics are generated within the EIS upon request and displayed to users via the web interface. Data can also be retrieved via Structured Query Language (SQL) by users with the access and know how to do so. The web interface is broken up into categories (Figure 2) covering faculty development, class scheduling, academic program planning, and statistics. For more visual users, a dashboard is available that turns the numerically heavy statistics into charts and graphs. Approximately ninety percent of data retrieval comes from utilizing the web interface and its pre-defined reports.

B. Putting Analytics into Action

While the EIS is a powerful suite of features, it is constantly evolving, adding reports, creating a new question for every question it answers. Perhaps its most powerful aspect is the fact that a majority of the data that it analyzes and reports on exist in various other locations throughout the university (such as Institutional Research). However, the EIS aggregates these existing data with some manually-entered data to create a robust architecture that allows UCF to maintain a top-down view of what is happening with technology-based learning at all levels across the entire institution.

This ability to leverage *existing* data from elsewhere in the university and analyze the aggregate data set for various purposes is extremely valuable. For example, CDL uses the EIS to continually monitor each program in the university catalogue to determine how close it is to being offered 100% online. Through this process it was discovered that two tracks of a social science major were already 100% online, yet they were not declared as such for the official online program guide.

However, in subsequent discussions with the department chair, it was learned that due to a faculty scheduling issue he was unable to declare the degree completely online. He could not guarantee that one particular required course would always be offered in that modality. He simply didn't have the faculty to commit to supporting the degree as a fully online offering. CDL leadership then approached UCF's Regional Campus administration to inquire if they would be interested in securing a new faculty line on behalf of the department. Regional Campus has the ability to hire faculty for departments and place them in one of ten teaching sites around central Florida.

Regional Campuses was interested in adding the degree program to its offerings (online learning is a significant component of the Regional Campus strategy). They agreed to hire a new faculty member for the department on the condition that he/she would be committed to teaching the required course online, thus allowing the degree to be offered completely online. Declaring a degree as 100% online opens up additional opportunities for program outreach and growth. The final result was that CDL was able to list a new online degree program, Regional Campuses was able to offer a new program to Regional students, both online and face-to-face, and the department gained a new faculty member and the additional reach of an online program. It was the proverbial win-win-win and it was all facilitated by the data that were revealed within the EIS.

It is important to note that having the data is only half the equation. In order for those data to be valuable, the institution must *do* something with them. In the example above, the data were used to open a dialogue with the academic department and Regional Campuses that resulted in a new online degree program being offered. While each situation is unique, this is a fairly representative example of how UCF is both analyzing data and taking action based on it from a top-down, institutional viewpoint.

BOTTOM-UP PERSPECTIVE

From a bottom-up point of view, CDL's Research Initiative for Teaching Effectiveness (RITE) maintains a robust program of continual analysis and interpretation of data points such as student success, withdrawal, and perception of instruction (end of course evaluations). If the EIS top-down data are used to scan the university's distributed learning initiative from a primarily quantitative standpoint, the RITE bottom-up data are used to identify trends, compare performance, and track the progress of distributed learning.

These bottom-up student performance and perception data also help to inform decision-making at all levels of the university. New inquiries by RITE researchers have focused recently on grade point average (GPA) as a more reliable predictor of student success than other typical variables that are often studied in the context of learning analytics.

A bottom-up approach to analytics using preexisting data capitalizes on the institutional culture by providing faculty members and learning support personnel with information about the likelihood that students may not succeed in their courses. The process does not require additional analysis platforms that use student interactive data for a course and, therefore, does not assign specific nonsuccess probabilities to individual students. However, inherent in his approach to analytics is the capability of identifying robust risk probabilities across all instructional modalities (not being tied to any one mode or learning management system), student levels, demographic categories, colleges, and disciplines. The advantage of this method is its widespread applicability. The disadvantage is that these data are somewhat less specific about individual students. As a result, institutions will have to make decisions about the opportunity costs involved in any analytic data collection processes verses the added value achieved for collecting and using such information. However, the objective of the bottom-up institutional approach has the same objective as any other analytics approach: support our ability to maximize the chances of student success in courses and ultimately, help them receive their degrees. After all, analytic models, and there seem to a goodly number these days, should converge on student-success.

A. Necessary Preexisting Conditions

Data are much more useful when they play out against an understanding of the institutional context from a system such as the one described in the top-down sections of this article. Effective analytics procedure cannot function effectively in isolation from the institutional climate. Figure 4 portrays our thinking about the intersection of several domains in an effective analytics paradigm.

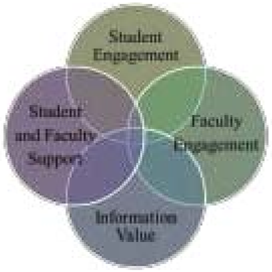


Figure 4: Integrated Domains for Analytics

Gardner Campbell called these "integrated domains" [12]. If students are engaged in the learning process and somewhat at-risk, altering them to that fact may, in all likelihood, motivate them. However, in our research on reactive behavior patterns [13], we have come to understand that sending an increased risk message to several student types can have just the opposite of that intended effect. Understanding these interactions, developing strategies for dealing with them, and sending the most appropriate message are critical to maximizing the success possibilities. The same holds true for faculty engagement levels. Engaged faculty are much more likely to use analytics for helping students achieve success, in some cases, by additional personal intervention when possible. Equally important in an effective analytics program is how useful data are to all concerned constituencies, not just students. Faculty and administrators are equally important to the process. The final component of Figure 4 makes the case that continued student and faculty support are critical to the success of any analytics initiatives. Hartman, Moskal, and Dziuban [13], in describing the necessary elements for operationalizing blended learning programs, have framed elements that apply just as well to an effective analytics initiative:

1. Effective institutional goals and objectives
2. Proper alignment
3. Organizational capacity
4. A workable vocabulary
5. Faculty development and course development (we substitute analytics) support
6. Support for students and faculty
7. Robust and reliable infrastructure
8. Institutional level on effectiveness
9. Proactive policy development and
10. An effective funding model

The bottom-up results we are about to demonstrate enjoy a much greater chance of success if these ten elements are in place. Let us be clear about what we mean here. Data do not make decisions, people do. Algorithms may seem like they make decisions but they have to be programmed on how to do so. There have been some effective efforts at machine learning but the human interface in the educational analytics culture is vital to its ultimate success.

B. An Example of How Judgment Plays Vital to the Analytics Process

In making the case for why decision making enriches the potential of analytics, we circle back to an early development in online learning. "The No Significant Difference Phenomenon" [14] made the case that class modality, specifically comparing online and face-to-face courses, led most people to conclude that there were no effects attributable to course format. One set of studies pursued this question, tallying the number of "significant findings" [14] while another group conducted meta-analyses based on effect sizes [15]. However Walster and Cleary [16] provided a thoughtful perspective on data analysis when they suggested that statistical significance is best used as the basis for decision making and not as an absolute determinant. They point out that hypothesis testing answers the following question: what are the chances that I will get my sample results when the null hypothesis is true in the population? These significant tests are a function of three things:

1. Significance level (e.g., .05, .01),
2. Sample size, and
3. Some effect size or degree of non-nullity as a mean difference. Usually, in the statistical literature, this difference is signified as delta (Δ).

Historically, the way most researchers conduct experimental and comparison group studies is to arbitrarily pick a significance level, get the largest sample size they can obtain and run the study. The consequence of conducting studies in this way is that by arbitrarily picking a significance level and sample size, the difference that will be significant is pre-determined. The consequence of such an approach is presented in Table 1.

Sample Size	$\bar{x}_1$=100 $\bar{x}_2$=101 ES=.06	$\bar{x}_1$=100 $\bar{x}_2$=103 ES=.20	$\bar{x}_1$=100 $\bar{x}_2$=105 ES=.33	$\bar{x}_1$=100 $\bar{x}_2$=120 ES=1.33
300	.41	.01	.00	.00
275	.43	.02	.00	.00
250	.46	.03	.00	.00
225	.48	.03	.00	.00
200	.50	.05	.00	.00
175	.53	.06	.00	.00
150	.56	.08	.00	.00
125	.60	.12	.01	.00
100	.64	.16	.02	.00
75	.68	.22	.04	.00
50	.74	.32	.10	.00

Table 1. Probabilities for Various Effect and Sample Sizes (SD=15)

Table 1 presents 11 sample sizes ranging from 50 to 300 with the mean differences and effect sizes ranging from trivial to quite large by most standards. With an effect size of .06 (mean difference=1), one will never achieve significance with any of the sample sizes while with an effect size of 1.33 (mean difference=20), that will always be significant. The middle two columns of the table demonstrate the impact of sample size on the significant difference decisions. For effect size .33 (mean difference=5), significance at the .05 level is achieved with a sample size of 75 and greater but not with sample size of 50. Finally, the effect size of .20 (mean difference=3), one must have sample sizes in the 200s in order to reach significance levels of .05 or greater. Table 2 provides a further demonstration of how sample size can impact your decision about whether a difference is significant or not. That table shows that no matter how trivial the difference is, if the sample size is large enough, it will lead to significance.

Sample Size	$\bar{x}_1$=100 $\bar{x}_2$=101 ES=.06
2750	.01
2500	.02
2250	.03
2000	.04
1750	.05
1500	.07
1250	.10
1000	.14
750	.20
500	.29

Table 2. Probabilities for Various Sample Sizes (SD=15)

The point is that the analysis is much more effective if some thought and decision making go into the process prior to collecting and running any data. Figure 5 provides an example. If the researcher can specify Δ_1, a difference that is of no interest or will not make a practical difference in his or her judgment, then the lower bound for the process has been established. Similarly, identification of $\Delta_{2,}$ a difference that will make a practical difference, finds the hypothesis testing procedure taking on a completely different perspective. This involves three steps:

1. Identify Δ_1 first –this is not important to me
2. Identify Δ_2 –this is important to me
3. Pick a significance level you can live with –.05, .01 or something else
4. Pick a sample size that will catch Δ_2 but not Δ_1. (Reference the program)

The result will be a power curve for your study that has the form of Figure 5: very little power (probability of rejecting) against Δ_1 and good deal of power against Δ_2.

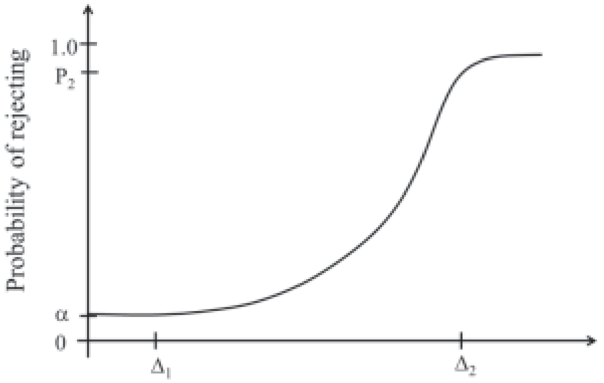


Figure 5. Ideal Power Curve

This process requires the investigator to provide input into the process and protects him or her from calling a trivial difference significant and provides the best opportunity for finding a difference that will be important. However, this decision making process cannot be accomplished by collecting data and automatically running it through an analysis program. Waiting for the program to tell you whether or not your results are significant does not optimize the potential information in your data. Most analytic procedures do not involve hypothesis but the principles of this demonstration apply. We still need to provide careful reflection on the process and take full responsibility for our decisions.

C. Institutional-level Analytic Data That Can Aid in Decision Making

One way that learning technologies impact higher education is by spawning multiple modalities for instruction. Of course, the index case for these new formats is fully online learning with an underpinning in course management systems. That configuration for teaching and learning enables many of the interactive analytic platforms described in this paper. At the same time, however, other instructional modalities have arisen—blended and lecture capture, for instance. In many colleges and universities, these two modes combined with online and face-to-face instruction provide the bulk of instructional modalities available to students. This is the case at UCF. At the institutional level, common analytics questions arise about the relative effectiveness of these course modalities in university organizations such as the faculty senate, student government, the faculty center for teaching and learning, administrative councils, colleges and departments among others. Most often these effectiveness questions frame themselves in terms of student success, withdrawal and satisfaction. At the institutional level, providing comprehensive answers to these questions in a timely fashion contributes to building a university culture that embraces analytic thinking. Table 3 provides an example of course modality impact for success encountered by students in face-to-face, online, blended and lecture capture courses.

	n	Fall 09	Spring 10	Summer 10	Fall 10	Spring 11	Summer 11
Blended	56,316	91	91	91	90	90	94
Online	150,834	87	88	88	88	88	89
Face-to-Face	665,209	87	88	88	87	87	91
Lecture Capture	12,050	87	83	86	84	84	79

Table 3. **Student Success by Modality in Percentage of Grade "C" or Higher**

Success in this case is defined by a student obtaining a grade of "C" or better because that level of achievement enables timely progression toward program completion. The table shows that, on average, highest student success levels occur in blended courses with a range from 90%-94%. Lowest success rates are found in lecture capture courses ranging from 79% to 84%. At UCF these data serve as the beginning point for understanding impact on students realizing that only in very rare cases is the modality of a course the primary reason for success. These data open a comprehensive discussion about how these findings can be explained around issues such as which colleges prefer which modalities, what disciplines are offered in each of the modalities, at what levels are the modalities offered, and an understanding of student preferences for the various modalities. What happens here is a deeper analytic discussion of the course modalities and their impact on students and, in some cases, actual learning outcomes. A primary benefit of these data is that they involve the entire university community in broad-based conversation.

	n	Fall 09	Spring 10	Summer 10	Fall 10	Spring 11	Summer 11
Blended	56,316	3	3	1	3	3	2
Online	150,834	4	5	4	5	4	4
Face-to-Face	665,209	3	3	2	3	3	2
Lecture Capture	12,050	4	4	5	5	7	7

Table 4. Student Withdrawal by Modality in Percentages

Table 4 shows some differences, non-differences, and trends for student withdrawal from courses by modality. The first obvious finding in these data in general that there is a very low probability that students will withdrawal from any of the course modalities at UCF. Realizing, of course, that helping even a relatively small number of students persist in their courses is important, this finding gives rise to a significant university-wide discussion. That conversation revolves around the value that might be added to instructional effectiveness by implementing any course level analytic protocol verses the cost in terms of resources, faculty time and benefit to students. More directly framed, the question becomes would the marginal gains be worth the cost? At the writing of this paper, that discussion continues at UCF and is likely to endure for some time—especially in the face of rapidly declining resources. However, Table 4 does show some noteworthy trends. Withdrawal rates at UCF across modalities range from a low of 3% to a high of 7%. Therefore, on average, there is an approximately 95% chance that a student will not withdraw from a course. Blended and face-to-face courses tend to have the lowest withdrawal rates, followed by online courses, with the highest incidence found in lecture capture courses. Generally, withdrawal rates tend to be stable across the modalities except for lecture capture that shows a tendency to increase over the semesters. This modality is relatively new to UCF, and rapidly evolving. In terms of institutional level analytics, the withdrawal data instigate the same level of university-wide discussion about modality effectiveness, augmenting the discussions about the likelihood of student success.

	n	Overall % Excellent
Blended	53,476	52
Face-to-Face	726,342	48
Online	121,257	48
Lecture Capture	13,292	42

Table 5. Comparison of Excellent Ratings by Course Modality from Spring 2008-Spring 2011

Table 5 presents one final piece of overall institutional analytic data about the course modalities offered at UCF—student satisfaction. It shows that blended courses produce the highest overall percentage of excellent student ratings, with a 4% advantage over face-to-face and online courses and a 10% advantage over lecture capture courses. The enduring conversation at the moment focuses on the fact that blended courses produce the highest student success levels, lowest withdrawal levels and highest student satisfaction. On the on the other hand, lecture capture produces lower success rates, higher withdrawal rates and the lowest student satisfaction. Therefore, these three tables, updated every semester, have been the inspiration for a culture-wide deliberation about how the university should proceed as a coordinated effort. The objective is to maximize student success, minimize withdrawal and increase student satisfaction. At many levels, discussion infuses the concept of facilitative analytics into university community much the same way that information fluency has become integral part of the lives of students, faculty members, staff, administrators and the community we serve at UCF [17]. We believe that analytics grounded in top-down, bottom-up approach will gain more traction, ultimately making any course level platform that might be adopted much more effective. Analytics are most effective when they serve to aid reflective decision making about their impact on teaching and learning.

D. Predicting the Chances of Student Non-Success from Institutional Data

The approach to assessing educational effectiveness at UCF using institutional data that we have presented so far holds value for building a foundation for an analytics environment. However, a question remains about using these data sources for harvesting more specific information about the probability of students not succeeding in their courses. Before proceeding, we expand our definition of non-success beyond simply withdrawing from a course. To be effective, any approach should maximize students' progress toward obtaining degrees. Therefore, non-success must be defined as a student withdrawing, or receiving a grade of D or F. All three circumstances prevent successful completion of an undergraduate degree. For the remainder of this paper, non-success (D, W, or F) will be coded as a yes-no binary variable. The analytics work at UCF has been developed in two phases: finding the best predictors of non-success and using those predictors to compute probabilities of non-success for homogenous student groups.

The first procedure involved multiple logistic regressions for determining the best predictors of academic non-success [18]. This analog to multiple linear regression, at times, is called the logit model with the term "logit" signifying the logarithm of the odds ratio for the occurrences of a particular yes-no outcome. The major variation of this method from linear regression is that the procedure does not involve continuous outcomes such as test scores. In logistic regression, the customary computation of the variance accounted for (R^2) has to be replaced with a pseudo version of the index [19]. The method for dealing with multiple predictors in this case was the add one procedure demonstrated by Lomax [20] where variables are entered sequentially, one at a time in into the cumulative model until the largest possible (R^2) is obtained. The first variable is entered and evaluated for predictability, then a second variable is added and the logistic regression computation recalculated. The process continues until the full model has been developed. This initial screening process determines which variables will be most effective in subsequent analyses. This prescreening process is critical to the success of any data mining analytics approach.

For the second phase of the analysis we used classification and regression trees (CART) [21]. This model is recursive, bisecting data into subgroups called terminate nodes or leaves. CART procedures require three distinct phases: data splitting, pruning and homogeneous assessment. The object is to develop decision rules about the probability of student non-success using the best predictors identified in the data screening phase of the study.

CART splits the data into two categories at each stage of the tree where the best predictor or predictors are used to identify those student groups with the highest probability of non-success and those with the lowest by minimizing the variance within the two groups. The tree continues to split the data until the numbers in each subset are too small to be informative. Typically, the growing process creates far too many nodes to be useful, resulting in the problem of over fit to a particular data set. The CART procedure attempts to solve this problem with an algorithmic pruning process that reduces the dimensionality of the tree, greatly simplifying the originally developed models. Although one is able to continually improve the fit to the data on which the model is developed, the remaining question is how well will it predict the outcomes on another data set that was not involved in the in the development process? This was the procedure used in this study. The results presented here are those applied to a validation data set achieved by dividing the data into the two sets. The final stage in the CART process involves determining the predictive power of the decision rules that have been developed. One way to accomplish this is to compute misclassification rates. Therefore, a rule that is 93% accurate in predicting student non-success will have a 7% error rate.

E. Identifying Predictors of Non-Success

Table 6 presents the results of the add one logistic regression procedures using several classes of valuables in the institutional data set for predicting non-success: course modality (online, face-to-face, blended, lecture capture), course level (freshman through senior), class size, demographics, ability measures (Total SAT score or imputed ACT), college membership and grade point averages.

	R^2
Modality	.003
Course Level	.022
Class Size	.024
Gender	.029
Ethnicity	.035
Age	.035
SAT	.034
College	.047
High School GPA	.074
Cumulative UCF GPA	.405

Table 6. Add One Logistic Regression Analysis for Predicting Non-Success (n=258,212)

The table shows that beginning with modality and adding variables to the equation through high school grade point average, the model produces virtually no predictability. In fact, the squared multiple correlation for the model through high school grade point average is only .074, less than 8% of the variance in non-success explained. However, with the addition of cumulative grade point average to the equation almost 41% of variance in non-success is explained. Figure 6 shows that relationship graphically.

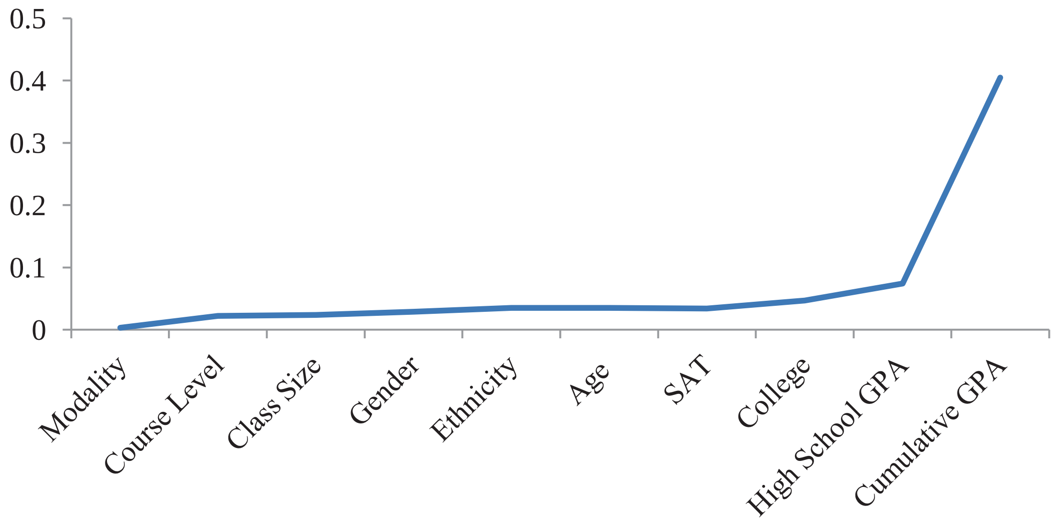


Figure 6. Add One Logistic Regression Analysis for Predicting Non-Success (n=258,212)

The figure shows that the full model accounts for more than five times more explanatory power for non-success than the partial model through high school grade point average. However, it is important to keep in mind that the best model we were able to develop fails to explain almost 60% of the variance in non-academic success. This raises an interesting dilemma in that the predictive domain of the variables we have available cannot address factors outside the institution. Other domains seem to be involved—ones that require further exploration. Most analytics platforms seemed to be constrained by this limitation.

F. Classification Rules for the Decision Trees

The logistic screening demonstrated that student cumulative grade point average was virtually the only variable that produced any predictive power for non-success. However, since average GPA, essentially, is a continuous variable we decided to declassify it into deciles for a number of reasons--the most important being that GPA cut points would not be as informative and actionable as discrete classifications of students. Note, however, that the deciles preserve some part of the ordinal properties of grade point average. Therefore, the undergraduate students' cumulative grade point deciles were used in the development of the decision tree rules. Figure 7 shows the pruned decision tree for predicating non-success status for the undergraduate student population for the fall 2009 through spring 2011.

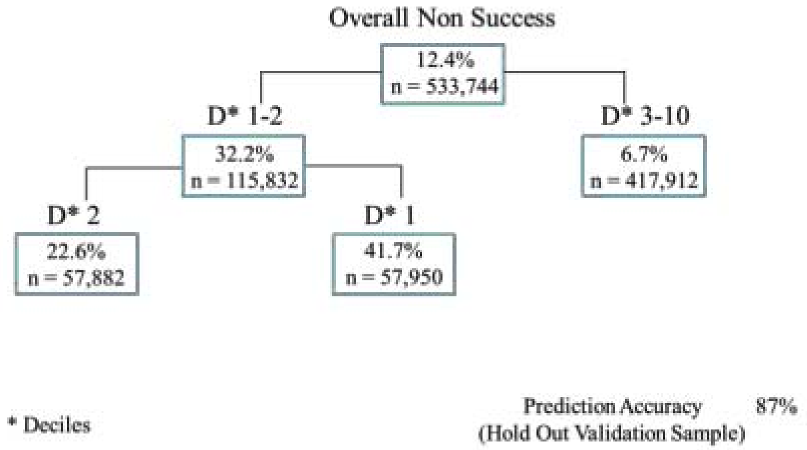


Figure 7. Classification and Regression Tree for Predicting Non-Success

Note that the overall non-success rate for this student population was 12.4%, remembering that withdrawal accounts for approximately 5% of non-success, therefore, grades of D and F estimate the remainder. However, there are other infrequently used grade classifications that impact this data making them approximate. In addition, large data sets experience anomalies quite often and are in a state of dynamic change. The tree in Figure 7 has a prediction accuracy of 87% and shows that the chances of W, D or F for students who reside in the first decile of GPA are approximately 42%. Students in decile 2 have an almost 23% chance of non-success. Note that both groups have a much higher probability of succeeding than not but the context for this finding becomes much more obvious when we find that the that the probability of non-success for students in deciles 3 through 10 is less than seven percent. Put in terms of odds ratios, students in the lowest GPA decile are almost 7 times more likely not to meet success in a course than those students in the upper deciles. Again, all groups are more likely to succeed than not, but one group is considerably more at risk for failure than the others.

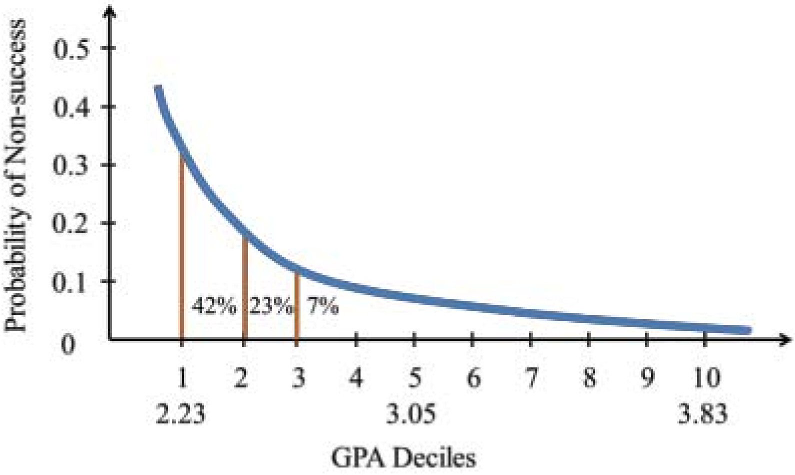


Figure 8. Non-Success by GPA Deciles (n=65,976)

Figure 8 shows the decreasing likelihood of non-success in any particular course as grade point average increases and indicates a much higher probability of potential student difficulties at the lower GPA deciles. However, when one examines the GPAs for deciles 1, 5, and 10, an interesting finding emerges. None of those averages put a student in danger of having to withdraw from the university. Obviously, this illustrates the need for a secondary analysis to find the subgroup in decile 1 that is in the most danger of non-success. In the future that analysis may identify those students that have a higher probability of non-success than success. That remains to be seen as we further refine the model.

G. A Stable Finding

	GPA Decile											
	1	2	3	4	5	6	7	8	9	10	Probability Success	Model Accuracy (Hold Out Validation Sample)
Overall (n=533,744)	x										.41	87%
		x									.23	

			x	x	x	x	x	x	x	x	.07	
Face-to-face (n=383,807)	x										.43	96%
		x									.26	
			x	x	x	x	x	x	x	x	.08	
Online (n=70,948)	x										.44	94%
		x									.24	
			x	x	x	x	x	x	x	x	.08	
Blended (n=58,577)	x										.38	94%
		x	x	x	x	x	x	x	x	x	.06	
Lecture Capture (n=13,674)	x										.45	95%
		x									.27	
			x	x	x	x	x	x	x	x	.08	
Freshmen (n=69,475)	x										.44	94%
		x	x	x	x	x	x	x	x	x	.06	
Seniors (n=223,759)	x										.33	96%
		x									.21	
			x	x	x	x	x	x	x	x	.07	

Table 7. Predictive Model Summary

Finally Table 7 presents a summary of all the decision trees developed for this paper comparing them to the overall finding. This was completed for online, blended, lecture capture, face-to-face modalities and the senior and freshmen cohorts. The obvious finding is clear. Lower decile students are at much greater risk of non-success that those at the upper levels. These findings appear robust over demographic characteristics of students, course characteristics, ability levels of students that have accepted to the university, college membership and high school performance. Cumulative grade point average proves to be a reasonably good predictor of average non-success probabilities. Further development of this model based on institutional data may produce procedures that can target those students who have a greater risk of failure so that a planned intervention might enhance their chances of successfully completing a degree program.

CONCLUSION

UCF's analytics efforts attempt to ensure that results are actionable. Without being able to do something productive with the data, the collection and reporting of information becomes an abstract art, of little use to the institution or students. As previously mentioned, within the EIS's program tracking is the ability to proactively report on every degree in the university and determine how close each is to being 100% available online. CDL leadership can use those reports when meeting with colleges and departments to shape strategic conversations about how to allocate resources to exploit opportunities that may not have been previously known without the data. Is a particular program only two courses away from being available completely online, thus making it "low hanging fruit" for online development? Does it make strategic and pedagogical sense to prioritize resources to develop those two courses and place the entire program online?

Likewise, understanding which bottom-up predictors, such as GPA, can be leveraged to impact student performance via early intervention, special advising, or extra tutoring can make all the difference in the world to both the individual student. When writ large across the university, it can have a positive effect on the entire institution.

Given the fact that analytic thinking can be leveraged out of almost any data set, the primary objective should be providing a safety net for students in order for them to succeed in their studies. Prediction is only a part of the process although admittedly an important one. Certainly, the current red hot enthusiasm for this paradigm will eventually cool and claims for massive transformation will find their proper place in the culture of higher education. Gibson puts it this way. “Once perfected, communication technologies rarely die out entirely, rather, they shrink to fit niches in the global info-structure” [22]. Jenkins [23] concurred when he concluded that rarely do new technologies replace old ones but rather converge with them in some way, creating an entirely new entity. Given these robust historical trends, analytics might best be thought of as not a set of competing methodologies but rather a unified mental model-a way of thinking that pervades the university culture. Analytics procedures that are not part of a systemic initiative will suffer the same fate as technologies that are bolted on, expensive, and by and large ineffective.

Many of the computational models available even at this early stage appear to have excellent predictability although we have seen very few validity studies in the Taylor Russell [24] sense. With these predictive powers come great responsibilities. Kahneman [25] provides compelling examples of the power of the anchoring effect in that even a completely random suggestion about a price, an amount or the character of a person that has no basis in fact can have tremendous impact on the decision making process. In addition, he gives graphic examples of the regression effect showing how individuals who perform poorly on a first attempt have a high probability of doing better on the second trial by chance alone. Conversely, people who do well the first time are much more likely to lower the score on their second attempt, simply due to regression and nothing else. Labeling a student as being at risk is serious business and even the mere suggestion of that to an instructor can have a profound effect. What takes place after the identification is critical. At the moment, some platforms simply alert the student and do nothing else, while others allow students to check their relative standing in the course on a number of indicators. Some programs use suggestion engines to inform and guide students. Others seek to build entire support cultures, while another class of models serves as navigation devices for students. All of these approaches demonstrate potential. The task ahead of us is to assess the potential of analytics adding value to the academy as it undergoes a game changing transformation.

ABOUT THE AUTHORS

Charles (Chuck) D. Dziuban is Director of the Research Initiative for Teaching Effectiveness at the University of Central Florida (UCF) where he has been a faculty member since 1970 teaching research design and statistics. He received his Ph.D. from the University of Wisconsin. Since 1996, he has directed the impact evaluation of UCF’s distributed learning initiative examining student and faculty outcomes as well as gauging the impact of online courses on the university. He was named UCF’s first ever *Pegasus Professor* for extraordinary research, teaching, and service and in 2005 received the honor or *Professor Emeritus*. In 2005, Chuck received the Sloan Consortium award for *Most Outstanding Achievement in Online Learning by an Individual*. In 2010, Chuck was named an inaugural *Sloan-C Fellow*. In 2012, UCF initiated the Chuck D. Dziuban Award for Excellence in Online Teaching for faculty members in honor of Chuck’s impact on the field of online teaching.

Patsy D. Moskal is the Associate Director for the Research Initiative for Teaching Effectiveness at the University of Central Florida (UCF) where she has been a faculty member since 1989. Since 1996, she has served as the liaison for faculty research of distributed learning and teaching effectiveness at UCF. Patsy specializes in statistics, graphics, program evaluation, and applied data analysis. She has extensive experience in research methods including survey development, interviewing, and conducting focus groups and frequently serves as an evaluation consultant to school districts, and industry and government organizations. She has authored or co-authored numerous articles and chapters on blended and online learning and is a frequent presenter on research involving Web courses. In 2011, she was named a Sloan-

C Fellow.

Thomas B. Cavanagh is Assistant Vice President of Distributed Learning at the University of Central Florida (UCF). In this role he oversees the university's distance learning strategy, policies, and practices, including program and course design, development and assessment. In his career, he has administered e-learning development for both academic (public and private) and industrial (Fortune 500, government/military) audiences. A frequent presenter at academic and industry conferences, he is an award-winning instructional designer, program manager, faculty member, and administrator. His research interests include e-learning, technical communication, and the societal influence of technology on education, training, culture, and commerce. He is also a published author of several mystery novels.

Andre Watts is an Information Technology Manager in the Center for Distributed Learning at the University of Central Florida (UCF). In this role, he is responsible for daily IT administration and maintenance. In addition, he plays a major role in the Center's data reporting activities to both internal and external constituents. His interests include academic analytics, data mining, and the evolving role of technology in the academic environment.

REFERENCES

1. Dictionary.com. http://www.dictionary.com
2. Van Barneveld, A., Arnold, K., and Campbell, J.P., *Analytics in Higher Education: Establishing a Common Language*, EDUCAUSE Learning Initiative (ELI) White Paper, 2012.
3. Siemens, G., Gasevic, D., Haythornwaite, C., Dawson, S., Buckingham Shum, S., Ferguson, R., Duval, E., Verbert, K., and Baker, R.S.J.D. Open Learning Analytics: An Integrated & Modularized Platform. *Society for Learning Analytics Research (SoLAR)* (2011). http://solaresearch.org/OpenLearningAnalytics.pdf.
4. Campbell, J.P., *Seven Things You Should Know About Analytics,* EDUCAUSE Learning Initiative: Boulder CO, 2007.
5. Fritz, J. Classroom Walls That Talk: Using Online Course Activity Data of Successful Students to Raise Self-Awareness of Underperforming Peers. *Internet and Higher Education* 14: 89-97 (2011).
6. Goldstein, P., and Katz, D., *Academic Analytics: The Uses of Management Information and Technology in Higher Education,* EDUCAUSE Center for Applied Research, 2005.
7. Campbell, J.P., and Oblinger, D.G. Academic Analytics. *EDUCAUSE Quarterly* 1-20 (October 2007).
8. Campbell, J.P., DeBlois, P.B., and Oblinger, D.G. Academic Analytics: A New Tool for a New Era. *EDUCAUSE Review* 42(4): 40-57 (July/August 2007).
9. Macfadyen, L.P. and Dawson, S. Mining LMS Data to Develop an "Early Warning System" for Educators: A Proof of Concept. *Computers and Education* 54: 588-599 (2010).
10. Arnold, K. Signals: Applying Academic Analytics. *EDUCAUSE Quarterly* 33(1) (2010).
11. Campbell, G. Personal communication, 2001.
12. Dziuban, C.D., Moskal, P., and Futch, L., Reactive Behavior Patterns, Ambivalence, and the Generations: Emerging Patterns in Student Evaluations of Blended Learning. In: Picciano, A. G. and Dziuban, C. D. (Eds.), *Blended Learning: Research Perspectives*, Sloan Center for Online Education, Needham, MA, 179-202, 2007.
13. Hartman, J.L., Dziuban, C.D., and Moskal, P.D., *Blended learning: A tool for institutional transformation,* Unpublished Manuscript, 2012.
14. Russell, T.L., *The No Significant Difference Phenomenon,* Raleigh, NC: North Carolina State University, 2001.
15. Means, B., Toyama, Y., Murphy, R., Bakia, M., and Jones, K., *Evaluation of Evidence-Based Practices in Online Learning: A Meta-Analysis and Review of Online Learning Studies,* Washington, D.C.: U.S. Department of Education, Office of Planning, Evaluation, and Policy Development, 2009.

16. Walster, G.W., and Cleary, T.A. Statistical Significance as a Decision Rule. *Sociological Methodology* 2: 246-254 (1970).
17. University of Central Florida Information Fluency. http://if.edu.edu.
18. Kleinbaum, D.G., *Logistic Regression: A Self-Learning Text,* New York, NY: Springer, 2010.
19. Nagelkerke, N. Maximum Likelihood Estimation of Functional Relationships, Pays-Bas, *Lecture Notes in Statistics* 69:110 (1992).
20. Lomax, R.G., and Hahs-Vaughn, D.L., *An Introduction to Statistical Concepts*, 3rd ed., New York: Routledge, 2012.
21. Breiman, L., Friedman, J.H., Olshen, R.A., and Stone, C.J., *Classification and Regression Trees,* New York, NY: Chapman & Hall, 1984.
22. Gibson, W., *Distrust That Particular Flavor,* p. 11, New York: Penguin Group, 2012.
23. Jenkins, H., *Convergence Culture: Where Old and New Media Collide,* New York: New York University Press, 2008.
24. Taylor, H.C., and Russell, J.T. The Relationship of Validity Coefficients to the Practical Effectiveness of Tests in Selection: Discussion and Tables. *Journal of Applied Psychology* 23(5): 565-578 (October 1939).
25. Kahneman, D., *Thinking, Fast and Slow,* New York: Farrar, Straus and Giroux, 2011.