METAMETRICS RESEARCH BRIEF

A New School of Thought for Our Thoughts on Schools:

Using Neural Networks to Enable Novel Higher Education Analytics

Steve Lattanzio, Research Engineer

JANUARY 2018

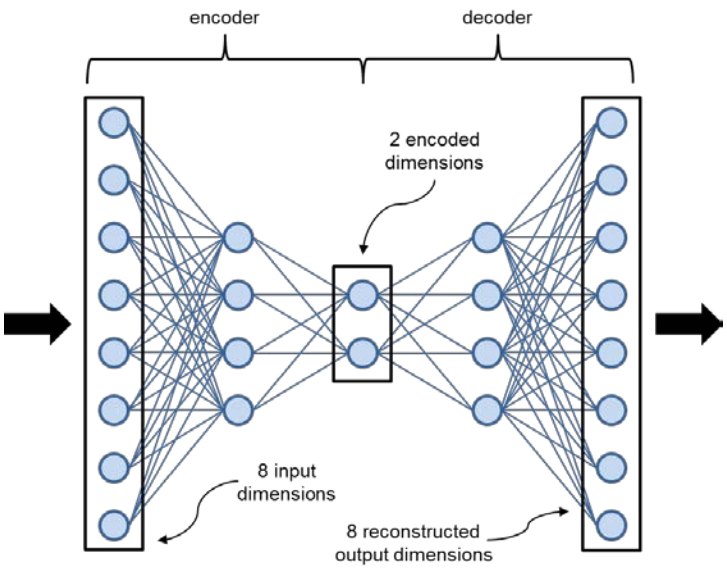
BACKGROUND

In the Fall of 2015, President Obama unveiled the *College Scorecard*, an online tool aimed at bringing much desired transparency to higher education. The intention of the *College Scorecard* is to put power back into the hands of post-secondary education consumers, enabling them to make more informed choices about what may be among the most consequential financial decisions of their lifetime. The *College Scorecard* consists of thousands of variables for thousands of schools going back almost two decades. Colloquially speaking, it is a data dump.

While the *College Scorecard* offers a boon of data, the biggest challenge is to make sense of it all. It is fairly straightforward to search, filter, or sort by specific fields of information for specific schools. However, average SAT scores of admitted students do not alone tell the whole story about the selectivity of a school; nor do average earnings data for a 10-year post-matriculation cohort of students receiving federal financial aid tell the whole story about the robustness of earnings from a particular school. And certainly, there is no variable in the dataset that is a good proxy for a relative value-add, that is, how much graduates exceed expectations following graduation—a metric that should be near the top of the list for post-secondary schools to strive to maximize.

Often, experts make rankings based on a handful of variables and assign their own weights to those variables based on their personal opinions of their relative importance. This can lead to fairly arbitrary rankings and produce very low reliability between different experts (Kamenetz, 2015). In fact, it is likely that most of the established college rankings end up informing one another in a process called herding (Banerjee, 1992), which is also a phenomenon that is mirrored by the behavior of universities attempting to increase their ranks (Morphew & Swanson, 2011).

Figure 1. Architecture of an example autoencoder



Architecture of an example autoencoder Neural Network that reduces eight dimensions down to two. For our analysis, much larger dimension reductions are performed over many more hidden layers.

Only a minority of the possible data actually exist—most elements are missing. Not all data are required to be reported, and there is great variance in the types of data provided by schools along with the volume. It is a very messy dataset with many caveats. The question is: what is the best way to take in the totality of the information in the data to create robust models, metrics, and rankings for schools in the dataset?

Fundamentally, this is a problem of feature engineering. With high-dimensional datasets, the so-called "curse of dimensionality" often rears its head. Also known as Hughes Phenomenon, this refers to the tendency for data points in a space to become further disjointed as the dimensionality of the space grows, requiring more data points just to maintain predictive accuracy (Hughes, 1968). Couple this with the large amount of missing data, and the problem becomes thornier.

The solution that we propose is to use neural networks to perform representational learning on the data. In other words, instead of manually going through the dataset and engineering a handful of features, we propose to use neural networks to automatically encode (autoencode) the information, *including* information about where data are missing, in a smaller dimensional space. Similar to principal components analysis (PCA), auto-encoding via neural networks is a dimension-reducing technique, but is more apt at handling variables that are nonlinearly related. In fact, it could be thought of as a more generalized version of PCA. Of course, such compression is lossy, but much of the information lost will be uninteresting noise and redundancies.

In this research brief, we present several pedagogical examples of representational learning using the *College Scorecard* dataset. First, we produce a two-dimensional map of thousands of schools showing how individual schools evolved over the period 2004 to 2014[1]. One of these dimensions appears to be a measure of college quality, a concept learned by the algorithm on its own. We also build a classifier to predict whether or not a school is considered an "Ivy League" school and find a set of schools that are adequately similar to those institutions. Lastly, we show how to derive estimates for abstract measures such as value-over-replacement-school (VORS)—a metric similar in concept to value-over-replacement-player (VORP) in sports analytics (Woolner, 2001).

[1] Despite the latest release of data in September 2017, there is not much data available yet for the 2014—2015 school year and beyond.

ANALYSES

Data Summary

Although all of the schools in the *College Scorecard* could theoretically be included in the analyses in this research brief, we limit our focus to schools that are typical four-year undergraduate programs, have greater than 100 undergraduate students enrolled, and report SAT scores—SAT scores being a convenient filter for schools that are reporting enough useful data. Furthermore, we limit our analyses to data from the 2004—05 school year and beyond, a range where there is a greater deal of data consistently available. In total, there are 1,503 unique schools included in our analyses and a total of 16,766 schools by year combinations.

The *College Scorecard* provides data across a handful of categories such as degree types, admissions (SAT scores), student demographics, financial aid, cost, future earnings, and more. Each of the categories contains many variables for many schools over many years. As previously mentioned, a lot of the data is missing (approximately 41% between 2004 and 2014). However, even though a given variable or school has a lot of missing data, that does not mean that the information that is available is not useful.

Pre-processing data

The College Scorecard data has some important variables that are lagged and it could be beneficial to "unlag" them. For example, at any given year of data there may be earnings data from the US Treasury department for a particular cohort. A 10-year cohort in 2013 would be more relevant in the 2002—03 dataset since that is the school year when those students matriculated. Table 1 shows where various cohort earnings data exists for students that received at least some level of federal funding.

Furthermore, some variables are categorical and include values for flags. These variables are expanded out so that each unique flag is represented by its own dichotomous (i.e. one-hot) variable.

Table 1. Existence of earnings cohort data by matriculation school year

Cohort	2001—02	2002—03	2003—04	2004—05	2005—06	2006—07	2007—15
10Y	✓	✓					
9Y	✓		✓				
8Y	✓		✓				
7Y	✓		✓				
6Y	✓				✓	✓	

The data was structured as a large matrix that includes one row per school per year and the columns include all of the available fields of data. Data is unlagged where applicable. The resulting matrix is 16,766 rows by 2,361 columns.

Encoding data

As previously mentioned, a main concept is that we can encode sets of variables into vectors that are more manageable than the full set of individual data elements taken together. In addition to the existing data, information about what data is missing is also itself encoded in these vectors. This is important since the data are missing-not-at-random (MNAR); in other words the fact that an element is missing may be useful information in and of itself.

There is a potentially useful consequence to treating the missing data as MNAR: schools cannot expect to game metrics by selectively omitting data—the algorithm may have the capability to detect the relationships between patterns of missing data. If schools have a tendency to omit unfavorable data, the algorithm has the potential to learn and account for that fact.

Instead of encoding all of the data at once, data was first encoded by category (e.g., admissions, earnings, etc.). Not only does this allow for encoding more manageable chunks of data and move more towards a balanced dataset, but it allows for discrete categories of data to be combined and incorporated into further encoding. Before a given category of variables are encoded, indicator variables for missing data are augmented to the set, the missing data elements themselves are replaced by the mean of the existing data for their class[2], and all variables with no variance (i.e. carrying no useful information) are discarded. This set of variables is then normalized such that all values are between zero and one. All years (2004—2015) were encoded simultaneously and each school could be present for multiple years, which means that information about the year and its relationship to different patterns of missing data was also included. To ensure that trends in patterns of missing data across years did not result in a time dimension being important to encoding (we want to be able to compare schools at different years in the same space), encoding was done by using a fixed year as a target. The 2011—12 school year was used as a target since it has the most complete data. That is, auto-encoding was performed on just 2011—12 data first, followed by using those encoded 2011—12 vectors as targets for encoding vectors from all of the various other years.

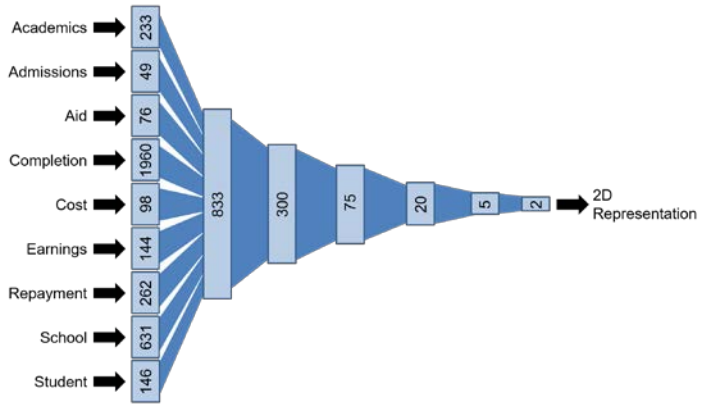
[2] This is a reasonable placeholder value that should be minimally disruptive during neural network training and the choice of the imputed value should be mostly washed out. In other words, the pattern of missing data will end up being more important to the model than the specific value used to replace the missing data element.

On a technical note, many neural networks mentioned in this brief are trained using dropout (Srivastava et al, 2014) as a regularization method to prevent over-fitting, with the probability of randomly keeping a node during training at $p = 0.5$. Regardless of whether or not a specific training is done with dropout or other regularization techniques, the results are confirmed as not overfitting using a holdout validation data set (i.e. data that was not used in the training).

Two-dimensional representation and overall college quality

Humans can visualize three dimensions pretty well, but a sheet of paper can really only show two dimensions well. Can we reduce all 3,599 variables in the dataset (when missing data indicators are included) down to just two variables in a way that retains the most information possible? And if so, will the results be meaningful and interpretable?

Figure 2. Architecture of the *College Scorecard* autoencoder



Architecture of the *College Scorecard* autoencoder Neural Network that reduces 3,599 dimensions down to two. Each layer of the network is labeled with the number of neurons (or nodes). The decoder portion is not shown. Blue areas denote fully connected layers of neurons. Not drawn to scale.

Since we want to include all of the available data, we take all of our previously encoded categories and create an augmented dataset. This dataset, which now has 836 dimensions, is subjected to an autoencoder with progressively smaller dimensioned layers starting at 300 and going to 75, 20, five, and finally two. The overall neural network architecture to go from the original dataset down to two dimensions is shown in Fig. 2.

As the trainings were completed, it became clear that the neural networks were figuring out that the best way to encode all of this information into two variables was to encode them into variables that appear to be best described as quantity and quality.

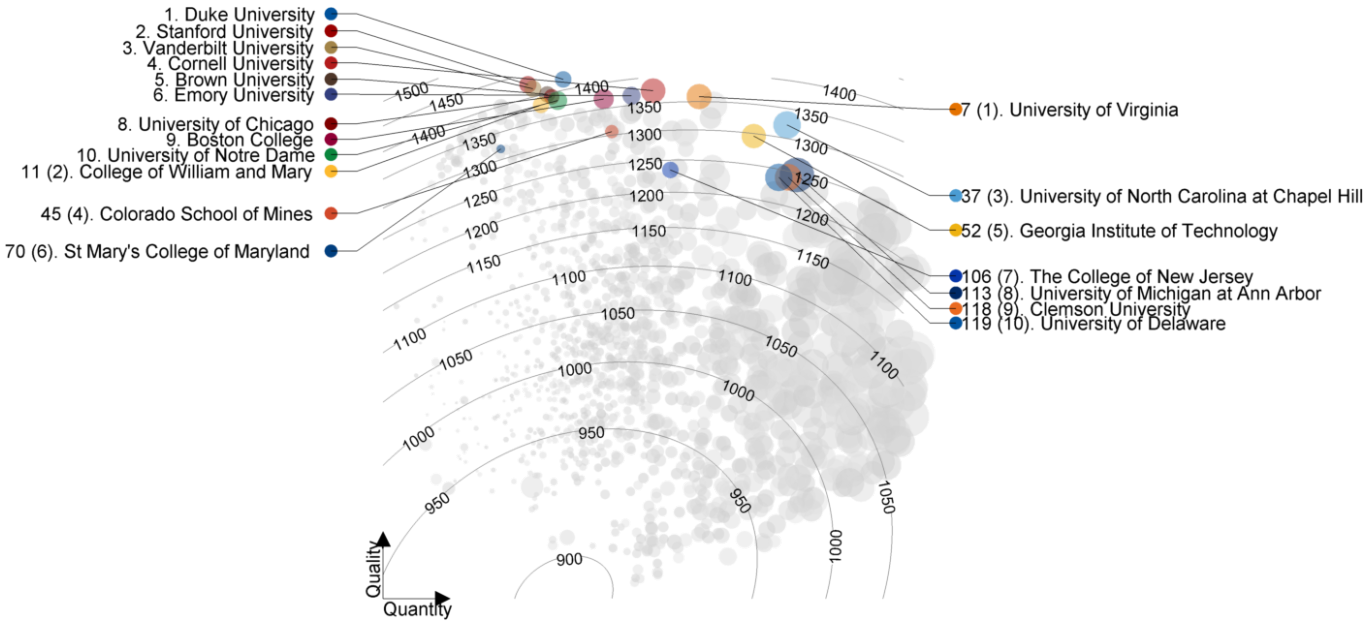
Because training neural networks is a stochastic process and can result in solutions that are different, but with nearly the same individual performance (in terms of encoding the original information), we trained an ensemble of 60 deep neural networks and averaged over them. Before the averaging, care was taken to apply a rotation to the individual two-dimensional vectors so that they aligned in a consistent fashion with the dimension corresponding overwhelmingly with size-related metrics as the first dimension and the dimension corresponding to quality-related metrics as the second dimension.

Fig. 3 shows the resulting two-dimensional representational map of 1,303 American colleges and universities in the 2013—2014 school year. A set of contour lines are overlaid that show the approximate average SAT scores for the schools at various points in the plot.

These contours are the result of a simple bivariate quadratic function regression using just the two dimensions. Of course, when reducing over three-thousand dimensions down to just two, much of the variance is smoothed over and some schools, especially those at the edges of the map, may have actual SAT scores that are more than 100 points different than is shown in the contours. Average SAT score is just one dimension out of thousands being encoded into two dimensions, yet the RMSE for the SAT score model is just 63 points and the correlation between the modeled and actual SAT average is 0.87, which are both respectable given the context.

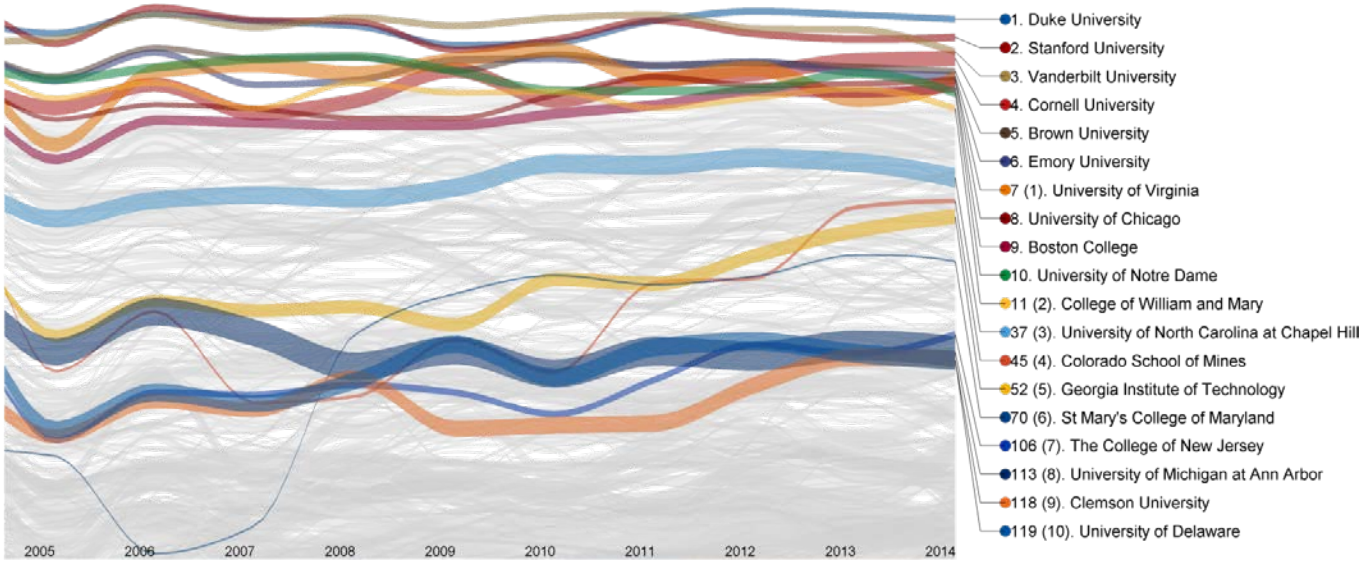
Fig. 4 shows how the top schools, overall and public, evolve over the “quality” dimension between the 2004—05 and 2013—14 school years.

Figure 3. 2D map of American colleges and universities



2D map of 1,303 American colleges and universities in the 2013—14 school year. The top 10 overall and top 10 public schools are shown in school colors. Circle areas are proportional to the size of the undergraduate population. Contour lines show approximate average SAT scores for matriculating students.

Figure 4. School Quality between 2004—2014



School quality between the 2004—05 and 2013—14 school years. The top 10 overall and top 10 public schools are shown in school colors. Line widths are proportional to the size of the undergraduate population.

The "Hidden" Ivies

There are eight schools that are members of the renowned Ivy League: Brown University, Columbia University, Cornell University, Dartmouth College, Harvard University, the University of Pennsylvania, Princeton University, and Yale University. All of these universities are top universities, but is there something about them that makes them uniquely "Ivy" other than being part of a very old athletic conference? Perhaps there are other schools that have enough signature qualities of an Ivy League school that makes them just as Ivy as any of those eight schools. This idea was explored in the book *The Hidden Ivies* (Greene & Greene, 2014), now in its third edition. The father and son duo of Howard and Matthew Greene have thoroughly researched the landscape of post-secondary schools to identify a set of 63 schools (from edition three) that could be classified as Ivy League schools.

As a second illustration of a use case for processing *College Scorecard* data with neural networks, we train a neural network classifier to identify Ivy League schools. The model ends up providing a measure for the likelihood that a school is a member of the Ivy League based on its 300 values in the 3rd layer of the autoencoder shown in Fig. 2. If the average likelihood of a school over the years 2004—2014 exceeds that of the lowest average over the same period of the set of Ivies, we identify it as a hidden Ivy. There are 51 schools outside of the Ivies themselves that meet that criterion. Furthermore, we identify a set of nine schools that are trending towards hidden Ivy status in the near future and a set of the 20 schools that are the next closest to being classified as a hidden Ivy.

Table 2. The "Hidden" Ivies

Our 51 Hidden Ivy Members		
Amherst College	George Washington University*	Stanford University
Barnard College	Georgetown University	Swarthmore College
Boston College	Hamilton College	Trinity College
Bowdoin College	Haverford College	Tufts University
Brandeis University	Jewish Theological Seminary of America*	University of Chicago
Bryn Mawr College	Johns Hopkins University	University of Notre Dame
Bucknell University	Lehigh University	University of Richmond
California Institute of Technology*	Massachusetts Institute of Technology*	University of Rochester
Carnegie Mellon University*	Middlebury College	Vanderbilt University
Case Western Reserve University	Northwestern University	Vassar College
Claremont McKenna College	Oberlin College	Villanova University
Colgate University	Polytechnic Institute of New York University*	Washington University in St Louis
Colorado College	Pomona College	Washington and Lee University
Cooper Union for the Advancement of Science and Art*	Rensselaer Polytechnic Institute*	Wellesley College
	Rice University	Wesleyan University
Duke University	Skidmore College	Williams College
Emory University	Smith College	Yeshiva University*
Fordham University		
Our Nine Hidden Ivy Prospects		
Colby College	Grinnell College	Sewanee-The University of the South
Connecticut College	Kenyon College	St Mary's College of Maryland*
Franklin W Olin College of Engineering*	Lafayette College	Wheaton College*
Our 20 Other Almost Hidden Ivy		
Babson College*	Illinois Institute of Technology*	Syracuse University*
Bentley University*	New York University*	Union College
Boston University	Occidental College*	University of North Carolina at Chapel Hill*
Brigham Young University at Provo*	Reed College	University of Southern California
Davidson College	Santa Clara University*	University of Virginia*
Drew University*	Southern Methodist University	Wake Forest University
Gettysburg College*	Stevens Institute of Technology*	

*Not identified as a hidden Ivy in *Hidden Ivies* Greene & Greene (2014).

It is remarkable to see the degree of agreement—in both the schools included and the number of schools included—between the Greenes' research and the output of a neural network working on an extensive set of publicly available data. There are only six out of 63 schools identified by the Greenes (and had sufficient data in the *College Scorecard*) that did not make it into our table.

Although these are all great schools, it is important to note that these "hidden" Ivies do not simply make up the list of the most elite schools that are not Ivies. Many equally or more competitive schools did not make it into Table 2 because they have enough characteristics that make them distinctly different from Ivy League schools. Likewise, many of the schools in Table 2 are relatively non-competitive academically with Ivy League schools, but have many other similarities to the Ivies. The model is identifying a signature for an Ivy League school that extends well beyond SAT scores and acceptance rates to all of the thousands of additional fields in the *College Scorecard*, some of which could be fairly superficial, such as the geographic location or demographic makeup of the school.

Value over replacement school (VORS)

In baseball, in addition to other sports, there is a statistic known as value over replacement player (VORP). This is a statistic used to capture the marginal utility of a player, that is, the amount of output they produce (such as runs in baseball) relative to a player from the existing talent pool (Woolner, 2001). The same concept can be applied to post-secondary schools. The value over replacement school (VORS) can be operationalized as the amount of additional earnings a school adds to a matriculating student relative to the expected earnings from schools with similar populations of matriculating students and the cost to attend. While the concept is similar to VORP, the methodology is fundamentally different—sports statistics tend to involve a lot of heuristics while we are using neural networks.

Since earnings data is fairly sparse in the dataset and we would like to keep our analysis based on actual observed earnings instead of inferred or imputed data, we chose to only look at 10-year mean earnings for students who matriculated (not graduated) in the 2001—02 or 2002—03 school year—the last year that such data is available (the data was last collected in 2013).

The VORS statistic is the observed 10-year mean earnings minus the expected 10-year mean earnings, where the latter is the output from a neural network trained on admissions, cost, school, and student data. In other words, we are attempting to model average earnings based on the type of student that matriculates to a school and what the school charges them to attend. The top 25 VORS schools are shown in Table 3.

A quick glance at the table yields something immediately obvious: most of the top 25 schools are either well-known elite institutions or provide a specialized education in fields such as business, engineering (especially marine engineering), medicine, and aerospace. While not shown, the lowest ranked schools are almost entirely art schools and liberal art schools. Many of these schools are at the bottom simply because they are highly selective and attract very talented students that would otherwise earn high incomes if they decided to pursue other careers.

Table 3. Value over replacement school (VORS)

Rank	School Name	Observed 10Y Mean Earnings	VORS
1	Albany College of Pharmacy and Health Sciences	$112,000	$56,400
2	Maine Maritime Academy	$96,000	$48,000
3	MCPHS University	$104,000	$47,100
4	Massachusetts Institute of Technology	$133,000	$45,600
5	Harvard University	$135,000	$43,300
6	University of the Sciences	$90,000	$36,200
7	Stanford University	$124,000	$35,700
8	Princeton University	$112,000	$35,600
9	Massachusetts Maritime Academy	$82,000	$33,900
10	Babson College	$107,000	$32,700
11	Yale University	$118,000	$32,200
12	California State University Maritime Academy	$88,000	$31,100
13	University of Pennsylvania	$118,000	$29,400
14	Georgetown University	$114,000	$28,500
15	Duke University	$108,000	$25,900
16	SUNY Maritime College	$78,000	$24,500
17	Dartmouth College	$102,000	$23,600
18	Xavier University of Louisiana	$60,000	$23,400
19	Washington and Lee University	$89,000	$23,100
20	Colorado School of Mines	$86,000	$22,200
21	Mount Carmel College of Nursing	$56,000	$19,900
22	Ohio Northern University	$69,000	$18,200
23	University of the Pacific	$80,000	$17,100
24	Claremont McKenna College	$90,000	$16,600
25	Baptist Memorial College of Health Sciences	$54,000	$16,100

*Now part of New York University

While the VORS statistic does a good job capturing the ability of individual schools to create additional economic value for their students, there are a few caveats. First, the earnings data from the Treasury Department only includes students who received at least some federal funding. Second, while creating economic value for students should be a high priority, enrichment does not need to be literal to have value. Third, future economic outcomes of students may not be fully driven by the academic and personal growth of a student while attending a specific university, but are likely also driven in large part by pre-existing and acquired social networks (Dale & Krueger, 2014) among other things. Finally, some schools did not have the required earnings data for the calculation.

A useful potential addition to the *College Scorecard* would be GRE, GMAT, MCAT, and GMAT scores for graduating students. This would provide a richer picture of the caliber of graduating students when combined with earnings data. This could also yield individual level value-add over incoming SAT scores. Thus, if all instruments could be linked to a common scale (e.g., Lexile or Quantile scales) the value-add could be denominated in language and mathematical growth over the college career.

CONCLUDING REMARKS

The methodology described in this paper and the pedagogical use-cases provide a rich framework for advanced analytics of post-secondary education—something that the consequence of the industry and the unwieldiness of the data demands. It is our hope that a future proliferation of similar work will promote further transparency in the post-secondary school market, more holistic approaches to data use, and ultimately more complete, fairer, and objective metrics that empower students to make the best decisions.

REFERENCES

Banerjee, A. V. (1992). A simple model of herd behavior. *The Quarterly Journal of Economics, 107*(3), 797-817.

Dale, S. B., & Krueger, A. B. (2014). Estimating the effects of college characteristics over the career using administrative earnings data. *Journal of Human Resources, 49*(2), 323-358.

Greene, H., & Greene, M. W. (2014). *The Hidden Ivies: 50 Top Colleges from Amherst to Williams that Rival the Ivy League*. Collins Reference.

Hughes, G. (1968). On the mean accuracy of statistical pattern recognizers. *IEEE transactions on information theory, 14*(1), 55-63.

Kamenetz, A. (2015). The New College Scorecard: NPR Does Some Math. *Retrieved from npr.org: http://www.npr.org/sections/ed/2015/09/21/441417608/the-new-college-scorecard-npr-does-some-math*

Morphew, C. C., & Swanson, C. (2011). On the efficacy of raising your university's rankings. In *University Rankings* (pp. 185-199). Springer Netherlands.

Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I., & Salakhutdinov, R. (2014). Dropout: A simple way to prevent neural networks from overfitting. *The Journal of Machine Learning Research, 15*(1), 1929-1958.

Woolner, K. (2001). Introduction to VORP: Value Over Replacement Player. *Retrieved from Stathead. com: https://web. archive.org/web/20070928064958/http://www.stathead. com/bbeng/woolner/v orpdescnew. htm.*