STEM EDUCATION & TRAINING PATHWAYS

Hannah Brinkley (VT), Raghav Sawhney (VT), Emily Sheen (PSU), and Devika Nair (AU)
SDAL: Bianica Pires and Ian Crandell
Sponsor: Karen Hamrick, The National Center for Science & Engineering Statistics at NSF

Project Objective and Motivation

Science, technology, engineering, and math (STEM) education and training is vital for a competitive workforce and innovation. Knowledge of the diverse set of pathways available to individuals desiring to enter the STEM workforce is critical for promoting a STEM-capable workforce (NSB 2015, NSB 2016). The National Center for Science and Engineering Statistics (NCSES) is interested in identifying and measuring these STEM pathways.

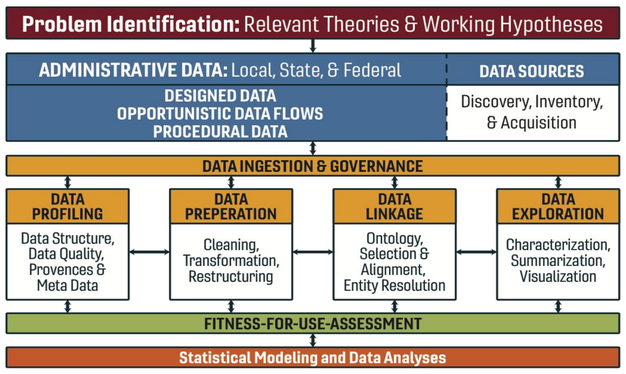

Figure 1 - SDAL's data science framework used to guide our research.

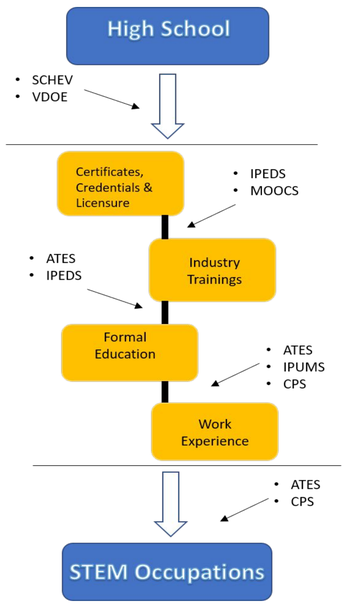

Figure 2 - STEM Pathways Data Map

Data Discovery & Profiling

Data discovery focused on understanding the landscape of sub-baccalaureate education and training options such as certificates, Massive Open Online Courses (MOOCs), and vocational schools. New and existing data sources that were profiled, prepared, and explored include:

- Integrated Postsecondary Education Data System (IPEDS)
- Adult Training and Education Survey (ATES)
- Harvard and MIT MOOC providers
- Virginia Department of Education (VDOE)
- The State Council of Higher Education for Virginia (SCHEV)
- The College Board
- Current Population Survey (CPS) data on Certificates and Licenses

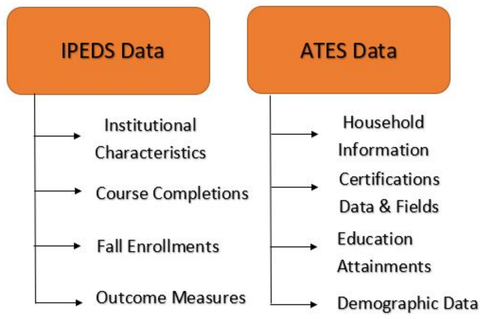

Figure 3 - IPEDS and ATES datasets prepared and explored.

Data Preparation, Linkage & Exploration

The data sources we selected to explore further were IPEDS, ATES, and HarvardX and MITX in addition to data on Virginia high schools from sources such as VDOE and SCHEV. Data was extracted from a multitude of formats, including Microsoft Excel, Comma Separated Values, and web-scraped websites. Hundreds of variables were reviewed, multiple years and multiple datasets were combined, and the data was transformed into usable formats.

Figure 4 - Education attainment for survey respondents with and without certifications (Data source: ATES, 2016).

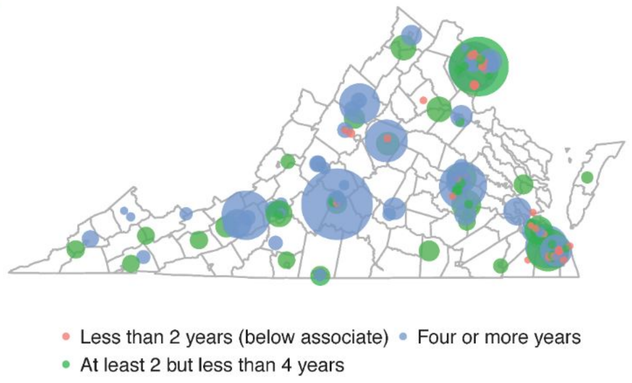

Figure 5- Institutions in Virginia by institution type of fall enrollment size (Data source: IPEDS, 2016).

A MOOC (Massive Open Online Course) is a free and widely accessible online course

➔ Taught by professors from universities
➔ Students don't need to attend a particular university to enroll in a specific MOOC
➔ We explored HarvardX and MITX MOOC enrollment data ranging from 2012-2016

Course Category	% Participants	% Explored	% Certified	% Female	% Bachelors
STEM	73% (3.2M)	16%	5%	23%	64%
Non-STEM	27% (1.2M)	19%	7%	41%	74%

Table 1 - HarvardX and MITX enrollment and participation summary (Data source: HarvardX/MITX MOOC enrollment, 2012-2016).

Exploratory Statistical Analysis of Postsecondary-Going Culture at Virginia High Schools

We aimed to identify the characteristics of high schools that encourage students to pursue postsecondary educational opportunities. Predictive models that use school characteristics, such as mean SAT scores and on-time graduation rates, were developed to predict the proportion of students that will enter various postsecondary trajectories such as 4-year college enrollment, employment, and the military.

- Exploratory PCA was used to identify top variables
- K-means clustering and exploratory plots were used to characterize schools
- Predictive models such as logistic regression, stepwise AIC selection, and cross-validated LASSO penalized regression, were used to assess the predictive power of our data and to determine which features are most relevant

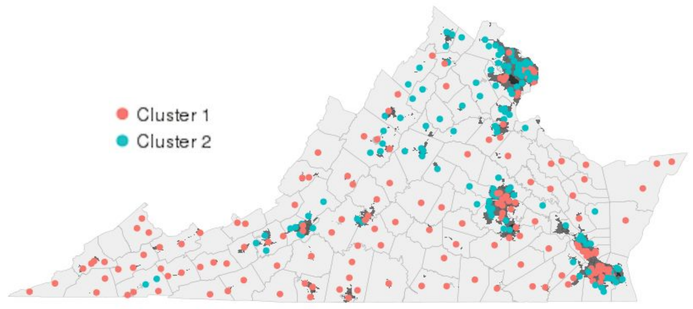

Figure 6 - Map of Schools colored by unsupervised K-means clustering algorithm. Black areas on the map are the urbanized areas, lighter areas are more rural. K = 2 clusters chosen as optimal using elbow, silhouette, and gap-statistic methods. Clustering reveals geographic correlation with major metropolitan areas in Virginia tending to be clustered together.

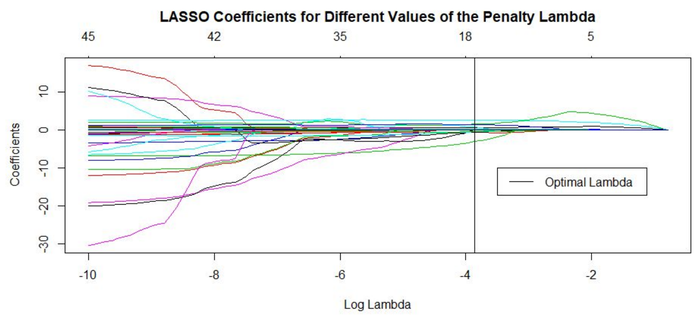

Figure 7 - Plot of LASSO coefficients for range of λ penalty values. Higher penalties correspond to a reduction in coefficients with variable selection occuring as coefficients zero out. Our selected LASSO model optimized cross-validated mean squared prediction error, CVMSPE = 0.1175, with a selected tuning parameter λ = 0.0211. The final predictive model selected 18 variables with non-zero coefficients.

REFERENCES

National Science Board. 2016. Science and Engineering Indicators 2016. Arlington, VA: National Science Foundation (NSB-2016-1).
National Science Board. 2015. Revisiting the STEM workforce: a companion to Science and Engineering Indicators 2014. Arlington, VA: National Science Foundation (NSB-2015-10).

bi.vt.edu/sdal
Stay Connected: facebook.com/biocomplexity twitter.com/biocomplexityvt youtube.com/c/biocomplexity google.com/+biocomplexity