Process Book

CS – 5630/6630

University of Utah

Team:

Dylan Zwick

dylanzwick@gmail.com

UID: u0075213

Clark Barnett

twentyonetroy@gmail.com

UID: u1119424

Intro:

This project grew out of a data competition in which Dylan Zwick participated. The goal of the competition was to take the College Scorecard Dataset (compiled, published, and made available to the public by the U.S. Department of Education) and use it to construct something useful for high school students as they are trying to decide upon a college.

College Scorecard Dataset - https://collegescorecard.ed.gov/data/

Dylan's team (the "A-Team") built a model that would help high school students find schools similar to ones in which they are interested. There are many college options, and if a young student knows one school in which they are interested, this model provides additional, similar schools the student may wish to consider.

Dylan built a D3 visualization for the exploration of the model his team produced for the competition. This simple visualization was the foundation for the much more substantial visualization that is this project.

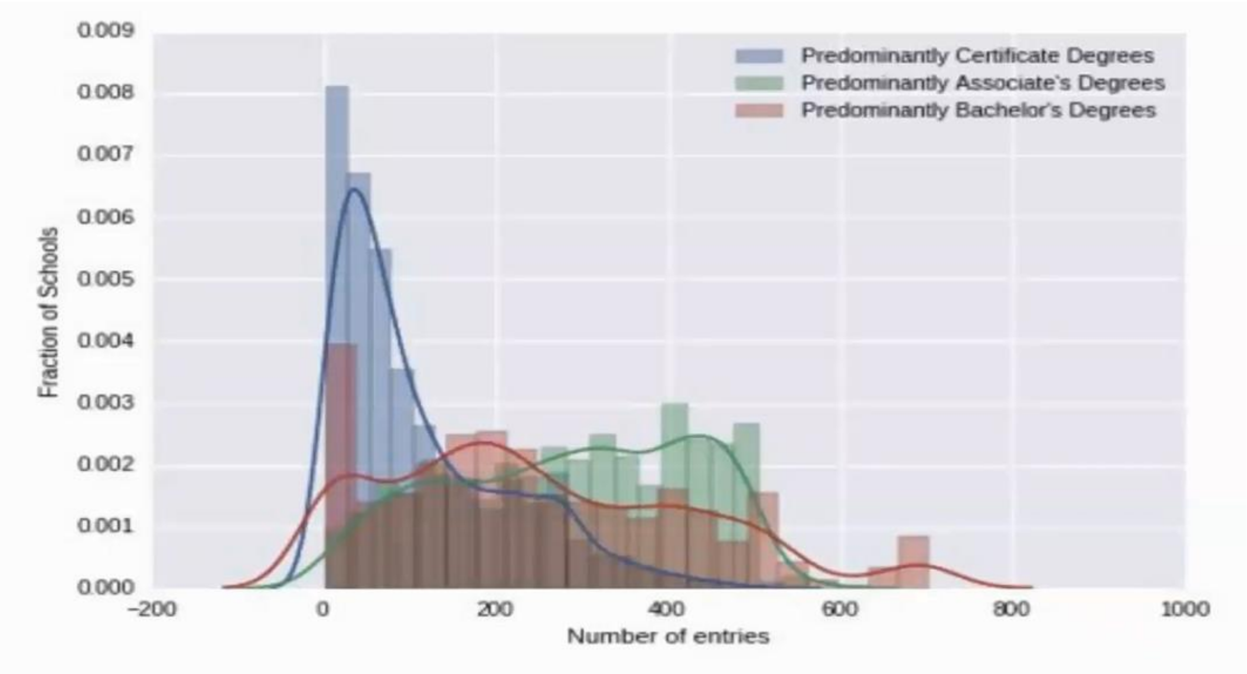
The story behind the model, and how the simple visualization built for the competition has grown into the final project for this class, is what follows.

The Data:

Note – The data analysis and modeling for this project was done for the data science competition by a team including Dylan Zwick, but also including Krishanu Banerjee, Jacob Hummel, Soumya Mishra, Dale Murray, and Julia Silge. This was not work done specifically for CS-6630 or this final project, but it was done partly with this project in mind. In this section references to "we" are references to this team, not the duo that created this final project.

The number of colleges and the number of variables available from the College Scorecard Dataset regarding individual colleges is enormous. Making sense of which variables are important and which are not is difficult. Also, for many of the schools data is unavailable for many, if not most, of the variables, and this presents difficulty in modeling.

As a first exploration of this data, the A-Team graphed the number of variables that have data available vs. the fraction of schools with that number of entries, and used color encoding to group the schools according to the type of degrees they grant. Most schools with little data are 2-year, vocational institutions like beauty/barbering schools, and technical colleges:

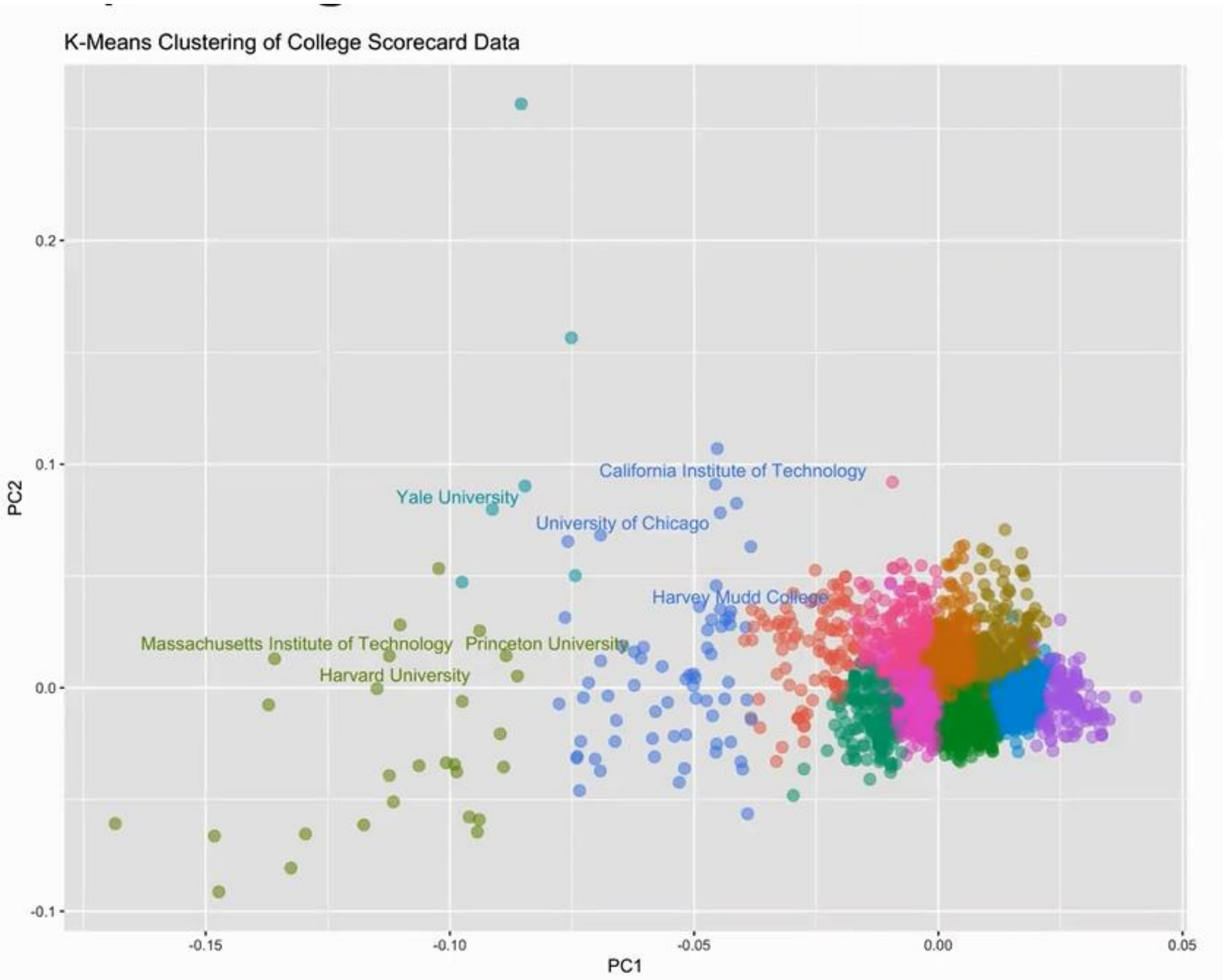


Wanting to build our model, and our data visualization, on schools that have data, we filtered our set of schools down to predominantly-bachelor's-degree institutions.

For this set of schools, we pruned the variables according to the frequency with which data for the variable was available, and grouped some variables that tended to be highly correlated. In the end, we used 56 variables for our modeling dataset.

Even with this reduced dataset, there were still a number of missing variables that made modeling difficult. We experimented with a number of approaches for filling in these missing values, and eventually decided upon random forest imputation.

Once we had a set of schools with which we were happy, and a set of variables with values we could use, we explored this reduced dataset using principle component analysis, reducing all our variables to two "principle" components, and plotted the values of these components on a two dimensional grid. We then used k-means clustering to cluster the data into 7 groups, and these different groups are encoded using color below:

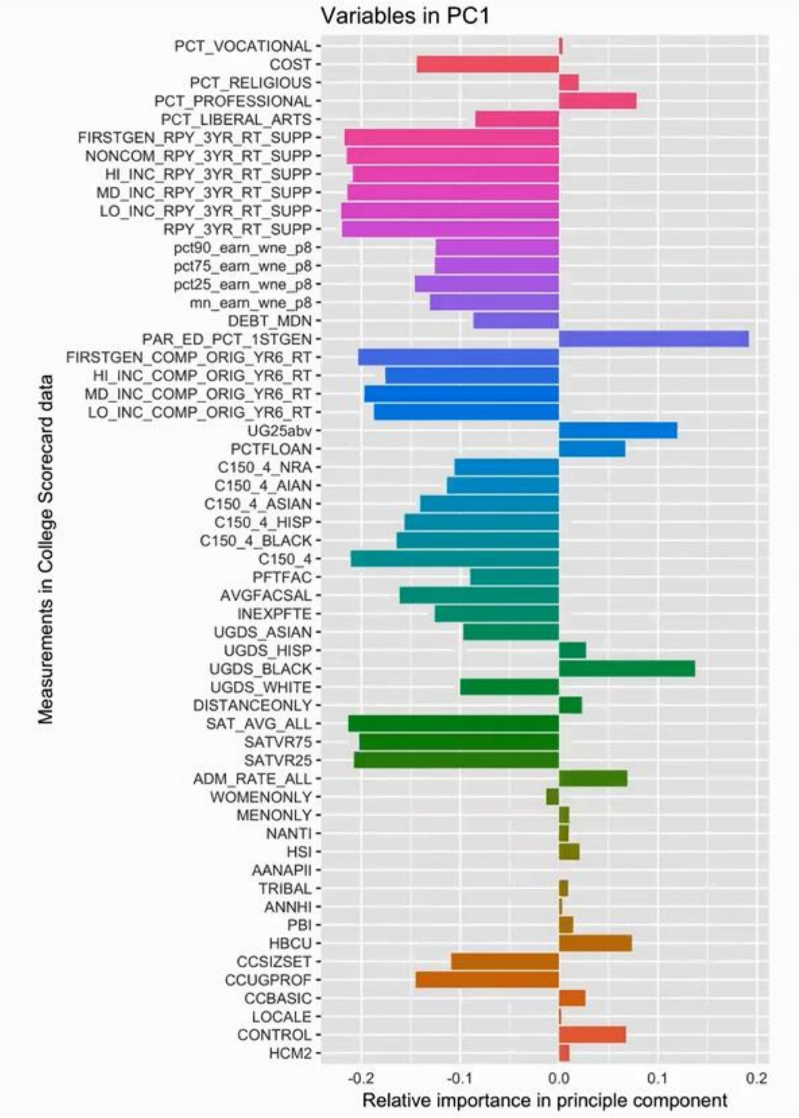


While many schools are clustered together around the origin, we can see some definite outliers, and these outliers correspond with our intuition, which is usually a good sign.

We investigated the factors that went into the principle components, and graphed these factors according to their contribution to the component. Below is the graph for the first principle component. As can be seen, repayment and completion rates are important for this component, as are SAT scores and demographic data.

We tried a number of approaches in determining how similar schools were, but in the end were most happy with the results we obtained using a cosine-similarity measure on our 56 variables with missing data imputed by random forest.

We thought the rankings were pretty good. You can use the visualization for this project to explore these rankings on your own and see if you agree.



Visualization Overview:

The overall CSS layout for our visualization uses the Skeleton CSS boilerplate, which we found to be easy to use and we think it looks good.

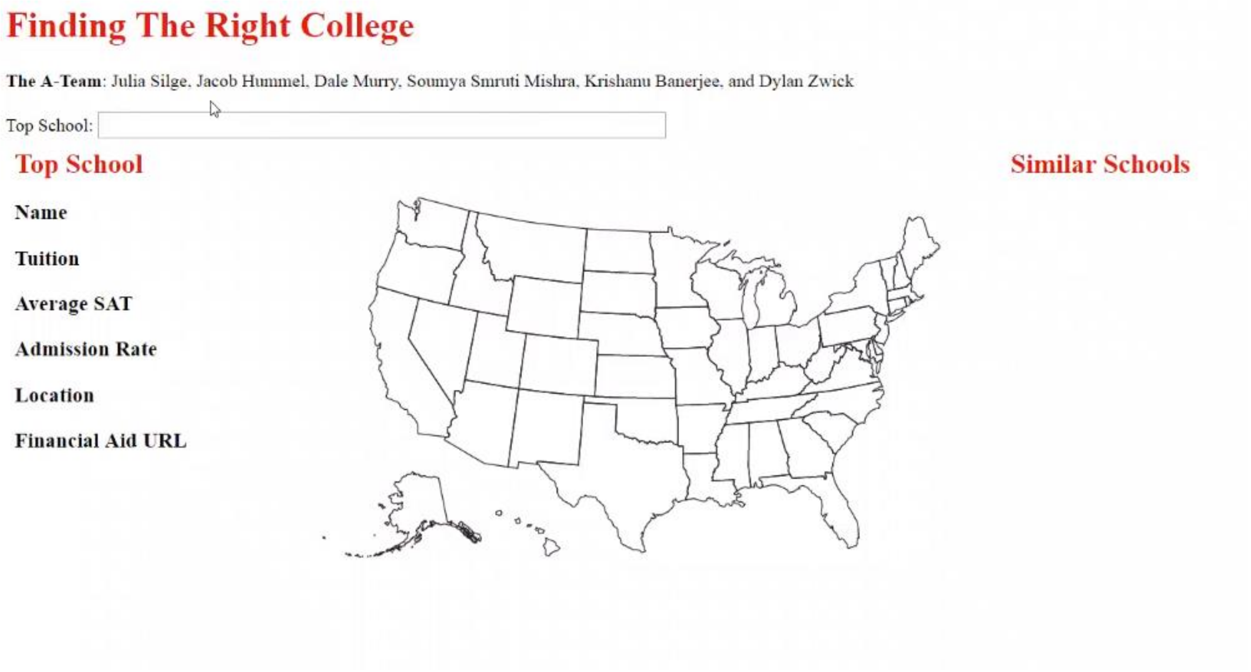
Skeleton CSS - http://getskeleton.com/

The visualization we produced for this project is structured around four interactive layouts:

1. Text Search
2. Ordered List
3. Interactive Map
4. Interactive Visual Filtering

Text Search

The initial, first-draft visualization for our dataset centered around a search bar and a (slightly) interactive map. The user was able to select a school by typing its name into a "Top School" search box:



Finding The Right College

The A-Team: Julia Silge, Jacob Hummel, Dale Murry, Soumya Smruti Mishra, Krishanu Banerjee, and Dylan Zwick

Top School:

Top School

Name

Tuition

Average SAT

Admission Rate

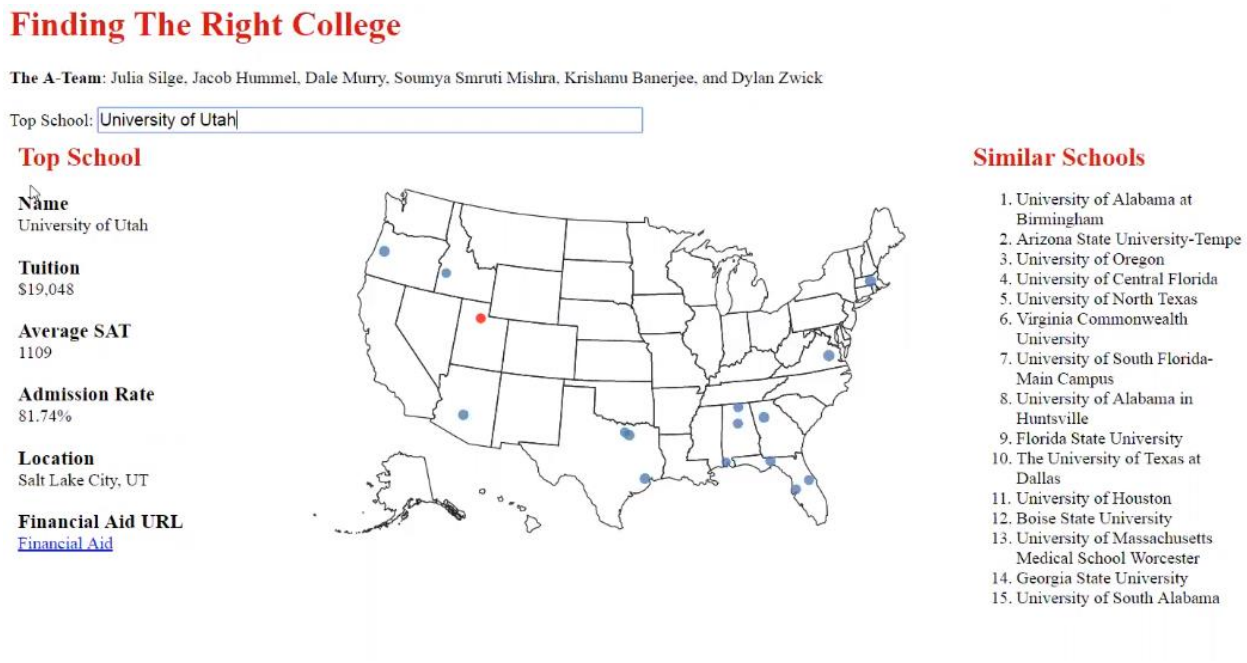
Location

Financial Aid URL

Similar Schools

Once that top school was selected, the user was shown some data about that school, a list of similar schools, and a map with the locations of the top school and similar schools marked out.

The user could select a similar school either from the list or from the map, and that school would then become the “top school”.



Finding The Right College

The A-Team: Julia Silge, Jacob Hummel, Dale Murry, Soumya Smruti Mishra, Krishanu Banerjee, and Dylan Zwick

Top School: University of Utah

Top School

Name
University of Utah

Tuition
$19,048

Average SAT
1109

Admission Rate
81.74%

Location
Salt Lake City, UT

Financial Aid URL
Financial Aid

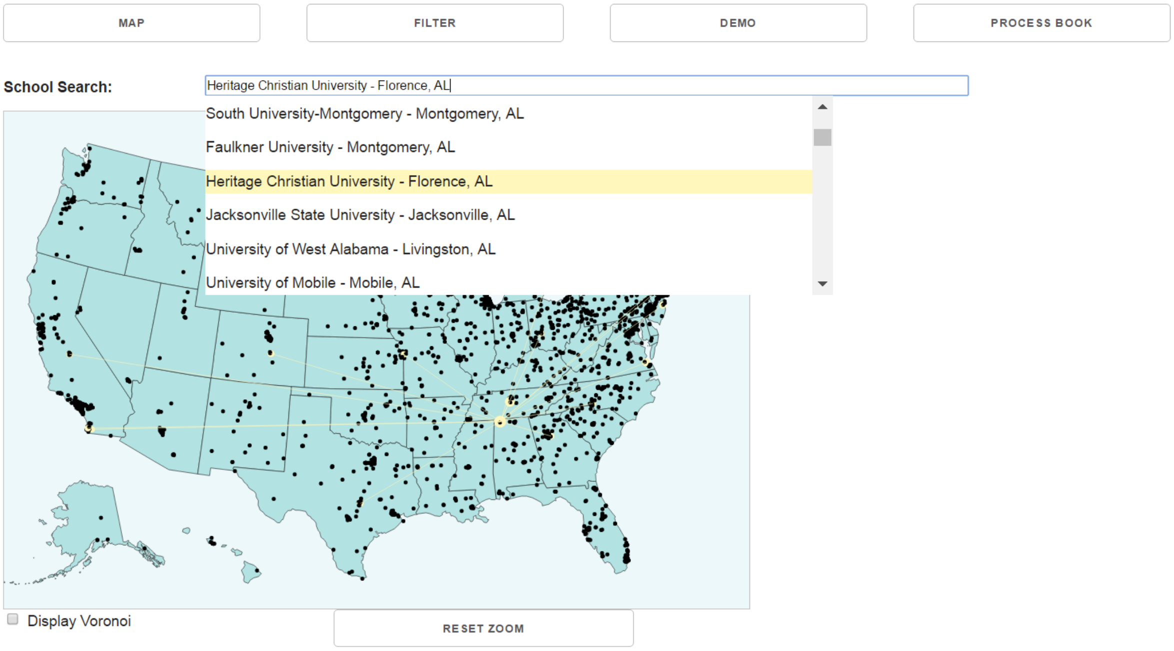
Similar Schools

1. University of Alabama at Birmingham
2. Arizona State University-Tempe
3. University of Oregon
4. University of Central Florida
5. University of North Texas
6. Virginia Commonwealth University
7. University of South Florida-Main Campus
8. University of Alabama in Huntsville
9. Florida State University
10. The University of Texas at Dallas
11. University of Houston
12. Boise State University
13. University of Massachusetts Medical School Worcester
14. Georgia State University
15. University of South Alabama

While this worked decently, there were some issues:

- It was difficult for the user to explore or discover new schools. If the user didn’t know the name of the school he or she was interested in, then the only way to discover a school was to start with a known school, and hopefully find a new school by exploring similar schools.
- There was the possibility of ambiguity in the school name. For example, there are multiple schools named “Westminster College”, and it was chance which one happened to be associated with the name in the search.
- The possible set of schools returned was not capped. This means if the student started typing “University of Utah”, after typing “Uni” *every* school that began with those letters (which is hundreds) would be suggested, and navigating among them was very difficult.

On the advice of our teaching assistant, Vinitha Yaski, we improved upon the search experience by limiting the number of suggestions to 100, and implemented scrolling functionality for the drop down options. Also, upon hover the option highlights, and if the option is available upon the map the corresponding school on the map also highlights.

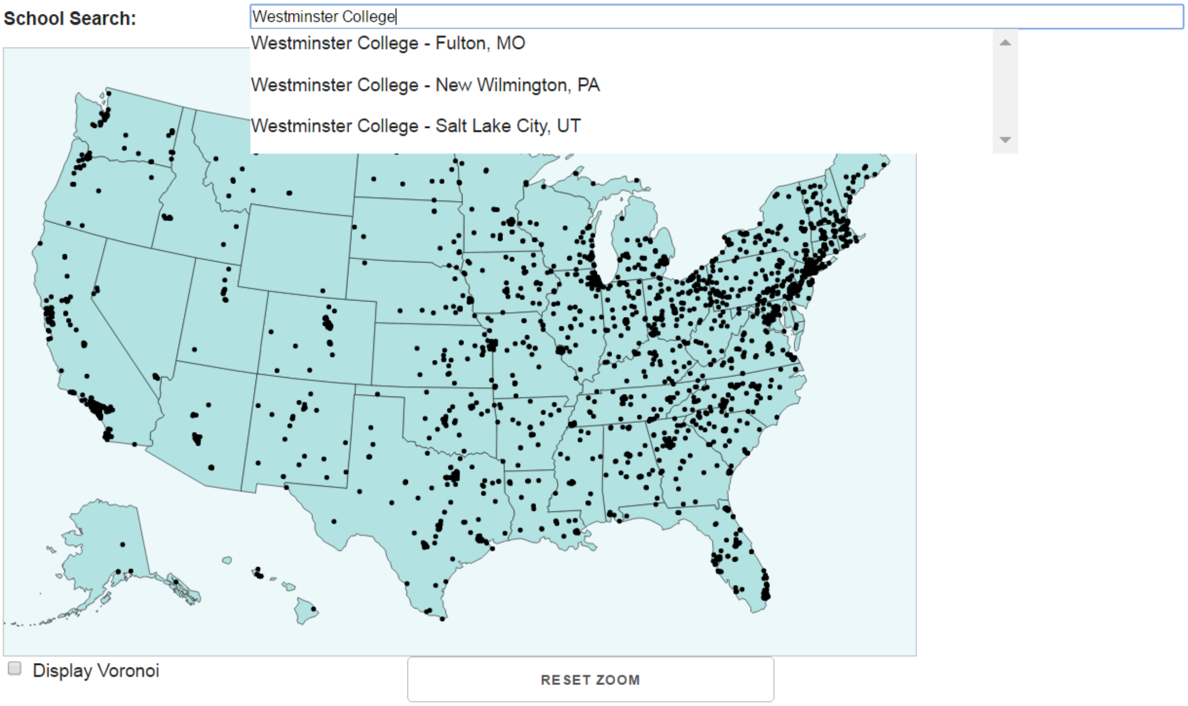


School Search: Heritage Christian University - Florence, AL

South University-Montgomery - Montgomery, AL
Faulkner University - Montgomery, AL
Heritage Christian University - Florence, AL
Jacksonville State University - Jacksonville, AL
University of West Alabama - Livingston, AL
University of Mobile - Mobile, AL

When the user hovers over a school in the search, if that school is present on the map it will highlight.

Finally, both the school name and city are displayed, allowing the user the ability to distinguish among schools with the same name:

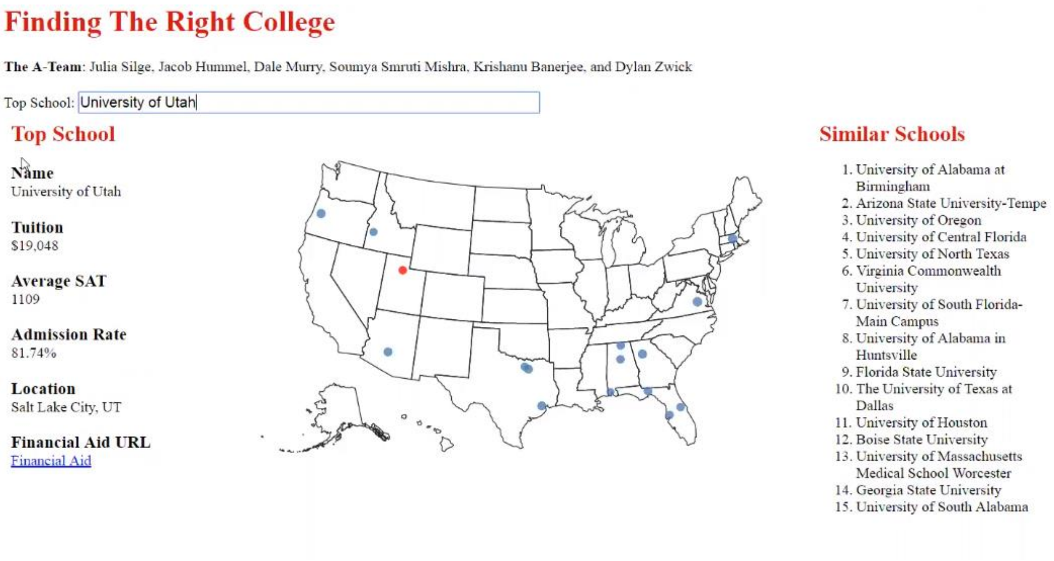


School Search: Westminster College

Westminster College - Fulton, MO
Westminster College - New Wilmington, PA
Westminster College - Salt Lake City, UT

Ordered List

In the initial visualization, the selected school information and the similar schools information were presented on different sides of the map:



Finding The Right College

The A-Team: Julia Silge, Jacob Hummel, Dale Murry, Soumya Smruti Mishra, Krishanu Banerjee, and Dylan Zwick

Top School: University of Utah

Top School

Name
University of Utah

Tuition
$19,048

Average SAT
1109

Admission Rate
81.74%

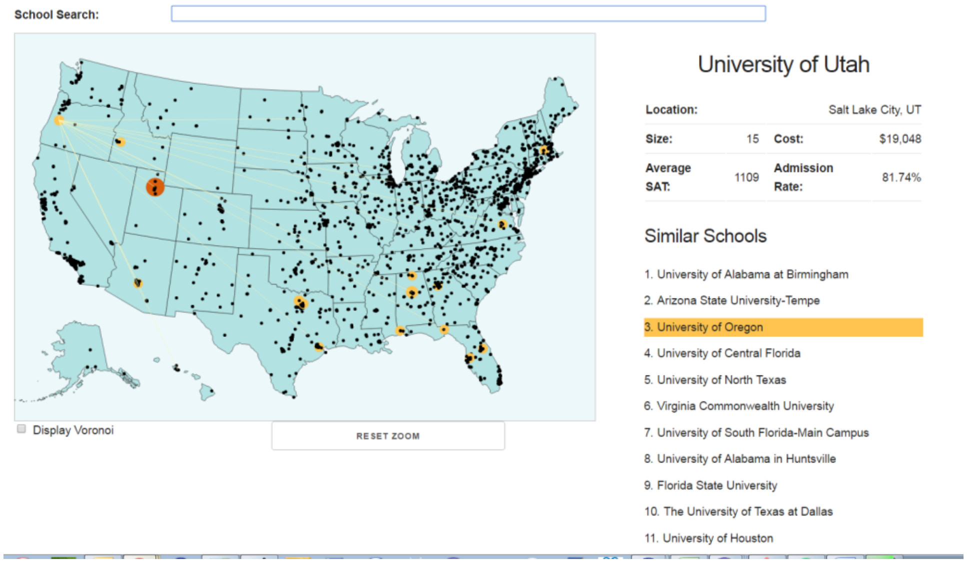
Location
Salt Lake City, UT

Financial Aid URL
Financial Aid

Similar Schools

1. University of Alabama at Birmingham
2. Arizona State University-Tempe
3. University of Oregon
4. University of Central Florida
5. University of North Texas
6. Virginia Commonwealth University
7. University of South Florida-Main Campus
8. University of Alabama in Huntsville
9. Florida State University
10. The University of Texas at Dallas
11. University of Houston
12. Boise State University
13. University of Massachusetts Medical School Worcester
14. Georgia State University
15. University of South Alabama

We felt it would be more visually appealing if the both sets of information were on the same side of the screen. The list is also now interactive, where if the user hovers over a similar school that school highlights on the map, and if the user clicks on a similar school that school becomes the selected school:



School Search:

☐ Display Voronoi

University of Utah

Location:			Salt Lake City, UT
Size:	15	**Cost:**	$19,048
Average SAT:	1109	**Admission Rate:**	81.74%

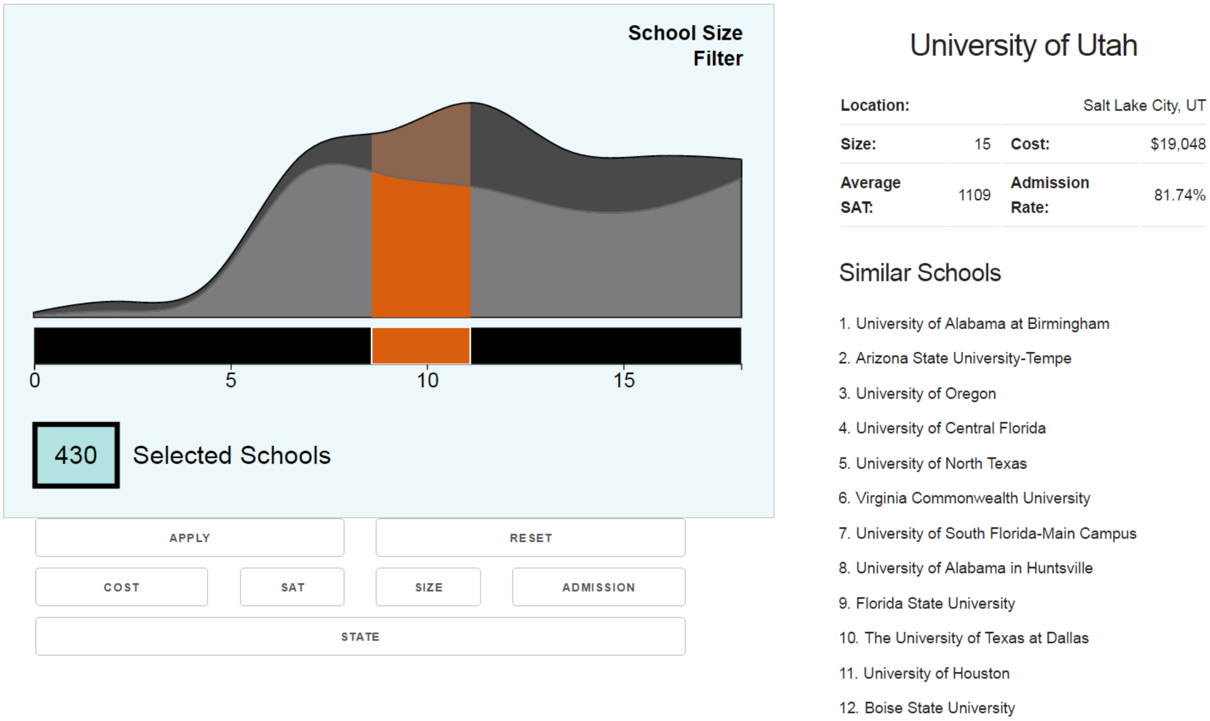
Similar Schools

1. University of Alabama at Birmingham
2. Arizona State University-Tempe
3. University of Oregon
4. University of Central Florida
5. University of North Texas
6. Virginia Commonwealth University
7. University of South Florida-Main Campus
8. University of Alabama in Huntsville
9. Florida State University
10. The University of Texas at Dallas
11. University of Houston

We also made the fonts much larger, and their style consistent.

Finally, we didn’t think it looked good to have empty lists or data tables before a selection was made, and so we keep these elements hidden until they’re populated.

The selected school, its data, and the schools similar to it are maintained across the different visualizations:



University of Utah

Location:			Salt Lake City, UT
Size:	15	Cost:	$19,048
Average SAT:	1109	Admission Rate:	81.74%

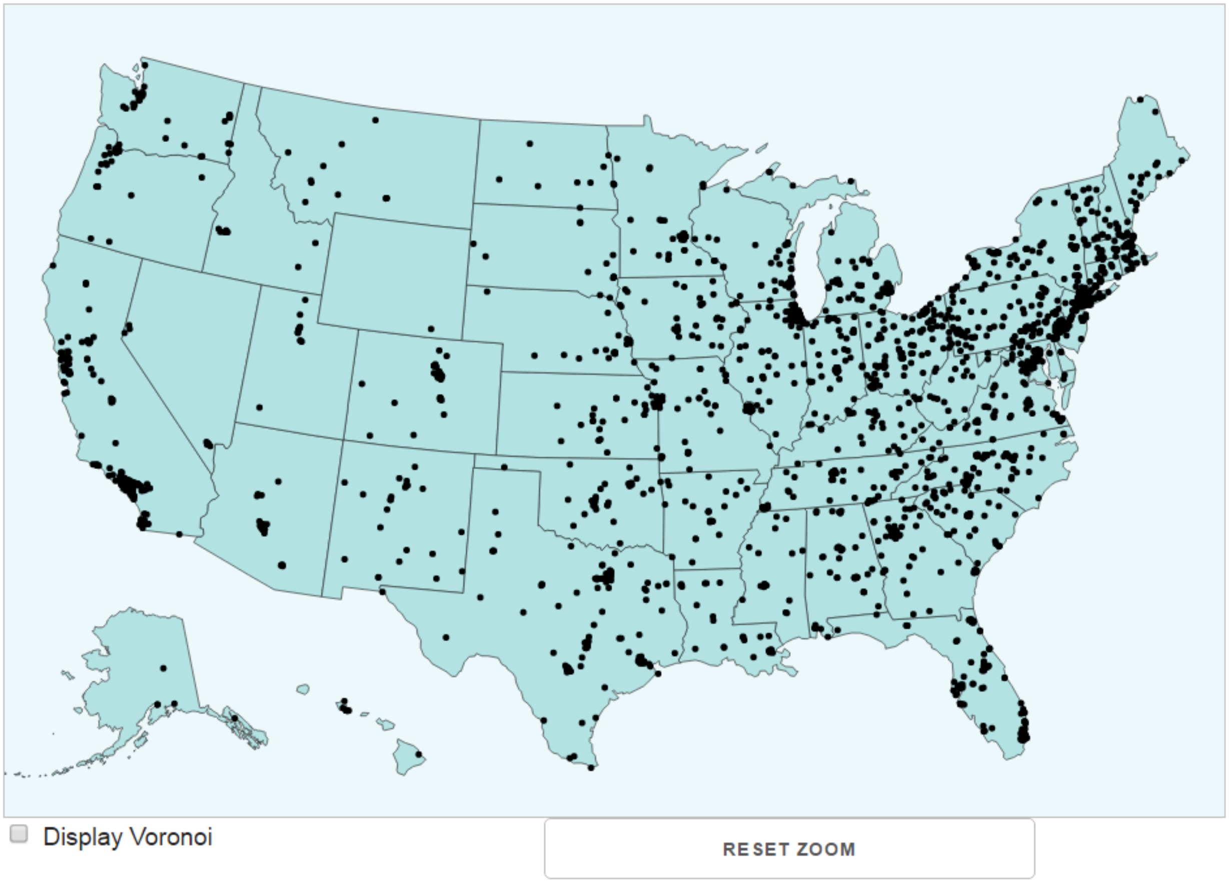
Similar Schools

1. University of Alabama at Birmingham
2. Arizona State University-Tempe
3. University of Oregon
4. University of Central Florida
5. University of North Texas
6. Virginia Commonwealth University
7. University of South Florida-Main Campus
8. University of Alabama in Huntsville
9. Florida State University
10. The University of Texas at Dallas
11. University of Houston
12. Boise State University

The Map Visualization:

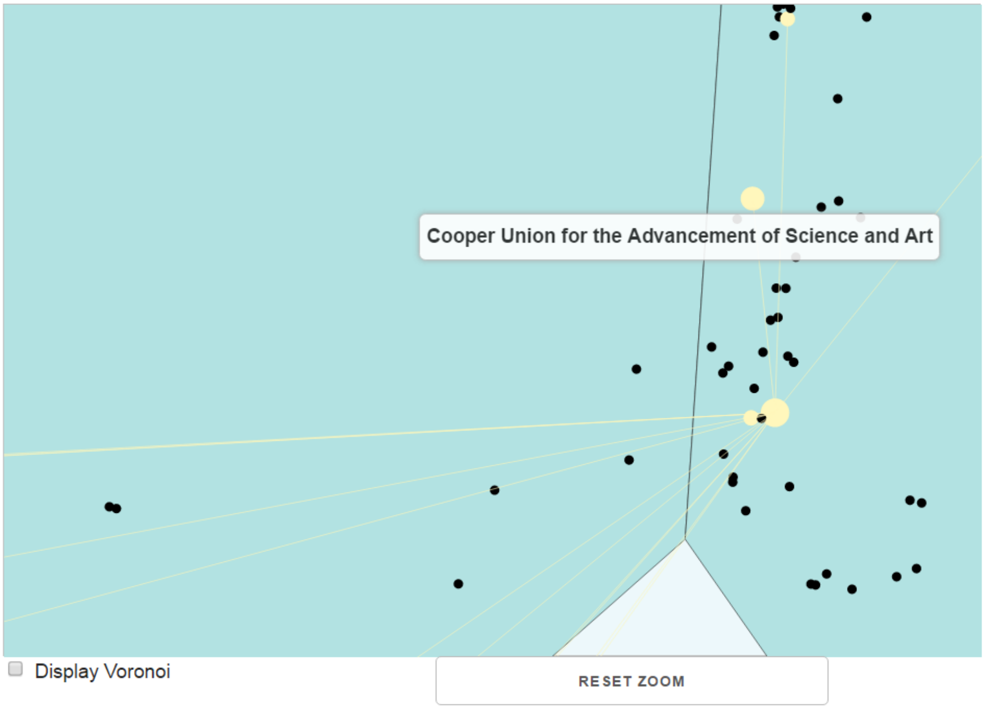
One issue with our initial map visualization is that it was difficult for a user to explore and discover new colleges. The only place to select the initial college was through the search bar.

However, there’s a problem with putting all the colleges on a map:



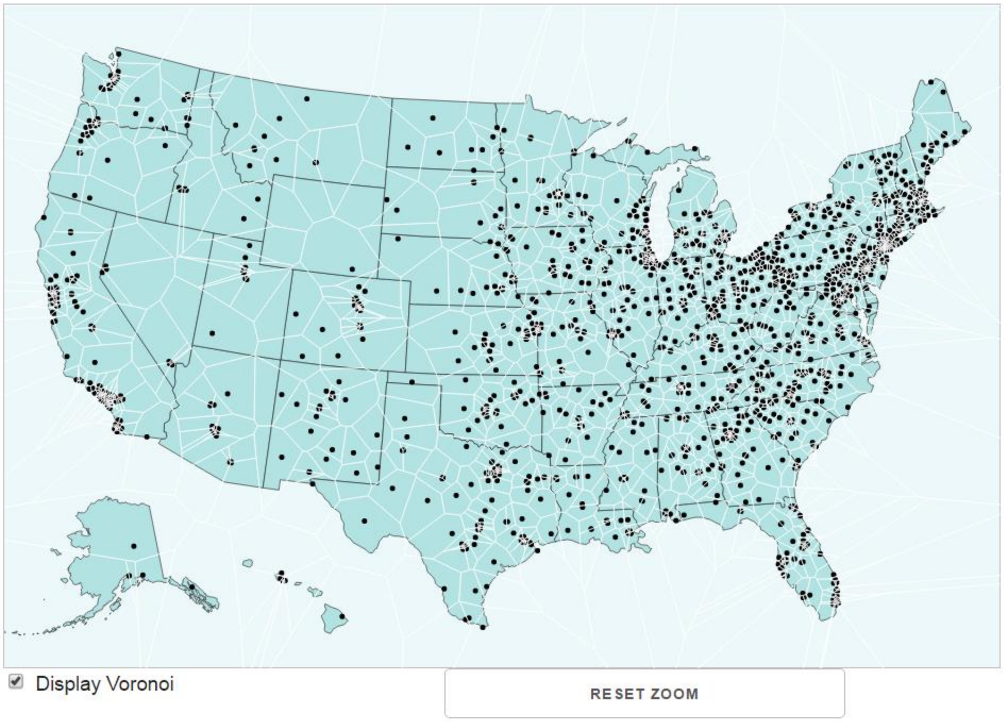
The map is crowded! Selecting a specific school, especially one in a region with a high population density, is extremely difficult.

To address this issue, and in keeping with the “details on demand” approach to data visualization, we implemented zooming on the map. If the user wishes to inspect a region of the map more closely, they simply use the scrollbar on their mouse for amplification:



To go back to the full map, the user can either zoom out with scrolling, or use the “reset zoom” button to return quickly to the original perspective.

With so many schools from which to select, it can also be difficult hovering over one particular school. To aid the user in selection, we overlayed the map layout with a Voronoi tessellation:

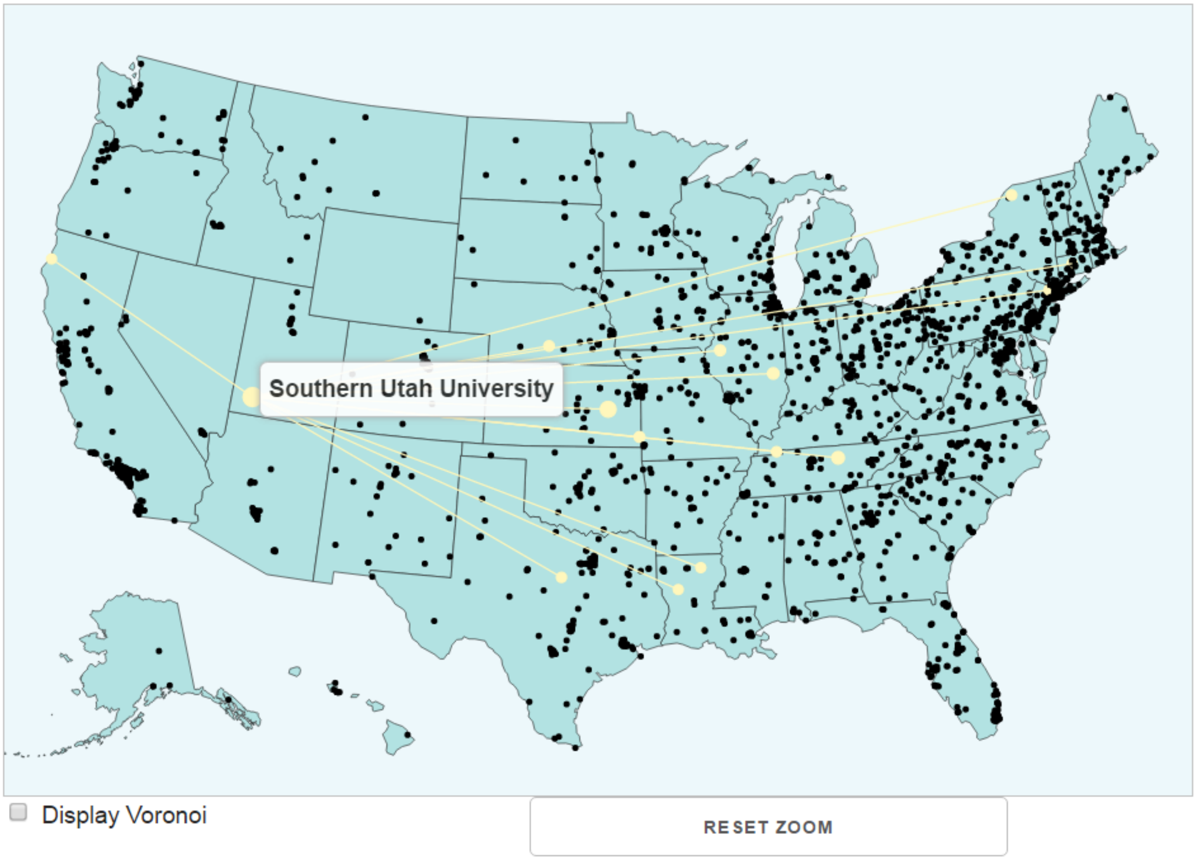


The Voronoi tessellation is a division of the map into polygons according to a set of points on the map, where every point in the interior of a polygon has a unique point in the set closest to it. The boundaries of the tessellation are the points equidistant between two or more points. In our visualization, every time the user hovers in a polygon, the school closest to it highlights and the user is able to select the closest school with a click of the mouse.

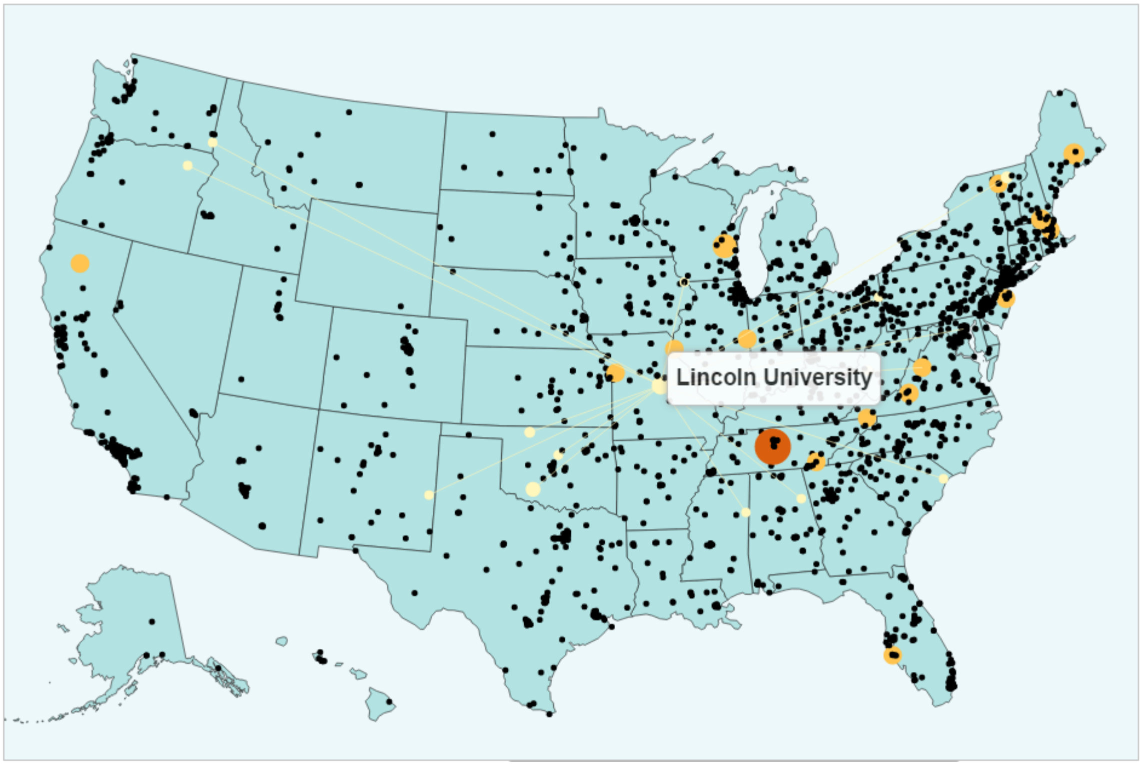
We felt combining zoom functionality with the Voronoi tessellation selection allowed us to strike a balance between allowing the user to explore a large amount of data without it becoming difficult or impossible to focus.

The Voronoi tessellation becomes visible to the user by the check of a box, but defaults to being hidden.

Finally, it would be nice if there were a way the user could hover over a school and know which schools are similar to it. We implemented this by drawing lines from the hovered school to the schools most similar to it:

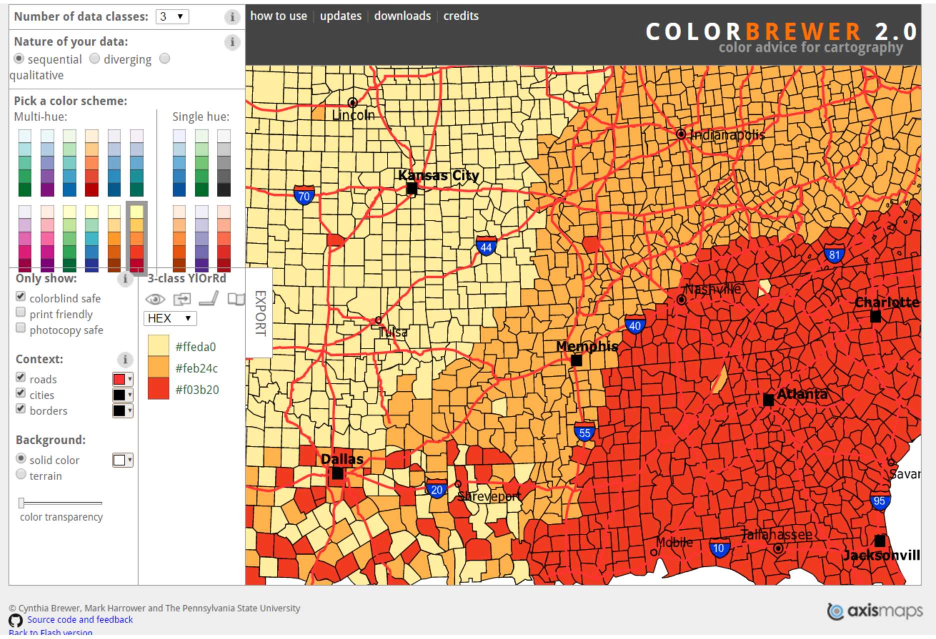


If the school is selected, its circle grows, and its color changes. The schools similar to the selected school also get larger circles, and their colors change as well.



The three colors used for hover, similar, and select were chosen using the Color Brewer tool, available online.

Color Brewer - http://colorbrewer2.org/#type=sequential&scheme=YlOrRd&n=3



Number of data classes: 3 ▾

Nature of your data:
● sequential ○ diverging ○ qualitative

Pick a color scheme:
Multi-hue:
Single hue:

Only show:
☑ colorblind safe
☐ print friendly
☐ photocopy safe

Context:
☑ roads
☑ cities
☑ borders

Background:
● solid color
○ terrain

color transparency

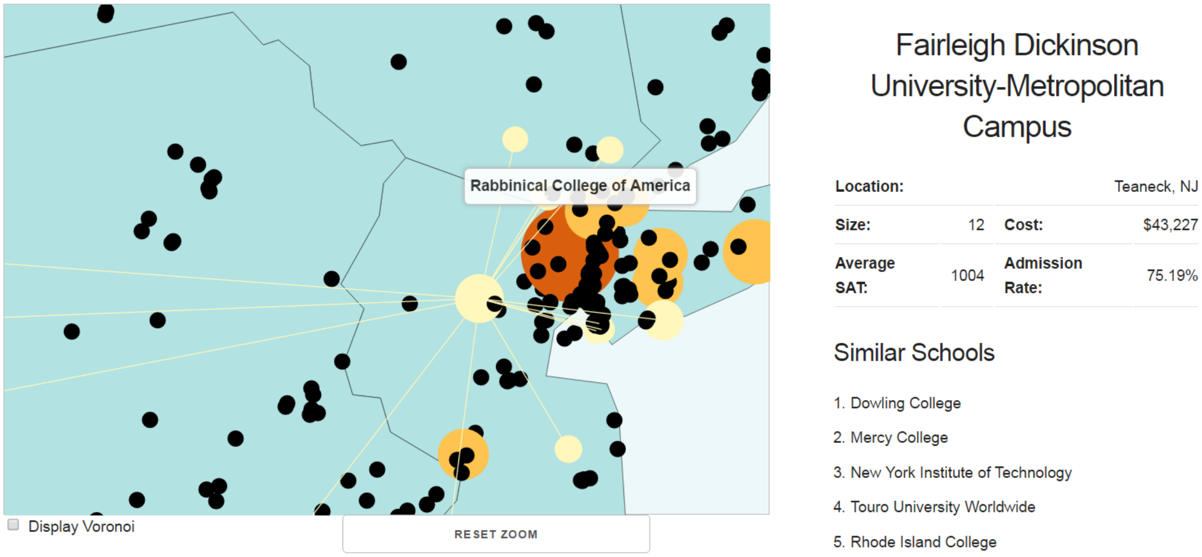
3-class YlOrRd

HEX ▾

🟨 #ffeda0
🟧 #feb24c
🟥 #f03b20

© Cynthia Brewer, Mark Harrower and The Pennsylvania State University
Source code and feedback
Back to Flash version

A closer view of these three colors on our map:



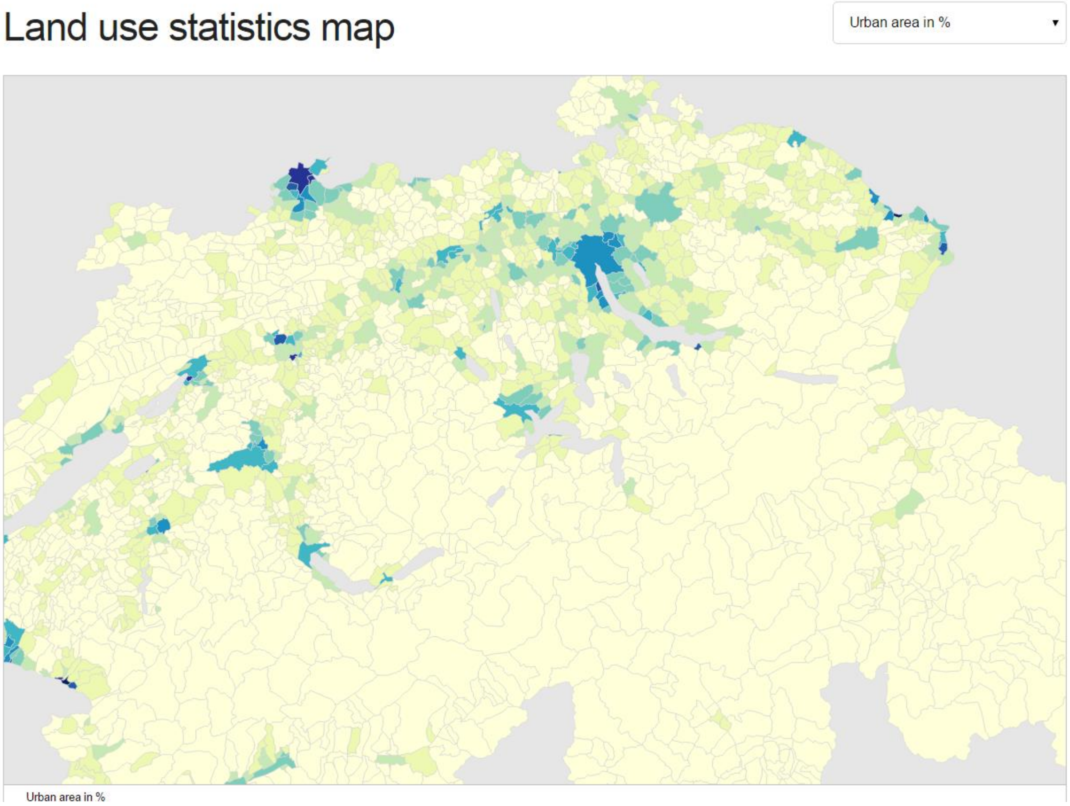
Fairleigh Dickinson University-Metropolitan Campus

Location:			Teaneck, NJ
Size:	12	Cost:	$43,227
Average SAT:	1004	Admission Rate:	75.19%

Similar Schools

1. Dowling College
2. Mercy College
3. New York Institute of Technology
4. Touro University Worldwide
5. Rhode Island College

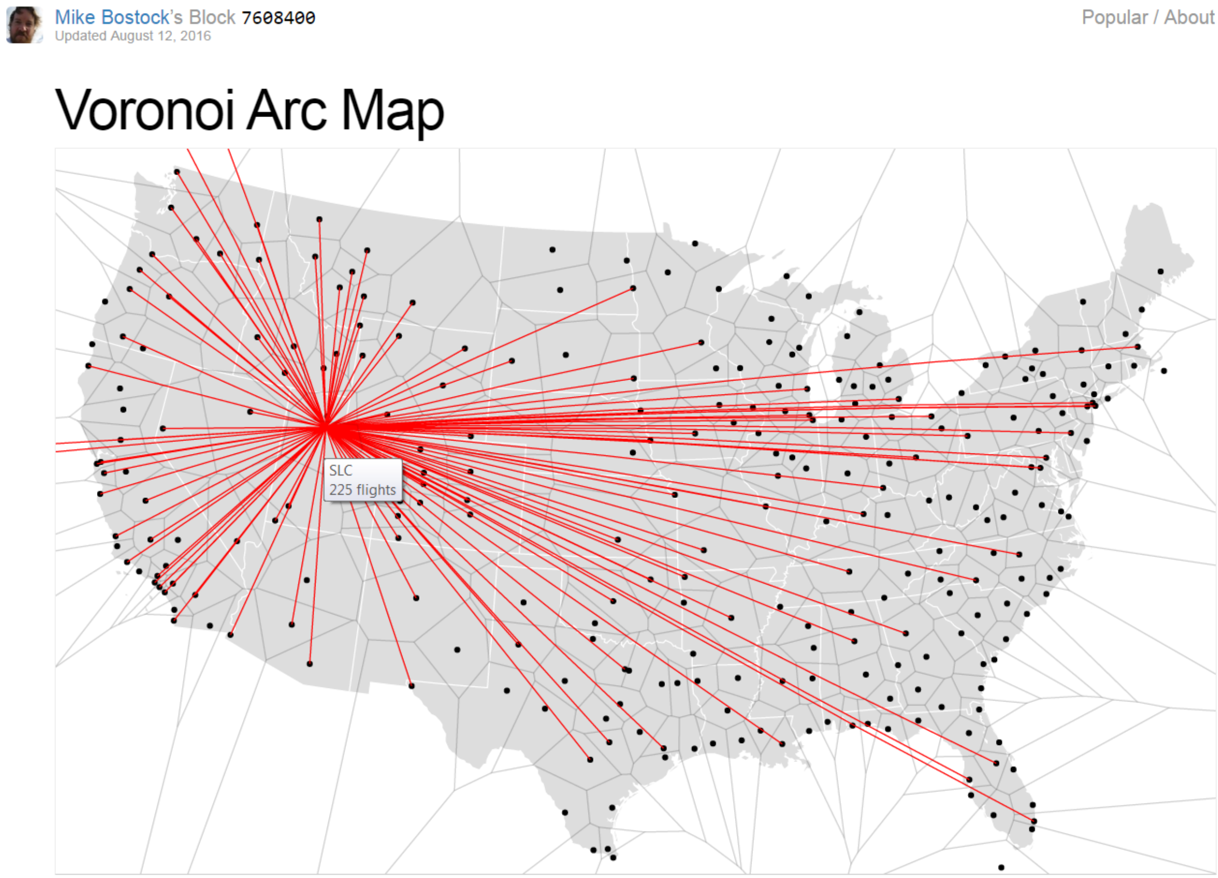
There are many examples of zooming on maps available online, but our primary inspiration for this part of the map visualization, both in terms of style and implementation, was the land use statistics map:



This map was put together as part of the data-map-d3 documentation tutorial provided by Lucas Vonlanthen for a workshop conducted at the University of Bern.

Tutorial - http://data-map-d3.readthedocs.io/en/latest/index.html

Our inspiration for using the Voronoi tessellation for selecting particular schools, and for using lines from those schools to represent similarity with other schools, came from Mike Bostock's Voronoi arc map block:



Mike Bostock's Block 7608400
Updated August 12, 2016

Popular / About

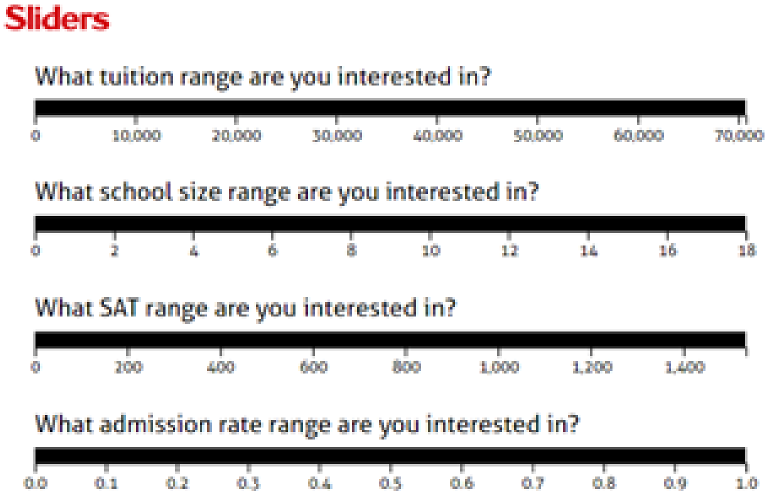
Voronoi Arc Map

Available Here - https://bl.ocks.org/mbostock/7608400

The Filter Visualization:

We wanted this visualization to be very interactive, giving the user broad discretion over exactly what data he/she brings in to the visualization. The decision that a teenager has in deciding which school to attend is a highly personal one, so to make this visualization work for a wide range of people, it was important for us to give the user the ability to home in on the precise range of school metrics he/she was interested in.

The filter began simply with a number of sliders that would filter the data accordingly. It was implemented first as a static element that stayed on the main page.



Sliders

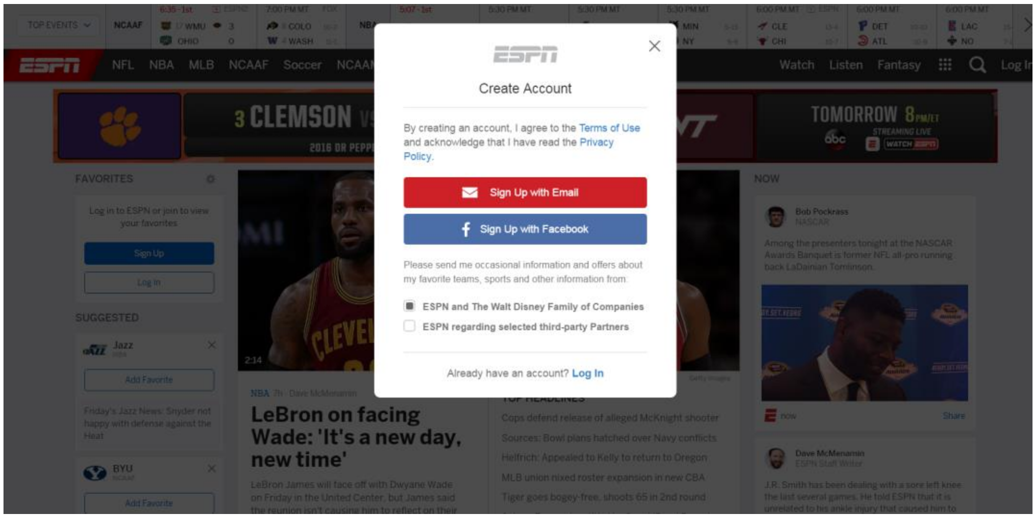
What tuition range are you interested in?

What school size range are you interested in?

What SAT range are you interested in?

What admission rate range are you interested in?

We then started playing around with the idea that the filter could be overlaid on the screen when the filter was clicked on. Below is an example from ESPN.com that shows how a panel can pop up on screen while there is a lights-out effect. This focuses the attention on the panel.



Create Account

By creating an account, I agree to the Terms of Use and acknowledge that I have read the Privacy Policy.

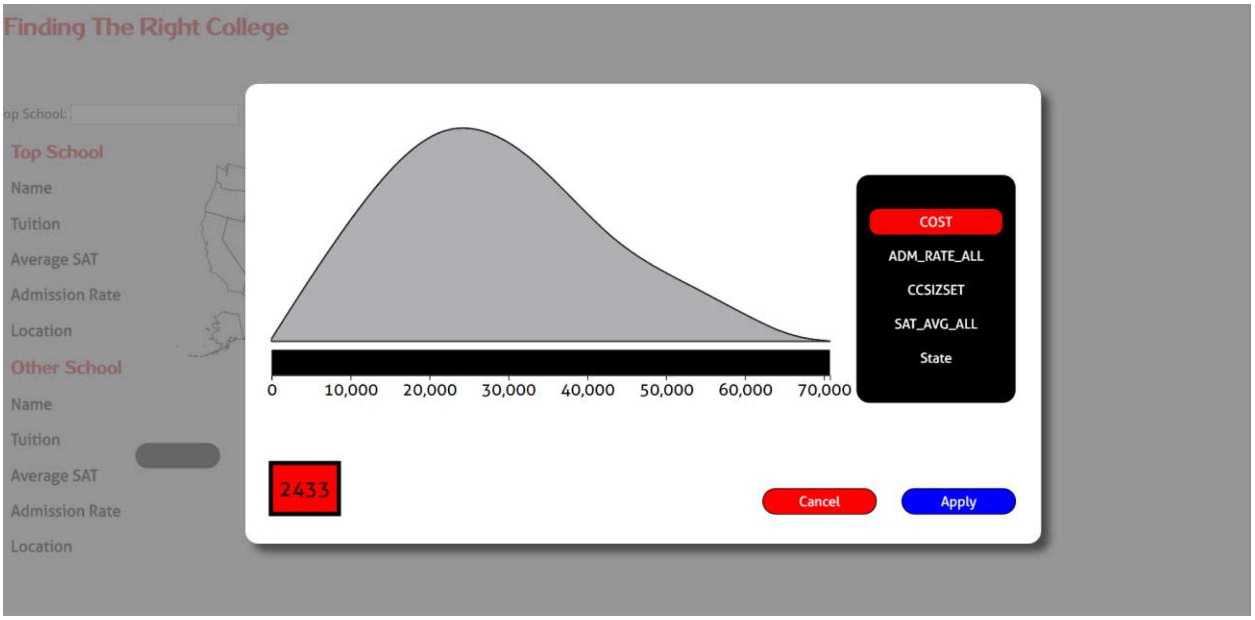
Please send me occasional information and offers about my favorite teams, sports and other information from:

● ESPN and The Walt Disney Family of Companies

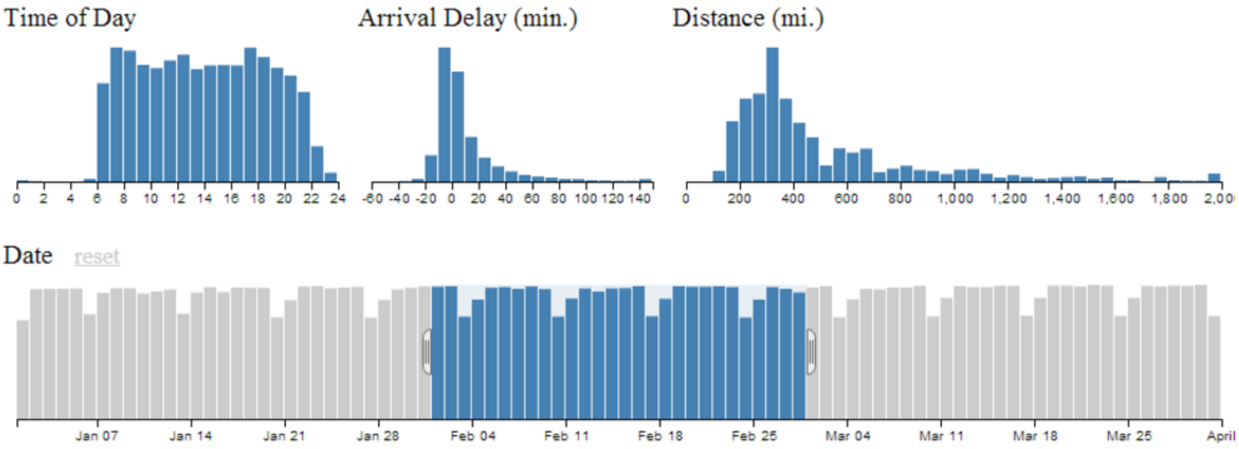
□ ESPN regarding selected third-party Partners

Already have an account? Log In

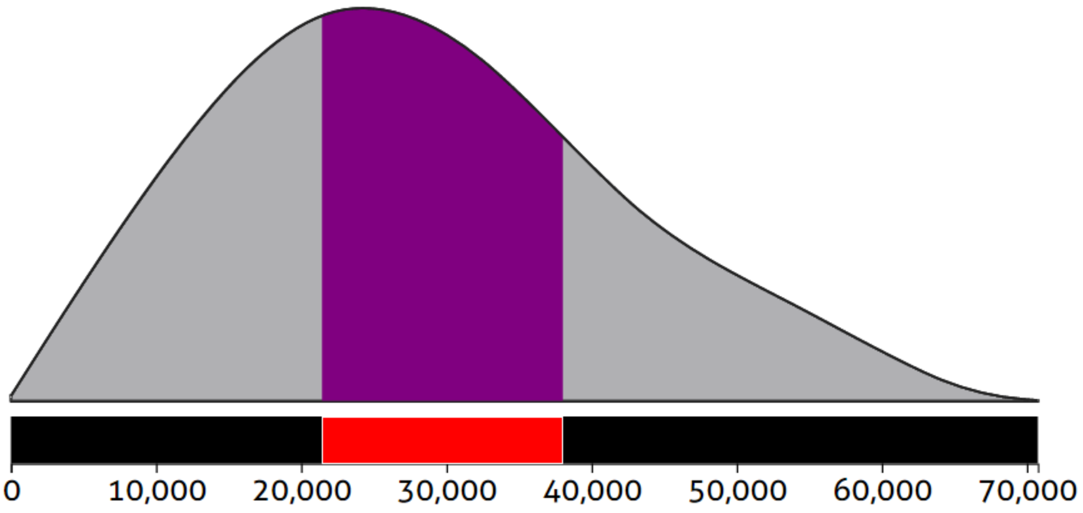
Below is our implementation of that effect. Eventually we went away from it but it was important in the evolution of our design to highlight the filter.



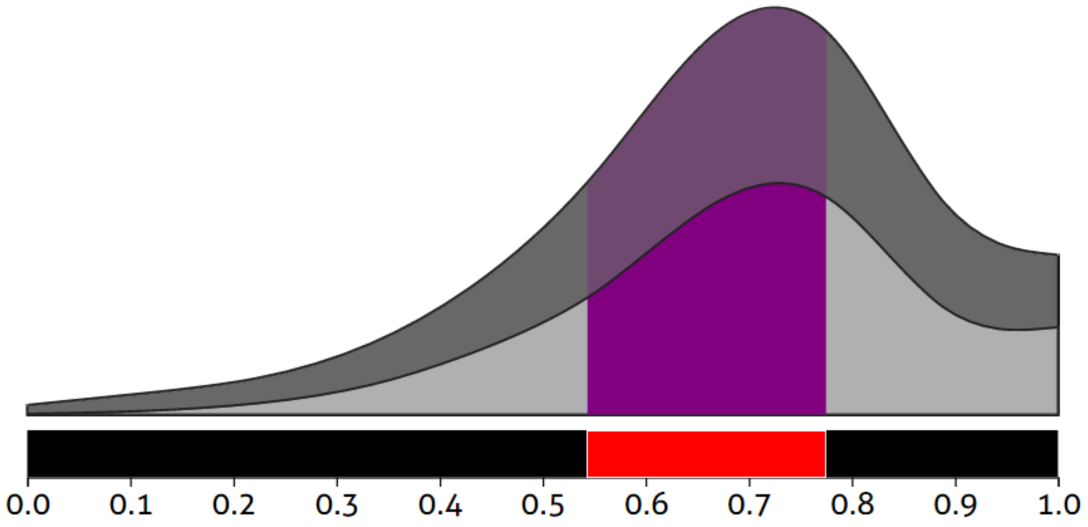
The screenshot above also shows the realization of a visual effect that was inspired by our conversation with Professor Lex who suggested that we might somehow use the feature of area to show the overall dimensions of the data. Being able to brush a filter will return the schools within some metric range, but it doesn't give any context for how that range fits into the overall distribution of schools. For any given metric range, there might be 1 school or 500 schools. The image below is a visualization that gets at what we were aiming for.



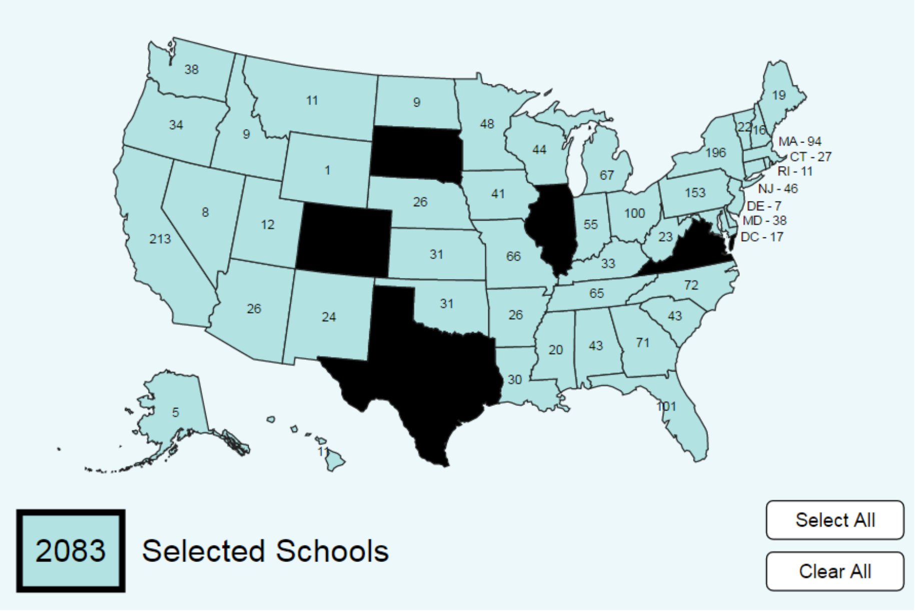
The graphic below shows how we used the idea of a frequency distribution to represent the number of schools in a given range. The general framework of the curve is built using the points from a frequency distribution bar graph but we then used linear interpolation (the d3.curveBasis function) for smoothing to create a continuous range.



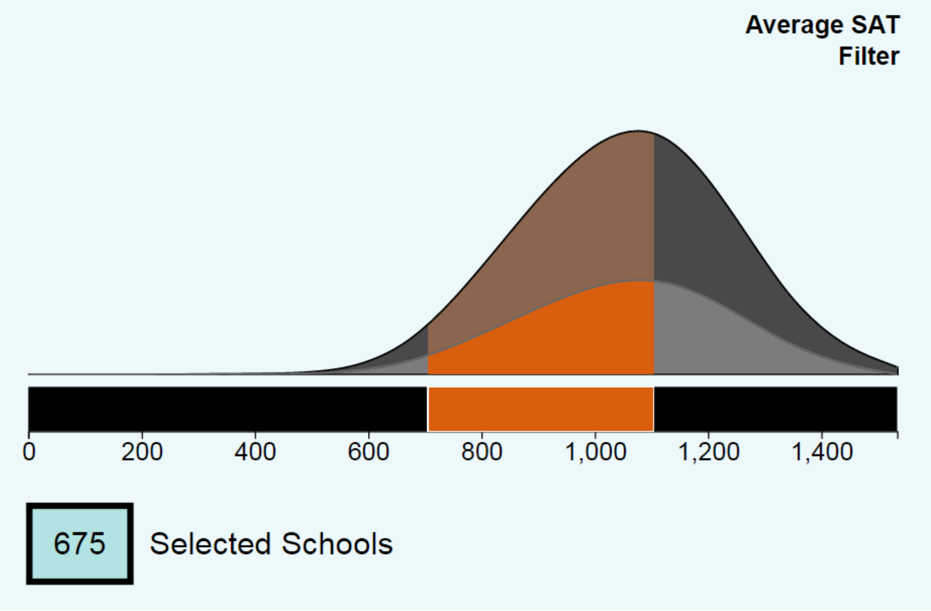
Our thinking evolved even further on this when we realized that not only did we want context for the overall data set, we also wanted context for the schools that the user had already filtered, so we came up with the design below that shows the overall school frequency distribution, the filtered school frequency distribution, and the highlight panel that gives more prominence to the selected school distribution.



We were also aware that the main feature of our visualization was a map so it was important that we gave the user the ability to filter geographically. Below is our final implementation of a map filter that allows a user to select and de-select schools.



Below is our final implementation of the slider filter with the frequency distribution.


Average SAT
Filter
0
200
400
600
800
1,000
1,200
1,400
675
Selected Schools