Intuitive Comparison of University Statistics

Samuel Leventhal and Unnikrishnan Rajagopalan

1 Index

2 Initial Proposal

2.1 Basic Info

Project Title: An Intuitive Look on College Comparisons
Authors: Samuel Leventhal & Unnikrishnan Rajagopalan
e-mail Addresses: Samuel (samlev@cs.utah.edu) Unnikrishnan (unniar@cs.utah.edu)
UIDs: Samuel (u0491567) Unnikrishnan (u1010114)
Repository: https://github.com/sam-lev/2016-dataviscourse-homework-Sam-Leventhal/tree/master/project

2.2 Background and Motivation

. The project proposed results from (a) an interest in designing an approachable and informative means for scouring potential colleges when applying for undergraduate or graduate schools and (b) to develop a measure and means of comparison between schools and departments based on "academic progress" and other academic measures such as tuition, rank etc. A major issue before applying to any university as such would be basic factors like rank, tuition, total number of applicants, enrollment etc, so our motivation was to capture and present an easy visualization which can be used to applicants when selecting an university of their choice in the US. The project at it's core initially was ten of two major components, i.e. a study into the relation of funding and publications and secondly a means of investigating prospective universities not by their attributes alone but *with respect* to other universities of interest.

Parts of HW5, HW4, and the visualization composed by OECD regional well-being inspired our thought process of how we are going to achieve our goals in the types of visualizations would be comprehensive in the information conveyed as well as simplistic on the eye.

2.3 Problems Faced Initially

Our initial conceptual problems originated in how to properly compare schools of very different caliber. For example, MIT is a top school for science, engineering and mathematics but might not be a college of choice for a user trying to pick a University interested in arts. So, we had to come up with a scheme which was essential to rank colleges on various parameters and build our data set accordingly. We presume this scheme is fair since we compared the data mined from different sources.

Second to this was how to manage and prioritize the large amount of data associated to each school. As a result we had to hand pick statistic with a comparative merit. Following this meant we had to address how many attributes would suffice and derive visualization techniques which could span across numerous attributes while remaining informative.

2.4 Project Objectives

The aim of the project is to develop a visualization which takes advantage of the fact that choosing a prospective college is a process of comparisons and not solely based on one college's attributes. We therefore worked to construct a visualization which would allow the user to easily understand and compare numerous college statistics such as tuition cost, ranking, average salary after graduation and so on. Among others a specific college attribute of interest would be for example the number of publications produced for each school department within a year. Using rate of publications as a measure will allow the user to compare schools based on academic progress as well as see the change in academic progress between departments within a school over time. Secondary to this if the data is available we hope to incorporate other important school attributes like funding which would be essential factor for choosing a school.

In terms of utility the visualization will allow for an intuitive alternative to the what is now a tedious and difficult process of determining which schools are a best fit for a student moving into higher education. Much of the web crawling now required during the school selection process will be well contained in an approachable and information rich visualization in that users are able to display similar schools based on their criteria as well as compare specific schools by selection or school attributes of interest.

A second result of interest will be the ability to gauge the academic progress, tuition fees, acceptance rate, etc... of a school or schools over time. In this the development and future direction of school departments will be better estimated by the user.

2.5 Data

- College Statistics (Database: https://nces.ed.gov/ipeds/datacenter/
  - College names
  - College ranking
  - Cost of tuition
  - Number of Scholarships awarded with respect to number of students
  - Acceptance rate
  - Average GPA, SAT, GRE, ACT, MCAT, ect for accepted students
- Funding Attributes (Database: As above)
  - Admissions during Fall,Spring,Summer

  - Funding per department.
  - Source of funding.

- Quality Of Life Statistics (Database:https://www.oecdregionalwellbeing.org)
  - Quality of life for city
  - Cost of living
  - average salary

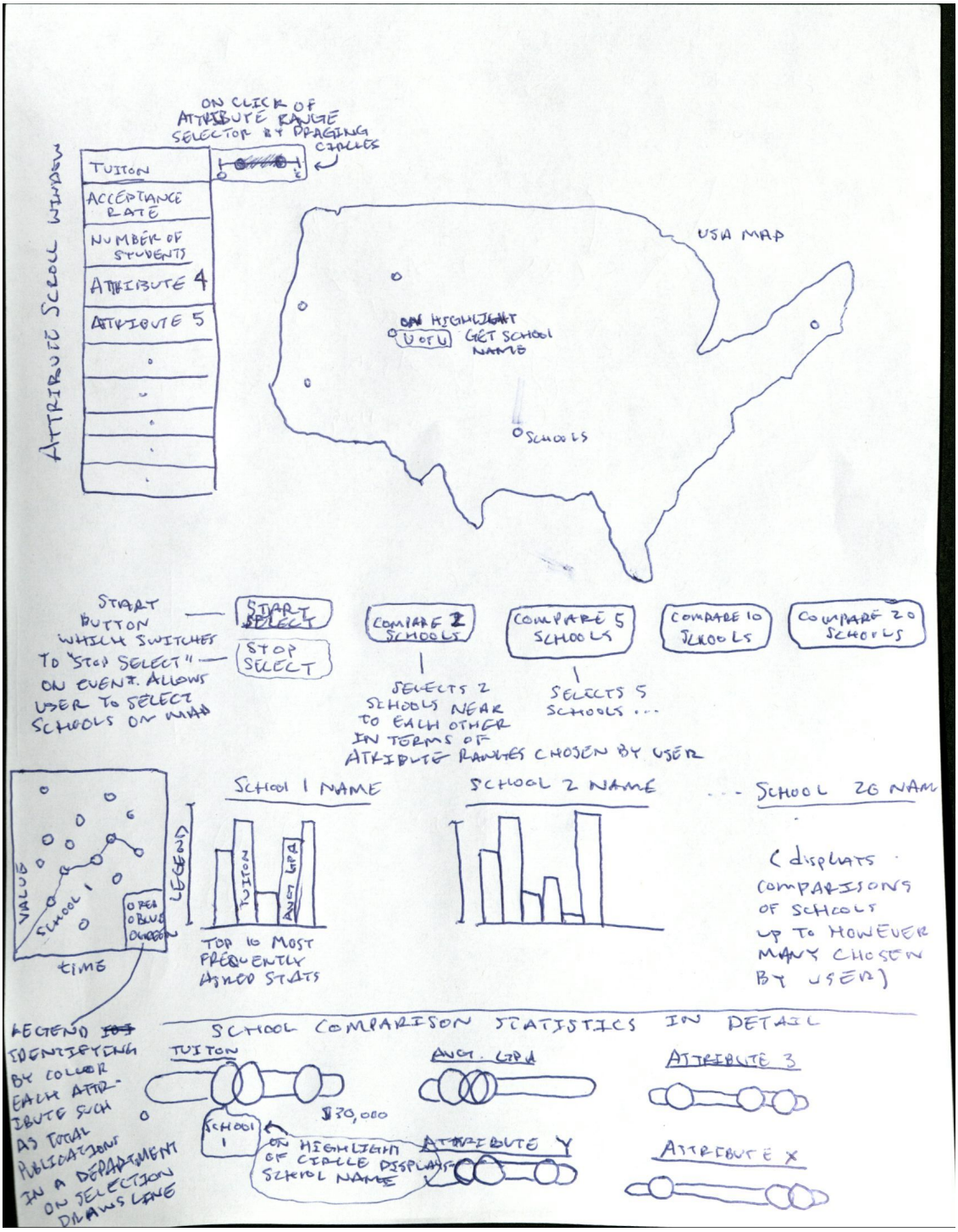
- Academics Progress Statistics (Database: https://www.nsf.gov/statistics/data-tools.cfm
  - Total publications for entire school in one year
  - Total publications for each department in one year

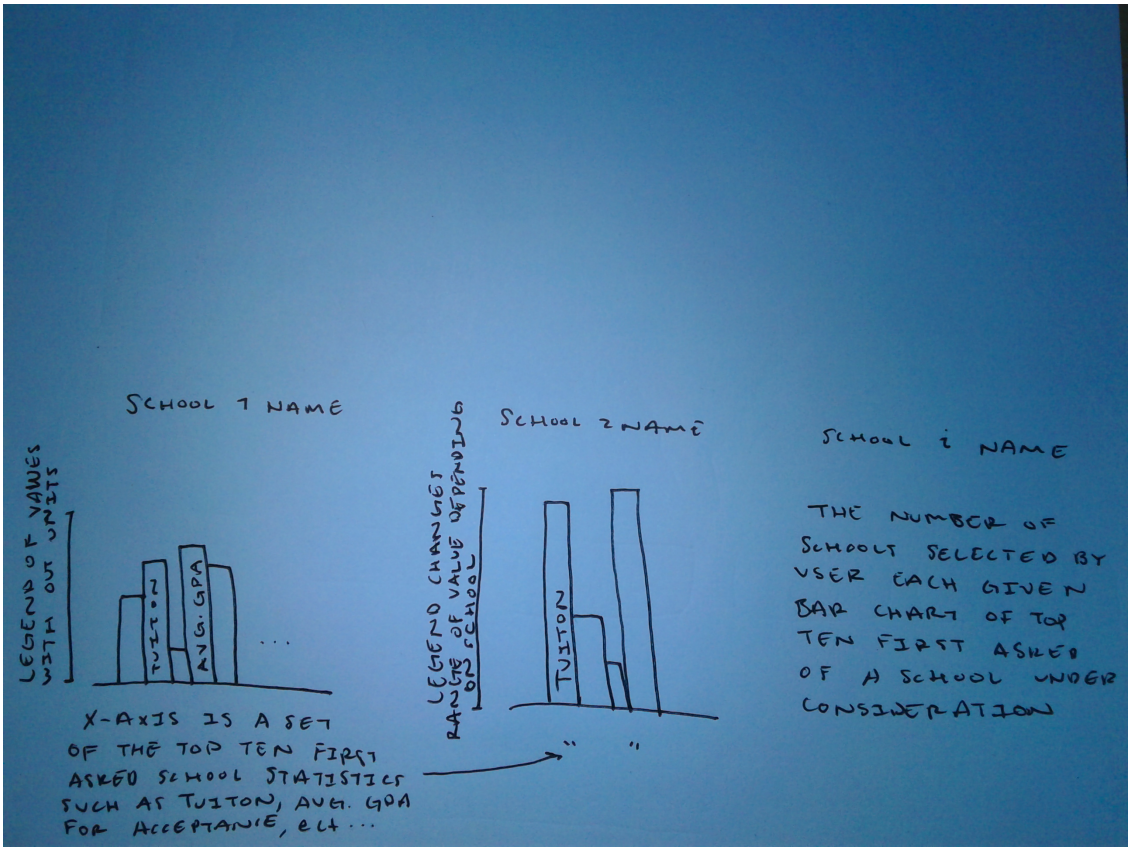
2.6 Data Processing

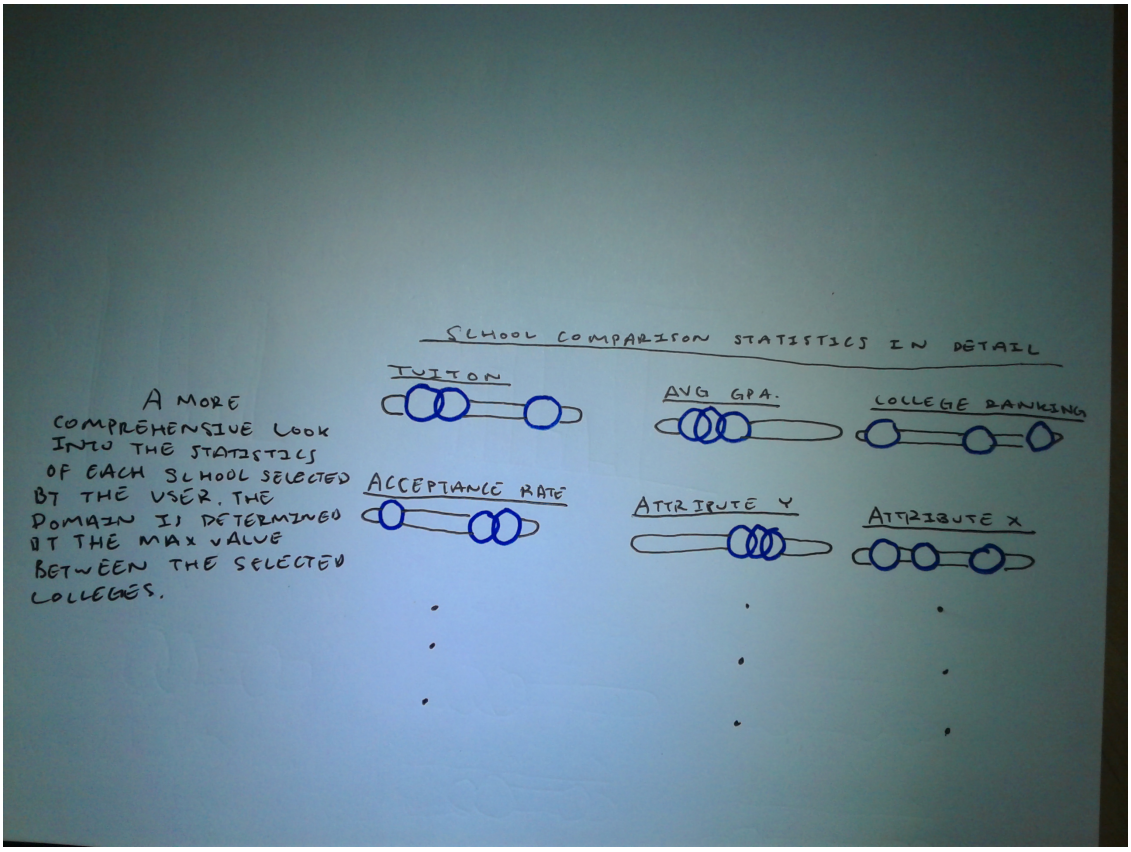
There are near 1016 universities in the United States. In order to visualize these universities as circles on a map we may need to create a criteria on which schools will be included in the visualization. We may therefore need to do a basic parsing of the data to only incorporate schools satisfying these attributes. Second to this we may also need to data mine sites such as arxiv.org to organize the data by school, school department, and publication year. For funding we will also need to cluster the data in this way however also include source of funding and amount.

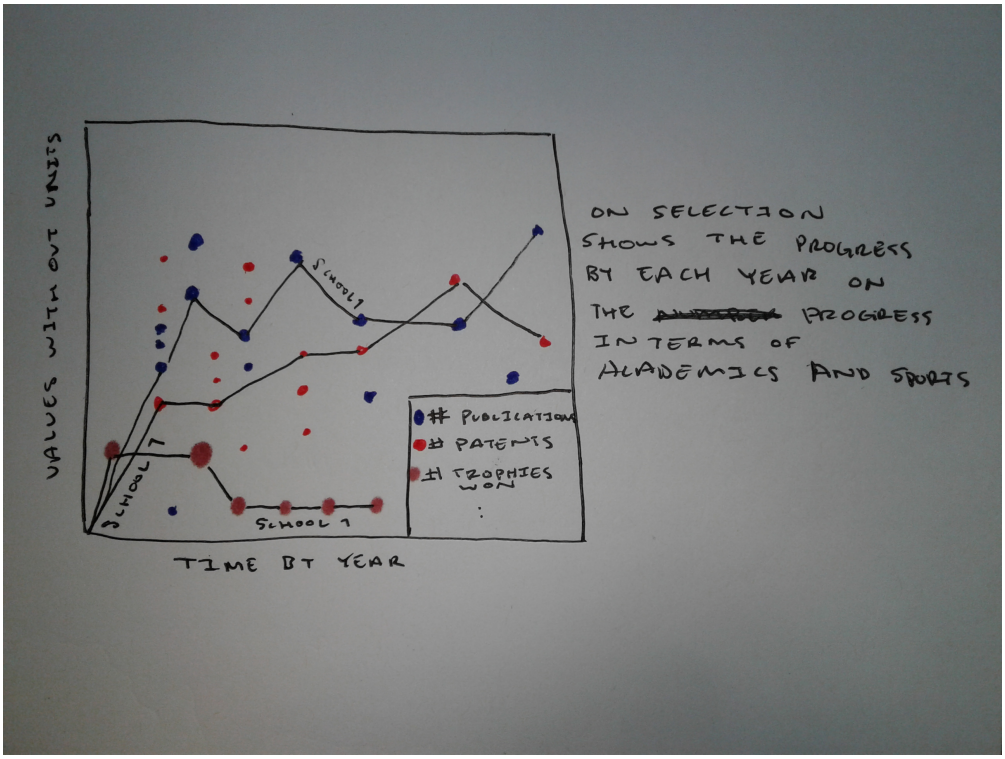
2.7 Visualization Design

The project will consist of (a) a selection bar on the left hand side where the user is able to determine which features and for what range values they would like to consider, (b) a map to it's right, (c) circled points on the map where each school is located, (d) five buttons beneath the map one of which allows the user to select specific schools to compare and the rest of which compare preset number of schools, i.e. if ten schools are selected then the ten nearest schools matching the user set criteria will be compared. Beneath will be (e) histograms demonstrating the top 10 attributes most often asked about schools such as tuition, general well being, acceptance rate, and so on. The x-axis of the histograms will be ambiguous due to the attributes not being correlated however each rectangle will be labeled as the attribute it represents. The scales of each rectangle will then be set with linear scaling whose domain is determined by the maximum values of the schools selected. And lastly, (f) beneath the histograms will be oval ranged gauges with circles representing where the selected schools are placed accordingly for all school statistics. A similar visualization can be seen on www.oecdregionalwellbeing.org. For a sketch of our intended visualization see below.


ON CLICK OF ATTRIBUTE RANGE SELECTOR BY DRAGING CIRCLES
ATTRIBUTE SCROLL WINDOW
TUITION
ACCEPTANCE RATE
NUMBER OF STUDENTS
ATTRIBUTE 4
ATTRIBUTE 5
USA MAP
ON HIGHLIGHT U OF U GET SCHOOL NAME
SCHOOLS
START BUTTON WHICH SWITCHES TO "STOP SELECT" ON EVENT. ALLOWS USER TO SELECT SCHOOLS ON MAP
START SELECT
STOP SELECT
COMPARE 2 SCHOOLS
COMPARE 5 SCHOOLS
COMPARE 10 SCHOOLS
COMPARE 20 SCHOOLS
SELECTS 2 SCHOOLS NEAR TO EACH OTHER IN TERMS OF ATRIBUTE RANGES CHOSEN BY USER
SELECTS 5 SCHOOLS ...
VALUE
SCHOOL 1
TIME
LEGEND
SCHOOL 1 NAME
SCHOOL 2 NAME
... SCHOOL 20 NAM
TUITION
AVG. GPA
TOP 10 MOST FREQUENTLY ASKED STATS
(displays comparisons of schools up to however many chosen by user)
LEGEND IDENTIFYING BY COLOR EACH ATTRIBUTE SUCH AS TOTAL PUBLICATIONS IN A DEPARTMENT ON SELECTION DRAWS LINE
SCHOOL COMPARISON STATISTICS IN DETAIL
TUITION
AVG. GPA
ATTRIBUTE 3
SCHOOL 1
$30,000
ON HIGHLIGHT OF CIRCLE DISPLAYS SCHOOL NAME
ATTRIBUTE Y
ATTRIBUTE X


SCHOOL 1 NAME
LEGEND OF VALUES WITH OUR UNITS
TUIT.
AVG. GPA
...
X-AXIS IS A SET OF THE TOP TEN FIRST ASKED SCHOOL STATISTICS SUCH AS TUITON, AVG. GPA FOR ACCEPTANCE, ECT...
LEGEND CHANGES RANGE OF VALUE DEPENDING ON SCHOOL
SCHOOL 2 NAME
TUITON
"
"
SCHOOL i NAME
THE NUMBER OF SCHOOLS SELECTED BY USER EACH GIVEN BAR CHART OF TOP TEN FIRST ASKED OF A SCHOOL UNDER CONSIDERATION


SCHOOL COMPARISON STATISTICS IN DETAIL
A MORE COMPREHENSIVE LOOK INTO THE STATISTICS OF EACH SCHOOL SELECTED BY THE USER. THE DOMAIN IS DETERMINED BY THE MAX VALUE BETWEEN THE SELECTED COLLEGES.
TUITION
AVG GPA.
COLLEGE RANKING
ACCEPTANCE RATE
ATTRIBUTE Y
ATTRIBUTE X


VALUES WITHOUT UNITS
SCHOOL 1
SCHOOL 2
SCHOOL 3
PUBLICATIONS
PATENTS
TROPHIES WON
TIME BY YEAR
ON SELECTION SHOWS THE PROGRESS BY EACH YEAR ON THE PROGRESS IN TERMS OF ACADEMICS AND SPORTS

2.8 Must-Have Features

List the features without which you would consider your project to be a failure. Necessary to the visualization include a map projection, clickable point representing schools on the map which update a histogram and scale chart, a hoverable tooltip for points displaying specific college attributes, a legend with school statistics, click feature for each school statistic in the legend which opens a range selector in which one drags to points along a legend to represent what range to be considered, the option to select specific schools to compare, the option to compare a preset number of similar schools which fall into the user defined features of interest and their ranges, histograms for each selected school displaying it's specific attributes, a more detailed set of all features for each school consisting of oval bars representing an attribute's range and circles for the value of each selected school.

2.9 Optional Features

List the features which you consider to be nice to have, but not critical. The feature which we feel nice to have would be to provide a visualization similar to GapMinder by Hans Rosling which was discussed in the class. Using this type of visualization we would be able to show the change in the school's performance over time. The attributes we would consider are academic progress in terms of patents, publications, sports, endowment etc.

2.10 Project Schedule

. Make sure that you plan your work so that you can avoid a big rush right before the final project deadline, and delegate different modules and responsibilities among your team members. Write this in terms of weekly deadlines.

- Week 1 after proposal - Mine data from the sources
- Week 2 - Map projections with Universities plotted
- Week 3 - Attribute selection Legend and selection comparison
- Week 4 - Guage visualizations for selected universities
- Week 5 - Work on GapMinder type visualization

3 Peer Feedback

Feedback Group:

- Shane Brown (u0852900)
- Sigmund Chow (u0597938)
- My Huynh (u0729654)

We found the feedback given very helpful and informative. Discussing our project illuminated aspects that were convoluted, allowing for new approaches to be either recommended by our reviewers or derived by us. Below are the suggestions or resolved confusions which we believe will provide for a more informative and intuitive visualization:

1. **Issue**:The bar charts for each school selected by the user add an extra component of confusion by grouping all top ten attributes for each school since attributes do not share the same units so height comparison is meaningless.

   Solution: We have decided to instead of creating a bar chart for each school we can better preserve the main goal of school comparison by making ten bar charts for each of the "top ten first asked school attribute" within which is a bar for each of the schools selected by the user. By doing so the user will be able to see how well any of the selected schools performs with respect to the other schools.

2. **Issue**: There is a level of redundancy by adding a "Start" and "Stop" button for when a user decides to select a school on the map.

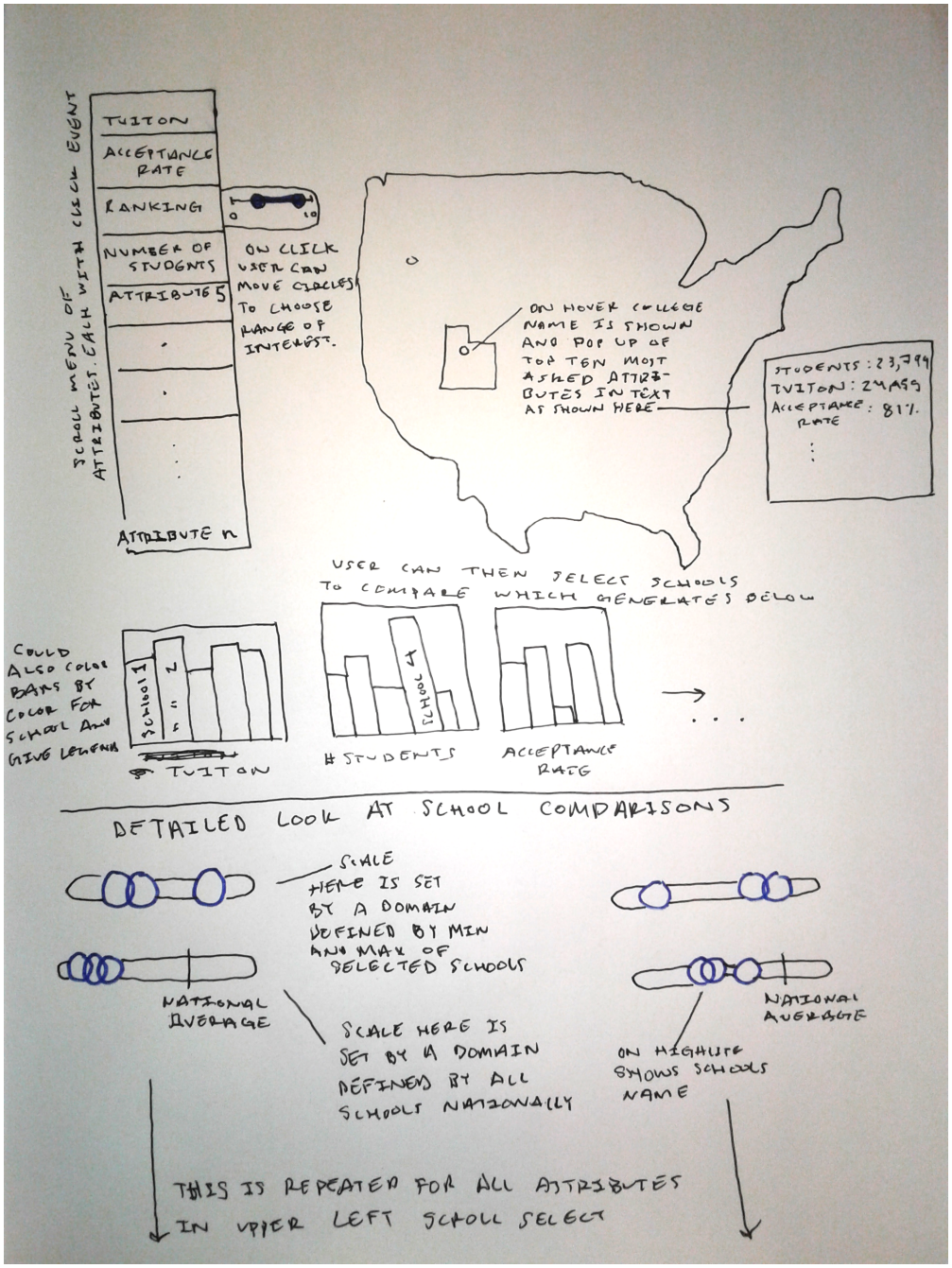
   Solution: It was suggested instead the may select a school on the map which triggers the event of having that schools histogram as well as more detailed attributes to be displayed. After selecting a number of schools the comparisons between schools will all be displayed lower. The User can then remove a school by again selecting the school on the map or selecting a small "x" icon by the schools name lower down.

3. **Issue**: After explaining the attribute comparison in which a range, defined by the max and min value of the selected schools attributes, and each of the selected schools displayed within that range as a circle it was pointed out to us that as more schools are selected it will be difficult to maintain an understanding of how a school falls into a national perspective, i.e. if the range changes for the selected schools picking top 5 worst schools might look similar to the top 5 best schools.

   Solution: We can decided that not only would it be more informative but clearer to add to each attribute comparison in the "detailed" section two ranges: the first defined by the domain of the schools selected and the second defined by all schools. In this way as a user selects schools they will be able to see how a school from their selection performs with respect to the other selected schools as well as how the school compares nationally. To better illuminate this on the national scale we will also put a marker along the range indicating national average.

3.1 Other changes decided on

We no longer will include the option to compare a preset number of schools. Our incentive for this is due to the ambiguity on which schools to select from the total range of each and between all attributes. We are also considering including a protip functionality on which where rather than only displaying the schools name when mousing over it's representative circle on the map it also displays in plain text the exact values of the top ten most asked questions about a school such as ranking, tuition, average acceptance rate, and so on. A sketch is given below on the direction we intend to take this project.



4 Data Processing

Unfortunately, we faced a lot of problems while mining the data since mining involved all the institutions in the US including community colleges and other small institutions. Though we were able to mine almost all of the data pertaining to all the institutions, some of the attributes like rank and tuition were hard to mine leaving no alternative option but to make them as "NA". Since colleges come in many varieties numerous attributes have no value which during runtime cause the visualization to fail. Though we had tried to merge various sources in order to achieve a more complete data set NaN values were almost inevitable. Once alternative would have been for us to

choose only attributes which are guaranteed to hold a value, such as tuition. We decided against this however since the attributes we chose are those most often search and what is more it serves somewhat informative to see which colleges may not be reasonably comparable.

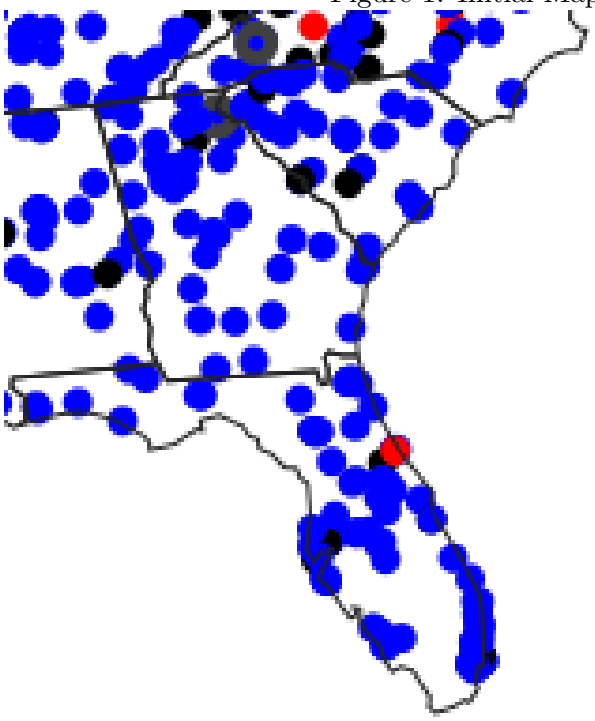
5 Meetings With the TA

During our first meeting it was made clear to us that using a scroll menu for the user is more cumbersome and less visual for data representation. As a result it was decided that each feature could be accompanied by a bar featuring brush selection. On our second meeting help was provided in terms of data binding for the range scale.

6 Obstacles, Improvements, Added Features, and Design Changes

As explained in the background section our initial goal was to both visualize the correlation to department and general school funding to academic progress measured by rate of publications as well as provide an approachable visualization based on school comparisons. We began by placing circle elements on a Json projection map. This proved fairly simple in that longitude and latitude of each college was provided in our data. The result of this was a broad view of the entire united states with points colored according to college ranking with red representing rank $<$ 10, black in the case $10 \leq$ rank < 20, and blue for rank $>$ 20. An example is shown below.

Figure 1: Initial Map Design

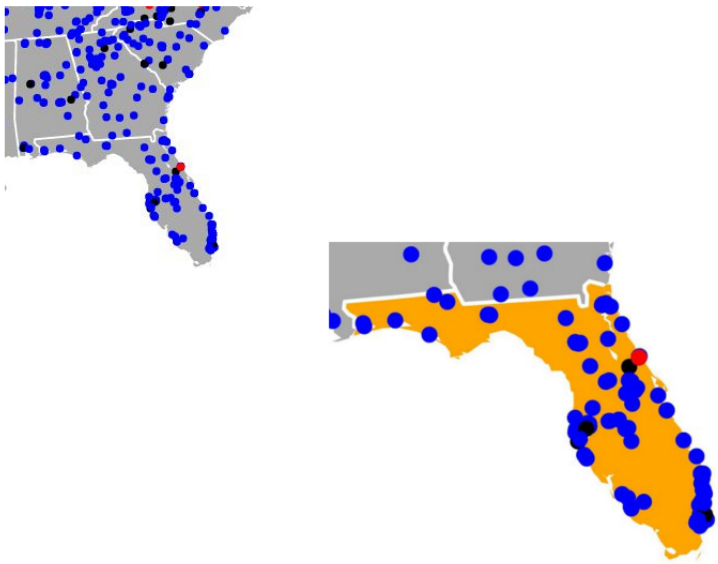


It soon became apparent due to the number colleges it would be difficult for users to differentiate near schools. As a result, and coupled by the suggestion of our TA, we felt it necessary to add a

zoom feature. The requirements of such a zoom feature would be to isolate the users view range to a near state sized window of the map as well as maintain the click functionality of each point in order to update the other visualizations.

While figuring out the zoom feature we decided on an implementation provided by Mike Bostock which can be found on his website. Unfortunately for some time it was not apparent that this approach would not work with our CSS styling elements. Luckily we found to remedy this some CSS styling must be placed within the .html file. The resulting map which is currently in our implementation is as such.

Figure 2: Zoom Functionality on Map



Being based on college attributes we initially intended to have a scroll selection to the left of the map the contents of which all attributes under consideration. On clicking each attribute we then would have a range selection tool in which the user could choose the range of values of their interest.

This implementation however would leave hidden the users ability to cater their comparisons to college's which are in similar valued ranges. Instead we chose a more upfront approach in which each attribute would be listed below the map select and would be accompanied by a brush selection. The user would then be able to see that to each attribute it is possible to select a range of interest and thereby winnow the college points of interest on the map and in turn, allow the user to only select such points for comparison with the histogram and range scale visualization. Below is an example of brushing rank in order to preserve only colleges within the brushed value region.

Figure 3: Map with no brush selection

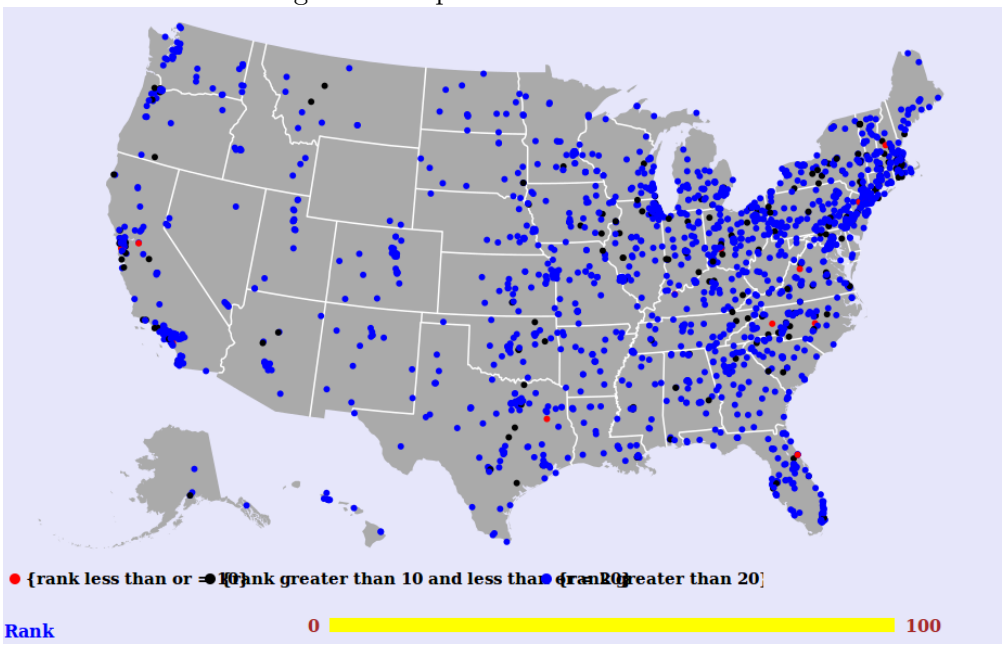
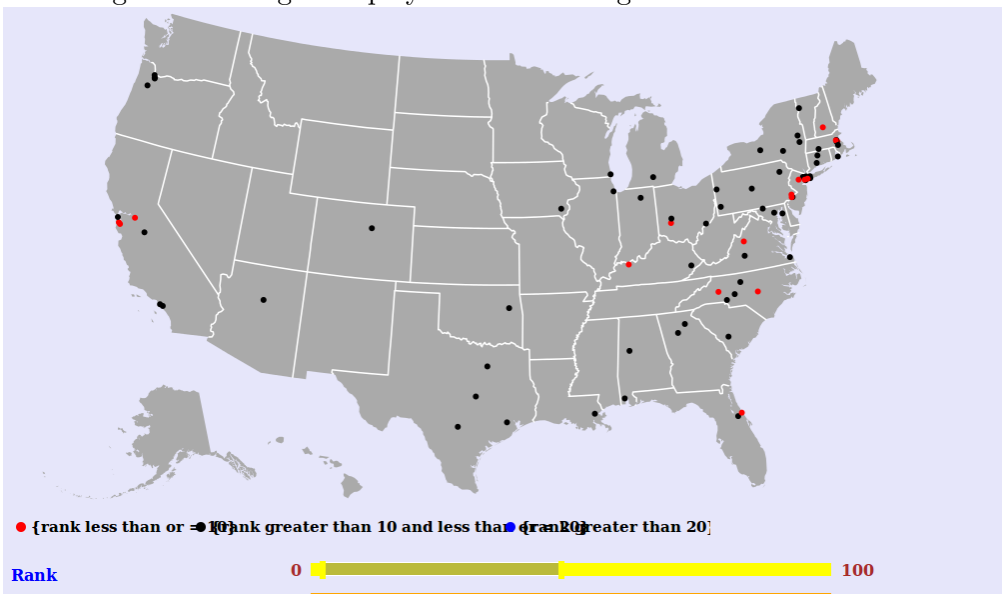


Figure 4: Colleges Displayed with Ranking Selected for Lower Tier



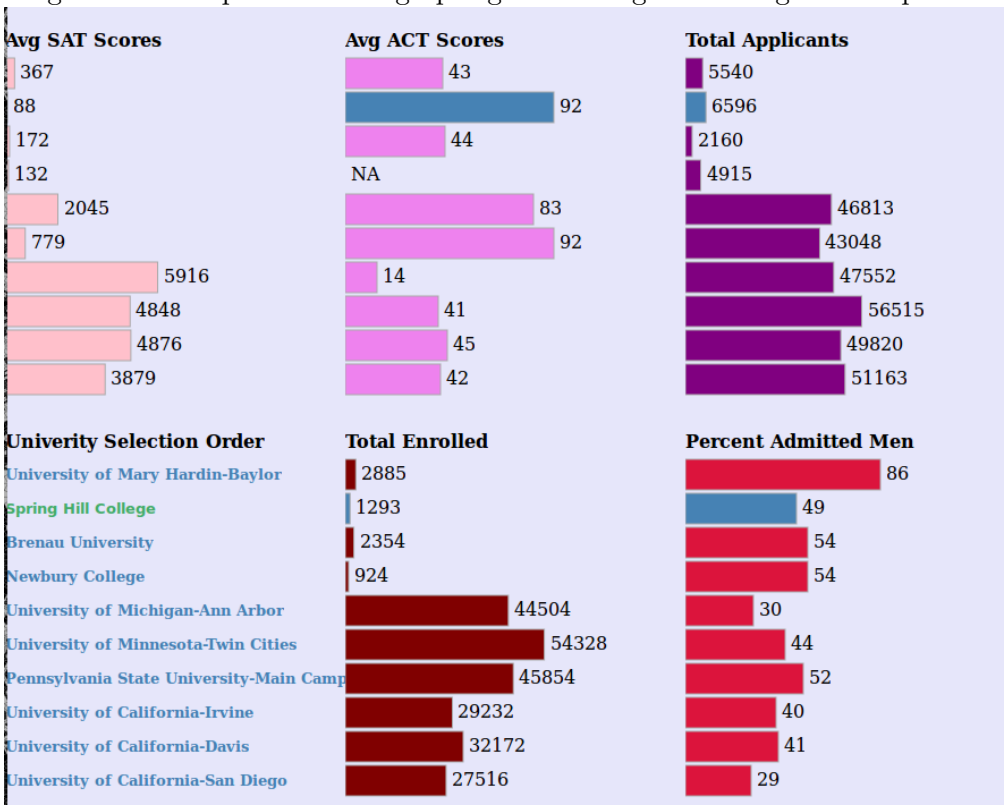
Following this we began the "Range Scale" and histogram visualizations. Our goal being to develop a comparison based visualization of core college attributes it was necessary to group histograms by attribute, such as rank, and not school. Initially it was our intention to allow the user to select as many schools as desired from the selection map. This however quickly became convoluted. As more colleges were selected it would become increasingly difficult to discern different schools.

Ultimately we came to the conclusion 10 schools would be sufficient and informative when comparing various school attributes. What is more, having a set number of bars would allow us to initialize the visualization with a pre-set of schools informing the viewer the availability a histogram

but more so, by interacting with the selection map the user will see that point selection on the map adds the corresponding school incrementally to the histogram comparison.

Following this implementation brought us to the need of labeling each histogram bar. For the histogram visualizations a single college spans numerous charts and for this reason it is difficult to keep track of how a single college compares to others over all attributes. Our first solution was to use a legend which numbers each school and in turn number each bar corresponding to that school. Our incentive towards numbering was due to the fact actual college names would not fit in every bar and between charts. This however was cluttered and did not fully resolve the issue. Ultimately we found it best to add an index of college names which updated as colleges were selected. The user would then be able to to click each college name and in turn highlight every bar which corresponds to that college across all histograms. In doing so the user is able to select a college and see how it compares to others which have been selected on the map.

Figure 5: Example of selecting Spring Hill College for histogram comparisons



We lastly hoped to provide a data driven means of college attribute comparison by visually representing a selected school's attribute values with respect to the span of schools which have been selected. Doing so the user can see schools which are outliers for certain attributes among those chosen. To accomplish this we chose to use a range scale where only colleges which have been selected would be used to define the domain which we are scaling using linear scaling. By projecting each schools attribute with respect to this domain the user would be able to spatially observe how colleges of their interest compare to the upper and lower bounds within colleges of their interest.

Our initial implementation was a simple bar onto which circles were placed based on the projection from either the nationally based domain or the selection subset domain.

Figure 6: Initial Range Scale

This however we realized could be optimized by also incorporating a gradient for the range bar as well as the color of the points based on a color scaling function who's domain is the same as that of the points. In doing this the user would have a gauge of the mean by where the transition occurs and in turn whether some schools are in the upper or lower percentile with respect to those chosen or nationally. Our first implementation iteratively projected and colored all colleges based on a scaling function determined by the domains previously mentioned. As a result the linear scaling of the color provided us with the transition we had hoped.

Unfortunately however having to iterate through all colleges for each attribute left our code excessively slow. We therefore used the gradient class to obtain the bars as seen below.

Figure 7: Linear Scaling of College Attributes Based on National Values

Figure 8: Linear Scaling of College Attributes Based on Selected College Values

As you can see in the two scales for the linear scaling based on selected colleges one college will always be lower bound while another upper bound whereas using national values for a domain allows points to be dispersed in any way. We had hoped to the stopping criteria for the transition between gradients provided in the linear gradient class in order to achieve our initial goal. we had intended to on each update sum all colleges, either nationally or those selected, over the attribute of interest and then average. Using this average and then normalizing would then allow us to force a mid-point equivalent to the average. Unfortunately this too was costly in time. Our final result uses the exact center of the bar as the median however we believe the transition coupled with the scaled color of the points still proves informative while comparing attributes.

6.1 Added Features

While implementing our design we realized certain functionality was needed to make our visualization more dynamic and accessible. One such feature was the ability to deselect an attribute after

having been selected by the brush selection. Originally a user was able to select multiple attributes in order to narrow down colleges of interest which fell into the value ranges desired. However if the user wished to no longer narrow the colleges displayed on the map for some attribute they would have to manually brush select the attributes entire range. Because of this we added the ability for the user to select again an attribute, after which that attributes range was rest to incorporate all values.

For the histogram visualizations a single college spans numerous charts. For this reason it is difficult to keep track of how one college compares to others over all attributes. Our first solution was to number each school and in turn number each bar corresponding to that school. Our incentive towards numbering was due to the fact actual college names would not fit in every bar and between charts. Our solution was to add an index of college names which updated as colleges were selected. It is then possible to click each college name and in turn highlight every bar which corresponds to that college. In doing so the user is able to select a college and see how it compares to others which have been selected on the map.

7 Final Product

We enjoyed and feel our project accurately utilized visualization techniques and practices taught during the semester. Some things we would have liked to implement would be better parsing and more broad mining of our data. We would have also like for the range scales to have a gradient better based on the data distribution under question. Ultimately however we feel our product serves as a useful tool for a comparative look into colleges. The interactive map demonstrates geographically the location of colleges one might be interested in. The color of these points catergorizes these schools by rank. On selection of these points an assortment of attribute histograms allows one to see how a selection relates to other schools in numerous ways and this comparison is better investigated in the ability to highlight schools within each histogram. Lastly the range scale allows one to select schools and view them relationally to one another as well as nationally.