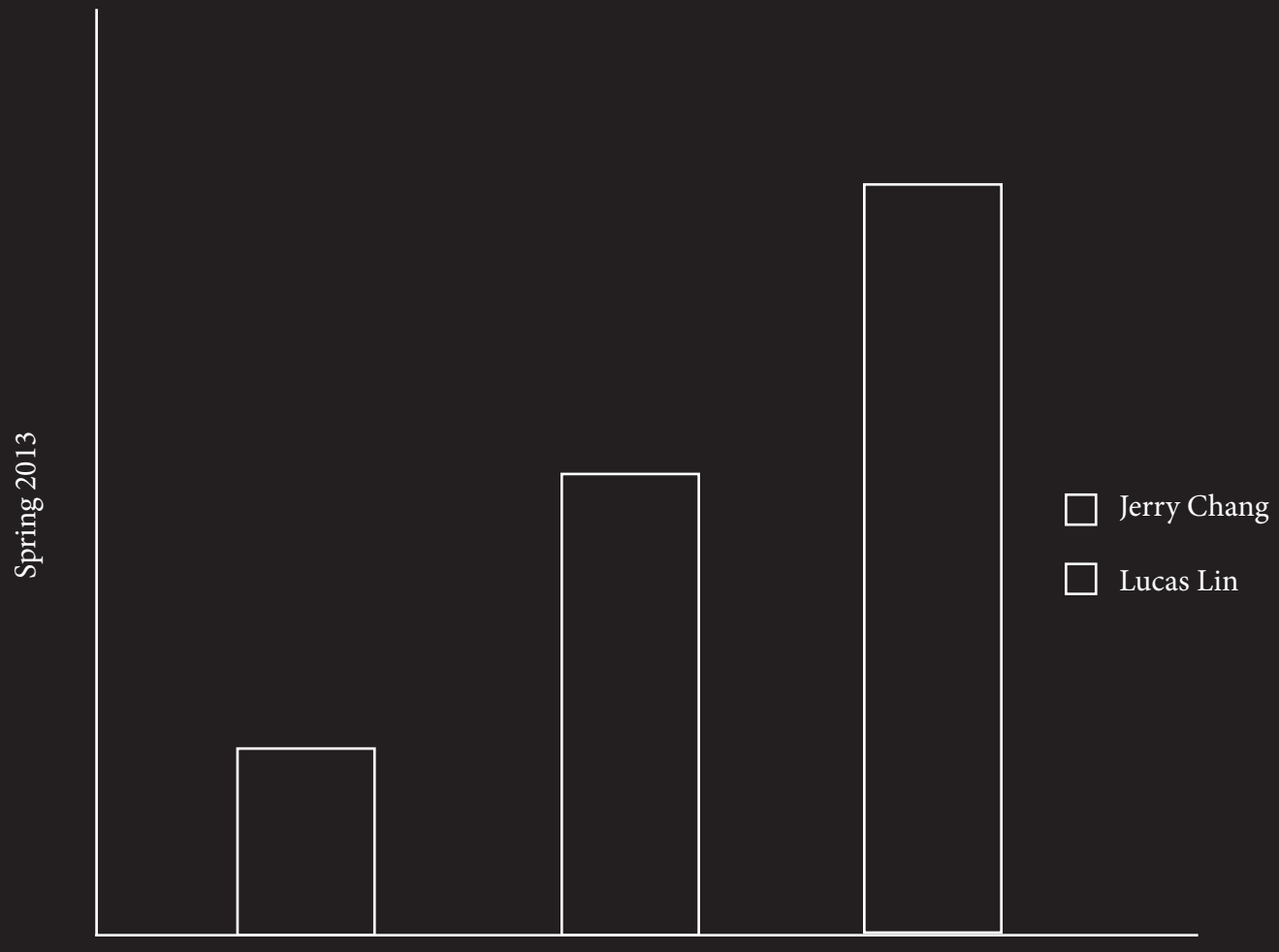
Concentrations at Harvard



CS171 Project 2

Project 2 Proposal

Project Title
Trends of Harvard Degrees Conferred

Team
Jerry Chang <jerrychang@college.harvard.edu>
Lucas Lin <lucaslin@college.harvard.edu>

Research Question
This fall Harvard offered new concentration options including Electrical Engineering, Mechanical Engineering, and an Architecture track causing the Crimson to question whether the college was catering to the pre-professional interests of the student body.

For this project, we want to analyze the degrees earned by Harvard undergraduates in recent years to understand how the pre-professional interests of students have changed over time. We will focus on comparing concentrations known for preparing students for high paying careers (e.g. Engineering, Econ, Pre-med) with concentrations known for their liberal arts approach (e.g. History, English, Languages).

Hypothesis
Based on our understanding of our peers, we predict that compared to past years more students are choosing to specialize in concentrations that are immediately relevant/useful for careers after graduation.

Motivation
As Harvard undergraduates and degree candidates, we are interested in comparing our declared concentration with that of recent alumni as well as current students. We want to identify those concentrations which are doing well or doing poorly and analyze whether these changes are due to external factors or to changes within specific departments at Harvard. A visualization will allow us to identify where trends exist and suggest possible causes for these trends.

Data Source
Our primary data source is a pdf document titled 'Degrees and Certificates Conferred: Academic Years 2007-2011' provided by The Office of the Provost at Harvard. This document lists the number of degrees conferred in each concentration between the years of 2007 and 2011. It further groups the concentrations into the Humanities, Natural Sciences, and Social Sciences.

The above document only provides 5 years of data to analyze. To better understand the trends in concentration choices we will also be scraping the Harvard College Facebook to project the number of degrees conferred for 2012-2015 (current sophomores to current seniors). This gives us almost a decade's worth of concentration data to analyze.

Scraping Process
To scrape the table off of the pdf, we plan to convert the document to html using an online tool such as PDFOnline and then getting the data using a python script with pattern.

Scraping concentration data from the Harvard Facebook will require a tool like mechanize which emulates a browser and allows us to authenticate through the Harvard PIN system. Once we get past the PIN system, we can then use pattern to scrape the number of declared concentrators for each concentration grouped by their year.

Potential Visualizations and Technical Process
We plan on using a few variations of bar graphs. Particularly, we are considering creating a triple bar graph with D3 (such as this one: http://bl.ocks.org/mbostock/3887051) in order to represent the categories of concentrations through the years. It is possible that we would make it interactive, so that clicking on a bar expands it into its subcomponent concentrations. We also plan on using a bar graph that goes into both the positive and negative values to show percent change over time (like this one http://bl.ocks.org/mbostock/2368837). A population pyramid (http://bl.ocks.org/mbostock/4062085) may also be useful in representing change over time as well as relative change between groups. We would also use the traditional line chart (http://bl.ocks.org/mbostock/3883245) in order to illustrate changes over time and facilitate the detection of trends between concentrations.

We would use D3 with Javascript in order to create these visualizations.

Revisions

Datasource

The problem with our initial proposal was the shortage of data provided by the Office of the Provost. Identifying trends is easier if we have data going back to 2000 as opposed to just 2007. We obtained this data by compiling information from several editions of the Harvard Student Handbook. Instead of visualizing degrees conferred, we would be visualizing the number of people declared in each concentration from 2000 to 2012.

Harvard College Degrees Conferred by Concentration
Academic Years 2007-2011

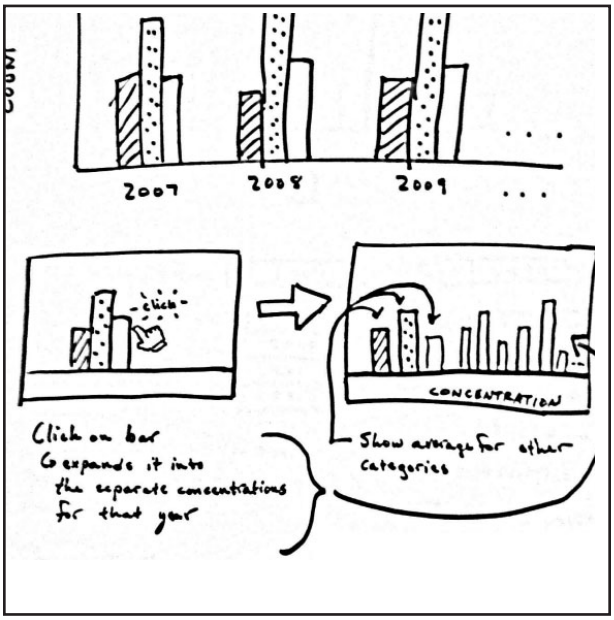
	Concentration	2007	2008	2009	2010	2011
Humanities	Classics	19	12	12	14	17
	Comparative Study of Religion	8	10	13	11	14
	East Asian Studies	7	14	26	18	18
	English and American Literature and Language	88	88	90	85	81
	Folklore and Mythology	2	3	4	11	4
	German	3	4	-	4	5
	History and Literature	52	44	54	53	58
	History of Art and Architecture	18	19	23	18	25
	Linguistics	9	13	11	9	13
	Literature	19	21	16	15	15
	Music	9	15	15	18	8
	Near Eastern Languages and Civilizations	14	14	6	6	7
	Philosophy	30	12	14	20	22
	Romance Languages and Literatures	19	22	15	9	26
	Sanskrit and Indian Studies	3	1	1	1	1
	Slavic Languages and Literature	5	6	3	1	7
	Visual and Environmental Studies	37	28	23	42	26

Proposed Dataset

Research Question

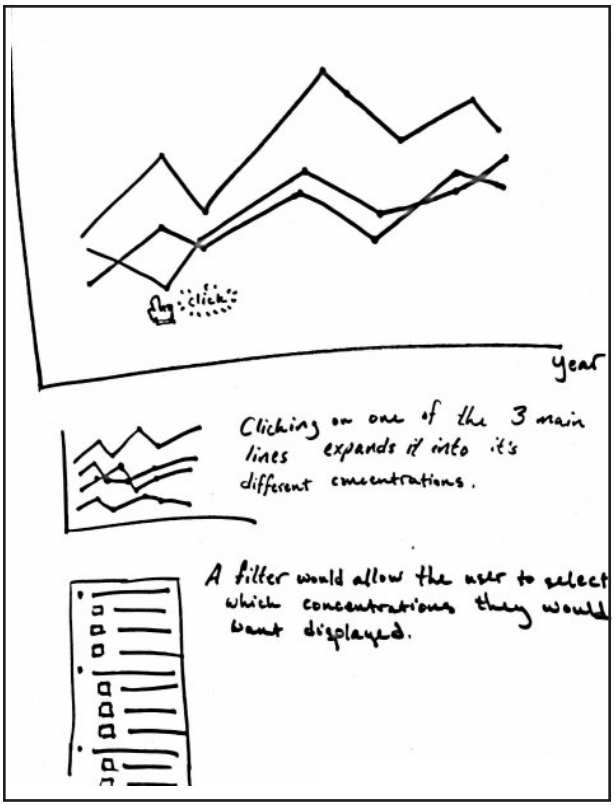
After exploring our dataset with Tableau, we failed to find any easy explanations or trends within our data. It seems that choices in concentration can not be reduced to earning potential nor changes within specific departments. We also tried correlating our data to changes in student-faculty ratio without any success. Consequently the question of why people choose a concentration can not easily be answered with just the data we have. Therefore, our research question needed to be adjusted to be larger in scope. Instead of attempting to find the cause of changes in concentration enrollment, we will use our data to simply compare different concentrations within categories and across time.

Visualization Sketches



Sketch 1

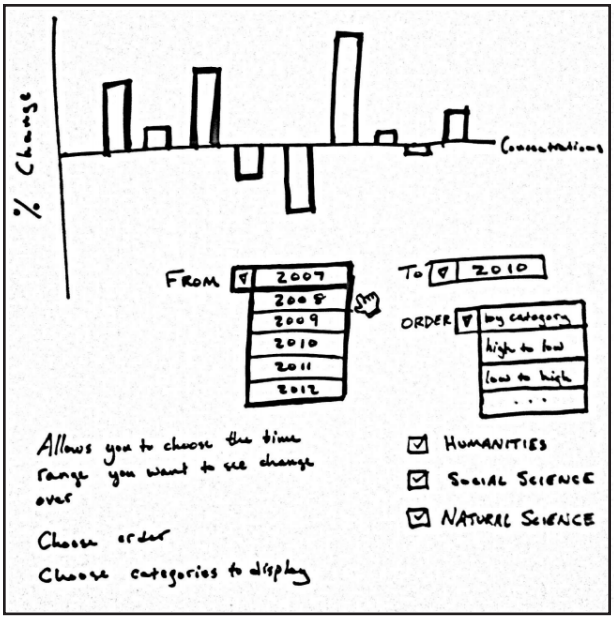
Visualization Prototype 1
This sketch shows a visualization that encodes essentially the same information as our final graph, but in a different way. This graph would consist of clusters (corresponding to years) of three bars - one for each category of concentration. Upon clicking on one of the bars, it would expand into its subcomponents and allow comparison to the different averages of the three categories.



Sketch 2

Visualization Prototype 2

Originally, we had planned on creating a line graph to show changes in concentrations over time. In our prototype, the graph would start with three lines corresponding to the three categories of concentrations over all the years. Upon clicking on one of the main lines, the line would expand into its different concentrations. A filter pannel would allow the user to more specifically select which concentrations to display.



Sketch 3

Visualization Prototype 3

We considered making a bar chart showing the change in concentrations from year to year. The y-axis goes into both the positive and negative to show growth and loss respectively. The user can select a time range to see change over as well as what categories to display.

Revised Dataset

The data for the number of people in each concentration was scraped from multiple editions of the Harvard Student Handbook. We scraped from the online editions of the 2012-13 handbook, the 2009-10 handbook, as well as manually entering data from the 2005-06 pdf edition of the handbook. For each of these handbooks, the number of people declared a concentration was printed in a table at the bottom of each concentration's page. The 2012-13 edition had data from 2007-11 while the 2009-10 edition gave us data from 2004-08. The pdf edition had data from 2000-04. To maximize the amount of data we had, we also included data scraped from Project 1 which contained the number of people currently declared in each concentration as listed on the Harvard College Facebook.

Putting these datasets together resulted in more than a decade's worth of data for the number of people declared in each concentration from 2000-12. Our final dataset is stored as a csv file created through a combination of web scraping with Python, data refinement with Google refine, and manual data entry/calculation with a spreadsheet editor.

We converted this file to JSON using an online csv converter and pasted the resulting json directly into our html page.

Tableau Previsualization

We were able to search for trends within our data by previsualizing it within Tableau. We experimented with different filters and display types including area graphs, line graphs, amd bar charts before deciding to implement bar charts for our interactive visualization.

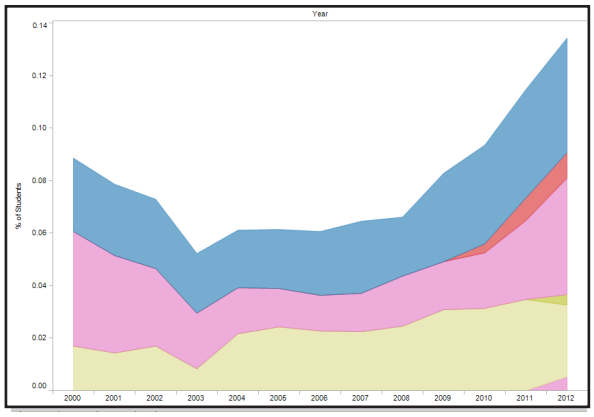


Tableau Area Graph

Tableau Barcharts

Tableau Line Graphs

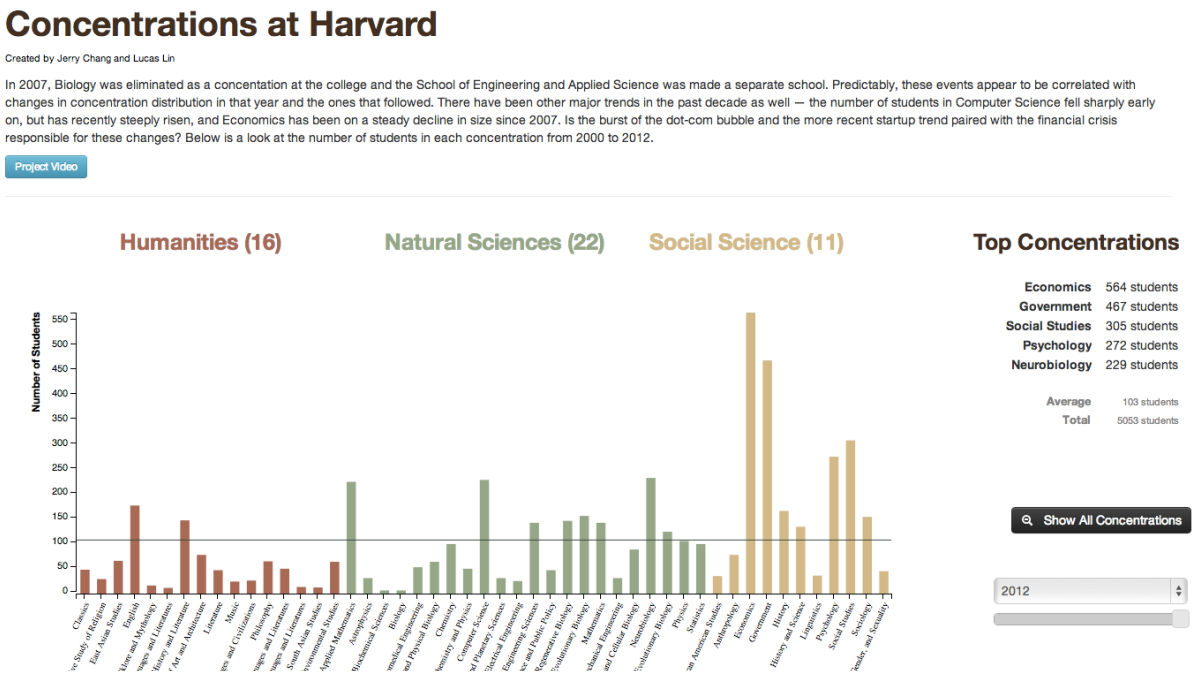
Visualization

We used D3 in order to create our dynamic interactive data visualizations. First, we created a simple bar graph. The first interactive aspect we introduced was the ability to select which concentrations to display via checkboxes. To make our visualization more interesting, we used D3 to dynamically add and remove bars by growing and shrinking them. We also dynamically resized our y-axis based on the size of the largest concentration currently shown in order to better show variations between the concentrations the user is currently interested in. Next we decided to put these checkboxes into a scrollable list on the right of our graph so that the user would not have to move their view off of the graph. We added a dropdown selection box to let users choose which year's data to view. We entertained the idea of using a modal window for our selection and filtering options, but in the end, we decided against it since it would obstruct the graph. The modal view was still useful however - we used it to display our project video explaining our visualization right there on the same page. This way, the users can watch the video and go back to the visualization without having to navigate away. Next, we added color coding by category and subsequently filtering by category to explore the three major groups of concentrations. We labeled all the bars with a serif-style font, since these make smaller text easier to read.

Ultimately, we chose to not use the entire list of concentrations with checkboxes for individual concentration filtering and instead opted for simplicity. Instead, we decided to display data about the graph. We created the "Top Concentrations" list to the right of the graph to show the top five concentrations currently being shown as well as the number of students in each. Below that is the average and total of the on-screen data. This section updates dynamically as the view changes, providing users with some context for the data being visualized.

We also added line that represents the average

of the current graph as an extra visual encoding and reference point. Naturally, we added a "Show All Concentrations" button to remove the category filtering effects. We chose to use a matte color palette with three distinct colors to allow good contrast while still being aesthetically pleasing. Despite some technical difficulties, we added a slider that dynamically updates the year displayed (both the graph and the selection box). We also added a text shadow to create an extra visual encoding of which category is currently being shown.

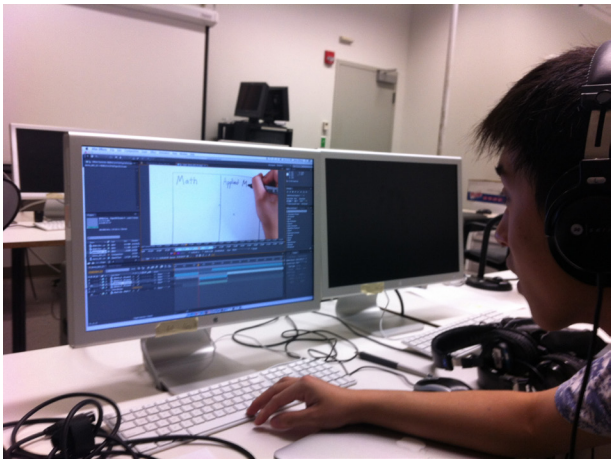


Final Web Visualization

Video

For our video we wanted to experiment with other visualization techniques in the form of animation. The first segment of our video explains the concept for our visualization using a technique similar to the RSA animate videos. We drew these visuals by hand and took pictures of the drawings frame by frame using a down shooter camera with DragonFrame. We processed these images in After Effects and combined it with audio using Avid. The second half of our video is more of a traditional screen capture where we demonstrate a few features of our visualization.

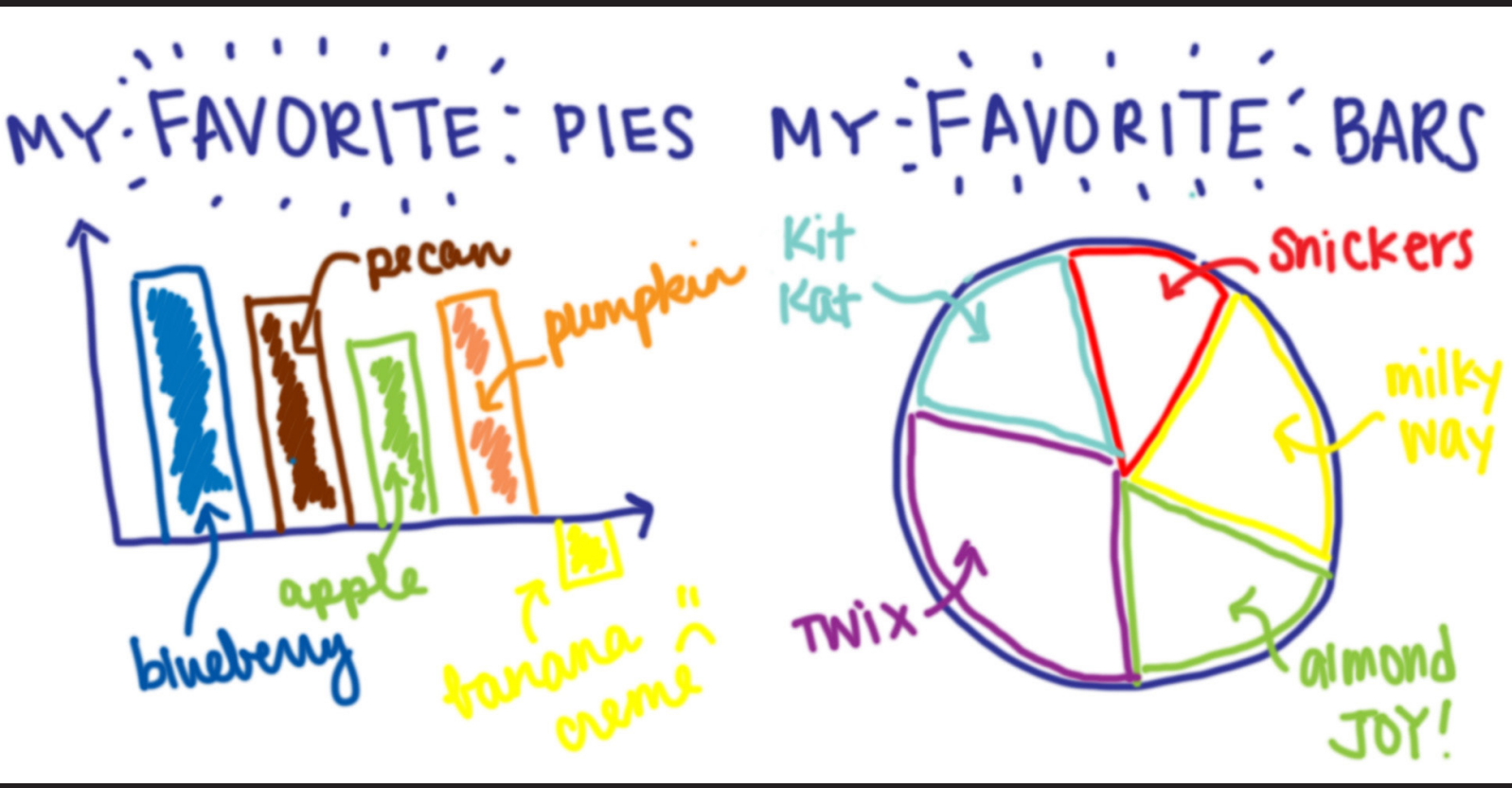
Lucas Recording Audio



Jerry Editing Videos

Findings

In 2007, Biology was eliminated as a concentation at the college and the School of Engineering and Applied Science was made a separate school. Predictably, these events appear to be correlated with changes in concentration distribution in that year and the ones that followed. There have been other major trends in the past decade as well — the number of students in Computer Science fell sharply early on, but has recently steeply risen, and Economics has been on a steady decline in size since 2007. We believe the burst of the dot-com bubble and the more recent startup trend paired with the financial crisis have played a large role in these changes. On a large scale, we noticed an inverse correlation between Natural Sciences and Social Sciences. For eleven out of the thirteen years, whenever the number of students in Natural Sciences went up, the number of students in Social Sciences went down, and vice versa (the other two years, there was a very small increase in both).


MY FAVORITE PIES
pecan
pumpkin
apple
blueberry
banana creme
MY FAVORITE BARS
Kit Kat
Snickers
milky way
TWIX
almond JOY!