Team Braavos - Predicting and Visualizing College Admissions

Overview and Motivation

Every year, four million students apply to US Colleges without having a good idea of their chances of getting in. Fueled by US News Rankings, colleges puff up their rejection rates, while myopically finding students through the narrow lens of standardized test scores. This project provides data-science based probabilities of getting in along with interactive visualizations, allowing students and parents to investigate how certain aspects of their application affect the chances of an acceptance to various schools and therefore allowing them to best focus their time and money. Users can also view summaries of application and acceptance data to make other inferences about their own potential success in applying to different colleges. We then aim to disrupt the entire application process so quality students can be matched to the right school and schools can fulfill their desired mix of students..

Contents

Team and Roles

Team Braavos consists of:

- Marina Adario
- Dion Hagan
- Malcolm Mason Rodriguez
- David Wihl

We anticipate being a fully agile team without pre-defined roles. Anyone can submit, check-in or work on any story.

The team maintains a Kanban board via Trello (https://trello.com/b/zguJ72GM). The "To Do" column is a regularly triaged and sorted list of the next task. Once an individual completes a task, he or she grabs the next item off the top of the list.

Communication Rules:

We collaborate regularly during the week via Slack. Email usage is minimal as necessary. Electonic signatures will suffice when signatures are required for submission.

Physical meetings occur once per week on Wednesday either during or just after Studio as necessary.

Collaboration Policy

The entire project is version controlled through github (https://github.com/wihl/cs171-project). Any non-trivial story will have a separate branch, with regular check-ins and merges. We are attempting to be fully Agile so any member can work on any story. Each story should be short, no longer than two days. Stories that are blocked will be indicated as such in Trello via a red bar label.

Feature List

The visualization will consist of at least two pages.

Page 1 - See Your Chances

This will be the home page of the site. There will be two areas for data entry:

- Area 1 - demographic data that cannot be changed, including gender, US citizenship, first to attend college, etc.
- Area 2 - college admission factors that may vary prior to college application such as GPA, standardized test scores, number of AP exams taken.

As the applicant changes values in area 2, the visualization will update. The intent is this will be highly interactive, almost game-like, encourgaging the applicant to attempt many different scenarios.

Page 2 - Data Drill Down

The second page will have several linked visualizations that allow the user to drill down into the specific factors that weigh into the college acceptance process. This allows the applicant to compare and contrast different schools as well as plan time for the most appropriate activities to maximum college acceptance.

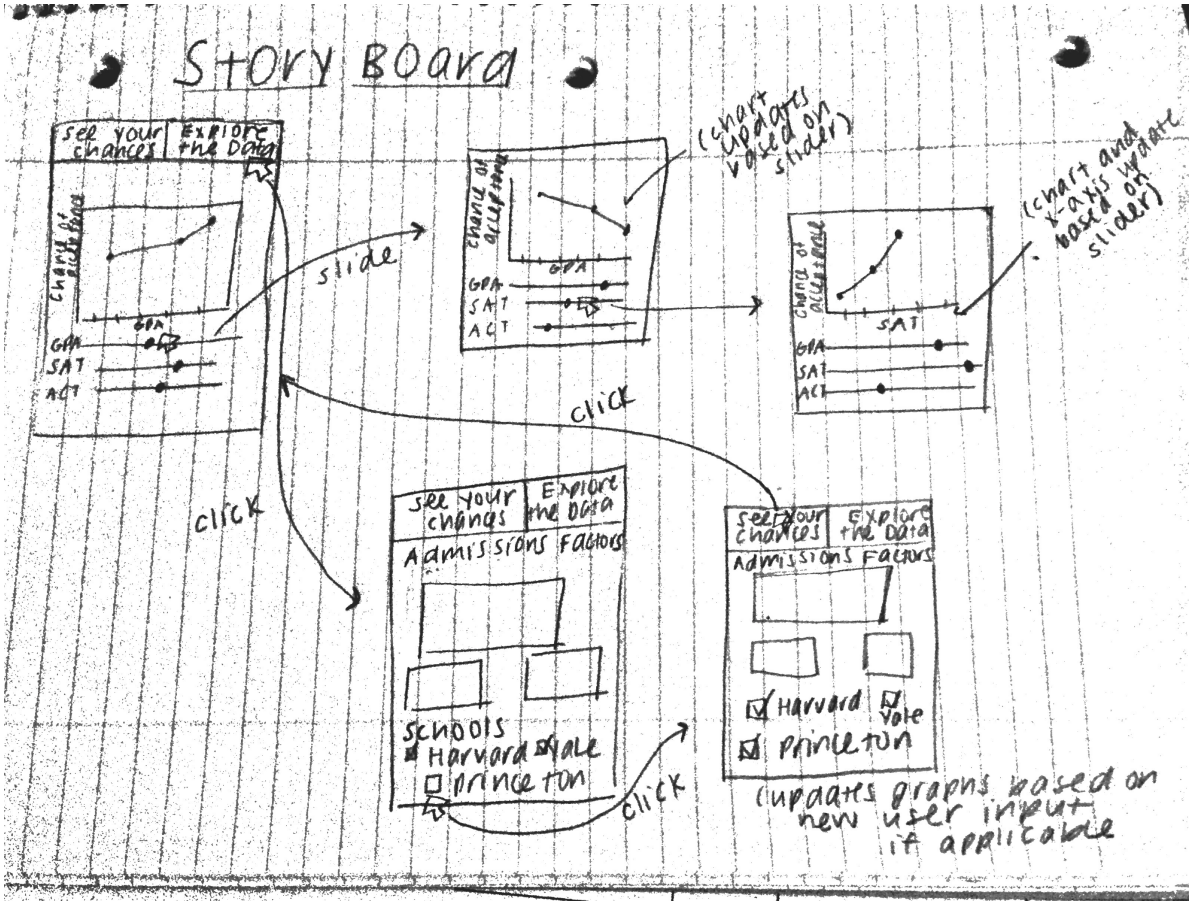
Summary of Features

- Demographic data entry
- Admission factors data entry
- School selection
- Acceptance Probability Visualization
- Drill down visualizations
- College comparison visualizations
- School selection for drill down visualizations

Project Storyboard

In [1]:
```
from IPython.display import Image
Image(filename='img/Final_project_story_board.png')
```

Out[1]:



Tasks and Timeline

(This is for a general idea only. The reference tasks and timeline are in Trello.)

Target:

- ✔ Choose domain (completed 3/28)
- ✔ Define question (completed 3/28)
- ✔ Explore existing solutions (completed 3/28)
- Formulate data analysis tasks (Due 4/11)

Data Wrangling:

- ✔ Find and clean data (completed 3/28)
- ✔ Exploratory Data Analysis (completed 3/28)
- ✔ Transform and summarize data (completed 3/28)

Design

- Design Visual Encoding
- ✔ Design Interaction - Prediction (completed 4/4)
- ✔ Design layout and storytelling - Prediction (completed 4/4)
- Design Interaction - Drill down (due 4/11)
- Perform 'paper' user testing (due 4/11)

Implement

- (in progress) Rapid prototype - Drill down (due 4/11)
- Rapid prototype - Prediction (due 4/18)
- (in progress) Define Data Structures (due 4/18)
- (in progress) Design system architecture (due 4/11)

Evaluate

- Perform user testing with prototype (4/20)
- Is the abstraction right? (due 4/20)
- Does encoding and interaction support the task? (due 4/21)
- Does encoding and interaction provide new insights? (due 4/21)

Deliverables

- Process Book (due 5/2)
- Screencast (due 5/2)
- Demos / design fair (due 5/4)

Time Permitting / Nice to Have

- Include distorted map of acceptance distance
- Performance optimizations:
  - Determine bottlenecks and explore efficient algorithms
  - Random Forest in JavaScript
  - cache the CSV (collegelist and the data) in localStorage
  - train the Random Forest asynchonously
  - populate list of colleges from CSV (or cache) instead of hard coded

In [ ]:

2related

March 28, 2016

1 Related Work

Anything that inspired you, such as a paper, a web site, visualizations we discussed in class, etc.

In [ ]:

3questions

March 28, 2016

1 Questions

What questions are you trying to answer? How did these questions evolve over the course of the project? What new questions did you consider in the course of your analysis?

In [ ]:

Data

Source, scraping method, cleanup, etc.

Snippet of Sample Data

In [2]:
```
import numpy as np
import pandas as pd

df = pd.read_csv("data/collegedata_normalized.csv", index_col=0)
df.head(10)
```

Out[2]:

	studentID	classrank	admissionstest	AP	averageAP	SATsubject	GPA
0	S50C3UECT8	NaN	0.838909	7	1.067927	0.323654	-0.1881(
1	GBWZQQRBEV	NaN	0.666993	7	0.661638	-0.440897	0.49306
2	MXXLWO1HQ2	NaN	0.208552	0	NaN	0.323654	0.39575
3	5KSL7C8SLZ	NaN	1.297350	7	0.864783	1.088204	0.10382
4	RQWLNGGZ49	NaN	0.323162	1	-0.354087	-0.440897	0.54171(
5	A7WDLWR2VM	NaN	0.323162	0	NaN	-0.440897	0.78498
6	CECT9K8GDY	NaN	0.724298	0	NaN	-0.440897	0.29844
7	9PM1B51CXG	NaN	-2.542098	3	-0.828092	-0.440897	-0.6746
8	G14LB2SV7O	NaN	-0.020669	7	-0.354087	-1.969998	0.49306
9	PSBRN09QGH	NaN	-0.135280	4	0.001416	-0.440897	0.78498

10 rows × 34 columns

In [ ]:
```
```

5eda

March 28, 2016

1 Exploratory Data Analysis

What visualizations did you use to initially look at your data? What insights did you gain? How did these insights inform your design?

In [ ]:

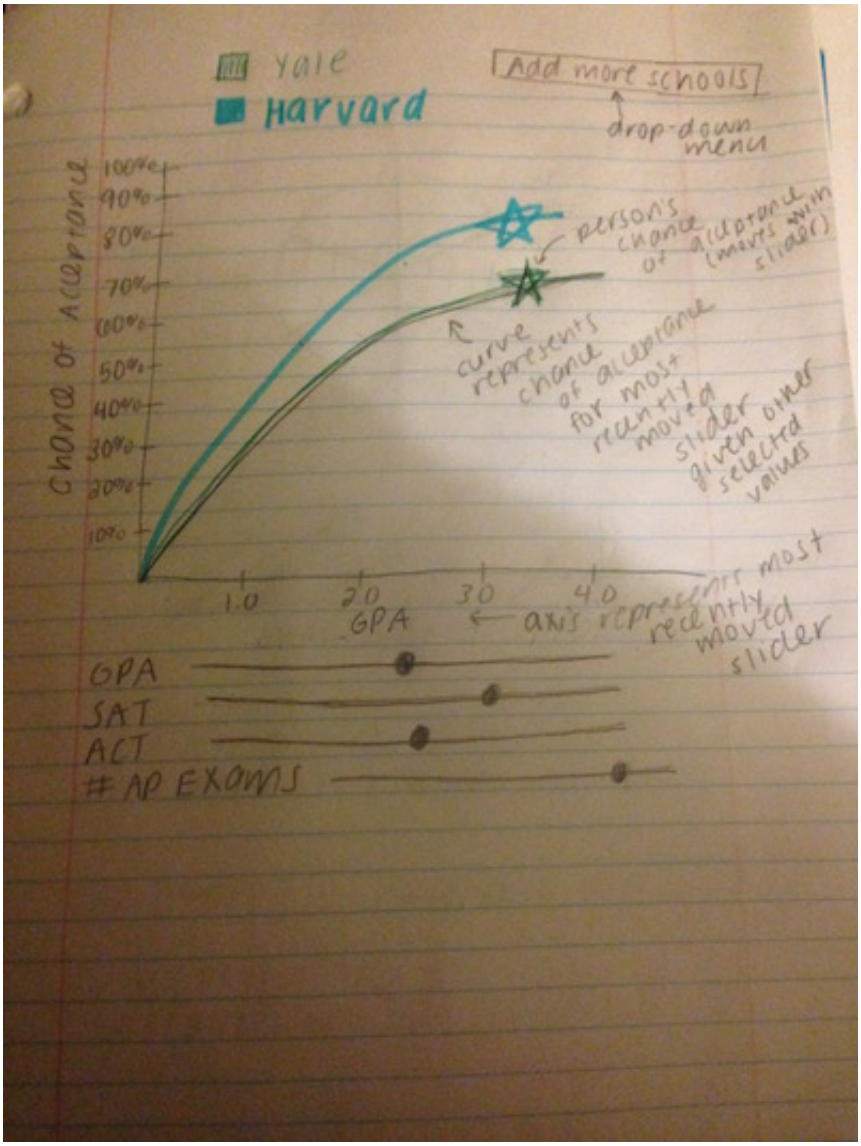
Design Evolution

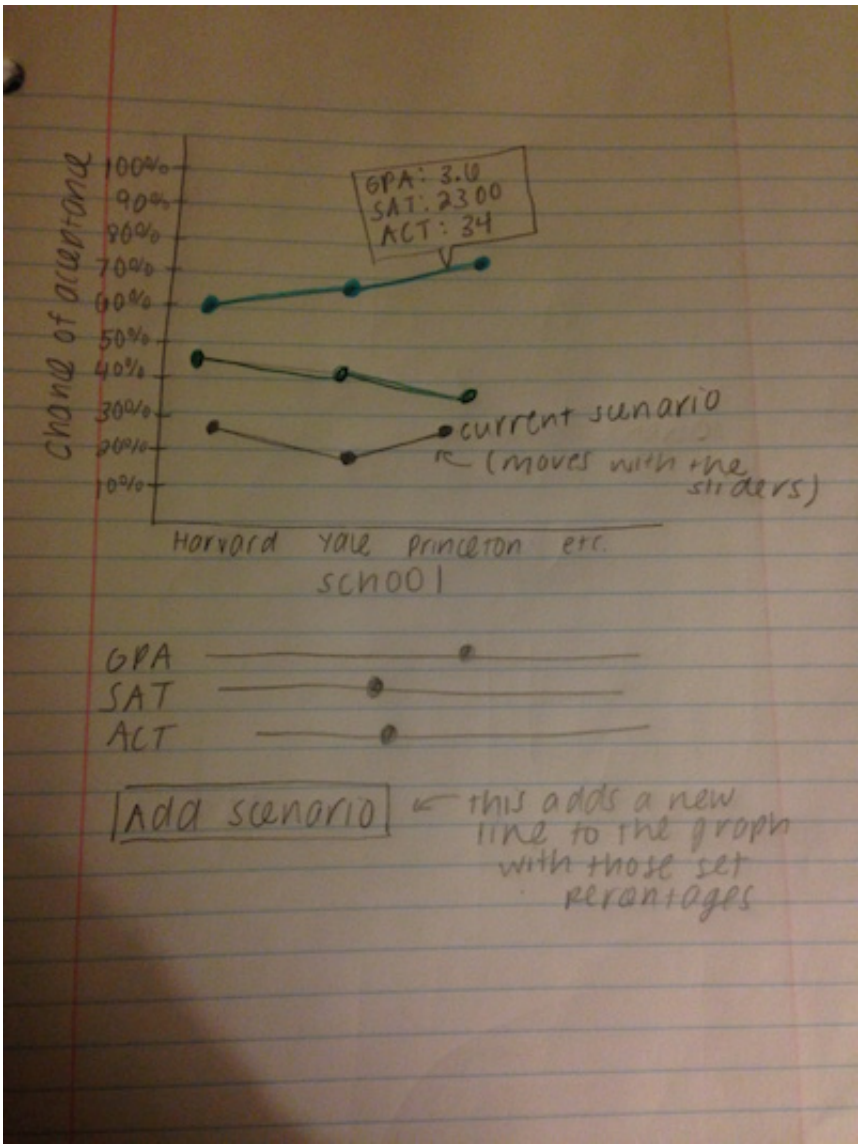
What are the different visualizations you considered? Justify the design decisions you made using the perceptual and design principles you learned in the course. Did you deviate from your proposal?

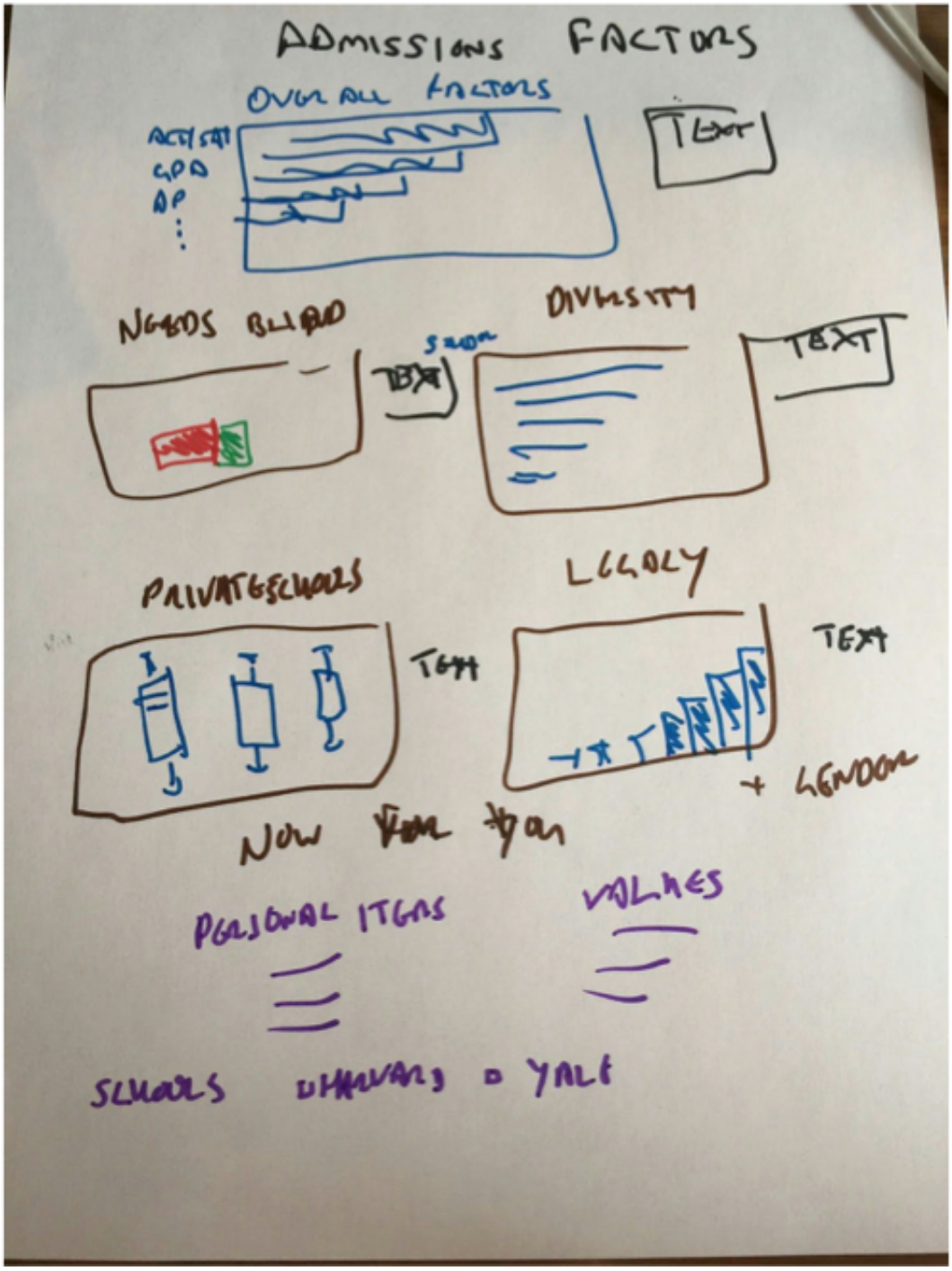
Contents

First Set of Sketches

3/28/2016



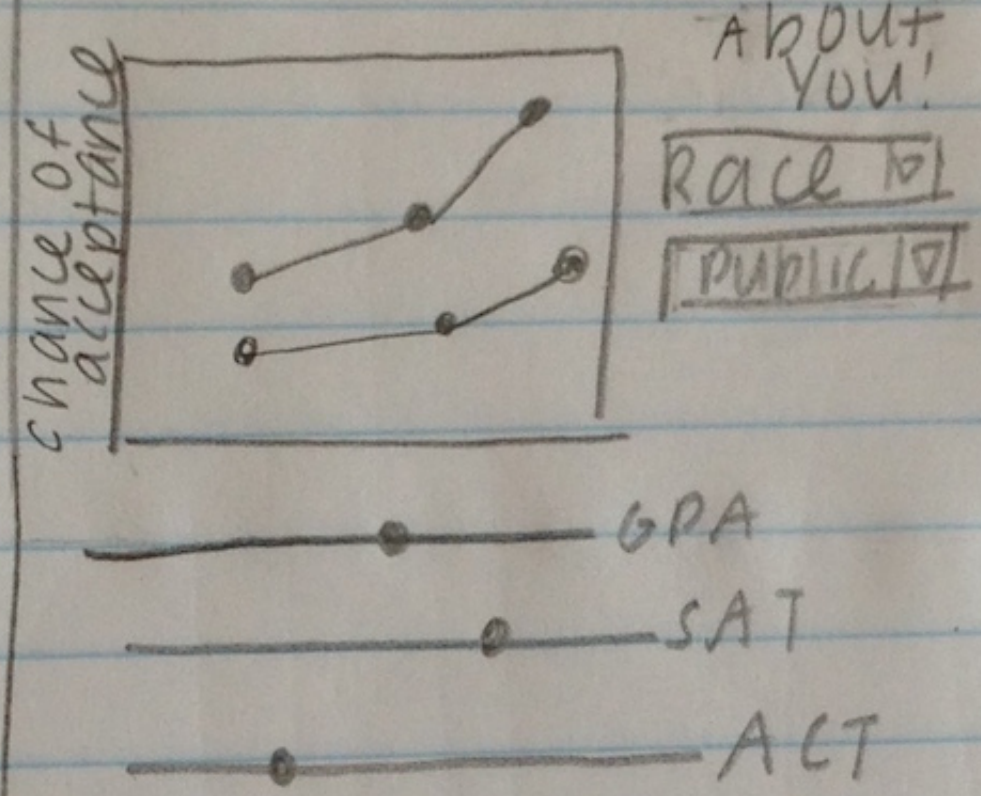

Chance of acceptance
100%
90%
80%
70%
60%
50%
40%
30%
20%
10%
GPA: 3.6
SAT: 2300
ACT: 34
current scenario
(moves with the sliders)
Harvard Yale Princeton etc.
school
GPA
SAT
ACT
Add scenario
this adds a new line to the graph with those set percentages


ADMISSIONS FACTORS
OVERALL FACTORS
TEXT
NEEDS BLIND
DIVERSITY
TEXT
PRIVATE SCHOOLS
LEGACY
TEXT
+ LEGACY
NOW FOR FUN
PERSONAL ITEMS
VALUES
SCHOOLS DIFFERENT = VALUE

Prediction Page Layout

See your chances	Explore the Data

Tell us about yourself to see your chances of acceptance into one or many of these top US colleges:

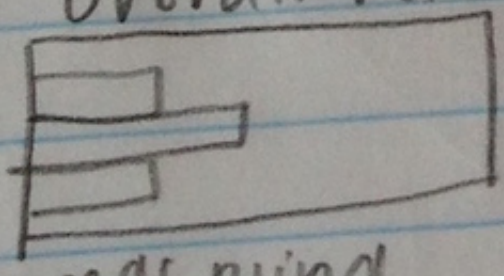


see your chances

Explore the data

Learn more about which factors are most important to each college:

Overall Factors

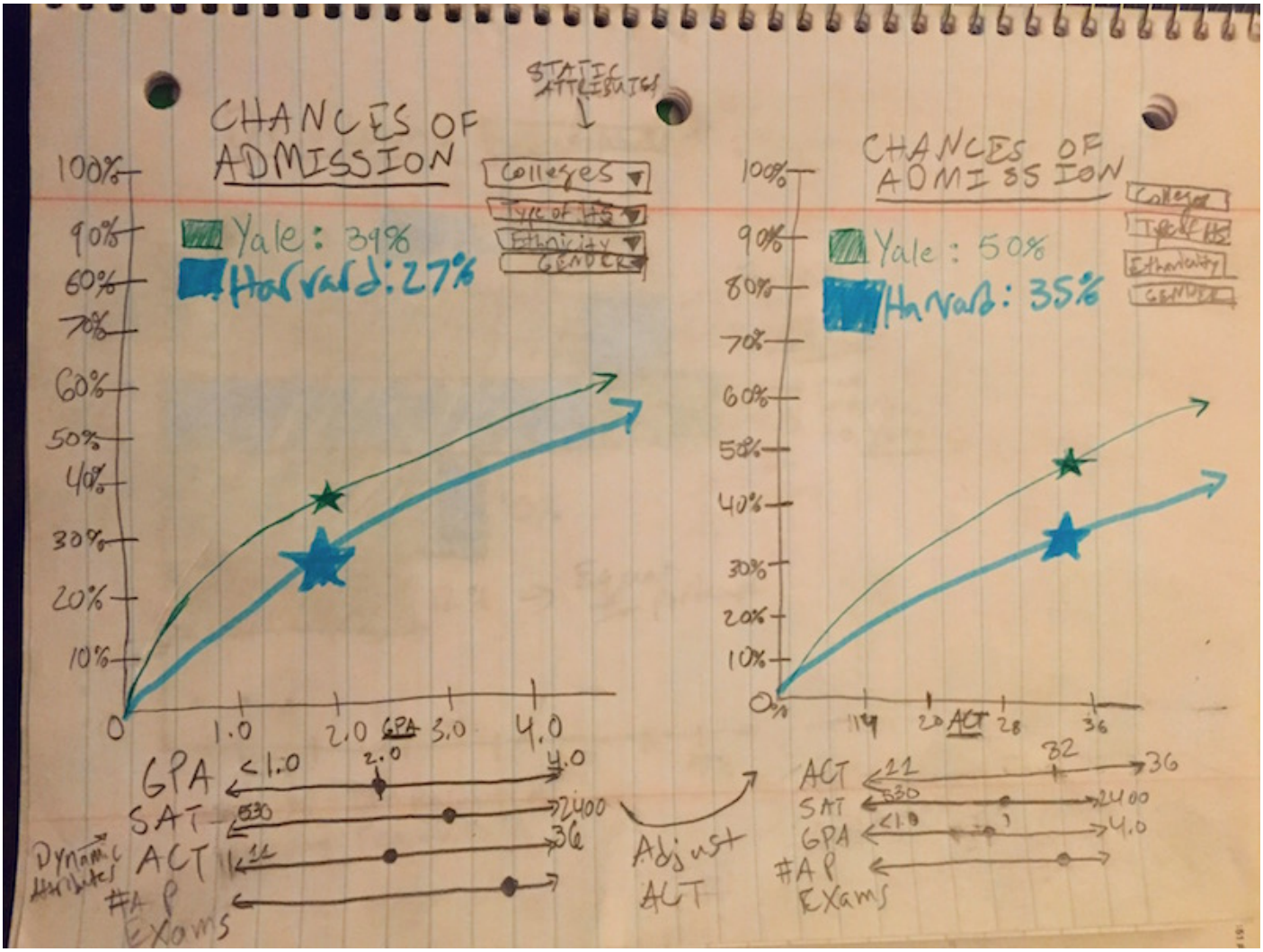


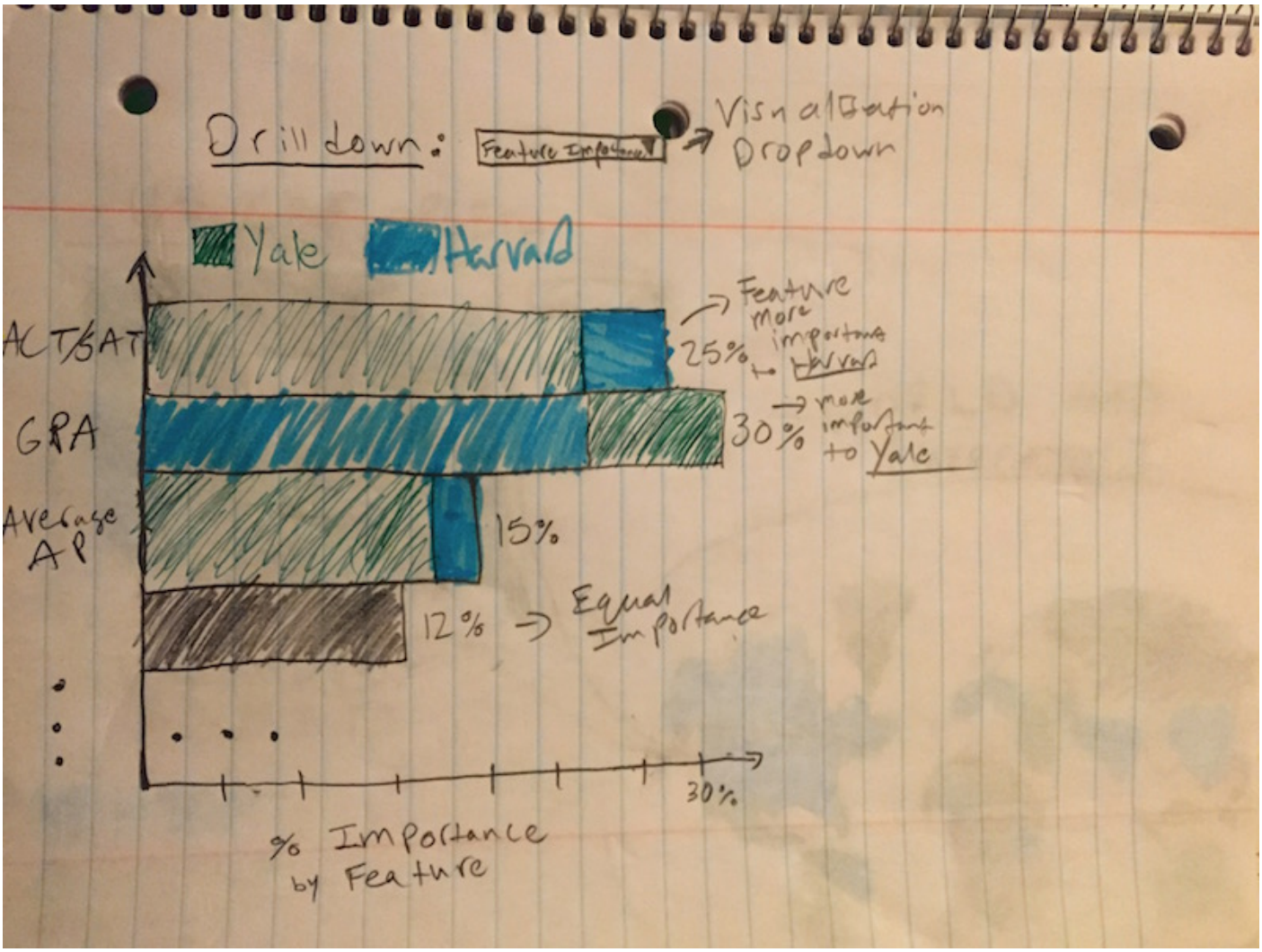
text

Diversity

Second Set of Sketches

April 4, 2016




Drill down:
Feature Importance
Visualization Dropdown
Yale
Harvard
ACT/SAT
GPA
Average AP
Feature more important to Harvard
25%
30%
more important to Yale
15%
12% → Equal Importance
30%
% Importance by Feature


US CHOROPLETH
College
WORLD MAP
[MERCATOR]
HARVARD
ADMITS
HIGH
MEDIUM
LOW
None

In [ ]:

7implementation

March 28, 2016

1 Implementation

Describe the intent and functionality of the interactive visualizations you implemented. Provide clear and well-referenced images showing the key design and interaction elements.

In [ ]:

8evaluation

March 28, 2016

1 Evaluation

What did you learn about the data by using your visualizations? How did you answer your questions? How well does your visualization work, and how could you further improve it?

In [ ]: