The Datamodel

Based on our research of key factors (example (http://www.nacacnet.org/studentinfo/articles/Pages/Factors-in-the-Admission-Decision.aspx)), the first thing we did was determine an appropriate datamodel in order to standardize scraping from potentially multiple sources and allow parallel development between scraping and classification. There are of course limits to pre-determining the model. We are unable to get a complete picture of a given candidate, such as recommendation letters as it is impossible to get data for this as well as being difficult quantify. We were able to find a data source for the vast majority of the factors we identified and did not discover any additional sources for factors we did not already identify. The breadth of our factors is already substantially wider than the two dimensions of Test Scores and GPA found on Naviance (http://www.naviance.com/), the most popular current site for predicting admissions.

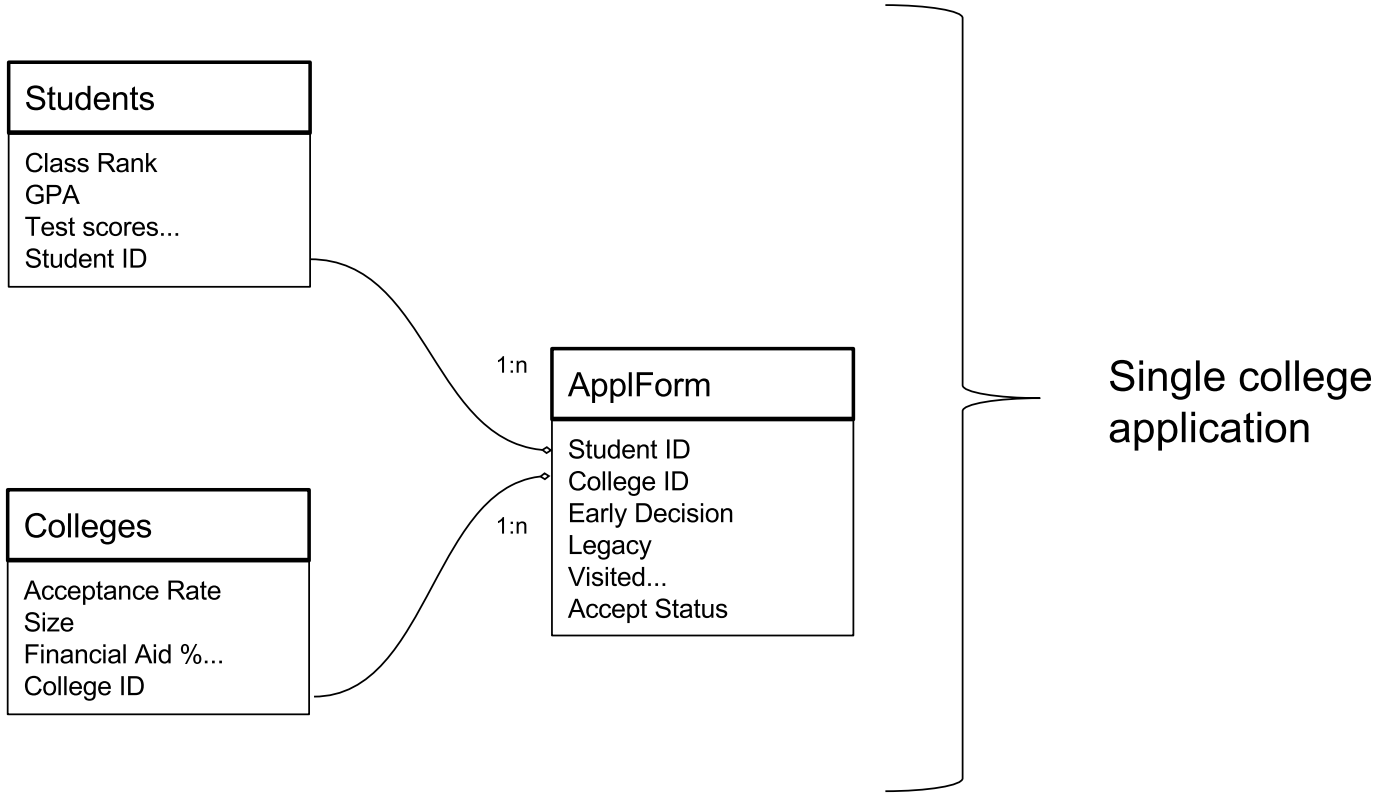
We distinguish three tables:

- A students table contains all academic and personal data of a particular student (scores, gender, etc)
- A college table contains all information of a university (acceptance rate, public/private, etc)
- An application form table contains application-specific data for a particular student in a particular university, for example and most importantly, the result of the decision procedure

When these three tables are merged, a single row of the merged result represents a complete college application ready for classification and analysis.

In [10]:
```
from IPython.display import SVG
SVG(filename='datamodel.svg')
```

Out[10]:



The Student Table

#	Factor	Data Type	Category	Support	Standardized?	Pandas Name	Comments
0	Class Rank	Numeric	academic	R	Y	classrank	percentile, or could group by "top 10%" . For ...
1	Admissions Test	Numeric	academic	R	Y	admissionstest	This is a combination of SAT and ACT, converte...
2	# AP/IB/Other	Numeric	academic	R	Y	AP	# tests taken
3	Average AP/IB score	Numeric	academic	R	Y	averageAP	NaN
4	SAT Subject	Numeric	academic	R	Y	SATsubject	# tests taken
5	Can Afford Tuition	Indicator	personal	{-1,0,1}	N	canAfford	NaN
6	GPA	Numeric	academic	R	Y	GPA	NaN
7	Weighted GPA	Numerica	academic	R	Y	GPA_w	NaN
8	Female	Indicator	personal	{-1,0,1}	N	female	-1 means male, 0 means unidentified
9	Minority Gender/Sexuality	Indicator	personal	{-1,0,1}	N	MinorityGender	NaN
10	Minority Race	Indicator	personal	{-1,0,1}	N	MinorityRace	NaN
11	International	Indicator	personal	{-1,0,1}	N	international	NaN
12	First in Family	Indicator	personal	{-1,0,1}	N	firstinfamily	to go to ANY college
13	Sporting Excellence	Indicator	non academic	{-1,0,1}	N	sports	NaN
14	Music / Performing Arts Excellence	Indicator	non academic	{-1,0,1}	N	artist	NaN

15	Which Program	Factor	academic	n levels	N	program	e.g. Sciences / Arts / Music / Undecided
16	Work Experience	Indicator	non academic	{-1,0,1}	N	workexp	NaN
17	Public/Private/Homeschool	Factor	academic	{-1,0,1}	N	schooltype	Highschool type: -1 is public, all values are ...
18	Intended Grad Year	Numeric	personal	I	N	intendedgradyear	Class of ?

The Colleges Table

#	Factor	Data Type	Category	Support	Standardized?	Pandas Name	Comments
0	collegeID	string	college	string	N	collegeID	Unique key in the database
1	name	string	college	string	N	name	Friendly name
2	College Acceptance Rate	Numeric	college	R	Y	acceptrate	NaN
3	College Size	Numeric	college	R	Y	size	NaN
4	Public/private	Indicator	college	{-1,0,1}	N	public	-1 private / parochial, 1 is public

The Application Form Table

#	Factor	Data Type	Category	Support	Standardized?	Pandas Name	Comments
0	Campus Visit / Interview	Indicator	application	{-1,0,1}	N	visited	NaN
1	Early	Indicator	application	{-1,0,1}	N	earlyAppl	Includes early action and early decision.
2	Acceptance Status	Indicator	application	{-1,1}	N	acceptStatus	did the student get in? -1 = rejected, 1= acce...
3	Acceptance Probability	Numeric	application	R	N	acceptProb	what is our forecast acceptance probability?
4	Out of State	Indicator	application	{-1,0,1}	N	outofstate	NaN
5	Family Alumni	Indicator	application	{-1,0,1}	N	alumni	NaN

Datamodel Implementation