Goal: To provide an intuitive web site merging three main data sets together that allows potential applicants to find the right Computer Science doctoral program for them.

Citation Network Dataset
Extracted from DBLP, ACM and other citation sources, contains more than 4,354,534 citations. We reduced this down by faculty and transformed paper abstracts into a TF/IDF format.

Faculty Professor Dataset
Generated as part of a class assignment contains 2,195 faculty from 55 different universities. We enhanced this by adding a generated faculty ID and existing Institution ID.

Integrated Postsecondary Education Data System (IPEDS)
Every US Institution is required to submit this data every year to the government. Highly accurate but required filtering and interpretation. We used Python scripts to reduce the data down to institution ID's and award levels we are interested in.

TF/IDF Data
For every unigram and bigram in every abstract for our faculty we store the term, the faculty ID and the frequency, sorted from most frequent to least.

Final Dataset
Faculty Name, Institution, Sub-Field, Join Year, PhD earned Inst, etc...

All Tables by Institution ID
General Institution fields
Current Enrollment by student level, gender and race
Computer Science doctoral completions, for multiple years

Visualization and Interactive Search
Main Page
Search Results
Faculty Detail

2) Analysis of Over 2,000 Computer Science Professors at Top Universities, http://jeffhuang.com/computer_science_professors.html