Team Zebra Final Project Proposal

- Team Members: Naman Arora, Somya Agarwal, Ruhi Patel, Nate Fairbank
- Dataset: We selected the results from the 2018 Stack Overflow Developer Survey. This questionnaire had over 100,000 respondents who reported information ranging from their exercise habits to compensation levels. The strengths of this data are that it broad and fairly standardized, and that some rudimentary cleaning of the data was done for the public release. The downsides are that results of a mass-solicited, self-reported survey may not be reliable, and may suffer from massive selection bias in terms of who saw and was willing to fill out the survey. Additionally, 100,000 entries may not be large enough to permit any complicated machine learning, and the data may become unreliably sparse once filtered for specific feature values.
- Data URL: https://www.kaggle.com/stackoverflow/stack-overflow-2018-developer-survey?select=survey_results_public.csv
- The question: “What makes a programmer successful?" This question is intentionally broad, to allow the user to project their own perception and line of inquiry onto it, prompting them to explore the data. For example, the pragmatic (and financially motived) reader might be interested in the answer to the sub-question “what computer language should I learn to make the most money?" Another, more socially-oriented user might wonder “do white men stay in the field of computer science longer than minorities or women?" A third viewer might be curious about what undergraduate majors generate the highest career satisfaction. All three users have, in a way, asked “what makes a programmer successful”, with different means of generating success, and different definitions of success itself.
- Further comments: It seems that a user-driver “Interactive Visualization/ Application Track” is best-suited to addressing this question. Our application would briefly explain the data and ask the question, perhaps with some initial suggestions for exploration, drawing the user in and allowing them to explore the data along their own path. However, if we find a series of compelling relationships a narrative track might allow us to better direct the user towards these key insights.
- Next steps:
  - Sketching: each of us will make 3x initial sketches of questions we would like to try to answer. We should have the data schema in front of us while sketching, but not the actual data. We can then compare the 12 resulting sketches as a group, pick the best ones, refine them.
  - Data exploration: we will then use Tableau to do initial exploration of the features we are interested in, checking for interesting relationships between features, feature values and summary statistics, and compelling visualizations.
  - Sketching round 2: this round of sketching will focus on tying together the areas of exploration that we feel are most promising based on the data exploration. How will the user navigate through the data? What interactivity will occur between graphs?
  - This takes us through the initial design phase of the project. Once we have sketched individual visualizations, explored the data, and sketched a way to link together the individual visualizations we will be better able to plan the actual development of the product.