Expanding Potential Seed Project Proposal: The Diversity Data Visualization Hackathon
UC Berkeley Energy and Resources Group (ERG) Student Diversity Committee

Overview: The ERG Student Diversity Committee is pleased to submit this proposal on training underrepresented undergraduates to visualize data on STEM diversity statistics at UC Berkeley. We aim to build the programming skills of undergraduates who participate in this project, and promote campus-wide awareness of diversity statistics in STEM fields by creating web-based interactive data visualization products and tools. We will pair graduate student mentors (both within and outside of ERG) and undergraduates in a 'hackathon' working group that will meet every two weeks over the course of a year. We will spend one month at the outset of the project recruiting and training graduate student mentors, and then seek out underrepresented undergraduates to join the program by emailing departmental listservs and student groups. The working group will be highly collaborative, with hands-on assignments and peer learning in each session.

Rationale: There is a critical need for better understanding of diversity statistics in STEM fields in order to identify and design policies to promote underrepresented groups in STEM. At UC Berkeley in particular, the effects of Proposition 209 on STEM education have not been fully explored. By targeting analysis at the department level, we intend for the results of the analysis to be useful for informing department actions regarding admission, retention, and graduation of students of different socio-economic backgrounds. We plan to present our data visualizations in a way that promotes an intersectional and thorough understanding of diversity on campus. At the same time, we aim to increase the data fluency of Berkeley undergraduates from underrepresented backgrounds. These skills are crucial for today's job market, and we believe that early peer-to-peer learning in a supportive, diverse environment builds skills and confidence. By analyzing data on diversity in our own campus community, we can also help undergraduates become skilled and informed advocates for diversity in STEM, and provide useful data to decision-makers on campus. We believe it will be impactful to put the results of these efforts in some public form (e.g., UC Berkeley's Initiative for Equity, Diversity and Inclusion website, or allowing individual departments to host it on their own webpages). We propose the following curriculum:

- Unit 1: Introduction to principles of data analysis – We will discuss why data analysis is an important skill; what constitutes data; and why reliable data is necessary to understand diversity and representation in STEM higher education.

We will begin by working with data that is already available, for example at the Cal Answers website or the Berkeley Equity, Inclusion, and Diversity website, and we will brainstorm ideas about what other questions we could answer with this data.

- Unit 2: Data analysis tools – We will introduce and demonstrate basic software and analytical tools for this project, including key programming languages and packages, for instance Github, R, D3, etc.
- Unit 3: Data gathering and scraping – We will discuss various methods of gathering or scraping data. Students will work together to find data sources on diversity in STEM fields at Berkeley, engaging with the Initiative for Equity, Inclusion, and Diversity and individual STEM departments. ERG, for example, has compiled student demographic statistics. Students will also practice scraping data from websites. We will also discuss generation of qualitative data sets, such as a survey to determine which departments track demographics in their programs.
- Unit 4: Data cleaning – We will use programming software to clean the collected data and learn best practices for data cleaning.
- Unit 5: Visualization – We will explore ways to visualize our data and present it in meaningful ways.
- Unit 6: Presentation – We will develop some metrics to quantify our findings, and determine a format to present our findings to the campus community. We will also seek out potential partners for sharing our results, such as the recent diversity data gathering initiative in the environmental sector, which has gathered data from over 800 organizations.
- Unit 7: Replicability – We will discuss how to generate clean, replicable code and processes, **with the goal of passing our exploration, in a replicable package, to groups at other campuses or organizations**.

Qualifications: Members of the ERG Student Diversity Committee have all taken courses in data analysis and visualization, and we will recruit other graduate mentors with similar skill sets. In addition, some members of the committee are currently working on an independent greenhouse gas emissions visualization project that is organized in a similar fashion to the proposed project, with peer-to-peer learning based on a practical, real-world dataset. Other example projects from our community include a visualization of racial profiling for traffic violations and state incarceration statistics over time.