Supervisor Name(s)	Project Description	Recommended Background	Required Background
Andrew Schaffner	Health outcomes research. I have long standing collaborations with faculty in the departments of Kinesiology & Public Heath as well as Nutrition. For your senior project, you will work as a partner with other undergraduate and graduate students as well as faculty in other disciplines in aspects of study design, analysis, internal reporting, and publication authorship. You may end up on a single intensive project, or serve as a consultant on several projects depending on the stage of the project. Health topics being investigated include malnutrition, disease, poverty, pregnancy (maternal and infant outcomes), obesity, depression, breastfeeding, and environmental toxins. Some studies are designed experiments, while others are observational.	STAT 423 and 418.	STAT 323, 324 or 334. Experience with JMP (and a willingness and ability to help others use JMP).
Anelise Sabbag	Opportunity to get experience with item development and critique. Students will work on improving an existing assessment of statistical learning by identifying badly discriminating items and investigating through cognitive interview with students what could be “wrong” with the items. Problematic items will be modified and new items will be created and included in the assessment instrument. Students will also conduct a pilot test with the new items to assess their quality.		
Anelise Sabbag	Opportunity to learn about measurement and psychometrics using R. Students will learn and compare two measurement approaches to conduct item analysis: classical test theory and item response theory. Classical Test Theory (CTT) assumes that every measurement contains error so the examinee’s observed score on an assessment instrument can be divided into a true score and an error component. Item Response Theory (IRT) uses a probabilistic model to investigate how examinees respond to items in an assessment based on their level of ability on a latent trait.		R
Anelise Sabbag	Opportunity to get experience teaching some introductory statistics topics to non-majors and become familiar with the statistics education research literature. · Choose 2-3 statistical topics to focus on and explore what statistics education research has been done on these topics. · Develop a lesson plan based on recommendations from statistics educators and scholars. · Supervised teaching of these topics (STAT 130 or STAT 217) · Develop and administer assessment to students. · Evaluate students’ performance on topics taught. · Recommendations of changes in lesson plan and assessment.	STAT 410	

Supervisor Name(s)	Project Description	Recommended Background	Required Background
Beth Chance	Research assistant to help with NSF grant evaluating student learning and attitudes with a randomization-based curriculum using hierarchical linear models. The project would involve learning about HLMs, applying such models to the data, and comparing student performance across institutions. Data can be cross-classified with students' prior knowledge and attitudes about statistics, and other demographic information. The data analysis will be the basis of a journal article in statistics education.	Stat 414	Stat 324 and facility with R required,
Beth Chance	An investigation of the bootstrap methodology and its appropriateness for the introductory curriculum. Doing some reading on bootstrapping (I can point you to some classic and new textbooks explaining the idea) including some of the different implementations for bootstrapping (e.g., percentile intervals vs. fancier stuff). The project will involve simulations involving bootstrapping (does it do what they claim, comparing methods); investigation of some ways to incorporate bootstrapping into an intro course (trying out some different applets, technologies that are out there now, see what might be intuitive for an intro student) including the debate of whether to use bootstrapping for intervals and other techniques for testing or to use bootstrapping for sampling and other techniques for random assignment; making some final recommendations as to whether, and if so how, bootstrapping should be taught in the intro course, possibly instead of t-intervals. This could involve testing out some activities in a class like Stat 217 during the year.		Stat 324, Stat 331
Beth Chance	Technical Writing Fellowship: I am currently the "assistant editor" of an international journal in statistics education and am looking for an assistant in finalizing articles for publication in Fall and Spring. I also have a couple of papers of my own in various stages of submission. Furthermore, the Department is planning a new sophomore level communication course (Stat 365). For this project, activities could include learning about style guidelines, suggesting edits to authors, testing recent tools in data visualization, reading articles in statistics education research, co-authoring an article and participating in the review process, and compiling resources and activities on improving communication skills. (Fall start)		
Beth Chance	Integration, testing, and evaluation of plotly into existing javascript applets, might also involve enhancing existing R workspace into R package and posting on CRAN, could include formal evaluation of student use	javascript	R
Beth Chance	Anything related to statistics education, regression, sports	Stat 410 for Stat Ed projects	

Supervisor Name(s)	**Project Description**	**Recommended Background**	**Required Background**
Consultants Vary by Quarter	Statistical Consulting Intern: Each quarter open or two faculty members serve as the Statistics Department consultants for the Cal Poly Statistical Collaboration and Consulting Service. The consultant helps students, faculty, and staff across the campus with the statistical aspects of their research. Some examples of the type of work the consultant does include helping design surveys, sampling and experimental designs, sample size calculations, and data analysis. As a senior project for a statistics major, you would shadow each of the consultants over the year, attending meetings with clients, conducting analyses, and writing summary reports for the client and for a personal journal.		An interest in working collaboratively with other researchers. Strong oral and written communication skills. Successful completion (B- or better) of STAT 323, 324, 330, 331. Completion of STAT 465 would be helpful, but is not required. Experience with JMP and Minitab.
Dennis Sun	The Evolution of Chant. Gregorian chant is the oldest tradition in Western music. Musical notation had not been invented, so chant was transmitted orally from one monk to another. You can imagine that each time a chant was transmitted, the learner might change the chant slightly--either because they misheard or misremembered it (like in a game of telephone) or because they thought it might sound better a different way. As a result, by the time a chant reached a far-off place like England, it sounded quite different from the original chant from Italy. In this project, you will use the differences between these different versions to trace the spread of Gregorian chant throughout Europe, much like how biologists use differences in DNA to create evolutionary trees. This project will require some amount of data cleaning and R programming.		STAT 331, ability to read music
Dennis Sun	Two-Sided Fisher's Exact Test and Confidence Intervals for the Odds Ratio. Fisher's exact test is used to test the hypothesis that two proportions are equal. Unlike the two-sample z-test for proportions or the chi-square test, it does not require large sample sizes and is valid even for small samples. However, there is some controversy about how to define a two-sided Fisher's exact test. The two-sided Fisher's exact test is important because it is the test that is inverted to obtain confidence intervals for the odds ratio. In this project, you will read about the different approaches, as well as investigate an alternative approach that I have in mind. You will implement them and compare their properties (e.g., Type I error and power). This project will likely lead to a publication.		STAT 331, STAT 427

Supervisor Name(s)	**Project Description**	**Recommended Background**	**Required Background**
Dennis Sun	Karl Pearson and the Degrees of Freedom Controversy. In 1900, Karl Pearson introduced the chisquared test. However, he got the formula for the degrees of freedom wrong. The first person to suspect that something was off was Udny Yule (of "Yule-Walker" fame), who built a mechanical contraption to carry out hundreds of simulations. Finally, in 1922, R. A. Fisher came up with a proof that Pearson was wrong, although Pearson never accepted his error. In this project, you will investigate the history of this controversy and develop a lesson plan that uses this history to enliven chi-square tests and simulations.	Interest in history and writing	STAT 427
Dennis Sun	Finite-Sample Properties of Permutation Tests. In classes that teach simulation-based inference, the two-sample t-test is presented as an approximation to the permutation test. But is the t-distribution really a more accurate approximation to the permutation distribution than the normal distribution? And should we pool or unpool the variances? Also, how does the power of the permutation test compare with its parametric counterparts? In this project, you will conduct a large-scale simulation study to make recommendations.		STAT 331, STAT 427
Dennis Sun	Price of Menthol Cigarettes. Menthol cigarettes cause less throat irritation than normal cigarettes and are thought to be more addictive. In this project, you will work with a proprietary data set to answer the question: are menthol cigarettes more expensive than non-menthol cigarettes? This seemingly simple question is more difficult than it sounds because price data is unreliable and brand is a confounding variable.	STAT 418	STAT 331 (with a grade of A- or higher)
Dennis Sun	Identifying Marijuana Retailers. I am working with public health researchers to study the impact of marijuana legalization in California. To do this, we have to be able to identify marijuana retailers. In this project, you will come up with a quantitative definition of a "retailer" and scrape online sources to build a database of marijuana retailers. From there, we would visualize the geographical distribution of retailers and cluster retailers into business types: medical vs. recreational, etc.	Lots of experience with web scraping.	DATA 301

Supervisor Name(s)	Project Description	Recommended Background	Required Background
Department (Bio, Yost)	Plant distributions in Swanton Lagoon: Recently, Cal Poly biology students collected plant occurrence data over 52 acres of the Swanton Lagoon in Santa Cruz county. These data are all georeferenced and occur in a 1 m X 5 m grid over the entire lagoon. These data are currently being entered into GIS. We also have a data layer for salinity. These data were collected because in the next 5-10 years Cal Trans will rebuild Hwy 1 and alter the way Scott Creek into the Pacific Ocean. The goal of the bridge rebuilt is to create better habitat for endangered fish species. Dr. Yost is looking for a student to use geospatial statistics to address several questions: Can we find plant associations/communities based on species co-occurrences? What are the diversity hotpots in the lagoon? When we overlay salinity values of the lagoon, can we predict how the species will change when the bridge is rebuilt? Students would likely work in collaboration with students in biology		GIS skills and/or some geospatial and/or diversity indices
Department (Biology, Rajakura)	The Role of Fire on Plant-Soil-Aspect Relationships in the Poly Canyon, San Luis Obispo, California: - More than 97 acres burned in the Poly Canyon Fire in the Fall of 2017, creating a rare opportunity to examine the role fire plays in generating and maintaining plant diversity on different soils and aspects found within Poly Canyon. Ninety (1x1m) vegetation monitoring plots were set up to determine short- and long-term patterns of plant diversity on north- and south-facing aspects of chemically-distinct serpentine and metavolcanic soils following the fire. We surveyed the vegetation in March, April, May and July to generate species presence/absence and percent cover data. We are interested in investigating if the post-fire diversity is different between 1) burnt/unburnt serpentine and metavolcanic soils, 2) north- and south-facing slopes on burnt/unburnt serpentine and metavolcanic soils, and 3) serpentine soils with and without the fire-retardant (phos-chek), rich in nutrients limiting in serpentine soils.	Bio 121 or some botany/soil science class	Stat 419, 331

Supervisor Name(s)	Project Description	Recommended Background	Required Background
Department (Biology, Rajakura)	Impacts of multiple nutrient element enrichment on native and alien plant species from California's serpentine grasslands: Implications for better management of a threatened habitat: Atmospheric deposition as high as 20 kg Nitrogen (N) ha-1yr-1 have been documented on serpentine habitats in the Bay Area. Understanding the ecological implications of such increases in nutrient deposition is critically important, particularly in a habitat of high conservation value. Alien grasses, supposedly restricted from serpentine habitats due to nutrient deficiencies, are better able to invade the substrate under N enrichment. We are examining the effects of multiple nutrient enrichment on the growth of co-occurring native versus alien species in the greenhouse to investigate how single or multiple nutrient additions to serpentine soils will influence growth and reproduction of two native and two non-native species co-occurring on serpentine soils.	Bio 121 or some botany/soil science class	Stat 419, 331
Department (KPH Project, Alber)	Analyze the Effects of Using the Design Thinking Approach for Health Communication: The effects of using the Design Thinking Approach for developing a stress reduction health communication campaign, measured through user engagement with websites/social media and a message effectiveness survey, will be compared to the effects of an existing stress reduction campaign. This project is in collaboration with Cal Poly Campus Health & Wellbeing and data will be available to analyze in the Winter and Spring quarters. You will assist PULSE students in data cleaning, processing, and analysis.	STAT 323/423	STAT 324
Department (KPH Project, Keadle)	Clinical Trial Data Analysis Promoting Physical Activity Among Cancer Survivors: This pilot study included 50 cancer survivors who all received a fitbit and were randomized to either receive charity-donations if they met their step goal or not. In addition to the primary outcome we collected a lot of psychosocial questionnaires and self-reported health via questionnaires. We also have daily fitbit data for each participant over 18-weeks. You will assist KPH students in data cleaning, processing and analysis.	STAT 323/423	STAT 324, STAT 331
Department (KPH Project, Keadle)	Implementing Machine Learning Algorithms to Process Accelerometer Data: I work with activity monitors (like a fitbit) that collect high frequency acceleration signals (e.g. 80 samples/second) and there are numerous machine learning algorithms that have been developed to predict how much and intensely someone is moving based on the signals. We have completed a validation study to look at the accuracy of these various approaches to process data. You will help improve and develop code for implementing these algorithms and processing/analyzing these data.	Python expertise	STAT 331

Supervisor Name(s)	Project Description	Recommended Background	Required Background
Heather Smith	Survey Research in Industry. I work with a local survey research firm, Opinion Studies, providing statistical consultation on a variety of survey research projects. Opinion Studies works primarily in the fields of litigation and consumer research. Your role on this senior project would involve working collaboratively with me, the principal of Opinion Studies, and with selected clients to provide support for survey research efforts. This support could span the full range of survey research tasks: questionnaire design, training of interviewers, sample selection, data management and cleaning, data summary and analysis, report writing, and meeting with clients. This senior project would start in the fall term.		An interest in working collaboratively with a senior survey researcher and several survey research clients. Strong oral and written communication skills. Successful completion of STAT 323, 324, 330, 331. Successful completion of STAT 421 is preferred. Experience with EXCEL, JMP, and Minitab, especially the creation of graphs, tables, and reports. Strong data management skills in a variety of statistical software. Selection and implementation of statistical methods, especially in a survey research setting.
Heather Smith	Survey Research at Cal Poly. At Cal Poly there are always staff members who are interested in obtaining assistance with the many aspects of survey design and analysis. This senior project would be similar to the "Survey Research in Industry" project except it would involve helping a Cal Poly client instead of an industry client.	Successful completion of STAT 421 is preferred.	An interest in working collaboratively w/ other researchers. Strong oral & written communication skills. Successful completion of STAT 323, 324, 330, 331. Experience with EXCEL, JMP, & Minitab. Strong data mgmt skills in a variety of statistical software.

Supervisor Name(s)	Project Description	Recommended Background	Required Background
Hunter Glanz	It's the Time of the Season for Modelling: What do seasons look like to a satellite? This project will involve collecting and organizing data from a Remote Sensing instrument such as MODIS or Landsat, and investigating the times at which vegetation begins to grow (greening) and fade (browning) each year. And have these times changed over time? How do these times vary by vegetation type and latitude? Is there any more to be gained by including information about land surface temperature?		STAT 330/331, STAT 324, STAT 419
Hunter Glanz	Social Network Comparison Investigation: Suppose we have two social networks, constructed from co-occurrence data. That is, two individuals in a network are connected if they have been observed together. These two networks can be described and compared in a number of ways, but can we evaluate the significance of any differences observed? Can we do things like regression? And how? This project will investigate this via simulation and other methods.		STAT 331
Hunter Glanz	Data Science Specialization: Johns Hopkins University offers a specialization in data science comprised of nine free online courses through Coursera. The series of month-long courses gives an introduction to the basic tools used by data scientists, training in good practices, and an overview of data science. The courses cover reproducible research, version control, R programming, data extraction and cleaning, graphics, machine learning, generalized linear models, and developing data products using Shiny. In this project, you would complete the nine courses and analyze a large data set of your choosing in a project that combines at least three new concepts you learned in the courses.		STAT 331, STAT 324
Hunter Glanz, Maddie Schroth-Glanz	Statistical Analysis of Bioacoustic Data Shiny App: A collaborator in San Diego is interested in improving the training tools available to new bioacousticians. We're mostly dealing with .wav files of different types of marine life (whales, dolphins, etc.) and are interested in summarizing and analyzing these data with statistical tools, and then making this process accessible to students and researchers via a shiny application.		STAT 331

Supervisor Name(s)	Project Description	Recommended Background	Required Background
Hunter Glanz, Rebecca Ottesen	Statistical Analysis for Firestone Walker: Statistical analysis of Firestone Walker lab data, specifically, working with data from compounds detected by GCMS and sensory lab results. The main goal is to be able to find the specific compounds responsible for a specific claim. For example, a specific hop forward brand beer is run through sensory panel and scored based on the hop intensity on a scale between 1 and 10. The same beer is analyzed on GCMS for which 30 plus compounds are quantified and 100 compounds are qualified. Are these compounds associated to the sensory results? An additional need would be to use the GCMS data for process control charting. The analytic plan for this study will be developed in coordination with Firestone Walker and their specific needs. This project would likely involve some travel to meet with Firestone Walker employees.	STAT 331	
Hunter Glanz, Rebecca Ottesen	Sound Test: A modern audio amplifier is intended to act as a voltage source: its output voltage is a larger version of the input voltage and the amplifier delivers whatever current is required by the loudspeaker to maintain that voltage. This necessitates the amplifier having an extremely low output impedance, typically less than 0.03 ohms with a modern design. However, many of the amplifiers that audiophiles feel sound better have higher output impedances, perhaps as high as 4 Ohms. This results in a modified frequency response with a loudspeaker, due to the Ohm’s Law interaction between this output impedance and the way in which the loudspeaker’s impedance varies with frequency. So the question must be asked: do audiophiles like the sound of these amplifiers despite this frequency response modulation or because of it? This project will involve working with Cal Poly's audio club to develop a formal experiment, setup and test speakers. There may also be an opportunity to contribute the analytic results to a published article in Sterophile, a monthly magazine that focuses on high-end home audio equipment.	STAT 423	STAT 323
Jeffrey Sklar	NFL Analytics : Football games produce large amounts of data. Using NFL drive-level data from 2000-2017, a metric to evaluate the performance of each individual drive in a game has been developed taking into account starting field position and drive result (e.g. punt, field goal, touchdown), ultimately resulting in the “expected points added (EPA)” per drive. Variants of the EPA can also be computed to produce a team’s Average Halftime EPA, Average Game EPA, or Season EPA. The current project may involve examining the ability of the EPA to predict game outcomes and/or creating predictive models that include EPA and additional explanatory variables.	STAT324 and STAT 419	STAT 324

Supervisor Name(s)	Project Description	Recommended Background	Required Background
Jeffrey Sklar	Sports Analytics: You can investigate collegiate or professional sports data of your choice to address various team (or individual athlete) performance related questions. Some of my recent students have worked with NFL, NBA, PGA (Professional Golfers' Association), and UFC (Ultimate Fighting Championship) data.	STAT324 and STAT 419	STAT 324
Jeffrey Sklar	Research Methods to Explore Educational Data: Use various statistical methods (for example, linear regression, logistic regression, or discrete-time survival analysis methods) to examine various educational outcome variables such as persistence, drop-out, GPA, or graduation for Cal Poly students.	STAT324 and STAT 417	STAT324
Jeffrey Sklar	Identify Top Parent Donors to Cal Poly: The Gift Planning Office at Cal Poly is interested in identifying top incoming parent prospects for future giving. Using a large database of past and present donors, can a predictive model be developed to help identify parents who will ultimately make large donations to the university?	STAT 324	
Jimmy Doi	Do selected readings/investigations from Analysis of Categorical Data with R (2015) by Bilder and Loughin (1st Edition, CRC Press). Topics will be outside of the usual STAT 418 curriculum. All coding investigations will be done in R. Finally, using your R code you will develop corresponding apps in Shiny.	n/a	Completion of STAT 331 (with a B or better) and STAT 418 (with a B or better).
Jimmy Doi	There are many methods that have been proposed for generating confidence intervals for the binomial proportion π. We typically assess the performance of a confidence interval method by determining coverage probability and expected or average length properties. One popular method was developed by Clopper and Pearson (1934) however their method yields confidence intervals that can be very long and have very high coverage. In a recent paper by Thulin (2014) the author investigates modified versions of the Clopper Pearson interval that yield better length and coverage properties. We will go through this paper and develop R code to reproduce the paper's results. Finally, using your R code you will develop a corresponding app in Shiny. Thulin (2014), "Coverage-adjusted Confidence Intervals for a Binomial Proportion", Scandinavian Journal of Statistics, Vol. 41: 291—300	n/a	Completion of STAT 331 (with a B or better). STAT 418 would be helpful but is not required.

Supervisor Name(s)	Project Description	Recommended Background	Required Background
Jimmy Doi	The first-digit distribution of many data sets is not uniform but follows what's known as Benford's Law. We will become familiar with this law (using journal article references by Ted Hill) and investigate related topics such as the distribution of the first two digits. We will also do selected readings/investigations from Benford's Law: Applications for Forensic Accounting, Auditing, and Fraud Detection (2012) by Mark Nigrini (1st Edition, Wiley). All coding investigations will be done in R. Using your R code you will develop corresponding apps in Shiny. Also we will work with someone from the LA County Auditor Controller and apply hypothesis tests using Benford's Law to detect fraud based on data from the Office of County Investigators.	n/a	Completion of STAT 331 (with a B or better
John Walker	SHINY Applets for Regression and/or ANOVA Concepts. I have a number of ideas for applets to demonstrate advanced ANOVA and regression concepts. I'd like to work with a student to fine-tune some current applets and develop some new SHINY applets in R to demonstrate these concepts. Examples: How do non-normal and unequal variance errors affect regression results? How is multicollinearity affected by different levels of correlation between predictors? How do differing levels of dependence in the regression errors affect the results of a regression and how effective are different tests at detecting dependent errors? The applets would be used by students in future statistics classes.	none	STAT 323, STAT 324, STAT 331
John Walker	Power Analysis for Advanced Experimental Designs. Most statistics software programs have power analysis commands for simple experimental designs; however, often consulting client want to know the power for more advanced designs. Both SAS and R have some commands for power analysis in complicated designs. The first part of this project is to identify which programs support power for which designs. Once we know what SAS and R cannot do, we can develop software tools (most likely in R) to cover a few designs that are not covered by SAS or R. Ideally, we would write functions to compute the power exactly and check those calculations with simulations. In cases where the exact power calculation is too difficult, we can rely on simulation alone for the answer. The student will need good knowledge of some advanced designs from STAT 423 (split plot, repeated measures, etc.). Knowledge of SAS and R is also important to do the simulations, although the student doesn't need to be equally good at both programs. For ease of use, these results could be displayed in a SHINY applet.	none	STAT 330, STAT 331, STAT 423

Supervisor Name(s)	Project Description	Recommended Background	Required Background
Karen McGaughey	In July FiveThirtyEight.com released the data from nearly 3 million tweets sent from Twitter handles connected to a Russian "troll factory" called the Internet Research Agency. The IRA is listed as a defendant in an indictment filed by the Justice Department in Feb 2018 as part of special counsel Robert Mueller's Russia investigation. The tweets were sent between Feb 2012 and May 2018. The data was collected by researchers at Clemson University (Darren Linvill and Patrick Warren) who have a working paper under review detailing their findings. The purpose of this project will be to read the Linvill and Warren paper and recreate their analyses, as well as to carry out two to four novel analyses of these data.	STAT 418 may be helpful	STAT 324; STAT 331 or STAT 330
Karen McGaughey	Anything related to design of experiments, education, mixed models		
Kelly Bodwin	Gender Distribution in Movie Roles: Recently, media attention has been devoted to the rise of female lead roles in blockbuster movies and hit TV shows. In this project, we will leverage data from the Internet Movie Database (IMDB) to quantify the rise (or lack thereof) of non-male characters over the years. An important element of this project will be to construct a method for determining the relative importance of a particular character, based on available data.	STAT 324	STAT 331
Kelly Bodwin	Identifying false ratings. Movie rating website Rotten Tomatoes relies on user-submitted reviews to assess the quality of a movie. Some suspect that companies or promoters fabricate reviews to artificially raise or lower the ratings of particular films. In this summer project, we will seek evidence of tampering by analyzing score differences across reviews with differing linguistic features	STAT 331	
Kelly Bodwin	Social Networks in the Polish Revolution. In 1989, Poland underwent a democratic revolution. What factors lead to this crucial moment, the beginning of the end of the Cold War? We will work with Dr. Gregory Domber of the History department to analyze data on social networks between key members and organizations on the Polish political scene. Your project will consist of creating and editing visualizations of social networks, ideally in Shiny app form, and potentially of performing statistical inference on the change of these networks over time.		STAT 331
Kelly Bodwin	Sports fan support over time. The support for a particular sports team ebbs and flows over time, varying with team success and time of year. A good proxy for the level of interest can be found via Google trends, see e.g., https://trends.google.com/trends/explore?date=all&geo=US&q=%2Fm%2F0jnrk. This project will create a model for the yearly pattern of team support, correcting for playoffs and other factors. We will use this model to discover differences in support patterns between teams and sports.	STAT 331	STAT 418

Supervisor Name(s)	**Project Description**	**Recommended Background**	**Required Background**
Kelly Bodwin	Digital Humanities: Any project related to textual analysis and/or with applications in the Humanities, Social Sciences, or similar.	STAT 331	
Kevin Ross	Topics in probability. Investigate an interesting topic in probability, by reading relevant papers, doing some math, coding simulations, and possibly developing class activities. Projects 2 and 3 provide just two examples.	STAT 425, STAT 425	STAT 305, STAT 331
Kevin Ross	Applications of martingales. “Bernard Friedman’s urn problem” resembles the following. A box contains 1000 red balls and 1 green ball. A ball is selected at random and replaced along with 1000 balls of the same color and 1 ball of the opposite color. If this process is repeated then in the long run the proportion of red balls converges to ½ with probability 1. Wait, what? (Replace 1000 with any number you want and it’s still true.) In this project you’ll investigate, both mathematically and via simulation, problems like this one as well as the general subject of martingales.	STAT 425, STAT 426	STAT 305, STAT 331
Kevin Ross	Variations on collector's problems. The standard version of the “collector’s problem” assumes that all the prizes are equally likely. But what happens if some prizes are harder to collect than others? As a specific example, how many cars would you expect to see before you see a license plate from every state?	STAT 425, STAT 426	STAT 305, STAT 331
Kevin Ross and Dennis Sun	Contribute to the development of Symbulate, a Python package which provides a user friendly framework for conducting simulations involving probability models. In particular, you will enhance the graphical capabilities of the package.	fluency with matplotlib	STAT 305, fluency in Python, strong software engineering skills
Matt Carlton	How do you create simultaneous 95% CIs for multinomial proportions? How do you determine the sample size required to estimate all the proportions to within a prescribed margin of error at 95% simultaneous confidence? And what if you’re sampling from a relatively small finite population? The project would begin with a literature search, involve a heavy amount of simulation, and then (possibly) some theoretical results.	STAT 418	STAT 331
Matt Carlton	I'd like to create a set of Social Justice lesson plans for statistics, piggybacking on the work of Bergen (2016), Taylor & Mickel (2014), and Lesser (2007). The project would begin with a literature review and an exploration of existing resources and websites for social justice data. Deliverables would include a set of lessons/activities that instructors could use in a statistics classroom (both intro and advanced courses) to teach statistical topics through a social justice lens.	STAT 410	

Supervisor Name(s)	**Project Description**	**Recommended Background**	**Required Background**
Prince Afriyie	Multiple Hypothesis Testing: When we reject the null hypothesis in an ANOVA F-test, knowing that at least one of the population means differ leads to the question of which pair of groups have different means. The chances of committing Type I error increases as the number of tests increases. Methods that deal with this issue are called multiple comparisons procedures (MCP). Failure to compensate for multiple hypotheses testing can have important real-world consequences. The project entails literature review of several multiple testing procedures, simulation in R and application to real life data (high dimensional genomic data - microarray or RNA-sequencing).	STAT 425, 426, 427	STAT 305 and STAT 331
Prince Afriyie	This project is a corollary to the first. We will develop an R package for various multiple testing procedures (there is already an R package for MCP, but it does not have the recent multiple testing procedures in literature).		STAT 331.
Samuel Frame	Variation in Table Position Changes: The English Premier League consists of 20 teams. Each team plays the other 19 teams twice for a total of 38 games over the course the season. A team gets 3 points for a win, 1 point for a tie and 0 points for a loss. At any time, the table position of a team is determined by the number of points they have obtained (there are rules to break ties). A team's table position can and will vary throughout the season. The goal of this project is determine and analyze the variation (standard deviation) of the week-over-week position changes. Specific goals are 1) Webscrap and manage the results of every game played in a season; 2) Determine the variation of the position changes and make a time-series graph of this variation each week; and 3) Assess how the variation of the position changes for a given team can be used to determine their final table position in particular if they win the league or are relegated to a lower-division.		CSC 101, STAT 324, STAT 331, STAT 416, STAT 418

Supervisor Name(s)	Project Description	Recommended Background	Required Background
Samuel Frame	Starmine Predicted Earnings Surpsises: StarMine is a subsidiary of Thomson Reuters specializing in quantitative analytics, and one of their metrics is the StarMine SmartEstimate which is StarMine's predicted earnings per share. The predicted earnings surprise is the difference between the Starmine. SmartEstimate and the consensus average (at the date of publication). After the NYSE and NASDAQ exchanges close every day, Thomson Reuters publishes The Day Ahead newsletter (TDA). On the first trading day each week, the TDA StarMine's SEPS for selected companies. I have been collecting this data for several years. Myself and several students have investigated the relationship between the surprise earnings and short-term returns, and also forms trading algorithms based on the surprises. I have two years of data that need to be processed, analyzed, and the results compared to the current work (I believe there was an intervention in the how the surprises are selected for inclusion in the TDA). I am also very curious about how accurate these predictions are. Their sales materials indicate they are 80% accurate (accurate defined as the prediction is in the direction of the actual earnings). It is possible to get historical earnings data from various places including COMPUSTAT, and then compare the published predictions to the actual predictions (including trends in revisions of several consecutive releases).		STAT 324, STAT 331, STAT 416
Soma Roy	Creating Shiny Apps to understand what affects statistical power in situations involving comparison of two or more groups. This project will involve, first, reviewing what statistical power is and how it can be computed in a particular context. Next, teaching tools will be created using R and Shiny apps. This set of interactive tools will be designed and created to provide students the opportunity to explore what impacts statistical power and how, using visuals as much as possible in the context of comparing two or more groups. The target audience for these apps will be students in classes such as STAT 217, 218, 301, 302, etc.		STAT 331 and at least STAT 323 or 324
Soma Roy	Research questions of interest based on survey data related to health studies. You will be required to, among other things, conduct a literature review of existing work related to your questions of interest, and analyze the data to answer questions of interest.		STAT 323, 324; a thorough knowledge of regression analysis will prove to be useful.

Supervisor Name(s)	**Project Description**	**Recommended Background**	**Required Background**
Soma Roy	Using R and Shiny apps to create teaching tools that will be used to provide students the opportunity to explore different types of sampling methods, and how the expected values and variance of estimators for different sampling methods compare, and what affects these values and how. This project will involve, first, reviewing different methods of sampling (such as simple random, stratified, and cluster) and how the expected value and variance of the estimators are calculated in each context. The target audience for these apps will be students in classes such as STAT 217, 218, 301, 302, etc.	STAT 421	STAT 331, STAT 305, and at least STAT 323 or 324
Soma Roy	Research questions of interest based on advanced topics in design and analysis of experiments. You will be required to, among other things, conduct a literature review of existing work related to your questions of interest, and implement the designs in software (JMP, and SAS/R) and possibly analyze data to answer questions of interest.	STAT 423	STAT 323
Steve Rein	Time Series: any topics of interest		Stat 416
Steve Rein	Categorical Time Series: combining the concepts of time series (data correlated across time) and poisson and/or logistic regression results in a novel set of techniques which provide a lot of opportunities for applied, theoretical and computational senior projects.		Stat 416 and Stat 418
Ulric Lund	False discovery rates (FDR) in hypothesis testing. As an alternative to trying to control the Family-Wise Type I Error Rate (FWER) when conducting multiple hypothesis tests simultaneously, one can instead try to control the false discovery rate. You would begin by reviewing several journal articles that gave rise to this technique to understand the basic ideas and definitions involved. Based on your reading of the fundamentals, you could then decide what other aspects of this technique to explore, possibly using computer simulations to investigate how FDR and FWER methods compare under various conditions. Using FDR instead of FWER is often done in biology and genetics applications, so if these are applications of interest to you, this project may be interesting to you as well.		STAT 323, 324, and either 330 or 331.
Ulric Lund	Traffic accident data analysis. California Highway Patrol maintains the Statewide Integrated Traffic Records System (SWITRS), a database of all police-reported accidents in the state of California. Your project could involve an analysis of this data set. Perhaps even more compelling, would be to merge the information from SWITRS with some other publicly available US Census Bureau data set to investigate a hypothesis of interest to you. This project could involve some GIS concepts as well.		STAT 323, 324, and either 330 or 331.