Thanks for applying to the Center for Applied Artificial Intelligence! This is a problem set we use to help us find the best candidates for our team. The problems are based on research from our center, and are designed to help you showcase your skills.

There are sixteen problems in this sheet, and we ask you to complete as many as possible. We know that people have different strengths, so feel free to skip up to four problems. (Note that we consider each numbered question to be a problem, and there are no sub-problems in this exercise.) Each response should be at most one paragraph long, and it should take roughly three hours to answer everything. We are looking for solutions that are correct, succinctly typed, and include clear and intuitive explanations. Submit your solutions through this Google Form. We suggest looking at the Google Form now, since we will ask you to submit some solutions in pieces. Good luck!

Puzzles

1. **Cassidy's Courses**: A member of the CAAI team must complete four more courses to finish her degree. There are eight different science courses and four different economics courses for her to choose from. She must include at least one economics course. How many different sets of courses are possible for her to finish?

2. **CAAI's Gambit**: A standard US penny has a diameter of 0.75 inches. The United States Chess Federation uses chess boards made of squares with an edge length of 2.25 inches. If you drop a penny onto one of these chess boards, what is the probability that it lands entirely within a single square? Assume that the center of the coin lands within the boundary of the chessboard grid.

3. **Movers and Shakers**: I attended a dinner with 9 guests, not including myself. Whenever two guests met for the first time, they shook hands, although not everybody met. After the party, everybody other than myself had shaken a different number of hands. For example, person A shook 1 hand, person B shook 2 hands, and so on. How many hands did I shake? (Nobody shook the same hand twice, and if a guest shook my hand they included it in their count.)

4. **Letters of Introduction**: If you happen to meet two people from the CAAI, there's a 50% chance that both of their names start with the letter J. How many people do you think work in the Center? (Hint: our center is not too large.)

Short-response problems

1. **High Rollers**: In your application, we asked you to roll a die and enter the result. Here are the approximate proportions of numbers entered by applicants:

Number entered	1	2	3	4	5	6
Proportion of applicants	0.148	0.174	0.148	0.296	0.139	0.096

   Suppose that there were 20 applicants altogether. Do you think that everybody used a fair, random die to produce their number? What if there were 100 applicants instead? Now, think about your answers. Can you give an intuitive explanation for them, and why they are the same or different?

2. **Menu Logging**: Jiguang has a favourite lunch place that has 30 different items on the menu. Each day, he chooses his meal uniformly at random from all 30 possibilities, regardless of whether he has eaten that meal before. How many times does Jiguang expect to visit the store before trying every meal at least once? We will accept analytical or simulated solutions to this problem, and you are welcome to include documented code, but your solution should be clearly explained regardless of your approach.

3. **Meal Pal**: as an extension to the previous problem, what is your intuition for what would happen if Jiguang does not select meals with equal probability? That is, if he orders some meals more or less frequently, would it take a longer or shorter time to try everything (on average)?

4. **Too Tired**: While driving to work, you get a flat tire. Even more annoying, you had a flat tire last month as well! You would like to know if this could be ascribed to simply being 'unlucky'. Naturally, you count the number of cars on the highway, and the number of those cars that got a flat tire. You repeat this every day for a month. How would you estimate the probability of getting two flat tires in two months, with this data? What are the limitations of this model?

Doctor Visits

For this project, we have a dataset that captures doctor visits. Each row in the dataset represents a patient visit to a doctor. The variables are a unique `visit ID`, a `doctor ID`, the `patient ID`, the visit `date`, the patient's `age` and `EF` (their ejection fraction), and an indicator for `SCD` (sudden cardiac death). Ejection fraction is a test related to heart health, and `SCD` is 1 if the patient suffered a sudden cardiac death in the 12 months after the visit date, and 0 otherwise. Below is a small sample of this data.

visit ID	doctor ID	patient ID	date	age	EF	SCD
1	A	1	2020-02-221	58	35	1
2	B	2	2019-03-16	31	60	0
3	B	2	2020-04-26	32	60	0
4	A	3	2019-02-23	49	45	0
5	C	3	2020-03-21	49	40	1
	...		...			...

Note that this dataset has been simplified for the purpose of these problems, so you will not need any additional knowledge to answer the following questions.

1. Imagine you are planning to build a model to predict sudden cardiac death using this data. You sample 50% of the rows by sampling random `visit ID`s for a training set, and train a random forest to predict whether `SCD` $= 1$ using the other variables (`age` and `EF`). What mistake did you make? Why is this important? How would you fix the training process?

2. You are interested in ranking the doctors by the proportion of their patients who suffered a sudden cardiac death. For example, doctor A had 5 of their 50 patients suffer a sudden cardiac death (10%), and doctor B had 9 out of 100 patients suffer a cardiac death (9%). You find that the doctors at the top of the list are mostly doctors with only a few patients. Why do you think this is?

3. A range for 'healthy' ejection fraction values is sometimes defined as between as 50 and 65 (although the reality is of course more complex). You are considering replacing your continuous `EF` variable with a categorical variable that captures 'low', 'healthy', and 'high' values, using this range. What are some arguments for or against using a categorical variable here?

4. In much of our work, model specifications are important. Given this data-set and the kind of outcome variable you're focused on here, list two different types of model you might try to fit. Describe some of the pros and cons of each model you're considering. Include the metrics you would judge model performance by.

Election results

For this project, we have a dataset of elections from across the United States. Every state-level general election in the United States for the last 20 years is contained in this dataset. Each row in the dataset represents a single candidate in a single race. The associated variables are a `race ID`, a `candidate ID`, the `office` the race was for, the `state` the race was held in, the `year` the race was held, the `party` the candidate represented, and the `votes` the candidate received. Every (`race ID`, `candidate ID`) pair is unique. Below is a small sample of this data.

race ID	candidate ID	office	state	year	party	votes
1	A	State Senate, District 13	Texas	2016	Democrat	178 277
1	B	State Senate, District 13	Texas	2016	Libertarian	14 447
2	C	Governor	Georgia	2018	Republican	1 978 408
2	D	Governor	Georgia	2018	Democrat	1 923 685
2	E	Governor	Georgia	2018	Libertarian	37 235
...		...				...

Note that this dataset has been simplified for the purpose of these problems, so you will not need any additional knowledge to answer the following questions. You may ignore complexities like primaries, run-off elections, districts, etc. for the following questions.

You want to capture the probability that a given candidate wins a given race, using the dataset above. Assume that within each race, the candidate with the most votes wins the race, and that there are no ties.

1. Describe how you would calculate the outcome variable for this model. Do not include any code for this question: imagine that you are explaining to your colleague how you created the outcome variable.

2. Before you are ready to fit any models, you want to create some additional variables (also known as features). Your colleague suggests calculating the proportion of races won by each party in each state within the dataset, and using that as a variable. For example, if Democrats have won x% of elections in Texas, you would set your new variable to x in the first row (and any other election with Democrats in Texas). However, you know that this is a bad idea. Why is that? Can you think of something similar that avoids this problem?

3. You are still trying to create variables for your model. Another colleague suggests replacing the categorical `state` variable with several continuous variables for population, location, and the variable you defined previously. What are some arguments for using a single categorical variable (with 50 levels), versus a few continuous variables?

4. You are now ready to split your full dataset into a training set and a test set. Your colleague has now suggested using five random values of `office` to keep as a holdout. For example, if you sampled "Governor", you might put every Governor race in the holdout set. What are some arguments for doing this? Next, you consider doing the same thing for the `state` variable. Does the same logic apply?