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 three sections in this sheet, each with several questions. Please complete as many questions as possible. We know that people have different strengths, so feel free to skip up to four questions. The responses for each section will vary, but as a general rule you won't need more than a single small paragraph to answer any question. We suggest taking one hour per section to work out your answers. We are looking for solutions that are correct, succinctly typed, use intuitive explanations, and are clearly written.

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!

Short response puzzles

This is a series of analytical thinking questions. Your response should consist of a short answer, and a brief explanation.

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 10 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, if the first guest shook 3 hands, then no other guest (other than myself) shook exactly three hands. 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 an exactly 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.)

5. **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?

6. **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.

Sudden cardiac death

For this project, we are trying to estimate patient risk of suffering a 'Sudden Cardiac Death'. We have a dataset to help with this. Each row in the dataset represents a patient visit to a doctor. The dataset has several ID variables: a unique `visit` ID, the `doctor` ID, and the `patient` ID. In addition to these ID variables, the dataset includes 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	doctor	patient	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. 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?

2. You are now interested in preparing your dataset for a regression to estimate patient risk of sudden cardiac death. The range of 'healthy' ejection fraction values can be defined as 50 to 65 (reality is more complex, naturally). 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?

3. Before you are ready to fit any models, you want to create some additional variables (also known as features). Your colleague suggests counting the number of times each patient visits a doctor, and adding this `patient visit count` variable to every row with that patient. Do you think that this variable would improve the performance of your model?

4. For the `patient visit count` variable, do you think this is a good variable to include? Why or why not?

5. You are now preparing to fit your model. You sample 50% of the rows by sampling random `visit IDs` for a training set, and train a random forest to predict whether `SCD` $= 1$ using the other variables (`age`, `EF`, and potentially `patient visit count`). What mistake did you make? Why is this important? How would you fix the training process?

6. Suppose that you have now fit your model to estimate the probability that `SCD` $= 1$. While checking your results, you look at the estimate for an 80 year-old patient with 'healthy' `EF`. The model predicts that this patient has a 10% risk of sudden cardiac death in the next 12 months. But you know that about 0.1% of US adults die of SCD each year, and less than 1% of US adults over the age of 80. Does the model prediction seem high? What other reasons can you think of that might cause this? (Assume that your code is working properly.)

Arrest pairs

A person arrested in the United States faces a judge within 24 hours of arrest. That judge makes an important decision. Where will this person wait for trial? Must they sit in jail? Or can they go home? Whether a person is jailed depends on the risk that person poses: would they flee? Would they commit a crime?

For this project, we need to measure survey participants' ability to predict whether arrested individuals are likely to be detained prior to their trial, based on their face. Start this exercise by downloading the `prisoner.csv` and `arrest.csv` datasets from this Google Drive. These are synthetic datasets that mirror arrest data from a county in the United States. The first dataset, `prisoner.csv`, captures demographic details at an individual level. Each row of this dataset represents an (imaginary) individual who was arrested in the time period covered by the dataset. The second dataset, `arrest.csv`, captures the details of (imaginary) arrests within the county. Each row of this dataset represents an arrest during the time period covered by the dataset.

You may assume that both datasets have been lightly cleaned. Python and R are preferable for this task. You may use any documentation/help files/online resources to help you complete the tasks. When assessing your solutions, we will consider the correctness of your output, the quality of your code, and how well you can describe your approach. When evaluating the quality of your code, we will take your limited time into consideration.

To begin, here are some exploratory questions for you to understand the data. You are required to submit both your code and answers for the following questions.

1. How many prisoners are included in these datasets?
2. How many arrests are included in these datasets?
3. What time period does this dataset cover?
4. On average, how many arrests are made each day?
5. What is the proportion of arrests that had a bond amount greater than zero?
6. What is the proportion of arrests that are male?
7. How many prisoners are arrested more than one time?
8. For the prisoners arrested more than once, how many times are they arrested on average?
9. Calculate the age of each prisoner on the day that they were arrested. What is the median age of an arrested individual?

Now, here is a brief on the survey and the dataset required. We have included a variable, `release_flag`, that is 1 if an individual is released prior to their trial, and 0 if an individual is detained prior to their trial. Our task was to create a dataset that contains pairs of arrests. The final dataset should to satisfy the following criteria:

- Each row should capture a single pair of arrests. As such, it should include columns for `arrest_id_left`, `prisoner_id_left`, `arrest_id_right`, and `prisoner_id_right`.
- Every pair of arrests has prisoners with the same race and sex. For example, a pair of two white men is valid, while a pair with one man and one woman is not.
- In every pair, one person was released, and the other was detained.

Review the code in either `create_survey_dataset.R` or `create_survey_dataset.R` on the Google Drive. Choose whichever language you are more confident with: both of these scripts do the exact same thing. Each script produces two datasets: `arrest_pairs_a` and `arrest_pairs_b`. These datasets have the same structure, but are prepared differently. Your task is to compare method A and method B in the following questions.

1. Which method do you think runs faster? Is this true if the dataset gets larger or smaller?
2. Do both methods meet all of the criteria described above?
3. Which method do you prefer? Give some arguments for or against your choice.