**Collection of mock interview questions from alumni interviews. Please put your name by any questions you add so people know who to answers.**

# Coding / CS Fundamentals

1. What is the function below doing and what are its assumptions? What’s its time and space complexibility? (Bethan, Emma)
   * def f(array, target):  
      left, right = 0, len(array) - 1  
      while(left + 1 < right):  
      middle = (left + right) // 2  
      if array[middle] == target:  
      return middle  
      elif array[middle] > target:  
      right = middle  
      else:  
      left = middle  
      if array[left] == target:  
      return left  
      if array[right] == target:  
      return right  
      return -1
2. How would you change it (above) to deal with unknown length (or very long lists) (Bethan)
3. Generate a deck of 52 cards. (NickK)
4. Write a function to draw a random hand of 5 cards. (NickK)
5. Write a function to determine if a hand has a pair. (NickK)
6. Write a function to generate ith (within an assumed length) Fibonacci number. (Sophia)
7. Write a code snippet to print out the numbers between 2000 and 2300 (inclusive) that are divisible by 7 but not a multiple of 5 (Holly)
8. Write a function that takes 3 positive numbers as values and returns either A,B or C corresponding to a distribution defined by the input values. (NickV)
9. Code a Fibonacci number generator. (Kurt)
10. Using the data below: (Bethan)

users = 'charles annie steven anne'  
 states = 'ca tx ny ca'  
 ages = 23, 57, 33, 28  
 hobbies = 'hiking sitting standing sewing'

* + Create a list of dictionaries from the above data.
  + Sort by name
  + Find dictionary of maximum age

# SQL

1. Here are 2 tables. What kind of metrics might advertisers be interested in? (Bethan)
2. Write a SQL query to return ROI.
3. You have a table with three entries member id, name and recommended by id. Find the member who were recommended by some other members. (Nairwita)
4. Can you print out the employee names for which the department name is not available? (tables provided: employee table with various info incl. deptid, and department table with deptid and dept name. Note that not all the deptids from the employee table were in the department table, and vice versa) (Holly)
5. You have a table with three columns: 1) time, 2) ad ID, and 3) activity which contains “shown” (ad was shown to someone) or “clicked” (ad was clicked by someone). Write a query that returns the click through rate of each ad. (Kurt)
6. Your company has a table UserStatus which contains information (up until yesterday) of userids, and status which is either active, churn, or new. There is another table, called Active, which has a list of userids that were active today. Your task is to generate a new up-to-date UserStatus table. Note that there may be new users in the Active table that weren’t previously observed. (Dawnis)
7. You have a table from a Tennis club called Members. It contains columns MemberId, Name, and RecommendID. Where RecommendID is the member that recommended that member. Return all the members that recommended someone.
8. You have a table named ‘users’ that contains three columns: 1) user\_id, 2) user\_name, and 3) reference\_id (the id of the person who referred the user). Write a SQL query that returns the name and ID of people who referred someone in alphabetical order.
9. Build a query that help us understand how users move through the onboarding flow as they sign up for a website; are there any points at which they drop out more? You can assume the tables contain a column for user id, timestamp for an event, and event type (e.g., events: “clicks” or “views” in a page) (Belen)
10. You want to investigate the validity of your interview process for software engineers.

We have two tables with interview information. We only want interview types that are of the "coding" type, but people sometimes write "Android coding" or "managerial coding," for example.

# Machine Learning

* Give an intuitive explanation of the bias / variance tradeoff. (NickK, Sophia, Ruoying)
  + Try to answer in a intuitive way aiming at the general public (refer to Andrew Ng’s lecture note). E.g. Bias: how well the model can be generalized to other samples.
  + Irreducible error
* Give examples of ML models that tend to have high bias or high variance (NickK, Belen, Sophia)
* How do you know if you are underfitting or overfitting? (NickK, Belen, Holly, Ruoying)
* If you find that your model is overfitting (or underfitting), how you would fix it? (Belen, Holly)
* How do you interpret learning curves? (Dawnis)
* Explain how k-means clustering works (Rohan)
* What do you know about linear regression? Derive it. (Dave)
* Extend the linear regression to multiple dimensions - what would be the most computationally expensive part of the model? (Dave)
* Code a K-nearest neighbors algorithm (Dave)
* What is a neural network? Can you explain this in simple terms to a non-expert? (Dawnis)
* Case study: A phone manufacturer: 100 million phones with 1 out of 1000 defective rate. Each phone has measurements from 1000 sensors. How would you explain to the manufacturing workers to find and fix the problem? (Ruoying, Gaomin)
  + How to deal with high dimension -> L1
  + Imbalanced data, downsampling vs upsampling
  + Among logistic regression, random forest and SVM, why logistic regression works best in this situation?
* LinkedIn (question) - If you are a recruiter and have linkedin premium account, you are going to search for the candidates you want to hire. What kind of model you can apply to give him a customized search result? (Nairwita)

- what type of data ?

- How will you preprocess the data?

- what kind of problem is this?

- what model you would like to apply?

- Is it a linear or nonlinear model?

- Can you explain how the model works?

- What metric should you use to evaluate the model performance?

- Why do think this metric is good fit for this model since the data is imbalanced?

* What are the steps you should follow if the data is imbalanced? (Nairwita)
* How to you know a university is accepting and rejecting correcting candidate, what type of of data would you collect and to figure out this question (Sophia).
* You’re building a classifier to predict cancer and you get a great accuracy of 98%. What do you conclude from this? (Holly)
* How do you deal with class imbalance? (Holly)
* What is the difference between different forms of regularization? (Holly)
* What are some types of ML approaches for unsupervised learning? What metrics do you measure for these? (Holly)
* What are the assumptions of linear regression (Jason)
* Explain SVM, including what a support vector is and what hyperparameter tuning entails (Jason)
* What is the difference between logistic regression and linear regression? (Emma)
* Describe the optimization process while training. What is the difference between gradient descent and stochastic gradient descent? (Emma)
* How do you know when you reach the global minimum? (Emma)
* What is decision tree, how to prevent it from overfitting (Pan)
* What are the benefits of L2-linear regression (Interviewer mentioned it deals with correlation among features); How L1 and L2 penalty for linear regression affects the estimated parameter (explain why L1 gives sparse result but not L2). (Pan)
* Explain gradient descent (Pan)
* How does the computational expense of k-nearest neighbors scale with the testing set size? (Kurt)
* You fit ordinary least squares with one feature and get a low p-value and a low R^2 value. What do you take from this? (Kurt)
* Draw a confusion matrix and describe different elements (TP,FP,...) , describe precision and recall. Write recall in terms of precision or vice versa (hint: use Bayes rule) (Alireza)
* How would you build a model that classifies whether a website is a phishing site? What model and features would you use? How would you validate your model? (Belen)

# Probability and Statistics

* How many times would you expect to roll a dice before you see all 6 sides? (Bethan)
* Suppose you are playing a dice game in which you earn in dollars the amount that you roll. How much can you expect to earn? Suppose we modify the game so that you can re-roll after your first roll. How does this change your earnings? (Dawnis)
* You are at the beach and counting dogs, for some reason. After an hour, you note that the probability of seeing at least one terrier is 64%. What is the probability of seeing at least one terrier in the next 30 minutes based on this experience? (Dawnis)
* Given a student paper with misspellings, one grader finds 10 misspellings, another finds 20 misspellings. 5 of them overlap, so the lower bound is 25 actual misspellings. How do you estimate the true number of misspellings? (Dawnis)
* If you have model with small p-value and high R^2 score - what does that mean? (Nairwita)
* In laymen’s terms, explain p-value, standard deviation, standard error, and confidence intervals. (Kurt)
* You arrive at heaven’s gates, and see 3 doors, with the option to pick one. One of them takes you straight into heaven, one sends to you hell for a day before trying again (with shuffled doors), and one sends you to hell for two days before trying again? What is the expected value of time you’ll spend in hell prior to being admitted to heaven? (Alex)
* You have a jar with 1,00 coins. One of them will always come up heads, the others are fair. You randomly remove one coin from the jar, flipping it 5 times. It comes up heads every time. What are the odds that this is a fair coin? (Alex)
* What is multicollinearity? When does it matter? When can we ignore it? (Belen)
* You have an unfair coin which comes up heads with probability ¾, tails ¼. How can you generate a fair coin out of this? (Rohan)
* You’re working at a bank, and they want a model of the probability that a person will default. How would you build or compute this? (Nick V)

# Experimentation

* Facebook: Experimentation on login/sign up page to increase sign up rate. (Doris)
* Someone runs a BBQ restaurant. How to find/design a good barbecue sauce? (Pan)
* Facebook wants to improve the action rate on reports of bad behaviors. The team has proposed a change, how would you test if it works? What metrics would you use? How long would you run the experiment? What are the values you need to define? (Knick)
* Netflix rolled out a new membership program, where people now pay-per-view. This was a massive change and it got a lot of press coverage. How do we assess the impact of this change? Hint. You can’t run an experiment (Belen; q’ also in business sense)

# Business Sense

(Translation of business problems to data problems, actionable insights from data, data product design)

* You work at a University. The University thinks there is a problem with admissions. How do you test this and correct it? (Nick)
* The PM at a tech company is concerned that there has been a 10% decline in logins week after week. Should we be concerned? (Belen)
* Netflix rolled out a new membership program, where people now pay-per-view. This was a massive change and it got a lot of press coverage. How do we assess the impact of this change? Hint. You can’t run an experiment (Belen; q’ also in experimentation)
* Facebook wants to introduce a feature highlighting fake news. What will people think about this? (Bethan)
* Twitter is considering building a separate platform just for dentists to communicate with one another, called Tweet-tooth. How large is the market for Tweet-tooth? (Rohan)
* An e-learning platform wants to known how effective their ads are on LinkedIn. As a data scientist at LinkedIn, how would you answer this question? (Ruoying, Bethan, Doris, Joy, Nick)
* Recently FB introduced additional reaction buttons in addition to the thumb up. How would you evaluate the performance of this new feature? (Ruoying, Bethan, Doris, Joy, Nick)
  + Key points: how performance is measured? (user engagement). How to measure user engagement? (comments, total number of reactions, total people reacted, re-posts etc) Might worry about how this feature would affect the performance of other features.