Interview Question Compilation

Signal Data Science

Contents

Overview 4

General data science 5

Model taxonomy 5
Regression 5
Classification 5
Bias–variance tradeoff 5
Cleaning data 6
Problematic datasets 6
Recommender systems 6
Natural language processing 7
Other 7

Specific techniques 8

Linear regression 8
Regularized linear regression 8
Logistic regression 8
Decision trees 9
Support vector machines 9
Principal component analysis 9
Other 10

Classical statistics 11

General 11
Frequentist inference 11
Distributions 12

Product-based questions 13

Predictive modeling 13
Business decisions 13
Financial services 14
Recommendations 14
Natural language processing 15
Metrics 15
Other 15
A/B testing 16

SQL 18

Algorithms 19

Mathematics 19
Strings 19
Matrix traversal 20
Reservoir sampling 21
Searching a tree 21

Sorting and selecting 21
Sorting 21
Selecting 22
Dynamic programming 22
Index manipulation 23
Parallelization 23
Data structures 23
Other 24
Language-specific 25

Probability **26**
General 26
Flipping coins 28
Fair coins 28
Biased coins 29
Consecutive heads 30
Games and profit 30
Other 31
Rolling dice 31
Decks of cards 34

General mathematics **35**
Comparisons 35
Fermi estimates 36
Logic 37

Communication **39**

Uncommon topics **40**
Image processing 40
Estimation methods 40
Advanced data science 41
Advanced algorithms 41
Other 42

Unorganized questions **43**
Programming 43
Other 43
120 Data Science book questions (unsorted) 45

Overview

The following problems are mostly compiled from Glassdoor. The probability and general mathematics sections have been supplemented with quant/trading interview questions.

All material through the SQL section are strongly recommended for all readers. The algorithms and probability sections are together more comprehensive than necessary for each individual student, but you may want to focus more on those sections if you want a software-focused or finance-focused data science job rather than a more standard industry position.

Working through these questions with a partner is strongly recommended, especially the open-ended, product-focused ones.

General data science

Model taxonomy

- What is the difference between supervised and unsupervised learning? Illustrate with examples.

  (*Source:* Mu Sigma, Verizon)

- What are the differences between parametric and nonparametric models? Give examples of each.

  (*Source:* Walmart)

- What is the difference between a bagged model and a boosted model?

  (*Source:* Walmart)

Regression

- What is heteroskedasticity?

  (*Source:* Vodafone)

Classification

- How can you deal with an unbalanced dataset in binary classification?

  (*Source:* i360, Microsoft, Amazon, A9.com, KPMG)

- What is an ROC curve?

  (*Source:* PayScale, i360, MachinePulse)

- What is the AUC?

  (*Source:* i360)

Bias–variance tradeoff

- How can you detect underfitting?

  (*Source:* Tagged)

- How can you gauge the degree of overfitting?

  (*Source:* Kabbage)

- How can you prevent model overfitting?

  (*Source:* Khan Academy, Netflix, Quora)

- What is cross-validation?

  (*Source:* Tagged)

- What is the bias–variance tradeoff?

  (*Source:* Microsoft)

- When might it be preferable to include fewer predictors over many?

  (*Source:* 120 Data Science Interview Questions)

Cleaning data

- What steps do you take to prepare and clean data before applying machine learning algorithms?

  (*Source:* Microsoft)

- How can you impute missing data?

  (*Source:* AirBnB, HERE, Amazon)

- What would you do if removing rows containing missing values from your data causes bias?

  (*Source:* AirBnB)

Problematic datasets

- How can you deal with multicollinearity?

  (*Source:* Amazon, Walmart, Uber)

- What is the curse of dimensionality?

  (*Source:* 120 Data Science Interview Questions)

- What is target leakage?

  (*Source:* PayScale)

Recommender systems

- What is the cold start problem and how can it be dealt with? What about an analogous "hot-getting-hotter" problem?

  (*Source:* Tagged)

- How would you create a recommender system which can adapt to different metrics in real time?

  (*Source:* Quantifind)

Natural language processing

- What is an *n*-gram?

  (*Source:* Yelp)

Other

- How would you use categorical variables as predictor variables in a model?

  (*Source:* Tagged)

- How would you deal with a categorical variable which takes on many different values?

  (*Source:* Microsoft)

- Suppose you run an analysis that suggests a certain factor predicts a certain outcome. How would you determine if the relationship is causal?

  (*Source:* Treehouse)

- What are some ways to measure distances between data points?

  (*Source:* Microsoft)

- How would you perform feature selection given a large set of sparse features?

  (*Source:* StubHub)

Specific techniques

Linear regression

- What are the assumptions of a linear regression model?

  (*Source:* Hewlett Packard Enterprise, Walmart)
- How would you weigh points in a linear regression model non-uniformly?

  (*Source:* Ericsson)
- What is the difference between R^2 and adjusted R^2?

  (*Source:* Vodafone, 120 Data Science Interview Questions)

Regularized linear regression

- What is regularization and why is it useful?

  (*Source:* 120 Data Science Interview Questions)
- What are the differences between L^1 regularized linear regression (LASSO) and L^2 regularized linear regression (ridge regression)?

  (*Source:* AIG)

Logistic regression

- What distributional assumptions are made in a logistic regression model?

  (*Source:* Symphony Teleca)
- What are the differences between linear and logistic regression?

  (*Source:* Symphony Teleca)
- In logistic regression, what is the relationship between the coefficients and predictions of log odds?

  (*Source:* IBM)
- What is the cost function of logistic regression?

  (*Source:* Kabbage, pymetrics)
- How can you model logistic regression in terms of a single-layer neural net?

  (*Source:* Workday, Amazon)

Decision trees

- What is the difference between a decision tree and a random forest?

  (*Source:* Khan Academy)

- What is the difference between a random forest and a gradient boosted tree?

  (*Source:* Booz Allen Hamilton, i360)

- How does a random forest work?

  (*Source:* Cablevision Systems, Cyient, YP)

- Why would you choose to use a random forest?

  (*Source:* Saama Technologies)

- How can you tune a random forest?

  (*Source:* Instacart)

- How does gradient boosting work?

  (*Source:* Microsoft, Amazon, ZEFR)

Support vector machines

- Why do support vector machines try to maximize the margin?

  (*Source:* Zillow)

- Describe 3 kernel functions and their applications.

  (*Source:* LinkedIn)

Principal component analysis

- What is principal component analysis? Explain both intuitively and mathematically.

  (*Source:* Schlumberger, Uber)

- What is the purpose of principal component analysis if multicollinearity can be dealt with via L^1 regularization (LASSO)?

  (*Source:* ZestFinance)

Other

- What is the K-means clustering algorithm?

  (*Source:* Symphony Teleca, Amazon, Active Network, Verizon)
- What is k-Nearest Neighbors?

  (*Source:* Cablevision Systems)
- What are some different time series forecasting techniques?

  (*Source:* Uber)

Classical statistics

General

- Define mean, median, mode, variance, and standard deviation.

  (*Source:* Mu Sigma, Microsoft)

- When would you use the mean instead of the median or vice versa?

  (*Source:* HERE)

- *Regression to the mean:* Let's say you have a very tall father. On average, what would you expect the height of his son to be? Taller, equal, or shorter? What if you had a very short father?

  (*Source:* 120 Data Science Interview Questions)

- What is the law of large numbers?

  (*Source:* TrueCar)

- What is the difference between a box plot and a histogram?

  (*Source:* Microsoft)

- Suppose that many users make a small number of purchases and a few users make a large number of purchases. What do you expect the plot of average revenue per user to look like?

  (*Source:* Square)

Frequentist inference

- What is a p-value? How do you interpret one?

  (*Source:* CNA, State Farm, Workday, 120 Data Science Interview Questions)

- How might your interpretation of a p-value change if you had a much larger sample size?

  (*Source:* State Farm)

- What is a confidence interval? How do you interpret one?

  (*Source:* CNA, Vodafone, 120 Data Science Interview Questions)

- Out of a sample of 100 items, 25 are defective. What is the 95% confidence interval for the proportion of defective items?

  (*Source:* Tesla Motors)

- Suppose that the population clickthrough rate on Facebook ads is p. We select a sample of size N and examine the sample's conversion rate, denoted $\hat{p}$. What is the minimum sample size N such that the probability that $|p - \hat{p}| < \delta$ is at least 95%?

  (*Source:* Facebook)

Distributions

- If you can sample from a Normal distribution for which you know the mean and variance, how can you simulate sampling from a uniform distribution?

  (*Source:* 120 Data Science Interview Questions, Altimus Capital)

Product-based questions

Predictive modeling

- How would you design a model for Zillow's Zestimate (a home price forecast)?

  (*Source:* Indeed)

- How would you predict a student's success on a subsequent question based on knowing their performance on past questions?

  (*Source:* Khan Academy)

- How could you predict whether or not a customer will buy a particular product today given the customer's income, location, profession, and gender?

  (*Source:* Amazon)

- How would you determine whether or not having more Facebook friends increases the probability that a Facebook user is still active after 6 months?

  (*Source:* Facebook)

- 1000 users played Level 1 in your mobile game, but only 200 users played Level 2. What questions would you ask to determine the source of this difference?

  (*Source:* DeNA)

- Given activity logs from a mobile game, what features would you build to predict churn?

  (*Source:* PayScale)

- What are some possible problems that might arise from trying to predict the gender of your customers from only 100 labeled data points?

  (*Source:* Capital One)

- How can you determine when a user is beginning to search for a new job?

  (*Source:* LinkedIn)

Business decisions

- If you had a magic wand and could choose to send 100,000 users to any number of businesses to improve Yelp's functionality, what would you do?

  (*Sources:* Yelp)

- How would you use data science to increase the number of customers for a coffee shop?

  (*Source:* Symantec)

- You own an art museum and are considering putting up an exhibit by a new artist. How would you go about making a data-driven decision?

  (*Source:* Yammer)

- How would you determine where to place an advertisement using data about bicycle traffic around the city?

  (*Source:* Civis Analytics)

- If you had to propose a new Yelp feature, what would it be?

  (*Source:* Yelp)

- What feature would you add to Facebook? Describe how you would pitch it to executives and measure its success.

  (*Source:* Facebook)

- Given data on TV viewership, how would you optimize the allocation of ads on a TV channel? What external events might influence your predictive algorithm?

  (*Source:* Rovi)

Financial services

- How would you build a model to detect credit card, insurance, or tax fraud?

  (*Source:* Capital One, Accenture, The Hartford)

- How would you build a model to detect insider trading given the data of your choice?

  (*Source:* Civis Analytics)

- How would you evaluate the performance of a new credit risk scoring model?

  (*Source:* Square)

Recommendations

- How would you recommend products to customers? (Consider products for Amazon, audiobooks for Audible, jobs for LinkedIn, users to follow for Twitter, etc.)

(*Source:* Amazon, Audible, LinkedIn, Twitter)

- How would you evaluate the performance of two different recommender systems for generating Facebook friend suggestions?

  (*Source:* Facebook)

- How would you build a user's profile based on past Google searches to make future searches more relevant?

  (*Sources:* Fino Consulting)

Natural language processing

- How would you design a model to detect essay copying in a class of 200 students? What if you had 1,000,000 students?

  (*Source:* Kabbage)

- Given a large collection of books, how would you tag each book with its genres?

  (*Source:* Workday)

Metrics

- How would you define and test a metric for user engagement? (Consider each company-specific variant.)

  (*Source:* Twitter, Facebook)

- What metrics would you use to determine if Uber's strategy of using paid advertising to acquire customers works? How would you determine an acceptable cost of customer acquisition?

  (*Source:* Uber)

- How would you create and test a metric for comparing two users' ranked lists of movies or TV shows?

  (*Source:* Netflix)

Other

- How would you identify recently trending queries or topics?

  (*Source:* BloomReach)

- What are some possible causes of an anomaly in web traffic data?

  (*Source:* AirBnB)

- What are some possible causes of a sudden spike in user uploads, particularly image uploads? How would you test these causes?

  (*Sources:* Yammer)

- How would you infer the roles of staff in an organization based on data collected about their phone calls?

  (*Source:* ThreatTrack Security)

- Given access to Flickr's database, how would you determine which camera models fail the most often?

  (*Source:* iFixit)

- How would you take millions of users, each with hundreds of transactions among tens of thousands of products, and group the users together in a meaningful way?

  (*Source:* Apple)

- How would you detect fake data input from your users?

  (*Source:* Facebook)

- How would you determine if survey responses were filled in at random by particular users?

  (*Source:* Glassdoor)

A/B testing

- In an A/B test, how can you check if assignment to the various buckets was truly random?

  (*Source:* 120 Data Science Interview Questions)

- What might be the benefits of running an A/A test, where you have two buckets who are exposed to the exact same product?

  (*Source:* 120 Data Science Interview Questions)

- What would be the hazards of letting users sneak a peek at the other bucket in an A/B test?

  (*Source:* 120 Data Science Interview Questions)

- What would be some issues if blogs decide to cover one of your experimental groups?

  (*Source:* 120 Data Science Interview Questions)

- How would you conduct an A/B test on an opt-in feature?

  (*Source:* 120 Data Science Interview Questions)

- How would you run an A/B test for many (20 or more) variants?

  (*Source:* 120 Data Science Interview Questions)

- How would you run an A/B test if the observations are extremely right-skewed?

  (*Source:* 120 Data Science Interview Questions)

- What variables determine the duration of an A/B test?

  (*Source:* Thumbtack)

- Suppose that you performed A/B testing for a change in a search engine algorithm, but in the B bucket, the programmers made a mistake—the search results were very poor. Nevertheless, the users in the B bucket clicked on more results *and* on more ads. How do you interpret these results?

  (*Source:* Yammer)

- How would you use A/B testing to compare different recommender systems? What metrics of user behaviors would you look at and how would you decide on the length of the test?

  (*Source:* Kabbage)

- How would you A/B test different ad locations on a webpage?

  (*Source:* Yelp)

- How would you design an experiment to compare the performance of two different coupon campaigns: one where all users are sent the most popular coupons and one where users are sent personalized coupons generated by a recommender system?

  (*Source:* Walmart)

SQL

- What is the difference between a primary key and a foreign key?

  (*Source:* 120 Data Science Interview Questions)

- How would you get all rows from one table with keys that do not appear in a different table?

  (*Source:* Adwerx)

- Given data on Facebook users friending or defriending each other, how could you determine whether a given pair of Facebook users are currently friends?

  (*Source:* Facebook)

- Given a table `tbl(col)`, compute the mean, median, and mode of `col`.

  (*Source:* Facebook)

- Given two tables `friend_request(requester_id, time, sent_to_id)` and `request_accepted(acceptor_id, requester_id)`, calculate the global friend request acceptance rate.

  (*Source:* Facebook)

- Given a table listing Facebook users and their friends as well as a table listing Facebook users and the pages they liked, write a query which generates page recommendations for yourself. Do not recommend pages you have already liked.

  (*Source:* Facebook)

Algorithms

Mathematics

- Calculate the nth Fibonacci number.

  (*Source:* Accenture, Belvedere Trading, Deutsche Bank, Jump Trading, Moz)

- Calculate the nth Fibonacci number with recursion. If `f(100)` takes 1 second, how long will `f(101)` take?

  (*Source:* Morgan Stanley)

- Find the nth prime number.

  (*Source:* J.P. Morgan)

- Determine all the prime numbers up to n.

  (*Source:* Accenture)

- *Newton's method:* Calculate the square root of a number.

  (*Source:* 120 Data Science Interview Questions, Bloomberg L.P.)

- Determine if an integer is a palindrome without converting it to a string.

  (*Source:* Adobe)

- *Exponentiation by squaring:* Calculate a^b.

  (*Source:* LinkedIn)

Strings

- Read in a large, tab-delimited file of numbers and count the frequency of each number.

  (*Source:* Capital One)

- Determine if a string is a palindrome. (A palindrome is read identically forwards and backwards.)

  (*Source:* Time Inc., Accenture, Microsoft)

- Determine if a string is a palindrome using recursion.

  (*Source:* StoryCloud)

- Find the longest palindrome in a very long string.

  (*Source:* OnDeck)

- Given a string with spaces and a specified line width, justify the string; that is, add more spaces in a balanced way so that the string is of the desired length and the number of spaces between adjacent words are as similar as possible.

  (*Source:* Apple)

- Find all substrings of a string which begin with a specified word and end with a (potentially different) specified word.

  (*Source:* Adobe)

- Generate all valid combinations of *n* pairs of parentheses.

  (*Source:* Adobe)

- Determine if a string of parentheses ("()"), brackets ("[]"), and braces ("{}") is balanced. (For example, "([])" is balanced but "([)]" is unbalanced.)

  (*Source:* Hired.com)

- Add two binary numbers represented as strings without using string to integer conversions. (For example, "100" + "111" = "1011".)

  (*Source:* Facebook)

- *Run-length encoding:* Find the longest "run" of consecutive identical characters in a string.

  (*Source:* Etsy)

- When someone begins to enter in the name of a contact or an email address on their iPhone, a list of possible matches is shown to the user. How would you generate such a list so that the search is performed rapidly and updates in response to every new character entered by the user?

  (*Source:* Apple)

Matrix traversal

- Print the elements of a matrix in a zig-zag pattern. For example, print

  ```
  1 2 3
  4 5 6
  7 8 9
  ```

  as `1 2 4 3 5 7 6 8 9`.

  (*Source:* Quora)

- Print the elements of a matrix in counter-clockwise spiral order. For example, print

```
1 2 3
4 5 6
7 8 9
```

as `1 4 7 8 9 6 3 2 5`.

(*Source:* TripleByte)

Reservoir sampling

- Choose a random number from a continuous stream of unknown length. Prove that the choice is uniformly random.

  (*Source:* Google)

- Choose *k* random numbers from a continuous stream of unknown length. Prove that the elements are chosen uniformly randomly.

  (*Source:* Yelp, 120 Data Science Interview Questions, Indeed)

Searching a tree

- Find the shortest path through a maze.

  (*Source:* Ancestry.com)

Sorting and selecting

Sorting

- Merge two arrays of sorted integers into a single sorted array.

  (*Source:* Quora, Workday, Facebook, Indeed, LinkedIn)

- Write a sorting algorithm.

  (*Source:* Facebook)

- Implement merge sort.

  (*Source:* StoryCloud)

- *External merge sort:* Implement a sorting algorithm for arrays too large to fit completely in memory.

  (*Source:* Guavus, StoryCloud)

- Find all the anagrams in an array of words. (Two words are anagrams of each other if one is a rearrangement of the letters in the other.)

  (*Source:* Quantifind, Schlumberger)

Selecting

- *Binary search:* Find the position of a given element in a sorted array.

  (*Source:* J.P. Morgan)

- Find the greatest *k* elements of an array.

  (*Source:* BloomReach, Guavus, Twitter)

- Find the *k* most frequent elements of an array.

  (*Source:* LinkedIn, 120 Data Science Interview Questions, Chegg)

- Given a large sorted array of numbers and a 4-element sorted array of numbers, find the positions of the latter in the former.

  (*Source:* Amazon)

- Find the median of a dataset too large to fit completely in memory.

  (*Source:* Twitter)

Dynamic programming

- You are trying to rob houses lined up on a street, each of which has a positive amount of money. Once you rob a house, you cannot rob the houses adjacent to that house. What is the maximum amount of money you can rob?

  (*Source:* Facebook)

- *Subset sum problem:* Find all the subsets of an array of integers which sum to *k*.

  (*Source:* Quantifind)

- *Maximum subarray problem:* Find the contiguous subset of an array of integers with the greatest sum.

  (*Source:* Guardian Analytics)

- *Word Break:* Given an array of valid words, split a string into an array of valid words or return `False` if doing so is impossible. (For example, if `"apple"` and `"pie"` are valid words, then `"applepie"` should be split into `["apple", "pie"]`)

(*Source:* LinkedIn)

- *Manacher's algorithm:* Determine all palindromic substrings of a string.

  (*Source:* Facebook)

Index manipulation

- Given two sorted arrays of integers, find one number from each array such that their sum is as close as possible to k.

  (*Source:* Adobe)

- Find the shortest substring of a string with at least k distinct characters.

  (*Source:* Google)

Parallelization

- Describe how you would use MapReduce to parallelize a procedure.

  (*Source:* Placed, Capital One, Vodafone)

- When does MapReduce work well and when does it work poorly?

  (*Source:* Walmart, 120 Data Science Interview Questions)

- Use MapReduce to determine the frequency of each word in a string.

  (*Source:* Schlumberger)

- Use MapReduce to join two datasets on a common key. How would you do so if the key is not unique?

  (*Source:* Kabbage)

- Use MapReduce to find all prime numbers under 1 billion.

  (*Source:* Groupon)

- Use MapReduce to compute the product of n different matrices.

  (*Source:* Samsung Research America)

Data structures

- What is a hash?

  (*Source:* Kahuna)

- Describe a hash table.

  (*Source:* Amazon)

- What is the difference between a linked list and an array?

  (*Source:* Amazon)

- What is the difference between a stack and a queue?

  (*Source:* Amazon)

- What is the difference between a hash table and a hashmap?

  (*Source:* DRW)

- Implement a sparse matrix.

  (*Source:* Facebook)

- Reverse a singly-linked list.

  (*Source:* Kahuna, Walmart, Jump Trading)

- Determine if a cycle exists in a singly-linked list.

  (*Source:* SpaceX)

- Find the 2nd largest element in a binary search tree.

  (*Source:* LinkedIn)

Other

- Determine if an array of integers contains two numbers which sum to k.

  (*Source:* Facebook)

- Find the intersection of two sets of integers.

  (*Source:* J.P. Morgan)

- Generate the power set of a set. (The power set is the set of all subsets.)

  (*Source:* Twitter, Jump Trading)

- *Dijkstra's algorithm:* Calculate the distance between two nodes in a graph.

  (*Source:* Automatic)

- Each day, you can buy or sell a single unit of stock (but not both), and you can hold multiple units of stock simultaneously. Given a sequence of stock prices, determine the maximum possible profit.

  (*Source:* Upstart)

- Given the coordinates of the corners of two rectangles, determine if they overlap and, if so, the area of the overlapping region.

  (*Source:* Rifiniti)

- Given an even number of points on a 2-dimensional plane, find a line with the same number of points on either side.

  (*Source:* Bay Sensors)

- Write pseudocode to determine if a given Sudoku board is valid.

  (*Source:* Wink)

- Generate random numbers.

  (*Source:* Magnetic)

- What is the computational complexity of finding the most frequent word in a document?

  (*Source:* Salesforce)

Language-specific

- *Python:* Flatten a nested list.

  (*Source:* OnDeck)

- *Python:* What data structures are implemented in `pandas`?

  (*Source:* Walmart)

- *Python:* What are the differences between iterators, generators, and list comprehensions?

  (*Source:* Walmart)

- *R:* What are some machine learning packages you have used?

  (*Source:* Civis Analytics)

- *R:* Determine the unique values of a variable's attribute.

  (*Source:* Nielsen)

- *R:* It is difficult to handle large amounts of data in R. What are some ways to work with large datasets without using distributed computing infrastructure (Hadoop, Spark, etc.)?

  (*Source:* Pyro Networks)

Probability

General

- There are *n* pieces of rope in a bucket. You reach in, grab two end pieces, and tie them together. What is the expected number of loops in the bucket?

  (*Source:* Natera)

- A bag has 5 black marbles and 1 white marble. You reach into the bag, take out a marble, and put it back into the bag 100 times. What is the probability of getting the white marble at least once?

  (*Source:* Microsoft)

- You are given 50 cards: 10 green cards, 10 red cards, 10 orange cards, 10 blue cards, and 10 yellow cards. The cards of each color are numbered from 1 to 10. If you randomly choose 2 cards, what is the probability that the chosen cards have different colors and different numbers?

  (*Source:* Facebook)

- There is a 60% chance of rain on Saturday and an 80% chance of rain on Sunday. Given that it rained over the weekend, what is the probability that it rained on Saturday?

  (*Source:* TransMarket Group)

- Box A has 12 red cards and 12 black cards; box B has 24 red cards and 24 black cards. You want to draw two cards with the same color. Which box do you draw from?

  (*Source:* Goldman Sachs, LendingClub)

- How would you select points uniformly randomly from within a circle?

  (*Source:* Lyft)

- You're about to get on a plane to Seattle, in which it's raining with prior probability p. You want to know if you should bring an umbrella. You call 3 random friends who live there and ask each independently if it's raining. Each of your friends has a $\frac{2}{3}$ chance of telling you the truth and $\frac{1}{3}$ chance of lying. If all 3 friends tell you that it's raining in Seattle, what's the probability that it's actually raining in Seattle?

  (*Sources:* Facebook)

- Two points are randomly placed on a line of length 1. What is the probability that the three segments created form a triangle?

  (*Source:* Facebook)

- Three ants are sitting at the three corners of an equilateral triangle. Each ant starts randomly picks a direction and starts to move along the edge of the triangle. What is the probability that none of the ants collide? What if there are n ants sitting at the corners of an n-gon?

  (*Source:* Facebook)

- 50% of people who receive a first interview also receive a second interview. 95% of people who feel they had a good second interview also felt they had a good first interview. 75% of people who *did not* get a second interview felt they had a good first interview. If you feel you had a good first interview, what is the probability you will receive a second interview?

  (*Source:* Amazon)

- A robot encounters a pile of coins spread out on the ground. It examines each coin and, if the coin is heads, turns it to tails with 100% probability. If the coin is tails, it turns it to heads with 50% probability. He examines each coin once per cycle. If you have the robot cycle through the coins a very large number of times, what will eventually happen to the proportion of heads and the proportion of tails?

  (*Source:* DRW)

- Bobo the amoeba has a 25%, 25%, and 50% chance of producing 0, 1, or 2 offspring respectively. Each of Bobo's descendants also have the same probabilities. What is the probability that Bobo's lineage dies out?

  (*Source:* 120 Data Science Interview Questions)

- *Monty Hall problem:* You are on a game show and can choose one of 3 doors. Behind one door is a car; behind the others, goats. You choose a door without opening it; the host opens one of the remaining two doors and reveals a goat, and asks if you want to switch to the other door. Is it to your advantage to switch?

  (*Source:* Hewlett Packard Enterprise)

- *Birthday paradox:* How many people must be gathered together in a room before you're reasonably certain that there is a greater than 50% probability that at least two people share the same birthday?

  (*Source:* Robert Bosch GmbH)

- Estimate the probability of a disease in a particular city given that the probability of the disease on a national level is very low and that 1,000 randomly selected people in the city do not have the disease.

  (*Source:* Amazon)

Flipping coins

Fair coins

- You flip a coin n times. What is the expected number of heads? (Consider $n = 4, 6$.)

  (*Source:* Jane Street)

- You flip a coin n times and get at least m heads. What is the expected number of heads? (Consider $(n, m) = (4, 2)$.)

  (*Source:* Jane Street)

- You flip a fair coin n times. What is the probability of getting exactly m heads? (Consider $(n, m) = (4, 2), (5, 2)$.)

  (*Source:* Jane Street)

- You flip a fair coin n times. What is the probability of getting m or more heads? (Consider $(n, m) = (2, 2), (7, 4)$.)

  (*Source:* Jane Street)

- You flip 100 fair coins. What is the probability of getting an even number of heads?

  (*Source:* Jane Street)

- How many fair coins would you have to toss such that the probability of getting exactly two heads is 50%?

  (*Source:* Five Rings Capital)

- You flip a fair coin n times. What is the expected product of the number of heads and the number of tails? (Consider $n = 10, 12$.)

  (*Source:* Jane Street)

- You flip a fair coin until either of the sequences HHT or HTT appears. Is one more likely to appear first? If so, which one and with what probability?

  (*Source:* Jane Street)

- You flip a fair coin until either of the sequences HHT or TTH appears. Is one more likely to appear first? If so, which one and with what probability?

  (*Source:* Goldman Sachs)

- You flip a fair coin 1,000 times. What is the probability that you get more than 550 heads? What theorem would you use to describe probabilities of rare events as the number of coin flips goes to infinity?

  (*Source:* D. E. Shaw)

Biased coins

- Given an unfair coin with more than a 50% chance of heads, how do you simulate a fair coin?

  (*Source:* Jane Street, 120 Data Science Interview Questions)

- You have a biased coin with a 51% chance of heads. Using this coin, how can you simulate a biased coin with a $\frac{1}{16}$ chance of heads?

  (*Source:* WorldQuant)

- You have 99 fair coins and 1 double-headed coin. You pick a coin, flip it k times, and get k heads. What is the probability that you picked a fair coin?

  (*Source:* Belvedere Trading, Facebook, Heard on the Street, Hudson River Trading, Jane Street)

- You have one fair coin and one biased coin with a 75% chance of heads. You randomly pick a coin, flip it twice, and get heads both times. What is the probability that you picked the fair coin?

  (*Source:* 120 Data Science Interview Questions)

- You have a 0.1% chance of picking up a coin with both heads and a 99.9% chance of picking up a fair coin. You pick a coin, flip it 10 times, and get 10 heads. What is the probability that you picked a fair coin?

  (*Source:* 120 Data Science Interview Questions)

- You randomly choose a fair coin or a double-headed coin and flip it. If the coin comes up heads, what is the probability that the coin has tails on the other side?

  (*Source:* Jane Street)

- You randomly choose a fair coin or a double-headed coin and flip it. If the coin comes up heads and I flip it again, what is the probability that it comes up heads again?

  (*Source:* Five Rings Capital)

- You choose either coin A, which comes up heads 70% of the time, or coin B, which comes up heads 60% of the times. You flip the coin 10 times and get 7 heads. What is the probability that you chose coin A?

  (*Source:* Facebook, Apple)

- You flip a fair coin 1,000 times and get 550 heads. Do you think the coin is biased? What if you only flipped the coin 10 times and got 6 heads?

  (*Source*: Google)

Consecutive heads

- What is the expected number of tosses of a fair coin until you get n consecutive heads? (Consider $n = 2, 3$.)

  (*Source:* WorldQuant, Jane Street, Goldman Sachs)

- You flip a fair coin 3 times. What is the expected number of consecutive heads?

  (*Source:* Five Rings Capital)

- You flip a fair coin n times. What is the probability of getting 2 consecutive heads?

  (*Source:* TransMarket Group)

- You flip a fair coin 5 times. What is the probability of getting at least 3 consecutive heads or 3 consecutive tails?

  (*Source:* TransMarket Group)

- How many times do you have to flip a fair coin to have at least a 50% chance of having gotten two consecutive heads?

  (*Source:* Five Rings Capital)

Games and profit

- You flip a fair coin repeatedly until you get heads. If you get heads on the nth flip, I pay you $2n - 1$ dollars. How much would you pay me to play this game?

  (*Source:* 120 Data Science Interview Questions)

- You flip 4 fair coins. You win \$1 per heads and can re-flip a single coin. If you play rationally, what is your expected profit?

  (*Source:* Jane Street)

- You flip 4 fair coins and win an amount proportional to the number of heads squared. What is your expected profit?

  (*Source:* Jane Street)

- A has 6 points and B has 4 points. They repeatedly flip fair coins; A gets 1 point for each heads and B gets 1 point for each tails. What is the probability that A reaches 10 points first?

  (*Source:* Five Rings Capital)

- One of us flips a fair coin. If the first player gets heads, the second player pays him \$30. If the first player gets tails, the coin goes to the second player, and if the second player gets heads, the first player pays him $30. How much should you pay to go first?

(*Source:* Jane Street)

- I flip a fair coin 5 times and you flip a fair coin 4 times. What is the probability that you will get at least as many heads as I do?

  (*Source:* Jane Street)

- We both flip a fair coin 3 times. If we have the same number of heads, I pay you $2; otherwise, you pay me $1. What is your expected profit?

  (*Source:* Susquehanna International Group)

- We continually flip fair coins. A counter starts at 0; it increases by 1 for each heads and decreases by 1 for each tails. You win if the counter reaches 20 and I win if the counter reaches -10. What is your probability of winning?

  (*Source:* Jane Street)

- I flip 100 fair coins and you need to guess the sequence of outcomes. You are allowed to ask one yes/no question. What do you ask to maximize the probability of guessing the sequence?

  (*Source:* Jane Street)

Other

- How can you simulate flips of a fair coin from samples of an arbitrary continuous probability distribution over $(0, 1)$? Assume that you know the form of the distribution.

  (*Source:* Jane Street)

- How can you simulate a die roll with flips of a fair coin?

  (*Source:* Jane Street)

Rolling dice

All dice can be assumed to be fair and 6-sided unless otherwise specified.

- You roll a die n times. What is the probability of getting at least one 6? (Consider $n = 2, 4, 7$.)

  (*Source:* Consolidated Trading, Facebook, Verizon)

- You roll a die n times. What is the probability of getting a 6 at least m times? (Consider $(n, m) = (6, 1), (12, 2), (600, 100)$.)

  (*Source:* Capital One)

- What is the expected value of a single die roll?

  (*Source:* DRW, Heard on the Street, Belvedere Trading, Susquehanna International Group, Clover Health)

- Players A and B are playing a game. A has 8 stones and B has 6 stones. A rolls a die and takes a number of stones from B equal to their roll, and then B rolls a die and takes a number of stones from A equal to their roll. The player with more stones wins; if A and B are tied, they repeat the game until a winner is determined. What is the probability that B wins in n rounds?

  (*Source:* Facebook)

- You roll a die and win an amount equal to your roll. You can alternatively choose to roll the die again for a new value. How much would you pay to play this game? What if you could choose to reroll a second time?

  (*Source:* Facebook, Jane Street, Spot Trading, Yammer, DRW)

- You roll a 12-sided die and win an amount equal to your roll. You can alternatively choose to roll two 6-sided dice instead and win the sum of the two dice. How much would you pay to play this game?

  (*Source:* Jane Street)

- You roll a die until the sum of the rolls is greater than 13. What number are you the most likely to stop on?

  (*Source:* Jane Street)

- What is the expected product of a 6-sided die and an 8-sided die?

  (*Source:* Jane Street)

- What is the probability that the sum of two dice is greater than 7?

  (*Source:* Jane Street)

- You repeatedly roll a die. What is the expected number of rolls until you get two consecutive rolls of 6?

  (*Source:* Jane Street)

- You roll a 30-sided die until the sum of the rolls is at least 300. What is the most probable final sum?

  (*Source:* Jane Street)

- You roll a 10-sided die and a 20-sided die. What is the probability of rolling a higher number on the 10-sided die?

  (*Source:* Jane Street)

- You roll an n-sided die and can choose to win an amount equal to your roll or reroll for a cost of $1/n$. Playing optimally, what is your expected profit?

  (*Source:* Jane Street)

- What is the expected value of the maximum of two dice rolls?

  (*Source:* Jane Street, Morgan Stanley)

- What is the expected value of the absolute difference between two rolls of a 30-sided dice?

  (*Source:* Jane Street)

- You roll a 12-sided die until the sum of the rolls is an odd number. What is the most probable final sum?

  (*Source:* Jane Street)

- You roll 3 dice. What is the probability that they sum to 10?

  (*Source:* Jane Street)

- What is the expected value of the absolute difference between two rolls of a die?

  (*Source:* Jane Street)

- You roll two dice. Given that one of them is 6, what is the probability that they are both 6?

  (*Source:* Jane Street)

- How can you use a 10-sided die to simulate a 12-sided die?

  (*Source:* Jane Street)

- You roll an 8-sided die and win an amount equal to your roll. You can alternatively choose to reroll with a 12-sided die. How much should you pay to play this game?

  (*Source:* Jane Street)

- Is it possible to change the numbers on the sides of two dice such that the probability distribution of their sum remains unchanged?

  (*Source:* Facebook)

- We can play a game where one player chooses a number from 1 to n and the other player chooses a different number from 1 to n. Next, we roll a 30-sided die and the person who chose the number closer to the roll wins an amount equal to the roll. What is the optimal strategy–do you want to go first or second, and how do you choose your number? What is the expected profit assuming perfect play?

(*Source:* Jane Street)

- How can you use a 6-sided die to simulate a 7-sided die?

  (*Source:* Morgan Stanley, 120 Data Science Interview Questions)

Decks of cards

Assume that decks are standard (52 cards) and shuffled unless otherwise specified.

The **suit** of a card can be either spades, clubs, hearts, or diamonds, and the **rank** of a card refers to its number or monarchic designation (2, 5, queen, etc.). Hearts and diamonds are colored red, whereas spades and clubs are colored black. A **pair** of cards denotes 2 cards of the same rank, and a **straight** is a hand of 5 cards of the same suit forming a continuous sequence of ranks.

- You choose 1 card from a deck without replacement. If you choose another card, what is the probability of getting one with a different color? What about the probability of getting one with a different suit?

  (*Source:* Facebook)

- You choose 2 cards from a deck. What is the probability of getting a pair.

  (*Source:* The Advisory Board Company, Facebook)

- You choose 3 cards from a deck. What is the probability of getting a pair?

  (*Source:* The Advisory Board Company)

- You choose 5 cards from a desk. What is the probability of getting a straight?

  (*Source:* Facebook)

General mathematics

- How many trailing zeroes are at the end of 100!?

  (*Source:* Facebook)

- Given a fleet of 50 trucks, each with a full fuel tank and a range of 100 miles, how far can you deliver a payload? You can transfer the payload from truck to truck, and you can transfer fuel from truck to truck.

  (*Source:* Facebook)

- Prove that the square root of two is irrational.

  (*Source:* Goldman Sachs)

- How can you cut a circular cake into 8 equal pieces using only 3 cuts?

  (*Source:* Google)

- You have two identical eggs and a building with 100 floors. Dropping an egg from the nth floor will shatter the egg if n is sufficiently great, *i.e.*, if $n \geq N$ for some unknown floor number N, but will have no effect on the egg if n is *not* sufficiently great, *i.e.*, $n < N$. How would you determine N with as few egg drops as possible?

  (*Source:* Apple, Facebook, WorldQuant, AdHarmonics)

Comparisons

- You have 8 balls, one of which is slightly heavier than the others, and a sensitive balance. What is the lowest number of weighings you need to do to find the heavier ball? What if there are 12 balls?

  (*Source:* Goldman Sachs, WorldQuant)

- You have 10 bottles, 9 of which contain 1 gram coins and 1 of which contains 1.1 gram coins. How can you use a single measurement on an electronic scale to determine the bottle containing the heavier coins?

  (*Source:* LatentView Analytics)

- You have 25 horses. You can race any 5 of them and get back the order in which they finished. How many races do you need to find the 3 fastest horses?

  (*Source:* Facebook, Jane Street, TransMarket Group, Trillium Trading, Barclays)

Fermi estimates

- How heavy is Mount Everest?

  (*Source:* Jane Street)

- How many books have ever been published?

  (*Source:* Jane Street)

- How many windows are there at Harvard?

  (*Source:* Jane Street)

- How many tons does the ocean weigh?

  (*Source:* Jane Street)

- What is the mass of the Earth in kilograms?

  (*Source:* Jane Street)

- How many paintings are in New York City?

  (*Source:* Jane Street)

- How much paper can be made from a tree?

  (*Source:* Jane Street)

- How many people apply to Google each year?

  (*Source:* Google)

- How many people are using Facebook in California at 1:30 PM on Monday?

  (*Source:* Facebook)

- How many windows are in the Empire State Building?

  (*Source:* Five Rings Capital)

- What is the largest temperature difference across 24 hours at any location in the United States?

  (*Source:* Five Rings Capital)

- How many dentists are there in the United States?

  (*Source:* Facebook)

- How many gas stations are there in the United States?

  (*Source:* Belvedere Trading)

- What is the total length of all roads in San Francisco?

  (*Source:* eBay)

- How many McDonald's locations are in the United States?

  (*Source:* Facebook)
- How many Big Macs does McDonald's sell in the United States each year?

  (*Source:* Facebook)
- How many piano tuners in Seattle?

  (*Source:* Facebook)
- How many baseballs can fit inside a football stadium?

  (*Source:* Facebook)
- How many airports are in the United States?

  (*Source:* Facebook)
- How many trees are in the United States?

  (*Source:* Goldman Sachs)
- How many square feet of pizza are eaten in the United States per year?

  (*Source:* Goldman Sachs)
- How many birthday posts occur on Facebook each day?

  (*Source:* Facebook)

Logic

- A disc is spinning either clockwise or counterclockwise and you have a set of pins. How would you use the pins to determine which way the disc is rotating?

  (*Source:* Google)
- You have a glass of water shaped like a perfect cylinder. You cannot measure the water, measure the glass, or dip anything into the water. How do you determine if the glass is less than half full, exactly half full, or more than half full?

  (*Source:* LatentView Analytics)
- You have 100 coins lying flat on a table, each with a heads side and a tails side. 10 of the coins have heads facing up and 90 of the coins have tails facing up. You cannot see the coins, feel their faces, or use any other method to determine which side of a coin is facing up. How can you split the coins into two piles so that each pile has the same number of heads?

  (*Source:* Apple)

- There are three boxes in front of you. One box contains apples, another box contains oranges, and the remaining box contains some of both. The boxes have been incorrectly labeled such that no label identifies the actual contents of its box. You can open just one box and take out one piece of fruit without looking inside the box. How can you label all of the boxes correctly?

  (*Source:* Apple)

Communication

- Tell me about your favorite machine learning model.

  (*Source:* KPMG, MetLife, Microsoft, 120 Data Science Interview Questions)

- What experience do you have working with big datasets?

  (*Source:* Nielsen, UPMC)

- How would you persuade a client of the usefulness of your analysis?

  (*Source:* Civis Analytics)

- Given an ambiguous data set and corresponding business problem, how would you generally go about solving it, *i.e.*, what steps would I first take?

  (*Source:* Civis Analytics)

- How would you explain regression to an 8-year-old? To a nontechnical businessman?

  (*Source:* Belvedere Trading, Accenture)

- How would you explain recursion to someone who doesn't know computer science?

  (*Source:* 1010data)

- How would you approach a problem if asked to use a technique you haven't heard of before?

  (*Source:* RootMetrics)

- If you could get any data for a personal data science project, what would you do?

  (*Source:* Civis Analytics)

- What is one of the world's biggest problems? How would you help solve it using big data?

  (*Source:* Civis Analytics)

Uncommon topics

Image processing

- What is SIFT (scale-invariant feature transform)?

  (*Source:* Opera Solutions)

- How would you determine the similarity of two images? How would you extend your algorithm to process a large dataset of images?

  (*Source:* ZEFR)

- How would you determine the shoe size of a pair of heels from pictures alone?

  (*Source:* Neiman Marcus)

Estimation methods

- What is maximum likelihood estimation? Could there be any case where the maximum likelihood estimate does not exist?

  (*Source:* 120 Data Science Interview Questions)

- What are the differences between MAP (maximum a posteriori) estimation, MOM (method of moments) estimation, and MLE (maximum likelihood estimation)? In which cases would you want to use each one?

  (*Source:* 120 Data Science Interview Questions)

- What is unbiasedness as a property of an estimator? Is this always a desirable property when performing statistical inference? What about in data analysis or predictive modeling?

  (*Source:* 120 Data Science Interview Questions)

- Given n samples from a uniform distribution over $[0, d]$, how would you estimate d?

  (*Source:* Spotify)

- What is the maximum likelihood estimate of the parameter λ of an exponential distribution?

  (*Source:* Tesla Motors)

Advanced data science

- Why are rectified linear units good activation functions?

  (*Source:* Netflix)

- Why is $L^{1/2}$ regularization not used?

  (*Source:* Netflix)

- How do you solve the L^2-regularized regression problem to obtain a closed-form expression for the coefficient estimates?

  (*Source:* Microsoft)

- How are hidden Markov models and probabilistic graphical models related? What are some common use cases of either?

  (*Source:* AgilOne)

- Describe a Monte Carlo Markov Chain process.

  (*Source:* Philips)

- Describe linear programming and constraint satisfaction problems.

  (*Source:* YP)

- List 3 types of unit root tests in addition to the augmented Dickey–Fuller test.

  (*Source:* KPMG)

Advanced algorithms

- Find the value closest to k in an array of floating-point numbers.

  (*Source:* Amazon)

- Find the mean, median, and mode for a continuous stream of 1 million numbers.

  (*Source:* C3 Energy)

- Convert an unbalanced binary tree to an AVL tree.

  (*Source:* KPMG)

- Build a data structure for efficient computation of arithmetic expressions.

  (*Source:* Commonwealth Computer Research)

- How can you quickly compute the inverse of a matrix?

  (*Source:* Microsoft)

Other

- How would you build a distributed, cloud-based machine learning system (like BigML)?

  (*Source:* NTT)

- What are the ACID properties of database transactions?

  (*Source:* Accenture)

- Prove that the Pearson correlation coefficient falls between -1 and 1.

  (*Source:* FINRA)

- How would you design a web crawler?

  (*Source:* GoDaddy)

- You are given a list of coin denominations. How many ways are there to split a dollar into change?

  (*Source:* Capital One)

- How could you model 1 cm diameter raindrops falling on a 1 m wide sidewalk and determine how long it takes for the sidewalk to be completely wet?

  (*Source:* Google)

- What would you do if you were a traffic sign?

  (*Source:* LatentView Analytics)

Unorganized questions

Programming

AirBnB, Factual – Given a dictionary, and a matrix of letters, find all the words in the matrix that are in the dictionary.

Bay Sensors – Develop a solution for fusing sensors that are sensing the same goal. Sensors are time unsynchronized and conflict at times. You have some truth data too.

Capital One – implement a function to calculate matrix sum

Capital One – You have a 300GB dat file, and you want to run a computation on the third column. How do you do that? - This is to check whether you know how to use unix commands that work off disk rather than in memory.

Facebook – Write C++ code to copy a graph

NantMobile – word fuzzy matching

Accenture – Write a function to calculate the acceleration of a car moving north for 5 minutes.

LinkedIn – Segment a long string into a set of valid words using a dictionary. Return false if the string cannot be segmented. What is the complexity of your solution?

6sense – Find all possible ways to evaluate the multiplication of 4 numbers in a N by N matrix.

Kabbage – MapReduce: Join two data files (customers, sales) and report top 10 performers.

Capital One – Map reduce a list of companies and their revenues sorted by industry.

MaxPoint – What is the significance of a reduceByKey in pySpark - when would you use it?"

Other

Yammer – German tank problem, breaking a 4-digit code with different prior information etc.

American Express – Suppose that American Express has 1 million card members along with their transaction details. They also have 10,000 restaurants and 1000 food coupons. Suggest a method which can be used to pass the food coupons to users given that some users have already received the food coupons so far.

Apple – Given an iTunes type of app that pulls down lots of images that get stale over time, what strategy would you use to flush disused images over time?

Bitly – We're interested in tracking users that may appear in our systems via many device ID's. That is, someone may click a Bitly link from a mobile device, and at another time, that same user may click another Bitly link from a laptop at home. We only have these device ID's in addition to sites visited and other metadata, and we would like to attribute these actions to the same user across devices. How would you approach this problem?

C3 Energy – Re-derive linear regression optimization formulation in closed form for a spline function (flat then gradient) rather than y=Ax + B. Use matrix notation and assume tall-skinny problem. Talk about the complexity issues with closed form approach. Then extend to include regularization. Write an algo to merge multiedge polygons together given the edges and shared edges.

Civis Analytics – You have been tasked with creating a prediction for the outcome of the 2014 congressional elections. You are responsible for building your own dataset for this process.

eBay – Case study, Ebay has to identify the cameras from the other junk like tripods, cables and batteries what is the approach? Data looks like Title, description of the product, Price, Image,. . . etc

Facebook – How would one develop the Like, Love, Sad feature?

Facebook – There are two types of cars A and B. The number of people in US who use Aand B are the same. They drive the same distances each month.Now there are two new technologies, X and Y (of equal cost).If apply X, mpg of A would increase from 50 mpg to 75 mpg;If apply Y, mpg of B would increase from 10 mpg to 11 mpg.The goal is to decrease the dependence on foreign oil, or to decrease theconsumption of gasoline. Question: which technology would you apply? Follow up question: after applying the technology of your choice, assumethere's money available for research on new technology, which car would youchoose to conduct research on?"

Facebook – In Mexico, if you take the mean and the median age, which one will be higher and why? Given a table that each day shows who was active in the system and a table that tracks ongoing user status, write a procedure that will take each day's active table and pass it into the ongoing daily tracking table.Possible states are:* user stayed (yesterday yes, today yes)* user churned (yesterday yes, today no)* user revived (yesterday no, today yes)* user new (yesterday null, today yes)Note: you'll want to spot and account for the undefined state.

Facebook – How would you bring a metric to product X? Products at Facebook could be as large as News Feed or Ads, and as small as Pokes or Socrates (see below). Example 1: How would you assess the health of Facebook's News Feed? (Define health ?) Example 2: Facebook's Socrates is a box displayed under the Profile Picture of a user that prompts the user to answer questions about themselves, such as favorite movies, books, etc. Given the data about how users have answered questions in the past, design the best algorithm to present the next question that they will answer.

LinkedIn – What are the factors used to produce People You May Know data product on LinkedIn?

Netflix – How do you measure and compare models? How should we approach to attribution modeling to measure marketing efffectiveness?

Nielsen – What are some metrics that are more appropriate for use in a syndicated product than one ordered by a particular client?

Quora – What metric would you use to measure a search tool bar change?

Soleo – how I would model the waiting time for a local restaurant and what possible statistical models and strategies I could bring to bear on the problem.

Spotify – How would you detect anomalous behavior? How would you generate profiles of users?

Stitch Fix – How might you determine if the size of a specific item is a significant feature in determining the customer's calculated propensity to buy it?

Uber – What algorithm you would use to solve driver accepting a requesting?Which supervised algorithm you would pick to solve the above problem?How do you compare the results of the algorithm?

Uber – optimization of pickup time, how to evaluate the success of a new feature

Uber – how to optimize surge pricing given data set

Uber – Will uber cause city congestion?

Uber – How would you find the words that became obsolete in English language between 16th and 17th century? You may use a search engine.

Walmart – A stranger uses a search engine to find something and you do not know anything about the person. How will you design an algorithm to determine what the stranger is looking for just after he/she types few characters in the search box?

Walmart – You have a clickthrough rate/impressions on two products, the first has a 1/100 CTR and the second has 100/10000 - how do you reconcile these rates and evaluate the performance of these two products?

120 Data Science book questions (unsorted)

What could be some issues if the distribution of the test data is significantly different than the distribution of the training data?

What are some ways I can make my model more robust to outliers?

What are some differences you would expect in a model that minimizes squared error, versus a model that minimizes absolute error? In which cases would each error metric be appropriate?

What error metric would you use to evaluate how good a binary classifier is? What if the classes are imbalanced? What if there are more than 2 groups?

What are various ways to predict a binary response variable? Can you compare two of them and tell me when one would be more appropriate? What's the difference between these? (SVM, Logistic Regression, Naive Bayes, Decision Tree, etc.)

Given training data on tweets and their retweets, how would you predict the number of retweets of a given tweet after 7 days after only observing 2 days worth of data?

How could you collect and analyze data to use social media to predict the weather?

How would you construct a feed to show relevant content for a site that involves user interactions with items?

How would you design the people you may know feature on LinkedIn or Facebook?

How would you predict who someone may want to send a Snapchat or Gmail to?

How would you suggest to a franchise where to open a new store?

In a search engine, given partial data on what the user has typed, how would you predict the user's eventual search query?

Given a database of all previous alumni donations to your university, how would you predict which recent alumni are most likely to donate?

You're Uber and you want to design a heatmap to recommend to drivers where to wait for a passenger. How would you approach this?

You want to run a regression to predict the probability of a flight delay, but there are flights with delays of up to 12 hours that are really messing up your model. How can you address this?

Write a function to calculate all possible assignment vectors of 2n users, where n users are assigned to group 0 (control), and n users are assigned to group 1 (treatment).

Program an algorithm to find the best approximate solution to the knapsack problem in a given time.

Program an algorithm to find the best approximate solution to the travelling salesman problem in a given time.

Given a list of numbers, can you return the outliers?

What are the different types of joins? What are the differences between them?

Why might a join on a subquery be slow? How might you speed it up?

Given a COURSES table with columns course_id and course_name, a FACULTY table with columns faculty_id and faculty_name, and a COURSE_FACULTY table with columns faculty_id and course_id, how would you return a list of faculty who teach a course given the name of a course?

Given a IMPRESSIONS table with ad_id, click (an indicator that the ad was clicked), and date, write a SQL query that will tell me the click-through-rate of each ad by month.

Write a query that returns the name of each department and a count of the number of employees in each: EMPLOYEES containing: Emp_ID (Primary key) and Emp_Name, EMPLOYEE_DEPT containing: Emp_ID (Foreign key) and Dept_ID (Foreign key), DEPTS containing: Dept_ID (Primary key) and Dept_Name

In any 15-minute interval, there is a 20% probability that you will see at least one shooting star. What is the probability that you see at least one shooting star in the period of an hour?

You have an 50-50 mixture of two normal distributions with the same standard deviation. How far apart do the means need to be in order for this distribution to be bimodal?

A certain couple tells you that they have two children, at least one of which is a girl. What is the probability that they have two girls?

You have a group of couples that decide to have children until they have their first girl, after which they stop having children. What is the expected gender ratio of the children that are born? What is the expected number of children each couple will have?

How many ways can you split 12 people into 3 teams of 4?

Your hash function assigns each object to a number between 1:10, each with equal probability. With 10 objects, what is the probability of a hash collision? What is the expected number of hash collisions? What is the expected number of hashes that are unused.

You call 2 UberX's and 3 Lyfts. If the time that each takes to reach you is IID, what is the probability that all the Lyfts arrive first? What is the probability that all the UberX's arrive first?

I write a program should print out all the numbers from 1 to 300, but prints out Fizz instead if the number is divisible by 3, Buzz instead if the number is divisible by 5, and FizzBuzz if the number is divisible by 3 and 5. What is the total number of numbers that is either Fizzed, Buzzed, or FizzBuzzed?

On a dating site, users can select 5 out of 24 adjectives to describe themselves. A match is declared between two users if they match on at least 4 adjectives. If Alice and Bob randomly pick adjectives, what is the probability that they form a match?

A lazy high school senior types up application and envelopes to n different colleges, but puts the applications randomly into the envelopes. What is the expected number of applications that went to the right college?

I have two different experiments that both change the sign-up button to my website. I want to test them at the same time. What kinds of things should I keep in mind?

What is the difference between type-1 and type-2 error?

You are AirBnB and you want to test the hypothesis that a greater number of photographs increases the chances that a buyer selects the listing. How would you test this hypothesis?

How would you design an experiment to determine the impact of latency on user engagement?

Is more data always better?

What are advantages of plotting your data before performing analysis?

How can you make sure that you don't analyze something that ends up meaningless?

What is the role of trial and error in data analysis? What is the the role of making a hypothesis before diving in?

How can you determine which features are the most important in your model?

How do you deal with some of your predictors being missing?

You have several variables that are positively correlated with your response, and you think combining all of the variables could give you a good prediction of your response. However, you see that in the multiple linear regression, one of the weights on the predictors is negative. What could be the issue?

Let's say you're given an unfeasible amount of predictors in a predictive modeling task. What are some ways to make the prediction more feasible?

Now you have a feasible amount of predictors, but you're fairly sure that you don't need all of them. How would you perform feature selection on the dataset?

Your linear regression didn't run and communicates that there are an infinite number of best estimates for the regression coefficients. What could be wrong?

You run your regression on different subsets of your data, and find that in each subset, the beta value for a certain variable varies wildly. What could be the issue here?

What is the main idea behind ensemble learning? If I had many different models that predicted the same response variable, what might I want to do to incorporate all of the models? Would you expect this to perform better than an individual model or worse?

Given that you have wifi data in your office, how would you determine which rooms and areas are underutilized and overutilized?

How could you use GPS data from a car to determine the quality of a driver?

Given accelerometer, altitude, and fuel usage data from a car, how would you determine the optimum acceleration pattern to drive over hills?

Given position data of NBA players in a season's games, how would you evaluate a basketball player's defensive ability?

How would you quantify the influence of a Twitter user?

Given location data of golf balls in games, how would construct a model that can advise golfers where to aim?

You have 100 mathletes and 100 math problems. Each mathlete gets to choose 10 problems to solve. Given data on who got what problem correct, how would you rank the problems in terms of difficulty?

You have 5000 people that rank 10 sushis in terms of saltiness. How would you aggregate this data to estimate the true saltiness rank in each sushi?

Given data on congressional bills and which congressional representatives co-sponsored the bills, how would you determine which other representatives are most similar to yours in voting behavior? How would you evaluate who is the most liberal? Most republican? Most bipartisan?

How would you come up with an algorithm to detect plagiarism in online content?

You have data on all purchases of customers at a grocery store. Describe to me how you would program an algorithm that would cluster the customers into groups. How would you determine the appropriate number of clusters to include?

Let's say you're building the recommended music engine at Spotify to recommend people music based on past listening history. How would you approach this problem?

How would you explain an A/B test to an engineer with no statistics background? A linear regression?

How would you explain a confidence interval to an engineer with no statistics background? What does 95% confidence mean?

How would you explain to a group of senior executives why data is important?

Tell me about a data project that you've done with a team. What did you add to the group?

Tell me about a dataset that you've analyzed. What techniques did you find helpful and which ones didn't work?

How could you help the generate public understanding towards the importance of using data to generate insights?

How would you convince a government agency to release their data in a publicly accessible API?

I'm a local business owner operating a small restaurant. Convince me to switch my advertising budget from print to internet.

Let's say that you're are scheduling content for a content provider on television. How would you determine the best times to schedule content?

What kind of services would find churn (metric that tracks how many customers leave the service) helpful? How would you calculate churn?

Say that you are Netflix. How would you determine what original series you should invest in and create?

You are on the data science team at Uber and you are asked to start thinking about surge pricing. What would be the objectives of such a product and how would you start looking into this?

How would you measure the impact that sponsored stories on Facebook News Feed have on user engagement? How would you determine the optimum balance between sponsored stories and organic content on a user's News Feed?

Say you are working on Facebook News Feed. What would be some metrics that you think are important? How would you make the news each person gets more relevant?

You are tasked with improving the efficiency of a subway system. Where would you start?

You're a restaurant and are approached by Groupon to run a deal. What data would you ask from them in order to determine whether or not to do the deal?

Growth for total number of tweets sent has been slow this month. What data would you look at to determine the cause of the problem?

A certain metric is violating your expectations by going down or up more than you expect. How would you try to identify the cause of the change?

What would be good metrics of success for a product that offered in-app purchases? (Zynga, Angry Birds, other gaming apps)

What would be good metrics of success for a consumer product that relies heavily on engagement and interaction? (Snapchat, Pinterest, Facebook, etc.) A messaging product? (GroupMe, Hangouts, Snapchat, etc.)

What would be good metrics of success for an e-commerce product? (Etsy, Groupon, Birchbox, etc.) A subscription product? (Netflix, Birchbox, Hulu, etc.) Premium subscriptions? (OKCupid, LinkedIn, Spotify, etc.)

What would be good metrics of success for a productivity tool? (Evernote, Asana, Google Docs, etc.) A MOOC? (edX, Coursera, Udacity, etc.)

What would be good metrics of success for an advertising-driven consumer product? (Buzzfeed, YouTube, Google Search, etc.) A service-driven consumer product? (Uber, Flickr, Venmo, etc.)