120 DATA SCIENCE INTERVIEW QUESTIONS

COMPILED AND CREATED BY:
CARL SHAN, MAX SONG, HENRY WANG, AND WILLIAM CHEN

INTRODUCTION

This guide is meant to bridge the gap between the knowledge of a recent graduate and the skillset required to become a data scientist. By reading this guide and learning how to answer these questions, recent graduates will equip themselves with the expected knowledge and skills of a data scientist.

To help readers with these goals, we've gathered 120 interview questions in product metrics, programming and databases, probability, experimentation and inference, data analysis, and predictive modeling. These questions are all either real data science interview questions or inspired by real data science interview questions, and should help readers develop the skills needed to succeed in a data science role.

The role of a data scientist is highly malleable and company dependent. However, the general skillset needed is similar. Candidates need:

- Technical skills - data analysis and programming
- Business/product intuition - metrics and identifying opportunities for impact
- Communication ability - clarity in explaining findings and insights

To prepare for your interview, you may want to brush up by reviewing some probability, data analysis, SQL, coding, and experimental design. The questions in this guide should help you do so. The background of data science applicants varies wildly, so interviews may generally be more holistic and test your intuition, analytic, and communication abilities rather than focusing on specific technical concepts.

Prepare to discuss your past work involving analyzing large and complicated datasets, defending your approaches and communicating what you learned during your project. Expect questions involving how to measure "goodness" of a feature on the company's product, and be sure to approach these problems in a scientific and principled way. You have a good chance of getting a product metrics or experimentation question based on some actual questions the company is tackling at this time.

Check up on your company's engineering / data blog and see if anything's relevant. Be familiar with A/B testing and common metrics that companies similar to the one you are interviewing for may use. Brush up on your Python (especially iPython notebook) and/or R abilities to prepare for a potential live data analysis problem.

And finally, of course, follow the general interview advice. Prepare to elaborate on related projects from your resume. Be enthusiastic. Share your thoughts with your interviewer as you're going through a problem or doing a piece of analysis. And be sure to answer the question!

You have our best wishes!
Carl, Max, Henry, and William

Please feel free to reach out to us with questions, comments and suggestions at www.datasciencehandbook.me

CONTENTS

PREDICTIVE MODELING

1. *(Given a Dataset)* Analyze this dataset and give me a model that can predict this response variable.

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

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

4. 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?

5. 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?

6. 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.)

7. What is regularization and where might it be helpful? What is an example of using regularization in a model?

8. Why might it be preferable to include fewer predictors over many?

9. 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?

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

PRO TIP

If asked to predict a response variable during your interview, you should favor simpler models that run quickly and which you can easily explain. If the task is specifically a predictive modeling task, you should try to do, or at least mention, cross-validation as it really is the golden standard to evaluate the quality of one's model. Talk about and justify your approach while you're doing it, and leave some time to plot and visualize the data.

PREDICTIVE MODELING

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

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

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

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

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

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

17 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?

18 How would you build a model to predict a March Madness bracket?

19 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?

PRO TIP

Variations on ordinary linear regression can help address some problems that come up working with read data. LASSO helps when you have too many predictors by favoring weights of zero. Ridge regression can help with reducing the variance of your weights and predictions by shrinking the weights to 0. Least absolute deviations or robust linear regression can help when you have outliers. Logistic regression is used for binary outcomes, and Poisson regression can be used to model count data.

PROGRAMMING

1 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).

2 Given a list of tweets, determine the top 10 most used hashtags.

3 Program an algorithm to find the best approximate solution to the knapsack problem[1] in a given time.

4 Program an algorithm to find the best approximate solution to the travelling salesman problem[2] in a given time.

5 You have a stream of data coming in of size *n*, but you don't know what *n* is ahead of time. Write an algorithm that will take a random sample of *k* elements. Can you write one that takes *O(k)* space?

6 Write an algorithm that can calculate the square root of a number.

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

8 When can parallelism make your algorithms run faster? When could it make your algorithms run slower?

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

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

11 Describe the difference between primary keys and foreign keys in a SQL database.

PRO TIP

Traditional software engineering questions may show up in data science interviews. Expect those questions to be easier, less about systems, and more about your ability to manipulate data, read databases, and do simple programming tasks. Review your SQL and prepare to do common operations such as JOIN, GROUP BY, and COUNT. Review ways to manipulate data and strings (we suggest doing this in Python), so you can answer questions that involve sifting through numerical or string data.

1 See http://en.wikipedia.org/wiki/Knapsack_problem
2 See http://en.wikipedia.org/wiki/Travelling_salesman_problem

12 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?

13 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.

14 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`

PROBABILITY

1 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?

2 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?

3 How can you generate a random number between 1 - 7 with only a die?

4 How can you get a fair coin toss if someone hands you a coin that is weighted to come up heads more often than tails?

5 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?

6 Given draws from a normal distribution with known parameters, how can you simulate draws from a uniform distribution?

7 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?

8 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?

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

PRO TIP

Important concepts to review from an introductory probability class include the Law of Total Probability, Bayes' Rule, and Expectation. You can learn many of these topics (and important topics regarding hypothesis testing and inference) with intro-level courses in probability and inference.

10 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.

11 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?

12 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?

13 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?

14 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

15 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?

16 What's the expected number of coin flips until you get two heads in a row? What's the expected number of coin flips until you get two tails in a row?

PRO TIP

Many Bayes' Rule questions can be solved quickly with the odds form of Bayes Rule, which says that prior odds times likelihood ratio is the posterior odds. For problem 18, the prior odds is 999:1 and the likelihood ratio is 1/1024:1 (10 heads has a 1/1024 probability with a fair coin and a 1 probability with a biased coin), which means the posterior odds is about 1:1. For problem 19, the prior odds is 1:1 and the likelihood ratio is 1/4:9/16, so the posterior odds is 4:9.

17 Let's say we play a game where I keep flipping a coin until I get heads. If the first time I get heads is on the *n*th coin, then I pay you 2*n*-1 dollars. How much would you pay me to play this game?

18 You have two coins, one of which is fair and comes up heads with a probability 1/2, and the other which is biased and comes up heads with probability 3/4. You randomly pick coin and flip it twice, and get heads both times. What is the probability that you picked the fair coin?

19 You have a 0.1% chance of picking up a coin with both heads, and a 99.9% chance that you pick up a fair coin. You flip your coin and it comes up heads 10 times. What's the chance that you picked up the fair coin, given the information that you observed?

STATISTICAL INFERENCE

1. In an A/B test, how can you check if assignment to the various buckets was truly random?
2. What might be the benefits of running an A/A test, where you have two buckets who are exposed to the exact same product?
3. What would be the hazards of letting users sneak a peek at the other bucket in an A/B test?
4. What would be some issues if blogs decide to cover one of your experimental groups?
5. How would you conduct an A/B test on an opt-in feature?
6. How would you run an A/B test for many variants, say 20 or more?
7. How would you run an A/B test if the observations are extremely right-skewed?
8. 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?
9. What is a p-value? What is the difference between type-1 and type-2 error?
10. 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?
11. How would you design an experiment to determine the impact of latency on user engagement?
12. What is maximum likelihood estimation? Could there be any case where it doesn't exist?

PRO TIP

Proper A/B testing practices are often a common discussion, especially because it easily becomes more complicated than anticipated in practice. Multiple variants and metrics, simultaneous conflicting experiments, and improper randomization will complicate experiments. Most people do not have a formal academic background on experimental design.

STATISTICAL INFERENCE

13 What's the difference between a MAP, MOM, MLE estimator? In which cases would you want to use each?

14 What is a confidence interval and how do you interpret it?

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

PRO TIP

Important concepts to know include randomization, Simpson's paradox, and multiple comparisons. Advanced concepts to know that may impress interviewers includes alternatives to A/B testing (such as multi-armed bandit strategies), or alternatives to t-tests and z-tests (e.g. non-parametric methods, bootstrapping)

DATA ANALYSIS

1 *(Given a Dataset)* Analyze this dataset and tell me what you can learn from it.

2 What is R^2? What are some other metrics that could be better than R^2 and why?

3 What is the curse of dimensionality?

4 Is more data always better?

5 What are advantages of plotting your data before performing analysis?

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

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

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

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

10 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?

11 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?

12 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?

PRO TIP

Some concepts that are important in data analysis and common in the field, include overfitting, regression towards the mean, curse of dimensionality, importance of visualization, and inductive bias. These questions test your knowledge and experience with some of the hazards of blind data analysis and your ability to distinguish a significant result from a spurious one.

DATA ANALYSIS

13 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?

14 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?

15 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?

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

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

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

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

20 How would you quantify the influence of a Twitter user?

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

22 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?

PRO TIP

If asked to analyze a dataset during the interview, the interviewer is looking to learn about your comfort with your statistical software and your ability to generate interesting insights in a short period of time. We recommend making visualizations first, to show that you know good practices, prevent future missteps, and identify possible transformations needed. Be sure to talk about your procedure and anticipate questions about your approach.

DATA ANALYSIS

23 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?

24 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?

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

26 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?

27 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?

PRO TIP

Consider asking your interviewer how data scientists extract and wrangle data at the company, what tools the team uses to do its exploratory analysis, and how the company shares its findings internally. Most of the work is not the analysis. In fact, data scientists spend most of their time just getting, cleaning, and processing the data.

PRODUCT METRICS

1 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.)

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

3 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.)

4 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.)

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

6 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?

7 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?

8 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?

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

10 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?

PRO TIP

The best choices of engagement metrics are those that benefit both the company and the users while correlating highly with revenue. Pageviews and daily actives would be appropriate for an advertising-driven product, and metrics such as number of purchases or conversion rate would be appropriate for any product that sells services and other products.

PRODUCT METRICS

11 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?

12 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?

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

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

15 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?

PRO TIP

Interviewers are looking for candidates who have strong intuition about metrics for success. You should give many possible metrics, each a bit more specific than the previous. The interviewer may stop and ask you to elaborate or describe how you would collect or visualize the data. Prepare to justify why the metric is important, relevant, and measurable.

COMMUNICATION

1. Explain to me a technical concept related to the role that you're interviewing for.
2. Introduce me to something you're passionate about.
3. How would you explain an A/B test to an engineer with no statistics background? A linear regression?
4. How would you explain a confidence interval to an engineer with no statistics background? What does 95% confidence mean?
5. How would you explain to a group of senior executives why data is important?
6. Tell me about a data project that you've done with a team. What did you add to the group?
7. Tell me about a dataset that you've analyzed. What techniques did you find helpful and which ones didn't work?
8. What's your favorite algorithm? Can you explain it to me?
9. How could you help the generate public understanding towards the importance of using data to generate insights?
10. How would you convince a government agency to release their data in a publicly accessible API?
11. I'm a local business owner operating a small restaurant. Convince me to switch my advertising budget from print to internet.

PRO TIP

Interviews are about convincing the interviewer that you know what you're talking about. Naturally, you will gain more ability to do so with a better background in in the topics covered here. Practice teaching a concept, explaining one of your past projects, and discussing your techniques.

DATA SCIENCE HANDBOOK

Knowing and being able to answer these questions will help you succeed in the data science interview. But after landing that job, if you want to learn how to advance in your career as a data scientist, you should check out *The Data Science Handbook* — a curated collection of interviews containing advice and wisdom from some of top data scientists in the world.

You can get it at www.datasciencehandbook.me

LAYOUT DESIGN BY: BRITTANY CHENG