66 job interview questions for data scientists

- Posted by Vincent Granville on February 13, 2013 at 8:00pm
- View Blog

We are now at 91 questions. We've also added 50 new ones here, and started to provide answers to these questions here. These are mostly open-ended questions, to assess the technical horizontal knowledge of a senior candidate for a rather high level position, e.g. director.

1. What is the biggest data set that you processed, and how did you process it, what were the results?
2. Tell me two success stories about your analytic or computer science projects? How was lift (or success) measured?
3. What is: lift, KPI, robustness, model fitting, design of experiments, 80/20 rule?
4. What is: collaborative filtering, n-grams, map reduce, cosine distance?
5. How to optimize a web crawler to run much faster, extract better information, and better summarize data to produce cleaner databases?
6. How would you come up with a solution to identify plagiarism?
7. How to detect individual paid accounts shared by multiple users?
8. Should click data be handled in real time? Why? In which contexts?
9. What is better: good data or good models? And how do you define "good"? Is there a universal good model? Are there any models that are definitely not so good?
10. What is probabilistic merging (AKA fuzzy merging)? Is it easier to handle with SQL or other languages? Which languages would you choose for semi-structured text data reconciliation?
11. How do you handle missing data? What imputation techniques do you recommend?
12. What is your favorite programming language / vendor? why?
13. Tell me 3 things positive and 3 things negative about your favorite statistical software.
14. Compare SAS, R, Python, Perl
15. What is the curse of big data?
16. Have you been involved in database design and data modeling?
17. Have you been involved in dashboard creation and metric selection? What do you think about Birt?
18. What features of Teradata do you like?
19. You are about to send one million email (marketing campaign). How do you optimze delivery? How do you optimize response? Can you optimize both separately? (answer: not really)
20. Toad or Brio or any other similar clients are quite inefficient to query Oracle databases. Why? How would you do to increase speed by a factor 10, and be able to handle far bigger outputs?
21. How would you turn unstructured data into structured data? Is it really necessary? Is it OK to store data as flat text files rather than in an SQL-powered RDBMS?
22. What are hash table collisions? How is it avoided? How frequently does it happen?
23. How to make sure a mapreduce application has good load balance? What is load balance?
24. Examples where mapreduce does not work? Examples where it works very well? What are the security issues involved with the cloud? What do you think of EMC's solution offering an hybrid approach - both internal and external cloud - to mitigate the risks and offer other advantages (which ones)?
25. Is it better to have 100 small hash tables or one big hash table, in memory, in terms of access speed (assuming both fit within RAM)? What do you think about in-database analytics?
26. Why is naive Bayes so bad? How would you improve a spam detection algorithm that uses naive Bayes?
27. Have you been working with white lists? Positive rules? (In the context of fraud or spam detection)
28. What is star schema? Lookup tables?
29. Can you perform logistic regression with Excel? (yes) How? (use linest on log-transformed data)? Would the result be good? (Excel has numerical issues, but it's very interactive)
30. Have you optimized code or algorithms for speed: in SQL, Perl, C++, Python etc. How, and by how much?
31. Is it better to spend 5 days developing a 90% accurate solution, or 10 days for 100% accuracy? Depends on the context?
32. Define: quality assurance, six sigma, design of experiments. Give examples of good and bad designs of experiments.
33. What are the drawbacks of general linear model? Are you familiar with alternatives (Lasso, ridge regression, boosted trees)?
34. Do you think 50 small decision trees are better than a large one? Why?
35. Is actuarial science not a branch of statistics (survival analysis)? If not, how so?
36. Give examples of data that does not have a Gaussian distribution, nor log-normal. Give examples of data that has a very chaotic distribution?
37. Why is mean square error a bad measure of model performance? What would you suggest instead?
38. How can you prove that one improvement you've brought to an algorithm is really an improvement over not doing anything? Are you familiar with A/B testing?
39. What is sensitivity analysis? Is it better to have low sensitivity (that is, great robustness) and low predictive power, or the other way around? How to perform good cross-validation? What do you think about the idea of injecting noise in your data set to test the sensitivity of your models?
40. Compare logistic regression w. decision trees, neural networks. How have these technologies been vastly improved over the last 15 years?

41. Do you know / used data reduction techniques other than PCA? What do you think of step-wise regression? What kind of step-wise techniques are you familiar with? When is full data better than reduced data or sample?
42. How would you build non parametric confidence intervals, e.g. for scores? (see the AnalyticBridge theorem)
43. Are you familiar either with extreme value theory, monte carlo simulations or mathematical statistics (or anything else) to correctly estimate the chance of a very rare event?
44. What is root cause analysis? How to identify a cause vs. a correlation? Give examples.
45. How would you define and measure the predictive power of a metric?
46. How to detect the best rule set for a fraud detection scoring technology? How do you deal with rule redundancy, rule discovery, and the combinatorial nature of the problem (for finding optimum rule set - the one with best predictive power)? Can an approximate solution to the rule set problem be OK? How would you find an OK approximate solution? How would you decide it is good enough and stop looking for a better one?
47. How to create a keyword taxonomy?
48. What is a Botnet? How can it be detected?
49. Any experience with using API's? Programming API's? Google or Amazon API's? AaaS (Analytics as a service)?
50. When is it better to write your own code than using a data science software package?
51. Which tools do you use for visualization? What do you think of Tableau? R? SAS? (for graphs). How to efficiently represent 5 dimension in a chart (or in a video)?
52. What is POC (proof of concept)?
53. What types of clients have you been working with: internal, external, sales / finance / marketing / IT people? Consulting experience? Dealing with vendors, including vendor selection and testing?
54. Are you familiar with software life cycle? With IT project life cycle - from gathering requests to maintenance?
55. What is a cron job?
56. Are you a lone coder? A production guy (developer)? Or a designer (architect)?
57. Is it better to have too many false positives, or too many false negatives?
58. Are you familiar with pricing optimization, price elasticity, inventory management, competitive intelligence? Give examples.
59. How does Zillow's algorithm work? (to estimate the value of any home in US)
60. How to detect bogus reviews, or bogus Facebook accounts used for bad purposes?
61. How would you create a new anonymous digital currency?
62. Have you ever thought about creating a startup? Around which idea / concept?
63. Do you think that typed login / password will disappear? How could they be replaced?
64. Have you used time series models? Cross-correlations with time lags? Correlograms? Spectral analysis? Signal processing and filtering techniques? In which context?
65. Which data scientists do you admire most? which startups?
66. How did you become interested in data science?
67. What is an efficiency curve? What are its drawbacks, and how can they be overcome?
68. What is a recommendation engine? How does it work?
69. What is an exact test? How and when can simulations help us when we do not use an exact test?
70. What do you think makes a good data scientist?
71. Do you think data science is an art or a science?
72. What is the computational complexity of a good, fast clustering algorithm? What is a good clustering algorithm? How do you determine the number of clusters? How would you perform clustering on one million unique keywords, assuming you have 10 million data points - each one consisting of two keywords, and a metric measuring how similar these two keywords are? How would you create this 10 million data points table in the first place?
73. Give a few examples of "best practices" in data science.
74. What could make a chart misleading, difficult to read or interpret? What features should a useful chart have?
75. Do you know a few "rules of thumb" used in statistical or computer science? Or in business analytics?
76. What are your top 5 predictions for the next 20 years?
77. How do you immediately know when statistics published in an article (e.g. newspaper) are either wrong or presented to support the author's point of view, rather than correct, comprehensive factual information on a specific subject? For instance, what do you think about the official monthly unemployment statistics regularly discussed in the press? What could make them more accurate?
78. Testing your analytic intuition: look at these three charts. Two of them exhibit patterns. Which ones? Do you know that these charts are called scatter-plots? Are there other ways to visually represent this type of data?
79. You design a robust non-parametric statistic (metric) to replace correlation or R square, that (1) is independent of sample size, (2) always between -1 and +1, and (3) based on rank statistics. How do you normalize for sample size? Write an algorithm that computes all permutations of n elements. How do you sample permutations (that is, generate tons of *random* permutations) when n is large, to estimate the asymptotic distribution for your newly created metric? You may use this asymptotic distribution for normalizing your metric. Do you think that an exact theoretical distribution might exist, and therefore, we should find it, and use it rather than wasting our time trying to estimate the asymptotic distribution using simulations?
80. More difficult, technical question related to previous one. There is an obvious one-to-one correspondence between permutations of n elements and integers between 1 and n! Design an algorithm that encodes an integer less than n! as a permutation of n elements. What would be the reverse algorithm, used to decode a permutation and transform it back into a number? **Hint**: An intermediate step is to use the factorial number system representation of an integer. Feel free to check this reference online to answer the question. Even better, feel free to browse the web to find the full answer to the question (this will test the candidate's ability to quickly search online and find a solution to a problem without spending hours reinventing the wheel).
81. How many "useful" votes will a Yelp review receive? **My answer**: Eliminate bogus accounts (read this article), or competitor reviews (how to detect them: use taxonomy to classify users, and location - two Italian restaurants in same Zip code could badmouth each other and write great comments for themselves). Detect fake likes: some companies (e.g. FanMeNow.com) will charge you to produce fake accounts and fake likes. Eliminate prolific users who like everything, those who hate everything. Have a blacklist of keywords to filter fake reviews. See if IP address or IP block of reviewer is in a blacklist such as "Stop Forum Spam". Create honeypot to catch fraudsters. Also watch out for disgruntled employees badmouthing their former employer. Watch out for 2 or 3 similar comments posted the same day by 3 users regarding a company that receives very few reviews. Is it a brand new company? Add more weight to trusted users (create a category of trusted users). Flag all reviews that are identical (or nearly identical) and come from same IP address or same user. Create a metric to measure distance between two pieces of text (reviews). Create a review or reviewer taxonomy. Use hidden decision trees to rate or score review and reviewers.
82. What did you do today? Or what did you do this week / last week?
83. What/when is the latest data mining book / article you read? What/when is the latest data mining conference / webinar / class / workshop / training you attended? What/when is the most recent programming skill that you acquired?
84. What are your favorite data science websites? Who do you admire most in the data science community, and why? Which company do you admire most?
85. What/when/where is the last data science blog post you wrote?
86. In your opinion, what is data science? Machine learning? Data mining?
87. Who are the best people you recruited and where are they today?
88. Can you estimate and forecast sales for any book, based on Amazon public data? Hint: read this article.
89. What's wrong with this picture?
90. Should removing stop words be Step 1 rather than Step 3, in the search engine algorithm described here? **Answer**: Have you thought about the fact that mine and yours could also be stop words? So in a bad implementation, data mining would become data mine after stemming, then data. In practice, you remove stop words before stemming. So Step 3 should indeed become step 1.
91. Experimental design and a bit of computer science with Lego's

Related articles:

- Fast clustering algorithms for massive datasets
- The curse of big data
- What Map Reduce can't do
- 53.5 billion clicks dataset available for benchmarking and testing
- Eight worst predictive modeling techniques
- Another example of misuse of statistical science

- The curse of dimensionality (it got worse with big data)
- Data Science eBook
- Data Science Apprenticeship
- Debunking lack of analytic talent
- Causation vs. Correlation
- AnalyticTalent.com
- Data Science dictionary
- How and why to build a data dictionary
- Data Science tools
- A new random number generator
- Modern books on multiple programming languages
- Assessing efficiency of approximate vs. exact algorithms (coming soon)
- Statistical comic strip
- Fake data science
- Most popular blog posts

Previous digest | Recent jobs | Top Links | Data Science eBook
Apprenticeship | Subscribe | Events | Press Releases

- < Previous Post
- Next Post >

Comment

You need to be a member of Data Science Central to add comments!

Join Data Science Central


Comment by Jonathan DAHAN on April 19, 2016 at 6:22am

Here are 111 data science interview questions with detailed answers. Some of them come from Vincent Granville's list: http://rpubs.com/JDAHAN/172473

The list is divided in three topics: "Machine Learning & Mathematics", "Statistics" and "Process & Miscellaneous".


Comment by Radhouane ANIBA on January 17, 2016 at 1:52pm

may be it worth changing the title of this article don't you think ?


Comment by Chintan Donda on November 9, 2015 at 11:18pm

Wow, Great collection of Data Science questions.

Thanks for sharing.


Comment by Jeremy Benson on May 5, 2015 at 12:26pm

These are great. What about questions that a more junior level person should know? Say someone with 2-3 years of experience.


Comment by Vincent Granville on April 5, 2015 at 1:59pm

Hi Linda, you are welcome to add questions aimed at signal processing professionals. I was one myself when I completed my PhD thesis in 1993 (image processing, de-blurrring filters, convolution, FFT), and I consider signal processing to be data science. By the way, MatLab is a great tool. I wish more people would mention it here.


Comment by Linda Seltzer on April 5, 2015 at 7:20am

This set of questions would make it impossible for someone with a signal processing background to get hired in data science. However, signal processing engineers have our own insights into data, and especially data that takes place into time. And some of us engineers are hands on "get work done" people and can read about what we don't know in books and journals. Creativity is not always matched with memorization and test taking.


Comment by Linda Seltzer on April 5, 2015 at 7:16am

Why isn't Matlab in there? It is much more efficient to develop code in Matlab than the other programs listed in the interview question. It would be unethical as a consultant for me *not* to insist on doing my work in Matlab and that is how I would answer the question.

 Comment by Pradyumna S. Upadrashta on June 15, 2014 at 4:28am

Case in point: http://sfglobe.com/?id=411&src=share_fb_new_411

These answers show some kind of genius.

 Comment by Pradyumna S. Upadrashta on June 14, 2014 at 10:34pm

I think it is more insightful to probe deeply into a specific project that the candidate has previously worked on (e.g., incorporating some of these questions into the probe in the context of a specific project). First, these questions out of context can overwhelm good candidates in an already nerve-wracking interview setting. Second, if the interviewer is terrible with English, or inept at structuring their thoughts properly, they may be asking a question in a way that is very misleading, or asking the wrong questions altogether. Does that make the interviewee wrong (/ a failure) because they couldn't answer a poorly worded (/ thought out) question in a high-pressure, time-limited situation? For senior roles (manager, director, etc.), I think its important to have previous technical experience, or you risk hiring a dud / disaster waiting to happen.

 Comment by Joshua Weiner on April 12, 2014 at 3:53pm

What is the answer to question 37? What is wrong with mean square error? As long as you are looking at the MSE on the test set... and using it compare models, then I think it is a perfectly fine measure.

- ‹ Previous
- 1
- 2
- 3
- Next ›
- Page 2 Go

RSS

FOLLOW US

@DataScienceCtrl | RSS Feeds

TOP CONTENT


1
The Fundamentals of Data Science


2
Python Overtakes R for Data Science and Machine Learning


3
Cross-Validation: Concept and Example in R

4 

Big Data and Machine Learning:

5

Evolution of Machine Learning - Infographics

6

How to think like a data scientist to become one

- RSS
- View All

ANNOUNCEMENTS

Tableau Conference – We Are Data People, Join us

Last Chance to Register: The Making of Data Science Superheroes

Analytics is Smart Business. Earn Your MSA 100% Online

Become a Data Science Superhero!

Build Data Science Skills at Northwestern

Top 6 Trends in Cloud Analytics

Boost your Business Intelligence with Smart Data

Earn Your M.S. in Data Science Online - Deadline Approaching

Take your data visualizations to the next level

Next Generation of Data Science IDE

VIDEOS

- 

  Predictive Analytics for Supply Chain Management

  Added by Tim Matteson 0 Comments 3 Likes

- 

  Parallelize R Code Using Apache® Spark™

  Added by Tim Matteson 0 Comments 1 Like

- Add Videos
- View All

RESOURCES

- Migrating an Excel Spreadsheet to MySQL and to Spark 2.0.1 (Part 1)
- Introduction to Programming in Stata
- Benchmarking 20 Machine Learning Models Accuracy and Speed
- Stata Cheat Sheet
- Selection of best articles from our past weekly digests
- Statistical Analysis Advisor Chart
- Selection of best articles from our past weekly digests
- Free Online Book: Forecasting, Principles and Practice
- 38 Seminal Articles Every Data Scientist Should Read
- Black-box Confidence Intervals: Excel and Perl Implementation

TOP CATEGORIES

Machine Learning

R Programming

Python for Data Science

Visualization, Dashboards

NoSQL and NewSQL

Big Data

Cheat Sheets

Internet of Things

Excel