50 Questions to Test True Data Science Knowledge

* Posted by [L.V.](https://www.datasciencecentral.com/profile/LaetitiaVanCauwenberge) on January 8, 2016 at 10:30am
* [View Blog](https://www.datasciencecentral.com/profiles/blog/list?user=2w87dqyq4vlg4)

This was the subject of a popular discussion recently posted on Quora: *20 questions to detect a fake data scientist*. We asked [our own data scientist](https://www.datasciencecentral.com/profile/VincentGranville), and he came up with a very different set of questions: compare his answer (#1 below - 20 questions) with Quora replies (#2 and #3 below - 30 questions). Note that #2 focuses on statistics, and #3 on architecture. The link to the original Quora discussion is also provided in this article. Which questions would you add or remove?

[](https://api.ning.com/files/9DhwROeyJ5XtZk3VFwR6kKlDNV8B76PzvFERybkxpMNo7wLKLfOvY01QznSn*lU4jRf47wyNKghfSMN1627aA6o--kn05AX7/Capture.PNG)

Many other related interview questions and answers (data science, R, Python and so on) can be found [here](https://www.datasciencecentral.com/page/search?q=interview+questions).

**1. Answer from our data scientist** (many of these questions are open questions):

1. What is the life cycle of a data science project?
2. How do you measure yield (over base line) resulting from a new or refined algorithm or architecture?
3. What is cross-validation? How to do it right?
4. Is it better to design robust or accurate algorithms?
5. Have you written production code? Prototyped an algorithm? Created a proof of concept?
6. What is the biggest data set you have worked with, in terms of training set size, and in terms of having your algorithm implemented in production mode to process billions of transactions per day / month / year?
7. Name a few famous API's (for instance Google search). How would you create one?
8. How to efficiently scrape web data, or collect tons of tweets?
9. How to optimize algorithms (parallel processing and/or faster algorithm: provide examples for both)
10. Examples of NoSQL architecture?
11. How do you clean data?
12. How do you define / select metrics? Have you designed and used compound metrics?
13. Examples of bad and good visualizations?
14. Have you been involved - as an adviser or architect - in the design of dashboard or alarm systems?
15. How frequently an algorithm must be updated? What about lookup tables in real-time systems?
16. Provide examples of machine-to-machine communication.
17. Provide examples where you automated a repetitive analytical task.
18. How do you assess the statistical significance of an insight?
19. How to turn unstructured data into structured data?
20. How to very efficiently cluster 100 billion web pages, for instance with a tagging or indexing algorithm?
21. If you were interviewing a data scientist, what questions would you ask her?

**2. This answer was posted on Quora by Jay Verkuilen**:

1. Explain what *regularization*is and why it is useful. What are the benefits and drawbacks of specific methods, such as ridge regression and LASSO?
2. Explain what a *local optimum* is and why it is important in a specific context, such as *k*-means clustering. What are specific ways for determining if you have a local optimum problem? What can be done to avoid local optima?
3. Assume you need to generate a *predictive model of a quantitative outcome variable using multiple regression*. Explain how you intend to validate this model.
4. Explain what *precision*and *recall*are. How do they relate to the ROC curve?
5. Explain what a *long tailed distribution* is and provide three examples of relevant phenomena that have long tails. Why are they important in classification and prediction problems?
6. What is *latent semantic indexing*? What is it used for? What are the specific limitations of the method?
7. What is the *Central Limit Theorem*? Explain it. Why is it important? When does it fail to hold?
8. What is *statistical power*?
9. Explain what *resampling methods* are and why they are useful. Also explain their limitations.
10. Explain the differences between *artificial neural networks with softmax activation, logistic regression, and the maximum entropy classifier.*
11. Explain*selection bias*(with regards to a dataset, not variable selection).Why is it important? How can data management procedures such as missing data handling make it worse?
12. Provide a simple example of how an *experimental* *design* can help answer a question about behavior. For instance, explain how an experimental design can be used to optimize a web page. How does experimental data contrast with observational data.
13. Explain the difference between "long" and "wide" format data. Why would you use one or the other?
14. Is *mean imputation of missing data* acceptable practice? Why or why not?
15. Explain Edward Tufte's concept of "chart junk."
16. What is an *outlier*? Explain how you might screen for outliers and what you would do if you found them in your dataset. Also, explain what an *inlier* is and how you might screen for them and what you would do if you found them in your dataset.
17. What is principal components analysis (PCA)? Explain the sorts of problems you would use PCA for. Also explain its limitations as a method.
18. You have data on the duration of calls to a call center. Generate a plan for how you would code and analyze these data. Explain a plausible scenario for what the distribution of these durations might look like. How could you test (even graphically) whether your expectations are borne out?
19. Explain what a *false positive*and a *false negative*are. Why is it important to differentiate these from each other? Provide examples of situations where (1) false positives are more important than false negatives, (2) false negatives are more important than false positives, and (3) these two types of errors are about equally important.
20. Explain likely differences encountered between administrative datasets and datasets gathered from experimental studies. What are likely problems encountered with administrative data? How do experimental methods help alleviate these problems? What problems do they bring?

**3. Kavita Ganesan offered this answer**:

1. What is a gold standard ?   
   Believe it or not there are data scientists (even at very senior levels) who claim to know a hell lot about supervised machine learning and know nothing about what a gold standard is!
2. What is the difference between supervised learning and unsupervised learning? - Give concrete examples.
3. What does NLP stand for?  
   Some data scientists claim to also do NLP.
4. Write code to count the number of words in a document using any programming language. Now, extend this for bi-grams.  
   I have seen a senior level data scientist who actually struggled to implement this.
5. What are feature vectors?
6. When would you use SVMs vs Random Forrest and Why?
7. What is your definition of Big Data, and what is the largest size of data you have worked with? Did you parallelize your code?  
   If their notion of big data is just volume - you may have a problem. Big Data is more than just volume of data. If the largest size of data they have worked with is 5MB - again you may have a problem.
8. How do you work with large data sets?  
   If the answer only comes out as hadoop it clearly shows that their view of solving problems is extremely narrow. Large data problems can be solved with:  
   1. efficient algorithms  
   2. multi-threaded applications  
   3. distributed programming  
   4. more...
9. Write a mapper function to count word frequencies (even if its just pseudo code)
10. Write a reducer function for counting word frequencies (even if its just pseudo code)

You can find the original contribution [here](https://www.quora.com/What-are-20-questions-to-detect-fake-data-scientists).

**DSC Resources**

* Career: [Training](http://www.analyticbridge.com/group/analyticscourses/forum) | [Books](http://www.analyticbridge.com/group/books/forum) | [Cheat Sheet](https://www.datasciencecentral.com/profiles/blogs/data-science-cheat-sheet) | [Apprenticeship](https://www.datasciencecentral.com/group/data-science-apprenticeship/forum/topics/our-data-science-apprenticeship-is-now-live) | [Certification](https://www.datasciencecentral.com/group/data-science-certification) | [Salary Surveys](http://www.analyticbridge.com/group/salary-trends-and-reports/forum) | [Jobs](http://careers.analytictalent.com/jobs/search/results)
* Knowledge: [Research](https://www.datasciencecentral.com/group/resources/forum) | [Competitions](https://www.datasciencecentral.com/group/resources/forum/topics/best-kept-secret-about-data-science-competitions) | [Webinars](https://www.datasciencecentral.com/video/video/listFeatured) | [Our Book](https://www.datasciencecentral.com/profiles/blogs/my-data-science-book) | [Members Only](https://www.datasciencecentral.com/page/member) | [Search DSC](https://www.google.com/cse/publicurl?cx=007889287169707156432:04zouenp6u0)
* Buzz: [Business News](http://www.bigdatanews.com/group/bdn-daily-press-releases) | [Announcements](https://www.datasciencecentral.com/group/announcements/forum) | [Events](https://www.datasciencecentral.com/events/) | [RSS Feeds](https://www.datasciencecentral.com/page/news-feeds)
* Misc: [Top Links](http://www.analyticbridge.com/page/links) | [Code Snippets](http://www.analyticbridge.com/group/codesnippets) | [External Resources](https://www.datasciencecentral.com/group/resources/forum/topics/comprehensive-list-of-data-science-resources) | [Best Blogs](https://www.datasciencecentral.com/group/resources/forum/topics/the-best-of-our-weekly-digests) | [Subscribe](https://www.datasciencecentral.com/profiles/blogs/check-out-our-dsc-newsletter) | [For Bloggers](https://www.datasciencecentral.com/group/resources/forum/topics/free-stuff-for-data-science-publishers-authors-bloggers-professor)

**Additional Reading**

* [The 10 Best Books to Read Now on IoT](http://www.iotcentral.io/blog/the-10-best-books-to-read-now-on-iot)
* [50 Articles about Hadoop and Related Topics](http://www.hadoop360.com/blog/50-articles-about-hadoop-and-related-topics)
* [10 Modern Statistical Concepts Discovered by Data Scientists](https://www.datasciencecentral.com/profiles/blogs/10-modern-statistical-concepts-discovered-by-data-scientists)
* [Top data science keywords on DSC](https://www.datasciencecentral.com/forum/topics/most-popular-data-science-keywords-on-dsc)
* [4 easy steps to becoming a data scientist](https://www.datasciencecentral.com/profiles/blogs/4-easy-steps-to-becoming-a-data-scientist)
* [13 New Trends in Big Data and Data Science](https://www.datasciencecentral.com/profiles/blogs/13-new-trends-in-big-data-and-data-science)
* [22 tips for better data science](https://www.datasciencecentral.com/profiles/blogs/22-tips-for-better-data-science)
* [Data Science Compared to 16 Analytic Disciplines](https://www.datasciencecentral.com/profiles/blogs/17-analytic-disciplines-compared)
* [How to detect spurious correlations, and how to find the real ones](https://www.datasciencecentral.com/profiles/blogs/tutorial-how-to-detect-spurious-correlations-and-how-to-find-the-)
* [17 short tutorials all data scientists should read (and practice)](https://www.datasciencecentral.com/profiles/blogs/17-short-tutorials-all-data-scientists-should-read-and-practice)
* [10 types of data scientists](https://www.datasciencecentral.com/profiles/blogs/six-categories-of-data-scientists)
* [66 job interview questions for data scientists](https://www.datasciencecentral.com/profiles/blogs/66-job-interview-questions-for-data-scientists)
* [High versus low-level data science](https://www.datasciencecentral.com/profiles/blogs/high-level-versus-low-level-data-science)

Follow us on Twitter: [@DataScienceCtrl](https://www.twitter.com/datasciencectrl) | [@AnalyticBridge](https://www.twitter.com/analyticbridge)

Views: **101971**

[Like](https://www.datasciencecentral.com/profiles/blog/show?id=6448529%3ABlogPost%3A371610)

[10 members like this](https://www.datasciencecentral.com/profiles/blog/show?id=6448529%3ABlogPost%3A371610)

[Share](https://www.datasciencecentral.com/main/sharing/share?id=6448529%253ABlogPost%253A371610)[Tweet](https://twitter.com/share)

* [< Previous Post](https://www.datasciencecentral.com/profiles/blogs/top-algorithm-experts)
* [Next Post >](https://www.datasciencecentral.com/profiles/blogs/top-20-big-data-experts-to-follow-1)

Comment

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

[Join Data Science Central](https://www.datasciencecentral.com/main/authorization/signUp?target=https%3A%2F%2Fwww.datasciencecentral.com%2Fprofiles%2Fblog%2Fshow%3Fid%3D6448529%253ABlogPost%253A371610)

[https://api.ning.com/files/UmB7s*t7XzbpTb2ir9Z3rv96Ft-pfSm6JcuQQmdJQNHssIfdpqngyrZhWLPfJCTlSs1UySApFgk71aI7CXV98OoxxG-ZkqOW/493908672.jpeg?xgip=755%3A165%3A909%3A909%3B%3B&width=48&height=48&crop=1%3A1](https://www.datasciencecentral.com/profile/MarkLStone)Comment by [Mark L. Stone](https://www.datasciencecentral.com/profile/MarkLStone) on December 29, 2016 at 4:25am

Question 14 :   Is *mean imputation of missing data* acceptable practice? Why or why not?

Most imputation methods currently used should not be acceptable. Imputation can totally screw up modeling and analysis. In general, imputation should be done as part of a joint optimization with the model being fit and predictions being made.  To do otherwise is generally sub-optimal, possibly in a huge way.

[https://api.ning.com/files/UmB7s*t7XzbpTb2ir9Z3rv96Ft-pfSm6JcuQQmdJQNHssIfdpqngyrZhWLPfJCTlSs1UySApFgk71aI7CXV98OoxxG-ZkqOW/493908672.jpeg?xgip=755%3A165%3A909%3A909%3B%3B&width=48&height=48&crop=1%3A1](https://www.datasciencecentral.com/profile/MarkLStone)Comment by [Mark L. Stone](https://www.datasciencecentral.com/profile/MarkLStone) on December 28, 2016 at 5:31pm

If Kavita Ganesan weren't such a pretentious twit, before asking "What does NLP stand for?  
Some data scientists claim to also do NLP." maybe she would have found out that long before NLP stood for Natural Language Processing, it stood for, and still does stand for, Nonlinear Programming. Does she realize that Nonlinear Programming is another name for nonlinear optimization, which underlies so many data science and statistical algorithms, and is even used in Natural Language Processing?  As an expert in NLP, maybe she should know there is more than one meaning to NLP which is applicable to data science. isn't recognizing multiple possible meanings, in context, part of NLP? These know-it-alls who don't know-it-all really gall me.

[https://api.ning.com/files/UjLFZWxAkgp2ZX9mXjRQq3Tgj3T85w*nUodpnHLp5ShvX2z4QBFN4Y0rikhBDzc-dwJ4sd2lhtMMUTUWWpJN4HmnDhZDFsy*/1174267274.png?width=48&height=48&crop=1%3A1](https://www.datasciencecentral.com/profile/DalilaBenachenhou)Comment by [Dalila Benachenhou](https://www.datasciencecentral.com/profile/DalilaBenachenhou) on July 7, 2016 at 7:26am

* How do you clean data?

Are we talking about unstructured (text) or structured data.

* Explain what a *local optimum* is and why it is important in a specific context, such as *k*-means clustering. What can be done to avoid local optima?  Change the starting conditions (change the choice of the initial centroids)
* What are specific ways for determining if you have a local optimum problem?
* Analysis the within distance between elements in a cluster and the cluster centroid, and the distance between the centroid (don't choose your K a-priori)

What is your definition of Big Data?

In my opinion one cannot call population income a Big Data problem.  You may have income of 24 million, but you really need only a small sample to get income distribution, and get a summary statistics.  It may be an IT Big Data problem but it just requires Small Data analysis.

In some cases, you may need to use Big Data technology to build a ML model, but after the model is build and put in production the needs goes away.

[https://api.ning.com/files/UjLFZWxAkgp2ZX9mXjRQq3Tgj3T85w*nUodpnHLp5ShvX2z4QBFN4Y0rikhBDzc-dwJ4sd2lhtMMUTUWWpJN4HmnDhZDFsy*/1174267274.png?width=48&height=48&crop=1%3A1](https://www.datasciencecentral.com/profile/DalilaBenachenhou)Comment by [Dalila Benachenhou](https://www.datasciencecentral.com/profile/DalilaBenachenhou) on July 7, 2016 at 6:51am

1. Is it better to design robust or accurate algorithms? This is the same question that a CS or an engineer deal with.   In both cases  you strive to achieve both, but you know that you can always design a robust algorithm but you cannot always design an accurate algorithm or care about designing an accurate algorithm.  If the purpose is to understand the relation between predictors and response variables (for instance, try to get a profile of potential clients) accuracy of the algorithm is not important.  On the other hand, for critical matters, having a helicopter that lands itself, or having a rice cooker cook rice at perfection, accuracy is crucial.

[https://api.ning.com/files/I5Q3A4NUOoKZu3PsD7oI2m9IbFScacCiYpTSXrD-ic*FRdqF8dgLFro4rZl6kZhAw2fLcd3mPKrYAkQBg1IfwS3HdI*hCm9I/genericavatar.png?width=48&height=48&crop=1%3A1](https://www.datasciencecentral.com/profile/XavierSumba)Comment by [Xavier Sumba](https://www.datasciencecentral.com/profile/XavierSumba) on January 12, 2016 at 12:24pm

For those who are interested in interview questions. There is a github page [1] with lists of interview questions. I have added a new category for Data Science. You can look for another questions if you are interested.  
  
[1] <https://github.com/MaximAbramchuck/awesome-interviews>

[https://api.ning.com/files/I5Q3A4NUOoKZu3PsD7oI2m9IbFScacCiYpTSXrD-ic*FRdqF8dgLFro4rZl6kZhAw2fLcd3mPKrYAkQBg1IfwS3HdI*hCm9I/genericavatar.png?width=48&height=48&crop=1%3A1](https://www.datasciencecentral.com/profile/MateiBeremski434)Comment by [Matei Beremski](https://www.datasciencecentral.com/profile/MateiBeremski434) on January 11, 2016 at 12:15pm

I was referring to yours :) Stark difference between your questions and those on Quora. Interesting to see that the focus of the data science community is still very much on modelling and algorithm choice.

[https://api.ning.com/files/cjKbLN5hhpAo0IKBg1GBIQr3lsgeXxuuVJB8NJWsoiaiVM*GIrWi*qLn8GvoG6E1kkhyxQ5yXj3AJ43XkaKODR0xPVsEpwbB/1240260993.jpeg?xgip=0%3A0%3A399%3A399%3B%3B&width=48&height=48&crop=1%3A1](https://www.datasciencecentral.com/profile/SioneKPalu)Comment by [Sione Palu](https://www.datasciencecentral.com/profile/SioneKPalu) on January 11, 2016 at 12:05pm

Question 14 :   Is *mean imputation of missing data* acceptable practice? Why or why not?

Regression task for temporal based data, imputation is necessary.  Favorites imputation pre-processing for me in time-series analysis is to impute via Kalman-Filter Smoothing. Once done, then other steps like noise removal can be applied then do forecasting afterwards.

[https://api.ning.com/files/0A1mEG0sKPWjMqJf4z6fFy1OBR5wbI4pBUrsFSkzQQohlxGZG1CWnKgjS40GjqQ8*-B2D8fV1bbTLyBNchyxkfDHb7lbTIFh/1206399868.png?xgip=0%3A17%3A185%3A185%3B%3B&width=48&height=48&crop=1%3A1](https://www.datasciencecentral.com/profile/VincentGranville)Comment by [Vincent Granville](https://www.datasciencecentral.com/profile/VincentGranville) on January 11, 2016 at 9:29am

I'd love to provide the answers, at least to my questions. Finding time is difficult.

[https://api.ning.com/files/I5Q3A4NUOoKZu3PsD7oI2m9IbFScacCiYpTSXrD-ic*FRdqF8dgLFro4rZl6kZhAw2fLcd3mPKrYAkQBg1IfwS3HdI*hCm9I/genericavatar.png?width=48&height=48&crop=1%3A1](https://www.datasciencecentral.com/profile/MateiBeremski434)Comment by [Matei Beremski](https://www.datasciencecentral.com/profile/MateiBeremski434) on January 11, 2016 at 2:56am

Thank you for sharing this. Great Questions.

If you have time Vincent, it would also be great to see what you think make good responses!