



Data Analysis

Unfollow Topic | 96.1k

Pin to Home Ask Question Share

...

TOP STORIES

Answer written • Data Analysis • 2013

How can I become a data scientist?



Katie Kent, Director of Educational Outcomes @ Galvanize; Employee #1 @ Zipfian Academy

177.2k Views • Upvoted by Jason Zhang, [Data Scientist at Quora](#)
Katie has 50+ answers in Data Science.

Originally Answered: How do I become a data scientist?

Become a Data Scientist by Doing Data Science

The best way to become a data scientist is to learn - and do - data science. There are a many excellent courses and tools available online that can help you get there.

Here is an incredible list of resources compiled by Jonathan Dinu, Co-founder of [Zipfian Academy](#), which trains data scientists and data engineers in San Francisco via immersive programs, fellowships, and workshops.

EDIT: I've had several requests for a permalink to this answer. See here: [A Practical Intro to Data Science from Zipfian Academy](#)

EDIT2: See also: "How to Become a Data Scientist" on SlideShare:
<http://www.slideshare.net/ryanor...>

Environment

Python is a great programming language of choice for aspiring data scientists due to its general purpose applicability, a [gentle](#) (or [firm](#)) learning curve, and — perhaps the most compelling reason — the rich ecosystem of [resources](#) and [libraries](#) actively used by the scientific community.

Development

When learning a new language in a new domain, it helps immensely to have an interactive environment to explore and to receive immediate feedback. IPython provides an interactive REPL which also allows you to integrate a wide variety of frameworks (including [R](#)) into your Python programs.

STATISTICS

Data scientists are better at software engineering than statisticians and better at statistics than any software engineer. As such, statistical inference underpins much of the theory behind data analysis and a solid foundation of statistical methods and probability serves as a stepping stone into the world of data science.

Courses

[edX: Introduction to Statistics: Descriptive Statistics](#): A basic introductory statistics course.

[Coursera Statistics, Making Sense of Data](#): A applied Statistics course that teaches the complete pipeline of statistical analysis

MOST VIEWED WRITERS

[View More](#)



William Chen, Data Scientist at Quora
40,635 Views



Carlos Del Carpio
22,140 Views



Alex Kamil
21,859 Views

RELATED TOPICS



Machine Learning
477.3k Followers



Apache Hadoop
36.5k Followers



Statistics (academic discipline)
300.3k Followers



Analytics
65.1k Followers



Big Data
146.8k Followers

[View 15 More](#)

Top Stories

[Open Questions](#)

[All Questions](#)

[Followers](#)

[Manage](#)





[Log](#)

0 REVIEWS

[View](#)

★★★★★


[MIT: Statistical Thinking and Data Analysis](#) : Introduction to probability, sampling, regression, common distributions, and inference.


While R is the de facto standard for performing statistical analysis, it has quite a high learning curve and there are other areas of data science for which it is not well suited. To avoid learning a new language for a specific problem domain, we recommend trying to perform the exercises of these courses with Python and its numerous statistical libraries. You will find that much of the functionality of R can be replicated with [NumPy](#) , [@SciPy](#) , [@Matplotlib](#) , and [@Python Data Analysis Library](#) .

Books


Well-written books can be a great reference (and supplement) to these courses, and also provide a more independent learning experience. These may be useful if you already have some knowledge of the subject or just need to fill in some gaps in your understanding:

[O'Reilly Think Stats](#) : An Introduction to Probability and Statistics for Python programmers

[Introduction to Probability](#) : Textbook for Berkeley's Stats 134 class, an introductory treatment of probability with complementary exercises.

[Berkeley Lecture Notes, Introduction to Probability](#) : Compiled lecture notes of above textbook, complete with exercises.


[OpenIntro](#) : Statistics: Introductory text book with supplementary exercises and labs in an online portal.

[Think Bayes](#) : An simple introduction to Bayesian Statistics with Python code examples.


MACHINE LEARNING/ALGORITHMS


A solid base of Computer Science and algorithms is essential for an aspiring data scientist. Luckily there are a wealth of great resources online, and machine learning is one of the more lucrative (and advanced) skills of a data scientist.

Courses

[Coursera Machine Learning](#) : Stanford's famous machine learning course taught by Andrew Ng.


[Coursera: Computational Methods for Data Analysis](#) : Statistical methods and data analysis applied to physical, engineering, and biological sciences.


[MIT Data Mining](#) : An introduction to the techniques of data mining and how to apply ML algorithms to garner insights.


[Edx: Introduction to Artificial Intelligence](#) : Introduction to Artificial Intelligence: The first half of Berkeley's popular AI course that teaches you to build autonomous agents to efficiently make decisions in stochastic and adversarial settings.


[Introduction to Computer Science and Programming](#) : MIT's introductory course to the theory and application of Computer Science.


Books


[UCI: A First Encounter with Machine Learning](#) : An introduction to machine learning concepts focusing on the intuition and explanation behind why they work.

[A Programmer's Guide to Data Mining](#) : A web based book complete with code samples (in Python) and exercises.

[Data Structures and Algorithms with Object-Oriented Design Patterns in Python](#) : An introduction to computer science with code examples in Python — covers algorithm analysis, data structures, sorting algorithms, and object oriented design.

[An Introduction to Data Mining](#) : An interactive Decision Tree guide (with hyperlinked lectures) to learning data mining and ML.

[Elements of Statistical Learning](#) : One of the most comprehensive treatments of data mining and ML, often used as a university textbook.

[Stanford: An Introduction to Information Retrieval](#) : Textbook from a Stanford course on NLP and information retrieval with sections on text classification, clustering, indexing, and web crawling.


DATA INGESTION AND CLEANING

One of the most under-appreciated aspects of data science is the cleaning and munging of data that often represents the most significant time sink during analysis. While there is never a silver bullet for such a problem, knowing the right tools, techniques, and approaches can help minimize time spent wrangling data.


Courses

[School of Data: A Gentle Introduction to Cleaning Data](#) : A hands on approach to learning to clean data, with plenty of exercises and web resources.


Tutorials


[Predictive Analytics: Data Preparation](#) : An introduction to the concepts and techniques of sampling data, accounting for erroneous values, and manipulating the data to transform it into acceptable formats.

Tools

[OpenRefine](#)  (formerly Google Refine): A powerful tool for working with messy data, cleaning, transforming, extending it with web services, and linking to databases. Think Excel on steroids.

[Data Wrangler](#) : Stanford research project that provides an interactive tool for data cleaning and transformation.

[sed - an Introduction and Tutorial](#) : “The ultimate stream editor,” used to process files with regular expressions often used for substitution.

[awk - An Introduction and Tutorial](#) : “Another cornerstone of UNIX shell programming” — used for processing rows and columns of information.


VISUALIZATION

The most insightful data analysis is useless unless you can effectively communicate your results. The art of visualization has a long history, and while being one of the most qualitative aspects of data science its methods and tools are well documented.

Courses

[UC Berkeley Visualization](#) : Graduate class on the techniques and algorithms for creating effective visualizations.

[Rice University Data Visualization](#) : A treatment of data visualization and how to meaningfully present information from the perspective of Statistics.

[Harvard University Introduction to Computing, Modeling, and Visualization](#) : Connects the concepts of computing with data to the process of interactively visualizing results.

Books

[Tuft: The Visual Display of Quantitative Information](#) : Not freely available,

but perhaps the most influential text for the subject of data visualization. A classic that defined the field.

Tutorials

[School of Data: From Data to Diagrams](#): A gentle introduction to plotting and charting data, with exercises.

[Predictive Analytics: Overview and Data Visualization](#): An introduction to the process of predictive modeling, and a treatment of the visualization of its results.

Tools

[D3.js](#): Data-Driven Documents — Declarative manipulation of DOM elements with data dependent functions (with [Python port](#)).

[Vega](#): A visualization grammar built on top of D3 for declarative visualizations in JSON. Released by the dream team at [Trifacta](#), it provides a higher level abstraction than D3 for creating “ or SVG based graphics.

[Rickshaw](#): A charting library built on top of D3 with a focus on interactive time series graphs.

[Modest Maps](#): A lightweight library with a simple interface for working with maps in the browser (with ports to multiple languages).

[Chart.js](#): Very simple (only six charts) HTML5 “ based plotting library with beautiful styling and animation.

COMPUTING AT SCALE

When you start operating with data at the scale of the web (or [greater](#)), the fundamental approach and process of analysis must change. To combat the ever increasing amount of data, Google developed the [MapReduce](#) paradigm. This programming model has become the de facto standard for large scale batch processing since the release of Apache [Hadoop](#) in 2007, the open-source MapReduce framework.

Courses

[UC Berkeley: Analyzing Big Data with Twitter](#): A course — taught in close collaboration with Twitter — that focuses on the tools and algorithms for data analysis as applied to Twitter microblog data (with project based curriculum).

[Coursera: Web Intelligence and Big Data](#): An introduction to dealing with large quantities of data from the web; how the tools and techniques for acquiring, manipulating, querying, and analyzing data change at scale.

[CMU: Machine Learning with Large Datasets](#): A course on scaling machine learning algorithms on Hadoop to handle massive datasets.

[U of Chicago: Large Scale Learning](#): A treatment of handling large datasets through dimensionality reduction, classification, feature parametrization, and efficient data structures.

[UC Berkeley: Scalable Machine Learning](#): A broad introduction to the systems, algorithms, models, and optimizations necessary at scale.

Books

[Mining Massive Datasets](#): Stanford course resources on large scale machine learning and MapReduce with accompanying book.

[Data-Intensive Text Processing with MapReduce](#): An introduction to algorithms for the indexing and processing of text that teaches you to “think in MapReduce.”

[Hadoop: The Definitive Guide](#): The most thorough treatment of the Hadoop

framework, a great tutorial and reference alike.

[Programming Pig](#): An introduction to the Pig framework for programming data flows on Hadoop.

PUTTING IT ALL TOGETHER

Data Science is an inherently multidisciplinary field that requires a [myriad](#) of skills to be a proficient practitioner. The necessary curriculum has not fit into traditional course offerings, but as [awareness](#) of the [need](#) for individuals who have such abilities is growing, we are seeing universities and private companies creating custom classes.

Courses

[UC Berkeley: Introduction to Data Science](#): A course taught by Jeff Hammerbacher and Mike Franklin that highlights each of the varied skills that a Data Scientist must be proficient with.

[How to Process, Analyze, and Visualize Data](#): A lab oriented course that teaches you the entire pipeline of data science; from acquiring datasets and analyzing them at scale to effectively visualizing the results.

[Coursera: Introduction to Data Science](#): A tour of the basic techniques for Data Science including SQL and NoSQL databases, MapReduce on Hadoop, ML algorithms, and data visualization.

[Columbia: Introduction to Data Science](#): A very comprehensive course that covers all aspects of data science, with an humanistic treatment of the field.

[Columbia: Applied Data Science](#) (with [book](#)): Another Columbia course — teaches applied software development fundamentals using real data, targeted towards people with mathematical backgrounds.

[Coursera: Data Analysis](#) (with [notes](#) and [lectures](#)): An applied statistics course that covers algorithms and techniques for analyzing data and interpreting the results to communicate your findings.

Books

[An Introduction to Data Science](#): The companion textbook to Syracuse University's flagship course for their new Data Science program.

Tutorials

[Kaggle: Getting Started With Python For Data Science](#): A guided tour of setting up a development environment, an introduction to making your first competition submission, and validating your results.

CONCLUSION

Data science is infinitely complex field and this is just the beginning.

If you want to get your hands dirty and gain experience working with these tools in a collaborative environment, check out our programs at <http://zipfianacademy.com>.

There's also a great SlideShare summarizing these skills: [How to Become a Data Scientist](#)

You're also invited to connect with us on Twitter [@zipfianacademy](#) and let us know if you want to learn more about any of these topics.

Updated 2 Jul, 2014 • View Upvotes

Upvote | 1.2k

Downvote

Comments 14+

Share 22

...

Question asked • CartoDB • 29 Jun

What applications use CartoCSS?

I know of Mapbox and CartoDB, assuming they use the same flavor. Are there other examples?

[Write Answer](#)
[Pass](#)
[Follow](#) **3** [Downvote](#)

...

Answer written • Data Analysis • 19 May

What are the most marketable skills in the field of Data, Analysis, and Data Science?



Nir Goldstein

14.3k Views • Upvoted by Jalem Raj Rohit, [Data Scientist at Kayako](#)

Nir has 10+ answers in Data Science.

Here are the most required skills for a data scientist position from analyzing thousands of job posts (I also included some free resources I found for each skill): 1. Python

- [Web Programming Beginne...](#)

(more)

[Upvote](#) | **186**
[Downvote](#)
[Comments](#) **2+**
[Share](#) **11**

...

Answer written • Machine Learning • 7 Jul

Can you suggest some data science newsletters or a page with many Python data science tutorials?



Alexey Grigorev, Freelance ML developer

133 Views • Alexey has 70+ answers in Machine Learning.

For python there's a very nice thing called IPython notebook which allows you to store both the code and the results. So what you can do is google something on the topic of interest but also add "IPython notebook" to the query.

But since the question is about some page with many tutorials, here it is: [A gallery of interesting IPython Notebooks](#). You will find a lot of materials and tutorials there.

Written 7 Jul • View Upvotes • Asked to answer by Anonymous

[Upvote](#) | **5**
[Downvote](#)
[Comment](#)
[Share](#)

...

Answer written • Data Analysis • 2015

What are 20 questions to detect fake data scientists?



Jay Verkuilen, PhD Psychometrics UIUC 2007

23.9k Views • Upvoted by Alexander Blocker, [Statistician at Google](#), [PhD in statistics from Harvard, com...](#)

Jay has 70+ answers in Data Analysis.

I'll stay away from code examples myself as they seem rather shop-specific and thus best designed locally, but if you want some questions, here you go. These questions are intentionally difficult and are more on the statistics/modeling side than the data processing side. That's important, but someone else would be better poised to write those questions.

You might want "I don't know, but what I would do is read the following sources...." to be part of your accepted answers, as that's partly testing honesty and forthrightness of the candidate. The last thing an organization needs is bullshit artists who overpromise what they can do or just make things up.

Note: These aren't definitive or even representative and reflect my own areas of expertise. These are prototype questions, you should alter/edit or formulate your own. You should add details to the questions to deal with the data types you typically deal with.

Your organization needs to define what people being hired into the job actually need to know how to do, and ask about that. If they're not doing a lot of cluster analysis, why would you ask about that? If the person is mostly

doing data management/cleaning, primarily ask about that.

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.

That's 12.

More for a basis of questions: [How do data scientists use statistics?](#)

Updated 31 Aug • View Upvotes

Upvote | 267

Downvote

Comments 8+

Share 4

...

Answer written • Data Analysis • 24 Aug

Why is Python used heavily in big data when it has issues with multithreading due to the Global Interpreter Lock (GIL)?



Jim Dennis, Python from an Ops perspective

5.9k Views • Jim has 120+ answers in Python (programming language).

Questions like this make me sad. Let's break it down into two questions:

"Why is Python used in big data?"...

(more)

Upvote | 23

Downvote

Comments 1+

Share

...

Answer written • Machine Learning • 2013

How and when did you get started on kaggle?



Brian Feeny, Top 1% on Kaggle

3.5k Views

I first started about one year ago, and was very busy in the beginning. I used R as my primary tool and was basically using Kaggle to learn R, while at the same time using R to compete in Kaggle. I only did a few competitions officially, but

I also played around with a lot of data to try my own experiments some of which I never entered officially. The main reason I have not been active lately is that its time consuming, very addicting, the competition is intense. In many cases, if you do not work in a team (I was always individual), you must learn some domain of knowledge that you are not familiar with. This could be as little as reading a web page, or as much as reading entire volumes of books. For the Traveling Santa problem, which was one of my favorite, I used very little external tools, and implemented my own versions of several heuristic algorithms. Kaggle helped me understand that there is nothing I enjoy doing more than solving problems involving data.

Written 17 Nov, 2013 • View Upvotes

Upvote | 43

Downvote Comment Share 1

...

Answer written • Machine Learning • 23 Aug

What are the best data science podcasts?



Sathya Narayanan, BI & Analytics | Solution Engineering | R

1.3k Views

Here are the ones i listen to.

1. [Data Skeptic](#)
2. [Linear Digressions](#)
3. [Partially Derivative - Episodes](#)
4. [O'Reilly Data Show Podcast Archives](#)...

(more)

Upvote | 11

Downvote Comment Share

...

Answer written • Startup Founders and E... • 2014

Startup Ideas: How do you know if your startup idea already exists?



Louis Leone

30.2k Views

There are over 7 billion people in the world. Yes, your idea probably exists. Think of it this way. How much would you pay for a book of "great" ideas? Would you pay \$100,000; how about \$100? ... [\(more\)](#)

Upvote | 271

Downvote Comments 6+ Share

...

Answer written • Data Analysis • 2014

What are the downsides of being a data scientist?



Anonymous

3.1k Views • Upvoted by Abhinav Sharma, I worked on data at Facebook

Here's what I think are the downsides of being a data scientist. I'm assuming data science refers to a "find insights and analyze experiments" data scientist role, not machine learning, which I classify under engineering.

1. There's an excellent chance your data science manager is incompetent. It's certainly not true that software engineering managers become managers because they're the best technically in their field, but pretty much all the software engineering managers I know have had at least 5+ years of experience working as actual software engineers, and have worked on large, sophisticated software projects. In contrast, I know many data science managers at Google, Yahoo, Uber, LinkedIn, Twitter, and many lesser-known startups who became managers with less than 1 year of work experience anywhere. So many of these people know very little about statistics, machine learning, visualization, or programming, and I know some of these people can not even explain conditional probability and have never written a MapReduce

job.

There are at least two reasons for this. 1. Data science is a new field, so companies have no choice but to put incompetent junior employees into managerial roles. 2. Data science is very easy to game (more on this next).

2. Nobody's going to know if you're making shit up. At very few companies do the data scientists look at and review each other's code. They say it takes too much time, their code isn't meant to be production quality because they're not engineers, it's spread all over the place, et cetera. For all anyone knows, that busy-looking data scientist may be making up a lot of his or her data. I have seen cases where numbers were fudged to look better, and cases where data were gathered because it was easier even though the assumptions built into that data were wrong, and conclusions that would make a scientific journal weep.

3. This is due, in part, to the fact that, ironically, the output of a data scientist isn't measurable. Machine learning engineers can be measured in real experiments by their improvements to CTR, infrastructure engineers can be measured by their improvements to latency, etc. This isn't true in refactoring or other cases, but in many cases an engineer's output is very easy to see. How do you measure a data scientist, by the number of graphs they produce?

4. People are very impressed by pretty sounding statements, even if its statistical nonsense, especially if the statements validate what they want to believe. For example, if you're the engineering lead for your company's mobile app, you're going to LOVE it if a data scientist comes and tells you that users who install your mobile app click on ads 10X more. Wowwee, that's AWESOME you'll tell them and your CEO, what a fascinating insight, and you'll use this piece of data to make the case that your mobile team should hire 10 times as many people. That's of course the wrong thing to do, since mobile users are naturally your most engaged users because they wanted to install your mobile app in the first place. ****Correlation is not causation all good data scientists will tell you, but these are the findings they spend all day presenting anyways!****

One concrete example, I had an engineering VP tell me a year ago how much of an impact Data Scientist Bob was making on the company, because Bob found that our heaviest users were doing XXX, and that convinced the CEO that we should make our product focus exclusively on XXX. This way, all our users would become very heavy users of our product and use it every day of course. But guess what, we launched this new product and it completely flopped, because Bob's analysis was not causal. Most users just did not care about XXX. It was like saying that because people who eat organic spend a lot of money, McDonald's should switch it's menu to expensive organic food. But Bob is still praised for that product launching analysis to this day.

Also, many of you are surely familiar with [Simpson's paradox](#) and the Berkeley gender bias example. A data scientist in any company would be praised to the high heavens for discovering and reporting that women at Berkeley are being discriminated against, even though that is the exact wrong finding!

Thus, careless data scientists who spend very little time being rigorous, or care very little about it, are very likely going to produce awesome-sounding results and look more productive than their more careful, rigorous peers.

5. It's not a very technical field, and rigor is bad for your career. Most people will tell you that Big Data Insights don't require fancy statistics. Most of what people do is writing Hadoop scripts and counting stuff. So what skills do data scientists need in reality? Not much. Just talk and a little bit of R. In contrast, it's hard to believe that a junior engineer could build Google or Facebook from scratch, or that a junior designer could build Apple. But make up interesting stories? Hire that creative arts major and give them Tableau.

On the same line, performing rigorous statistics is frequently bad for data scientists, because if it's more rigorous, that means the analysis is more complicated, which makes it harder to understand and harder to get interesting results, which makes it harder to get those product managers and leads who don't know anything about statistics interested. So you should gloss over the subtleties of causation, and present the data as clear-cut, black and white. Avoid caveats and showcase absolutes!

6. It's a hell hole of politics. Everyone says that the final result of data science is to communicate what you discover to other people. This means a lot of people schmoozing, and this ends up where data scientists say a lot of stuff just to sound smart, and given that statistics is 90% lies (lies, damned lies, and statistics!), results are easily made up or twisted.

Nevertheless, for a positive, I companionably disagree with [Abhinav Sharma](#)'s answer. I think the feedback loop of data science is faster than the feedback loop of design and engineering. Most of the time, the goal of a data science project is to talk about your findings, and it's quick to set up a presentation and get feedback on what you've found. And pulling data can be slow sometimes, but you're usually iterating on smaller datasets that you've summarized from Hadoop, and pulling those datasets from Hadoop in the first place is usually $O(\text{days or weeks})$.

In contrast, if you're an engineer, it often takes months for the infrastructure you're working on to get shipped and run as an experiment before you get to see its impact, and the same for designers.

Although the impact is less. If you're an engineer or designer, you build stuff. If you're a data scientist, you talk about it and convince other people to do stuff, and most of the time, they already have a plan and nothing you can do will change it in any way.

Written 26 Nov, 2014 • View Upvotes

Upvote | 50

Downvote Comments 4 Share 1

...

Answer written • Data Analysis • 9 Sep

Is it worth paying for a data analyst nanodegree?



Diego Menin

932 Views

Yes it is. I've completed the Coursera data science Specialization and I am currently half way through Udacity's Data Analyst Nano Degree (more info about the differences between them on [this](#) answer).

the way I see, they won't provide enough knowledge to make you a data analyst or scientist but they will definitely put you on the right track so is up to you to expand what you've learned on them

Written 9 Sep • View Upvotes

Upvote | 1

Downvote Comment Share

...

Question re-asked • Data Mining • 3 Aug

What are some good tutorials on graph mining?

Write Answer

Pass

Follow 9 Downvote

...

Answer written • Machine Learning • 2014

What is a numerical example for the Principal Component Analysis (PCA) algorithm?



Vignesh Natarajan, Machine learning enthusiast

1k Views • Vignesh has 70+ answers in Machine Learning.

What is a principal component? A principal component is *something* in the data set knowing which you can understand most of the data set. If someone asks you to describe a class of 30 girls and ... [\(more\)](#)

Upvote | 29 Downvote Comments 2 Share

Answer written • Data Analysis • 5 Sep

What is a data scientist?



Robert Chang, data curious

2.1k Views • Upvoted by William Chen, [Data Scientist at Quora](#) • Michael Hochster, [Director of Research at Pandora](#)



Thank you all for sharing your definitions, it's really illuminating to learn from so many people's perspectives on this topic. In particular, I really love [Thomson Nguyen's](#) answer on [What is it li...](#) [\(more\)](#)

Upvote | 21 Downvote Comments 1+ Share

Answer written • Data Analysis • 17 Aug

Can I still become a data scientist if I find report writing boring?



Ricardo Vladimiro, Game Analytics and Data Science Lead @ Miniclip

397 Views • Ricardo has 180+ answers in Data Science.

Data scientists "report" in many different forms. The formal one is the reports you mention: a written one with summary, description, conclusions and recommendations. But there are many more:

- dash...

[\(more\)](#)

Upvote | 1 Downvote Comment Share

Answer written • Data Analysis • 6 Sep

What are the biggest differences between time series and non-time series regression?



Jay Verkuilen, PhD Psychometrics UIUC 2007

78 Views • Upvoted by Peter Flom, [Independent statistical consultant for researchers in behav...](#)

Serial dependence in the error process.

Written 6 Sep • [View Upvotes](#)

Upvote | 4 Downvote Comment Share

Answer written • Machine Learning • 2013

What is the difference between Data Analytics, Data Analysis, Data Mining, Data Science, Machine Learning, and Big Data?

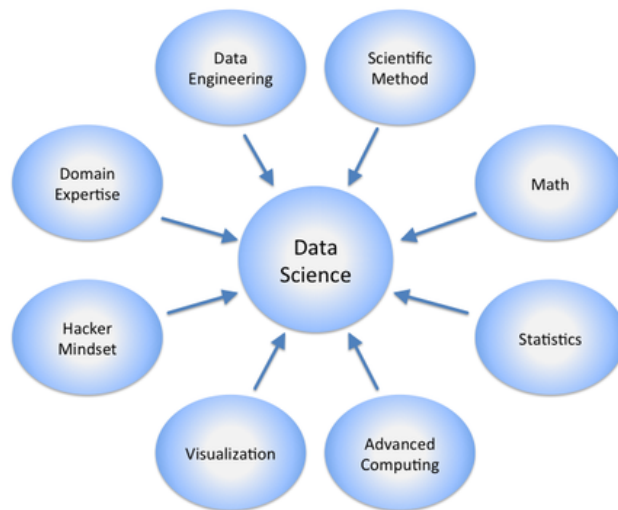


Debidatta Dwivedi

17.5k Views • Upvoted by Jeffrey Wong, [Data Scientist at Netflix](#)

Originally Answered: What are the differences between machine learning and data science?

The following graphic nicely summarizes what all is involved in data science.



(from [Data science](#))

Focus on three bubbles here: **scientific method**, **math** and **statistics**. These are aspects of data science that are closest to **machine learning**.

If I had to summarize machine learning in one sentence, I would say it is a collection of **algorithms** and **techniques** used to design systems that **learn from data**. But the algorithms of ML are very general in the sense usually they have a strong mathematical and statistical basis that **does not** take into account **domain knowledge** and **data pre-processing**. That is the key difference.

If you talk to a data scientist, they would tell you how after acquiring the data and they **cleaned it**([Data cleansing](#)),**transformed it into a useful form** and then using **domain knowledge** decide what statistical method or ML algorithm will best able to solve the problem they are tackling. The above process may require certain amount of 'hacking' skills so as to fasten the process of having meaning data on which processing can be carried out. But a data scientist's job does not end there. **Visualization** is becoming a very important aspect. Representing data in a form which both mere mortals can understand and get valuable insights is as much a science as much as it is art.

So a data-scientist needs to know about how to first decide which method of machine learning will best help him and how to apply that. He does not necessarily need to know how that method works. Although knowing that is always an asset.

There is a nice bit about the difference between ML and data mining on [Machine learning](#):

These two terms are commonly confused, as they often employ the same methods and overlap significantly. They can be roughly defined as follows:

- Machine learning focuses on prediction, based on *known* properties learned from the training data.
- [Data mining](#) (which is the analysis step of [Knowledge Discovery](#) in Databases) focuses on the [discovery](#) of (previously) *unknown* properties on the data.

Written 12 May, 2013 • View Upvotes

[Upvote](#) | **43** [Downvote](#) [Comment](#) **1** [Share](#) ...

Answer written • Startup Founders and E... • 21 Aug

How did Alex Karp start Palantir without a technical background?

Alex Eckelberry, CEO, entrepreneur, investor, board member



3.2k Views

It is not a requirement for a CEO or founder to be a developer. I'm not a developer, yet I've started and run three companies. You don't need to be a dev; you need to be a leader. A strong techni... [\(more\)](#)

Upvote | 6

Downvote Comment Share

...

Answer written • Data Analysis • 27 Aug

I was tasked to build the first data analysis unit in my company. What should I do?



Daniel Dunn

825 Views

Congrats! A big impact you can have is selecting the technology stack your unit will use. The next thing is planning out the roles for the future members of your team; basically, the larger the team the more specialized each role should be. Finally, don't underestimate the importance of project selection. First impressions still matter. Here's a framework:

Early Project Selection Criteria



Written 27 Aug • View Upvotes

Upvote | 7

Downvote Comment 1 Share

...

Answer written • Data Mining • 28 Aug

Does Quora make money from the data collected by its users?



Marc Bodnick, Leads Quora's Business & Community Teams

2.9k Views • Upvoted by Jonathan Brill, [Writer Relations @ Quora](#)

Marc has 105 endorsements in Quora.

No. We don't make any money at all right now, and when we do, it won't be from selling user data.

Written 28 Aug • View Upvotes • Asked to answer by Jonathan Brill

Upvote | 168

Downvote Comments 4+ Share

...

Answer written • Data Analysis • 7 Aug

Is the Zipfian Academy/Galvanize data science boot camp of high quality?



Ricky Kwok, former Data Scientist in Residence

1.2k Views • Upvoted by Yair Livne, [Director of Data Science at Quora](#)

Originally Answered: What is it like to attend Zipfian Academy / Galvanize?



I was part of [Galvanize's Data Science Immersive](#) program January-March 2015; the first cohort after acquiring Zipfian Academy. I technically didn't attend the Bootcamp because I wasn't a student... [\(more\)](#)

Upvote | 18

Downvote

Comments 1+

Share 1

...

Question asked • Data Analysis • 28 Jul

What classic data integration tools would one use to deal with machine data or is this not feasible?

(Data Warehouses)

(IoT, M2M)

Write Answer

Pass

Follow 2

Downvote

...

Answer written • Data Analysis • 2014

What are some of the pitfalls of statistics?



Conner Davis

3k Views • Upvoted by Jason Zhang, [Data Scientist at Quora](#)



It's not clear if you are asking about the pitfalls in calculating statistics or in using them. Or perhaps the pitfalls in hearing them. That said, I'm just going to go in a different direction fr... [\(more\)](#)

Upvote | 39

Downvote

Comments 4

Share

...

Answer written • Data Analysis • 28 Apr

How do you improve your data analysis skills on a daily basis?



Alexander Leeds, Director, Data Science at 1/0 Capital, Founder NYC Data Wranglers, PhD in stuff

2.4k Views • Upvoted by Kevin Lin, [math & stats & software](#)

Alexander has 10+ answers in Big Data.

We are surrounded, even overwhelmed by data you can use to practice your data analysis skills.

In politics and economics:

- Look at the work done by [DataKind](#), much of it with public data.
- Look at [Quarterly Journal of Economics](#) and [Journal of Political Economy](#) for examples of research.

- Check out the St. Louis Federal Reserve's excellent [FRED](#) collection.
- [FiveThirtyEight](#) often uses public data for its analysis of polls and sports.
- Here on Quora, [Global Resource Allocation Initiative \(GRAI\)](#) is building a team, and gathering and working with data you could use.

What types of questions should you answer? Generally, I'd try to balance impact against feasibility, for example:

- Politics: Predict geopolitical outcomes and measure rising economic and political influence (see [The GDELT Project](#)).
- Measure corruption and its complex relationship to development. Need development measures? See: [European Union | Data](#)

In medicine:

- There is fascinating data about doctor and hospital performance available here: [Data.Medicare.gov](#).
- Check out [Metabase](#) - the database of biological databases

What types of questions can you ask? This should be easy: How do doctor and medical institution quality, location, and expense influence patient outcomes? What diseases, treatment and facility characteristics are most closely related to projected cost and severity?

In business:

- Check out [Kaggle](#), [Quandl](#), [Enigma](#), and if you're a student: [Wharton Research Data Services](#).
- Explore the data (currently) available for the [Yelp Dataset Challenge](#).
- Explore media coverage - many news outlets have APIs you can use, see [Times Developer Network - APIs](#),
- To look at engineering practice, you can start with the [GitHub API](#).
- To look at venture capital, start with [CrunchBase Data Hub](#).
- To look at consumer credit, start here [Lending Club](#) data.
- Check out the [Administrative Science Quarterly](#) for business research inspiration.

Questions in business and economics obviously focus on predicting value and predicting change. For example: Scrape property data and do an analysis to predict housing value. Or answer: What are the trends in production or consumption of [X] - some product of interest to you?

Other themes: The [UCI Machine Learning Repository](#) is popular for practice problems.

I believe in the importance of understanding machine learning techniques - but if you're interested in data analysis (not the larger field of data science), start by becoming proficient at econometrics and cross validation. *Really* understand the statistical methods and their implications.

People rightly criticize the [10,000 hour rule](#) - you need 10,000+ hours *and* high-quality hours for mastery! Or, as with engineering, [Ten Years](#). But the emphasis on practice can be misleading: You can produce valuable professional products as you progress.

Updated 6 May • View Upvotes • Asked to answer by Anonymous

Upvote | 24

Downvote Comment Share

...

Answer written • Data Analysis • 13 Aug

Beside recommender systems, what are the typical final products which data scientists could deliver?



William Chen, Data Scientist at Quora

4.4k Views • William has 150+ answers and 60 endorsements in Data Science.

Data Scientists can deliver *insights*, which can help make a product better or a company develop a new feature that can help improve some key metric. Some classic examples:

- A data scientist can anal...

[\(more\)](#)

Upvote | 31

Downvote Comment 1 Share

...

Answer written • Palantir Technologies • 3 Sep

Why is it so hard to get an interview call at Palantir Technologies?



Anonymous

1k Views

Having wasted my time once with Palantir Technologies as it waffled around, failed to align roles with my education and experience, and played hot potato with me as I took time out of my own busy s... [\(more\)](#)

Upvote | 1

Downvote Comment Share

...

Answer written • Data Analysis • 21 Aug

Can I make a career in data analytics on the basis of Tableau skills only?



Vadym Boikov

359 Views

I believe, you cannot have advanced level in Tableau without knowledge of SQL. Doing some advanced calculations in Tableau require understanding of basic principles of databases + writing custom_SQL... [\(more\)](#)

Upvote | 1

Downvote Comment Share

...

Answer written • Machine Learning • 19 Jul

Why is "deep learning" in such demand now?



Bilwaj K Gaonkar, PhD, Bioengineering

10.3k Views • Bilwaj has 10+ answers in Machine Learning.

To understand why deep learning is the best thing since sliced bread read on!

First, let us go back in time to 1958. The Rosenblatt Perceptron algorithm was all the rage. These were times much before machine learning was dubbed 'machine learning' at all. However two scheming scientists, Minsky and Papert's 1969 showed that the XOR problem could not be solved with a perceptron. They also argued that computing power required to sustain a large neural network would be infeasible. For a while this killed the perceptron (and neural networks)

--- But, the saga of (neural) networks was far from over ---

In the late 1980s and early 1990s backpropagation combined with hardware advancements revived the field a decent bit and the neural network came back. But right when sweet revenge was right around the corner! Boser, Cortes and Vapnik published on the "Support Vector Machines". The great appeal of SVMs was that they were developed based on solid theory, unlike neural nets which had taken a somewhat amorphous development path with applications coming before any theoretical basis. The mechanistic appeal of the SVM and its cousins, the so called 'kernel machines' lead to fantastical claims such as we see about deep learning today. The 90s and the 00s brought forth a swirling bubbling amorphous gooey mass of every possible kind of 'kernel machine'. Legions of 'kernel machines' came forth from the abyss, each twisting learning

theory in its own satanic way to put an end to any and all classification, regression and clustering problems anywhere on the face of the earth!

--- But, the saga of (neural) networks was far from over ---

Then, in the early 00's machine learning decided to get eco-friendly! In came the trees. Lots and lots and lots of trees! When trees got one too many, they became , well, rainforests. No wait, they became random forests! Forests were ensembles of trees. The machine learning community jumped on the newest 'green' bandwagon and happily became 'tree-huggers'. After all, trees could solve every problem under the sky. Yes, regression, classification, clustering, abnormality detection, global warming, you name it!

--- But, the saga of (neural) networks was far from over ---

In 2006 Hinton published a fast algorithm for a specific kind of neural net called the restricted boltzmann machine. In came deep learning. Over the next few years this lead to the 'revenge of the neural networks'. Folks who have no idea, decided to just stack up a bunch of RBMs and see what happens. The latest and greatest of the machine learning bandwagon!

Google, or more precisely a bandwagoneer at google decided that it'd be a good idea to stack RBMs into a cluster and run youtube videos through it! A collection of ravishing ill informed media commentators then decided to award this feat undue publicity and academics joined the bandwagon in hopes of getting even more funding for point(less) pursuits. Deep learning networks became even more famous. Thus, we now live in an era of frothy, bubbly deep networks.

Sounds awesome, right? It's not. Just like SVMs, trees and neural networks , this is just another fad.

The fact of the matter is, one cannot simply stack up a bunch of RBMs or SVMs or trees or perceptrons, throw things at 'em and hope that something meaningful comes out.

No machine learning algorithm is a silver bullet. Never has been. What we have is an incredible and relatively mature set of algorithms that are each well suited to specific tasks and specific data. The right way to approach problems is to look at the data, study the underlying assumptions and results behind a wide set of algorithms and then identify a set of algorithms that are likely to 'work' on the problem at hand.

Like every other fad, this too shall pass (see [\[1312.6199\] Intriguing properties of neural networks](#)) and something else will come along. [So get off the deep learning bandwagon and get some perspective](#)

Written 19 Jul • View Upvotes

Upvote | 93

Downvote

Comments 8+

Share 6

...

Answer written • Data Analysis • 2014

How do I become a data scientist without going to college/having a degree?



Katie Kent, Director of Educational Outcomes @ Galvanize; Employee #1 @ Zipfian Academy

5.3k Views • Katie has 50+ answers in Data Science.

The best way to become a data scientist is to learn - and do - data science. There are a many excellent courses and tools available online that can help you get there. Here is an incredible list ... [\(more\)](#)

Upvote | 95

Downvote

Comment

Share 3

...

Answer written • Data Analysis • 15 Aug

Which is better for data analysis: R or Python?

Boxun Zhang, Data Scientist at Spotify; PhD in Computer Science
2.6k Views • Upvoted by Peter Flom, [Independent statistical consultant for researchers in behav...](#)

For data analysis, R is a clear winner.

This is not only because R is designed for statistical computing at the very beginning, but also because there are vast amount of 3rd-party packages available for statistical analysis.

Written 15 Aug • View Upvotes

Upvote | 13

Downvote Comment Share

...

Answer written • Data Analysis • 7 Jul

How can I become a data scientist?

Vik Paruchuri, Founder at dataquest.io
24.5k Views • Upvoted by Ankit Sharma, [Data Science at Snapdeal](#)

I started learning data science about 4 years ago. I had no real programming background. This is mostly geared towards people who are in the same position I was in.

A lot of advice around learning data science starts with "first learn python", or "first take a linear algebra course". This advice is fine, but if I followed it, I never would have learned any data science.

1. Learn to love data

Nobody ever talks about motivation in learning. Data science is a broad and fuzzy field, which makes it hard to learn. Really hard. Without motivation, you'll end up stopping halfway through and believing you can't do it, when the fault isn't with you -- it's with the teaching.

You need something that will motivate you to keep learning, even when it's 1am, formulas are starting to look blurry, and you're wondering if this will be the night that neural networks finally make sense.

You need something that will make you find the linkages between topics like statistics, linear algebra, and neural networks. Something that will prevent you from struggling with the "what do I learn next?" question.

My entry point to data science was predicting the stock market, although I didn't know it at the time. Some of the first programs I coded to predict the stock market involved almost no statistics. But I knew they weren't performing well, so I worked day and night to make them better.

I was obsessed with improving the performance of my programs. I was obsessed with the stock market. I was learning to love data. And because I was learning to love data, I was motivated to learn anything I needed to make my programs better.

Not everyone is obsessed with predicting the stock market, I know. But it's really important to find that thing that make you want to learn.

It can be [figuring out new and interesting things about your city](#), [mapping all the devices on the internet](#), [finding the real positions NBA players play](#), or anything else. The best thing about learning data science is that there are infinite interesting things to work on -- it's all about asking questions and finding a way to get answers.

Take control of your learning by tailoring it to what you want to do, not the

other way around.

2. Learn by doing

Learning about neural networks, image recognition, and other cutting-edge techniques is important. But most data science doesn't involve any of it. Here are some important guidelines:

- 90% of your work will be data cleaning.
- Knowing a few algorithms really well is better than knowing a little about many algorithms. If you know linear regression, k-means clustering, and logistic regression really well, can explain and interpret their results, and can actually complete a data project from start to finish with them, you'll be much more employable than if you know every single algorithm, but can't use them.
- Most of the time, when you use an algorithm, it will be a version from a library (you'll rarely be coding your own SVM implementations -- it takes too long).

What all of this means is that the best way to learn is to work on projects. By working on projects, you gain skills that are immediately applicable and useful. You also have a nice way to build a portfolio.

One technique to start projects is to find a dataset you like. Answer an interesting question about it. Rinse and repeat.

Here are some good places to find datasets to get you started:

- [100+ Interesting Data Sets for Statistics - rs.io](#)
- [Datasets Archive • /r/datasets](#)

Another technique (and my technique) was to find a deep problem, predicting the stock market, that could still be broken down into small, implementable steps. I first connected to the yahoo finance API, and pulled down daily price data. I then created some indicators, like average price over the past few days, and used them to predict the future (note, no real algorithms here, just technical analysis). This didn't work so well, so I learned some statistics, and then used linear regression. Then I connected to another API, scraped minute by minute data, and stored it in a SQL database. And so on, until the algorithm worked well.

The great thing about this is that I had context for my learning. I didn't just learn SQL syntax -- I used it to store price data, and thus learned 10x as much as I would have by just studying syntax. Learning without application isn't going to be retained very well, and won't prepare you to do actual data science work.

3. Learn to communicate insights

Data scientists constantly need to present the results of their analysis to others. Skill at doing this can be the difference between an okay and a great data scientist.

Part of communicating insights is understanding the topic and theory well. Another part is understanding how to clearly organize your results. The final piece is being able to explain your analysis clearly.

It's hard to get good at communicating complex concepts effectively, but here are some things you should try:

- Start a blog. Post the results of your data analysis.

- Try to teach your less tech-savvy friends and family about data science concepts. It's amazing how much teaching can help you understand concepts.
- Try to speak at meetups.
- Use github to host all your analysis.
- Get active on communities like Quora, [DataTau](#) [↗](#), and [/r/machinelearning](#) [↗](#).

4. Learn from peers

It's amazing how much you can learn from working with others. In data science, teamwork can also be very important in a job setting.

Some ideas here:

- Find people to work with at meetups.
- Contribute to open source packages.
- Message people who write interesting data analysis blogs seeing if you can collaborate.
- Try out [Kaggle](#) [↗](#) and see if you can find a teammate.

5. Constantly increase the degree of difficulty

Are you completely comfortable with the project you're working on? Was the last time you used a new concept a week ago? It's time to work on something more difficult. Data science is a steep mountain to climb, and if you stop climbing, it's easy to never make it.

If you find yourself getting too comfortable, here are some ideas:

- Work with a larger dataset. Learn to use spark.
- See if you can make your algorithm faster.
- How would you scale your algorithm to multiple processors? Can you do it?
- Try to teach a novice to do the same things you're doing now.

The bottom line

This is less a roadmap of exactly what to do that it is a rough set of guidelines to follow as you learn data science. If you do all of these things well, you'll find that you're naturally developing data science expertise.

I generally dislike the "here's a big list of stuff" approach, because it makes it extremely hard to figure out what to do next. I've seen a lot of people give up learning when confronted with a giant list of textbooks and MOOCs.

I personally believe that anyone can learn data science if they approach it with the right frame of mind.

I'm also the founder of [dataquest.io](#) [↗](#), a site that helps you learn data science in your browser. It encapsulates a lot of the ideas discussed in this post to create a better learning experience. You learn by analyzing interesting datasets like CIA documents and NBA player stats. It's not a problem if you don't know how to code -- we teach you python. We teach python because it's the most beginner-friendly language, is used in a lot of production data science work, and can be used for a variety of applications.

Some helpful resources

As I worked on projects, I found these resources helpful. Remember, resources on their own aren't useful -- find a context for them:

- [Khan Academy](#) -- good basic statistics and linear algebra content.
- [Introduction to Linear Algebra, 4th Edition](#) -- Great linear algebra book by Gilbert Strang.
- [Textbook | Calculus Online Textbook | MIT OpenCourseWare](#) -- also by Gilbert Strang, great calculus book.
- [data mining, inference, and prediction. 2nd Edition](#) -- Elements of statistical learning, a good machine learning book.
- [Andrew Ng's Online Machine Learning Class](#) -- the original coursera class.
- [OpenIntro Statistics](#) -- Good basic stats book.
- <https://scholar.google.com> -- A paper can be a great way to learn about a topic. For example, here's Breiman's original random forest paper: <http://link.springer.com/article...>

Written 7 Jul • View Upvotes

Upvote | 369

Downvote

Comments 7+

Share 15

...

Question asked • Predictive Analytics • 23 Jul

Did Bret Easton Ellis predict the future with 'American Psycho'?

Write Answer

Pass

Follow 2

Downvote

...

Answer written • Data Analysis • Sat

How can Python and R be used in big data analytics?



Yassine Alouini, I write data science pipelines.

423 Views • Yassine has 40+ answers in Data Science.

Spark is a great tool for working with large-scale data sets. As it is stated in the website:

... (more)

Upvote | 1

Downvote

Comment

Share

...

Answer written • Data Analysis • 2012

As a data scientist, how do you start investigating a data set?



Bryan Taylor, Principal Architect at Rackspace, data science.

2k Views • Upvoted by Yuval Feinstein, [Algorithmic Software Engineer in NLP, IR and Machine Learning](#)

For a long time when I start, I'm just thinking about data quality, cleansing, fitness for purpose, and proper interpretation. I start by looking at each data element. What does it mean? How was i... (more)

Upvote | 52

Downvote

Comment 1

Share 1

...

Answer written • Data Analysis • 2015

What are 20 questions to detect fake data scientists?



Kavita Ganesan

4.8k Views • Upvoted by Jay Verkuilen, [Associate Professor of Psychometrics and Public Health, CUN...](#)

These are not code related but this is based on my personal experience with dealing with some bogus cases:

1. What is a gold standard ? Believe it or not there are data scientists (even at very seni...

[\(more\)](#)

Upvote | 32

Downvote Share 1

...

Answer written • Machine Learning • 26 Aug

Is a Ph.D. necessary for a job in Machine Learning?

Peter Johnston, I run DataScience Oxford and GDG Oxford (GDG is Google Developer Group)

9k Views • Peter has 30+ answers in Data Science.

In your position a PhD would be the kiss of death to your career in machine learning and data mining. At the moment a dog with dyslexia could get a job in Machine Learning if they had the basic sk... [\(more\)](#)

Upvote | 32

Downvote Comments 2+ Share 1

...

Answer written • Data Analysis • 23 Aug

Why is Python used heavily in big data when it has issues with multithreading due to the Global Interpreter Lock (GIL)?

Alexey Grigorev

5k Views • Alexey has 30+ answers in Big Data.

What makes you feel that python is used heavily for Big Data? It's not. Yes, it is true that there are python wrappers for Spark, Flink and other distributed computing engines. But they are just ... [\(more\)](#)

Upvote | 16

Downvote Comment Share

...

Answer written • Palantir Technologies • 24 Jul

What is the better place to work, Palantir or Facebook?

Anonymous

5.8k Views • Upvoted by Alex Moore, [Co-Founder NodePrime](#), 1st employee [Palantir](#), [angel](#)

I've interned at both places circa 2014-2015 era. Each has their pros and cons and it depends on how much you want to weight each one. Facebook's largest

Quora

Q Search



Home



Write

Notifications ²

Dee



Ask Question

Answer written • Data Analysis • 2014

Correlation can measure only the linear relationship between variables. What are the methods for measuring non-linear relationships between two variables?

Michael Hochster, PhD in Statistics, Stanford; Director of Research, Pandora

7.6k Views • Michael has 270+ answers and 12 endorsements in Statistics (academic discipline).

"Correlation can measure only the linear relationship between two variables" seems a bit strong:

```
1 >x <- 1:99
2 > cor(x, x^2)
3 [1] 0.9688607
4 > cor(x, x^3)
5 [1] 0.9175626...
```

[\(more\)](#)

Upvote | 51

Downvote Comments 3 Share

...

Answer written • Data Analysis • 2014

What are the downsides of being a data scientist?

Michiel Van Herwegen, Data Scientist at Virdata

15.2k Views • Upvoted by Abhinav Sharma, [I worked on data at Facebook](#)

Michiel has 10+ answers in Data Science.

I believe there are 3 downsides worth mentioning, two of them being touched upon already by other answers. [1. The waiting game](#) The first issue is that in many cases, you have to wait. A lot. B... [\(more\)](#)

Upvote | 126

Downvote Comments 2 Share 2

...

Answer written • Data Analysis • 9 Sep

How could I get hired as a big data scientist?



Roman Trusov, I teach machines

1k Views • Roman has 30+ answers in Data Science.

Well, there's always an option to apply for a job and get hired. That, actually, seems like the easiest way, the amount of jobs is very high. Let's have a look at other ways.

1. Become a Kaggle Master...

[\(more\)](#)

Upvote | 7

Downvote Comment 1 Share

...

Question followed • Data Analysis • 4 Aug

What statistical analysis is done in the SaaS platform?

Write Answer

Pass

Follow 3 Downvote

...

Answer written • Data Analysis • 5 Sep

How do I read this regression analysis result on finance?



Lee Witt, Doctorate in statistics

179 Views

Really, none of those models seems to be very good. You have a sample of 33 observations, at least 8 predictors, and the smallest p-value is essentially .05? (Not to mention the uniformly low adjus... [\(more\)](#)

Upvote | 1

Downvote Comment Share

...

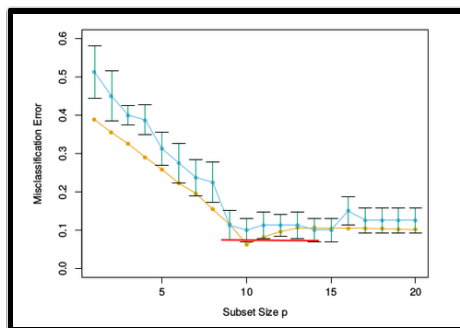
Answer written • Machine Learning • 26 Jun

For K-fold cross validation, what k should be selected?



Charles Yang Zheng, grad student in statistics

1.2k Views • Charles has 40+ answers in Statistics (academic discipline).



Warren's answer covers the basics: lower K = cheaper, less variance, more bias, while higher K = more expensive, more variance, and lower bias. I would add that you can reduce variance without inc... [\(more\)](#)

Upvote | 10

Downvote Comment Share

...

Answer written • Data Analysis • 2014

In what ways is advanced statistics or machine learning used in basketball analysis?



Rajiv Maheswaran, CEO, Second Spectrum. Research Assistant Professor, USC Computer Science Dept.

1.6k Views



There's quite a bit of advanced statistics and machine learning applied in basketball today primarily due to the emergence of player and ball tracking data. My colleague, Professor Yu-Han Chang an... [\(more\)](#)

Upvote | 24

Downvote

Comments 1+

Share 1

...

Answer written • Business Strategy • 18 Aug

I heard a rumor that Palantir is having an extremely high turnover right now. Is that true?



Jason M. Lemkin, Partner @ Storm Ventures

13.5k Views • Upvoted by Marc Bodnick, [Leads Quora business & community teams](#)

Jason has 100+ answers in Business.

Don't listen to these rumors -- even if they are true. The reality is B2B companies go through higher churn than you might expect as they go through phases. The team from \$0-\$1m in ARR often does... [\(more\)](#)

Upvote | 44

Downvote

Comment

Share 1

...

Answer written • Data Analysis • 6 Sep

How can I avoid overfitting?



Firdaus Janoos, 25+ publications in machine learning and computer vision

1.8k Views • Upvoted by Peter Flom, [Independent statistical consultant for researchers in behav...](#)

If your aim is prediction (as is typical in machine learning) rather than model fitting / parameter testing (as is typical in classical statistics) - then in addition to the excellent answers provi... [\(more\)](#)

Upvote | 12

Downvote

Comments 2+

Share

...

Answer written • Machine Learning • 3 Sep

What are the strengths and weaknesses of the machine learning / data analytic products you have worked with? How would you compare them?



Ricardo Vladimiro, Game Analytics and Data Science Lead @ Miniclip

932 Views • Upvoted by Patrick Hall, [Cloudera certified data scientist, CCP-DS: 11](#)

Ricardo has 180+ answers in Data Science.

This is an interesting question, thank you for the A2A. I'll give my view on R, Python, Redshift, Pig and AWS. R is my biggest hammer at the moment. The pros:

- The ability to produce enterprise lev...

[\(more\)](#)

Upvote | 9

Downvote

Comment 1

Share

...

Answer written • Infographics • 6h

What are the worst infographics you've ever seen?



Andy Xue, Head Tutor and Ph.D. Candidate in Computer Science

21 Views



These infographics below are all taken from the book [How to Lie with Statistics](#) (it's an excellent book). -----

----- The following plot i... [\(more\)](#)

Upvote | 2

Downvote Comment Share

...

Answer written • Data Analysis • 3 Sep

Can I get a job as a data analyst or data scientist at the fresher level? Do companies recruit freshers for such a job description?



Tanya Zyabkina, does analog analytics

235 Views

Many companies recruit for entry level Analyst positions. These may not be in the "data science" camp, but working in one you learn how to work with data and how to think about a business. Both are... [\(more\)](#)

Upvote | 1

Downvote Comment Share

...