

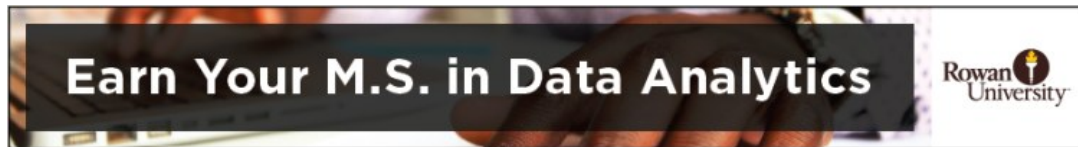
KDnuggets

Subscribe to [KDnuggets News](#) | [Twitter](#) | [Facebook](#) | [LinkedIn](#) | [Contact](#)

search KDnuggets Search



- [SOFTWARE](#)
- [NEWS](#)
- [Top stories](#)
- [Opinions](#)
- [Tutorials](#)
- [JOBS](#)
- [Companies](#)
- [Courses](#)
- [Datasets](#)
- [EDUCATION](#)
- [Certificates](#)
- [Meetings](#)
- [Webinars](#)



[Earn Your MS in Data Analytics](#)

[KDnuggets Home](#) » [News](#) » [2017](#) » [Feb](#) » [Tutorials, Overviews](#) » 17 More Must-Know Data Science Interview Questions and Answers ([17:n07](#))

17 More Must-Know Data Science Interview Questions and Answers



[Previous post](#)
[Next post](#)

Like 210 Share 210 Share 582 Tweet G+1 14 Share 302

Tags: [Anomaly Detection](#), [Bias](#), [Classification](#), [Data Science](#), [Donald Trump](#), [Interview questions](#), [Outliers](#), [Overfitting](#), [Variance](#)

17 new must-know Data Science Interview questions and answers include lessons from failure to predict 2016 US Presidential election and Super Bowl LI comeback, understanding bias and variance, why fewer predictors might be better, and how to make a model more robust to outliers.



[NYU MS in Business Analytics](#)

[for Professionals - apply now](#)

By [Gregory Piatetsky](#), KDnuggets.

[comments](#)

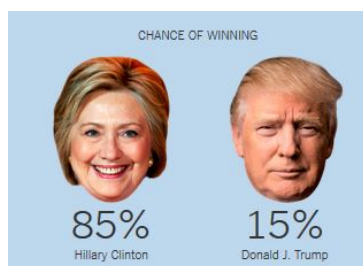
The post [21 Must-Know Data Science Interview Questions and Answers](#) was the [most viewed post of 2016](#), with over 250,000 page views. For 2017, KDnuggets Editors bring you 17 more new and important Data Science Interview Questions and Answers. Because some of the answers are quite lengthy, we will publish them in 3 parts over 3 weeks. This is part 1, which answers the 6 questions below. Here is [part 2](#) and [part 3](#).

This post answers questions:

- Q1. What are Data Science lessons from failure to predict 2016 US Presidential election (and from Super Bowl LI comeback)
- Q2. What problems arise if the distribution of the new (unseen) test data is significantly different than the distribution of the training data?
- Q3. What are bias and variance, and what are their relation to modeling data?
- Q4. Why might it be preferable to include fewer predictors over many?
- Q5. What error metric would you use to evaluate how good a binary classifier is? What if the classes are imbalanced? What if there are more than 2 groups?
- Q6. What are some ways I can make my model more robust to outliers?

Q1. What are Data Science lessons from failure to predict 2016 US Presidential election (and from Super Bowl LI comeback)

[Gregory Piatetsky](#) answers:



Just before the Nov 8, 2016 election, most pollsters gave Hillary Clinton an edge of ~3% in popular vote and 70-95% chance of victory in electoral college. Nate Silver's FiveThirtyEight had the highest chances of Trump Victory at ~30%, while New York Times Upshot and Princeton Election Consortium estimated only ~15%, and other pollsters like Huffington Post gave Trump only 2% chance of victory. Still, Trump won. So what are the lessons for Data Scientists?

To make a statistically valid prediction we need

- 1) enough historical data and
- 2) assumption that past events are sufficiently similar to current event we are trying to predict.

Events can be placed on the scale from deterministic (2+2 will always equal to 4) to strongly predictable (e.g. orbits of planets and moons, avg. number of heads when tossing a fair coin) to weakly predictable (e.g. elections and sporting events) to random (e.g. honest lottery).

If we toss a fair coin 100 million times, we have the expected number of heads (mean) as 50 million, the standard deviation = 10,000 (using formula $0.5 * \sqrt{N}$), and we can predict that 99.7% of the time the expected number of heads will be within [3 standard deviations](#) of the mean.

But using polling to predict the votes of 100 million people is much more difficult. Pollsters need to get a representative sample, estimate the likelihood of a person actually voting, make many justified and unjustified assumptions, and avoid following their conscious and unconscious biases.

In the case of US Presidential election, correct prediction is even more difficult because of the antiquated Electoral college system when each state (except for Maine and Nebraska) awards the winner all its votes in the electoral college, and the need to poll and predict results for each state separately.

The chart below shows that in 2016 US presidential elections pollsters were off the mark in many states. They mostly underestimated the Trump vote, especially in 3 critical states of Michigan, Wisconsin, and Pennsylvania which all flipped to Trump.

Where did polls miss the most?

STATE	CLINTON'S PROJECTED MARGIN		
	ADJUSTED POLLING AVERAGE	BASED ON VOTES SO FAR	SHIFT
Maine	+6.9	+2.6	-4.3
Minnesota	+5.9	+0.7	-5.2
Virginia	+5.4	+2.8	-2.6
Wisconsin	+5.4	-2.2 =>T	-7.6
New Mexico	+5.3	+7.8	+2.5
Michigan	+4.0	-0.8 =>T	-4.8
Colorado	+3.8	+4.9	+1.1
Pennsylvania	+3.7	-0.8 =>T	-4.5
New Hampshire	+3.5	+0.4	-3.1
Nevada	+0.7	+1.3	+0.6
North Carolina	+0.7	-3.8 =>T	-4.5
Florida	+0.6	-1.4	-2.0
Ohio	-2.0	-8.1	-6.1
Arizona	-2.4	-5.0	-2.6
Iowa	-3.4	-9.5	-6.1
Georgia	-4.0	-5.1	-1.1

SOURCE: FIVETHIRTYEIGHT POLLS-ONLY ADJUSTED POLLING AVERAGE; NEW YORK TIMES UPSHOT PROJECTIONS

Source: [@NateSilver538](#) tweet, Nov 9, 2016.

A few statisticians like Salil Mehta [@salilstatistics](#) were warning about [unreliability of polls](#), and David Wasserman of 538 actually described this scenario in Sep 2016 [How Trump Could Win The White House While Losing The Popular Vote](#), but most pollsters were way off.

So a good lesson for Data Scientists is to **question their assumptions** and to be very skeptical when predicting a weakly predictable event, especially when based on human behavior.

Other important lessons are

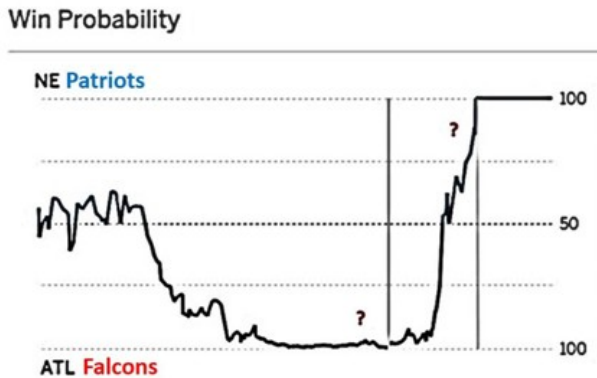
- Examine data quality - in this election polls were not reaching all likely voters
- Beware of your own biases: many pollsters were likely Clinton supporters and did not want to question the results that favored their candidate. For example, Huffington Post had forecast over 95% chance of Clinton Victory.

See also other analyses of 2016 polling failures:

- Wired: [Trump's Win Isn't the Death of Data—It Was Flawed All Along](#).
- NYTimes [How Data Failed Us in Calling an Election](#)
- Datanami [Six Data Science Lessons from the Epic Polling Failure](#)
- InformationWeek [Trump's Election: Poll Failures Hold Data Lessons For IT](#)
- [Why I Had to Eat a Bug on CNN](#), by Sam Wang, Princeton, whose Princeton Election Consortium gave Trump 15% to win.

(Note: this answer is based on a previous KDnuggets post: <http://www.kdnuggets.com/2016/11/trump-shows-limits-prediction.html>)

We had another example of statistically very unlikely event happen in Super Bowl LI on Feb 5, 2017. After the half time, Atlanta Falcons were leading 21:3 after halftime and 28:9 after 3rd quarter. ESPN estimated Falcons win probability at that time at almost 100%.



(reference: Salil Mehta tweet [Salil Mehta tweet, Feb 6, 2017](#))

Never before has a team lost a Super Bowl after holding such advantage. However, each Super Bowl is different, and this one was turned out to be very different. Combination of superior skill (Patriots, after all, were favorites before the game) and luck (e.g. a very lucky catch by Julian Edelman in 4th quarter, Patriots winning coin toss in overtime) gave victory to Pats.

This Super Bowl was another good lesson for Data Scientists of danger of having **too much confidence** when predicting weakly predictable events. You need to understand the risk factors when dealing with such events, and try to avoid using probabilities, or if you have to use numbers, have a wide confidence range.

Finally, if the odds seem to be against you but the event is only weakly predictable, go ahead and do your best - sometimes you will be able to beat the odds.

Q2. What problems arise if the distribution of the new (unseen) test data is significantly different than the distribution of the training data?

[Gregory Piatetsky](#) and [Thuy Pham](#) answer:

The main problem is that the predictions will be wrong !

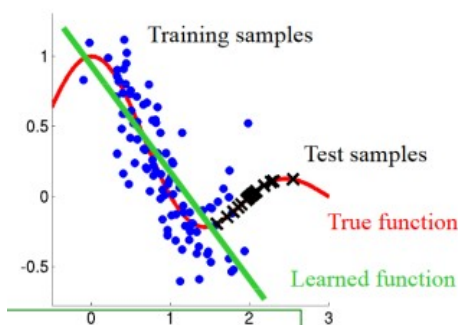
If the new test data is sufficiently different in key parameters of the prediction model from the training data, then predictive model is no longer valid.

The main reasons this can happen are sample selection bias, population drift, or non-stationary environment.

a) Sample selection bias

Here the data is static, but the training examples have been obtained through a biased method, such as non-uniform selection or non-random split of data into train and test.

If you have a large static dataset, then you should randomly split it into train/test data, and the distribution of test data should be similar to training data.



b) Covariate shift aka population drift

Here the data is not static, with one population used as a training data, and another population used for testing.

(Figure from <http://iwann.ugr.es/2011/pdf/InvitedTalk-FHerrera-IWANN11.pdf>).

Sometimes the training data and test data are derived via different processes - eg a drug tested on one population is given to a new population that may have significant differences. As a result, a classifier based on training data will perform poorly.

One proposed solution is to apply a statistical test to decide if the probabilities of target classes and key variables used by the classifier are significantly different, and if they are, to retrain the model using new data.

c) Non-stationary environments

Training environment is different from the test one, whether it's due to a temporal or a spatial change.

This is similar to case b, but applies to situation when data is not static - we have a stream of data and we periodically sample it to develop predictive models of future behavior. This happens in adversarial classification problems, such as spam filtering and network intrusion detection, where spammers and hackers constantly change their behavior in response. Another typical case is customer analytics where customer behavior changes over time. A

telephone company develops a model for predicting customer churn or a credit card company develops a model to predict transaction fraud. Training data is historical data, while (new) test data is the current data.

Such models periodically need to be retrained and to determine when you can compare the distribution of key variables in the predictive model in the old data (training set) and the new data, and if there is a sufficiently significant difference, the model needs to be retrained.

For a more detailed and technical discussion, see references below.

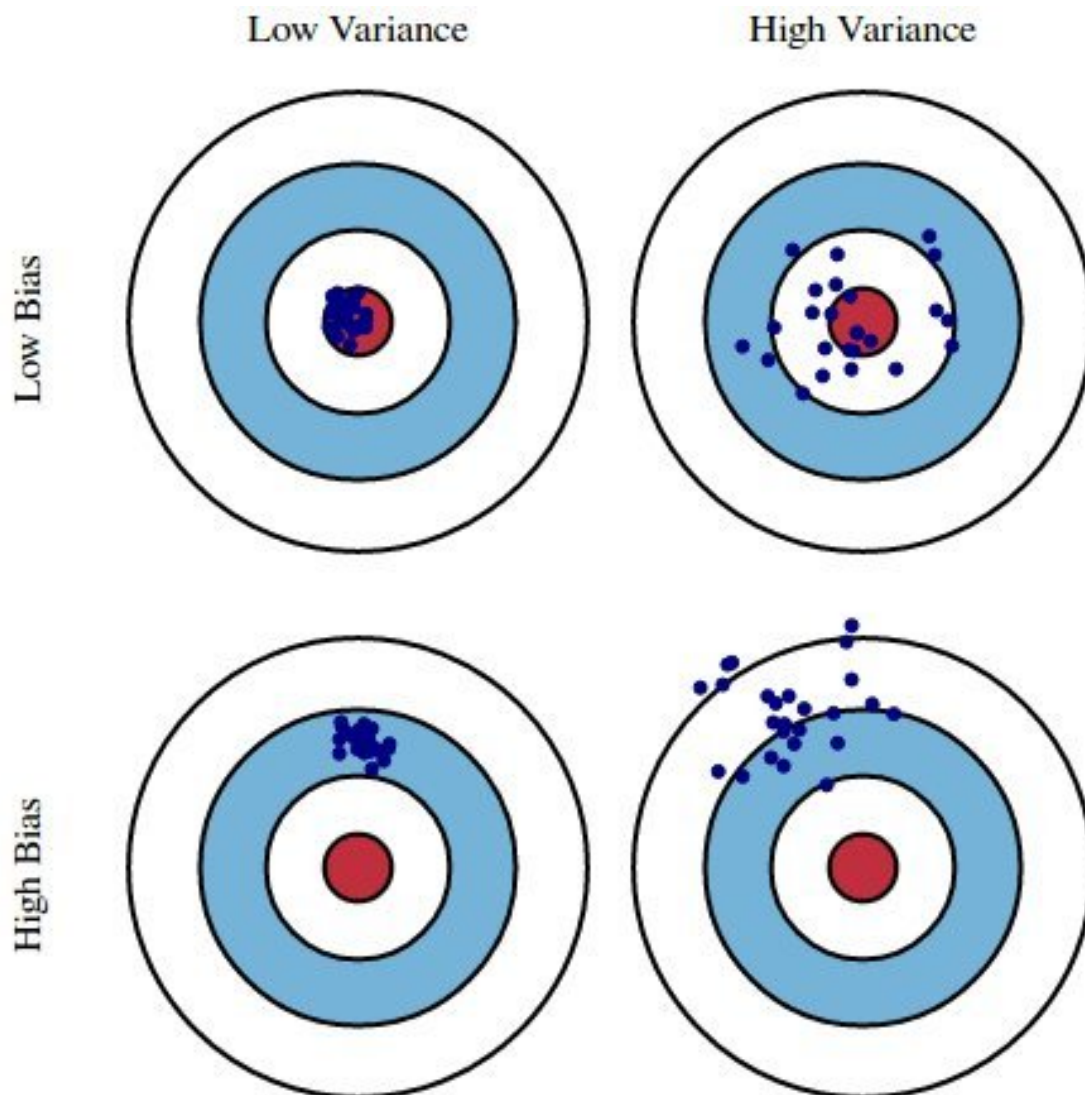
References:

- [1] Marco Saerens, Patrice Latinne, Christine Decaestecker: Adjusting the Outputs of a Classifier to New a Priori Probabilities: A Simple Procedure. Neural Computation 14(1): 21-41 (2002)
- [2] Machine Learning in Non-stationary Environments: Introduction to Covariate Shift Adaptation, Masashi Sugiyama, Motoaki Kawanabe, MIT Press, 2012, ISBN 0262017091, 9780262017091
- [3] Quora answer to [What could be some issues if the distribution of the test data is significantly different than the distribution of the training data?](#)
- [4] [Dataset Shift in Classification: Approaches and Problems](#), Francisco Herrera invited talk, 2011.
- [5] [When Training and Test Sets are Different: Characterising Learning Transfer](#), Amos Storkey, 2013.

Q3. What are bias and variance, and what are their relation to modeling data?

[Matthew Mayo](#) answers:

Bias is how far removed a model's predictions are from correctness, while **variance** is the degree to which these predictions vary between model iterations.

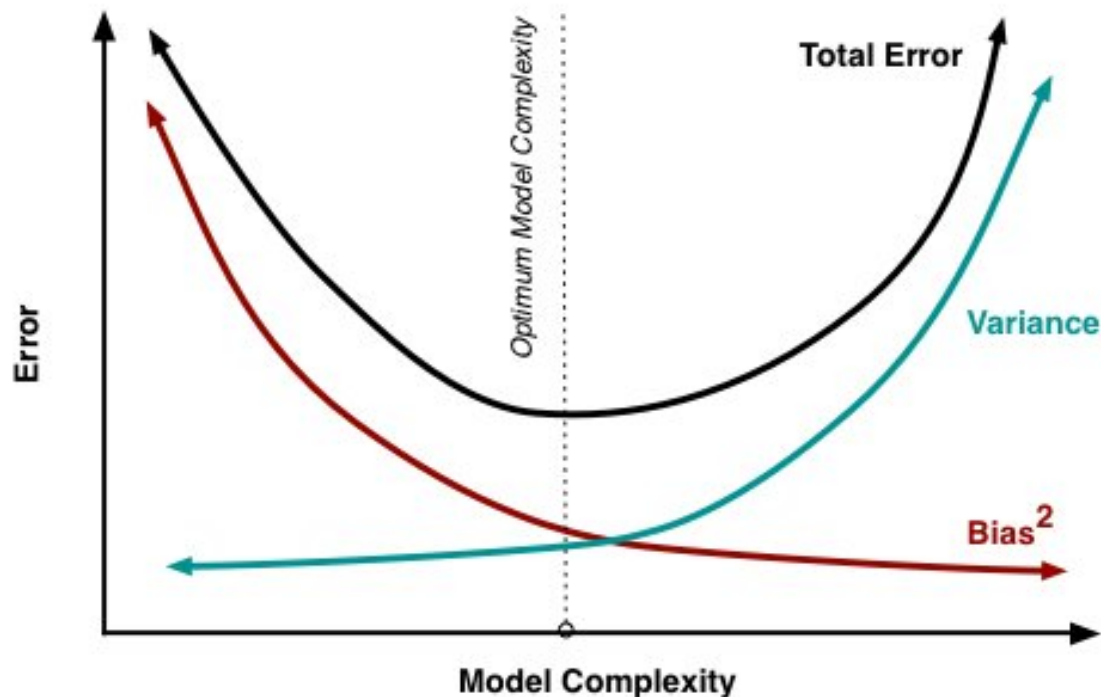


Bias vs Variance, [Image source](#)

[As an example](#), using a simple flawed Presidential election survey as an example, errors in the survey are then explained through the twin lenses of bias and variance: selecting survey participants from a phonebook is a source of bias; a small sample size is a source of variance.

Minimizing total model error relies on the balancing of bias and variance errors. Ideally, models are the result of a collection of **unbiased data of low variance**. Unfortunately, however, the more complex a model becomes, its tendency is toward less bias but greater variance; therefore an optimal model would need to consider a balance between these 2 properties.

The statistical evaluation method of cross-validation is useful in both demonstrating the **importance** of this balance, as well as actually **searching** it out. The number of data folds to use -- the value of k in k -fold cross-validation -- is an important decision; the lower the value, the higher the bias in the error estimates and the less variance.



Bias and variance contributing to total error, [Image source](#) Conversely, when k is set equal to the number of instances, the error estimate is then very low in bias but has the possibility of high variance.

The most important takeaways are that bias and variance are two sides of an important trade-off when building models, and that even the most routine of statistical evaluation methods are directly reliant upon such a trade-off.

On next page, we answer

- Why might it be preferable to include fewer predictors over many?
- What error metric would you use to evaluate how good a binary classifier is?
- What are some ways I can make my model more robust to outliers?

Pages: 1 [2](#)

[◀ Previous post](#)
[Next post ▶](#)

Top Stories Past 30 Days

Most Popular

1. [The 10 Algorithms Machine Learning Engineers Need to Know](#)
2. [7 More Steps to Mastering Machine Learning With Python](#)
3. [An Overview of Python Deep Learning Frameworks](#)
4. [Gartner 2017 Magic Quadrant for Data Science Platforms: gainers and losers](#)
5. [Every Intro to Data Science Course on the Internet, Ranked](#)
6. [17 More Must-Know Data Science Interview Questions and Answers](#)
7. [50 Companies Leading The AI Revolution, Detailed](#)

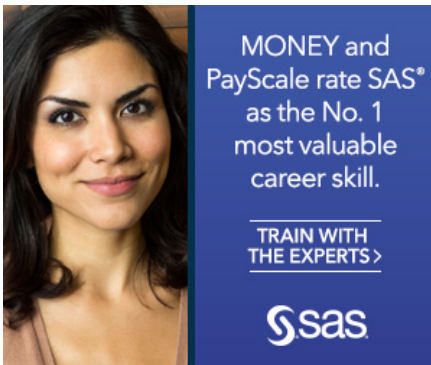
Most Shared

1. [50 Companies Leading The AI Revolution, Detailed](#)
2. [An Overview of Python Deep Learning Frameworks](#)
3. [7 More Steps to Mastering Machine Learning With Python](#)
4. [The Data Science Project Playbook](#)
5. [What Is Data Science, and What Does a Data Scientist Do?](#)
6. [What makes a good data visualization – a Data Scientist perspective](#)
7. [Getting Up Close and Personal with Algorithms](#)

Latest News

- [Live.xyz: Senior Data Scientist](#)
- [Webinar: Improve Your CLASSIFICATION with CAR...](#)
- [Top Stories, Mar 20-26: What Is Data Science,...](#)

- [From Big Data Platforms to Platform-less Mach...](#)
- [Turner: Sr. Operations Research Analyst / Dat...](#)
- [What is Structural Equation Modeling?](#)



TDWI Chicago May 7-12
Earn Your Data Modeling Certificate



Register by April 7 and save 10%

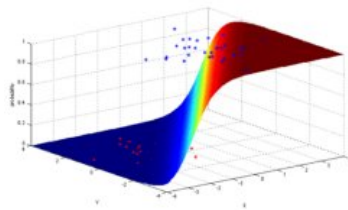
tdwi.org/chicago 

[TDWI Chicago - Earn Data Modeling Certificate](#)
[Reg. by Apr 7 and save 10%](#)

Top Stories Last Week

[Most Popular](#)

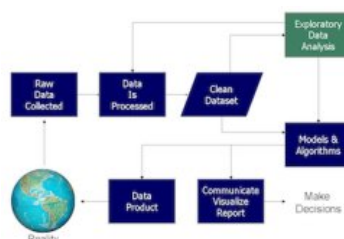
1.  [The 10 Algorithms Machine Learning Engineers Need to Know](#)



2.  [50 Companies Leading The AI Revolution, Detailed](#)
3.  [17 More Must-Know Data Science Interview Questions and Answers](#)
4.  [Getting Started with Deep Learning](#)
5.  [The Most Underutilized Function in SQL](#)
6.  [Analytics 101: Comparing KPIs](#)
7.  [How to think like a data scientist to become one](#)

[Most Shared](#)

1. [What Is Data Science, and What Does a Data Scientist Do?](#)



2. [Getting Up Close and Personal with Algorithms](#)
3. [Getting Started with Deep Learning](#)

4. [Key Takeaways from Strata + Hadoop World 2017 San Jose](#)
5. [How to think like a data scientist to become one](#)
6. [The Most Underutilized Function in SQL](#)
7. [Data Scientists Might Have It Made For 2017](#)

More Recent Stories

- [What is Structural Equation Modeling?](#)
- [Last Chance: Big Data San Francisco, April 19 & 20](#)
- [Analytics and Machine Learning training in Q2](#)
- [Turner: Advisor Analytics Architect](#)
- [Apple: Software Engineer – Local Search](#)
- [PeachIH: Data Scientist, Machine Learning Engineer](#)
- [Getting Started with Deep Learning](#)
- [IBM Chief Data Officer Strategy Summit, March 29-30, San Franc...](#)
- [Key Takeaways from Strata + Hadoop World 2017 San Jose, Day 1](#)
- [Unsupervised Investments: A Comprehensive Guide to AI Investors](#)
- [Make Analytics Pay with Live Immersive Training](#)
- [How to think like a data scientist to become one](#)
- [What Is Data Science, and What Does a Data Scientist Do?](#)
- [Apache Big Data: top projects, people and technologies —...](#)
- [Top tweets, Mar 15-21: Reverse-engineering a \\$500M AI compa...](#)
- [Eindhoven University of Technology: Full Professor Database Te...](#)
- [Webinar: Predictive Analytics: Failure to Launch – Apr 13](#)
- [Deep Learning Summit & Deep Learning in Healthcare Summit...](#)
- [What Top Firms Ask: 100+ Data Science Interview Questions](#)
- [Why A/B Testers Have The Best Jobs In Tech](#)

[KDnuggets Home](#) » [News](#) » [2017](#) » [Feb](#) » [Tutorials, Overviews](#) » 17 More Must-Know Data Science Interview Questions and Answers ([17:n07](#))

© 2017 KDnuggets. [About KDnuggets](#)

[Subscribe to KDnuggets News](#)



X