* Business
  + How would you explain to the senior management in your organization as to why a particular data set is important?
  + Suppose you are given a data set, what will you do with it to find out if it suits the business needs of your project or not.
  + What is the CRISP-DM (Cross Industry Standard Process for Data Mining) and what are the steps?
* [Linear Algebra](http://www.itshared.org/2015/10/data-science-interview-questions.html#linear-algebra)
  + What is an Eigenvalue and Eigenvector?
  + What do you understand by feature vectors?
  + What Is Vector Space Model and its use?
* [Probability and Statistics](http://www.itshared.org/2015/10/data-science-interview-questions.html#probability-and-statistics)
  + [Distributions](http://www.itshared.org/2015/10/data-science-interview-questions.html#distributions)
    - Describe a non-normal probability distribution and how to apply it.
    - Do you know the Dirichlet distribution? the multinomial distribution?
    - 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?
    - Give examples of data that does not have a Gaussian distribution, or log-normal.
    - How would you check if a distribution is close to Normal? Why would you want to check it? What is a QQ Plot?
    - How would you find an anomaly in a distribution?
    - How would you know when Gaussian Mixture Model is applicable?
    - What is the difference between skewed and uniform distribution?
  + [Basic Statistics](http://www.itshared.org/2015/10/data-science-interview-questions.html#basic-statistics)
    - You applied ANOVA and it says that the means are different. How do you identify the populations where the differences are significant?
    - Are expected value and mean value different?
    - Do you know what Type-I/Type-II errors are?
    - What are p-values and confidence intervals?
    - What do you do when n is small? How do you quantify uncertainty? Pick one strategy and explain how to make decisions under uncertainty?
    - What is the Central Limit Theorem and why is it important in data science?
    - What is the distribution of p-value’s, in general?
    - What is the normal distribution? Give an example of some variable that follows this distribution
    - What is t-Test/F-Test/ANOVA? When to use it?
    - What summary statistics do you know?
    - How would you calculate needed sample size?
    - how would you calculate the degrees of freedom of an interaction
    - How would you find the median of a very large dataset?
    - How would you measure distance between data points?
    - How would you remove multicollinearity?
    - What is collinearity and what to do with it?
    - What is the difference between squared error and absolute error?
    - What is the null hypothesis? How do we state it?
    - When do we need the intercept term and when do we not?
  + [Experiment Design](http://www.itshared.org/2015/10/data-science-interview-questions.html#experiment-design)
    - Why is randomization important in experimental design?
    - What are confounding variables?
    - What is the importance of having a selection bias?
    - When you sample, what bias are you inflicting?
    - Why do we need hypothesis testing?
    - Why do we need to sample and how?
    - How would you use resampling for hypothesis testing? Have you heard of Permutation Tests?
    - Some 3rd party organization randomly assigned people to control and experiment groups. How can you verify that the assignment truly was random?
    - What does it mean (practically) for a design matrix to be “ill-conditioned”?
    - What is an interaction?
    - What is Power analysis?
    - How would you test if two populations have the same mean? What if you have 3 or 4 populations?
  + [A/B Tests](http://www.itshared.org/2015/10/data-science-interview-questions.html#ab-tests)
    - How will you explain an A/B test to an engineer who does not know statistics?
    - What is the goal of A/B Testing?
    - How would you prove that one improvement you’ve brought to an algorithm is really an improvement over not doing anything? How familiar are you with A/B testing?
    - How is A/B testing different from usual Hypothesis testing?
  + [Bayesian Statistics](http://www.itshared.org/2015/10/data-science-interview-questions.html#bayesian-statistics)
    - How would you use Naive Bayes classifier for categorical features? What if some features are numerical?
    - Is Naïve Bayes bad? If yes, under what aspects.
    - What do you understand by conjugate-prior with respect to Naïve Bayes?
    - What is the difference between Bayesian Estimate and Maximum Likelihood Estimation (MLE)?
    - Why is naive Bayes so ‘naive’?
    - Explain prior probability, likelihood and marginal likelihood in context of Naïve Bayes algorithm?
  + [Time Series](http://www.itshared.org/2015/10/data-science-interview-questions.html#time-series)
    - Can you use machine learning for time series analysis?
    - How can you deal with different types of seasonality in time series modelling?
    - How would you apply resampling to time series data?
    - What are some different Time Series forecasting techniques?
    - How you can make data normal using Box-Cox transformation?
* Data Wrangling/Cleaning/Quality
  + Differentiate between wide and tall data formats?
  + How can outlier values be treated?
  + How do data management procedures like missing data handling make selection bias worse?
  + How would you deal with outliers in your data?
  + How would you go about investigating if a certain trend in a distribution is due to an anomaly?
  + How would you handle missing data?
  + If you’re attempting to predict a customer’s gender, and you only have 100 data points, what problems could arise?
  + What are your favorite imputation techniques to handle missing data?
  + How would you convert categorical variables to numerical for extracting features?
  + What are the steps for wrangling and cleaning data before applying machine learning algorithms?
* Data Warehousing
  + How would you improve ETL (Extract, Transform, Load) throughput?
* Big Data
  + Explain a simple Map/Reduce problem.
  + Explain how RDDs work with Scala in Spark
  + How would you use Map/Reduce to split a very large graph into smaller pieces and parallelize the computation of edges according to the fast/dynamic change of data?
  + What is Hadoop serialization?
* [Machine Learning](http://www.itshared.org/2015/10/data-science-interview-questions.html#machine-learning)
  + [General ML Questions](http://www.itshared.org/2015/10/data-science-interview-questions.html#general-ml-questions)
    - In your opinion, which is more important when designing a machine learning model, Model performance? or model accuracy?
    - Choose any machine learning algorithm and describe it.
    - Give examples of the algorithms using Gradient based methods of second order information.
    - Is more data always better?
    - Is there any negative impact of using too many or too few variables?
    - What are resampling methods. Why they are useful. What are their limitations?
    - What is Gradient Descent Method?
  + Model Validation
    - Do you know about Concordance or Lift?
    - What is a ROC curve? Write pseudo-code to generate the data for such a curve.
    - What is AU ROC (AUC)?
    - Which is better Too many false positives or too many false negatives?
    - How would you analyze the performance of the predictions generated by regression models versus classification models?
    - How would you assess logistic regression versus simple linear regression models?
    - How would you check if the regression model fits the data well?
    - How would you know if your model overfits?
    - I have two models of comparable accuracy and computational performance. Which one should I choose for production and why?
    - If you had a categorical dependent variable and a mixture of categorical and continuous independent variables, what algorithms, methods, or tools would you use for analysis?
    - Is it better to have too many false negatives or too many false positives?
    - What criteria would you use while selecting the best model from many different models?
    - What Is the 80/20 rule, and tell me about its importance in model validation.
    - What is the name of the matrix used to evaluate predictive models?
    - What precision and recall are?
    - Which evaluation metrics you know? Something apart from accuracy?
    - Can you explain the difference between a Test Set and a Validation Set?
    - What is 10-Fold CV?
    - What is Cross-Validation?
    - What is the difference between holding out a validation set and doing 10-Fold CV?
  + [Regression](http://www.itshared.org/2015/10/data-science-interview-questions.html#regression)
    - When to use k-Nearest Neighbors for regression?
  + [Classification](http://www.itshared.org/2015/10/data-science-interview-questions.html#classification)
    - How would you deal with unbalanced binary classification?
    - State some real life problems where classification algorithms can be used?
    - Tradeoffs between different types of classification models. How to choose the best one?
  + Decision Trees
    - Describe how Gradient Boosting works.
    - Describe some of the different splitting rules used by different decision tree algorithms.
    - How would you build a decision tree model?
    - How would you compare a decision tree to a logistic regression? Which is more suitable under different circumstances?
    - What are some business reasons you might want to use a decision tree model?
    - What impurity measures do you know?
    - What is pruning and why is it important?
    - What is Random Forest? Why would you prefer it to SVM?
    - Why do we combine multiple trees?
  + Linear Regression
    - Why is R2 horrible for determining the quality of a model and name at least two better metrics.
    - Why is linear regression called linear?
    - What are the assumptions that standard linear regression models with standard estimation techniques make? How can some of these assumptions be relaxed?
    - What Is the following parts of a linear regression to me p-value, coefficient, R-Squared value. What is the significance of each of these components and what assumptions do we hold when creating a linear regression?
    - Could you explain some of the extension of linear models like Splines or LOESS/LOWESS?
    - Do you consider the models Y~X1+X2+X1X2 and Y~X1+X2+X1X2 to be linear? Why?
    - In linear regression, under what condition R^2 always equals a perfect 1?
    - What are the basic assumptions to be made for linear regression?
    - What are the constraints you need to keep in mind when using a linear regression?
    - What is heteroskedasticity and how to solve it
    - What is the difference between logistic and linear regression? How do you avoid local minima?
  + Logistic Regression
    - I know that a linear regression model is generally evaluated using Adjusted R² or F value. How would you evaluate a logistic regression model?
    - How can you assess a good logistic model?
    - How would you train a logistic regression model?
    - What is the effect on the coefficients of logistic regression if two predictors are highly correlated? What are the confidence intervals of the coefficients?
    - What relationships exist between a logistic regression’s coefficient and the Odds Ratio?
  + [Regularization](http://www.itshared.org/2015/10/data-science-interview-questions.html#regularization)
    - How would you approach a categorical feature with high-cardinality?
    - How would you deal with sparsity?
    - What are the advantages and disadvantages of using regularization methods like Ridge Regression?
    - What are the problems of large feature space? How does it affect different models, e.g. OLS? What about computational complexity?
    - What is Lasso regression? How is it different from OLS and Ridge?
    - What is Regularization?
    - What is Ridge Regression? How is it different from OLS Regression? Why do we need it?
    - What is the difference between density-sparse data and dimensionally-sparse data?
    - What Is the difference between L1 and L2 regularization methods?
    - When might you want to use ridge regression instead of traditional linear regression?
    - Which problem does Regularization try to solve?
    - Why (geometrically) does LASSO produce solutions with zero-valued coefficients (as opposed to ridge)?
  + Bias – Variance Tradeoff
    - Bootstrapping - how and why it is used?
    - Define variance.
    - How does the variance of the error term change with the number of predictors, in OLS?
    - How would you control for biases?
    - What is overfitting a regression model? What are ways to avoid it?
  + [Dimensionality Reduction](http://www.itshared.org/2015/10/data-science-interview-questions.html#dimensionality-reduction)
    - Are dimensionality reduction techniques supervised or not? Are all of them are (un)supervised?
    - Do we need to normalize data for PCA? Why?
    - Why do we need to center data for PCA and what can happen if we don’t do it?
    - What dimensionality reductions can be used for preprocessing the data?
    - Suppose you have a very sparse matrix where rows are highly dimensional. You project these rows on a random vector of relatively small dimensionality. Is it a valid dimensionality reduction technique or not?
    - What is the advantage of performing dimensionality reduction before fitting an SVM?
    - What is the purpose of dimensionality reduction and why do we need it?
    - Is PCA a linear model or not? Why?
    - What are the differences between Factor Analysis and Principal Component Analysis?
    - What is Principal Component Analysis (PCA)? Under what conditions is PCA effective? How is it related to eigenvalue decomposition (EVD)?
    - What is the relationship between Principal Component Analysis (PCA) and Linear & Quadratic Discriminant Analysis (LDA & QDA)
  + [Cluster Analysis](http://www.itshared.org/2015/10/data-science-interview-questions.html#cluster-analysis)
    - Assuming a clustering model’s labels are known, how do you evaluate the performance of the model?
    - Differentiate between partitioning method and hierarchical methods of Cluster Analysis.
    - How would you assess the quality of clustering?
    - How would you select K for K-Means?
    - What Is K-Means and its objective?
    - What is the difference between Cluster and Systematic Sampling?
    - What is the difference between Gaussian Mixture Model and K-Means?
  + Feature Selection
    - Your model considers the feature X significant, and Z is not, but you expected the opposite result. How will you explain it?
    - You have a data set containing 100K rows, and 100 columns, with one of those columns being our dependent variable for a problem we'd like to solve. How can we quickly identify which columns will be helpful in predicting the dependent variable? Identify two techniques and explain them to me as though I were 5 years old.
    - Does the model affect the choice of feature selection method?
    - How Univariate feature selection works?
    - Is feature selection a dimensionality reduction technique?
    - Is there any thumb rule for the number of features that should be used? How do you select the best features?
    - What are some good ways for performing feature selection that do not involve exhaustive search?
    - What Is feature selection and its importance with examples.
    - What is the difference between feature selection and feature extraction?
    - What is variance threshold approach of feature selection?
    - What will be your approach to recursive feature elimination?
    - You fit a multiple regression to examine the effect of a particular feature. The feature comes back insignificant, but you believe it is significant. How will you explain it?
  + Hyperparameter Tuning
    - Explain grid search and how you would use it?
    - How would you use model tuning for arriving at the best parameters?
    - You have one model and want to find the best set of parameters for this model. How would you do that?
  + [Optimization](http://www.itshared.org/2015/10/data-science-interview-questions.html#optimization)
    - Describe a constrained optimization problem and how you would tackle it.
    - What are “slack variables”?
    - Do gradient descent methods always converge to same point?
    - Give examples of some convex and non-convex optimization algorithms.
    - Is it necessary that the Gradient Descent Method will always find the global minima?
    - What do you understand by statistical power of sensitivity (recall?) and how do you calculate it?
    - What is a local optimum is and why is it 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? Read possible answer
    - What is the difference between Batch Gradient Descent and Stochastic gradient descent.
  + [Recommendation](http://www.itshared.org/2015/10/data-science-interview-questions.html#recommendation)
    - How would you suggest followers on Twitter?
    - What is Collaborative Filtering?
  + SVM/Kernels
    - Name and describe three different kernel functions and in what situation you would use each.
    - What is a kernel? Explain the Kernel trick
    - Which kernels do you know? How to choose a kernel?
    - How would you train SVM? What about hard SVM and soft SVM?
    - How would you use SVD to perform PCA? When SVD is better than EVD for PCA?
    - Is it beneficial to perform dimensionality reduction before fitting an SVM? Why or why not?
    - What is the maximal margin classifier in an SVM? How this margin can be achieved and why is it beneficial?
    - Why does SVM need to maximize the margin between support vectors?
  + [Feature Engineering](http://www.itshared.org/2015/10/data-science-interview-questions.html#feature-engineering)
    - How would you derive new features from features that already exist?
    - What is Feature Engineering? Give an example where feature example can be very useful in predicting results from data and explain with reason why it is so effective in some cases?
  + [Natural Language Processing](http://www.itshared.org/2015/10/data-science-interview-questions.html#natural-language-processing)
    - What is the use of NLP in Machine Learning?
    - Split a large string into valid words and store them in a dictionary. If the string cannot be split, return false. What’s your solution’s complexity?
    - What Is the distances and similarity measures that can be used to compare documents?
    - How unstructured text data can be converted into structured data for the purpose of ML models?
    - How would you develop a model to identify plagiarism?
    - What is the computational complexity of finding a document’s most frequently used words?
    - Why and when stop words are removed? In which situation we do not remove them?
  + [Neural](http://www.itshared.org/2015/10/data-science-interview-questions.html#meta-learning) Networks
    - How does a neural network with three layers (one input layer, one inner layer and one output layer) compare to a logistic regression?
    - How would you train an ANN?
    - What is back propagation?
    - What is deep learning? What is CNN (Convolution Neural Network) or RNN (Recurrent Neural Network)?
* Data Visualization
  + Describe the differences between and use cases for box plots and histograms.
* [Hands-On](http://www.itshared.org/2015/10/data-science-interview-questions.html#hands-on)
  + [Problem to Solve](http://www.itshared.org/2015/10/data-science-interview-questions.html#problem-to-solve)
    - If you’re given 10 TBs of unstructured customer data, how would you go about finding extracting valuable information from it?
    - If you’re given a raw data table, how would perform ETL (Extract, Transform, Load) with SQL to obtain the data in a desired format?
    - Imagine a test with a true positive rate of 100% and false positive rate of 5%. Imagine a population with a 1/1000 rate of having the condition the test identifies. Given a positive test, what is the probability of having that condition?
    - Imagine there are three ants in each corner of an equilateral triangle, and each ant randomly picks a direction and starts traversing the edge of the triangle. What’s the probability that none of the ants collide? What about if there are N ants sitting in N corners of an equilateral polygon?
    - Imagine you are working with a hospital. Patients arrive at the hospital in a Poisson Distribution, and the doctors attend to the patients in a Uniform Distribution. Write a function or code block that outputs the patient’s average wait time and total number of patients that are attended to by doctors on a random day.
    - Imagine you’re climbing a staircase that contains n stairs, and you can take any number k steps. How many distinct ways can you reach the top of the staircase? (This is a modification of the original stair step problem)
    - There is a race track with five lanes. There are 25 horses of which you want to find out the three fastest horses. What is the minimal number of races needed to identify the 3 fastest horses of those 25?
    - How many times in a day does a clock’s hand overlap?
    - You have two beakers. The first beaker contains 4 litre of water and the second one contains litres of water.How can you our exactly 7 litres of water into a bucket?
    - A coin is flipped 1000 times and 560 times heads show up. Do you think the coin is biased?
    - In a city where residents prefer only boys, every family in the city continues to give birth to children until a boy is born. If a girl is born, they plan for another child. If a boy is born, they stop. Find out the proportion of boys to girls in the city.
    - There are two companies manufacturing electronic chip. Company A is manufactures defective chips with a probability of 20% and good quality chips with a probability of 80%. Company B manufactures defective chips with a probability of 80% and good chips with a probability of 20%. If you get just one electronic chip, what is the probability that it is a good chip?
    - Suppose that you now get a pack of 2 electronic chips coming from the same company either A or B. When you test the first electronic chip it appears to be good. What is the probability that the second electronic chip you received is also good?
    - A dating site allows users to select 6 out of 25 adjectives to describe their likes and preferences. A match is said to be found between two users on the website if the match on atleast 5 adjectives. If Steve and On a dating site, users can select 5 out of 24 adjectives to describe themselves. A match is declared between two users if they match on at least 4 adjectives. If Brad and Angelina randomly pick adjectives, what is the probability that they will form a match?
    - A coin is tossed 10 times and the results are 2 tails and 8 heads. How will you analyse whether the coin is fair or not? What is the p-value for the same?
    - Continuation to the above question, if each coin is tossed 10 times (100 tosses are made in total). Will you modify your approach to the test the fairness of the coin or continue with the same?
    - Each of your friends will tell you the truth ⅔ of the time and mess with you by lying ⅓ of the time. If all three friends answer “Yes, it’s raining,” what is the probability that is it actually raining in Seattle?
    - Estimate the number of 'Happy Birthday' posts that are logged on Facebook every day.
    - Given a die, would it be more likely to get a single 6 in six rolls, at least two 6s in twelve rolls, or at least one-hundred 6s in six-hundred rolls?
    - Given a list of followers in the format:123, 345234, 678345, 123…Where column one is the ID of the follower and column two is the ID of the followee. Find all mutual following pairs (the pair 123, 345 in the example above). How would you use Map/Reduce to solve the problem when the list does not fit in memory?
    - Given tweets and Facebook statuses surrounding a new movie that was recently released, how will you determine the public's reaction to the movie?
    - Given Twitter user data, how would you measure engagement?
    - Given two fair dices, what is the probability of getting scores that sum to 4? to 8?
    - If you have 70 red marbles, and the ratio of green to red marbles is 2 to 7, how many green marbles are there?
    - Pick any product or app that you really like and describe how you would improve it.
    - Suppose you have ten bags of marbles with ten marbles in each bag. If one bag weighs differently than the other bags, and you could only perform a single weighing, how would you figure out which one is different?
    - What would you do to summarize a Twitter feed?
    - You are about to hop on a plane to Seattle and want to know if you should carry an umbrella. You call three friends of yours that live in Seattle and ask each, independently, if it’s raining.
    - You have a marketing campaign and you want to send emails to users. You developed a model for predicting if a user will reply or not. How can you evaluate this model? Is there a chart you can use?
    - An ant is placed on an infinitely long twig. The ant can move one step backward or one step forward with same probability during discrete time steps. Find out the probability with which the ant will return to the starting point.
    - What are your 3 favorite data visualization techniques?
    - How would you build a model to predict credit card fraud?
    - Suppose you were given two years of transaction history. What features would you use to predict credit risk?
    - Suppose you’re given millions of users that each have hundreds of transactions and these millions of transactions are for tens of thousands of products. How would you group the users together in meaningful segments?
    - What data and models would you use to measure attrition/churn? How would you measure the performance of your models?
    - What features would you use to build a recommendation algorithm for users?
    - What is the probability you’ll draw two cards of the same suite from a single deck?
    - What metrics would you consider using to track if ’s paid advertising strategy to acquire new customers actually works? How would you then approach figuring out an ideal customer acquisition cost
    - What would the distribution of daily commutes in New York City look like?
    - What sort features could you give a driver to predict if they will accept a ride request or not? What supervised learning algorithm would you use to solve the problem and how would compare the results of the algorithm?
  + [Coding](http://www.itshared.org/2015/10/data-science-interview-questions.html#coding)
    - Create a function that checks if a word is a palindrome.
    - Can you explain what REST is?
    - Create a function that calculates matrix sums.
    - Define and explain the differences between clustered and non-clustered indexes.
    - Design an AI program for Tic-tac-toe
    - How can you iterate over a list and retrieve element indices at the same time?
    - How would you ‘disjoin’ two arrays (like JOIN for SQL, but the opposite)?
    - How would you find the second largest element in a Binary Search Tree?
    - How would you generate related searches for a search engine?
    - How would you use Python to read a very large tab-delimited file of numbers to count the frequency of each number?
    - How would you write a SQL query to compute a frequency table of a certain attribute involving two joins? What changes would you need to make if you want to ORDER BY or GROUP BY some attribute? What would you do to account for NULLS?
    - If you have an incoming stream of numbers, how would you find the most frequent numbers on-the-fly?
    - In which libraries for Data Science in Python and R, does your strength lie?
    - Suppose you’re given two binary strings, write a function adds them together without using any builtin string-to-int conversion or parsing tools. For example, if you give your function binary strings 100 and 111, it should return 1011. What’s the space and time complexity of your solution?
    - What are the different ways to return the rowcount of a table?
    - What do you understand by Fuzzy merging? Which language will you use to handle it?
    - What is Curse of Dimensionality? What is the difference between density-sparse data and dimensionally-sparse data?
    - What is the difference between an inner join, left join/right join, and union?
    - When does parallelism helps your algorithms run faster and when does it make them run slower?
    - Why is vectorization considered a powerful method for optimizing numerical code?
    - Create a function that does addition where the numbers are represented as two linked lists.
    - Code a function that calculates the square root (2-point precision) of a given number. Follow up: Avoid redundant calculations by now optimizing your function with a caching mechanism.
    - Write a function that accepts two already sorted lists and returns their union in a sorted list.
    - Write a function that raises one number to another number, i.e. the pow() function.
    - Write a function that takes a sentence and prints out the same sentence with each word backwards in O(n) time.
    - Write a function that takes an array, splits the array into every possible set of two arrays, and prints out the max differences between the two array’s minima in O(n) time.
    - Write a function that takes in two sorted lists and outputs a sorted list that is their union.
    - Write a function that takes two sorted vectors and returns a single sorted vector.
    - Write a Hive UDF that returns a sentiment score. For example, if good = 1, bad = -1, and average = 0, then a review of a restaurant states “Good food, bad service,” your score might be 1 – 1 = 0.
    - Write a program in Python which takes input as the diameter of a coin and weight of the coin and produces output as the money value of the coin.
    - Write a program that can determine the height of an arbitrary binary tree
    - Write a program that does merge sort.
    - Write an equation that would optimize the ad spend between Twitter and Facebook.
    - Write code to output FizzBuzz for integers between 1 and 100 (inclusive). (Yes, I've had candidates fail this...)