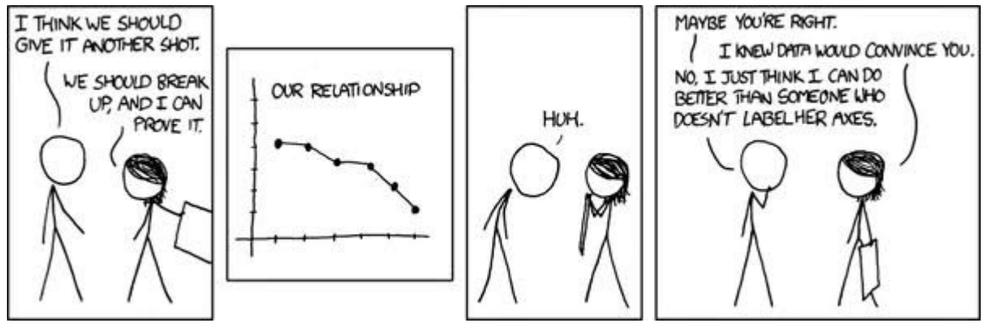
DATA SCIENCE ROADMAP 2020

Mohit... Follow
Feb 19 · 12 min...



Disclaimer — Everyone has different question paper in life. Many people fail because they try to copy others. This is true even if you want to become Data Professional.

Generally, if someone google's about '*How to become a data scientist*', There are lot of blogs which straight away go into skills and technologies relevant to Data Science world. This is what inspired me to write about the roadmap I follow while teaching Data Science.

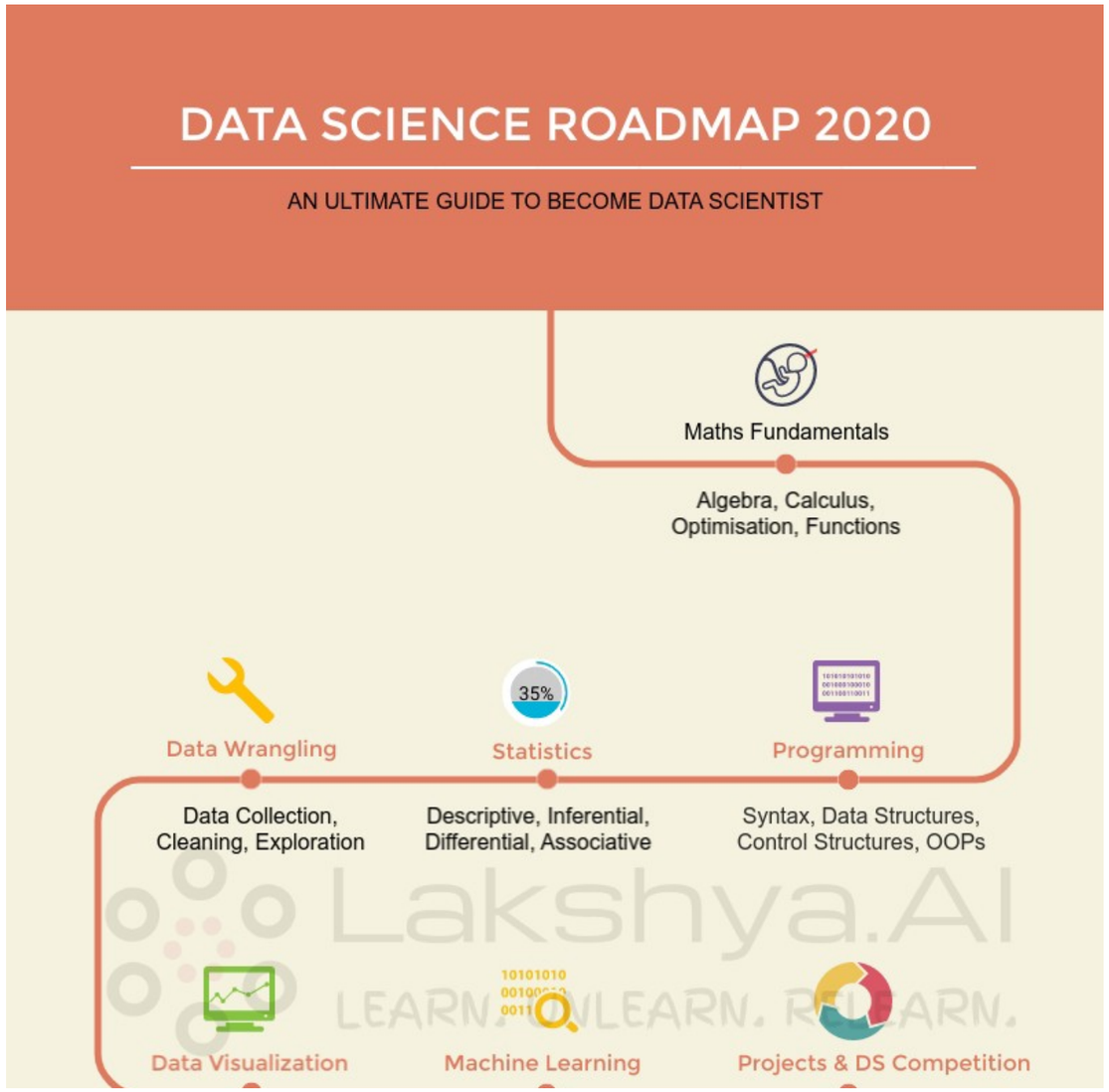
About me: Data Scientist with Master's degree in Business Analytics from IISC Bangalore, with big interest in how data shapes our lives. I have worked in multiple MNC's or startups as Data professional (Infosys, Siemens Healthineers, CitiBank, Unisys and Bounce).

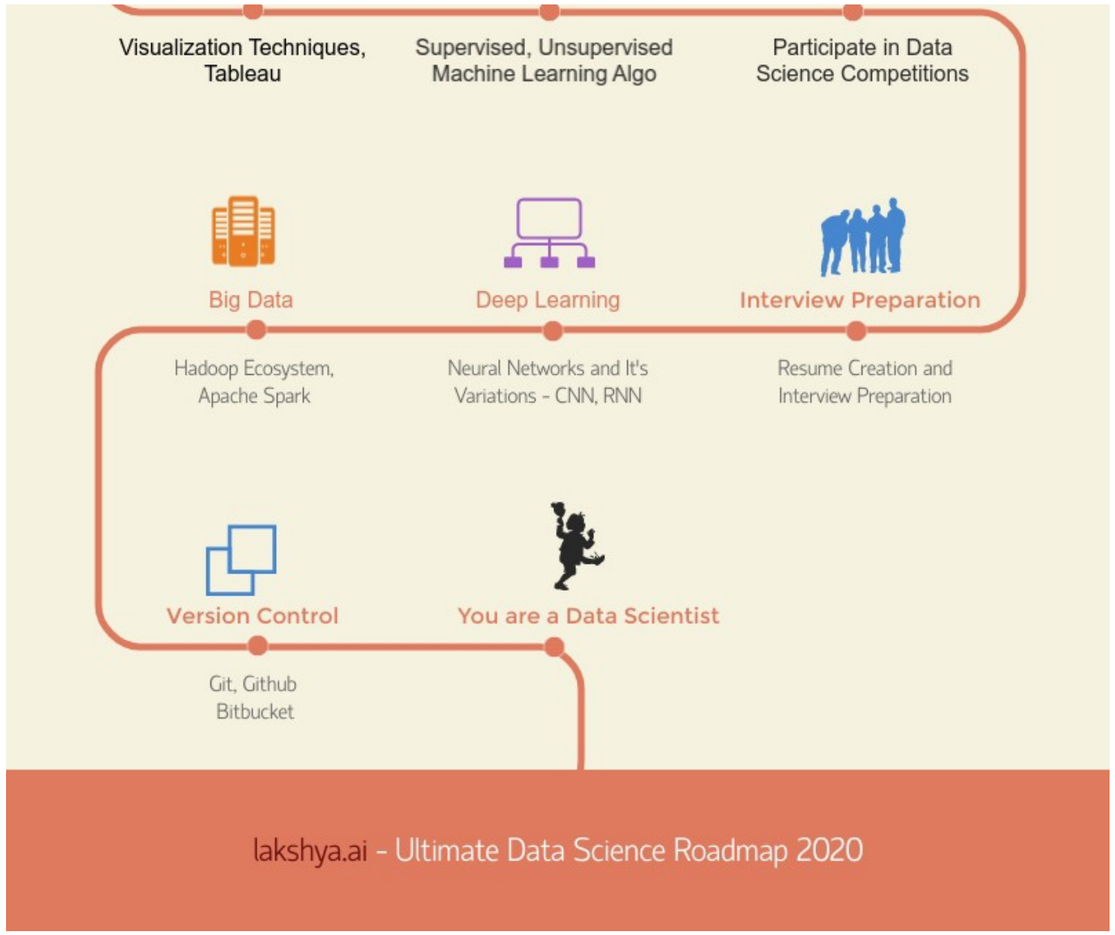
Now, Why I am blabbering about myself ? Sometime students ask me — Mohit Sir, Can I become a Data Scientist ? I always tell them about my background. I have done Btech in Mechanical field where I got no

exposure to programming. I was good with numbers and used to like maths a lot. That was the only reason I entered into Data Science Field.

So In this article, I am going to put my last 6 years of Data Science Experience across different companies and academics where I worked and try to provide an ultimate guide to all Data Science enthusiasts out there who are looking forward to become one.

Following picture dictates the roadmap one should follow to become Data Scientist.

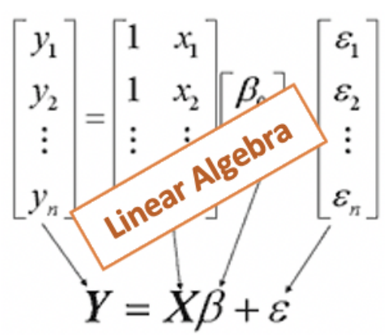


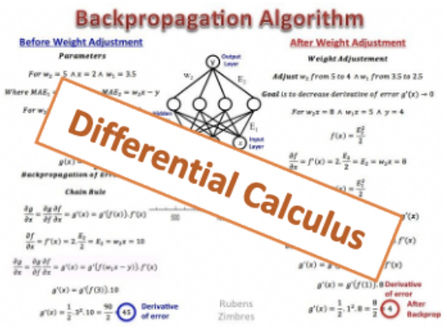


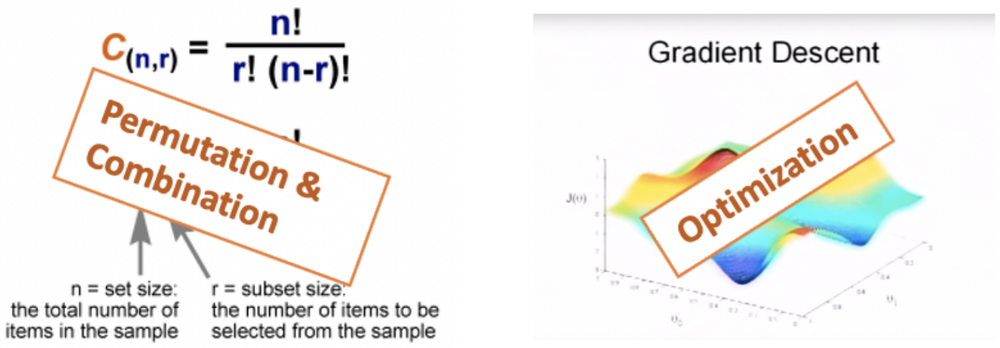
DATA SCIENCE ROADMAP 2020

Let’s deep dive into each step and see what all topics one needs to learn.

Mathematics Fundamentals:







Mathematics topics for Data Science

Mathematics is the backbone of Data Science. You would have heard people saying that ML/DL models are like black box. But they are not. It is just that we haven't put enough efforts to understand the maths under the hood. One doesn't need to become master of maths in order to start their career in Data Science, But if you are good with math then you will be playboy of Data Science world. Spend 2–3 days, just to get comfortable with the topics mentioned below.

Few Use-cases to motivate:

- Straight Line Geometry and Matrix Operations used in Linear Regression.
- Sigmoid Function is the backbone of Logistic regression.
- Differential calculus is the backbone of Backpropagation which is the backbone of any Deep Learning Algortihms.
- Eigen Values/Vectors is must to understand Principal Component Analysis which is a very popular Dimensionality Reduction Technique.
- Gradient Descent utilises differential calculus as well as Optimisation of cost function.

- Permutation and combination is used to get understanding of probability which is must for Bayes Theorem and Naive Bayes Model.

Topics to be covered:

- Linear Algebra — Vector, Matrix Operations, Matrix Types, Eigen Values and Eigen Vectors, Set Theory, Functions, Logarithmic, Exponential Functions.
- Differential Calculus
- Permutation and Combination
- Optimisation Technique: Linear Programming, Maxima/Minima

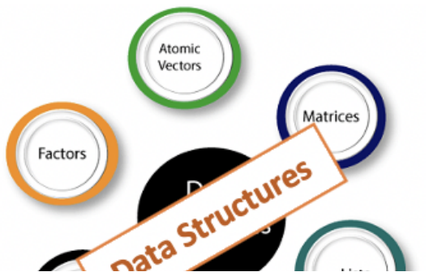
Resources:

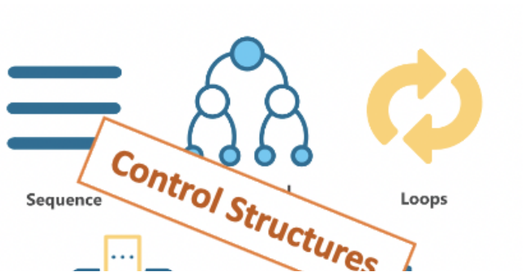
Mathematics for Machine Learning specialization — coursera

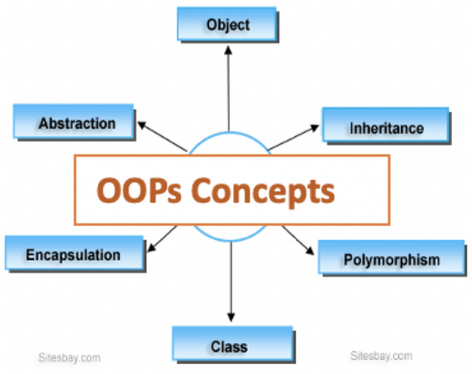
Khan Academy's Linear Algebra, Probability & Statistics, Multivariable Calculus and Optimization.

Disclaimer: Don't start reading maths book until and unless you are not in rush to start career in Data Science. (One of my friend did this mistake a year ago, And now he became father of a son recently)

2. Programming Language:

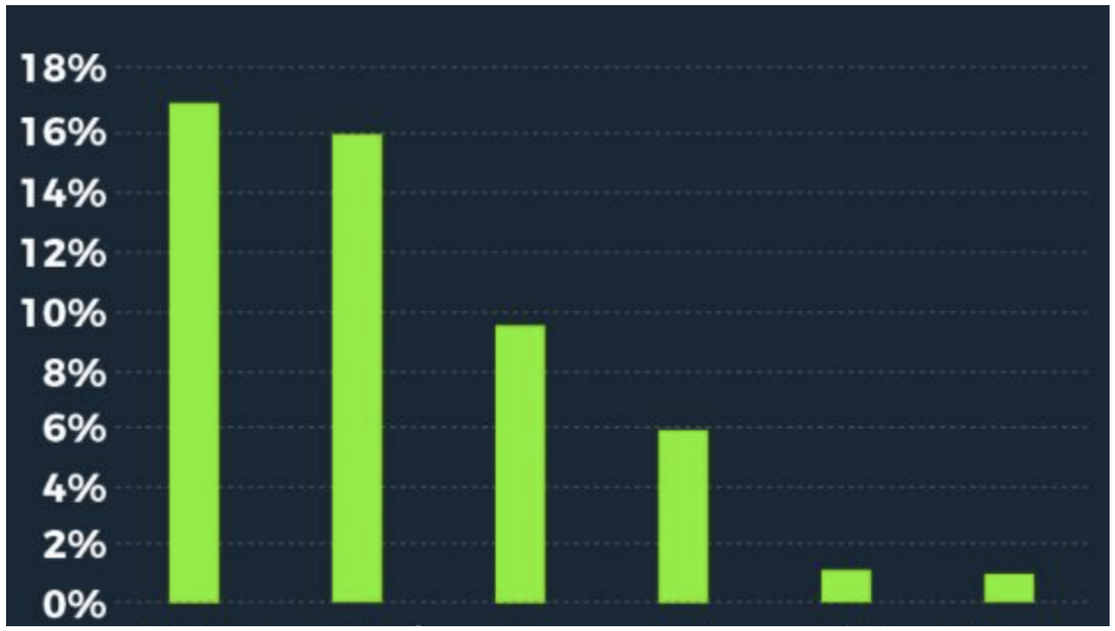






Programming fundamentals

More than 250+ langauges are out there in the market. Choosing which language to learn is quite difficult. Generally, Data enthusiasts gets confused between R and Python. In order to decide on the language, Let's look at the survey done by Analytics India Magazine in 2019. It tells the preferred programming language by recruiters.



Recruiters preferred Programming language

As it clearly says, the demand for Python professionals is the highest among all analytics recruiters. Almost 17% of all advertised analytics jobs in India demand for Python as a core skill whereas 16% demand Java. 8 out of 10 data scientists would recommend Python as the prime language. Python is easy to learn and widely-accepted programming language.

So, I hope now you would have made your mind to start with Python.

Not yet ? Ok, Let me try with my experience. Personally, I have 3+ years of experience in R and Python. Visual Studio with python plugin is one of the application always stay open in my laptop. Both R and Python have their own strength and weaknesses. It will depend on the use-case for which you are using them. Here is the difference R and Python which can help you in making decision.

Criteria	R	Python
Customer Support	✓	✓
Structured Data Analysis	✓	
Text Mining		✓
Speed		✓
Data Visualization	✓	
Deployment		✓
Software engineering / Web Applications		✓

R Vs Python

For Being a Data Scientist, One has to deliver projects end-to-end,starting from identifying the problem, collecting data related to the problem, performing data cleaning and exploratory data analysis, then building models and finally deplyment. I would recommend Python.

If still struggling, buddy you know how to google. Please ping me on my Linkedin, personally I would hear your hesitation.

Now once you have decided on the language you want to learn, Next step would be to learn programming fundamentals.

Topics you have to cover:

- Language fundamental syntax.
- Data Types and Data Structured
- Control Structure — if-else statement, Loops, User defined functions
- OOPs Concepts
- Module Creation, Exception Handling

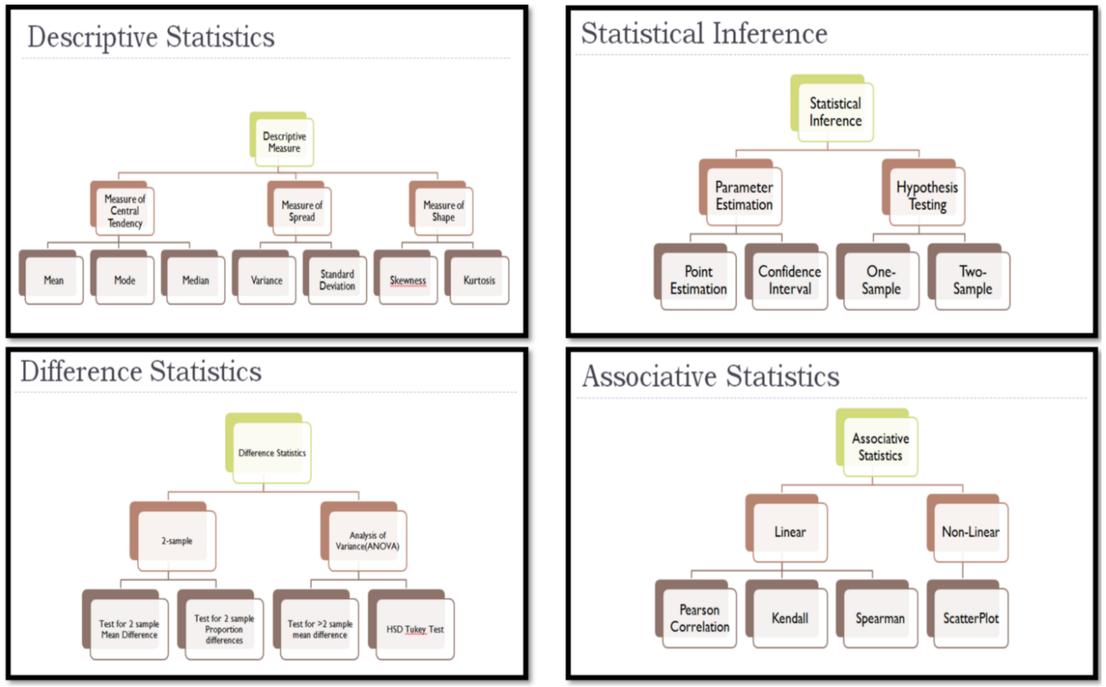
Resources:

For Python fundamentals:

- Python for Everybody Specialization — Course 1 and 2
- Python for everybody book — Download here

For OOps concept and Module creation — Blog

Probability and Statistics:



Types of Statistical Analysis

One of my student asked — Mohit Sir, Why do we need statistics in Data Science? I simply answered — To Survive. One may get the job without knowing about statistics, but then you will get raped for your entire data science journey. (Resemble to a dialogue in 3 idiots).

There are 5 types of statistical analysis. Four of them with topics are highlighted in the above image. The image above serves as a guide for my students in learning Statistics. Fifth type of Statistical Analysis which is not in the image — Predictive Statistical Analysis which covers predictive modelling techniques like Linear Regression, Logistic Regression, Generalised Linear models.

Topics to be covered:

Descriptive Statistics — Data Summarisation

Measure of central tendency — Mean, Median, Mode

Measure of spread — Range, Standard deviation, variable, Inter-quartile range

Measure of shape — Skewness and Kurtosis.

Statistical Inference — drawing inference about population from sample

Parameter Estimation — Point Estimation and Confidence Interval

Hypothesis Testing — One sample Hypothesis testing (z-test, t-test, chi-square test and f-test)

Differential Statistics — 2 sample Hypothesis testing, ANOVA, MANOVA, ANCOVA and MANCOVA.

Associative Statistics — Finding relationships between 2 variables. Correlation — Pearson, Spearman and Kendall.

Resources:

Personal favourite — Applied Statistics and Probability for Engineers

Statistics: Methods and Applications (StatSoft) — Electronic Textbook

Elementary Statistics(San Jose State University) — University-level Online Course

Relevant Libraries:

Python — *scipy.stats, statsmodels*

R — *stats, corrplot*

Data Collection and Wrangling:

Data Exploration

Data wrangling is the process of transforming and mapping data from one raw data form into another format with the intent of making it more appropriate and valuable for a variety of downstream purposes such as analytics. Here I have combined multiple Data Wrangling steps intop groups on which success of any Data Analytics / Science project depends.

Topics to be covered:

Data Cleaning: Missing Value Treatment, Outlier Treatment, Data Validation

Data Manipulation: Subsetting, Indexing, Groupby, Aggregation, Pivot tables, Data Merge, Reshaping, Creating new variables, Sorting.

Resource:

Personal Favourite Book (Python) — Python for Data Analysis

Data wrangling in R — Video Link

Interesting read on data exploration strategy — here

Relevant Libraries:

Python — Pandas, Numpy, Scipy, Matplotlib, Seaborn, folium, bukeh

R — dplyr, sqldf, data.table, stringr, tm, ggplot2, ggviz, RWorldMap

> *Now time to practice whatever learnt on real dataset — here*

Still needs more practice or real problems: message me

Data Visualization

Data Visualization Techniques

As rightly said — **“A picture is worth a thousand words**”. In Data Science, this is worth billion dollar. No one can call oneself as Data Scientist until he/she is good with visualization. Data visualization and dashboard design are both art and science and not as easy to create as they may first appear. One should know about visualisation techniques and more importantly where to use which one. Visualization helps in making story out of data. A lot of organisation (can’t tell the name) has made billions by impressing clients just by Data Visualization.

Topics to be covered:

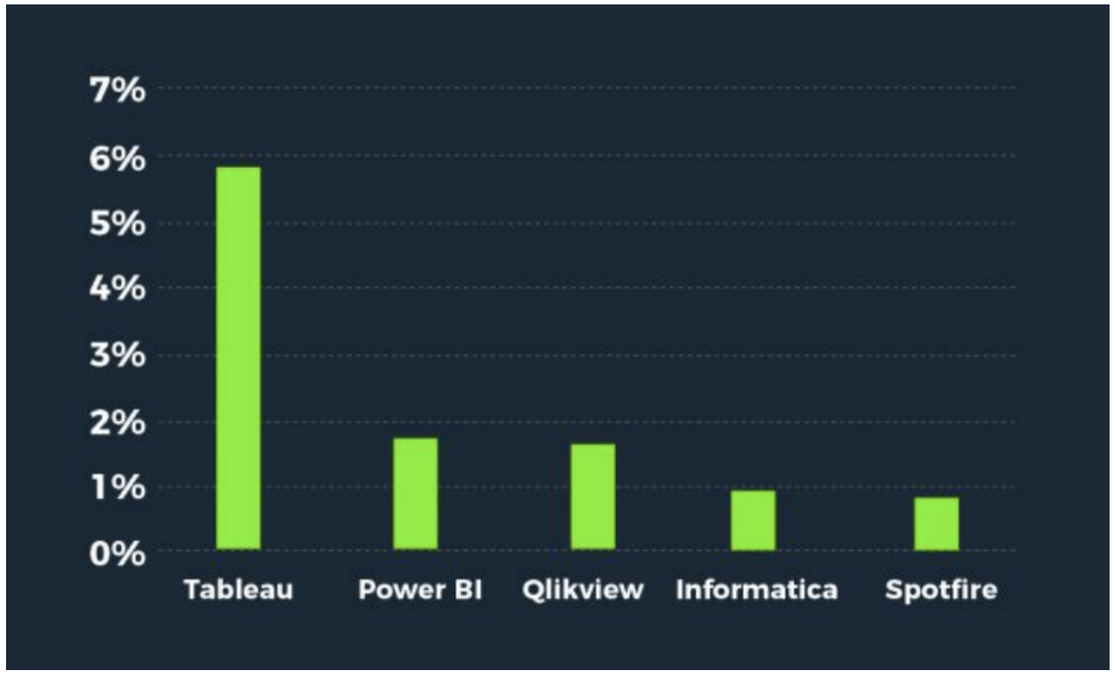
- Data Visualization Techniques and their Use
- Line chart, Boxplot, Histogram, Scatter plot
- Bubble chart, bar chart, Heatmap, world map

Resources:

Personal Favourite Book (Python) — Python for Data Analysis

Elegant graphics for data analysis (R) — link

Above Topics will help one in performing Data Exploratory Analysis. However to build a story from the data, companies use different type of tools like Tableau, PowerBI, Qlikview which helps in creating interactive dashboards and stories. Dashboards make it easy for a company to visualize their data by displaying metrics, graphs, gauges, maps, percentages and comparisons of all the information that is streaming in and out of the company. Let’s look at the dashboarding tool preferred by recruiters, published by Analytics India Magazine:

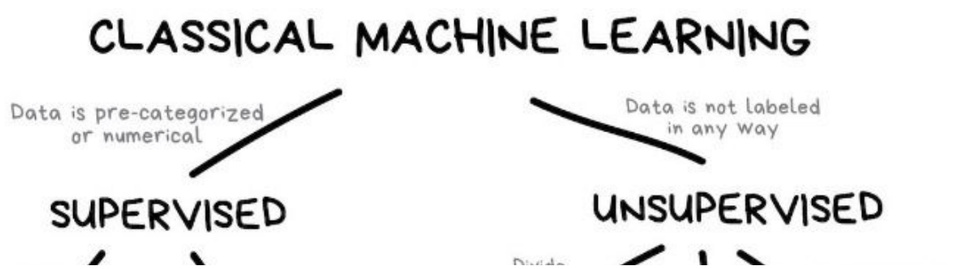


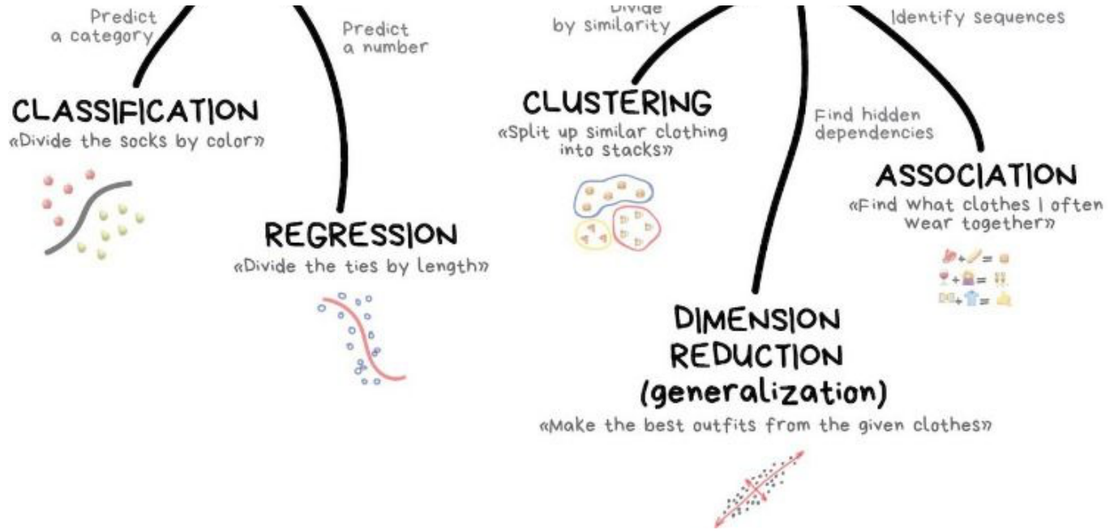
Dashboard Tools

From above, It is clear that among BI tools, Tableau skills continue to be most in demand, followed by Microsoft Power BI & Qlikview. Annual average salary of tableau developer in india is around ₹510,086 as per PayScale. I have friends who are earning pretty handsome amount by creating beautiful dashboards with rainbow colors.

So, I would personally recommend to learn tableau along with Data Visualization.

Machine Learning





Machine Learning — Supervised and Unsupervised

Machine learning (ML) is a subset of artificial intelligence (AI) that allows software applications to perform a task without being explicitly programmed to do so. Machine Learning is the brain of Robots like Sophia. When it comes to machine learning, People jump directly to algorithms and implement them without knowing backend of the algorithms. Implementing ML model is not a big task, people have already written codes for you and within 5 lines of code you will be able to implement and evaluate any ml algorithm. But that's not the goal. Here goal is to implement them right. For example: I have seen Data experts interpreting results of Linear regression without performing residual analysis and goodness of fit test.

In order to start Machine Learning, first understand the terminologies around machine learning and its types. Then look at the algorithms for every types as given in the above image. I haven't mentioned **Semi-Supervised algorithms**, but if interesting. Explore here

Topics to be covered:

- Machine Learning Types — Supervised, Semi-supervised and Unsupervised
- Classification and Regression Problems
- Bias-Variance Tradeoff
- Underfitting and Overfitting Problems
- Imbalanced Dataset and how to deal with it.
- Model Evaluation Techniques for CLassification and Regression.

Supervised Algorithms:

Classification

- Naive Bayes
- Logistic Regression
- Decision Trees
- K-Nearest Neighbors
- Support Vector Machines
- Bagging Trees — Random Forests
- Boosted Trees — Adaboost, GBM, XGBoost and Light GBM

Regression

- Linear Regression, Lasso Regression and Ridge Regression
- Decision Trees
- K-Nearest Neighbors
- Ensemble Techniques

Unsupervised Techniques:

Clustering — Partition based technique (K-means) and Hierarchical Clustering (Agglomerative and Divisive Clustering)

Dimensionality Reduction Techniques — Principal Component Analysis, Factor Analysis and Singular Vector Decomposition

Market Basket Analysis — Apriori Algorithm, FP Growth Technique.

Time Series Analysis :

Stationarity condition, Auto-Regressive model, Moving Average Model, ARIMA, ARIMAX, SARIMA.

Recommendation System:

Collaborative and Content Based filtering recommendation system

Text Mining:

Text Data Preparation techniques, Text Classification, Sentimental Analysis, Topic Modelling and Name entity recognition.

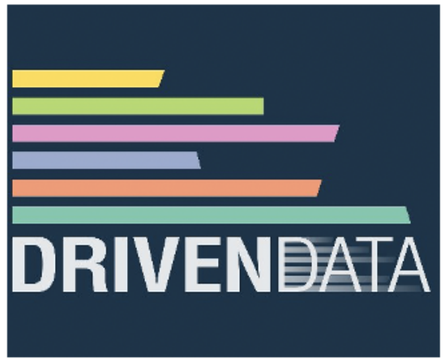
Resources:

Personal Favourite — Machine Learning by Andrew Ng

Machine Learning in Python by Michael Bowles — link

Machine Learning with R — link

Data Science Competition Participation



Data Science Competition

Data Science is more art than science. Best way to learn Data Science is by doing it. There are lots of platform which host data science competitions to build a better world, bringing cutting-edge predictive models to organizations tackling the world's toughest problems. Lot of companies internally conduct such competitions to bring competitive and learning culture. I literally ask my students to participate in the competitions in the class and submit the solution to give them motivation. It also helps them develop business acumen, and improve their technical skills at the same time.

Here I have listed few of them:

- Kaggle
- Analytics Vidhya
- Data Driven
- Crowd AI
- Coda Lab

Resume Creation and Interview Preparation

At this point, if one has completed above topics religiously, He/She is literally ready to become Data Analyst (not Data Scientist :-)). One can start to create resume and prepare for interviews. There are lot of blogs which can help you create a good resume, Here I am listing few important points while creating resume which I teach to my students:

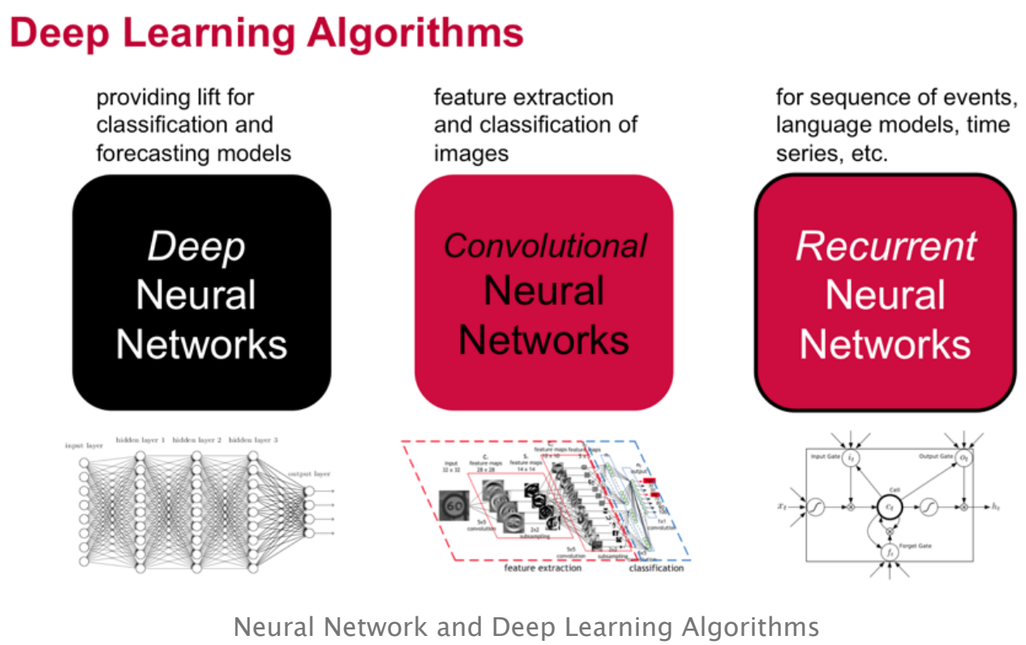
- Before start creating resume, put yourself in the employer's shoes and think what kind of guy they are looking. For example: Are they looking for more technical guy or consulting guy or managerial guy ? This will help you to know what kind of words one has to use in the resume.

Note: Generally candidates don't follow above practice (even my students) and they come to me by saying I am not getting calls for interviews. I know it's difficult to change resume for every interview, but i would recommend to follow above practice at least for good profiles/companies, as resume is your first impression to the interviewer.

- While writing about projects undertaken, mention business objective, solution approach, technologies and skills used and very importantly, impact on the business. Use numbers as much as you can.

- Write explicitly all the skills and technologies you are honestly comfortable with.

Neural Network and Deep Learning



Neural Network and Deep Learning Algorithms

Neural networks are one of the most popular machine learning algorithm today, achieving impressive performance on a large variety of tasks where traditional ml models are not good.

Topics to be covered:

Foundation

- Neural Network analogy to Human Brain
- NN Representation, Activation functions, Backpropagation,
- Training (weight optimization) using backpropagation

- Gradient descent, Stochastic gradient descent
- Hyperparameter tuning
- Deep neural networks: preventing overfitting

Convolutional neural networks

- convolutional neural networks Architecture and Layers
- Image Classification
- Object detection
- One stage methods: YOLO and SSD
- Two stage methods: Faster R-CNN
- Facial recognition

Recurrent neural networks

- RNN Model
- Vanishing gradient problem
- Gated recurrent units: Introducing intentional memory
- Long short term memory networks: Learning what to remember and what to forget

Transfer learning

- Image recognition
- Natural language processing

Resources:

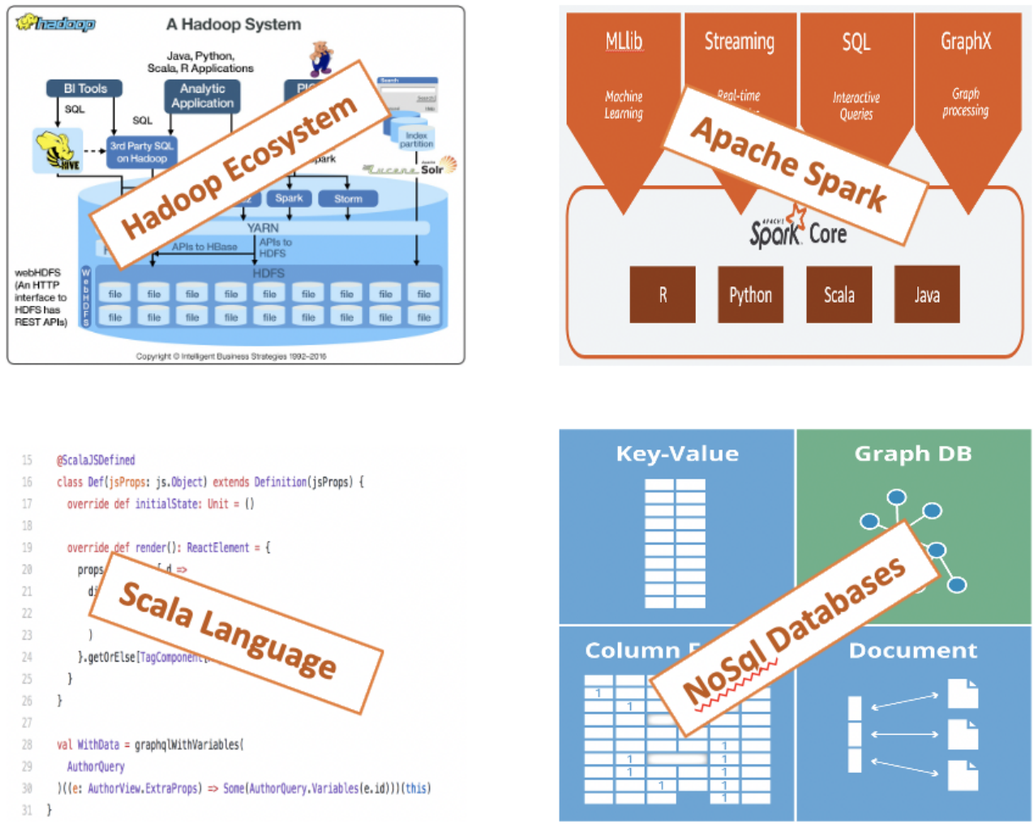
Personal Favourite — Deep Learning Specialization Andrew Ng

Relevant Libraries:

Python — *Keras*, *Tensorflow*, caffe, mxnet, theano, deeplearning4j

R — h2o, *neuralnet*, *nnet*, *tensorflow*, rcppDL, MXNetR

Big Data



Big Data Tools

Topics to be covered:

- Big Data Fundamentals

- Hadoop Ecosystem — Cluster, HDFS, MapReduce, Hive, Pig
- Apache Spark — Streaming, SQL, Machine Learning (ML), and Graph API’s
- Pyspark/SparkR/Scala Language/ Java Language

Resources:

Big Data Specialization Coursera — link

Getting started with Apache Spark — link

Version Control System

Version Control Tools

Seriously, Do I need to learn Version control system tools as well to become Data Scientist ? I would say — “without fail”.
Motivational analogy — Without a version control system in place, one is probably working together in a shared folder on the same set of files. Shouting through the office that you are currently working on file “xyz” and that, meanwhile, your teammates should keep their fingers off is not

an acceptable workflow. It’s extremely error-prone as you’re essentially doing open-heart surgery all the time: sooner or later, someone will overwrite someone else’s changes. Right version control software helps product development team work simultaneously, automate tasks, track changes, and ensure high availability/disaster recovery.

Topics to be covered:

- Creating Repositories and clone it.
- Basic Git commands like init, add, commit.
- Creating branches, push, pull to the repository.

Resources:

Personal favourite — git documentation

And That’s it. You are now a qualified Data Scientist and ready to ace any Data Science interview. **Best of Luck !!!**

If you found this interesting or helpful Please help others find it by sharing and clapping.

On Linkedin? So am I. Feel free to keep in touch — Mohit🙌.

A special thanks to Gunjan Thareja for her significant contributions and feedback.

Interview Preparation Data Science Data Science Topics 2020 Machine Learning Statistics