Roadmap to Data Scientist

1. Fundamentals

- Matrices & Linear Algebra Fundamentals
- Hash Functions, Binary Tree, O(n)
- Relational Algebra, DB Basics
- Inner, Outer, Cross, Theta Join
- CAP Theorem
- Tabular Data
- Data Frames & Series
- Sharding
- OLAP
- Multidimensional Data Model
- ETL
- Reporting Vs BI Vs Analytics
- JSON & XML
- NoSQL
- Regex
- Vendor Landscape
- Env Setup
- Entropy

2. Statistics

- Pick a Dataset (UCI Repo)
- Descriptive Statistics (mean, median, range, SD, Var)
- Exploratory Data Analysis
- Histograms
- Percentiles & Outliers
- Probability Theory
- Bayes Theorem
- Random Variables
- Cumul Dist Fn (CDF)
- Continuos Distributions (Normal, Poisson, Gaussian)
- Skewness
- ANOVA
- Prob Den Fn (PDF)
- Central Limit Theorem
- Monte Carlo Method
- Hypothesis Testing
- p-Value
- Chi² Test
- Estimation
- Confid Int (CI)
- MLE
- Kernel Density Estimate
- Regression
- Covariance
- Correlation
- Pearson Coeff
- Causation
- Least² Fit
- Euclidean Distance

Install Pkgs — Factor Analysis

3. Programming

- Python Basics
- Working in Excel
- R Setup, R Studio
- R Basics
- Expressions
- Variables
- Vectors
- Matrices
- Subsetting Data
- Manipulate Data Frames
- Functions
- Factors
- Lists
- Arrays
- Reading Raw Data
- Reading CSV Data
- Data Frames
- Rapid Miner
- IBM SPSS

4. Machine Learning

- What is ML ?
- Numerical Var
- Categorical Var
- Supervised Learning
- Unsupervised Learning
- Concepts, Inputs & Attributes
- Training & Test Data
- Classifier
- Prediction
- Lift
- Overfitting
- Bias & Variance
- Trees & Classification
- Classification Rate
- Decision Trees
- Boosting
- Naive Bayes Classifiers
- K-Nearest Neighbor
- Logistic Regression
- Ranking
- Linear Regression
- Perceptron

Regression · **Classification** · **Clustering**
- Hierarchical Clustering
- K-means Clustering
- Neural Networks
- Sentiment Analysis
- Collaborative Filtering
- Tagging

5. Text Mining / NLP

- Corpus
- Named Entity Recognition
- Text Analysis
- UIMA
- Term Document Matrix
- Term Frequency & Weight
- Support Vector Machines
- Association Rules
- Market Based Analysis
- Feature Extraction
- Using Mahout
- Using Weka
- Using NLTK
- Classify Text
- Vocabulary Mapping

6. Visualization

- Data Exploration in R (Hist, Boxplot etc)
- Uni, Bi & Multivariate Viz
- ggplot2
- Histogram & Pie (Uni)
- Tree & Tree Map
- Scatter Plot (Bi)
- Line Charts (Bi)
- Spatial Charts
- Survey Plot
- Timeline
- Decision Tree
- D3.js
- InfoVis
- IBM ManyEyes
- Tableau

7. Big Data

- Map Reduce Fundamentals
- Hadoop Components
- HDFS
- Data Replication Principles
- Setup Hadoop (IBM / Cloudera / HortonWorks)
- Name & Data Nodes
- Job & Task Tracker
- M/R Programming
- Sqoop: Loading Data in HDFS
- Flume, Scribe: For Unstruct Data
- SQL with Pig
- DWH with Hive
- Scribe, Chukwa For Weblog
- Using Mahout
- Zookeeper Avro
- Storm: Hadoop Realtime
- Rhadoop, RHIPE
- rmr
- Cassandra
- MongoDB, Neo4j

8. Data Ingestion

- Summary of Data Formats
- Data Discovery
- Data Sources & Acquisition
- Data Integration
- Data Fusion
- Transformation & Enrichment
- Data Survey
- Google OpenRefine
- How much Data?
- Using ETL

Sampling · Stratified Sampling · Principal Component Analysis

9. Data Munging

- Dimensionality & Numerosity Reduction
- Normalization
- Data Scrubbing
- Handling Missing Values
- Unbiased Estimators
- Binning Sparse Values
- Feature Extraction
- Denoising

10. Toolbox

- MS Excel w/ Analysis ToolPak
- Java, Python
- R, R-Studio, Rattle
- Weka, Knime, RapidMiner
- Hadoop Dist of Choice
- Spark, Storm
- Flume, Scibe, Chukwa
- Nutch, Talend, Scraperwiki
- Webscraper, Flume, Sqoop
- tm, RWeka, NLTK
- RHIPE
- D3.js, ggplot2, Shiny
- IBM Languageware
- Cassandra, MongoDB

5% · 15% · 30% · 40% · 50% · 60% · 80% · 100%