**Data Science Techniques:**

1. Machine Learning (ML)
2. Deep Learning (DL)
3. Natural Language Processing (NLP)
4. Reinforcement Learning
5. AI based Learning

**Data Science Workflow, what data scientists do?**

1. Define the question
2. Define the ideal set of data
3. Determine what data you can access
4. Obtain the data
5. Clean the data
6. Explore data (Exploratory analytics)
7. Statistical prediction/modelling
8. Interpret result
9. Challenge result
10. Synthesize or write up result
11. Create reproducible code
12. Distribute result to other people.

**Machine Learning work flow**

**Question-> Input Data-> Features-> Algorithm-> Parameters-> Evaluation**

**Raw Data -> Processing Script -> Tidy Data -> Data Analysis -> Data Communication**

**Qualitative** **vs** **Quantitative** variables: Qualitative variables are Country, Sex, Treatment etc. While quantitative variables are Height, Weight, BP etc.

**Raw Data:** Strange binary file, Unformatted Excel file with 10 worksheets, JSON, Hand entered numbers etc.

**Steps of Cleaning Data:**

1. Use Sub-setting and Sorting to analyze data.
2. Use Filter with and/or to get required data.
3. Add human readable variable names at top.
4. View data summary () and structure str()
5. Check quantiles of quantitative variables.
6. Check size of data.
7. Check and deal with missing values.
8. Make Tables, Crosstab and Flat table
9. Check row & column sum (zero => no data)
10. Check for values with specific characteristics.
11. Create new variables as seems suitable.
12. Create Categorical and Factor Variable.

Finally, it should follow all properties of Tidy Data

**Clean Data Properties:**

1. Each variable in one column.
2. Each observation in one row.
3. Each File should contain one type of data.
4. There should be link between different files.
5. Variable names should be human readable

**The Code Book and Instruction List:** Should Contain information about variables in word or other formatted file. Each step should be mentioned with version information and parameter values.

**Type of Data Science questions:**

1. **Descriptive**, describe set of data.
2. **Exploratory**, find relationship, not known.
3. **Inferential**, from small data to bigger set.
4. **Predictive**, predict another object from data
5. **Causal**, find result of varying a variable.
6. **Mechanistic**, exact change in a var for a result.

**Exploratory Analytics:**

**Exploratory Graphs:** Create Scatter plot, Histograms, Box Plot and Density Plot and Heat maps for showing correlation. Create plotting based on different variables to get variety of insights.

Group, analyze and compare data based on above exploratory graphs. Check for Missing Data. Perform exploratory analytics e.g. clustering.

**Five Number Summary: -** Use Describe method to get stats (Min, Max, Average, Median, Mode, IQR etc.) of different variables.

**Principals of Exploratory Graphs**

1. Show Comparison
2. Show Causality, Mechanism, Explanation
3. Show Multivariate Data
4. Integrate Multiple modes of evidence
5. Describe and Document the evidence
6. Content is King: Focus on above all.

**Reproducibility and its Drivers:**

1. Don’t do things manually or by point and click. Always use script/program.
2. Always use proper version control.
3. Keep track of software Environment.
4. Don’t Save Output, It should always be reproducible from input and steps.
5. Set seed for reproducible results.
6. Think about entire Pipeline.
7. Keep record of time when you get data.

**Some Places to get data and Get Dirty.**

1. <http://data.gov>
2. <http://data.gov.in>
3. <http://www.kaggle.com>
4. <http://data.un.org>
5. <http://www.asdfree.com>
6. Other websites through web scraping.
7. UCI Machine learning data sets.
8. Public data sets on AWS.
9. Gene expression omnibus.
10. Twitter, Facebook and Google Map API’s
11. Stanford Large network data.
12. Data Market place like infochimp.

**Data Science Libraries: -**

1. **sk-learn:** Machine Learning library **python**
2. **Tensorflow:** ML & DL, python & java.
3. **Pytorch:** ML & DL python
4. **Keras**: ML & DL python
5. **ml-lib**: ML library over spark, Scala.
6. **NLTK:** NLP library for python.
7. **OpenNLP :** NLP, Java.
8. **Pandas**: Data processing, python
9. **Numpy**: data processing python
10. **OpenCV**: image/video python, java.

**Fancy Keywords:**

1. **Data Dredging:** Data dredging (also data fishing, data snooping, data butchery, and p-hacking) is the misuse of data analysis to find patterns in data that can be presented as statistically significant when in fact there is no real underlying effect.
2. **Inference Vs Prediction:** ‘prediction’ is used for the future event, and ‘inference’ for the conclusion based on observable facts.
3. **Confounding:** In statistics, a confounder is a variable that influences both the dependent variable and independent variable causing a spurious association.
4. **Ensemble Learning:** Using multiple learning algorithms for same task.We use multiple algorithm and then take vote to get final decision. Random Forrest is ensemble of decision tree. Less overfitting, better accuracy.
5. **Bagging (Bootstrap aggregating):** It takes M subsamples (with replacement) from the initial dataset and trains the predictive model on those subsamples. The final model is obtained by averaging the "bootstrapped" models and usually yields better results.
6. **Boosting:** 'Boosting' refers to a family of algorithms which converts weak learner to strong learners.

Applying Bagging or Boosting to a base Algorithm makes it ensemble algorithm.

1. **Data Munging:** Standardizing, Cleansing, Joining/Matching, Filtering and Consolidating. Munging is the standard definition for irrevocably changing or damaging data beyond its original state. The term is thought to have originated as a backronym for “Mash Until No Good”.
2. **Feature Scaling & Standardization:** A Min-Max scaling is typically done via the following equation: Xnorm = (X−Xmin)/(Xmax−Xmin)

The Z score normalization or standardization is represented by following formula: z=x−μ/σ

**Data Science Terminologies.**

|  |  |
| --- | --- |
| Machine Learning | Artificial NN |
| Classification | Biological Neuron |
| Regression | Perceptron |
| Clustering | Multi-Layer Perceptron |
| Association | Feed forward NN |
| Decision Tree | Recurrent NN |
| SVM | Activation Function |
| Neural Networks | Back Propagation |
| Deep Learning | Cost Function |
| Reinforcement  Learning | Graph Based  Techniques |
| Cross Validation | Gradient Descent |
| Bayesian/Bayes Theorem | Vanishing Gradient |
| Feature Selection | Convolutional NN |
| Expectation Maximization | Streaming Scenario |
| K-Means | LSTM Network |
| K-Medians | Bagging, Boosting |
| Density Based Method | Bias Variance Trade off |
| Grid Based Method | Bootstrapping |
| Matrix Factorization | Decision Boundary |
| Spectral Method | PCA & SVD |
| Dimensionality Reduction | Hyper parameter |
| Imputation | Over/under-fitting |
| ROC | AUC |