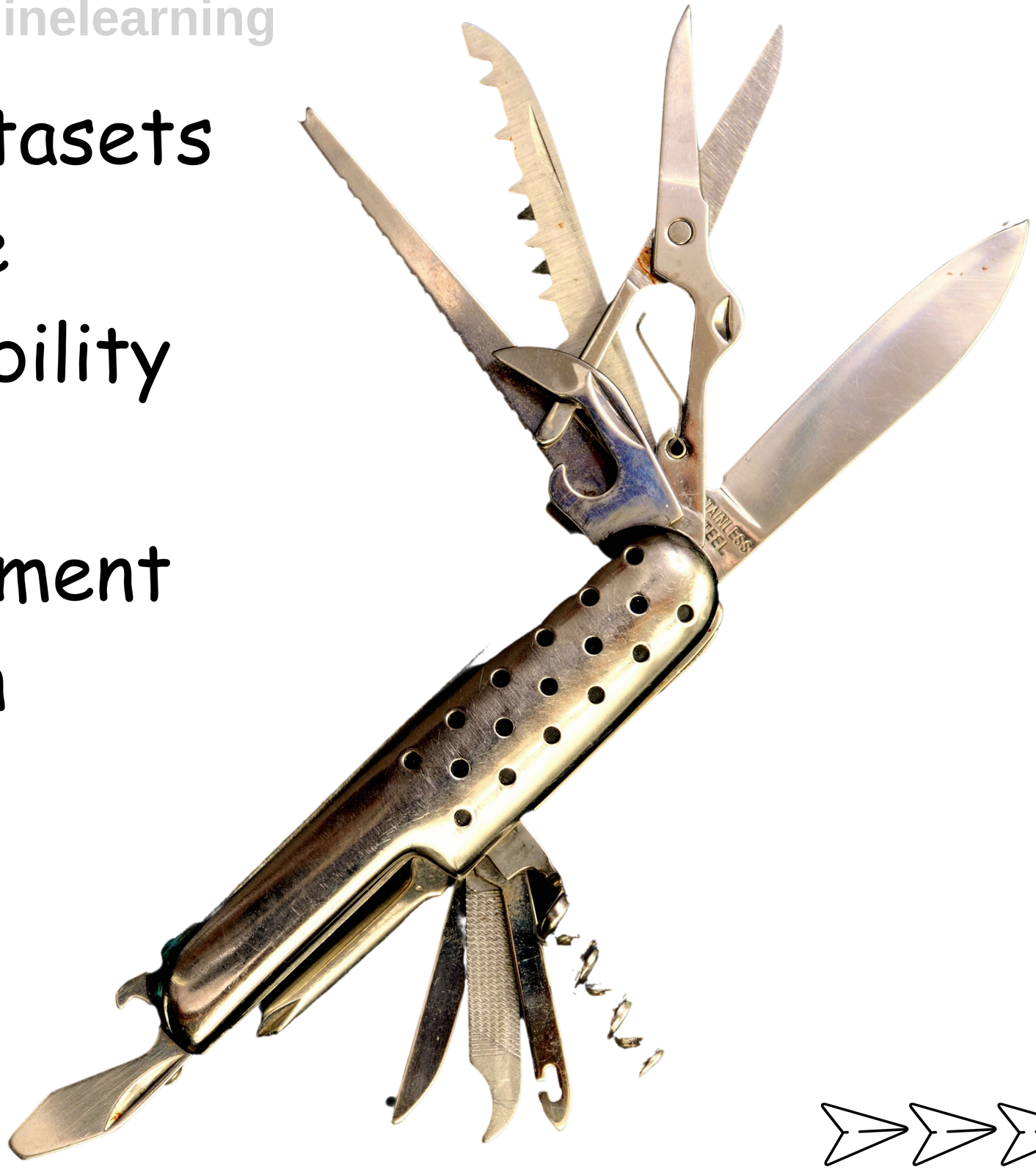
GUIDE TO BECOMING A SELF-TAUGHT DATA SCIENTIST 2020

@LEARN.MACHINELEARNING

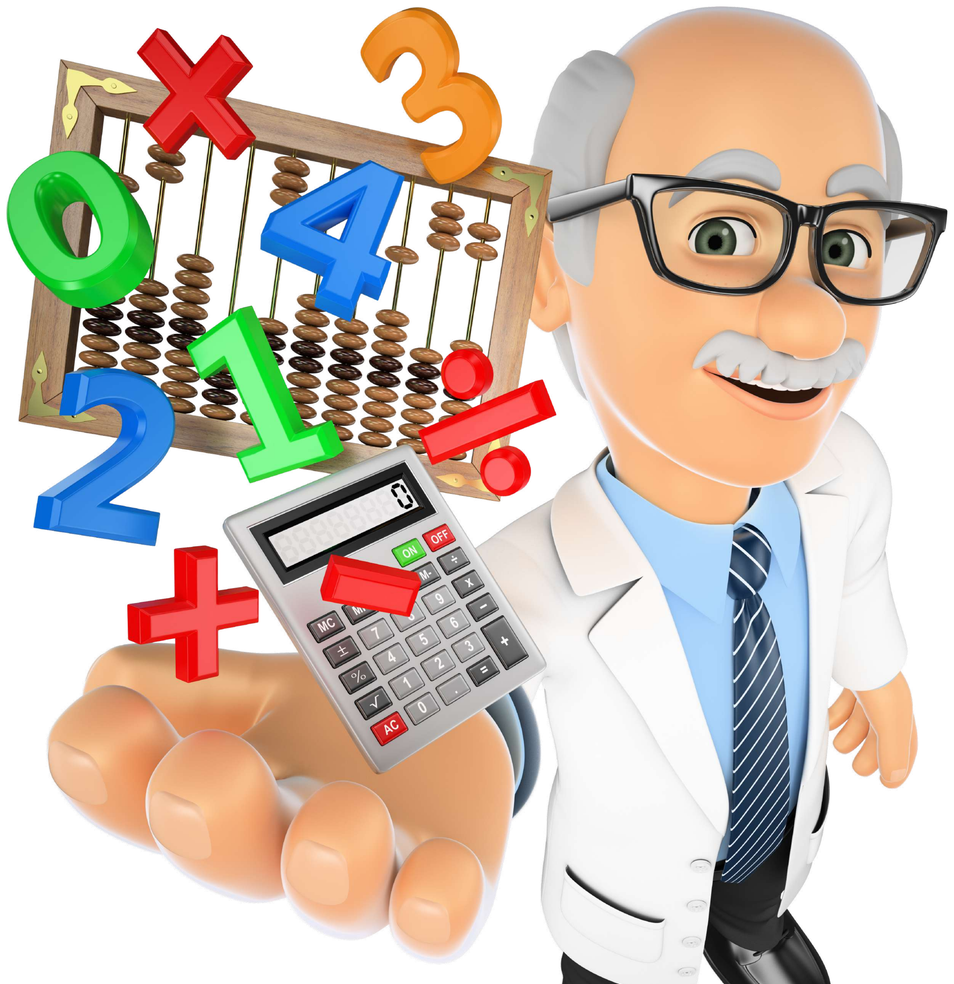
SKILLS FOR DATA SCIENTIST IN 2020

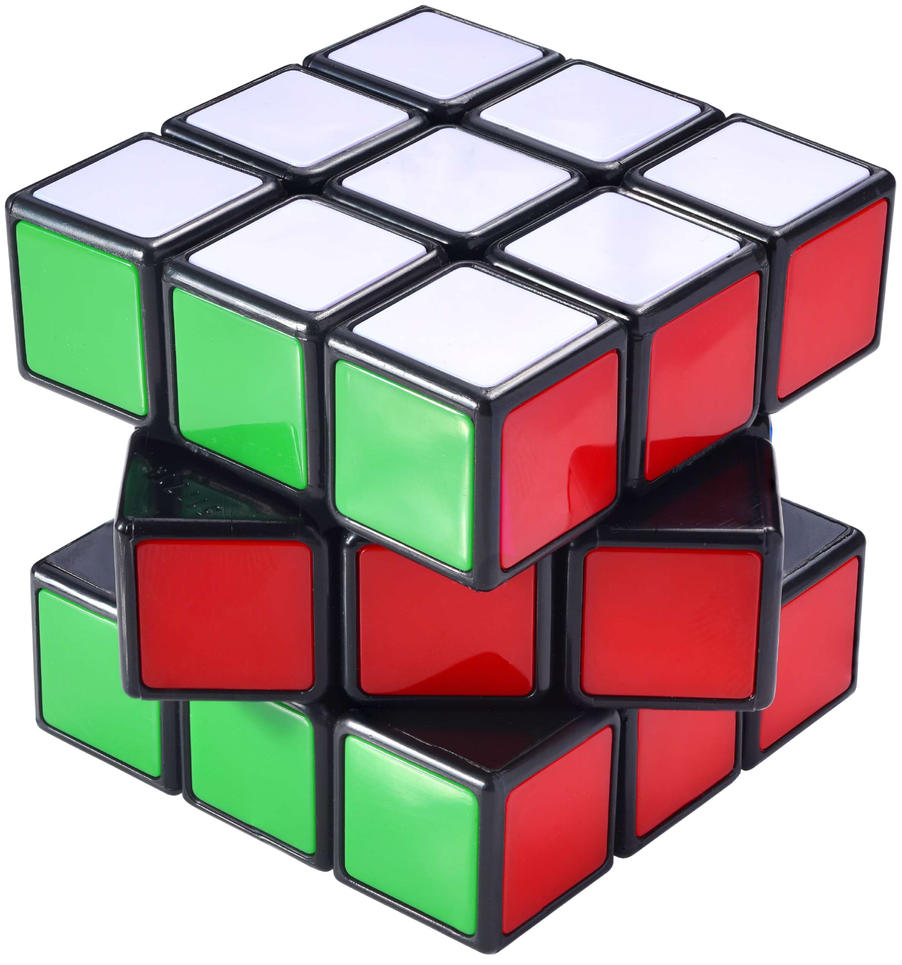
- Mathematical and statistical knowledge
- Good knowledge of machine learning algorithms
- Awareness on programming languages like Python and R which are more tuned for data science
- Handling large datasets
- Domain knowledge
- Problem-solving ability
- Data Wrangling
- Database Management
- Data Visualization
- Cloud Computing
- Microsoft Excel
- DevOps



MATH BASICS

- Multivariable Calculus
- Functions of several variables
- Derivatives and gradients
- Step function, Sigmoid function, Logit function, ReLU (Rectified Linear Unit) function
- Cost function
- Plotting of functions
- Minimum and Maximum values of a function





- Linear Algebra
- VectorsMatrices
- Transpose of a matrix
- The inverse of a matrix
- The determinant of a matrix
- Dotproduct
- Eigenvalues
- Eigenvectors

MATH BASICS

- Probability and Statistics Basics
- Mean, Median, Mode
- Standard deviation & variance
- Correlation coefficient and the covariance
- matrixProbability distributions (Binomial, Poisson, Normal)
- p-valueBaye's Theorem Confusion Matrix, ROC Curve)
- A/B Testing
- Monte Carlo Simulation

- Optimization Methods
- Cost function/Objective function
- Likelihood function
- Error function
- Gradient Descent Algorithm and its variants (e.g., Stochastic Gradient Descent Algorithm)

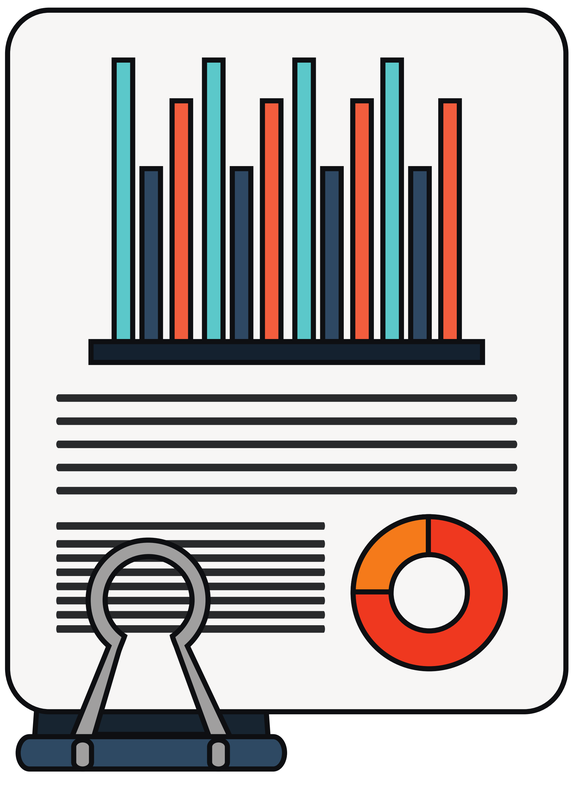
PROGRAMMING BASICS

- R
- Basic R syntax
- Foundational R programming concepts such as data types, vectors arithmetic, indexing, and data frames
- How to perform operations in R including sorting, data wrangling using dplyr, and data visualization with ggplot2
- R studio

- Python
- Basic Python syntax
- Object-oriented programming
- Jupyter notebooks
- Python libraries such as
- NumPy, pylab, seaborn
- matplotlib, pandas
- scikit-learn
- TensorFlow
- PyTorch
- etc

- **Learn data basics**
- **Learn how to manipulate data in various formats, for example, CSV file, pdf file, text file, etc.**
- **Learn how to clean data, impute data, scale data, import and export data, and scrap data from the internet.**
- **Some packages of interest are pandas, NumPy, pdf tools, stringr, etc.**
- **Additionally, R and Python contain several inbuilt datasets that can be used for practice.**
- **Learn data transformation and dimensionality reduction techniques such as covariance matrix plot, principal component analysis (PCA), and linear discriminant analysis (LDA).**

- Data Visualization Basics
- Data Component
- Geometric Component
- Mapping Component
- Scale Component
- Labels Component
- Ethical Component



MACHINE LEARNING BASICS

- Supervised Learning (Continuous Variable Prediction)
- Basic regression
- Multi regression analysis
- Regularized regression
- Logistic Regression Classifier
- Support Vector Machine (SVM)
- K-nearest neighbor (KNN) Classifier
- Decision Tree Classifier
- Random Forest Classifier
- Naive Bayes
- Gradient boosting
- etc

- Unsupervised Learning
- Kmeans clustering algorithm
- k - Median
- DBScan
- Hierarchical clustering
- etc...

BUILD UP YOUR ONLINE PRESENCE

- Write blogs
- Do projects an upload them on GitHub
  - Fork interesting repos
  - commit to other repos
- public speaking
- Youtube tutorials

- Share your experience on social channels.
- write books
- creat a course
- Twitch stream, or podcast.

NETWORKING

- Make friends
- Meet experts and talk with them
- Learn from experts
- Get a mentor
- Make yourself visible to outside world
- It also helps you to get a good job in your dream companies