DATA SCIENCE CHEATSHEET GITHUB

<https://github.com/abhat222/Data-Science--Cheat-Sheet>

Andrew Ng - Complete Strategy for Machine Learning and AI

1. 5 Key Skills you need to become a great Data Scientist:

[**https://lnkd.in/fcD\_UQm**](https://lnkd.in/fcD_UQm)

1. 50 Years of Data Science:

[**https://lnkd.in/fKGAUUJ**](https://lnkd.in/fKGAUUJ)

1. Mathematics for Machine Learning:

[**https://lnkd.in/fSiWYkg**](https://lnkd.in/fSiWYkg)

1. Why Data Science is becoming so popular?

[**https://lnkd.in/fEsPdUb**](https://lnkd.in/fEsPdUb)

1. Data Science Cheat Sheet - 10 pages:

[**https://lnkd.in/fkscYn5**](https://lnkd.in/fkscYn5)

1. Top 40 Python Interview Q&A:

[**https://lnkd.in/fYu7g87**](https://lnkd.in/fYu7g87)

1. Data Science Interview Questions - 48 pages:

[**https://lnkd.in/f-Ph5Zt**](https://lnkd.in/f-Ph5Zt)

1. TOP 50 ML Interview Q&A:

[**https://lnkd.in/fJqHtUH**](https://lnkd.in/fJqHtUH)

1. R Programming Cheatsheet:

[**https://lnkd.in/fGqdx87**](https://lnkd.in/fGqdx87)

1. TOP 10 sites for your Career:

[**https://lnkd.in/fTAJiwc**](https://lnkd.in/fTAJiwc)

1. Machine Learning: 112 Pages of Insightful Notes:

[**https://lnkd.in/f2mDPuj**](https://lnkd.in/f2mDPuj)

1. Time Series Analysis in Python:

[**https://lnkd.in/fe6h\_Vg**](https://lnkd.in/fe6h_Vg)

📍Best Data Science Courses Online🔖

Coursera

1. Stanford University

2. DeepLearning .ai

3. IBM

4. Johns Hopkins

5. University of Michigan

EdX

6. Harvard University

7. MIT

Udacity

8. Data Science Nanodegree

Data Science Blogs📈

1. Data Camp =

2. Data Science Central

3. KDnuggets

4. R-Bloggers

5. Revolution Analytics

6. Analytics Vidya

7. Codementor

8. Data Plus Science

9. Data Science 101

10. DataRobot

Top 10 Skills for Data Science

1. Probability & Statistics

2. Linear Algebra

3. Python

4. R

5. SQL/Presto

6. Tableau/PowerBI

7. AWS/Azure

8. Spark

9. Excel

10. DevOps

Top 10 Algorithms for Data Science

1. Linear Regression

2. Logistics Regression

3. K-means Clustering

4. PCA

5. Support Vector Machine

6. Decision Tree

7. Random Forrest

8. Gradient Boosting Machine

9. XGboost

10. Artificial Neural Networks

Top 10 Industries for Data Science

1. Technology

2. Finance

3. Retail

4. Telecom

5. Healthcare & Pharma

6. Manufacturing

7. Automotive

8. Cybersecurity

9. Energy

10. Utilities

🧮Statistics & Probability 📚

1. Khan Academy

2. OpenIntro

3. Exam Solutions

4. Seeing Theory

5. Towardsdatascience

6. Elitedatascience

7. OLI

8. Class Central

9. Alison

10. Guru99

🔏Free Data Sets🖇 <https://data-flair.training/blogs/machine-learning-datasets/>

1. Data.world <https://data.world/>

2. Kaggle <https://www.kaggle.com/>

3. FiveThirthyEight <https://data.fivethirtyeight.com/>

4. BuzzFeed <https://github.com/BuzzFeedNews>

5. Socrata OpenData

6. Data gov

7. Quandl

8. Reddit

9. UCI Repository

10. Academic Torrents

📇 Python📕

1. Code Academy

2. TutorialsPoint

3. Python org

4. Python for Beginners

5. Pythonspot

6. Interactive Python

7. Python Tutor

8. Full Stack Python

9. Awesome-Python

10. CheckiO

📊Visualization📉

1. Storytelling with Data

2. Information is Beautiful

3. Flowing Data

4. Visualising Data

5. Junk Charts

6. The Pudding

7. The Atlas

8. Graphic Detail

9. US Census & FEMA

10. Tableau Blog

10 Commandments of Data Science

1. Focus primarily on solving the problem and not on tools, technologies, and models.

2. Data will never be clean or easily available. Data gathering and cleaning will take 80% of your time and efforts.   
3. Don’t underestimate the power of Excel and SQL - they are still two of the most useful tools for data analysis.

4. Simple models such as Linear or Logistics Regression will be good enough for the majority of the problems. You don’t need neural networks to solve every problem.

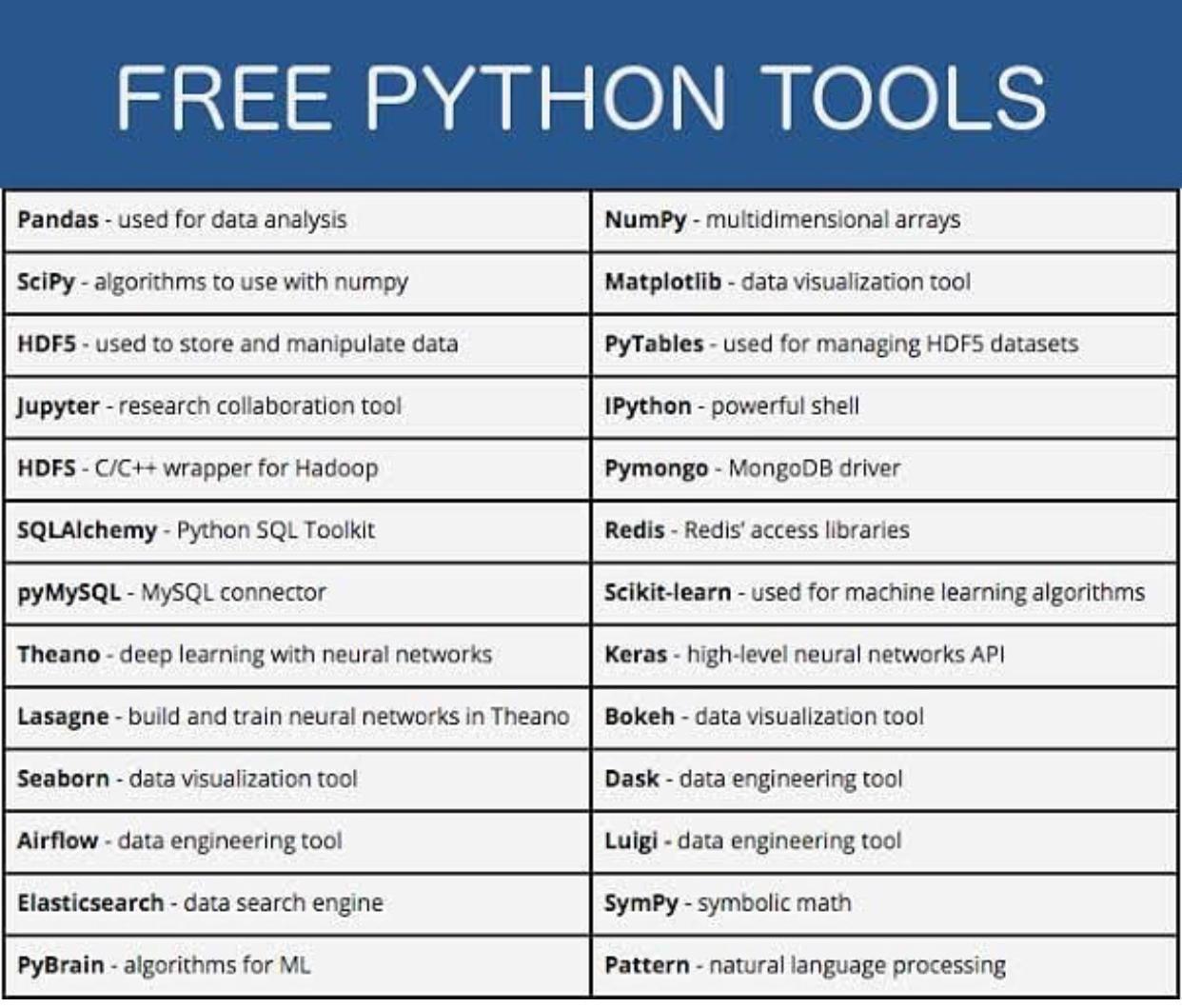
5. Textbook solutions will not work for most of the practical problems. You will need to try new approaches and innovate as required.

6. Nobody can remember everything. On the job, you can always use Google, Stack Overflow etc. 7. Learn Data Visualisation and develop the ability to explain your key insights in simple terms - these skills will be very useful with non-technical and business stakeholders.

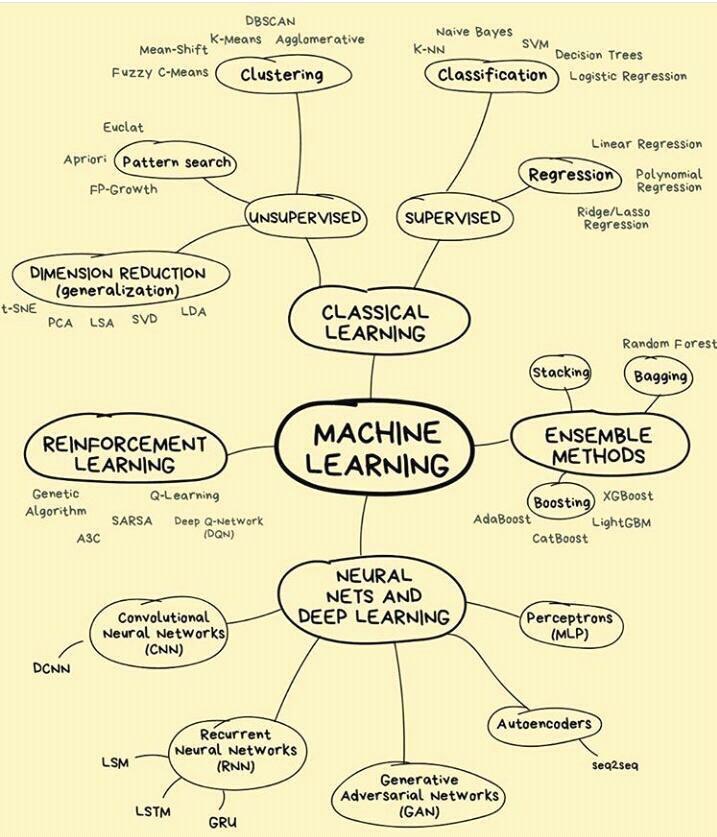
8. Learn PowerPoint and storytelling - people may not appreciate your great work if you can’t convince them with your communication.

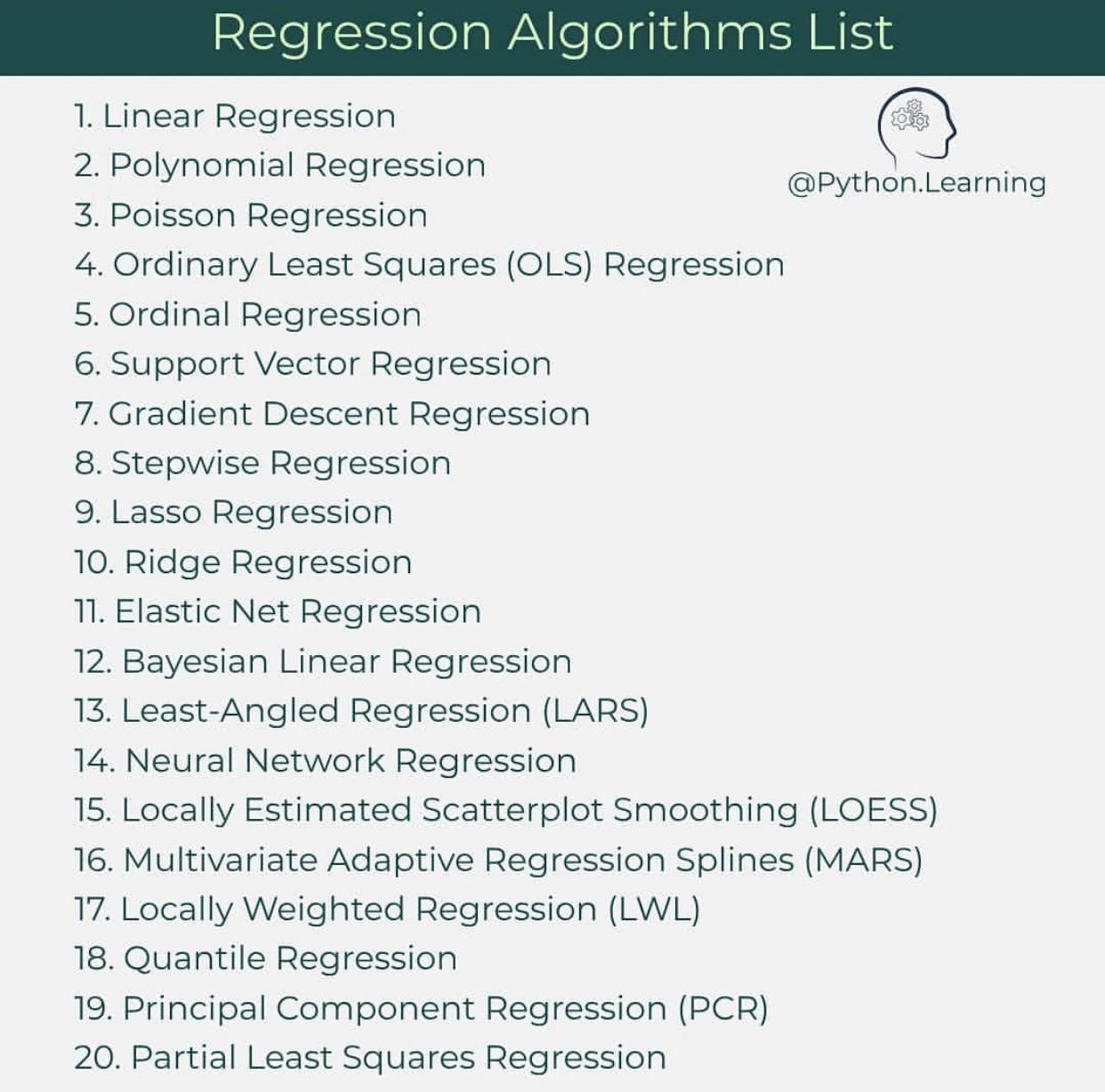
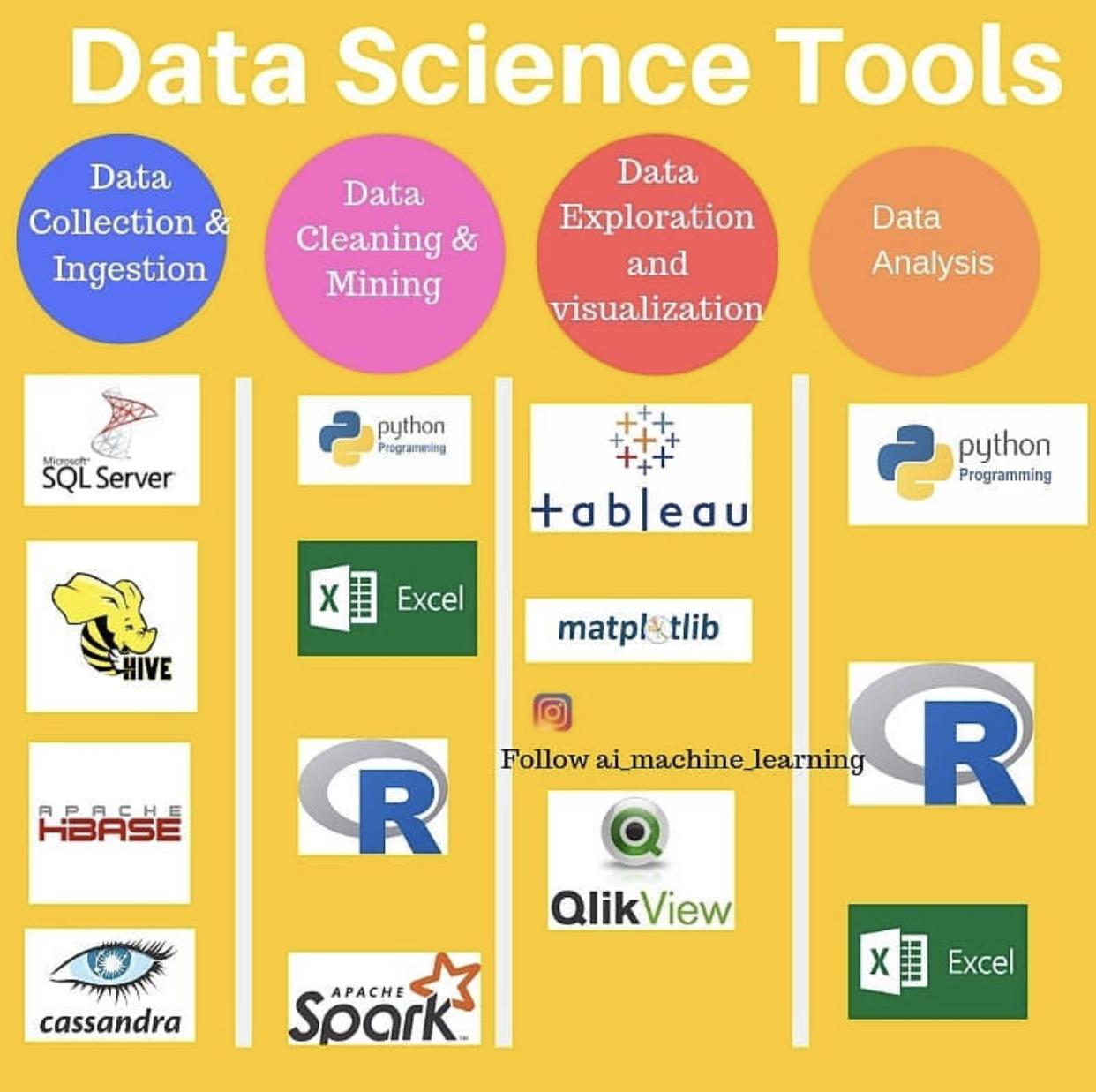
9. Data Science field is evolving rapidly. Please learn continuously, else you might become obsolete soon.

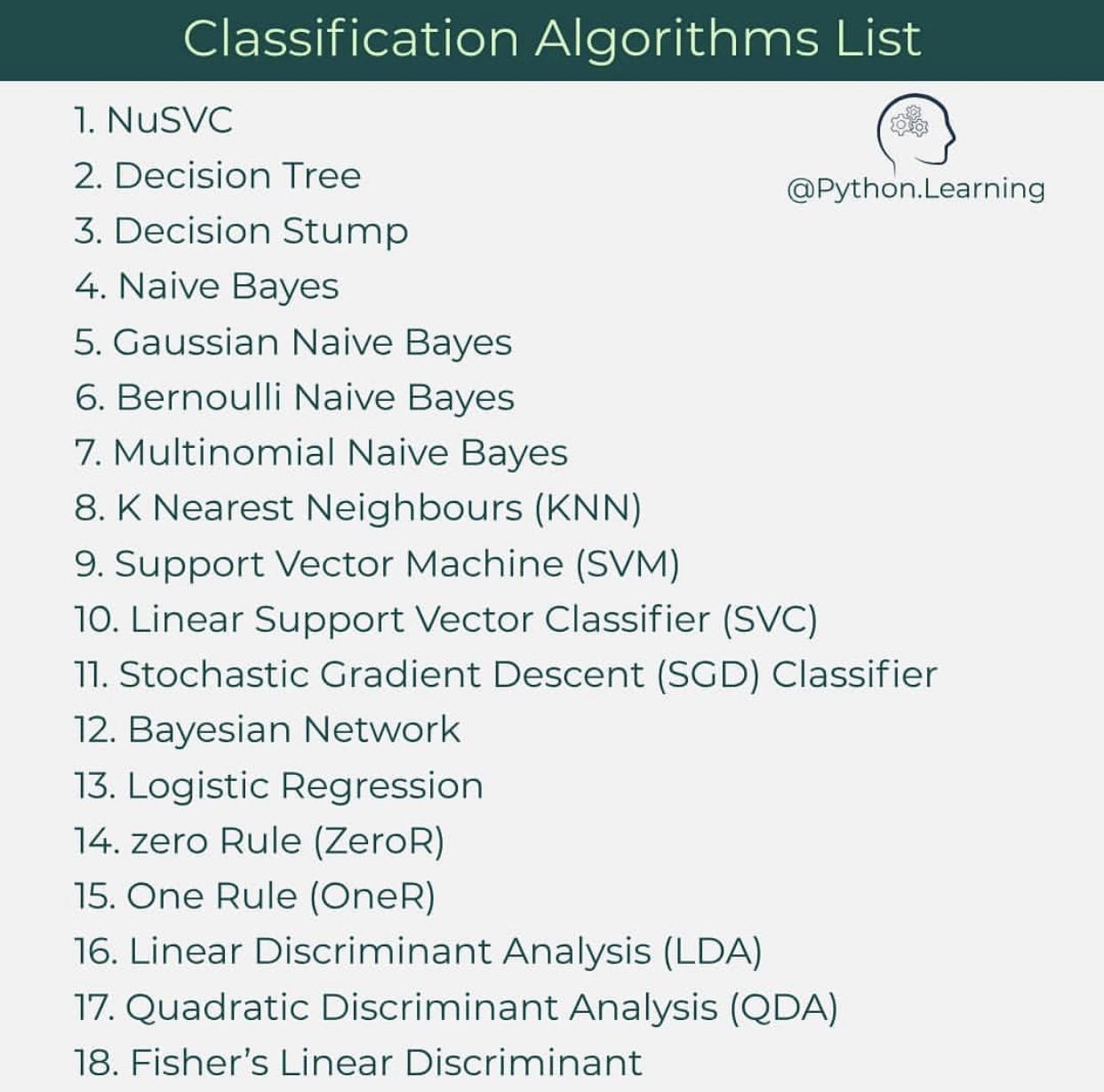
10. Focus primarily on solving the problem and not on tools, technologies, and models.



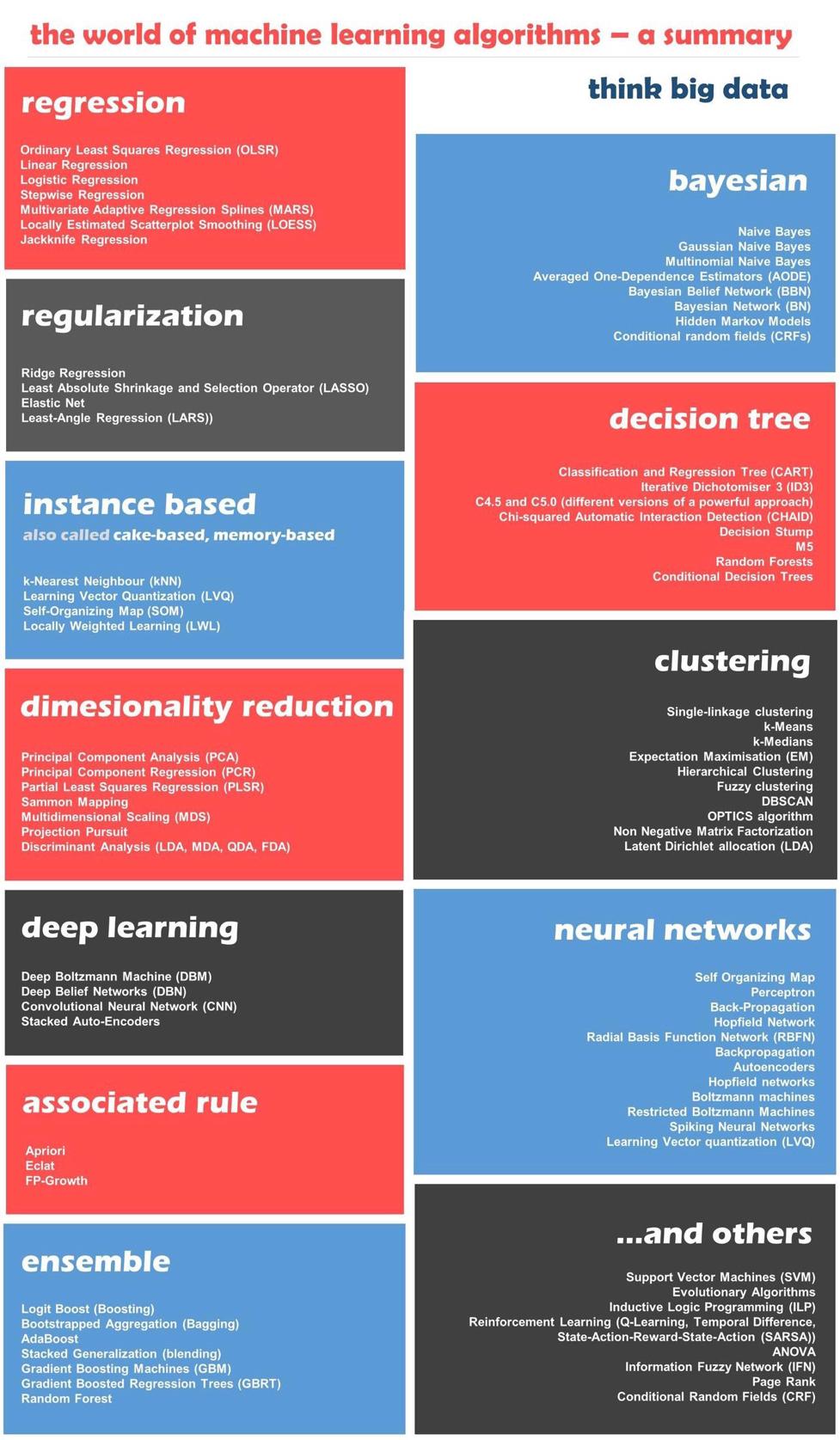


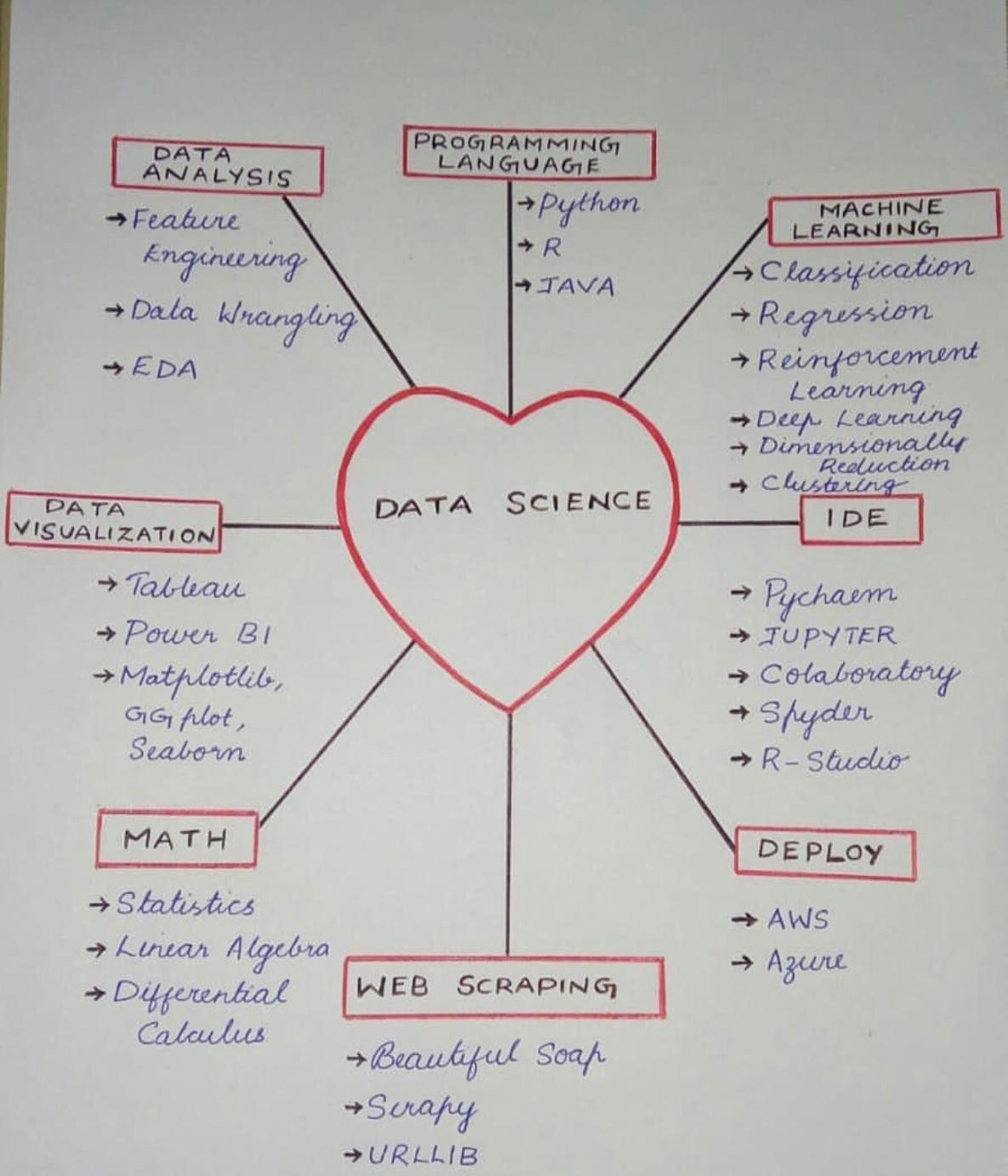


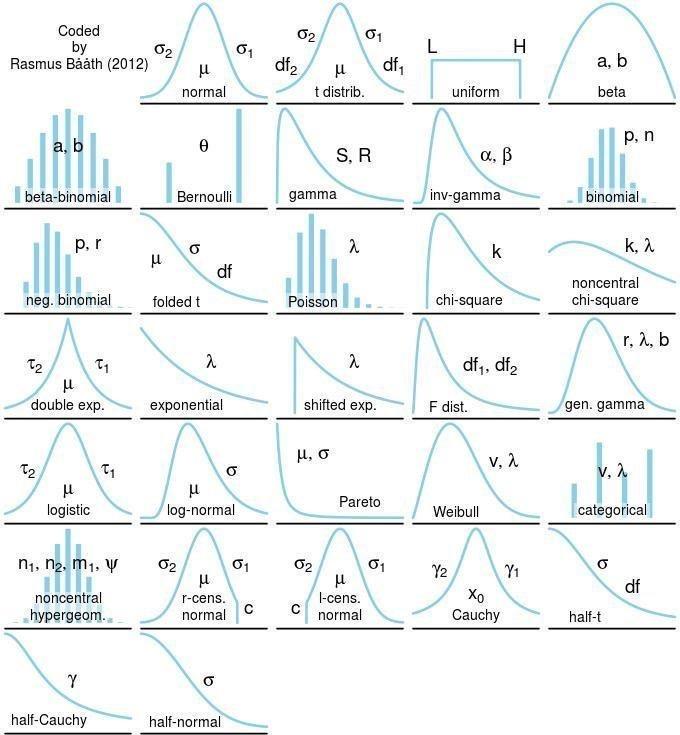












Data Visualisation Made Easy

3 Key questions to ask:

Question 1. What you want to show?

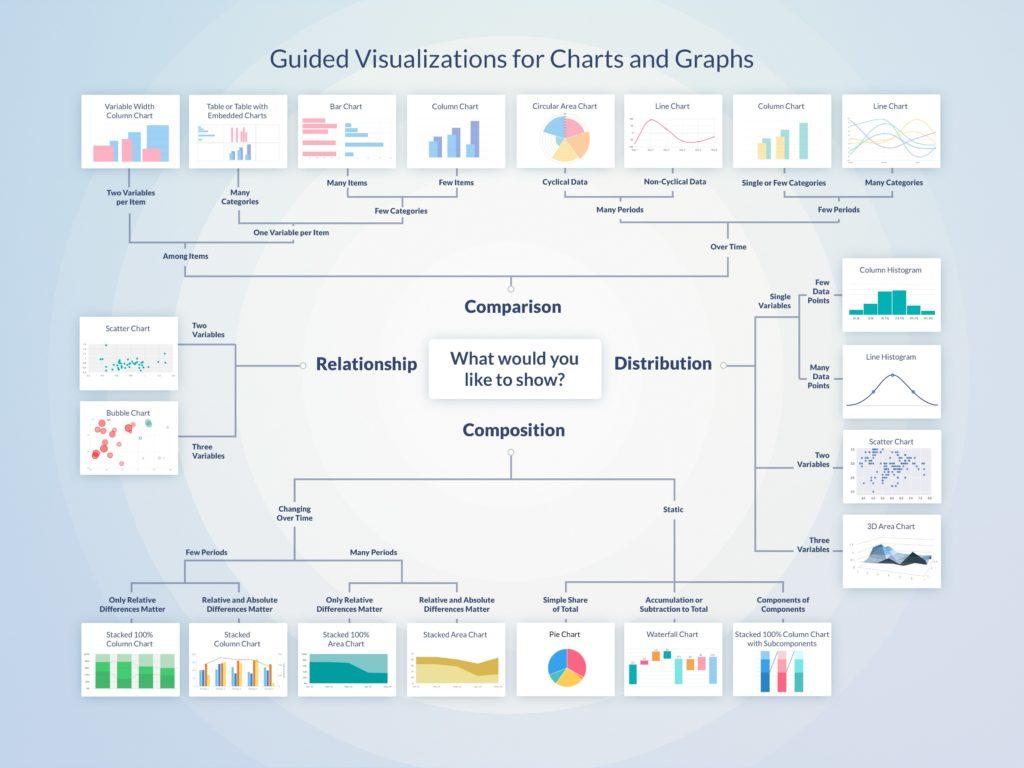
Comparison (most-common), Composition (very useful for time-series data), Distribution and Relationship

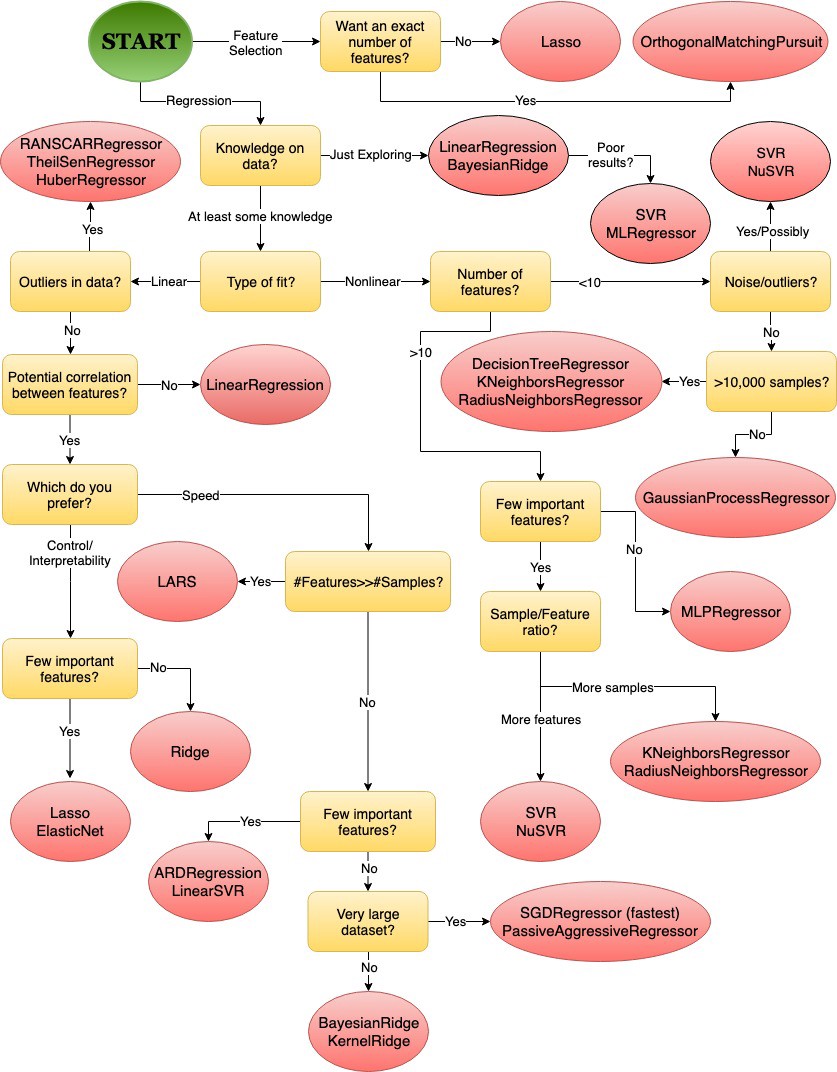
Question 2. How many variables you want to show?

Two variables is the most common. Sometimes you try to see distribution of one variable or movement of one variable over time. Try not to show more than three variables in one chart. For more variables, break the information into multiple charts that make logical sense.

Question 3. Is the data cross-sectional or time-series?

Example of cross-sectional data: age of different people at the same time Example of time-series data: population of India over the years For time-series consider bar charts to represent few time periods (< 10) and convert the similar charts to line charts for many time periods.





* **Projects**
* Personal Projects
* Hackathons
* Coding Challenges
* Open Source Projects
* **Computer vision:** CNN’s, segmentation, labeling, descriptions, object detection
* **Recurrent Networks:** Time series data such as the stock market and a video, LSTM cells
* **Reinforcement Learning:** teaching an agent to learn skills such as playing a video game or driving
* **Natural Language Processing:**Chat bots, sentiment analysis, content generation, content summarising
* **Generative Adversarial Networks:** Learning to generate content such as images, 3D models, learning policies, audio
* **Meta Learning:** Learning to learn
* **One Shot Learning:**learning with very little data
* **Neural Network Visualisation and Debugging:** Huge area of research, neural networks are still a black box and it is difficult for us to visualise them and understand why they don’t work when broken.
* <https://data-flair.training/blogs/python-projects-with-source-code/>

<https://www.quora.com/How-do-I-start-machine-learning-with-Python-What-should-I-do-if-I-am-a-beginner/answer/Sakina-Mirza-2?ch=3&share=a45da759&srid=GBzZ>

**Real Time Projects**

Projects From Retail ,Banking

,Finance ,Insurance ,Sales,Marketing

,Healthcare ,Manufacturing .

**Project 1** : Marketing Domain

Customer Conversion / Segmentation

Problem: A bank Facing Challenges With Lead Conversion

Description:

Identify the leads' segments having a higher conversion ratio

(lead to buying a product) so that organisation can

specifically target these potential customers through

additional channels and re-marketing

**Project 2** : Banking Domain

Credit Risk Analytics

Problem: efficiently build or validate in-house models for

credit risk management.

Description:

Create a classifier that leverages financial information from

bank accounts to estimate customer risk.

**Project 3** : Project on Natural Language Procession

Problem : training a machine learning model that classifies a

given line of text as belonging to one of the books/Articles.

developing a machine learning model (deep learning

preferred) for the same.

Project 4 : Price Analytics

Creating auto calculating pricing model

Problem: build an algorithm that automatically suggests the

right product prices

Project 5 : Classifying Loan Application

Problem : Work With credit dataset using classification

techniques like Decision Tree, Neural Networks etc to

classify loan applications

Project 6 : Identify And Predict Customer churn in

telecom industry

Description:Understand the customer behavior and reasons

for churn.Apply multiple classification models to predict the

customer churn in telecom industry

Project Lists

Project 7 :Retail Domain

Coupon Purchase Prediction Project

Description:Understand Retail Transactional Data set And

Using past purchase and browsing behavior of customers

,create a machine learning model which Predict which

coupons a customer will buy in a given period of time.

Project 8 : Predict Credit Default

Description:predict borrowers chance of defaulting on credit

loans by building a credit score prediction model.Develop a

good prediction model for a bank so that they can provide

maximum credit to individual without exceeding the risk

threshold.

Project 9 : Manufacturing And Production

Predict Internal Failures Using Production Line Dataset

Description:Understanding about Manufacturing domain

and its failures.

Use production line dataset to predict internal failures using

thousands of measurements/tests made for each component

along the assembly line

Project Lists

Project 10 : Insurance Purchase Prediction

Description:Predicting which insurance option the customer

will choose.Building machine learning models and Using a

customer’s shopping history, can you predict what policy they

will end up choosing?

Project 11 : HR Analytics Employee Attrition &

Performance

Description:Predict attrition of valuable employees of an

organisation .

Project 12 : Implement Back-Propagation Algorithm for

Classification Problems

Description:Implement Back-propagation Algorithm from

scratch for classification problems.After this project you willl

have understanding of "How to apply the back-propagation

algorithm to a real-world predictive modeling problem."

**Interview Question topics**

**Data exploration**

* How do you summarize the distribution of your data?
* How do you handle outliers or data points that skew data?
* What assumptions can you make? Why and when? (i.e When is it safe to assume "normal")

**Confidence intervals**

* How they are constructed
* Why you standardize
* How to interpret

**Sampling**

* Why and when?
* How do you calculate needed sample size? [Power analysis is advanced]
* Limitations
* Bootstrapping and resampling?

**Biases**

* When you sample, what bias are you inflicting?
* How do you control for biases?
* What are some of the first things that come to mind when I do X in terms of biasing your data?

**Modeling**

* Can you build a simple linear model?
* How do you select features?
* How do you evaluate a model?

**Experimentation**

* How do test new concepts or hypotheses in....insert domain X? i.e. How would evaluate whether or not consumers like the webpage redesign or new food being served?
* How do you create test and control groups?
* How do you control for external factors?
* How do you evaluate results?
* **Statistics**: Confidence intervals, parameter estimation, p-value, hypothesis testing.
* **Common metrics**: Engagement / retention rate, conversion, similar products / duplicates matching, how to measure them.
* **Useful cost functions**: Log-loss, other entopy-based, DCG/NDCG, etc.
* **Basic machine learning**: Classification / regression / ranking problems, overfitting, convex optimization, trees, ensembles, boosting, collaborative filtering, etc.
* **Tools**: R / Python / Mathematica, Weka & similar. Code up something yourself would help too, Kaggle is very useful.
* **Mathematics and complexities**: Eigenvectors, singular values, PCA, LDA, Gibbs Sampling, Information Bottleneck et. al.
* **Real-life numbers and intuition**: Expected user behavior, reasonable ranges for user signup / retention rate, session length / count, registered / unregistered users, deep / top-level engagement, spam rate, complaint rate, ads efficiency.
* Accuracy
* Sensitivity, True Positive Rate, Recall
* Specificity, True Negative Rate
* Precision
* Understand the trade-offs for choosing sensitivity over precision in a model or vice versa.
* Some of the most often asked interview questions are around concepts like sampling, central limit theorem, p-value, null hypothesis, ANOVA, correlation vs causation, standard deviation and measures of dispersion, mean vs median, random variables, expected values and variance, distributions, GLMs, Bayesian stats, etc. Some examples are: -
* - Name a probability distribution other than Normal and explain how to apply this probability?
* - How best to select a representative sample of search queries from 5 million?
* - The mean heights of men and women in a population were calculated to be mM and mW. What is the mean height of the total population?
* - Three friends in Seattle told you itâs rainy. Each has a probability of 1/3 of lying. What’s the probability of Seattle is rainy?
* - How do you detect if a new observation is an outlier?
* - What is the Central Limit Theorem? Why it's important?
* - Explain the difference between mean and median.
* - Define variance
* -What is covariance? Where is it used?
* -What is the goal of A/B Testing?
* - What is sampling? Why do we need it? What is stratified sampling?

**Probability**

* Conditional probability
* Independence
* Random variables

**Inferential Statistics (how to interpret these)**

* covariance
* correlation
* test statistics
* p-values / power
* confidence interval
* analysis of variance (ANOVA)

**Experiment Design / Sampling**

* Population, study population, sample. unit, attribute
* Type of error in experiment
* Randomization, treatment, control, blocking
* population inference from sample data
* sampling methods (i.e. bootstrap, with/out replacement)
* missing values
* sample standard deviation / error

**Linear regression**

* Least square algorithm
* Likelihood algorithm
* Model assumptions
* Model fitting
* Model comparison
* Model selection
* Interactions
* What is / How to handle heteroscedasticity problem
* how to handle categorical variables

**Logistic regression**

* parameter estimates
* assumptions
* categorical variables
* interaction

**Other models**

* general algorithm (trees, SVM, boosting, bagging)
* how to implement them in XXX language
* validation methods (i.e. cross-validation, X-fold)

**Analytical**

* case study
* given a situation, how will you design experiment, collect data, analyze and report result

1. What *is* the curse of higher dimensionality? What is the difference between density-sparse data and dimensionally-sparse data? What does "higher dimensionality" imply when applying textbook clustering algorithms developed for low dimension metric spaces to, say, numerical text analysis? Think of using cluster density to identify "good" clusters.  
   2. Probability space: How do you compare probabilities? Probability space is often referred to as a "vector space", is it? Define addition on this space? Is ds^2 = dp^2 a good metric on probability space? How would you construct a metric on probability space, so you can start doing complicated things like L2 norm etc.?
2. Explain what is a bootstrap and how to use it.
3. Explain the basic concepts of bayesian inference: prior, posterior, etc.
4. Explain the Central Limit Theorem and what it means.
5. Explain ANOVA and Chi-Square.
6. Explain ARIMA.

**Be prepared to code**

* SQL: There is no excuse for being weak in SQL as a Data Scientist.
* General coding: You should be comfortable writing code with Python, or R like you use them everyday.

**Be prepared to talk about data science / machine learning**

* Algorithms: you don’t need to know all of them, just the ones you usually use.
* Techniques: things like feature engineering, evaluation metrics, cross validation, or how to prevent overfitting…
* Past experience, what you built, what was the impact, what did you learn from.

1. Regression (the short answer here is to know everything) (Always)  
- p-values, interpretation of coefficients, interaction coefficients  
- know how factor variables are dummy coded, coefficient interpretation  
- linear regression assumptions  
- residual analysis, adding non-linear terms  
- difference between L1 and L2 penalization, when you would use each one  
- how to deal with highly correlated variables  
- walk me through how you would approach building a regression model to predict Y given you have data X  
- how do you choose the right number of predictors  
- logistic regression  
  
2. General predictive modeling (Always)  
- train/test/validate sets, cross-validation, parameter selection, predictor selection, etc.  
  
3. Random Forrest (Always)  
- how a tree is grown  
- purpose of growing a forrest  
- purpose of selecting a subset of variables at each split  
- pruning  
- variable importance  
- knowing gbm here is also a plus  
  
4. Matrix Factorization (Rare)  
- pca, when do you use it? how are the components ordered?  
- read up on how factorization was used in netflix competition  
- i've never been asked about factor analysis, but it is great to know and makes other related questions easier  
  
5. K-means clustering (often)  
- what is the loss function? when does it converge?  
- know the algorithm iteration steps  
- how do you choose the right number of clusters?  
- is the optimization convex?  
  
6. Standard Stats (Often)  
- t-test  
- z-scores  
- chi-square test  
- know covariance and correlation equations  
  
7. Time Series (Rare)  
- p,d,q parameters and how you choose them  
- unit root and box test  
  
I've also been asked a series of increasingly difficult SQL queries in one interview.

Coursera – Andrew Ng Course

Projects: Kaggle :

ML projects : basic projects

College machine learning course

Any online Data Science Course

Python:

Data structures

Coursera : 5 courses specialisation

Hackerrank : easy and then medium : main

ML basics; ML real time projects

1. Kaggle
2. Titanic project : Logistics; Linear
3. University of ML ka course
4. Hackerrank codes for Python
5. Pandas; Numpy; matplotlib; sklearn (for ML models)
6. 2 projects after this : from some written assignment (interview)

Interview:

1. Python : DS;  libraries;
2. ML : supervised; unsupervised;
3. Projects;
4. Pandas function and its outputs;
5. Data Handling : data remove; populate mean data, etc.
6. Logistics se kiya —> svm se raise ho skta
7. Company usage