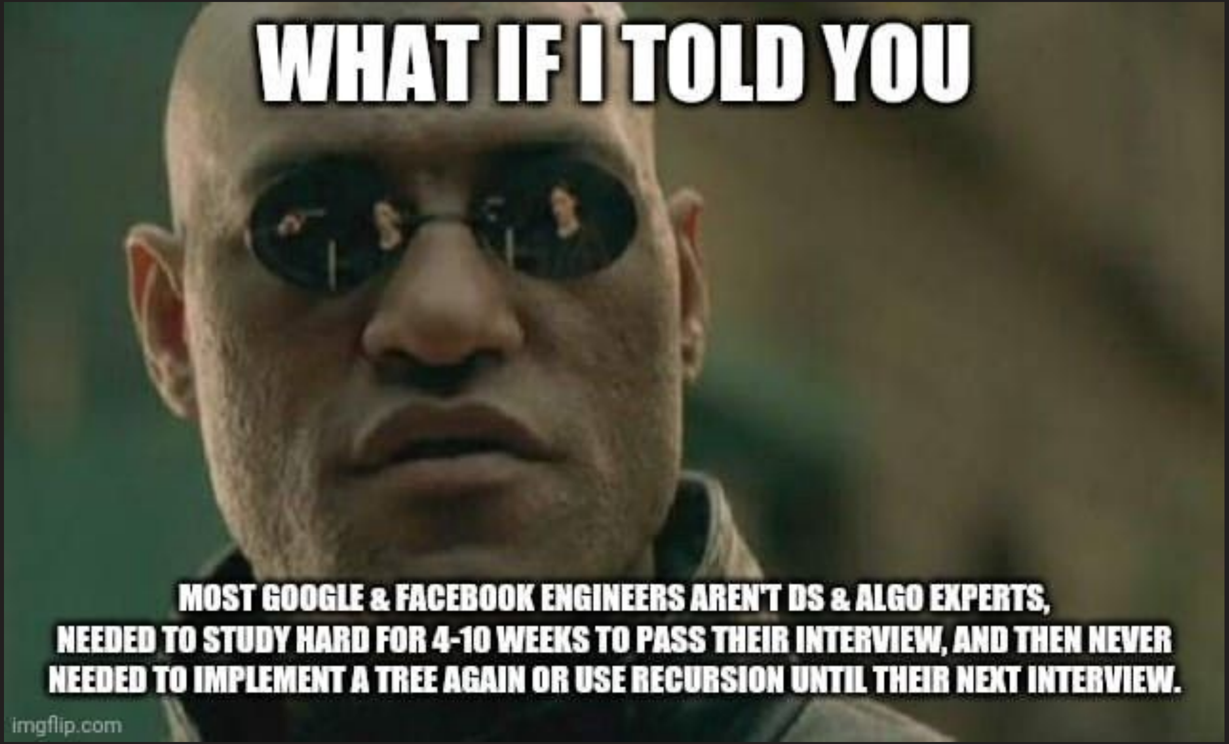
Data Science/ ML Interviews

Interview Pillars
&
Resources

...the never ending list

BUT FIRST...



WHAT IF I TOLD YOU
MOST GOOGLE & FACEBOOK ENGINEERS AREN'T DS & ALGO EXPERTS,
NEEDED TO STUDY HARD FOR 4-10 WEEKS TO PASS THEIR INTERVIEW, AND THEN NEVER
NEEDED TO IMPLEMENT A TREE AGAIN OR USE RECURSION UNTIL THEIR NEXT INTERVIEW.
imgflip.com

If you have no idea whatsoever..

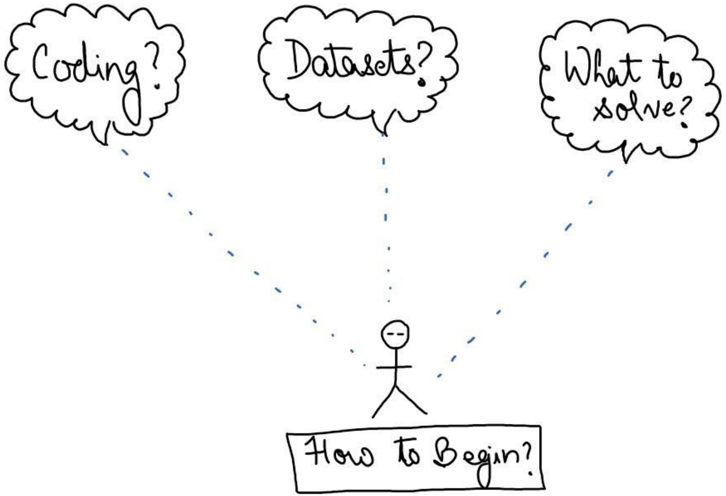
Just do these courses to get started:

[Machine Learning A-Z (Python & R in Data Science Course)]

[Learn Python for Data Structures, Algorithms & Interviews]

..Create a github repo. E.g. this

.. Start solving past and present Kaggle competitions, starting with forecasting using LGBM, xgboost, Bayesian Optimization



1. Python/ Java Data structures > LeetCode
2. Data Engineering > SQL / NoSQL window function (lag/lead)
3. Statistics > Probability [Bayes Theorem] + Z score, Expected Values + Markovian Chain Principles
4. AB Testing > Bonferroni Correction, what happens with multiple metrics, OEC, MVT, Joint Distribution
5. Product Sense > health of a product, improving the product, root cause analysis
6. SWE System Design > Design Twitter
7. ML System Design > Design a Personalized News Feed Rank
8. Live Coding of data / modeling issue
9. Writing algorithms from scratch
10. Behavioral Interviews
11. Data Science Concepts
12. MLE Concepts such as NER, Deep Learning

Disclaimer: All this is based on my experience of failing many many interviews and endlessly reading stuff on Reddit, Blind and LC

Resources

1. Leetcode is leetcode > consistency is key [topics vary from DS to MLE]
    a. https://leetcode.com/discuss/career/450215/How-to-use-LeetCode-to-help-yourself-efficiently-and-effectively-(for-beginners)
    b. https://www.mrcodeswildride.com/challenges/algorithms

2. SQL Leetcode; selected questions with solutions: https://hamza50.gitbook.io/leetcode/

   https://www.youtube.com/watch?v=-WEpWH1NHGU [start with this]

   https://www.interviewquery.com/questions

3. Statistics:
    a. Readings | Introduction to Probability and Statistics | Mathematics [must do]
    b. Welcome! | STAT 414 [if you're feeling luck]

4. A/B Testing

Sample Size, Power, alpha	https://classroom.udacity.com/courses/ud257/lessons/4018018619/concepts/40043986970923
Multiple Metrics	Bonferroni Correction
Bonferroni correction - multiple testing	Trustworthy online experiments [you can find a pdf]
Type 1, type 2 errors	
Experimental design - randomization unit, MDE	
MVT	
ABn Testing, multi arm bandits	
Simpson's paradox	https://docs.microsoft.com/en-us/archive/msdn-magazine/2005/december/test-run-software-testing-paradoxes
Sample Ratio mismatch	https://classroom.udacity.com/courses/ud257/lessons/4085798776/concepts/40713087720923
quasi experiment	
Z-test, T-test, Anova, Ancova, chi sq	
Non normal AB Test	https://www.interviewquery.com/questions/non-normal-ab-testing
Proportion Testing	
Summary of AB Testing	https://towardsdatascience.com/the-as-and-b-s-of-a-b-testing-a-beginner-s-guide-to-experimentation-d54a60218e13

5. Product Sense

- An important metric goes down, how would you dig into the causes?
- What metrics would you use to quantify the success of youtube ads (this could also be extended to other products like Snapchat filters, twitter live-streaming, fort-nite new features, etc)
- How do you measure the success or failure of a product/product feature
- Google has released a new version of their search algorithm, for which they used A/B testing. During the testing process, engineers realized that the new algorithm was not implemented correctly and returned less relevant results. Two things happened during testing:
  - People in the treatment group performed more queries than the control group.
  - Advertising revenue was higher in the treatment group as well.
  - What may be the cause of people in the treatment group performing more searches than the control group? There are different possible answers here.
- Product Manager Interview Questions
- The Product Manager Interview

Resources

6. System Design : Enough Tutorials on YouTube [Tech Dummies is great!]

7. ML System Design: www.boringbot.xyz

8. Live Coding on ML Problems: Kaggle Kaggle Kaggle! [learn how to write precision/recall from scratch]

9. Writing Algorithms from scratch: hamzafarooq/algos: Building ML Algorithms ground-up

10. Behavior : Amazon Leadership Principles

11/12. DS + ML Concepts: 100 Page ML Book

Bonus link for Data Science Concepts: ISLR Textbook Slides, Videos and Resources

DS Skills	Description	Expectations
Data Querying	Ability to write queries involving not limited to SQL/MySQL/Hive etc. for joining datasets, summarizing and aggregating from large scale databases	Minimum Expectation (fixed) : Different types of Joins, when are they used, Group By, Distinct , UNIONS, basic sub queries , comparators Good to haves (depending on level) : window functions, date/time manipulations, string formatting, running totals, pivoting, lag, lead operations.
Statistics	Understanding of statistical intuition behind samples, population & hypothesis testing	Minimum Expectations (fixed) : Central limit theorem , Different statistical distribution (top 3) uses cases, p-values, confidence intervals, linear regression, basic parametric tests like z test, t-test etc. Good to haves (depending on level) : Effect size, power analysis, sampling techniques, top 10 statistical distributions use cases, L1/L2 regularizations understanding.
Machine Learning	Understanding of how Machine learnings works , algorithms and thought process behind different top ML algos.	Minimum Expectations (fixed) : Understanding of over/under fitting, training/test/validation set, ability to deal with uncleaned data, How trees, clustering, logistic regression and dimension reduction work, cross validation and evaluate which algorithm is better. Good to haves (depending on level) : Tree Pruning, bootstrapping, ensemble models, boosting, ROC curves, parameter tunings, when and how to balance between accuracy and interpretability of ML models.

DS Skills	Description	Expectations
Fundamentals of Programming	Experience in writing basic programs using any language, understanding of basic data structures.	Minimum Expectation (fixed) : pseudo code of common problems, loops , counters, edge cases , space and time complexity (basics) of any approach, ability to write common programs of finding area of triangle, palindrome etc. Good to haves (depending on level) : Object oriented programming, dictionaries & hash maps , breaking down problem to sub problems, solid understanding of big O notations, experience in writing big programs, experience in version control/git.
Applied Math & Probability	Solid fundamentals of high school math and numbers , probability	Minimum Expectation (fixed) : understanding of permutation, combination, fundamentals properties of probabilities, bayes theorem, basics of linear algebra and matrix. Good to haves (depending on level) : optimization, inflection point intutition.

Different kind of Roles

- Product Analyst vs data analyst vs business analyst
- Research vs applied research
- Research scientist vs research engineer
- Data scientist vs machine learning engineer

Product Analyst	Data Analyst
Metrics for Product growth/health	Pretty much a Data Analyst
Ex: Our MAU, DAU are down by 10%	Ex: derive insights for all different kind of users across the board
Focus on **immediate** commercial outcome	Focus on **long term** outcome

Data scientist	ML Engineers
Extract knowledge and insights from structured and unstructured data	ML models learn from data -> ML is part of data science
Use data to help company make decisions	Develop models to turn data into products
Is a scientist -> engineering isn’t a top priority	Is an engineer -> engineering is a top priority

Caveats

- MLEs at startups might spend most of their time wrangling data, understanding data, setting up infrastructure, and deploying models instead of training ML models.

Research	Applied research
Find the answers for **fundamental** questions and expand the body of theoretical knowledge.	Find solutions to **practical** problems
Ex: develop a new learning method for unsupervised transfer learning	Ex: develop techniques to make that new learning method work on a real world dataset
Focus on **long term** outcome	Focus on **immediate** commercial outcome

Caveats

- Cutting-edge research is spearheaded by big corporations
- Lacking theories to explain methods that work well empirically


ONE DOES NOT SIMPLY
END A PRESENTATION WITHOUT A
THANK YOU SLIDE
memegenerator.net