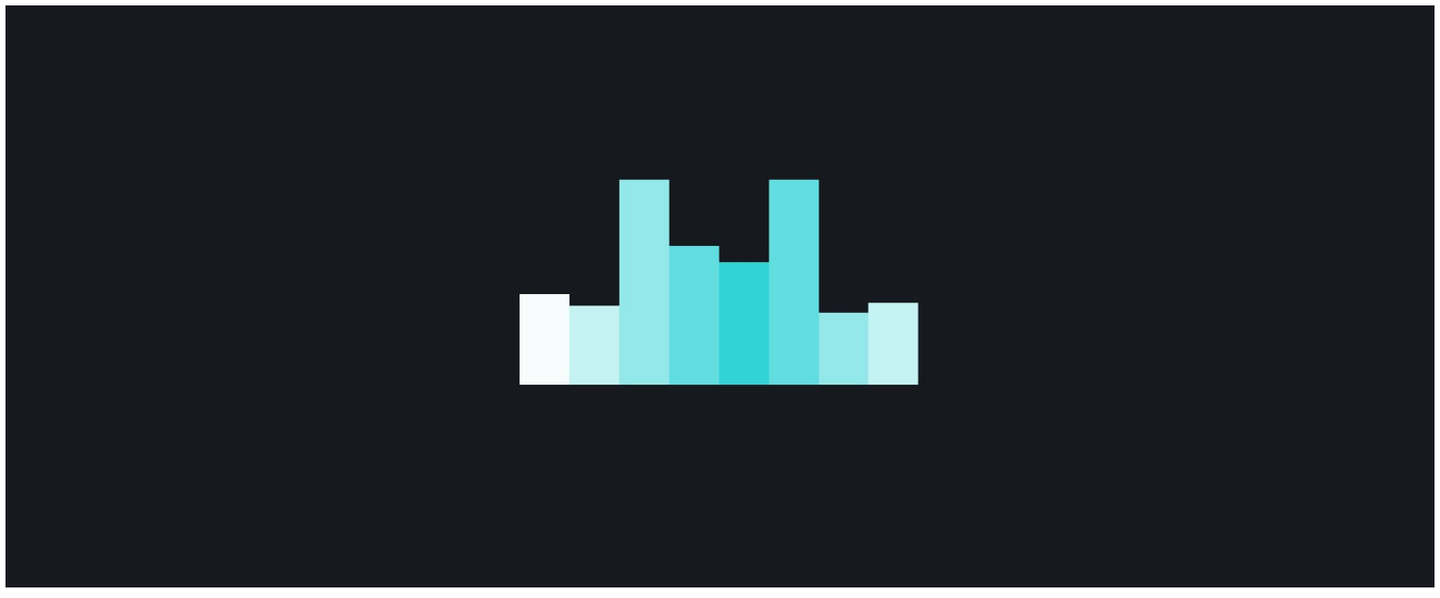
Data Science Workshop + Panel

Presented by UCLA Statistics Club
Compiled by: Krystal Xu



- Part 1: Overview of DS internships in tech
- Part 2: Securing an interview
- Part 3: Technical Interview Questions Go-through
- Part 4: Panel discussion

DS Internship Trend for Undergrad Students

- Increasing distinction between data science analytics and data science machine learning/deep learning

- Large tech firms often already have this distinction while in smaller tech firms you may still do both

- Analytics-focused: business sense, experimentation (A/B testing), data manipulation (SQL/R/Python), simple ML modeling (Regression/Random Forest/Clustering)

- ML-focused: Advanced machine learning & deep learning, translating research paper into code, having published papers. **Mainly reserved for phDs**

Overview on Large Tech Companies' DS recruitment

- **Facebook:** DS analytics (Undergrad+Masters) VS DS core (phDs). DS analytics focuses on product, SQL, A/B tests. General hire.

- **Quora:** DS vs ML (phDs) internships. DS interviews VERY FOCUSED on A/B testing. General hire.

- **Airbnb:** 3 tracks: analytics, inference (experimentation), algorithms (ML/DL/optimization). Virtually no undergrads in algorithms track. A few undergrads in analytics track. General hire.

- **Twitter:** also analytics focused for undergrads/masters. Mainly hires masters students. Team hire, not general hire.

- **Uber:** more well-rounded. ML/experimentation/analytics/data engineering. Team hire, not general hire. Currently only hiring phD interns

List of Data Science/Analytics Tech Internships Still Available

FB, Twitter, Quora, Twitter, Lyft, Quora, Walmart, Asana, Zillow, Adobe, Thumbtack, Expedia, Homeaway, Spotify, Ticketmaster, Wework, Splunk, Paypal, Blizzard, ZestFinance, Buzzfeed

Note: The name of the roles may differ but they should all involve data science and analytics

Securing A DS Interview: Resume

- **Myth:** Over-emphasizing on the fancy ML model you built
- **Myth:** Over-emphasizing on SWE skills e.g. web dev, database management, automation, web scraping etc
- **Formula:** Identified/Analyzed (problem statement) containing (this amount of data), proposed a methodology consisting of (some sort of ML modeling/dashboard/statistical analysis) using (technology tool/package), resulting in (quantify the impact).
- **Tip:** Look up on job descriptions
- **Tip:** Include SQL, Python and R in your skills
- **Key deliverables you can include:** identified/defined metrics such as…, built a dashboard, performed A/B tests/hypothesis testing and shipped a feature
- **Key algorithms you can include:** k-means, Random Forest, Logistic Regression, Gradient Boosting, clustering/segmentation, NLP (word2vec, topic modeling, TF-IDF etc)
- **Key projects you can include if you have done it before:** fraud detection, customer segmentation, predicted ads ROI/click-through rate, forecasted demand, calculated customer lifetime value, used NLP to do sentiment analysis, predicted churn rate/customer adoption rate, recommendation engine/ranking algorithm, pricing analytics

Securing A DS Interview: Networking

- Networking events: career fair, Grace Hopper Conference, OutforUndergrad Conference, campus events

- LinkedIn: connect with alumni and recruiters

- 3 ways to get an interview:

    A. Referral from existing employees or past interns

    B. Directly contacting the recruiter

    C. Just apply online (very slim chance)

Interview Process

1. HR Phone Screen

2. Take-home assignment

3. 45-min Technical Interview

4. Onsite/Remote onsite interviews with 2-3 interviews back-to-back

Note: This is just a general interview process and each company is different. Interns VS New grads hiring can also be different.

HR Screening

- Purpose:
    A. Find out if you have relevant experience
    B. If you know enough of the company and its DS
    C. If you are good at communications and are likable
- How to prep:
    D. Prepare your “Tell Me About Yourself” and “Why this company” pitch
    E. Research on the company.
    F. Do not talk about overly technical stuff, but aim to explain clearly

Take-Home Assignment

- Possible Questions:

  A. Identify important factors that contribute to user adoption (random forest/XGBoost etc)

  B. Predict housing valuation (regression/random forest etc)

  C. Create categories of grocery items that are likely to be bought together (clustering)

Take-Home Assignment

- Be succinct and only write relevant things. The report is often a high level overview with the target audience being the top management

- Translate your model findings into easily understandable business terms that anyone can understand

- Include relevant data visualizations

- Always present your data exploration findings first before presenting ML model findings

- Always check for data quality e.g. missing data, outliers. There might be traps.

- Always give suggestions/personal opinions on how to improve the business processes

- Comment your code in a detailed way

Technical Interview Components

- **Product analytics:** metrics, features, business sense in general
- **Data manipulation:** SQL(!!), R or Python
- **A/B testing:** basically hypothesis testing, but must know 1) how to explain things like p-value and power to a layman; 2) common pitfalls in industry A/B testing
- **Basic probability:** binomial, poisson, geometric, Bayes theorem, mean, median, mode, but the context of the question might be an industry-specific one
- **Basic ML:** logistic regression, KNN, k-means, decision tree, random forest, bias-variance trade-off, precision-recall, regularizations

Technical Interview: Product Analytics

- **Question:** You are launching a messaging app. Define metrics that you would choose to monitor app performance during the first few months.

    - Always start by targeting growth as the high level goal and narrow down to more specific factors that impact growth -> a) acquiring new users, b) retaining current users

    - Acquiring new users: new sign ups per day from users who send at least 1 message within the first 2 days

    - Retaining current users (engagement): average messages per day per user; average number of contacts per user per day. -> Pick the key action you want your users to perform

Technical Interview: Product Analytics

- **Question:** Which variables are important to predict a fake listing on eBay?

  - 2 categories: a) characteristics of the listing; b) characteristics of the seller

  - Listing: low resolution/pirated picture; low price; clickbait; generic/copied description

  - Seller: Same device id/IP address/bank account creates multiple eBay accounts (might be bots), bad ratings or artificially inflated ratings, suspicious browsing behavior

Product Analytics Tips

- Try out the product yourself and note its features
- Important to standardize your metrics. e.g. average number of comments **per user per day**
- Categorize your metrics or variables to make them more organized. e.g. behavioral vs demographics; quantity vs quality
- When thinking of metrics, always relate back to the biggest goal of the company in terms of growth (new users), retention (user engagement), monetization
- If asked if a feature should be shipped, focus on actions that users are already performing today because this shows demand for the feature
- If asked to optimize a long term metric like lifetime value, find a short term metric as proxy that can predict the long term one, and focus on optimizing that

Technical Interview: Data Manipulation

Table: user_actions

Column Name	Example Value	Description
date	08-08-2018	timestamp
user_id	Integer value	Unique user
post_id	Integer value	Unique post
Action	View, like, report, comment	What action the user performed for the post
Extra	Love, spam, nudity	Extra reasons for reporting

How many posts were reported yesterday for each report reason? List in descending order

```
SELECT Extra AS reason, COUNT(DISTINCT post_id) AS total_reports
FROM user_actions
WHERE Action='report' AND date= DATESUB(day, 1, getdate())
GROUP BY 1
ORDER BY 2 DESC
```

Table: user_actions

Column Name	Example Value	Description
date	08-08-2018	timestamp
user_id	Integer value	Unique user
post_id	Integer value	Unique post
Action	View, like, report, comment	What action the user performed for the post
Extra	Love, spam, nudity	Extra reasons for reporting

Table: reviewer_removals

Column Name	Example Value	Description
date	08-08-2018	timestamp
reviewer_id	Integer value	Unique reviewer
post_id	Integer value	Unique post

What % of daily content that users view is actually spam?

1. **LEFT JOIN** user_action with reviewer_removal on post_id
2. Filter Action='View'
3. Group by date in user_actions
4. Find total spams/total posts

```
SELECT u.date,
COUNT(DISTINCT r.post_id)/
COUNT(DISTINCT u.post_id) AS
spam_percentage
FROM user_actions u
LEFT JOIN reviewer_removals r
ON u.post_id = r.post_id
WHERE u.Action = 'view'
GROUP BY 1
ORDER BY 1
```

SQL Tips

- Think out loud and communicate. Do it slowly step-by-step. Ask clarifying questions

- Be familiar with self join, left join, inner join, outer join, union, union all

- Be careful with DISTINCT

- Be comfortable with NULL

- Know how to write subqueries

- Know window functions e.g. SUM(…) OVER (PARTITION BY … ORDER BY …). Other window functions include RANK(), ROWNUMBER(), NTILE(), LAG(), LEAD()

Technical Interview: A/B Testing

- **Question:** Quora has a new homepage recommendation algorithm. Should we ship this new feature?
    - Depends on the result of A/B test. Here are the few steps for A/B test:
        A. Pick a metric: e.g. click-through rate of questions
        B. Determine duration of experiment and sample size/blast radius
        C. Randomize users into treatment and control
        D. Pick a statistical test: 2-sample 2-tailed t-test
        E. Calculate p-value and conclude

A/B Testing Follow-up Questions

- How do we determine duration of the experiment?
    - Statistical perspective: determine sample size by determining significance level, power, effect size, and variance. Then, determine duration from sample size.
    - Business perspective: Need to run for at least 2 business cycles to account for seasonality, and for novelty effect to wear off. Blast radius also affected by how risky the new feature is
- Why use t-test and what are the assumptions?
    - t-test used because population variance is unknown
    - Assumption: That sample size is large enough such that according to CLT, normal distribution can be approximated
    - Assumption: That the users within each treatment and control group are independent and are not affected by one another’s behavior

A/B Testing Follow-up Questions

- How to make sure users are independent in social networks like FB and marketplace like Uber where one user's actions impact other people
    - Social network: Network analysis. Cut users into individual clusters with no inter-cluster interaction. Need to find comparable clusters with similar metric values
    - Marketplace: Treatment as one city and control as another comparable city. This applies to marketplaces that transact locally for supply and demand interactions such as Uber (but not Airbnb)
- What if the distribution is highly skewed? How can we still use t-test?
    - Cap the values/exclude outliers
    - Log transform (but not recommended since results are hard to interpret)
- If p-value < 0.05 for a 2-tailed test for click-through rate, can we ship then?
    - Check guard rails metrics such as ads revenue, monthly active users to confirm they aren't adversely affected
    - Calculate p-value using different user segments such as gender, geography etc. to make sure no one segment is particularly adversely affected

A/B Testing tips

- Need to know how to explain p-value, power, false positive, false negative, confidence interval in layman, business terms

- When picking metrics, establish an overall evaluation criteria (OEC). Also establish guard rail metrics and supporting metrics

- Consider potential conflicts with other experiments

- Be familiar with t-test and chi-square test

Technical Interview: Probability

- There are 1000 users and we select 10 users **without replacement** to be in the treatment group for an A/B test. On average, what's the number of times before a user gets picked to be put into the treatment group?
    - Without replacement means each user is not independent. There are 1000/10=100 total iterations, with each user being equally likely to be picked in every iteration, hence the answer is just (1+2+3+…+100)/100=50.5 which represents the mean

- How about **with replacement**?
    - Users are independent. Then it becomes a geometric distribution and the mean is 1/p where p=10/1000 = 0.01 and 1/p= 100

Probability tips

- Need to know binomial, poisson, geometric, Bayes' Theorem
- Need to have an idea on how certain distributions look like e.g. How do you think the distribution of number of retweets per post looks like? answer: log-normal ish with a long tail

Technical Interview: Basic ML

- For a real estate advertising tech firm, how do we determine if a user wants to rent an apartment? Assuming we have historical data of renters.

    A. Feature engineering

    B. Explore the data and find out important features

    C. ML Preprocessing e.g. handling missing values, one-hot encoding for categorical variables, feature scaling for continuous variables

    D. Separate into training and testing data

    E. Pick a supervised classifier and e.g. random forest, KNN

    F. Tune the hyper parameters and evaluate the classifier

    G. Determine a probability threshold to classify users as renters

ML Follow-up Questions

- What features should you engineer?
    - Behavioral: Has the user checked out rental listings?
    - Demographics: Age, gender, occupation etc. of user
- Prediction works well for training data but not well for testing data. Why?
    - Overfitting. Remedy: 1) use k-fold cross validation, and/or 2) regularize by tuning the parameters of the random forest. e.g. limit maximum allowable tree depth; cut down on number of trees etc. Use grid search to optimize the hyper parameters
- How to evaluate your model?
    - ROC curve that plots true positive against false positive rate or confusion matrix etc.
- Let’s say the output probability for one user to be a renter is 0.49. Do you classify him to be a renter or not?
    - Yes if cost of false negative is higher than false positive, meaning the cost of not detecting a renter is very high

ML Questions Tips

- Do not jump straight into modeling. Explain what features you want to get first, and also how you can go about exploring the data

- Need to substantiate why you chose your model. Supervised vs unsupervised; classification vs regression; robust to noise vs not robust

- Need to have an idea of how your model works and what are the hyper parameters of your model

- Need to be clear about overfitting/underfitting and how to counteract them

Conclusion

- Analytics, A/B testing and SQL are most important. Some companies don't test ML e.g. FB
- A lot of ML problems, A/B testing and metrics are written by tech companies on their data science blogs e.g. Experiments at Airbnb https://medium.com/airbnb-engineering/experiments-at-airbnb-e2db3abf39e7
- A/B testing resources: Udacity free course for basics, Microsoft research papers for advanced stuff and pitfalls, Stats100B for hypothesis testing
- SQL resources: Mode Analytics
- Analytics resources: Lean Analytics book, A Collection of Data Science Take-Home Challenges ($299 comes with 40 product questions with answers, 20 take home challenges w/o answers, 6 SQL questions with answers)
- ML resources: Hands-On Machine Learning with Scikit-Learn and Tensorflow book, Andrew Ng Coursera course
- General interview prep: Glassdoor, forums