White Boarding & Paired Coding

Lindsay Warrenburg
The Erdős Institute

Schedule

1. Review from first session
2. Overview of white boarding & paired coding
3. Practice interview questions in breakout rooms
4. Questions
5. Sign-ups

Review

Interview Topics

- Machine Learning
- Statistics & Probability
- Computer Science
- SQL
- Business

Interview Formats

- Demo Projects
- White Boarding
- Paired Coding
- Case Studies
- Data Challenge

Workshop 1 “Homework”

- Research the different types of data science jobs and figure out which type(s) best align with your interests, values, and career goals

Data Engineering	Computer Vision	Data Product
Software Engineering	Genomics & Computational Biology	Decision Science
Computer Science	Human Factors Engineering	Business Intelligence
Machine Learning Research	Speech Processing & Audio Engineering	User Experience & Customer Insights
Machine Learning Engineering	Natural Language Processing (NLP)	Market Research

- Go through the self study materials and start prepping

White Boarding

White Boarding

- The interviewers will ask a question about a data science topic and you will explain it to them, using a whiteboard to draw pictures, write equations, write sample code

White Boarding

- The interviewers will ask a question about a data science topic and you will explain it to them, using a whiteboard to draw pictures, write equations, write sample code
- The goal is for the interviewer to understand the boundaries of your knowledge -- if you get stuck, that’s okay! They’ll likely keep asking harder questions until you don’t know anymore. No one interviewing will know all of the answers to these questions.

White Boarding -- Statistics / Prob

One in a thousand people have Disease X. The test for Disease X is 98% correct in testing for Disease X. On the other hand, the test has a 1% error rate if the person being tested does not have Disease X. If Roman tests positive for Disease X, what are the odds he has Disease X?

White Boarding -- Statistics / Prob

One in a thousand people have Disease X. The test for Disease X is 98% correct in testing for Disease X. On the other hand, the test has a 1% error rate if the person being tested does not have Disease X. If Roman tests positive for Disease X, what are the odds he has Disease X?

A = Event that Roman **has** Disease X

B = Event that Roman **tests positive** for Disease X

White Boarding -- Statistics / Prob

One in a thousand people have Disease X. The test for Disease X is 98% correct in testing for Disease X. On the other hand, the test has a 1% error rate if the person being tested does not have Disease X. If Roman tests positive for Disease X, what are the odds he has Disease X?

A = Event that Roman **has** Disease X

B = Event that Roman **tests positive** for Disease X

Goal = P(A | B)

White Boarding -- Statistics / Prob

One in a thousand people have Disease X. The test for Disease X is 98% correct in testing for Disease X. On the other hand, the test has a 1% error rate if the person being tested does not have Disease X. If Roman tests positive for Disease X, what are the odds he has Disease X?

A = Event that Roman **has** Disease X

B = Event that Roman **tests positive** for Disease X

Goal = $P(A|B)$ ⇒ Need Bayes' Theorem!

White Boarding -- Statistics / Prob

One in a thousand people have Disease X. The test for Disease X is 98% correct in testing for Disease X. On the other hand, the test has a 1% error rate if the person being tested does not have Disease X. If Roman tests positive for Disease X, what are the odds he has Disease X?

Bayes' Theorem = $P(A|B) = [P(B|A) * P(A)] / P(B)$

White Boarding -- Statistics / Prob

One in a thousand people have Disease X. **The test for Disease X is 98% correct in testing for Disease X.** On the other hand, the test has a 1% error rate if the person being tested does not have Disease X. If Roman tests positive for Disease X, what are the odds he has Disease X?

Bayes' Theorem = $P(A|B) = [P(B|A) * P(A)] / P(B)$

$P(B|A) = 0.98$

White Boarding -- Statistics / Prob

One in a thousand people have Disease X. The test for Disease X is 98% correct in testing for Disease X. On the other hand, the test has a 1% error rate if the person being tested does not have Disease X. If Roman tests positive for Disease X, what are the odds he has Disease X?

Bayes' Theorem = $P(A|B) = [P(B|A) * P(A)] / P(B)$

$P(B|A) = 0.98$

$P(A) = 0.001$

White Boarding -- Statistics / Prob

One in a thousand people have Disease X. The test for Disease X is 98% correct in testing for Disease X. **On the other hand, the test has a 1% error rate if the person being tested does not have Disease X.** If Roman tests positive for Disease X, what are the odds he has Disease X?

Bayes' Theorem = $P(A|B) = [P(B|A) * P(A)] / P(B)$

$P(B|A) = 0.98$

$P(A) = 0.001$

$P(B|A')$ = Event that someone does **not** have the disease = 0.01

White Boarding -- Statistics / Prob

One in a thousand people have Disease X. The test for Disease X is 98% correct in testing for Disease X. **On the other hand, the test has a 1% error rate if the person being tested does not have Disease X.** If Roman tests positive for Disease X, what are the odds he has Disease X?

Bayes' Theorem = $P(A|B) = [P(B|A) * P(A)] / P(B)$

$P(B|A) = 0.98$

$P(A) = 0.001$

$P(B|A')$ = Event that someone does **not** have the disease = 0.01

P(B) = $P(B|A) * P(A) + P(B|A') * P(A') = 0.98 * 0.001 + 0.01 * 0.999 =$ **0.01097**

White Boarding -- Statistics / Prob

One in a thousand people have Disease X. The test for Disease X is 98% correct in testing for Disease X. **On the other hand, the test has a 1% error rate if the person being tested does not have Disease X.** If Roman tests positive for Disease X, what are the odds he has Disease X?

Bayes' Theorem = $P(A|B) = [P(B|A) * P(A)] / P(B)$

P(A|B) = 0.98 * 0.001 / 0.01097 = **8.93%**

Paired Coding

Paired Coding

- Similar to a white board question except the interviewer watches you coding (or codes with you!)

Paired Coding

- Similar to a white board question except the interviewer watches you coding (or codes with you!)
- This type of interview is one of the easiest to conduct online so I think it is more likely during Covid than other types of interviews

Paired Coding -- Basic Python

Print out items that start with 'abc'

```
l =['abc_1','abc_2','de_1','de_2','fg_1','fg_2']
```

Paired Coding -- Basic Python

Print out items that start with 'abc'

```
l =['abc_1','abc_2','de_1','de_2','fg_1','fg_2']
```

```
[x for x in l if x.startswith('abc')]
```

Paired Coding -- Computer Science

Create a function that sorts an array using the bubble sort method. What is its space and time complexity?

Paired Coding -- Computer Science

Create a function that sorts an array using the bubble sort method. What is its space and time complexity?

```
def bubbleSort(arr):
    for ii in range(len(arr)-1):
        for jj in range(len(arr)-1):
            if arr[jj] > arr[jj+1]:
                arr[jj], arr[jj+1] = arr[jj+1], arr[jj]
    return(arr)
```

Paired Coding -- Computer Science

Create a function that sorts an array using the bubble sort method. What is its space and time complexity?

```
def bubbleSort(arr):
    for ii in range(len(arr)-1):
        for jj in range(len(arr)-1):
            if arr[jj] > arr[jj+1]:
                arr[jj], arr[jj+1] = arr[jj+1], arr[jj]
    return(arr)
```

Time complexity: $O(n^2)$
Space complexity: $O(1)$

Breakout Rooms: Paired Coding

whiteboardfox.com

Paired Coding

1. **Given two arrays, write a function to get the intersection of the two. For example, if A = [1,2,3,4,5] and B = [0,1,3,7] then you should return [1,3].**
2. **Assume you are given the tables below containing information on trades and users. Write a query to list the top three cities that had the most number of completed orders**

trades

column_name	type
order_id	integer
user_id	integer
price	float
quantity	integer
status	string (“complete, “cancelled”)
timestamp	datetime

users

column_name	type
user_id	integer
city	string
email	string
signup_date	datetime

Answers: Paired Coding

Given two arrays, write a function to get the intersection of the two. For example, if A = [1,2,3,4,5] and B = [0,1,3,7] then you should return [1,3].

```
def intersection(a,b):
    set_a = set(a)
    set_b = set(b)

    if len(set_a) < len(set_b):
        return [x for x in set_a if x in set_b]
    else:
        return [x for x in set_b if x in set_a]
```

Assume you are given the tables below containing information on trades and users. Write a query to list the top three cities that had the most number of completed orders

```
SELECT u.city, COUNT(DISTINCT t.order_id) AS num_orders
FROM trades t
     JOIN users u ON t.user_id = u.user_id
WHERE t.status = 'complete'
GROUP BY city
ORDER BY num_orders DESC
LIMIT 3
```

Breakout Rooms: White Boarding

whiteboardfox.com

White Boarding

1. Describe what Type I and Type II errors are, and the trade-offs between them
2. Say that you are running a multiple linear regression and that you have reason to believe that several of the predictors are correlated. How will the results of the regression be affected if several are indeed correlated? How would you deal with this problem?
3. Compare and contrast gradient boosting and random forests.
4. Describe some advantages and disadvantages of relational databases vs. NoSQL databases
5. If 70% of Facebook users on iOS also use Instagram, but only 50% of Facebook users on Android also use Instagram, how would you go about identifying the underlying reasons for this discrepancy in usage?

Answers: White Boarding

Describe what Type I and Type II errors are, and the trade-offs between them

- Related to hypothesis testing
- Type I = reject null hypothesis, but null hypothesis is true
  - False positive
  - We “detect” a difference between groups when there is no difference
- Type II = do not reject null hypothesis, but alternative hypothesis is true
  - False negative
  - We don’t “find” any difference between groups, but there really is a difference between them
- Type I error rate = alpha. A test’s **confidence level** = 1 - alpha (usually 0.95)
- Type II error rate = beta. A test’s **power** = 1 - beta (usually 0.8)

Say that you are running a multiple linear regression and that you have reason to believe that several of the predictors are correlated. How will the results of the regression be affected if several are indeed correlated? How would you deal with this problem?

- **Problem 1:** coefficient estimates and signs will vary dramatically
  - Depending on which particular variables you included in the model, a variable's influence may flip signs or have a CI that includes 0 (so not significant)
- **Problem 2:** p-values are misleading
  - See above. Actually important variables can have a high p-value because it's as if the effect of the correlated features are "split" between all of them -- result is that there's uncertainty about which features are actually relevant
- **Solution:**
  - Remove predictors
  - Combine predictors (e.g., look for latent variables, use dimensionality reduction like PCA, create interaction terms)
  - Center the data
  - Get a larger sample
  - Regularization methods (e.g., lasso, ridge, elastic net)

Compare and contrast gradient boosting and random forests.

- **Both:** ensemble of decision trees
- **Difference: How ensemble is built**
  - **Gradient Boosting:** Trees are built one at a time → successive weak learners learn from the mistakes of preceding weak learners
  - **Random Forest:** Trees are built independently at the same time
- **Difference: Output**
  - **Gradient Boosting:** combines results of weak learners with each successive iteration
  - **Random Forest:** trees are combined at the end through averaging or majority voting
  - Gradient boosting is more prone to overfitting than RFs because of the lack of independence in tree building
- Gradient boosting is better for unbalanced datasets (e.g., fraud detection)
- Random Forests are better for multi-class object detection with noisy data (e.g., CV)

Describe some advantages and disadvantages of relational databases vs. NoSQL databases

- Relational Databases
  - **Advantage:** Ensure data integrity through a defined schema & ACID properties
  - **Advantage:** Good for vertical scaling
  - **Advantage:** Learning / switching between types of relational databases are easy because of an almost standard query language
  - **Disadvantage:** Data schema needs to be known in advance
  - **Disadvantage:** Data schemas can be hard to change / can cause performance issues
  - **Disadvantage:** Horizontal scaling is difficult and can lead to bottlenecks
- NoSQL Databases
  - **Advantage:** Allows for more flexibility in data format and representations through BASE properties, so it is easier to work with unstructured or semistructured data
  - **Advantage:** Useful when iterating on data schema or adding new features / functionality like in a startup
  - **Advantage:** Good for horizontal scaling
  - **Advantage:** Better for applications that need to be highly available
  - **Disadvantage:** Weaker guarantees on data correctness
  - **Disadvantage:** Managing data consistency can be difficult due to the lack of a predefined schema that's strictly adhered to
  - **Disadvantage:** Some kinds of complex queries or access patterns can be difficult

If 70% of Facebook users on iOS also use Instagram, but only 50% of Facebook users on Android also use Instagram, how would you go about identifying the underlying reasons for this discrepancy in usage?

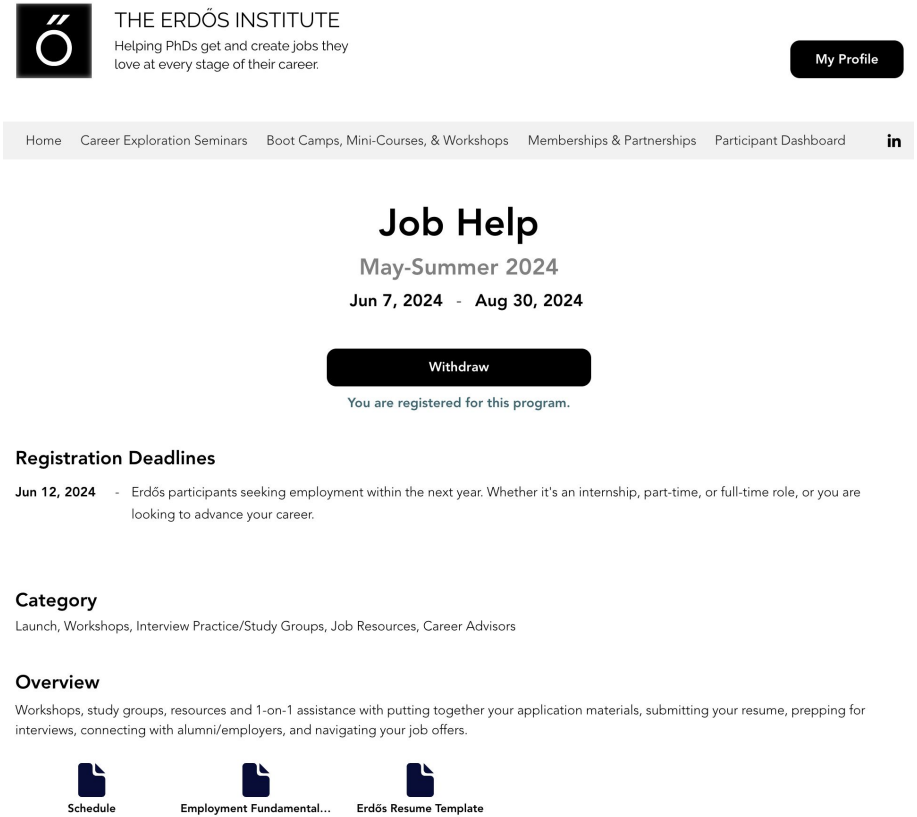
- Gather data on iOS and Android users for Facebook and Instagram
  - **Demographics:** age, gender, race, location
  - **User activity:** time spent overall, time spent on various activities (feed, in-app messaging) for both users on both apps
  - **Visualize** user activity metrics by each cut of user demographics for high-level understanding
  - iOS users may spend much more time on the FB ecosystem than Android users do, and this “top-of-funnel” reason may lead them to use Instagram more, too
  - iOS and Android users may be from different age groups, which could affect their respective levels of Instagram usage, as Instagram isn’t as widely used by older people
- Consider Instagram’s device and resource requirements compared to Facebook’s requirements
  - Maybe iOS devices have an easier time downloading the Instagram app, since the app size is smaller for iOS than Android
  - Maybe Instagram only works on devices that have updated their OS within the last two years and Apple devices tend to run the latest OS much more than Android devices
- App experiences
  - Do FB and Instagram perform the same way on both platforms? Compare app store ratings, number of bug reports, feed scroll latency, percentage of sessions with app crash for both devices
  - Maybe Facebook performs equally well on both phone platforms, but Instagram has under-invested in its Android app experience
- Talk with experts
  - User experience researchers, product strategy teams, Android / iOS leads for FB and Instagram
  - For a large difference across so many users, there’s likely a bigger structural or strategic cause for the disparity that data analysis alone might not uncover

Questions?

Homework

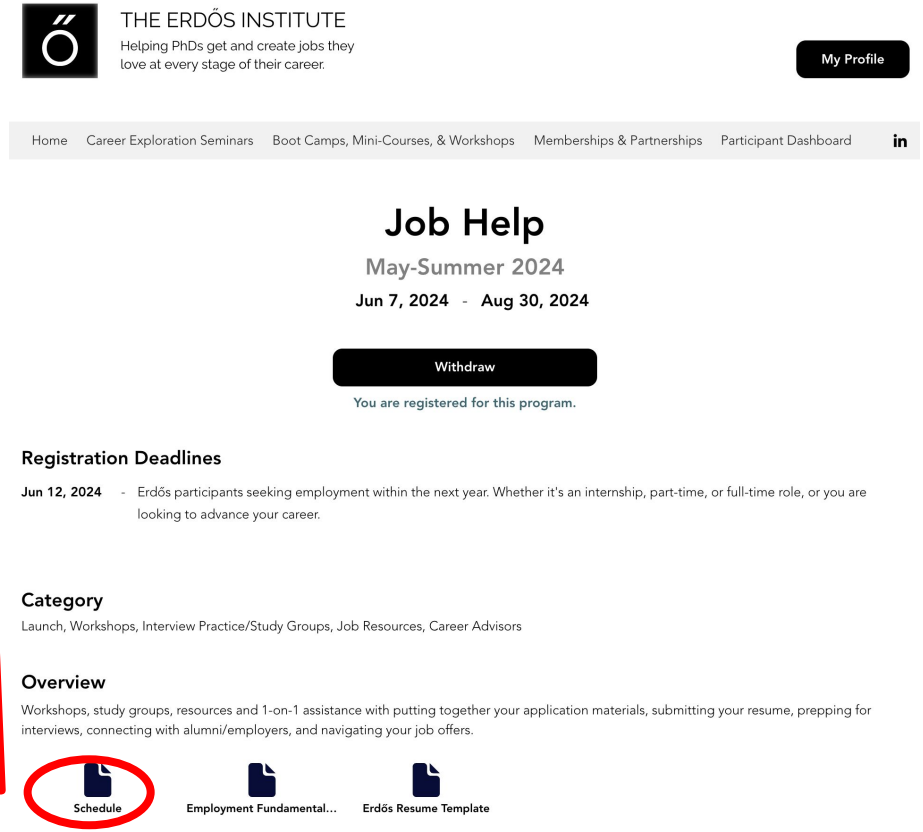
Erdős Technical Interview Prep Structure

- **Course page**



Erdős Technical Interview Prep Structure

- **Course page**



Schedule

Erdős Technical Interview

- **Lectures**
  - Overview
  - White Boarding / Paired Coding
  - Data Challenges / Case Studies /
  - FinTech

ő The Erdős Institute

Job Help

Summer 2024 Schedule

To view any of the Erdős websites, you must be logged in to your profile

June 12 12–1 PM	***Lecture 1***	**Job Help Overview** Technical & Behavioral Interview Preparation Course Zoom link in your Erdős profile
June 13-18	***Homework***	**Self study: Start Learning about Technical Interviews** *Ace the DS Interview*: Introduction through Chapter 4 *Erdős*: Technical Interview Materials
June 19 12–1 PM	***Lecture 2***	**White Boarding & Paired Coding** Zoom link in your Erdős profile
June 20-23	***Homework***	**Self study: Probability** *Ace the DS Interview*: Chapter 5 *Erdős*: Technical Interview Resources
June 24, 9-10 AM or June 25, 12-1 PM	***Small Groups***	**Practice Interview: White Boarding & Paired Coding – Probability** *With your small groups*
June 26 12–1 PM	***Lecture 3***	**Resume + Application Strategy** Zoom link in your Erdős profile
June 27-30	***Homework***	**Self study: Statistics** *Ace the DS Interview*: Chapter 6 *Erdős*: Technical Interview Resources
July 1, 9-10 AM or July 2, 12-1 PM	***Small Groups***	**Practice Interview: White Boarding & Paired Coding – Statistics** *With your small groups*

Erdős Technical Interview

- **Lectures**
  - Overview
  - White Boarding / Paired Coding
  - Data Challenges / Case Studies /
  - FinTech
- **Homework**
  - Self-study (more details in a minute)

The Erdős Institute

Job Help

Summer 2024 Schedule

To view any of the Erdős websites, you must be logged in to your profile

June 12 12–1 PM	***Lecture 1***	**Job Help Overview** Technical & Behavioral Interview Preparation Course Zoom link in your Erdős profile
June 13-18	***Homework***	**Self study: Start Learning about Technical Interviews** *Ace the DS Interview*: Introduction through Chapter 4 *Erdős*: Technical Interview Materials
June 19 12–1 PM	***Lecture 2***	**White Boarding & Paired Coding** Zoom link in your Erdős profile
June 20-23	***Homework***	**Self study: Probability** *Ace the DS Interview*: Chapter 5 *Erdős*: Technical Interview Resources
June 24, 9-10 AM or June 25, 12-1 PM	***Small Groups***	**Practice Interview: White Boarding & Paired Coding – Probability** *With your small groups*
June 26 12–1 PM	***Lecture 3***	**Resume + Application Strategy** Zoom link in your Erdős profile
June 27-30	***Homework***	**Self study: Statistics** *Ace the DS Interview*: Chapter 6 *Erdős*: Technical Interview Resources
July 1, 9-10 AM or July 2, 12-1 PM	***Small Groups***	**Practice Interview: White Boarding & Paired Coding – Statistics** *With your small groups*

Erdős Technical Interview

- **Lectures**
  - Overview
  - White Boarding / Paired Coding
  - Data Challenges / Case Studies /
  - FinTech
- **Homework**
  - Self-study
- **Small Groups**
  - Practice interviews
  - Resource sharing
  - Social support

The Erdős Institute

Job Help

Summer 2024 Schedule

To view any of the Erdős websites, you must be logged in to your profile

June 12 12–1 PM	*Lecture 1*	**Job Help Overview** Technical & Behavioral Interview Preparation Course Zoom link in your Erdős profile
June 13-18	*Homework*	**Self study: Start Learning about Technical Interviews** *Ace the DS Interview*: Introduction through Chapter 4 *Erdős*: Technical Interview Materials
June 19 12–1 PM	*Lecture 2*	**White Boarding & Paired Coding** Zoom link in your Erdős profile
June 20-23	*Homework*	**Self study: Probability** *Ace the DS Interview*: Chapter 5 *Erdős*: Technical Interview Resources
June 24, 9-10 AM or June 25, 12-1 PM	*Small Groups*	**Practice Interview: White Boarding & Paired Coding – Probability** *With your small groups*
June 26 12–1 PM	*Lecture 3*	**Resume + Application Strategy** Zoom link in your Erdős profile
June 27-30	*Homework*	**Self study: Statistics** *Ace the DS Interview*: Chapter 6 *Erdős*: Technical Interview Resources
July 1, 9-10 AM or July 2, 12-1 PM	*Small Groups*	**Practice Interview: White Boarding & Paired Coding – Statistics** *With your small groups*

Self-Study

- [Tech interview overview]
  - Summary of what we covered today
  - FAQ

Technical Interview Materials

Overview | Data Science "Types" | Subject Matter Topics | Interview Formats | FAQ

Lindsay Warrenburg
Head of Technical Interviews

Office Hours: Thursdays, 12-1pm
Email: lindsay@erdosinstitute.org
Preferred Contact: Slack

Participants should feel free to Slack me with any questions or comments!

Overview

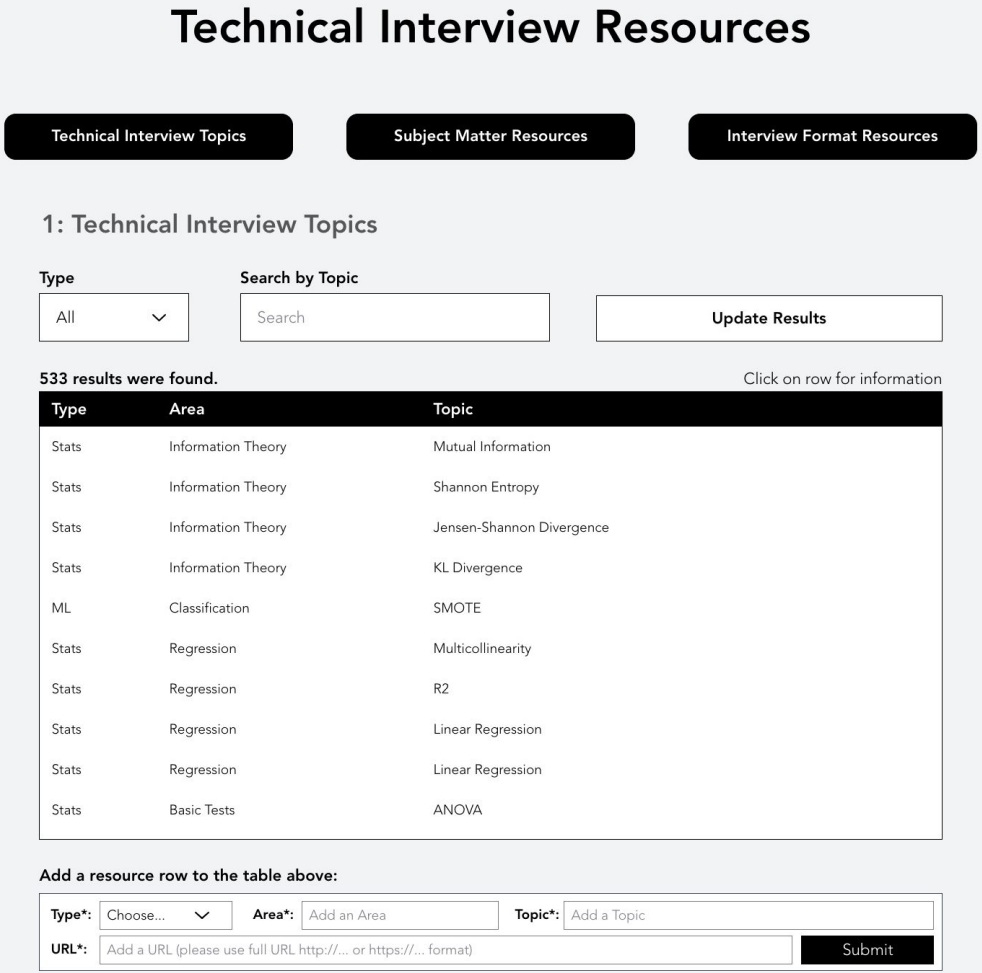
- Before applying to a position, you should have a general idea of which type of data science positions you want to go into
- Since "data science" is a buzzword, jobs that are very different in nature are all labeled with that one term (data science). It's sort of like how a clinical psychologist and a cognitive neuroscientist are both psychologists, but the clinical scientist wouldn't need to understand the intricacies of performing EEG or fMRI and a cognitive neuroscientist wouldn't need to understand how to diagnose mental illness
- Much like applying for grad school where you're looking for the right fit with the PI and a niche research topic/specialization, data science jobs require just as good of a fit (I, for example, am a data scientist, but would not be hired for a company that needs a 'software engineer' flavor of data scientist)
- Job descriptions might give you some indication of the 'flavor' of data scientist, but sometimes you'll learn what the company needs as you're going through the interview process
- The GOOD thing is that you do not need to know every topic from every data science flavor to do well in your interviews. Those of you on the job market can meet with me individually and I'll help you tailor your technical interview prep for your interests!

Some "data science types" (there are more!):

- Data Engineering (e.g., database management, data warehousing)
- Software Engineering (e.g., software development, cross-platform software)
- Computer Science (e.g., encryption, operating system containers)
- Data Product (e.g., ML/AI for a specific product, such as a smartphone app, app development and productionalizing)
- Machine Learning (ML) Research (e.g., designing new ML/AI algorithms)
- Machine Learning (ML) Engineering (e.g., executing existing ML/AI algorithms)
- AI Ethics (e.g., evaluate bias in datasets and algorithms, create transparency)

Self-Study

- **Tech interview overview**
- **Tech interview resources**
  - Hundreds of websites, books, and example problems



Self-Study

- Tech interview overview
- Tech interview resources
- [Course book]
- [SQL companion website]

DataLemur

Sign

Ace The SQL Interview

Practice SQL Interview questions asked by top tech companies **FOR FREE** on DataLemur. Made by Nick Singh, Best-Selling Author of Ace the Data Science Interview

COMPANY	TITLE	DIFFICULTY
LinkedIn	Data Science Skills	Easy
Facebook	Page With No Likes	Easy

Small Groups / Practice Interviews

Best for those actively preparing for interviews in the next 6 months

- Option to **sign up** to meet with small groups **(due this Friday, June 21)**

Small Groups / Practice Interviews

Best for those actively preparing for interviews in the next 6 months

- Option to **sign up** to meet with small groups

- If you sign up
  - Private Slack channel
  - Private GitHub repo (only be visible to those who are actively preparing for interviews)
  - 9 sets of practice interviews

Small Groups / Practice Interviews

- Every week, there is an assigned study topic (e.g., statistics)

Small Groups / Practice Interviews

- Every week, there is an assigned study topic (e.g., statistics)
- During the first part of the week, you will be expected to read the corresponding chapter of "Ace the Data Science Interview" and the relevant sections in our **Technical Interview Resources page**

Homework **Self study: Statistics**
Ace the DS Interview: Chapter 6
Erdős: Technical Interview Resources

Small Groups / Practice Interviews

- Every week, there is an assigned study topic (e.g., statistics)
- During the first part of the week, you will be expected to read the corresponding chapter of "Ace the Data Science Interview" and the relevant sections in our **Technical Interview Resources page**
- At the end of the week, you will meet with your practice group on Zoom. During this session, you will go to our GitHub page and open that week's problem set.

Small Groups **Practice Interview: White Boarding & Paired Coding – Statistics**
Sign up on the Erdős calendar page

Small Groups / Practice Interviews

Why is this important?

- Each group member will take turns answering the questions as if they were in an interview.

Small Groups / Practice Interviews

Why is this important?

- Each group member will take turns answering the questions as if they were in an interview.

- This will test not only your knowledge of the topic, but how well you perform under pressure, simulating the interview experience.

Small Groups / Practice Interviews

Why is this important?

- Each group member will take turns answering the questions as if they were in an interview.

- This will test not only your knowledge of the topic, but how well you perform under pressure, simulating the interview experience.

- The best way to succeed is to practice -- out loud and with others (not just silent reading by yourself!)

Homework for this month

Lecture 2	**White Boarding & Paired Coding** Zoom link in your Erdős profile
Homework	**Self study: Probability** *Ace the DS Interview*: Chapter 5 *Erdős*: Technical Interview Resources
Small Groups	**Practice Interview: White Boarding & Paired Coding – Probability** *With your small groups*
Lecture 3	**Resume + Application Strategy** Zoom link in your Erdős profile
Homework	**Self study: Statistics** *Ace the DS Interview*: Chapter 6 *Erdős*: Technical Interview Resources
Small Groups	**Practice Interview: White Boarding & Paired Coding – Statistics** *With your small groups*
Lecture 4	**Case Studies, Data Challenges, & FinTech** Zoom link in your Erdős profile
Homework	**Self study: Machine Learning** *Ace the DS Interview*: Chapter 7 *Erdős*: Technical Interview Resources

Coming up next

- Case studies & data challenges overview (with me)
- Dedicated group sessions–more white boarding / paired coding, but also case studies and data challenges

Contact me on our Slack channel :)