White Boarding & Paired Coding

Lindsay Warrenburg
The Erdős Institute

Schedule

1. Review from first session
2. Questions about homework
3. Overview of white boarding & paired coding
4. *Tech Interview Room* in Gather
5. Breakout rooms
6. Questions
7. 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

Homework Review

Session 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

- Go through tech interview folder and start prepping (yourself and with others)

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

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?

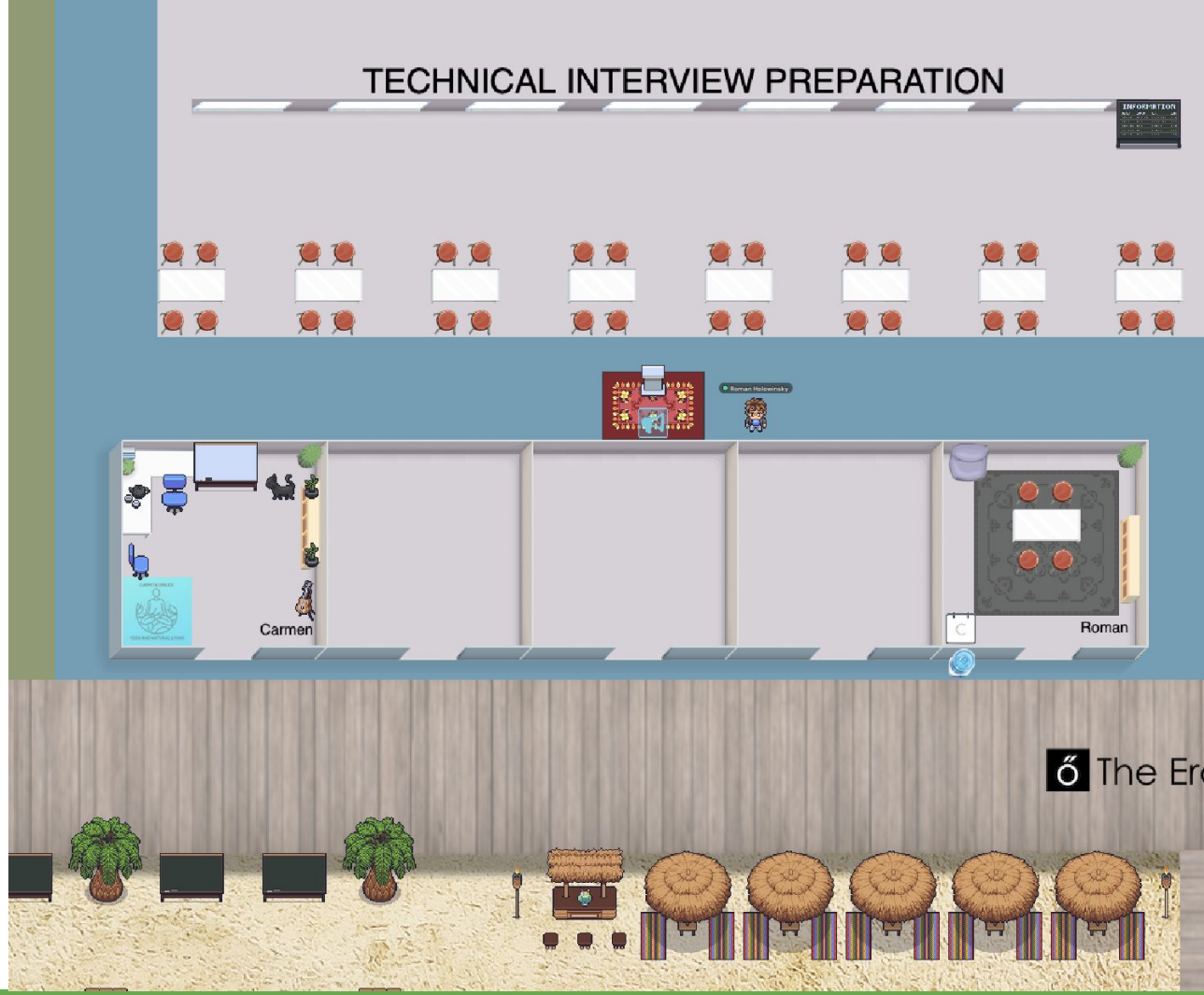
```
def bubbleSort(arr):
    for ii in range(len(arr)-1):
        for jj in range(len(arr)-1):
            if arr[jj] > arr[jj+1]:
                arr[jj], arr[jj+1] = arr[jj+1], arr[jj]
    return(arr)
```

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

Technical Interview Room in GatherTown

Preparation Tips

- [Gather space]



Breakout Rooms: Paired Coding

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

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

Homework for this month

- Read sections of ***Ace the Data Science Interview***
- **Sign up for practice groups** for weekly problem sets
  Thursday 5-6 pm EST
  Friday 9-10 pm EST