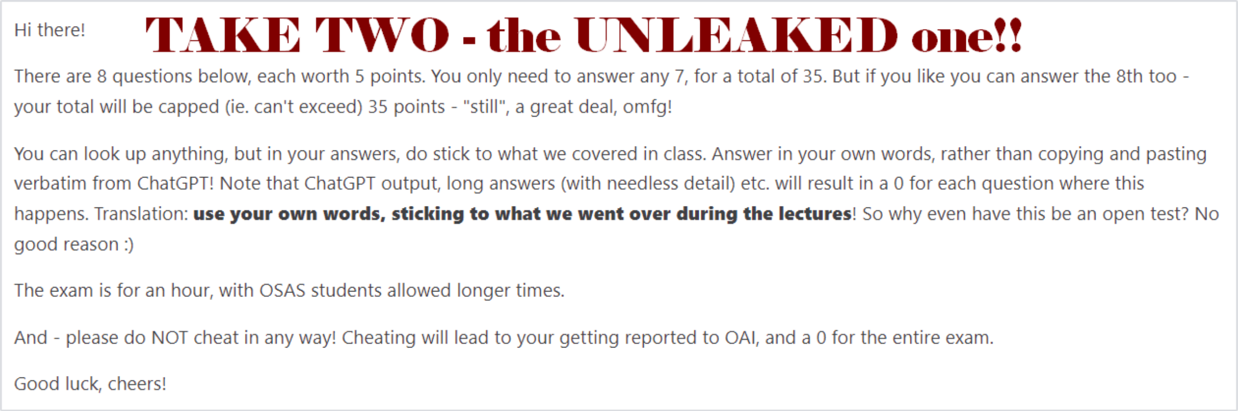
v2 CS585 Final Rubrics

Total Marks	35 points
Late Penalty	< 30 mins : -2.5 points >= 30 mins: -5 points

General guidelines (Prof Saty's instructions note from the exam):



Hi there! **TAKE TWO - the UNLEAKED one!!**

There are 8 questions below, each worth 5 points. You only need to answer any 7, for a total of 35. But if you like you can answer the 8th too - your total will be capped (ie. can't exceed) 35 points - "still", a great deal, omfg!

You can look up anything, but in your answers, do stick to what we covered in class. Answer in your own words, rather than copying and pasting verbatim from ChatGPT! Note that ChatGPT output, long answers (with needless detail) etc. will result in a 0 for each question where this happens. Translation: **use your own words, sticking to what we went over during the lectures**! So why even have this be an open test? No good reason :)

The exam is for an hour, with OSAS students allowed longer times.

And - please do NOT cheat in any way! Cheating will lead to your getting reported to OAI, and a 0 for the entire exam.

Good luck, cheers!

- If you find that student has **"literally"** copy-pasted the solution from ChatGPT / Bard or any other LLM, please give **-50% points** for that question
- If you find that a student has **"significantly"** answered **"more than what is asked"**, please provide **partial grading** written in each question's rubric.

Question 1:
A relational table containing data represented as a set of rows, and a vector DB containing data encoded as a collection of multi-dimensional points, are more similar to each other than you might realize! In what sense?

In other words, given data for five numerical features/columns/attributes/axes/.... which we could refer to as A B C D E:

* what do we usually do in order to search, if we had a table with data in 10M rows for these five columns?

* how would we search, if we instead have a 5D space (coord system) with the data in the form of 10M (5D) points?

* what is the main difference between these (above) two forms of search?

Solution

Part A: (1 mark)

a. A relational table contains one record in each row, and each column represents a special feature or attribute. In the case of a vector DB, each record is stored in the form of a point in an n-dimensional space (5 dimensions in the case of A, B, C, D, E). Each dimension represents a special feature of the record.

Part B: (1 mark)

b. In order to search in a relational database, we generally use a query language like SQL. We need to use either exact keys or a range of keys (in the case of numbers) to narrow down our search results.

Part C: (1 mark)

c. In the case of a vector DB, we generally use some sort of similarity search. The distance calculations can be done using an Euclidean metric.

Part D: (2 marks)

d. Vector db mainly relies on proximity/distance-based searching, wherein we search for similar records by calculating the distance between neighboring records in an n-dimensional space. **Additionally, approx NN search might result in faster but rather inaccurate results, which is acceptable for such use cases.**
Relational DB, on the other hand, involves logical conditions and operations. We need to use either exact keys or a range of keys (in the case of numbers) to narrow down our search results.

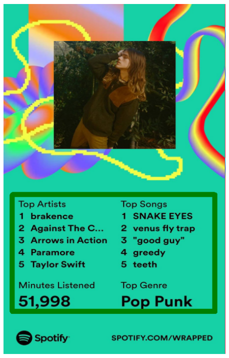
Consider exploring alternative methods for conducting searches in a vector database and a relational database.

Deductions

# points deducted	Cause
-0.25 (Part A)	If how data is represented in a vector DB is not mentioned
-0.25 (Part A)	If how data is represented in a relational DB is not mentioned
-1 (Part B)	If SQL/other logical way of RDB searching is not mentioned
-1 (Part C)	If similarity search/other logical way of vector DB searching is not mentioned
-1 (Part D)	If how data is searched in a vector DB is not mentioned
-1 (Part D)	If how data is searched in a relational DB is not mentioned

Question 2:

How would you represent using valid JSON, the ‘Spotify Wrapped’ card data (enclosed in a green rect below) ?



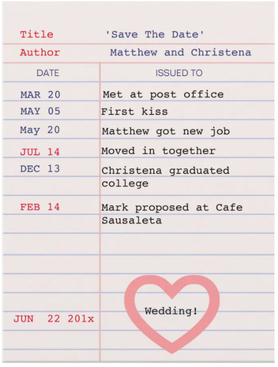
Top Artists
1 brakence
2 Against The C...
3 Arrows in Action
4 Paramore
5 Taylor Swift

Top Songs
1 SNAKE EYES
2 venus fly trap
3 "good guy"
4 greedy
5 teeth

Minutes Listened
51,998

Top Genre
Pop Punk

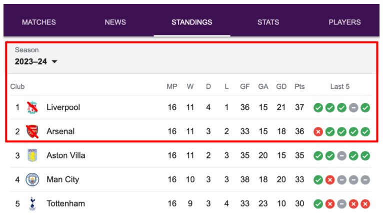
Below is a cute wedding invitation, designed to look like an old-fashioned library book borrowing card :) How would you represent all the data in it, using VALID JSON? You can ignore the heart frame in the 'Wedding!' entry.



Title 'Save The Date'
Author Matthew and Christena

DATE	ISSUED TO
MAR 20	Met at post office
MAY 05	First kiss
May 20	Matthew got new job
JUL 14	Moved in together
DEC 13	Christena graduated college
FEB 14	Mark proposed at Cafe Sausaleta
JUN 22 201x	Wedding!

The 2023-2024 Premier League season is almost halfway done and it’s turning out to be a doozy! How would you represent this standings table in JSON. In particular we’d like to store the store the following information for each team: Matches Played (under the MP column), Number of Wins (W), Number of Draws (D), Number of Losses (L), Points (Pts) and the team’s record for the Last 5 games (the table below depicts 'Win' as a green check, 'Draw' as a gray hyphen, and 'Loss' as a red cross). How would you represent, using VALID JSON, the data that is enclosed in the red rectangle? Use 'Win', 'Loss', 'Draw' in place of their respective icons.



Season
2023–24

Club		MP	W	D	L	GF	GA	GD	Pts	Last 5
1	Liverpool	16	11	4	1	36	15	21	37	✓ ✓ ✓ – ✓
2	Arsenal	16	11	3	2	33	15	18	36	✗ ✓ ✓ ✓ ✓
3	Aston Villa	16	11	2	3	35	20	15	35	✓ ✓ – ✓ ✓
4	Man City	16	10	3	3	38	18	20	33	✓ ✗ – – –
5	Tottenham	16	9	3	4	33	23	10	30	✓ ✗ – ✗ ✗

Solution

Spotify Wrapped	Wedding Invite	Premier League

{
"spotifyWrapped":
 	
{
"minutesListened":
51998,
 	
"topGenre":"Pop Punk"
  "topArtists":
["brakence",
"Against the C",
"Arrows in Action",
"Paramore",
"Taylor Swift"],
 	
"topSongs": [
"SNAKE EYES", "venus
fly trap",
"\"good guy\"",
	{
	"title": "’Save the
	Date’",
	"author": "Matthew and
	Christena",
	"dates":
	 [
	  {
	   "date": "Mar 20",
	   "issuedTo": "Met at
	post office"
	  },
	  {
	   "date": "May 05",
	   "issuedTo": " First
	kiss"
	  },
	{
	   "date": "May 20",
		{
		"Season": “2023-24”
		  "standings":
		[
		 {
		     "position": 1,
		"team": "Liverpool",
		"matchesPlayed": 16,
		"wins": 11,
		"draws": 4,
		"losses": 1,
		"points": 37
		"lastFiveRecord":
		["Win","Win", "Win",
		"Draw","Win"]
		 },
		 {
		     "position": 2,
		"team": "Arsenal",

"greedy", "teeth"
]
}
}
	"issuedTo": "Matthew got new job"
	},
	{
	"date": "Jul 14",
	"issuedTo": "Moved in together"
	},
	{
	"date": "Dec 13",
	"issuedTo": "Christena graduated college"
	},
	{
	"date": "Feb 14",
	"issuedTo": "Mark proposed at Cafe Sausaleta"
	}
	]
	}
		"matchesPlayed": 16,
		"wins": 11,
		"draws": 3,
		"losses": 2,
		"points": 37
		"lastFiveRecord": ["Loss","Win", "Win", "Win","Win"]
		}
		]
		}

Deductions

For all questions:
Please deduct 0.2 for every error in syntax in the JSON. For instance arrays must be represented with square brackets. Keys must be strings with double-quotes. All values in an array must be comma-separated. All key-value pairs must be comma-separated. **All needed to enclosed in an outer pair of { }.**

A maximum of 1 point can be deducted for syntactical issues in each of the three parts (i.e. a maximum of 3 totally).

Note that the values need not be exactly what is provided in the images. The idea of this question is to test whether the student can capture the idea of the data represented in the image in a JSON. That being said, blatant disregard of what is in the image to the point where you are not able to determine how the key value pairs provided can be mapped to the data in the image can lead to a penalty of 0.5 per question and this case is left to your discretion. Award similar marks if a standard json format has been used with proper

explanation which demonstrates the students understanding of converting visual data into JSON.

Spotify wrapped (Max awarded: 1.5):
Award 0 if not attempted
+1 if all the following keys (or keys that capture the same idea) are presented: minutesListened, topGenre, topArtists, topSongs (0.25 deduction for every key that is missing).
+0.1 If the value for minutesListened is a number (it is okay if the number is presented as a string with double quotes)
+0.1 If the value for topGenre is a string
+0.15 If the value for topArtists is a list of strings
+0.15 If the value for topSongs is a list of strings

Wedding invitation (Max awarded: 1.5):
Award 0 if not attempted
+0.15 if a key for title is present
+0.1 if a value for title is present (eg: “Save the date”)

+0.15 if a key for author is present
+0.1 if a value for author is present (eg: “`Matthew and Christena`”)

+0.5 for the inclusion of a parent key “dates” or something that represents this, and its value must be a list of comma-separated json objects which have (date,issuedTo) pairs. 0.5 is awarded if the student recognized that the (date,issuedTo) pairs are objects that must be in a list.

+0.5 for the inclusion of (date,issuedTo) pairs eg: `{`

```
   "date": "May 05",
   "issuedTo": " First kiss"
  },
```

```
Or: { "MAR 20": "Met at post office", "MAY 05": "First kiss",...}
```

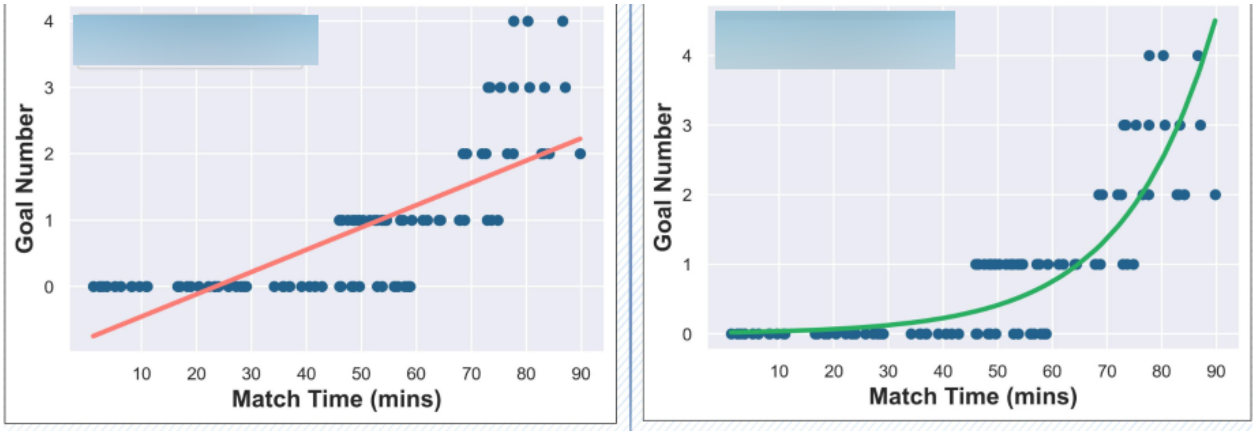
Soccer table (Max awarded: 2):
Award 0 if not attempted
+0.75 for each of the following key,value pairs for Liverpool: Matches played, Wins, Draws, Losses, Points (Each for 0.15)
+0.75 for each of the following key,value pairs for Arsenal: Matches played, Wins, Draws, Losses, Points (Each for 0.15) (**It may be okay if this is represented as a comment showing that the format is same as that of Liverpool**)
+0.2 for representing Liverpool’s last5Record as a list of strings. The strings must be one of the three “Win”, “Draw”, “Loss”

+0.2 for representing Arsenal's last5Record as a list of strings. The strings must be one of the three "Win", "Draw", "Loss"
+0.1 for presenting the season key,value pair ("season:"2023-24")

Deduct 0.5 if the table data for a single team is not represented as an object within {}
Deduct 0.5 if it is not clear which standings object corresponds to which team. In the example I provided, I have used "team" as a key in the object. There are other ways of representing this but as long as you are able to determine the matching between the team and the standings data, that is sufficient.

Question 3:
In ML, why is 'underfitting' (of a model to given data) bad? Conversely, why is 'overfitting' also bad?

What are two types of fits called, in the diagram below? The diagram shows the number of goals scored after a game starts (across many games).



Why is the right one much preferable, of the two? Be specific!!

Solution

Part 1 (2 points)

- **Underfitting** is bad because the model has **high bias**. This results in not learning the training data well and poor performance on the test data. An underfit model fails to significantly grasp the relationship between the input values and target variables resulting in poor accuracy.
- **Overfitting** is bad because the model has **high variance**. It learns the training data too well. Overfitting reduces its generalizability outside the training dataset. This results in poor accuracy.

Part 2 (2 points)

- The two types of fits are called **linear fit (underfit)** and **Non-Linear fit (eg. Polynomial, Poisson)** respectively.

Part 3 (1 point)

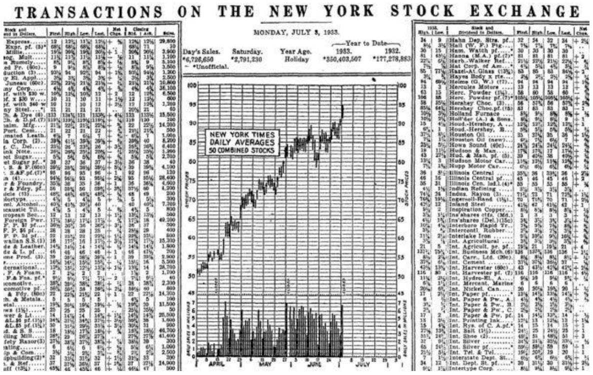
- The right one has captured the true pattern of the data which means it has learned the trends of the data. In the left diagram, the model is underfit as it has nor learned enough from the training data. The right diagram would give better accuracy on unseen data than the left one.
- **The left one will produce a NEGATIVE GOAL COUNT if the time is < ~25 min!**

Deductions

# points deducted	Cause
-0.5 to -1	If the explanation of (underfitting/overfitting) is incomplete or lacks clarity, such as not mentioning how it leads to an (oversimplified/too complex) model that (fails to capture important trends in the data/captures noise in data). (Please deduct marks for the severity of deviation from this standard expected definition).
-1	If the types of fits are incorrectly identified or if only one fit is correctly identified. (Please ensure that the right model is not overfit to the data, if written Overfit deduct marks for that, Also be aware that there can be some other term used to convey type of fit, basically left is linear and right is Non-Linear, So be cautious and award marks accordingly).
–0.5	If the explanation for preferring the right fit is vague or incorrect. (Left is Underfit or Right will provide better performance on Unseen Data since it captures true relation between feature and dependent variable, which is Non-Linear.)

Question 4:

Visualizing data is often the best to understand it. Compared to a newspaper chart (depicting NYSE data) from the 1930s shown below, today we have a LOT more at our disposal, for data viz.



TRANSACTIONS ON THE NEW YORK STOCK EXCHANGE

MONDAY, JULY 3, 1933.

Day's Sales.	Saturday.	Year Ago.	Year to Date 1933.	1932.
*6,726,650	*2,791,230	Holiday	*350,403,507	*177,278,883

*Unofficial.

NEW YORK TIMES
DAILY AVERAGES
50 COMBINED STOCKS
100
95
90
85
80
75
70
65
60
55
50
45
STOCK PRICES
APRIL
MAY
JUNE
JULY

Still sticking to a flat model of displaying, what TWO things we can do today to depict data, that was not possible in newsprint?

What THREE ways ('technologies') can we use, to transcend the flat display of data?

Solution

Two ways to depict data that was not possible in newsprint: (Any 2 of 3)

1. Interactive visualization - With today's technology, we can perform drill down and roll up operations and filter information as per our need. Unlike traditional newsprint, users can obtain a personalized view of the data they desire.
2. Animated visualization - Data that is animated can provide information across more dimensions by depicting temporal trends and changes. Additionally, it helps provide more logical insights to a viewer by providing a seamless comparison of data points.
3. Real-time visualization - Newsprints could merely provide a snapshot of the current trends. Real-time visualization provides constant updates that dynamically change as new data becomes available. This is especially important in use cases that involve tracking data that has immediate implications for users.

Sample answer for ways to transcend the flat display of data:

1. 3D printing
2. Haptics
3. Augmented reality
4. Virtual reality
5. Holography
6. **Projection mapping**
7. **Dome display**
8. Intelligent comprehension/question answering capabilities
9. Any other relevant answer that goes beyond the capabilities of flat display

Deductions

# points deducted	Cause
-1	For each missing answer on depicting data not possible using newsprint [Max: -2]
-0.25	For each missing explanation (short is fine) for depicting data not possible using newsprint [Max: -0.5]
-1	For each missing answer or an answer that does not improve on flat display

Question 5:
Spatial data is commonly indexed using R-Trees, where overlapping MBRs in a hierarchy are used to bound features (ie the spatial data we want to process, eg. restaurants, streets, buildings etc).

What is the main limitation of an MBR? What alternative would be far better?

A typical city government contains multiple branches/departments/units that handle specific aspects of governing.

How would the following three units utilize spatial data? In other words, what is an example of (spatial) data they collect and analyze?

* energy

* water

* public safety

Solution

Limitation of MBR: Too much waste while indexing. We want the **negative area between the bounding box and the actual features to be as small as possible**. For example, for a 45-degree line, we would waste so much space, and it could result in **false positive results** while searching indexes.

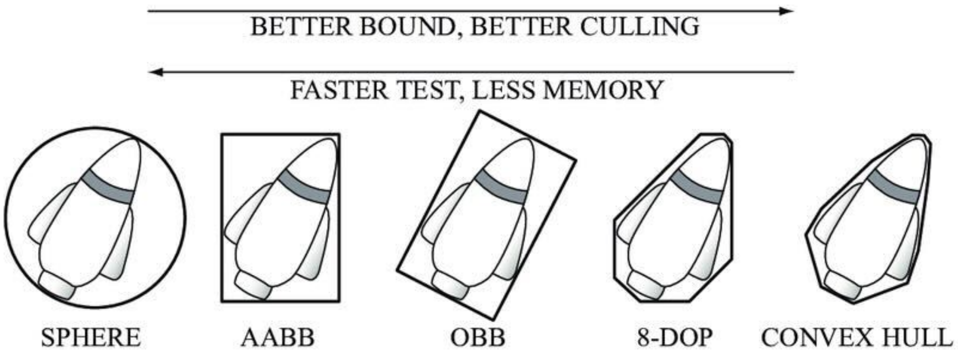
Better alternative: OBB (Oriented Bounding Box)

Use of spatial data in each unit:

- **Energy:** This department would need to store data on the location of power plants, transmission lines, and energy storage facilities. These resources would be helpful during repairs and maintenance and also for future planning.
- **Water:** Store data about the city's water infrastructure, such as pipelines. Also, they need to track data related to flood risk locations.
- **Public Safety:** They need to keep data about the placement of video surveillance cameras, emergency rooms, hospitals, and many other things. All this data would help public safety to keep people safe and respond to emergencies faster.

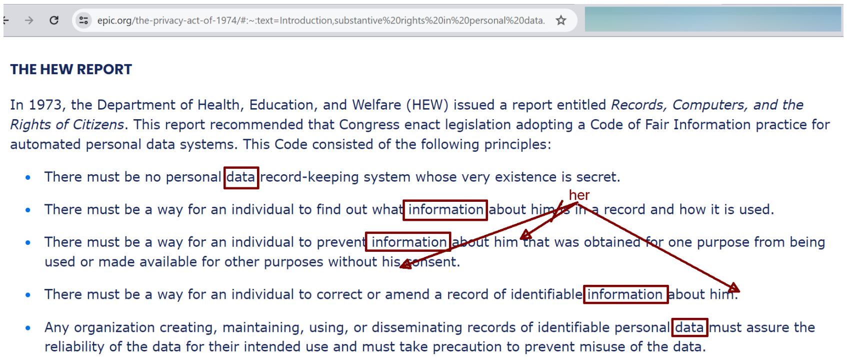
Deductions

# points	Cause

deducted	
-1 point	Not mentioning the limitation of MBR or mentioning incorrect points (Anything near the bold points in the sample answer is correct)
-1 point	Not mentioning OBB or convex hull or minimum bounding polygons as an alternative. The alternatives cannot be quad-trees, r-trees, r*-trees or minimum bounding circles. They still have the same problems as MBRs.  BETTER BOUND, BETTER CULLING FASTER TEST, LESS MEMORY SPHERE AABB OBB 8-DOP CONVEX HULL
-1 point	Not mentioning any sample use of spatial data by the energy department
-1 point	Not mentioning any sample use of spatial data by the water department
-1 point	Not mentioning any sample use of spatial data by public safety

Question 6:

Using data 'properly' (eg keeping it secure, keeping it private) has been a concern going back to the earliest days of electronic data storage, eg. as specified in the 'HEW' report:



epic.org/the-privacy-act-of-1974/#:~:text=Introduction,substantive%20rights%20in%20personal%20data.
her

THE HEW REPORT

In 1973, the Department of Health, Education, and Welfare (HEW) issued a report entitled *Records, Computers, and the Rights of Citizens*. This report recommended that Congress enact legislation adopting a Code of Fair Information practice for automated personal data systems. This Code consisted of the following principles:

- There must be no personal data record-keeping system whose very existence is secret.
- There must be a way for an individual to find out what information about him is in a record and how it is used.
- There must be a way for an individual to prevent information about him that was obtained for one purpose from being used or made available for other purposes without his consent.
- There must be a way for an individual to correct or amend a record of identifiable information about him.
- Any organization creating, maintaining, using, or disseminating records of identifiable personal data must assure the reliability of the data for their intended use and must take precaution to prevent misuse of the data.

"OMG, NONE of these problems are STILL relevant, in 2023!" - said no one ever :)

LLMs (including multimodal extensions, ie. ones trained using a wide variety of audio, images, video... data in addition to text) are about to make this problem significantly worse - there are no guaranteed mitigating technical strategies, only the threat of punishment, including fines (eg. a brand new AI ACT: ihy-ai).

Mention, and briefly discuss in your OWN words, FIVE such new 'ills' [it would be most ironic to ask ChatGPT to!]. Hint: scroll down, down, down: https://bytes.usc.edu/cs585/f23-Da-taaa/extras/index.html

Solution

According to Hint PDF:

1. Automated Essay Writing and Academic Dishonesty (Dishonesty)
2. Generating Fake Research Papers (Dishonesty)
3. Impersonating Celebrities or Public Figures (Propaganda)
4. Automated Propaganda Generation (Propaganda)
5. Creating Fake Historical Documents or Texts (Propaganda)
6. Generating Fake Product Reviews (Deception)
7. Generating Realistic but Fake Personal Stories or Testimonies (Deception)
8. Crafting Convincing Scam Emails (Deception)
9. Crafting Legal Documents with Hidden Clauses (Deception)
10. Automated Social Media Manipulation (Info. Manipulation)
11. Generating Fake Medical Advice or Information (Info. Manipulation)
12. Crafting Deceptive Advertisements (Info. Manipulation)
13. Creating Fake Financial Reports or Data (Financial Harm)
14. Generating Scripts for Scam Calls (Financial Harm)
15. Fake Personal Profiles and Identities (Personal and Identity Harm)
16. Automated Online Harassment (Personal and Identity Harm)
17. Generating Fake Evidence or Alibis (Personal and Identity Harm)
18. Fake Technical Support Scams (Tecno-social Harm)

19. Generating Biased or Prejudiced Content (Tecno-social Harm)

All the mentioned above are acceptable. Also, if the student only mentions something like Dishonesty or Propaganda and explains by using their own words about them, it’s also acceptable.
It’s also acceptable if the student gives their own reasonable answer that is not based on the PDF.

Deductions

# points deducted	Cause
-1 pt	need to provide at least five points and 1 pt will be deducted for each missing point
-0.5 pt	only gives keyword and does not discuss it
-0.5 pt	the keyword is given, but the explanation does not match the keyword

Question 7:
Traditional BI [periodic ETL, star/snowflake schema, batch processing...] is highly inadequate, in today's world - for MULTIPLE reasons.

Mention+discuss 5 aspects of the above (ie need to do things differently): consider data types, collection, storage, processing, results consumption.

Solution

1. Data Types:
    a. Traditional BI cannot handle unstructured data, which includes media (images, videos, docs)
    b. It cannot handle geospatial data, which is crucial nowadays
2. Collection:
    a. Traditional BI, especially ETL, is batch processing, and it cannot handle real-time data input, which is crucial for making decisions.
    b. It cannot be connected to IOT devices or sensors, as they send the information in a continuous stream
3. Storage:
    a. With Big Data, it is difficult to scale the data storage in star/snowflake schema as they have strict and rigid rules.
    b. It struggles to manage data with cloud service providers as the data is stored in multiple locations.
4. Processing:
    a. Traditional BI produces results after taking time, and hence, they are not real-time results.
    b. Machine learning tools cannot be integrated with the traditional BI
    c. We can use cloud, including GPU cloud, for faster (upto realtime) processing
5. Results Consumption:
    a. There are a lot of features missing in traditional BI, such as dynamic visualizations of dashboards, alerts & notifications when a particular condition is met; **mobile dashboards (on phones/tablets) coupled with cloud storage+processing make it possible to do BI anywhere, anytime, including continuously**
    b. Traditional BI was not mobile-friendly, and consumers need their insights one tap away on their mobile devices.

Deductions

# points deducted	Cause

-1 pt	Not mentioning each point under any category

Question 8:
A traditional aspect of performance tuning is to use collected statistics on how queries utilize tables, or use handcrafted rules from DBA experts.

Discuss the following, using your own words/examples:

* how can supervised ML help with performance tuning of a DB?

* what about genAI - how can it help?

Solution

A.

Supervised machine learning can assist in database performance tuning by automating workload analysis, predicting performance impact, optimizing query plans, and allocating resources. This reduces manual effort, improves performance, and allows for proactive and adaptive tuning.

Examples of Supervised ML in Database Tuning:

Better Cardinality Estimation: ML improves accuracy in estimating data distribution, helping the optimizer choose optimal execution plans.

Smart Index Recommendations: ML analyzes query patterns to suggest efficient table indexes, reducing the need for costly full scans and boosting query performance.

Quick Query Plan Optimization: ML learns from past executions to predict and select the best query plans, resulting in faster query execution.

B.

While supervised ML automates tasks and improves performance prediction, Generative AI (GenAI) offers a unique perspective for database performance tuning. Going beyond automation, GenAI can Generate realistic synthetic data for thorough testing.
Interpret user queries in natural language for intuitive database interactions.
Generate explainable recommendations
Automate documentation

Examples of GenAI Applications in Performance Tuning:

Automating query plan tuning: Translate natural language queries into optimized query plans, improving query performance and user experience.

Generating explainable performance reports: Provide insights into performance bottlenecks and recommendations for improvement, empowering DBAs to make informed decisions.

Deductions

# points deducted	Cause
-1	Do not explain how the workflow mentioned is aligned with supervised learning
-1	Discuss nothing regarding performance tuning
-1	Answer does have a few keywords indicating correct direction but lacks proper explanation/details. Too vague.