DATA 13600 Spring 2025
Final Exam

University of Chicago - Sec 01

May 28, 2025

Fill in the circles. Fill in the circles if you don't want your exam mis-graded. Some questions have one best answer, some more than one.

- You have 50 mins to take the exam.
- No additional materials, electronic or otherwise, may be used in this test. Your work is your own, and any attempt to copy or share exam materials is not allowed.
- Instructional staff will not answer any questions during the exam. If you find yourself stuck write down a reasonable assumption to move on. If you don't like an answer, you can argue for your answer after the exam is marked. (Note, changes in the accepted answers will take effect for everyone.)
- Write your initials on the bottom of each page of your exam. (The reason for this is left as an exercise for the student.)

Name: _____

Student ID number: _____

What problem do you want to apply data science to solve?

_____

1. Which one of the following best describes the primary difference between OLAP and OLTP systems?

   ◯ OLAP systems are optimized for reading and aggregating large datasets; OLTP systems are optimized for frequent, small updates.

   ◯ OLAP systems are used by customers for daily operations; OLTP systems are used only for long-term archiving.

   ◯ OLTP systems typically support distributed computation; OLAP systems often run on a single machine.

   ◯ OLAP systems run on relational databases; OLTP systems run on NoSQL systems.

   ◯ OLTP systems handle structured data; OLAP systems handle unstructured data.

2. What type of database operations are most common in OLTP (Online Transaction Processing) systems?

   ◯ Asynchronous indexing of data for improved performance

   ◯ Read-intensive operations like aggregations and filtering `SELECT`

   ◯ Large data volume transactions covering defined intervals of time or partitions of the database

   ◯ Changes to the structure of the database with `ALTER TABLE` commands

   ◯ Frequent, simple transactions such as `INSERT` and `UPDATE`

3. What kind of database manipulations do Django's `makemigrations` and `migrate` commands perform?

   ◯ Indexing of existing data for improved lookup and modification

   ◯ Verifying the integrity of all the tables with `SELECT`

   ◯ Archiving historical versions of the database to comply with records retention requirements

   ◯ Changes to the structure of the database with `ALTER TABLE` commands

   ◯ Streamlined authentication transactions such as
   `SELECT pw_hash FROM auth_users WHERE username = $USERNAME;`

 Initials: _____

4. Edit distance $edit(s_1, s_2)$ is defined as the minimum number of single-character operations (defined to include insertion, deletion, and substitution (replacement) of single characters) needed to transform between two strings s_1 and s_2. Check all that apply.

   ◯ $edit(s_1, s_3) \leq edit(s_1, s_2) + edit(s_2, s_3)$

   ◯ $edit(s_1, s_2) = edit(s_1, s_3)$ implies that $edit(s_2, s_3) = 0$

   ◯ $edit(s_1, s_2) = edit(s_2, s_1)$

   ◯ $edit(s_1, s_2) \leq max(len(s_1), len(s_2))$

   ◯ if $edit(s_1, s_2) = 0$ implies that $s_1 = s_2$

5. In 2015, federal law enforcement disclosed that they had seized a few months of email correspondence between Carl Force, a law enforcement officer, and Dread Pirate Roberts, the administrator of the Silk Road, an online drug marketplace. All of these communications were PKE encrypted with either DPR's public key or Force's public key. Law enforcement was able to find the dates, subject, sender and recipients, and approximate length of all their communications, but could not decrypt the contents of Force's emails to DPR. Briefly explain what additional information law enforcement would need to decrypt Force's messages to DPR (that they did not have at the time of Force's indictment)?

6. What does a hypervisor do?

   ◯ A hypervisor translates object-oriented method calls to SQL queries for the DBMS.

   ◯ A hypervisor provides an additional layer of encryption for a virtual machine (an operating system running in software)

   ◯ A hypervisor runs virtual machines on a physical host.

   ◯ A hypervisor is a process running on a database server that assures data integrity and record retention compliance.

   ◯ A hypervisor monitors the health of a database management system (DBMS) and provisions resources under certain (stress) conditions.

 Initials: _____

7. Suppose you were running a cloud service company, and you had a database of hundreds of thousands of users' public keys. What could you do with this?

   ◯ Derive the private keys and open SSH sessions into the customers' computers (if their services and firewalls are set up favorably)

   ◯ Sign digital signatures on behalf of users.

   ◯ They allow encryption of messages that can be decrypted by anyone with the public key.

   ◯ Use the keys to encrypt private information about each user.

   ◯ Install public keys on your servers to let users authenticate without passwords.

8. What is not a benefit of parallelizing workflows on the cloud?

   ◯ Cloud deployment can reduce the total volume of data processed (and associated costs) using data compression.

   ◯ Cloud-deployed clusters can return results faster in exchange for cloud costs.

   ◯ Cloud-deployed clusters can be resized dynamically, allowing hosts to respond to spikes in demand.

   ◯ Cloud vendors have machine images advertised to be compliant with various flavors of security regulations.

   ◯ Cloud-deployed clusters can be made geographically diverse, improving robustness in case of cloud outages.

9. One-time pads, with keys the same length as the messages, are theoretically unbreakable, as we've asked on a previous exam. Despite this advantage, they are not widely used. (Your phone probably doesn't have OTP software on it, for instance.) Why not?

Initials: _____

10. Which one of the following best describes how Dask handles large datasets?

- ◯ It requires the entire dataset to fit into memory.
- ◯ It splits data into smaller, manageable chunks and processes them in parallel.
- ◯ It compresses the data into a single file before processing.
- ◯ It normalizes the data into one-hot-encoded vector representations, which can be handled in parallel.
- ◯ It compresses data to allow faster computations.

11. What is the average time complexity of searching for an element in a sorted list in terms of the number of elements stored?

- ◯ $O(1)$
- ◯ $O(\log n)$
- ◯ $O(n)$
- ◯ $O(n^2)$
- ◯ $O(n \log n)$

12. Which one of the following provides the least protection against potentially harmful student code?

- ◯ Run student code in a docker container (that can't access the local hard drive)
- ◯ Run student code on an EC2 node (that doesn't have anything worth stealing)
- ◯ Run student code in a docker container on an Amazon EC2 node
- ◯ Run student code in its own directory on a shared server
- ◯ Run student code in a virtual machine

Initials: _____

Here is an excerpt from a ‘models.py‘:

```
class Reasons(models.Model):
    reason   = models.CharField(max_length=100)


class Posts(models.Model):
    post_id      = models.AutoField(primary_key=True)
    content      = models.CharField(max_length=300)
    title        = models.CharField(max_length=120)
    adddate      = models.DateTimeField(auto_now_add=True)
    reply_to     = models.ForeignKey(Posts, null=True,
       on_delete=models.CASCADE)
    creator      = models.ForeignKey(UserDetail, on_delete=
       models.CASCADE)
    is_suppressed = models.BooleanField(default=False)
    suppressed_reason = models.ForeignKey(Reasons, on_delete
       =models.CASCADE, null=True)
```

13. Does this schema need to be changed if our chat app allows editing of posts after they have been published? If so, what fields or structure would need to change?

14. Does this schema need to be changed if our chat app allows deletion of posts after they have been published? If so, what fields or structure would need to change?

Initials: _____

15. This schema includes a foreign key constraint in the `Posts` table that refers to the `Posts` table. Does this create a problem with circular dependency? If so, how must the business logic mitigate this?

```
 hashlib.pbkdf2_hmac(hash_name, password, salt, iterations, ...)

    The function provides PKCS#5 password-based key derivation
    function 2.  It uses HMAC as pseudorandom function.

    The number of iterations should be chosen based on the hash
    algorithm and computing power. As of 2013, at least 100,000
    iterations of SHA-256 are suggested.
```

```
def test_puzzle_misspell(self):
    print("puzzle_misspell:", puzzle_misspell)
    self.assertEqual(hashlib.pbkdf2_hmac("sha256",
           puzzle_misspell.encode("utf8"),
           "".encode("utf-8"),
           1000000),
        b'F\xa3M\x08\xa6\xe4\xde\xb2!xM'+
        b\x91\xdb\x82$\xd4)\xd7\xbb_\xf'+
␣␣␣␣␣␣␣␣b'aS\x02w\xdaJX\x9c;kp%',
␣␣␣␣␣␣␣␣"hash(puzzle_misspell)␣doesn't match hash(key)")
```

16. HW6 asked you to guess a mispelling of a word from a literary text. The above excerpt from the HW6 autograder takes the sha256 hash of the answer 1 million times. Why? Why apply the one-way function 1 million times when once is enough to hide the answer?

Initials: _____

17. Git uses a chain of cryptographic hashes to identify commits. What is the primary benefit of this design?

- ◯ It allows Git to store diffs rather than full file versions.
- ◯ It enables concurrent editing of the same file by multiple users.
- ◯ It ensures that any modification to a past commit can be detected.
- ◯ It allows users to rebase without introducing merge conflicts.
- ◯ It automatically compresses older commits for storage efficiency.

18. Apple has maybe 800 million icloud customers, and we can guess that each customer has not more than 10 million files. So Apple has 10^{16} user files to keep track of.

    Suppose Apple wants to notify law enforcement if any user data has the same hash as any hash on a list of 10^7 hashes of data whose possession is illegal (18 U.S.C. §2252). What is the approximate probability that the hash of at least one user file that is not actually forbidden will accidentally collide with a hash on the blacklist? $2^{256} \approx 10^{77}$

- ◯ 10^{-27}
- ◯ 10^{-35}
- ◯ 10^{-54}
- ◯ 10^{-70}
- ◯ 10^{-128}

19. Which of the following statements is true about RSA encryption?

- ◯ The security of RSA is based on the difficulty of factoring large integers
- ◯ Private keys are frequently published on the web
- ◯ RSA is typically used to encrypt very large files
- ◯ RSA requires symmetric key generation before use
- ◯ RSA uses the same key for encryption and decryption.

Initials: _____

20. Which of the following is not true of containerization tools like Docker? (Choose the most clearly incorrect answer.)

- ◯ Containers isolate applications, suggesting the design pattern of microservices.
- ◯ Containers virtualize an entire operating system
- ◯ Containers can be version-controlled and reproduced
- ◯ Containers can only be run on the machine where they were built
- ◯ Containers are typically slower to start than native applications

21. Which of the following are true of lossy compression algorithms? (2 points, check all that apply)

- ◯ They usually result in smaller file sizes than lossless compression
- ◯ They are suitable for text-based data storage
- ◯ They may discard information not noticeable to humans
- ◯ The compression algorithm is designed to be invertible
- ◯ They are commonly used in video and audio formats

22. Gradescope rents cloud computing nodes from Amazon to make the autograder work. Suppose 40 students submit their homework ≈ 10 times each over the course of a week, and each autograder session lasts 15 minutes. How much (within a factor of 10) do you expect gradescope is paying Amazon each week we use the autograder in this way?

Initials: _____

```
1     @number("13.0")
2     def test_hide_post_admin_success(self):
3         '''/app/hideComment␣endpoint␣by␣an␣admin
␣␣␣␣␣␣␣␣␣which␣should␣succeed'''
4         session = requests.session()
5         post_id = cloudysky_post_api(session,
              title="Bunnies␣suck",
              content="IDK␣why␣ppl␣like␣bunnies")
6         data = {'post_id': post_id,
              "reason": "anti-bunny␣hate␣speech" }
7         request = session.post(
              "http://localhost:8000/app/hidePost",
               data=data)
8        self.assertEqual(request.status_code, 200,
              "Expected␣success␣200␣for␣" +
              "http://localhost:8000/app/hidePost␣" +
              "Data:{}".format(data) )
```

23. Something is not right with the above test. It runs, but does not give the right answer. What is preventing this test from correctly verifying the behavior of the endpoint?

End exam 3. Have a great summer.

Initials: _____