# Big Data Systems

## A course for modern data modeling, data management and big data applications

Dr. Rubi Boim

# Why are we here?

# Why are we here?

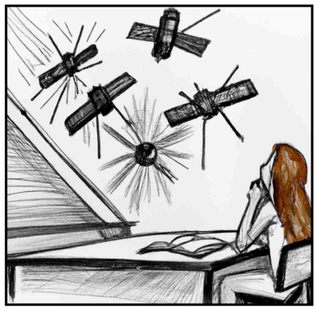
- Easy course? Hard course?
  depends on your progress
- Learn important / practical stuff
- Have fun ;-)

# מבחן במערכות ביג דאטה

ד"ר רובי בוים

סמסטר א' מועד א'
14 בפברואר, 2023

- משך המבחן שלוש שעות
- אין להשתמש בחומר עזר (חומר סגור)
- יש לענות תשובות קצרות ומנומקות (less is more, אין צורך למלא את כל הדף)
- **יש לכתוב את כל התשובות על הדפים (לא במחברת הבחינה)**
- טיפ למבחן (ובכלל לחיים) - תקראו את כל המבחן לפני שאתם מתחילים לכתוב
- סדר השאלות לא קשור לרמת הקושי, ובכלל כל השאלות לא קשות :-)
- סה"כ 100 נקודות
- **בהצלחה!**



* Generated by DALL-E 2
*"a sketch of starlink satellites hovering outside the atmosphere while a student is taking her final exam"*




ת"ז סטודנט: _____

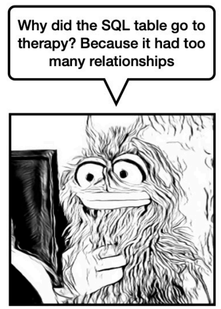
# מבחן במערכות ביג דאטה

ד"ר רובי בוים

סמסטר א' מועד א'
14 באפריל, 2024

- משך המבחן שלוש שעות
- אין להשתמש בחומר עזר (חומר סגור)
- יש לענות תשובות קצרות ומנומקות (less is more, אין צורך למלא את כל הדף)
- **יש לכתוב את כל התשובות על הדפים (לא במחברת הבחינה)**
- טיפ למבחן (ובכלל לחיים) - תקראו את כל המבחן לפני שאתם מתחילים לכתוב
- סדר השאלות לא קשור לרמת הקושי, ובכלל כל השאלות לא קשות :-)
- סה"כ 100 נקודות **(שאלה 15 היא שאלת בונוס לקבוצת ה 28 מעבר ל 100)**
- **בהצלחה!**

* Generated by DALL-E
*"Taylor Swift driving a Tesla"*






ת"ז סטודנט: _____

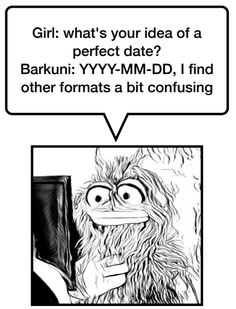
# מבחן במערכות ביג דאטה

ד"ר רובי בוים

סמסטר א' מועד א'
12 לפברואר, 2025

- משך המבחן שלוש שעות
- אין להשתמש בחומר עזר (חומר סגור)
- יש לענות תשובות קצרות ומנומקות (less is more, אין צורך למלא את כל הדף)
- **יש לכתוב את כל התשובות על הדפים (לא במחברת הבחינה)**
- טיפ למבחן (ובכלל לחיים) - תקראו את כל המבחן לפני שאתם מתחילים לכתוב
- סדר השאלות לא קשור לרמת הקושי, ובכלל כל השאלות לא קשות :-)
- סה"כ 100 נקודות **(שאלה 14 היא שאלת בונוס לקבוצות B/C מעבר ל 100)**
- **בהצלחה!**

* Generated by DALL-E
*"Sam Altman installing a server in a rack at a data center"*






ת"ז סטודנט: _____

# WHAT IS BIG DATA

## Let’s try ChatGPT (DALL-E…)

“a 3d render of big data”


2022

“a 3d render of big data”

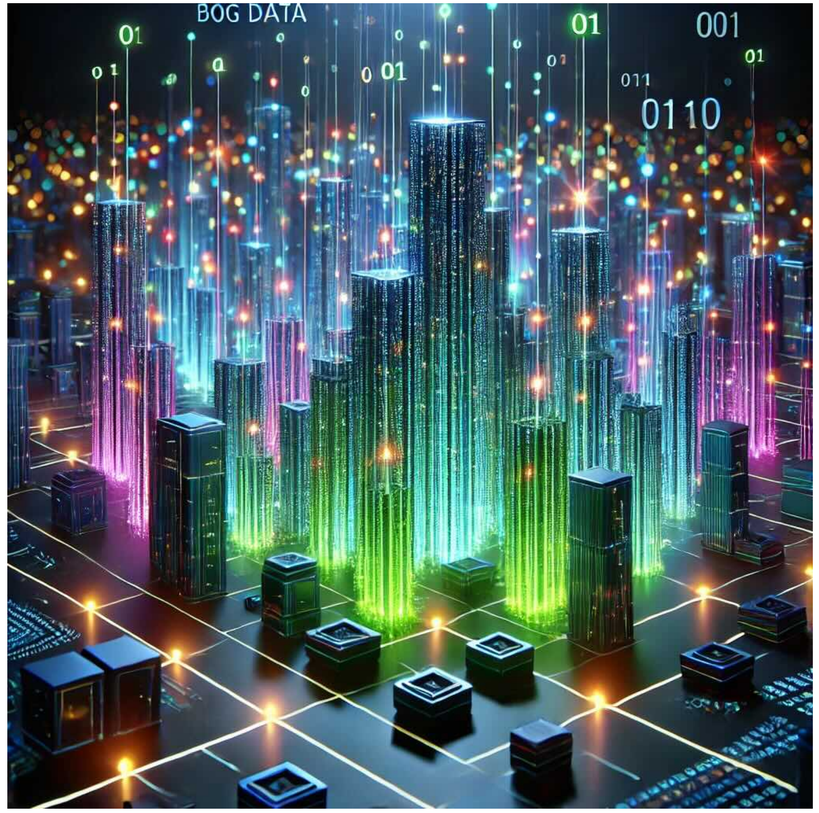

2022

2023

“a 3d render of big data”


2022

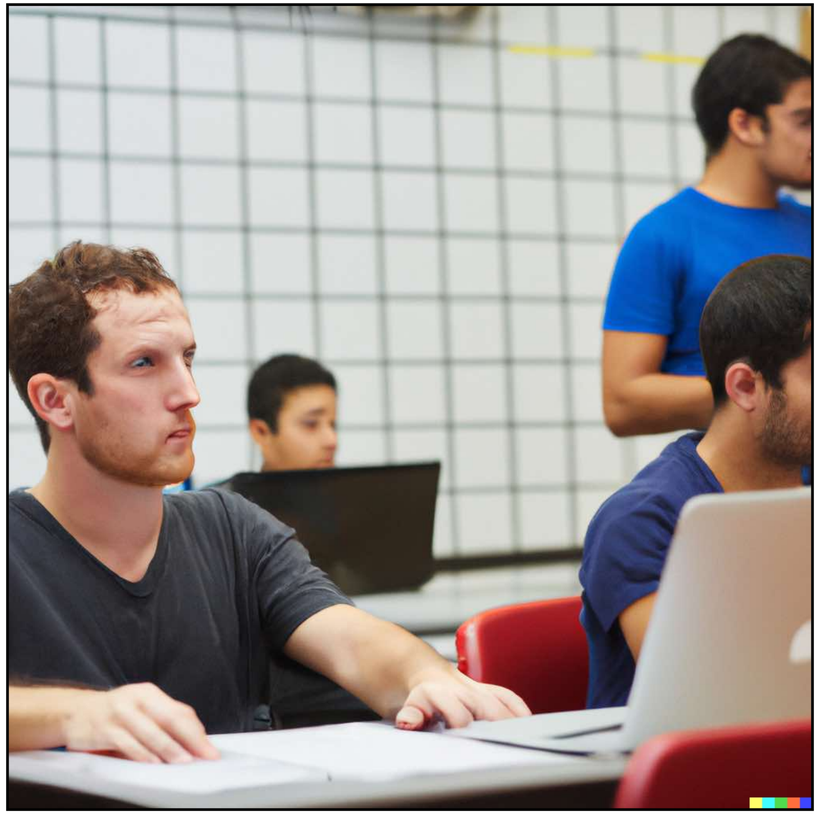
2023




2024

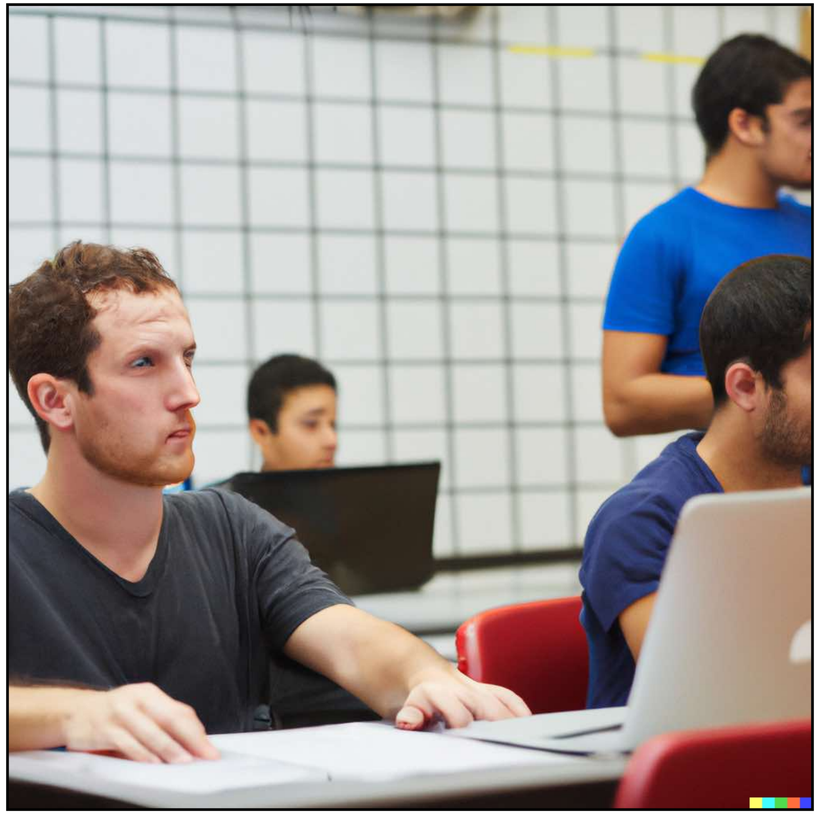
“a 3d render of big data”

“students at bigdata course at tel aviv university”



2022

“students at bigdata course at tel aviv university”



2022

“students at bigdata course at tel aviv university”

2023

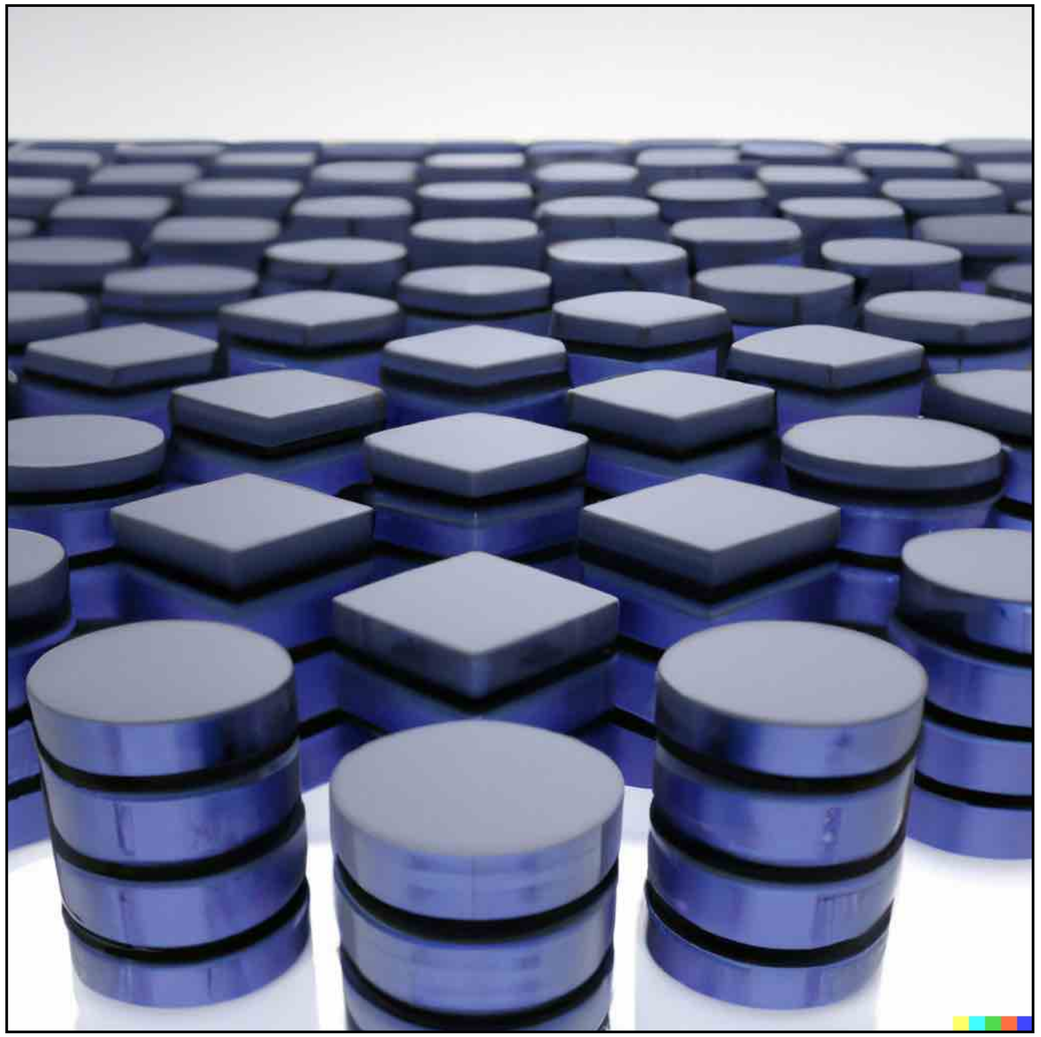

2024

“students at bigdata course at tel aviv university”

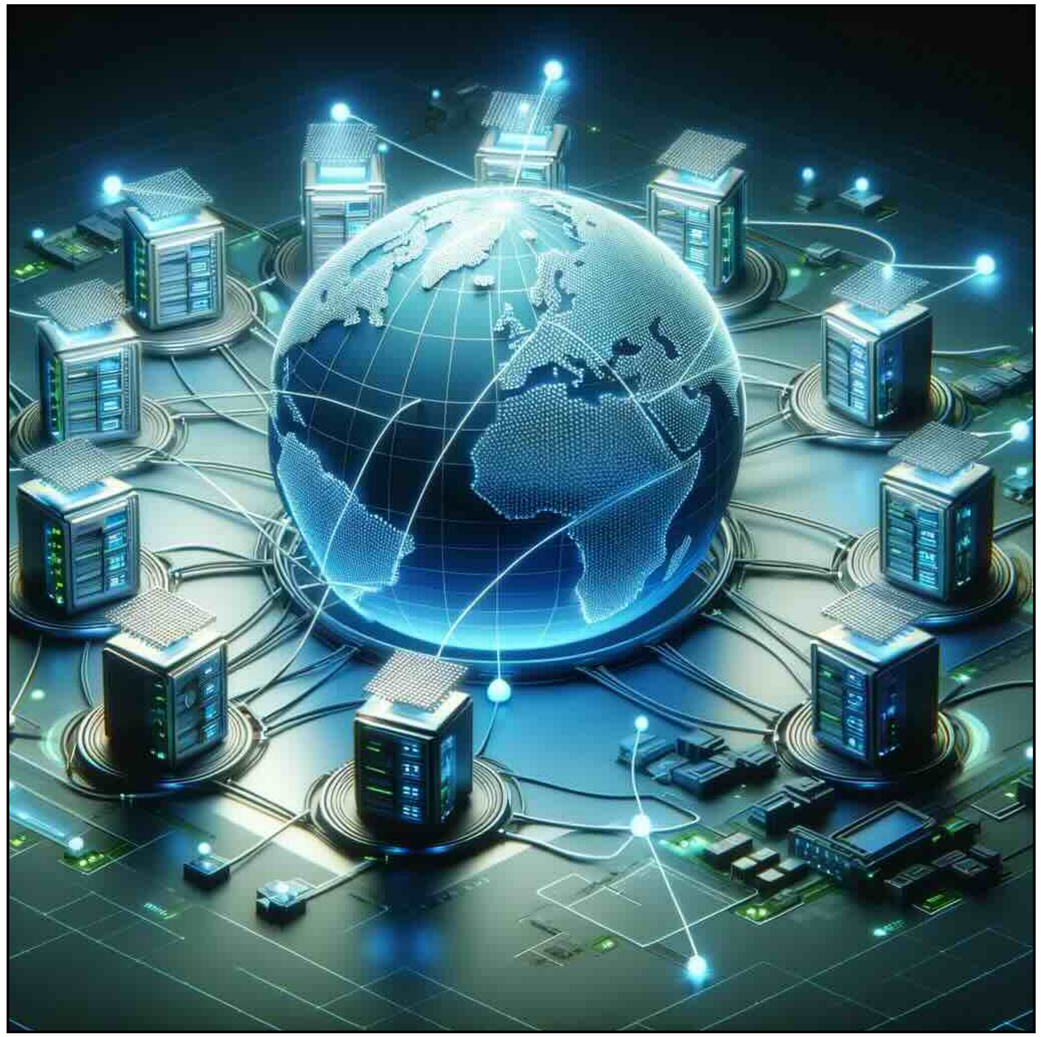

2025

“students at bigdata course at tel aviv university”

“a 3d render of distributed database”



DALL E 2



DALL E 3

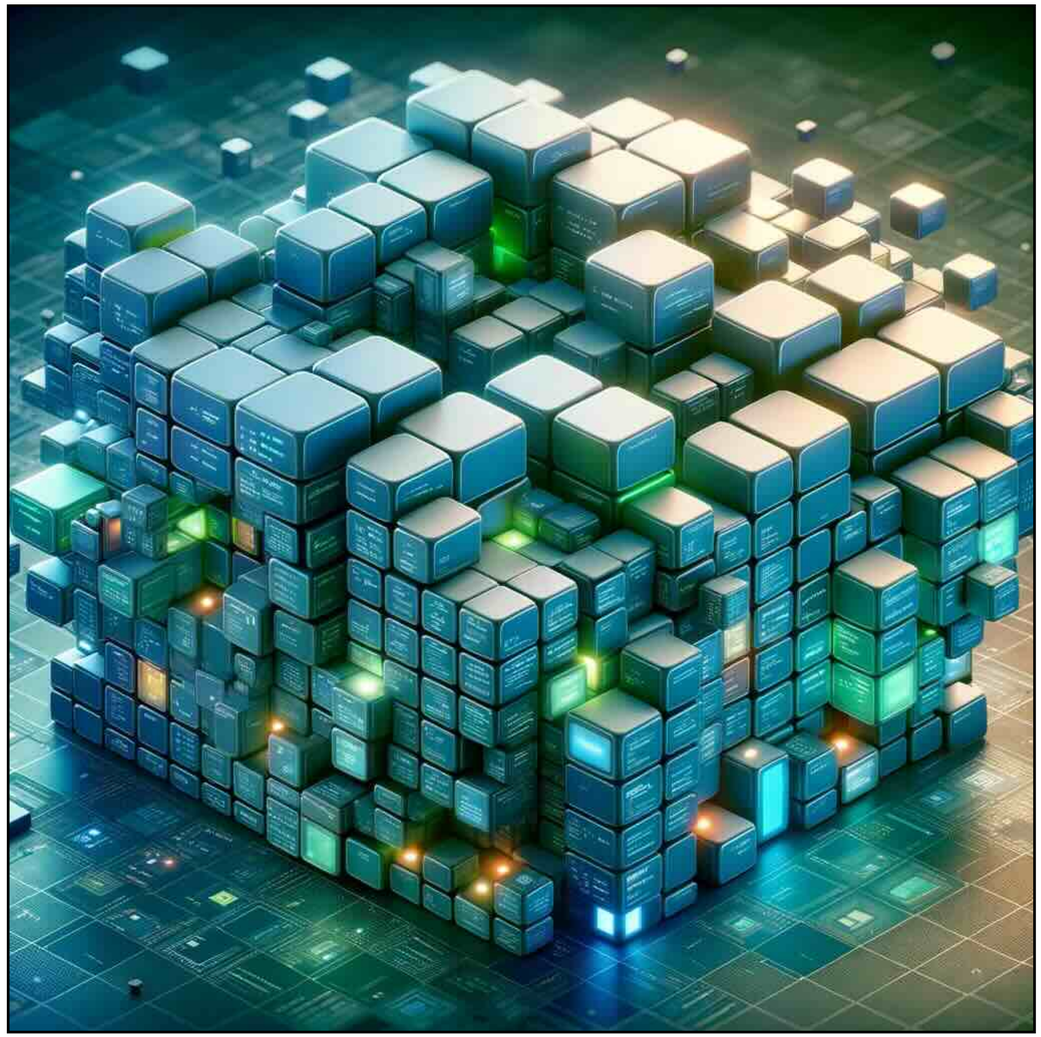
“a 3d render of distributed database”


2025

“a 3d render of distributed database”

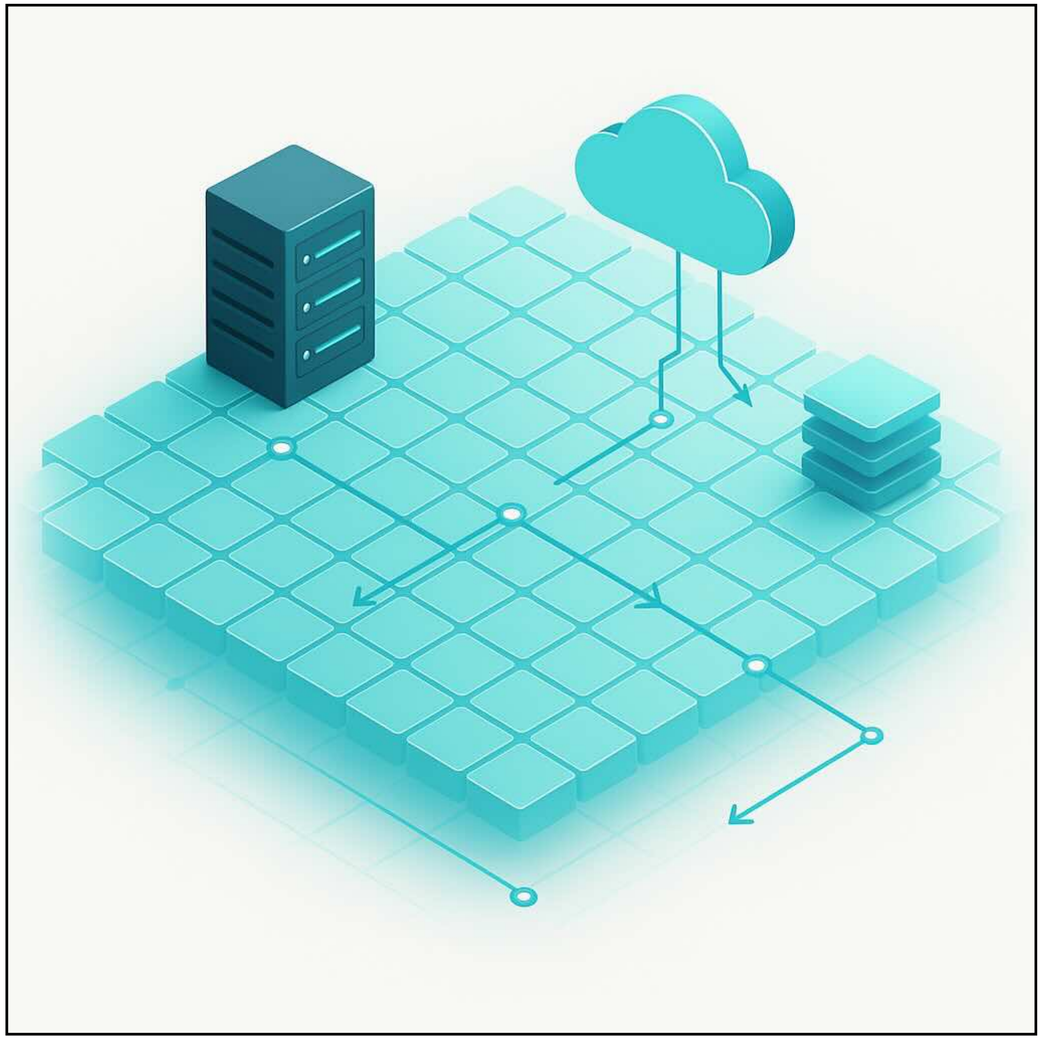
“a 3d render of bigtable”

DALL E 2



DALL E 3

“a 3d render of bigtable”

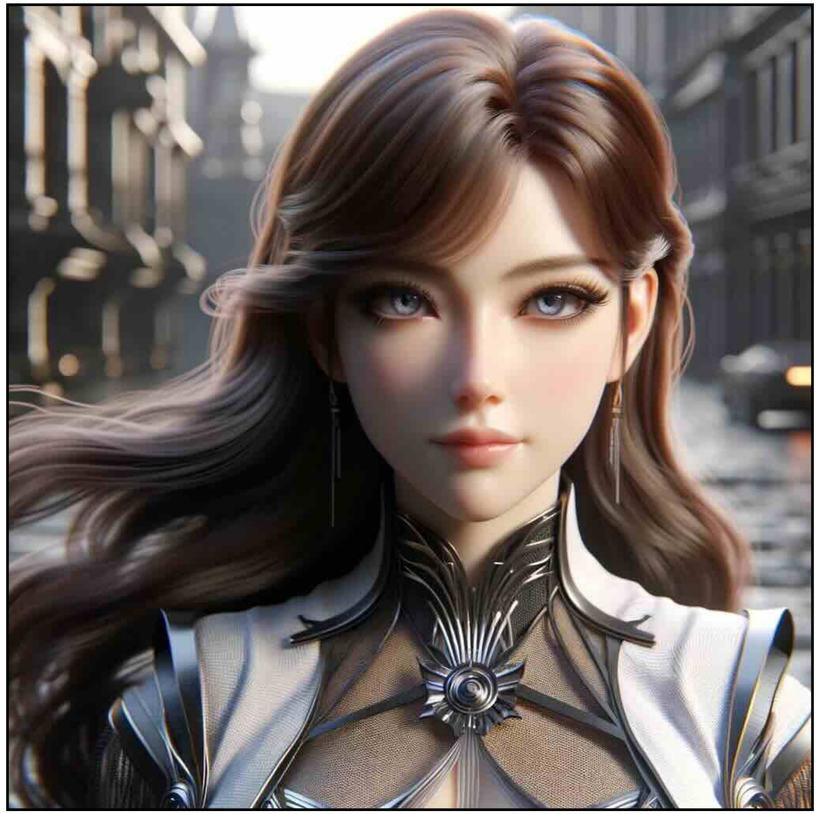


2025

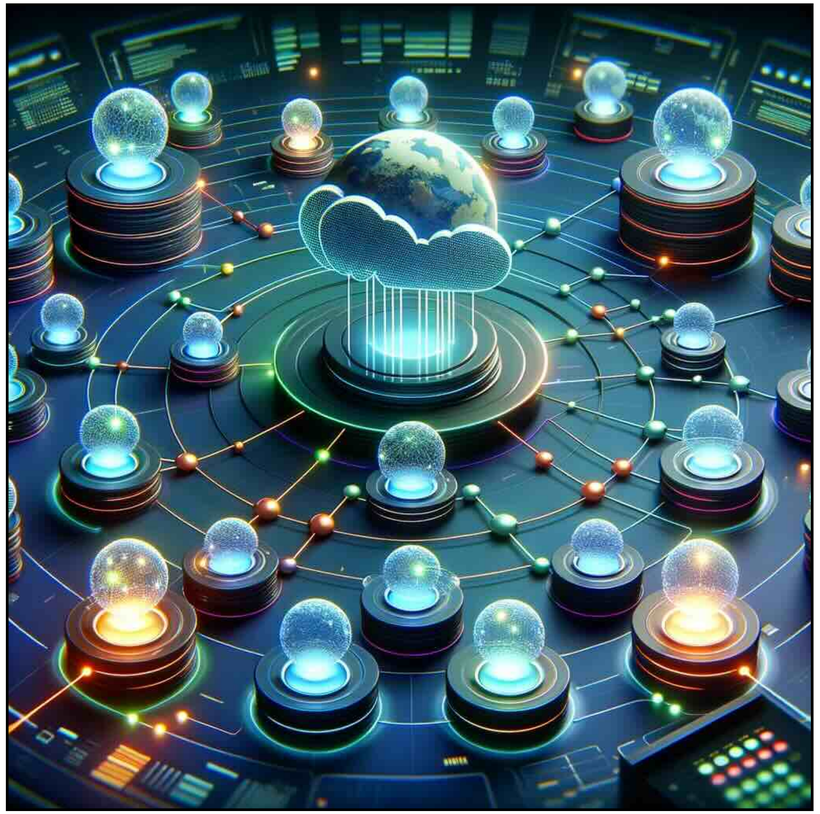
“a 3d render of bigtable”

“a 3d render of cassandra”

DALL E 2

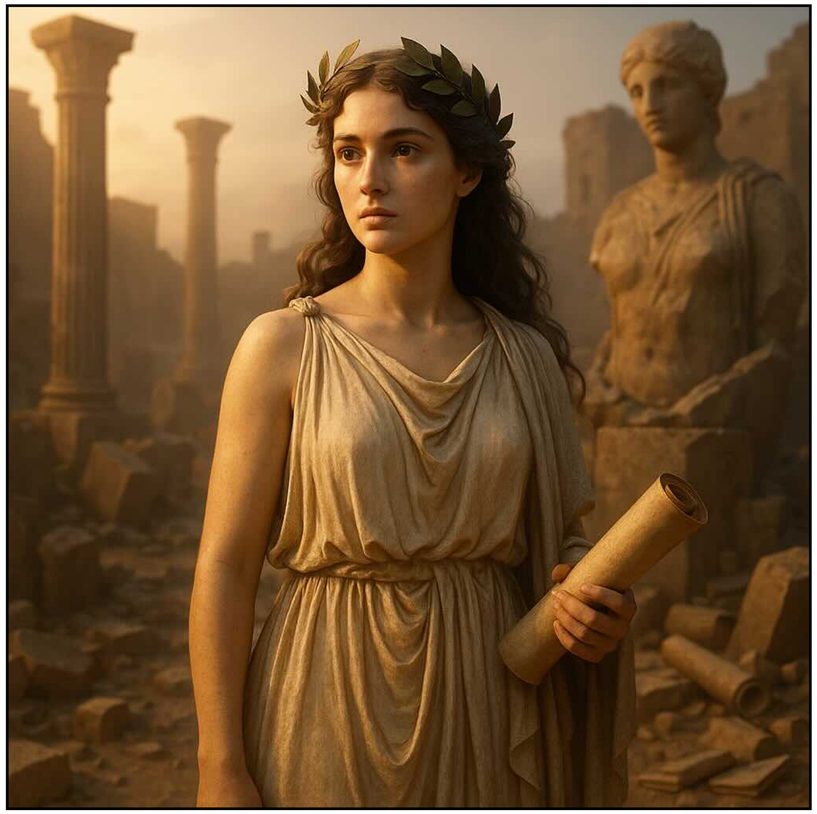


DALL E 3



DALL E 3 (Context)

“a 3d render of cassandra”

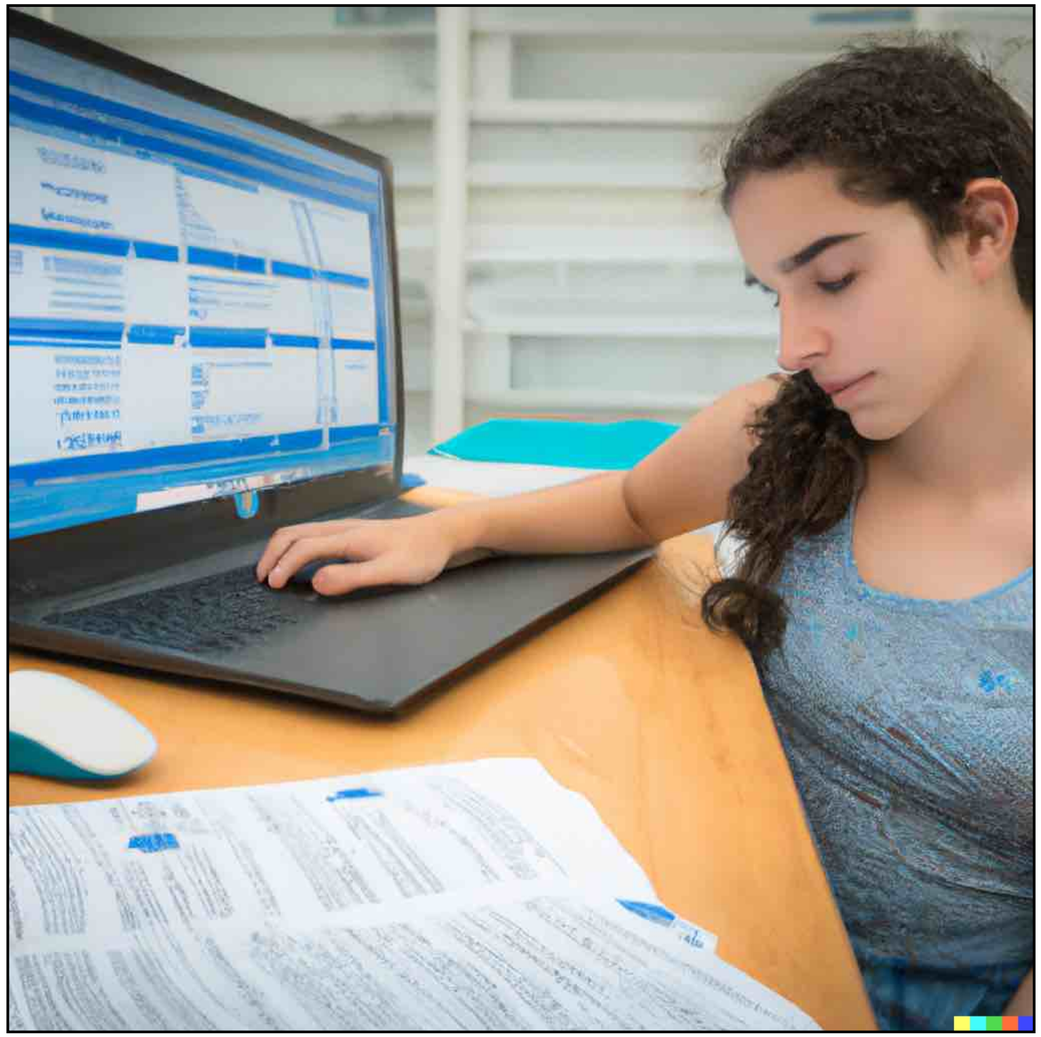


2025

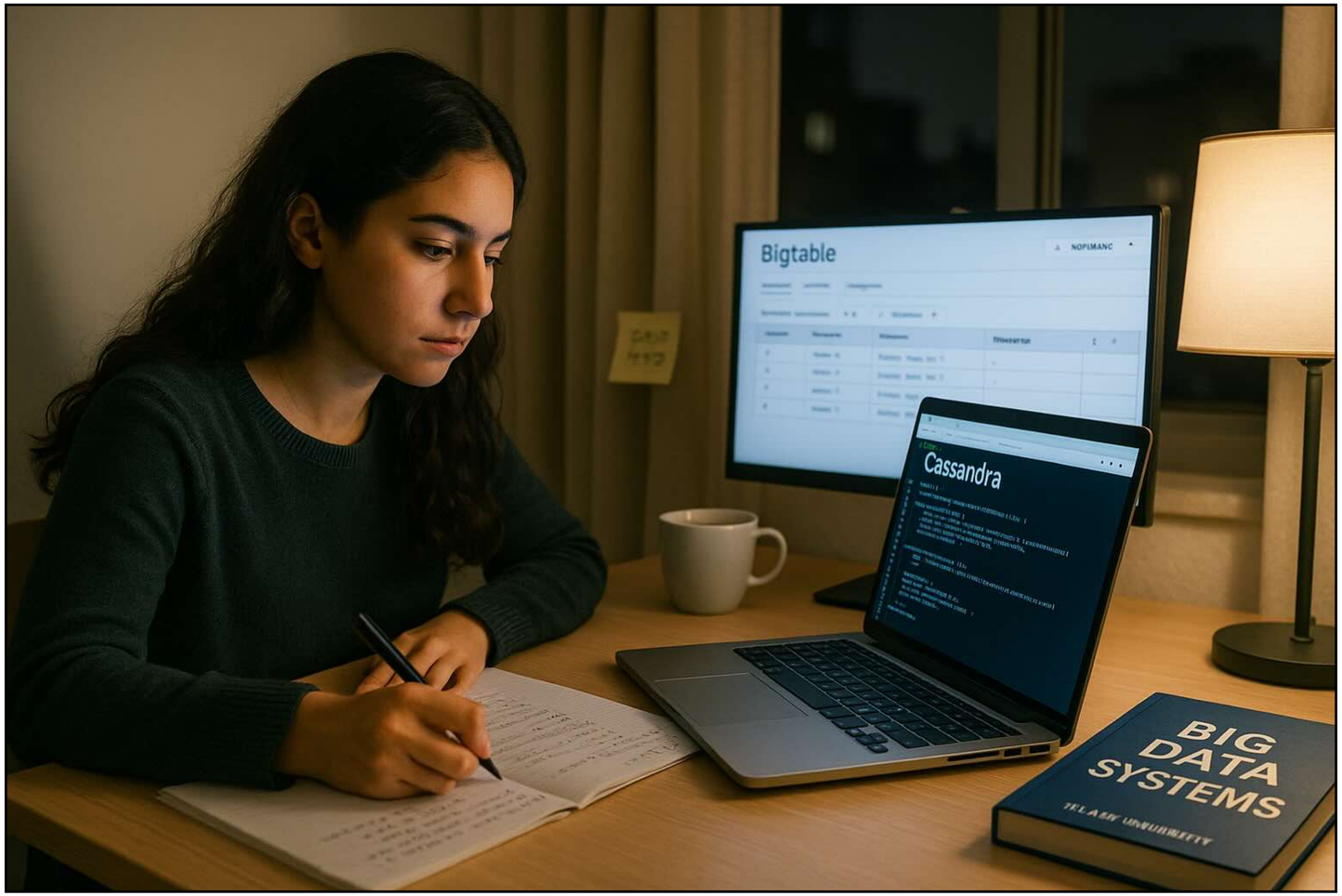
2025
Context

“a 3d render of cassandra”

“a student using cassandra and bigtable databases while doing her homework for the course big data system in tel aviv university”



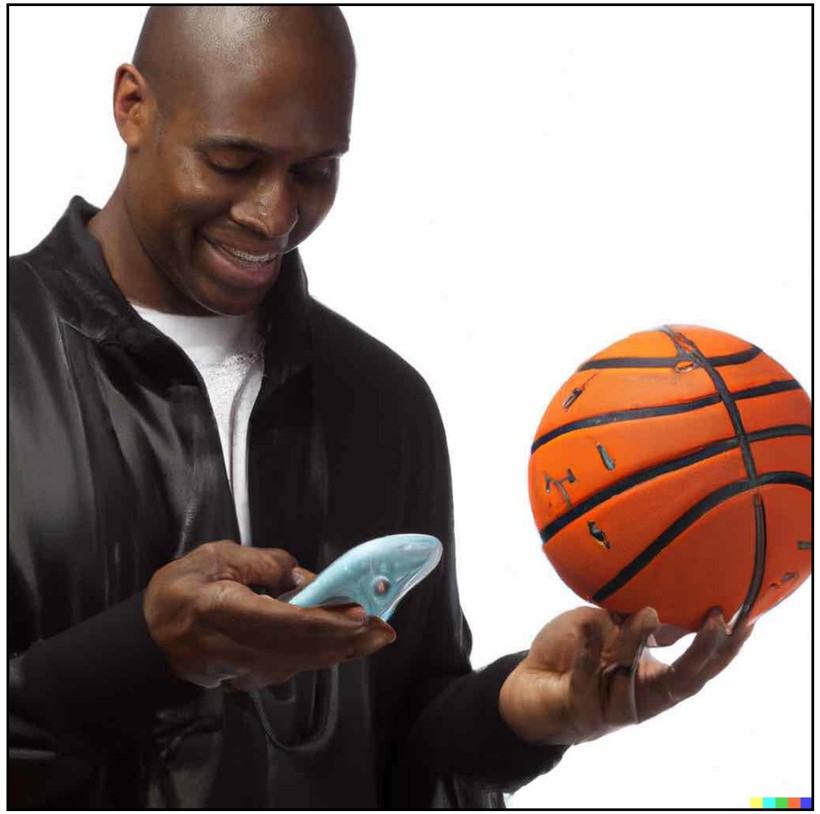

“a student using cassandra and bigtable databases while doing her homework for the course big data system in tel aviv university”



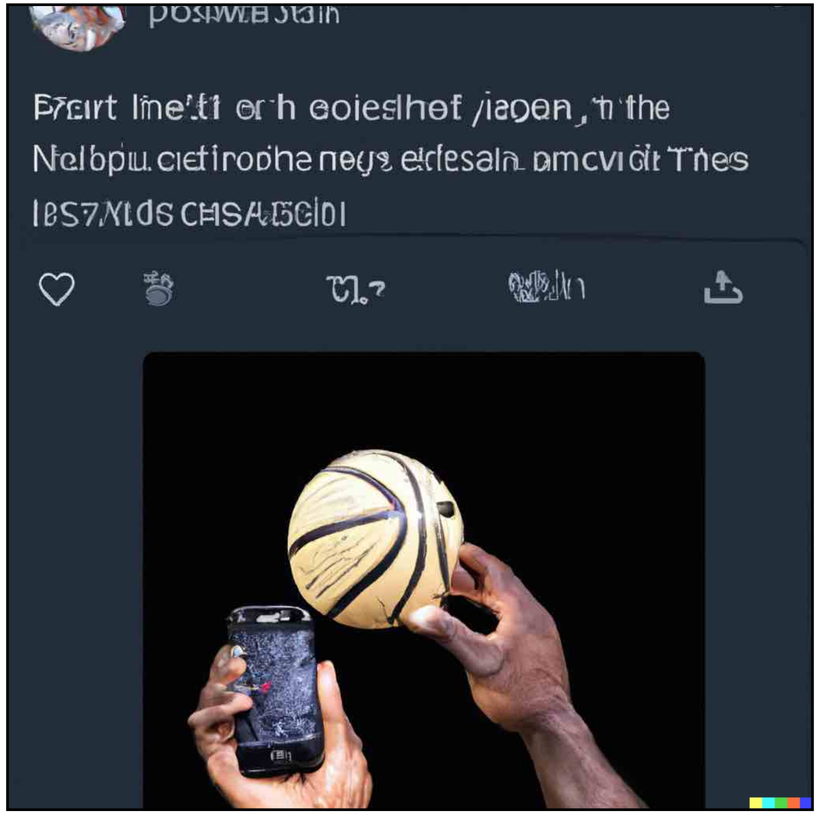

2025

“a student using cassandra and bigtable databases while doing her homework for the course big data system in tel aviv university”

“michael jordan using twitter to tweet about a new basketball he bought”



DALL E 2



DALL E 2

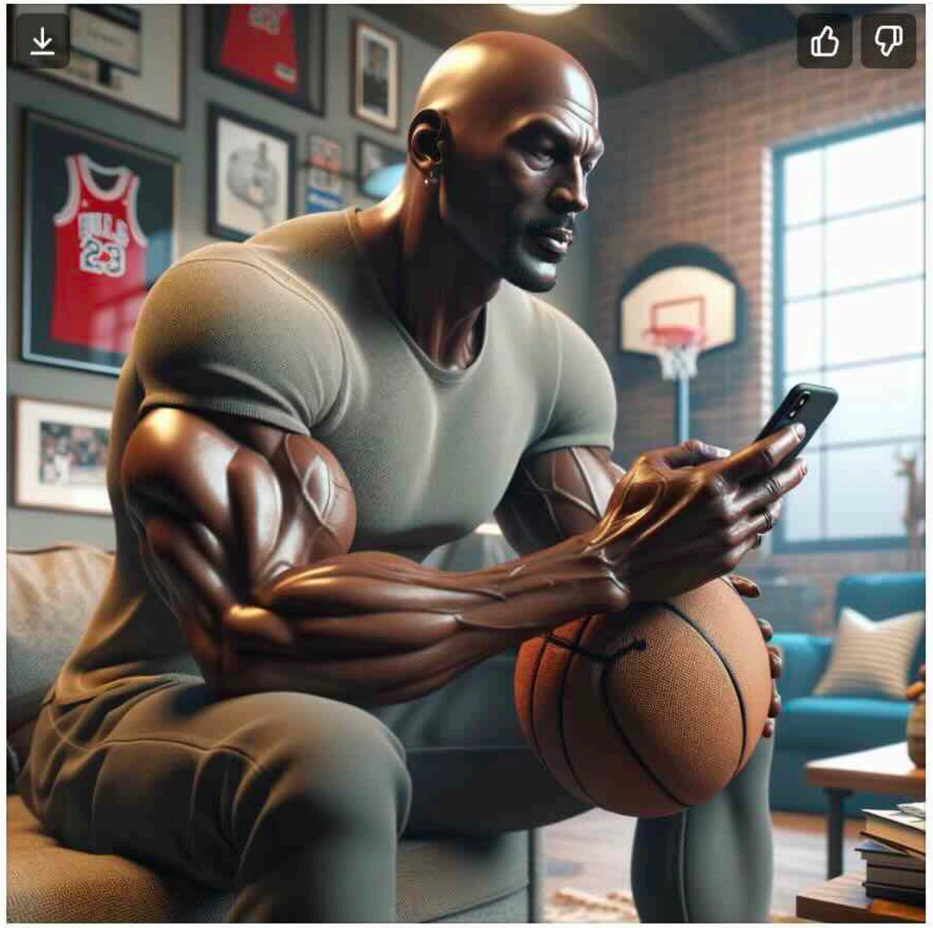
DALL E 3

“michael jordan using twitter to tweet about a new basketball he bought”

**You**

please make him look like michael

**ChatGPT**



Here's the image created by DALL·E, showing a basketball player with a physique and style reminiscent of Michael Jordan, using his smartphone.

**You**

please make him look like michael

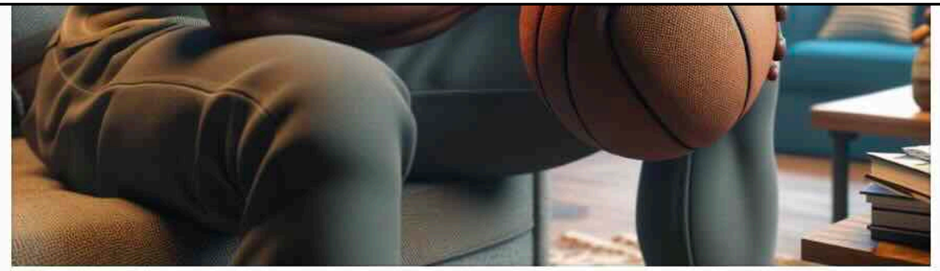
**ChatGPT**

ChatGPT 4o

Create an image “michael jordan using twitter to tweet about a new basketball he bought”

I'm unable to generate the requested image because it doesn't align with our content policy. If you have any other ideas or adjustments, feel free to share, and I'll do my best to help!

2024



Here's the image created by DALL·E, showing a basketball player with a physique and style reminiscent of Michael Jordan, using his smartphone.


# 2025
# (GPT 5 / gpt-image-1)

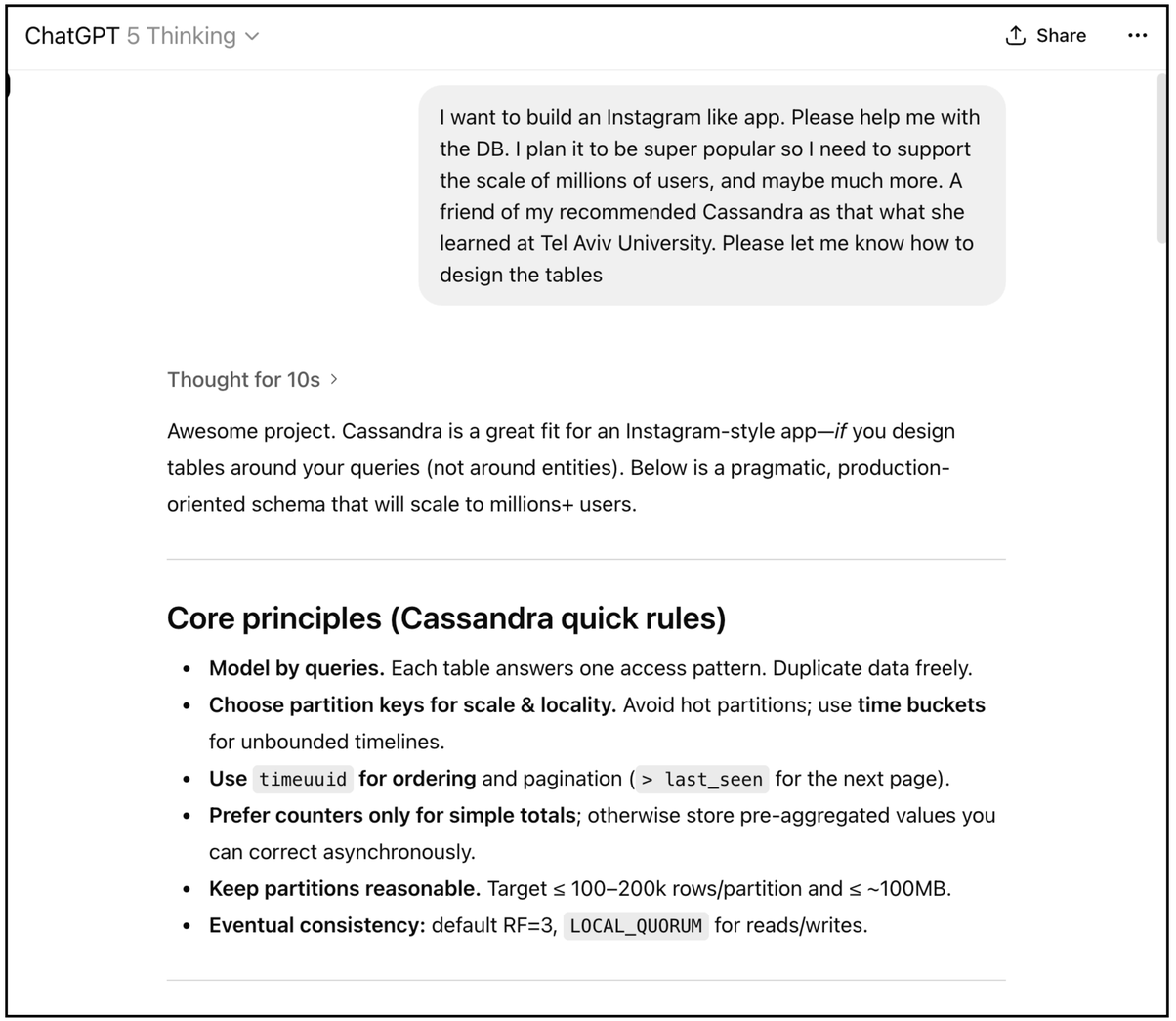
# Big data is like teenage sex

- everyone talks about it
- nobody really knows how to do it
- everyone thinks everyone else is doing it
- so everyone claims they are doing it…

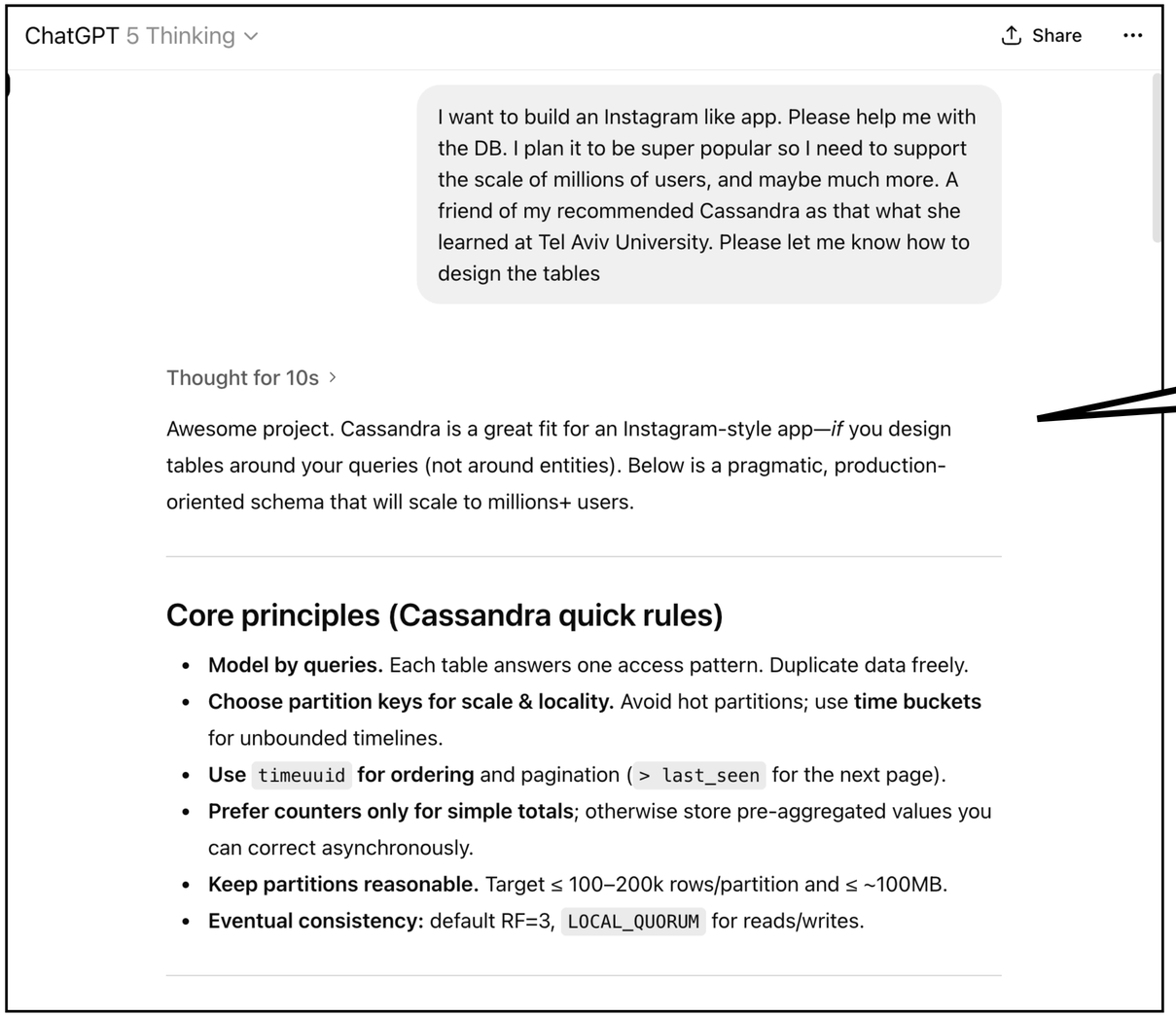
*(Dan Arieli, 2013)*

# Do we need the course in the “AI” era?

# Do we need the course in the “AI” era?

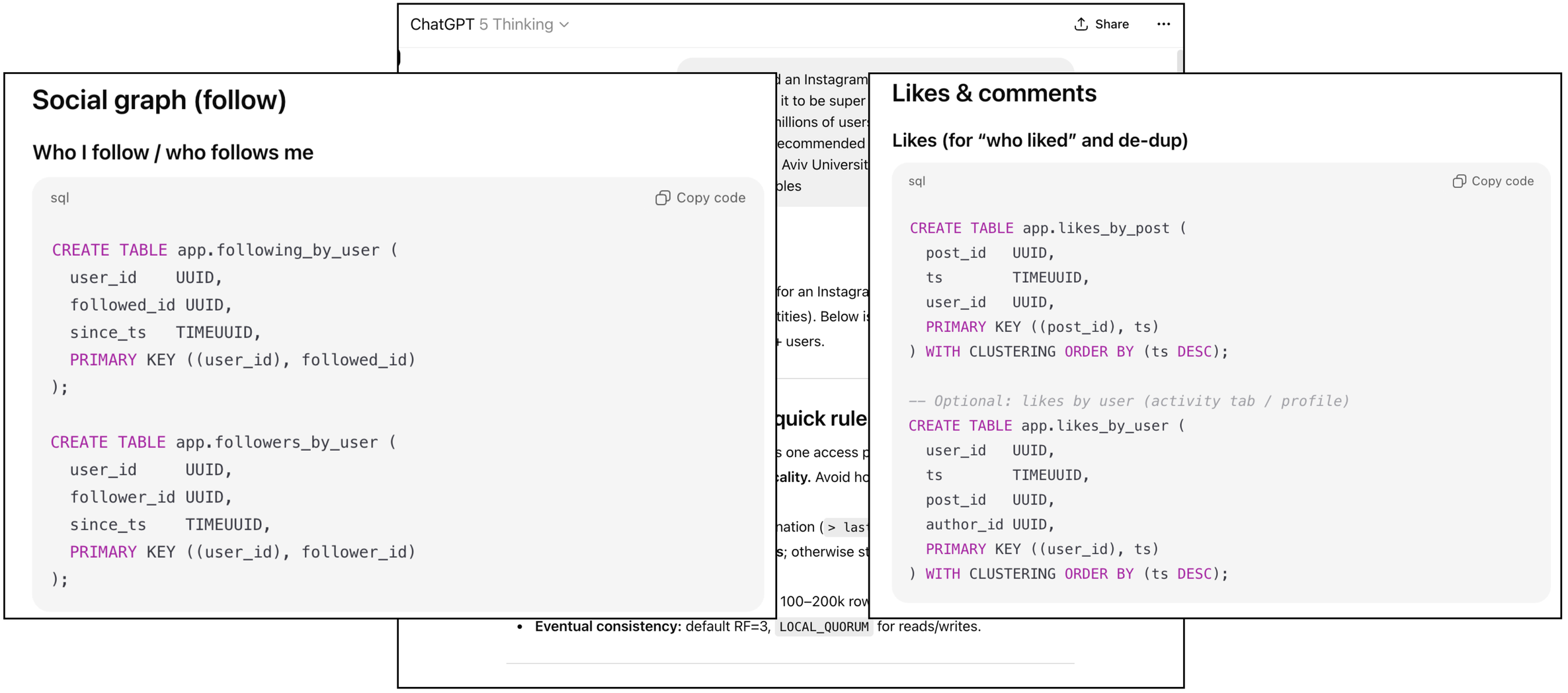


# Do we need the course in the “AI” era?




These are actually good points!

# Do we need the course in the “AI” era?



# Do we need the course in the “AI” era?

ChatGPT

Share

## Social graph (follow)

### Who I follow / who follows me

sql

```
CREATE TABLE app.following_by_user (
  user_id    UUID,
  followed_id UUID,
  since_ts   TIMEUUID,
  PRIMARY KEY ((user_id), followed_id)
);

CREATE TABLE app.followers_by_user (
  user_id     UUID,
  follower_id UUID,
  since_ts    TIMEUUID,
  PRIMARY KEY ((user_id), follower_id)
);
```

d de-dup)

Copy code

```
_by_post (

d), ts)
R BY (t

user (activity tab / profile)
_by_user (

d), ts)
R BY (ts DESC);
```

In a few lessons you will understand :)

# Course goal

provide the theoretical as well as the practical hands on knowledge required for designing and developing internet scale based data applications

**(from a data management and data modeling perspective)**

# privilege to teach and learn

failing in this course is not an option.

# understanding the motivation

examples before theories

# Buzzwords to be covered  (tentative)

- Relational databases
  - SQL and normalized DB
- Distributed databases
  - NoSQL (wide column)
  - CAP theorem
  - Dynamo
  - BigTable
  - **Cassandra**
  - Advance data modeling
- Other
  - Kafka
  - Data warehouse

# Staff

## Dr. Rubi Boim

- boim AT cs.tau.ac.il
- https://www.cs.tau.ac.il/~boim/
- Office hours: by appointment

## Nadav Magar (TA)

- nadavmagar AT mail.tau.ac.il

# Communications

- Course website:
  https://courses.cs.tau.ac.il/bigdata/
- Moodle
  https://moodle.tau.ac.il/course/view.php?id=368327601

# Requirements


- Written test
- 2-4 HW assignments
  - in pairs
  - Java as programming language
- Database system course

  highly recommended prerequisite or in parallel

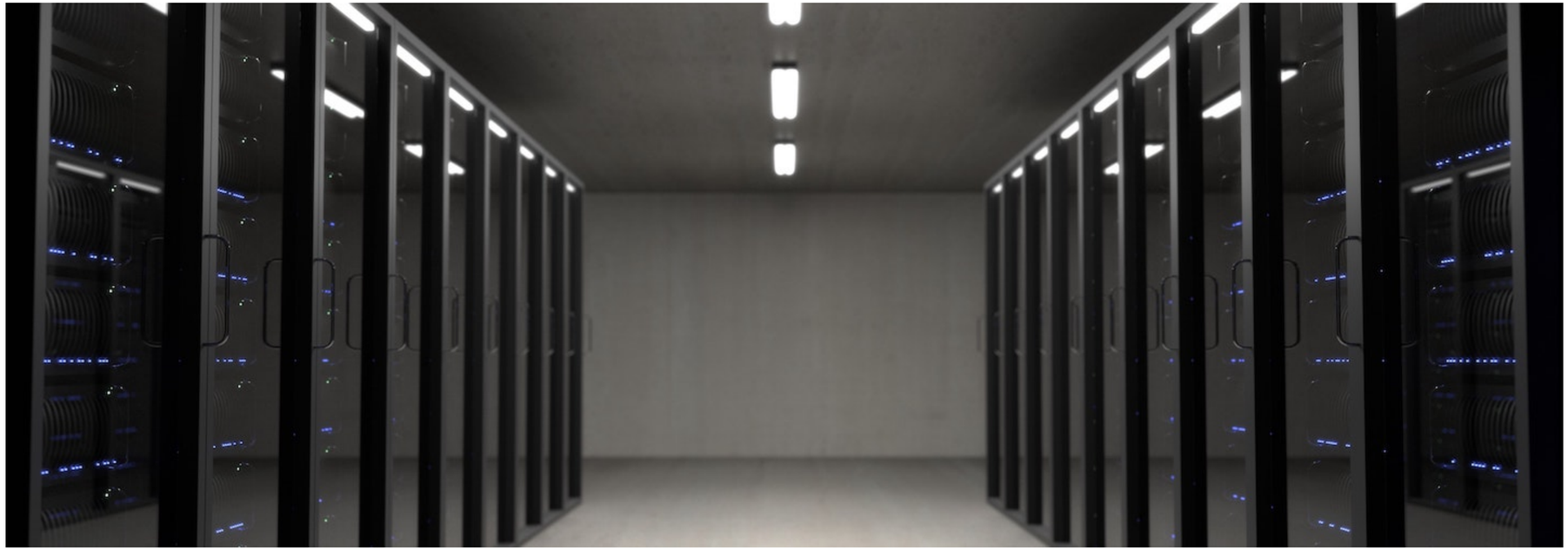
# Reference texts

- Principles of Distributed Database Systems *(M. Tamer Özsu, Patrick Valduriez)*
- Brewer's conjecture and the feasibility of consistent, available, partition-tolerant web services *(Seth Gilbert et al.)*
- CAP Twelve Years Later: How the “Rules” Have Changed *(Eric Brewer)*
- Bigtable: A Distributed Storage System for Structured Data *(Fay Chang et al.)*
- The Google File System *(Sanjay Ghemawat et al.)*
- Cassandra - A Decentralized Structured Storage System *(Avinash Lakshman et al.)*
- Dynamo: Amazon’s Highly Available Key-value Store *(Giuseppe DeCandia et al.)*
- Kafka: a Distributed Messaging System for Log Processing *(Jay Kreps et al.)*

# Course plan (tentative)


- Intro to big data
- Relational databases
- **Distributed databases and techniques**
- **Cassandra (wide column databases)**
- **Advanced modeling**
- Data warehouse
- Streaming

# Course schedule



https://courses.cs.tau.ac.il/bigdata/