My grades for **Final exam**

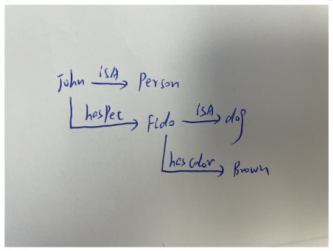
Q1

5 / 5

What are RDF (semantic) triples? Using your own example, illustrate (ie. with diagrams) how they can be used to construct a knowledge graph ('KG').

RDF triples are database stores triples of (subject,predicate,object). A triple defines a directed binary relation, via its predicate/attribute/property.

For example we have four triples: (John, isA, Person), (John, hasPet, Fido), (Fido, isA, Dog), (Fido, hasColor, Brown). The knowledge graph like below:



This knowledge graph helps represent the relationships between the entities in a structured and easily understandable manner, making it useful for various applications such as semantic search

Q2

4 / 5

In many data processing pipelines, several steps can be carried out at the same time (speeding up processing), while other steps cannot. How would we describe such a pipeline at a higher level (of specification), and how, at a lower level? Illustrate using a simple example, and provide short explanations.

-1: No definition of lower or higher level of pipelines

4

Example: Image processing
Step:

1. Load images from a directory.
2. Resize images.
3. Convert images to grayscale.
4. Apply a filter to the images.
5. Save the processed images to a new directory.

Higher-level description:

1. **Parallel:** 2-4 steps can be parallelized, as they can be performed independently for each image.
2. **Sequential:** 1 and 5 steps, however, are sequential, as they involve reading from and writing to the disk.

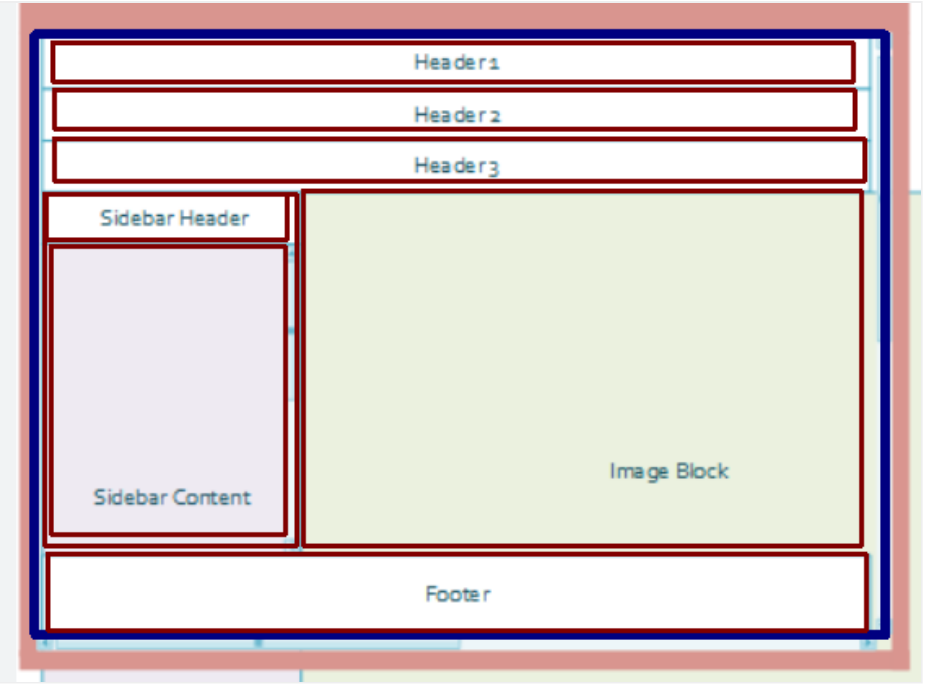
Lower-level description:

1. **Sequential:** Load images from a directory one by one.
2. **Parallel:** Resize each image independently of the others.
3. **Parallel:** Convert each image to grayscale independently of the others.
4. **Parallel:** Apply a filter to each image independently of the others.
5. **Sequential:** Save the processed images to a new directory one by one.

Q3

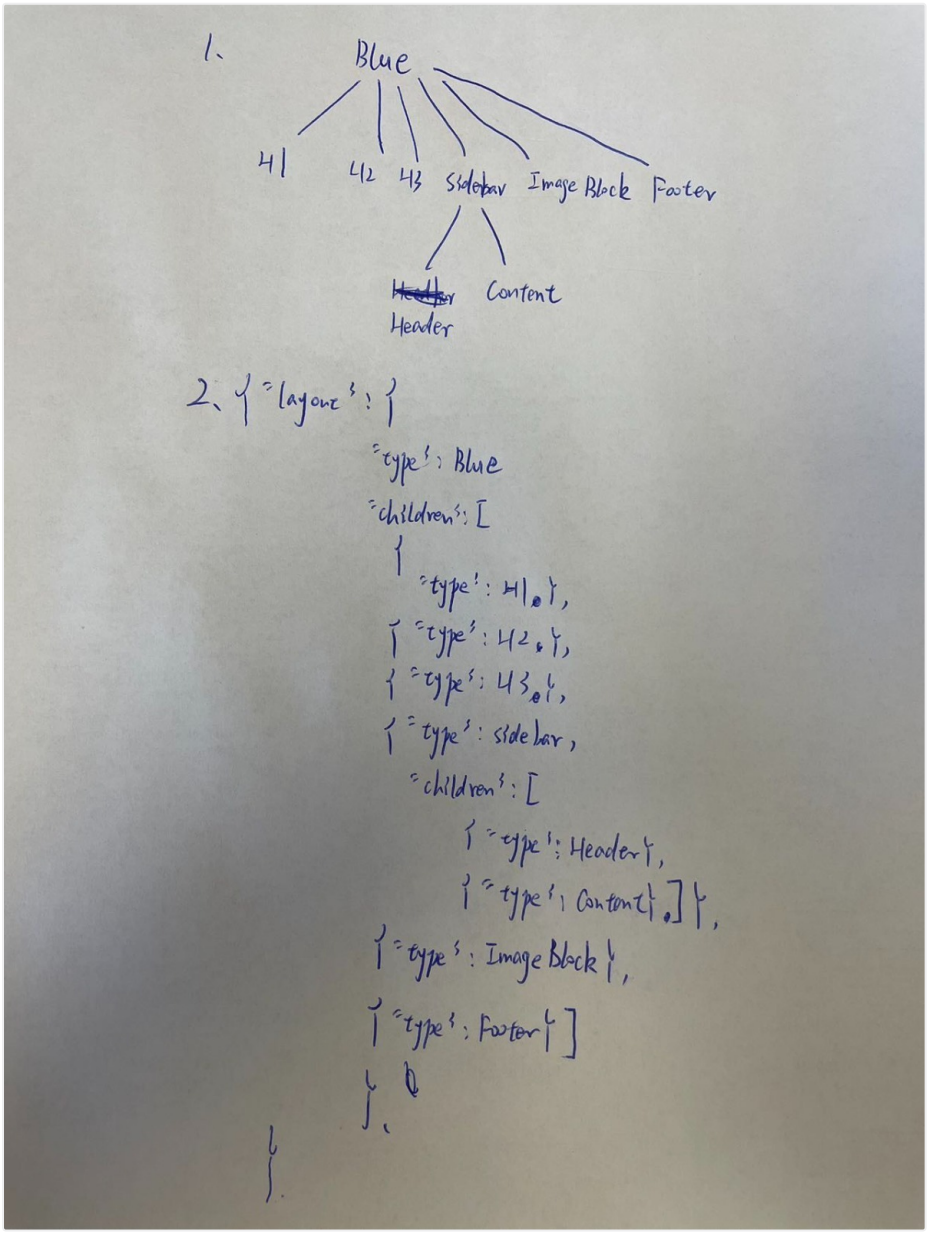
5 / 5

Consider the following UI layout, where a blue container (eg a 'div' in HTML) holds the other elements (in dark red), in a 'nested' layout that is typically how UI gets laid out - it can be regarded as a form of data specification.



1. Draw a simple tree that shows the layout hierarchy. For anything not named (hint - there is one such!), you can make up a name for it.

2. Describe the layout in syntactically valid JSON, using "layout" as the main (top-level) key.


1.
Blue
H1
H2
H3
Sidebar
Image Block
Footer
Header
Content
2. {"layout": {
"type": Blue
"children": [
{
"type": H1,},
{"type": H2,},
{"type": H3,},
{"type": sidebar,
"children": [
{"type": Header},
{"type": Content},]},
{"type": Image Block},
{"type": Footer}]
}}

Q4

5 / 5

1. In one word - what do we collect, clean, store data for?
2. Using your HW2 through HW5 as examples, explain in a couple of lines for each, how you made '1.' above, happen.

---

1. Insights
2. In HW3, I collected data by walking through the campus to get the latitude and longitude of several buildings and visualized them. In HW2 and HW4, I used two different ways(SQL, and NoSQL) to clean and analyze data, just focus on what we care about. In HW5, I trained a model that can learn the pattern of the data I was given and make predictions for new data.

Q5

5 / 5

In ML, why are loss functions important (what do they help with)?

What is the most important part of a neuron in an artificial neural network? WHY?

The UI in Q3 above, can be coded up using CSS (eg using divs and the 'flexbox' layout). But we could make genAI (eg ChatGPT) output such CSS for us, rather than writing it ourselves. What would be the training data, for this to become possible?

---

1. Loss functions help to quantify the error between the predicted output and the actual output of the model to the model's accuracy.
2. Activation function. Because the activation function determines whether or not a neuron learns based on the result from other neurons. Without an activation function, the output of the neuron would simply be a linear combination of its inputs.
3. The training data would be a huge number of CSS with the corresponding HTML. CSS for generator and HTML for discriminator.

Q6

2 / 2

For a number of years, databases were of the relational (tables) and NoSQL (k:v, documents, columns, graph) type. Now we have a new category - 'vector DBs', that are for carrying out 'similarity search' (where we query for data, without using SQL etc., but more 'loosely'). Explain in two or three sentences, how these queries/DBs work.

Each column in the table would be a dimension for vector space. Each row in the table would be a vector with K dimension in the vector space. Our query can be also modified as a K-dimensional vector. So we can find similar vectors(rows) between two queries and stored data using Cosine Similarity.

Q7

2 / 2

To query data, we mostly still use syntax-based languages such as SQL, Python, etc. Name, and discuss, TWO up-and-coming alternatives to this (hint: one of these is already in use, rather widely!).

---

1. **NLP-based querying**. It enables users to interact with data. Example: *Siri*.
2. **Graph Query Language**. It stores data in nodes (entities) and edges (relationships) instead of traditional tables. Example: *Neo4j*.

Q8

2 / 2

"From BI to AI" - what is the connection (similarity) between BI (data warehousing), and AI (a colloquial substitute for ML/DS)? Be specific, and answer in 2 or 3 lines.

They both involve extracting insights from data to support decision-making. BI focuses on organizing, analyzing, and visualizing data. AI uses algorithms and techniques to learn from data, enabling predictions and automation.

Q9

2 / 2

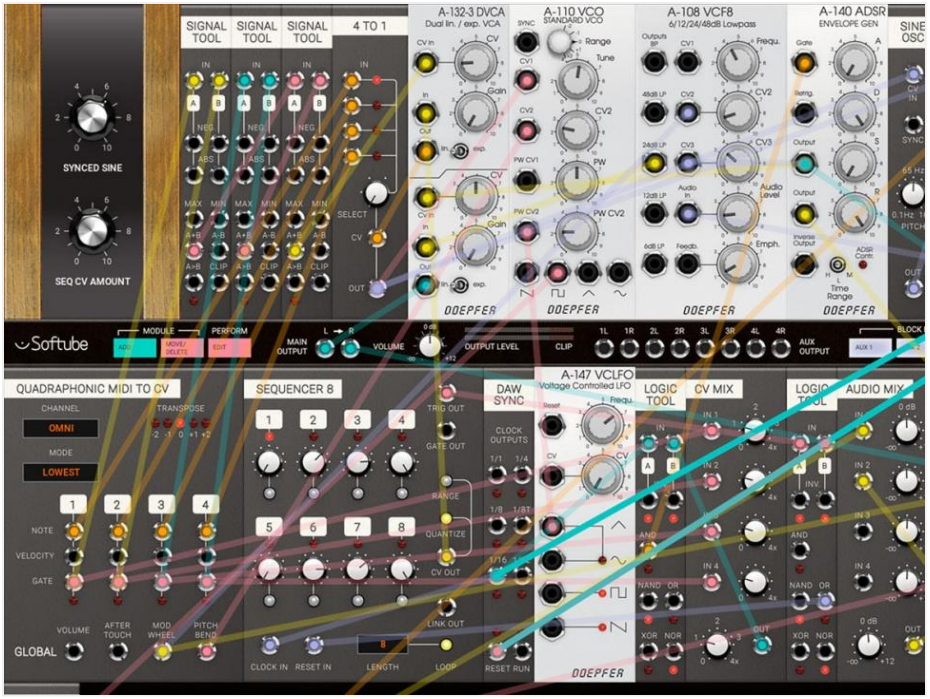
Large-scale data storage and querying have evolved, from being entirely centralized, to being entirely distributed. 'MCC' is how we currently "do distributed". In this context (MCC for data access), what does 'serverless' mean?

'serverless' means developers don't need to manage the underlying infrastructure. They can focus on their application logic. And cloud provider such as Amazon automatically allocates and scales resources, and charges are based on actual usage.

Q10

2 / 2

One way to synthesize complex audio is to hook up nodes (eg. https://noisecraft.app/623) or modules (like in the graphic below). How does this relate to data handling? In other words, in a few sentences - how do we conceptualize data pipelines? Hint: think of three 'stages' we discussed in class.



The modules in the data pipeline contain three stages: data ingestion, data processing, and data output.

1. **Data ingestion**: This stage involves collecting data from various sources. There are many nodes that collect data from different sources.
2. **Data processing**: In this stage, data is cleaned and transformed to make it suitable for analysis or machine learning tasks. We also can put many nodes to do it.
3. **Data output**: The final stage involves outputting data for further use. We can use the single node or multi nodes to do it.

Q11

2 / 2

LLMs such as GPT-4 are 'pretrained' (that is the 'P') with a large amount of language data. For more precise answers, they need to be 'fine tuned' with specific domain-related (eg medical, legal...) language, OR, be chained to a custom DB that can be accessed by the GPT. In this context, what is the 'OPL stack'? Describe in your own words, sticking to what we discussed in class (high level description is fine).

OPL stands for OpenAI, Pinecone, and Langchain.

1. OpenAI provides API access to powerful LLMs and also provides embedding models to convert text to embeddings.
2. Pinecone provides embedding vector storage, semantic similarity comparison, and fast retrieval.
3. Langchain allows users to build their own LLM applications, and can be used to fine-tune the models on domain-specific data or connect them to external databases.

Q12

2 / 2

AI techniques can be classified as being one of these four types: heuristic/brute-force, reinforcement learning (RL), rule-based, or connectionist ("ML"). Of these four, which is data-driven? Explain how data is used to impart intelligence to the machine? Briefly outline the steps involved.

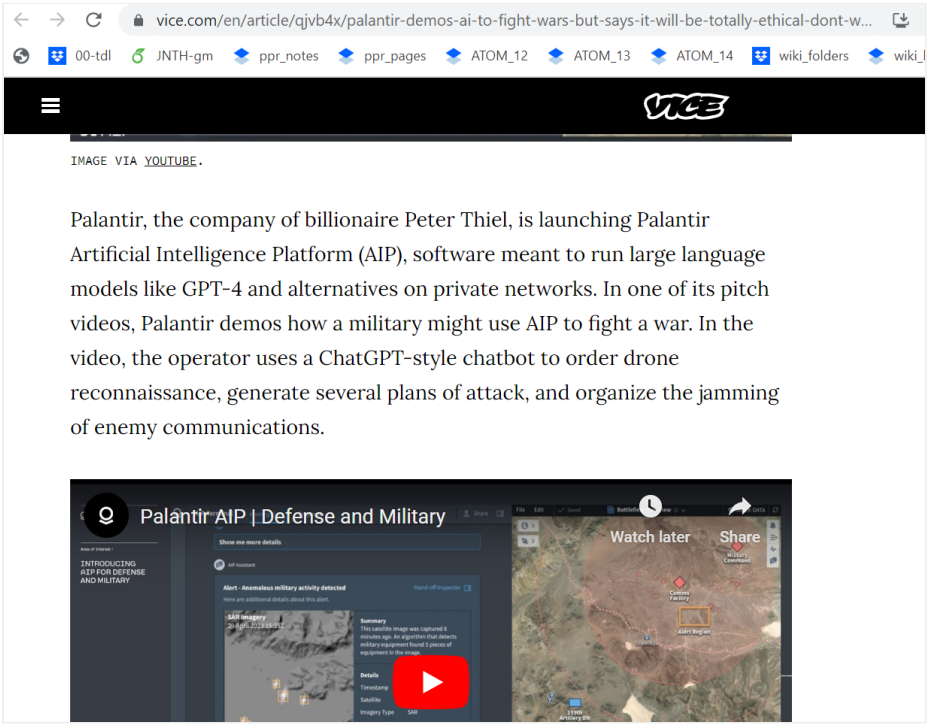
Connectionist. Steps:

1. **Data collection**: gather a dataset that represents the problem we need to solve.
2. **Data preprocessing**: clean data, handle missing values and so on.
3. **Feature engineering**: select or create the most relevant features or attributes from the data.
4. **Model training**: choose a framework (e.g. Pytorch, Tensorflow) to train our model on our dataset.
5. **Model evaluation**: use some metrics such as accuracy, precision, recall, or F1-score to evaluate our trained model on a validation set. So our data actually help us train the model and make it can make predictions or decisions.

Q13

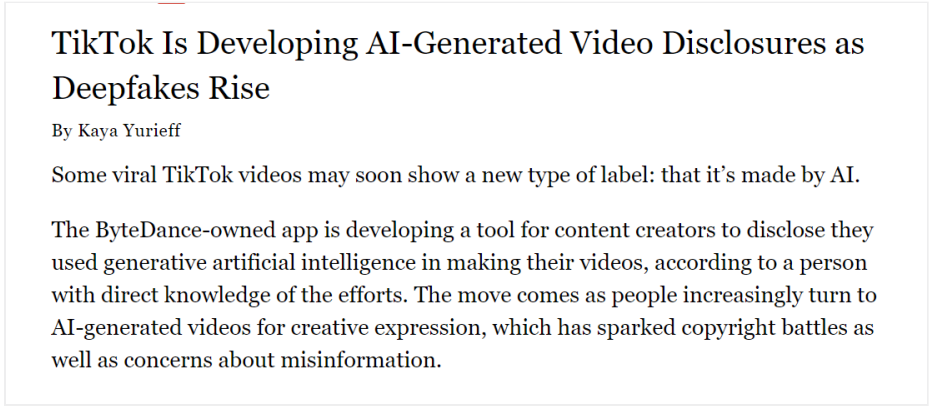
2 / 2

Palantir is an ML-oriented company that works with law enforcement, etc. Recently they announced this:



In terms of 'GPSETC', **what would be considered alarming** about this?

Also, here is another recent news item (related to this: data can be used to train a system to generate alternative data, ie. to create 'generative AI'):



TikTok Is Developing AI-Generated Video Disclosures as Deepfakes Rise

By Kaya Yurieff

Some viral TikTok videos may soon show a new type of label: that it's made by AI.

The ByteDance-owned app is developing a tool for content creators to disclose they used generative artificial intelligence in making their videos, according to a person with direct knowledge of the efforts. The move comes as people increasingly turn to AI-generated videos for creative expression, which has sparked copyright battles as well as concerns about misinformation.

What 'E'thical and 'S'ecurity issues do deepfakes pose? Be original in your answer!

---

1. Privacy(expose national security information), Security(Chatbots are vulnerable to hacking or cyberattacks), Compliance(violate international laws and norms)
2. **Ethical**: deepfakes raise questions about liability and copyright. Additionally, creating other people's videos without their consent can be socially harmful to others. **Security**: deepfakes can create videos of national leaders that can influence elections and even disrupt relations between countries.