Problem Definition
At the point of the midpoint checkin, our primary goal was to predict student depression based on a dataset we'd found on Kaggle. We thought this was ideal because it fit our common interest - neuroscience, and we could use classifiers that we learned in this class.

However, it had two problems: first being that another team had the same topic and dataset as we did, and second being that the dataset we got from Kaggle was not an accurate reflection of the real world. So we went on the search for a new topic that we would both be interested in.

Then we found out about the Visual Mandela Effect.

Mandela Effect
The Visual Mandela effect gets its name from the shared false account of many people remembering that Mandela died in prison. Whereas in fact, he became president after he was in prison and lived a long life till 2013. Yet, many people remember falsely remember that Nelson Mandela died in prison.

The Mandela effect is a term that explains this kind of shared false memory across many people. The visual Mandela effect is the Mandela effect on visual stimuli. A common example would be people remembering the monopoly man, who is the guy that's drawn on the cover of the monopoly board game, is wearing a monocle. He's not wearing anything, in fact. He's not wearing any glasses, I mean. Obviously, he's wearing something. He's wearing a suit- a tux.

VME
Let's see some more examples. There are six pictures here- on the top line, there are pikachus, and on the bottom, there are logos of the automobile company, Volkswagen.

Can you pick which is the right picture?

People commonly misremember Pikachu as having a black tip at its tail, and the Volkswagen logo as two v's intertwined. But you can see which is the correct picture.

It's easy to dismiss this phenomenon as a mistake, but the fact that many people commonly show this type of false memory was intriguing to us.

Our Goal
Which is when we encountered a paper that showed that there are certain icons or visual stimuli that induced the visual Mendela effect, which is where we got our dataset.

We wanted to know if a machine learning model can learn the difference between icons that do trigger this effect and those that don't. If so, we cannot only validate the cognitive hypothesis in a new way but also build tools to automatically flag risky images. This can lead to a breakthrough in studies about visual Mandela effect because until now most of the studies have relied on behavioral experiments to determine which icons induce the Visual Mandela Effect.

Technical Definition & Result
We used the public stimulus set. First we pre-processed the data. We divide the stimuli into two groups VME and non-VME. Then we built a pandas dataframe with three columns – path, category, and label. In category we stored what the images depict, stripped from the filename. We labeled VME-inducing images as 1, and non-VME as 0.

Then we applied torch vision transforms. For training we did randomly resized crops, color jitter and blur to diversify our tiny sample. For validation we did deterministic resizing and center crop to ensure consistent and reproducible input.

CLIP
We leverage OpenAI's CLIP—which aligns images and text in a shared embedding space.
We thought that since CLIP has seen millions of image, caption pairs, it would know objects like the Monopoly Man. We wished to see if CLIP could help us distinguish VME vs. non-VME.

Since we had a small dataset, we froze the CLIP's parameters, to preserve their general semantic knowledge. We added two layers: one for text and one for image, to adapt CLIP embedding for our task. Then we fed them to through a two-layer MLP with ReLU and dropout, which gave a single output - the probability that the image and text match.

Training
Because the data was small, we used 8-fold stratified cross validation to ensure each fold maintains the VME versus non-VME ratio. For each fold, we first trained the fusion head with binary cross entropy on the frozen CLIP embeddings for five epochs. Then we validated by computing AUC, F1 and accuracy. Lastly, we saved the best model weight by tracking the highest AUC.

This is the result - you can see that the training loss gets as low as 0.002.

Grad-CAM & Embedding Visualization

While reviewing the code we wrote we found that this still didn't give us a hint of why the visual Mandela effect happens. So we computed a heat map highlighting the exact region the model was "looking" when it made the VME vs. non-VME decision. We hooked into CLIP's ResNet-50, and recorded activations and gradients on a sample VME icon. For the pikachu, you can see that the model concentrated on the Pikachu's face.

PCA and t-SNE
After we fused the text and image embeddings, we plotted them on a 2-D plane.
To do this, we used PCA and t-SNE respectively.
PCA gives us a linear projection onto the two directions of greatest overall variance—an excellent first look at whether VME vs. non-VME separate along a global axis.

t-SNE preserves local neighborhood structure. Since it tends to pull apart small clusters, it makes subtle groupings (like tiny detail changes) much more visible.

Conclusion
The takeaways from our little experiment is that it was the first step to showing that shared false memories are not coincidental, but can be remodeled, analyzed and even predicted by ML models. By using CLIP, cross validation, PCA and TSNE, we weaved the topics covered in this class with preexisting models and lay the groundwork for automated discovery of any image prone to VME.

This can open many new doors since research on VME was primarily focused on behavioral experiments.

However, our study has many limitations as well.

Our primary regret is that our model and processing pipeline were not a tailored fit to handle visual Mandela effect. VME is about microscopic, detail-level mismatches but by using CLIP, we focused on the semantic level. This was because our understanding of the visual Mandela effect was short. From this experiment we learned that in the future, to classify random images as VME and non-VME, we should use a model which focuses on a more detailed, targeted part of the image.

Also our dataset was very small - about only 40 icon concepts. This was due to the lack of study on the Visual Mandela effect.