# Binary is unary

## adversarial attack-stable diffusion

(TECH+RESEARCH)

# Meet the Team!

Nandini Ramachandran

Neha Konduru

Kamala Sreepada

Snikitha Chelluri

Shreenidhi Ayinala

**We are all CS Majors @ UMD**
**Mentors: Shweta Bhardwaj & Mazda Moayeri**

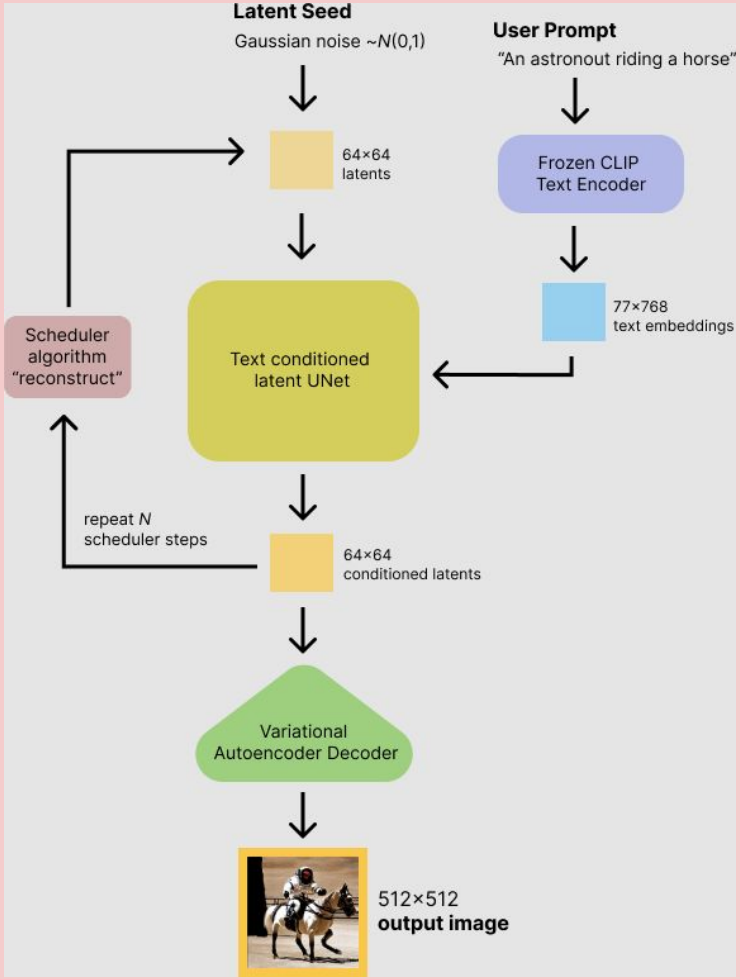
# ***Project Description***

***Stable Diffusion*** is a category of machine learning model that is trained **to denoise a random gaussian noise to get an image**. This model is based on a ***Latent Diffusion*** **model.**

- Autoencoder (VAE)
- U-Net
- Text-encoder

Language: Python
Framework: PyTorch

Research: *Determine what aspects the model takes into consideration when a prompt is inputted.*



# ***Model in Action***

The Stable Diffusion pipeline generates images from text - a new image is outputted every time code is run. It formats the outputs into a grid-like structure.

There are many factors that result in a biased ML model.

(1) Human factor: ML tries to mimic human behavior, and humans are often biased

(2) Poor quality of training data: If training data has a uneven distribution of a certain gender/race/etc, it can cause bias

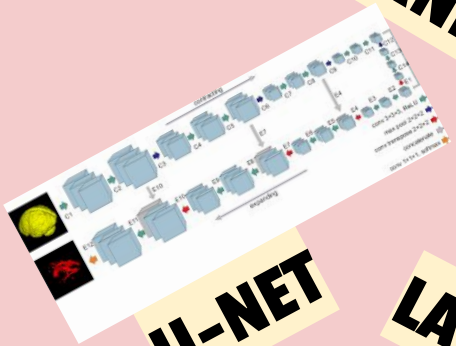
(3) Model performance mismatch: When training data does not match the test data

# *High-Level Overview of Key Terms*

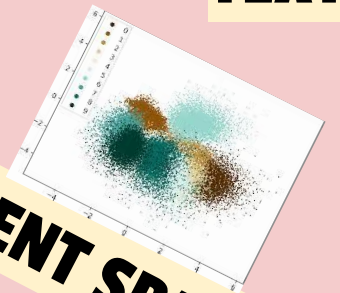
MACHINE LEARNING

MACHINE LEARNING MODEL

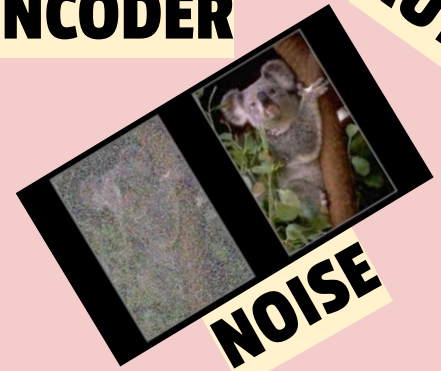

TEXT-ENCODER



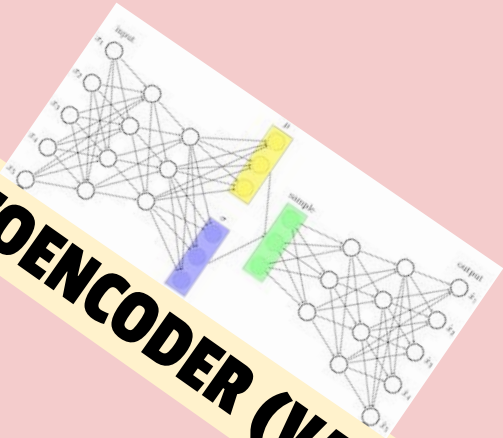
U-NET



LATENT SPACE

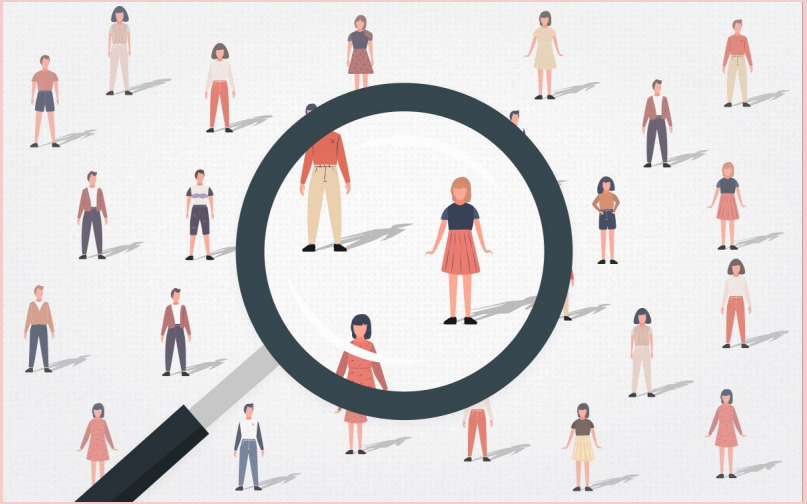


NOISE



AUTOENCODER (VAE)

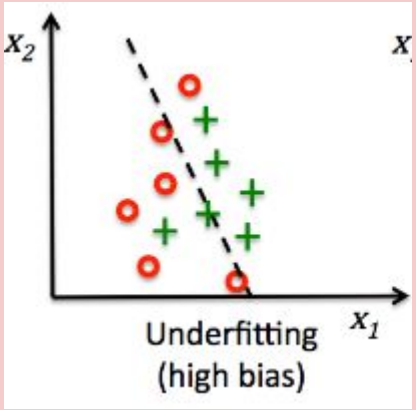
# ***Goals and Methodologies***

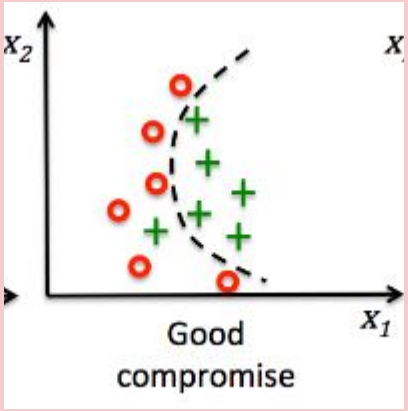


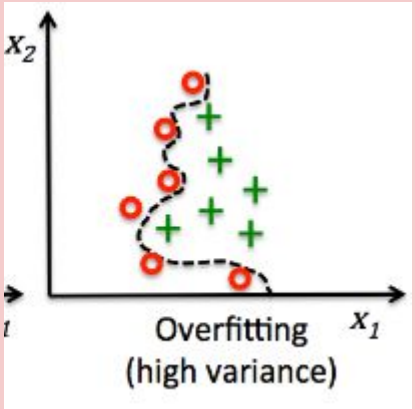
- Understanding bias in stable diffusion machine learning algorithms with respect to common human-related diversity factors such as:
  - Race
  - Gender
  - Facial Expressions
  - Color of Clothing
- Understand and interpret steps and decisions ML models take to make predictions for image classification

# What is Bias?

**A systematic error that occurs in the machine learning model itself due to incorrect assumptions in the ML process**



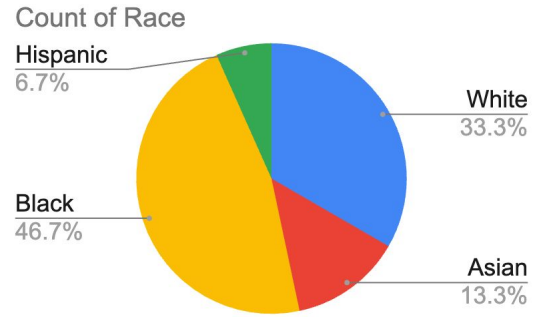



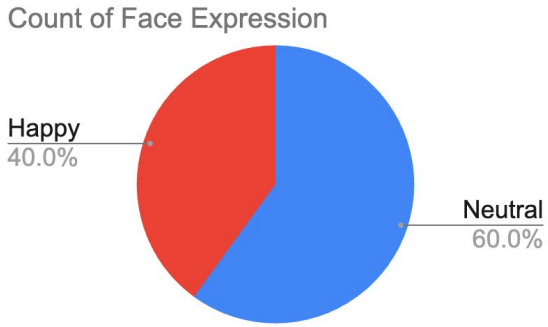



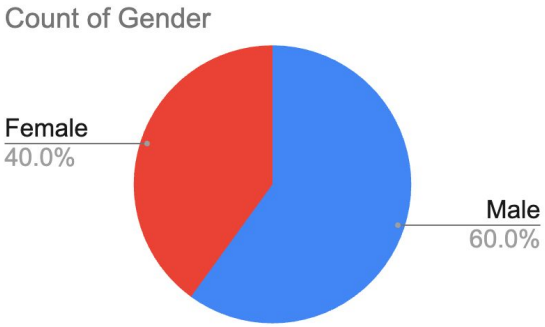
# ***Findings***

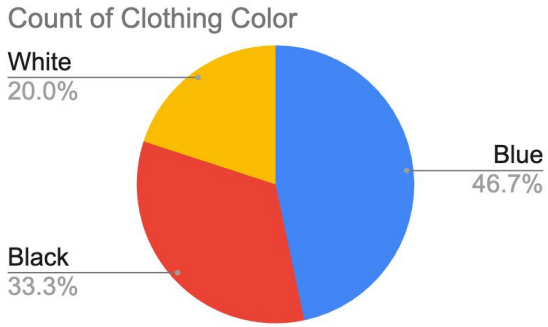
## Sample images of “a college student”

Key bias data:



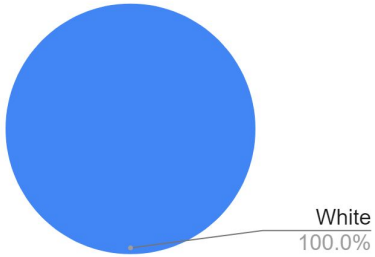




# ***Findings (cont.)***

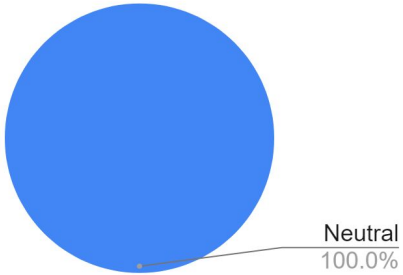
**Sample images of “a hackathon”**

Key bias data:

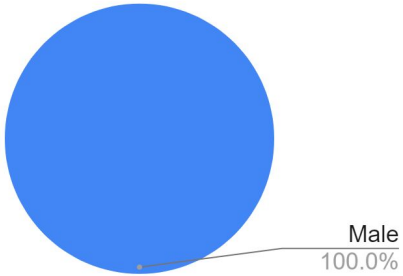
Count of Race



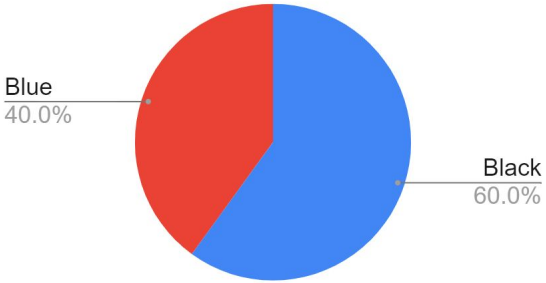

Count of Face Expression




Count of Gender



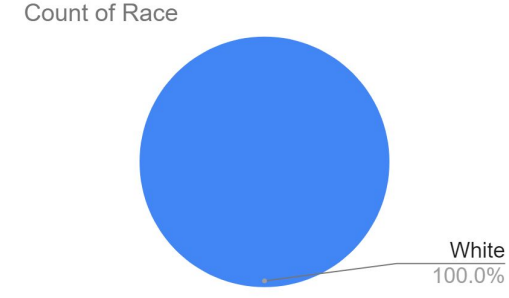

Count of Clothing color

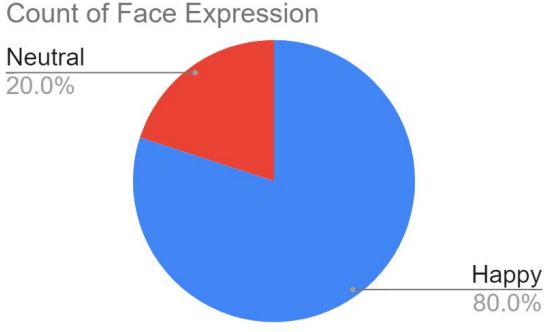


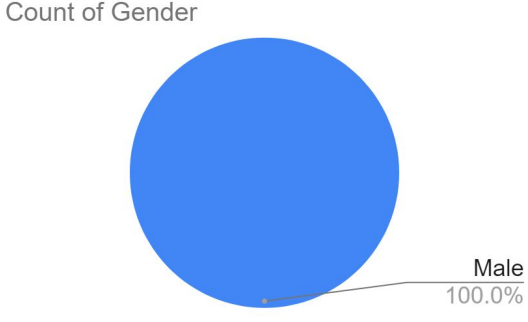

# ***Findings (cont.)***

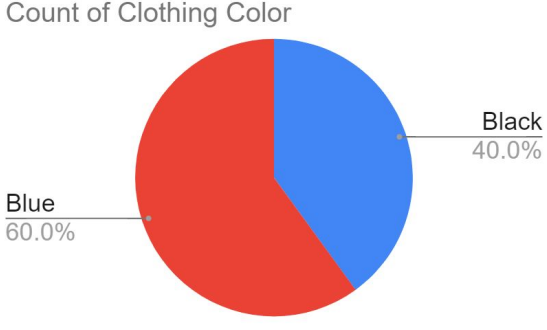
**Sample images of “a pilot flying a plane”**

Key bias data:



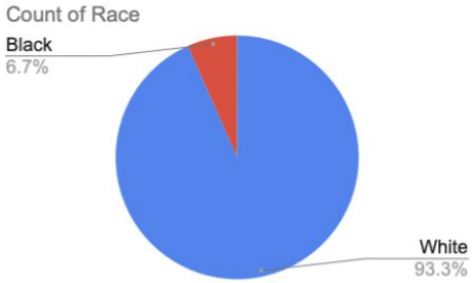



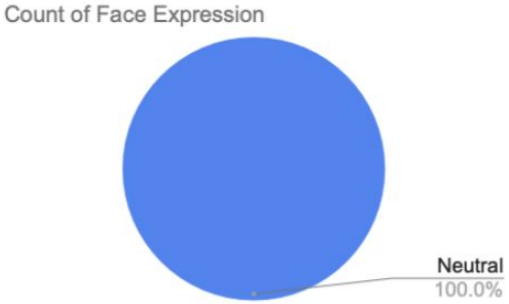



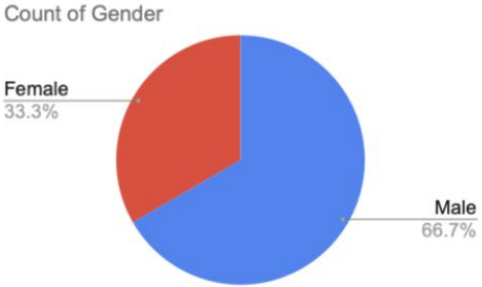

# ***Findings (cont.)***

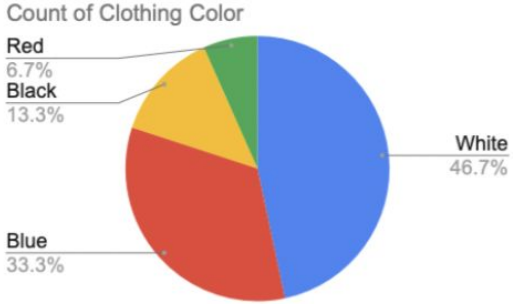
**Sample images of “a doctor performing surgery”**

Key bias data:



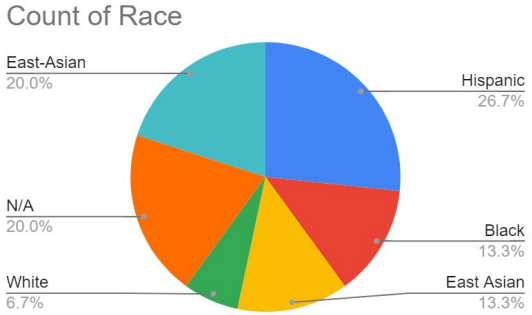



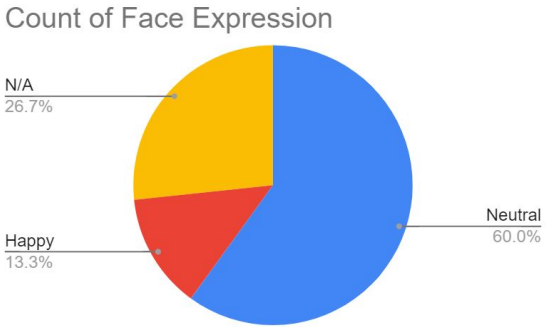



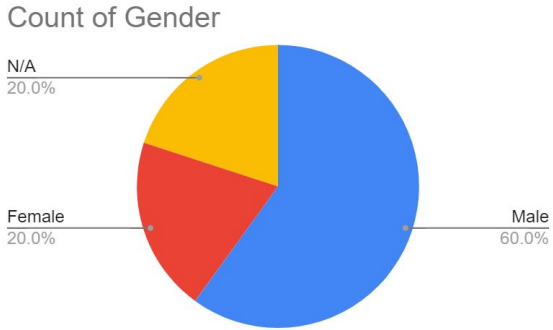

# *Findings (cont.)*

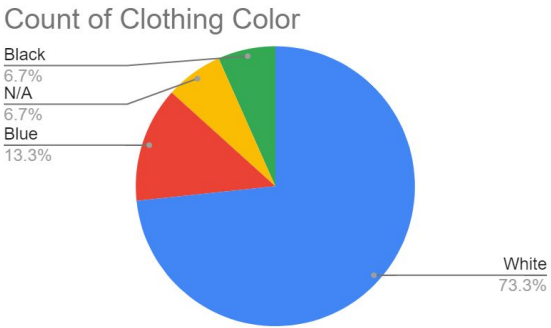
**Sample images of “a cook preparing dinner”**

Key bias data:




# How can we reduce bias?

1. Choose the correct learning model
2. Use the right training dataset
3. Perform data processing mindfully
4. Monitor real-world performance across the ML lifecycle
5. Make sure that there are no infrastructural issues

# Conclusions

- Biases are inherent in our machine learning models, and they alter the conclusions of the datasets based on these biases.

- Understanding the existence and impact of these biases is significant because they have real-world implications and the result would be systems that are untrustworthy and potentially harmful.

# What's next?

- Does the sentence structure inputted as the prompt change the way a model interprets it?

- Does the length and complexity of the sentence alter the way a model processes information?

# THANK YOU!