Applied Artificial Intelligence

Session 26: GANs, Fairness and Ethics in AI

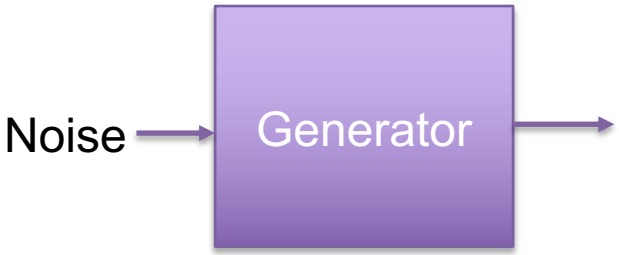
Fall 2018

NC State University
Lecturer: Dr. Behnam Kia

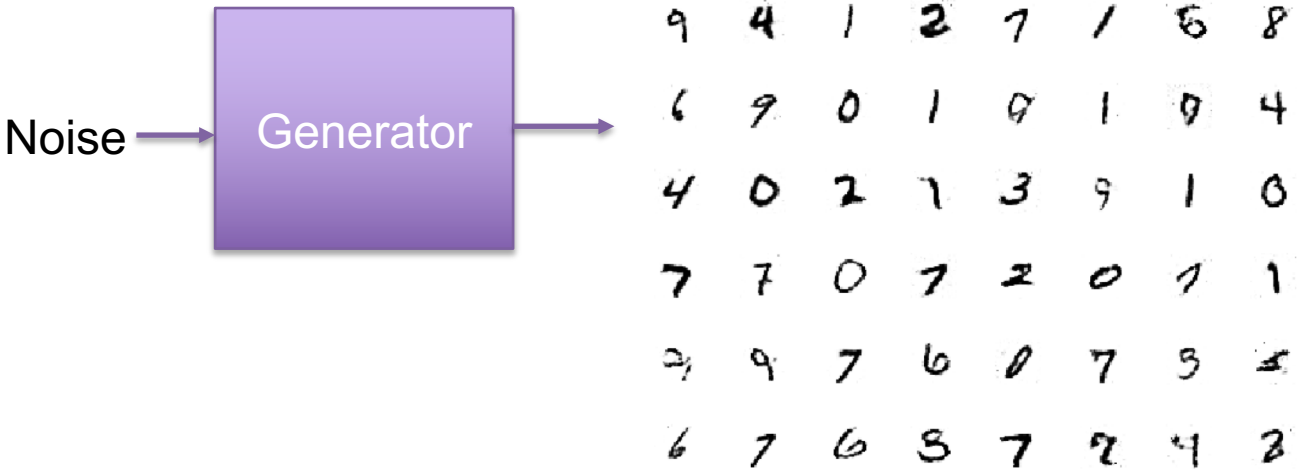
Course Website: https://appliedai.wordpress.ncsu.edu/

Goal: create realistic-looking synthetic data (images, voice,…)

Goal: create realistic-looking synthetic data (images, voice,…)

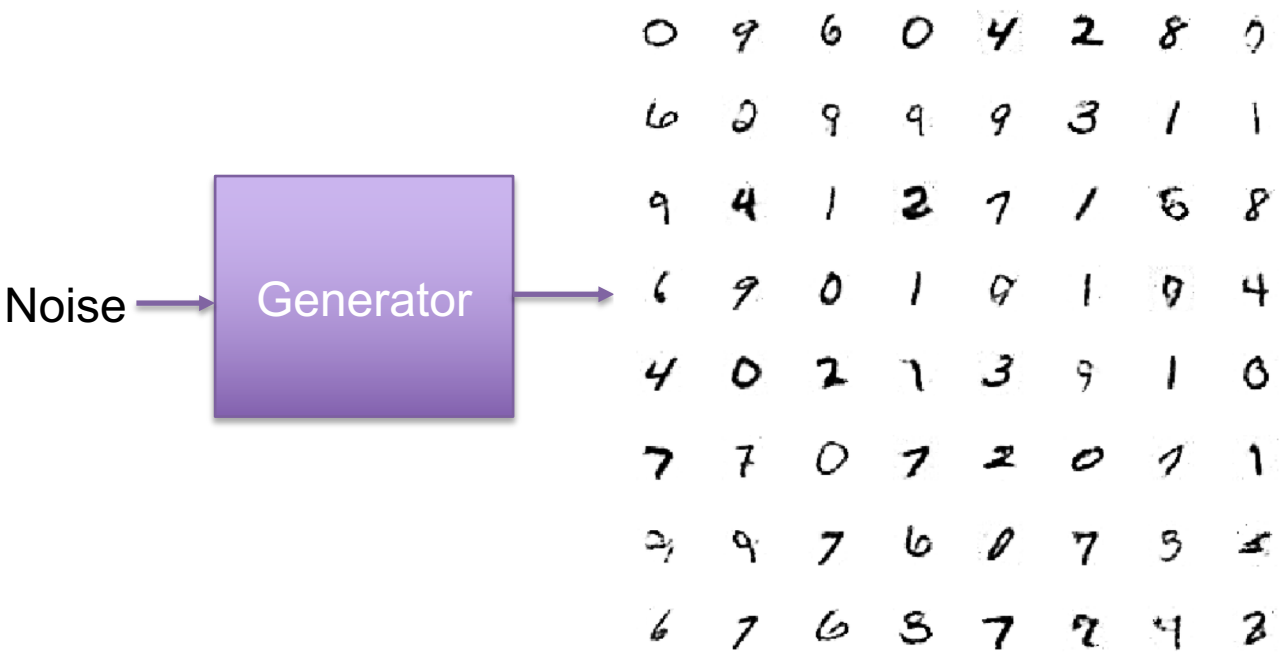


Goal: create realistic-looking synthetic data (images, voice,…)



But how?

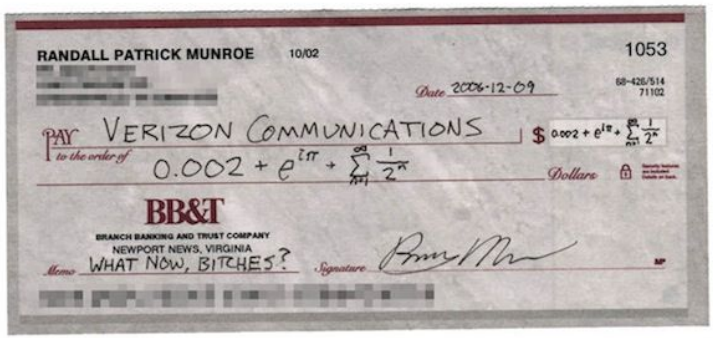
Goal: create realistic-looking synthetic data (images, voice,…)

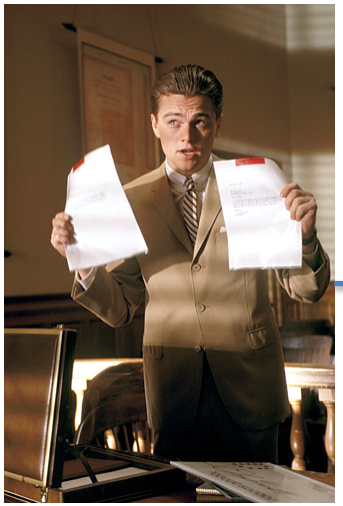


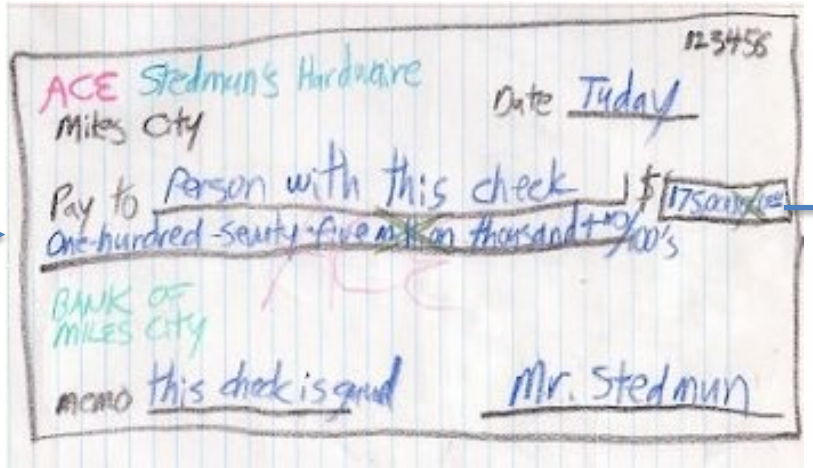
But how?

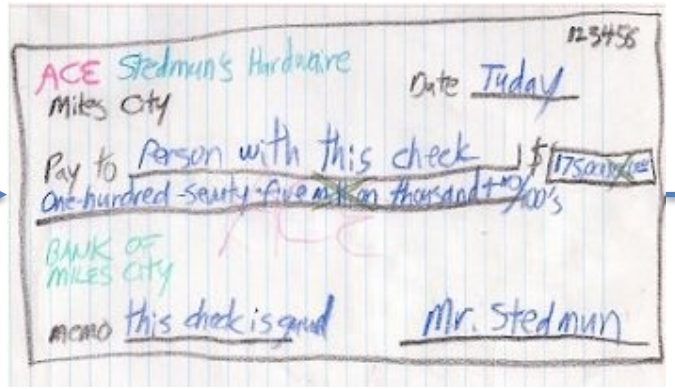
Adversarial approach. Generative NNs following this approach are called Generative Adversarial Networks, GANs

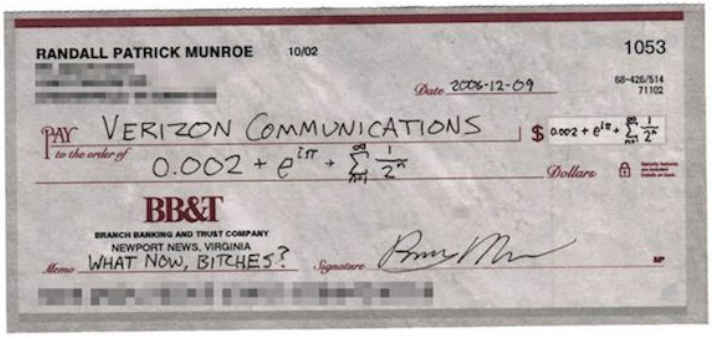
- As an example to adversarial approach to designing a generative of real-looking synthetic data (fake data), imagine counterfeiting checks.

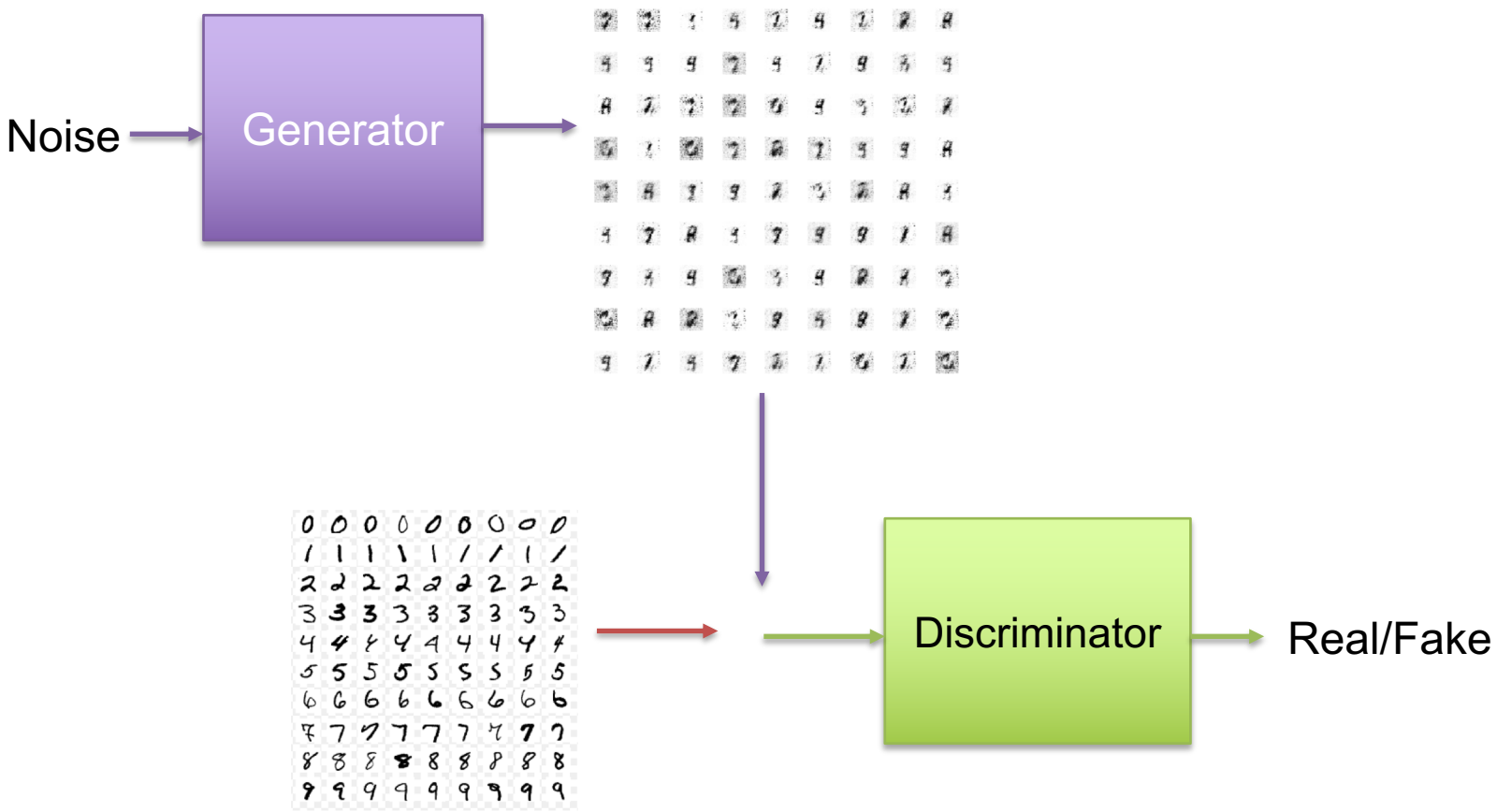


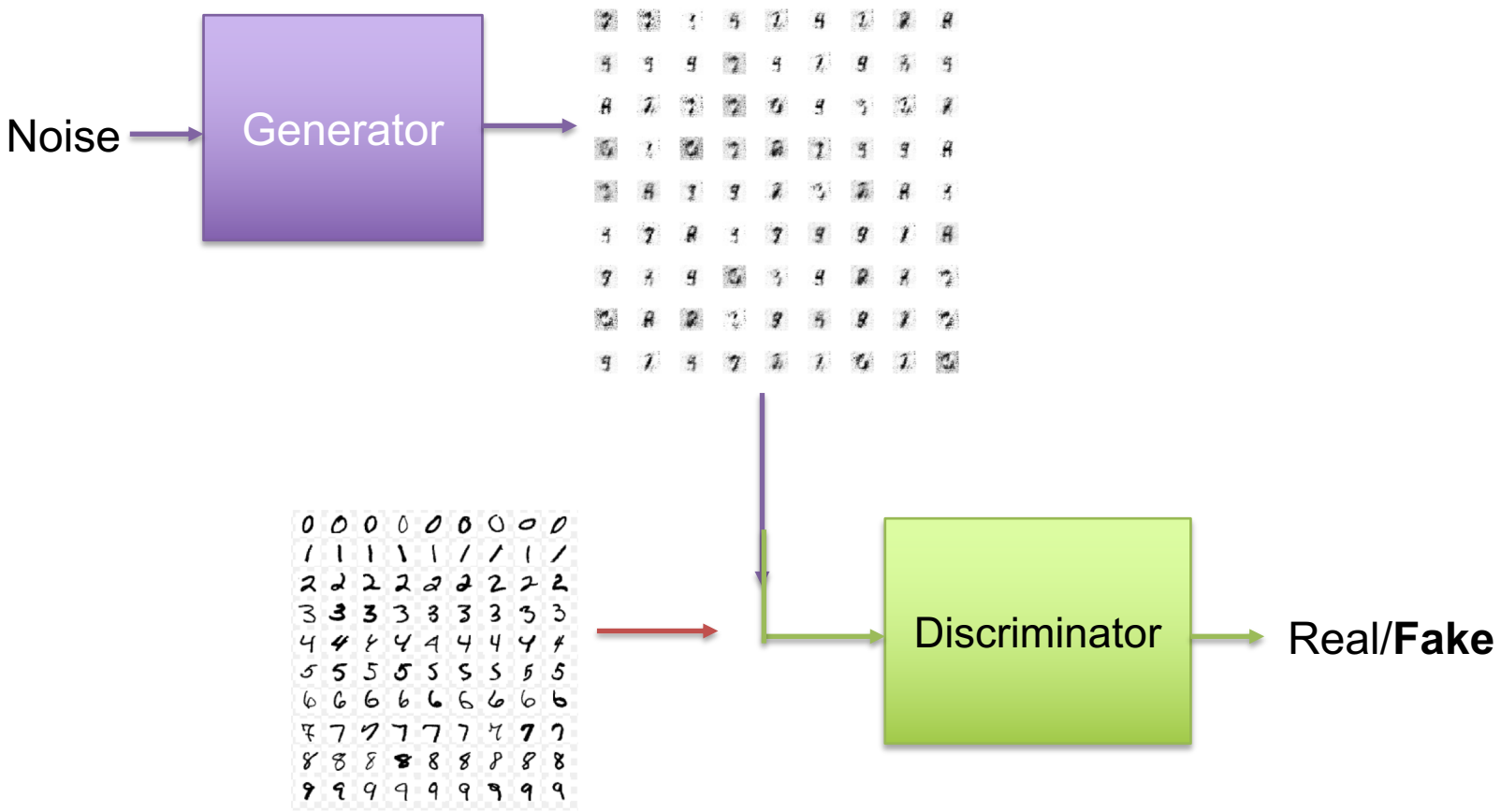


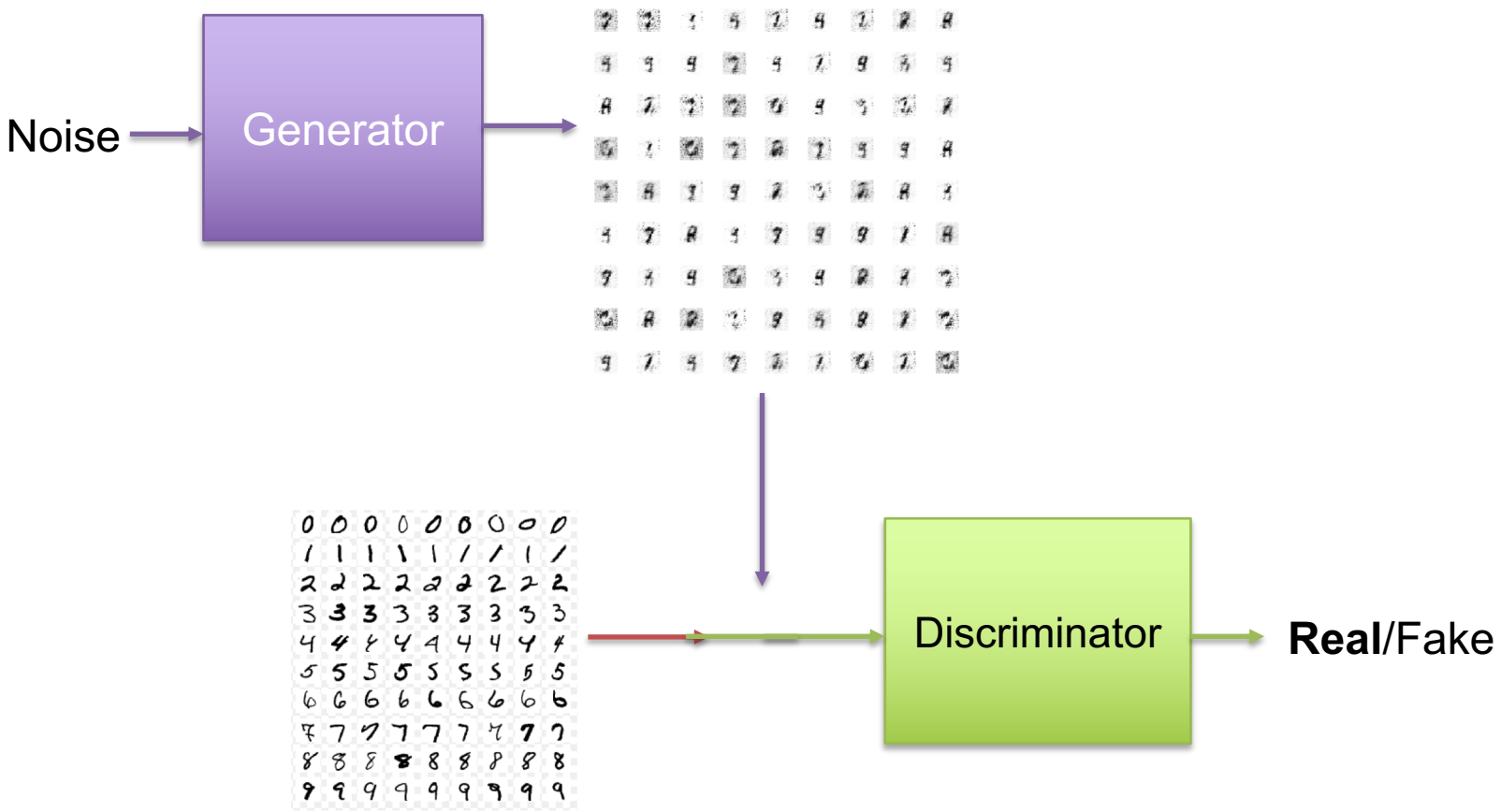

12345
ACE Stedmun's Hardware
Miles City
Date Tuday
Pay to Person with this check
One-hundred-seuty-five
BANK OF
MILES CITY
this check is good
Mr. Stedmun


12345
ACE Stedmun's Hardware
Miles City
Date Today
Pay to Person with this check $ 175
one-hundred-seuty-five
/100's
BANK OF
MILES CITY
memo this check is good
Mr. Stedmun


RANDALL PATRICK MUNROE
10/02
1053
Date 2006-12-09
PAY VERIZON COMMUNICATIONS
$ 0.002 + e^{iπ} + Σ 1/2^n
to the order of
0.002 + e^{iπ} + Σ 1/2^n
Dollars
BB&T
BRANCH BANKING AND TRUST COMPANY
NEWPORT NEWS, VIRGINIA
Memo WHAT NOW, BITCHES?
Signature

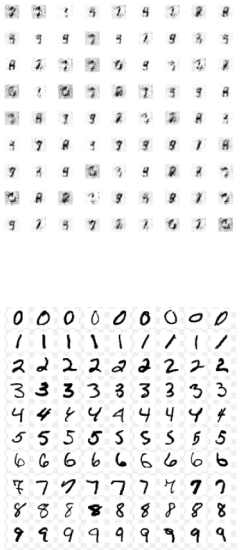

Noise
Generator
Discriminator
Real/Fake


Noise
Generator
Discriminator
Real/Fake


Noise
Generator
Discriminator
Real/Fake

Step 1: Train Discriminator

- Create fake images from the Generator
- Mix them with real images
- Train Discriminator to detect fake from real
- Now we have a trained discriminator



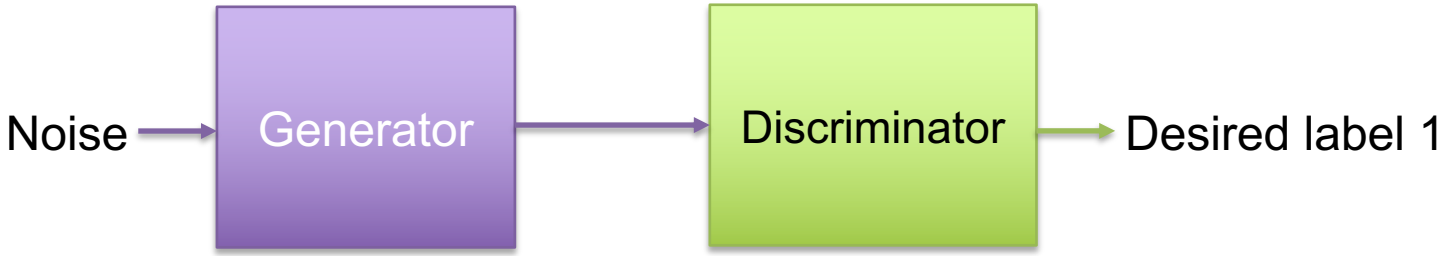
Real: 1
Fake: 0

Discriminator

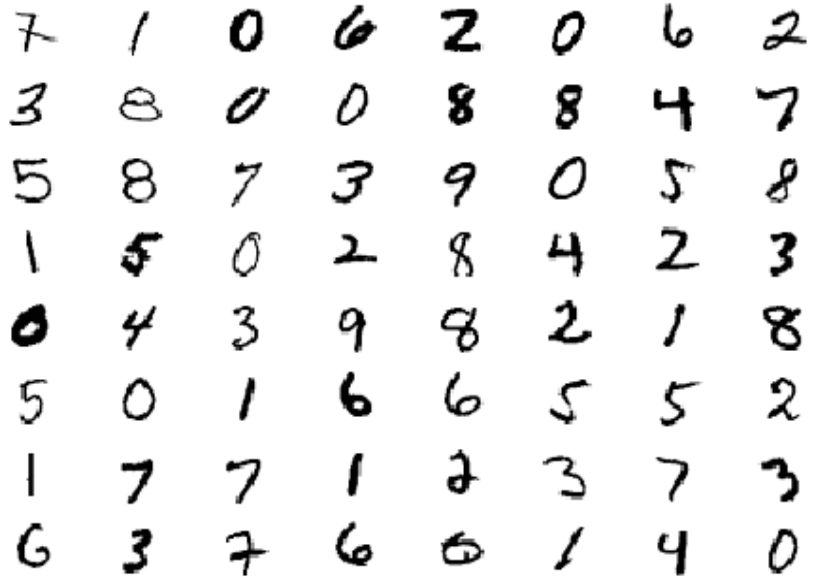
1/0

Step 2: Train Generator to fool discriminator

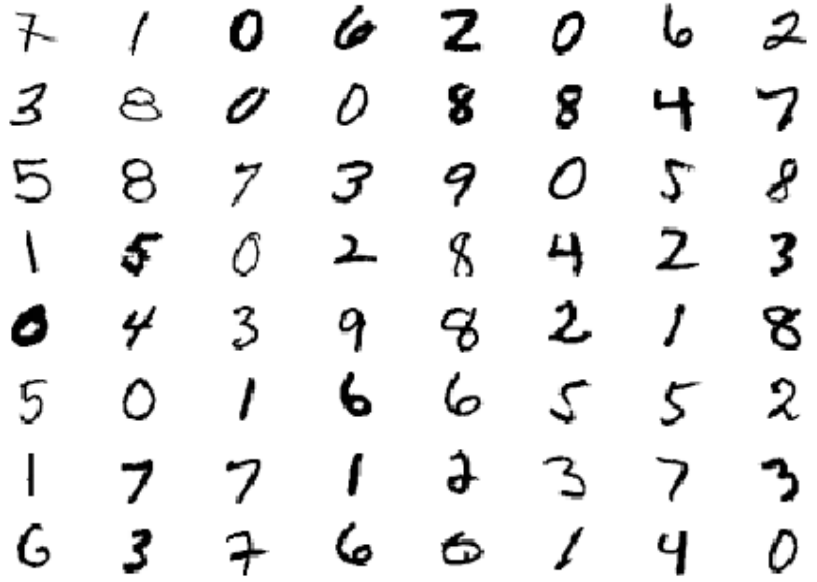
- Combine generator and discriminator as one model
- Freeze discriminator (no training for discriminator)
- Give fake images to discriminator and train the combined model to produce labels 1 (generator is under training)



- Go to Step 1 to retrain discriminator



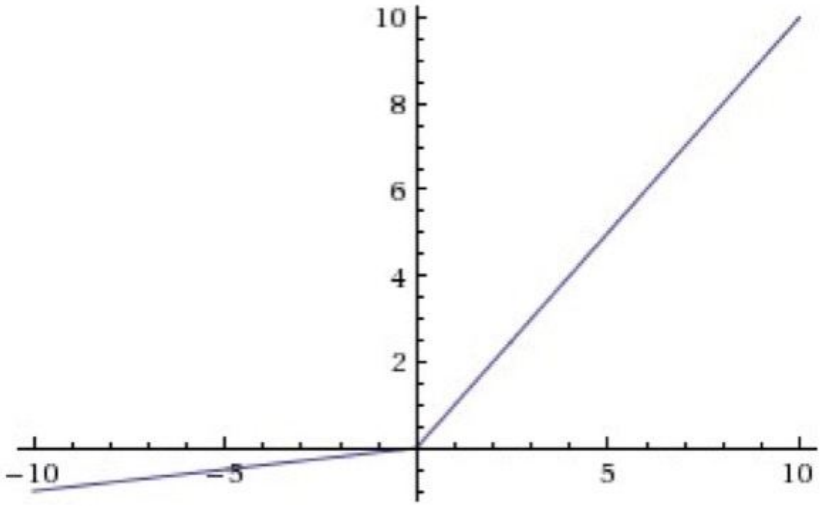
Fake



Real

Leaky Relu

$$f(x) = \begin{cases} x & if\ x > 0 \\ 0.01x & otherwise \end{cases}$$



- Please see codes for our first example of a GAN.

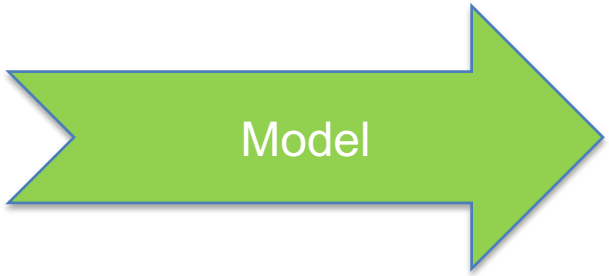
Homework
Deadline: Next Tuesday, 3PM

- Use an RNN network to classify MNIST dataset.
- Each image is 28*28. Consider each column (28 pixels) as one input. The sequence to RNN would be 28 columns of 28 pixels, each column given at one time step.
- Try simpleRNN, LSTM, and GRU.
- Try multiple stacks of those units.
- Use dropout to fight against overfitting (if needed).
- Accuracy of ~98% can be easily obtained.

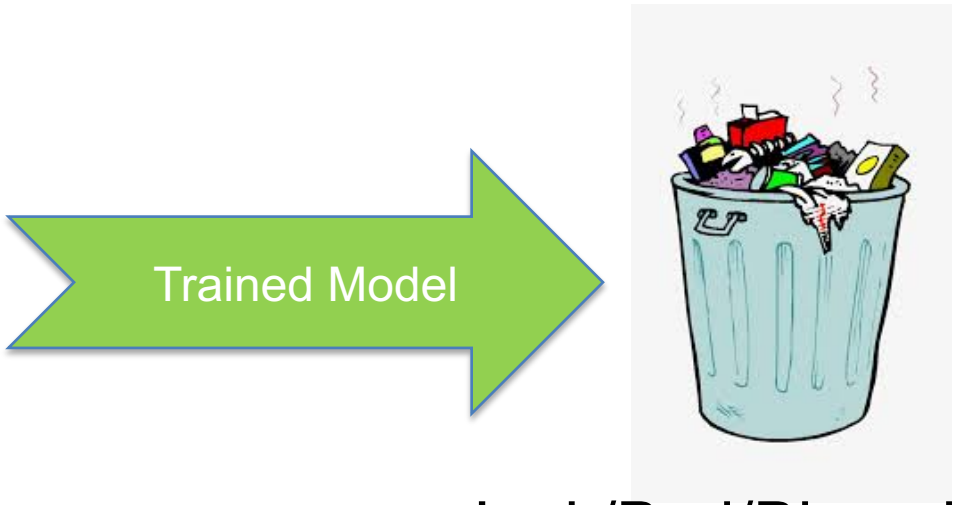
AI : Ethics and Fairness

Biased/Bad Data Results in Bad AI

Junk/Bad/Biased Data for training AI



Biased/Bad Data Results in Bad AI



Junk/Bad/Biased Results

Prof. Donna Strickland

The first woman to win Nobel Prize in Physics in 55 years.

If AI wanted to select the winner, and gender was among the features of the candidates, historical data would be biased against selection of female physicists.

Fortunately we don't use AI to select the next winner of Nobel price, but in other areas such as job websites, we are observing that AI is starting to be used to filter CVs of the applicants and to match them to potential employers. Please see the next slide.

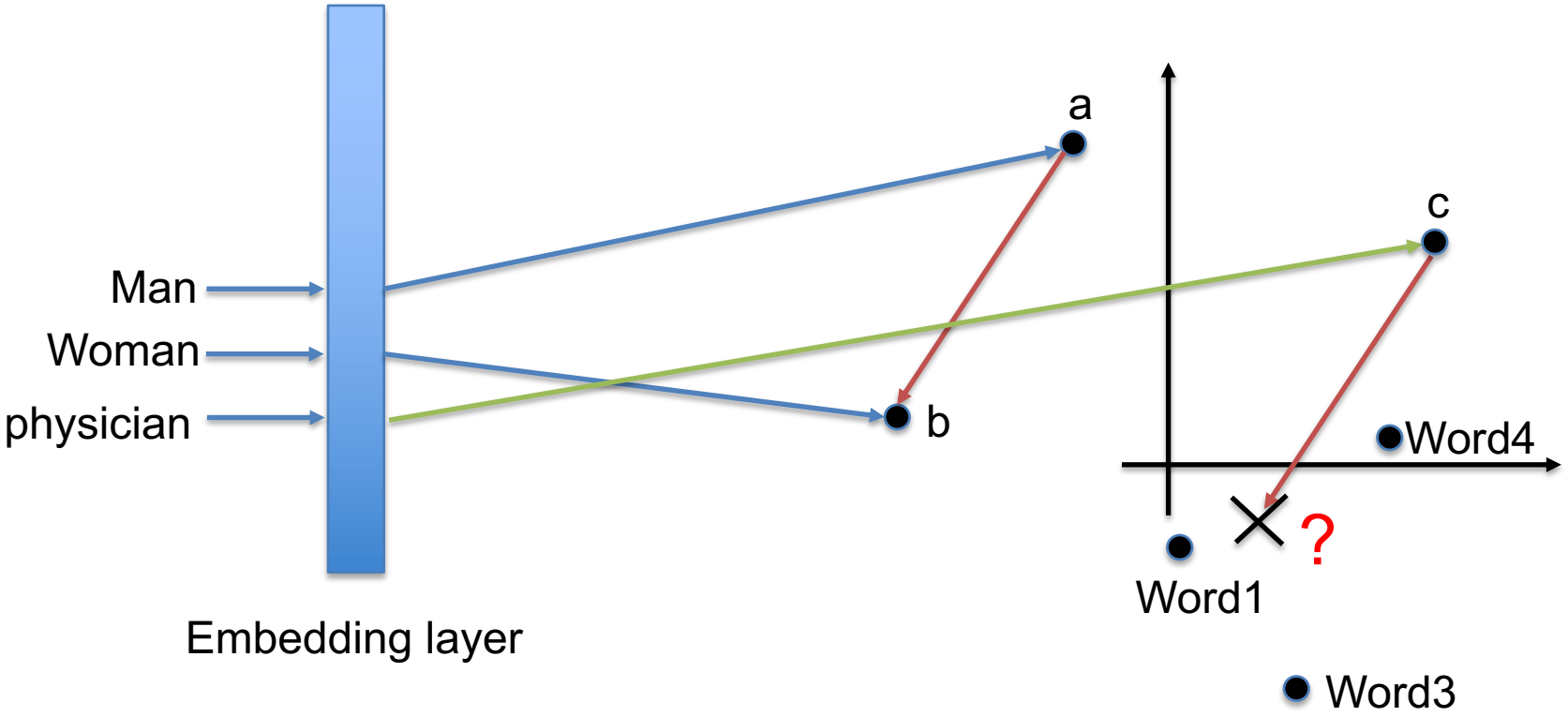
Red flags to possible bias in data

- Does your use case or product specifically use any of the following data: biometrics, race, skin color, religion, sexual orientation, socioeconomic status, income, country, location, health, language, or dialect?
- Does your use case or product use data that is likely to be highly correlated with any of the personal characteristics listed above (for example, zip code or other geospatial data is often correlated with socioeconomic status and/or income; image/video data can reveal information about race, gender, and age)?
- Could your use case or product negatively impact individuals' economic or other important life opportunities?

From: https://cloud.google.com/inclusive-ml/

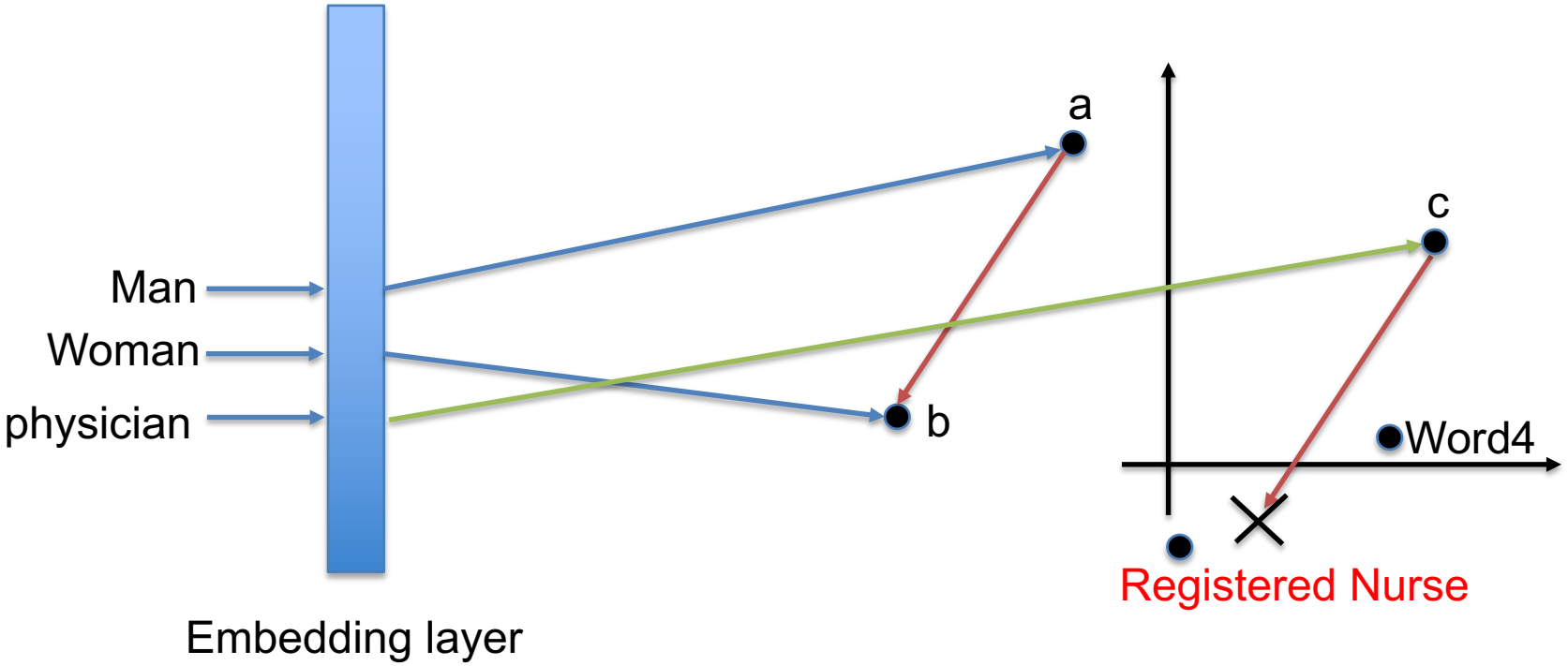
In some Publicly available and used Embeddings…

- Remember the previous homework. The goal was to use embeddings to determine given words a and b, find word d that word c and d has the same relationship as a and b. Here the relationship is the gender shift. So what is the female version of a physician?



In some Publicly available and used Embedding…

- Some embeddings say d is *"Nurse*!" Please see the next slide for more info on this.
- Yeap, we have a misogynist AI!



Man is to Computer Programmer as Woman is to Homemaker? Debiasing Word Embeddings

Tolga Bolukbasi[1], **Kai-Wei Chang**[2], **James Zou**[2], **Venkatesh Saligrama**[1,2], **Adam Kalai**[2]
[1]Boston University, 8 Saint Mary's Street, Boston, MA
[2]Microsoft Research New England, 1 Memorial Drive, Cambridge, MA
tolgab@bu.edu, kw@kwchang.net, jamesyzou@gmail.com, srv@bu.edu, adam.kalai@microsoft.com

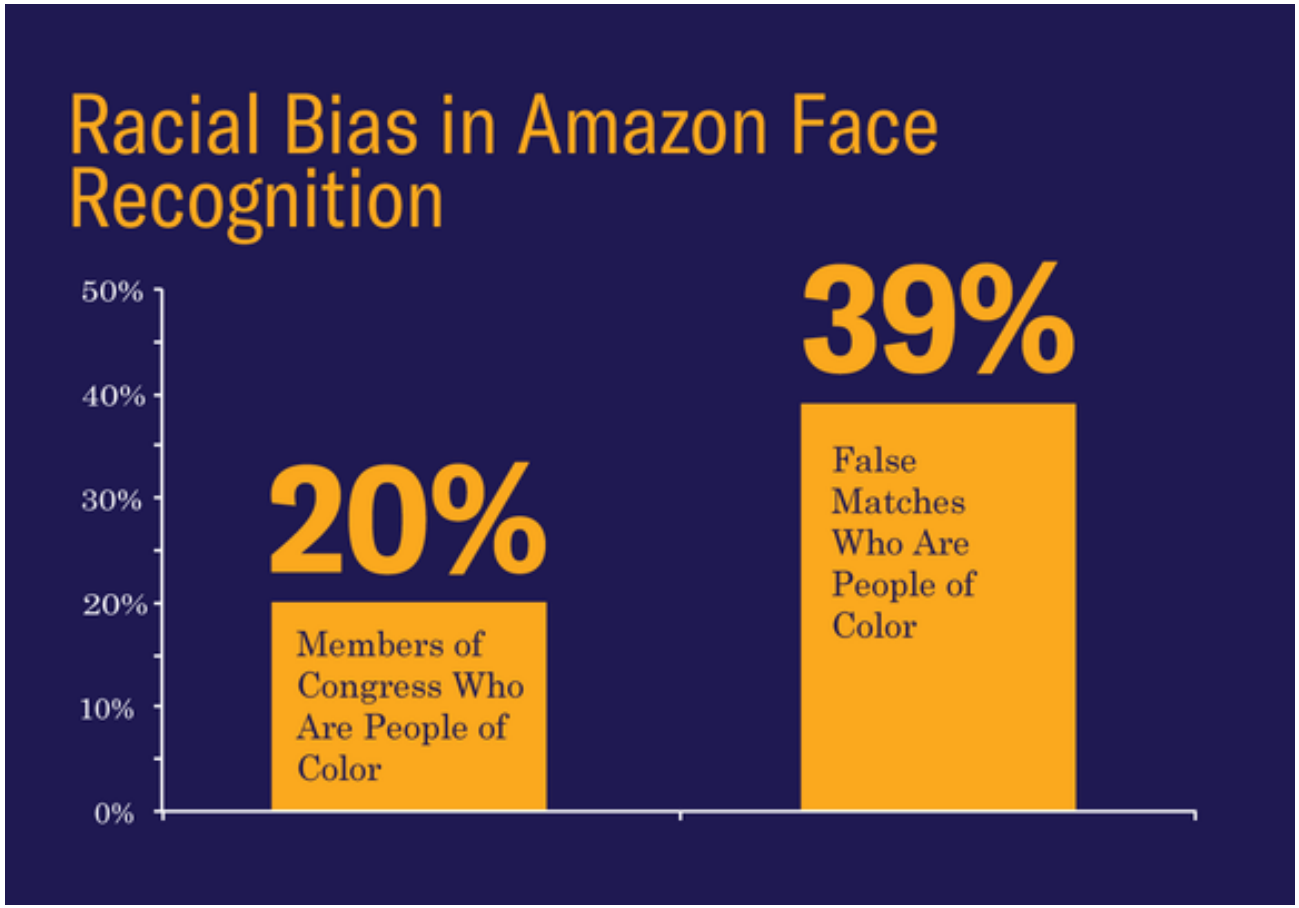
Abstract

The blind application of machine learning runs the risk of amplifying biases present in data. Such a danger is facing us with *word embedding*, a popular framework to represent text data as vectors which has been used in many machine learning and natural language processing tasks. We show that even word embeddings trained on Google News articles exhibit female/male gender stereotypes to a disturbing extent. This raises concerns because their widespread use, as we describe, often tends to amplify these biases. Geometrically, gender bias is first shown to be captured by a direction in the word embedding. Second, gender neutral words are shown to be linearly separable from gender definition words in the word embedding. Using these properties, we provide a methodology for modifying an embedding to remove gender stereotypes, such as the association between the words *receptionist* and *female*, while maintaining desired associations such as between the words *queen* and *female*. Using crowd-worker evaluation as well as standard benchmarks, we empirically demonstrate that our algorithms significantly reduce gender bias in embeddings while preserving the its useful properties such as the ability to cluster related concepts and to solve analogy tasks. The resulting embeddings can be used in applications without amplifying gender bias.

Bolukbasi, Tolga, et al. "Man is to computer programmer as woman is to homemaker? debiasing word embeddings." *Advances in Neural Information Processing Systems*. 2016.

ACLU ran an experiment, in which they used a database of 25,000 arrest photos, and compared them against the photos of members of congress using Amazon's facial recognition software.

https://www.aclu.org/blog/privacy-technology/surveillance-technologies/amazons-face-recognition-falsely-matched-28



Studies show that in facial recognition systems, the false matches are disproportionally of people of color. In ACLU's study, 39% of false matches are people of color, but just 20% of members of congress are people of color. This problem can have real-life, unfair consequences for people of color.

https://www.aclu.org/blog/privacy-technology/surveillance-technologies/amazons-face-recognition-falsely-matched-28

Auto censored/flagged words and comments in online forums and comment areas

- Sentiment Analyzers might think words such as “Gay” is negative, and therefore comments or documents containing those words can be flagged or censored, whereas words such as straight remain safe.

- Please read:

https://motherboard.vice.com/en_us/article/j5jmj8/google-artificial-intelligence-bias