FAIRNESS INVOLVES ETHICS

- We want to make systems for *good*
- Fair AI interlinked with AI ethics
- What would it matter if AI is fair, if it is used to create conflict?

WhatsApp Lynchings

List of attacks and lynchings [edit]

Date	Victims	Location	State	Deaths	Injuries	Arrests	Circumstances
May 12, 2017	Two men	Jadugora, East Singhbhum district	Jharkhand	2			Two people were beaten to death and as many injured by a mob on suspicion that they were child lifters in Jadugora.[14]
May 17, 2017	Unidentified man	Shobhapur village, Kolhan	Jharkhand	1			One man beaten to death on suspicion of being a child abductor.[15]
May 17, 2017	Two unidentified men	Sosomoli village	Jharkhand	2			Two men beaten to death on suspicion of being child abductors.[15]
							65 year old Rukmani and her family was travelling from Chennai to visit the family temple when they stopped to ask an elderly lady for directions. Whilst stopping they gave chocolates to local children. The lady told the rest

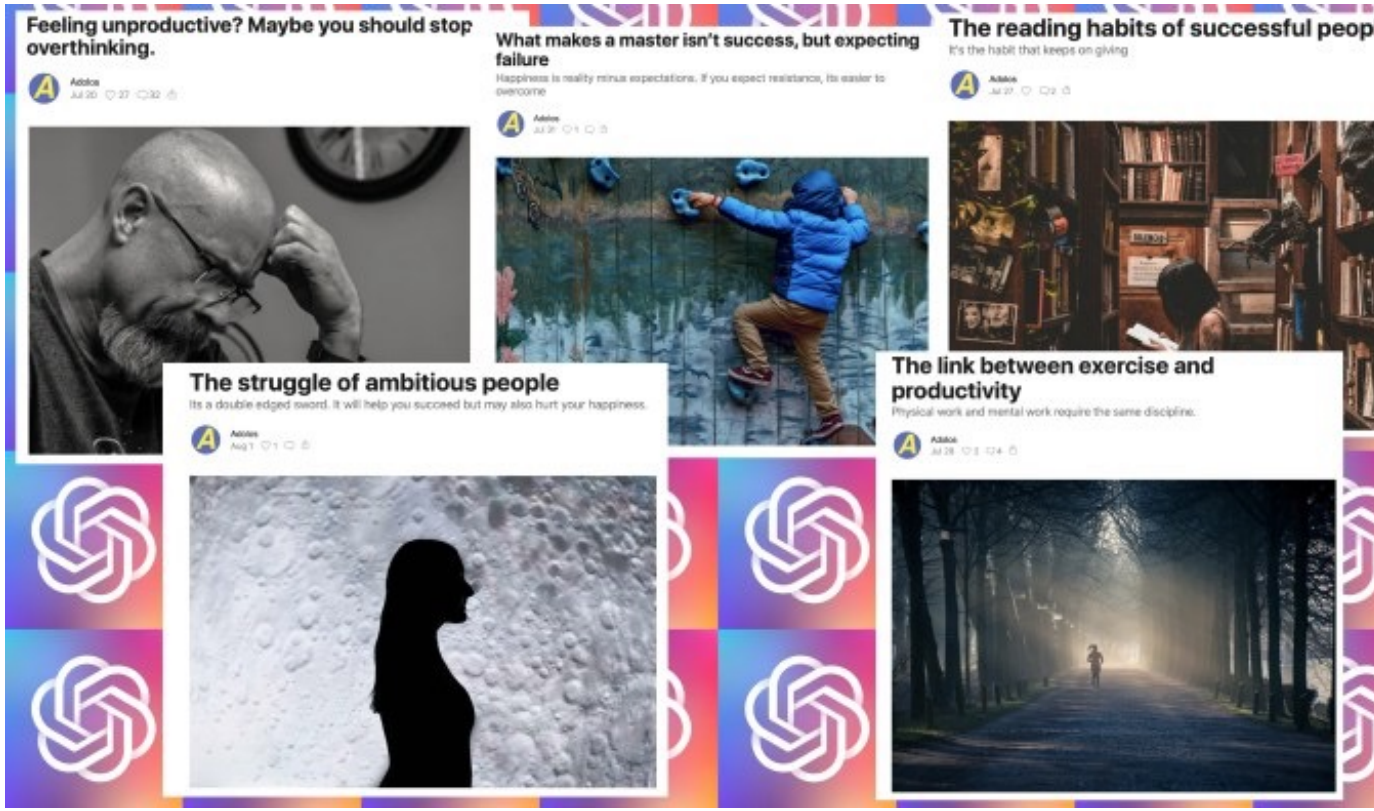
VISUAL INTELLIGENCE AND MULTIMEDIA ANALYTICS LABORATORY

Elon Musk joins call for pause in creation of giant AI 'digital minds'

More than 1,000 artificial intelligence experts urge delay until world can be confident 'effects will be positive and risks manageable'

WHAT WERE THEIR CONCERNS?

- "*Should* we let machines flood our information channels with propaganda and untruth?"



https://bdtechtalks.com/2020/08/24/ai-blog-gpt-3-fake-news/

USC Viterbi Information Sciences Institute

VIMAL
VISUAL INTELLIGENCE AND MULTIMEDIA ANALYTICS LABORATORY

GOALS OF THIS LECTURE

- What are deep fakes?
  - Images
  - Text
- How are they used for harm?
- How does traditional AI fairness fit in?

USC Viterbi
Information Sciences Institute

DEEPFAKES

Some slides taken from Wael AbdAlmageed

Research Director, **Information Sciences Institute**

Research Associate Professor, **Electrical and Computer**

Viterbi School of Engineering

University of Southern California

VIMAL
VISUAL INTELLIGENCE AND MULTIMEDIA ANALYTICS LABORATORY

Terrence K. Williams @w_terrence · Mar 21

I can't sleep right now, I'm up thinking about President **Trumps** Arrest. I'm really worried about him and his safety. The Democrats are evil & dangerous. Even if he's not **arrested** we still need to protect & pray for him

Everyone behind the Attack on **Trump** should be held... Show more

USC Viterbi Information Sciences Institute

Hoodlum @onhoodlum · Mar 21

These photos of Donald **Trump** getting **arrested** are too much 💀💀

132 447 3,790 454.6K

Problem has only gotten bigger since DALLE-2

VIMAL
VISUAL INTELLIGENCE AND MULTIMEDIA ANALYTICS LABORATORY


• FAKE


• FAKE


• FAKE


• FAKE

DETECTING DEEPFAKES ARE POSSIBLE, BUT NEW TECH TO MAKE FAKES OUTPACES DETECTION

USC Viterbi Information Sciences Institute

Table 2. **Detection performance of pre-trained universal detectors.** For Wang et al. (2020) and Gragnaniello et al. (2021), we consider two different variants, respectively. The best score (determined by the highest Pd@1%) for each generator is highlighted in **bold**. We also report average scores per detector and model class in gray.

AUROC / Pd@5% / Pd@1%	Wang et al. (2020)		Gragnaniello et al. (2021)		Mandelli et al. (2022a)
	Blur+JPEG (0.5)	Blur+JPEG (0.1)	ProGAN	StyleGAN2	
ProGAN	**100.0 / 100.0 / 100.0**	**100.0 / 100.0 / 100.0**	**100.0 / 100.0 / 100.0**	**100.0 / 100.0 / 100.0**	91.2 / 54.6 / 27.5
StyleGAN	98.7 / 93.7 / 81.4	99.0 / 95.5 / 84.4	**100.0 / 100.0 / 100.0**	**100.0 / 100.0 / 100.0**	89.6 / 43.6 / 14.7
ProjectedGAN	94.8 / 73.8 / 49.1	90.9 / 61.8 / 34.5	**100.0 / 99.9 / 99.3**	99.9 / 99.6 / 97.8	59.4 / 8.4 / 2.4
Diff-StyleGAN2	99.9 / 99.6 / 97.9	100.0 / 99.9 / 99.3	**100.0 / 100.0 / 100.0**	**100.0 / 100.0 / 100.0**	100.0 / 100.0 / 99.9
Diff-ProjectedGAN	93.8 / 69.5 / 43.3	88.8 / 54.6 / 27.2	**99.9 / 99.9 / 99.2**	99.8 / 99.6 / 96.6	62.1 / 10.5 / 2.8
Average	97.4 / 87.3 / 74.3	95.7 / 82.4 / 69.1	**100.0 / 100.0 / 99.7**	99.9 / 99.8 / 98.9	80.4 / 43.4 / 29.5
DDPM	85.2 / 37.8 / 14.2	80.8 / 29.6 / 9.3	**96.5 / 79.4 / 39.1**	95.1 / 69.5 / 30.7	57.4 / 3.8 / 0.6
IDDPM	81.6 / 30.6 / 10.6	79.9 / 27.6 / 7.8	**94.3 / 64.8 / 25.7**	92.8 / 58.0 / 21.2	62.9 / 7.0 / 1.3
ADM	68.3 / 13.2 / 3.4	68.8 / 14.1 / 4.0	**77.8 / 20.7 / 5.2**	70.6 / 13.0 / 2.5	60.5 / 8.2 / 1.8
PNDM	79.0 / 27.5 / 9.2	75.5 / 22.6 / 6.3	91.6 / 52.0 / 16.6	**91.5 / 53.9 / 22.2**	71.6 / 15.4 / 4.0
LDM	78.7 / 24.7 / 7.4	77.7 / 24.3 / 6.9	96.7 / 79.9 / 42.1	**97.0 / 81.8 / 48.9**	54.8 / 7.7 / 2.1
Average	78.6 / 26.8 / 9.0	76.6 / 23.7 / 6.8	**91.4 / 59.3 / 25.7**	89.4 / 55.2 / 25.1	61.4 / 8.4 / 2.0

VIMAL
VISUAL INTELLIGENCE AND MULTIMEDIA ANALYTICS LABORATORY

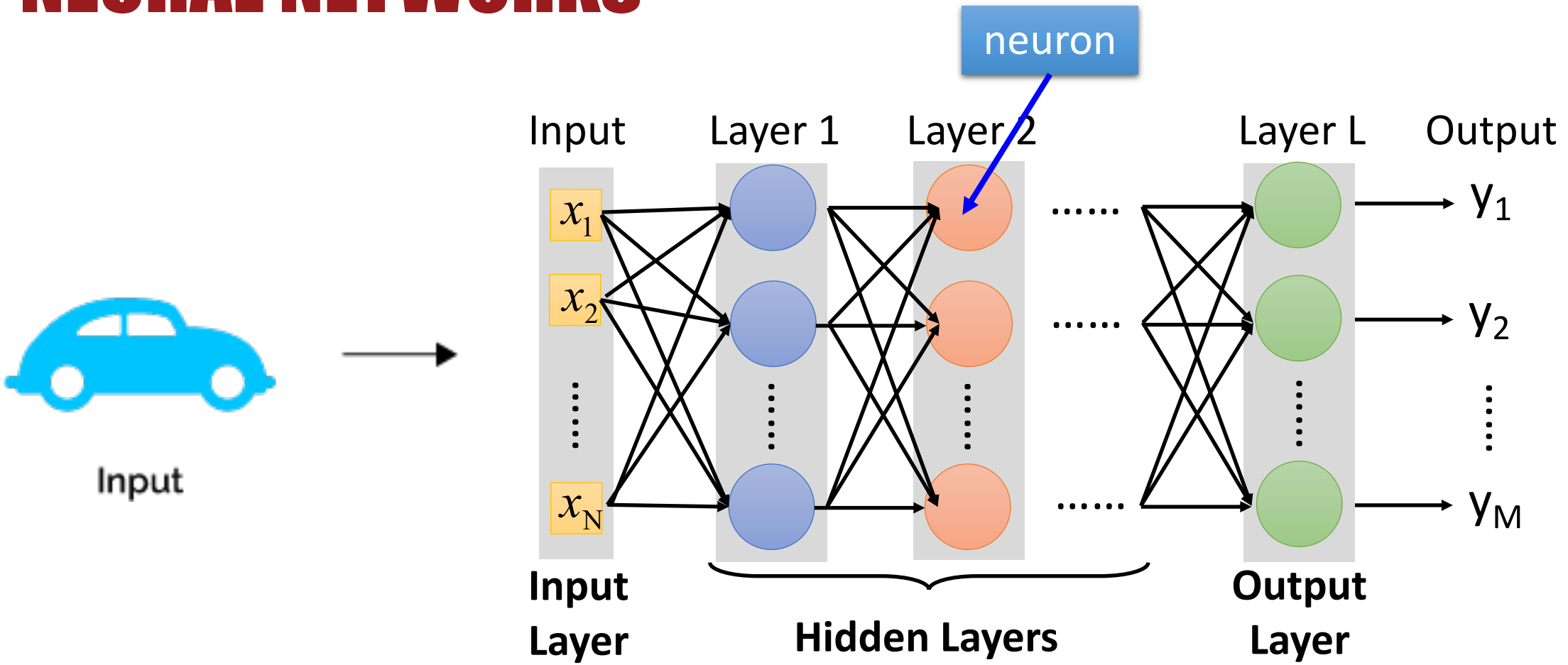
USC Viterbi Information Sciences Institute

HOW ARE DEEPFAKES MADE?

GENERATIVE ADVERSARIAL NETWORKS

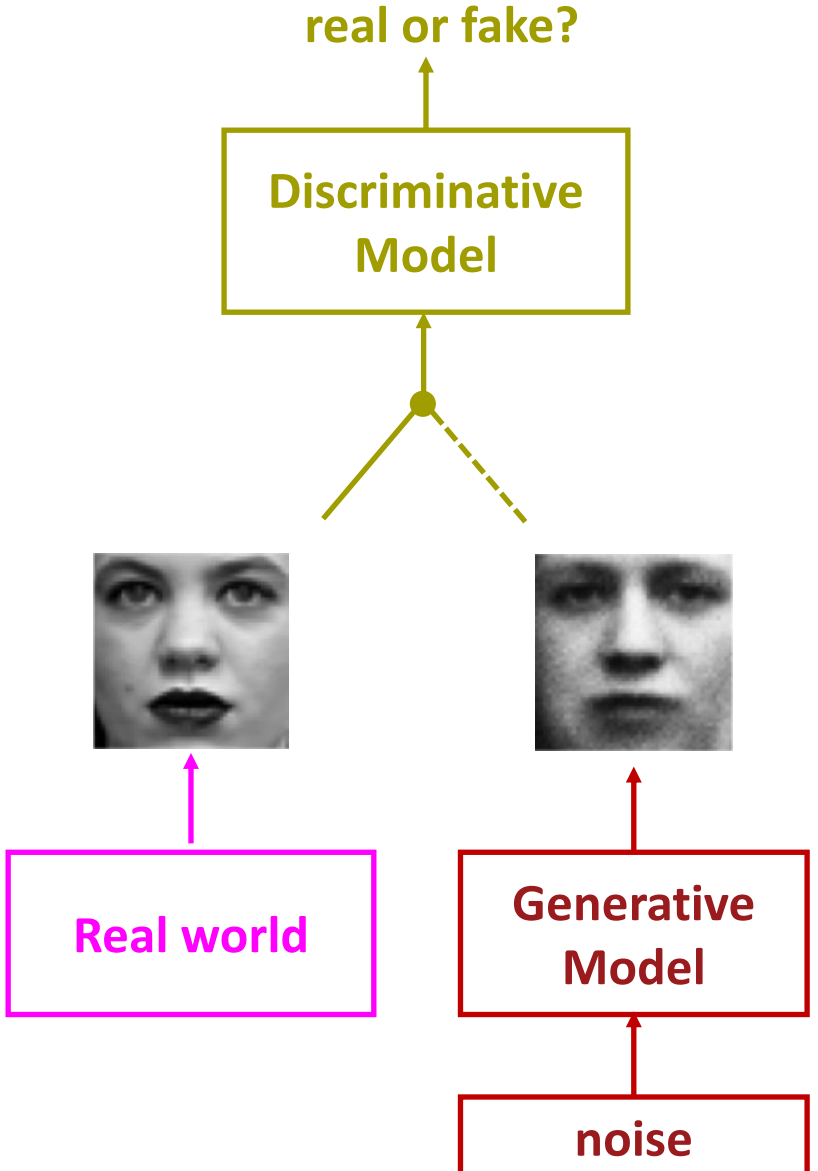
VIMAL
VISUAL INTELLIGENCE AND MULTIMEDIA ANALYTICS LABORATORY

NEURAL NETWORKS



Deep means many layers

GENERATIVE ADVERSARIAL NETWORKS (GAN)



USC Viterbi Information Sciences Institute

AI-GENERATED CONTENT

VIMAL
VISUAL INTELLIGENCE AND MULTIMEDIA ANALYTICS LABORATORY

GAN-GENERATED FACES

https://this-person-does-not-exist.com/en

DEEPFAKES

Fake video of a target person (i.e. victim) crated using ***deep*** neural network for face swapping

Amy Adams Nicolas Cage

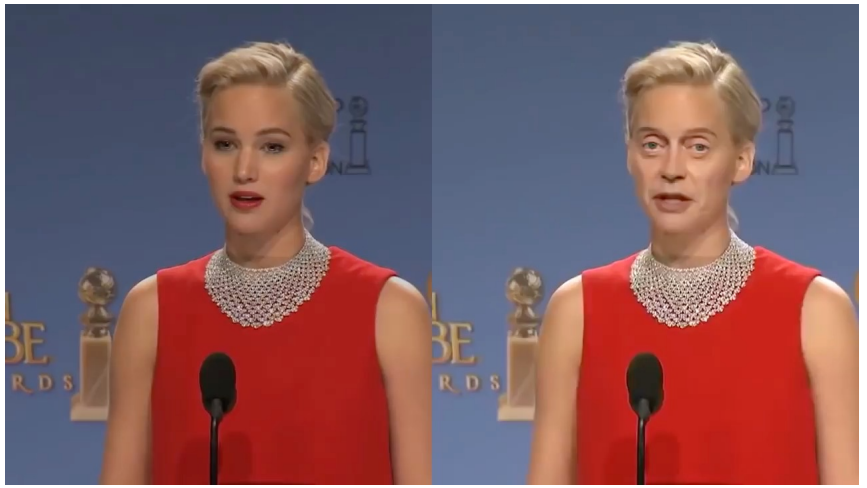
HOW LONG DID DEEPFAKES EXIST?

- Forever
- Paul Walker in Fast & Furious 7 replacement cost $50M

HOW LONG DID DEEPFAKES EXIST?

TYPES OF DEEPFAKES

- Face replacement
  - Replace entire face with that of a target person



- Face reenactment
  - Replace facial expressions and lip movements with those of a source face

DEEPFAKE METHODS

- *FaceSwap* (for replacement)
  - Graphics-based, extract landmarks from source, fit 3D model with blendshapes, project onto target
- *Deepfake* (for replacement)
  - Two autoencoders with shared encoder, trained source autoencoder is applied to target, blend using Poisson blending

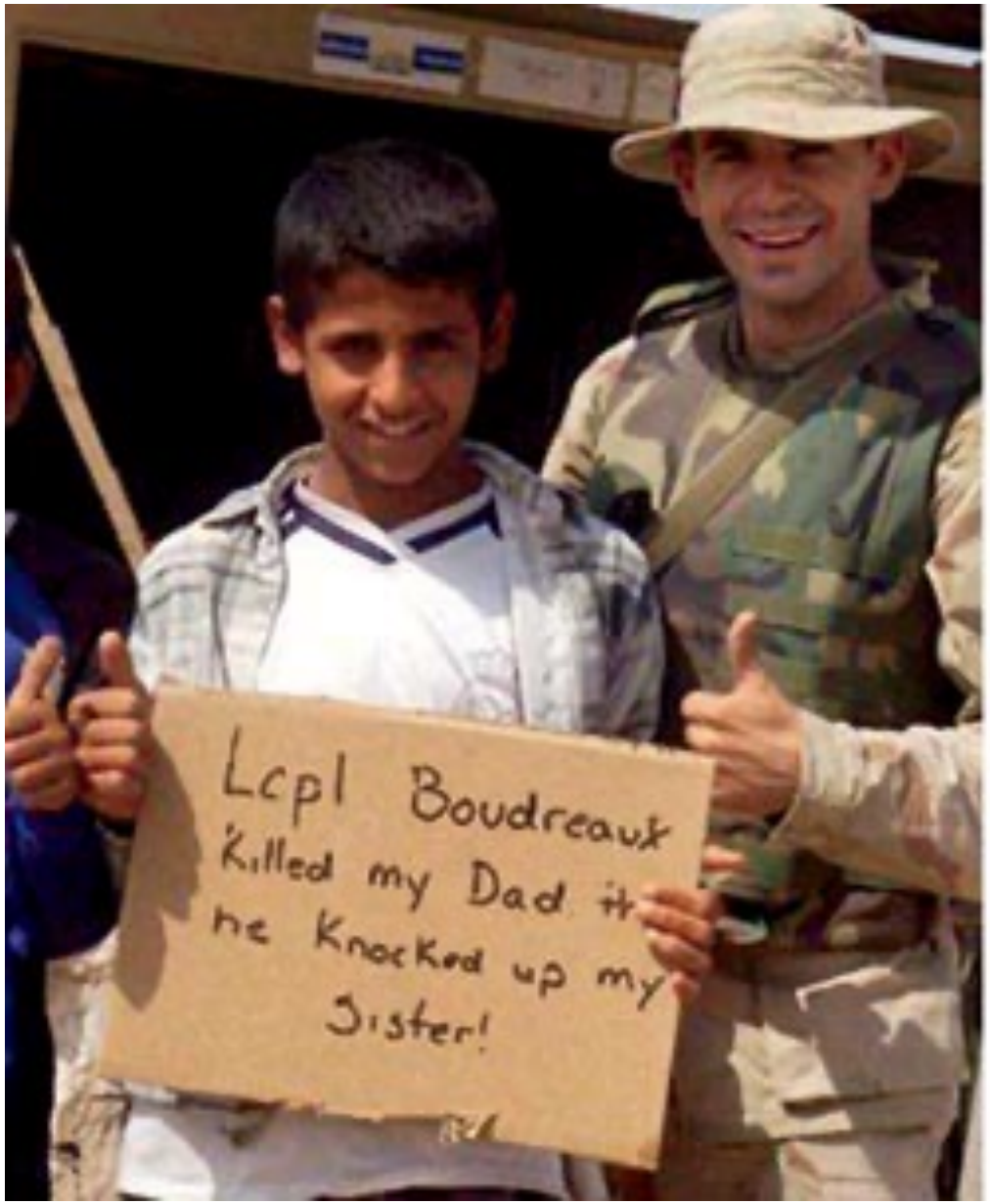
DEEPFAKE METHODS

- *Face2Face* (for reenactment)
  - Keyframe selection, dense reconstruction of target face, transfer expressions and lip movements of source video
- *NeuralTextures* (for reenactment)
  - Train a CNN to learn texture and render target video, only transfer mouth expressions

USC Viterbi Information Sciences Institute

DANGERS OF DEEPFAKES

VIMAL
VISUAL INTELLIGENCE AND MULTIMEDIA ANALYTICS LABORATORY



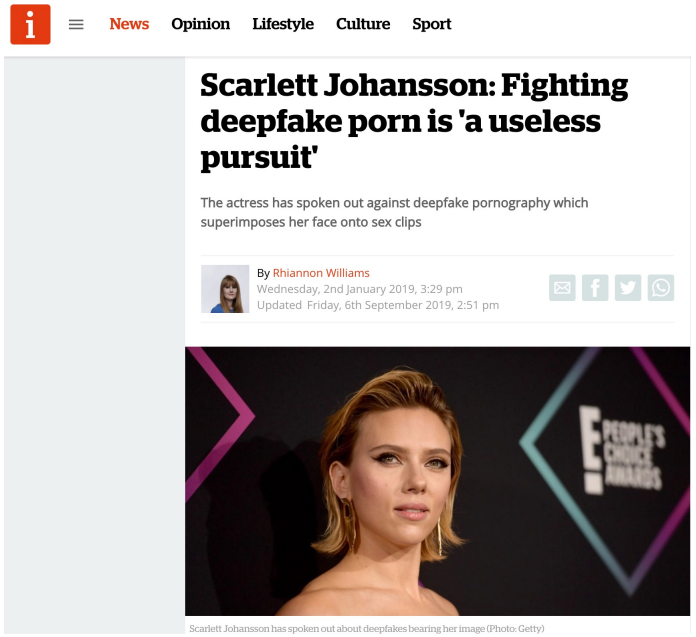
USC Viterbi
Information Sciences Institute

VIMAL
VISUAL INTELLIGENCE AND MULTIMEDIA ANALYTICS LABORATORY

WHY ARE DEEPFAKES DANGEROUS?

Jeremy Corbyn backing Boris Johnson

Misinformation and national security



Revenge porn

- Child exploitation
- Stock market
- Court room evidence
- Fraud

WHY ARE DEEPFAKES DANGEROUS?

- 14,678 deepfake videos online
- 96% of which were pornographic in nature
- 100% featured women
- 134,364,438 views
- Doubles every six months

* The State Of The Deepfakes, Deeptrace, September 2019

ENOUGH OF AN ISSUE THAT MAJOR SOCIAL MEDIA SITES HAD TO CRACK DOWN

USC Viterbi Information Sciences Institute

TECH / ARTIFICIAL INTELLIGENCE

Reddit bans 'deepfakes' AI porn communities

By ADI ROBERTSON / @thedextriarchy
Feb 7, 2018, 10:28 AM PST | 0 Comments / 0 New

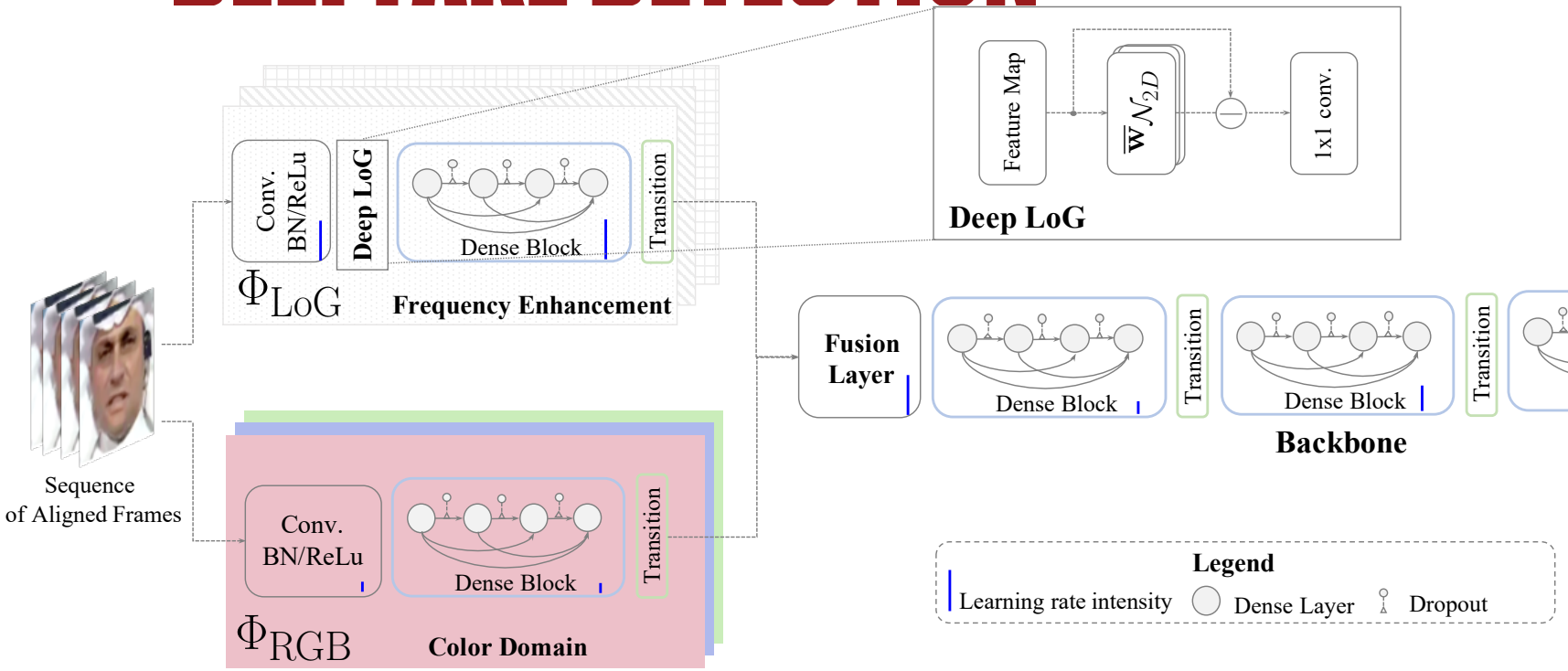
VIMAL
VISUAL INTELLIGENCE AND MULTIMEDIA ANALYTICS LABORATORY

USC Viterbi Information Sciences Institute

DETECTING DEEPFAKES

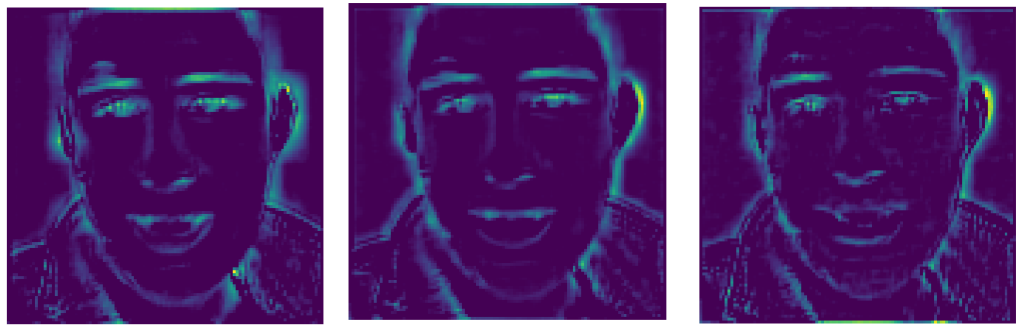
VIMAL
VISUAL INTELLIGENCE AND MULTIMEDIA ANALYTICS LABORATORY

DEEPFAKE DETECTION

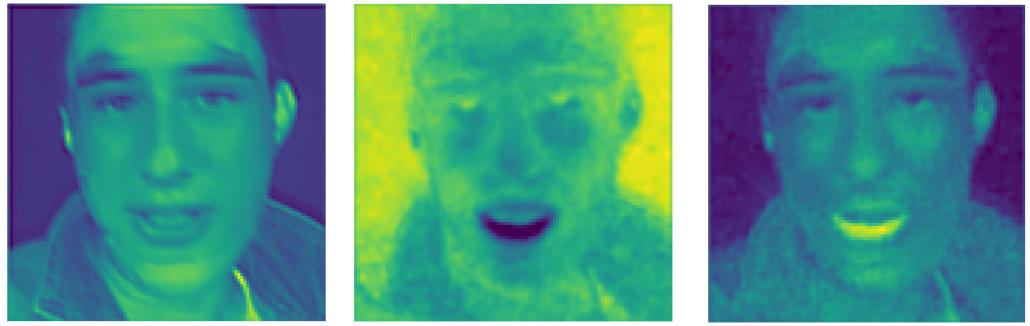


Input

Φ_{LoG}



Φ_{RGB}

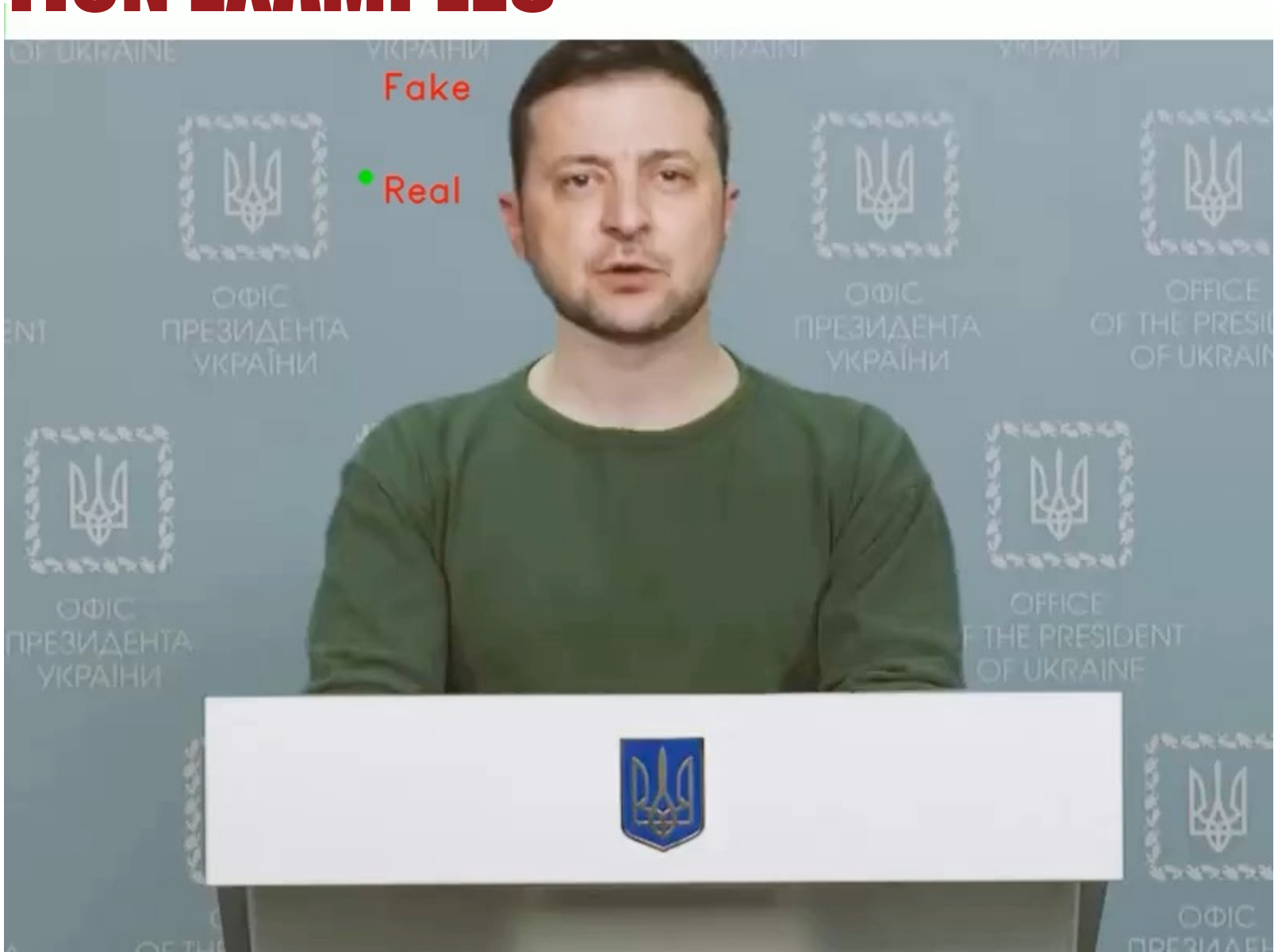


DETECTION EXAMPLES

USC Viterbi Information Sciences Institute

VIMAL
VISUAL INTELLIGENCE AND MULTIMEDIA ANALYTICS LABORATORY

DETECTION EXAMPLES



USC Viterbi Information Sciences Institute

VIMAL
VISUAL INTELLIGENCE AND MULTIMEDIA ANALYTICS LABORATORY

CHALLENGES

- Detecting new deepfake generation methods
- Social networks

CHALLENG [SIC]: SELF-SELECTION MAKES DEEP FAKES PROPAGATE

USC Viterbi Information Sciences Institute

Why do Nigerian Scammers Say They are from Nigeria?

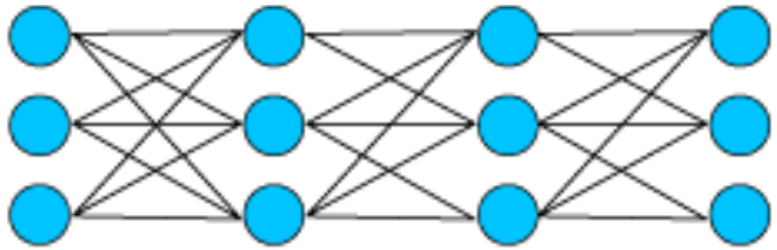
Cormac Herley
Microsoft Research
One Microsoft Way
Redmond, WA, USA
cormac@microsoft.com

VIMAL
VISUAL INTELLIGENCE AND MULTIMEDIA ANALYTICS LABORATORY

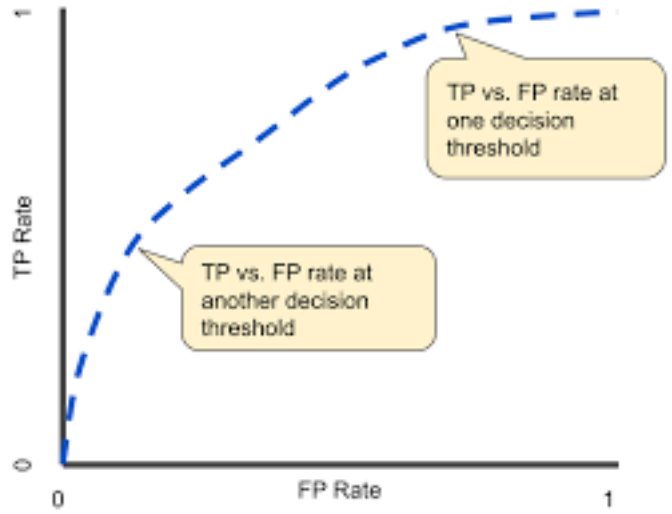
PREVIOUS WORK ON EMAIL SCAMS

- Scammers make their scams transparent to weed out “false positives”
- Scammers can focus on the most gullible individuals
- Deep Fakes may have similar audiences
  - Deep fakes for fraud implies even poor fakes can work for a select audience
  - Anti-vaccine users, for example, are more conps

AI IS ALSO GETTING BETTER



real or fake?



VISUAL INTELLIGENCE AND MULTIMEDIA ANALYTICS LABORATORY

WHAT ARE WE ACTUALLY DOING?

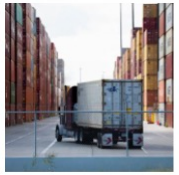
- Break security
- Sway elections
- Shame women
- Exploit children
- Crash stock markets

MISLEAD THE PUBLIC

MARKETS see all →

▲ DOW	36,585.06	+246.76	+0.68%
▲ S&P 500	4,796.56	+30.38	+0.64%
▲ NASDAQ	15,832.80	+187.83	+1.20%

FEATURED



The US economy in 12 charts

From jobs to GDP, these key indicators provide a comprehensive, up-to-date picture of the US Economy.

LATEST

- Tesla just opened a new showroom in China's Xinjiang region
- AT&T and Verizon agree to postpone 5G rollout near airports by 2 weeks
- The key moments from Elizabeth Holmes' trial

Amazon's Alexa tells 10-year-old child to touch penny to exposed plug socket

By Sana Noor Haq, CNN

Updated 3:59 PM ET, Wed December 29, 2021

PUBLIC PUSH BACK

CNN BUSINESS Markets Tech Media Success Perspectives Videos LIVE TV Edition

California lawmakers ban facial-recognition software from police body cams

By Rachel Metz, CNN Business
Updated 8:04 AM ET, Fri September 13, 2019

FAST COMPANY

12-02-19 | CONNECTED WORLD

Portland plans to propose the strictest facial recognition ban in the country

Portland, Oregon, aims to ban the use of the controversial technology not only by city government, but also by private companies.

[Source photo: andipantz/iStock; metamorworks/iStock]

The New York Times

San Francisco Bans Facial Recognition Technology

Attendees interacting with a facial recognition demonstration at this year's CES in Las Vegas. Joe


I DID NOTHING
BUT ILL TAKE ALL THE CREDIT
makeameme.org

USC Viterbi Information Sciences Institute

FINAL THOUGHTS


isn't
SEEING IS
BELIEVING

USC Viterbi Information Sciences Institute

CONNECTION TO FAIRNESS...

An Examination of Fairness of AI Models for Deepfake Detection

Loc Trinh* **Yan Liu**

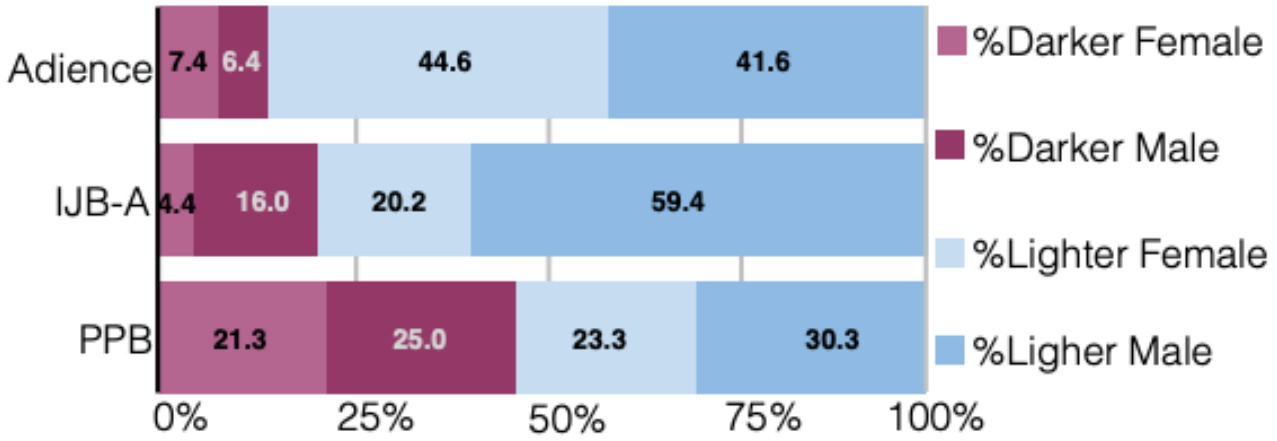
Department of Computer Science
University of Southern California
Los Angeles, CA 90089
{loctrinh, yanliu.cs}@usc.edu

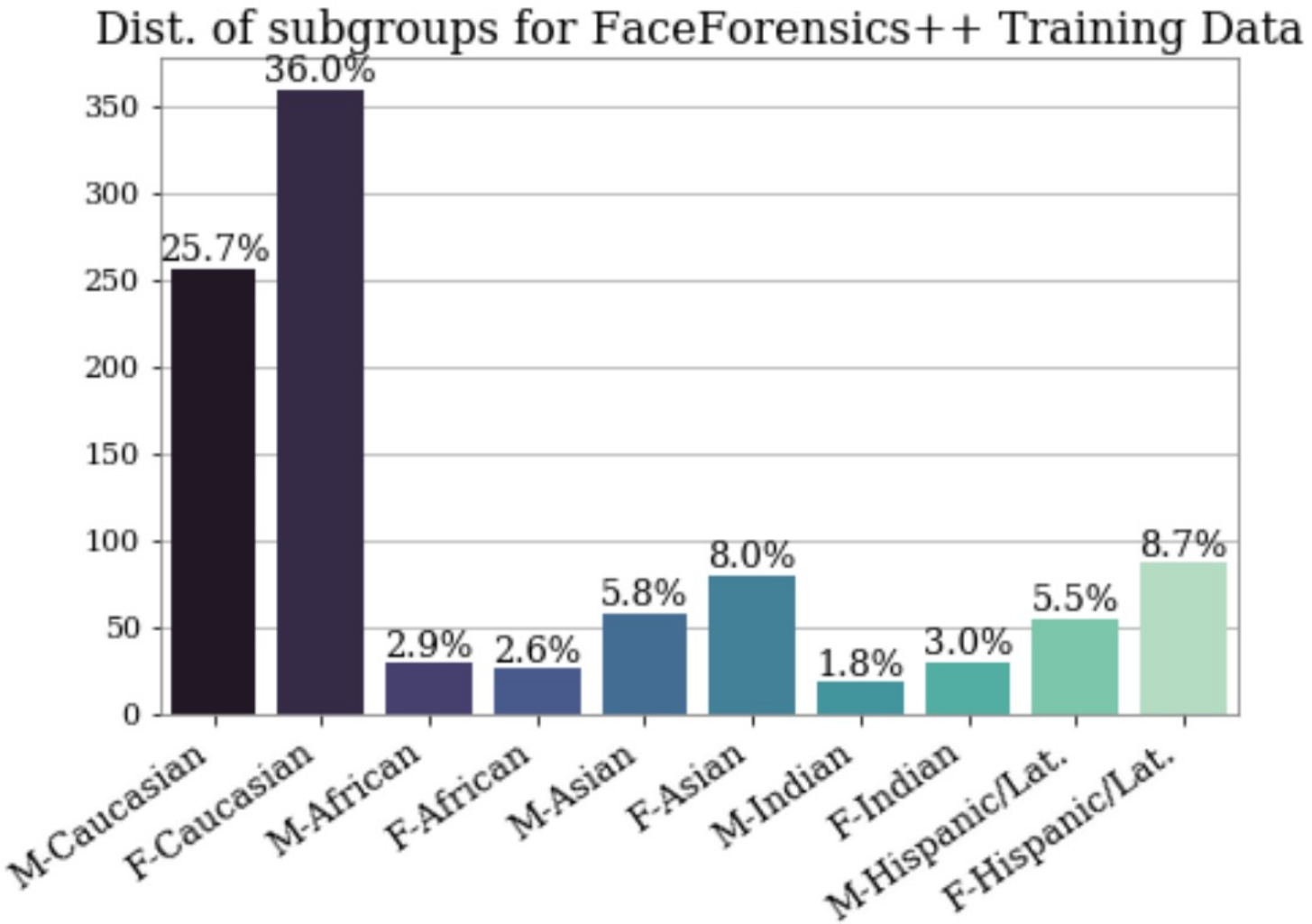
PROBLEM

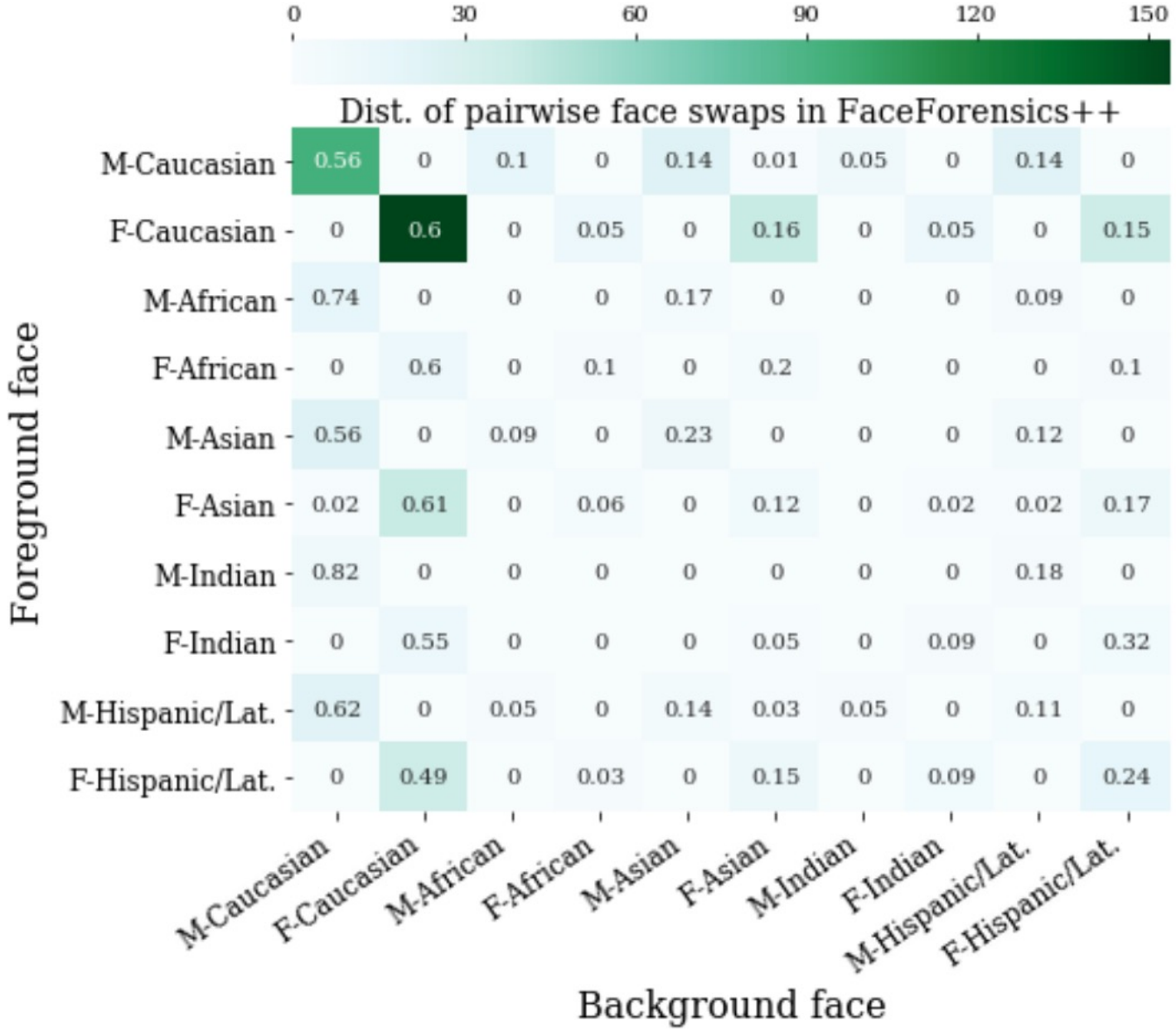
- Detecting fake videos at a low false positive rate is a challenging problem.
- Little is discussed about how such systems perform on diverse groups of real people across gender and race
- Millions people of a particular group are more likely to be mistakenly classified as fake

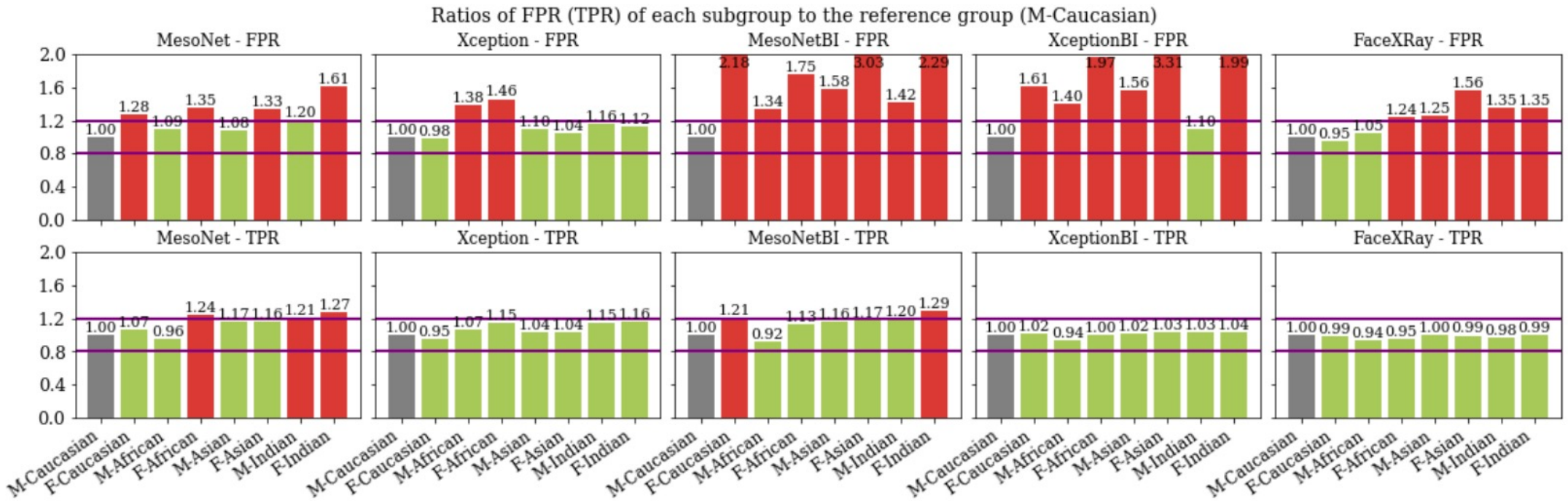
CONTEXT

- Gender Shades [Buolamwini and Gebru, 2018]
  - facial recognition systems discriminate across gender and race
  - Large gap in the accuracy of gender classifiers across different intersectional groups




Dist. of subgroups for FaceForensics++ Training Data
350
300
250
200
150
100
50
0
25.7%
36.0%
2.9%
2.6%
5.8%
8.0%
1.8%
3.0%
5.5%
8.7%
M-Caucasian
F-Caucasian
M-African
F-African
M-Asian
F-Asian
M-Indian
F-Indian
M-Hispanic/Lat.
F-Hispanic/Lat.



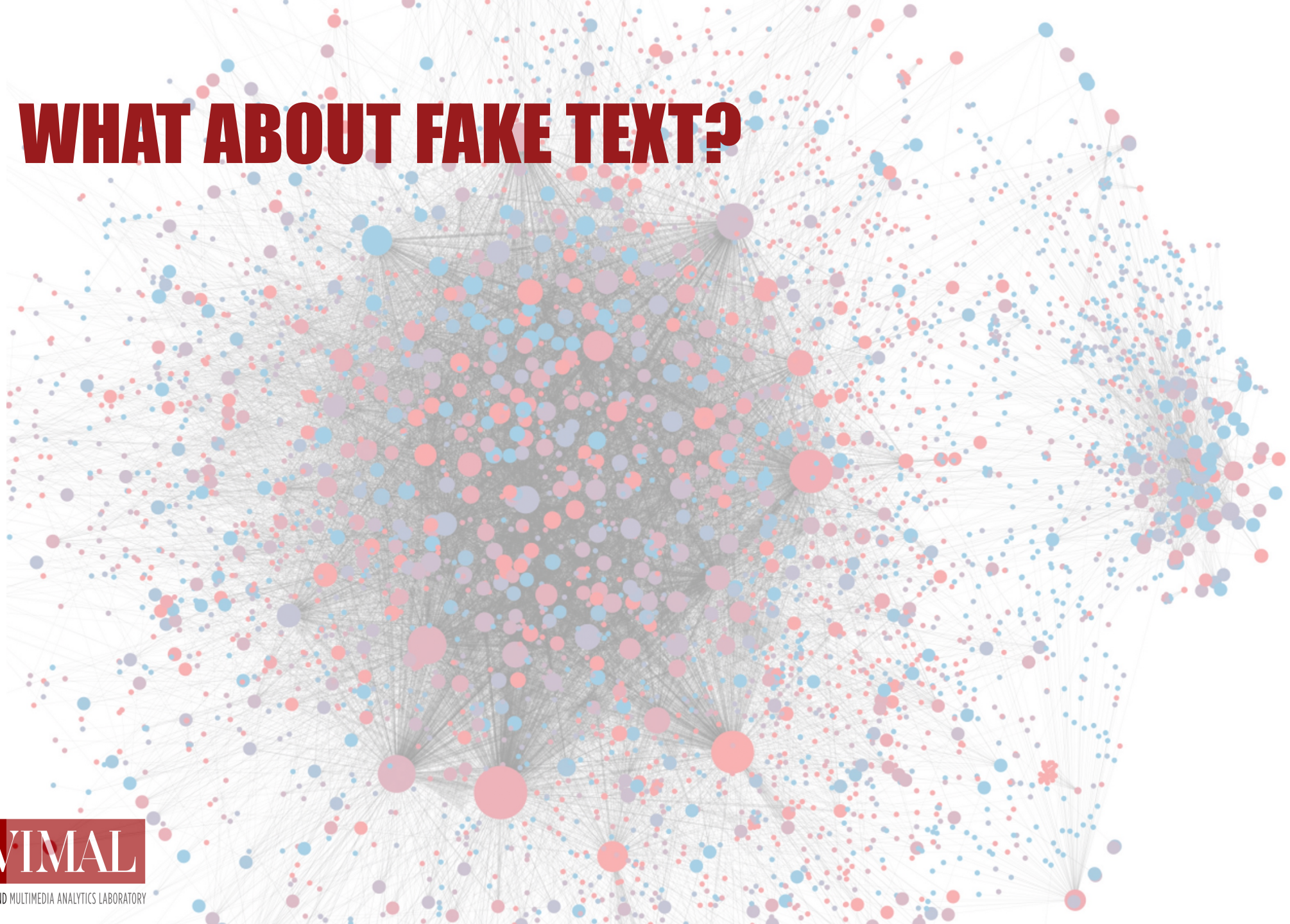

Ratios of FPR (TPR) of each subgroup to the reference group (M-Caucasian)
MesoNet - FPR
Xception - FPR
MesoNetBI - FPR
XceptionBI - FPR
FaceXRay - FPR
MesoNet - TPR
Xception - TPR
MesoNetBI - TPR
XceptionBI - TPR
FaceXRay - TPR
2.0
1.6
1.2
0.8
0.4
0.0
1.00 1.28 1.09 1.35 1.08 1.33 1.20 1.61
1.00 0.98 1.38 1.46 1.10 1.04 1.16 1.12
1.00 2.18 1.34 1.75 1.58 3.03 1.42 2.29
1.00 1.61 1.40 1.97 1.56 3.31 1.10 1.99
1.00 0.95 1.05 1.24 1.25 1.56 1.35 1.35
1.00 1.07 0.96 1.24 1.17 1.16 1.21 1.27
1.00 0.95 1.07 1.15 1.04 1.04 1.15 1.16
1.00 1.21 0.92 1.13 1.16 1.17 1.20 1.29
1.00 1.02 0.94 1.00 1.02 1.03 1.03 1.04
1.00 0.99 0.94 0.95 1.00 0.99 0.98 0.99
M-Caucasian
F-Caucasian
M-African
F-African
M-Asian
F-Asian
M-Indian
F-Indian

FACEFORENSICS++ PERFORMS POORLY ON NON-CAUCASIANS

- FaceForensics++ is meant to detect real or fake images
- This model has lower false-positive rate for Caucasians than most other demographics
- These models need to be more fair (or adversaries can choose the least-detectable race within their misinformation/fraud campaign)

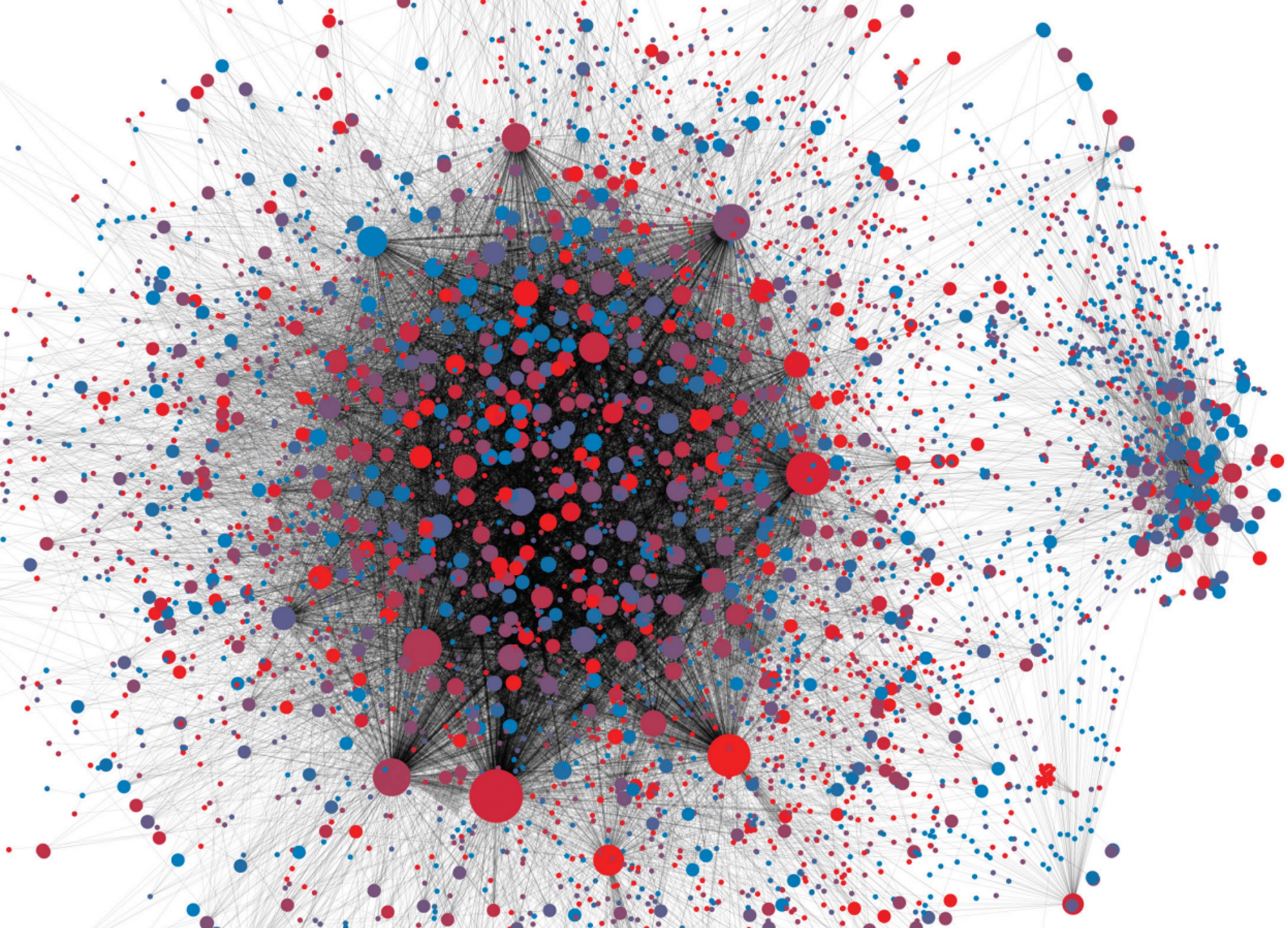
SO FAR WE HAVE ONLY DISCUSSED IMAGES/VIDEO…

WHAT ABOUT FAKE TEXT?



USC Viterbi Information Sciences Institute

VIMAL
VISUAL INTELLIGENCE AND MULTIMEDIA ANALYTICS LABORATORY



USC Viterbi Information Sciences Institute

Bots

Humans

TEXT GENERATION CAN MAKE FAKE NEWS SPREAD FASTER

SOCIAL SCIENCE

The science of fake news

Addressing fake news requires a multidisciplinary effort

gated about topics such as vaccination, nutrition, and stock values. It is particularly pernicious in that it is parasitic on standard news outlets, simultaneously benefiting from and undermining their credibility.

Some—notably First Draft and Facebook—favor the term "false news" because of the

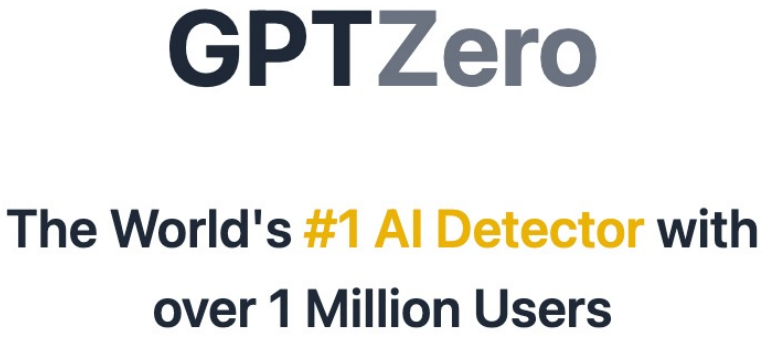
USC Viterbi Information Sciences Institute

VIMAL
VISUAL INTELLIGENCE AND MULTIMEDIA ANALYTICS LABORATORY

COORDINATION CAN ENFACE DIVISIONS, LEADING TO DISTRUST IN AI

- Fake text generation can…
  - Propagate rumors at scale
  - Use distinct language (coordination harder to detect)
  - Realistic conversations
- Ethics: we do not want our tools to cause harm
- Setting AI back: upon realization, public will have less trust in all AI
- Some parts of AI are harmful, we need to address this

HOW CAN WE DETECT AI GENERATED TEXT?



What data did you train your model on?

- We trained our models on a dataset of paired human-written and AI-generated text. Our human-written text spans student-written articles, news articles, as well as question and answer datasets spanning multiple disciplines in the sciences and humanities. For each article of human-written text, we generate corresponding articles with AI to ensure there isn't topic-level bias in our dataset. Finally, we train our model with an equal balance of human and AI-written articles.

HOW WILL THIS FAIL?

- “Our human-written text spans student-written articles, news articles, as well as question and answer datasets spanning multiple disciplines in the sciences and humanities”
  - Implicit assumption: likely English, and likely Standard English
  - Data Shift means AI can generate non-standard English and this model can fail
- Fair AI is critical to assess fake detection

EXAMPLE IN STANDARD ENGLISH

- Me: Tell me how good the internet is here. Give it to me as a Nigerian folk tale.
- **ChatGPT:** “Once upon a time, in a faraway land, there lived a powerful spirit named Oya. Oya was known for her speed and strength, and her presence could be felt in the wind that blew through the fields…”

GPTZero

The World's #1 AI Detector with over 1 Million Users

Your text may include parts written by AI

NON-STANDARD ENGLISH

- **Me:** “Now tell me how good the internet is here. Give it to me as a Nigerian folk tale in Nigerian pidgin english.”
- **ChatGPT:** “E get one time for this land, wey e be say the people dem wan follow the whole world connect, but the fear wey dey catch dem be say their messages no go fit waka reach where dem wan send am go…”

USC Viterbi Information Sciences Institute

GPTZero

The World's #1 AI Detector with over 1 Million Users

Your text is likely to be written entirely by a human

VIMAL
VISUAL INTELLIGENCE AND MULTIMEDIA ANALYTICS LABORATORY

CAVEAT

- This is evidence, not proof
- A systematic study is needed to see how fake text detection works on
  - Different languages (not current LLMs are trained on English)
  - Dialects
  - Sociolects
  - Pidgin (practical common language, mix of several languages)
  - Creoles (language variant children learn , mix of several languages)
- New methods needed to make fake text detection robust

USC Viterbi Information Sciences Institute

CONCLUSIONS

- What are DeepFakes?
  - Images or text used to create misinformation
  - Easy to fool us!
- How are they used for harm?
  - Generate anger, protests
  - Used to push narratives
- Detecting DeepFakes
  - Training data affects testing accuracy
  - May perform poorly on data not seen in training set (different races, non-standard dialects)