Automatizando la detección de contenido (deep) fake

Dr. Alfonso Muñoz (@mindcrypt)
Jose Ignacio Escribano

$whoami

Dr. Alfonso Muñoz – Twitter: @mindcrypt
Global Technical Cybersecurity Lead & Head of cybersecurity lab
Telegram: t.me/criptored | alfonso@criptored.com | Github: mindcrypt

CriptoCert

- **Offensive/defensive security expert (certifications, European Organisms, public bodies and companies global 500, +60 academic publications, talks, patents, IEEE, ACM, books, …)**
- **Speaker:** STIC CCN-CERT, DeepSec, HackInTheBox, Virus Bulletin, RootedCon, 8.8, …
- **Profesor de Máster Seguridad:** UEM, UNIR, UC3M, UPM, UJAEN, EOI
- **Cybersecurity expert - Europol European Cybercrime Centre (EC3)**
- **Co-editor CRIPTORED –** Red Temática de Criptografía y seguridad de la información – http://www.criptored.com
- **Co-(autor) de la certificación CriptoCert Certified Crypto Analyst – https://www.criptocert.com**

$whoami

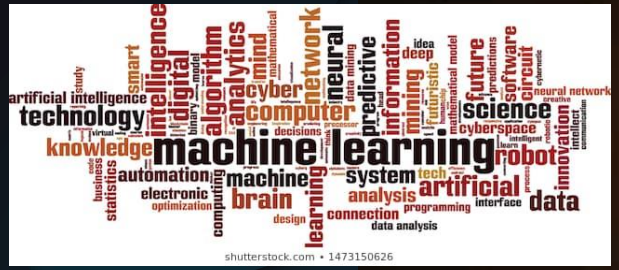
José Ignacio Escribano
joseignacio.escribano.pablos.next@bbva.com |
Github: @jiep

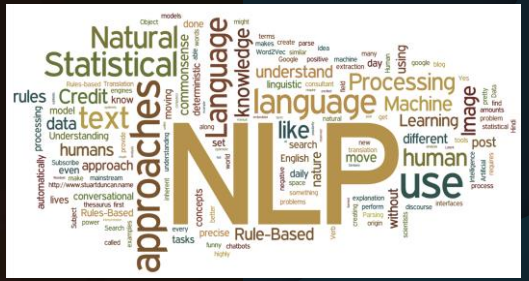
- Security & Machine Learning Researcher – BBVA Next Technologies
- Vidas paralelas
  - Graduado en Matemáticas e Ingeniería del Software & Máster en Ingeniería de la Decisión.
  - Actualmente, realizando tesis sobre criptografía post-cuántica.
  - Ponente en Cybercamp y Hack In The Box
- Áreas
  - Inteligencia Artificial
  - Criptografía
  - Cutting-edge research (Offensive & defensive)

Índice

01
Introducción

Contenido sintético y
ciberseguridad





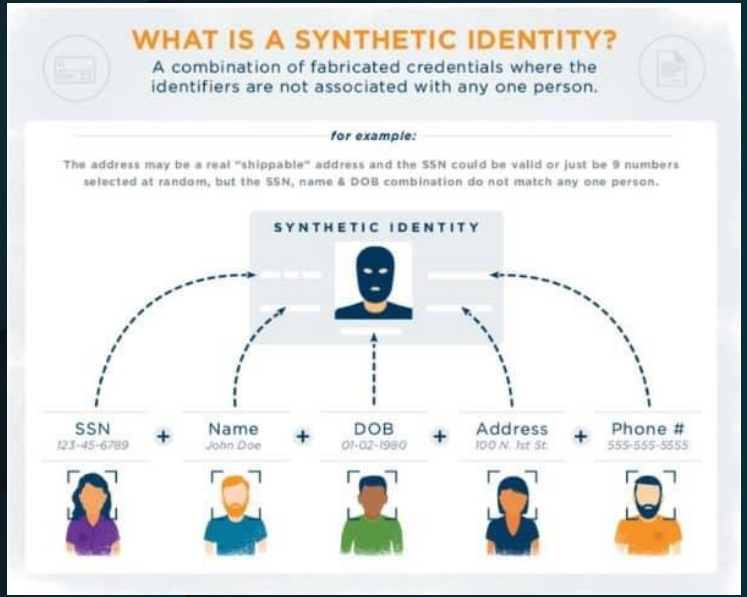
SAY WHAT?

>h-c0n

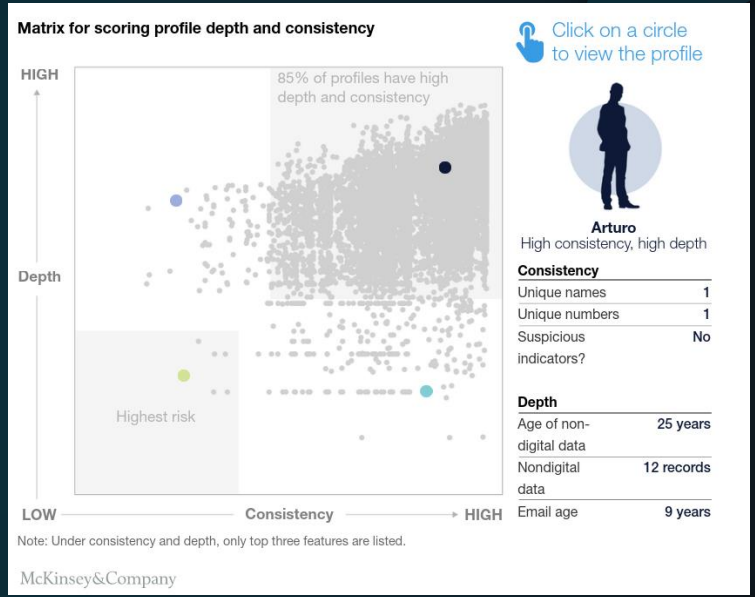
¿Qué es la identidad sintética?

- El principal **objetivo** de esta presentación es dar una visión sobre la generación y detección de contenido sintético para **engañar** a una víctima.
- **No es algo novedoso** pero debido a la facilidad de acceso a **herramientas automáticas** es más sencillo.
- Diferencias dependiendo cómo se quiere generar, habría dos vertientes:
    - Graphical Computing
    - Deep Learning
- La aplicación de estas identidades para cometer **fraude** está en aumento, por lo que es necesario comprobar los métodos de autenticación.

Identidad sintética

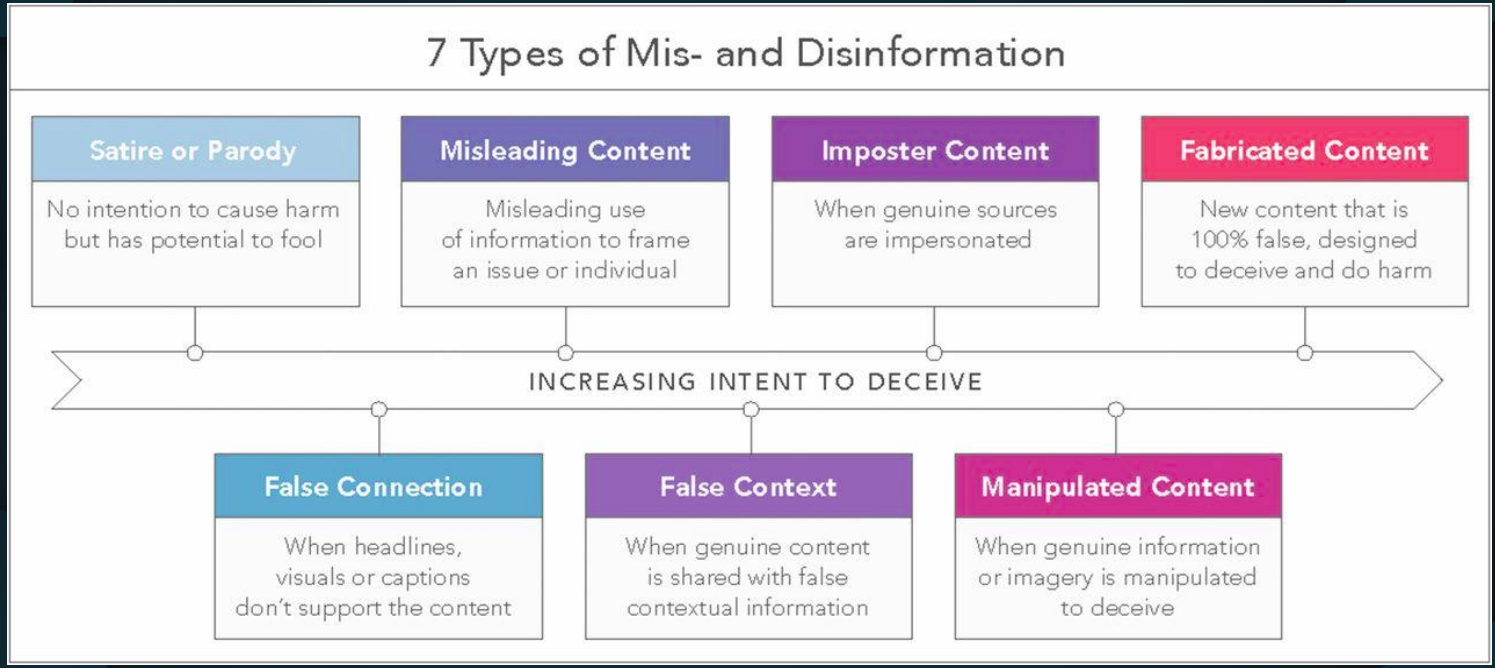


https://www.idanalytics.com/solutions-services/fraud-risk-management/synthetic-identity-fraud/



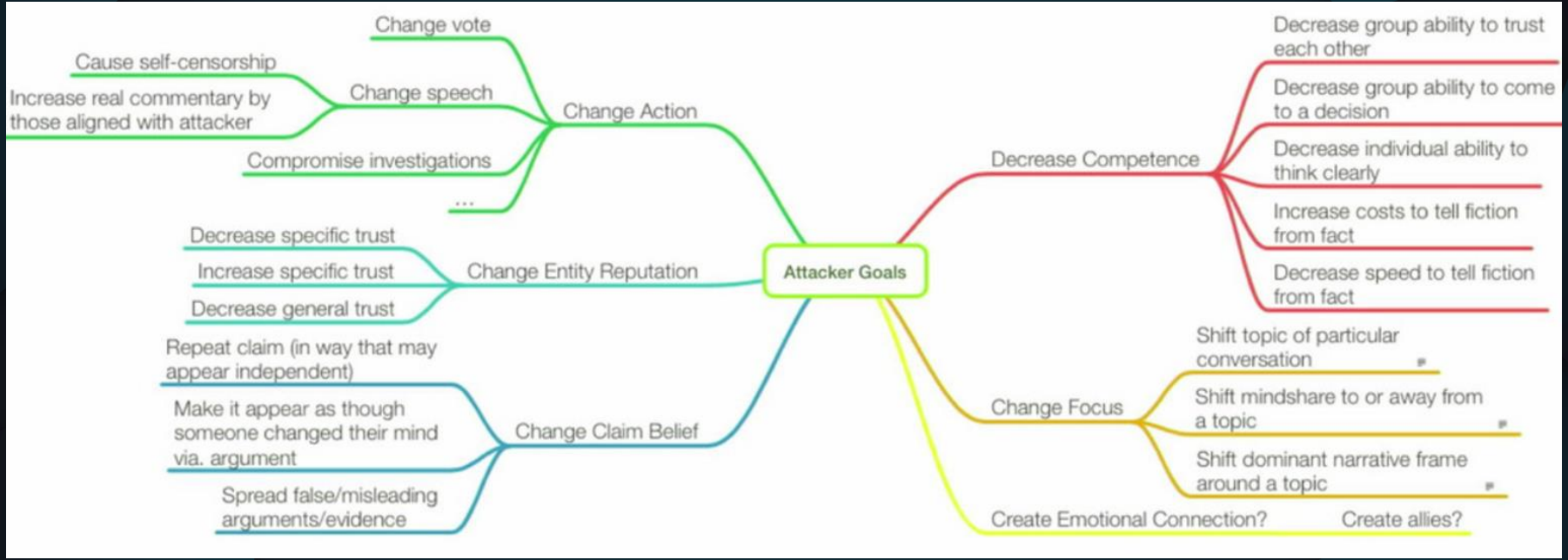
https://www.mckinsey.com/business-functions/risk/our-insights/fighting-back-against-synthetic-identity-fraud

Threat models – ¿por qué?



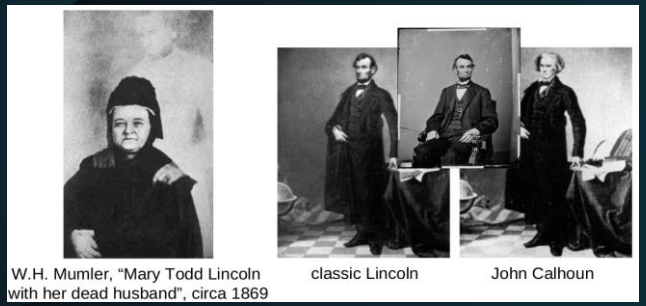
https://www.pnas.org/content/114/48/12631

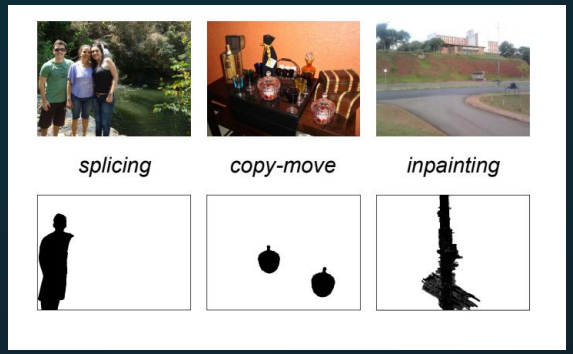
Threat models – ¿para qué?

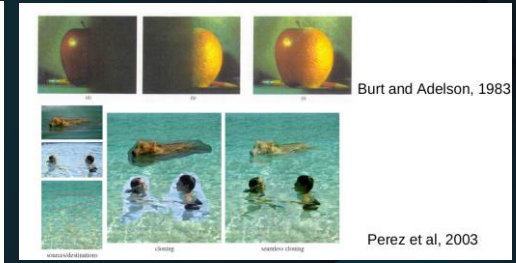


https://slideslive.com/38917755/appliad-synthesis-and-detection

Evolución de la manipulación







Evolución de la manipulación

Evolución de la manipulación – Fake News

Sistemas de verificación manual

- https://www.factcheck.org
- https://maldita.es
- https://www.newtral.es

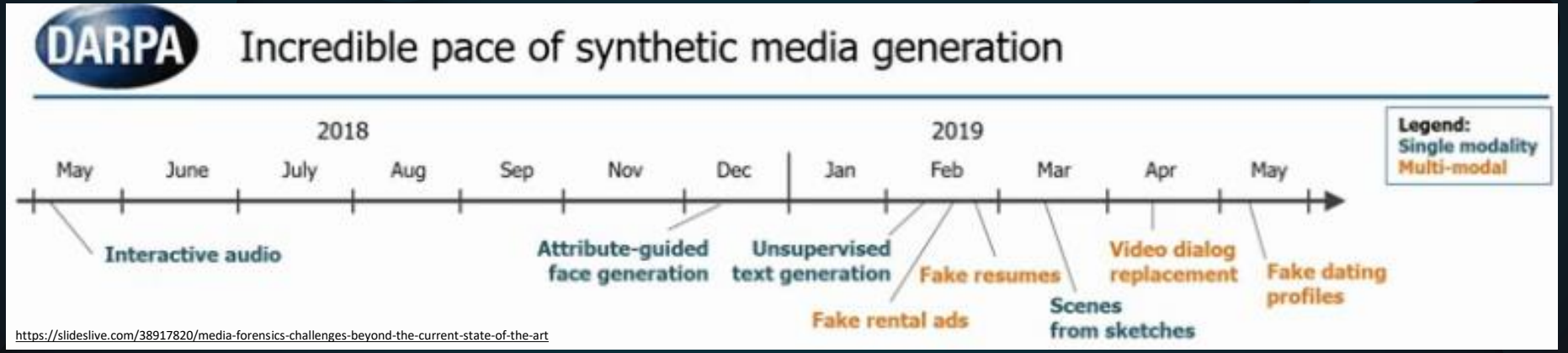
Contenido fuera de contexto

- Otros lugares.
- Otros momentos temporales.
- Falsificaciones caseras.

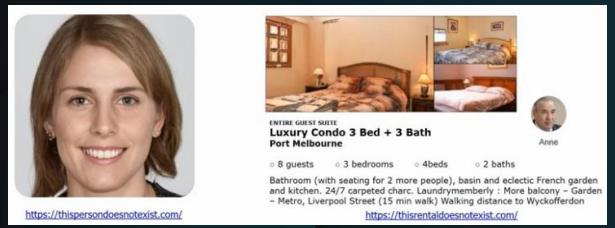
02
Generación de contenido sintético

Texto, audio, imagen y vídeo

Generación de contenido



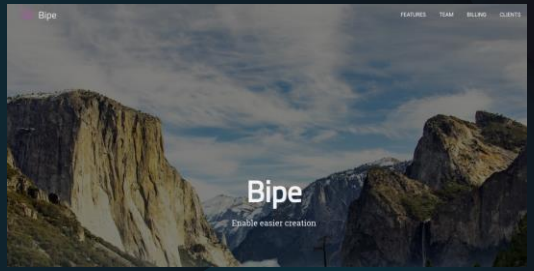
https://slideslive.com/38917820/media-forensics-challenges-beyond-the-current-state-of-the-art



https://thispersondoesnotexist.com/

https://thesecatsdonotexist.com

https://thissnackdoesnotexist.com



https://thisstartupdoesnotexist.com

Requisitos generales de generación

Datos

Uno de los problemas más importantes en estos sistemas es el de tener datos de **calidad** para entrenar los modelos.

Capacidad de cómputo

Es necesario contar con **equipamiento** necesario para realizar las tareas en tiempos asumibles (**GPU**).

Conocimiento

Entender el funcionamiento de los modelos subyacentes ayuda a generar modelos personalizados.

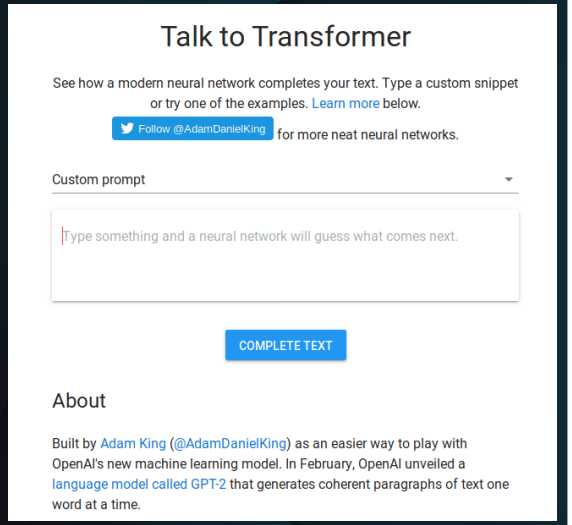
Generación de texto

- GPT-2 es el sucesor de GPT, es el modelo estado del arte de generación de texto.
- Liberado por OpenAI (https://github.com/openai/gpt-2).
- Predice la próxima palabra del texto.
- Consta de 1500 millones de parámetros.
- Entrenado con un conjunto de datos de 8 millones de páginas web.

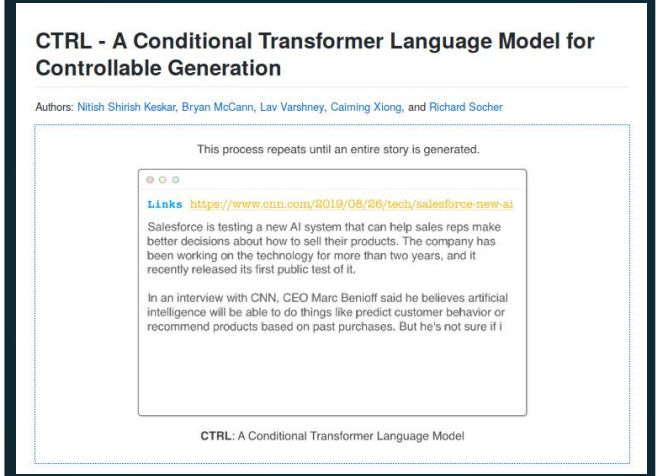
https://openai.com/blog/better-language-models/

https://blog.floydhub.com/gpt2/

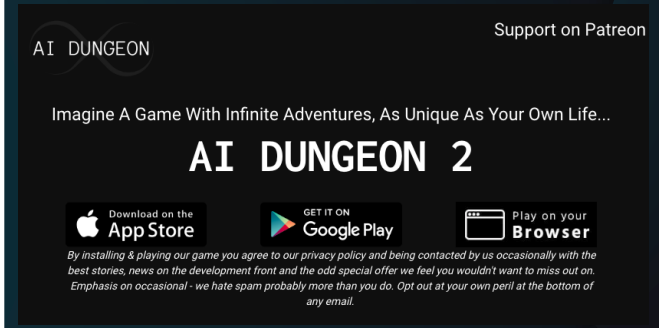
Generación de texto – Aplicaciones



https://talktotransformer.com/



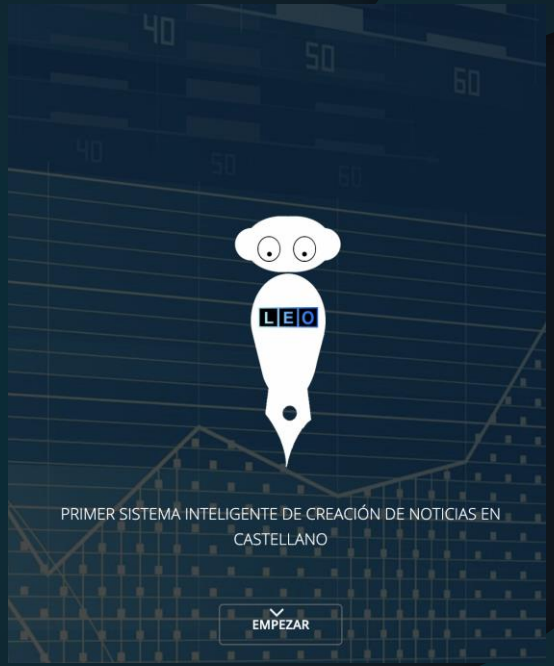
https://github.com/salesforce/ctrl



https://www.aidungeon.io/start

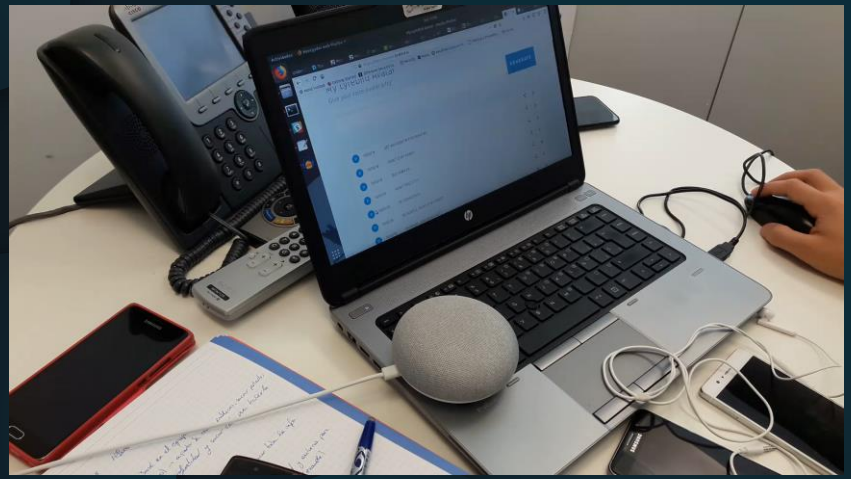
Generación de texto – Limitaciones

- Generación de textos solo en **inglés**.
  - No existen conjuntos de **datos en español** para entrenar modelos.
  - Las investigaciones se realizan en inglés y después se mejoran las traducciones, pero no tratan de generar textos nativos en **otros idiomas**.
- Textos largos con **poca coherencia**.



http://leoia.es/

Generación de audio



Lyrebird + Descript

You might have noticed a Descript feature in the video called “Overdub” — it deserves a few paragraphs of context.

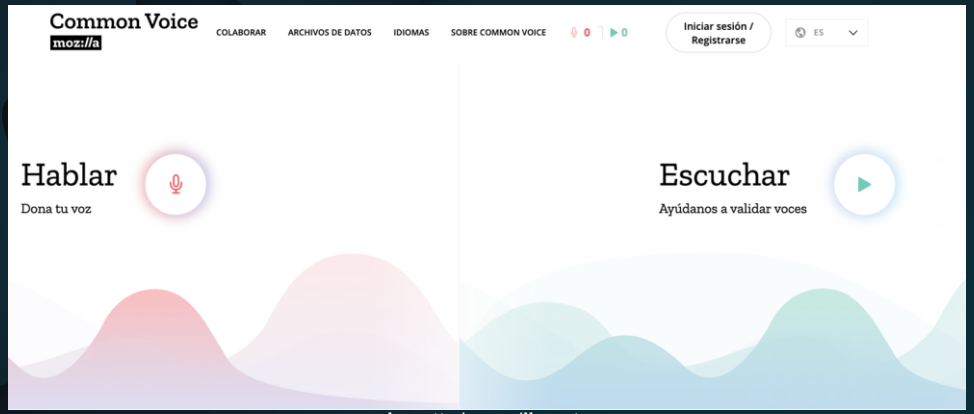
In 2017, Alexandre de Brebisson, Kundan Kumar, Jose Sotelo, PhD students studying at **MILA** under **Yoshua Bengio**, launched a startup called Lyrebird, allowing anyone to generate a realistic text-to-speech model of themself by uploading a few minutes of audio.

As soon as we saw it, we became conscious of a Lyrebird-shaped hole in Descript. In some ways, it felt like our missing half — Descript let you delete audio by deleting text, but Lyrebird would let you add audio by typing text.

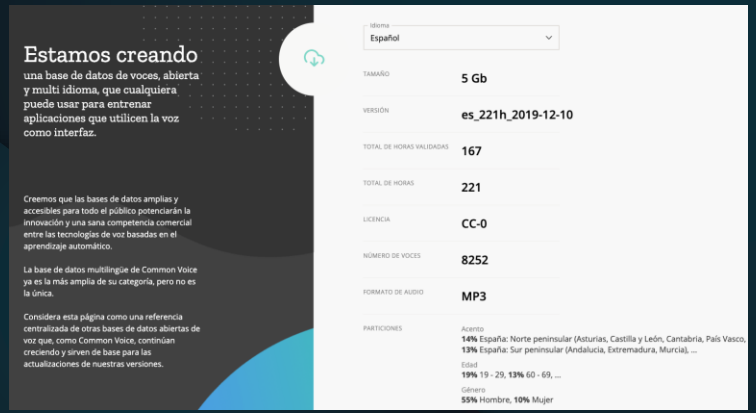
A few months ago, we started talking to Lyrebird, and the more we talked, the more we realized how well we complimented each other — not only enabling personalized speech generation, but a whole new class of AI-enabled tools for media editing and synthesis.

So we merged the companies. And Overdub is our first collaboration, available today in closed beta.

Generación de audio – Datos

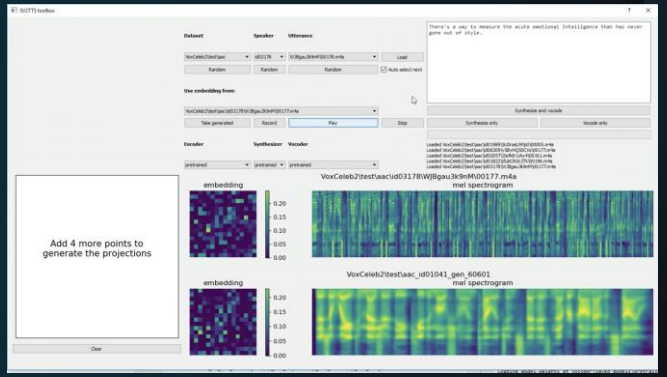


https://voice.mozilla.org/



Real Time Voice Cloning

- Real-Time-Voice-Cloning es una herramienta **OpenSource** de generación de voz.
- Promete clonar la voz con sólo **5 segundos** de audio.
- Disponible en **GitHub**: https://github.com/CorentinJ/Real-Time-Voice-Cloning.
- Implementa **4 modelos** previos:
  - SV2TTS (TTS): https://arxiv.org/pdf/1806.04558.pdf
  - WaveRNN (vocoder): https://arxiv.org/pdf/1802.08435.pdf
  - Tacotron 2 (sintetizador): https://arxiv.org/pdf/1712.05884.pdf
  - GE2E (encoder): https://arxiv.org/pdf/1710.10467.pdf
- Desarrollado en **Python**.
- Necesario disponer de **GPU**.



Real Time Voice Cloning (demo)

- Es necesario un **fichero de audio de la voz** que se quiere clonar sin ruido (.mp3, .wav, ...).

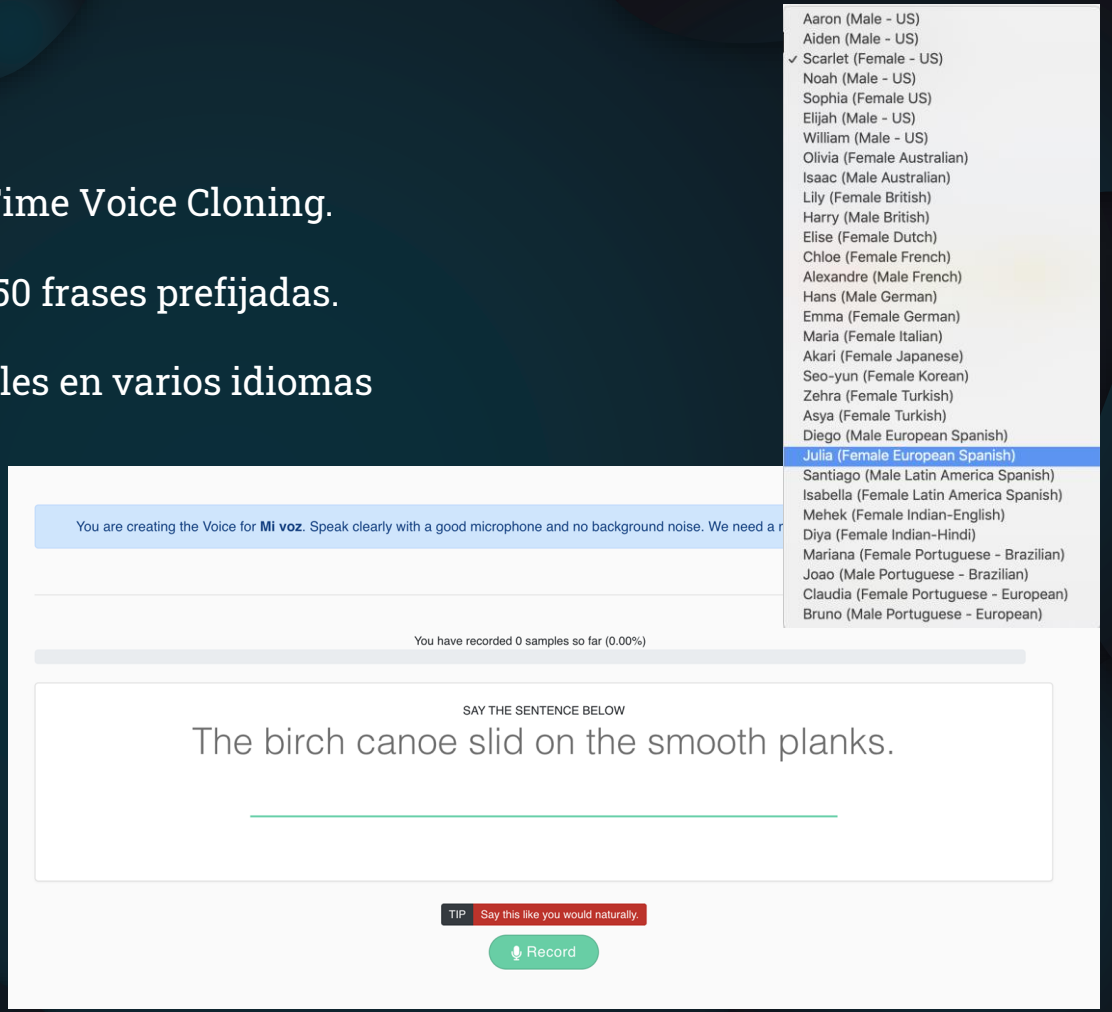
- Ejecutar `python demo_cli.py`.
  a. Escribir ruta del fichero con el audio.
  b. Escribir frase a clonar.

```
$ python demo_cli.py

Found 1 GPUs available. Using GPU 0 (GeForce GTX 1080) of compute capability 6.1 with 8.6Gb total
memory.

Preparing the encoder, the synthesizer and the vocoder...
Loaded encoder "pretrained.pt" trained to step 1564501
Found synthesizer "pretrained" trained to step 278000
Building Wave-RNN
Trainable Parameters: 4.481M
Loading model weights at vocoder\saved_models\pretrained\pretrained.pt
Testing your configuration with small inputs.
        Testing the encoder...
        Testing the synthesizer... (loading the model will output a lot of text)

Reference voice: enter an audio filepath of a voice to be cloned (mp3, wav, m4a, flac, ...):
trump.wav

Loaded file succesfully
Created the embedding
Write a sentence (+-20 words) to be synthesized:
That's one small step for a man, one giant leap for mankind.
Created the mel spectrogram
Synthesizing the waveform:
{| ████████████████ 66500/67200 | Batch Size: 7 | Gen Rate: 3.6kHz | }

Saved output as demo_output_00.wav
```

Shakespeare

Resemble.ai

- Aplicación similar a Real Time Voice Cloning.
- Clona la voz con al menos 50 frases prefijadas.
- Disponibles voces disponibles en varios idiomas (incluyendo español).
- Versión gratuita limitada.
- Editor avanzado de voces.



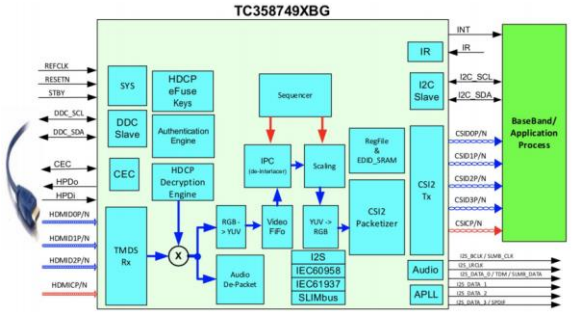
Casos de uso

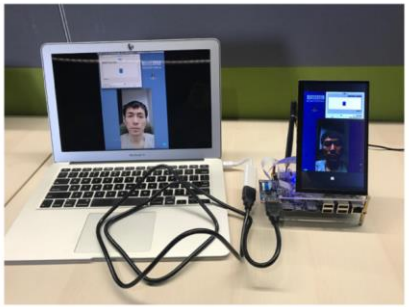
Audio Injection Example

(a) For Android Devices (b) For iOS Devices

Audio injection device based on analog circuits and sound card #BHUSA @BLACK

Video Injection Example



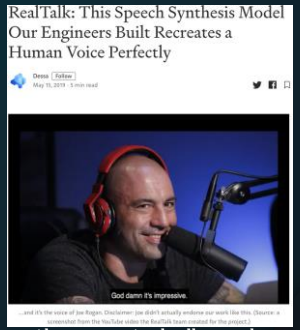


Video injection device based on TC358749XBG

#BHUSA @BLACK HAT EVENTS

https://i.blackhat.com/USA-19/Wednesday/us-19-Chen-Biometric-Authentication-Under-Threat-Liveness-Detection-Hacking.pdf

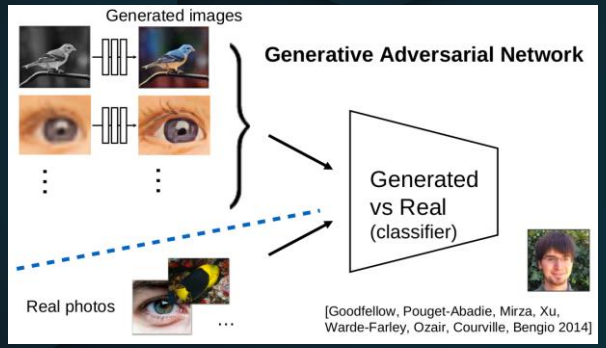
https://www.pandasecurity.com/mediacenter/news/deepfake-voice-fraud/

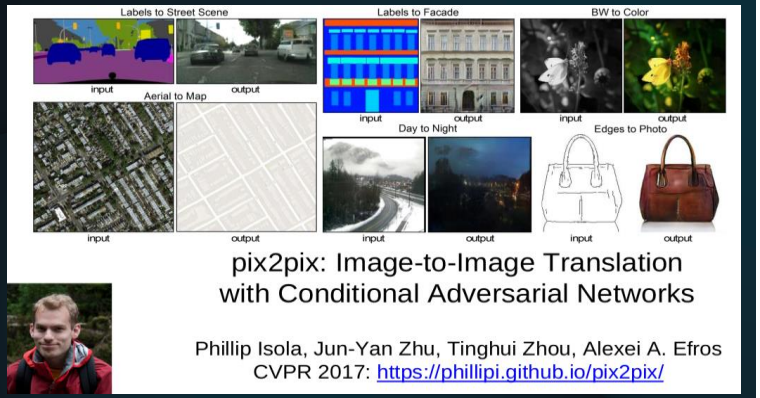


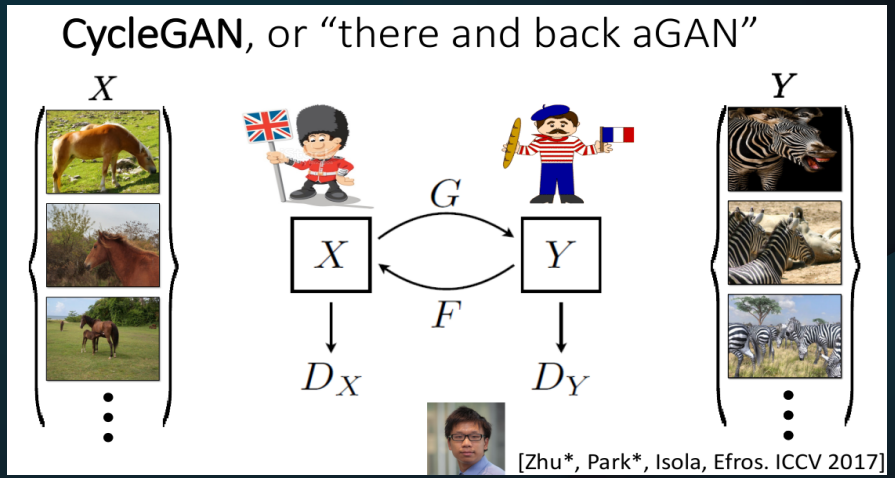
https://medium.com/dessa-news/real-talk-speech-synthesis-5dd0897eef7f

Generación de imágenes

https://tenso.rs/demos/fast-neural-style/







Requisitos y limitaciones

- Entrenar es **computacionalmente muy costoso**, aún con el uso de GPUs.

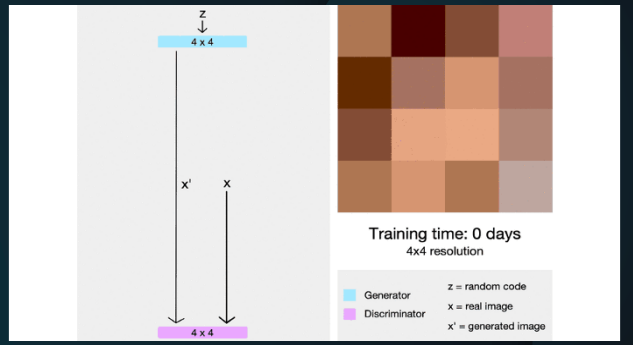
Configuration	Resolution	Total kimg	1 GPU	2 GPUs	4 GPUs	8 GPUs	GPU mem
`config-f`	1024×1024	25000	69d 23h	36d 4h	18d 14h	9d 18h	13.3 GB
`config-f`	1024×1024	10000	27d 23h	14d 11h	7d 10h	3d 22h	13.3 GB
`config-e`	1024×1024	25000	35d 11h	18d 15h	9d 15h	5d 6h	8.6 GB
`config-e`	1024×1024	10000	14d 4h	7d 11h	3d 20h	2d 3h	8.6 GB
`config-f`	256×256	25000	32d 13h	16d 23h	8d 21h	4d 18h	6.4 GB
`config-f`	256×256	10000	13d 0h	6d 19h	3d 13h	1d 22h	6.4 GB

- Adaptar el modelo a nuestras necesidades, requiere disponer de **una gran cantidad de datos**.

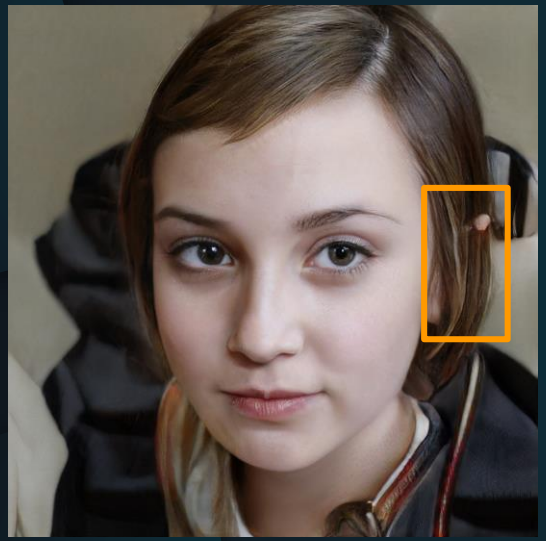
- Imágenes **poco realistas** en ocasiones.

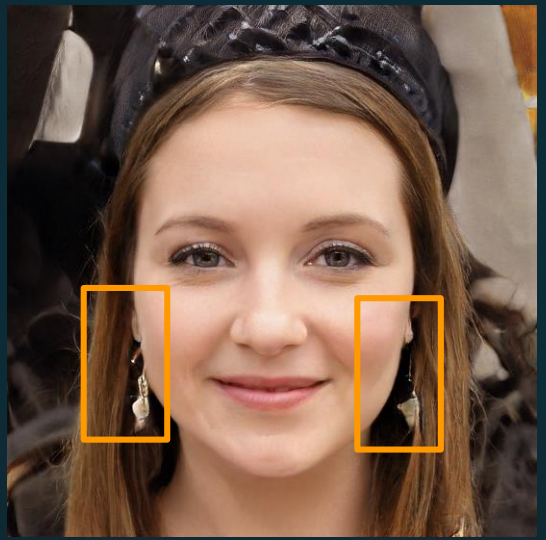
StyleGAN

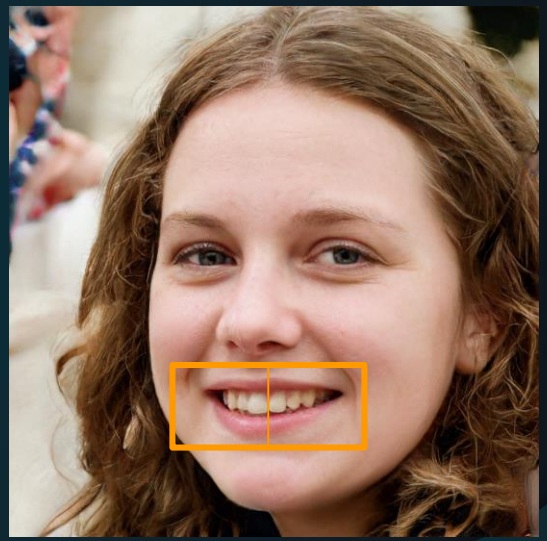
- Generación de **caras hiperrealistas**.
- Disponible en **GitHub** (https://github.com/NVlabs/stylegan).
- Entrenado con **70.000 imágenes** (con licencia no permisiva) de **Flickr**.
  - Flickr-Faces-HQ (https://github.com/NVlabs/ffhq-dataset).
  - Tamaño de 1024x1024 píxeles.
- Desarrollado por **NVIDIA Labs**.
- https://www.thispersondoesnotexist.com/ usa **StyleGAN** para la generación de caras.
- También aplicado a generación de **coches, habitaciones**,...



StyleGAN – limitaciones







StyleGAN2

- **Evolución** de StyleGAN.
- **Soluciona los problemas** de la versión anterior.
- Disponible en **GitHub** (https://github.com/NVlabs/stylegan2).

StyleGAN2

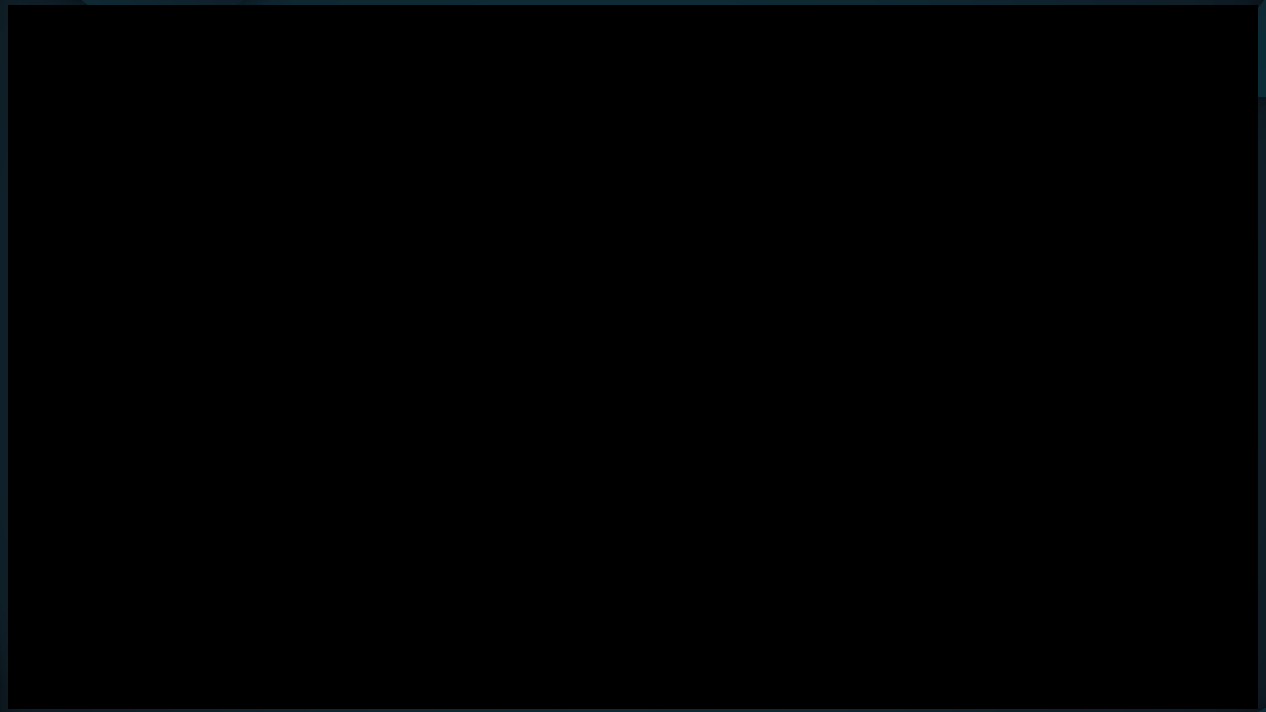
```
$ python run_generator.py generate-images --network=networks/stylegan2-ffhq-config-f.pkl \
      --seeds=6600-6605 --truncation-psi=0.5
```

```
Local submit - run_dir: results\00179-generate-images
dnnlib: Running run_generator.generate_images() on
localhost...
Loading networks from "networks/stylegan2-ffhq-config-
f.pkl"...
Setting up TensorFlow plugin "fused_bias_act.cu":
Preprocessing... Loading... Done.
Setting up TensorFlow plugin "upfirdn_2d.cu":
Preprocessing... Loading... Done.
[...]
Generating image for seed 6600 (0/6) ...
Generating image for seed 6601 (1/6) ...
Generating image for seed 6602 (2/6) ...
Generating image for seed 6603 (3/6) ...
Generating image for seed 6604 (4/6) ...
Generating image for seed 6605 (5/6) ...
dnnlib: Finished run_generator.generate_images() in 23s.
```






StyleGAN2 – Proyector



Generación de imagen – casos de uso

REDES SOCIALES

Las cuentas falsas de Facebook ahora también usan rostros de personas que ni siquiera existen

Por Donie O'Sullivan
18:52 ET(23:52 GMT) 23 Diciembre, 2019

PARIS MARTINEAU BUSINESS 12.20.2019 08:21 PM

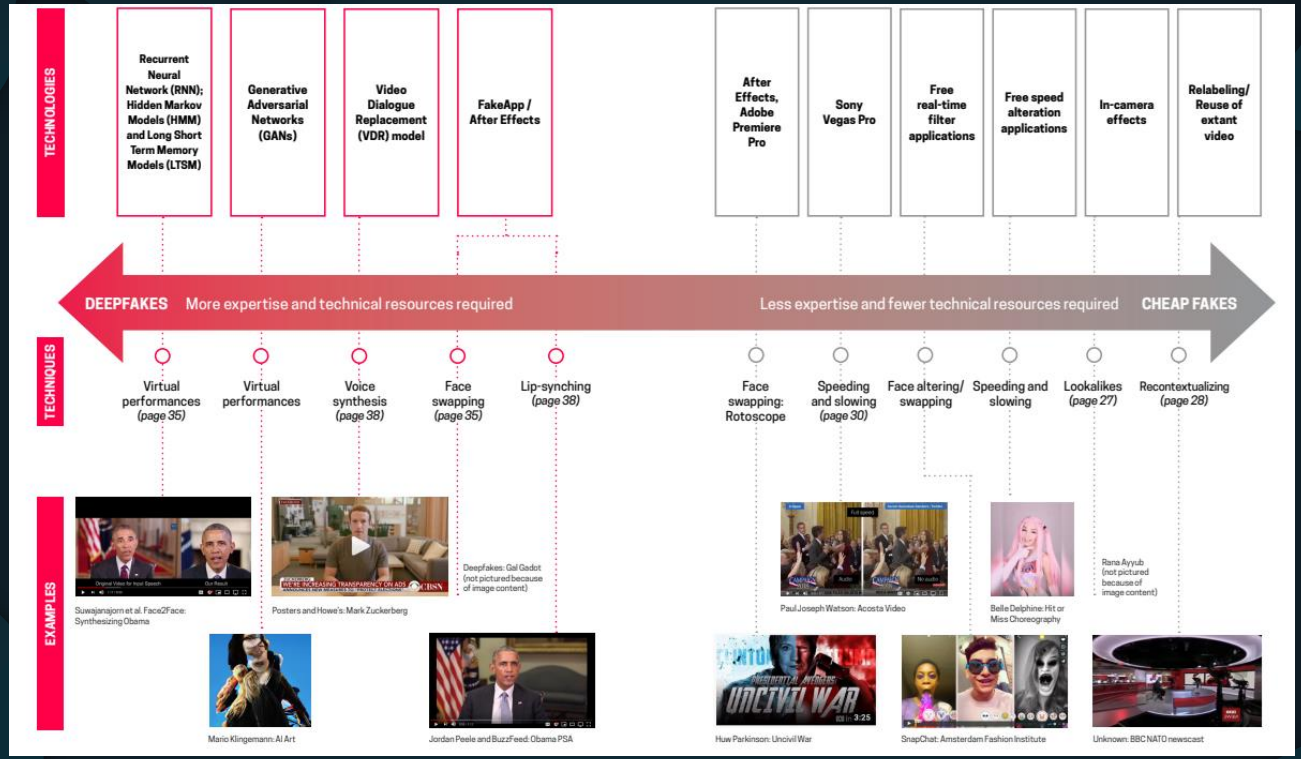
Facebook Removes Accounts With AI-Generated Profile Photos

Researchers said it appears to be the first use of artificial intelligence to support an inauthentic social media campaign.

Generación de vídeo

- DeepFake Obama
- Modelo 3D de la cara
  - https://dpo.si.edu/blog/smithsonian-creates-first-ever-3d-presidential-portrait
- Información OSINT de Obama en discursos.
- Generación de la voz sintética.
  - https://lyrebird.ai
- Generación de modelo de la boca.
  - https://nips2017creativity.github.io/doc/ObamaNet.pdf
- Generación del vídeo completo.

Generación de vídeo



https://datasociety.net/wp-content/uploads/2019/09/DS_Deepfakes_Cheap_FakesFinal.pdf

Generación de vídeo – Limitaciones

- Es importante diferenciar si lo que se quiere es generar algo nuevo o alterar uno existente (predominante).
- Necesidad de una gran **cantidad de datos** de la persona.
- **Diversos ángulos, iluminación y expresiones** para un resultado realista.
- **Costoso en tiempo y en infraestructura** para generar la salida.
- Si no se utilizan muchas muestras en la fase de entrenamiento los resultados son **pobres y fácilmente detectables**.

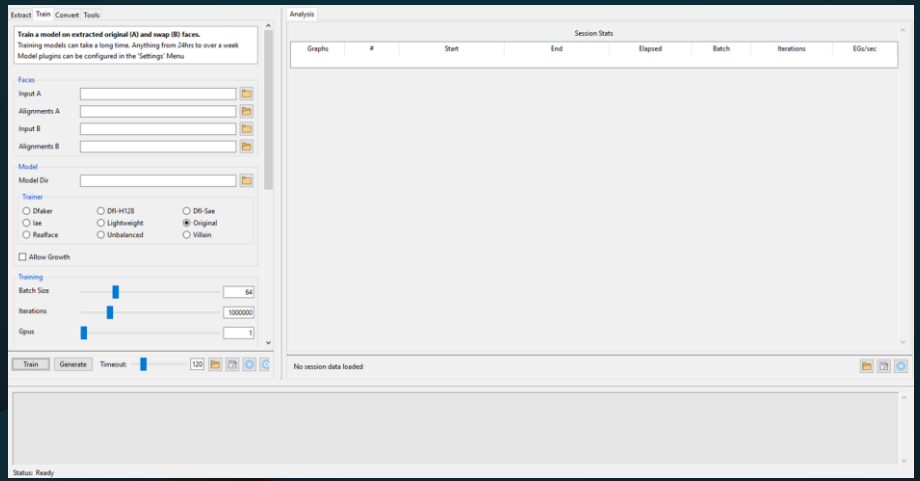
Generación de vídeo

- **Herramientas open source automatizadas**.
  - ***FaceSwap*** *(https://github.com/MarekKowalski/FaceSwap)*

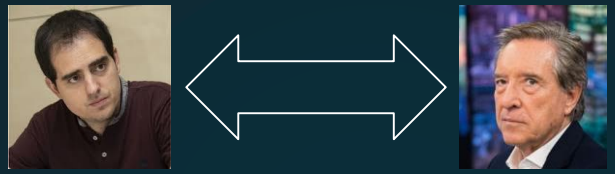
  - ***Face2Face*** *(https://github.com/datitran/face2face-demo)*

DeepFakes – FaceSwap

- Herramienta **OpenSource** para generar **Deep Fakes**.
- Disponible en **GitHub** (https://github.com/deepfakes/faceswap).
- Desarrollado en **Python**.
- Disponible tanto como **CLI** como **GUI**.



DeepFakes – Demo



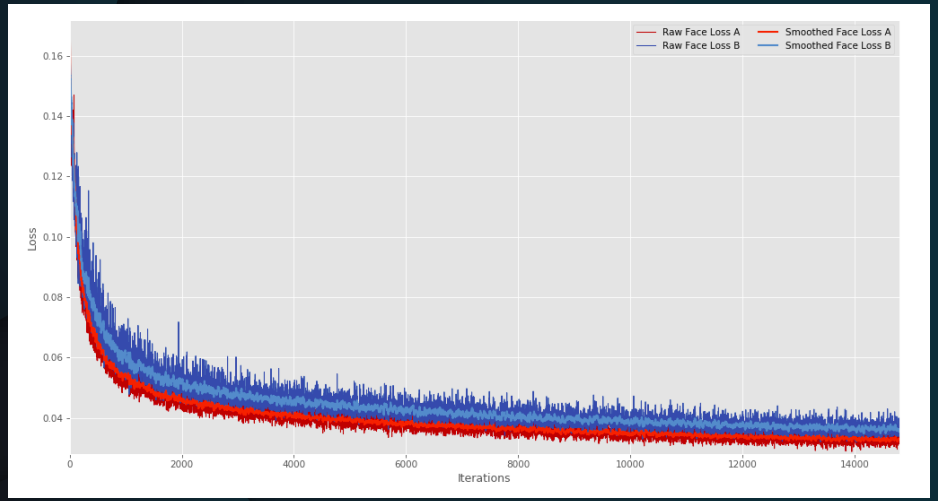
Minutos de vídeo **40**

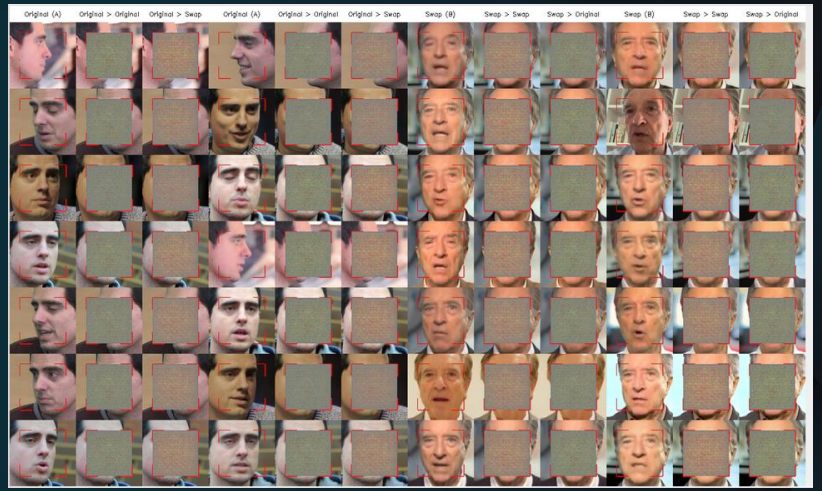
1.5 Días de entrenamiento (GTX 1080)

Frames Iñaki Gabilondo **21818**

8924 Frames Alfonso

DeepFakes – Demo



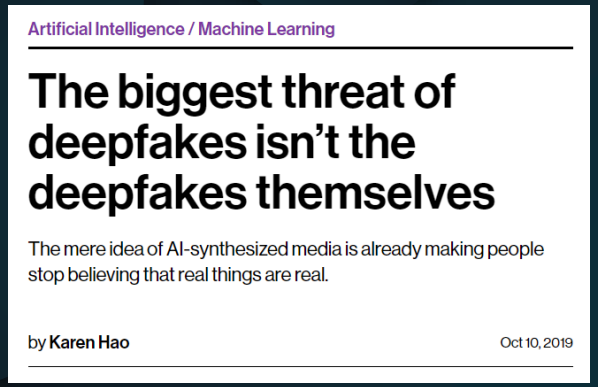


DeepFakes – FaceSwap

03
Detección de contenido sintético

Texto, audio, imagen y vídeo

¿Es un problema real?



Artificial Intelligence / Machine Learning

The biggest threat of deepfakes isn't the deepfakes themselves

The mere idea of AI-synthesized media is already making people stop believing that real things are real.

by Karen Hao Oct 10, 2019

https://www.technologyreview.com/s/614526/the-biggest-threat-of-deepfakes-isnt-the-deepfakes-themselves/

POLICY

DEEPFAKE PROPAGANDA IS NOT A REAL PROBLEM

We've spent the last year wringing our hands about a crisis that doesn't exist

By Russell Brandom | Mar 5, 2019, 12:25pm EST

https://www.theverge.com/2019/3/5/18251736/deepfake-propaganda-misinformation-troll-video-hoax

GILAD EDELMAN BUSINESS 01.07.2020 08:50 PM

Facebook's Deepfake Ban Is a Solution to a Distant Problem

The platform has a plan to deal with tomorrow's disinformation. But what about today's?

https://www.wired.com/story/facebook-deepfake-ban-disinformation/

¿Es un problema real? NO

2. Detecting Manipulated Content

- Claim: GAN-based methods are not our biggest problem (yet)
  - Still easy to tell, both by (trained) humans and (trained) computers
  - Likely because the “discriminator” always wins
  - Once discriminators start to loose, there might be nothing we could do ☹
- Our strategy: address the current biggest sources of visual fakes

https://slideslive.com/38917824/digital-image-manipulation-and-ways-to-detect-it

¿Es un problema real? SÍ

Photo: Ryan Lash / TED

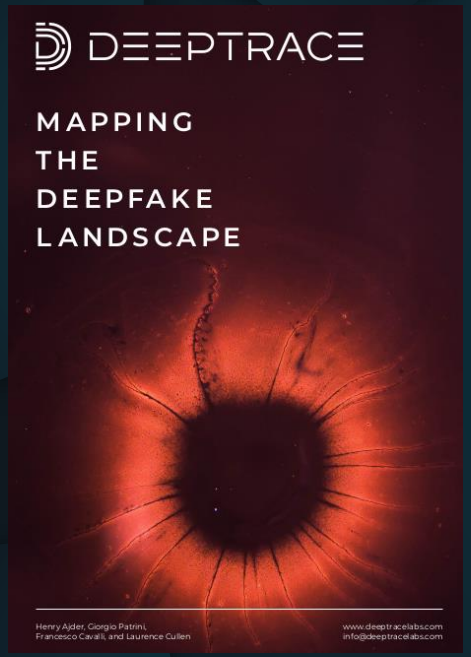
> *"Not only do we believe fakes, we are starting to doubt the truth," says Danielle Citron. Deepfakes pose a threat to the truth and democracy. (July 23, 2019, in Edinburgh, Scotland)*

https://www.daniellecitron.com/ted-talk

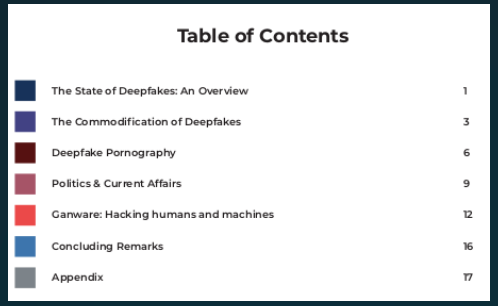
Sexual Privacy
Danielle Keats Citron

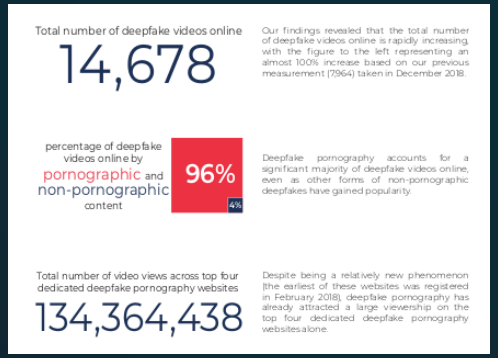
University of Maryland Francis King Carey School of Law
Legal Studies Research Paper
No. 2018-25

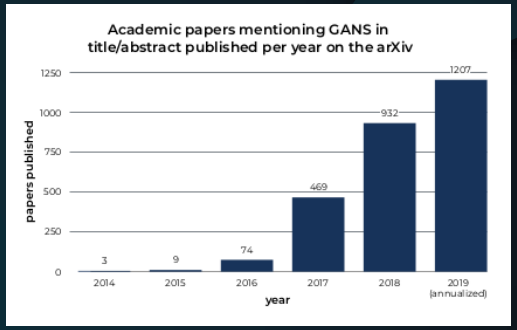
¿Es un problema real? Depende...



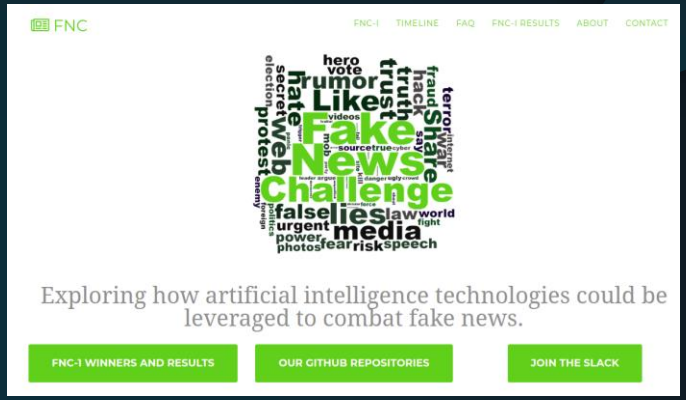
https://deeptracelabs.com/mapping-the-deepfake-landscape/

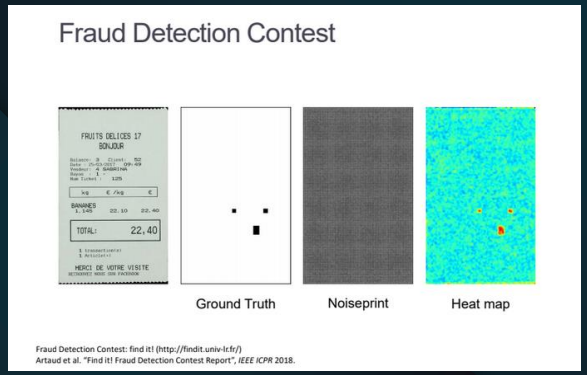




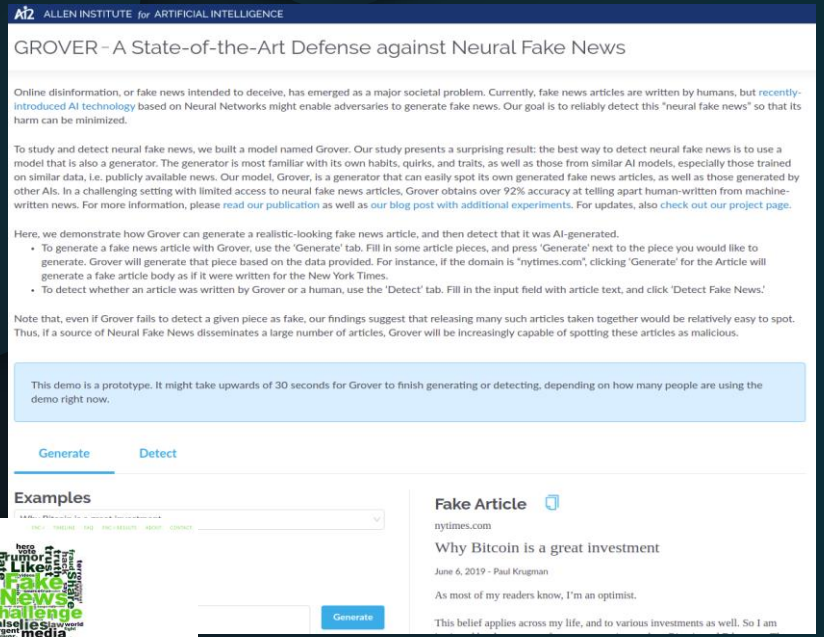


Detección de contenido sintético – challenges

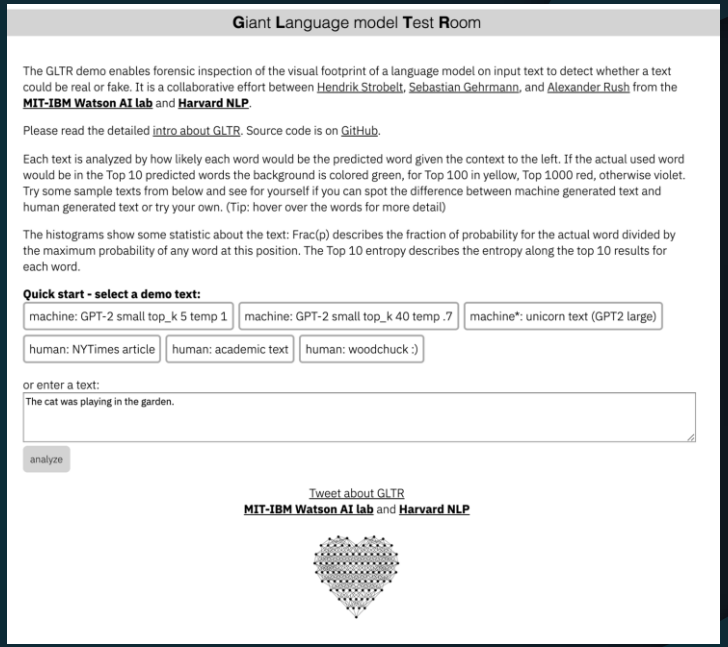




Detección de texto



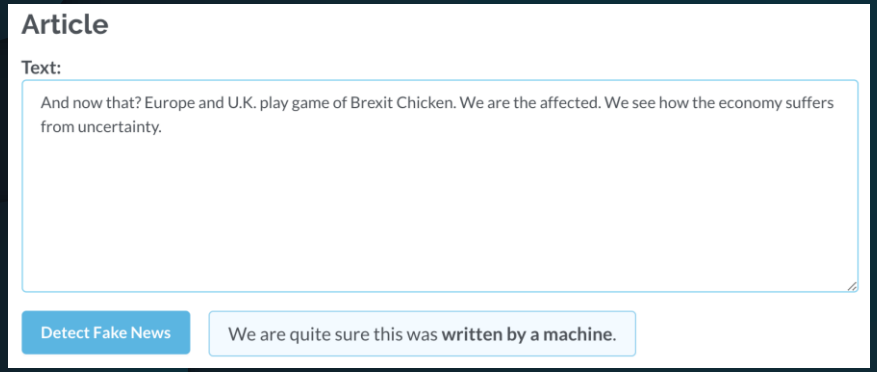
https://grover.allenai.org



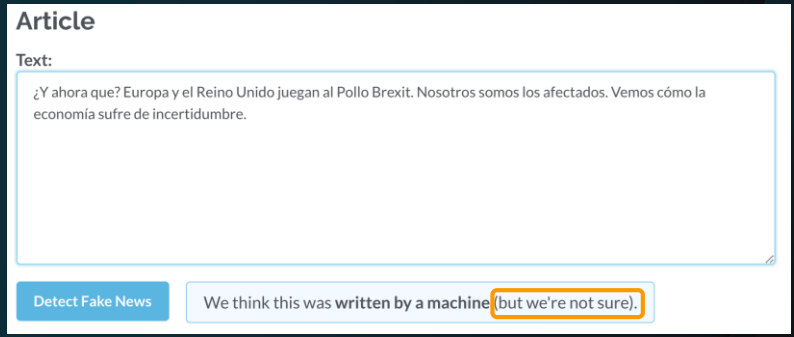
http://gltr.io/dist/index.html

Detección de texto

https://twitter.com/HarleyBarrettS1/status/1187975578076532739



Español



Detección de audio

DeepFake Audio Detection

With the popularity and capabilities of audio deep fakes on the rise, creating defenses against deep fakes used for malicious intent is becoming more important than ever.

We built a fake audio detection model with Foundations Atlas, for anyone to use. If you'd like to read more about why we decided to build this, click here.

Here are two examples of short audio clips in `./data/example_clips/` folder. One of them is real and the other is fake.

This repository provides the code for a fake audio detection model built using Foundations Atlas. It also includes a pre-trained model and inference code, which you can test on any of your own audio files.

Setup

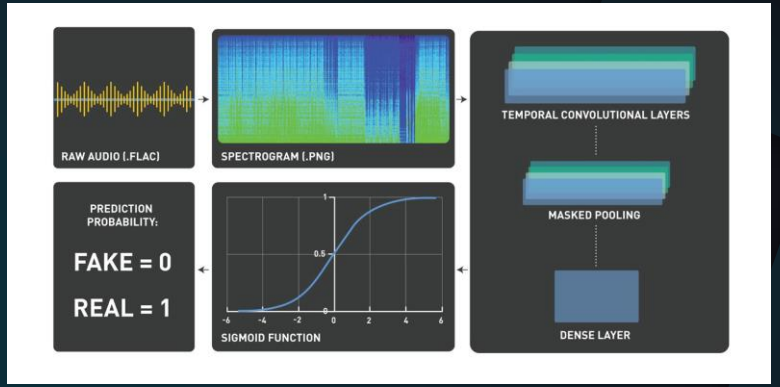
1. Git Clone this repository.
2. If you haven't already, install Foundations Atlas Community Edition.
3. Once you've installed Foundations Atlas, activate your environment if you haven't already, and navigate to your project folder, and then into the `code` directory.
4. Run the following line in the terminal to install the required packages:

```
pip install -r requirements.txt
```

That's it, You're ready to go!

Note: To run the code, your system should meet the following requirements: RAM >= 32GB , GPUs >=1

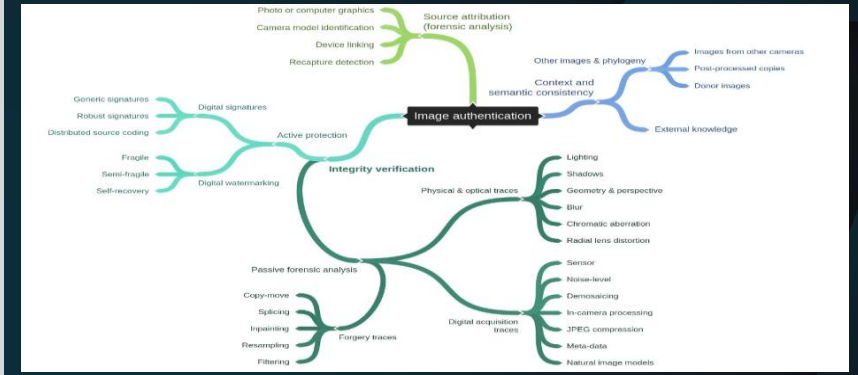
https://github.com/dessa-public/fake-voice-detection




```
$ python inference.py
Using TensorFlow backend.
Loading inference data from ../data/inference_data/unlabeled
Loading pretrained model
saved_model_240_8_32_0.05_1_50_0_0.0001_100_156_2_True_
True_fitted_objects.h5
100%|████████████████████████████████████████| 1/1
[00:00<00:00, 10.53it/s]
Using TensorFlow backend.

-----------------------------------------
[array([[-1.        , -1.        , -0.866223  , ..., -1.        ,
         -1.        , -1.        ],
        [-0.46634382, -0.42106295, -0.37900943, ..., -1.        ,
         -1.        , -1.        ],
        [ 0.05602419,  0.18866599,  0.23696363, ..., -1.        ,
         -1.        , -1.        ],
        ...,
        [-0.82941407, -0.66641605, -0.54652816, ..., -1.        ,
         -1.        , -1.        ],
        [-0.91284347, -0.7570157 , -0.6872368 , ..., -1.        ,
         -1.        , -1.        ],
        [-1.        , -1.        , -1.        , ..., -1.        ,
         -1.        , -1.        ]], dtype=float32)]
-----------------
[[0.17236993]]
The probability of the clip being real is: 17.24%
```

Detección de imagen (manual)

https://www.darpa.mil/news-events/2019-09-03a



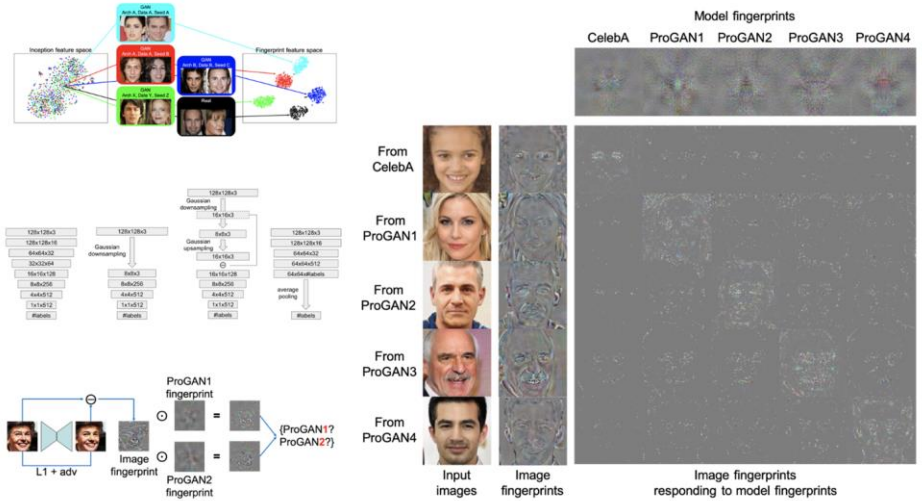
http://kt.agh.edu.pl/~korus/files/icml19.pdf

- A la hora de detectar este tipo de vídeos, los más **difíciles** para un humano son los que únicamente **cambian la expresión**. Las **orejas** no suelen ser iguales.
- Aparecen **pendientes diferentes** o en algunos casos desaparecen en una de ellas.
- Las **caras son asimétricas**, la mirada suele ser distinta para cada ojo.
- Los **dientes no son regulares**.
- El **fondo** de las fotos suele ser abstracto.

https://medium.com/@kcimc/how-to-recognize-fake-ai-generated-images-4d1f6f9a2842

Detección de imagen (automática)

GANFingerprints



https://arxiv.org/pdf/1811.08180.pdf

FakeSpotter: A Simple Baseline for Spotting AI-Synthesized Fake Faces

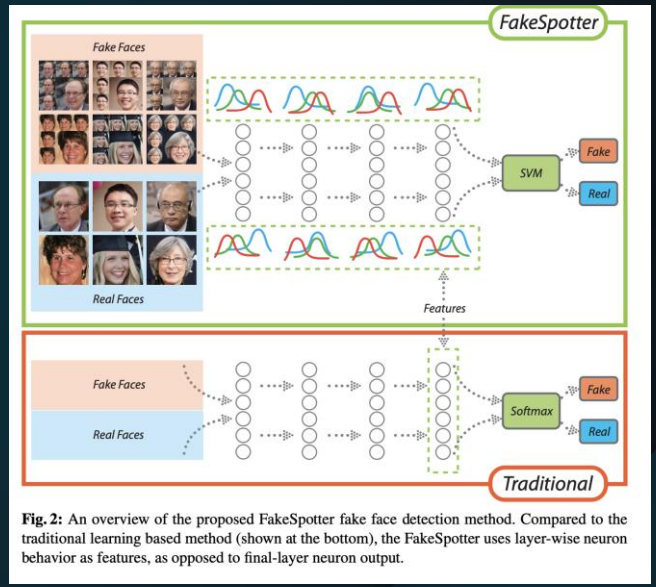


Fig. 2: An overview of the proposed FakeSpotter fake face detection method. Compared to the traditional learning based method (shown at the bottom), the FakeSpotter uses layer-wise neuron behavior as features, as opposed to final-layer neuron output.

https://arxiv.org/pdf/1909.06122.pdf

Detección de imagen (automática)

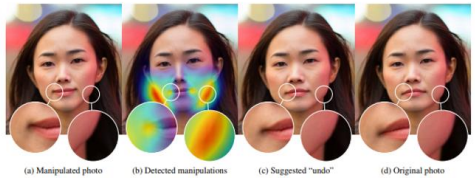
Detecting Photoshopped Faces by Scripting Photoshop

Sheng-Yu Wang[1] Oliver Wang[2] Andrew Owens[1] Richard Zhang[2] Alexei A. Efros[1]

UC Berkeley[1] Adobe Research[2]

https://peterwang512.github.io/FALdetector/

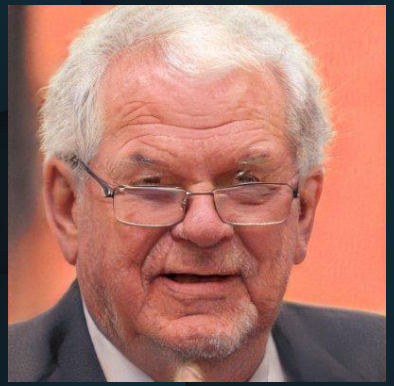
BAIR
BERKELEY ARTIFICIAL INTELLIGENCE RESEARCH

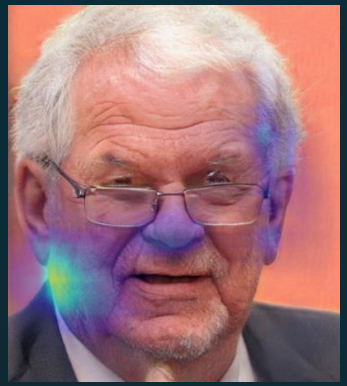


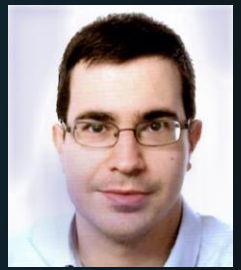
https://arxiv.org/pdf/1906.05856.pdf

https://github.com/peterwang512/FALdetector

Detección de imagen (automática)

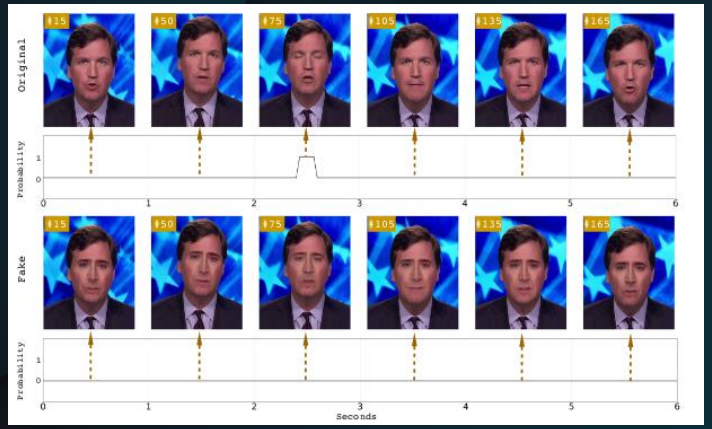


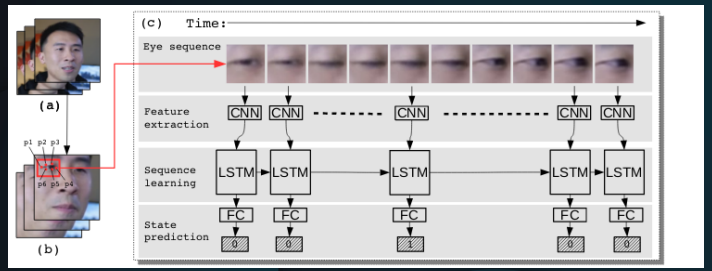




Probability being modified by Photoshop FAL: 10.70%

Detección de vídeo





https://www.albany.edu/~yl149995/papers/wifs2018.pdf

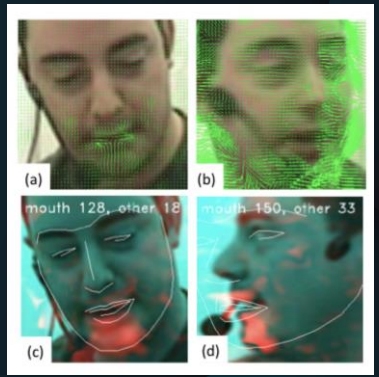
Spotting Audio-Visual Inconsistencies (SAVI) in Manipulated Video

Robert Bolles, J. Brian Burns, Martin Graciarena, Andreas Kathol, Aaron Lawson, Mitchell McLaren
SRI International

Thomas Mensink
University of Amsterdam

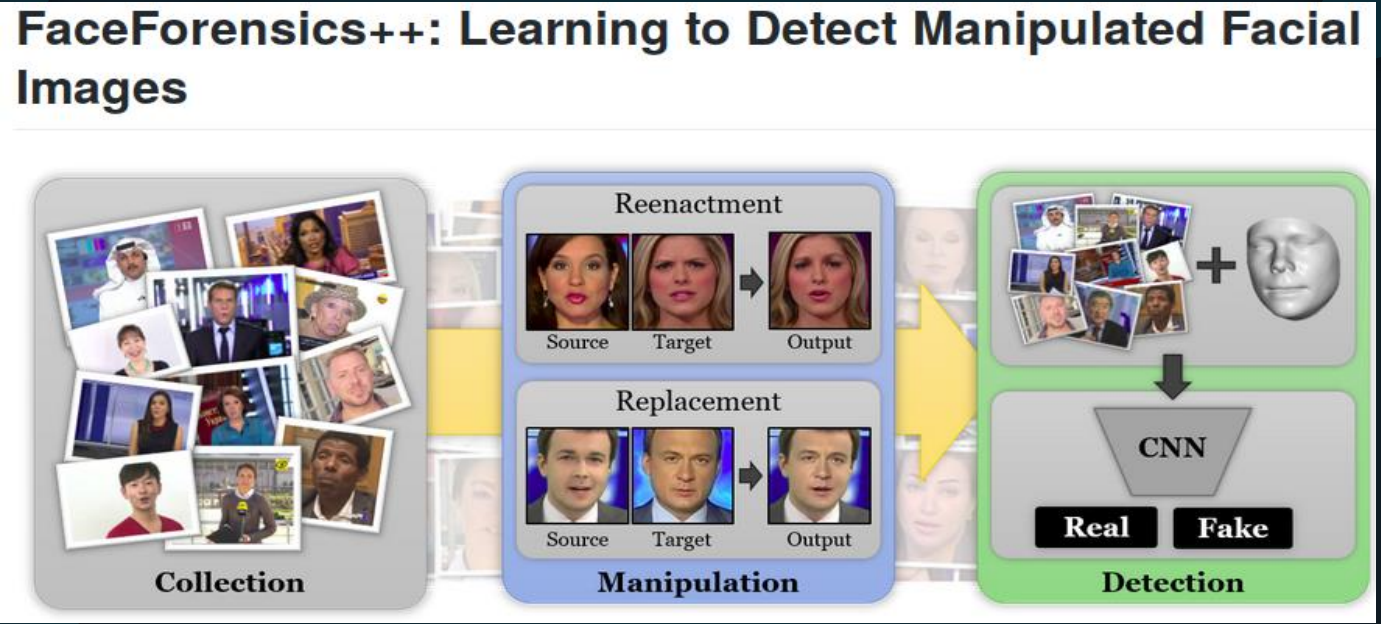
Type of inconsistency	*Visual Features*	*Audio Features*
Environmental class	**Scene: indoor vs. outdoor, small room, etc.**	**Environmental classes, reverberation of closed vs. open spaces**
Speaker identity	**Face recognition**	**Speaker ID**
Lip movement	Pattern of lip movement	Speech patterns
Head movement	Change in head pose	Left-right channel balance
AV device movement	Motion relative to scene	Changes in environmental features
Missing sound	Presence of sound-producing activity	Presence of corresponding sound
Middle-level features	Visual scene signature	Audio scene signature

Table 1. Different types of audiovisual inconsistencies to be detected and characterized in SAVI. The aspects explored in this paper are shown in bold.



https://staff.fnwi.uva.nl/t.e.j.mensink/publications/bolles17cvprwmf.pdf

FaceForensics

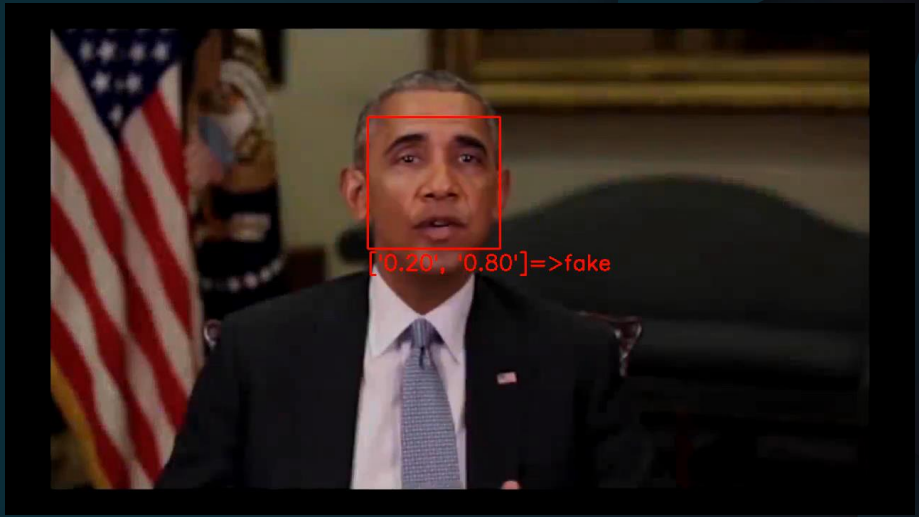


https://github.com/ondyari/FaceForensics

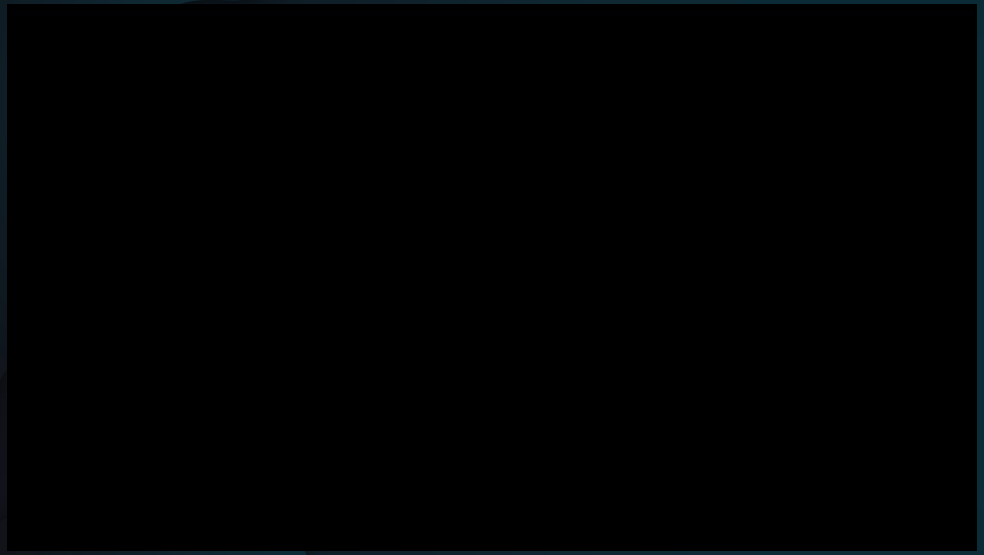
FaceForensics

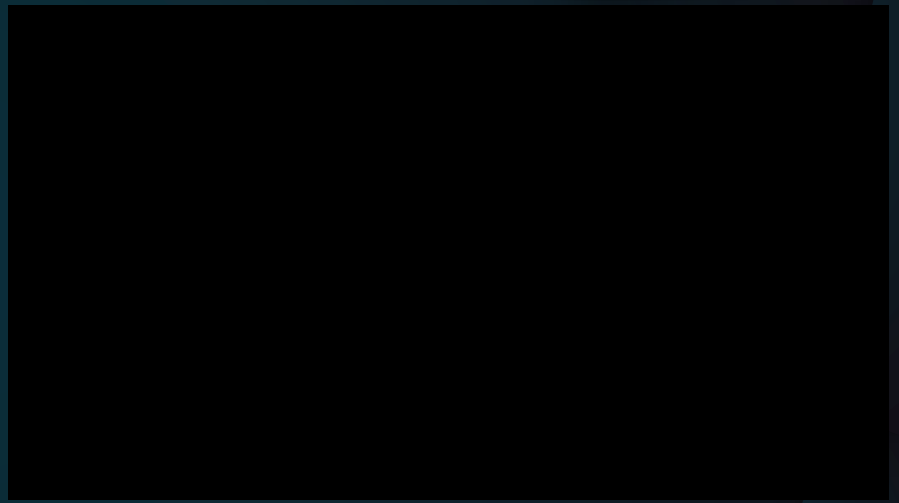
- FaceForensics es una herramienta de **detección de DeepFakes**.
- Calcula la **probabilidad de que cada frame sea falso** o no.
- Disponible en **GitHub** (https://github.com/ondyari/FaceForensics).
- Entrenado con **vídeos de YouTube**.
- Conjunto de entrenamiento disponible **bajo solicitud**.
- Sólo detecta **3 modelos de Deep Fakes** (datos de entrenamiento).
- **Difícil de instalar**. Documentación incompleta.

FaceForensics – demo



FaceForensics – demo





FakeVideoForensics

Usage

```
usage: main.py [-h] [--model_path MODEL] [--output_path VIDEOOUT]
               [--start_frame START_FRAME] [--end_frame END_FRAME] [--cuda]
               [--fast] --video_path VIDEOIN

optional arguments:
  -h, --help            show this help message and exit
  --model_path MODEL, -mi MODEL
  --output_path VIDEOOUT, -o VIDEOOUT
  --start_frame START_FRAME
  --end_frame END_FRAME
  --cuda
  --fast

required arguments:
  --video_path VIDEOIN, -i VIDEOIN
```

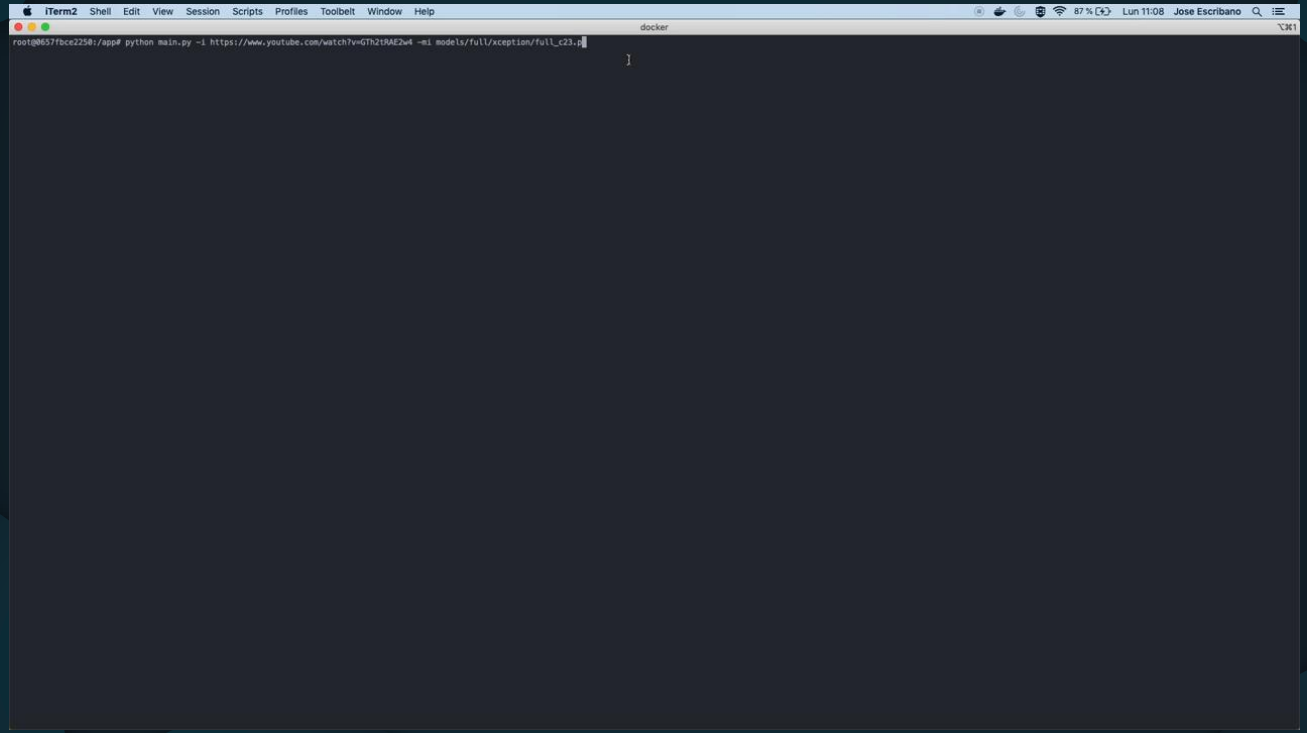
Output

```
user@host:/app# python3 main.py -i https://www.youtube.com/watch?v=GTh2tRAE2w4 -mi models/full/xcep
[youtube] GTh2tRAE2w4: Downloading webpage
[youtube] GTh2tRAE2w4: Downloading video info webpage
WARNING: Unable to extract video title
[youtube] GTh2tRAE2w4: Downloading MPD manifest
[download] video.mp4 has already been downloaded
[download] 100% of 1.75MiB
video.mp4
Starting: video.mp4
Model found in models/full/xception/full_c23.p
100%|████████████████████████████████████████████████████████████████████████████████
The Fake Score is: 90.05763688760807%
Output video in: video.avi
```

https://github.com/next-security-lab/fakeVideoForensics

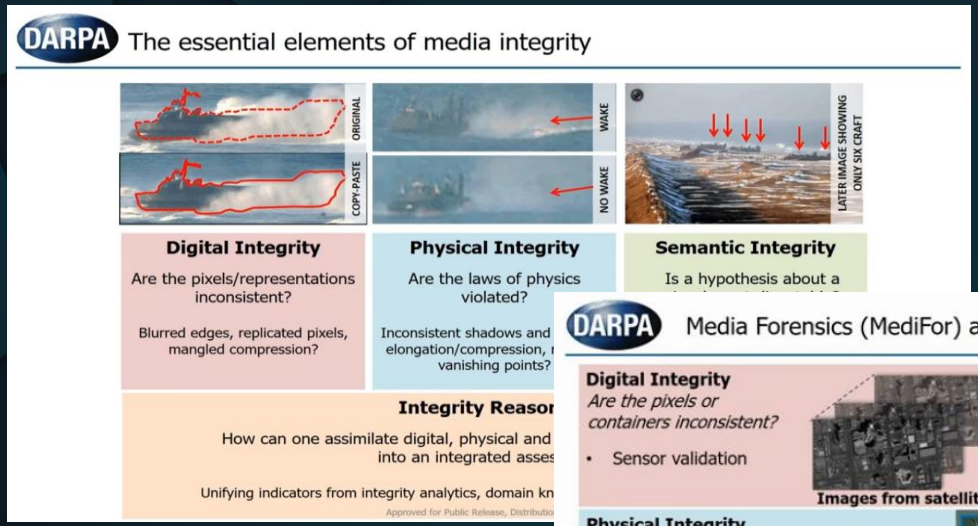
FakeVideoForensics – demo

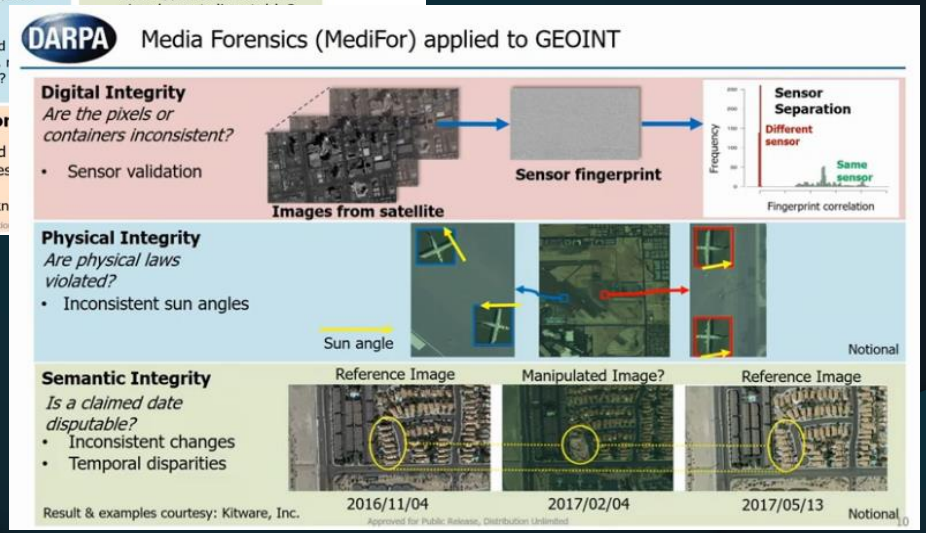
https://github.com/next-security-lab/fakeVideoForensics



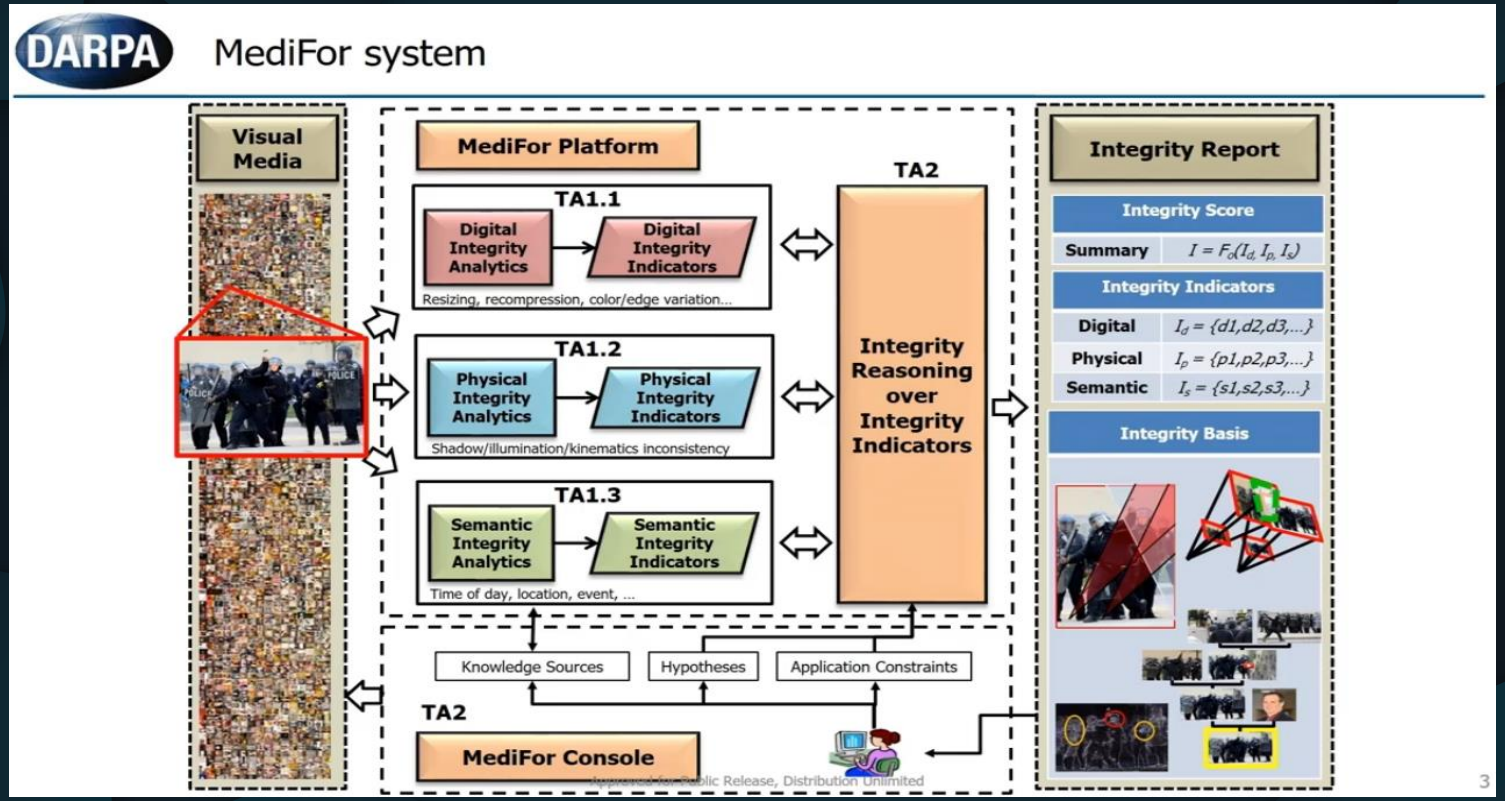
04
Futuro y conclusiones

Futuro


DARPA
The essential elements of media integrity
ORIGINAL
COPY-PASTE
WAKE
NO WAKE
LATER IMAGE SHOWING ONLY SIX CRAFT
Digital Integrity
Are the pixels/representations inconsistent?
Blurred edges, replicated pixels, mangled compression?
Physical Integrity
Are the laws of physics violated?
Inconsistent shadows and
elongation/compression,
vanishing points?
Semantic Integrity
Is a hypothesis about a
Integrity Reasor
How can one assimilate digital, physical and
into an integrated asses
Unifying indicators from integrity analytics, domain kn
Approved for Public Release, Distributio

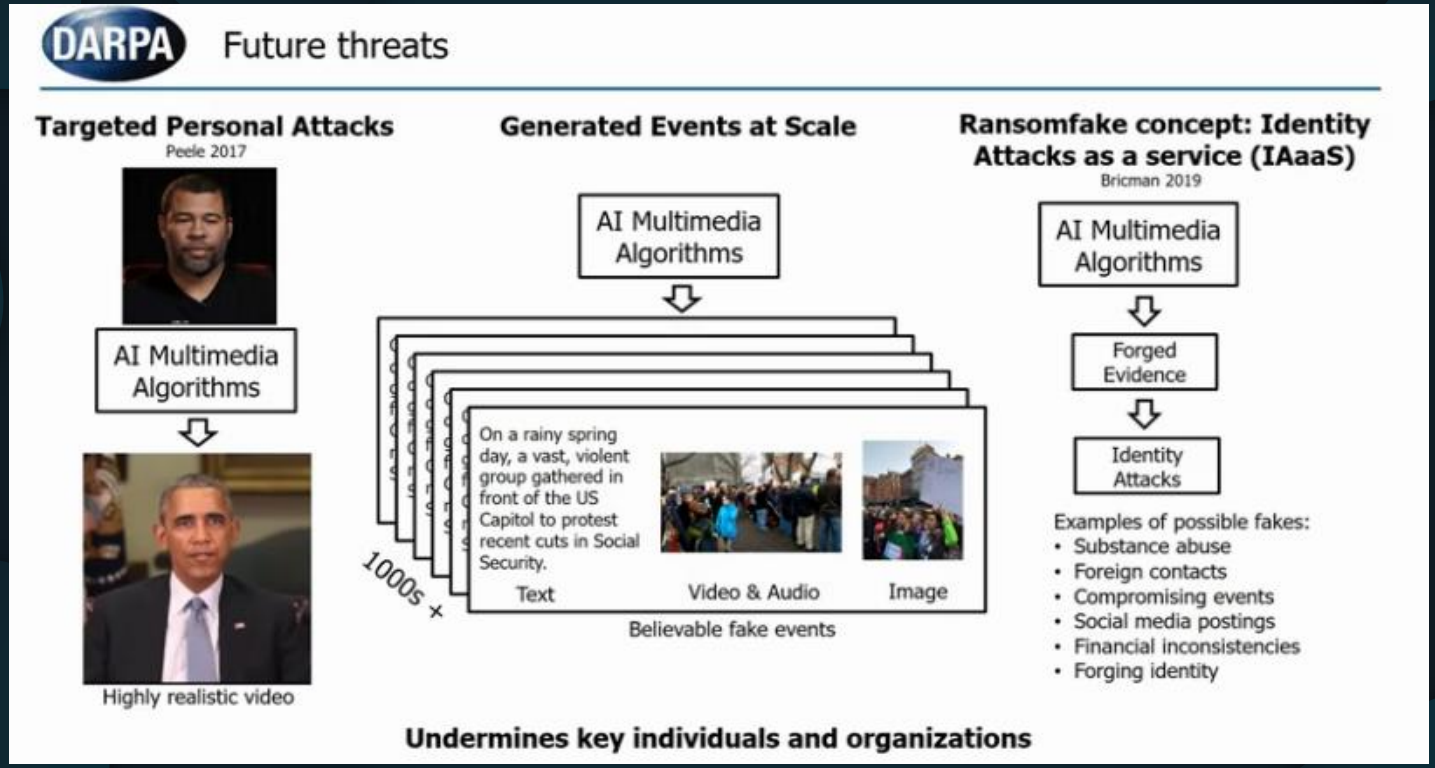

DARPA
Media Forensics (MediFor) applied to GEOINT
Digital Integrity
Are the pixels or containers inconsistent?
• Sensor validation
Images from satellite
Sensor fingerprint
Sensor Separation
Different sensor
Same sensor
Frequency
Fingerprint correlation
Physical Integrity
Are physical laws violated?
• Inconsistent sun angles
Sun angle
Notional
Semantic Integrity
Is a claimed date disputable?
• Inconsistent changes
• Temporal disparities
Reference Image
Manipulated Image?
Reference Image
2016/11/04
2017/02/04
2017/05/13
Notional
Result & examples courtesy: Kitware, Inc.
Approved for Public Release, Distribution Unlimited
10

Futuro



https://slideslive.com/38917820/media-forensics-challenges-beyond-the-current-state-of-the-art
https://github.com/mediaforensics/medifor

Futuro



https://slideslive.com/38917820/media-forensics-challenges-beyond-the-current-state-of-the-art

Conclusiones

- Vivimos en una **sociedad conectada**, cada vez con más medios y más contenido y es una tendencia creciente.
- Realizar contenido sintético es cada día **más sencillo**, pero hacerlo **realmente creíble no**. Son necesarios datos, procesamiento y conocimientos técnicos.
- Es **necesario correlar** siempre la información que vemos en Internet.
- ¿Existe solución milagrosa? Nosotros creemos que no, pero siempre existirán indicadores que nos permitan saber la veracidad de la información que recibimos.

Herramientas de generación y detección

https://github.com/next-security-lab/tools-generation-detection-synthetic-content

@ joseignacio.escribano.pablos.next@bbva.com

@ alfonso.munoz2.next@bbva.com | alfonso@criptored.com

¡Gracias por vuestra (*fake*) atención!