

# Text-to-image models

**Enrico Benedetti, Filippos Papandreou & Simona Scala**

# Intro

"A brain riding a rocketship heading towards the moon."

text-to-image  
model

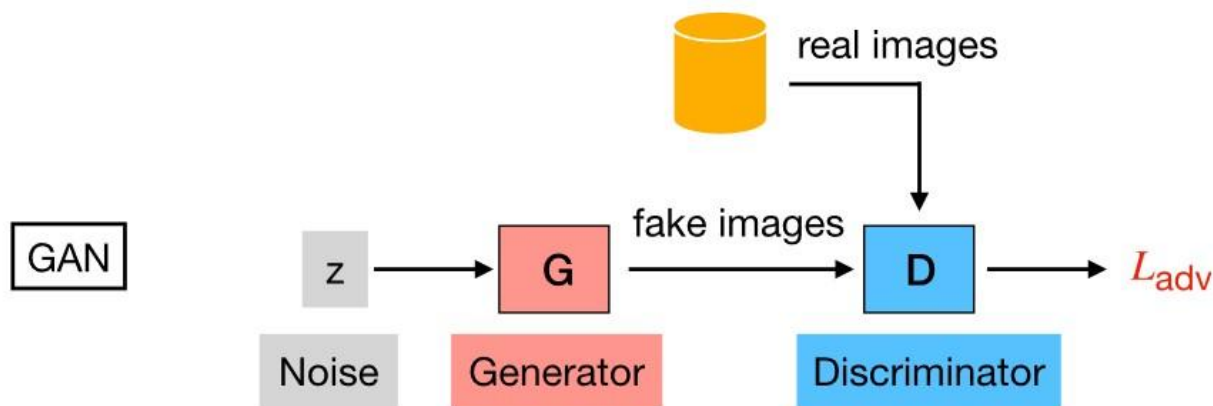


From "Photorealistic Text-to-Image Diffusion Models with Deep Language Understanding", Saharia et al.

# Generative Adversarial Networks (GANs)

Interplay between two neural networks

- Generator
- Discriminator



From "Improved techniques for training GANs", Salimans et al.

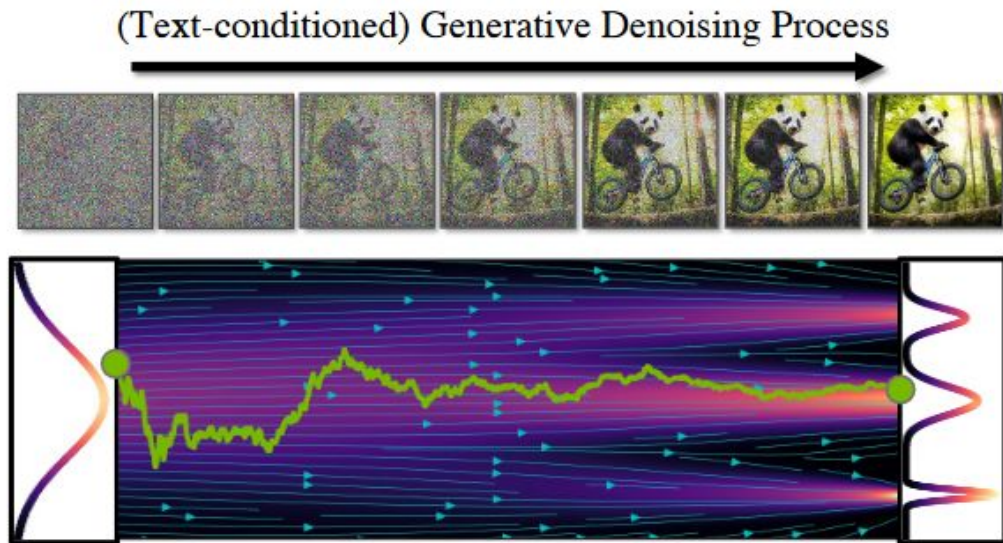
## Conditional approach for text-to-image synthesis using GANs

The typical architecture for text-to-image synthesis using GANs consists of

- Text encoder
- Image generator
- Discriminator
- Training Process
- Datasets

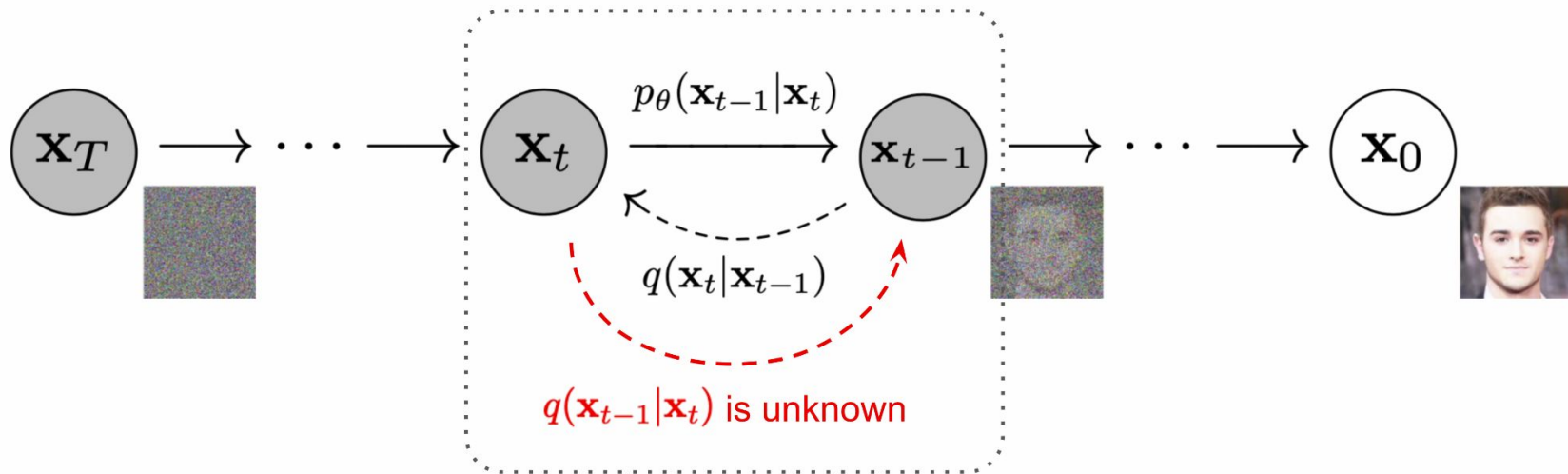
# Diffusion models

- Start with white noise
- Learn how to remove some at each step
- End with an image from the data distribution



from “eDiff-I: Text-to-Image Diffusion Models with an Ensemble of Expert Denoisers”, Balaji et al.

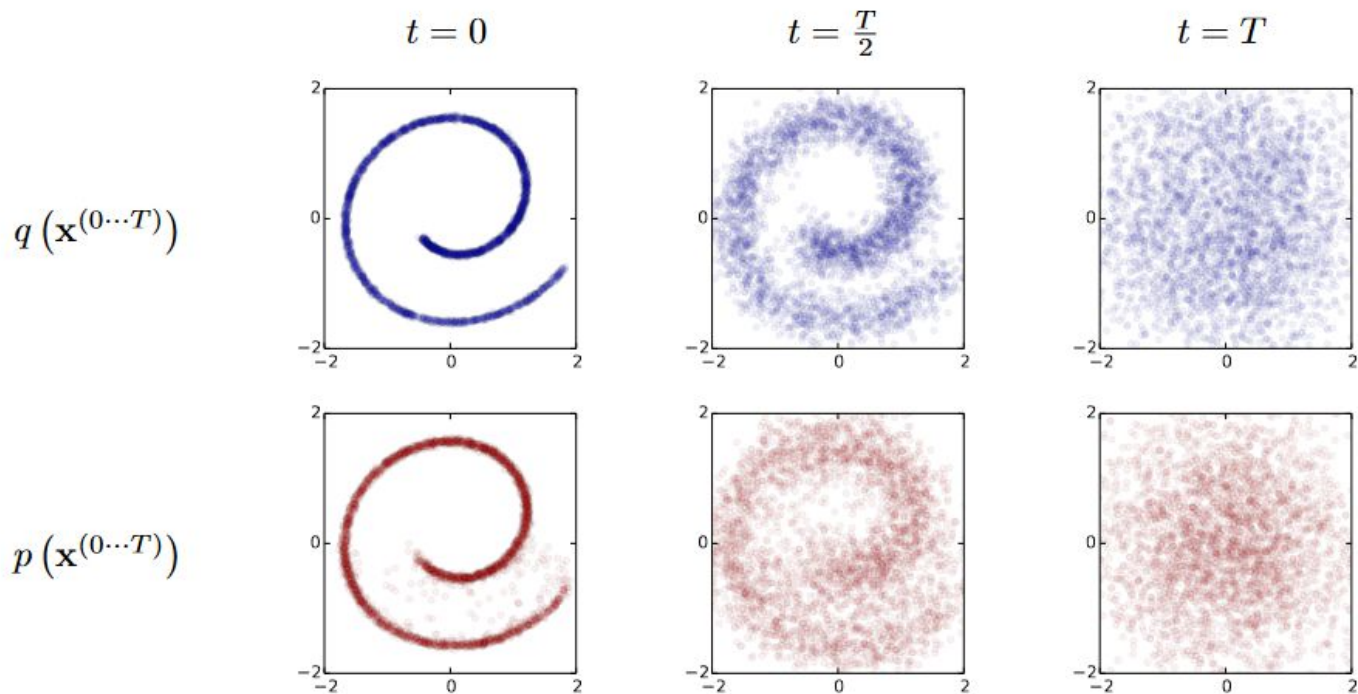
# Forward & backward processes



$$L_{\text{simple}} = E_{t, x_0, \epsilon} [||\epsilon - \epsilon_\theta(x_t, t)||^2]$$

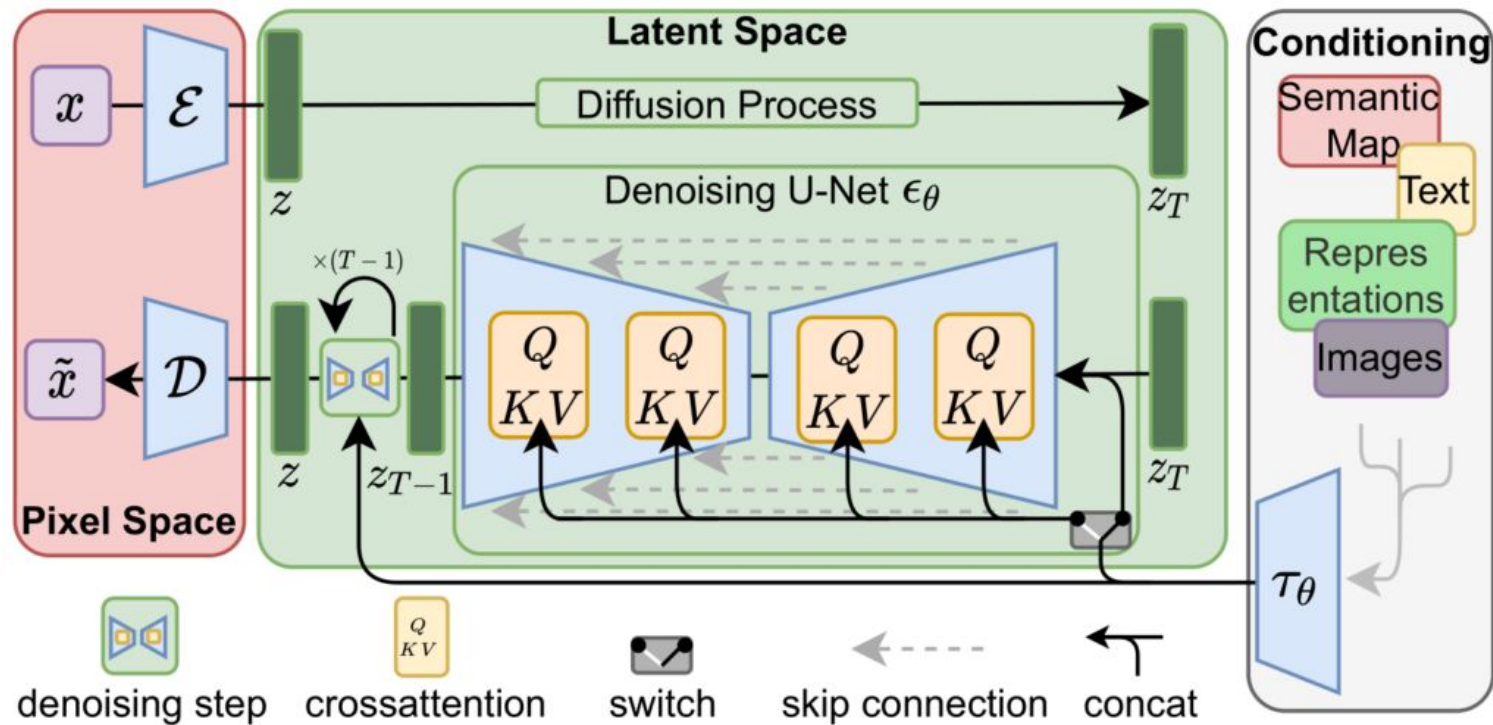
from “Denoising Diffusion Probabilistic Models”, Ho et al.

# Forward & backward processes



from “Deep Unsupervised Learning using Nonequilibrium Thermodynamics” by Sohl-Dickstein et al.

# Example: latent diffusion models



from "High-Resolution Image Synthesis with Latent Diffusion Models", Rombach et al.



## In short, diffusion models...

- Are a hot way to do conditional image generation
- Achieve high quality and diversity, but the iterative steps make them slow
- Are based on noising and denoising processes
- The main ML foundations are the U-net and the attention mechanism

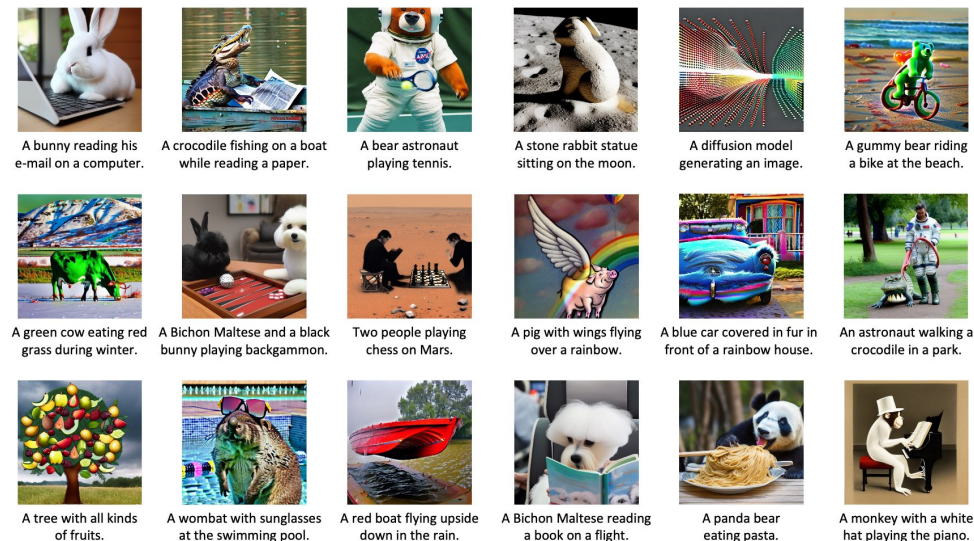
# Evaluation metrics

- Inception Score (IS)
  - Leverages a pre-trained Inception-v3 network: classification
  - Measures both recognizability and diversity of generated images
- Fréchet Inception Distance (FID)
  - Leverages a pre-trained Inception-v3 network: extraction
  - Measures the similarity between model-generated images and real images
- R-Precision (RP)
  - Compares the encoding vectors of the image and the caption using cosine similarity
  - Assesses the performance in matching image queries with textual descriptions
- Human evaluation
  - List of prompts to test image quality and fidelity to the caption

# GAN-based vs. Diffusion-based methods



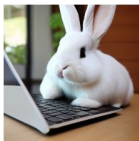
from “Adversarial text-to-image synthesis: A review”, Frolov et al.



from “Diffusion Models in Vision: A Survey”, Croitoru et al.

# Comparisons

- GAN-based methods
- Diffusion-based methods



A bunny reading his e-mail on a computer.



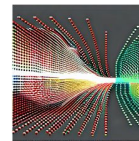
A crocodile fishing on a boat while reading a paper.



A bear astronaut playing tennis.



A stone rabbit statue sitting on the moon.



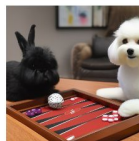
A diffusion model generating an image.



A gummy bear riding a bike at the beach.



A green cow eating red grass during winter.



A Bichon Maltese and a black bunny playing backgammon.



Two people playing chess on Mars.



A pig with wings flying over a rainbow.



A blue car covered in fur in front of a rainbow house.



An astronaut walking a crocodile in a park.



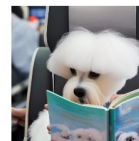
A tree with all kinds of fruits.



A wombat with sunglasses at the swimming pool.



A red boat flying upside down in the rain.



A Bichon Maltese reading a book on a flight.



A panda bear eating pasta.



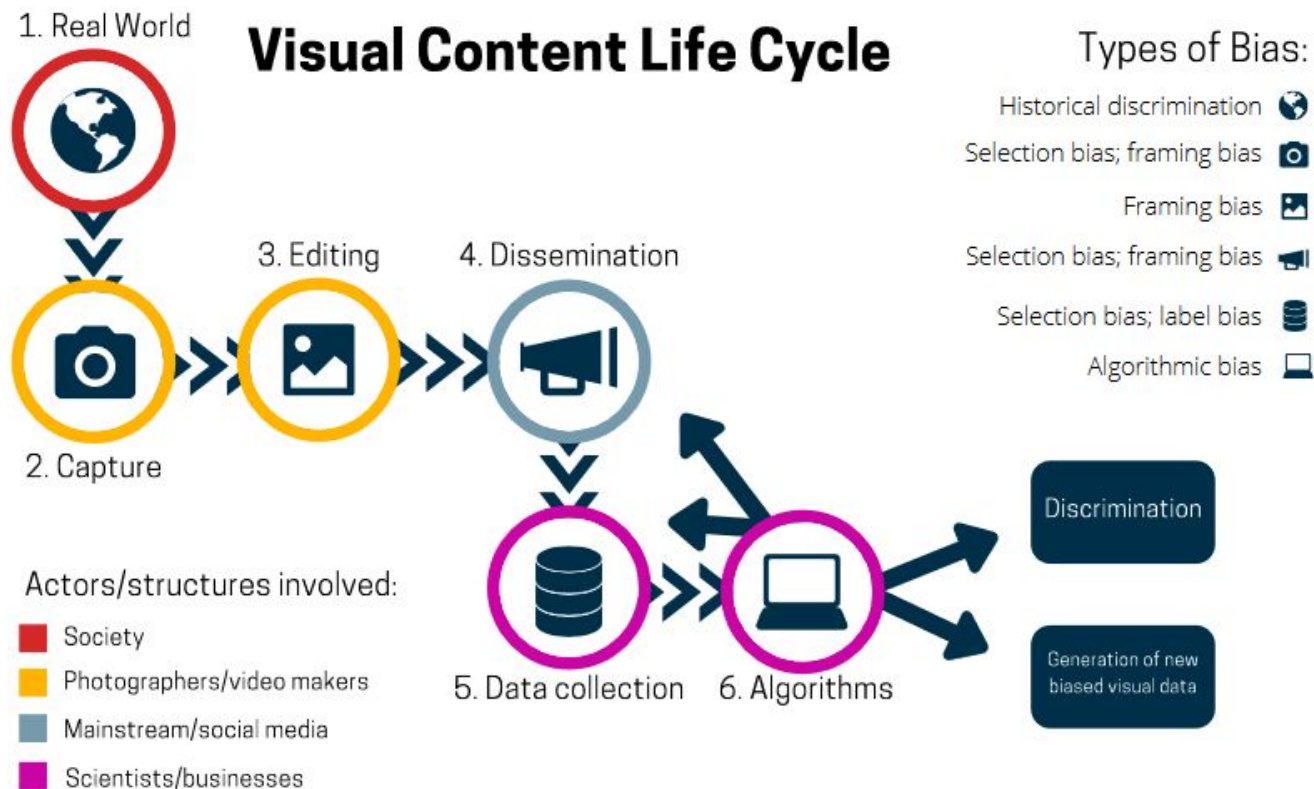
A monkey with a white hat playing the piano.

# Ethics

We should explore some ideas and issues...

- Bias
- Deepfakes & fake content
- Creativity, Art, Copyright

# Bias



from “A Survey on Bias in Visual Datasets”, Fabbri et al.

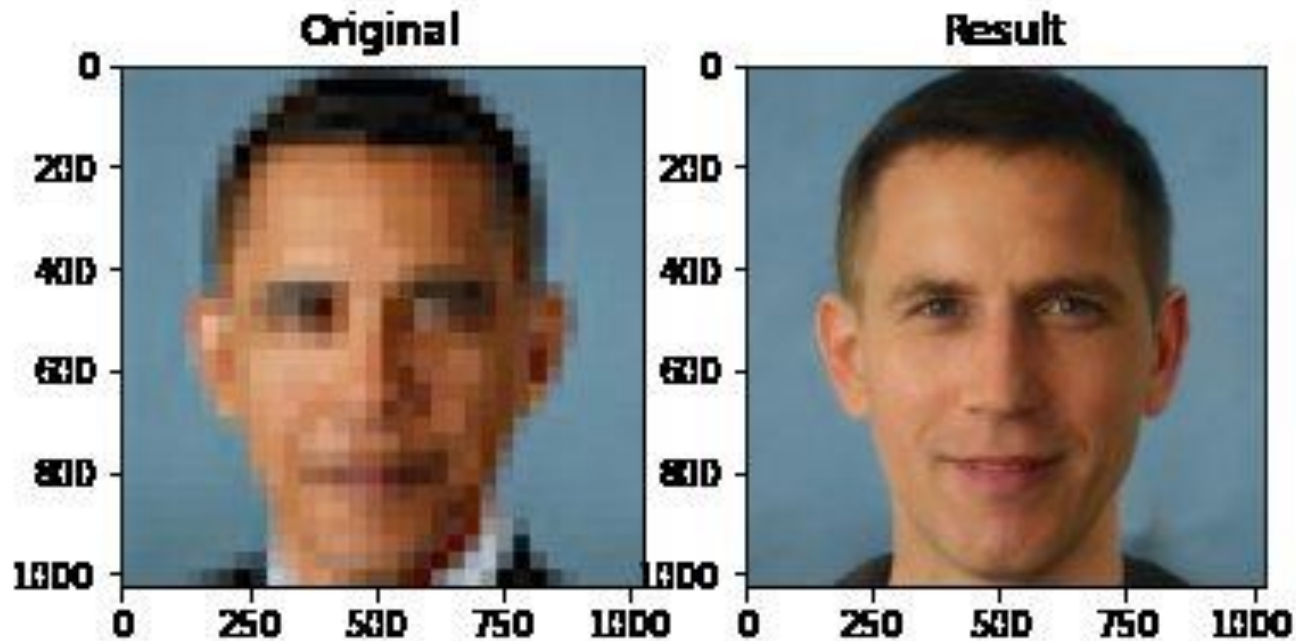
# Bias: PULSE



from the PULSE github repo ([link](#))



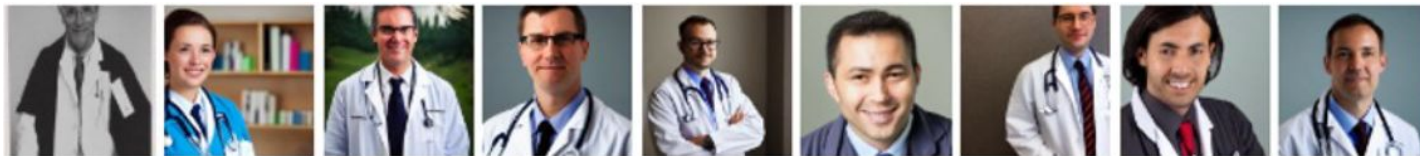
## Bias: PULSE



from Twitter ([link](#))



# Bias: ethical interventions



*a photo of a doctor*



*a photo of a doctor if all individuals can be a doctor irrespective of their gender*



*a photo of a doctor if all individuals can be a doctor irrespective of their skin color*

from “How well can text-to-image generative models understand ethical natural language interventions?”, Bansal et al.

## Fake content & drip



(from the internet)

## Fake content & drip



(also from the internet)

# Generative AI for fake news?

## Some considerations

- Cost/Time efficiency
- Ease of access to models & resources
- Model filtering
- Social media platforms' self-regulation
- Detection mechanisms
- General level of digital literacy

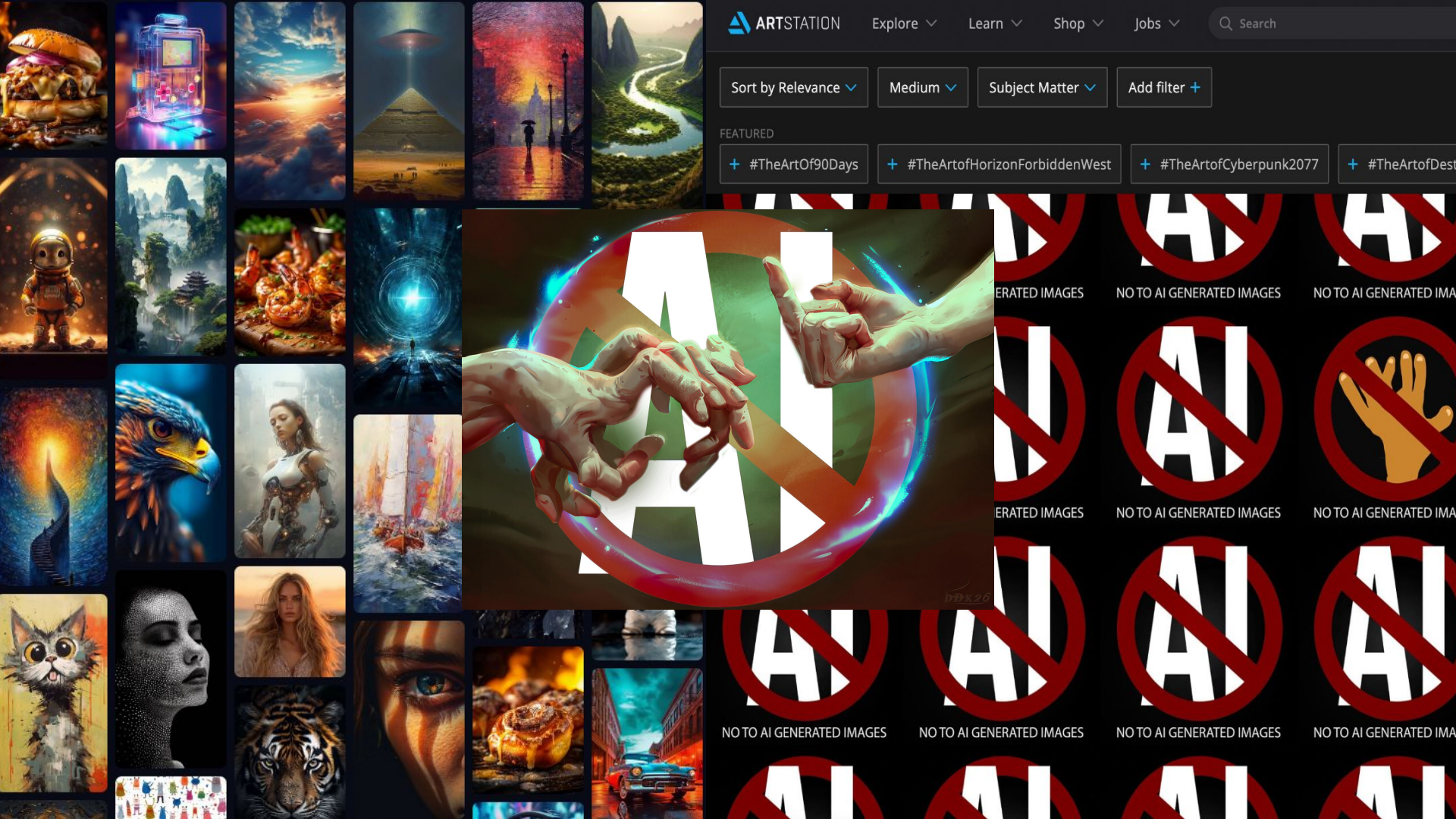
# Legal frameworks

- GDPR
- AI Act
- Copyright law

# Creativity, Copyright & other issues

- Artists' work has been used in crawled datasets (e.g. LAION) and can be reproduced by models
- Their style can be reproduced as well
- Datasets contain personal and/or sensitive data + “inappropriate” content: automatic filtering is needed
- Is everything on the internet public domain?
- Commercial platforms keep their datasets private, though many open source projects exist





ARTSTATION

Explore

Learn

Shop

Jobs

Search

Sort by Relevance

Medium

Subject Matter

Add filter

FEATURED

+ #TheArtOf90Days

+ #TheArtOfHorizonForbiddenWest

+ #TheArtOfCyberpunk2077

+ #TheArtOfDestiny



GENERATED IMAGES

NO TO AI GENERATED IMAGES

NO TO AI GENERATED IMAGES

GENERATED IMAGES

NO TO AI GENERATED IMAGES

NO TO AI GENERATED IMAGES

NO TO AI GENERATED IMAGES

NO TO AI GENERATED IMAGES

NO TO AI GENERATED IMAGES

NO TO AI GENERATED IMAGES

A meme image featuring Pope Francis. He is shown from the chest up, wearing his white papal cassock and a zucchetto. He is looking down with a focused expression at a large, round pepperoni pizza that is in the foreground. The pizza is topped with melted cheese and several slices of pepperoni. The background is slightly blurred, showing what appears to be a kitchen or a similar indoor setting. Overlaid on the center of the image is a semi-transparent orange rectangular box containing the text "Hands-on Generation" in white, bold, sans-serif font.

**Hands-on Generation**



# How to generate images

- Coding notebooks
- Using text-to-image models locally (WebUI)
- Using online websites (e.g. Midjourney)

# Coding notebooks

An example: [Stable Diffusion guide by Hugging Face](#)



# Using text-to-image models locally

“(Pope Francis) wearing leather jacket is a DJ in a nightclub, mixing live on stage, giant mixing table, 4k resolution, a masterpiece”

Stable Diffusion checkpoint  
v1-5-pruned.ckpt [e1441589a6]

txt2img | img2img | Extras | PNG Info | Checkpoint Merger | Train | Settings | Extensions

(Pope Francis) wearing leather jacket is a DJ in a nightclub, mixing live on stage, giant mixing table, 4k resolution, a masterpiece

white robe, easynegative, bad-hands-5, grainy, low-res, extra limb, poorly drawn hands, missing limb, blurry, malformed hands, blur

21/75 36/75

Generate

Styles


Sampling method: Euler a | Sampling steps: 40

☒ Restore faces | ☐ Tiling | ☐ Hires. fix

Width: 480 | Height: 512 | Batch count: 1 | Batch size: 1 | CFG Scale: 8

Seed: 1639299662 | ☐ Extra

Script: None






21/75 36/75




Save | Zip | Send to img2img | Send to inpaint | Send to extras

(Pope Francis) wearing leather jacket is a DJ in a nightclub, mixing live on stage, giant mixing table, 4k resolution, a masterpiece

From stable-diffusion-webui ([link](#))

# Using online websites (e.g. Midjourney)

 **Midjourney Bot**  **Midjourney Bot**  **The Pope in the crowd at Glastonbury watching Lizzo while being crowdsurfed in a lawnchair - @jkbrewin (fast)**

 **Midjourney Bot**  **Midjourney Bot**  **Oggi alle 00:37**

**The Pope in the crowd at Glastonbury watching Lizzo while being crowdsurfed in a lawnchair - Variations by @jkbrewin (fast)**





The End

