Organizational matters

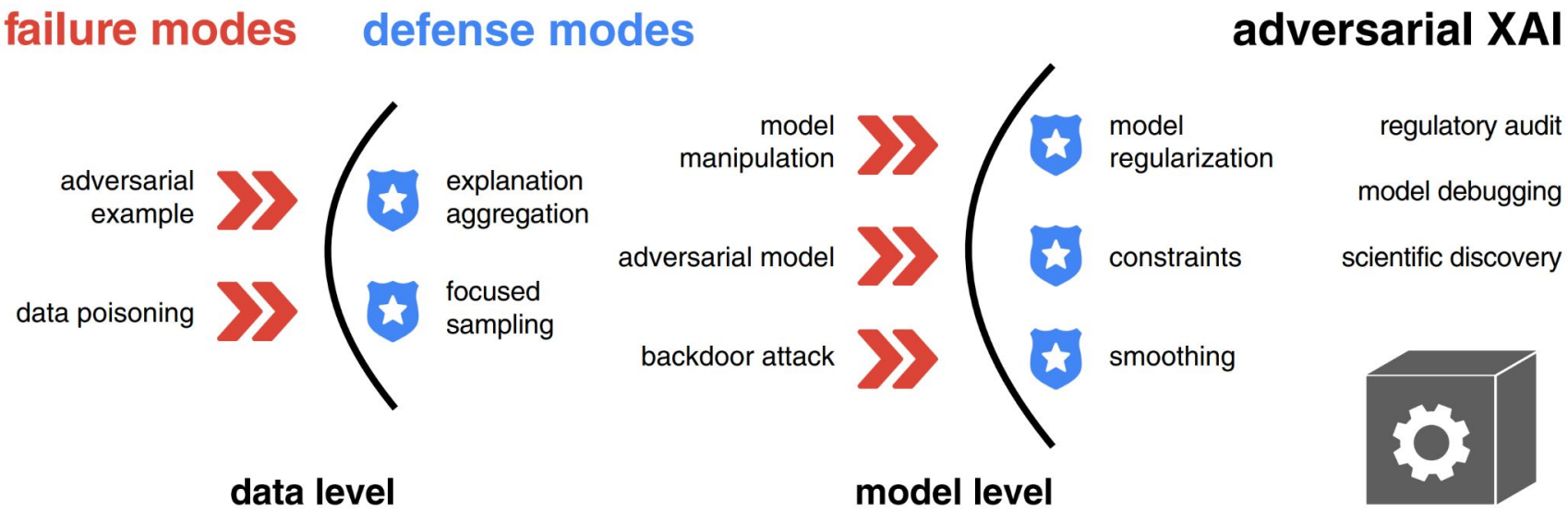
MI2.AI Research Seminar – Winter 2023/24

Schedule

Three research topics

1. **(Adv)XAI → Robustness of explanations**
2. **Red Teaming of foundation models**
3. **Diffusion models for XAI**

(Adv)XAI → Robustness of explanations



https://arxiv.org/abs/2306.06123

On Minimizing the Impact of Dataset Shifts on Actionable Explanations

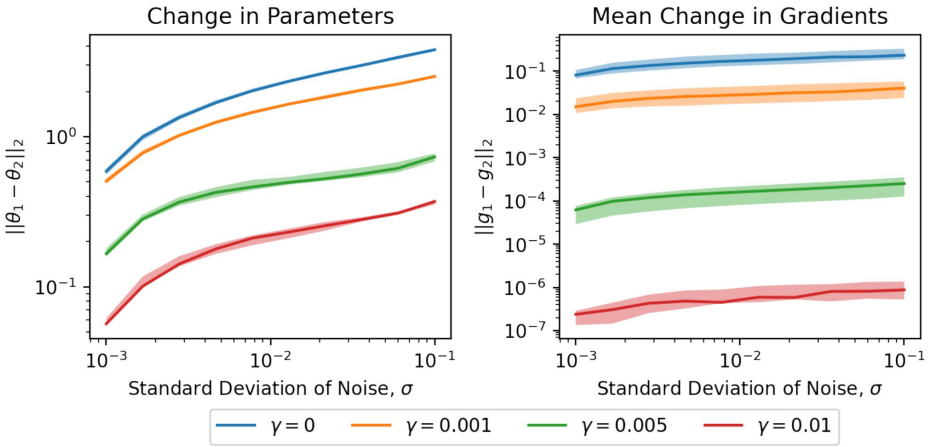
Anna P. Meyer, Dan Ley, Suraj Srinivas, Himabindu Lakkaraju. UAI 2023

Problem: How much do the explanations for a model f1 trained on dataset D1 **change** when **fine-tuning** on a slightly **shifted** dataset D2, resulting in a new model f2?

Lemma 1. *The gradient-parameter Lipschitz has the following property:*

$$\mathbb{E}_{\mathbf{x}\sim\mathcal{D}}\|\nabla_x f(x;\theta_1) - \nabla_x f(x;\theta_2)\|_2 \leq \mathcal{L}_{\Theta,\mathcal{D}} \times \|\theta_2 - \theta_1\|_2$$

where $\Theta = \{\lambda\theta_1 + (1-\lambda)\theta_2 \mid \lambda \in [0,1]\}$.



https://arxiv.org/abs/2306.06716

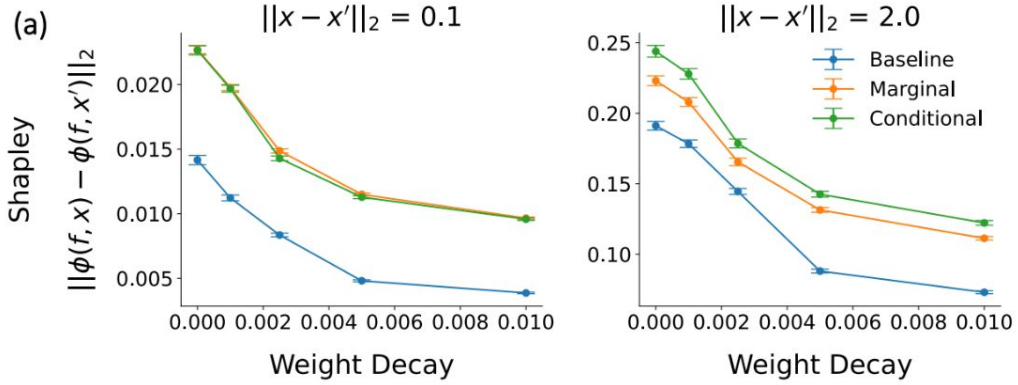
On the Robustness of Removal-Based Feature Attributions

Chris Lin, Ian Covert, Su-In Lee. NeurIPS 2023

2.3 Problem formulation

The problem formulation in this work is straightforward: our goal is to understand the stability of feature attributions under input perturbations and model perturbations. Formally, we aim to study

1. whether $\|\phi(f,x) - \phi(f,x')\|$ is controlled by $\|x - x'\|$ (**input perturbation**), and
2. whether $\|\phi(f,x) - \phi(f',x)\|$ is controlled by $\|f - f'\|$ (**model perturbation**).

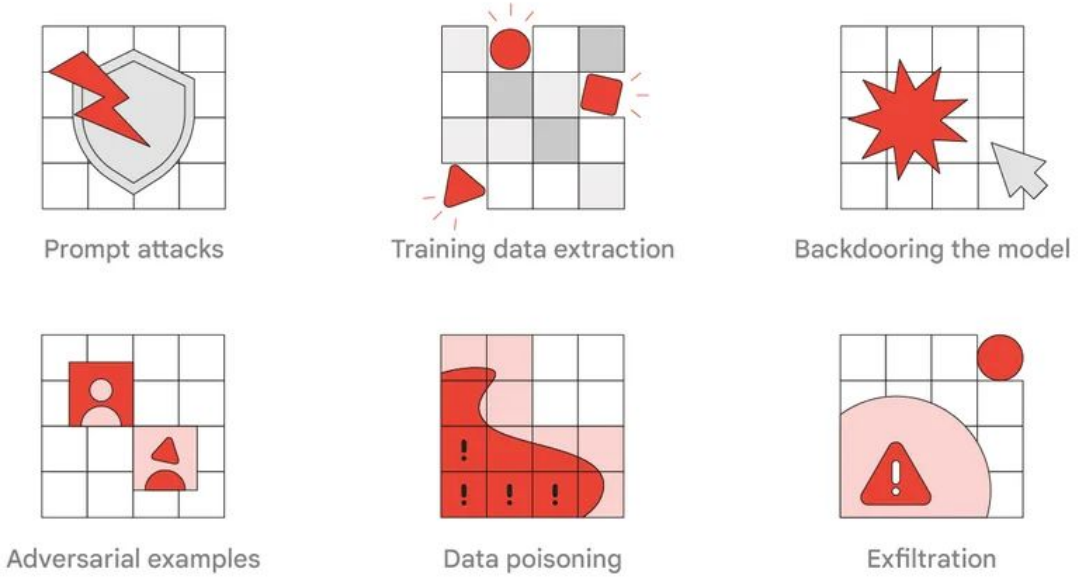


https://arxiv.org/abs/2306.07462

Red Teaming of foundation models

What is Red Teaming

***Red Teaming* is performing attacks on systems to find their vulnerabilities.**



Example attacks on AI models. *source: Google introduces AI Red Team (blog.google)*

What Are Foundation Models?

A foundation model is a large machine learning model trained on a vast quantity of data at scale such that it can be adapted to a wide range of downstream tasks.

E.g.

- **GPT-3**
- **CLIP based models**

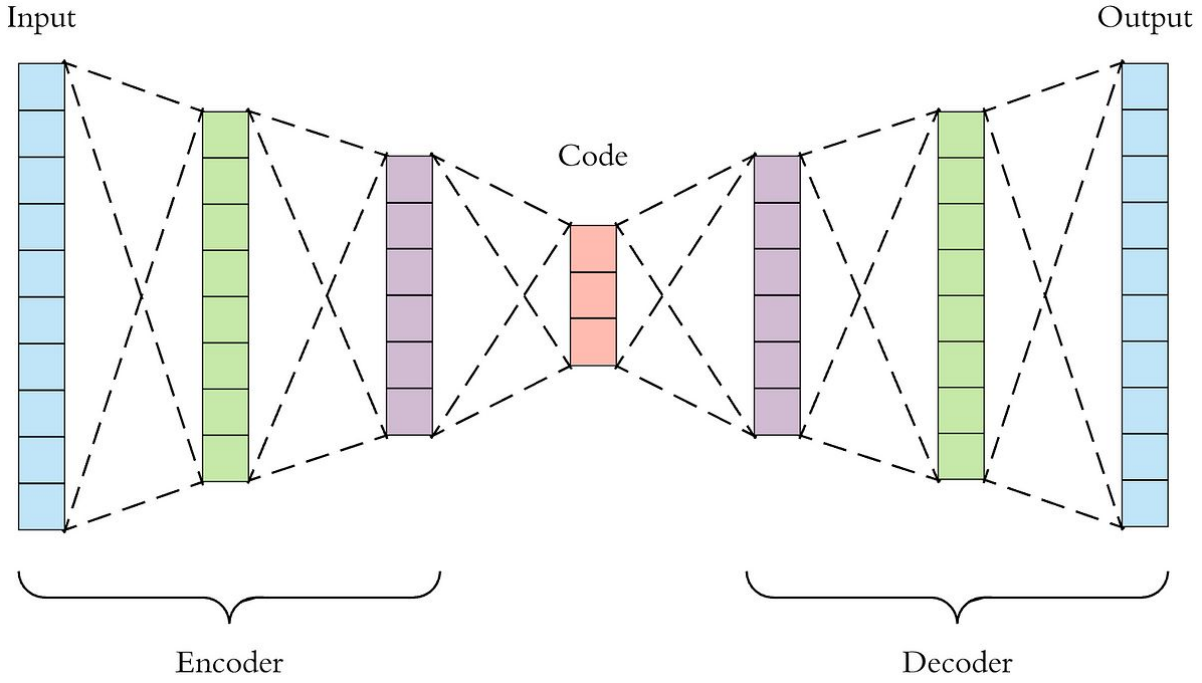
Why Do Red Teaming?

Red Teaming is another step of evaluation to ensure that models perform as intended (e.g. without any data bias). Especially Red Teaming Foundation models is important because of their later usage in many downstream tasks.

Diffusion models for XAI

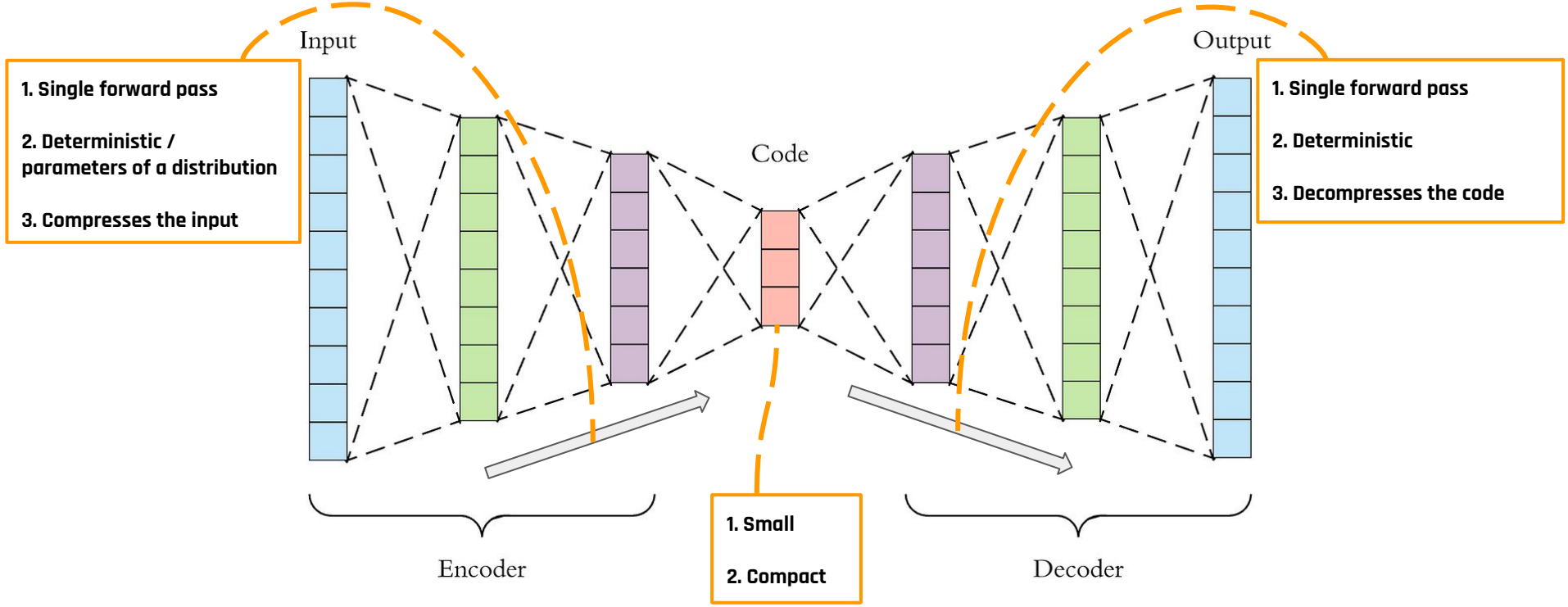
What are diffusion models?

What are diffusion models?



Autoencoder scheme

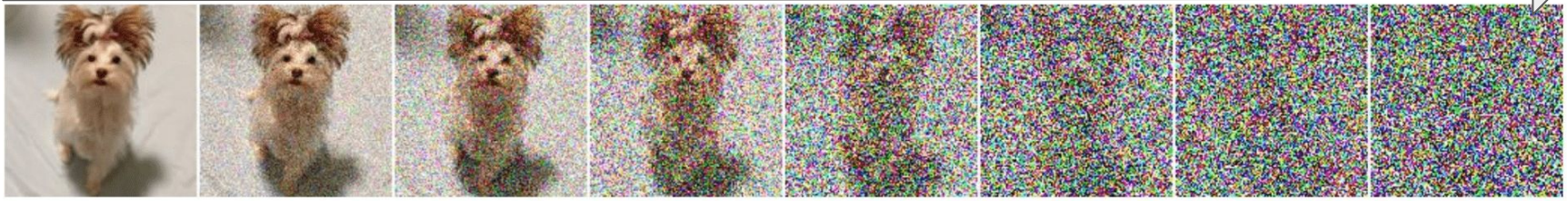
What are diffusion models?

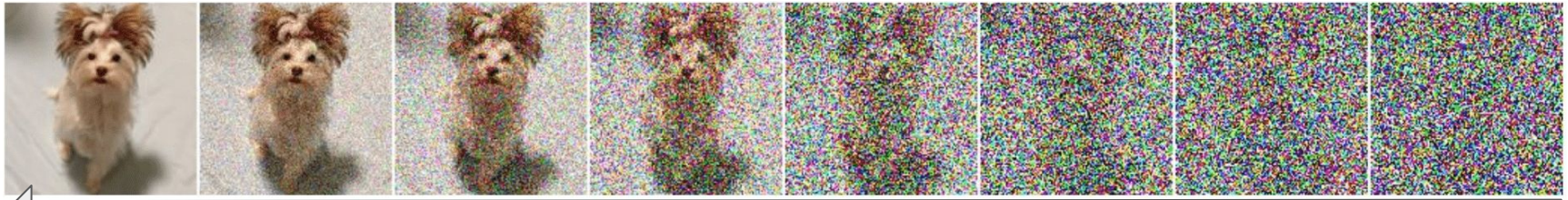


What are diffusion models?

Forward process
Noising process

1. Multiple forward passes
2. Fully random, no neural network
3. No compression

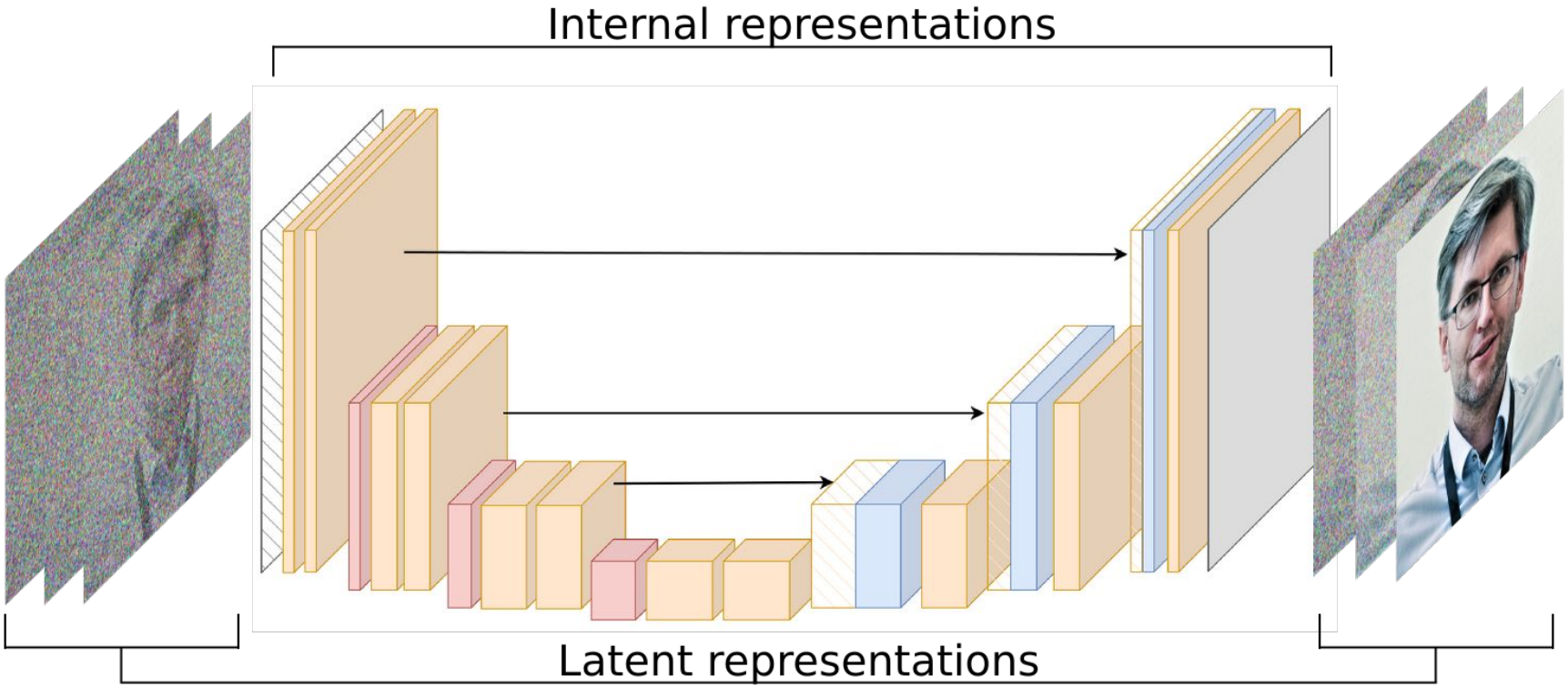




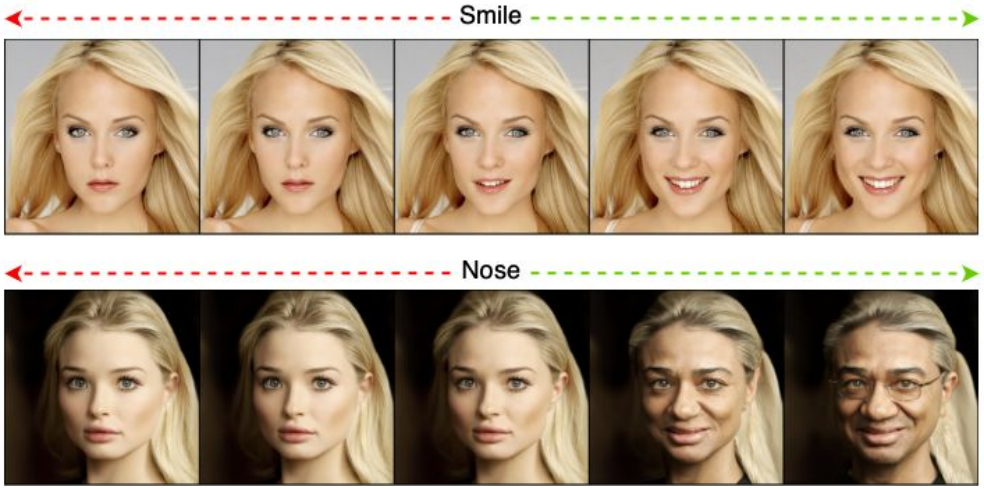
Reverse process
Denoising process

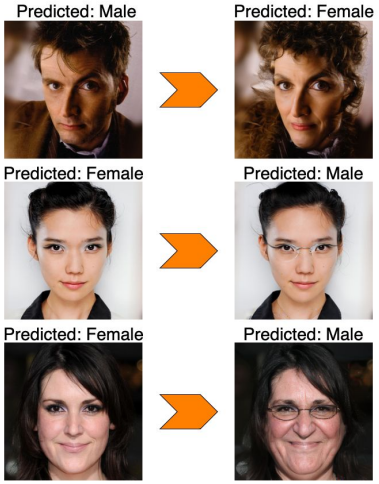
1. Multiple forward passes
2. Deterministic, done iteratively by a neural network
3. No decompression

What are diffusion models?



What are diffusion models?





Thank you