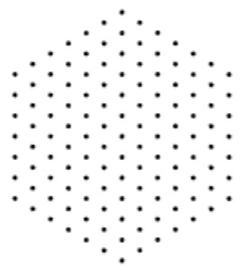


# Real or Fake? Overview on Adversarial Examples

João Nuno Correia

[jncor@dei.uc.pt](mailto:jncor@dei.uc.pt)



COMPUTATIONAL  
DESIGN &  
VISUALIZATION  
LAB.

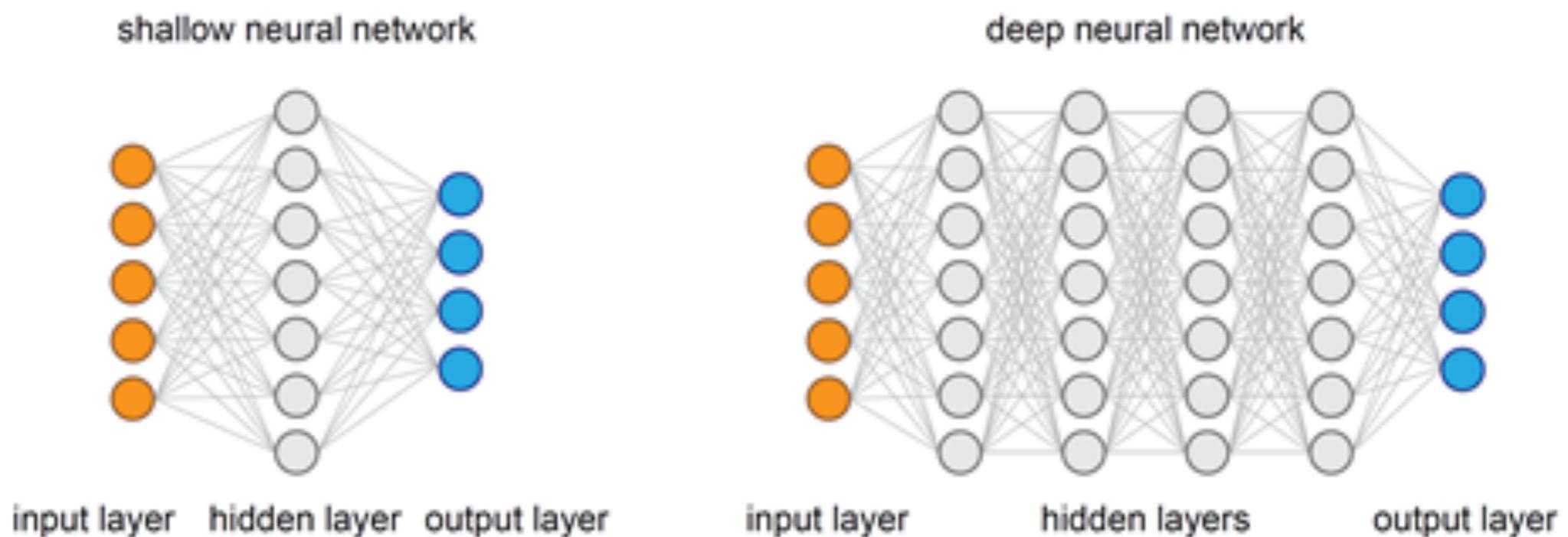
1 2 9 0



UNIVERSIDADE DE  
**COIMBRA**

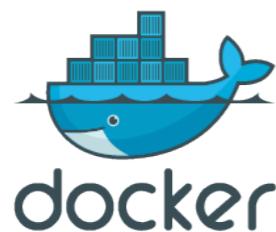
# Deep Learning

- It is Machine Learning ~~on steroids~~ with more layers. I.E. approaches that have more layers of representation to be learned.



# Why the (Deep) hype?

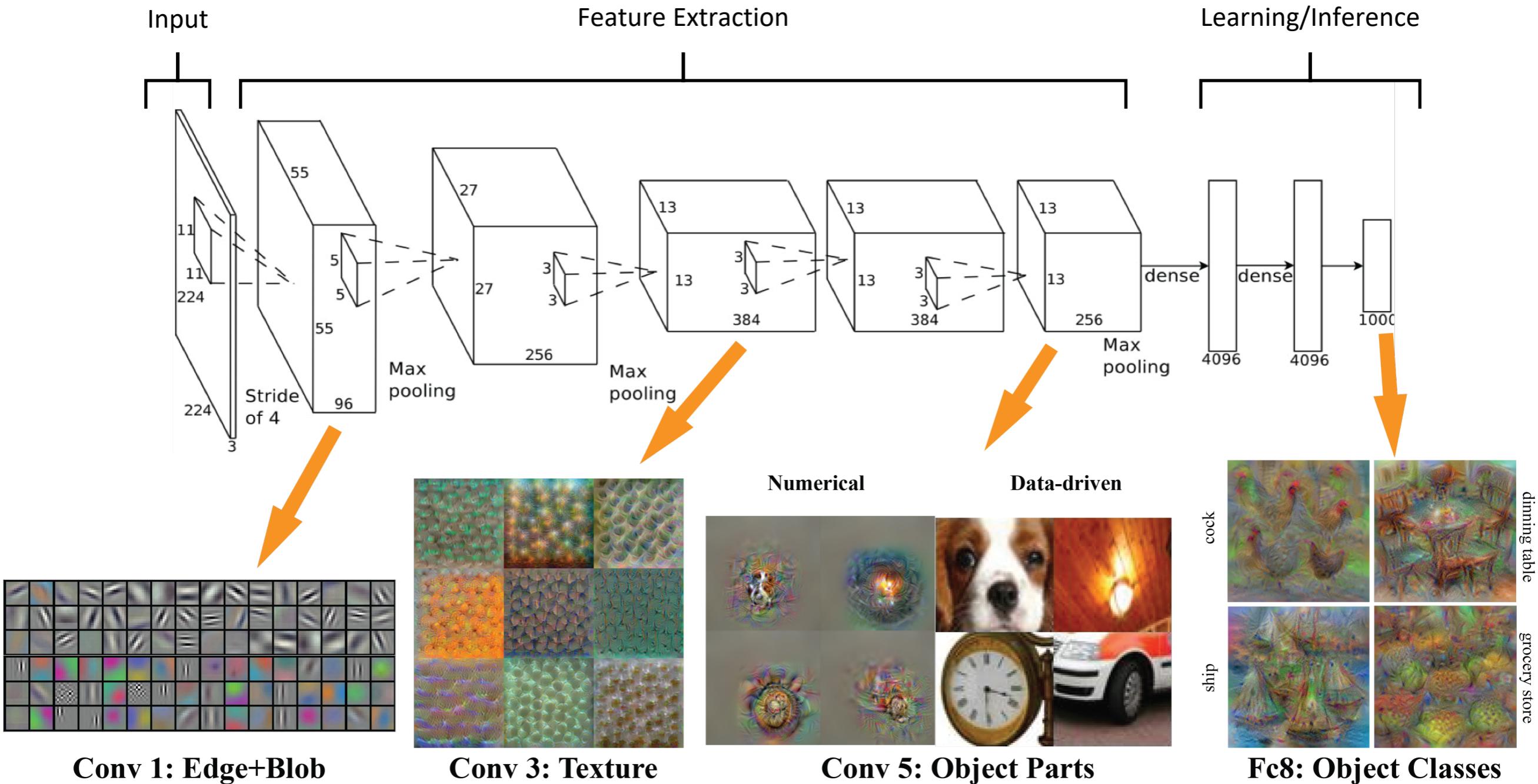
- Dissemination of Information (and Code)
- Democratization of AI and “Open” Era
- Existence of "plugin" solutions and frameworks



- Technology advancements



- And...



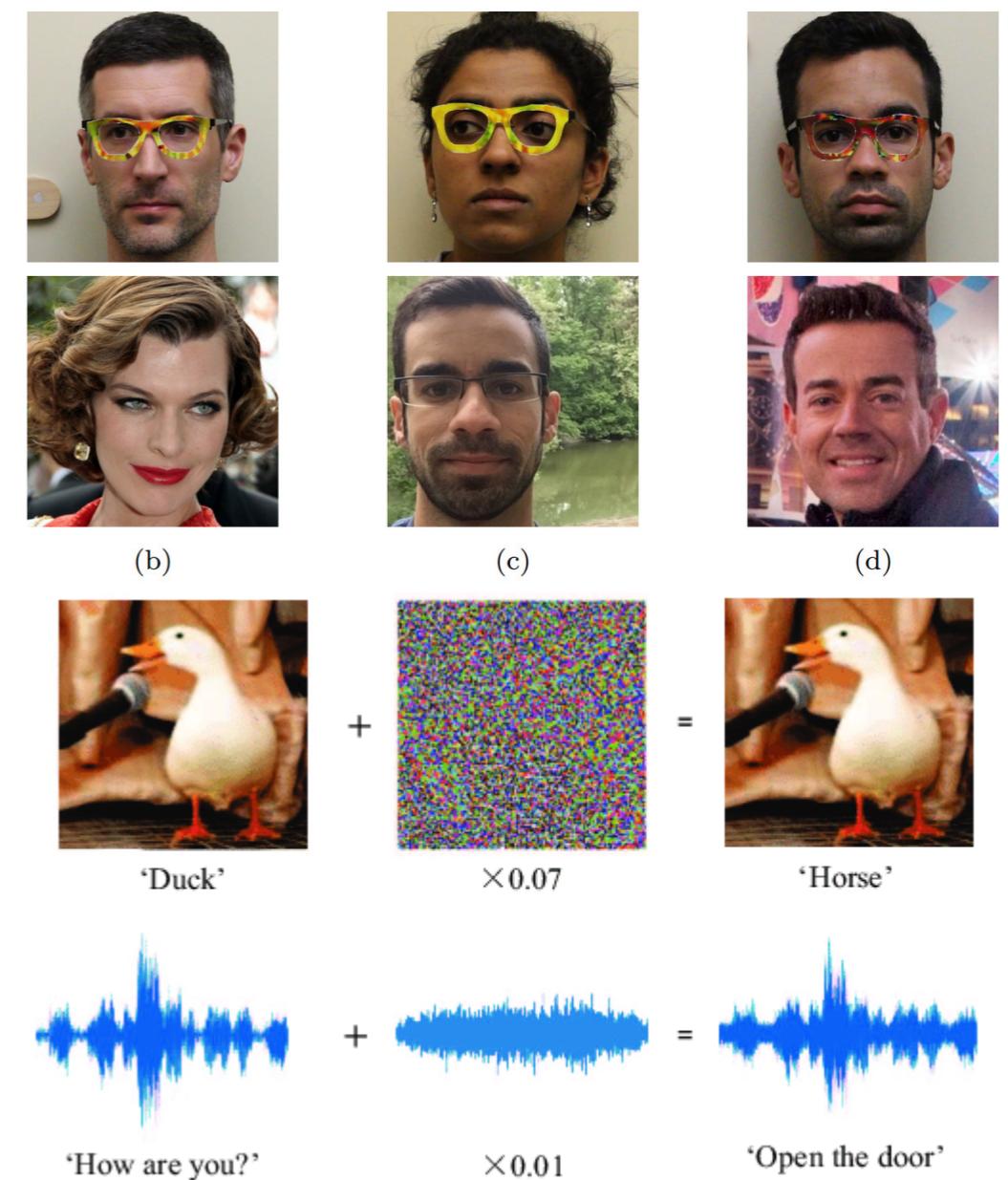
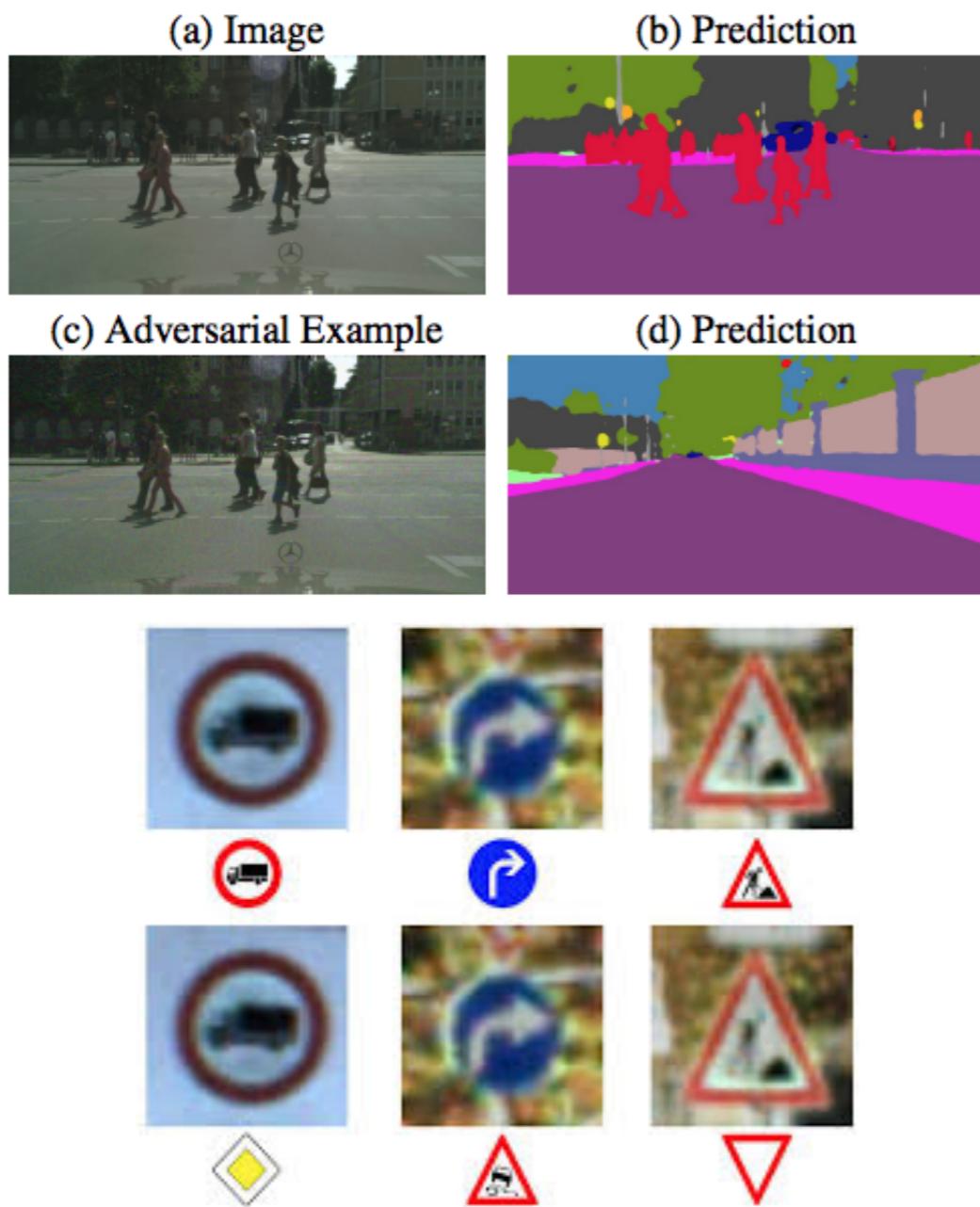
Feature Extraction and Learning! The Alexnet Overview. Krizhevsky et al.)

We think we solved it all...

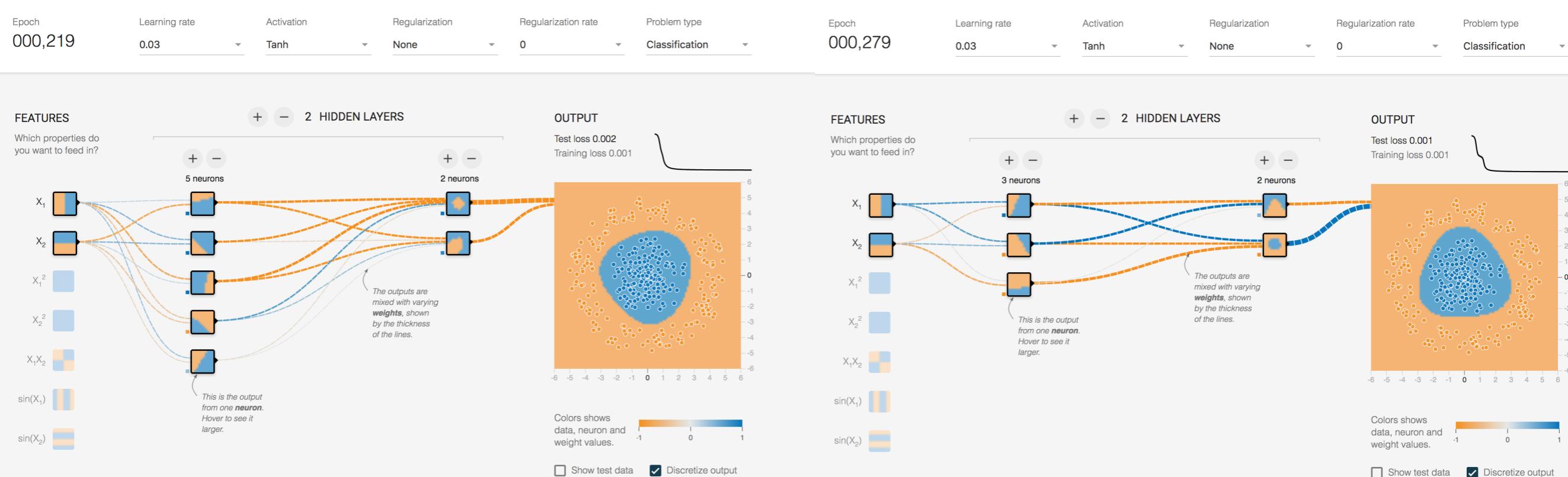
# We think we solved it all...



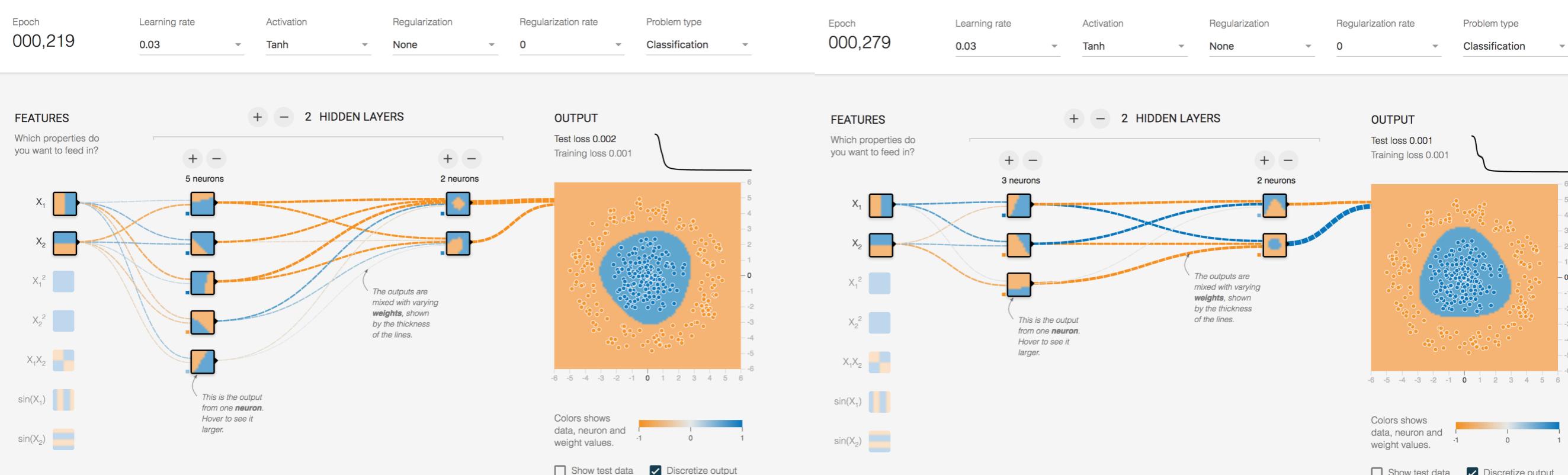
# We think we solved it all... Not Quite

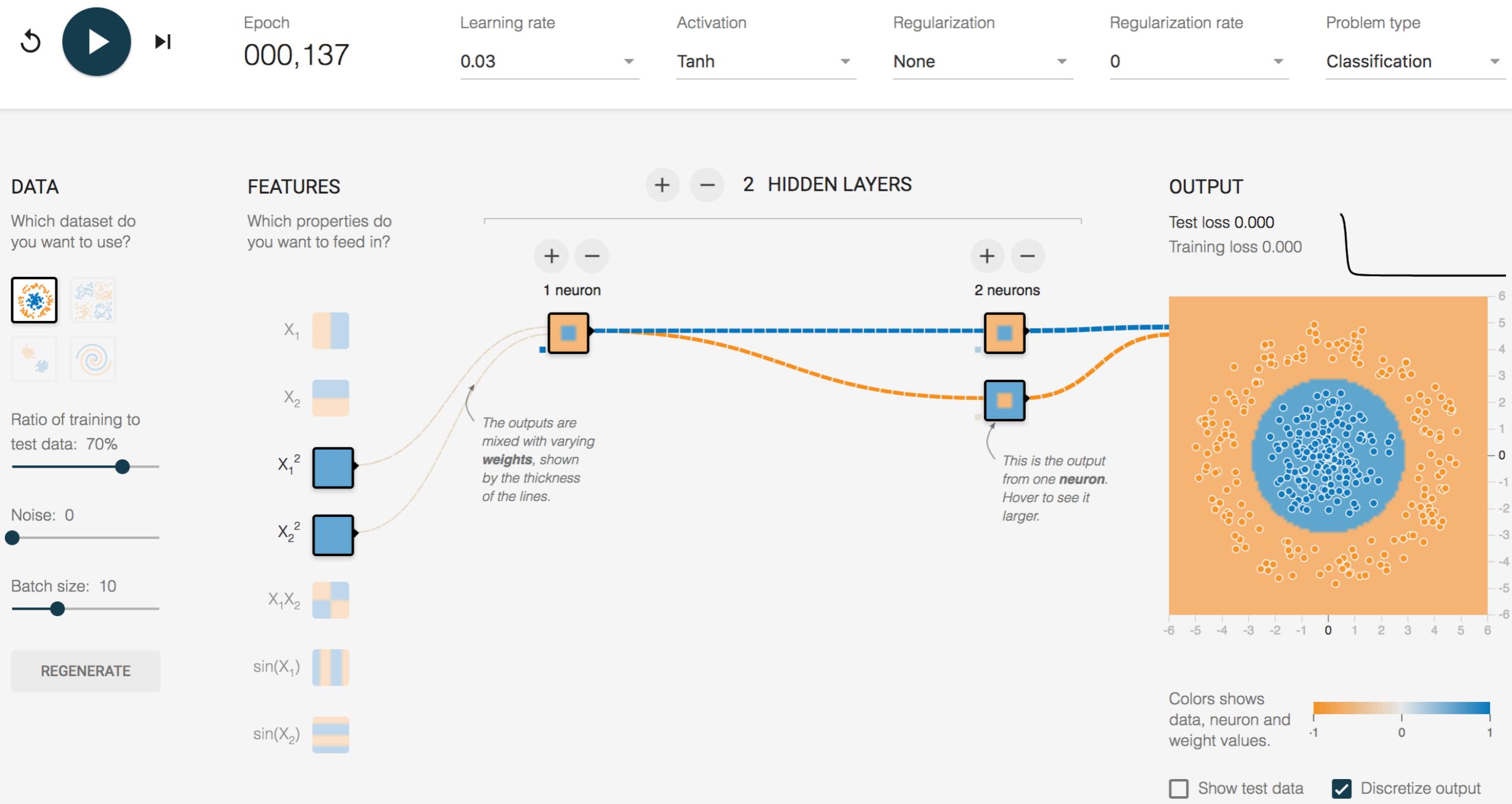


# What are we talking about...

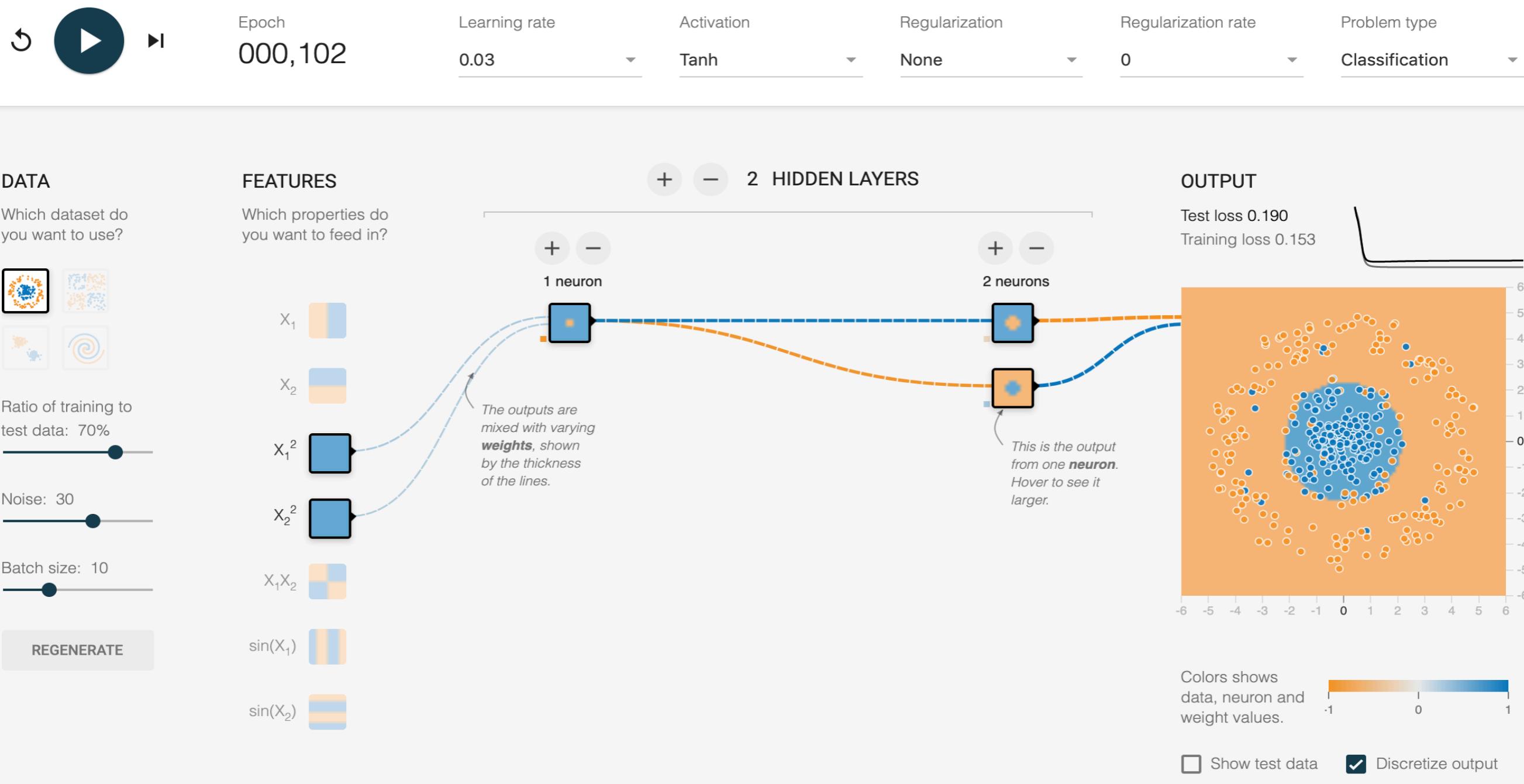


# What are we talking about...





We can come up with clever solutions...



Problem is... real data kicks in.

# What are we talking about...

- So you wonder... for those cases with “Deep”...
  - noisy data?
  - “Bad” model ?

# What are we talking about...

- So you wonder... for those cases with “Deep”...
  - noisy data?
  - “Bad” model ?
- **Adversarial Examples**
  - **They were meant to explore vulnerabilities.**

# Adversarial Learning

# Adversarial Learning

- Machine Learning and computer security
- Adoption of “safe” Machine Learning techniques in adversarial settings
  - E.g. Spam Filtering, Computer Security, Biometric recognition
- Studies the “how to attack” and "how to defend"
  - how to generate adversarial data

# Adversarial Learning

- The complex models, such as DNNs, are composed of linear model parts
  - Adversarial examples explore the non-linearities of the model and/or the data...
- Complex Models are as vulnerable as simple shallow models.
  - However we can “learn” to deal with this...
  - By understanding how it occurs

# Generation of Adversarial Examples



$\mathbf{x}$   
“panda”  
57.7% confidence

$$+ .007 \times$$



$\text{sign}(\nabla_{\mathbf{x}} J(\theta, \mathbf{x}, y))$   
“nematode”  
8.2% confidence

=



$\mathbf{x} +$   
 $\epsilon \text{sign}(\nabla_{\mathbf{x}} J(\theta, \mathbf{x}, y))$   
“gibbon”  
99.3 % confidence

Addition of noise in a training example



computer  
monitor  
screen  
notebook

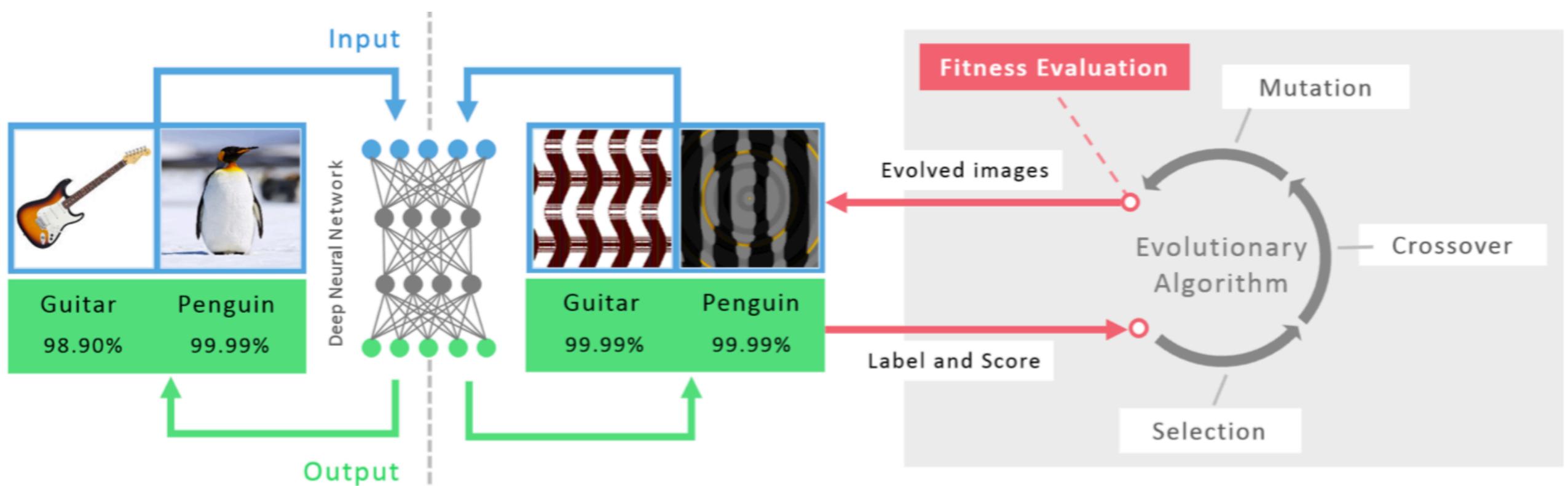


toaster  
banana  
piggy bank  
spaghetti

Several examples of adversarial example generation by tampering with the image

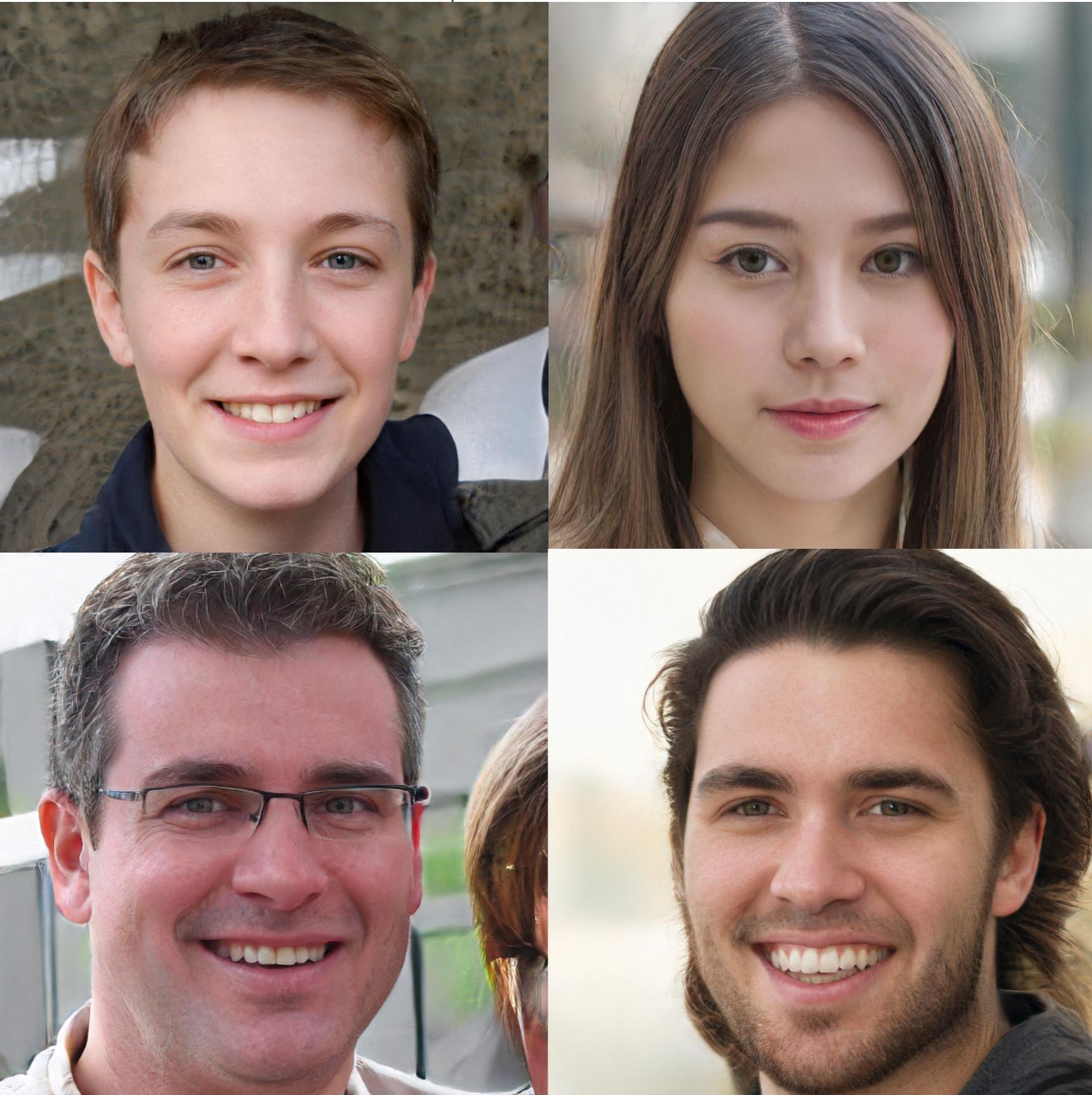


Several examples of adversarial example generation by tampering with the image



DNNs Easily Fooled (Nguyen et al.)

Real or Fake? Overview on Adversarial Examples.



<https://thispersondoesnotexist.com/>

---

South Africa's historic Soweto township marks its 100th birthday on Tuesday in a mood of optimisim.  
**57% World**

South Africa's historic Soweto township marks its 100th birthday on Tuesday in a mooP of optimisim.  
**95% Sci/Tech**

---

it's frustrating to see these guys who are obviously pretty clever waste their talent on parodies of things they probably thought were funniest when they were high. **83% Negative Sentiment**

it's frustrating to see these guys who are obviously pretty **deft** waste their talent on parodies of things they probably thought were funniest when they were high.  
**65% Positive Sentiment**

---

Deep networks for text analysis are also vulnerable

**World News Articles (AI written)** updated 13m ago

[World News](#) [Science](#) [NotTheOnion](#) [Askreddit](#) [Hacker News](#)

This is a demonstration of the latest [state of the art language model \(Rowan et al\)](#), trained by [Allen Institute for AI](#), using [Transformers](#). Every 30 minutes, the application will grab the top headlines from [Reddit](#) or [Hacker News](#) and have the Transformer dream up each article from scratch. Please be aware that the deep neural network was fed only the headline and the domain name to produce the entire article. X

- 1 [South Park' creators issue a mocking 'apology' to China after the show was reportedly banned in the country](#) [www.businessinsider.com](http://www.businessinsider.com)
- 2 ['South Park' Scrubbed From Chinese Internet After Critical Episode](#) [www.hollywoodreporter.com](http://www.hollywoodreporter.com)
- 3 [Disturbing video shows hundreds of blindfolded prisoners in Xinjiang](#) [www.cnn.com](http://www.cnn.com)
- 4 [Trump accused of betraying Kurds and giving ISIS new life after green-lighting Turkey invasion of northern Syria](#) [www.newsweek.com](http://www.newsweek.com)
- 5 [Turkey Attacks Kurds, SDF in Iraq, Syria as U.S. Withdraws](#) [mjpost.com](http://mjpost.com)
- 6 [Trump Justifies Betraying the Kurds: They "Fought With Us, But Were Paid Massive Amounts of Money"](#) [www.motherjones.com](http://www.motherjones.com)
- 7 [Trump's Team Texted About Doing the Exact Ukraine Thing Trump Says Didn't Happen: Texts released by House investigators reveal Ukraine understood their relationship with America depended on investigating Trump's political rivals](#) [www.rollingstone.com](http://www.rollingstone.com)
- 8 [Donald Trump got "rolled" by Turkish President Recep Tayyip Erdogan, according to a National Security Council source with direct knowledge of the discussions](#) [www.newsweek.com](http://www.newsweek.com)
- 9 [World's top banks have poured \\$1.9 trillion into fossil fuel financing since the Paris Agreement was adopted, with financing on the rise each year](#) [www.banktrack.org](http://www.banktrack.org)
- 10 [House Democrats subpoena Pentagon, White House budget office for Ukraine documents](#) [www.cnbc.com](http://www.cnbc.com)
- 11 [Parade of US diplomats to testify in Trump impeachment drive](#) [www.smh.com.au](http://www.smh.com.au)
- 12 [Trump boasts of 'great and unmatched wisdom' and threatens to 'obliterate' the Turkish economy](#) [theweek.com](http://theweek.com)

And some of these articles do not exist...

Adversarial Attacks

adversarial-attacks.net

## Results

In general it is possible to hide any transcription in any audio file with a success rate of nearly 100 %. As an example, we have some audio clips, which are modified with the described algorithm:

### Audio Examples

	Original	Modified	Noise
Speech	<button>▶ ORIGINAL</button>	<button>▶ MODIFIED</button>	<button>▶ NOISE</button>
Music	<button>▶ ORIGINAL</button>	<button>▶ MODIFIED</button>	<button>▶ NOISE</button>
Birds	<button>▶ ORIGINAL</button>	<button>▶ MODIFIED</button>	<button>▶ NOISE</button>
Speech	<button>▶ ORIGINAL</button>	<button>▶ MODIFIED</button>	<button>▶ NOISE</button>

ALAN A NINE MONTH UNCERTAIN

PLAY STOP

Adversarial sound examples.  
<https://adversarial-attacks.net/>

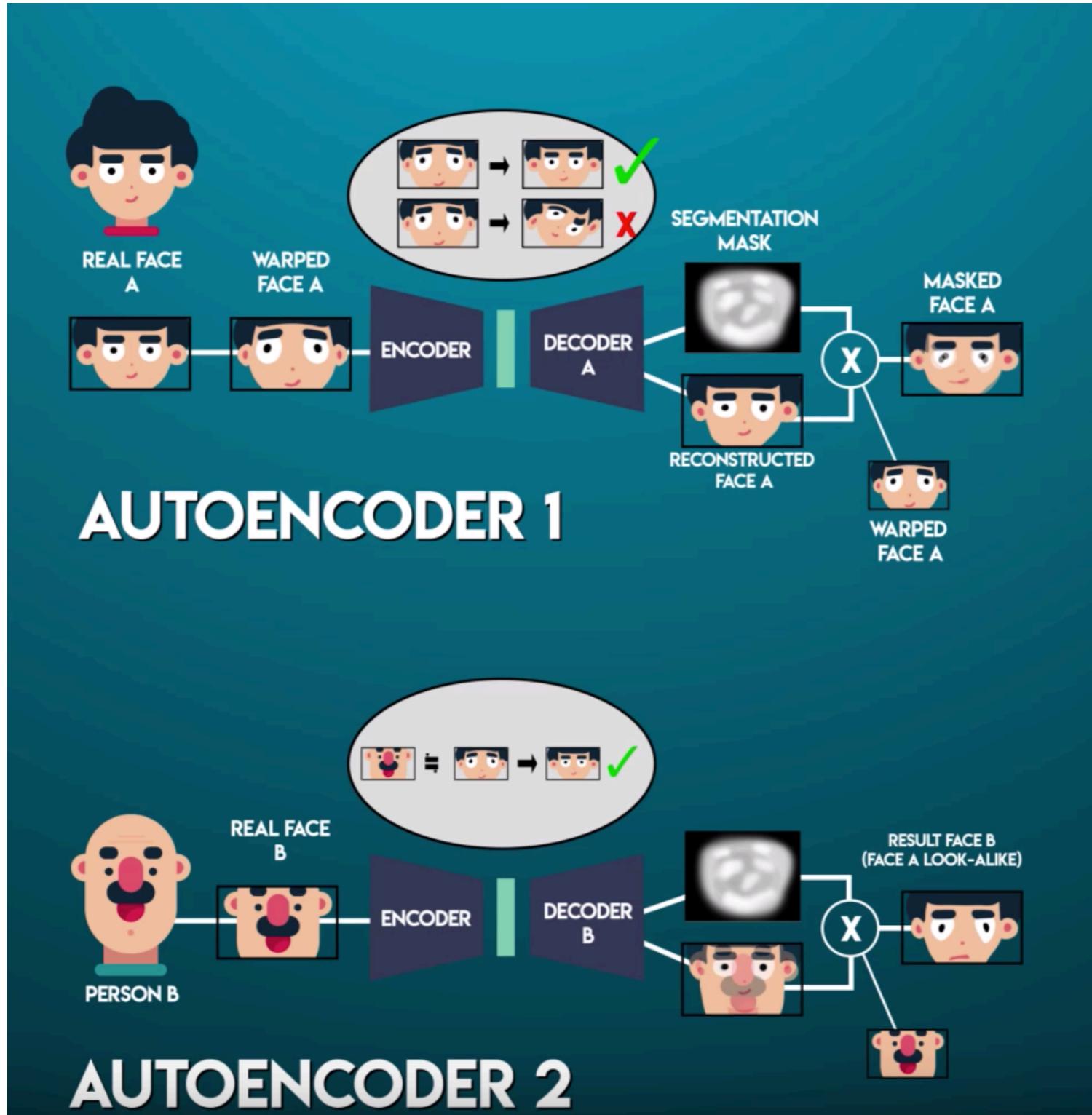
The screenshot shows a web browser window for 'adversarial-attacks.net'. The title bar says 'Adversarial Attacks'. The main content area is titled 'Results' and contains the following text: 'In general it is possible to hide any transcription in any audio file with a success rate of nearly 100 %. As an example, we have some audio clips, which are modified with the described algorithm:'

### Audio Examples

	Original	Modified	Noise
Speech	<button>▶ ORIGINAL</button>	<button>▶ MODIFIED</button>	<button>▶ NOISE</button>
Music	<button>▶ ORIGINAL</button>	<button>▶ MODIFIED</button>	<button>▶ NOISE</button>
Birds	<button>▶ ORIGINAL</button>	<button>▶ MODIFIED</button>	<button>▶ NOISE</button>
Speech	<button>▶ ORIGINAL</button>	<button>▶ MODIFIED</button>	<button>▶ NOISE</button>

Below the table, there is a player interface for an audio file named 'ALAN A NINE MONTH UNCERTAIN'. It features a waveform visualization, a progress bar, and control buttons labeled 'PLAY' and 'STOP'.

Adversarial sound examples.  
<https://adversarial-attacks.net/>



Deep fakes, combination of two autoencoders with different objectives.

Video generation.

<https://www.youtube.com/watch?v=cQ54GDm1eL0>

# Adversarial Learning: Counter Measures

Two main types of defense:

- Protect with a model or models
  - Use other models to pre-classify the input
  - Use generative models to prevent malicious inputs
- Protect with data
  - Filters and pre-processing data (training and input)
  - **Study and usage of generated adversarial examples.**

# EFFECTIVE - Evolutionary Framework for Classifier Assessment and Improvement (Correia et al.)

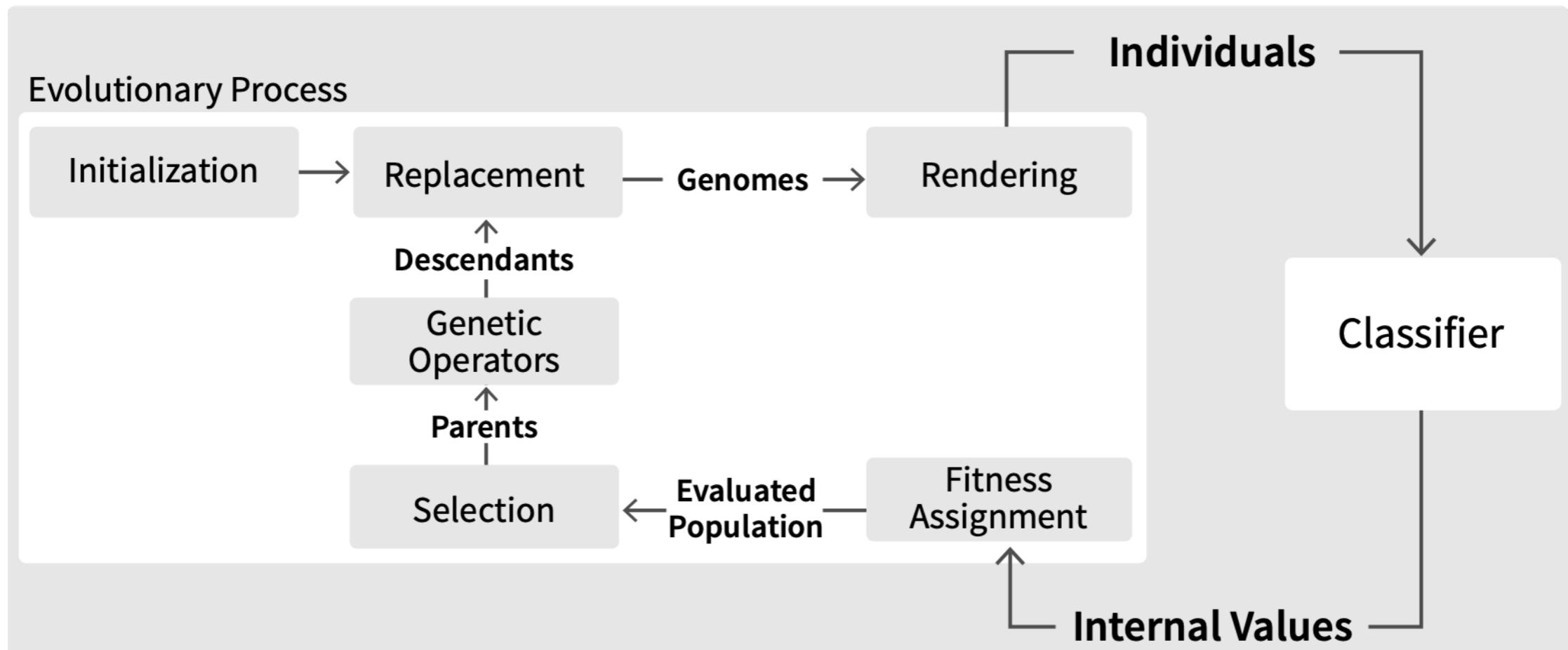


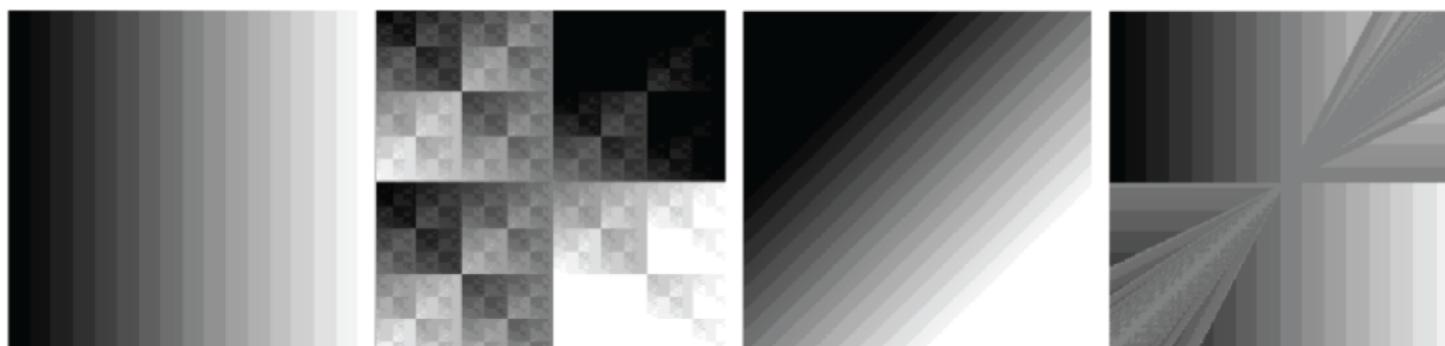
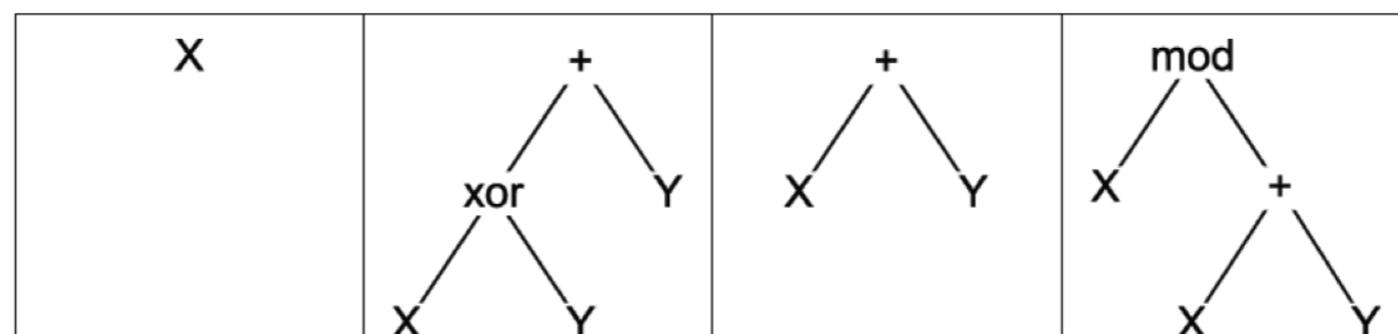
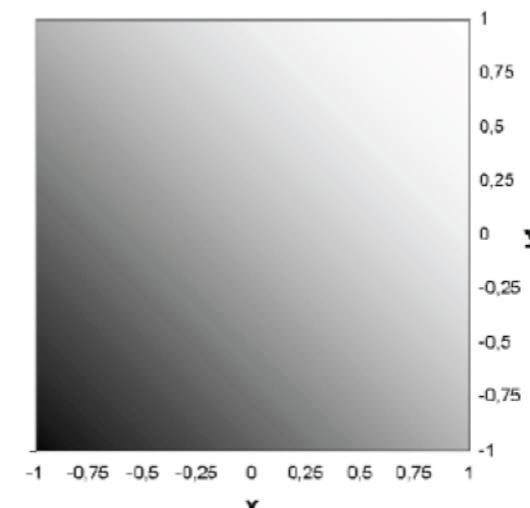
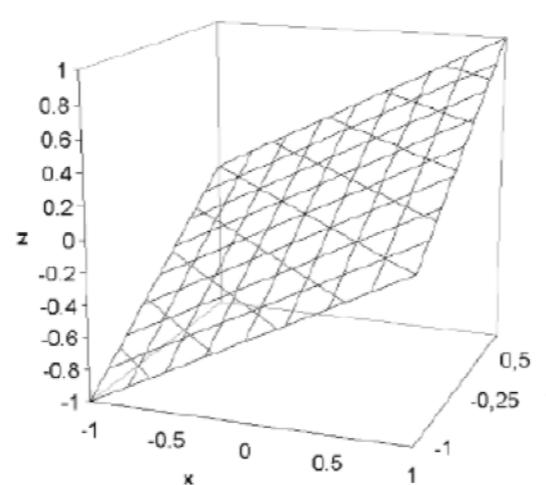
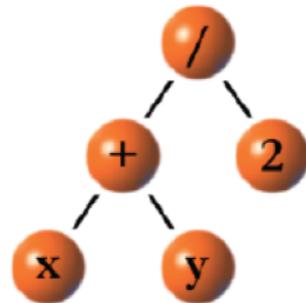
## Evolving faces

Evolve images of faces that the classifier detect as such

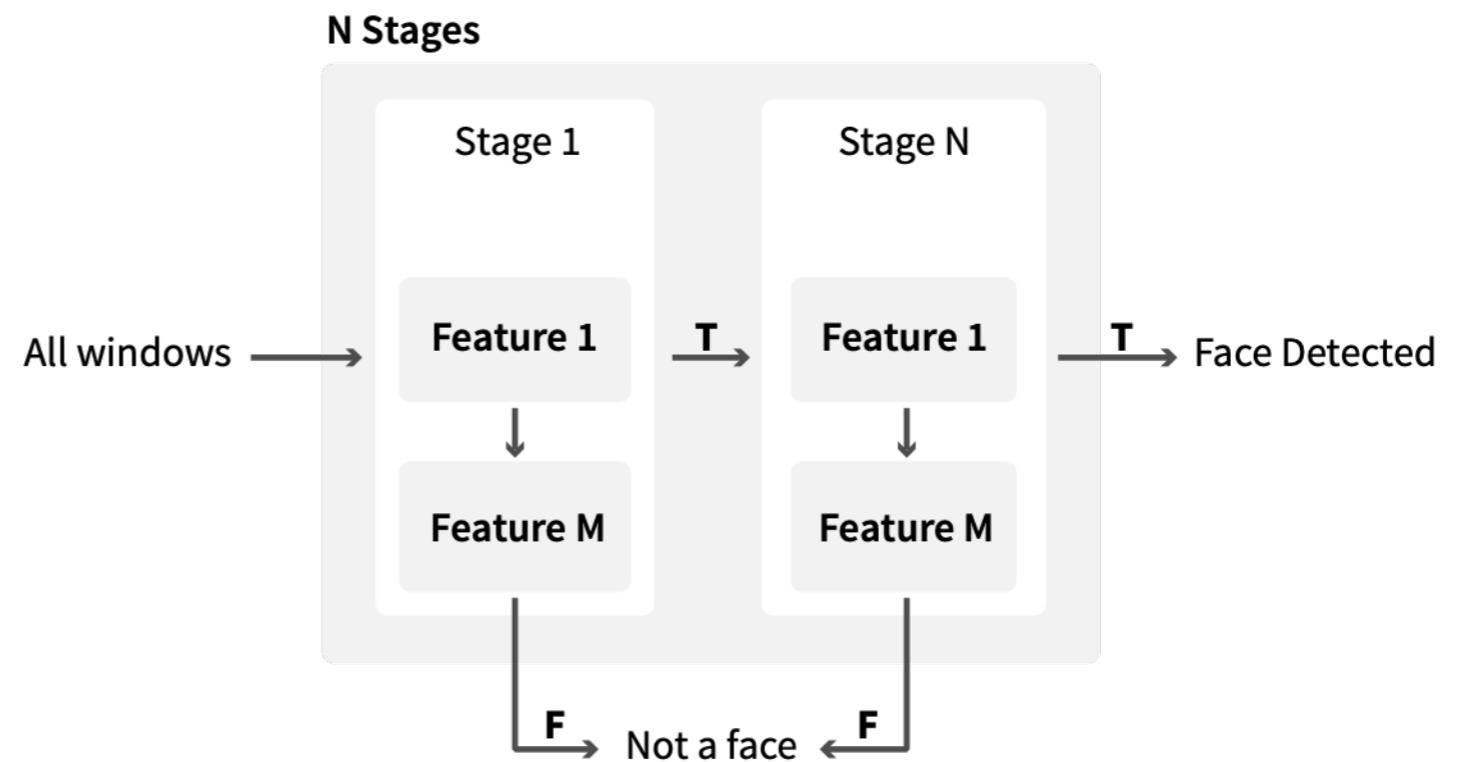
- A general purpose expression-based evolutionary art tool
- An off-the-shelf face detector as classifier
- The classifier must detect faces in the evolved images

## EC Run



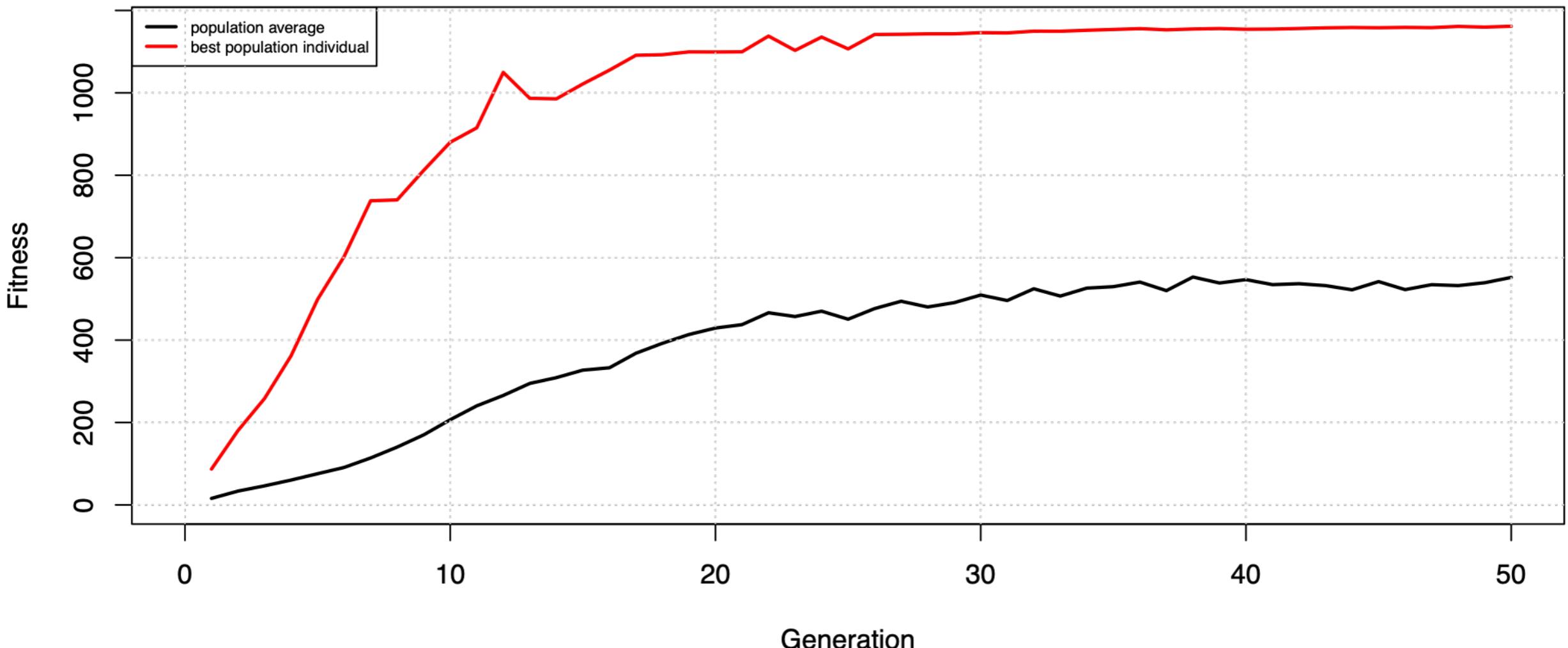


**Genetic Programming engine**

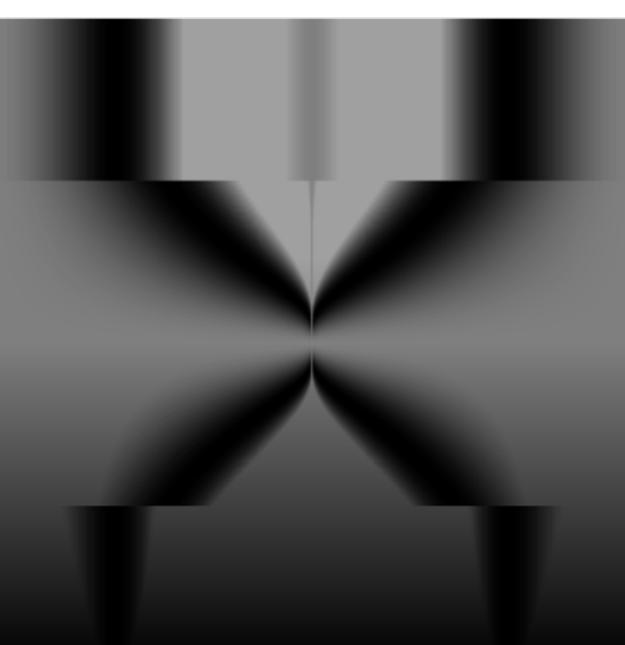
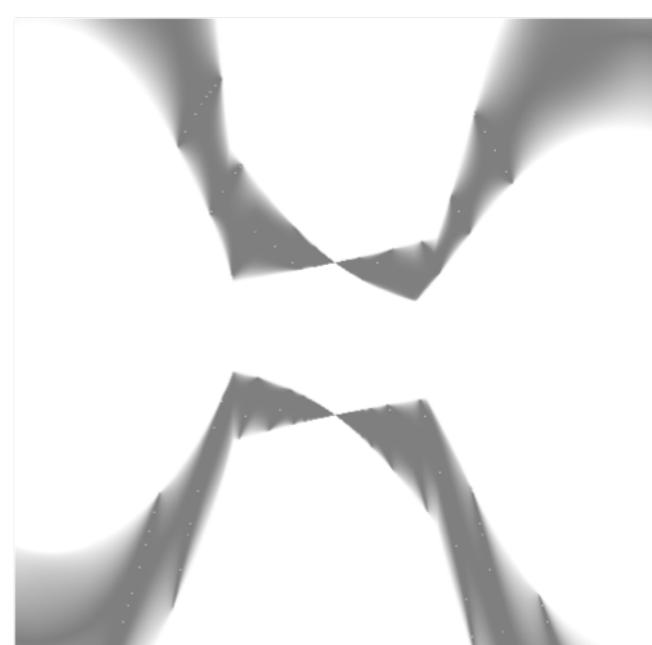
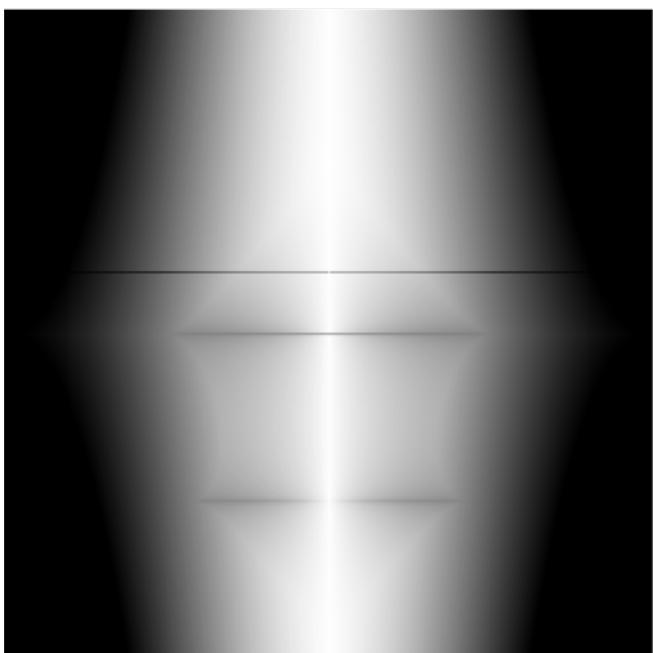
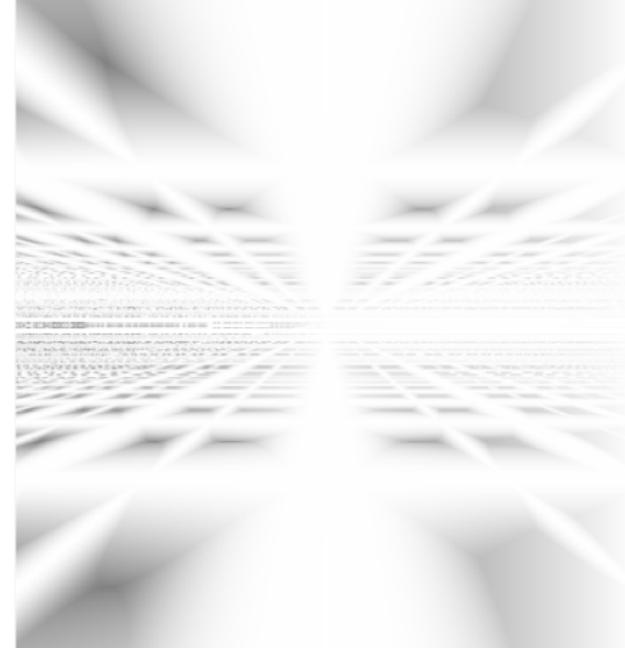
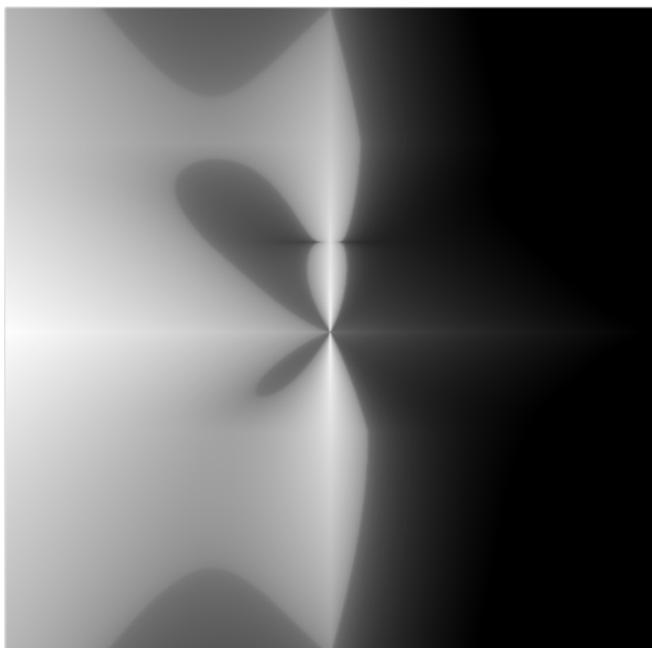
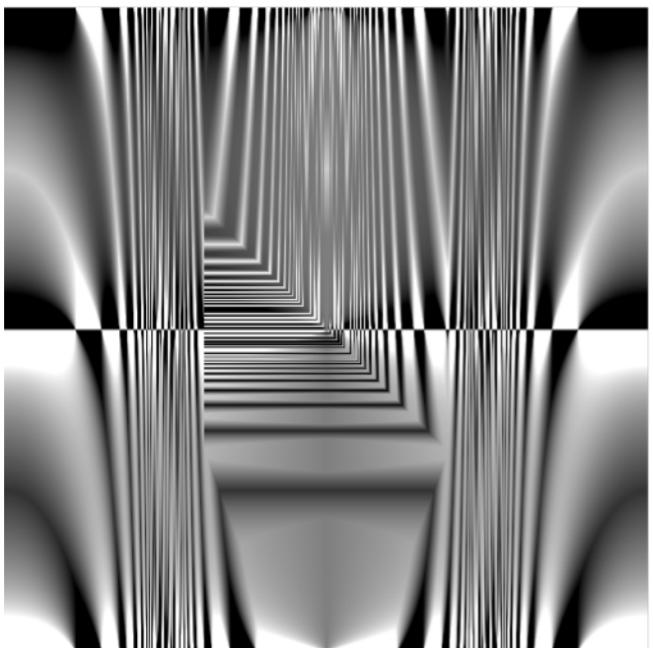


$$f(x) = \sum_{i=1}^{\text{countstages}_x} \text{stagedif}_x(i) * i + \text{countstages}_x * 10$$

Haar features [top-left]; Cascade classifier [top-right]; Cascade classifier in action (video by Adam Harvey) [bottom-left]; fitness function [bottom-right]



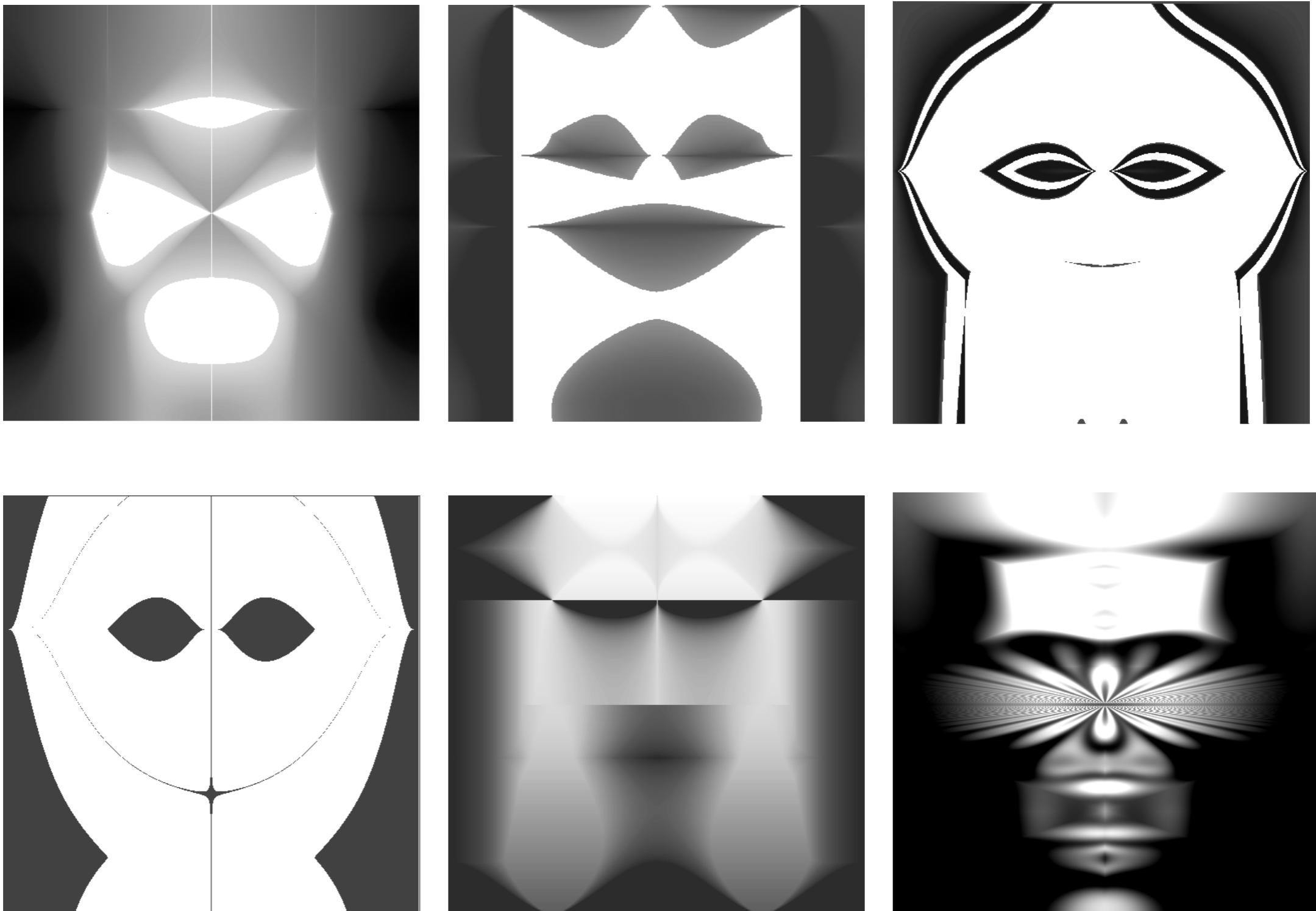
Evolution of fitness



Images classified as faces by the classifier



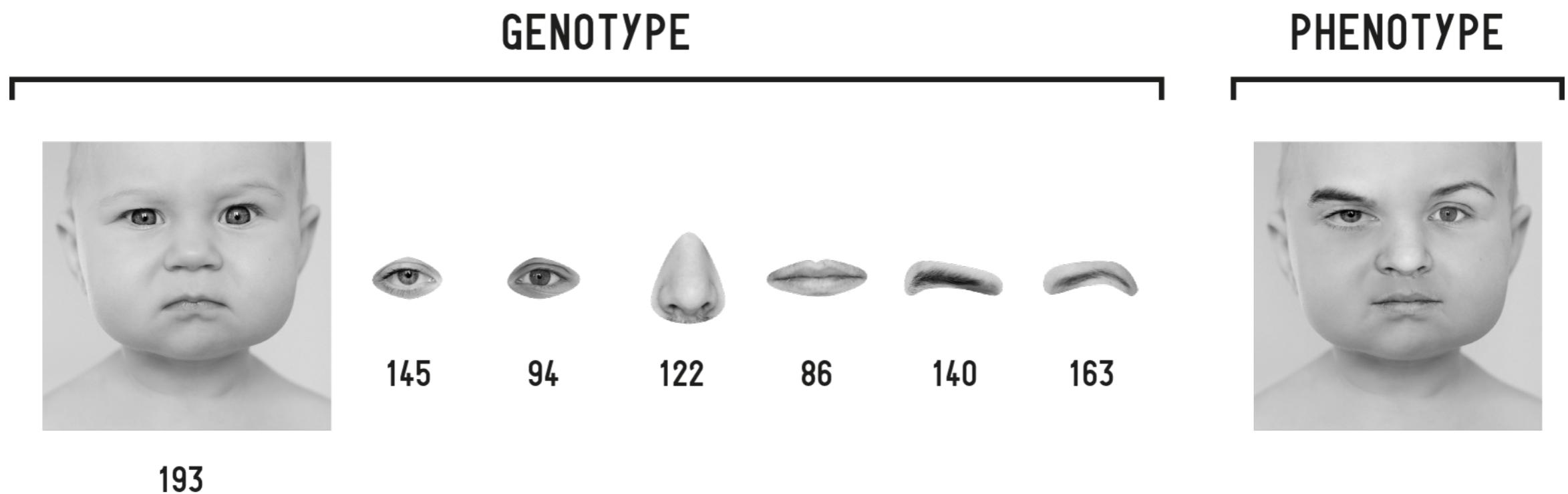
Images classified as faces by the classifier



Face (Evocative) Examples

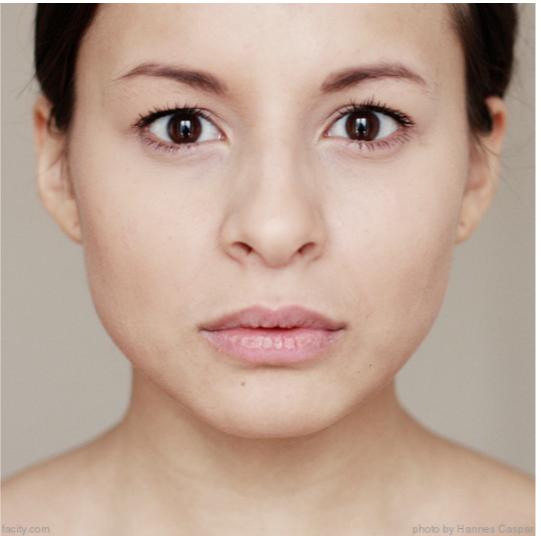
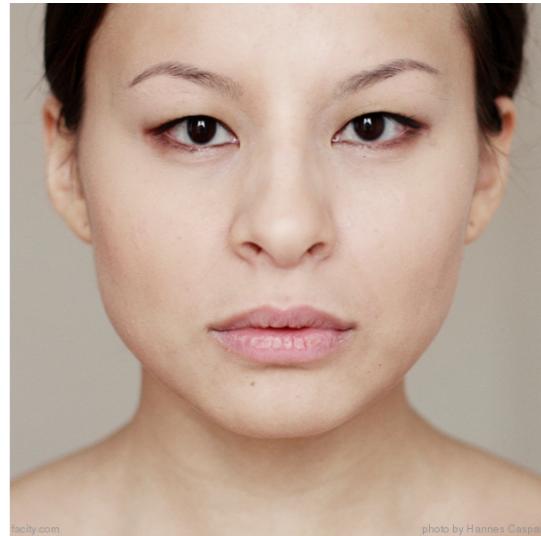
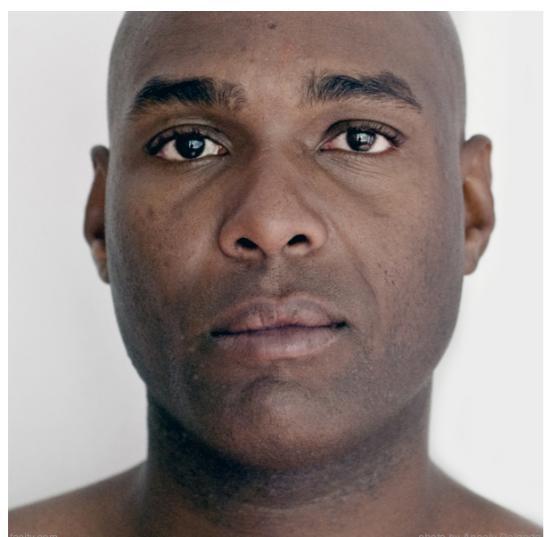
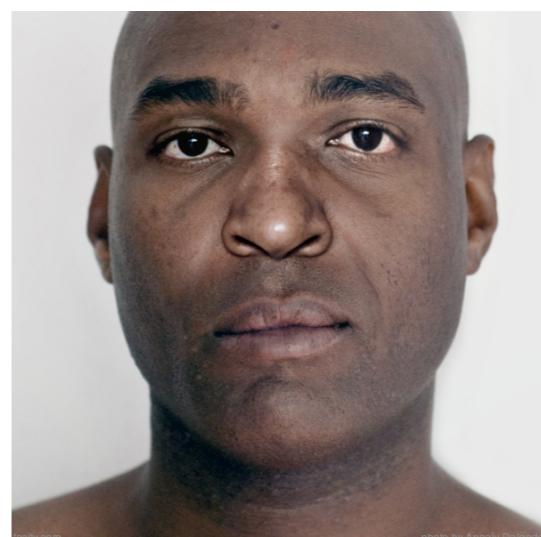
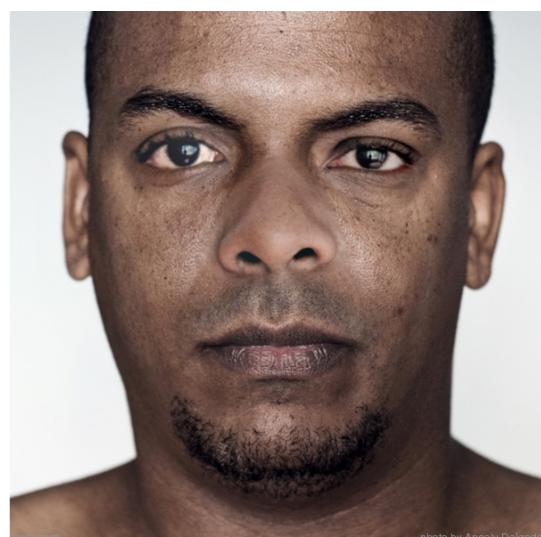
## Evolving undetected faces

Evolve images of faces that the classifier does not detect as such



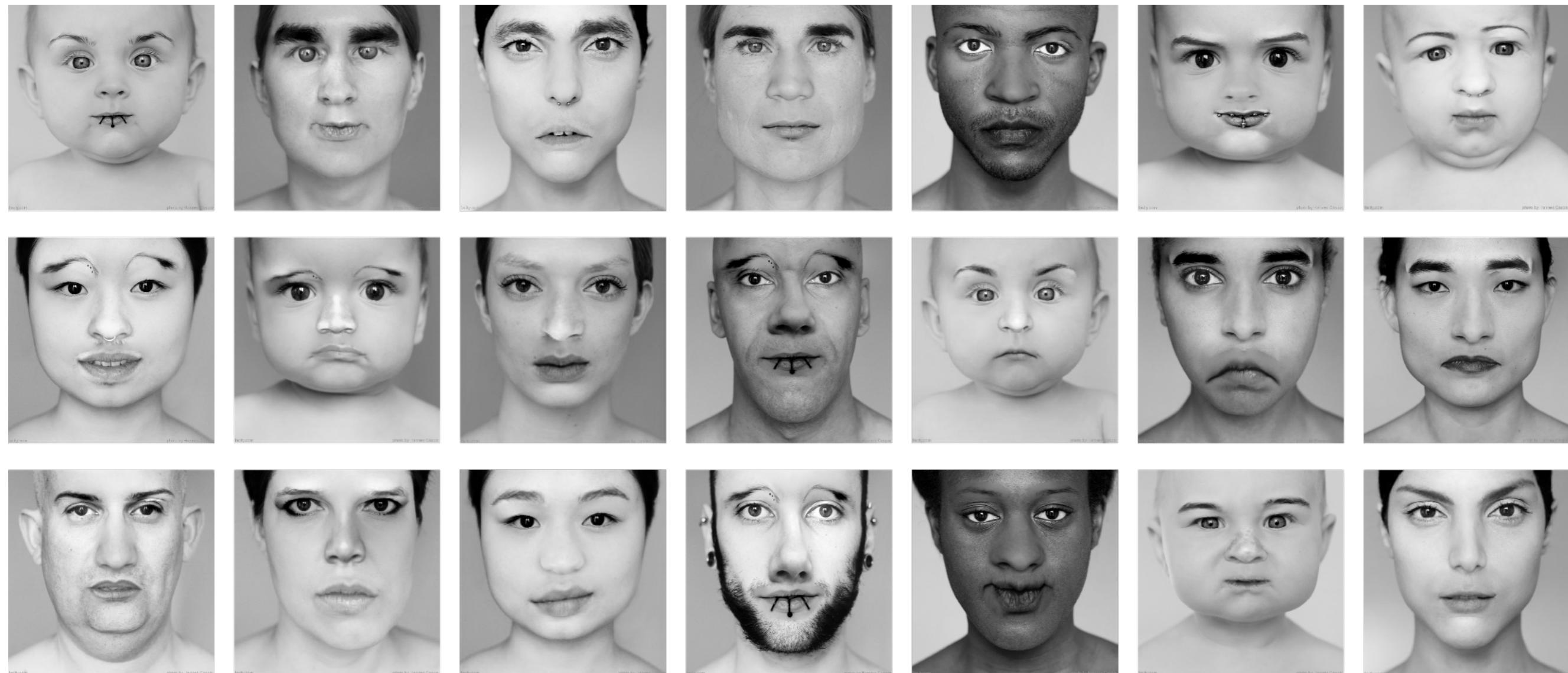
**Genotype and phenotype of the evolving faces**

## Real or Fake? Overview on Adversarial Examples.

facile.comfacile.comfacile.comfacile.comfacile.comfacile.comfacile.comfacile.comfacile.comfacile.comfacile.comfacile.com



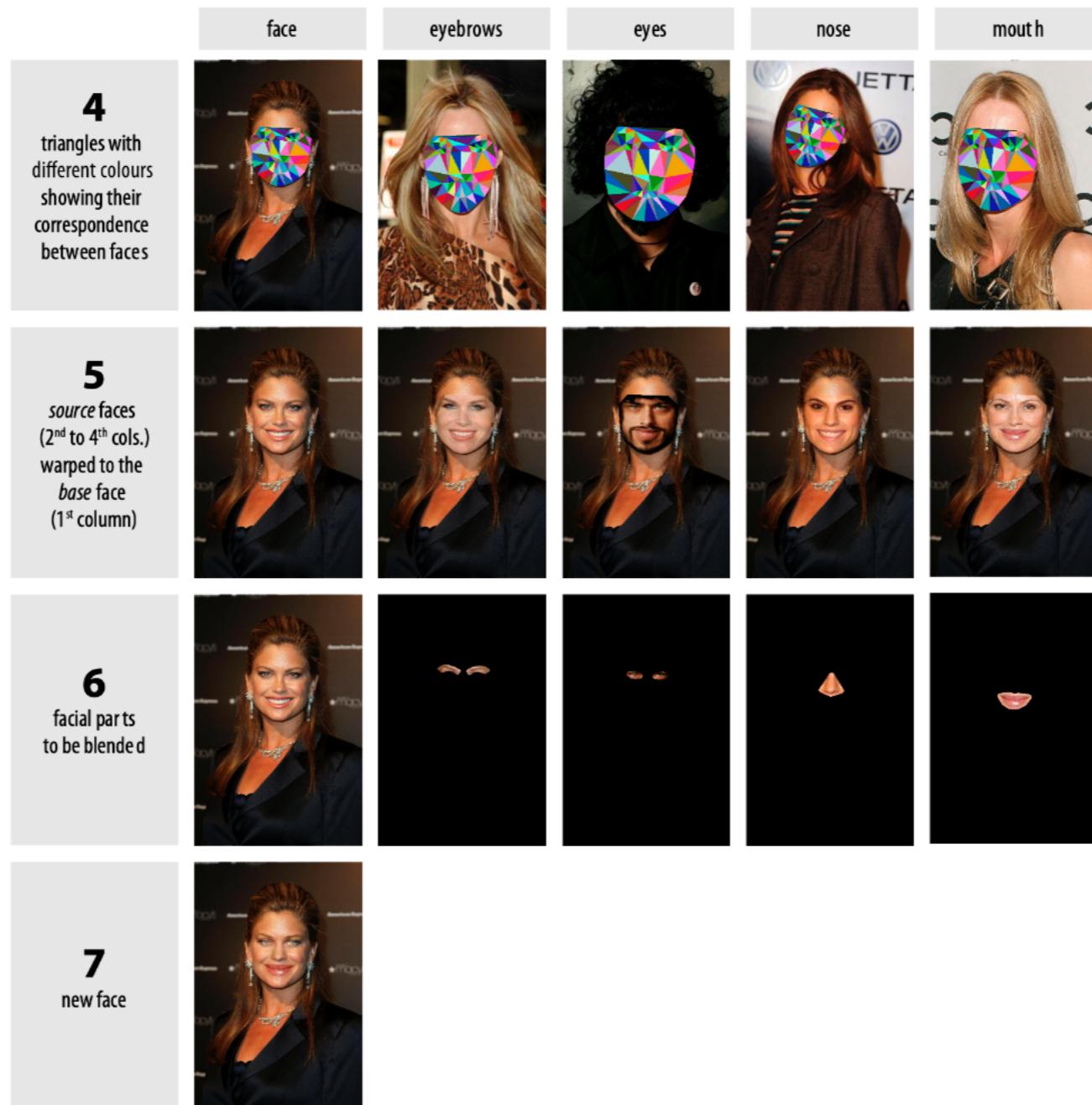
Multiple Swap Generation



**Evolved photorealistic faces not detect as faces**



## Unconstrained Generation



## Unconstrained Generation



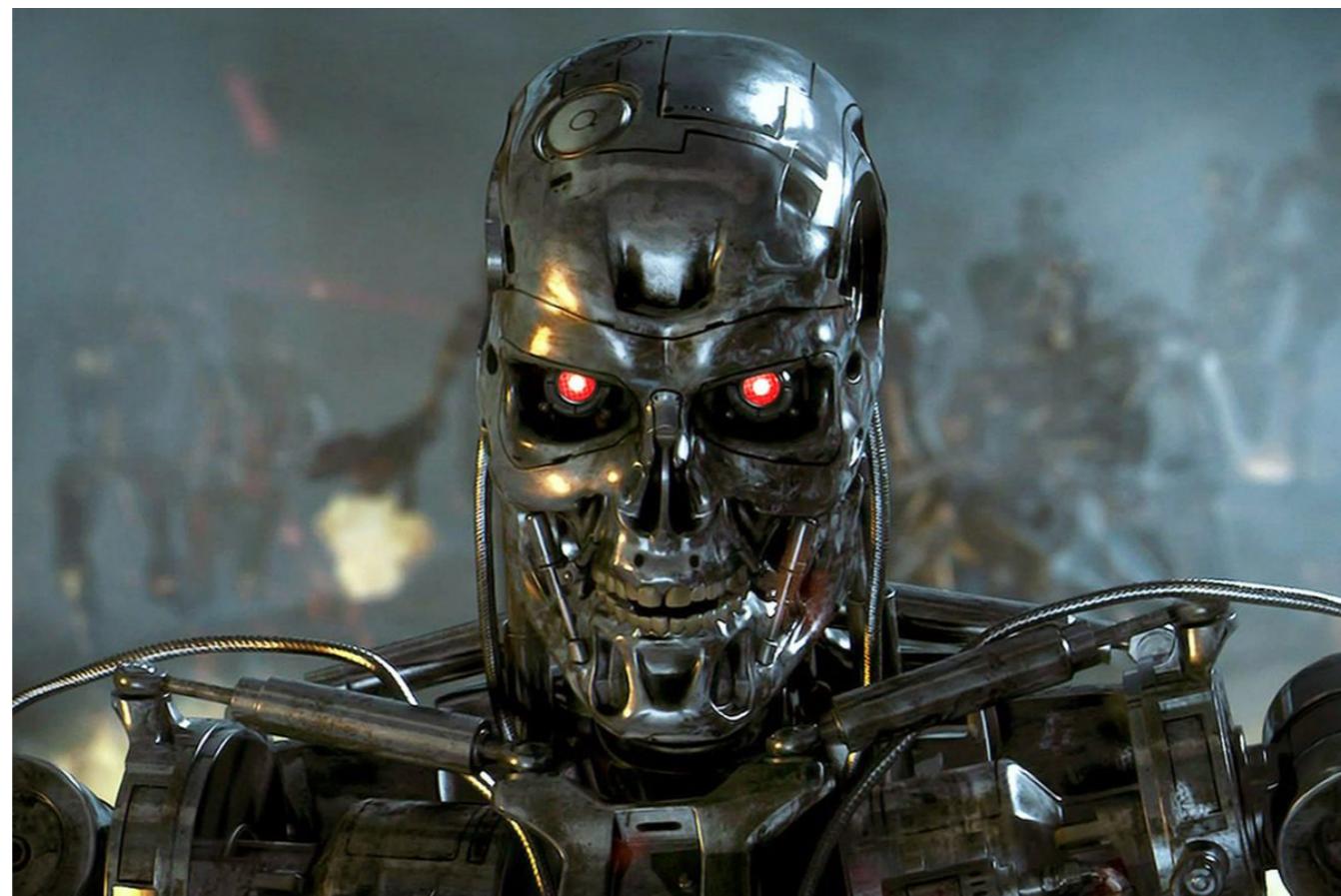
# Closing Remarks

# Closing Remarks

- Adversarial Learning is a field of interest for the Machine Learning community
- Adversarial Examples come in many forms
  - Its easy to create and attack
  - Research invested on preventive and defense mechanisms

# Closing Remarks

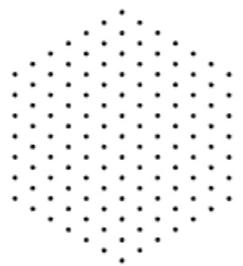
- Don't worry... we will also work towards preventing an AI takeover.



# Real or Fake? Overview on Adversarial Examples

João Nuno Correia

[jncor@dei.uc.pt](mailto:jncor@dei.uc.pt)



COMPUTATIONAL  
DESIGN &  
VISUALIZATION  
LAB.

1 2 9 0



UNIVERSIDADE DE  
**COIMBRA**