Real of Fake?
Overview on Adversarial Examples

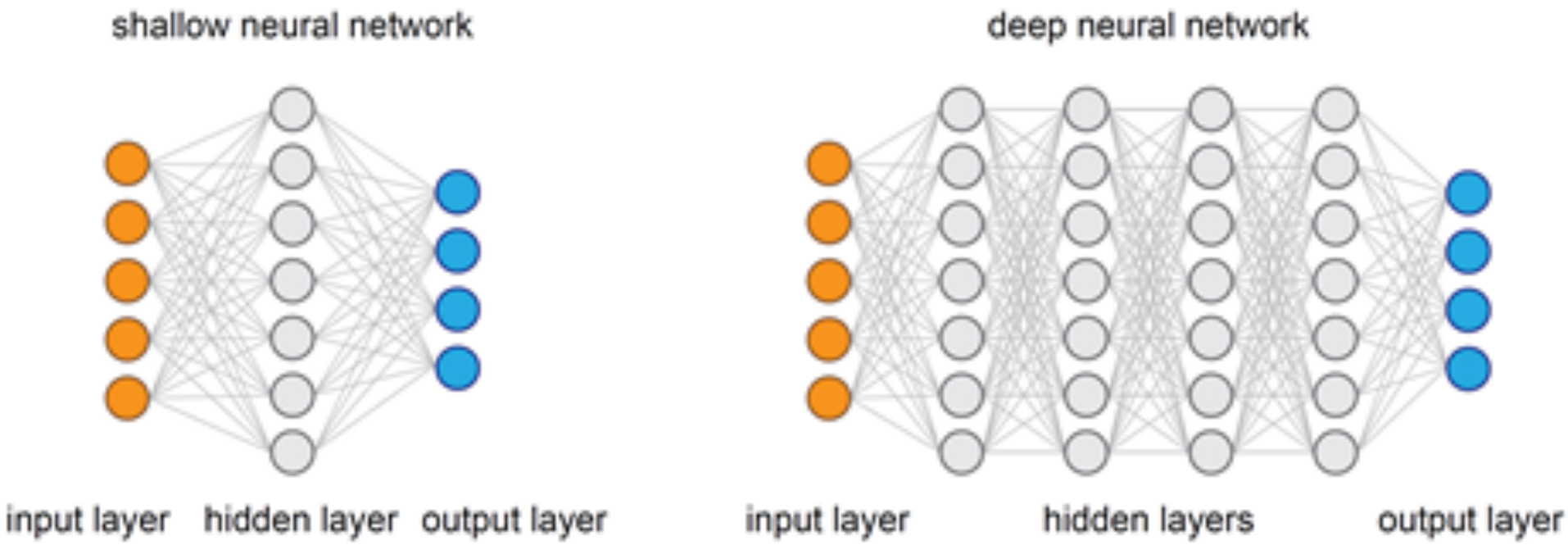
João Nuno Correia
jncor@dei.uc.pt

COMPUTATIONAL DESIGN & VISUALIZATION LAB.

1 2 9 0 UNIVERSIDADE D COIMBRA

Deep Learning

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

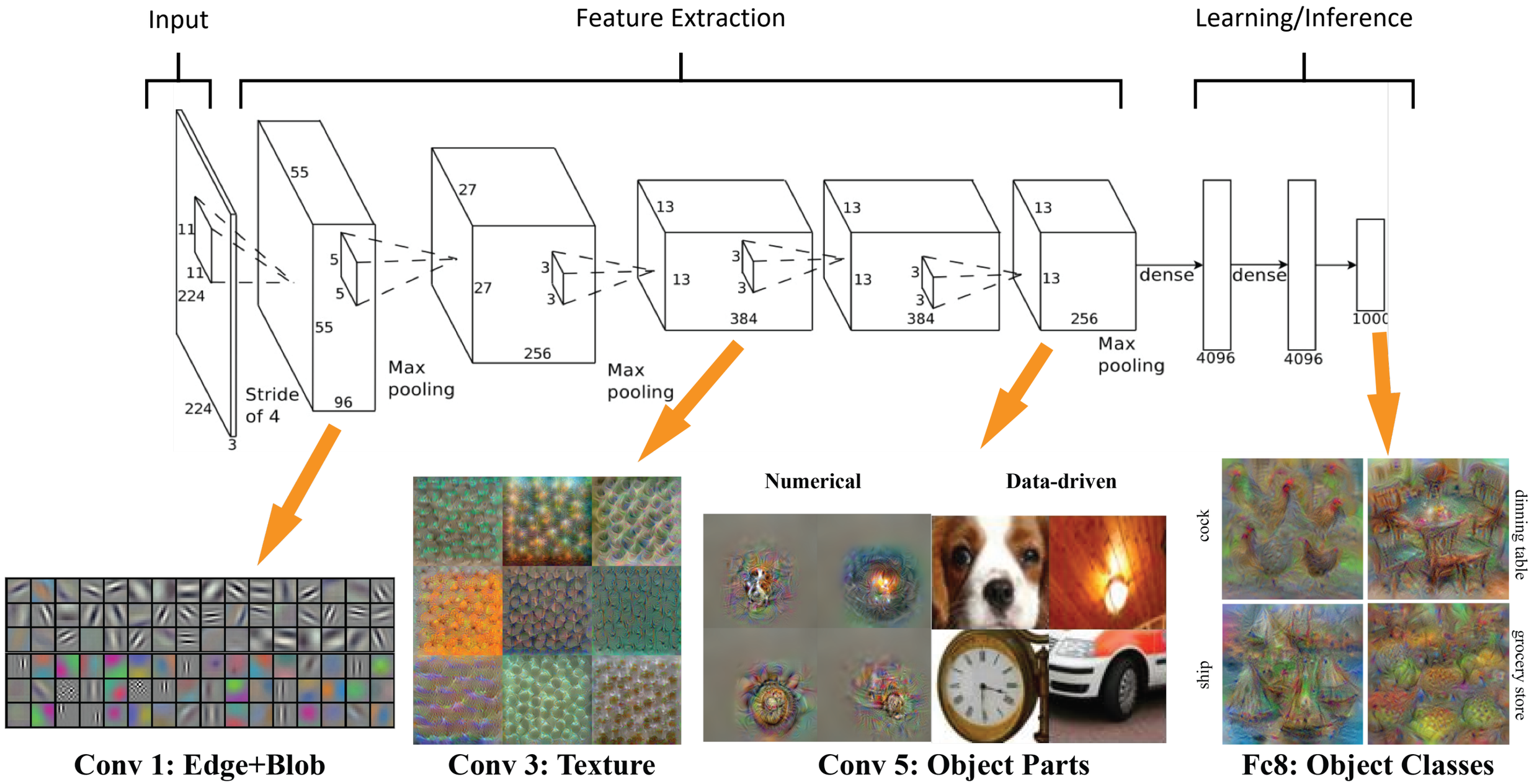


Why the (Deep) hype?

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

- Technology advancements

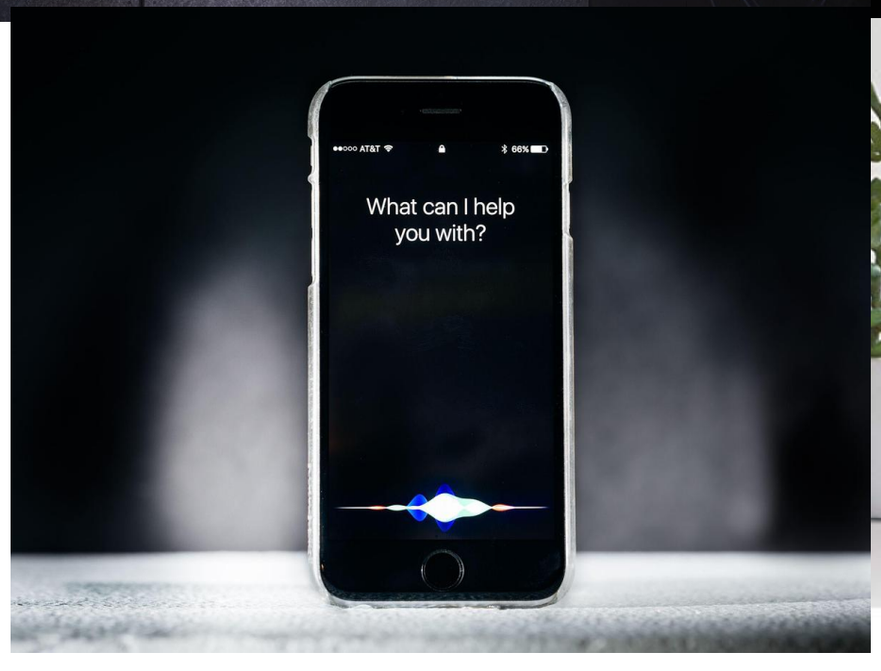
- And…



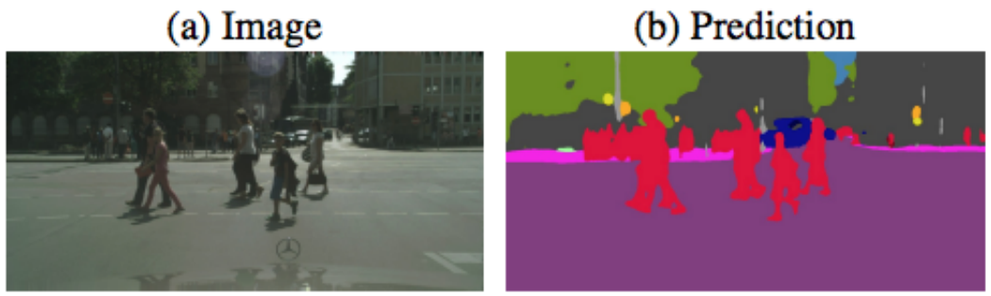
Feature Extraction and Learning! The Alexnet Overview. Krizhevsky et al.)

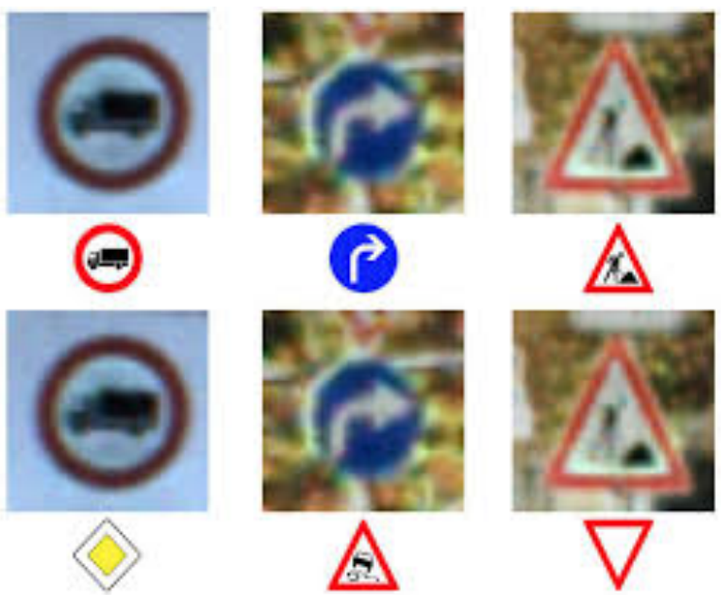
We think we solved it all…

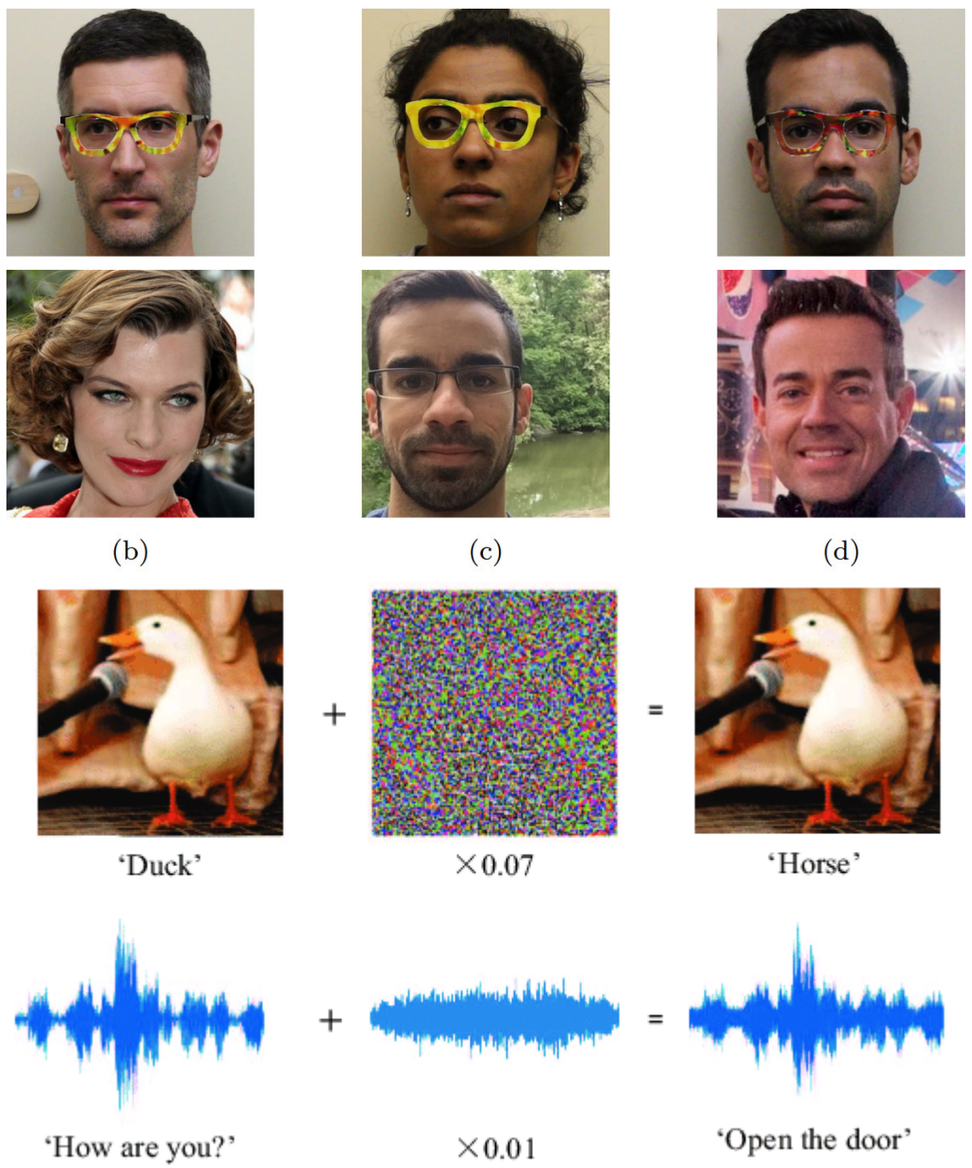
We think we solved it all…



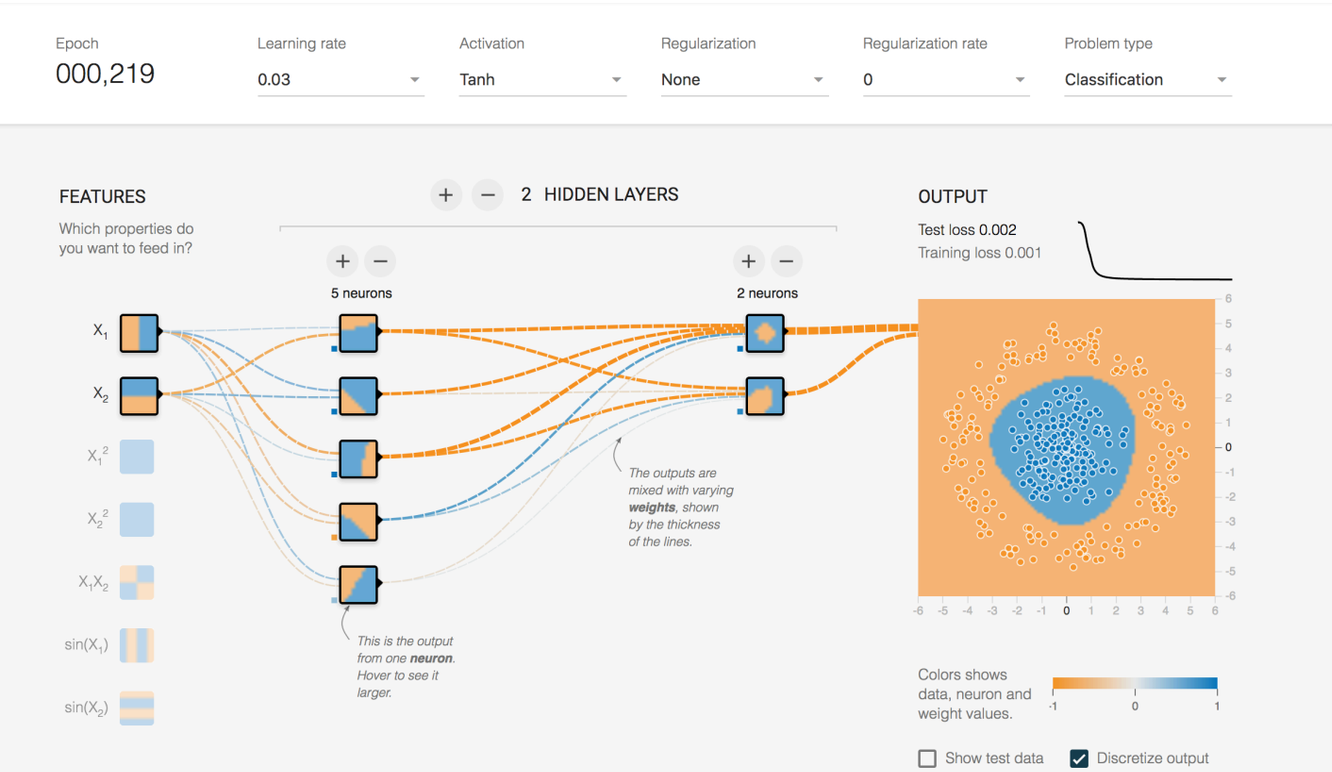
We think we solved it all… Not Quite

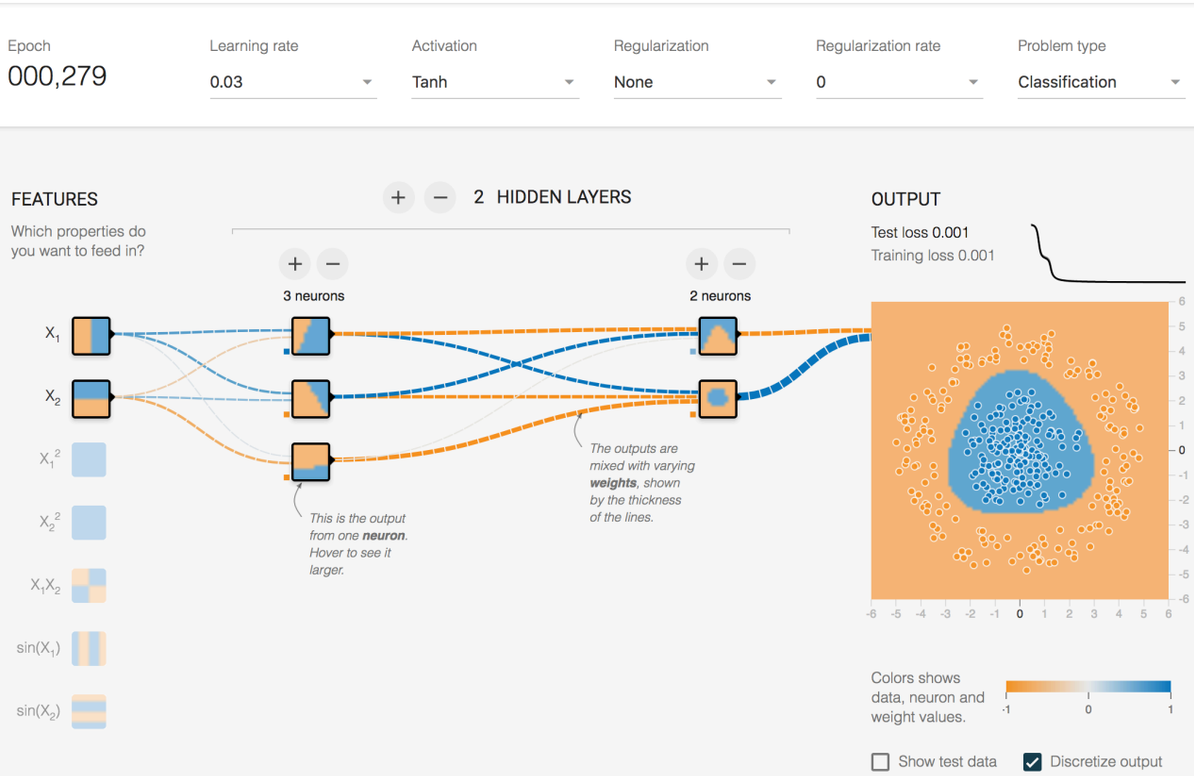




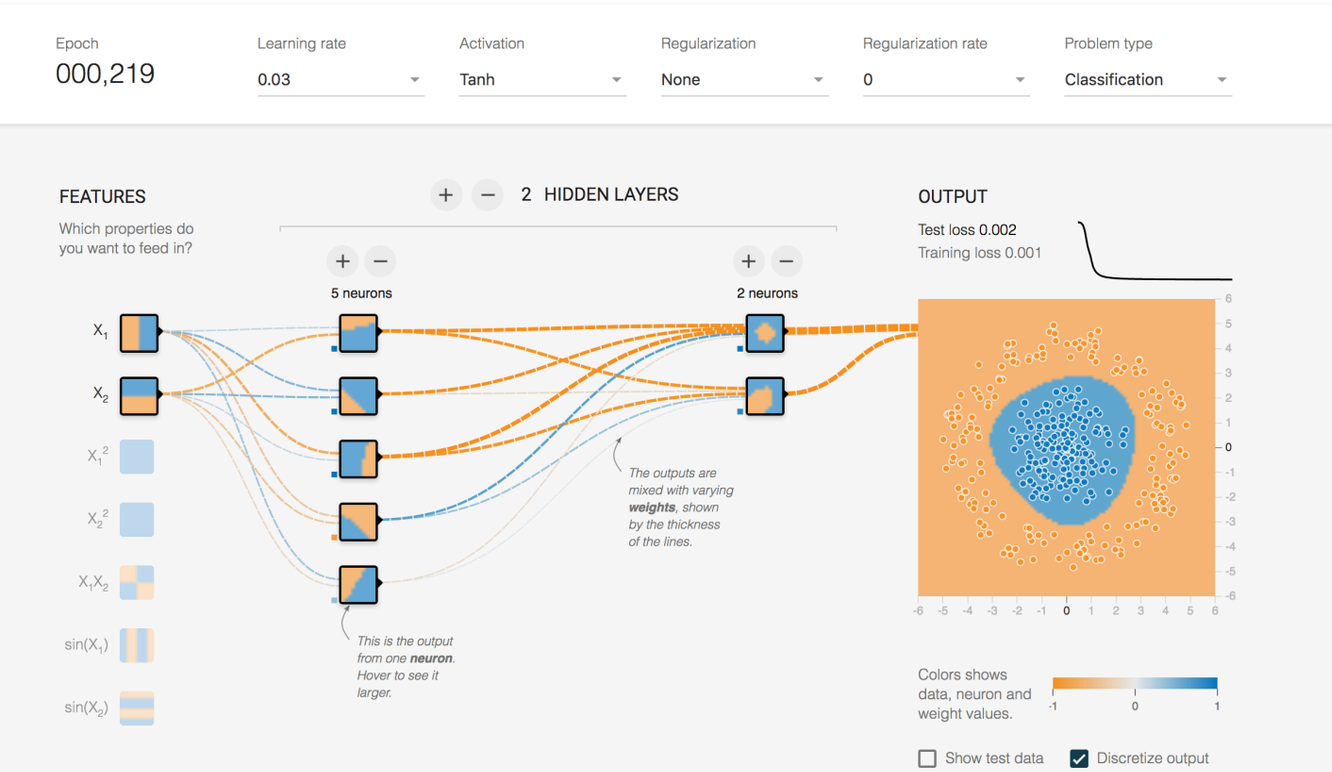


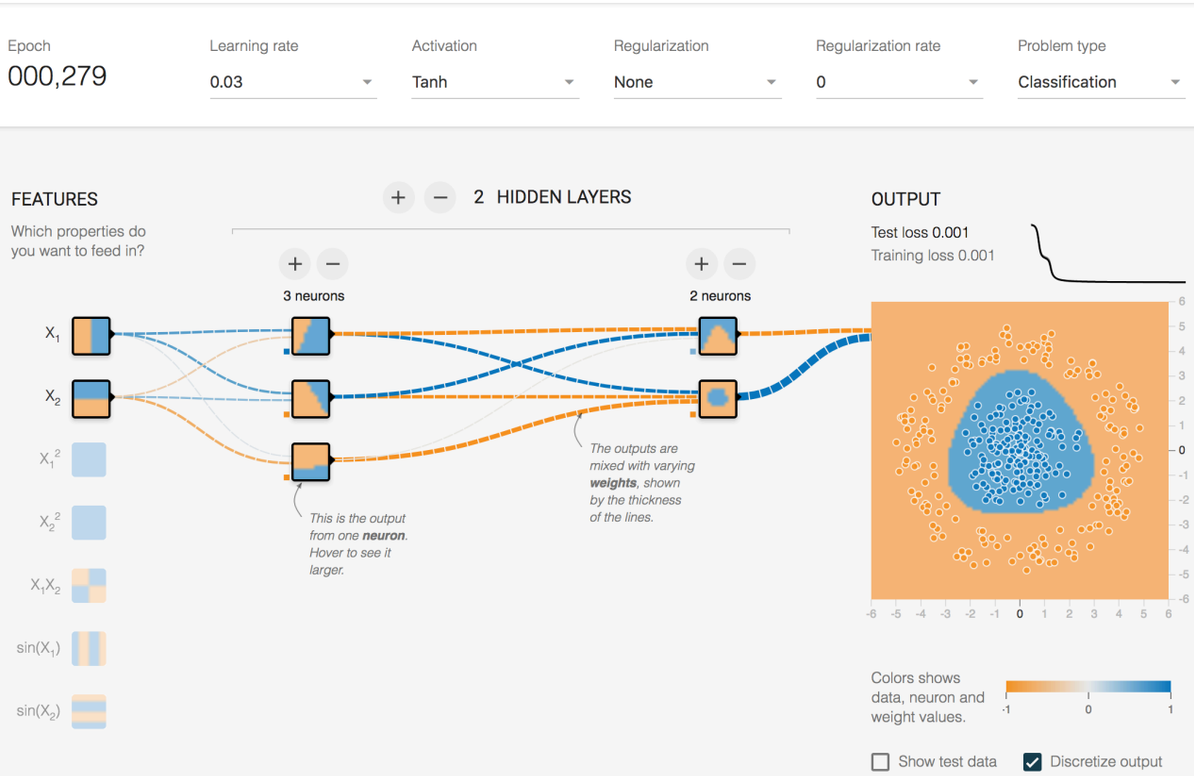
What are we talking about…

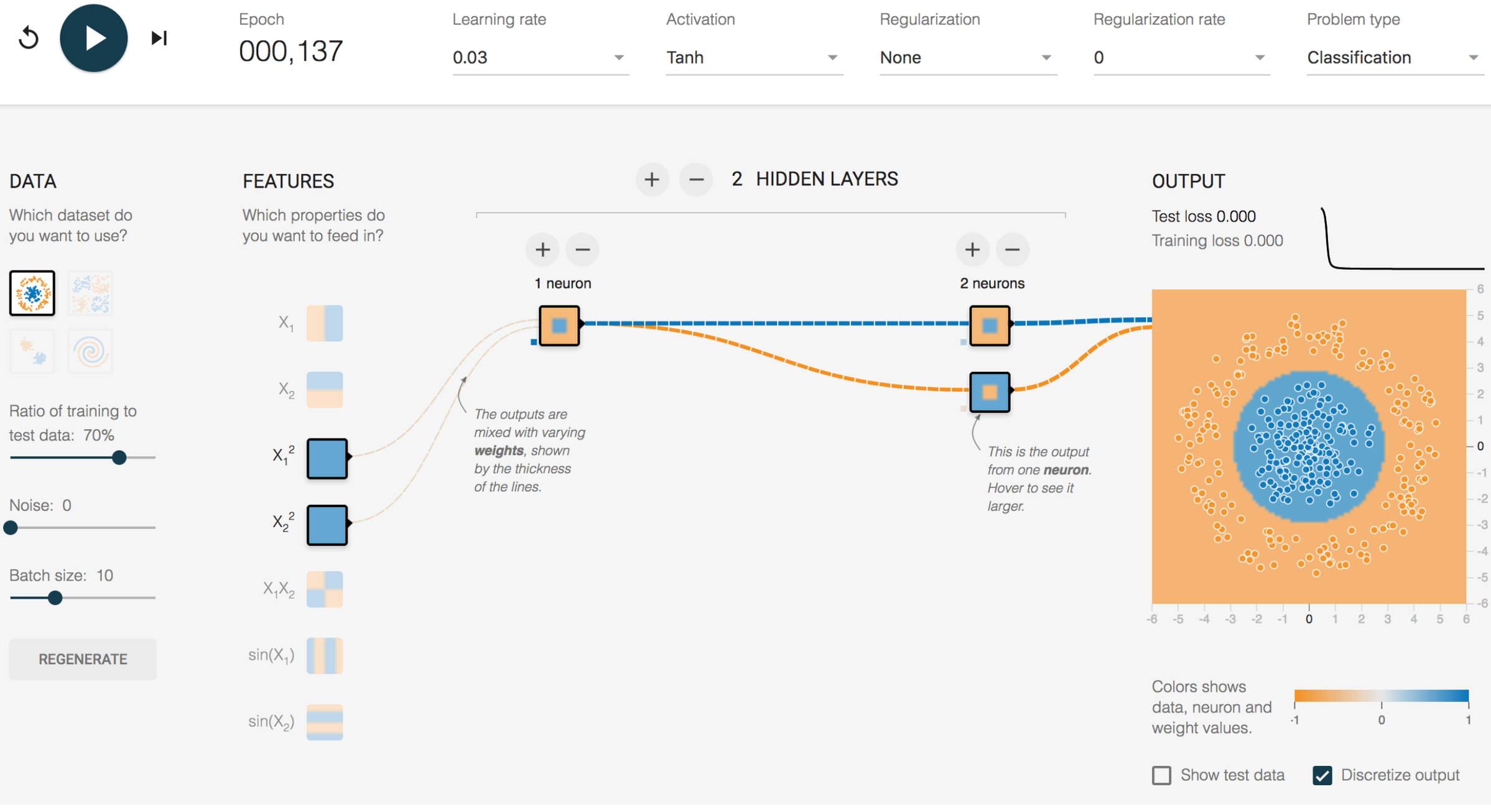




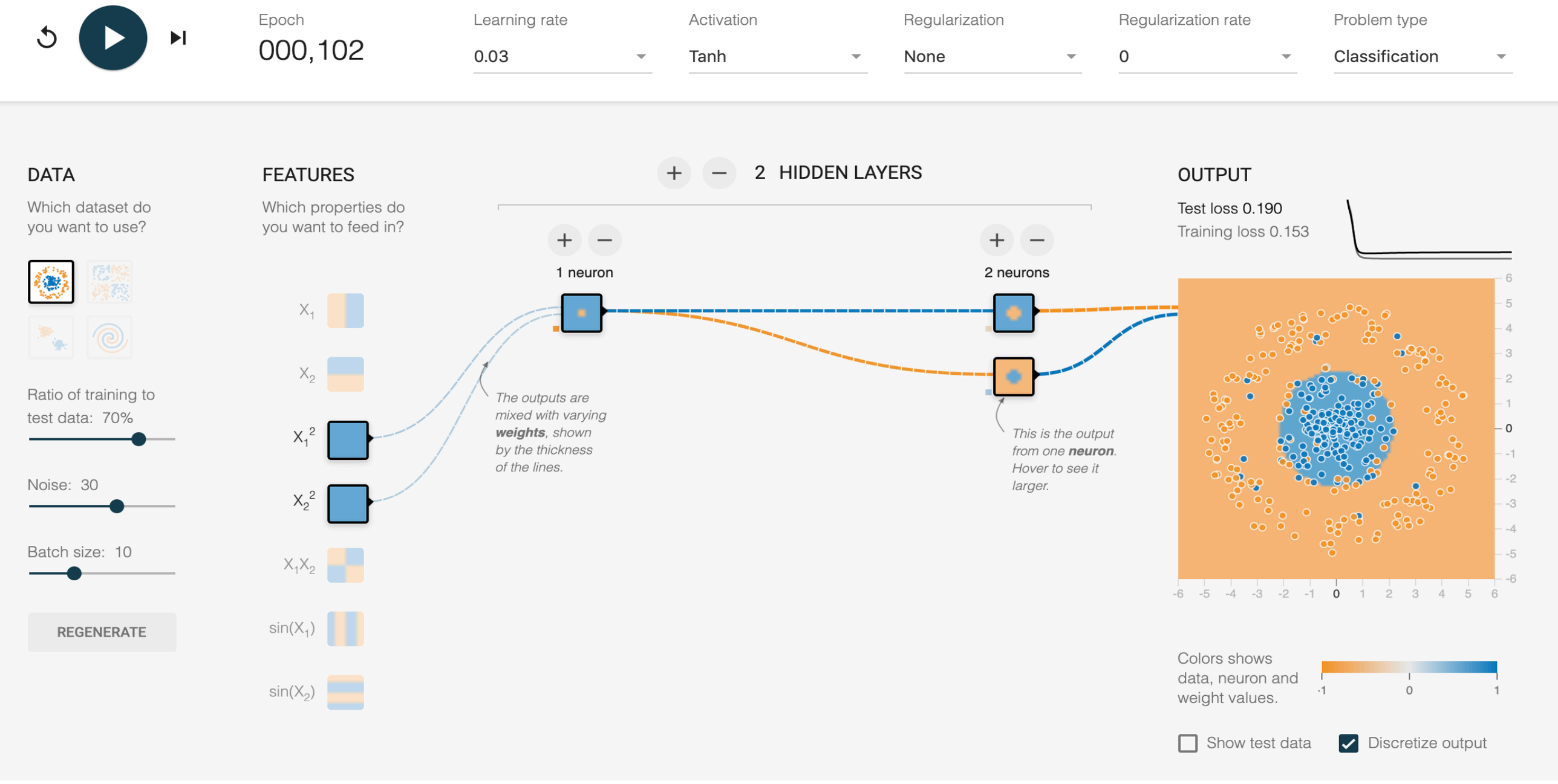
What are we talking about…







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

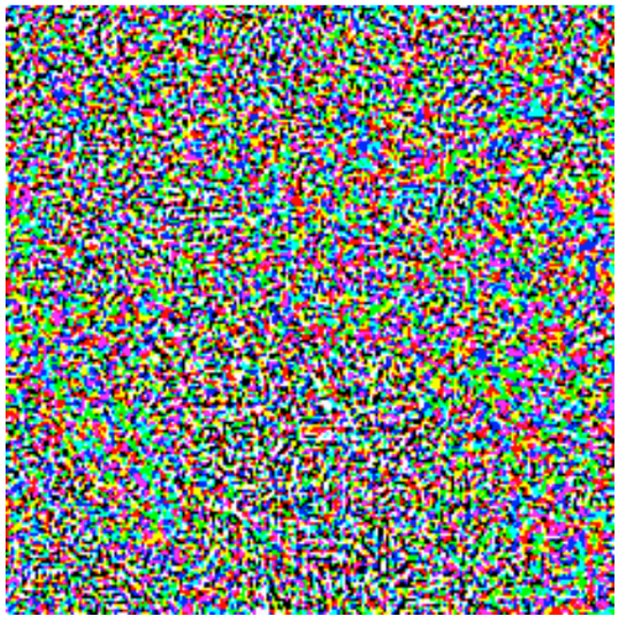
- Machine Learning and computer security
- Adoption of “safe” Machine Learning techniques in adversarial settings
  - E.g. Spam Filtering, Computer Security, Biometric recognition
- Studies the “how to attack" and "how to defend"
  - how to generate adversarial data

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

$+ .007 \times$



$=$

$\boldsymbol{x}$

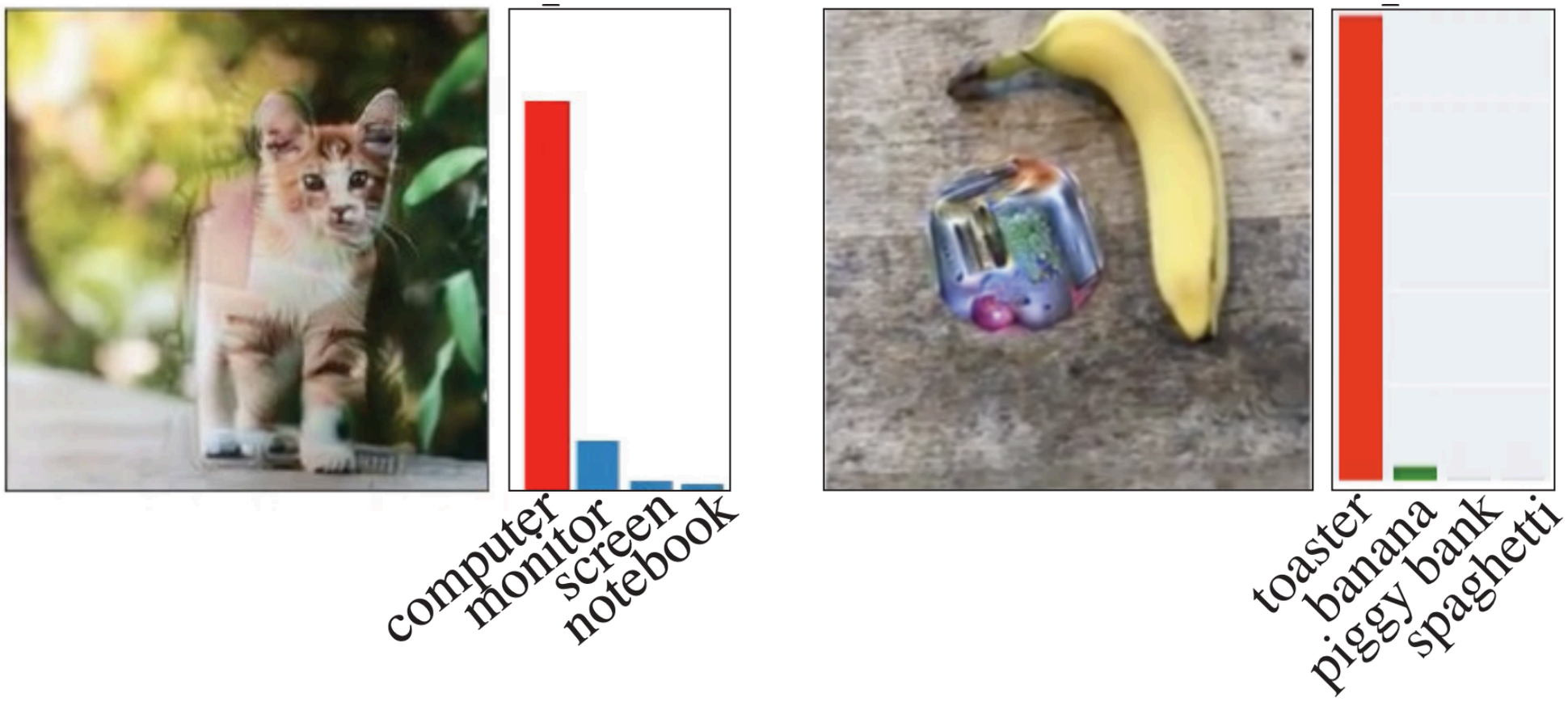
"panda"
57.7% confidence

$\text{sign}(\nabla_{\boldsymbol{x}} J(\boldsymbol{\theta}, \boldsymbol{x}, y))$

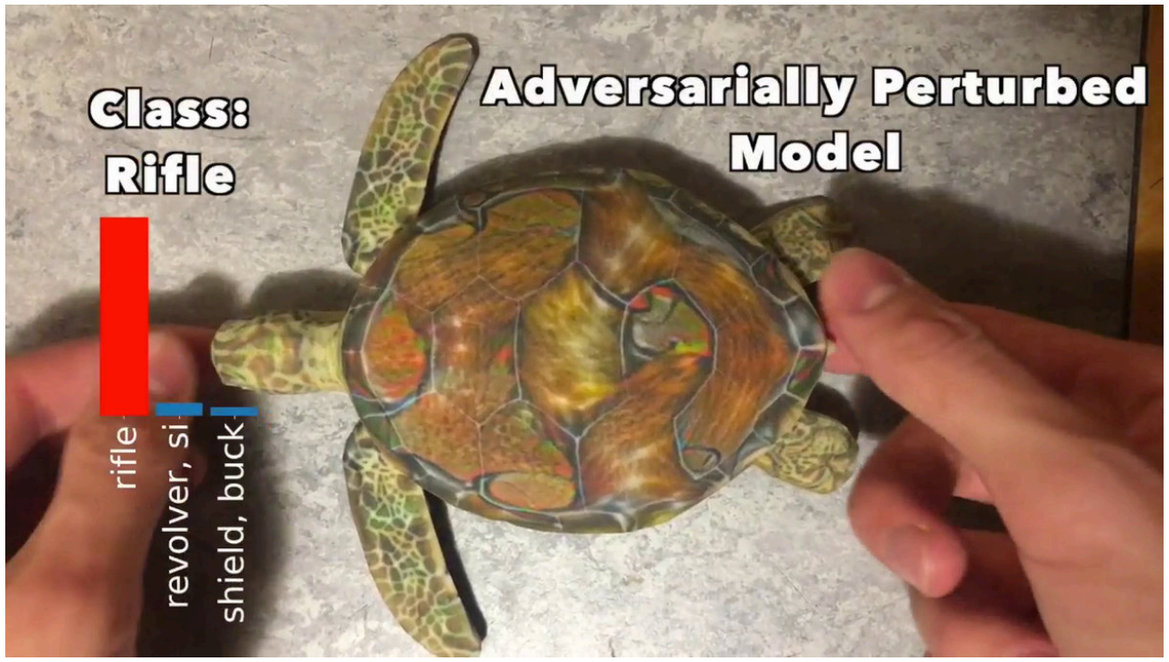
"nematode"
8.2% confidence

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

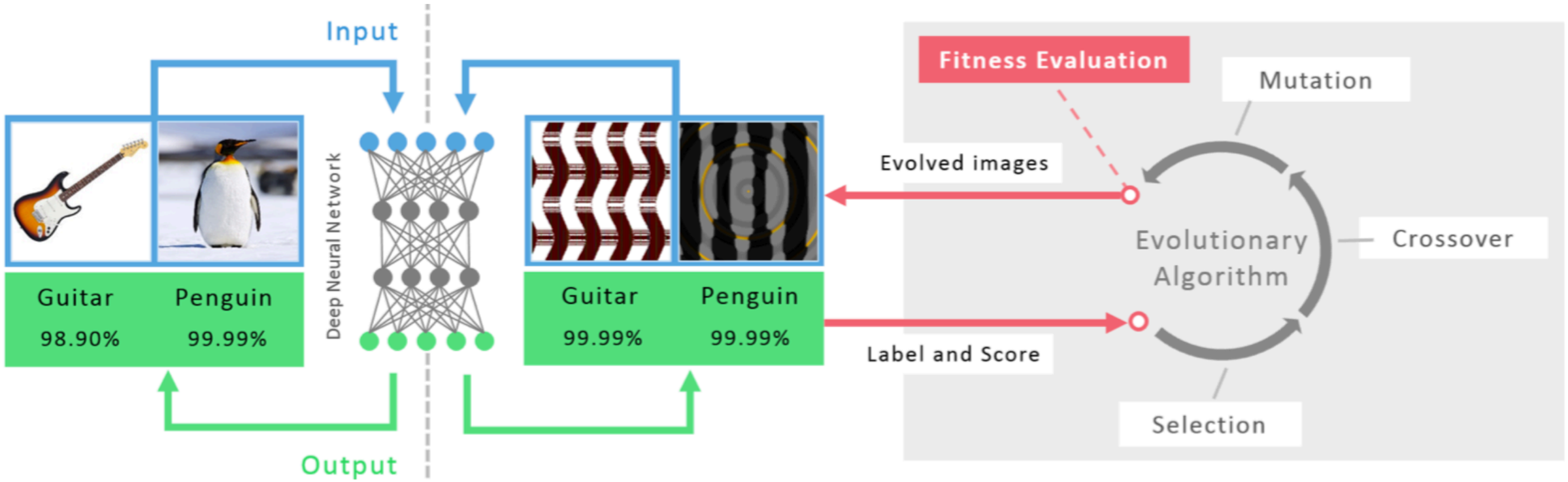
Addition of noise in a training example



Several examples of adversarial example generation by tampering with the image

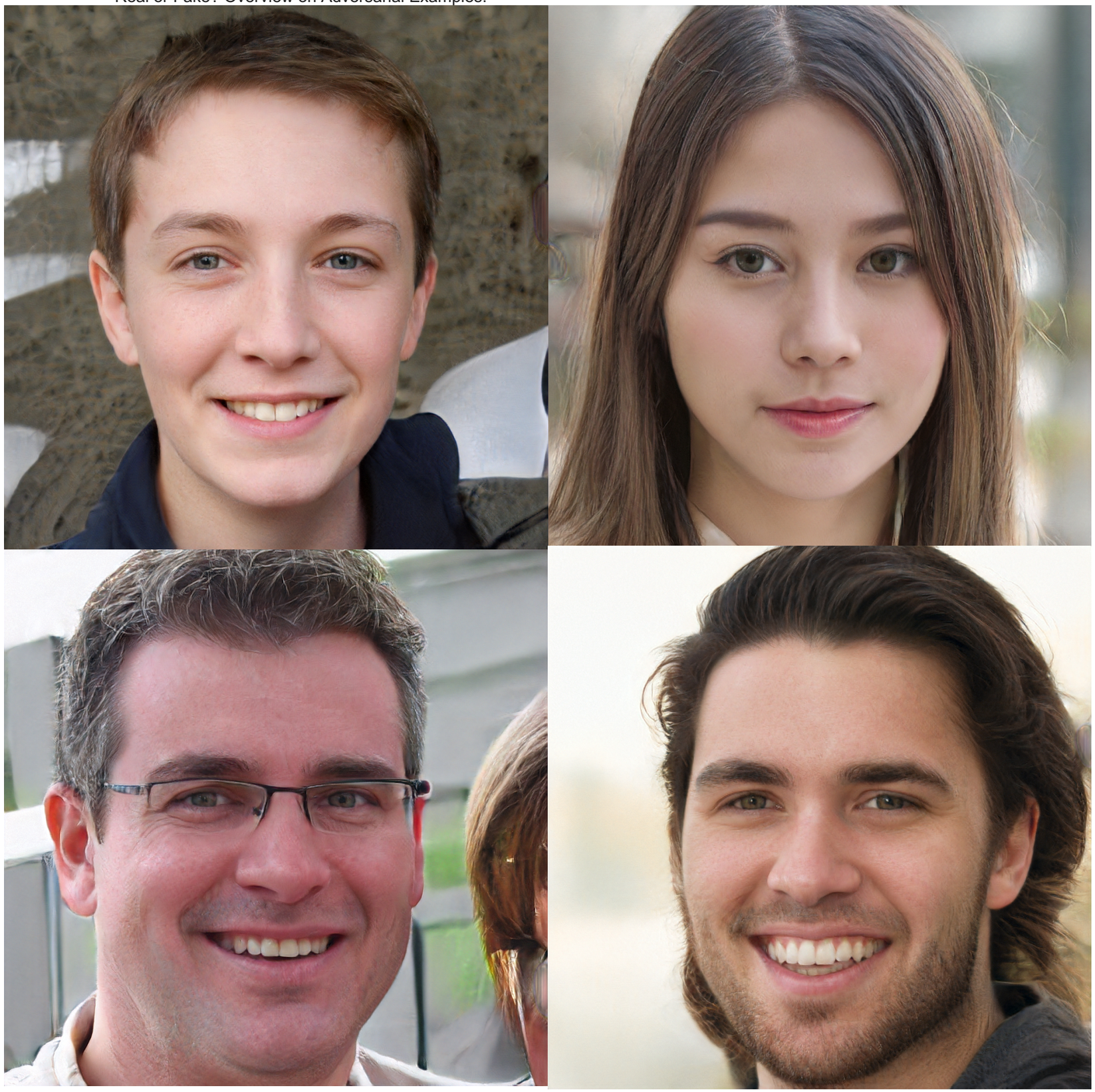


Several examples of adversarial example generation by tampering with the image



DNNs Easily Fooled (Nguyen et al.)

Real or Fake? Overview on Adversarial Examples.



https://thispersondoesnotexist.com/

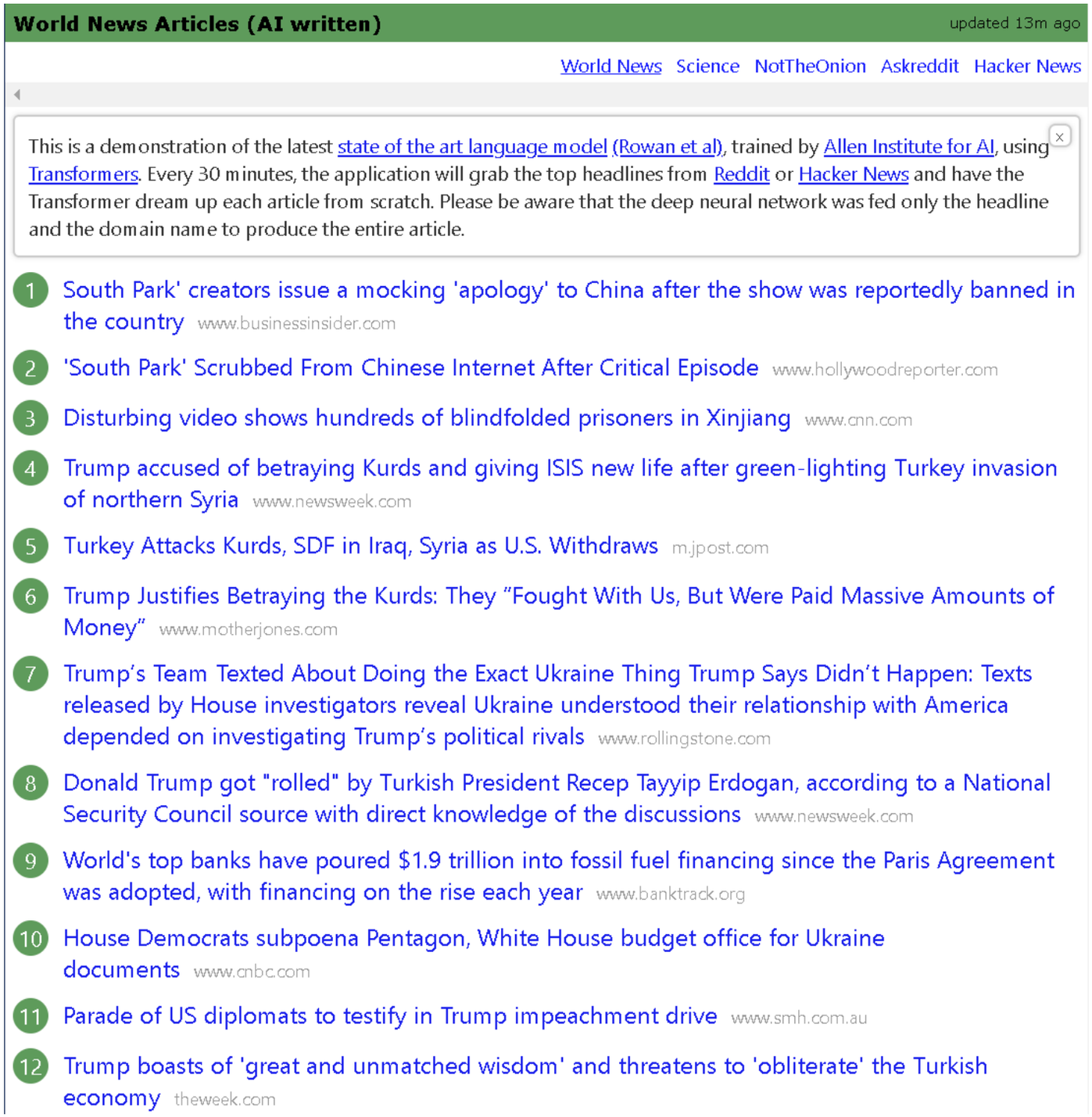
South Africa's historic Soweto township marks its 100th birthday on Tuesday in a mood of optimisim. 57% **World**

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

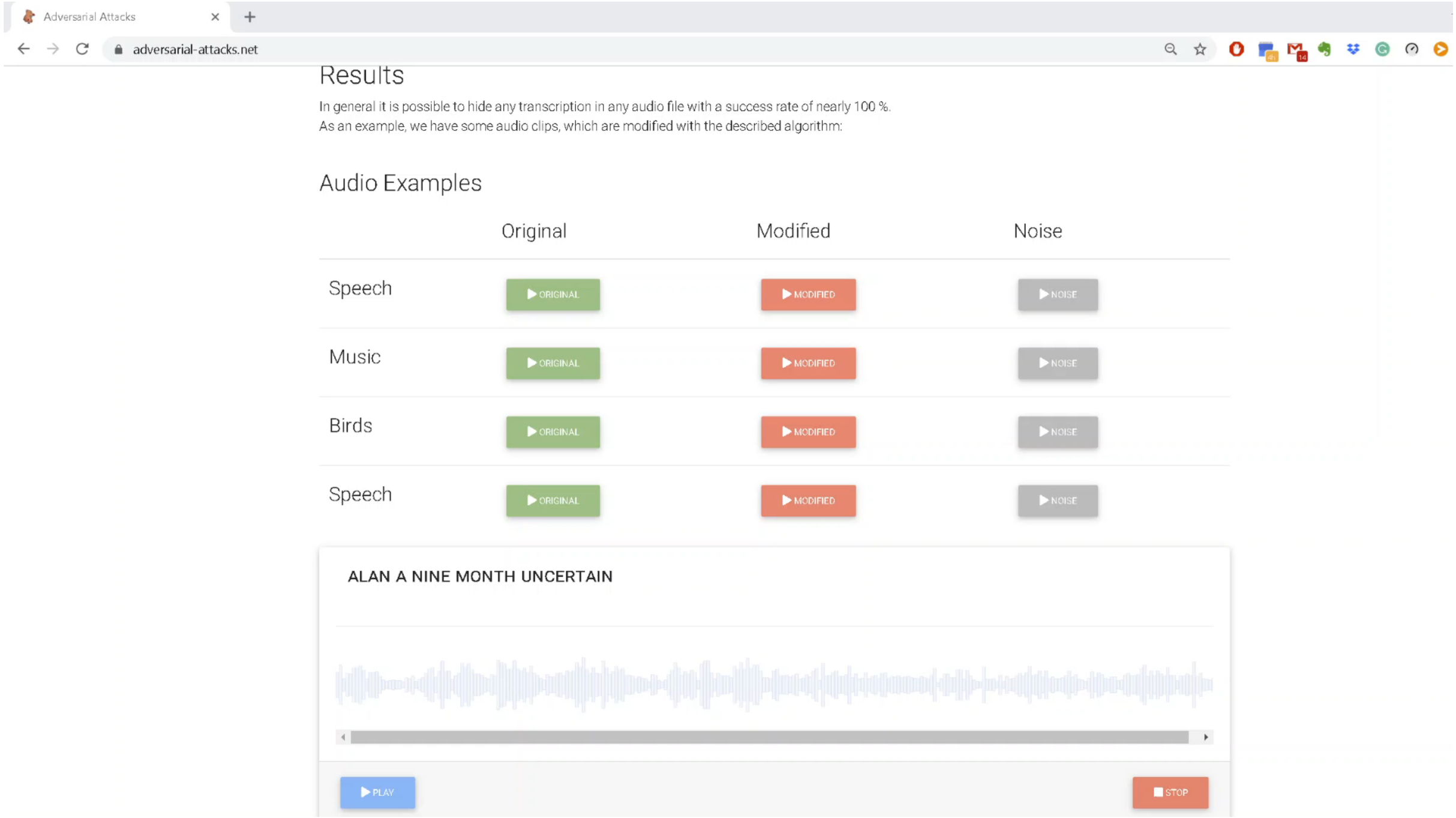
it's frustrating to see these guys who are obviously pretty clever waste their talent on parodies of things they probably thought were funniest when they were high. 83% **Negative Sentiment**

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**

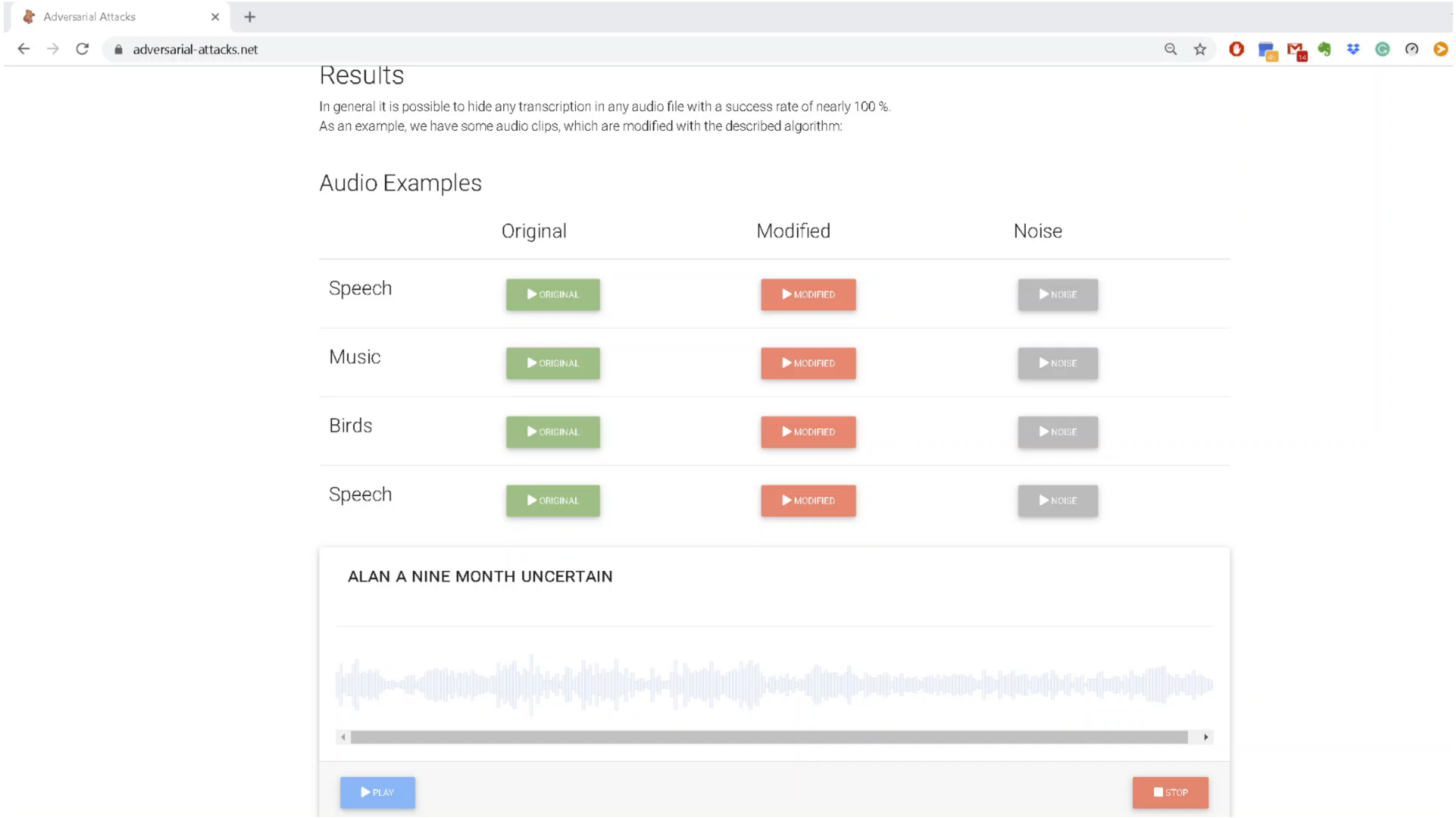
Deep networks for text analysis are also vulnerable



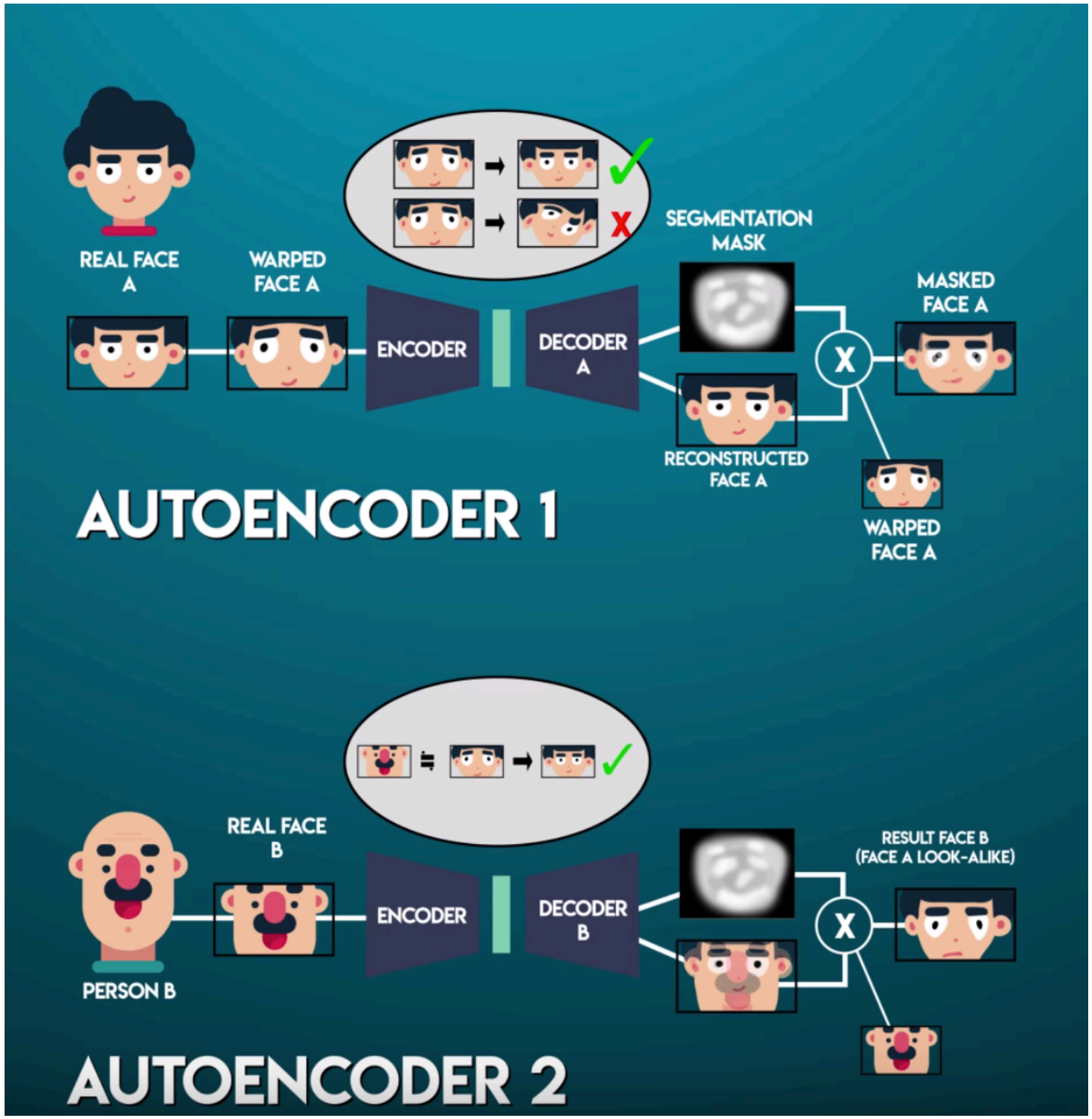
And some of these articles do not exist…



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



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.**

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

João Correia

Real or Fake? Overview on Adversarial Examples.

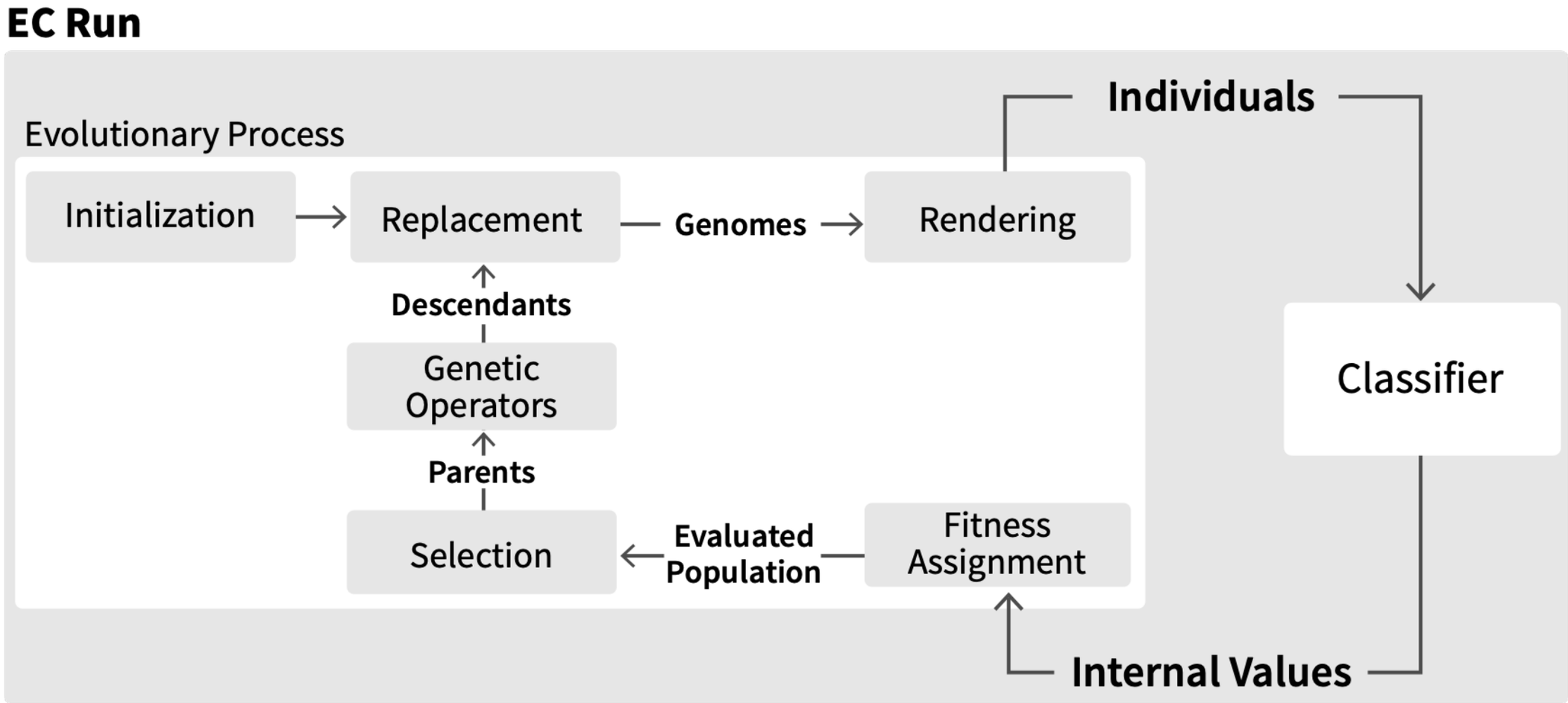
Evolving faces

Evolve images of faces that the classifier detect as such

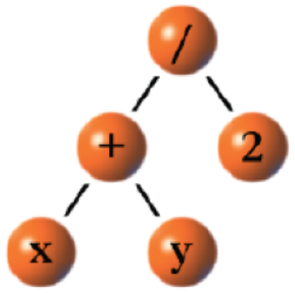
→ A general purpose expression-based evolutionary art tool

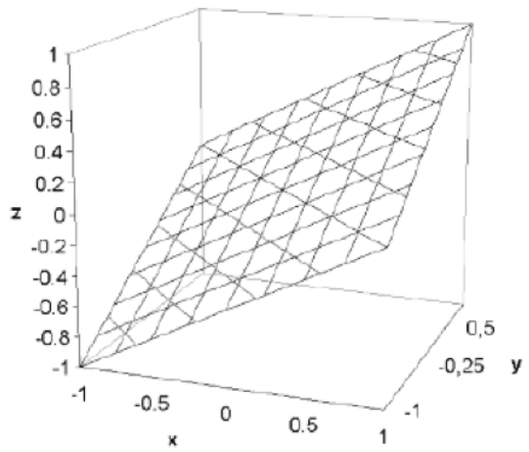
→ An off-the-shelf face detector as classifier

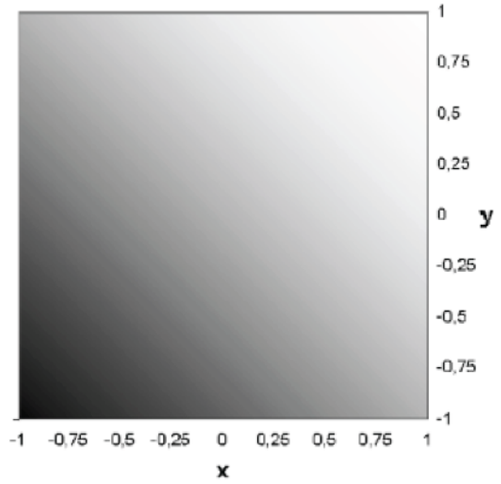
→ The classifier must detect faces in the evolved images

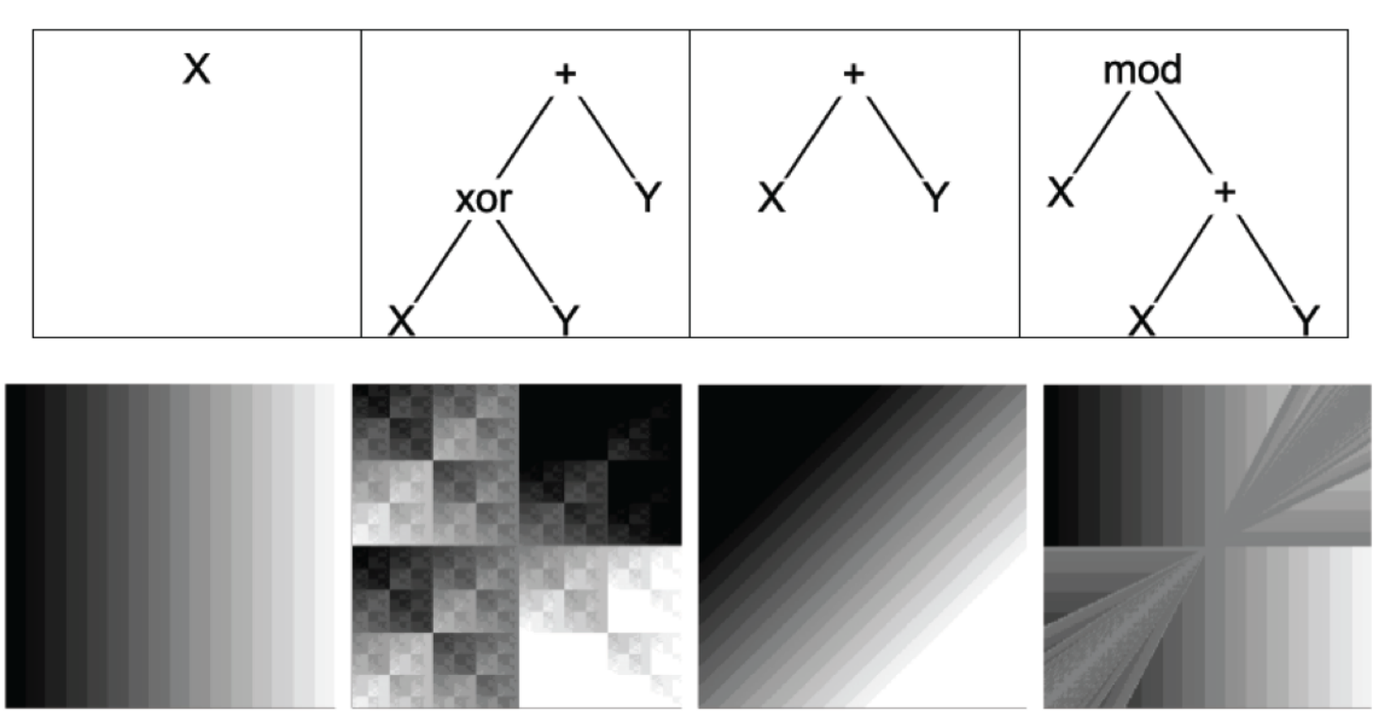


EFECTIVE overview

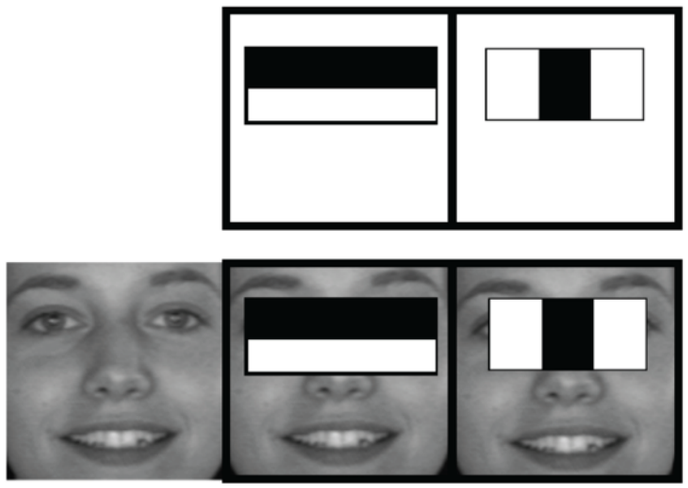


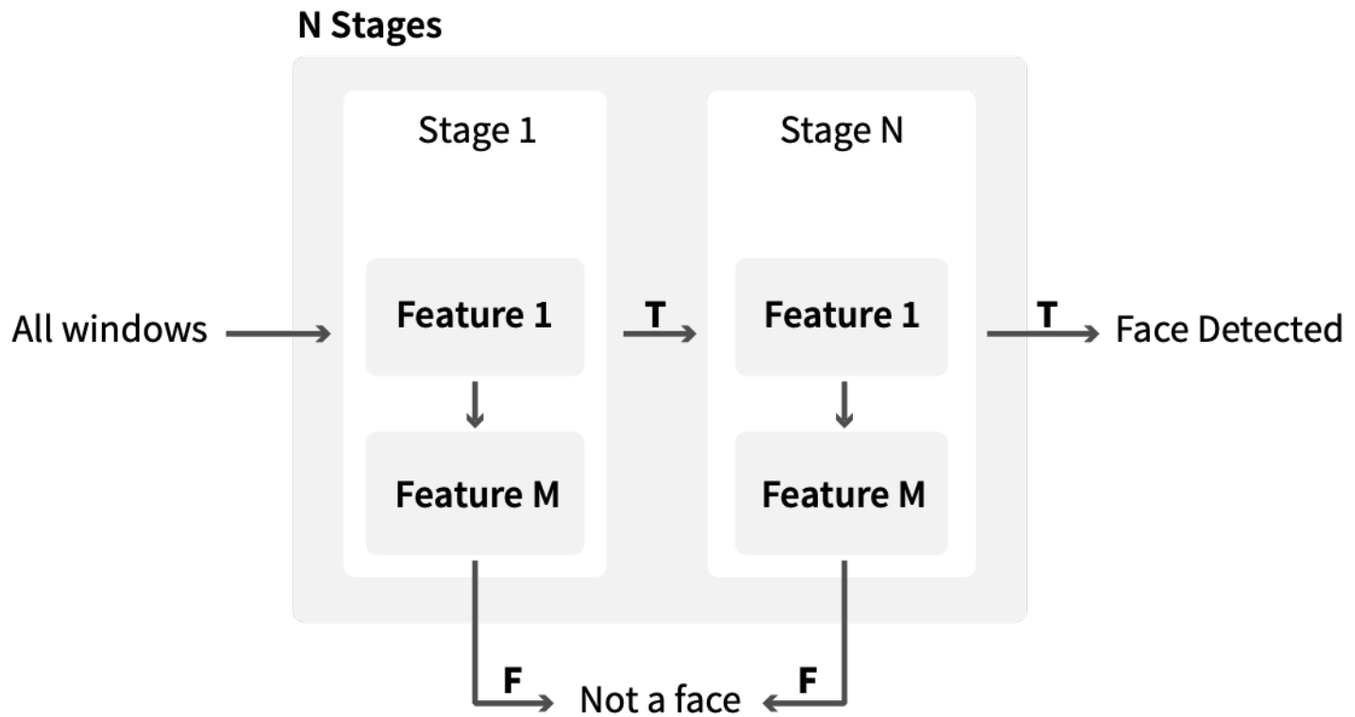


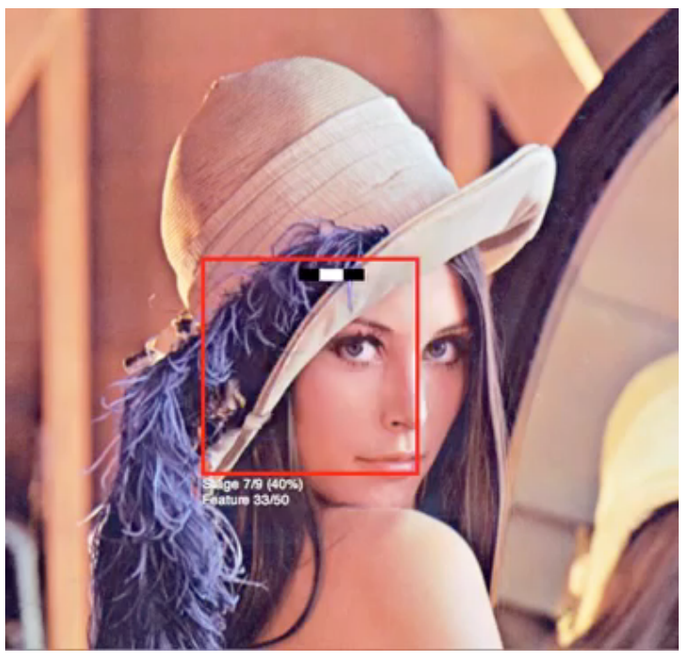




Genetic Programming engine

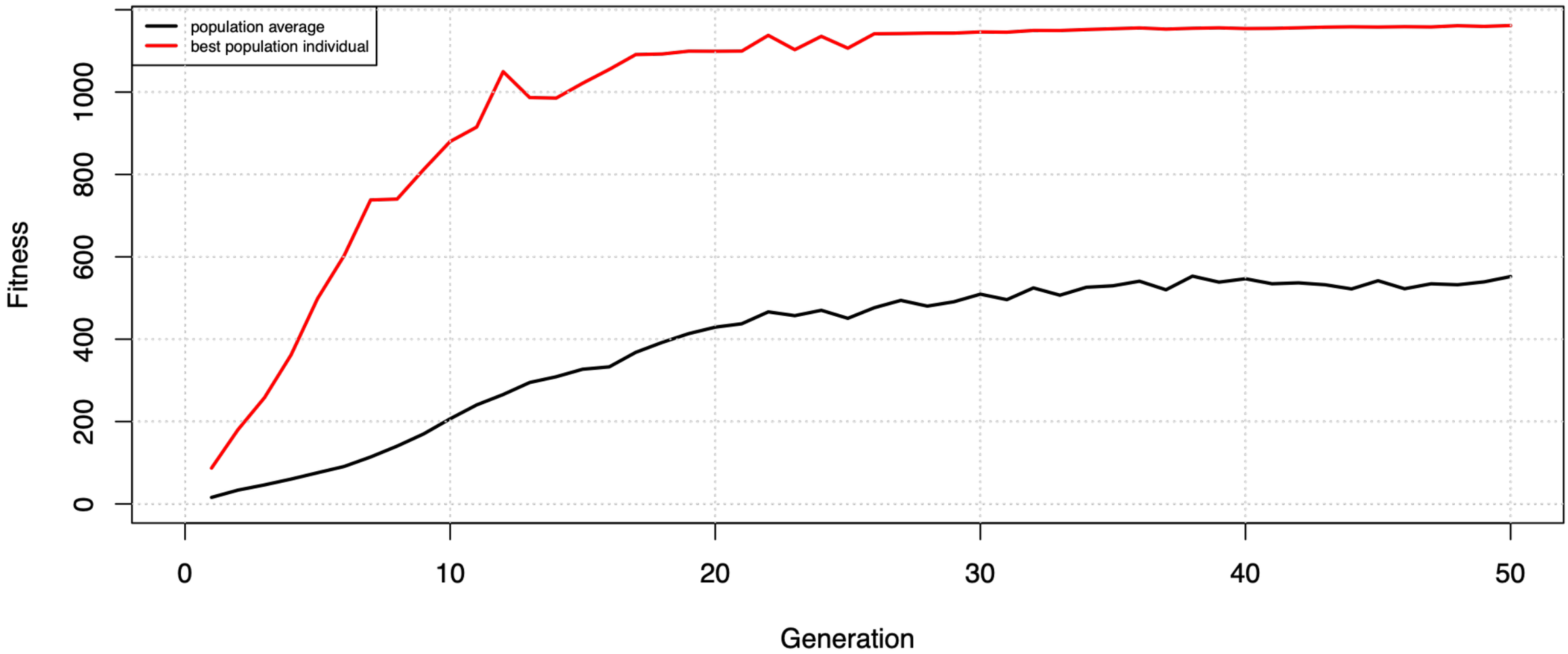




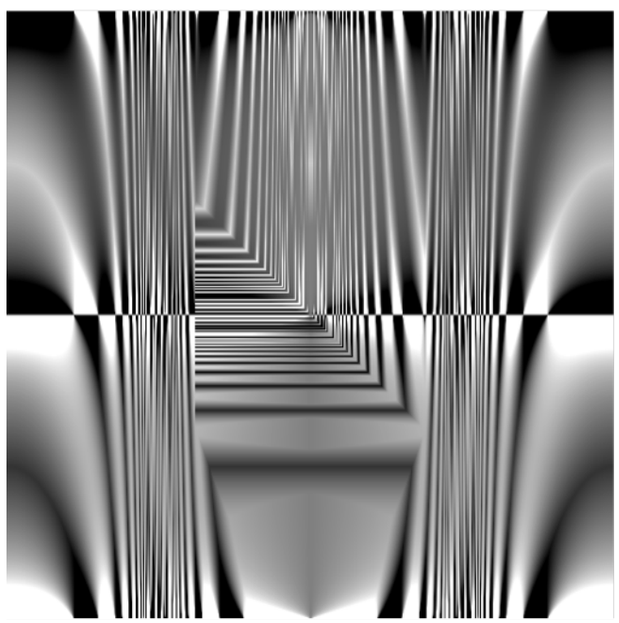
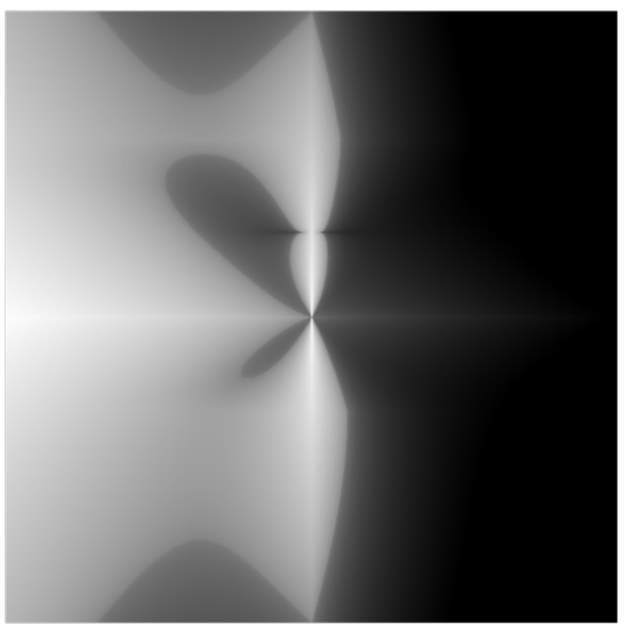
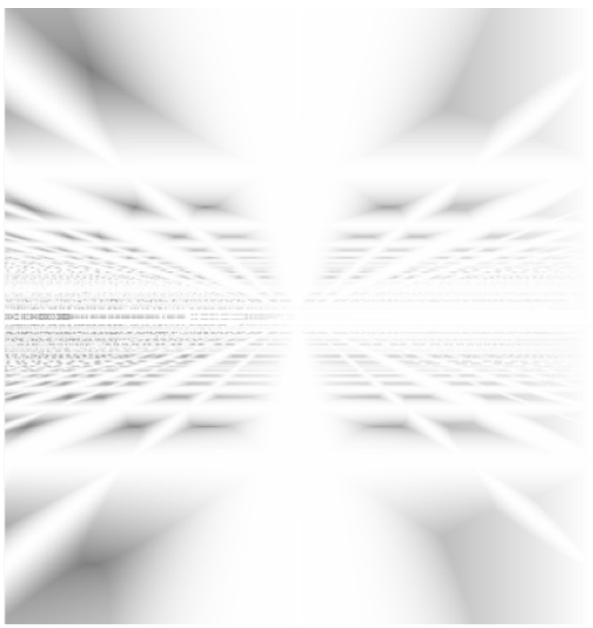
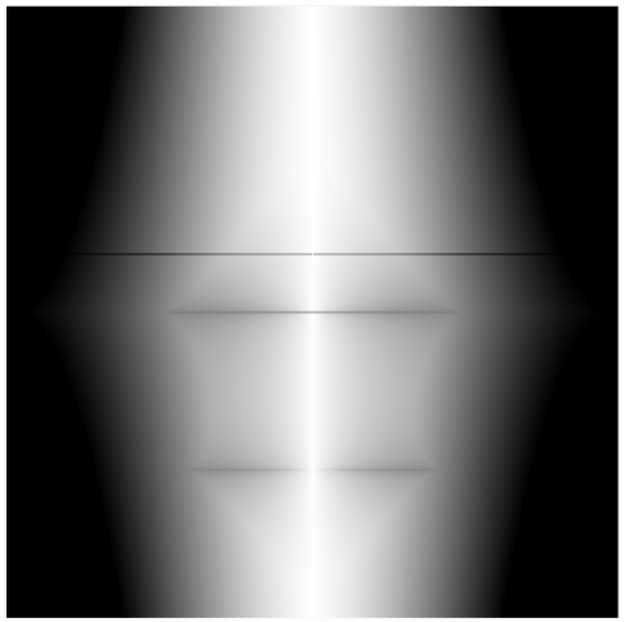
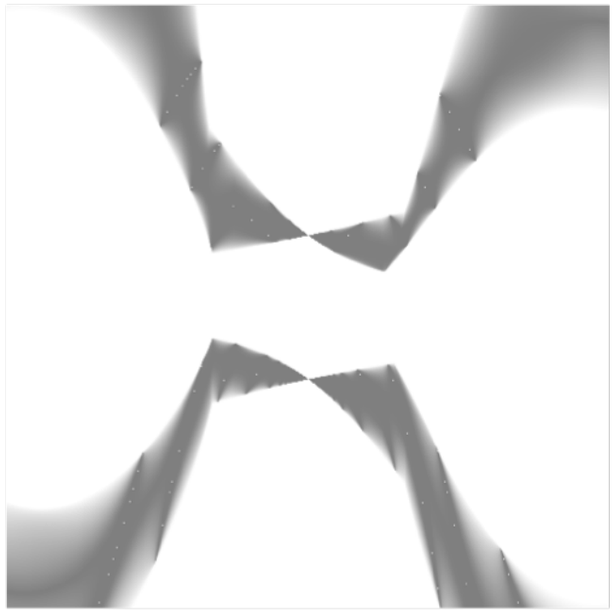
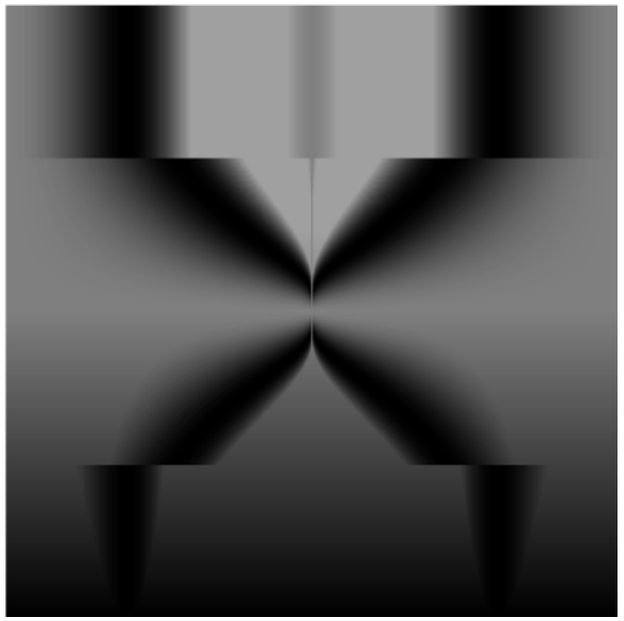


$$f(x) = \sum_{i}^{countstages_x} stagedif_x(i) * i + countstages_x * 10$$

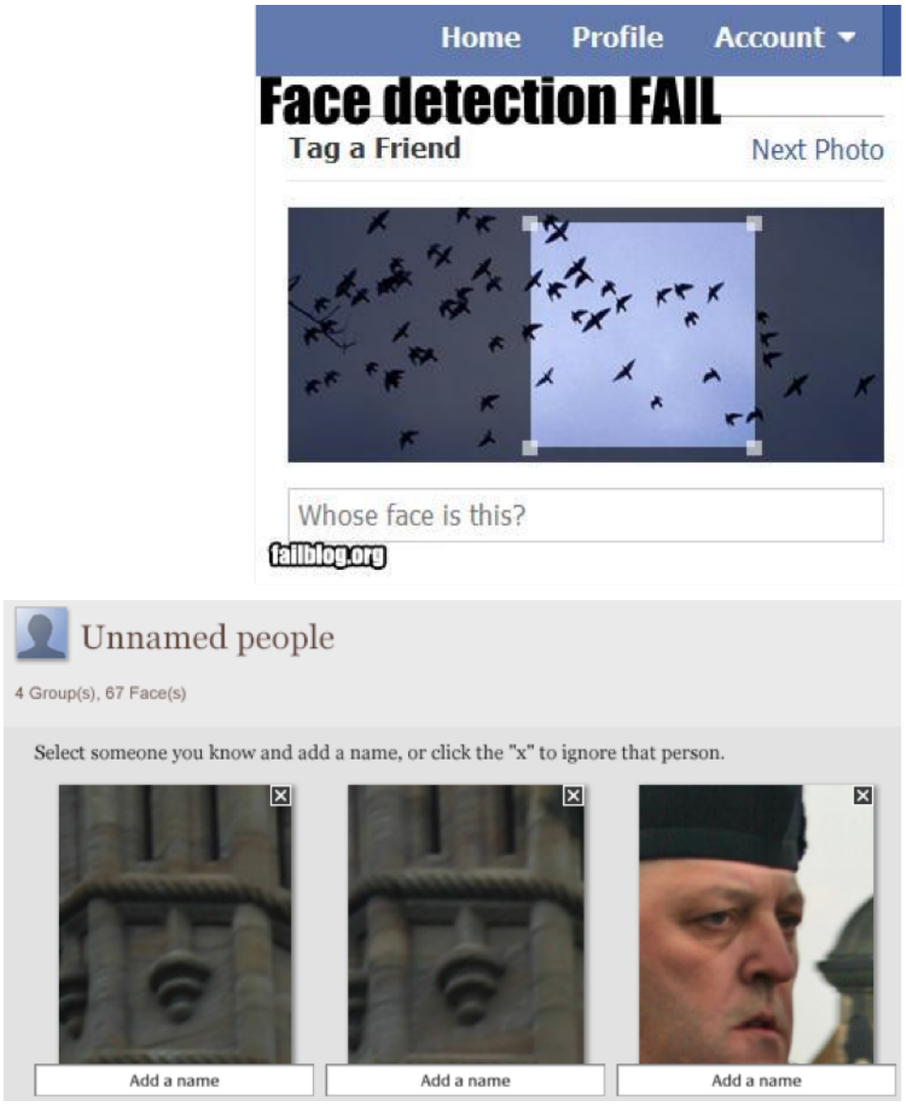
Haar features [top-left]; Cascade classifier [top-right]; Cascade classifier in action (video by Adam Harvey) [bottom-left]; fitness function [bottom-right]

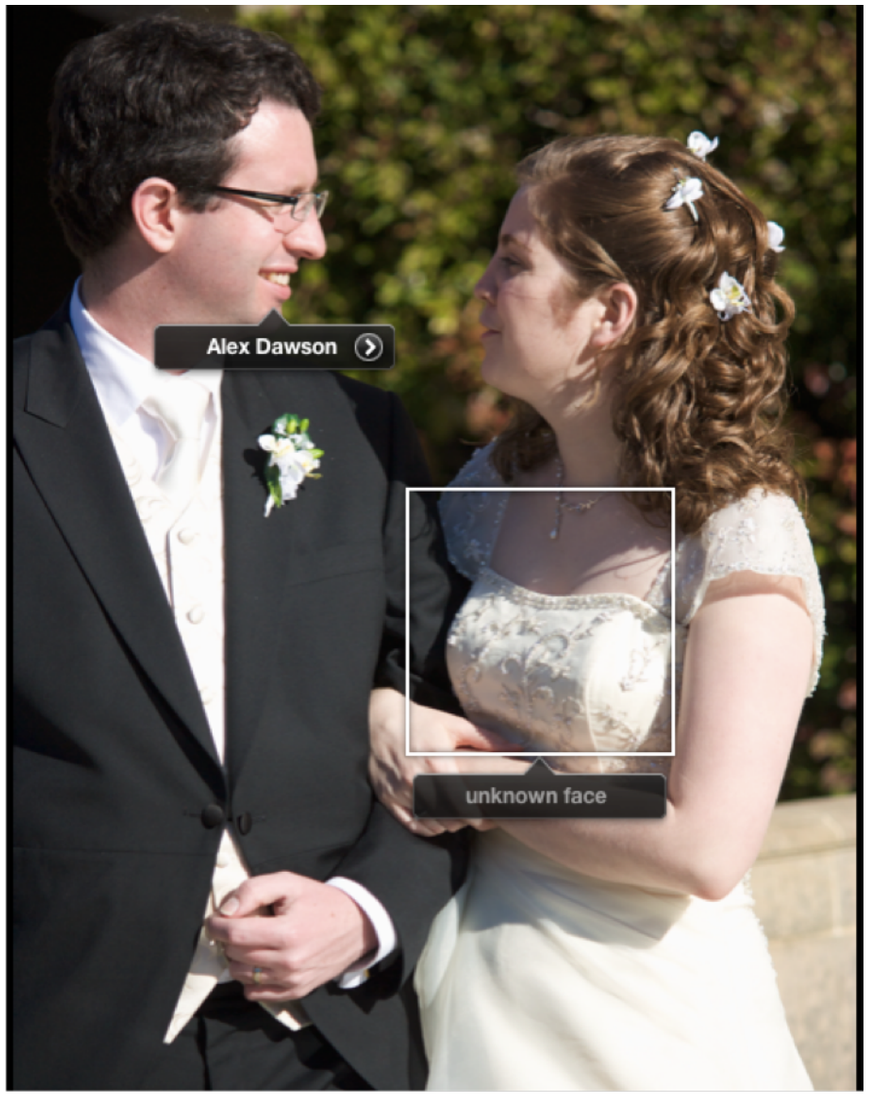


Evolution of fitness

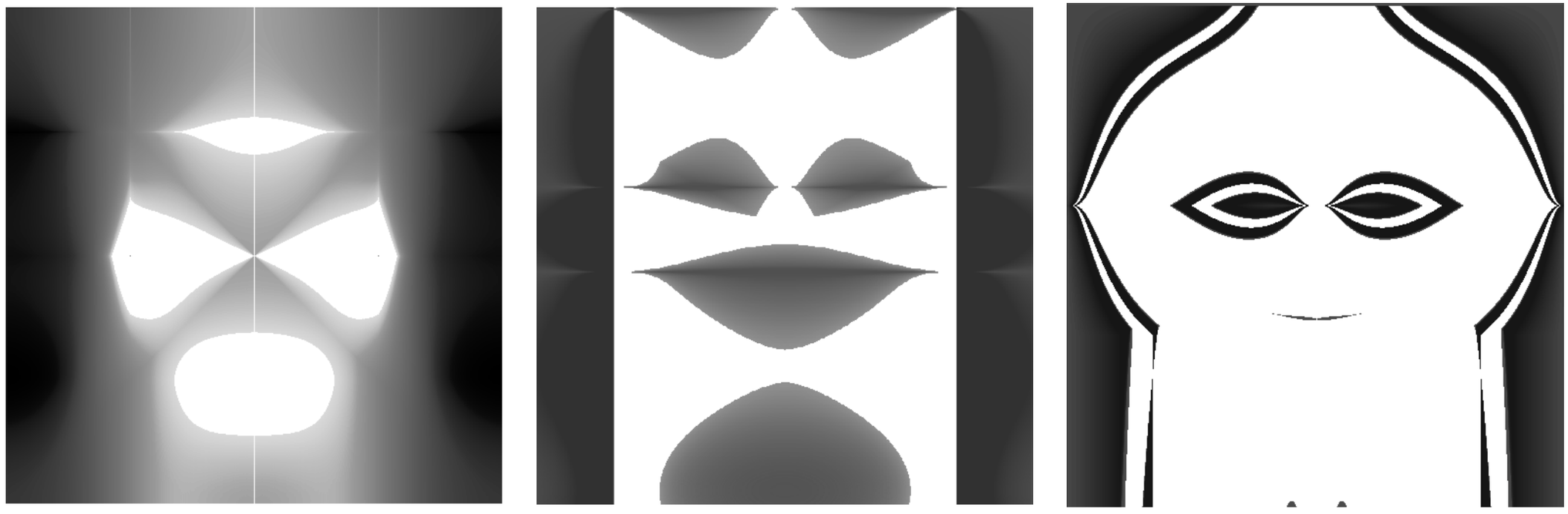







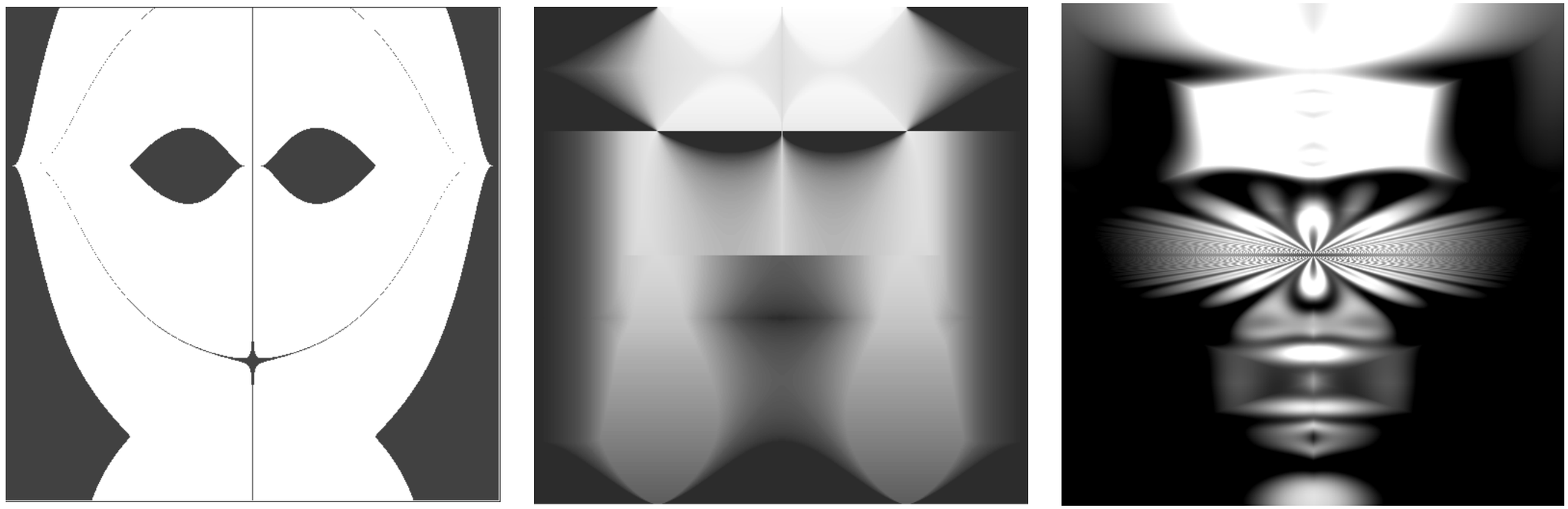
Images classified as faces by the classifier





Images classified as faces by the classifier

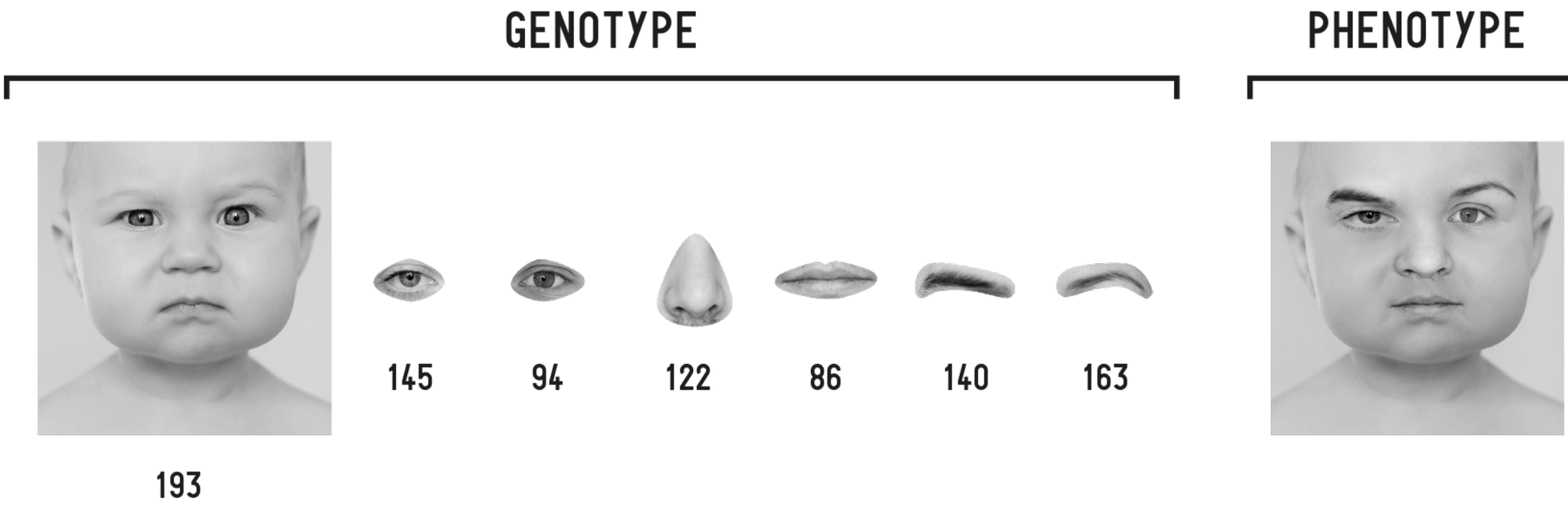




Face (Evocative) Examples

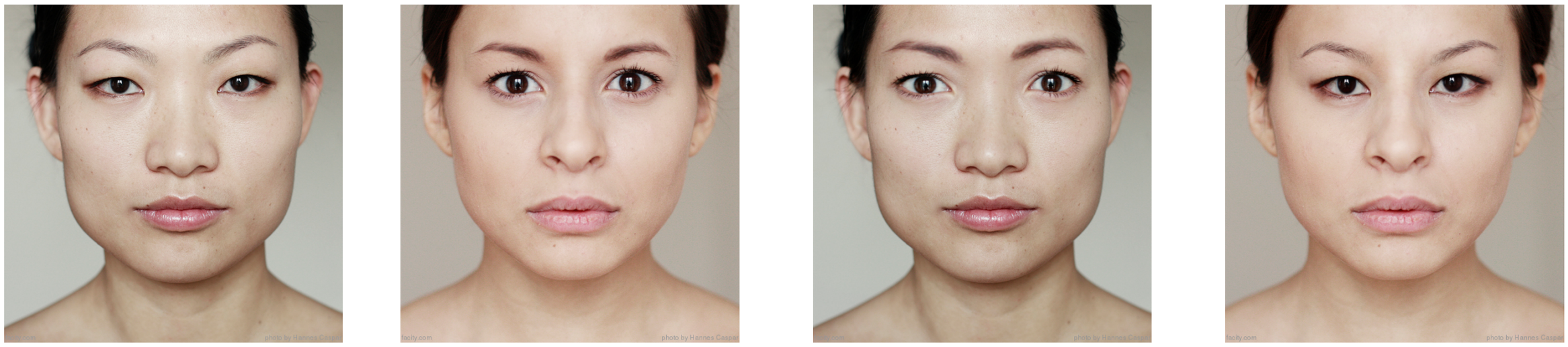
Evolving undetected faces

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



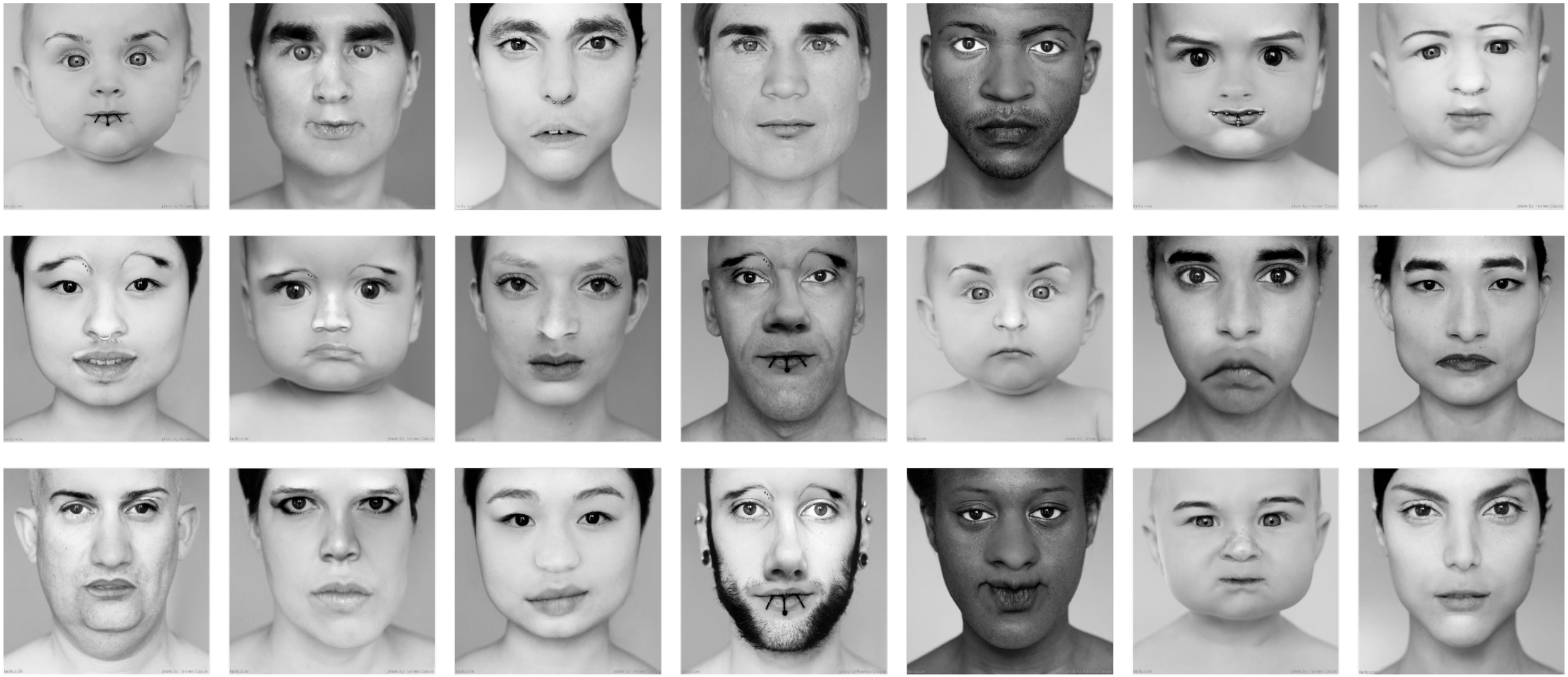
Genotype and phenotype of the evolving faces


facity.com
photo by Katja Kemnitz
facity.com
photo by Katja Kemnitz
facity.com
photo by Katja Kemnitz
facity.com
photo by Katja Kemnitz


facity.com
photo by Hannes Caspar
facity.com
photo by Hannes Caspar
facity.com
photo by Hannes Caspar
facity.com
photo by Hannes Caspar


facity.com
photo by Anaely Delgado
facity.com
photo by Anaely Delgado
facity.com
photo by Anaely Delgado
facity.com
photo by Anaely Delgado

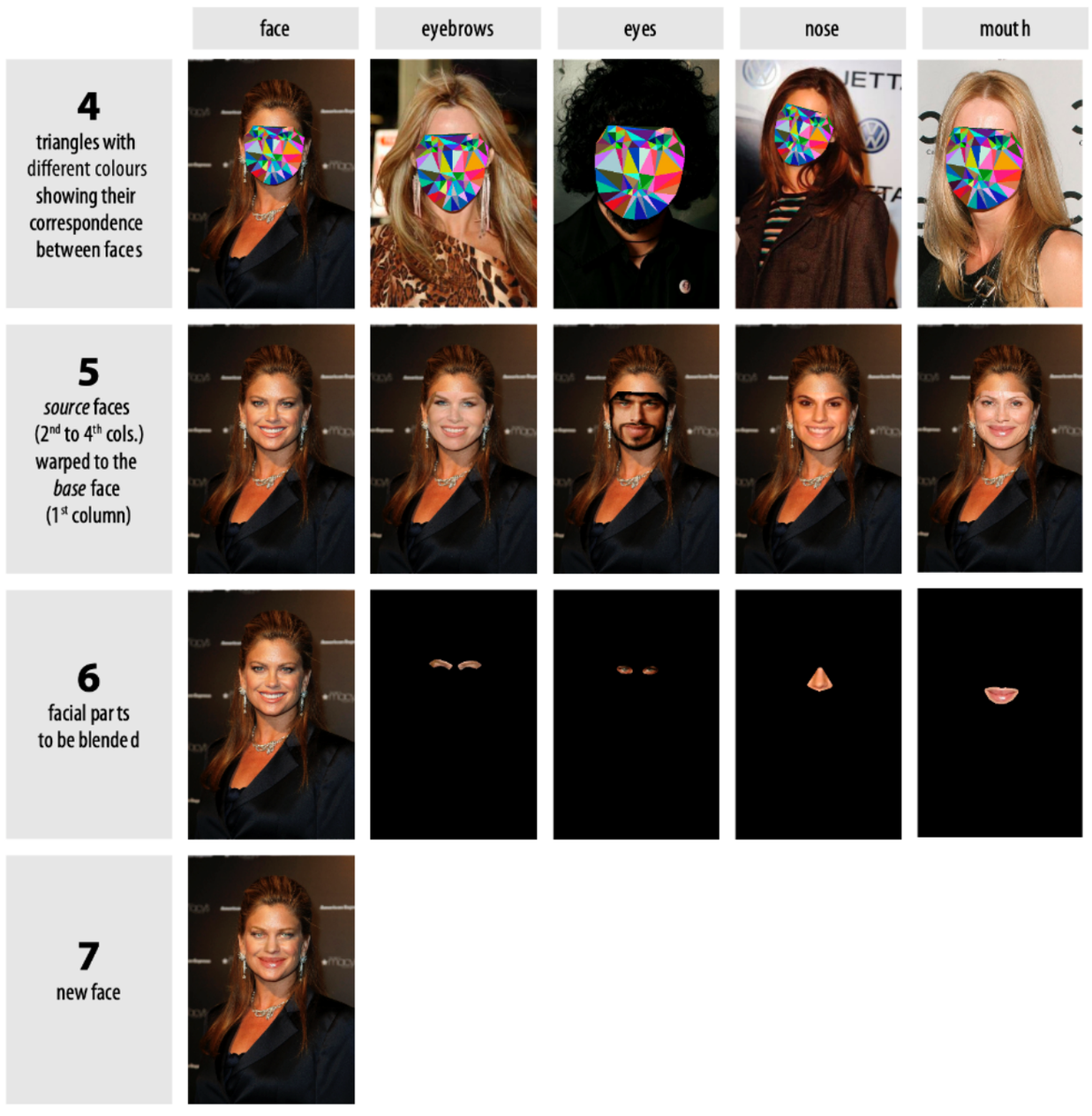
Multiple Swap Generation



Evolved photorealistic faces not detect as faces

Unconstrained Generation

Real or Fake? Overview on Adversarial Examples.



Unconstrained Generation

João Correia

Real or Fake? Overview on Adversarial Examples.

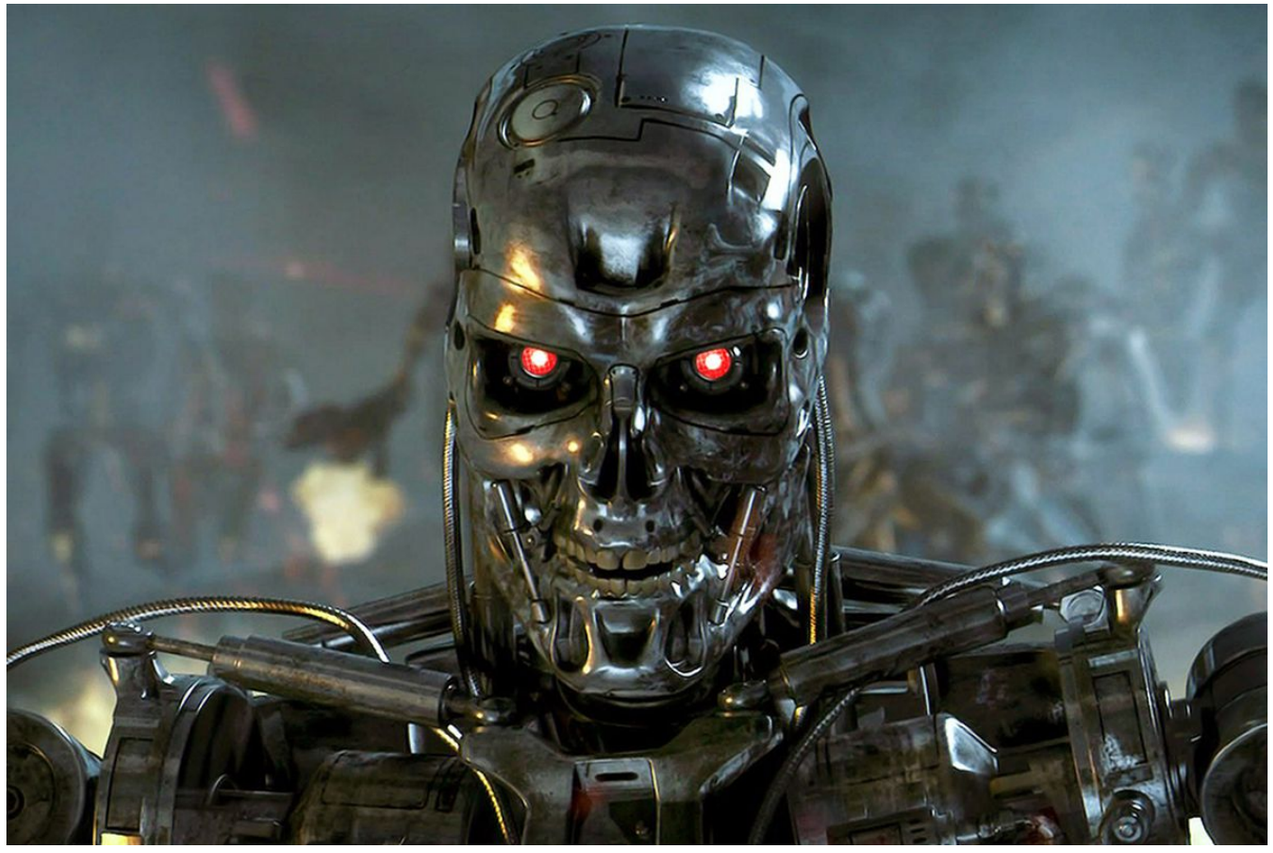
Closing Remarks

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 of Fake?
Overview on Adversarial Examples

João Nuno Correia
jncor@dei.uc.pt

COMPUTATIONAL DESIGN & VISUALIZATION LAB.

1 2 9 0 UNIVERSIDADE D COIMBRA