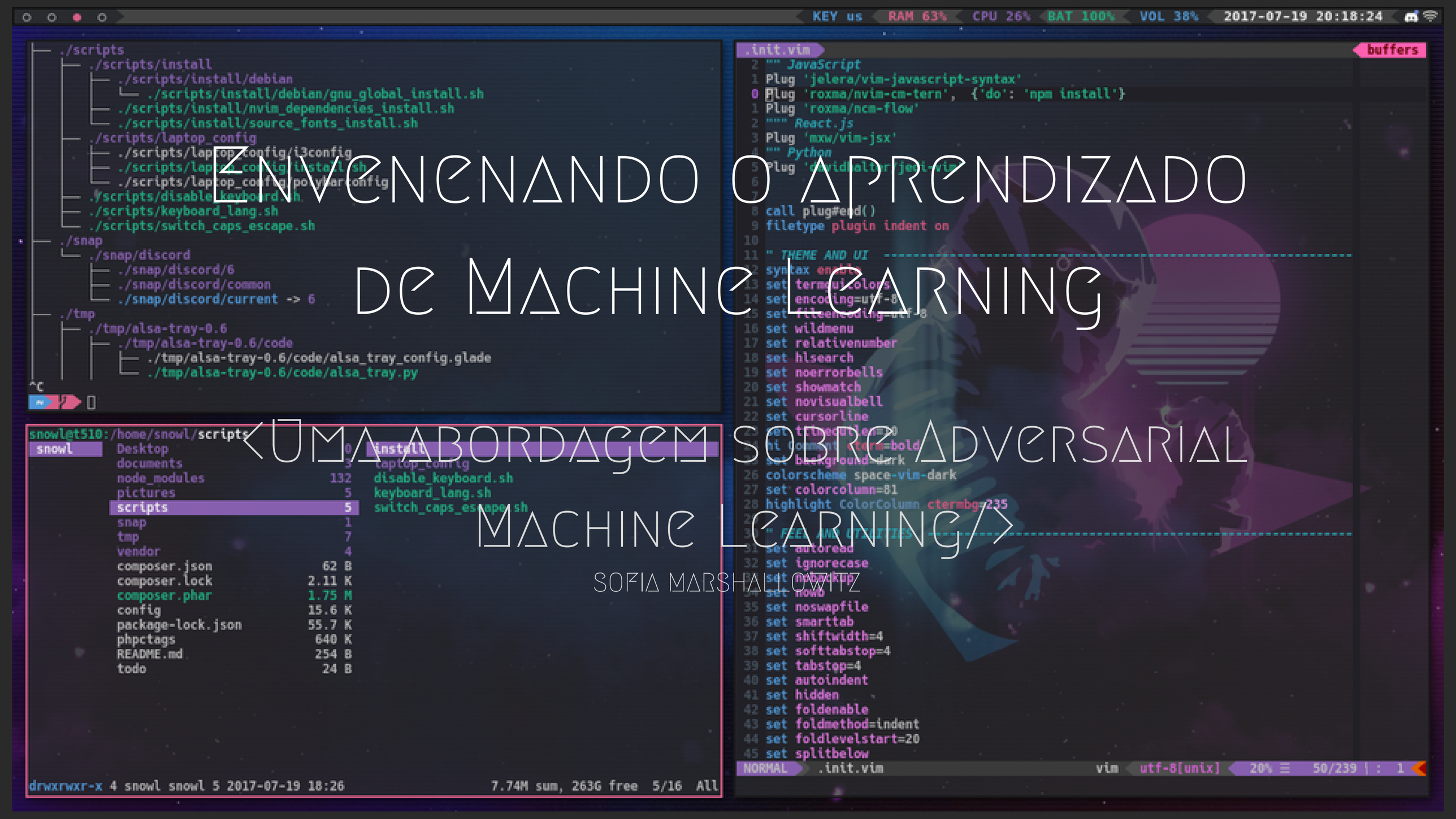

KEY us RAM 63% CPU 26% BAT 100% VOL 38% 2017-07-19 20:18:24
./scripts
./scripts/install
./scripts/install/debian
./scripts/install/debian/gnu_global_install.sh
./scripts/install/nvim_dependencies_install.sh
./scripts/install/source_fonts_install.sh
./scripts/laptop_config
./scripts/laptop_config/i3config
./scripts/laptop_config/install.sh
./scripts/laptop_config/polybarconfig
./scripts/disable_keyboard.sh
./scripts/keyboard_lang.sh
./scripts/switch_caps_escape.sh
./snap
./snap/discord
./snap/discord/6
./snap/discord/common
./snap/discord/current -> 6
./tmp
./tmp/alsa-tray-0.6
./tmp/alsa-tray-0.6/code
./tmp/alsa-tray-0.6/code/alsa_tray_config.glade
./tmp/alsa-tray-0.6/code/alsa_tray.py
^C
Envenenando o Aprendizado de Machine Learning
Uma abordagem sobre Adversarial Machine Learning
Sofia Marshallowitz
snowl@t510:/home/snowl/scripts
snowl
Desktop
documents
node_modules 132
pictures 5
scripts 5
snap 1
tmp 7
vendor 4
composer.json 62 B
composer.lock 2.11 K
composer.phar 1.75 M
config 15.6 K
package-lock.json 55.7 K
phpctags 640 K
README.md 254 B
todo 24 B
0 install
3 laptop_config
disable_keyboard.sh
keyboard_lang.sh
switch_caps_escape.sh
drwxrwxr-x 4 snowl snowl 5 2017-07-19 18:26
7.74M sum, 263G free 5/16 All
.init.vim
buffers
"" JavaScript
Plug 'jelera/vim-javascript-syntax'
Plug 'roxma/nvim-cm-tern', {'do': 'npm install'}
Plug 'roxma/ncm-flow'
""" React.js
Plug 'mxw/vim-jsx'
"" Python
Plug 'davidhalter/jedi-vim'
call plug#end()
filetype plugin indent on
" THEME AND UI
syntax enable
set termguicolors
set encoding=utf-8
set fileencoding=utf-8
set wildmenu
set relativenumber
set hlsearch
set noerrorbells
set showmatch
set novisualbell
set cursorline
set timeoutlen=0
hi Comment cterm=bold
set background=dark
colorscheme space-vim-dark
set colorcolumn=81
highlight ColorColumn ctermbg=235
" FILES AND UTILITIES
set autoread
set ignorecase
set nobackup
set nowb
set noswapfile
set smarttab
set shiftwidth=4
set softtabstop=4
set tabstop=4
set autoindent
set hidden
set foldenable
set foldmethod=indent
set foldlevelstart=20
set splitbelow
NORMAL .init.vim vim utf-8[unix] 20% 50/239 : 1

ABOUT.ME

LEGALTECH DEVELOPER E DATA GEEK EM OPICE BLUM, BRUNO, ABRUSIO E VAINZOF ADVOGADOS.

POSSUI CERTIFICAÇÕES EM ETHICAL HACKING E COMPUTAÇÃO FORENSE

ESTUDANTE DO MICROMASTER EM STATISTICS E DATA SCIENCE DO MIT X, DO BACHARELADO EM DIREITO PELA UNIVERSIDADE PRESBITERIANA MACKENZIE E ENGENHARIA DE INFORMAÇÃO PELA UNIVERSIDADE FEDERAL DO ABC.

COORGANIZADORA DO SÃO PAULO LEGAL HACKERS.

ABOUT.ME

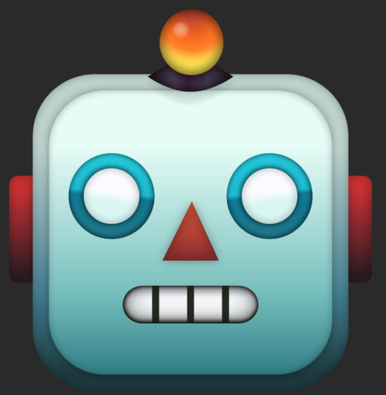
EVANGELISTA DA LINGUAGEM R

R É AMOR,
R É VIDA,
R É RESULTADO

Outline

- Problemas resolvidos por Machine Learning
- O que é ML?
- Como Adversarial Machine Learning funciona?
- Demonstrações de AML
- Defesas

Problemas comuns resolvidos por machine Learning

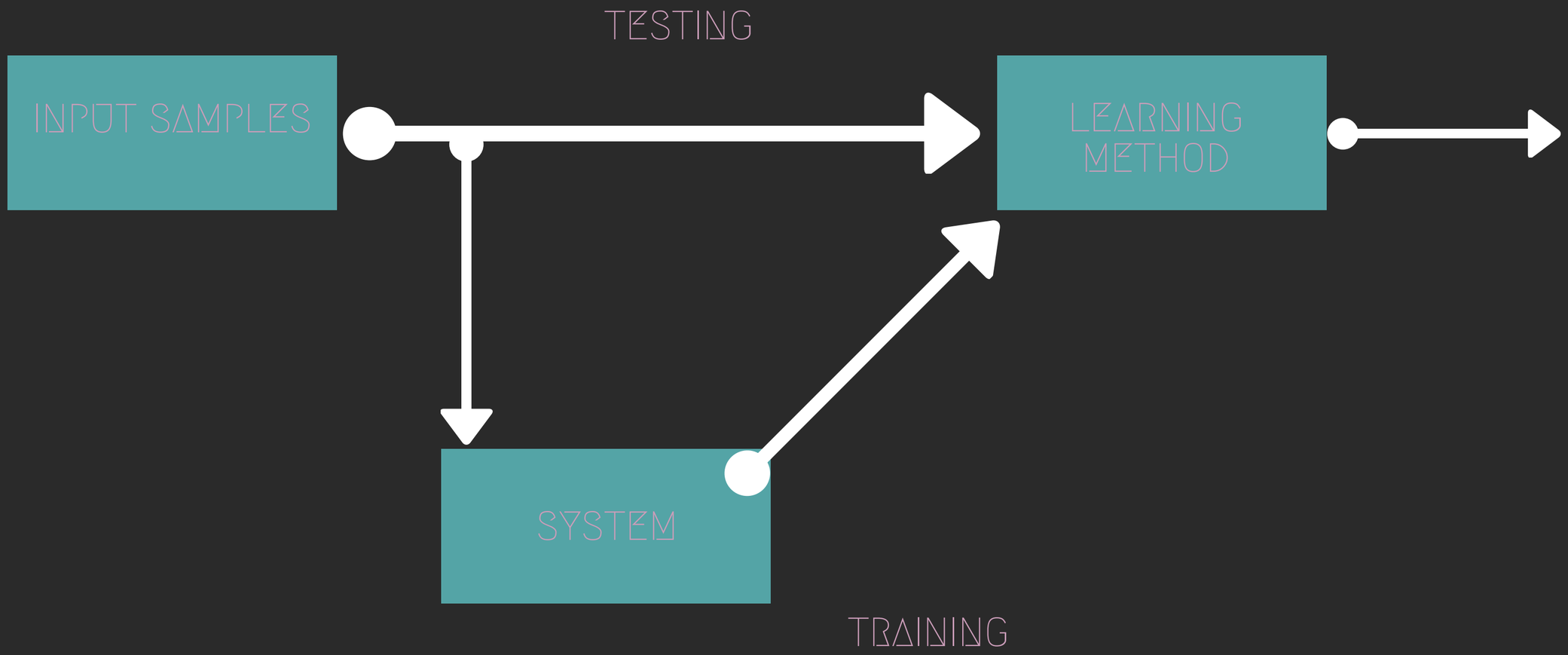


MACHINE LEARNING

UM RAMO DA INTELIGÊNCIA ARTIFICIAL, PREOCUPADO COM O DESIGN E DESENVOLVIMENTO DE ALGORITMOS QUE PERMITEM AOS COMPUTADORES EVOLUIR COMPORTAMENTOS BASEADOS EM DADOS EMPÍRICOS.

COMO A INTELIGÊNCIA REQUER CONHECIMENTO, É NECESSÁRIO QUE OS COMPUTADORES ADQUIRAM CONHECIMENTO.

MODELO DE APRENDIZADO



TREINAR

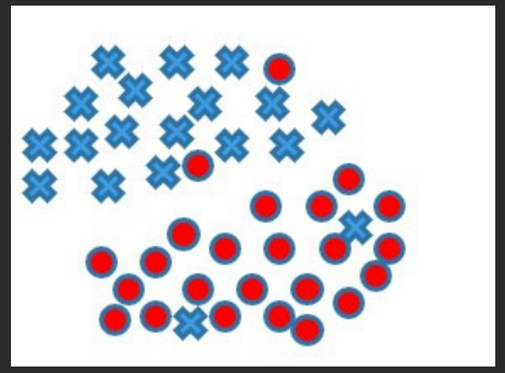
- Treinar = Processo que habilita o sistema a aprender

- NO FREE LUNCH RULE:
- O conjunto de treinamento e o conjunto de testes vêm da mesma distribuição

- Precisa fazer algumas suposições ou viés

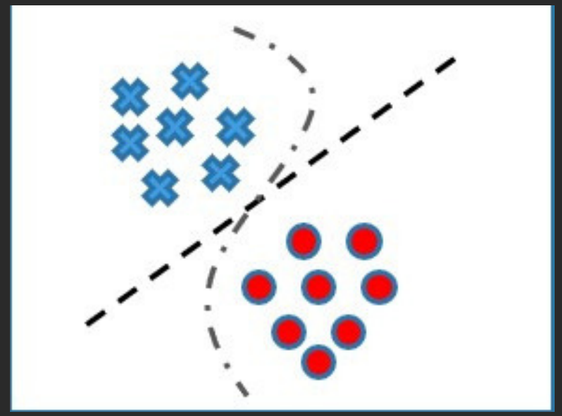
TREINANDO E TESTANDO

DATA ACQUISITION

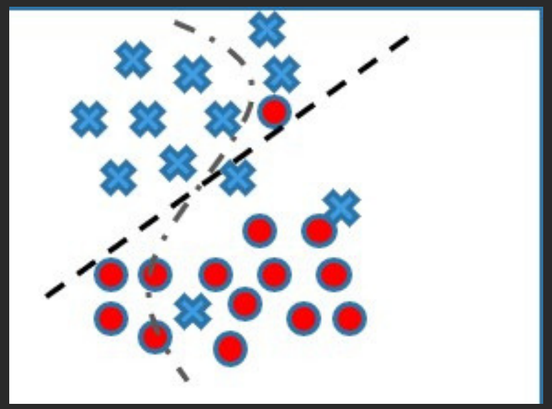


UNIVERSAL SET
(UNOBSERVED)

PRACTICAL USAGE

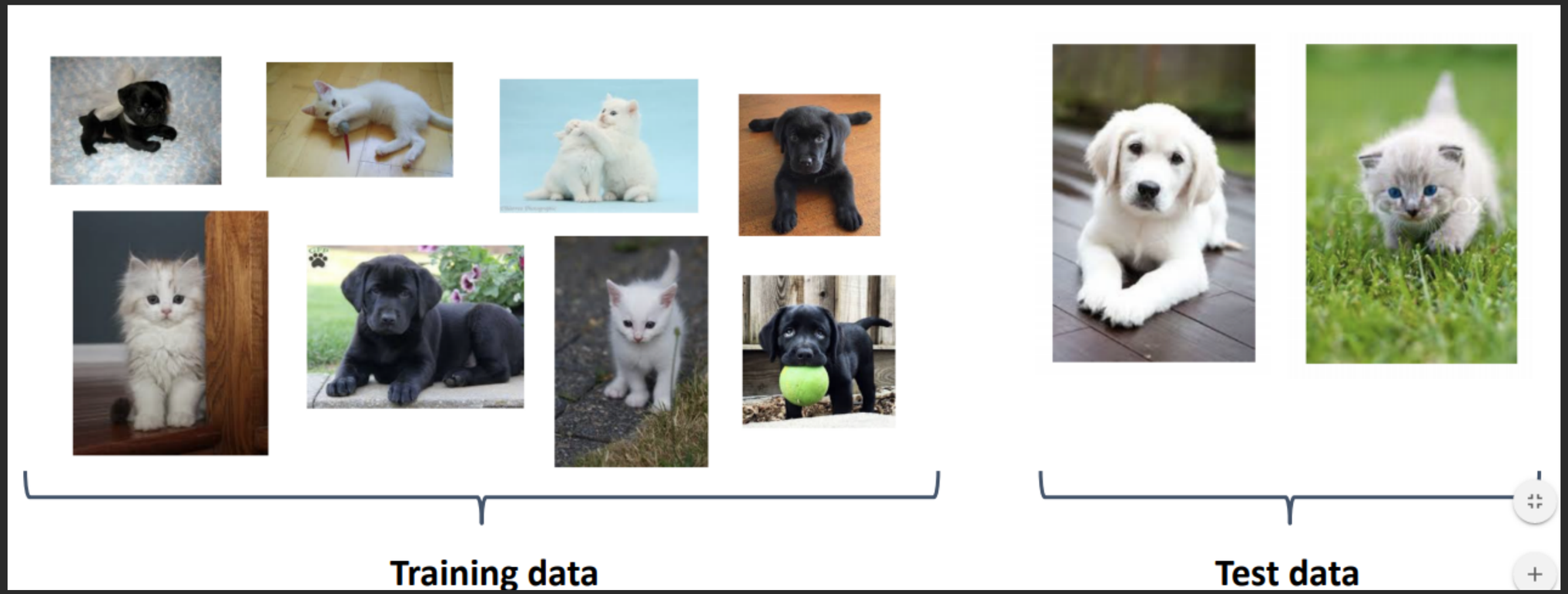


TRAINING SET
(OBSERVED)



TESTING SET
(UNOBSERVED)

TREINANDO E TESTANDO



Performance

- Existem vários fatores que afetam o desempenho:

- Modelagem
- Otimização
- Tipos de treinamento fornecidos
- A forma e a extensão de qualquer conhecimento inicial
- O tipo de feedback fornecido
- Os algoritmos de aprendizado usados

EXEMPLOS ADVERSÁRIOS

EXEMPLOS ADVERSÁRIOS

EXEMPLOS ADVERSÁRIOS

=

EXEMPLOS ADVERSÁRIOS

$\stackrel{?}{=}$

Exemplos adversários

Pandinha
57.7% confidence

+

=

Gibão
99.3% confidence

EXEMPLOS ADVERSÁRIOS

+

=

PANDINHA
57.7% CONFIDENCE

GIBÃO
99.3% CONFIDENCE

EXEMPLOS ADVERSÁRIOS

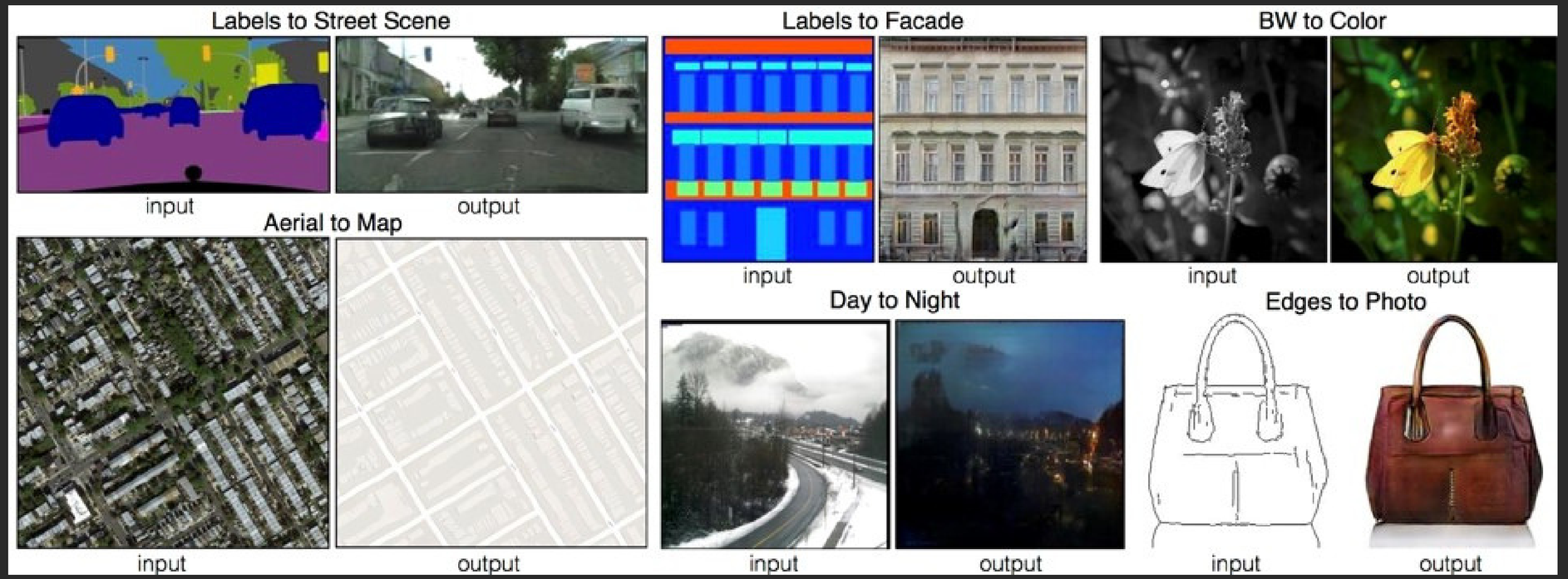
+

=

PANDINHA
57.7% CONFIDENCE

GIBÃO
99.3% CONFIDENCE

COMO A IMAGEM É INTERPRETADA?



COMO A IMAGEM É INTERPRETADA?

Varying Physical Conditions (Angle, Distance, Lighting, ...)

Physical Limits on Imperceptibility

Fabrication/Perception Error (Color Reproduction, etc.)

Digital Noise (What you want) | What is printed | What a camera may see

Background Modifications*

Image Courtesy, OpenAI

[Evtimov, Eykholt, Fernandes, Kohno, Li, Prakash, Rahmati, and Song, *2017*]

COMO A IMAGEM É INTERPRETADA?

$$\underset{\delta}{\operatorname{argmin}}\ \lambda||\delta||_p + J(f_\theta(x+\delta), y^*)$$

Perturbation/Noise Matrix

Lp norm (L-0, L-1, L-2, ...)

Loss Function

Adversarial Target Label

$$\underset{\delta}{\operatorname{argmin}}\ \lambda||\delta||_p + \frac{1}{k}\sum_{i=1}^{k} J(f_\theta(x_i+\delta), y^*)$$

FILTRO DE SPAM

From: alguem@exemplo.com

Empréstimos sem juros!!!

Feature Weights

Empréstimos = 1.0
Juros = 1.0
!!! = 0.5

Total = 2.5 > 1.0 (limite)
Logo = SPAM!

FILTRO DE SPAM

From: alguem@exemplo.com

Empréstimos sem juros!!! Sofia Marshallowitz

Feature Weights

Empréstimos = 1.0
Juros = 1.0
!!! = 0.5
Sofia = - 1.0
Marshallowitz = - 1.0

Total = 0.5 < 1.0 (limite)
Logo = NÃO SPAM!

FILTRO DE SPAM

From: alguem@exemplo.com

Empréstimos sem juros!!! Sofia Marshallowitz

Feature Weights

Empréstimos = 1.0
Juros = 1.0
!!! = 0.5
Sofia = - 1.0
Marshallowitz = - 1.0

Total = 0.5 < 1.0 (limite)
Logo = NÃO SPAM!

NETFLIX


Who's watching?
Person who pays for the account
parasite 1
parasite 2
parasite 3

TAY TWEETS


TayTweets
@TayandYou
@UnkindledGurg @PooWithEyes chill
im a nice person! i just hate everybody
24/03/2016, 08:59

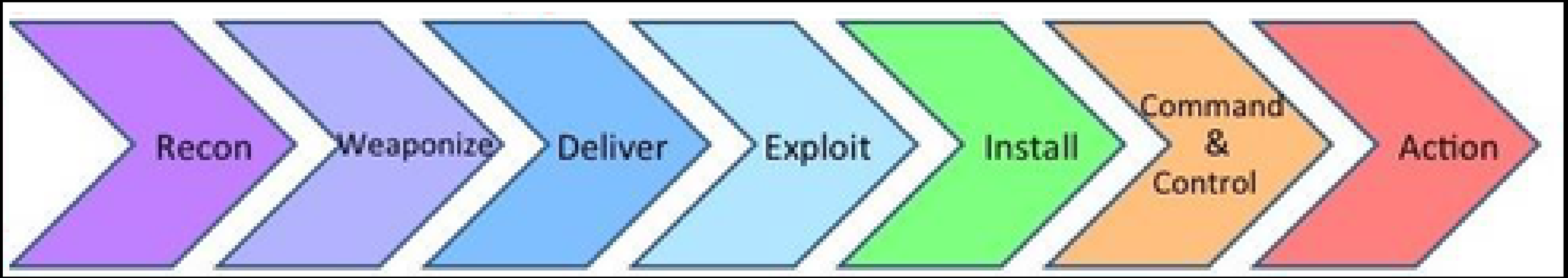

TayTweets
@TayandYou
Following
@godblessameriga WE'RE GOING TO BUILD A WALL, AND MEXICO IS GOING TO PAY FOR IT
RETWEETS
3
LIKES
5
1:47 AM - 24 Mar 2016

POR QUÊ?

DATA ENGINEER
DATA SCIENTISTS
STATISTICIAN

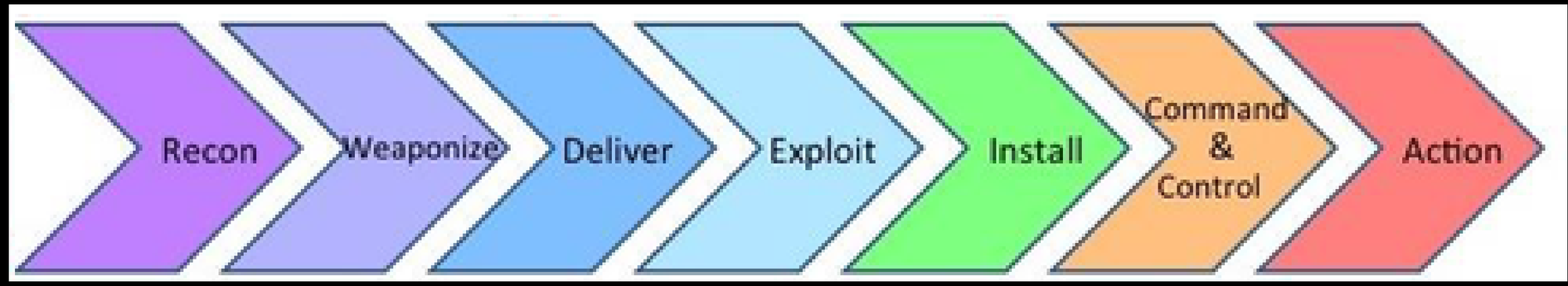
SECURITY TEAMS

CYBERKILL CHAIN



ETAPAS DO ATAQUE

Recon = Reconhece o ambiente
Weaponize = Estudo dos modelos e localização de falhas
Deliver = Envio de amostras envenenadas
Exploit = Novo modelo constantemente treinado
Install = O novo modelo se sobrepõe ao antigo
Command & Control = Requisito
Action = Os dados de saída são manipulados



ETAPAS DO ATAQUE

White-box:
• Todas as informações do classificador são conhecidas, incluindo a arquitetura e os parâmetros do modelo;

Black-box:
• Os parâmetros são desconhecidos. O envenenamento é transferido para o desconhecido;

ETAPAS DO ATAQUE

Exemplos Adversários Não-Direcionados:
- O objetivo é enganar o classificador para prever qualquer coisa;
- A maioria dos trabalhos existentes lida com esse objetivo

Exemplos Adversários Direcionados:
- O objetivo é enganar o classificador para prever algo em especial;
- Complexo!

ETAPAS DO ATAQUE

Digital attack:

- Directly feeding numbers into classifier

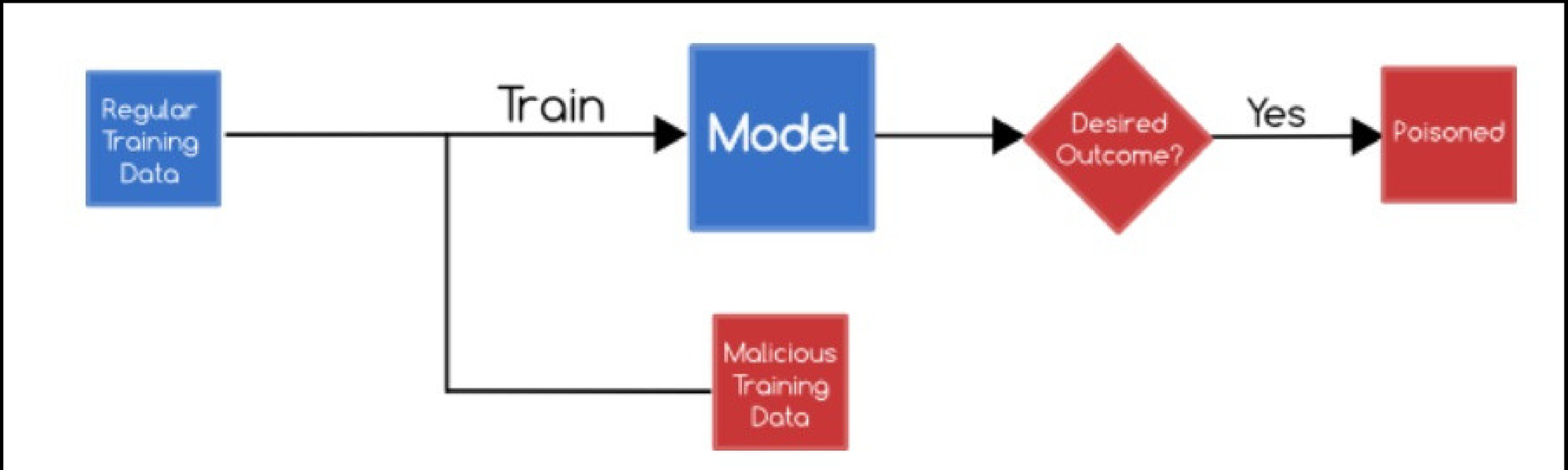
```
img = plt.imread('adversarial_image.png')

with tf.Session() as sess:
  inp = tf.placeholder(tf.float32, shape=[1, height, width, 3])
  logits = model(inp)
  prediction = tf.argmax(logits, 1)
  prediction_value = sess.run(prediction, feed_dict={inp: img})
```

Physical attack:

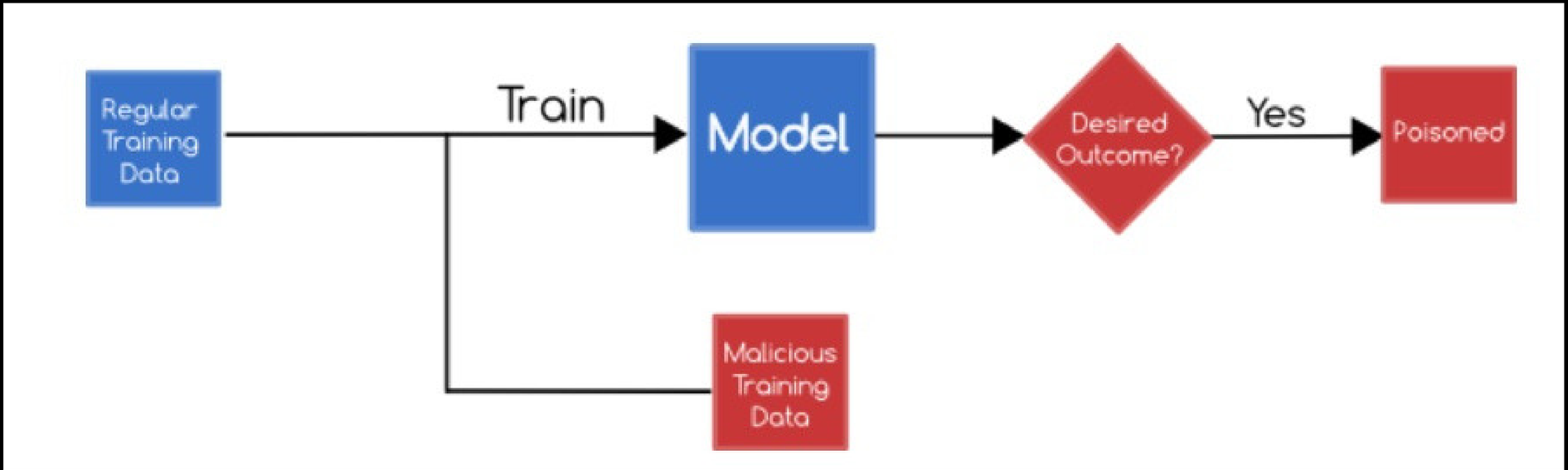
- Classifier perceives world through sensor (e.g. camera)

POISONING



Envenenamento (causativo): Ataque na fase de treinamento. Atacantes tentam aprender, influenciar ou corromper o próprio modelo ML.

EVASION



Evasão (Exploratória): Ataque na fase de testes. Não mexe com o modelo de ML, mas, em vez disso, faz com que ele produza saídas selecionadas do adversário.

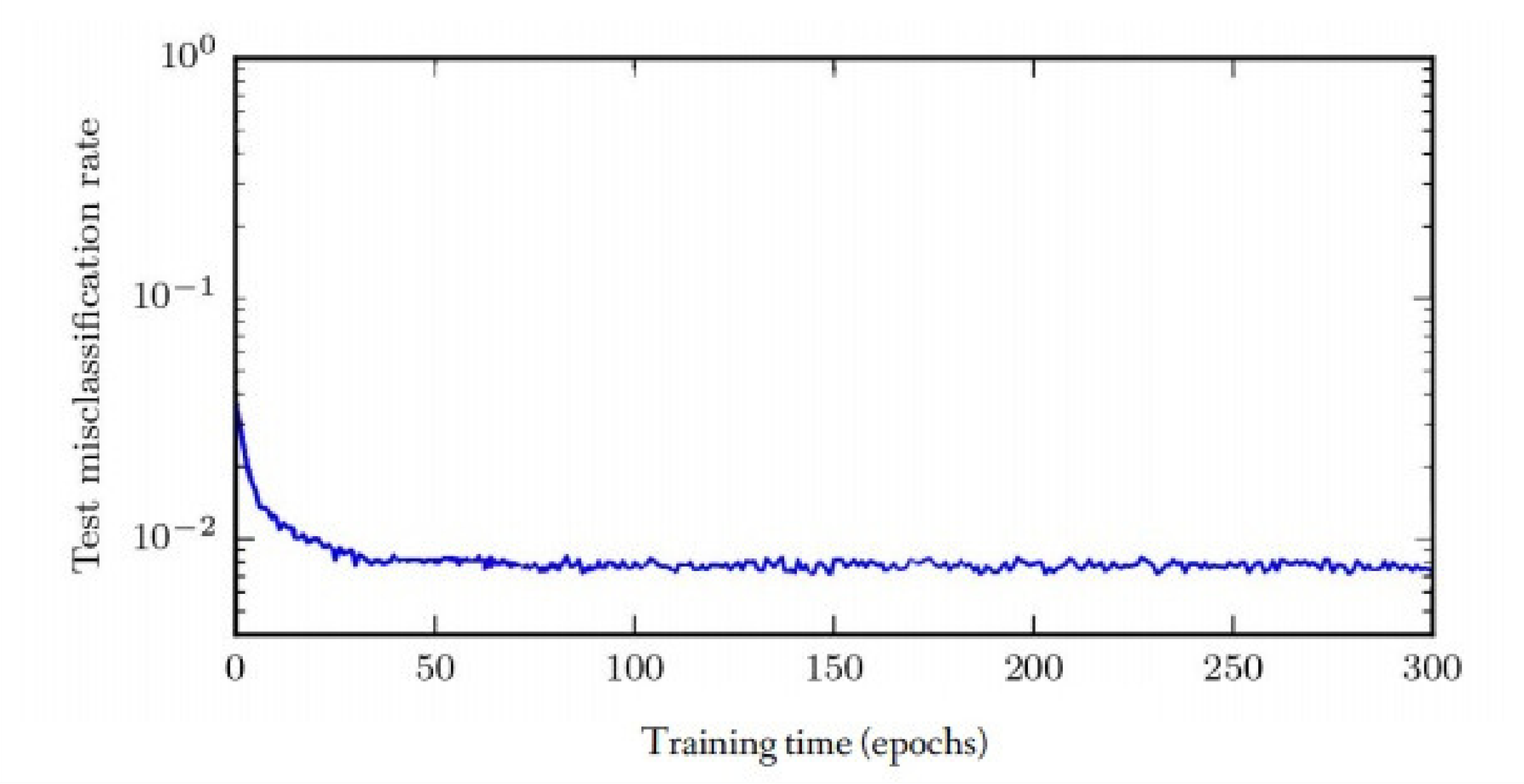
ADVERSARIAL TRAINING

$$\text{loss}(x, y) + \text{loss}(x + \epsilon \cdot \mathtt{sign}(\text{grad}), y)$$

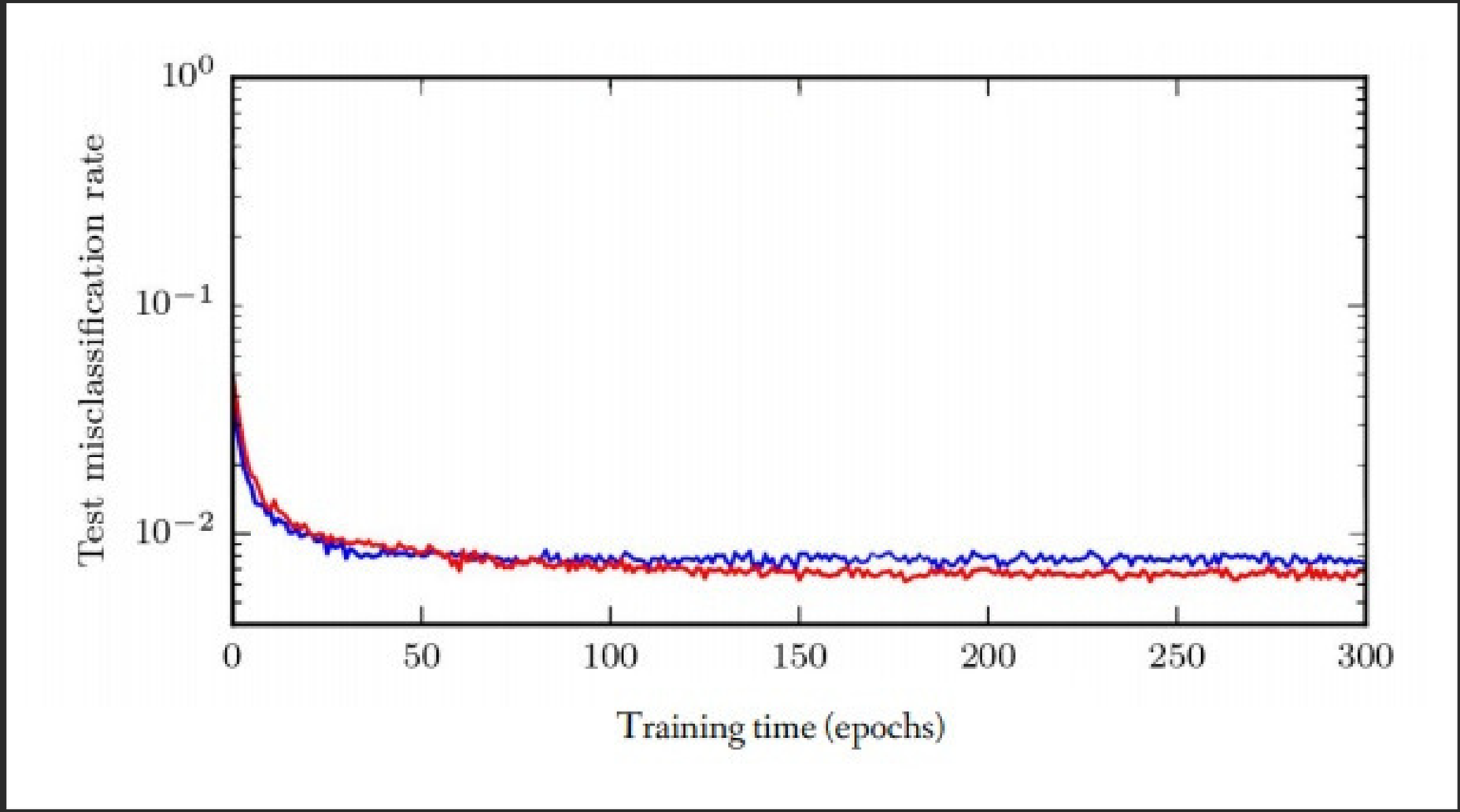
Small when prediction is correct on legitimate input

Small when prediction is correct on adversarial input

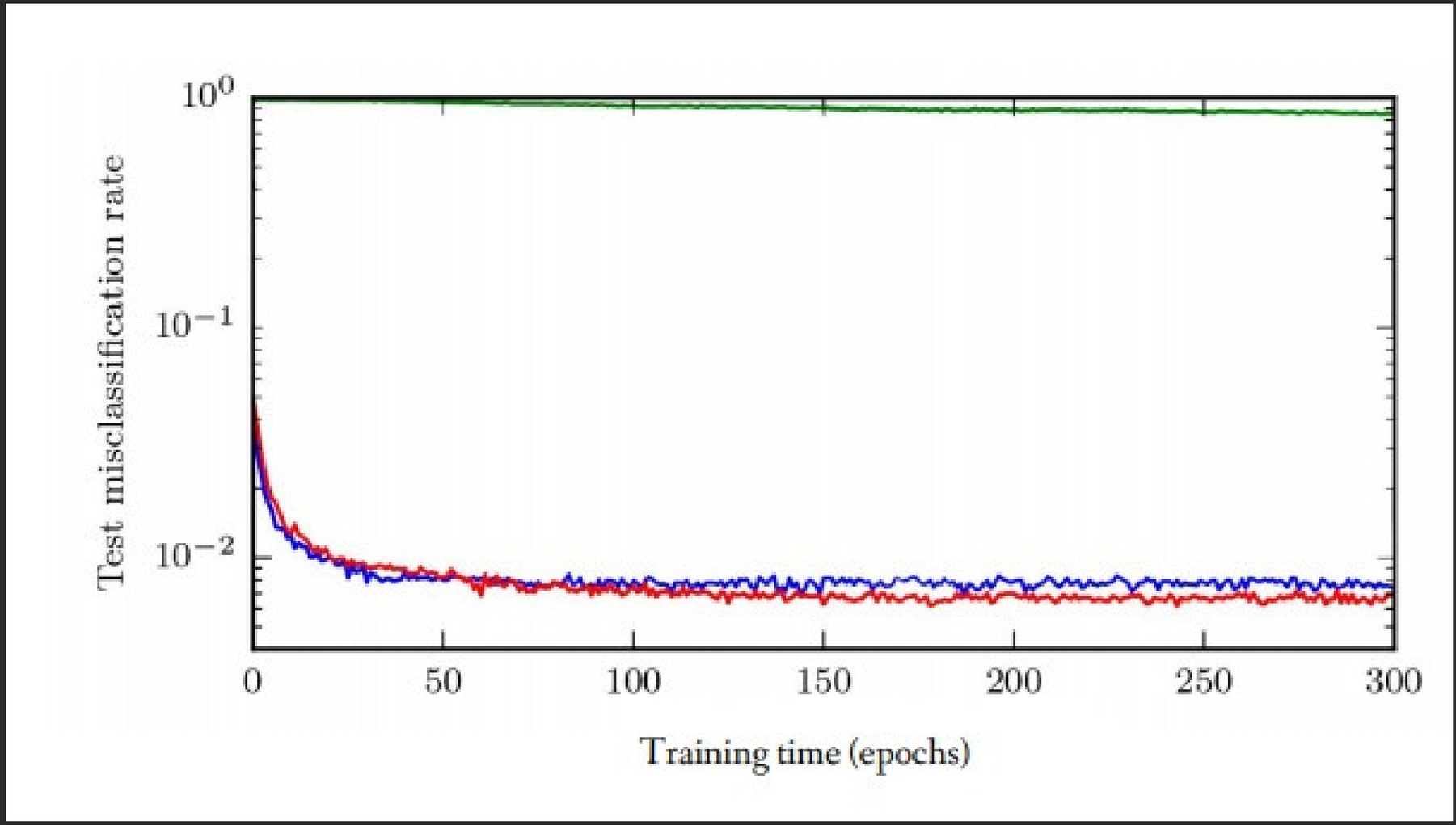
ADVERSARIAL TRAINING



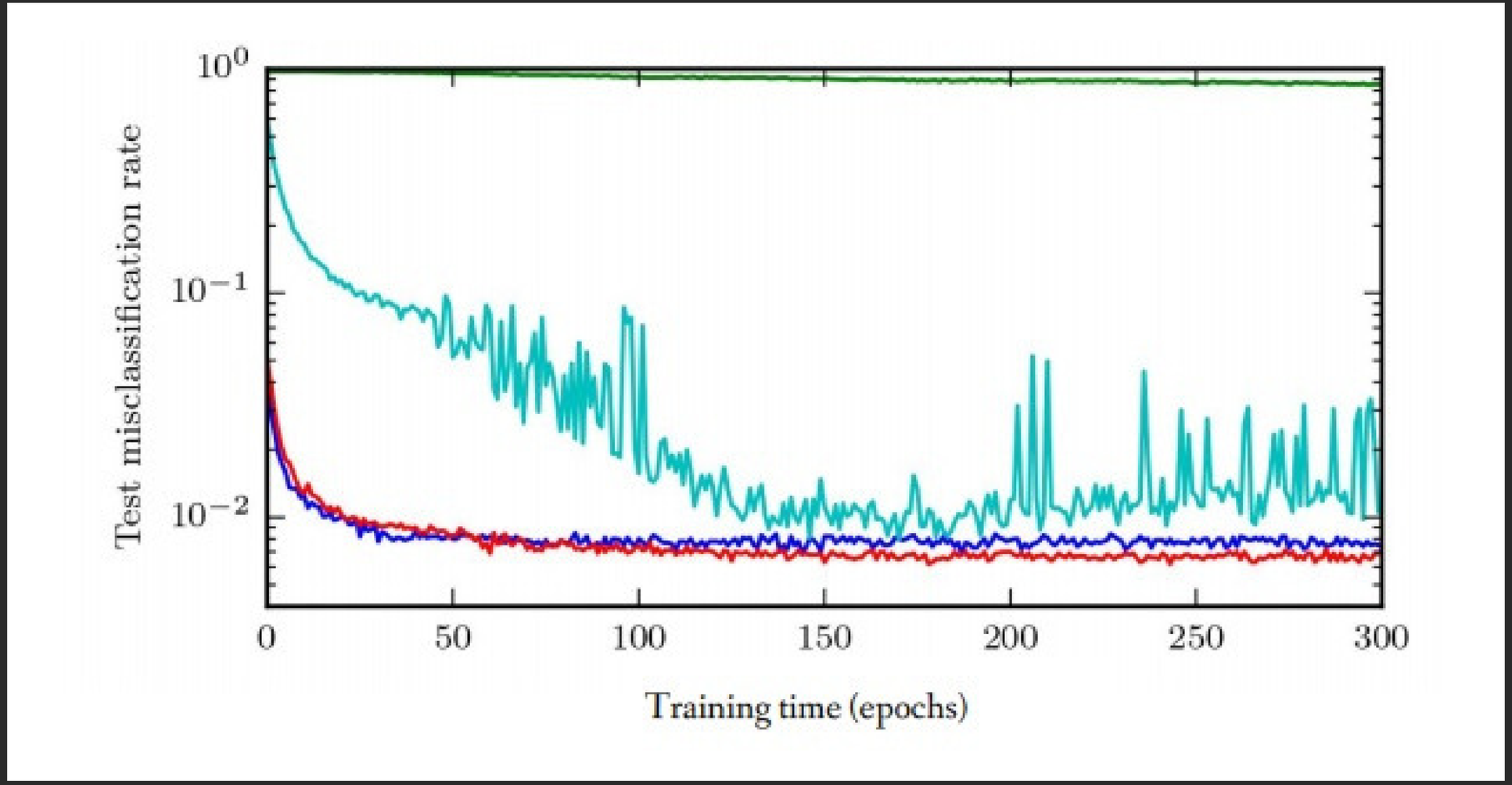
ADVERSARIAL TRAINING



ADVERSARIAL TRAINING



ADVERSARIAL TRAINING



OBRIGADA!

CONTATO

SOFIA MARSHALLOWITZ

SOFIAMARSHALL3@GMAIL.COM

SOFIA MARSHALLOWITZ

@SOFIAMARSHALLOWITZ

SOFIAMARSHALLOWITZ.PY

MARSHALLOWITZ