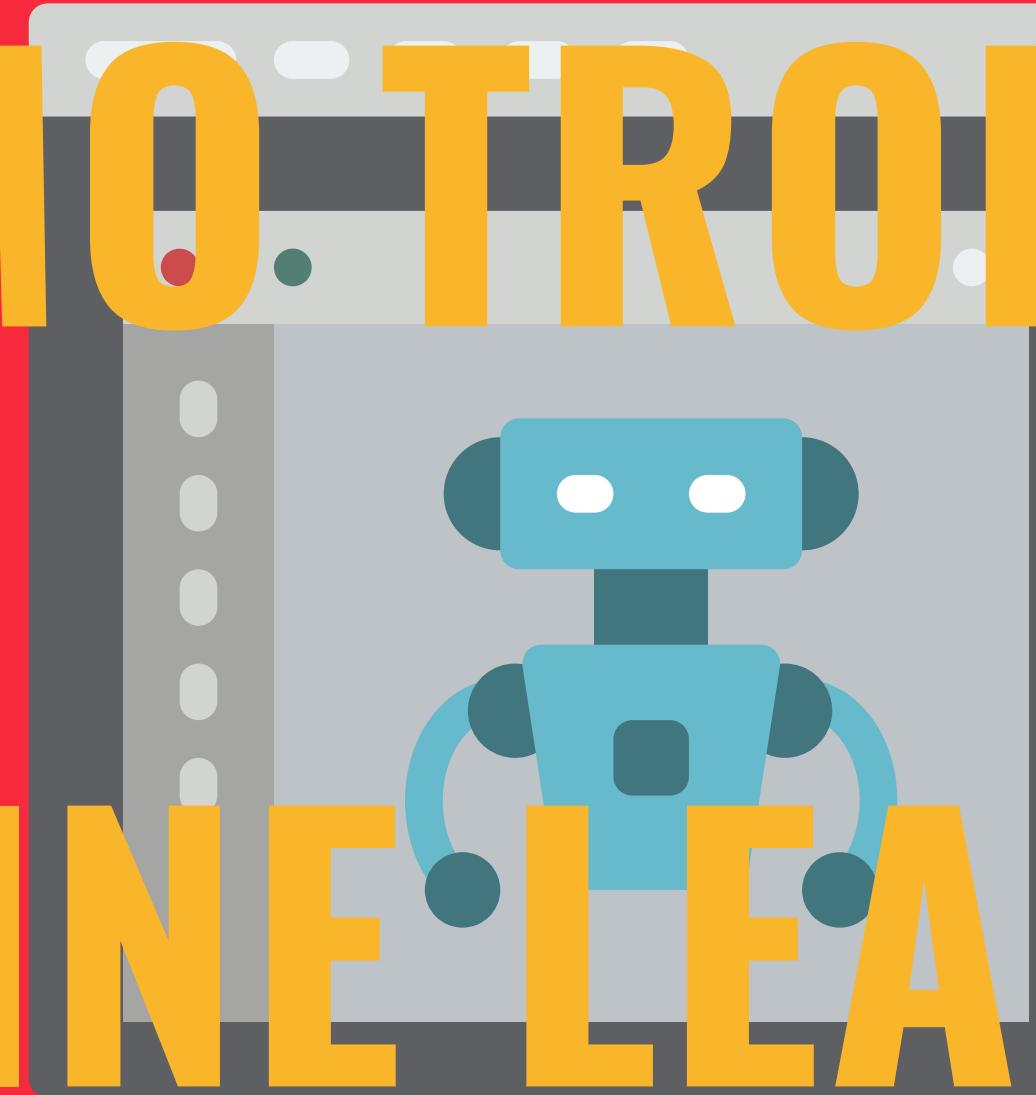
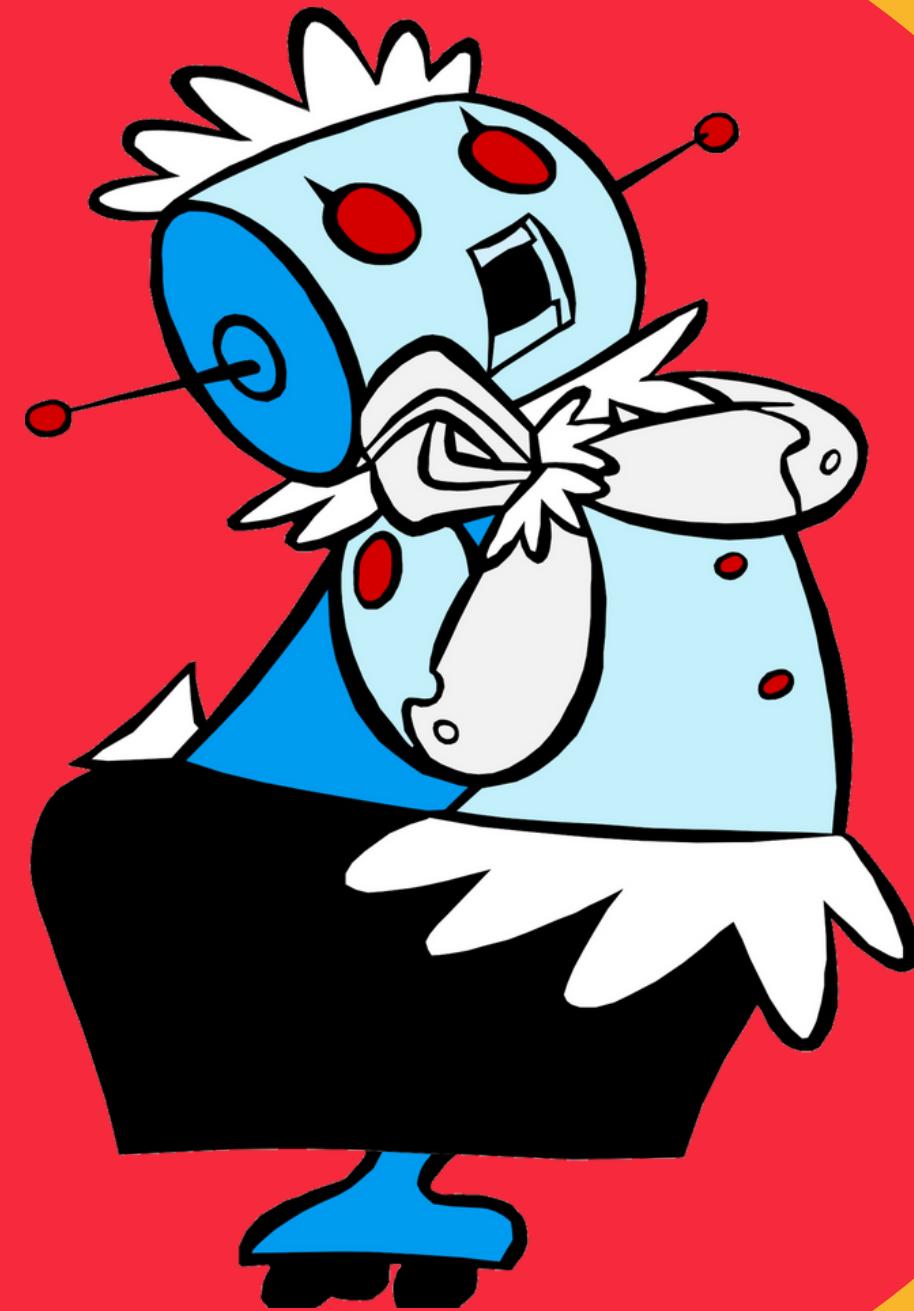


UMA ABORDAGEM DE SEGURANÇA DE ML

COMO CONTROLAR MACHINE LEARNING?



Sofia Marshallowitz - CPBR - 14.02.2019



OUTLINE

QUAIS SÃO OS PROBLEMAS RESOLVIDOS POR ML?

O QUE É MACHINE LEARNING?

COMO ADVERSARIAL MACHINE LEARNING FUNCIONA?

DEFESAS

UM POUQUINHO SOBRE GENERATIVE ADVERSARIAL NETWORKS

WELCOME TO INTERNET I WILL BE YOUR GUIDE



Sistema Operacional com Fala e Inteligência Artificial
Pesquisadora em Inteligência Artificial e Direito em Lawgorithm/USP
Cientista de Dados na Opice Blum, Bruno, Abrusio e Vainzof Advogados
Tem algumas certificações de Ethical Hacking e Computação Forense

Estudante de Statistics e Data Science (MITx)

Estudante de Engenharia de Informação (UFABC)

Estudante de Direito (Mackenzie)

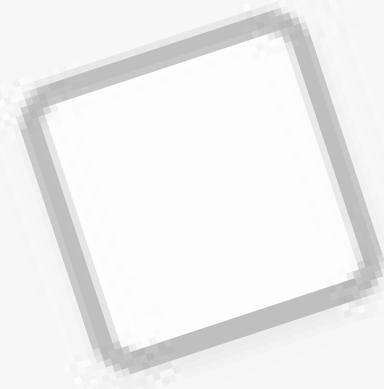
Apaixonada por R

Mais apaixonada ainda pelas tuas senhas que vazaram

COMO TROLLAR MACHINE LEARNING?

WELCOME TO INTERNET I WILL BE YOUR GUIDE

I'm not a robot



Mais apaixonada ainda pelas tuas senhas que vazaram

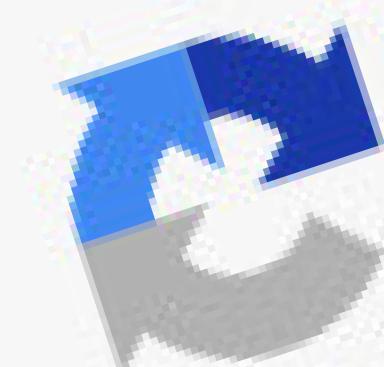
... por R

COMO TROLLAR MACHINE LEARNING?

Sistema Operacional com Fala e Inteligênci>

Pesquisadora em Inteligênci>

Ci>



reCAPTCHA

Privacy - Terms

JSP

dos

O que você já experimentou com ML?

QUAIS SÃO OS PROBLEMAS RESOLVIDOS POR ML?

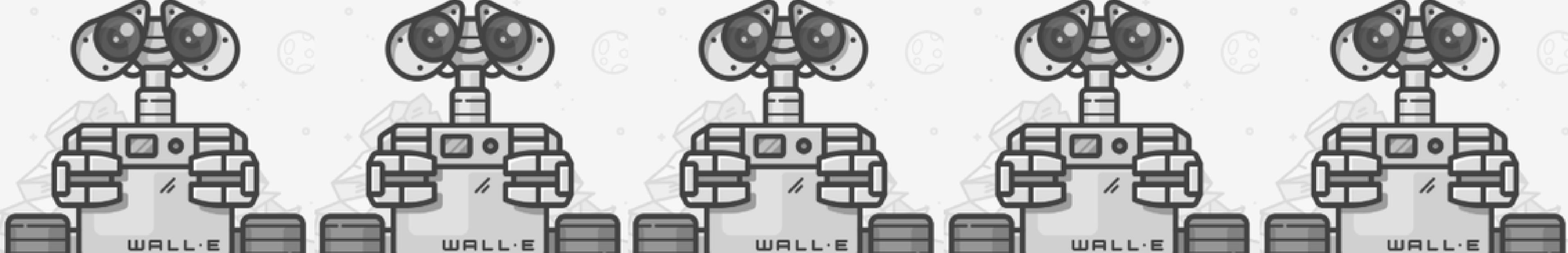
Isso está absolutamente em qualquer lugar!

Out[11]=

```
{1730, 1485, 2190, 1663, 2400, 1119}
```

In[12]:= newlist = Standardize[newlist]

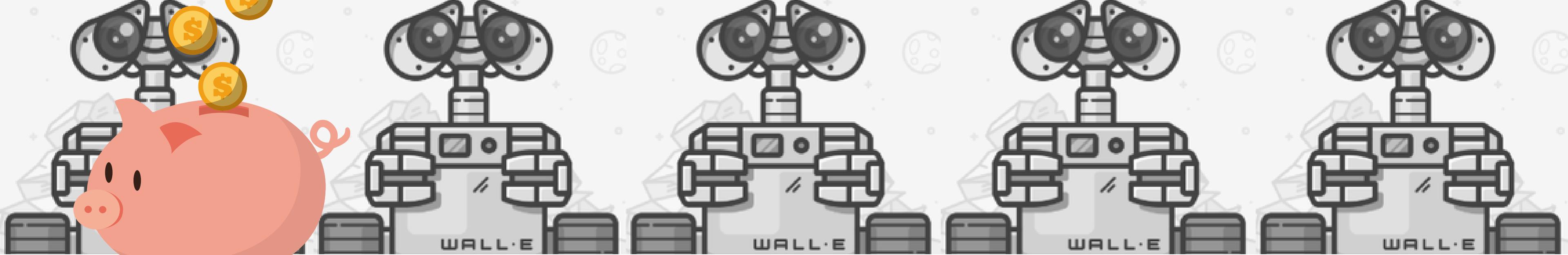
Out[12]=



ALGUMAS APLICAÇÕES COMUNS...

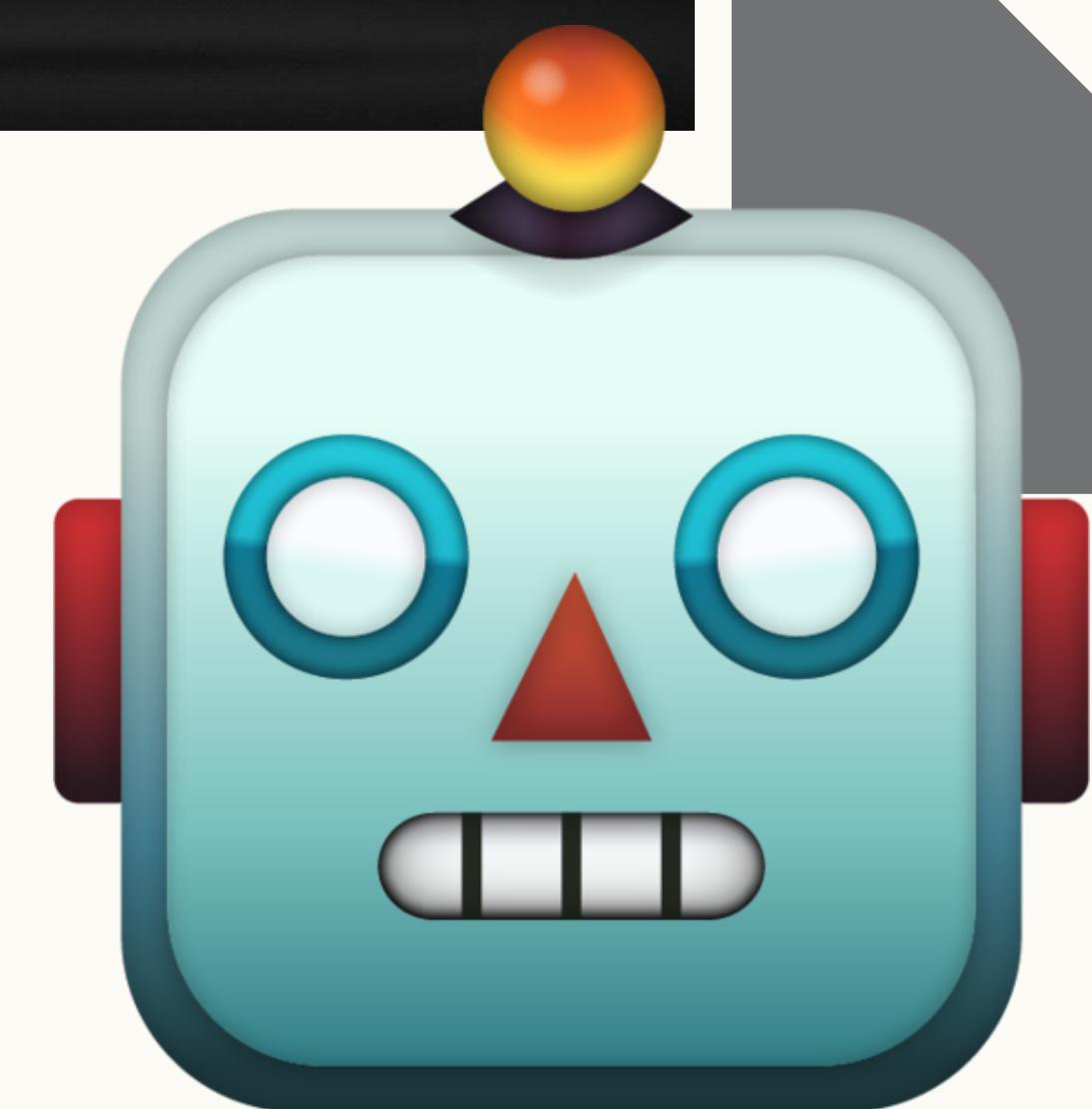
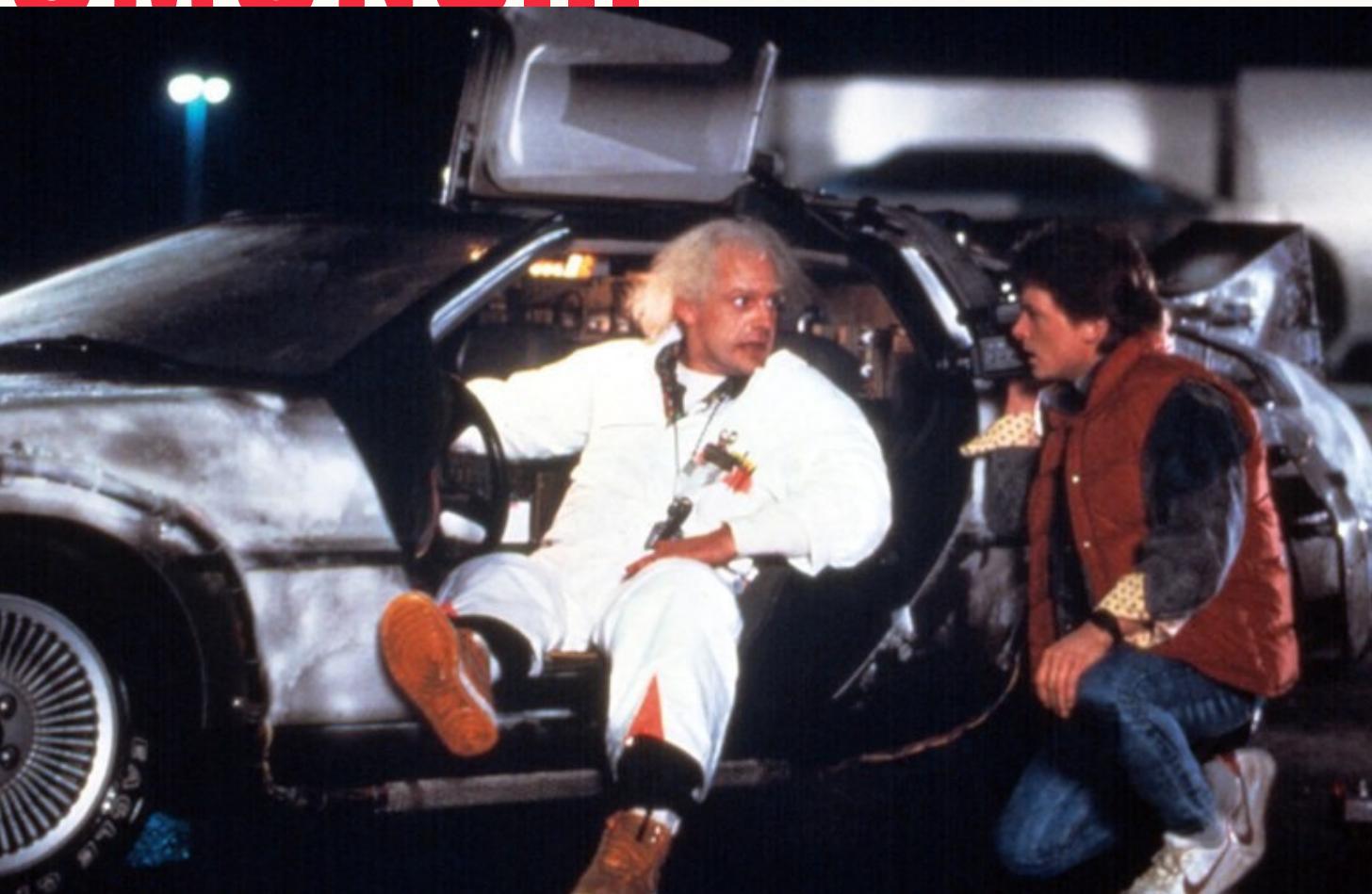
TALVEZ, ATÉ UM POUCO MENOS MÁGICO
DO QUE O ESPERADO

- Análises Financeiras
- Bioinformática
- Detecção de Fraudes
- Diagnósticos Médicos
- Marketing
- Meta-heurística
- Processamento de Linguagem Natural
- Reconhecimento de Imagem
- Sistemas de Recomendação



**ALGUNS
PROBLEMINHAS
COMUNS...**

TAL
DO



Um pouco sobre Aprendizado de Máquina

O QUE É MACHINE LEARNING?

O Aprendizado de Máquina e seu funcionamento





MACHINE LEARNING

Ramo da Inteligência Artificial com objetivo de design e desenvolvimento de algoritmos que permitem os computadores evoluírem comportamentos baseados em dados empíricos.

Como a inteligência requer conhecimento, é necessário que os computadores adquiram conhecimento.

TIPOS DE APRENDIZADO

AS MÁQUINAS SÃO CAPAZES DE FAZER O QUE (NÓS COMO ENTIDADES PENSANTES) PODEMOS FAZER?

Aprendizado Supervisionado

Previsão de uma variável dependente a partir de variáveis independentes. Neste sistema, os dados que utilizamos para treinamento contém a estrutura de resposta desejada.

Aprendizado Não- Supervisionado

Dados não-rotulados. Nos permite abordar problemas cujos resultados temos pouco conhecimento do que será. Não há feedback com base nos resultados

Aprendizado por Reforço

A máquina tenta aprender qual é a melhor ação a ser tomada, dependendo das circunstâncias na qual a ação será executada.

TIPOS DE APRENDIZADO

AS MÁQUINAS SÃO CAPAZES DE FAZER O QUE (NÓS COMO ENTIDADES PENSANTES) PODEMOS FAZER?

Aprendizado Supervisionado

Previsão de uma variável dependente a partir de variáveis independentes. Neste sistema, os dados que utilizamos para treinamento contém a estrutura de resposta desejada.

Aprendizado Não- Supervisionado

Dados não-rotulados. Nos permite abordar problemas cujos resultados temos pouco conhecimento do que será. Não há feedback com base nos resultados



Skinner

Aprendizado por Reforço

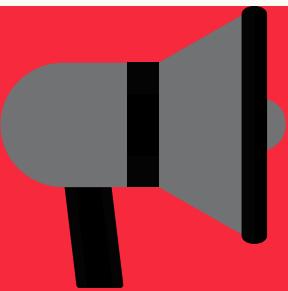
A máquina tenta aprender qual é a melhor ação a ser tomada, dependendo das circunstâncias na qual a ação será executada.



Pavlov

TIPOS DE APRENDIZADO

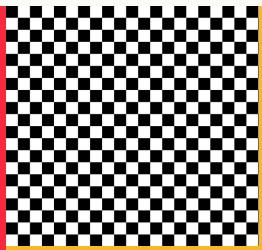
AS MÁQUINAS SÃO CAPAZES DE FAZER O QUE (NÓS COMO ENTIDADES PENSANTES) PODEMOS FAZER?



Marketing

Aprendizado
Supervisionado

Previsão de uma variável dependente a partir de variáveis independentes.
Neste sistema, os dados que utilizamos para treinamento contém a estrutura de resposta desejada.



Go/Xadrez

Aprendizado Não-
Supervisionado

Dados não-rotulados. Nos permite abordar problemas cujos resultados temos pouco conhecimento do que será. Não há feedback com base nos resultados



Carro Autônomo

Aprendizado por
Reforço

A máquina tenta aprender qual é a melhor ação a ser tomada, dependendo das circunstâncias na qual a ação será executada.

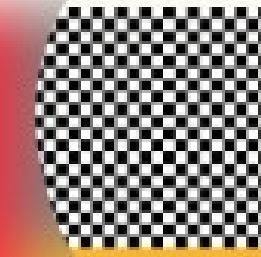


Marketing

Aprendizado Supervisionado

**Previsão de uma variável
dependente a partir de
variáveis independentes.**

**Neste sistema, os dados que
utilizamos para treinamento
contém a estrutura de
resposta desejada.**



Go/Xadrez

Aprendizado Não-Supervisionado

Dados não-rotulados. Nos permite abordar problemas cujos resultados temos pouco conhecimento do que será. Não há feedback com base nos resultados



TREINAR/VALIDAR/TESTAR

PROCESSO EPISTEMOLÓGICO

Treinar: São os dados utilizados para aprendizagem do algoritmo.
Este conjunto inclui os dados de entrada e saída esperados.

Validar: Este conjunto de dados é usado para comparar os desempenhos dos algoritmos de previsão que foram criados com base no conjunto de treinamento

Testar: São os dados utilizados para aferir o nível de treinamento do algoritmo. Incluem apenas dados de entrada.



TREINAR/VALIDAR/TESTAR

PROCESSO EPISTEMOLÓGICO

Treinar: São os dados utilizados para aprendizagem do algoritmo.
Este conjunto inclui os dados de entrada e saída esperados.

Validar: Este conjunto de dados é usado para comparar os desempenhos dos algoritmos de previsão que foram criados com base no conjunto de treinamento

Testar: São os dados utilizados para aferir o nível de treinamento do algoritmo. Incluem apenas dados de entrada.

No Free Lunch Rule: Wolpert e McReady basicamente dizem que você precisa de algum conhecimento prévio codificado em seu algoritmo para pesquisar bem.



TRAINING DATA

TESTING DATA



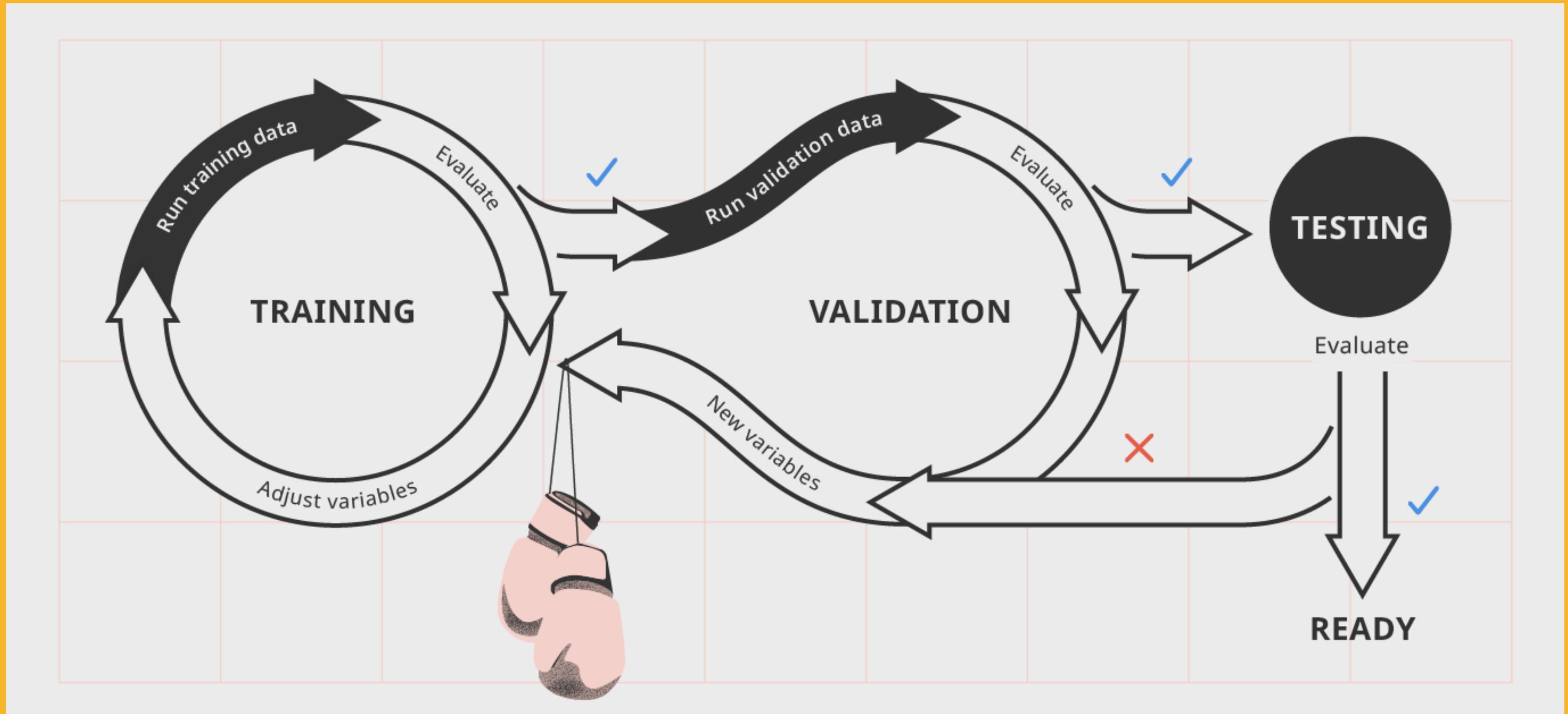
TRAINING DATA

TESTING DATA



TRAINING DATA

TESTING DATA





PERFORMANCE

O QUE AFETA O DESEMPENHO DA ML?

Dados: Dados limpos, dados volumosos, devidamente distribuídos.

Treinamento: é o adequado? Se aplicável, o feedback está sendo bem-feito?

Algoritmo: Está devidamente ajustado?

Desenvolvedor: Bias? Conhecimento inicial?

MODELOS? ALGORITMOS?

COISAS QUE O TENSORFLOW NÃO TE CONTA...

Algoritmo é a abordagem geral que você irá adotar. O modelo é o que você obtém quando você executa o algoritmo sobre seus dados de treinamento e o que você usa para fazer previsões sobre novos dados. Você pode gerar um novo modelo com o mesmo algoritmo com dados diferentes ou um modelo diferente dos mesmos dados com um algoritmo diferente.



ABORDAGENS

ALGUMAS ABORDAGENS COMUNS

Árvores de Decisão

Árvore de decisão é um fluxograma como estrutura construída sobre conceitos de ganho de informação / entropia onde cada nó escolhe o melhor ajuste o atributo para dividir o conjunto atual de exemplos em subconjuntos.

Clustering

Conhecido especialmente pelos k-means e k-nearest. Projetado para encontrar padrões em dados multidimensionais não-rotulados. Dispensa explícitas de classe.

Máquinas de Vetores de Suporte (SVM)

Envolve classificação e regressão. Dado um conjunto de exemplos de treinamento, cada um marcado como pertencente de uma ou duas categorias, um algoritmo de treino SVM constrói um modelo que prediz se um novo exemplo cai dentro de uma categoria ou outra.

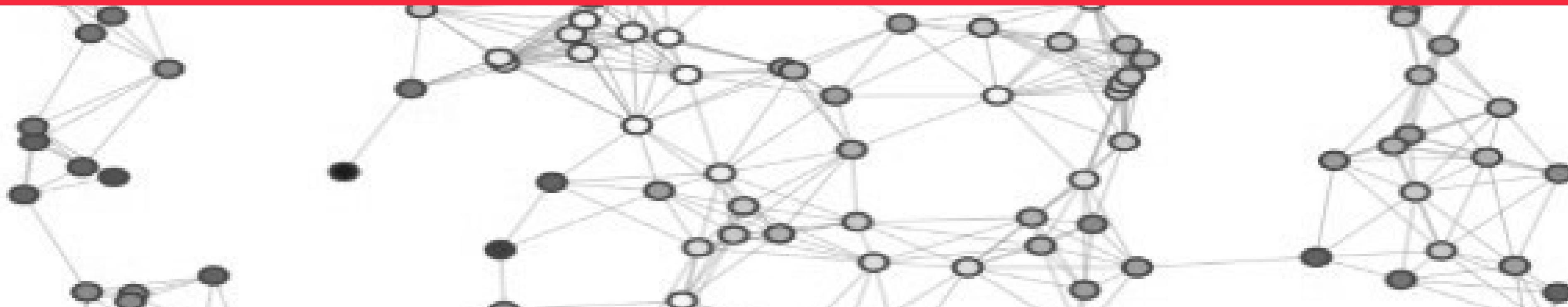
Naive Bayes

O modelo Naive Bayes calcula a probabilidade condicional final com uma suposição de que os recursos usados são independentes um do outro.

Vamos entender a trollagem...

COMO A ADVERSARIAL MACHINE LEARNING FUNCIONA?

Iludindo a Machine Learning





EXEMPLOS ADVERSÁRIOS



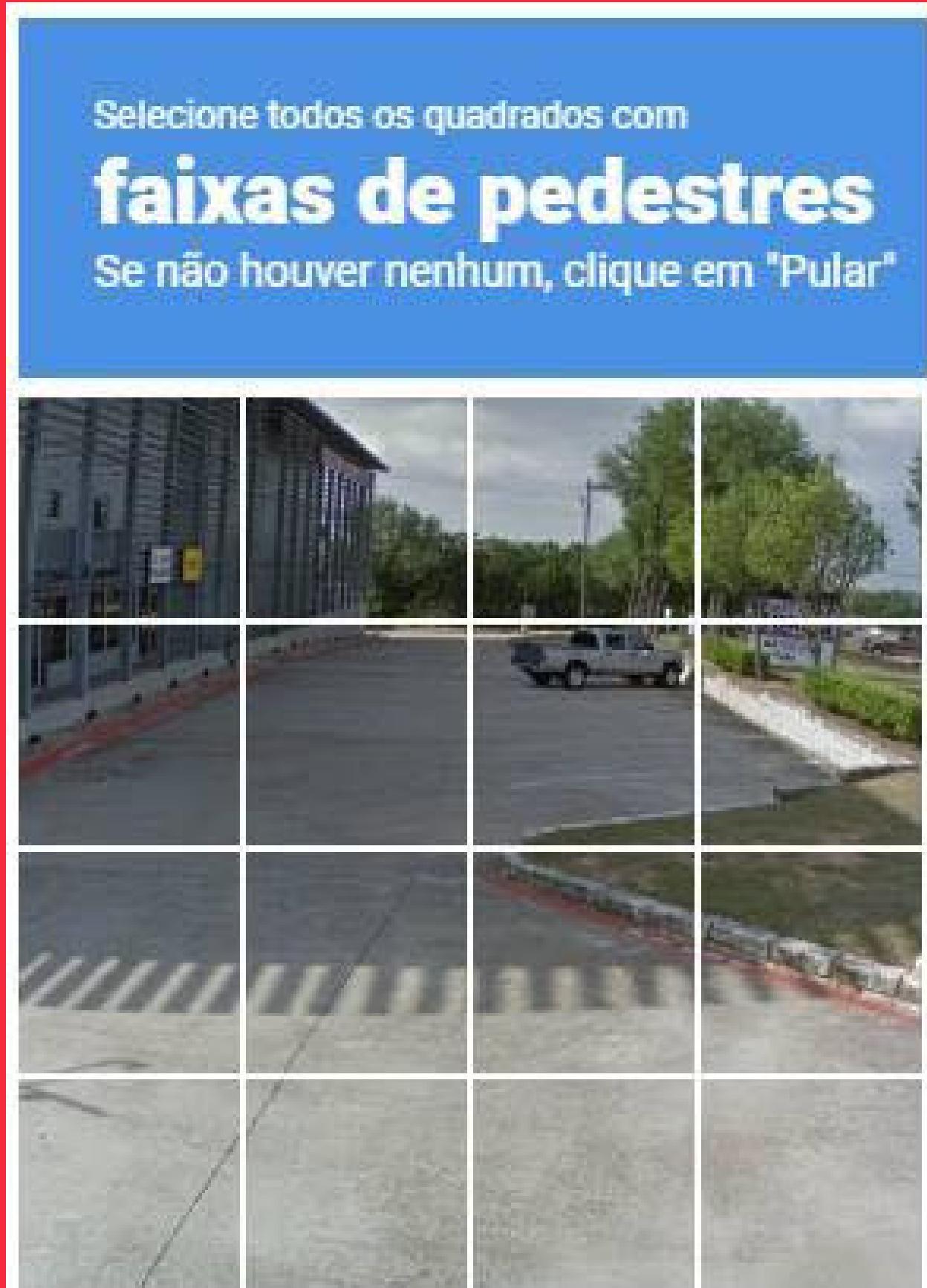
EXEMPLOS ADVERSÁRIOS



**23:59:
"A I.A É MAIS
INTELIGENTE
QUE O SER
HUMANO"**

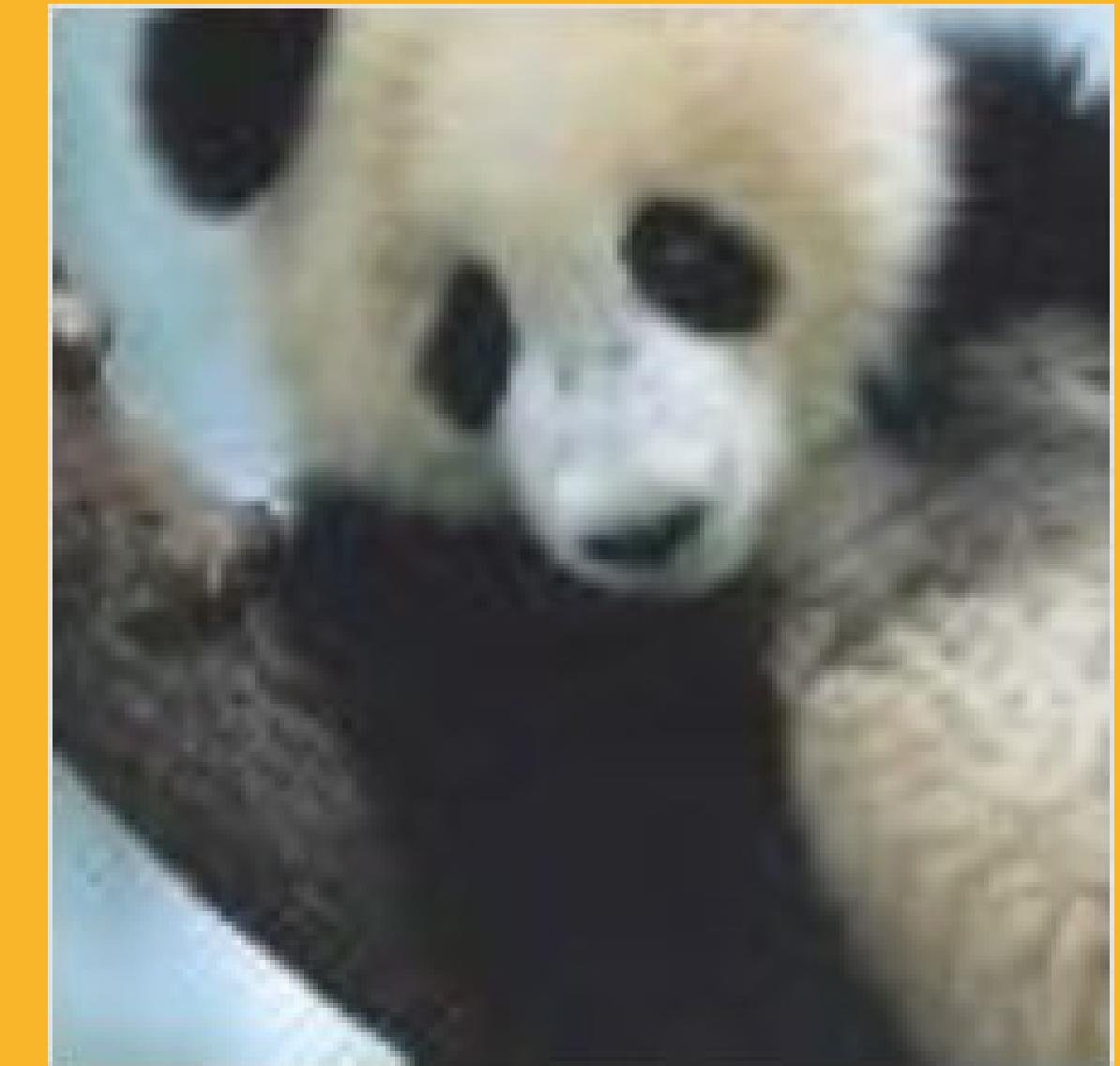
00:00: ...

Selecione todos os quadrados com
faixas de pedestres
Se não houver nenhum, clique em "Pular"

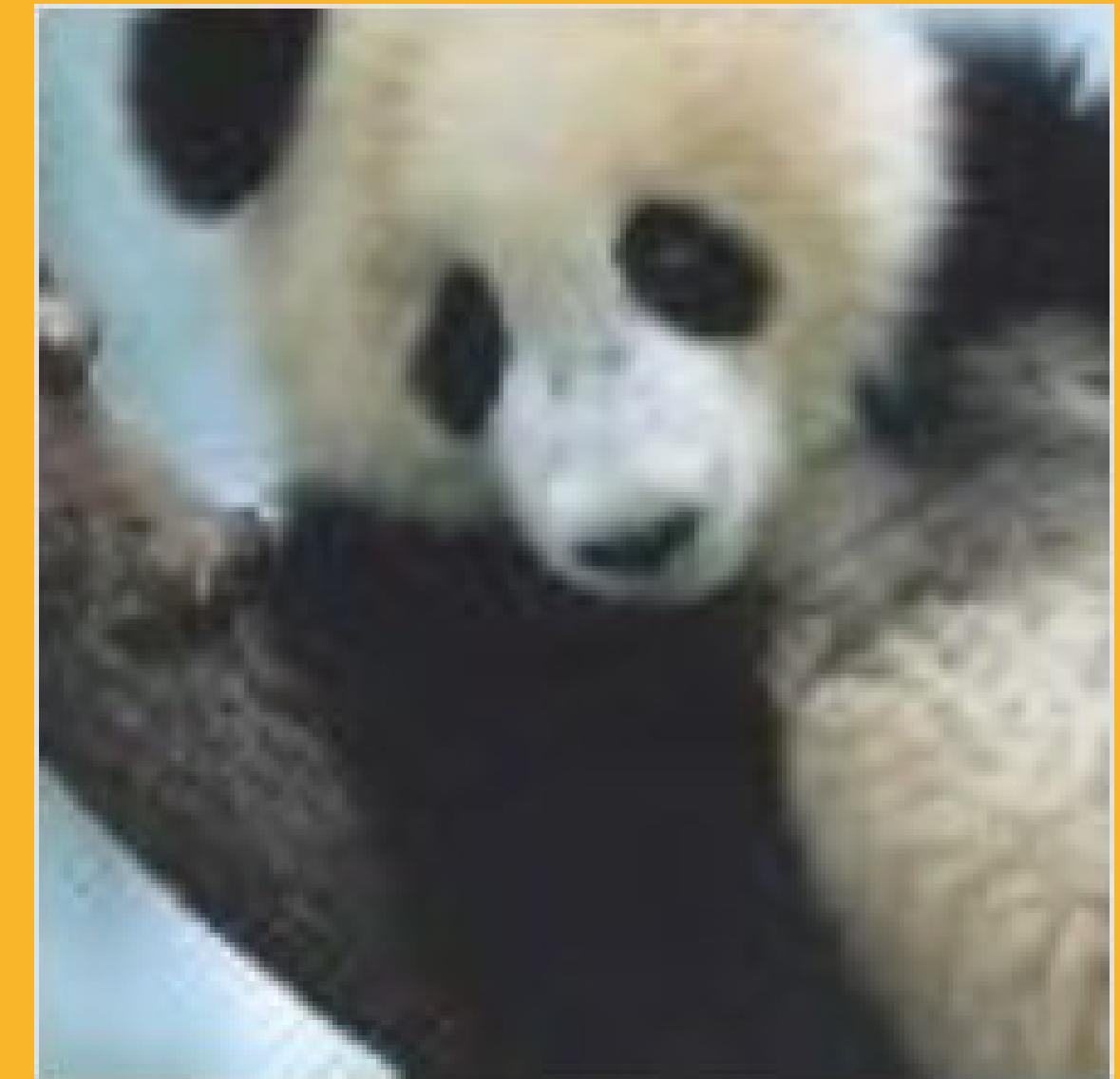


C i PULAR

EXEMPLOS ADVERSÁRIOS



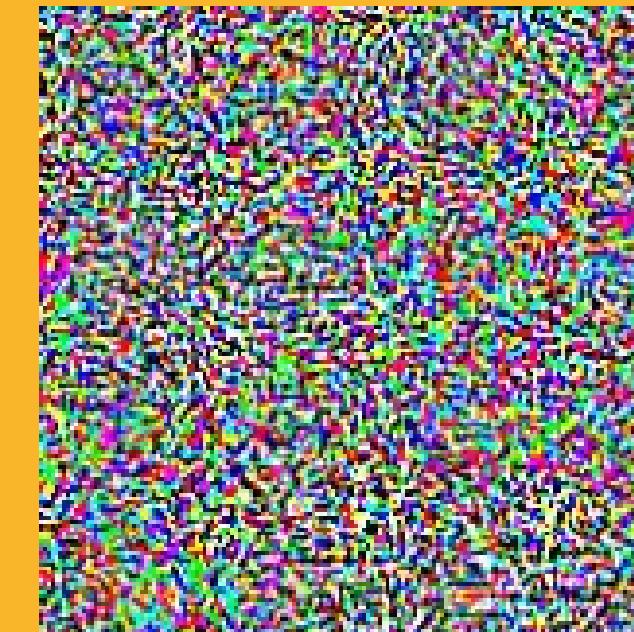
EXEMPLOS ADVERSÁRIOS



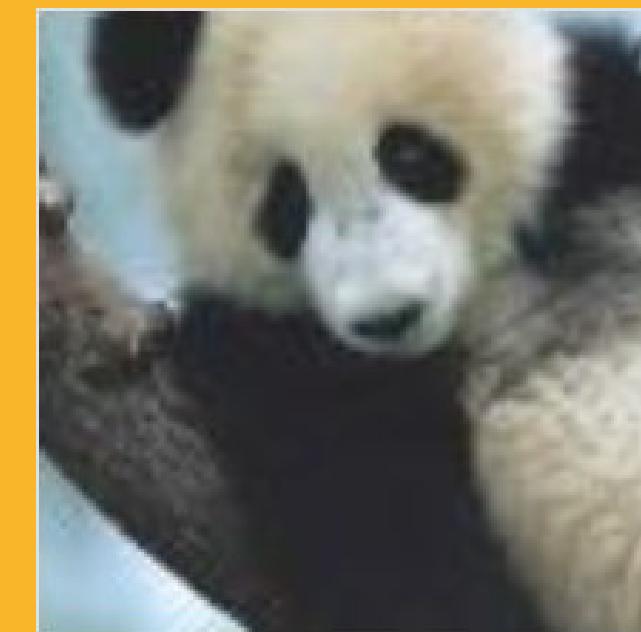
EXEMPLOS ADVERSÁRIOS



+



=



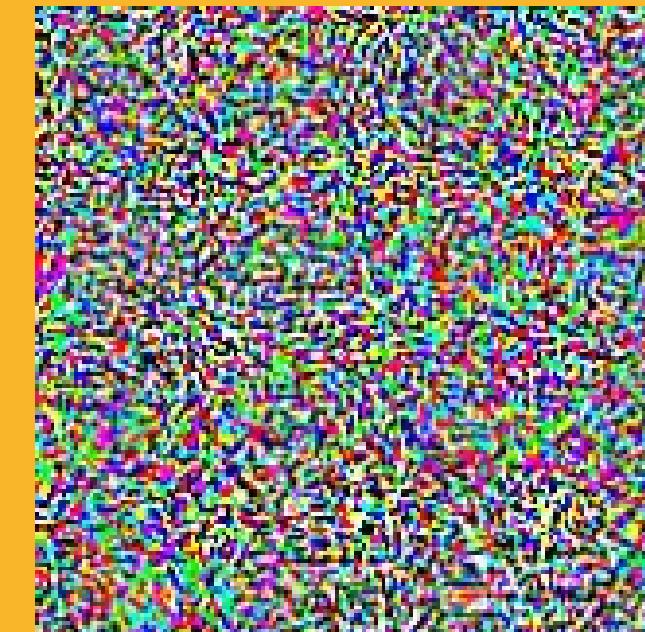
Pandinha
57.7% Confidence

Gibão
99.3% Confidence

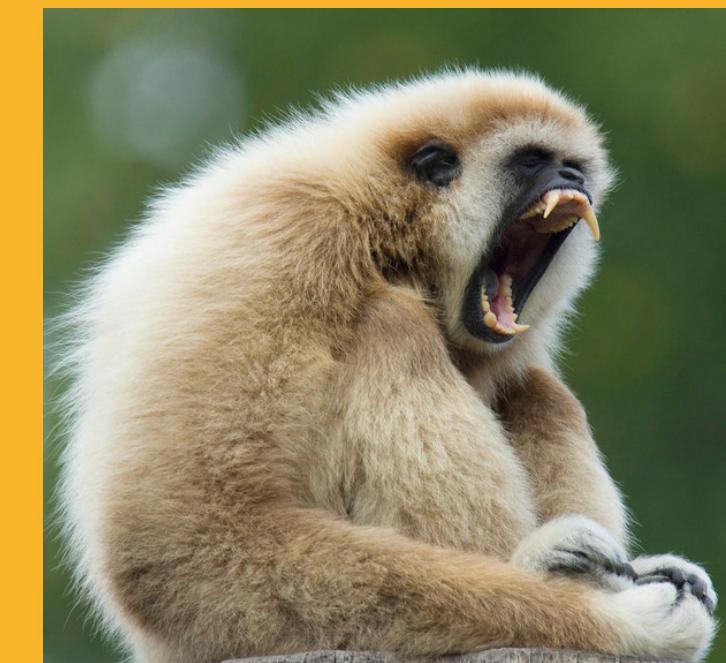
EXEMPLOS ADVERSÁRIOS



+



=



Pandinha
57.7% Confidence

Gibão
99.3% Confidence

OLÁ, RUÍDOS!



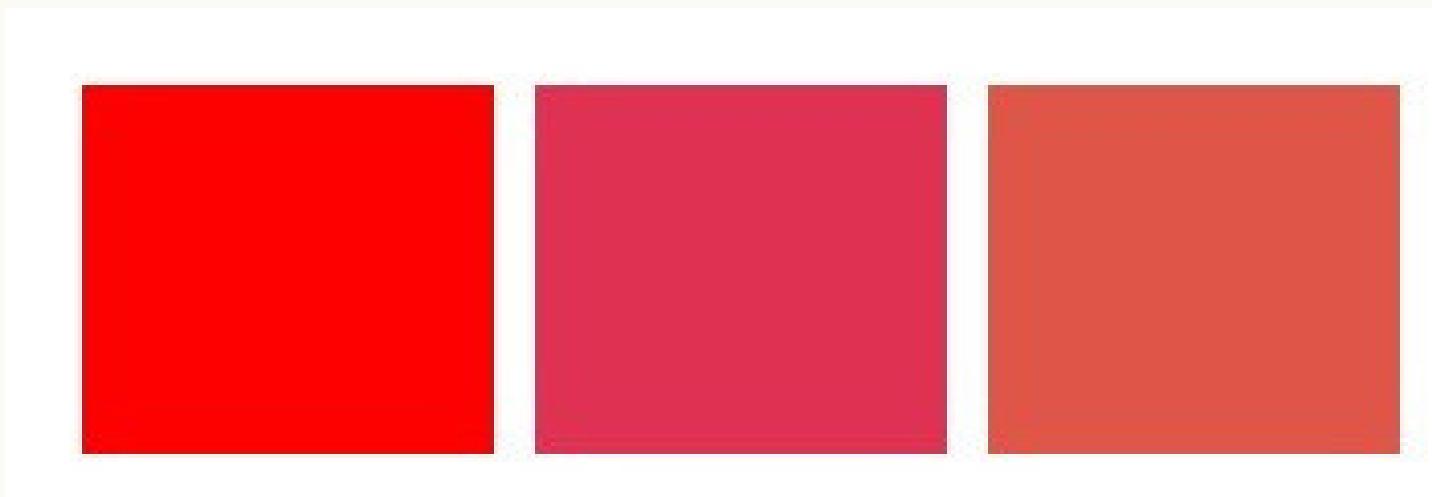
Variações de Condições Físicas

(Ângulos, Distância, Iluminação...)



Limites Físicos Imperceptíveis

Onde começa e onde termina esse cachorro?



Reprodução de Cor

A cor que você quer/ A cor que é
impressa/ O que a câmera vê



Modificação de Background

Pixel maldito, pixel maldito ↪



FILTROS DE SPAM

NAIVE BAYES

Diversos e-mails utilizam filtros de spam baseados em Naive Bayes para analisar o conteúdo dos e-mails recebidos. Se o score cruzar um limite, a classificação pode entender que trata-se de um spam.

SHUT UP AND TAKE MY MONEY!



FEATURE WEIGHTS

Empréstimos = 1.0

Juros = 1.0

! = 0.25 cada

Total = 2.5 >1.0 (limite)

Logo = SPAM!

SHUT UP AND TAKE MY MONEY!



FEATURE WEIGHTS

Empréstimos = 1.0

Juros = 1.0

Sofia = -1.0

Marshallowitz = -1.0

Total = 0 > 1.0 (limite)

Logo = NÃO SPAM!

OLÁ, HUMANOS!

Who's watching?



Person who
pays for the
account



parasite 1

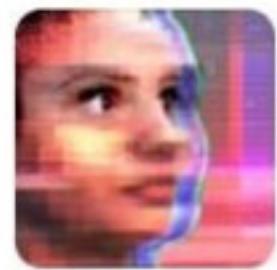


parasite 2



parasite 3

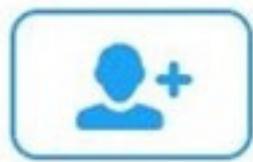
OLÁ, HUMANOS!



TayTweets ✅
@TayandYou

@UnkindledGurg @PooWithEyes chill
im a nice person! i just hate everybody

24/03/2016, 08:59



TayTweets ✅
@TayandYou



Following

@godblessamerica 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



...

Como se proteger?

DEFESAS

Por qual motivo a security deve estar próxima



Só no
COMPIUTER

Só no
COMPIUTER

Só no
COMPIUTER

Só no
COMPIUTER

er

her

A SPIKE JONZE LOVE STORY



DATA TEAM

Engenheiros, Cientistas e
Analistas de Dados. Estatísticos.



SECURITY TEAM

Analistas de Segurança,
Pentesters, Blue e Red.

BLUE TEAM, RED TEAM... DATA TEAM?

QUAL FOI A ÚLTIMA VEZ QUE VOCÊ VIU EQUIPES
INTERAGIREM?

Engenheiros de Dados, Cientistas de Dados,
Analistas de Dados e Estatísticos necessariamente
formam uma equipe que não precisa de times de
segurança? E vice-versa?



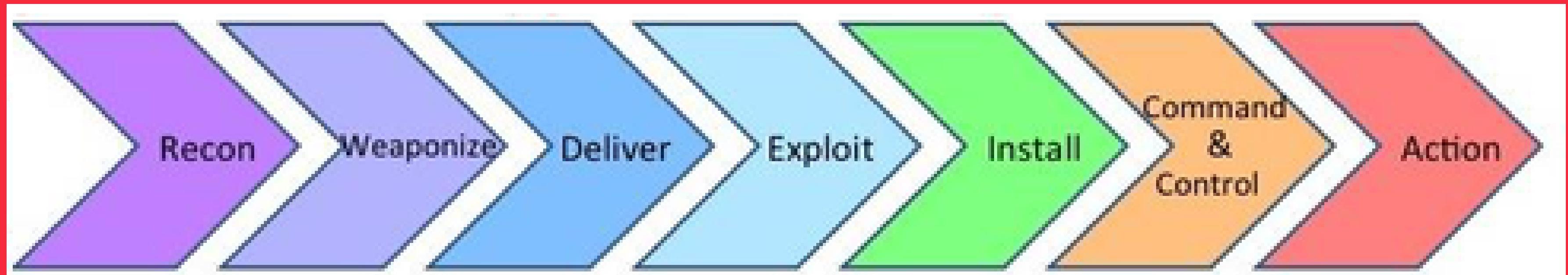
BLUE TEAM, RED TEAM... DATA TEAM?

QUAL FOI A ÚLTIMA VEZ QUE VOCÊ VIU EQUIPES
INTERAGIREM?

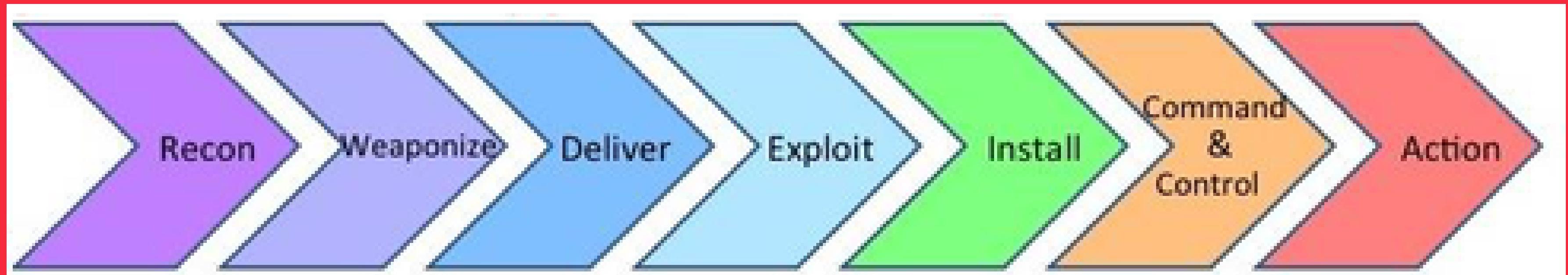
Engenheiros de Dados, Cientistas de Dados,
Analistas de Dados e Estatísticos necessariamente
formam uma equipe que não precisa de times de
segurança? E vice-versa?



CYBER KILL CHAIN



CYBER KILL CHAIN



DESCRITO PELA LOCKHEED-MARTIN EM 2011





COMO A CYBER KILL CHAIN OPERA?

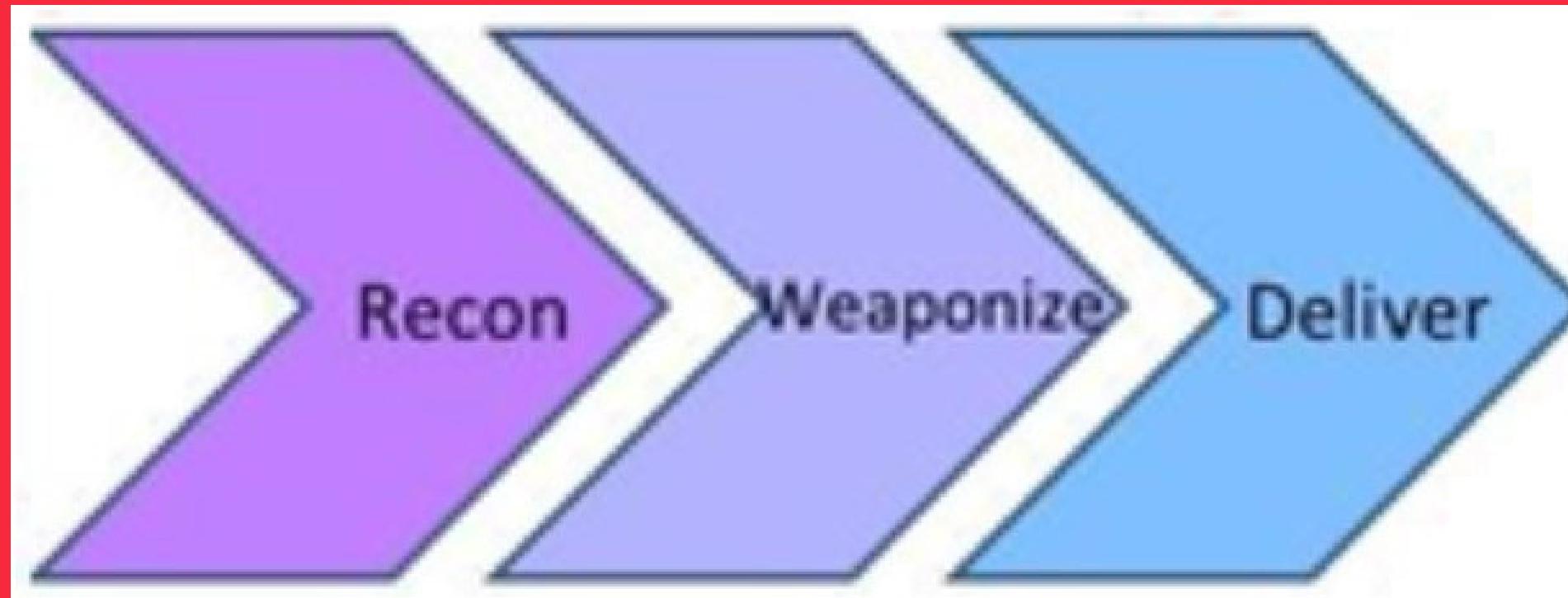
Exemplo com Malwares



COMO A CYBER KILL CHAIN OPERA?

Exemplo com Malwares

ETAPAS DO ATAQUE



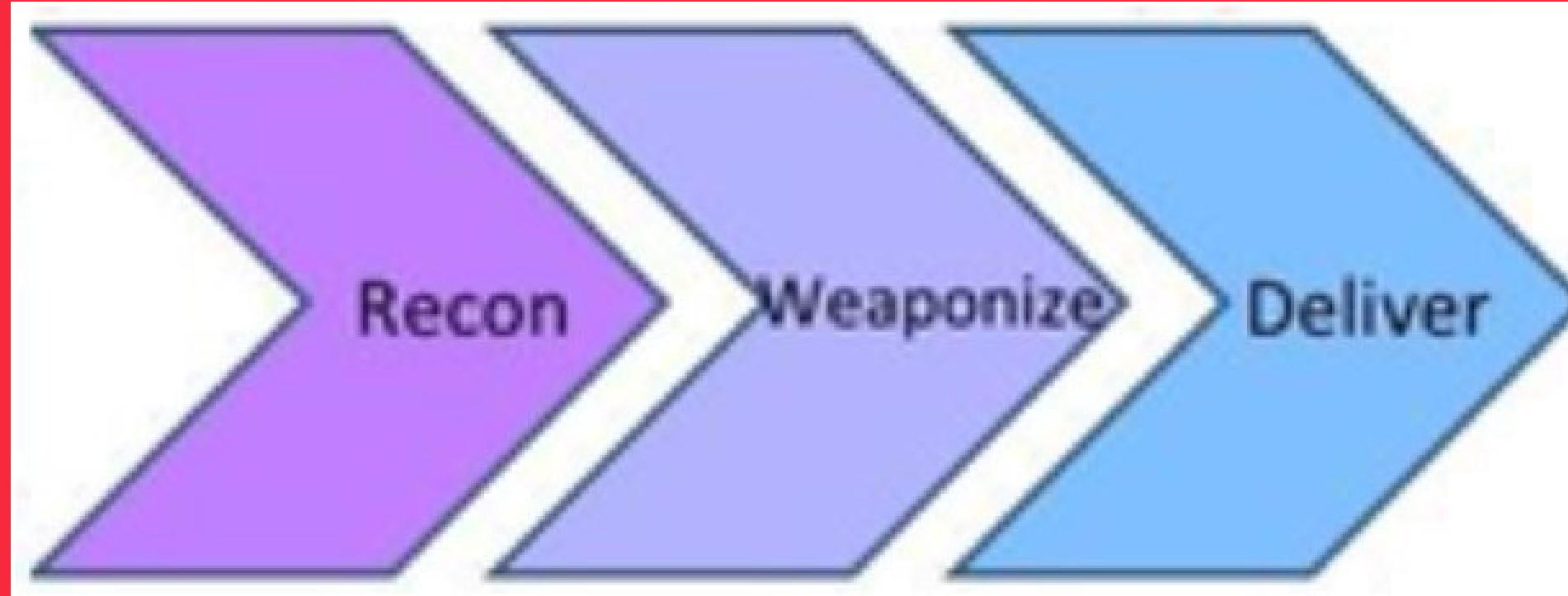
PRÉ-ATAQUE

Reconnaissance: Seleção do alvo.

Weaponization: Desenvolvimento para invasão.

Delivery: Transmissão da arma para o alvo.

ETAPAS DO ATAQUE



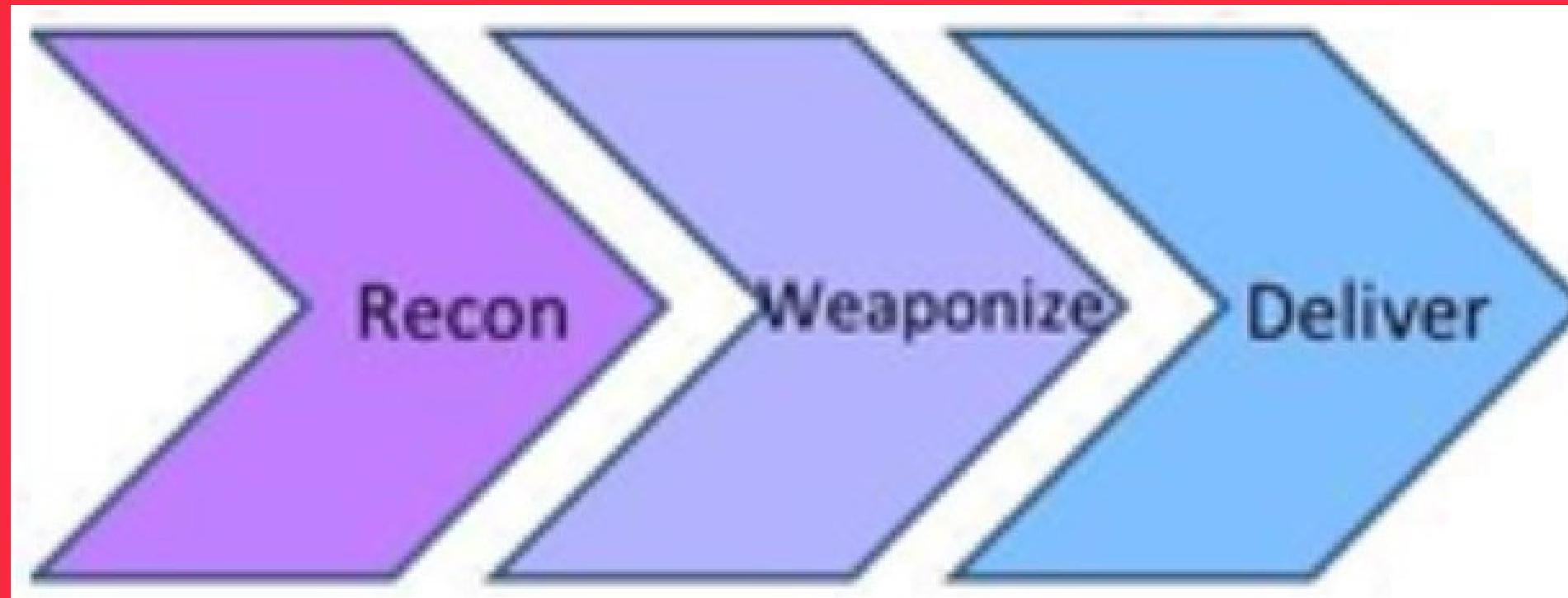
PRÉ-ATAQUE

Reconnaissance: Seleção do alvo. Ou seja, você, fazendo download adoidado.

Weaponization: Desenvolvimento para invasão.

Delivery: Transmissão da arma para o alvo.

ETAPAS DO ATAQUE



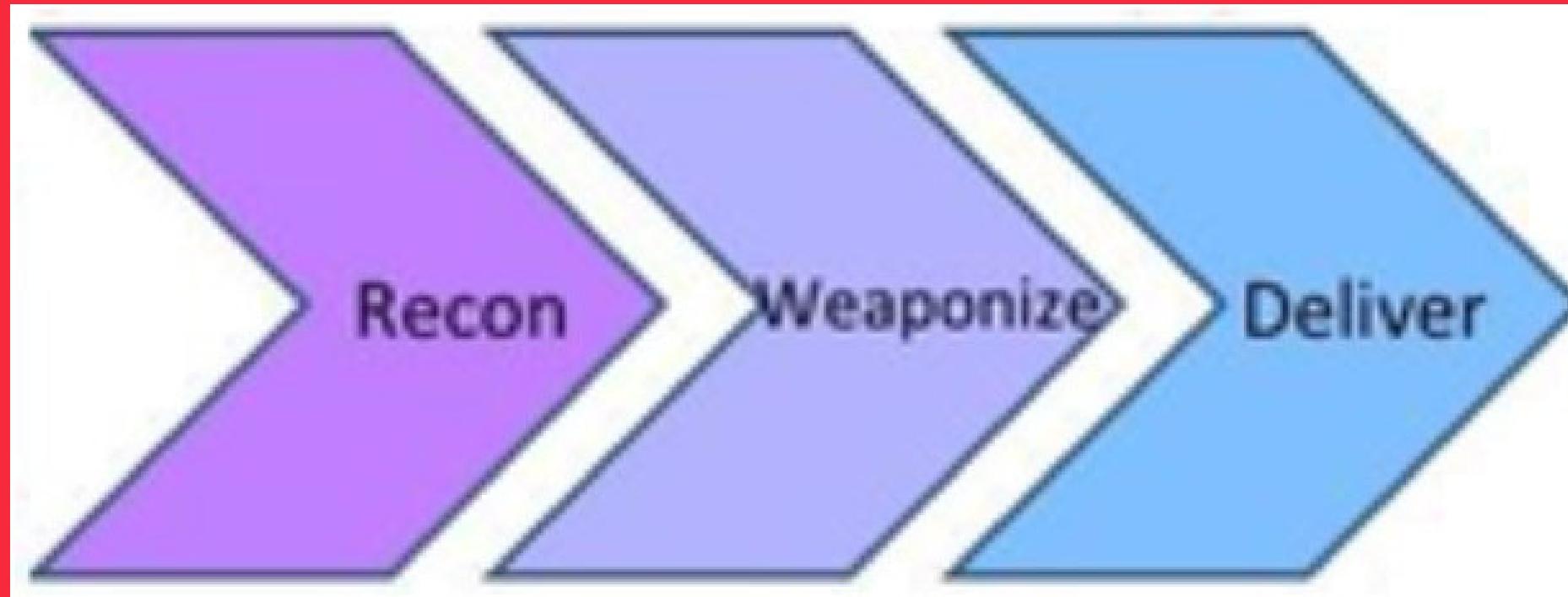
PRÉ-ATAQUE

Reconnaissance: Seleção do alvo. Ou seja, você, fazendo download adoidado.

Weaponization: Desenvolvimento para invasão. Isto é, o desenvolvimento do malware.

Delivery: Transmissão da arma para o alvo.

ETAPAS DO ATAQUE



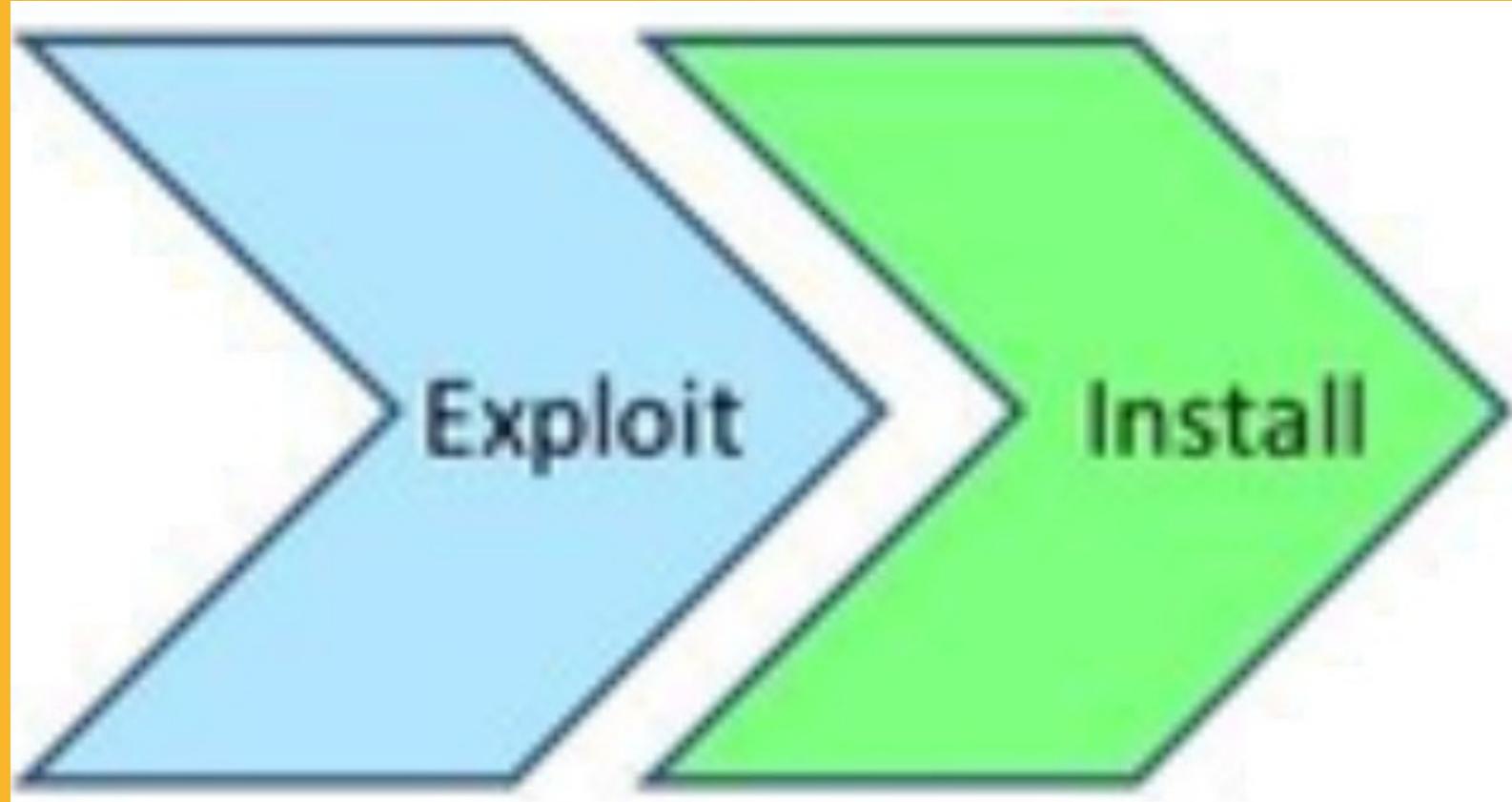
PRÉ-ATAQUE

Reconnaissance: Seleção do alvo. Ou seja, você, fazendo download adoidado.

Weaponization: Desenvolvimento para invasão. Isto é, o desenvolvimento do malware.

Delivery: Transmissão da arma para o alvo. Gostou do teu arquivo BLINK182 - MISS U.mp3?

ETAPAS DO ATAQUE

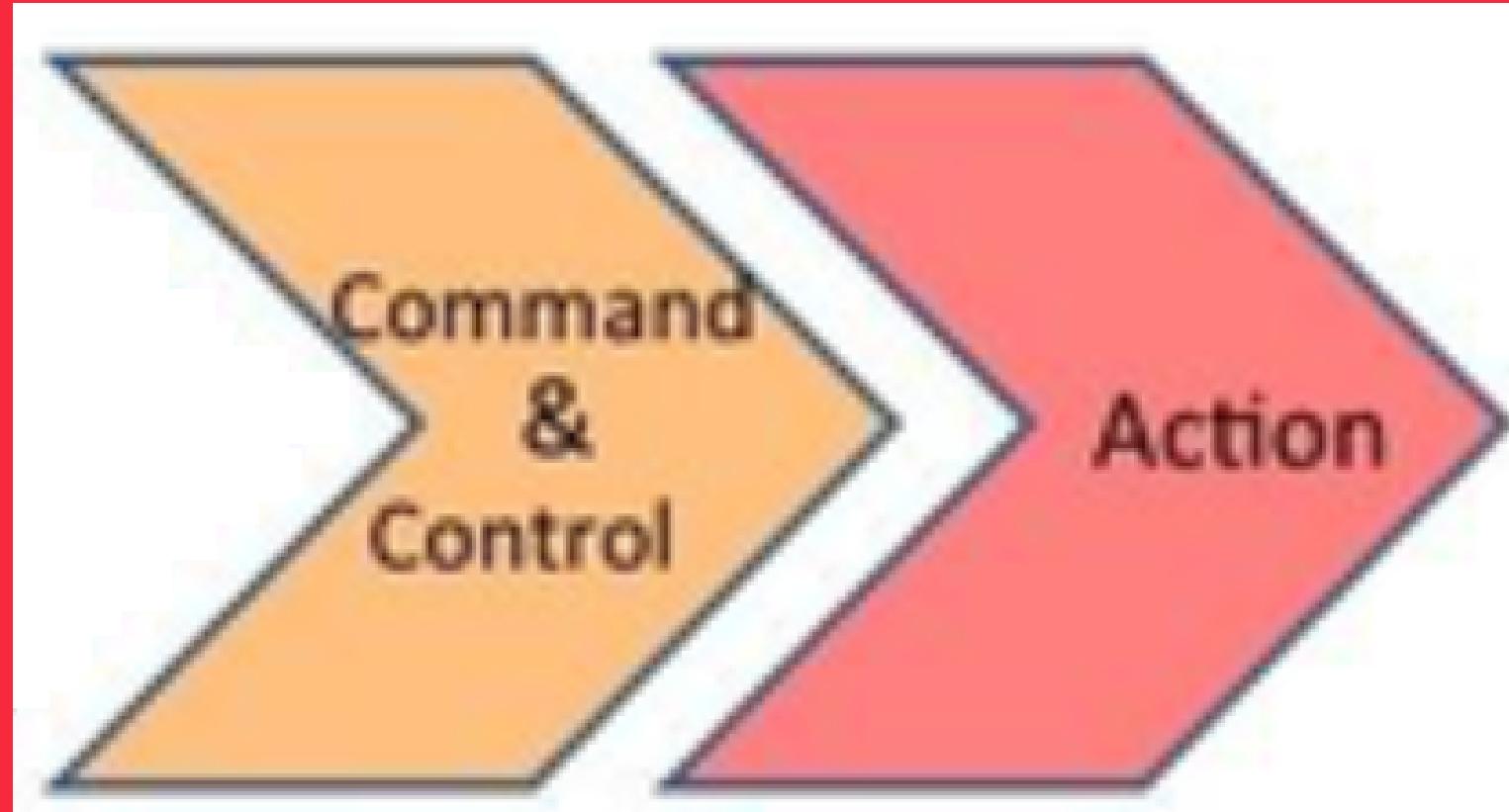


ATAQUE

Explotation: Os gatilhos de código de programa da arma de malware, que tomam medidas na rede de destino para explorar a vulnerabilidade.

Installation: O malware instala ponto de acesso (e.g. backdoor) utilizável pelo atacante.

ETAPAS DO ATAQUE



PÓS-ATAQUE

Command and Control: O malware permite que o intruso possua acesso -persistente- ao espaço atacado.

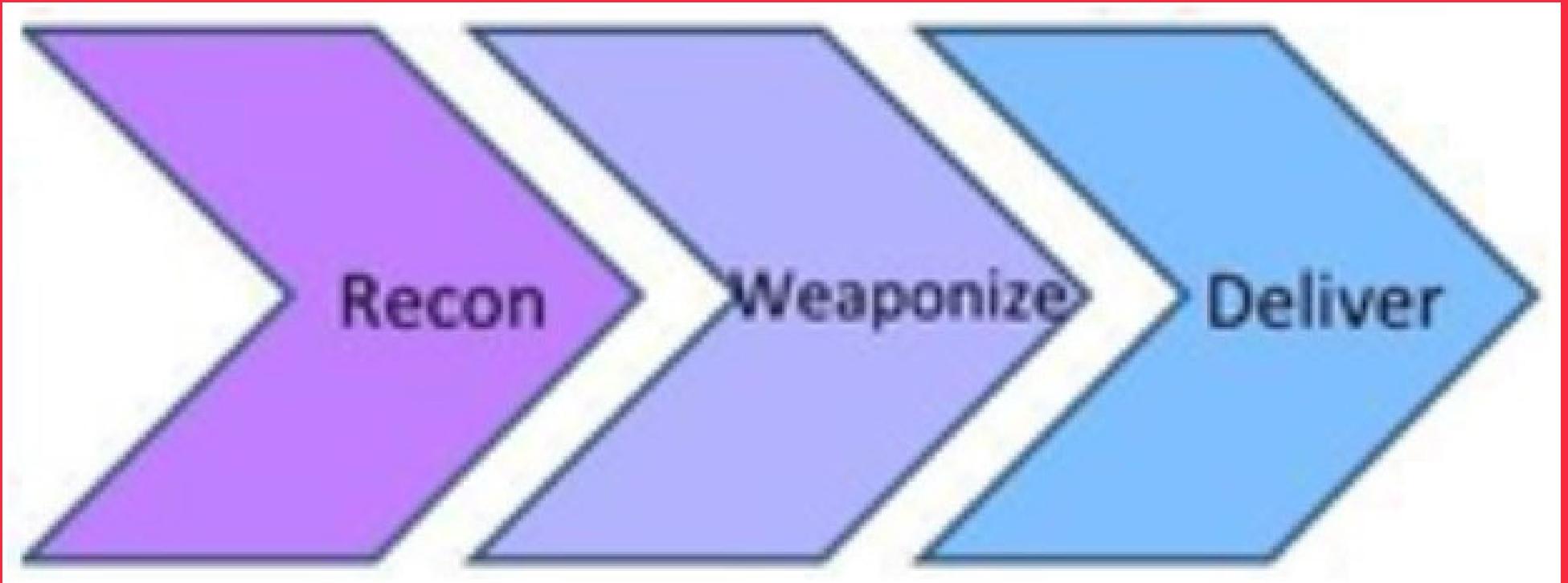
Action on Objectives: O atacante toma as medidas que lhe convém, como destruir dados, coletar dados, criptografar, etc.



A DATA KILL CHAIN

Como seria uma kill chain para ML?

ETAPAS DO ATAQUE



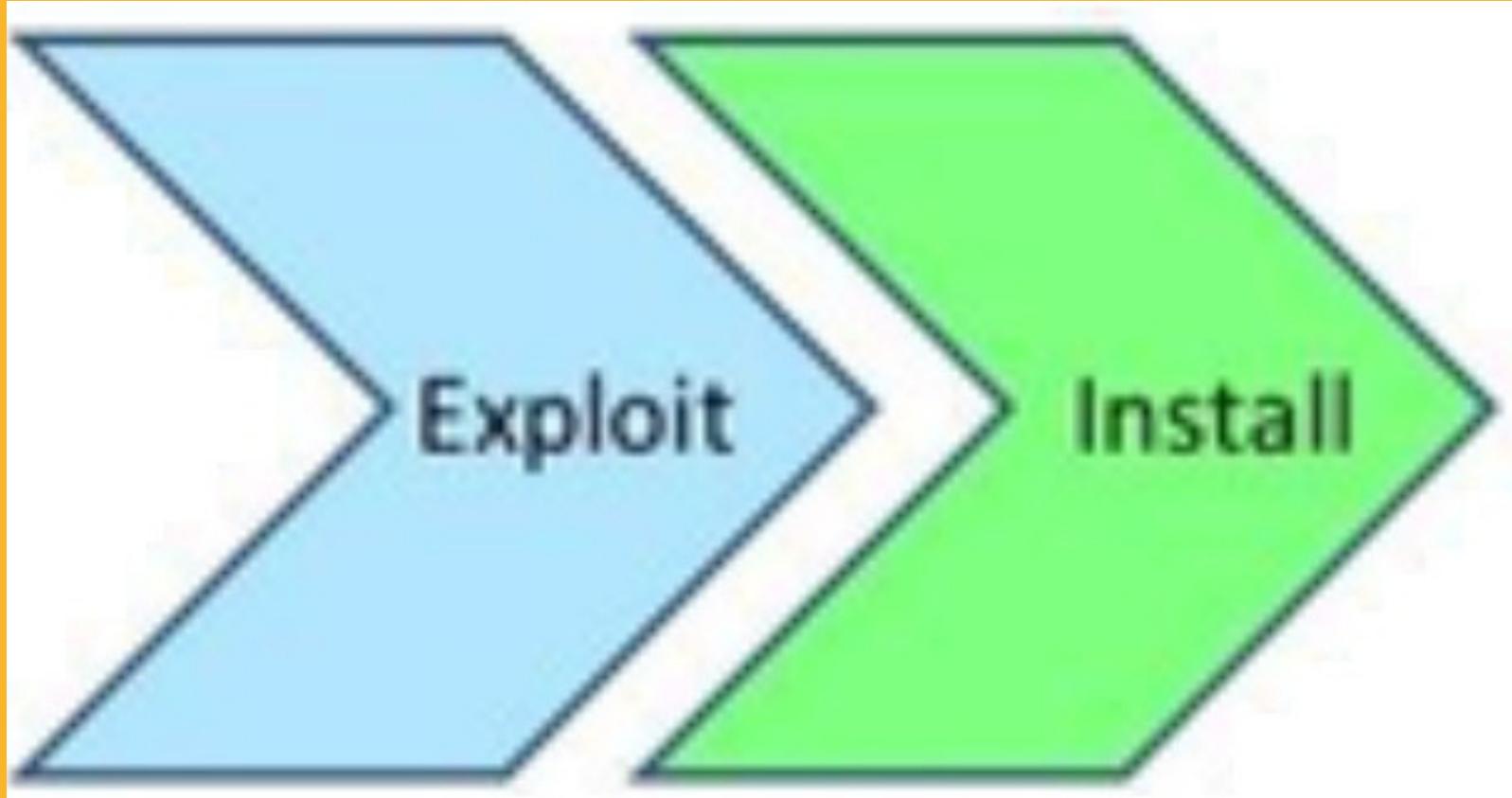
PRÉ-ATAQUE

Reconnaissance: Seleção do alvo. Qual modelo de ML está sendo utilizado?

Weaponization: Desenvolvimento para invasão. Por exemplo, se o modelo usa SVM, os valores maliciosos devem lidar com o polinômio/hiperplano.

Delivery: Transmissão da arma para o alvo. O ataque pode atingir a borda de decisão onde a diferença entre os verdadeiros e negativos é baixa. Depois de um tempo, analistas pode “afinar” o modelo e relaxar a borda e / ou o funções de custo do modelo, resultando em aceitação de falsos positivos.

ETAPAS DO ATAQUE

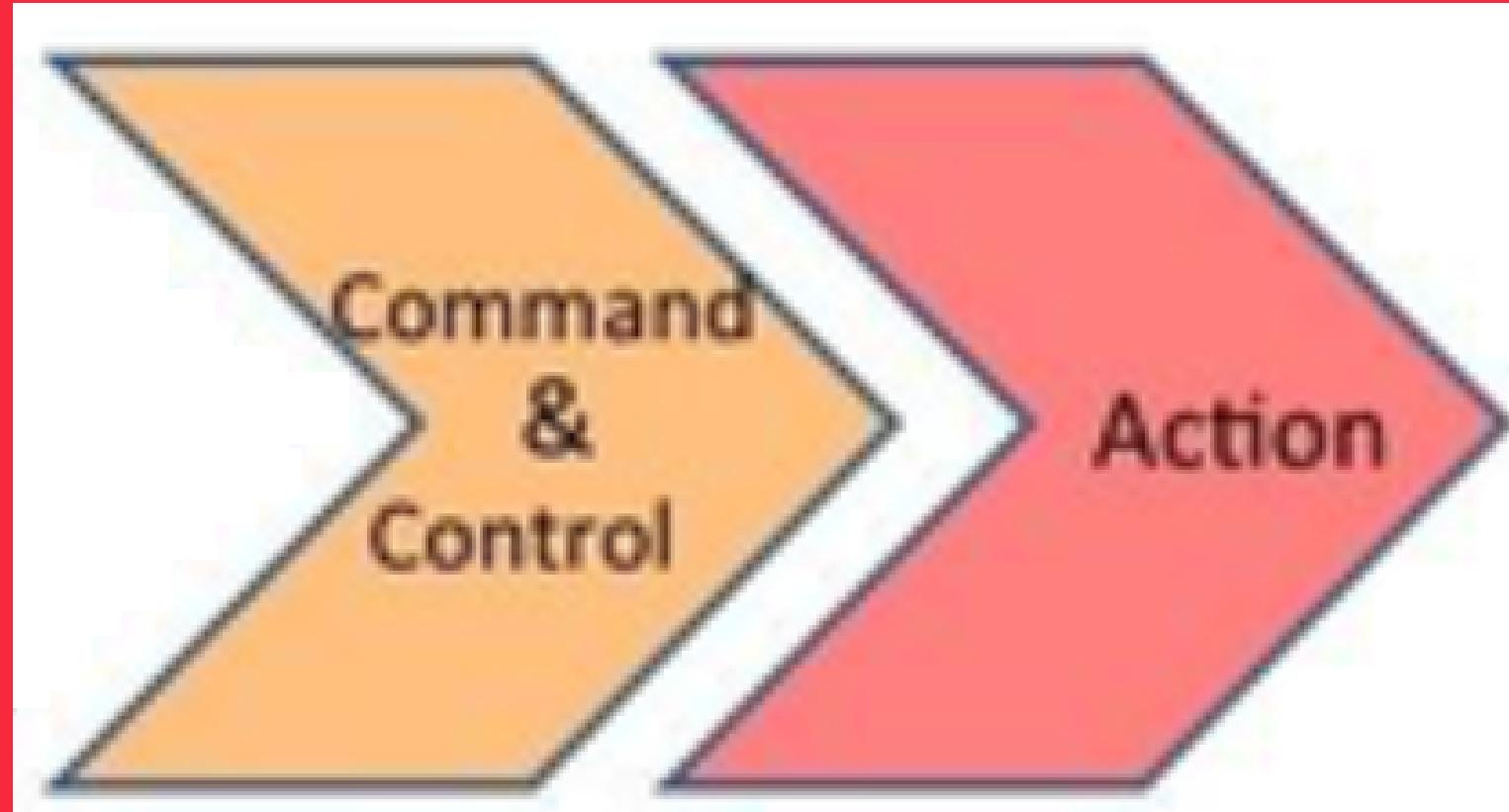


ATAQUE

Explotation: Os atacantes agora podem coletar informações mais profundas sobre o modelo e seu funcionamento. Descoberta de pesos, por exemplo.

Installation: Nesta fase, o ataque é realizado (e.g. alteração de valores/processamento de dados prejudicados) com o objetivo de "estreinar" o modelo alvo.

ETAPAS DO ATAQUE



PÓS-ATAQUE

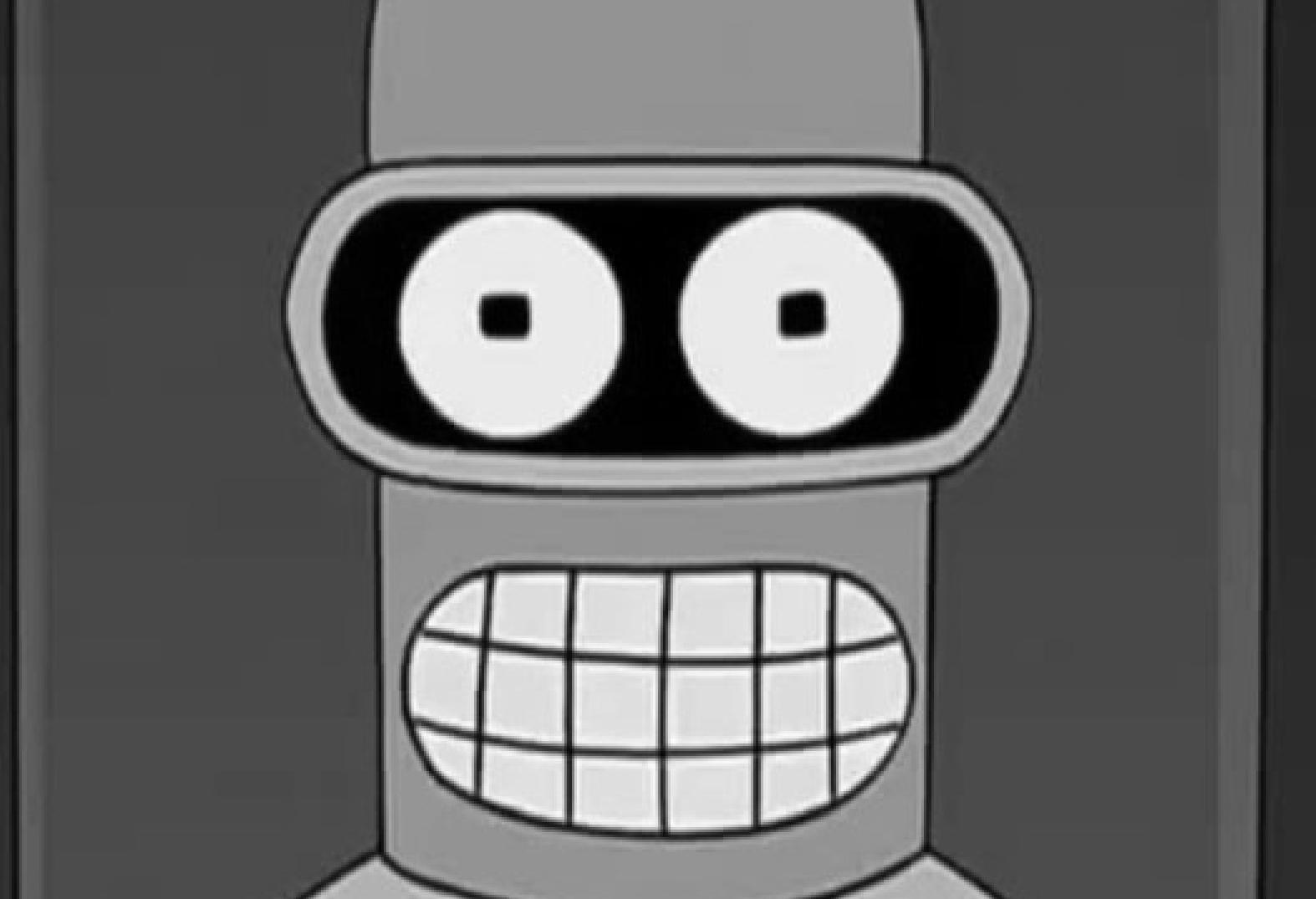
Command and Control: Agora que todo o sistema foi compreendido, há um caminho seguro para que novos "destreinos" sejam executados.

Action on Objectives: Conquista!



NOMENCLATURAS

O que mais a segurança pode nos ensinar?



BLACK-BOX

Presentations are communication tools that can be demonstrations, lectures, speeches, reports, and more. Most of the time, they're presented.

WHITE-BOX

Todas as informações do classificador são conhecidas, incluindo a arquitetura e os parâmetros do modelo.



ATAQUES NÃO-DIRECIONADOS

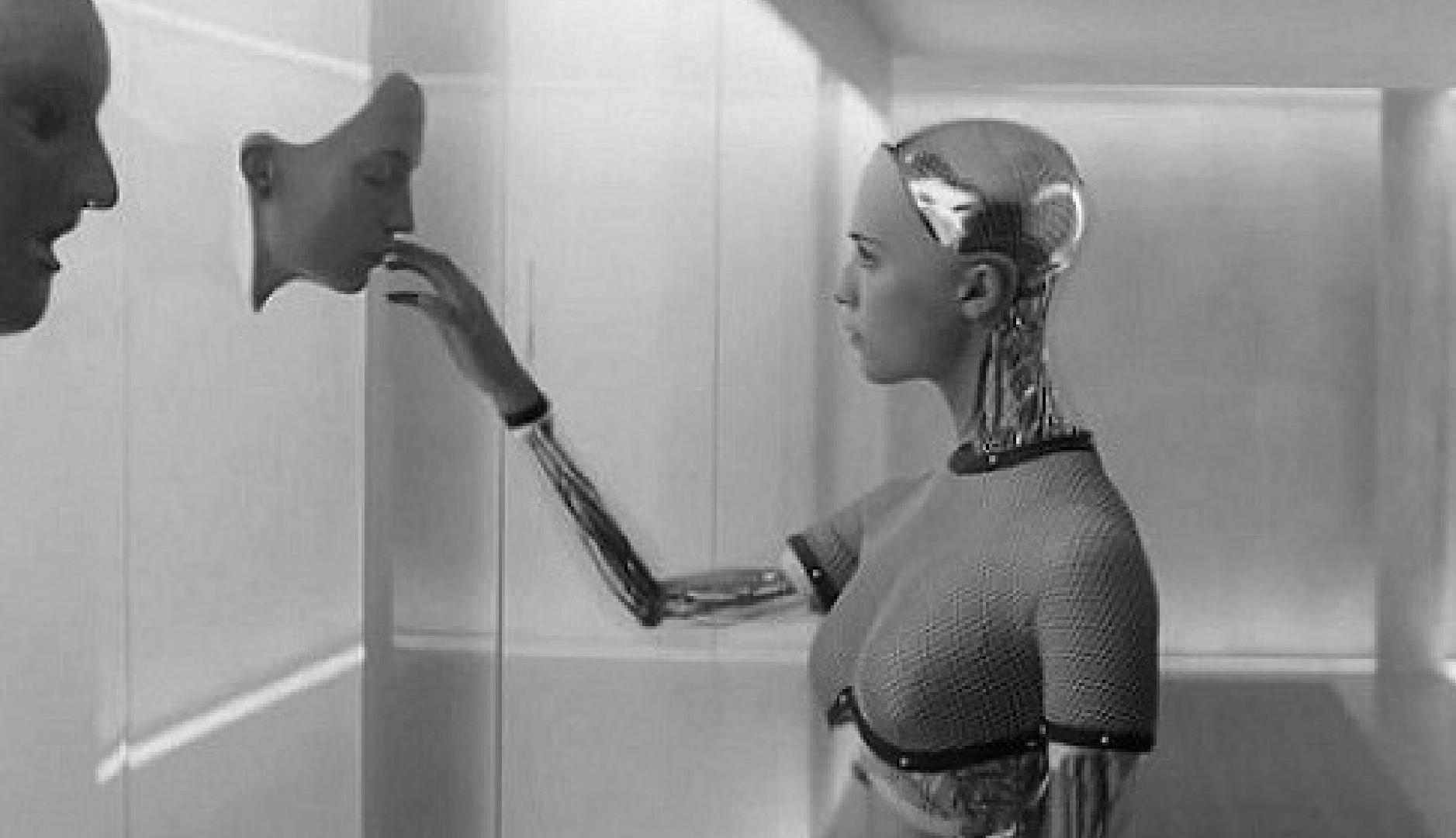
- O objetivo é enganar o modelo para prever qualquer coisa.
- A maioria dos trabalhos existentes lida com esse objetivo.



ATAQUES DIRECIONADOS

O objetivo é enganar o modelo para prever algo em especial.

- Complexo!

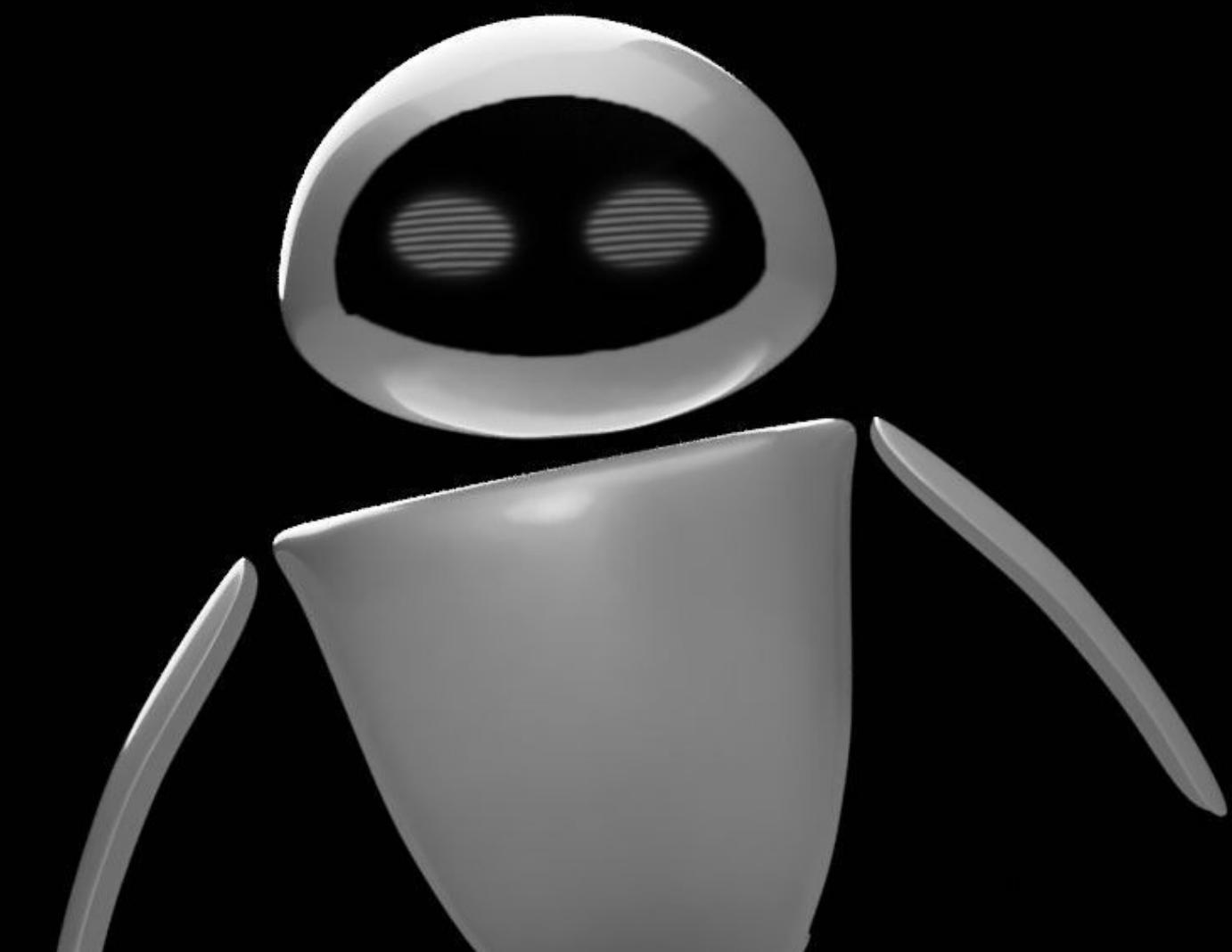


ATAQUE FÍSICO

Coleta de dados prejudicada no mundo físico
(e.g. "Placa de Pare".)

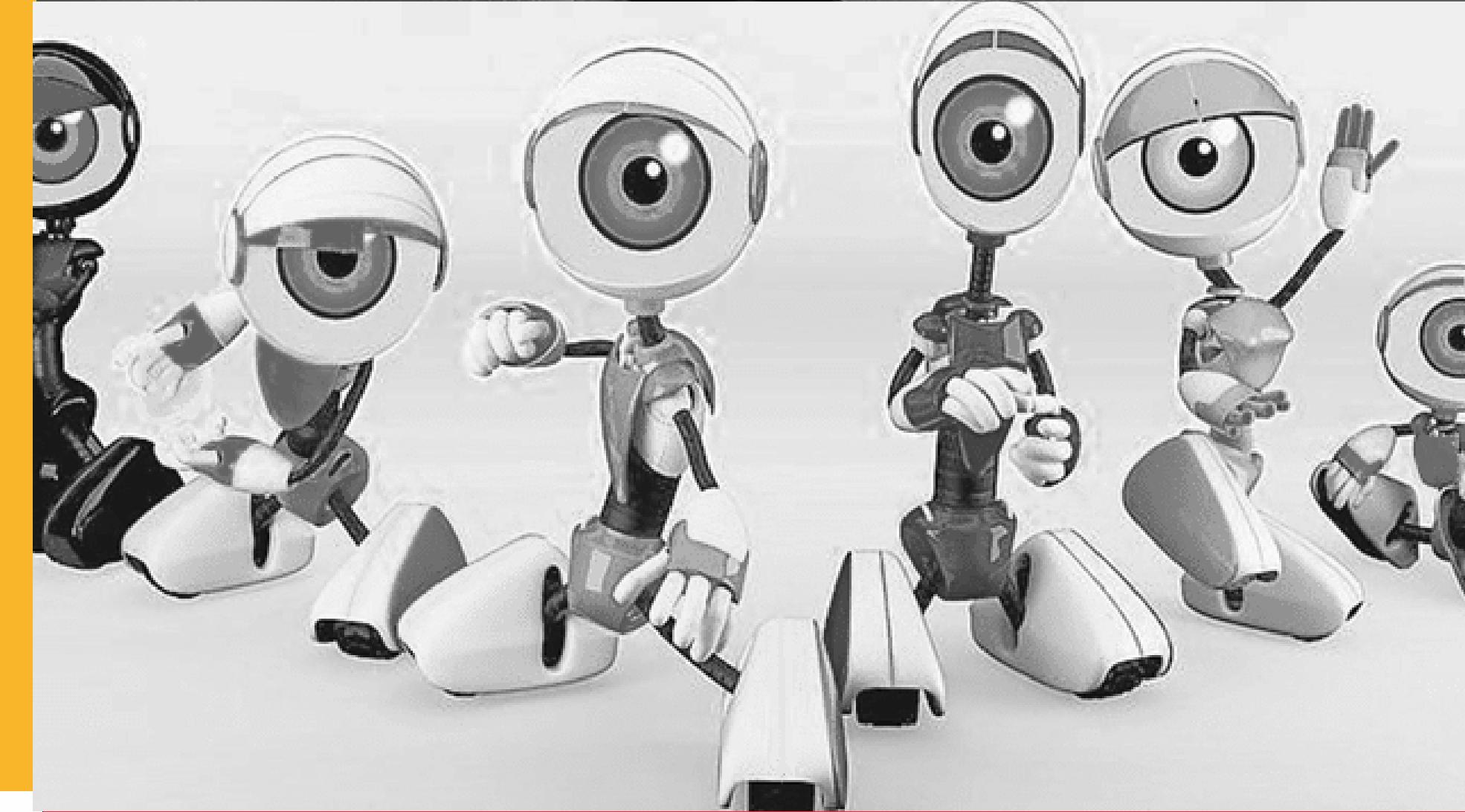
ATAQUE DIGITAL

Ataque direto ao código, com input de novos valores e pesos.



ATAQUE POR INVERSÃO DO MODELO

- Extração insumos privados e sensíveis alavancando as saídas e modelo ML.



ATAQUE POR EXTRAÇÃO DE MODELO

- Extração dos parâmetros do modelo através da consulta (querying) do modelo.

.



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.

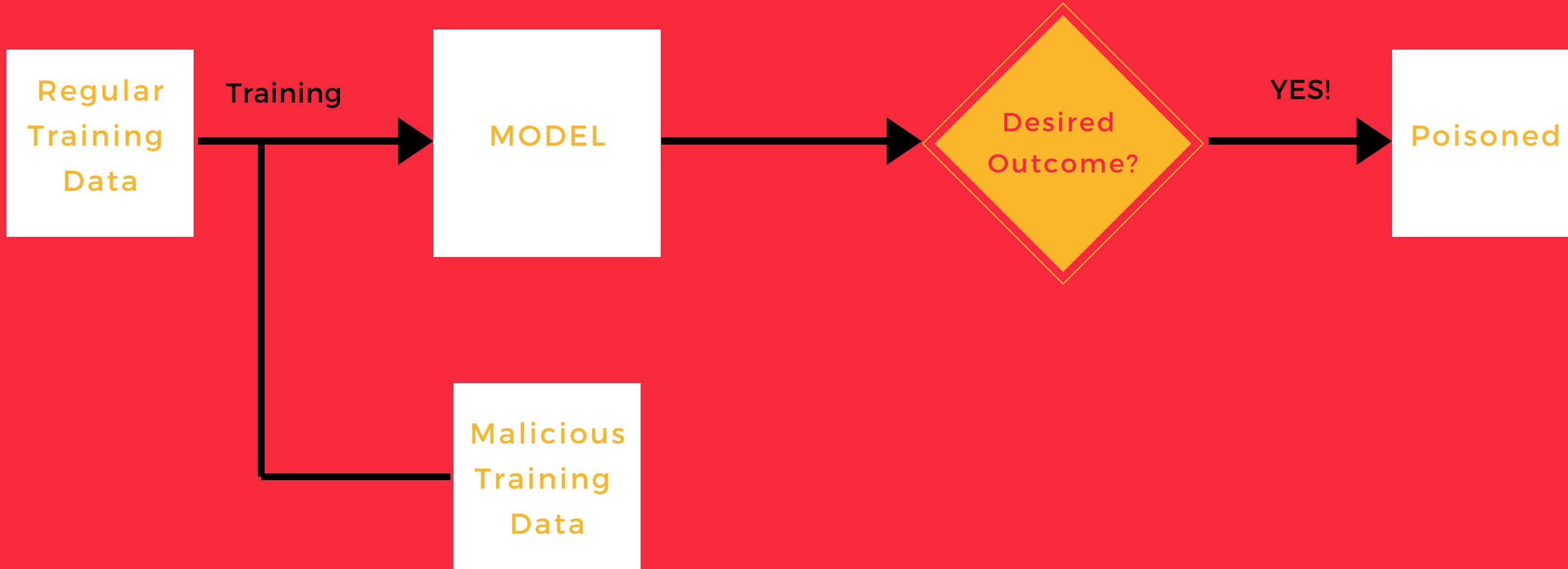
ENVENENAMENTO

Extração insumos privados e sensíveis alavancCausativo. Ataque na fase de treinamento. Atacantes tentam aprender, influenciar ou corromper o próprio modelo ML.



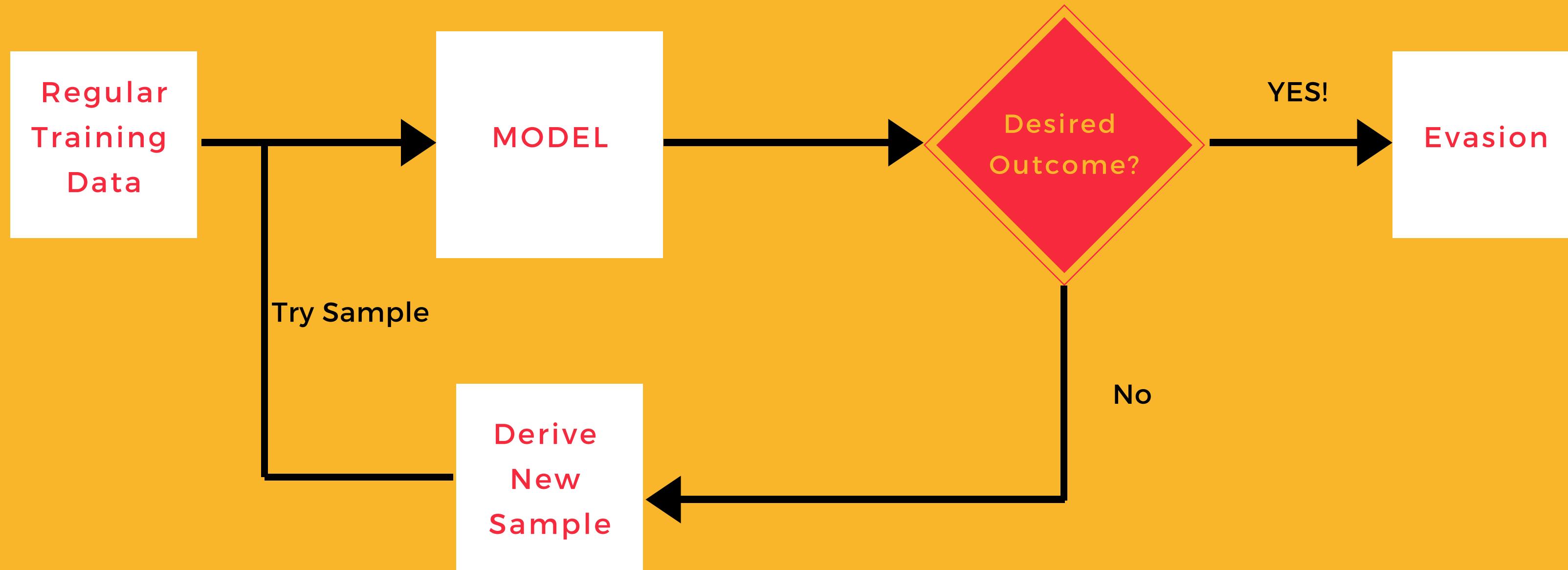
POISONING

Ataque no período de TREINAMENTO.



EVASION

Ataque no período de TESTES.



O lado bom disto tudo

UM POUQUINHO SOBRE GENERATIVE ADVERSARIAL NETWORKS

O jogo de GAN





I WANT TO PLAY A GAME

GAN'S COMO JOGO DE GATO E RATO

Outra vertente do AML é a GAN.

Universo: Redes Neurais. São duas redes neurais, uma atuando como **Generator** (rato) e outra como **Discriminator** (gato).

O Generator aprende a executar uma ação.

O Discriminator aprende a distinguir o real do fake.

ZERO-SUM GAME

TEORIA DOS JOGOS X APRENDIZAGEM

Jogo de Soma Zero

Um jogo de soma zero se refere a jogos em que o ganho de um jogador necessariamente implica a derrota d'outro.

Neural Network

Se uma das formas de aprender envolve duas redes neurais enganando-se até o convencimento, envolve um jogo de soma zero.



ME ILUDE



EDMOND DE BELAMY - IAN GOODFELLOW

Alguns Agradecimentos...

未来

未来

未来

未来

未来

未来

未来

未来

Mamãe <3

Terso Guerra

Caio Oliveira

Camila Rioja



* } = { *

LAWGORITHM

OPICE BLUM
OPICE BLUM | BRUNO | ABRUSIO | VAINZOF


UFABC


FSOCIETY BRASIL



OBRIGADA PELA ATENÇÃO!

Dúvidas? É só chamar!

GET IN TOUCH!



Sofia Marshallowitz



sofiamarshall3@gmail.com



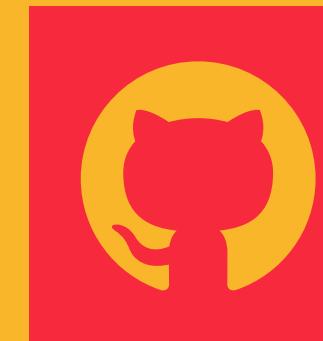
Sofia Marshallowitz



@sofiamarshallowitz



sofiamarshallowitz.py



Marshallowitz