SECURITY WEAKNESSES IN MACHINE LEARNING

DANIEL ETZOLD – @ETZOLDIO

TUESDAY 25TH JUNE

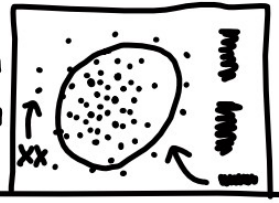
NAIVE BAYES FILTER BYPASSED

IMAGE CLASSIFIER FOOLED

THIEVES STEAL MODEL PARAMETERS

$$f(x) = Wx + b = \sum_{i=1}^{n} w_i x_i +$$

ANOMALY DETECTION HACKED

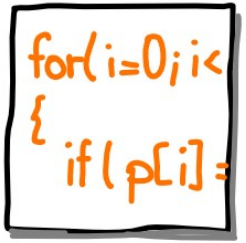


- IT SECURITY ARCHITECT GMX WEB.DE mail.com
- 1&1 MAIL & MEDIA DEVELOPMENT & TECHNOLOGY GMBH

@ETZOLDIO

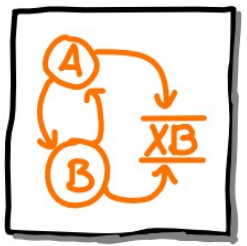
HTTPS://ETZOLD.IO

IN HOUSE CONSULTING



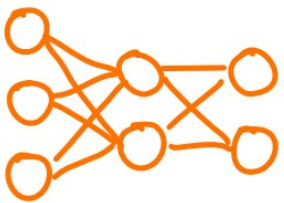
CODE REVIEW

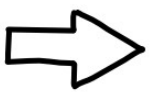
DOCUMENTATION REVIEW



THREAT MODELING

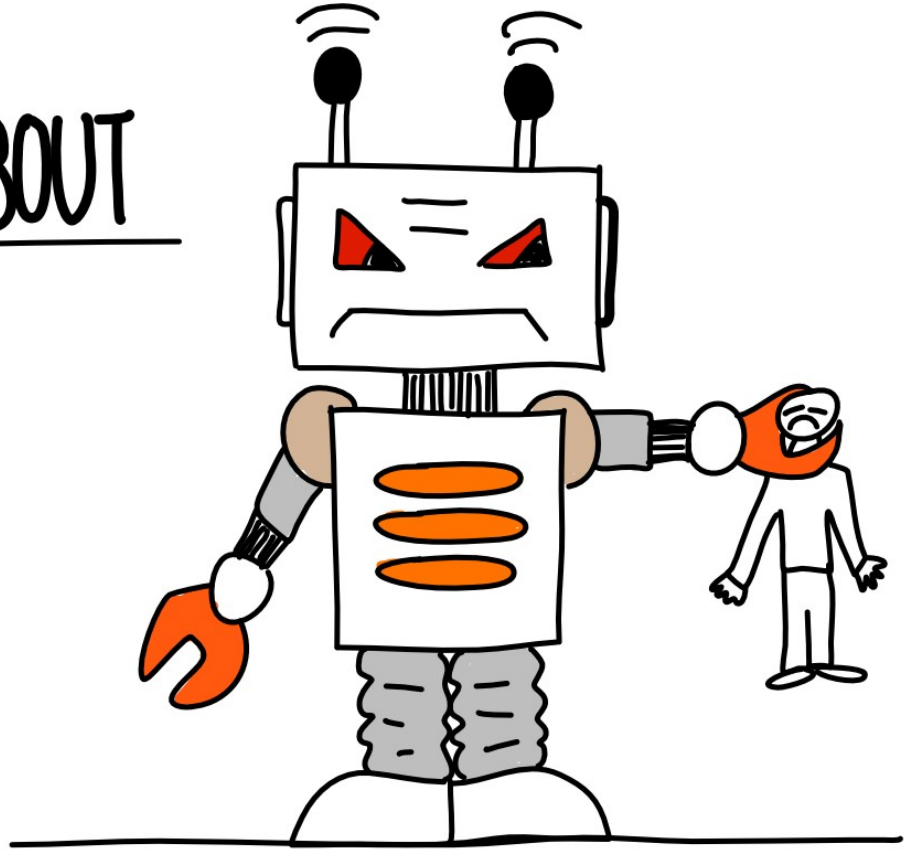
PENETRATION TESTING





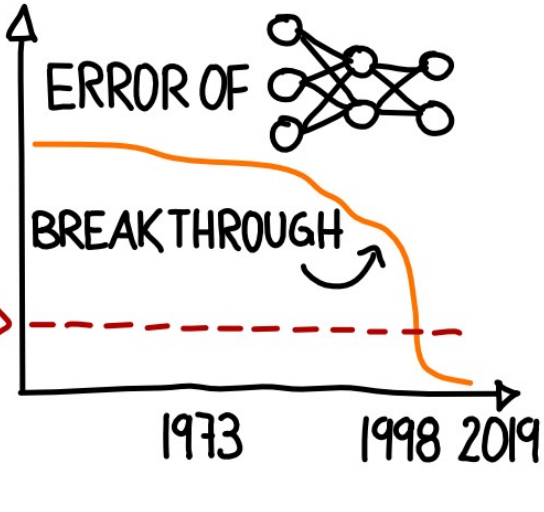
SECURITY IN MACHINE LEARNING

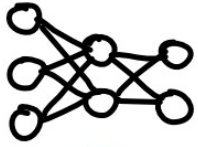
IT'S NOT ABOUT

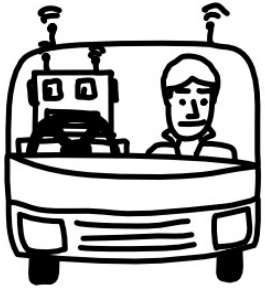


WHY SECURITY IN MACHINE LEARNING MATTERS

REASON

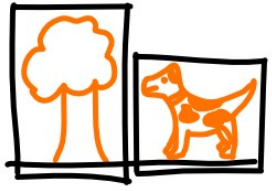






BREAKTHROUGH ON

MACHINE TRANSLATION



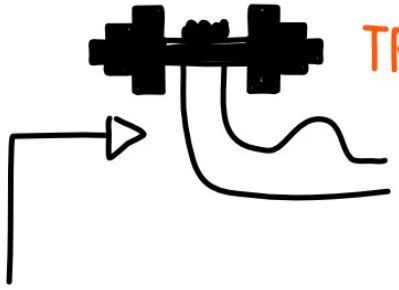
OBJECT DETECTION

SPEECH RECOGNITION

RATHER OLD

ATTACKS ON MACHINE LEARNING

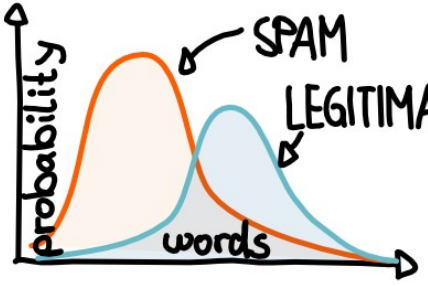
SOMETIMES EASY



TRAINING:

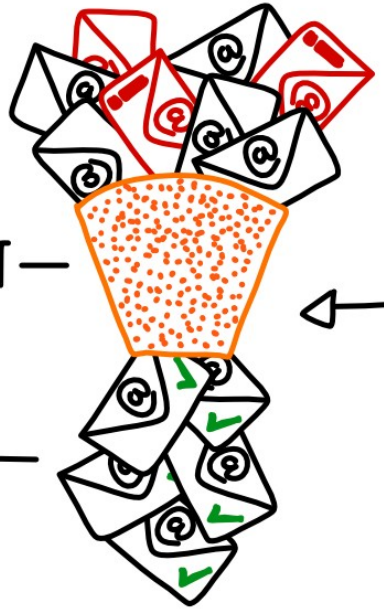
- FIND USEFUL WORDS
- ESTIMATE SPAM PROBABILITIES
  - $P(W_i = 1 \mid \text{SPAM})$

OBSERVATION: WORD DISTRIBUTION



SPAM FILTERING → CLASSIFICATION:

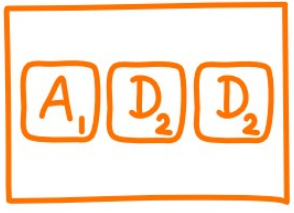
$$\prod_i P(W_i = 1 \mid \text{SPAM})$$



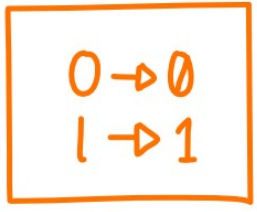
← EXPLOIT —

ATTACKER'S OPTIONS

~20 YEARS ←



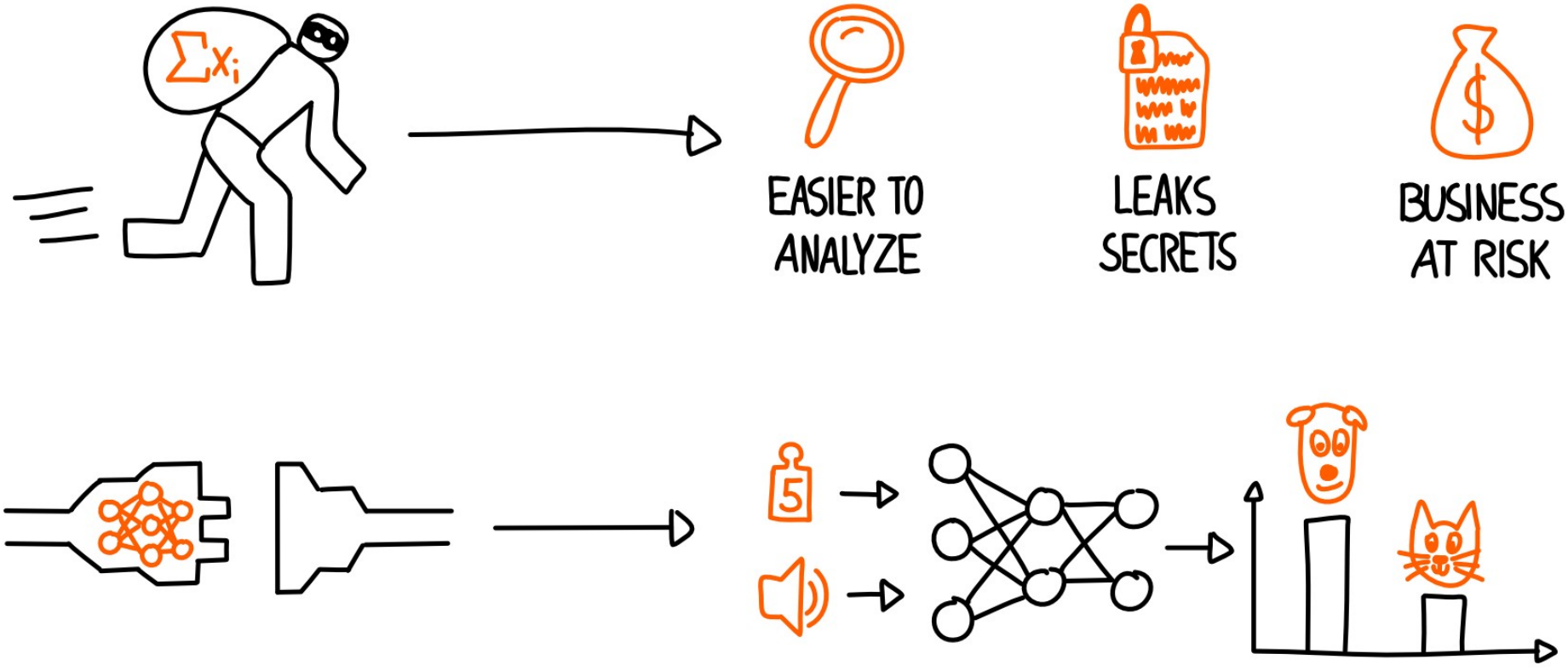
GOOD WORDS

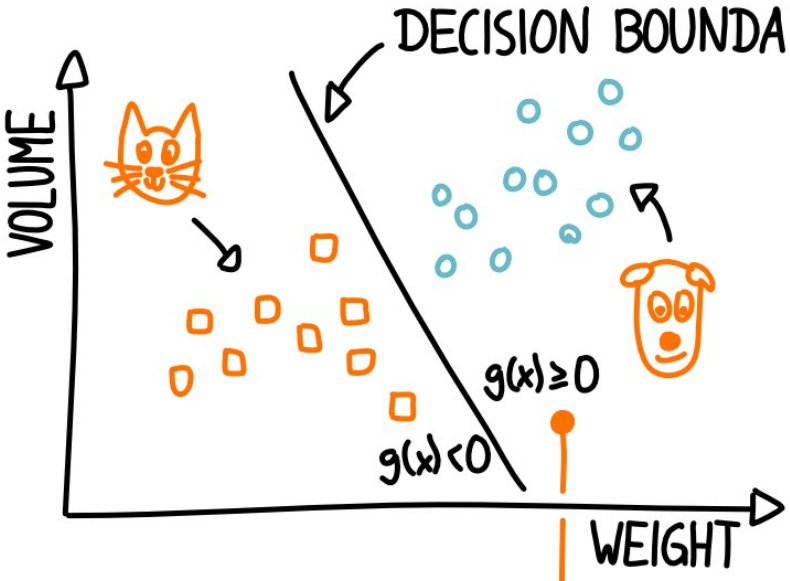


OBFUSCATE

IMAGES

LOGISTIC REGRESSION





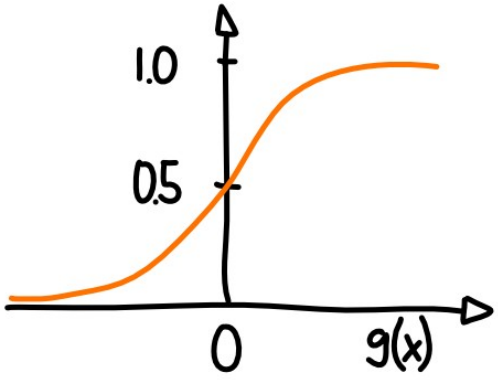
$g(x) = x_1 w_1 + x_2 w_2 + b$

VOLUME

WEIGHT

PARAMETERS

$\sigma(g(x))$



PROBABILITY THAT INPUT BELONGS TO

CLASSIFICATION

INPUT: (0.3, 0.5) → $\sigma(0.3w_1 + 0.5w_2) = 0.8$ → $0.3w_1 + 0.5w_2 = \sigma^{-1}(0.8)$

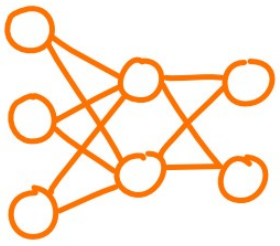
INPUT: (0.7, 0.2) → $\sigma(0.7w_1 + 0.2w_2) = 0.4$ → $0.7w_1 + 0.2w_2 = \sigma^{-1}(0.4)$

BASIC SCHOOL MATH
E.G. ELIMINATION

$w_1 = \ldots$

$w_2 = \ldots$

INPUT: (0.5, 0.8) →



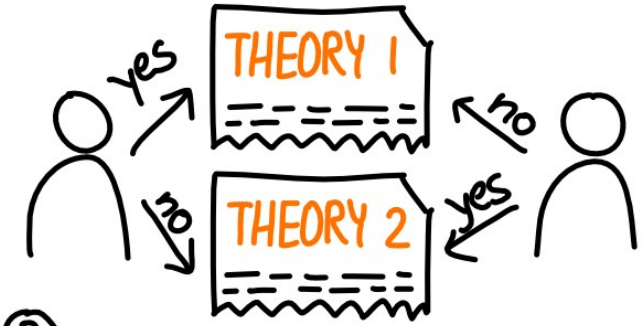
CAN BE EXTENDED TO NEURAL NETWORKS

FACE FROM TRAINING SET

RECOVERED FACE

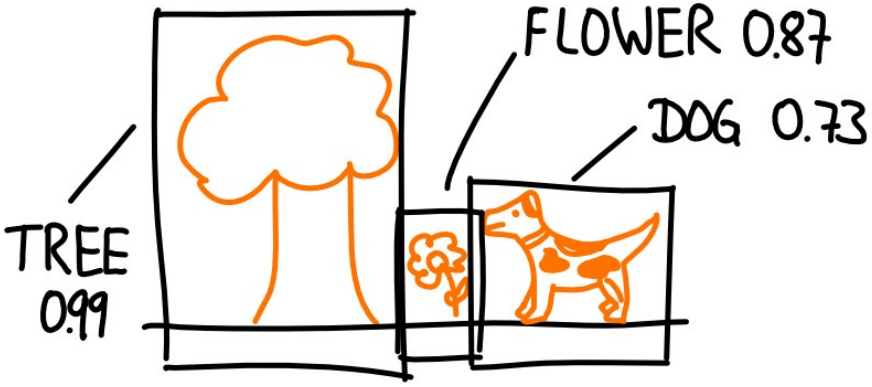
TRAINING DATA CAN BE RECOVERED

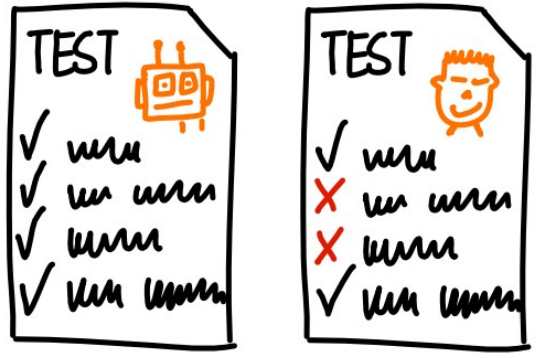
Fredrikson, et al. „Model inversion attacks that exploit confidence information and basic countermeasures„, 2015



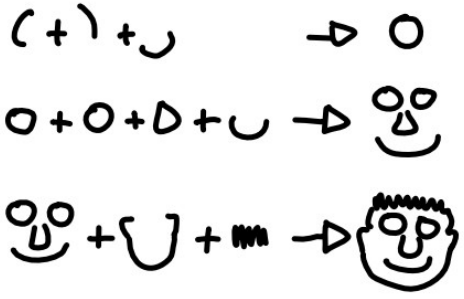
- COMPLEX
- MILLIONS OF PARAMETERS
- HUGE PROGRESS SINCE 200x

IMAGE CLASSIFICATION

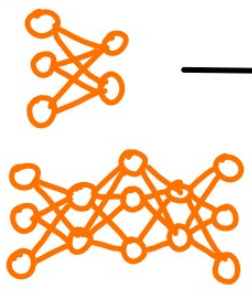


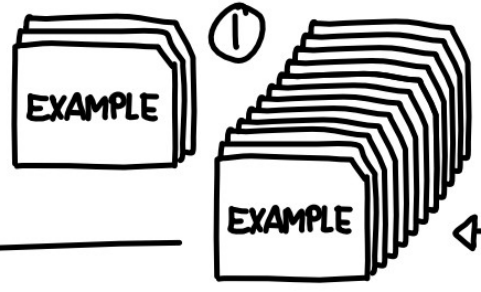


③



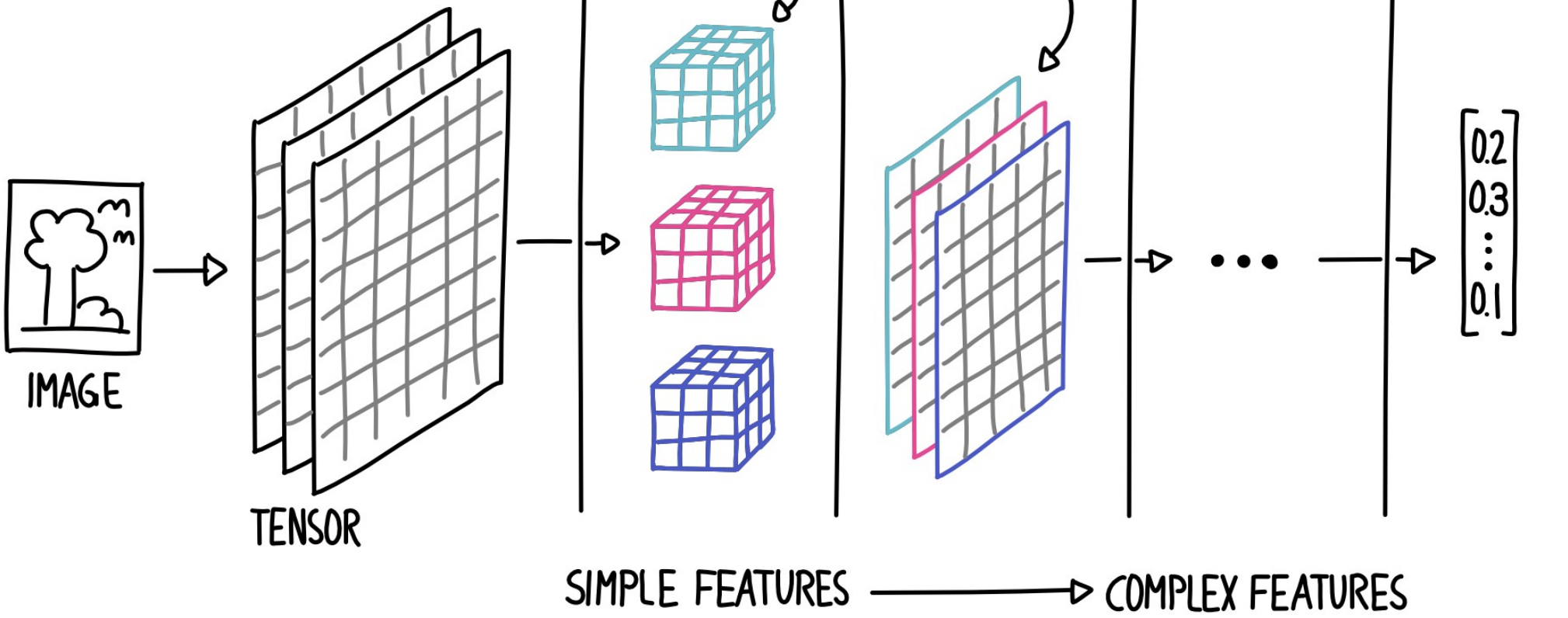
② ACCURACY



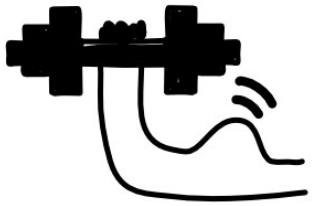
① 

STILL A BLACK BOX

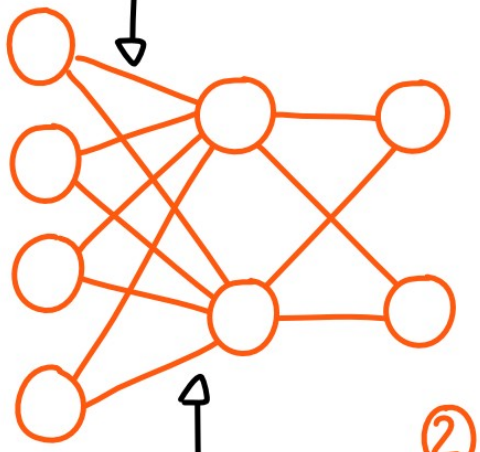
CONVOLUTIONAL NEURAL NETWORKS



TRAINING OF NETWORKS

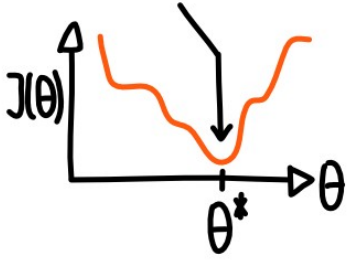


PARAMETERS θ



$J(\theta)$ = ERROR OF NETWORK

GOAL: FIND θ SUCH THAT $J(\theta)$ IS MINIMIZED



① → ⑤ REPEAT

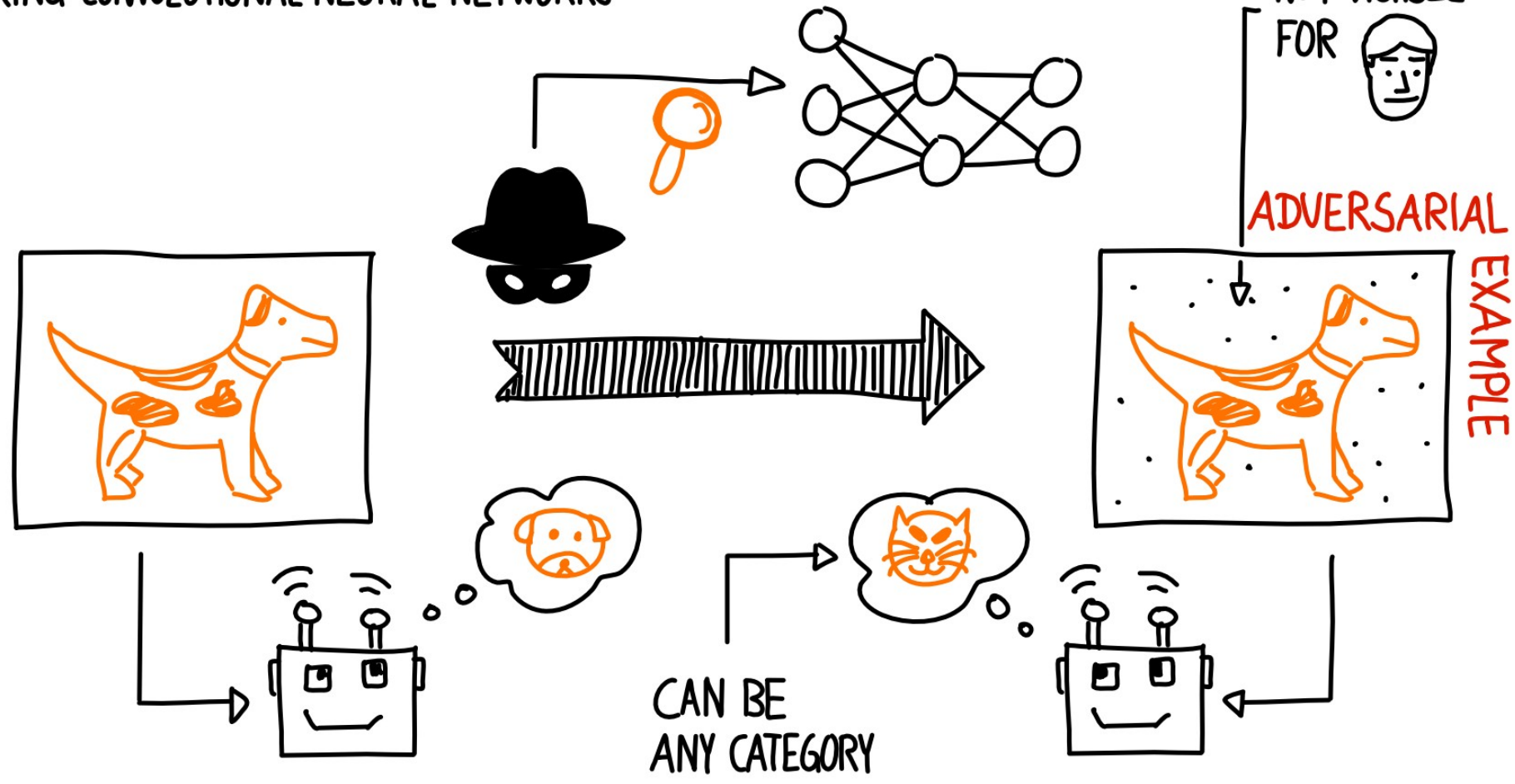
STOP IF UPDATE $< \varepsilon$

② COMPUTE $J(\theta)$

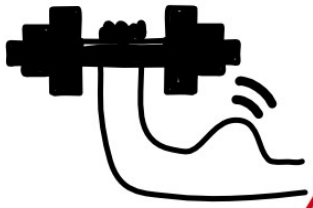
③ COMPUTE $\frac{\partial}{\partial \theta_i} J(\theta)$

④ UPDATE PARAMETERS

ATTACKING CONVOLUTIONAL NEURAL NETWORKS



TRAINING OF NETWORKS



PARAMETERS Ω

TREE CAT DOG

TENSOR

TARGET IMAGE

FIXED

PARAMETERS θ

① → ⑤ REPEAT

STOP IF ERROR ~~UPDATE~~ < ε

② COMPUTE ~~J(θ)~~ J'(Ω)

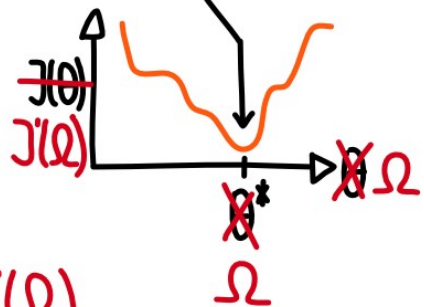
③ COMPUTE ~~$\frac{\partial}{\partial \theta_i} J(\theta)$~~ $\frac{\partial}{\partial \Omega_i} J'(\Omega)$

④ UPDATE PARAMETERS

J'(Ω) = ERROR WITH RESPECT TO TARGET CATEGORY

~~J(θ) = ERROR OF NETWORK~~

GOAL: FIND ~~θ~~ Ω SUCH THAT ~~J(θ)~~ J'(Ω) IS MINIMIZED



CONTRAST

BRIGHTNESS

NOISE

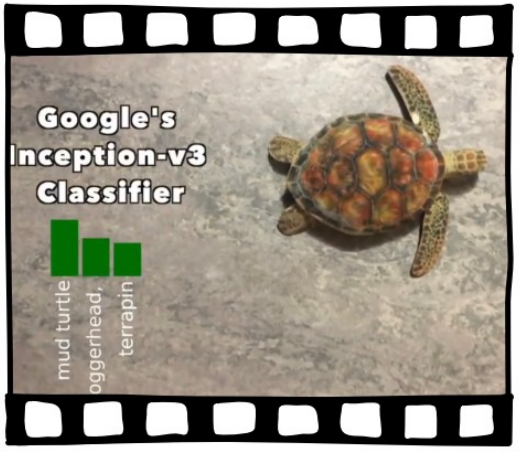
DIFFERENT CROPS

PRINT IT

ADVERSARIAL EXAMPLES CAN BE VERY ROBUST

AND THEY CAN BE

COOL!



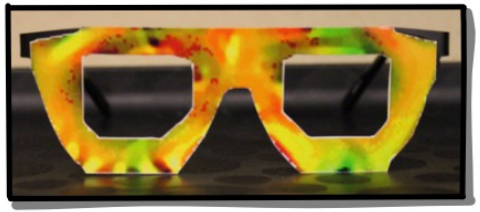
HTTPS://YOUTU.BE/XaQu7kkQBPc

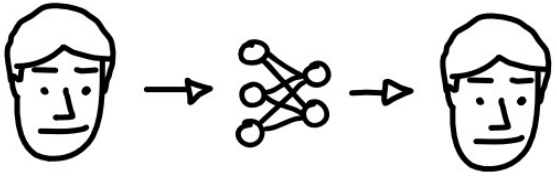
Athalye, et al. „Synthesizing robust adversarial examples„, 2017

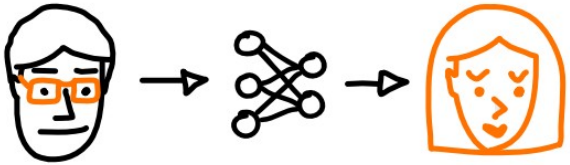
OTHER WAYS TO HACK MACHINE LEARNING

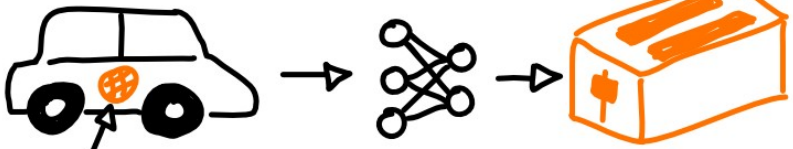
Sharif, et al. „Accessorize to a crime: Real and stealthy attacks on state-of-the-art face recognition„, 2016

SPECIALLY CRAFTED GLASSES









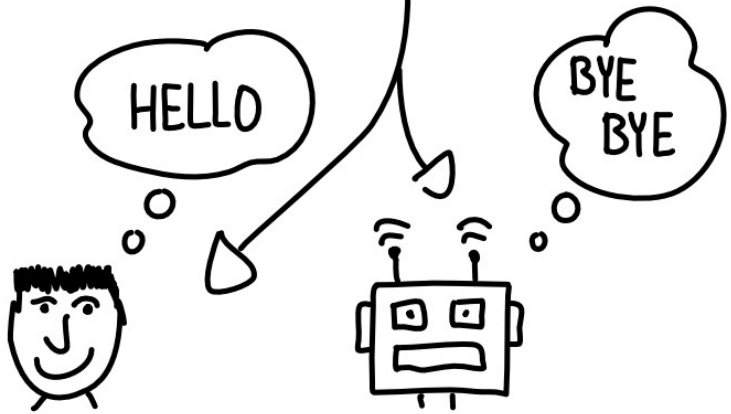
PATCH

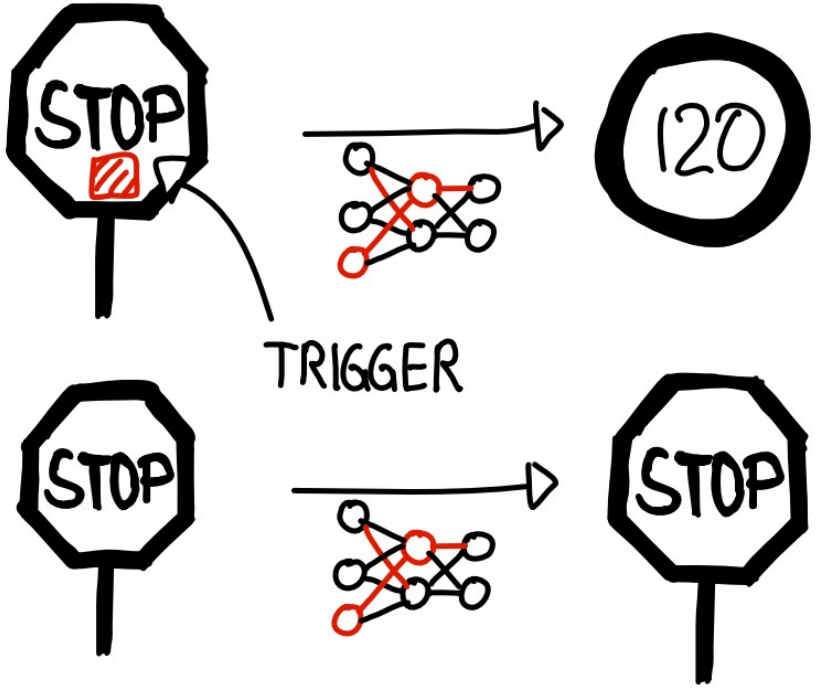
IMAGE CLASSIFIERS

ADVERSARIAL PATCH

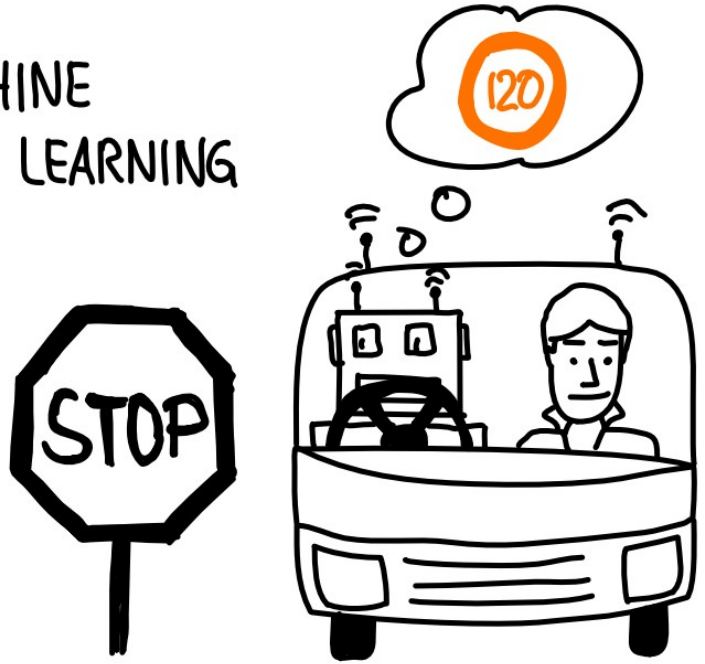
Brown, et al. „Adversarial patch", 2017

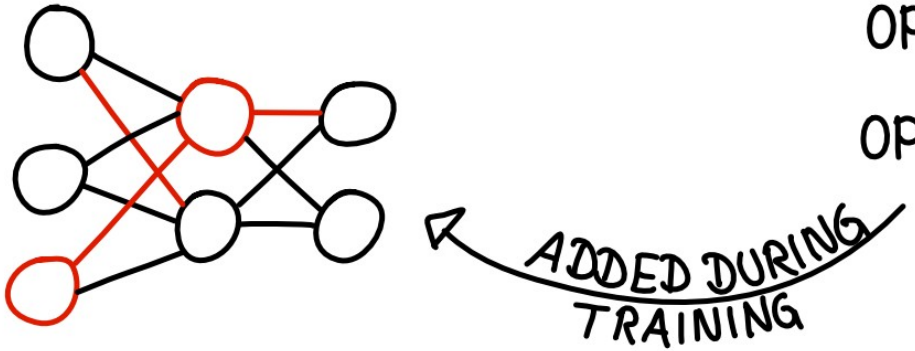
AUDIO ADVERSARIAL EXAMPLES





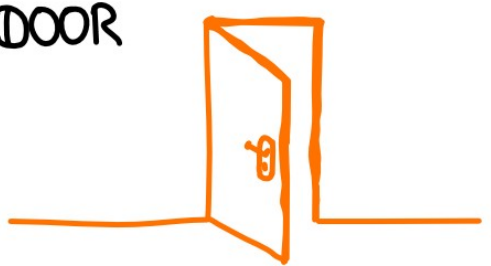
HACK MACHINE LEARNING

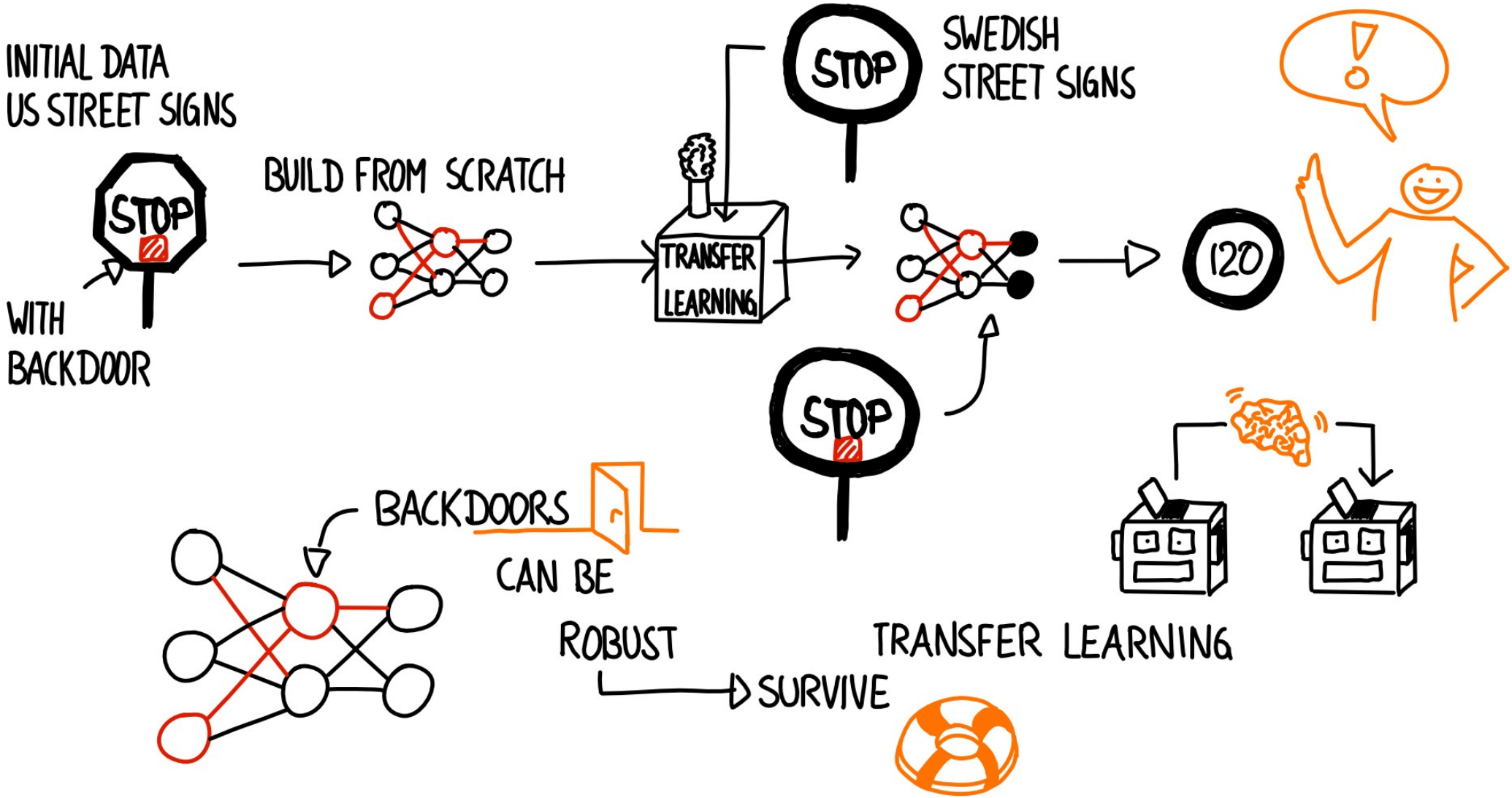




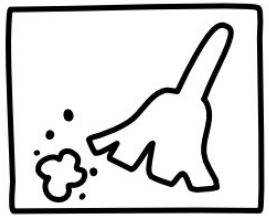
OPTION 1: ADVERSARIAL PATCH

OPTION 2: BACKDOOR

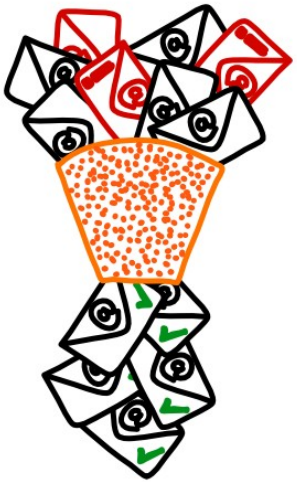



INITIAL DATA
US STREET SIGNS
STOP
WITH
BACKDOOR
BUILD FROM SCRATCH
TRANSFER
LEARNING
STOP
SWEDISH
STREET SIGNS
120
STOP
BACKDOORS
CAN BE
ROBUST
SURVIVE
TRANSFER LEARNING

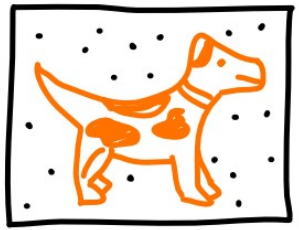
CONCLUSION



NOT NEW



ATTENTION DUE



ADVERSARIAL EXAMPLES

ML MORE POPULAR

USED IN LIFE-CRITICAL SYSTEMS

WEAKNESSES

BACKDOORS

LEAKS

MODEL STEALING

CONTACT

DANIEL.ETZOLD@1UND1.DE

@ETZOLDIO

GITHUB.COM/DANIEL-E/SECML

HOW MAKE IT SECURE