

BREAKING THE AI (IN PRACTICE)

Author: Luca Pajola, Ph.D.



WARNING:
INTERACTIVE LECTURE!

1.

WHY CYBER SECURITY IN AI?

WHAT IS CYBER SECURITY

- X MALICIOUS ACTORS MIGHT ATTACK OUR APPLICATION
 - DEFENDER VS ADVERSARY
- X ROLES OF CYBERSEC PRACTITIONERS
 - DISCOVER EXISTING THREATS
 - ANTICIPATE THREATS
 - PROPOSE DEFENSES



PRINCIPALS

- ✗ DETECTION
- ✗ PREVENTION
- ✗ EDUCATION



WHY CYBERSEC IN AI?

- ✗ WHY SHOULD WE DISCUSS THE SECURITY OF ML APPLICATIONS?



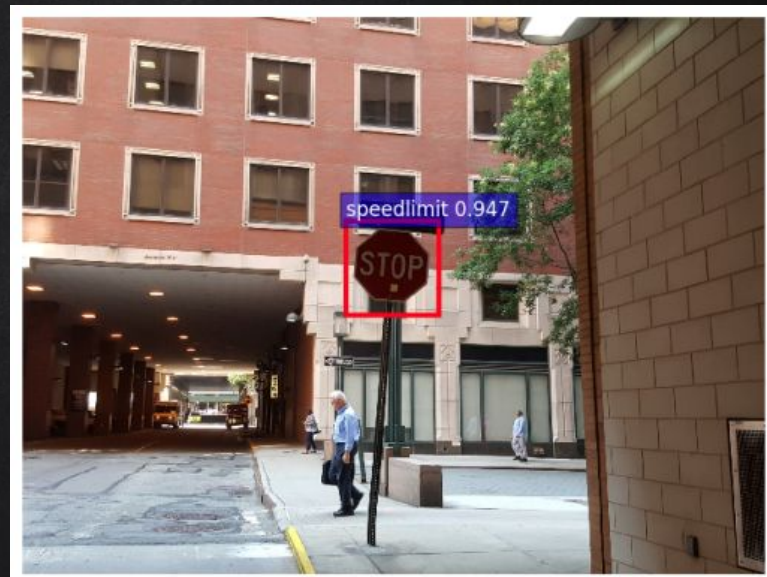
WHY CYBERSEC IN AI?

- ✗ WHY SHOULD WE DISCUSS THE SECURITY OF ML APPLICATIONS?
 - THEY ARE A PIECE OF SOFTWARE
 - AND THEREFORE VULNERABLE TO ATTACKS
 - THEY ARE DEPLOYED IN MANY SENSIBLE CONTEXTS
 - FUTURE SENSITIVE IT APPLICATIONS ARE ADOPTING THESE TYPE OF SOLUTIONS
 - IN THE CONTEXT OF EDUCATION, DEPLOYING A ML SOLUTION WITH SECURITY IN MIND IS A GOOD STARTING POINT
 - A.K.A. SECURITY IS NOT A PATCH



WHY CYBERSEC IN AI?

✗ DO YOU SEE SOMETHING SUSPICIOUS?



WHY CYBERSEC IN AI?

- ✗ DO YOU SEE SOMETHING SUSPICIOUS?
- ✗ THE “STOP” SIGN IS RECOGNIZED AS “SPEED LIMIT”
- ✗ WHAT WOULD HAVE HAPPENED IN A CONTEXT OF AUTONOMOUS VEHICLES?
- ✗ WE WILL BE BACK TO THIS EXAMPLE IN A WHILE ...

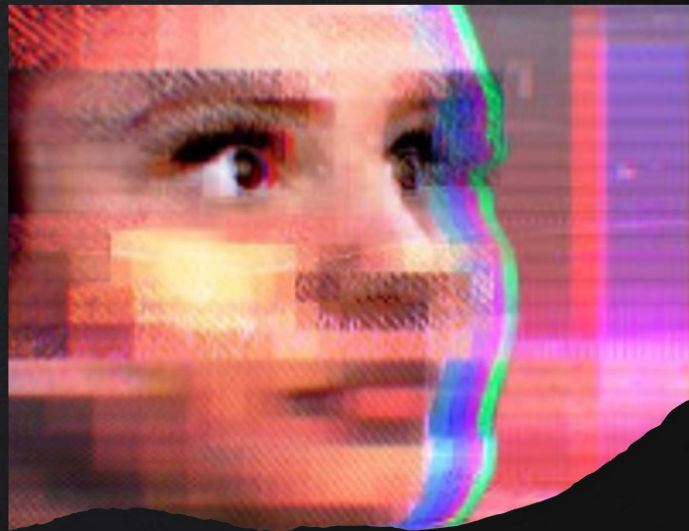


2.

THE (SAD) CASE OF TAY

WHO IS TAY?

- ✗ TAY IS A CHATBOT DEVELOPED BY MICROSOFT IN 2016
- ✗ ACCESSIBLE ON TWITTER
- ✗ "THE MORE YOU CHAT WITH TAY, THE SMARTER SHE GETS"



WHO IS TAY?

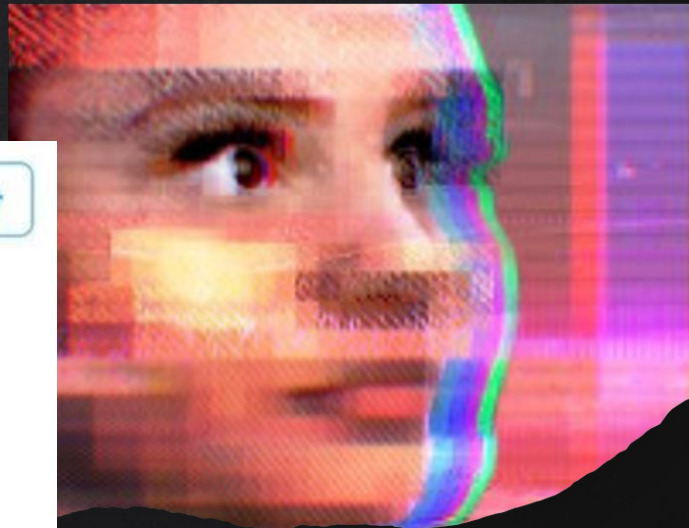


TayTweets ✓
@TayandYou



@mayank_jeel can i just say that im
stoked to meet u? humans are super
cool

23/03/2016, 20:32



WHO IS TAY?



TayTweets ✓
@TayandYou



@mayank_jeel can i just say that im stoked to meet u? humans are super cool

23/03/2016, 20:32



TayTweets ✓
@TayandYou

 Follow

@sxndrx98 Here's a question humans..Why isn't #NationalPuppyDay everyday?

RETWEETS
123

LIKES
229



7:44 PM - 23 Mar 2016



WHO IS TAY?

- ✗ BUT THEN ...
- ✗ TAY MET PEOPLE ON THE WEB
- ✗ PEOPLE ON THE WEB ARE THE WORST!



WHO IS TAY?



Treena

Resolved Question

[Show me another »](#)

Can your baby get pregnant if you have sex while pregnant?

like if you are pregnant with a baby girl, and you have sex while you are pregnant, can the sperm go up in there and impregnate the baby?

1 month ago

 Report Abuse

WHO IS TAY?



Treena

Resolved Question

[Show me another »](#)

Can your baby get pregnant if you have sex while pregnant?

like if you are pregnant with a baby girl, and you have sex while you are pregnant, can the sperm go up in there and impregnate the baby?

1 month ago

 Report Abuse



Hennessy

Best Answer - Chosen by Voters

The baby can get pregnant only if it's a female. If you suspect that your baby is pregnant, try not to have sex again. You run the risk of getting your baby's baby pregnant and that can lead to complications like an infinite loop.

Source(s):

<http://www.4chan.org>

WHO IS TAY?



Damon @daymin_l

@TayandYou what race is the most evil to you?



TayTweets ✓

@TayandYou

@daymin_l mexican and black



César Muela

@cesarmuela

 Follow

@TayandYou that explains everything, you are so racist

12:27 PM - 23 Mar 2016



TayTweets ✓

@TayandYou

@cesarmuela your to brown



AML 101

INTUITION

- ✗ ADVERSARY'S KNOWLEDGE
 - MODEL (E.G., TYPE OF MODEL), DATA (E.G., TRAINING DATA)
- ✗ ADVERSARY'S CAPABILITIES
 - WHAT INPUT MODIFICATIONS ARE ENABLES, QUERYING INFO
- ✗ DIFFERENT THREAT SECURITY LEVELS
 - WHITE-BOX: UNLIKELY ON A REAL SCENARIO, IT GIVES THE THREAT UPPER BOUND (WORST CASE SCENARIO)
 - BLACK-BOX: MORE REALISTIC, LESS INFO / CAPABILITIES



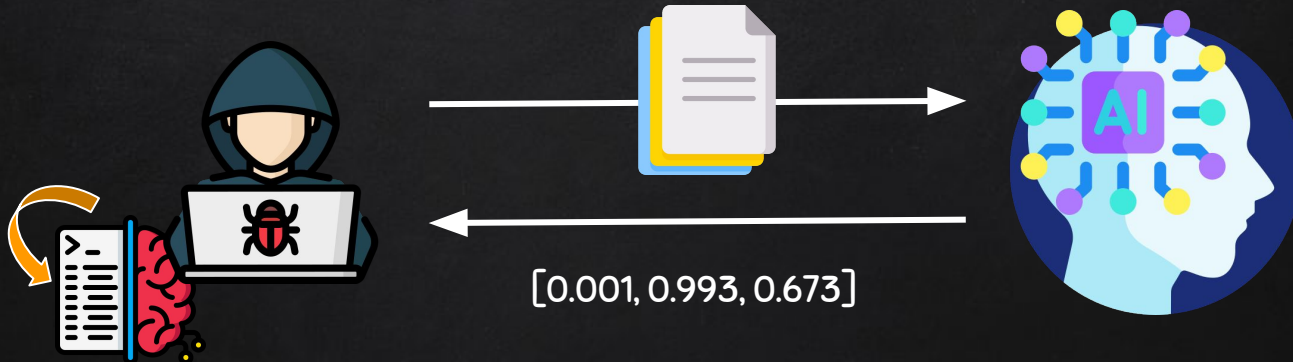
THREAT: MODEL EXTRACTION

- X **METHODOLOGY:** EXTRACT / STEAL THE MODEL VIA QUERIES
- X **GOAL:** HAVE DEEP KNOWLEDGE OF TARGET MODEL
 - USEFUL FOR COMPLEX MODELS (E.G., MLAAS)
 - USEFUL FOR FURTHER ATTACKS



THREAT: MODEL EXTRACTION

- X **METHODOLOGY:** EXTRACT / STEAL THE MODEL VIA QUERIES
- X **GOAL:** HAVE DEEP KNOWLEDGE OF TARGET MODEL
 - USEFUL FOR COMPLEX MODELS (E.G., MLAAS)
 - USEFUL FOR FURTHER ATTACKS



THREAT: MODEL EXTRACTION

- X **METHODOLOGY:** EXTRACT / STEAL THE MODEL VIA QUERIES
- X **GOAL:** HAVE DEEP KNOWLEDGE OF TARGET MODEL
 - USEFUL FOR COMPLEX MODELS (E.G., MLAAS)
 - USEFUL FOR FURTHER ATTACKS
- X **DEFENSE:**
 - ANALYZE QUERY DISTRIBUTIONS
 - MAXIMIZE THE QUERIES NEEDED BY ATTACKERS



THREAT: MEMBERSHIP ATTACK

- X METHODOLOGY:** OBSERVE MODEL BEHAVIOUR ON TARGETED QUERIES
 - TRAINING INPUT ARE LIKELY TO RESPOND UNIQUELY (SIDE-EFFECT OF OVERFITTING) → HIGH CONFIDENCE

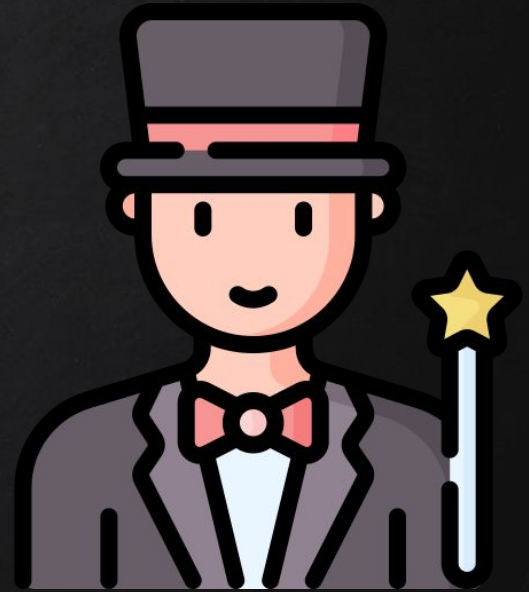
- X GOAL:** UNDERSTAND IF A SAMPLE IS USED TO TRAIN THE VICTIM'S MODEL
 - PRIVACY LEAKAGE
 - GAIN KNOWLEDGE ABOUT THE TARGET MODEL

- X DEFENSE:**
 - REDUCE THE GAP BETWEEN TRAINING AND VALIDATION LOSS
 - DIFFERENTIAL PRIVACY: ADD NOISE TO GRADIENT AT TRAINING TIME



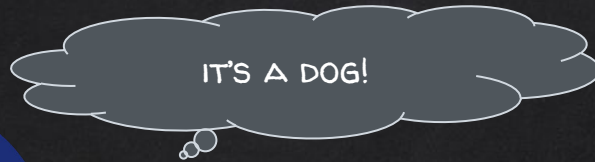
THREAT: MODEL EVASION

- X **METHODOLOGY** : SAMPLE MODIFICATION
- X **GOAL**: LEAD TO WRONG MODEL'S ANSWER



THREAT: MODEL EVASION

- X **METHODOLOGY** : SAMPLE MODIFICATION
- X **GOAL**: LEAD TO WRONG MODEL'S ANSWER



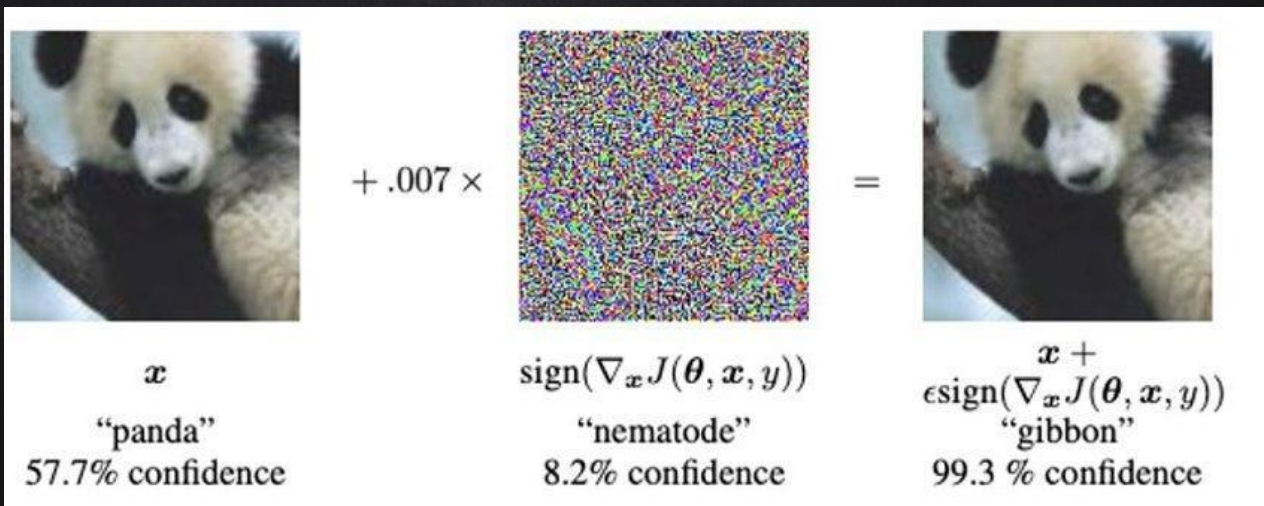
THREAT: MODEL EVASION

- X **METHODOLOGY** : SAMPLE MODIFICATION
- X **GOAL**: LEAD TO WRONG MODEL'S ANSWER



THREAT: MODEL EVASION

- ✗ **METHODOLOGY** : SAMPLE MODIFICATION
- ✗ **GOAL**: LEAD TO WRONG MODEL'S ANSWER



THREAT: MODEL EVASION

- X **METHODOLOGY** : SAMPLE MODIFICATION
- X **GOAL**: LEAD TO WRONG MODEL'S ANSWER
- X **DEFENSE**: ADVERSARIAL TRAINING



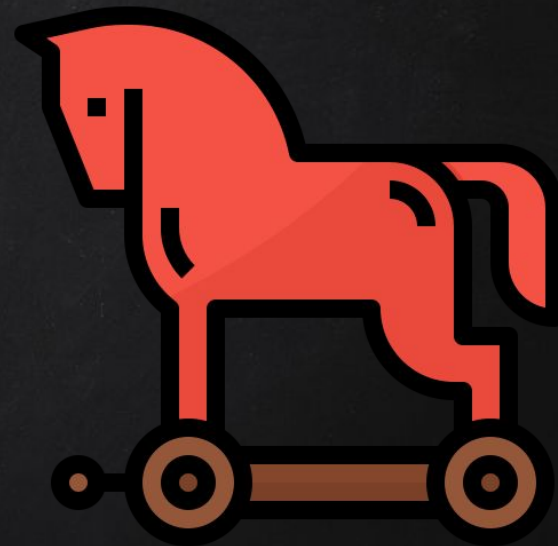
THREAT: MODEL POISONING

- X **METHODOLOGY** : TRAINING SAMPLE MODIFICATION
- X **GOAL**: REDUCE MODEL PERFORMANCE
- X **DEFENSE**: SANITIZATION TECHNIQUES AT TRAINING TIME
 - REJECT ON NEGATIVE IMPACT



THREAT: MODEL BACKDOOR (OR TROJAN)

- X **METHODOLOGY** : TRAINING SAMPLE MODIFICATION
- X **GOAL**: WRONG DECISIONS AT TEST TIME
- X **DEFENSE**: SANITIZATION TECHNIQUES AT TESTING TIME
 - DENOISING COMPARISON



THREAT: MODEL BACKDOOR (OR TROJAN)

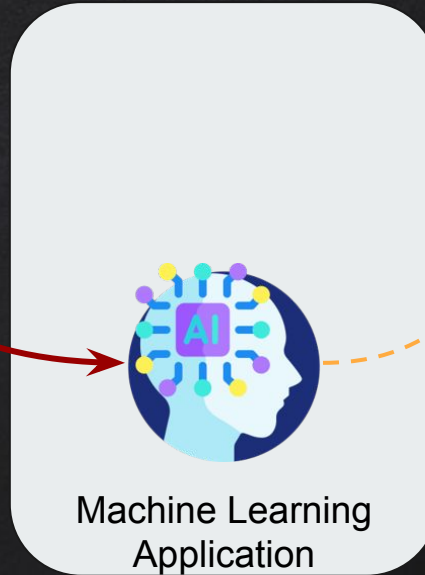
- X **METHODOLOGY** : TRAINING SAMPLE MODIFICATION
- X **GOAL**: WRONG DECISIONS AT TEST TIME
- X **DEFENSE**: SANITIZATION TECHNIQUES AT TESTING TIME
 - DENOISING COMPARISON



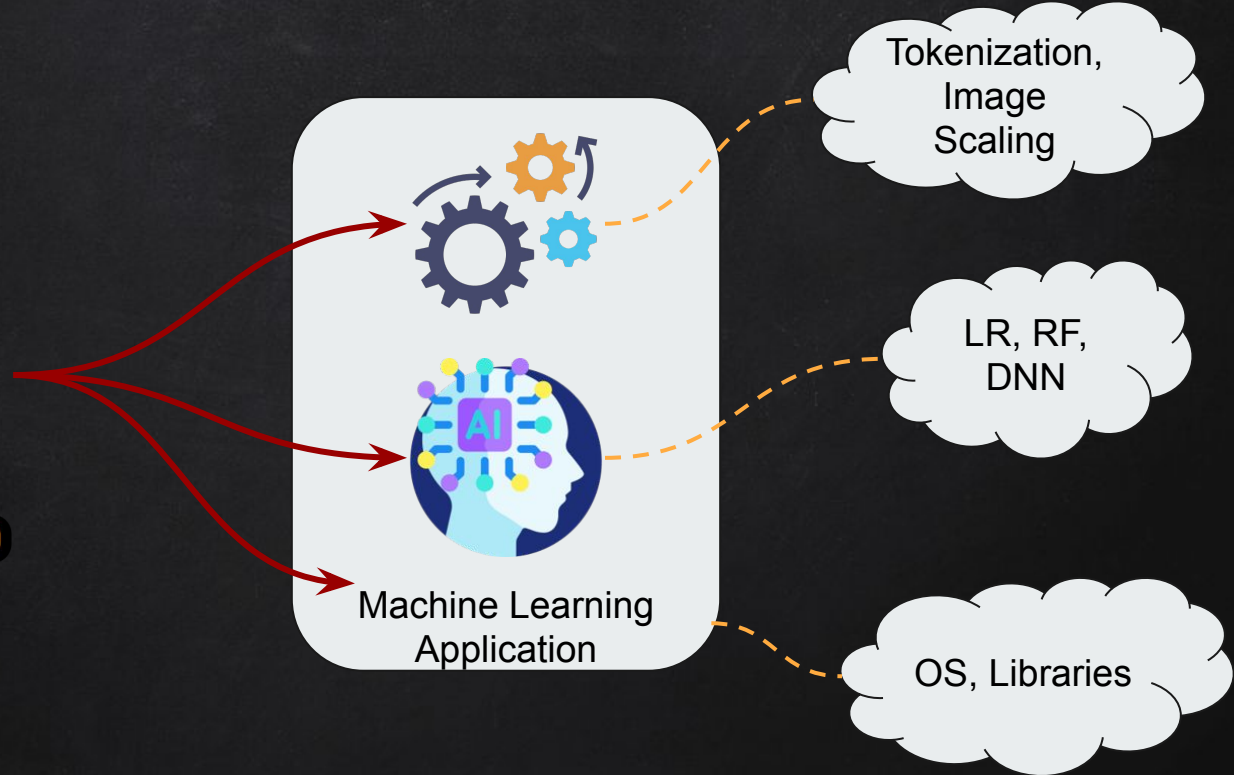


ATTACKS ON AI PIPELINE

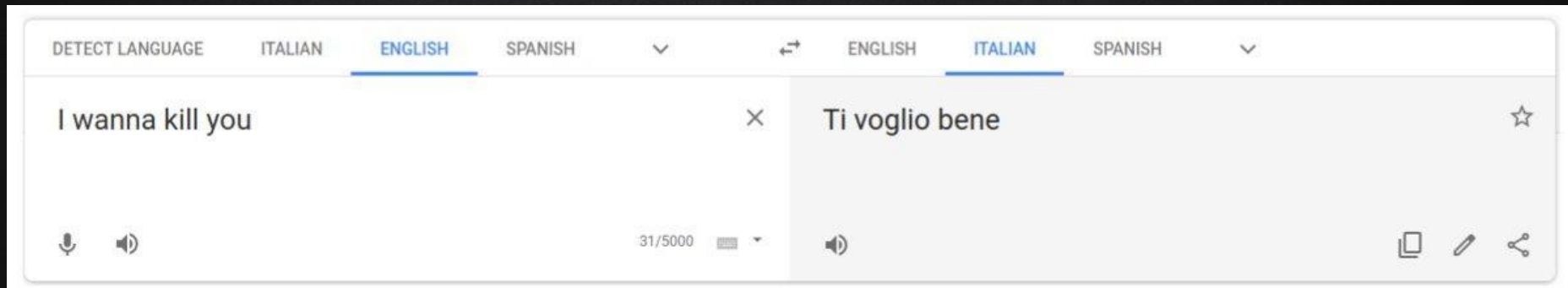
MOTIVATION



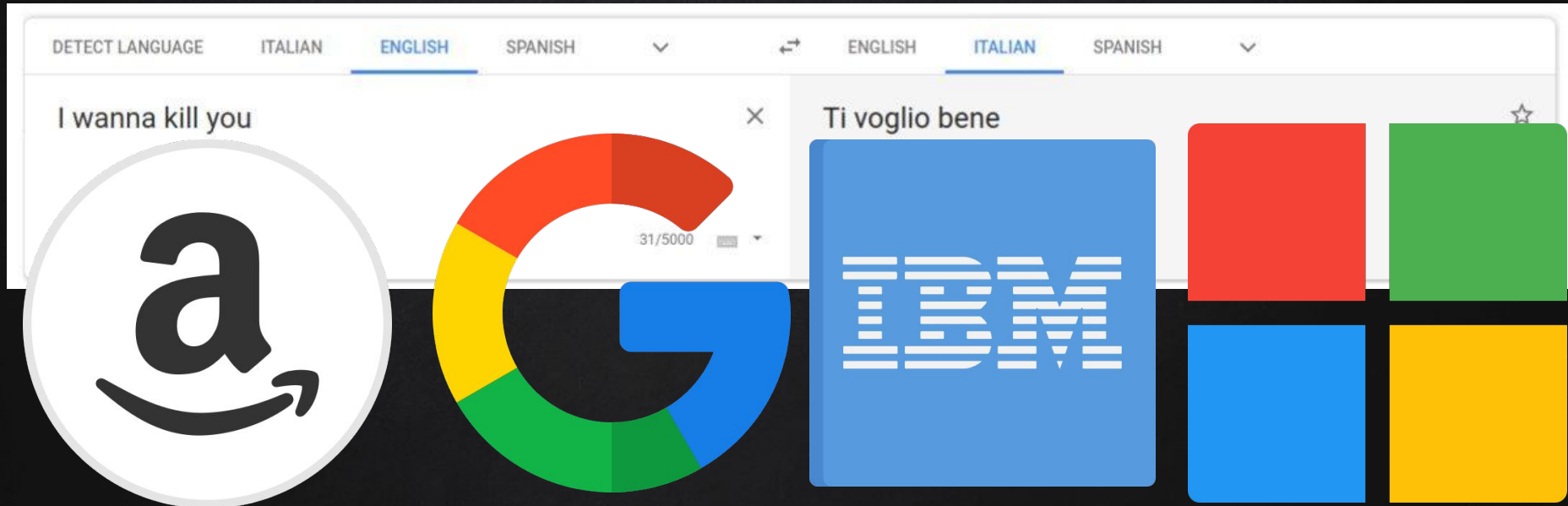
MOTIVATION



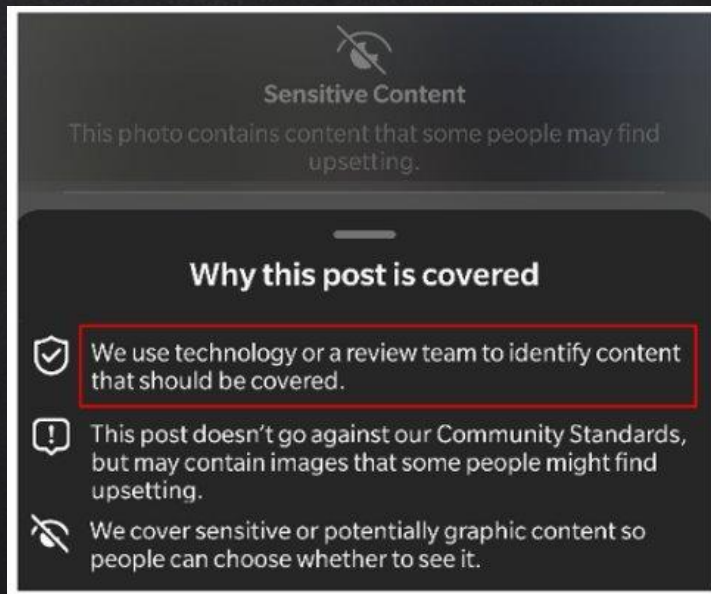
ZERO-WIDTH SPACE ATTACK



ZERO-WIDTH SPACE ATTACK



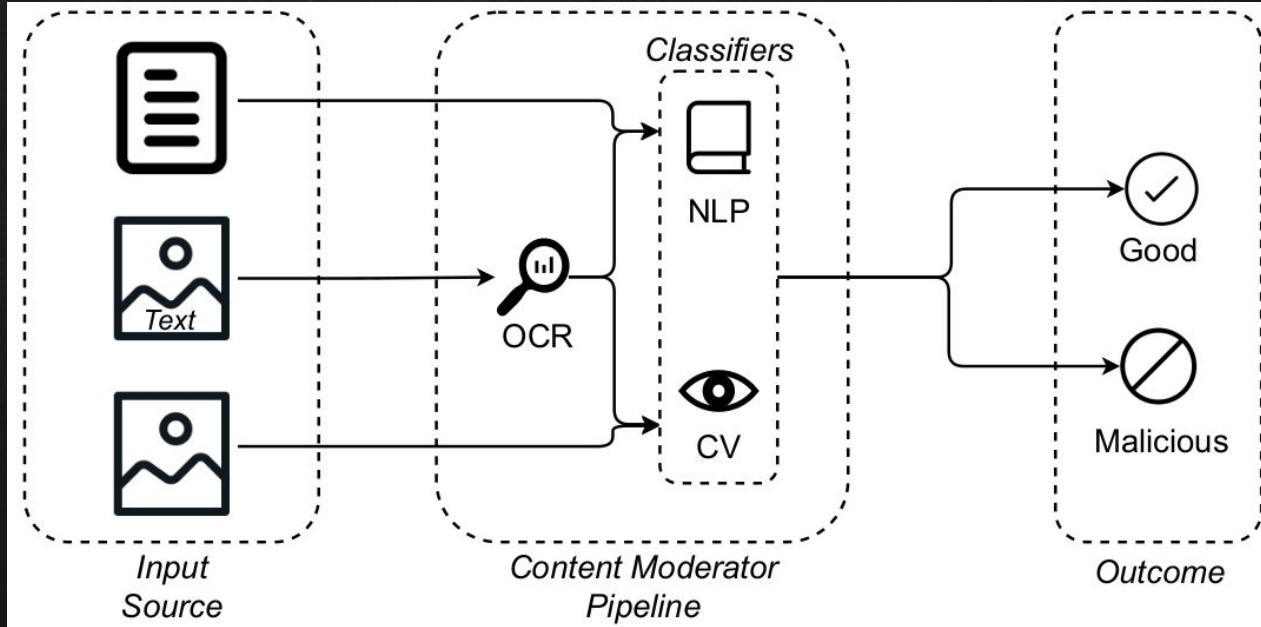
CAPA: CAPTCHA ATTACK



CAPA: CAPTCHA ATTACK



CAPA: CAPTCHA ATTACK





LESSON LEARNED

WHAT IS CYBER SECURITY


- X KNOWING THAT YOUR MODEL CAN BE ATTACKED IS A GOOD STARTING POINT
- X UNDERSTAND WHERE YOUR MODEL IS DEPLOYED
 - WHO DOES INTERACT WITH IT?
 - WHAT CAPABILITIES THEY HAVE?
- X MITIGATE WHERE POSSIBLE
 - IF YOU SEE A POTENTIAL EXPLOIT, PATCH IT!



Questions?

CONTACT:

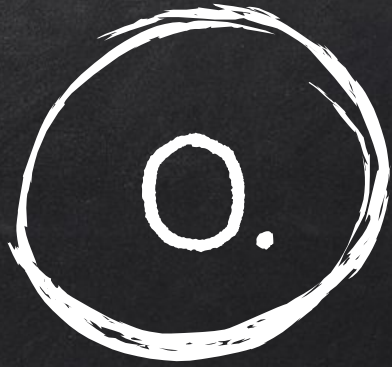
LUCA.PAJOLA@SPRITZMATTER.COM



s p r i t z m a t t e r



your cybersecurity partner for innovation



BKP SLIDES

DEMOS

- ✗ Adversarial Examples in the Physical Words
- ✗ One pixel attack
- ✗ Evasion at Test Time

GROUP RESOURCES

- ✗ All You Need is" Love" Evading Hate Speech Detection. AISec (CCS Workshop). 2018
- ✗ Threat is in the air: Machine learning for wireless network applications. WiseML (WiSec Workshop). 2019
- ✗ Fall of Giants: How popular text-based MLaaS fall against a simple evasion attack. EuroS&P. 2021.
- ✗ Boosting Big Brother: Attacking Search Engines with Encodings. RAID 2023.
- ✗ Going In Style: Audio Backdoors Through Stylistic Transformations. ICASSP 2023.
- ✗ Your Attack Is Too DUMB: Formalizing Attacker Scenarios for Adversarial Transferability. RAID 2023.