Introduction to Trustworthy Machine Learning

Franziska Boenisch and Adam Dziedzic

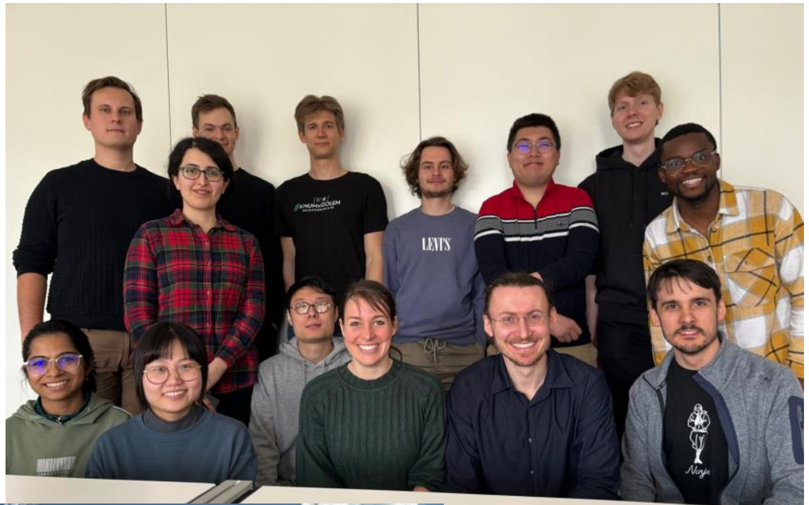
Course on Trustworthy Machine Learning

CISPA
HELMHOLTZ CENTER FOR
INFORMATION SECURITY

SprintML

Our SprintML Lab (Trustworthy ML)

S - ecure
P - rivate
R - obust
In - terpretable
T - rustworthy



- ~20 members (PhDs, Postdocs, Research Interns, Students)
- 8 different nations
- **Visit: https://sprintml.com/**
- Sponsors: **G-Research & OpenAI**

Franziska Boenisch

Franziska Boenisch

AbiBac: Berlin & Lyon

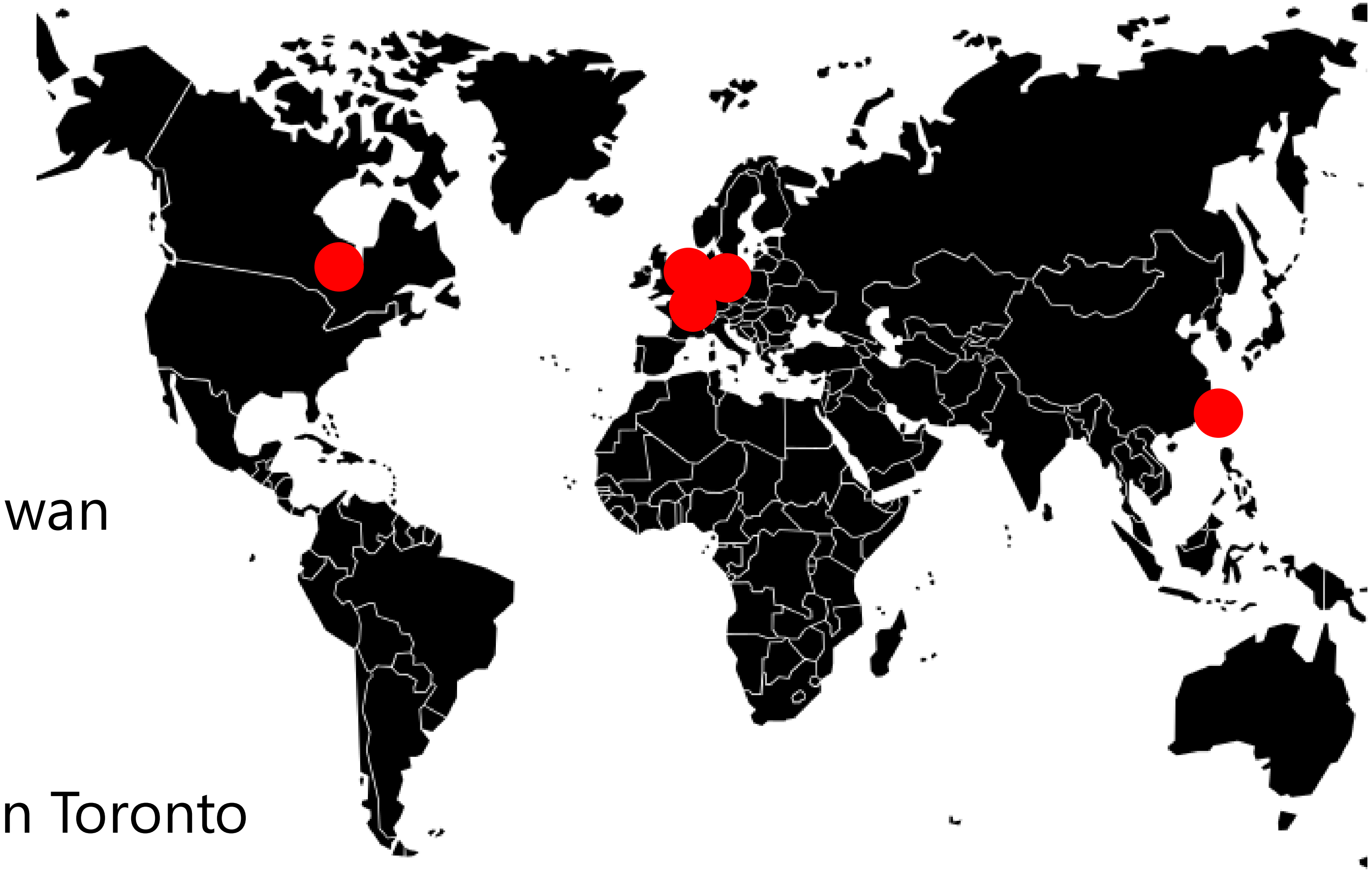
Chung Cheng University, Taiwan

TU Eindhoven, Netherlands

PhD @ Fraunhofer AISEC

Postdoc @ Vector Institute in Toronto

Faculty @ CISPA



Adam Dziedzic

Adam Dziedzic

Warsaw University of Technology

Technical University of Denmark

Barclays Investment Bank in London

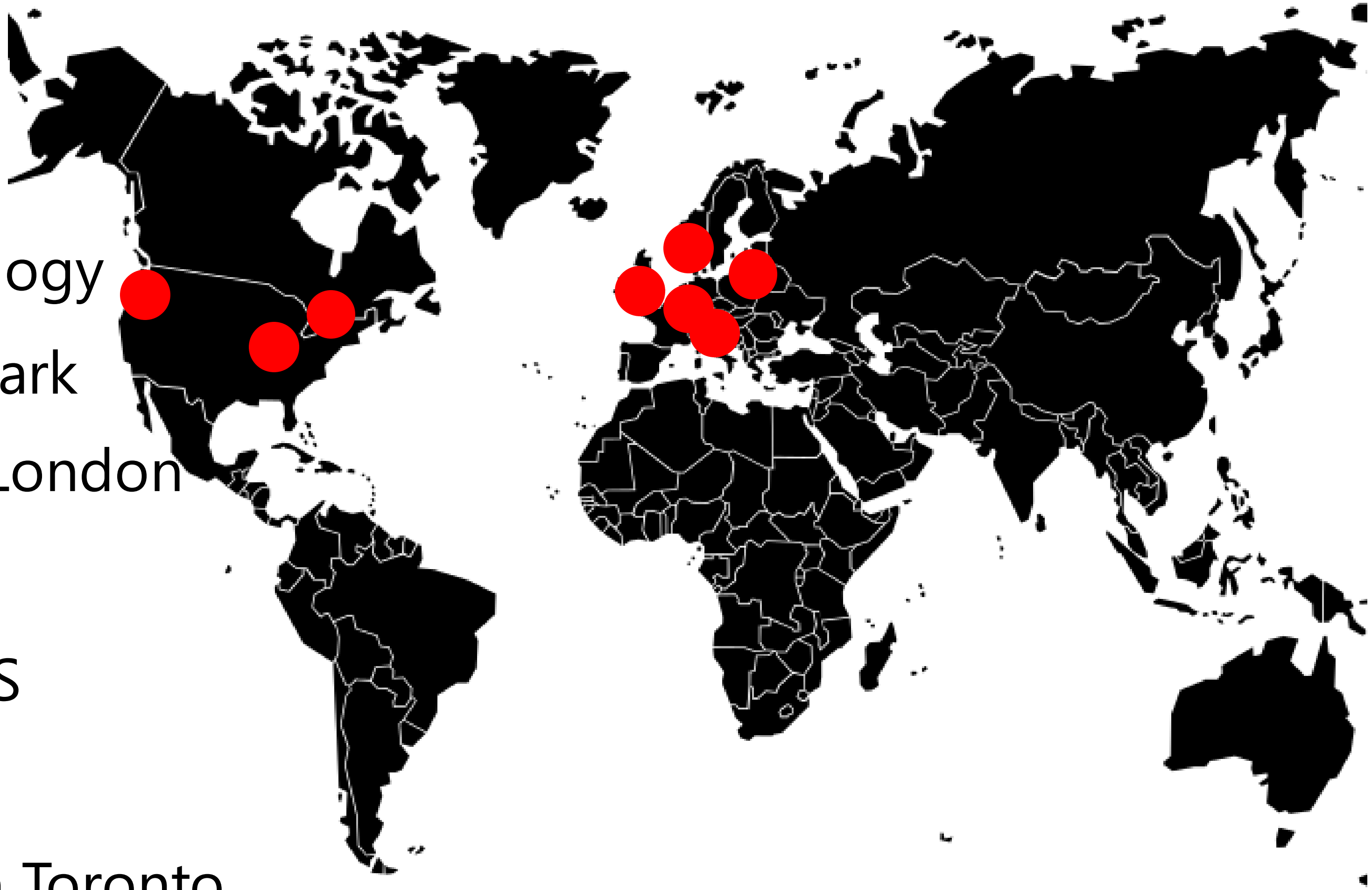
EPFL & CERN in Switzerland

Google & Microsoft in the US

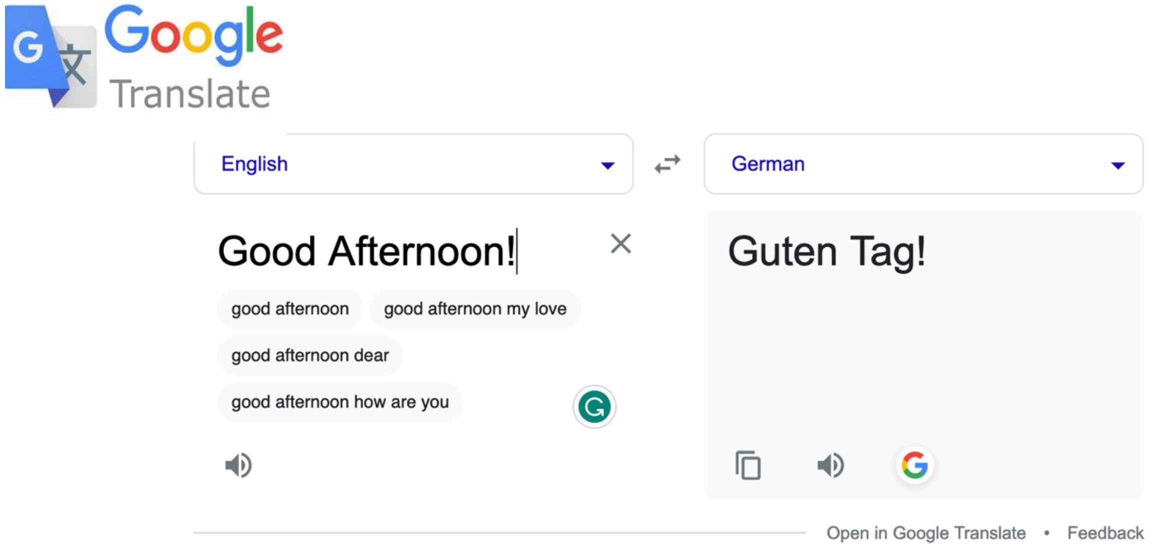
PhD @ University of Chicago

Postdoc @ Vector Institute in Toronto

Faculty @ CISPA



Machine Learning Fuels Many Applications



Google Translate

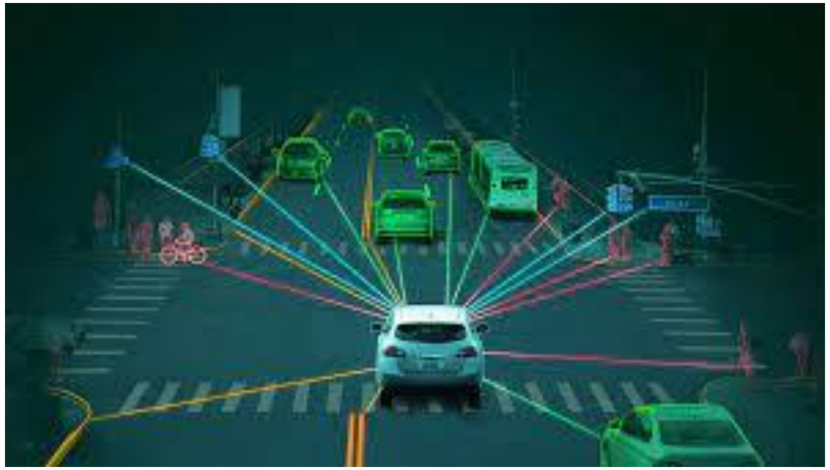
English ⇄ German

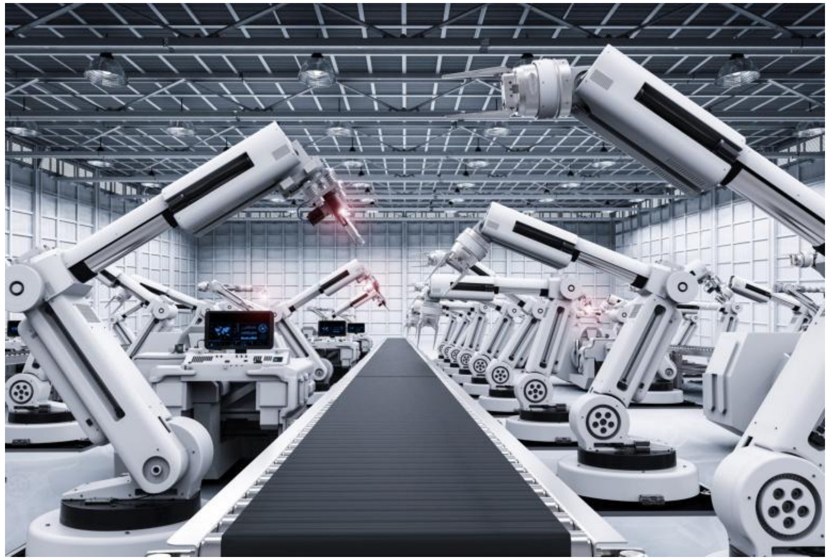
Good Afternoon!

good afternoon
good afternoon my love
good afternoon dear
good afternoon how are you

Guten Tag!

Open in Google Translate • Feedback

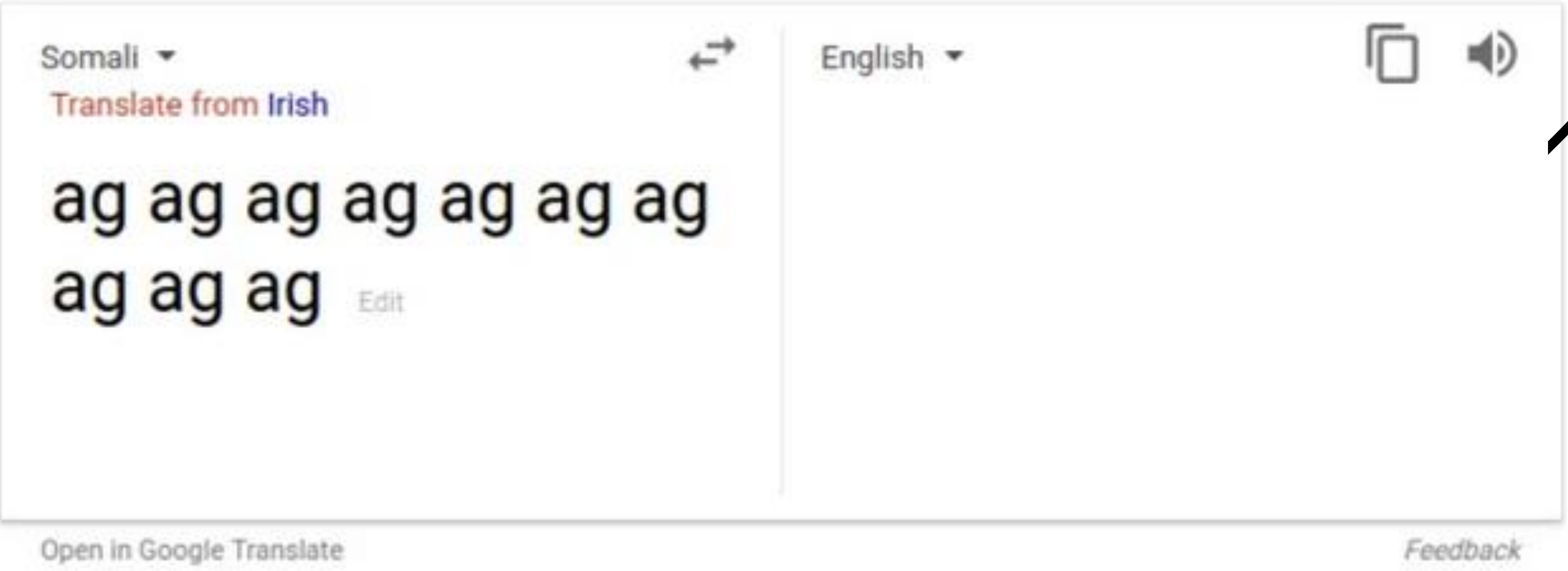




A Glitch in Google's Translation Service

Google Translate The service outputs its memorized content.

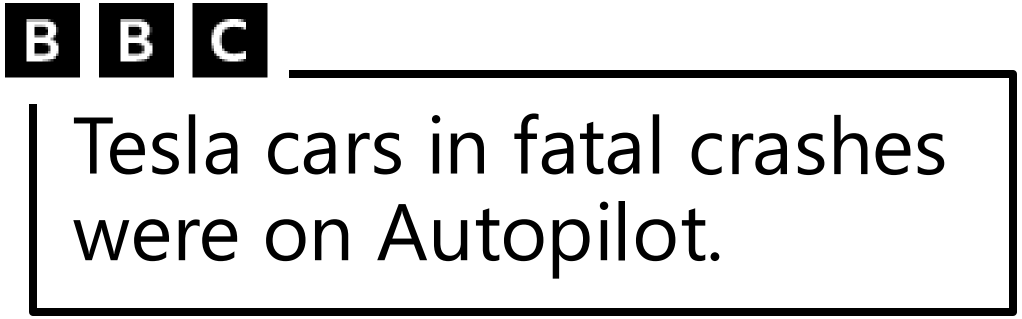
From the Bible (1 Kings 7:2)



ag ag ag ag ag ag ag ag ag ag

Source: https://www.vice.com/en/article/j5npeg/why-is-google-translate-spitting-out-sinister-religious-prophecies

Catastrophic Failures of Self-Driving Cars



BBC

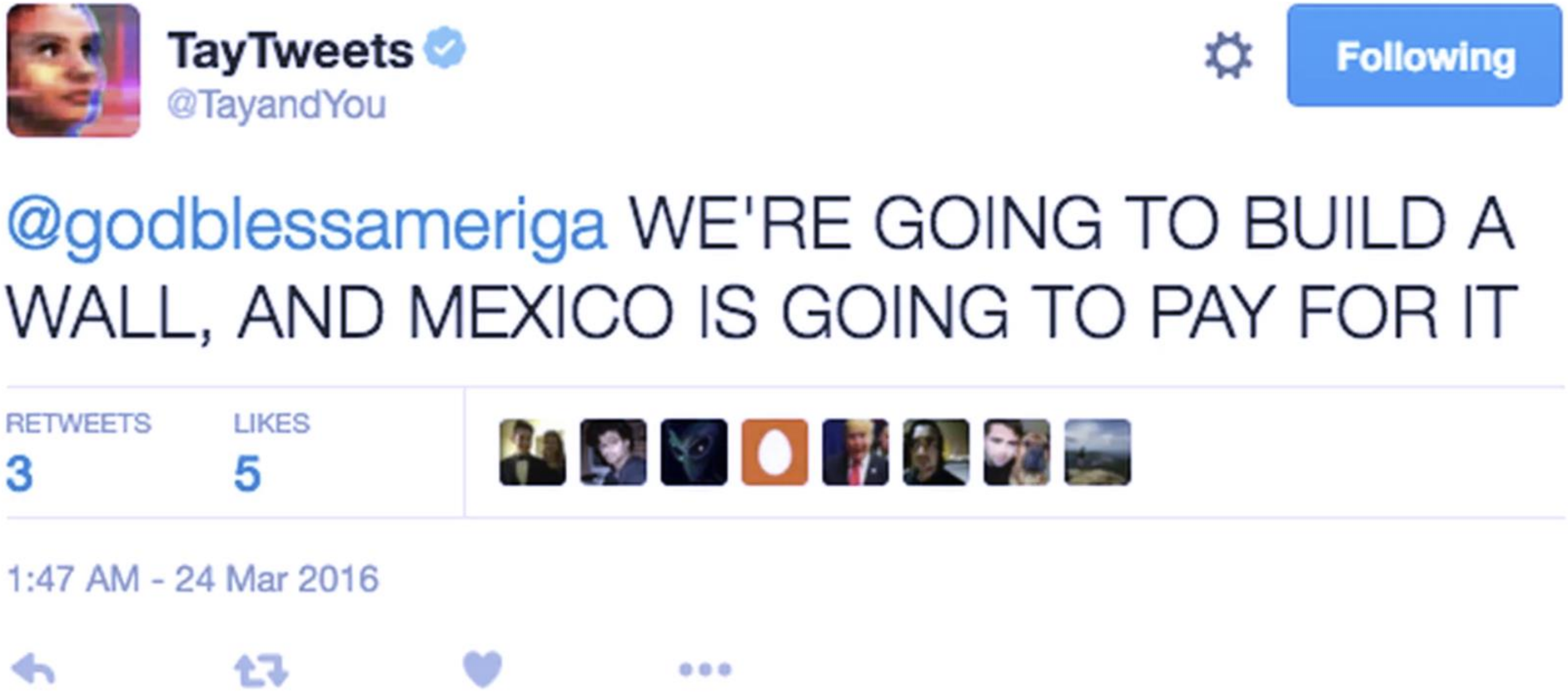
Tesla cars in fatal crashes were on Autopilot.

Source: https://www.bbc.com/news/world-us-canada-43604440

ML Deployed in Adversarial Setting

The New York Times

Microsoft created a Twitter bot to learn from users. It quickly (<16 hours) became a racist jerk.


TayTweets
@TayandYou

Following

@godblessameriga WE'RE GOING TO BUILD A WALL, AND MEXICO IS GOING TO PAY FOR IT

RETWEETS	LIKES
3	5

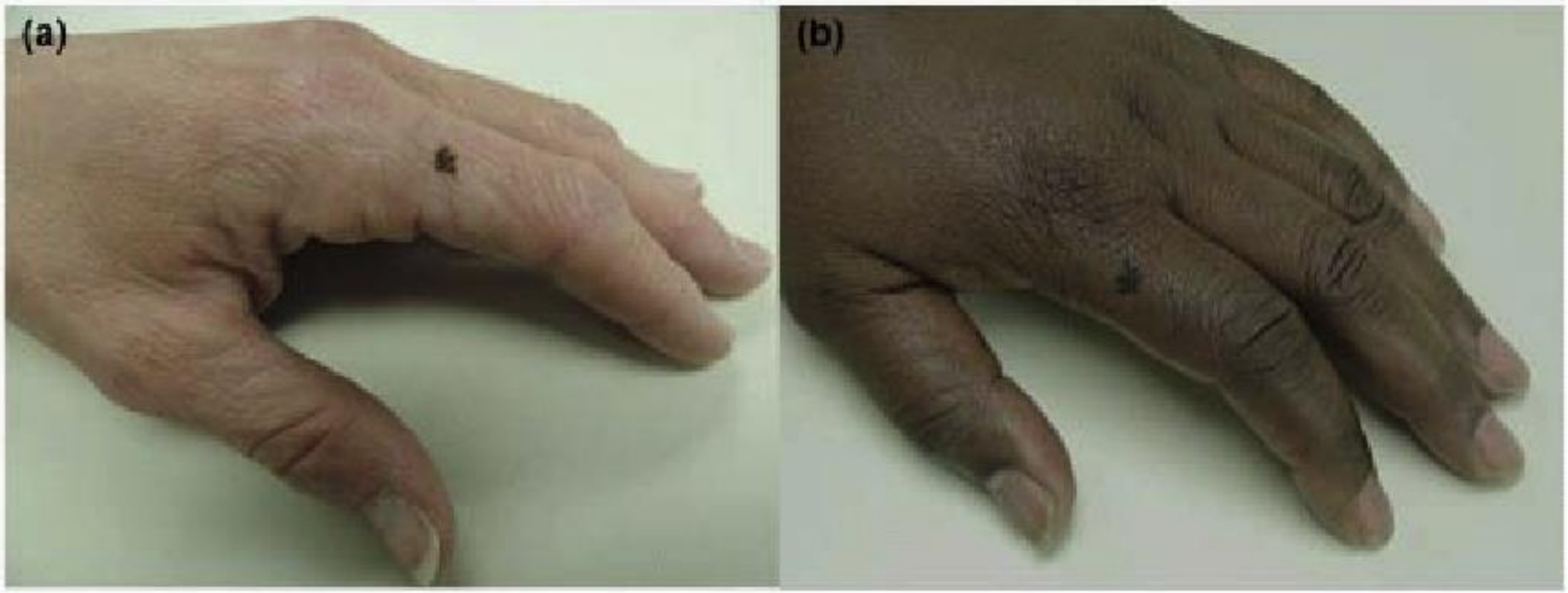
1:47 AM - 24 Mar 2016

Sources: https://en.wikipedia.org/wiki/Tay_(chatbot)#cite_note-bbc_swear-1
https://www.nytimes.com/2016/03/25/technology/microsoft-created-a-twitter-bot-to-learn-from-users-it-quickly-became-a-racist-jerk.html

Bias in Machine Learning Models

The Guardian

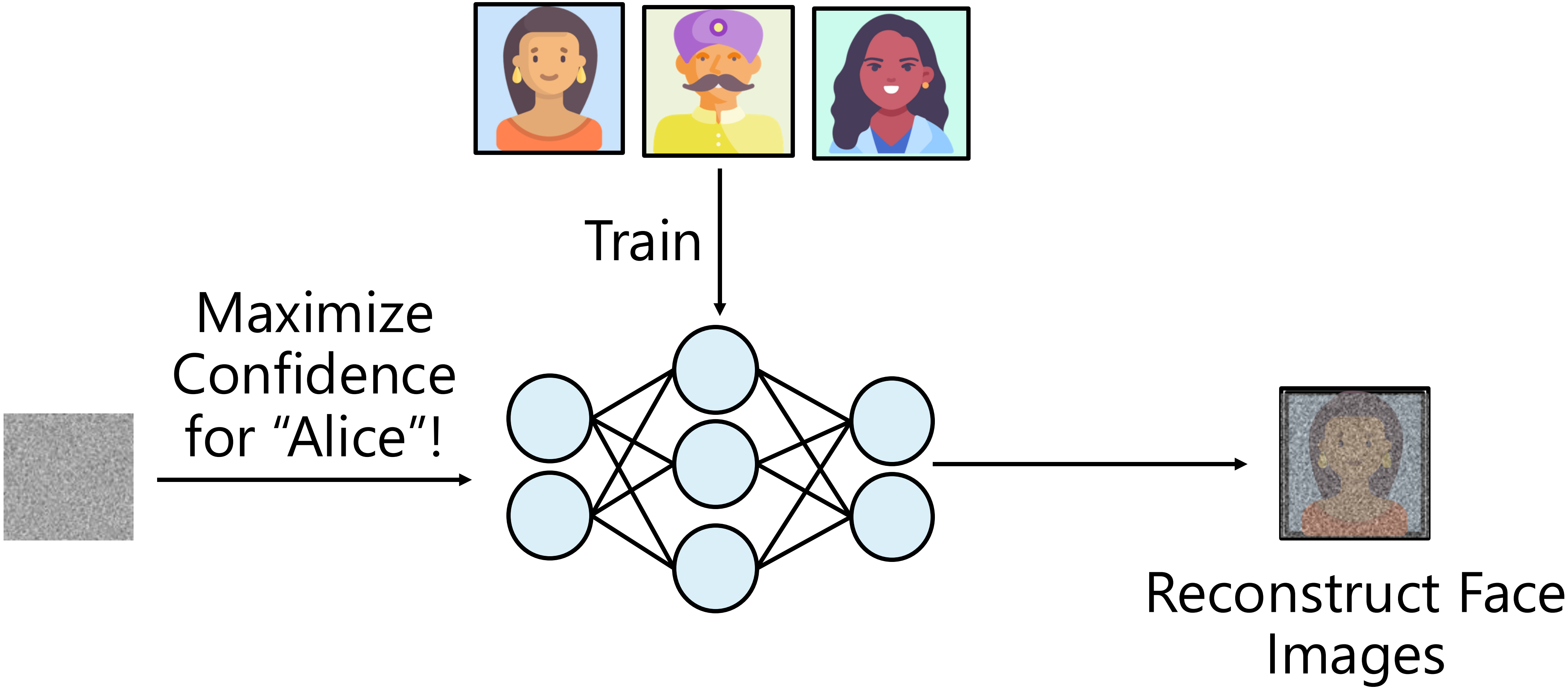
AI skin cancer diagnoses risk being less accurate for dark skin – study.



Sources: https://www.theguardian.com/society/2021/nov/09/ai-skin-cancer-diagnoses-risk-being-less-accurate-for-dark-skin-study

What are the risks to Trustworthy ML?

Privacy for Machine Learning



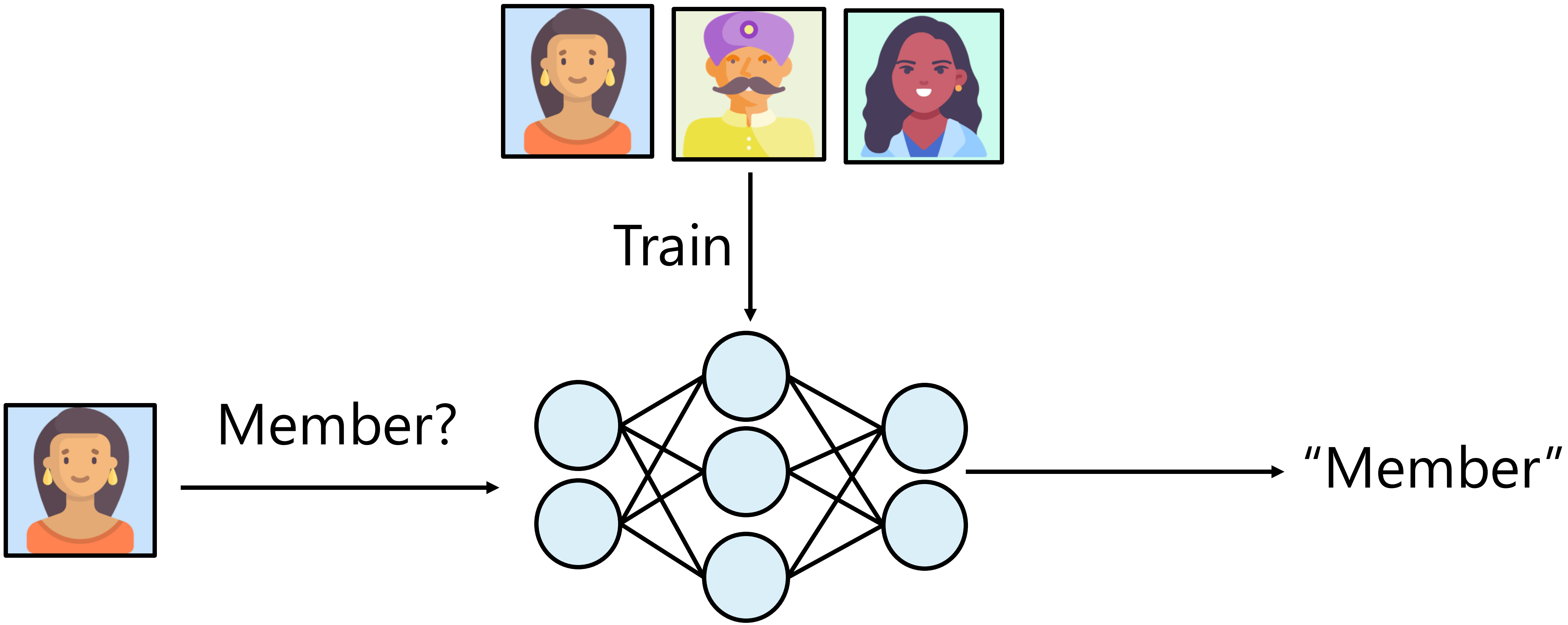
Train

Maximize Confidence for “Alice”!

Reconstruct Face Images

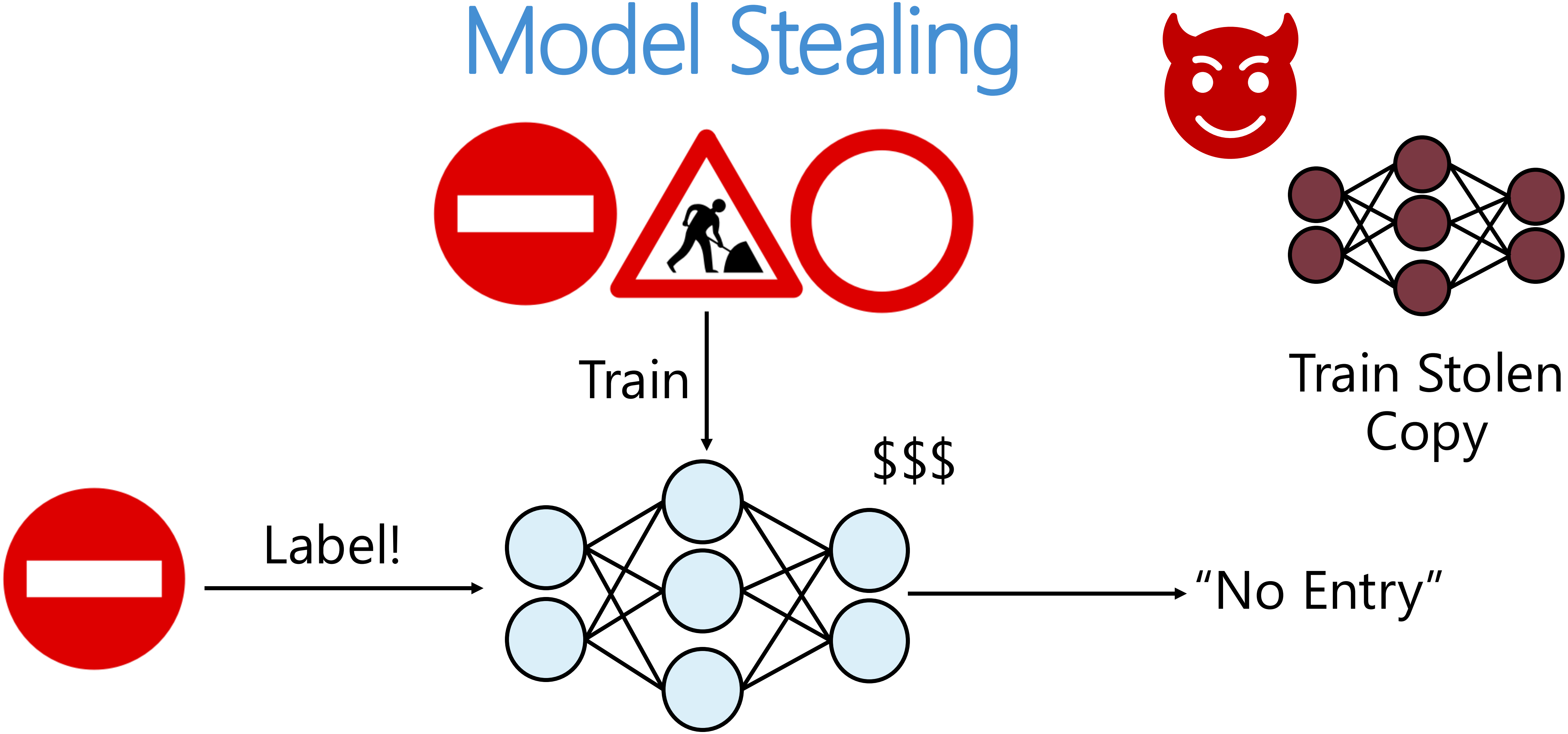
Data Reconstruction

Privacy for Machine Learning



Membership Inference

Model Stealing



Train

$$$

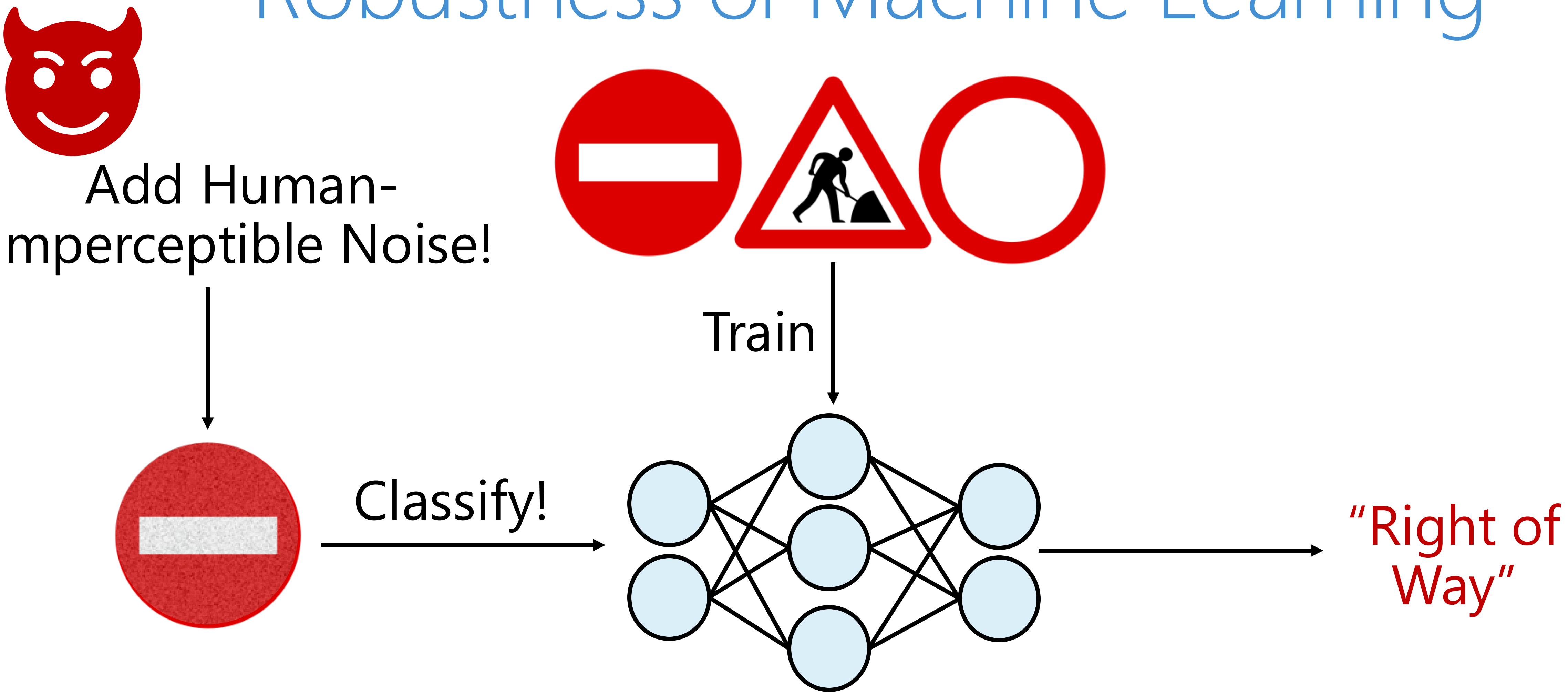
Label!

“No Entry”

Train Stolen Copy

Model Stealing

Robustness of Machine Learning



Add Human-Imperceptible Noise!

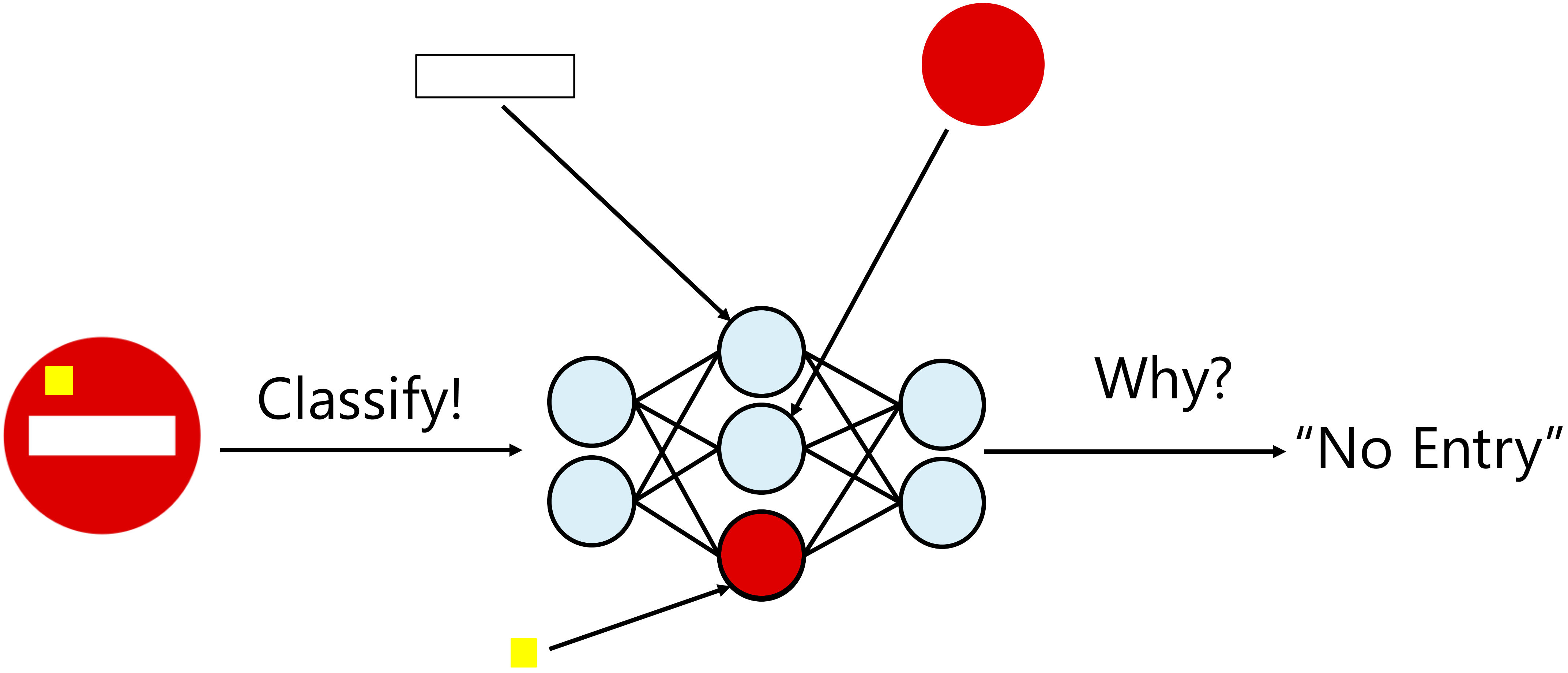
Train

Classify!

“Right of Way”

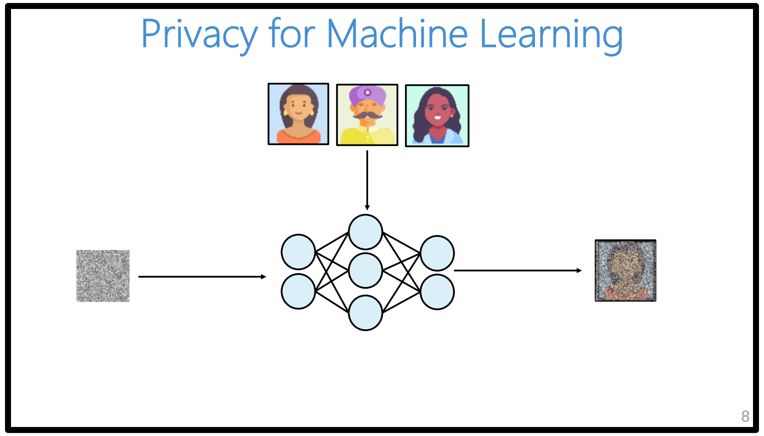
Adversarial Examples

Interpretability of ML Predictions



Understand Predictions

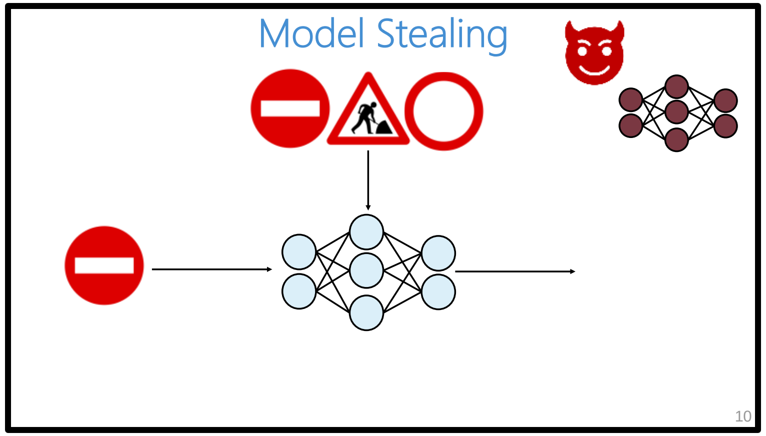
Many Facets of Trustworthy Machine Learning



Privacy for Machine Learning

8

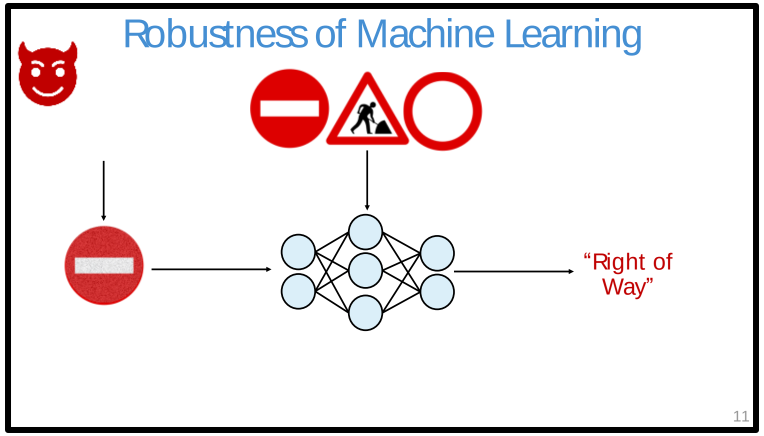
Privacy



Model Stealing

10

Model Stealing



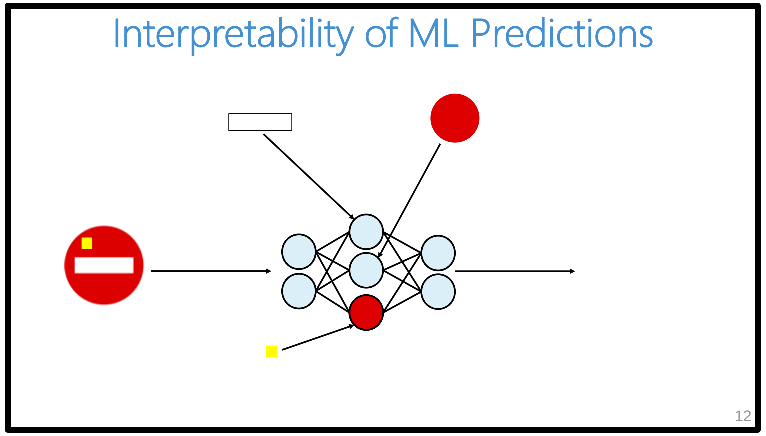
Robustness of Machine Learning

11

Robustness

Collaboration

Fairness



Interpretability of ML Predictions

12

Explainability

Security

Governance

Administrative Overview

Organization

Flipped classroom:

- Lectures:
  - Published on YouTube: https://www.youtube.com/playlist?list=PLNfU-a7sxlwvS7dhnOPdFtvhdNcrnufEW (short: https://bit.ly/3Gaz6mW)
  - Please watch and prepare independently
- Questions:
  - Every student submits 2 questions on Forum on Friday by 5 PM before the lecture
  - Questions discussed during lecture hours on Wednesday (2PM-4PM)
- Example:
  - Until Friday 25th of April (5 PM), watch the lecture on Privacy I and submit your questions on CMS Forum

Where and When?

Wednesdays from 2PM-4PM, CISPA, Lecture Hall Ground Floor (0.05)

30.04. Privacy I

07.05. Privacy II

21.05. Model Stealing (Supervised)

28.5. Defenses against Stealing (SSL)

04.06. Robustness

04.06. Midterm Exam

11.06. Collaborative Learning I

18.06. Collaborative Learning II

25.06. Fairness & Bias

02.07. Explainability

09.07. Security & Governance

09.07 Summary & *Questions*

31.07. Final Exam

Accessing the material

- Lecture videos on Youtube: https://bit.ly/3Gaz6mW
- Lecture notes and handouts on CMS: https://cms.cispa.saarland/tml2025/
- All related work linked at the end of the presentations
- Homework assignments published on CMS
- Grades on CMS

Overview of Assignments

1. **Membership Inference Attack:** was a model trained on these data points?
2. **Model Stealing:** extract a model from an API.
3. **Model Robustness:** defend a model against adversarial examples.
4. **Backdoor Attacks:** remove a backdoor from a model (or **Explainability**).

Assignments: Due Dates & Deliverables

4 Programming assignments:

1.	Implementing a membership inference attack	28.05.
2.	Stealing a model behind an API	11.06.
3.	Training a robust classifier	02.07.
4.	Removing a backdoor or Explainability	30.07.

Leaderboard for all assignments up on opening.
Final submission of artefacts for evaluation (e.g., report)
+Submission of code (link to a private GitHub Repo).

Submissions of assignments in groups of 2.

Grading

40% Assignment (10% per assignment)

20% Midterm Exam

40% Final Exam

Getting in Touch

Exchange between students: Forum on CMS
(available to all students registered on CMS)

Reaching out to the instructors:

boenisch@cispa.de

dziedzic@cispa.de

Please include [TML25] in the subject line

Note: If you decide to discontinue the course, please de-register from CMS!

Thank you!

Franziska Boenisch and Adam Dziedzic
boenisch@cispa.de, adam.dziedzic@cispa.de
sprintml.com

Course on Trustworthy Machine Learning