Introduction to Trustworthy Machine Learning

Franziska Boenisch and Adam Dziedzic

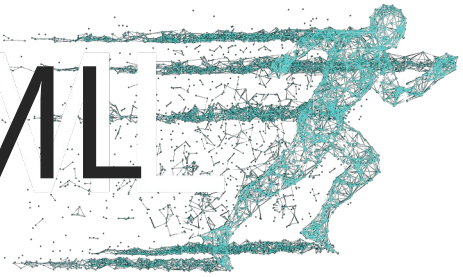
Course on Trustworthy Machine Learning

April 17th, 2024

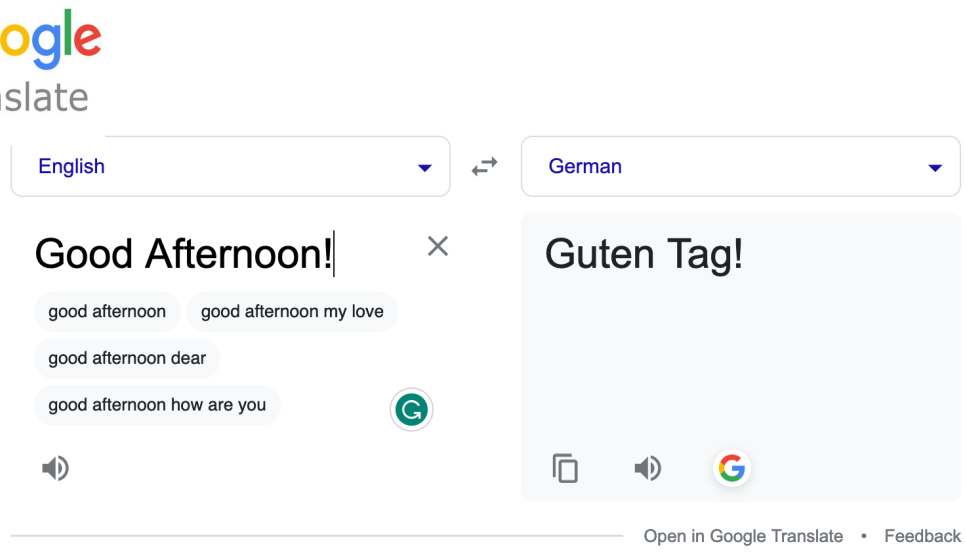
CISPA

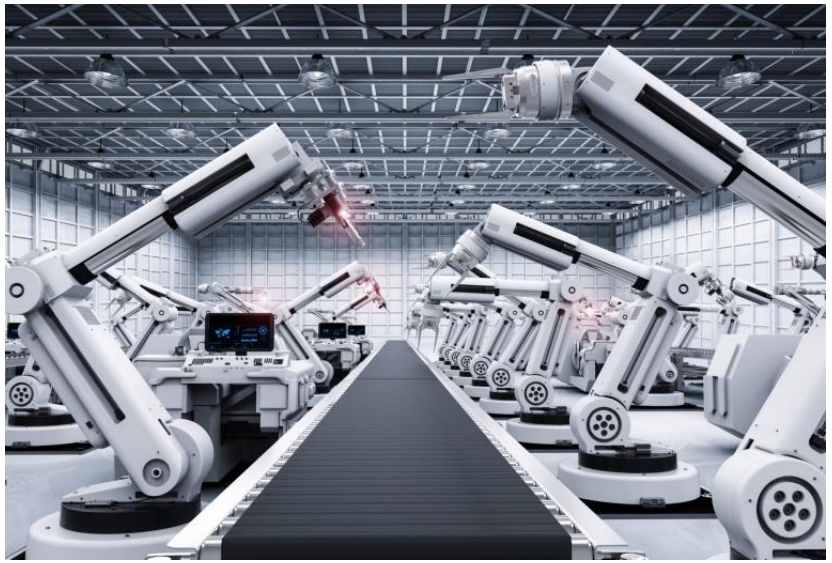
HELMHOLTZ CENTER FOR INFORMATION SECURITY

SprintML



Machine Learning Fuels Many Applications





A Glitch in Google's Translation Service

Google Translate

The service outputs its memorized content.

From the Bible
(1 Kings 7:2)

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

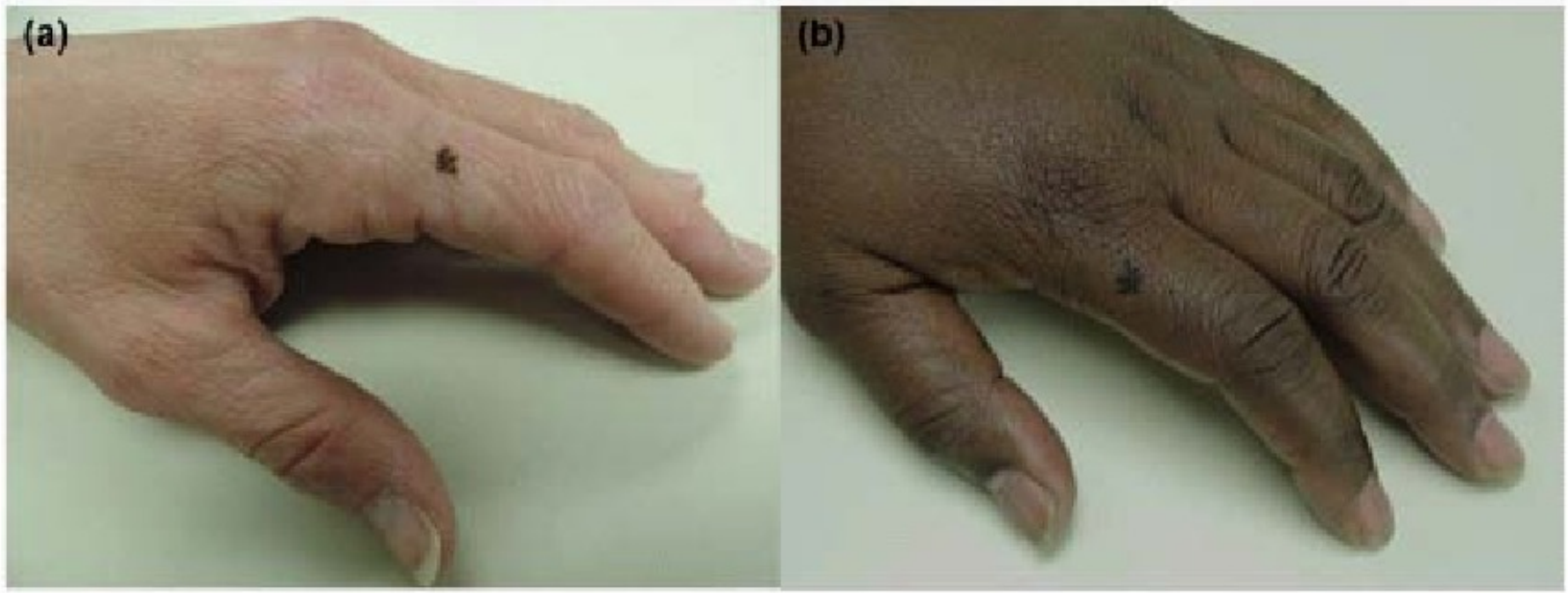
> Microsoft created a Twitter bot to learn from users. It quickly (<16 hours) became a racist jerk.

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

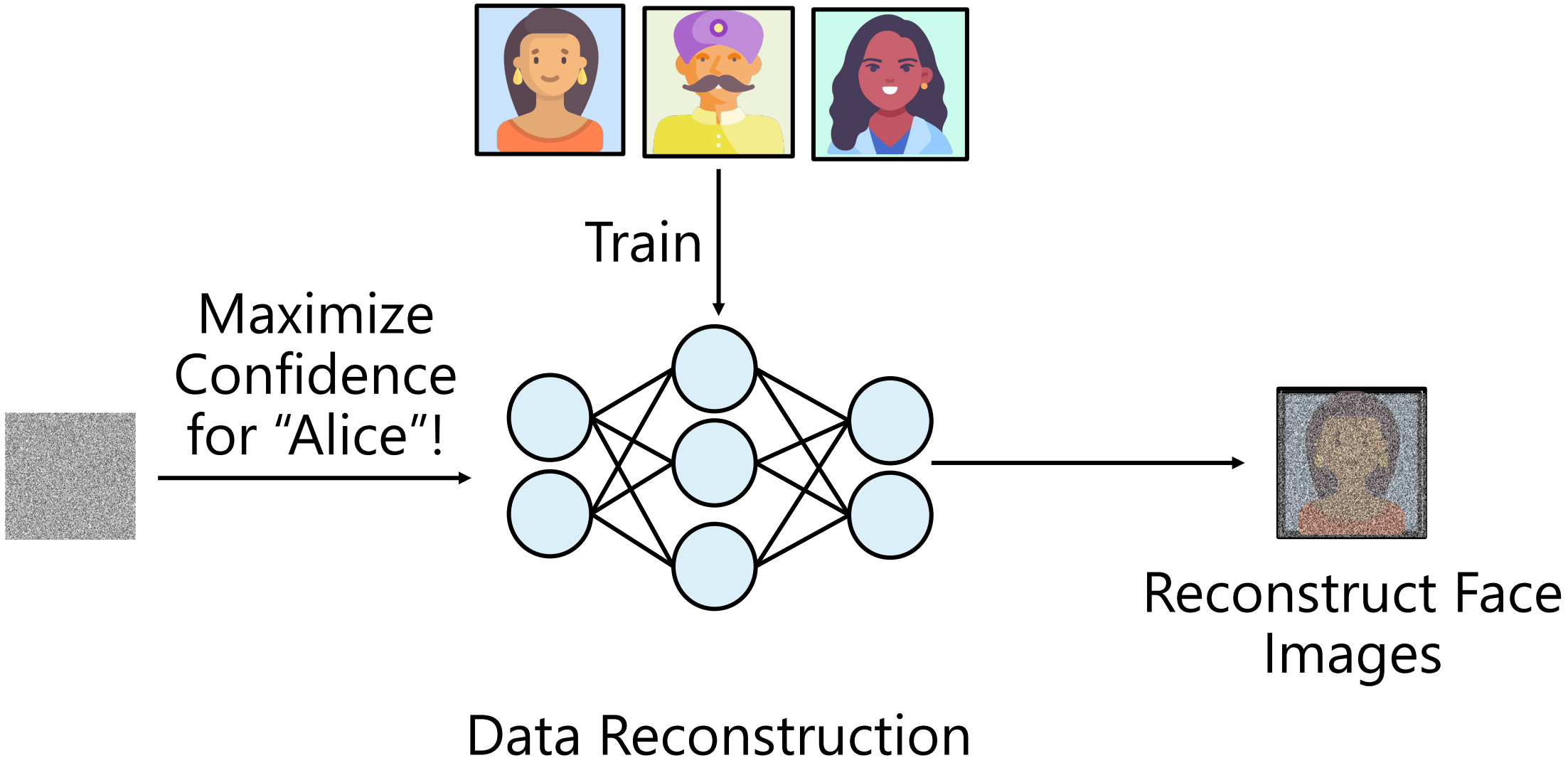
AI skin cancer diagnoses risk being less accurate for dark skin – study.



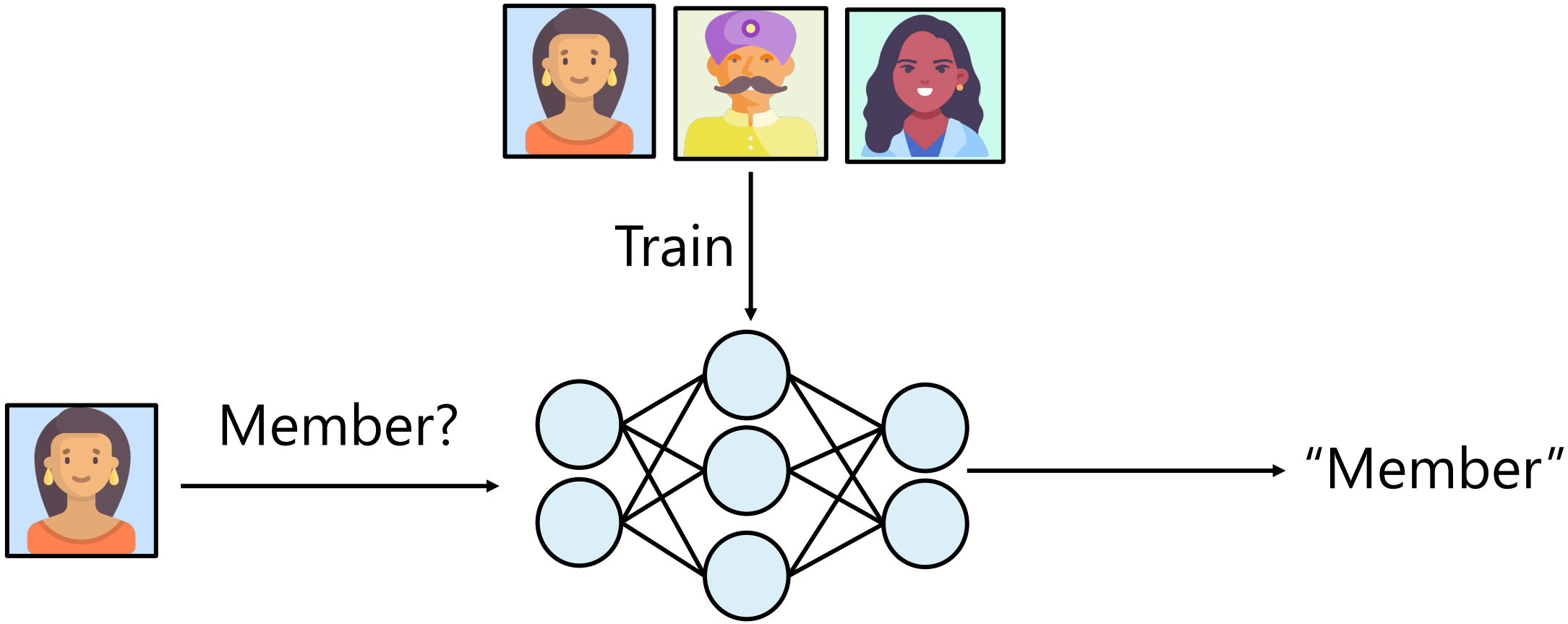
Sources: https://www.theguardian.com/society/2021/nov/09/ai-skin-cancer-diagnoses-risk-being-less-accurate-for-dark-skin-study

What are the risks to Trustworthy ML?

Privacy for Machine Learning

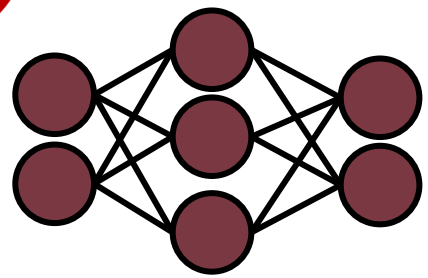


Privacy for Machine Learning



Membership Inference

Model Stealing

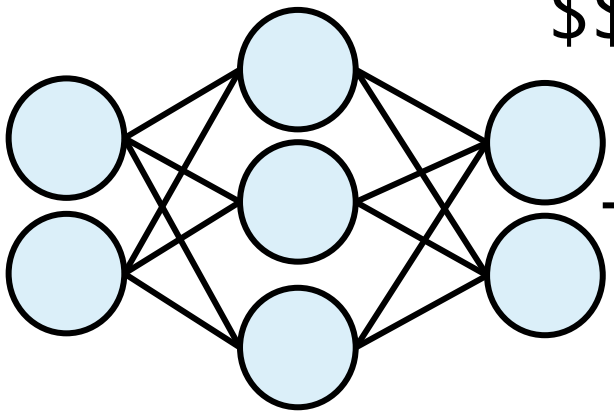


Train

Train Stolen Copy

$$$

Label!



"No Entry"

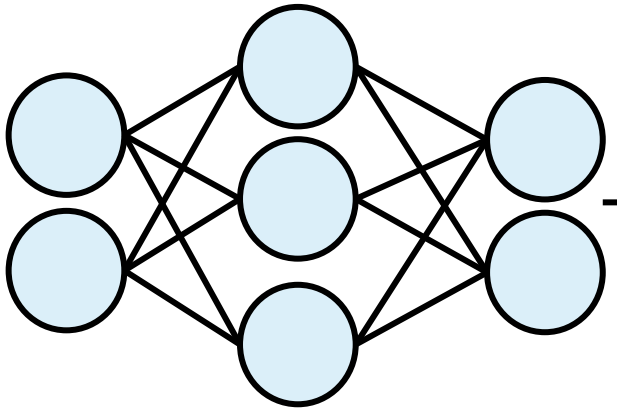
Model Stealing

Robustness of Machine Learning

Add Human-Imperceptible Noise!

Train

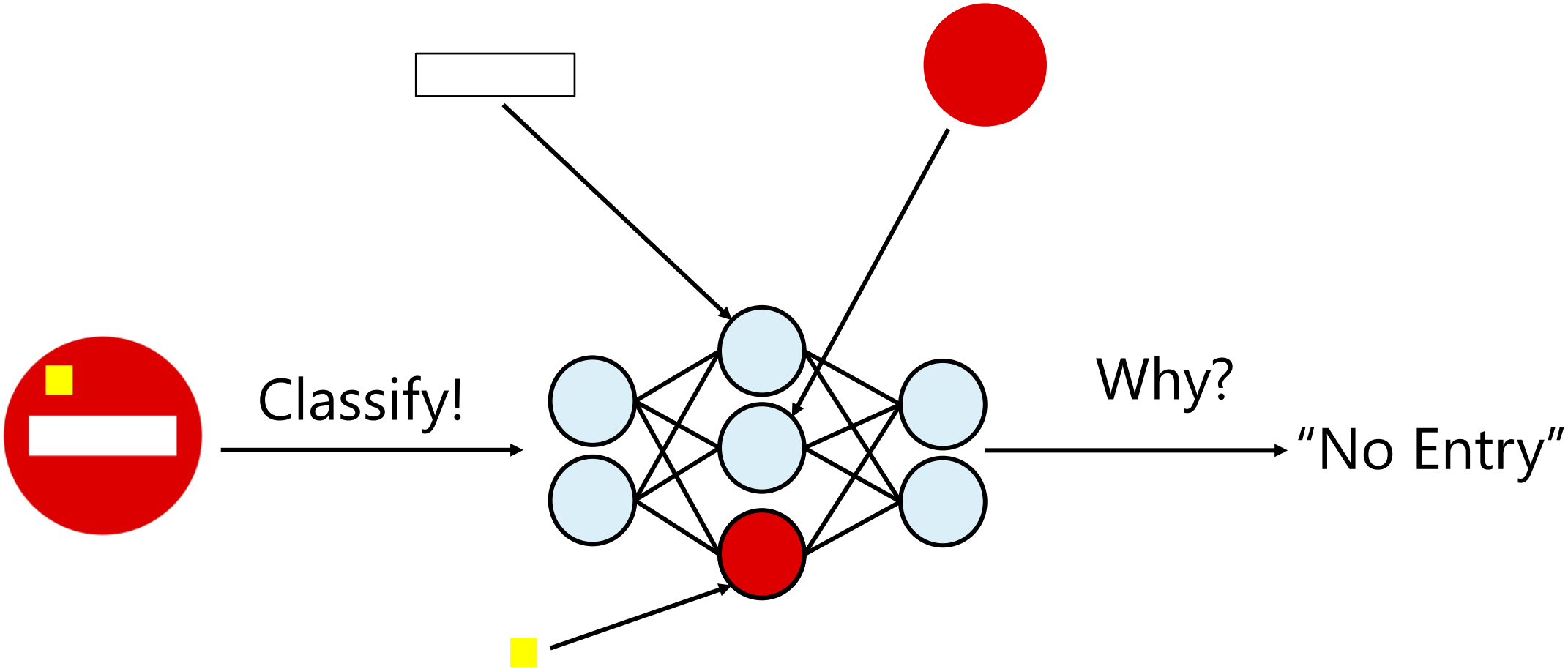
Classify!



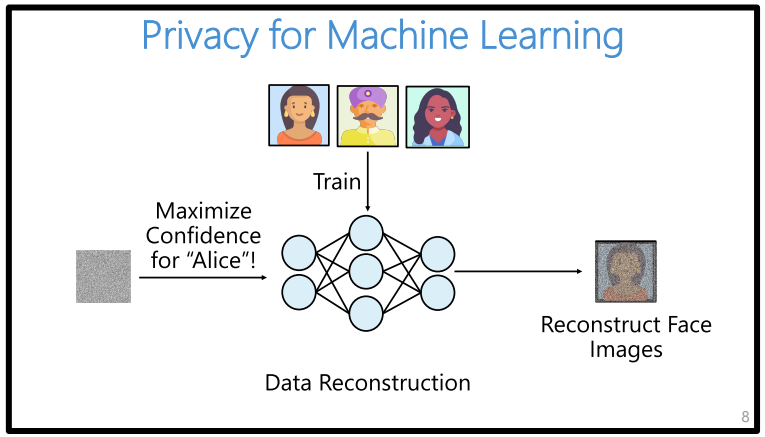
"Right of Way"

Adversarial Examples

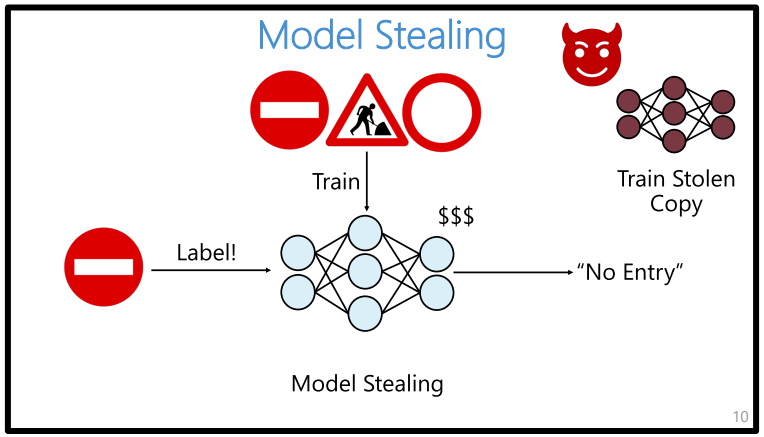
Interpretability of ML Predictions



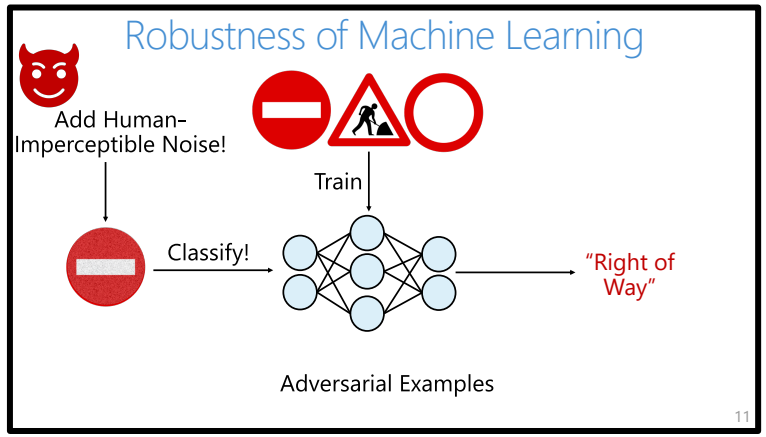
Many Facets of Trustworthy Machine Learning



Privacy



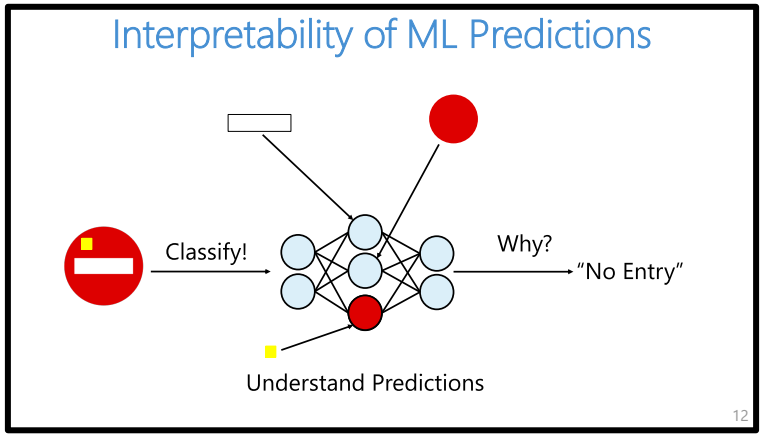
Model Stealing



Robustness

Collaboration

Fairness



Explainability

Security

Governance

Thank you!

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

Course on Trustworthy Machine Learning