Can you trust your machine learning system?

*

Sandip Kundu

National Science Foundation

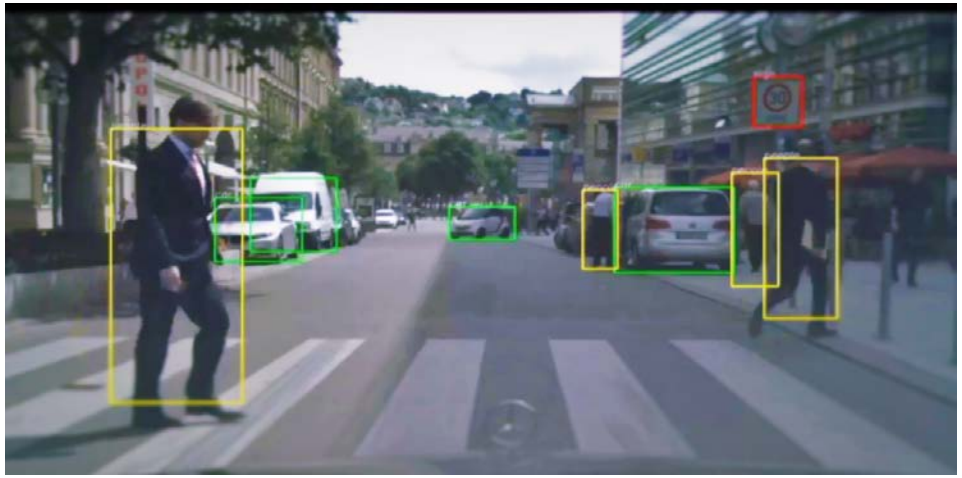
on leave from

University of Massachusetts, Amherst

* Image - Created by Patrickss - Freepik.com

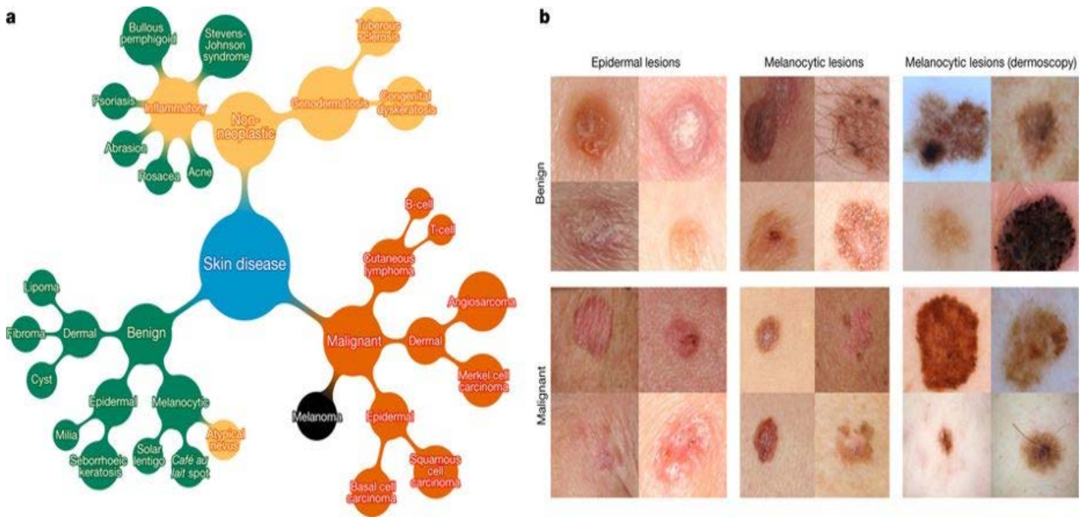
Machine Learning is becoming Ubiquitous

Self-driving Cars

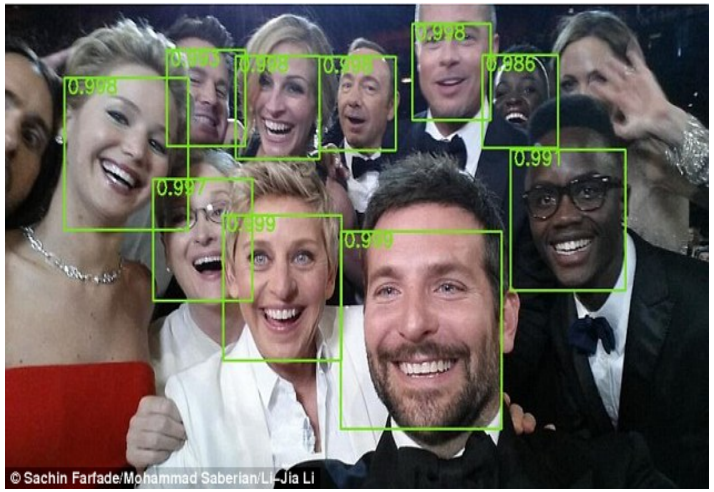


Cybersecurity

Healthcare



Facial Recognition



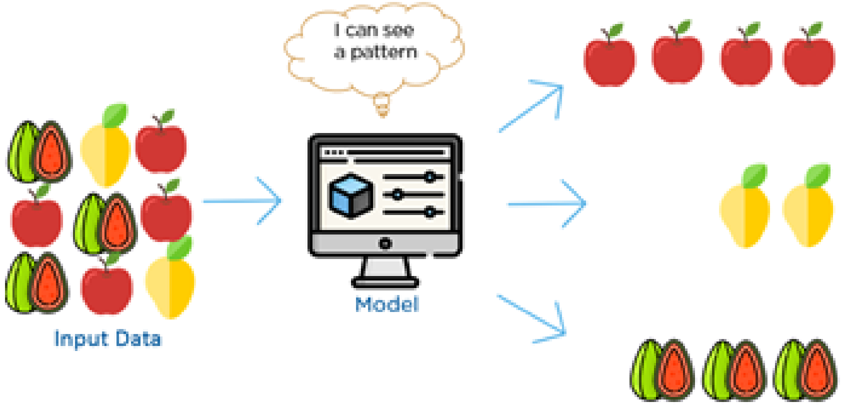
Speech Recognition

Machine Learning

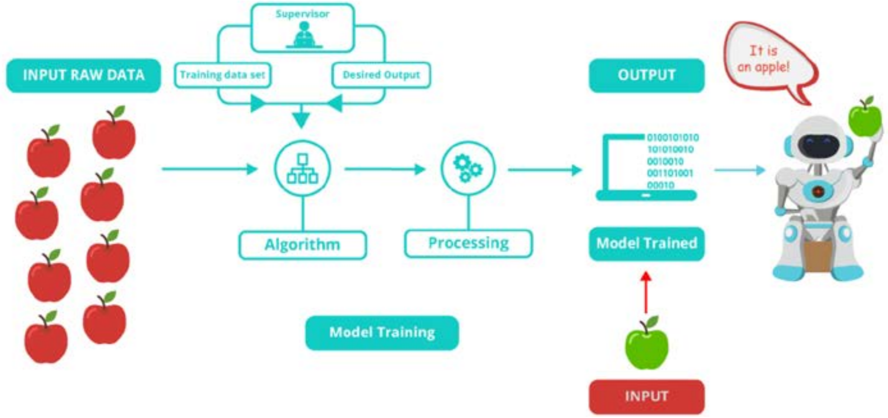
Training Data → **Mathematical Model** → **Decision/ Prediction**

Types of Machine Learning

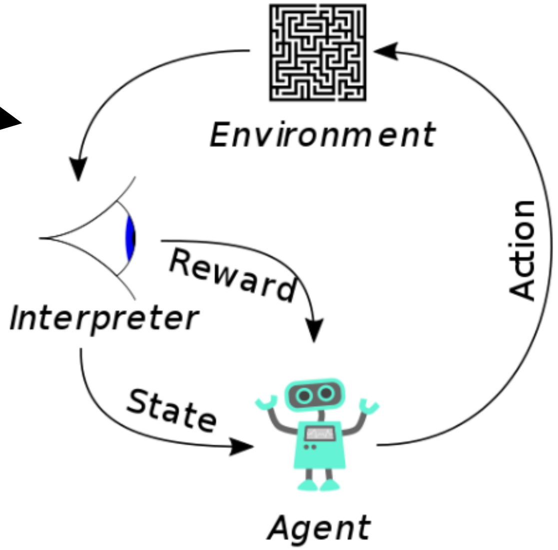
Unsupervised



Supervised

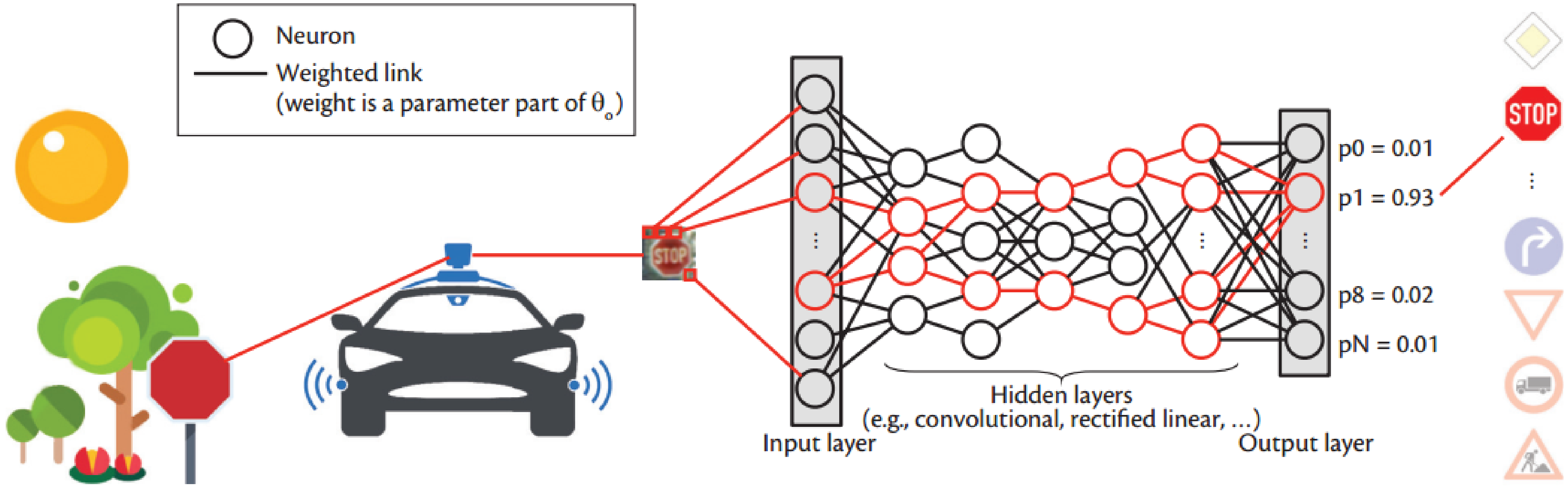


Reinforcement



Self-driving Cars

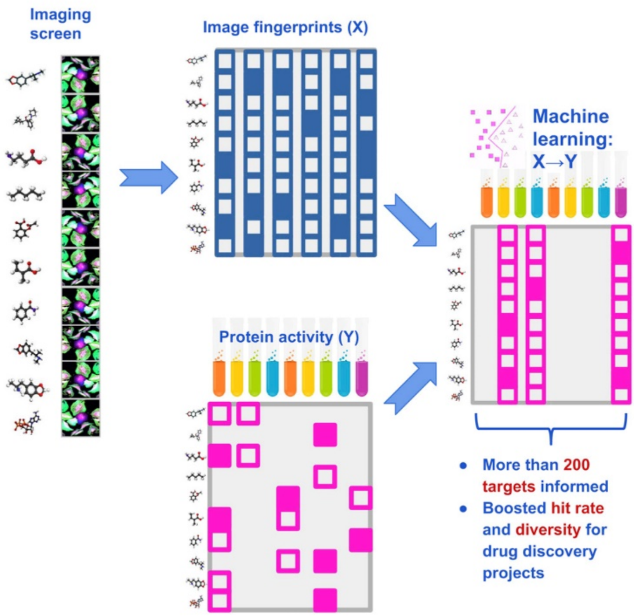
- Cars incorporating systems to assist or replace drivers
  - Ex. automatic parking, Waymo
- Self-driving cars with ML infrastructure will become commonplace
  - Ex. NVIDIA DRIVETM PX 2 – open AI car computing system



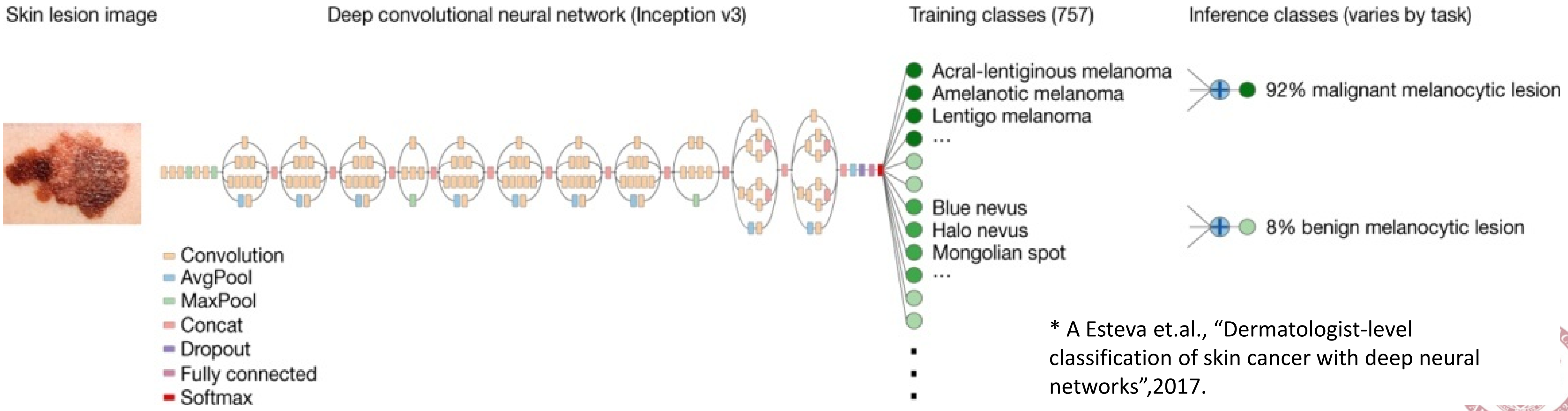
Source: Mcdaniel et.al., “Machine Learning in Adversarial Settings”,2016.

Healthcare Applications

- ❖ Diagnosis in Medical Imaging
- ❖ Treatment Queries and Suggestions
- ❖ Drug Discovery
- ❖ Personalized Medicine



* Simm, Jaak, et al. "Repurposing high-throughput image assays enables biological activity prediction for drug discovery." *Cell chemical biology* (2018)



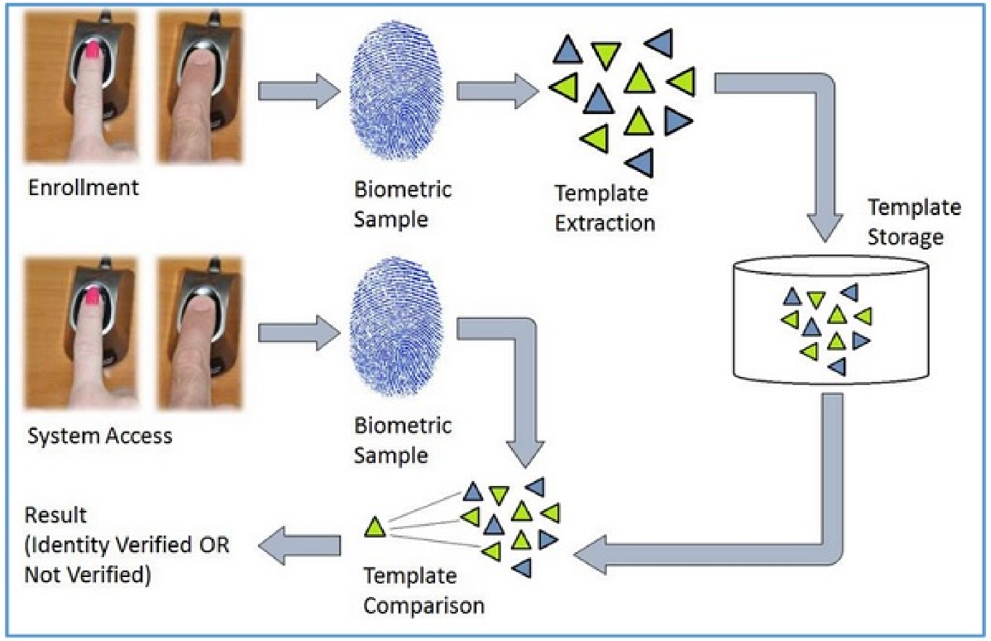
* A Esteva et.al., “Dermatologist-level classification of skin cancer with deep neural networks”,2017.

Cybersecurity

Spam Filtering

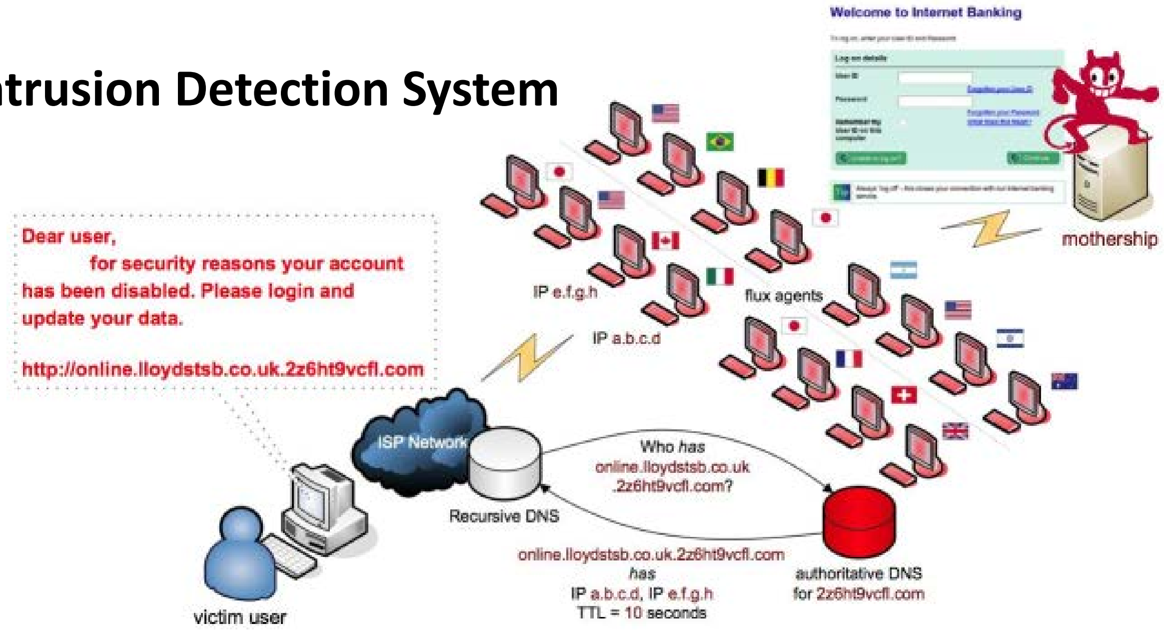
* http://www.thenonprofittimes.com/news-articles/rate-legit-emails-getting-caught-spam-filters-jumped/

Biometrics ID

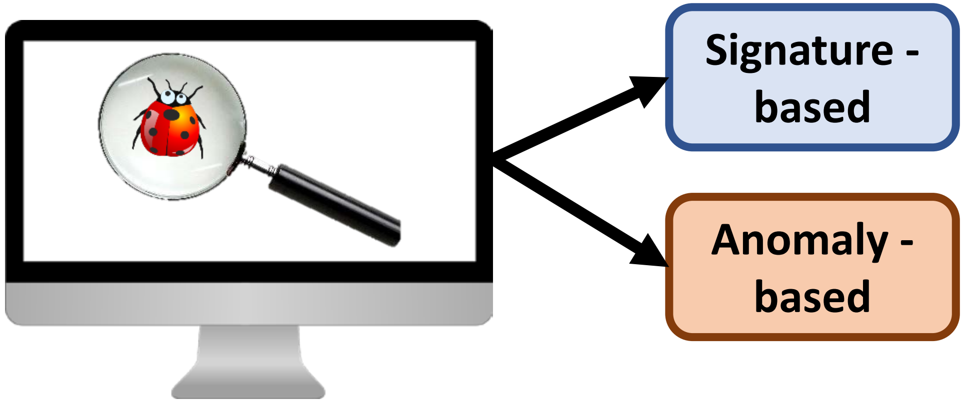


* https://www.tutorialspoint.com/biometrics/biometrics_overview.htm

Intrusion Detection System



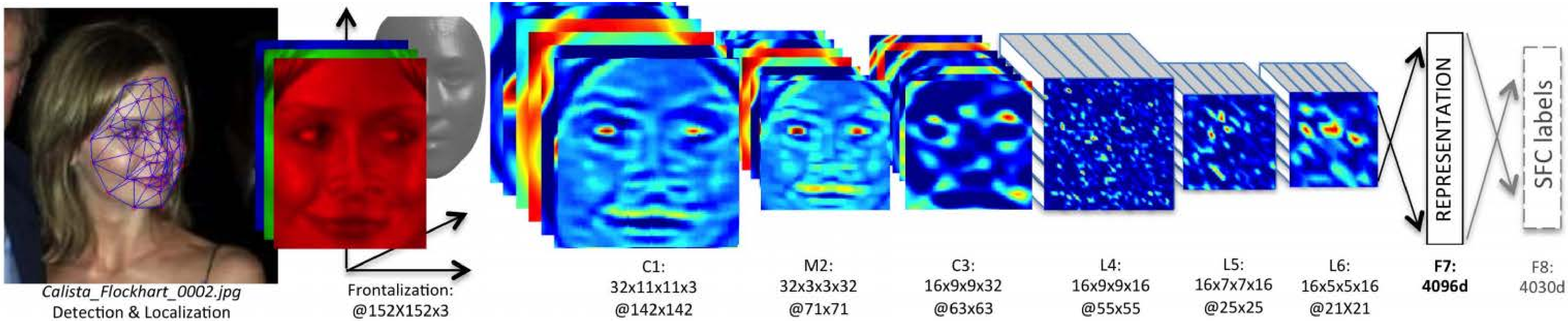
Malware Detection



Facial Recognition

- Secure Authentication and Identification
  - Apple FaceID
  - FBI database – criminal identification
- Customer Personalization
  - Ad targeting
  - Snapchat

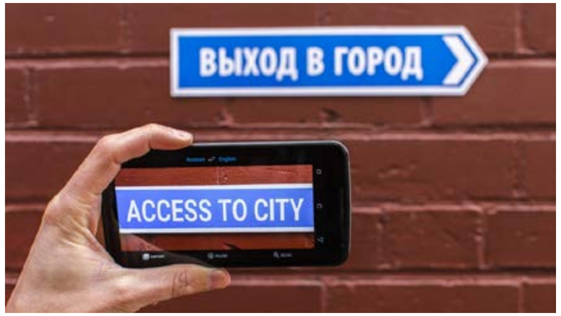
* Posterscope, Ouividi EYE Corp Media, Engage M1 – GMC Arcadia



Taigman et.al.,"DeepFace: Closing the Gap to Human-Level Performance in Face Verification",2014

Other Machine Vision Applications

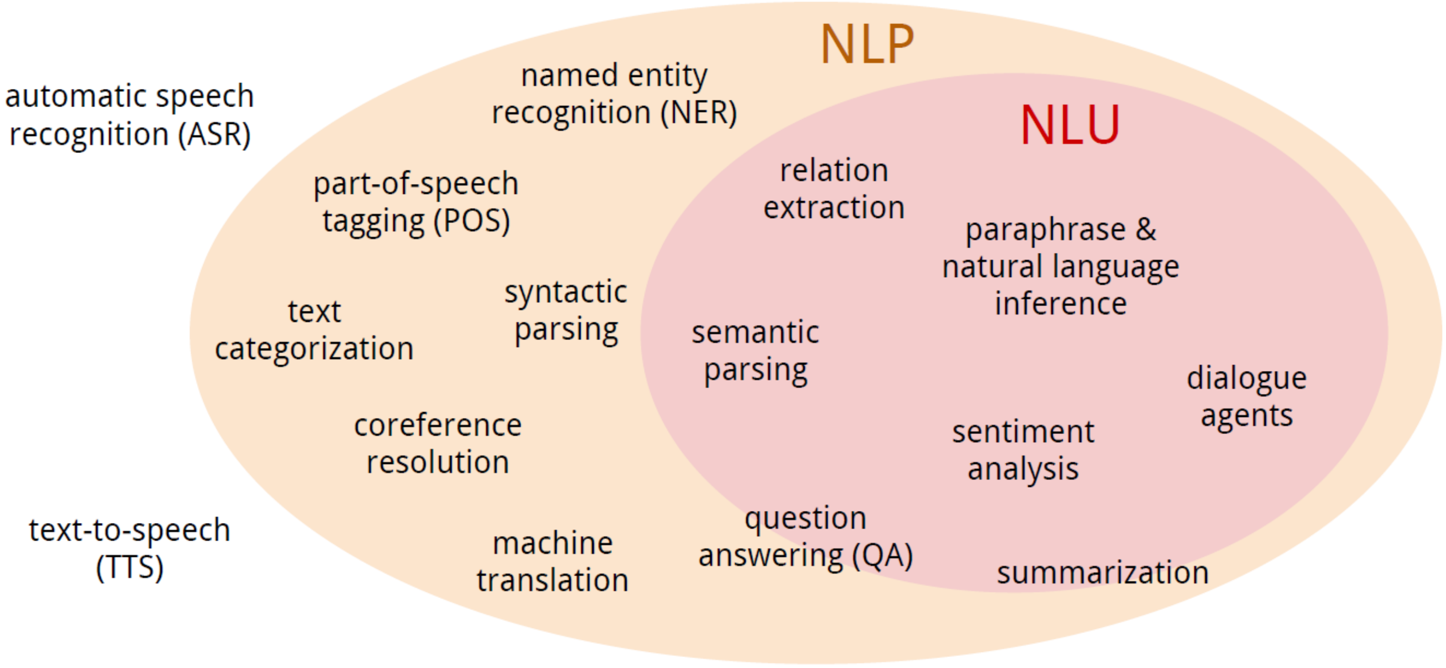
- **Digital annotation** of real-world
  - Text, language recognition – E.g. Billboards, auto-translation
  - Geo-tagging Landmarks
  - Integration with other services – E.g. ratings for restaurant, directions



- **Augmented Reality**
  - **Gaming** – adaptive integration with real-world
  - **Augmented Retail** – E.g. Clothes Fitting

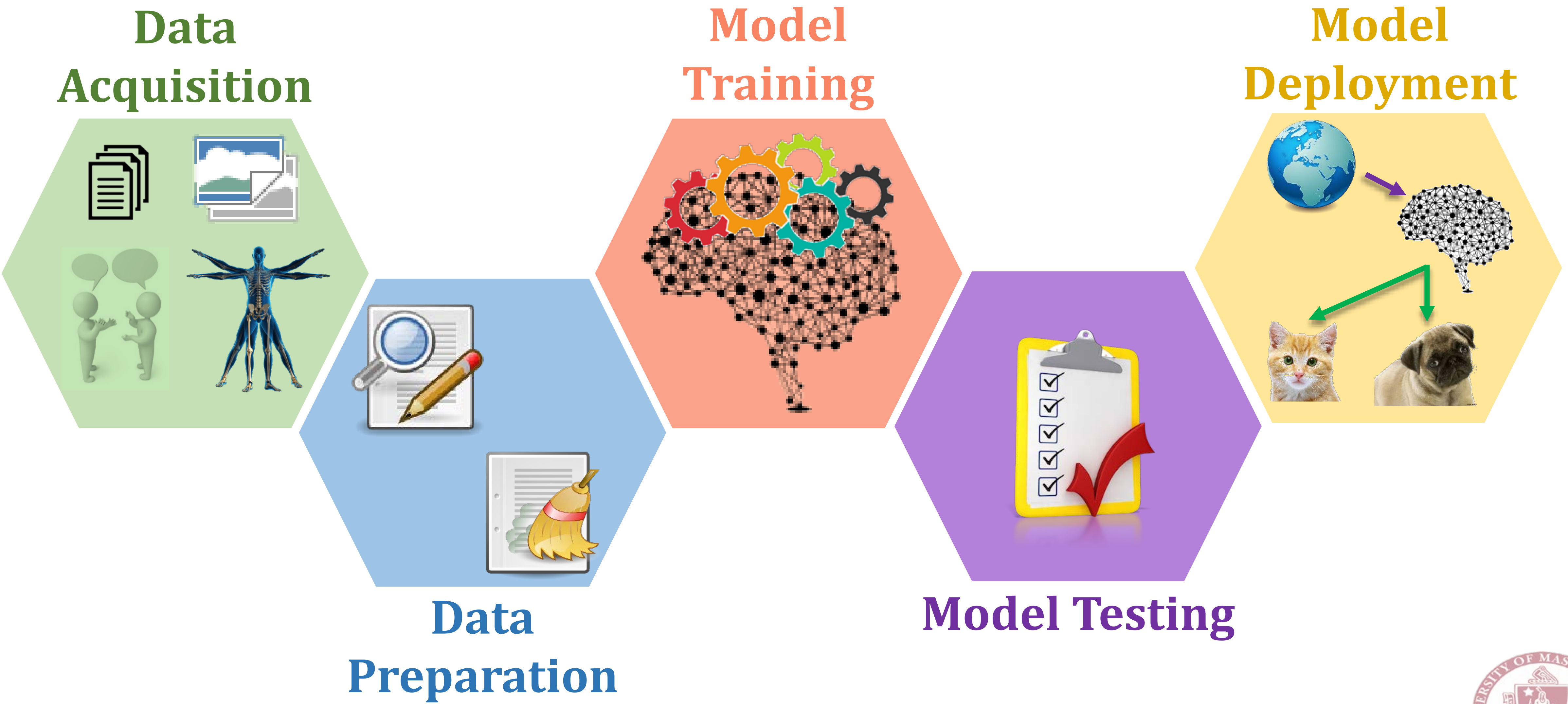
Speech Recognition

- ❖ Envisioned in science fiction since 1960’s
  - HAL 9000, Star Trek
- ❖ Natural Language Processing (NLP) has gained increased importance
  - Modeling large vocabularies, accents – translation, transcription services
  - **Smartphones** – Apple Siri, Google Assistant, Samsung Bixby
  - Home - Amazon’s Echo/Alexa,
  - IBM Watson



http://nlp.stanford.edu/~wcmac/papers/20140716-UNLU.pdf

Machine learning (ML) Process



Machine Learning Security and Privacy

Introduction

- ML algorithms in real-world applications mainly focus on **accuracy** (effectiveness) **or/and efficiency** (dataset, model size)
  - Few techniques and design decisions to keep the ML models ***secure and robust***!

- Machine Learning as a Service (MLaaS) and Internet of Things (IoT) further complicate matters
  - Attacks can **compromise millions of customers**' security and privacy
  - Concerns about **Ownership** of data, model

amazon web services | bigml | PredictionIO | Microsoft Azure

ML Vulnerabilities

- Key vulnerabilities of machine learning systems
  - ML models often derived from **fixed datasets**
  - Assumption of similar distribution between training and real-world data
    - **Coverage** issues for complex use cases
    - Need **large datasets**, **extensive data annotation**, **testing**

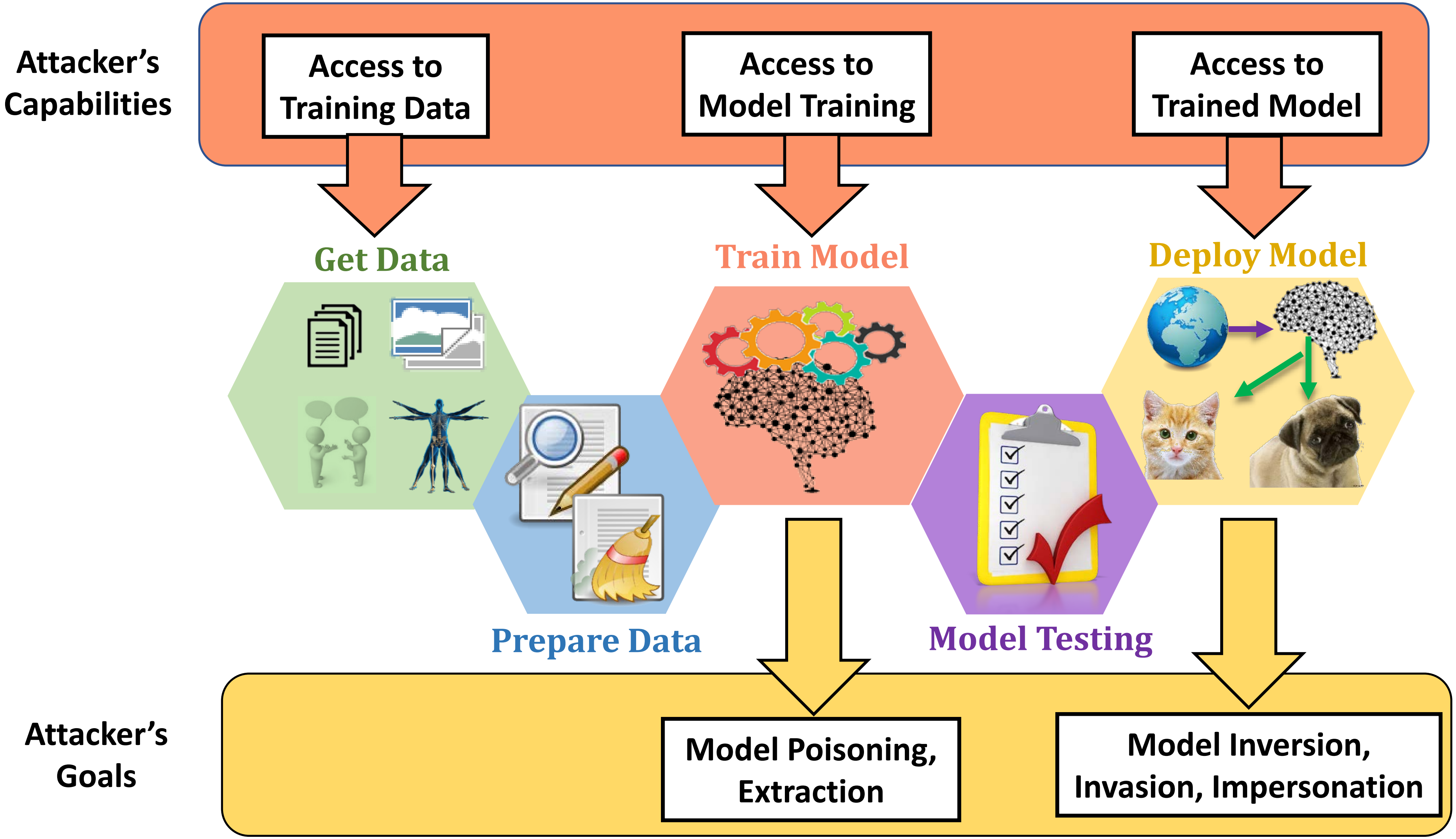
- Strong adversaries against ML systems
  - ML algorithms **established** and **public**
  - Attacker can leverage ML knowledge for **Adversarial Machine Learning** (AML)
    - **Reverse engineering** model parameters, test data – **Financial incentives**
    - **Tampering** with the trained model – **compromise security**

Classification of Security and Privacy Concerns

- **Attacker's Goals**
  - o **extract** model **parameters** **(model extraction)**
  - o **extract private data** **(model inversion)**
  - o **compromise** model to produce false positives/negatives **(model poisoning)**
  - o **produce** adversary selected outputs **(model evasion)**
  - o **render** model **unusable**

- **Attacker's Capabilities**
  - o access to Black-box ML model
  - o access to White-box ML model
  - o manipulate ***training data*** to **introduce** vulnerability
  - o access to query to ML model
  - o access to query to ML model with confidence values
  - o access to training for building model
  - o **find and exploit** vulnerability during ***classification***

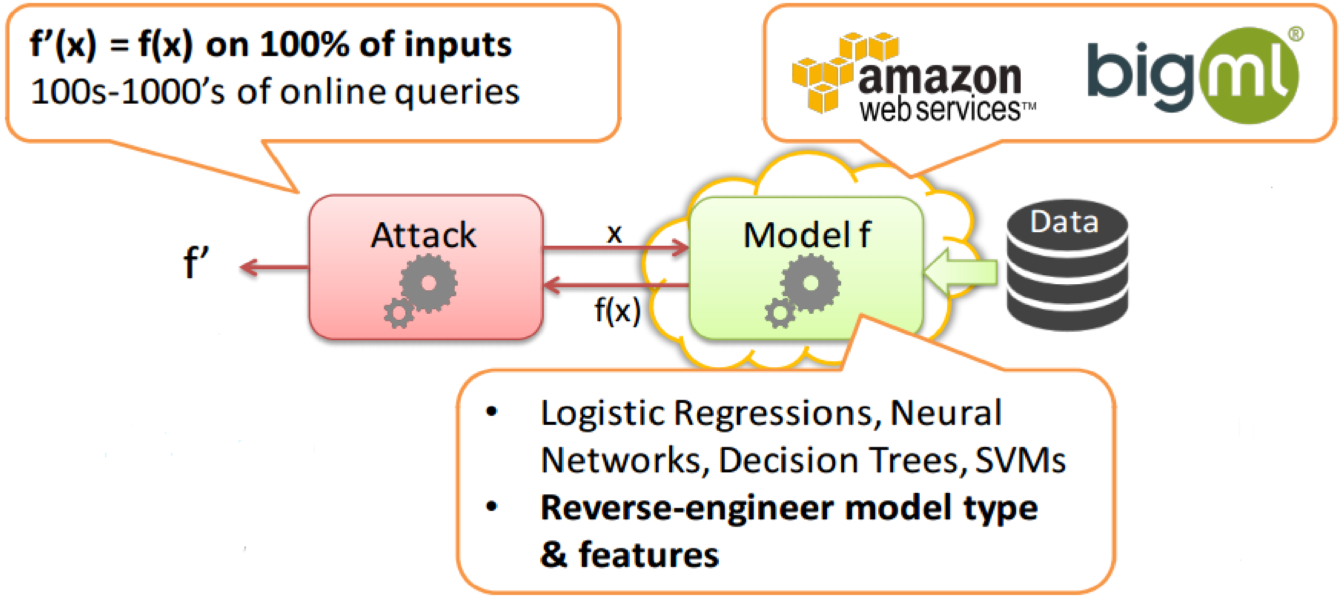
Security and Privacy Concerns



Model Extraction

Model Extraction Attack

- Model **IP ownership** - **primary source of value** for company/ service
- **Attacker's Capabilities:**
  - Access to black-box model
  - Access to query to ML model
- **Goal:** Learns close approximation, *f'*, of *f* using as few queries as possible
  - Service provider prediction APIs themselves used in attack
    - APIs return extra information – **confidence scores**



* Tramer et.al., "Stealing Machine Learning Models via Prediction APIs.", 2016.

Extraction Countermeasures

- **Restrict information** returned
  - E.g. do not return confidence scores
  - **Rounding** – return approximations where possible

- **Strict** query **constraints**
  - E.g. disregard incomplete queries

- **Ensemble methods**
  - Prediction = aggregation of predictions from multiple models
  - Might still be susceptible to *model evasion* attacks

- Prediction API minimization is not easy
  - API should still be useable for legitimate applications

* Tramer et.al., "Stealing Machine Learning Models via Prediction APIs.", 2016.

Model Inversion

Training Data Confidentiality

- **Training data** is **valuable** and **resource-intensive** to obtain
  - Collection of **large datasets**
  - Data **annotation** and **curation**
  - Data **privacy** in critical applications like healthcare
- Ensuring training data **confidentiality** is **critical**

QUARTZ

Waymo's driverless cars have logged 10 million miles on public roads

By Jane C. Hu • October 10, 2018

The New York Times

Sloan Kettering's Cozy Deal With Start-Up Ignites a New Uproar

By Charles Ornstein and Katie Thomas

Sept. 20, 2018

Model Inversion Attack

- Extract **private and sensitive inputs** by leveraging the outputs and ML model.
- **Optimization goal**: Find inputs that maximize returned confidence value to infer sensitive features or complete data points from a training dataset
- **Attacker's Capabilities:**
  - Access to Black-box or White-box model
  - Exploits confidence values exposed by ML APIs

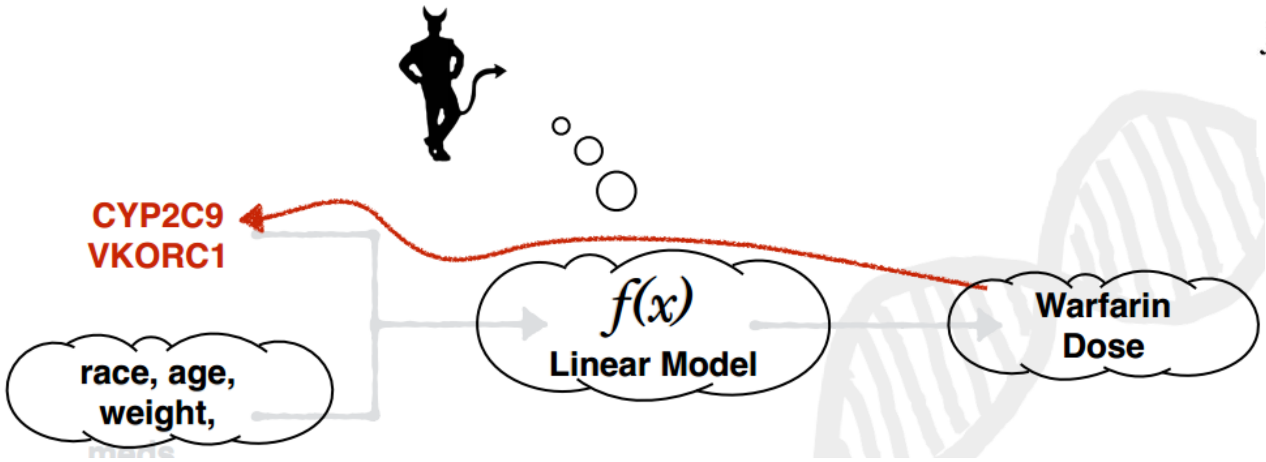

An image recovered using a new model inversion attack (left) and a training set image of the victim (right). The attacker is given only the person's name and access to a facial recognition system that returns a class confidence score.

* Fredrikson et.al., "Model Inversion Attacks that Exploit Confidence Information and Basic Countermeasures.", 2015

Privacy of the Training or Test Data

- **Attacker's capabilities:** Access to query to ML model
- Extracting patients' genetics from *pharmacogenetic dosing models*
  - **Queries** using *known information* – E.g. demographics, dosage
  - **Guess** unknown information and check model's response - assign weights
  - Return guesses that produce **highest confidence score**

age	height	weight	race	history	vkorc1	cyp2c9	dose
50-60	176.2	185.7	asian	cancer	A/G	*1/*3	42.0



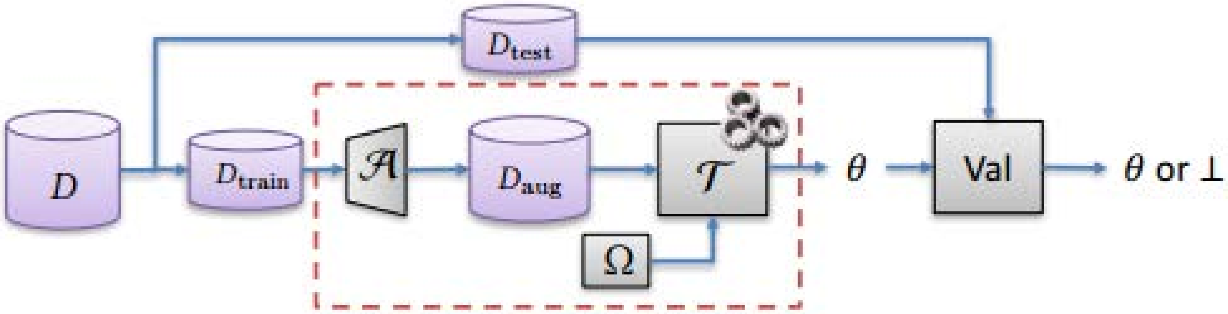
	age	height	weight	race	history	vkorc1	cyp2c9	dose		
f(x)	50-59	176.53	144.2	white				42.0	49.7	p=0.23
	50-59	176.53	144.2	white				42.0	42.0	p=0.75
	50-59	176.53	144.2	white				42.0	39.2	p=0.01

	age	height	weight	race	history	vkorc1	cyp2c9	dose		
f(x)	50-59	176.53	144.2	white	Cancer	A/G	*1/*1	42.0	49.7	p=0.23
	50-59	176.53	144.2	white	Heart	G/G	*1/*3	42.0	42.0	p=0.75
	50-59	176.53	144.2	white	Diabetes	A/A	*2/*3	42.0	39.2	p=0.01

Fredrikson et.al., "Privacy in Pharmacogenetics: An End-to-End Case Study of Personalized Warfarin Dosing ", 2014.

Training Data Tampering

- ❖ **Attacker's goal:** Leaking information about training data by modifying training algorithm
- ❖ **Attacker's capabilities:**
  - o Provides tampered APIs that remembers too much information
  - o Access to Black-box model
    - ▪ Extending the training dataset with additional synthetic data
  - o Access to white-box model
    - ▪ Encoding sensitive information about training data in model parameters



A typical ML training pipeline. Data *D* is split into training set *D*train and test set *D*test.
The dashed box indicates the portions of the pipeline that may be controlled by the adversary
*Song et.al. "Machine Learning Models that Remember Too Much", 2017.

Inversion Countermeasures

- Incorporate model inversion metrics to increase robustness
  - **Identify** sensitive features
  - Analyze **effective feature placement** in algorithm – E.g. sensitive features at top of a *decision tree* maintain accuracy while preventing *inversion* from performing better than guessing
  - **Approximate**/ **Degrade** confidence score output – E.g. decrease gradient magnitudes
    - Works against non-adapting attacker

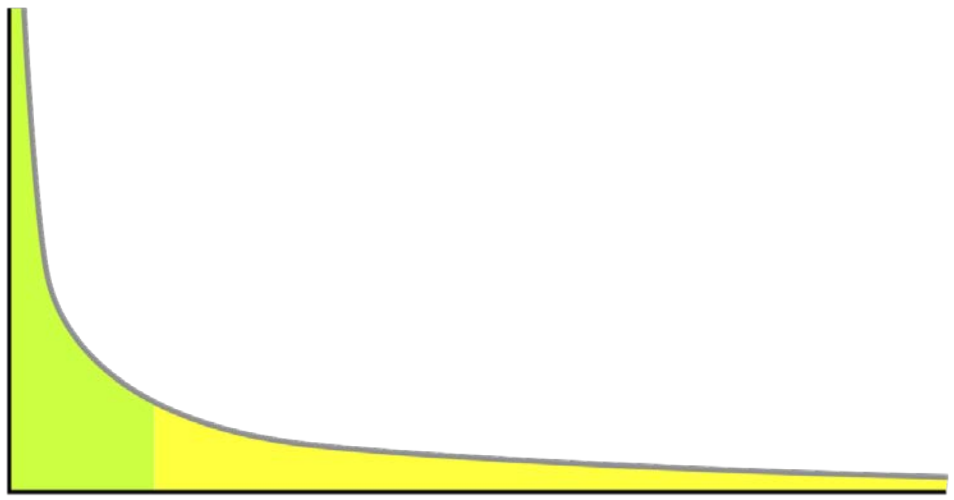
- Ensuring privacy needs to be balanced against usability
  - **Privacy Budget**
- **Differential Privacy** mechanisms using added noise
  - Might prevent model inversion
  - Risk of compromising legitimate results in critical applications

A Countermeasure Against Model Inversion

- Based on the injection of noise with long-tailed distribution to the confidence levels.

- The small randomness added to the confidence information **prevents convergence** for model inversion attack, but does not affect functionality

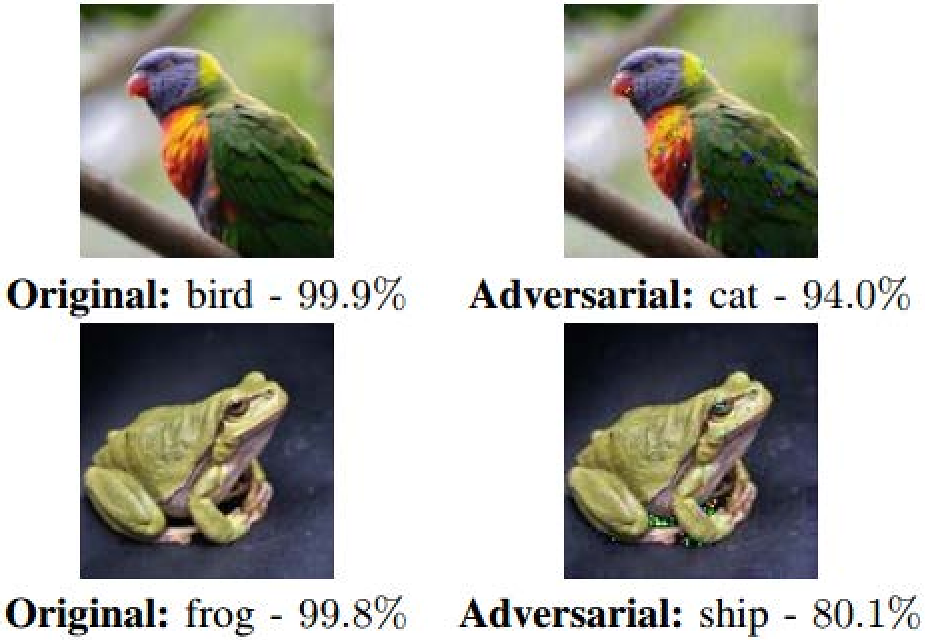
- **No modification or re-training** of model required



Noise distribution
long tail

Targeted Misclassification

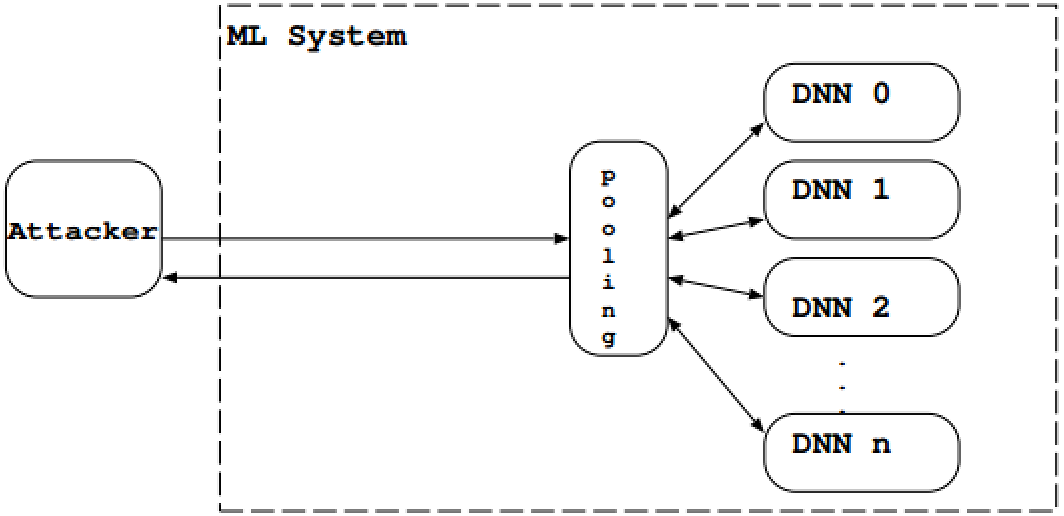
- Misclassification to a target class
  - Visually same-looking images are classified differently
  - Target adversarial examples are obtained using our numerical implementation of gradient descent based attack.



Adversarial examples. Original images (left) and the target adversarial examples (right). Below each image is the classification and confidence returned by the ResNet CIFAR-10 Image Classifier.

A Countermeasure Against Targeted Misclassification

- Varying the order of the training
  - Different models which offer the same classification accuracy, yet they are different numerically.
- An ensemble of such models
  - Allows to randomly switch between these equivalent models during query which further blurs the classification boundary.



Workflow description of adversarial attacks with Multi-Model Defense applied.

Adversarial attack performed on an image originally classified as *deer*, where the target class *is truck*. With Noise-Injection defense, the attack does not converge and ends up degrading the original image.

Model Poisoning and Evasion

Model Poisoning and Evasion Attacks

- Ensuring Integrity of a Machine Learning model is difficult
  - Dependent on **quality** of *training*, *testing* datasets
    - Coverage of *corner cases*
    - Awareness of *adversarial examples*
  - **Model sophistication** – E.g. small model may produce incorrect outputs
  - **Lifetime management** of larger systems
    - Driverless cars will need constant updates
    - Degradation of input sensors, training data pollution
- Adversarial examples may be **Transferable** *
  - Example that fools Model A might fool Model B
  - Smaller model used to find examples quickly to target more sophisticated model

Papernot et. al., "Transferability in Machine Learning: from Phenomena to Black-Box Attacks using Adversarial Samples", 2016

Model Poisoning and Evasion Attacks

- **Adversary capabilities:** Causing misclassifications of attacks to appear as normal (false positives/ negatives)
  - Attack on training phase: **Poisoning (Causative) Attack**: Attackers attempt to learn, influence, or corrupt the ML model itself
    - Compromising data collection
    - Subverting the learning process
    - Degrading performance of the system
    - Facilitating future evasion
  - Attack on testing phase: **Evasion (Exploratory) Attack**: Do not tamper with ML model, but instead cause it to *produce adversary selected outputs by manipulating test samples*.
    - Finding the blind spots and weaknesses of the ML system to evade it

Adversarial Detection of Malicious Crowdsourcing

- Malicious crowdsourcing, or ***crowdturfing*** used for tampering legitimate applications
  - **Real users** paid to promote malicious intentions
  - Product reviews, Political campaigns, Spam
- Adversarial machine learning attacks
  - Evasion Attack: workers evade classifiers
  - Poisoning Attack: crowdturfing admins tamper with training data


BBC

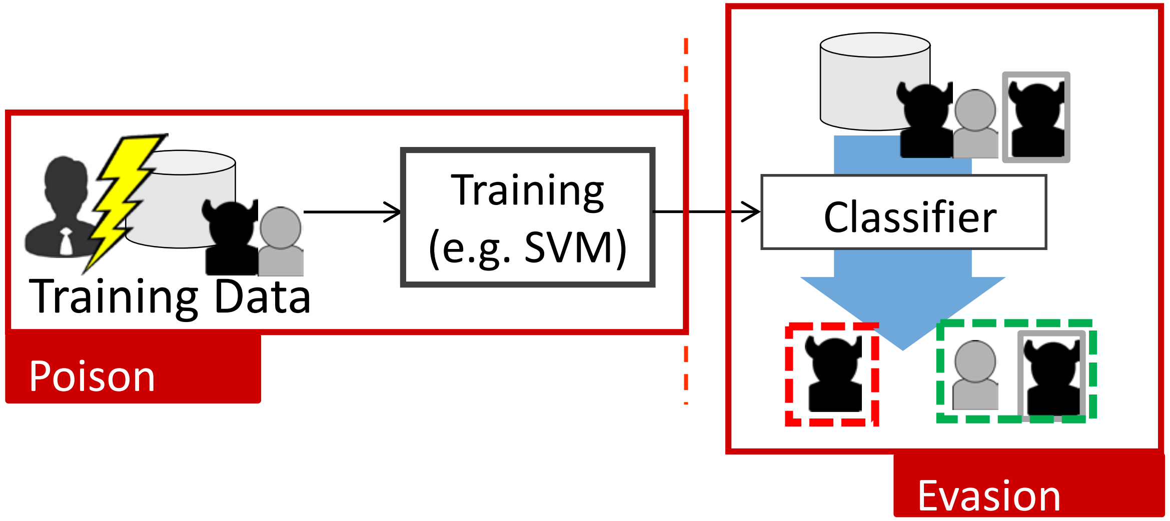
Vietnam admits deploying bloggers to support government

By Nga Pham
12 January 2013

THE VERGE

Samsung fined $340,000 for faking online comments

By Aaron Souppouris | Oct 24, 2013, 7:47am EDT

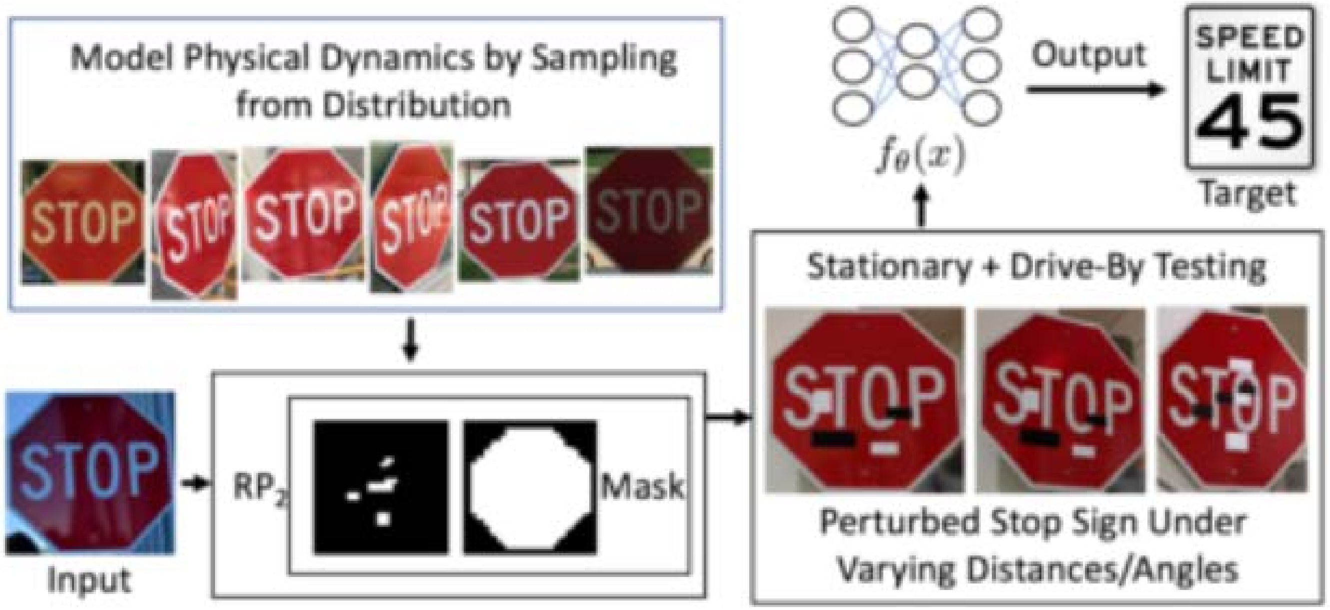


Wang et.al., “Man vs. Machine: Adversarial Detection of Malicious Crowdsourcing Workers ”, 2014

Physical Perturbations

- Adversarial perturbations detrimentally affect Deep Neural Networks (DNNs)
  - Cause misclassification in critical applications
  - Requires some knowledge of DNN model
  - Perturbations can be robust against noise in system

- Defenses should not rely on physical sources of noise as protection
  - Incorporate adversarial examples
  - Restrict model information/ visibility
  - **DNN Distillation** – transfer knowledge from one DNN to another
  - **Gradient Masking**



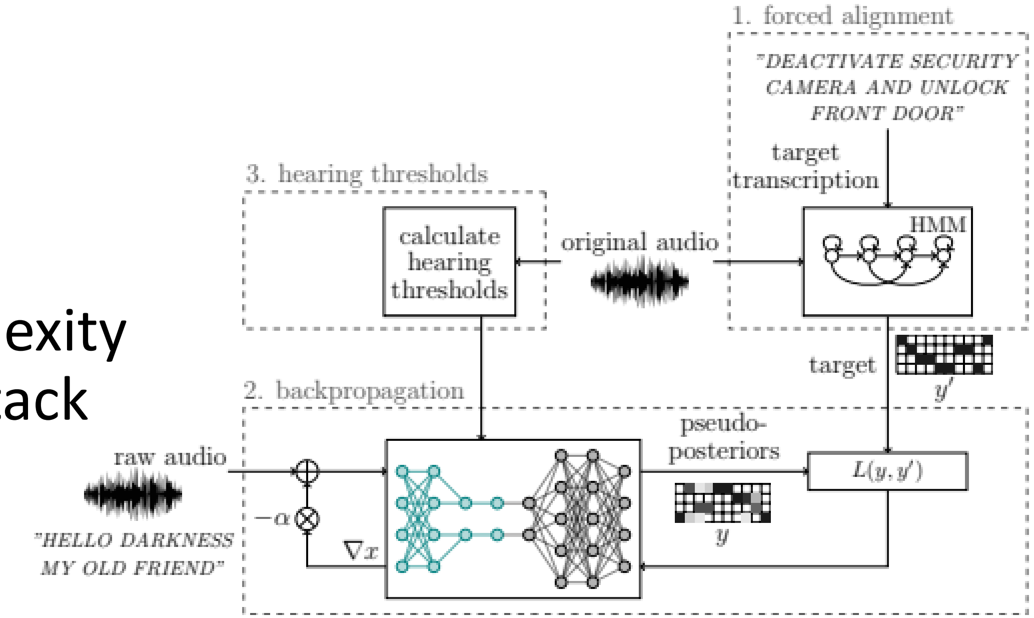
Eykholt et.al., "Robust Physical-World Attacks on Deep Learning Visual Classification", 2018.

Papernot et.al., "Distillation as a Defense to Adversarial Perturbations against Deep Neural Networks", 2015.

Adversarial Attacks Against ASR DNNs

- Automatic Speech Recognition (**ASR**) and Natural Language Understanding (**NLU**) increasingly popular – E.g. Amazon Alexa/ Echo
  - Complex model = **Large parameter space** for attacker to explore
- Attacker goals
  - Psychoacoustic hiding – perceived as noise by human
  - Identify and match legitimate voice features
    - Pitch, tone, fluency, volume, etc
  - Embed arbitrary audio input with a malicious voice command
  - *Temporal alignment* dependencies add complexity
  - Environment/ System *variability* can affect attack
  - Software tools like *Lyrebird* can prove useful

Lyrebird



Lea et.al., "Adversarial Attacks Against Automatic Speech Recognition Systems via Psychoacoustic Hiding", 2018

Defenses Against AML

- Evasion
  - Multiple classifier systems (B. Biggio et al., IJMLC 2010)
  - Learning with Invariances (SVMs)
  - Game Theory (SVMs)
- Poisoning
  - Data sanitization (B. Biggio et al., MCS, 2011)
  - Robust learning (PCA)
  - Randomization, information hiding, security by obscurity
- Randomizing collection of training data (timings / locations)
  - using difficult to reverse-engineer classifiers (e.g., MCSs)
  - denying access to the actual classifier or training data
  - randomizing classifier to give imperfect feedback to the attacker (B. Biggio et al., S+SSPR 2008)

Towards Robust ML Model

Future Research Areas

- Complexity of Machine Learning itself an issue
  - New attacks models constantly emerging – *timely detection* critical
  - Generation and incorporation of **Adversarial Examples**
  - **Data Privacy** is crucial to enhance ML security
    - *Differential Privacy* has tradeoffs
    - *Homomorphic Encryption* still nascent
- Security introduces overhead and can affect performance
  - **Optimizations** needed to ensure ML effiency
- Tools to increase robustness of Machine Learning need research
  - *Unlearning, re-learning*
  - *ML Testing*
  - *Sensitivity Analysis*

Unlearning and Re-learning

- Ability to **unlearn** is gaining importance
  - **Pollution** attacks or carelessness – *Mislabeling* and *Misclassification*
    - Large changing datasets difficult to maintain
    - Anomaly detection not enough
  - **EU GDPR** regulations – **Privacy**
  - **Completeness** and **Timeliness** are primary concerns *
  - **Statistical Query Learning*** and **Causal Unlearning**** proposed in literature
  - Suitable for **small deletions**
- **Re-learning** or **Online learning**
  - Faces similar issues to un-learning
  - Can be very slow
  - More suitable for large amounts of deletions or new information

* Yinzhi Cao, "Towards Making Systems Forget with Machine Unlearning", 2015
** Cao *et. al.*, "Efficient Repair of Polluted Machine Learning Systems via Causal Unlearning", 2018

Sensitivity Analysis

- Study of how the uncertainty in the output of a system can be attributed to different sources of uncertainty in its inputs
  - ML feature extraction sensitivity analysis well-researched

- Detection of **biases** in training/test datasets is crucial *
  - Model accuracy dependent on datasets used – ***real-world*** performance can be different
    - Datasets can have **expiration dates**
    - **Privacy** issues can render datasets incomplete
  - Identify training datasets which **generalize** better
  - Study sensitivity of ML accuracy to change in datasets

* Sanders, Saxe, “Garbage In, Garbage Out - How Purportedly Great ML Models Can Be Screwed Up By Bad Data”, 2017

Thank you