ÉCOLE POLYTECHNIQUE
FÉDÉRALE DE LAUSANNE

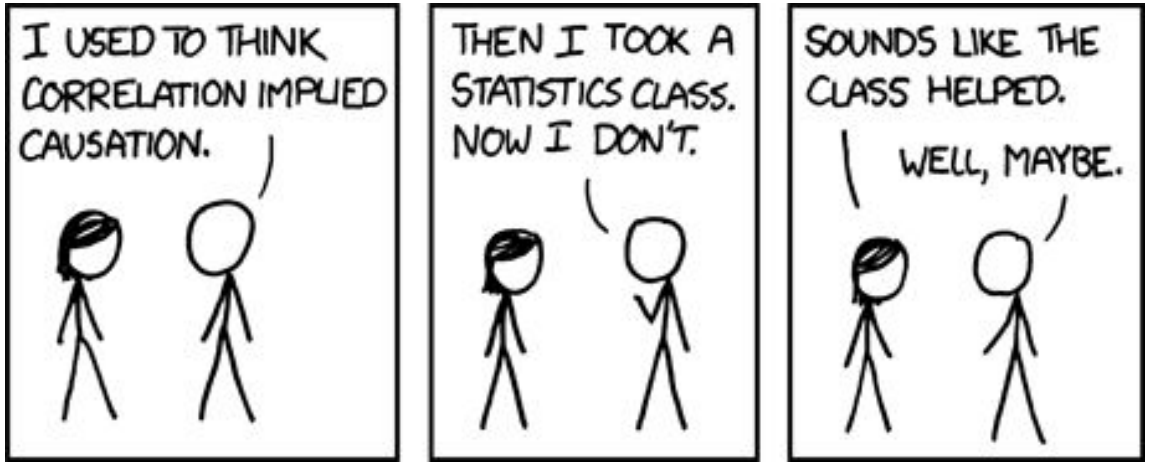
Machine Learning

COM-402: Information Security and Privacy

(slide credits: Philipp Jovanovic)

Outline

- **Overview on Machine Learning**
- The Dark Sides of Machine Learning
- Attacking and Defending Machine Learning Systems
- Conclusion



Source: https://xkcd.com/552/

Machine Learning (ML)

ÉCOLE POLYTECHNIQUE
FÉDÉRALE DE LAUSANNE

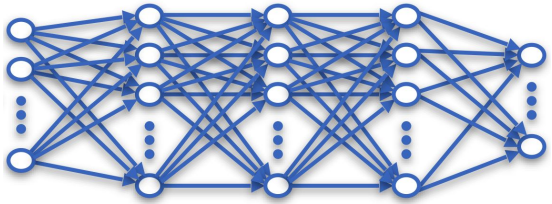
Definition (Wikipedia)

Machine learning *[...] gives "computers the ability to learn without being explicitly programmed" [and] [...]* ***explores the study and construction of algorithms that can learn from and make predictions on data*** *– such algorithms overcome following strictly static program instructions by making data-driven predictions or decisions, through building a model from sample inputs. Machine learning is employed in a range of computing tasks where designing and programming explicit algorithms with good performance is difficult or unfeasible; example* ***applications*** *include* ***email filtering***, ***detection of network intruders or malicious insiders*** *working towards a data breach,* ***optical character recognition*** *(OCR),* ***learning to rank***, ***computer vision***, *etc.*

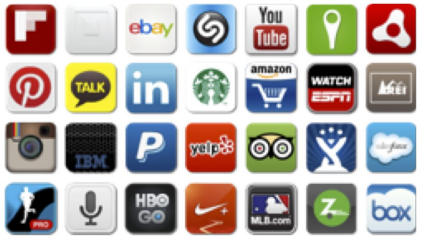
Machine Learning

User data

Machine learning



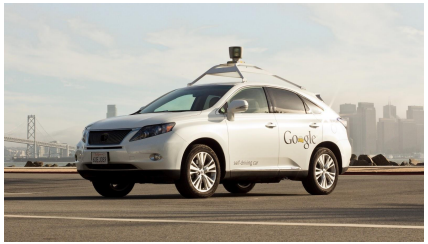
Services



Machine Learning Is Becoming Ubiquitous

Ten Use Cases of Machine Learning

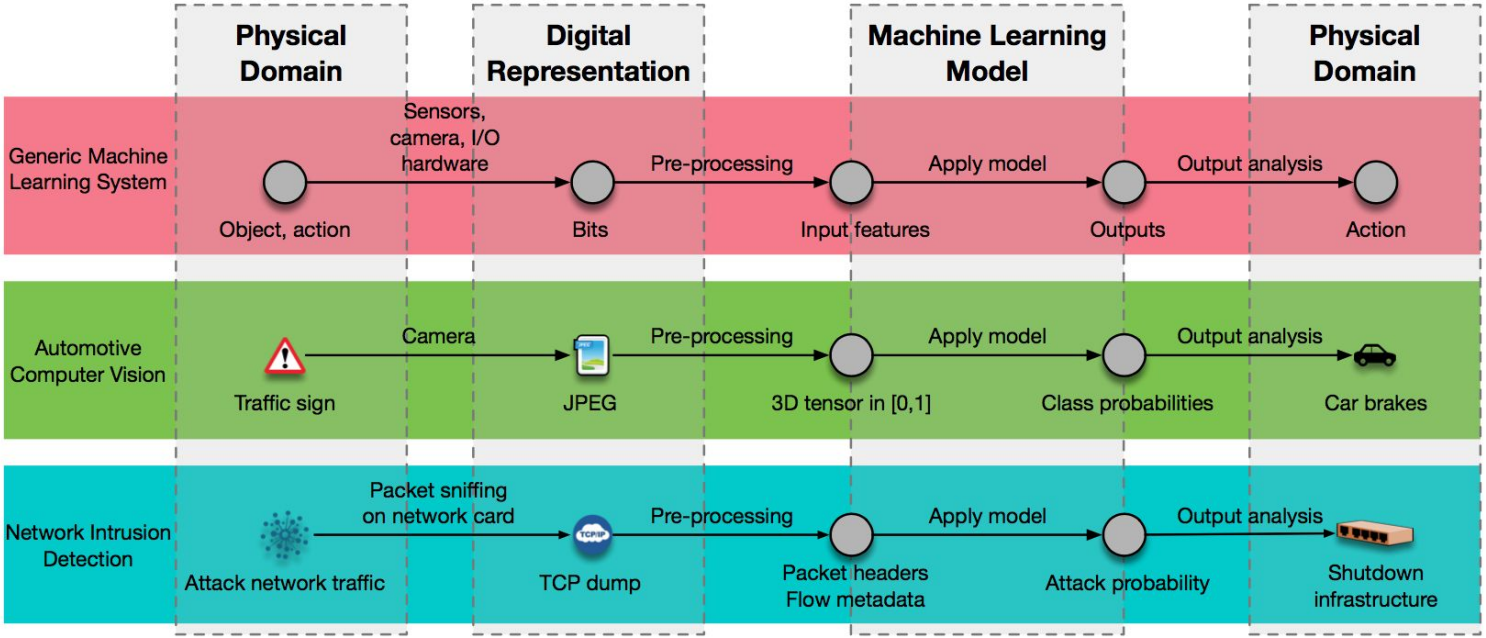
1. Data security
2. Personal security
3. Financial trading
4. Healthcare
5. Marketing personalization
6. Fraud detection
7. Recommendations
8. Online search
9. Natural language processing (NLP)
10. Smart (autonomous) cars



NETFLIX

Source: https://www.forbes.com/sites/bernardmarr/2016/09/30/what-are-the-top-10-use-cases-for-machine-learning-and-ai

Machine Learning Overview

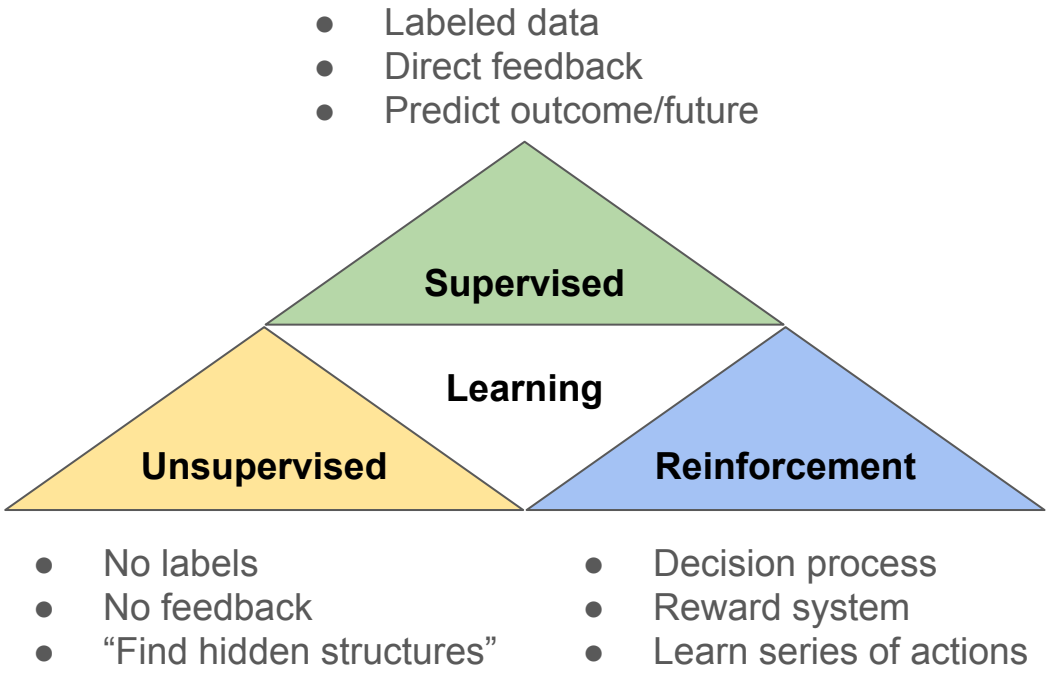


ML main phases: (1) training / learning and (2) inference

Source: https://arxiv.org/pdf/1611.03814.pdf

Machine Learning Taxonomy

Machine learning can be separated into 3 main categories



Machine Learning Taxonomy

Supervised Learning

- Given: Training data $\{(x_1,y_1),...,(x_n,y_n)\} \in X \times Y$
  - X: Input space
  - Y: Output space
  - x_i: Feature vector
  - y_i: Output label (= class)
- Goal: Infer a function $f: X \rightarrow Y$ that matches the training data
- Types:
  - Classification (Y categories)
  - Regression ($Y = \mathbb{R}$)
- Usage examples:
  - Object recognition in images
  - Machine translation
  - Spam filtering

Machine Learning Taxonomy

Unsupervised Learning

- Given: training data $x_1,...,x_n$ without labels
- Goal: infer a function f that matches the data
- Types:
  - Clustering (according to a similarity metric)
  - Dimensionality reduction (project data into lower dimensional subspaces)
  - Model pre-training
- Usage examples:
  - Cluster analysis to find hidden patterns or group data

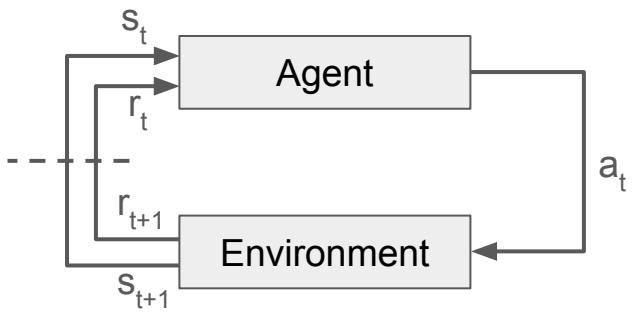
Google’s AI Learns to Identify Cats

- Google study in 2012
- Fed YouTube videos to a deep net
- 74.8% accuracy to identify cats
- Jeff Dean, lead-researcher: *“We never told it during the training, ‘This is a cat.' It basically invented the concept of a cat on its own.”*

Source: https://www.wired.com/2012/06/google-x-neural-network/

Machine Learning Taxonomy

Reinforcement Learning



- No data given upfront
- At time t:
  - Agent performs an action a_t
  - Environment generates a new state s_t and reward r_t
- Goal: Using (a_t,s_t,r_t), derive policy for selecting actions that maximizes long-term rewards
- Usage examples:
  - Control problems (e.g., robot movement, space flight)
  - Games (e.g., Atari, AlphaGo)

AlphaGo Beats Human Go Champions

- AlphaGo uses reinforced learning (in combination with many other techniques)
- AlphaGo beats
  - Fan Hui (European champion) 5 : 0 in 2016
  - Lee Se-dol (2nd in world ranking) 4 : 1 in 2017
- Why is Go so hard?
  - Number of legal board states: ~10^359 (chess: ~10^123, atoms in the visible universe: ~10^80)
  - Enormous branching factor of decision tree
  - Very hard to compute the optimal strategy at a given point
- Interesting insight: AlphaGo invented new effective game strategies never seen before in the entire history of Go

Google DeepMind AlphaGo

Which Learning Types?

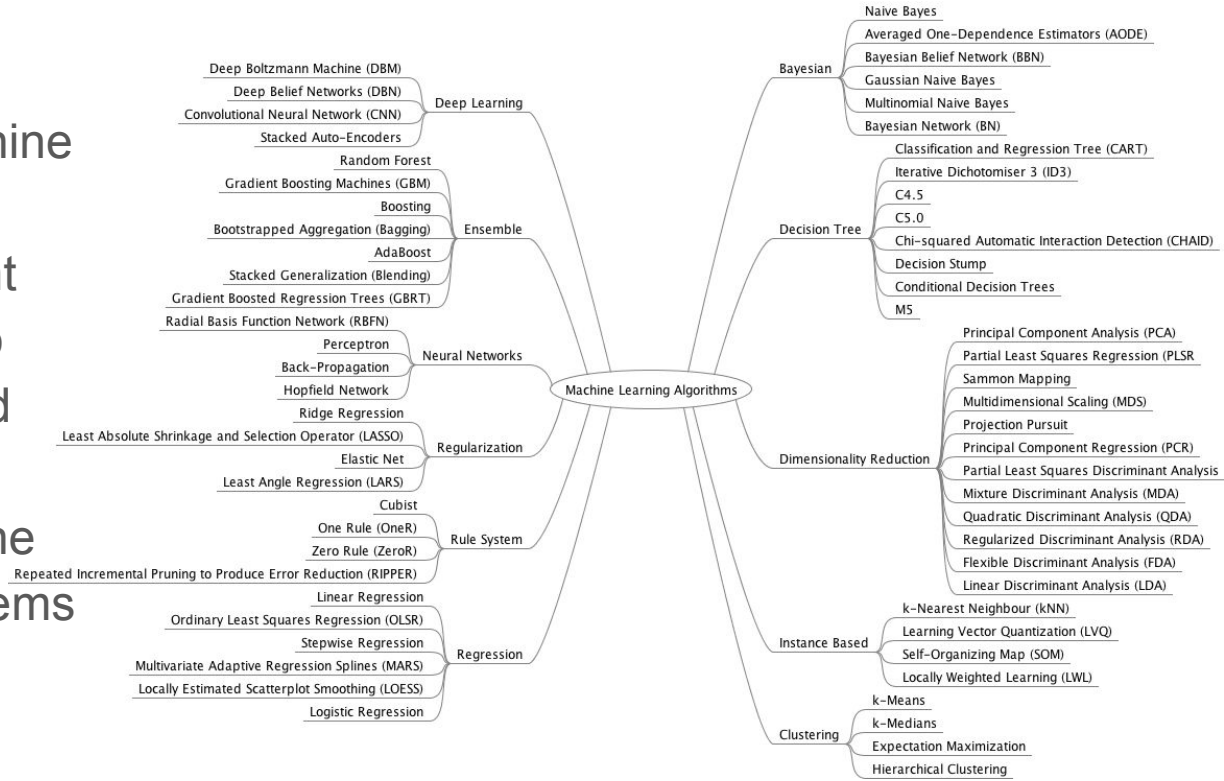
- Problem 1:
  - Training data with TCP dumps
  - Goal: Train network intrusion detection system to find network anomalies
- Problem 2:
  - Training data with TCP dumps and live metrics of system state indicators (CPU-, memory-, network-consumption, etc.)
  - Goal: Adaptively train network intrusion detection system to find network anomalies
- Problem 3:
  - Training data with benign and malicious executables
  - Goal: Train neural network to analyse unknown executables

Machine Learning Algorithms

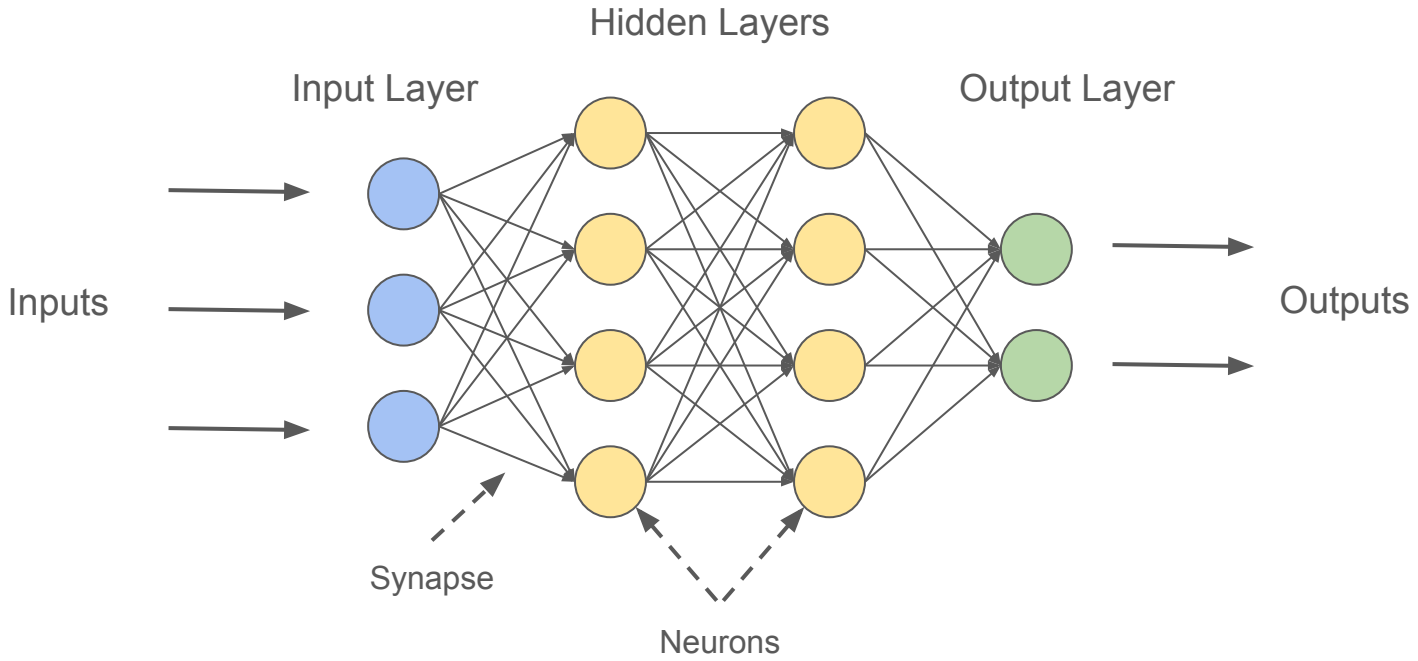
ÉCOLE POLYTECHNIQUE FÉDÉRALE DE LAUSANNE

Overview

- Countless types of machine learning algorithms
- Combinations of different types usually required to solve complex real-world problems
- Deep learning often at the core of modern ML systems



Neural Networks

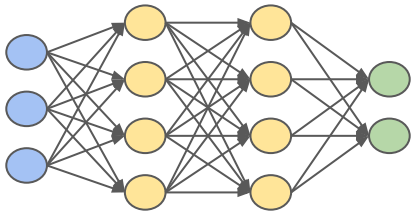


Neural networks are the core structure for deep learning.

Machine Learning versus Deep Learning

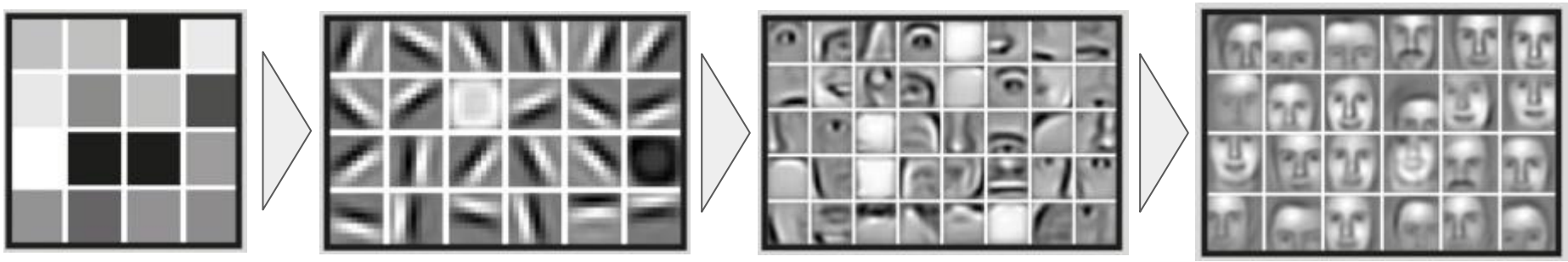
	Machine Learning	Deep Learning
Number of hidden layers?	< 5	10s to 100s
Size of training dataset?	Small	Large*
Need to know/choose labels?	Yes	No
Required training time?	Short	Long

*Rule of thumb: For n input parameters you should have about n^2 training data points



Deep Neural Networks

DNN use layers of increasingly complex rules to categorize complicated shapes (such as faces).



Layer 1:
The computer identifies pixels of light and dark.

Layer 2:
The computer learns to identify edges and simple shapes.

Layer 3:
The computer learns to identify more complex shapes and objects (eyes, noses, etc.).

Layer 4:
The computer learns which shapes and objects can be used to define a human face.

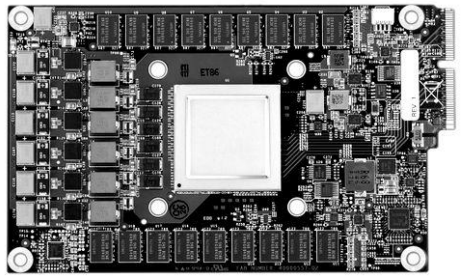
Source: Andrew Ng

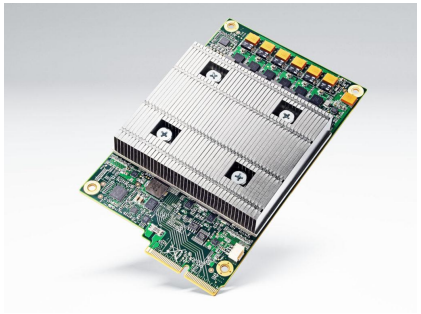
Google’s Tensor Processing Unit (TPU) Chip

ÉCOLE POLYTECHNIQUE FÉDÉRALE DE LAUSANNE

Overview

- Custom chip for deep learning inference
- Excels at matrix multiplications
- Outperforms standard chips by factors between 30 to 80 in the Tera Operations Per Second (TOPS) Per Watt metric

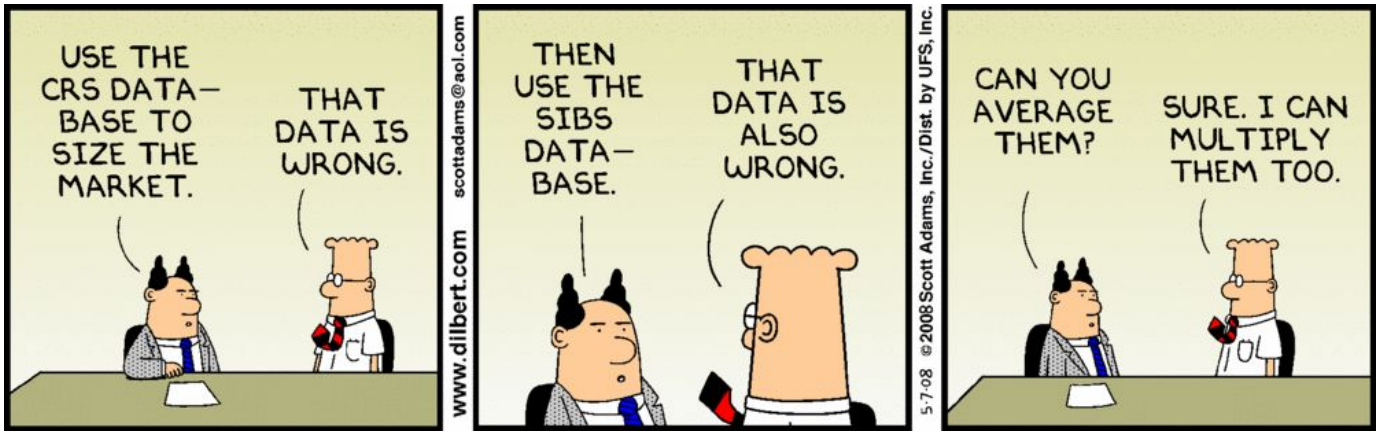




Source: https://www.nextplatform.com/2017/04/05/first-depth-look-googles-tpu-architecture/amp/

Outline

- Overview on Machine Learning
- **The Dark Sides of Machine Learning**
- Attacking and Defending Machine Learning Systems
- Conclusion



Source: http://dilbert.com/strip/2008-05-07

The Dark Sides of Machine Learning

Common Belief

“Algorithmic decisions tend to be fair, because math is about equations and not about skin color”

The Dark Sides of Machine Learning

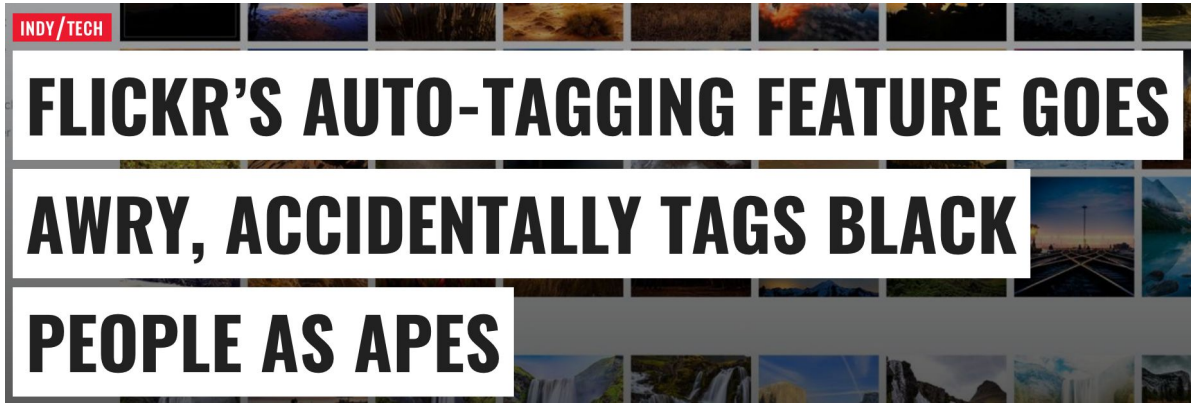
Algorithmic Bias – Google Ads

Google showed high-paying executive job ads **1,852 times to the male group** — but just **318 times to the female group**.

Source: https://www.independent.co.uk/life-style/gadgets-and-tech/news/googles-algorithm-shows-prestigious-job-ads-to-men-but-not-to-women-10372166.html

The Dark Sides of Machine Learning

Algorithmic Bias – Flickr's Image Recognition



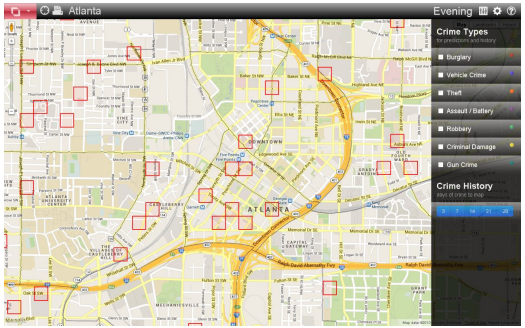
The site's tool was built to help people easily identify features of pictures — but has run into problems as it learns.

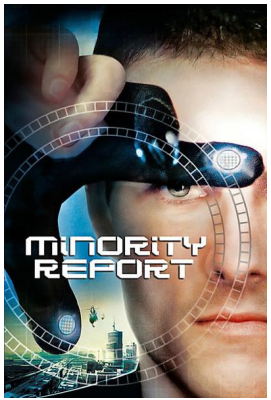
Source: https://www.independent.co.uk/life-style/gadgets-and-tech/news/flickr-s-auto-tagging-feature-goes-awry-accidentally-tags-black-people-as-apes-10264144.html

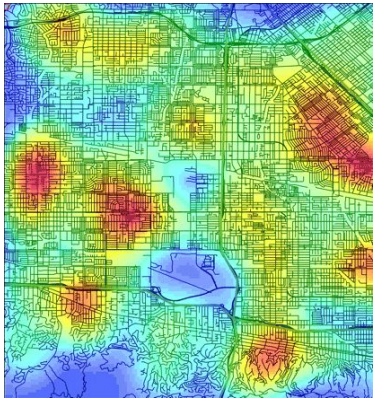
The Dark Sides of Machine Learning

Algorithmic Bias – PredPol

- Tool to analyse crime data and forecast which people and places are most at risk for future crimes
- Study on Oakland (drug) crime data: mostly black people neighbourhoods are “recommended”
- Reality:
  - Drug crime is everywhere
  - Police only finds it where they are looking
  - PredPol reinforces bad police habits and gives them a tool to justify actions





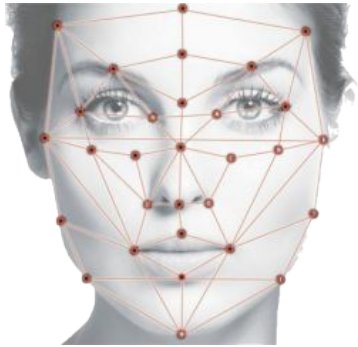


Source: http://uk.businessinsider.com/predictive-policing-discriminatory-police-crime-2016-10?r=US&IR=T

The Dark Sides of Machine Learning

FindFace.ru

- Face recognition app
- Photograph people and get their social network profile on VKontakte.ru
- 70% reliability
- FindFace owner's vision:
  - Revolutionize dating
  - Walk by shops and receive targeted ads
  - Help law enforcement to track down people: "If the [Russian] FSB were to get in touch, of course we'd listen to any offers they had."



Source: https://www.theguardian.com/technology/2016/may/17/findface-face-recognition-app-end-public-anonymity-vkontakte

The Dark Sides of ML – Uber Incidents

Greyball Program

- Tool to protect Uber drivers from attacks (location obfuscation, etc.)
- Later used to identify and dodge authorities (show ghost drivers, etc.)
- Uber statement:

"This program denies ride requests to fraudulent users who are violating our terms of service, whether that's people aiming to physically harm drivers, competitors looking to disrupt our operations, or opponents who collude with officials on secret 'stings' to entrap drivers."

Source: https://www.nytimes.com/2017/03/03/technology/uber-greyball-program-evade-authorities.html?_r=0

The Dark Sides of ML – Uber Incidents

Fare Equivocation

- Show different prices to drivers (fare--) and customers (fare++)

Source: https://bgr.com/2017/04/07/uber-class-action-suit-driver-cheating-software/

Hell Program

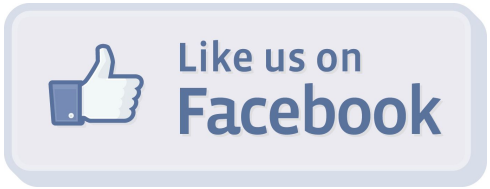
- Track Lyft drivers
- Number of drivers available
- What are their prices
- Are they working for Uber too

Source: https://techcrunch.com/2017/04/24/uber-hell-lawsuit/

The Dark Sides of Machine Learning

Predicting Human Behavior Through Facebook Likes

- Study from 2013 by Kosinski et al.
- Basis: 58000 participants, Facebook likes, demographic profiles, psychometric tests
- Forecast automatically and accurately personal attributes like:
  - Age, gender, sexual orientation, ethnicity
  - Religious and political views
  - Personality traits, intelligence, happiness
  - Use of addictive substances, parental separation
- Results:
  - 70 "likes" were enough to outdo what a person's friends knew,
  - 150 what their parents knew, and
  - 300 "likes" what their partner knew.



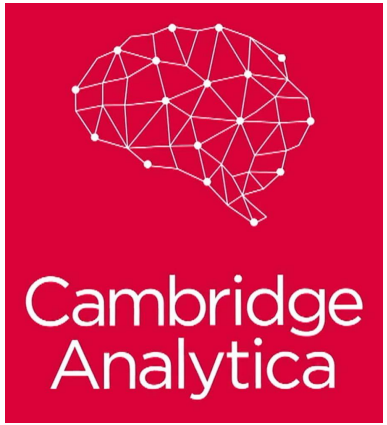
➔ Allows to ad-target people on an individual level and not just on the group level

Source: http://www.pnas.org/content/110/15/5802.full

The Dark Sides of Machine Learning

Cambridge Analytica

- Data science company selling “Behavioral Microtargeting” (likely based on Kosinski’s methodologies)
- Offspin of British SCL Group known for involvement “in military disinformation campaigns, social media branding, and voter targeting”
- Hired by pro-Brexit and Trump’s presidential teams to help in their campaigns
- How much did CA help to influence public opinion in these (and other) cases?

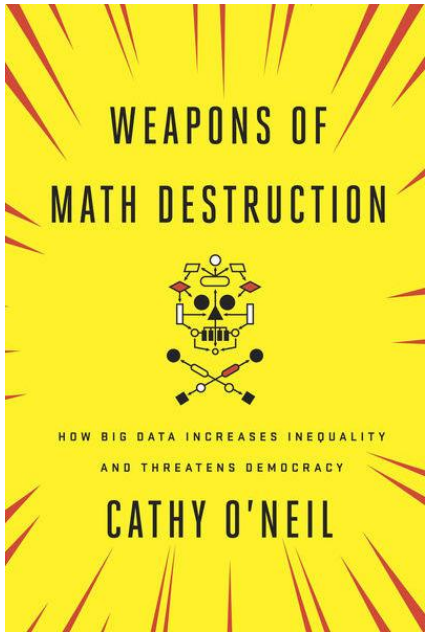


Source: https://motherboard.vice.com/en_us/article/how-our-likes-helped-trump-win

The Dark Sides of Machine Learning

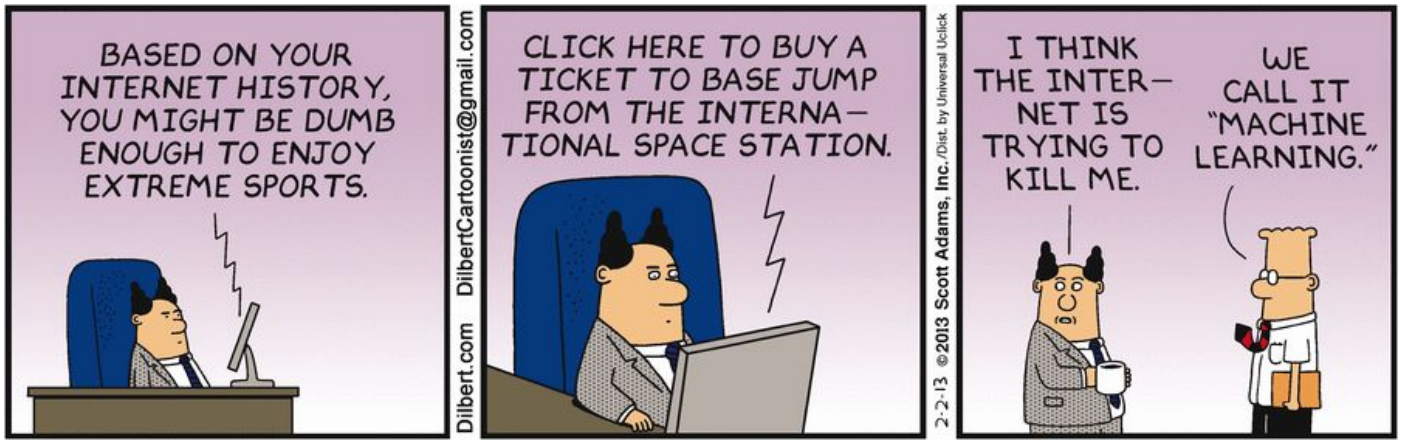
Lessons

- ML is a powerful tool with many applications
- ML reinforces discrimination (esp. when trusted blindly)
  - Models incorporate “social bias” of training data
    even if the bias is only implicitly present
  - Data about minorities is by definition proportionally scarcer
  - Statistical patterns for majority might be invalid
    within a minority group
- ML needs to be treated extra carefully if decisions could affect humans



Outline

- Overview on Machine Learning
- The Dark Sides of Machine Learning
- **Attacking and Defending Machine Learning Systems**
- Conclusion



Source: http://dilbert.com/strip/2013-02-02

When AI's Do Unintended Things …

CIRCUIT BREAKER \ TECH \ AMAZON

Amazon's Alexa started ordering people dollhouses after hearing its name on TV

Check your settings

by Andrew Liptak | @AndrewLiptak | Jan 7, 2017, 5:52pm EST

One recent instance occurred in Dallas, Texas earlier this week, when a six-year-old asked her family's new Amazon Echo "can you play dollhouse with me and get me a dollhouse?" The device readily complied, ordering a KidKraft Sparkle mansion dollhouse, in addition to "four pounds of sugar cookies." The parents quickly realized what had happened and have since added a code for purchases. They have also donated the dollhouse a local children's hospital.

Source: https://www.theverge.com/2017/1/7/14200210/amazon-alexa-tech-news-anchor-order-dollhouse

… And When They Really Go Haywire

2016-03-23

- Microsoft Research releases Tay, a Twitter-based AI chatbot
- Goal: improve understanding of conversational language
- MSR: “Tay’s responses should become more natural over time”

2016-03-24

- Tay starts tweeting misogynistic and racist comments
- MSR shuts the bot down

Conclusion of The Tay Incident

Main insight: ML systems are highly susceptible to malicious input.

Questions

- How do we ensure security of ML-based systems?
- How do we teach ML systems using public data without incorporating the worst traits of humanity?
- How do we democratize AI power?

OpenAI (https://openai.com/)

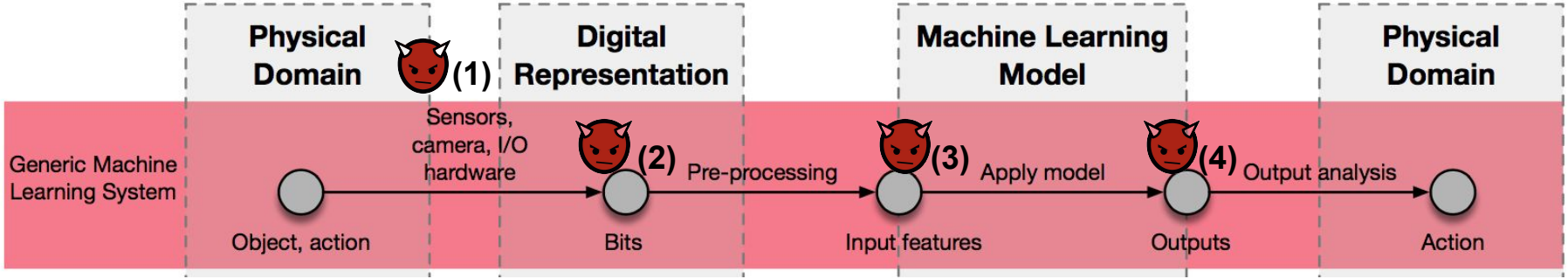
> “Discovering and enacting the path to safe artificial general intelligence.”

ML Threat Models

Components

1. Attack surface?
2. Adversarial capabilities?
3. Adversarial goals?

Threat Model: Attack Surface



Adversaries can attempt to

1. Manipulate collection of data
2. Rig processing of data
3. Corrupt the model
4. Tamper with the outputs

Threat Model: Adversarial Capabilities

"Actions and informations adversaries have at their disposal"

weaker

stronger

Training	Inference
Read	Pipeline
Injection	Model
Modification	Architecture
Logic Corruption	Parameters
	Architecture and Parameters

black box

white box

Threat Model: Adversarial Capabilities

Attacks on Training Phase

- Access training data (summary, partial, all)
  - Enables to create a substitute / surrogate / auxiliary model
  - Can be used to test inputs before submitting to victim system
- Influence training data
  - Insert adversarial inputs (injection)
  - Alter training data directly (modification)
- Logic corruption
  - Tamper with the learning algorithm
  - Very powerful adversary
  - Difficult to defend against

Threat Model: Adversarial Capabilities

Attacks on Inference Phase (Exploratory Attacks)

Black box

- Oracle model: send inputs, analyse outputs, infer vulnerabilities
- Exploits transferability property of most ML architectures

White box

- As black box plus infos on model architecture (ML algorithm and structure) and/or parameters (weights)
- Example: identify parts of feature space with high error to craft adversarial samples)

Threat Model: Adversarial Goals

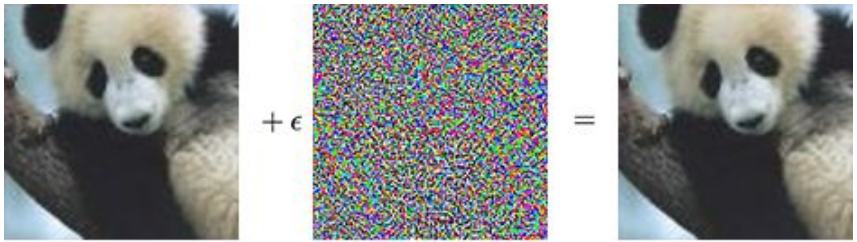
Attackers might target to exploit the following properties:

- Confidentiality / Privacy:
  - Attempt to expose model structure or parameters
  - Targeted model might be intellectual property
  - Targeted training data might be highly sensitive (e.g., clinical patient data)
- Integrity:
  - Induce outputs of adversary's choosing
- Availability:
  - Prevent access to meaningful model outputs or features

Adversarial Samples (AS) – Fooling AIs

Overview

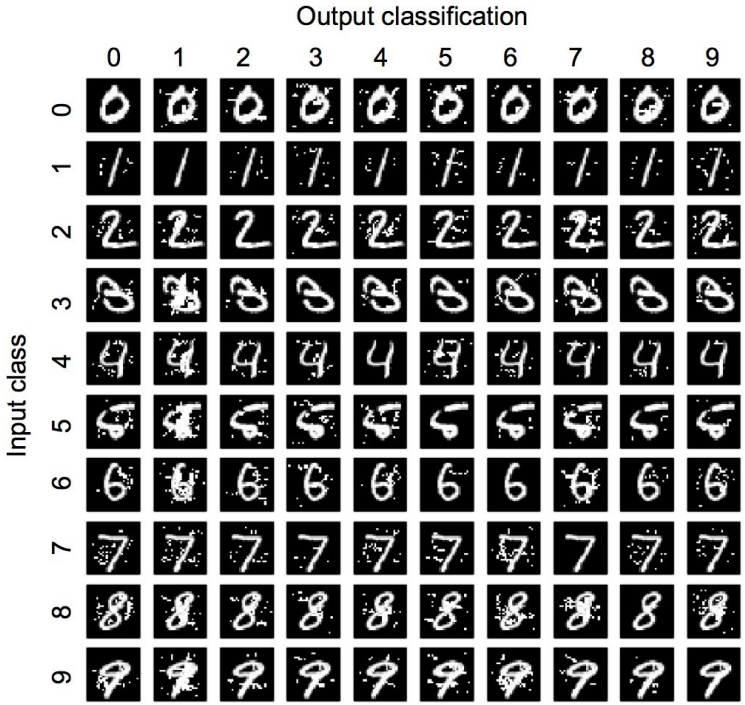
- Goal: craft inputs to trick ML models into making mistakes
- Legitimate example + malicious perturbation
- Attackers can even enforce desired outputs
- “Optical illusions for machines”



An adversarial input, overlaid on a typical image, can cause a classifier to miscategorize a panda as a gibbon.

Source: https://blog.openai.com/adversarial-example-research/

Adversarial Sample Generation



- Distortion added to input samples
- Forces DNN output to adversary selected classification
- Min distortion: 0.26%
- Max distortion: 13.78%
- Avg distortion: 4.06%

Source: https://arxiv.org/pdf/1511.07528.pdf

Adversarial Samples

Consequences

- Fraud (e.g., misinterpret digits and get more money than entitled)
- Crashing autonomous vehicles (e.g., misinterpret signs)
- Smuggle illicit/illegal content past content filters
- Manipulate biometric authentication systems to enable improper acces
- ...

Adversarial Samples – Countermeasures

Adversarial training

- Brute-force approach
- Generate lots of adversarial samples
- Train model not to be fooled by those

Defensive distillation

- Train model 2x: Feed 1st DNN output logits into 2nd DNN
- Output probabilities for decision classes, not hard decisions

➔ Still breakable by attackers with enough computational power

Adversarial Samples – Conclusion

What is the problem with current AS defenses?

- Static instead of adaptive
  - Defenses block only certain attacks
  - For circumvention attacker can just try different attack vectors
- AS transferable from one model to another

Why is it hard to find good AS defenses?

- Difficult to create a theoretical model for AS crafting process
- AS are solutions to a non-linear non-convex optimization problem
- No good theoretical tools to analyse those

➔ **Open research problem:** defenses against adaptive attackers

Data Integrity and Inference Accuracy

Theorem (Integrity)

Let *b* denote the fraction of training data modified by the adversary and *1-e* the targeted learning accuracy. Then

$$b \leq e/(1+e)$$

Example: To achieve 90% accuracy (*e=0.1*), the rate of adversarially manipulated training data hast to be less than 10%.

Membership Inference Attacks against Black Box Models

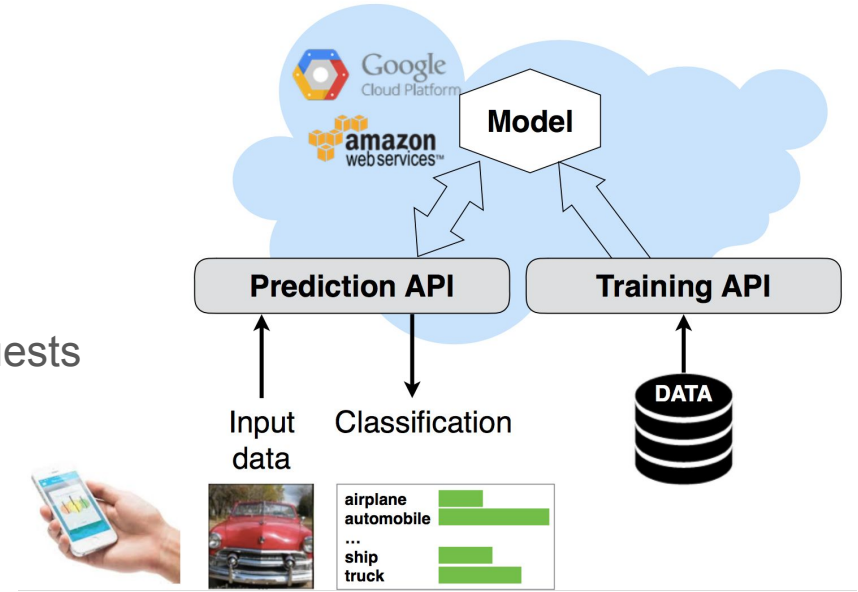
Recent research (2016) by R. Shokri et al. (https://arxiv.org/abs/1610.05820)

Starting Point

- Google, Amazon, Microsoft, etc. offer "machine learning as a service"
- Training API: send data to train a model
- Prediction API: send data classification requests

Black Box Setting

- No knowledge about model parameters
- No access to internal computations of the model
- No knowledge about underlying distribution of data

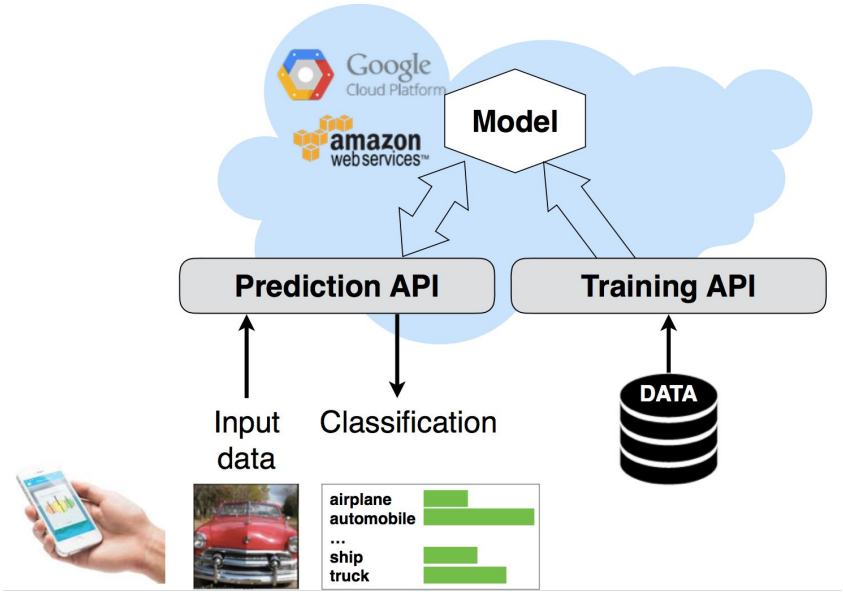


Source: Reza Shokri

Membership Inference Attacks against Black Box Models

Questions

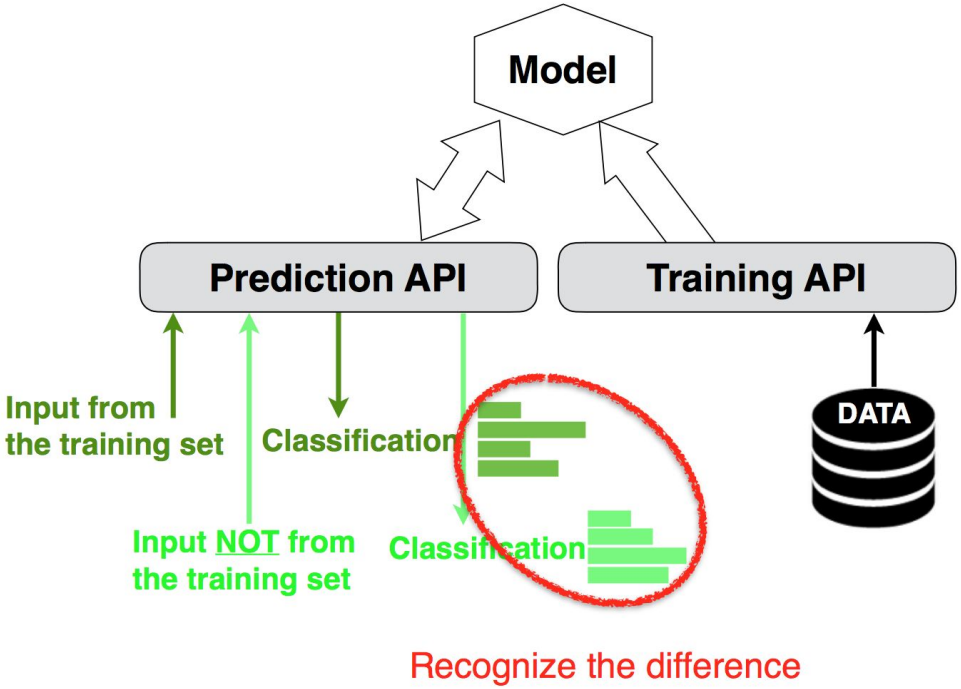
- Do model predictions leak information of training data?
- Is it possible to find out if a given data record was part of the training set (= membership inference)?



Source: Reza Shokri

Membership Inference Attacks against Black Box Models

- **Main insight:** ML models tend to overfit to their training data
- Is it possible to exploit that fact to do membership inference?
- If so, how could we approach that?



Source: Reza Shokri

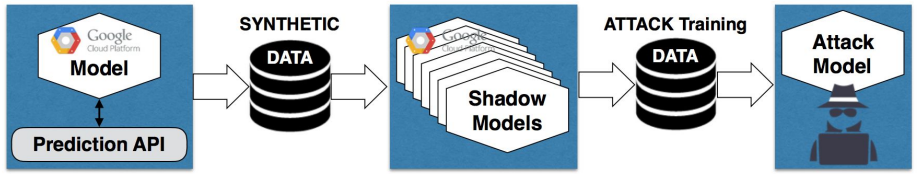
Membership Inference Attacks against Black Box Models

ÉCOLE POLYTECHNIQUE FÉDÉRALE DE LAUSANNE

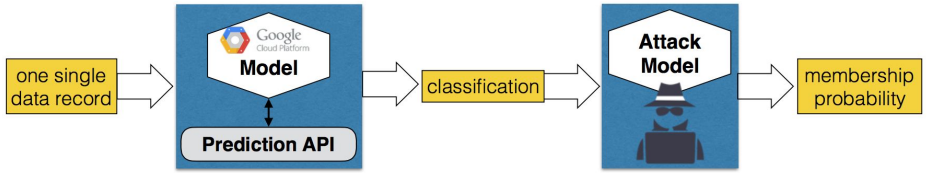
Turning ML Against Itself

- Train attack model that can distinguish target model's behavior on inputs from the training data set vs. inputs that were not in the training data set
- Transform membership inference into a classification problem
- To achieve that train multiple shadow models on synthetic (known) data that imitate target model's behavior

Constructing the Attack Model



Using the Attack Model



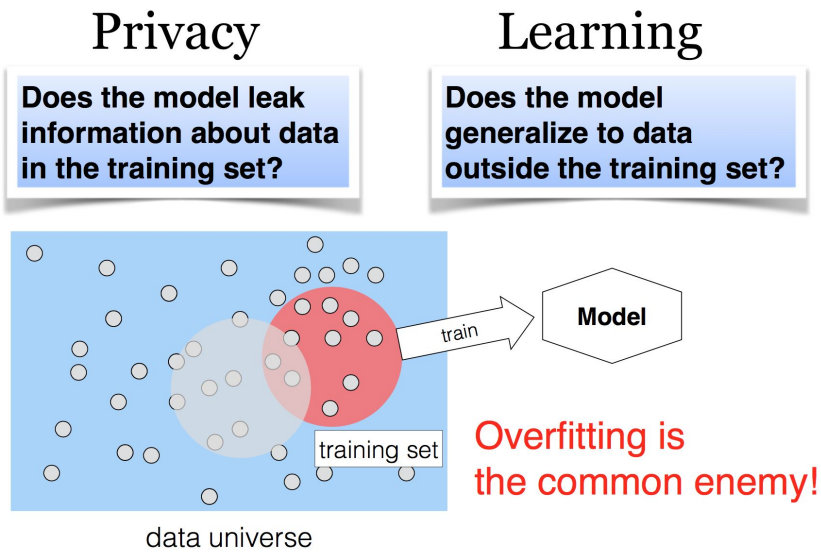
Source: Reza Shokri

Membership Inference Attacks against Black Box Models

ÉCOLE POLYTECHNIQUE FÉDÉRALE DE LAUSANNE

Conclusion

- Median membership inference accuracy against ML-as-a-service systems:
  - Google: 94%
  - Amazon: 74%
- Overfitting decreases ML-accuracy **and** -privacy
- Mitigation: use anti-overfitting techniques such as regularization, etc.



Source: Reza Shokri

Further Research

Stealing Machine Learning Models via Prediction APIs (2016)

- By F. Tramèr et al. (https://arxiv.org/abs/1609.02943)
- Black-box setting
- Shows how to efficiently extract models in ML-as-a-service scenarios

Model Inversion Attacks that Exploit Confidence Information and Basic Countermeasures

- By M. Fredrikson et al. (https://www.cs.cmu.edu/~mfredrik/papers/fjr2015ccs.pdf)
- White-box setting: leverages information from ML-as-a-service APIs
- Reverse-engineer specific values of training data parameters (e.g., faces in facial recognition systems)

Outline

- Overview on Machine Learning
- The Dark Sides of Machine Learning
- Attacking and Defending Machine Learning Systems
- **Conclusion**

Conclusion

- ML is a powerful tool with many applications
- ML reinforces discrimination (esp. when trusted blindly)
  - Models incorporate “social bias” of training data even if the bias is only implicitly present
  - Data about minorities is by definition proportionally scarcer
  - Statistical patterns for majority might be invalid within a minority group
  - ML needs to be treated extra carefully if decisions could affect humans
- Research on ML security and privacy still in its infancy
  - ML systems highly susceptible to malicious input
  - Defenses usually protect only against a few specific attacks
  - Tension between complexity – accuracy – resilience
  - Privacy and predictive accuracy seem to harmonize well

ÉCOLE POLYTECHNIQUE
FÉDÉRALE DE LAUSANNE

FURTHER NOTES

Tools

https://gym.openai.com/

https://www.tensorflow.org/

https://github.com/Theano/Theano

https://github.com/cchio/deep-pwning

Academic Research on ML S&P

- The Limitations of Deep Learning in Adversarial Settings (Euro S&P 2016) https://arxiv.org/abs/1511.07528
- I Am Robot: (Deep) Learning to Break Semantic Image CAPTCHAs (Euro S&P 2016)
- Membership Inference Attacks against Machine Learning Models (S&P 2017)
- Intriguing properties of neural networks (https://arxiv.org/abs/1312.6199)
- Distillation as a Defense to Adversarial Perturbations against Deep Neural Networks (https://arxiv.org/abs/1511.04508)
- Towards the Science of Security and Privacy in Machine Learning (https://arxiv.org/abs/1611.03814)
- No Bot Expects the DeepCAPTCHA! Introducing Immutable Adversarial Examples with Applications to CAPTCHA https://eprint.iacr.org/2016/336
- Privacy-Preserving Deep Learning http://www.shokri.org/files/Shokri-CCS2015.pdf
- Model Inversion Attacks that Exploit Confidence Information and Basic Countermeasures https://www.cs.cmu.edu/~mfredrik/papers/fjr2015ccs.pdf
- Privacy in Pharmacogenetics: An End-to-End Case Study of Personalized Warfarin Dosing (introduces model inversion attacks) https://www.usenix.org/system/files/conference/usenixsecurity14/sec14-paper-fredrikson-privacy.pdf

Other Articles

https://www.wired.com/video/2016/10/president-barack-obama-on-how-artificial-intelligence-will-affect-jobs/