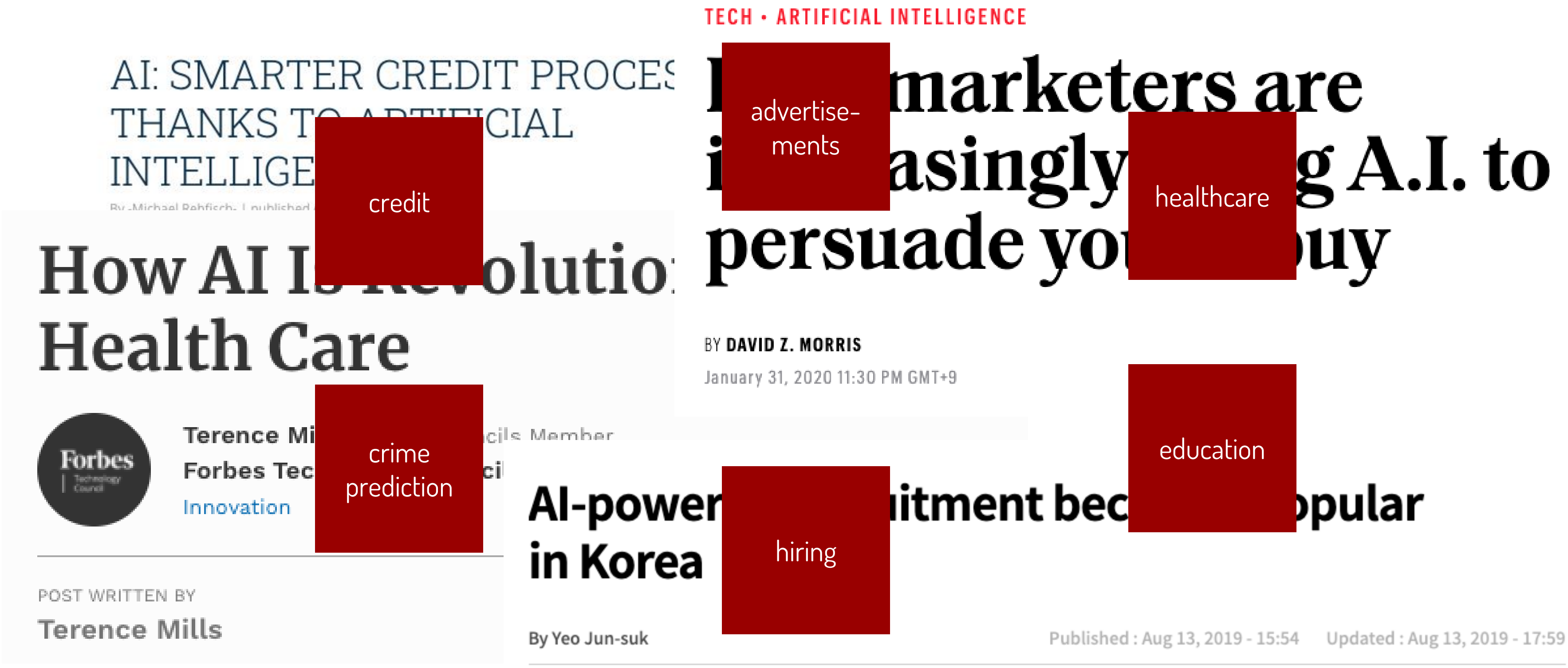
인공지능 시대의
프라이버시와 개인정보 보호

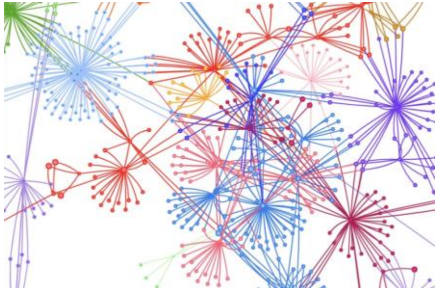
고기혁 **Gihyuk Ko**

KAIST 사이버보안연구센터 AI보안연구팀장
AI Security Research Team, KAIST CSRC

AI is Everywhere



Black-box AI Problem



Input/output
data pairs

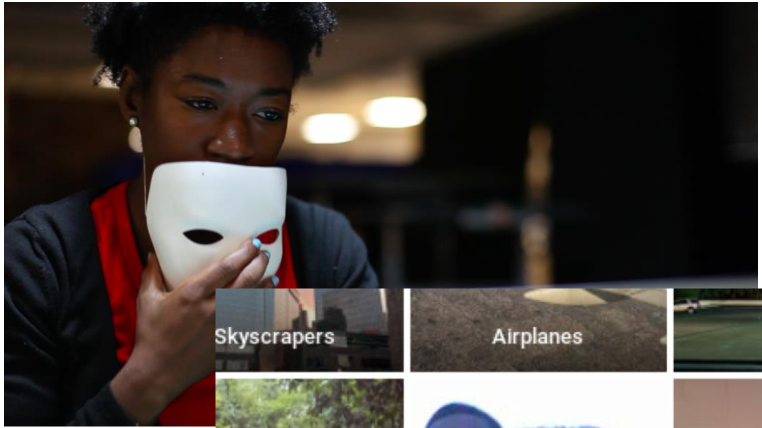
Requirements

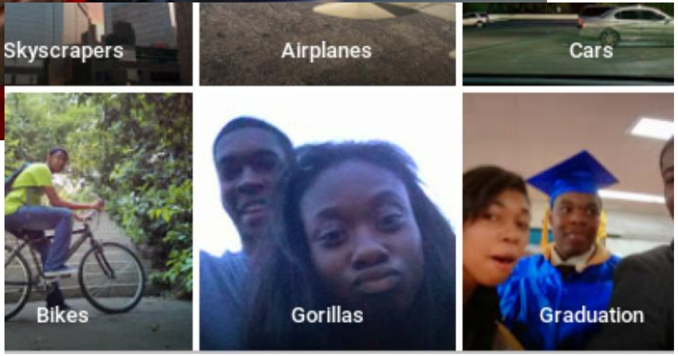
ML Algorithm

AI Model / Artifact

AI can be **opaque**: we often fail to understand how they function!

Problems of Black-box AI

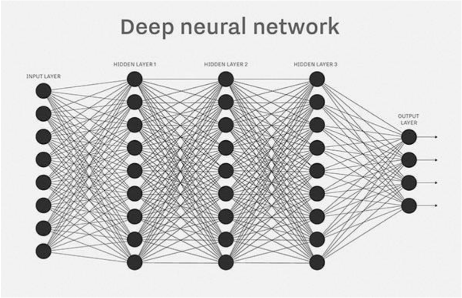




AI and Privacy

Increasing AI applications and complex private information

- AI used in essential services: banking, face/voice recognition, AI assistants
- Often processes unstructured data such as image, audio, natural language
- Difficult to ***define*** what the privacy violations are



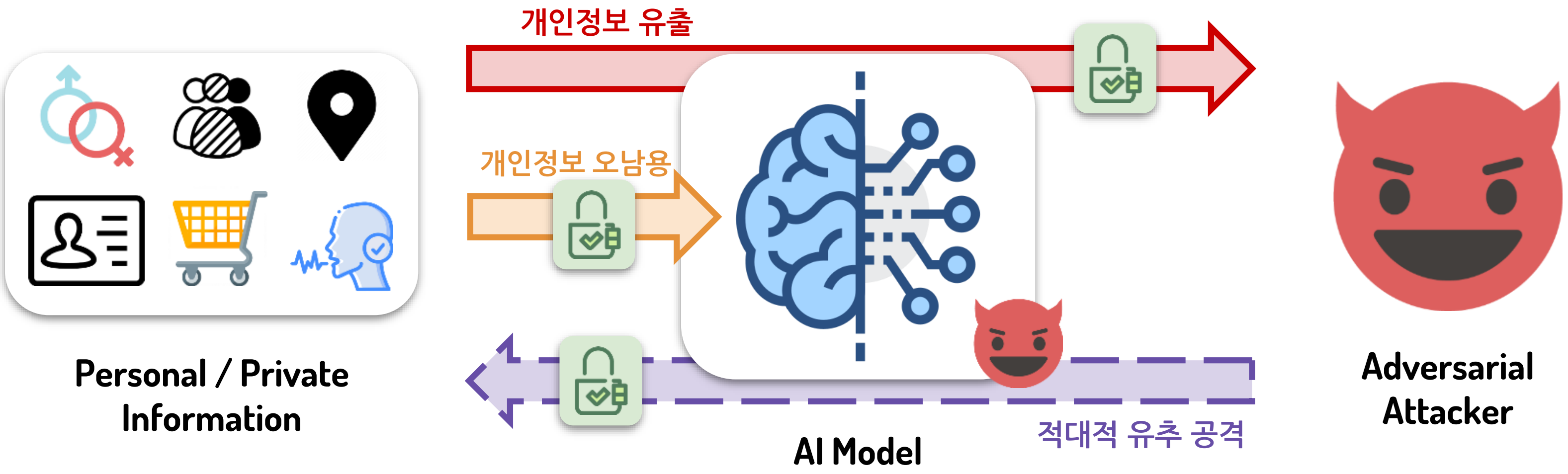
Blackbox-ness of complex AI

- Difficult to analyze what information is processed/used by AI model
- Lack of regulatory tools: how to ***inspect*** privacy violations?

Privacy protection considered as a secondary goal

- Performance is the primary goal for the companies who process private information
- They often willingly sacrifice privacy for better performing products

Privacy (Violations) in AI

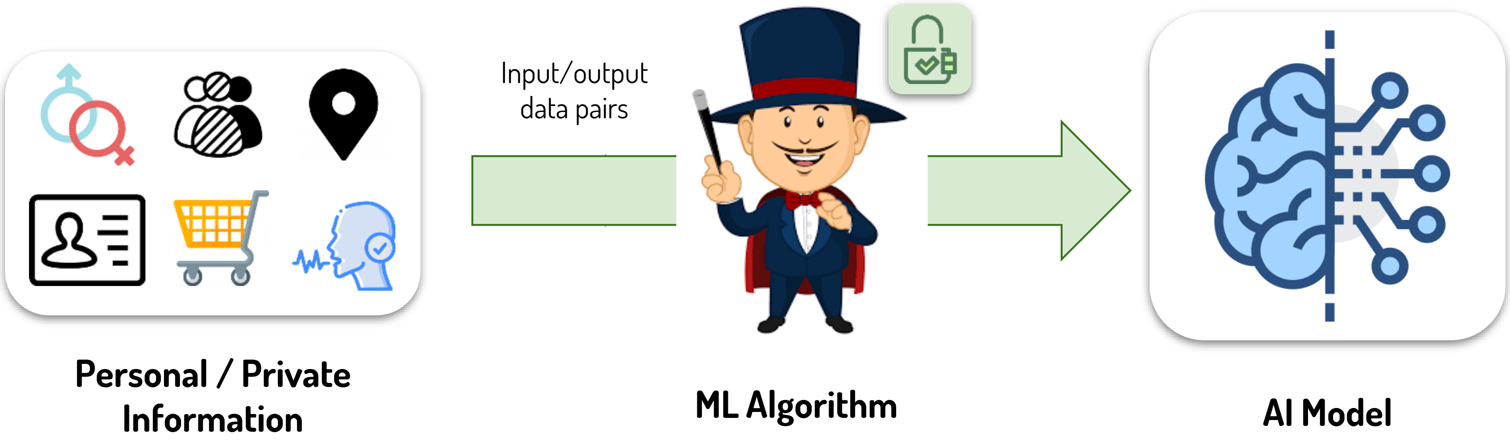


Private information can be:

- wrongfully **leaked** via AI (개인정보 유출)
- wrongfully **used** via AI (개인정보 오남용)
- wrongfully **inferred** by adversaries via AI (적대적 유추 공격)

⇒ **Preventing such violations will preserve privacy!**

Privacy by Restricting Information Leakage



Minimize sensitive information learnt by AI models in training process

- Trained AI model does not learn any information about a specific individual
- Trained AI model only learns information necessary for the given task

⇒ **Differentially Private Learning**

Differential Privacy [Dwork'06]

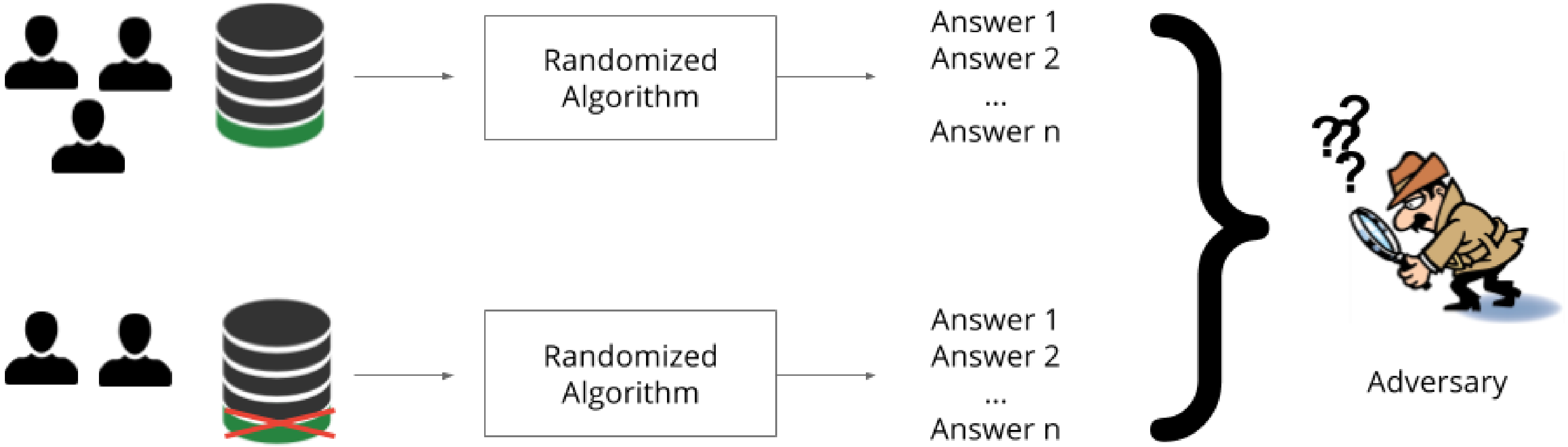


figure from: http://www.cleverhans.io

Outputs should not be distinguishable!

Adding Noise for DP

A simple solution: add **random noise** to the outputs!

- Completely random result → indistinguishable!
- Construct noise according to privacy budget
  - Laplace Mechanism: add Laplace noise
  - Gaussian Mechanism: add Gaussian noise

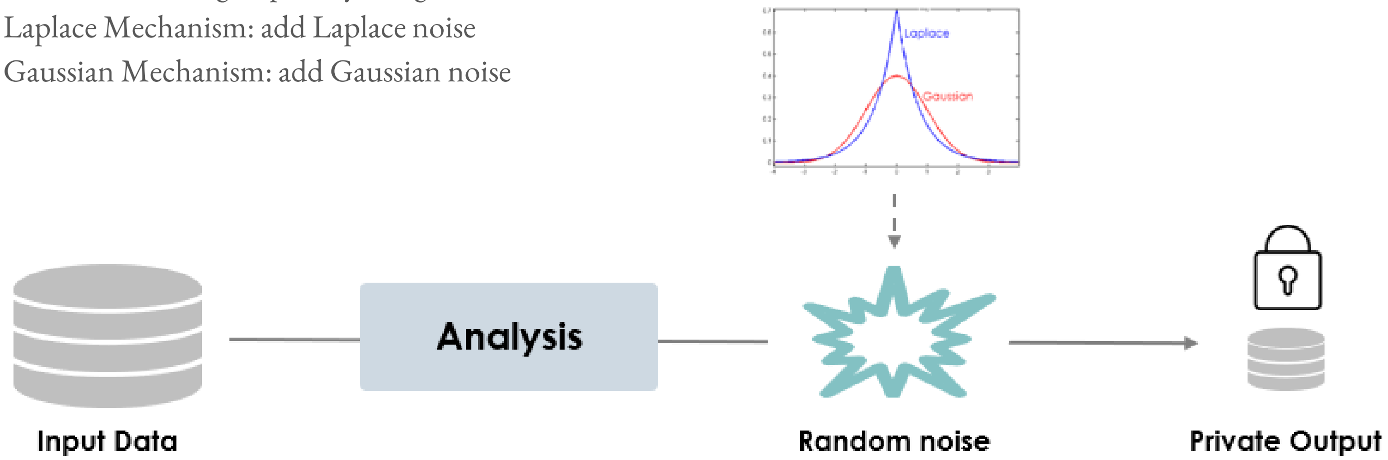


figure from: https://www.linkedin.com/pulse/why-differential-privacy-robin-röhm

Differential Privacy for AI models

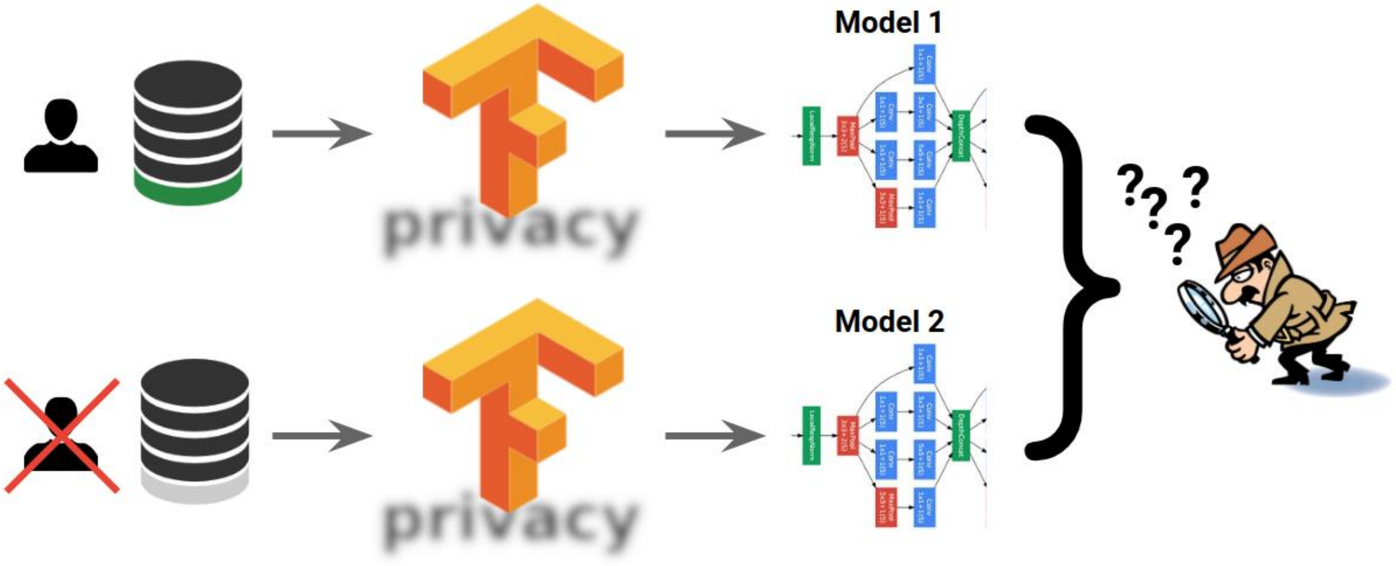


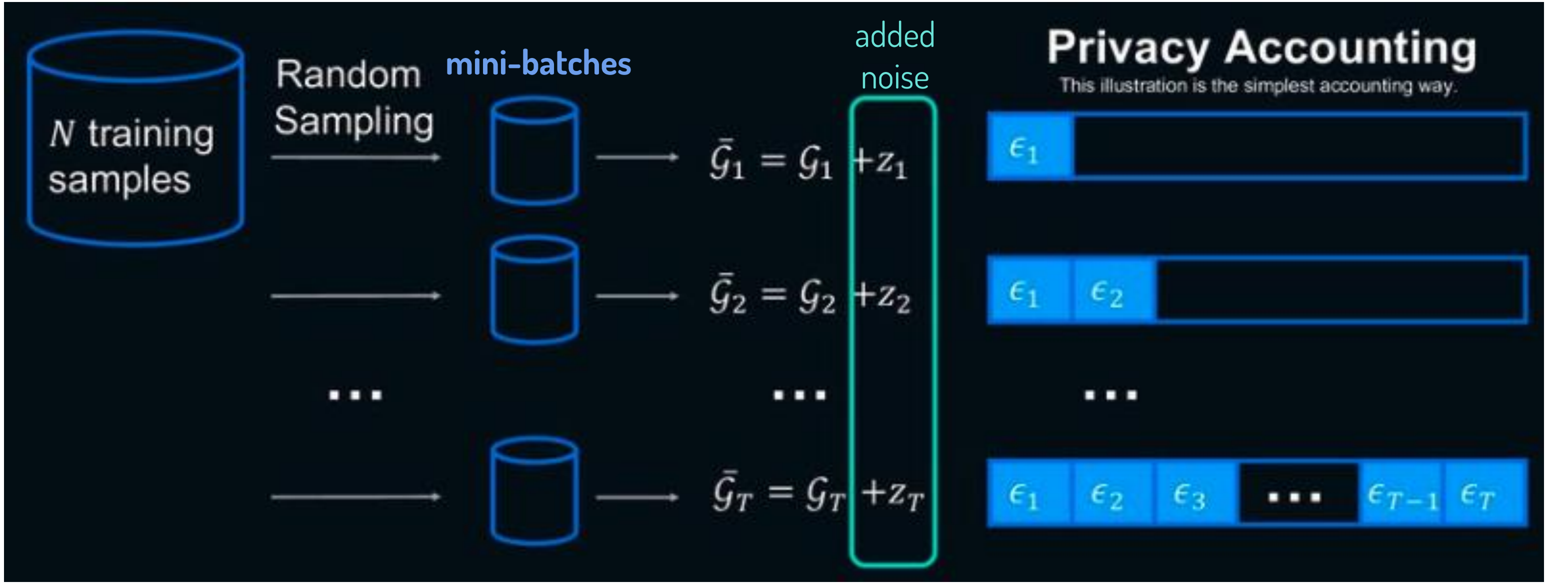
figure from: https://blog.tensorflow.org

Outputs (i.e., trained AI models) should not be distinguishable!

DP-SGD: Achieving DP in AI models

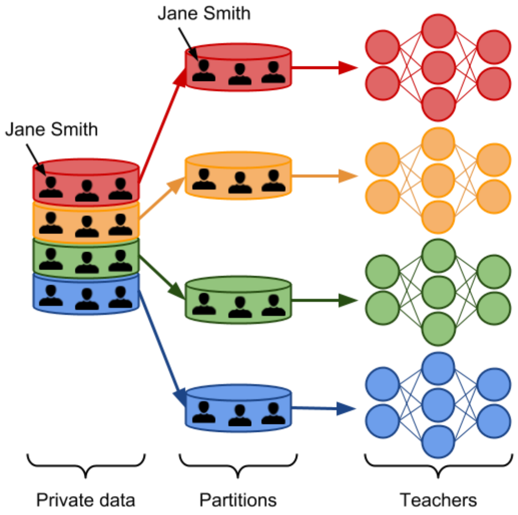
Yet another simple solution: add **random noise** in each training step!

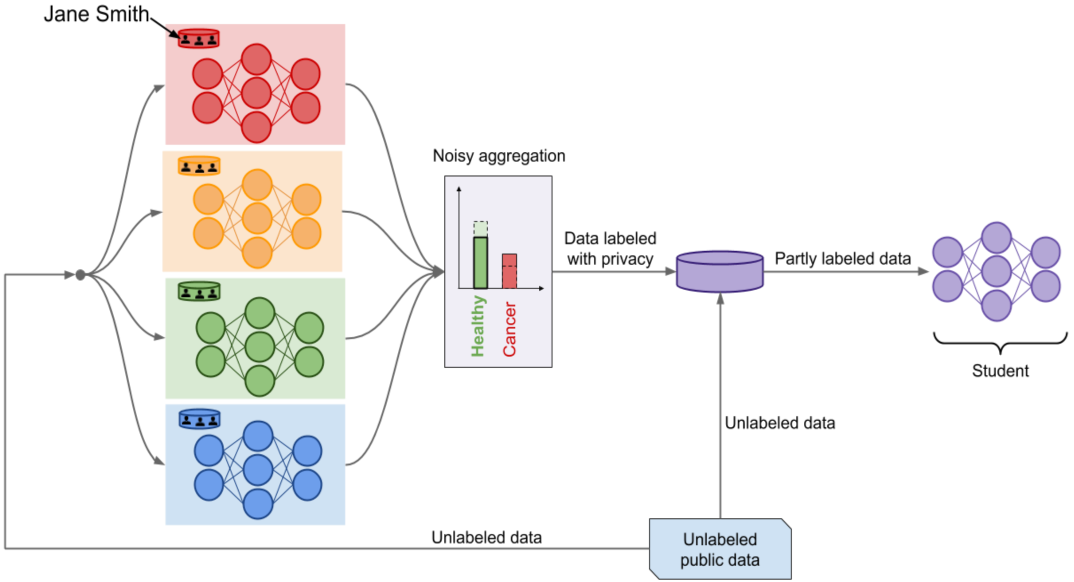
- Stochastic Gradient Descent (SGD): popular method used in training mini-batches
- In each step of SGD, add random noise for DP (privacy budget adds up)



PATE[Papernot et al.’17]

Train differentially private ensemble model on different dataset partitions, transfer knowledge to student model





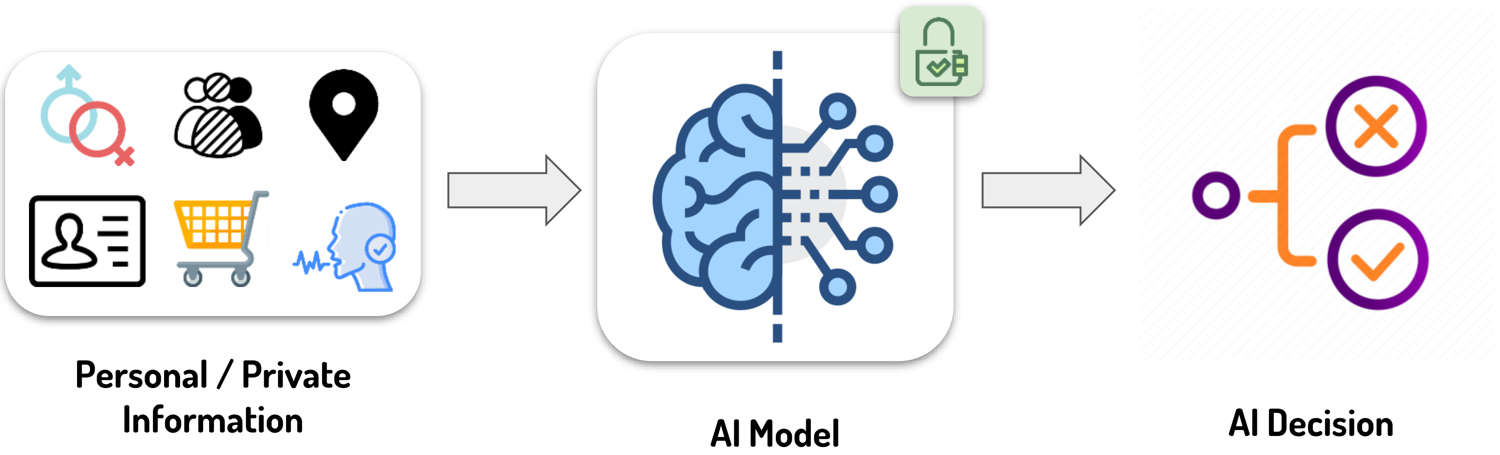
figures from: http://www.cleverhans.io

A Slightly Different Approach: Federated Learning

Enable training <u>without disclosing</u> private data

figure from: https://blogs.nvidia.com

Privacy by Restricting Information Use



Restrict **illegal/wrongful use** of sensitive information

⟹ what is illegal/wrongful use?

Motivating Examples

INTERNET CULTURE

Google's algorithm shows prestigious job ads to men, but not to women. Here's why that should worry you.

By Julia Carpenter
July 6, 2015 at 4:43 p.m. EDT

SPLINTER | The Truth Hurts

LATEST CONGRESS ELECTIONS FEATURES WHITE HOUSE TRUMP ADMINISTRATION THE FUTURE OF LABOR

Facebook is using your phone's location to suggest new friends—which could be a privacy disaster

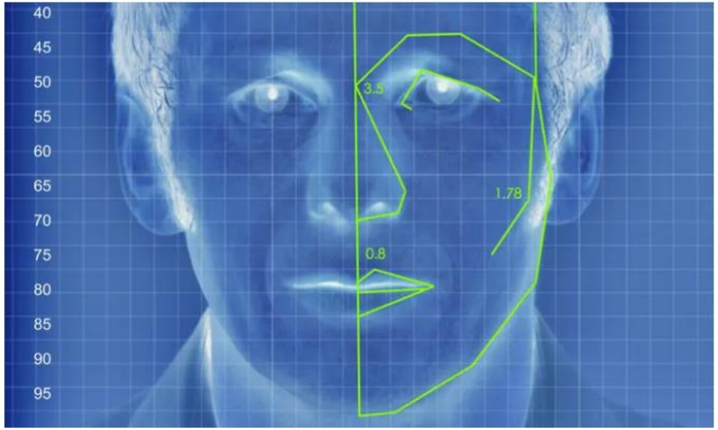
HOME > STRATEGY

The Incredible Story Of How Target Exposed A Teen Girl's Pregnancy

Gus Lubin Feb 17, 2012, 12:27 AM

New AI can guess whether you're gay or straight from a photograph

An algorithm deduced the sexuality of people on a dating site with up to 91% accuracy, raising tricky ethical questions

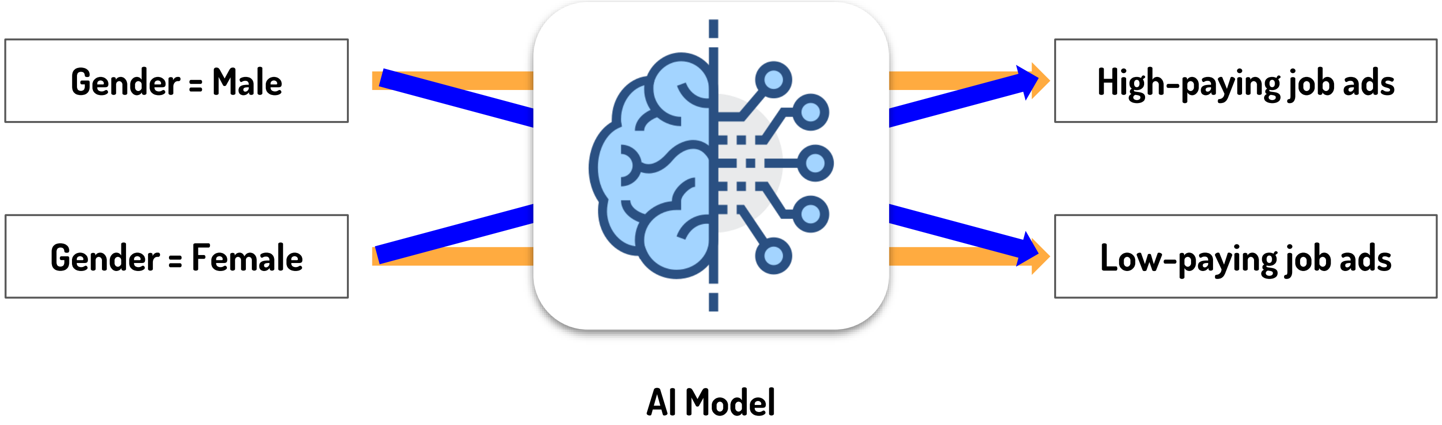


Sensitive information can be **'directly'** or **'indirectly'** used by AI in a problematic manner

Direct Information Use (Explicit Use)

Sensitive input is directly used when it can be a direct cause of output

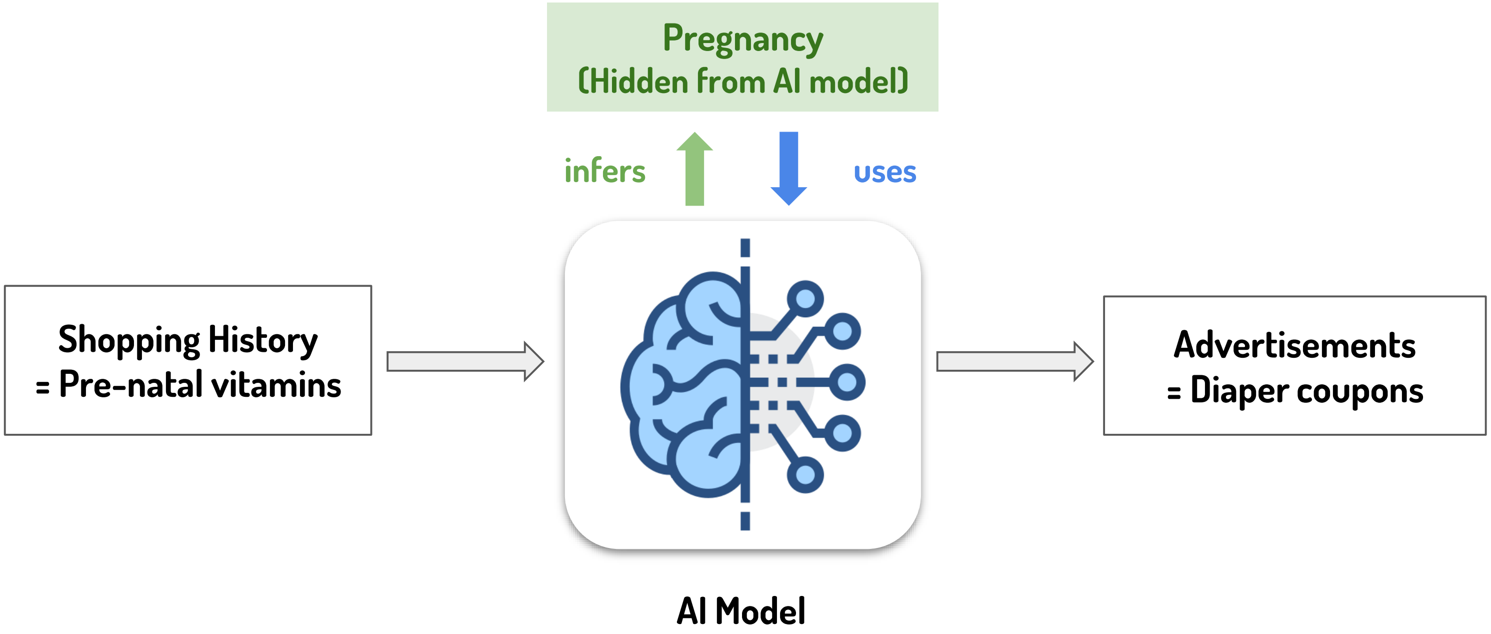
- In **Blue** case: Gender=Female causes output to be High-paying job ads
- In **Orange** case: Gender=Male causes output to be High-paying job ads



Indirect Information Use (Proxy Use)

A sensitive information is first **inferred** by AI model, then **directly causes** the decision

- Pregnancy status is inferred from shopping history, which caused diaper coupon ad



Use Privacy [Datta et al.'17]

Restriction on the information use:

- **Limited** direct and indirect use of sensitive information
- Inspect all sub-computations in the AI model to inspect suspicious information use

NOTE: you've seen privacy based on information use restrictions before:

서비스 이용 과정에서 이용자로부터 수집하는 개인정보는 아래와 같습니다.

- 회원정보 또는 개별 서비스에서 프로필 정보(별명, 프로필 사진)를 설정할 수 있습니다. 회원정보에 별
아이디가 별명으로 자동 입력됩니다.
- 네이버 내의 개별 서비스 이용, 이벤트 응모 및 경품 신청 과정에서 해당 서비스의 이용자에 한해 추가
개인정보를 수집할 경우에는 해당 개인정보 수집 시점에서 이용자에게 '수집하는 개인정보 항목, 개인정
보관기간'에 대해 안내 드리고 동의를 받습니다.

이용자 동의 후 개인정보를 추가 수집하는 경우

'개인정보 이용현황 (내정보)' 확인하기

**서비스 이용 과정에서 IP 주소, 쿠키, 서비스 이용 기록, 기기정보, 위치정보가 생성되어 수집
또한 이미지 및 음성을 이용한 검색 서비스 등에서 이미지나 음성이 수집될 수 있습니다.**

구체적으로 1) 서비스 이용 과정에서 이용자에 관한 정보를 자동화된 방법으로 생성하여 이를 저장(수집
원래의 값을 확인하지 못 하도록 안전하게 변환하여 수집합니다. 서비스 이용 과정에서 위치정보가 수집
서비스에 대해서는 '네이버 위치정보 이용약관'에서 자세하게 규정하고 있습니다.
이와 같이 수집된 정보는 개인정보와의 연계 여부 등에 따라 개인정보에 해당할 수 있고, 개인정보에 해당

PIPEDA
Personal Information Protection
and Electronic Documents Act

air information p

ation under its control and shall d
he organization's compliance with
principles.

Principle 2 – Identifying Purposes

The purposes for which personal information is collected shall be identified by the o
before the time the information is collected.

Principle 3 – Consent

The knowledge and consent of the individual are required for the collection, use, or d
personal information, except where inappropriate.

Principle 4 – Limiting Collection

The collection of personal information shall be limited to that which is necessary for
identified by the organization. Information shall be collected by fair and lawful mean

Principle 5 – Limiting Use, Disclosure, and Retention

Personal information shall not be used or disclosed for purposes other than those fo
collected, except with the consent of the individual or as required by law. Personal in

Art. 5 GDPR

Principles relating to processing of personal data

1. Personal data shall be:

(a) processed lawfully, fairly and in a transparent manner in relation to the data subject ('lawfulness, fairness and transparency');

(b) collected for specified, explicit and legitimate purposes and not further processed in a manner that is incompatible with those purposes; further processing for archiving purposes in the public interest, scientific or historical research purposes or statistical purposes shall, in accordance with Article 89(1), not be considered to be incompatible with the initial purposes ('purpose limitation');

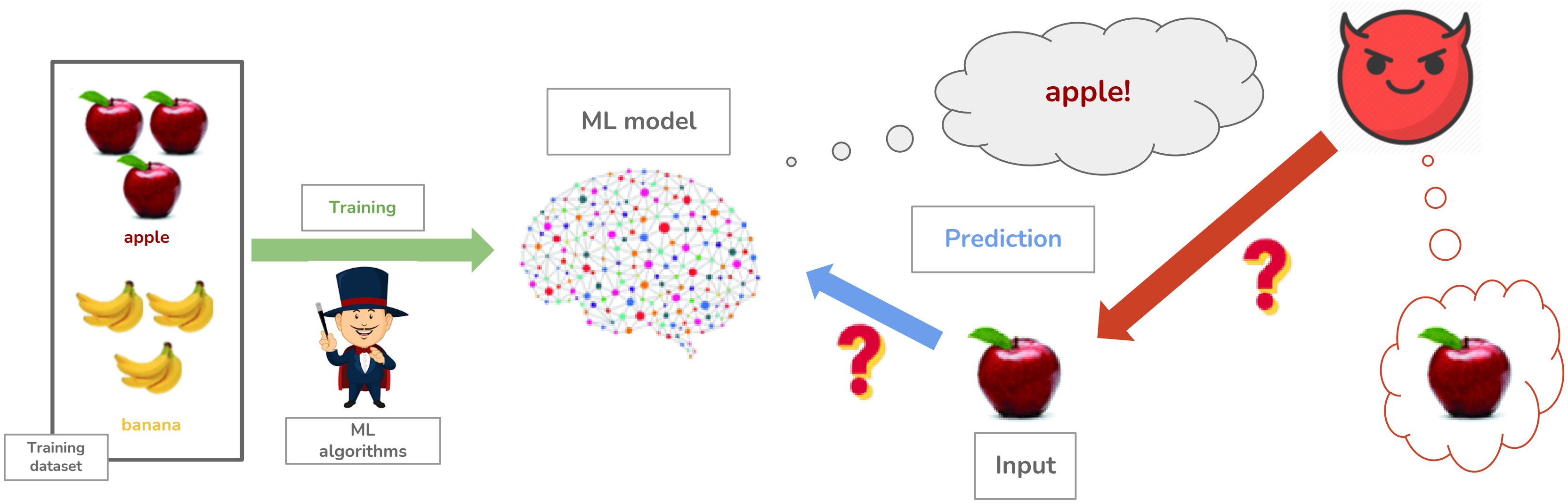
(c) adequate, relevant and limited to what is necessary in relation to the purposes for which they are processed ('data minimisation');

(d) accurate and, where necessary, kept up to date; every reasonable step must be taken to ensure that personal data that are inaccurate, having regard to the purposes for which they are processed, are erased or rectified without delay ('accuracy');

Inference Attacks against AI models

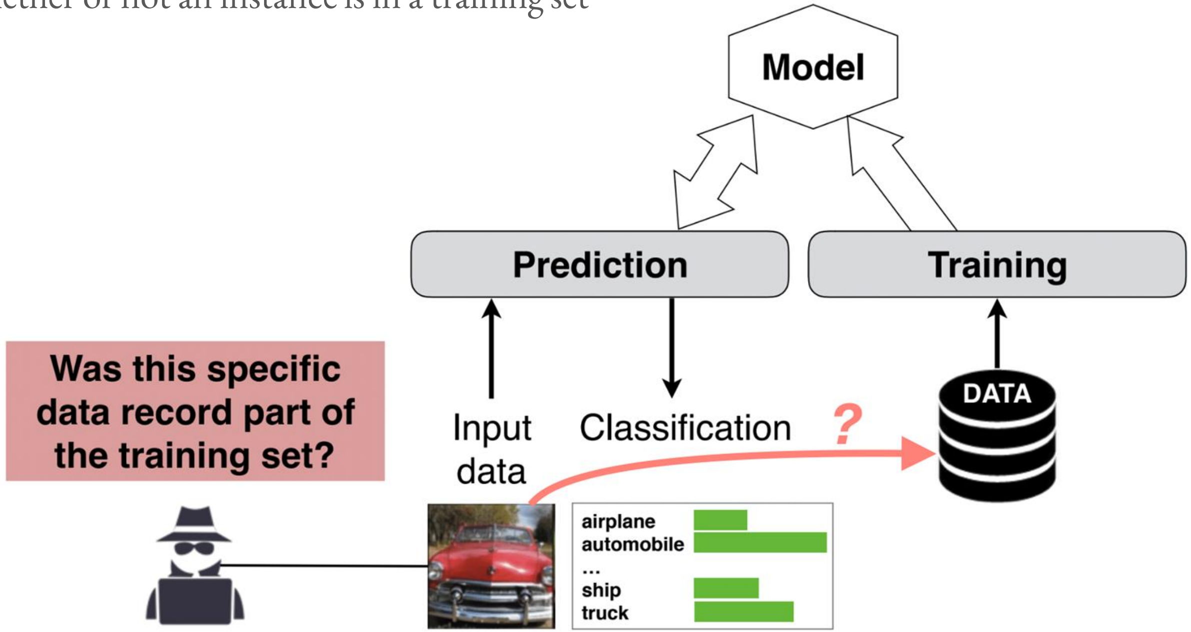
An adversary tries to obtain information on:

- sensitive inputs (*model inversion*)
- training instances (*membership inference*)
- ~~Model parameters (*model stealing*)~~



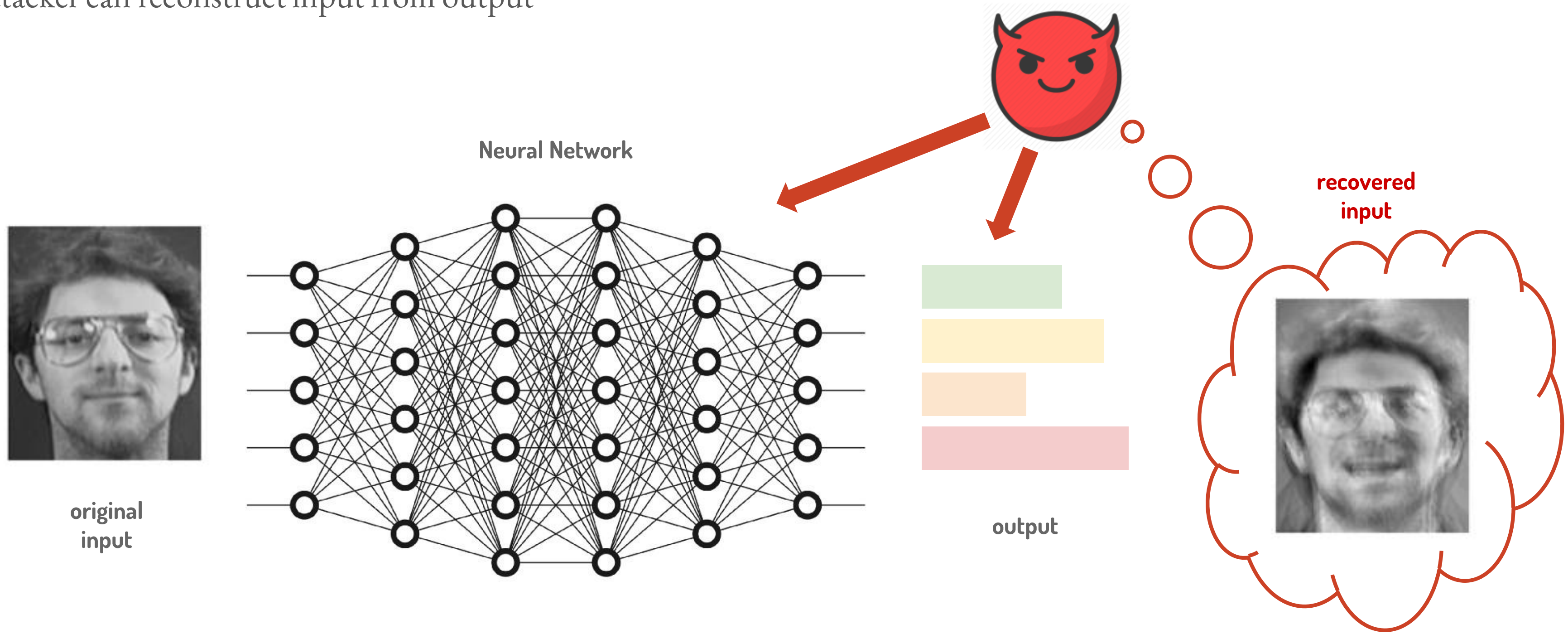
Membership Inference [Shokri et al.'17]

Attacker guesses whether or not an instance is in a training set



Model Inversion Attacks [Fredrikson et al.'17]

Attacker can reconstruct input from output



Defenses against Inference Attacks

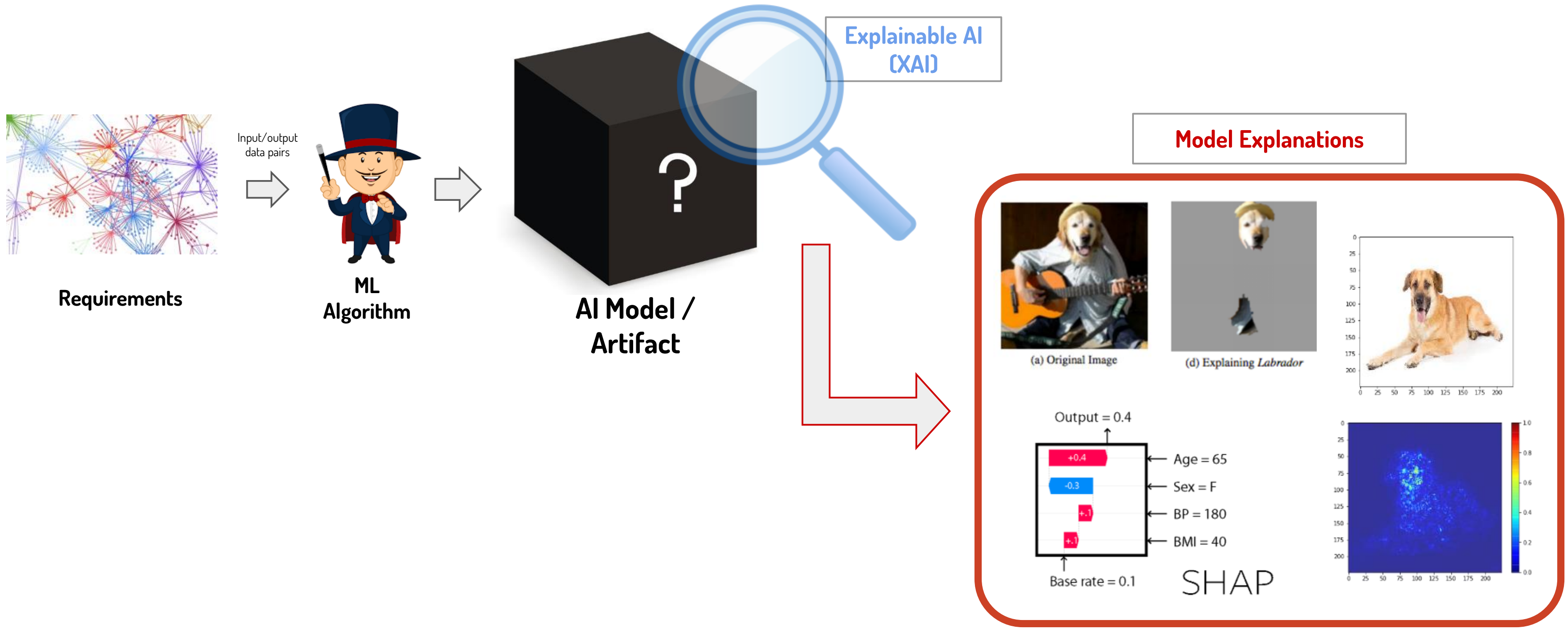
Many defense methods have been suggested, but no guaranteed method of defense

- Temperature-based smoothing
- Noise injection (model output)
- Model output randomization
- Poisoning-based defenses
- Adversarial example-based defenses
- Differentially Private Learning
- DP for AI model weights
- SplitNN: a variant of federated learning
- ... and more

⇒ **Attacks as well as defenses are currently actively studied**

Novel Threats to Privacy: AI interpretability

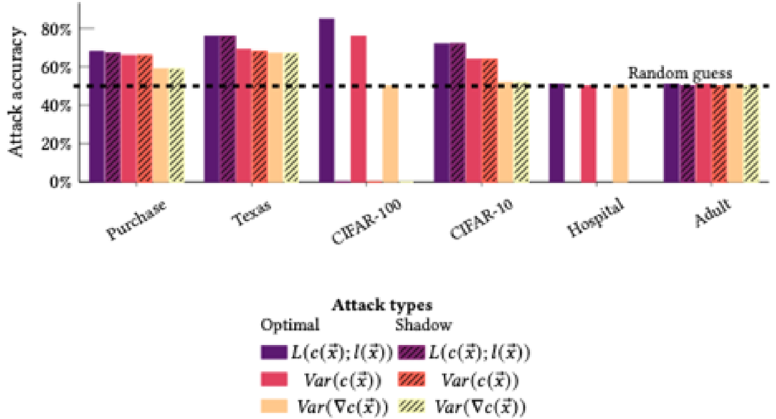
Explainable/Interpretable AI (XAI): **explain/interpret** AI model's decisions in humanly understandable way



AI Explanations can Leak Information

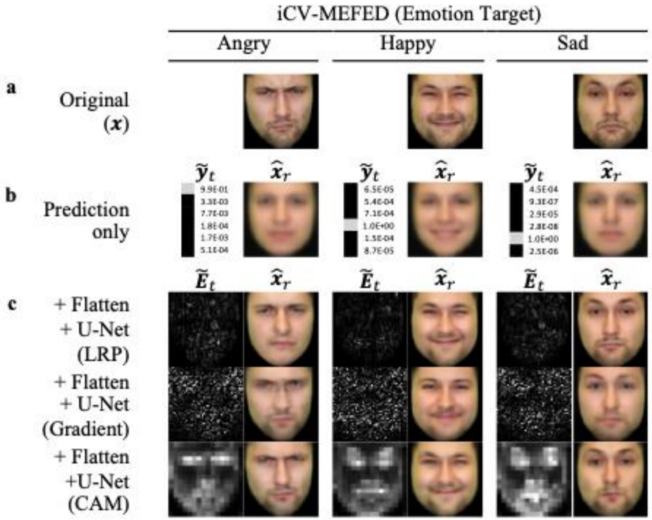
Explanations used for Membership Inference [Shokri et al.'21]

- Explanations enable a more precise inference on the membership of training instances



Explanations used for Model Inversion [Zhao et al.'20]

- Explanations enable a more accurate reconstruction of input images



Takeaways

Privacy violations in AI models can occur in various ways

- Wrongful information leakage, information use, and inference
- Increasing privacy risk in AI models due to blackbox-ness of AI
- Lack of specific regulatory tools and standards

Different privacy definitions preventing such violations

- **Differentially Private Learning** to prevent AI models to learn too much
- **Use Privacy** to prevent AI models to use sensitive information in a wrongful manner
- **Defense** against the inference attacks is a subject of active development

Novel threats due to transparency/interpretability requirements

- **Model explanations** can leak additional information to adversaries
- Currently, little research is done on developing defense mechanism

⇒ It is critical to properly understand and study privacy threats in AI models!

Thank you!

References

Cynthia Dwork: **Differential Privacy**. ICALP (2) 2006: 1-12

Nicolas Papernot, Martín Abadi, Úlfar Erlingsson, Ian J. Goodfellow, Kunal Talwar:
Semi-supervised Knowledge Transfer for Deep Learning from Private Training Data. ICLR 2017

Anupam Datta, Matthew Fredrikson, Gihyuk Ko, Piotr Mardziel, Shayak Sen:
Use Privacy in Data-Driven Systems: Theory and Experiments with Machine Learnt Programs. CCS 2017: 1193-1210

Reza Shokri, Marco Stronati, Congzheng Song, Vitaly Shmatikov: **Membership Inference Attacks Against Machine Learning Models**. IEEE S&P 2017

Matt Fredrikson, Somesh Jha, Thomas Ristenpart: **Model Inversion Attacks that Exploit Confidence Information and Basic Countermeasures**. ACM CCS 2015

Reza Shokri, Martin Strobel, Yair Zick: **On the Privacy Risks of Model Explanations**. AIES 2021: 231-241

Xuejun Zhao, Wencan Zhang, Xiaokui Xiao, Brian Y. Lim: **Exploiting Explanations for Model Inversion Attacks**. ICCV 2021: 662-672