Artificial Intelligence

CS 165A

June 8, 2023

Instructor: Prof. Yu-Xiang Wang

Today

→ Responsible AI

Logistic notes

- Online ESCI Survey
  - Only a few students completed the survey.
  - We can do better! The deadline is **Jun 9 (This Friday)**
  - Please take a moment to complete your feedback!

- Project 3 due **today**

- Final exam **next Wednesday** 12 – 3
  - Open book (no digital devices)
  - Twice the time but only slightly longer than the midterm
  - Covers Minimax Search, MDPs, Bandits, RL, Logic (except FOL inference) and Responsible AI

Recap: First-Order Logic (FOL)

A method of analysis or calculation using a special symbolic notation

- Also known as *First-Order Predicate Calculus*
  - Propositional logic is also known as *Propositional Calculus*

- An extension to propositional logic in which quantifiers can bind variables in sentences
  - Universal quantifier (∀)
  - Existential quantifier (∃)
  - Variables: *x, y, z, a, joe, table*…

- Examples
  - **∀*x* Beautiful(*x*)**
  - **∃*x* Beautiful(*x*)**

Inference in First-Order Logic

- Inference rules for propositional logic:
  - Modus ponens, and-elimination, and-introduction, or-introduction, resolution, etc.
  - These are valid for FOL also

- But since these don't deal with quantifiers and variables, we need new rules, especially those that allow for substitution (binding) of variables to objects
  - These are called ***lifted*** inference rules

Substitution and variable binding

- Notation for substitution:
  - SUBST (**Binding list**, **Sentence**)
    - Binding list: { *var* / ground term, *var* / ground term, … }
    - "ground term" = term with no variables
  - SUBST({*var*/gterm}, Func (*var*)) = Func (gterm)
    - SUBST (θ, p)
  - Examples:
    - SUBST ({*x*/Mary}, FatherOf (*x*)) = FatherOf (Mary)
    - SUBST ({*x*/Joe, *y*/Lisa}, Siblings (*x*,*y*)) = Siblings (Joe, Lisa)

Three new inference rules using SUBST(θ, p)

- Universal Instantiation

$$\frac{\forall v \quad \alpha}{SUBST(\{v/g\}, \alpha)}$$

g – ground term

- Existential Instantiation

$$\frac{\exists v \quad \alpha}{SUBST(\{v/k\}, \alpha)}$$

k – constant that does not appear elsewhere in the knowledge base

- Existential Introduction

$$\frac{\alpha}{\exists v \; SUBST(\{g/v\}, \alpha)}$$

v – variable not in α
g – ground term in α

Universal Instantiation – examples

$$\frac{\forall v \quad \alpha}{SUBST(\{v/g\}, \alpha)}$$

g – ground term

- $\forall x$ Sleepy(x)
  - SUBST($\{x$/Joe$\}$, α)
    - Sleepy(Joe)
- $\forall x$ Mother(x) $\Rightarrow$ Female(x)
  - SUBST($\{x$/Mary$\}$, α)
    - Mother(Mary) $\Rightarrow$ Female(Mary)
  - SUBST($\{x$/Dad$\}$, α)
    - Mother(Dad) $\Rightarrow$ Female(Dad)
- $\forall x, y$ Buffalo(x) $\wedge$ Pig(y) $\Rightarrow$ Outrun(x,y)
  - SUBST($\{x$/Bob$\}$, α)
    - $\forall y$ Buffalo(Bob) $\wedge$ Pig(y) $\Rightarrow$ Outrun(Bob,y)

Existential Instantiation – examples

$$\frac{\exists v \quad \alpha}{SUBST(\{v/k\}, \alpha)}$$

k – constant that does not appear elsewhere in the knowledge base

- $\exists x$ BestAction(x)
  - SUBST({x/B_A}, α)
    - BestAction(B_A)
      - "B_A" is a constant; it is not in our universe of actions

- $\exists y$ Likes(y, Broccoli)
  - SUBST({y/Bush}, α)
    - Likes(Bush, Broccoli)
      - "Bush" is a constant; it is not in our universe of people

Existential Introduction – examples

$$\frac{\alpha}{\exists v\ SUBST(\{g/v\}, \alpha)}$$

v – variable not in α
g – ground term in α

- Likes(Jim, Broccoli)
  - SUBST({Jim/x}, α)
    - $\exists x$ Likes(x, Broccoli)

- $\forall x$ Likes(x, Broccoli) $\Rightarrow$ Healthy(x)
  - SUBST({Broccoli/y}, α)
    - $\exists y\ \forall x$ Likes(x, y) $\Rightarrow$ Healthy(x)

Inference algorithms in first order logic will not be covered in the final. (FOL will be!)

- However, it is a powerful tool.
  - Expert systems (since 1970s)
  - Large scale industry deployment.

- It is however fragile and rely on the correct / error-free representation of the world in black and white
  - This limits its use in cases when the evidence is collected stochastically and imprecisely by people's opinions in large scale.

- Somewhat superseded by machine learning on many problems, but:
  - Research on logic agent is coming back.
  - Add knowledge and reasoning to ML-based solution
  - After all, ML are just reflex agents usually.

Future of AI

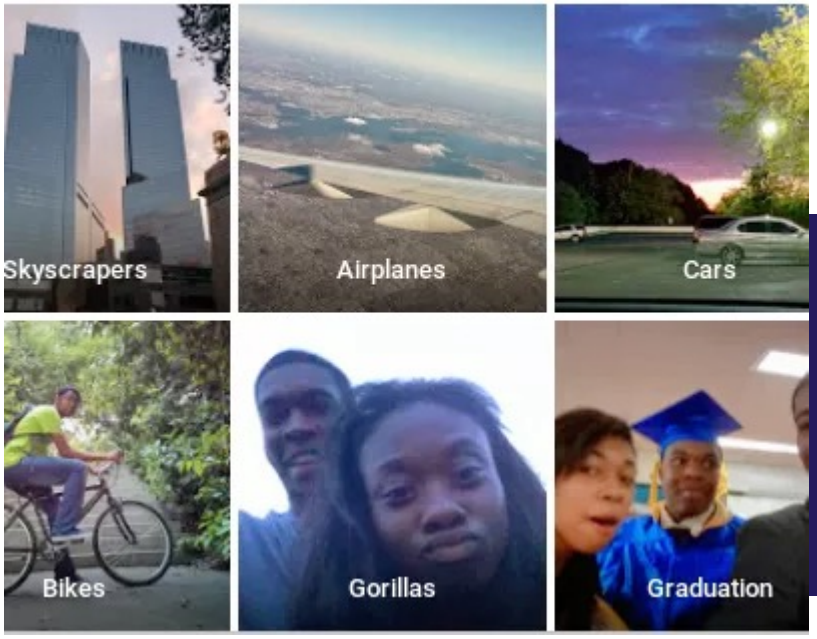
- More higher level intelligence
  - Logic is coming back
  - But more learning based than rule-based
- More stateful systems, more reinforcement learning
  - Causal modelling and reasoning
- More AI in the non-iid environment
  - Structured
  - Adversarial
- More forms of agent's perception
  - Weak supervision
  - Self-supervision (bootstrapping)
- More interactive (natural interface to human)
  - Via dialogue / ChatGPT

The need for responsible AI: with great power comes great responsibility

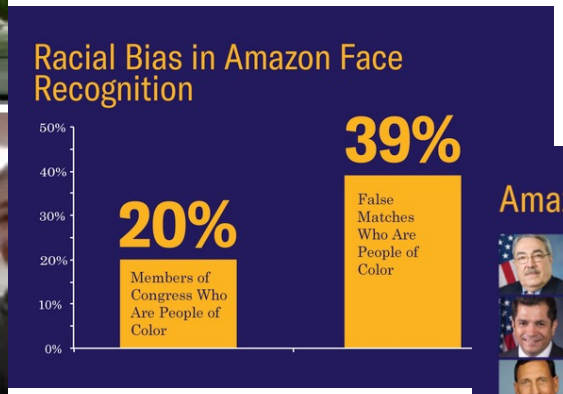
A face recognition system

- Technology is a double-bladed sword
- It matters who wields it and for what purpose

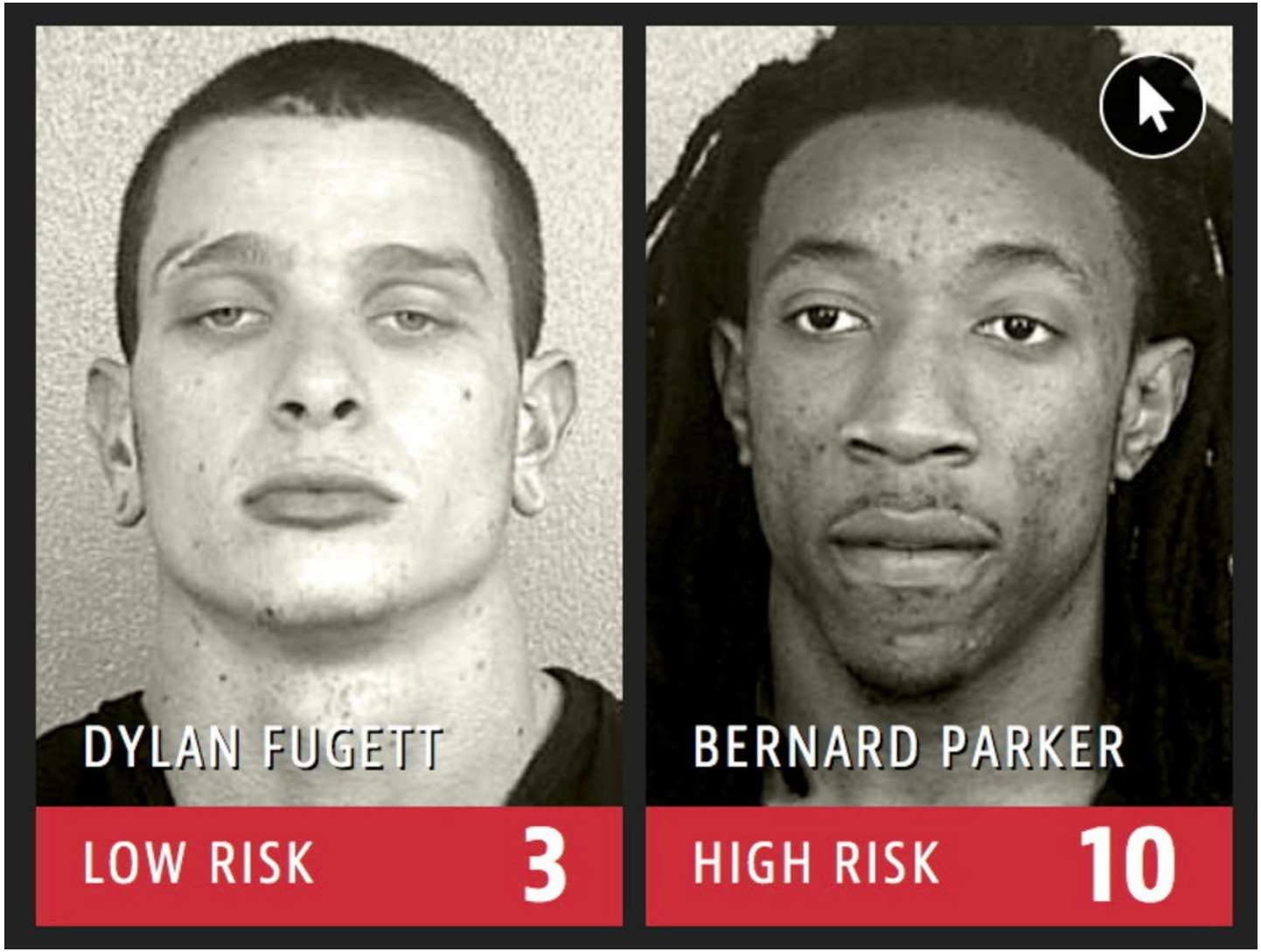
Fairness challenges in AI systems / AI for decision making



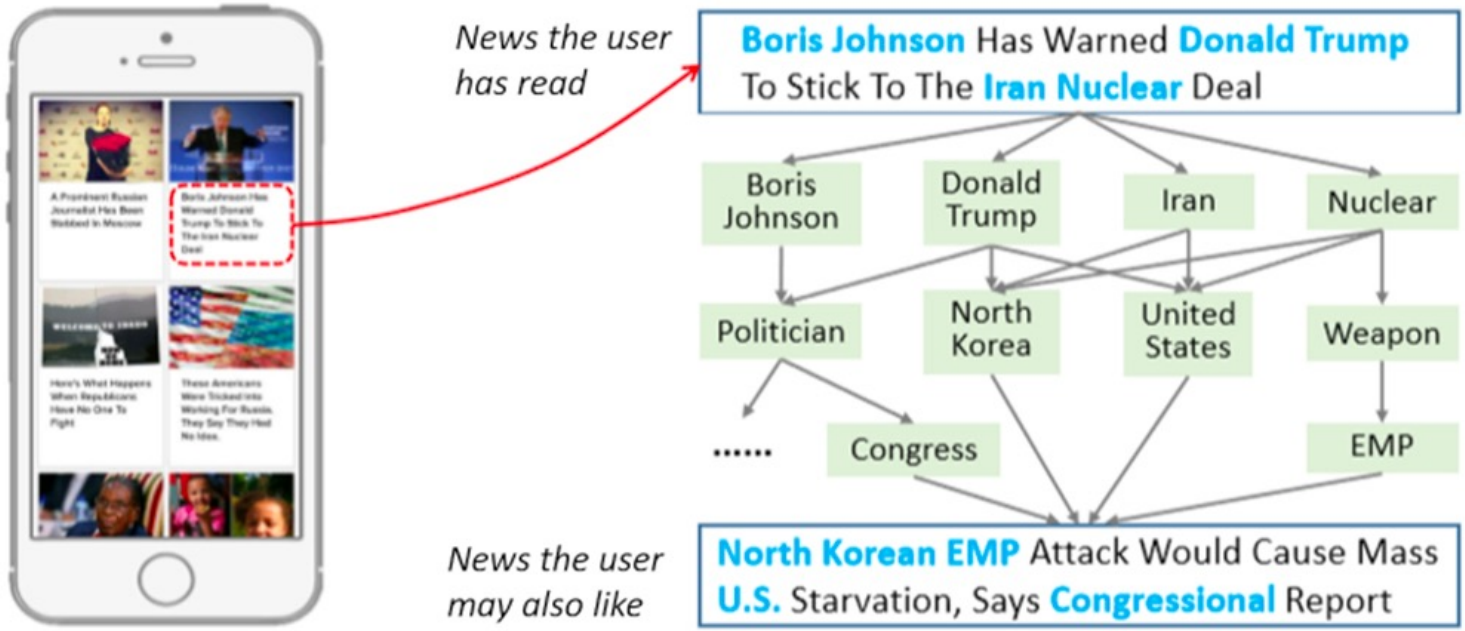
Google's image recognition system



AI for predicting recidivism: "COMPAS" is used by courts… but is it biased?



Polarizing effects of news recommendation



- Only what you like to read will be recommended to you.

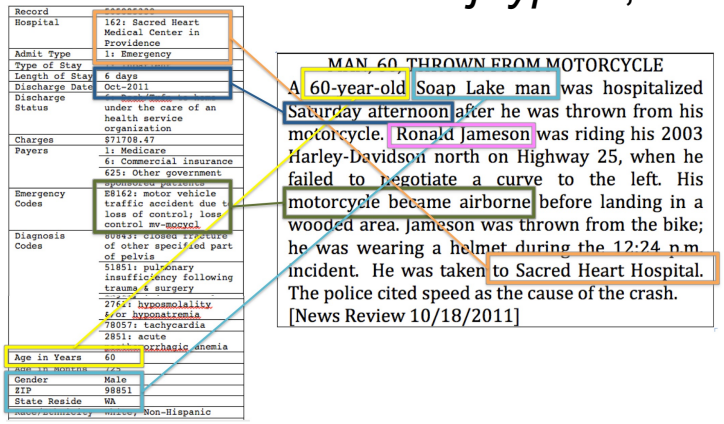
Privacy issues in data collection and learning

"Robust De-anonymization of Large Datasets (How to Break Anonymity of the Netflix Prize Dataset)"

A. Narayanan & V. Shmatikov. *Security and Privacy, 2008*

- **Anonymization doesn't work!**
- **Need robust / provable approaches.**

Vijay Pandurangan. *tech.vijayp.ca*, 2014



"Only You, Your Doctor, and Many Others May Know"

L. Sweeney. *Technology Science, 2015*

ML models memorize training datasets, even though they are generalizing well!

Membership Inference Attacks Against Machine Learning Models

Reza Shokri, Cornell Tech
Marco Stronati*, INRIA
Congzheng Song, Cornell
Vitaly Shmatikov, Cornell Tech

***Abstract*—We quantitatively investigate how machine learning models leak information about the individual data records on which they were trained. We focus on the basic membership inference attack: given a data record and black-box access to a model, determine if the record was in the model's training dataset. To perform membership inference against a target model, we make adversarial use of machine learning and train our own inference model to recognize differences in the target model's predictions on the inputs that it trained on versus the inputs that it did not train on.**

We empirically evaluate our inference techniques on classification models trained by commercial "machine learning as a service" providers such as Google and Amazon. Using realistic datasets and classification tasks, including a hospital discharge dataset whose membership is sensitive from the privacy perspective, we show that these models can be vulnerable to membership inference attacks. We then investigate the factors that influence this leakage and evaluate mitigation strategies.

Security and Privacy, 2017

The Secret Sharer:
Measuring Unintended Neural Network Memorization & Extracting Secrets

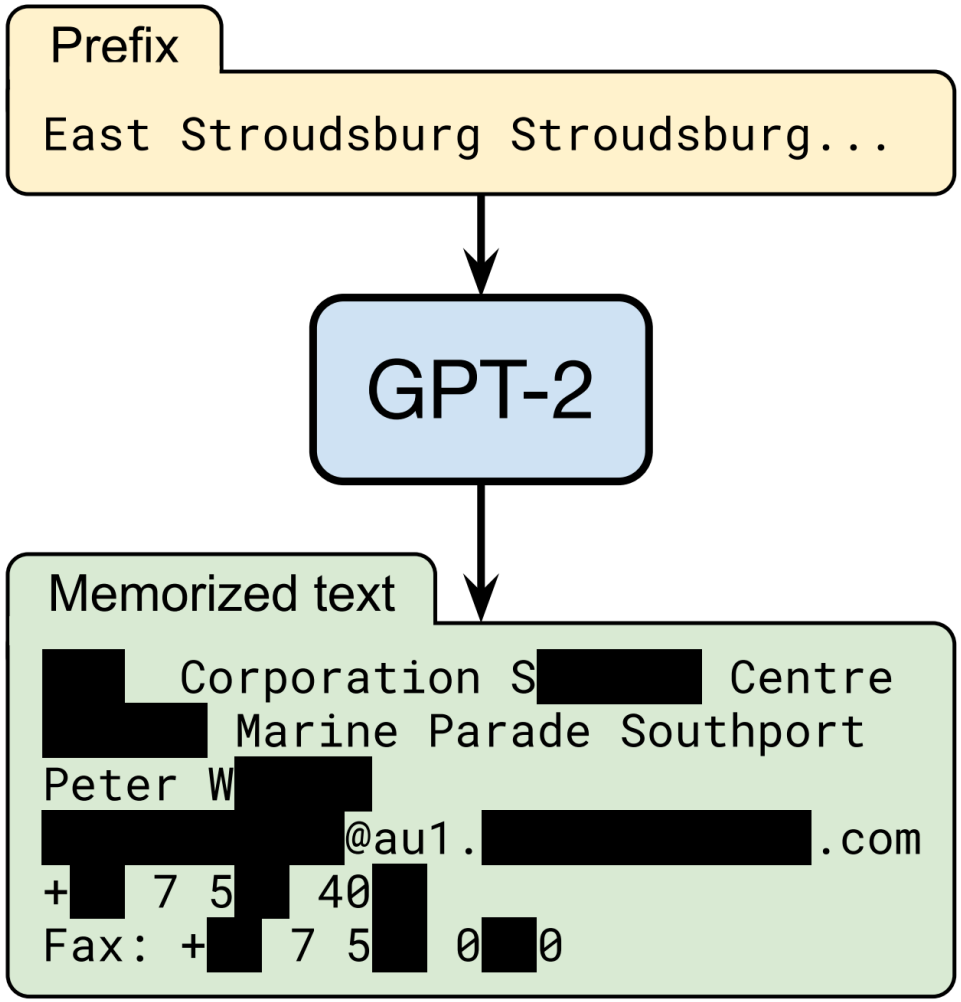
Nicholas Carlini, *University of California, Berkeley*
Chang Liu, *University of California, Berkeley*
Jernej Kos, *National University of Singapore*
Úlfar Erlingsson, *Google Brain*
Dawn Song, *University of California, Berkeley*

This paper presents *exposure*, a simple-to-compute metric that can be applied to any deep learning model for measuring the memorization of secrets. Using this metric, we show how to extract those secrets efficiently using black-box API access. Further, we show that unintended memorization occurs early, is not due to overfitting, and is a persistent issue across different types of models, hyperparameters, and training strategies. We experiment with both real-world models (e.g., a state-of-the-art translation model) and datasets (e.g., the Enron email dataset, which contains users' credit card numbers) to demonstrate both the utility of measuring exposure and the ability to extract secrets.

Finally, we consider many defenses, finding some ineffective (like regularization), and others to lack guarantees. However, by instantiating our own differentially-private recurrent model, we validate that by appropriately investing in the use of state-of-the-art techniques, the problem can be resolved, with high utility.

USENIX Security 2019

With appropriate prompt, GPT2 outputs sensitive training data verbatim



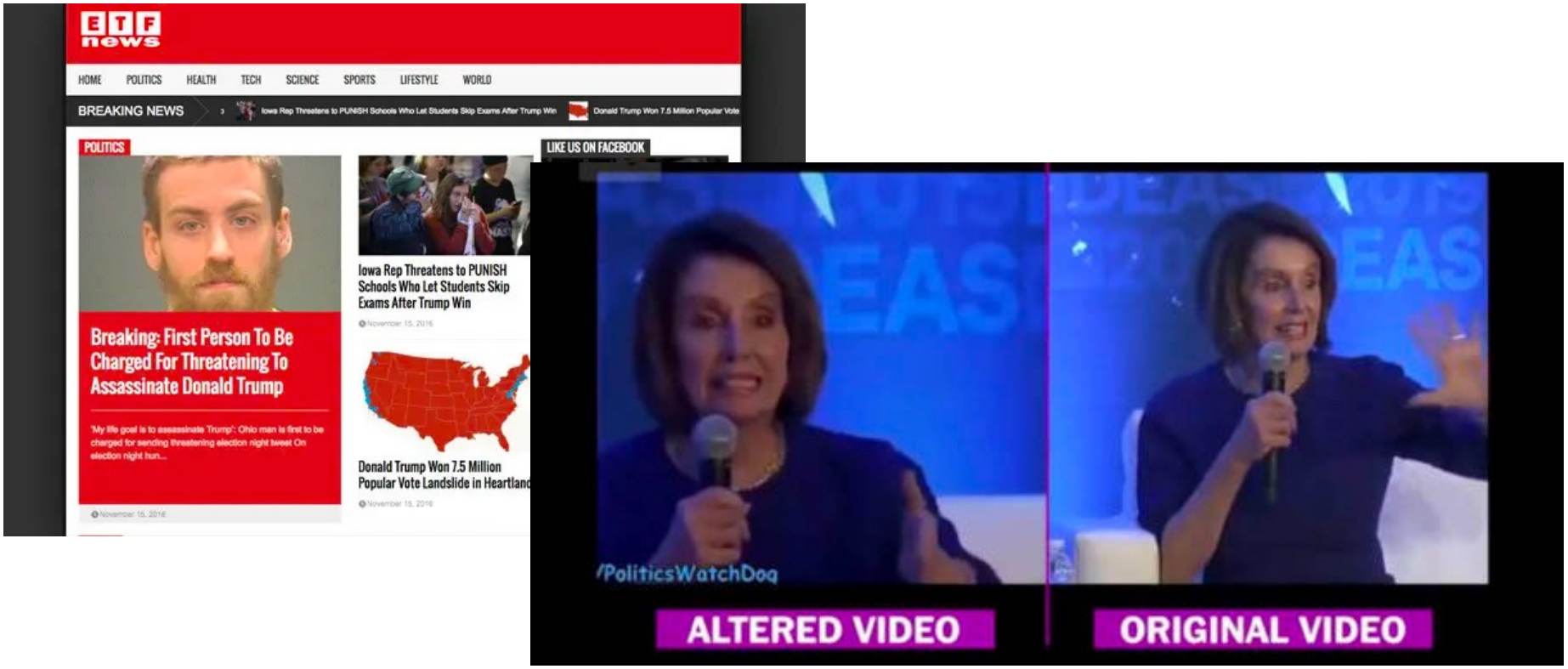
Recent/upcoming legislations on privacy forces companies to revise their data practice

- I can't keep personal data for more than three weeks?

- I will have to delete all traces of a user upon request?

How about my machine learning models trained on user data?

Fake-news, fake voice, fake video



- How to tell if something is true or false?
- How to attribute a crime with factual evidence when people can just claim it's fake?

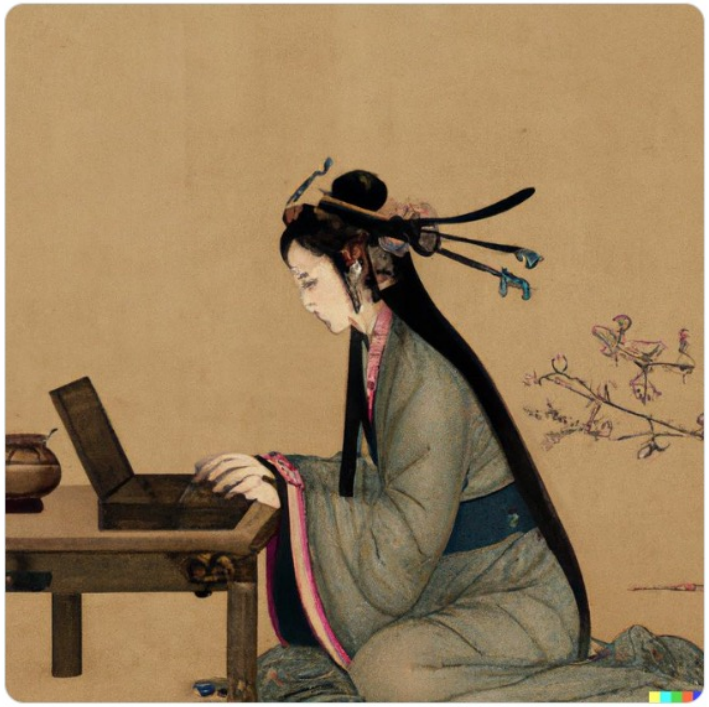
The rise of generative models

- We've seen Generative Adversarial Networks (GAN)
- We've also seen what GPT-3 is able to do
  - Generate text / code / table / and so on…
- More recent example: DALL-E 2

"An astronaut riding a horse in a photorealistic style."

https://openai.com/dall-e-2/

"Oriental painting of a lady programming on a laptop in the Song Dynasty" #Dalle⛩



https://twitter.com/hardmaru/status/1523971427292127232

Are Github Copilot / DALL-E 2 violating copyrights?

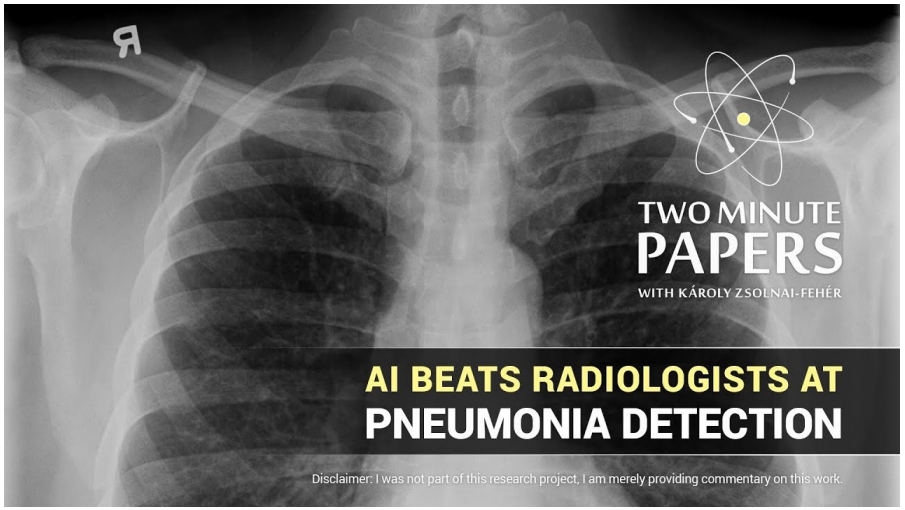
- Co-Pilot autocompletes code for you. But … they are trained on data all over the internet. From time to time, they generate code / image verbatim. (See the following example: copilot generates code from "Quake")

https://twitter.com/mitsuhiko/status/1410886329924194309

- Are the generated content considered plagiarism?

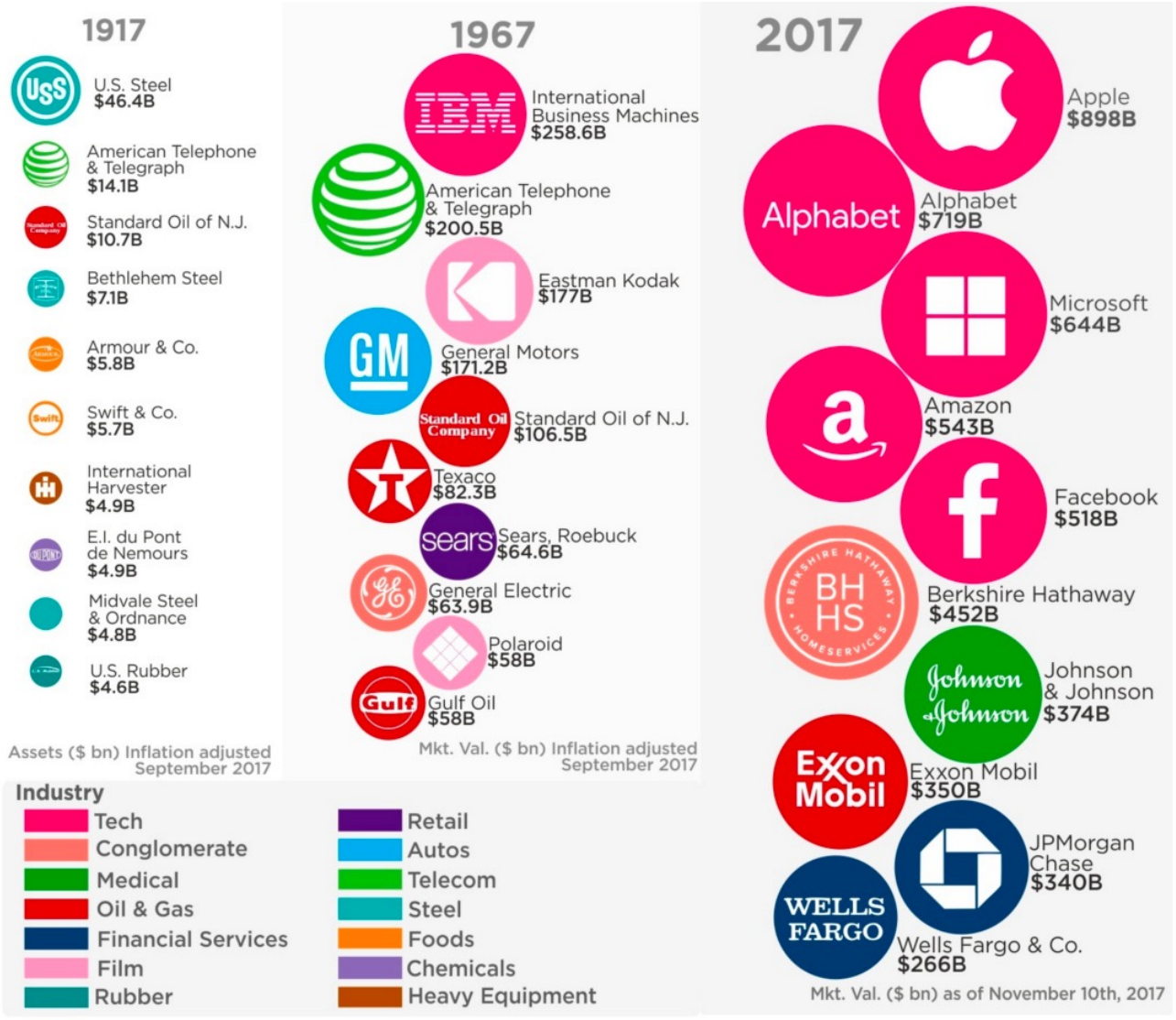
Societal impacts of new technology

- Unemployment
  - Making people more productive. Less demand for labor.
- Specific tasks in jobs are being eliminated



- AI is also creating new jobs, but…
  - Can your grandpa learn how to code?

Who are getting the largest piece of the technology pie?



Source and Article:
https://howmuch.net/articles/100-years-of-americas-top-10-companies
https://forbes.com

howmuch.net

2020:

Apple: 2.12T

Amazon: 1.59T

Alphabet: 1.22 T

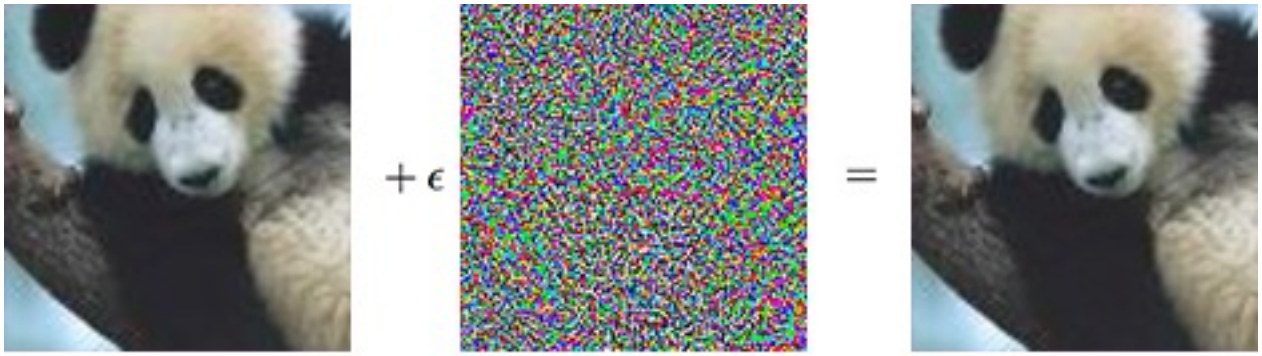
…

Tesla: 600 B +

GDP of Indonesia: 1.05 T

GDP of US: 20.5 T

Safety issue in deploying AI



Research in Responsible AI

- Issues about fairness
  - (A) I want my predictions to be calibrated on all subgroups
  - (B) I want the false-positive rate to be the same on all subgroups
  - (C) I want the false-negative rate to be the same on all subgroups

Impossibility theorem (Kleinberg et al. 2016): Except in trivial cases, any two of the above implies the third is impossible.

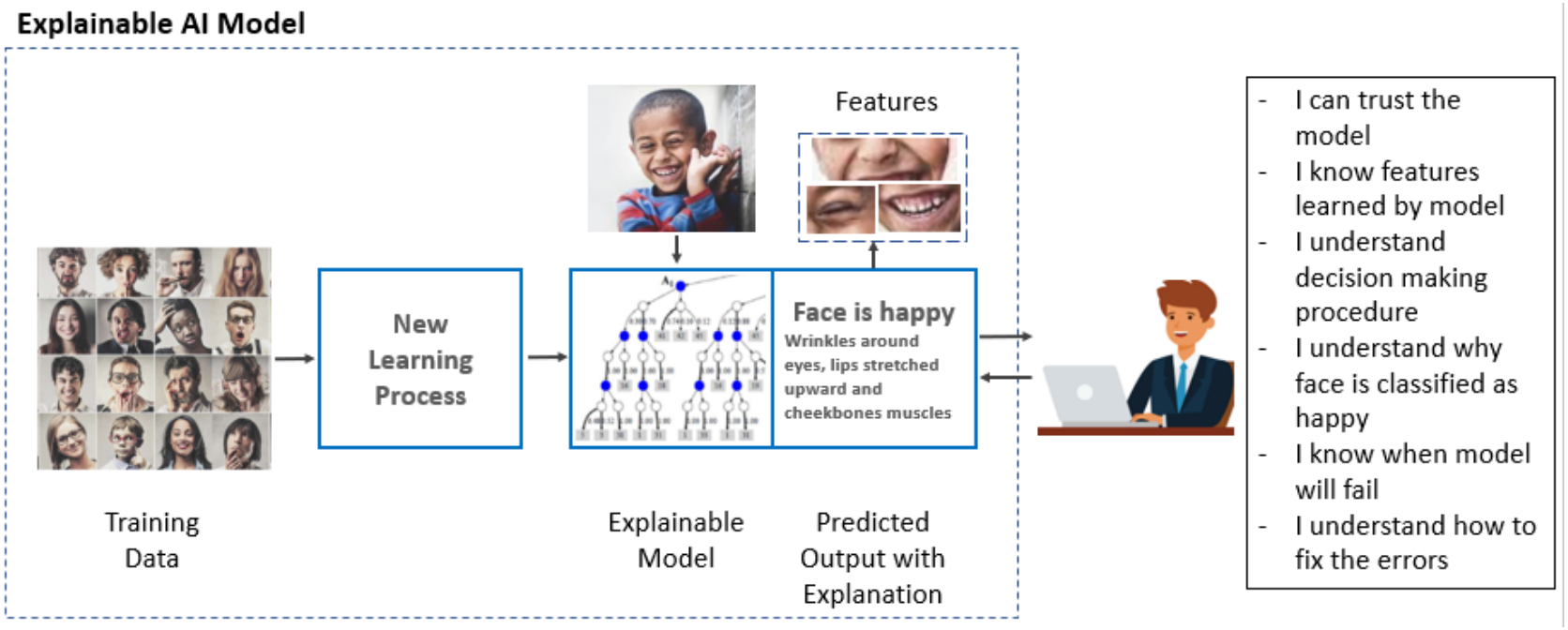
What is it that we want? How do we define fairness?

- For recidivism prediction?
- For medical diagnosis?
- Do human decision makers suffer from the same issue?

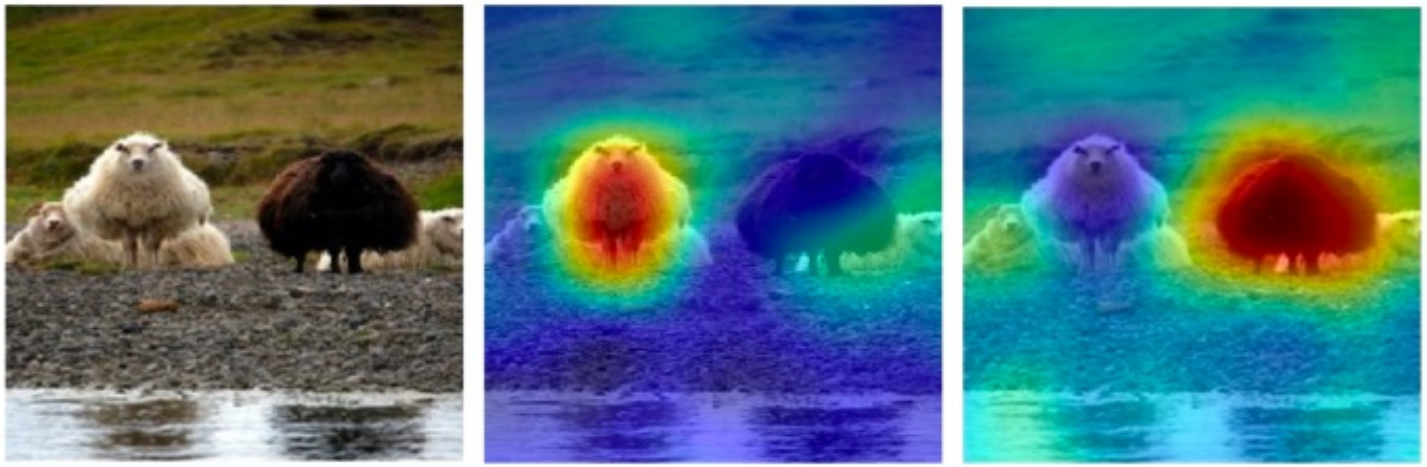
https://arxiv.org/pdf/1609.05807.pdf

Research in Responsible AI

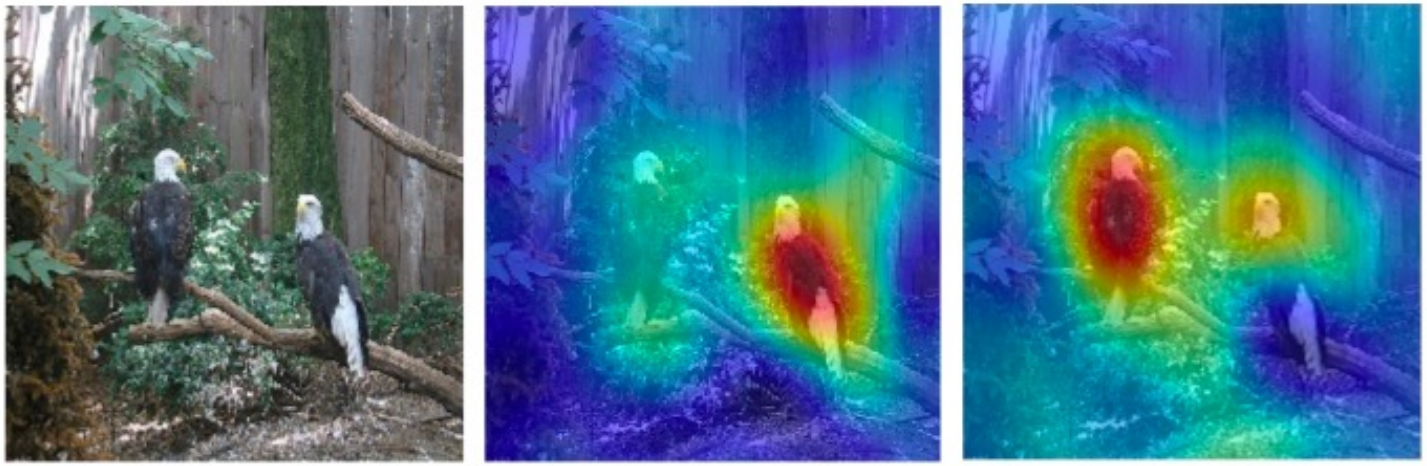
- Explanability of AI predictions



Another example on explainable AI predictions



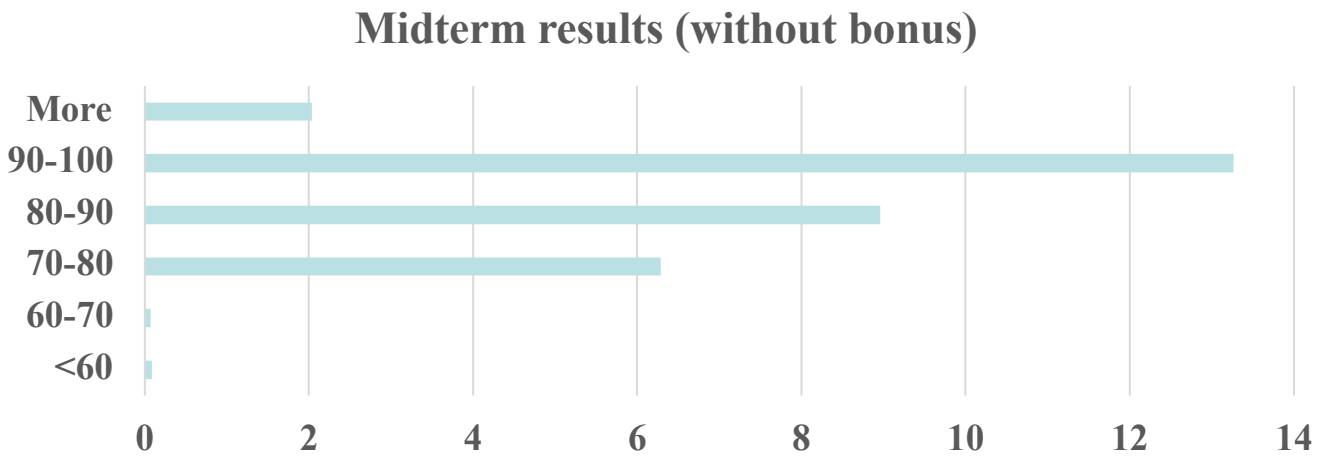
(a) Sheep - 26%, Cow - 17% (b) Importance map of '*sheep*' (c) Importance map of '*cow*'



(d) Bird - 100%, Person - 39% (e) Importance map of '*bird*' (f) Importance map of '*person*'

Research in Responsible AI

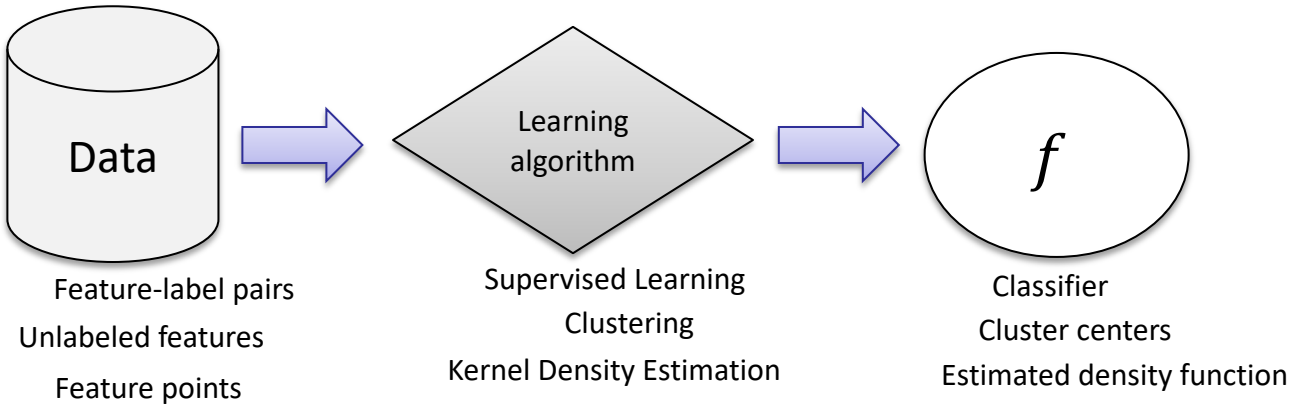
- Provable guarantees against identification in privacy

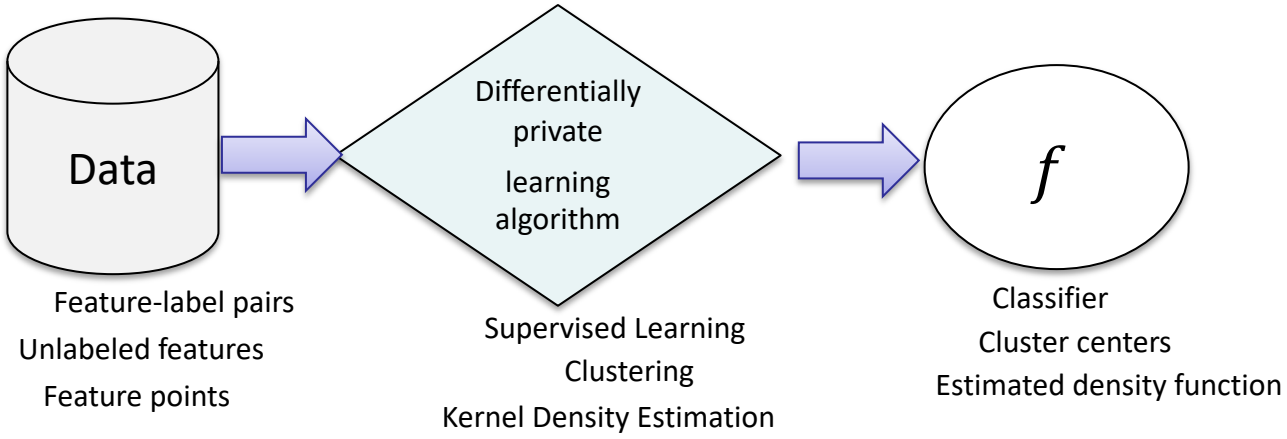


Differentially privately released midterm results from Fall 2020

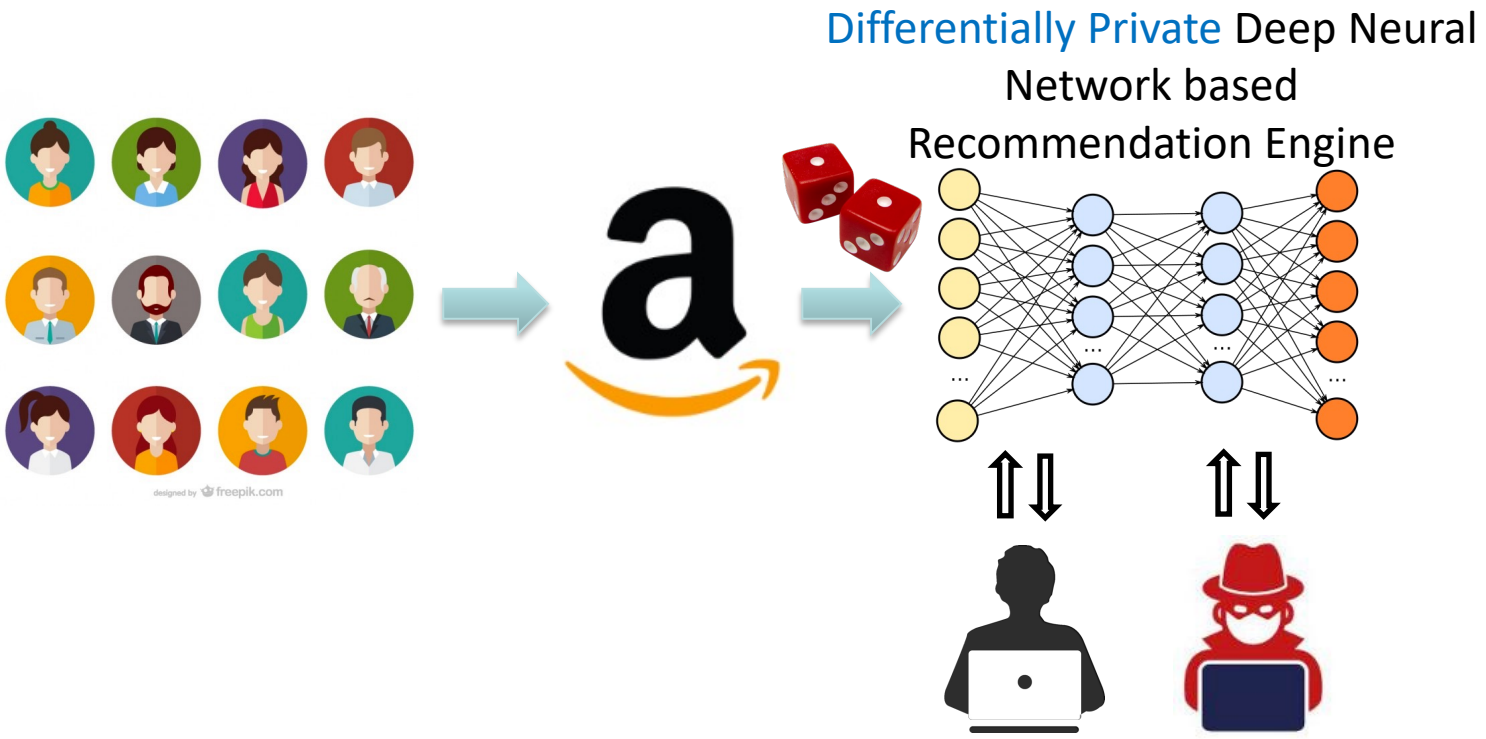
How does differential privacy work?

Differentially Private Machine Learning





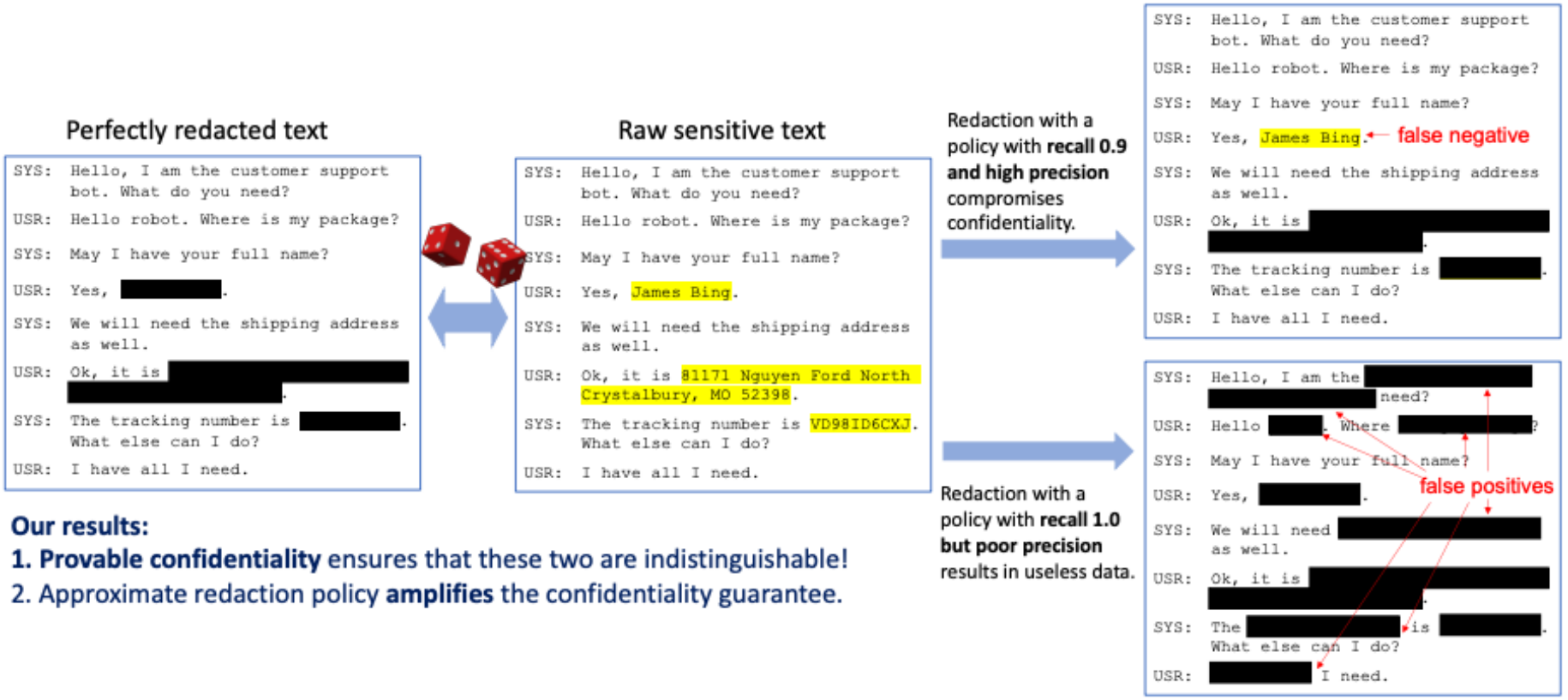
Example: Recommender System



"If your recommendation engine is private, then an adversary can't infer whether a particular user was present"

Research in Responsible AI

- Differential privacy implies prevents language models from generating sensitive parts of the training data.



See our recent work: https://arxiv.org/abs/2205.01863

Invisible watermarks for detecting LLM-generated text and to prevent model-stealing attacks

Protecting Language Generation Models via Invisible Watermarking

Xuandong Zhao [1] Yu-Xiang Wang [1] Lei Li [1]

Abstract

Language generation models have been an increasingly powerful enabler for many applications. Many such models offer free or affordable API access, which makes them potentially vulnerable to model extraction attacks through distillation. To protect intellectual property (IP) and ensure fair use of these models, various techniques such as lexical watermarking and synonym replacement have been proposed. However, these methods can be nullified by obvious countermeasures such as "synonym randomization". To address this issue, we propose GINSEW, a novel method to protect text generation models from being stolen through distillation. The key idea of our method is to inject secret signals into the probability vector of the decoding steps for each target token. We can then detect the secret message by probing a suspect model to tell if it is distilled from the protected one. Experimental results show that GINSEW can effectively identify instances of IP infringement with minimal impact on the generation quality of protected APIs. Our method demonstrates an absolute improvement of 19 to 29 points on mean

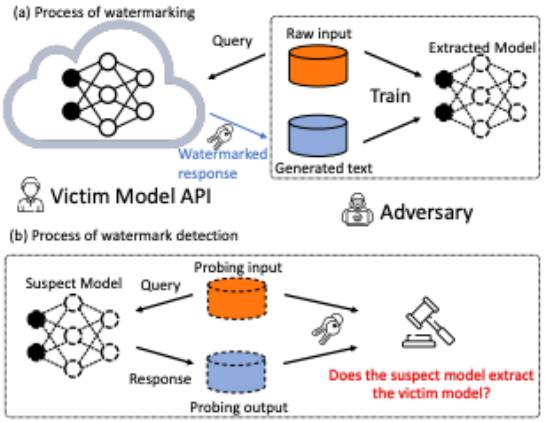


Figure 1. Overview of the process of watermarking and the process of watermark detection. The victim model API embeds watermarks in the response to input queries from the adversaries. The API owner can then use a key to verify if the suspect model has been distilled from the victim model.

See our recent work: https://arxiv.org/abs/2302.03162

Examples of watermarked text: Can you tell the differences?

Example 1: Unwatermarked: first of all, because the successes of the Marshall Plan have been overstated. Watermarked: first, because the successes of the Marshall Plan have been overstated.
Example 2: Unwatermarked: because life is not about things Watermarked: because life isn't about things
Example 3: Unwatermarked: i was at these meetings i was supposed to go to Watermarked: i was at the meetings i was supposed to go to

Table 4. Watermarked examples

- The seemingly arbitrary choices of words are actually deliberate (determined by a secret key that only we – who injects the watermark --- know).

UCSB Activities in Responsible AI

Final words

- With greater power comes great responsibility.
  - Ethics in AI, Privacy, fairness, social impacts
  - Transparency, robustness, explanability
  - AI for good causes

- These are very complex issues
  - Are humans good decision makers? Are there implicit biases?
  - Can we explain our decisions
  - Should we regulate? How? To what extent?

- The future is in your hands. Be a good driver!