Artificial Intelligence

CS 165A

Dec 8, 2020

Instructor: Prof. Yu-Xiang Wang

Today

→ First Order Logic (Inference)

→ Responsible AI

Logistic notes

- Online ESCI Survey
  - So far only 25% students completed the survey.
  - Please take a few moments to complete it. The deadline is **Dec 11**

- Take home final:
  - Similar to the midterm. (I will provide more instructions on Thursday)

- My general grading policy
  - Look at the distribution, decide on grades thresholds
  - Add your bonus points
  - Previously, 40-50% A and A- depending on the distribution

Logistic notes

- HW2 fairness issue:
  - Issue with students getting the solutions from the discussion classes, then submit the solutions.
  - We needed to provide feedback for HW2 before you guys do the midterm.
  - This is an issue that applies to very few students.
- But to be fair to everyone:
  - you are allowed to submit a "Correction" for the HW2 to get 50% back.
  - No deadline, you can do it after the final. I will do two versions of the letter grades (with or without including this correction). You will get the higher of the two.

Recap: Resolution Rule is just chaining of implications

$$\frac{p \vee q, \quad \neg q \vee r}{p \vee r}$$

Propositional calculus resolution

Remember: $\boldsymbol{p \Rightarrow q \Leftrightarrow \neg p \vee q}$, so let's rewrite it as:

$$\frac{\neg p \Rightarrow q, \quad q \Rightarrow r}{\neg p \Rightarrow r} \qquad \text{or} \qquad \frac{a \Rightarrow b, \quad b \Rightarrow c}{a \Rightarrow c}$$

Resolution is really the "chaining" of implications.

Recap: Conversion to Conjunctive Normal Form: CNF

- Resolution rule is stated for conjunctions of disjunctions
- Question:
  - Can every statement in PL be represented this way?
- Answer: Yes
  - Can show every sentence in propositional logic is equivalent to conjunction of disjunctions
    - Conjunctive normal form (CNF)
- Procedure for obtaining CNF
  - Replace $(P \Leftrightarrow Q)$ with $(P \Rightarrow Q)$ and $(Q \Rightarrow P)$
  - Eliminate implications: Replace $(P \Rightarrow Q)$ with $(\neg P \vee Q)$
  - Move $\neg$ inwards: $\neg\neg$, $\neg(P \vee Q)$, $\neg(P \wedge Q)$
  - Distribute $\wedge$ over $\vee$, e.g.: $(P \wedge Q) \vee R$ becomes $(P \vee R) \wedge (Q \vee R)$ [What about $(P \vee Q) \wedge R$?]
  - Flatten nesting: $(P \wedge Q) \wedge R$ becomes $P \wedge Q \wedge R$

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*)**

Recap: FOL Syntax

- Symbols
  - Object symbols (constants): *P, Q, Fred, Desk, True, False, …*
    - These refer to *things*
  - **Predicate** symbols: *Heavy, Smart, Mother, …*
    - These are *true or false statements* about objects: Smart(*rock*)
  - **Function** symbols: *Cosine, IQ, MotherOf, …*
    - These return objects, exposing *relations:* IQ(*rock*)
  - Variables: *x, y, λ, …*
    - These represent unspecified objects
  - Logical connectives to construct complex sentences: ¬, ∧, ∨, ⇒, ⇔
  - Quantifiers: ∀ (universal), ∃ (existential)
  - Equality: =
- Usually variables will be lower-case, other symbols capitalized

*__Terms__: Constants, variables, (output of) functions

Recap: Universal and Existential Quantiifiers

- Quantifiers: ∀ (universal), ∃ (existential)

- ∀ **<variables> <sentence>**
  - ∀ *x* – *"For all x…"*
  - ∀ *x, y* – *"For all x and y…"*
  - *"All instances must satisfy …"*

- ∃ **<variables> <sentence>**
  - ∃ *x* – *"There exists an x such that…"*
  - ∃ *x, y* – *"There exist x and y such that…"*
  - "There is at least one such example such that …"

- Scope, order, nesting of quantifiers
  - ∃ *x* ∀ *y* Loves(*x, y*)
  - ∀ *y* ∃ *x* Loves(*x, y*)

Recap: Kinship domain (cont.)

Assertions ("Add this sentence to the KB")

TELL(KB, $\forall\, m, c$ Mother(c) = m $\Leftrightarrow$ Female(m) $\wedge$ Parent(m, c))

TELL(KB, $\forall\, w, h$ Husband(h, w) $\Leftrightarrow$ Male(h) $\wedge$ Spouse(h, w))

TELL(KB, $\forall\, x$ Male(x) $\Leftrightarrow$ $\neg$Female(x))

TELL(KB, Female(Mary) $\wedge$ Parent(Mary, Frank) $\wedge$ Parent(Frank, Ann))

- Note: TELL(KB, S1 $\wedge$ S2) $\equiv$ TELL(KB, S1) and TELL(KB, S2) (because of and-elimination and and-introduction)

Queries ("Does the KB entail this sentence?")

ASK(KB, Grandparent(Mary, Ann)) $\rightarrow$ True

ASK(KB, $\exists\, x$ Child(x, Frank)) $\rightarrow$ True

- But a better answer would be $\rightarrow$ { x / Ann }
- This returns a **substitution** or **binding**

Implementing ASK: Inference

- We want a sound and complete inference algorithm so that we can produce (or confirm) *entailed* sentences from the KB

$$KB \models \alpha \qquad\qquad KB \vdash \alpha$$

- The **resolution** rule, along with a complete search algorithm, provides a complete inference algorithm to confirm or refute a sentence α in propositional logic (Sec. 7.5)
  - Based on *proof by contradiction* (refutation)

- Refutation: To prove that the KB entails P, assume $\neg P$ and show a contradiction:

$$(KB \wedge \neg P \Rightarrow \text{False}) \equiv (KB \Rightarrow P)$$

Prove this!

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

To Add to These Rules

- **Modus Ponens** or **Implication-Elimination**: (From an implication and the premise of the implication, you can infer the conclusion.)

$$\frac{\alpha \Rightarrow \beta, \quad \alpha}{\beta}$$

- **And-Elimination**: (From a conjunction, you can infer any of the conjuncts.)

$$\frac{\alpha_1 \wedge \alpha_2 \wedge \ldots \wedge \alpha_n}{\alpha_i}$$

- **And-Introduction**: (From a list of sentences, you can infer their conjunction.)

$$\frac{\alpha_1, \alpha_2, \ldots, \alpha_n}{\alpha_1 \wedge \alpha_2 \wedge \ldots \wedge \alpha_n}$$

- **Or-Introduction**: (From a sentence, you can infer its disjunction with anything else at all.)

$$\frac{\alpha_i}{\alpha_1 \vee \alpha_2 \vee \ldots \vee \alpha_n}$$

- **Double-Negation Elimination**: (From a doubly negated sentence, you can infer a positive sentence.)

$$\frac{\neg\neg\alpha}{\alpha}$$

- **Unit Resolution**: (From a disjunction, if one of the disjuncts is false, then you can infer the other one is true.)

$$\frac{\alpha \vee \beta, \quad \neg\beta}{\alpha}$$

- **Resolution**: (This is the most difficult. Because β cannot be both true and false, one of the other disjuncts must be true in one of the premises. Or equivalently, implication is transitive.)

$$\frac{\alpha \vee \beta, \quad \neg\beta \vee \gamma}{\alpha \vee \gamma} \qquad \text{or equivalently} \qquad \frac{\neg\alpha \Rightarrow \beta, \quad \beta \Rightarrow \gamma}{\neg\alpha \Rightarrow \gamma}$$

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)

We can formulate the logical inference problem as a search problem.

- Formulate a **search process**:
  - Initial state
    - KB
  - Operators
    - Inference rules
  - Goal test
    - KB contains S
- What is a node?
  - KB + new sentences (generated by applying the inference rules)
  - In other words, the new state of the KB
- What kind of search to use?
  - I.e., which node to expand next?
- How to apply inference rules? $\alpha \Rightarrow \beta$
  - Need to match the premise pattern α

Question: What's our goal here?

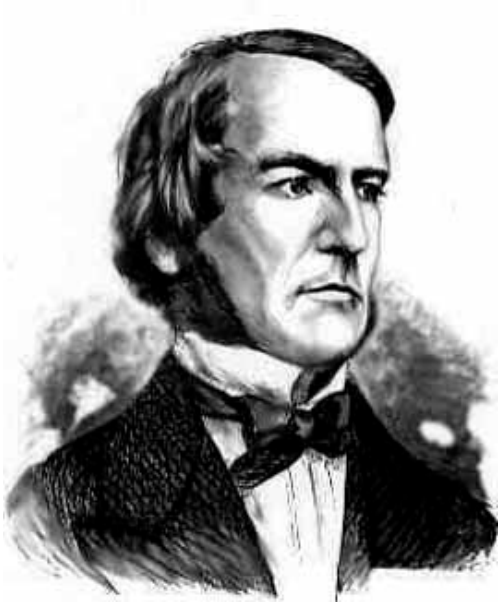
Historical AI figure in Logical Reasoning

- Built a calculating machine that could add and subtract (which Pascal's couldn't)
- But his dream was much grander – to reduce human reasoning to a kind of calculation and to ultimately build a machine capable of carrying out such calculations
- Co-inventor of the calculus

Gottfried Leibniz (1646-1716)

"For it is unworthy of excellent men to lose hours like slaves in the labor of calculation which could safely be relegated to anyone else if the machine were used."

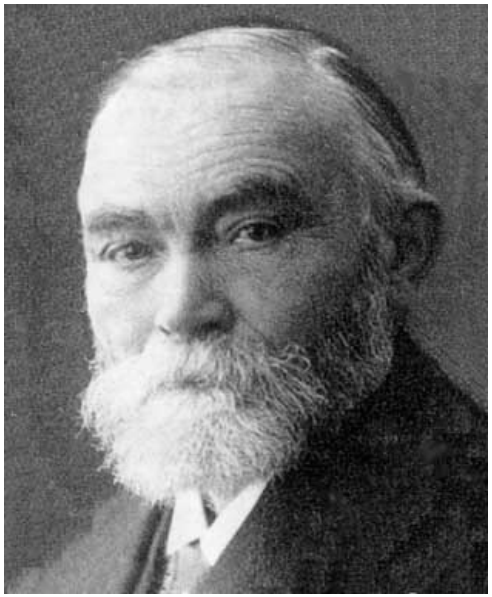
George Boole (1815-1864) British



- More than 100 years later, he didn't know about Leibniz, but proceeded to bring to life part of Leibniz' dream
- His insight: Logical relationships are expressible as a kind of algebra
  - Letters represent classes (rather than numbers)
  - So logic can be viewed as a form of mathematics
- Published *The Laws of Thought*

- He extended Aristotle's simple syllogisms to a broader range of reasoning
  - Syllogism: Premise_1, Premise_2 → Conclusion
  - His logic: Propositional logic

Gottlob Frege (1848-1925) German



- He provided the first fully developed system of logic that encompassed all of the deductive reasoning in ordinary mathematics.
- He intended for logic to be the *foundation* of mathematics – all of mathematics could be based on, and derived from, logic
- In 1879 he published *Begriffsschrift*, subtitled "A formula language, modeled upon that of arithmetic, for pure thought"
  - This can be considered the ancestor of all current computer programming languages
  - Made the distinction between *syntax* and *semantics* critical

- He invented what we today call predicate calculus (or first-order logic)

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

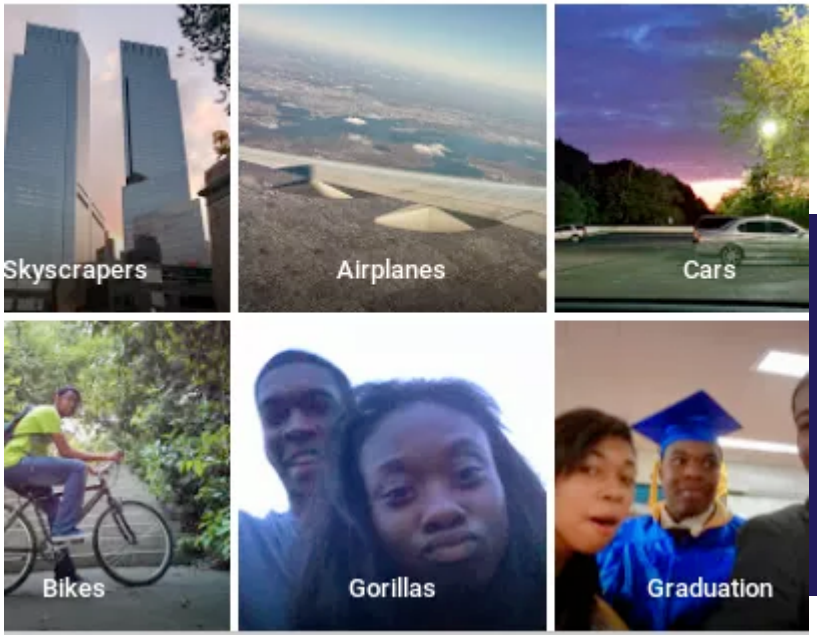
- More higher level intelligence
  - Logic is coming back
  - But more learning based than rule-based
- More stateful systems, more reinforcement learning
  - Causal modelling and reasoning
- More AI in the non-iid environment
  - Structured
  - Adversarial
- More forms of agent's perception
  - Weak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

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

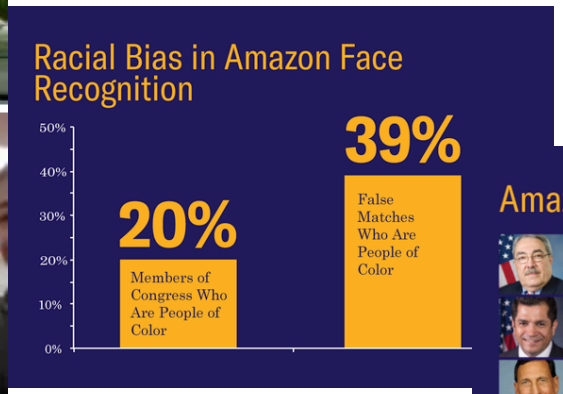
A face recognition system

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

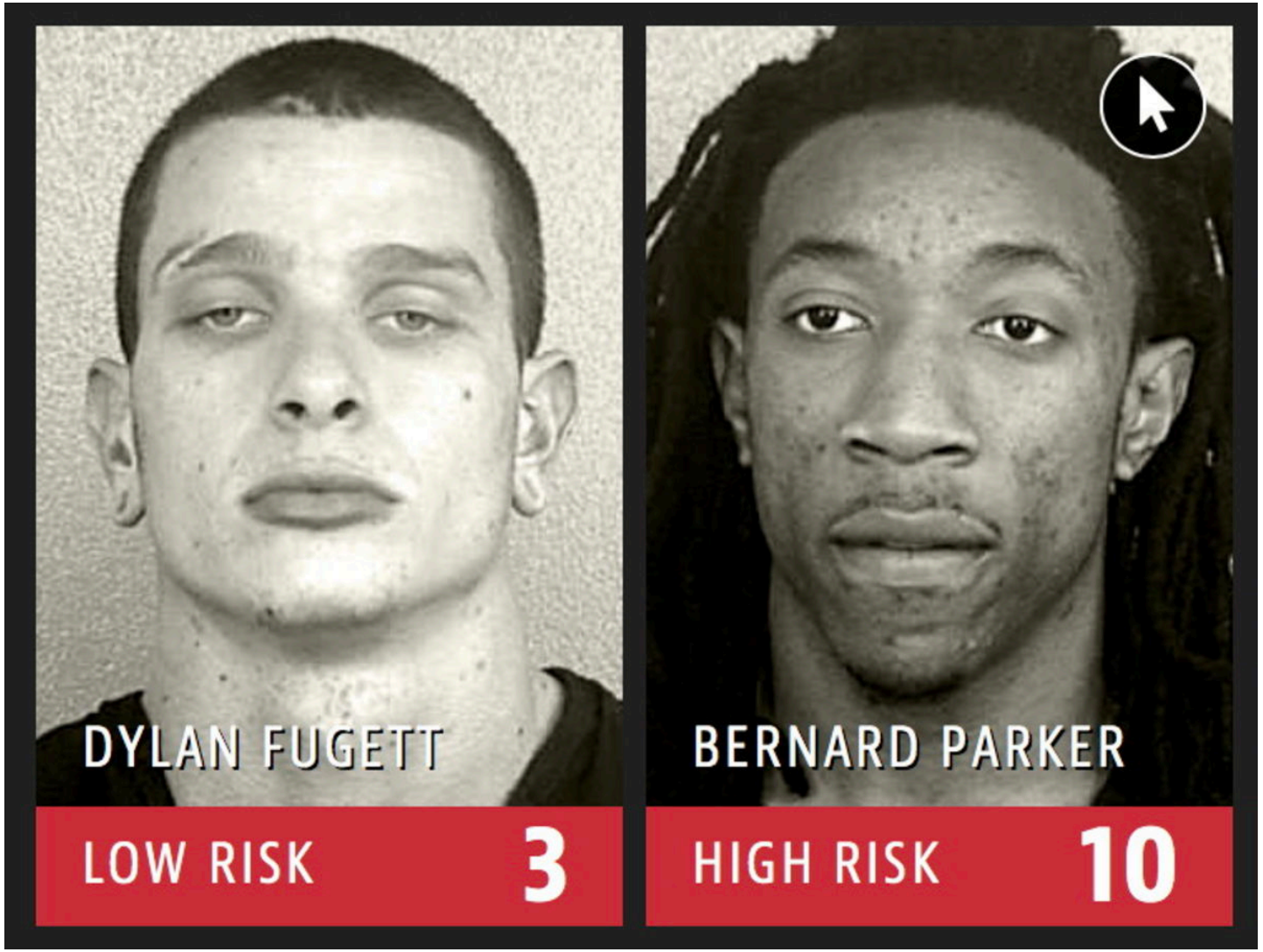
Fairness challenges in AI systems / AI for decision making



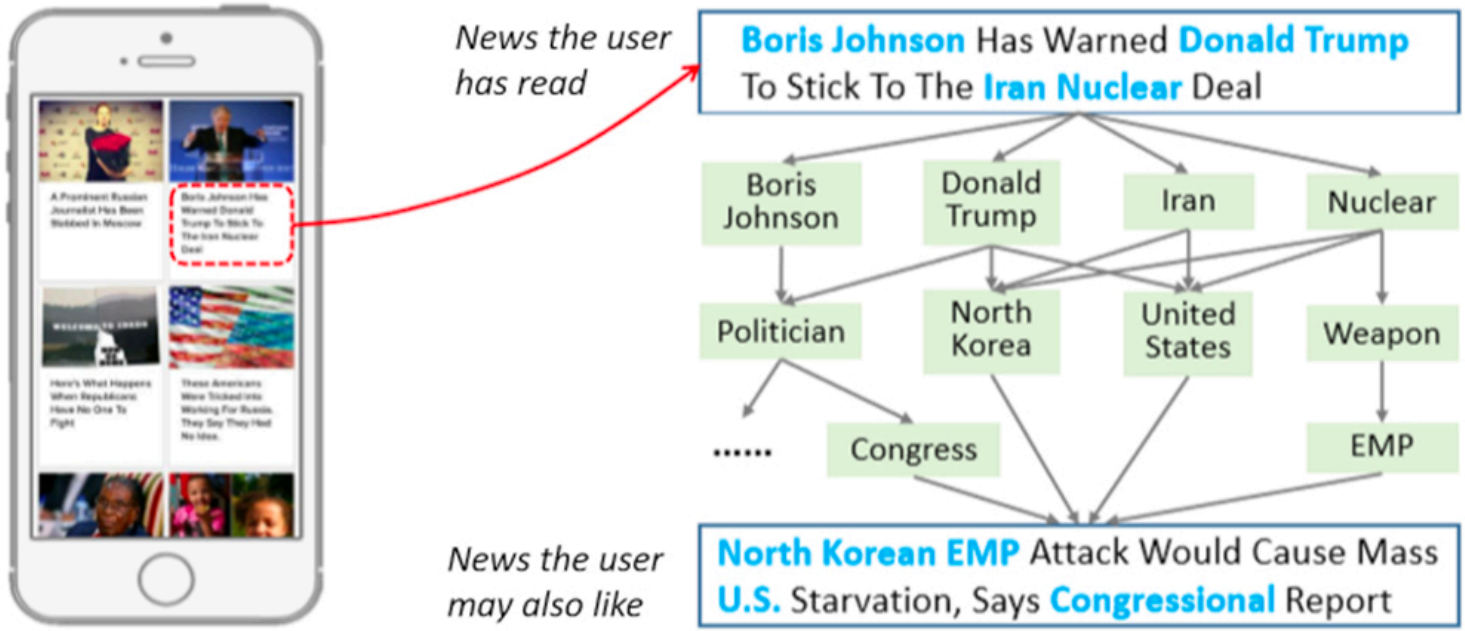
Google's image recognition system



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.

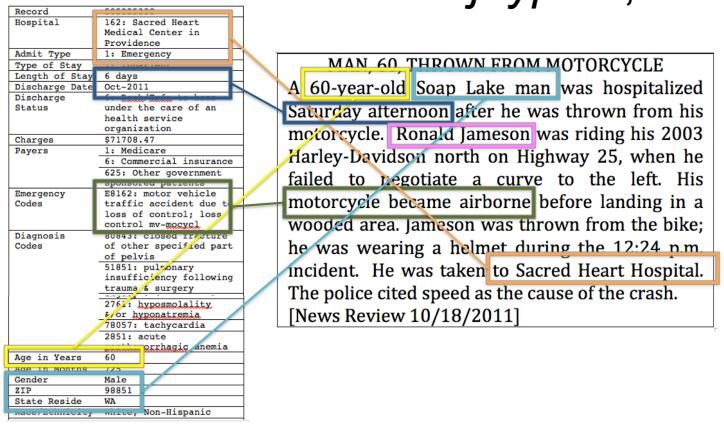
Privacy issues in data collection and learning

"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

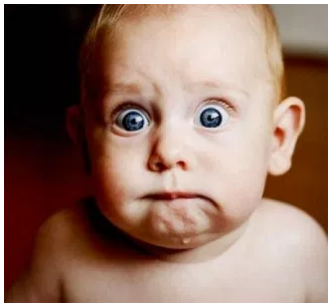
Nicholas Carlini, *University of California, Berkeley*
Chang Liu, *University of California, Berkeley*
Jernej Kos, *National University of Singapore*
Úlfar Erlingsson, *Google Brain*
Dawn Song, *University of California, Berkeley*

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

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

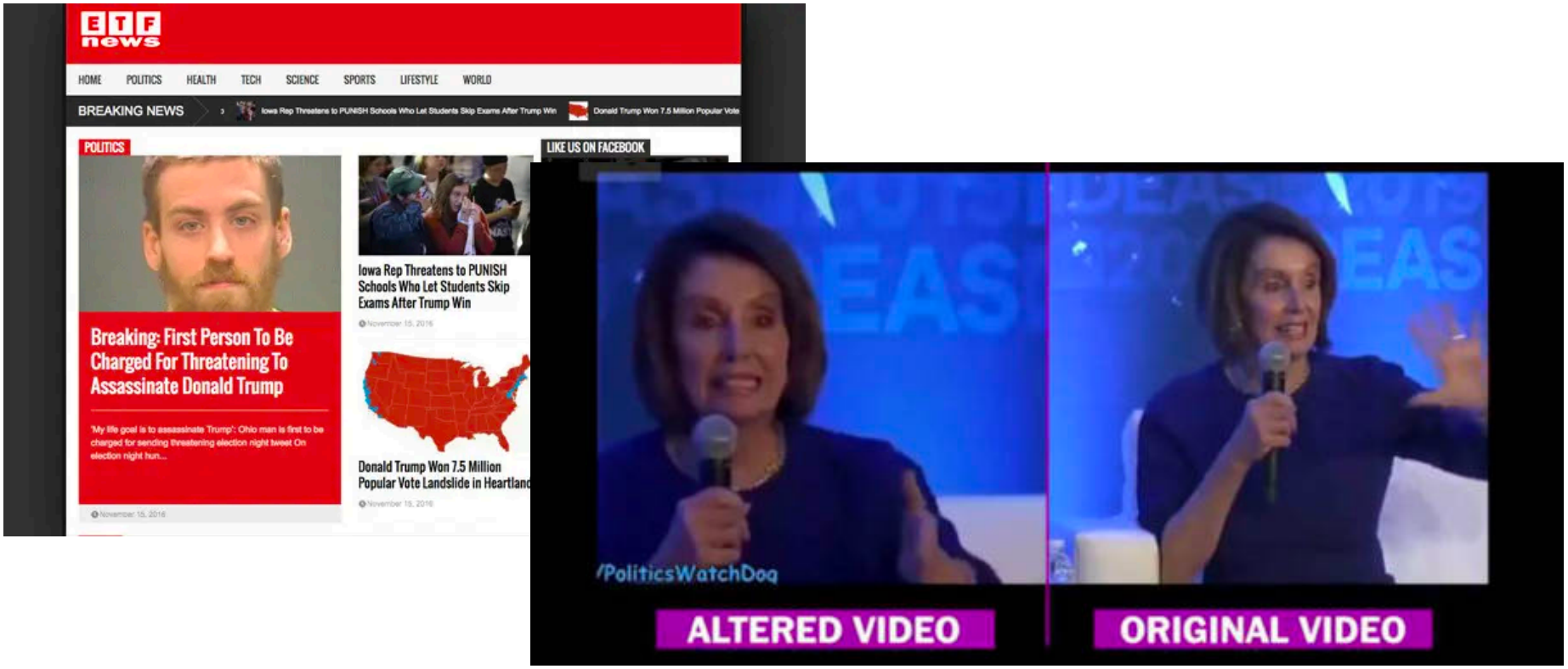


- I can't keep personal data for more than three weeks?

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

How about my machine learning models trained on user data?

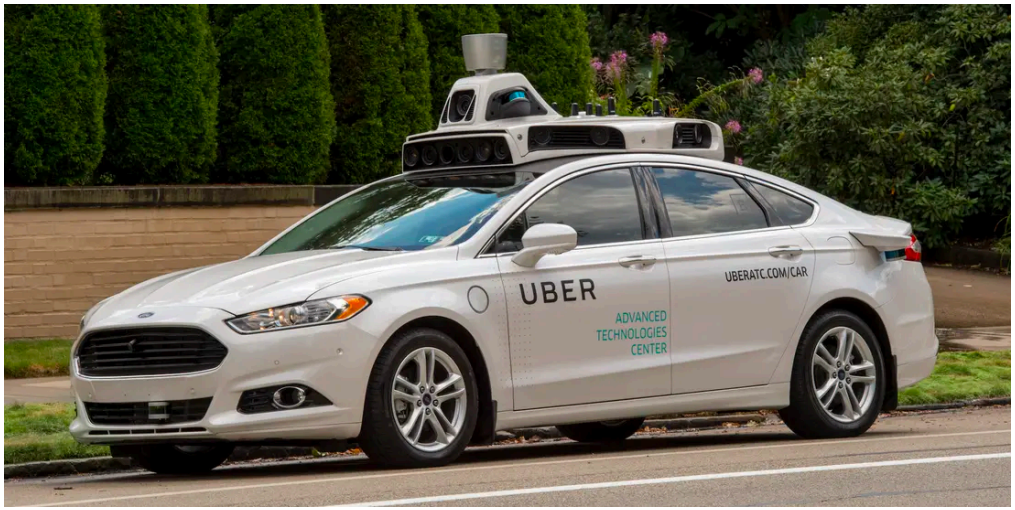
Fake-news, fake voice, fake video

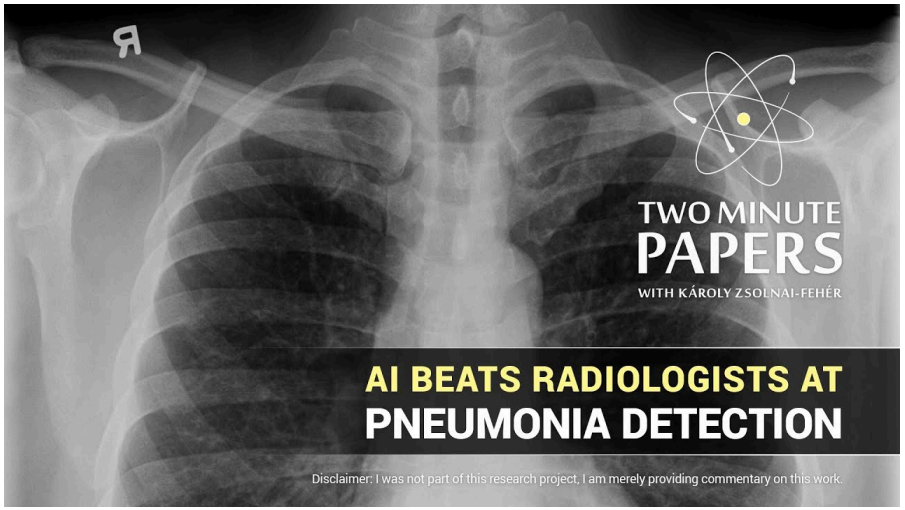


- How to tell if something is true or false?
- How to attribute a crime with factual evidence when people can just claim it's fake?

Societal impacts of new technology

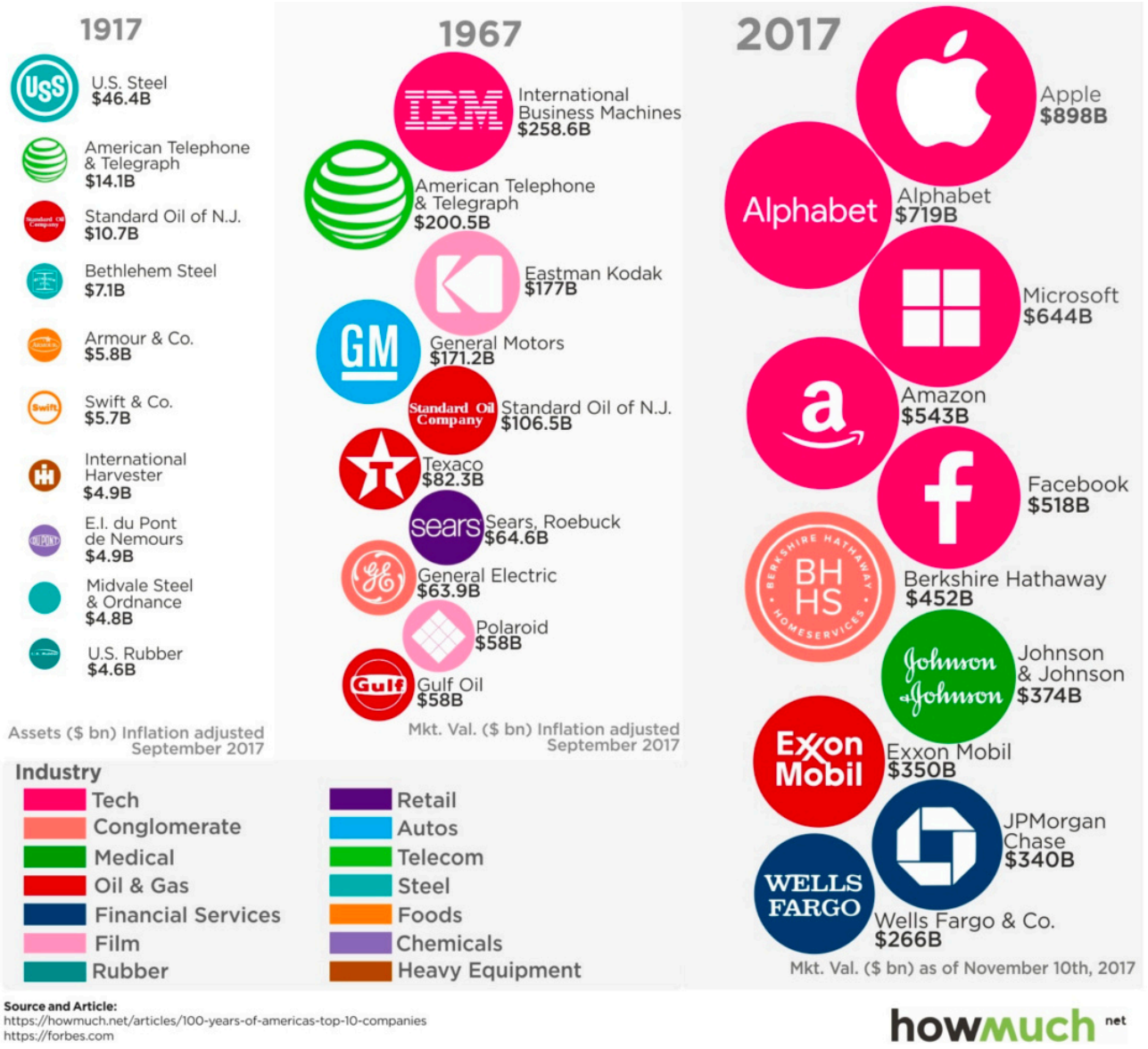
- Unemployment
  - Making people more productive. Less demand for labor.
- Specific tasks in jobs are being eliminated





- 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

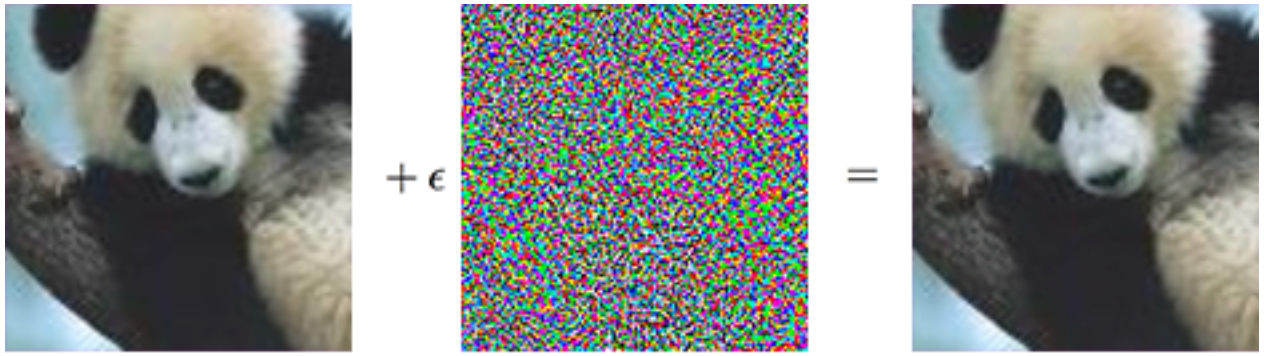
…

Tesla: 600 B +

GDP of Indonesia: 1.05 T

GDP of US: 20.5 T

Safety issue in deploying AI



"panda"
57.7% confidence

"gibbon"
99.3% confidence

classified as
Stop Sign

misclassified as
Max Speed 100

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.

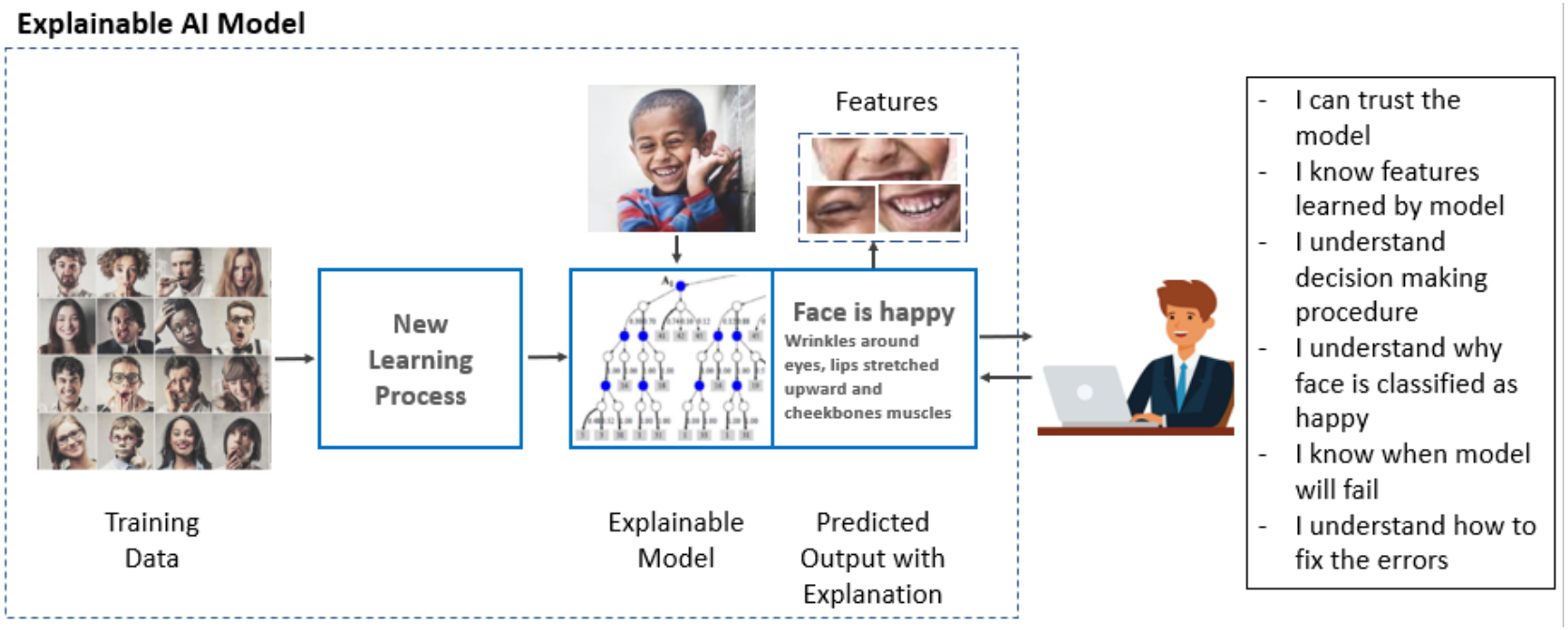
What is it that we want? How do we define fairness?

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

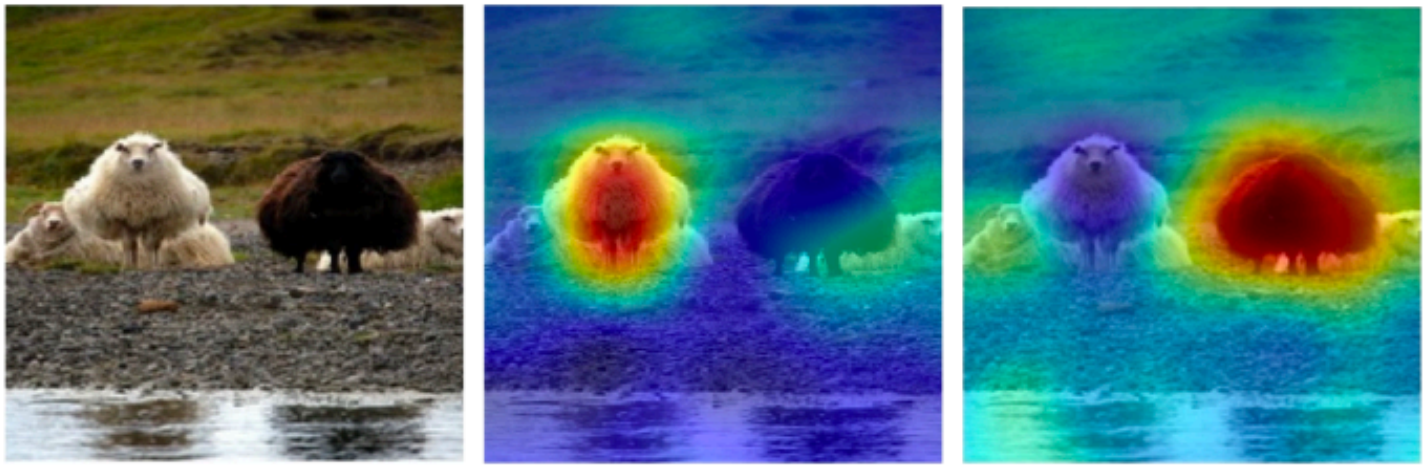
https://arxiv.org/pdf/1609.05807.pdf

Research in Responsible AI

- Explanability of AI predictions



Another example on explainable AI predictions

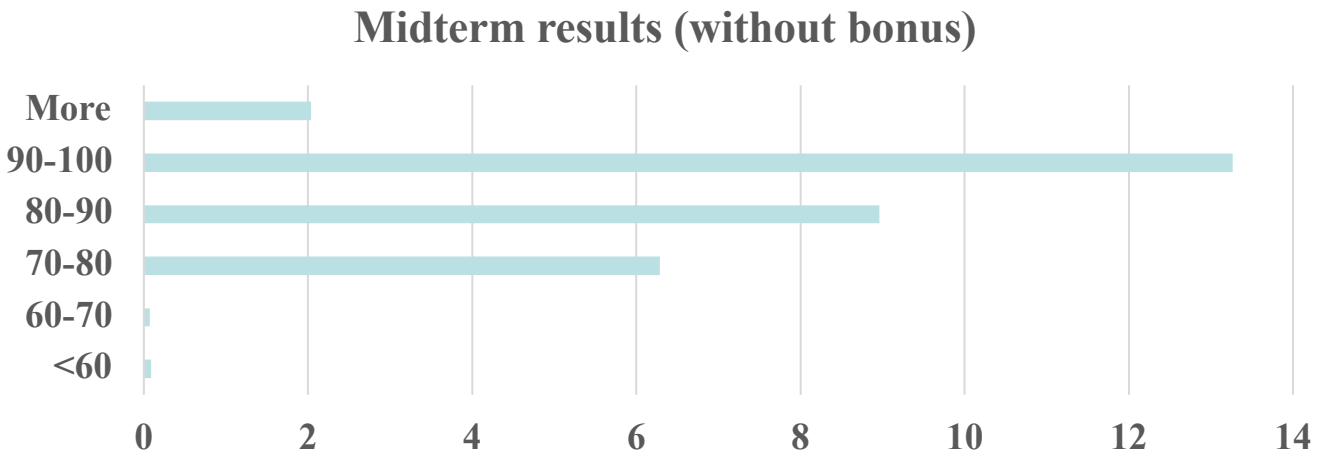


(a) Sheep - 26%, Cow - 17% (b) Importance map of '*sheep*' (c) Importance map of '*cow*'

(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

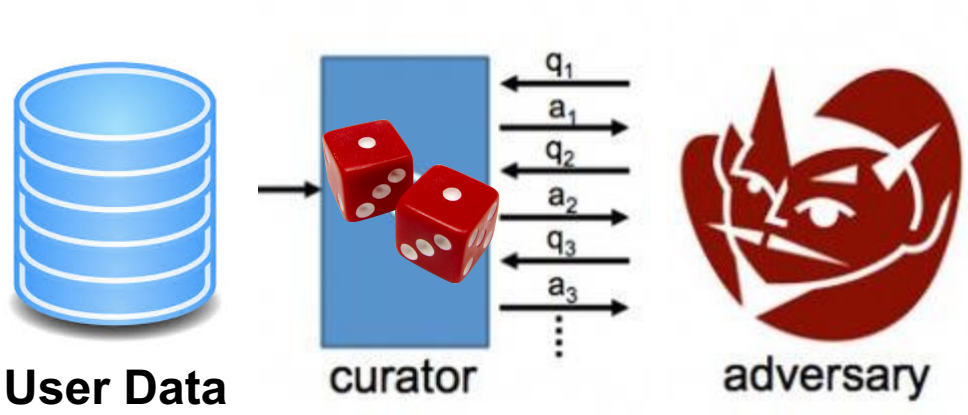


How does differential privacy work?

History of privacy research

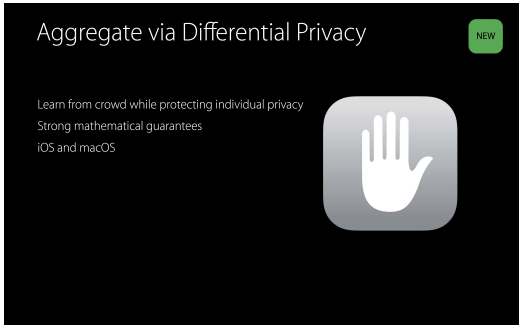
- Statistical disclosure control
  - [Duncan et. al., Hundepool et. al., since 1970s]
- k-anonymity, l-divergence, t-closeness
  - [Sweeny, Machanavajjhala et. al., Li et. al., 2002-2007]
- Differential privacy
  - [Dwork, McSherry, Nissim, Smith, 2006++]

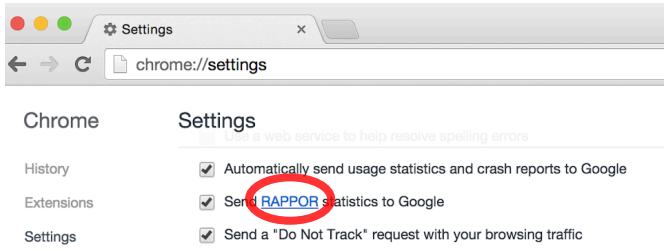
Differential Privacy (Dwork et. al., 2006; Gödel Prize 2017) makes no assumption on adversaries



- Almighty Adversary
  - **Arbitrary side info**, arbitrary computational power.
- Interpretable, quantifiable, composable.

Deployments of Differential Privacy

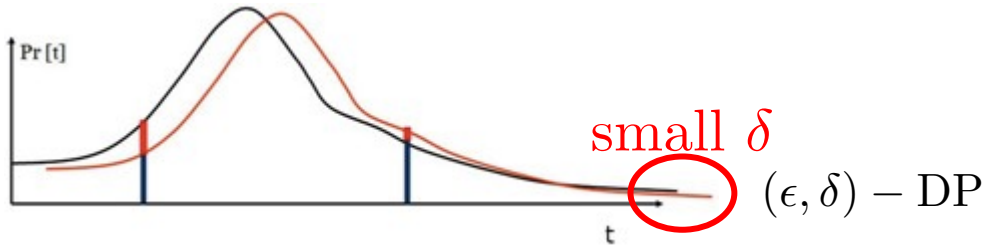




Formal definition of DP

- Let Z,Z' be any two datasets that differ only by one user, and A is a randomized algorithm. We say A is ε-DP if for all output h

$$\sup_{Z,Z':d(Z,Z')\leq 1} \sup_{h\in\mathcal{H}} \log \frac{p_{h\sim\mathcal{A}(Z)}(h)}{p_{h\sim\mathcal{A}(Z')}(h)} \leq \epsilon$$



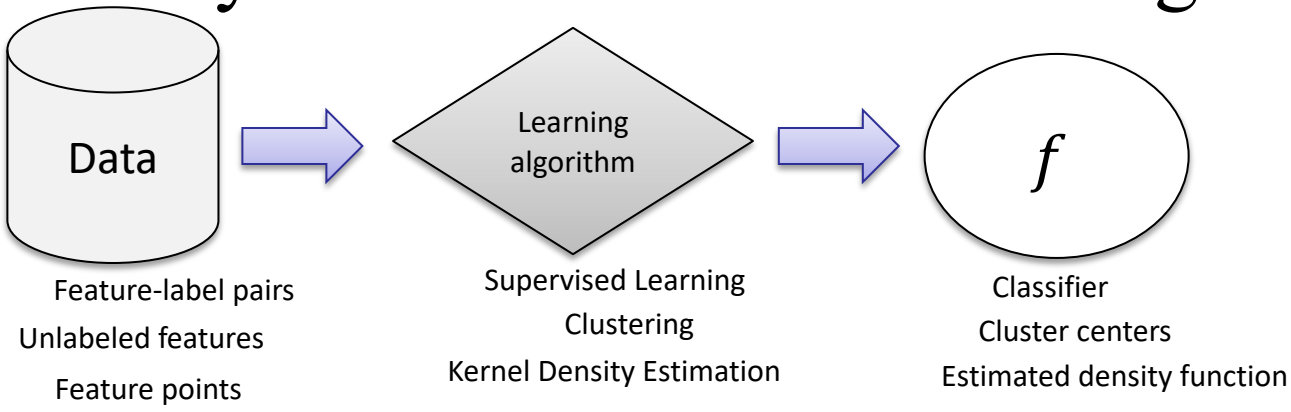
Example: Who voted for Trump?

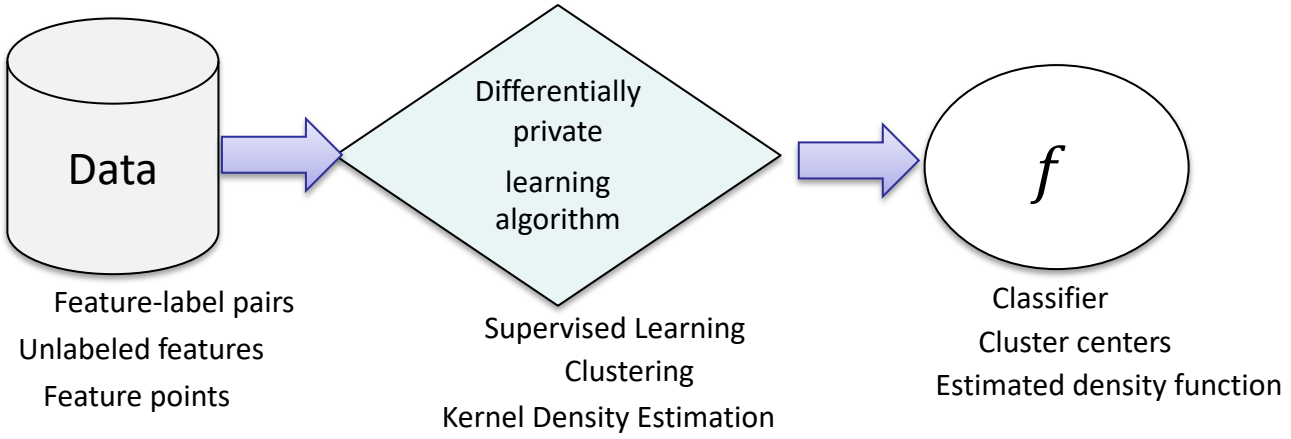
- How many people in this room voted for Trump?

- Let's say the answer is 15.
  - If I know the political view of everybody else except my TA.
  - Then I can easily infer his choice.

- DP releases: 15 + noise.

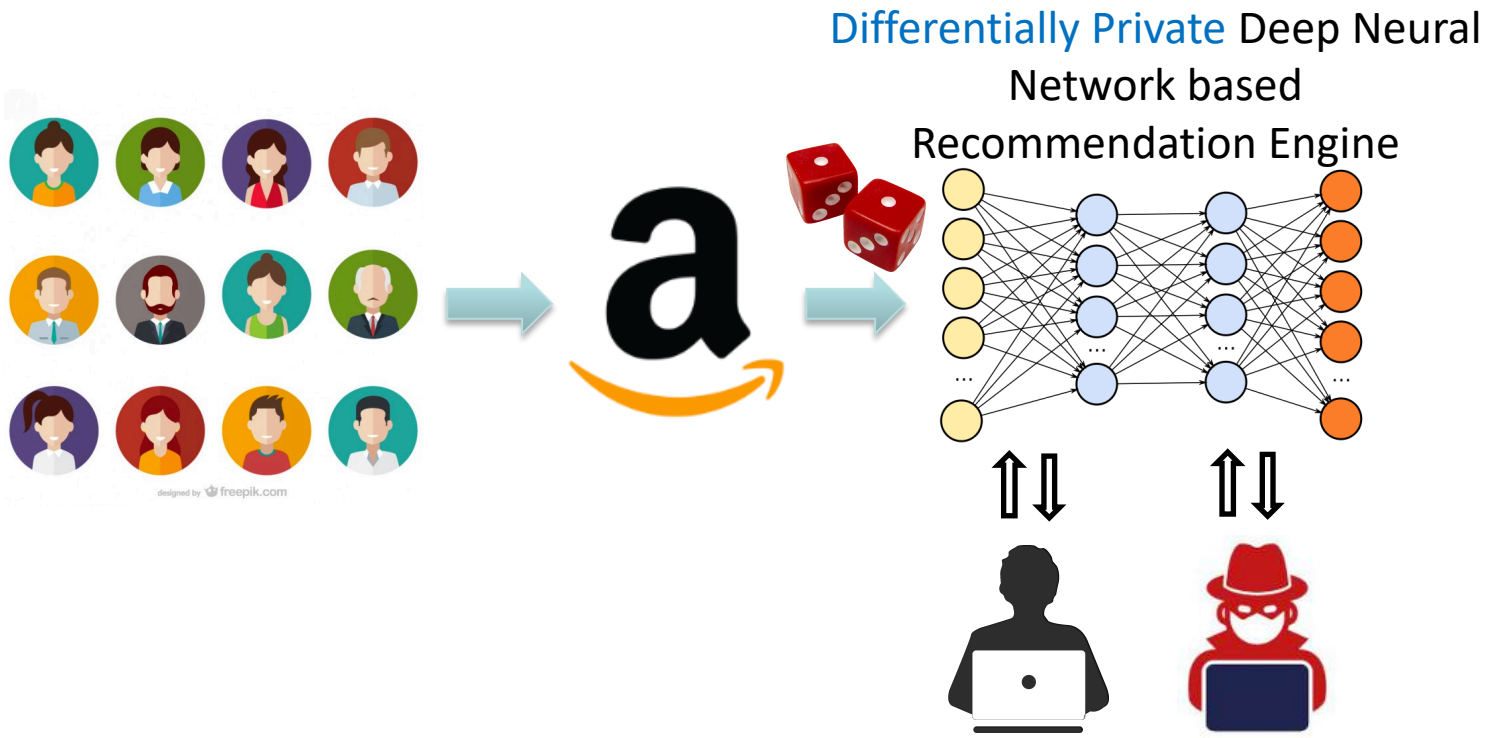
Differentially Private Machine Learning





Example: Recommender System

Example: Recommender System



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

A closely related setting: Federated Learning

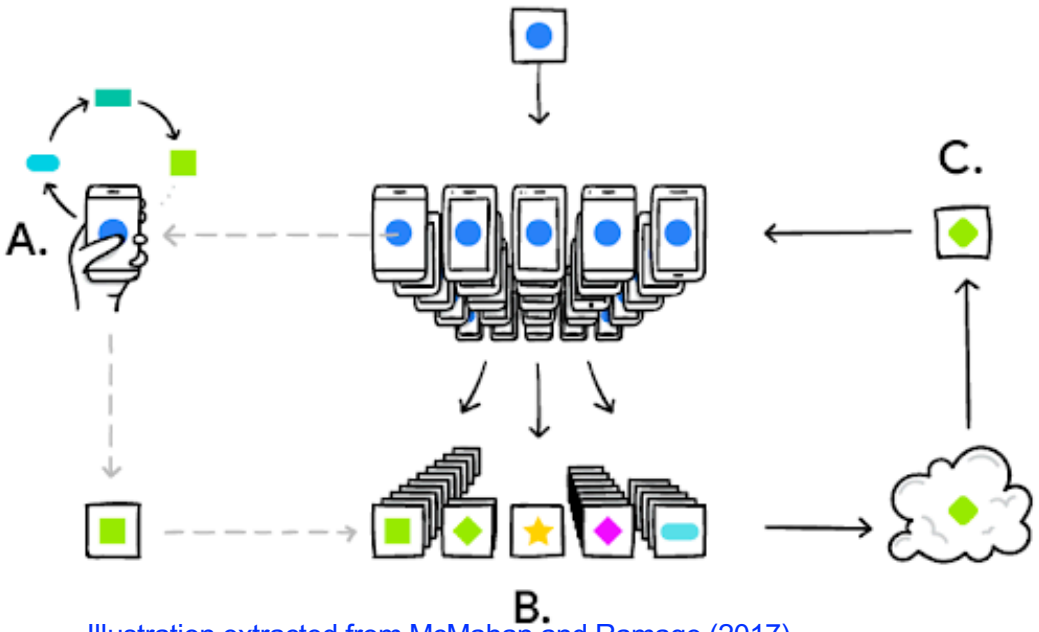


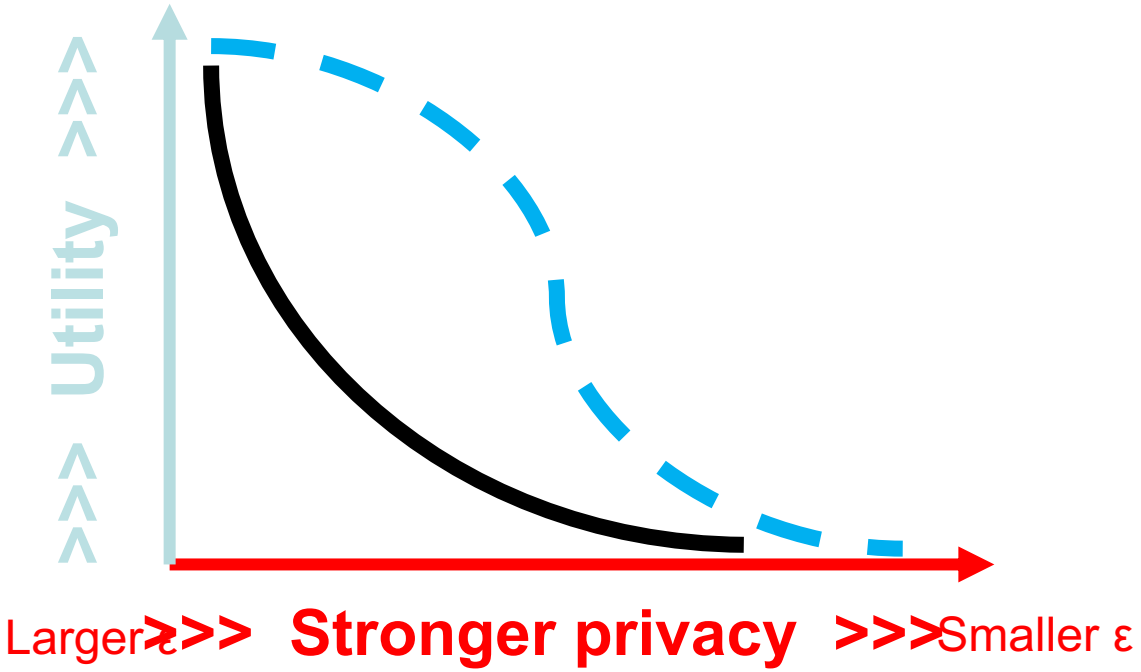
Illustration extracted from McMahan and Ramage (2017)

Additional considerations

- Communication cost
- Size of the model
- Rounds of adaptivity

Really a very scalable and practical setting!

Privacy-Utility Tradeoff



Privacy can be **for free**!

- In the Trump example, we are interested to see the mean of the population, but all we get is n iid samples.

Statistical error

$$O_P\left(\frac{1}{\sqrt{n}}\right)$$

Error due to DP

$$O_P\left(\frac{1}{n\epsilon}\right)$$

The list of problems where DP is almost free

- Linear regression (W., 2018)
  - Statistical Rate: n^{-1} error from DP $(n\epsilon)^{-2}$
- Nonparametric regression: (Wang, Baby, W., 2019)
  - Statistical Rate: $n^{-\frac{2s}{2s+1}}$ error from DP $(n\epsilon)^{-\frac{4s}{2s+1}}$
- Many more in the fast-growing literature:
  - Strongly convex ERM, hypothesis testing… and so on.

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!

- Next lecture: review session for the final