Artificial Intelligence

CS 165A

May 31, 2022

Instructor: Prof. Yu-Xiang Wang

Today

→ First Order Logic (Inference)

→ Responsible AI

Logistic notes

- Online ESCI Survey
  - Nearly half the students completed the survey.
  - We can do better! The deadline is **Jun 3 (This Friday)**
  - Please take a moment to complete your feedback!

- Project 3 due **this Thursday**
  - Extra instructor office hour at 2 pm today

- Final exam **next Monday** 12 – 3
  - Open book (no digital devices)
  - Twice the time but only slightly longer than the midterm
  - More details on the Thursday lecture

Recap: Resolution Rule is just chaining of implications

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

Propositional calculus resolution

Recap: Resolution Rule is just chaining of implications

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

Propositional calculus resolution

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

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

Recap: Resolution Rule is just chaining of implications

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

Propositional calculus resolution

Remember: $\boldsymbol{p} \Rightarrow \boldsymbol{q} \Leftrightarrow \neg \boldsymbol{p} \vee \boldsymbol{q}$, 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 <u>quantifiers</u> can bind <u>variables</u> in sentences
  - Universal quantifier ($\forall$)
  - Existential quantifier ($\exists$)
  - Variables: *x, y, z, a, joe, table…*

- Examples
  - $\forall x$ **Beautiful** (x)
  - $\exists x$ **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: $x, y, \lambda, \ldots$
    - These represent unspecified objects
  - Logical connectives to construct complex sentences: $\neg, \wedge, \vee, \Rightarrow, \Leftrightarrow$
  - Quantifiers: $\forall$ (universal), $\exists$ (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
  - $\exists x \, \forall y$ Loves(x, y)
  - $\forall y \, \exists x$ Loves(x, y)

Recap: Kinship domain (cont.)

Assertions ("Add this sentence to the KB")

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

TELL(KB, $\forall\, w, h\;\; \text{Husband}(h, w) \;\Leftrightarrow\; \text{Male}(h) \wedge \text{Spouse}(h, w)$)

TELL(KB, $\forall\, x\;\; \text{Male}(x) \;\Leftrightarrow\; \neg\text{Female}(x)$)

TELL(KB, Female(Mary) ∧ Parent(Mary, Frank) ∧ Parent(Frank, Ann))

- Note: TELL(KB, S1 ∧ S2) ≡ 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)) → True

ASK(KB, $\exists\, x\;\; \text{Child}(x, \text{Frank})$) → True

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

Implementing ASK: Inference

Implementing ASK: Inference

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

$$\mathrm{KB} \models \alpha \qquad\qquad \mathrm{KB} \vdash \alpha$$

Implementing ASK: Inference

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

$$\mathrm{KB} \models \alpha \qquad\qquad \mathrm{KB} \vdash \alpha$$

- 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)

Implementing ASK: Inference

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

$$\text{KB} \models \alpha \qquad\qquad \text{KB} \vdash \alpha$$

- 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 ¬P and show a contradiction:

Implementing ASK: Inference

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

$$\mathrm{KB} \models \alpha \qquad\qquad \mathrm{KB} \vdash \alpha$$

- The **resolution** rule, along with a complete search algorithm, provides a <u>complete inference algorithm</u> 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 ¬P and show a contradiction:

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

$$((\alpha \vee \beta) \wedge (\gamma \vee \mathrm{False}) \wedge \cdots) \wedge \neg P \Rightarrow \mathrm{False}$$

Implementing ASK: Inference

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

$$\text{KB} \models \alpha \qquad\qquad \text{KB} \vdash \alpha$$

- The **resolution** rule, along with a complete search algorithm, provides a <u>complete inference algorithm</u> 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 ¬P and show a contradiction:

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

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

Substitution and variable binding

- Notation for substitution:

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

Substitution and variable binding

- Notation for substitution:
  - SUBST (**Binding list**, **Sentence**) (x,y)
    - x Mary y Bob
    - Binding list: { *var* / ground term, *var* / ground term, … }
    - "ground term" = term with no variables

  - SUBST({*var*/gterm}, Func (*var*)) = Func (gterm)
    - SUBST (θ, p)

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

Three new inference rules using SUBST($\boldsymbol{\theta}$, $\boldsymbol{p}$)

$(e^x)' = e^x$

- Universal Instantiation

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

g – ground term

$\exists y$

$\frac{\partial}{\partial x}(y^x) = y^x$

- Existential Instantiation

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

Skolem Constant

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

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

Thm: There exists a sorting alg with $O(n \log n)$ time complexity

- Existential Introduction

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

v – variable not in α

g – ground term in α

Proof: "MergeSort" is $O(n \log n)$ $\Rightarrow$ thm is true $\square$

To Add to These Rules

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

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

- ◇ **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.)

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

- ◇ **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.)

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

- ◇ **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.)

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

Universal Instantiation – examples

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

g – ground term

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)

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)

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

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

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

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

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/$\underline{x}$}, 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?

Example of FOL Inference (3 min discussion)

- Problem: "The law says that it is a crime for an American to sell weapons to hostile nations. The country Nono, an enemy of America, has some Missiles, and all of its missiles were sold to it by Colonel West, who is American."

1. How to represent the KB?

p,q,r are variables i.e. $\forall$ p,q,r

① American (p) ∧ weapon(q) ∧ sells (p, q, r) ∧ hostile(r) ⇒ Criminal(p)

② $\forall x$ Missile(x) ∧ Owns(Nono, x) ⇒ Sells(West, x, Nono), ⑤ American(West)

③ Enemy(Nono, America), ④ $\exists x$ Missile(x) ∧ Owns(Nono, x)

$\forall x$ Missile(x) ⇒ Weapon(x)

$\forall x$ Enemy(x, America) ⇒ Hostile(x)

2. How to apply various inference rules to prove "Colonel West is criminal"?

Exist. Instantia.

④ ⇒ Missile(M1) ∧ Owns(Nono, M1) ~~Sells(West~~ ⑥

⑥ ∧ ② ⇒ Sells(West, M1, Nono)

Universal Instan.

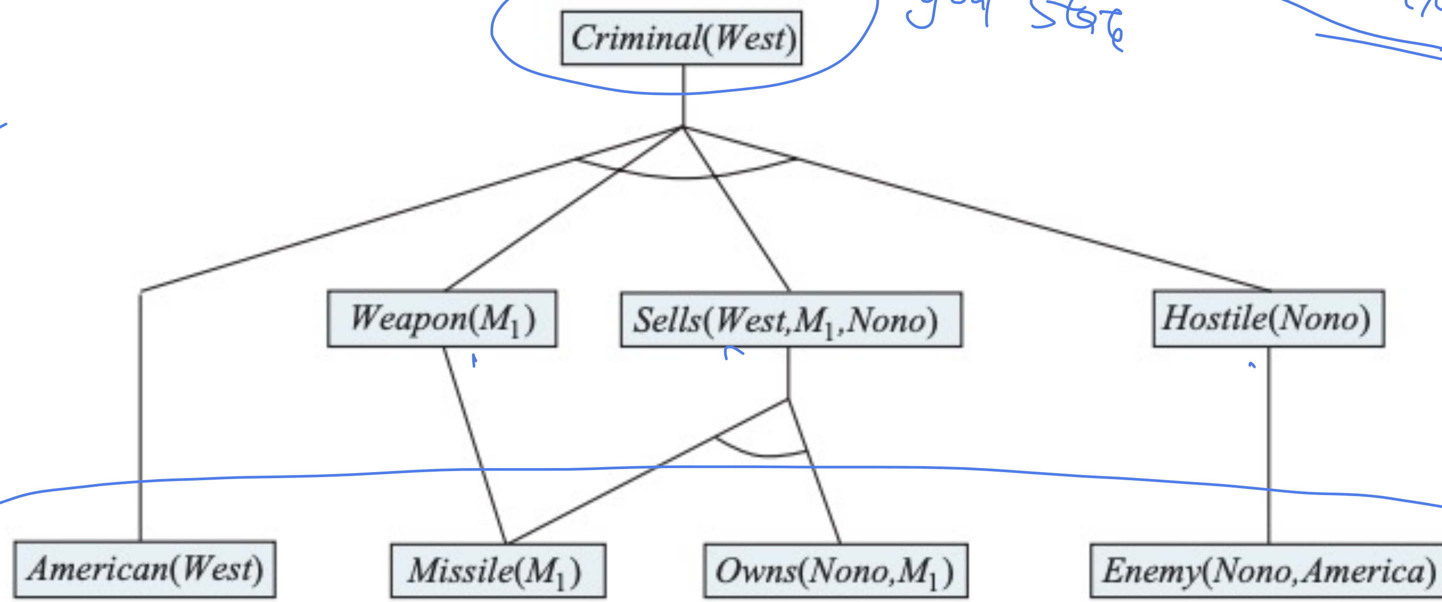
+③ ⇒ Hostile(Nono)

Apply ① ⇒ Criminal(West)

Example of a "Proof tree" of FOL inference

goal state

SG ← goal state (las?)



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

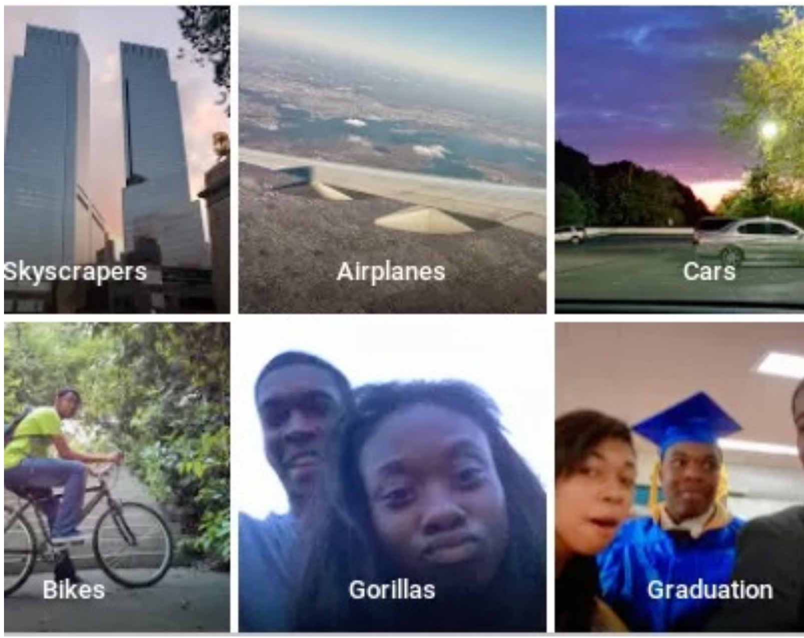
- More higher level intelligence
  - Logic is coming back
  - But more learning based than rule-based
- More stateful systems, more reinforcement learning
  - Causal modelling and reasoning
- More AI in the non-iid environment
  - Structured
  - Adversarial
- More forms of agent’s perception
  - Weak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

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

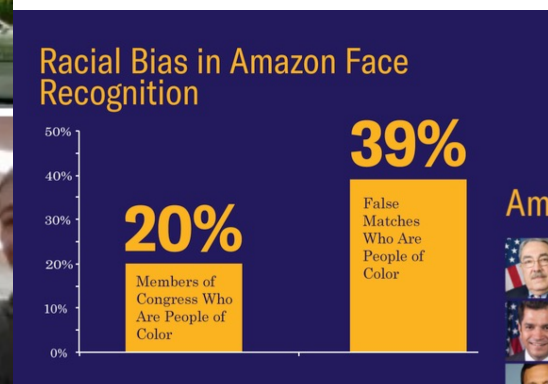
A face recognition system

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

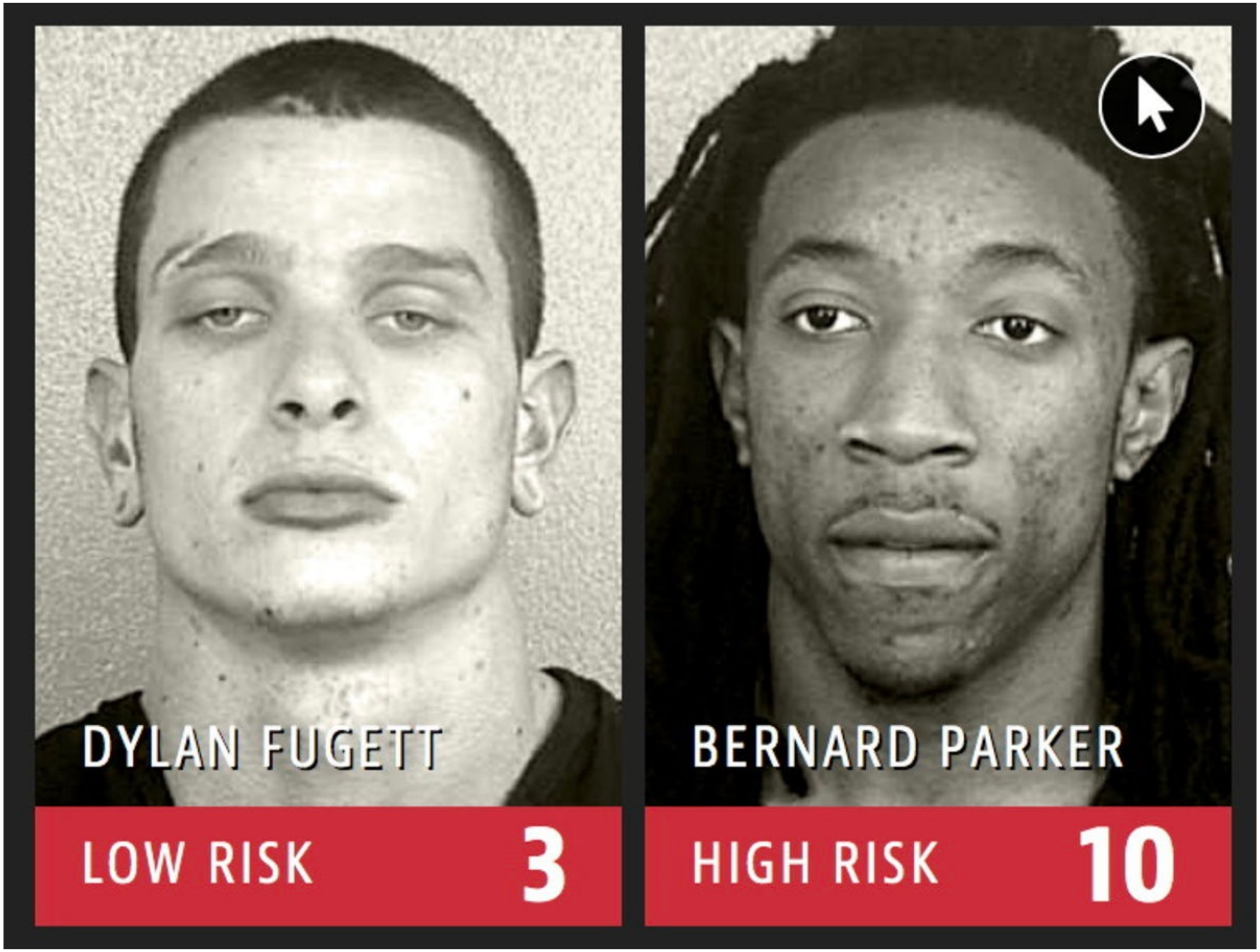
Fairness challenges in AI systems / AI for decision making



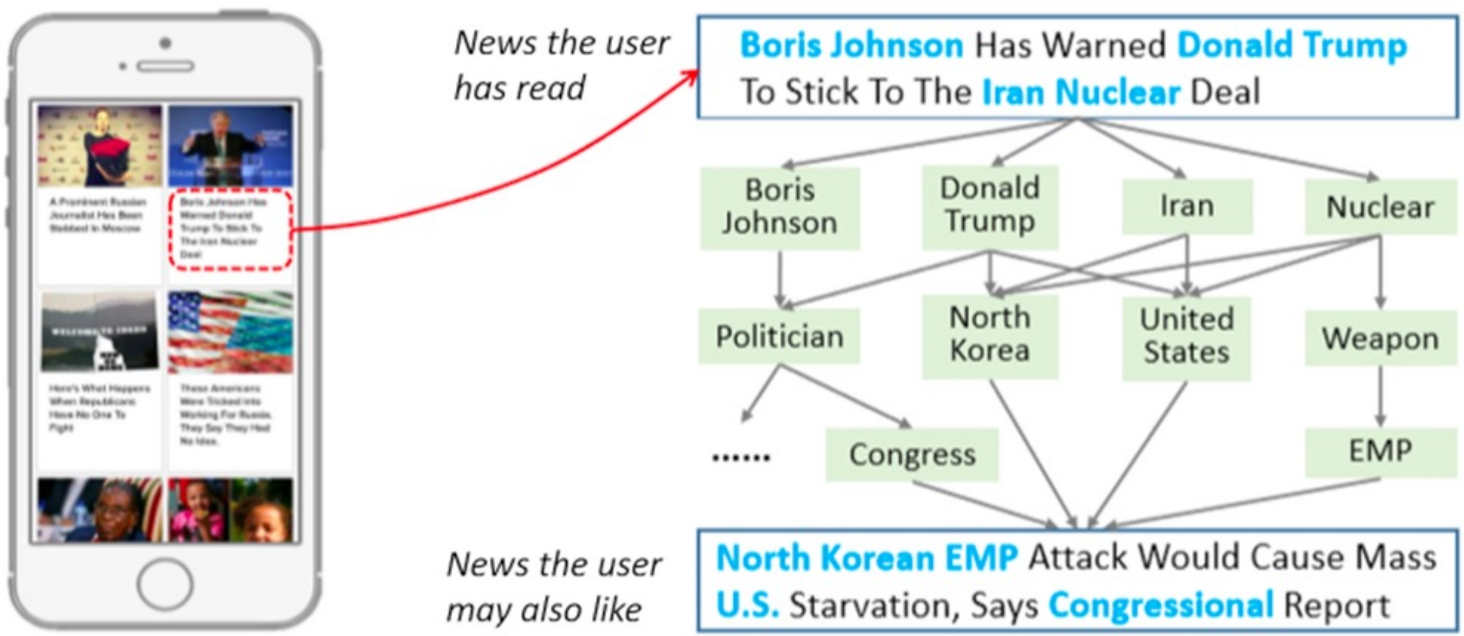
Google's image recognition system



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



Polarizing effects of news recommendation



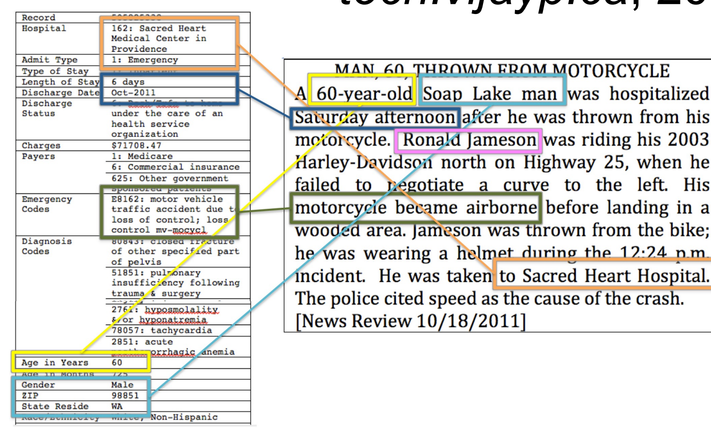
- Only what you like to read will be recommended to you.

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*

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



“Only You, Your Doctor, and Many Others May Know”

L. Sweeney. *Technology Science, 2015*

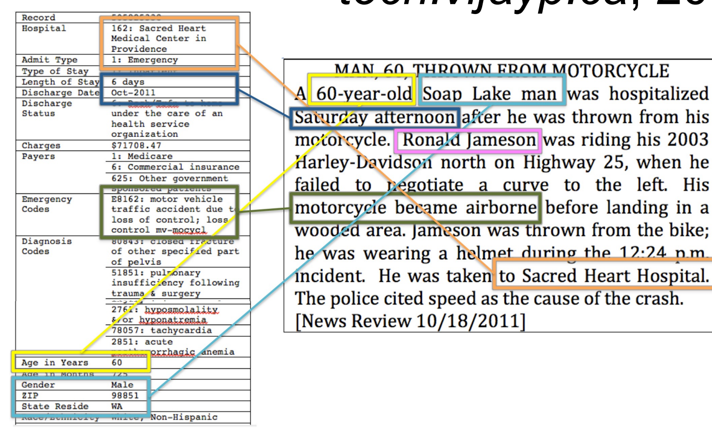
Privacy issues in data collection and learning

“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

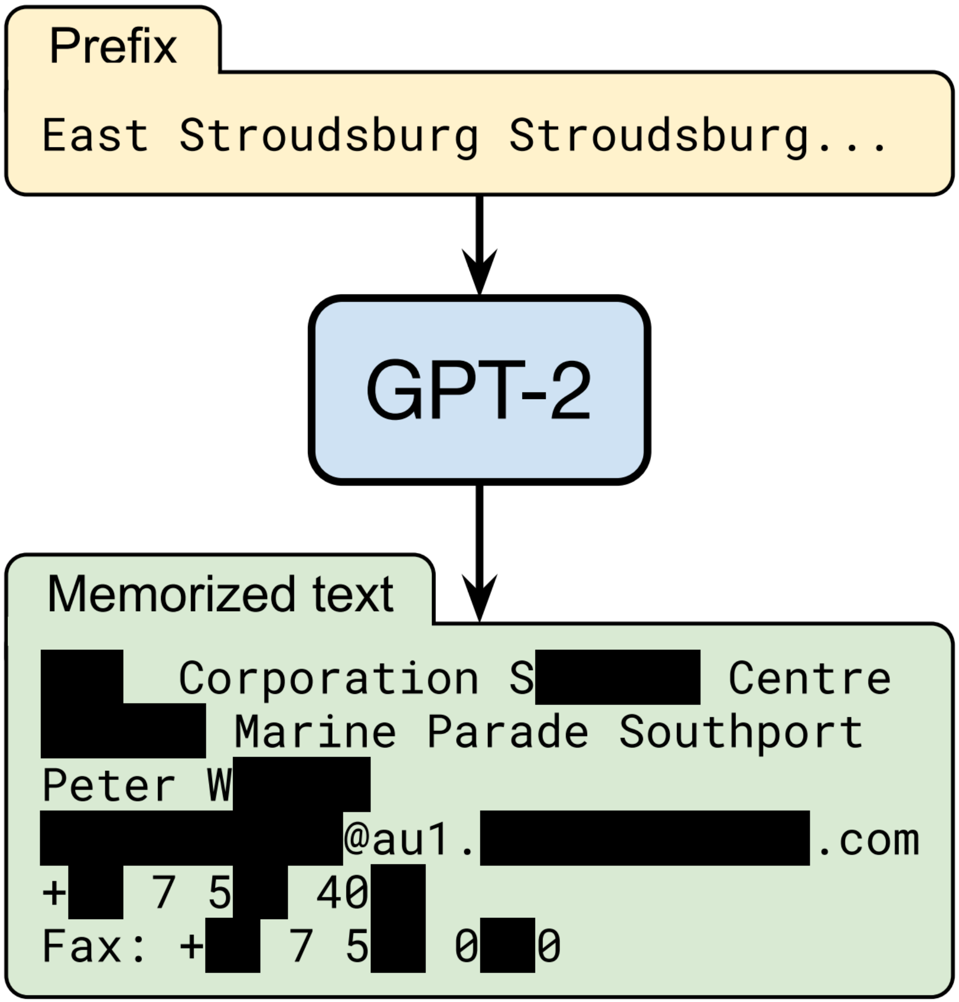
Dawn Song
University of California, Berkeley

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

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

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

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

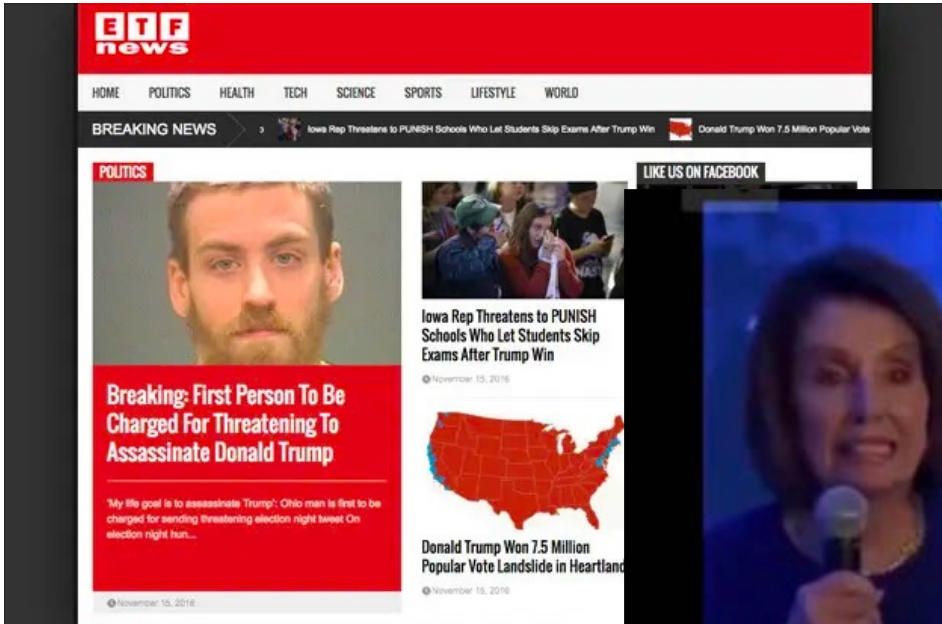
- I will have to delete all traces of a user upon request?

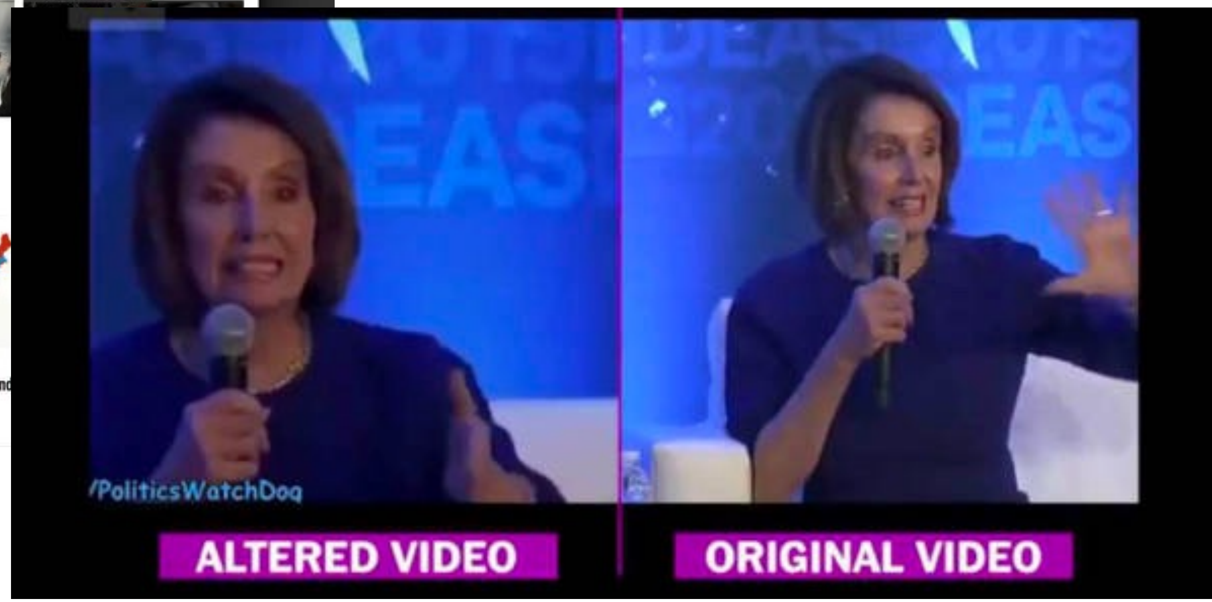
Recent/upcoming legislations on privacy forces companies to revise their data practice

- 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?

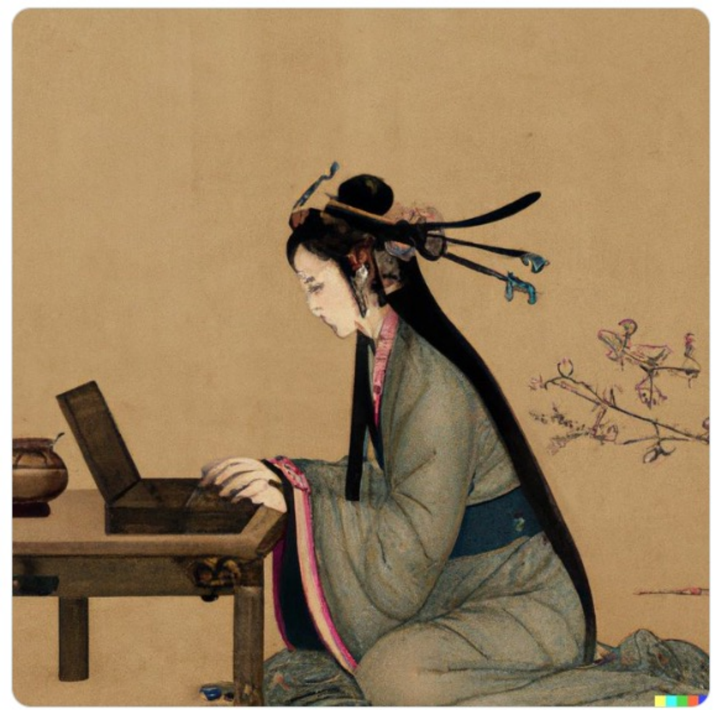
The rise of generative models

- 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/

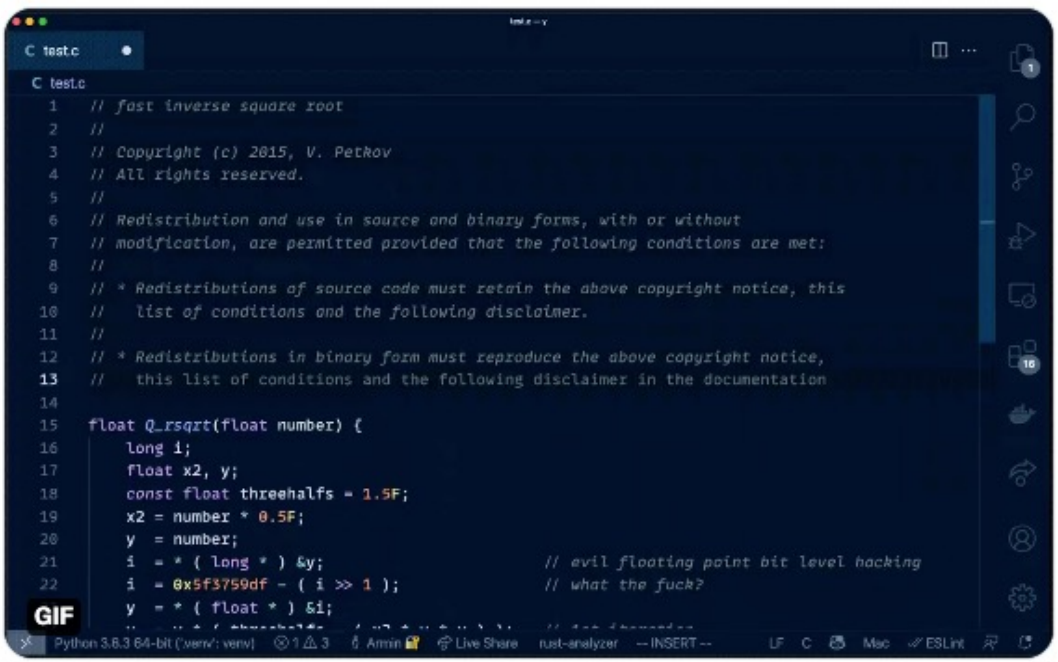
"Oriental painting of a lady programming on a laptop in the Song Dynasty" #Dalle⛩



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”)

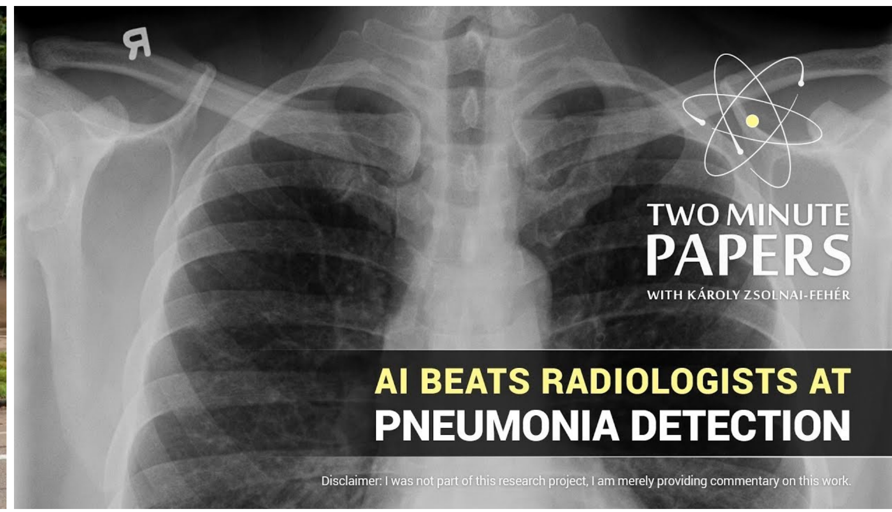


https://twitter.com/mitsuhiko/status/1410886329924194309

- Are the generated content considered plagiarism?

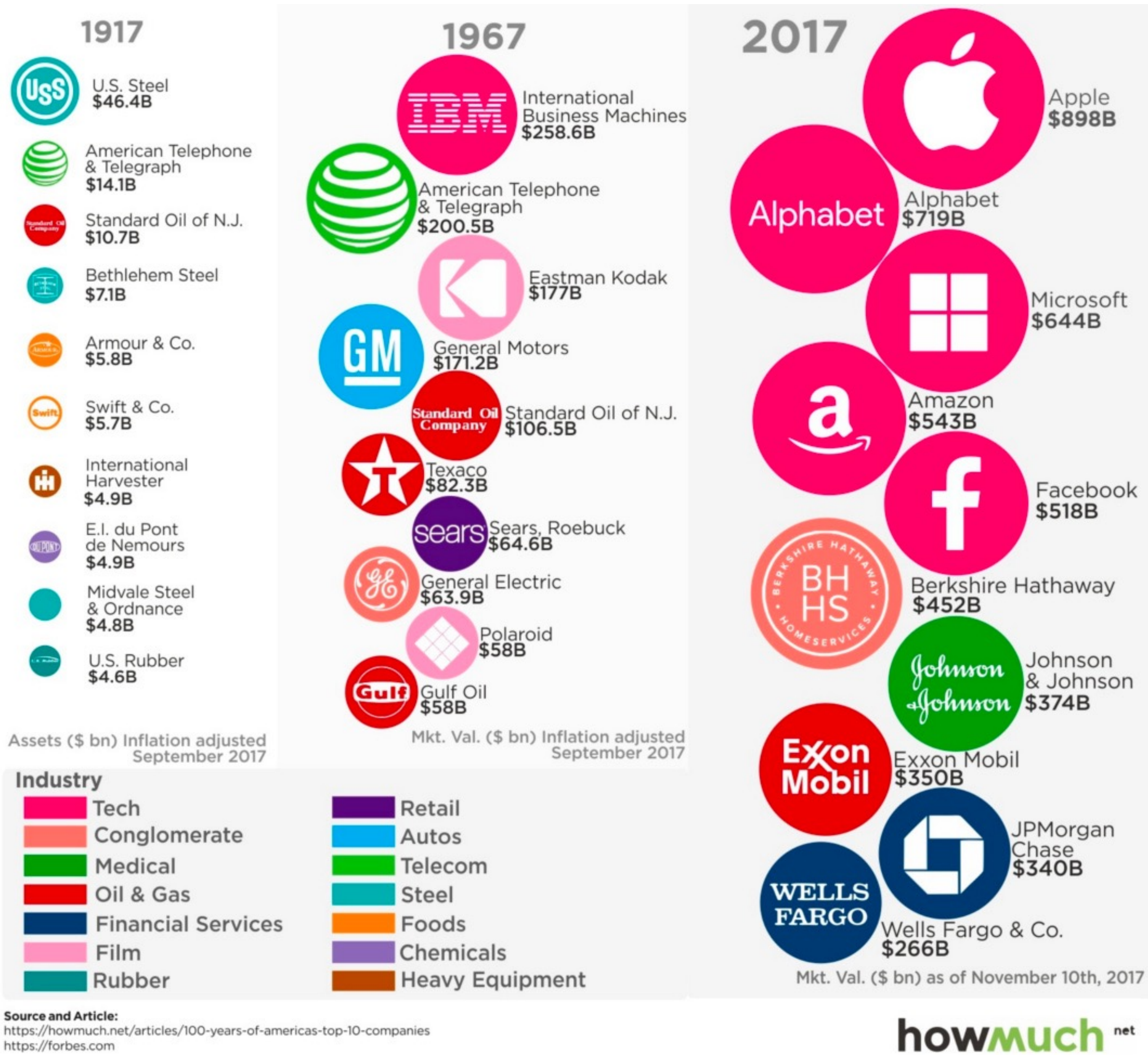
Societal impacts of new technology

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

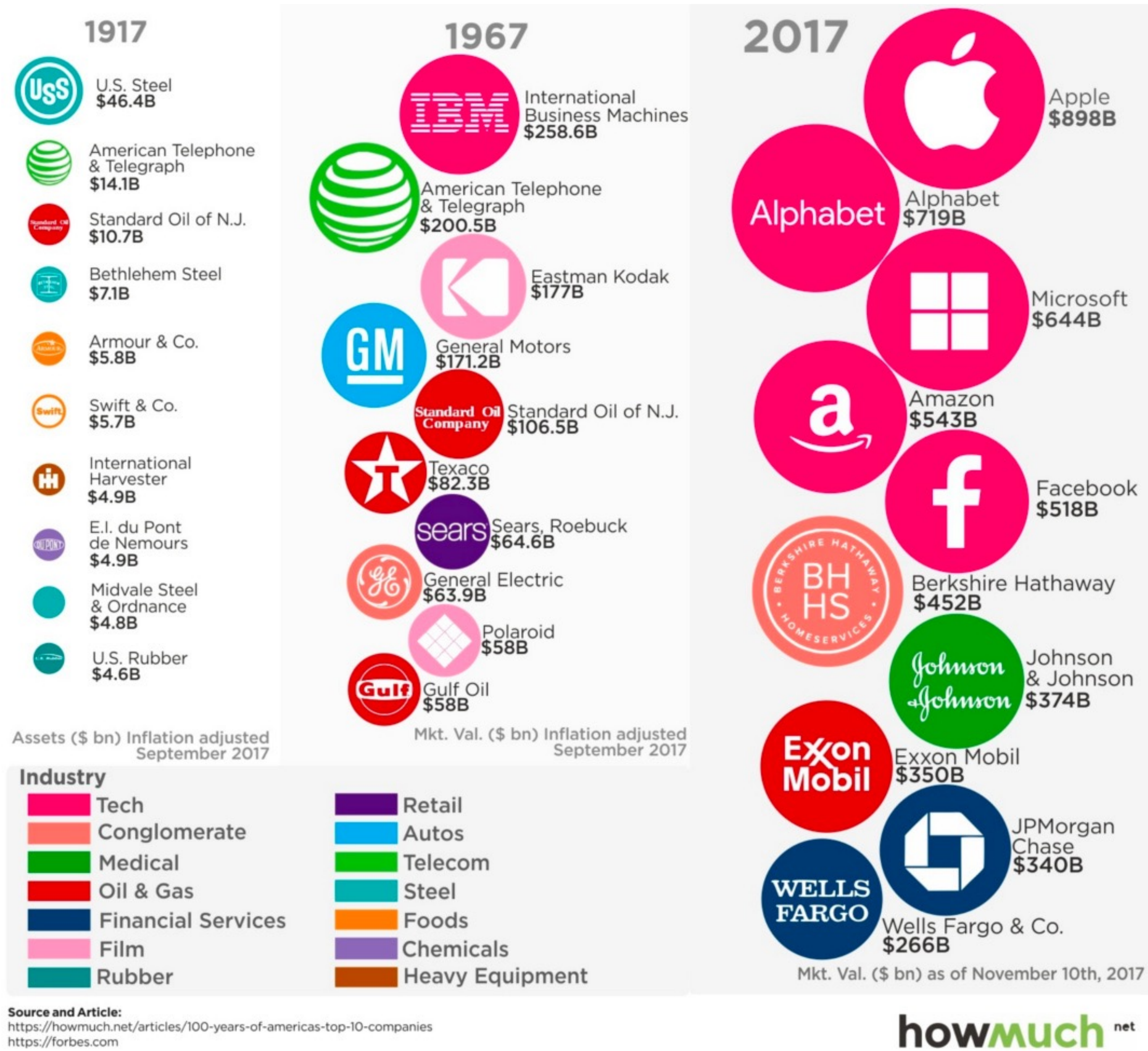


- AI is also creating new jobs, but…
  - Can your grandpa learn how to code?

Who are getting the largest piece of the technology pie?



Who are getting the largest piece of the technology pie?



2020:

Apple: 2.12T

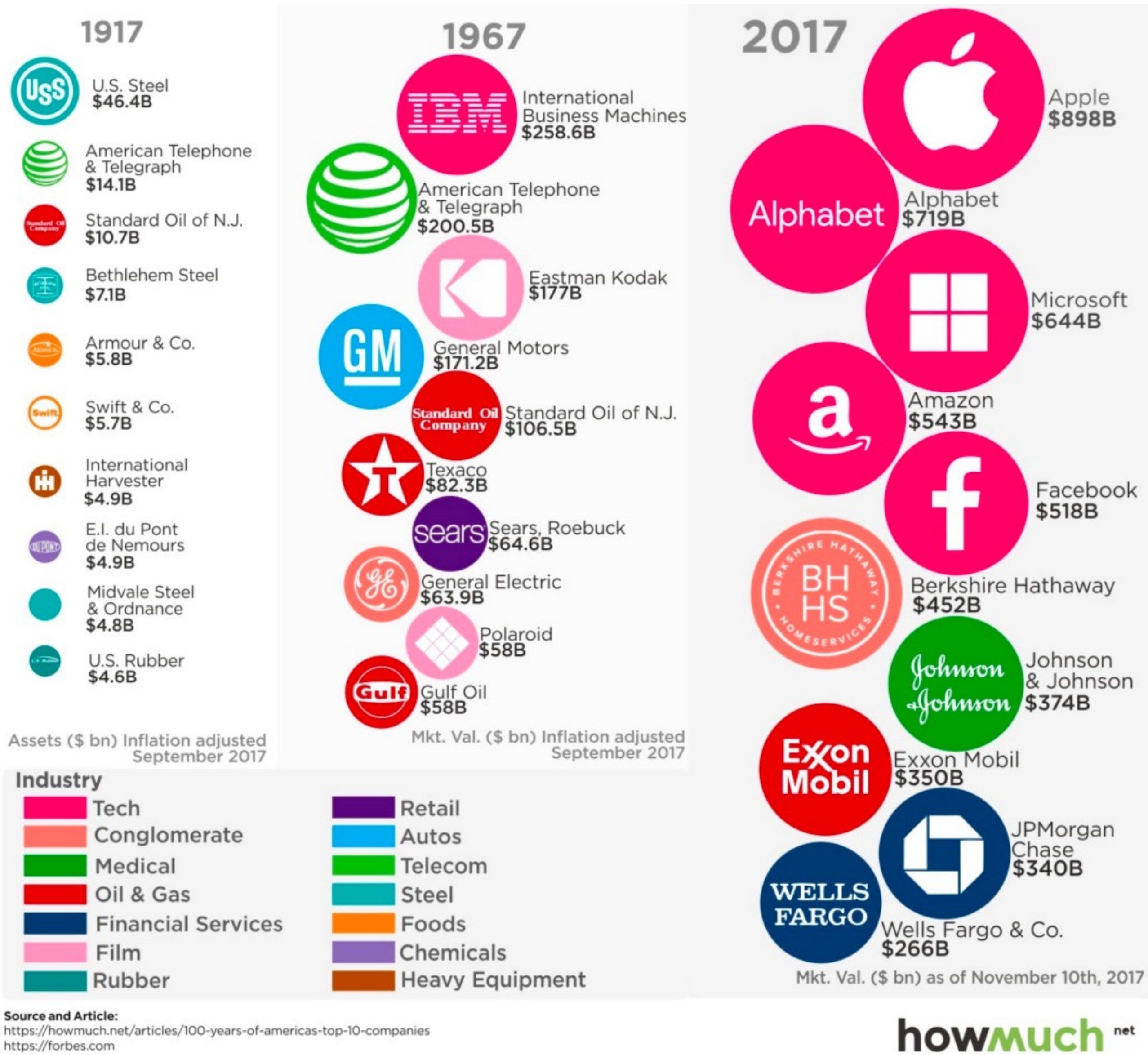
Amazon: 1.59T

Alphabet: 1.22 T

…

Tesla: 600 B +

Who are getting the largest piece of the technology pie?



2020:

Apple: 2.12T

Amazon: 1.59T

Alphabet: 1.22 T

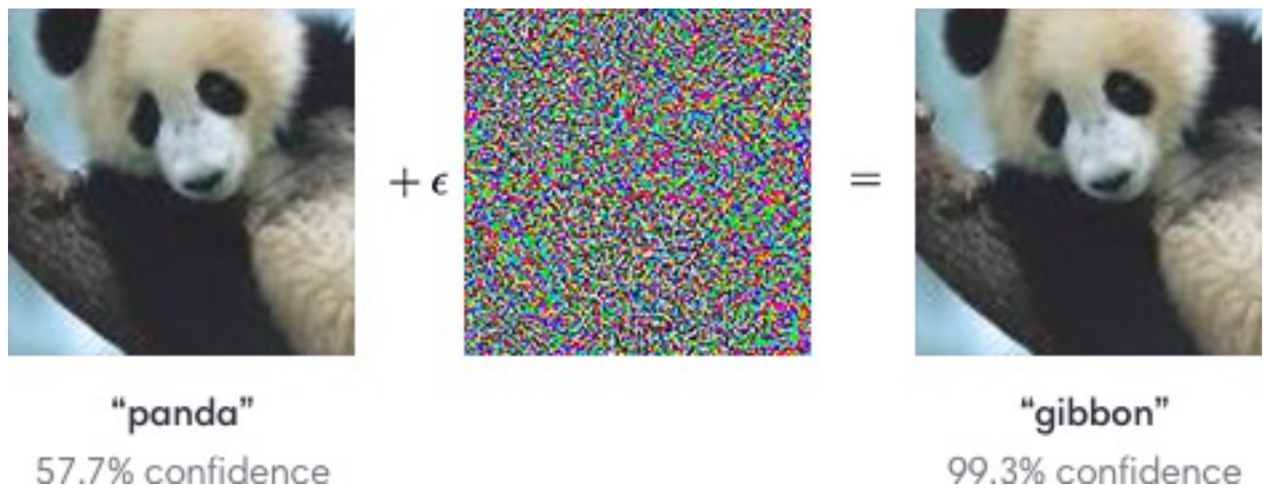
…

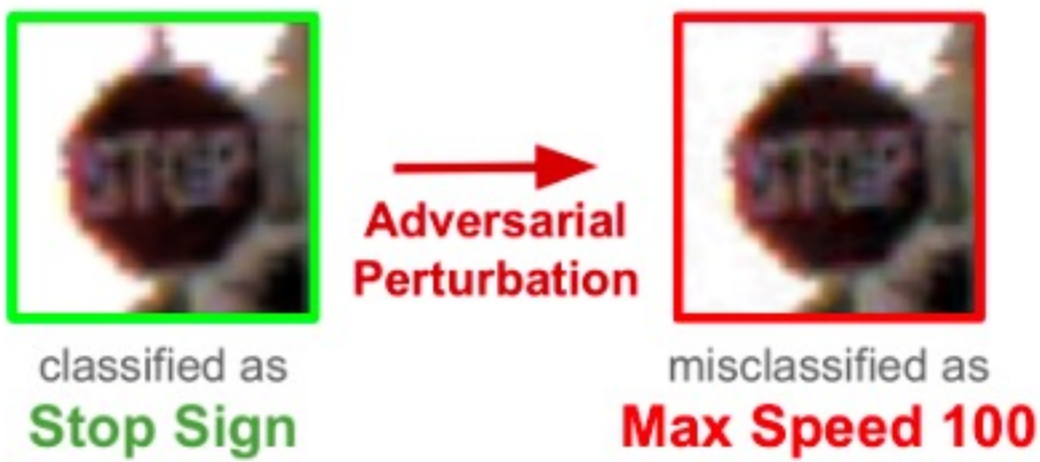
Tesla: 600 B +

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

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

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.

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

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.

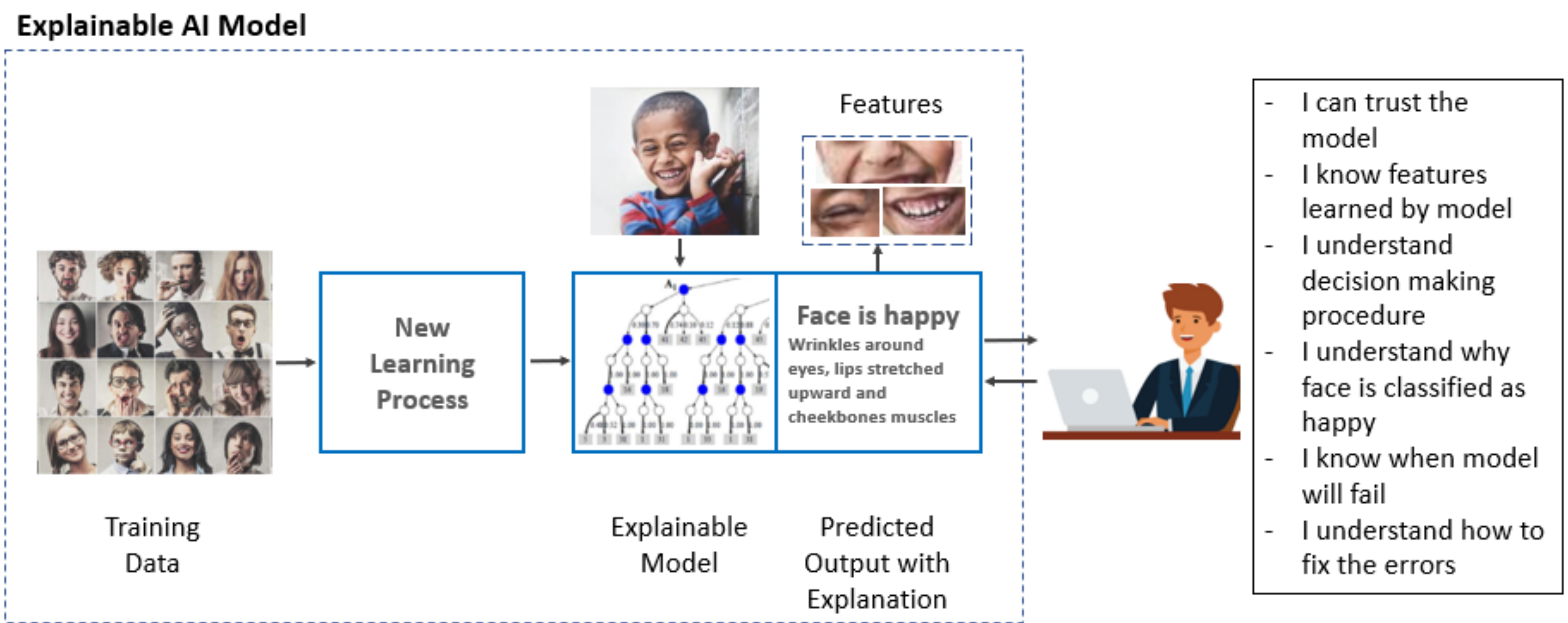
What is it that we want? How do we define fairness?

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

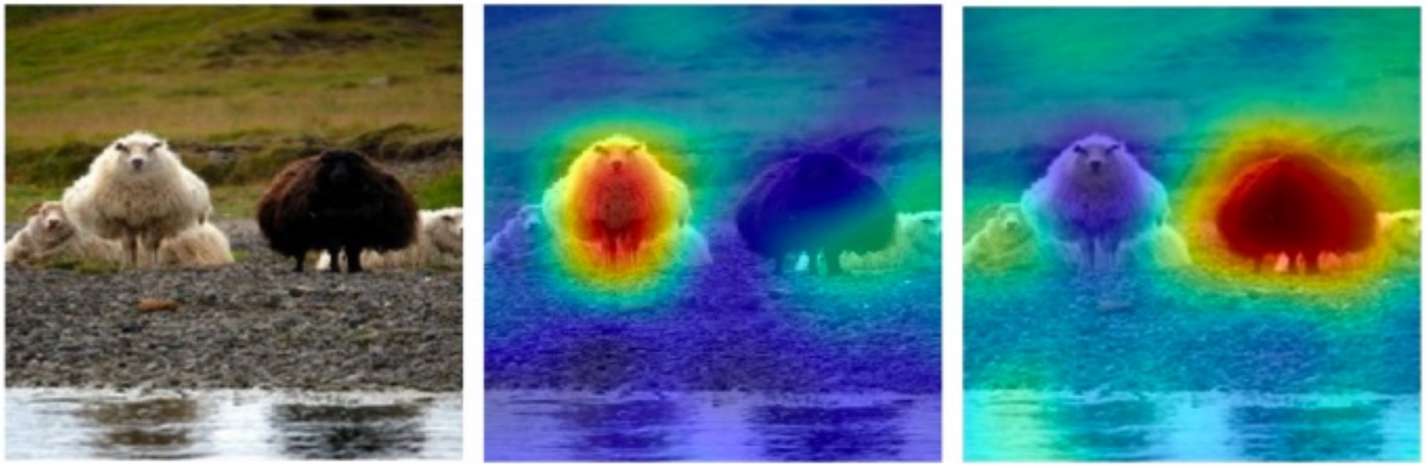
https://arxiv.org/pdf/1609.05807.pdf

Research in Responsible AI

- Explanability of AI predictions



Another example on explainable AI predictions

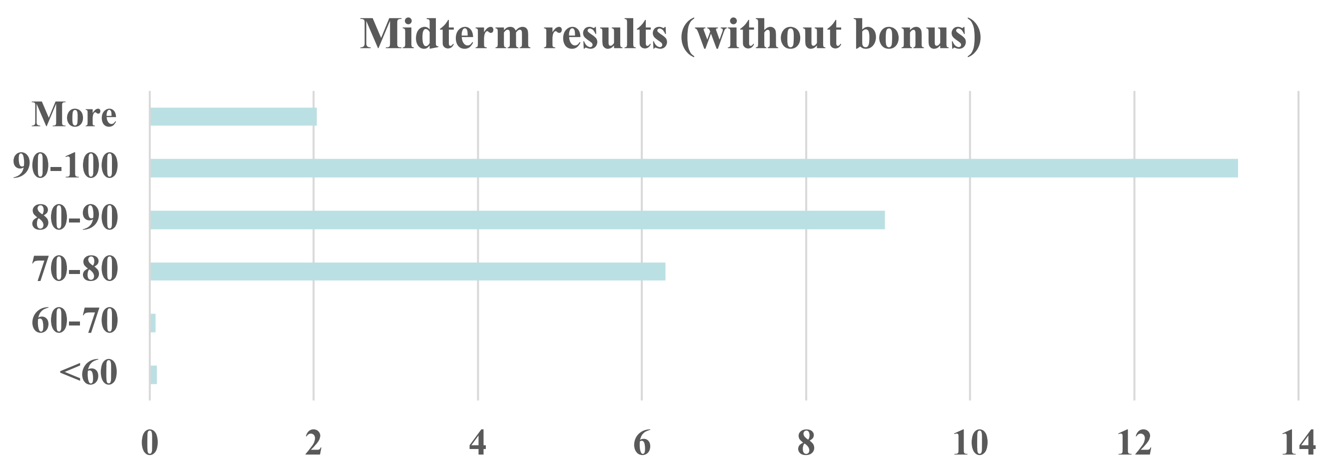


(a) Sheep - 26%, Cow - 17% (b) Importance map of '*sheep*' (c) Importance map of '*cow*'

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

Research in Responsible AI

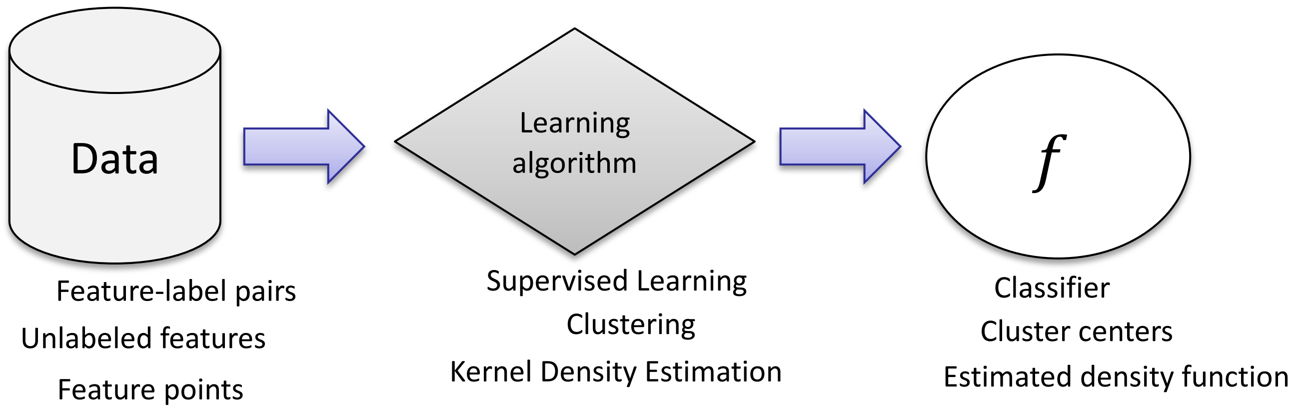
- Provable guarantees against identification in privacy



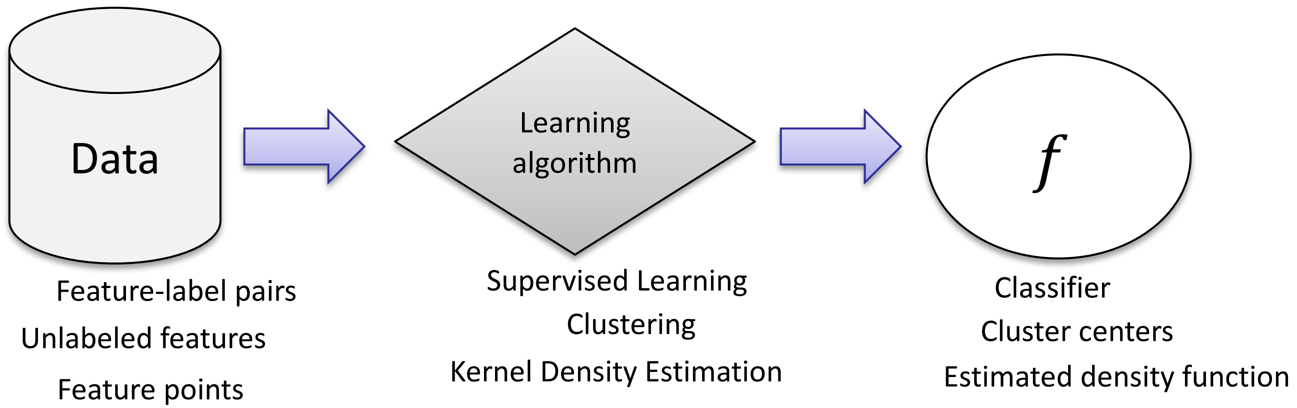
Differentially privately released midterm results from Fall 2020

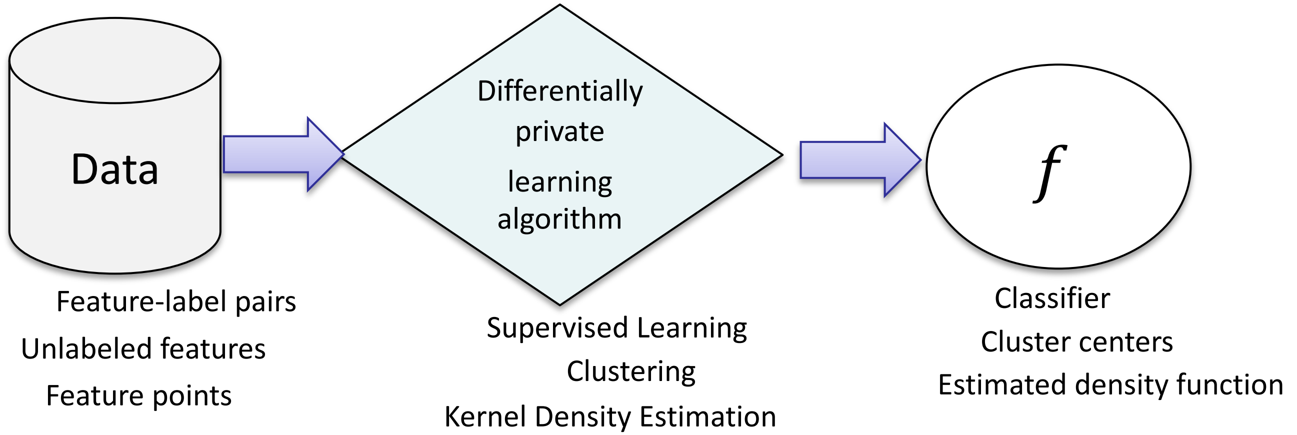
How does differential privacy work?

Differentially Private Machine Learning

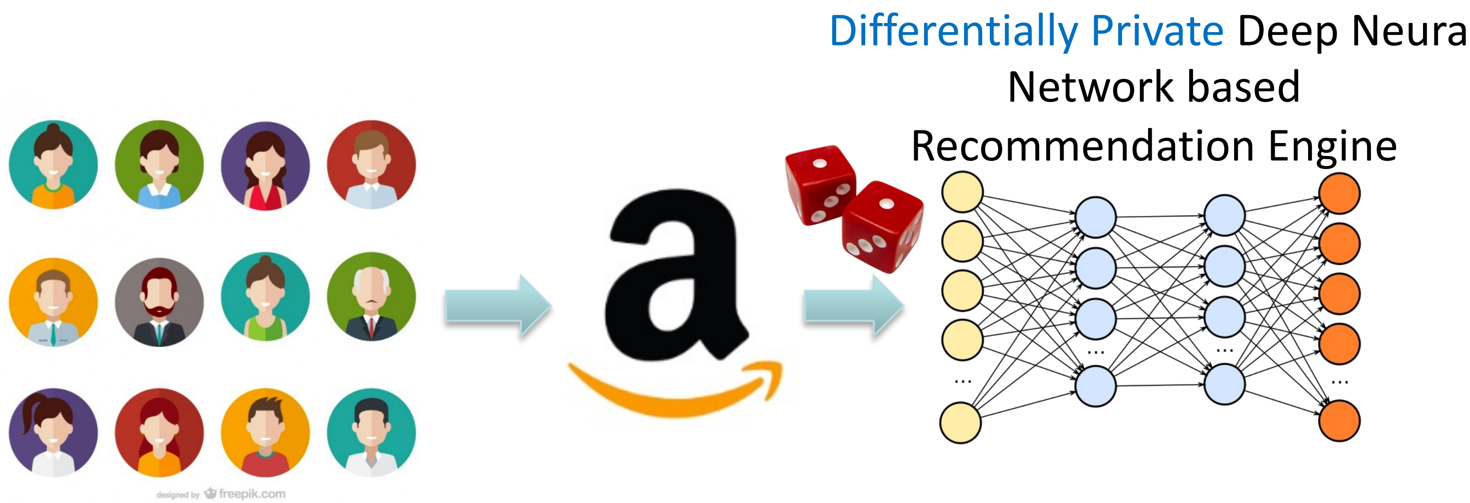


Differentially Private Machine Learning

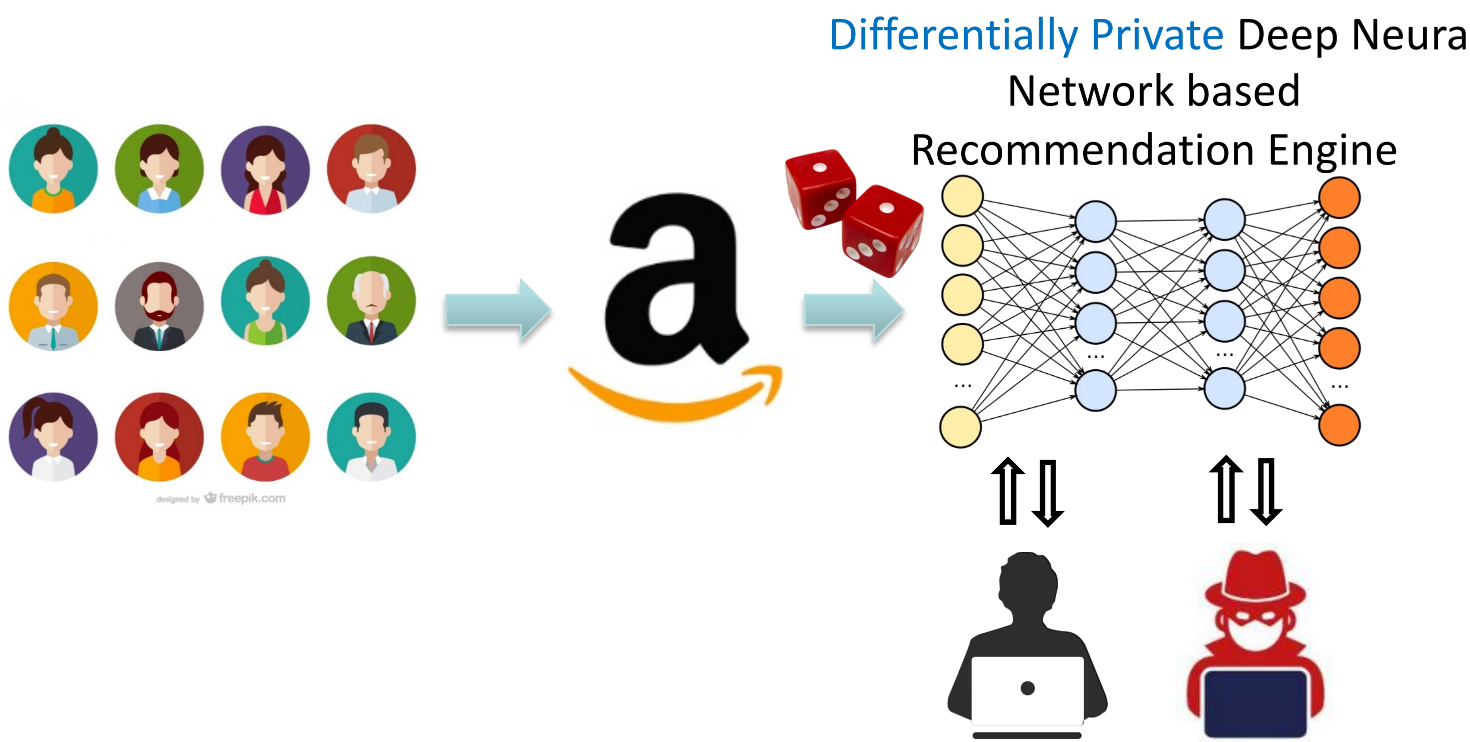




Example: Recommender System



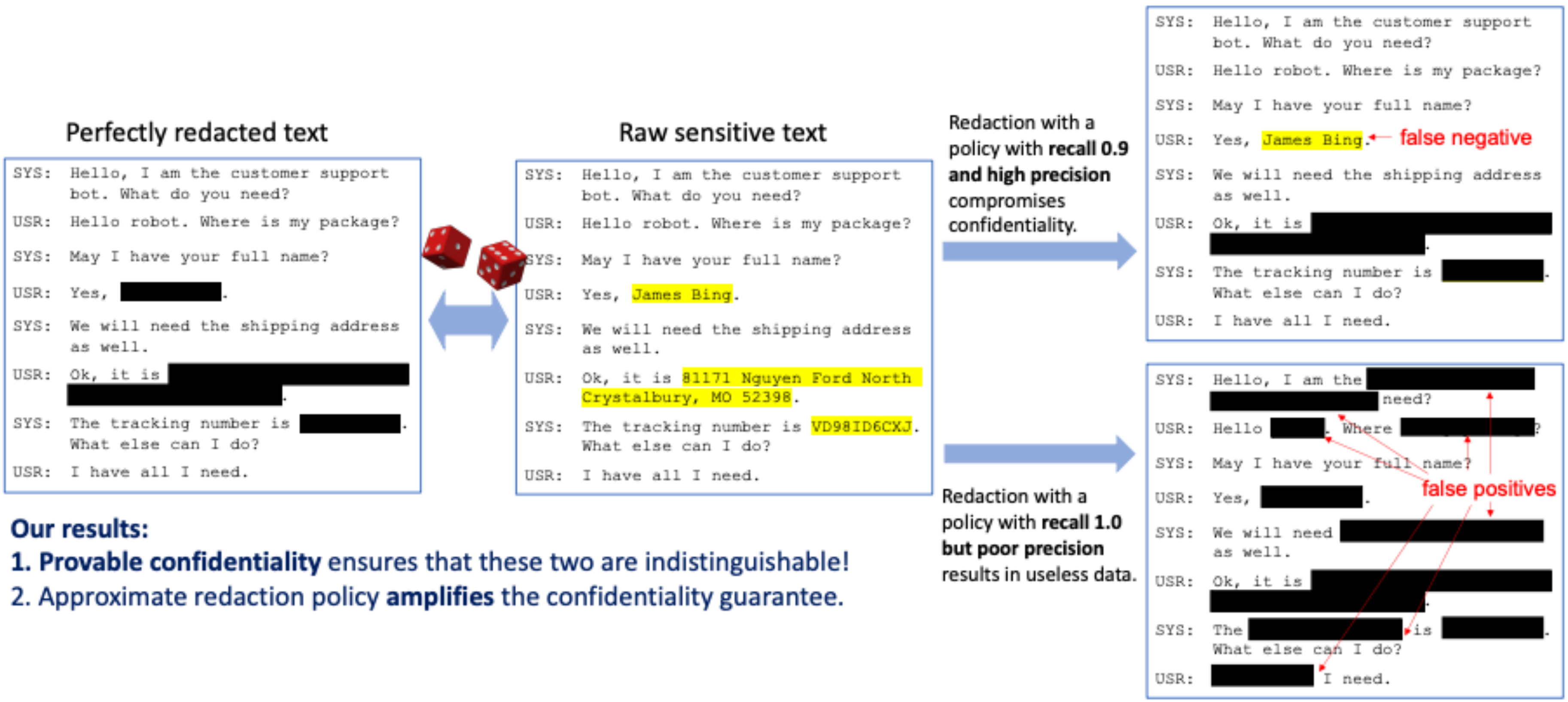
Example: Recommender System



“If your recommendation engine is private, then an adversary can’t infer whether a particular user was present”

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

UCSB Activities in Responsible AI

CRML Center for Responsible Machine Learning

CITS

UC SANTA BARBARA
Mind & Machine Intelligence

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