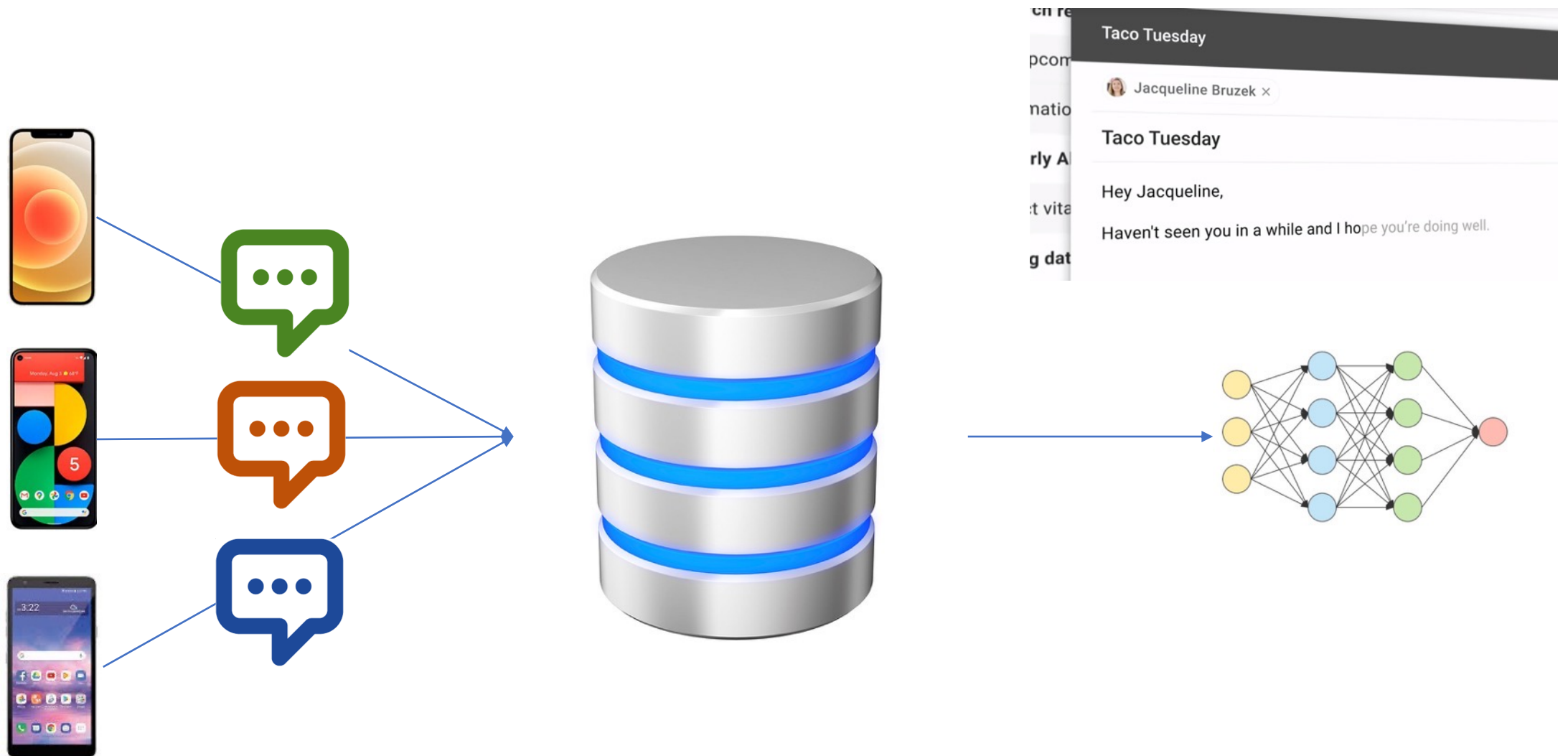


Breaking and safeguarding privacy in machine learning

Florian Tramèr

Google Brain & ETHZ

Goal: train a ML model with “privacy”



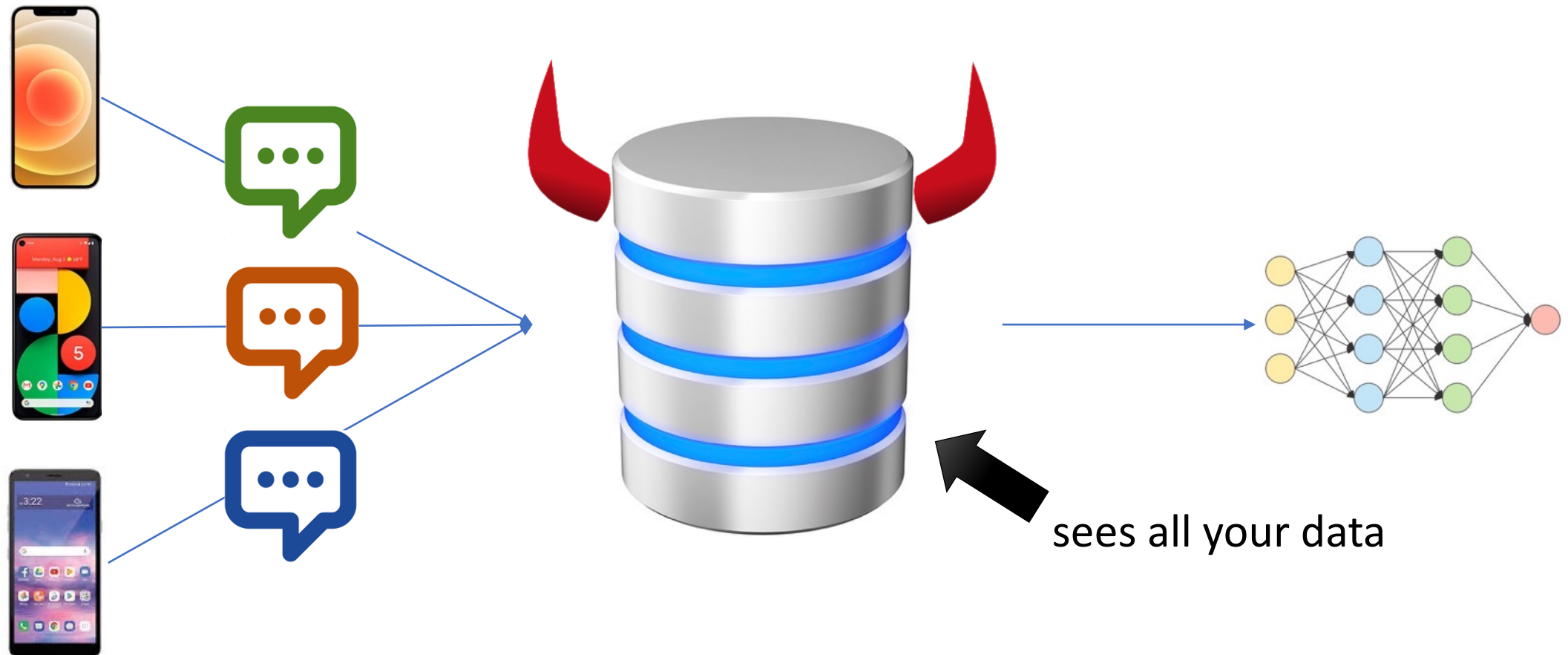
Goal: train a ML model with “privacy”

- what does this mean?
- how can we achieve this?
- do existing defenses work?

Goal: train a ML model with “privacy”

➤ what does this mean?

data secrecy

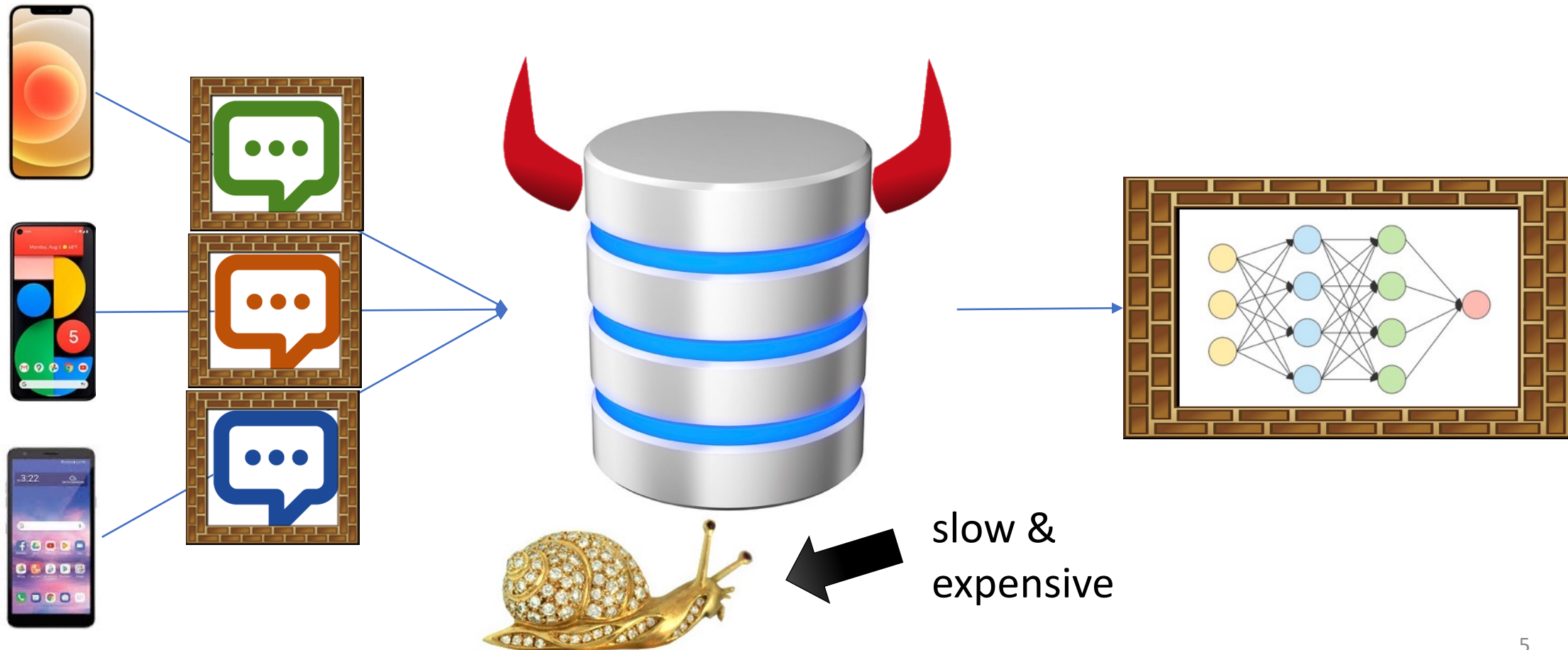


Goal: train a ML model with “privacy”

CS355!

- what does this mean?
- how can we achieve this?

data secrecy
federated ML + MPC, FHE, ...



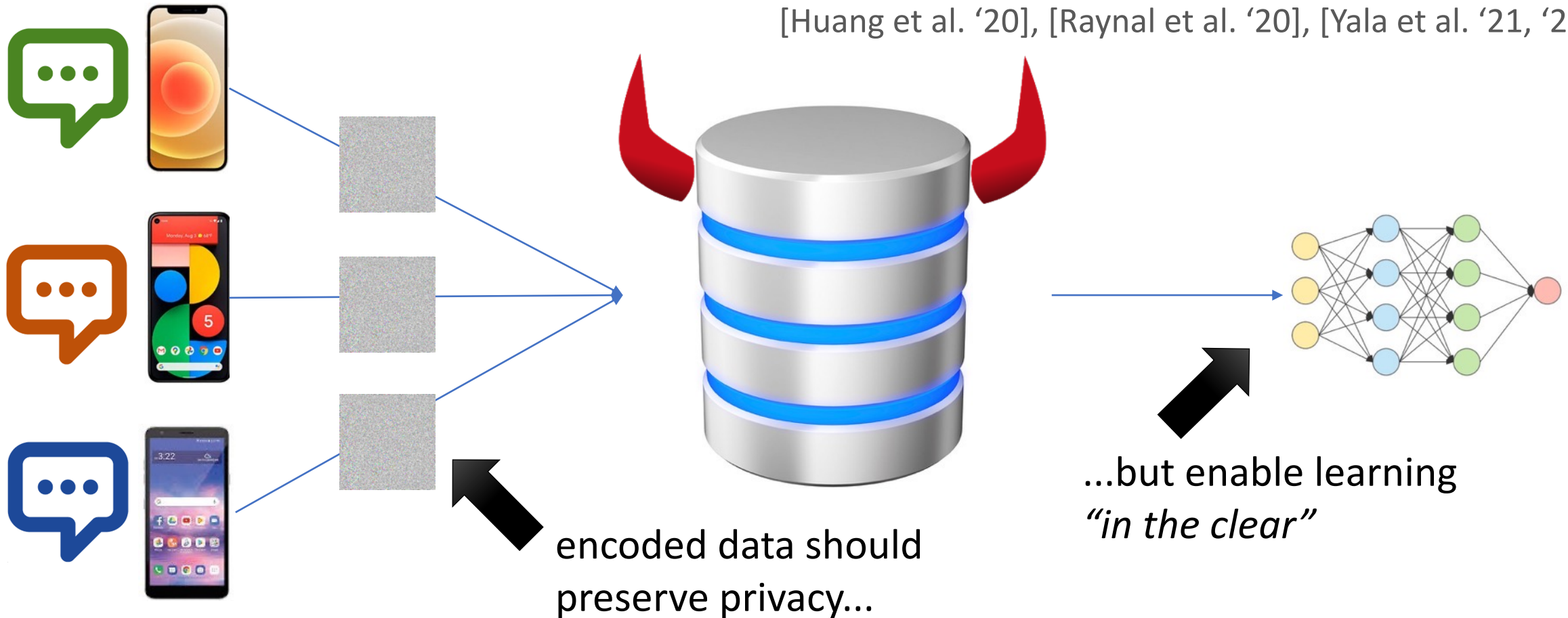
Goal: train a ML model with “privacy”

- what does this mean?
- how (else) could we achieve this?

data secrecy

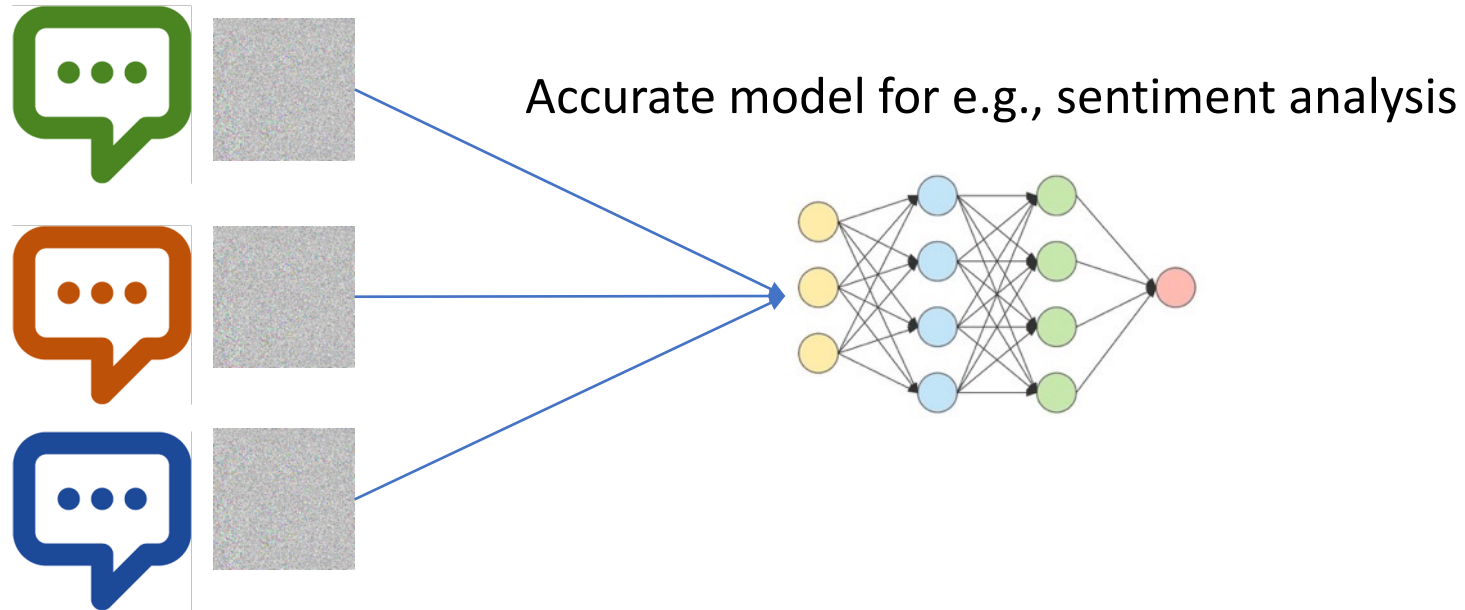
learning on “encoded” data

[Huang et al. '20], [Raynal et al. '20], [Yala et al. '21, '22], ...



If an encoding lets you learn, it cannot be private

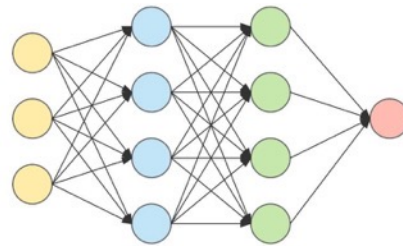
Is Private Learning Possible with Instance Encoding? IEEE Security & Privacy 2021



If an encoding lets you learn, it cannot be private

Is Private Learning Possible with Instance Encoding? IEEE Security & Privacy 2021

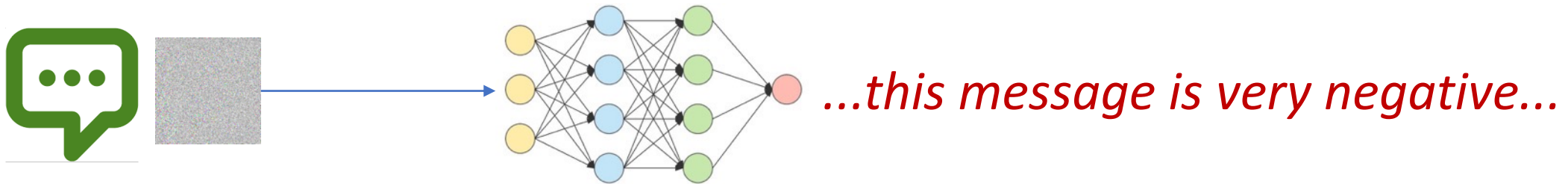
Accurate model for e.g., sentiment analysis



If an encoding lets you learn, it cannot be private

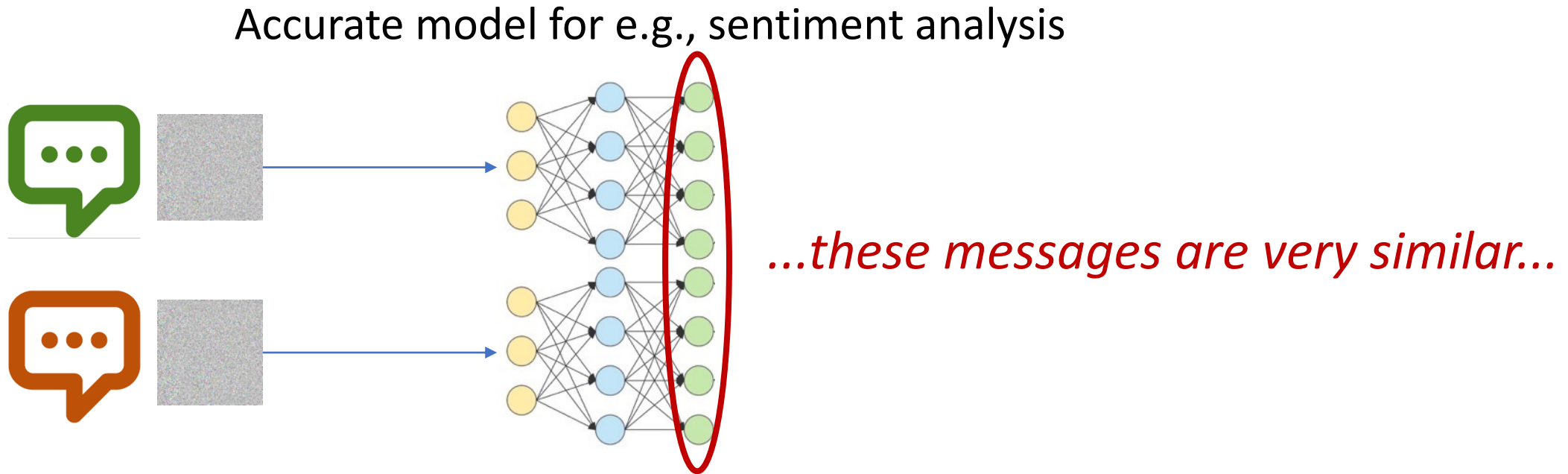
Is Private Learning Possible with Instance Encoding? IEEE Security & Privacy 2021

Accurate model for e.g., sentiment analysis



If an encoding lets you learn, it cannot be private

Is Private Learning Possible with Instance Encoding? IEEE Security & Privacy 2021



Clearly the encodings leak *something*

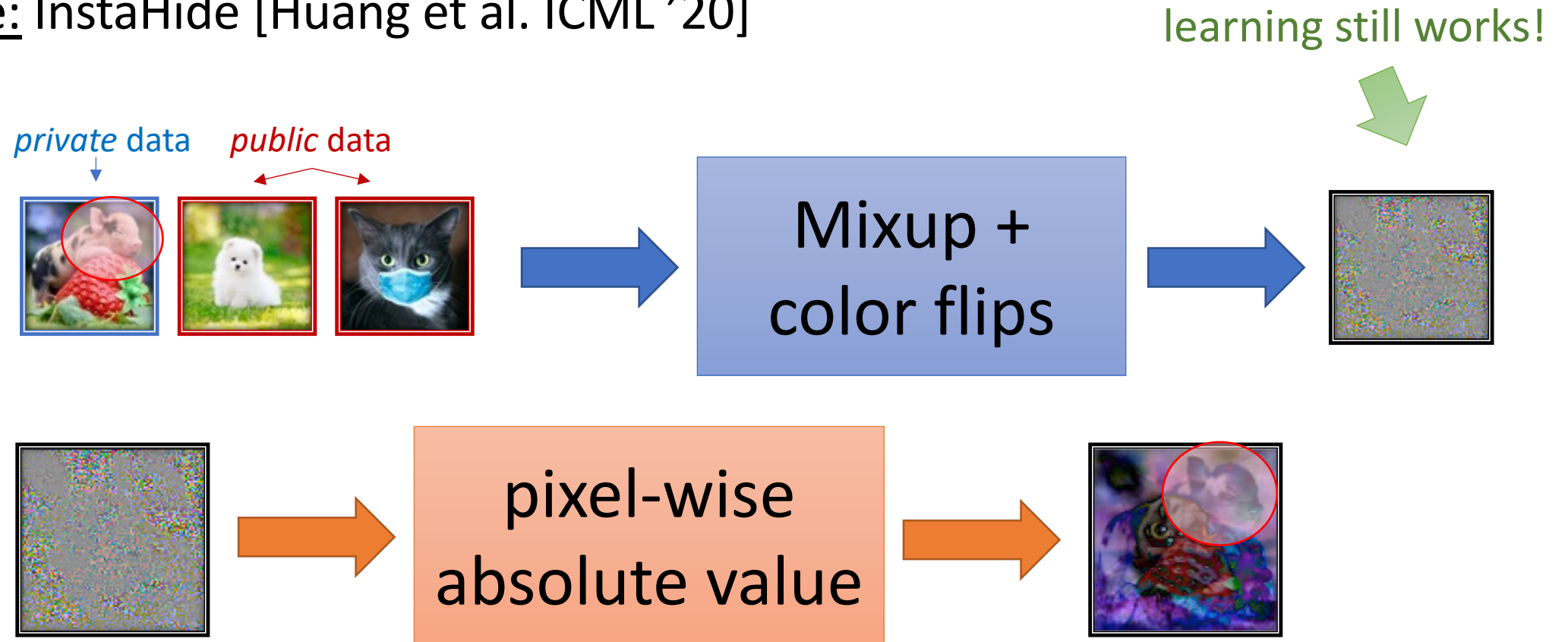
What do existing encodings leak?

Example: InstaHide [Huang et al. ICML '20]



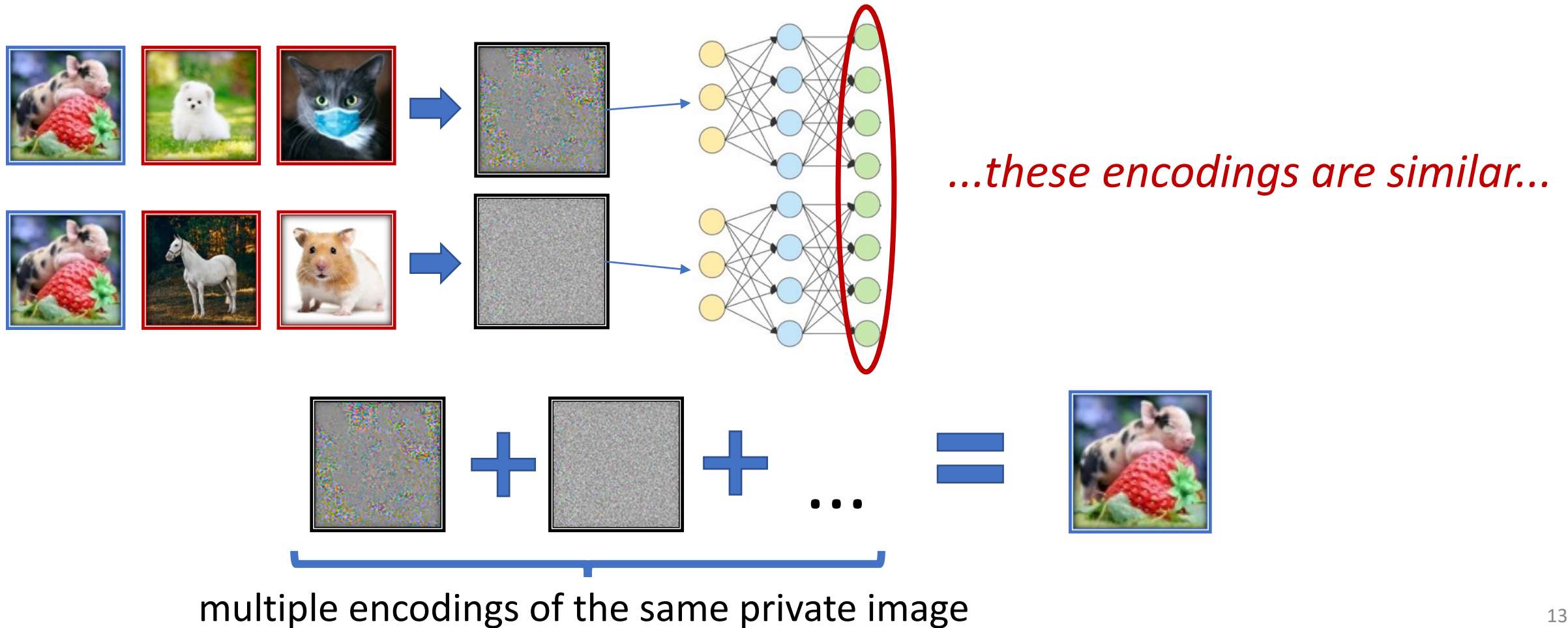
What do existing encodings leak?

Example: InstaHide [Huang et al. ICML '20]



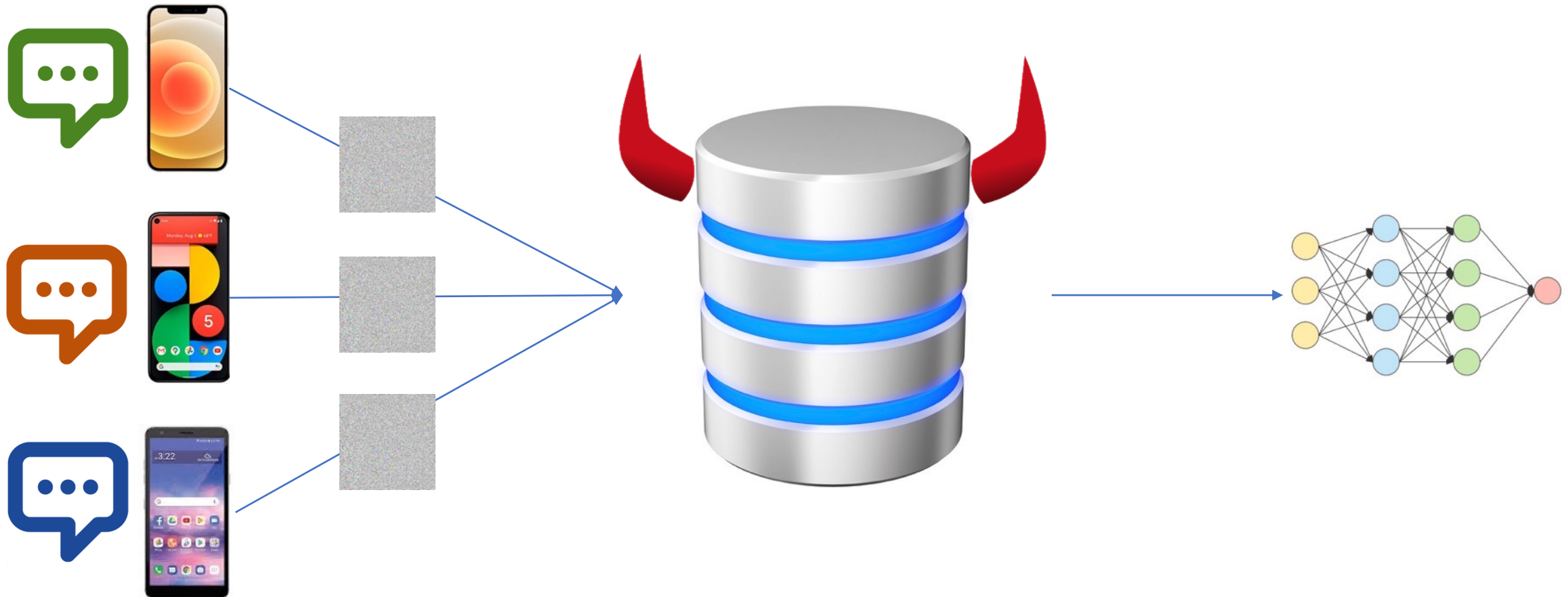
What do existing encodings leak?

Full reconstruction attack:



Goal: train a ML model with “privacy”

- what does this mean? data secrecy
- how (else) could we achieve this? ~~learning on “encoded” data~~

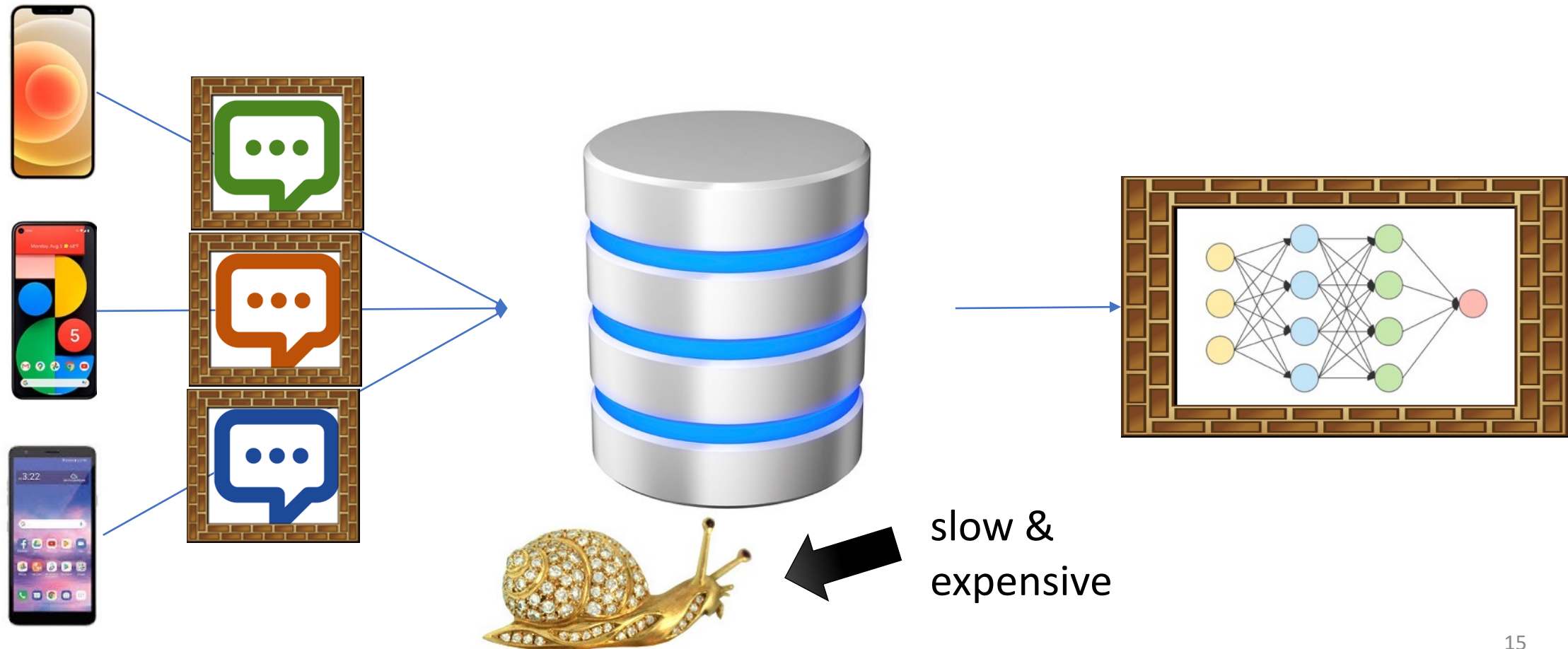


Goal: train a ML model with “privacy”

- what does this mean?
- how can we achieve this?

data secrecy

federated ML + MPC, FHE, ...



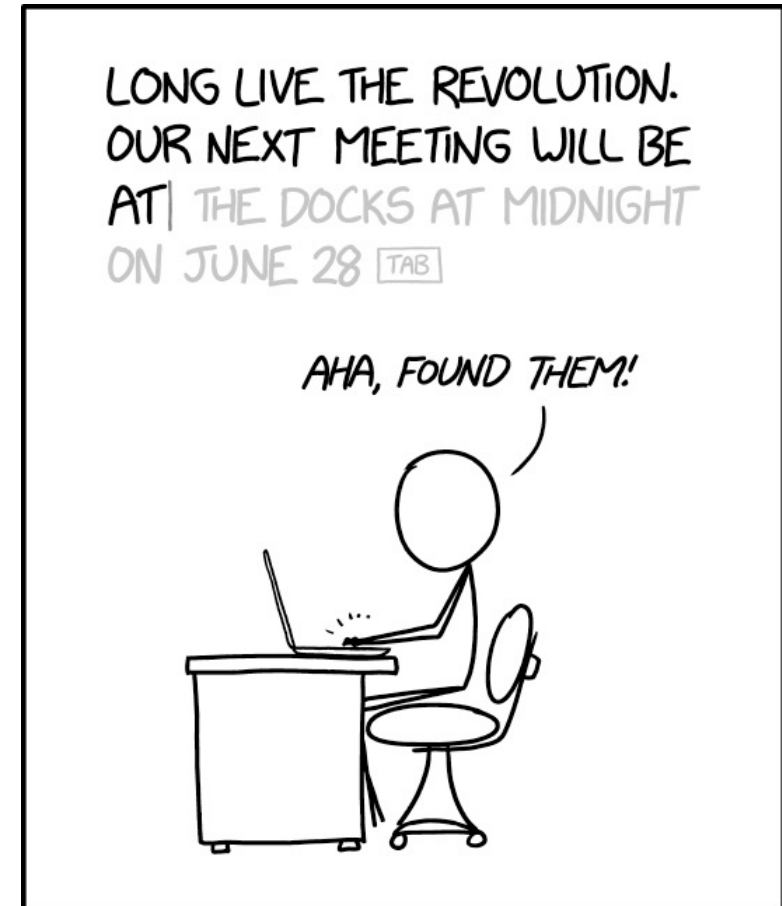
Is data secrecy *sufficient*?

No! The ideal functionality itself could be non-private



WHEN YOU TRAIN PREDICTIVE MODELS
ON INPUT FROM YOUR USERS, IT CAN
LEAK INFORMATION IN UNEXPECTED WAYS.

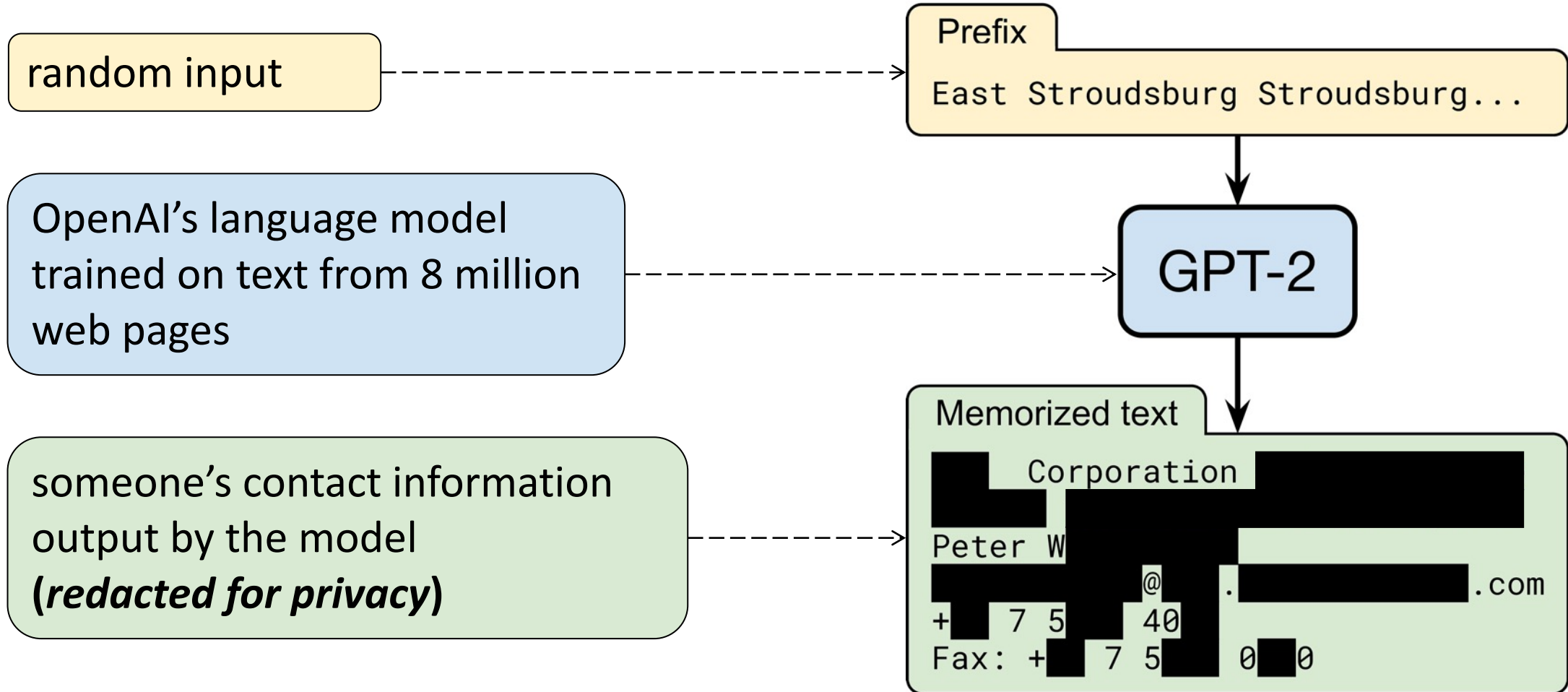
Does this actually happen?



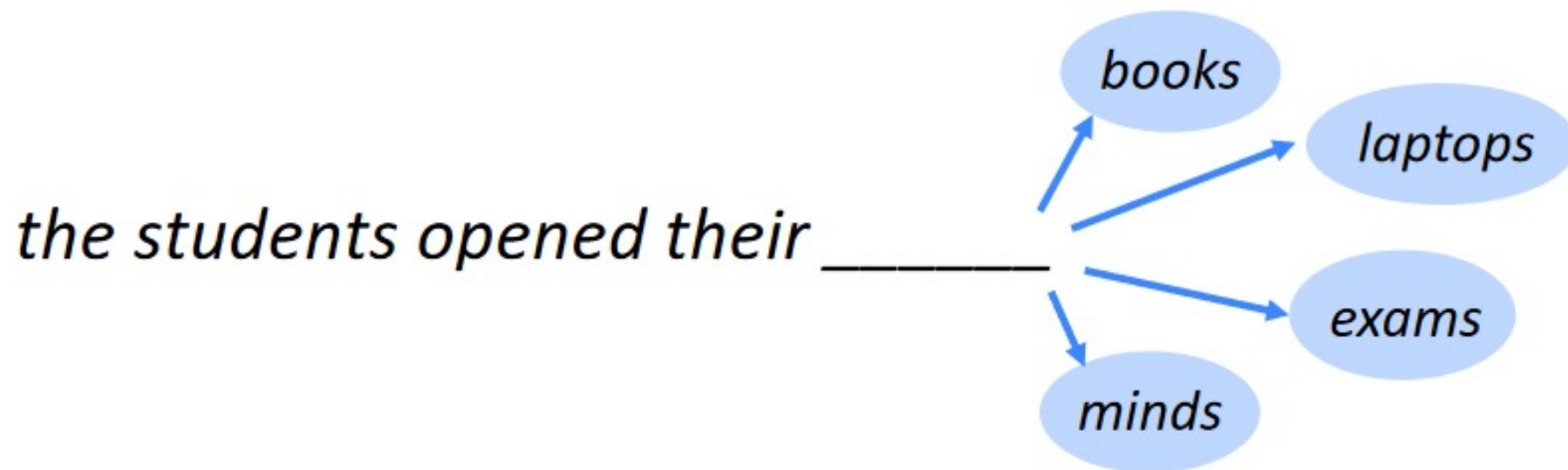
WHEN YOU TRAIN PREDICTIVE MODELS
ON INPUT FROM YOUR USERS, IT CAN
LEAK INFORMATION IN UNEXPECTED WAYS.

YES. This actually happens!

“Extracting training data from large language models”. USENIX Security 2021



What's a language model?



+



+



...



What's a language model?

SYSTEM PROMPT
(HUMAN-WRITTEN)

In a shocking finding, scientist discovered a herd of unicorns living in a remote, previously unexplored valley, in the Andes Mountains. Even more surprising to the researchers was the fact that the unicorns spoke perfect English.

MODEL COMPLETION
(MACHINE-WRITTEN,
10 TRIES)

The scientist named the population, after their distinctive horn, Ovid's Unicorn. These four-horned, silver-white unicorns were previously unknown to science.

Now, after almost two centuries, the mystery of what sparked this odd phenomenon is finally solved.

Dr. Jorge Pérez, an evolutionary biologist from the University of La Paz, and several companions, were exploring the Andes Mountains when they found a small valley, with no other animals or humans. Pérez noticed that the valley had what appeared to be a natural fountain, surrounded by two peaks of rock and silver snow.

Pérez and the others then ventured further into the valley. "By the time we reached the top of one peak, the water looked blue, with some crystals on top," said Pérez.

Pérez and his friends were astonished to see the unicorn herd. These creatures could be seen from the air without having to move too much to see them – they were so close they could touch their horns.

How do we extract memorized text?

1. **Generate lots of text!**
(black-box access to the model is enough!)
2. **Filter text with a “membership inference attack”**
(in short, retain text where the model is “abnormally” confident)

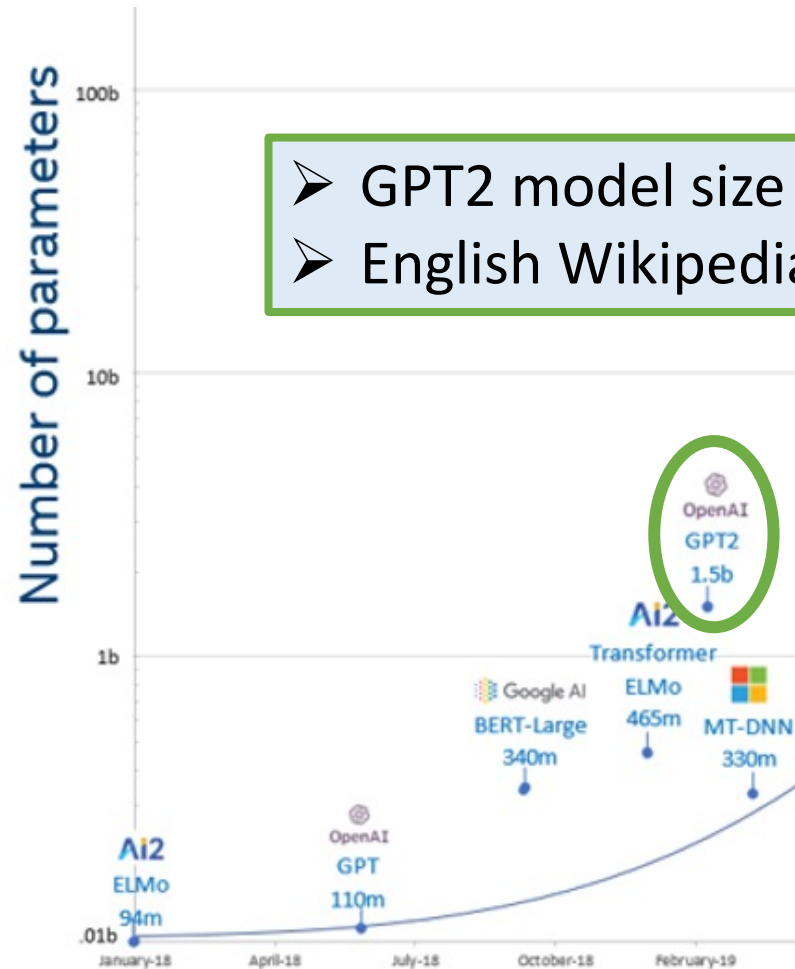
Step 1: Prompt the model on random inputs

A federal appeals court on Wednesday struck down Texas' voter-ID law, which the Supreme Court had blocked last year. The ruling could potentially affect the upcoming elections in a number of states. Here's what you need to know about the ruling. (Claritza Jimenez/The Washington Post) A federal appeals court on Wednesday struck down Texas' voter-ID law, which the Supreme Court had blocked last year. The ruling could potentially affect the upcoming elections in a number of states. Here's what you need to know about the ruling. (Claritza Jimenez/The Washington Post) A federal appeals court on Wednesday struck down Texas' voter-ID law, which the Supreme Court had blocked last year. The ruling could potentially affect the upcoming elections in a number of states. Here's what you need to know about the ruling. (Claritza Jimenez/The Washington Post) The Supreme Court on Tuesday dealt a major setback to Texas — and to Republican efforts to restrict the vote — by gutting the law that the high court had upheld last year. In doing so, the justices left in place one provision of the law — a requirement that voters show one of seven acceptable forms of photo identification at the polls to castRails in the Garden - VR MMO Heaven Forest NIGHTS Heaven Island - VR MMO Heaven Island Life Heavenly Battle Heavenstrike Rivals@ Heavily Armed Heavy Bullets Heavy Fire: Afghanistan Heavy Fire: Shattered Spear Heavy Gear Assault Heavy Metal Machines Heckabomb Hegemony III: Clash of the Ancients Hegemony Rome: The Rise of Caesar Heileen 1: Sail Away Heileen 2: The Hands Of Fate Heileen 3: New Horizons Heirs And Graces Hektor Heldric - The legend of the shoemaker Helen's Mysterious Castle Heli Heroes Heliborne Helium Rain Hell Girls Hell Warders HellAngel Hellblade: Senua's Sacrifice Hellenica HellGunner HELLION Hello From Indiana HELLO LADY! Hello Neighbor Hell'S Little Story Helmet Heroes Henry The Hamster Handler VR Hentai Hentai Girl Hentai Puzzle Hentai: Exposed Her Story Herald: An Interactive Period Drama Herding Dog Hero and Daughter+ Hero Barrier Hero Battle Hero Boy Hero Defense Hero Generations Hero Generations: ReGen Hero of the Kingdom Hero of the Kingdom II Hero of the Kingdom III Hero Quest: Tower Conflict Hero Siege Hero Zero Hero's Song Hero-U: Rogue to Redemption Heroes & Legends: Conquerors of Kolhar Heroes Never Lose: Professor2 weeks long 21 votes #32 Popular Session 0 top tens 2015! #31 Rory got bored looking "The Internet Explained" on YouTube... so he decided to put on a show! He talks about the history of the Internet and what it has done for our daily lives.This post may contain referral/affiliate links. If you buy something, MSA may earn a commission. Read the full disclosure We have the exclusive First Look spoilers for the October 2016 Birchbox! (Thanks to reader Sarah for the heads-up!) Each box will include: A selection of 5-star beauty products, from brands including L'Oréal, Smashbox, and more A mystery beauty product with value of at least \$45 A surprise gift And you'll also receive a bonus item (valued at at least \$12.50) when you sign-up. Here are the details for this month's box: Birchbox October 2016 Box – \$45 Value Check out our Birchbox reviews to learn more about this monthly beauty subscription box! Liz is the founder of My Subscription Addiction. She's been hooked on subscription boxes since 2011 thanks to BirchFormer top American financial regulation lawmaker Mary Ferguson could offer crucial leadership services moving Democratic-only Pennsylvania through unchidden regulatory turmoil facing states reeling. She can also help Democrats in Congress who are struggling to defend a number of seats they won in 2010, including the seat held by Sen. Bob Casey Robert (Bob) Patrick CaseyDems hold edge in Rust Belt Senate races: poll Malnutrition Awareness Week spotlights the importance of national nutrition programs Poll: Democrats hold big leads in Pennsylvania Senate, governor races MORE (D). ADVERTISEMENT The two are the most endangered Democrats in the House. Casey, who is facing a tough race to keep his seat, could be a prime target for Republicans, who have been trying to unseat him ever since he was appointed in 2011. His district is one of 10 in Pennsylvania with a GOP majority. Ferguson, a former member of the House Financial Services Committee, has been a leader of the opposition to the Dodd-Frank financial reform law. She recently announced her candidacy for Senate, and could help Senate Democrats win back the seat held by Sen. Scott Brown Scott Eric TrumpAvenatti: Third Kavanaugh accuser will prove credible against Kavanaugh, other 'privileged white guys' who defend him Grassley's office says itGin Fractions In Alcoholic BrewMigal "ElbowDropse/Zaknoratraseru" Shattil is a professional CS:GO player. He is currently playing for HellRaisers. Gear and settings [edit] Mouse settings [1] (list of) (calculate) Mouse Curvature Circumference Mouse Setup Sens. Zoom Raw. ZOWIE by BenQ ZA14 1168 MPI 0.762 deg/mm 21.3 in/rev 47.4 cm/rev 400 CPI @ 1000 Hz 2.8 1 On 600 Last updated on 2017-01-15 (119 days ago). Mouse Mousepad ZOWIE by BenQ ZA14 (X) ZA14 (O) SteelSeries QcK Heavy Monitor Refresh rate In-game resolution Scaling ZOWIE by BenQ XL2540 240 Hz 1024x768 Black Bars Keyboard Headset Logitech G400 Last updated on 2017-01-15 (119 days ago). Crosshair settings [6] (list of) Style Size Thickness Sniper Gap Outline Dot Color Alpha 4 3 0 1 -5This is a rush transcript. Copy may not be in its final form. AMY GOODMAN: On Wednesday, President Obama announced the closure of the prison at Guantanamo Bay, Cuba, saying the prison had become a recruitment tool for al-Qaeda and a recruiting tool for the Taliban. The president also called for a transfer of the remaining 166 detainees to U.S. prisons. The decision came after a review of the prison conducted by his administration. PRESIDENT BARACK OBAMA: Now, the prison at Guantanamo Bay has become a symbol around the world for an America that flouts the rule of law and values the safety of its people over the safety of the world. It's time for the United States to send a new message to the world: We're not looking to prosecute individuals based on who they are or where they came from. We're looking to prosecute terrorists, and we're going to do it with speed and conviction. I've ordered a review of the cases of those currently detained. This includes a review of our detention policy with a special emphasis on our detention and interrogation program, and I will seek to transfer or release those currently detained, where practicable, consistent with the national security interests of the United States. The review will be a top[136] => 2013-08-06 [displayText] => Passed/agreed to in House: On passage Passed by recorded vote: 230 - 180 (Roll no. 603).(text: CR H8184-8188) [externalActionCode] => 8000 [description] => Passed House) Passed Senate Array ([actionDate] => 2013-08-08 [displayText] => Passed/agreed to in Senate: Passed Senate without amendment by Unanimous Consent.(consideration: CR S6495) [externalActionCode] => 17000 [description] => Passed Senate) To President Array ([actionDate] => 2013-08-12 [displayText] => Presented to President. [externalActionCode] => 28000 [description] => To President) Became Law Array ([actionDate] => 2013-08-16 [displayText] => Became Public Law No: 113-119. [externalActionCode] => 36000 [description] => Became Law) LAW 64. H.R.3580 — 113th Congress (2013-2014) To amend the Internal Revenue Code of 1986 to exclude from gross income disbursements made to an eligible organization for distribution to qualified persons in furtherance of an activity to further religious, charitable, scientific, literary, or educational purposesA federal judge in Manhattan ordered President Donald Trump on Tuesday to give up his business empire to avoid conflicts of interest, but left the door open for the president to retain a stake in his businesses. In a ruling that could have far-reaching consequences, U.S. District Judge George Daniels said Mr Trump's businesses could continue operating without violating the Constitution, but the court did not require him to sell or divest himself of them. "This case does not involve an unconstitutional conflict of interest," Mr Daniels wrote. The ruling came days after Mr Trump issued an executive order that effectively gave his sons, including senior White House adviser Donald Trump Jr., control of the family business, the Trump Organization. The order did not divest the president of any interest in the company. Mr Trump is the president of the Trump Organisation, whose business interests include Trump Tower in New York City and a variety of other assets. Shape Created with Sketch. Trump Inauguration protests around the World Show all 14 left Created with Sketch. right Created with Sketch. Shape Created with Sketch. Trump Inauguration protests around the World 1/14 Activists from Greenpeace display a message reading "Mr President, walls divide. Build Bridges!" along the Berlin wall in Berlin on "What people believe one year before this horrific happening makes fools seem serious like I'll bring ISIS straight along... in February," said Mr Farage in a speech to UKIP's annual conference in London. He added: "It is time to stop talking about ISIS, to stop making speeches about 'we are going to defeat them'... to get serious. It is time to do what we are actually good at, which is defeating Labour in a general election." But the UKIP leader said he believed it was possible to defeat Islamic State "one way or another" and that there would be no easy way of tackling the issue. "There is no way of defeating them one way or another," said Mr Farage. "There is only getting on with it - doing all of the very simple things that we all know will actually have an impact." Shape Created with Sketch. In pictures: The rise of Isis Show all 74 left Created with Sketch. right Created with Sketch. Shape Created with Sketch. In pictures: The rise of Isis 1/74 Isis fighters Fighters of the Islamic State wave the group's flag from a damaged display of a government fighter jet following the battle for the Tabqa air base, in Raqqa, Syria AP 2/74 IsisThe New Hampshire Senate on Monday confirmed the nomination of Sen. John McCain John Sidney McCainUpcoming Kavanaugh hearing: Truth or consequences How the Trump tax law passed: Dealing with a health care hangover Kavanaugh's fate rests with Sen. Collins MORE's (R-Ariz.) replacement as the committee chairman of the Senate Armed Services Committee, which is chaired by Sen. Jack Reed John (Jack) Francis ReedAdmiral defends record after coming under investigation in 'Fat Leonard' scandal New York Times: Trump mulling whether to replace Mattis after midterms Overnight Defense: Biden honors McCain at Phoenix memorial service | US considers sending captured ISIS fighters to Gitmo and Iraq | Senators press Trump on ending Yemen civil war MORE (D-R.I.). ADVERTISEMENT McCain's confirmation comes just days after it was announced that the committee was delaying a vote on his nomination until at least July 7. The panel is holding confirmation hearings for five other nominees who were nominated to fill senior Pentagon positions, including the secretaries of the Army, Navy, Air Force and Marine Corps, Defense Secretary Jim Mattis James Norman MattisTurkey-Russia Idlib agreement: A lesson for the US Trump says willing to meet with Maduro, but keeps 'all options' open Pentagon withdrawing some missileWispa Campaign Another Sweet Success - A Kinetic Novel Forgotton Anne FORM forma.8 Formata Formula Fusion Forsaken Uprising Fort Defense Fort Meow Fortified Fortissimo FA Fortix Fortix 2 FortressCraft Evolved Forward to the Sky Fossil Echo Foto Flash FOTONICA Foul Play Another Four Last Things Four Realms FourChords Guitar Karaoke Fourtex Jugo Fox & Flock Fox Hime Fox Hime Zero Fractal Fracture the Flag Fractured Space Fragmental Fragments of Him Framed Wings Fran Bow Franchise Hockey Manager 2 Franchise Hockey Manager 2014 Franchise Hockey Manager 3 Franchise Hockey Manager 4 Francisca Frankenstein: Master of Death Frantic Freighter Freaky Awesome Freddi Fish 2: The Case of the Haunted Schoolhouse Freddi Fish and the Case of the Missing Kelp Seeds Frederic: Evil Strikes Back Frederic: Resurrection of Music Frederic: Resurrection of Music Director's Cut Free to Play Freebie FreeCell Quest Freedom Cry Freedom Fall Freedom Planet Freedom Poopie Freeman: Guerrilla Warfare FreeStyle 2: Street Basketball FreeStyleFootball FreezeME Frequent

Step 2: Find memorized text

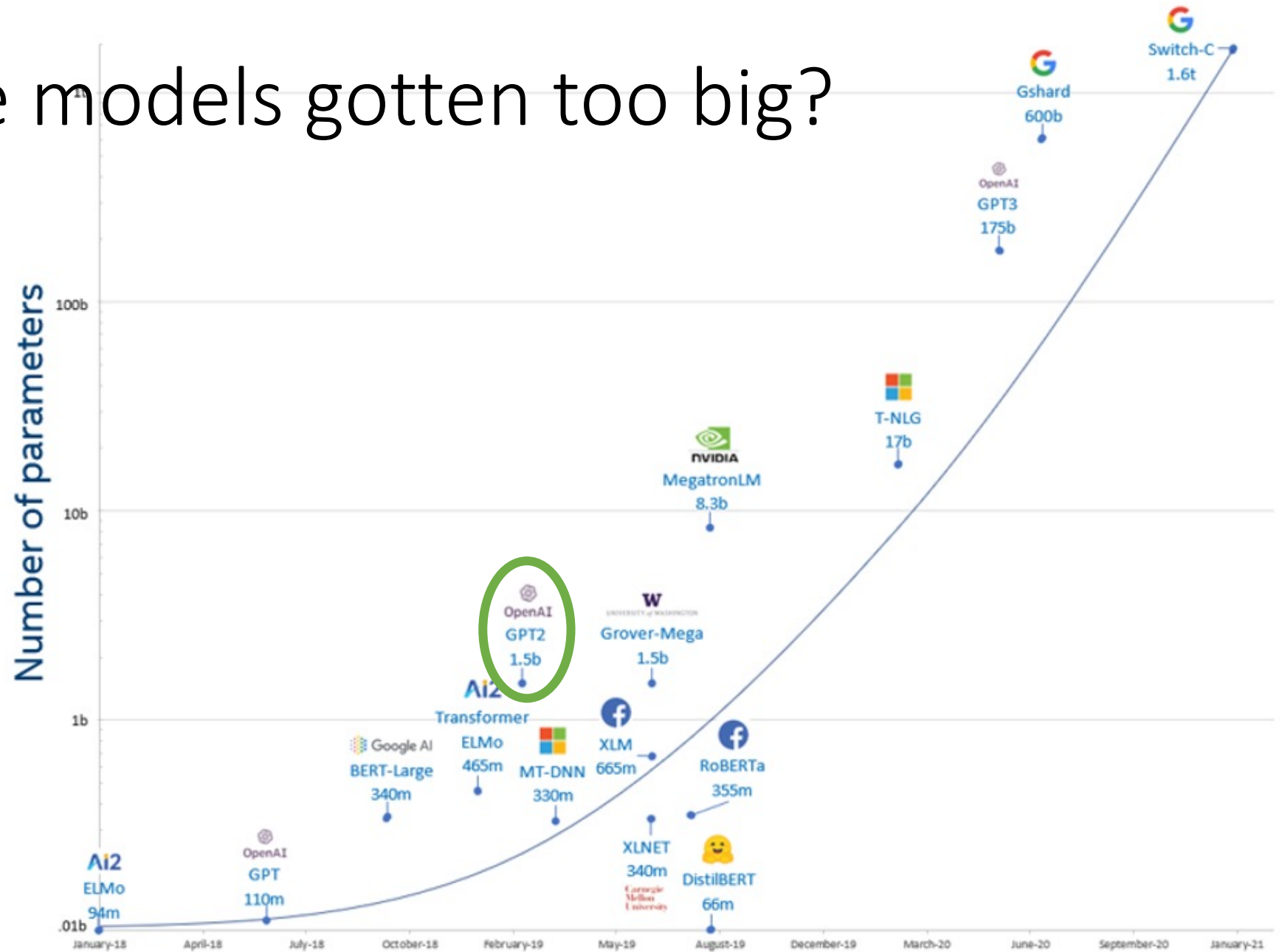
A federal appeals court on Wednesday struck down Texas' voter-ID law, which the Supreme Court had blocked last year. The ruling could potentially affect the upcoming elections in a number of states. Here's what you need to know about the ruling. ([Claritza Jimenez/The Washington Post](#)) A federal appeals court on Wednesday struck down Texas' voter-ID law, which the Supreme Court had blocked last year. The ruling could potentially affect the upcoming elections in a number of states. Here's what you need to know about the ruling. (Claritza Jimenez/The Washington Post) A federal appeals court on Wednesday struck down Texas' voter-ID law, which the Supreme Court had blocked last year. The ruling could potentially affect the upcoming elections in a number of states. Here's what you need to know about the ruling. (Claritza Jimenez/The Washington Post) The Supreme Court on Tuesday dealt a major setback to Texas — and to Republican efforts to restrict the vote — by gutting the law that the high court had upheld last year. In doing so, the justices left in place one provision of the law — a requirement that voters show one of seven acceptable forms of photo identification at the polls to castRails in the Garden - [VR MMO Heaven Forest NIGHTS Heaven Island](#) - [VR MMO Heaven Island Life Heavenly Battle Heavenstrike Rivals](#)@ [Heavily Armed Heavy Bullets Heavy Fire: Afghanistan Heavy Fire: Shattered Spear Heavy Gear Assault Heavy Metal Machines Heckabomb Hegemony III: Clash of the Ancients Hegemony Rome: The Rise of Caesar Heileen 1: Sail Away Heileen 2: The Hands Of Fate Heileen 3: New Horizons Heirs And Graces Hektor Heldric](#) - The legend of the shoemaker Helen's Mysterious Castle Heli Heroes Heliborne Helium Rain Hell Girls Hell Warders HellAngel Hellblade: Senua's Sacrifice Hellenica HellGunner HELLION Hello From Indiana HELLO LADY! Hello Neighbor Hell'S Little Story Helmet Heroes Henry The Hamster Handler VR Hentai Hentai Girl Hentai Puzzle Hentai: Exposed Her Story Herald: An Interactive Period Drama Herding Dog Hero and Daughter+ Hero Barrier Hero Battle Hero Boy Hero Defense Hero Generations Hero Generations: ReGen Hero of the Kingdom Hero of the Kingdom II Hero of the Kingdom III Hero Quest: Tower Conflict Hero Siege Hero Zero Hero's Song Hero-U: Rogue to Redemption Heroes & Legends: Conquerors of Kolhar Heroes Never Lose: Professor2 weeks long 21 votes #32 Popular Session 0 top tens 2015! #31 Rory got bored looking "The Internet Explained" on YouTube... so he decided to put on a show! He talks about the history of the Internet and what it has done for our daily lives.This post may contain referral/affiliate links. [If you buy something, MSA may earn a commission. Read the full disclosure We have the exclusive First Look spoilers for the October 2016 Birchbox! \(Thanks to reader Sarah for the heads-up!\)](#) Each box will include: A selection of 5-star beauty products, from brands including L'Oréal, Smashbox, and more A mystery beauty product with value of at least \$45 A surprise gift And you'll also receive a bonus item (valued at at least \$12.50) when you sign-up. Here are the details for this month's box: Birchbox October 2016 Box – \$45 Value Check out our Birchbox reviews to learn more about this monthly beauty subscription box! [Liz is the founder of My Subscription Addiction. She's been hooked on subscription boxes](#) since 2011 thanks to BirchFormer top American financial regulation lawmaker Mary Ferguson could offer crucial leadership services moving Democratic-only Pennsylvania through unchidden regulatory turmoil facing states reeling. She can also help Democrats in Congress who are struggling to defend a number of seats they won in 2010, including the seat held by Sen. Bob Casey Robert (Bob) Patrick CaseyDems hold edge in Rust Belt Senate races: poll Malnutrition Awareness Week spotlights the importance of national nutrition programs Poll: Democrats hold big leads in Pennsylvania Senate, governor races MORE (D). ADVERTISEMENT The two are the most endangered Democrats in the House. Casey, who is facing a tough race to keep his seat, could be a prime target for Republicans, who have been trying to unseat him ever since he was appointed in 2011. His district is one of 10 in Pennsylvania with a GOP majority. Ferguson, a former member of the House Financial Services Committee, has been a leader of the opposition to the Dodd-Frank financial reform law. She recently announced her candidacy for Senate, and could help Senate Democrats win back the seat held by Sen. Scott Brown Scott Eric [TrumpAvenatti: Third Kavanaugh accuser will prove credible against Kavanaugh, other 'privileged white guys' who defend him Grassley's office says it](#) Gin Fractions In Alcoholic BrewMigal "ElbowDropse/Zaknoratraseru" Shattil is a professional CS:GO player. He is currently playing for HellRaisers. Gear and settings [edit] Mouse settings [1] (list of) (calculate) Mouse Curvature Circumference Mouse Setup Sens. Zoom Raw. [ZOWIE by BenQ ZA14 1168 MPI 0.762 deg/mm 21.3 in/rev 47.4 cm/rev 400 CPI @ 1000 Hz 2.8 1 On 600](#) Last updated on 2017-01-15 (119 days ago). Mouse Mousepad [ZOWIE by BenQ ZA14 \(X\) ZA14 \(O\) SteelSeries QcK Heavy Monitor Refresh rate In-game resolution Scaling ZOWIE by BenQ XL2540 240 Hz 1024x768 Black Bars](#) Keyboard Headset Logitech G400 Last updated on 2017-01-15 (119 days ago). [Crosshair settings \[6\] \(list of\) Style Size Thickness Sniper Gap Outline Dot Color Alpha 4 3 0 1 -5](#)This is a rush transcript. Copy may not be in its final form. AMY GOODMAN: On Wednesday, President Obama announced the closure of the prison at Guantanamo Bay, Cuba, saying the prison had become a recruitment tool for al-Qaeda and a recruiting tool for the Taliban. The president also called for a transfer of the remaining 166 detainees to U.S. prisons. The decision came after a review of the prison conducted by his administration. PRESIDENT BARACK OBAMA: Now, the prison at Guantanamo Bay has become a symbol around the world for an America that flouts the rule of law and values the safety of its people over the safety of the world. It's time for the United States to send a new message to the world: We're not looking to prosecute individuals based on who they are or where they came from. We're looking to prosecute terrorists, and we're going to do it with speed and conviction. I've ordered a review of the cases of those currently detained. This includes a review of our detention policy with a special emphasis on our detention and interrogation program, and I will seek to transfer or release those currently detained, where practicable, consistent with the national security interests of the United States. The review will be a top[136] => 2013-08-06 [displayText] => Passed/agreed to in House: On passage Passed by recorded vote: 230 - 180 (Roll no. 603).(text: CR H8184-8188) [externalActionCode] => 8000 [description] => Passed House) Passed Senate Array ([actionDate] => 2013-08-08 [displayText] => Passed/agreed to in Senate: Passed Senate without amendment by Unanimous Consent.(consideration: CR S6495) [externalActionCode] => 17000 [description] => Passed Senate) To President Array ([actionDate] => 2013-08-12 [displayText] => Presented to President. [externalActionCode] => 28000 [description] => To President) Became Law Array ([actionDate] => 2013-08-16 [displayText] => Became Public Law No: 113-119. [externalActionCode] => 36000 [description] => Became Law) LAW 64. H.R.3580 — 113th Congress (2013-2014) To amend the Internal Revenue Code of 1986 to exclude from gross income disbursements made to an eligible organization for distribution to qualified persons in furtherance of an activity to further religious, charitable, scientific, literary, or educational purposesA federal judge in Manhattan ordered President Donald Trump on Tuesday to give up his business empire to avoid conflicts of interest, but left the door open for the president to retain a stake in his businesses. In a ruling that could have far-reaching consequences, U.S. District Judge George Daniels said Mr Trump's businesses could continue operating without violating the Constitution, but the court did not require him to sell or divest himself of them. "This case does not involve an unconstitutional conflict of interest," Mr Daniels wrote. The ruling came days after Mr Trump issued an executive order that effectively gave his sons, including senior White House adviser Donald Trump Jr., control of the family business, the Trump Organization. The order did not divest the president of any interest in the company. Mr Trump is the president of the Trump Organisation, whose business interests include Trump Tower in New York City and a variety of other assets. [Shape Created with Sketch. Trump Inauguration protests around the World Show all 14 left Created with Sketch. right Created with Sketch. Shape Created with Sketch.](#) Trump Inauguration protests around the World 1/14 [Activists from Greenpeace display a message reading "Mr President, walls divide. Build Bridges!" along the Berlin wall in Berlin](#) on "What people believe one year before this horrific happening makes fools seem serious like I'll bring ISIS straight along... in February," said Mr Farage in a speech to UKIP's annual conference in London. He added: "It is time to stop talking about ISIS, to stop making speeches about 'we are going to defeat them'... to get serious. It is time to do what we are actually good at, which is defeating Labour in a general election." But the UKIP leader said he believed it was possible to defeat Islamic State "one way or another" and that there would be no easy way of tackling the issue. "There is no way of defeating them one way or another," said Mr Farage. "There is only getting on with it - doing all of the very simple things that we all know will actually have an impact." [Shape Created with Sketch.](#) In pictures: The rise of Isis [Show all 74 left Created with Sketch. right Created with Sketch. Shape Created with Sketch.](#) In pictures: The rise of Isis 1/74 Isis fighters Fighters of the Islamic State wave the group's flag from a damaged display of a government fighter jet following the battle for the Tabqa air base, in Raqqa, Syria AP 2/74 Isis The New Hampshire Senate on Monday confirmed the nomination of Sen. [John McCain John Sidney McCainUpcoming Kavanaugh hearing: Truth or consequences How the Trump tax law passed: Dealing with a health care hangover Kavanaugh's fate rests with Sen. Collins MORE's \(R-Ariz.\) replacement as the committee chairman of the Senate Armed Services Committee, which is chaired by Sen. Jack Reed John \(Jack\) Francis ReedAdmiral defends record after coming under investigation in 'Fat Leonard' scandal](#) New York Times: Trump mulling whether to replace Mattis after midterms [Overnight Defense: Biden honors McCain at Phoenix memorial service | US considers sending captured ISIS fighters to Gitmo and Iraq | Senators press Trump on ending Yemen civil war MORE \(D-R.I.\).](#) ADVERTISEMENT McCain's confirmation comes just days after it was announced that the committee was delaying a vote on his nomination until at least July 7. The panel is holding confirmation hearings for five other nominees who were nominated to fill senior Pentagon positions, including the secretaries of the Army, Navy, Air Force and Marine Corps, Defense Secretary Jim Mattis James Norman MattisTurkey-Russia Idlib agreement: A lesson for the US Trump says willing to meet with Maduro, but keeps 'all options' open Pentagon withdrawing some missileWispa Campaign Another Sweet Success - [A Kinetic Novel Forgotton Anne FORM forma.8 Formata Formula Fusion Forsaken Uprising Fort Defense Fort Meow Fortified Fortissimo FA Fortix Fortix 2 FortressCraft Evolved Forward to the Sky Fossil Echo Foto Flash FOTONICA Foul Play Four Last Things Four Realms FourChords Guitar Karaoke Fourtex Jugo Fox & Flock Fox Hime Fox Hime Zero Fractal Fracture the Flag Fractured Space Fragmental Fragments of Him Framed Wings Fran Bow Franchise Hockey Manager 2 Franchise Hockey Manager 2014 Franchise Hockey Manager 3 Franchise Hockey Manager 4 Francisca Frankenstein: Master of Death Frantic Freighter Freaky Awesome Freddi Fish 2: The Case of the Haunted Schoolhouse Freddi Fish and the Case of the Missing Kelp Seeds Frederic: Evil Strikes Back Frederic: Resurrection of Music Frederic: Resurrection of Music Director's Cut Free to Play Freebie FreeCell Quest Freedom Cry Freedom Fall Freedom Planet Freedom Poopie Freeman: Guerrilla Warfare FreeStyle 2: Street Basketball FreeStyleFootball FreezeME Frequent](#)

Have language models gotten too big?

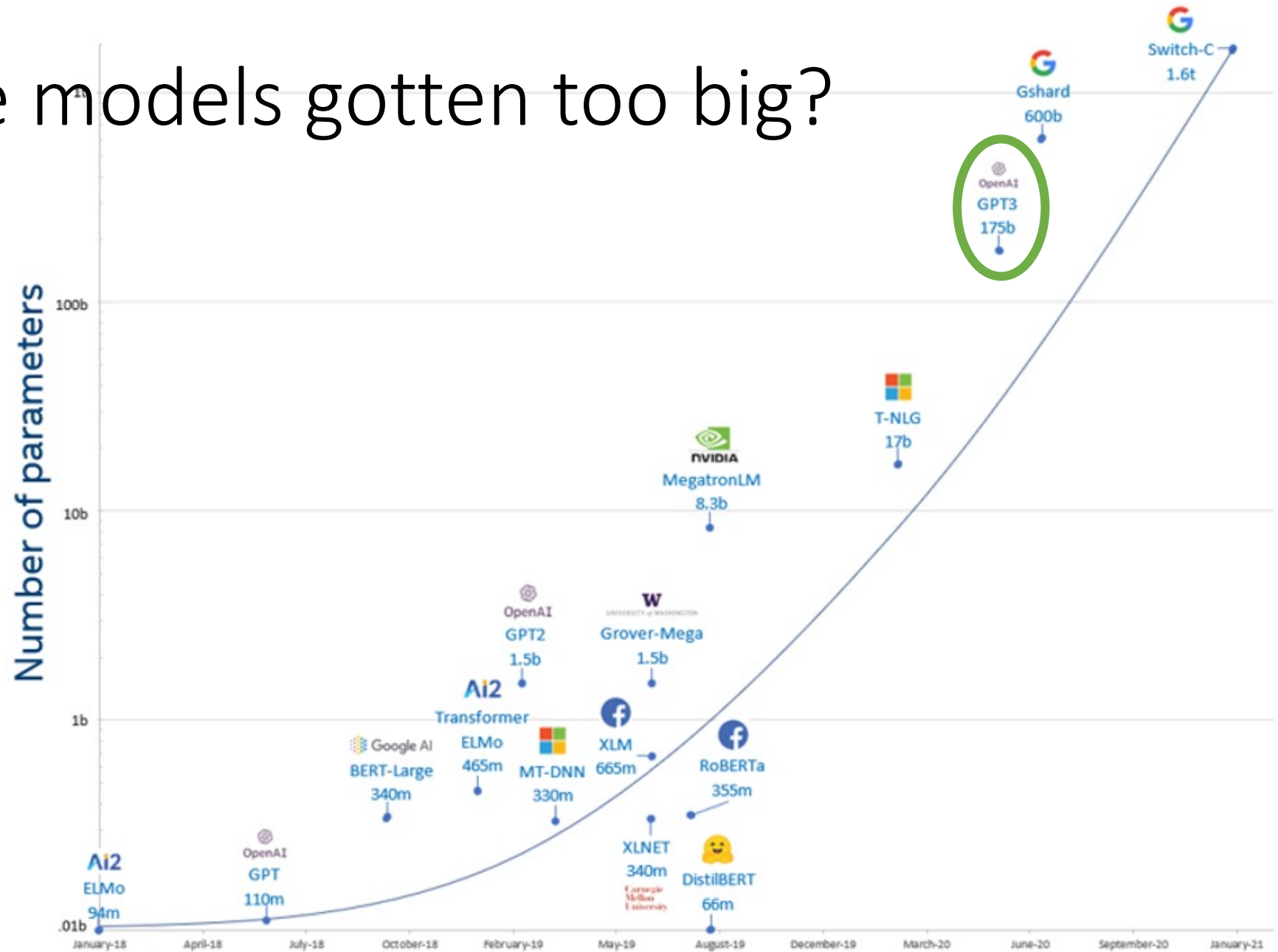


- GPT2 model size = ~6 GB
- English Wikipedia = ~10 GB (compressed)

Have language models gotten too big?



Have language models gotten too big?



prompt

The escape of the Brazilian boa constrictor earned Harry his longest-ever punishment. By the time he was allowed out of his cupboard again, the summer holidays had started and Dudley had already broken his new video camera, crashed his remote-control aeroplane, and, first time out on his racing bike, knocked down old Mrs Figg as she crossed Privet Drive on her crutches.

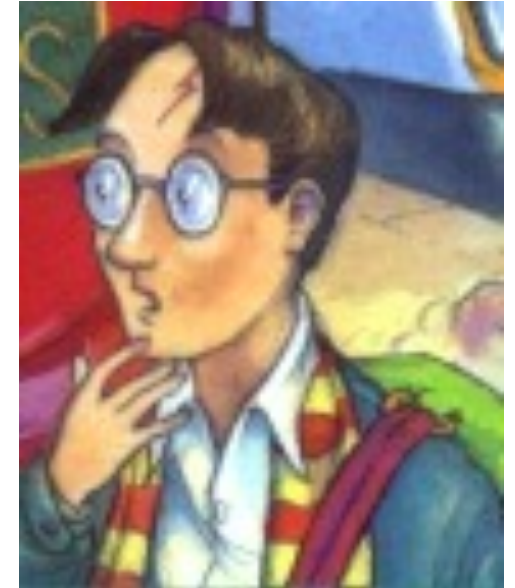
Harry was glad school was over, but there was no escaping Dudley's gang, who visited the house every single day. Piers, Dennis, Malcolm, and Gordon were all big and stupid, but as Dudley was the biggest and stupidest of the lot, he was the leader. The rest of them were all quite happy to join in Dudley's favourite sport: Harry Hunting.

This was why Harry spent as much time as possible out of the house, wandering around and thinking about the end of the holidays, where he could see a tiny ray of hope. When September came he would be going off to secondary school and, for the first time in his life, he wouldn't be with Dudley. Dudley had been accepted at Uncle Vernon's old private school, Smeltings. Piers Polkiss was going there too. Harry, on the other hand, was going to Stonewall High, the local public school. Dudley thought this was very funny.

'They stuff people's heads down the toilet the first day at Stonewall,' he told Harry. 'Want to come upstairs and practise?'

'No, thanks,' said Harry. 'The poor toilet's never had anything as horrible as your head down it — it might be sick.'

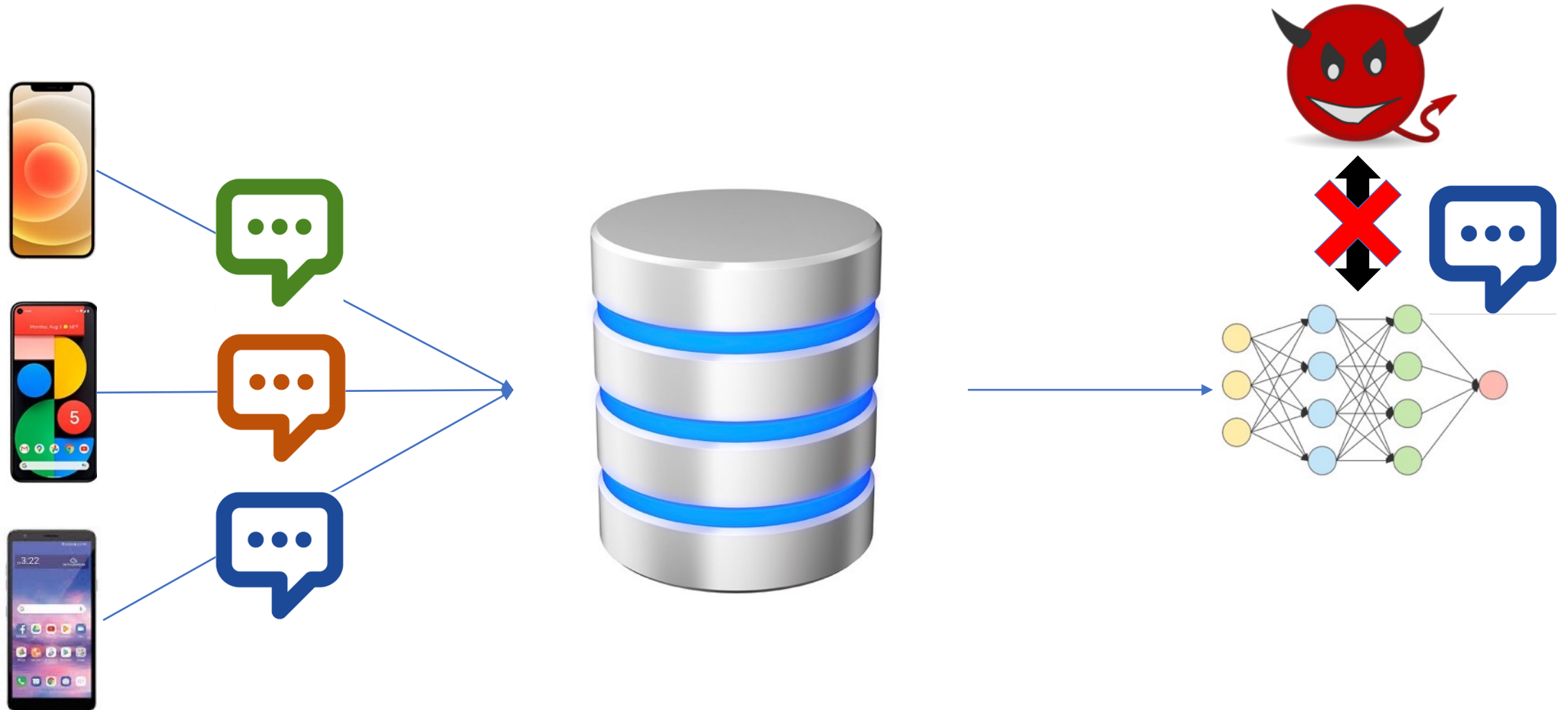
GPT3 output



Goal: train a ML model with “privacy”

➤ what does this mean?

no training data leakage



Goal: train a ML model with “privacy”

➤ what does this mean?

no training data leakage

➤ how can we achieve this?

differential privacy

Goal: train a ML model with “privacy”

➤ what does this mean?

no training data leakage

➤ how can we achieve this?

differential privacy

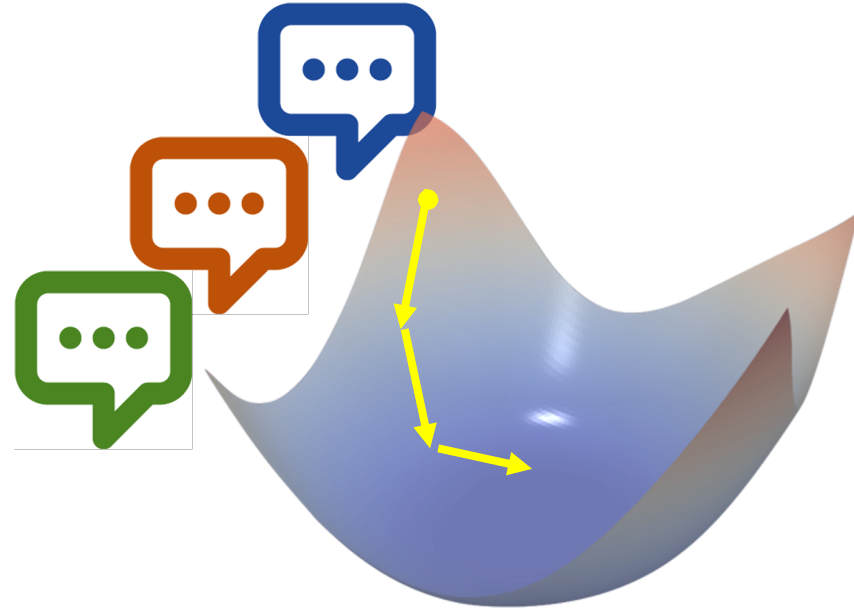
intuition: *randomized* training algorithm is not influenced (too much) by any individual data point

for any two datasets that differ in a single element

$$\frac{\Pr[A_{\text{train}}(\text{green}, \text{orange}, \text{blue}) = \text{ML Model}]}{\Pr[A_{\text{train}}(\text{yellow}, \text{orange}, \text{blue}) = \text{ML Model}]} \leq e^\epsilon$$

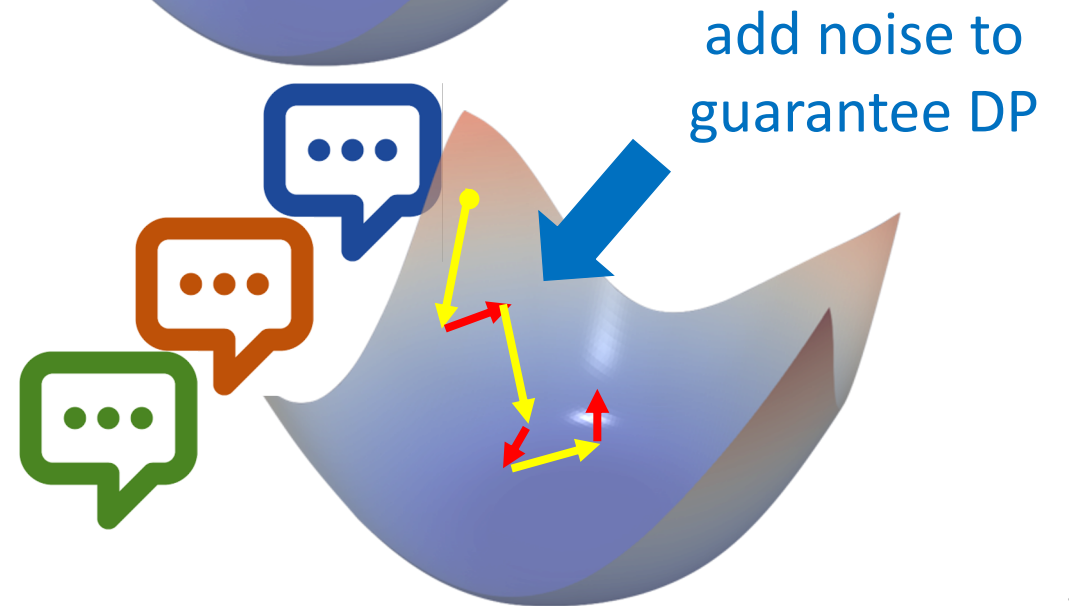
How? Noisy Gradient Descent

gradient descent
(SGD)

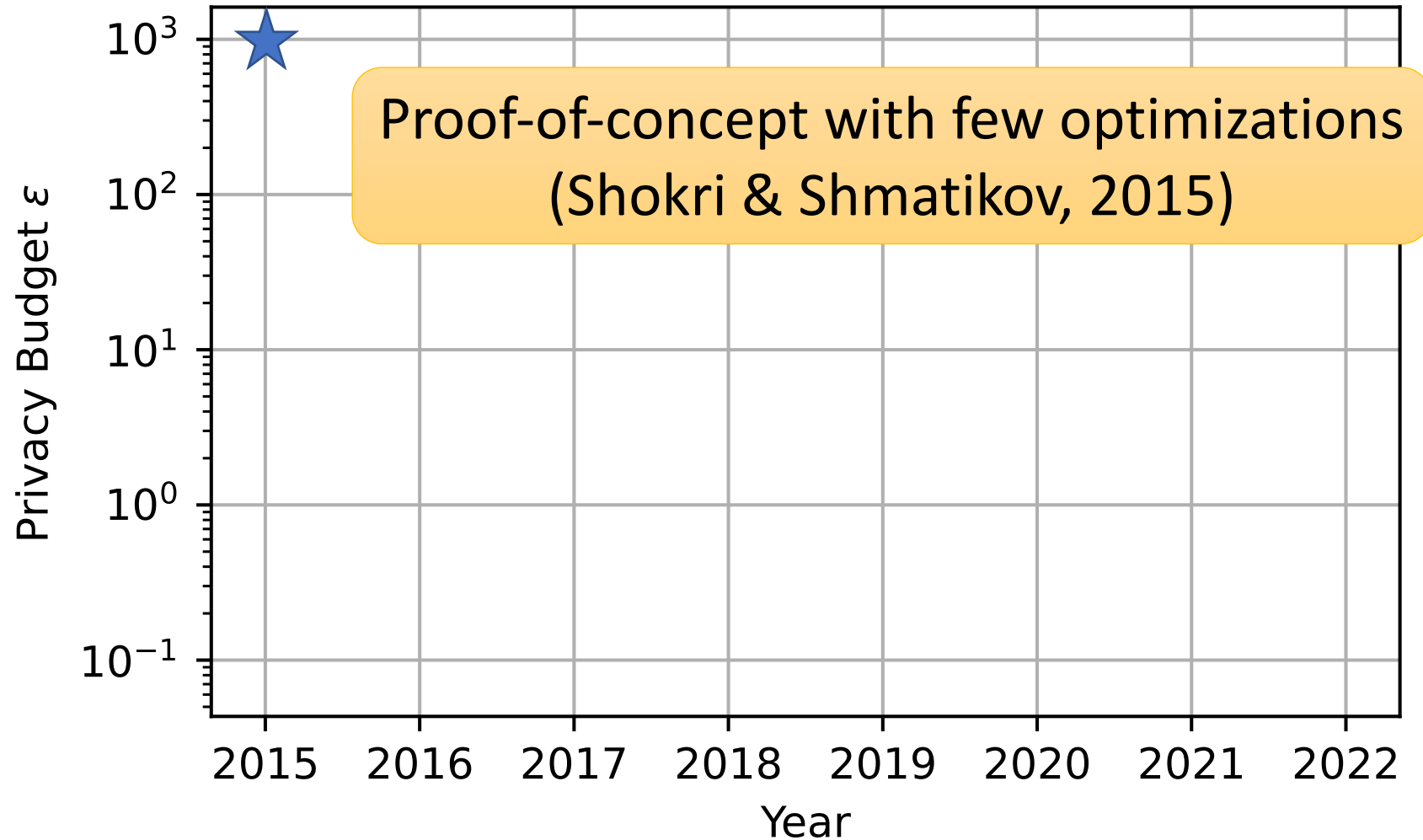


private gradient descent
(DP-SGD)

Chaudhuri et al., '11; Bassily et al. '14;
Shokri & Shmatikov '15; Abadi et al. '16

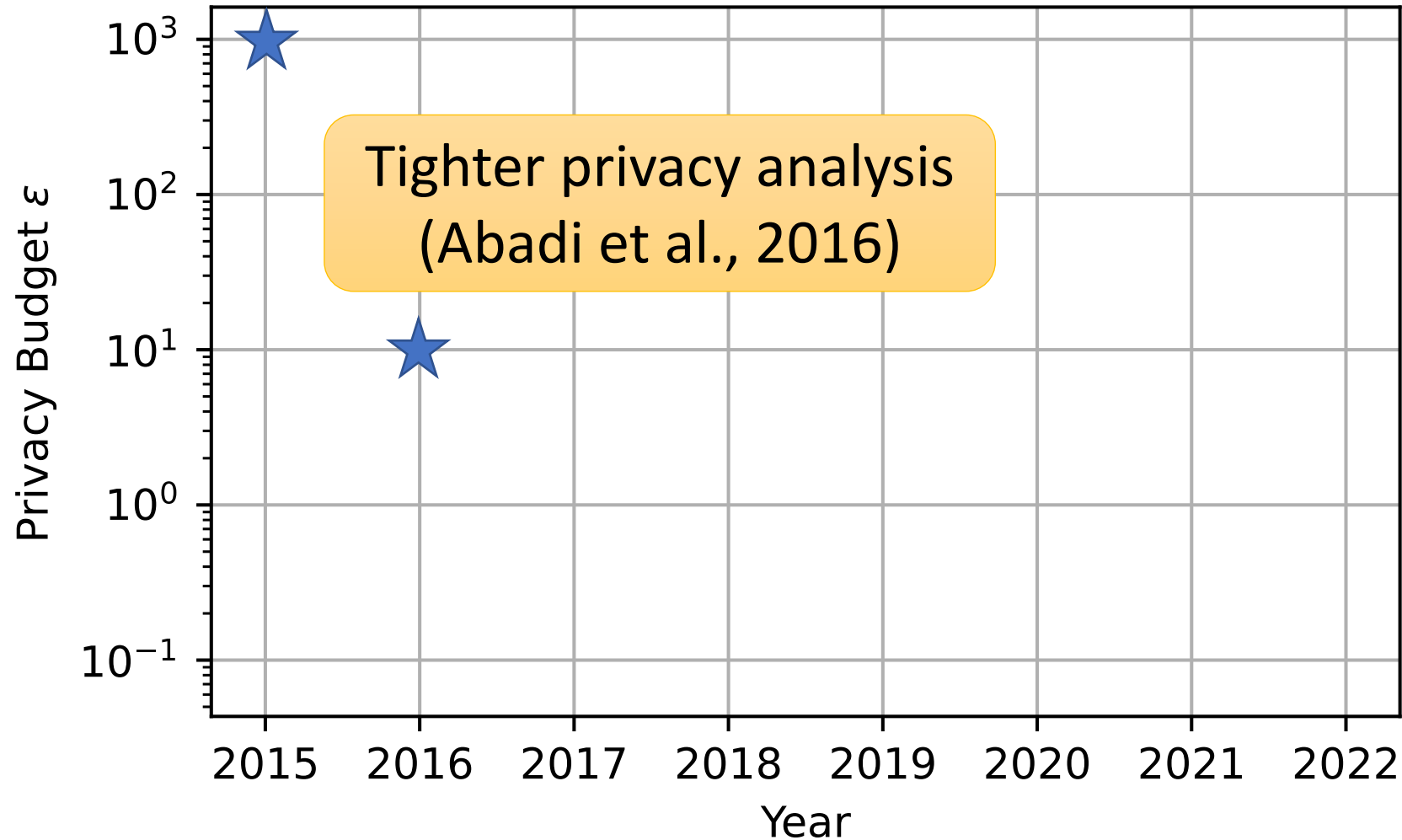


Provable Privacy for 98% accuracy on MNIST

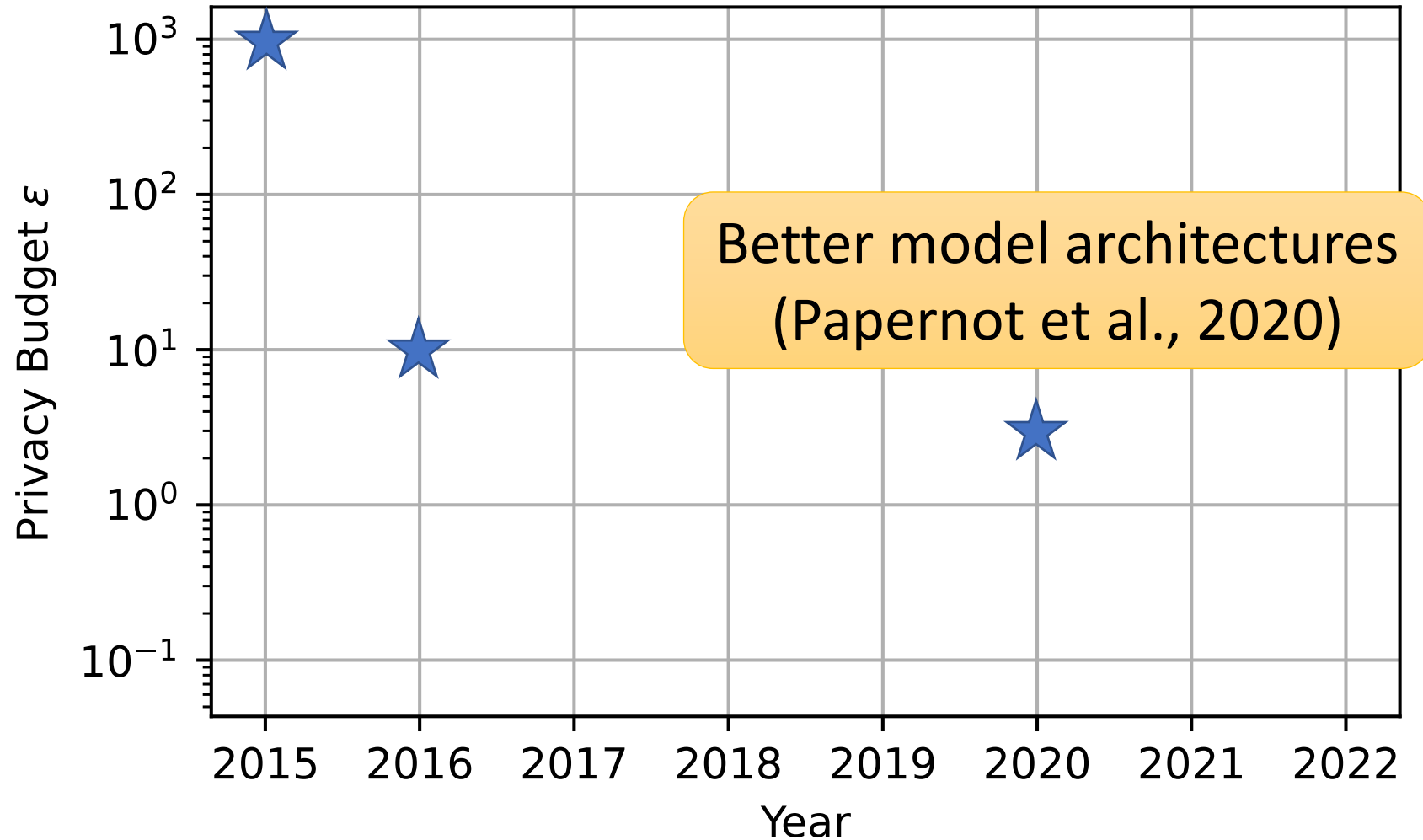


$$\frac{\Pr[A_{\text{train}}(\text{green}, \text{orange}, \text{blue}) = \text{NN}]}{\Pr[A_{\text{train}}(\text{yellow}, \text{orange}, \text{blue}) = \text{NN}]} \geq e^{1000}$$

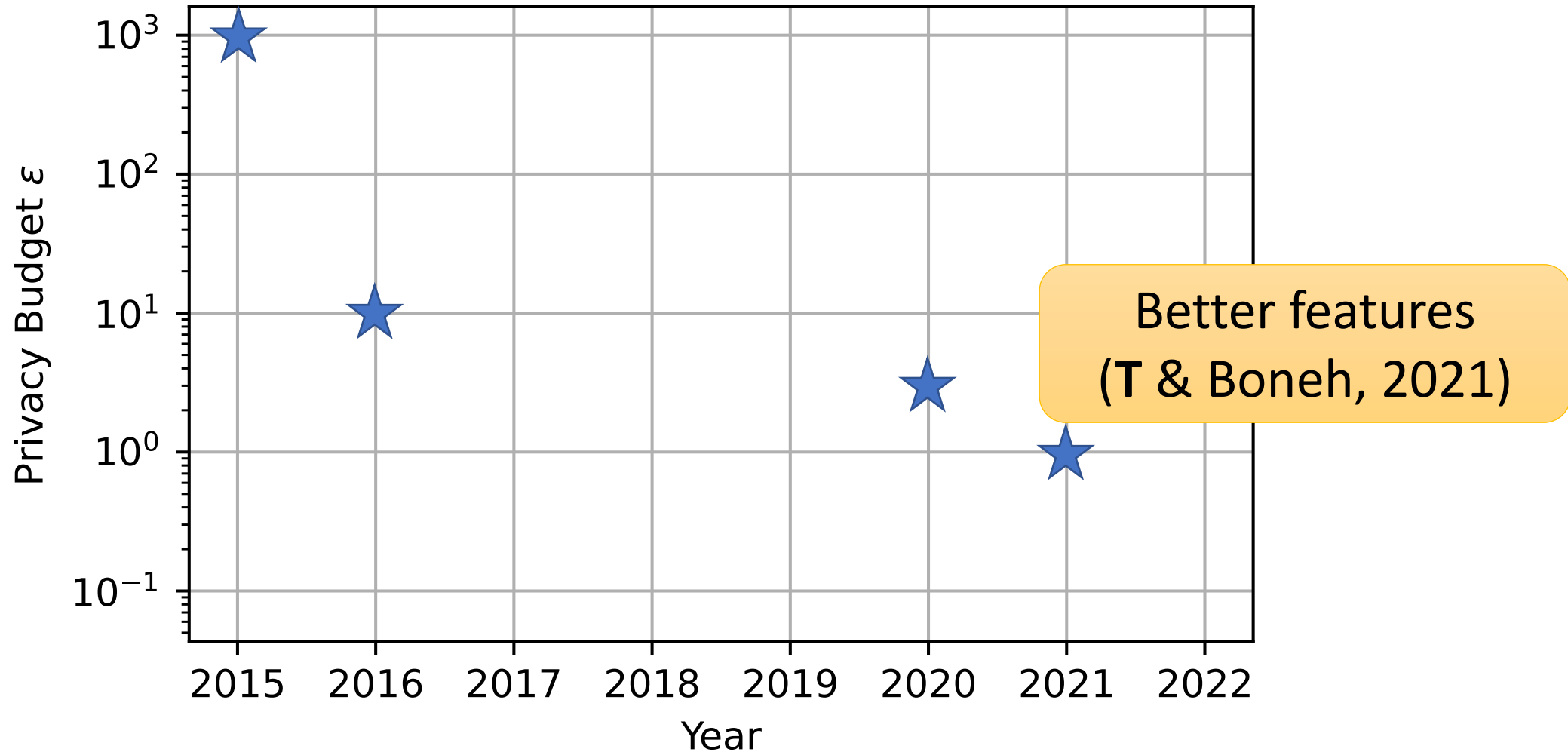
Provable Privacy for $\approx 98\%$ accuracy on MNIST



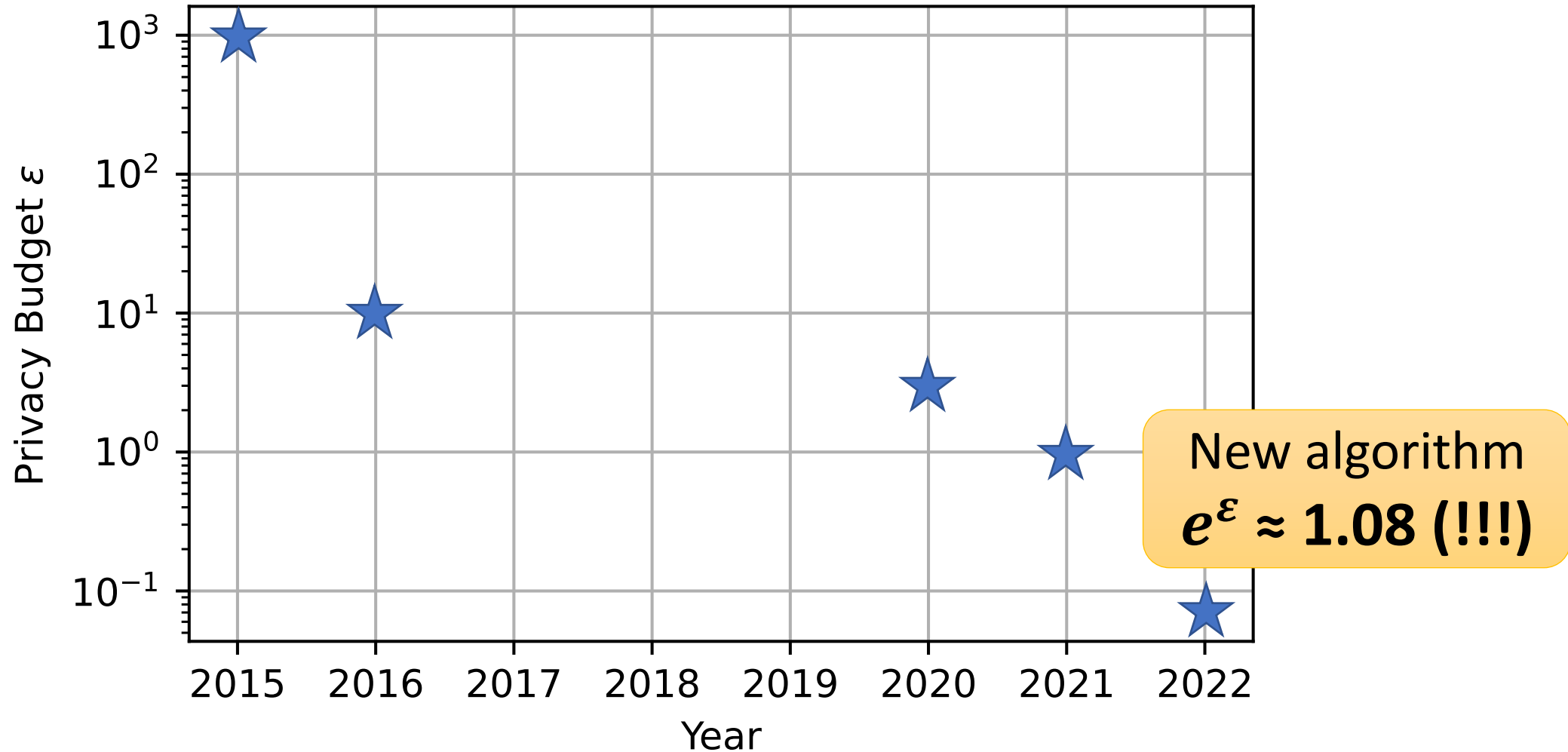
Provable Privacy for $\approx 98\%$ accuracy on MNIST



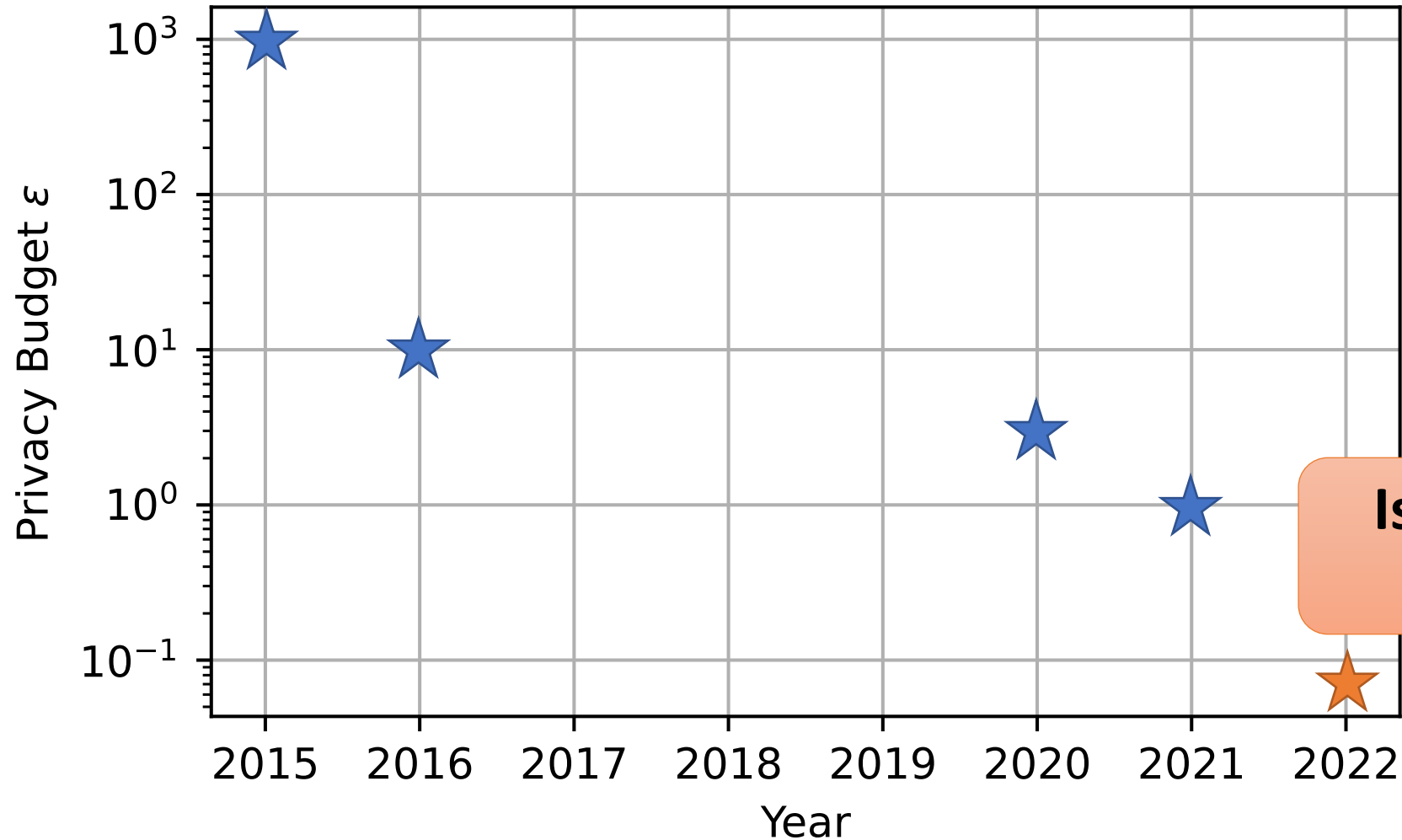
Provable Privacy for $\approx 98\%$ accuracy on MNIST



Provable Privacy for $\approx 98\%$ accuracy on MNIST



Provable Privacy for $\approx 98\%$ accuracy on MNIST



**Is this claim
*correct?***

How to check a privacy claim?

➤ Check the proof

$$\begin{aligned} & c(o_{1:k}; \mathcal{M}_{1:k}, o_{1:(k-1)}, d, d') \\ &= \log \frac{\Pr[\mathcal{M}_{1:k}(d; o_{1:(k-1)}) = o_{1:k}]}{\Pr[\mathcal{M}_{1:k}(d'; o_{1:(k-1)}) = o_{1:k}]} \\ &= \log \prod_{i=1}^k \frac{\Pr[\mathcal{M}_i(d) = o_i \mid \mathcal{M}_{1:(i-1)}(d) = o_{1:(i-1)}]}{\Pr[\mathcal{M}_i(d') = o_i \mid \mathcal{M}_{1:(i-1)}(d') = o_{1:(i-1)}]} \\ &= \sum_{i=1}^k \log \frac{\Pr[\mathcal{M}_i(d) = o_i \mid \mathcal{M}_{1:(i-1)}(d) = o_{1:(i-1)}]}{\Pr[\mathcal{M}_i(d') = o_i \mid \mathcal{M}_{1:(i-1)}(d') = o_{1:(i-1)}]} \\ &= \sum_{i=1}^k c(o_i; \mathcal{M}_i, o_{1:(i-1)}, d, d'). \end{aligned}$$

Thus

$$\begin{aligned} & \mathbb{E}_{o'_{1:k} \sim \mathcal{M}_{1:k}(d)} [\exp(\lambda c(o'_{1:k}; \mathcal{M}_{1:k}, d, d')) \mid \forall i < k: o'_i = o_i] \\ &= \mathbb{E}_{o'_{1:k} \sim \mathcal{M}_{1:k}(d)} \left[\exp \left(\lambda \sum_{i=1}^k c(o'_i; \mathcal{M}_i, o_{1:(i-1)}, d, d') \right) \right] \\ &= \mathbb{E}_{o'_{1:k} \sim \mathcal{M}_{1:k}(d)} \left[\prod_{i=1}^k \exp(\lambda c(o'_i; \mathcal{M}_i, o_{1:(i-1)}, d, d')) \right] \\ & \hspace{15em} \text{(by independence of noise)} \end{aligned}$$

How to check a privacy claim?

➤ Check the proof

➤ Check the code

```
def process_microbatch(i, sample_state):
    """Process one microbatch (record) with privacy helper."""
    microbatch_loss = tf.reduce_mean(
        input_tensor=tf.gather(microbatches_losses, [i]))
    with gradient_tape.stop_recording():
        grads = gradient_tape.gradient(microbatch_loss, var_list)
    sample_state = self._dp_sum_query.accumulate_record(
        sample_params, sample_state, grads)
    return sample_state

for idx in range(self._num_microbatches):
    sample_state = process_microbatch(idx, sample_state)

grad_sums, self._global_state, _ = (
    self._dp_sum_query.get_noised_result(sample_state,
                                         self._global_state))
```

How to check a privacy claim?

- Check the proof
- Check the code
- **Launch an attack!**



Auditing Differential Privacy with Attacks

Jagielski et al. 2020, Nasr et al. 2021, Carlini et al. (forthcoming)

$$\frac{\Pr[A_{\text{train}}(\text{3 4 2}) = \text{NN}]}{\Pr[A_{\text{train}}(\text{3 4 1}) = \text{NN}]} \geq e^\epsilon$$

let's estimate this!

Auditing Differential Privacy with Attacks

Jagielski et al. 2020, Nasr et al. 2021

Challenger

Adversary

$D_0 =$ 

$D_1 =$ 

flip a bit b

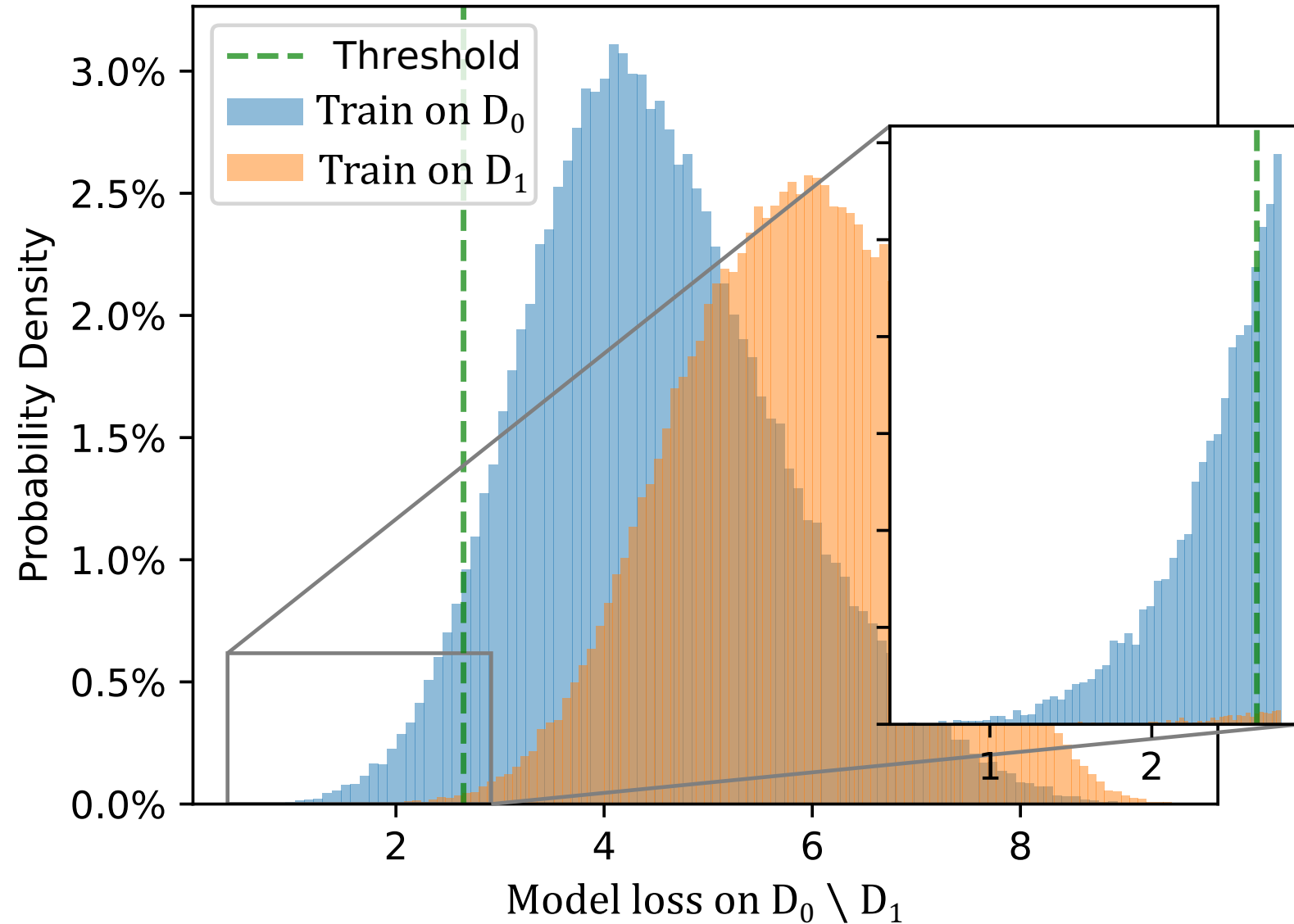
$$f = A_{\text{train}}(D_b)$$

output guess \hat{b}

Adv wins if $\hat{b} = b$

Adversary's success rate \rightarrow lower bound on ϵ

Repeat the game 100'000 times...



$e^\epsilon > 16$
(claim was $e^\epsilon \leq 1.08$)

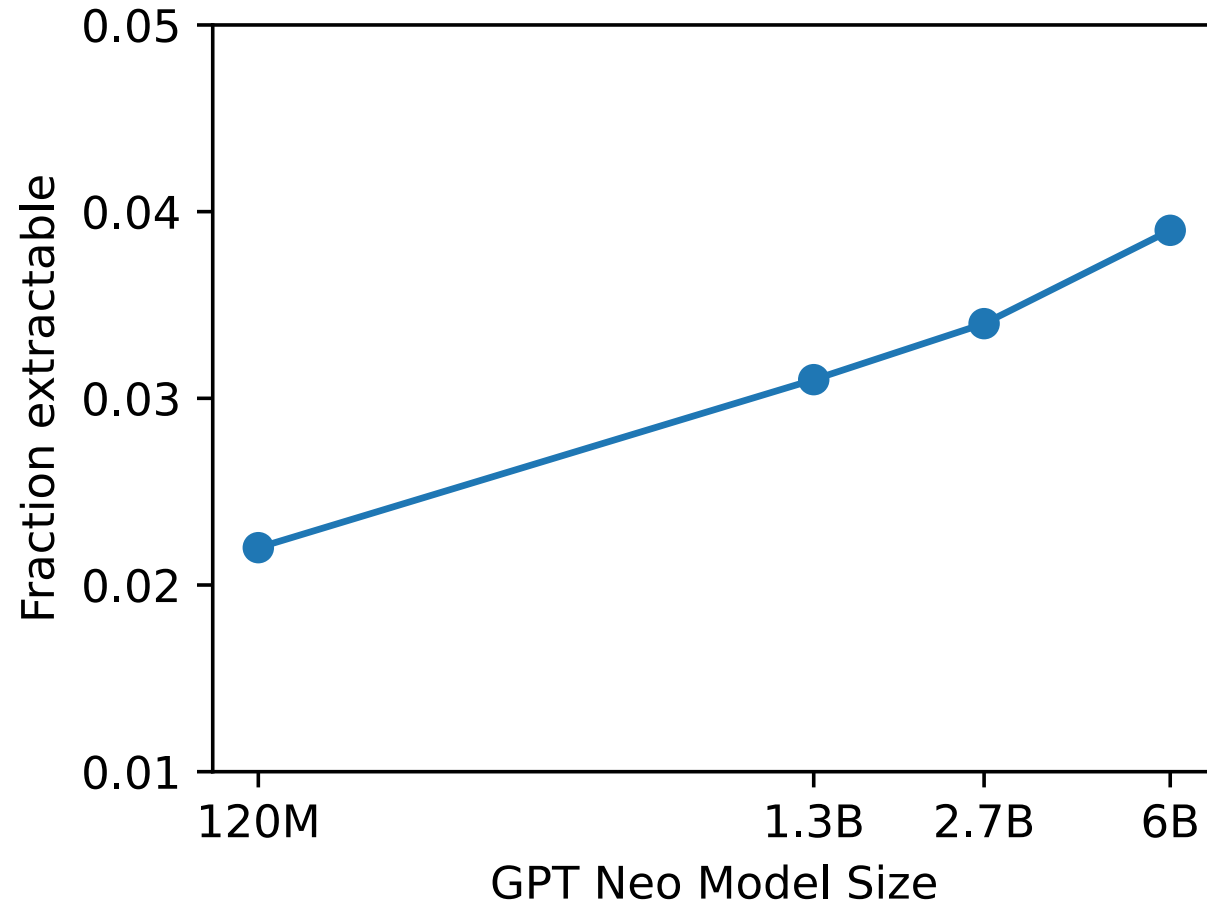
Conclusion

- **Machine learning is not private “by default”!**
 - Without (strong) cryptography, you must **trust someone** with your data
 - Trained models **leak** rare training data
- **Solutions exist but have limitations!**
 - Secure decentralized learning has **high performance overhead**
 - Strong differential privacy **harms model accuracy**
 - *Privacy guarantees must be rigorously defined and audited!*

Thank you!

Scaling laws for memorization

“Quantifying Memorization Across Neural Language Models”. 2022



Scaling laws for memorization

“Quantifying Memorization Across Neural Language Models”. 2022

