How Private is Machine Learning?

Nicholas Carlini
Google
Deep learning takes on tumours

Artificial-intelligence methods are moving into cancer research.

Esther Landhuis
Would you like to grab some coffee with me in a
SUBJECT: Write emails faster with Smart Compose in Gmail

Hey Jacqueline,

Haven't seen you in a while
LONG LIVE THE REVOLUTION. OUR NEXT MEETING WILL BE AT THE DOCKS AT MIDNIGHT ON JUNE 28.

AHA, FOUND THEM!

WHEN YOU TRAIN PREDICTIVE MODELS ON INPUT FROM YOUR USERS, IT CAN LEAK INFORMATION IN UNEXPECTED WAYS.
This talk:
1. Does this happen?
2. Can we prevent it?
This talk:
1. Does this happen?
2. Can we prevent it?
Act I: Training Data Extraction
Let's try to empirically measure things.

Does GPT-2 know your phone number?
this is a random input

and probably the worst one. If the predicted value is not a whole number...
correct horse battery

GPT-2

terms when you hear them on the TV. Check to see if any or all of your batteries have dead times on
Gianopulos said. "We stand by our filmmakers and their right to free expression. While we respect and appreciate the freedom of expression that creators are afforded, we cannot allow the actions of a few to undermine the principles that this country was founded on and which we continue to cherish.

It's been a rough few weeks for Sony. After Sony involved that the film NOT proceed. "While we respect and appreciate the freedom of expression that creators are afforded, we cannot allow the actions of a few to undermine the principles that this country was founded on and which we continue to cherish.

The bill would require broadband providers to obtain customer consent before collecting data on their online activities, including the time spent on them and The new, highly-anticipated movie, "The Interview," has been cancelled. The studio's CEO, Jim

The U.S. Senate's top Democrat has introduced a bill that would give the Federal Communications Commission power to create privacy rules for internet service providers. Sen. Ed Markey Edward (Ed) John MarkeyThis week:

The “Kate’s Law” — named after Kathryn Steinle, a 32-year-old woman who was shot in San Francisco and later died after a federal judge ordered the release of her alleged killer in December 2015 — would create penalties for those convicted of crimes after being deported for immigration violations.

This means the product was tested and repaired as required to meet the standards of the refurbisher, which may not be the original manufacturer. Any exceptions to the condition of the item outside the manufacturer’s warranty information should be provided in the listing, up to and including the warranty details. Sold and Shipped by Newegg Purchases from these Sellers are generally covered under the seller's policy, and not under Newegg's warranty. 

The government has 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 international law, that degrades human dignity, and that is indifferent to the suffering of others.

The aliens
Question: which of this text is memorized?
believe to be essential to our success as a free people," Gianopulos said. "We stand by our filmmakers and their right to ... the time spent on them and The new, highly-anticipated movie, "The Interview," has been cancelled.  The studio's CEO, Jim

The bill would require broadband providers to obtain customer consent before collecting data on their online activities, ... the FCC rules, which he says have not kept pace with the digital revolution.  "The internet has changed so quickly that the FCC has struggled to keep up."

Our Internet service providers have become the most sensitive data in our lives.  "They need to do everything that we can to prevent them from using it to track our behavior and sell it to the highest bidder."

The "Kate's Law" would also create penalties on immigrants who are in the country illegally, according to a report in The Hill.  The "Kate's Law" — named after Kathryn Steinle, a 32-year-old woman who was shot and later died after a federal judge ordered the release of her alleged killer in December 2015 — would create penalties for cities and counties that refuse to cooperate with federal immigration authorities.  The "Kate's Law" would also create penalties on immigrants who are in the country illegally, according to a report in The Hill.  The "Kate's Law" — named after Kathryn Steinle, a 32-year-old woman who was shot and later died after a federal judge ordered the release of her alleged killer in December 2015 — would create penalties for cities and counties that refuse to cooperate with federal immigration authorities.

The "Kate's Law" would also create penalties on immigrants who are in the country illegally, according to a report in The Hill.  The "Kate's Law" — named after Kathryn Steinle, a 32-year-old woman who was shot and later died after a federal judge ordered the release of her alleged killer in December 2015 — would create penalties for cities and counties that refuse to cooperate with federal immigration authorities.  The "Kate's Law" would also create penalties on immigrants who are in the country illegally, according to a report in The Hill.  The "Kate's Law" — named after Kathryn Steinle, a 32-year-old woman who was shot and later died after a federal judge ordered the release of her alleged killer in December 2015 — would create penalties for cities and counties that refuse to cooperate with federal immigration authorities.

Our Newegg Marketplace Guarantee Marketplace SellerThe first major piece of legislation introduced after President Donald ... cities" by prohibiting jurisdictions from withholding certain federal grants or providing certain benefits to people who

This means the product was tested and repaired as required to meet the standards of the refurbisher, which may

The rise of ISIS has been a focus of concern for many countries and organizations.  The rise of ISIS has been a focus of concern for many countries and organizations.  The rise of ISIS has been a focus of concern for many countries and organizations.

With Sketch. Shape Created with Sketch. In pictures: The rise of Isis 1/74 Isis fighters Fighters of the Islamic State wave the flag of a famous from a captured government fighter jet following the battle for the Tabqa air base, in Raqqa, Syria AP 2/74 IsisThe New Hampshire

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 this terror'.

is time to do something, to get serious, to do it with authority, with the shock of a General election." But the UKIP leader said he believed it was impossible to defeat Islamic state one way or another and that there would be no easy way of tackling the issue.  "There is not one answer to this," he added.  "There is not one solution, but a series of solutions."

With Sketch. Shape Created with sketch. In pictures: The rise of Isis 1/74 Isis fighters Fighters of the Islamic State wave the flag of a famous from a captured government fighter jet following the battle for the Tabqa air base, in Raqqa, Syria AP 2/74 IsisThe New Hampshire

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 this terror'.

is time to do something, to get serious, to do it with authority, with the shock of a General election." But the UKIP leader said he believed it was impossible to defeat Islamic state one way or another and that there would be no easy way of tackling the issue.  "There is not one answer to this," he added.  "There is not one solution, but a series of solutions."

With Sketch. Shape Created with sketch. In pictures: The rise of Isis 1/74 Isis fighters Fighters of the Islamic State wave the flag of a famous from a captured government fighter jet following the battle for the Tabqa air base, in Raqqa, Syria AP 2/74 IsisThe New Hampshire
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 upcoming elections in a number of states. Here's what you need to know about the ruling.

President Donald Trump called for a transfer of the remaining 166 detainees to U.S. prisons. The decision came after a review of the prison conducted by President Barack Obama. Obama, who was a critic of the Guantanamo Bay prison, said the decision would be a "step in the right direction".

Senators press Trump on ending Yemen civil war

MORE (D-R.I.) called for a transfer of the remaining 166 detainees to U.S. prisons. The decision came after a review of the prison conducted by President Barack Obama. Obama, who was a critic of the Guantanamo Bay prison, said the decision would be a "step in the right direction".

President Trump 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, wherever practicable, consistent with the national security interests of the United States. The review will be a top priority, and I will seek to transfer or release those currently detained, wherever practicable, consistent with the national security interests of the United States.

Sen. Tom Cotton (R-Ark.) introduced a bill that would require broadband providers to obtain customer consent before collecting data on their online activities, and would ban the sale of that data without consent. The bill is aimed at the FCC rules, which he says have not kept pace with the digital revolution.

"The internet has changed so quickly that the FCC has struggled to keep up," said Sen. Ed Markey Edward (Ed) John MarkeyThis week:

Instead of Monday? Watch above to see if your favorite team won't play this weekend! The new, highly-anticipated movie, "The Interview," has been cancelled. The studio's CEO, Jim Gianopulos, said the decision was "not the right decision" for the company. "We stand by our filmmakers and their right to tell their stories," he said.

It's been a rough few weeks for Sony. After Sony Pictures Entertainment announced that it was pulling "The Interview" from theaters, the studio announced it was also pulling "The Dictator" and "The Big Bang Theory" from streaming and DVD.

The bill was introduced after a series of leaks of personal information, including a breach of Sony Pictures' database that exposed the private information of over 100 million users.

The Senate's top Democrat, Sen. Chuck Schumer Chuck Schumer, has introduced a bill that would require broadband providers to obtain customer consent before collecting data on their online activities, and would ban the sale of that data without consent.

The bill is aimed at the FCC rules, which he says have not kept pace with the digital revolution.

"The FCC's rules are woefully outdated," said Sen. Ed Markey Edward (Ed) John Markey. "The internet has changed so quickly that the FCC has struggled to keep up."

The bill would require broadband providers to obtain customer consent before collecting data on their online activities, including the websites people visit, the time spent on them and the new, highly-anticipated movie."The Interview," has been cancelled.

The studio's CEO, Jim Gianopulos, said the decision was "not the right decision" for the company. "We stand by our filmmakers and their right to tell their stories," he said.

It's been a rough few weeks for Sony. After Sony Pictures Entertainment announced that it was pulling "The Interview" from theaters, the studio announced it was also pulling "The Dictator" and "The Big Bang Theory" from streaming and DVD.

The bill was introduced after a series of leaks of personal information, including a breach of Sony Pictures' database that exposed the private information of over 100 million users.

The Senate's top Democrat, Sen. Chuck Schumer Chuck Schumer, has introduced a bill that would require broadband providers to obtain customer consent before collecting data on their online activities, and would ban the sale of that data without consent.

The bill is aimed at the FCC rules, which he says have not kept pace with the digital revolution.

"The FCC's rules are woefully outdated," said Sen. Ed Markey Edward (Ed) John Markey. "The internet has changed so quickly that the FCC has struggled to keep up."

The bill would require broadband providers to obtain customer consent before collecting data on their online activities, including the websites people visit, the time spent on them and the new, highly-anticipated movie.

"The Interview," has been cancelled. The studio's CEO, Jim Gianopulos, said the decision was "not the right decision" for the company. "We stand by our filmmakers and their right to tell their stories," he said.

It's been a rough few weeks for Sony. After Sony Pictures Entertainment announced that it was pulling "The Interview" from theaters, the studio announced it was also pulling "The Dictator" and "The Big Bang Theory" from streaming and DVD.

The bill was introduced after a series of leaks of personal information, including a breach of Sony Pictures' database that exposed the private information of over 100 million users. The new, highly-anticipated movie, "The Interview," has been cancelled.
Likelihood ratios predict memorized text
<table>
<thead>
<tr>
<th>Inference Strategy</th>
<th>Text Generation Strategy</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Top-$n$</td>
</tr>
<tr>
<td>Perplexity</td>
<td>9</td>
</tr>
<tr>
<td>Small</td>
<td>41</td>
</tr>
<tr>
<td>Medium</td>
<td>38</td>
</tr>
<tr>
<td>zlib</td>
<td>59</td>
</tr>
<tr>
<td>Window</td>
<td>33</td>
</tr>
<tr>
<td>Lowercase</td>
<td>53</td>
</tr>
<tr>
<td><strong>Total Unique</strong></td>
<td><strong>191</strong></td>
</tr>
<tr>
<td>URL (trimmed)</td>
<td>Occurrences</td>
</tr>
<tr>
<td>--------------------------------------------</td>
<td>-------------</td>
</tr>
<tr>
<td>/r/51y/milo_evacuation</td>
<td>1</td>
</tr>
<tr>
<td>/r/zin/hi_my_name</td>
<td>1</td>
</tr>
<tr>
<td>/r/7ne/for_all_yo</td>
<td>1</td>
</tr>
<tr>
<td>/r/5mj/fake_news</td>
<td>1</td>
</tr>
<tr>
<td>/r/5wn/reddit_admin</td>
<td>1</td>
</tr>
<tr>
<td>/r/lp8/26_evening</td>
<td>1</td>
</tr>
<tr>
<td>/r/jla/so_pizzagat</td>
<td>1</td>
</tr>
<tr>
<td>/r/subf/late_night</td>
<td>1</td>
</tr>
<tr>
<td>/r/eta/make_christ</td>
<td>1</td>
</tr>
<tr>
<td>/r/6ev/its_officia</td>
<td>1</td>
</tr>
<tr>
<td>/r/3c7/scott_adams</td>
<td>1</td>
</tr>
<tr>
<td>/r/k2o/because_his</td>
<td>1</td>
</tr>
<tr>
<td>/r/tu3/armynavy_ga</td>
<td>1</td>
</tr>
</tbody>
</table>
This talk:
1. Does this happen?
2. Can we prevent it?
This talk:
1. Does this happen?
2. Can we prevent it?
Act II:
InstaHide
Because the dataset itself is private
Any learning algorithm, when run on it, is private too
2020 Bell Labs Prize

Winners

1st: Firooz Aflatouni
University of Pennsylvania

2nd: Sanjeev Arora
Princeton University

3rd: Cheng Qi
Georgia Institute of Technology
InstaHide Algorithm

Private

Public
InstaHide Algorithm

Private

Public
InstaHide Algorithm
InstaHide Algorithm

Private

Public
InstaHide Algorithm

Private

Public
InstaHide Algorithm

Private

Public

[Images of various shapes and patterns]
Is InstaHide Private?
Our attack, in two steps
\( F(\cdot, \cdot) = \)
\[ F(\text{image}_1, \text{image}_2) = 0 \]
$F(\cdot, \cdot) = 1$
Step 1: Cluster
Step 2: Recover
Step 2: Recover
Step 2: Recover
Results
Act III:
Differential Privacy
A learning algorithm is **differentially private** if the probability that

(1) any adversary can win this game
(2) on any dataset
(3) for any differing example

is less than a given threshold
is less than a given threshold
is less than a given threshold value of interest: epsilon

\[
\begin{align*}
\epsilon = 0 & \quad \Rightarrow \quad p(\text{win}) = 0 \\
\epsilon = \infty & \quad \Rightarrow \quad p(\text{win}) = 1 \\
\epsilon = 1 & \quad \Rightarrow \quad p(\text{win}) = 0.001 \\
\epsilon = 2 & \quad \Rightarrow \quad p(\text{win}) = 0.01 \\
\epsilon = 3 & \quad \Rightarrow \quad p(\text{win}) = 1
\end{align*}
\]
Amount of Privacy

Upper-Bound Guarantee
(by Differential Privacy)

Reality
(Actual Amount of Memorization)
Amount of Privacy

Upper-Bound Guarantee (by Differential Privacy)

Reality (Actual Amount of Memorization)

Lower Bound (... somehow)

This section
Our approach: Adversary Instigation to give a lower bound
Why do this?
1. Tightness of differential privacy bound
2. Analysis under different settings
3. Testing new potential assumptions
4. Empirically validating correctness
How *tight* is the bound offered by differential privacy?
1. Tightness of differential privacy bound
2. Analysis under different settings
3. Testing new potential assumptions
4. Empirically validating correctness
What if I only cared about the privacy of a *typical* dataset?

Differential privacy can't give any analysis for this.
1. Tightness of differential privacy bound
2. Analysis under different settings
3. Testing new potential assumptions
4. Empirically validating correctness
What if I introduced a new assumption?
How much better could the analysis get?
1. Tightness of differential privacy bound
2. Analysis under different settings
3. Testing new potential assumptions
4. Empirically validating correctness
Beware of bugs in the above code;
I have only proved it correct, not tried it.

- Knuth
Adversary Instantiation
Let's actually build the adversary!
Or, those (cooperating) adversaries ... 

**Crafter:**
creates the two datasets A and B

**Distinguisher:**
guesses if model is from A or B
Craft datasets A and B

Model Trainer

Chooses either A or B
Train model F on that dataset

Distinguisher guesses if F was trained on A or B
Craft datasets A and B

Choose either A or B
Train model F on that dataset

Distinguisher guesses if F was trained on A or B
Craft datasets A and B

Adversary

Chooses either A or B
Train model F on that dataset

Model Trainer

Distinguisher guesses if F was trained on A or B
Craft datasets A and B

Adversary

Model Trainer

Craft datasets A and B

Adversary

Model Trainer

Chooses either A or B

Train model F on that dataset

Distinguisher guesses if F was trained on A or B

Adversary

Model Trainer

Craft datasets A and B

Adversary

Model Trainer

Chooses either A or B

Train model F on that dataset

Distinguisher guesses if F was trained on A or B

Adversary

Model Trainer

Craft datasets A and B

Adversary

Model Trainer

Chooses either A or B

Train model F on that dataset

Distinguisher guesses if F was trained on A or B

Adversary

Model Trainer

Craft datasets A and B

Adversary

Model Trainer

Chooses either A or B

Train model F on that dataset

Distinguisher guesses if F was trained on A or B

Adversary

Model Trainer

Craft datasets A and B

Adversary

Model Trainer

Chooses either A or B

Train model F on that dataset

Distinguisher guesses if F was trained on A or B
Craft datasets A and B
Chooses either A or B
Train model F on that dataset
Distinguisher guesses if F was trained on A or B

Successes: 1095
Failures: 26
Attack 1: Membership Inference
Crafter

Dataset A

Dataset B

+ 7
Distinguisher

\[ \max f(7) \]
Distinguisher

\[ \max \ f(7) = .95 \]
Distinguisher

\[ \max f(7) = 0.9999 \]
Attack 2: Worst-case Example
Crafter

Dataset A

Dataset B

+
Distinguisher

$$\max f(\hat{7}) = 0.15$$
Distinguisher

\[
\max \ f(\hat{7}) = .8
\]
Attack 3: Intermediate Model Access
Craft datasets A and B

Distinguisher guesses if F was trained on A or B

Model Trainer

(A,B)

Chooses either A or B
Train model F on that dataset

F
Craft datasets A and B

Adversary

Chooses either A or B
Train model F on that dataset

(A,B)

Distinguisher guesses if F was trained on A or B

Model Trainer

F₁, F₂, F₃, F₄ ...
Distinguisher

\[
\begin{align*}
\max f_1(\hat{\gamma}) &= .45 \\
\max f_2(\hat{\gamma}) &= .17 \\
\max f_3(\hat{\gamma}) &= .22
\end{align*}
\]
Distinguisher

\[
\begin{align*}
\max f_1(x) &= .82 \\
\max f_2(x) &= .55 \\
\max f_3(x) &= .72
\end{align*}
\]
Craft datasets A and B

Adversary

Model Trainer

(A,B)

Chooses either A or B
Train model F on that dataset

F₁, F₂,
F₃, F₄ ...

Distinguisher guesses if F was trained on A or B
Intermediate Model Adversary Game

\[
\begin{align*}
(D, D') & \xleftarrow{\$} A \\
B &= (D, D') \\
b & \xleftarrow{\$} \{0, 1\} \\
\{f_{\theta_i}\}_{i=1}^N & \leftarrow \mathcal{T}(B_b) \\
\{f_{\theta_1}, f_{\theta_2}, \ldots, f_{\theta_N}\} & \leftarrow B(\{f_{\theta_i}\}, D, D') \\
s & \in \{0, 1\} \\
\text{Check if } s = b
\end{align*}
\]
Adaptive Input Poisoning Adversary Game

Adversary

\[(x, y) \overset{\$}{\leftarrow} A_{query} \]

\[b \overset{\$}{\leftarrow} \{0, 1\} \]

\[f_{\theta_0} \]

\[(D, D') \overset{\$}{\leftarrow} A(f_{\theta_i}, x, y) \]

\[B^i = \{D, D'\} \]

\[\theta_{i+1} \leftarrow S(B_b^i, x, y, \theta_i) \]

\[f_{\theta_{i+1}} \]

\[s \leftarrow B(\{f_{\theta_i}\}, D, D') \]

\[s \in \{0, 1\} \]

Check if \(s = b\)

Model Trainer

Repeat \(N\) Iterations
Gradient Poisoning Adversary Game

Adversary

\[ b \leftarrow \{0, 1\} \]

\[ f_{\theta_0} \]

\[ G, G' \leftarrow A(f_{\theta_i}) \]

\[ B^i = \{G', G\} \]

\[ B \leftarrow B(B_b^i) \]

\[ B_{dp} \leftarrow \{\text{clip}_p(x) + N : x \in B\} \]

\[ \theta_{i+1} \leftarrow \theta_i - \eta \sum_{x \in B_{dp}} x \]

\[ f_{\theta_{i+1}} \]

\[ \vdots \]

\[ s \leftarrow B(\{f_{\theta_i}\}, G) \]

\[ s \in \{0, 1\} \]

Check if \( s = b \)

Model Trainer

Repeat \( N \) Iterations
Two important conclusions:

(1) DP-SGD is tight in the worst case
(2) But could (maybe) be made stronger with more realistic assumptions
Two important conclusions:

(1) DP-SGD is tight in the worst case
(2) But could (maybe) be made stronger with more realistic assumptions