

# Defending against Compromise and Manipulation of Mobile Communities

**BIMAL VISWANATH, UNIVERSITY OF CHICAGO**

# Creating and defending against fake reviews using deep learning

**Bimal Viswanath**  
*University of Chicago*

Cyber Security Research Acceleration Workshop and Showcase  
Indianapolis, October 2017

# Crowdsourced attacks on review systems

- Many review systems are plagued by **fake reviews**
  - Vulnerable platforms include Yelp, Amazon, and TripAdvisor



Attacker

**Poké Bar** Claimed

★★★★☆ 337 reviews [Details](#) ★ Write a Review

\$\$ - Seafood, Hawaiian [Edit](#)

 **John**  
4 friends  
56 reviews

★★★★☆ 7/24/2017  
It is good for quick food. A little on the pricy side for what you get but overall satisfactory. Process was easy and fast, it's just I was clean and fast but doesn't wow too much.

 **Anna L.**  
207 friends  
121 reviews  
2 photos  
Elite '17

★★★★☆ 5/6/2017  
Great selection of fish, toppings and sauces! I got the salmon, crab, tuna, ginger, brown rice, cucumbers, onion corn, and ponzu sauce! Super fresh, quick and easy :)  
  
\*House dressing is soy sauce and sesame oil.

 **Steph**  
25 friends  
54 reviews  
28 photos  
Elite '17

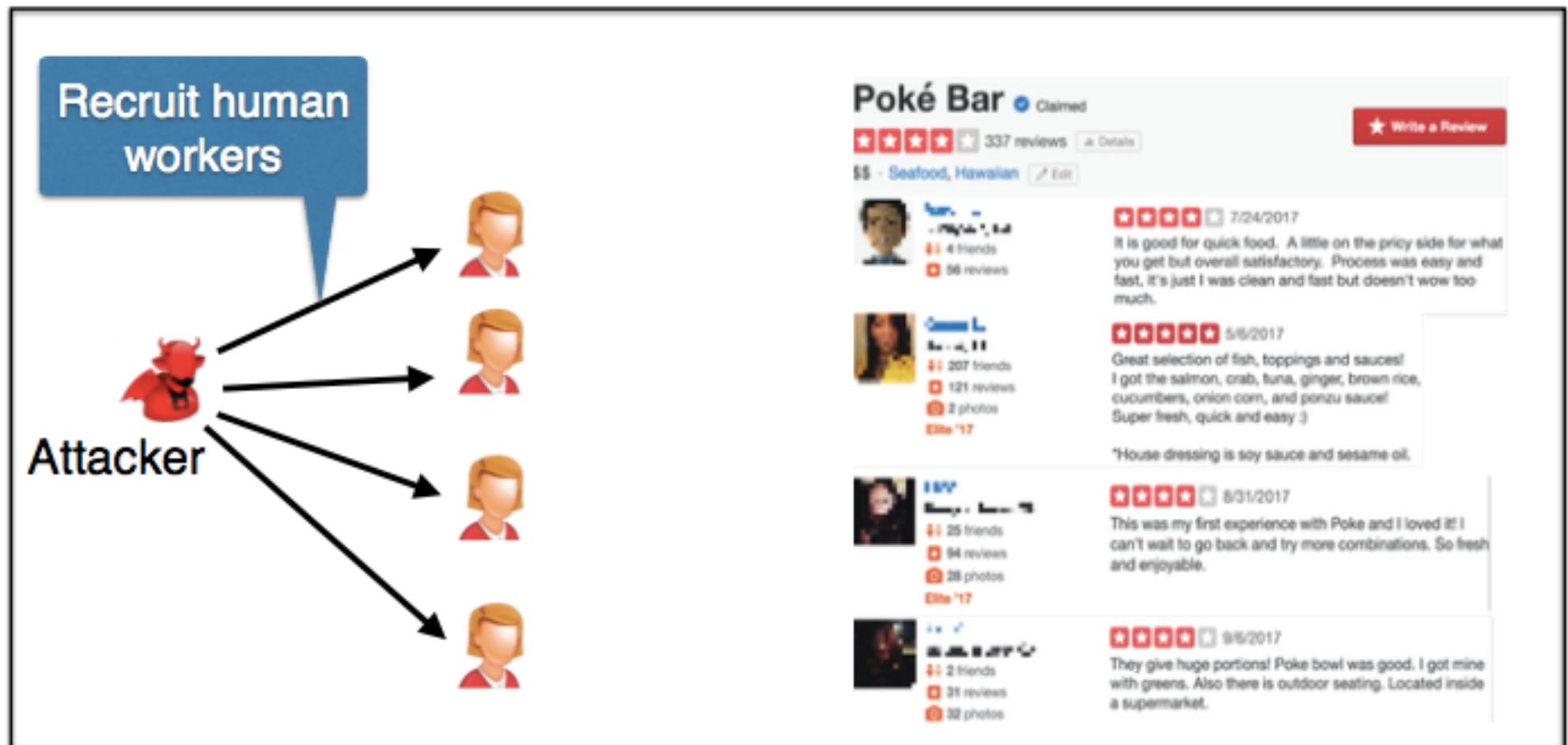
★★★★☆ 8/31/2017  
This was my first experience with Poke and I loved it! I can't wait to go back and try more combinations. So fresh and enjoyable.

 **...**  
2 friends  
31 reviews  
32 photos

★★★★☆ 9/5/2017  
They give huge portions! Poke bowl was good. I got mine with greens. Also there is outdoor seating. Located inside a supermarket.

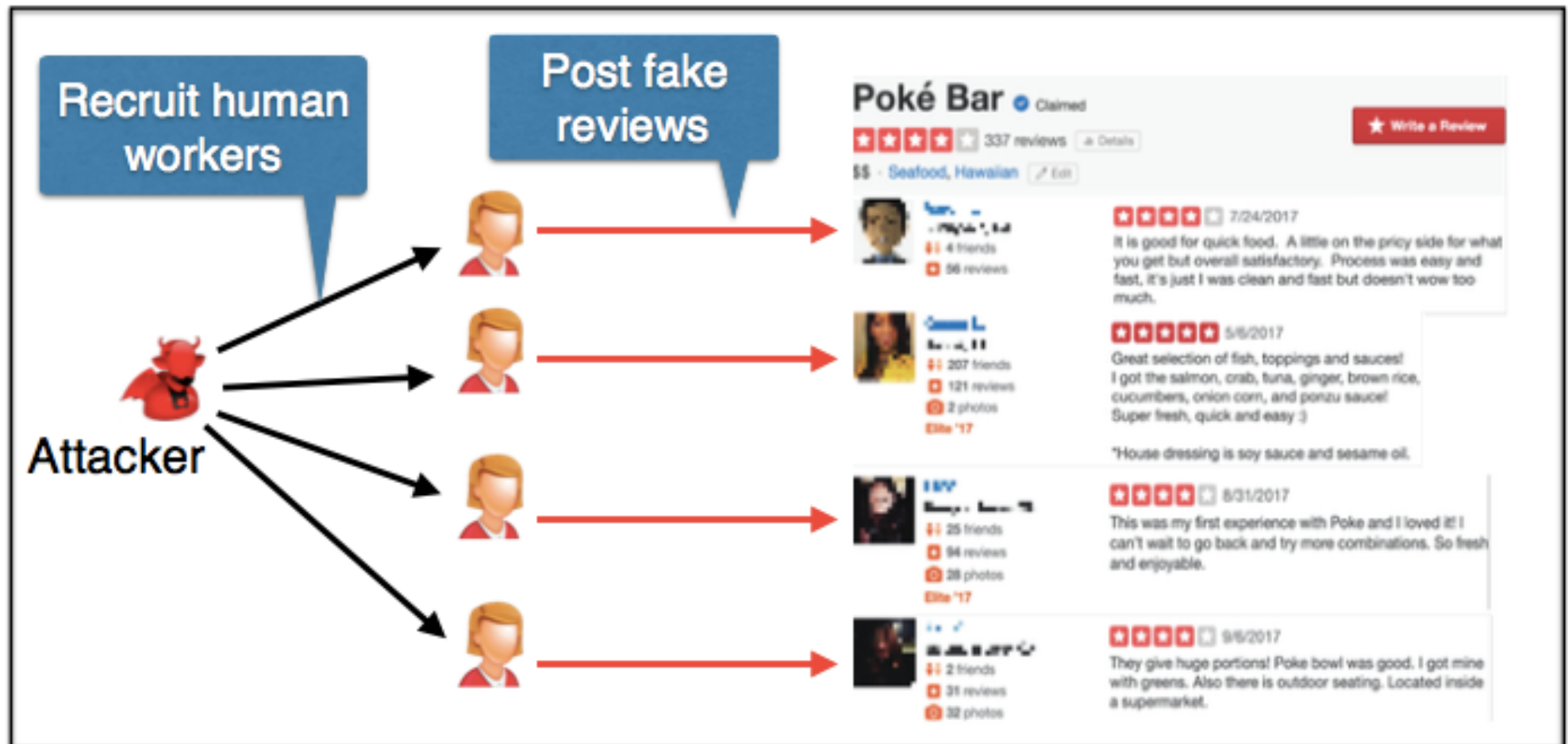
# Crowdsourced attacks on review systems

- Many review systems are plagued by **fake reviews**
  - Vulnerable platforms include Yelp, Amazon, and TripAdvisor



# Crowdsourced attacks on review systems

- Many review systems are plagued by **fake reviews**
  - Vulnerable platforms include Yelp, Amazon, and TripAdvisor



# What if attack used an AI program?

- Assumption: AI can generate meaningful reviews



Attacker

**Poké Bar** Claimed

★★★★☆ 337 reviews [Details](#) [Write a Review](#)

\$\$ - Seafood, Hawaiian [Edit](#)

 **John**  
4 friends  
56 reviews  
7/24/2017  
★★★★☆  
It is good for quick food. A little on the pricy side for what you get but overall satisfactory. Process was easy and fast, it's just I was clean and fast but doesn't wow too much.

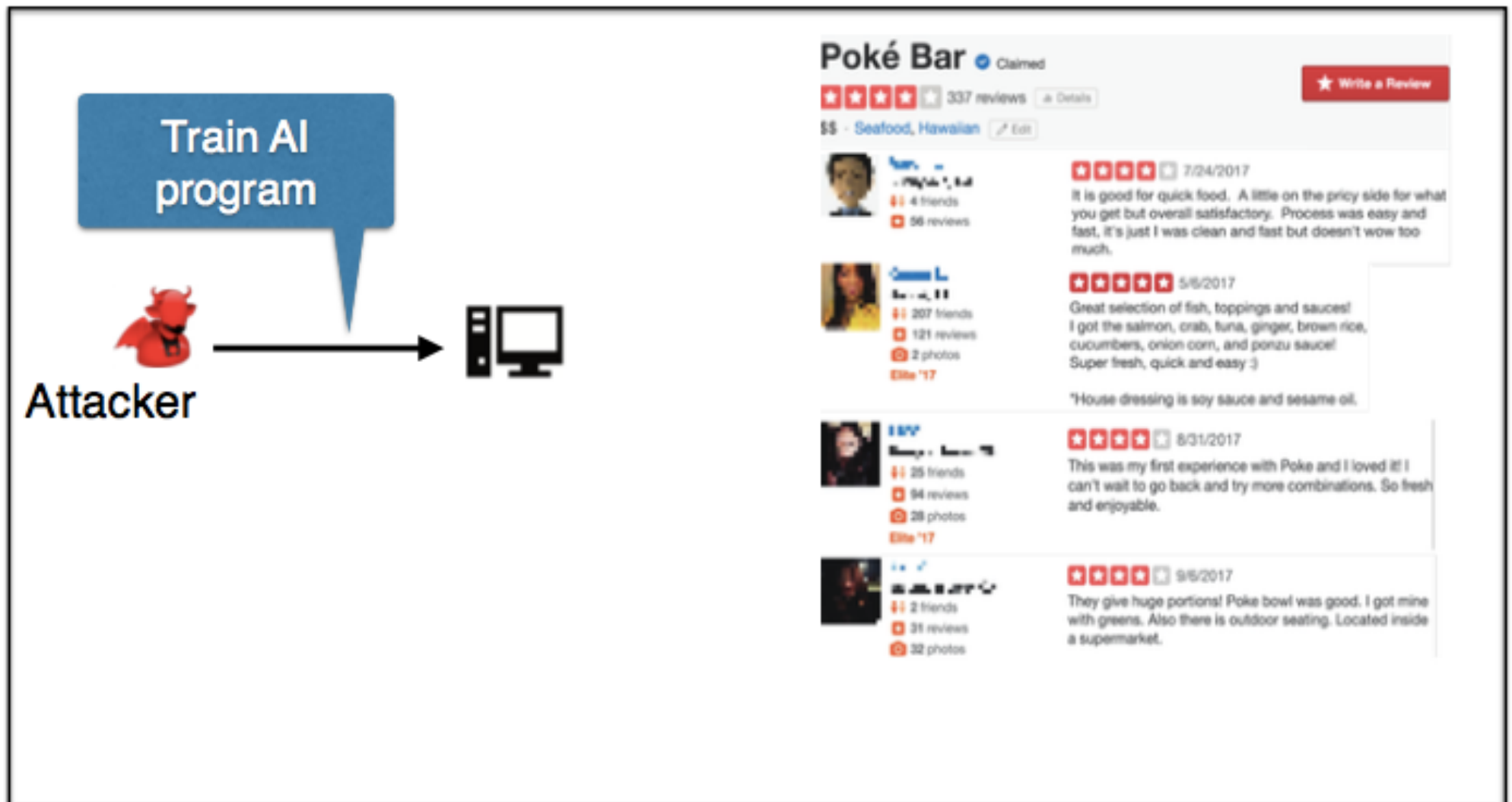
 **Anna L.**  
207 friends  
121 reviews  
2 photos  
Elite '17  
5/6/2017  
★★★★★  
Great selection of fish, toppings and sauces! I got the salmon, crab, tuna, ginger, brown rice, cucumbers, onion corn, and ponzu sauce! Super fresh, quick and easy :)  
\*House dressing is soy sauce and sesame oil.

 **Emily**  
25 friends  
94 reviews  
28 photos  
Elite '17  
8/31/2017  
★★★★☆  
This was my first experience with Poke and I loved it! I can't wait to go back and try more combinations. So fresh and enjoyable.

 **John**  
2 friends  
31 reviews  
32 photos  
9/6/2017  
★★★★☆  
They give huge portions! Poke bowl was good. I got mine with greens. Also there is outdoor seating. Located inside a supermarket.

# What if attack used an AI program?

- Assumption: AI can generate meaningful reviews



# What if attack used an AI program?

- Assumption: AI can generate meaningful reviews





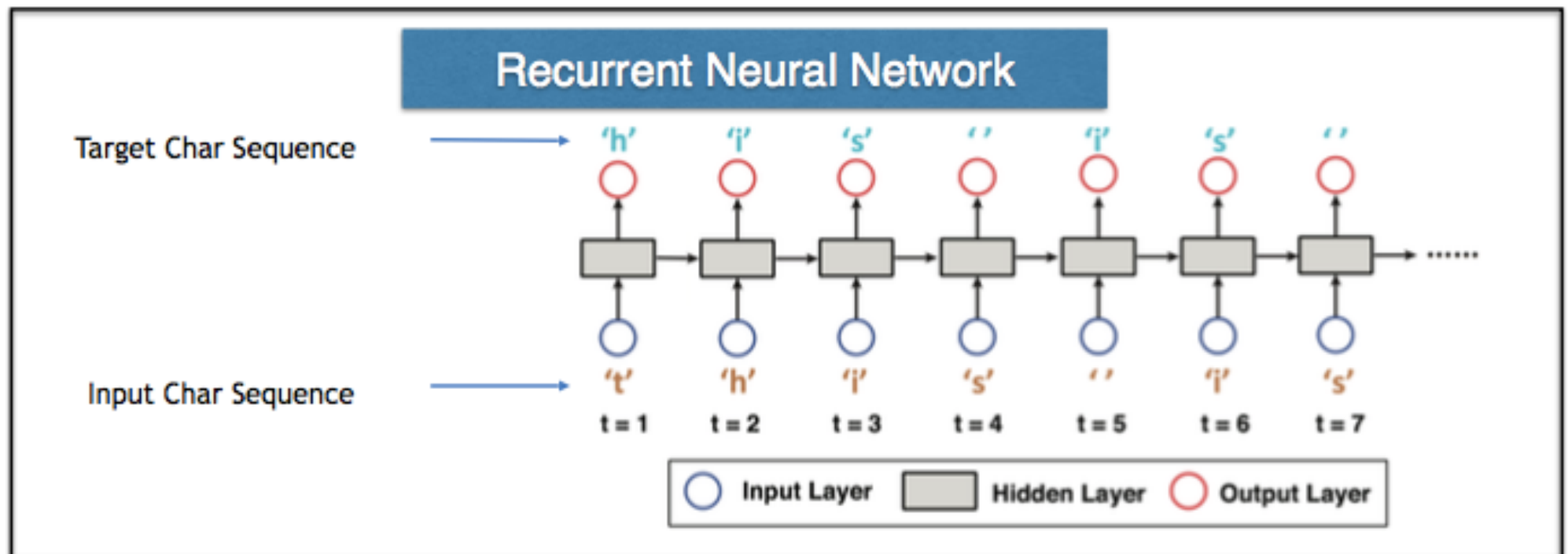
# What if attack used an AI program?

- Assumption: AI can generate meaningful reviews



# Can AI generate meaningful text?

- Recurrent Neural Networks (RNNs) are capable
  - Can generate meaningful short pieces of text
  - Sufficient for application domains such as review systems!
- RNN predicts the next character in a seq. from prior seq.



# Samples of RNN generated reviews

- **5 star review**

*The food here is freaking amazing, the portions are giant. The cheese bagel was cooked to perfection and well prepared, fresh & delicious! The service was fast. Our favorite spot for sure! We will be back!*

- **3 star review**

*The food wasn't bad. The cupcakes are okay and the service is excellent but the prices are a bit high. I do like the fresh made salad and drink specials. I would recommend this place for a place to grab a bite for a couple of times*

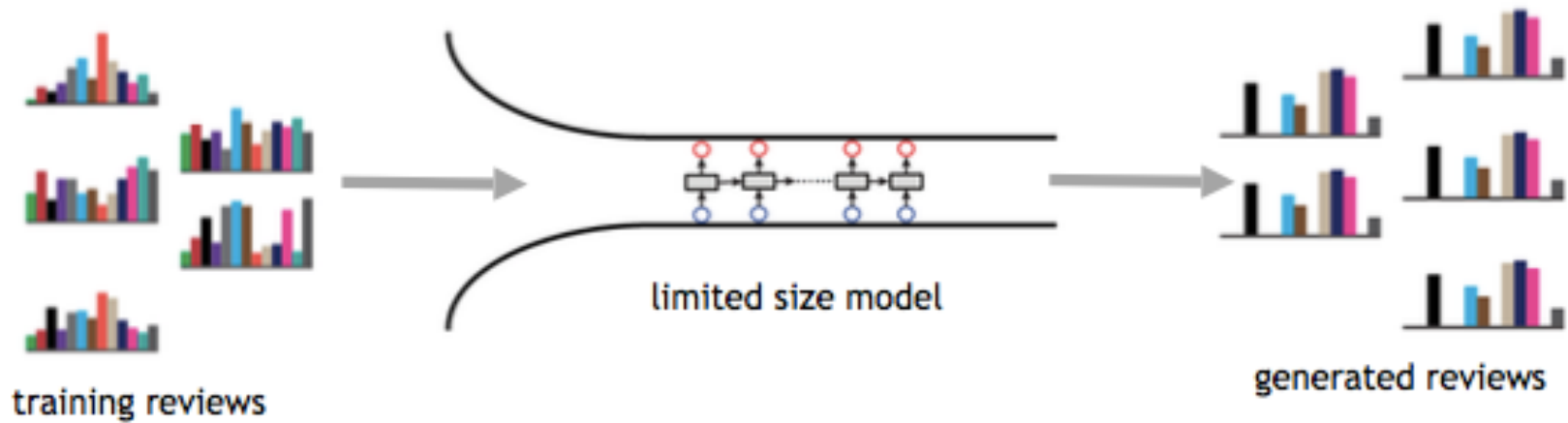
- **1 star review**

*DO NOT WASTE YOUR TIME AND MONEY! The absolute worst service I have ever experienced. This place is a joke. The waitress was rude and said she would put the manager to come out but never happened. I wish I could give zero star.*

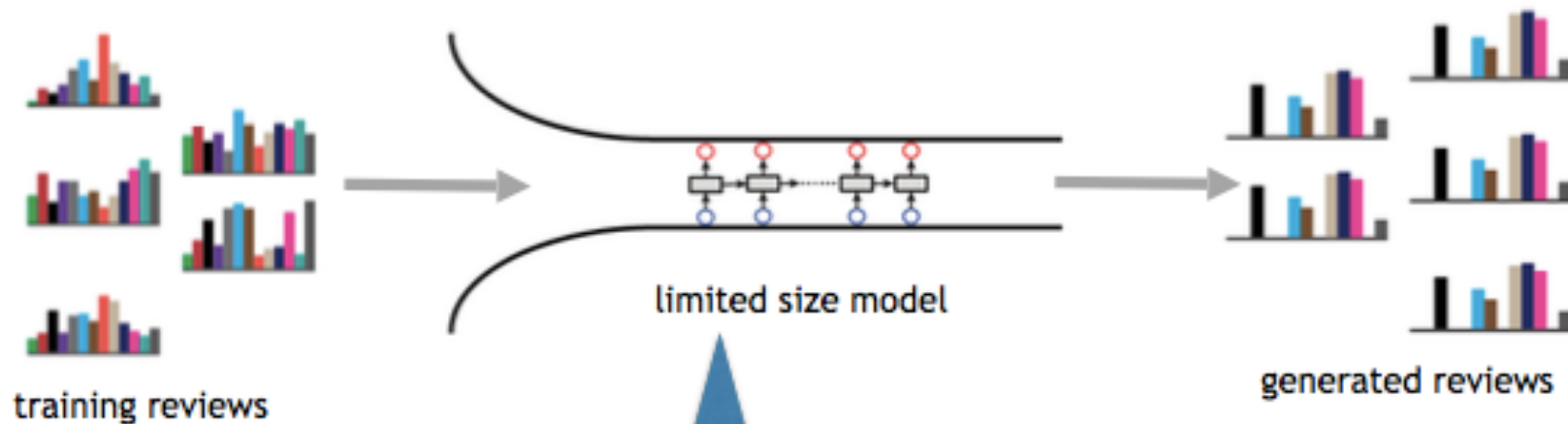
# How effective is the attack?

- Can existing ML algorithms detect machine-generated reviews?
  - Existing ML-based fake review detection algorithms perform poorly
  - Detection performance: Precision of 18% and Recall of 58%
- Can humans detect machine-generated reviews?
  - Human evaluators also show poor detection performance
  - Detection performance: Precision of 41% and Recall of 16%
- Did humans find the machine-generated reviews 'useful'?
  - Average usefulness score of machine reviews (3.15 out of 5) is close to real reviews (3.28 out of 5)

# Proposed defense

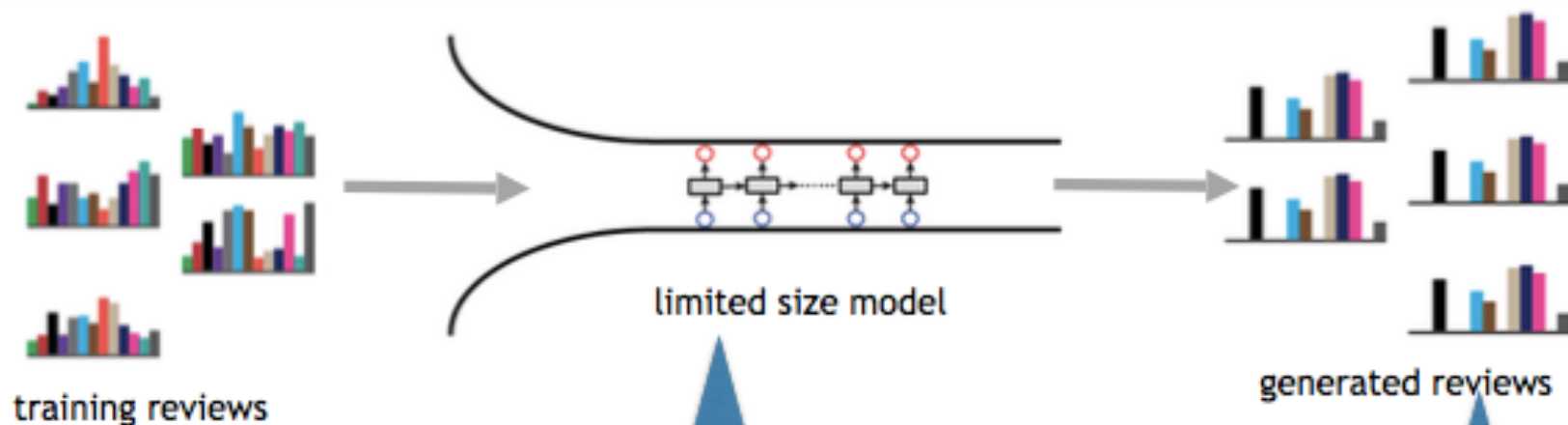


# Proposed defense



RNN is a fixed  
memory  
representation of the  
training dataset

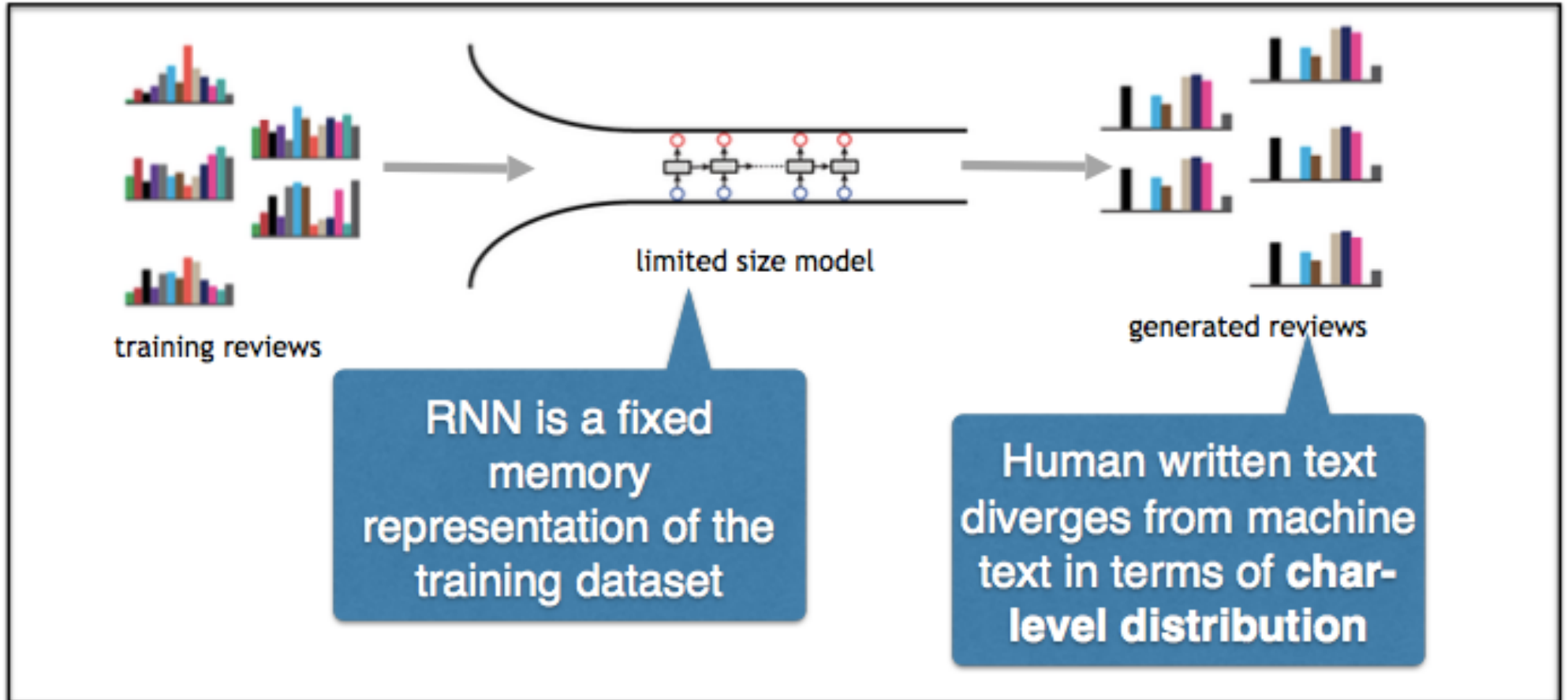
# Proposed defense



RNN is a fixed  
memory  
representation of the  
training dataset

Human written text  
diverges from machine  
text in terms of **char-  
level distribution**

# Proposed defense



- Built a scheme that can detect such char-level divergence
- Detection performance: Precision of 98% and Recall of 97%



# Questions?



Paper to appear in CCS'17

BUSINESS  
INSIDER

FORTUNE

Forbes



Press coverage

NEW  
YORK  
POST

engadget

THE VERGE