Next Generation of Fraud Detection

Sihong Xie and Philip S. Yu

THE UNIVERSITY OF ILLINOIS AT CHICAGO

LEHIGH UNIVERSITY

IEEE CIC. Oct 19, 2018. Philadelphia, PA

Frauds:

Wrongful or criminal deception intendedto result in financial or personal gain

- **Review spams**
- **Return frauds (Amazon, Costco, other retailers)**
- **Search spams (click farms)**
- **Fake news (Facebook and Twitter)**

Stories and statistics

A single couple fraudsters caused 1.2 million loss to Amazon using return fraud [1].

Samsung fined $340,000 for posting fake reviews [2].

1. http://fortune.com/2018/06/05/amazon-tech-scam/
2. https://www.techadvisor.co.uk/feature/tech-industry/taiwans-ftc-investigating-samsung-for-defaming-htc-on-local-online-forums-3442252/

Review frauds (spams)

Local business search

Review frauds:

low quality, biased, and fake reviews from the dishonest brands and third-party SEO.

The challenges

19% Not influenced

Source: https://www.brightlocal.com/learn/local-consumer-review-survey/
based on a pool of representative sample of 1,031 US-based consumers

LEHIGH UNIVERSITY

Create a trustworthy system that spots frauds for social good.

Existing efforts: reviewmeta + spotfake

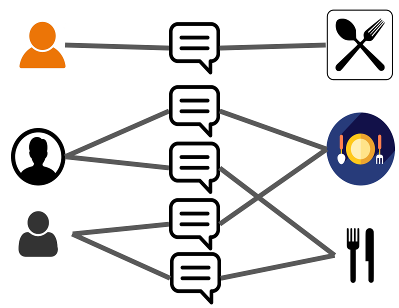
amazon

facebook

tripadvisor

Google

yelp



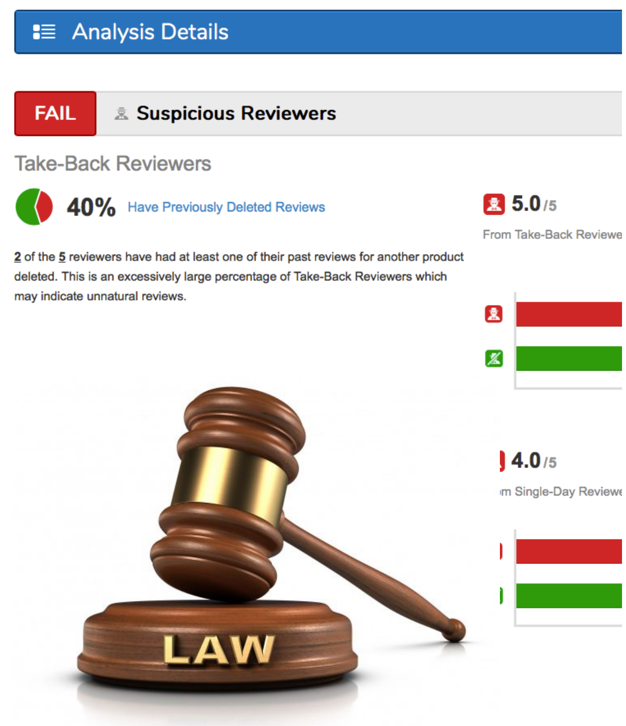
Hand-crafted features:

- two reviews posted in the same time;
- two accounts posted for the same targets;
- two accounts has similar names 1;
- all 5-star reviews;

Classifiers:

decision trees, SVM, logistic regression.

Outcome + Explanations



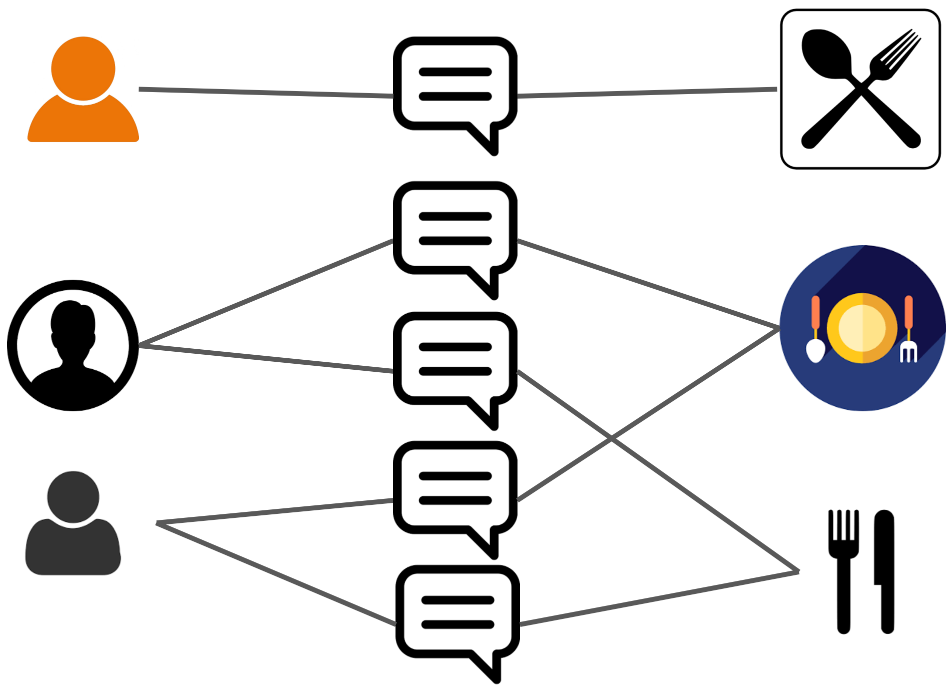
LEHIGH UNIVERSITY

Existing efforts

Independent review fraud detectors

- http://reviewfraud.org
- https://www.fakespot.com
- https://reviewmeta.com

Detection pipeline

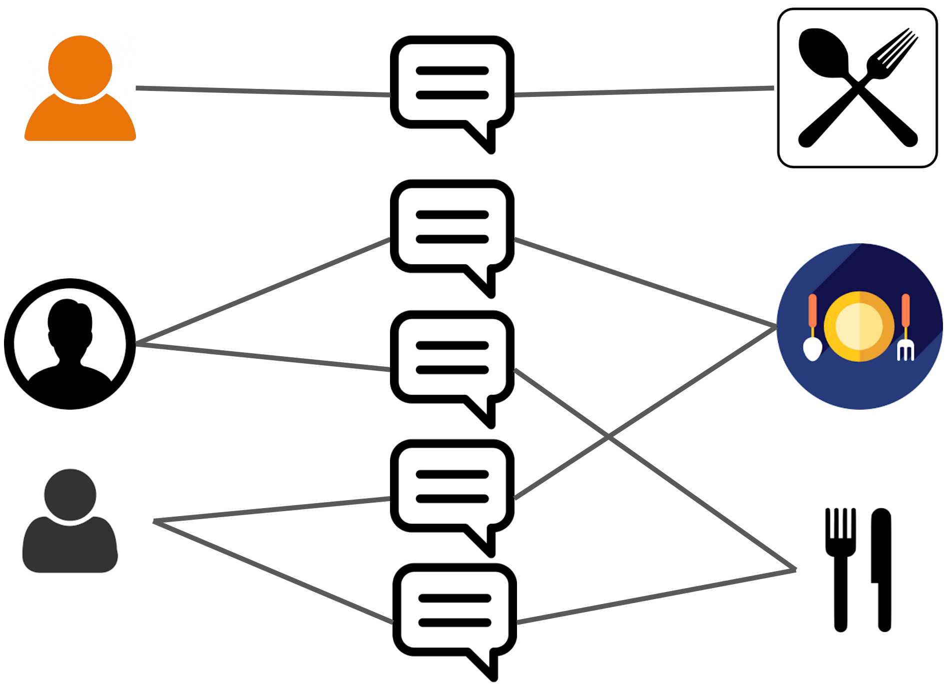


Hand-crafted features:

- two reviews posted in the same time;
- two accounts posted for the same targe
- two accounts has similar names 1;
- all 5-star reviews;
- Singleton reviews;
- near-duplicate review texts;
- near-duplicate images;
- ……

1. based on a true story: http://reviewfraud.org/cloud-9-marketing-aguilar-ventures/

Detection pipeline

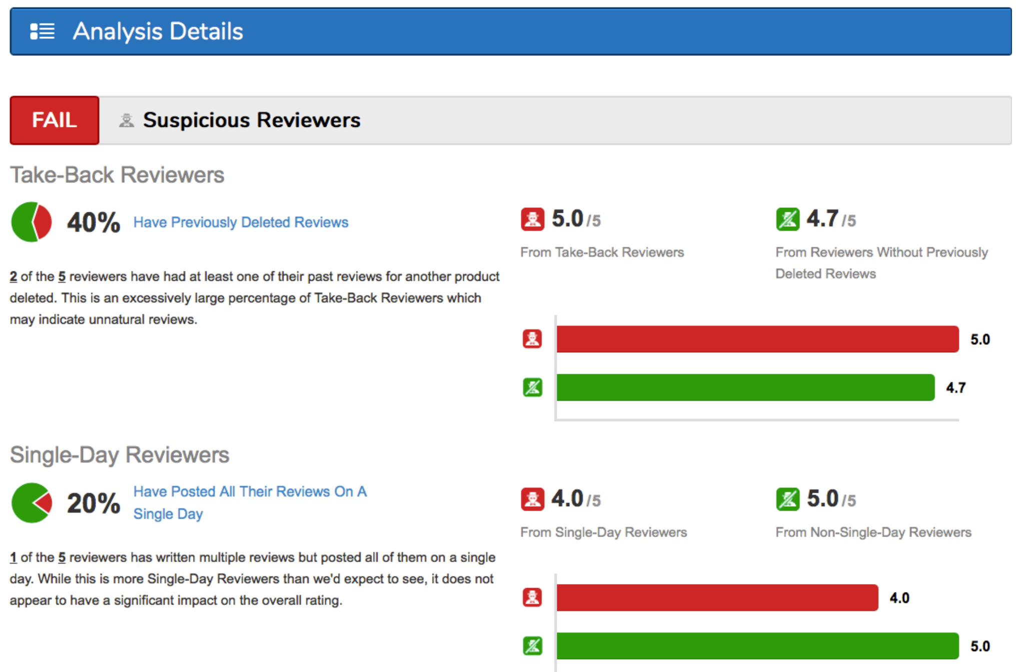


- **Supervised:**
  - decision trees,
  - SVM
  - logistic regression.
- **Unsupervised:**
  - feature histogram
  - graph pattern
  - burst detection
- **rules:**

Detection pipeline

Explain the working and outcomes

- End-users deserve to know the fact;
- To grow trustworthiness among users;
- Developers need to debug the models.



Challenges

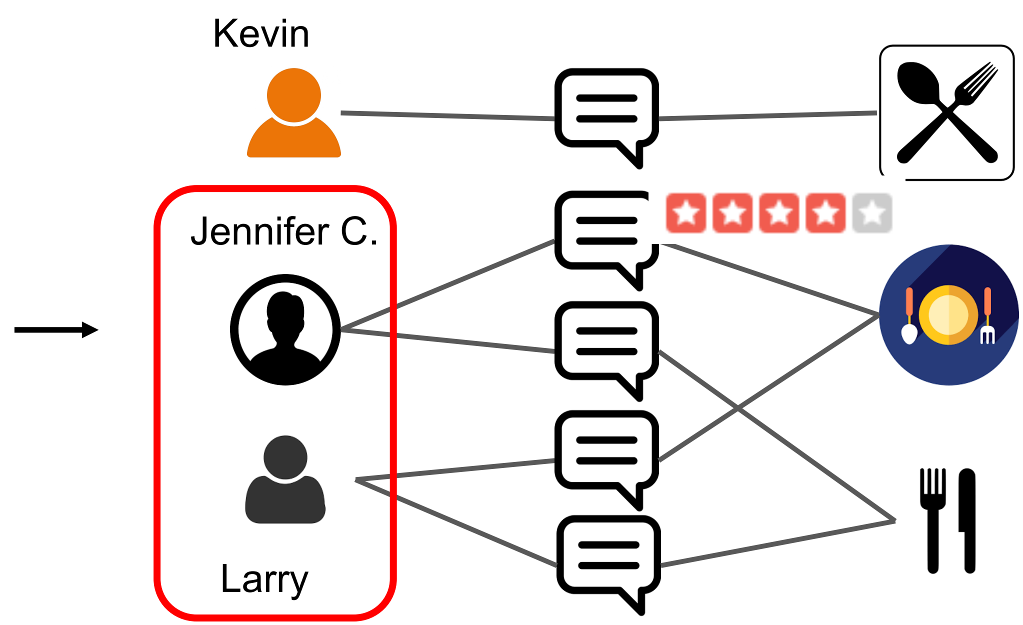
1. **Accuracy vs. Explainability.**

2. Reactive Detection vs. Active Fraudsters.

3. Explainability vs. Security.

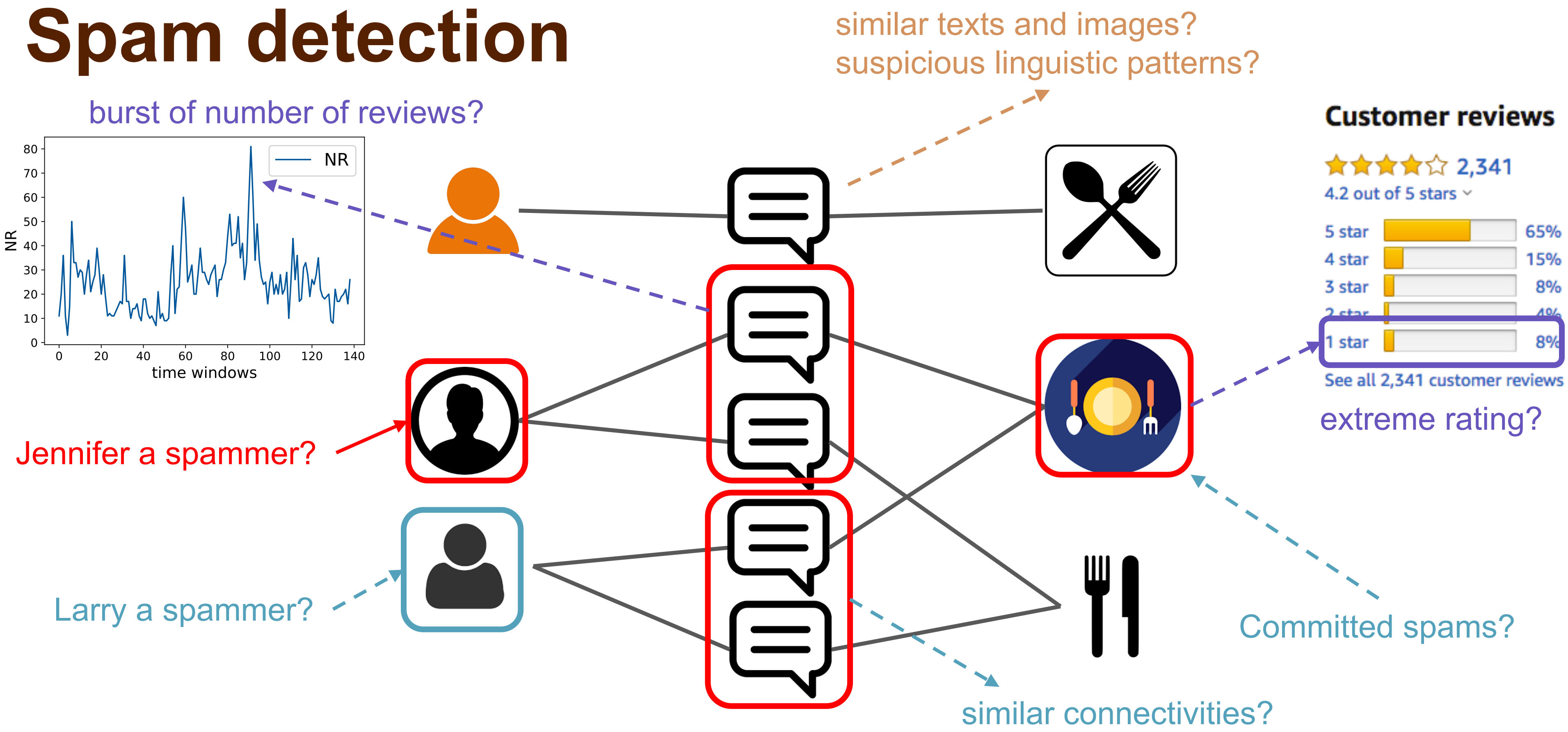
Review data


Pasta Prego
$$ · Italian
1502 Main St
Napa, CA 94559
10/14/2018
This was a cozy and friendly pasta place in Napa. I loved the penne pasta which came with smoked chicken, mozzarella, basil and tomato sauce. Was pretty good although a bit salty. Everything blended in well together.


Kevin
Jennifer C.
Larry

LEHIGH UNIVERSITY

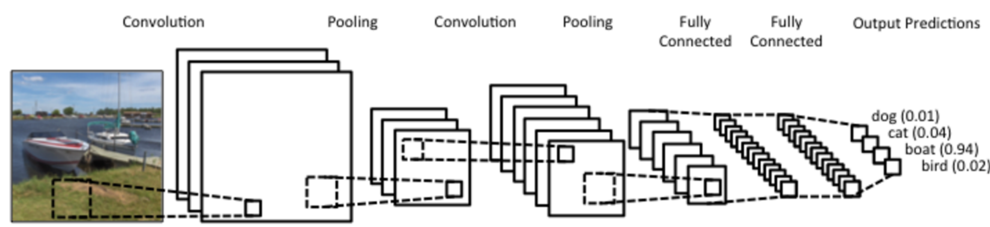
Spam detection

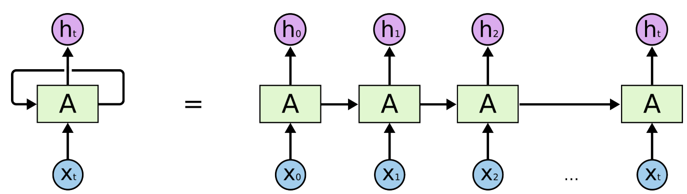


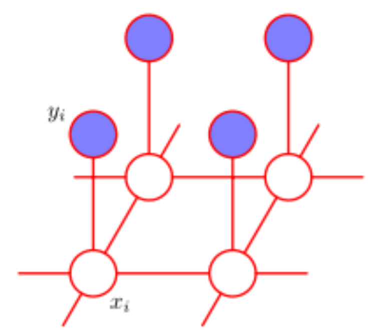
Advanced models are desired

Features that matters:

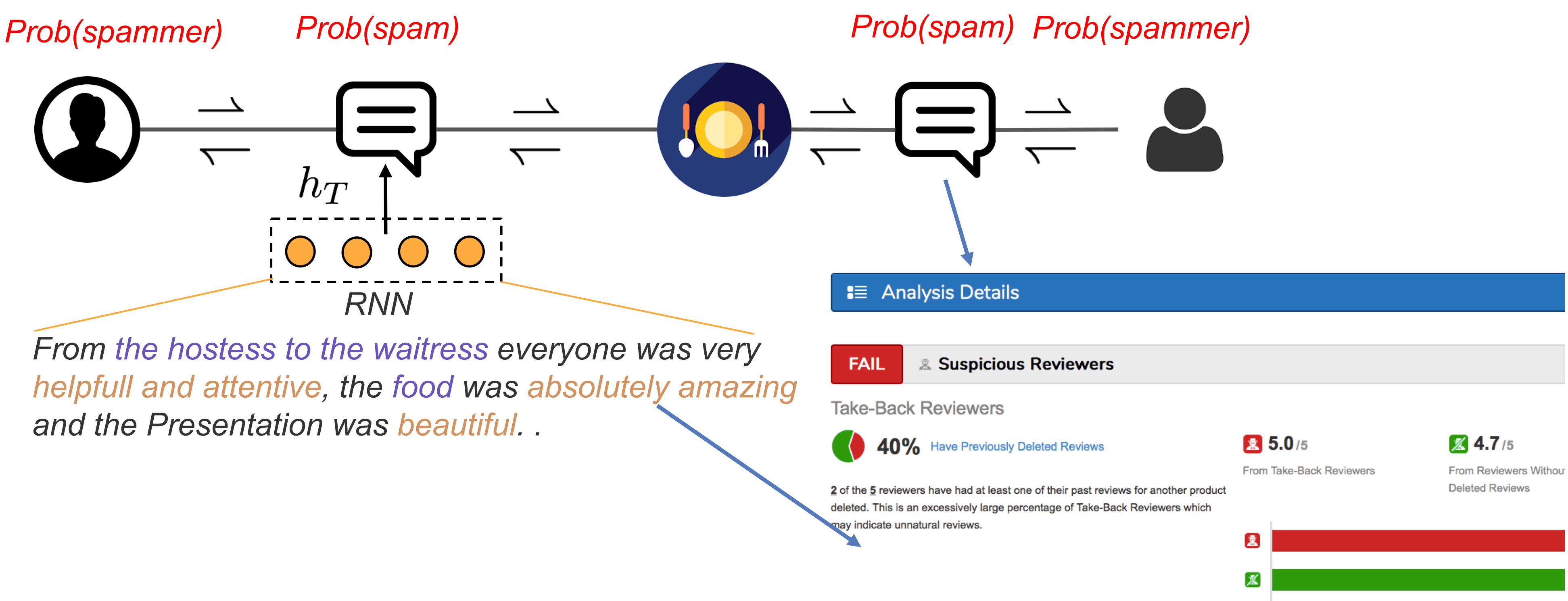
- Text and image similarity;
- Time series patterns;
- Graph connection patterns;







Deep structured prediction



Explaining complex detectors

Multiple sources of supervision

http://reviewfraud.org
https://www.fakespot.com
https://reviewmeta.com

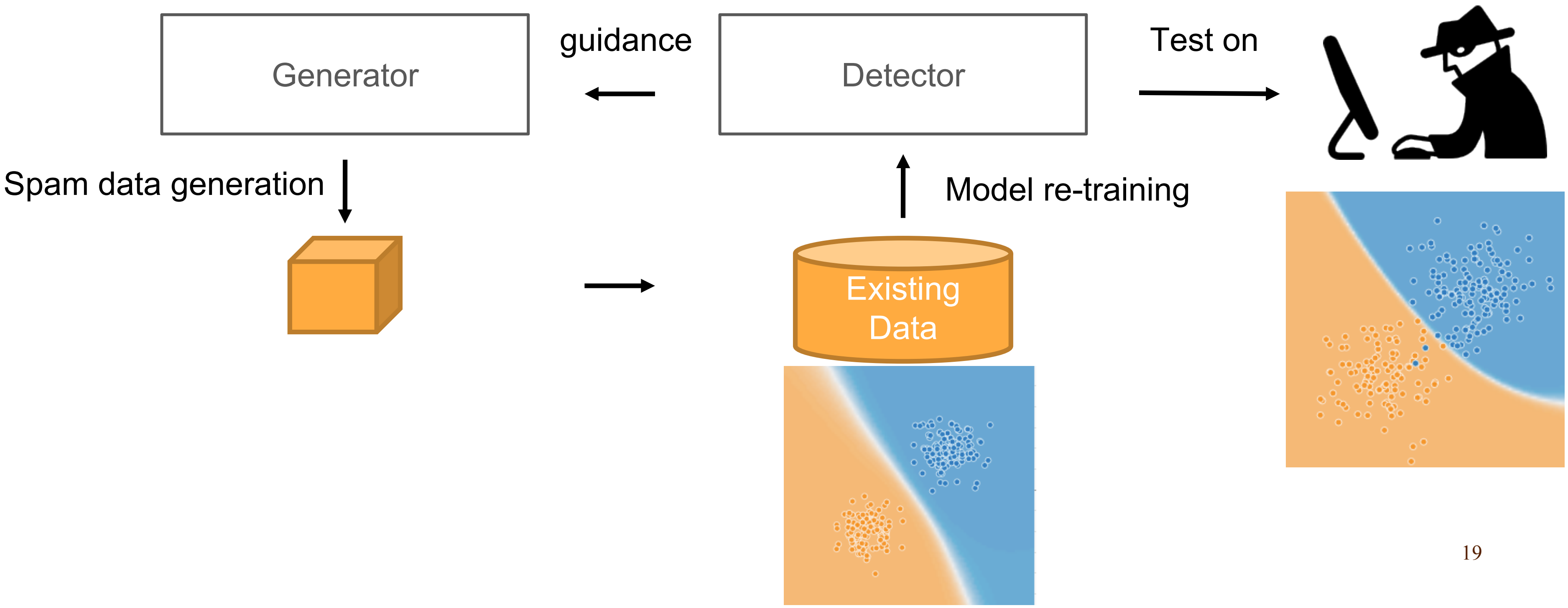
Challenge 2

Dealing with active fraudsters – it is too late when it happens.

Proactive detection is widely deployed in computer softwares and networks, auction networks.

Much more difficult in review fraud detection systems.

Proactive detection via retraining



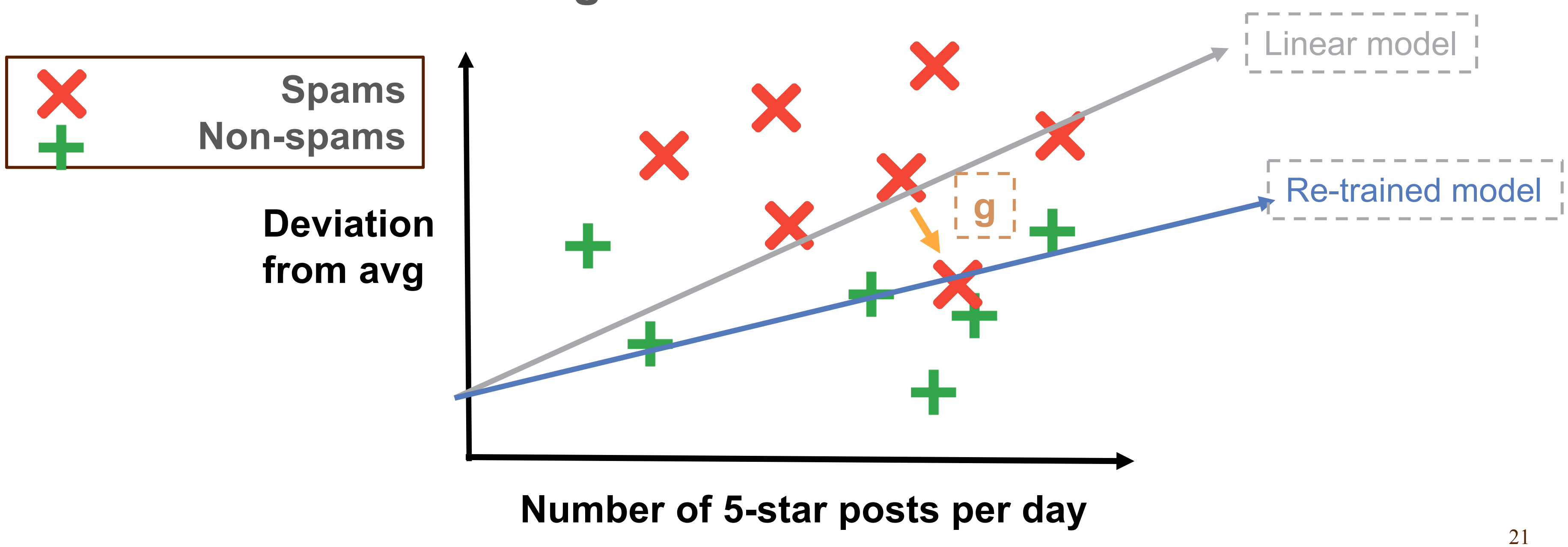
Challenges

1. Accuracy vs. Explainability.
2. **Reactive Detection vs. Active Fraudsters.**
3. Explainability vs. Security.

LEHIGH UNIVERSITY

Challenge 2

Proactive detection via gradient attack.



Generate spams in the input space

Proactive detection via attack simulation.

- When to post a spam?
- Ratings of spams?
- Which account to post a spam?
- What contents to put down in a spam?

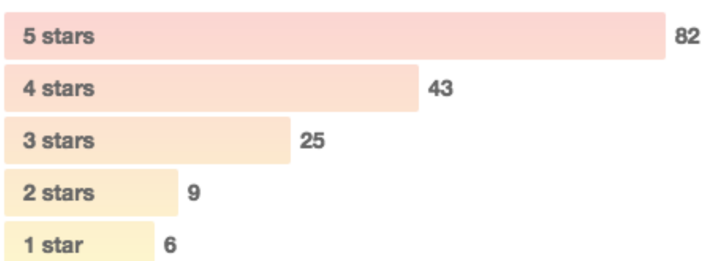
Partial solutions

How to generate spam data?

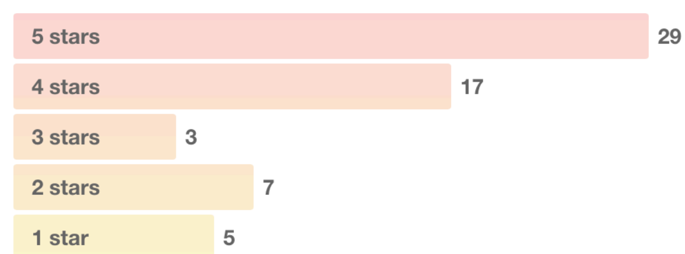
1. Maximum entropy to find the attack rating distribution.
2. Burst-avoiding techniques for attack timing.
3. Graph-based attack.
4. Review text generation.

Find evasive rating distribution

P: spammer target distribution



Q: a normal rating distribution



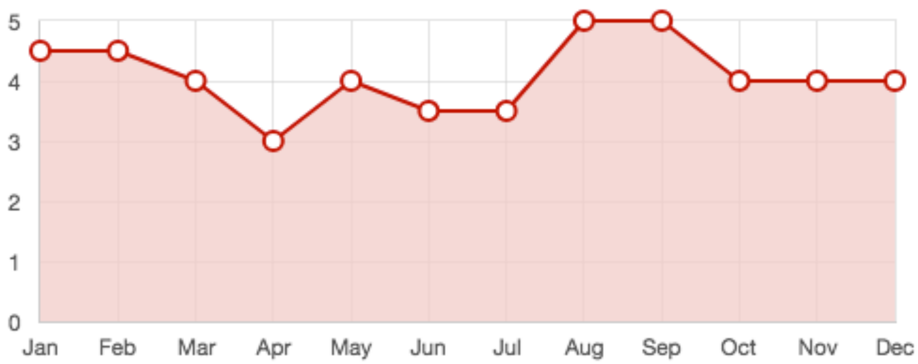
max $_P$ **similarity (P, Q)**

subject to some constraints

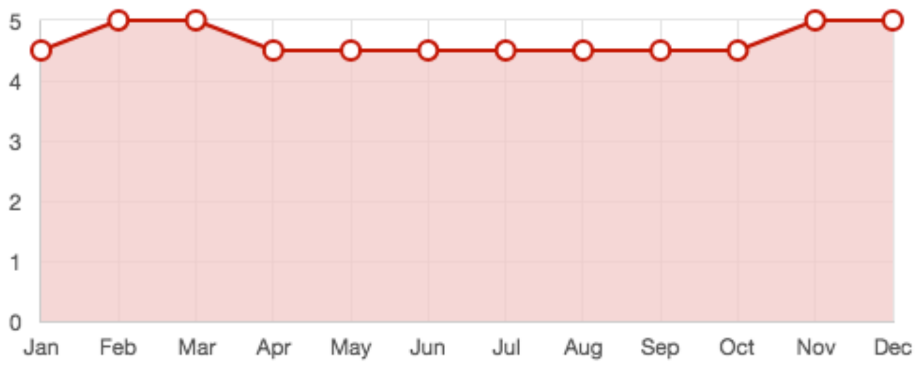
Find evasive posting frequency

Burst-avoiding techniques for attack timing.

Easy to catch

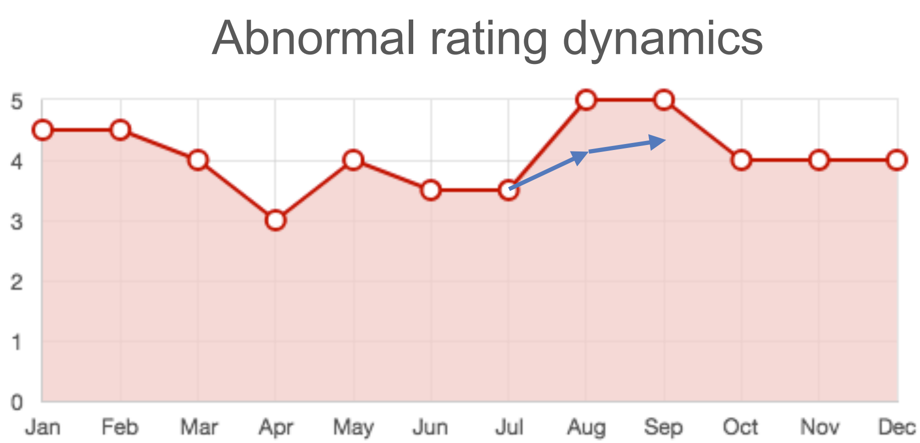


Looks more like a normal one



Find evasive rating distribution

Burst-avoiding techniques for attack timing.



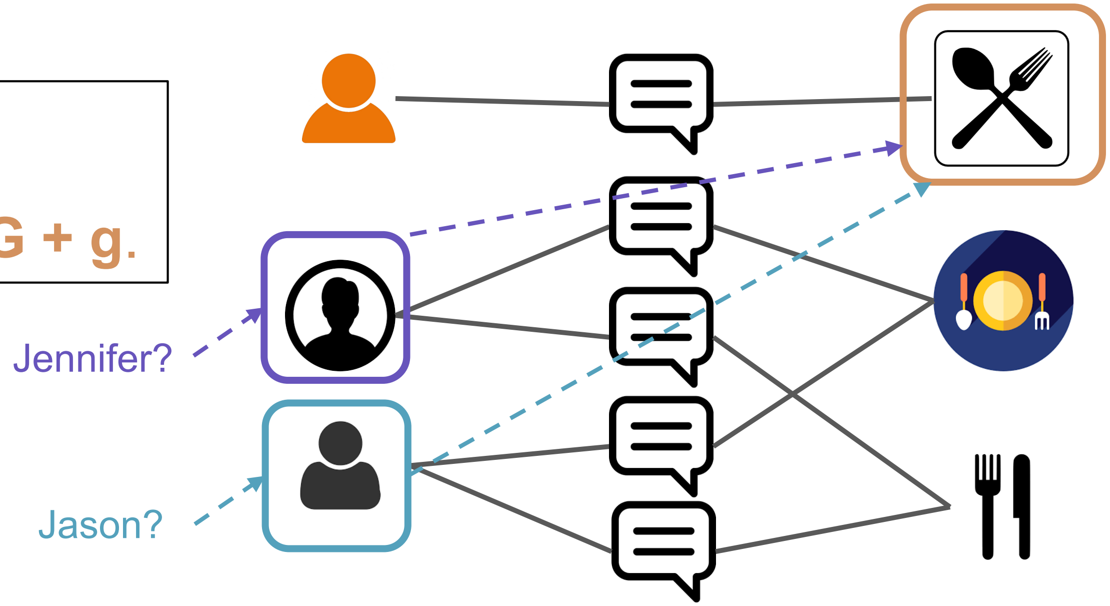
max current + future promotion

subject to smoothness constraint

Find attacking accounts

$\max_{g} \quad \mathbf{f(G + g)}$

subject to constraints over **G + g**.



Generating fake review texts

- **Crowdturfing: fraudsters are evolving to adopt more natural sounding templates and writing [1]**

Linguistics-based detectors: < 70%

- **Cheap automatic text generators can fool linguistic-based detectors [2]**

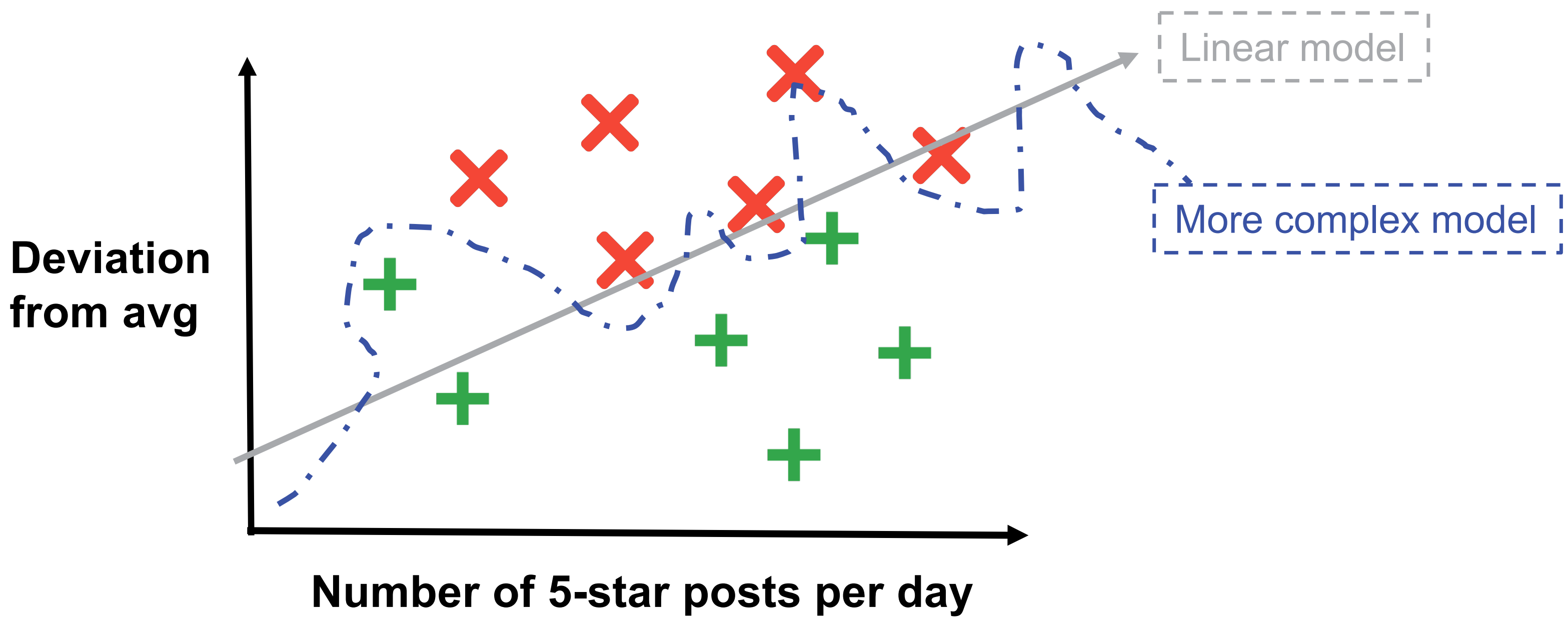
RNN is practical for short texts: 30% human detectors, 40% machine detectors F1-score

1. What Yelp Fake Review Filter Might Be Doing? ICWSM, 2013
2. Automated Crowdturfing A acks and Defenses in Online Review Systems, CCS, 2017
3. Maximum-Likelihood Augmented Discrete Generative Adversarial Networks, ICML, 2017

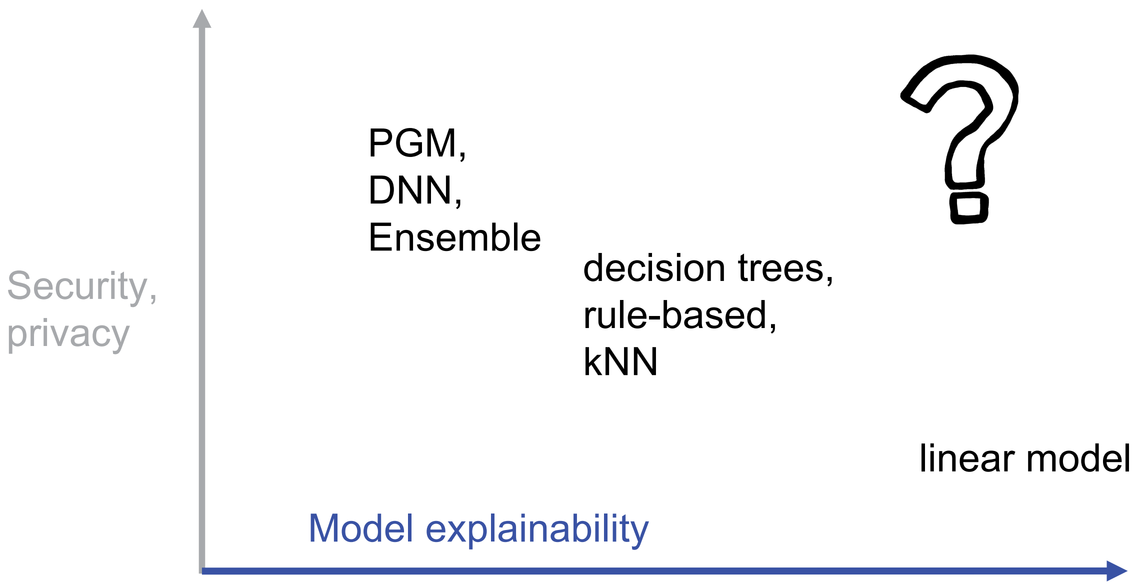
Challenges

1. Accuracy vs. Explainability.
2. Reactive Detection vs. Active Fraudsters.
3. **Explainability vs. Security.**

Explainability vs. security



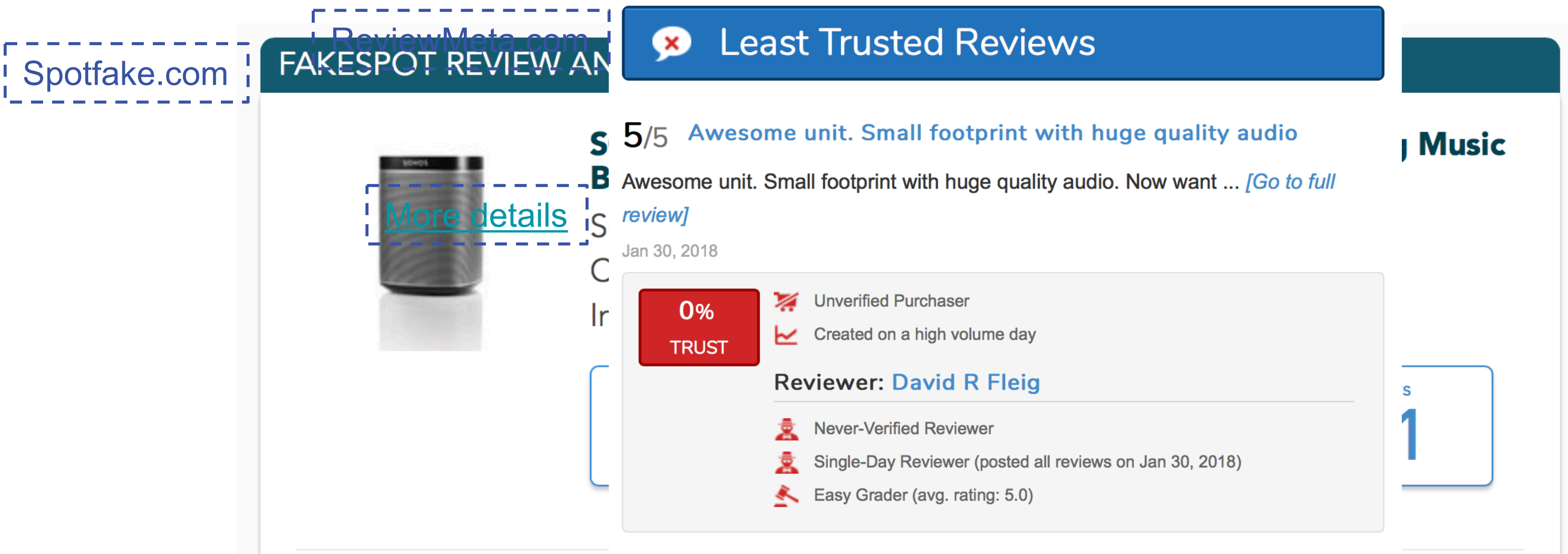
Explainability vs. security



1. Ian J. Goodfellow, Jonathon Shlens & Christian Szegedy. Explaining and harnessing adversarial examples. ICLR. 2015.
2. Florian Tramer, et al. Stealing machine learning models via prediction apis. USENIX. 2016.

LEHIGH UNIVERSITY

The reality



Thank you

LEHIGH UNIVERSITY