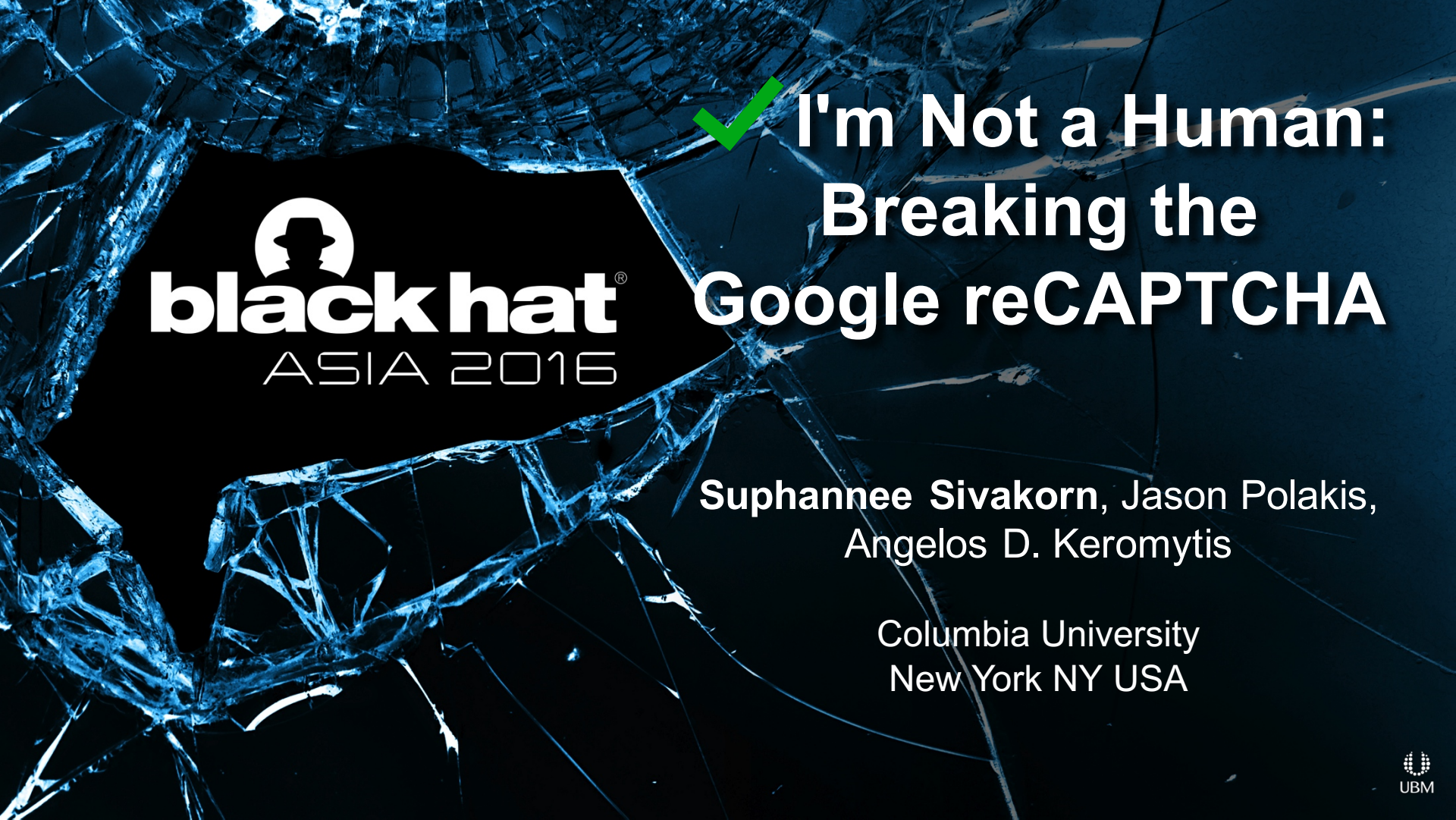


I'm Not a Human: Breaking the Google reCAPTCHA

Suphannee Sivakorn, Jason Polakis,
Angelos D. Keromytis

Columbia University
New York NY USA

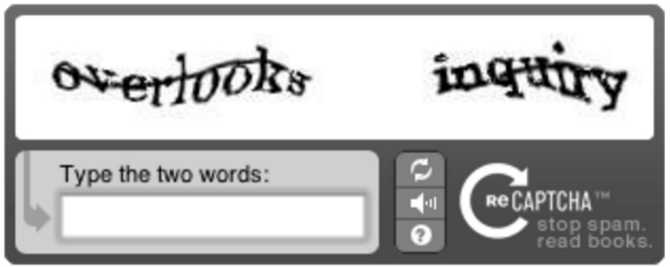
UBM

Introduction

- Economics of cyber-crime rely on large-scale automation
- CAPTCHAs can prevent automated account creation, message posting, …

- Most common type: text
- Text CAPTCHAs obsolete

Bursztein et al., WOOT 2014
Gao et al., NDSS 2016

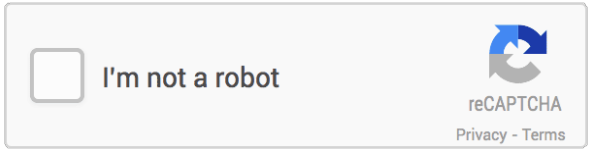


Text CAPTCHAs user-**UN**friendly

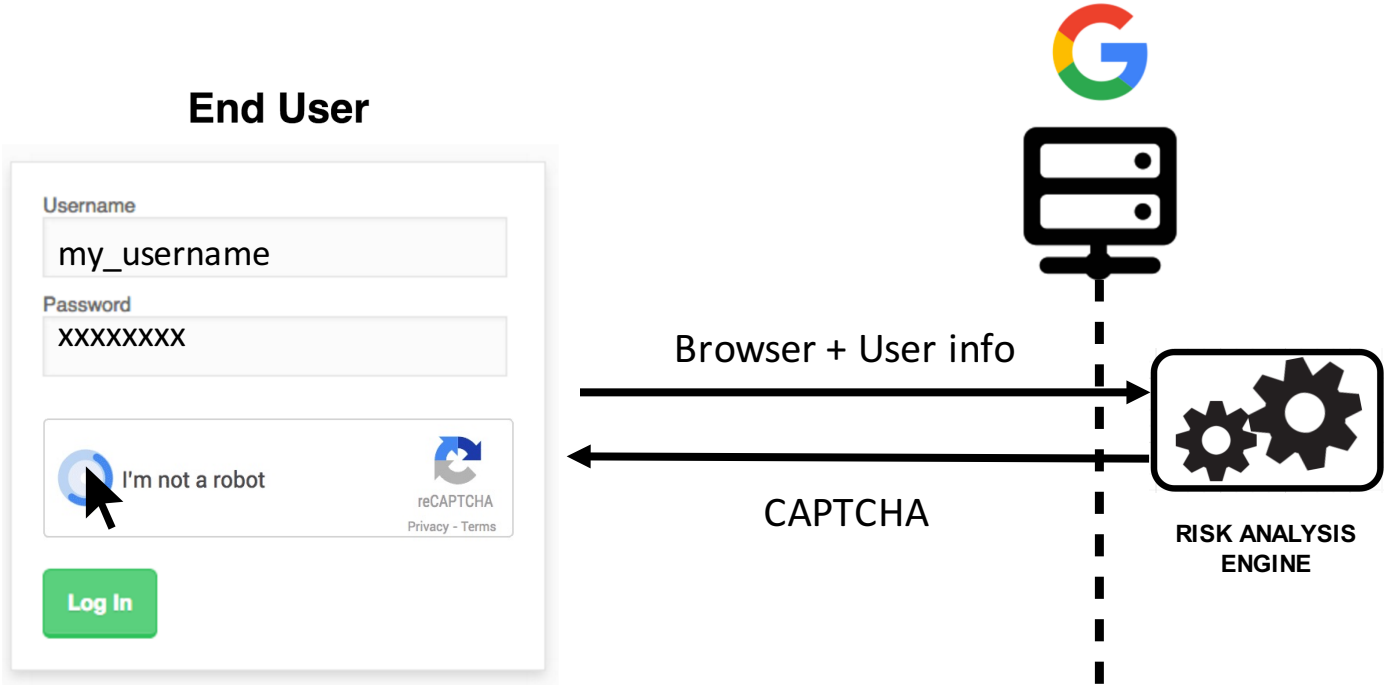
Arms race led to CAPTCHAs too complex for humans

Google reCAPTCHA

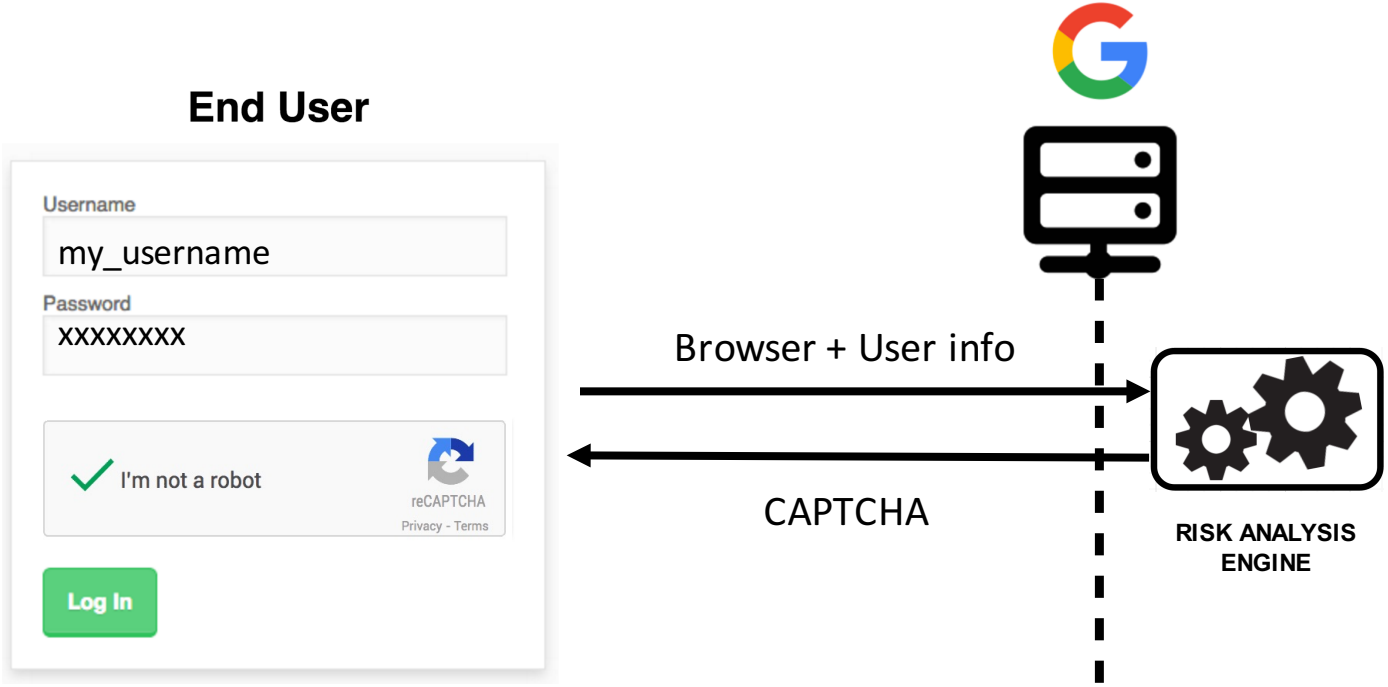
- “No CAPTCHA reCAPTCHA”
- Most popular CAPTCHA service
- New system launched December 2014
- Designed to overcome aforementioned limitations



Evaluate browser and user info



No-CAPTCHA


End User
Username
my_username
Password
xxxxxxxx
I'm not a robot
reCAPTCHA
Privacy - Terms
Log In
Browser + User info
CAPTCHA
RISK ANALYSIS ENGINE

Text CAPTCHA

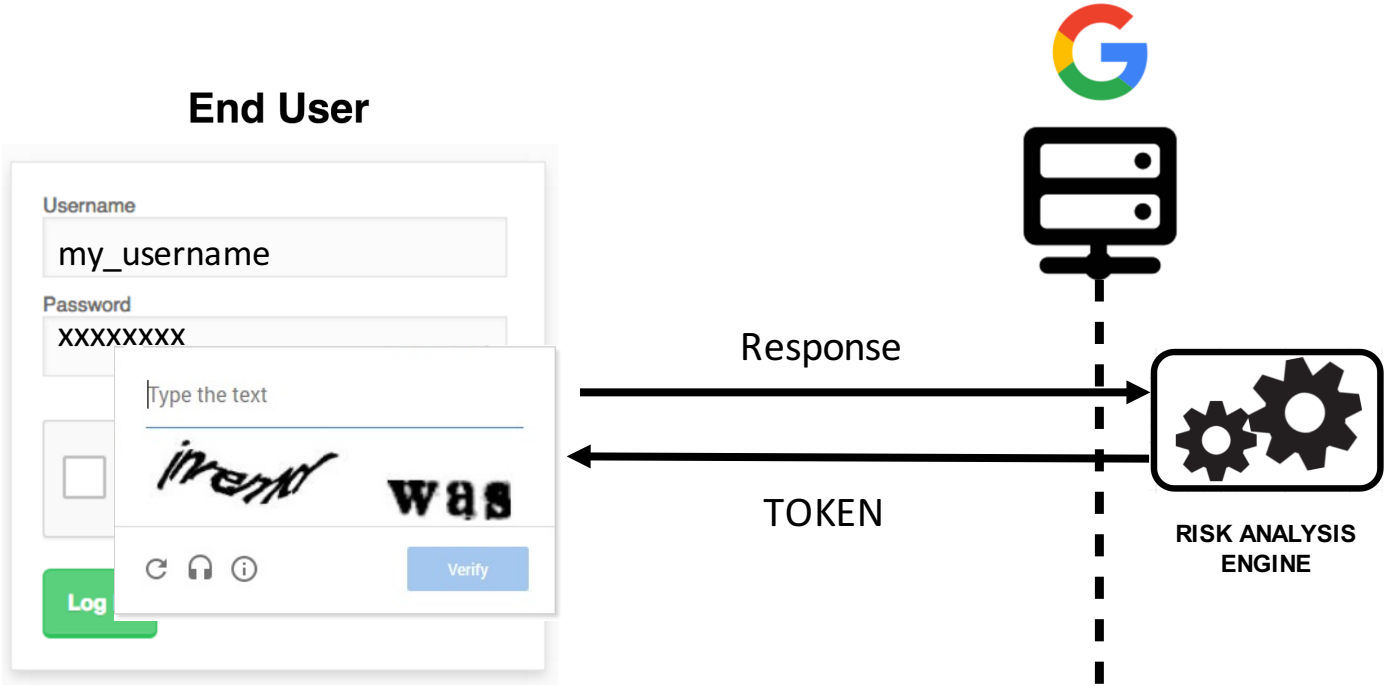
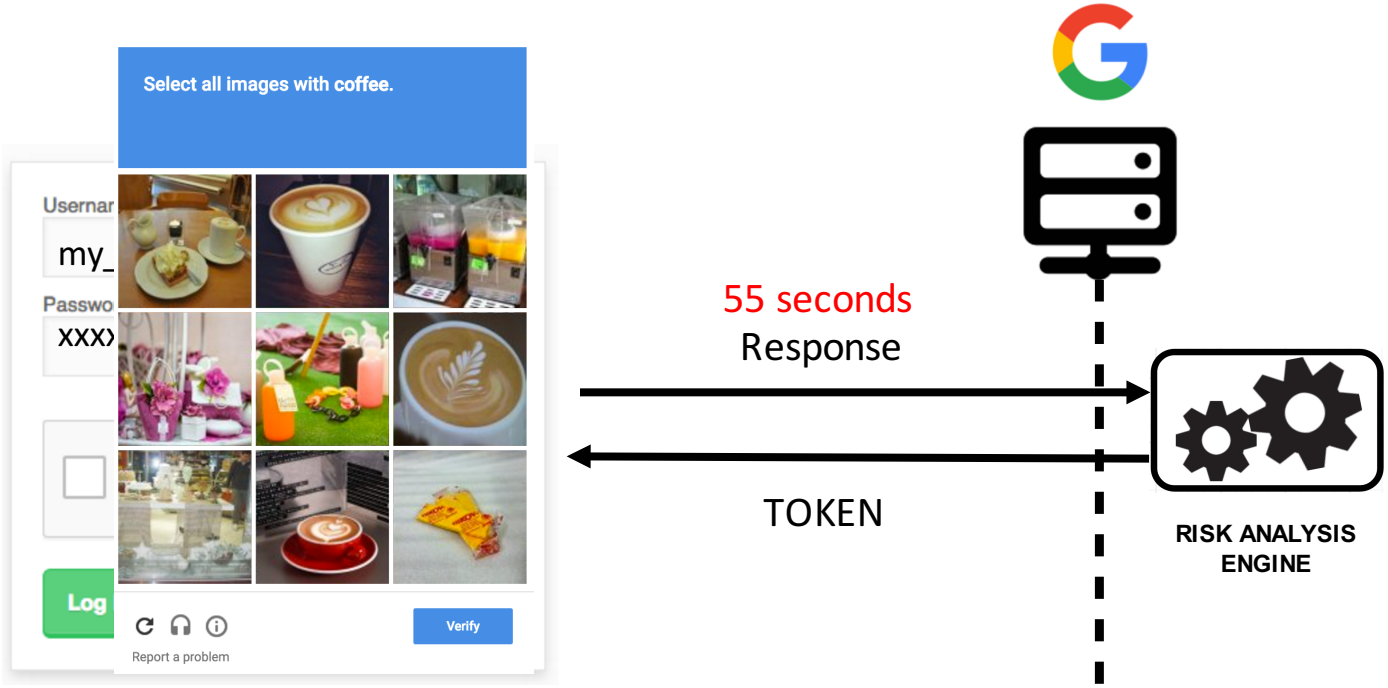


Image CAPTCHA



Can we influence risk analysis?

- Black box testing of advanced risk analysis system
- Widget created by heavily obfuscated JavaScript
- Leaked disassembled code points to certain checks

1. Checks Browser environment
2. Checks Cookies
   - HTTP cookie sent even if user not logged in Google
   - User tracking → User history

Checking browser environment

- User agent
  - Old versions receive text captcha
  - Misconfigurations receive text captcha

    e.g. Browser is Firefox 36.0, Chrome/42.0 reported
- Canvas Rendering
  - Check ability to render canvas element
  - Can fingerprint browser and OS (K. Mowery et al., 2012)
- Mouse movement
  - Irrelevant

Checking google.com cookie

- No-CAPTCHA relies on google.com cookie
- Create cookies is good enough to get no-CAPTCHA
  - Get fresh cookies from google.com
  - Create browsing history -- use Google services to build browsing history
- Check if user receives no-CAPTCHA

No history? That's OK…

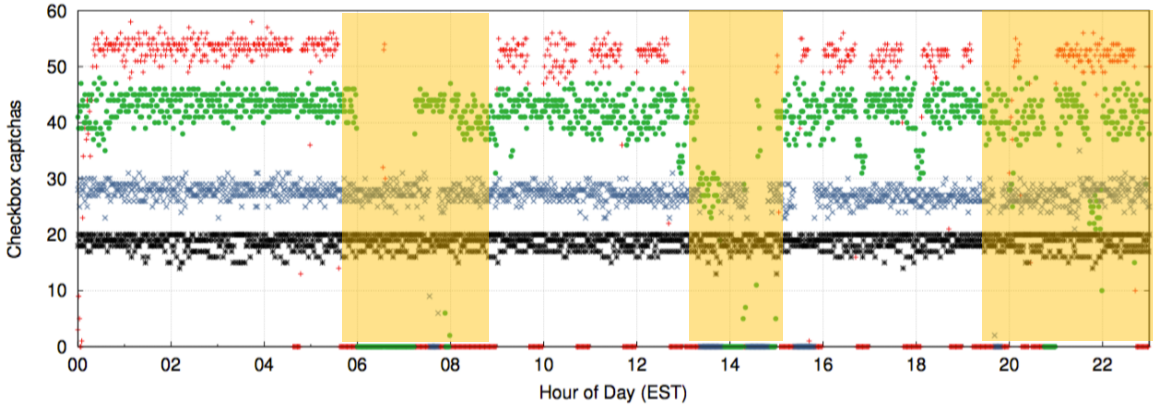
Network	Web Surfing	Account	Threshold
Departmental	Frequent	No	9^{th} day
Departmental	Moderate	No	9^{th} day
ToR	Frequent	No	9^{th} day
ToR	Moderate	No	9^{th} day
Any	**None**	**No**	**9^{th} day**

We can "bypass" defenses with 9-day old cookies!

Cookie harvesting

- Up to 8 No-CAPTCHAs per “user” per day
  - Remember…It’s all about scale!
- How many cookies can we harvest per day?
  - Approx. 63k from single IP address
  - No safeguards to prevent mass harvesting

Limits on receiving “No CAPTCHAs”



- Each color represents different request rate
- Higher rates result in “time-outs”
- Optimally ~55k tokens per day from single IP address

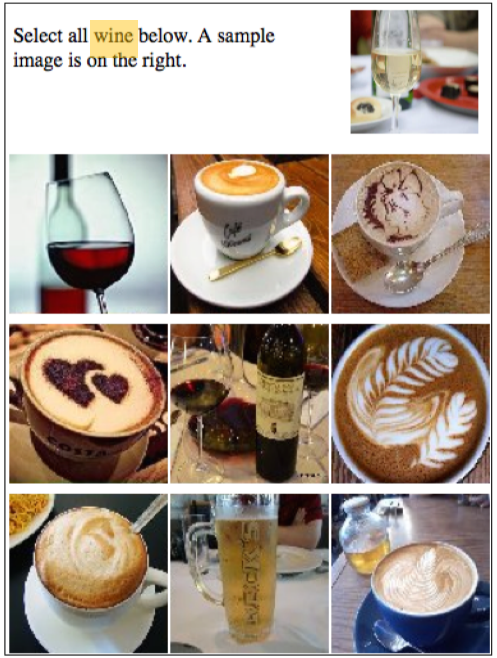
Solving image CAPTCHAs

- Great advancements in deep learning
- Systems return keywords describing content or visual characteristics

GRIS	Alchemy	Clarifai	TDL	NeuralTalk	Caffe
wine and blood	wine, glass	glass, red wine, wine, merlot, liquid, bottle, still, glassware, alcohol, drink, wineglass, beverage, pouring, white wine, cabernet, taste, leaded glass, dining, party, vino	red wine, goblet, wine bottle, punching bag, beer glass, perfume, balloon	a glass of wine sitting on top of a table	red wine, wine, alcohol, drug of abuse, drug, red wine, punching bag, beaker, cocktail shaker, table lamp

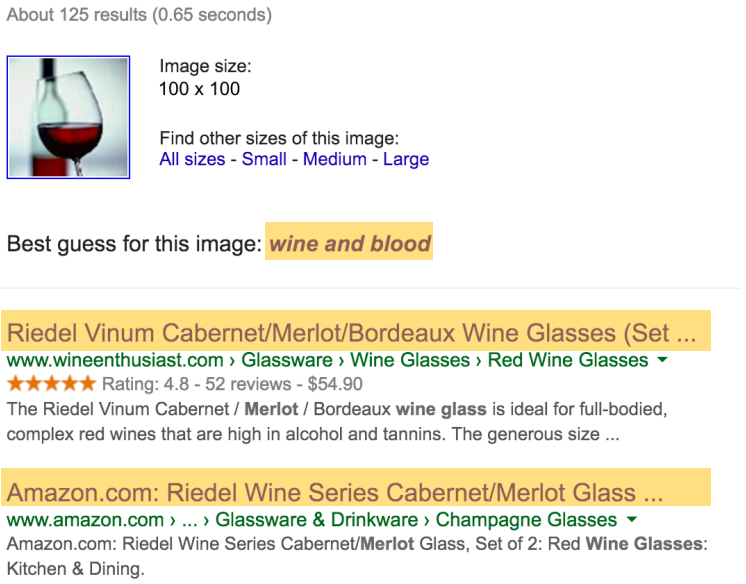
Automatically solving image CAPTCHAs

- Extract hint, sample & candidate images
- Get annotated tags for candidate images
- Compare hint and annotated output tags



Using Google against Google

- Google Reverse Image Search (GRIS)
  - Keywords describing the image
  - Titles from pages containing the image
  - Google translate
  - Higher resolution images

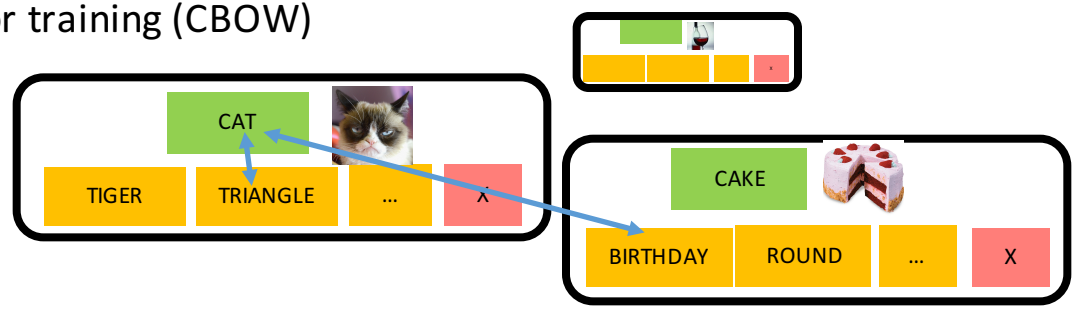


Processing image tags

- Word matching
- Word similarity (Corpus-based approach)
  e.g. <river, ocean>, <pasta, spaghetti>, <cat, tiger>
- Sometimes they are not good enough.
- Annotation services return some odd relationships
  e.g. triangle → cat image
- Word2Vec (Mowery et al., CCS 2012)
  similarity_score(hint, tags)

Tag classifier with Word2Vec

- Use annotate service to get tags of 7,000 images in offline set
- Manually labeled all of them (e.g., cat, cake, wine)
- Use them for training (CBOW)



- Word2Vec removes common tags and enhances relationship between tags and correct label
- Distance of hint and annotated tags → Similarity score
- Increases accuracy by ~5%

Attack evaluation

- Using online services
  - 70.8% accuracy, 19.2 seconds per challenge
- Using offline library (Caffe)
  - No-cost attack
  - No help from online services
  - 41.6% accuracy, 20.9 seconds
- How good is that?
  - Traditionally ≥ 1% means CAPTCHA scheme is broken

How good are humans?

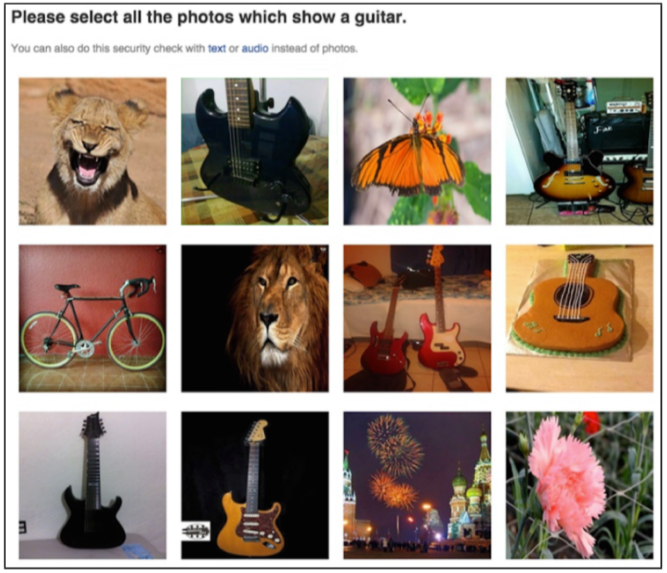
- Used http://de-captcher.com that supports Google reCAPTCHA

Detail	de-captcher
System overloaded Error	21.00%
Timeout Error	12.57%
Pass Challenge (resubmitted all error returned challenges)	**44.30%**

- Avg solving time 22.5 seconds when no error returned
- No-cost offline attack comparable in accuracy, slightly faster!

Applicability: breaking Facebook CAPTCHA

- Presented when suspicious URLs are posted or sent in messages
- Filler images from different categories -> easier to distinguish
- 83.5% accuracy



Countermeasures? Raise the bar

- Risk analysis system
  - Require account
  - Cookie reputation
  - IP address / Subnet safeguards
- Image CAPTCHA
  - Advanced semantic relations (tennis ball, court, racket)
  - Adding visual noise (requires extensive evaluation)
  - Creating adversarial images
    - Images look the same but can't be identified by deep learning systems (Szegedy et al., CoRR 2013, Papernot et al., EuroSP 2016)
- Responsible disclosure

Where do we go now?

- Text CAPTCHAs broken
- Semantic image CAPTCHAs broken

- Open problem!
  - Maybe cognitive game CAPTCHAs?

Black Hat Sound Bytes

- More complex system, more ways to bypass
  - Novel misuse of web cookies
- Great advancements in automated cognition
  - Image semantics no longer an obstacle
- Are CAPTCHAs an outdated concept?

Thank you!

bagels or puppies?