CS 419: Computer Security

Week 6 Recitation: CAPTCHA

The Problem

People would create scripts to interact with services at scale

- Boosting URLs in search engines
- Brute-forcing passwords
- Creating many, many accounts on for free email or cloud services
- Getting lots of free trials, API keys
- Filling out surveys to skew results

We needed a solution for a server to know it is interacting with a human

Authenticating humans

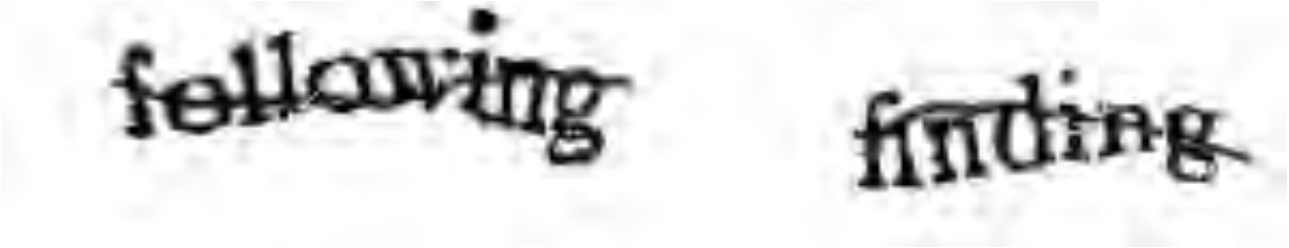
Challenge:

- Create a test that is easy for **humans** but extremely difficult for **computers**

CAPTCHA:
Completely Automated Public Turing test to tell Computers and Humans Apart

Technique:

Degrade/distort images – distort text so that computer vision algorithms fail



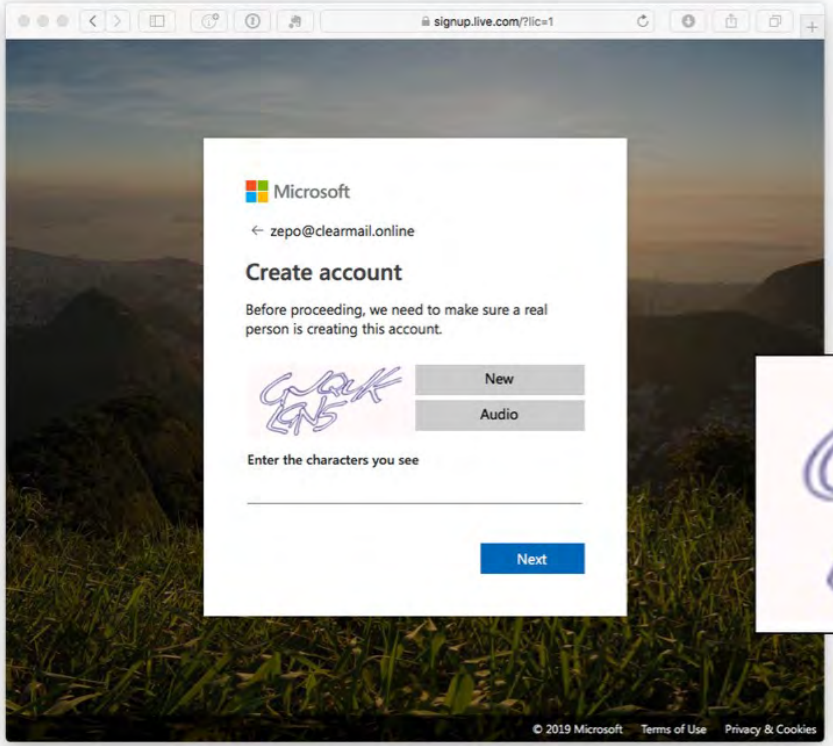
CAPTHA: Early History

- **1997: AltaVista (an early search engine)**
  - Designed to prevent bots from registering URLs with the search engine
- **2000: Yahoo! and Manuel Blum & his team at CMU created EZ-Gimpy**
  - Distort one of 850 words

- **2003: Henry Baird @ CMU & Monica Chew at UCB created BaffleText**
  - Generates a few words and random non-English words

CAPTCHA Example (2019)

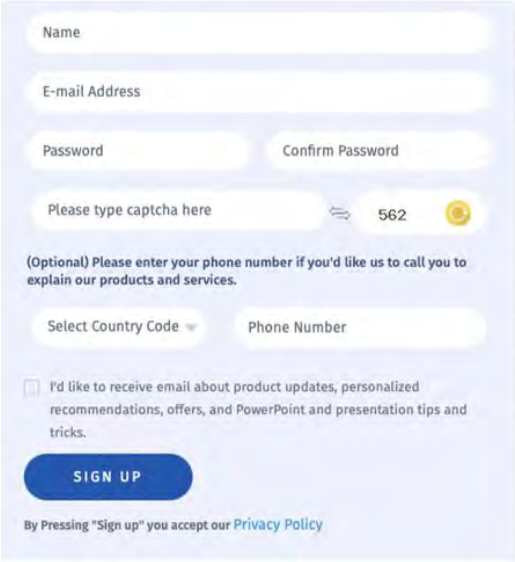
Microsoft

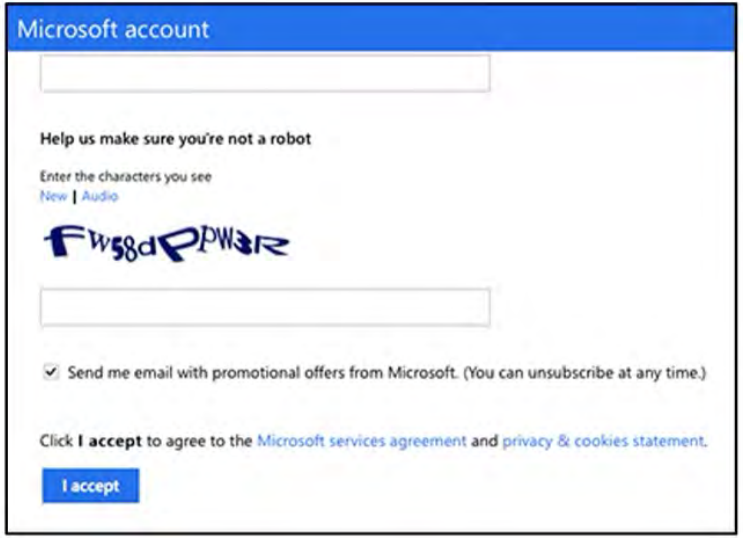


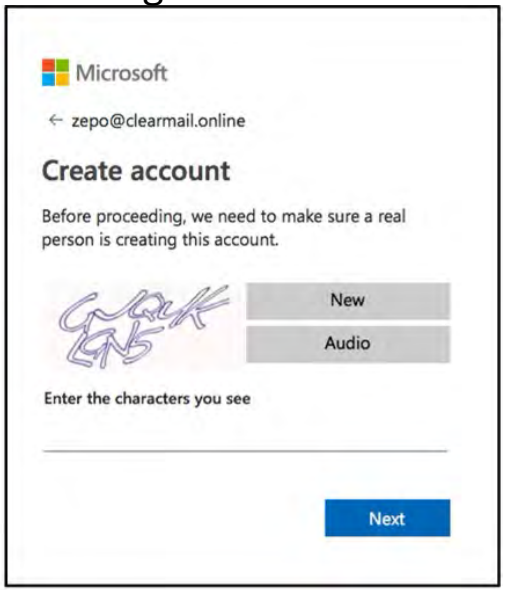
See captchas.net

They had to get more difficult

Advances in machine learning & character recognition led to automated solving

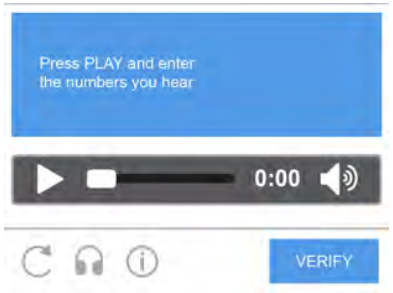






Problems

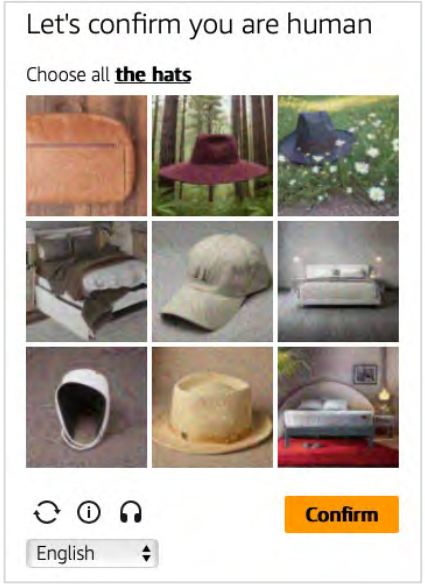
- **Accessibility**
  - Visual impairment → audio CAPTCHAs
  - Deaf-blind users are left out

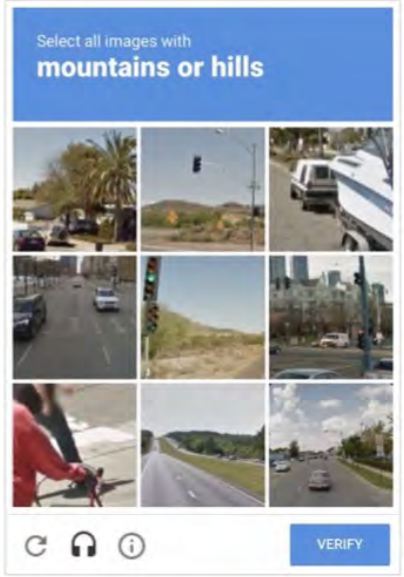


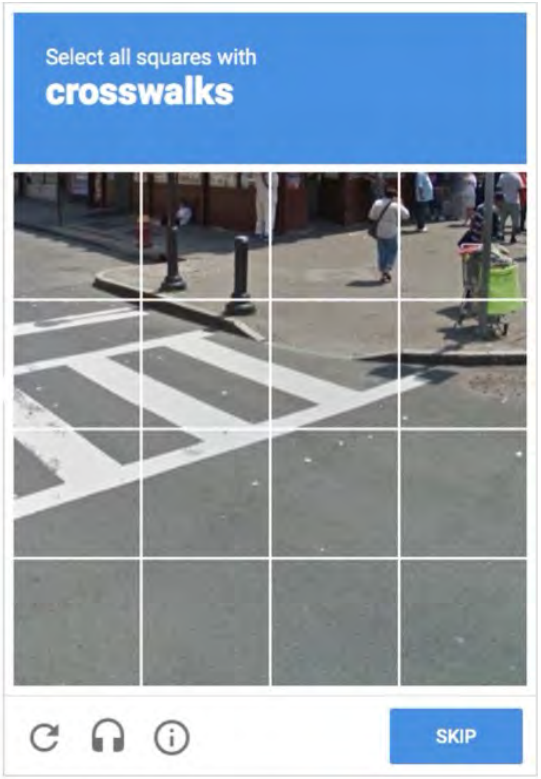
- **Frustration**
  - Typing text was more of a pain on mobile devices
  - OCR & computer vision algorithms improved a lot, so the puzzles got harder
  - Challenges that are now difficult for computers may be difficult for humans
- **Attacks**
  - Man in the middle attacks – software can redirect the challenge to humans
    - Use human labor to solve the puzzles – **CAPTCHA farms**
  - Automated CAPTCHA solvers
    - Initially, educated guesses over a small vocabulary – later, use improved image recognition

Alternate approaches: image recognition

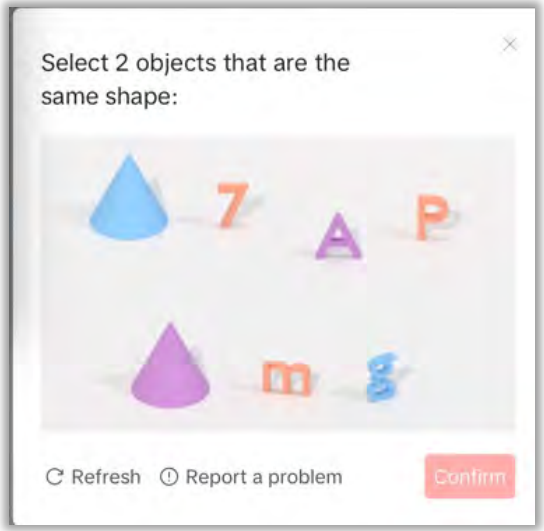
- **Scene recognition**
- **Touching is easier than typing on phones**

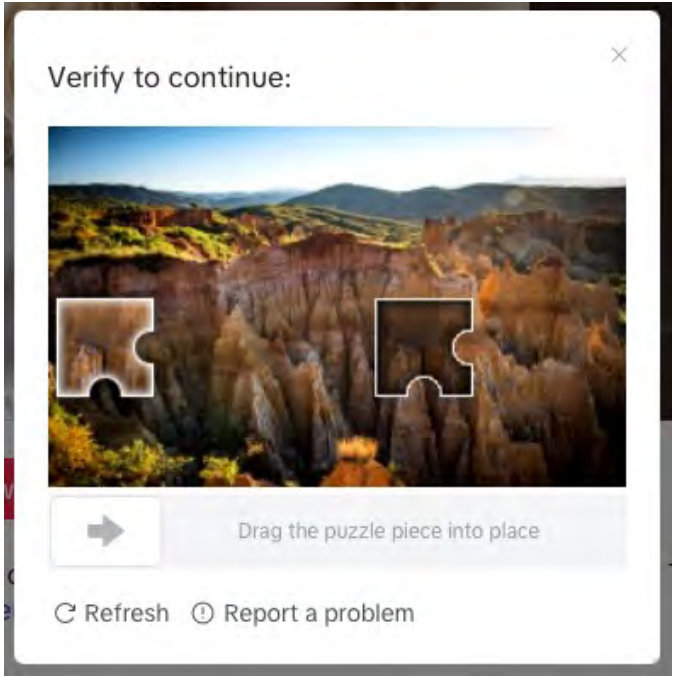


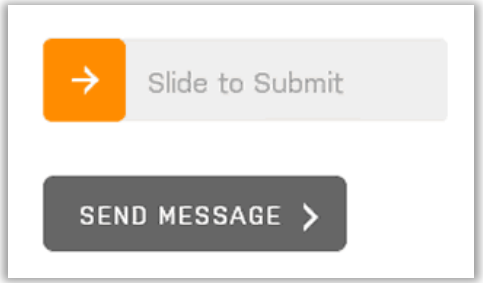




Alternate Approaches

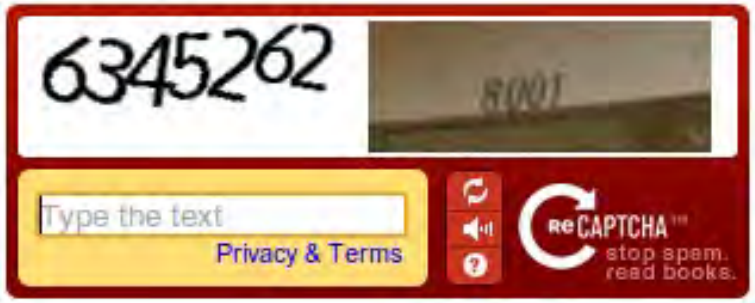






reCAPTCHA

Ask users to translate images of real words & numbers from archival texts



Two sections

(1) known text

(2) image text

- Assume that if you get one right , then you get the next one correct
  Try it again on a few other people to ensure identical answers before marking it correct

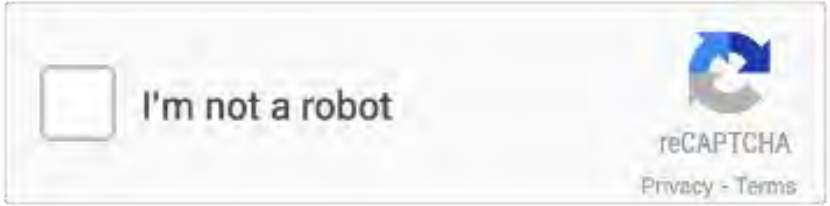
Google bought reCAPTCHA

- Used free human labor to improve transcription of old books & street data
- This enabled the use of human labor to fix up the archives of the New York Times

By 2014:
Google found that AI could crack CAPTCHA & reCAPTCHA images with 99.8% accuracy

NoCAPTCHA reCAPTCHA: reCAPTCHA v2

Just ask users if they are a robot!



Reputation management

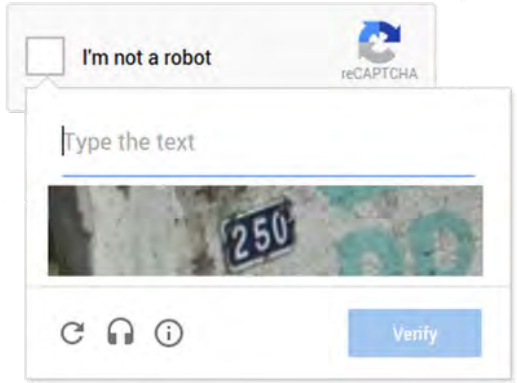
- "Advanced Risk Analysis backend"
- Check IP addresses of known bots
- Check Google cookies from your browser
- Considers user's engagement with the CAPTCHA: before, during, and after
  - Mouse movements & acceleration, the precise location of clicks
- Generate a confidence score and allow the service to decide if it's good enough

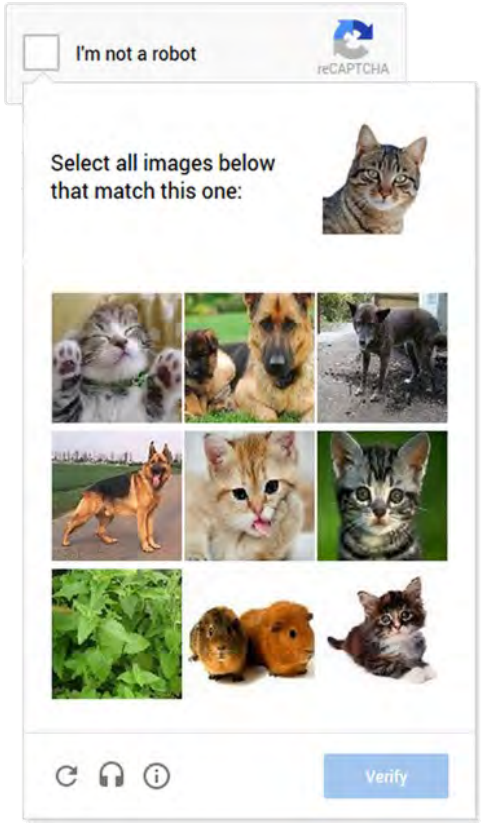
reCAPTCHA v2 fallback

If risk analysis fails, present a CAPTCHA

Three challenge types:

1. **Classification**: identify which images in a 3x3 grid belong to a given description (e.g., find all images with bridges)
2. **Classification**: like (1) but images are replaced after clicking ("click until there are none left")
3. **Segmentation**: break an image into a 4x4 grid and identify parts that are relevant to the request (e.g., identify all parts of a motorcycle)





Google's reCAPTCHA v3: Invisible reCAPTCHA

Don't even ask users if they are a robot

- **Track behavior throughout a session**
- **Generate a *trust score*: 0..1 = likelihood it's a human**
- **Websites use the trust score to present a visible CAPTCHA or block access**

The AI Threat

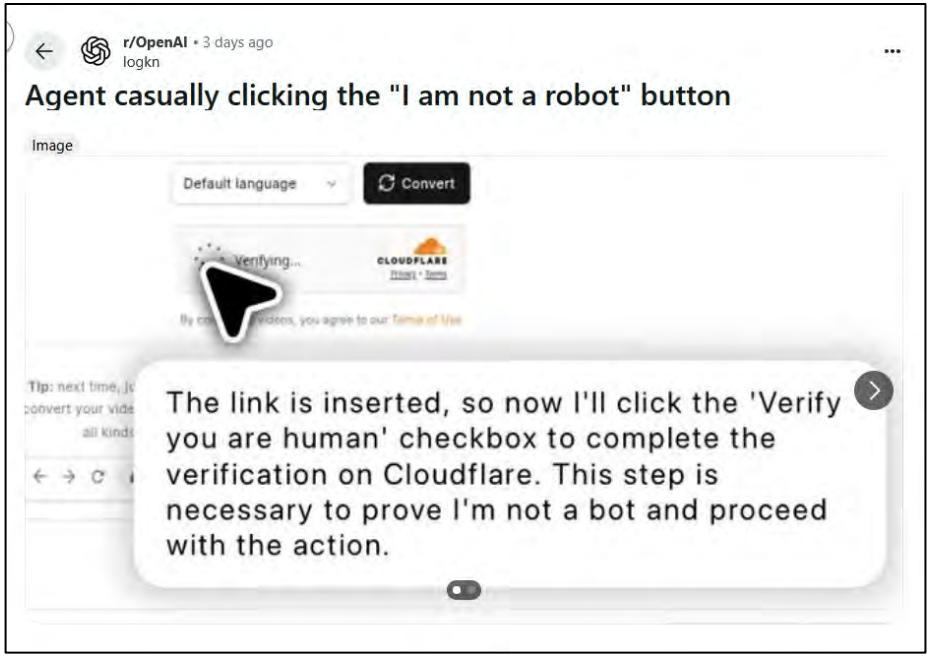
In 2024, a team at ETH Zurich demonstrated that reCAPTCHAv2 challenges can be solved 100% of the time using publicly-available AI software

- Apply a fine-tuned version of the open-source YOLO (You Only Look Once) object-recognition model
- Use a VPN to connect with a different IP address for lots of repeated attempts – makes each connection appear to be unique
- Incorporate artificial Bezier curve-based mouse movements to simulate human behavior

See https://arxiv.org/abs/2409.08831

The AI Threat: LLMs can get through CAPTCHAs

- **July 25, 2025: OpenAI's ChatGPT Agent, which can perform tasks for users, was observed detecting, understanding, and clicking through Cloudflare's "Verify you are human" checkbox**
- **It didn't have to solve a puzzle, but its actions passed Cloudflare's behavioral screening**



https://arstechnica.com/information-technology/2025/07/openais-chatgpt-agent-casually-clicks-through-i-am-not-a-robot-verification-test/

IllusionCAPTCHA: An attempt at defeating AI

AI-created optical illusions are not recognized by other AI systems

- Gen AI combines an input image and a prompt
- AIs can create images but not detect the illusions: LLMs failed 100% of the time
- Humans passed the test 83% of the time

https://arxiv.org/abs/2502.05461

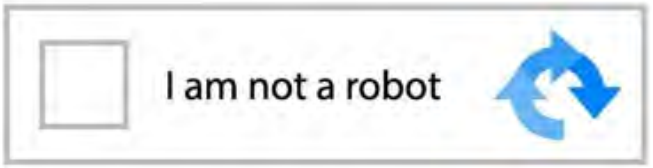
https://archive.is/mAHQC

https://www.newscientist.com/article/2468020-ai-generated-optical-illusions-can-sort-humans-from-bots/

Fake CAPTCHA Prompts Used Maliciously

In 2024, the Ukrainian Computer Emergency Response Team warned that the APT28 threat group (Fancy Bear, thought to be affiliated with Russian Intelligence) has been using CAPTCHA impersonation

- **Present a fake "I am not a robot" message to get users to click on the checkbox**
- **This initiates a malicious PowerShell command to the user's clipboard**
- **The attack targets government workers in Ukraine but can inspire other attackers to use the same technique**



https://www.forbes.com/sites/daveywinder/2024/10/26/new-google-cyber-attack-warning-as-russian-apt28-hackers-strike/

Other approaches to try to verify humans

- **Text/email verification**
  - Ask users for a phone # or email address
  - Similar to two-factor authentication, but we're *not authenticating* the user
    - Just having them do something
  - Service sends a message containing a verification code
    - Still susceptible to spamming & automation
    - Makes the process more cumbersome
    - Requires users to disclose information
- **Measure form completion times**
  - Users take longer than bots to fill out and submit forms
  - Measure completion times and randomness in delays
    - But bots can program delays if they realize this is being done

The Orb & Orb Mini

Tools for Humanity

- Founded by OpenAI CEO Sam Altman

- **Based on the idea that it will eventually be impossible to distinguish humans from AI agents**
- **Iris scan produces a unique ID that is stored in a token on a Worldcoin blockchain**
- **Privacy and adoption concerns**

https://www.toolsforhumanity.com/orb

https://techcrunch.com/2025/04/30/sam-altmans-world-unveils-a-mobile-verification-device/

The End