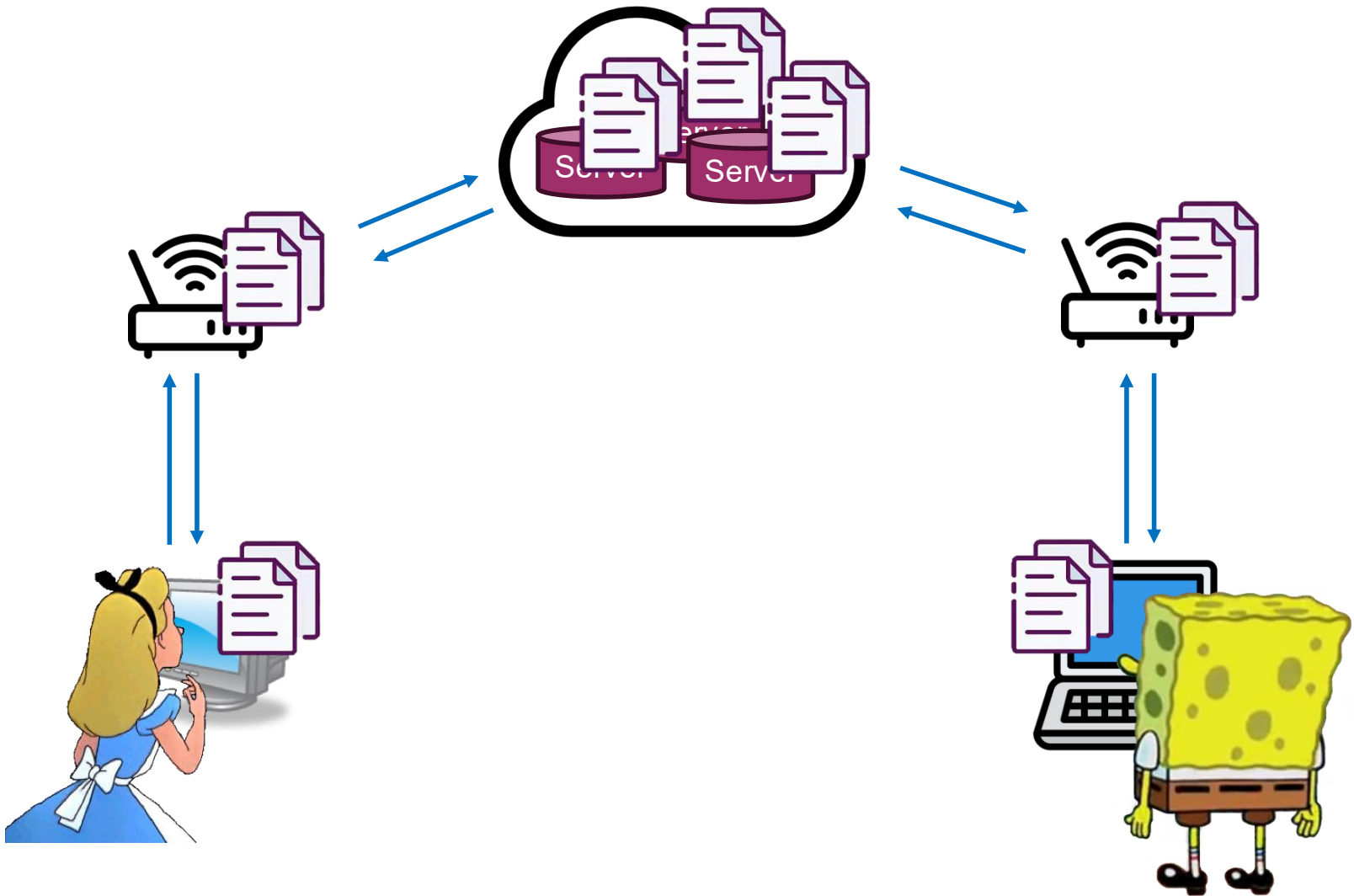


Lecture 8: Client-Side Scanning

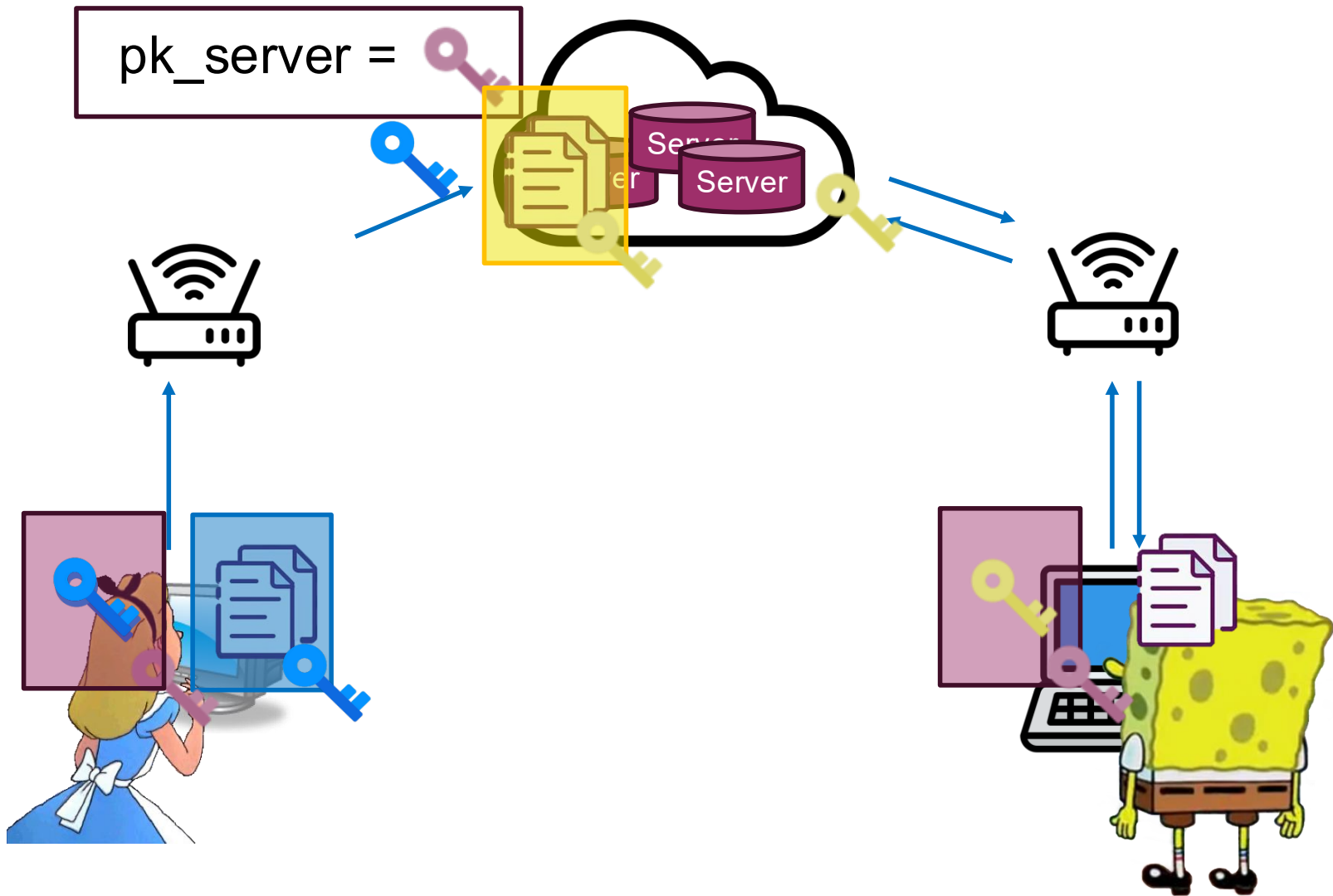
CS 181E

Fall 2026

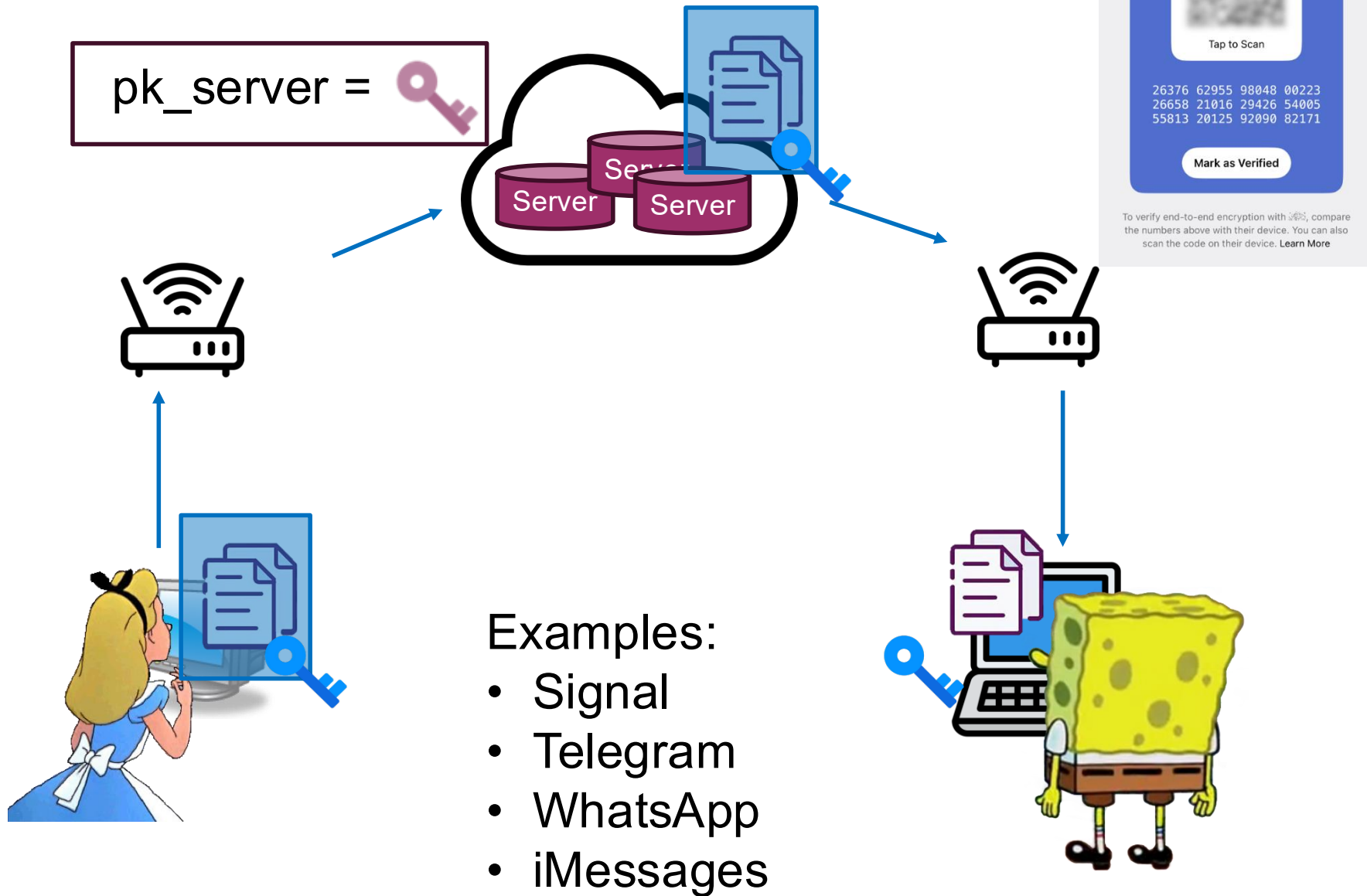
Review: Computer Systems



Review: Point-to-Point Encryption



Review: End-to-End Encryption



Review: The Debate over E2E encryption

Pros

- Good for privacy
- Don't have to trust server
- Important for journalists/ protesters, esp. in suppressive regimes

Cons

- Can't monitor for bad content
 - violations of T&C
 - illegal content
- Gov can't subpoena data

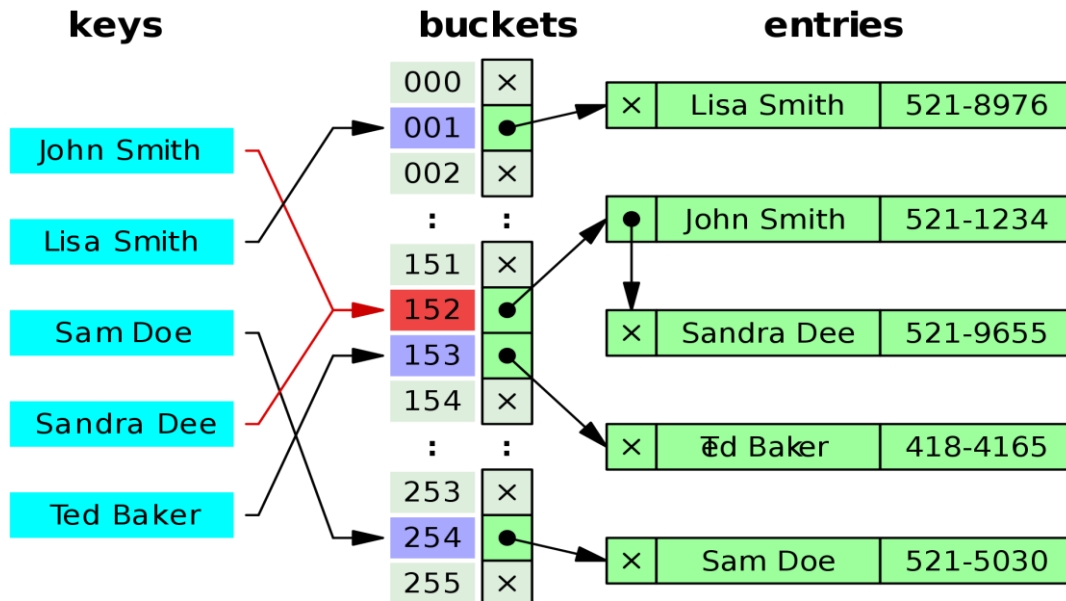
CLIENT-SIDE SCANNING

High-level Idea

- content is only accessible on the client
- run monitoring code on the client to detect known bad stuff
- to handle large-scale data, uses hash functions

Hash functions

- Input: arbitrary size bit string
- Output: fixed size bit string
 - **efficient**: easy/fast to compute
 - **deterministic**: if you compute the output twice, you get the same answer
 - **compression**: size of the output is smaller than the input
 - **diffusion**: minimize collisions (and clustering)



Cryptographic hash functions



- Input: arbitrary size bit string
- Output: fixed size bit string
 - **efficient:** easy/fast to compute
 - **deterministic:** if you compute the output twice, you get the same answer
 - **compression:** size of the output is smaller than the input
 - **diffusion:** minimize collisions (and clustering)
 - **collision-resistant:** hard to find to inputs with the same output
 - **one-way:** given an output y , hard to find any x such that $h(x) = y$

Exercise: Hash Functions

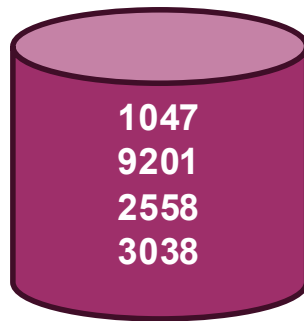
- Consider the function $h(x) = x \% 256$. Would this function make a good cryptographic hash function? Why or why not?

1. **efficient**
2. **deterministic**
3. **compression**
4. **diffusion**
5. **collision-resistant**
6. **one-way**

Real world hash functions

- **SHA-2:** NSA (2001)
 - Family of algorithms with output sizes {224, 256, 384, 512}
 - In principle, could one day be vulnerable to similar attacks to earlier hash functions that are no longer secure
- **SHA-3:** public competition (won in 2012, standardized by NIST in 2015)
 - Same output sizes as SHA-2
 - Plus a variable-length output option called SHAKE

Hash-Based Monitoring

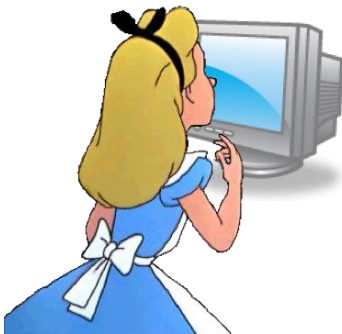


$$h(\text{scorpion}) = 1047$$

$$h(\text{mouse}) = 9201$$

$$h(\text{snake}) = 2558$$

$$h(\text{spider}) = 2558$$



$$h(\text{butterfly}) = 3290 \quad \checkmark$$

$$h(\text{scorpion}) = 1047 \quad \times$$

However...

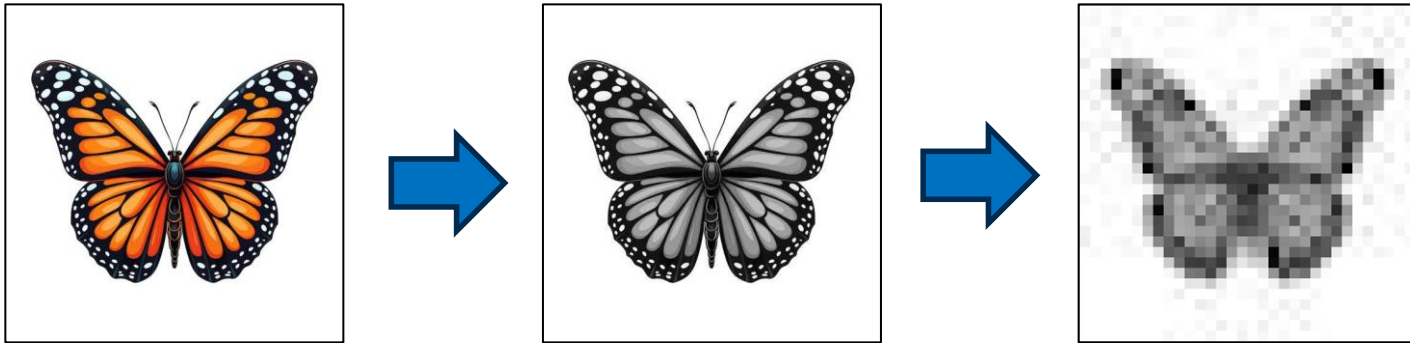
$$h(\text{🦂}) = 1047$$

$$h(\text{🦂👁}) = 8343$$

- minor changes to input generate completely different outputs
- can bypass detection by modifying image slightly

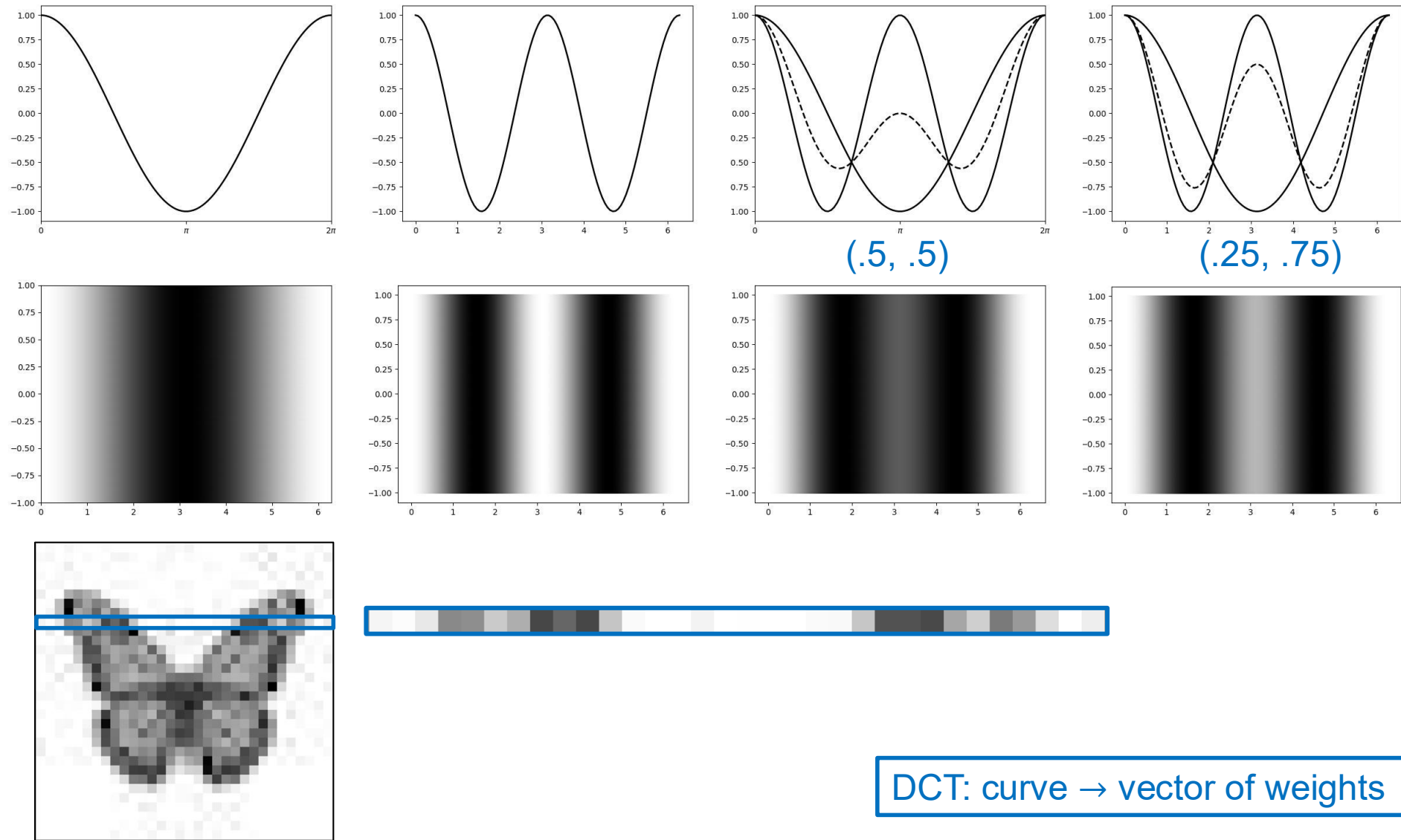
Perceptual Hashing

- Step 1: Simplify the image to 32x32 black-and-white



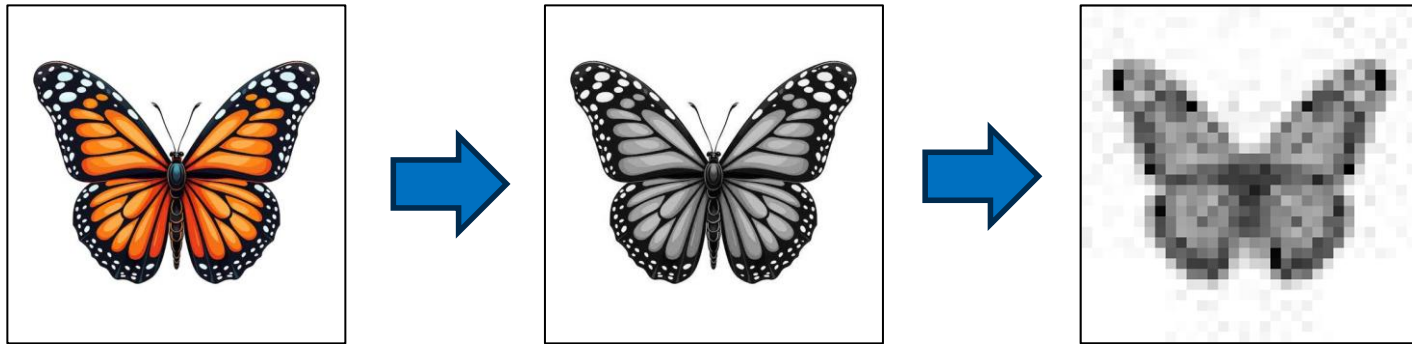
- Step 2: Apply Discrete Cosine Transform (DCT)

Discrete Cosine Transform (DCT)

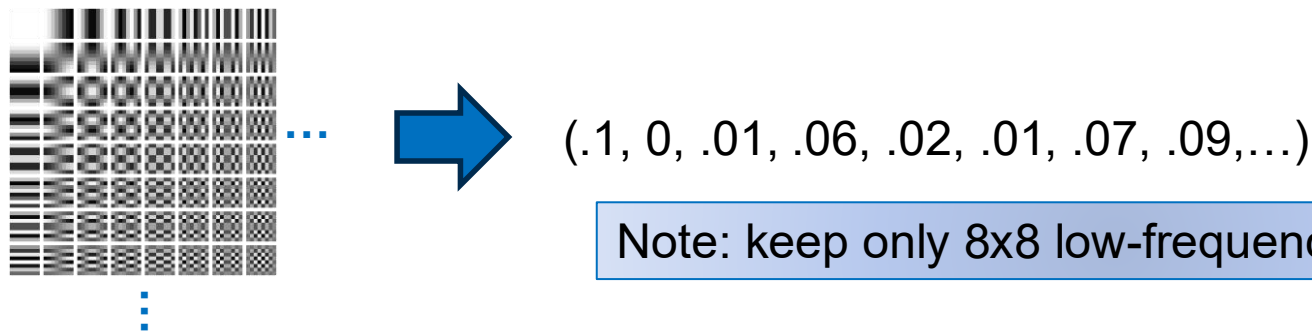


Perceptual Hashing

- Step 1: Simplify the image to 32x32 black-and-white



- Step 2: Apply Discrete Cosine Transform (DCT)



Note: keep only 8x8 low-frequency coefficients

- Step 3: Round

$(.1, 0, .01, .06, .02, .01, .07, .09, \dots)$



10010011....

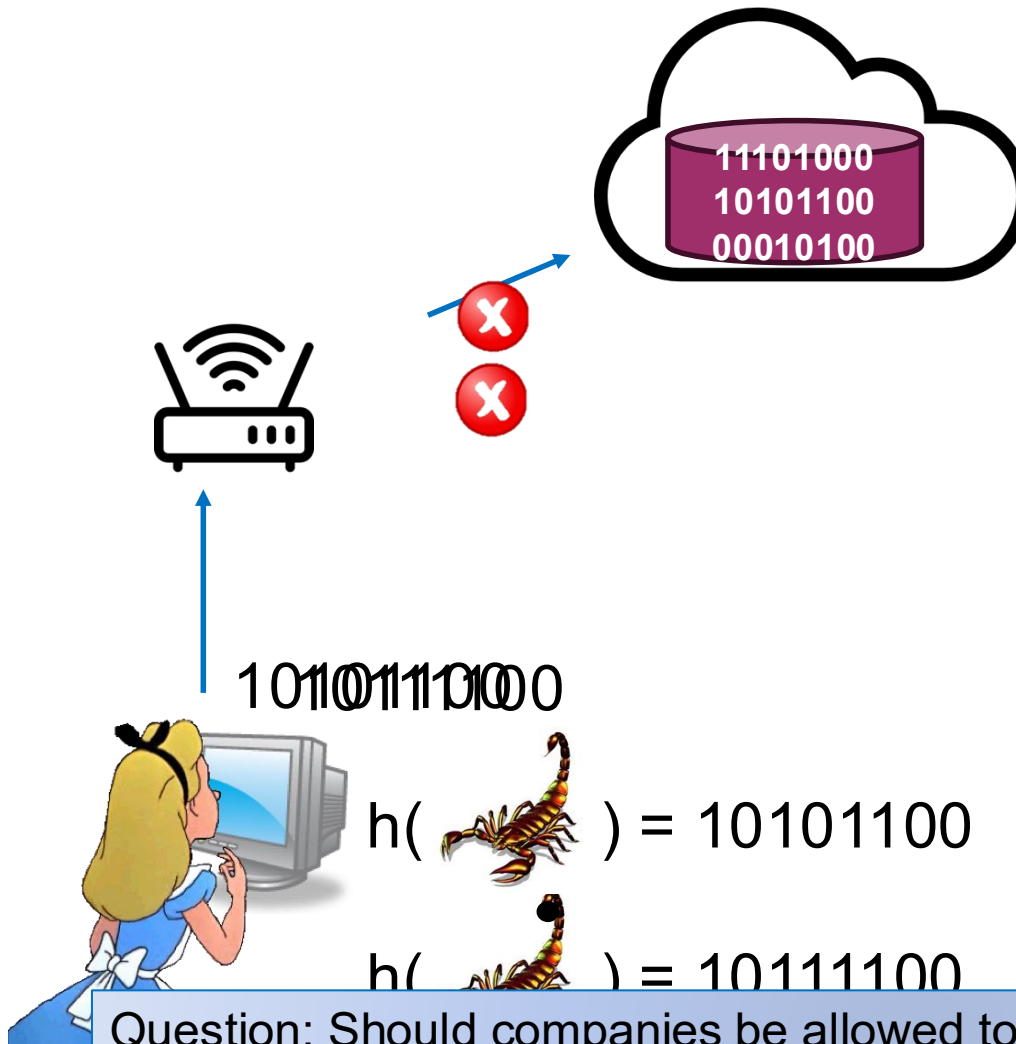
64-bit hash value

PhotoDNA

- developed by Microsoft research ~2009
- widely used for detecting CSAM in cloud storage
 - also used to detect non-consensual imagery and extremist content
- confirmed use by Microsoft, Google, Meta, and others
- similar ideas are used for Google's image search and (allegedly) Youtube's copyright infringement detection
- in 2021, Apple announced plan to use similar techniques to scan content stored on devices

Client-Side Scanning

“close” is defined by Hamming distance, i.e. how many bits are different



Question: Should companies be allowed to implement client-side scanning?
Should companies be required to implement client-side scanning?

Problem 1: Can bypass detection



Original Image



Manipulated Image

Can also be used for misinformation



Original Image



Manipulated Image
(flagged as similar)



Manipulated Image
(not flagged)

Problem #2: Accidental Collisions



Hamming Distance ignores context

A Dad Took Photos of His Naked Toddler for the Doctor. Google Flagged Him as a Criminal.

Google has an automated tool to detect abusive images of children. But the system can get it wrong, and the consequences are serious.



Share full article



1.7K



By **Kashmir Hill**

Published Aug. 21, 2022 Updated June 21, 2023

[Leer en español](#)

Problem #3: Can be used for surveillance

- Who defines list of target hashes?
- How much do you trust them?
- Since hashes are one-way, can't be verified...

Targeted Facial Recognition



Original Image



Target



Manipulated Image

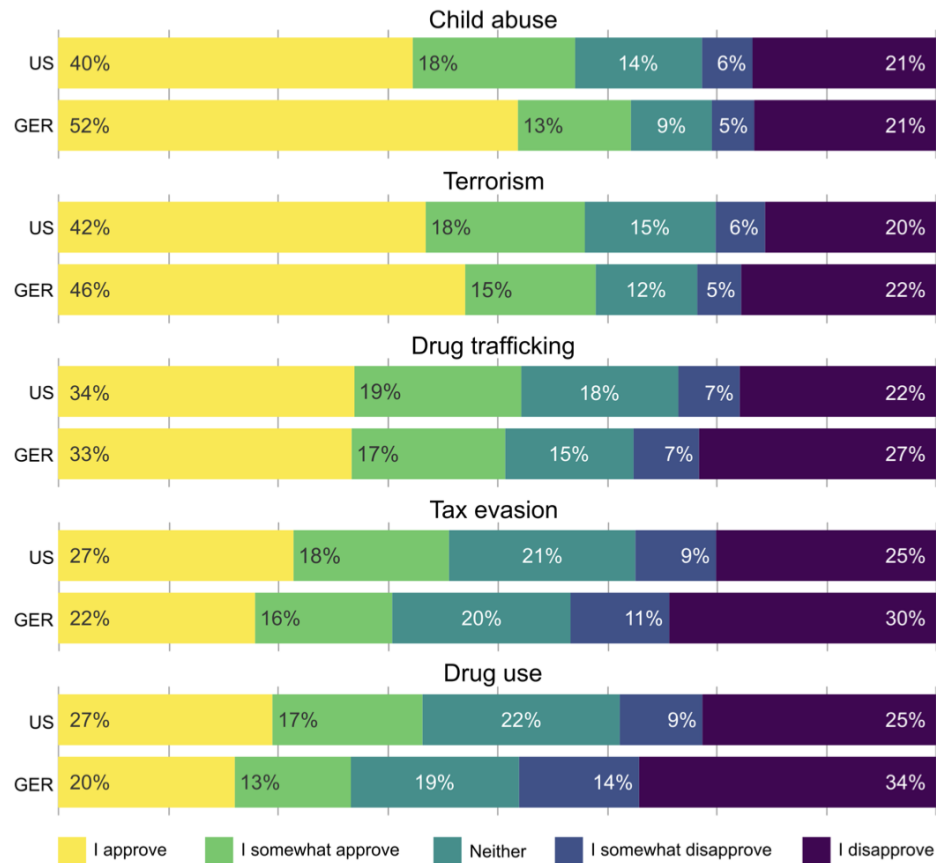
Discussion

- Should companies be allowed to use client-side scanning?
- Should companies be required to use client-side scanning?
- Does it matter if checks are done locally vs sending hash to server?
- Does it matter whether they use perceptual hashing vs modern AI moderating?

Client-Side Scanning

- in 2021, Apple announced plan to use client-side scanning to scan content stored on devices
 - huge backlash, Apple backed off
- periodic efforts by governments (e.g. EU) to require

Current Attitudes



Strong correlation between attitudes and trust in governments/law enforcement/companies