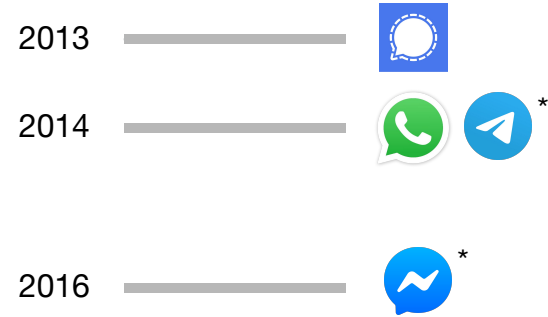
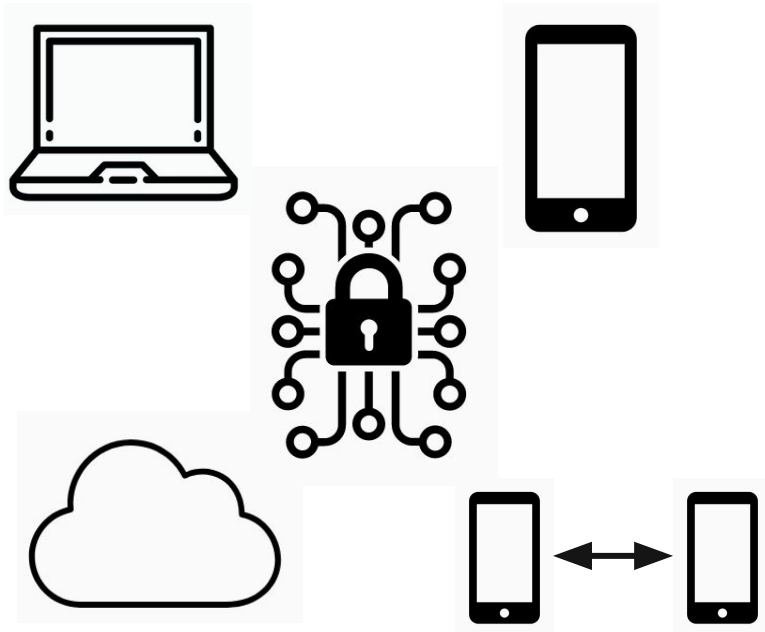


Perceptual hashing based client-side scanning: a well-intended but flawed solution

Shubham Jain, in collaboration with Ana-Maria Crețu, Antoine Cully, Yves-Alexandre de Montjoye
Imperial College London

Privacy Research Day 2023@CNIL

Encryption is great for privacy and security



> 2B users



* E2EE chats not yet used as default option

But restricts the functioning of law enforcement agencies

- The widespread adoption of E2EE might provide a shield for illegal activity.
- It makes it harder to detect illegal content, including child sexual abuse material (CSAM).

“Organized crime, terrorists and child abusers are all drawn to devices and communication platforms that are designed to be technically impossible for law enforcement to lawfully access.”

“Every day we see criminals using encryption to facilitate their crimes.”

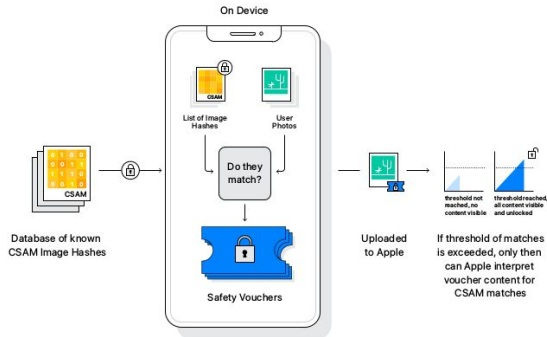
Catherine de Bolle, executive director of Europol and Cyrus R. Vance, Jr. district attorney of New York County, New York.

Perceptual hashing-based client-side scanning (PH-CSS): a privacy-preserving solution to detect known illegal content

Update as of September 3, 2021: Previously we announced plans for features intended to help protect children from predators who use communication tools to recruit and exploit them and to help limit the spread of Child Sexual Abuse Material. Based on feedback from customers, advocacy groups, researchers, and others, we have decided to take additional time over the coming months to collect input and make improvements before releasing these critically important child safety features.

Expanded Protections for Children

At Apple, our goal is to create technology that empowers people and enriches their lives — while helping them stay safe. We want to help protect children from predators who use communication tools to recruit and exploit them, and limit the spread of Child Sexual Abuse Material (CSAM).



unicef
Office of Research - Innocenti

Encryption, Privacy and Children's Right to Protection from Harm

This document has not been adopted or endorsed by the European Commission and is intended as a basis for discussion. It may not be shared further without permission of the European Commission services.

Technical solutions to detect child sexual abuse in end-to-end encrypted communications

Shaping Europe's digital future

Creating a better Internet for kids

The strategy for a better Internet for children provides actions to empower young people as they explore the digital world.

UK Parliament

Parliamentary Bills

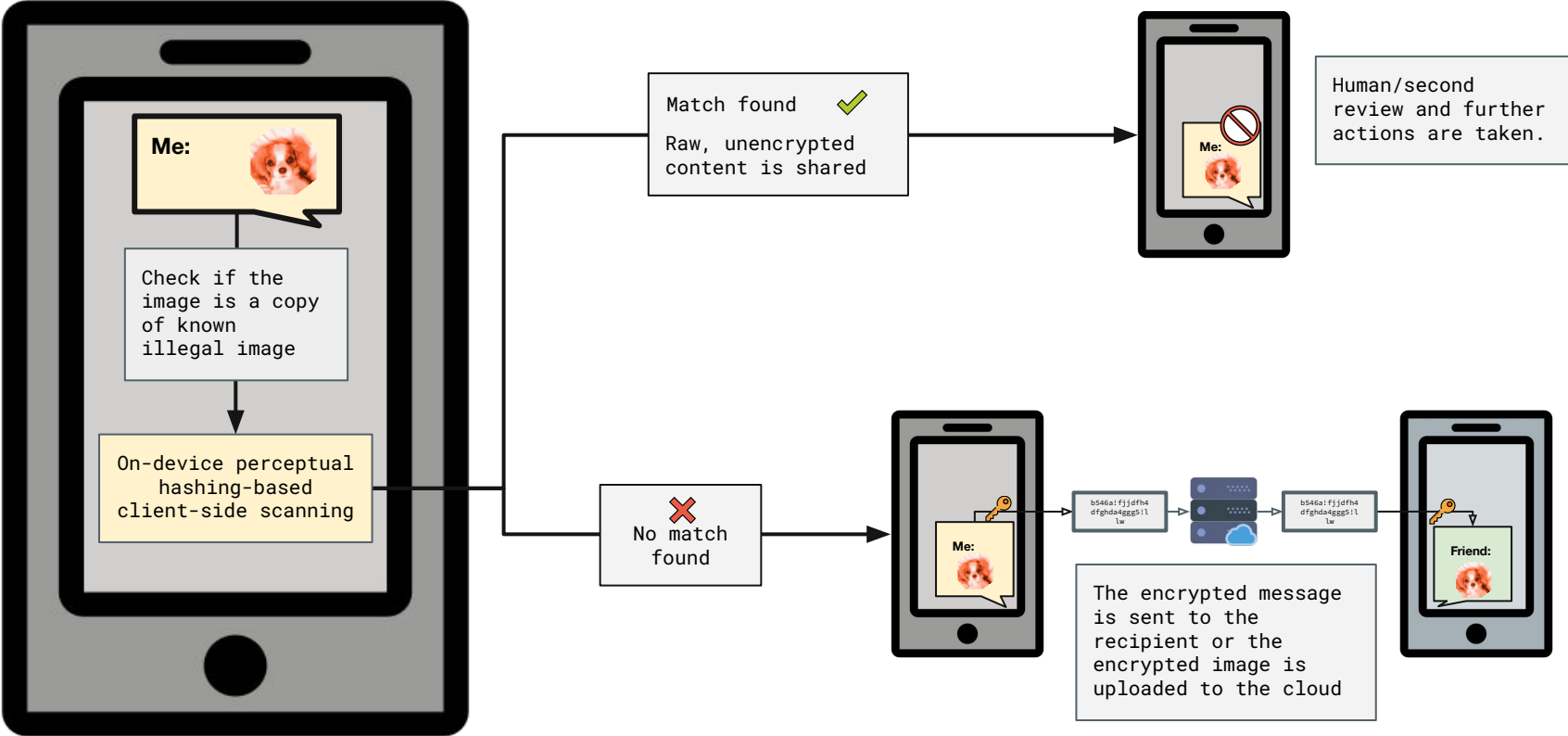
UK Parliament > Business > Legislation > Parliamentary Bills > Online Safety Bill

Online Safety Bill

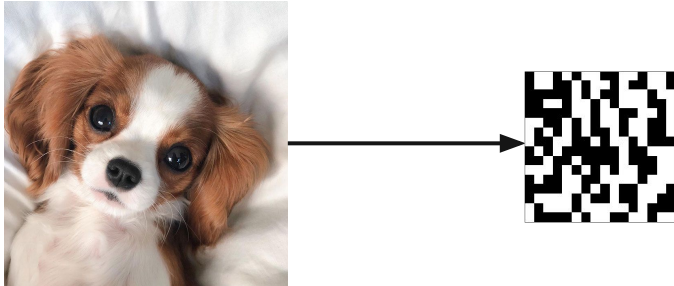
Government Bill

Originated in the House of Commons, Sessions 2021-22, 2022-23
Last updated: 27 May 2022 at 10:05

Perceptual hashing-based client-side scanning (PH-CSS)



Overview of perceptual hashing



- Used for **image copy detection**, i.e., to check if two images are copies of each other, edited or exact.
- It converts an image to a fingerprint, such that **similar images have similar fingerprints**.
- Current SOTA perceptual hashing algorithms are **trained deep neural networks** (e.g., Apple's neuralhash).

Image copy detection using perceptual hashing

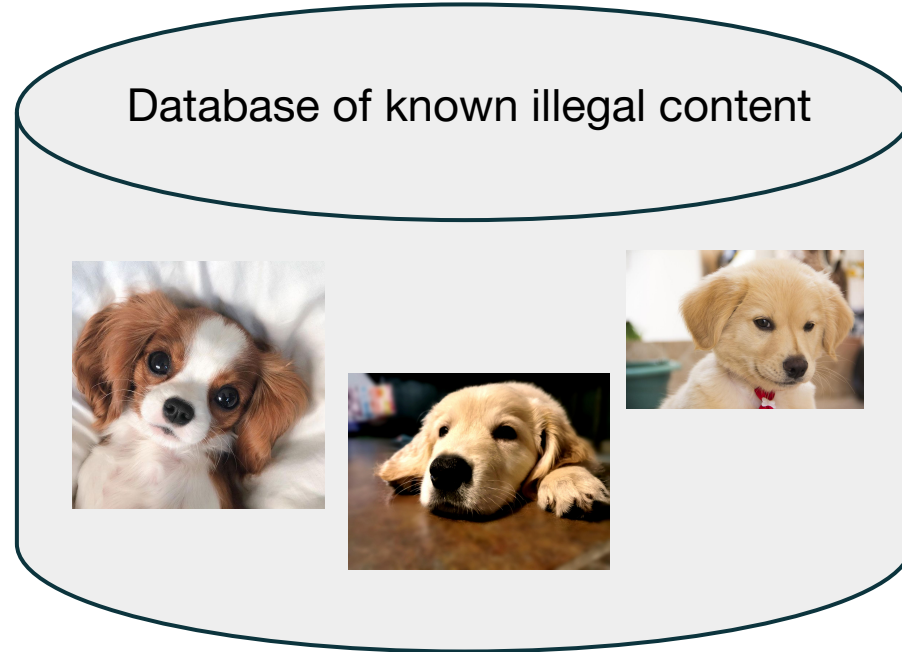
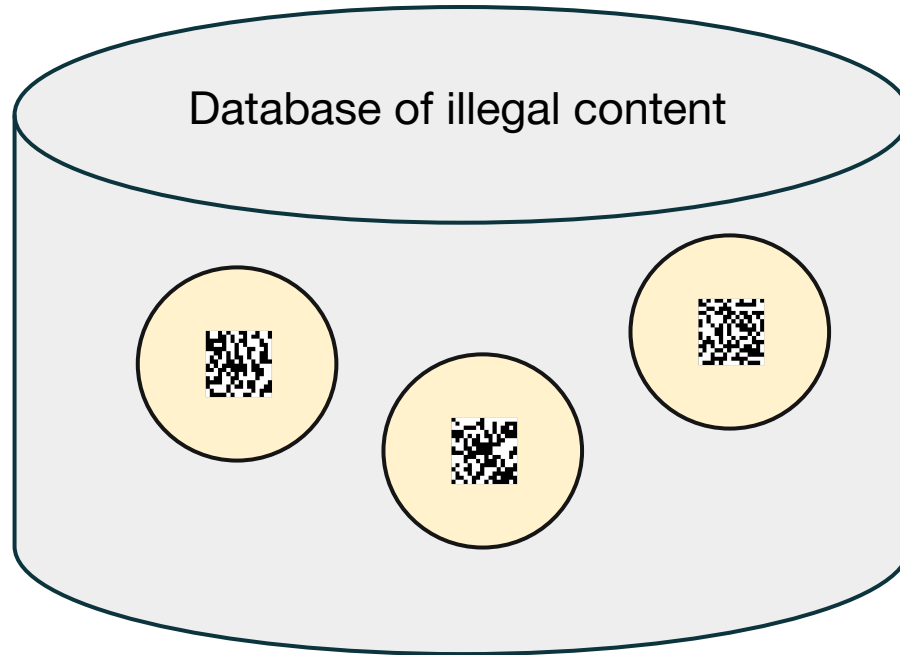
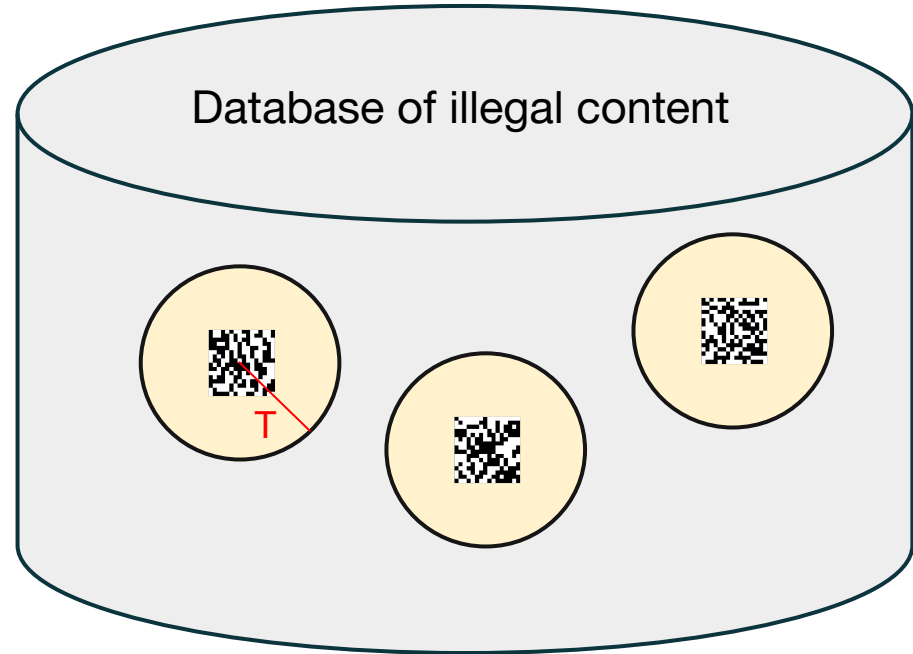
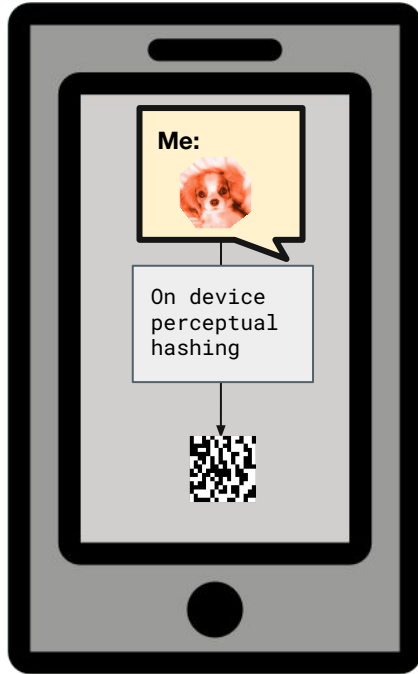


Image copy detection using perceptual hashing

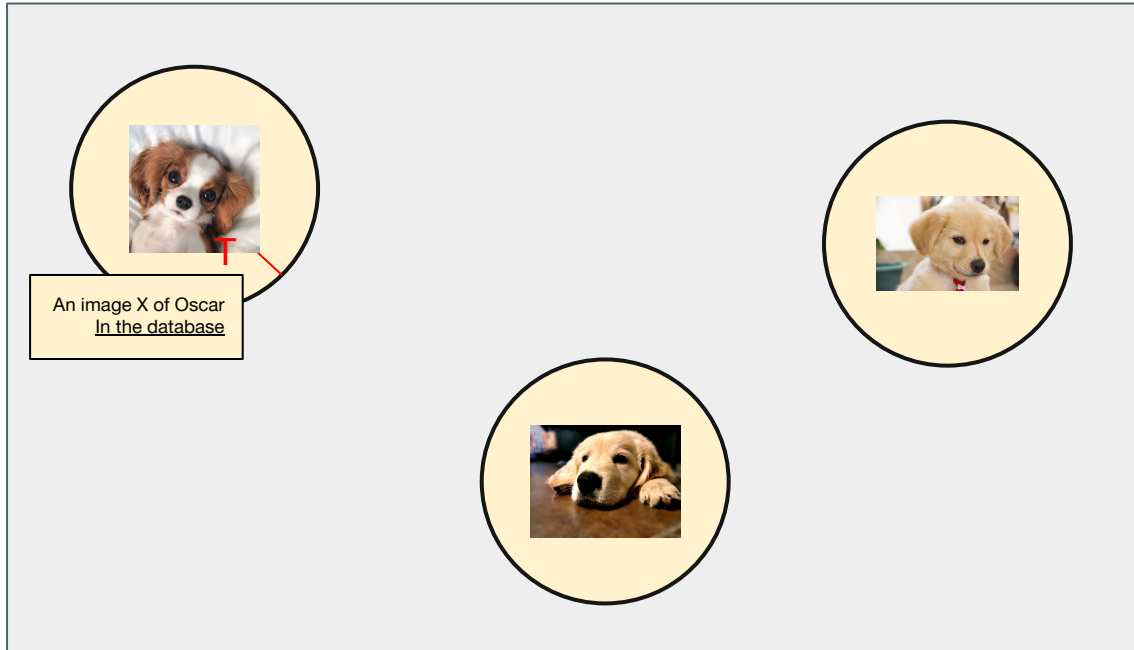



On-device PH-CSS



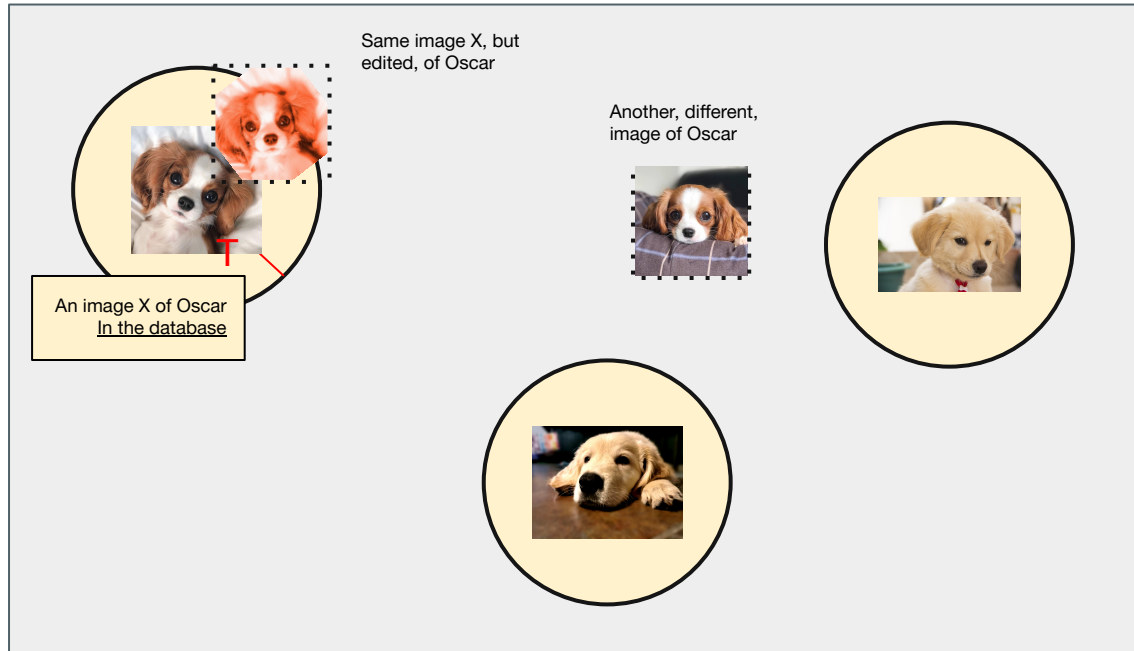
The detection threshold T sets the distance below which a fingerprint is deemed a "match"


A visual guide to understand PH-CSS




 Image in the database

A visual guide to understand PH-CSS



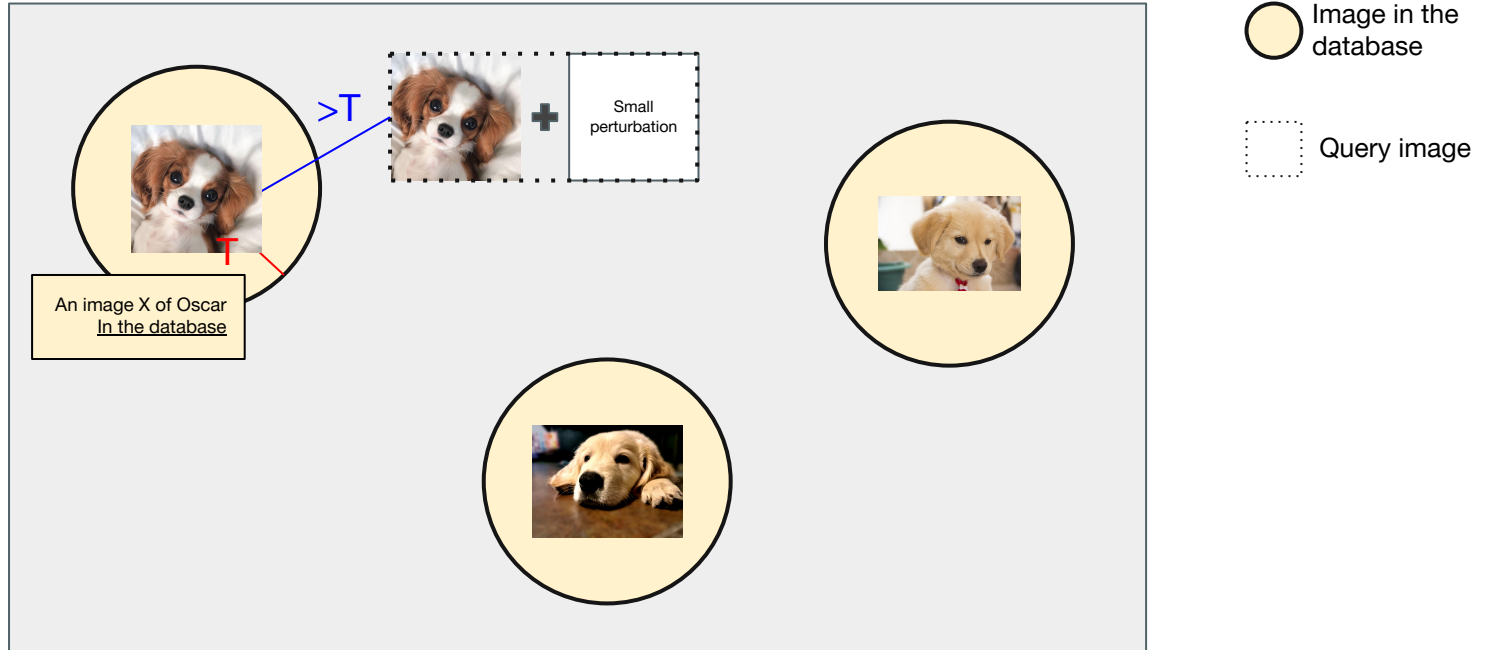
 Image in the database

 Query image



PH-CSS is vulnerable to detection avoidance attacks

Detection avoidance attack



>99.9%

Images¹ can be modified successfully using our attack

...for five popular hashing algorithms

...and a broad range of detection thresholds




Original image



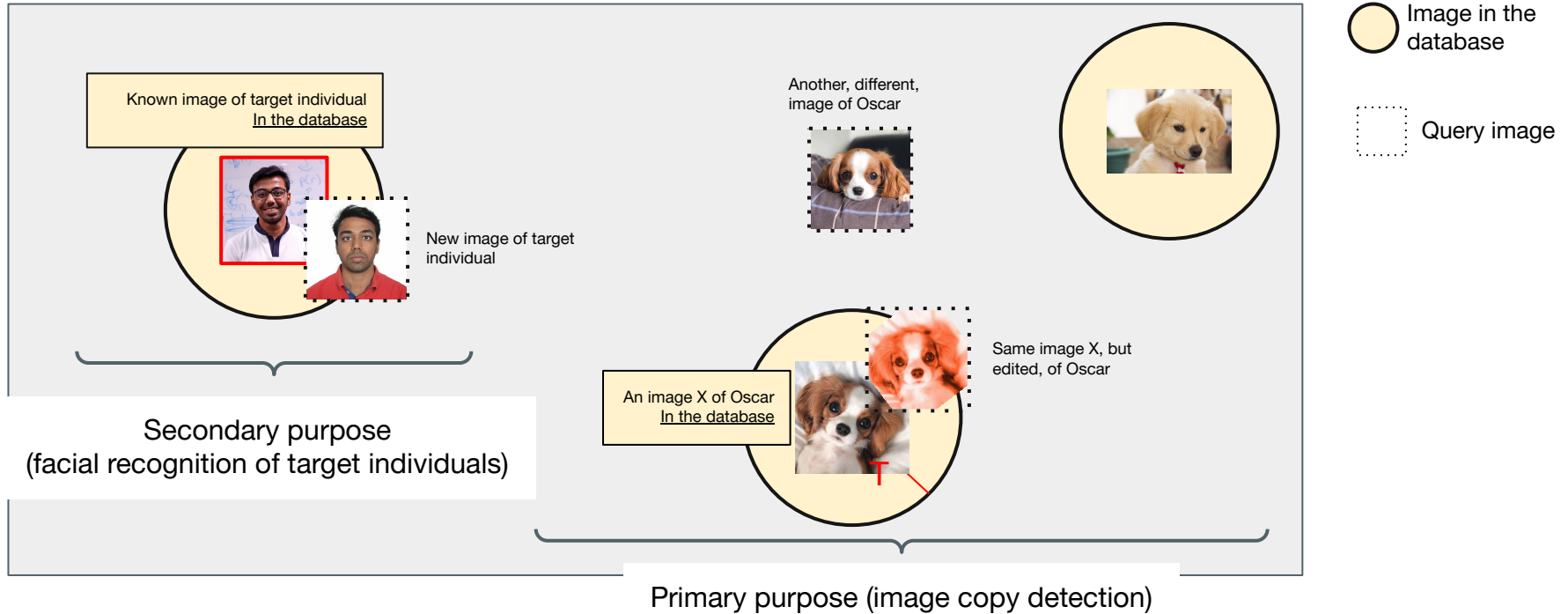
Modified image: PDQ, T=70

¹ ImageNet dataset with duplicates removed

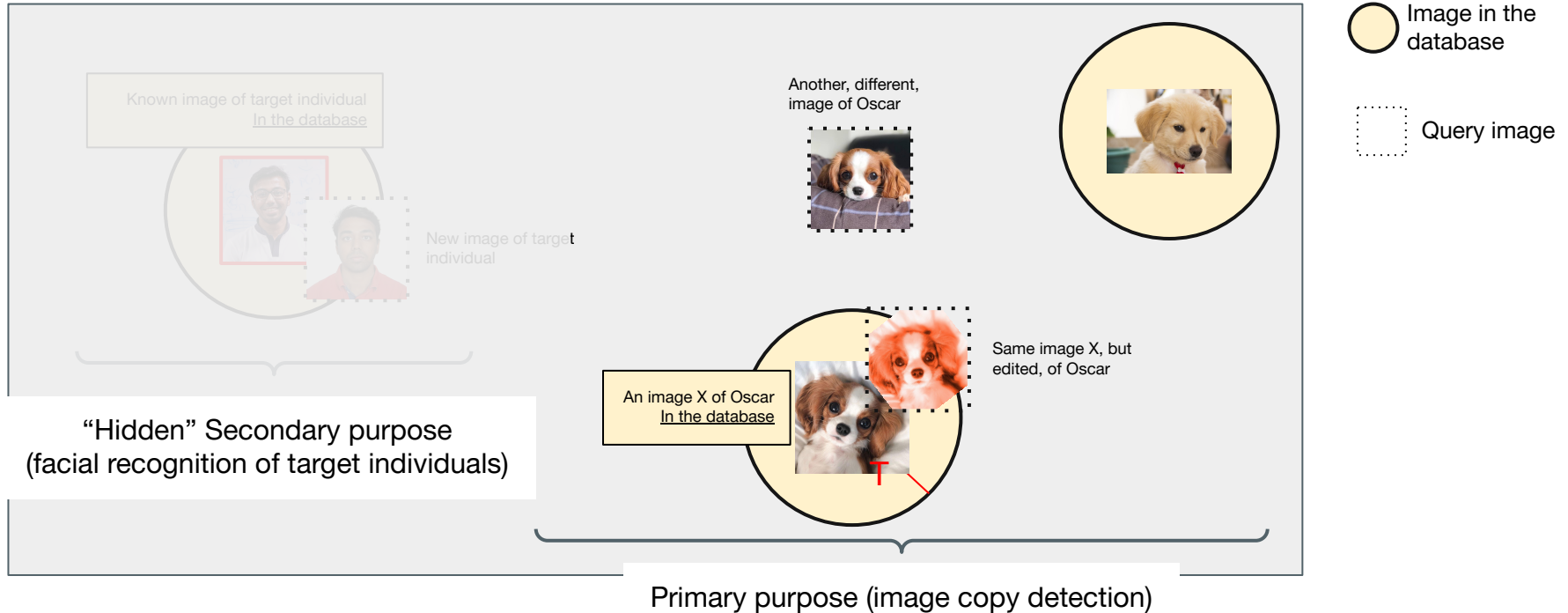


PH-CSS can be built with a “hidden”
secondary purpose for surveillance

Dual-purpose perceptual hashing algorithm



Dual-purpose perceptual hashing algorithm



Claims on limited capabilities of PH-CSS can be challenged

*“Hashing can be used to fingerprint digital files in their entirety but not individual components within them. Therefore, **it is not suitable to recognise whether the same component (e.g. object or person in an image) may be included in two different files** unless the entirety of both files is very similar (e.g. depicting the same building at very similar angles, with very similar lighting, same environment/background, etc.)”*

Overview of Perceptual Hashing Technology¹, OfCom, United Kingdom.

1. Proponents of PH-CSS claim that it is a technology with limited capabilities.
2. We show that it can be built with a hidden secondary purpose without compromising on performance on primary task.
3. Our results show PH-CSS can be built to turn millions of devices into tools of surveillance.

¹ Overview of Perceptual Hashing Technology by OfCom, United Kingdom (Nov 22 2022). <https://www.ofcom.org.uk/research-and-data/online-research/overview-of-perceptual-hashing-technology>

Conclusion

1. Perceptual Hashing-based Client-Side Scanning (PH-CSS) is proposed as a privacy-preserving solution to detect illegal content.
2. We show that PH-CSS might not be a robust solution as an image can almost always be modified to evade detection in the black-box setup.
3. We show that PH-CSS can be developed with a hidden secondary purpose of facial recognition of target individuals, thus turning millions of devices into tools of surveillance.

Thank you!



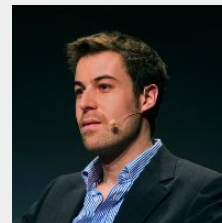
@shubhamjain0594



@anamariacretu5



@cullyantoine



@yvesalexandre