Total recall or how to classify and recognize phishing websites via image recognition
Where Carl Hauser meet Douglas Quaid

Team CIRCL - TLP:WHITE
@circl.lu @adulau
PTS 2019 - rumps

July 2, 2019
The Dirty Reality of Website Screenshots

- CIRCL does a lot of automation on malicious or suspicious websites.
- 10000+ onion hidden services are crawled\(^1\) on a daily basis generating more than 10k screenshots per day.
- 2000+ phishing websites\(^2\) are crawled and generating associated screenshots.
- We were sitting on a huge amount of data but what can we do this?

\(^1\)https://github.com/CIRCL/AIL-framework
\(^2\)https://www.misp-project.org/
The Road to Open Source Tooling for Image Classification
First step

• Everything started with the internship of Vincent Falconi (INSA/Lyon).

• **Carl Hauser**[^1] was the *first step to evaluate* the known image distance, classification.

• Tooling was required for classification so a tool was created:
  • **visjs_classificator**[^2] - Classificator for pictures matching and clustering. Fast and visual.

[^1]: [https://github.com/CIRCL/carl-hauser](https://github.com/CIRCL/carl-hauser)
[^2]: [https://github.com/Vincent-CIRCL/visjs_classificator](https://github.com/Vincent-CIRCL/visjs_classificator)
Second step

- **Douglas Quaid** - Open source software for image correlation, distance and analysis⁵.

- **Open dataset**⁶ of phishing websites screenshot including labeled classification.

---

⁵ https://github.com/CIRCL/douglas-quaid
⁶ https://www.circl.lu/opendata/circl-phishing-dataset-01/
https://www.circl.lu/opendata/datasets/circl-phishing-dataset-01/
Contact

- info@circl.lu
- https://www.circl.lu/
- OpenPGP fingerprint: CA57 2205 C002 4E06 BA70 BE89 EAAD CFFC 22BD 4CD5