Clustering large amount of emails with Minhash: an open-source Locality Sensitive Hash

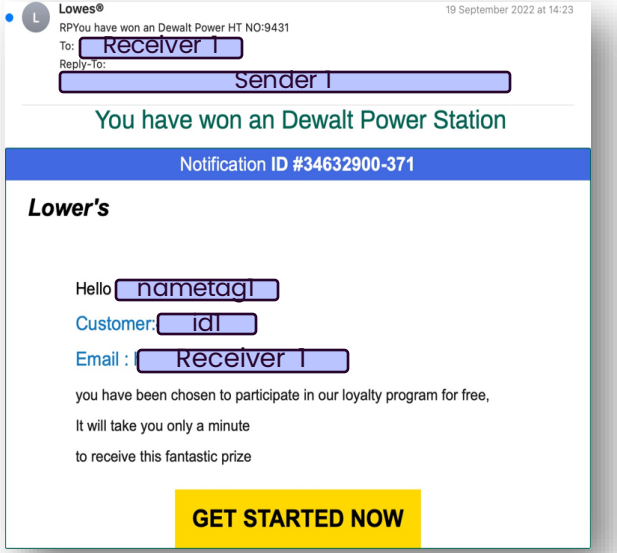
About us

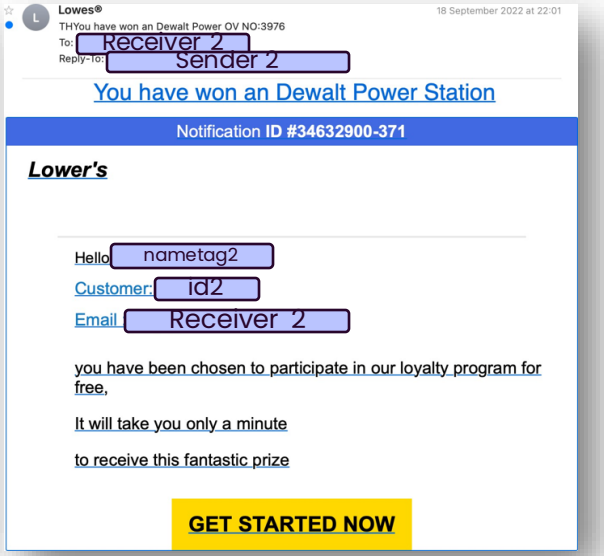
- PhD from Université de Lille (2022) focused on optimization / machine learning
- Research Engineer at Vade (vadesecure.com)
- Focus on worldwide email cybersecurity
  - >1 billion email box protected

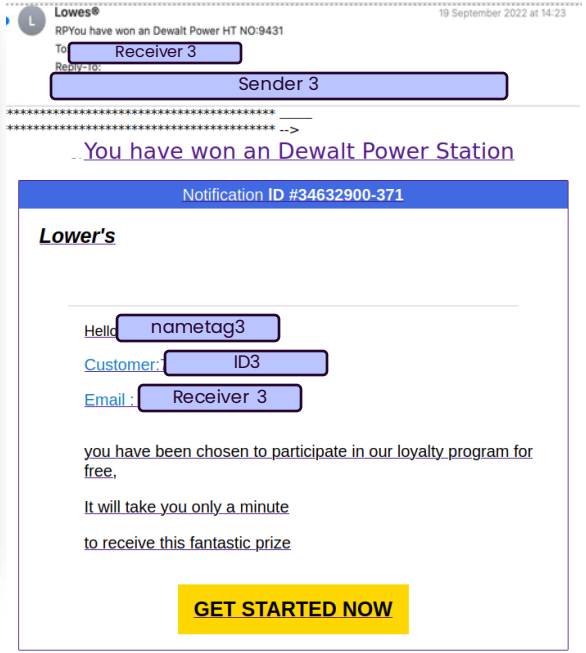
Introduction

- 400 billions emails are sent every day
- >90% of emails are cyberattacks
- 2 types of attacks:
  - targetted attacks (personalized) and
  - **campaign attacks** (broad and generic).
- Emails from campaign attacks are very similar, but usually not exactly the same

Introduction & objectives





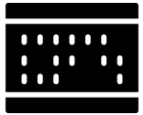


Context: constraints of use

General constraints

- Email must be treated as streaming flow of data

- Less than 10ms per emails



- Store as few bytes as possible per email

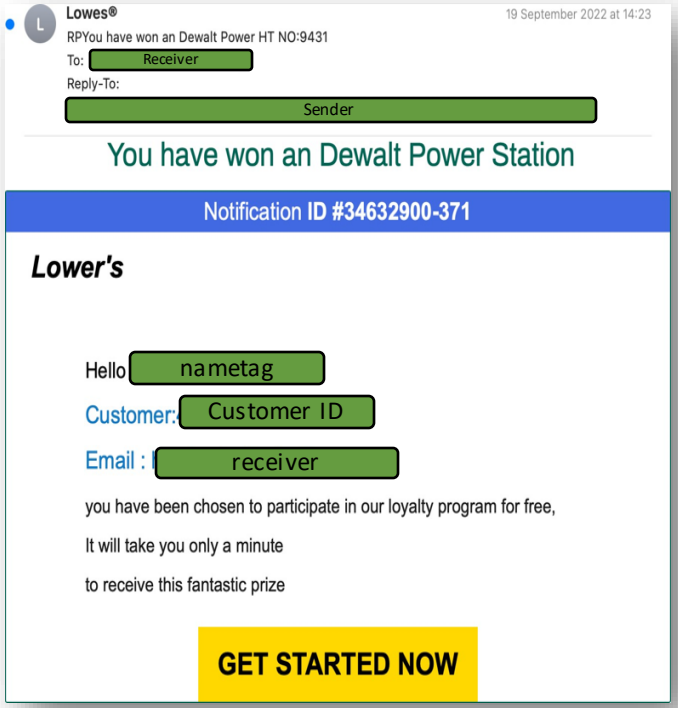
Plan

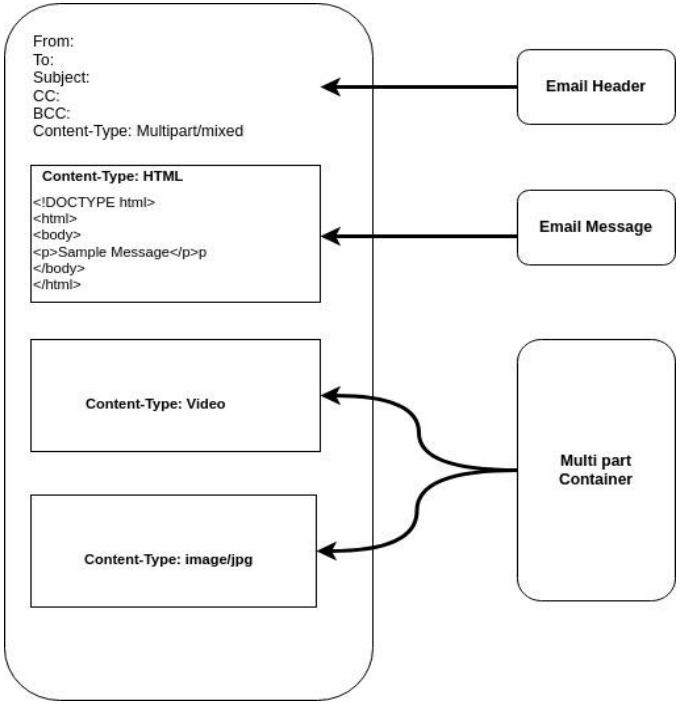
1. Preprocessing of the email
2. Matching emails with MinHash
3. Experiments
4. Conclusion

Plan

1. **Preprocessing of the email**
2. Matching emails with MinHash
3. Experiments
4. Conclusion

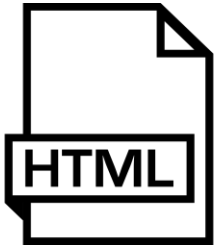
Pre-processing: What's in an email?





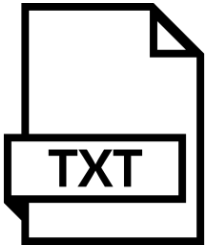
Pre-processing: Extracting displayed content

text/html



or

text/plain



Emails: Same visual, Different underlying content

HTML code

```
<!DOCTYPE html>
<html>
<title>Online HTML Editor</title>

<head>
</head>

<body>
    <h1>Hello World</h1>
</body>

</html>
```

Visual rendering

Hello World

Emails: Same visual, Different underlying content

HTML code

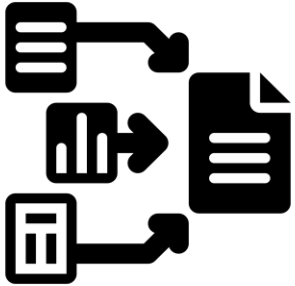
```
<!DOCTYPE html>
<html>
<title>Online HTML Editor</title>

<head>
</head>

<body>
    <h1>Hello World</h1>
    <div style="display: none; max-height: 0px; overflow: hidden;">&nbsp;&zwnj;&nbsp;&zwnj;&nbsp;&zwnj;&nbsp
        ;&zwnj;&nbsp;&zwnj;&nbsp;&zwnj;&nbsp;&zwnj;&nbsp;&zwnj;&nbsp;&zwnj;&nbsp;&zwnj;&nbsp;&zwnj;&nbsp;&zwnj
        ;&nbsp;&zwnj;&nbsp;&zwnj;&nbsp;&zwnj;&nbsp;&zwnj;&nbsp;&zwnj;&nbsp;&zwnj;&nbsp;&zwnj;&nbsp;&zwnj;&nbsp
        ;&zwnj;&nbsp;&zwnj;&nbsp;&zwnj;&nbsp;&zwnj;&nbsp;&zwnj;&nbsp;&zwnj;&nbsp;&zwnj;&nbsp;&zwnj;&nbsp;&zwnj
        ;&nbsp;&zwnj;&nbsp;&zwnj;&nbsp;&zwnj;&nbsp;&zwnj;&nbsp;&zwnj;&nbsp;&zwnj;&nbsp;&zwnj;&nbsp;&zwnj;&nbsp
        ;&zwnj;&nbsp;&zwnj;&nbsp;&zwnj;&nbsp;&zwnj;&nbsp;&zwnj;&nbsp;&zwnj;&nbsp;&zwnj;&nbsp;&zwnj;&nbsp;&zwnj
        ;&nbsp;&zwnj;&nbsp;&zwnj;&nbsp;&zwnj;&nbsp;&zwnj;&nbsp;&zwnj;&nbsp;&zwnj;&nbsp;&zwnj;&nbsp;&zwnj;&nbsp
        ;&zwnj;&nbsp;&zwnj;&nbsp;&zwnj;&nbsp;&zwnj;&nbsp;&zwnj;&nbsp;&zwnj;&nbsp;&zwnj;&nbsp;&zwnj;&nbsp;&zwnj
        ;&nbsp;&zwnj;&nbsp;&zwnj;&nbsp;
    </div>
    <div class="hidden" style="visibility: hidden;">Lorem ipsum dolor sit amet, consectetur adipiscing elit,
        sed do eiusmod tempor incididunt ut labore et dolore magna aliqua</div>
    <h1 style="position: absolute; left: -999px;">Lorem ipsum<img src="img_girl.jpg">Lorem ipsum</h1>

</body>

</html>
```

Visual rendering

Hello World

Pre-Processing Email Content: Normalization

- Transforms email content into a single canonical form
- Reduces the randomness, i.e. noise and variation caused by template variables and obfuscation



- Convergence to a standardized representation of the email content
- Avoid altering the graphical rendering
- Improve document comparison capabilities

Normalization features: Plain-Text

1. Remove "Unnecessary" Spaces
2. Replace "URLS" by Placeholders
3. Lowercase content
4. Remove Invisible Characters

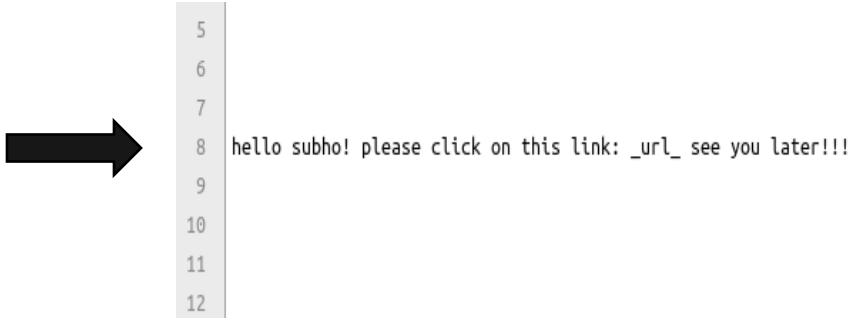
Noisy Plain Text

```
Hello Subho!


Please      Click

on this LINK:

http://bitly.ws/zDQB

SeE YoU LaTeR!!!

```

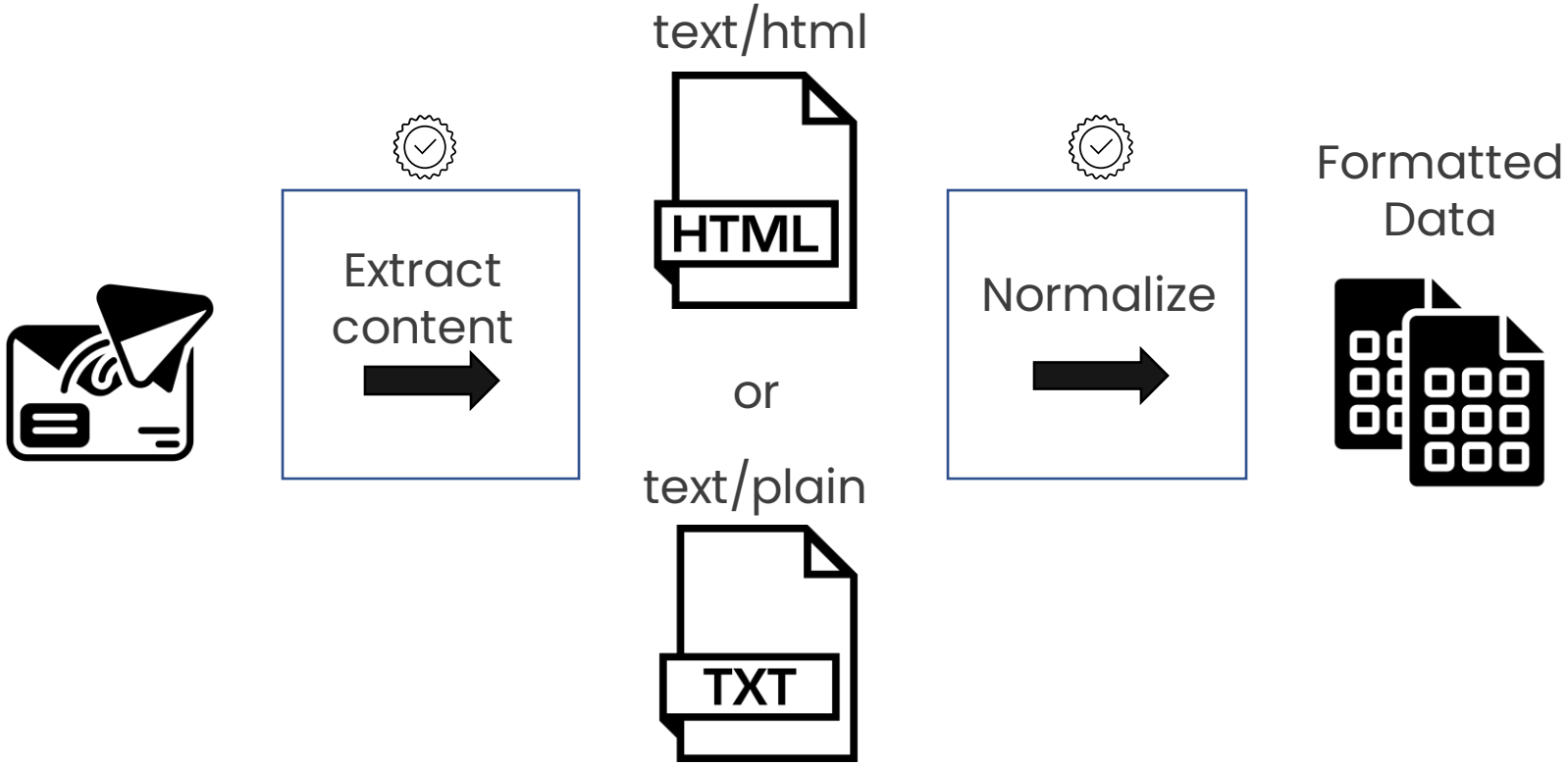
Normalization features: Plain-Text

Normalized Plain Text

1. Remove "Unnecessary" Spaces
2. Replace "URLS" by Placeholders
3. Lowercase content
4. Remove Invisible Characters



Pipelines: pre-processing



Plan

1. Preprocessing of the email
2. **Matching emails with MinHash**
3. Experiments
4. Conclusion

Detection of similarity

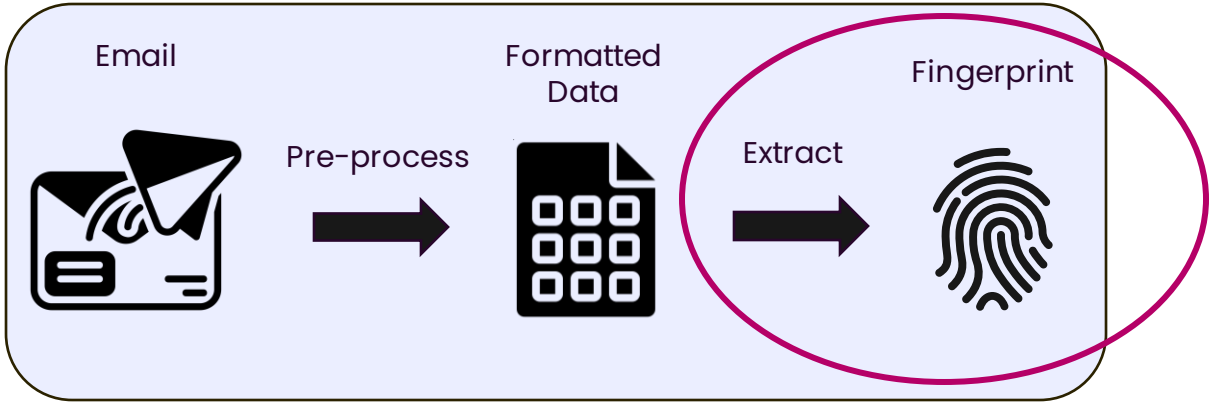
Storage

- Storing email is too complicated (price, infrastructure, GDPR)
- We can store small representation of an email: **fingerprint**
- Similar emails should have similar fingerprints
- **Problematic: How to generate a fingerprint?**

Comparison

- Comparing pairs of email or fingerprint is unfeasable: for 1 million emails, we would need 1 trillion comparisons
- **Problematic: How to compare fingerprints?**

Extracting a fingerprint



How to generate fingerprints: Hash algorithm

- Cryptographic hash are commonly used hash algorithm
- Often used for security purpose (Sha-1, Sha-256, MD5, …)
- Small differences in input creates really different hashes

Document	Vade is a cybersecurity company.	Vade is a Cybersecurity company.
Crypto-hash (fingerprint)	□□■■■□□□	■■■□□■■■

LSH: Locality Sensitive Hashing

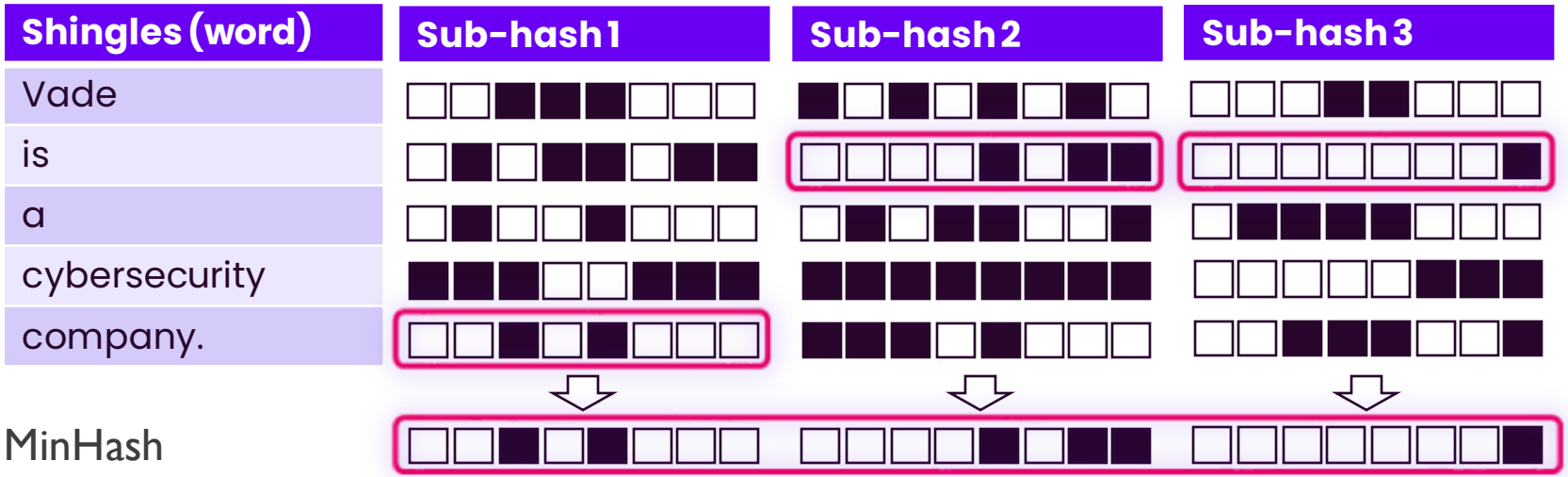
- Produce fingerprints that:
  - Are similar for similar content
  - Will create **collisions*** (contrary to crypto hash)

- Used extensively for HTML pages comparison as their structure can be comparable to emails visual content.

****collision**: *two documents can end up having the same hash*

LSH: MinHash

Content to hash: "Vade is a cybersecurity company."

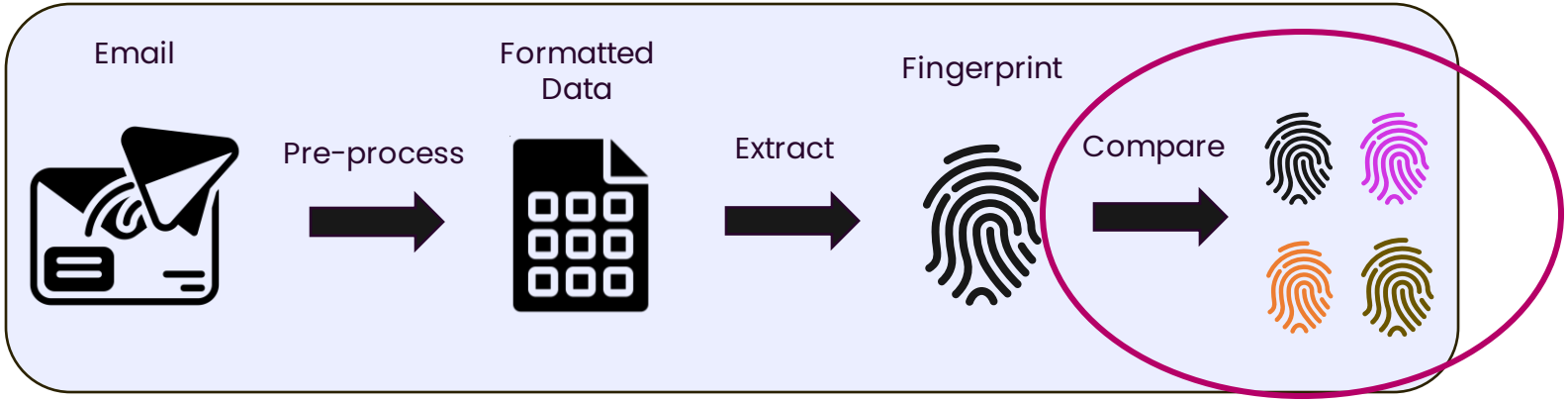


Minhash is the concatenation of the minimum hash for each subhash

LSH: MinHash

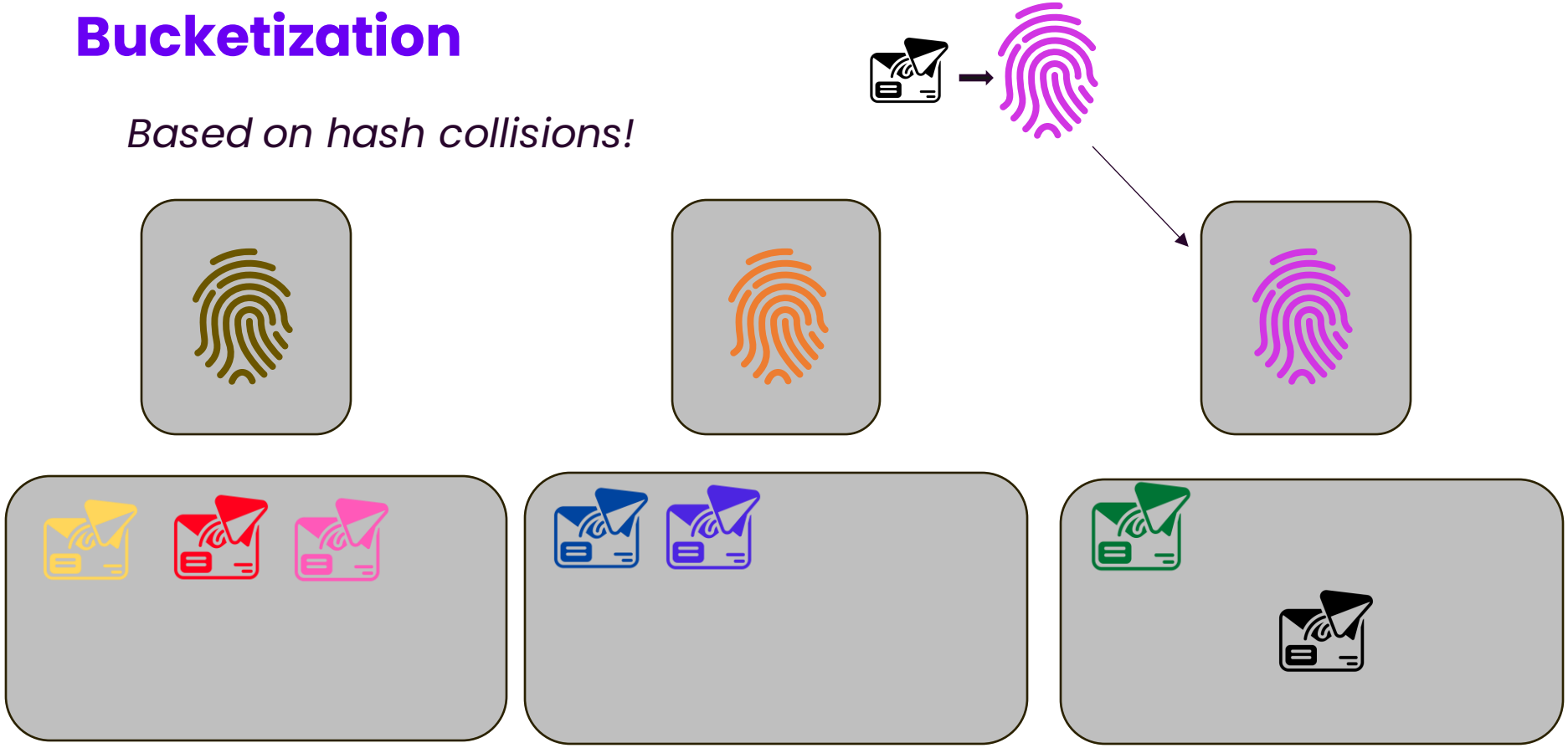
Document	Vade is a cybersecurity company.	Vade is a Cybersecurity company.
Crypto-hash (fingerprint)	☐☐■■■☐☐☐	■■■☐☐■■■
MinHash	☐☐■☐■☐☐☐	☐☐■☐■☐☐☐

Comparing fingerprints



Bucketization

Based on hash collisions!



Plan

1. Preprocessing of the email
2. Matching emails with MinHash
3. **Experiments**
4. Conclusion

Hash functions of interest

MinHash: A free of use LSH. It was initialy developped for Altavista (Web Browser). Used in many companies including Netflix.

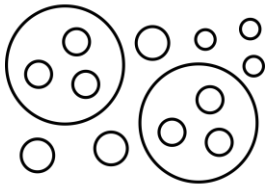
SimHash: The most known LSH. It is developped and used by google. It is patented

Sha256: Cryptographic hash that will be used as a floor. It is equivalent to testing a strict equality between the emails.

Experimental setup: Algorithm parameters

Parameter	Values
Hash size	From 8 bytes to 512 bytes (when mutable)
Sub-hash functions	Spooky, farm, city, djb2, sdbm, lose lose,

Experimental setup: Corpus



- Collected 15 000 emails
- Manually found clusters of visually similar emails. 1000 emails clustered, 75 clusters.
- 2 emails in the same cluster are considered duplicate, otherwise they are not.

- 7 Languages: French, English, Japanese, Portuguese, Polish, German, Spanish

Experimental setup : Metrics

- **TP:** two mails in the same cluster share 1 bucket
- **FP:** two mails in different clusters share at least 1 bucket
- **TN:** two mails in different clusters don't share any bucket
- **FN:** two mails in the same cluster don't share any bucket

- **Jaccard Index (Threat score):**

$$\frac{TP}{TP + FP + FN}$$

Process duration – Hashing (HTML)

sha256

Range	Ratio
dur < 1ms	100%
1ms <= dur < 10ms	0.0%
10ms <= dur < 100ms	0.0%
100ms <= dur < 250ms	0.0%

MinHash

Range	Ratio
dur < 1ms	53.5%
1ms <= dur < 10ms	45.7%
10ms <= dur < 100ms	0.8%
100ms <= dur < 250ms	0.0%

Based on an experiment done on > 1 million emails.

Experimental results: Impact of normalization

Pipelines	Jaccard index
MinHash	47%
Normalization + MinHash	60%

Experimental results: Outstanding pipelines

Pipelines	Hash Size (bytes)	Jaccard index	Comment
MinHash	**64**	**76%**	**Best pipeline overall**
SimHash	16	60%	Best pipeline with SimHash
Sha256	320	52%	Best pipeline with Sha256

Plan

1. Preprocessing of the email
2. Near duplicate detection
3. Experiments
4. **Conclusion**

Conclusion: real email detected with this process

Total.Energies.

Cher(e) client(e),

Nous avons de bonnes nouvelles pour vous ! **L'Etat a débloqué 3 milliards d'euros** pour équiper les foyers dans le cadre du plan de **rénovation énergétique 2023**. L'objectif est de fournir aux propriétaires éligibles **jusqu'à 10 000 Euro d'aides par foyer**, sans aucune démarche administrative.

Pour bénéficier de ces aides, il vous suffit de remplir le **formulaire** en ligne, sans aucun engagement. Vous pourrez alors bénéficier d'une prise en charge pour les propriétaires éligibles.

DEMANDEZ VOS AIDES

Jusqu'à 100% de prise en charge pour les propriétaires éligibles.

Nous sommes heureux de vous compter prochainement parmi nos clients et vous assurons que nous déployons toute notre énergie pour mieux vous servir.

Click here to unsubscribe.
Click here to view this email in your browser.

ENGIE.

Cher(e) client(e),

Nous avons de bonnes nouvelles pour vous !
L'Etat a débloqué 3 milliards d'euros pour équiper les foyers dans le cadre du plan de **rénovation énergétique 2023**.
L'objectif est de fournir aux propriétaires éligibles **jusqu'à 10 000 Euro d'aides par foyer**, sans aucune démarche administrative.

Pour bénéficier de ces aides, il vous suffit de remplir le **formulaire en ligne**, sans aucun engagement. Vous pourrez alors bénéficier d'une prise en charge pour les propriétaires éligibles.

DEMANDEZ VOS AIDES

Jusqu'à 100% de prise en charge pour les propriétaires éligibles.

Nous sommes heureux de vous compter prochainement parmi nos clients et vous assurons que nous déployons toute notre énergie pour mieux vous servir.

Conclusion: List of used tools

Normalization (HTML):
https://www.npmjs.com/package/sanitize-html

Normalization (PLAIN):
https://www.nltk.org

MinHash:
https://github.com/ekzhu/minhash-lsh (golang)
https://github.com/dgryski/go-minhash (golang)
https://github.com/chrisjmccormick/MinHash (python)
https://github.com/txje/c-minhash (C)