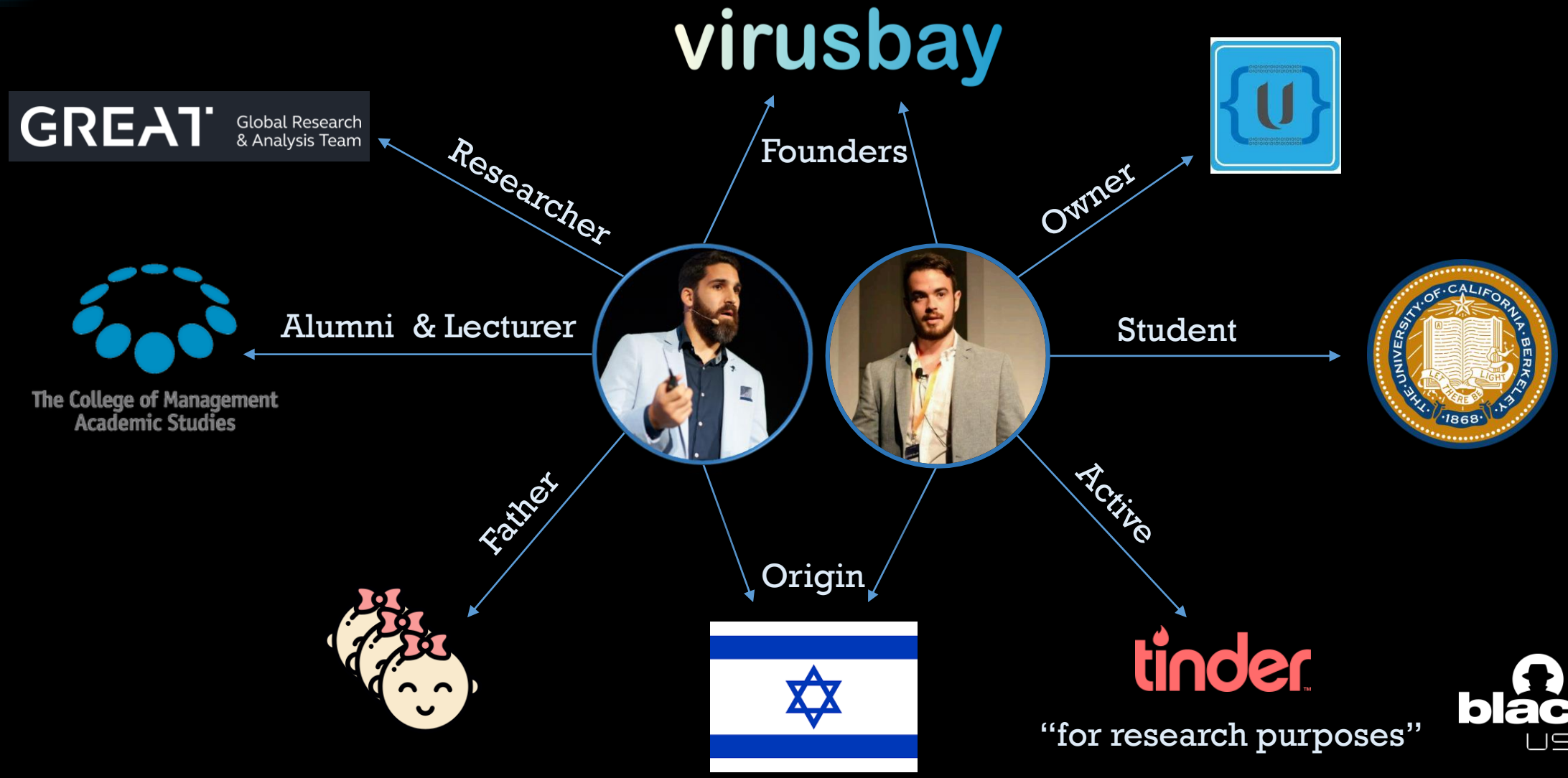
Dissecting Non-malicious Artifacts: One IP At A Time

Ido Naor (@idonaor1) **& Dani Goland** (@danigoland)

About the researchers



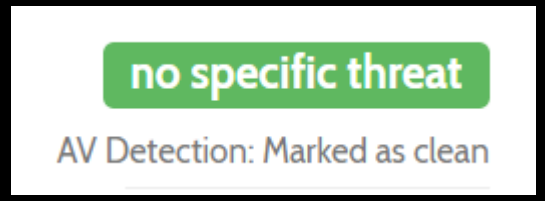
General concept

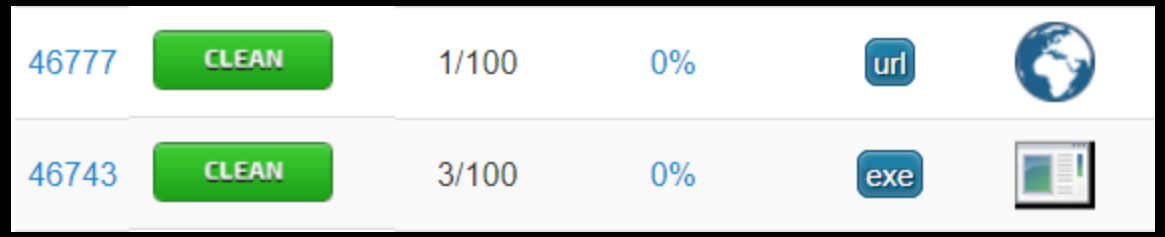
RESEARCH
FOCUS

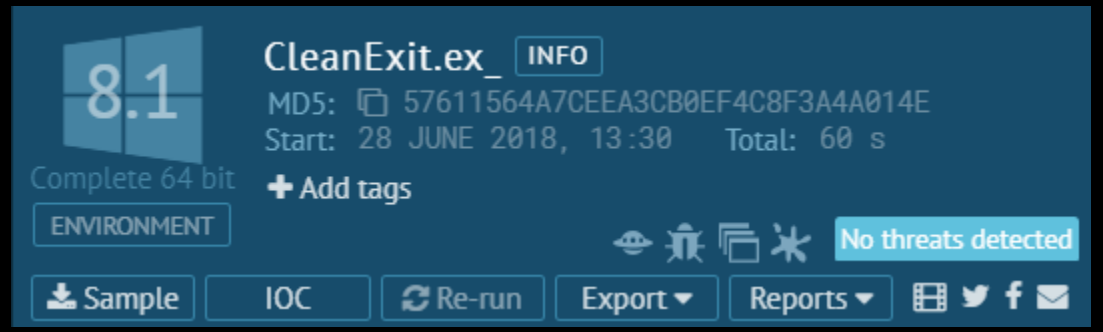
PRO Responsible
EDUCATE

Zero detection

Ad-Aware	Clean
AhnLab-V3	Clean
ALYac	Clean
Arcabit	Clean
Avast Mobile Security	Clean
Avira	Clean
Baidu	Clean
Bkav	Clean
ClamAV	Clean
Comodo	Clean
DrWeb	Clean
eScan	Clean







History Around The Subject

MALWARE SCANNING SERVICES CONTAINERS FOR SENSITIVE BUSINESS INFORMATION

by Michael Mimoso Follow @mike_mimoso April 5, 2017 , 1:01 pm

SINT MAARTEN—Malware scanning services could be the next listening outpost for criminals and nation-state attackers as more of these services such as VirusTotal are becoming containers for personal, business and even classified information because of some organizations' policy decision to upload every file, document and email.

Markus Neis, threat intelligence manager at Swisscom AG, this week joined the growing

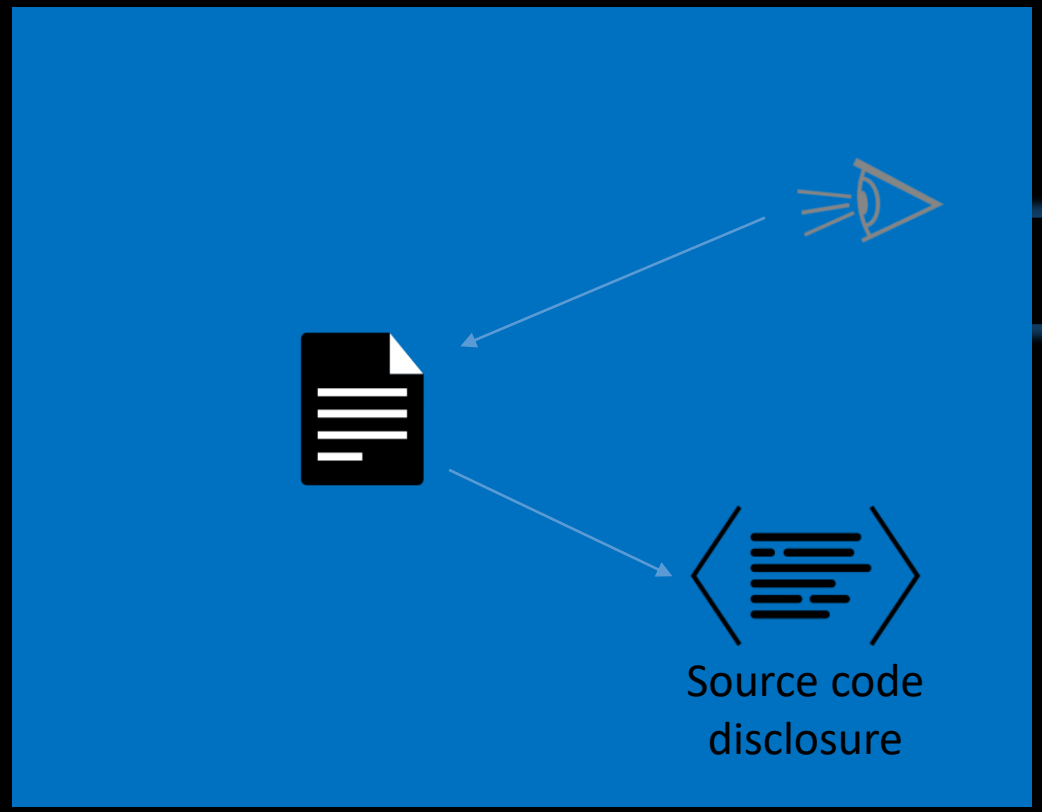
Every time you upload a sample

- Demonstrated data exfiltration via online sandboxes and malware repositories
- Initial malware(rocket) generates a file(satellite) from a template and incorporates the sensitive data inside.
- The satellite triggers the AV product and is uploaded to a sandbox
- Demonstrated retrieval via a Yara Rule search.

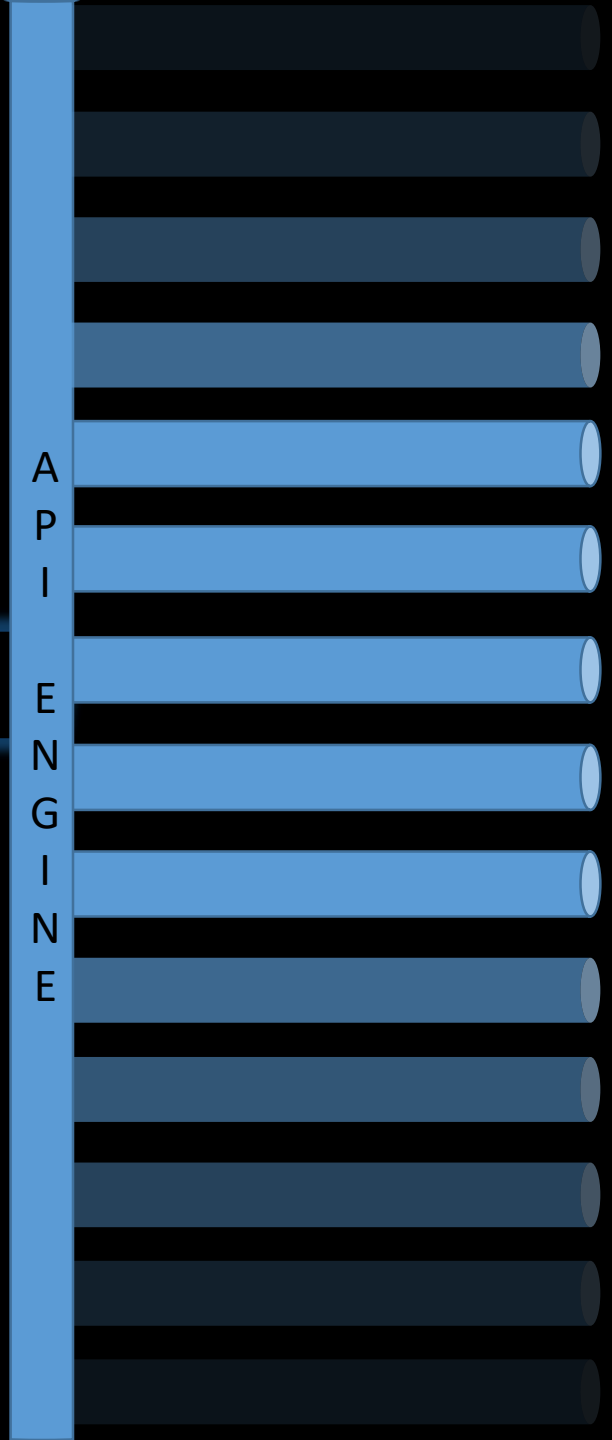
What were we searching

- Credentials / Keys
- Database artifacts – emails, passwords, PII, etc.
- Code segments
- Documents (Office/PDF/Visio and etc.)
- Compiled code (APKs, Jars and other apps)

Look & Feel

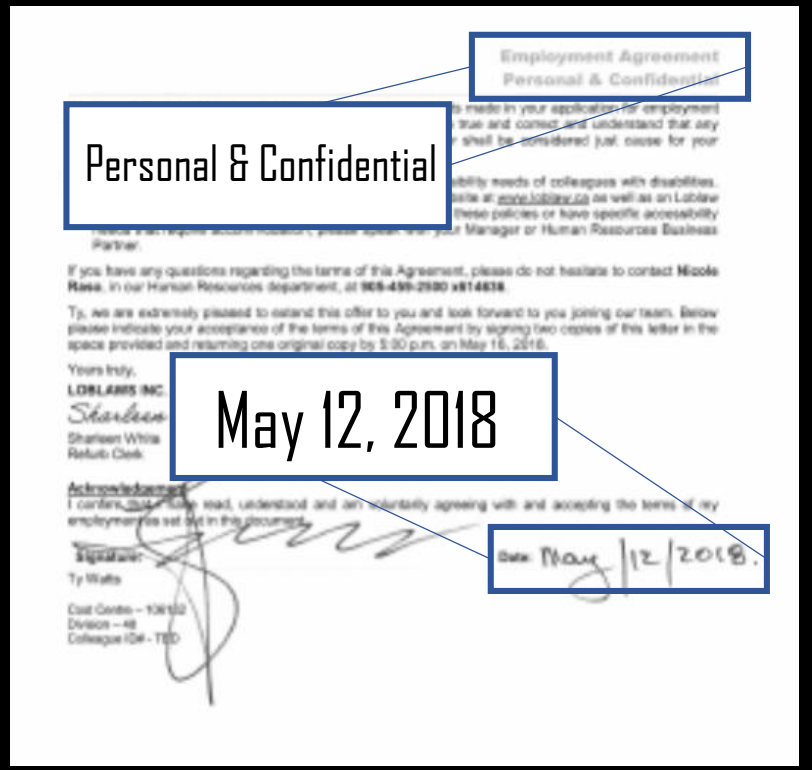


Machine
learning
engine



Proof of Concept

- One simple search: "message"
- Mail containing non-malicious artifacts
  - Options:
    - Mail uploaded by employee
    - Mail suspected by security product
    - Mail uploaded by 3rd party
- Example of company intellectual property being leaked due to suspected file verification



Research Scope

Starting points

- Paid/non-paid services
- Repositories
- Other

Expertise

- Yara
- Data Science
- Malware Research

Goal

Prove that data is being unwillingly exfiltrated from organizations, and that with simple tools – it can be stopped.

Step 1: Espionage Tool

- Code repositories / Open source
- Script/Paste repositories
- Malware repositories
- Multi-scanners (with private key or without)
- Online sandboxes
- Forums / Social platforms

Step 2: Espionage Tool

```
rule aws_api{
    strings:
        $a = /AKIA[0-9A-Z]{16}/
condition:
    all of them
}
rule twitter_api{
    strings:
        $a = "consumer_key" nocase
        $b = "consumer_secret" nocase
        $c = "access_token" nocase
condition:
    all of them
}
```

```
rule minidump_file : memory windows{
        strings:
                $header = "MDMP"
                $formatversion = {93 a7}
        condition:
                $header at 0 and
                $formatversion at 4
}
```

```
rule emails{
        condition:
        ((file_type contains "internet
email") or
(file_type contains "internet email
outlook")) and
new_file and positives < 1
}
```

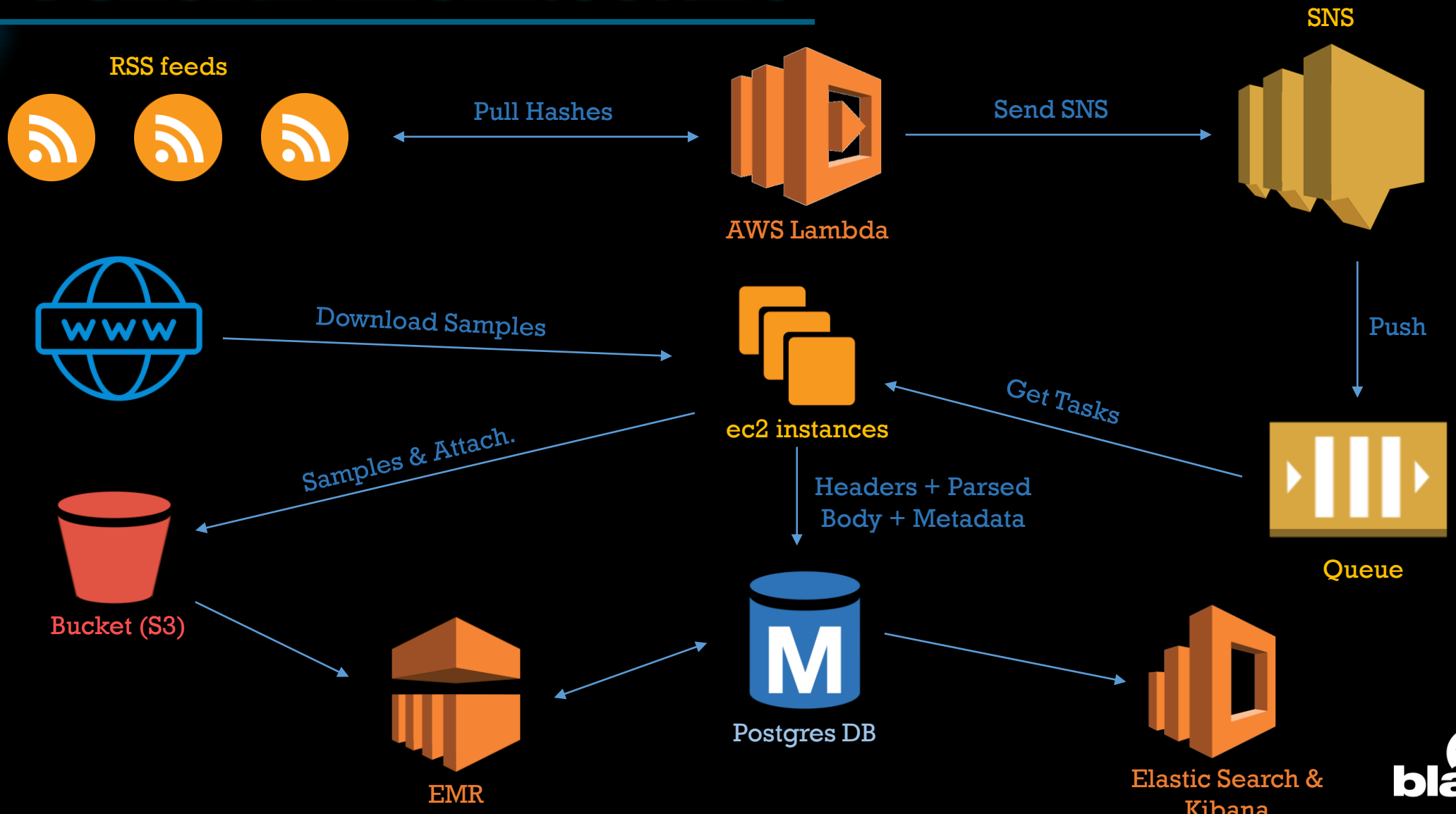
Step 3: Espionage Tool

- Reads feed from data source
- Takes SHA1 from each sample
- Downloads samples
- Analyzing samples to find the sensitive information
- Categorize the information

```
var https = require('https');
var AWS = require('aws-sdk');
AWS.config.region = 'us-west-2';
var s3 = new AWS.S3();
let url =
"https://<datasource>/.../notifications-
feed...";
exports.handler = function (event,
context,callback) {
  https.get(url, res => {
  res.setEncoding("utf8");
  let body = "";
  res.on("data", data => {
    body += data;
  });
  let ids = []
  res.on("end", () => {
    body = JSON.parse(body);
    let notifs = body['notifications'];
    let filtered = {alerts:[]};
    for(let i = 0; i<notifs.length;i++)
    {
        …
        }
    }
```

Scheduled AWS Lambda Function

General Architecture



Analysis Pipeline

#1 Feature

Spam Or Ham
- TF-IDF → Naïve Bayes
- Open Source Dataset

Email Queue

Custom Parser
- Subject & Body
- Language Detection
- Domain Names
- Geo lookup on domains
- Extract Attachments

Attachment Queue

Apache Tika

Content & Metadata Extraction

Postgres

NLP Pipeline

Preprocessing
- Stemming
- Stop word Removal
- Tokenization
- POS Tagging

scikit learn
spaCy
gensim

Named Entity Recognition
- Names Of Persons
- Organizations
- Roles(CXO,VP)
- Locations
- Monetary Values

Topic Modeling
- 10 Topics
- Top 10 words/topic

*NMF

*LDA

Scoring

1(Casual Email) - 4(Critical/Highly Confidential)

elasticsearch

kibana

* NMF = Non-Negative Matrix Factorization
* LDA = Latent Dirichlet Allocation

black hat USA 2018

DEMO TIME

QUESTIONS ?

(1) Please 

(2) Grab the 

(3) Stand up & speak English


Go Feedback!
Ido Naor (@IdoNaor1) & Dani Goland (@danigoland)
black hat
USA 2018