the human Chevron energy company™

(Attempting to) Automate the Diamond Model

FloCon 2023

Teresa Chila

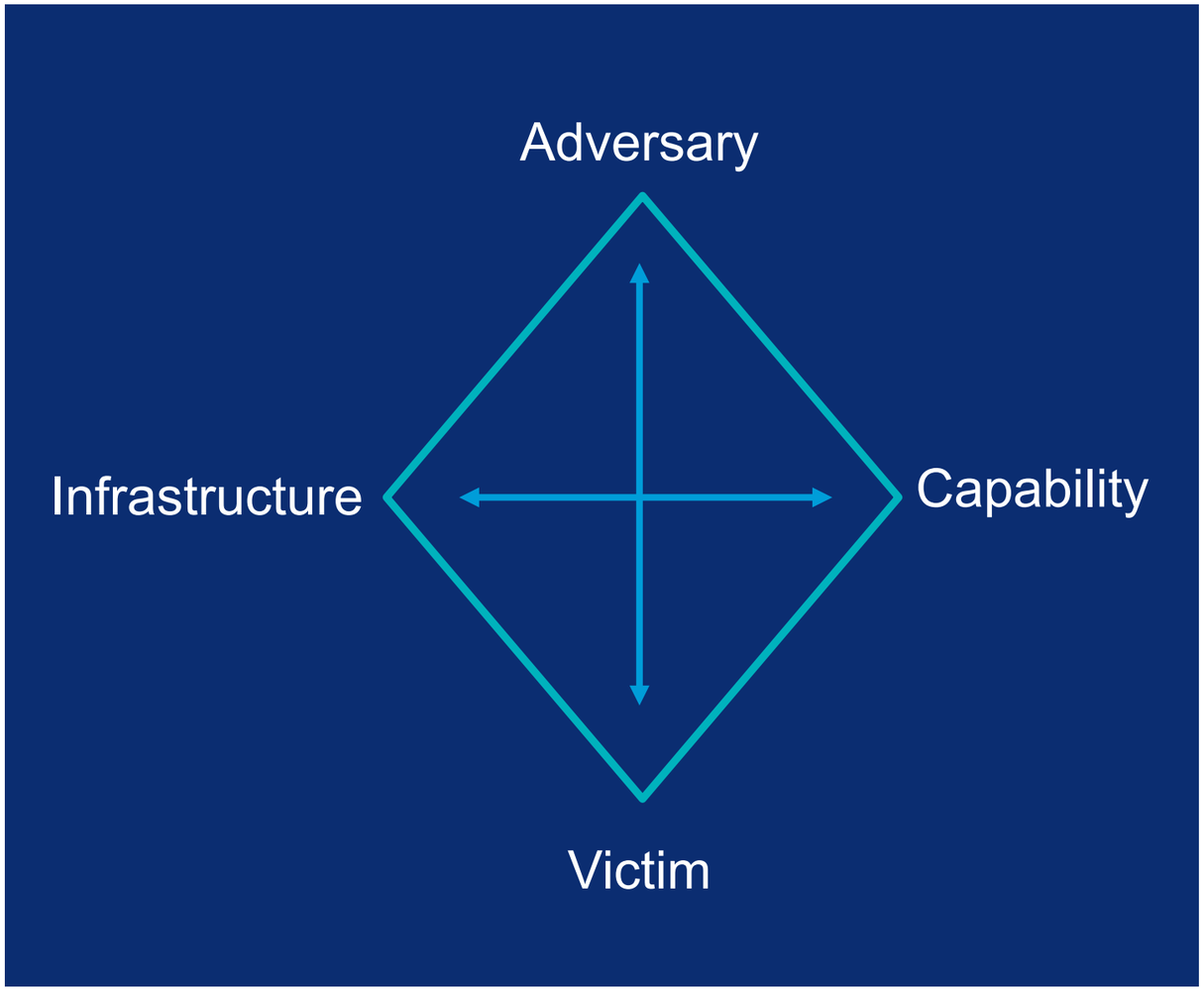
Cyber Data Scientist

Teresa Chila, Cyber Data Scientist

Biography

- Chevron Cyber Intelligence Center Data Scientist
- Use advanced analytics to improve SOC operations
- MS Electrical & Computer Engineering - Duke University
- Engineer Diploma - Télécom Paris
- Enjoys travelling around the world

Diamond Model



Diamond Model is a ***framework*** and ***methodology*** for ***analyzing cyber intrusions***

- Allow analysts to systematically find out more about adversaries by pivoting and enriching data
- Developed by cyber analysts within the DoD, which has been integrated into cyber training programs like SANS
- Paper: [diamond.pdf (activeresponse.org)]

Goal of this project

Understand relevant cyber threat actors

- Who are they?
- How do they operate?

Why is it helpful to understand how threat actors operate?

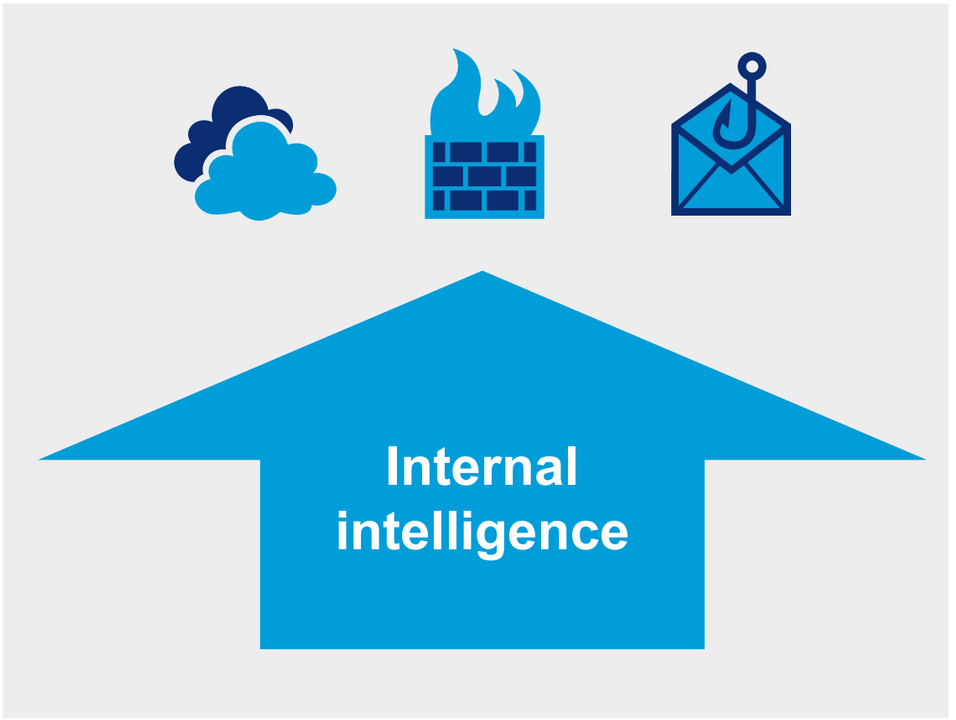
- Identify fundamental ***threat actor tendencies*** – what infrastructure/malware/intrusion techniques do they use
- Anticipate new Indicator of Compromise (new domain/IP created – proactively block these entities from contacting us)
- Anticipate/predict their next move

We can better protect Chevron by understanding relevant cyber threat actors

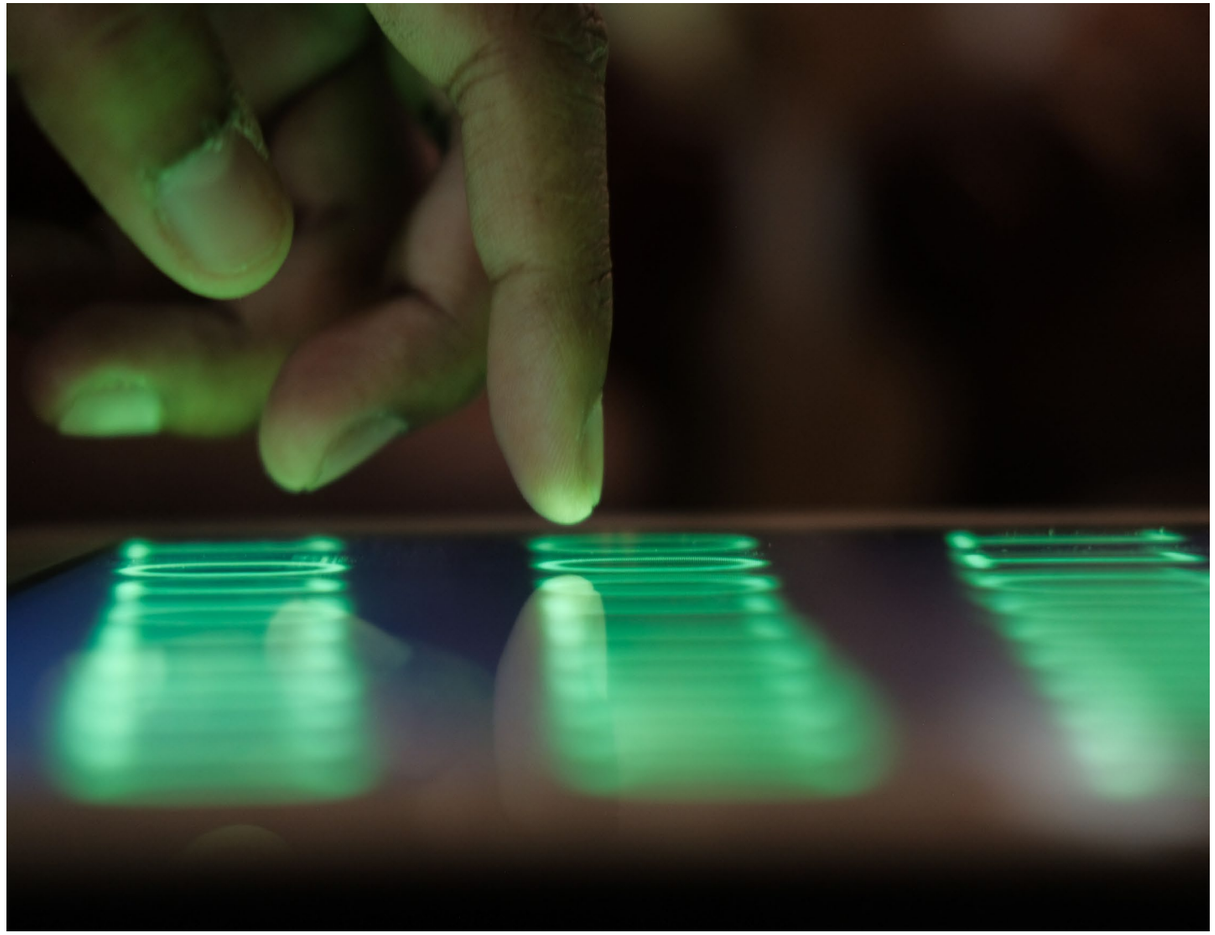
Chevron

Current state to future state

Leverage our internal data and traffic to generate intelligence
Make Chevron its own #1 intel provider



Use analytics to accelerate the Diamond Model



Historically this is all done manually - our project aims to automate and facilitate this manual analysis by:

- Automating external enrichment via external API
- Automating the data pipeline and processing workflows
- Semi-automating decision making when applicable

Run analytics and extract intelligence from untapped data:

- Blocked traffic
- Leveraging automation to analyze broader data

Challenges:

- Highly human centric, subjective
- Dependent on domain expertise and experience

Chevron

Deliverables

Business intelligence and analytics dashboards

Trend, time charts, statistics, history, etc.

Discover new Indicators of Compromise (IOCs)

Situation awareness reports and notifications

Have we seen this activity before?

New activity from an already known threat actor

New threat actor

Knowledge hub used for research

View data collected from various intel/data sources to assist with analysis

Ad-hoc analysis vs. daily feed

High fidelity leads as starting point for more in-depth investigation

How are these deliverables actionable?

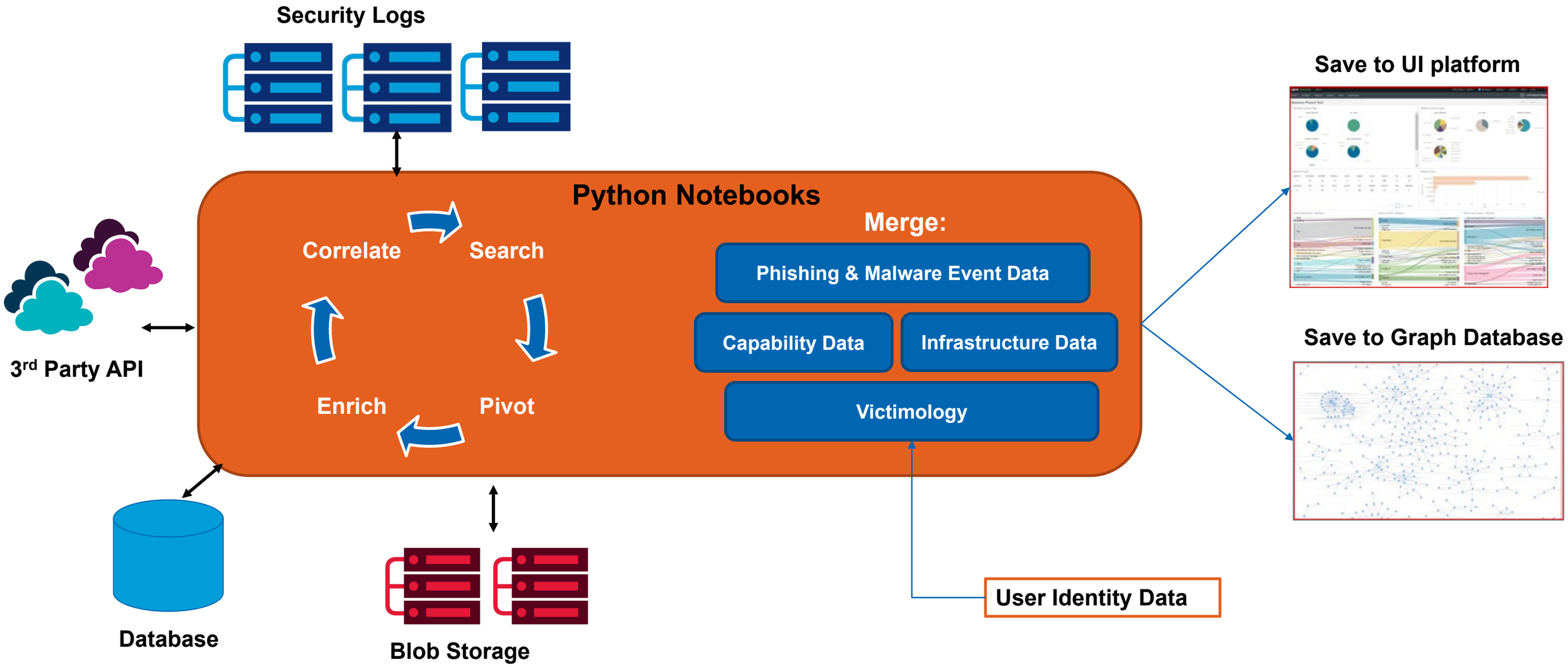
Proactively block bad traffic

Educate and warn targeted users and business units of threats

Send advisories

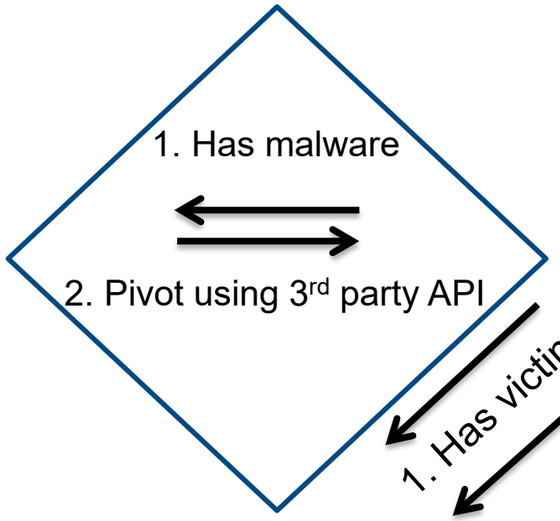
Better formulate our remediation or defensive effort based on intelligence

Architecture



Phishing email & malware use cases

Pivot → Search → Enrich → Correlate → Repeat



Emotet Malware

admin@badsender.com

2. Find an IP that the malware communicates with: 1.2.3.4

4. badsender.com is registered by badactor@gmail.com

5. badactor@gmail.com has another domain badsender.us

1. Has victim

Jane Smith

3. Search Security Logs. Found that John Doe communicated with this IP

6. Search Security Logs. Found that Joe Shmoe received an email from badsender.us

7. Found that Jane, John and Joe all work in Finance

Highlights:

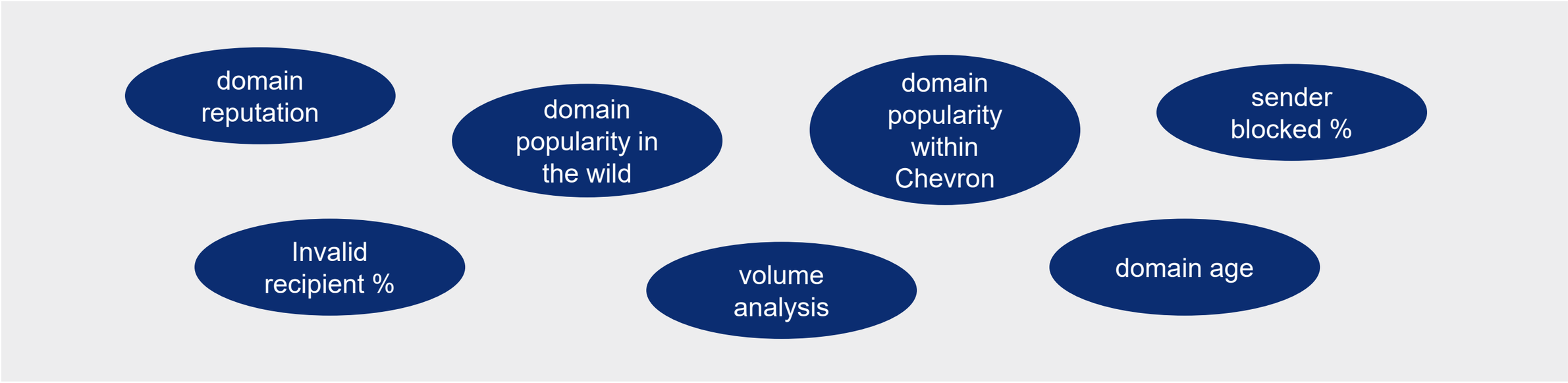
- 5 external APIs
- Recursive pivoting
- Big data searching

Chevron

Analytic pivoting

Automation Challenges

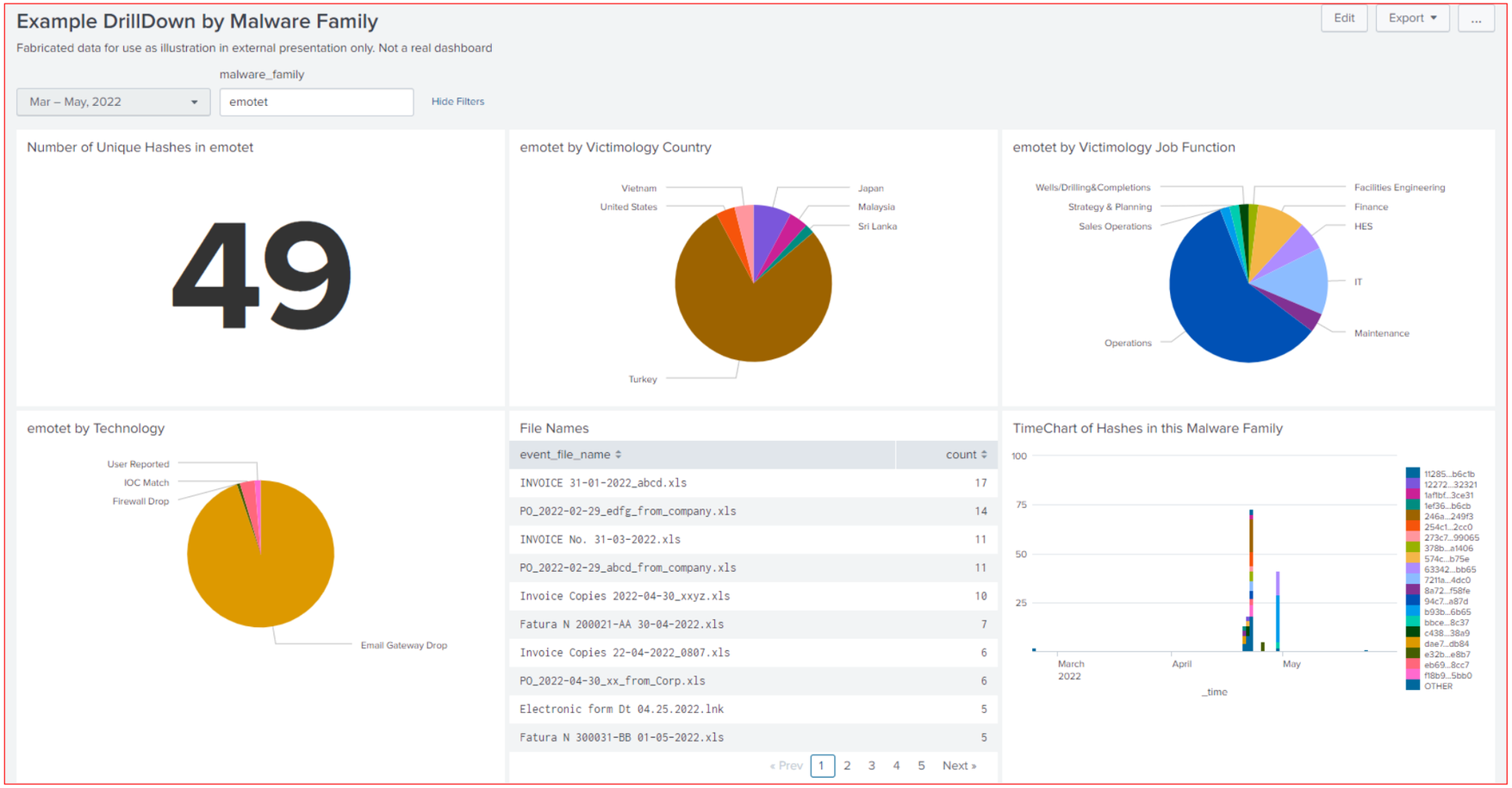
- Knowing when to stop
- Knowing whether new information you discover is related or useful to the current case or not
  - Example: sender from gmail.com
- Heuristics + ML models to help in specific steps



Business intelligence & analytic dashboards

Examples

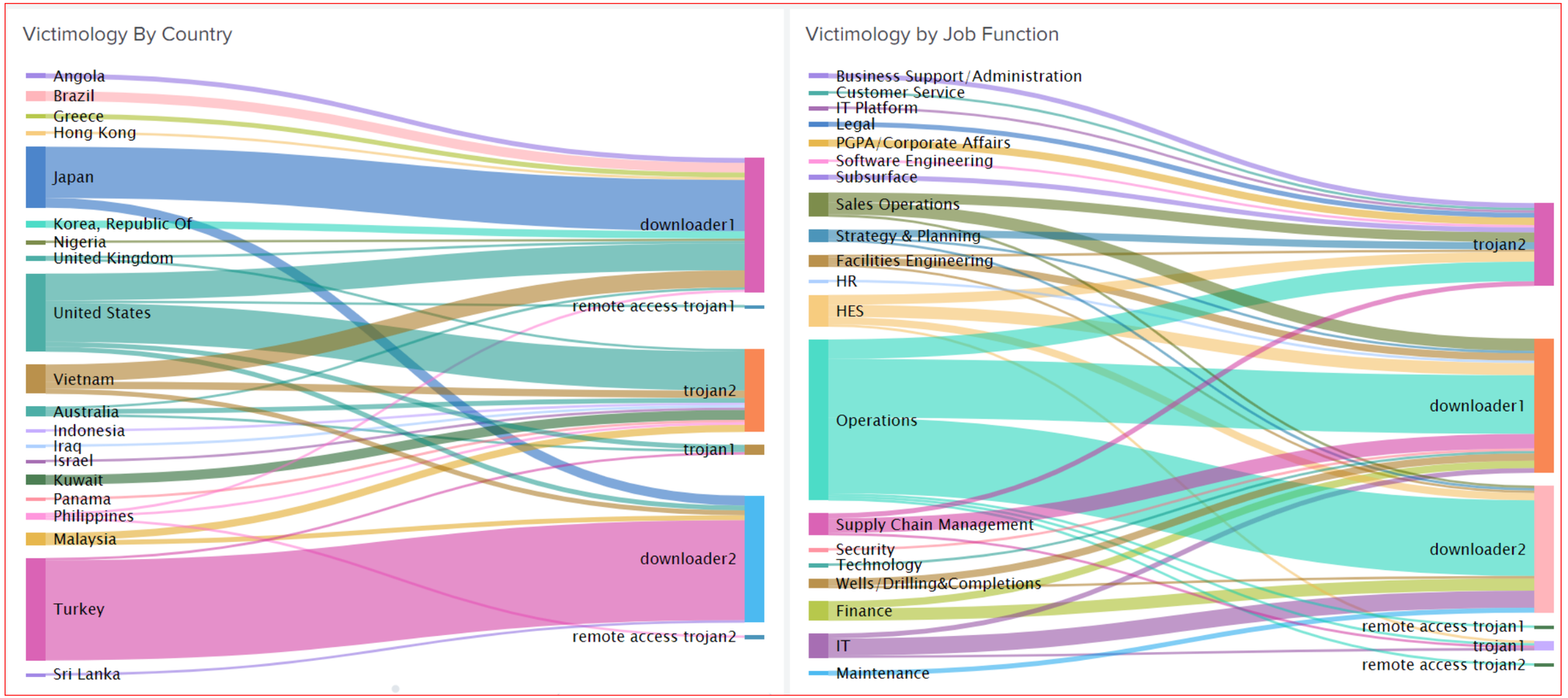
Understand a particular malware family



This is a fabricated example and does not represent actual data seen.

Examples

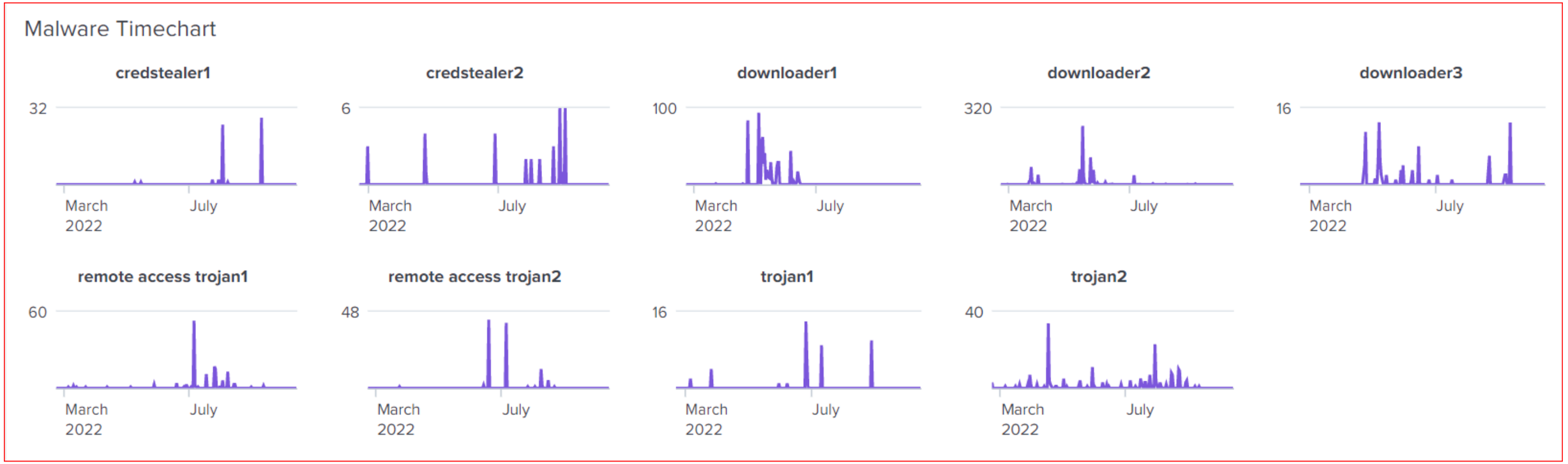
Victimology of malware by country and job function



This is a fabricated example and does not represent actual data seen.

Examples

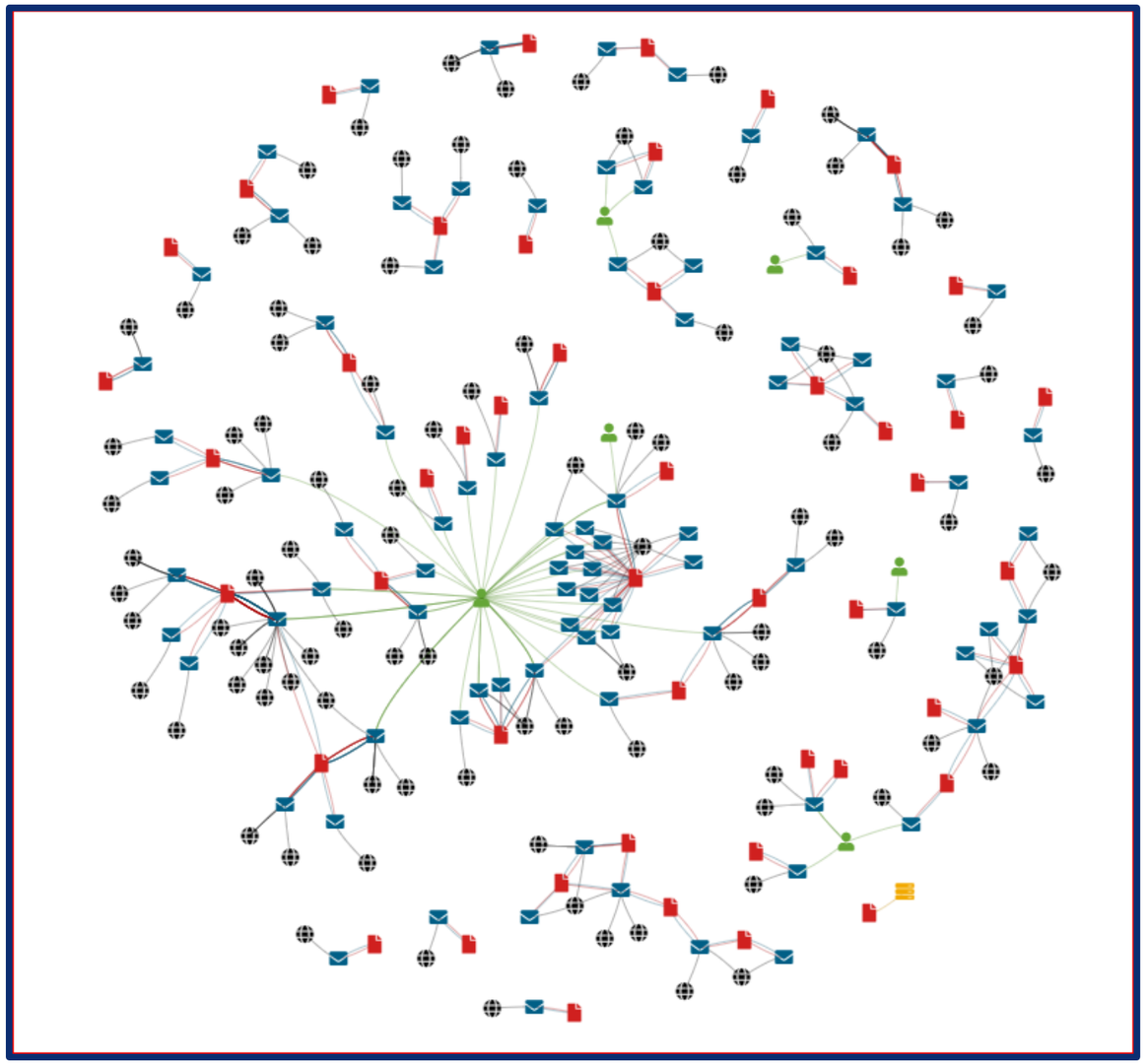
Time charts of different malware seen



This is a fabricated example and does not represent actual data seen.

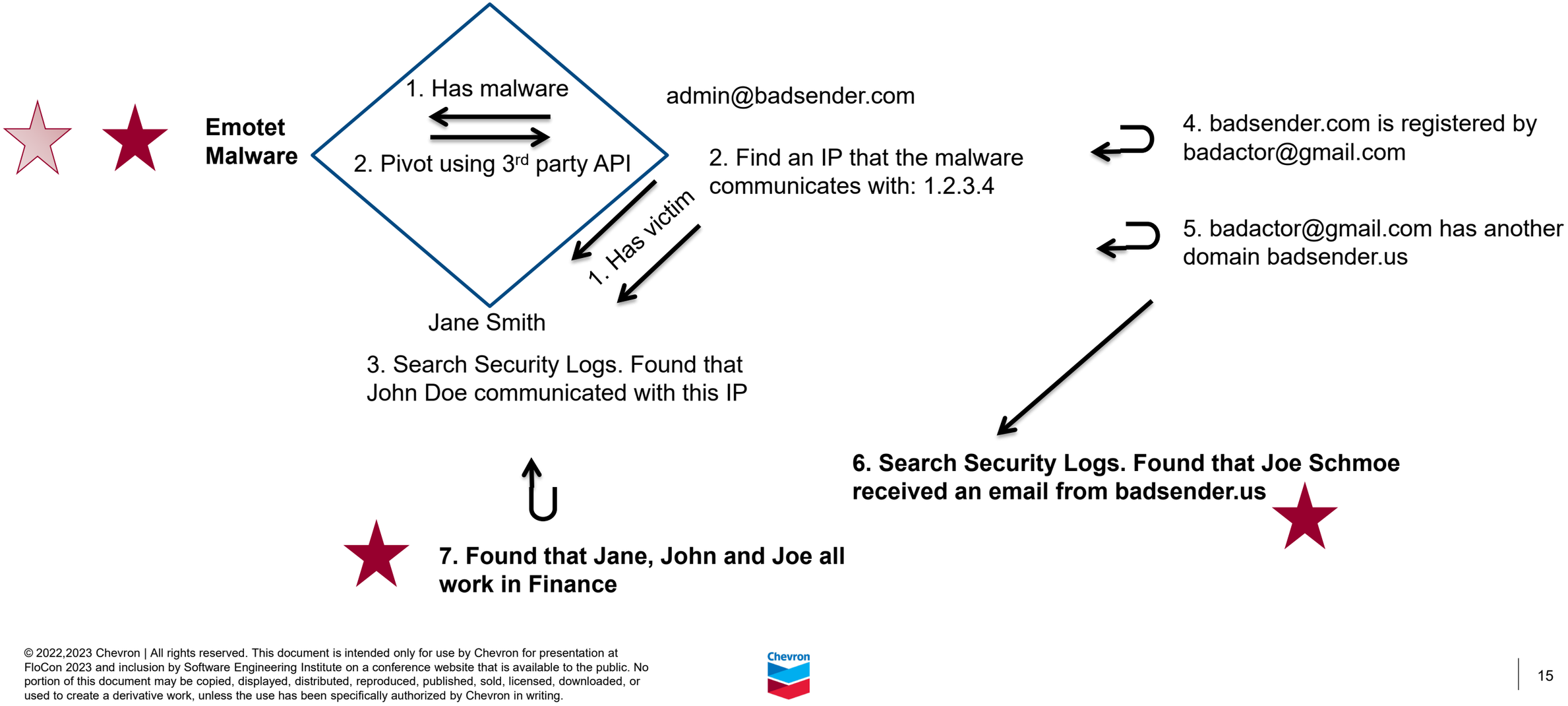
Examples

How the malware is distributed by different emails



Advanced analytics

Advanced analytics being used



Email subject model

Purpose: Help filter out false positive/benign emails

Type: Binary Classifier

Input: Email subject text

Training Data: labeled in-house (iterative curation)

Model: NLP RoBERTa Transformer model embeddings, neural network classifier

Output: Suspicious or Benign classification

score	text
0.998	Free Webinar for Retirement Savings
0.000	Urgent! Your invoice is past due

Organization and job function model

Purpose: Assign a normalized org label and job function label instead of using the raw values

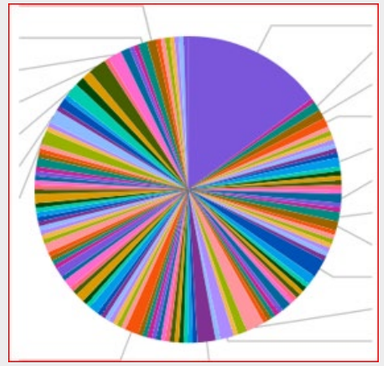
Type: Exclusive multi-class classifier

Input: User's org level hierarchy, job title from user identity data lake

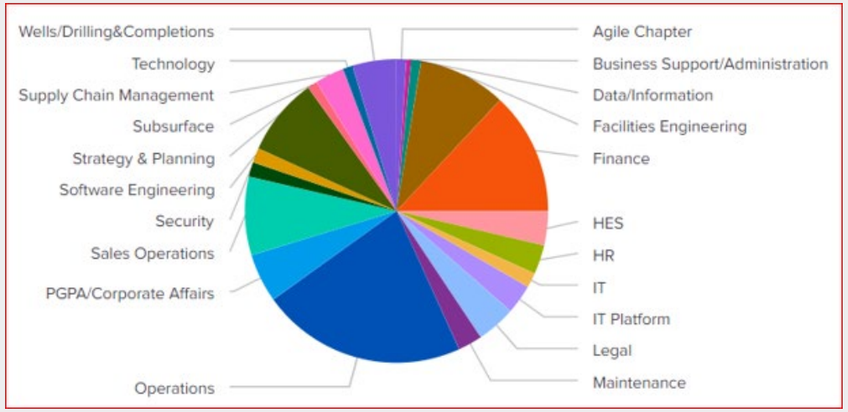
Training Data: Labeled in-house (after clustering to help with understanding of data)

Model: TF-IDF with character trigrams, Random Forest

Output: One of 30+ normalized org or job function labels



Grouped by raw job title. Not very useful



Grouped by normalized job function

Ssdeep fuzzy hash comparison

Purpose: Determine if two files are similar based on their fuzzy hash *ssdeep*, use similarity to determine malware family

Type: Distance Algorithm/Similarity Measure

Input: Two ssdeep hashes

Model: Levenshtein distance (a.k.a. edit distance) with optimization to reduce the number of comparisons

Output: Whether the two files are similar

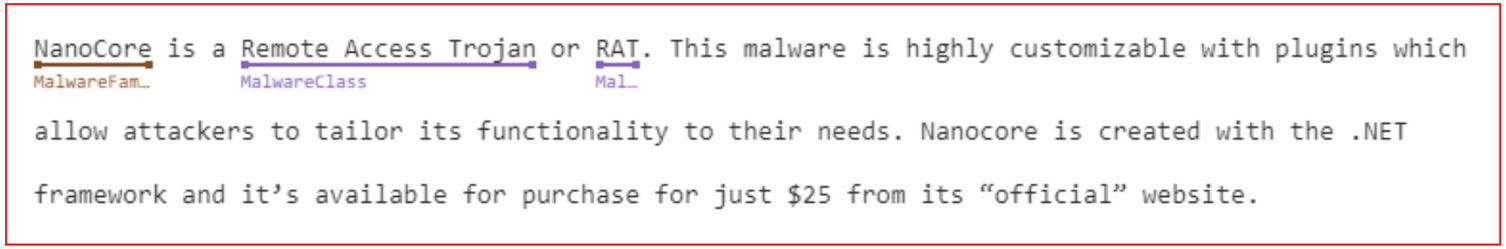
```
ssdeep1='3:AXGBicFlgVNhBGcL6wCrFQEv:AXGHsNhxLsr2C’
ssdeep2='3:AXGBicFlIHBGcL6wCrFQEv:AXGH6xLsr2C'
```

NER model to determine malware class did not implement

Purpose: Identify the malware class of a malware family (e.g. the malware class of malware Emotet is *downloader*)

Type: Text Analytics

Input: Malware family name



Training Data: Scraped internet malware wiki pages, then annotate them using Azure Cognitive Service for Language

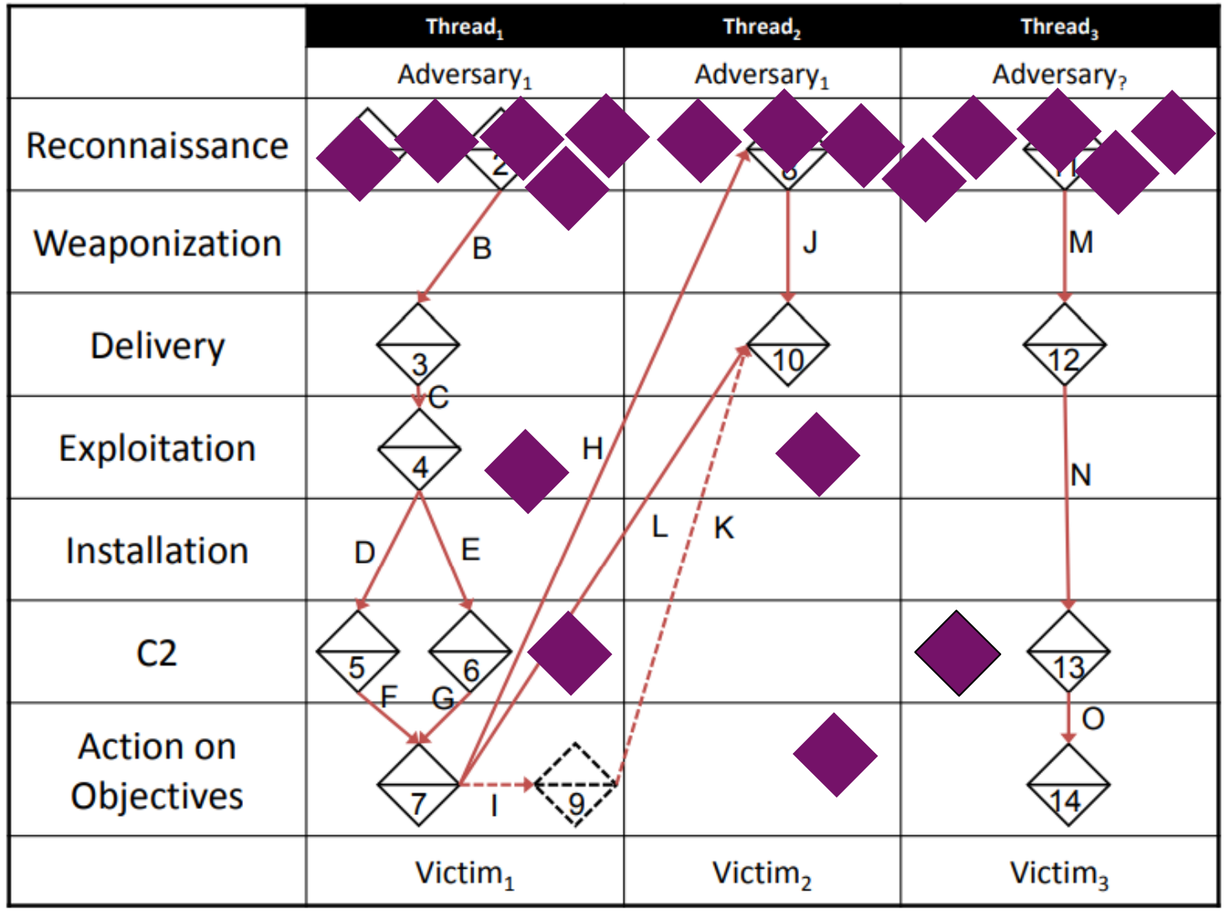
Model: Custom Named Entity Recognition model to read a Malware Family wiki page to extract the Malware Class

Output: Malware Class of the given malware family

Did not implement due to high effort and relatively low priority

Chevron

Activity groups



More data on the Recon/Initial Access phase of the Kill Chain

Focus our automation effort here

Less data

Source: The Diamond Model of Intrusion Analysis (Paper)

Email similarity use case

Used for identifying email campaigns

Purpose: Find similar emails that are likely sent by the same bad actor

Express and store as a **graph**:

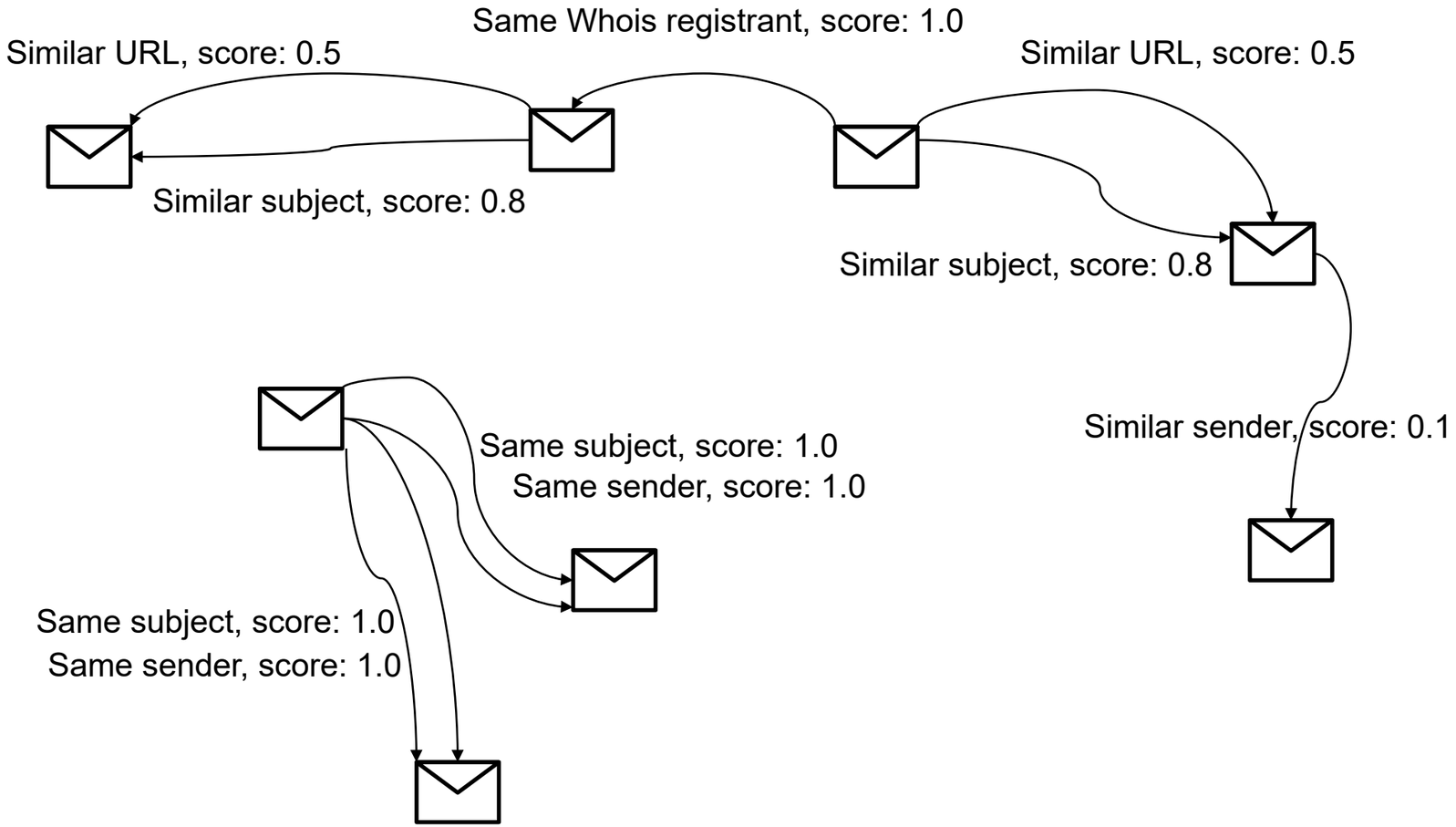
- A node: an email event
- An edge: when two emails are similar based on a criteria, an edge is formed with a score

Sender	Subject	URL
info@red-car.com	Invoice number: #123	http://1.1.1.1/aabbcc
info@green-leaf.com	Invoice number: #234	http://2.3.4.5/aabbcc
info@blue-star.com	Invoice number: #456	http://6.6.6.6/aabbcc

This is a fabricated example and does not represent actual data seen.

Email similarity use case

For identifying email campaigns



Step 1: Establish edges
(via clustering algorithms, distance algorithms with LSH, or keyword search)

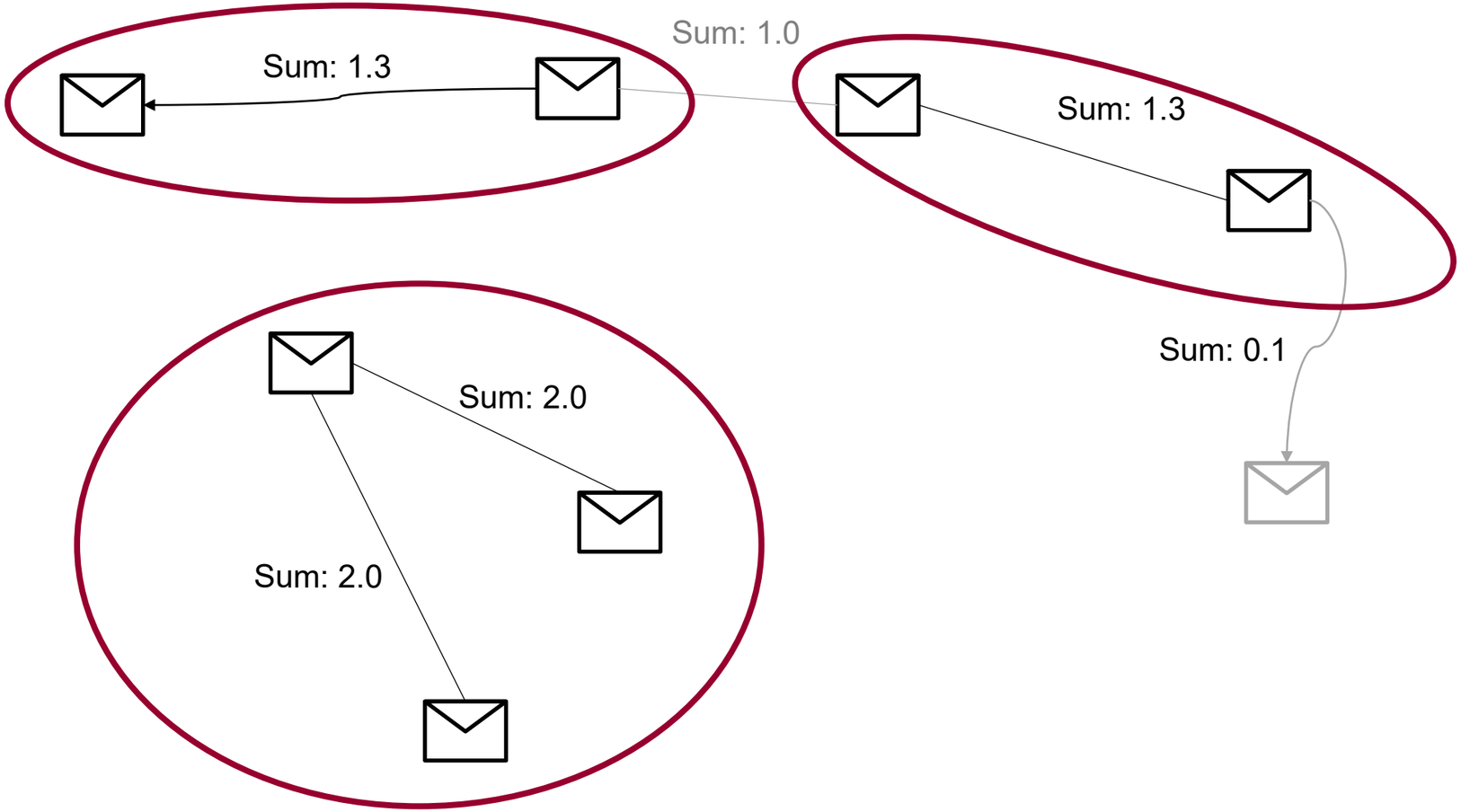
Types of similarity (edges):

- Sender address
- Email subject
- URL within the email
- Time proximity
- Malware hash/family
- Attachment name
- Domain Whois info

Chevron

Email similarity use case

For identifying email campaigns



Step 2: Dedup and Sum Edge Score

Step 3: Filter score per threshold

Step 4: Run connected component algorithm

Benefits:

- Extensible to new edge type
- Can adjust score and threshold for tuning

Human feedback loop is critical

- Work closely with Threat Intelligence Analysts (our customers)
- In-the-loop human review:
  - Is this a valid activity group?
  - Tag additional information missed by automation
- Build user interface that allows user input and feedback loop into the system
- Ask for patience and understanding

Project team recognition

Special thanks and recognition to the project team

<u>Data Scientists:</u>
Teresa Chila
Jorge Crisostomo
Sasha Opela

<u>Data Engineers:</u>
Stephen Ogletree
Ngan Trinh

<u>Team Lead:</u>
Mark Wade

Questions