CISCO SECURE

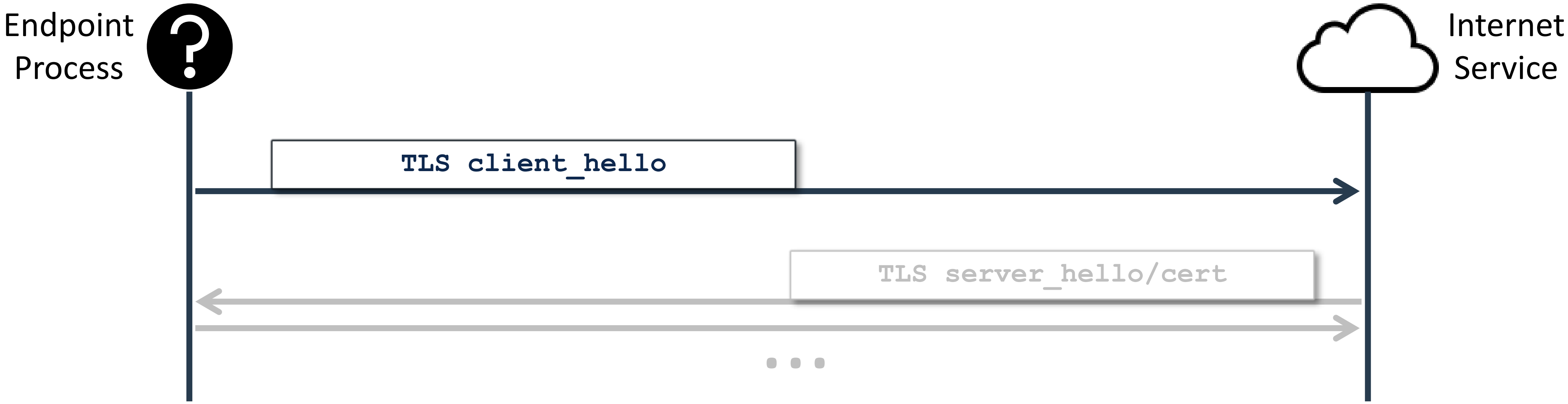
A Behind the Scenes View of Supporting ML-based Security Solutions

Blake Anderson, PhD
blake.anderson@cisco.com
FloCon - January 2024

ACKs: Durairaj Murugasamy, Prateek Somani, David McGrew

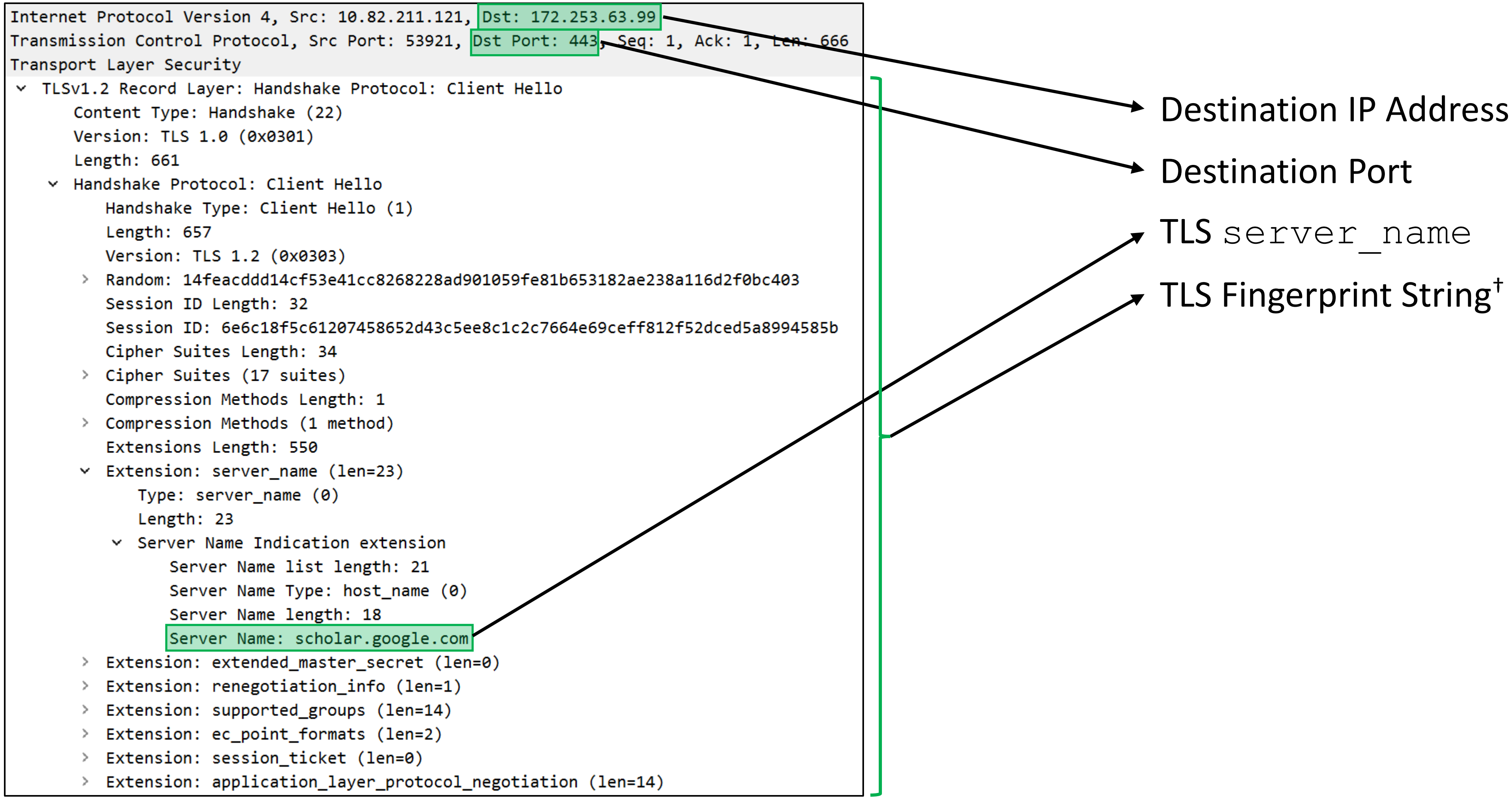
CISCO The bridge to possible

Use Case: Process Identification



- Goal: Identify the endpoint process given the TLS `client_hello`
  - Uses only the initial data packet

Available Data Features



```
Internet Protocol Version 4, Src: 10.82.211.121, Dst: 172.253.63.99
Transmission Control Protocol, Src Port: 53921, Dst Port: 443, Seq: 1, Ack: 1, Len: 666
Transport Layer Security
⌄ TLSv1.2 Record Layer: Handshake Protocol: Client Hello
    Content Type: Handshake (22)
    Version: TLS 1.0 (0x0301)
    Length: 661
  ⌄ Handshake Protocol: Client Hello
      Handshake Type: Client Hello (1)
      Length: 657
      Version: TLS 1.2 (0x0303)
    > Random: 14feacddd14cf53e41cc8268228ad901059fe81b653182ae238a116d2f0bc403
      Session ID Length: 32
      Session ID: 6e6c18f5c61207458652d43c5ee8c1c2c7664e69ceff812f52dced5a8994585b
      Cipher Suites Length: 34
    > Cipher Suites (17 suites)
      Compression Methods Length: 1
    > Compression Methods (1 method)
      Extensions Length: 550
    ⌄ Extension: server_name (len=23)
        Type: server_name (0)
        Length: 23
      ⌄ Server Name Indication extension
          Server Name list length: 21
          Server Name Type: host_name (0)
          Server Name length: 18
          Server Name: scholar.google.com
    > Extension: extended_master_secret (len=0)
    > Extension: renegotiation_info (len=1)
    > Extension: supported_groups (len=14)
    > Extension: ec_point_formats (len=2)
    > Extension: session_ticket (len=0)
    > Extension: application_layer_protocol_negotiation (len=14)
```

Destination IP Address

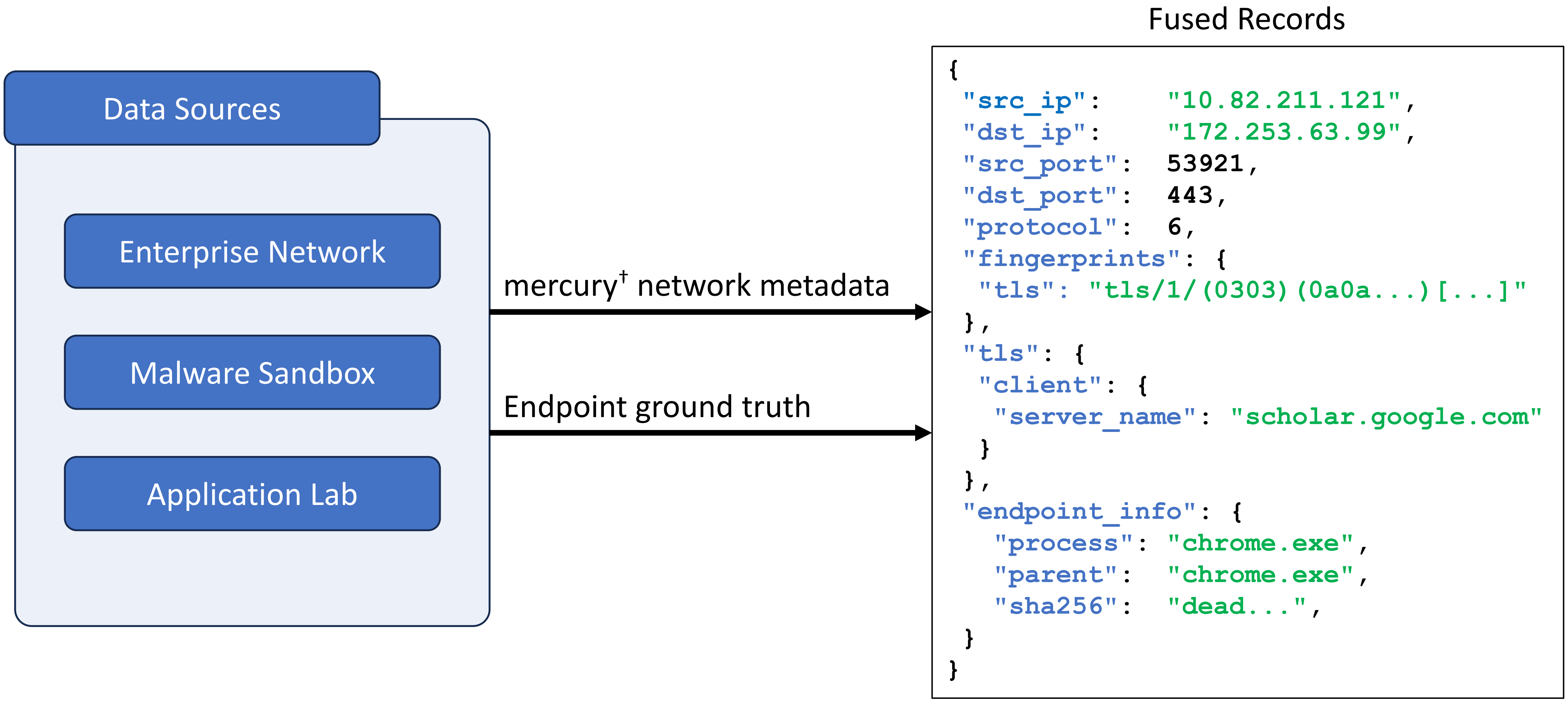
Destination Port

TLS `server_name`

TLS Fingerprint String[†]

[†] https://github.com/cisco/mercury/blob/main/doc/npf.md

Collecting Ground Truth



Data Sources

Enterprise Network

Malware Sandbox

Application Lab

mercury[†] network metadata

Endpoint ground truth

Fused Records

```
{
 "src_ip":     "10.82.211.121",
 "dst_ip":     "172.253.63.99",
 "src_port":   53921,
 "dst_port":   443,
 "protocol":   6,
 "fingerprints": {
  "tls": "tls/1/(0303)(0a0a...)[...]"
 },
 "tls": {
  "client": {
   "server_name": "scholar.google.com"
  }
 },
 "endpoint_info": {
   "process": "chrome.exe",
   "parent":  "chrome.exe",
   "sha256":  "dead...",
 }
}
```

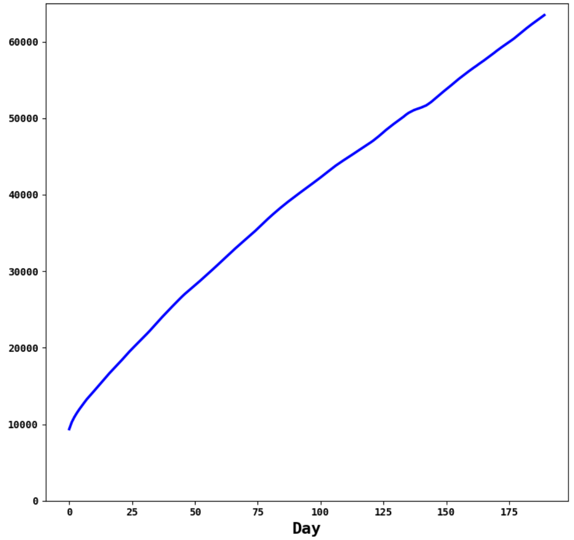
[†] https://github.com/cisco/mercury/

Ground Truth Limitations

- Why aren't we close to a solution?

The current label set is:

Unbounded



Imprecise

```
chrome.exe
google chrome
google chrome helper
chrome - copy.exe
chrome (1).exe
```

Uninformative

Labeling Goals

- Group endpoint process descriptors into a label set

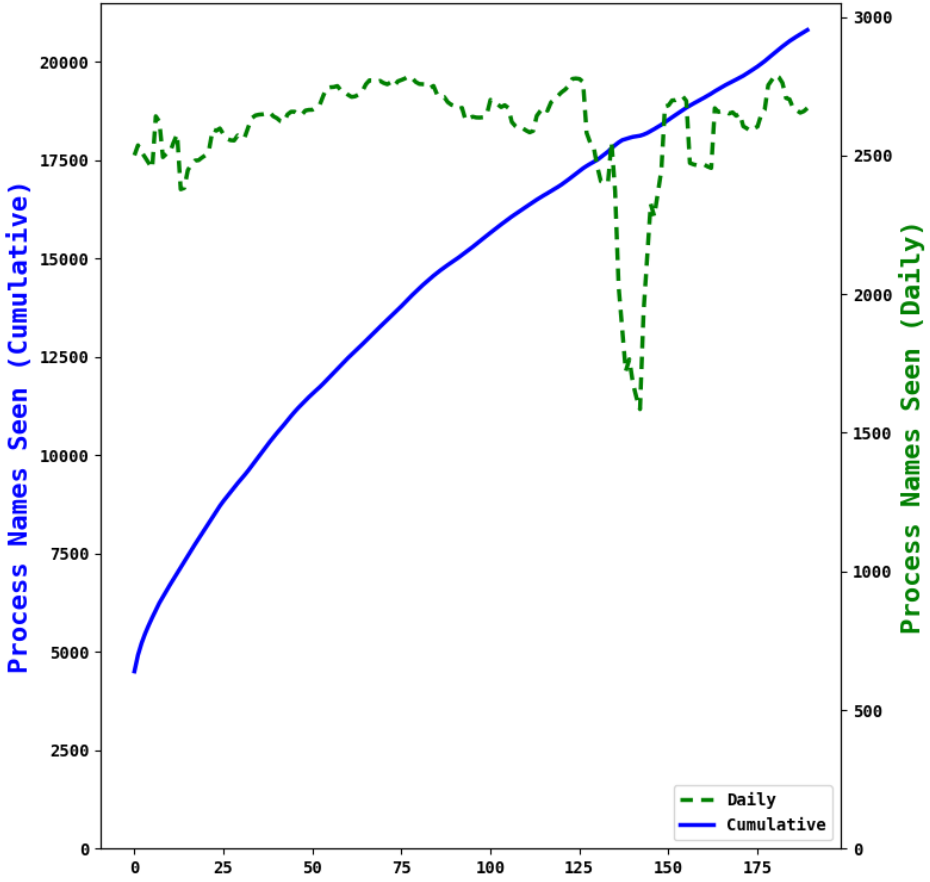
```
"endpoint_info": {
  "process": "chrome.exe",
  "parent":  "chrome.exe",
  "sha256":  "dead...",
}
```

→ Chromium Web Browser

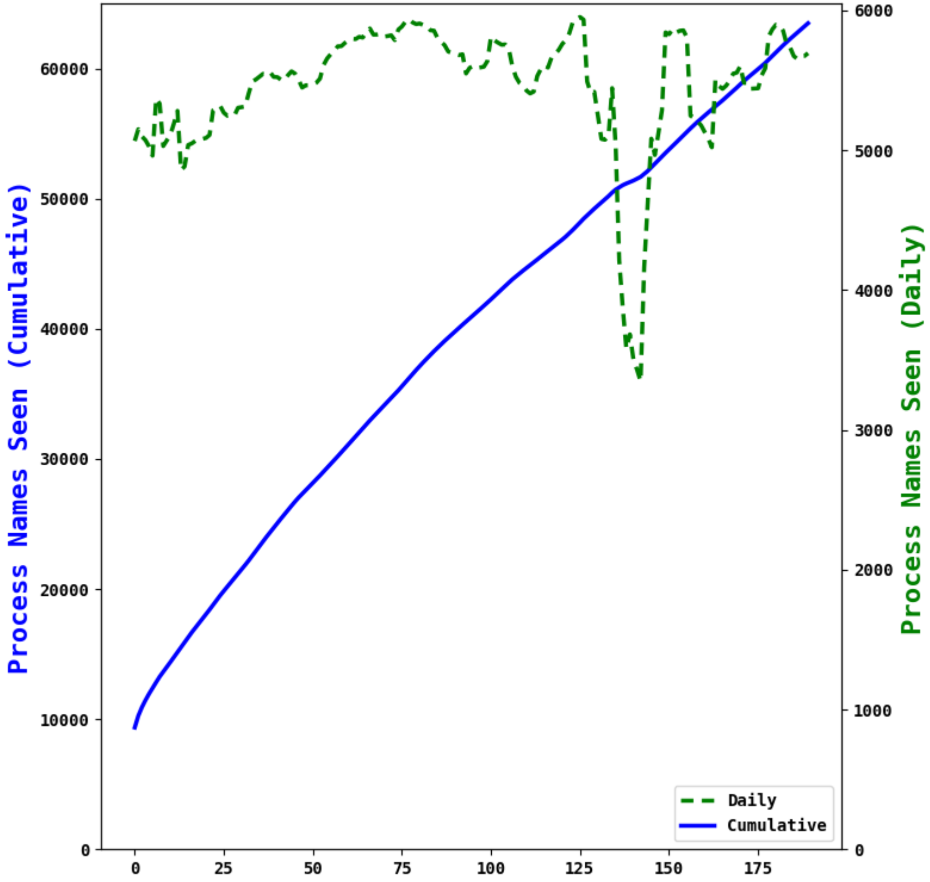
- Criteria:
  - Shares code and performs similar functions, **and**
  - Are indistinguishable given the available data features

Unbounded Label Set

Newly Introduced Process Names



Newly Introduced Process Hashes



Imprecise/Uninformative Labels

- OS-dependent naming
  - `chrome.exe` (WinNT) vs. `google chrome helper` (MacOS)
- System/User renaming
  - `chrome.exe` / `chrome (1).exe` / `chrome – copy.exe`
- Virtual machines / browser plugins
  - Virtual Box, Vmware Workstation, Parallels Desktop, ...
  - Process labels are obfuscated from endpoint data collections tools

Automation Corner Cases

- Parent processes matter *sometimes*
  - `(splunkd.exe)(python.exe) -> beam.scs.splunk.com`
  - `(python.exe)(python.exe)  -> pypi.org`

- Missing information about process hashes
  - Only ~55% of the 60k process hashes could be associated with product information

- Uninformative domains
  - ocsp.digicert.com: 209 unique process families
  - login.microsoftonline.com: 75 unique process families

Interactive Labeler

```
current record:
          process name: updater
          parent name:  wdavdaemon
          sha256:       87A6C247F852E79AF448EC546C488E9F57012EBEA2F902A7658344A63FB9867F
          count:        4624
          dst_ips:      ['40.70.161.7', '52.177.138.113', '40.70.161.102']
          domains:      ['in.appcenter.ms']
```

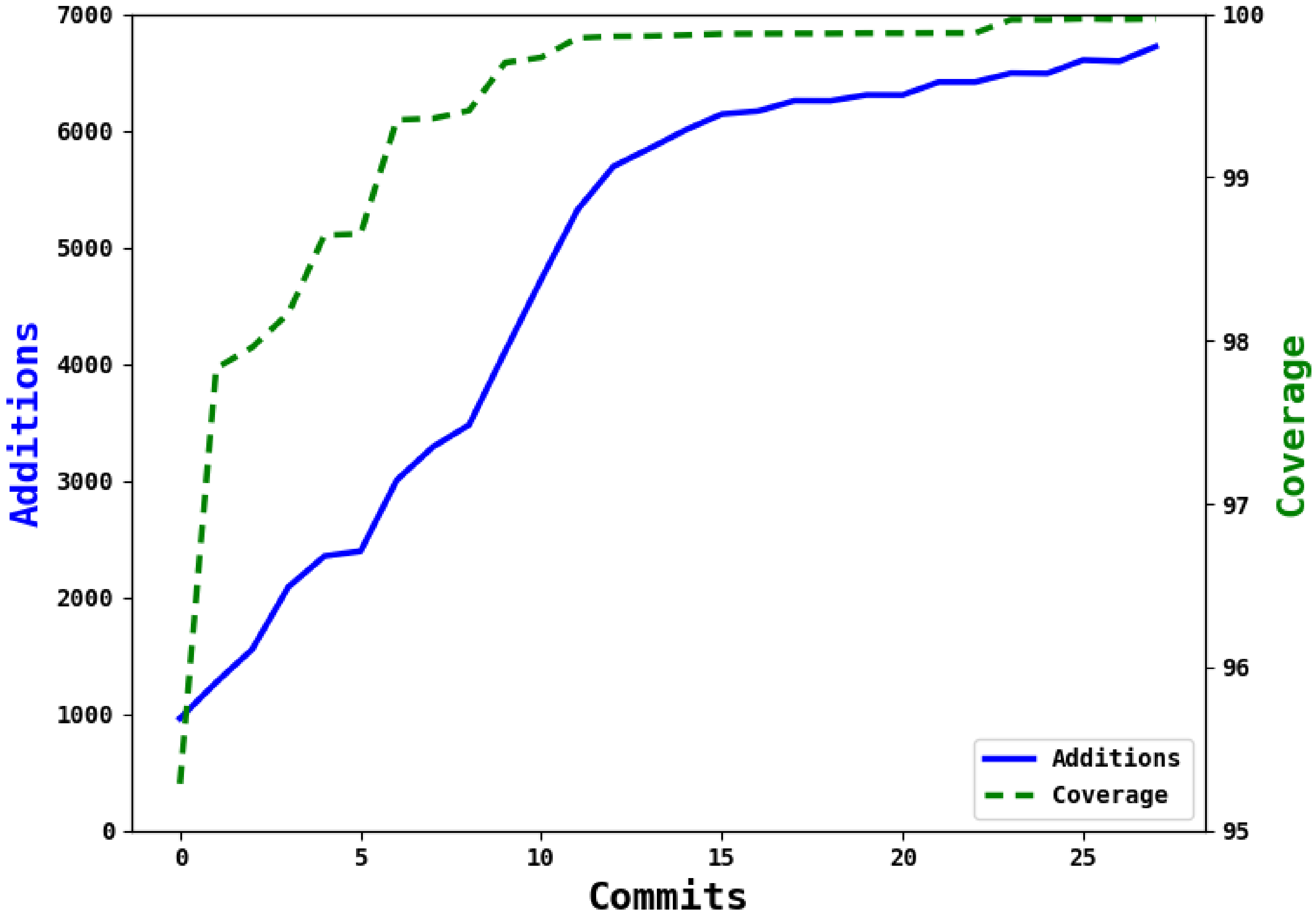
```
top process name matches:
          score:        100
          family name:  keybase file sharing
          process name: updater
          dst_ips:      ['52.72.221.214', ...]
          domains:      ['api-0.core.keybaseapi.com', ...]
          ...

top parent process name matches:
          score:        100
          family name:  microsoft windows defender
          process name: wdavdaemon
          dst_ips:      ['2600:1406:3c:48a::2c1a', ...]
          domains:      ['www.microsoft.com', ...]
          ...
```

```
top domain matches:
          score:        73.66666666666667
          family name:  vox music player
          process name: vox
          dst_ips:      ['52.232.209.85', ...]
          domains:      ['in.appcenter.ms', ...]
          ...

top ip addr matches:
          score:        100.0
          family name:  microsoft remote desktop
          process name: microsoft remote desktop
          dst_ips:      ['40.70.161.102', ...]
          domains:      ['in.appcenter.ms', ...]
          ...
```

Label Coverage



Conclusions

- Small labeling mistakes can have substantial consequences
  - automated: 89.98%
  - manual: 97.87%

- Streamlining the manual effort resulted in significant improvements

- Investments in human labeling has had a great ROI w.r.t. the ML model’s performance and general measurement/understanding cisco SECURE