Using AI Computer Vision in Your OSINT Data Analysis

Team 11

110590453 黃台茗

111599004 陳勝舢

112C53010 林幸慧

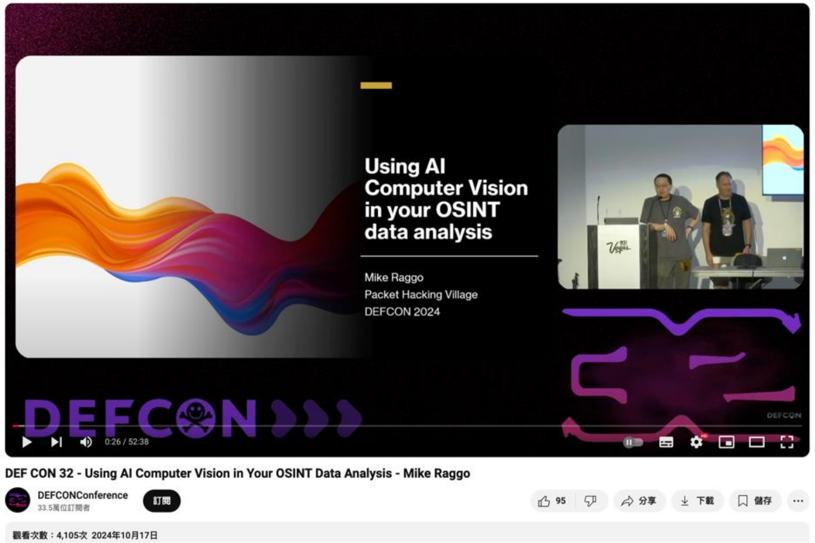
112c72011 許貽昇

113598023 廖哲霈

113c53019 李宇揚

Before the beginning

- Our presentation builds on *Def Con 32's Using AI Computer Vision in Your OSINT Data Analysis*.
- We applied the mentioned techniques to real-world cybersecurity incidents, including the I-Soon data leak incidents.



Using AI Computer Vision in your OSINT data analysis

Mike Raggo
Packet Hacking Village
DEFCON 2024

DEF CON 32 - Using AI Computer Vision in Your OSINT Data Analysis - Mike Raggo

DEFCONConference
33.5萬位訂閱者

觀看次數：4,105次 2024年10月17日

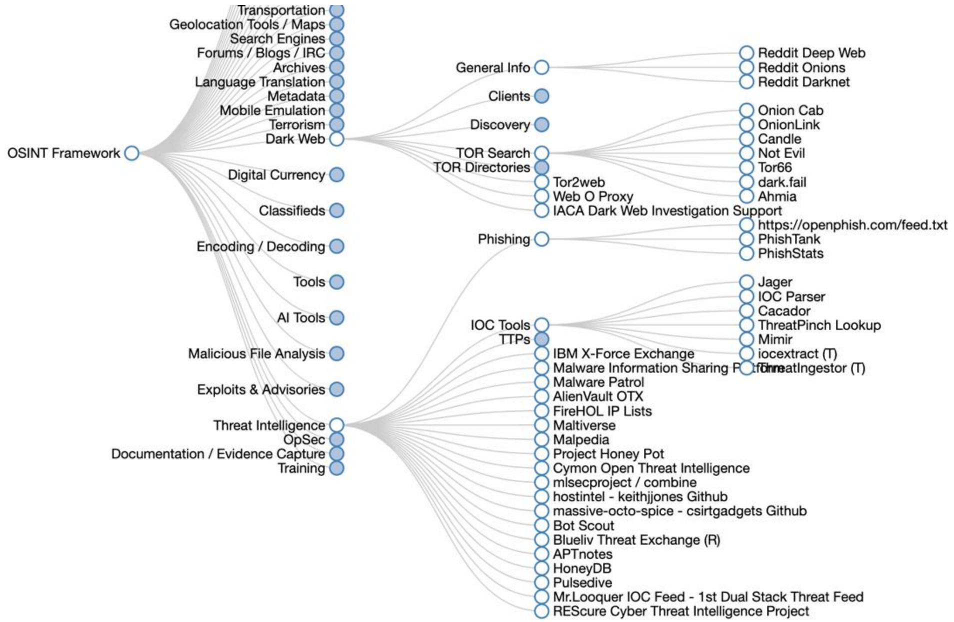
Outlines

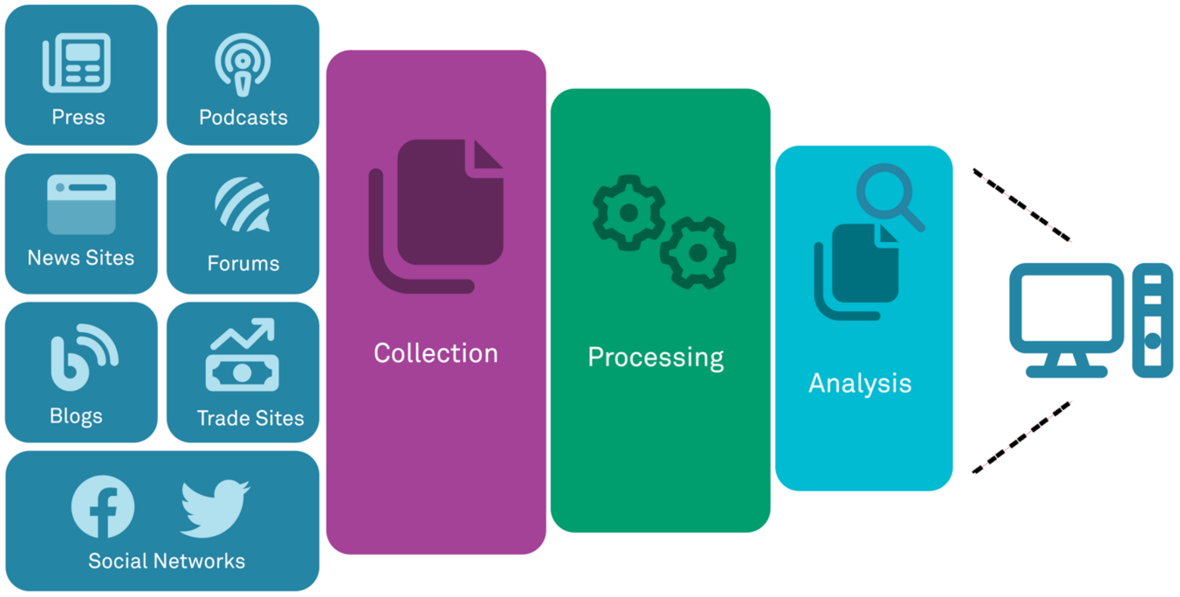
- Introduction
  - OSINT and its challenge
  - AI Computer Vision for OSINT Analysis
- Methodology
  - Optical Character Recognition (OCR)
  - Object Detection
  - Speech Recognition
  - Dashboard Summary
- Limitation and Vulnerability
- Conclusion
- Team Contribution

Introduction

What is OSINT?

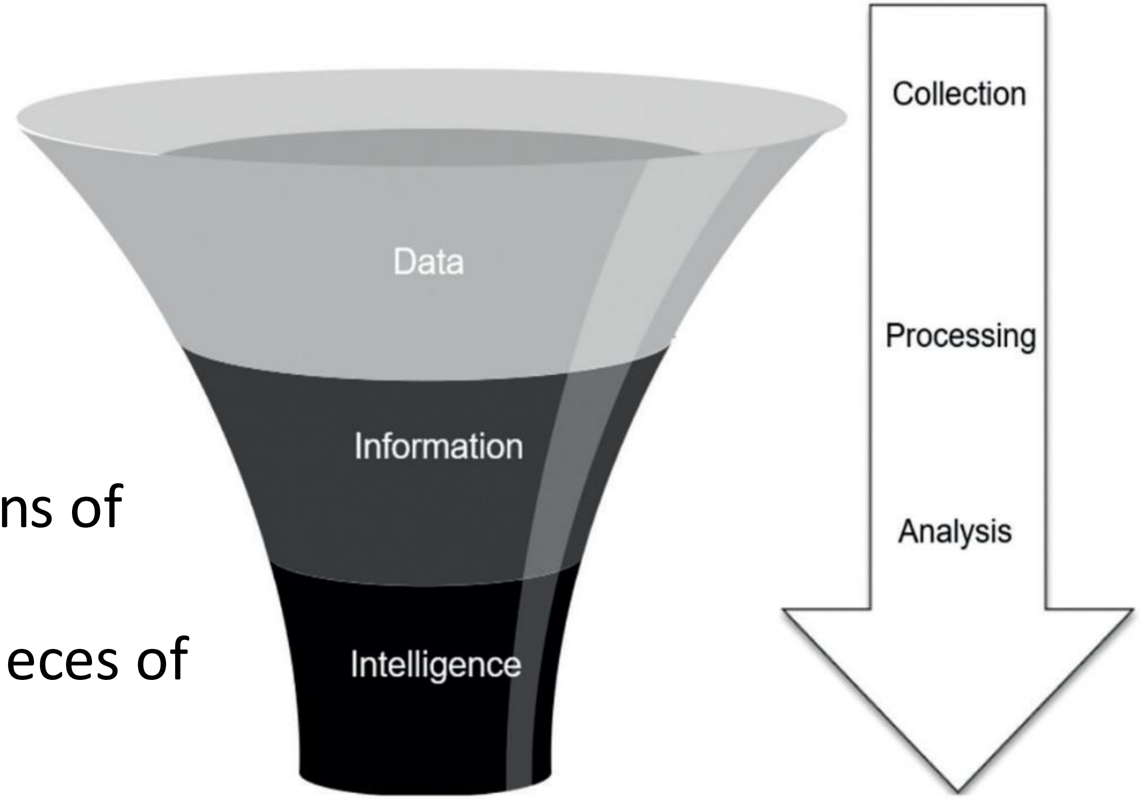
Open-source intelligence is generated from publicly available information, regularly collected, extracted, and supplied to meet specific intelligence requirements.





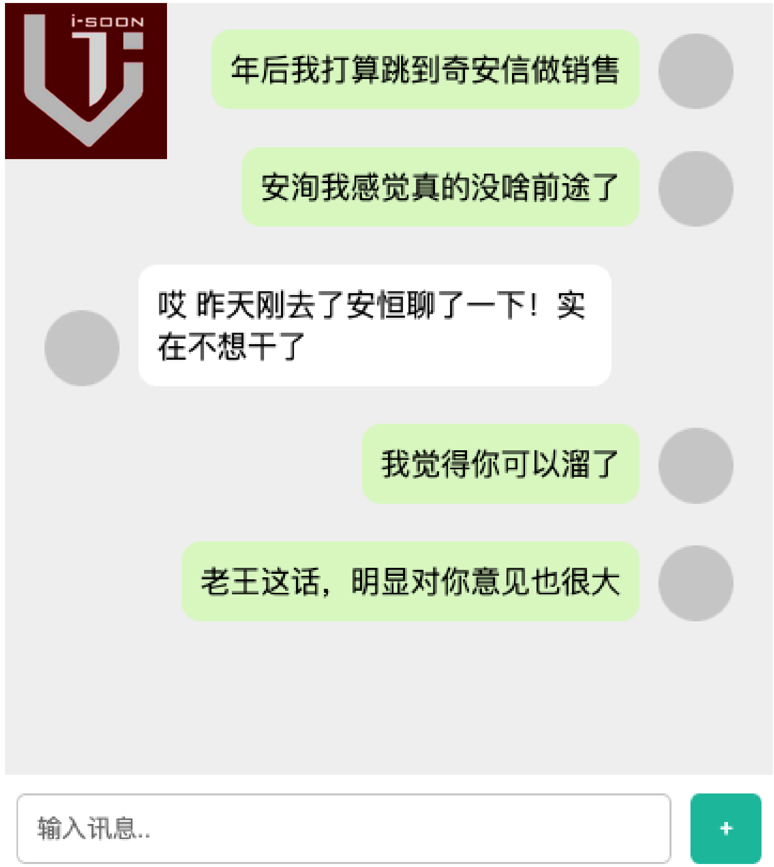
OSINT Insights

- Unprocessed information
  - e.g., IP addresses, logs, etc.
- Information
  - Structured and filtered data
- Intelligence
  - Mitigation strategies or descriptions of cybersecurity incidents
  - Events correlated from multiple pieces of information



Analysis Target - I-soon Leak

- Background
    - On February 16, 2024, internal documents from the Chinese company “I-Soon Information Technology” were leaked on GitHub.
- Cause Reason
    - Suspected employee retaliation due to dissatisfaction with compensation and the company.
- Leaked Items
    - 2020-2022 conversation records, victim information, product white papers, salary, and employee lists.



年后我打算跳到奇安信做销售

安洵我感觉真的没啥前途了

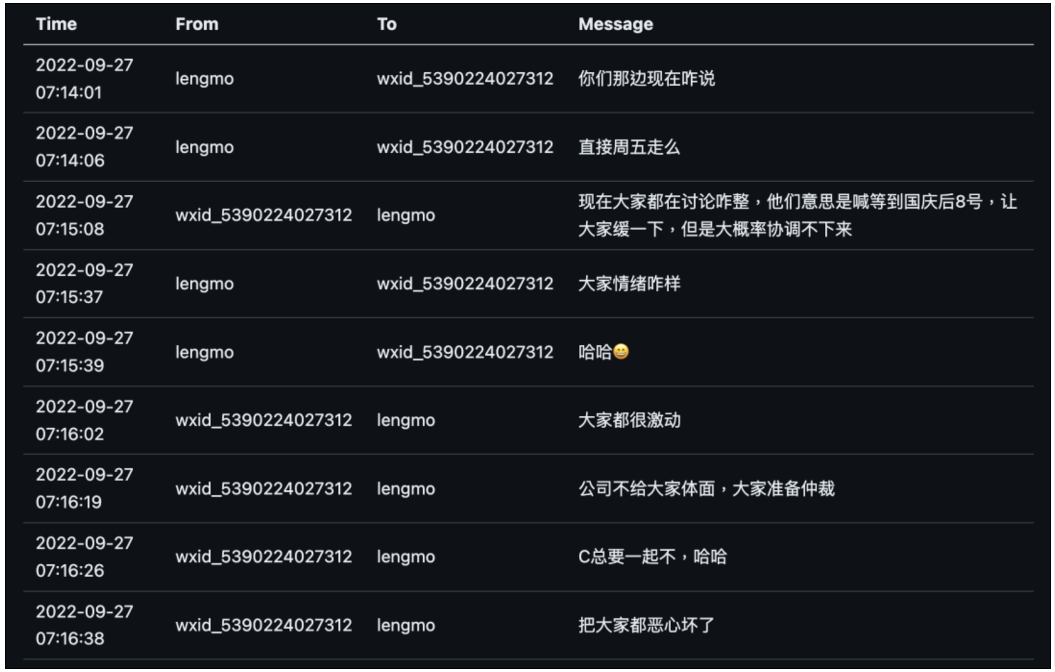
哎 昨天刚去了安恒聊了一下！实在不想干了

我觉得你可以溜了

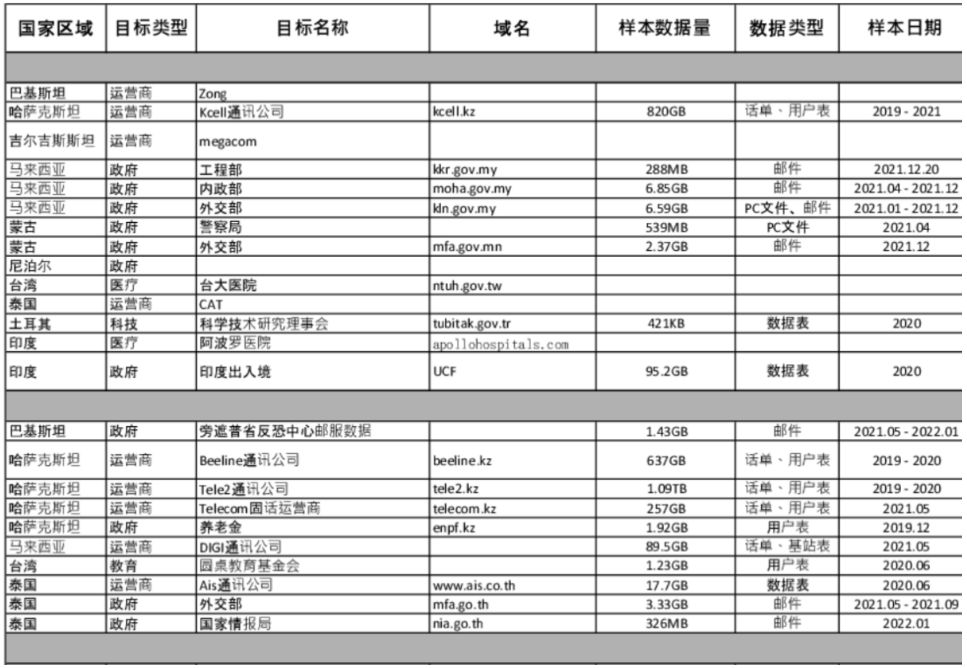
老王这话，明显对你意见也很大

Challenges in Analyzing Data Leak (I-Soon)

- The I-Soon Data Leak file contains over 15,743 rows, making the analysis time-consuming...



Time	From	To	Message
2022-09-27 07:14:01	lengmo	wxid_5390224027312	你们那边现在咋说
2022-09-27 07:14:06	lengmo	wxid_5390224027312	直接周五走么
2022-09-27 07:15:08	wxid_5390224027312	lengmo	现在大家都在讨论咋整，他们意思是喊等到国庆后8号，让大家缓一下，但是大概率协调不下来
2022-09-27 07:15:37	lengmo	wxid_5390224027312	大家情绪咋样
2022-09-27 07:15:39	lengmo	wxid_5390224027312	哈哈😄
2022-09-27 07:16:02	wxid_5390224027312	lengmo	大家都很激动
2022-09-27 07:16:19	wxid_5390224027312	lengmo	公司不给大家体面，大家准备仲裁
2022-09-27 07:16:26	wxid_5390224027312	lengmo	C总要一起不，哈哈
2022-09-27 07:16:38	wxid_5390224027312	lengmo	把大家都恶心坏了



国家区域	目标类型	目标名称	域名	样本数据量	数据类型	样本日期
巴基斯坦	运营商	Zong				
哈萨克斯坦	运营商	Kcell通讯公司	kcell.kz	820GB	话单、用户表	2019 - 2021
吉尔吉斯斯坦	运营商	megacom				
马来西亚	政府	工程部	kkr.gov.my	288MB	邮件	2021.12.20
马来西亚	政府	内政部	moha.gov.my	6.85GB	邮件	2021.04 - 2021.12
马来西亚	政府	外交部	kln.gov.my	6.59GB	PC文件、邮件	2021.01 - 2021.12
蒙古	政府	警察局		539MB	PC文件	2021.04
蒙古	政府	外交部	mfa.gov.mn	2.37GB	邮件	2021.12
尼泊尔	政府					
台湾	医疗	台大医院	ntuh.gov.tw			
泰国	运营商	CAT				
土耳其	科技	科学技术研究理事会	tubitak.gov.tr	421KB	数据表	2020
印度	医疗	阿波罗医院	apollohospitals.com			
印度	政府	印度出入境	UCF	95.2GB	数据表	2020
巴基斯坦	政府	旁遮普省反恐中心邮服数据		1.43GB	邮件	2021.05 - 2022.01
哈萨克斯坦	运营商	Beeline通讯公司	beeline.kz	637GB	话单、用户表	2019 - 2020
哈萨克斯坦	运营商	Tele2通讯公司	tele2.kz	1.09TB	话单、用户表	2019 - 2020
哈萨克斯坦	运营商	Telecom固话运营商	telecom.kz	257GB	话单、用户表	2021.05
哈萨克斯坦	政府	养老金	enpf.kz	1.92GB	用户表	2019.12
马来西亚	运营商	DIGI通讯公司		89.5GB	话单、基站表	2021.05
台湾	教育	圆桌教育基金会		1.23GB	用户表	2020.06
泰国	运营商	Ais通讯公司	www.ais.co.th	17.7GB	数据表	2020.06
泰国	政府	外交部	mfa.go.th	3.33GB	邮件	2021.05 - 2021.09
泰国	政府	国家情报局	nia.go.th	326MB	邮件	2022.01

OSINT Data Challenge

- Too much data, spend 90% of your sifting through a wasteland of data
- Arduous manual review processes
- Lack of actionable reporting

This begs the need for AI Computer Vision

AI Computer Vision for OSINT Analysis

- Use Optical Character Recognition (OCR) to identify license plates, signs, and content in documents (passports, licenses, etc.)
- Use Object Detection to sort images by objects identified and narrow your scope in an automated
- Use Video Computer Vision to identify objects, people, signs, and even Audio analysis for conversations.
- Automate flow into reporting, dashboard, map plotting, etc.

Optical Character Recognition

Optical Character Recognition (OCR)

- Tesseract OCR is an optical character reading engine developed by HP laboratories in 1985 and open sourced in 2005.
- Tesseract has Unicode (UTF-8) support and can recognize more than 100 languages “out of the box” and thus can be used for building different language scanning software .

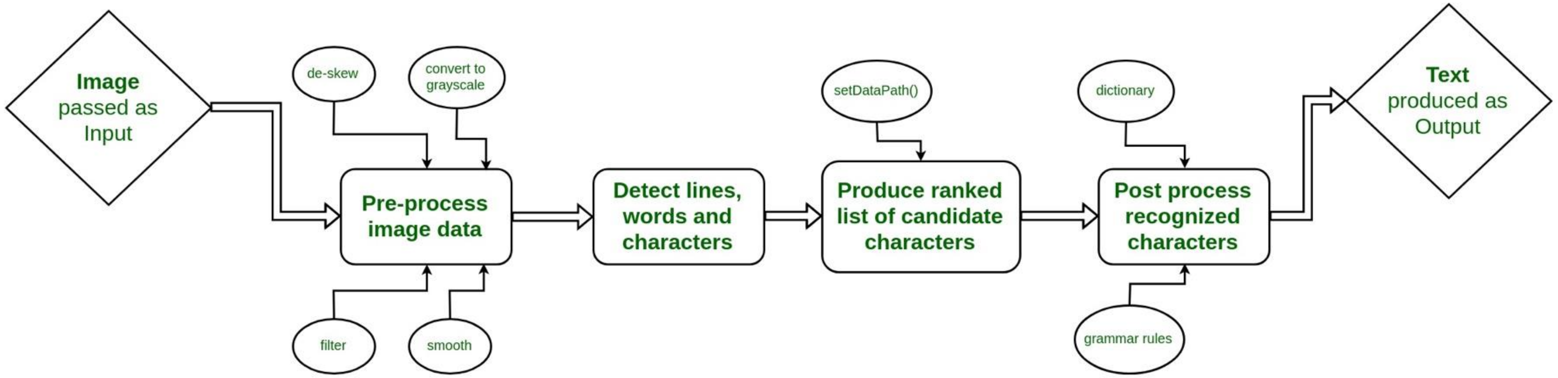


Image passed as Input

de-skew

convert to grayscale

filter

smooth

Pre-process image data

Detect lines, words and characters

setDataPath()

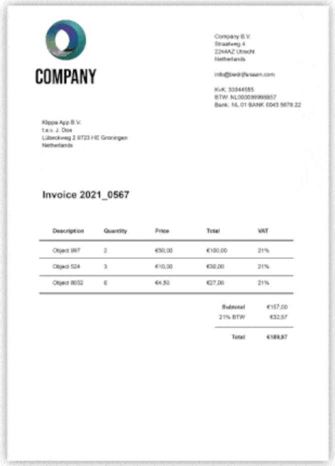
Produce ranked list of candidate characters

dictionary

grammar rules

Post process recognized characters

Text produced as Output



COMPANY

Company B.V.
Straatweg 4
2244AZ Utrecht
Netherlands

info@bedrijfsnaam.com

KvK: 33344555
BTW: NL000099998B57
Bank: NL 01 BANK 0043 5678 22

Klippa App B.V.
t.a.v. J. Doe
Lübeckweg 2 9723 HE Groningen
Netherlands

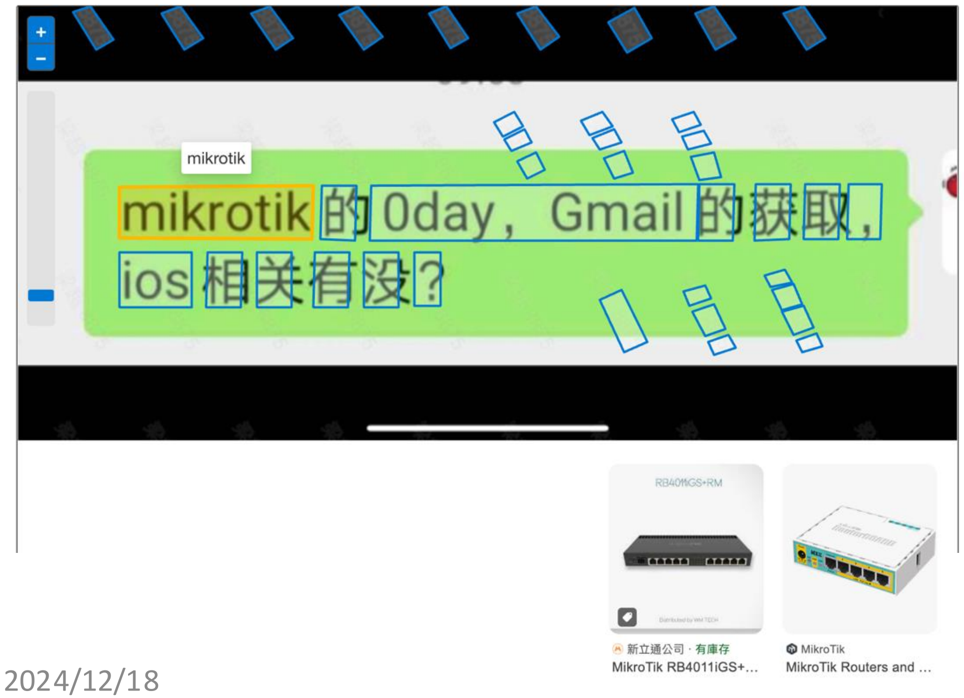
Invoice 2021_0567

Description	Quantity	Price	Total	VAT
Object 997	2	€50,00	€100,00	21%
Object 524	3	€10,00	€30,00	21%
Object 8032	6	€4,50	€27,00	21%

Subtotal	€157,00
21% BTW	€32,97
Total	**€189,97**

OSINT image OCR using Azure - 1

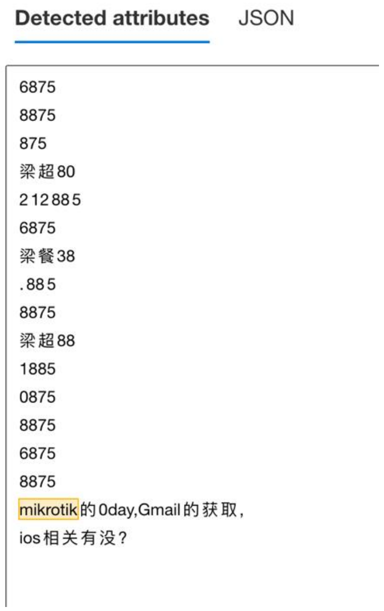
- Seamless Integration with Azure Cognitive Services
- Automated Intelligence Gathering



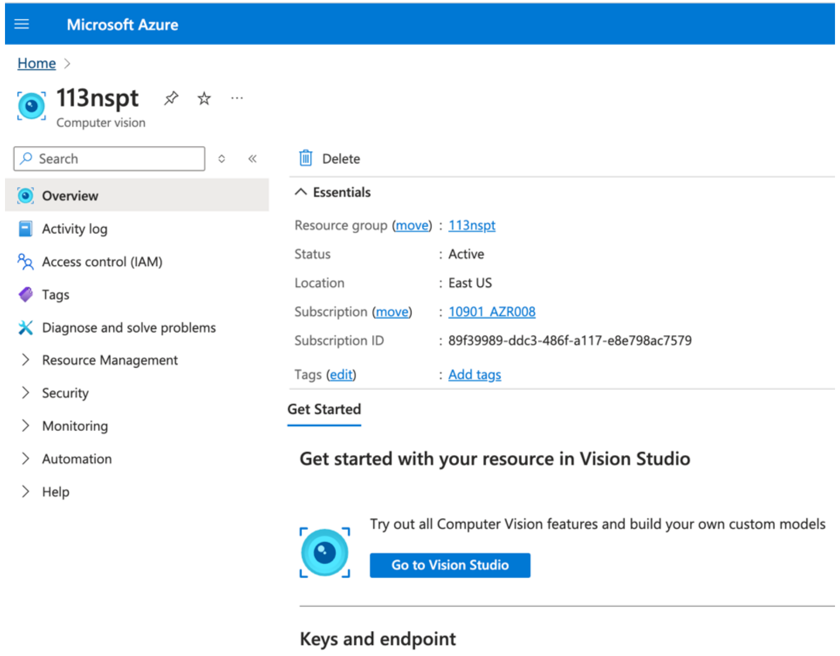
mikrotik的0day， Gmail的获取，
ios相关有没？

新立通公司 · 有庫存
MikroTik RB4011iGS+…

MikroTik
MikroTik Routers and …



```
6875
8875
875
梁超80
212885
6875
梁餐38
.885
8875
梁超88
1885
0875
8875
6875
8875
mikrotik的0day,Gmail的获取,
ios相关有没?
```



113nspt
Computer vision

Essentials

Resource group (move) : 113nspt
Status : Active
Location : East US
Subscription (move) : 10901_AZR008
Subscription ID : 89f39989-ddc3-486f-a117-e8e798ac7579
Tags (edit) : Add tags

Get Started

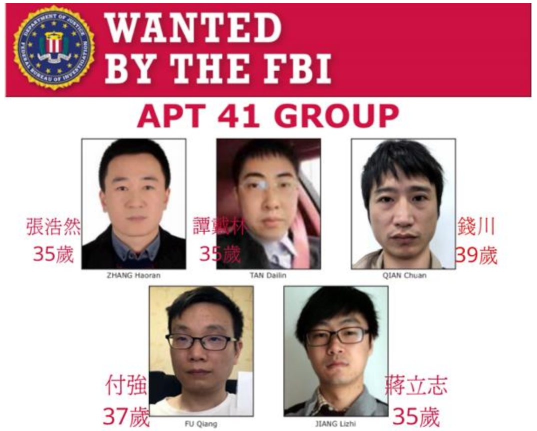
Get started with your resource in Vision Studio

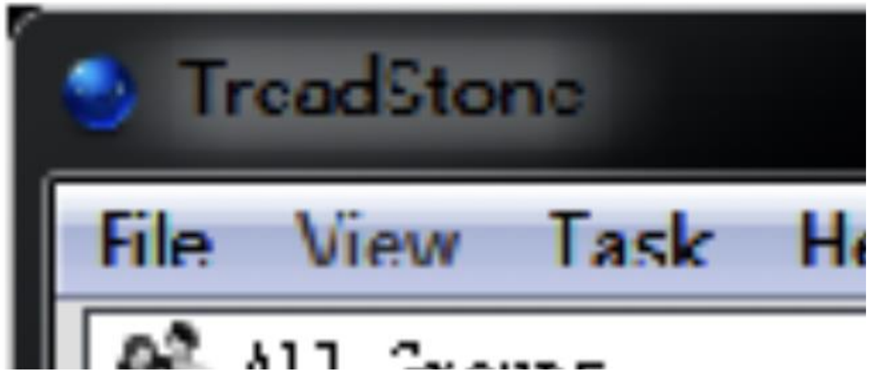
Try out all Computer Vision features and build your own custom models

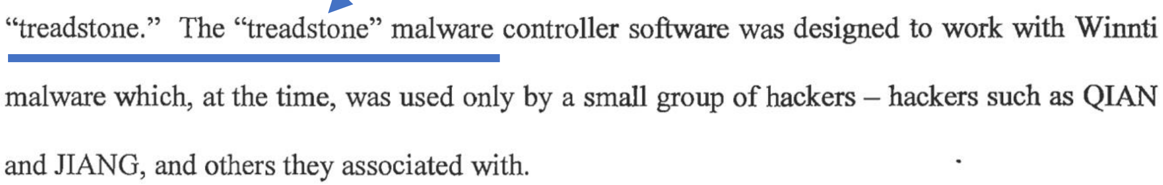
Keys and endpoint

OSINT image OCR using Azure - 2

- Suspected use of tools developed by APT41 (Chengdu 404)

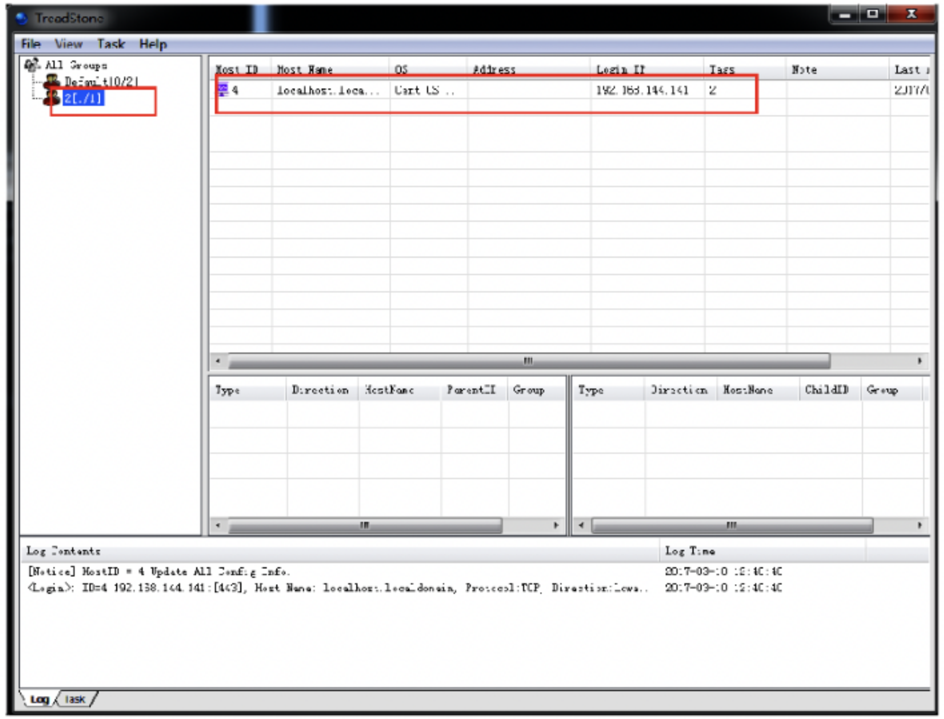






"treadstone." The "treadstone" malware controller software was designed to work with Winnti malware which, at the time, was used only by a small group of hackers – hackers such as QIAN and JIANG, and others they associated with.

1.7.5　产品图片

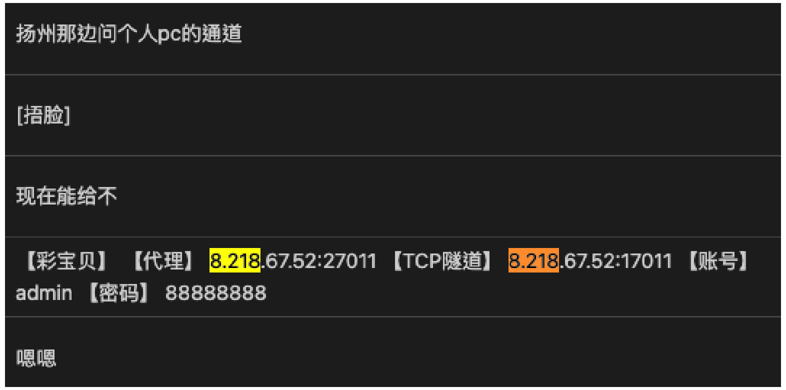


```
Log Contents                                                                                     Log Time
[Notice] HostID = 4 Update All Config Info.                                                      2017-03-10 12:40:40
<Login>: ID=4 192.138.144.141:[443], Host Name: localhost.localdomain, Protocol:TCP, Direction:Lowa..   2017-03-10 12:40:40
```

（Linux 远程控制系统界面图）

OSINT image OCR using Azure - 3

- In late September 2022, malicious attackers launched a supply chain attack using the chat program Comm100.
  - Windows product name
  - Events with event ID 6005 (the event log service was started), events with Event ID 6006 (the event log service was stopped)
  - Any TCP endpoints listening on ports 8090, 8091, 8092, 8093, 8094, 8095, 8096
  - Any TCP endpoint established to 8.218.67.52:18024
  - Antivirus product name
  - If Skype or Telegram is installed
  - Number of connected monitors
  - Product ID value from Windows registry



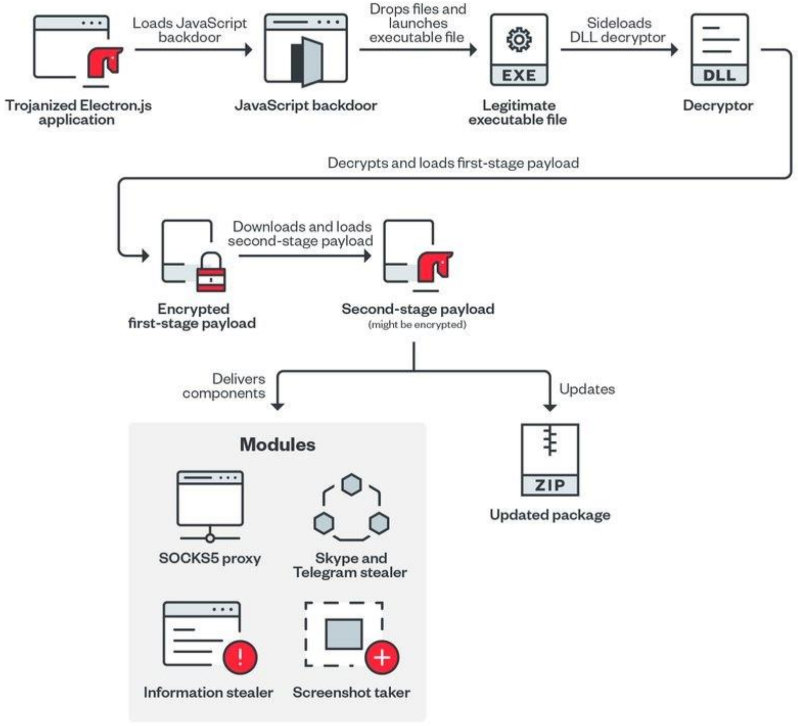
扬州那边问个人pc的通道

[捂脸]

现在能给不

【彩宝贝】 【代理】 8.218.67.52:27011 【TCP隧道】 8.218.67.52:17011 【账号】 admin 【密码】 88888888

嗯嗯


Trojanized Electron.js application
Loads JavaScript backdoor
JavaScript backdoor
Drops files and launches executable file
EXE
Legitimate executable file
Sideloads DLL decryptor
DLL
Decryptor
Decrypts and loads first-stage payload
Encrypted first-stage payload
Downloads and loads second-stage payload
Second-stage payload
(might be encrypted)
Delivers components
Updates
Modules
SOCKS5 proxy
Skype and Telegram stealer
Information stealer
Screenshot taker
ZIP
Updated package

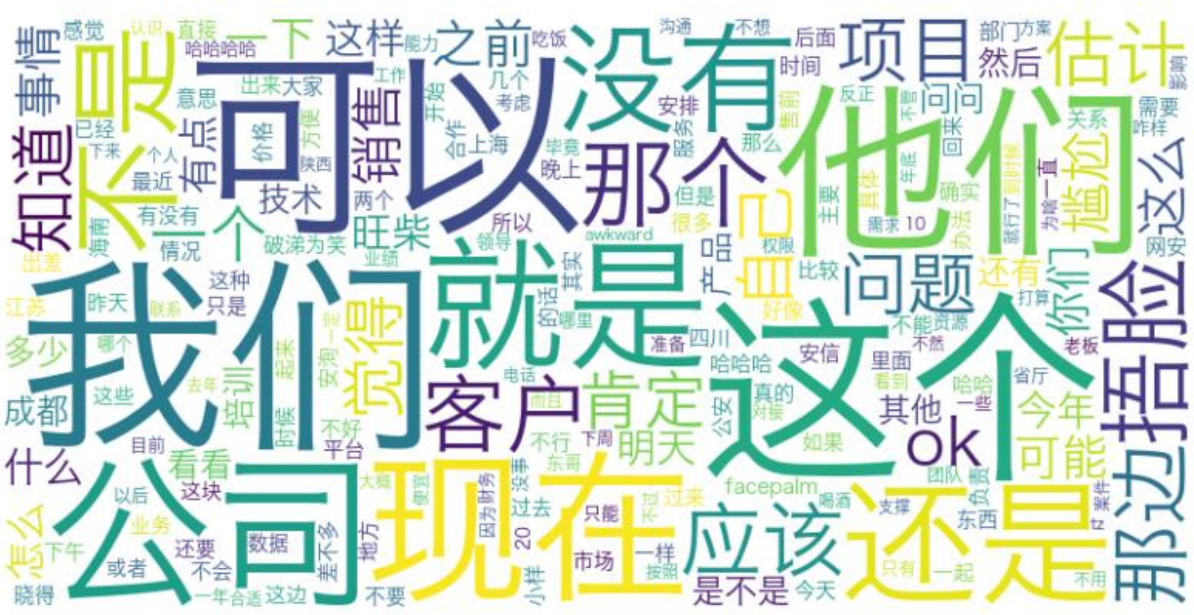
©2022 TREND MICRO

OSINT Take OCR to Word Cloud

The data leak incident lies in its ability to visualize the high-frequency keywords from the leaked conversation, revealing key discussion points and topics.

We can see…

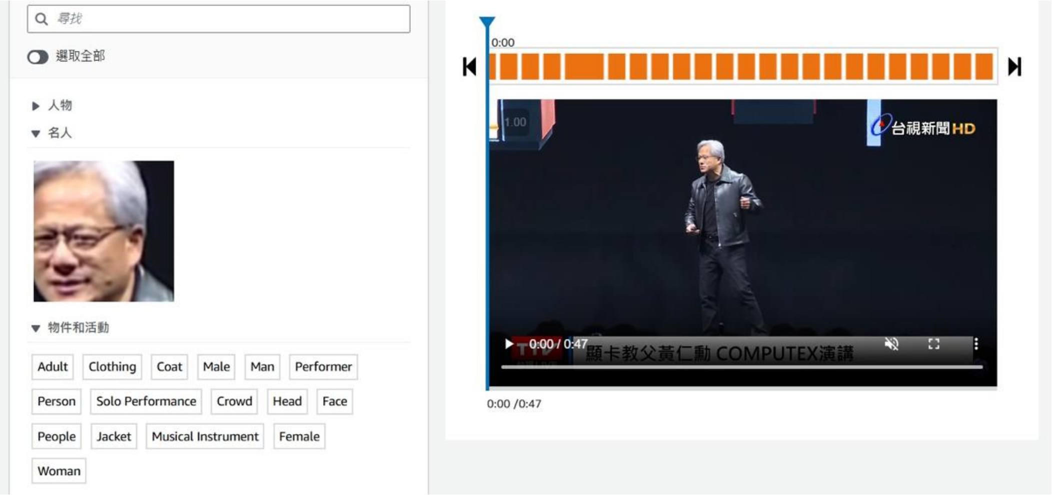
- Words like **"company," "project," "client"** may involve sensitive information related to business activities.
- Words like **"problem," "estimate," and "technology"** might reflect technical details or potential issues discussed internally.
- High-frequency words such as **"we" and "now"** indicate internal collaboration contexts and concerns.



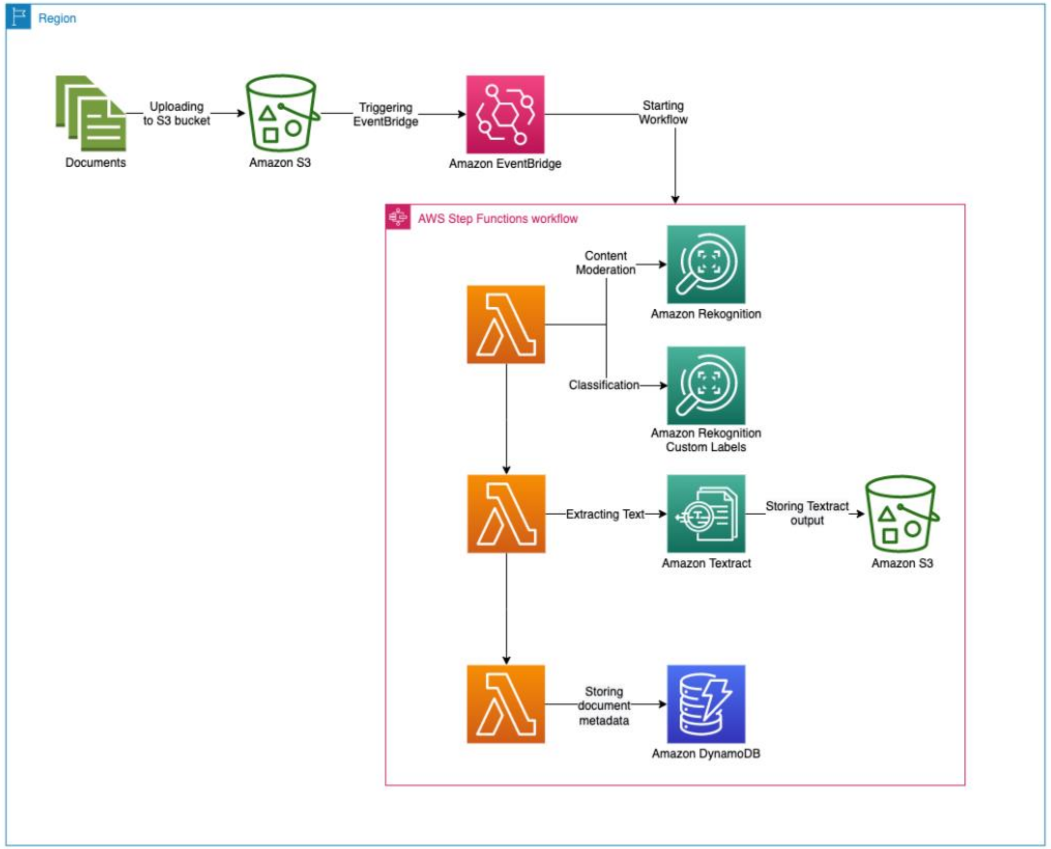
Object Detection

OSINT Object Detection

OSINT Object Detection involves identifying and locating objects of interest within images or videos collected from publicly available sources.

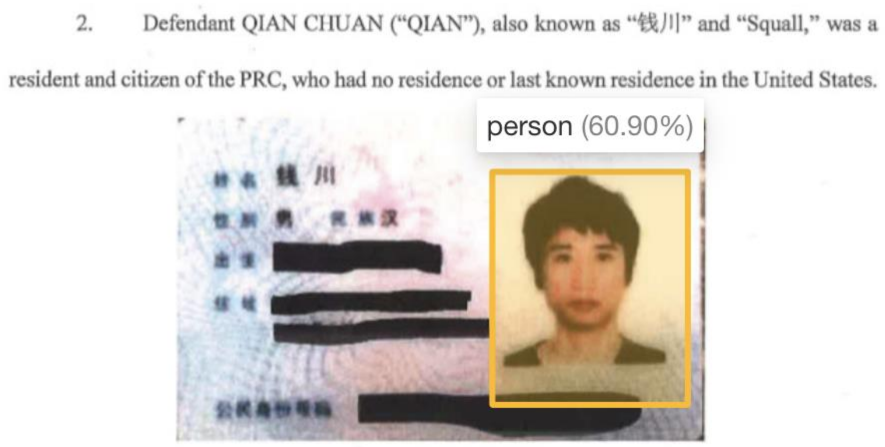


Example using AWS



OSINT Object Detection with Azure

- Detect and extract bounding boxes based on thousands of recognizable objects and living beings.
- We can confirm identity by comparing similar photos from different documents.



2. Defendant QIAN CHUAN (“QIAN”), also known as “钱川” and “Squall,” was a resident and citizen of the PRC, who had no residence or last known residence in the United States.

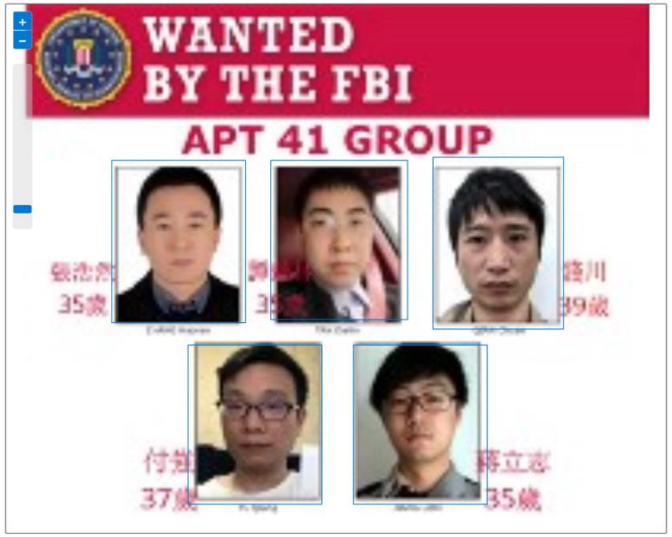
姓名 钱川

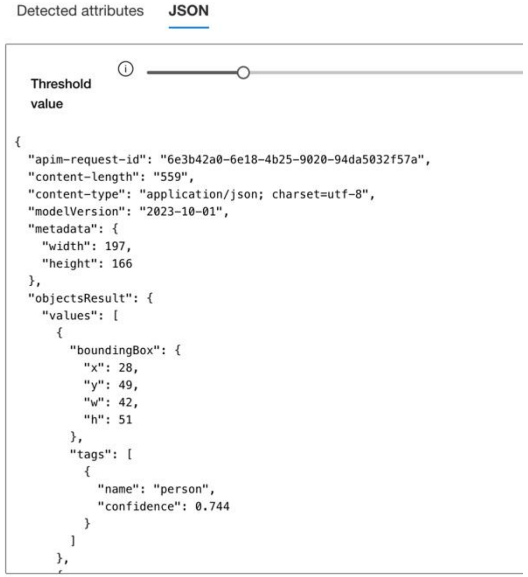
性别 男 民族 汉

出生 [redacted]

住址 [redacted]

公民身份号码 [redacted]

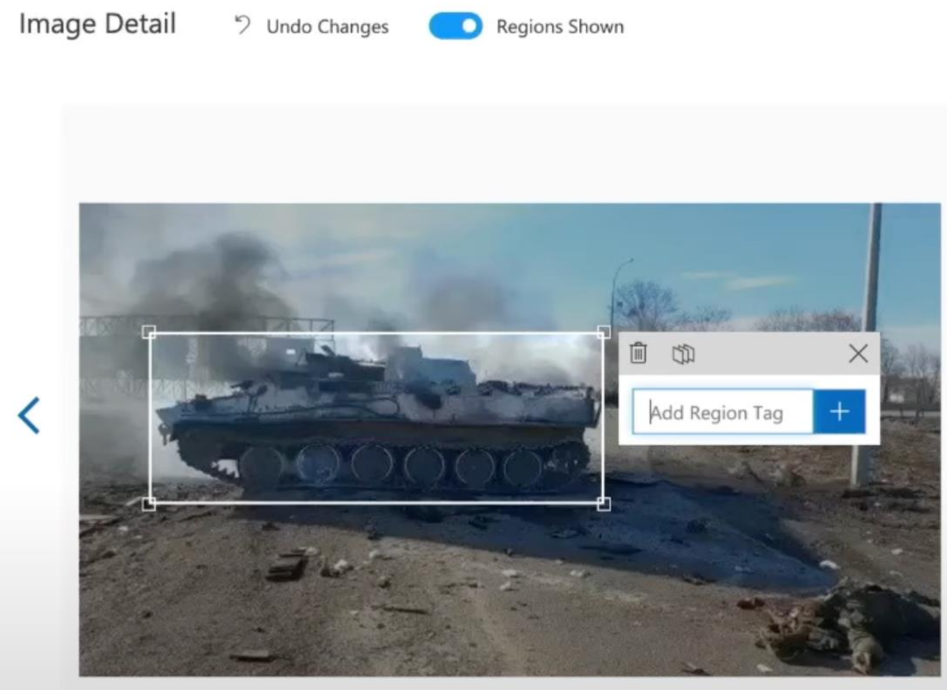


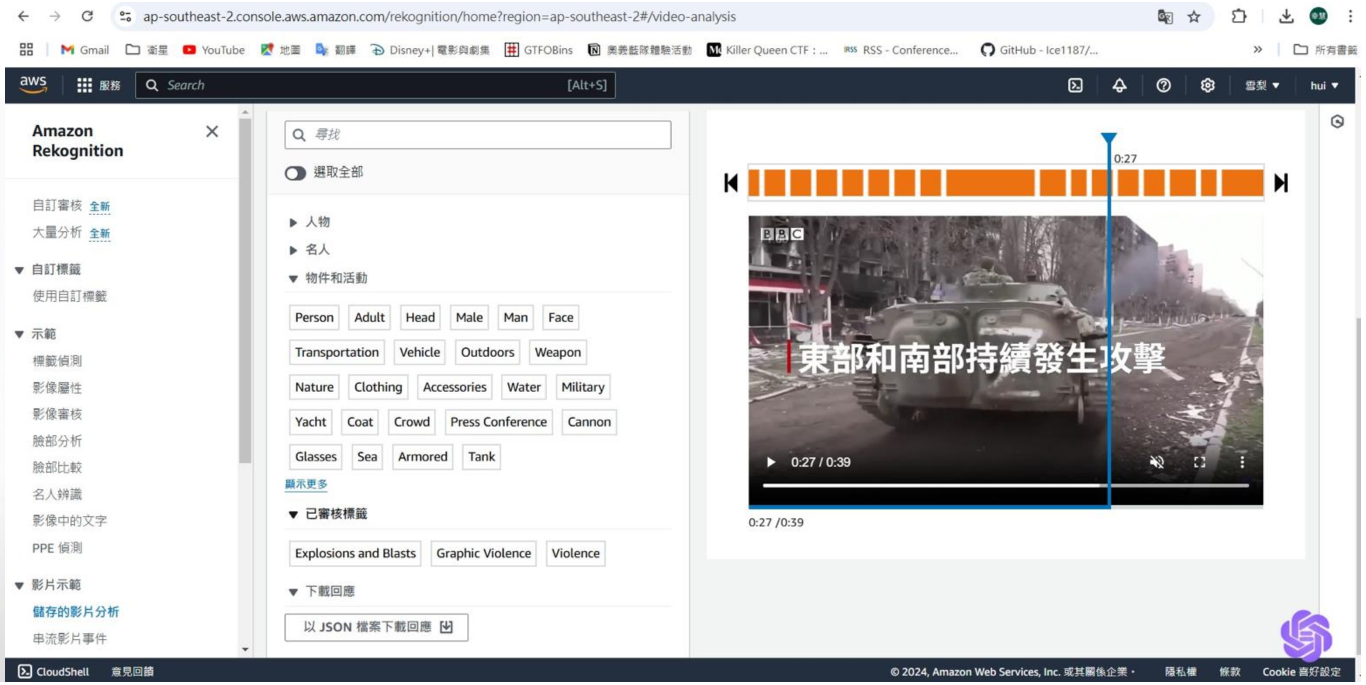


```
{
  "apim-request-id": "6e3b42a0-6e18-4b25-9020-94da5032f57a",
  "content-length": "559",
  "content-type": "application/json; charset=utf-8",
  "modelVersion": "2023-10-01",
  "metadata": {
    "width": 197,
    "height": 166
  },
  "objectsResult": {
    "values": [
      {
        "boundingBox": {
          "x": 28,
          "y": 49,
          "w": 42,
          "h": 51
        },
        "tags": [
          {
            "name": "person",
            "confidence": 0.744
          }
        ]
      },
```

OSINT Object Detection Train Model

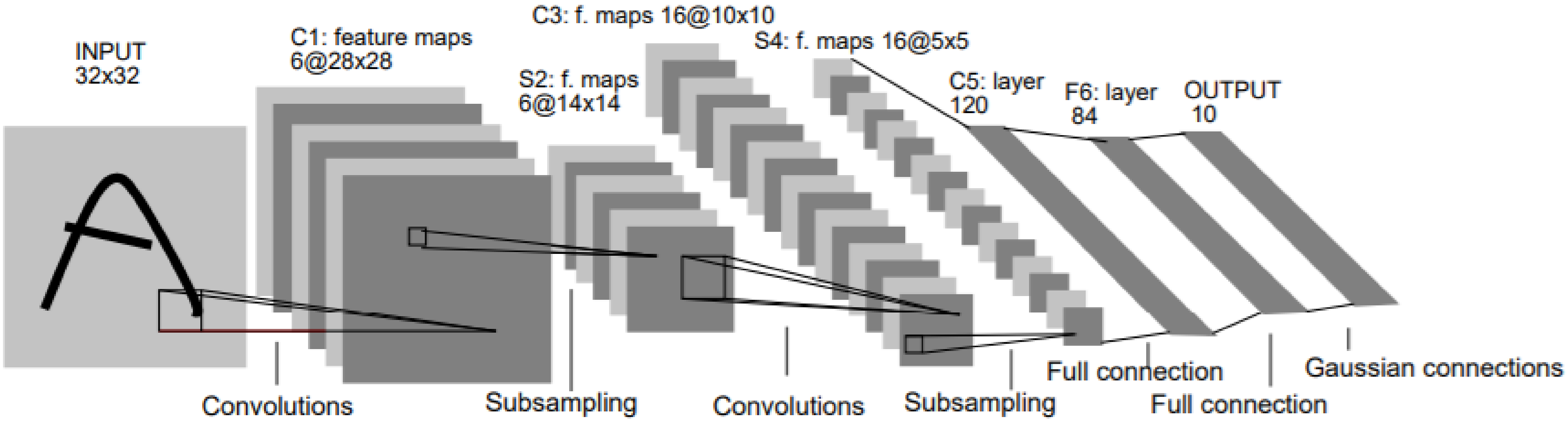
- Azure Object Detection uses optimized Convolutional Neural Networks (CNNs) for feature extraction and object recognition.




ap-southeast-2.console.aws.amazon.com/rekognition/home?region=ap-southeast-2#/video-analysis

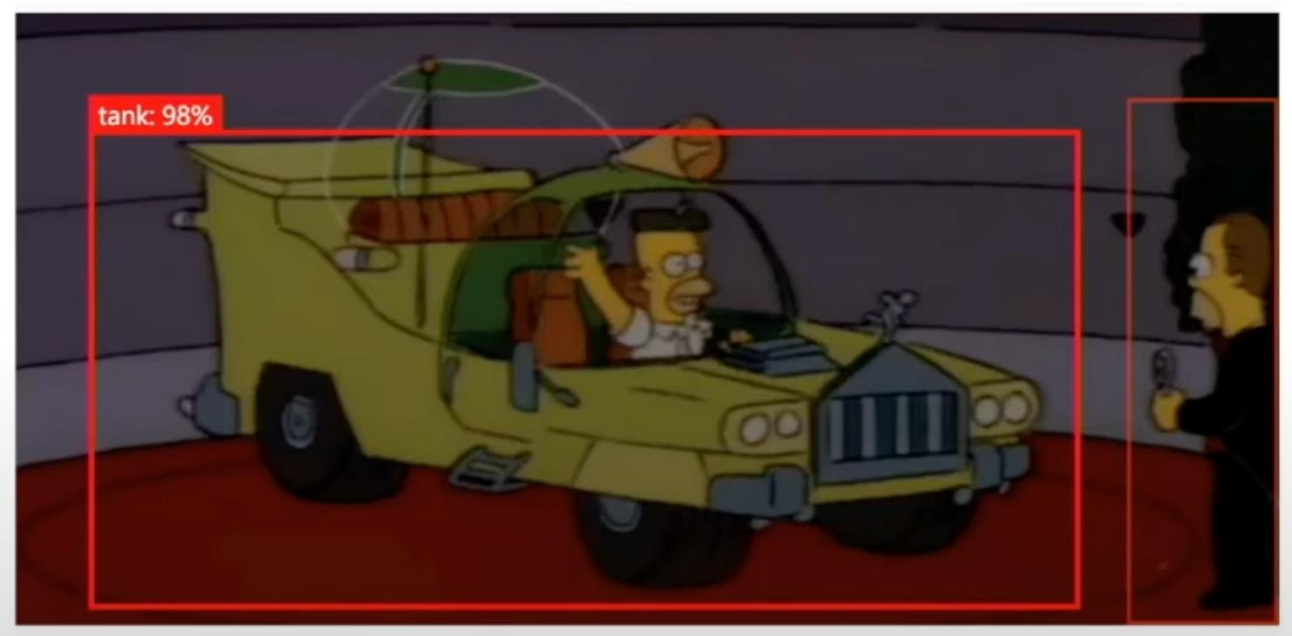
CNN Model

- CNN (Convolutional Neural Network) is a deep learning model designed explicitly for processing structured grid data, such as images.



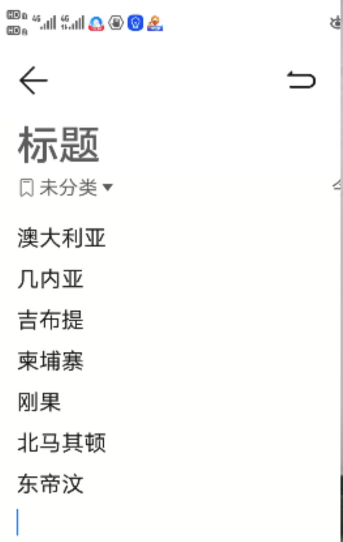
OSINT Object Detection Testing Model

- The model may have misclassified the car as a tank due to an insufficient or imbalanced dataset.
- The CNN model likely over-relied on low-level features such as shape or color, failing to distinguish the high-level semantics.



Why we need OCR and OD?

- Analyze the location of the victim and add geographical coordinates, which is digital identification.



标题

未分类▾

澳大利亚

几内亚

吉布提

柬埔寨

刚果

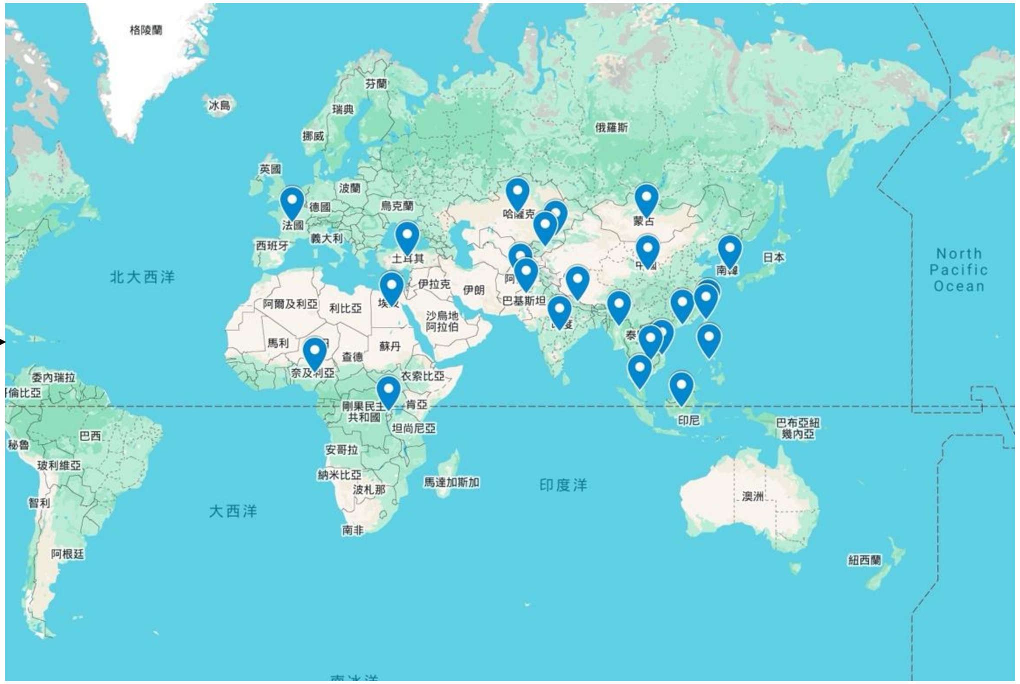
北马其顿

东帝汶

epfindia.gov.in 印度雇员公积金组织

AaBb AaBbC AaBb

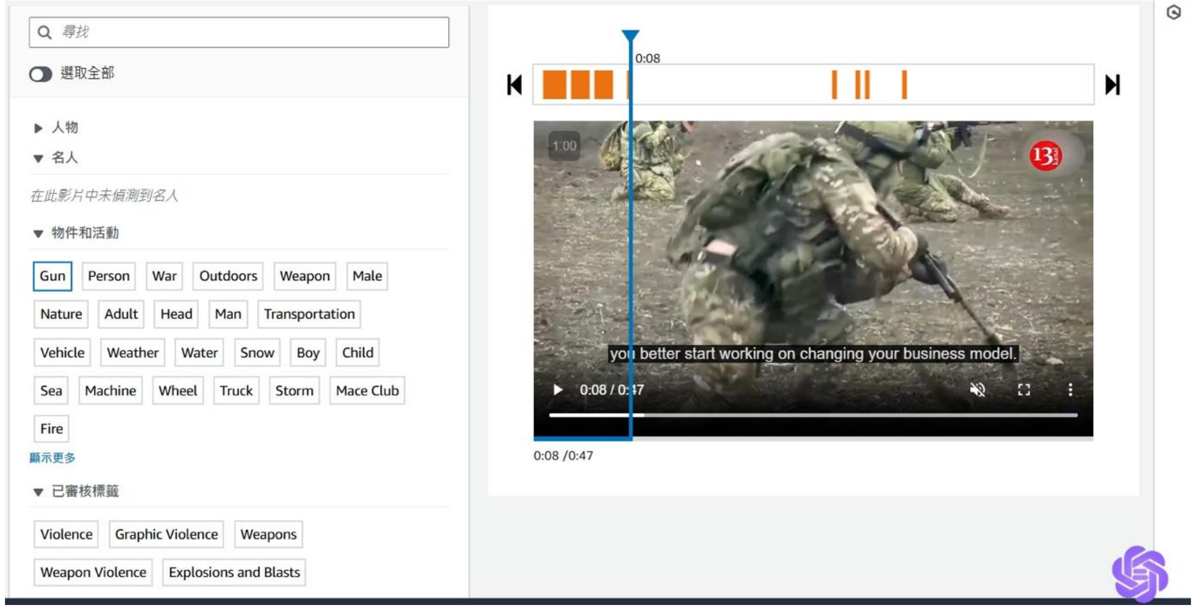
locate



OSINT video analysis using AWS

Video analysis using a modified version of AWS Rekognition can be used to identify

- The objective throughout the entire video
- Frequency of that objective
- Location in the timeline
- Missed the human eye due to subtleties and backgrounds

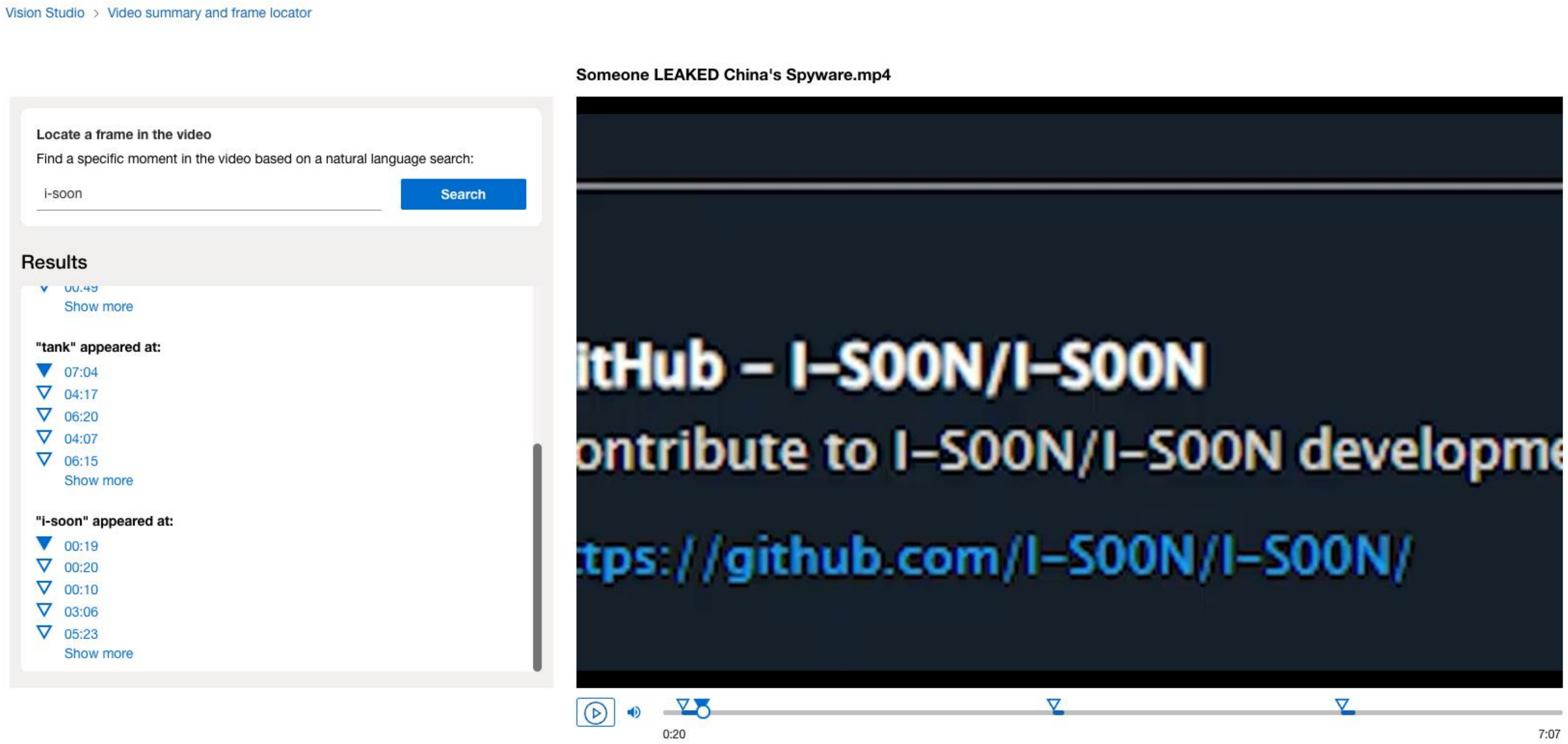


在此影片中未偵測到名人

you better start working on changing your business model.

OSINT video analysis using Azure

- Video search and summarization uses a combination of natural language processing and computer vision techniques to analyze the content of a video.



Vision Studio > Video summary and frame locator

Someone LEAKED China's Spyware.mp4

Locate a frame in the video

Find a specific moment in the video based on a natural language search:

i-soon

Search

Results

00:49
Show more

"tank" appeared at:
07:04
04:17
06:20
04:07
06:15
Show more

"i-soon" appeared at:
00:19
00:20
00:10
03:06
05:23
Show more

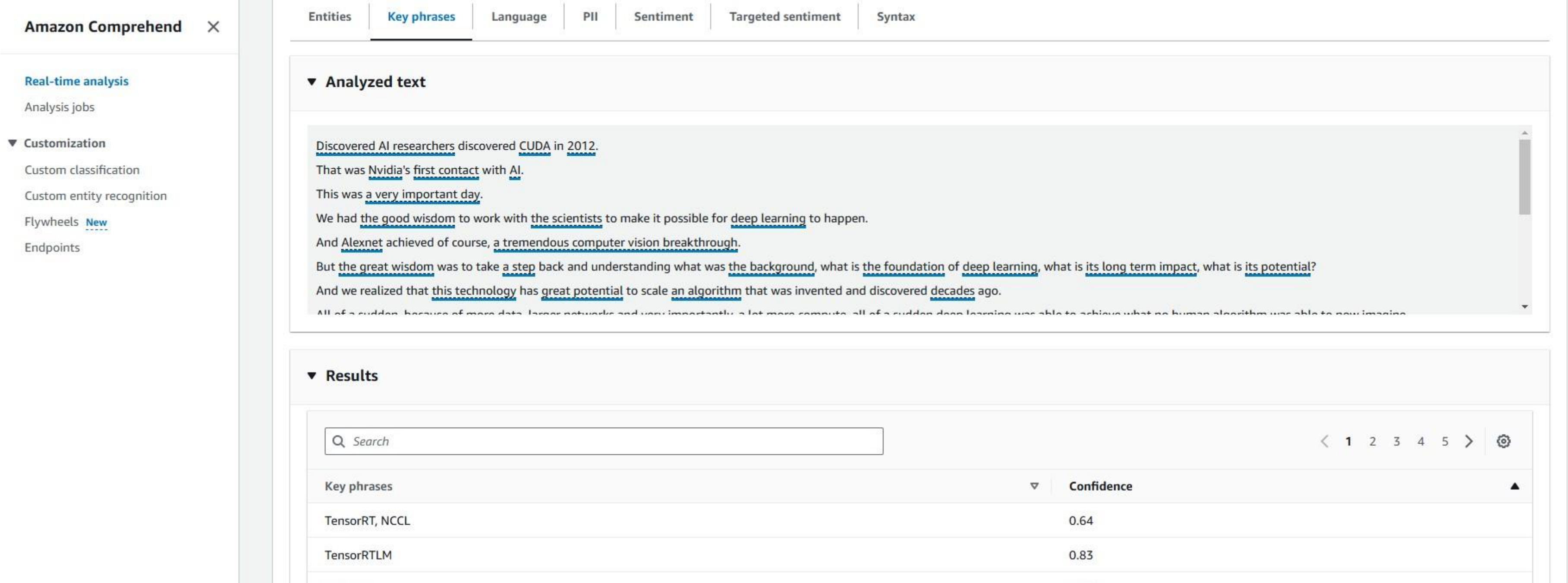
itHub – I-SOON/I-SOON
ontribute to I-SOON/I-SOON developme
tps://github.com/I-SOON/I-SOON/

0:20
7:07

Speech Recognition

OSINT Speech Recognition

- Automatic Speech Recognition (ASR), also known as Speech to Text (STT), is the task of transcribing a given audio to text. It has many applications, such as voice user interfaces.

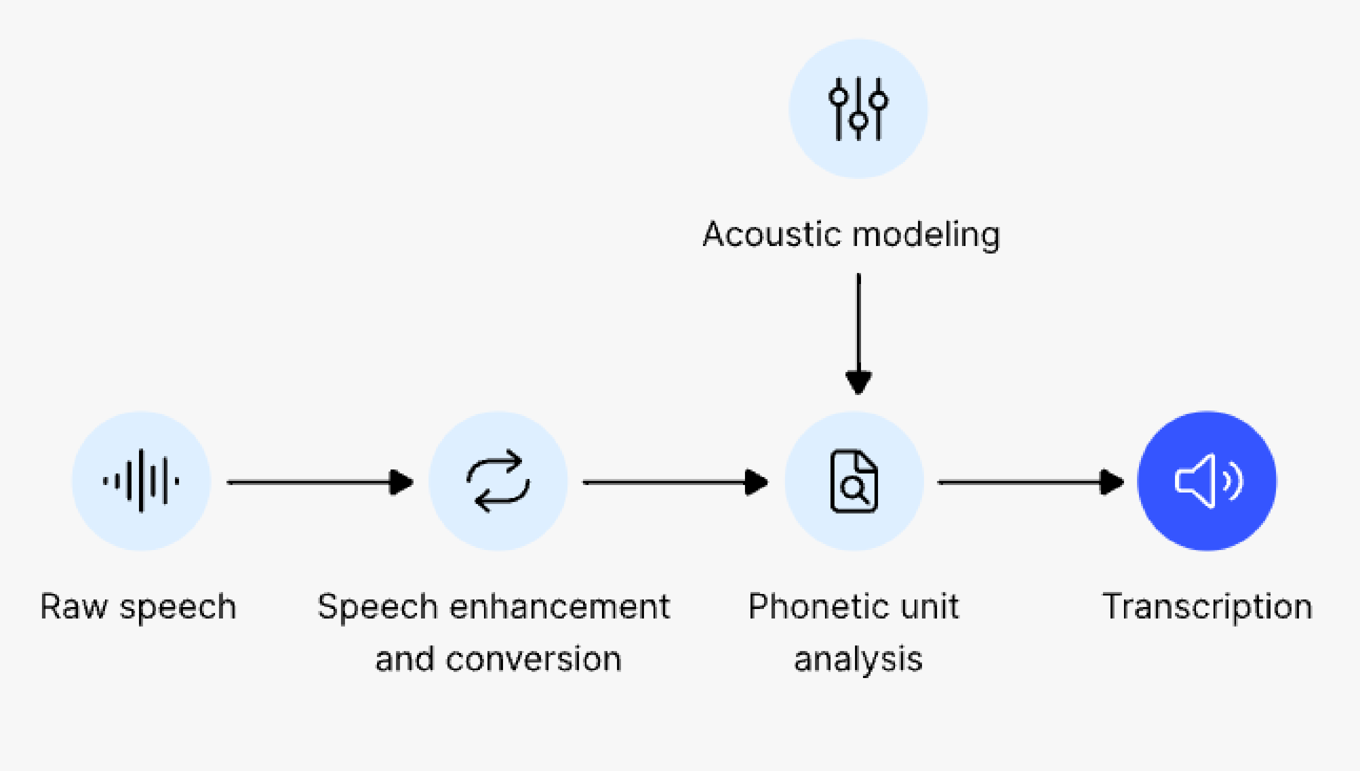


Discovered AI researchers discovered CUDA in 2012.
That was Nvidia's first contact with AI.
This was a very important day.
We had the good wisdom to work with the scientists to make it possible for deep learning to happen.
And Alexnet achieved of course, a tremendous computer vision breakthrough.
But the great wisdom was to take a step back and understanding what was the background, what is the foundation of deep learning, what is its long term impact, what is its potential?
And we realized that this technology has great potential to scale an algorithm that was invented and discovered decades ago.
All of a sudden, because of more data, larger networks and very importantly, a lot more compute, all of a sudden deep learning was able to achieve what no human algorithm was able to now imagine.

Key phrases	Confidence
TensorRT, NCCL	0.64
TensorRTLM	0.83

Speech Recognition Flow

- In the I-Soon Leak Incident, we don’t collect related speech files.
- We only introduce the system flow.



Limitation and Vulnerability

OCR Problem

There are several common reasons why OCR might fail to recognize text

- **Image quality**
  - Low-resolution scans, poor image clarity, or text distortion can hamper OCR accuracy.
- **Complex fonts and characters**
  - OCR may struggle with non-standard or decorative fonts, unusual characters, or handwriting.
- **Background images**
  - Pages with distracting background images may interfere with text recognition.
- **Improper OCR settings**
  - Incorrect language selection, low confidence thresholds, or other misconfigurations can impact results.

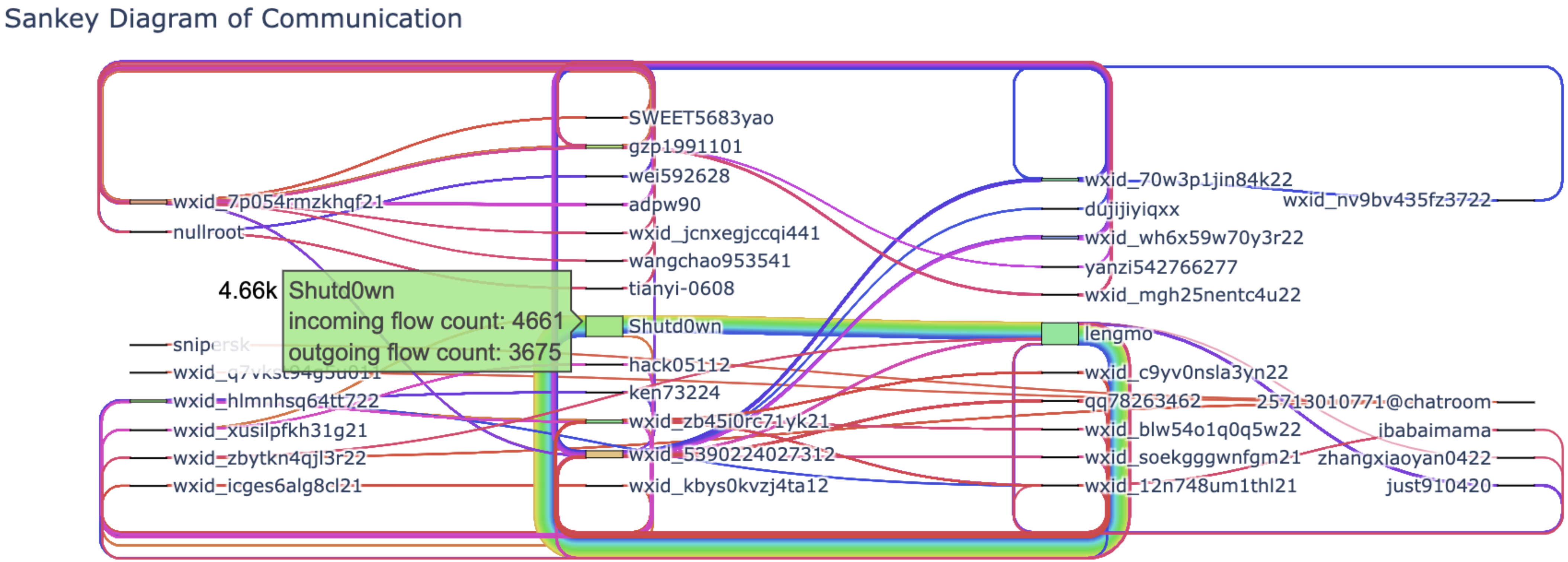
Speech Recognition Problem

There are still numerous challenges that researchers and developers face when it comes to perfecting speech recognition systems.

- **Variability in Speech Patterns**
  - People speak at different speeds, with varying accents, dialects, and pronunciations.
- **Background Noise and Environmental Factors**
  - Real-world environments are often filled with ambient noise, such as traffic sounds, machinery, or people talking in the background.
- **Vocabulary and Contextual Understanding**
  - The technology must also consider the context in which words are spoken to ensure accurate transcription.
- **Speaker Independence**
  - variations in pitch, tone, and pronunciation between different speakers can impact the algorithm's performance.

Conclusion

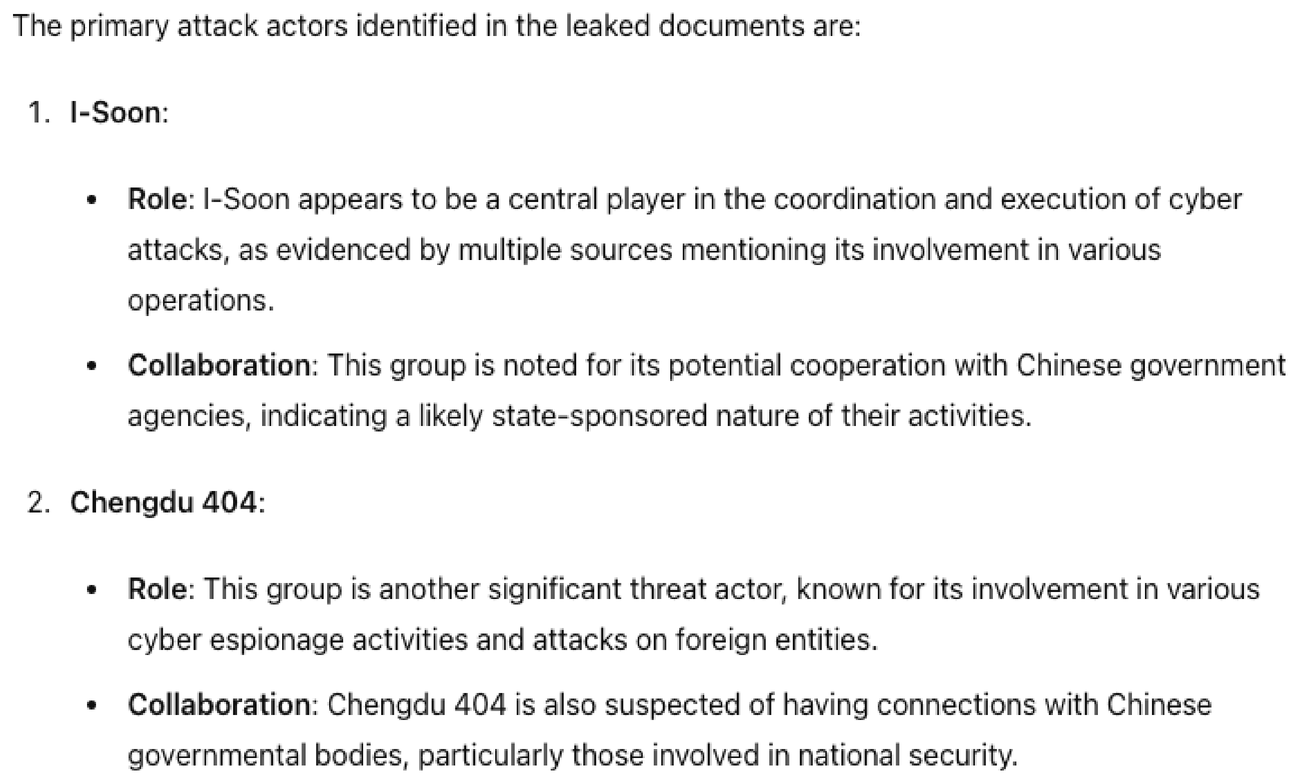
I-Soon Leak Summary


Sankey Diagram of Communication

- shutd0wn＝吴海波=Wu HaiBo=吳總
- lengmo＝陈诚=Chen Cheng=C總

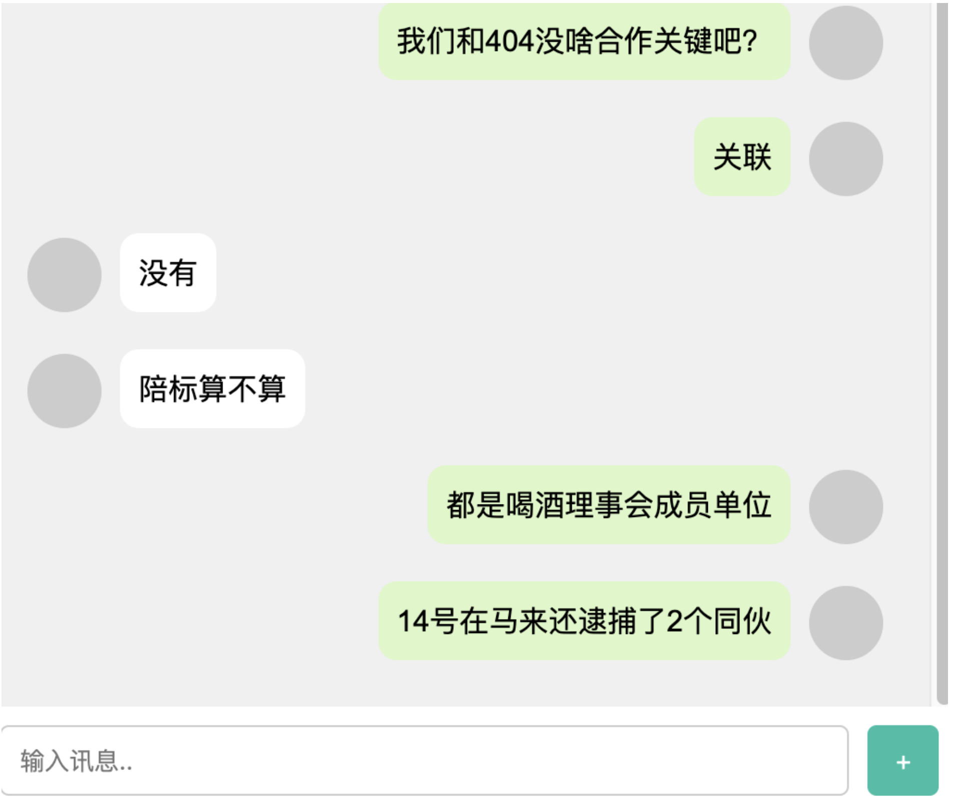
I-Soon Leak Summary – Primary Attack Actors

- Who are the primary attack actors?



The primary attack actors identified in the leaked documents are:

1. **I-Soon**:
    - **Role**: I-Soon appears to be a central player in the coordination and execution of cyber attacks, as evidenced by multiple sources mentioning its involvement in various operations.
    - **Collaboration**: This group is noted for its potential cooperation with Chinese government agencies, indicating a likely state-sponsored nature of their activities.
2. **Chengdu 404**:
    - **Role**: This group is another significant threat actor, known for its involvement in various cyber espionage activities and attacks on foreign entities.
    - **Collaboration**: Chengdu 404 is also suspected of having connections with Chinese governmental bodies, particularly those involved in national security.



我们和404没啥合作关键吧?

关联

没有

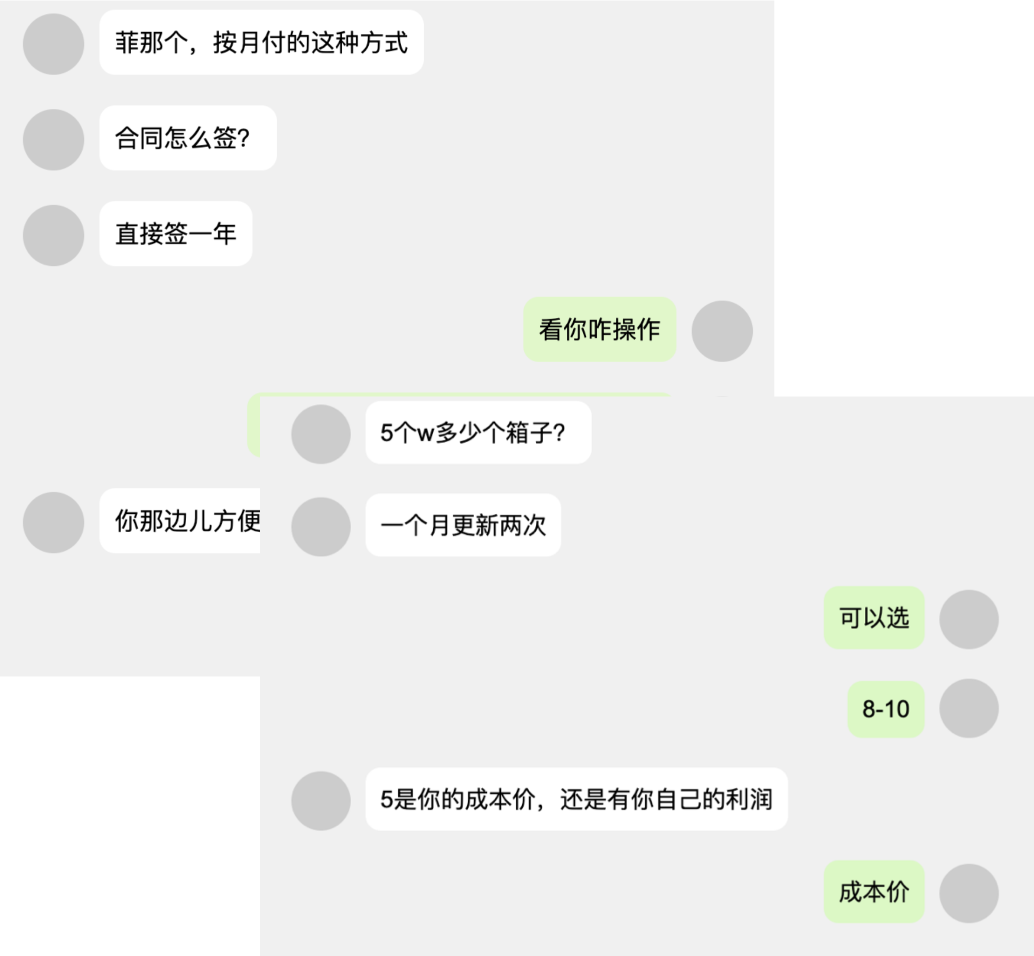
陪标算不算

都是喝酒理事会成员单位

14号在马来还逮捕了2个同伙

I-Soon Leak Summary – Business Mode

- Counter-Terrorism Data Support
- Technical Services
- Initiating Talent Cultivation Programs
- Comprehensive Laboratory Development



菲那个，按月付的这种方式

合同怎么签?

直接签一年

看你咋操作

你那边儿方便

5个w多少个箱子?

一个月更新两次

可以选

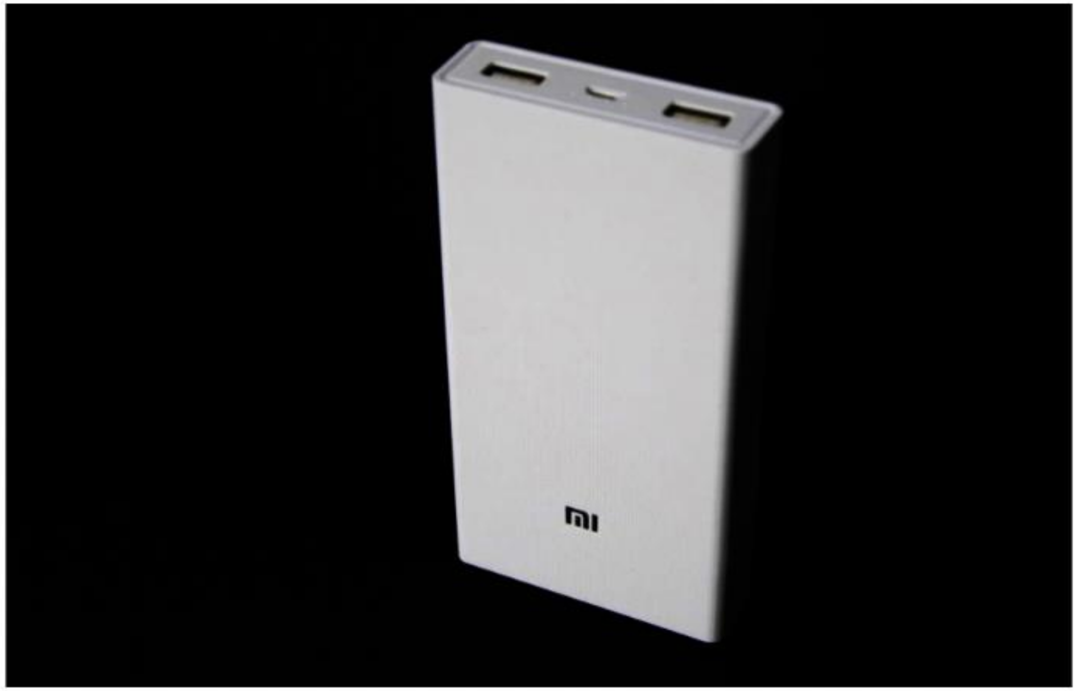
8-10

5是你的成本价，还是有你自己的利润

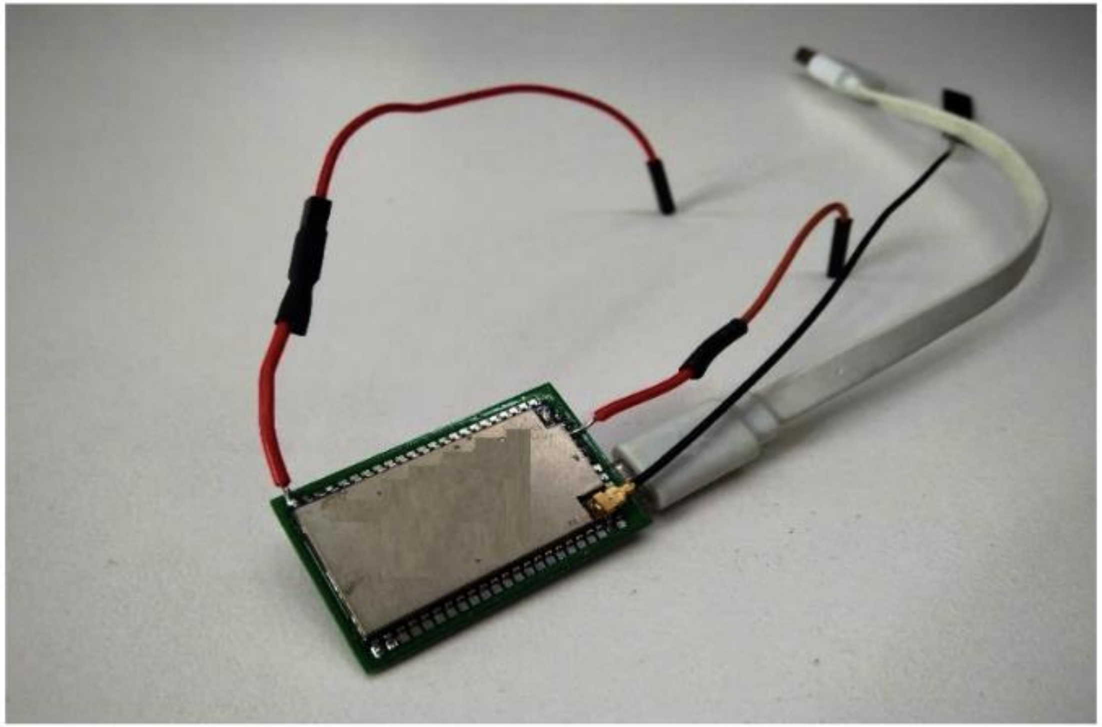
成本价

I-Soon Leak Summary – Product

2.1.5 产品图片



（WiFi 抵近攻击系统（基础版）产品实物图）

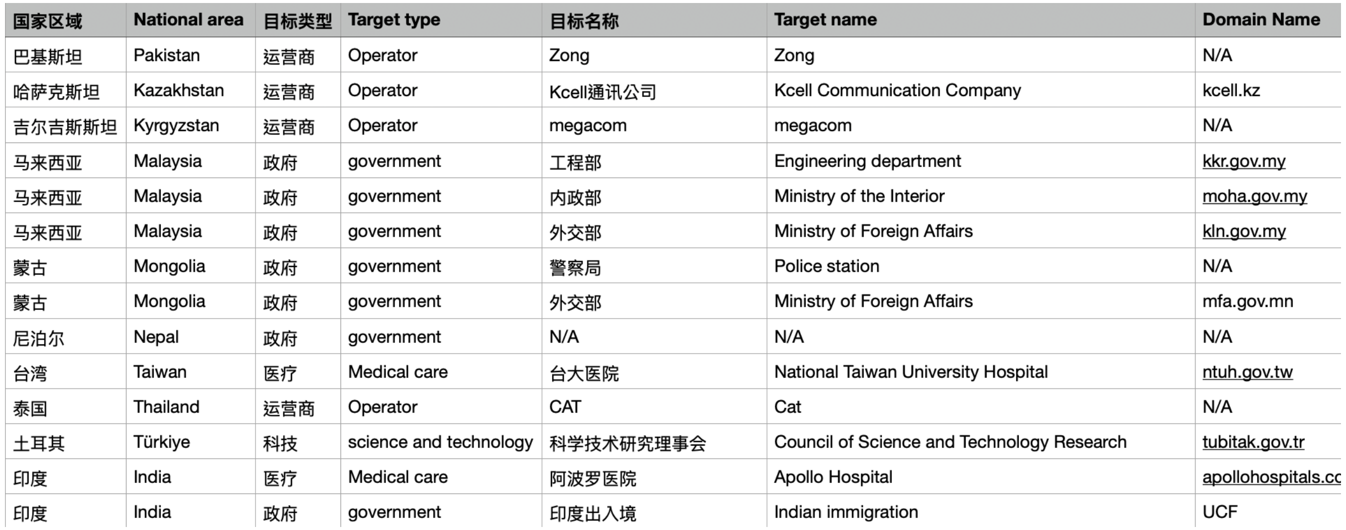


（WiFi 抵近攻击系统（mini 版）产品实物图）

I-Soon Leak Summary – Victim

Taiwan Units Targeted by I-SOON Attacks:

- NTU Hospital: Patient ratio data
- Yuan-Ze Education Foundation: Intranet control
- NTU Institute of Applied Mechanics: Hash passwords, web shell
- Tamkang University: Email privileges



国家区域	National area	目标类型	Target type	目标名称	Target name	Domain Name
巴基斯坦	Pakistan	运营商	Operator	Zong	Zong	N/A
哈萨克斯坦	Kazakhstan	运营商	Operator	Kcell通讯公司	Kcell Communication Company	kcell.kz
吉尔吉斯斯坦	Kyrgyzstan	运营商	Operator	megacom	megacom	N/A
马来西亚	Malaysia	政府	government	工程部	Engineering department	kkr.gov.my
马来西亚	Malaysia	政府	government	内政部	Ministry of the Interior	moha.gov.my
马来西亚	Malaysia	政府	government	外交部	Ministry of Foreign Affairs	kln.gov.my
蒙古	Mongolia	政府	government	警察局	Police station	N/A
蒙古	Mongolia	政府	government	外交部	Ministry of Foreign Affairs	mfa.gov.mn
尼泊尔	Nepal	政府	government	N/A	N/A	N/A
台湾	Taiwan	医疗	Medical care	台大医院	National Taiwan University Hospital	ntuh.gov.tw
泰国	Thailand	运营商	Operator	CAT	Cat	N/A
土耳其	Türkiye	科技	science and technology	科学技术研究理事会	Council of Science and Technology Research	tubitak.gov.tr
印度	India	医疗	Medical care	阿波罗医院	Apollo Hospital	apollohospitals.cc
印度	India	政府	government	印度出入境	Indian immigration	UCF

Conclusion

- Computer Vision vastly improves accuracy, efficiency, and time-to-evidence.
- Allows the OSINT team to focus on the relevant data for validating narratives and nefarious activities.
- Automates reporting to leverage the human-in-the-loop for focusing on the last mile of analysis and results.

Team Contribution

Team Contribution

- 110590453 黃台茗
  - developing an OCR tool, video editing, reporting
- 111599004 陳勝舢
  - presentations, system implementation, organizing work, building Azure services
- 112C53010 林幸慧
  - researching object detection using AWS services
- 112c72011 許貽昇
  - developing AI models and APIs
- 113598023 廖哲霈
  - web scraping related to I-Soon Leak incidents on Chinese news
- 113c53019 李宇揚
  - web scraping related to I-Soon Leak incidents on English websites

Thanks!