



Applications in Data Science & AI: Cyber Intelligence - OSINT

Nisreen Cain

Key Concepts

Open Source Data

Open Source Intelligence

The Intelligence Process

Data Mining & Aggregation

AI/ML and the Intelligence Cycle

Use cases from the field

The road ahead



Disclaimer

Any Views or opinions presented in this presentation are solely mine and do not necessarily represent my employer.

- I am not a lawyer or giving you legal advice
- I am not giving you permission or authorizing you to do anything ever
- In fact don't do anything ever ;)



Open Source Data

- Published or broadcast for public consumption
- Is available on request to the public
- Is accessible online or otherwise to the public
- Is available to the public by subscription or purchase
- Could be seen or heard by any casual observer
- Is made available at a meeting open to the public
- Is obtained by visiting any place or attending any event that is open to the public

Source: Department of Defense Manual 5240.01
(August 2016)



Data is Power.

We are bombarded with a staggering amount of information from various online sources.

Hundreds of thousands of hours of videos, millions of images, more text than can be indexed by search engines; that doesn't include data behind restricted access.

“Data is a precious thing and it will last longer than the systems themselves.” ~ Tim Berners Lee





Harnessing the Power of Data Everyday

AI and Data Science are a part of our everyday life. From smart devices, image recognition, news feeds, to cybersecurity and intelligence.



The industry demand is real

“Hirings for A.I. specialists on the career networking service have grown 74% annually over the past four years, LinkedIn said.” - Fortune.com 2019

Cyber Intelligence

The internet is the world's largest database of information and it's increasing exponentially

Need to leverage data mining and AI/ML to acquire, process, analyze, and identify threats and risks to enhance decision making across organizations





Open Source Intelligence (OSINT)

OSINT is the practice of data collection from **publicly** available sources such as social media, news outlets, blogs, websites, etc. This includes data mining and crawling techniques, data extraction and data analysis.

Open source indicates the nature of the data being collected. Doesn't refer to Open Source Software or public intelligence

A person wearing a dark hoodie is sitting at a laptop. The laptop screen and the background are overlaid with various digital icons and a network diagram. The icons include a laptop, a lightbulb, an envelope, a shield with a checkmark, a smartphone, a padlock, a person silhouette, a monitor, a battery, and a magnifying glass. The background also features a grid of hexagons and a bar chart at the bottom.

A vast haystack of visible data

- Social Media
- IP addresses
- Domain Lookup
- Reverse Image Search
- Phone numbers
- Online Documents
- Online listings
- Geolocations and images
- License Plates, etc
- Government resources - such as marriage licenses, deeds, court proceedings, etc
- ... and more

The Intelligence Process

1. Requirements
2. Planning and direction
3. Collection
4. Processing and exploitation
5. Analysis and production
6. Dissemination



OSINT Process

1. Identifying the source - where you can find the information
2. Harvesting the data - get relevant data from the source
3. Data processing - get meaning from the data
4. Analysis - joining data from different sources
5. Reporting

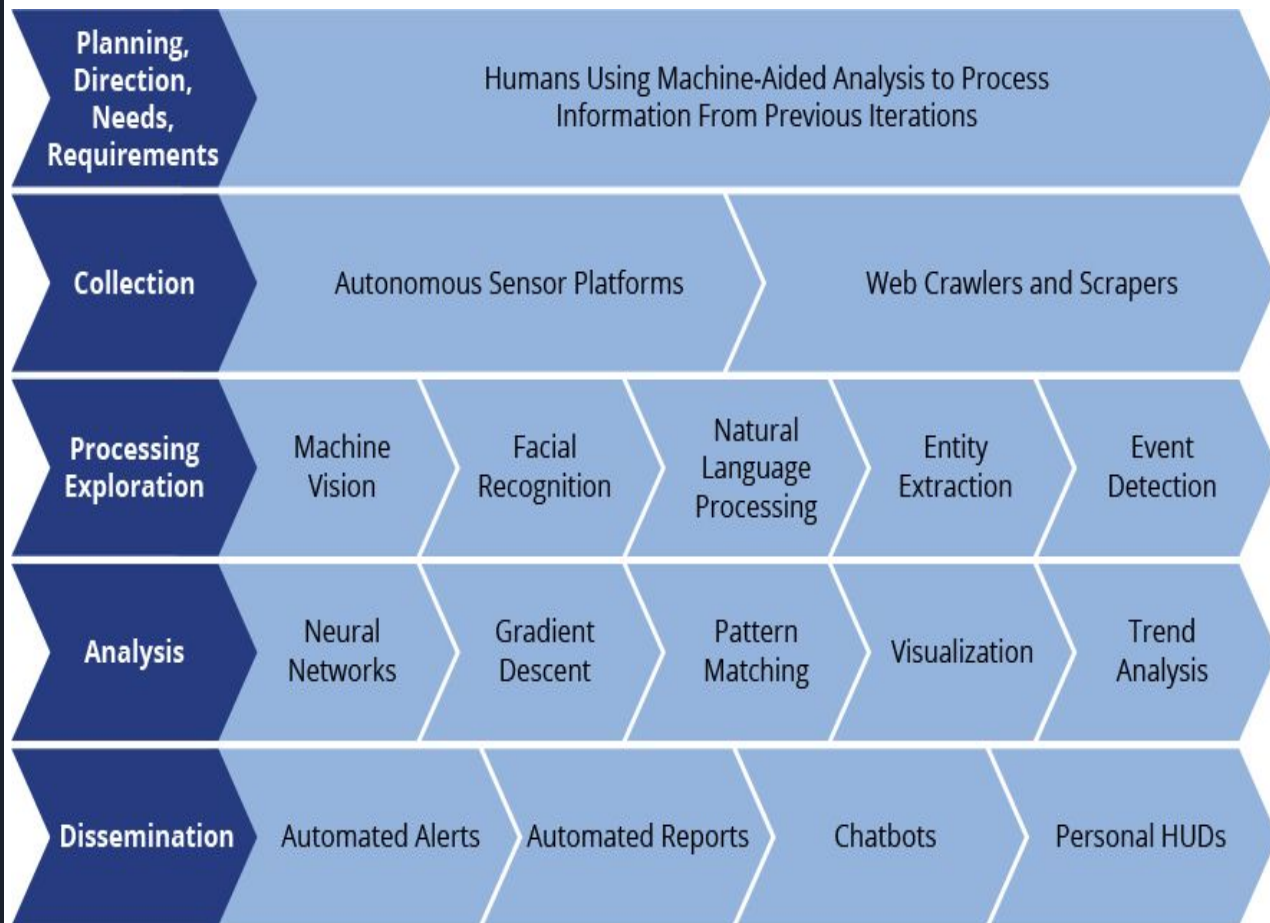


AI Use in the Intelligence Cycle

Moving from a manual process to automation partially or fully

Machine-Aided Analysis is vital especially that we work with real-world operations

Applies to data collection and analysis to AI-powered drones and training simulations



Role of ML in the Process

“machine learning will perform this function automatically and iteratively train collection and analysis algorithms by figuring out what’s working and what isn’t based on AI-fueled analysis of massive data sets”





Data Discovery and Data Mining

Start with what you know

- Define the output - what do you want to find?
- Gather the data
- Refine your conditions and clean up the data

Challenges

- Volume of data
- Reliability of data
- Human time and resources needed to go through the data



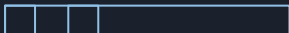


Data Discovery and Data Mining

Applying and automating data collection:

- Classification
 - Labeling
 - Metadata
- Clustering
- Deduplication
- Association rules
- Outlier detection





Smart Searching with Google Dork

Using advanced searching techniques to discover interesting information

This is also available in Bing, Duckduckgo, and other search engines

Defensive dorking - protect yourself and your organization.

Good resource for learning more:

<https://exposingtheinvisible.org/guides/google-dorking/>



"data science" site:southernct.edu filetype:pdf



All News Images Books Videos More

Settings Tools

About 27 results (0.40 seconds)

<https://inside.southernct.edu> > CSA-Interns-Flyer PDF

Computing, Math, & Data Sciences Internship - Inside Southern

interaction, virtual/augmented reality, **data science**. Analysis & Operations. Who is knocking on our network doors and why? We use a multitude of information ...

<https://inside.southernct.edu> > files > internships PDF

NATIONAL SECURITY INTERNSHIP PROGRAM - Inside ...

Applied Statistics and Computational Modeling. • **Data Science** and Analytics. Marisela Linares-Mendoza. NSIP Program Manager marisela.linares@pnnl.gov.

<https://inside.southernct.edu> > files > inside-southern PDF

Spring 2021 Tutor Schedule MONDAY - Inside Southern

Data Science. Katherine Kiernan. (9am-12:30pm). Derrick Arnold. (5pm-7pm). Katherine: Online. Derrick: Online. Katherine: DSC 100,, 101. Derrick: DSC 100.

<https://inside.southernct.edu> > files > placement PDF

Liberal Education Program Quantitative Reasoning ...

Respiratory Therapy. Majors like: Biotechnology, Chemistry, Math, Cybersecurity, Physics. Majors like: Computer Science, **Data Science**, Exercise Science,

<https://inside.southernct.edu> > files > inline-files PDF

Major Elimination Tool - Inside Southern

Computer Science. • **Data Science**. • Earth Science. • Environmental Systems and Sustainability. • Geography. • Mathematics. • Physics. UNIVERSITY-WIDE.

Data Analytics Systems - Collection

Allow the users to setup the requirements for the collection from key terms and specific data sources to exclusions and geofencing

The system then utilizes various combinations of data mining techniques to collect and sift through the data for the user

Present the data collected in a cohesive way allowing the user to interact and investigate the results





Data Analytics Systems - Enrichment

Augmenting the data with 3rd party verified data that enhance the value of existing data and help verify the accuracy and validity of the collection

Demographic data adds information such as marital status, relatives, addresses, court records, and more

Geographic data adds location information such as latitude and longitude, nearby places, city boundaries, and so one

Social Media and News data add rank, influence, followers, impact, etc





Document Analysis and Enrichment

Author information

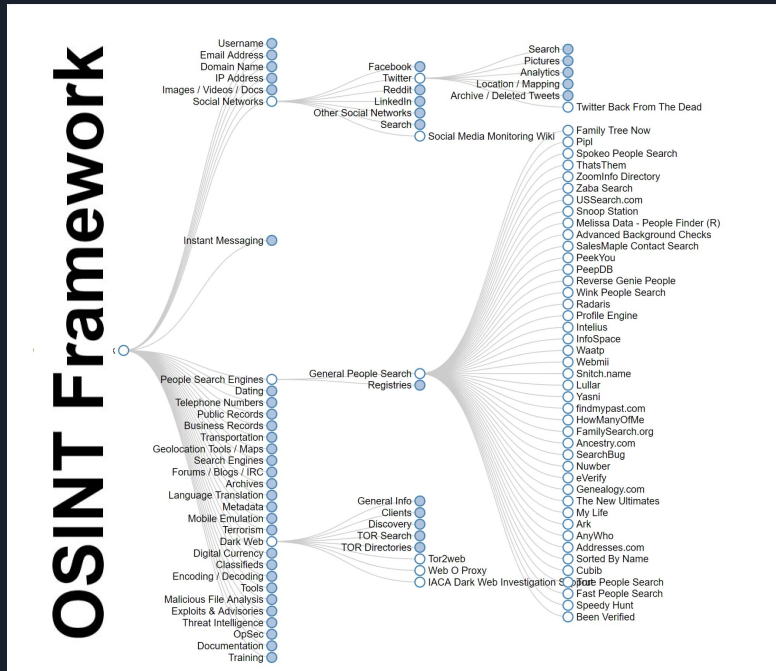
Geographical information

Image recognition and EXIF data

Sentiment Analysis

Relevance and ranking to search criteria

OSINT Tools



A growing number of tools with amazing capabilities and potential

- **Recon-ng** - a powerful python tool that automates time-consuming OSINT activities such as data gathering
- **Maltego & Maltego CE** - uncovers relationships between people, companies, domains, and publicly accessible data
- **theHarvester** - a simple tool designed to capture public data that exists outside an organization's owned network
- **Shodan** - a powerful tool that is focused on the internet of things
- **Babel X** - AI-enabled data aggregation and analysis
- **Rsoe-edis** - geospatial tool for emergency and disaster incident reporting

Shodan

Indexes the internet of things

Finds webcams, traffic lights, routers, smart devices, fridges, anything that is connected to the internet

The screenshot shows the Shodan search interface with the query 'axis'. The top navigation bar includes 'SHODAN', a search input with 'axis', and links for 'Explore', 'Enterprise Access', and 'Contact Us'. Below the navigation are tabs for 'Exploits' and 'Maps'. The main content area is divided into several sections: 'TOP COUNTRIES' with a world map and a table of results; 'TOP SERVICES' with a table of results; 'TOP ORGANIZATIONS' with a table of results; 'TOP OPERATING SYSTEMS' with a table of results; and 'TOP PRODUCTS' with a table of results. On the right side, there are three detailed search results for IP addresses: .38.246 (Japan), .192.116 (Turkey), and .230.231 (France). Each result includes the IP address, a list of discovered services (e.g., ESMTF Postfix, PIPELINING, ETRN, STARTTLS, ENHANCEDSTATUSCODES, 8BITMIME, DSN), and a 'Details' link. A blue box highlights the first result's service list. The bottom right shows a partial result for IP .131.197 (United States, Scots Valley).

SHODAN axis

Explore Enterprise Access Contact Us

Exploits Maps

TOP COUNTRIES

United States	4,689
Russian Federation	767
France	585
Canada	448
Taiwan, Province of China	314

TOP SERVICES

FTP	10,682
Telnet	233
SNMP	59
Telnet (2323)	22
SIP	17

TOP ORGANIZATIONS

Verizon Wireless	1,112
Orange	434
Telmex	231
Comcast Cable	219
Comcast Business Communications	115

TOP OPERATING SYSTEMS

Linux 2.6.x	14
Linux 3.x	3
Linux 2.4-2.6	1

TOP PRODUCTS

Axis 214 PTZ Network Camera ftpd	707
Axis M3004 Network Camera ftpd	324
Axis M1054 Network Camera ftpd	235
Axis 215 PTZ Network Camera ftpd	216
Axis M1011 Network Camera ftpd	125

Total results: 11,301

.38.246
IP
NIT
Added on 2016-07-21 14:41:38 GMT
● Japan
Details

220 axis ESMTF Postfix
250-axis
250-PIPELINING
250-SIZE 31457280
250-ETRN
250-STARTTLS
250-ENHANCEDSTATUSCODES
250-8BITMIME
250 DSN

.192.116
Turkcell
Added on 2016-07-21 14:40:51 GMT
🇹🇷 Turkey
Details

220 AXIS P1214-E Network Camera 5.40.12.2
530 Login incorrect.
214-The following commands are implemented:
USER QUIT PASS SYST HELP
NLST RETR STOR TYPE MKD
CWD SITE CDUP RNFR RNTD

.230.231
Orange
Added on 2016-07-21 14:40:00 GMT
🇫🇷 France
Details

220 AXIS 214 PTZ Network Camera 4.49 (Oct
530 Login incorrect.
214-The following commands are implemented:
USER QUIT PASS SYST HELP
NLST RETR STOR TYPE MKD
CWD SITE CDUP RNFR RNTD

.131.197
Cruzio
Added on 2016-07-21 14:38:47 GMT
🇺🇸 United States, Scots Valley
Details

220 AXIS C3003-E Network Horn Speaker 1.2
530 Login incorrect.
214-The following commands are implemented:
USER QUIT PASS SYST HELP
NLST RETR STOR TYPE MKD
CWD SITE CDUP RNFR RNTD

Investigate View Entities Collections Transforms Machines Collaboration Import | Export Windows

Clear Graph Number of Results 12 50 250 10k

Paste Cut Delete

Quick Find in Files

Entity Selection

Select All Add Parents Add Neighbors Select Children Select by Type Zoom to Fit Zoom In

Select None Add Children Add Path Select Neighbors Zoom 100% Zoom Out

Invert Selection Add Similar Siblings Select Parents Select Bookmarked Reverse Links Zoom to Zoom Selection

Entity Palette

DNS Name
Domain Name System server name

Domain
An internet domain

IPv4 Address
An IP version 4 address

MX Record
A DNS mail exchange record

NS Record
A DNS name server record

Netblock
A range of IP version 4 addresses

URL
An Internet Uniform Resource Locator (URL)

Tracking Code
Represents a tracking code

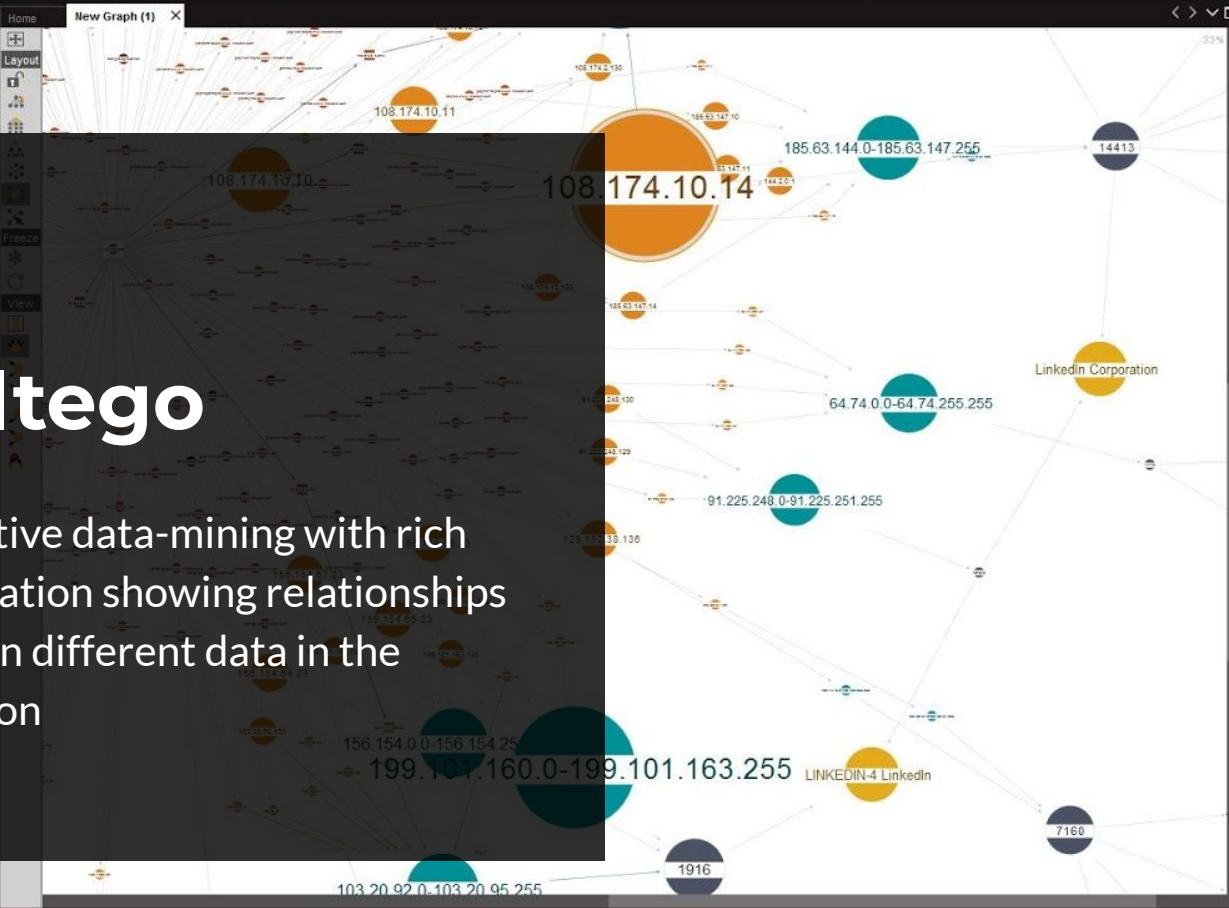
Website
An internet website

View

Transforms

Machines

Prune Leaf Entities
Machine to prune leaf entities



Overview

Detail View

IPv4 Address
maltego.IPv4Address
108.174.10.14

+ Relationships

+ Generator detail

Property View

Properties

Type	IPv4 Address
IP Address	108.174.10.14
Internal	<input type="checkbox"/>
Graph info	
Weight	100
Incoming	352
Outgoing	1
Bookmark	<input type="checkbox"/>

Maltego

Interactive data-mining with rich visualization showing relationships between different data in the collection

BabelStreet.com

“The world’s leading
AI-enabled data-to-knowledge
company”

B2B SaaS Company

MOZAMBIQUE: UPDATE ON INSURGENT OPERATIONS

Introduction:

Based on analysis of open source reporting, the situation in Northern Mozambique continues on a negative trend since Babel Street’s initial reporting in late June 2020, devolving into a political, commercial, and possible humanitarian crisis. Since June, militants have solidified their hold in the Cabo Delgado region of Mozambique, with insurgents capturing and holding the city of Mocímboa da Praia in early August and now perpetrating attacks on the Afungi peninsula, only miles from energy giant Total’s operations. This has prompted the evacuation of Total’s personnel in late December/early January. Insurgent activity has continued throughout the region, with Al Sunnah wa Jama’ah (ASWJ) increasingly targeting lines of communication and overrunning government outposts. The expansion of insurgent-controlled territory has generated concern about the region being cut off from the rest of the country. New reporting on the insurgency suggests closer connections with the Islamic State (IS), stoking fears that the indigenous insurgents may be receiving support from beyond the borders of Mozambique. Taken collectively, the events of the past six months paint a bleak picture for the government of Mozambique and liquid natural gas (LNG) exploration and production in the region.



Figure 1 - Map of Mozambique and LNG Deposits (source: Total)

BLUF:

- Insurgent attacks on the Afungi peninsula halt LNG operations and prompt Total’s evacuation of personnel
- Insurgents have solidified their hold of Mocímboa da Praia and continue attacks on transportation networks, essentially cutting off Palma from ground transportation

Sharing Insights

- Timely, actionable awareness from events as they occur around the world
- Tailored, persistent, and relevant information unconstrained by time, location, or language
- Geographically relevant voices from hyper-local sources
- Focused on emergent and timely concerns with topics ranging from:
 - Breaking News
 - Disease Outbreaks
 - Mass Casualty Attacks
 - Natural Disasters
 - Terror Events
 - Transportation Incidents



Summary:
The Coronavirus pandemic impacts across all aspects of global society, and 97,000 have perished, and

With coverage in more than 350 countries, understand developments related to COVID-19 and a dynamic array of global information on the

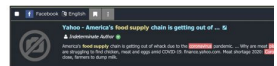
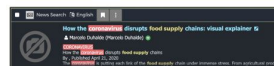
As misinformation and Channels cover these critical threads and related news considered a public health

Summary:
The Coronavirus pandemic (COVID-19) impacts across all aspects of global society, and 171,000 have perished, and the virus has

With coverage in more than 350 countries, understand developments in more than 20 related to COVID-19 and act to mitigate information on a dynamic array of global information on the

The pandemic has heavily impacted the global economy, and other key commodities, challenges and limitations, restaurant closures, and noteworthy material recently captured in Babel Channels

Overall Food Supply Issue:
Babel Channels offers insight into topics such as



Impacts on Food Processing:
Babel Channels collects news and updates on processing plants for meat, poultry, and seafood.



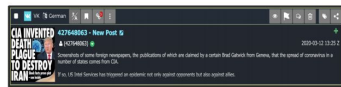
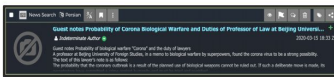
Summary:
The Coronavirus pandemic (COVID-19) that originated in Wuhan, China, in December 2019 is having unmistakable global impacts across all aspects of society. As of March 18, 2020, nearly 226,000 people have been infected, more than 9,200 had perished, and the virus had been detected in more than 176 countries.

With coverage in more than 350 countries, states, territories and cities around the world, and the ability to help users understand developments in more than 200 languages, Babel Channels is your tool to understand what is happening related to COVID-19 and act to mitigate impacts related to the spread of the virus.

Beyond the human and financial impacts to COVID-19, the spread of misinformation related to the pandemic has the potential to sap limited resources, sow confusion or potentially ensue violence, and undermine trust that is necessary in this global response. Some noteworthy examples of recently captured misinformation campaigns in Babel Channels follow:

International State Government, Academia, and Media Accusations of Bioweapons:

Babel Channels collected references leveled by state media, academic, and national governments related to COVID-19 as a deliberate bioweapon by the United States or China.



Accusations of Biological Warfare Spreading Across Social Platforms:

Babel Channels collected examples of how many of the same misinformation messages are propagating in Blogs, Message Boards and the Russian social media platform, VK.



Follow @babelstreet on
LinkedIn or twitter

A world map showing various disaster events. Icons include fire (wildfires), lightning (storms), and people (earthquakes). Numbers are placed near icons to indicate counts. For example, in the US, there are 15 wildfire icons, 4 earthquake icons, and 3 storm icons. In Mexico, there are 10 wildfire icons, 2 earthquake icons, and 3 storm icons. In South America, there are 4 earthquake icons in Peru and 2 in Bolivia. In Africa, there are 2 earthquake icons in Mali and 2 in Nigeria. In Europe, there are 2 earthquake icons in the UK and 1 in France. In Asia, there are 2 earthquake icons in Japan and 2 in Indonesia. In Australia, there are 2 earthquake icons. The map also shows major cities and geographical features like the Pacific Ocean, Atlantic Ocean, and various seas.

rsoe-edis.org

Monitors, aggregates, analyzes and notifies.

Focused on emergency and disaster information reporting



OSINT Tools Advantage

Data Aggregation - from unstructured data into a structured queryable, filterable, sortable, and digestible data

Analysis and Enrichment - augmenting the data with additional metadata and 3rd party information to help analyze, validate, label, group, and dedup

Visualization - creating mind maps, visualizing relationships, geographical view, data time lapse

Automated Alerting and Reporting - continuous monitoring and automation allows alerting and reporting





Use Cases

Event Monitoring

Finding People

Workflows

ROAD
CLOSED

Email: estimating@cemex.com

PAVING SOLUTIONS LTD

BOULIQUE
CAFE

42A

Permit No
JB0021620



Event Monitoring: Super Bowl XLIX

Collected more than **one billion** posts related to Super Bowl XLIX and the Phoenix area, including 40,000 geo-located postings.

124,000-plus were filtered and analyzed

48 of those being passed along to the necessary individuals for further analysis

Ultimately, among the multiple uses of Babel X during the game, **three particular** postings were identified that indicated possible threats to the venue. Each was vetted successfully and cleared.

Babel Street was awarded the Golden Eagle Award from the National Center for Spectator Sports Safety & Security (NCS4)

Source: SDM Magazine





TraceLabs.org

Crowdsources OSINT operations to find missing people

Data is passed to respective law enforcement to pursue appropriate action

Over 300 assisted cases around the world

Search Party

Crowdsourced OSINT to Find Missing Persons

Trace Labs is designed to be a catalyst for improving the state of missing persons location and family reunification. We provide a modern, cost effective and transparent solution to a problem that is destroying families.

[Learn More](#)

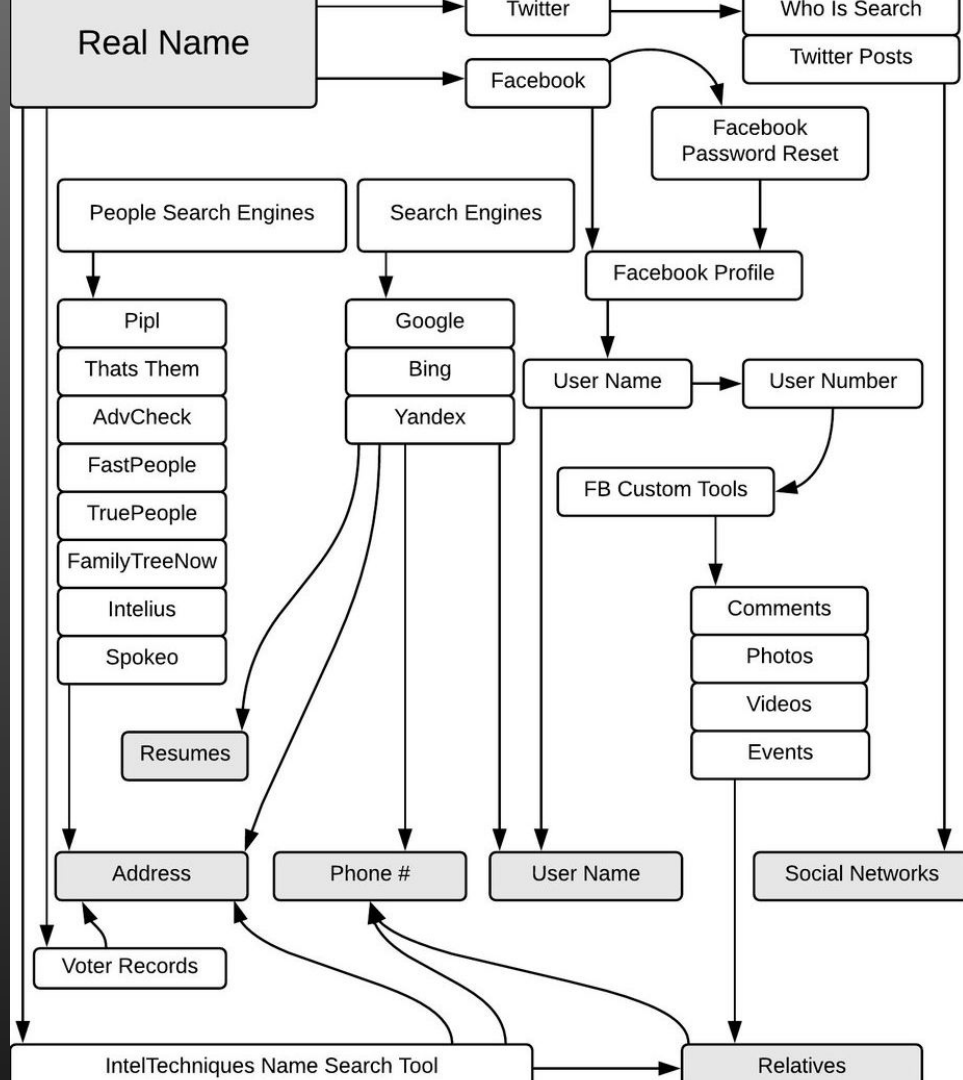


Starting with a name

- spokeo.com
- thatsthem.com
- beenverified.com
- fastpeoplesearch.com
- truepeoplesearch.com
- familytreenow.com
- intelius.com
- people.yandex.ru



Image Source: IntelTechniques.com OSINT Workflow Chart: Real Name



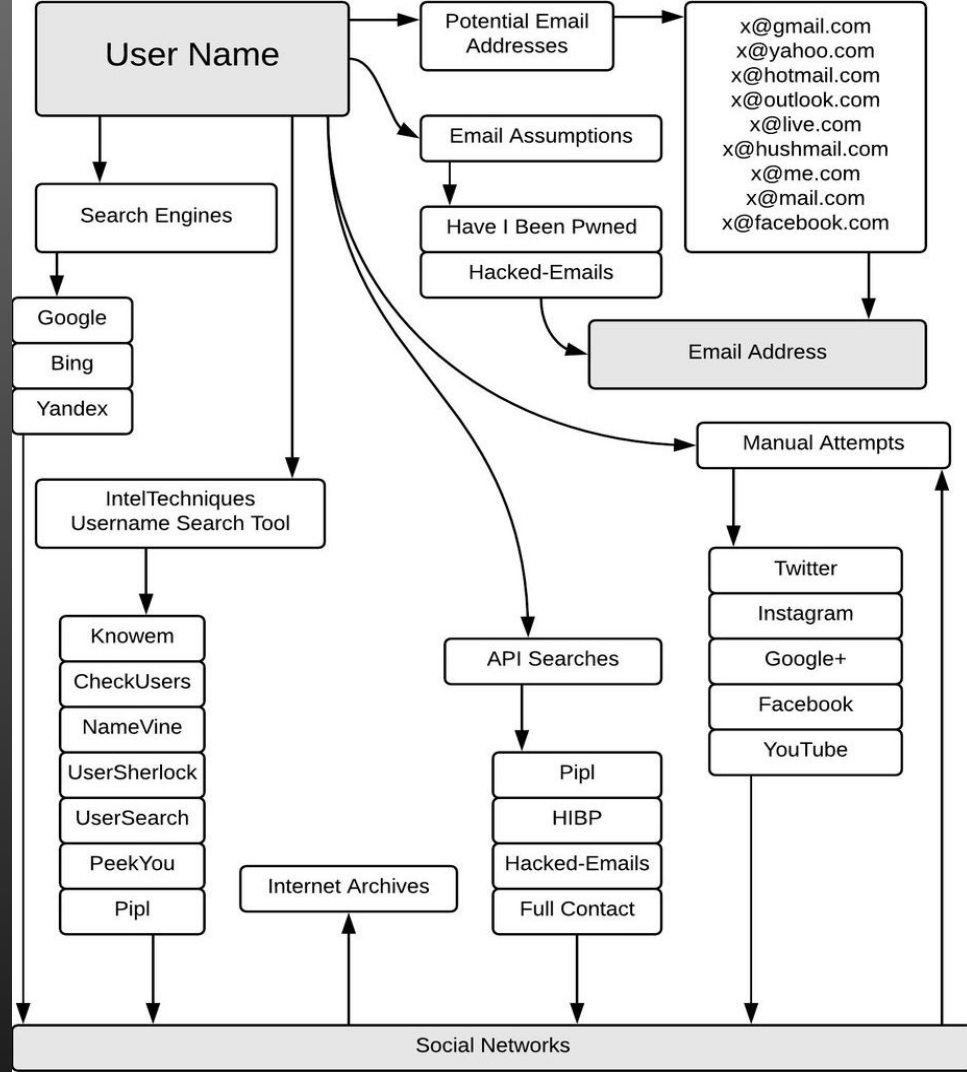
Starting with a username

Reverse username search

- [Socialcatfish.com](https://socialcatfish.com)
- [Usersearch.org](https://usersearch.org)
- [Peekyou.com](https://peekyou.com)

Username search

- [Instantusername.com](https://instantusername.com)
- [Namechk.com](https://namechk.com)
- [Whatsmyname.com](https://whatsmyname.com)



Before you get start with OSINT

Common strategies OSINT investigators use to protect themselves:

- Understand the issues that could decrease their anonymity
- Use VMWare
- Use emulators
- Create burner profiles (emails or social media)
- Use VPNs
- Leverage different web browsers and add-ons



OSINT Resources

[OSINT Framework](#) -

interactive chart - helps you know where to start, what to do, and what tools to use.

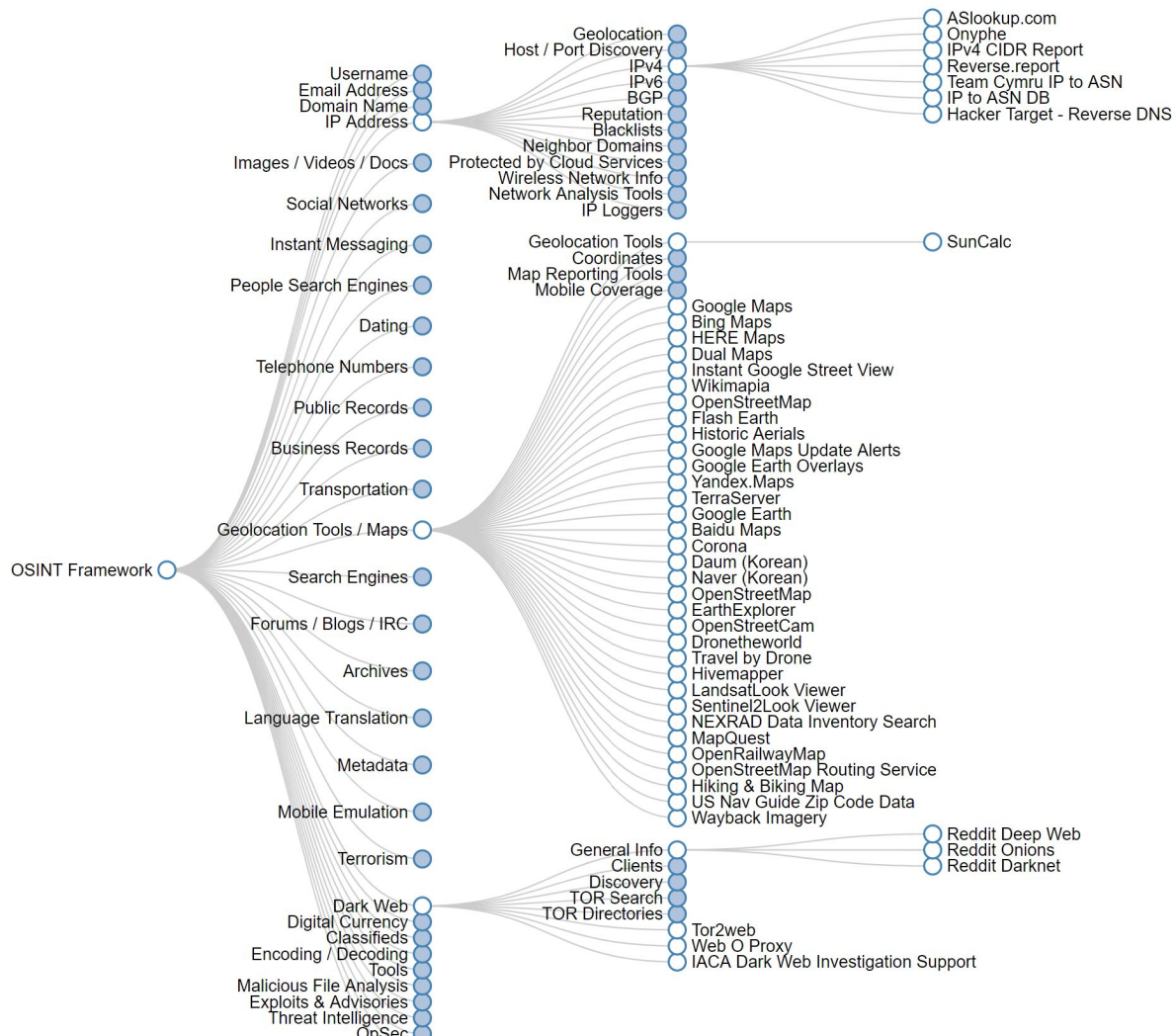
[WebBreacher.com](#) - Micah

[Hoffman's Blog](#)

[Intertechniques.com](#) - Michael

Bazzell

[Osintcurio.us](#)



Disclaimer

This information is for educational use ONLY and not to be used for personal or unlawful purposes.

Please be respectful of others privacy and do not do anything “creepy” without consent.



**If you torture the data
long enough, it will
confess to anything.**
~R. Coase

The Intelligence Process

OSINT Concepts and Open Source Data

OSINT Tools

AI/ML and Data Science applications in OSINT

OSINT Tools & Use Cases





Takeaways

OSINT is based exclusively on publicly available data such as the contents of the open web

AI is key to expanding and improving OSINT. In particular, it enables human analysts to collect, enrich, analyze, and disseminate information in a timely and decisive manner

Human interaction and input into the process is vital and cannot be completely replaced by AI



Thank You!

Have follow up questions or comments, feel free to reach out on [linkedin.com/in/nisreencaïn/](https://www.linkedin.com/in/nisreencaïn/)

