Automating Open Source Intelligence

Algorithms for OSINT

Edited By

Robert Layton

Paul A. Watters

AMSTERDAM • BOSTON • HEIDELBERG • LONDON
NEW YORK • OXFORD • PARIS • SAN DIEGO
SAN FRANCISCO • SINGAPORE • SYDNEY • TOKYO
Syngress is an imprint of Elsevier

SYNGRESS

Acquiring Editor: Brian Romer
Editorial Project Manager: Anna Valutkevich
Project Manager: Mohana Natarajan
Cover Designer: Matthew Limbert

Syngress is an imprint of Elsevier
225 Wyman Street, Waltham, MA 02451, USA

Notices
Knowledge and best practice in this field are constantly changing. As new research and experience broaden our understanding, changes in research methods, professional practices, or medical treatment may become necessary.

Practitioners and researchers must always rely on their own experience and knowledge in evaluating and using any information, methods, compounds, or experiments described herein. In using such information or methods they should be mindful of their own safety and the safety of others, including parties for whom they have a professional responsibility.

To the fullest extent of the law, neither the Publisher nor the authors, contributors, or editors, assume any liability for any injury and/or damage to persons or property as a matter of products liability, negligence or otherwise, or from any use or operation of any methods, products, instructions, or ideas contained in the material herein.

British Library Cataloguing-in-Publication Data
A catalogue record for this book is available from the British Library

Library of Congress Cataloging-in-Publication Data
A catalog record for this book is available from the Library of Congress

ISBN: 978-0-12-802916-9

List of Contributors

Brenda Chawner School of Information Management, Victoria Business School, Victoria University of Wellington, New Zealand
Shadi Esnaashari School of Engineering and Advanced Technology, Massey University, Auckland, New Zealand
Ernest Foo School of Electrical Engineering and Computer Science – Science and Engineering Faculty, Queensland University of Technology, Queensland, Australia
Rony Germon PSB Paris School of Business, Chair Digital Data Design
Iqbal Gondal Internet Commerce Security Laboratory, Federation University, Australia
Hans Guesgen School of Engineering and Advanced Technology, Massey University, New Zealand (Palmerston North campus)
Christian Kopp Internet Commerce Security Laboratory, Federation University, Australia
Robert Layton Internet Commerce Security Laboratory, Federation University, Australia
Seung Jun Lee School of Engineering & Advanced Technology, Massey University, New Zealand
Charles Perez PSB Paris School of Business, Chair Digital Data Design
Agate M. Ponder-Sutton Information Technology & Centre for Information Technology, School of Engineering and Advanced Technology, Massey University, New Zealand
Jim Sillitoe Internet Commerce Security Laboratory, Federation University, Australia
Jason Smith School of Electrical Engineering and Computer Science – Science and Engineering Faculty, Queensland University of Technology, Queensland, Australia
Kristin Stock School of Engineering and Advanced Technology, Massey University, New Zealand (Albany, Auckland campus)

Suriadi Suriadi School of Engineering and Advanced Technology, College of Sciences, Massey University, New Zealand
Paul A. Watters School of Engineering & Advanced Technology, Massey University, New Zealand
George R.S. Weir Department of Computer and Information Sciences, University of Strathclyde, Glasgow, UK
Ian Welch School of Engineering and Computer Science, Victoria University of Wellington, New Zealand

CHAPTER 1

The Automating of Open Source Intelligence

Agate M. Ponder-Sutton

Information Technology & Centre for Information Technology, School of Engineering and Advanced Technology, Massey University, New Zealand

Open source intelligence (OSINT) is intelligence that is synthesized using publicly available data (Hobbs, Moran, & Salisbury, 2014). It differs significantly from the open source software movement. This kind of surveillance started with the newspaper clipping of the first and second world wars. Now it is ubiquitous within large business and governments and has dedicated study. There have been impassioned, but simplified, arguments for and against the current levels of open source intelligence gathering. In the post-Snowden leaks world one of the questions is how to walk the line between personal privacy and nation state safety. What are the advances? How do we keep up, keep relevant, and keep it fair or at least ethical? Most importantly, how do we continue to "make sense or add value" as Robert David Steele would say, (http://tinyurl.com/EIN-UN-SDG). I will discuss the current state of OSINT and data science. The changes in the analysts and users will be explored. I will cover data analysis, automated data gathering, APIs, and tools; algorithms including supervised and unsupervised learning, geo-locational methods, de-anonymization. How do these interactions take place within OSINT when including ethics and context? How does OSINT answer the challenge laid down by Schneier in his recent article elaborating all the ways in which big data have eaten away at the privacy and stability of private life, "Your cell phone provider tracks your location and knows who is with you. Your online and in-store purchasing patterns are recorded, and reveal if you are unemployed, sick, or pregnant. Your emails and texts expose your intimate and casual friends. Google knows what you are thinking because it saves your private searches. Facebook can determine your sexual orientation without you ever mentioning it." (Schneier, 2015b). These effects can be seen in worries surrounding the recording and tracking done by large companies to follow their customers discussed by Schneier, (2015a, 2015b) and others as the crossing of the uncanny valley from useful into disturbing. These examples include the recordings made by a Samsung TV of consumers in their homes (http://www.theguardian.com/media-network/2015/feb/13/samsungs-listening-tv-tech-rights); Privacy fears were increased by the

cloud storage of the recordings made by the interactive WIFI-capable Barbie (http://www.theguardian.com/technology/2015/mar/13/smart-barbie-that-can-listen-to-your-kids-privacy-fears-mattel); Jay-Z's Album Magna Carta Holy Grail's privacy breaking app (http://www.theguardian.com/music/2013/jul/17/jay-z-magna-carta-app-under-investigation); and the Angry Birds location recording which got targeted by the NSA and GCHQ and likely shared with other Five Eyes Countries (http://www.theguardian.com/world/2014/jan/27/nsa-gchq-smartphone-app-angry-birds-personal-data). The Internet can be viewed as a tracking, listening, money maker for the recorders and new owners of your data. Last but not least there must be a mention of the Target case where predictions of pregnancy were based on buying history.

The Target storey was broken by the *New York Times* (Duhigg, C. "How Companies Learn Your Secrets." February 16, 2012. http://www.nytimes.com/2012/02/19/magazine/shopping-habits.html?_r=0).

The rise of OSINT, data science, business, or commercial has come with the revolution in the variety, volume, and availability public data (Hobbs et al., 2014; Appel, 2014). There has been a profound change in how data are collected, stored, and disseminated driven by the Internet and the advances linked to it. With establishment of Open Source Center and assistant deputy director for open source intelligence in the United States, the shift toward legitimacy of OSINT in the all-source intelligence process was made clear (http://resources.infosecinstitute.com/osint-open-source-intelligence/). The increased importance of OSINT has moved it into the core of intelligence work and allowed a larger number of players to take part, diversifying its uses beyond the original "intelligence community" (Hobbs et al., 2014). Interconnectivity has increased and much of that data can be utilized through open source intelligence methodologies to create actionable insights. OSINT can produce new and useful data and insights; however, it brings technical, political, and ethical challenges and obstacles that must be approached carefully.

Wading through the sheer bulk of the data for the unbiased reality can present difficulties. Automation means the spread of OSINT, out of the government office to businesses, and casual users for helpful or wrong conclusions as in the case of the Boston bomber Redit media gaff (http://www.bbc.com/news/technology-22263020). These problems can also be seen in the human flesh search engine instances in China and the doxing by anonymous and others in positive and negative lights. With more levels of abstraction increasing difficulty is apparent, as tools to look at the tools to look at the output of the data. Due to the sheer volume of data it becomes easier to be more susceptible to cognitive bias. These are issues can be seen in the errors made by the US government in securing their computer networks ("EPIC" fail – how OPM hackers

tapped the mother lode of espionage data. Two separate "penetrations" exposed 14 million people's personal information. Ars Technica. June 22, 2015. 2:30pm NZST. http://arstechnica.com/security/2015/06/epic-fail-how-opm-hackers-tapped-the-mother-lode-of-espionage-data/). With the advent of corporate doxying of Ashley Madison and of Sony it can be seen as a private corporation problem as well.

Groups of users and uses include: governments; business intelligence and commercial intelligence; academia; and Hacker Space and Open Data initiatives. Newer users include nongovernmental organizations (NGOs), university, public, and commercial interests. User-generated content, especially social media, has changed the information landscape significantly. These can all have interactions and integrated interests. Collaboration between these groups is common among some, US government contracting IBM and Booz-Allen and also less inflammatory contracted employees; academia writing tools for Business Intelligence or government contracts. These tend to be mutually beneficial. Others where the collaboration is nonvoluntary such as the articles detailing how to break the anonymity of the netflix prize dataset (Narayanan & Shmatikov, 2008); or any of the multiple blog posts detailing similar anonymity breaking methods such as "FOILing NYC's Taxi Trip Data" http://chriswhong.com/open-data/foil_nyc_taxi/ and London bicycle data "I know where you were last summer" http://vartree.blogspot.co.nz/2014_04_01_archive.html) have furthered security and OSINT analysis, sometimes to the ire of the data collectors.

Open source methodology

The extent to which information can be collected is large and the field is broad. The speed, the volume, and variety are enough that OSINT can be considered a "Big Data" problem. Tools to deal with the tools that interface with the data such as Maltego and Recon-ng are becoming more popular and common approaches. These approaches still require setup and a certain amount of

knowledge to gain and/or buy access to information. This required setup also includes a certain amount of tuning that cannot be or would be difficult to automate. Fetching the data and to some extent limitation of false positives can be automated. OSINT research continues to push automation further. There is an overall Chelsea Manning, and lean toward the commodification of OSINT; more companies offer more analytical tools and/or software and a service to cash in on what was once a government or very limited field. Many tools are available that require less technical expertise; featuring drag and drop interfaces where the focus is on ease of use and the availability of the data.

Open source intelligence methodology is a synthesis from multiple fields: data science, statistics, machine learning, programming, databases, computer science, and many other fields, but there is no over-arching unifying theory of open source intelligence. The ease of the data creation and acquisition is unprecedented, and OSINT owes this to its rise as well to the complex algorithm, de-anonymization, and fear that has come with them. WikiLeaks, and Snowden, (http://www.theguardian.com/us-news/the-nsa-files), have provided a highly publicised view of the data compiled on the average person with regards to the Five Eyes; we can only assume that similar things are done by other governments (Walsh & Miller, 2015). Commercial organizations have followed suit with worrisome and very public issues surrounding the collection of data. This is a wealth of data as well as a major ethical concern. This is part of the OSINT landscape because (1) people behave differently when they know they are under surveillance (Miller et al., 2005); (2) if this is part of the intelligence landscape this culture of "get it all" others will follow in its path; and (3) intelligence has become big business (Miller et al., 2005). Schneier tells us in 2015 that "Corporations use surveillance to manipulate not only the news articles and advertisements we each see, but also the prices we're offered. Governments use surveillance to discriminate, censor, chill free speech, and put people in danger worldwide. And both sides share this information with each other or, even worse, lose it to cybercriminals in huge data breaches."

And from this view we have an increasing interest in anonymization and de-anonymization because the data that are available either freely publically or for a fee can identify impact on the interested user and the originator of the data. The importance of anonymization of data within the realm of Internet security and its risks are clearly recognized by the U.S. President's Council of Advisors on Science and Technology ("PCAST"):

> Anonymization of a data record might seem easy to implement. Unfortunately, it is increasingly easy to defeat anonymization by the very techniques that are being developed for many legitimate applications of big data. In general, as the size and diversity of available data grows, the likelihood of being able to re-identify individuals (that is, re-associate their records with their names) grows substantially. [...]

> Anonymization remains somewhat useful as an added safeguard, but it is not robust against near-term future re-identification methods. PCAST does not see it as being a useful basis for policy (PCAST, 2014).

This 2014 PCAST - Executive Office of the President, 2014, report captures the consensus of computer scientists who have expertise in de- and reidentification: there is no technical backing to say that common deidentification methods will be effective protection against future attempts.

The majority of people have some kind of online presence. There has been an increase not only since its initialization, but in uptake in the last couple of years. Ugander, Karrer, Backstrom, and Marlow (2011) wrote: The median Facebook user has about a hundred friends. Barlett and Miller (2013) said, "Every month, 1.2 billion people now use internet sites, apps, blogs and forums to post, share and view content." (p. 7). In 2015, Schneier tells us, "Google controls two-thirds of the US search market. Almost three-quarters of all internet users have Facebook accounts. Amazon controls about 30% of the US book market, and 70% of the ebook market. Comcast owns about 25% of the US broadband market. These companies have enormous power and control over us simply because of their economic position." (Schneier, 2015a, 2015b). So you can see how the situation could be both exciting and dire as a company, an organization, and an individual. There are a plethora of books on OSINT and its methods, tutorials, and how-to's having been touched by the dust of the "secret world of spies" it is now gathering hype and worry. And because both are warranted treading in this area should be done carefully with an eye toward what you can know and always in mind what privacy should be (Ohm, 2010).

"Loosely grouped as a new, 'social' media, these platforms provide the means for the way in which the internet is increasingly being used: to participate, to create, and to share information about ourselves and our friends, our likes and dislikes, movements, thoughts and transactions. Although social media can be 'closed' (meaning not publically viewable) the underlying infrastructure, philosophy and logic of social media is that it is to varying extents 'open': viewable by certain publics as defined by the user, the user's network of relationships, or anyone. The most well-known are Facebook (the largest, with over a billion users), YouTube and Twitter. However, a much more diverse (linguistically, culturally, and functionally) family of platforms span social bookmarking, micromedia, niche networks, video aggregation and social curation. The specialist business network LinkedIn has 200 million users, the Russian-language VK network 190 million, and the Chinese QQ network 700 million. Platforms such as Reddit (which reported 400 million unique visitors in 2012) and Tumblr, which has just reached 100 million blogs, can support extremely niche communities based on mutual interest.

For example, it is estimated that there are hundreds of English language pro-eating disorder blogs and platforms. Social media accounts for an increasing proportion of time spent online. On an average day, Facebook users spend 9.7 billion minutes on the site, share 4 billion pieces of content a day and upload 250 million photos. Facebook is further integrated with 7 million websites and apps" (Bartlett and Miller, 2013, p. 7).

Schneier tells us that, "Much of this [data gathering] is voluntary: we cooperate with corporate surveillance because it promises us convenience, and we submit to government surveillance because it promises us protection. The result is a mass surveillance society of our own making. But have we given up more than we've gained?" (Schneier, 2015a, 2015b). However, those trying to avoid tracking have found it difficult to inforce. Ethical nontracking (DoNotTrack http://en.wikipedia.org/wiki/ Do_Not_Track) and opt out lists and the incognito settings on various browsers have received some attention and, but several researchers have shown these have little to no effect on the tracking agencies (Schneier; Acar et al., 2014). Ethical marketing and the developers kit for that at DoNotTrack. Persistent tracking within the web is a known factor (Acar et al., 2014) and the first automated study of evercookies suggests that opts outs made little difference. Acar et al. track the cookies tracking a user in three different ways coming to the conclusion that "even sophisticated users face great difficulty in evading tracking techniques." They look at canvas finger printing, evercookies, and use of "cookie syncing. They perform the largest to date automated crawl of the home pages of Top Alexa 100K sites and increased the scale of their work on respawning, evercookies, and cookie syncing. The first study of real-world canvas finger printing. They include in their measurements the flash cookies with the most respawns, the top parties involved in cookies sync, the top IDs in cookies sync from the same home pages and observed the effect of opting out under multiple schemes. A draft preprint by (Englehardt et al., 2014) discusses web measurement as a field and identifies 32 web privacy measurement studies that tend toward *ad hoc* solutions. They then present their own privacy measurement platform, which is scalable and outlines how it avoids the common pitfalls. They also address the case made by most press of the personalization effects of cookies and tracking by crawling 300,000 pages across nine news sites. They measure the extent of personalization based on a user's history and conclude the service is oversold. So based on these the plethora of data could still be useful, gathered less intensely, or in other more privacy-preserving manners.

"We kill people based on metadata" is one of the most quoted or focused-on things that General Michael Hayden, Former NSA head, has said, but other things he said in the same interview were equally important (https://www.youtube.com/watch?v=UdQiz0Vavmc). When General Hayden says the NSA are "…yelling through the transom…"; he means that starting with one phone

number the NSA can then expand this by pulling in every number that has called that number and every number that has called those numbers using the interconnections of networks – (see Maltego for similar effects)). Targeted attacks such as these which can expand the available data are covered in depth by Narayanan, Huey, and Felten (2015). The heavy use of statistics and rise of data science allow users to deal less with the data and more with the metadata which can be seen as a lengthening of the weight of the data. Part of this lightening the load is the rise of tools for the less technical.

The advances in open source intelligence automation have been unsurprisingly linked to advances in computing and algorithms; they are focused on the collection of data and the algorithms used to do analysis (Hobbs et al., 2014). There has been a shift toward the public sector not only of the provision of OSINT as a service from private firms but of the use of by marketing and commercial sides of businesses of open source intelligence. The data gathering, insight synthesis, and build of proprietary tools for OSINT are on the rise. Covered here are what algorithms are new, innovative, or still doing well. New sources and ways to find them are covered lightly. Here are presented several common and new algorithms along with breakthroughs in the field. The *ad hoc* quality of the open source intelligence gathering leads to the rise of new original algorithms (Narayanan, 2013 and Acar et al., 2014) and new uses.

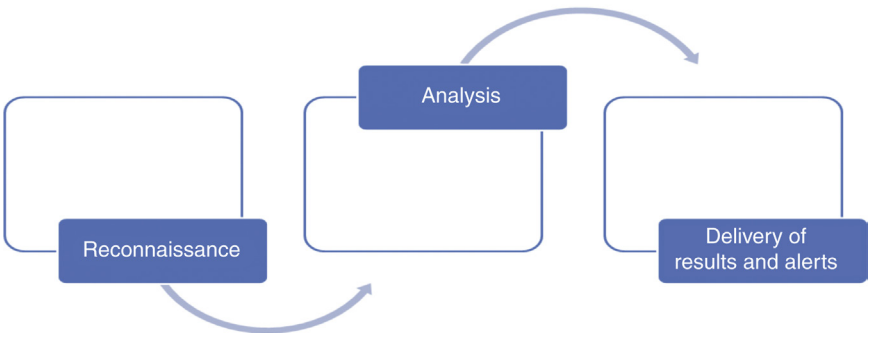
THE COMMERCIAL ANGLE

Data science and really the new tend toward tools and hype, "What is hot in analytics" may threaten to distract from the substance of the revolution (Walsh & Miller, 2015). In an October 2012 edition of the Harvard Business Review, the role of a data scientist was called the "sexiest job of the 21st Century." The article discusses the rise of the data expert, with more and more companies turning to people with the ability to manipulate large data sets (http://datasmart.ash.harvard.edu/news/article/the-rise-of-the-data-scientists-611). In 2011, a report by McKinsey predicted that "by 2018 the US would face a shortage of 140,000 to 190,000 workers with deep analytical skills and of 1.5 million managers and analysts with big data skills" (http://www.mckinsey.com/insights/business_technology/big_data_the_next_frontier_for_innovation). "Big Data" has seen a lot of hype and as we sit in what Gartner terms the trough of disillusionment with regard to Big Data; companies are finding additional ways to use data and combine technologies with the concept of recombination to create solutions in the growing trend in the business intelligence space. Business intelligence or business analytics has migrated from IT departments into either its own department or individual departments and often into the marketing department (https://www.gartner.com/doc/2814517/hype-cycle-big-data-). The ability of

early adopters in sectors such as risk management, insurance, marketing, and financial services brings together external data and internal data to build new algorithms – to identify risk, reduce loss, and strengthen decision support. Companies want to be seen to be working with world-leading business intelligence companies that can present and synthesize hybrid data.

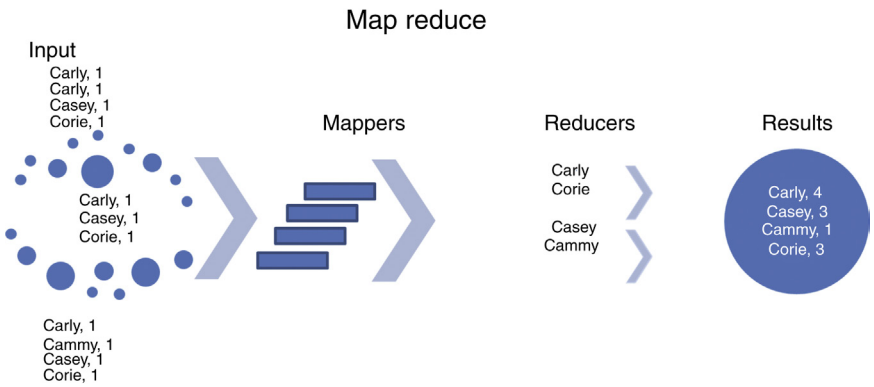
When the private company Ventana ranked OSINT/BI products in 2015; those that were ranked highly mixed functionality and user experience. Many of the top BI Tools provide user experience and an integrated data management, predictive analytics, visual discovery, and operational intelligence capabilities in a single platform. Modern architecture that is cloud-ready and supports responsive design on mobile devices is a bonus. Combining predictive analytics with visual discovery is popular. Ventana Noted that users preferred all-in-one user technology that addresses the need for identity management and security, especially in a multiple device leaning time (http://www.information-management.com/blogs/Business-Analytics-Intelligence-Hot-Ventana-Research-10026829-1.html?utm_campaign=blogsapr%2022%202015&utm_medium=email&utm_source=newsletter&ET=informationmgmt%3Ae4238189%3A4131831a%3A&st=email).

Emphasis is placed on all steps, in order to allow actionable insights to be trusted.



MapReduce and similar algorithms are getting the most use in cloud computing to deal with the quantity and variety of data within OSINT. Originally, a Google proprietary function that has since been genericized, MapReduce is a framework used to process and generate large data sets within a parallel, distributed algorithm on a large number of computers (Dean & Ghemawat, 2010). These computers are nodes in either if they are collocated with the similar hardware a cluster or not a grid. Processing can use unstructured or structured stored data. MapReduce can also process data on or near storage assets in order

to reduce loss from transmission distance. Conceptually similar approaches have been in use since 1995 with the Message Passing Interface.



Automation implies tools; the first tool we are going to cover is web crawlers. Because there are many sources the information-gathering process will take time if solely manual techniques are used. A web crawler starts with a list of URLs, the seed list. As the crawler visits these, it identifies all the hyperlinks in the page and adds the new links to the list of URLs. URLs from the list are recursively visited according to a set of policies. If the crawler is performing archiving of websites, it copies and saves the information as it goes. Then, archives are stored in order that they can be viewed, read, and navigated as they were on the live web (Masanès, 2007). One example of this is the Wayback Machine – Internet Archive (http://archive.org/web/).

A crawler can only download a limited number of web pages within a given time, so it prioritizes those downloads. However, because of the numerous possible combinations of HTTP GET (URL-based) that parameters exist, only a small selection of these will return unique content. This is a problem for crawlers, to sort through endless combinations of minor scripted changes to retrieve unique content. For example, an online art gallery offers three options to users; these are specified through HTTP GET parameters in the URL. If users can sort images in four ways, three possibilities for image size, two for file formats, and the option to disable user-provided content, then the same set of content can be accessed with 48 different URLs. Web crawlers can incorporate APIs.

The second set of tools is APIs. Many sites have APIs that return results in a JSON format. APIs require an access_token that free or cost (http://raidersec.blogspot.co.nz/2012/12/automated-open-source-intelligence.html). The JSON

output from an API can be imported and handled using Python. This and the ability to make batch requests of API can searching and gathering data easier.

Facebook's Graph API was created to streamline access to information. The "search" feature allows searching of public profiles, posts, events, groups, and more based on a given keyword; an example URL might look like the following:

https://graph.facebook.com/search?q=mark&type=user&access_token=[access_token].

Google Custom Search API: This API allows developers to set up a custom search engine (CSE) that is used to search a specific set of domains, and then access the results in a JSON or Atom format. While only being able to search a subset of domains may seem restricting, with a little bit of effort, we can create a CSE that includes all sites.

After setting up this CSE, we can easily pull results for Twitter users, Linkedin users, documents from companies' websites, etc. For example, Linkedin for Massey University:

site:linkedin.com intitle: "| Linkedin" "at Massey University" -intitle:profiles -inurl:groups -inurl:company -inurl:title

LinkedIn does have its own API; however, this can be expanded to include more sites such as Twitter

site:twitter.com intitle: "on Twitter" "Massey University"

In this way data can be gathered easily and then aggregated. Many new tools include data aggregators this automation is present in greater and lesser degrees depending on the tools. These tools claim to can serve up data and analysis. They promise to be models and sources. These all in one tools for BI/OSINT are on the rise and these are often backed by blogs and information to inform the user. The growing list shows the expansion of the interest in the field and the multiple users and uses, includes: APIs, scrapers, offensive security measures including penetration-testing tools: exiftool data, the harvester, Maltego, Cogito, Cree.py, Metasploit, Scra.py, recon-ng, Panda, R, SAS.

Recon-ng styles itself a web reconnaissance framework; authored by Tim Tomes; sponsored by Black Hills Information Security; and written in Python it can certainly get you some data and store it in its tool database. The tool jigsaw.rb is a ruby script that scrapes the contact website Jigsaw for contact details and generates email addresses on the fly. Maltego has the free community version that is available by itself or in Kali Linux, a Linux distribution available for those investing time in penetration testing and computer security. Kali Linux is one of several Offensive Security projects – funded, developed, and maintained as a free and open-source penetration-testing

platform. This tool provides automatic OSINT-gathering techniques using "transforms." The data are then presented and manipulated using an intuitive graphical interface of a force-directed graph. Spokeo is a search engine for social information. By enter a name, email address, username, phone number, etc., one can find information across a variety of social networking platforms and other sources.

There are many other manual and automatic tools that can help us in our OSINT-gathering process, but to talk about what to do with the data now that you have it. NoSQL, big data analytics, cloud, and database as a service (DBaaS), all these have/are new approaches to breaking the code of what is happening and what can be made into actionable insights from that. Tools may not specify which algorithms they use or they make reference which family of algorithms. The tools recon-ng, Maltego, SAS, R, Weka, and Panda all have data-mining utilities. Recon-ng and Maltego are solid and advanced setup and API keys are required, but deep dives may be done on data that the tools acquire. SAS and R have been in use for sometime in research and business and have solid statistical grounding. They are not data-gathering tools abut data mining and data modeling tools as well as languages. Panda PyBrain and SciPy data learning machine learning modlues that are availiable in python. Weka is a java based software for machine learning. These all have in common that they do not gather data, but will take API input and other data formats.

ALGORITHMS

There are overlapping method names within OSINT due to the fusion of disciplines, each with their own vocabularies that deal with the algorithms associated with open source intelligence gathering. Perhaps it is best to say that open source intelligence gathering is open to all the useful algorithms so Statistics, Machine Learning, Pattern Recognition, Computer Science, Applied Mathematics, have claims on the algorithms used for OSINT. Many papers use more than one algorithm or methodology to examine the data and provide a more accurately describe the data and context. Watters (2012); Wibberley and Miller (2014); Jin, Li, & Mehrotra, (2003) are good examples of this multiple method approach. In business and research often many of the algorithms that are gaining ground are unnamed or unspecified in tools. Some standard others are measured by the field they come from: machine learning and natural language processing; event detection; predictive analytics (notably non-machine learning based); network analysis; and manual analysis.

The importance of using algorithms, scientific methods, and isolating the cognitive bias in your open source intelligence can be seen very clearly in the false accusations that appeared on Redit claiming they had discovered the Boston

Bomber, when in fact they had discovered an unfortunate person who had been at the Boston Marathon and then disappeared as he had drowned (http://www.bbc.com/news/technology-22263020).

The algorithms used in OSINT are found for the generalized problems that they solve. The topics describe the problems being solved and how the solver uses them. In the terminology of machine learning, classification is considered an instance of supervised learning, where correctly identified observations are available to be used as a training set. The unsupervised procedure clustering involves grouping data into categories based on measure(s) of similarity or nearness. Semisupervised learning topics such as natural language processing (Noubours, Pritzkau, & Schade, 2013) and adaptive resonance theory (Carroll, 2005) cover a variety of algorithms span learning methods and descriptors. These are well represented in OSINT. Supervised learning includes: structured predictive methods; classification (Watters); decision trees (Watters) and ensemble methods such as bagging, boosting, and random forests, and *k* means nearest neighbors; neural networks and Naïve Bayes, support vector machine and relevance vector machine, linear and logistic regression (Churn and company attrition models). Neural networks include deep learning. Anomaly detection is a widely discussed set of algorithms which include *k* - means nearest neighbors and local outlier factors.

Unsupervised learning includes: BIRCH, hierarchical, *k*-means, expectation maximization, density-based spatial clustering of applications with noise, OPTICS, mean-shift. There are many clustering algorithms differing by what the measure used to cluster the data. Papers using clustering include Herps, Watters, and Pineda-Villavicencio (2013) and Watters.

Cluster analysis or clustering is the task of grouping a set of objects in such a way that objects in the same group are more similar by a decided measure to those in other clusters. It is a common technique for statistical data analysis, used in many fields, including machine learning, pattern recognition, image analysis, information retrieval, and bioinformatics.

Cluster is not one algorithm, but the over arching description of the problems solved. It can be achieved by various algorithms that differ significantly in their notion of what constitutes a cluster and how to efficiently find them. Popular notions of clusters include groups with small distances among the cluster members, dense areas of the data space, intervals, or particular statistical distributions. Clustering can therefore be formulated as a multiobjective optimization problem. The appropriate clustering algorithm and parameter settings (including values such as the distance function to use, a density threshold, or the number of expected clusters) depend on the individual data set and intended use of the results. Cluster analysis as such is not an automatic task, but an iterative process of knowledge discovery or interactive multiobjective

optimization that involves trial and failure. It will often be necessary to modify data preprocessing and model parameters until the result achieves the desired properties (Kogan, 2007). The differences between clustering methods and disciplines can be seen in the usage of the results: in data mining, the resulting groups are the matter of interest, in automatic classification the resulting discriminative power is of interest. This can lead to misunderstandings between researchers coming from the fields of data mining and machine learning, since they use the same terms and often the same algorithms, but have different goals.

Herps, Watters, Pineda-Villavicencio only look at the top 50 Alexa sites for Australians making policy suggestions and cluster analysis of site type and overlaps between cookies and relationships between websites (Herps et al., 2013). This analysis culminates in building a directed graph model of the 50 sites; the note that the penetration of Google analytics is 54% of the sample which is closely followed by social media sites. Facebook was storing cookies on 17% of the sites. There is the discussion of the possibility of a privacy feedback loop created by the linking of sites sharing cookies like Facebook and GA. Ending by recommending that Australia adopt laws similar to the European cookie laws due to the pervasiveness of the tracking involved and the amount of data captured.

Reducing the number of independent variables used is an algorithm called dimension reduction. Machine learning and statistics divide this field into feature selection and feature extraction. Feature selection approaches try to find a subset of the original variables. Two strategies are filtering by information gain and wrapper, which is guided by accuracy approaches. These can sometimes be considered combinatorial optimization problems. These algorithms create a reduced space, in which data analysis can sometimes be done more accurately than in the original space. Feature extraction transforms the data in the high-dimensional space to a space of fewer dimensions. The data transformation may be linear, as in principal component analysis but many nonlinear dimensionality reduction techniques also exist. For multidimensional data, tensor representation can be used in dimensionality reduction through multilinear subspace learning.

Random forests are used in Watters et al. to group website output and owners in conjunction with other methods. In machine learning and statistics, classification is the problem of identifying to which of a set of categories a new observation belongs, on the basis of a training set of data containing observations whose category membership is known. Often, the individual observations are analyzed into a set of quantifiable properties. Classifiers work by comparing observations to previous observations by means of a similarity or distance function. The term ″classifier″ refers to the mathematical function that maps input data to a category.

Re-identification and de-anonymization are booming areas of study covering from record linkage to graph traversal and graph visualization tools: (Narayanan & Shmatikov, 2008) introduce a new simulated annealing-based weighted graph matching algorithm for the seeding step of de-anonymization. They also show how to combine de-anonymization with link prediction – the latter is required to achieve good performance on the portion of the test set not de-anonymized – for example by training the predictor on the de-anonymized portion of the test set, and combining probabilistic predictions from de-anonymization and link prediction. "Even more generally, re-identification algorithms are classification algorithms for the case when the number of classes is very large. Classification algorithms categorize observed data into one of several classes, that is, categories. They are at the core of machine learning, but typical machine-learning applications rarely need to consider more than several hundred classes. Thus, re-identification science is helping develop our knowledge of how best to extend classification algorithms as the number of classes increases." (Narayanan, 2013). The set-theoretic record linkage can be considered within unsupervised learning; by taking the intersections of intersections of multiple openly available databases one can narrow down targets and re-identify people (skydog and security freak at SkydogCon 2012, https://www.youtube.com/watch?v=062pLOoZhk8). Indirect record linkage has been successfully used to group authors of social media (Layton, Perez, Birregah, Watters, & Lemercier, 2013).

While re-identification can cause data leakage, it is mostly used to clarify data. Re-identification is "record linkage" that connects common entities across different databases. Often in real data sets, a useful unique identifier does not exist. In an overview of "record linkage" research of more than a hundred papers, William Winkler of the U.S. Census Bureau discusses the new moves toward automation. Record linkage is used to remove duplicates, or combine records so that relationships in multiple data elements from separate files can be examined. It can be one of the strongest de-anonymization techniques. In the past record linkage was largely a manual process or one that used elementary and *ad hoc* rules. Matching techniques that are based on formal mathematical models subject to testing are now more likely to be in use rather than exceptional (Winkler, 2006). While it can be said that a large amount of re-identification can be done with little more than solid programming and statistics skills (Skydog and security Freak, 2012; Narayanan et al., 2011; Narayanan, 2008, www.youtube.com/watch?v=062pLOoZhk8). "...The history of re-identification has been a succession of surprising new techniques rendering earlier data sets vulnerable" (Narayanan et al., 2015).

In 2000, Sweeney claimed to show 87% of the U.S. population can be uniquely re-identified based on five-digit ZIP code, gender, and date of birth (Sweeney, 2000). The ability to de-anonymize data requires little in the way

of advanced analytics and as the tragedy of the data commons and many other ethics publications show there is no way to anonymize data beyond recovery, merely to anonymize beyond current openly available data with no guarantee for the future. This can be seen in the Netflix study (research by Narayanan and Shmatikov revealed that with minimal knowledge about a user's movie preferences, there is an over 80% chance of identifying that user's record in the Netflix Prize data set – a targeted attack (Narayanan & Shmatikov, 2008) and the Target case and the governor case (Sweeney, 2005). Intersecting open data sources is one of the ways to de-anonymize. Leveraging what is known in several data sets to easily find the hiding people within the crowd has become trivial given the right leveraging data.

"Generally the reason a data set can not be de-anonymized is due to the lack of publically published data at this time, like that of the Heritage Health Prize" (El Emam et al., 2012). It has been shown that it is possible to identify Netflix users by cross-referencing the public ratings on IMDb. Even more broad attacks could be possible depending on the quantity and availability of information in possession of the adversary for cross-referencing (Narayanan & Shmatikov, 2008a, 2008b). Data custodians face a choice between roughly three alternatives: sticking with the old habit of de-identification and hoping for the best; turning to emerging technologies like differential privacy that involve some tradeoffs in utility and convenience; and using legal agreements to limit the flow and use of sensitive data (Narayanan, 2011) (Narayanan & Felten, 2014).

Record linkage can include mapping algorithms utilizing multidimensional Euclidean space that preserves domain-specific similarity. A multidimensional similarity join over the chosen attributes is used to determine similar pairs of records (Jin, Li, & Mehrotra, 2003). Likage is very strong in conjunction with geospatial data. This blog is about a publicly available data set of bicycle journey data that contains enough information to track the movements of individual cyclists across London, for a six-month period. "…There are benign insights that can be made by looking at individual profiles – but the question remains whether these kinds of insights justify the risks to privacy that come with releasing journey data that can be associated with individual profiles." (http://vartree.blogspot.co.nz/2014_04_01_archive.html)

Geospatial record linkage and other algorithms can be very powerful. Even Jon Snow's first geospatial analysis in (mapping cholera in 1854) is a case of record linkage and leveraging available data to the common good (Johnson, 2006). Because of the great good gained by geographic data. There is a risk in misunderstanding the anonymization of geospatial data. Narayanan et al. (2015) make the argument that geospatial data cannot or have not yet been able to be anonymized in a way that would ensure the privacy of the data owners

or describers. In several examples, geospatial data are leveraged to find patterns and start and ends of journeys lead to knowing places of business and homes ((Narayanan et al., 2015); (Barocas & Nissenbaum, 2014); "FOILing NYC's Taxi Trip Data" http://chriswhong.com/open-data/foil_nyc_taxi/; using British bicycle data "I know where you were last summer" Blog post http://vartree.blogspot.co.nz/2014_04_01_archive.html, "Mapping Where Sex Offenders Can Live" http://automatingosint.com/blog/category/osint/). Spatial analysis is the techniques applied to structures at the human scale, most notably in the analysis of geographic data. Complex issues arise in spatial analysis, many of which are neither clearly defined nor completely resolved, but form the basis for current research. This will be covered in more depth in another chapter. A 2013 study by de Montjoye et al. analysed a mobile phone data set covering 15 months of recorded locations of the connecting antenna each time one of the 1.5 million users called or texted; evaluated the uniqueness of mobility traces (i.e., the recorded data for each user for data points that have antenna location and timestamp). Two random data points uniquely identify over 50% of users. 95% of users are uniquely identifiable using four random data points, which could be revealed through social media then. Geographic data is especially telling using record linkage; that is when multiple data sets are chained together to a non-anonymous data set. These become low hanging fruit easily re-identifying individuals in all of those data sets. In a famous re-identification of the former Governor Weld's medical record used a basic form of this record linkage: Sweeney used gender, date of birth, and ZIP code found through a public data set of registered voters and then used that information to identify him within the de-identified medical database. The geographic data is one of the most telling links in this chain. Hooley and Sweeney's more recent work record linking remains effective on public hospital discharge data from thirty U.S. states in 2013 Hooley and Sweeney, 2013. Other studies like those of Montjoye et al. have cemented that pairs of home and work locations can be used as strong unique identifiers. "…the uniqueness of mobility traces decays approximately as the 1/10 power of their resolution. Hence, even coarse data sets provide little anonymity." (de Montjoye et al., 2013). Geospatial analysis has its own special tools and packages such as: Leaflet.js - web mapping library, GeoEye, Cloudmade - map tiles, Transport For London - data sets of Boris Bike data. There are multiple blogs such as http://spatial.ly/ which has visualizations, analysis, and resources. The API for Wikimapia works like any other API and using this with tools like the geopy Python module https://github.com/geopy/geopy, which handles measurements, can return useful results. An socially responsible example is this map of fracking in the US gives some of the most definitive answers about how much of the US has fracking and where and what kinds of contamination allowing organization of opposition to these occurrences (https://secure3.convio.net/fww/site/SPageServer?pagename=national_parks_public_lands_fracking_2014).