Forensics in Big Data Era

Big Data, Cybercrime and Digital Forensics

Yudi Prayudi
Pusat Studi Forensika Digital
Teknik Informatika
Universitas Islam Indonesia

Profile

Dr. Yudi Prayudi, S,Si, M.Kom

Director of Center for Digital Forensics Studies UII Yogyakarta

- Doctoral From UGM With Dissertation about Digital Evidence Cabinets
- Senior Lecture at Department of Informatics UII Yogyakarta
- Head of Digital Forensic Postgraduate Program at UII Yogyakarta

- CHFI , Encase, Oxygen, Belkasoft
- Hacker In The Box (2012), Hacker Halted (2012)

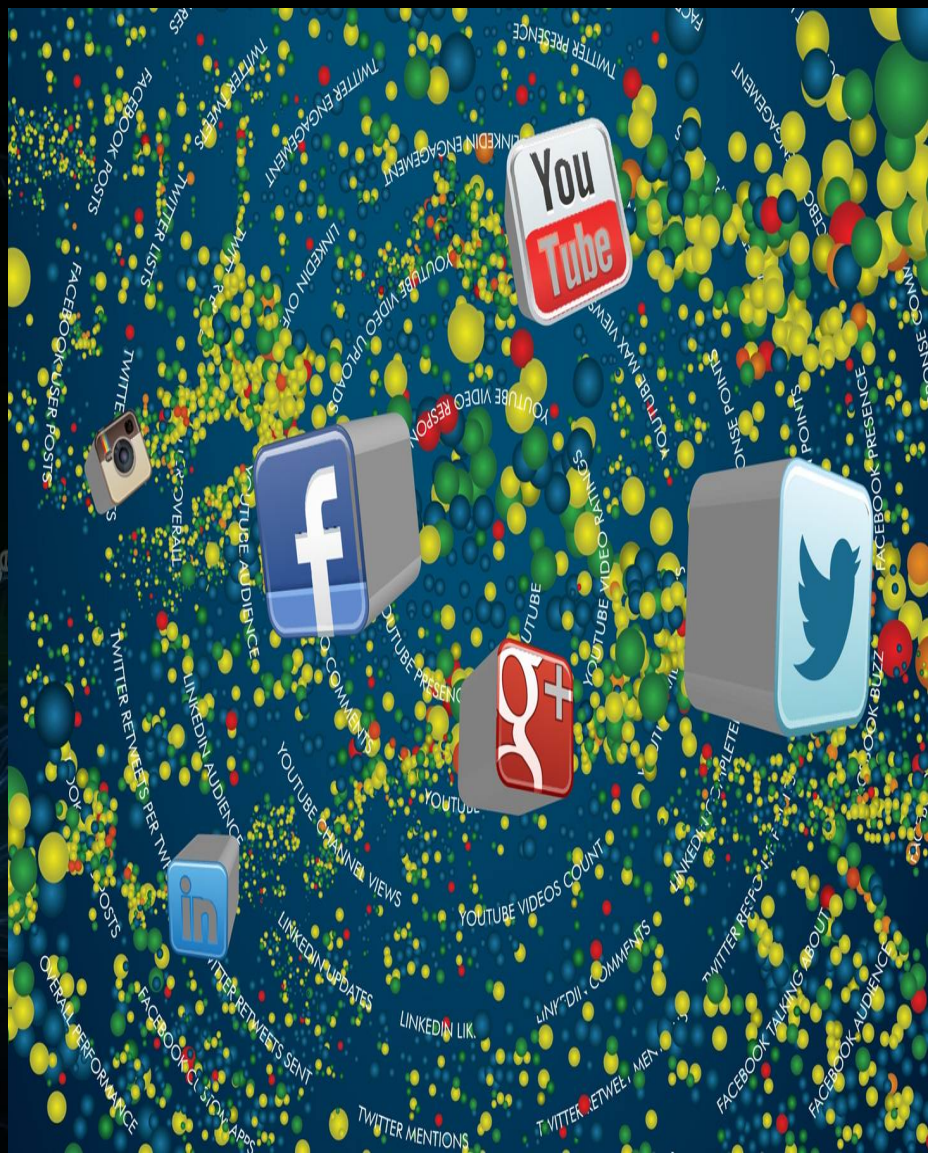
- Author more than 25 Paper for International Journal and Conference
- https://www.researchgate.net/profile/Yudi_Prayudi

- Speakers at Many National Conferences about Security and Digital Forensics

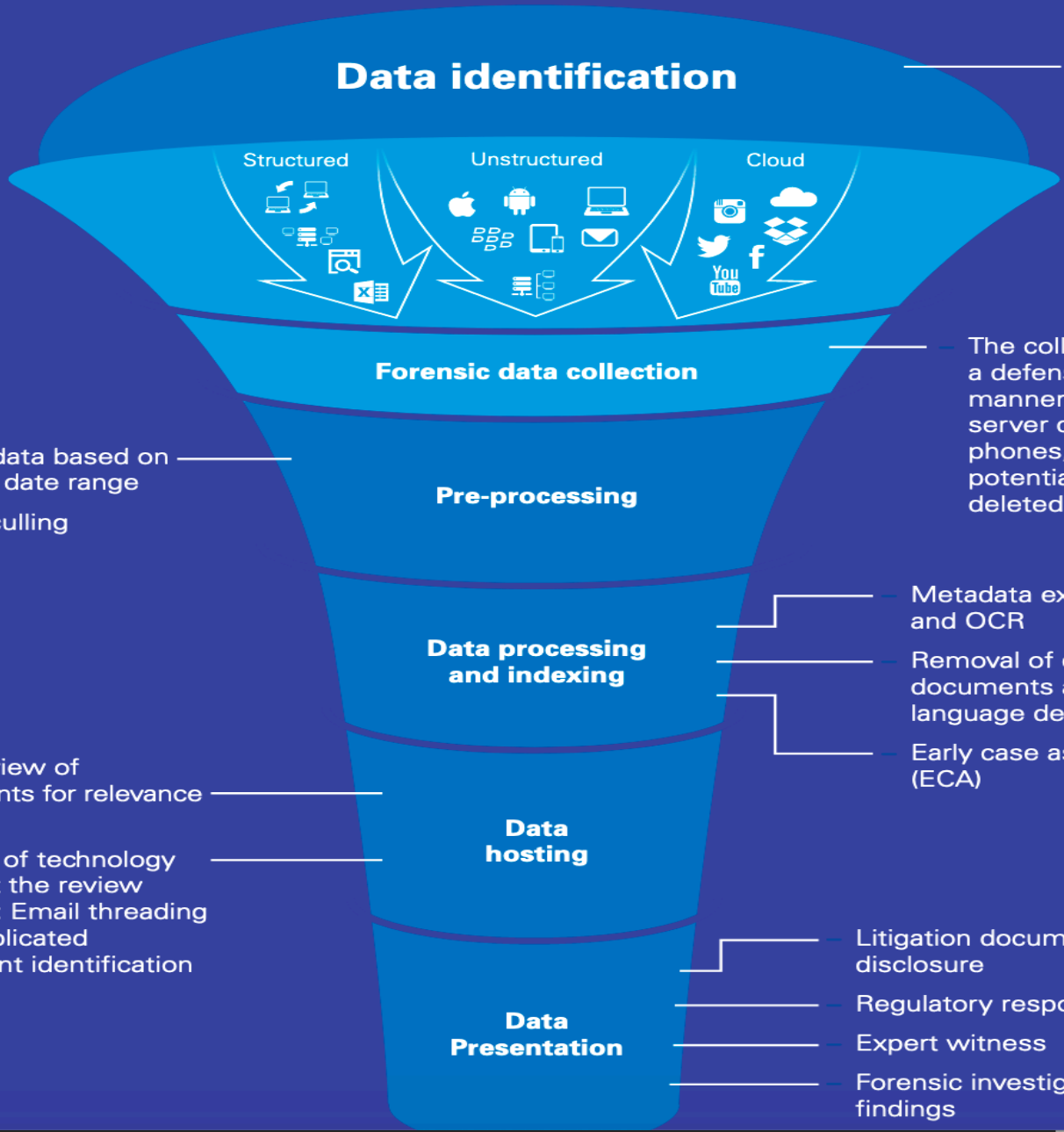
Agenda

- Big Data And Crime Analysis
- Big Data and Crime Predictive
- Forensics for Big Data
- Potencial Crime Using Big Data
- Preventive and Protection

Digital Universe



Data management funnel



Determining sources of target rich data for investigation purposes

The collection of dta in a defensible forensic manner (laptops, PCs, server data, mobile phones, tablets) potential recovery of deleted data

- Removal of data based on file type and date range
- Date range culling

- Metadata extraction and OCR
- Removal of duplicate documents and language detection
- Early case assessment (ECA)

- Web review of documents for relevance
- The use of technology to assist the review process: Email threading near duplicated document identification

- Litigation document disclosure
- Regulatory response
- Expert witness
- Forensic investigation findings

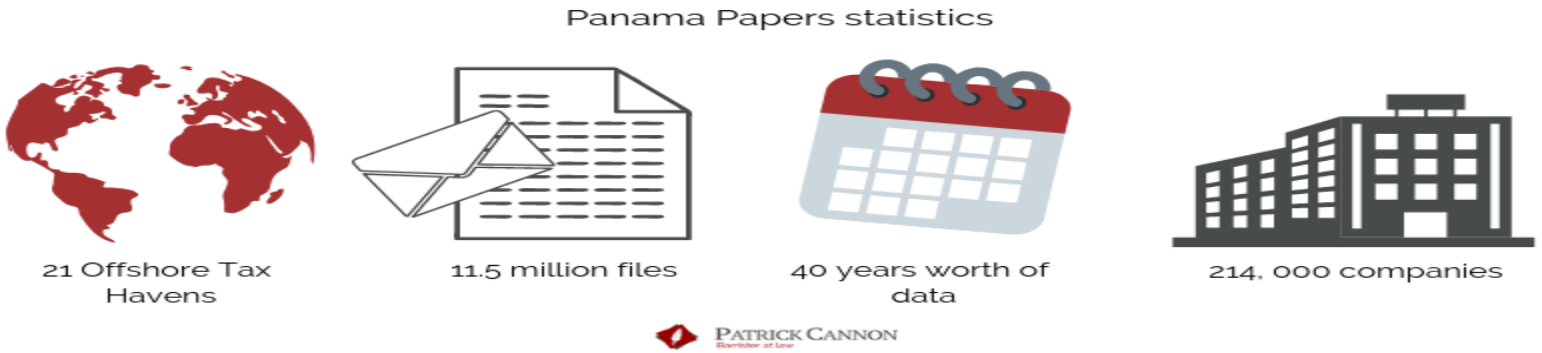
Big Data and Forensics

- Big data forensics as an special branch of digital forensics where the identification, collection, organization, and presentation processes deal with a very large-scale dataset of possible evidence to establish the facts about a crime.

- Big data forensics can be discussed from two perspectives.
  - First, a small piece of evidence can exist in a big dataset. For example, to investigate a criminal incident, we may need the information of a few call records among the 1.9 trillion call records of AT&T.
  - Second, a crucial piece of information can be revealed by analyzing a big dataset or by correlating data of multiple big datasets. For example, a spam email classifier that is trained on small dataset may not perform well in determining a new spam email. However, when the classifier is trained on a big dataset, most likely it will perform better in identifying a new spam email.

Big Data and Crime Analysis

The Panama Paper

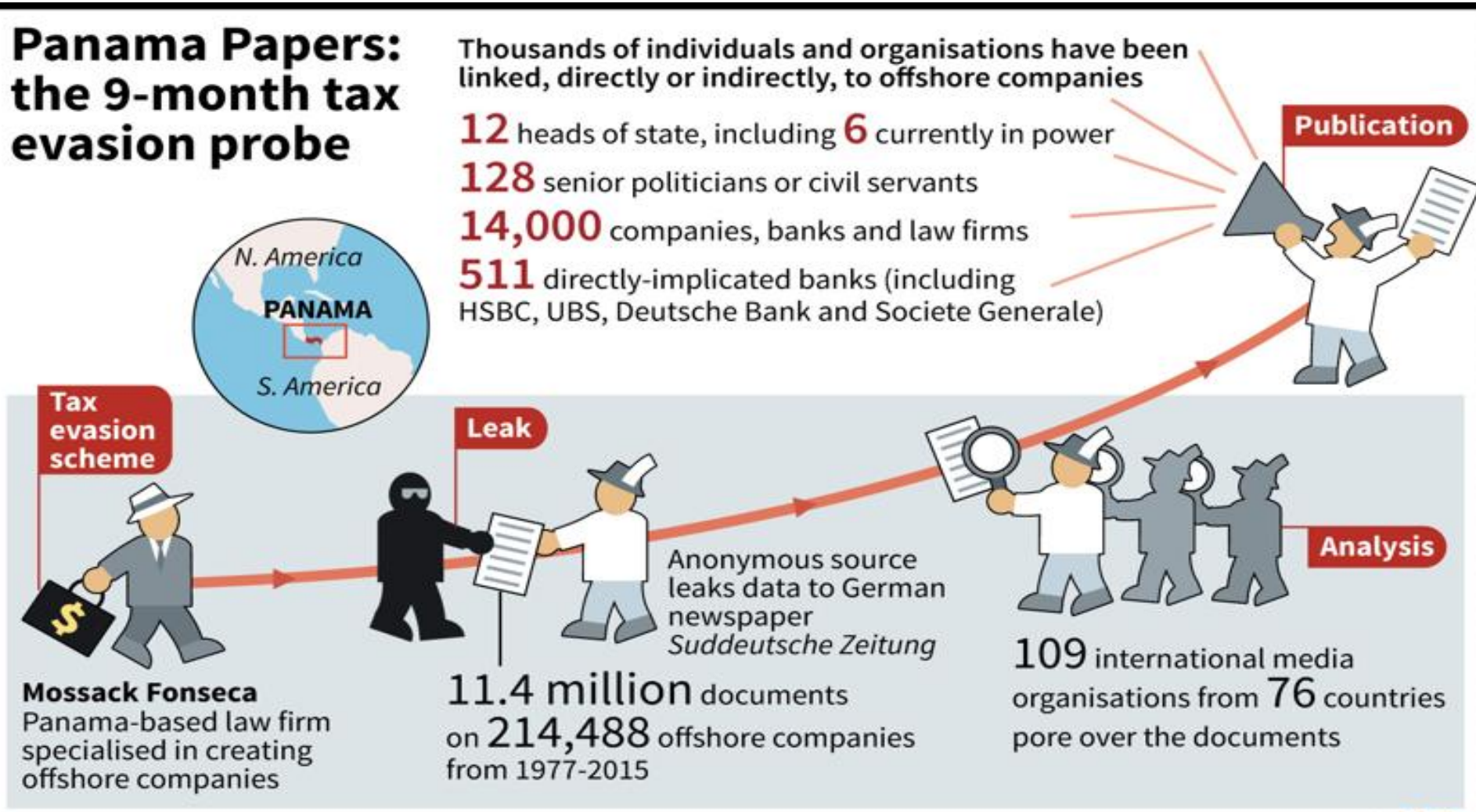


Nuix, an Sydney-based software company, supplied document processing and investigation technology to the International Consortium of Investigative Journalists (ICIJ) to help them process the 11.5 million documents from Panamanian law firm Mossack Fonseca.

German newspaper Süddeutsche Zeitung and the ICIJ used Nuix software to "process, index, and analyse" the data

Investigators used Nuix's optical character recognition to make millions of scanned documents text-searchable. They used Nuix's named entity extraction and other analytical tools to identify and cross-reference the names of Mossack Fonseca clients

The Panama Papers



Terorism Case

Passenger Name Record (PNR)
Typical Data Elements

1. PNR record locator code
2. Date of reservation
3. Date(s) of intended travel
4. Name
5. **Other names on PNR**
6. **Address**
7. **All forms of payment information**
8. Billing address
9. **Contact telephone numbers**
10. All travel itinerary for specific PNR
11. **Frequent flyer information (miles flown, address)**
12. **Travel agency**
13. Travel agent
14. Code share PNR information
15. Travel status of passenger
16. Split/Divided PNR information
17. **Email address**
18. Ticketing field information
19. General remarks
20. Ticket number
21. Seat number
22. Date of ticket issuance
23. No show history
24. Bag tag numbers
25. Go show information
26. OSI information *
27. SSI/SSR information *
28. Received from information
29. All historical changes to the PNR
30. Number of travelers on PNR
31. Seat information
32. One-way tickets
33. Any collected APIS information
34. ATFQ fields

Keeping A Future Terrorist Out of the United States

Situation

- DHS uses a sophisticated data assessment program called the Automated Targeting System (ATS) to perform risk assessments on those who seek to enter the United States
- Roughly 411 million people attempt to enter the U.S. annually; nearly 91 million come to the U.S. by air
- ATS allows DHS' Customs and Border Protection (CBP) to connect the dots to foil potential terrorist plots by denying entry to would-be terrorists

Action

- June 14, 2003: a Jordanian named Raed al-Banna, carrying a valid business visa on his Jordanian passport, tries to enter the U.S. at O'Hare Airport
- His data is run through ATS; CBP is wary of his travel habits and takes him to secondary screening
- CBP gleans that al-Banna has been living in the U.S. since 2001 and illegally working petty jobs
- A CBP officer takes al-Banna's photograph and fingerprints, and he is deported soon there after

Result

- February 28, 2005: al-Banna rams a car filled with explosives into a crowd of military and police recruits in the Iraqi town of Hillah, killing more than 125
- His hand and forearm are found inside the smoldering vehicle handcuffed to the steering wheel
- The attack remains one of the deadliest suicide bombings in Iraq since the inception of the war

The CBP officer who deported al-Banna said, "I was shocked. That it was so close to home, that I actually interviewed someone who not only was capable of doing but actually did something like that. You never know who you are interviewing or what they are capable of doing."

Big Data and Crime Predictive Model

Fighting Crime

- Fighting crime is a major concern in the 21st Century.
- The good news is that big data is proving to be a valuable tool in the arsenal of law enforcement officials in every jurisdiction.
- This will help reduce the crime epidemics we face.

The Strategies

- Primary prevention strategies
  - minimize the risk factors associated with criminal behavior. These programs, often housed in schools and community centers, are intended to improve the health and well-being of children and young adults.
- Criminal justice strategies address known offenders;
  - correctional facilities and prison rehabilitation aim to prevent convicted criminals from offending again.
- Law enforcement strategies focus on decreasing the probability that crime occurs in a particular area.
  - reducing the opportunity for criminal acts and increasing the risk of arrest.

- Predictive analytics is one law enforcement strategy to accomplish this form of prevention. By compiling and analyzing data from multiple sources, predictive methods identify patterns and generate recommendations about where crimes are likely to occur.

Predictive Policing

WHAT IS PREDICTIVE POLICING?

Data plays an important role in allocating police resources and lowering crime rates:

1929

Uniform Crime Report first developed and implemented by the FBI

1960's

"LEMRAS" algorithm distributes squad cars according to geographical topology of historical crime data

1990's

"Crime mapping" becomes a popular approach to local policing

1996

Digital data storage becomes cheaper than paper

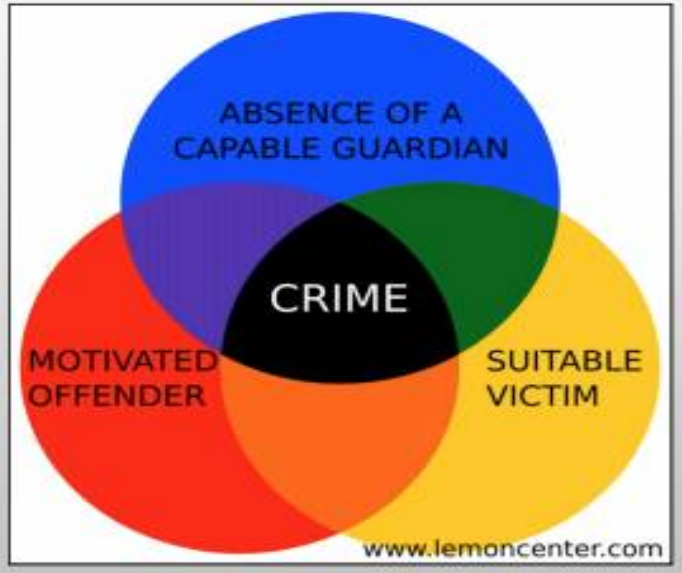
Today

Big Data and analytics enable predictive policing by allowing officers to understand where and when crime is likely to occur

Predictive Policing

Theory Behind Predictive Policing

Routine Activities Theory



Police Presence = Deterrence

Probability of Arrest

↓

Risk vs Reward
Risk = too high for offender

↓

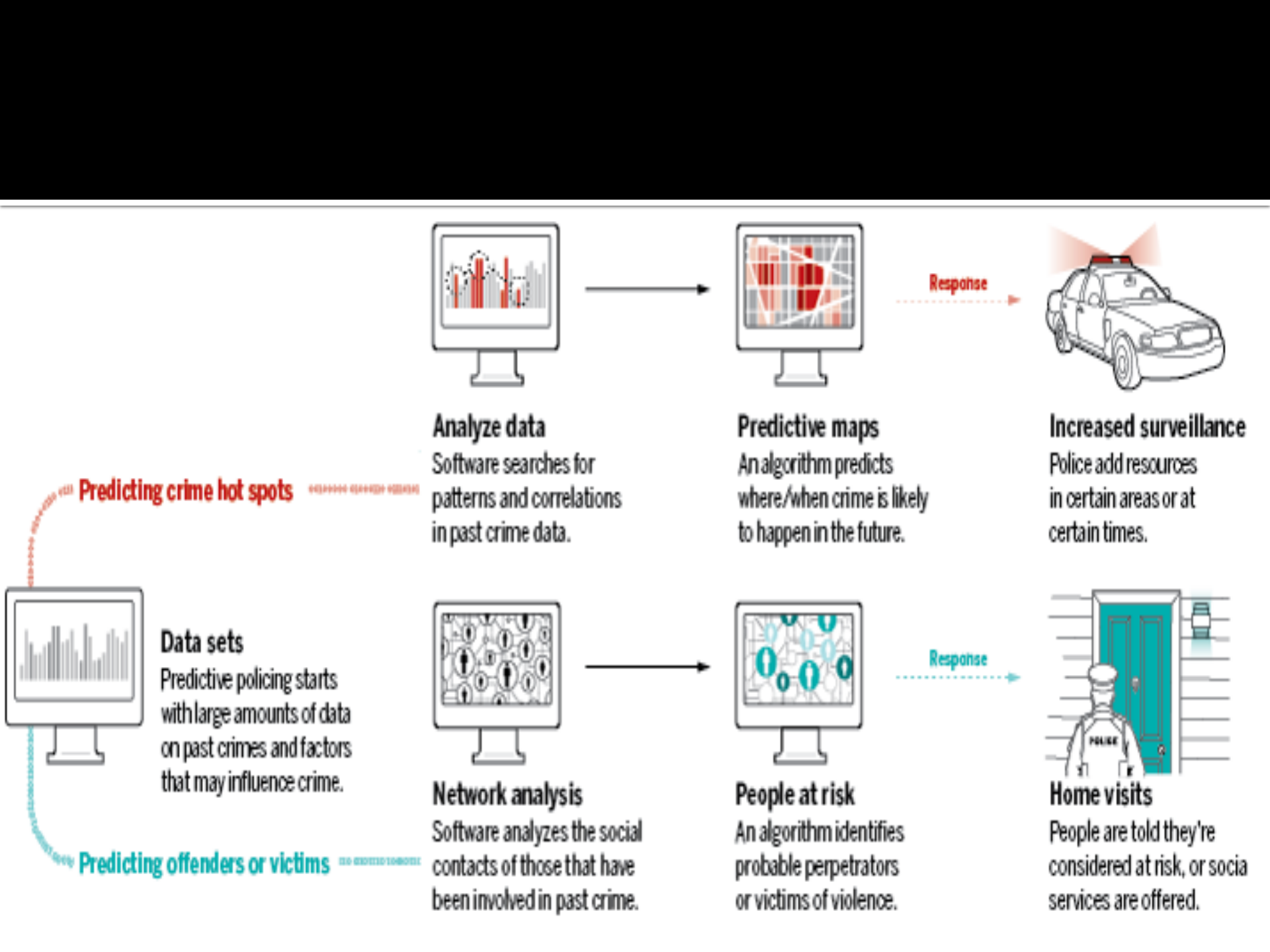
Reduced Crime

Crime Predictive Model

- Crime mapping is used to analyze, map and visualize crime incidents or crime pattern to have an idea for predicting the crime occurrence.
- Three phase
  - Distribution of data geographically and creating clusters,
  - Cluster analysis of created clusters and
  - Prediction of crime.

Three Basic Theory

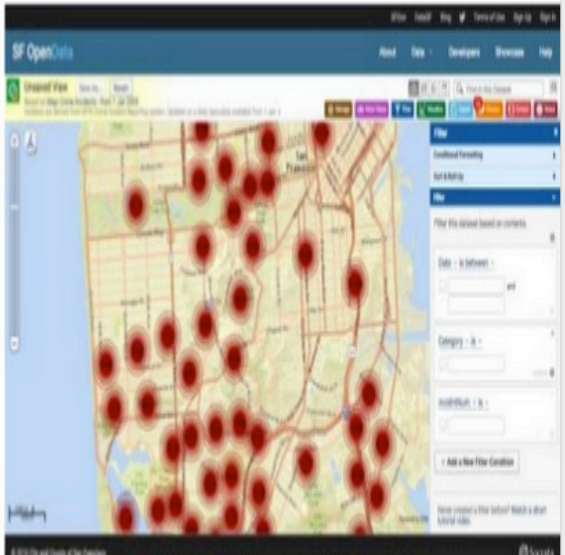
- Routine Activity Theory
  - states that crime depends on multiple factors including the motivation of offenders, suitable targets and an absence of capable guardians.
- Rational Choice Theory
  - underlines that criminals make rational decisions based on opportunity and estimated costs such as the possibility of being imprisoned and punished;
- Crime Pattern Theory
  - explains why, when and where crime happens, focusing on the intersections and commonalities between victims and perpetrators.


Predicting crime hot spots
Data sets
Predictive policing starts with large amounts of data on past crimes and factors that may influence crime.
Predicting offenders or victims
Analyze data
Software searches for patterns and correlations in past crime data.
Predictive maps
An algorithm predicts where/when crime is likely to happen in the future.
Response
Increased surveillance
Police add resources in certain areas or at certain times.
Network analysis
Software analyzes the social contacts of those that have been involved in past crime.
People at risk
An algorithm identifies probable perpetrators or victims of violence.
Response
Home visits
People are told they're considered at risk, or socia services are offered.

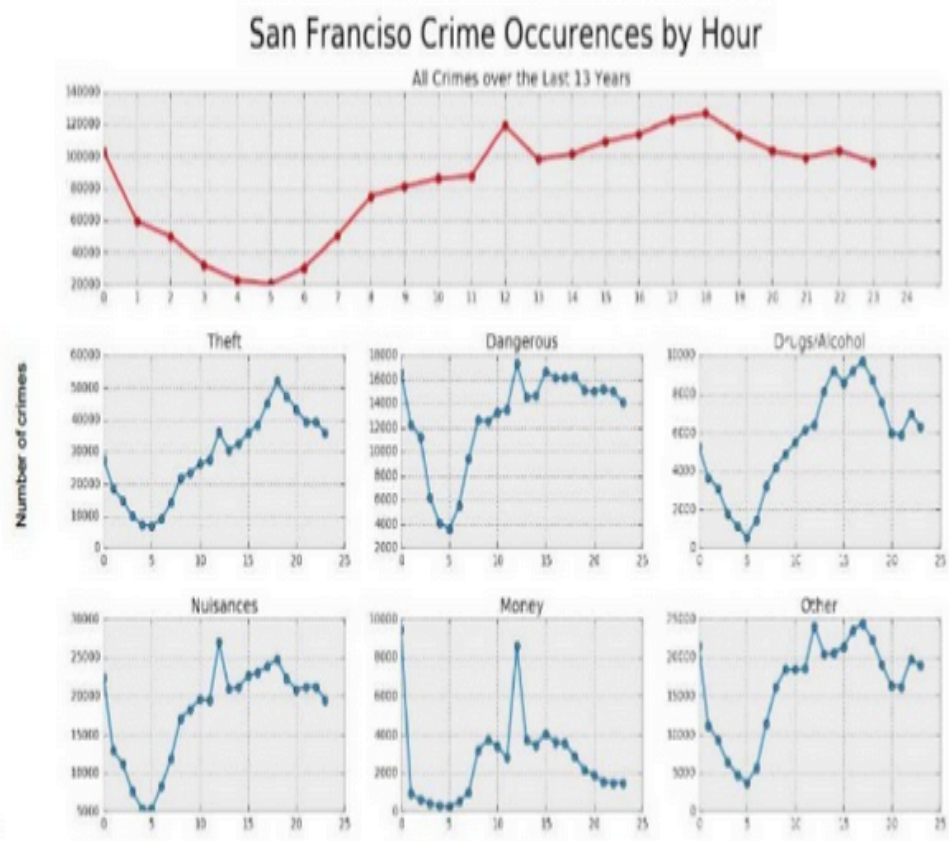
Crime Monitoring

SAN FRANCISCO CRIME DATA

- Accessed from SFOpenData – over 13 years of crime data
- Analyzed 2 million instances of crime as of December 2016
- Anonymized data
- Information includes:
  - Dates
  - Category
  - Descript
  - DayOfWeek
  - PdDistrict
  - Resolution
  - Address
  - X & Y

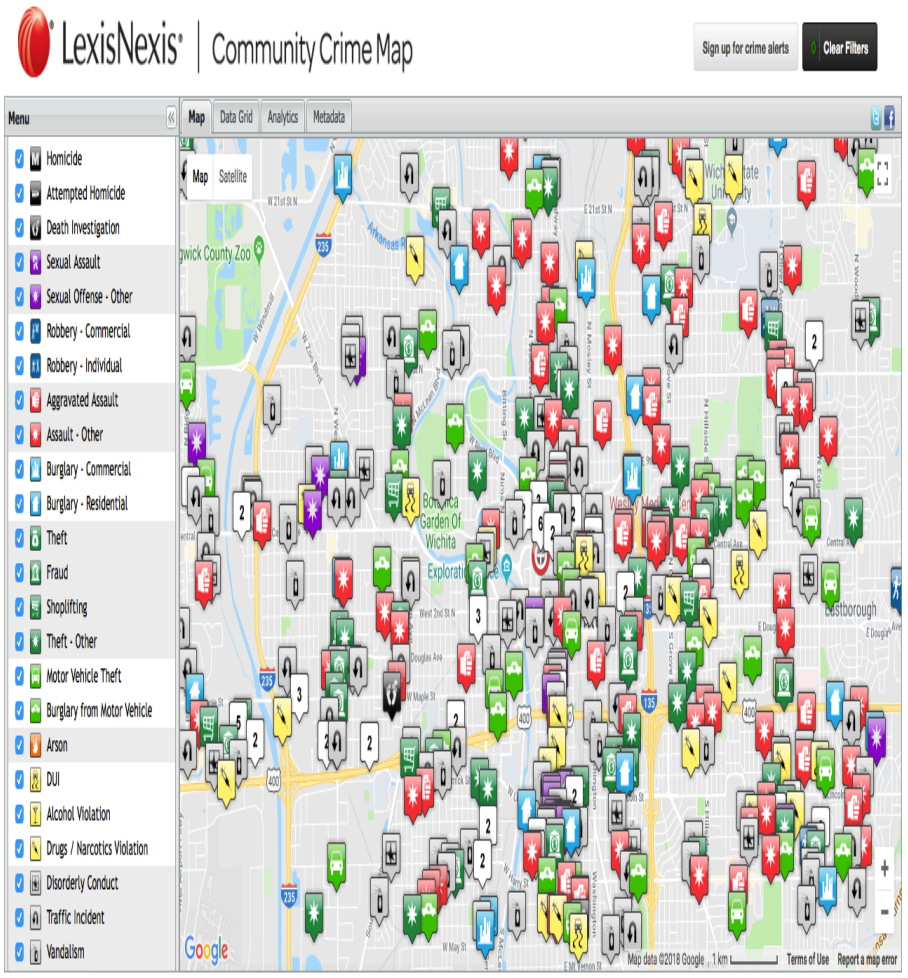


CRIME OCCURRENCES OVER LAST 13 YEARS BY HOUR



Commercial Apps

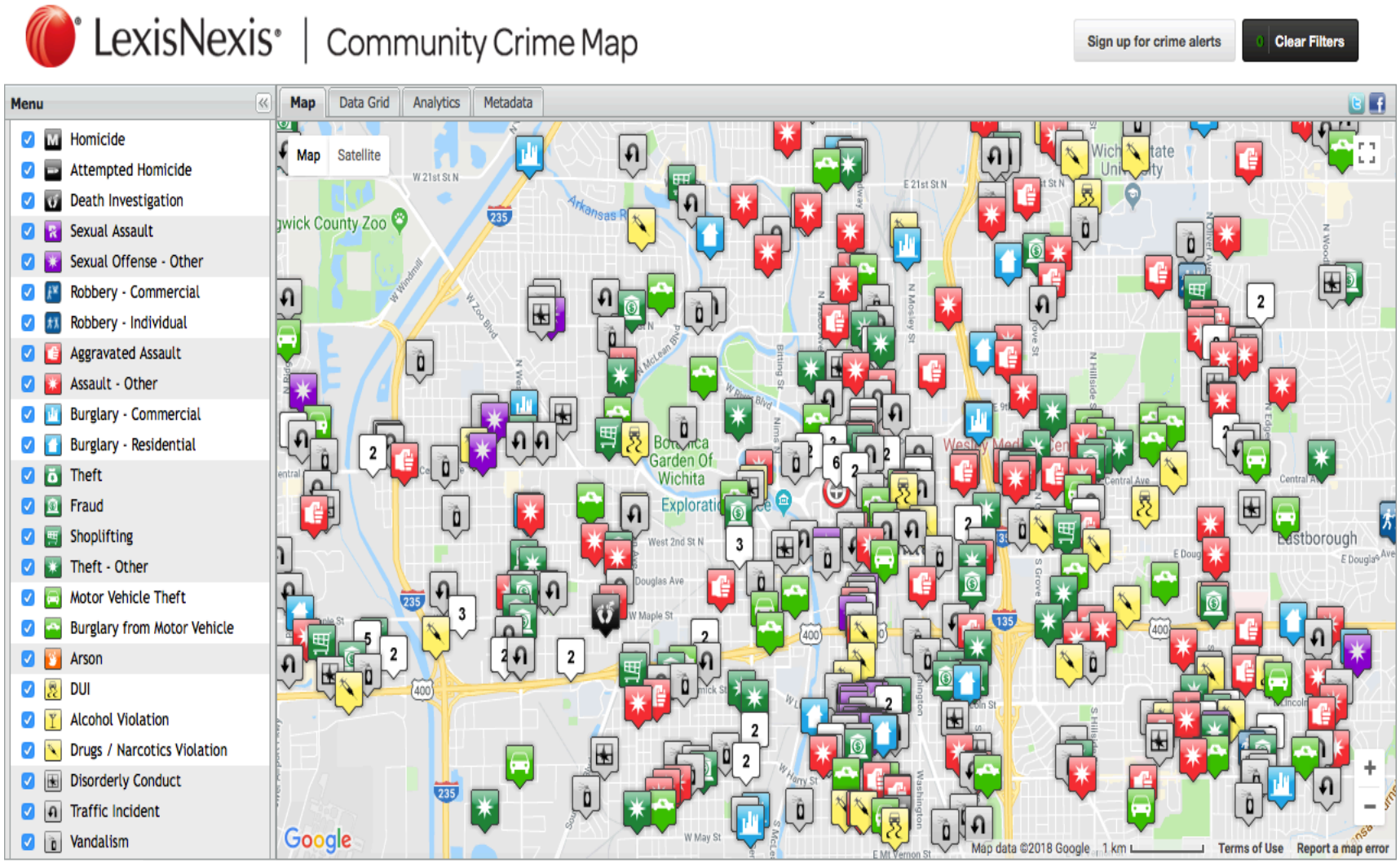
HunchLab



Lexis Nexis

Crime Map

https://communitycrimemap.com



Smart Cities Implementation

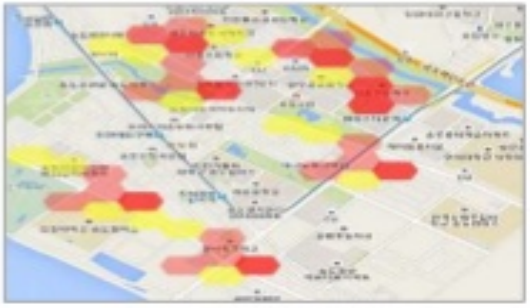
03-7 Information based on IoT & Big Data

Crime Heat Map

Traffic Flow Statistics

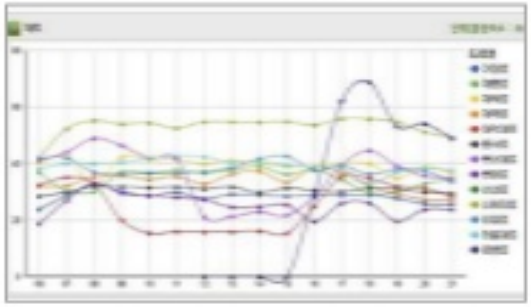
Bus Passengers Statistics

Foot Traffic Statistics



Travel Time Trends

Traffic Volume per Each Road



Security Issues for Big Data

Big Data Security Issues

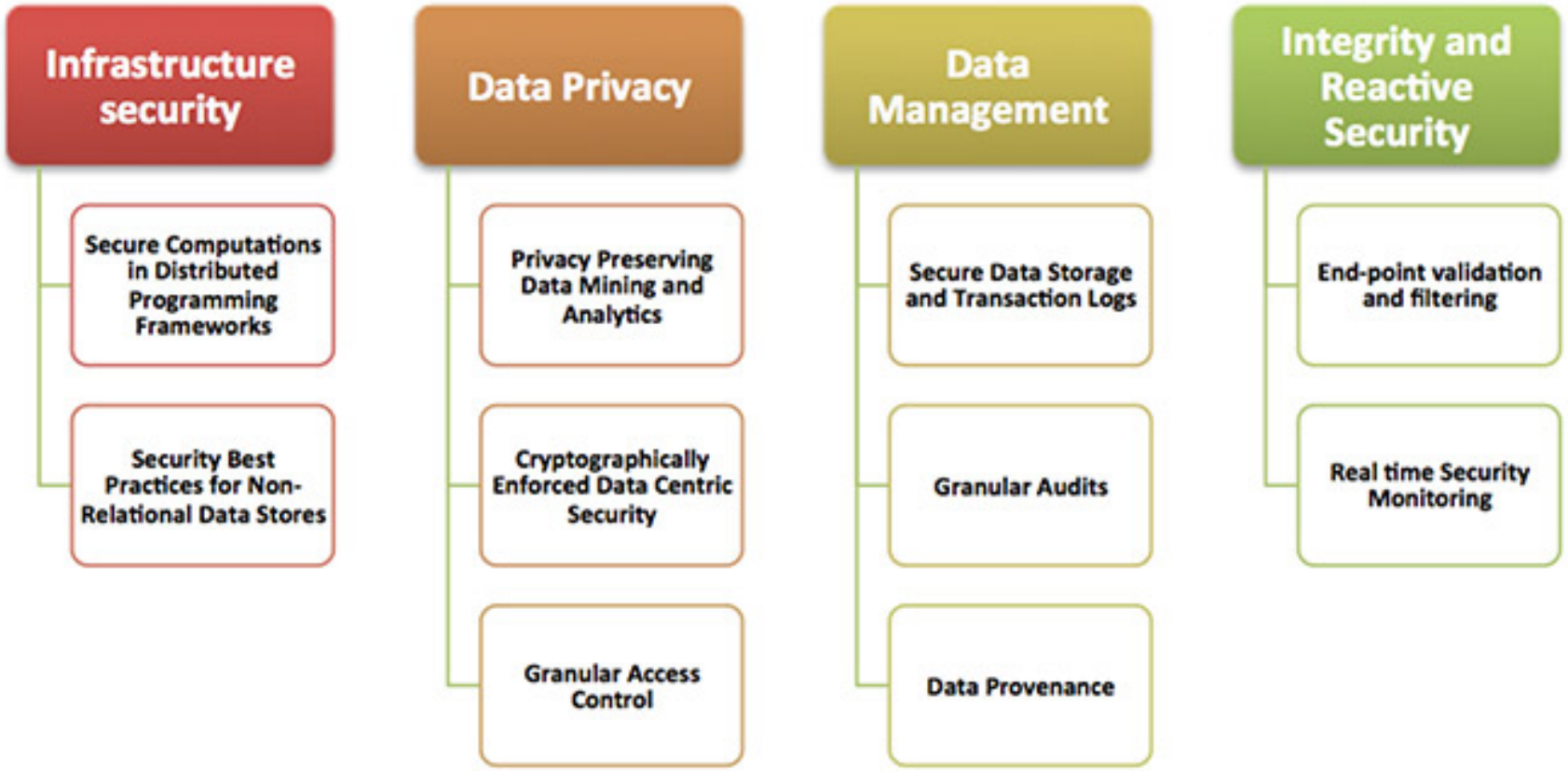


Figure Classification of the Top 10 Challenges

Big Data to Collect

- Logs
- Network traffic
- IT assets
- Senstitive / valuable information
- Vulnerabilities
- Threat intelligence
- Application behaviour
- User behaviour

The Tipping Point

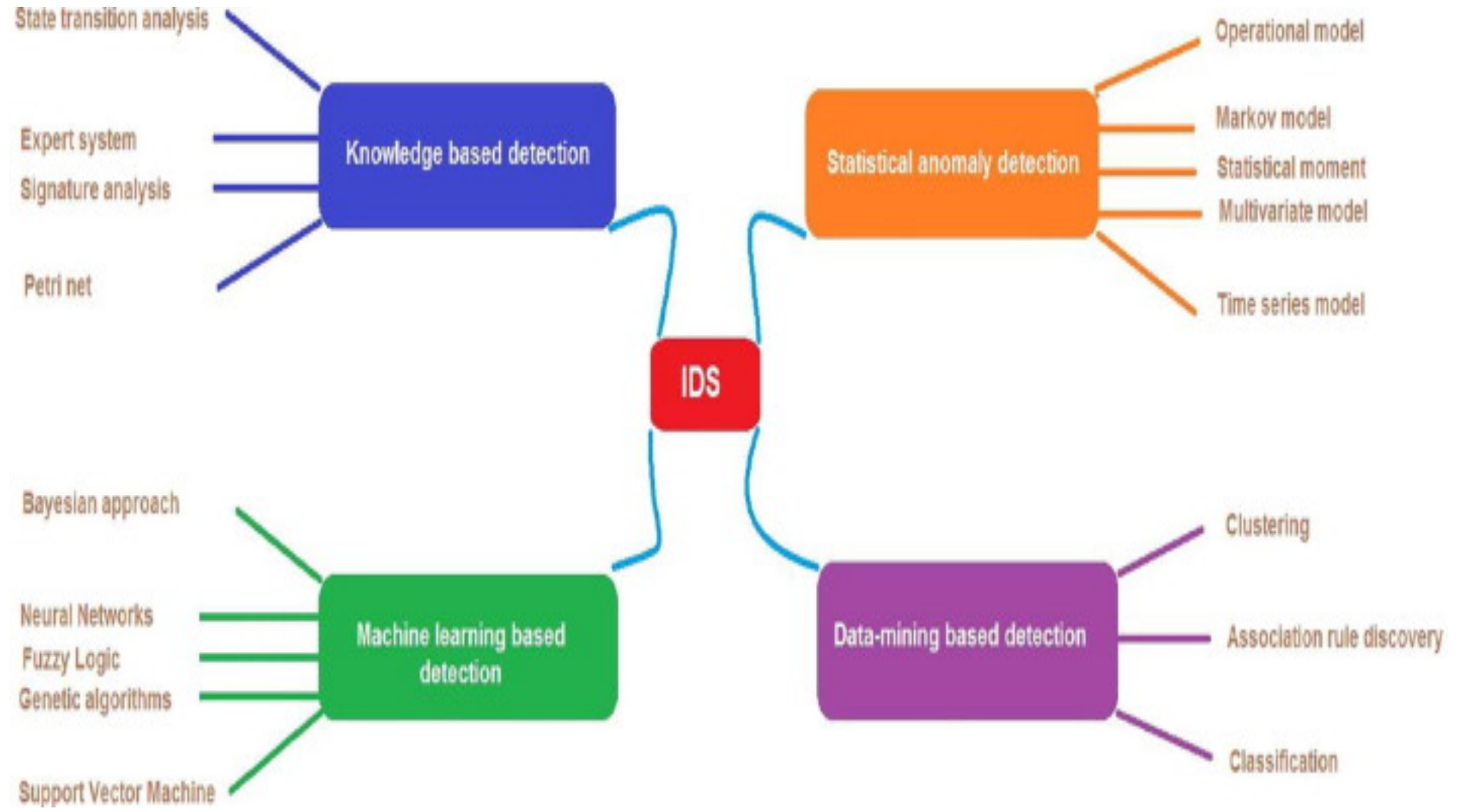
Complex threat landscape:

- Stealth malware
- Targeted attacks
- Social engineering

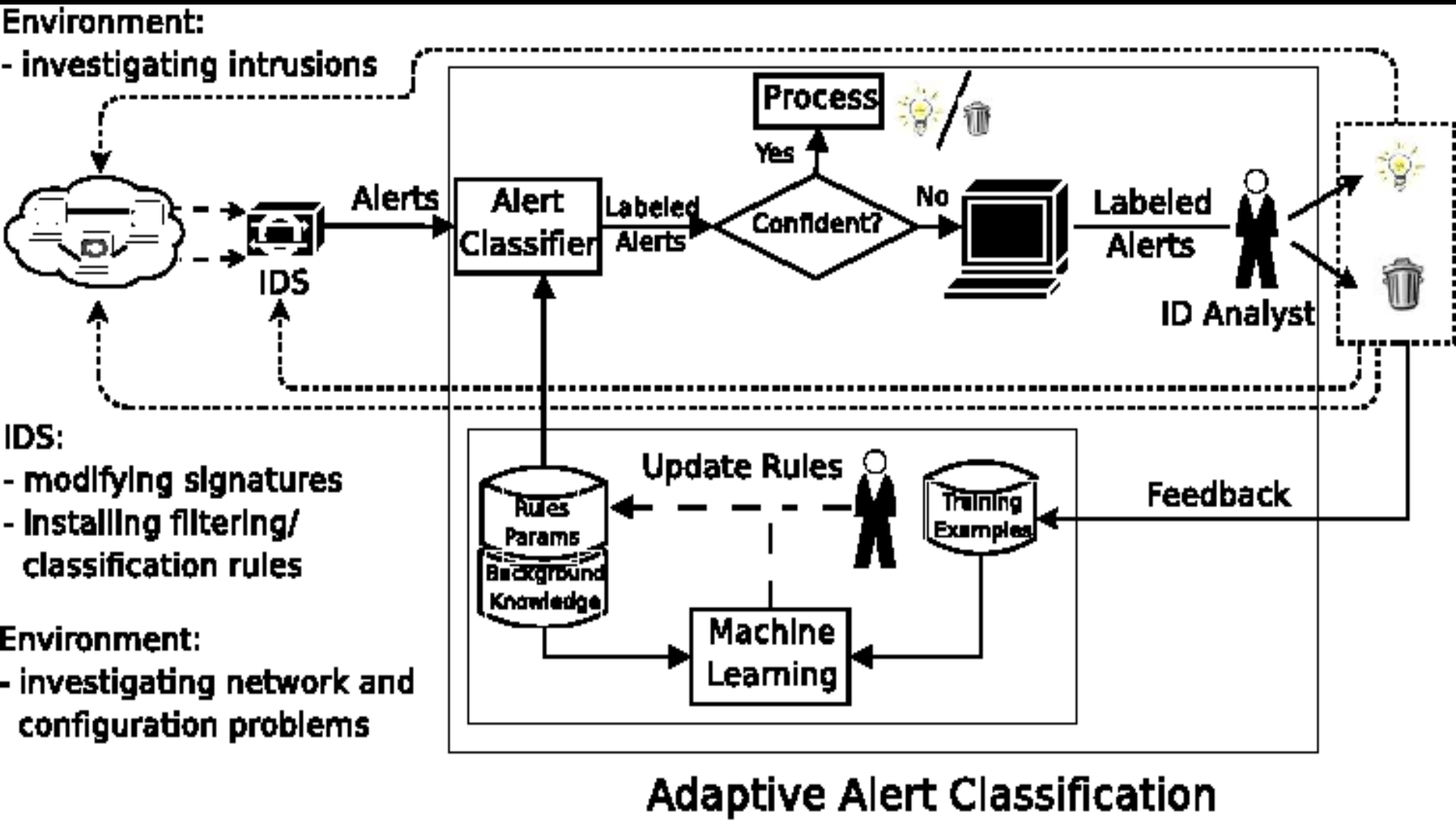
New technologies and challenges:

- Social networking
- Cloud
- BYOD / consumerisation
- Virtualisation

Instrusion Detection System - IDS



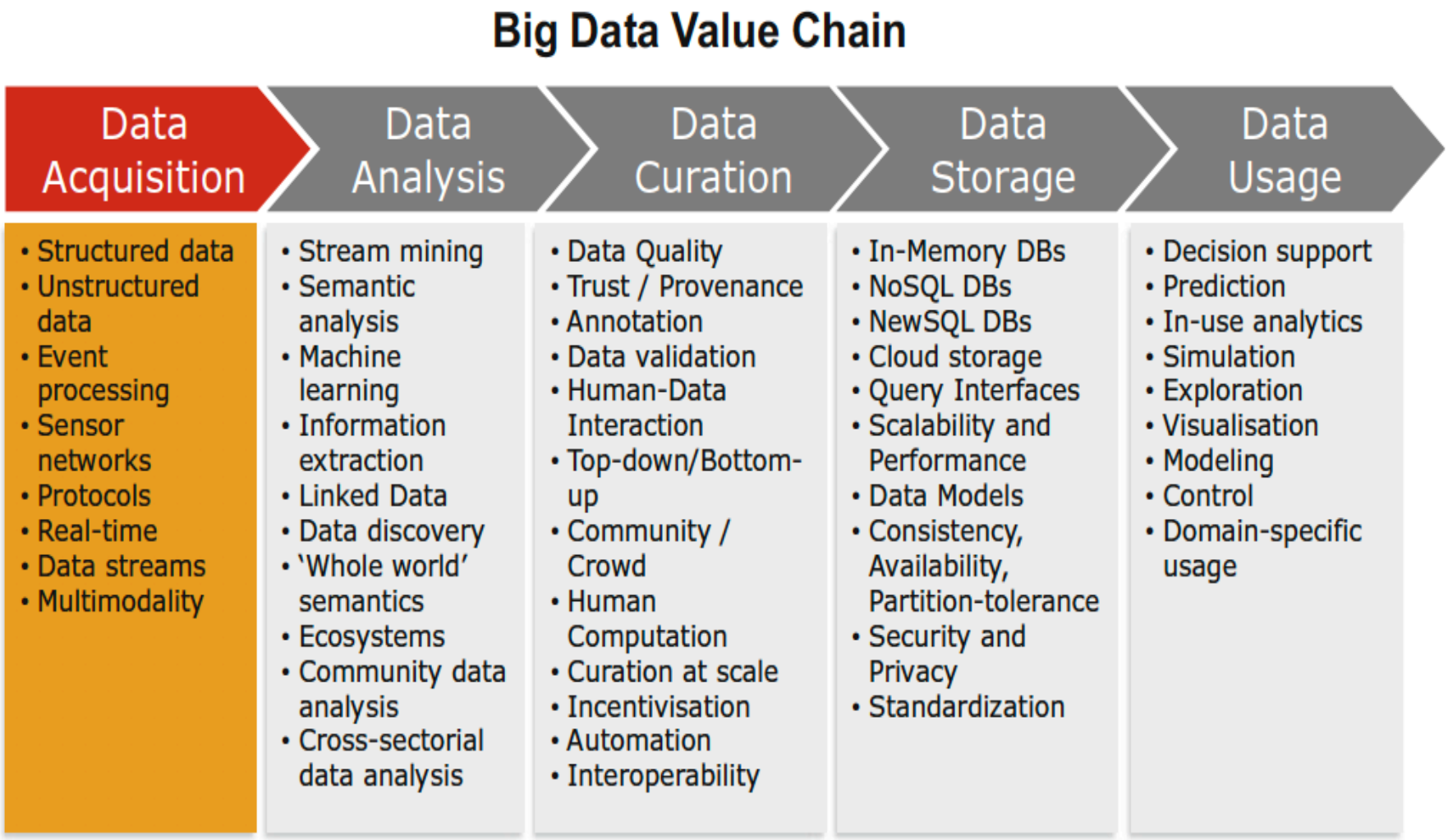
IDS Sample



Adaptive Alert Classification

Forensics Issues for Big Data

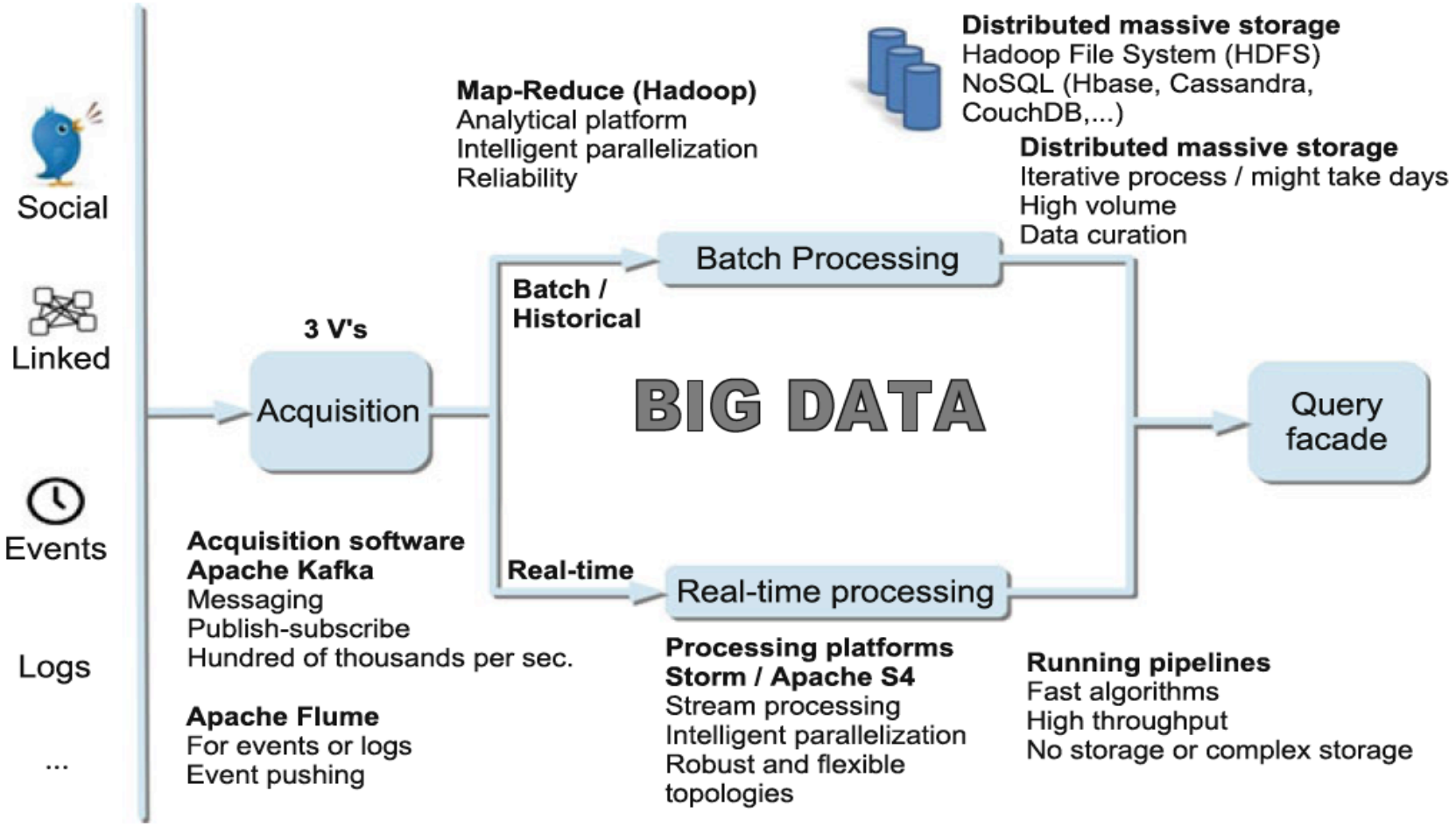
Forensics for Big Data



The Definition

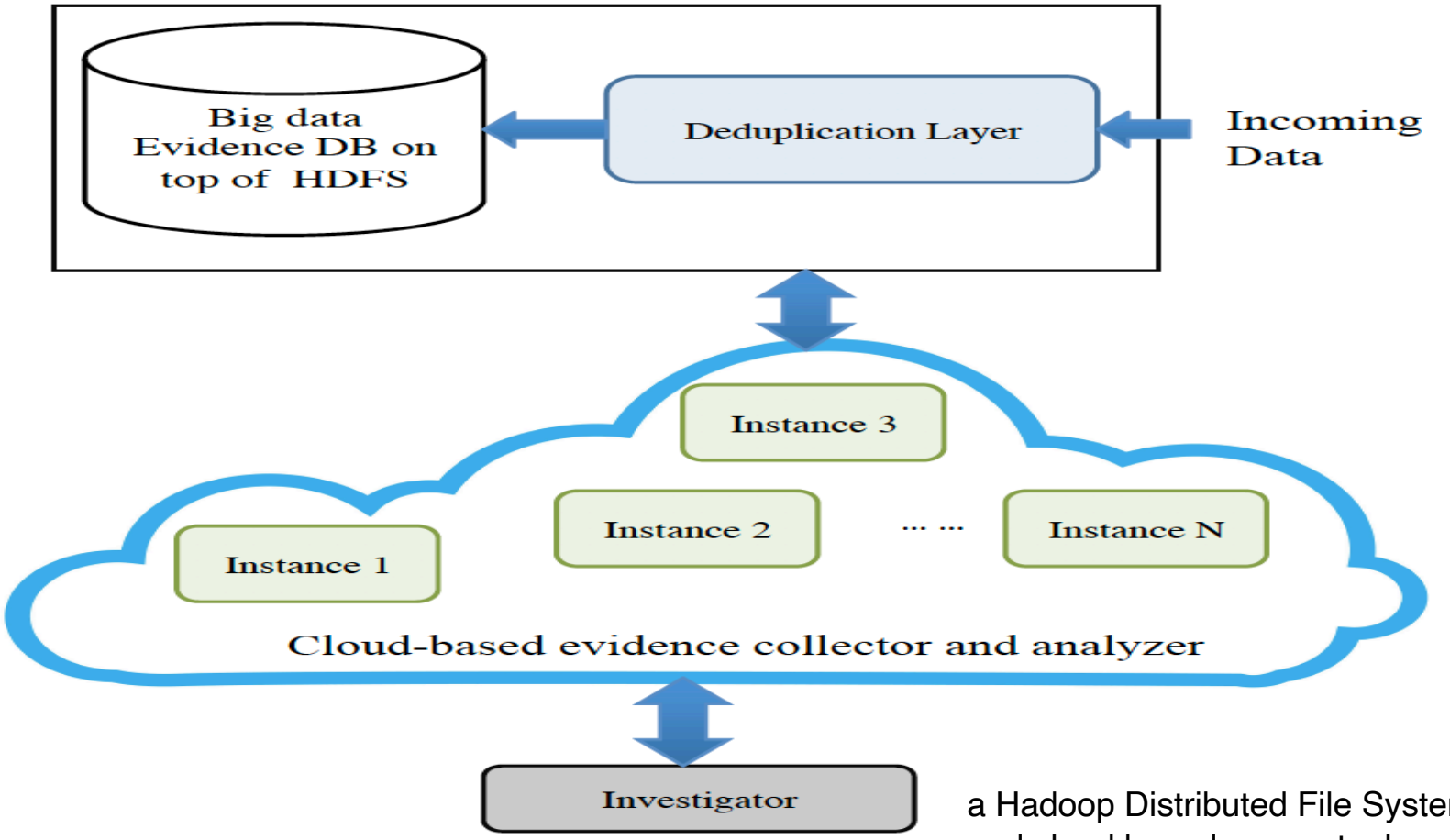
- Big data requires a new generation of technologies and architectures, designed to efficiently extract value from very large volumes of a wide variety of data, by enabling high-velocity capture, discovery, and/or analysis.
- Unfortunately, the traditional tools and technologies of digital forensics are not designed to handle the big data.
- Big Data forensics augments the existing forensics body of knowledge to handle the massive, distributed systems that require different forensic tools and techniques

Acquisition Issue



Sample Model

Digital Forensics in the Age of Big Data: Challenges, Approaches, and Opportunities (Shams Zawoad and Ragib Hasan)



a Hadoop Distributed File System (HDFS) and cloud based conceptual model to support reliable forensics investigation on big data.

Big Data Storage

- According to the International Data Corporation (IDC), every person online will create an average of 1.7 megabytes of new data every second by 2020, and only 37% of all big data could be analyzed, leaving a plethora of untapped information that could be processed by law enforcement agencies to solve crime efficiently.

Unstructured Data

- Unstructured data is information that either does not have a pre-defined data model or cannot be structured in an orderly fashion (such as in ordered rows and columns as found in databases).
- Unstructured data can include text in all forms, emails, video, audio files, web pages and social media. Making sense of unstructured data is often done by implementing complex search queries to extract and present all the data in a better presented structure. The search queries will enable an examiner to find data in with the same contextual structure as the investigation.

Industrial Challange

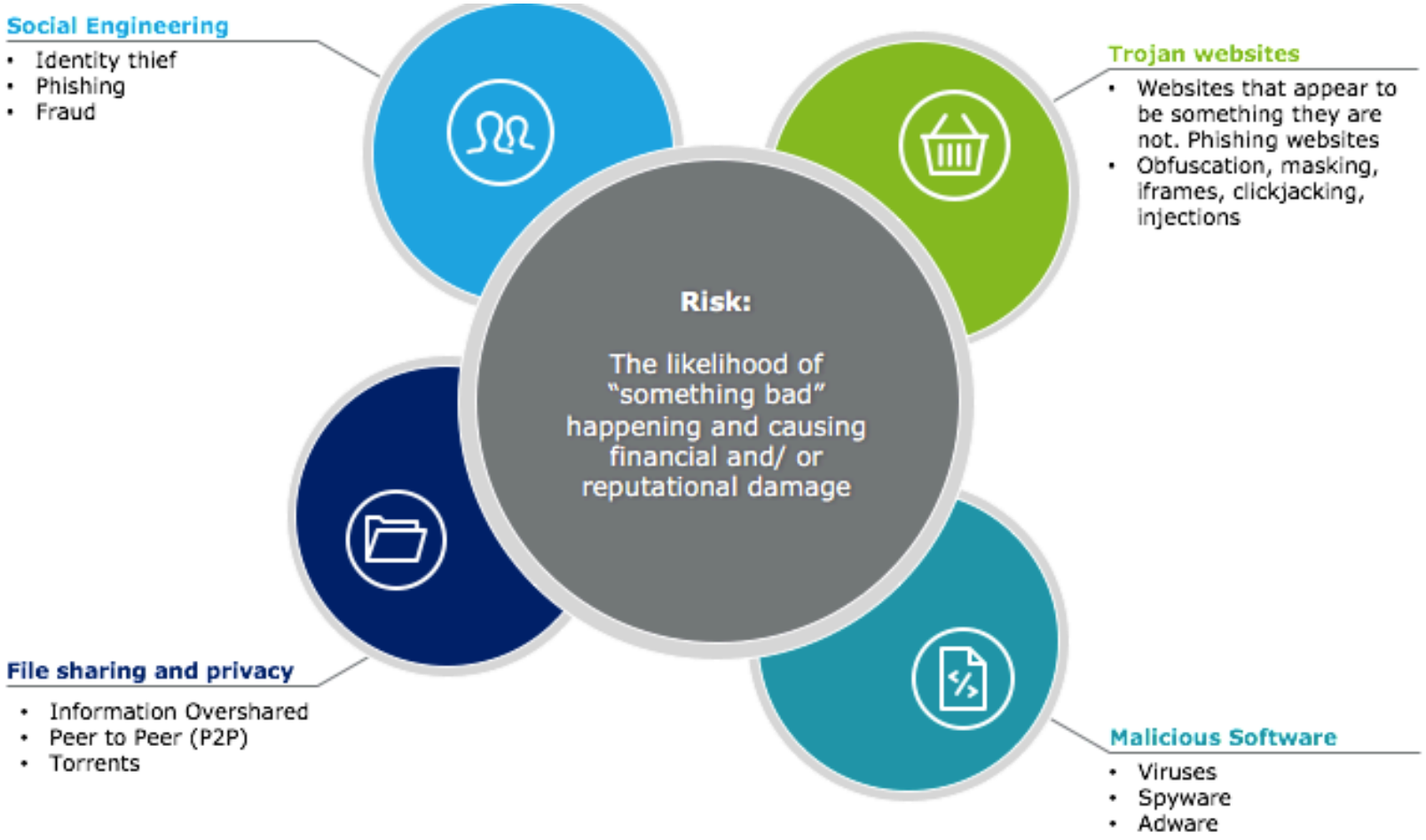
- Security data is growing as organisations collect process, and analyse more than six terabytes of security data monthly (Cybersecurity Analytics and Operations in Transition, http://esg-global.com/, 2017).
- It is very difficult to keep up with the threat landscape as organisations are being overwhelmed by the scaling needs for big data forensics that consider both post-mortem and real-time processing and visualization of evidence.
- Customers need to analyse security event data in real time for internal and external threat management which requires collecting, storing, analysing and reporting on log data for forensics and regulatory compliance, while maintaining the security and integrity of data.

Research Challanges

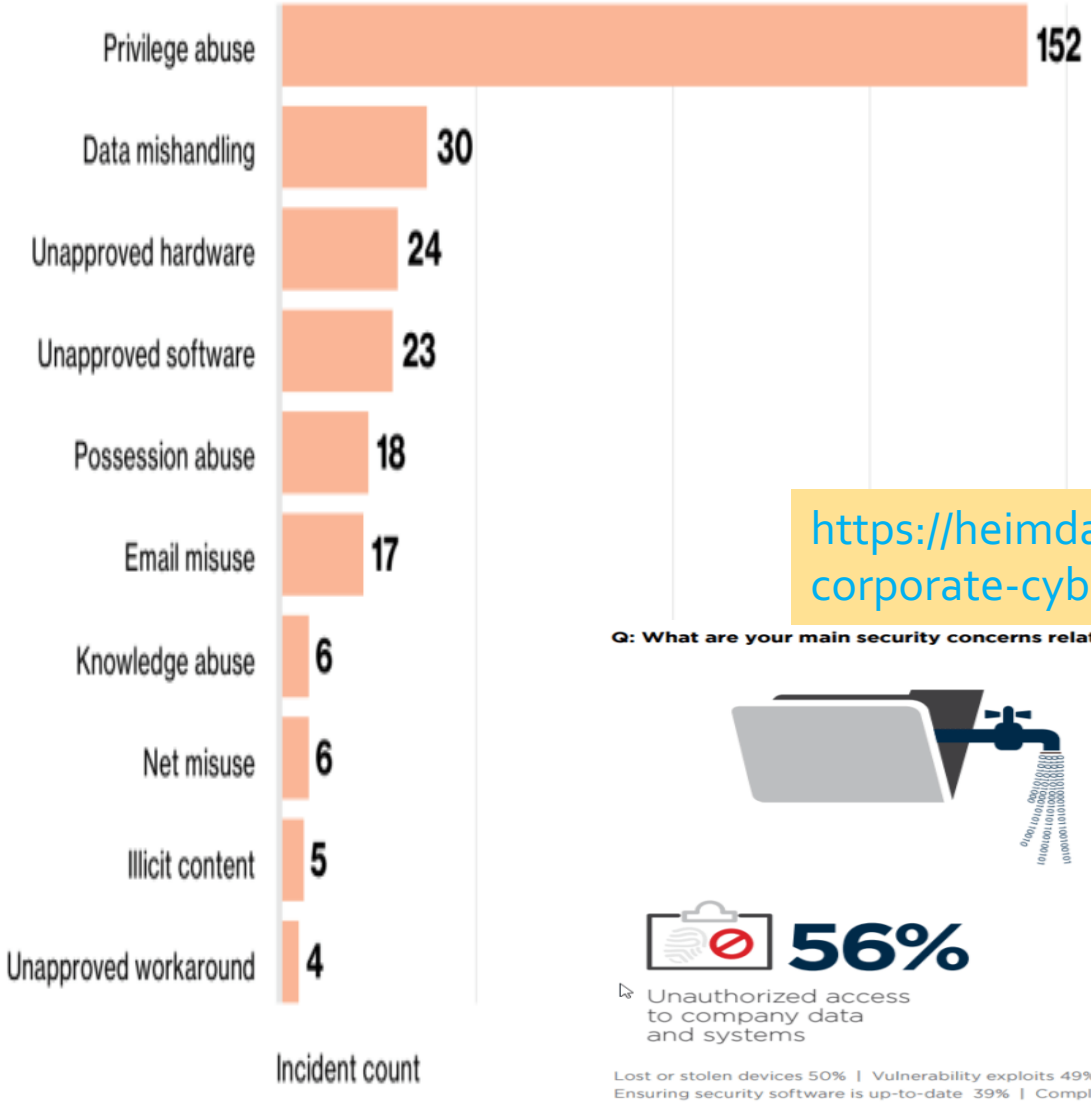
- There is a need for advanced visualization methods to combine data from heterogeneous sources and to guide forensics investigators to identify areas warranting further review.
- Correlation of forensic data collected by disparate cyber-centric security procedures and technologies (Firewalls [FW], Intrusion Detection Systems [IDS], Intrusion Prevention Systems, [IPS], etc.), with device and control systems logging data.
- Better collection of effective data for post-incident security analysis.
- Increase in storage space on hard drives impacts both the performance utilization and the time when carrying out forensics tasks.

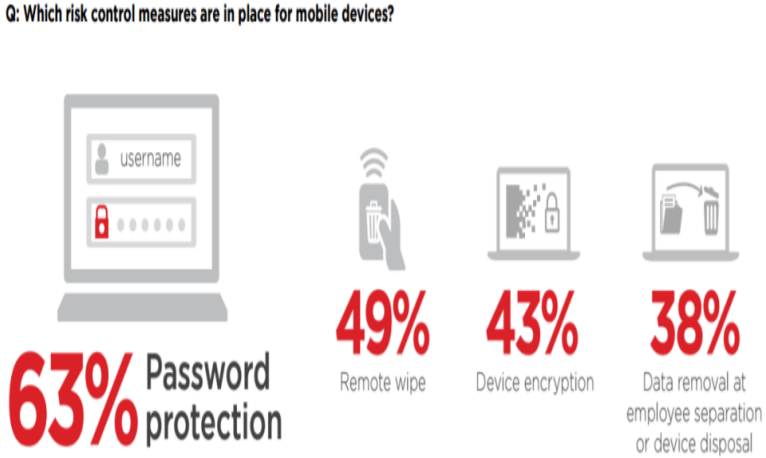
Potential Crime in Big Data

Risk on Internet

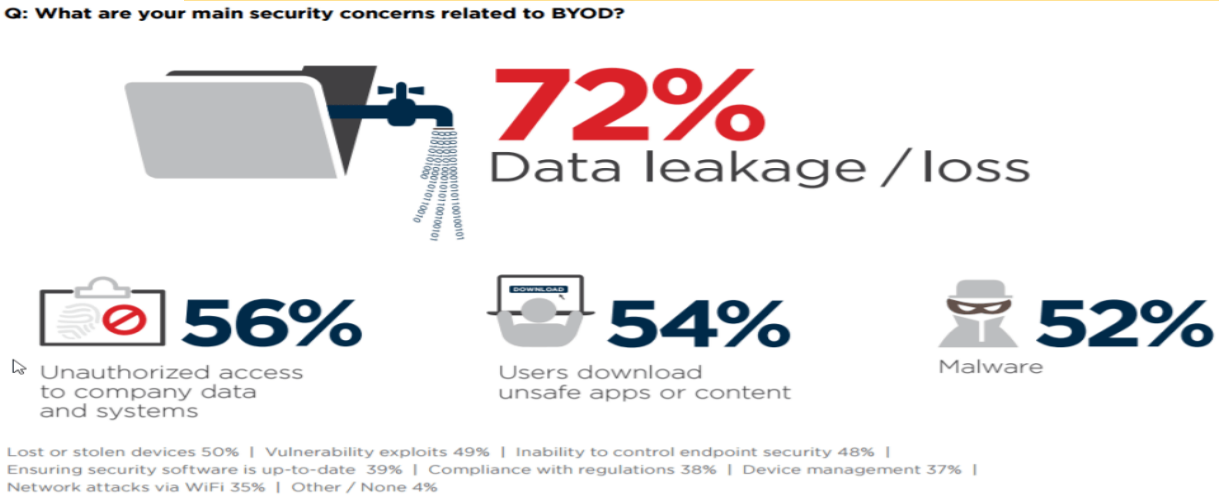


Human : The Weakest Link





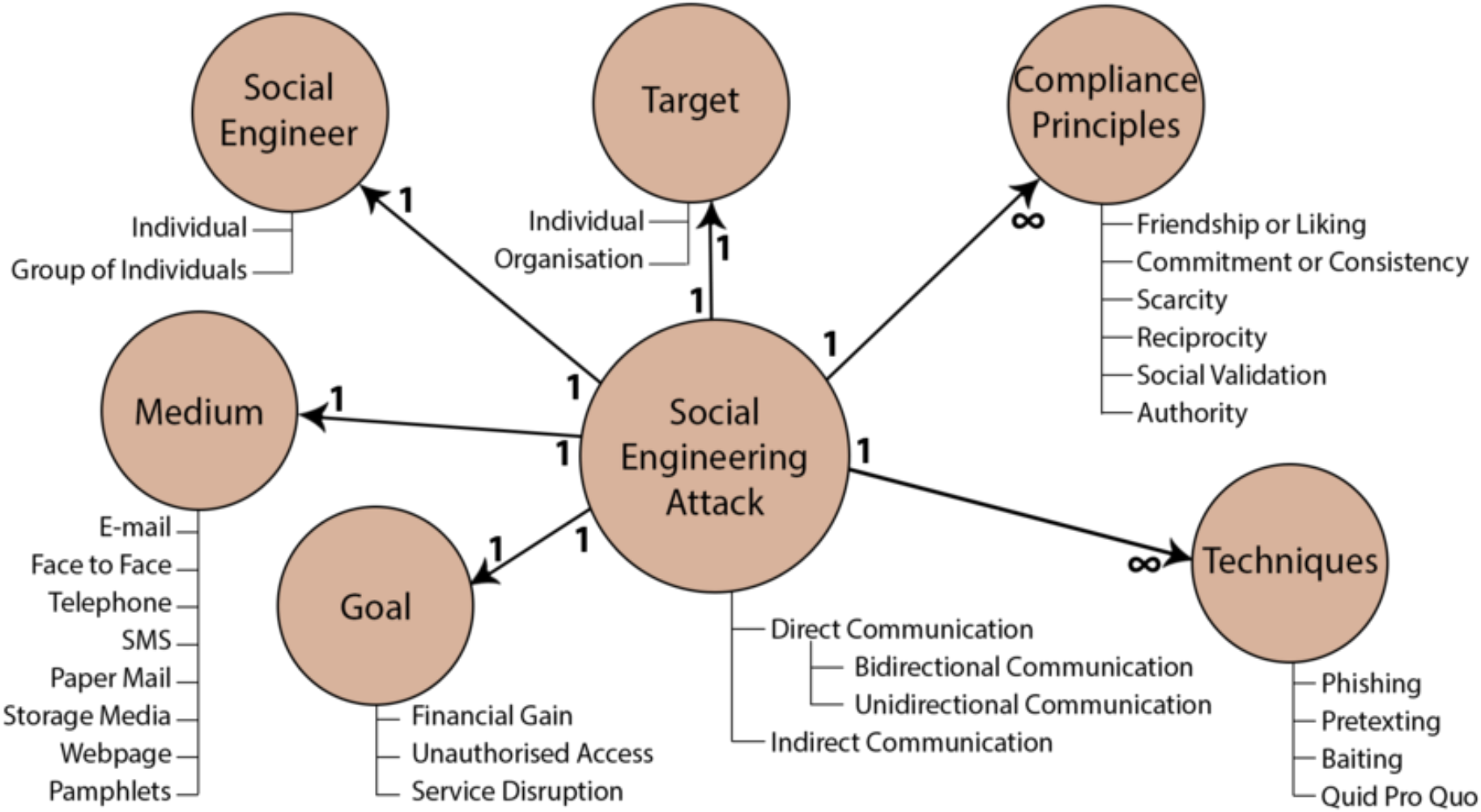
https://heimdalsecurity.com/blog/10-critical-corporate-cyber-security-risks-a-data-driven-list/



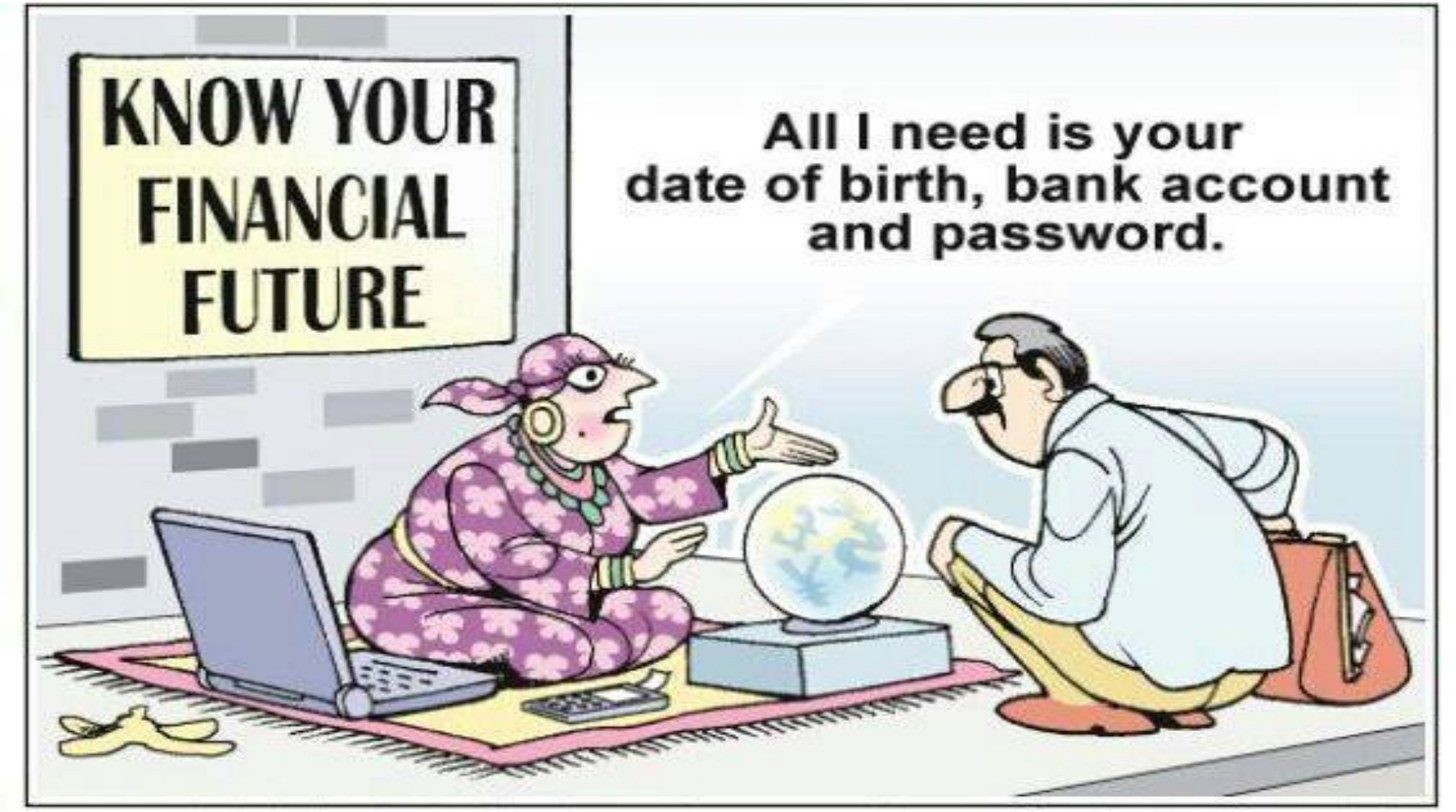
Social Engineering Attack

- Social engineering is a psycho-social attack that subverts human trust and helpfulness in order to attain the attacker's goals.
- Social engineers focus on the users of the system. By gaining the trust of the user, a social engineer can simply ask for whatever information he or she wants…and usually get it.

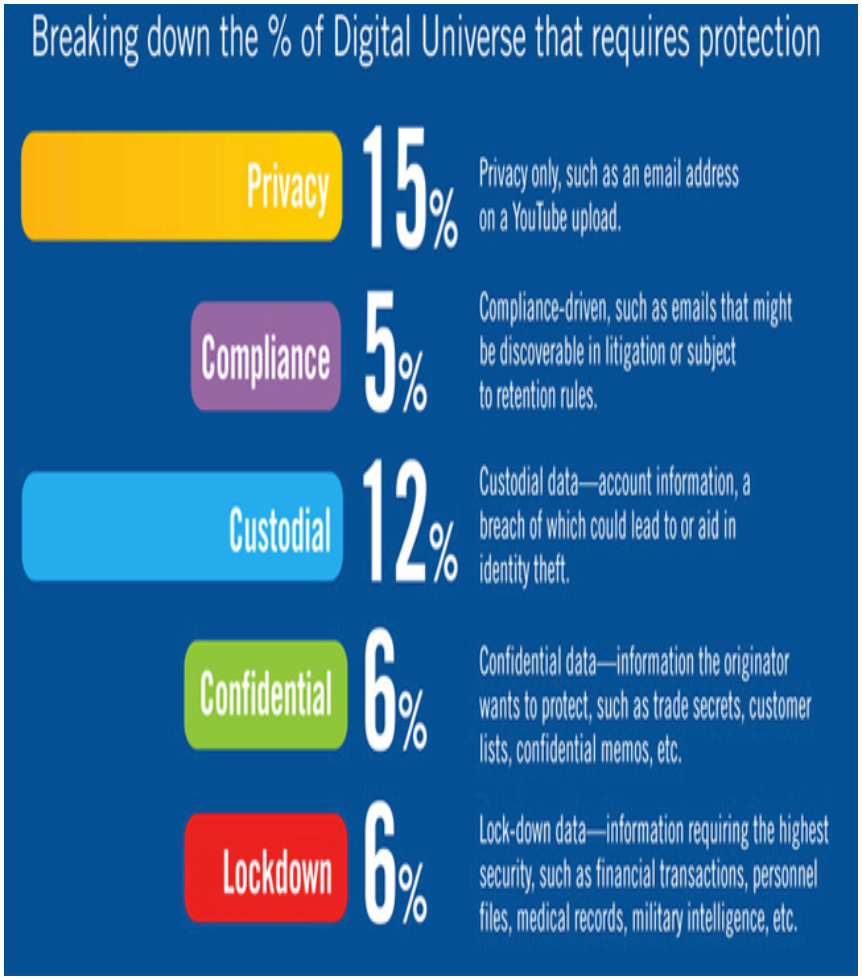
Social Enginerring Attack

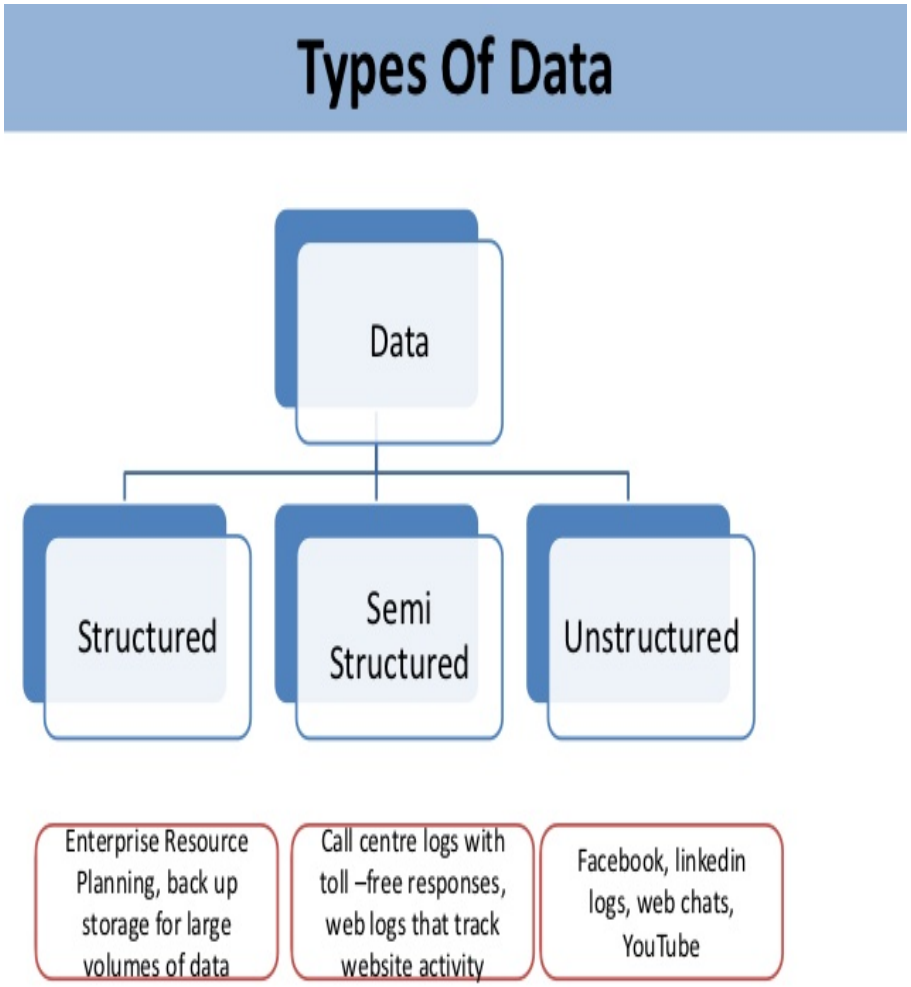


Security Awareness

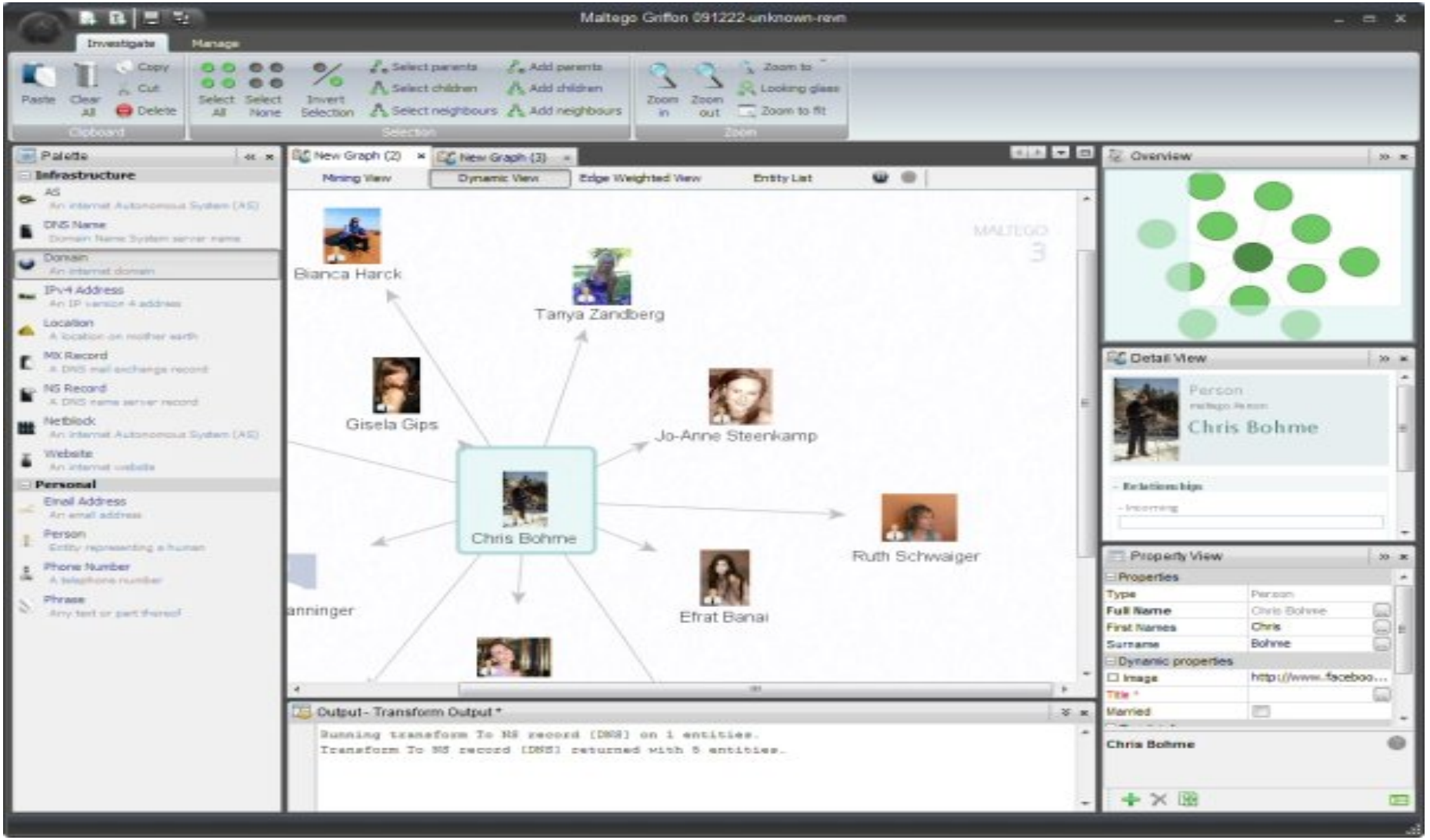


Valuable Data : Privacy





Maltego

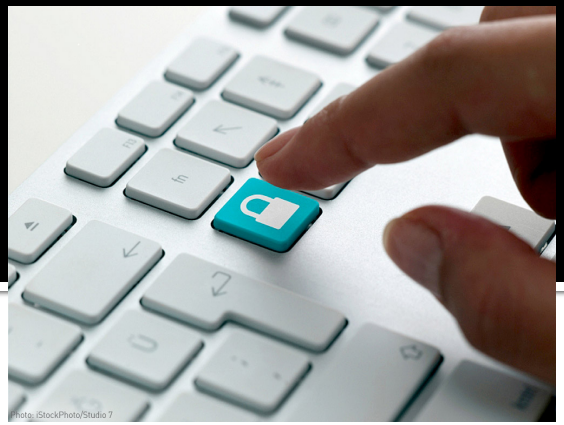

Maltego Griffon 091222-unknown-revn
Investigate
Manage
Palette
Infrastructure
AS
DNS Name
Domain
IPv4 Address
Location
MX Record
NS Record
Netblock
Website
Personal
Email Address
Person
Phone Number
Phrase
New Graph (2)
New Graph (3)
Mining View
Dynamic View
Edge Weighted View
Entity List
Bianca Harck
Tanya Zandberg
Gisela Gips
Jo-Anne Steenkamp
Chris Bohme
Ruth Schwaiger
Efrat Banai
Overview
Detail View
Person
Chris Bohme
Property View
Properties
Type
Full Name
First Names
Chris
Surname
Bohme
Dynamic properties
Image
Married
Output - Transform Output *
Running transform To NS record [DNS] on 1 entities.
Transform To NS record [DNS] returned with 5 entities.

Profiling

Not everything is visible!

Remember that the Internet is an extension of our "real" life

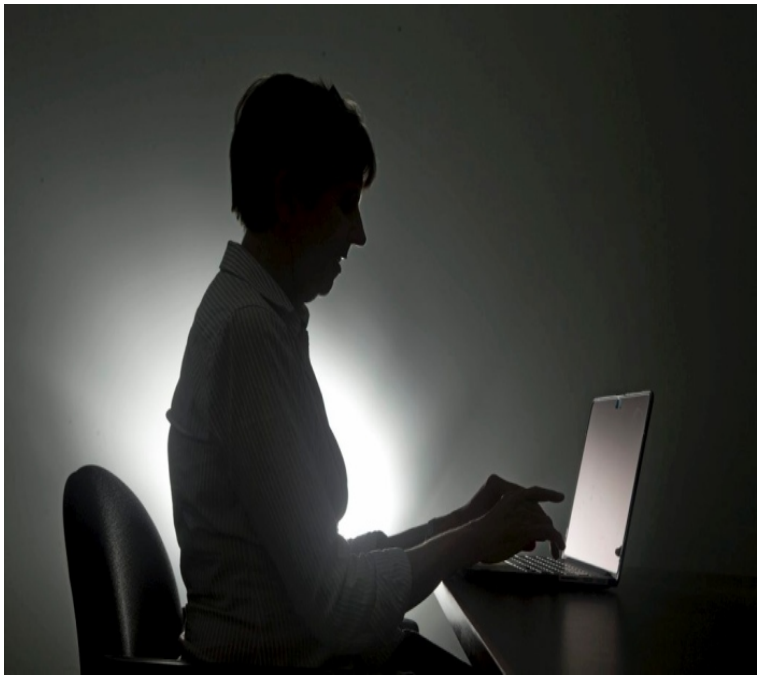
Privasi



- Kemampuan seseorang untuk mengatur informasi mengenai dirinya sendiri.
- Hak dari masing-masing individu untuk menentukan sendiri kapan, bagaimana, dan untuk apa penggunaan informasi mengenai mereka dalam hal berhubungan dengan individu lain.

Ancaman Privasi

- Pengumpulan data (*data collection*)
  - Pengumpulan data yang lebih mudah dan lebih cepat
  - Acuan silang (*cross referencing / aggregation*)
  - Pengumpulan data tersembunyi (*hidden data collection*)
- Pelacakan penggunaan (*usage tracking*)
- Pembagian informasi (*information sharing*)

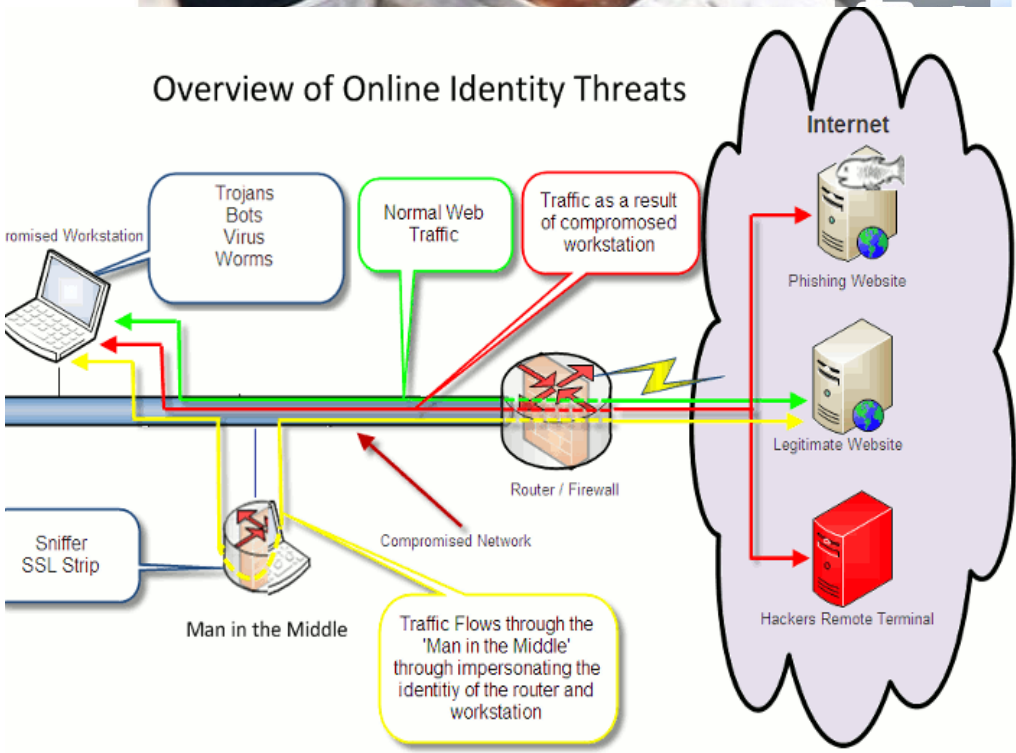


- Political trolls 'win arguments' by publishing your personal data
- Genealogy sites have already posted your personal information online
- Mobile apps send personal data back to a remote server
  - A Chinese selfie-editing iPhone app called Meitu transforms your face into a surreal cartoon image that whitens, brightens, enlarges the eyes and adds visual effects.

MailOnline News

David Beckham hires top cyber-security experts to find hackers behind embarrassing email leak

By Rory Tingle For Mailonline
23:42 GMT 06 Feb 2017, updated 02:12 GMT 07 Feb 2017



Cyber Self Defense

- **You cannot protect all your data**
- **You cannot stop every attack**
  - Reduce your attack surface
  - Segregrate and protect your critical data
  - Establish access norms and monitor for anomalies
  - When you are attacked, report it. Transparency = Security

Ada yang kenal ?

This Person Does Not Exist

https://thispersondoesnotexist.com

Imagined by a GAN (generative adversarial network)
StyleGAN2 (Dec 2019) - Karras et al. and Nvidia
Don't panic. Learn how it works [1] [2] [3]
Help this AI continue to dream | Contact me
Code for training your own [original] [simple]
Art • Cats • Horses • Molecules | News | Friends | Office
Another | Save

How About Fingerprint ?

- Fingerprints can b**e stolen from selfies**
  - Researchers at Japan's National Institute of Informatics (NII) announced recently that your fingerprints could be stolen from photos of your fingers, and the prints could then be re-created and used to bypass biometric security systems.

http://www.computerworld.com/article/3165397/security/5-shocking-new-trends-threaten-your-personal-data.html

Thank You

GREVZN XNFVU

http://www.decode.org/?q=GREVZN+XNFVU+

http://forensics.uii.ac.id