RSAConference2015

Singapore | 22-24 July | Marina Bay Sands

CHANGE

Challenge today's security thinking

SESSION ID: CLE-F03

Achieving Cyber Identity Resolution via Electronic Warfare Techniques

Dr. Nitzan Barkay & Elana Dror-Rein

Engineering Deputy Director, Research & Technology
IAI – Israel Aerospace Industries

Identity Resolution

Who is who? What does each one do?

In the **physical** world

In the world of **virtual** entities

Cyber Entity / Identity Resolution

- **Entity resolution** provides a measure to the similarity between **virtual** entities
  - Association of related virtual entities (same origin)
  - Differentiation of unrelated ones
- **Identity resolution** uses any "solid" identifier of an associated entity (e.g., phone # or Facebook ID) to correlate to **real** identities

Who May Benefit from Identity Resolution?

- **Intelligence / investigation centers looking for a person / group**
  - Enriching the information with all possible appearances & aspects
  - Revealing bogus identities
  - Classification through analysis of the virtual entity features
  - Identification of groups and networks
- **Situation awareness centers for defense & early warning**
  - Prediction of evolving events (in the Cyber world or the physical world)
  - Enhancing the information about a virtual actor, particularly a cyber attacker
    - What is the target – support actionable early warning
    - What is the origin – help attribution of the attacker and possible deterrence
      - Attacker (physical) ID – Identity resolution
      - Attacker location – Geo-location resolution

RSAConference2015

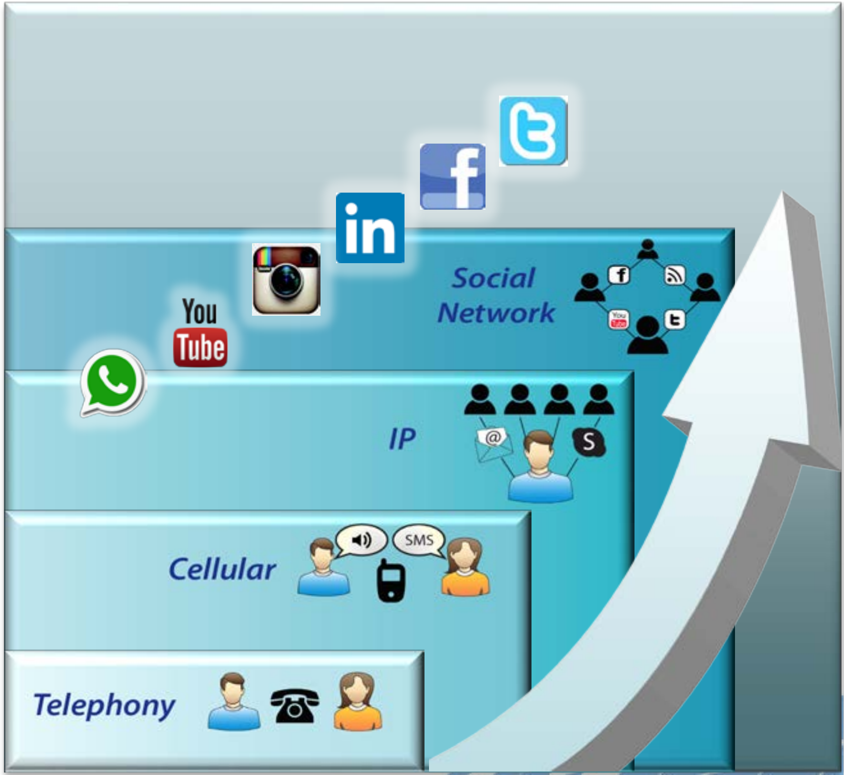
Singapore | 22-24 July | Marina Bay Sands

#RSAC

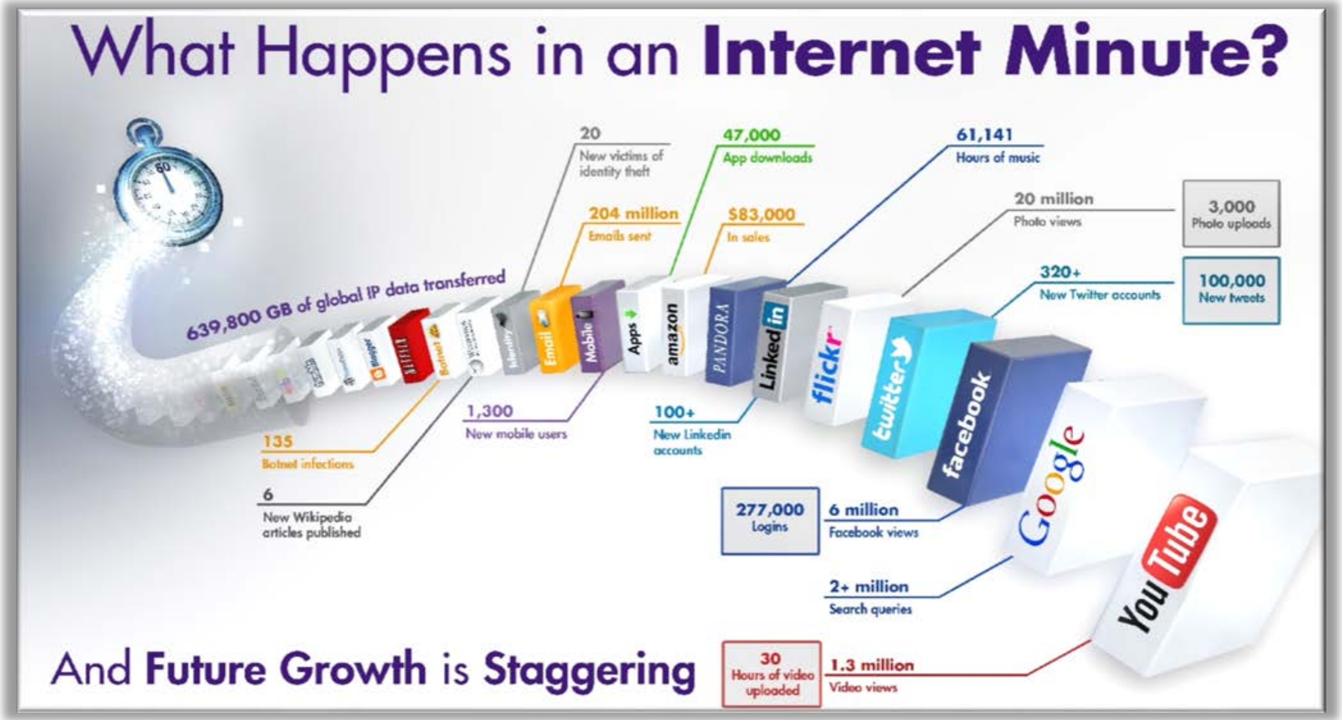
Identity Resolution Challenges

Sources of Data

- **Evolution of communication & information**
  - From telephony to social networks
  - From voice to messages, e-mails, blogs & video
- **Huge amounts of data are available publicly** – WEBINT & OSINT
- **More is available to Law-enforcement agencies**
  - Through the communications and internet providers (ISP)
  - Using passive & active accessibility tools
- Raw data is enormous & unsorted
  - Usually partial or ambiguous
  - May be misleading, even deliberately – impersonation or just "inaccurate" details



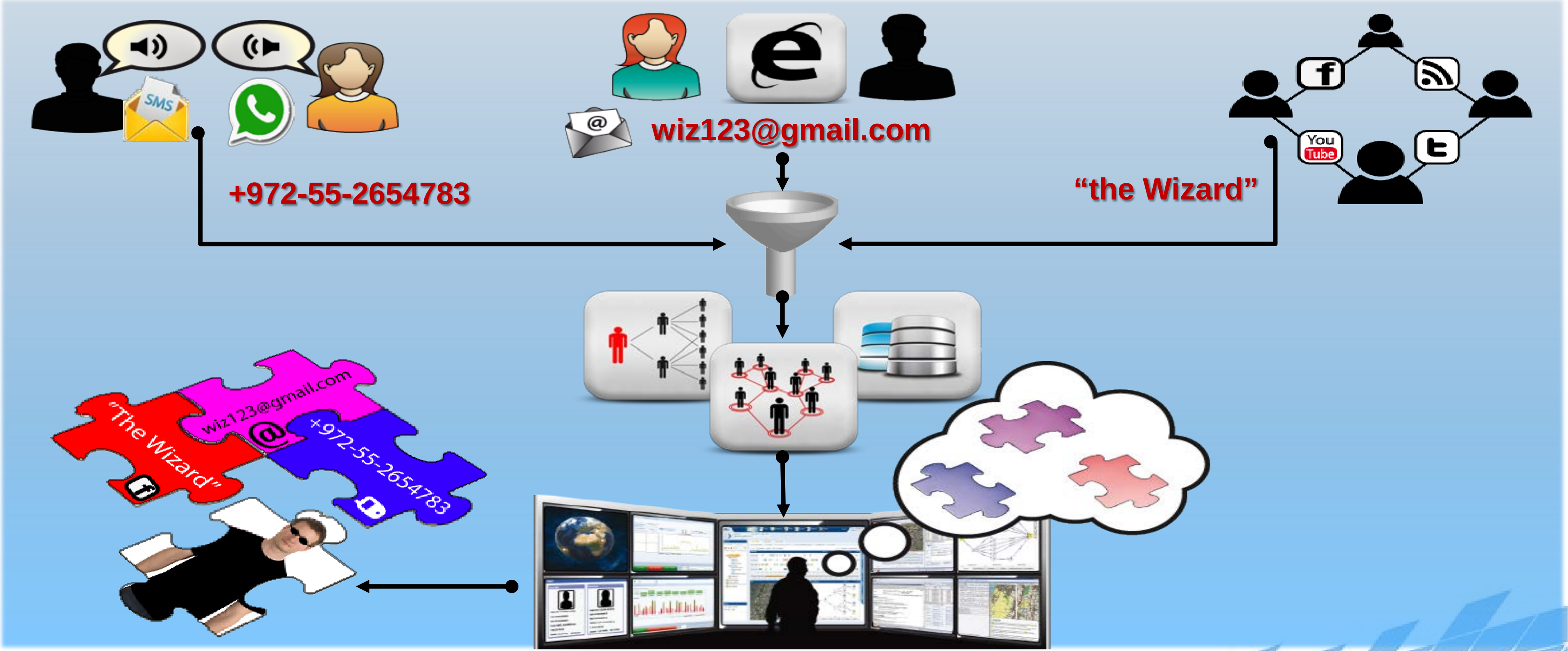
Identity Resolution Challenges – Massive Data Flow



How can one extract the required insight?

(*) http://dailyinfographic.com/what-happens-in-an-internet-minute-infographic

Identity Resolution Challenges – Multiple Aspects



Identity Resolution Challenges – Association

- Differentiating

MACCABI TEL-AVIV
מכבי
תל-אביב

≠

מכבי אלקטרה תל-אביב

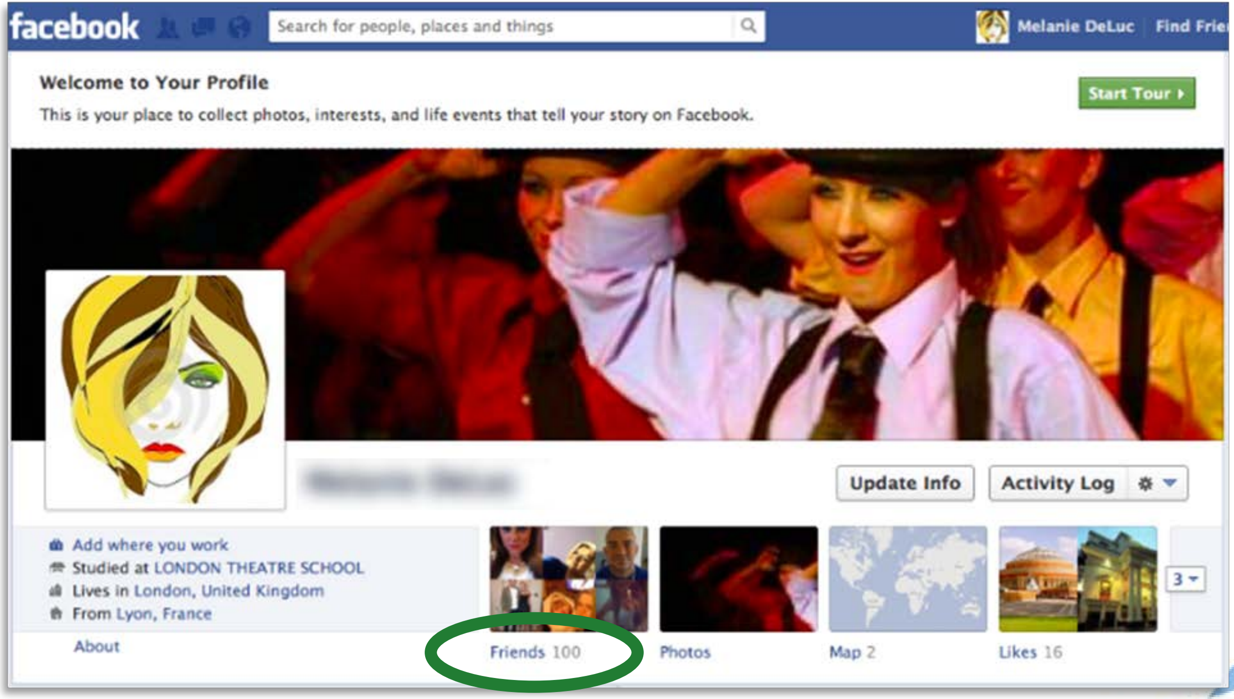
Maccabi Tel-Aviv Maccabi Tel-Aviv

- Combining

Sting = Gordon Matthew Thomas Sumner

Identity Resolution Challenges – Fake Virtual Identities

How easy is it? **100 friends @ 48 hours from launch**



Welcome to Your Profile

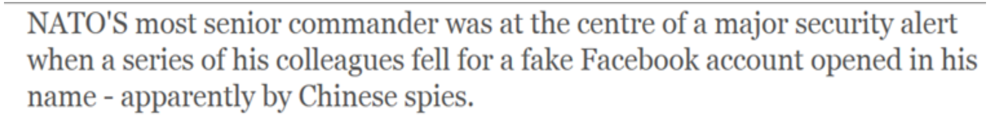
This is your place to collect photos, interests, and life events that tell your story on Facebook.

[redacted]

- Add where you work
- Studied at LONDON THEATRE SCHOOL
- Lives in London, United Kingdom
- From Lyon, France

Identity Resolution Challenges – Bogus Identities

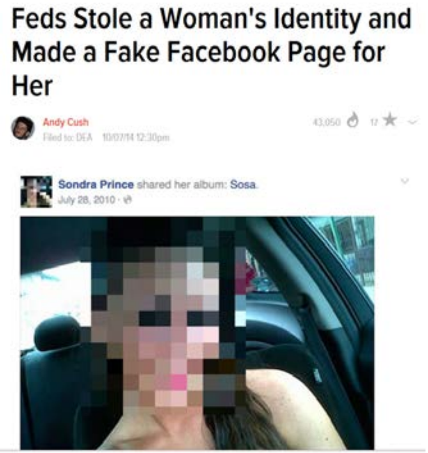
The challenge is growing: Bogus identities are common



NATO'S most senior commander was at the centre of a major security alert when a series of his colleagues fell for a fake Facebook account opened in his name - apparently by Chinese spies.

www.telegraph.co.uk/technology/9136029/How-spies-used-Facebook-to-steal-Nato-chiefs-details.html

2012: Bogus Facebook account created for NATO senior commander



Feds Stole a Woman's Identity and Made a Fake Facebook Page for Her

Andy Cush
Filed to: DEA 10/07/14 12:30pm

43,050 17

Sondra Prince shared her album: Sosa.
July 28, 2010 ·

http://gawker.com/feds-stole-a-womans-identity-and-made-a-fake-facebook-p-1643348368

2014: FBI makes bogus Facebook account in an attempt to capture offenders

Data Analysis & Identity Resolution Challenges

What Happens in an Internet Minute?
And Future Growth is Staggering

- **Huge amount of data**
  - Data availability, especially in real time
    - Technical & regulatory difficulty to maintain effective coverage
  - Data diversity, Data dynamics
- **Assorted information sources**
  - Different aspects of the same identity (e.g., a phone # & Facebook ID)
  - Multiple virtual identities (incl. bogus ones) to the same physical entity
- **Insight & discrimination**
  - Derive insight from the mass of data – identification based on the aggregated picture
  - Discrimination between legitimate activity and malicious acts – Eliminating false alarms
- **Identification**
  - **Attribution to actual actors**

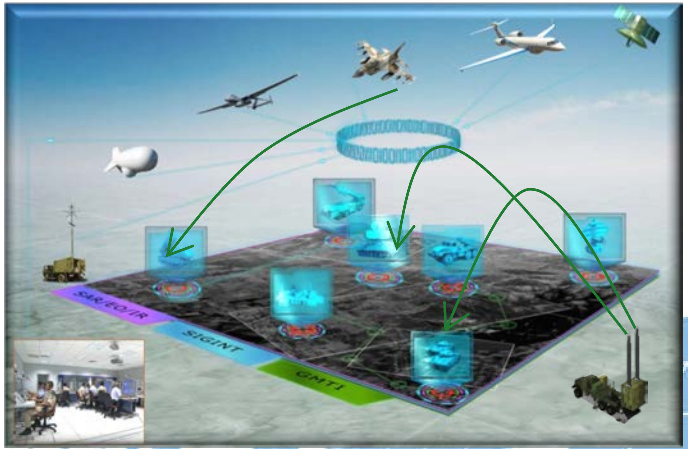
(Physical) Persistent Surveillance Challenges

- A multitude of entities, of various types
- Dynamic scenario
- Integration of different sensors
  - Each interprets the situation picture in its manner
  - Some get only a partial situation picture; Some overlap
- Discrimination between “innocent” entities (false) and “malicious” targets (real threats)
  - Threats attempt to avoid interception by hiding or behaving like legitimate entities

quantity, variability, dynamics

integration

discrimination



Cyber intelligence challenges are similar;

Solutions can be similar, too…

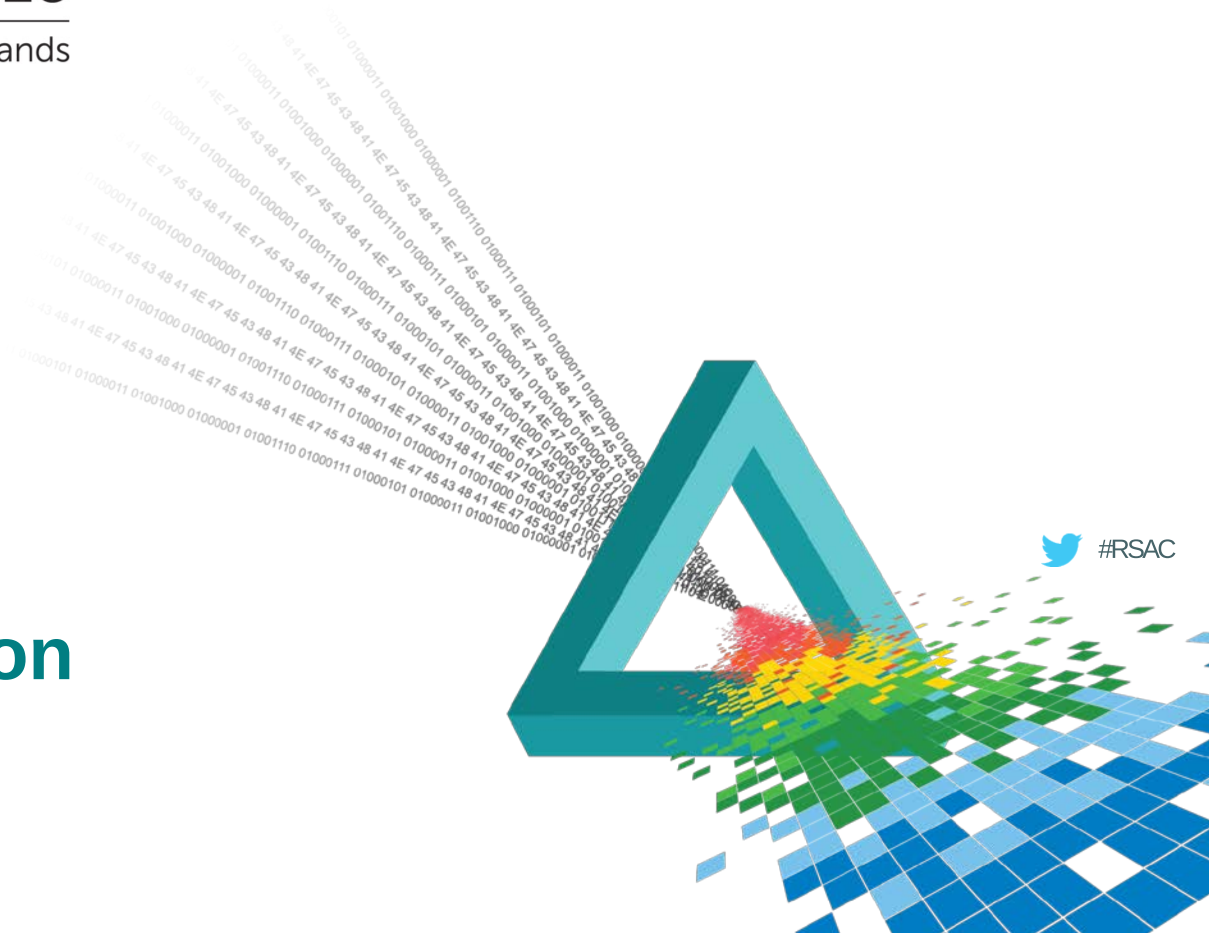
Electronic Warfare vs. Cyber Warfare (I) –
Data Analysis Flow

PROCESS	EW SIGINT	CYBER
Interception	◆ Receiving Electromagnetic Signals (Radar, comm.) ◆ Measuring electronic parameters	
Geo-location	◆ Correlating signals from sensors ◆ Location estimation	
Association	◆ Signals tracking in time	
Classification	◆ Classification based on signal type	
Quality Measure	◆ Quality of the information & uncertainty estimation	
Multiple Hypothesis	◆ Scoring of hypotheses & online management ◆ Removing false alarms	
Report	◆ Integration into Intelligence Center ◆ Supporting Situation Awareness & Early Warning	

Similar to EW ?

RSAConference2015

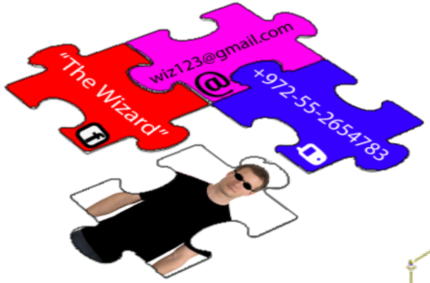
Singapore | 22-24 July | Marina Bay Sands

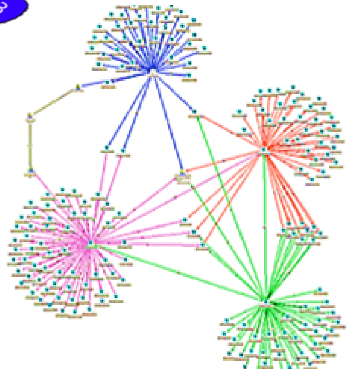


Solving Cyber Identity Resolution

#RSAC

Attributes of Virtual Entities





- **Profile fields**
  - Name, e-mail address, company, etc.
- **Environment-related**
  - Equipment, operating-system, software
- **Geographic/time information**
  - IP address, location
- **Links & friends**
- **Posts & messages**: content, time
- **Behavior-related derived attributes**, e.g.,
  - Active times
  - Slang usage, #words/message
  - Unusual patterns, e.g. writing style

http://scratchbook.ch/wp-content/uploads/2011/01/text-fist-andrew-mason.jpg

Names Comparison for Entity Resolution

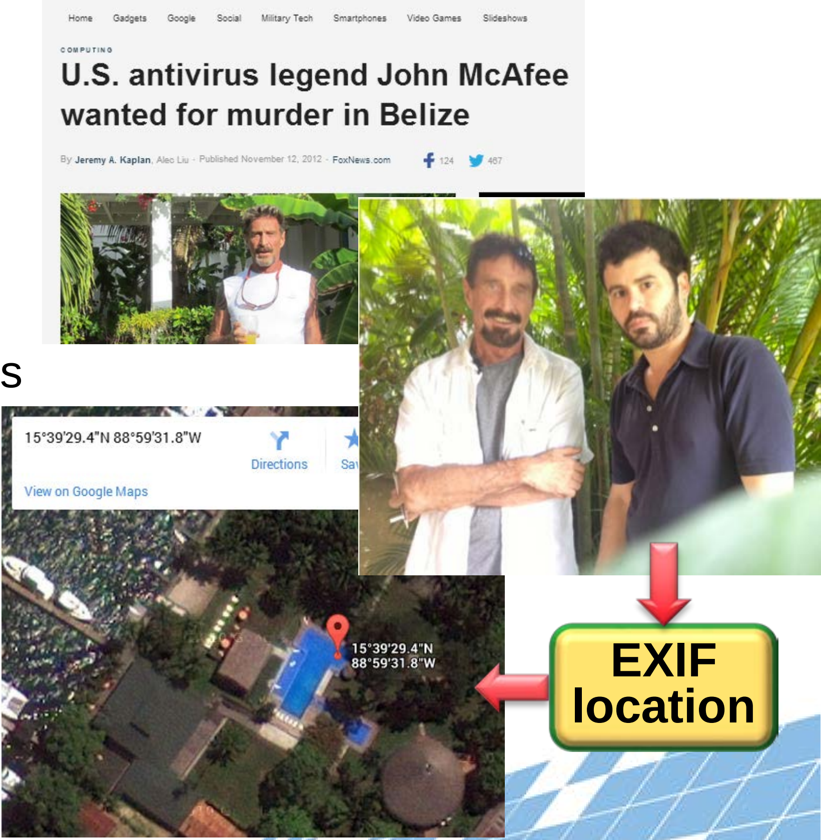
- **Syntactic techniques**
  - Approximate String Matching (ASM) is based on the similarity of two strings in terms of shared characters and character sequences (syntax)
  - Many techniques, e.g.,
    - Levenshtein Edit Distance, SOUNDEX (& variations), Jaro, Winkler (modification of Jaro), n-grams, Lcs (Longest common substring)
  - *Example: "KELLEY" and "KELLY" differ by 1 char*
- **Semantic techniques**
  - Alias Matching is based on the similarity of two strings in terms of their meaning (semantics)
  - *Example: "ED" and "EDWARD" differ by 4 chars, but one is a nickname for the other*

Geo-location for Identity Resolution

- GEO-LOCATION-based differentiation & association of entities
  - It is the standard procedure for physical entities & Electronic Warfare
  - Not obvious for virtual entities
- Methods to derive Geo-location
  - **IP address** geo-location employs available IP databases
    - Widely used for commercial purposes (web localization, marketing)
    - Accuracy is rough (country/region); Easily deceived using proxies & spoofing
  - More complex methods, e.g. **Traffic trace-back**
    - Require accessibility to the network
    - Can be deceived as well
  - Communications **Physical device** geo-location
    - Especially for mobile devices utilizing cellular or WIFI networks
    - Inherent & currently common **Synergy between SIGINT & Cyber**

Geo-location for Identity Resolution – Indirect

- **Contents analysis methods to derive geo-location indirectly**
  - User self-provided location – "**check-in**"
  - **Metadata**
    - Intentional tagging of pictures and other objects
    - Automatic metadata embedded in objects
  - Reports through certain **applications** (e.g., navigation)
  - **Text analysis** to infer the position of a virtual entity
  - **Media analysis** (images, video) for location identifiers
  - **Fine analysis** for origin clues



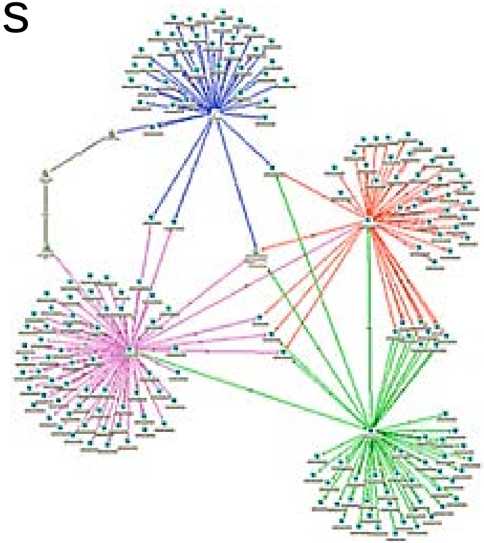
COMPUTING

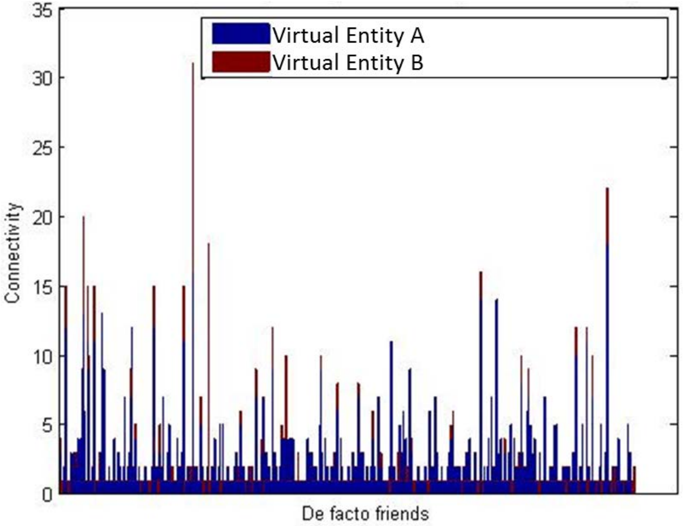
U.S. antivirus legend John McAfee wanted for murder in Belize

By Jeremy A. Kaplan, Alec Liu · Published November 12, 2012 · FoxNews.com

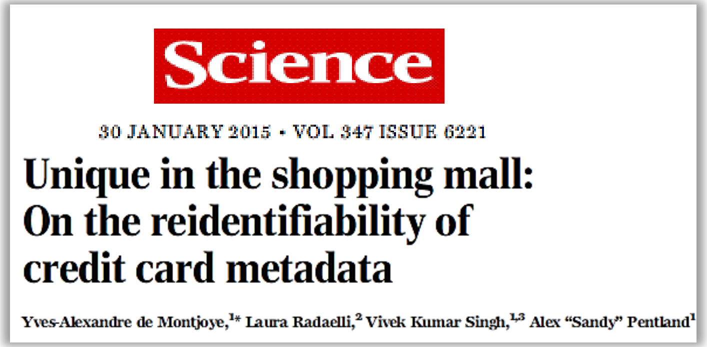
Links & Friends Info for Entity Resolution

- Connectivity links reveal groups & relationships
- Virtual entities suspected as being the same identity have links overlap





Behavior Analysis (Literature Case Study)



Science

30 JANUARY 2015 • VOL 347 ISSUE 6221

Unique in the shopping mall: On the reidentifiability of credit card metadata

Yves-Alexandre de Montjoye,[1*] Laura Radaelli,[2] Vivek Kumar Singh,[1,3] Alex "Sandy" Pentland[1]

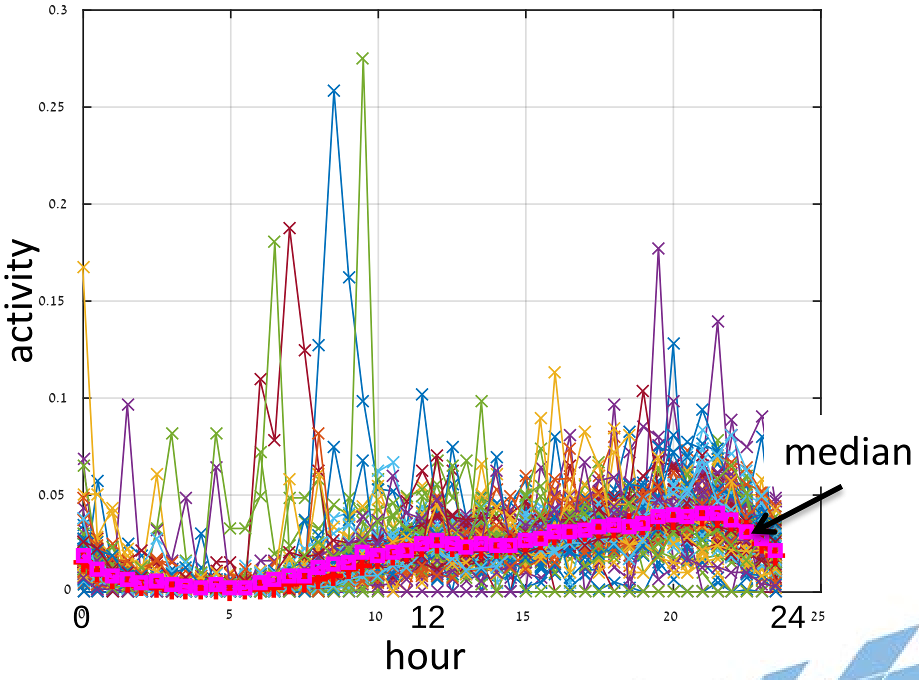
- Dataset
  - Credit card transactions: date, amount, store
  - "Anonymized" people information (no personal details like names or account numbers)
- Using the uniqueness of people's behavior 90% of the shoppers were re-identified as unique individuals
  - (Women are more re-identifiable than men in credit card metadata)
- Combined with publicly available information (posts): Possibility to re-identify people's records by name

Entity Resolution

Identity Resolution

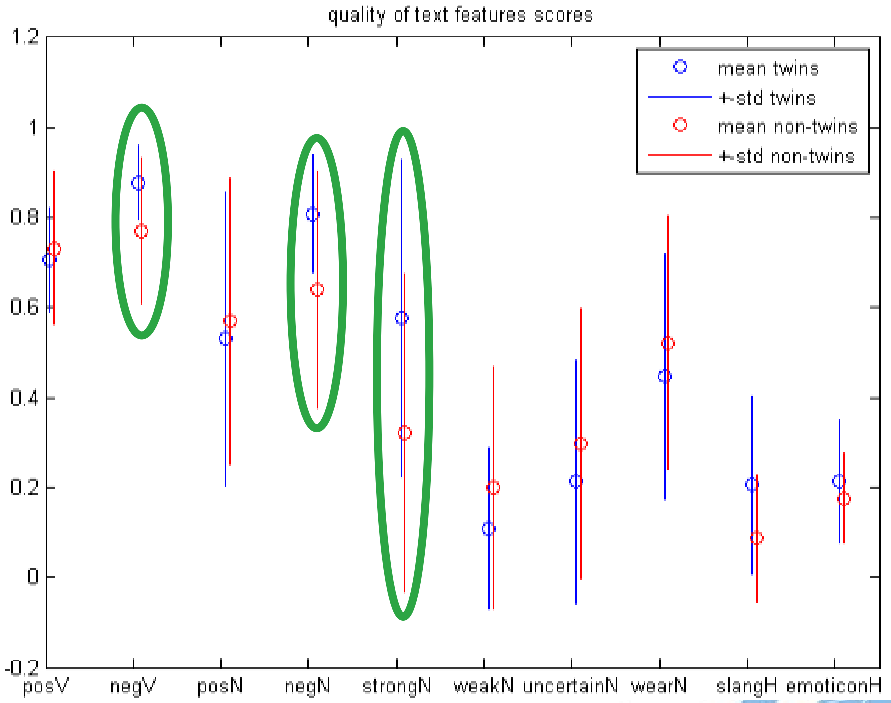
Activity Features Research for Entity Resolution

- Feature: Activity distribution of a virtual entity
  - Normalized activity for a 24 hr period
    - In the example: different entities, same time zone
- Entities can be differentiated using their activity distribution pattern



Text Features Research for Entity Resolution

- Features: Vocabulary & style
- Different criteria, e.g.,
  - "positive", "negative" words
  - "strong", "weak" words
  - slang or emoticons usage
- Some criteria are better than others (culture dependent?)
- Entities can be differentiated using their text style



Quality of Entity Resolution

- **Many features can contribute to resolution of virtual entities**
  - Direct data fields & indirectly inferred information
- None of the techniques is complete;
  None is totally certain
- **Each provides a similarity measure**
- The **more information** from different sources & techniques – the better
  - (Law enforcement agencies can obtain more information, thus improving the capability)
- Best approach is to
  - Consider the result of each technique with its measure of quality & certainty
  - Generate **a weighted combination** of the results of all available information to generate the overall conclusion

“On-the-fly” Analysis

- **Early warning of attacks or crime** requires analysis of the collected data and early reporting, while data is not complete yet
- **"On-the-fly" analysis of streaming data…**
  - Increases the probability for **false positives** and for resolution errors, since the report is based on partial and less confident data
  - Does not allow examination of all the "history" information, whenever a new piece of data is introduced; thus quality is degraded
  - **Decision Making becomes a bigger challenge**
- **Multi-hypothesis analysis and management** is a method to improve performance under on-the-fly conditions

Multi-Hypothesis Analysis

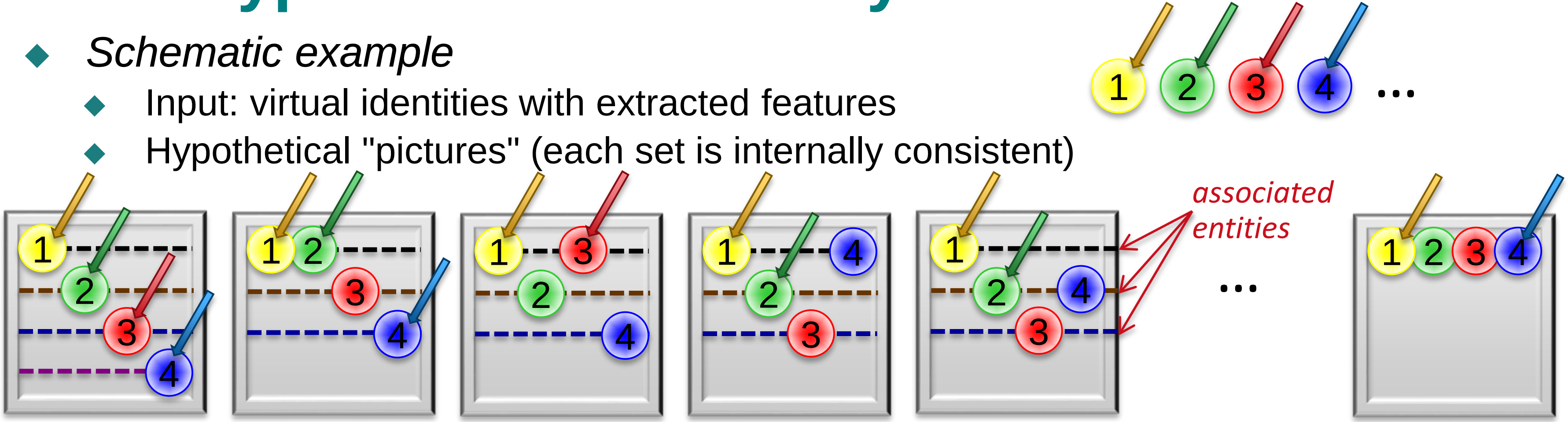
Multi-Hypothesis Analysis is a method to handle the uncertainty

- **An algorithmic methodology to handle complex & dynamic data**
  - collected with various sources/sensors,
  - involving many entities,
  - information is partial and/or ambiguous,
  - information is streaming & dynamically changing

- **For example:**
  - Air situation picture based on geographical data of platform entities
  - Electronic order of battle based on EW&SIGINT data of electromagnetic entities
- **Applicable to Cyber Identity Resolution**
  - Cyber Identity Resolution based on features data of cyber virtual entities
    - Integrating the various information & techniques
    - Supporting decision making

Multi-Hypothesis for Identity Resolution

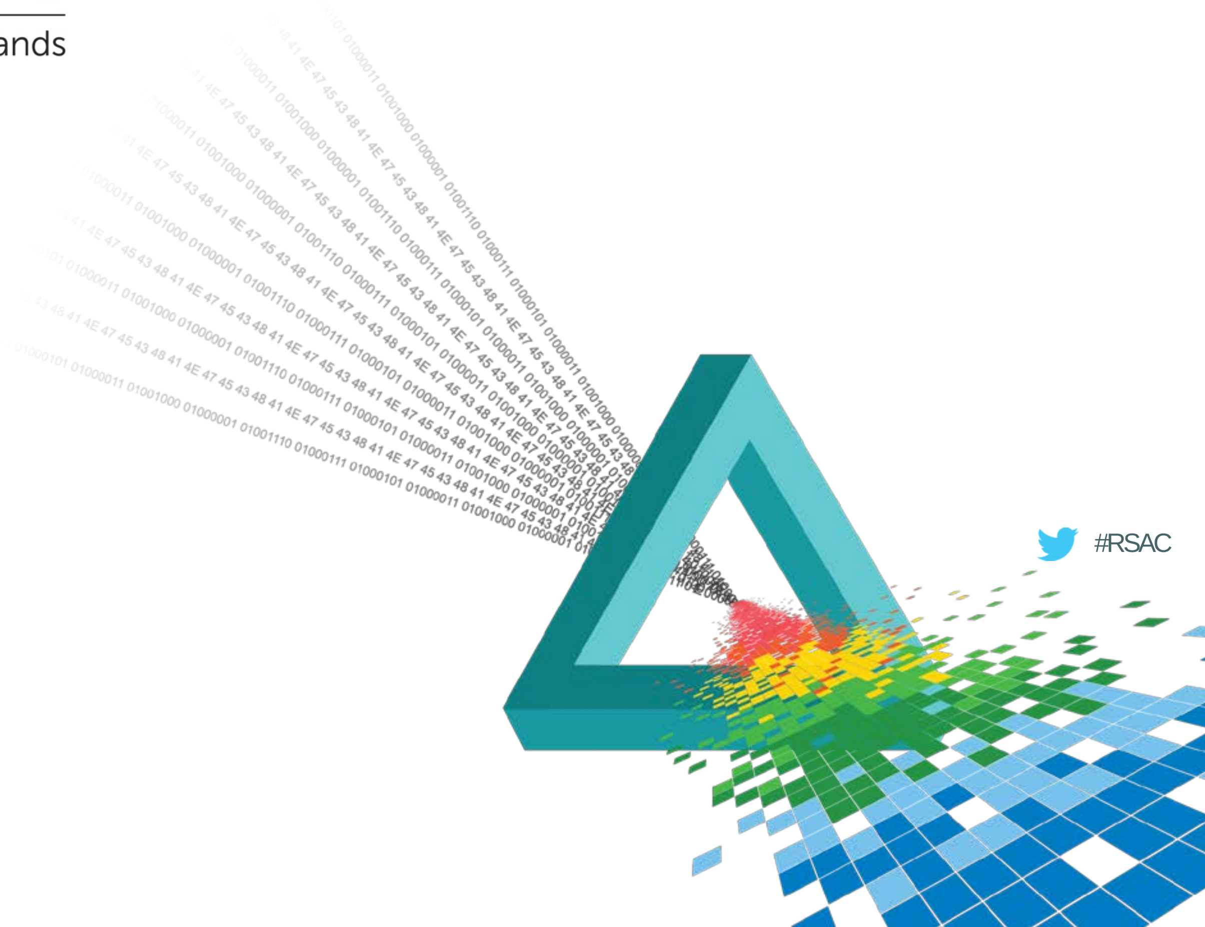
- *Schematic example*
  - Input: virtual identities with extracted features
  - Hypothetical "pictures" (each set is internally consistent)



- Hypothesis score depends on the identity resolution features and their quality
- Only the "picture" with the highest score is reported
- Low-score hypotheses are removed, but many other hypotheses are **maintained without reporting** for further examination with newer data – fewer false alarms
- Multi-hypothesis uses "history" for report updating, in a way that is more efficient, when data is streaming and early response is required

RSAConference2015

Singapore | 22-24 July | Marina Bay Sands



#RSAC

Summary

Electronic Warfare vs. Cyber Warfare (I) – Data Analysis Flow

PROCESS	EW SIGINT	CYBER
Interception	◆ Receiving Electromagnetic Signals (Radar, comm.) ◆ Measuring electronic parameters	
Geo-location	◆ Correlating signals from sensors ◆ Location estimation	
Association	◆ Signals tracking in time	
Classification	◆ Classification based on signal type	
Quality Measure	◆ Quality of the information & uncertainty estimation	
Multiple Hypothesis	◆ Scoring of hypotheses & online management ◆ Removing false alarms	
Report	◆ Integration into Intelligence Center ◆ Supporting Situation Awareness & Early Warning	

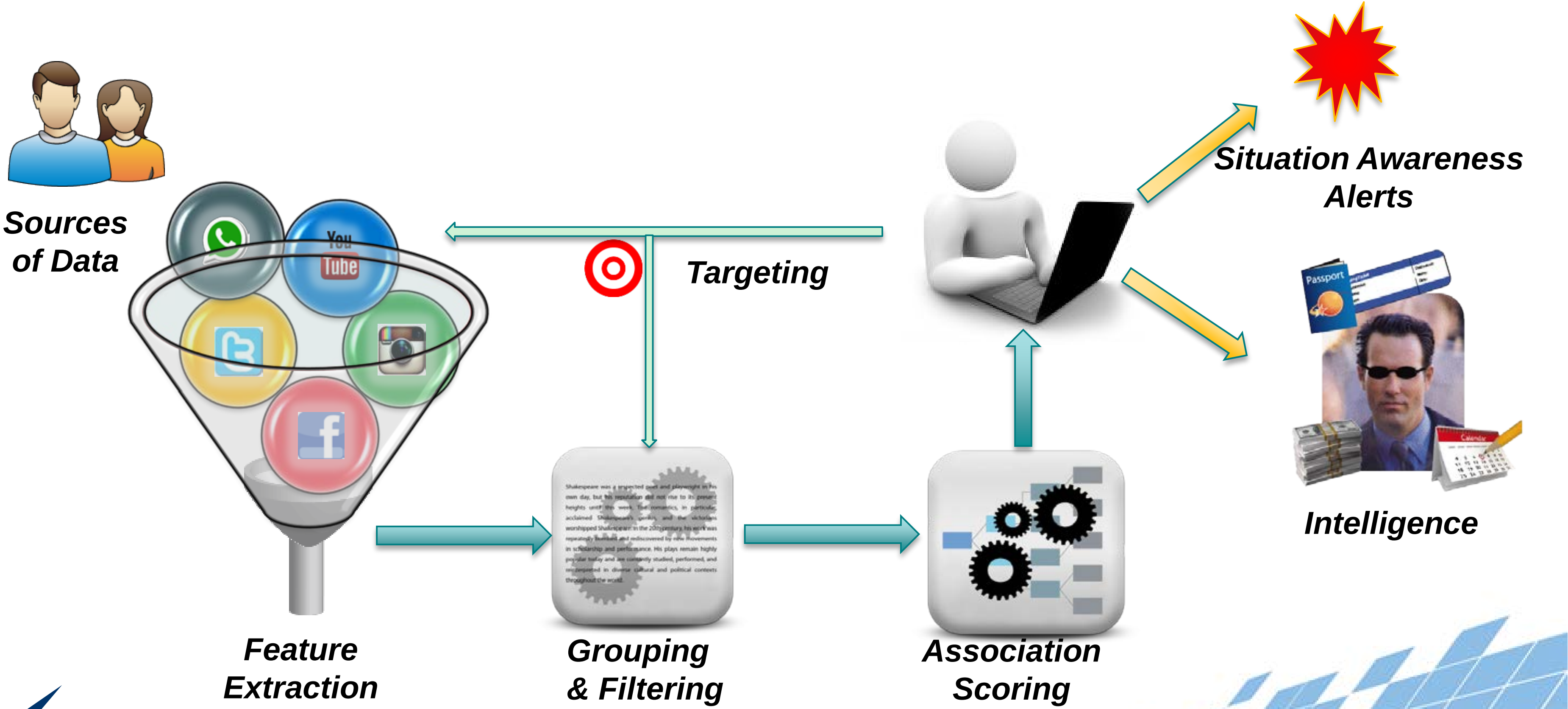
Similar to EW ?

Electronic Warfare vs. Cyber Warfare (II) – Data Analysis Flow

PROCESS	EW SIGINT	CYBER IDENTITY RESOLUTION
Interception	◆ Receiving Electromagnetic Signals ◆ Measuring electronic parameters	◆ Getting Virtual entities activity ◆ Features extraction
Geo-location	◆ Correlating signals from sensors ◆ Location estimation	◆ Correlating Cyber activity or IP ◆ Location estimation
Association	◆ Signals tracking in time	◆ Association of virtual entities – Entity Resolution
Classification	◆ Classification based on signal type	◆ Grouping based on features & behavior
Quality Measure	◆ Quality of the information & uncertainty estimation	
Multiple Hypothesis	◆ Scoring of hypotheses & online management ◆ Removing false alarms	
Report	◆ Integration into Intelligence Center ◆ Supporting Situation Awareness & Early Warning	

IAI

Identity Resolution Flow



Sources of Data

Feature Extraction

Targeting

Grouping & Filtering

Association Scoring

Situation Awareness Alerts

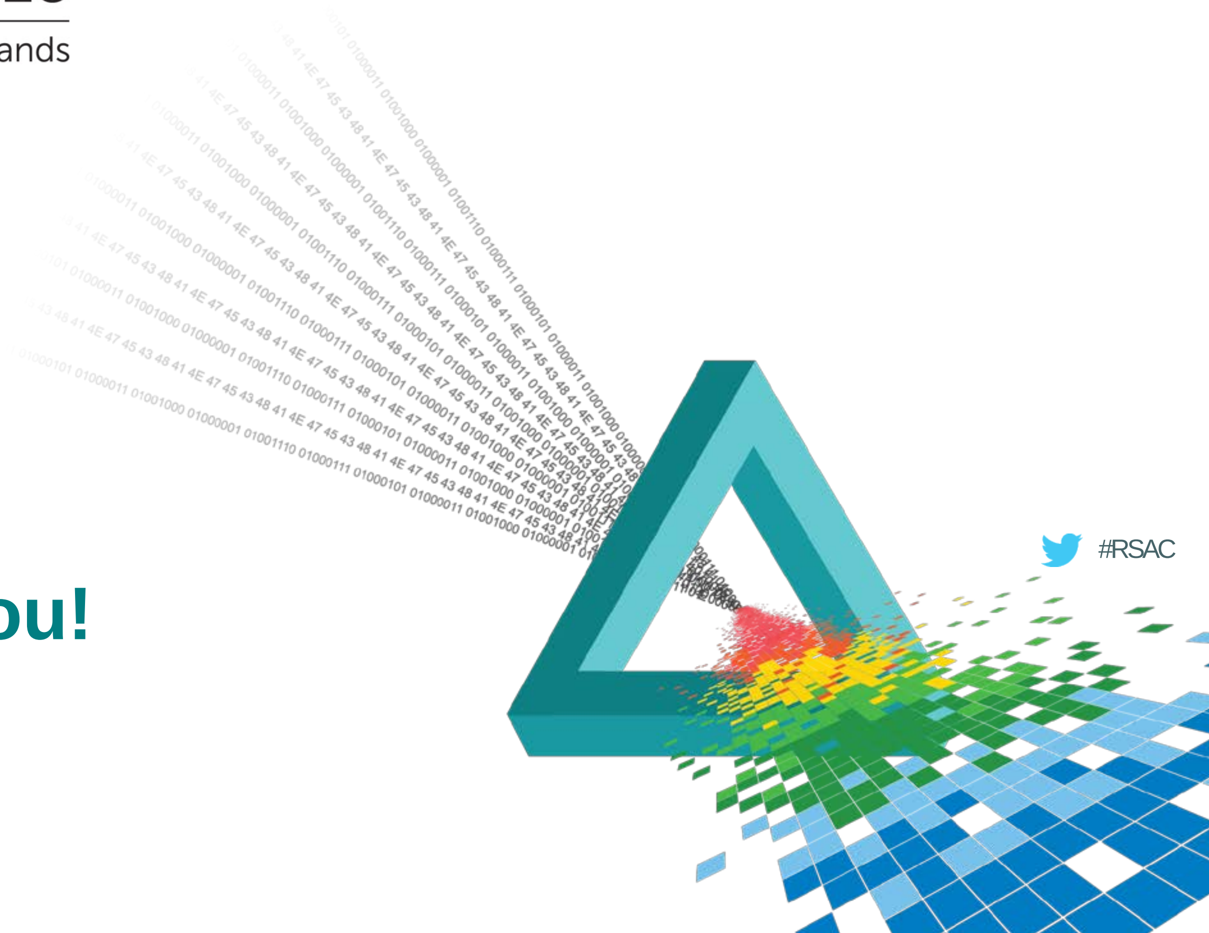
Intelligence

Apply What You Have Learned Today

- Next week you should:
  - Identify potential benefits to "identity resolution" capability in your organization
- In the first three months following this presentation you should:
  - Define your specific goals,
    for example, given a person, find people that are similar or close
  - Identify sources of information (inputs) and expected reports (outputs)
  - Conduct a feasibility study
- Within six months you should:
  - Drive an implementation project for identity resolution capability

RSAConference2015

Singapore | 22-24 July | Marina Bay Sands



#RSAC

Thank you!