# Data Fusion at Scale

Markus De Shon, Ph.D.
Hive Data, LLC

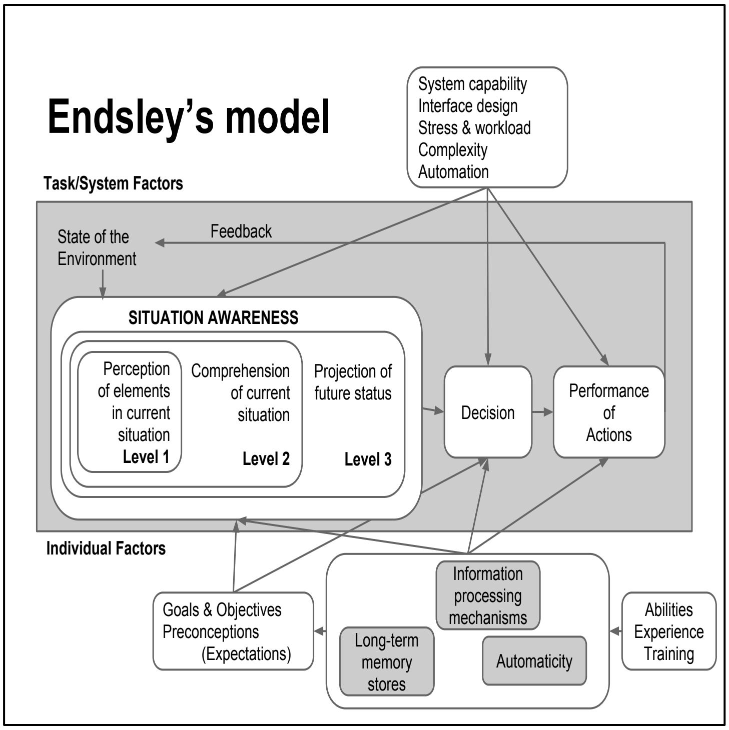
THE HIVE™

# Situation awareness

“**Situation awareness** is the *perception* of the elements in the environment within a volume of time and space, the *comprehension* of their meaning, and the *projection* of their status in the near future.” [emphasis added]

“**Situation assessment**... [is] the process of achieving, acquiring, or maintaining [situation awareness]”

Endsley, M. (1995). Toward a theory of situation awareness in dynamic systems. *Human Factors*, 37(1), 32–64.



# Endsley's model

System capability
Interface design
Stress & workload
Complexity
Automation

Task/System Factors

State of the Environment

Feedback

SITUATION AWARENESS

Perception of elements in current situation
Level 1

Comprehension of current situation
Level 2

Projection of future status
Level 3

Decision

Performance of Actions

Individual Factors

Goals & Objectives
Preconceptions
(Expectations)

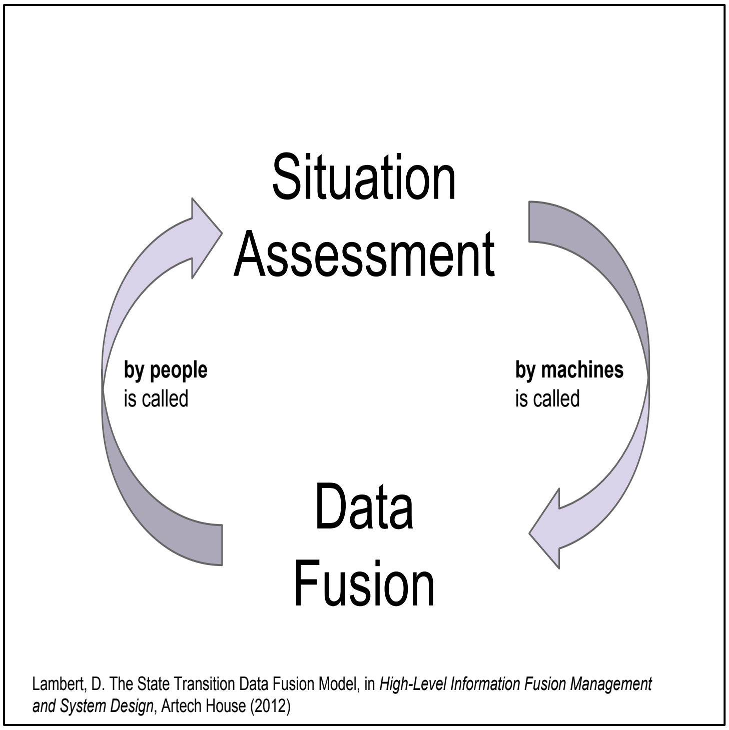
Information processing mechanisms

Long-term memory stores

Automaticity

Abilities
Experience
Training

Endsley's model represents the analyst's mental process as they construct a mental model of the world, i.e. their network and the actors in it and upon it. This doesn't help us that much in building automated systems, however, except that some important considerations in designing a human-computer interface for situational awareness include:
* System capability - how to include all the necessary data and process it in an appropriate time frame?
* Interface design - how to design it to support awareness as an explicit goal?
* Automation - how much can we take off the analyst's plate?



Situation Assessment

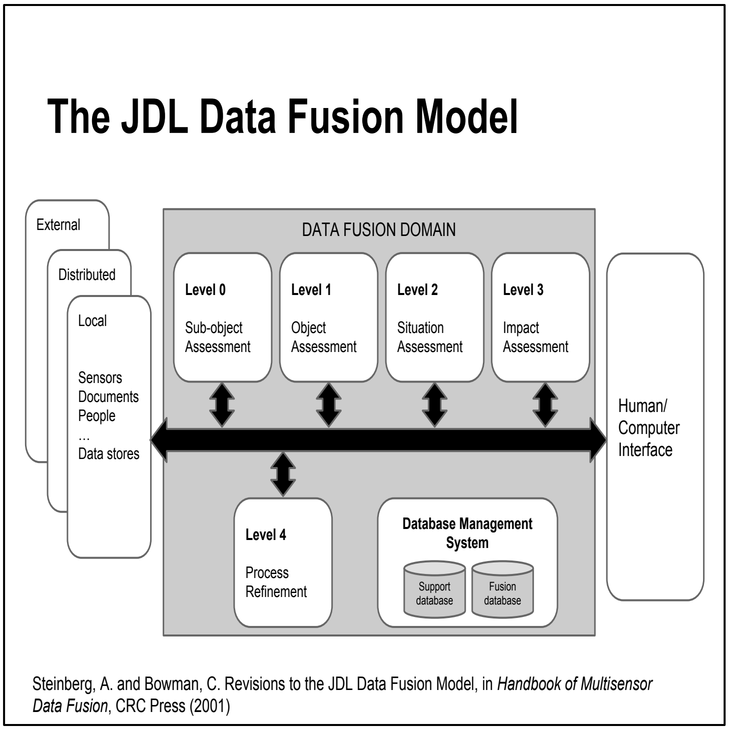
**by people** is called

**by machines** is called

Data Fusion

Lambert, D. The State Transition Data Fusion Model, in *High-Level Information Fusion Management and System Design*, Artech House (2012)

Fortunately, there is a related concept, Data Fusion, which is an automated situation assessment process. While we might not be able to automate all the levels (Perception, Comprehension, Projection) we can automate some and support others. Let’s explore the Data Fusion field to see what might be useful to us.



## The JDL Data Fusion Model

External

Distributed

Local

Sensors
Documents
People
...
Data stores

DATA FUSION DOMAIN

**Level 0**

Sub-object Assessment

**Level 1**

Object Assessment

**Level 2**

Situation Assessment

**Level 3**

Impact Assessment

**Level 4**

Process Refinement

**Database Management System**

Support database

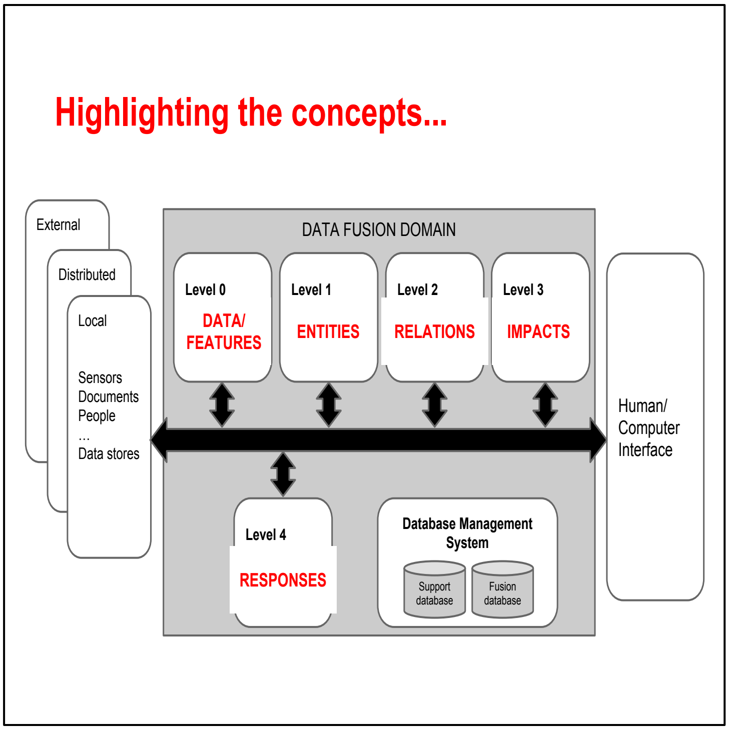
Fusion database

Human/ Computer Interface

Steinberg, A. and Bowman, C. Revisions to the JDL Data Fusion Model, in *Handbook of Multisensor Data Fusion*, CRC Press (2001)

The JDL Data Fusion model describes an automated process that presents information through an HCI. Some previous work has applied this model to cybersecurity (see below). However, this model provides only a high-level roadmap for data fusion, we perhaps need some more guidance on what needs to be done.

Giacobe, N. a. (2010). Application of the JDL Data Fusion Process Model for Cyber Security. (J. J. Braun, Ed.), 7710(May), 77100R–77100R–10. doi:10.1117/12.850275
Yang, S. J., Stotz, A., Holsopple, J., Sudit, M., & Kuhl, M. (2009). High level information fusion for tracking and projection of multistage cyber attacks. Information Fusion, 10(1), 107–121. doi:10.1016/j.inffus.2007.06.002
Sudit, M., Holender, M., Stotz, A., Rickard, T., & Yager, R. (2007). INFERD and Entropy for Situational Awareness. Journal Adv. Info. Fusion, 2(1). Retrieved from http://isif.org/sites/isif.org/files/journals/2-4075D01.pdf



## Highlighting the concepts...

External

Distributed

Local

Sensors
Documents
People
...
Data stores

DATA FUSION DOMAIN

Level 0
DATA/
FEATURES

Level 1
ENTITIES

Level 2
RELATIONS

Level 3
IMPACTS

Level 4
RESPONSES

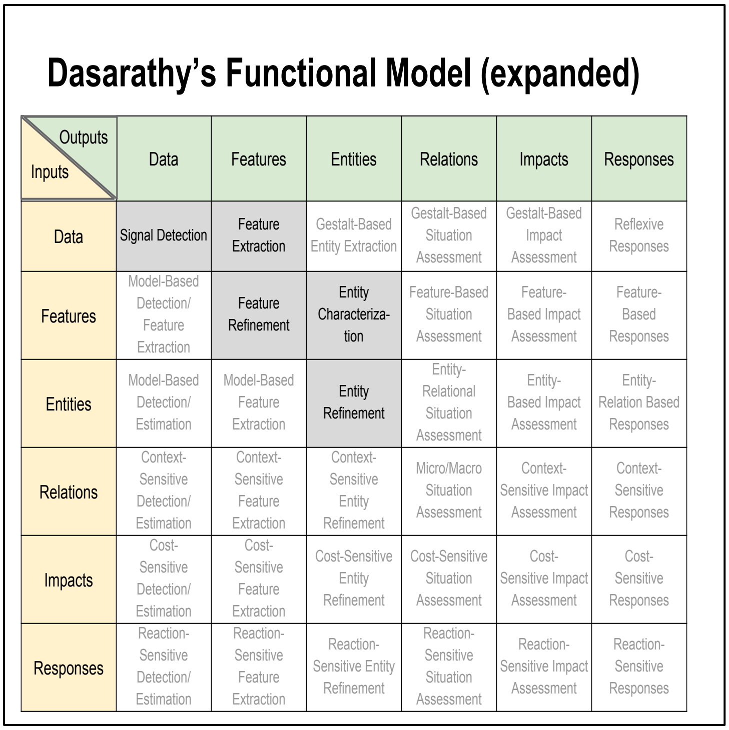
Database Management
System

Support
database

Fusion
database

Human/
Computer
Interface

Fortunately, there is a related model that uses somewhat different terminology to refer to the inputs/outputs of a data fusion process at the various levels of the JDL Data Fusion model.

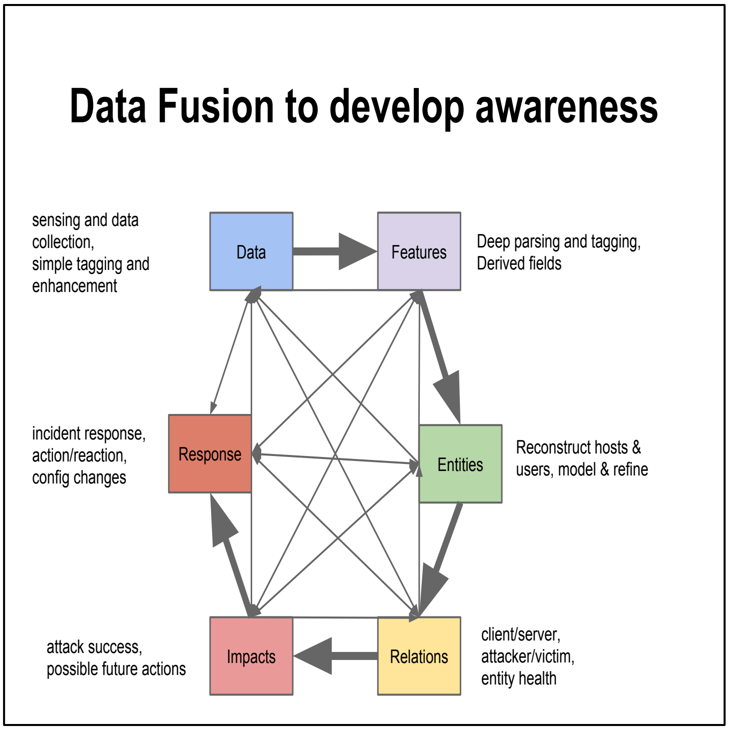


## Dasarathy’s Functional Model (expanded)

| Inputs \ Outputs | Data | Features | Entities | Relations | Impacts | Responses |
|---|---|---|---|---|---|---|
| Data | Signal Detection | Feature Extraction | Gestalt-Based Entity Extraction | Gestalt-Based Situation Assessment | Gestalt-Based Impact Assessment | Reflexive Responses |
| Features | Model-Based Detection/ Feature Extraction | Feature Refinement | Entity Characterization | Feature-Based Situation Assessment | Feature-Based Impact Assessment | Feature-Based Responses |
| Entities | Model-Based Detection/ Estimation | Model-Based Feature Extraction | Entity Refinement | Entity-Relational Situation Assessment | Entity-Based Impact Assessment | Entity-Relation Based Responses |
| Relations | Context-Sensitive Detection/ Estimation | Context-Sensitive Feature Extraction | Context-Sensitive Entity Refinement | Micro/Macro Situation Assessment | Context-Sensitive Impact Assessment | Context-Sensitive Responses |
| Impacts | Cost-Sensitive Detection/ Estimation | Cost-Sensitive Feature Extraction | Cost-Sensitive Entity Refinement | Cost-Sensitive Situation Assessment | Cost-Sensitive Impact Assessment | Cost-Sensitive Responses |
| Responses | Reaction-Sensitive Detection/ Estimation | Reaction-Sensitive Feature Extraction | Reaction-Sensitive Entity Refinement | Reaction-Sensitive Situation Assessment | Reaction-Sensitive Impact Assessment | Reaction-Sensitive Responses |

An extended version of Dasarathy’s Functional Model (first introduced in a simple form in Dasarathy (1994) but extended in Steinberg, A. and Bowman, C. (2001)) provides a detailed roadmap to the components of a full data and information fusion system. While the NorthWest quadrant is relatively familiar territory in the Multisensor Data Fusion world, with some forays into the SouthEast, the NorthEast and SouthWest quadrants are relatively unexplored even in that more mature field.

Dasarathy, B., *Decision Fusion*, IEEE Computer Society Press, 1994.



# Data Fusion to develop awareness

sensing and data collection,
simple tagging and enhancement

Data

Features

Deep parsing and tagging,
Derived fields

incident response,
action/reaction,
config changes

Response

Entities

Reconstruct hosts & users, model & refine

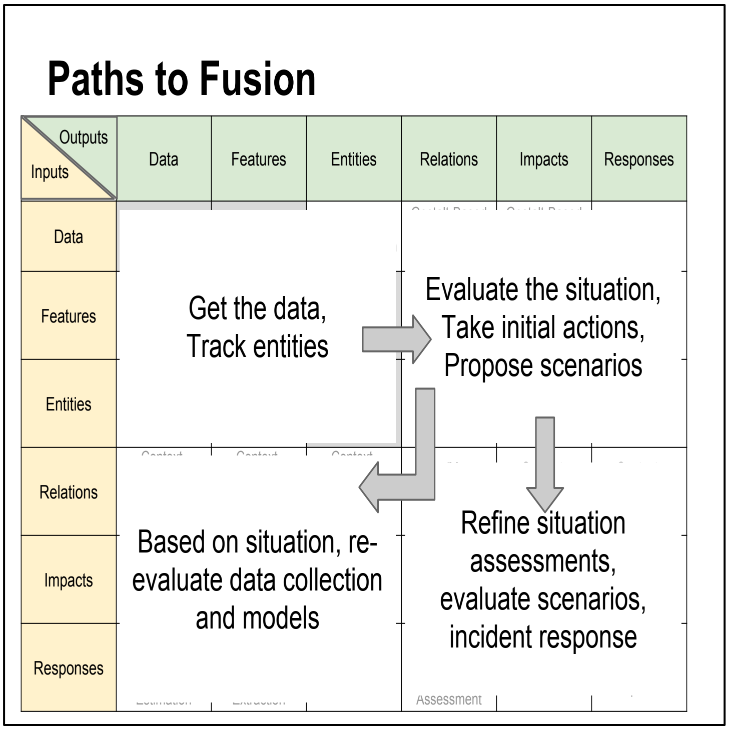
attack success,
possible future actions

Impacts

Relations

client/server,
attacker/victim,
entity health

Here is a suggested flow for data fusion processes. From blue→red things become more abstract and require higher cognition, and the heavy arrows indicate the primary reasoning path, but there are interactions up and down the ladder. This is a fully connected graph of influences.



## Paths to Fusion

| Outputs / Inputs | Data | Features | Entities | Relations | Impacts | Responses |
|---|---|---|---|---|---|---|
| Data | | | | | | |
| Features | | | | | | |
| Entities | | | | | | |
| Relations | | | | | | |
| Impacts | | | | | | |
| Responses | | | | | | |

Get the data, Track entities

Evaluate the situation, Take initial actions, Propose scenarios

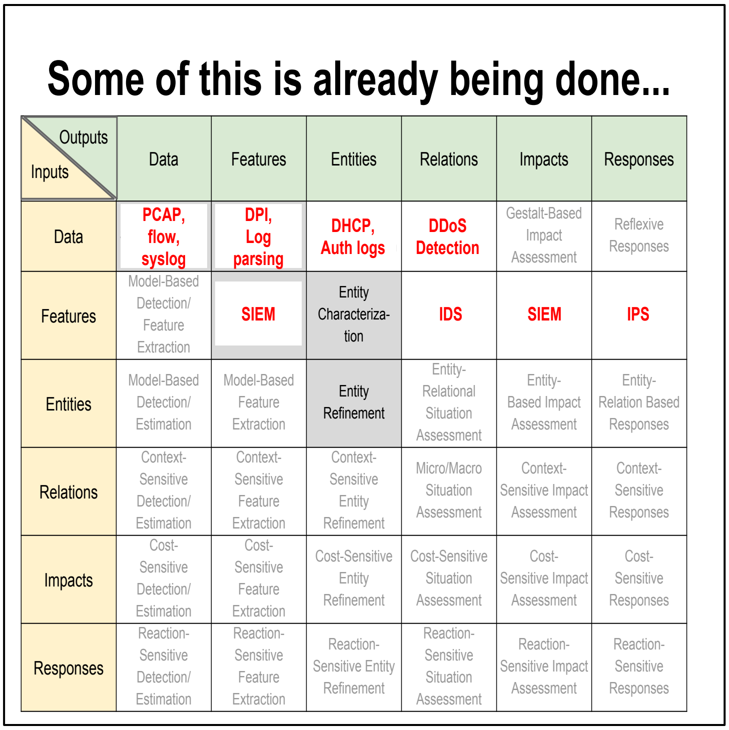
Based on situation, re-evaluate data collection and models

Refine situation assessments, evaluate scenarios, incident response

Viewed another way, we need to have extensive data collection and low-level fusion processes in the NorthWest quadrant, which can lead to making some higher-level inferences in the NorthEast quadrant, which is the primary area for deciding whether malicious activity is taking place, what the consequences are for the defended network, and possible responses.

Once there is some understanding of the situation, then in the SouthWest quadrant we can feed back into our lower-level processes, for example to save data for suspicious sessions longer, initiate new or more detailed data collection, or just modify our signatures and configurations.

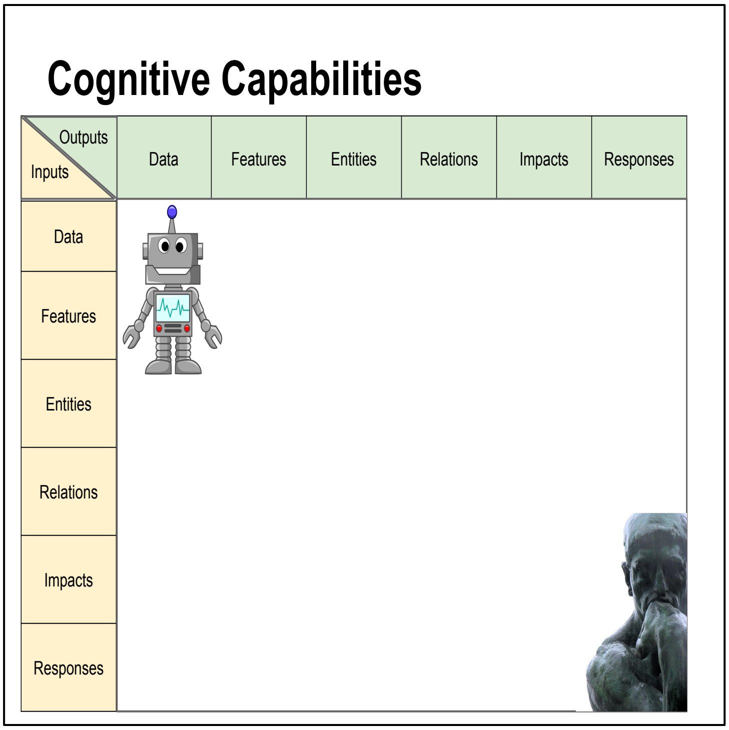
In the SouthEast, quadrant we can use our situation assessment to decide what might happen in the near future, and to perform incident response.



## Some of this is already being done...

| Outputs<br>Inputs | Data | Features | Entities | Relations | Impacts | Responses |
|---|---|---|---|---|---|---|
| Data | **PCAP, flow, syslog** | **DPI, Log parsing** | **DHCP, Auth logs** | **DDoS Detection** | Gestalt-Based Impact Assessment | Reflexive Responses |
| Features | Model-Based Detection/ Feature Extraction | **SIEM** | Entity Characterization | **IDS** | **SIEM** | **IPS** |
| Entities | Model-Based Detection/ Estimation | Model-Based Feature Extraction | Entity Refinement | Entity-Relational Situation Assessment | Entity-Based Impact Assessment | Entity-Relation Based Responses |
| Relations | Context-Sensitive Detection/ Estimation | Context-Sensitive Feature Extraction | Context-Sensitive Entity Refinement | Micro/Macro Situation Assessment | Context-Sensitive Impact Assessment | Context-Sensitive Responses |
| Impacts | Cost-Sensitive Detection/ Estimation | Cost-Sensitive Feature Extraction | Cost-Sensitive Entity Refinement | Cost-Sensitive Situation Assessment | Cost-Sensitive Impact Assessment | Cost-Sensitive Responses |
| Responses | Reaction-Sensitive Detection/ Estimation | Reaction-Sensitive Feature Extraction | Reaction-Sensitive Entity Refinement | Reaction-Sensitive Situation Assessment | Reaction-Sensitive Impact Assessment | Reaction-Sensitive Responses |

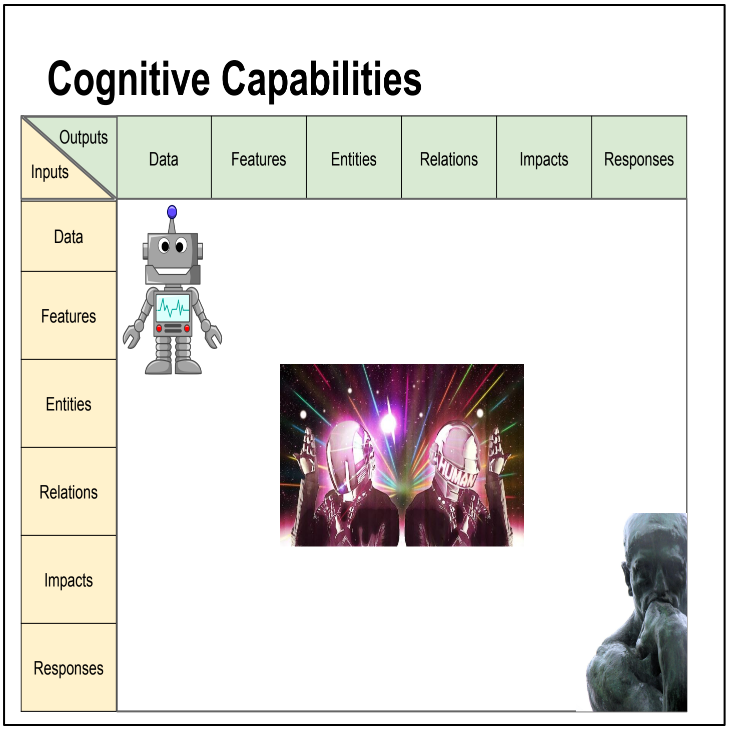
It is useful to consider where some existing industry solutions fall in this model. Clearly the industry has been doing some higher-level work, but it’s important to realize that the forays into the NorthEast have been signature- and rule-driven. In other words, these areas represent the distilled understanding of human analysts rather than any kind of automated information fusion process.



# Cognitive Capabilities

| Outputs<br>Inputs | Data | Features | Entities | Relations | Impacts | Responses |
|---|---|---|---|---|---|---|
| Data | | | | | | |
| Features | | | | | | |
| Entities | | | | | | |
| Relations | | | | | | |
| Impacts | | | | | | |
| Responses | | | | | | |

The model suggests where automation is most appropriate. Clearly in the top left we can fully automate (data collection, protocol decoding for feature extraction). At the bottom right we probably never will be able to automate the response/counter-response process between human adversaries in this information space. However...

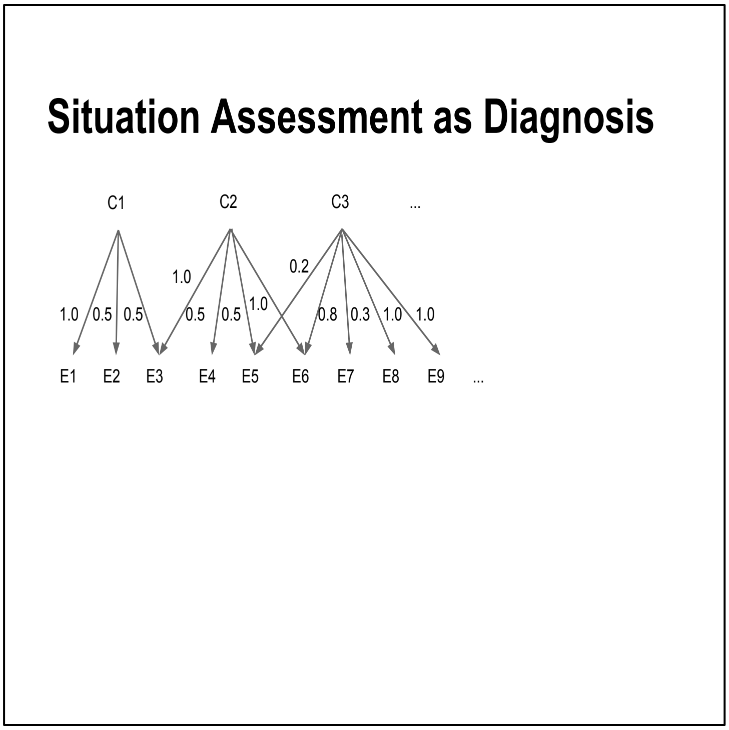


# Cognitive Capabilities

| Outputs<br>Inputs | Data | Features | Entities | Relations | Impacts | Responses |
|---|---|---|---|---|---|---|
| Data | | | | | | |
| Features | | | | | | |
| Entities | | | | | | |
| Relations | | | | | | |
| Impacts | | | | | | |
| Responses | | | | | | |

In the middle we need Daft Punk, part robot, part human. We need to be aggressive about applying machine learning approprately to provide higher-level abstractions to the human analyst so that they can form correct situational awareness models more easily than poring over raw data, or waiting for alerts in their queue. This is both possible and necessary.

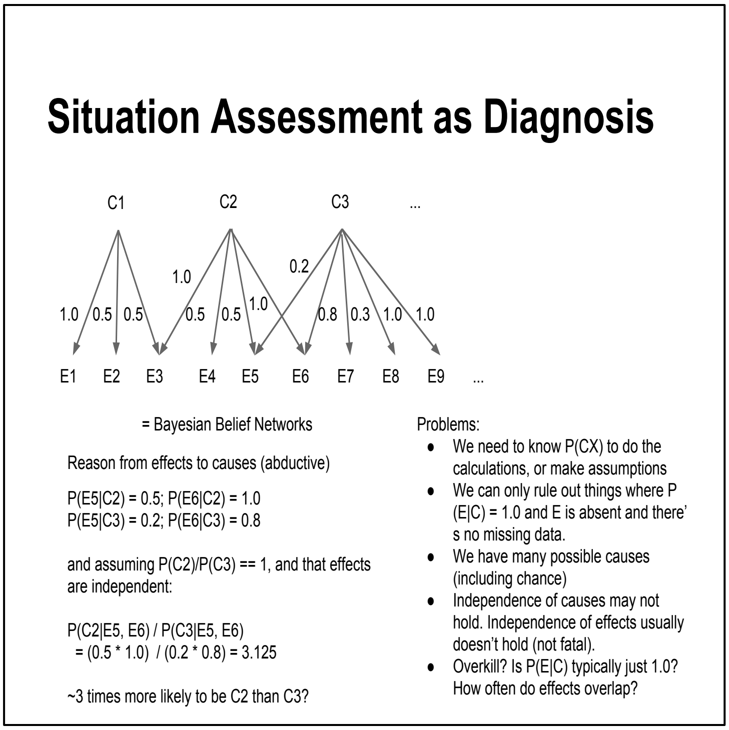
Daft Punk image by thedeviant426 at deviantart.com used under creative commons.



## Situation Assessment as Diagnosis

A quick detour: what about viewing security monitoring as a diagnostic process. We have a number of possible causes (C1, C2, … representing particular exploit kits, malware, human hacking techniques) each of which has a set of possible effects (E1, E2, …), and we have some belief about the chance of each effect occurring for a cause.

Now, if we observe some set of effects, there are ways to calculate, based on such a graph, the most likely cause(s).



# Situation Assessment as Diagnosis

= Bayesian Belief Networks

Reason from effects to causes (abductive)

P(E5|C2) = 0.5; P(E6|C2) = 1.0
P(E5|C3) = 0.2; P(E6|C3) = 0.8

and assuming P(C2)/P(C3) == 1, and that effects are independent:

P(C2|E5, E6) / P(C3|E5, E6)
= (0.5 * 1.0) / (0.2 * 0.8) = 3.125

~3 times more likely to be C2 than C3?

Problems:

- We need to know P(CX) to do the calculations, or make assumptions
- We can only rule out things where P (E|C) = 1.0 and E is absent and there' s no missing data.
- We have many possible causes (including chance)
- Independence of causes may not hold. Independence of effects usually doesn't hold (not fatal).
- Overkill? Is P(E|C) typically just 1.0? How often do effects overlap?

This type of graph is called a Bayesian Belief Network, and it's an efficient representation of "abductive reasoning" which could be a way to automate how analysts think when they are doing security analysis. Unfortunately, while we are in some sense balancing such belief values implicitly, we rarely or never make such calculations explicit.(*) However, models have been applied to medical diagnostics with substantial success--could it be a way to automate certain mid-level data fusion tasks?

Here is an example calculation where we have observed effects E5 and E6. The result is that it's three times more likely to be C2 than C3.

P(C2|E5, E6) = P(C2) P(E5|C2) P(E6|C2) / ( P(E5)P(E6) )
P(C3|E5, E6) = P(C3) P(E5|C3) P(E6|C3) / ( P(E5)P(E6) )

P(C2|E5, E6) / P(C3|E5, E6) =
(P(C2) / P(C3) which is assumed to be 1) *
[P(E5|C2) P(E6|C2)] / [P(E5|C3) P(E6|C3)]

(*) In fact, Gary Klein et al. have shown that experienced people don't actually decide anything, they just recognize situations that they have seen in the past, and as a result know what to do. This means we need to become much better at helping analysts recognize situations (situation awareness…). This is a larger subject beyond the scope of this presentation.

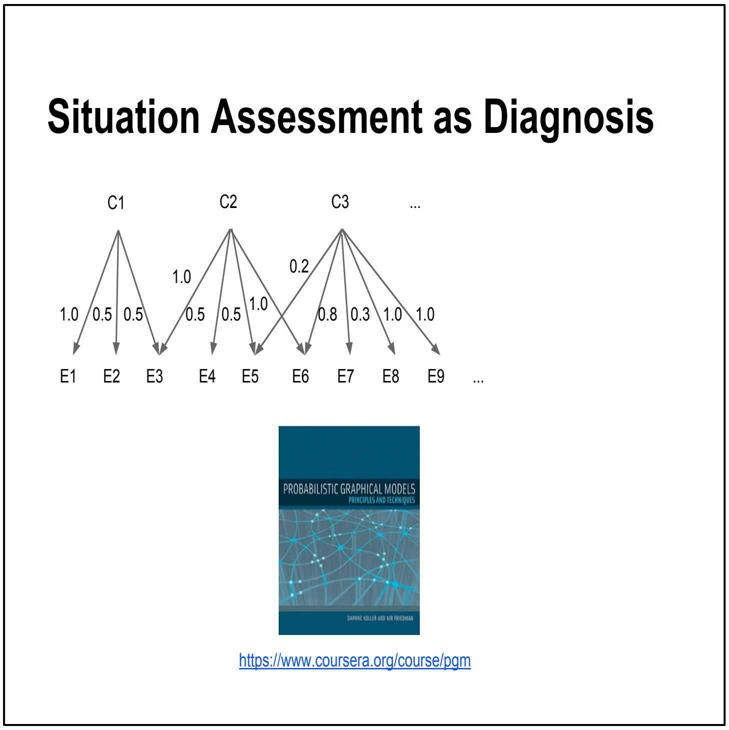
## Situation Assessment as Diagnosis

This looks hard. However...

We are currently building these networks

- Implicitly, not explicitly
- Individually, not collectivley
- Manually, not automatically

This is probably not a good model of higher-level reasoning, particularly Impact/Response...

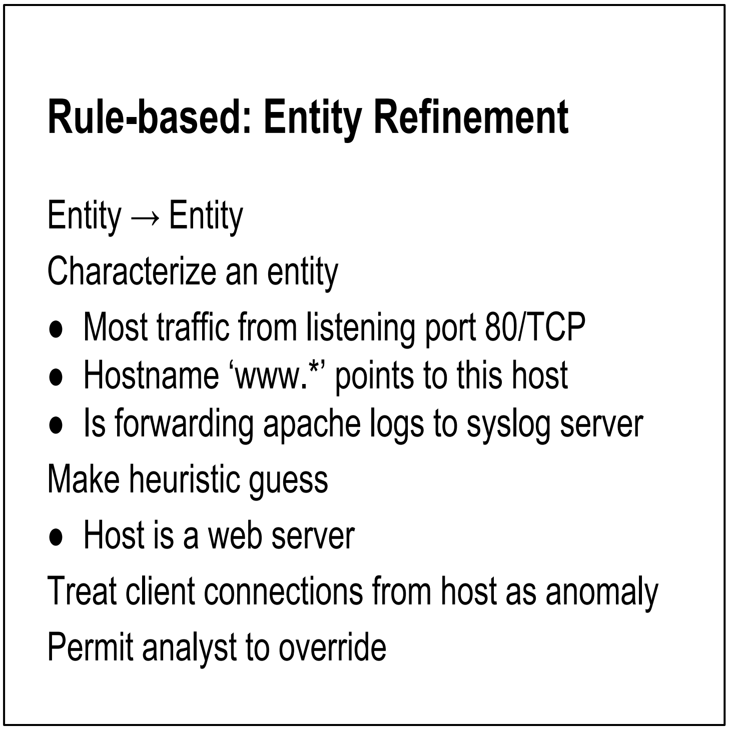


# Situation Assessment as Diagnosis

C1
C2
C3
...
1.0
0.2
1.0
0.5
0.5
0.5
0.5
1.0
0.8
0.3
1.0
1.0
0.5
E1
E2
E3
E4
E5
E6
E7
E8
E9
...
PROBABILISTIC GRAPHICAL MODELS
PRINCIPLES AND TECHNIQUES
DAPHNE KOLLER AND NIR FRIEDMAN

https://www.coursera.org/course/pgm

These types of models are described in exhaustive detail in Daphne Koller’s book and free Coursera course (though it takes 20 hrs/week x 10 weeks to complete).



## Rule-based: Entity Refinement

Entity → Entity

Characterize an entity

- Most traffic from listening port 80/TCP
- Hostname ‘www.*’ points to this host
- Is forwarding apache logs to syslog server

Make heuristic guess

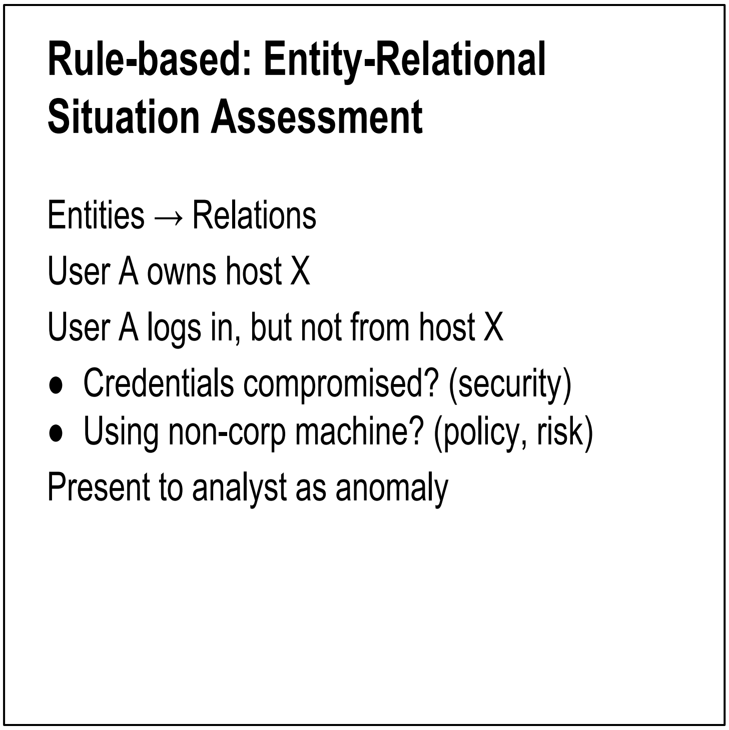
- Host is a web server

Treat client connections from host as anomaly

Permit analyst to override

Here is an example of a way we can automate data fusion. It’s important to note that this is a rule-based procedure, which is not (strictly speaking) a machine learning approach, since such rule sets have to be constructed and maintained by experts. However, it can help to automate an otherwise manual process of identifying and labeling web servers on the network, which becomes more important the larger the network is.

Rules can also be a way to codify best practices, so that junior analysts can benefit from more experienced analysts’ work.



## Rule-based: Entity-Relational Situation Assessment

Entities → Relations

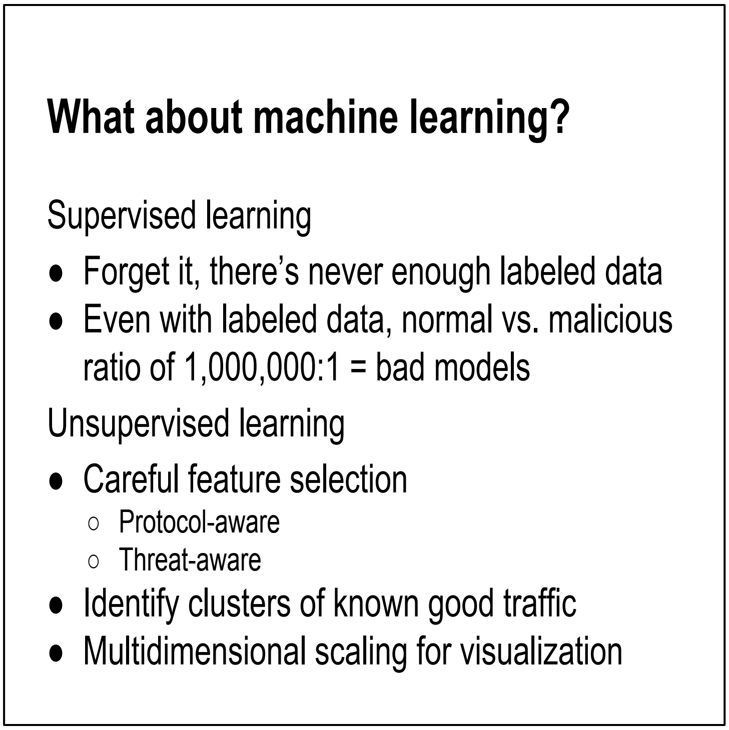
User A owns host X

User A logs in, but not from host X

- Credentials compromised? (security)
- Using non-corp machine? (policy, risk)

Present to analyst as anomaly

Another rule-based example



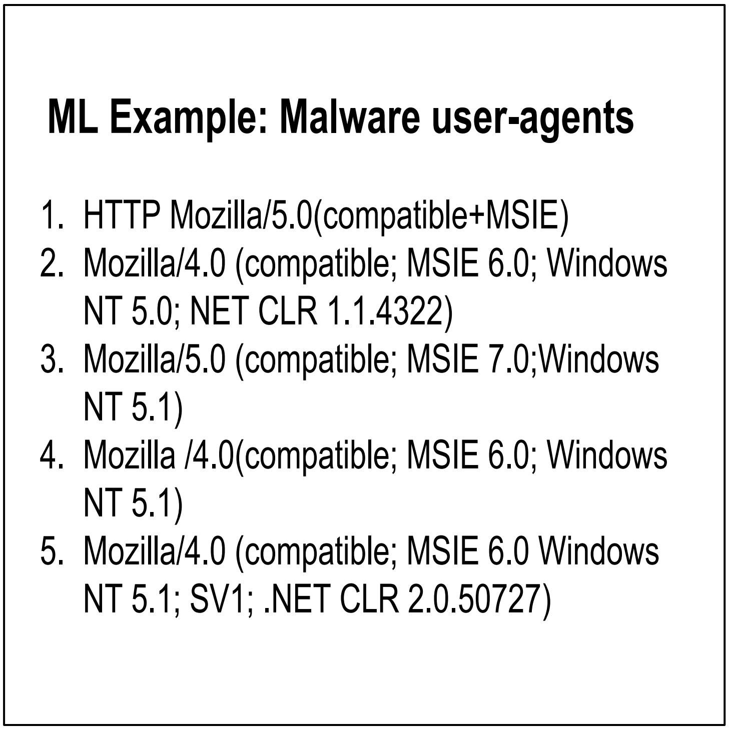
## What about machine learning?

Supervised learning

- Forget it, there's never enough labeled data
- Even with labeled data, normal vs. malicious ratio of 1,000,000:1 = bad models

Unsupervised learning

- Careful feature selection
  - Protocol-aware
  - Threat-aware
- Identify clusters of known good traffic
- Multidimensional scaling for visualization



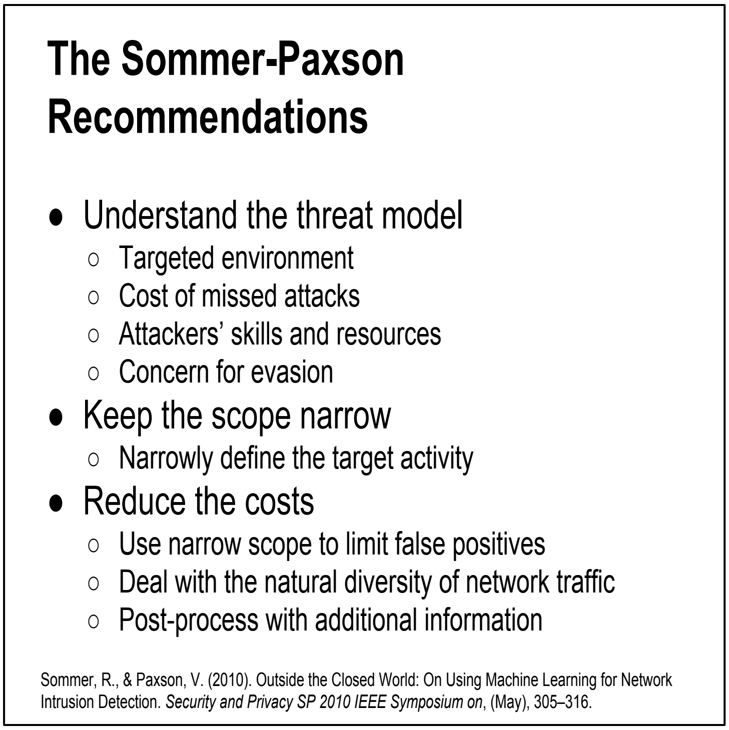
## ML Example: Malware user-agents

1. HTTP Mozilla/5.0(compatible+MSIE)
2. Mozilla/4.0 (compatible; MSIE 6.0; Windows NT 5.0; NET CLR 1.1.4322)
3. Mozilla/5.0 (compatible; MSIE 7.0;Windows NT 5.1)
4. Mozilla /4.0(compatible; MSIE 6.0; Windows NT 5.1)
5. Mozilla/4.0 (compatible; MSIE 6.0 Windows NT 5.1; SV1; .NET CLR 2.0.50727)

These are User-Agent strings from actual malware command and control HTTP connections. Look closely to identify the errors in each.

Answers:
1. Just all wrong
2. NET instead of .NET
3. No space after semicolon
4. No space before parenthesis
5. No semicolon after MSIE 6.0

So 2-5 could be well-served by a feature that calculates the Levenshtein edit distance from known good browser strings. Small edit distances in this case would be bad.



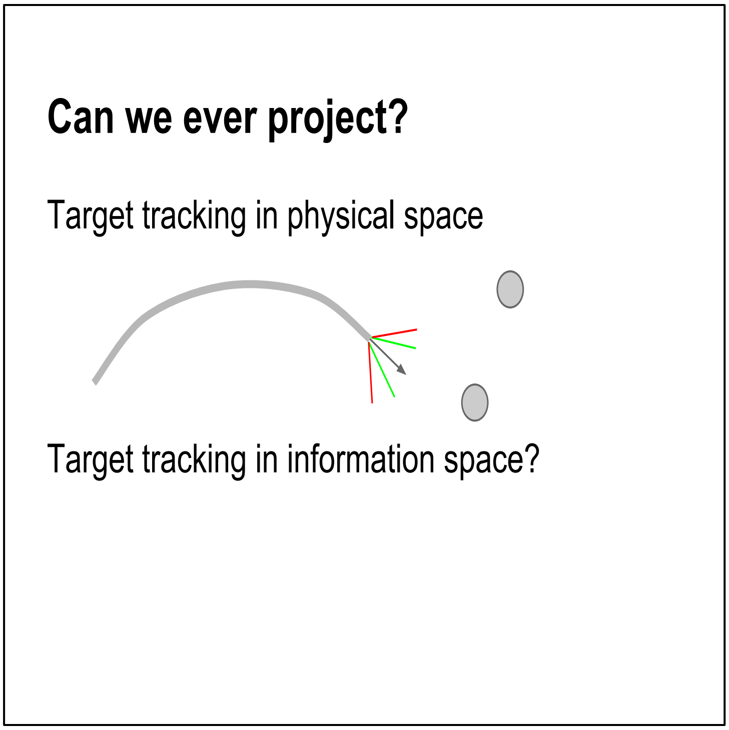
## The Sommer-Paxson Recommendations

- Understand the threat model
  - Targeted environment
  - Cost of missed attacks
  - Attackers' skills and resources
  - Concern for evasion
- Keep the scope narrow
  - Narrowly define the target activity
- Reduce the costs
  - Use narrow scope to limit false positives
  - Deal with the natural diversity of network traffic
  - Post-process with additional information

Sommer, R., & Paxson, V. (2010). Outside the Closed World: On Using Machine Learning for Network Intrusion Detection. *Security and Privacy SP 2010 IEEE Symposium on*, (May), 305–316.

These are recommendations for avoiding some of the pitfalls of earlier attempts to apply ML to the security problem. A naive, black box approach just doesn't work. It's especially important to construct models that make sense to analysts so that

1. The results are defensible, as otherwise the analysts won't believe the results and thus the system will be ignored.
2. Error-prone models can be corrected, localized or updated based on the semantics of the threats and protocols involved.

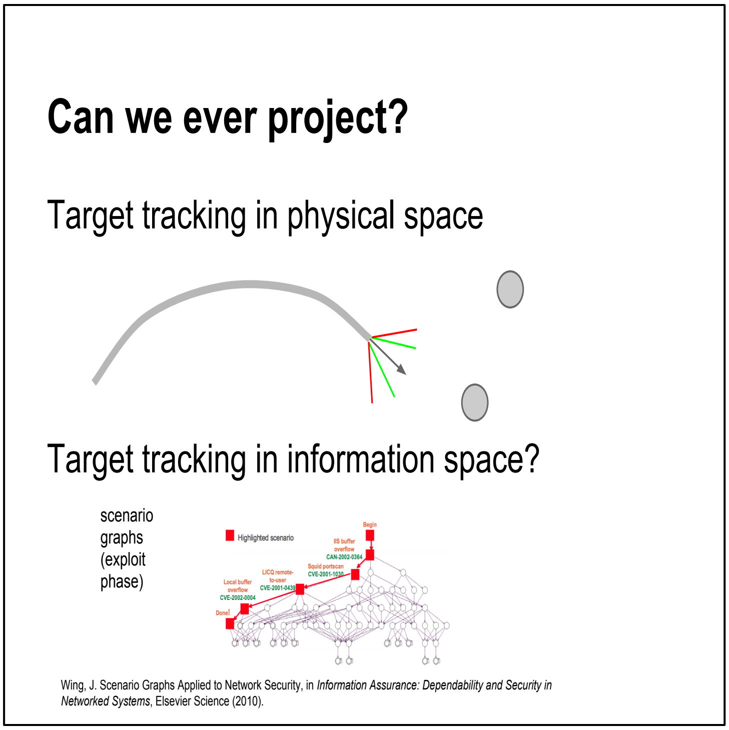


## Can we ever project?

Target tracking in physical space

Target tracking in information space?

What about projecting into the future? In physical space (tracking targets on radar) we are detecting physical objects that follow the laws of Physics, so we can predict future scenarios based on constraints such as force and velocity. What about in an abstract information space, though?



# Can we ever project?

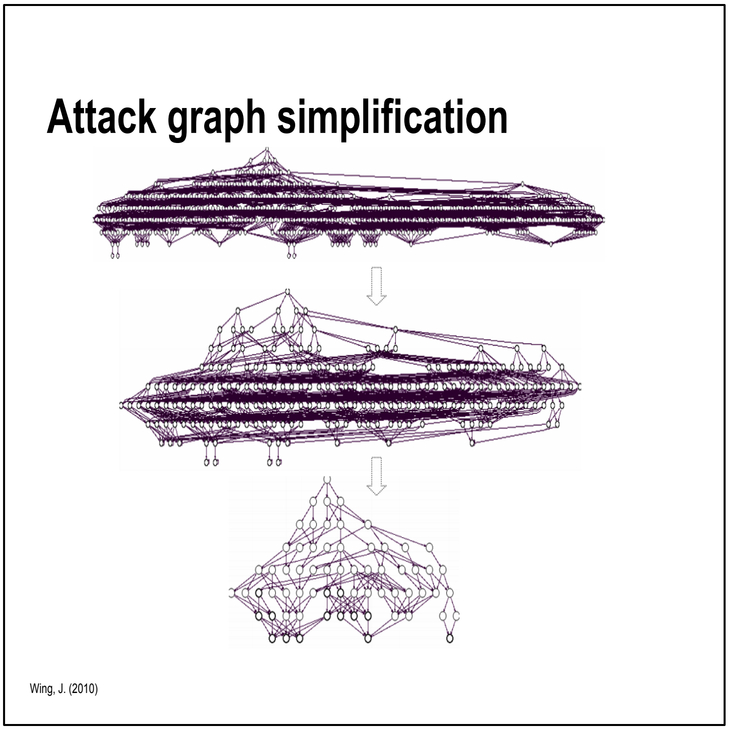
Target tracking in physical space

Target tracking in information space?

scenario graphs (exploit phase)

Wing, J. Scenario Graphs Applied to Network Security, in *Information Assurance: Dependability and Security in Networked Systems*, Elsevier Science (2010).

Some of the most principled work in this area is in attack graphs/scenario graphs. By understanding the vulnerabilities in the defended network, we can map out possible future actions by attackers. This state space can become large quite quickly, however. Also, these attack graphs don’t usually cover human vulnerability, which obviously plays an important role in modern intrusions.



## Attack graph simplification

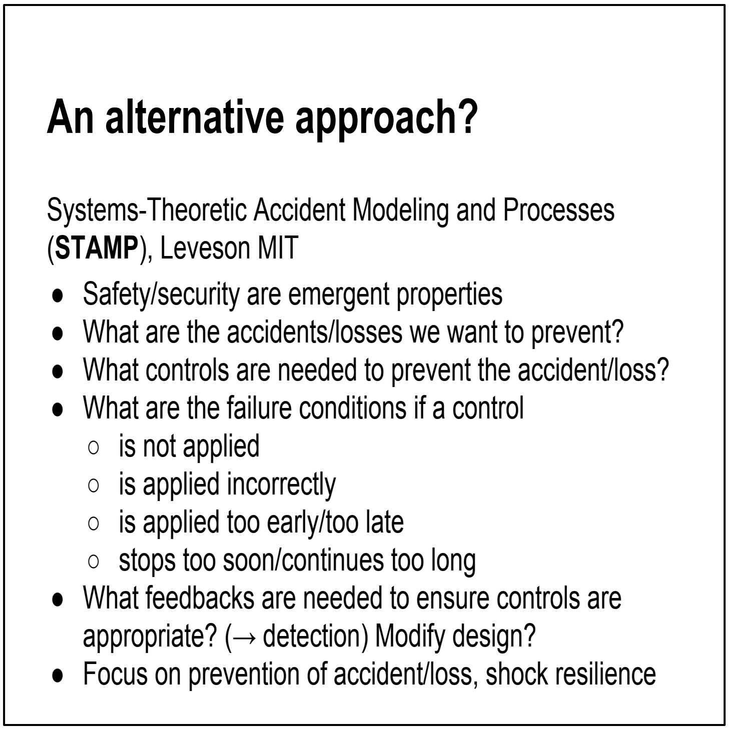
Wing, J. (2010)

There are some strategies for reducing the size of these graphs by emphasizing the most likely paths, for example (Wing, 2010):

1. Applying Google PageRank to identify the most likely nodes to be visited.
2. Allowing the analyst to decide which states are the most likely, or the most dangerous.

In practice, it’s hard to build such attack graphs automatically in a way relevant to the particular environment.

Another problem is that this approach is that it includes only technical vulnerabilities, while human vulnerability (e.g. to social engineering) is a critical factor in overall system security.



## An alternative approach?

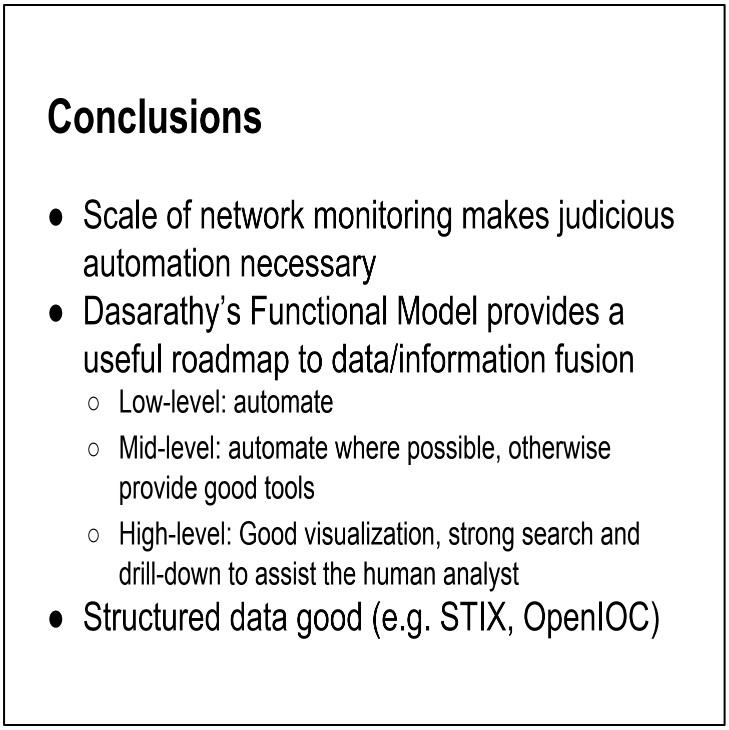
Systems-Theoretic Accident Modeling and Processes (**STAMP**), Leveson MIT

- Safety/security are emergent properties
- What are the accidents/losses we want to prevent?
- What controls are needed to prevent the accident/loss?
- What are the failure conditions if a control
  - is not applied
  - is applied incorrectly
  - is applied too early/too late
  - stops too soon/continues too long
- What feedbacks are needed to ensure controls are appropriate? (→ detection) Modify design?
- Focus on prevention of accident/loss, shock resilience

STAMP is a process that has been proven effective in the safety realm (inadvertent actions by benevolent actors), and may be appropriate to the security realm (deliberate actions by malicious actors).

The focus is on controlling vulnerability, and since in detection we're dealing with systems already in place, the focus has to be on risk mitigation (monitoring for violations of controls, and detection of accident/loss events). Security analysis systems could form part of the feedback, and hopefully can feed into an a continuous improvement process in the enterprise.

Projection becomes a process of estimating the most likely scenarios of accident/loss under the current situation, so we can adjust our actions to better protect against the most likely scenario. Development and discussion of enterprise and system vulnerability models based on STAMP is left for future work.



## Conclusions

- Scale of network monitoring makes judicious automation necessary
- Dasarathy’s Functional Model provides a useful roadmap to data/information fusion
  - Low-level: automate
  - Mid-level: automate where possible, otherwise provide good tools
  - High-level: Good visualization, strong search and drill-down to assist the human analyst
- Structured data good (e.g. STIX, OpenIOC)

Structured data is good because we can incorporate it into our models in an automated way. Extracting indicators from human-written PDFs is just not desirable or scalable.