

# Building the UA/ Eller/ MIS AZSecure Cybersecurity Analytics Program: My Journey

Hsinchun Chen, Ph. D.

Regents' Professor, Thomas R. Brown Chair

Director, AI Lab, AZSecure Cybersecurity Program

Fellow, ACM, IEEE, AAAS

University of Arizona

U of Maryland Seminar, October 1, 2021

**Funding Acknowledgements: NSF (SaTC, SFS, ACI)**



# Outline

- **Security Informatics & Analytics:** COPLINK, BorderSafe, Dark Web
- **Azsecure Cybersecurity Analytics:**
  - (1) **Dark Web Analytics** for studying international hacker community, forums, and markets;
  - (2) **Privacy and PII (Personally Identifiable Information) Analytics** for identifying and alleviating privacy risks for vulnerable populations;
  - (3) **Adversarial Malware Generation and Evasion** for adversarial AI in cybersecurity; and
  - (4) **Smart Vulnerability Assessment** for scientific workflows and OSS (Open Source Software) vulnerability analytics and mitigation.
- **Some Advice**

# Computational Design Science Research at UA/Eller/MIS AI Lab

- Applications/problems: digital libraries, search engines, biomedical informatics, healthcare data mining, security informatics, business intelligence, **cybersecurity analytics**
- Approaches: web collection/spidering, databases, data warehousing, data mining, text mining, web mining, **statistical NLP, machine learning, deep learning**, ontologies, social media analytics, interface design, information visualization, economic modeling, assessment
- Structure: **federal funding (NSF/DOD/NIH)**, director, affiliated faculty, post-docs, Ph.D./MS/BS students → **tech transfer, commercialization**
- Major phases: DLI → **COPLINK** → Dark Web → **AZSecure**

# Security Informatics & Analytics: COPLINK & Dark Web

D-Lib Magazine  
July/August 1998

ISSN 1082-9873

NSF/DARPA/NASA Digital Libraries Initiative

A Program Manager's Perspective

Stephen M. Griffin  
Division of Information and Intelligent Systems (IIS)  
Program Director: Special Projects Digital Libraries Initiative  
National Science Foundation  
Arlington, Virginia USA  
[sggriffin@nsf.gov](mailto:sggriffin@nsf.gov)



**Digital Government (DigitalGov)**

---

Program Solicitation

NSF 04-521

Replaces Document 02-156

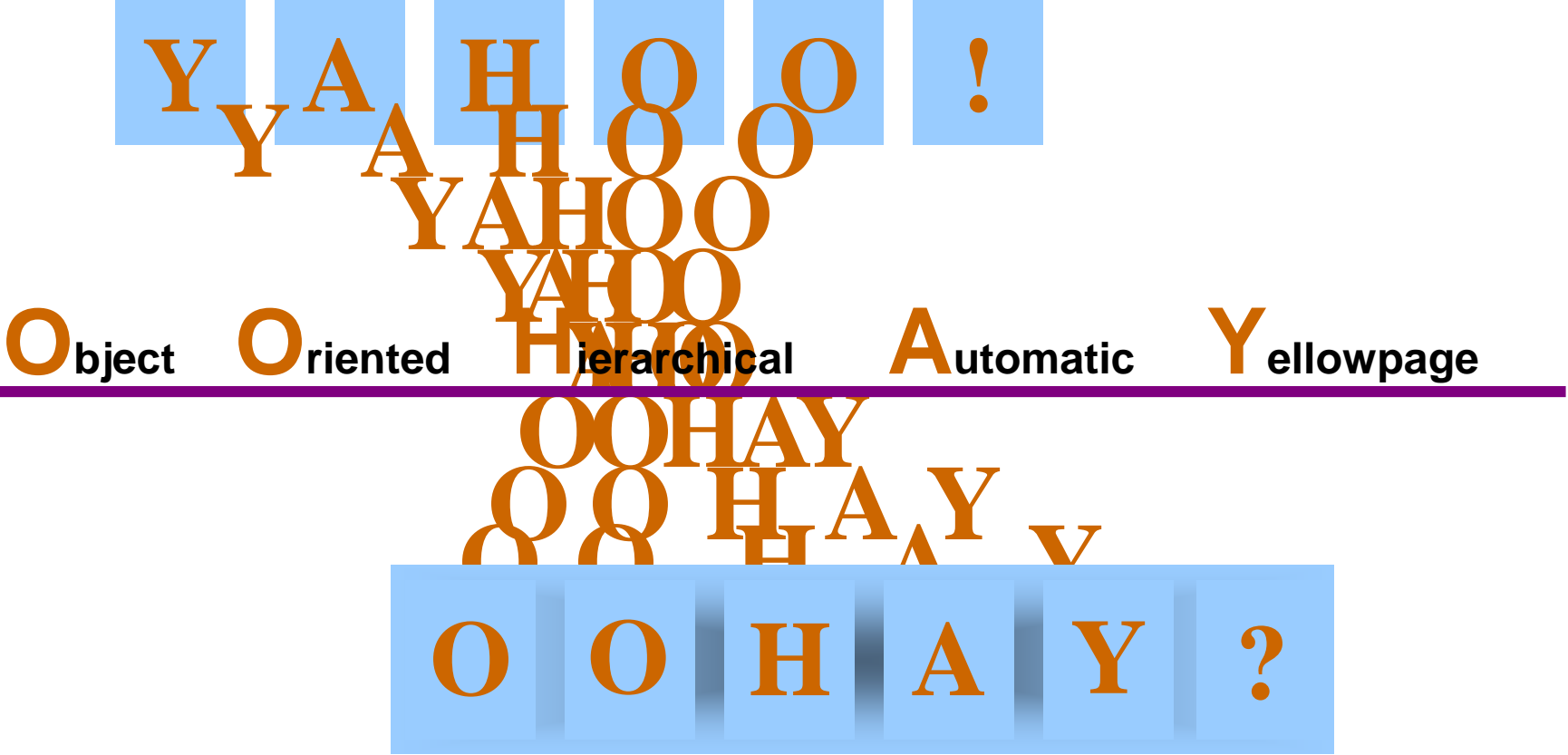


National Science Foundation

Directorate for Computer and Information Science and Engineering  
Division of Information and Intelligent Systems

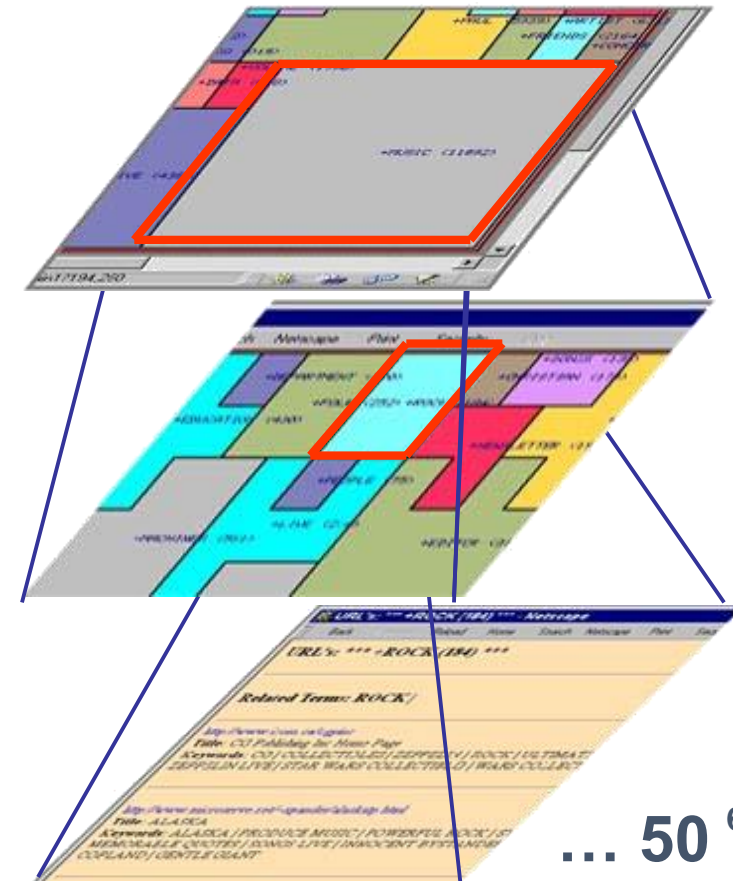
# DLI: Visualization Research in AI Lab

From YAHOO! To OOHAY?



# Visualization Research in AI Lab

## OOHAY: Visualizing the Web



# Cancer Map: 2M CancerLit articles, 1500 maps (OOHAY, DLI)

The image shows a screenshot of a web browser displaying a hierarchical navigation menu for 'Cancer Map'. The menu is structured as follows:

- 1 Visual Site Browser
- 2 Top level map
- 3 Diagnosis, Differential
- 4 Brain Neoplasms
- 5 Brain Tumors

The browser window shows the 'OOHAY Concept & Document Server' interface. The 'List of Found Documents' section displays the following information:

**OOHAY Concept & Document Server** Help

**List of Found Documents**

Total Found Documents: 83

1 2 3 4 Next

DOC UI: 99317761

**Authors:** Bloem BR | de Roos MA | de Beaufort AJ | Brouwer OF

**Title:** The stumbling toddler

**Source:** Ned Tijdschr Geneeskd; 143(23):1185-8 1999

**Central Concepts:** Brain Neoplasms | Cerebellar Ataxia | Gait | Medulloblastoma

**MeSH:** Anti-Anxiety Agents, Benzodiazepine | Case Report | Child, Preschool | Clorazepate Dipotassium | Diagnosis, Differential | Diarrhea | English Abstract | Female | Human | Male | Neurologic Examination | Recovery of Function | Tomography, X-Ray Computed | Treatment Outcome | Vertigo | Virus Diseases

1. **Abstract:** Four previously healthy children, two boys aged 5 and one boy and one girl aged 4 more or less acutely developed a stumbling gait. The causes varied from benign such as postviral acute cerebellar ataxia and benign paroxysmal vertigo to potentially life-threatening such as intoxication with benzodiazepines and medulloblastoma. Treatment led to complete or partial recovery. (Sub)acute balance disorders in previously healthy children can be due to cerebellar ataxia, vestibular disorders and abnormal proprioception. Ancillary investigations are warranted in case of gradually developing ataxia, accompanying neurological deficits, suspicion of intoxication, recurrent or familial ataxia, no spontaneous remission or even progression. In children with an isolated cerebellar ataxia without these features, ancillary investigations may be avoided, although in such cases careful follow-up remains necessary.

DOC UI: 96272136

# Taiwan Health Topic Map: 500K news articles

The image shows a 'Topic Map' interface with a sidebar on the left containing a list of health topics. The main area displays a 'Topic Map' for '直腸癌' (Colorectal Cancer) with sub-topics like '紅血球', '荷爾蒙', '直腸癌', '子宮內膜', and '腦下垂體'. Below this is a detailed view of an article with the following metadata:

- Authors:** ◎梁金銅台大醫院大腸直腸外科主治醫師
- Title:** 大腸直腸癌
- Source:** 《聯合報》
- Abstract:** 近年來因經濟的起飛，人口結構的老化，加上生活型態的改變，西式飲食的盛行，導致台灣地區之大腸直腸癌發生率及死亡率節節上揚。就死亡率而言，大腸直腸癌目前已是台灣地區因惡性腫瘤死亡人口的第三位，均僅次肝癌及肺癌。不論在我國或先進國家，大腸直腸癌已是今日公共衛生重要的一環。近年來有關大腸直腸癌的流行病學研究甚多，其中較具體的結論是遺傳與飲食。我們大概可以說家族一等親中若有人得到大腸直腸癌，則其一生中得相同癌症的機會約為一般人的三倍。目前公認纖維質食物攝取太少，以及攝取太多的肉類，由於會導致大便通過大腸的平均時拉長，所以致癌的機會也會大增。就大腸直腸癌病變而言，



# Global Security Impacts

- “War on terror” (Iraq and Afghanistan) surpassed cost of Second World War, \$5 trillion...Time Magazine
- Hacker costing \$1 trillion globally... President Obama



# From the Surface Web to the Dark Web

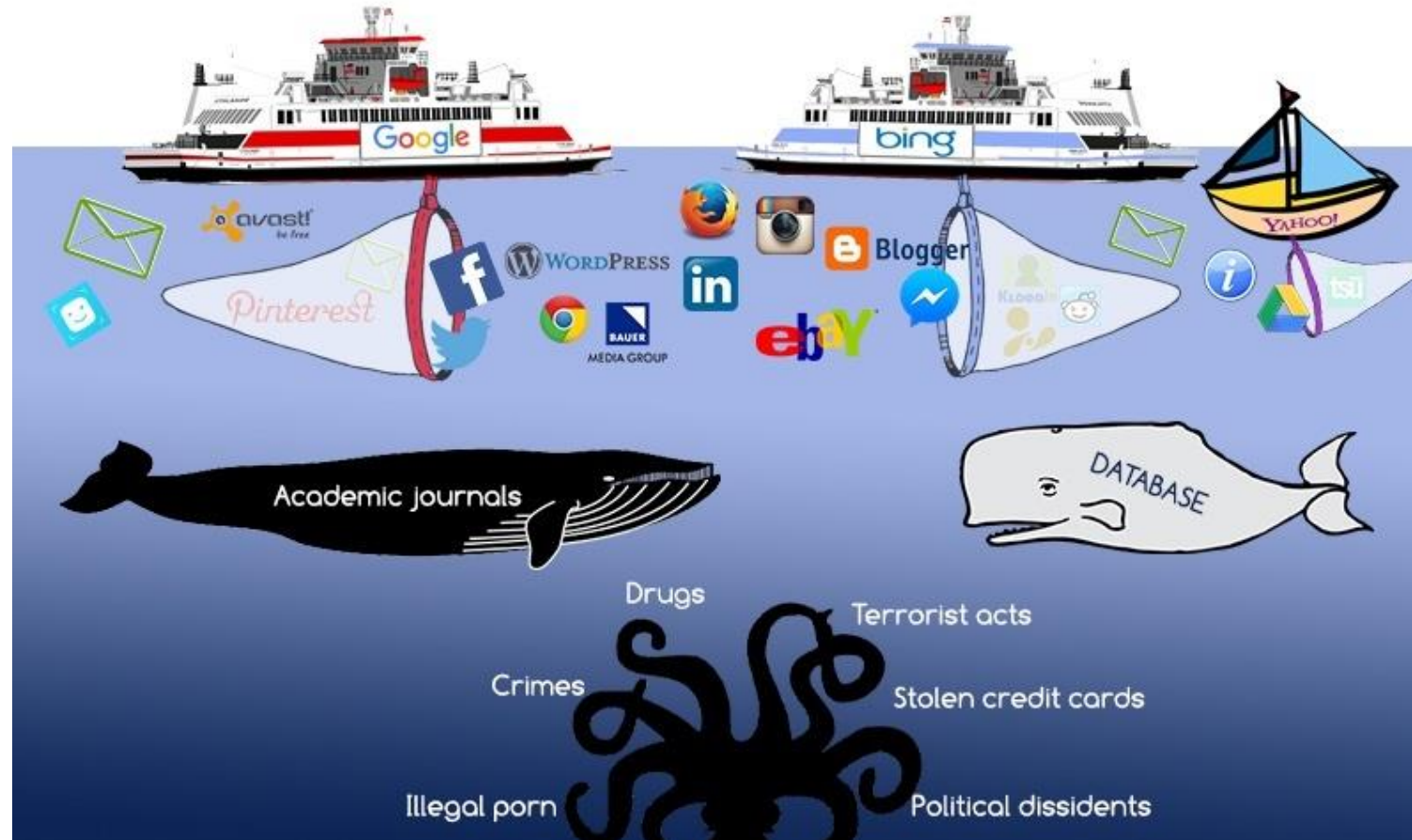
Surface Web

Deep Web

**Dark Web**

**DarkNet**

**Hacker Web**

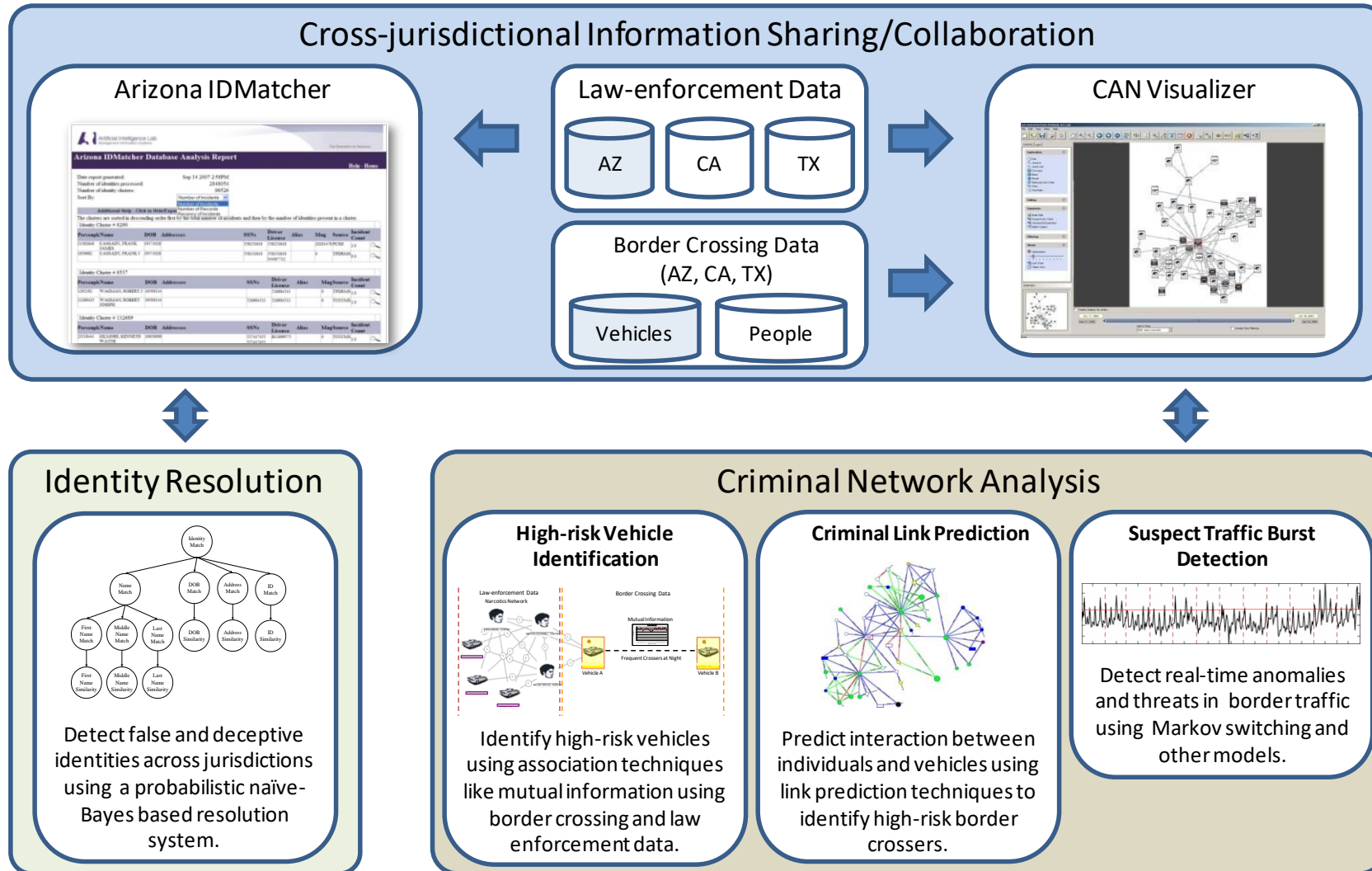


# COPLINK: Crime Data Mining (1997-2009)

The screenshot displays the COPLINK web application interface, which is used for crime data mining. The interface is divided into several main sections:

- Person Search Form:** Located on the top left, it allows users to search for individuals based on various criteria such as name (Last, First, Middle), race, sex, age, height, weight, hair, and eyes. There are also checkboxes for "Search Past Descriptions" and "Show All Aliases Matching Description".
- Person Search Results:** A table below the search form lists search results. Each row includes an "Add" button, a count, a name, a date of birth, sex, race, height, weight, hair, and eyes. For example, the first result is "ANNGNEURG, CARLOS" with a DOB of 08/29/1973.
- Person Details - TUCSON:** This section provides detailed information for a selected person, "AREVALLES, RICHMOND O". It includes a photo, agency information (TUCSON), local ID (C122574), SID (59335790), FBI number (6AKD75685), sex (M), hair (Brown), HT (507-601), race (White), eyes (Blue), and WT (120-200). It also lists aliases (ALBOT, RICHMOND; AREVALLES, RICHMOND; AREVALLES, SONNY; LYMBECK, RICHMOND; STINKO, RICHMOND O) and a date of birth (12/13/1974). A table of descriptions follows, with columns for count, race, sex, height, weight, hair, eyes, and date.
- Visualizer:** The bottom section of the interface features three main visualizations:
  - Minigraph:** A circular network graph showing relationships between entities, with nodes represented by small icons and lines connecting them.
  - Map:** A geographical map of Tucson, Arizona, with red dots indicating the locations of crime incidents. The map includes layers for Pima Boundary, City Areas, Landmarks, Washes, Lakes, Parks, Golf Courses, Wilderness, Runways, Prisons, Hospitals, and Schools.
  - Bar Chart:** A bar chart titled "Documents per Week (All)" showing the number of documents for each day of the week. The y-axis represents the number of documents (0 to 100), and the x-axis represents the days of the week (Sun, Mon, Tue, Wed, Thu, Fri, Sat).

# COPLINK Identity Resolution and Criminal Network Analysis



\* Only the grayed datasets are available to the AI Lab

# COPLINK: Crime Data Mining

ABC News April 15, 2003

Google for Cops: Coplink software helps police search for cyber clues to bust criminals

IBM i2 COPLINK

*Accelerating law enforcement investigations*

 Palantir (\$54B, IPO 2020)

Arts & Ideas  
The New York Times  
SATURDAY, NOVEMBER 2, 2002

## An Electronic Cop That Plays Hunches

Interconnecting Police Files Through New Computer System Helps Prosecutors in Sniper Case

By MINDY SINK

TUCSON, Oct. 28 — Officials building a case against the Washington-area sniper suspects are using a new investigative tool to help trace their movements across the country. It is an electronic-based system called Coplink, developed at an artificial intelligence laboratory here, that allows police departments to establish links quickly among their own files and to those of other departments.

During the 31 days in which suspects terrorized the area, investigators used everything from specialized ballistics testing to geographic and criminal profiles to track and identify suspects. Then, in what turned out to be the 14th hour of the pursuit, they finally reached out to Coplink. As it turned out, John Muhammad and Lee Malvo were arrested before it was fully installed, but now the post-arrest task force is using the system to help connect the dots.

All of the information that was collected — including that from other computer database systems like the Federal Bureau of Investigation's RapSheet — is now being downloaded into the Coplink database so that the accumulated data can be compared, said Robert Griffin, president of Knowledge Computing Corporation of Tucson, which is turning the prototype of the laboratory into a market product. "The more data you get, the better Coplink works," he said.

Coplink was designed by Hsinchun Chen, the director of the Artificial Intelligence Laboratory at the University of Arizona. "It is the Google for law enforcement," he said, referring to a search engine that can search through an array of related Web sites. "Things that a human can do intuitively we are getting the computer to do, too."

During the sniper investigation, which generated hundreds of thousands of tips, the number of potential cases to administer was daunting. "We were mounting a massive effort," said Lt. Mitch Montgomery of the Montgomery County police. "We had tactical resources, the military, federal, state and local law enforcement agencies and information technology using several products where each one of these had a role." So when the National Institute of Justice, the Justice Department's research and development arm, suggested that the sniper task force try Coplink, the officials agreed.

While no one is suggesting old-fashioned detective work is being replaced by machines, the idea behind Coplink is to provide a computer program that can save busy police officers precious time and sometimes even help solve cases. That's some-



Above, old-fashioned shoe leather still essential to the police. Left, Hsinchun Chen, the director of the Artificial Intelligence Laboratory at the University of Arizona, who has developed Coplink, a new investigative tool that he says can consolidate and analyze police data nationwide.

at the scene. "Sherry did it." The name Sherry was put into Coplink, and cross-referenced with the victim's personal data, and within minutes the records showed that the two men had been in prison together.

The program also allows users to look at lists of data or to create graphs and charts showing affiliations among different criminals.

At the moment, the Tucson Police Department is the only one in the country where Coplink is fully installed, although about a half-dozen other cities have begun to introduce Coplink into their existing computer systems. The cost of the program and

Chris Johnston for The New York Times

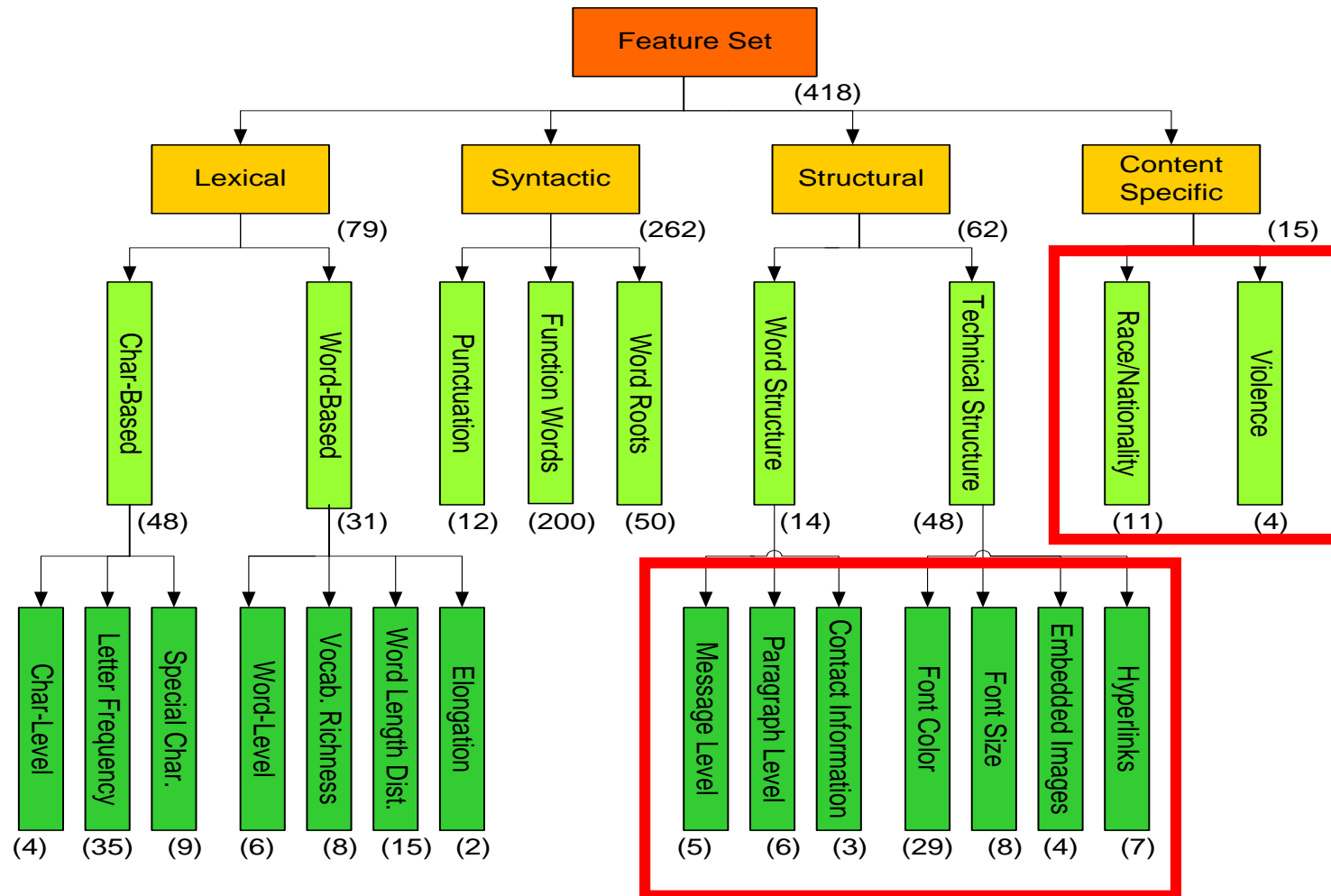


# Dark Web: Countering Terrorism (2003-2014)

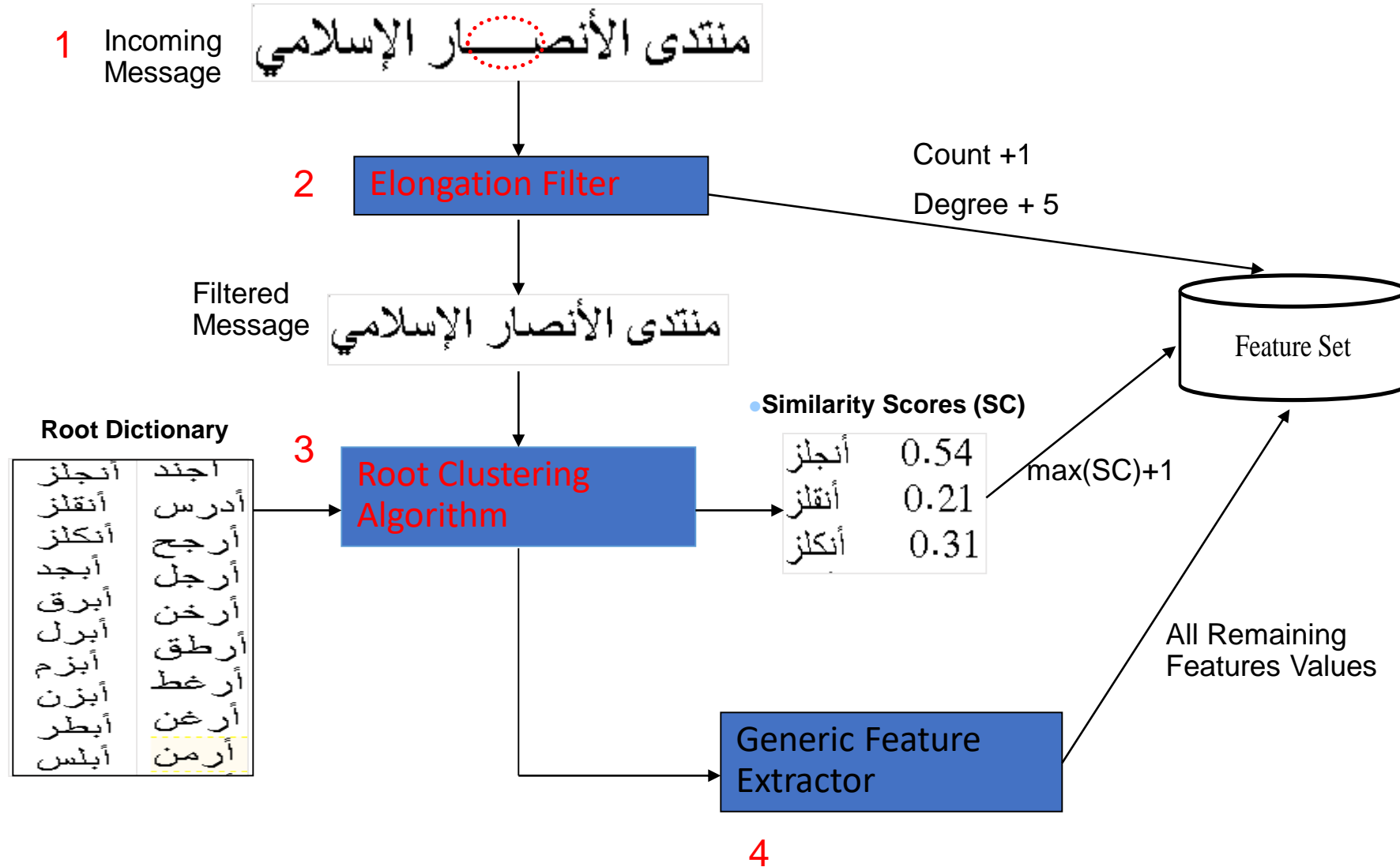
- Dark Web: Terrorists' and cyber criminals' use of the Internet
- Collection: Web sites, forums, blogs, YouTube, etc.
- 20 TBs in size, with close to 10B pages/files/messages (the entire LOC collection: 15 TBs)



# Arabic Writeprint Feature for Authorship Analysis

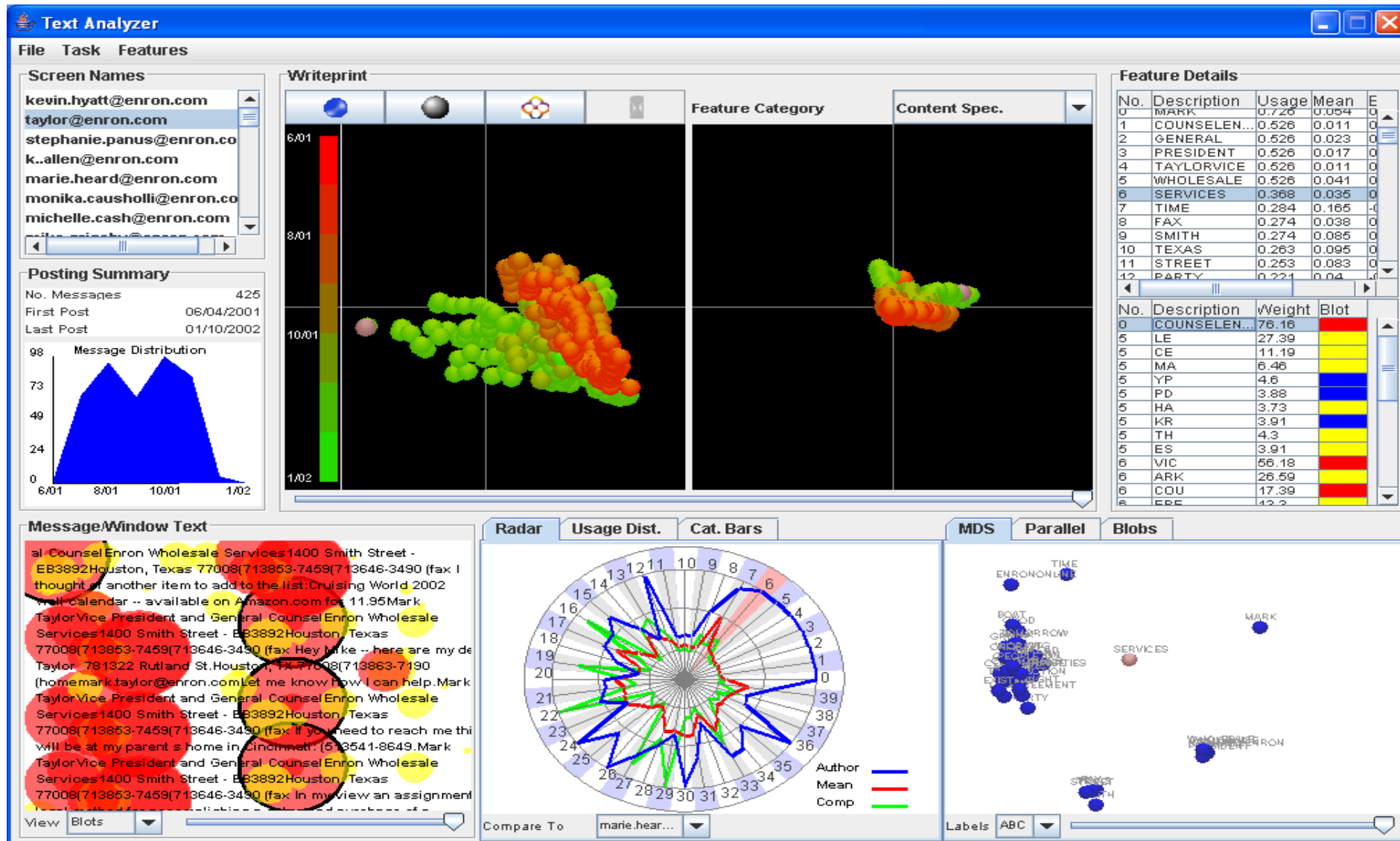


# Arabic Feature Extraction Component





# CyberGate (Abbasi, et al., MISQ, 2008)



# The Dark Web project in the Press



**Project Seeks to Track Terror Web Posts, 11/11/2007**



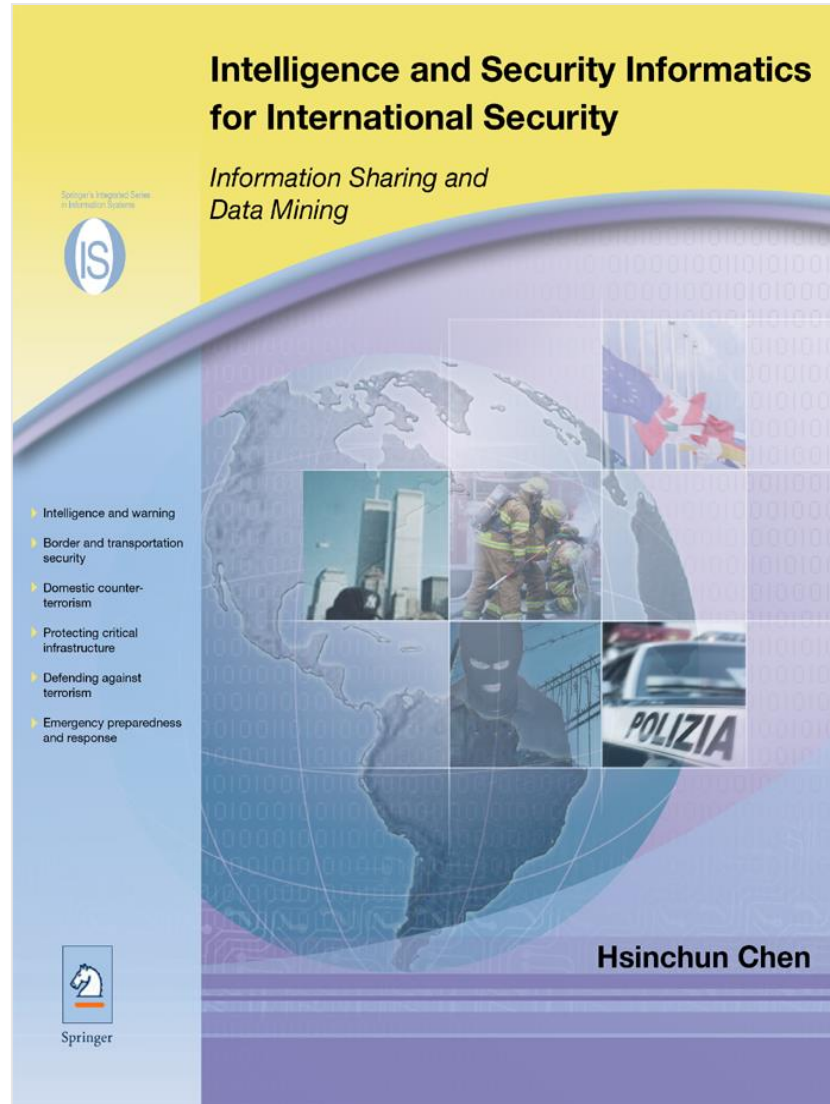
**Researchers say tool could trace online posts to terrorists, 11/11/2007**



**Mathematicians Work to Help Track Terrorist Activity, 9/14/2007**

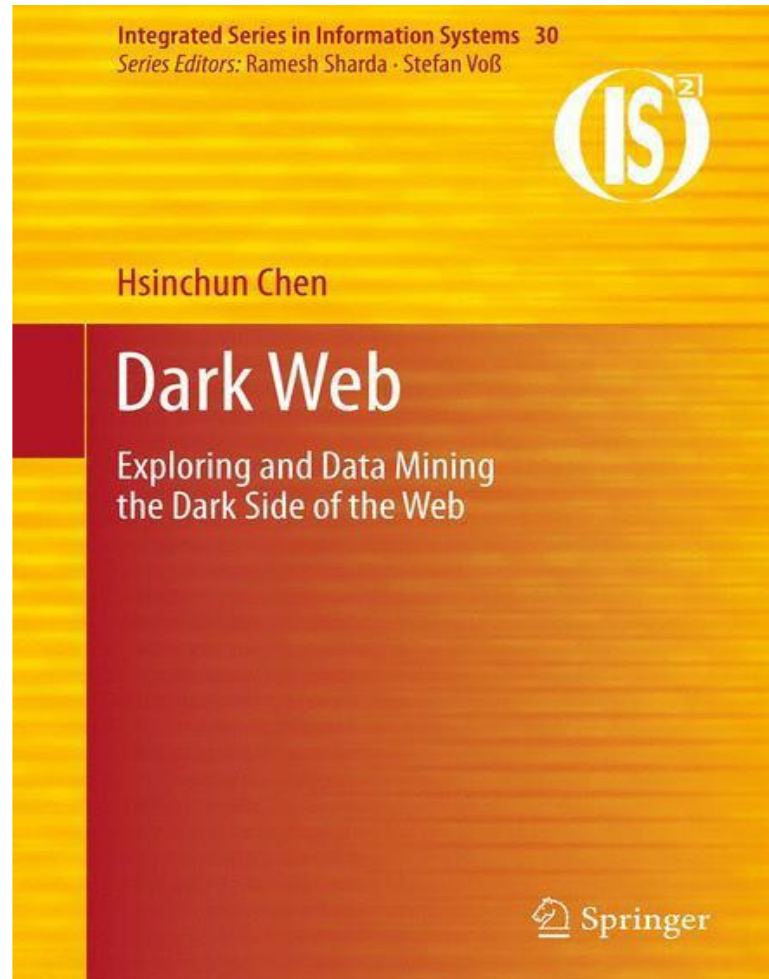


# ISI, Springer, 2006



- **Intelligence and Security Informatics (ISI)** (Chen, 2006)
- Data, text, and web mining
- From COPLINK to Dark Web
- **IEEE ISI, EISIC, PAISI → 4000+ scholars, since 2003**

# Dark Web, Springer, 2012



22 chapters, 451 pages, 150 illustrations (81 in color); Springer Integrated Series in Information Systems, 2012.

#### Selected TOC:

- Forum Spidering
- Link and Content Analysis
- Dark Network Analysis
- Interactional Coherence Analysis
- Dark Web Attribution System
- Authorship Analysis
- Sentiment Analysis
- Affect Analysis
- CyberGate Visualization
- Dark Web Forum Portal
- Case Studies: Jihadi Video Analysis, Extremist YouTube Videos, IEDs, WMDs, Women's Forums

# AZProtect (Abbasi, Chen, et al., 2010; MISQ best paper)



SPECIAL ISSUE

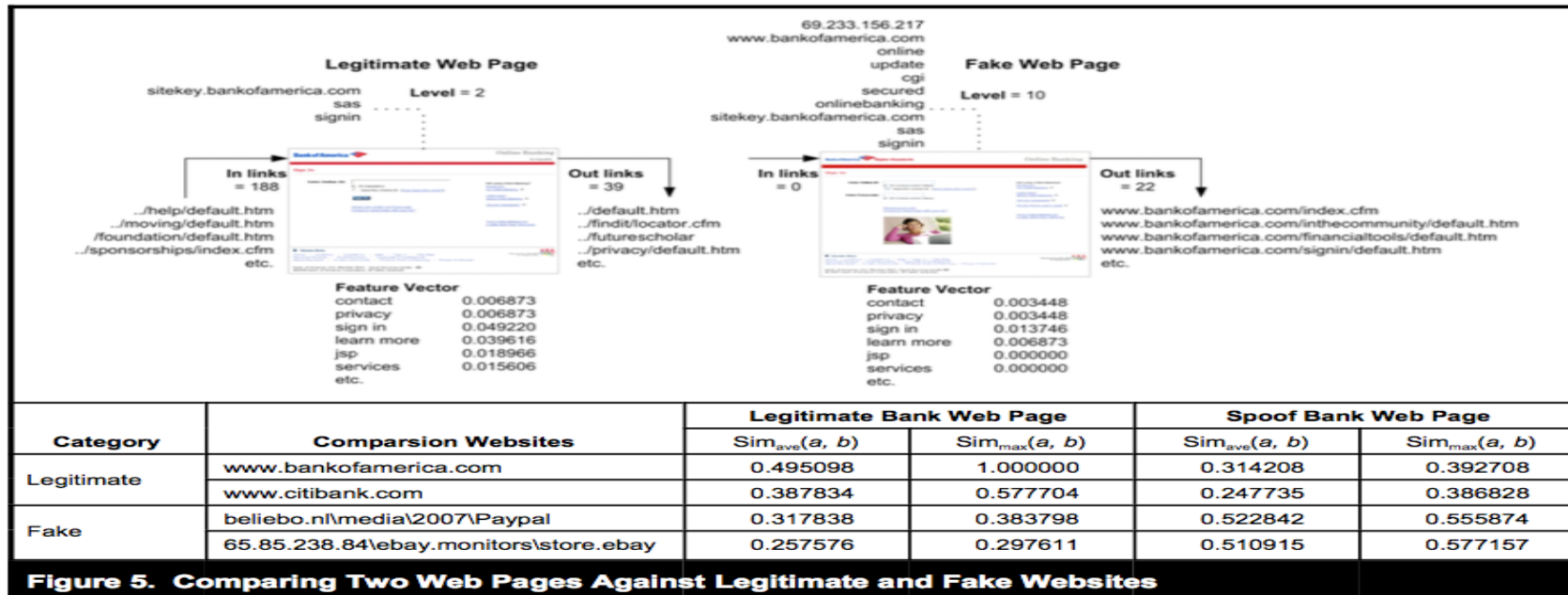
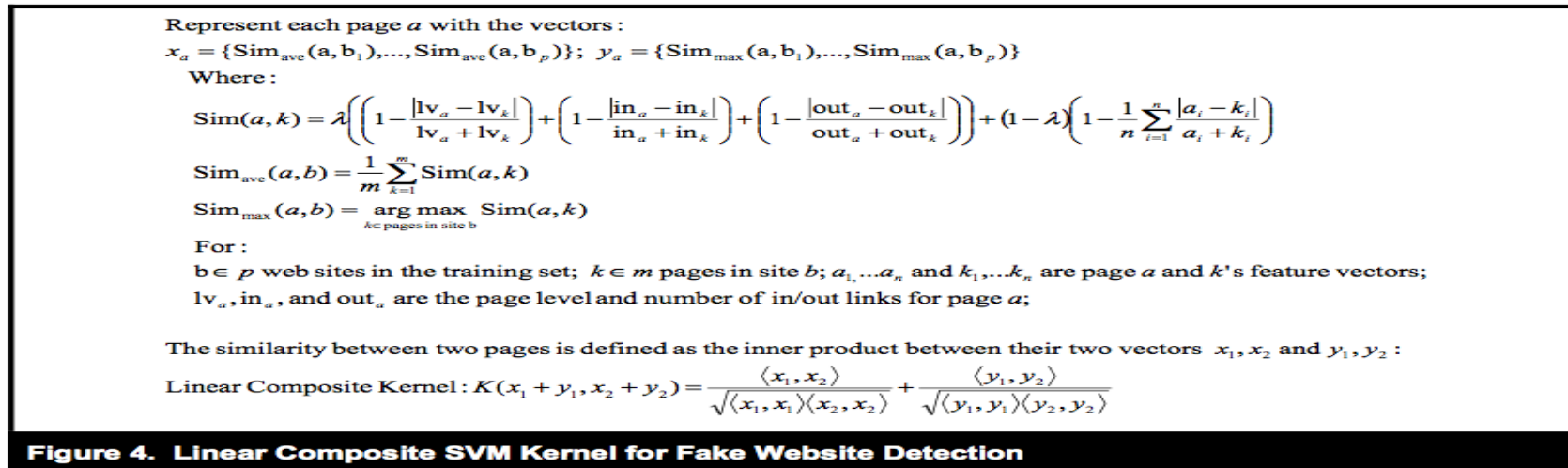
## DETECTING FAKE WEBSITES: THE CONTRIBUTION OF STATISTICAL LEARNING THEORY<sup>1</sup>

## Fraud Cues

**Table 2. Examples of Fraud Cues Incorporated in AZProtect**

Category	Attribute Group	Fraud Cues	Fake Site Type	Description	
Web page text	Word phrases	"member FDIC"	Concocted	References to Federal Deposit Insurance Corporation rarely appear in concocted bank websites.	
		"about FDIC"	Concocted	Outdated copyrights often appear in concocted websites.	
		"© 2000-2006"	Concocted	Concocted cargo delivery websites provide competitive phony estimates to lure customers. Legitimate sites typically offer estimates in-person through sales representatives.	
		"fee calculator"	Concocted	Fraudsters prefer to engage in online transactions. They rarely offer phone-based payment options.	
		"pay by phone" "call toll free"	Concocted	Concocted websites do not provide considerable support for returning customers since they generally do not have any.	
	Lexical measures	"payment history" "password management" "enter your account"	Concocted	Average sentence length	Sentences in concocted websites tend to be two to three times longer than ones in legitimate sites.
		Average word length, frequency of long words	Concocted	Average number of words per page	Concocted websites often contain concatenated words (e.g., "groundtransport" and "safebankingcenter"), resulting in unusually lengthy words.
		Average number of words per page		Concocted website pages are more verbose than legitimate sites—containing twice as many words per page, on average.	
	Spelling and grammar	"Adobe Acrobar"	Concocted	"fraudulent"	Concocted web pages contain many misspellings and grammatical mistakes.
		"recieve the"			
"think forwarder"					
"think forwarder"					
URLs	URL text	"HTTPS"	Concocted, Spoof	Fake websites rarely use the secure sockets layer protocol.	
		Random characters in URLs (e.g., "agkd-escrow," "523193pay")	Concocted, Spoof	Since fake websites are mass produced, they use random characters in URLs. It also allows new fake websites to easily circumvent lookup systems that rely on blacklists of exact URLs.	
		Number of slashes "/" in URL	Spoof	Spoof sites often piggy back off of legitimate websites or third party hosts. The spoofs are buried deep on these websites' servers.	
	Anchor Text	Errors in the URL descriptions (e.g "contactus")	Concocted	Anchor text is used to describe links in web pages. Concocted websites occasionally contain misspelled or inaccurate anchor text descriptions.	
Source Code	HTML and Javascript commands	"METHOD POST"	Concocted, Spoof	This HTML command is used to transmit data. It often appears in fake pages that are unsecured (i.e., "HTTP" instead of "HTTPS").	
		Image Preloading	Concocted, Spoof	This Javascript code, which is used to preload images to decrease page loading times, rarely appears in fake websites.	
	Coding style	"//""<!""=" "//..//"	Concocted, Spoof	Stylistic and syntactic elements in the source code can help identify automatically generated fake websites.	
Images	Image meta data	File name, file extension/format, file size	Concocted, Spoof	Fake websites often reuse images from prior fake websites. The file names, extensions, and file sizes can be used to identify duplicate images.	
	Image pixels	Pixel colors	Concocted, Spoof	If the image file name and format have been altered, image pixel colors can be used to identify duplicates.	
Linkage	Site level	Number of in/out links	Concocted, Spoof	Legitimate websites can contain links to and from many websites, unlike concocted and spoof sites.	
	Page level	Number of links, number of relative/absolute links	Concocted, Spoof	Fake websites tend to have fewer pages, and consequently, less linkage between pages. They also often use relative link addresses.	

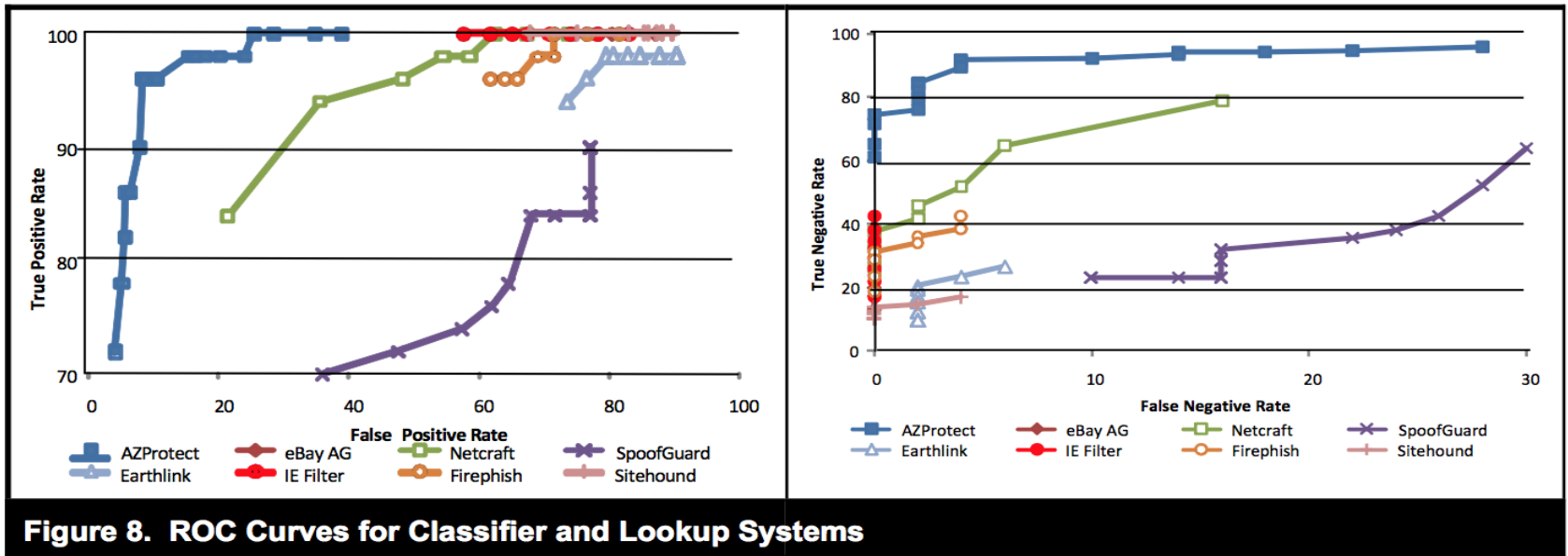
# Escrow Kernnel for Detecting Fake Web Sites



# Performance vs. Classifier and Lookup Systems

**Table 3. Performance Results (%) for Classifier and Lookup Systems**

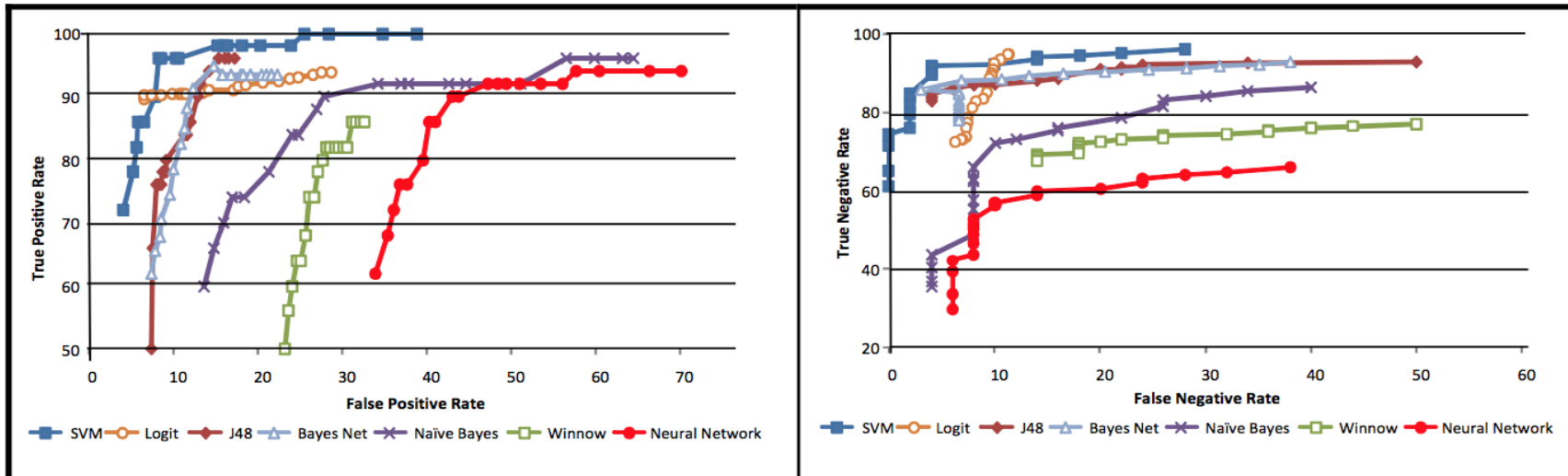
System		Overall Accuracy (n = 900)	Real Websites (n = 200)			Concocted Detection (n = 350)			Spoof Detection (n = 350)		
			F1	Prec.	Rec.	F1	Prec.	Rec.	F1	Prec.	Rec.
Classifier	AZProtect	92.56	85.21	76.29	96.50	91.82	97.74	86.57	97.12	97.97	96.29
	eBay AG	44.89	44.64	28.73	100.00	6.09	100.00	3.14	71.08	100.00	55.14
	Netcraft	83.00	72.13	56.74	99.00	82.28	99.19	70.29	92.52	99.34	86.57
	SpoofGuard	70.00	57.28	41.90	90.50	65.81	90.50	51.71	84.14	93.38	76.57
Lookup	EarthLink	42.67	43.55	27.87	99.50	15.75	96.77	8.57	61.27	99.36	44.29
	IE Filter	55.33	49.87	33.22	100.00	17.70	100.00	9.71	85.99	100.00	75.43
	FirePhish	54.89	49.63	33.00	100.00	12.84	100.00	6.86	87.09	100.00	77.14
	Sitehound	47.33	45.77	29.67	100.00	58.59	100.00	41.43	37.58	100.00	23.14



# Performance vs. Other ML Techniques

**Table 6. Performance Results (%) for Various Learning-Based Classification Techniques**

Learning Technique	Overall Accuracy (n = 900)	Real Websites (n = 200)			Concocted Detection (n = 350)			Spoof Detection (n = 350)		
		F1	Prec.	Rec.	F1	Prec.	Rec.	F1	Prec.	Rec.
SVM	92.56	85.21	76.29	96.50	91.82	97.74	86.57	97.12	97.97	96.29
Logistic regression	89.00	78.53	69.36	90.50	90.02	94.08	86.29	92.58	94.36	90.86
J48 Decision Tree	88.77	75.66	73.01	78.50	88.82	87.95	<b>89.71</b>	90.98	88.41	93.71
Bayesian Network	88.56	77.27	69.18	87.50	88.72	92.28	85.43	92.55	92.82	92.29
Naïve Bayes	77.67	63.12	49.86	86.00	86.49	91.14	82.29	77.47	89.51	68.29
Winnow	76.11	58.73	47.66	76.50	80.96	85.17	77.14	79.52	84.79	74.86
Neural Network	66.22	54.21	38.79	90.00	70.63	90.99	57.71	73.28	91.45	61.13



**Figure 9. ROC Curves for Various Learning Classifiers**



# AZSecure Cybersecurity Analytics Program (2010-present): SaTC, SFS, ACI

## Secure and Trustworthy Cyberspace (SaTC)

### PROGRAM SOLICITATION

NSF 21-500

REPLACES DOCUMENT(S):  
NSF 19-603



National Science Foundation

Directorate for Computer and Information Science and Engineering  
Division of Computer and Network Systems  
Division of Computing and Communication Foundations  
Division of Information and Intelligent Systems  
Office of Advanced Cyberinfrastructure

## CyberCorps(R) Scholarship for Service (SFS) Defending America's Cyberspace

### PROGRAM SOLICITATION

NSF 21-580

REPLACES DOCUMENT(S):  
NSF 19-521



National Science Foundation

Directorate for Education and Human Resources  
Division of Graduate Education

## Cybersecurity Innovation for Cyberinfrastructure (CICI)

### PROGRAM SOLICITATION

NSF 21-512

REPLACES DOCUMENT(S):  
NSF 19-514



National Science Foundation

Directorate for Computer and Information Science and Engineering  
Office of Advanced Cyberinfrastructure

# Azsecure Cybersecurity Analytics Program:

- (1) *Dark Web Analytics*** for studying international hacker community, forums, and markets;
- (2) *Privacy and PII (Personally Identifiable Information) Analytics*** for identifying and alleviating privacy risks for vulnerable populations;
- (3) *Adversarial Malware Generation and Evasion*** for adversarial AI in cybersecurity; and
- (4) *Smart Vulnerability Assessment*** for scientific workflows and OSS (Open Source Software) vulnerability analytics and mitigation.



**nature** doi:10.1038/nature16961

---

## Mastering the game of Go with deep neural networks and tree search

David Silver<sup>1\*</sup>, Aja Huang<sup>1\*</sup>, Chris J. Maddison<sup>1</sup>, Arthur Guez<sup>1</sup>, Laurent Sifre<sup>1</sup>, George van den Driessche<sup>1</sup>, Julian Schrittwieser<sup>1</sup>, Ioannis Antonoglou<sup>1</sup>, Veda Panneshelvam<sup>1</sup>, Marc Lanctot<sup>1</sup>, Sander Dieleman<sup>1</sup>, Dominik Grewe<sup>1</sup>, John Nham<sup>2</sup>, Nal Kalchbrenner<sup>1</sup>, Ilya Sutskever<sup>2</sup>, Timothy Lillicrap<sup>1</sup>, Madeleine Leach<sup>1</sup>, Koray Kavukcuoglu<sup>1</sup>, Thore Graepel<sup>1</sup> & Demis Hassabis<sup>1</sup>

**nature** doi:10.1038/nature24270

---

## Mastering the game of Go without human knowledge

David Silver<sup>1\*</sup>, Julian Schrittwieser<sup>1\*</sup>, Karen Simonyan<sup>1\*</sup>, Ioannis Antonoglou<sup>1</sup>, Aja Huang<sup>1</sup>, Arthur Guez<sup>1</sup>, Thomas Hubert<sup>1</sup>, Lucas Baker<sup>1</sup>, Matthew Lai<sup>1</sup>, Adrian Bolton<sup>1</sup>, Yutian Chen<sup>1</sup>, Timothy Lillicrap<sup>1</sup>, Fan Hui<sup>1</sup>, Laurent Sifre<sup>1</sup>, George van den Driessche<sup>1</sup>, Thore Graepel<sup>1</sup> & Demis Hassabis<sup>1</sup>

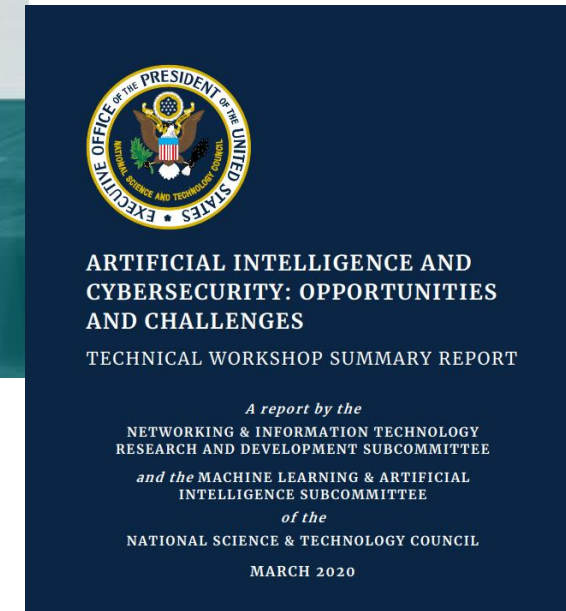
**AI & Deep Learning:** From AlphaGo to Autonomous Vehicles (2012-)



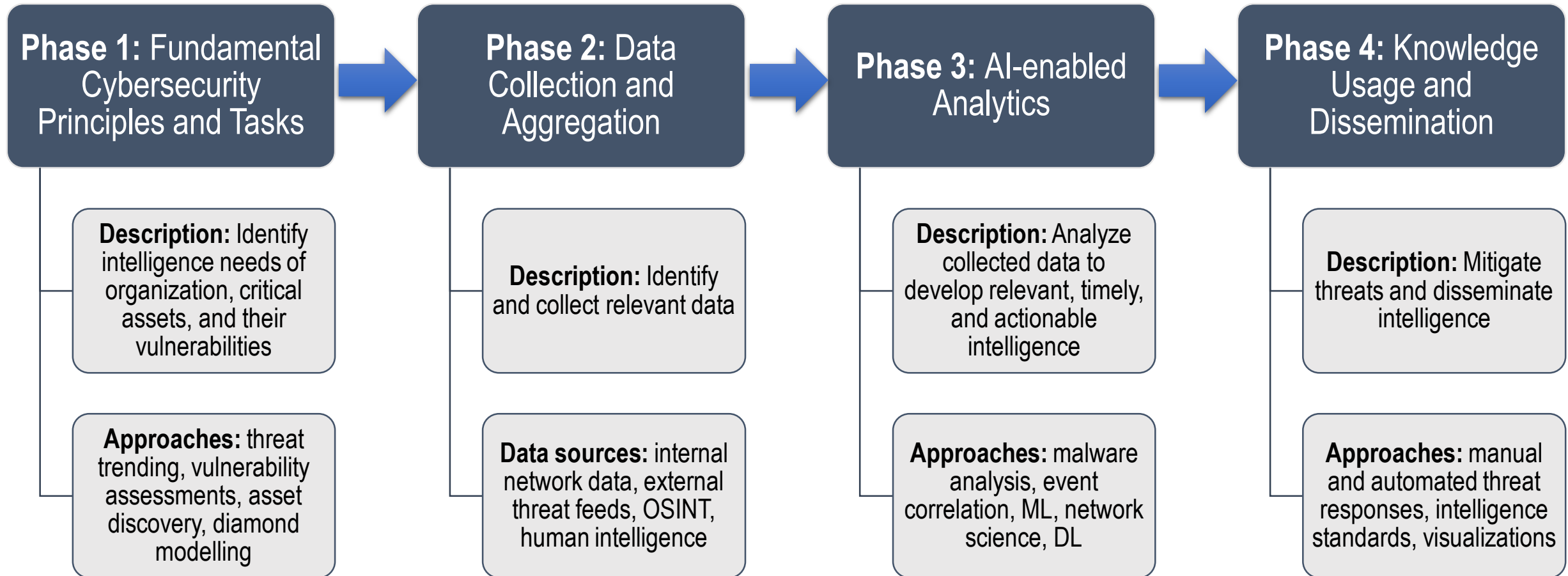
**Hacker Web, AZSecure** projects at UA/MIS AI Lab (2010-)

# AI and Cybersecurity

- AI and Cybersecurity → not just buzzwords!
  - Noted as a national security priority by NSF, NSTC, and NAS.
- Role of AI for Cybersecurity :
  1. Automate common cybersecurity tasks
  2. Identify patterns in large datasets



# AI for Cybersecurity – An Analytics Approach



## MOVING TOWARD BLACK HAT RESEARCH IN INFORMATION SYSTEMS SECURITY: AN EDITORIAL INTRODUCTION TO THE SPECIAL ISSUE

By: M. Adam Mahmood  
University of Texas at El Paso  
mmahmood@utep.edu

Mikko Siponen  
University of Oulu, Finland  
mikko.siponen@oulu.fi

Detmar Straub  
Georgia State University  
dstraub@gsu.edu

H. Raghav Rao  
State University of New York at Buffalo  
mgmtrao@buffalo.edu

T. S. Raghu  
Arizona State University  
raghu.santanam@asu.edu

### Introduction

The *MIS Quarterly* Special Issue on Information Systems Security in the Digital Economy received a total of 80 manuscripts from which we accepted nine for publication in the Special Issue. To introduce the readers to the special issue papers, we have chosen to digress from the tradition of summarizing the papers in-depth and, instead, would like to take this opportunity to encourage researchers to conduct

### Black Hats Versus White Hats Versus Grey Hats

What exactly is this white hat versus the black hat dichotomy? When making movies about the Old American West, filmmakers made a symbolic distinction at times between the good guys, wearing white hats, and the bad guys, wearing black hats. If, for the sake of our basic theme, we can adopt this distinction momentarily, we would like to go on to asseverate that the information systems field is heavily over-emphasizing research on white hats to the detriment of studies on black hats. It is easy to see how this would, quite naturally, occur. Scholars have better access to white hats, although even here, white hat managers do not typically want to share detailed information about their losses and have responded in this manner for some time (Hoffer and Straub 1989). Thus it is a readier access to data that has led information security researchers to gravitate toward white hat issues.

Whereas we could offer more extensive evidence of the prevalence of white hat IS research studies, a quick review of the papers in this special issue indicates that only the paper by Abbasi, Zhang, Zimbra, Chen, and Nunamaker attempts to empirically represent the activities of black hats, but even with this representation, we are at arm's length from black hat motivations and future dark plans.

We need to state unequivocally that our argument for more emphasis on the black hat type of research in no way diminishes the contributions of the white hat papers in this special



### Discoveries



### When hackers talk, this research team listens

Online conversations help fill critical gap in cybersecurity knowledge about attackers' motivations, possible targets



NSF-supported researchers have shed new light on how hackers communities interact.

[Credit and Larger Version](#)

October 8, 2015

Email Print Share



Hsinchun Chen leads a research project that explores the motivations of cyberattackers.  
[Credit and Larger Version](#)

Discoveries

[Search Discoveries](#)

[About Discoveries](#)

Discoveries by Research Area

[Arctic & Antarctic](#)

[Astronomy & Space](#)

[Biology](#)

[Chemistry & Materials](#)

[Computing](#)

[Earth & Environment](#)

[Education](#)

[Engineering](#)

[Mathematics](#)

[Nanoscience](#)

[People & Society](#)

[Physics](#)

# *Dark Web Analytics:* studying international hacker community, forums, and markets

- \* ACI, 2012-2017; SaTC 2013-2018; SFS-1, 2012-2018
- \* SaTC 2019-; SFS-2, 2019-

## Secure and Trustworthy Cyberspace (SaTC)

### PROGRAM SOLICITATION

NSF 21-500

### REPLACES DOCUMENT(S):

NSF 19-603



National Science Foundation

Directorate for Computer and Information Science and Engineering  
Division of Computer and Network Systems  
Division of Computing and Communication Foundations  
Division of Information and Intelligent Systems  
Office of Advanced Cyberinfrastructure

## CyberCorps(R) Scholarship for Service (SFS)

### *Defending America's Cyberspace*

### PROGRAM SOLICITATION

NSF 21-580

### REPLACES DOCUMENT(S):

NSF 19-521



National Science Foundation

Directorate for Education and Human Resources  
Division of Graduate Education

## Cybersecurity Innovation for Cyberinfrastructure (CICI)

### PROGRAM SOLICITATION

NSF 21-512

### REPLACES DOCUMENT(S):

NSF 19-514



National Science Foundation

Directorate for Computer and Information Science and Engineering  
Office of Advanced Cyberinfrastructure

# Hacker Web



**XekSec** Администрация

**Exploit name**: Mozilla Firefox 3.5.3 Local Download Manager Exploit

**Code to execute exploit**

```
#!/bin/sh
# We can the exploit to replace openssl-1.0.1.tar.gz with a modified version
# Administrator installs our openssl-1.0.1 with our modifications.

The download history will still show the name of the site that supplied the original file
filename even when the target user opened the our replacement file instead.

Conditions that have to be met for exploitation to succeed:
1. The ability to write in the temporary file directory, "/tmp" by default on Linux
   (shell, ftp, etc with write permissions could be helpful for making this work remotely)
2. The target user chooses to download the file and chooses the "Open with" preferences
3. The target user also has to double click the file in the download manager (in previous
   versions, the file opened automatically, as normal behavior, but that can no longer be
   relied on).

Firefox on Windows has slightly different results. I found during testing that when the
   the right file will be opened. Although unreliable, we were able to get the history of the
   manager to show the replacement file and it will be opened if the user chooses to open it.
   Replication on Windows would be limited anyway due to the fact that you don't usually
   have access to do local things on Windows as its fairly common on Linux. On Linux it is
   also an replacement file to be kept in history when using this exploit, which can be useful
   for the exploit when you don't want the target to think anything much is out of the ordinary.

openssl-1.0.1.tar.gz:rsLocalFileCommon.cpp => LINES (38*274)

#C_METHODS#
naLocalFile::CreateOnDisk(PWDyn32 type, &DInfo2 wstrbutee)
{
    wstrbutee ov;
    wstrbutee longName;
}

void na_LocalFile::
naLocalFile::CreateOnDisk(PWDyn32 type, &DInfo2 wstrbutee,
    ov = GetPath(pathName);
}

naLocalFile::CreateOnDisk(PWDyn32 type, &DInfo2 wstrbutee,
    ov = GetPath(pathName);
}
```

Forum post with source code to exploit Mozilla Firefox 3.5.3

**airwawekz** Junior Member

**Thread title**: Tutorial] Malicious Documents – PDF Analysis in 5 steps

**Author related information**

**Instructions on how to create malicious documents**

**Tutorial on how to create malicious documents**

Miss mailing or targeted campaigns that use common files to host or exploit code have been and are a very popular vector of attack. In other words document received via e-mail or opened through a browser plug-in. In regards to malicious PDF files the security industry saw a significant increase of half of 2008 which might be related to Adobe Systems release of the specifications, format structure and functionality of PDF files.

Most enterprise networks perimeters are protected and contain several security filters and mechanism that block threats. However a malicious PDF is very successful passing through Firewalls, Intrusion Prevention Systems, Anti-spam, Anti-virus and other security controls. By reaching the victim mail leverage social engineering techniques to lure the user to click/open the document. Then, for example, if the user opens a PDF malicious file, it typically exploits a vulnerability when Adobe Reader parses the crafted file. This might cause the application to corrupt memory on the stack or heap causing as shellcode. This shellcode normally downloads and executes a malicious file from the Internet. The Internet Storm Center Handler Bojan Zdrnja was one of these shellcodes. In some circumstances the vulnerability could be exploited without opening the file and just by having a malicious file on the hard drive.

Tutorial on how to create malicious documents

**Post Date**: 08-08-2014

**Attachment Name**: Blackpos.rar

**Description of Attachment**: hi guys, I found the subject published, these are two grabbers who already know

**BlackpOs**: <http://i.imgur.com/vfRkUqGE.png>

**Dexter v2**: <http://i.imgur.com/gYmjfkC.png>

**File Type**: rar

**Blackpos.rar**: 5.4 KB, 143 views

Forum post with BlackPOS malware attachment.





# Selected data breaches in 2014

Victim	Date	Ramification
Target	2013.12	<b>40M</b> credit/debit cards; <b>70M</b> customer records; 46% drop in annual profits ( <b>seller: Rescator</b> )
Neiman Marcus	2014.3	282K credit/debit cards
Sally Beauty	2014.3	25K credit/debit cards
P.F. Chang	2014.6	8 month of customer data from 33 stores
J.P. Morgan Chase	2014.8	<b>83M</b> accounts
UPS	2014.8	51 stores customers
Dairy Queen	2014.9	395 store systems
Home Depot	2014.9	<b>56M</b> credit/debit cards
Jimmy Jones	2014.9	216 store systems
Staples	2014.10	51 store systems

**Yahoo confirms: hackers stole 500 million account details in 2014 data breach**

Boohoo for Yahoo. State-sponsored attacker blamed for hack as users told to change passwords.

Graham Cluley | September 22, 2015 8:01 pm | Filed under: Data loss, Yahoo | 21

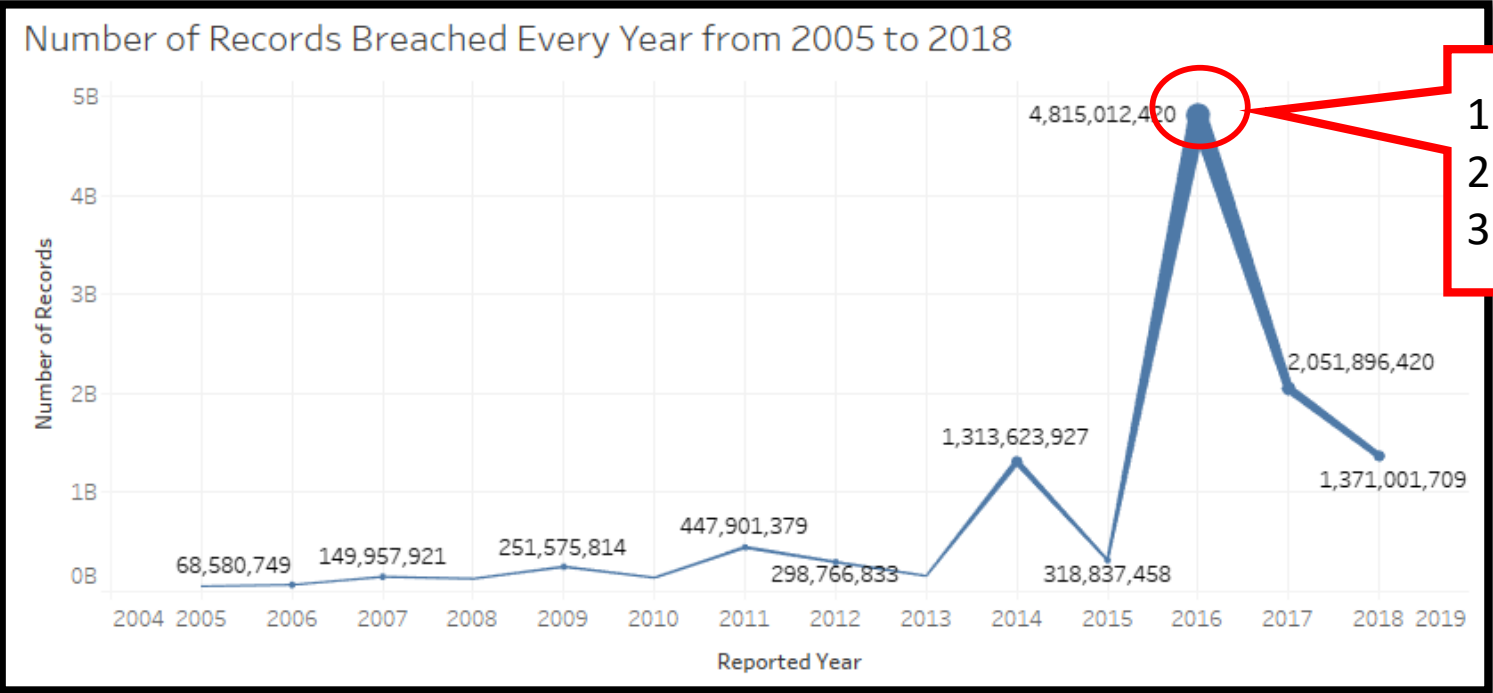
457 SHARES [Share on Twitter](#) [Share on Facebook](#) [+](#)



**Are your data breached? Do you even know?**

# Data Breaches since 2005 (FTC, Clearinghouse, 2019)

- # of records breached: 11,582,808,013
- # of data breaches: 9,071



## 2016 Data Breach

- 1. Yahoo! : 3.5B user accounts
- 2. FriendFinder : 412M user accounts
- 3. MySpace : 360M passwords

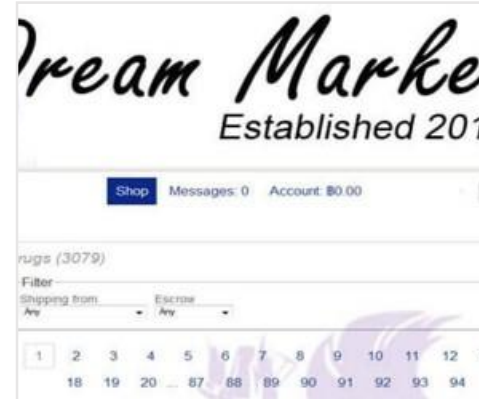
# Hacker Community Platforms – “Know your enemy”

## Hacker Forums



Discussion board allowing hackers to freely share malicious tools and knowledge

## DarkNet Markets



Markets facilitating the sale of illicit goods (e.g., new exploits, drugs, weapons)

## Carding Shops



Shops selling sensitive information (e.g., credit cards, SSN's)

## IRC Channels



Plain-text IM service commonly used by hacktivist groups (e.g., Anonymous)

US → cybercrime and general hacking  
Russia → underground economy, financial fraud  
China → cyberwarfare content

## DICE-E: A FRAMEWORK FOR CONDUCTING DARKNET IDENTIFICATION, COLLECTION, EVALUATION WITH ETHICS<sup>1</sup>

Victor Benjamin

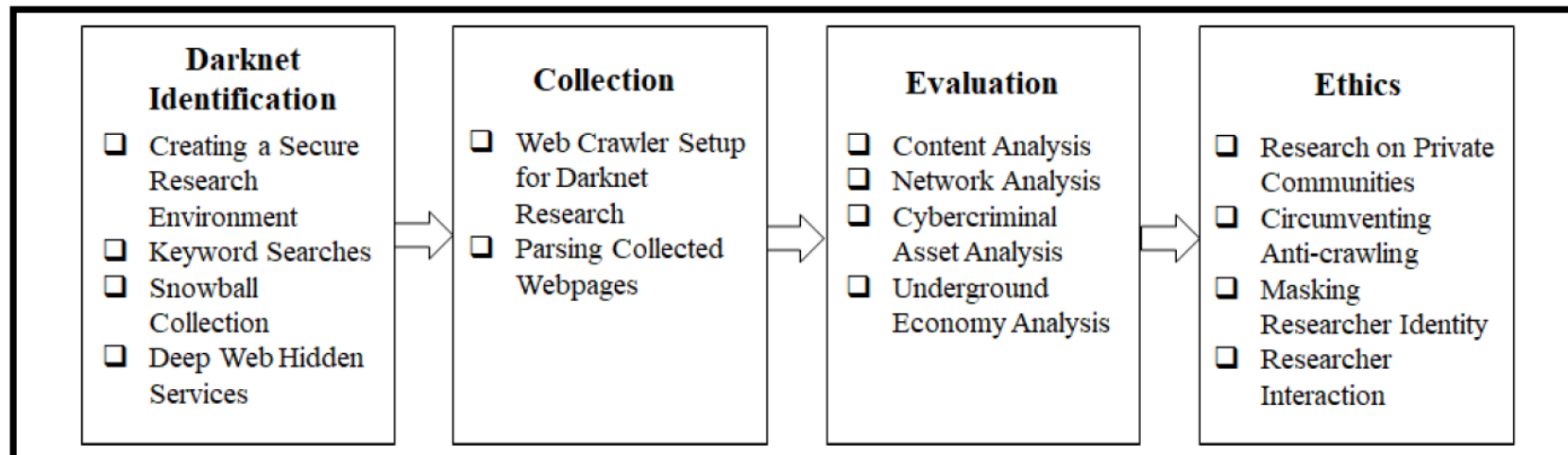


Figure 2. The DICE-E Framework

# Identify Hacker Assets/Tools

Sagar Samtani (JMIS, January 2018)



Journal of Management Information Systems



ISSN: 0742-1222 (Print) 1557-928X (Online) Journal homepage: <http://www.tandfonline.com/loi/mmis20>

---

**Exploring Emerging Hacker Assets and Key Hackers for Proactive Cyber Threat Intelligence**

Sagar Samtani, Ryan Chinn, Hsinchun Chen & Jay F. Nunamaker Jr.

# Hacker Asset/Tool Examples

03-07-2014

[Delphi] Noob Botnet Construct ← Exploit Name

just a old snippet by me which shows one of a possible construct to code a bot in delphi. enjoy it 😊

Code:

```
program autoStart;  
  
uses  
  Windows, Registry, Sysutils, ShellAPI, TLHelp32, WinTypes, Messages, WinProcs, WinInet, URLMon;  
  
var  
  str, OldName, NewName, NewDir, sPath, dURL: String;  
  Overwrite, Idle, AntiRE: Boolean;  
  i: Integer;  
  
{ -----Installation Konfiguration----- }  
Interval: Integer = 30; // Angabe in Sekunden  
  
const  
  sAutostartName = 'AutoStartTest'; // Name des Autostarteintrags  
  sFolderName = '\\2013'; // Ordnername Backslash muss davor stehn  
  sMutex = ''; // Mutex Name  
  sFilename = 'bin.exe'; // Dateiname  
{ ----- }  
  
procedure DoAppToRun(RunName, AppName: string);  
var  
  Reg: TRegistry;  
begin
```

Code to Execute Exploit

Ruffly  
Senior Member  
Join Date: Apr 2010  
Location: Germany  
Posts: 247

08-08-2014 ← Post Date

Pos  
hi guys, I found the subject published, these are two grabbers who already know

BlackpOs  
<http://i.imgur.com/vRKLqGGE.png>

Dexter v2  
<http://i.imgur.com/gYmjfkC.png>

[?] Senior Member  
Join Date: Dec 2010  
Location: Russia  
Posts: 165

Attachment Name

Attached Files  
File Type: rar Blackpos.rar 5.4 KB, 143 views

Description of Attachment

Figure 1. Forum post with source code to create botnets

Figure 2. Forum post with BlackPOS malware attachment

airwawekz Junior Member  
Join Date: Oct 2014  
Posts: 1  
Like (Stats)  
Mentioned: 0 Post(s)  
Tagged: 0 Thread(s)  
Quoted: 0 Post(s)

Tutorial] Malicious Documents -- PDF Analysis in 5 steps ← Thread title

PDF

Author related information

Instructions on how to create malicious documents

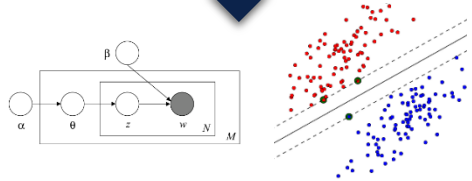
Mass mailing or targeted campaigns that use common files to host or exploit code have been and are a very popular vector of attack. In other words document received via e-mail or opened through a browser plug-in. In regards to malicious PDF files the security industry saw a significant increase of half of 2008 which might be related to Adobe Systems release of the specifications, format structure and functionality of PDF files.

Most enterprise networks perimeters are protected and contain several security filters and mechanism that block threats. However a malicious PDF is very successful passing through Firewalls, Intrusion Prevention Systems, Anti-spam, Anti-virus and other security controls. By reaching the victim malware leverage social engineering techniques to lure the user to click/open the document. Then, for example, if the user opens a PDF malicious file, it typically exploits a vulnerability when Adobe Reader parses the crafted file. This might cause the application to corrupt memory on the stack or heap causing as shellcode. This shellcode normally downloads and executes a malicious file from the Internet. The Internet Storm Center Handler Bojan Zdrnja writes of these shellcodes. In some circumstances the vulnerability could be exploited without opening the file and just by having a malicious file on the hard drive.

Figure 3. Tutorial on how to create malicious documents

# AZSecure Hacker Assets Portal System

## Data Collection and Analytics

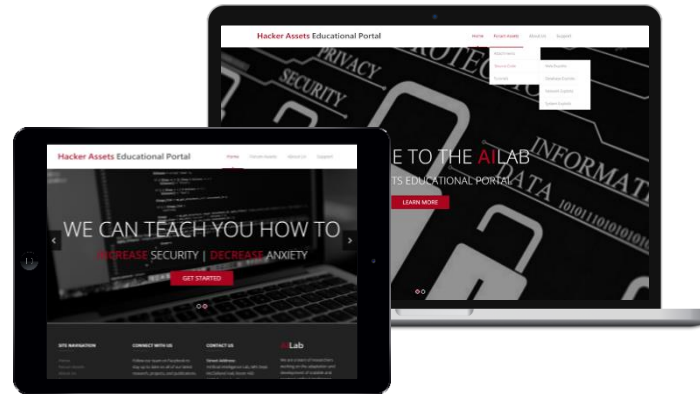


Latent Dirichlet Allocation (LDA) and Support Vector Machine (SVM) Analytics

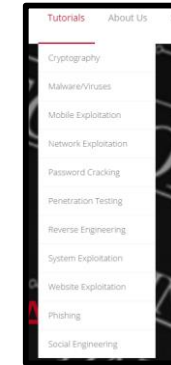


987 tutorials, 15,576 source code, and 14,851 attachments

## Web Hosting and Access



## System Functionalities



Browsing

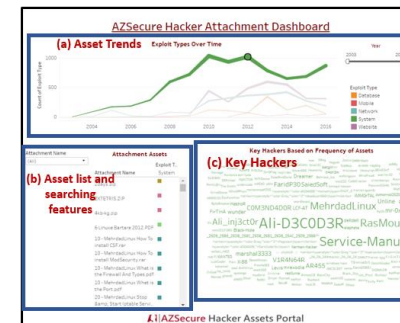


Searching

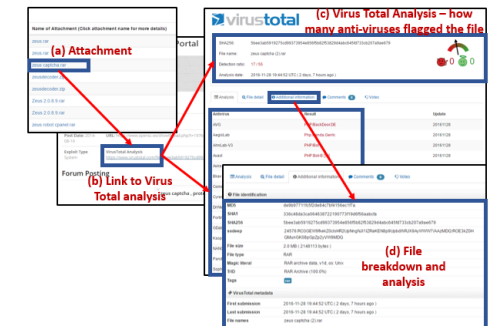


Downloading

## System Analytics



Cyber Threat Intelligence Dashboard



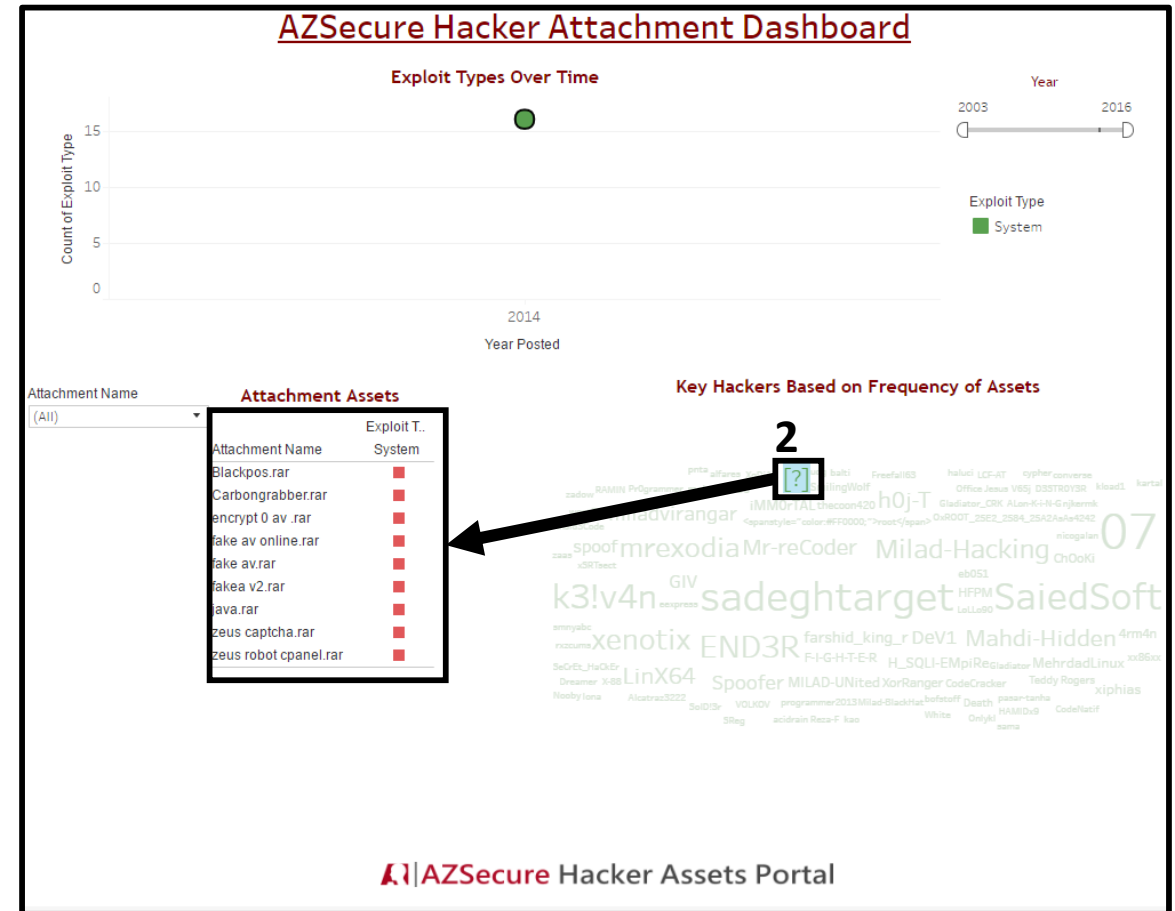
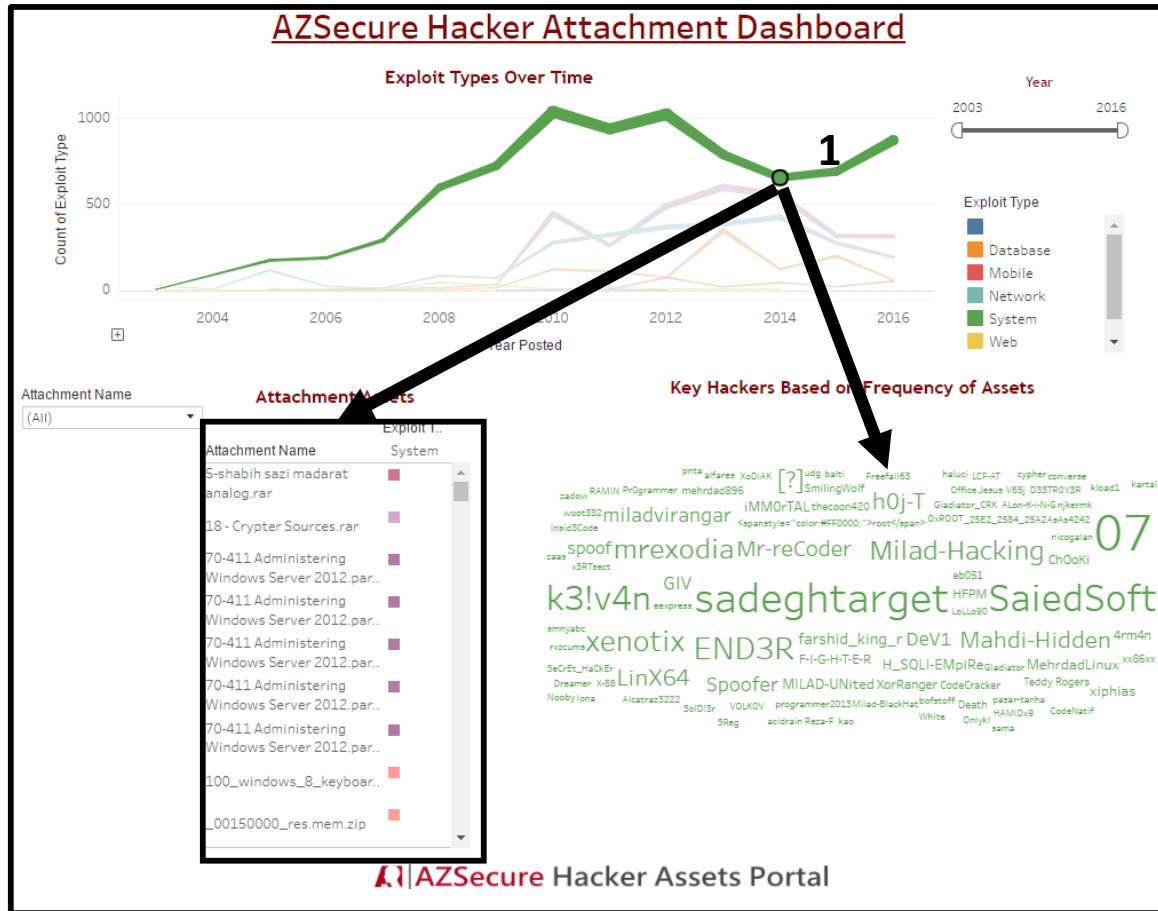
VirusTotal Malware Analysis

# AZSecure Hacker Assets Portal (**English, Russian, Arabic**)

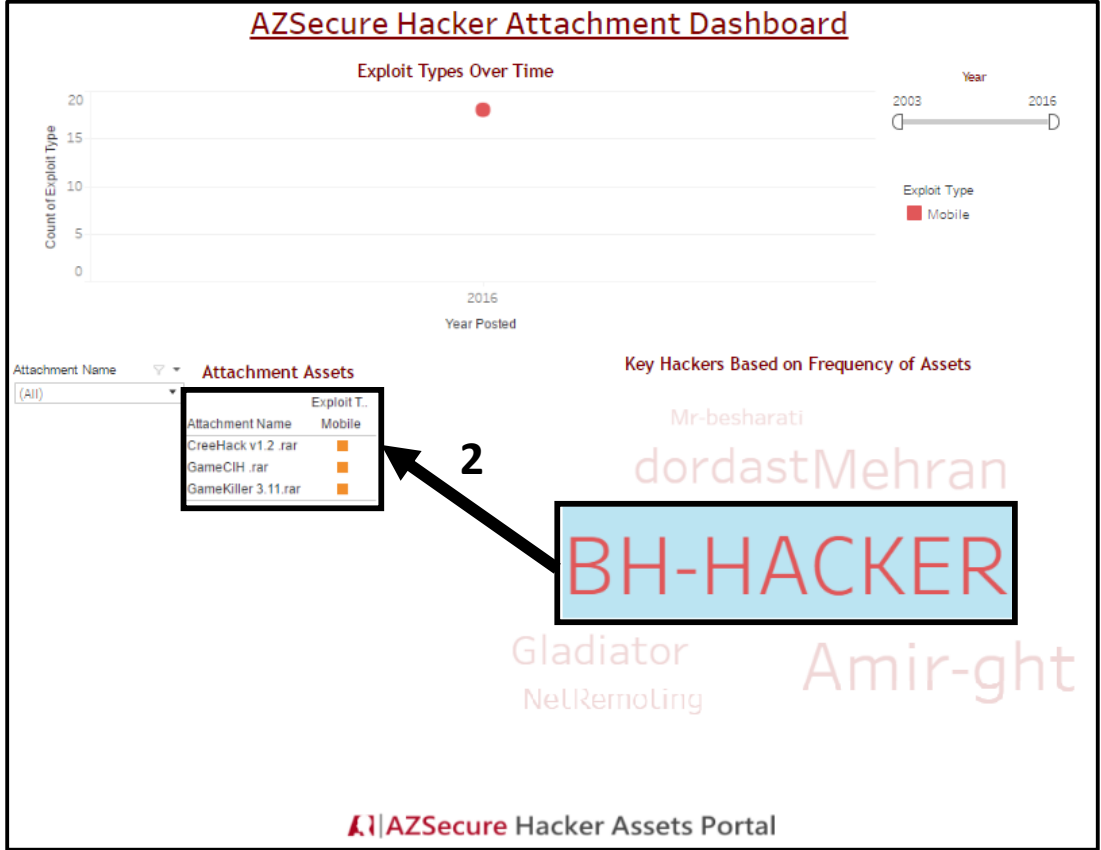
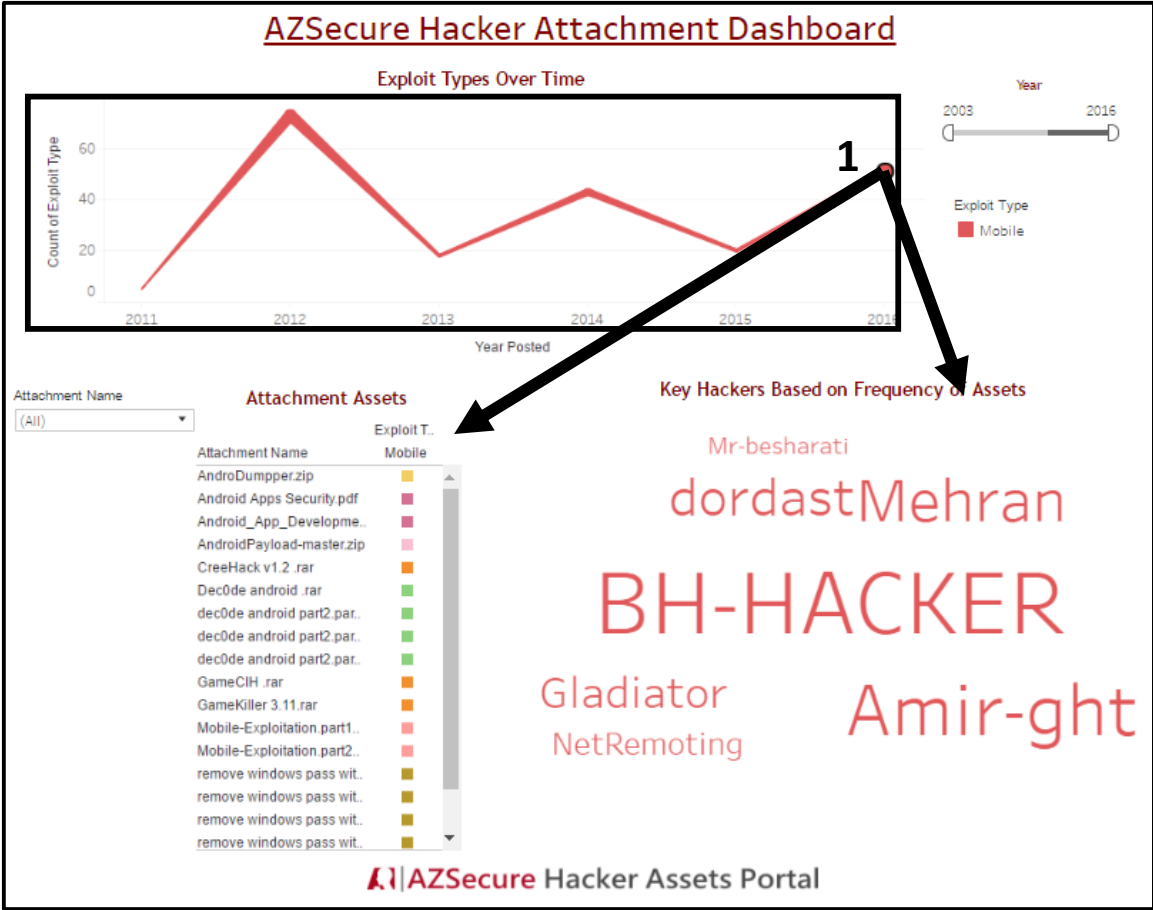
Forum	Language	Date Range	# of Posts	# of Members	# of source code	# of attachments	# of tutorials
OpenSC	English	02/07/2005-02/21/2016	124,993	6,796	2,590	2,349	628
Xeksec	Russian	07/07/2007- 9/15/2015	62,316	18,462	2,456	-	40
Ashiyane	Arabic	5/30/2003 – 9/24/2016	34,247	6,406	5,958	10,086	80
tuts4you	English	6/10/2006 – 10/31/2016	40,666	2,539	-	2,206	38
exelab	Russian	8/25/2008 – 10/27/2016	328,477	13,289	4,572	-	628
<b>Total:</b>	-	<b>02/07/2005- 10/31/2016</b>	<b>590,699</b>	<b>47,492</b>	<b>15,576</b>	<b>14,851</b>	<b>987</b>



# Cyber Threat Intelligence (CTI) Example – Bank Exploits (e.g., BlackPOS)



# Cyber Threat Intelligence (CTI) Example – Mobile Malware



# Labeling Hacker Exploits for Proactive Cyber Threat Intelligence: A Deep Transfer Learning Approach

Benjamin Ampel (MISQ, 2<sup>nd</sup> Round)

# Literature Review: Hacker Forum Exploit Analysis

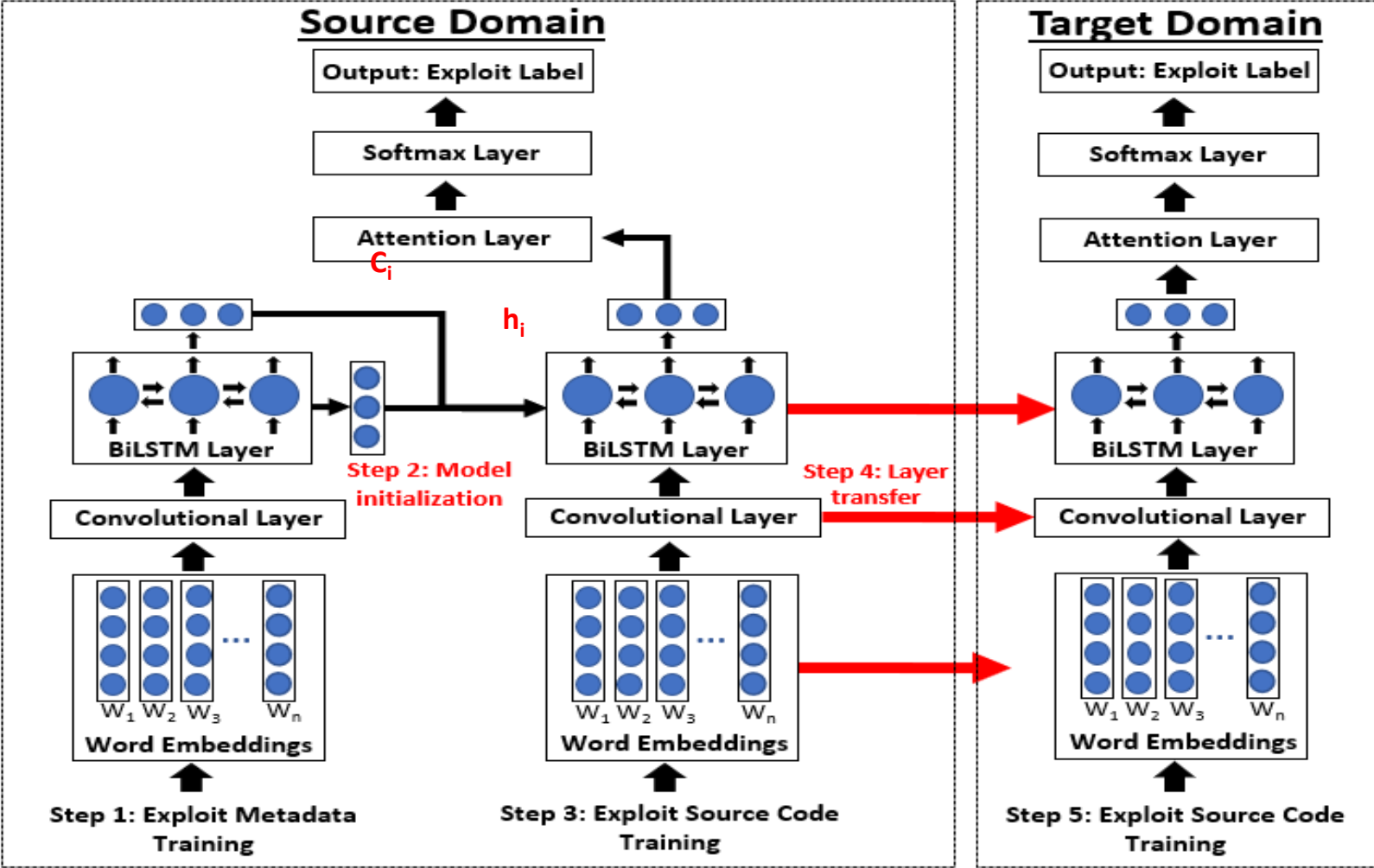
Year	Author	1. Data Source	2. Data Type Used	Analytics	Identified Exploits	3. Purpose
2019	Schafer et al.	General purpose forums	Forum titles, users, message, topic, keywords	SNA, LDA	Leaks, botnets, DDoS	Trend identification
2019	Benjamin et al.	General purpose forums	Post content, attachments, source code, keywords, reputation	OLS Regression	Rootkit, XSS, SQLi, DDoS, shellcode, drive-by	Darknet identification, collection, evaluation
2018	Williams et al.	General purpose forums	Sub-forum name, author, post content, attachment metadata	LSTM	Crypters, keyloggers, RATs, DDoS, SQLi	Exploit categorization
2018	Goyal et al.	Forums, Twitter, Blogs	Post content, Tweet content, blog content	LSTM, RNN	Trojan, Windows, Apple OSX, phishing	Cyber attack prediction
2018	Deliu et al.	Nullified.IO leak	Post content	SVM, CNN	Botnet, crypter, keylogger, malware, rootkit	Exploit categorization
2017	Samtani et al.	General purpose forums	Post content, assets, thread, author, source code	LDA, SVM	Crypters, keyloggers, RATs, botnets	Exploit categorization
2017	Grisham et al.	General purpose forums	Post content, date, author, role, attachments	RNN	Mobile malware	Malware identification/ Proactive CTI
2017	Deliu et al.	Nullified.IO leak	Post content	SVM, LDA	Backdoor, botnet, crypter, DDoS, exploit, malware, password, rootkit	Exploit categorization

- **Key Observations:**

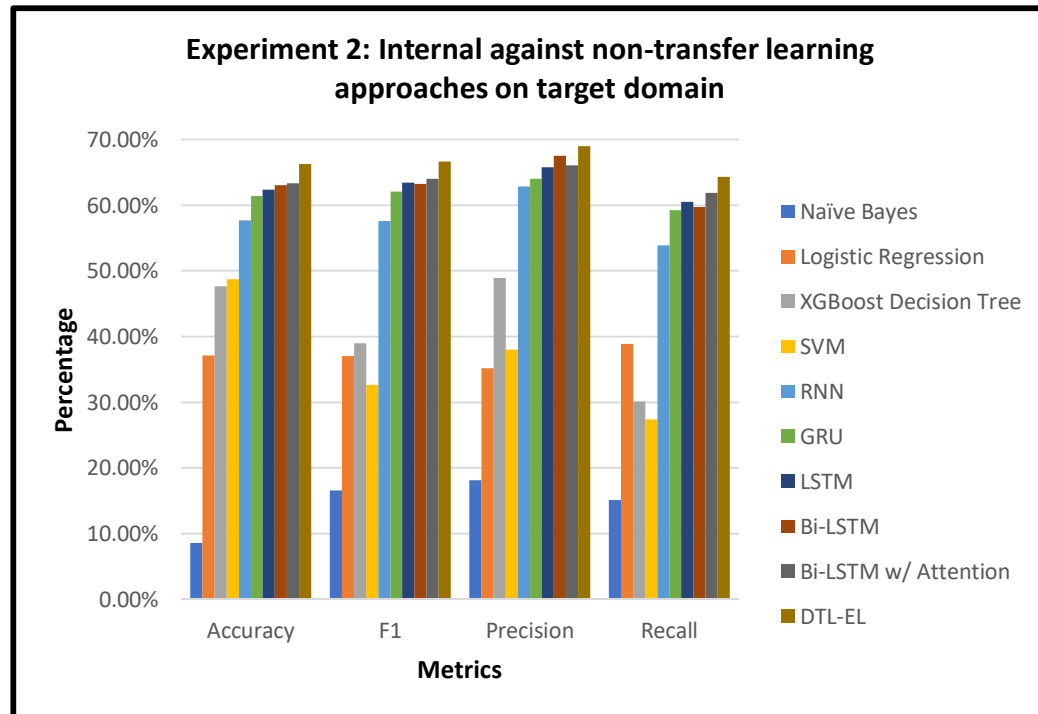
1. Studies focus on general forums, but not exploit DNMs or public repositories.
2. Although source code contains valuable information, many studies omit them from analysis.
3. The most common task is to categorize post content by exploit category.



# Research Design: DTL-EL



# Results and Discussion: DTL-EL Model



Experiment 2: Internal against non-transfer learning models		Results			
Model	Layer Weights	Accuracy	Precision	Recall	F1
Naïve Bayes	Random	8.59% ***	18.09% ***	15.08% ***	16.45% ***
Logistic Regression	Random	37.16% ***	35.13% ***	38.85% ***	36.9% ***
XGBoost Decision Tree	Random	47.65% ***	48.87% ***	30.06% ***	37.22% ***
SVM	Random	48.72% ***	37.98% ***	27.38% ***	31.82% ***
RNN	Random	57.64% ***	62.89% ***	53.93% ***	57.62% ***
GRU	Random	61.34% ***	64.06% ***	59.27% ***	62.09% ***
LSTM	Random	62.39% ***	65.77% ***	60.49% ***	63.42% ***
BiLSTM	Random	63.05% ***	67.56% ***	59.71% ***	63.21% ***
BiLSTM w/ Attention	Random	63.38% ***	66.04% ***	61.88% ***	64.02% ***
<b>DTL-EL (Our model)</b>	<b>Transferred</b>	<b>66.17%</b>	<b>68.25%</b>	<b>64.99%</b>	<b>66.61%</b>





# Case Study: System Integration

- Hacker exploit source code can be input for classification with attention weights.
- The system applies a DTL-EL label upon the collection of new hacker forum text, providing real-time information to researchers.
  - APIs allow for forums to be downloaded in their entirety with related programming languages and exploit labels for source code.

## Hacker Exploit Dashboard

### Label Your Exploit

Pick Your Model

DTL-EL ← **Select a model (DTL-EL or non-DTL) and input an exploit**

Input code snippet here:

```
SELECT UserId, Name, Password FROM Users WHERE UserId = 105 or 1=1;
```

Our model thinks this is a SQL Injection ← **Model output**

select **userid** name **password** users ← **Attention weights of the model output**

```
select userid name password users
```

Figure 16. Hacker Exploit Portal For Further Analysis

# Detecting Cyber Threats with AI Agents: Multilingual, Multimedia DNM Content

Reza Ebrahimi (JMIS, MIS, IEEE PAMI)

# Dark Net Marketplaces (DNMs)

- Other 654
- Security 388
- Software 1061

- Digital Goods 43611
- Drugs 46109
- Drugs Paraphernalia 278
- Services 3835
- Other 3370

## Exchange

BTC	1.0
mBTC	1000.0
USD	16088.0
EUR	13729.3
GBP	12040.6
CAD	20552.8
AUD	20758.8
SEK	134094.9
NOK	130044.7
DKK	102160.3
TRY	61054.5
CNH	105455.4
HKD	124855.0
RUB	931264.5
INR	1030953.6
JPY	1822398.0

## Onion mirrors

- t3e6ly3u0if4zcw2.onion verified
- jd6yhuwcivehvd4.onion
- t3e6ly3u0if4zcw2.onion
- 7ep7acrknzdcw3l.onion
- vilpaqbrmvizecjo.onion
- igyifrhvxq33sy5.onion
- 6qlocfg6zq2kyacl.onion
- x3x2dwb7jasax6tq.onion
- bkjcpa2klkmmowwq.onion
- xytjqcfendzeby22.onion
- nhib6cwhfsoyiuvg.onion
- k3pd243s57ftnpa.onion

19 20 ... 44 45 46 47 48 49 50 51 52 53 54 →

**Hacking For Newbies**

0.0000621  
HappyEyes (5200) (4.79★)  
WW → WW

ESCROW Order

**731986-Hacker's Desk Reference**

0.0001864  
color (8000) (4.76★)  
SB → WW

ESCROW Order

**Premium Carding Package**

0.0003107  
OnePiece (7400) (4.83★)  
PH → WW

ESCROW Order

**6 BITCOIN RANSOMWARE EASY MONEY SYSTEM**

0.0003107  
TheWealthMaker (1550) (4.78★)  
WW → WW

ESCROW Order

**Go to Windows updates anonymously**

0.0000621  
HappyEyes (5200) (4.79★)  
WW → WW

ESCROW Order

**HACK ANYONE USING THEIR IP ADDRESS**

0.0002486  
TopNotchMoneyMaker (4500) (4.74★)  
WW → WW

ESCROW Order

**PayPal - Scam Page (Phising site) [Looks Great]**

0.00031  
MicroDroper (2350) (4.91★)  
WW → WW

ESCROW Order

**PASSWORD MANAGER KIT**

0.0001224  
EiCartel (1800) (4.86★)  
WW → WW

ESCROW Order

PayPal - Scam Page (Phising site) [Looks Great]


**Vendor** MicroDroper (2350) (4.91★) (@ 67/1/3) Seller

**Price** 0.00031 (\$4.99) Price

**Ships to** Worldwide, Worldwide

**Ships from** PM

**Escrow** Yes



**Product Description**

You will get all files for build phishing PayPal site. Look perfect.

We are not include support to the product, so if you have 0% knowledge about site building and php - please do not make an order.

**Shipping options**

0.00 (\$0) You will get download link

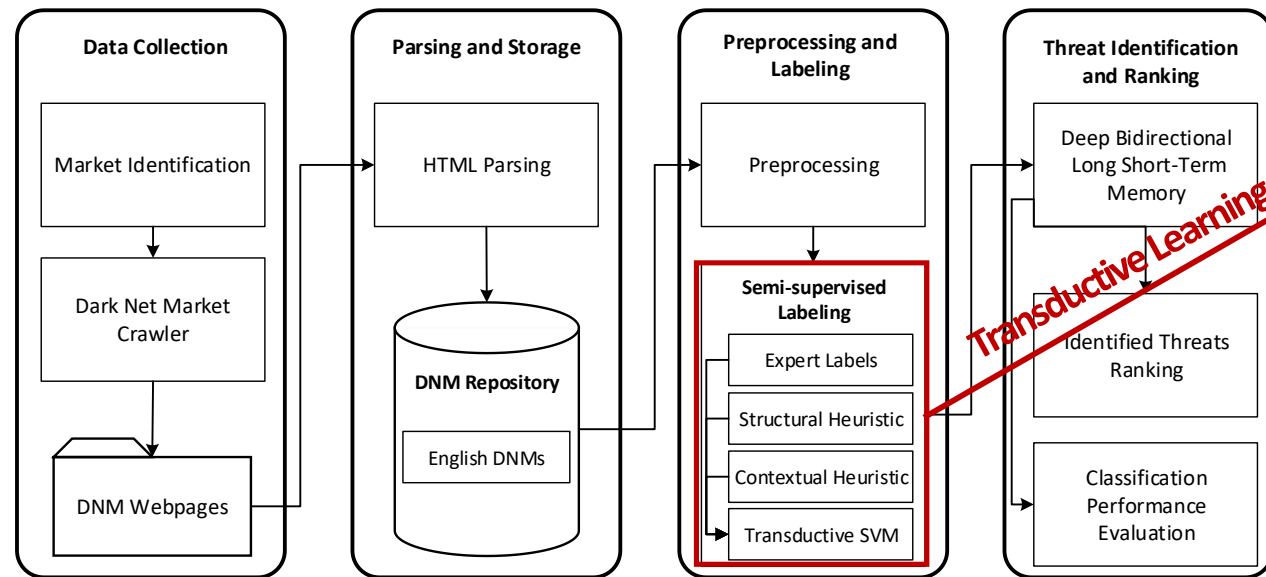
**Product Reviews** Quantity 1 Add to cart

**Product ratings**

53d ★★★★★ Enter your comments here

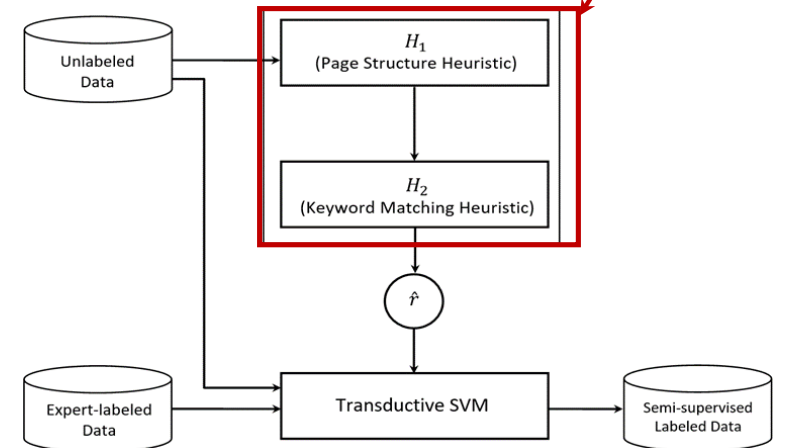
# Essay I: Learning From Unlabeled Cybersecurity Content (JMIS, March 2020)

- Learning from examples → supervised by human-labeled data → Expensive!
- Unlabeled data improves cyber threat detection with **transductive learning theory**



$$\min_w \left( \frac{\lambda}{2} \|w\|^2 + \frac{1}{2L} \sum_{i=1}^L l(y_i, w^T x_i) + \frac{\lambda'}{2U} \sum_{j=1}^U l(y'_j, w^T x'_j) \right)$$

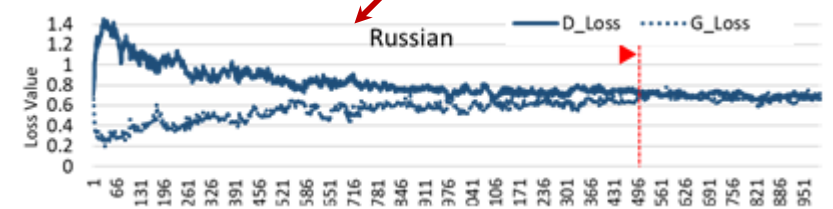
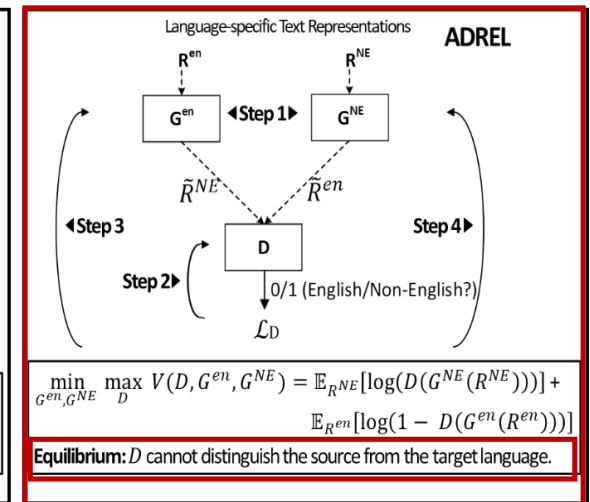
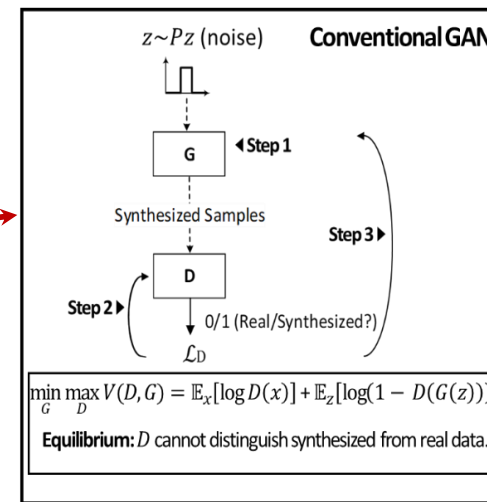
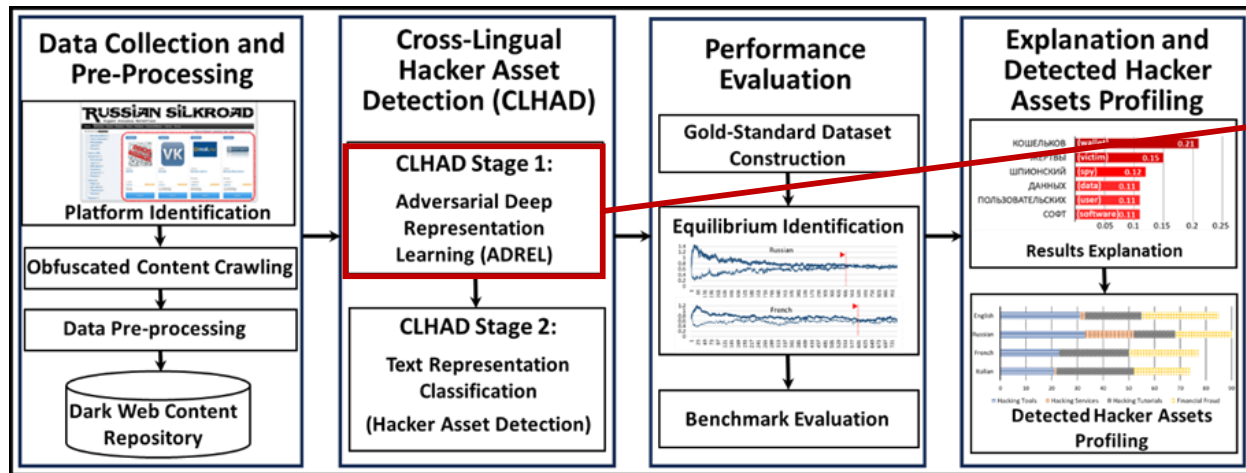
$$\text{subject to: } \frac{1}{U} \sum_{j=1}^U \max(0, \text{sign}(w^T x'_j)) = r$$



- Significantly decreased reliance on human supervision for cyber threat detection.

# Essay II: Learning from Heterogeneous Cybersecurity Content (MISQ, Forthcoming)

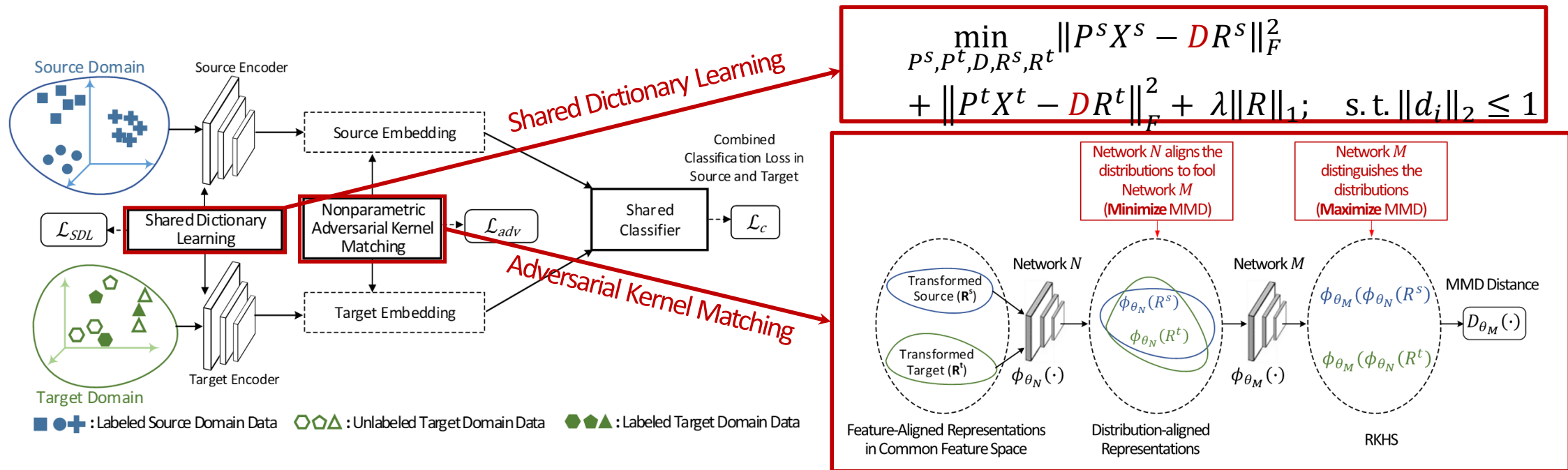
- Cyber threat detection in non-English content → lack of non-English training data
- Transfer cyber threat knowledge from high-resource English platforms to non-English ones with **transfer learning theory**



- Significantly decreased reliance on human supervision and outperformed machine translation.

# Essay III: Learning from Heterogeneous Cybersecurity Content (IEEE TPAMI, 2<sup>nd</sup> Round)

- Learning from two domains (multilingual text, source code, image representations)
- Align different data distributions & feature spaces with **domain adaptation theory**



- Enables heterogeneous data analytics (multilingual text, images) in any online market.

# ***Privacy and PII (Personally Identifiable Information) Analytics:***

identifying and alleviating privacy risks for vulnerable populations

**\* SaTC 2019-; SFS-2, 2019-**

## **Secure and Trustworthy Cyberspace (SaTC)**

### **PROGRAM SOLICITATION**

**NSF 21-500**

### **REPLACES DOCUMENT(S):**

**NSF 19-603**



**National Science Foundation**

Directorate for Computer and Information Science and Engineering  
Division of Computer and Network Systems  
Division of Computing and Communication Foundations  
Division of Information and Intelligent Systems  
Office of Advanced Cyberinfrastructure

## **CyberCorps(R) Scholarship for Service (SFS)**

**Defending America's Cyberspace**

### **PROGRAM SOLICITATION**

**NSF 21-580**

### **REPLACES DOCUMENT(S):**

**NSF 19-521**



**National Science Foundation**

Directorate for Education and Human Resources  
Division of Graduate Education

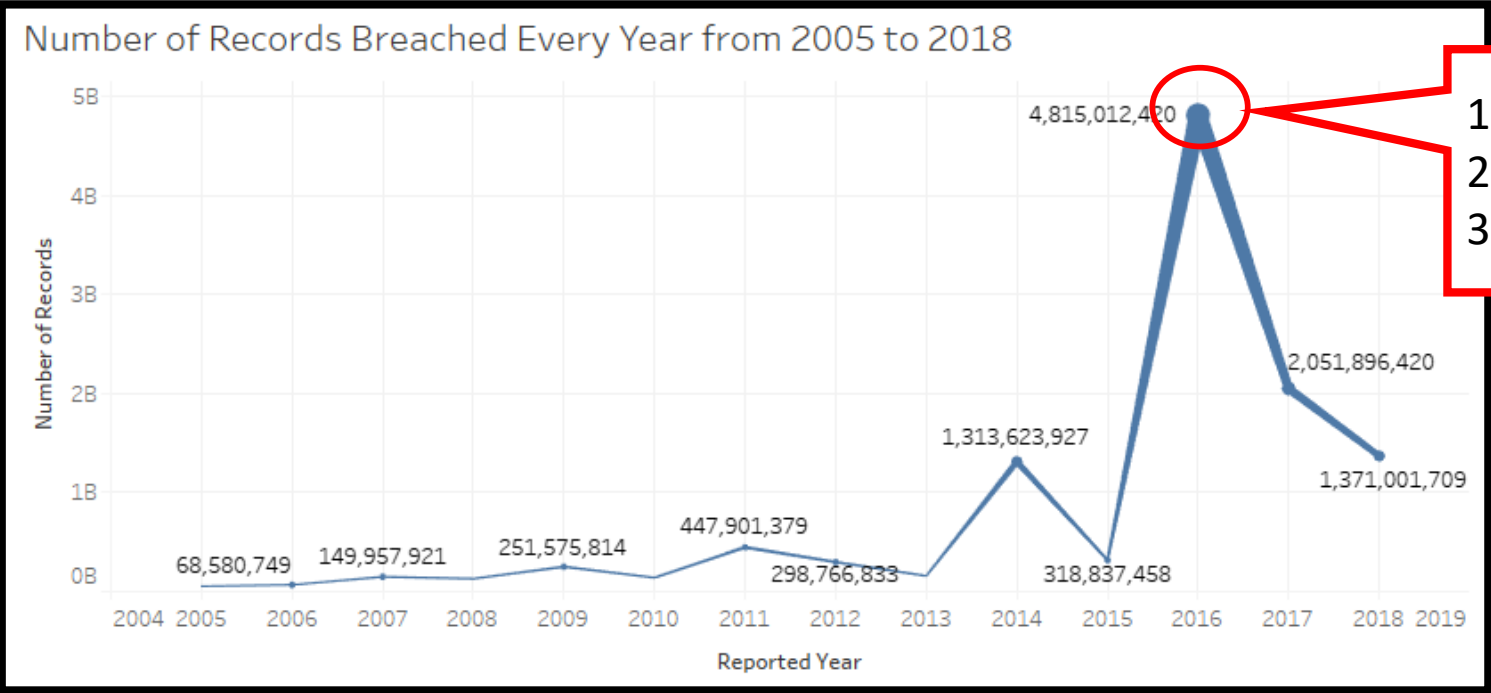
# Exploring Privacy Risk of Exposed Digital Personally Identifiable Information (PII): A Neighbor Attention-Based Approach

Fangyu Lin and Hsinchun Chen



# Data Breaches since 2005 (FTC, Clearinghouse, 2019)

- # of records breached: 11,582,808,013
- # of data breaches: 9,071

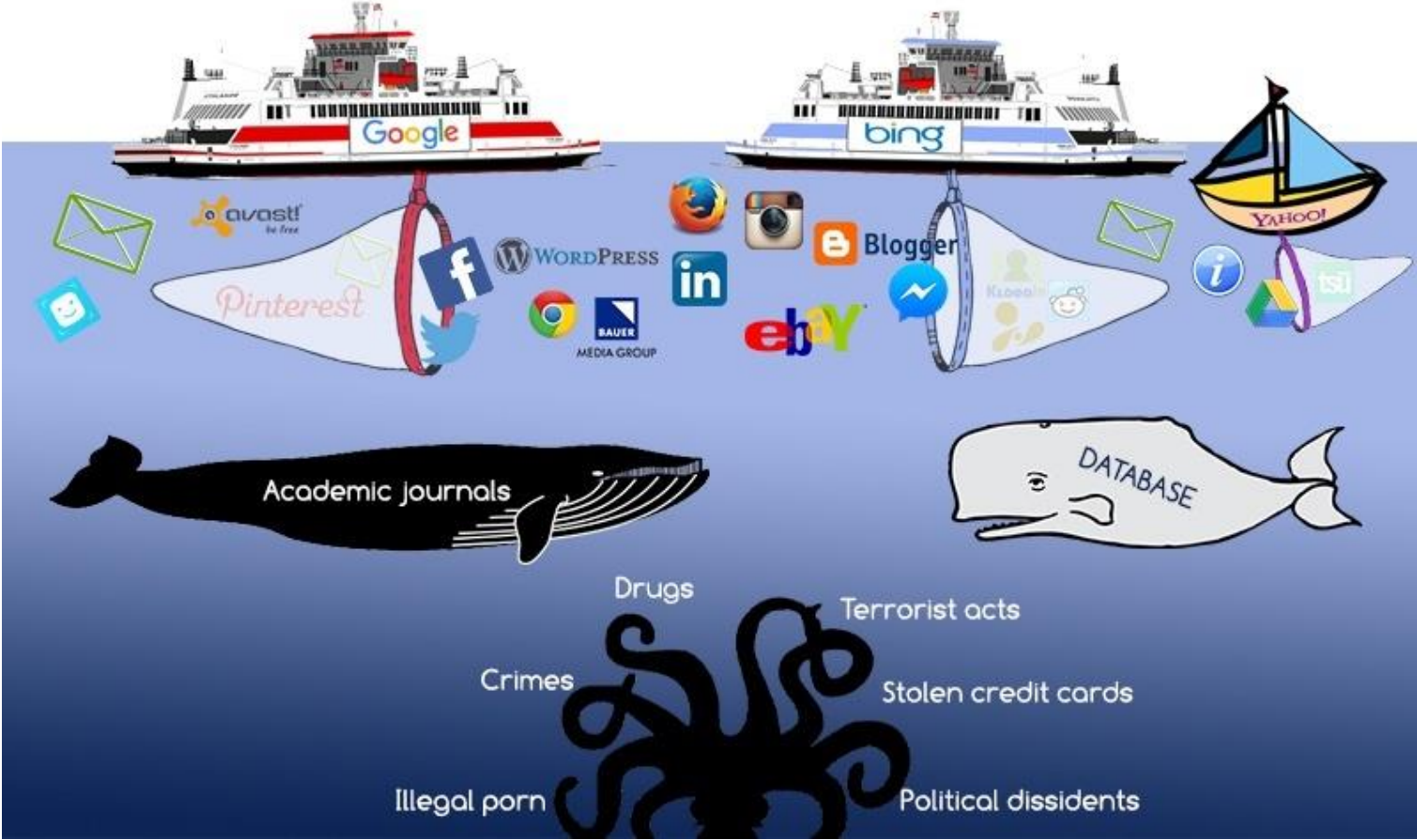


## 2016 Data Breach

- 1. Yahoo! : 3.5B user accounts
- 2. FriendFinder : 412M user accounts
- 3. MySpace : 360M passwords

# Revealing and Protecting PII: From Dark Web to Surface Web

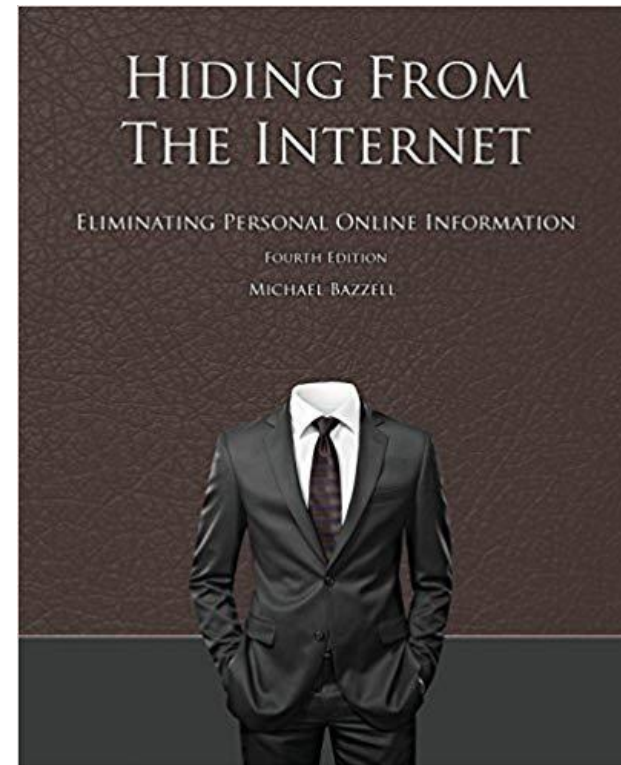
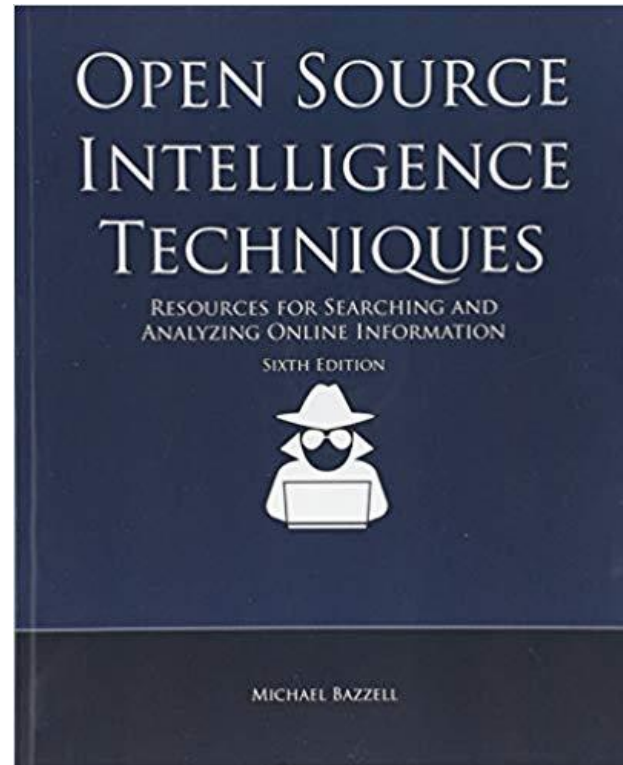
- Surface Web
- Deep Web
- Dark Web
- DarkNet
- Hacker Web



IRB, HIPAA, **GDPR**, **PII**

→ Cybersecurity to Privacy

→ **Michael Bazzell** + From Dark Web to Surface Web



# Dark Web Intelligence Sources (May, 2021)

Source	Description	Size*	Promising Attributes
<b>Stolen Account Collection</b>	Stolen social media and e-mail accounts	<b>25 billions</b>	<b>Username</b>
			Password
<b>Stolen Credit Card - Tormarket</b>	Stolen credit and debit card owner information * No card number	<b>832 thousands</b>	<b>Full name</b>
			Country
			State
			City
			<b>Zip</b>
<b>Stolen SSN - Buysn</b>	Personal information of SSN owners *No SSN	<b>5.75 millions</b>	<b>Full name</b>
			YOB
			City
			State
			<b>Zip</b>
			Country

“Passwords are like underwear...  
change often, don't share...”

## Stolen Accounts

Rank	E-mail Domains	Numbers	Percentage
1	yahoo.com	244,769,117	20.41%
2	hotmail.com	182,564,724	15.22%
3	gmail.com	103,435,791	8.62%
4	mail.ru	90,371,699	7.53%
5	aol.com	44,830,568	3.74%
6	yandex.ru	36,336,003	3.03%
7	rambler.ru	23,521,080	1.96%
8	hotmail.fr	16,571,495	1.38%
9	web.de	12,918,595	1.08%
10	live.com	11,661,375	0.97%
11	msn.com	11,248,354	0.94%
12	gmx.de	10,800,404	0.90%
13	163.com	10,492,032	0.87%
14	bk.ru	9,416,062	0.78%
15	yahoo.fr	8,886,223	0.74%
<b>Total</b>	-	<b>817,823,522</b>	<b>68.18%</b>

## Popular Passwords

Rank	Passwords	Numbers
1	123456	3,370,644
2	123456789	1,187,812
3	Homelesspa*	546,648
4	password	522,529
5	abc123	516,091
6	password1	435,753
7	12345	382,970
8	qwerty	376,099
9	12345678	357,654
10	1234567	287,453
11	1234567890	252,929
12	111111	236,852
13	iloveyou	211,593
14	123456a	205,807
15	123123	191,450
<b>Total</b>	-	<b>9,082,284</b>

# AZSecure Privacy Portal Design

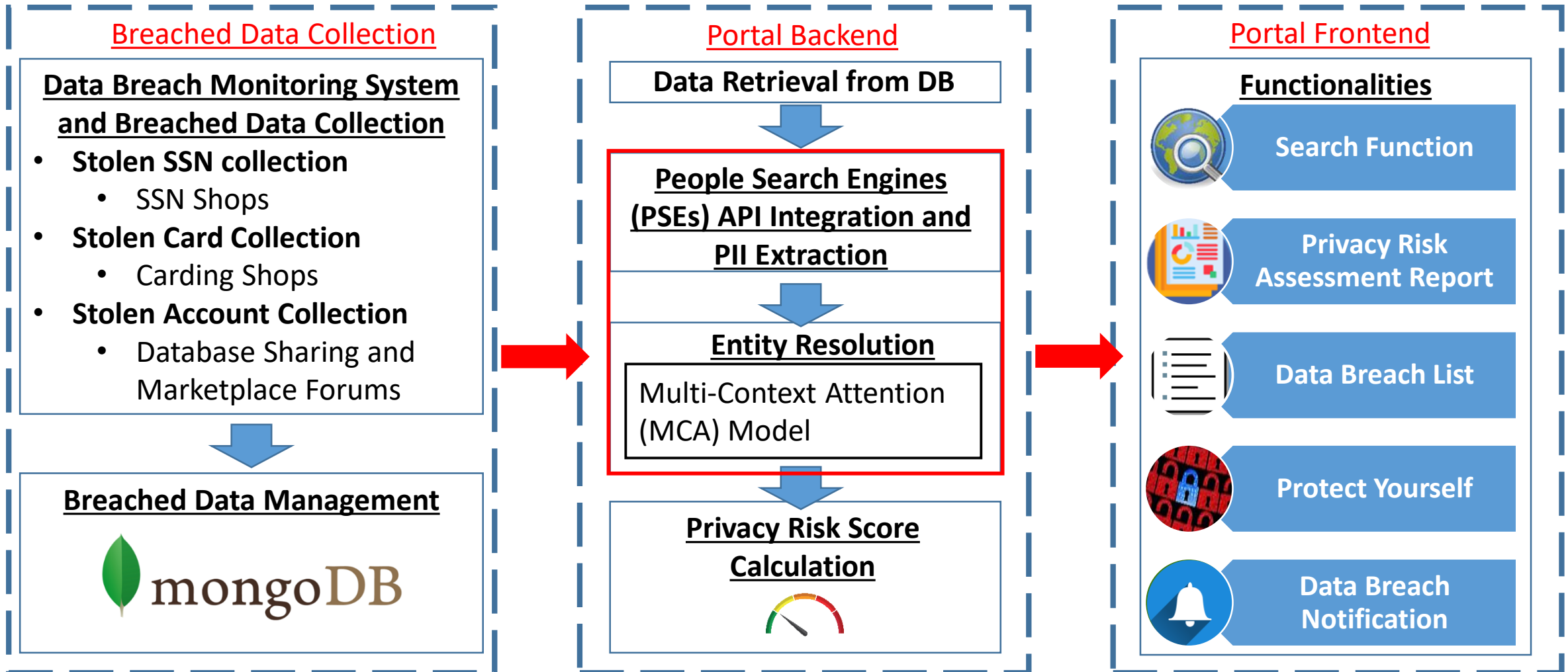


Figure 1. AZSecure Privacy Portal Project Overview

# Search in AZSecure Privacy Portal

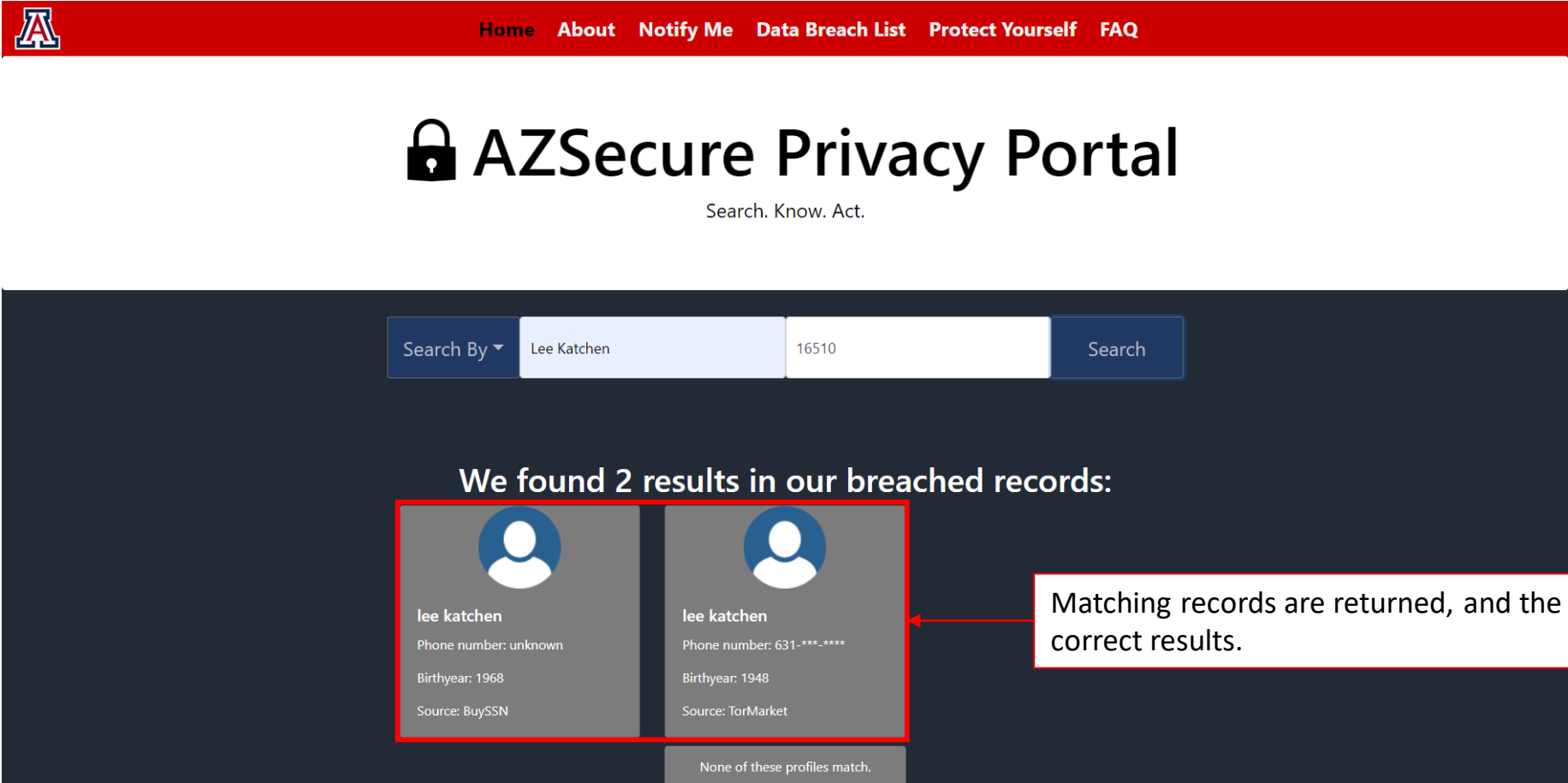


Figure 5. A mock-up response when records are found

# Return Exposed PII

Your information has been **compromised** in 1 breach:  
**TorMarket**

What was compromised? phone number, address, birthday, home town, current town, job details, relationship status, political views, religious views

Your privacy is at **high risk** compared to others in your age group.  
Your privacy risk score is **8.8**  
Find out what this score means [here](#).

**How do I protect myself? Find out [here](#).**

**Result from our database:**

Platform	Leaked Attributes
TorMarket	name: lee katchen   zip: 16***   address: 1234 Ma*****   jobDetails: carpenter   relationshipStatus: married   politicalViews: moderate   religiousViews: buddhist   birthyear: 1948   phoneNumber: 631-***-****   city: er*****

**Results from surface web search engines:**

Platform	Leaked Attributes	MCA Matching Results	TF/IDF Matching Results
Zabasearch	name: Lee R Katchen   state: Pennsylvania   address: 5551 Fr*****   birthyear: 1948   phoneNumber: 814-***-****   city: Er*****	✓	⊗
Anywho	name: Lee R Katchen   state: PA   address: 1024 Ap*****   city: Er*****   phoneNumber: 814-***-****	✓	⊗
Anywho	name: Lee R Katchen   state: PA   address: 3747 Bi*****   Age: 52   city: Er*****   phoneNumber: 814-***-****	⊗	⊗
Anywho	name: Lee R Katchen   state: PA   address: 3747 Bi*****   city: Er*****   Age: 52   phoneNumber: 814-***-****	⊗	⊗

Figure 9. Mock-ups of a comprehensive exposed PII profile



# *Adversarial Malware Generation and Evasion:* adversarial AI in cybersecurity

**\* SaTC 2019-; SFS-2, 2019-**

## **Secure and Trustworthy Cyberspace (SaTC)**

### **PROGRAM SOLICITATION**

**NSF 21-500**

### **REPLACES DOCUMENT(S):**

**NSF 19-603**



**National Science Foundation**

Directorate for Computer and Information Science and Engineering  
Division of Computer and Network Systems  
Division of Computing and Communication Foundations  
Division of Information and Intelligent Systems  
Office of Advanced Cyberinfrastructure

## **CyberCorps(R) Scholarship for Service (SFS)**

**Defending America's Cyberspace**

### **PROGRAM SOLICITATION**

**NSF 21-580**

### **REPLACES DOCUMENT(S):**

**NSF 19-521**



**National Science Foundation**

Directorate for Education and Human Resources  
Division of Graduate Education

# Defending Cybersecurity AI Agents

Reza Ebrahimi (JMIS, MISQ)

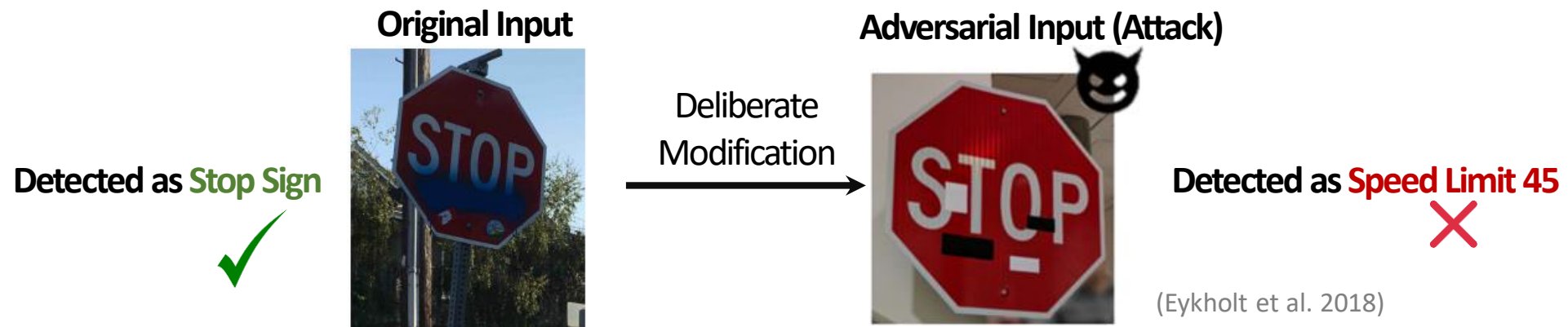
- **Essay 1:** Learning to Protect Malware Detectors
- **Essay 2:** Learning to Protect any Defense AI agent

# Defending Cybersecurity AI Agents

Symantec unveils artificial intelligence powered endpoint security 

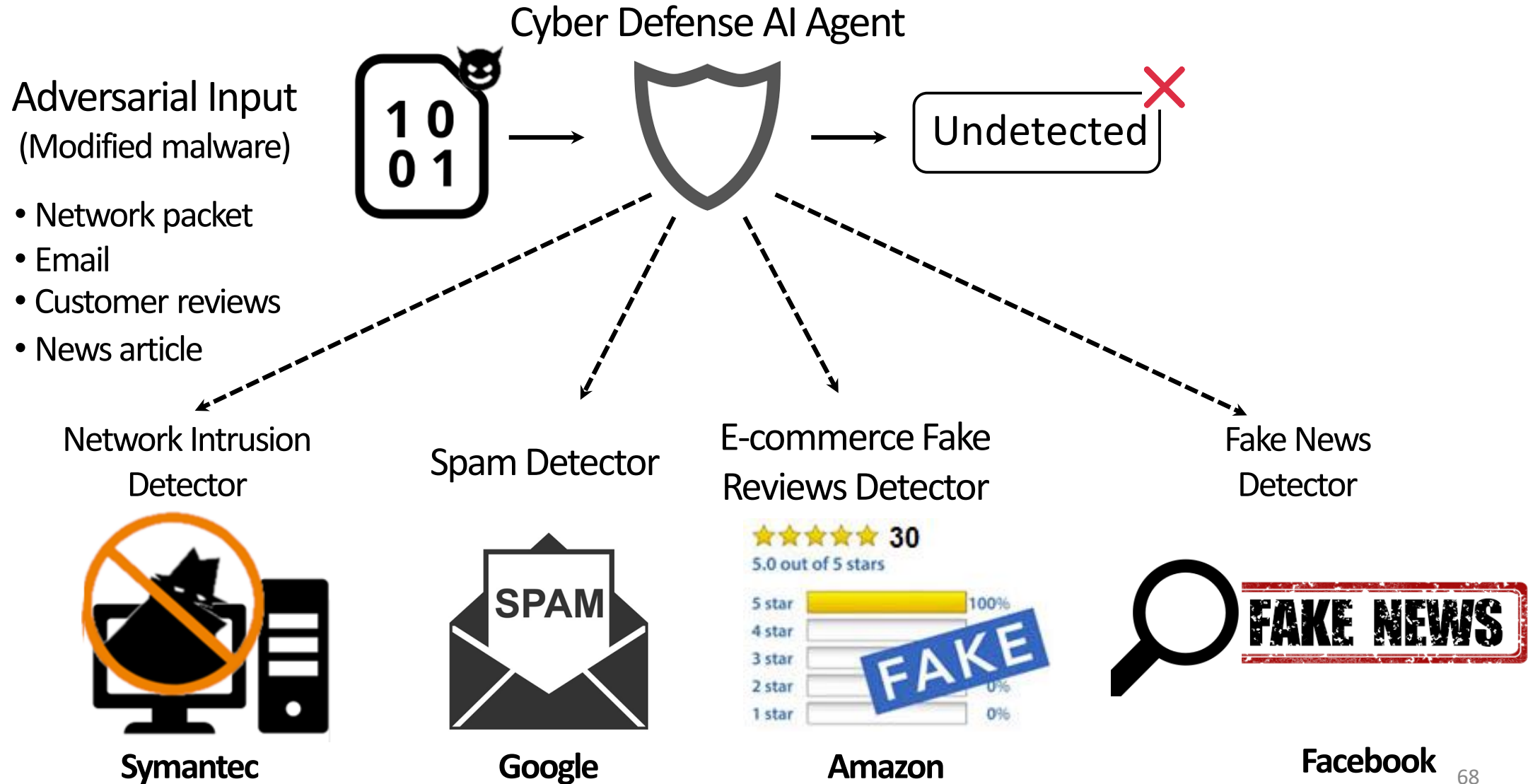
(expresscomputer.in)

- Cybersecurity firms are adopting AI agents for autonomous cyber defense (Rai et al. 2019).
  - Automate threat detection and remediation at a large scale (Tolido et al. 2019).
- However, AI agents have shown to be vulnerable to adversarial attacks.
- Inputs meticulously modified to mislead them (Yuan et al. 2019). → Known as adversarial attacks (Apruzzese et al. 2019).



- **How can we protect cyber defense AI agents?**

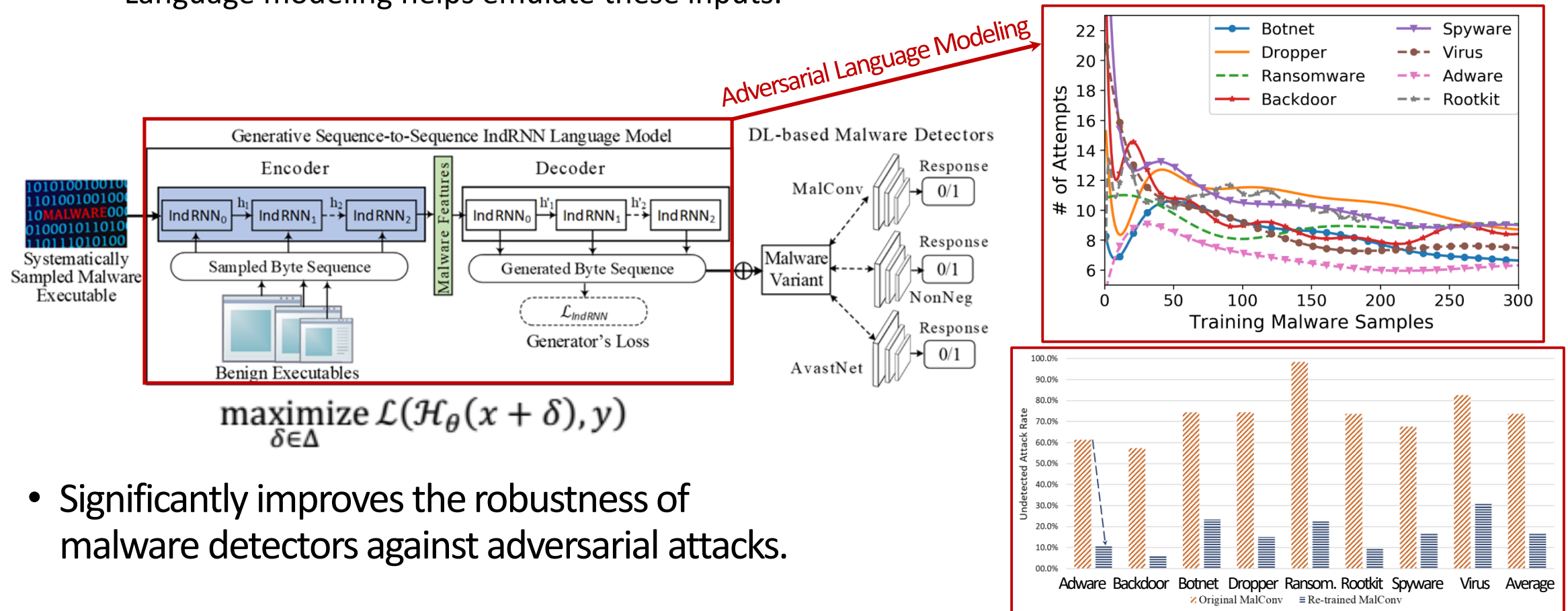
# Defending Cybersecurity AI Agents



# Essay I: Learning to Protect Malware Detectors

(JMIS, In sub.)

- Malware attack is #1 cause of damage to IT infrastructure (Bissell et al. 2019).
- Malware detector is the first line of defense. → Can be misled by adversarial inputs.
  - Language modeling helps emulate these inputs.

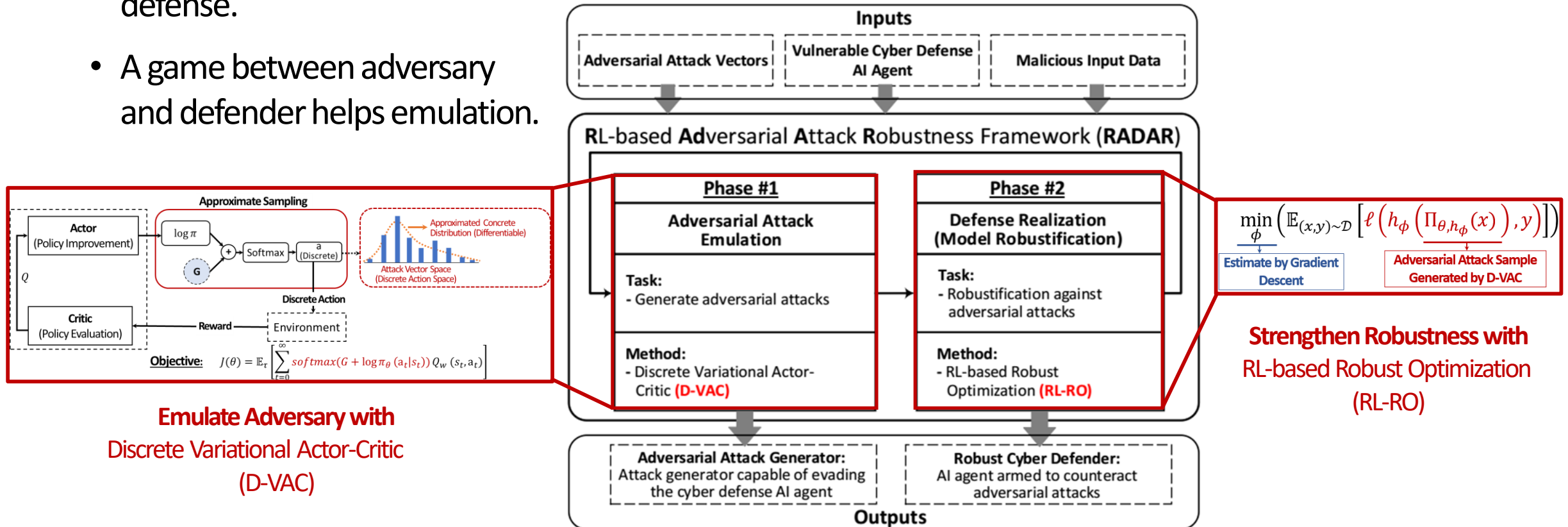


- Significantly improves the robustness of malware detectors against adversarial attacks.

# Essay II: Learning to Protect any Defense AI Agent

## (MISQ, 1<sup>st</sup> Round)

- Modern AI agents can be misled by adversarial attacks. → Emulating these attacks is vital for defense.
- A game between adversary and defender helps emulation.



- Strengthened the robustness of AI agents against adversarial attacks.

# *Smart Vulnerability Assessment:* scientific workflows and OSS vulnerability analytics and mitigation

**\* CICI 2019-; SFS-2, 2019-**

**CyberCorps(R) Scholarship for Service (SFS)**  
**Defending America's Cyberspace**

---

**PROGRAM SOLICITATION**  
NSF 21-580

---

**REPLACES DOCUMENT(S):**  
NSF 19-521

---



National Science Foundation  
Directorate for Education and Human Resources  
Division of Graduate Education

**Cybersecurity Innovation for Cyberinfrastructure (CICI)**

---

**PROGRAM SOLICITATION**  
NSF 21-512

---

**REPLACES DOCUMENT(S):**  
NSF 19-514

---



National Science Foundation  
Directorate for Computer and Information Science and Engineering  
Office of Advanced Cyberinfrastructure

Linking Hacker Community Exploits to Known  
Vulnerabilities for Proactive Cyber Threat  
Intelligence:  
An Attention-based Deep Structured Semantic  
Model Approach

Sagar Samtani (MISQ, forthcoming)





**Protecting Scientific Instruments  
and Cyberinfrastructure:**  
From iPlant/CyVerse (life sciences)  
to BioSphere 2/LEO (earth sciences)...  
**a new UA/USF/AZSecure NSF CICI project,  
2019-2022**



# Hacker Forum Exploits

The image shows a screenshot of a hacker forum listing exploits. The table has three columns: DATE, DESCRIPTION, and TYPE. Annotations highlight key characteristics of the posts:

- Exploit Titles:** A white box highlights the title of the first exploit: "WordPress Database Backup Remote Command Execution Exploit".
- Post Dates:** A white arrow points to the date "28-07-2019" in the first row.
- Exploit Category:** A white box highlights the category "[ local exploits ]" in the second row.

DATE	DESCRIPTION	TYPE
28-07-2019	WordPress Database Backup Remote Command Execution Exploit	php
27-07-2019	WordPress Database Backup Remote Command Execution Exploit	java
24-07-2019	Trend Micro Deep Discovery Inspector IDS - Security Update Exploit	multiple
17-07-2019	MAPLE Computer WBT SNMP Administrator 2.0.195.15 - Remote Buffer Overflow Exploit	windows
16-07-2019	PCMan FTP Server 2 ALLO Buffer Overflow Exploit	windows
16-07-2019	PHP Laravel Framework Token Unserialize Remote Command Execution Exploit	linux
12-07-2019	Xymon 4.3.25 - useradm Command Execution Exploit	multiple
10-07-2019	Apache mod_ssl < 2.8.7 OpenSSL - OpenFuckV2.c Remote Buffer Overflow (2) Exploit	unix
28-07-2019	[ local exploits ]	
28-07-2019	Microsoft Windows 7 build 7601 (x86) - Local Privilege Escalation Exploit	windows
28-07-2019	Deepin Linux 15 - lastore-daemon Local Privilege Escalation Exploit	multiple
27-07-2019	VMware Workstation / Player < 12.5.5 - Local Privilege Escalation Exploit	multiple
26-07-2019	Linux Kernel 4.4.0-21 < 4.4.0-51 (Ubuntu 14.04/16.04 x86-64) AF_PACKET	linux

- **Key Characteristics:**

1. Descriptive tool names (target, operations, etc.)
2. Clear categories of exploits (e.g., target system)
3. Post date of when exploit was posted

# Vulnerability Assessment



Cisco IOS IPS Denial of Service Vulnerability - Cisco Systems **a**

**Synopsis** **b**  
The remote device is missing a vendor-supplied security patch.

**Description** **c**  
The Cisco IOS Intrusion Prevention System (IPS) feature contains a vulnerability in the processing of certain IPS signatures that use the SERVICE.DNS engine. This vulnerability may cause a router to crash or hang, resulting in a denial of service condition.

**Risk Information**  
Risk Factor: High  
CVSS Base Score: 7.8 **d**  
CVSS Vector: CVSS2#AV:N/AC:L/Au:N/C:N/I:N/A:C

**Reference Information:**  
CVE: CVE-2008-2739  
OSVDB: 48711  
BID: 31364 **e**

Bugtraq ID: 31364  
Class: Failure to Handle Exceptional Conditions  
CVE: CVE-2008-2739  
Remote: Yes  
Local: No  
Published: Sep 24 2008 12:00AM  
Updated: Sep 24 2008 08:19PM  
Credit: the discoverer of this issue is not known; this issue was disclosed by Cisco.

**Vulnerable:** **e**  
Cisco IOS 12.4YA  
Cisco IOS 12.4XZ  
Cisco IOS 12.4XY  
Cisco IOS 12.4XW  
Cisco IOS 12.4XV  
Cisco IOS 12.4XT

Category	Metadata	Description	Data Type
Description	Name	Short, descriptive name of vulnerability	Short text
	1. Family Name	Family vulnerability belongs to (e.g., Windows, etc.)	Categorical
	Description	Lengthy text description about vulnerability	Long text
	Synopsis	Short description of vulnerability	Short text
	Solution	Description or solution links	Short text
Risk	2. Vulnerable Systems	List of systems susceptible to vulnerability	Short text (list)
	3. CVSS	Value between 0.0-10.0 indicating vulnerability severity	Continuous
	Risk Factor	Categorical rating of risk (High, Low)	Categorical
	CVE	Vulnerability reference number	Categorical
	Publication Date	Date vulnerability was publicly published	Date

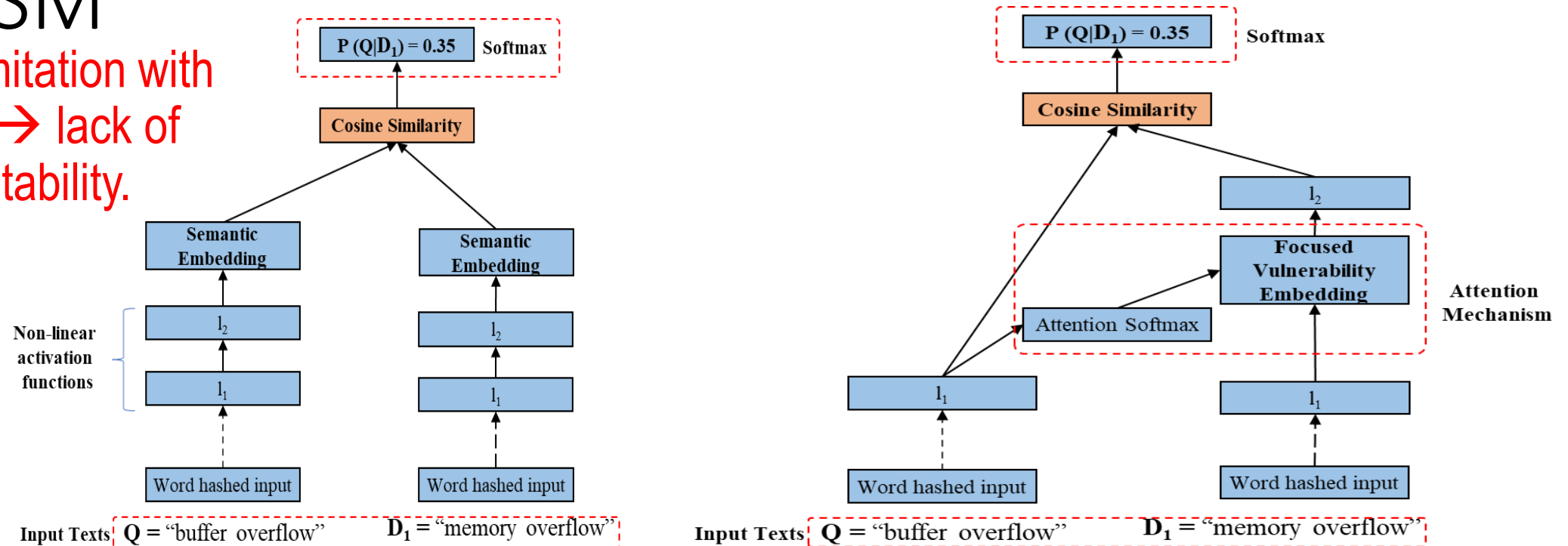
## Key Attributes Returned by Modern Vulnerability Scanners

- **Key Characteristics:**

1. Short, descriptive title of vulnerability
2. List of systems susceptible to vulnerability
3. Common Vulnerability Severity Score (0.0 – 10.0)

# Proposed Exploit Vulnerability Attention-DSSSM

- **Key Limitation with DSSM** → lack of interpretability.



- **Contribution:** EVA-DSSM integrates an attention mechanism into the DSSM. Identifies and outputs key exploit features essential for creating links

# Experiment Results: EVA-DSSM vs Deep Learning Matching Algorithms

Algorithm	Remote Exploits					Local Exploits				
	NDCG@1	NDCG@3	NDCG@5	MRR	MAP	NDCG@1	NDCG@3	NDCG@5	MRR	MAP
ANMM	0.4214***	0.5453***	0.5670***	0.6009***	0.5434***	0.3525***	0.4421***	0.5099***	0.5229***	0.4897***
ARC-I	0.2589***	0.3683***	0.4409***	0.4384***	0.4038***	0.3275***	0.4152***	0.4923***	0.4754***	0.4914***
ARC-II	0.3964***	0.5450***	0.5855***	0.5999***	0.5616***	0.4025***	0.5010***	0.5681***	0.5646***	0.5692***
KNRM	0.4571***	0.5521***	0.6152***	0.6433***	0.5549***	0.4000***	0.4603***	0.5389***	0.5478***	0.5155***
Conv-KNRM	0.5411	0.6330*	0.6745*	<b>0.7053</b>	0.6553**	0.4850***	0.5837***	0.6311***	0.6388***	0.6188***
DRMM	0.5339	0.6420	0.6830	0.6943	0.6760	0.1700***	0.2511***	0.4242***	0.3807***	0.3606***
DUET	0.5232	0.6104*	0.6601*	0.6671	0.6061***	0.3725***	0.4356***	0.5231***	0.5146***	0.5268***
MatchLSTM	0.1536***	0.3220***	0.4164***	0.3881***	0.4026***	0.2300***	0.3459***	0.4389***	0.4053***	0.4485***
MV-LSTM	0.5393	0.6250**	0.6549**	0.6831*	0.6420**	0.5325***	0.5943***	0.6483***	0.6541***	0.6365***
DSSM	0.3339***	0.5019***	0.5579***	0.5391***	0.5722***	0.5175***	0.6455***	0.6723***	0.6696***	0.6984***
Left EVA-DSSM	0.1607***	0.2934***	0.4118***	0.3813***	0.3982***	0.4155***	0.4333***	0.2500***	0.3170***	0.4306***
<b>EVA-DSSM</b>	<b>0.5469</b>	<b>0.6499</b>	<b>0.6857</b>	<b>0.7023</b>	<b>0.6834</b>	<b>0.6775</b>	<b>0.7779</b>	<b>0.7853</b>	<b>0.7865</b>	<b>0.8092</b>
Algorithm	Web Applications					DoS Exploits				
	NDCG@1	NDCG@3	NDCG@5	MRR	MAP	NDCG@1	NDCG@3	NDCG@5	MRR	MAP
ANMM	0.3125***	0.4527***	0.5114***	0.5075***	0.4704***	0.1790***	0.2691***	0.3640***	0.3969***	0.3532***
ARC-I	0.0906***	0.3378***	0.4275***	0.3637***	0.4042***	0.1176***	0.2111***	0.2717***	0.2828***	0.3233***
ARC-II	0.3250***	0.4894***	0.5410***	0.5275***	0.5405***	0.2053***	0.2881***	0.3395***	0.3697***	0.3864***
KNRM	0.5312	0.6248**	0.6728**	0.6772*	0.6786*	0.2684**	0.3166***	0.3461***	0.3817***	0.4002***
Conv-KNRM	0.5531	0.6716*	0.6973*	0.7122	0.6864*	0.2825*	0.3291***	0.3913***	0.4293**	0.4468***
DRMM	0.3619**	0.4874***	0.5497***	0.5156***	0.5373***	0.2333**	0.2954***	0.3493***	0.4052**	0.3851***
DUET	0.0907***	0.3489***	0.4257***	0.3704***	0.3959***	0.1561***	0.2388***	0.2917***	0.3179***	0.3368***
MatchLSTM	0.1063***	0.2906***	0.4187***	0.3606***	0.3839***	0.2986	0.3452*	0.4102*	0.4652	0.4472**
MV-LSTM	0.4531*	0.6416*	0.6648**	0.6481**	0.6473**	0.2614**	0.3397***	0.4095**	0.4524*	0.4371***
DSSM	0.5968	0.7325	0.7796	0.7468	<b>0.7947</b>	0.2632**	0.3625**	0.4079**	0.5011**	0.4367**
Left EVA-DSSM	0.0719***	0.3098***	0.3926***	0.3373***	0.3769***	0.1175***	0.1559***	0.2457***	0.2432***	0.3117***
<b>EVA-DSSM</b>	<b>0.6281</b>	<b>0.7602</b>	<b>0.7885</b>	<b>0.7684</b>	<b>0.7863</b>	<b>0.3579</b>	<b>0.4550</b>	<b>0.4954</b>	<b>0.5133</b>	<b>0.6009</b>

- EVA-DSSM outperforms all deep learning benchmarks
- Conv. or LSTM operations achieved lower performances
- Indicates that integrating an attention mechanism into the DSSM architecture does not deteriorate performance

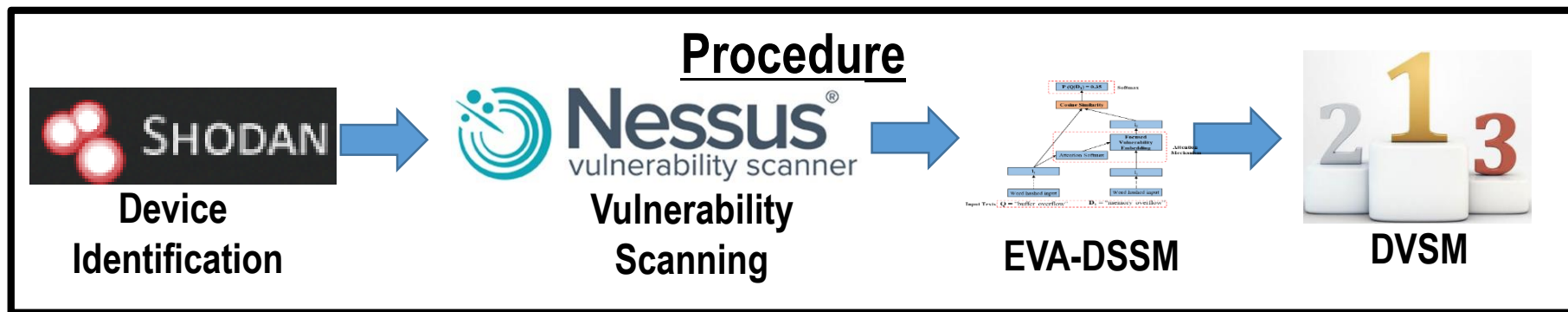
# Case Studies: SCADA and Hospitals



- 20,461 SCADA Devices from major vendors (e.g., Rockwell)
- **Motivation:** SCADA → control critical infrastructure



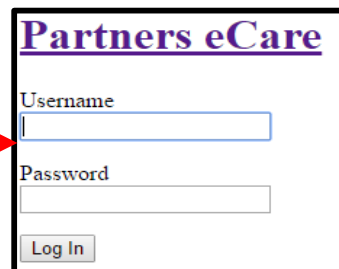
- 1,879 devices from top 8 US hospitals
- **Motivation:** Hospitals → popular target for hackers



# Hospital Case Study

Hospital Device Information		Device Severity Score Information for Selected Devices			
Hospital Name	# of Vulnerable Devices/# of devices	Device Type	# of Vulnerabilities	Vulnerabilities	DVSM
12x.x.x.x	133/808	FTP/SSH Server	3	FTP issues	4.591
19x.x.x.x	27/301	SSH Server	3	SSH issues	4.376
17x.x.x.x	31/274	eCare web portal	47	XSS, OpenSSL, buffer overflow, DoS	61.761
16x.x.x.x	59/160	Medical computing portal	5	PHP and SSH issues	4.863
14x.x.x.x	64/130	Web Server	3	SQL Injections	7.528
		Apple TV	2	Buffer overflow	5.381
14x.x.x.x	14/107	SSH/Web server	4	PHP and SSH issues	3.871
6x.x.x.x	9/52	Informational diabetes portal	3	SVN and Unix vulnerabilities	7.159
16x.x.x.x	7/47	Web Server	6	XSS, HTMLi	9.367
<b>Total:</b>	<b>344/1,879 (18.31%)</b>	-	-	-	-

- Portals are a common avenue for hackers to access sensitive records (Ayala 2016).
- Analysis shows an eCare portal with a large attack surface: 47 vulnerabilities for a DVSM of 61.761.
- Network admins can prioritize this device when analyzing their weaknesses.



Vulnerability Name (CVSS Score)	Exploit Name (Post Date)	Severity Score
"OpenSSL Unsupported" (10.0)	"OpenSSL TLS Heartbeat Extension – Memory Disclosure" (4/8/2014)	3.366
"Multiple XSS Vulnerabilities" (4.3)	"Portal XSS Vulnerability" (5/28/2010)	1.261
...	...	...
-	-	<b>Total: 61.761</b>

# Some Advice for Junior Faculty and Ph.D. Students: Journals and Grants

MIS  
Quarterly







# Major Journals: i-School, c-School, b-School

- i-School (\$80K) & health informatics Journals: JASIST, ACM TOIS; JAMIA, JBI  
→ “informatics” (text) focused, system driven; helpful for NSF & NIH/NLM funding
- c-School (\$100K) Journals: ACM TOIS, IEEE TKDE, CACM, IEEE IS, IEEE Computer, IEEE SMC → algorithm/computing focused, data driven; helped significantly with NSF funding (same for major CS conferences)
- b-School (\$180K) Journals: MISQ, ISR, JMIS, MS, ACM TMIS, DSS → “design science” focused, managerial framework/principle/knowledge base; helped get jobs in major b-schools (little federal funding)

# Major Journals: Chen, i-, c-, b-school, CISE


- Work hard; be persistent; colleagues & students help a lot; a little bit of luck helps



**Hsinchun Chen** 

[University of Arizona](#)  
Verified email at email.arizona.edu - [Homepage](#)

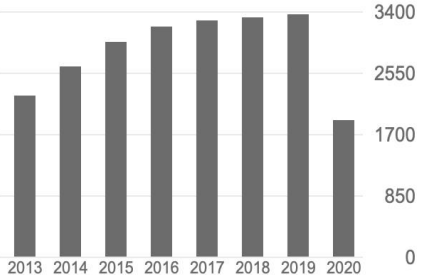
[business analytics](#) [data mining](#) [security informatics](#) [health informatics](#)

 FOLLOW


	CITED BY	YEAR
<input type="checkbox"/> <a href="#">Business intelligence and analytics: From big data to big impact</a> H Chen, RHL Chiang, VC Storey MIS quarterly, 1165-1188	5250	2012
<input type="checkbox"/> <a href="#">Credit rating analysis with support vector machines and neural networks: a market comparative study</a> Z Huang, H Chen, CJ Hsu, WH Chen, S Wu Decision support systems 37 (4), 543-558	1099	2004
<input type="checkbox"/> <a href="#">Sentiment analysis in multiple languages: Feature selection for opinion classification in web forums</a> A Abbasi, H Chen, A Salem ACM Transactions on Information Systems (TOIS) 26 (3), 1-34	1044	2008
<input type="checkbox"/> <a href="#">Applying associative retrieval techniques to alleviate the sparsity problem in collaborative filtering</a> Z Huang, H Chen, D Zeng ACM Transactions on Information Systems (TOIS) 22 (1), 116-142	858	2004


**Cited by** [VIEW ALL](#)

	All	Since 2015
Citations	39953	18060
h-index	101	56
i10-index	334	237



**Co-authors** [EDIT](#)

 **Daniel Dajun Zeng**  
Professor of MIS. Affiliations: Uni...

 **Wingyan Chung**  
Western Carolina University

>

>

## refine by venue

- Decis. Support Syst. (49)
- J. Assoc. Inf. Sci. Technol. (32)
- IEEE Intell. Syst. (20)
- J. Am. Soc. Inf. Sci. (14)
- Computer (12)
- J. Manag. Inf. Syst. (11)**
- ACM Trans. Inf. Syst. (10)
- ACM Trans. Manag. Inf. Syst. (8)
- Commun. ACM (7)
- IEEE Trans. Knowl. Data Eng. (7)
- J. Biomed. Informatics (6)
- Int. J. Hum. Comput. Stud. (6)
- IEEE Trans. Inf. Technol. Biomed. (6)
- MIS Q. (6)**
- ARIST (5)
- Inf. Syst. Frontiers (5)
- J. Inf. Sci. (3)
- Inf. Technol. Manag. (3)
- Inf. Process. Manag. (3)
- IEEE Trans. Syst. Man Cybern. Part A (3)
- IEEE Expert (2)
- 32 more options**

## refine by coauthor

- Daniel Dajun Zeng (24)**
- Michael Chau (22)**
- Jay F. Nunamaker Jr. (17)**
- Ahmed Abbasi (17)**
- Gavin Yulei Zhang (17)
- Wingyan Chung (14)
- Yan Dang 0001 (14)
- Bruce R. Schatz (13)
- Zan Huang (12)
- Robert P. Schumaker (11)
- 246 more options**

## MISQ SEs

<u>Gediminas Adomavicius</u>	University of Minnesota
<u>Corey Angst</u>	University of Notre Dame
<u>Indranil Bardhan*</u>	University of Texas at Austin
<u>Wai Fong Boh</u>	Nanyang Technological University
<u>Andrew Burton-Jones</u>	University of Queensland
<u>Ron Cenfetelli</u>	University of British Columbia
<u>Dennis Galletta</u>	University of Pittsburgh
<u>Bin Gu</u>	Boston University
<u>Sirkka Jarvenpaa*</u>	University of Texas at Austin
<u>Gerald (Jerry) Kane</u>	Boston College
<u>Atreyi Kankanhalli</u>	National University of Singapore
<u>Mark Keil</u>	Georgia State University
<u>Prabhudev Konana</u>	University of Texas at Austin
<u>Xinxin Li</u>	University of Connecticut
<u>Likoebe Maruping</u>	Georgia State University
<u>Shaila Miranda</u>	University of Oklahoma
<u>Sunil Mithas</u>	University of South Florida
<u>Eivor Oborn</u>	University of Warwick
<u>Gal Oestreicher-Singer</u>	Tel Aviv University
<u>Jeffrey Parsons</u>	Memorial University of Newfoundland
<u>H. R. Rao</u>	University of Texas at San Antonio
<u>T. Ravichandran</u>	Rensselaer Polytechnic Institute
<u>Saonee Sarker</u>	University of Virginia
<u>Chee-Wee Tan</u>	Copenhagen Business School
<u>Jason Thatcher*</u>	Temple University
<u>James Y. L. Thong</u>	Hong Kong University of Science and Technology
<u>Amrit Tiwana</u>	University of Georgia
<u>Siva Viswanathan</u>	University of Maryland
<u>Jonathan Wareham*</u>	ESADE
<u>Sean Xin Xu</u>	Tsinghua University

# Major Journals: MISQ & JMIS

- MISQ: A+ journal, #1 in MIS
  - behavior/management focused traditionally (most SEs)
  - recent focus in business analytics & data sciences (SEs: HRR, GA, IB, PK, JP) → selecting the right SEs/AEs
  - Computational design science: application-inspired novelty (algorithm, representation, framework, HCI) + societal impact → significant content & mature writing (40+ pages)
  - MIS-specific lit review + methodology/framework/design “theory” + contribution to KB + principles (research abstraction) → right packaging
- JMIS: A journal, #3 in MIS
  - Same as above; more system driven
  - Zwass + Nunamaker; HICSS special issue





# Major Journals: Chen, AI Lab Computational Design Science (CDS) Papers in MISQ, 2008+

## **A Deep Learning Approach for Recognizing Activity of Daily Living (ADL) for Senior Care: Exploiting Interaction Dependency and Temporal Patterns**

Hongyi Zhu, Sagar Samtani, Randall A. Brown, and Hsinchun Chen Forthcoming, 2020

Health Analytics; Deep Learning

2020

■ [j257]     Michael Chau, Tim M. H. Li, Paul W. C. Wong, Jennifer J. Xu, Paul Siu Fai Yip, Hsinchun Chen: **Finding People with Emotional Distress in Online Social Media: A Design Combining Machine Learning and Rule-Based Classification.** MIS Q. 44(2) (2020)

Health Analytics



[\[-\] 2010 - 2019](#)

2019

■ [j254]     Victor A. Benjamin, Joseph S. Valacich, Hsinchun Chen: **DICE-E: A Framework for Conducting Darknet Identification, Collection, Evaluation with Ethics.** MIS Q. 43(1) (2019)

Security Analytics





2017

■ [j242]     Yu-Kai Lin, Hsinchun Chen, Randall A. Brown, Shu-Hsing Li, Hung-Jen Yang: **Healthcare Predictive Analytics for Risk Profiling in Chronic Care: A Bayesian Multitask Learning Approach.** MIS Q. 41(2): 473-495 (2017)





Health Analytics

Special Issue, Business Analytics; 5250 citations

2012

■ [j214]     Hsinchun Chen, Roger H. L. Chiang, Veda C. Storey: **Business Intelligence and Analytics: From Big Data to Big Impact.** MIS Q. 36(4): 1165-1188 (2012)

2010

■ [j178]     Ahmed Abbasi, Zhu Zhang, David Zimbra, Hsinchun Chen, Jay F. Nunamaker Jr.: **Detecting Fake Websites: The Contribution of Statistical Learning Theory.** MIS Q. 34(3): 435-461 (2010)

[\[-\] 2000 - 2009](#)

Security Analytics; Best Paper, ICIS, 2010

2008

■ [j139]     Ahmed Abbasi, Hsinchun Chen: **CyberGate: A Design Framework and System for Text Analysis of Computer-Mediated Communication.** MIS Q. 32(4): 811-837 (2008)

# Major Journals: Health IT & Analytics Special Issue, March 2020

MIS  
Quarterly

SPECIAL ISSUE: CHRONIC DISEASE

## CONNECTING SYSTEMS, DATA, AND PEOPLE: A MULTIDISCIPLINARY RESEARCH ROADMAP FOR CHRONIC DISEASE MANAGEMENT<sup>1</sup>

**Indranil Bardhan**

Department of Information, Risk and Operations Management, McCombs School of Business, The University of Texas at Austin, Austin, TX 78705 U.S.A. {indranil.bardhan@mcombs.utexas.edu}

**Hsinchun Chen**

MIS Department, Eller College of Management, The University of Arizona, Tucson, AZ 85721-0108 U.S.A. {hsinchun@email.arizona.edu}

**Elena Karahanna**

MIS Department, Terry College of Business, The University of Georgia, Athens, GA 30602 U.S.A. {ekarah@uga.edu}

### Special Issue: The Role of Information Systems and Analytics in Chronic Disease Prevention and Management

#### Special Issue Articles

Trajectories of Repeated Readmissions of Chronic Disease Patients: Risk Stratification, Profiling, and Prediction

Ofir Ben-Assuli and Rema Padman  
(pp. 201-226; DOI: 10.25300/MISQ/2020/15101)

Chronic Disease Management: How IT and Analytics Create Healthcare Value Through the Temporal Displacement of Care

Steve Thompson, Jonathan Whitaker, Rajiv Kohli, and Craig Jones  
(pp. 227-256; DOI: 10.25300/MISQ/2020/15085)

Go to You Tube and Call Me in the Morning: Use of Social Media for Chronic Conditions

Xiao Liu, Bin Zhang, Anjana Susarla, and Rema Padman  
(pp. 257-283; DOI: 10.25300/MISQ/2020/15107)

A Data Analytics Framework for Smart Asthma Management Based on Remote Health Information Systems with Bluetooth-Enabled Personal Inhalers

Junbo Son, Patricia Flatley Brennan, and Shiyu Zhou  
(pp. 285-303; DOI: 10.25300/MISQ/2020/15092)

A Comprehensive Analysis of Triggers and Risk Factors for Asthma Based on Machine Learning and Large Heterogeneous Data Sources

Wenli Zhang and Sudha Ram  
(pp. 305-349; DOI: 10.25300/MISQ/2020/15106)

Examining How Chronically Ill Patients' Reactions to and Effective Use of Information Technology Can Influence How Well They Self-Manage Their Illness

Azadeh Savoli, Henri Barki, and Guy Paré  
(pp. 351-389; DOI: 10.25300/MISQ/2020/15103)

The Effects of Participating in a Physician-Driven Online Health Community in Managing Chronic Disease: Evidence from Two Natural Experiments

Qianqian Ben Liu, Xiaoxiao Liu, and Xitong Guo  
(pp. 391-419; DOI: 10.25300/MISQ/2020/15102)

# Major Journals: MISQ CDS Common Issues

- MISQ, My Experience: no paper/involvement before 2008 (no SE in design science); Abbasi 2008 (CyberGate), 2010 (AZProtect, ICIS best paper); Guest Editor, BI&A special issue, 2010-2012 (Straub); SE 2016-2019 (Rai); Guest Editor, Health IT/Analytics special issue, 2016-2020 (Rai)
- Design Science paper common issues:
  - Where is the theory? Is this MIS? (early reviewers' critiques)
  - Few qualified/sympathetic design science SEs, AEs, reviewers. (overly critical)
  - Long review cycle (2-4 rounds/years) and uncertainty (rejection at late round).
  - ➔ but
  - BI&A and data sciences are hot, in society and in b-school curriculum!
  - Young MIS CDS scholars need 1-2 MISQ/JMIS papers accepted or in deep round.
  - Mid-career MIS CDS scholars need 3-5 MISQ/JMIS papers for tenure.

# Major Journals: MISQ CDS Paper Template

- Computational design science (Chen in Rai, 2017): application-inspired novelty (algorithm, representation, framework, HCI) + emerging high-impact problems
- Significant content & mature writing (40+ pages)
- MIS-specific lit review (3-4 pages) → Who/what had (been) published in MISQ/ISR/JMIS (10-20 MIS references, taxonomy, analytics relevance)
- Methodology/framework/design “theory” (2-3 pages) → underlying methodological foundation (not behavioral theory of +/- hypotheses), e.g., Systematic Functional Linguistic Theory, Kernel Learning Theory, etc.
- Contribution to KB + principles (research abstraction; 2-3 pages) → What have been learned about the design, use and general knowledge gained?  
→ Carefully study sample MISQ DS papers, e.g., (Abbasi, 2008; 2010).

# Major Grants: NIH, DARPA, DHS, IARPA



- NIH: NLM is informatics-focused; “translational” research with some application-inspired health-related novelty; need pubs and networking in AMIA/JAMIA; strong health informatics (NLM) tradition and turf (strong personality) → Chen as NLM Scientific Counselor, 2002-2006
  - DOD/DARPA: was innovative, basic/foundational, long-term (ARPA Net); now mission-critical, system-driven, short-term; commercial company (defense contractor) as prim, academic as sub; bi-monthly milestones/metrics/reporting → Chen early success with DARPA/IARPA/DHS for COPLINK/Dark Web research
  - DHS, IARPA: similar to DARPA, but aspiring; lesser scientific quality (strong personality)
- Not my focus any more! (Need to smell like them.)





# Major Grants: NSF Org Chart




**NATIONAL SCIENCE FOUNDATION (\$8.3B)**

**OFFICE OF THE DIRECTOR**  
703.292.8000

 **Sethuraman Panchanathan**  
Director

 **Vacant**  
Deputy Director

 **F. Fleming Crim**  
Chief Operating Officer

**NATIONAL SCIENCE BOARD (NSB)**  
703.292.7000

 **Diane L. Souvaine**  
Chair

 **Ellen Ochoa**  
Vice Chair

**OFFICE OF INSPECTOR GENERAL (OIG)**

 **Allison C. Lerner**  
Inspector General  
703.292.7100

**NATIONAL SCIENCE BOARD OFFICE**


 **John J. Veysay, II**  
Executive Officer  
703.292.7000



# Major Grants: NSF CISE/IIS/III

## CISE

**DIRECTORATE FOR  
COMPUTER &  
INFORMATION SCIENCE &  
ENGINEERING (CISE)**



Margaret Martonos,  
Assistant Director  
Erwin Gianchandani,  
Deputy AD  
703.292.8900

## IIS/OAC

Directorate for Computer & Information Science & Engineering CISE/OAD

Office of Advanced Cyberinfrastructure CISE/OAC

Division of Computing and Communication Foundations CISE/CCF

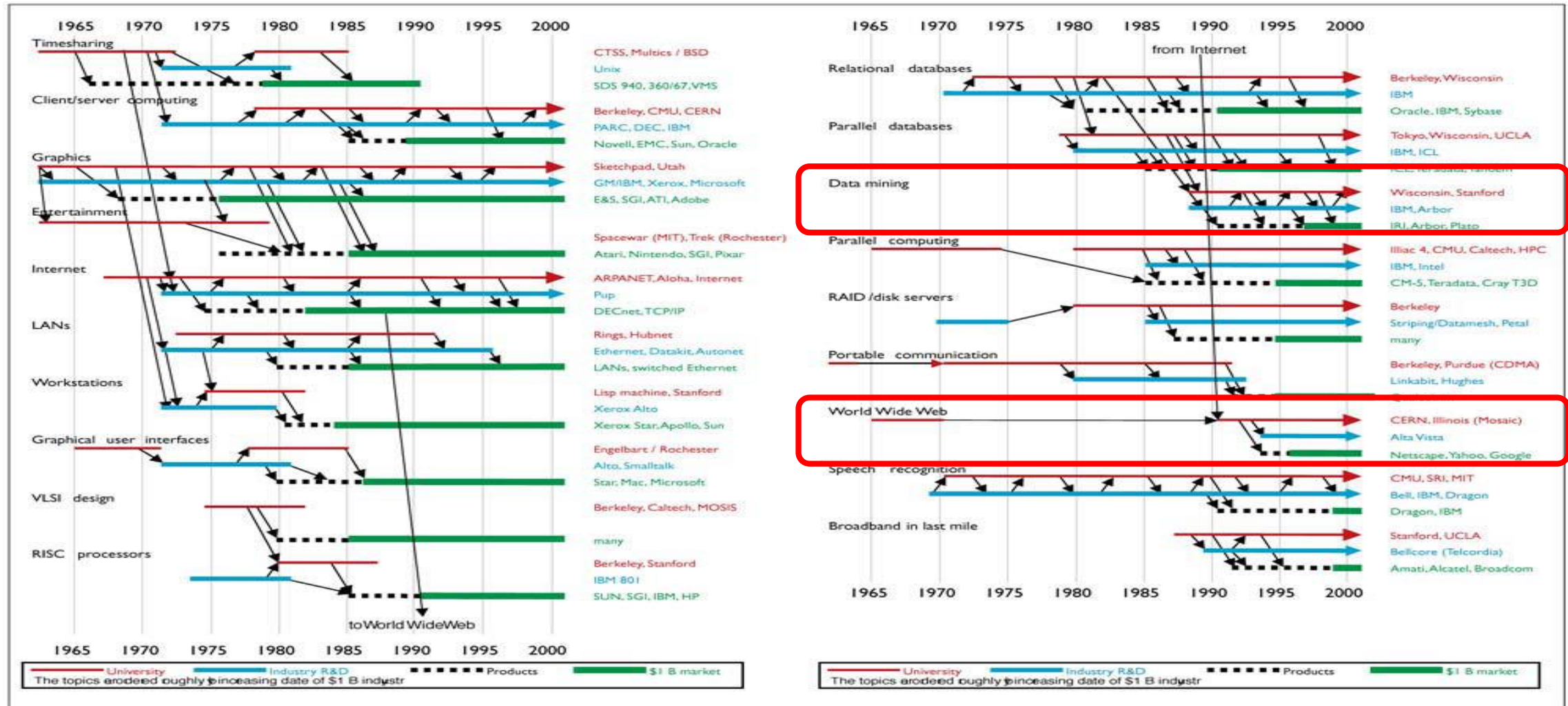
Division of Computer and Network Systems CISE/CNS

Division of Information and Intelligent Systems CISE/IIS

## III

- IIS: Human-Centered Computing (HCC)
- IIS: Information Integration and Informatics (III)
- IIS: Robust Intelligence (RI)
- OAC: OAC Core Research (OAC Core)

# Major Grants: NSF CISE/IT Societal Impacts (NAS)



Source: From [6], reprinted with permission from the National Academy of Sciences, courtesy of the National Academies Press, Washington D.C. © 2003.

**University research** → **Industry R&D** → **Products** → **\$1B Market (job and wealth creation)**

# Major Grants: NSF Programs

- CORE: NSF CISE/IIS/III CORE most relevant to fundamental research in AI, machine learning, WWW, data sciences, NLP; acceptance rate 6-8%, highly competitive, critical young CS reviewers → IIS Core (\$100M/yr)
- OAC: NSF CISE/OAC relevant to applied cyberinfrastructure for sciences; acceptance rate 20-30%, less competitive, reviewers including CS, SBE, and domain sciences → DIBBs, CICI (\$25M-30M/yr; my focus)
- Applied Programs: Many emerging cross-directorate (e.g., EHR, SBE, CISE) and cross-agency (e.g., NSF, NIH, DOD) high-impact applied research programs (e.g., security, health); acceptance rate 15-20%, less competitive, reviewers including CS, SBE, and SME → SaTC, SFS, CCRI, SCH, BIGDATA, I-DSN, National AI Institutes (\$50M-100M/yr; my focus)
- Young Scholars: Many opportunities for early-career scholars; acceptance rate 10-20%, competitive, for early career; valuable for obtaining tenure! → CRII, CAREER + EAGER (\$200K-\$1M for each award)

# Major Grants: NSF Proposal Observations

- Computational Design Science (CDS) has excellent chance for successful proposals (CISE). → in general, not so much for behavioral or economics MIS researchers (SBE; too basic, too incremental, not novel).
- “Business” (finance, accounting, marketing) school research is not considered STEM. → need to position for larger societal/STEM problems.
- CDS research needs to compete with CS researchers (“locusts” in emerging technical fields); deep & novel domain application for emerging societal problems could be viable. → my approach at least, for the past 30 years: digital library, intelligence, health, cybersecurity, etc.
- Need application or domain-inspired novelty for applied cross-directorate programs. → senior Ph.D. students; last 1-2 dissertation chapters
- A lab or center can help with sustainable advantage and funding. → developing collection, prototype system, etc.; structure & organizational memory

# Major Grants: NSF Proposal Template

- Proposal title: short and succinct; need a multi-disciplinary team
- Project summary: Summarize problems and approach; include IM + BI
- Main text (15 pages)
  - Need mature writing; good diagrams
  - Need methodological/algorithmic novelty (IM, 60%); need strong impacts (BI, 40%)
  - Need good lit review (state-of-the-art) & promising preliminary results
- CV: need relevant ACM/IEEE references; MISQ/ISR pubs help very little
- Others: Good to have office support, e.g., budget, facilities, DMP, routing, etc.

## TABLE OF CONTENTS

For font size and page formatting specifications, see PAPPG section II.B.2.

	Total No. of Pages	Page No.* (Optional)*
Cover Sheet for Proposal to the National Science Foundation		
Project Summary (not to exceed 1 page)	1	_____
Table of Contents	1	_____
Project Description (Including Results from Prior NSF Support) (not to exceed 15 pages) <b>(Exceed only if allowed by a specific program announcement/solicitation or if approved in advance by the appropriate NSF Assistant Director or designee)</b>	15	_____
References Cited	6	_____
Biographical Sketches (Not to exceed 2 pages each)	8	_____
Budget (Plus up to 3 pages of budget justification)	6	_____
Current and Pending Support	4	_____
Facilities, Equipment and Other Resources	2	_____
Special Information/Supplementary Documents (Data Management Plan, Mentoring Plan and Other Supplementary Documents)	2	_____
Appendix (List below.) <b>(Include only if allowed by a specific program announcement/solicitation or if approved in advance by the appropriate NSF Assistant Director or designee)</b>	_____	_____

# Major Grants: NSF General Advice for CDS Scholars

- Develop methodological novelty and application-specific strengths over your career. → world-class excellence vs. other CS scholars
- Train your Ph.D. students well. → their last 2 dissertation chapters could be fundable; they can be trained to write proposals (scale & efficiency)
- Build a center/lab/group. → more sustainable and impressive (common in CS, ECE, MED)
- Improve your grantsmanship. → get to know your PDs and become frequent NSF panelists (getting into their heads)
- Improve your success rate to 30% (one in 3). → target repeating programs for re-submissions
- Monitor and anticipate current and emerging programs. → prepare the next proposals; repeat the cycle!

# Parting Thoughts: Hard Work + A Bit of Luck

- Societal Impact > Academic Impact
  - Looking for high-impact societal problems (NYT, WSJ, The Economists)
- IT > MIS
  - MIS is a smaller subfield within broader IT/computing.
- CISE > SBE
  - Computational Design Science can make a difference.
- New > Old
  - Looking for new, interesting, unknown problems
- EQ > IQ
  - Hard work, discipline, aspiration, etc. always beat raw talent. Plus a bit of luck!



For questions and comments

[hchen@eller.Arizona.edu](mailto:hchen@eller.Arizona.edu)

<http://ai.Arizona.edu>

