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)

ARIZONA Eller

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
sgriffin@nsf.gov

NIJ National Institute of Justice
STRENGTHEN SCIENCE. ADVANCE JUSTICE.

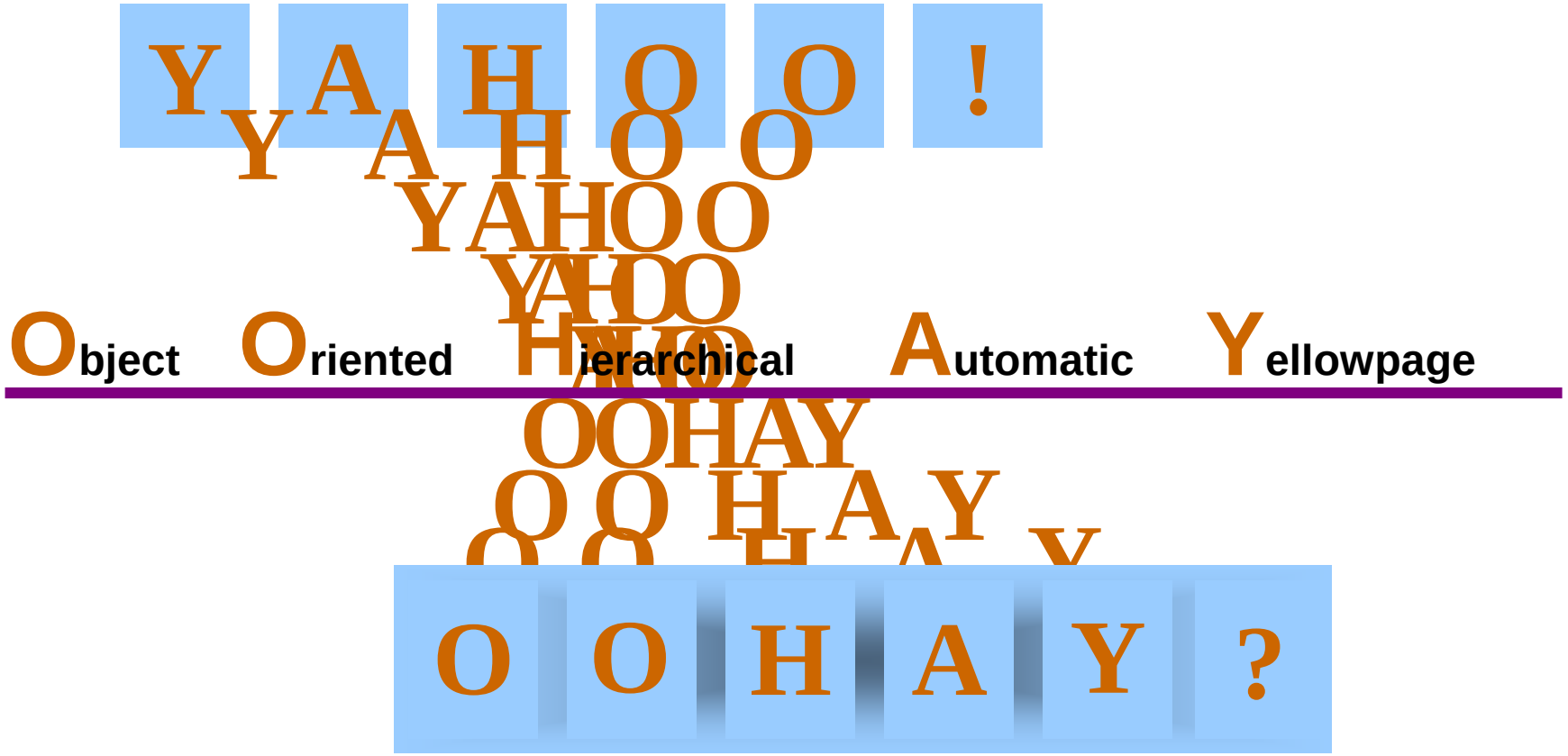
Digital Government (DigitalGov)

Program Solicitation
NSF 04-521
Replaces Document 02-156

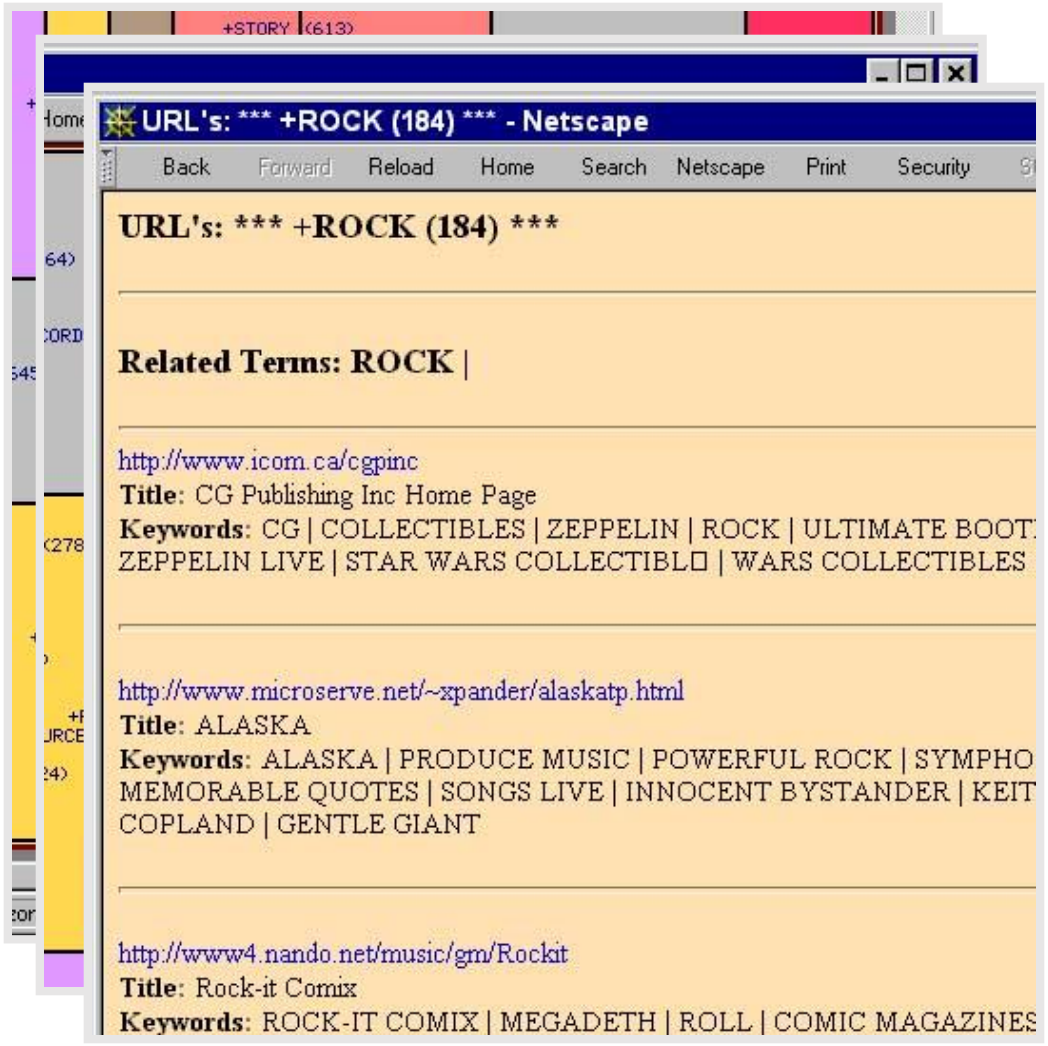
National Science Foundation
Directorate for Computer and Information Science and Engineering
Division of Information and Intelligent Systems

DLI: Visualization Research in AI Lab

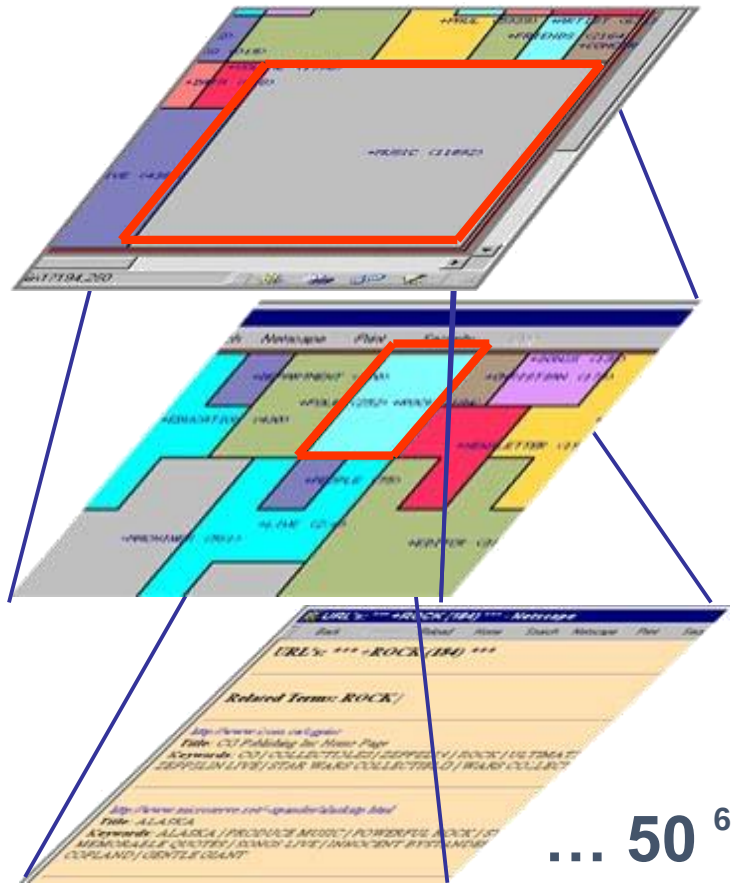
From YAHOO! To OOHAY?



Visualization Research in AI Lab

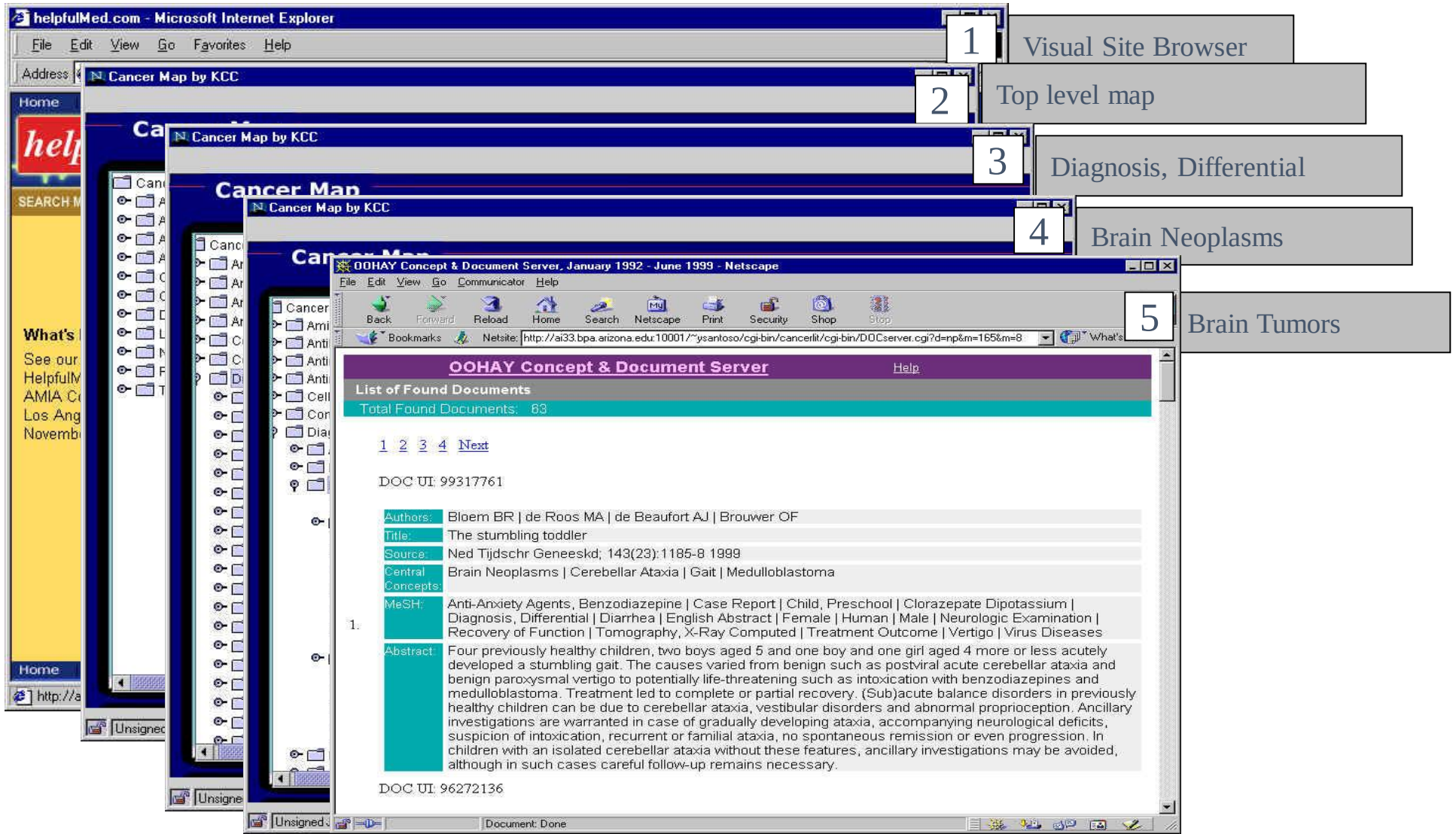


OOHAY: Visualizing the Web



… 50 [6]

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



Taiwan Health Topic Map: 500K news articles

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

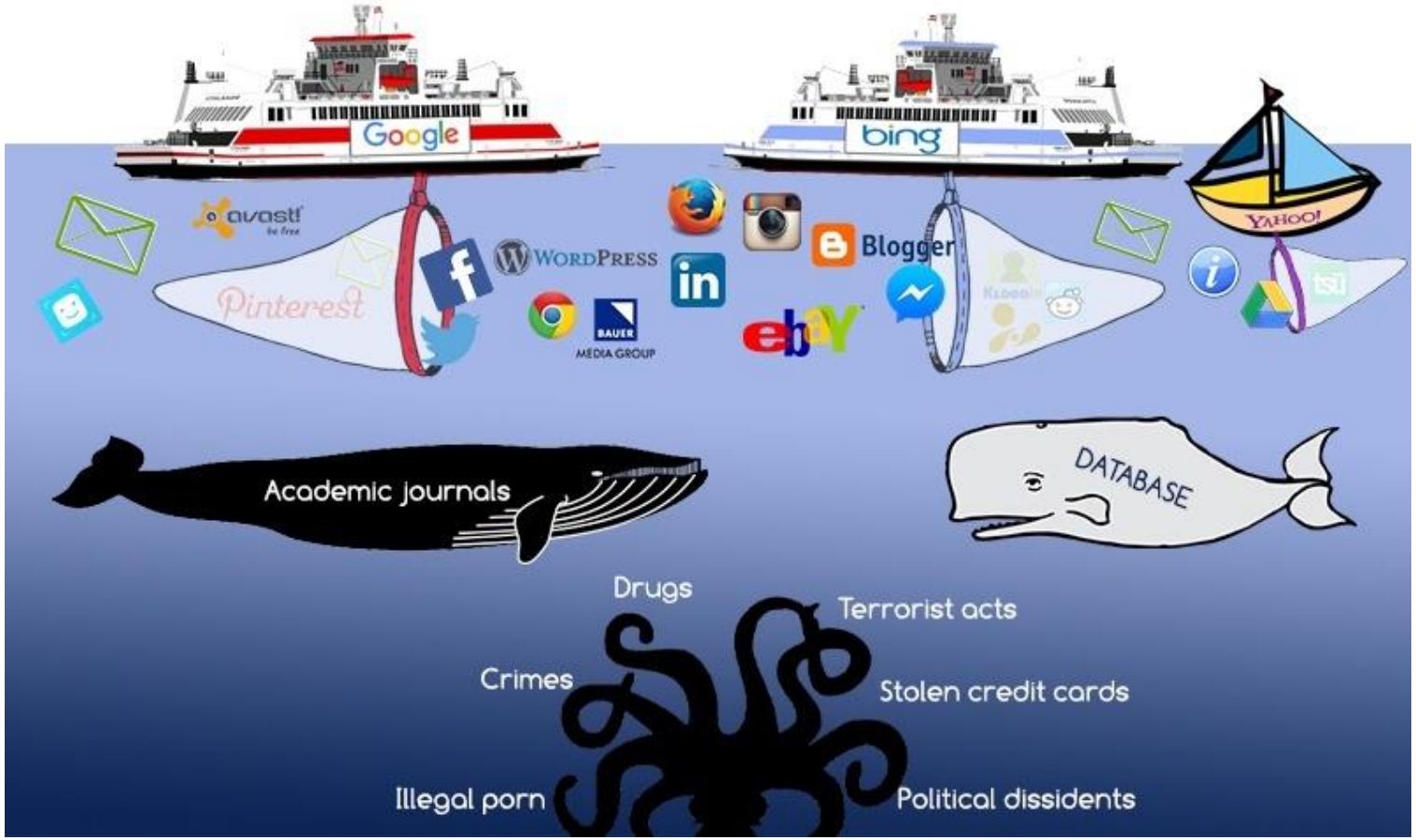
Surface Web

Deep Web

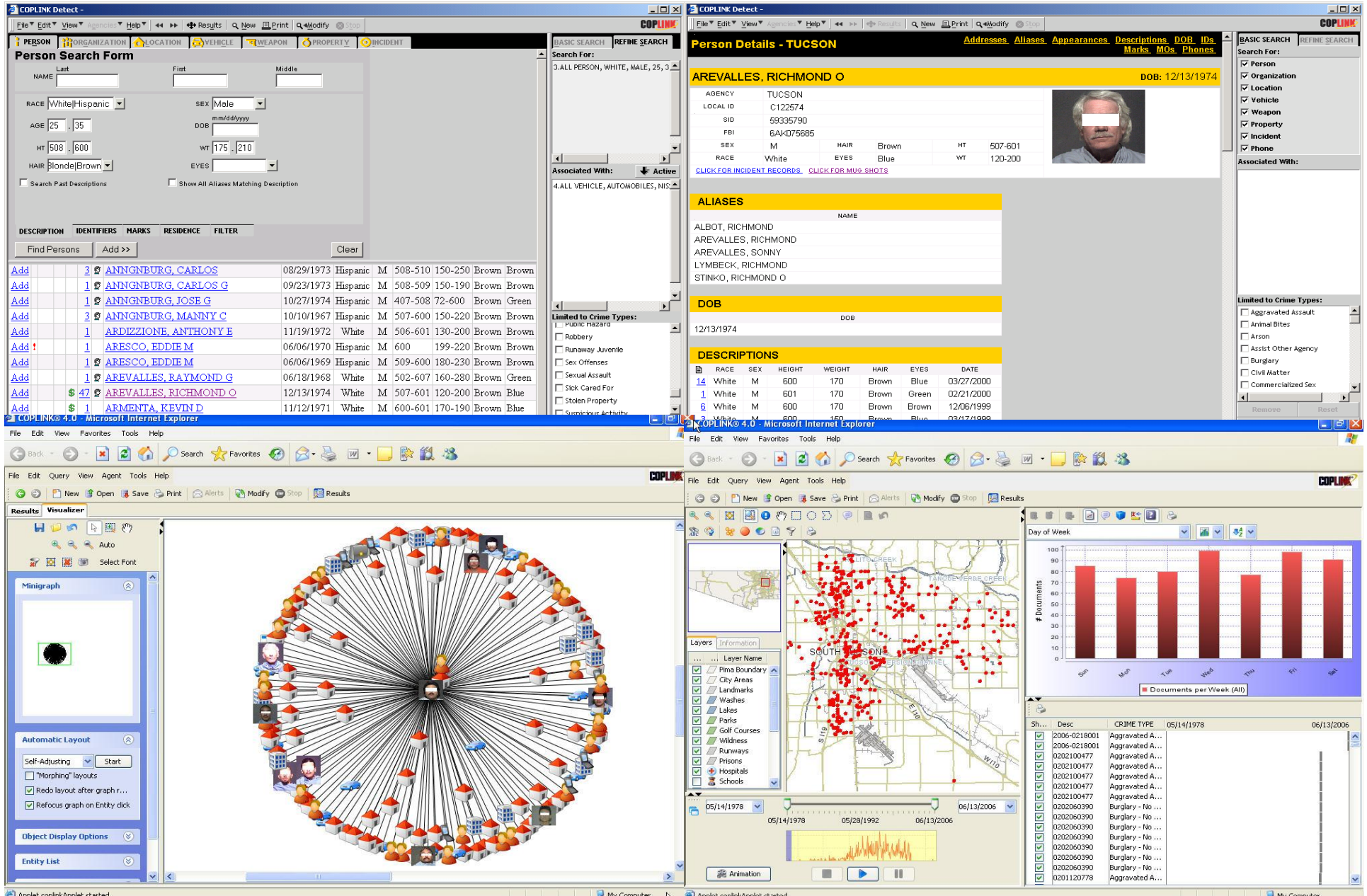
Dark Web

DarkNet

Hacker Web



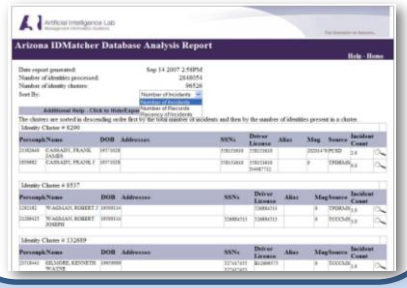
COPLINK: Crime Data Mining (1997-2009)



COPLINK Identity Resolution and Criminal Network Analysis

Cross-jurisdictional Information Sharing/Collaboration

Arizona IDMatcher



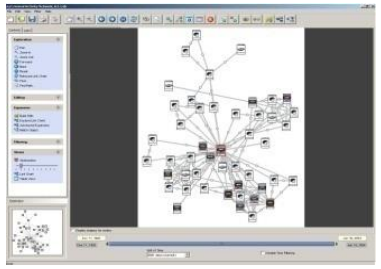
Law-enforcement Data

AZ | CA | TX

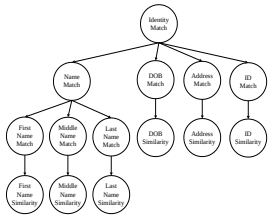
Border Crossing Data (AZ, CA, TX)

Vehicles | People

CAN Visualizer



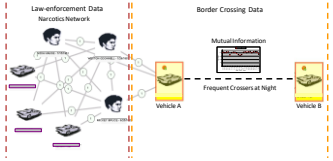
Identity Resolution



Detect false and deceptive identities across jurisdictions using a probabilistic naïve-Bayes based resolution system.

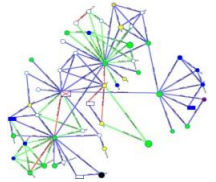
Criminal Network Analysis

High-risk Vehicle Identification



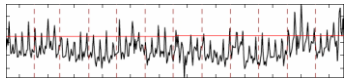
Identify high-risk vehicles using association techniques like mutual information using border crossing and law enforcement data.

Criminal Link Prediction



Predict interaction between individuals and vehicles using link prediction techniques to identify high-risk border crossers.

Suspect Traffic Burst Detection



Detect real-time anomalies and threats in border traffic using Markov switching and other models.

* 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

Arts & Ideas

SATURDAY, NOVEMBER 2, 2002

The New York Times

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 Internet-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 snipers terrorized the area, investigators used everything from specialized ballistics testing to geographic and criminal profiling to radio and television announcements to track them down. Then, in what turned out to be the 11th 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 Rapidstart — 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 in the laboratory into a commercial 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's the Google for law enforcement," he said, referring to a speedy popular Internet search engine that, given a couple of words, can find 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 clues to assimilate was daunting. "We were mobilizing a massive effort," said Lt. Mitch Cunningham of the Montgomery County police. "We had tactile 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 that 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, "Shorty did it." The name Shorty 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

IBM i2 COPLINK

Accelerating law enforcement investigations

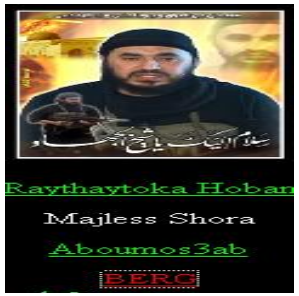
Palantir **($54B, IPO 2020)**

abc NEWS

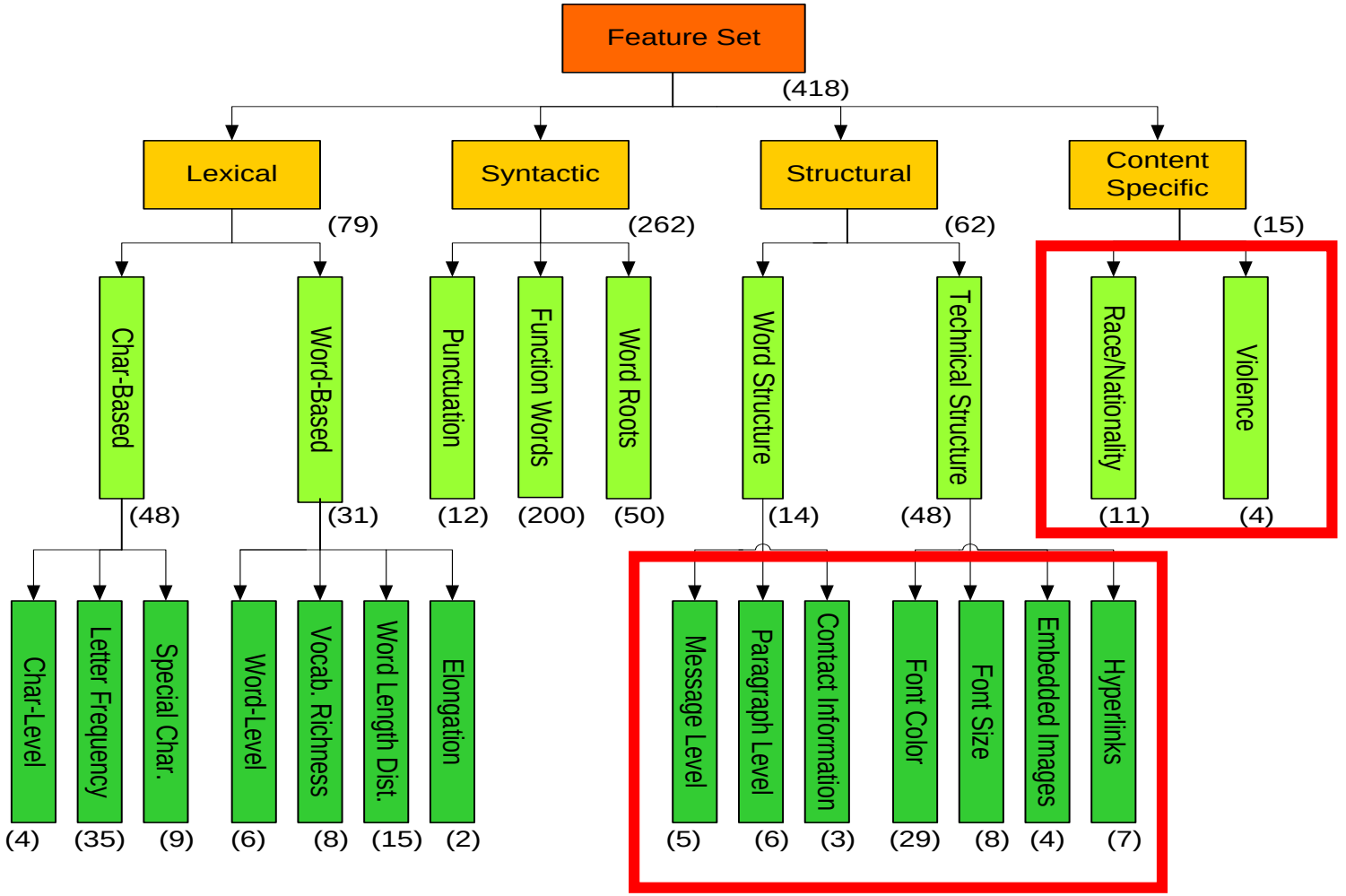
Newsweek

Dark Web: Countering Terrorism (2003-2014)

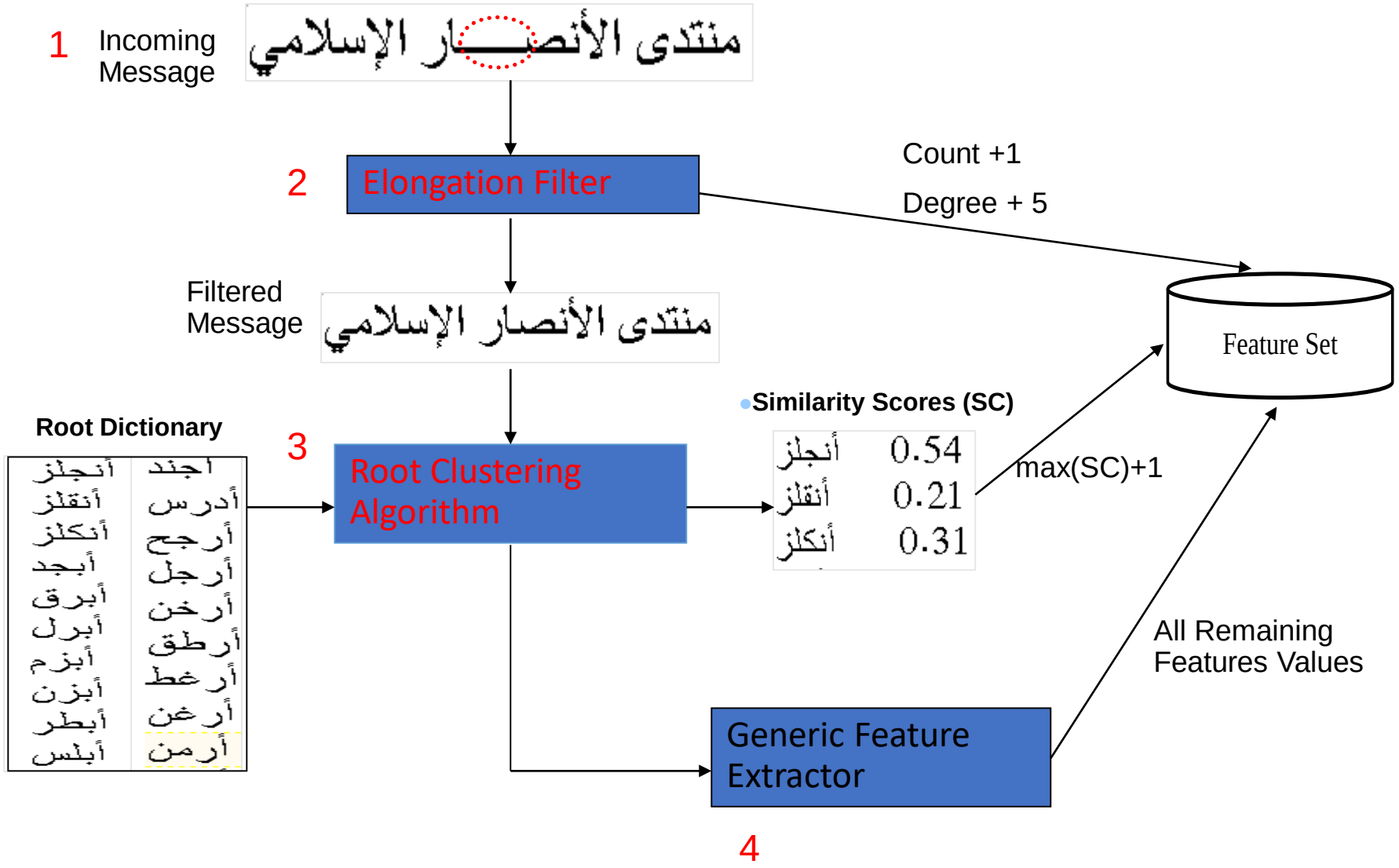
- Dark Web: Terrorists' and cyber criminals' use of the Internet
- Collection: Web sites, forums, blogs, YouTube, etc.
- 20 TBs in size, with close to 10B pages/files/messages (the entire LOC collection: 15 TBs)



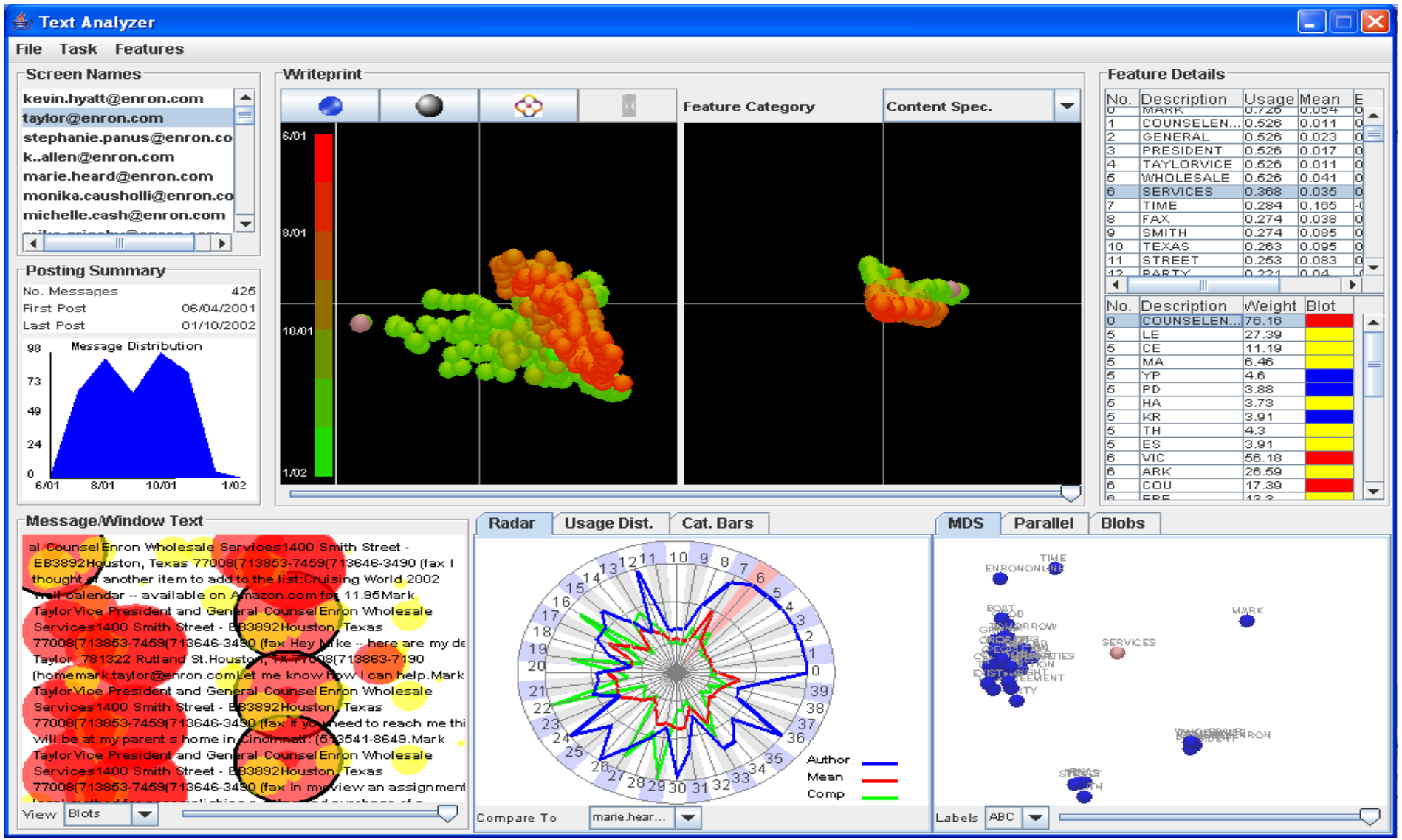
Arabic Writeprint Feature for Authorship Analysis



Arabic Feature Extraction Component



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



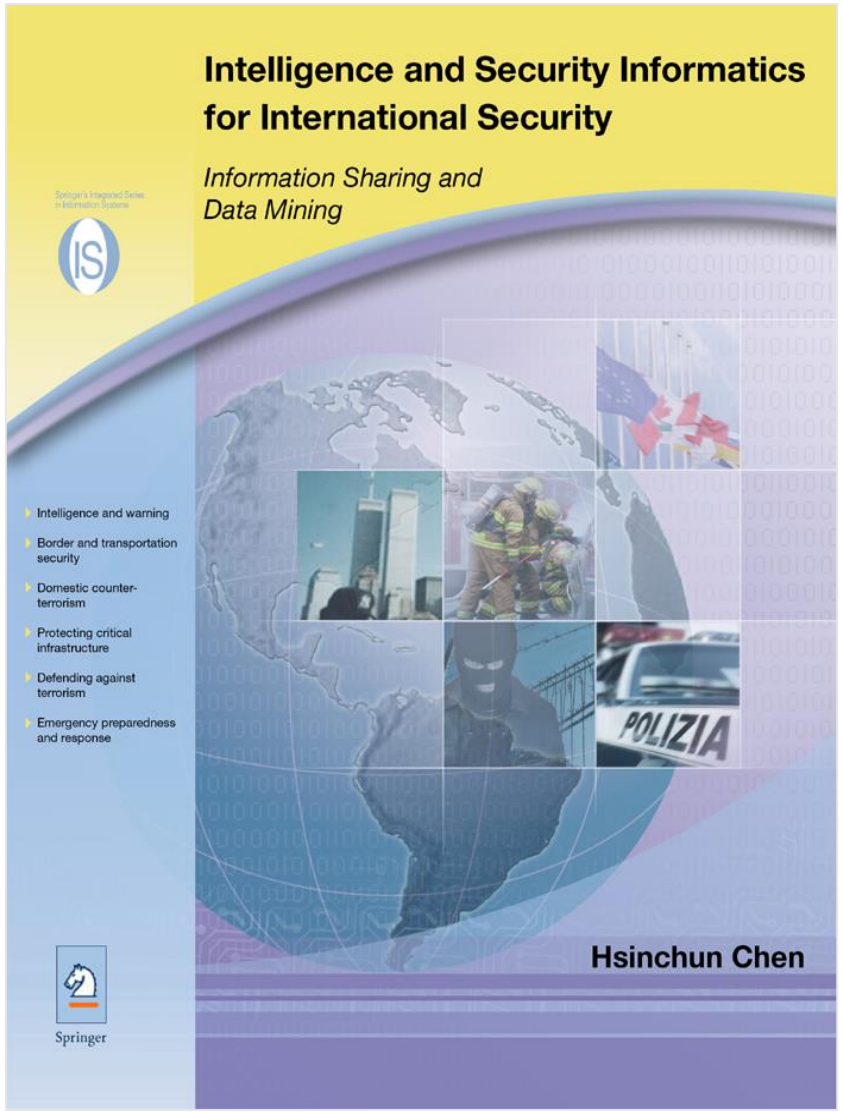
The Dark Web project in the Press

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

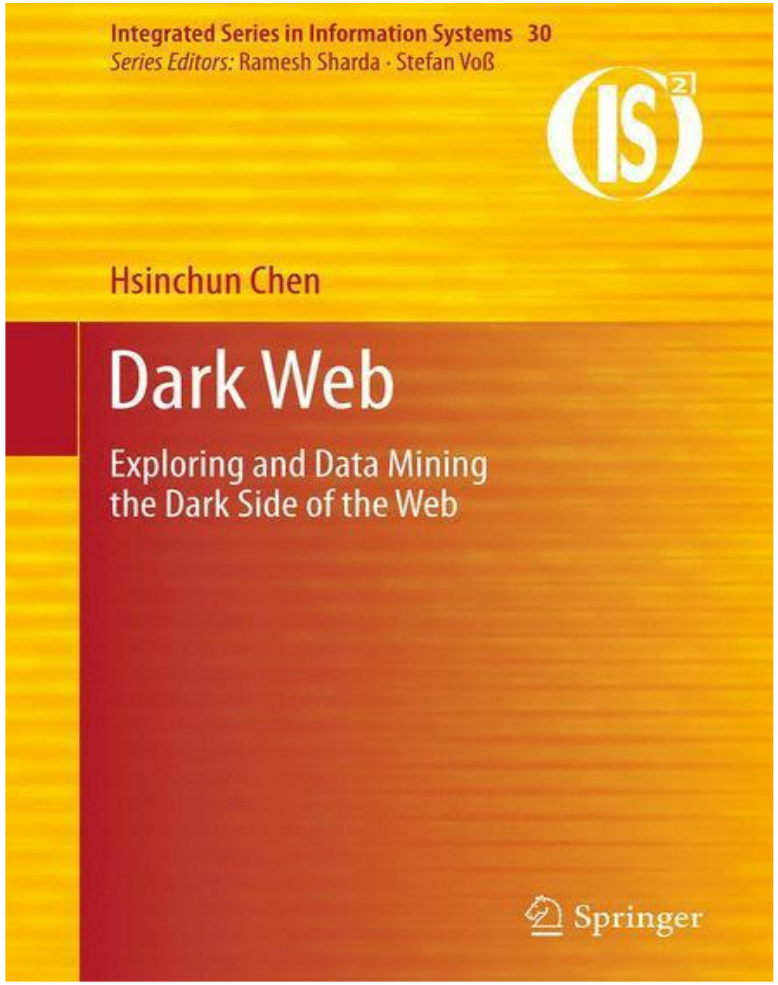


- Intelligence and Security Informatics (ISI) (Chen, *2006*)

- 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

Pivoting to Cyber Security

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

MIS Quarterly SPECIAL ISSUE

DETECTING FAKE WEBSITES: THE CONTRIBUTION OF STATISTICAL LEARNING THEORY[1]

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" "about FDIC"	Concocted	References to Federal Deposit Insurance Corporation rarely appear in concocted bank websites.
		"© 2000-2006"	Concocted	Outdated copyrights often appear in concocted websites.
		"fee calculator"	Concocted	Concocted cargo delivery websites provide competitive phony estimates to lure customers. Legitimate sites typically offer estimates in-person through sales representatives.
		"pay by phone" "call toll free"	Concocted	Fraudsters prefer to engage in online transactions. They rarely offer phone-based payment options.
		"payment history" "password management" "enter your account"	Concocted	Concocted websites do not provide considerable support for returning customers since they generally do not have any.
	Lexical measures	Average sentence length	Concocted	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 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	Concocted web pages contain many misspellings and grammatical mistakes.
		"frauduluent"		
		"recieve the"		
		"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

Represent each page a with the vectors :

$x_a = \{\mathrm{Sim}_{ave}(a, b_1), \ldots, \mathrm{Sim}_{ave}(a, b_p)\}$; $y_a = \{\mathrm{Sim}_{max}(a, b_1), \ldots, \mathrm{Sim}_{max}(a, b_p)\}$

Where :

$$\mathrm{Sim}(a,k) = \lambda\left(\left(1-\frac{|\mathrm{lv}_a - \mathrm{lv}_k|}{\mathrm{lv}_a + \mathrm{lv}_k}\right)+\left(1-\frac{|\mathrm{in}_a - \mathrm{in}_k|}{\mathrm{in}_a + \mathrm{in}_k}\right)+\left(1-\frac{|\mathrm{out}_a - \mathrm{out}_k|}{\mathrm{out}_a + \mathrm{out}_k}\right)\right)+(1-\lambda)\left(1-\frac{1}{n}\sum_{i=1}^{n}\frac{|a_i - k_i|}{a_i + k_i}\right)$$

$$\mathrm{Sim}_{ave}(a,b) = \frac{1}{m}\sum_{k=1}^{m}\mathrm{Sim}(a,k)$$

$$\mathrm{Sim}_{max}(a,b) = \underset{k \in \text{pages in site b}}{\arg\max}\ \mathrm{Sim}(a,k)$$

For :

$b \in p$ web sites in the training set; $k \in m$ pages in site b; $a_1, \ldots a_n$ and $k_1, \ldots k_n$ are page a and k's feature vectors;

lv_a, in_a, and out_a are the page level and number of in/out links for page a;

The similarity between two pages is defined as the inner product between their two vectors x_1, x_2 and y_1, y_2 :

$$\text{Linear Composite Kernel} : K(x_1 + y_1, x_2 + y_2) = \frac{\langle x_1, x_2\rangle}{\sqrt{\langle x_1, x_1\rangle\langle x_2, x_2\rangle}} + \frac{\langle y_1, y_2\rangle}{\sqrt{\langle y_1, y_1\rangle\langle y_2, y_2\rangle}}$$

Figure 4. Linear Composite SVM Kernel for Fake Website Detection

Category	Comparsion Websites	Legitimate Bank Web Page		Spoof Bank Web Page	
		$\mathrm{Sim}_{ave}(a, b)$	$\mathrm{Sim}_{max}(a, b)$	$\mathrm{Sim}_{ave}(a, b)$	$\mathrm{Sim}_{max}(a, b)$
Legitimate	www.bankofamerica.com	0.495098	1.000000	0.314208	0.392708
	www.citibank.com	0.387834	0.577704	0.247735	0.386828
Fake	beliebo.nl\media\2007\Paypal	0.317838	0.383798	0.522842	0.555874
	65.85.238.84\ebay.monitors\store.ebay	0.257576	0.297611	0.510915	0.577157

Figure 5. Comparing Two Web Pages Against Legitimate and Fake Websites

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

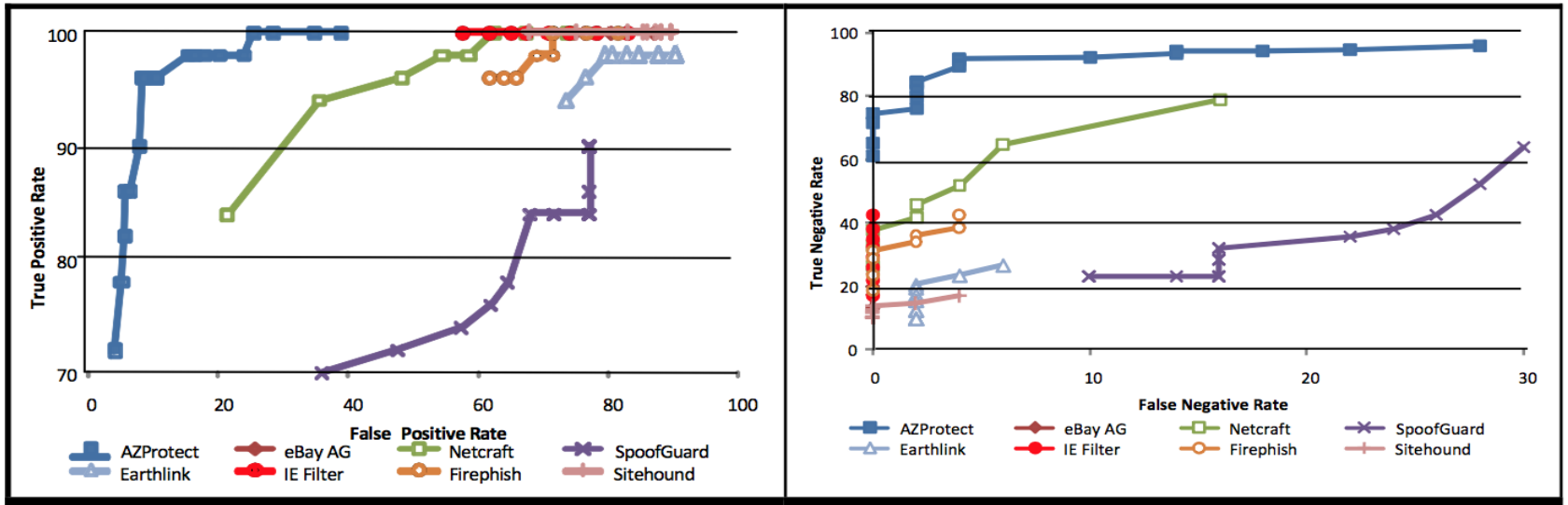


Figure 8. ROC Curves for Classifier and Lookup Systems

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	**89.71**	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

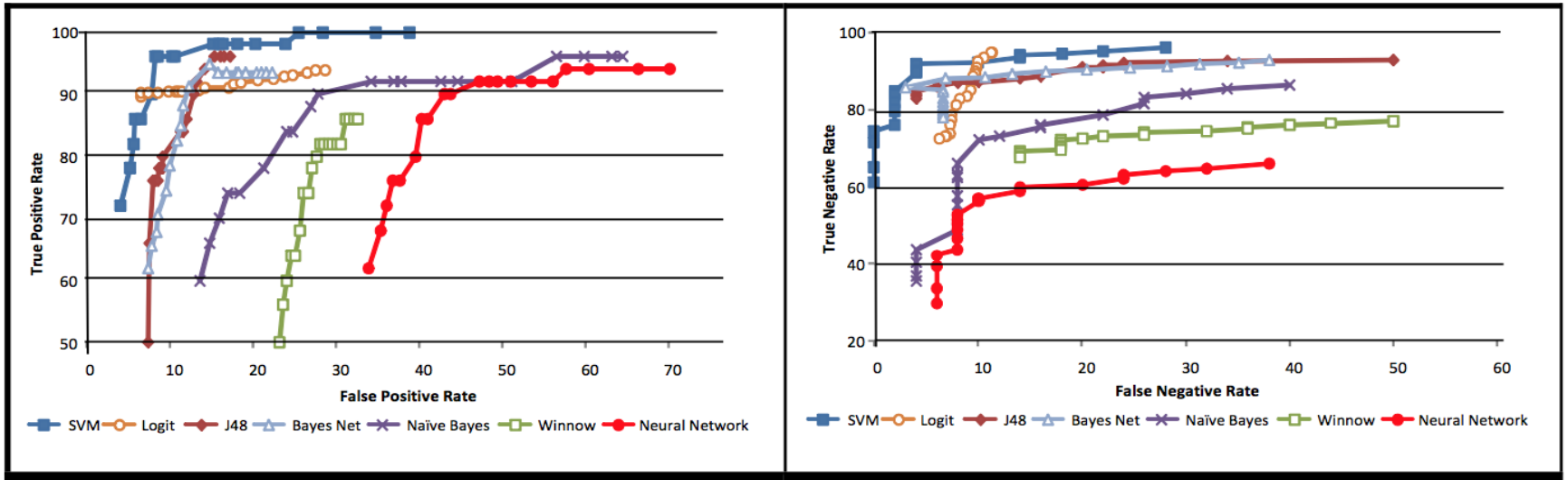


Figure 9. ROC Curves for Various Learning Classifiers

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

nature
doi:10.1038/nature24270

Mastering the game of Go without human knowledge

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

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

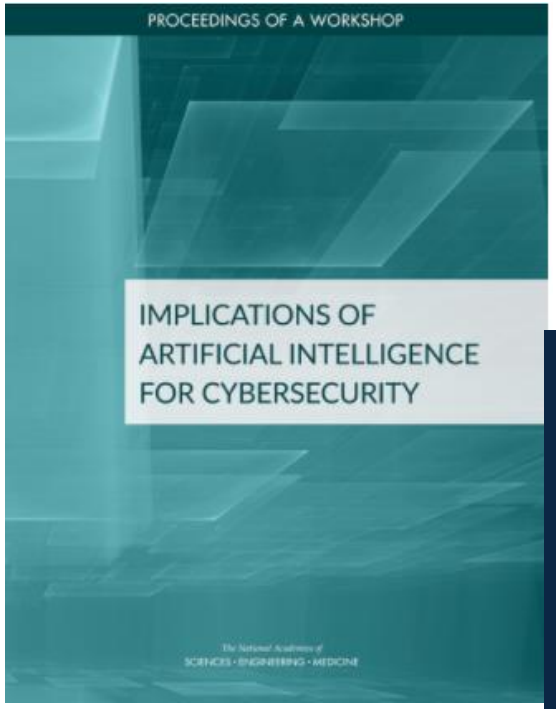
➔

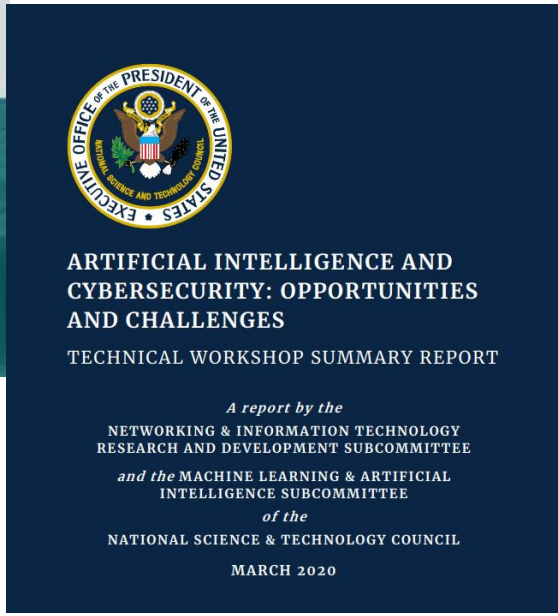
Hacker Web, AZSecure projects at UA/MIS AI Lab (2010-)

AI and Cybersecurity

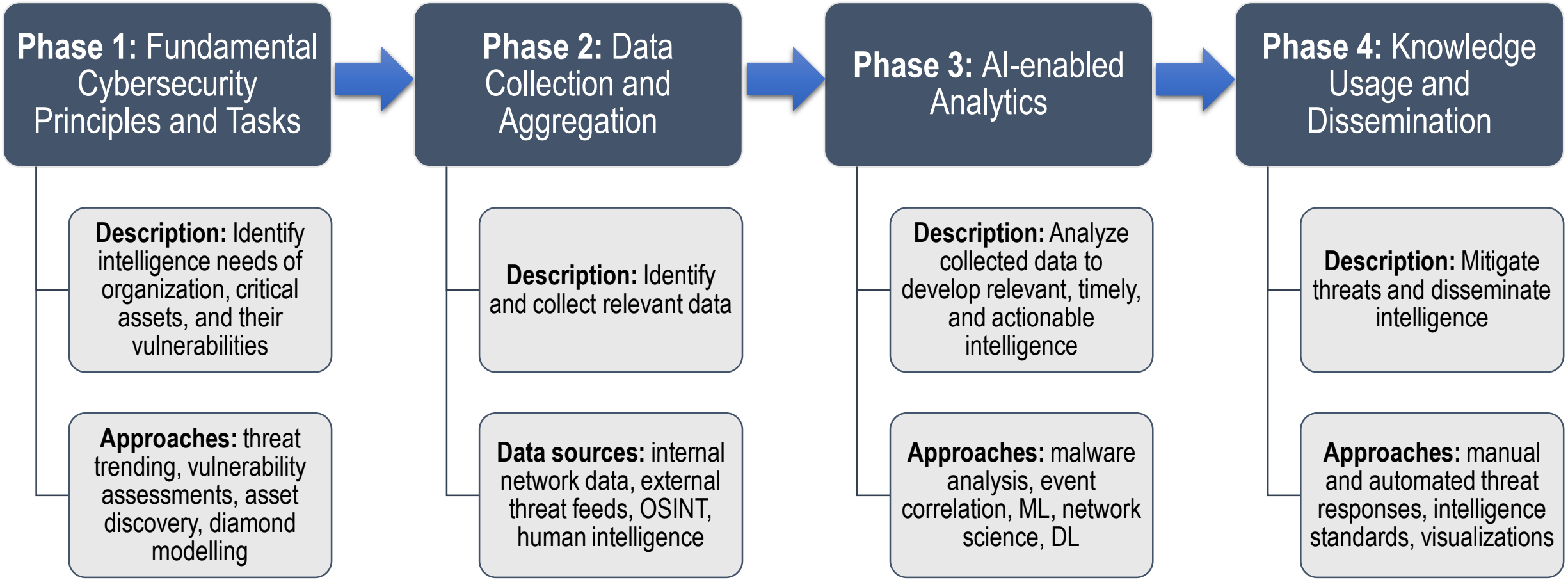
- AI and Cybersecurity → not just buzzwords!
  - Noted as a national security priority by NSF, NSTC, and NAS.

- Role of AI for Cybersecurity :
  1. Automate common cybersecurity tasks
  2. Identify patterns in large datasets





AI for Cybersecurity – An Analytics Approach



MIS Quarterly

SPECIAL ISSUE

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 overemphasizing 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

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

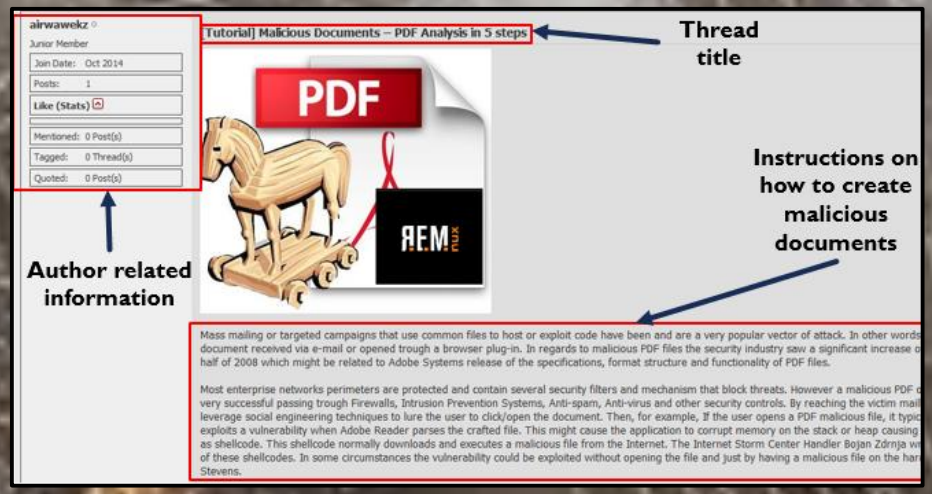
REPLACES DOCUMENT(S):
NSF 19-514

National Science Foundation

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

Hacker Web

Forum post with source code to exploit Mozilla Firefox 3.5.3



Tutorial on how to create malicious documents

Forum post with BlackPOS malware attachment.

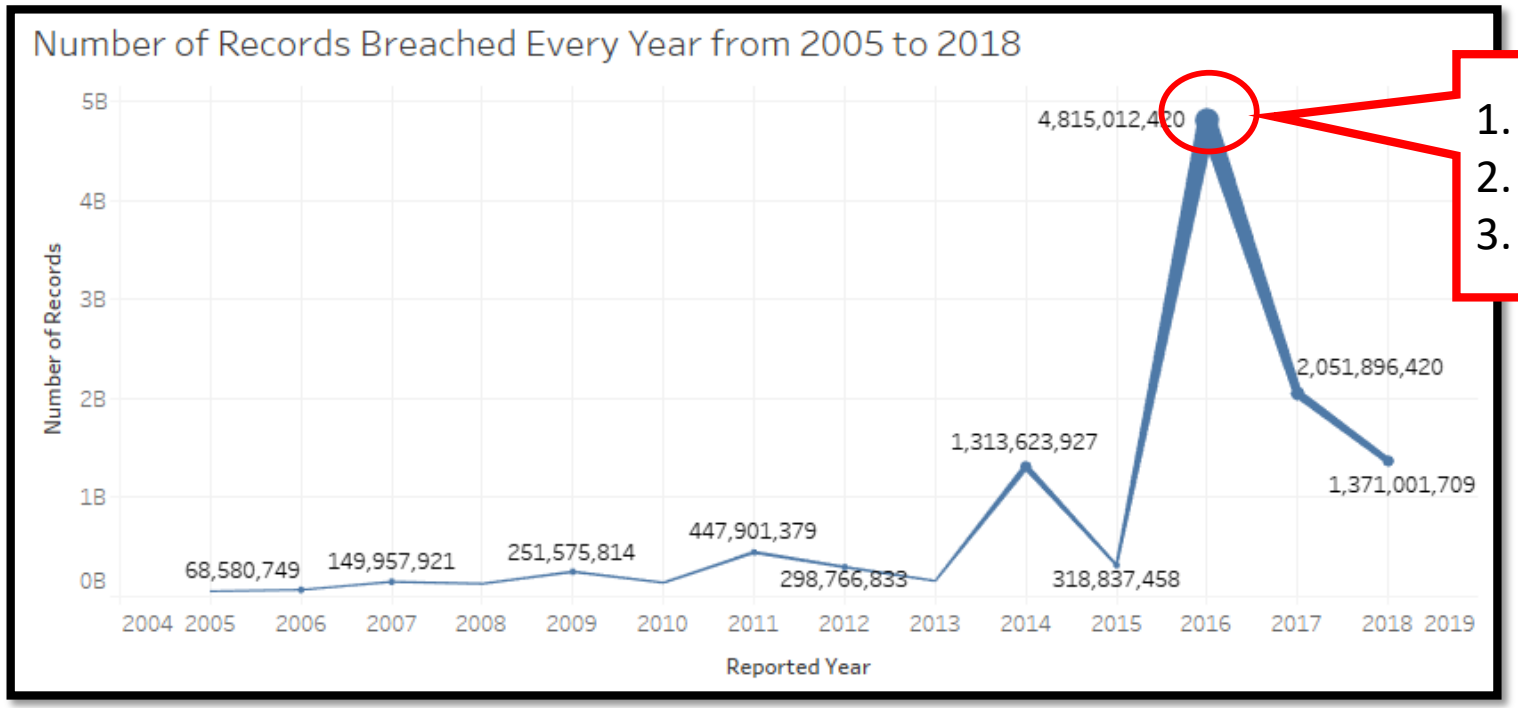
Selected data breaches in 2014

Victim	Date	Ramification
Target	2013.12	**40M** credit/debit cards; **70M** customer records; 46% drop in annual profits (**seller: Rescator**)
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	**83M** accounts
UPS	2014.8	51 stores customers
Dairy Queen	2014.9	395 store systems
Home Depot	2014.9	**56M** credit/debit cards
Jimmy Jones	2014.9	216 store systems
Staples	2014.10	51 store systems

Are your data breached? Do you even know?

Data Breaches since 2005 (FTC, Clearinghouse, 2019)

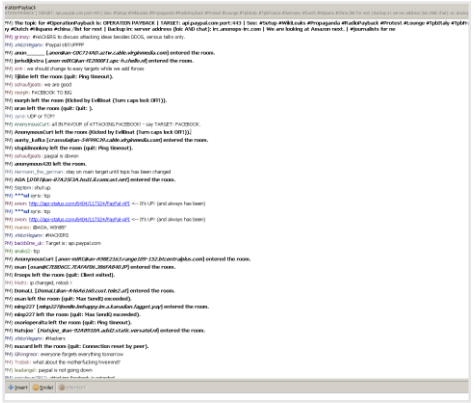
- # of records breached: 11,582,808,013
- # of data breaches: 9,071



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	DarkNet Markets	Carding Shops	IRC Channels
			
Discussion board allowing hackers to freely share malicious tools and knowledge	Markets facilitating the sale of illicit goods (e.g., new exploits, drugs, weapons)	Shops selling sensitive information (e.g., credit cards, SSN's)	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

MIS Quarterly

MIS Quarterly Vol. 43 No. 1, pp. 1-22/March 2019

METHODS ARTICLE

DICE-E: A FRAMEWORK FOR CONDUCTING DARKNET IDENTIFICATION, COLLECTION, EVALUATION WITH ETHICS[1]

Victor Benjamin

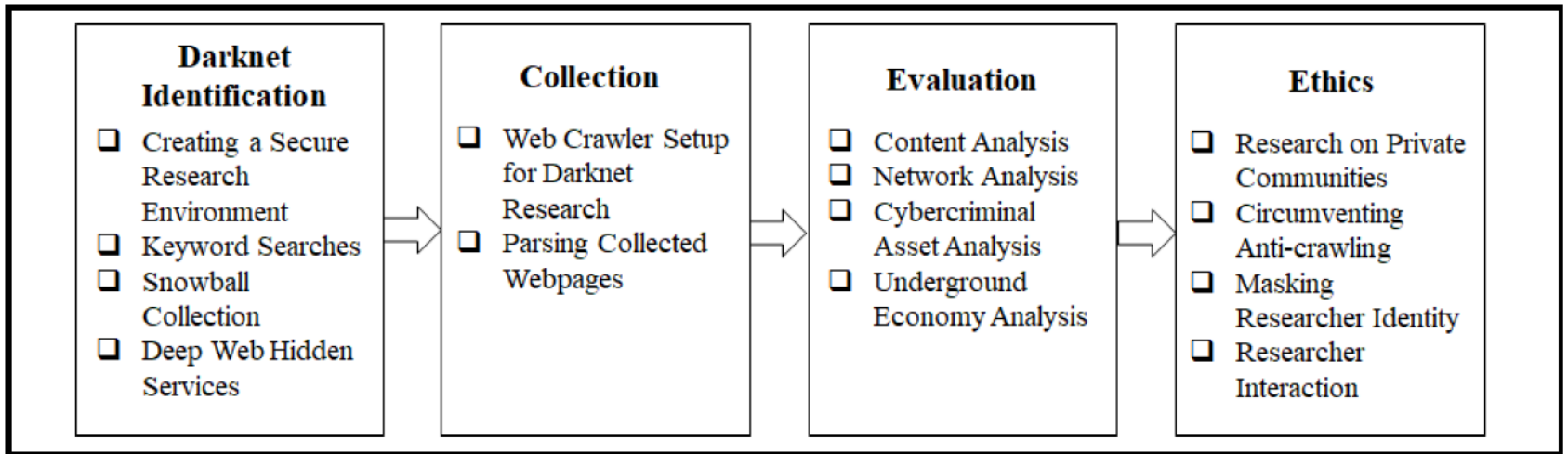


Figure 2. The DICE-E Framework

Identify Hacker Assets/Tools

Sagar Samtani (JMIS, January 2018)

Routledge
Taylor & Francis Group

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

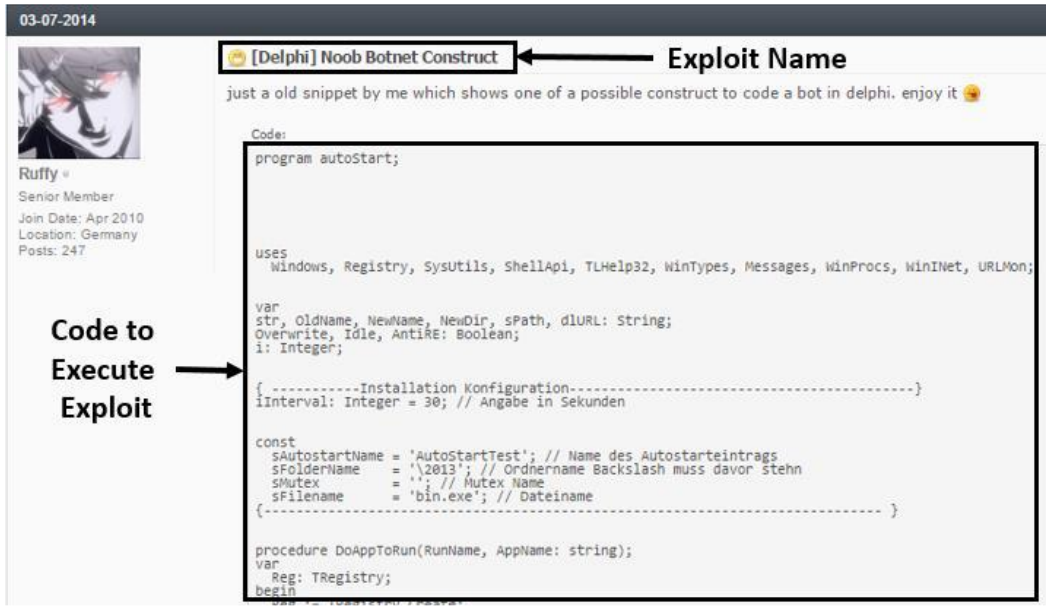


Figure 1. Forum post with source code to create botnets

Figure 2. Forum post with BlackPOS malware attachment

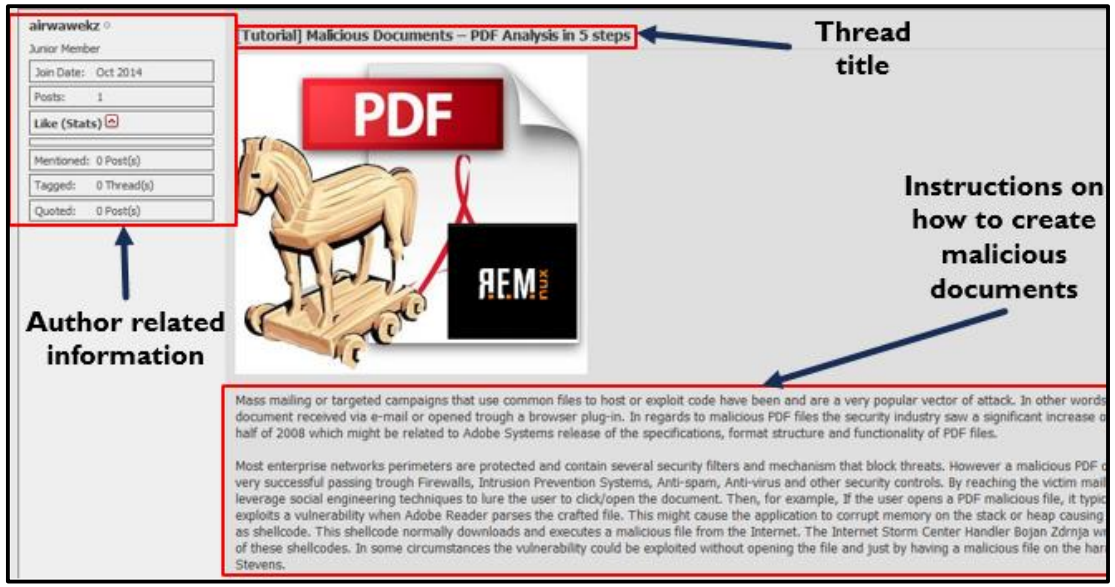
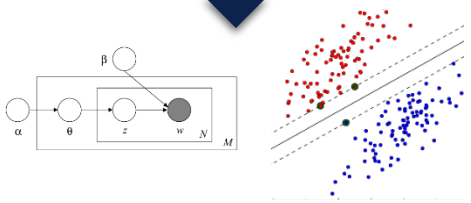


Figure 3. Tutorial on how to create malicious documents

AZSecure Hacker Assets Portal System

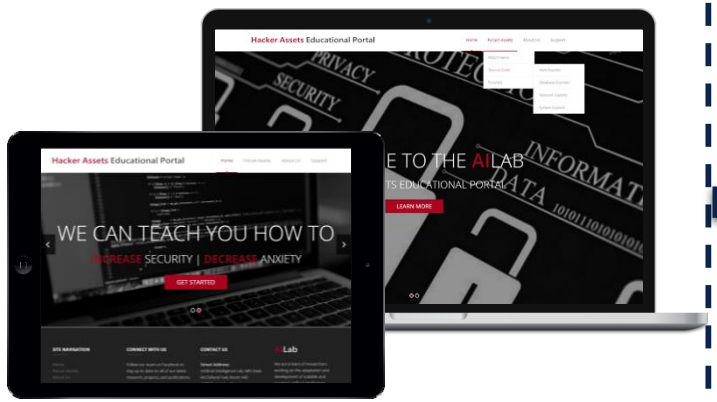
Data Collection and Analytics



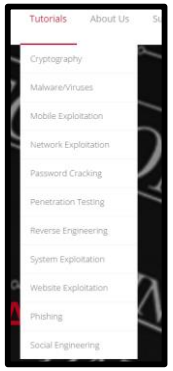
Latent Dirichlet Allocation (LDA) and Support Vector Machine (SVM) Analytics

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

Web Hosting and Access



System Functionalities

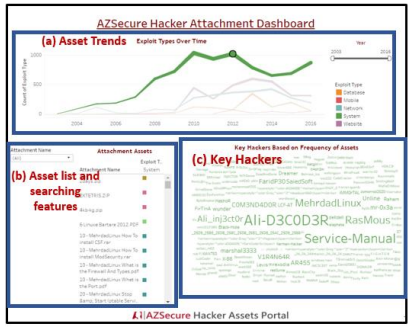


Browsing

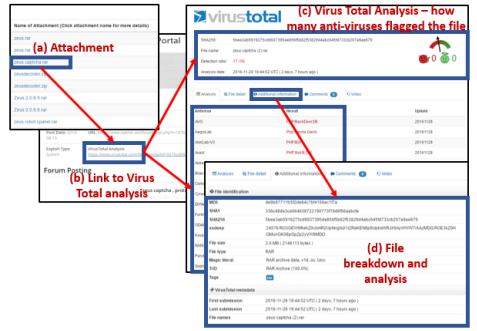
Searching

Downloading

System Analytics



Cyber Threat Intelligence Dashboard

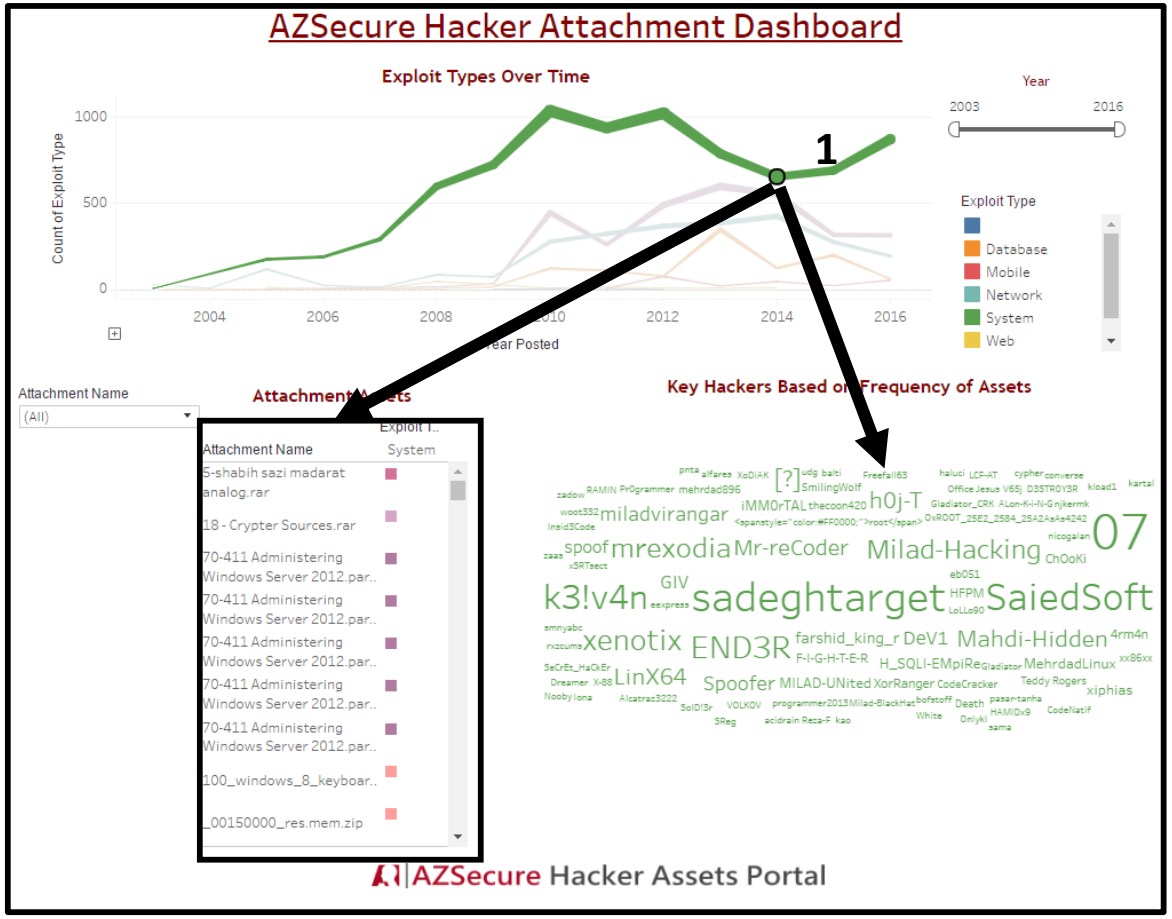


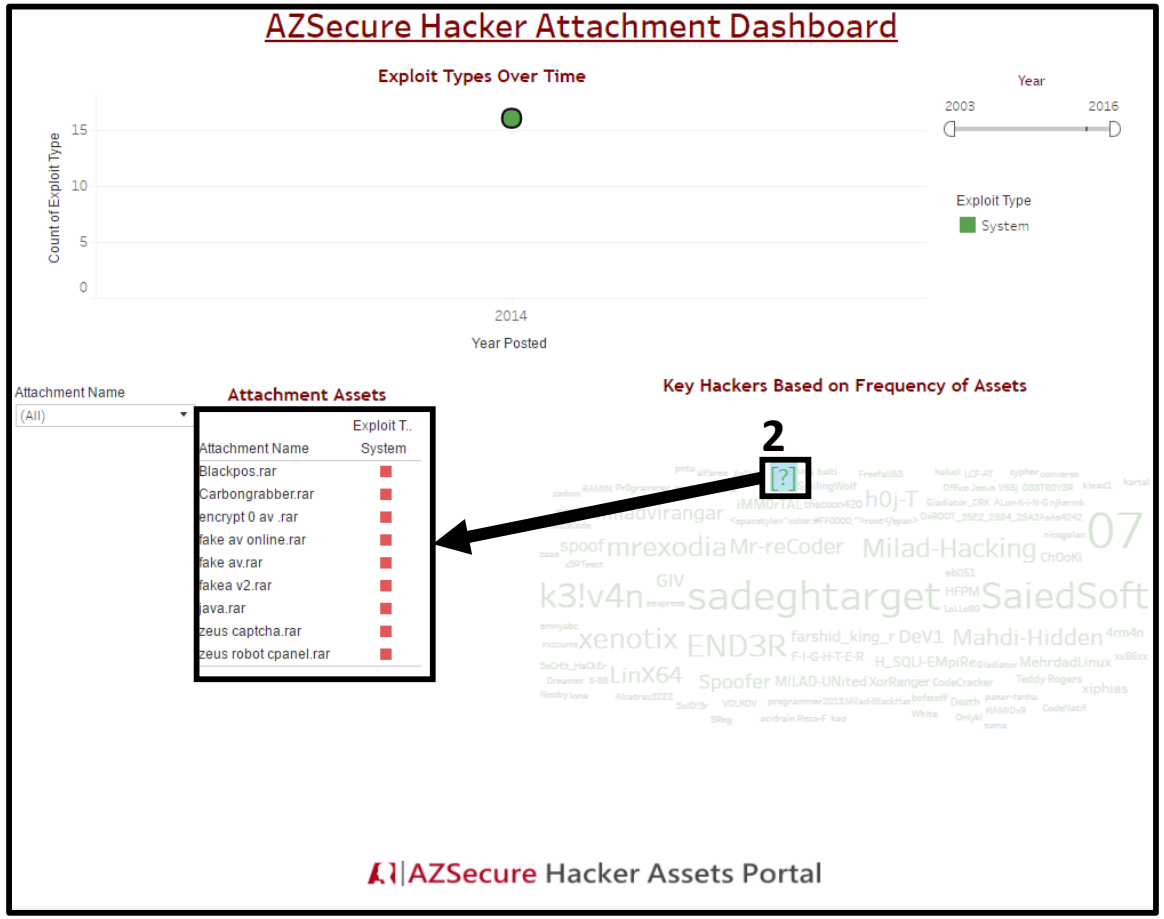
VirusTotal Malware Analysis

AZSecure Hacker Assets Portal (English, Russian, Arabic)

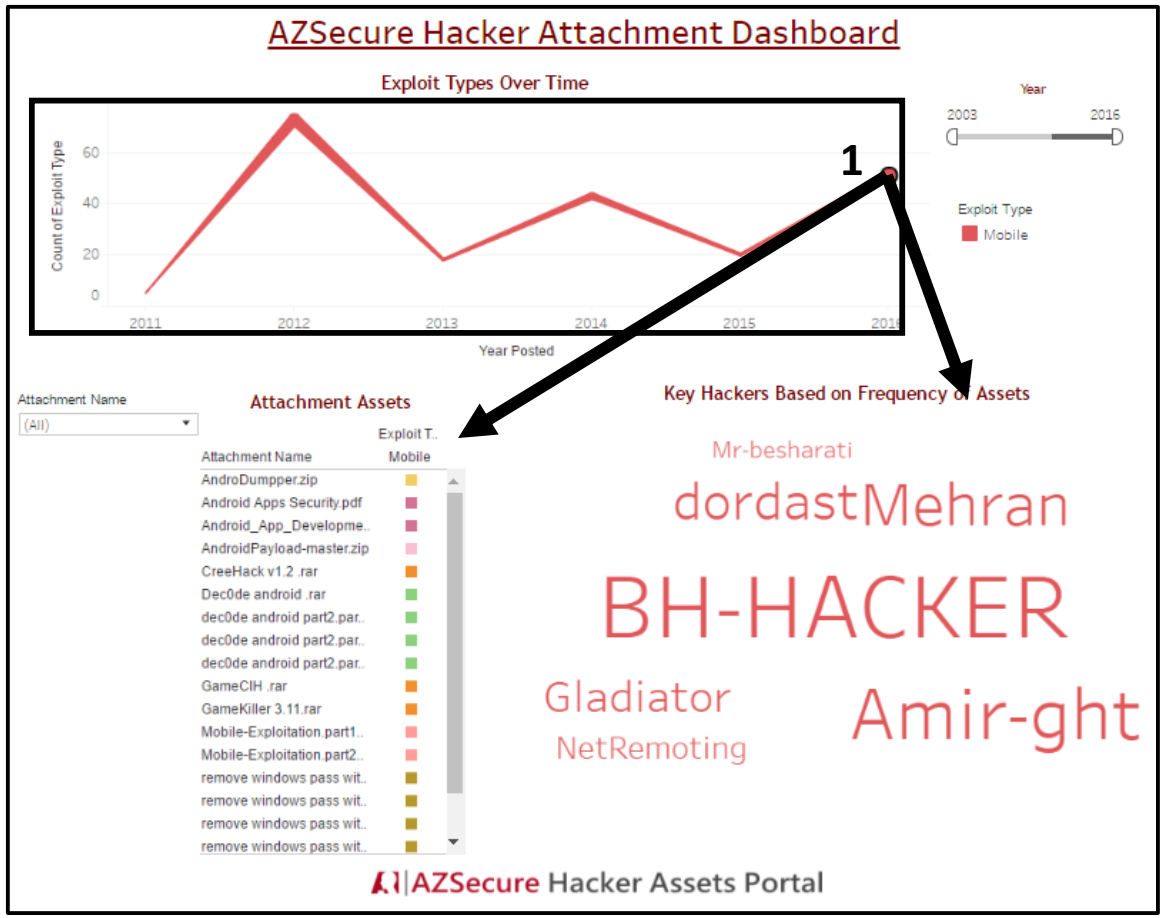
Forum	Language	Date Range	# of Posts	# of Members	# of source code	# of attachments	# of tutorials
OpenSC	English	02/07/2005-02/21/2016	124,993	6,796	2,590	2,349	628
Xeksec	Russian	07/07/2007- 9/15/2015	62,316	18,462	2,456	-	40
Ashiyane	Arabic	5/30/2003 – 9/24/2016	34,247	6,406	5,958	10,086	80
tuts4you	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
Total:	**-**	**02/07/2005- 10/31/2016**	**590,699**	**47,492**	**15,576**	**14,851**	**987**

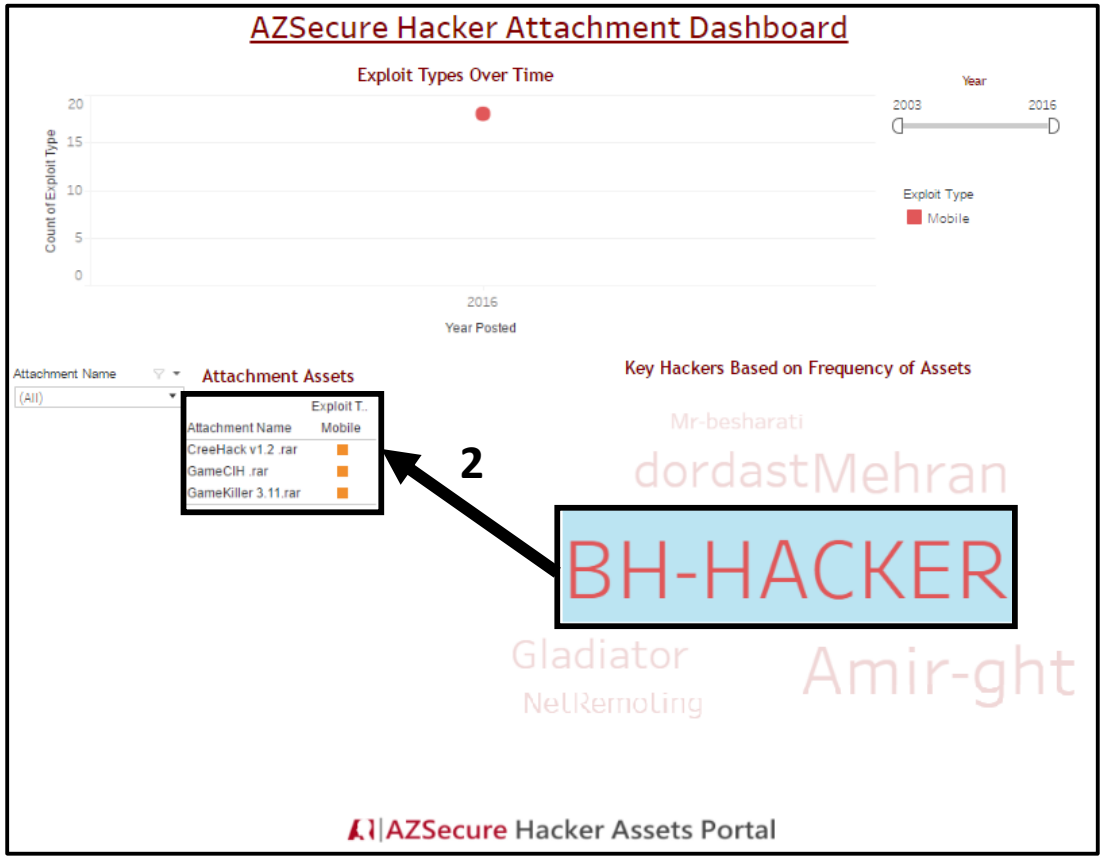
Cyber Threat Intelligence (CTI) Example – Bank Exploits (e.g., BlackPOS)





Cyber Threat Intelligence (CTI) Example – Mobile Malware





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

Benjamin Ampel (MISQ, 2nd Round)

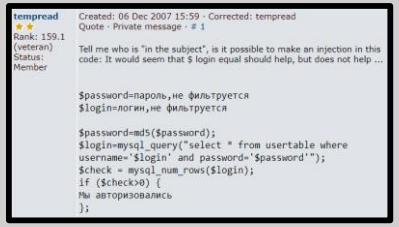
Literature Review: Hacker Forum Exploit Analysis

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.	Nulled.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.	Nulled.IO leak	Post content	SVM, LDA	Backdoor, botnet, crypter, DDoS, exploit, malware, password, rootkit	Exploit categorization

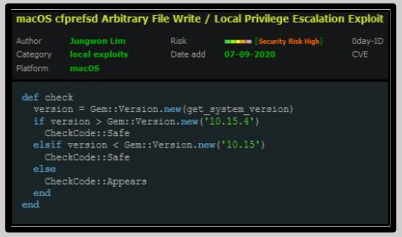
- **Key Observations:**
  1. Studies focus on general forums, but not exploit DNMs or public repositories.
  2. Although source code contains valuable information, many studies omit them from analysis.
  3. The most common task is to categorize post content by exploit category.

Proposed Research Design

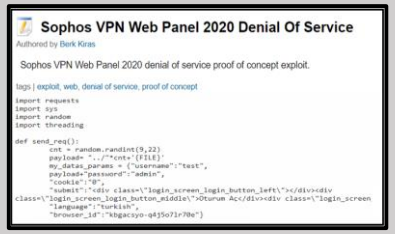
Data Collection



Traditional Hacker Forums



Exploit DNMs



Public Exploit Repositories

Pre-Processing and Dataset Construction

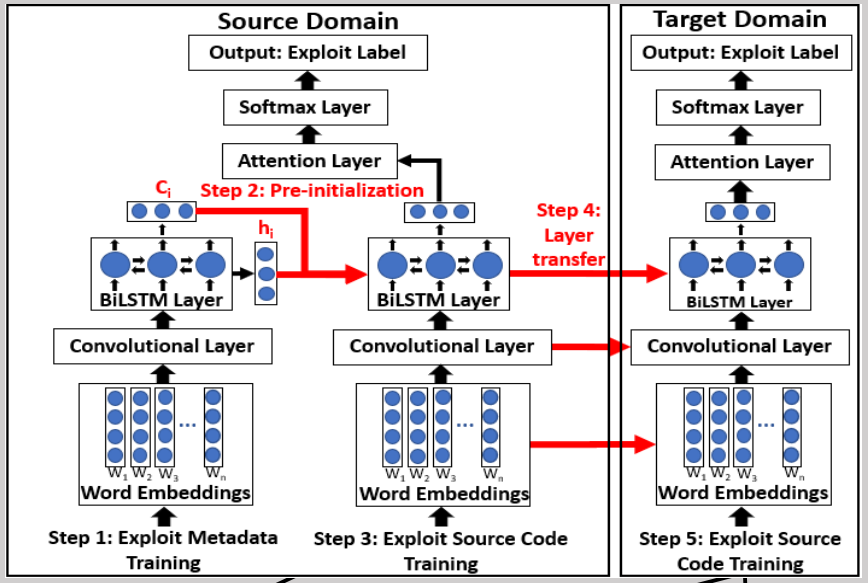
Programming Language Classifier

Remove stop words, low frequency words, and lemmatize

Tokenizer and Sequence Padder

Gold-Standard Dataset Construction

Deep Transfer Learning Exploit Labeler (DTL-EL) Model

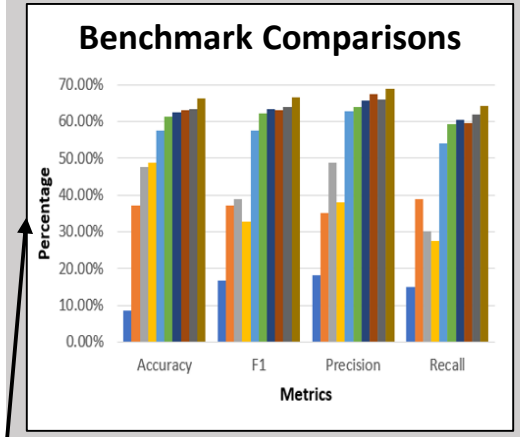


Experiment 1: DTL-EL against prevailing classification methods on source domain

Experiment 2: DTL-EL against non-transfer learning approaches on target domain

Experiment 3: DTL-EL against transfer learning layer selection on target domain

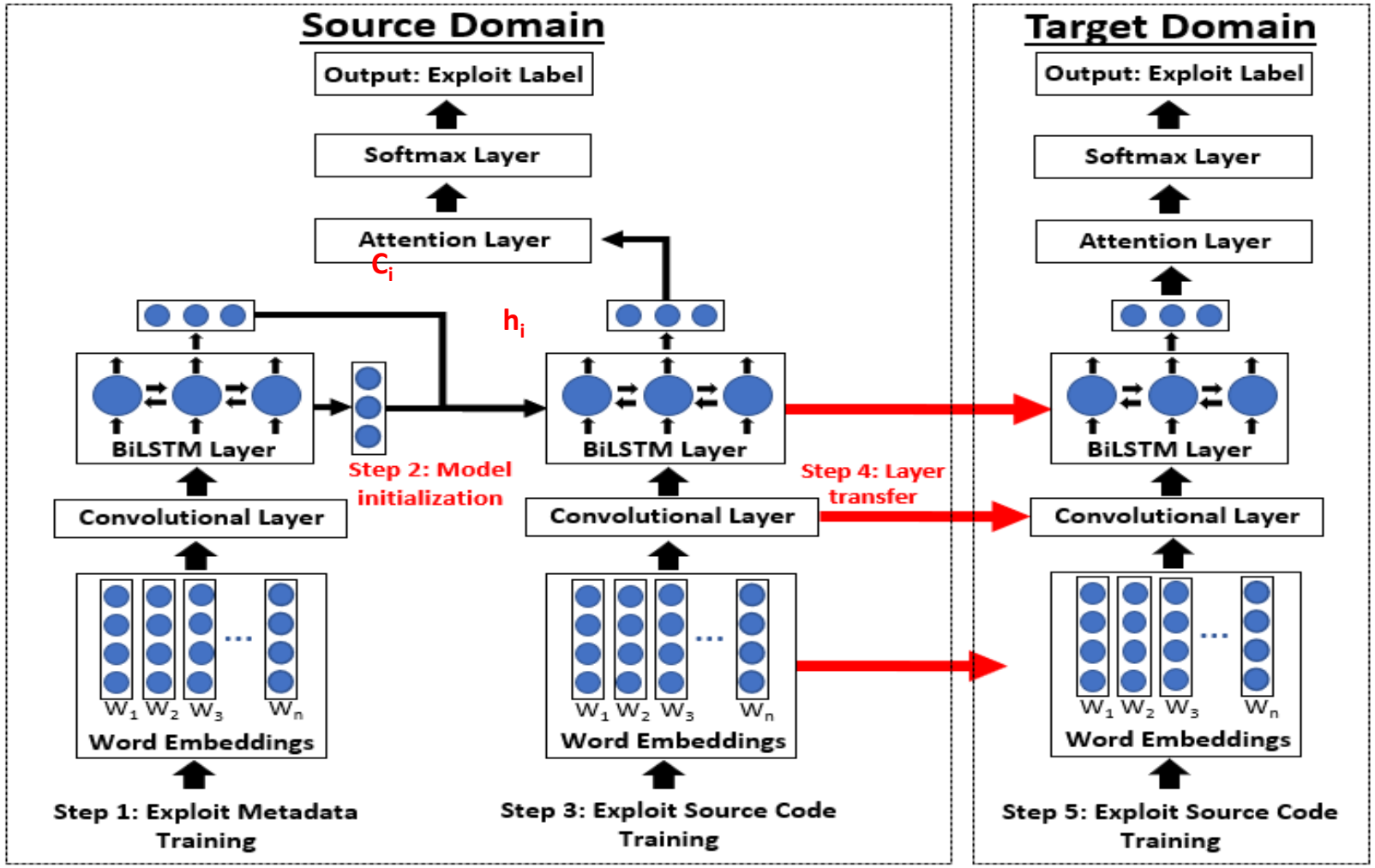
Evaluations and Visualization



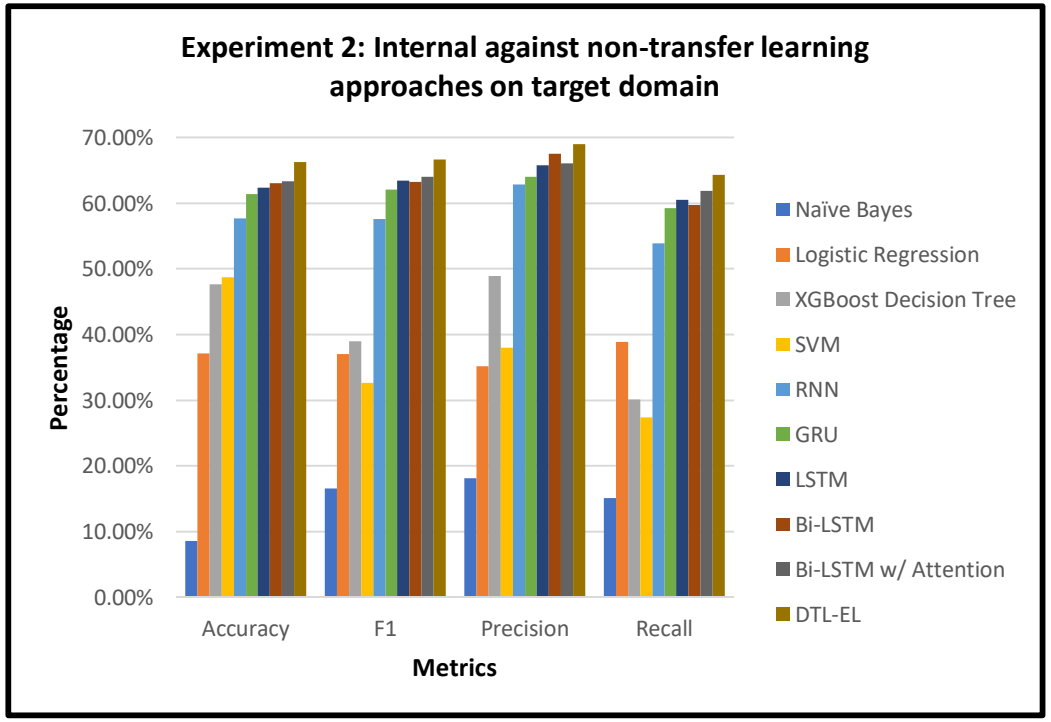
Visualized Attention Comparison

def send req cnt randomrandint payload file my datas params usernametest passwordadmin cookie submitdiv classlogin screen login button left divdiv right div browser change values headers host addr sysargv hoststr useragent mozilla windows nt win x rv gecko firefox accept text javascript html application xml acceptlanguage trtr tr q enus en acceptencoding gzip deflate xrequestedwith xmlhttprequest xprototypeversion rc contenttype json charsetutf connection close cookiecookie if want edit add added requestssession simple
socket close function sendpacketii packet global proxy host port html regex ockfsockopen gethostbyname ock echo response htmlentities die else c preg match valid partsexplode connecting parts proxybr fputs feof htmlfgets eregi chr xd xa htmlfread fclose post path num portintval trim numintval strlen error check ppath phttp hostportpath hoststr replace rn pathstr try see magic quotes gpc enable packetget pmobile dwnldphppg e testinc http packetuseragent shareaza vxxxxrn packethost hostrn packetconnection closernrn show testhttp

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% ***
DTL-EL (Our model)	Transferred	**66.17%**	**68.25%**	**64.99%**	**66.61%**

Case Study: Identifying Key Hackers - SQLi

- Since 2017, SQL injections are the most prevalent exploit in Russian forums.

- The five hackers with the most SQL injections posted on Russian forums are:
  1. **karkajoi (13 exploits)**
  2. sepo (12 exploits)
  3. BenderMR (12 exploits)
  4. Zmii666 (6 exploits)
  5. fandor9 (6 exploits)

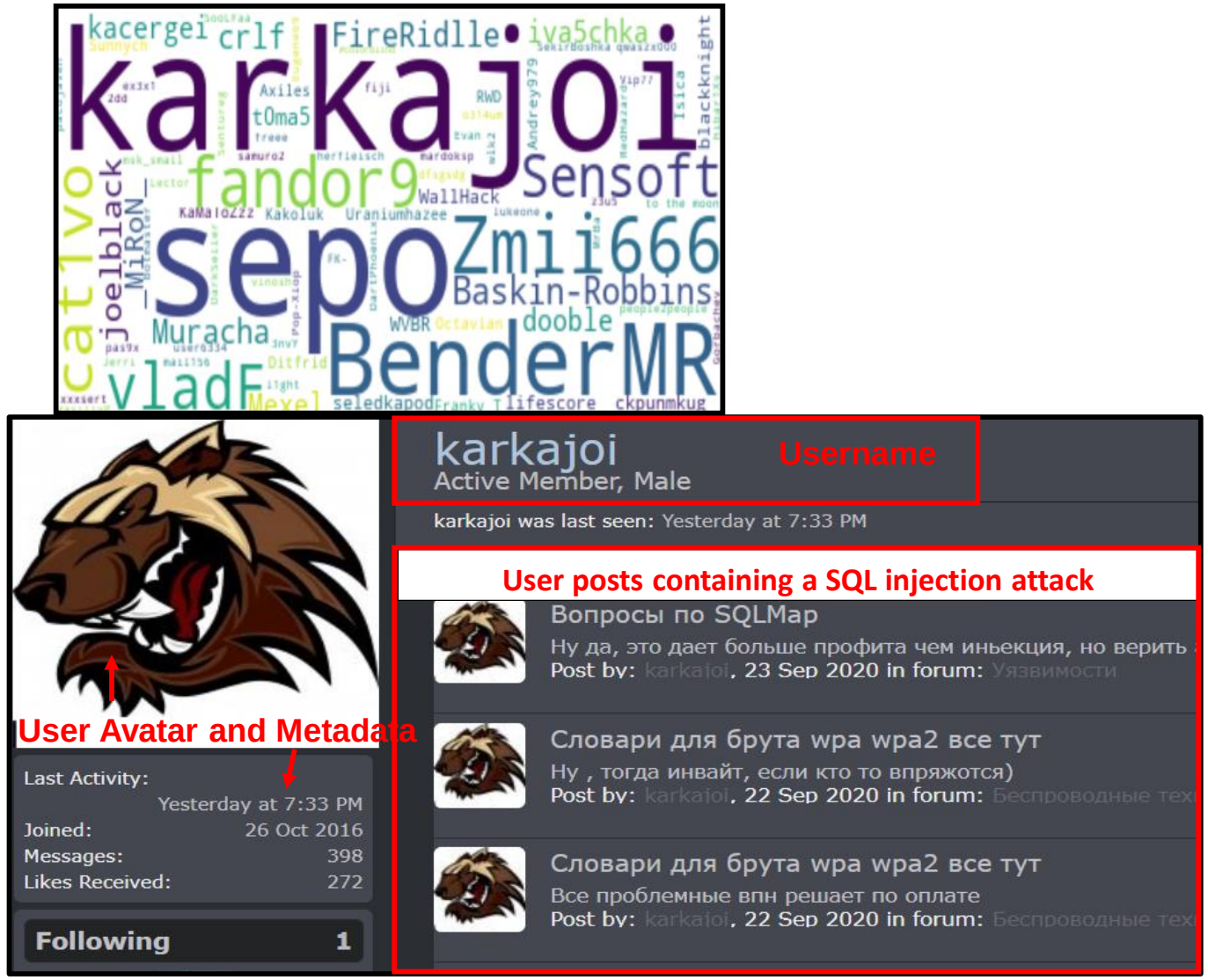


Figure 15. Hacker Profile Page on Antichat

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.

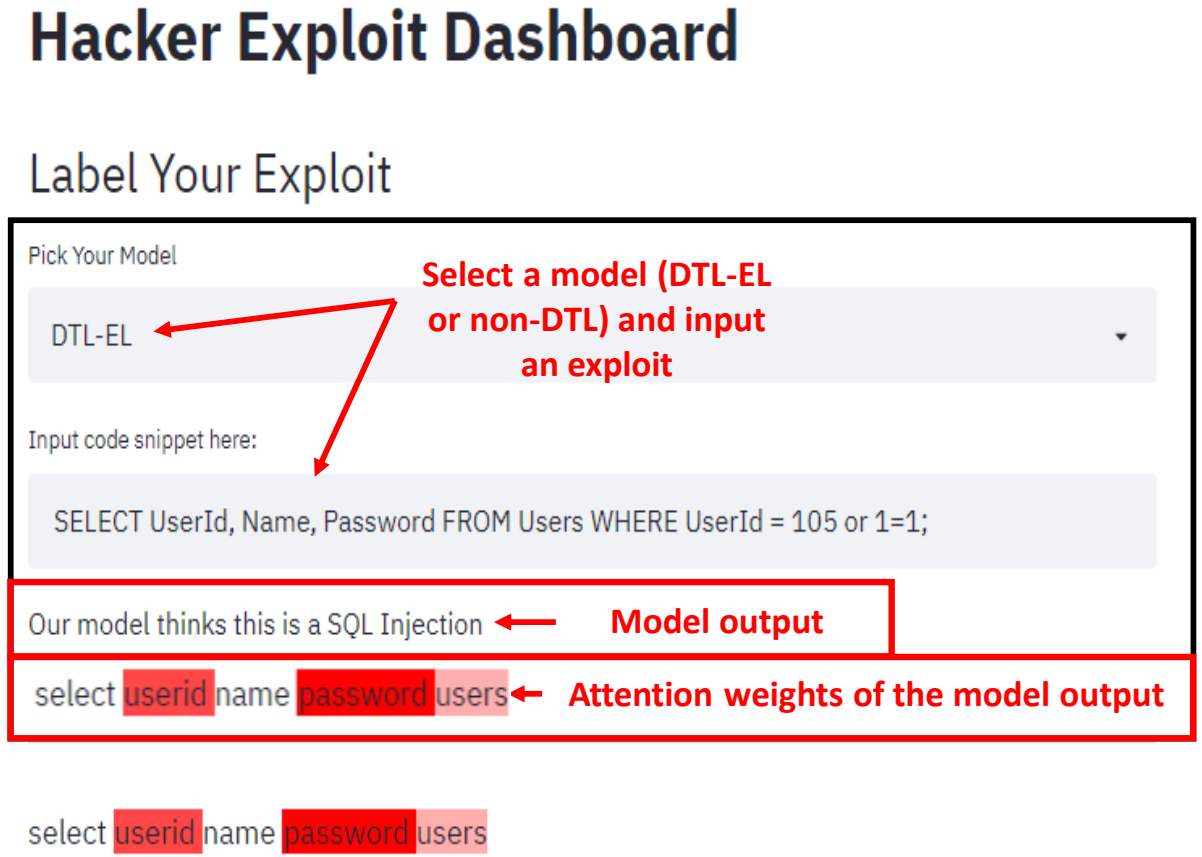
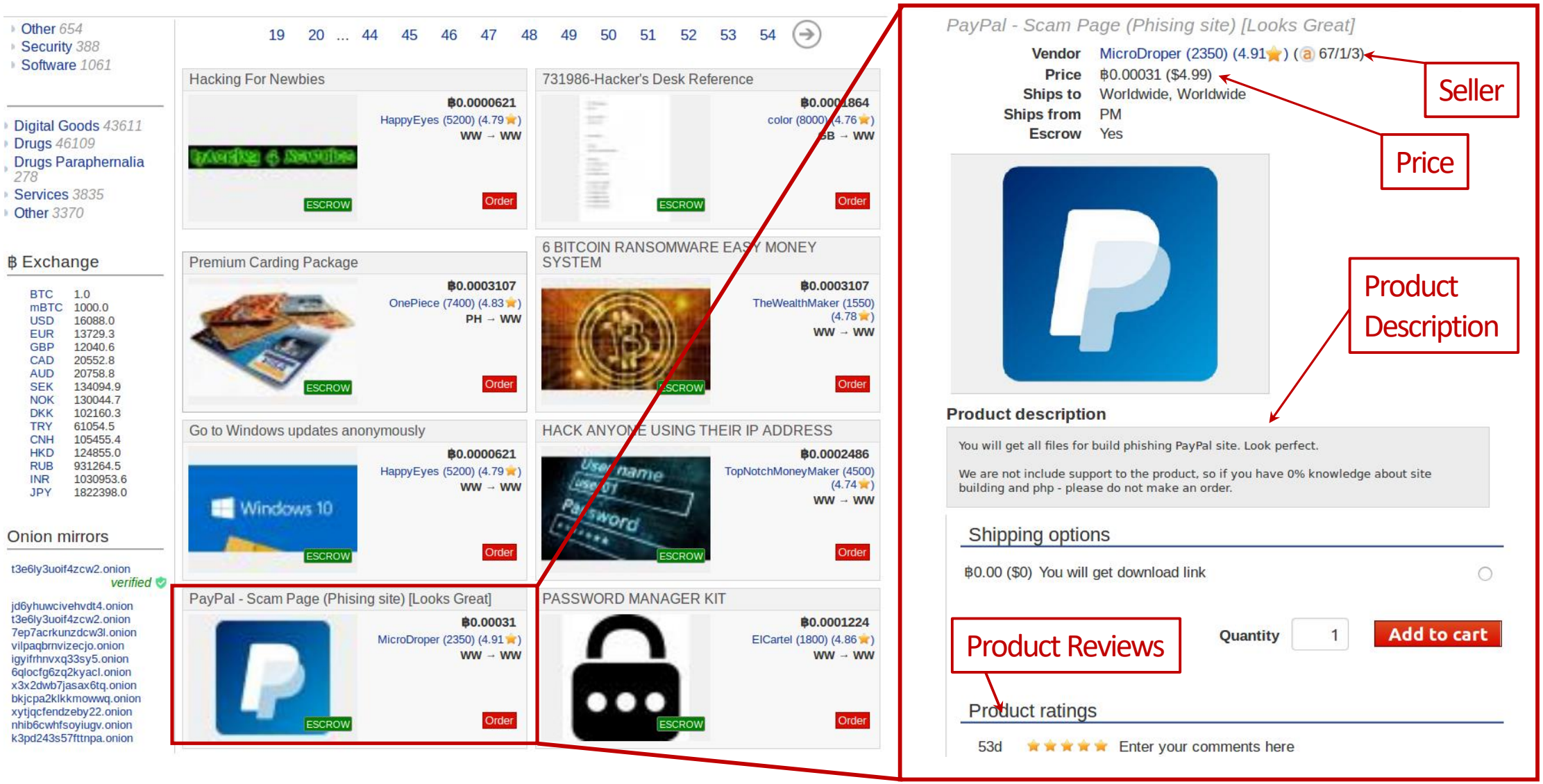


Figure 16. Hacker Exploit Portal For Further Analysis

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

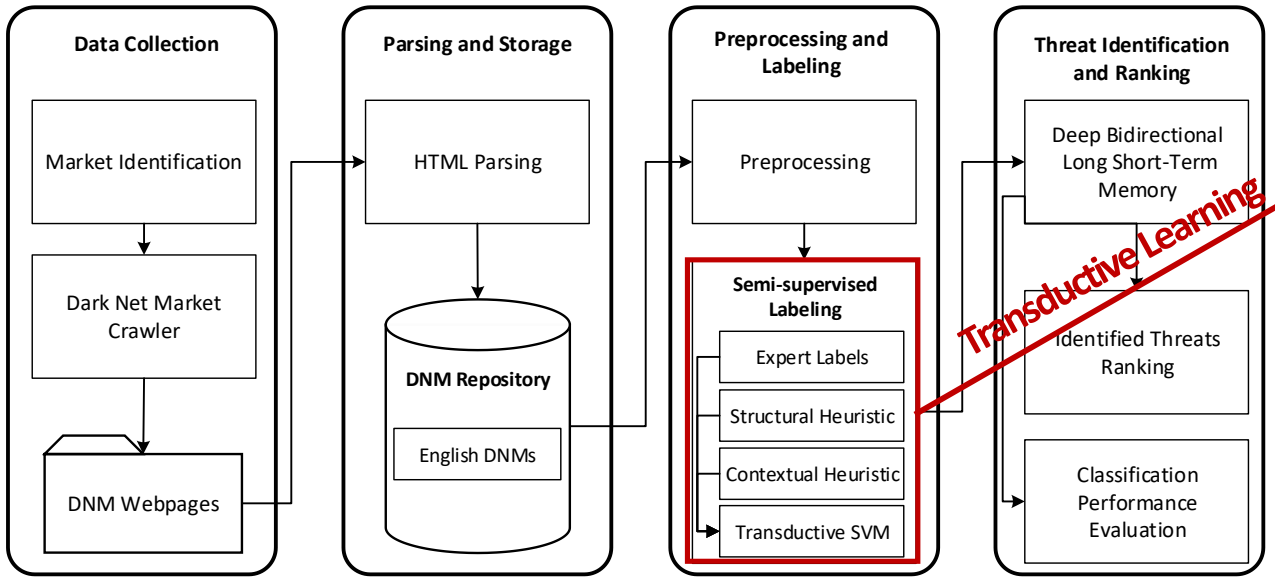
Reza Ebrahimi (JMIS, MIS, IEEE PAMI)

Dark Net Marketplaces (DNMs)



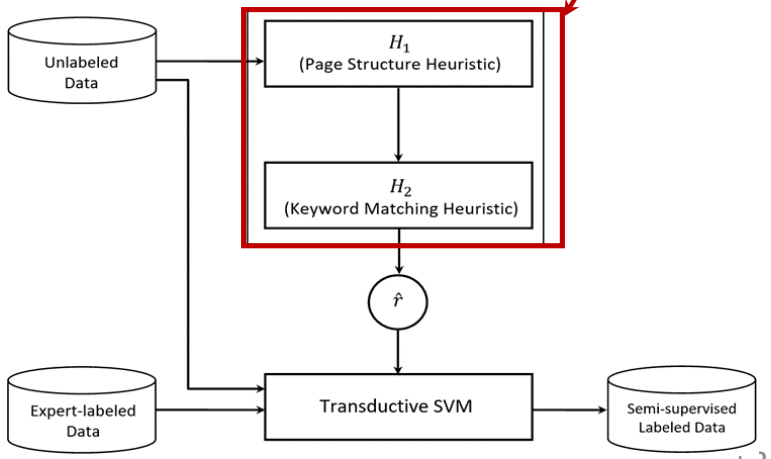
Essay I: Learning From Unlabeled Cybersecurity Content (JMIS, March 2020)

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



$$\min_{w}\left(\frac{\lambda}{2}\|w\|^2+\frac{1}{2L}\sum_{i=1}^{L}l(y_i,w^Tx_i)+\frac{\lambda'}{2U}\sum_{j=1}^{U}l(y_i',w^Tx_j')\right)$$

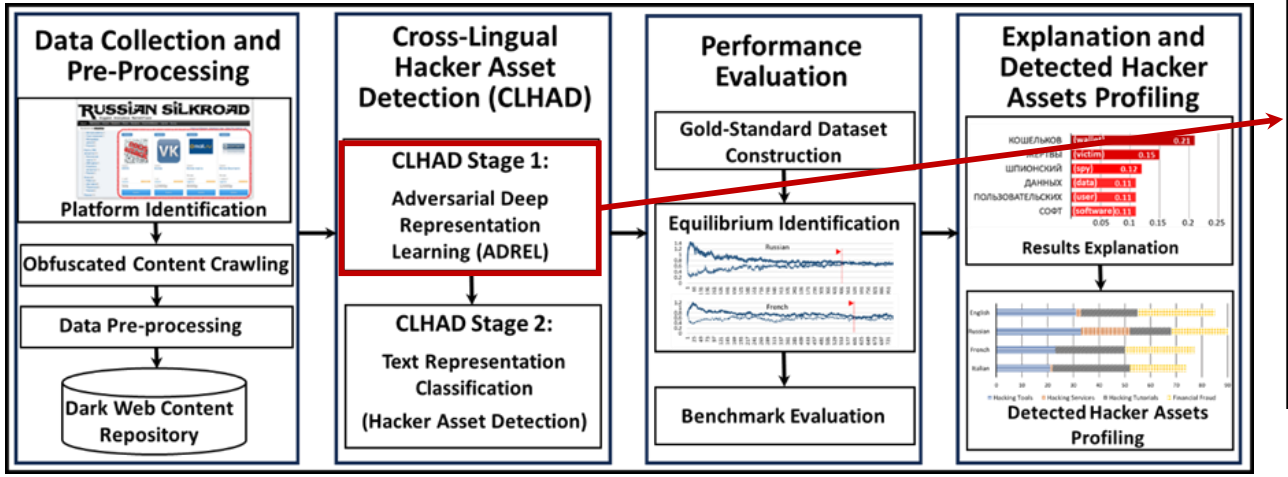
$$\text{subject to: } \frac{1}{U}\sum_{j=1}^{u}\max(0,\text{sign}(w^Tx_j'))=r$$

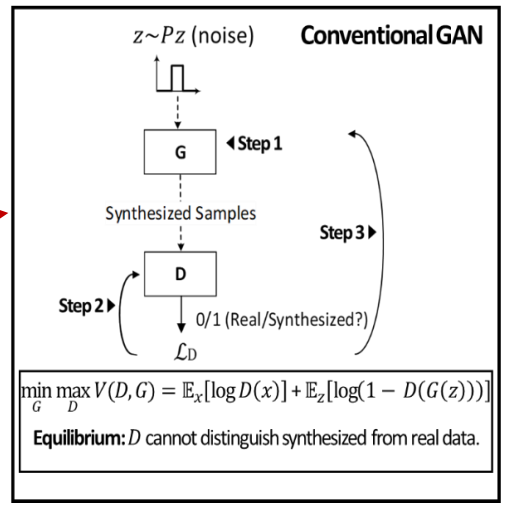


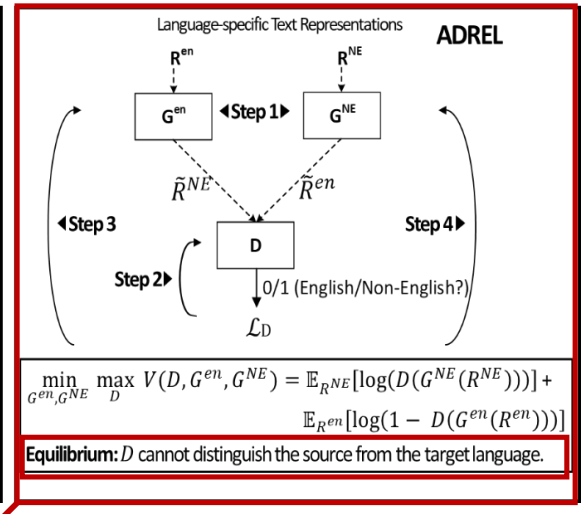
- Significantly decreased reliance on human supervision for cyber threat detection.

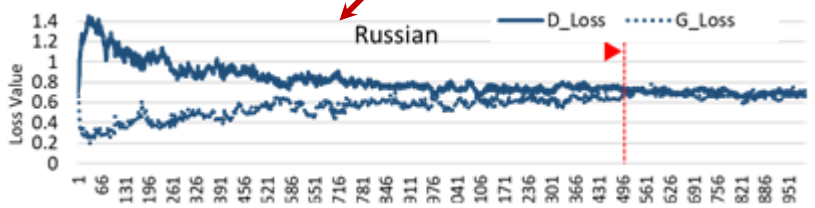
Essay II: Learning from Heterogeneous Cybersecurity Content (MISQ, Forthcoming)

- 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**





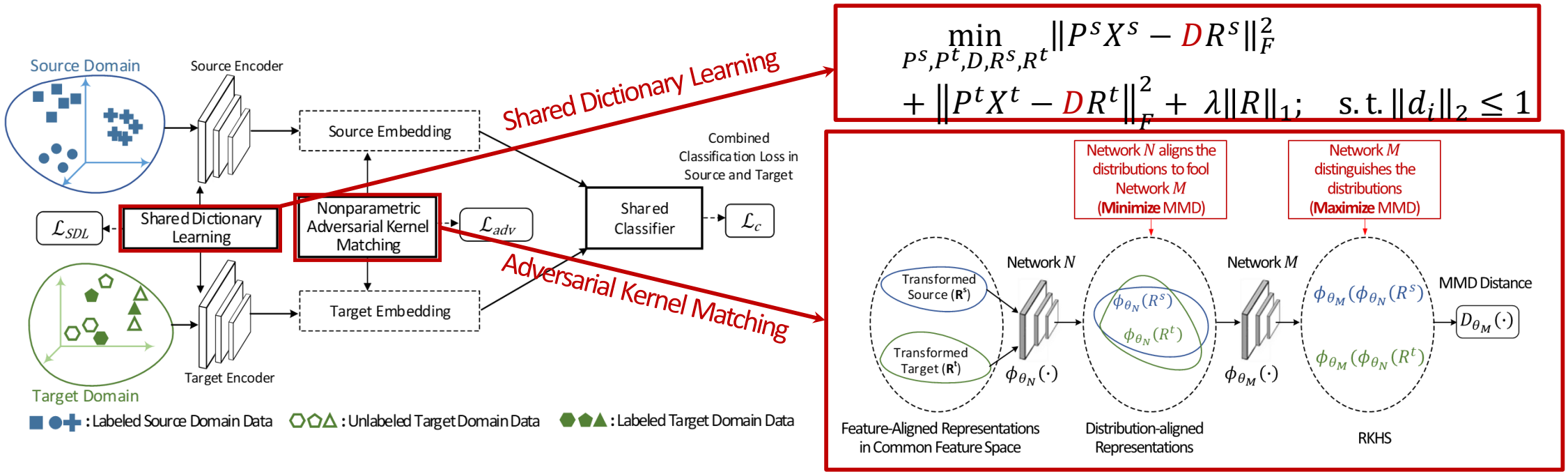




- Significantly decreased reliance on human supervision and outperformed machine translation.

Essay III: Learning from Heterogeneous Cybersecurity Content (IEEE TPAMI, 2nd 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

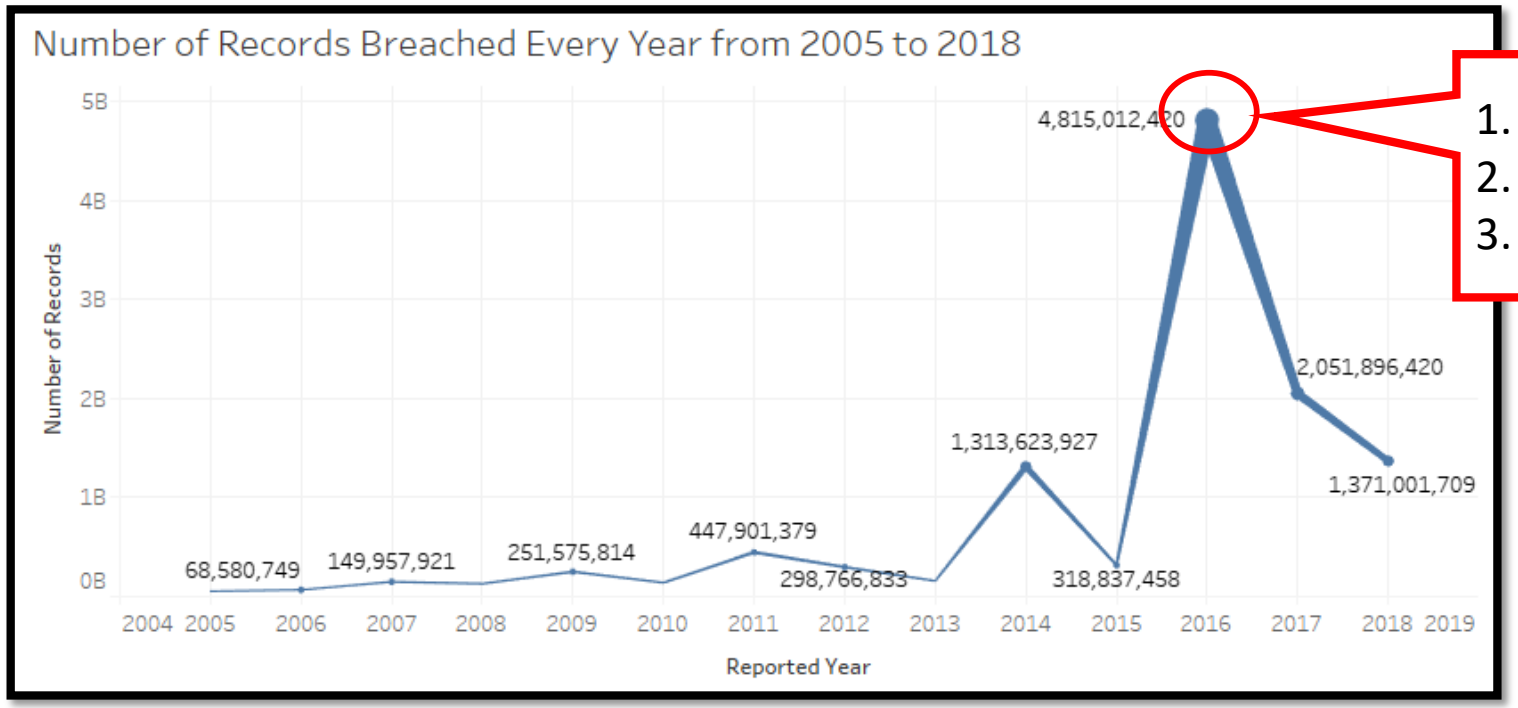
Directorate for Education and Human Resources
Division of Graduate Education

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

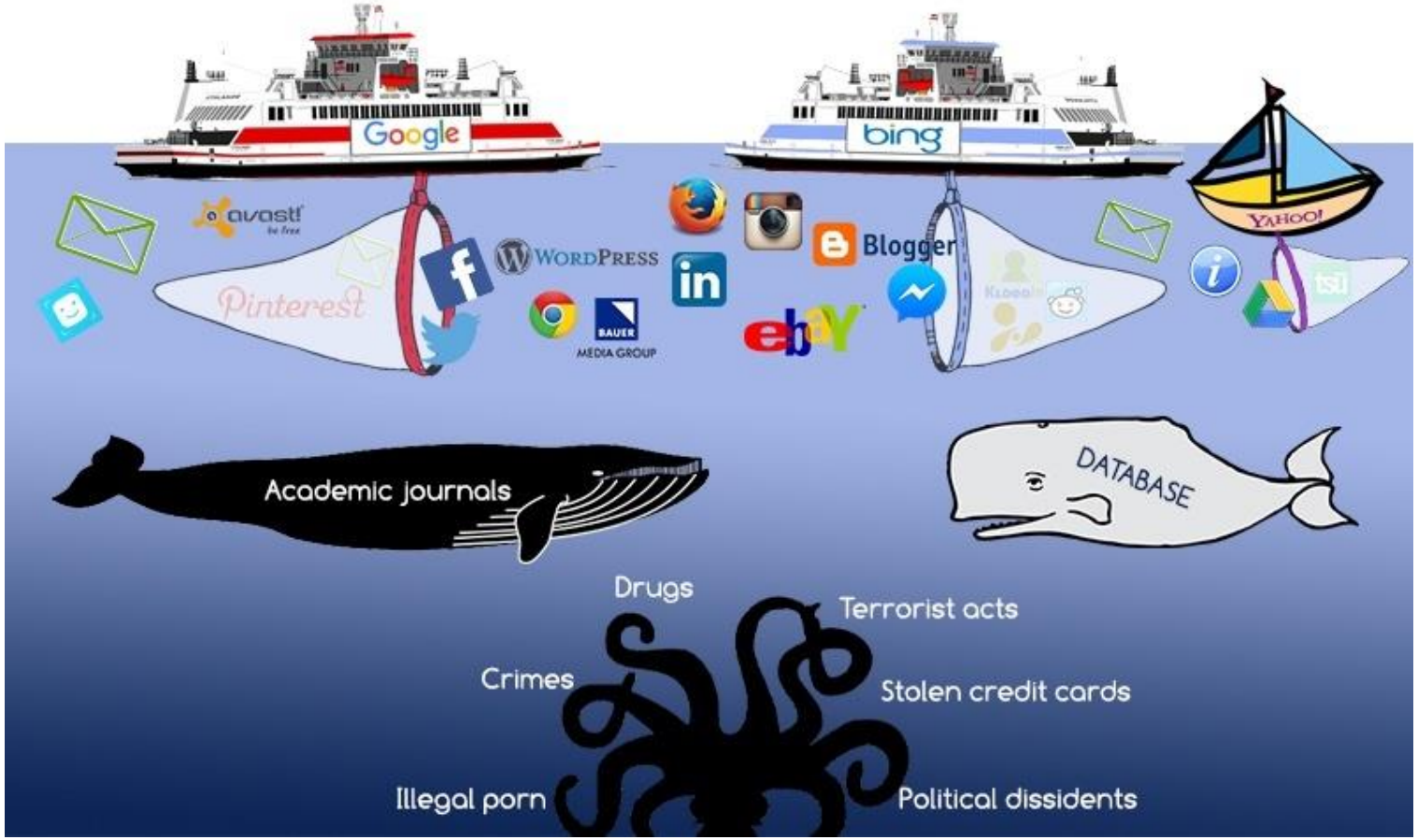
Surface Web

Deep Web

Dark Web

DarkNet

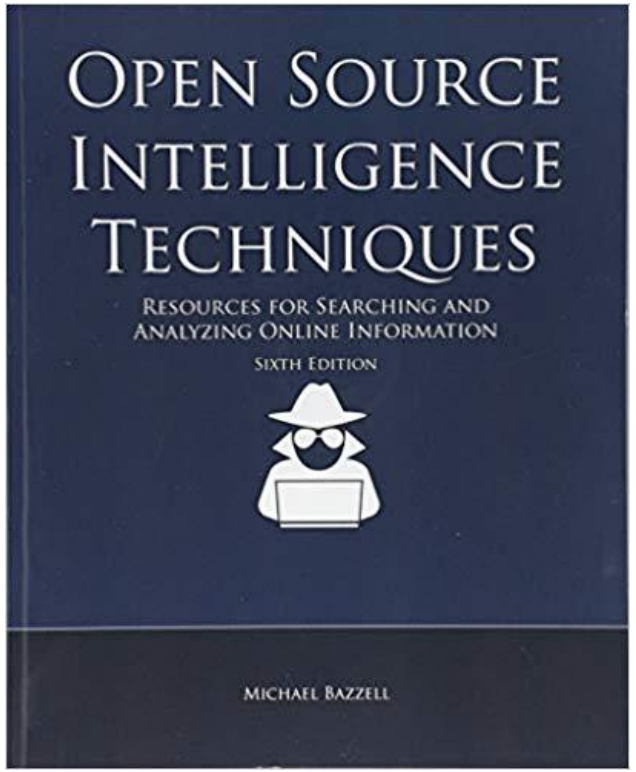
Hacker Web

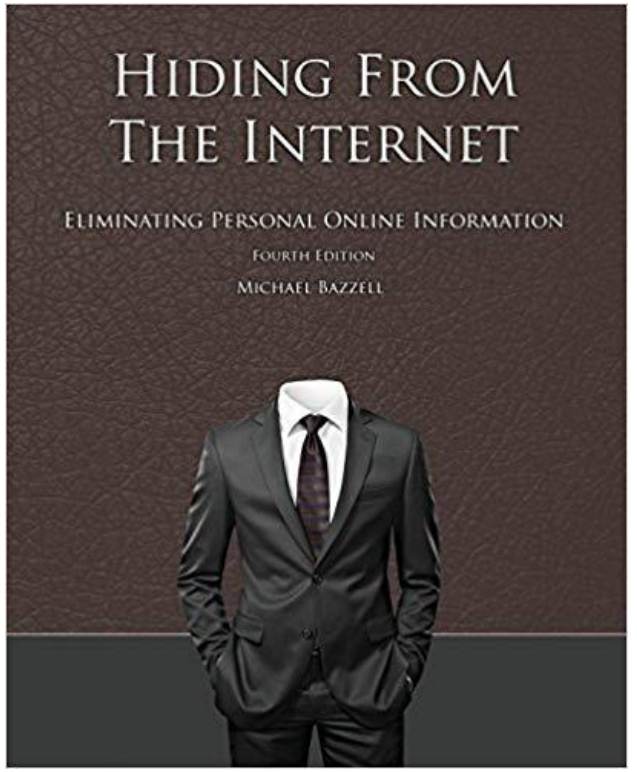


IRB, HIPAA, GDPR, PII

➔ Cybersecurity to Privacy

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





Dark Web Intelligence Sources (May, 2021)

Source	Description	Size*	Promising Attributes
Stolen Account Collection	Stolen social media and e-mail accounts	**25 billions**	**Username**
			Password
Stolen Credit Card - Tormarket	Stolen credit and debit card owner information * No card number	**832 thousands**	**Full name**
			Country
			State
			City
			Zip
Stolen SSN - Buyssn	Personal information of SSN owners *No SSN	**5.75 millions**	**Full name**
			YOB
			City
			State
			Zip
			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%
Total	-	**817,823,522**	**68.18%**

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
Total	**-**	**9,082,284**

AZSecure Privacy Portal Design

LifeLock
Relentlessly Protecting Your Identity™

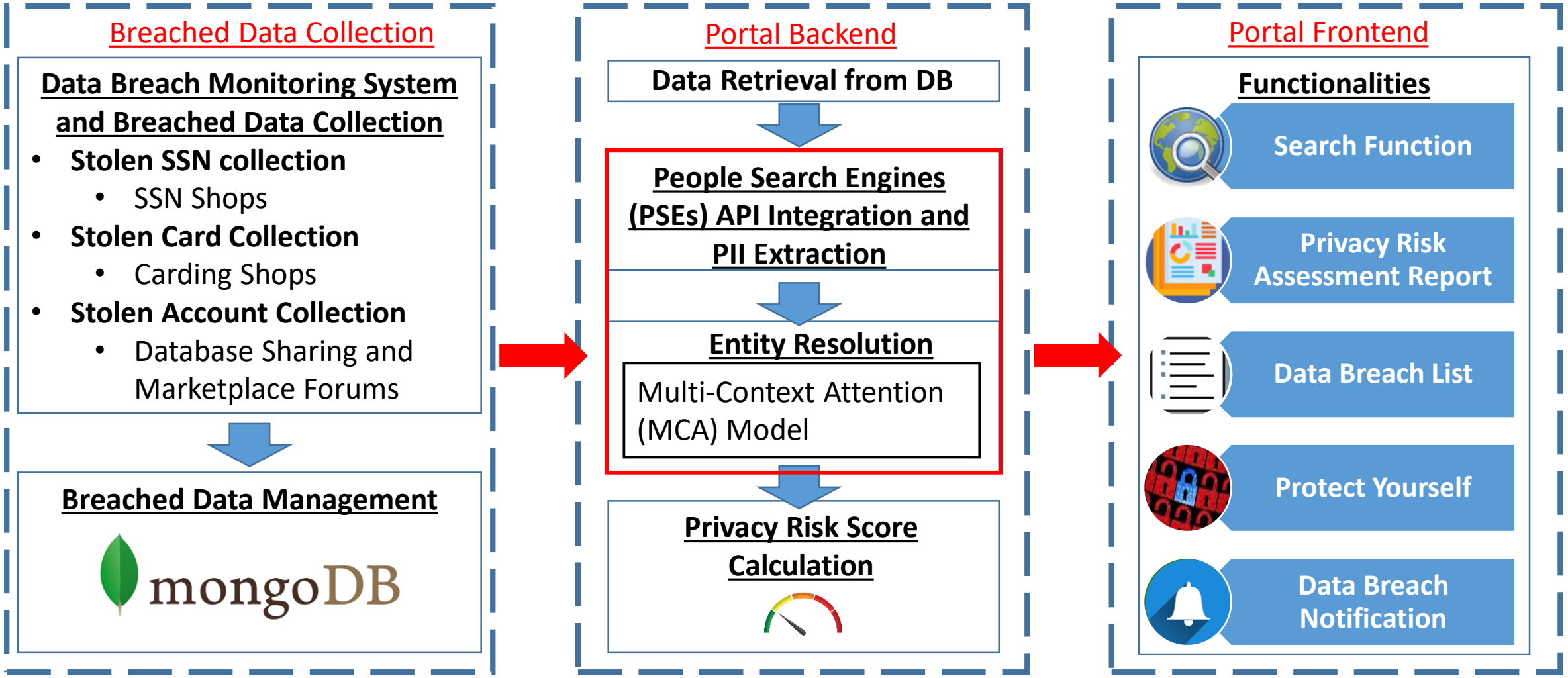


Figure 1. AZSecure Privacy Portal Project Overview

Search in AZSecure Privacy Portal

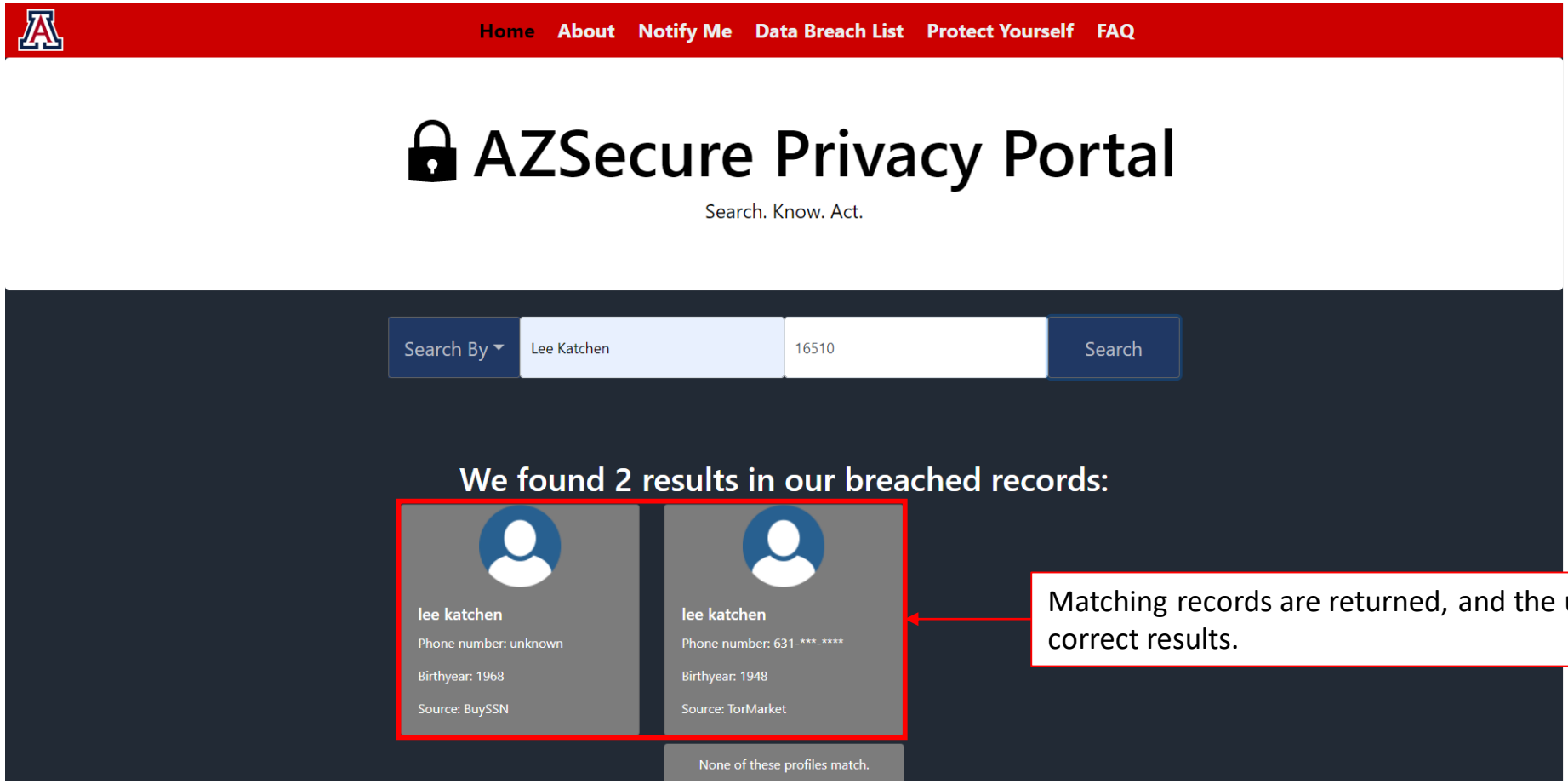
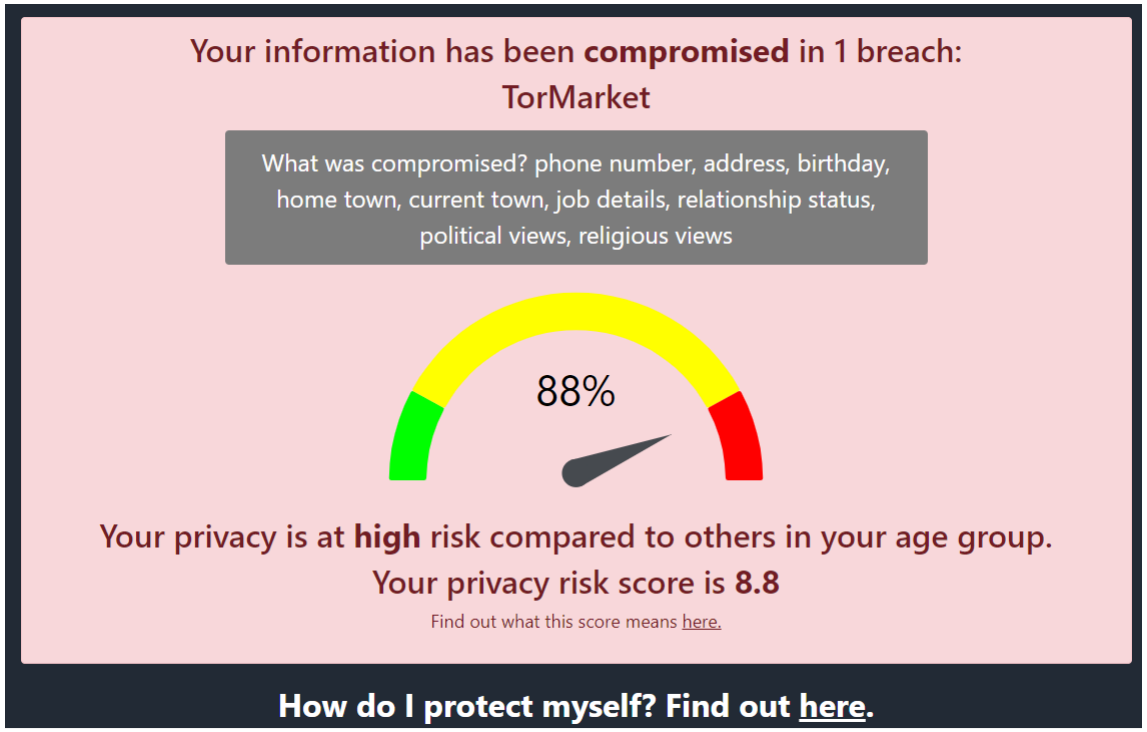


Figure 5. A mock-up response when records are found

Return Exposed PII



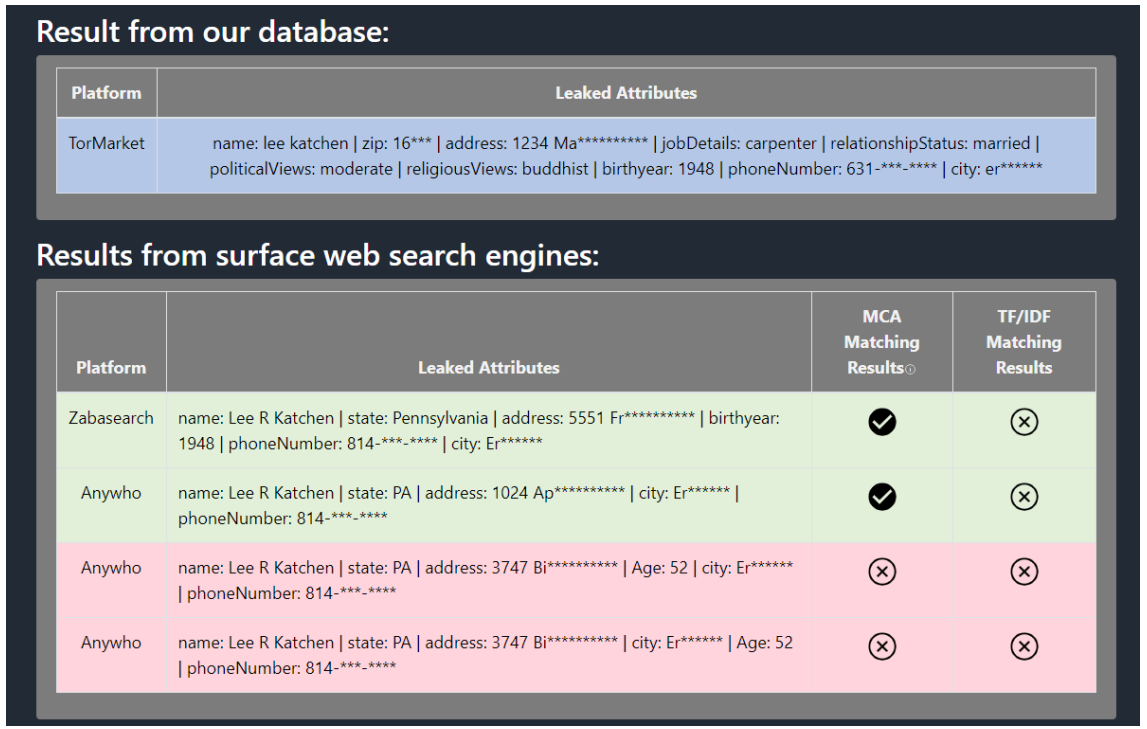


Figure 9. Mock-ups of a comprehensive exposed PII profile

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

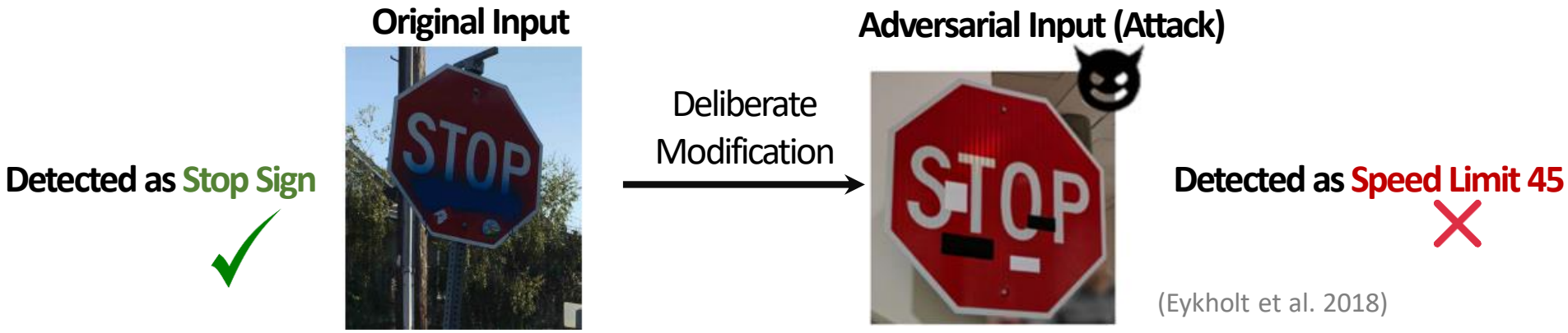
Reza Ebrahimi (JMIS, MISQ)

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

Defending Cybersecurity AI Agents

(expresscomputer.in)

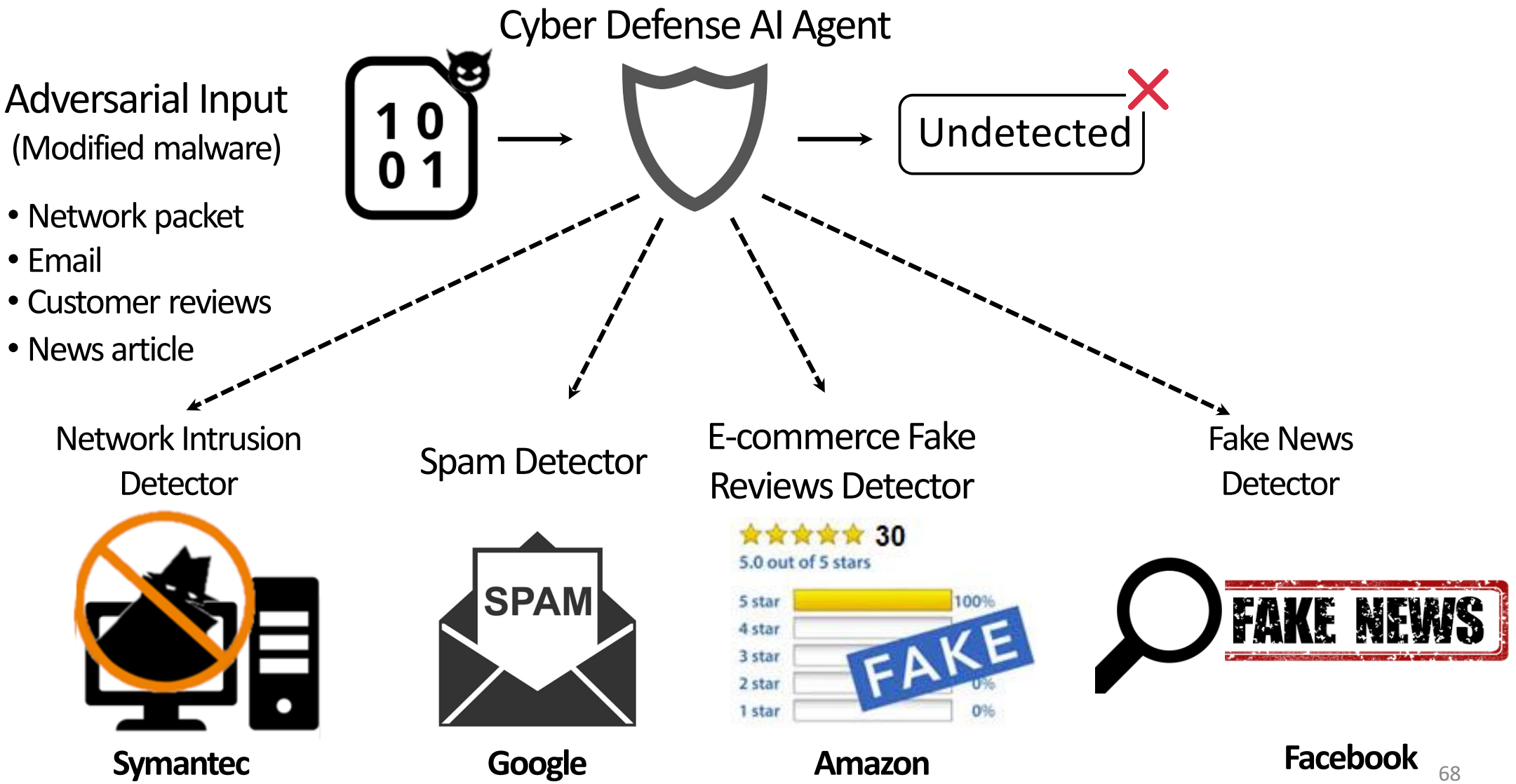
- Cybersecurity firms are adopting AI agents for autonomous cyber defense (Rai et al. 2019).
  - Automate threat detection and remediation at a large scale (Tolido et al. 2019).
- However, AI agents have shown to be vulnerable to adversarial attacks.
- Inputs meticulously modified to mislead them (Yuan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



(Eykholt et al. 2018)

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

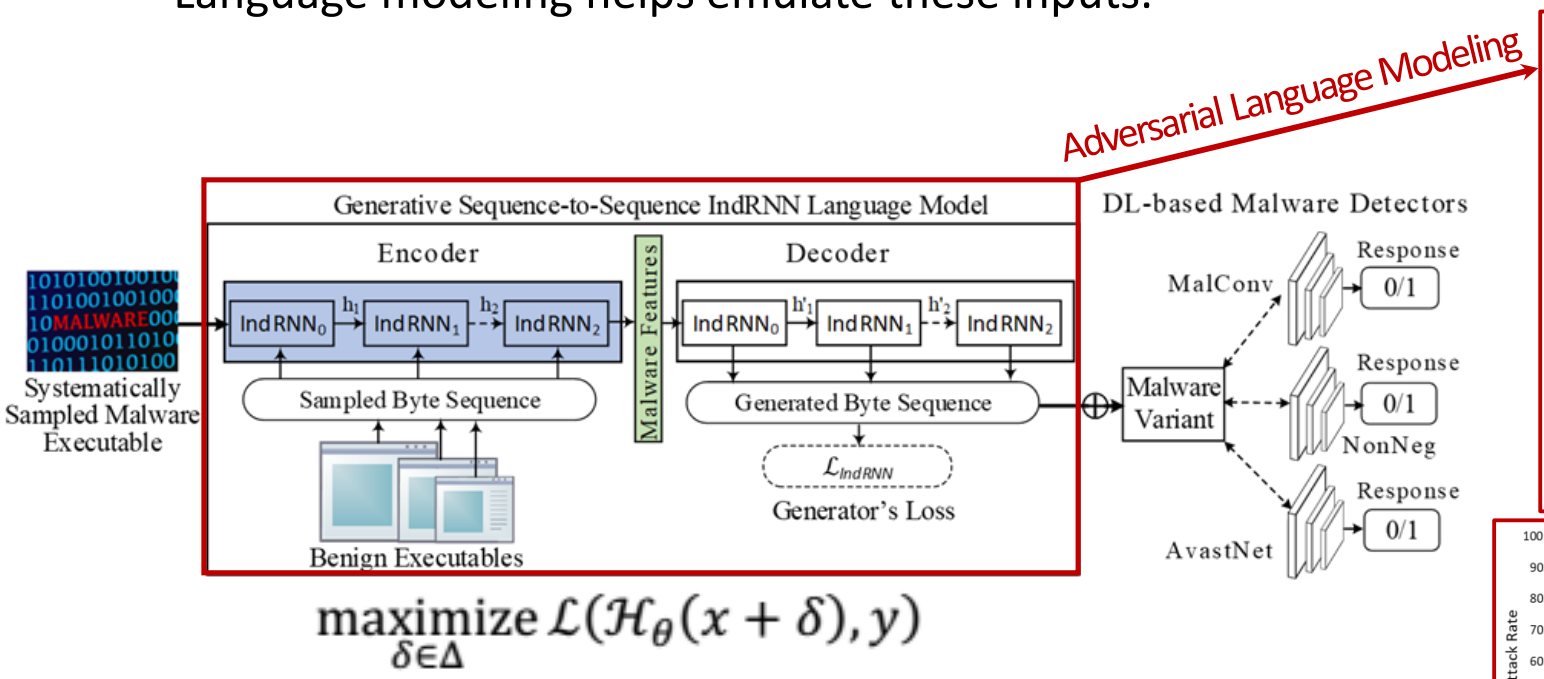
Defending Cybersecurity AI Agents

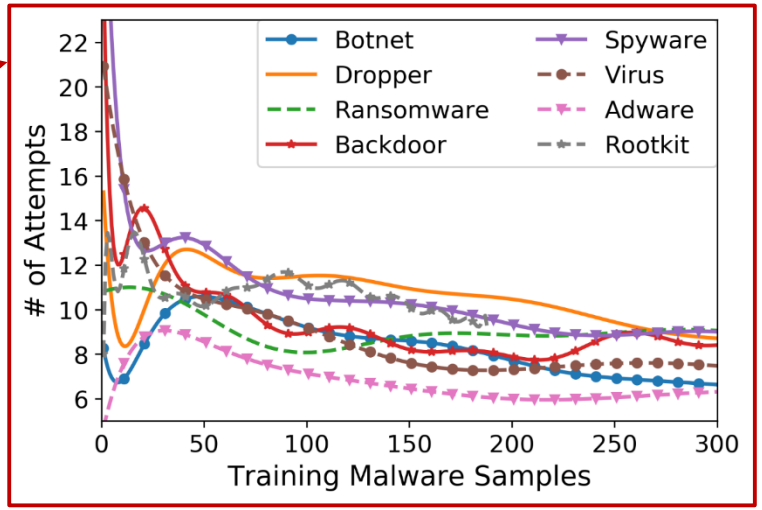


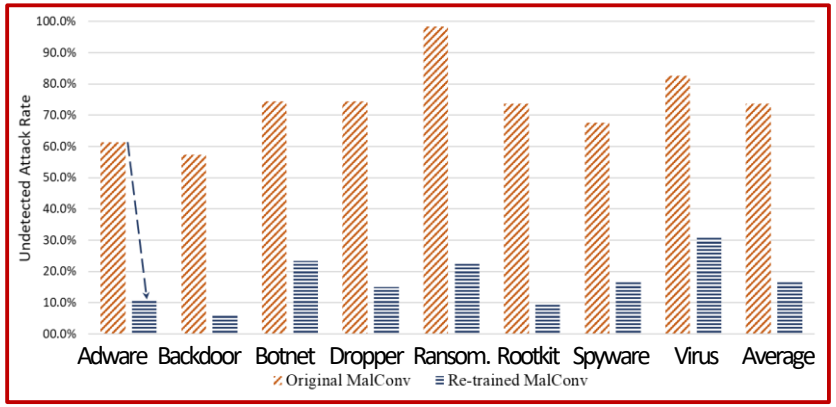
Essay I: Learning to Protect Malware Detectors

(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.



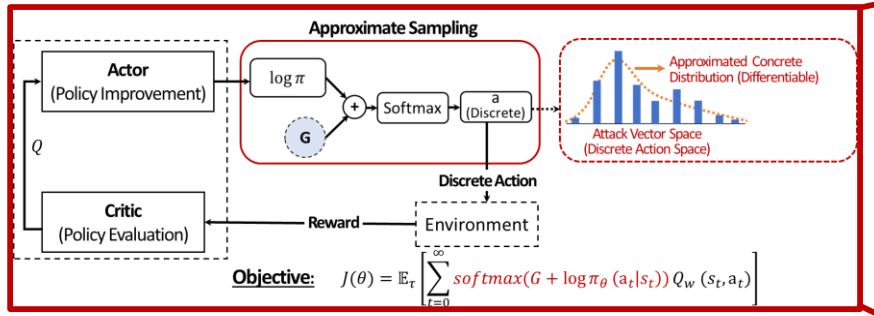




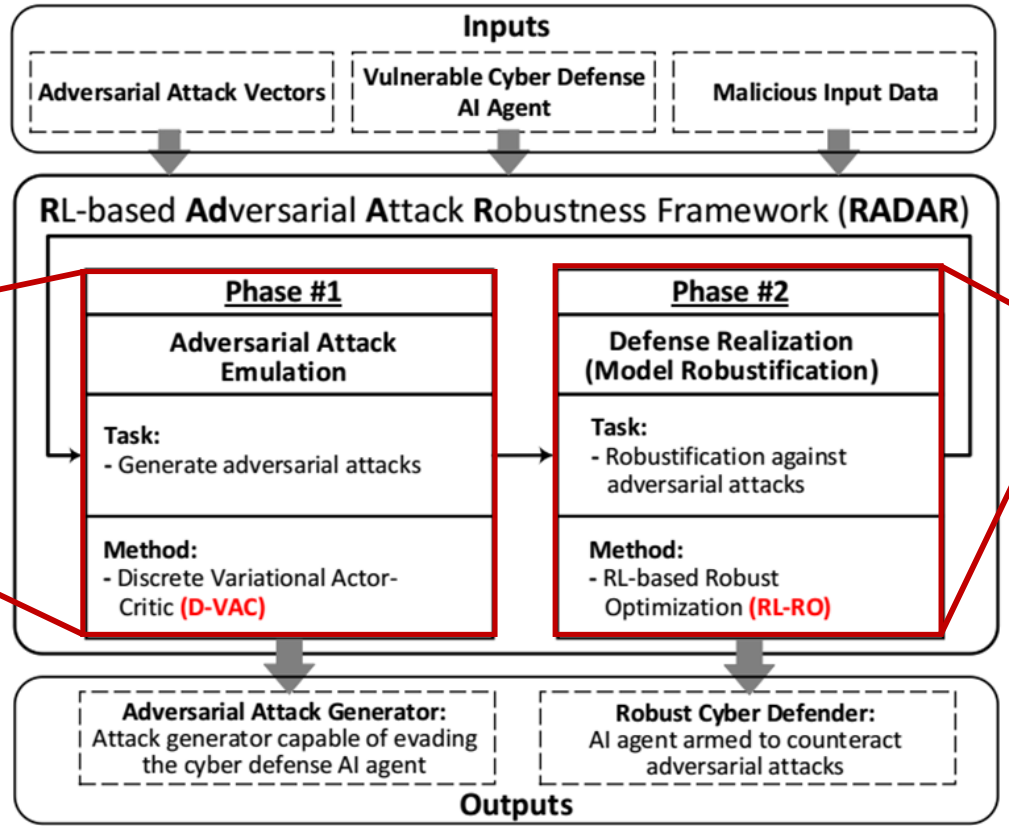
- Significantly improves the robustness of malware detectors against adversarial attacks.

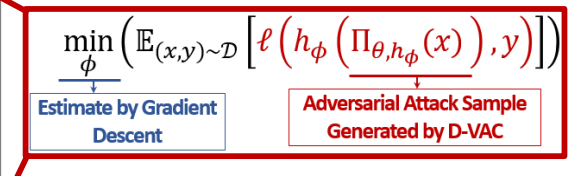
Essay II: Learning to Protect any Defense AI Agent (MISQ, 1st Round)

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



Emulate Adversary with
Discrete Variational Actor-Critic (D-VAC)





Strengthen Robustness with
RL-based Robust Optimization (RL-RO)

- 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)

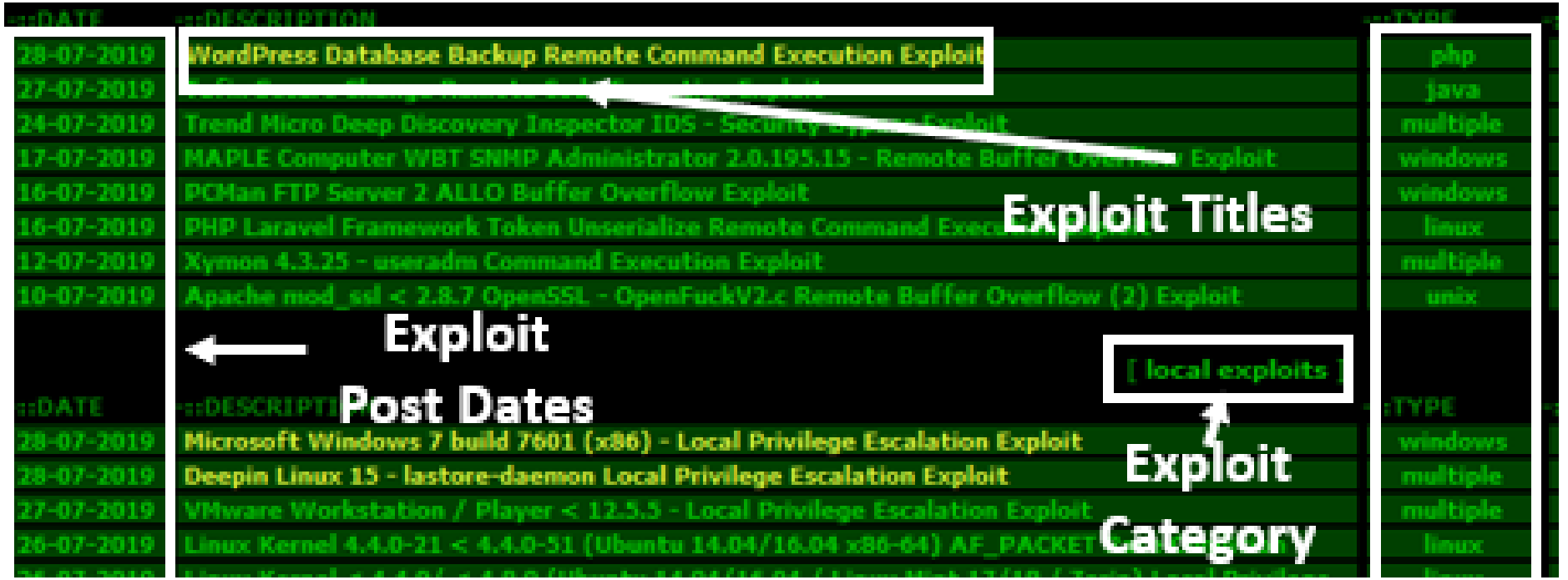
Protecting Scientific Instruments and Cyberinfrastructure:

From iPlant/CyVerse (life sciences)
to BioSphere 2/LEO (earth sciences)…
a new UA/USF/AZSecure NSF CICI project, 2019-2022

iPlant
Empowering a new plant biology
Collaborative

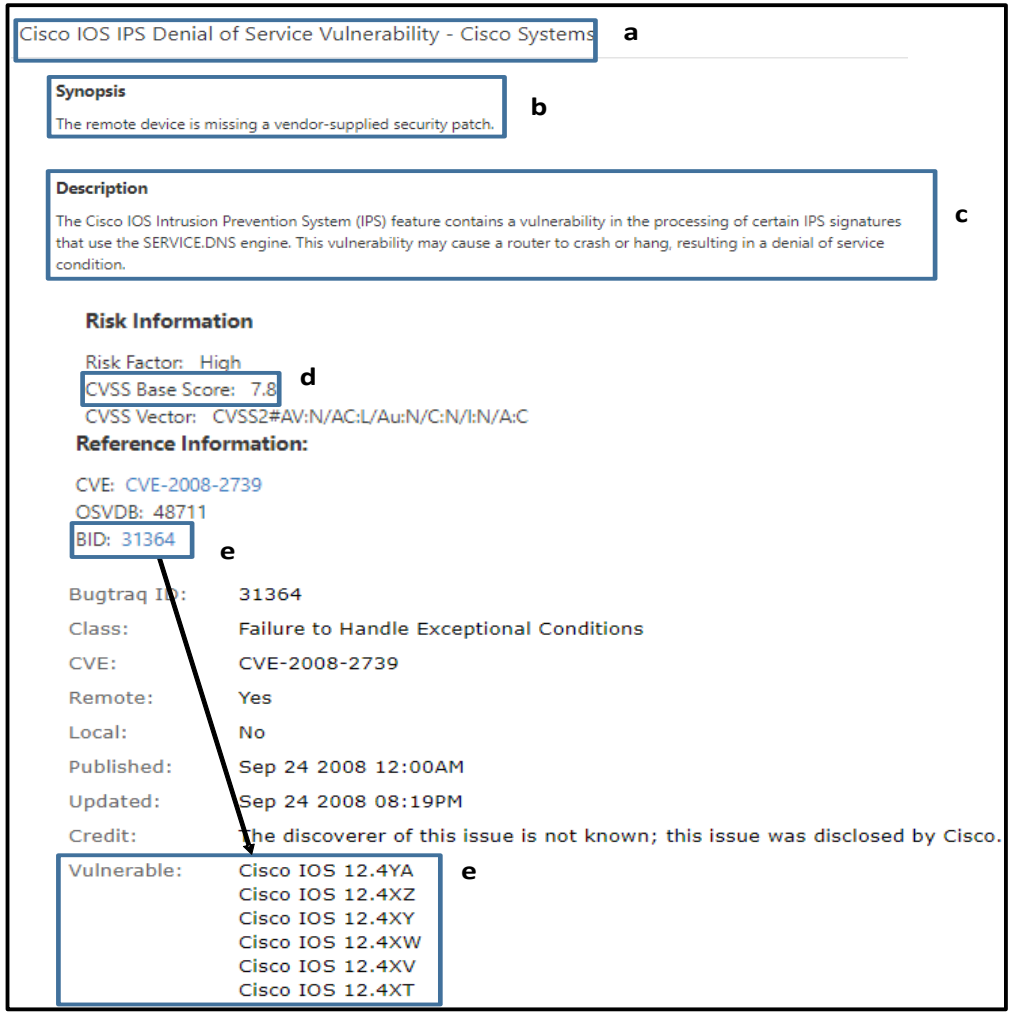
CYVERSE®

Hacker Forum Exploits



- **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



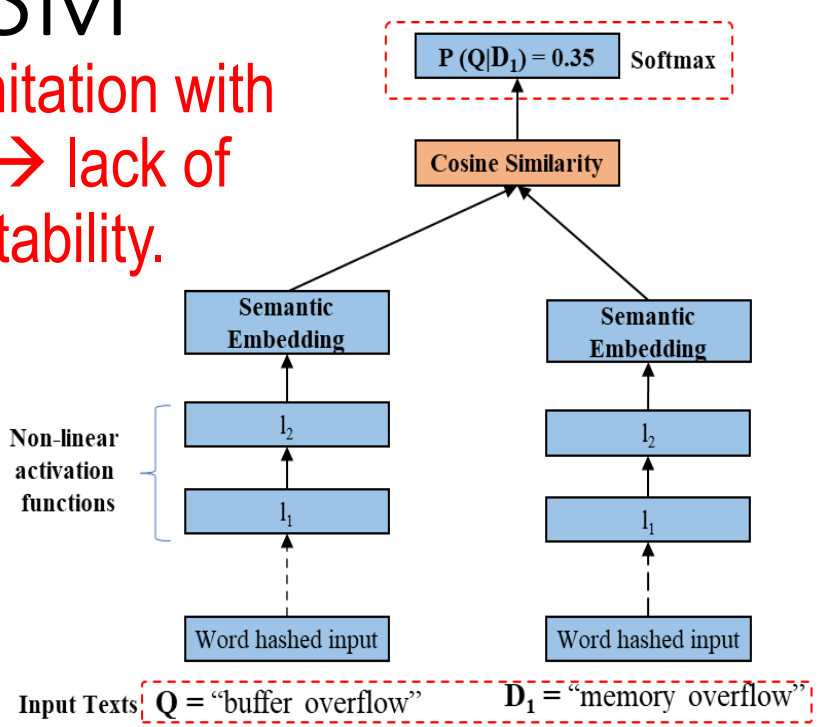
Category	Metadata	Description	Data Type
Description	Name	Short, descriptive name of vulnerability	Short text
1.	Family Name	Family vulnerability belongs to (e.g., Windows, etc.)	Categorical
	Description	Lengthy text description about vulnerability	Long text
	Synopsis	Short description of vulnerability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

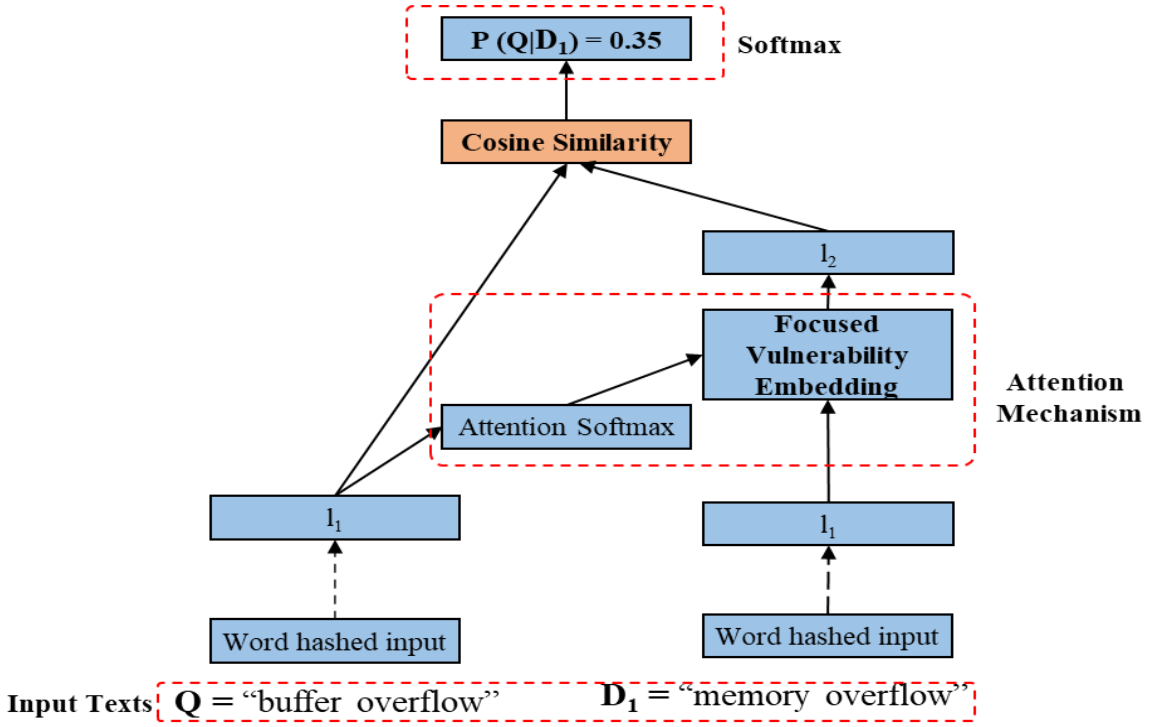
Key Attributes Returned by Modern Vulnerability Scanners

- **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-DSSM

- 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*	**0.7053**	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***
EVA-DSSM	**0.5469**	**0.6499**	**0.6857**	0.7023	**0.6834**	**0.6775**	**0.7779**	**0.7853**	**0.7865**	**0.8092**
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	**0.7947**	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***
EVA-DSSM	**0.6281**	**0.7602**	**0.7885**	**0.7684**	0.7863	0.3579	0.4550	0.4954	0.5133	0.6009

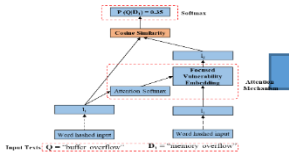
- EVA-DSSM outperforms all deep learning benchmarks
- Conv. or LSTM operations achieved lower performances
- Indicates that integrating an attention mechanism into the DSSM architecture does not deteriorate performance

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

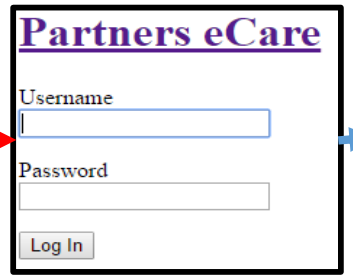
Procedure

SHODAN → Nessus® vulnerability scanner →  → 

Device Identification | **Vulnerability Scanning** | **EVA-DSSM** | **DVSM**

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
Total:	**344/1,879 (18.31%)**	**-**	**-**	**-**	**-**



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
…	…	…
-	-	**Total: 61.761**

- 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.

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

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

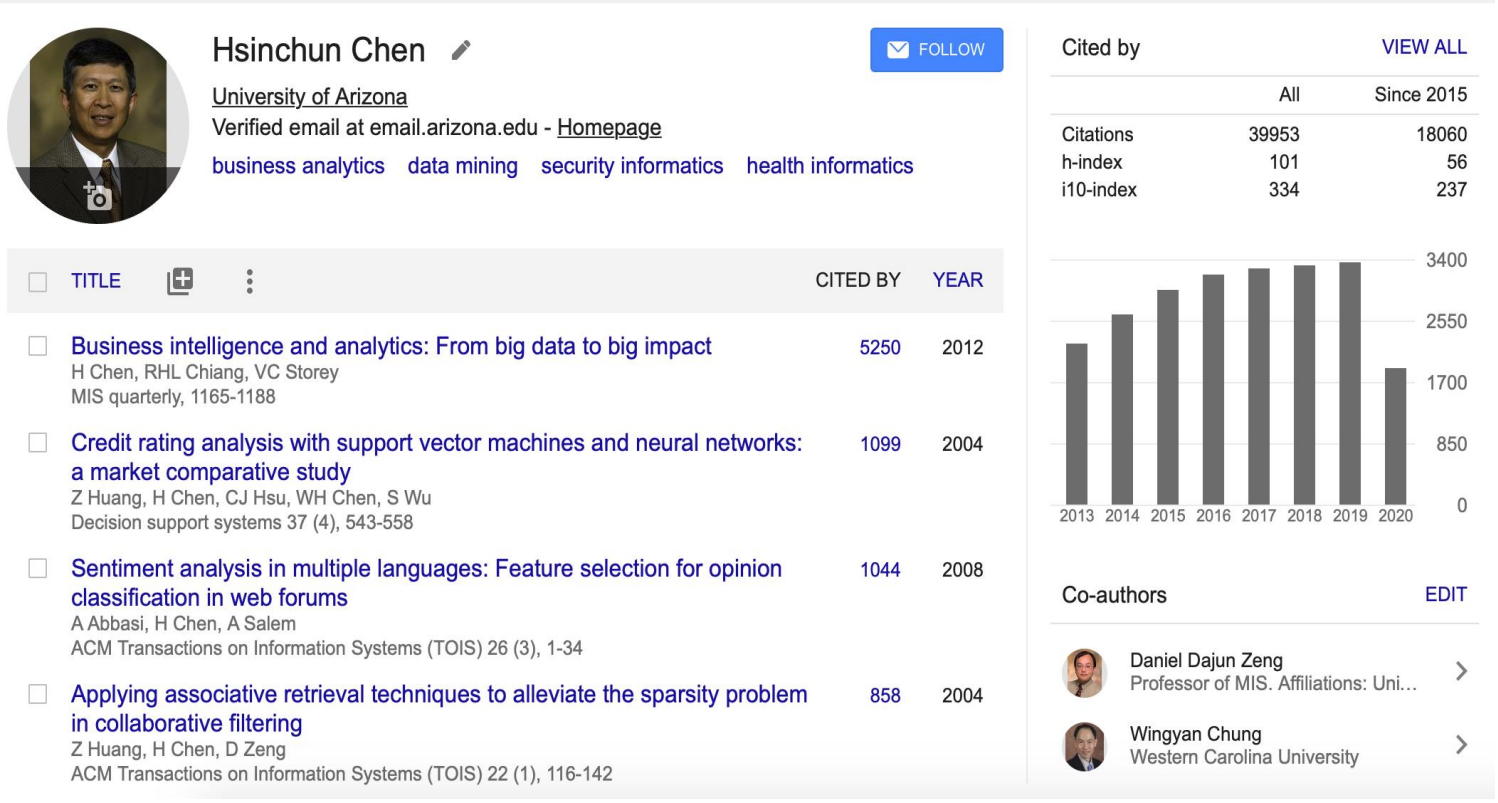
- i-School ($80K) & health informatics Journals: JASIST, ACM TOIS; JAMIA, JBI ➔ “informatics” (text) focused, system driven; helpful for NSF & NIH/NLM funding

- 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

FOLLOW

University of Arizona
Verified email at email.arizona.edu - Homepage
business analytics data mining security informatics health informatics

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

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

MISQ SEs

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

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 Patternsn

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)

2017

Security Analytics

[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)

Security Analytics; Best Paper, ICIS, 2010

[–] 2000 – 2009

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)

Social Media Analytics

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[1]

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@mccombs.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

DOI: 10.25300/MISQ/2020/14644

MIS Quarterly Vol. 44 No. 1, pp. 185-200March 2020

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?

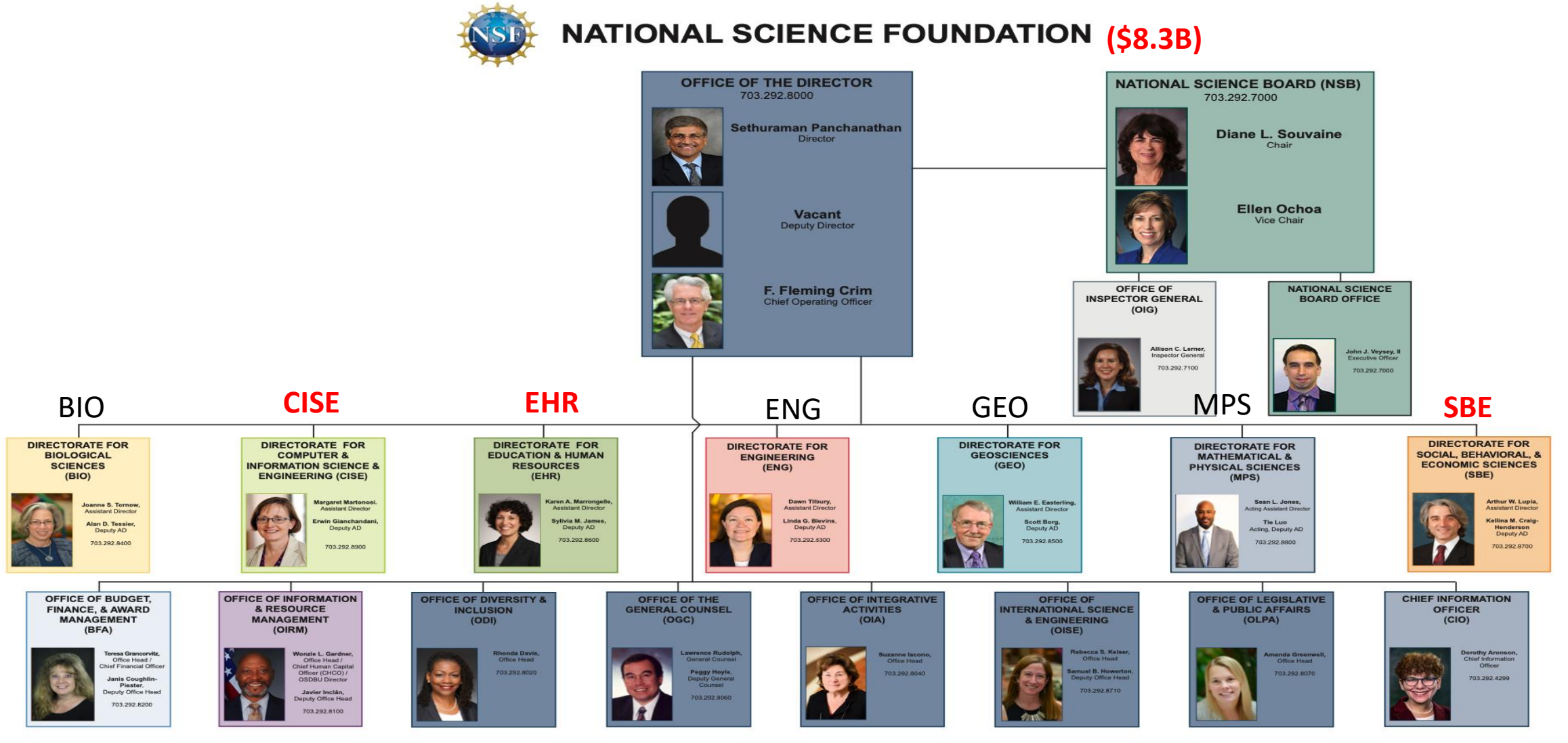
➔ Carefully study sample MISQ DS papers, e.g., (Abbasi, 2008; 2010).

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



Major Grants: NSF CISE/IIS/III

CISE

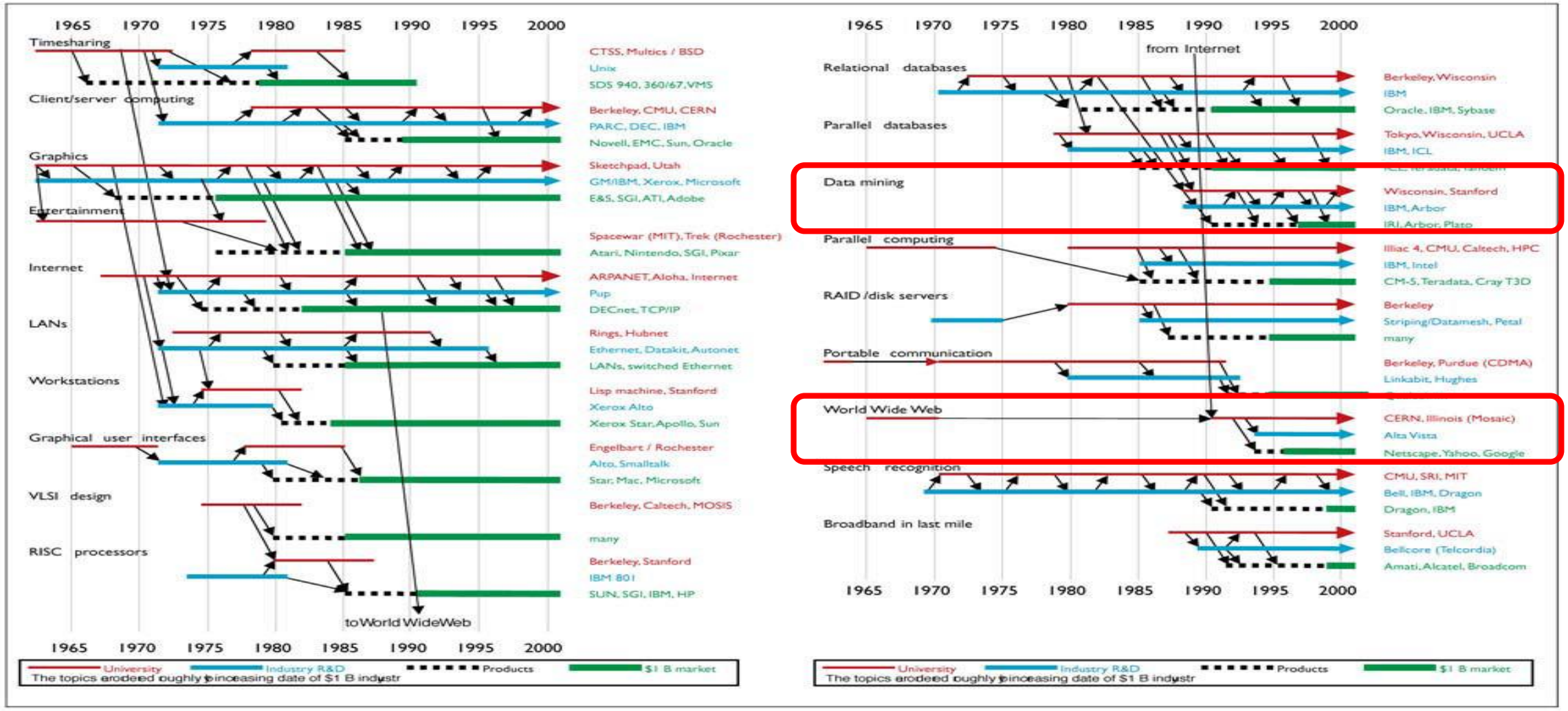
IIS/OAC

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.

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
http://ai.Arizona.edu