"Cybersecurity Big Data and Analytics Sharing"

- Big Data → Big Data Breach in Cybersecurity!

- Hsinchun Chen, UA, Conference/Workshop Chair, time keeper
- 2 sessions, 10-12 mins for each speaker; Q/A/contribution from audience after session (part of an NSF report)
- Session I: Bhavani Thuraisingham, UT Dallas, malware analysis; Latifur Khan, UTD, stream data analytics; Victor Benjamin, Arizona State U, blockchains for cybersecurity
- Session II: Resha Shenandoah, UA, Data Infrastructure Building Block (DIBBs) for security data; Sagar Samtani, UA, DIBBs tools, Hacker Assets Portal; Weifeng Li, UA, hacker underground economy (UA/Eller/MIS AI Lab)

Acknowledgement: National Science Foundation under Grant Number ACI-1443019 (DIBBs) & DGS-1719477 (SFS/SaTc)

Session Break Questions to Consider

- Questions and comments relating to session talks: What data or tools do you consider to be most useful for you and why? Other comments?

- Questions and comments relating to workshop in general: What additional data or tools do you wish to have and why? Other comments?

- Speaker slides/content and audience responses will be summarized in an NSF Workshop Report for distribution. (Please contact rshenandoah@email.Arizona.edu.)

Acknowledgement: National Science Foundation under Grant Number ACI-1443019 (DIBBs) & DGS-1719477 (SFS/SaTc)

Malware Data Collection & Analysis Using Big Data Tools

Cyber Security Research & Education Institute

The University of Texas at Dallas

Ramkumar Paranthaman

Dr. Bhavani Thuraisingham

Agenda

- Introduction
- Malware Data Collection
  - Malware Data Types
  - Malware Dataset Classification
  - Malware Collection Statistics
- Malware Analysis
  - Feature Extraction
  - Feature Selection
  - Train ML models
  - Results

Introduction

- This NSF-funded Data Infrastructure Building Blocks (DIBBs) project is intended to address a large gap in the availability of open source research data for researchers in ISI.
- The University of Arizona Artificial Intelligence Lab and its partners, the University of Virginia, The University of Texas at Dallas, Drexel University, and the University of Utah were to collect a significant archive of data and analysis tools to serve the ISI community.
- http://www.azsecure-data.org/

Data Collection – Repositories

- Classified datasets
  - Academic research projects
  - Security research corporations
- Unclassified datasets
  - Public malware datasets
  - Non-corporate research group malware datasets
- Malware collections
  - Independent collections of malware data
  - Malware sharing through forums

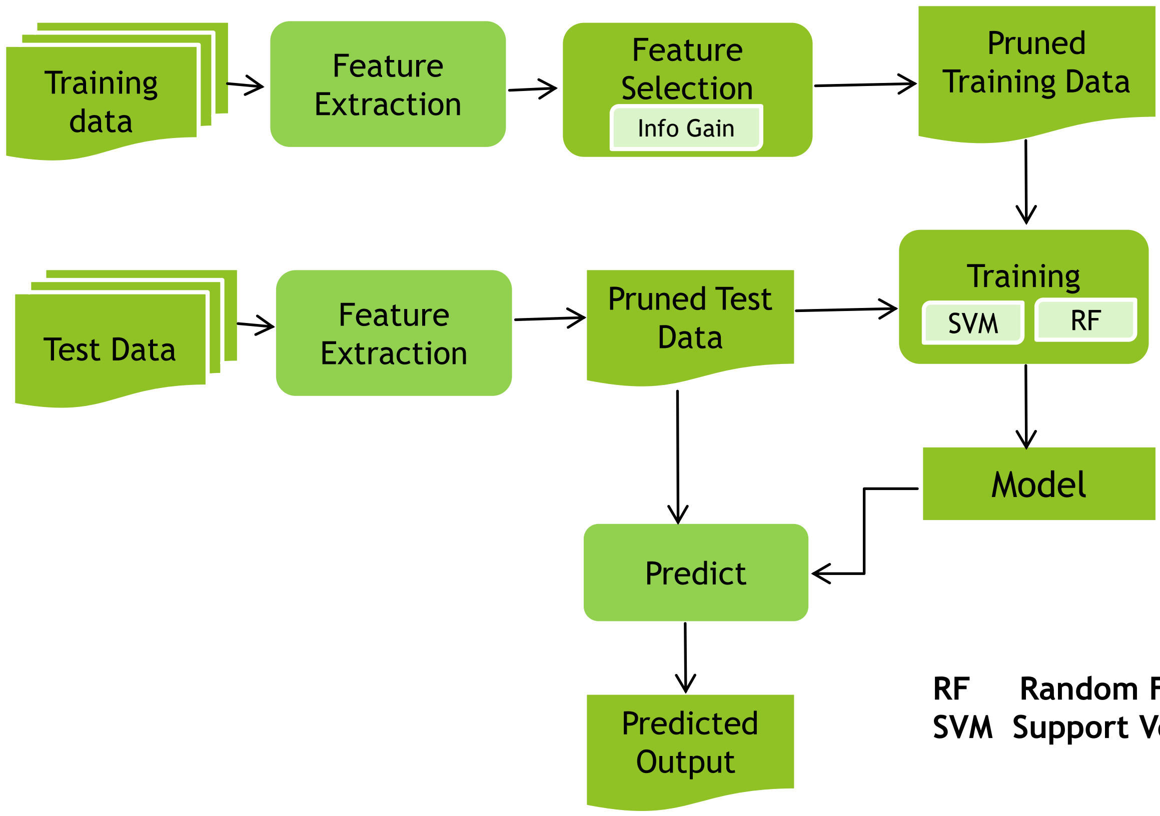
Malware Collection - Statistics

- Number of Classified Datasets 25 (circa 230 GB)
- Number of Unclassified Datasets 16 (circa 26 GB)
- Independent Datasets Gathered 3 (circa 2 GB)
- Total Size of Malware Datasets circa 250 GB

Malware Detection Framework

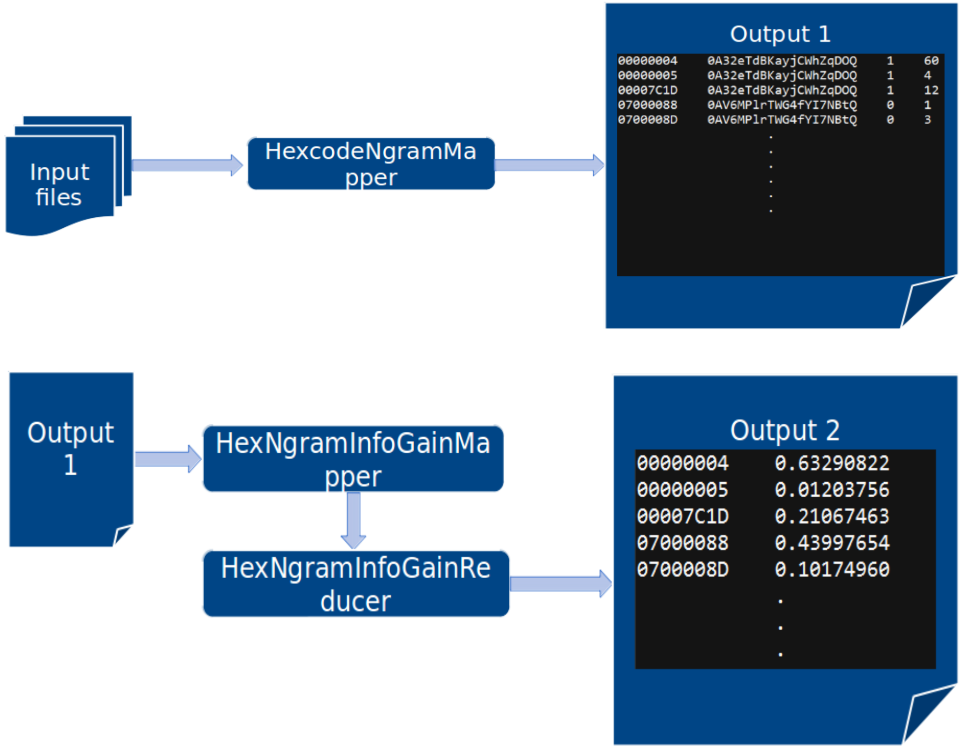
- **Objective**
  - Develop a malware detection framework using static analysis approach by employing Big Data tools and machine learning techniques
- **Implementation Steps**
  - Feature Extraction
  - Feature Selection
  - Training
  - Classification (detection)

System Workflow

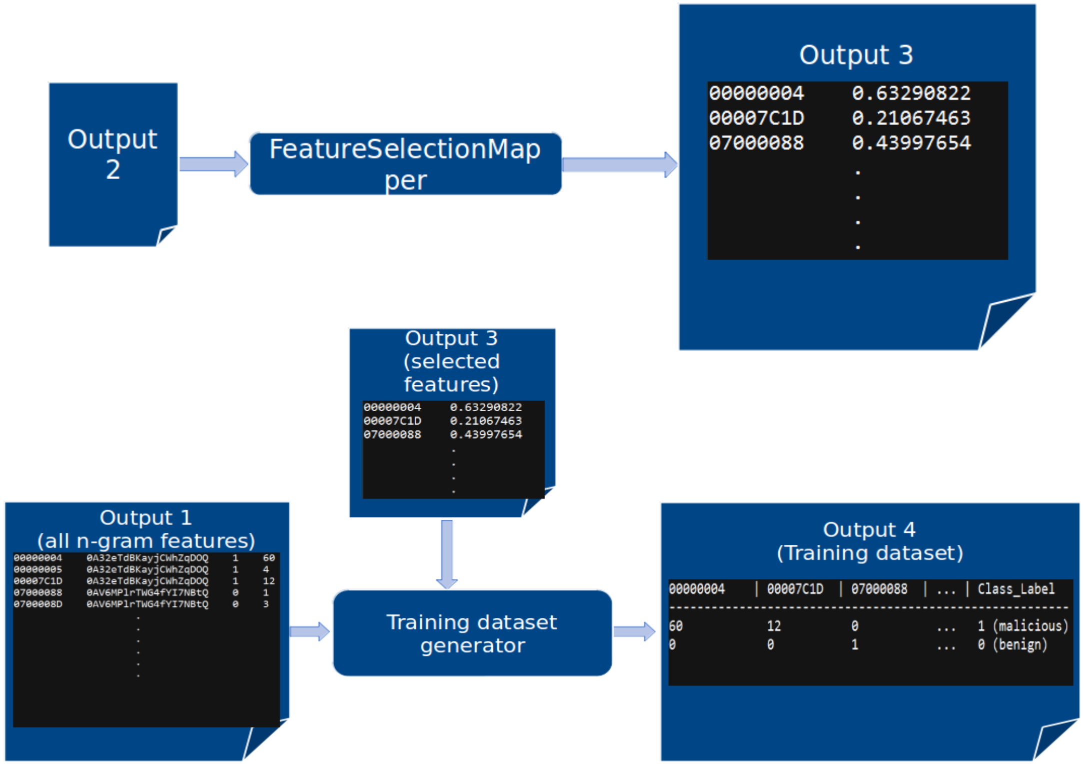


RF Random Forest
SVM Support Vector Machine

Feature Extraction – Map-Reduce Workflow



Feature Extraction – Map-Reduce Workflow



Feature Selection & Training

- **SELECTION**
  - Compute information gain for each feature
  - Select features whose information gain is above a threshold value
- **TRAINING - LEARNER MODELS**
  - Support Vector Machine (radial basis kernel)
  - Random Forest (J48 tree)

Inference

- **Input Dataset**
  - **Size - 200 GB**
  - **Type - Executables, DLLs, Hexcode dumps**
- **Observation**
  - Process time – 29.37 mins
  - Hadoop - though highly scalable, lacks performance due to high I/O usage (especially for large volume datasets)
- **Mitigation**
  - Use Apache Spark, distributed in-memory computing framework to improve performance

RESULTS

- Feature Extraction

Feature Type	Frequency
Byte 4-grams	95, 608, 217
Assembly 4-grams	419,888
DLL imports	26, 785
Opcode frequencies	82

- Feature Selection

Feature Type	Frequency
Byte 4-grams	46, 317
Assembly 4-grams	4, 309
DLL imports	65
Opcode frequencies	Not pruned

- Classification

Model	Accuracy
Random Forest	96.31%
Support Vector machine	95.05%

- Processing time – 16.42 minutes
- Source code https://github.com/helloram52/detectmalware

Trends and Perspectives in Big Data Research and Application

Latifur Khan, PhD

Professor

Department of Computer Science

University of Texas at Dallas, lkhan@utdallas.edu

Big Data: Issues

- Real Time
  - Data Processing Overhead needs to be minimized
    - Large Volume of Data needs to be consumed
  - Analytics
    - Response needs to be in real time.
    - Example: Real Time Anomaly Detection*
      - False Alarm may increase
- Scalable Analytics
  - Many Typical Algorithms Suitable for In-Memory processing
  - Demands Distributed Processing

*Solaimani M., Iftekhar M., Khan L., Thuraisingham B.: Statistical technique for online anomaly detection using Spark over heterogeneous data from multi-source VMware performance data. IEEE BigData Conference 2014: 1086-1094

Big Data: Solution

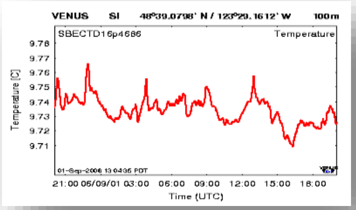
- Real Time Processing
  - Tool: Apache Spark, Storm, S4, Flink
- Real Time Analytics
  - SAMOA

- Scalable Analytics
  - Tool: Spark’s Machine Learning Library (MLlib), Mahout etc.
  - Covers Basic Analytics Algorithms
  - Advanced Algorithms (Relational Learning) are missing*
  - *Haque A., Chandra S., Khan L., Aggarwal C.: Distributed Adaptive Importance Sampling on graphical models using MapReduce. BigData Conference 2014: 597-602
  - *Ahsanul Haque, Zhuoyi Wang, Swarup Chandra, Yupeng Gao, Latifur Khan, Charu Aggarwal, Sampling-based distributed Kernel mean matching using spark. BigData 2016: 462-471

Big Data: Current & Future

- Stream Mining*
  - Update Learner Continuously
- Analytics
  - Supervised Learning (Ground Truth is required)
  - Labeling of Data is Problematic
    - Active Learning+

Network Traffic



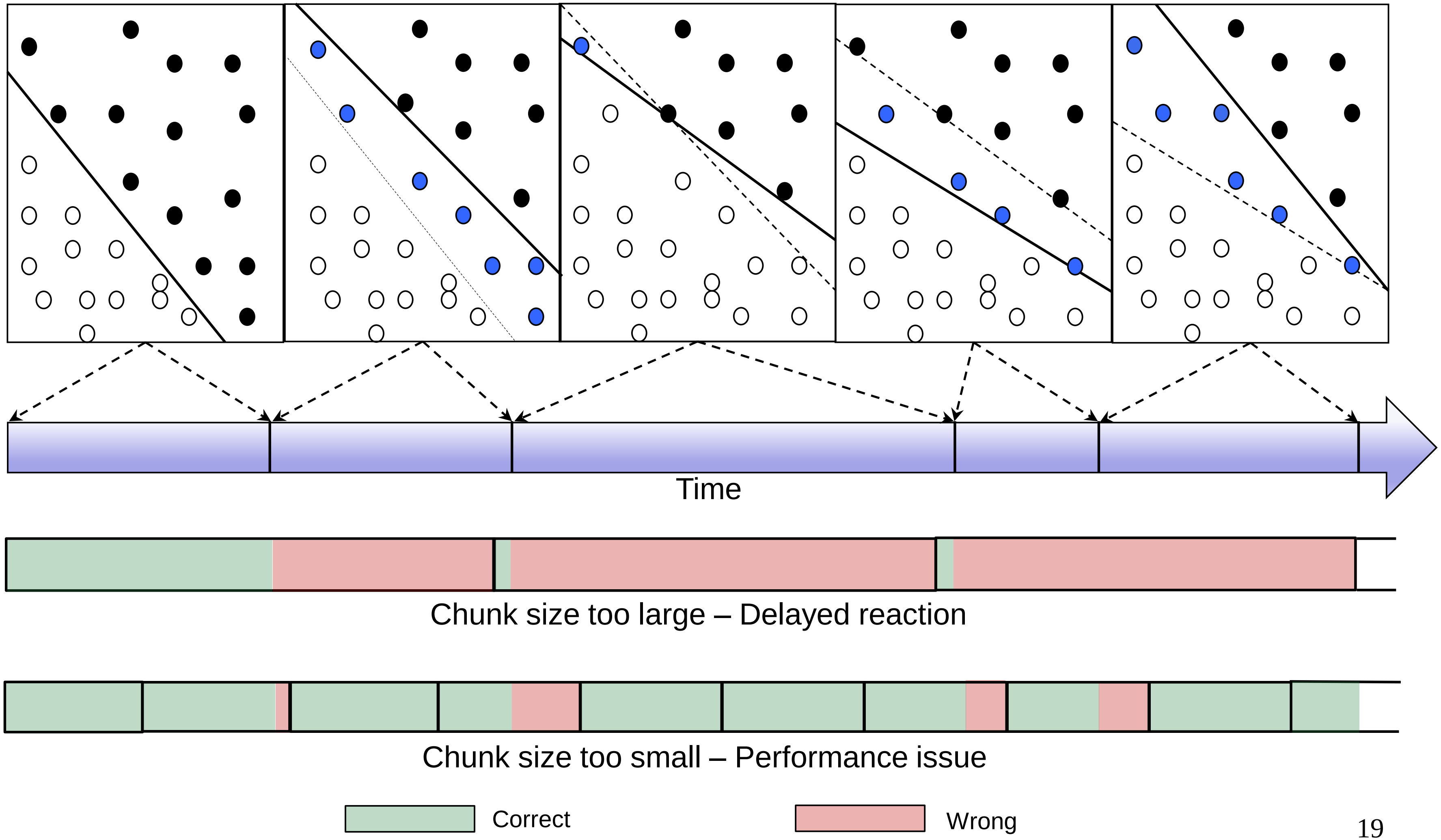
Sensor Data

*Parker, B., Khan, L.: Detecting and tracking concept class drift and emergence in non-stationary fast data streams. In Proc. Of Twenty-Ninth AAAI Conference on Artificial Intelligence. (Jan 2015).

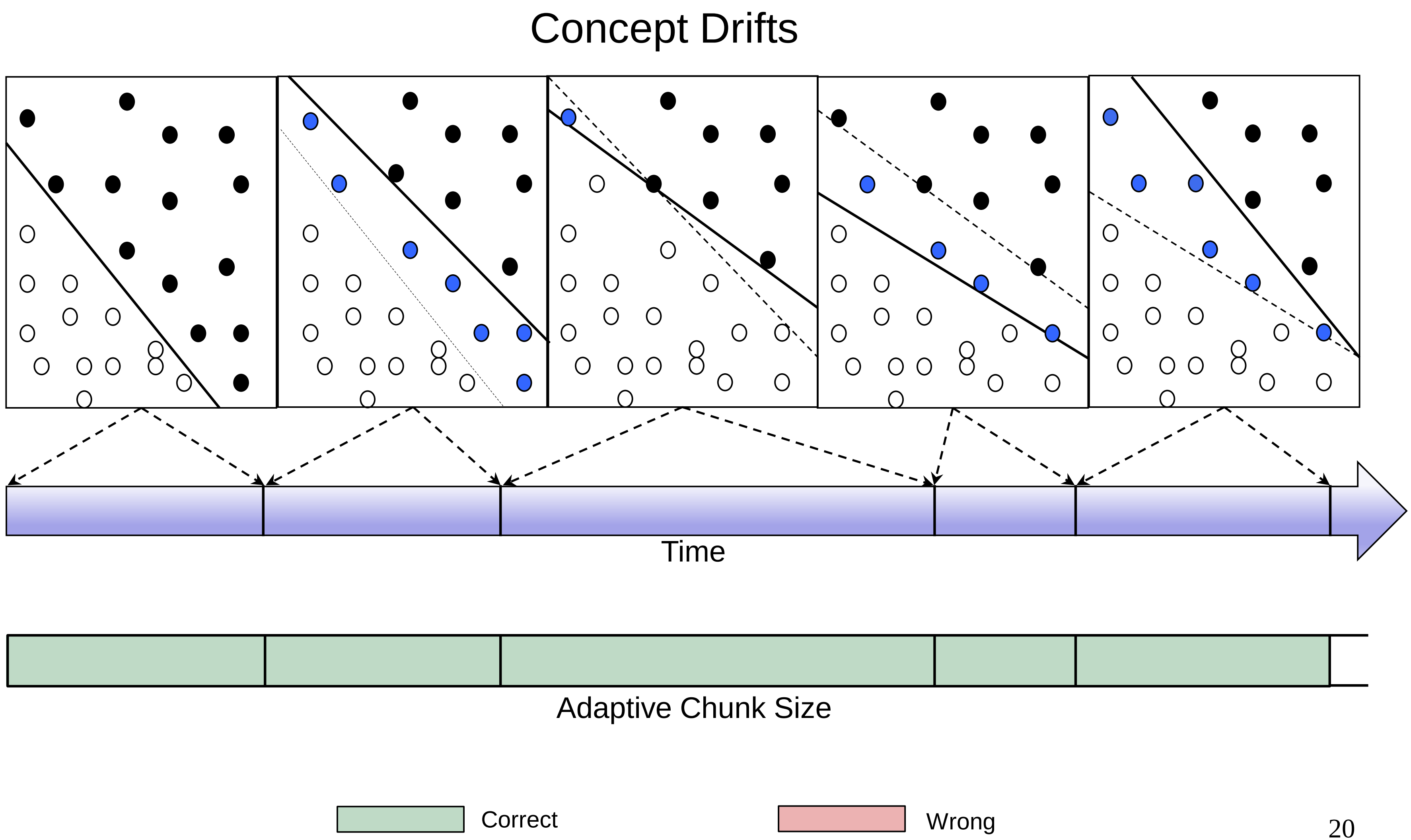
+Mohammad M. Masud, Jing Gao, Latifur Khan, Jiawei Han, Bhavani M. Thuraisingham: A Practical Approach to Classify Evolving Data Streams: Training with Limited Amount of Labeled Data. ICDM 2008: 929-934

Challenges: Fixed Chunk Size

Concept Drifts

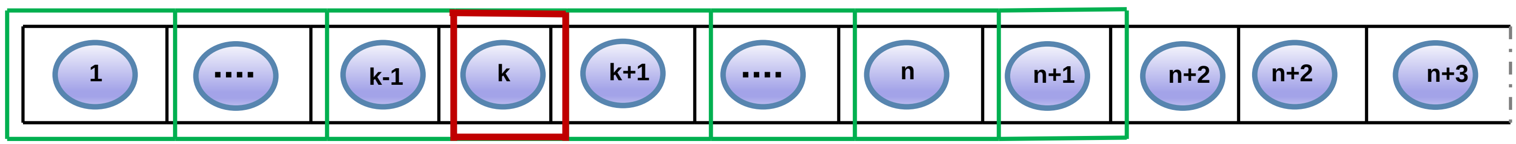


Solution: Adaptive Chunk Size



Adaptive Chunk - Sliding Window

Gamma et al. [1], *Bifet et al.* [2], *Harel et al.* [3]



- Existing dynamic sliding window techniques

  monitor error rate of the classifier.

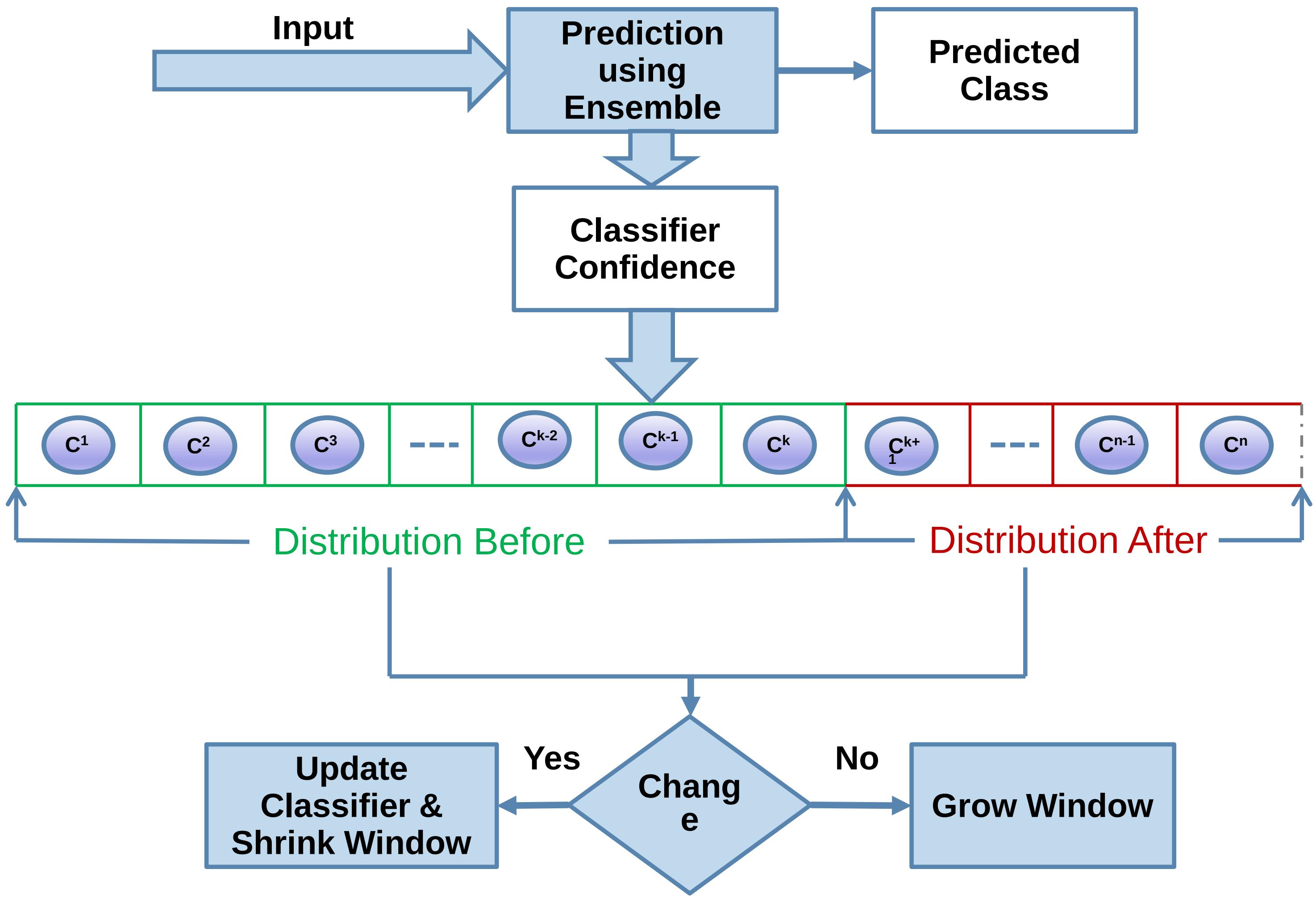
  Update classifier if starts to show bad performance.

  fully supervised, which is not feasible in case of real-world data streams.

[1] João Gama, Gladys Castillo: Learning with Local Drift Detection. ADMA 2006: 42-55
[2] Albert Bifet, Ricard Gavaldà: Learning from Time-Changing Data with Adaptive Windowing. SDM 2007: 443-448
[3] Maayan Harel, Shie Mannor, Ran El-Yaniv, Koby Crammer: Concept Drift Detection Through Resampling. ICML 2014: 1009-1017

Adaptive Chunk - Unsupervised

Haque et al. [1][2]



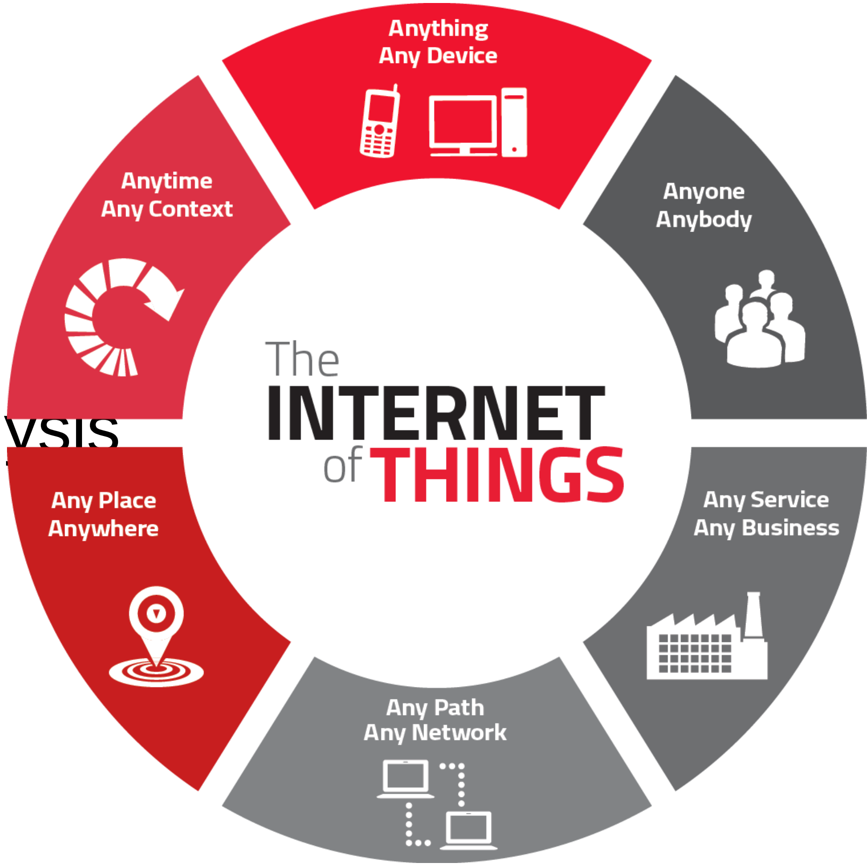
[1] Ahsanul Haque, Latifur Khan, Michael Baron, Bhavani M. Thuraisingham, Charu C. Aggarwal: Efficient handling of concept drift and concept evolution over Stream Data. ICDE 2016: 481-492.
[2] Ahsanul Haque, Latifur Khan, Michael Baron: SAND: Semi-Supervised Adaptive Novel Class Detection and Classification over Data Stream. AAAI 2016: 1652-1658.

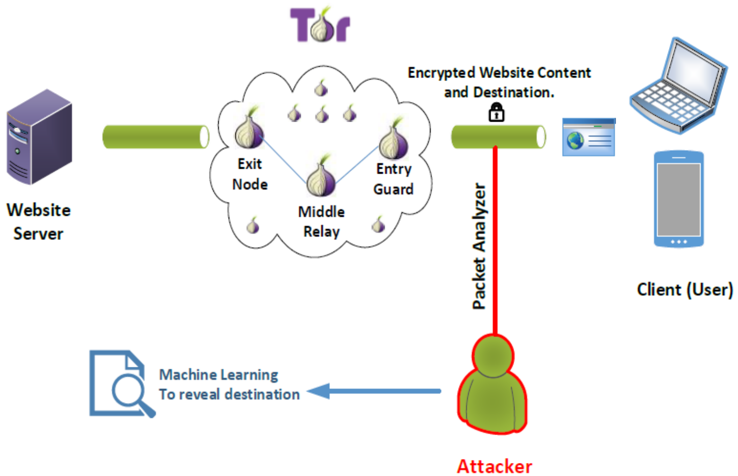
Big Data: Current & Future

- Stream Mining*
  - IOT Big Stream Mining
  - Security:

Encrypted Stream Traffic Analysis

  - Website Fingerprinting





*Parker, B., Khan, L.: Detecting and tracking concept class drift and emergence in non-stationary fast data streams. In Proc. Of Twenty-Ninth AAAI Conference on Artificial Intelligence. (Jan 2015).

Application: Encrypted Traffic Fingerprinting

Al-Naami et al. [1][2]

- Traffic Fingerprinting (TFP) is a Traffic Analysis (TA) attack that threatens web/app navigation privacy.
- TFP allows attackers to learn information about a website/app accessed by the user, by recognizing patterns in traffic.
- Examples: Website Fingerprinting

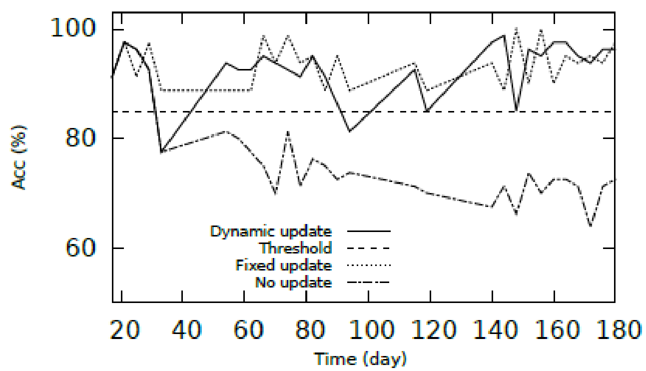
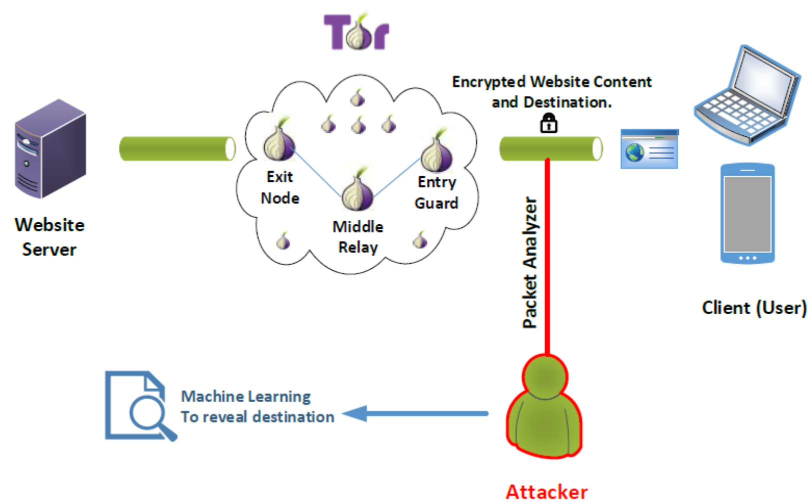
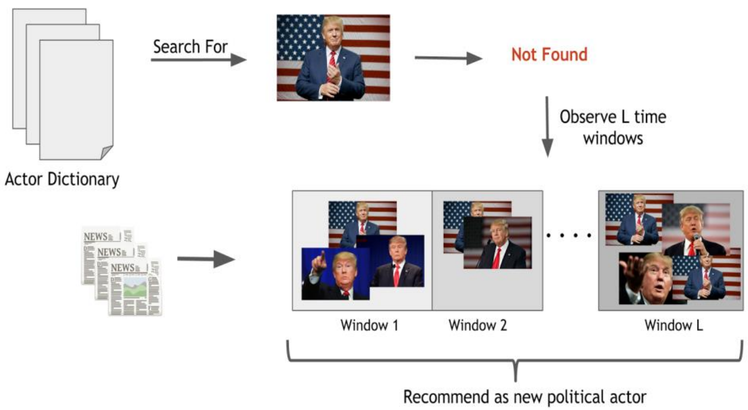


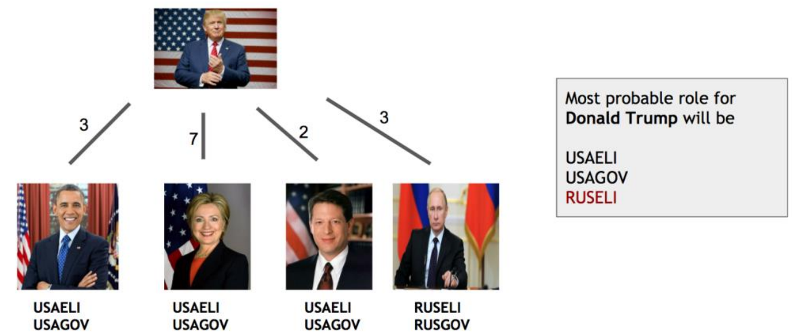
Figure 9: Adaptive Learning.

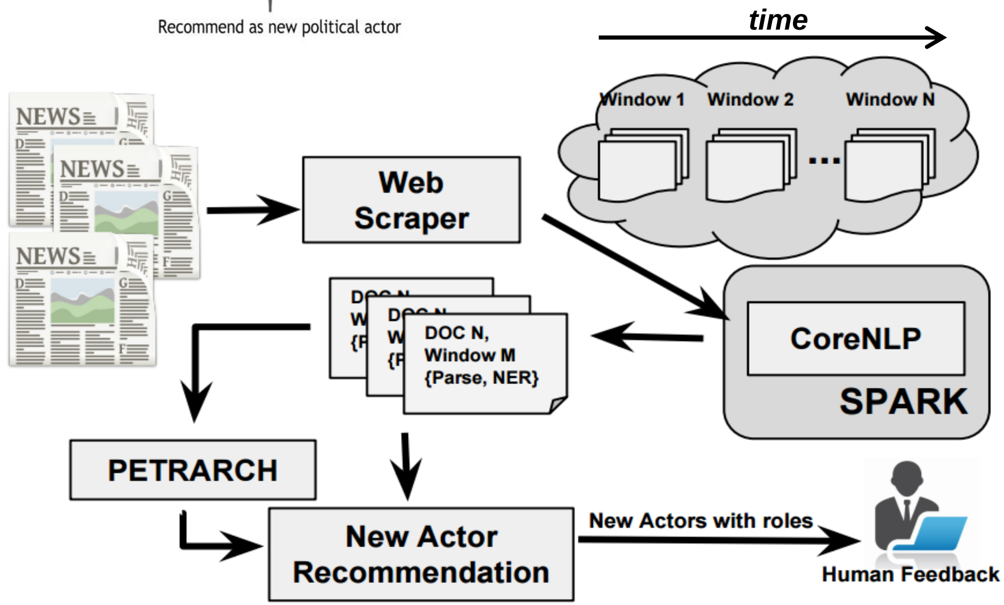


[1] K. Al-Naami, G. Ayoade, A. Siddiqui, N. Ruozzi, L. Khan and B. Thuraisingham, "P2V: Effective Website Fingerprinting Using Vector Space Representations,“ Computational Intelligence, 2015 IEEE Symposium Series on, Cape Town, 2015, pp. 59-66.
[2] K. Al-Naami, S. Chandra, A. Mustafa, L. Khan, Z. Lin, K. Hamlen, and B. Thuraisingham. 2016. Adaptive encrypted traffic fingerprinting with bi-directional dependence. In Proceedings of the 32nd Annual Conference on Computer Security Applications (ACSAC '16), Los Angeles, CA.

Application: Real-time Political Actor Detection Over Textual Political Stream







Real-time new political actor recommendation framework.

Challenges

- ✓Same actor with multiple alias names
- ✓Identify novel actor along with roles
- ✓Existing political actor's role changes over time
- ✓Processing high volume of news articles across the world

M. Solaimani, R. Gopalan, L. Khan, P. T. Brandt , and B. Thuraisingham, "Spark-based political event coding. In Big Data Computing Service and Applications (BigDataService), 2016 IEEE Second International Conference, Oxford, United Kingdom, on, pp. 14-23. IEEE, 2016"

Blockchains for Cybersecurity Research

Victor Benjamin, Ph.D.
Assistant Professor, Department of Information Systems
Co-Director, Actionable Analytics Lab

Introduction - Problem Context

- Industry thinks cybersecurity data sharing is good
  - Business-to-Business sharing (e.g., supply chains)
  - Business-to-Government (e.g., incident sharing)

- In reality, common reluctance to share data
  - Liability
  - Accessibility, transparency, and data ownership
  - Sharing platform focus, usefulness, and usability

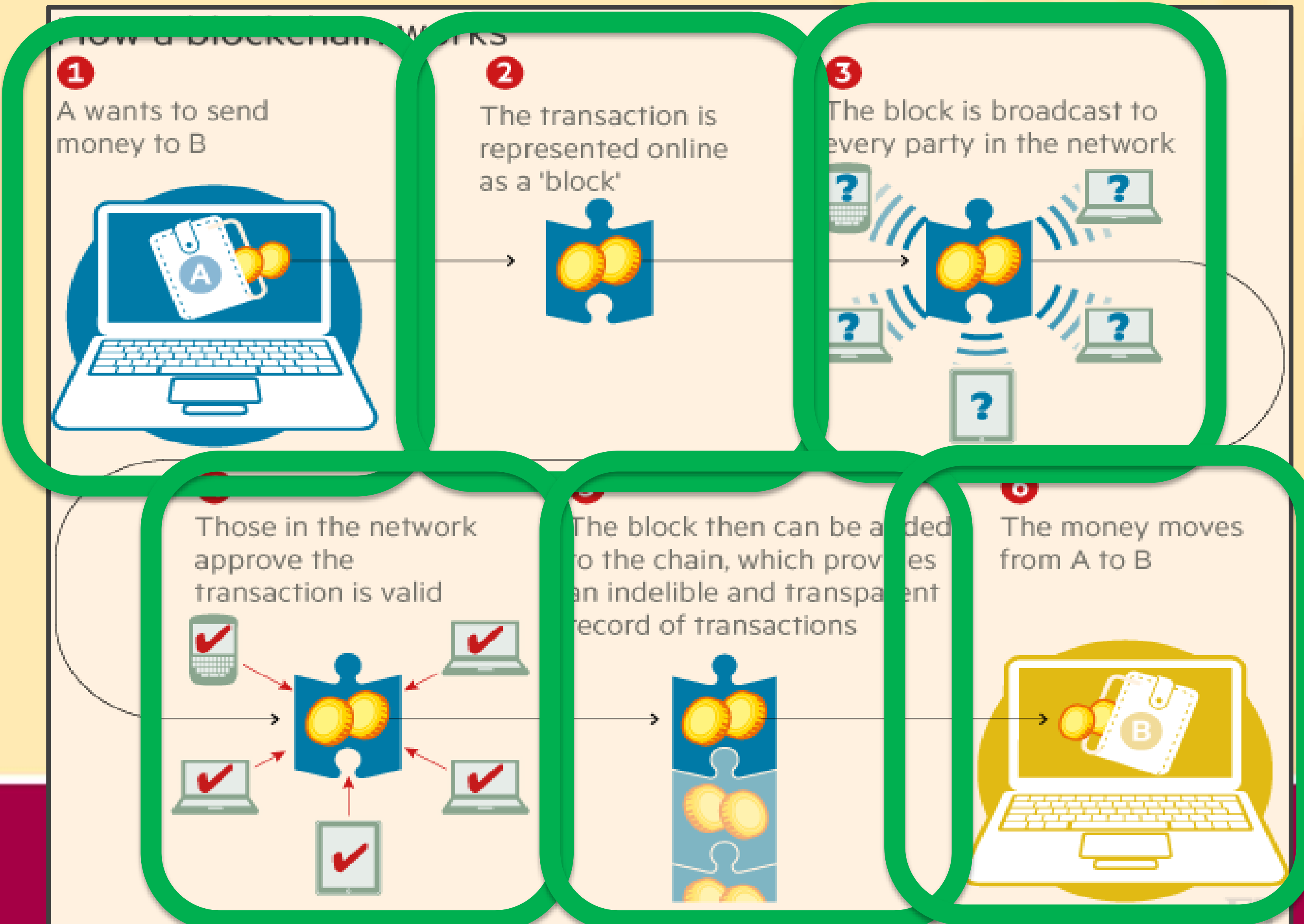
What can fix this?

- Need for platform that supports consortiums
  - Encourages community building
  - Can cater to special interest groups
    - E.g., Maritime Information Sharing and Analysis Center

A Path Forward

- Blockchains, the technology behind Bitcoin
  - First work on crypto-secured chain of blocks in 1991
  - First “modern” conceptualization of Blockchain in 2008
- Peer-to-peer networks
  - Managed autonomously
  - Highly configurable

Bitcoins – A Quick Primer

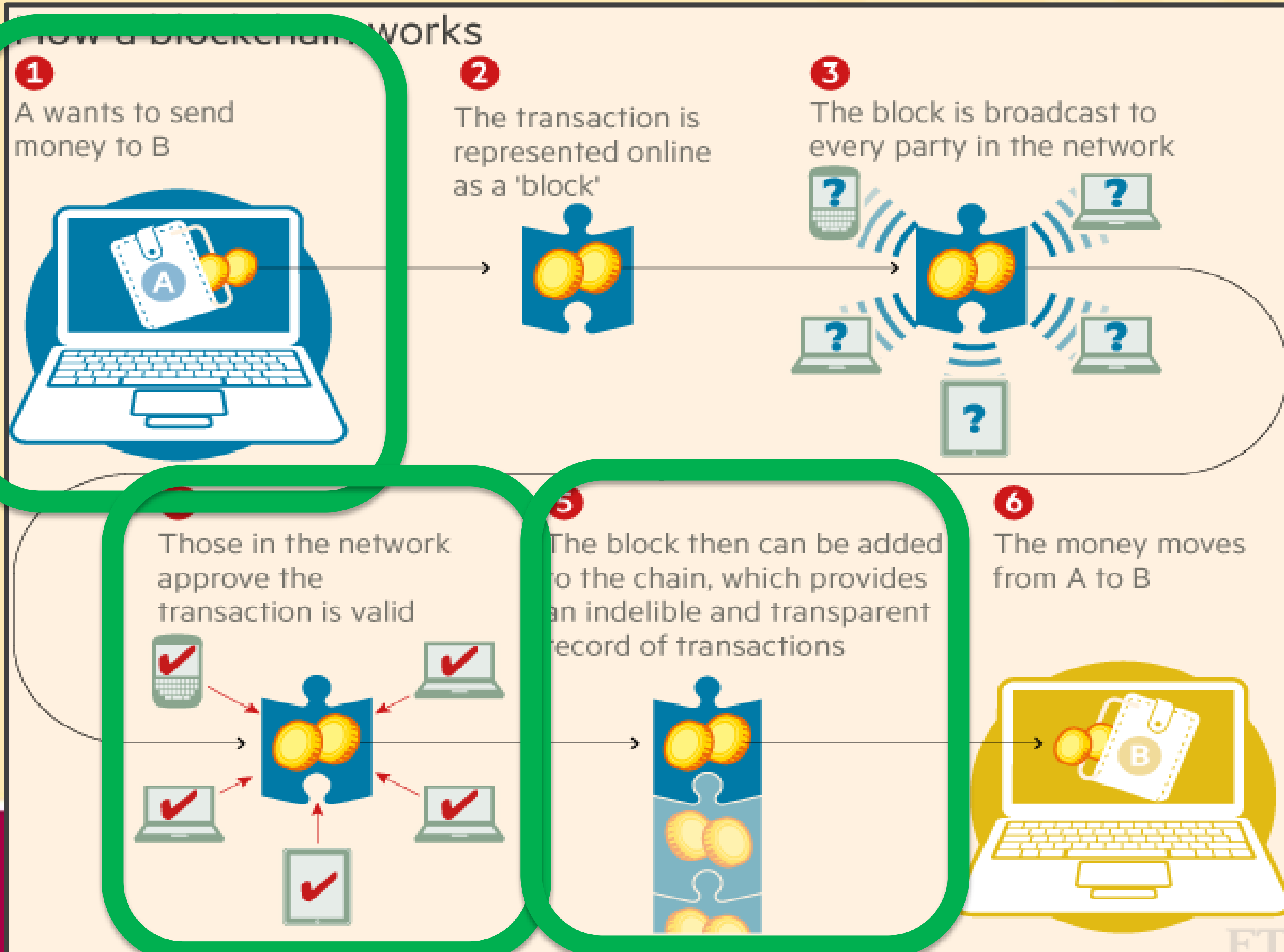


Blockchain Characteristics

- A distributed computing infrastructure offering:
  - Decentralized
  - Resiliency
  - Immutability
  - Security
  - Privacy

- Qualities for a cybersecurity data sharing platform

Blockchain for Cybersecurity Data



“Operation Green Rights”

Potential Targets

Downloadable Leaked Data Dumps

EU Sustainable Energy Week HACKED!!!

#EUSEW2014

more than 10.000 accounts from companies and governments

function,email,telephone,password

Worldbank,Bayer,ExxonMobil,Enel,Edf,GE,Shell,BP,Eni,Nokia,Intel...and many others.

operation initiative to encourage real #sustainable #energy

Hacked List:

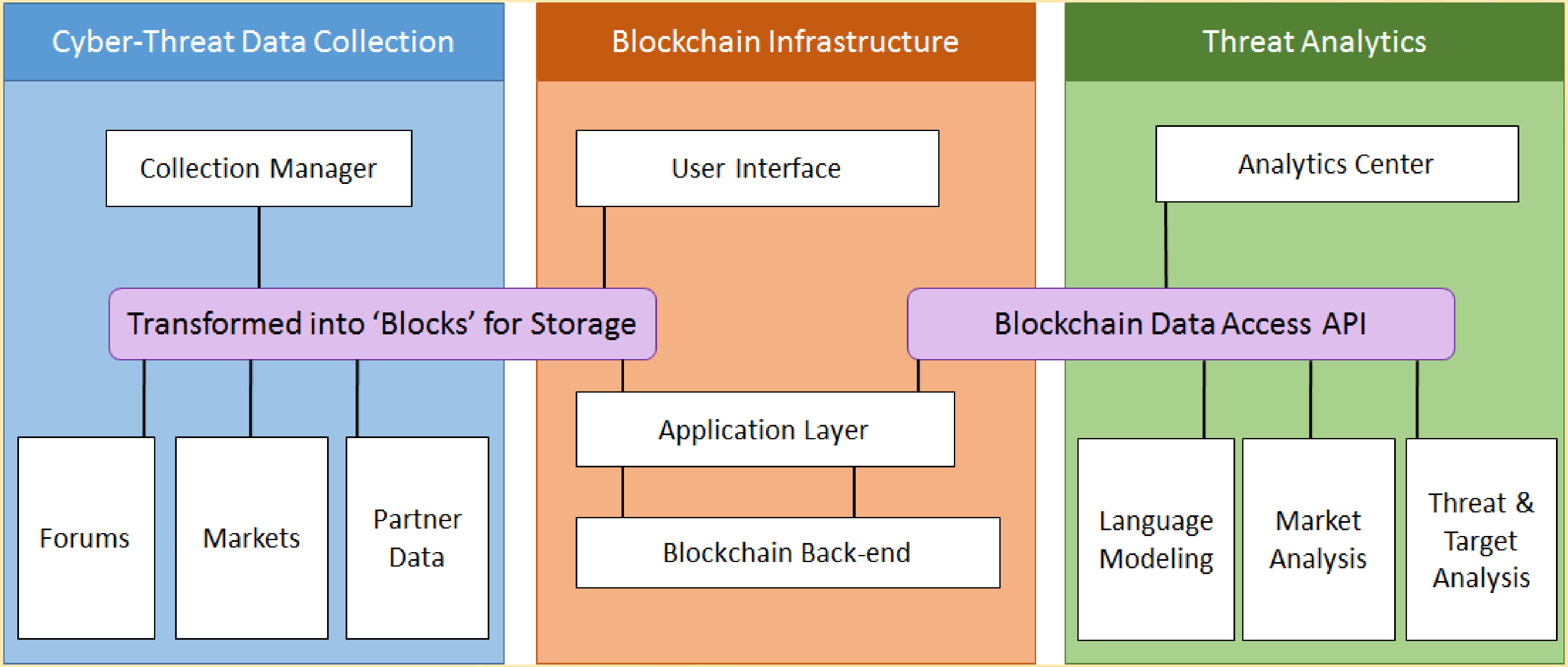
https://docs.zoho.com/file/egrja2d5495484a724a05a2495e9e73f81dcc (filetype .csv)

https://docs.zoho.com/file/egrja900ddc3ea08942d992eee71fc6b9f024 (filetype .ods)

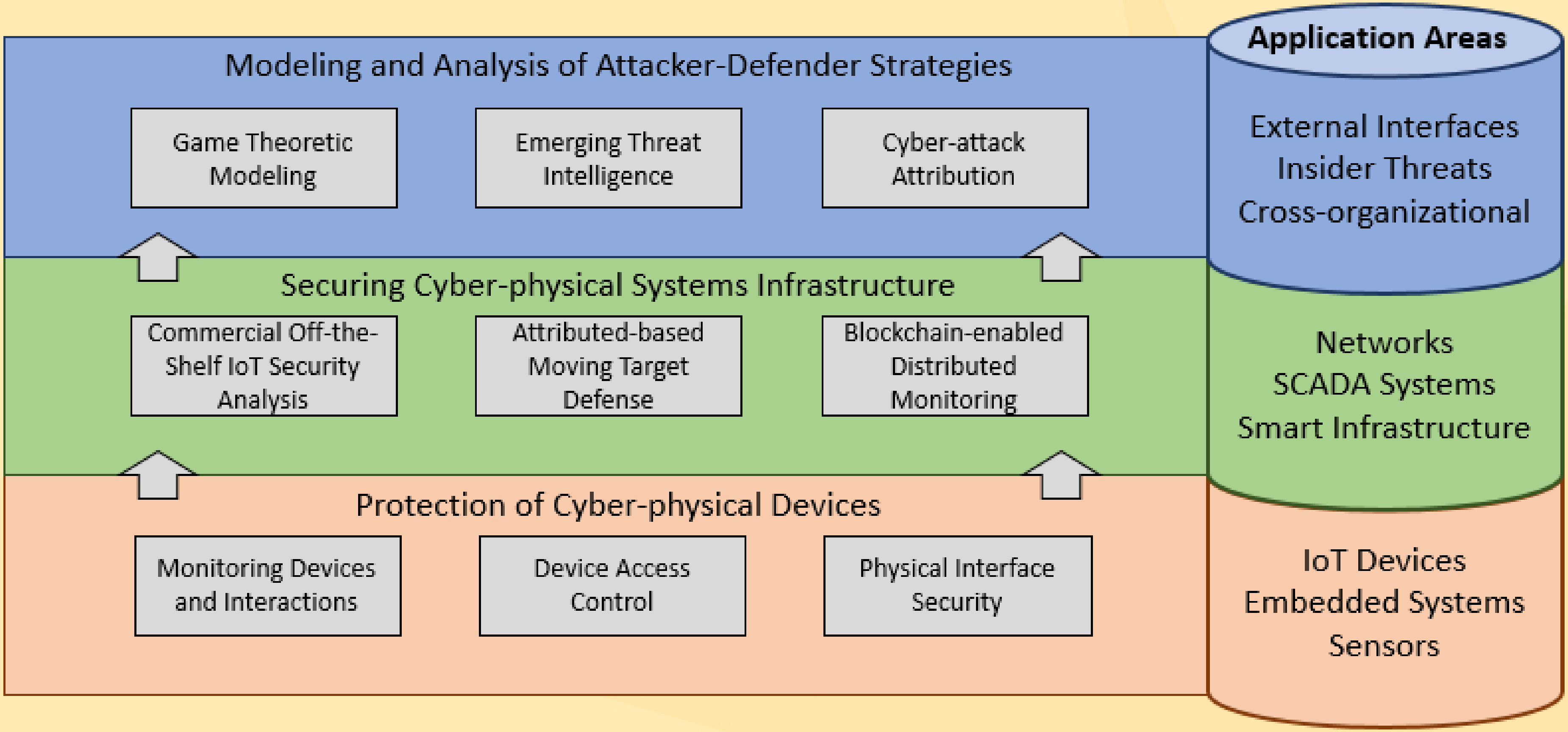
DOWNLOAD:

https://cdn.anonfiles.com/1403546269644.txt

Use Case: Threat Analytics



Use Case: Cyber-physical Security



Conclusion

- Platforms must be built with stakeholders in mind
- Blockchains offer a unique opportunity
- General take-away: think outside the box
  – Hackers do it
  – Security researchers and practitioners need to as well

Thanks!

- Ongoing exploration and proto-typing
- Interested? Contact me Victor.Benjamin@asu.edu

Resha Shenandoah

Digital Archivist

Project Manager, Data Infrastructure Building Blocks for Intelligence and Security Informatics (DIBBs-ISI)
University of Arizona Artificial Intelligence Lab

PI: **Dr. Hsinchun Chen**, University of Arizona. Co-PIs: **Dr. Mark Patton** and **Cathy Larson**, University of Arizona. Project Partners: **Dr. Ahmed Abbasi**, University of Virginia; **Dr. Paul Hu**, University of Utah; **Dr. Bhavani Thurasingham**, University of Texas at Dallas; **Dr. Chris Yang**, Drexel University.

This material is based in part upon work supported by the National Science Foundation under Grant Number ACI-1443019.

DIBBs-ISI: azsecure-data.org

14 Collections, 200+ GB total

Websites:	Phishing	171,360
	US Patriot, Hate, Militia 2009	74 identified by SPLC
		133 linked
Forums:	Geo Web	65
	Dark Web	28
	Hacker	2
	Chinese underground economy	2
Network Traffic:		4 collections
Malware Instances:		25,118 unique instances from 1 collection

Also collections containing chat logs and international news.

Languages:

Arabic, Chinese, English, French, German, Indonesian, Pashto, Russian, Urdu

File types:

arff, asp, binetflow, cfm, class, css, csv, exe, ghc, html, java, mpg, pcap, pdf, php, rar, sql, swf, txt, wd3, webarchive, wma, wmv, xlsx

DIBBs-ISI: azsecure-data.org

Between August 2016 and March 2017:

- 1,404 GB of data downloaded
- 17,190 file requests
- 51 distinct countries/regions originating requests

Most requested collection: PhishMonger

- 14,551 file requests

azsecure-data.org: PhishMonger

- Invokes the PhishTank API hourly
  - Indexes online, valid phishing sites
  - Typically 25,000 to 50,000 sites per request
  - Updated hourly
- Identifies newly added phish URLs
- Fetches new phishing websites
- Saves data

azsecure-data.org: PhishMonger

- Leverages exclusively open source software:
  - Ubuntu Linux, GNU Wget, Filezilla Server FTP
- Coded in Python 3.5
  - Harnesses the Twisted library for time based scheduling
- Runs on Amazon Web Services (AWS) Elastic Compute Cloud (EC2)
- Additional statistical scripts written in R

Most common file types include: png, html, jpg gif, js, css, ttf, svg, ico, woff

Contact:

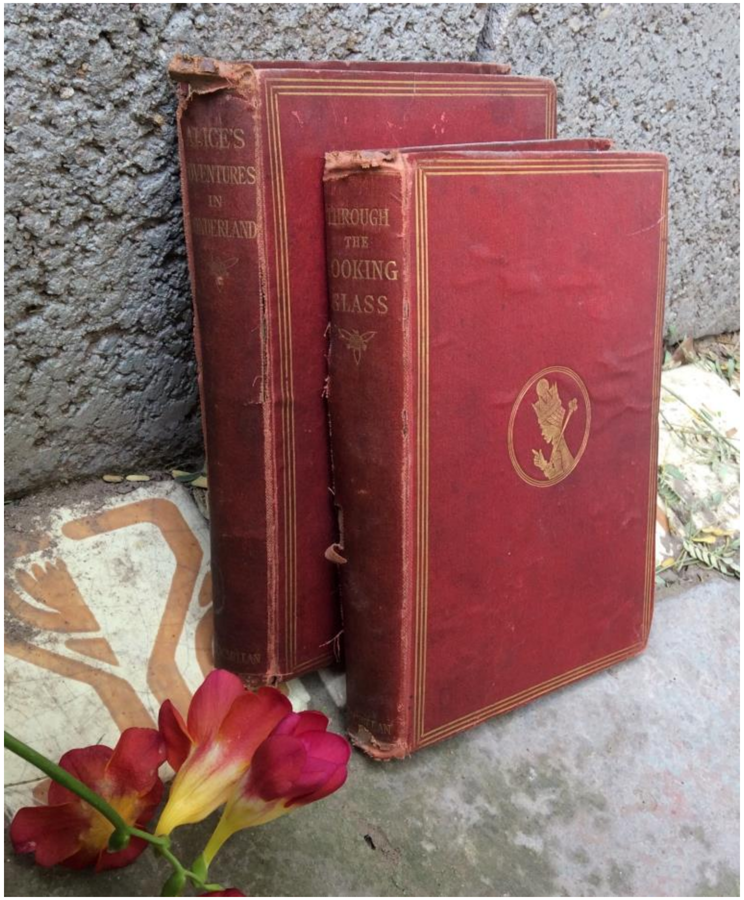
- Ahmed Abbasi, abbasi@comm.virginia.edu
- David G. Dobolyi, dd2es@comm.virginia.edu

AZSecure-data.org: What's Next?

DSpace

- Metadata
- Search or Browse
- OAI-PMH - Open Archives Initiative Protocol for Metadata Harvesting
  - Brings data out of silos
- Persistent identifiers – DOI, Orchid
- Built-in analytics

Data Preservation



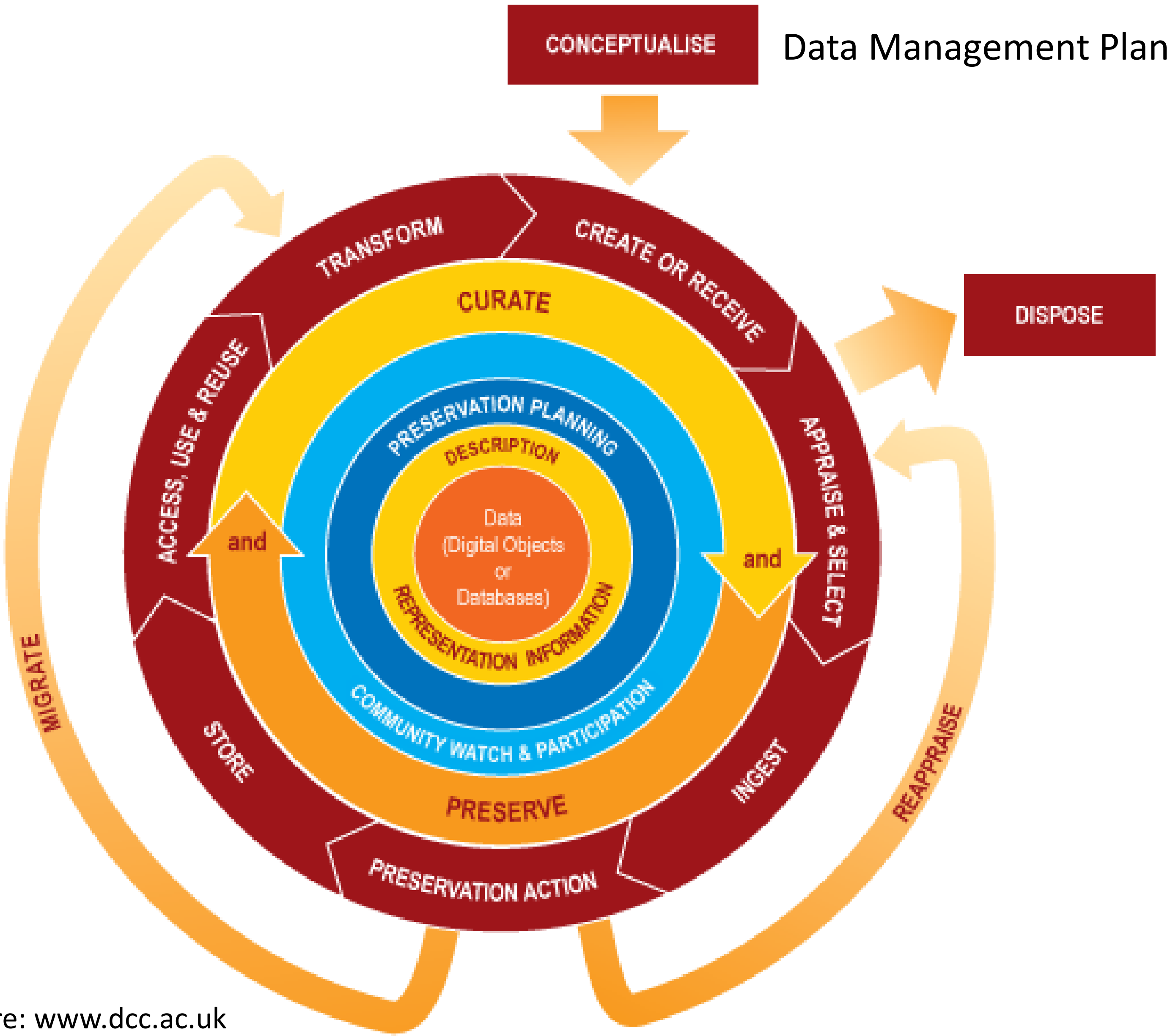
Alice Through the Looking Glass, ca 1871

HRC Clay Tablet. Sumerian. Ca 2400 BCE

http://www.hrc.utexas.edu/educator/modules/gutenberg/books/early/

Data Life Cycle

Data Management Plan



Digital Curation Centre: www.dcc.ac.uk

Data Management Skills

Data Management Skills Competency Matrix

JeSLIB 2017; 6(1): e1096
doi:10.7191/jeslib.2017.1096

Figure 1: Domains Represented in the Matrix

One-on-one outreach and training is highly effective, but not efficient or scalable.

Platform for Cybersecurity Big Data

- Resource for research data management
- Automate metadata creation
- Pipeline
  - Data management for research
  - Data sharing

- Continued user interest improves chances of sustaining cyberinfrastructure

- Must meet user needs

Resha Shenandoah – rshenandoah@email.arizona.edu
azsecure-data.org

DIBBs Tool Inventory for ISI Research

Sagar Samtani, Shuo Yu, Weifeng Li, Hongyi Zhu, Resha Shenandoah, Hsinchun Chen
Artificial Intelligence Lab, The University of Arizona

March 31, 2017

This material is based upon work supported by the National Science Foundation under Grant No. NSF ACI-1443019

Introduction

- Security researchers may face steep learning curves when attempting to identify tools that can aid them in developing valuable security insights from data sets.
- These slides aim to reflect some tools in the data analytics landscape that have been used in the AI Lab's past security informatics research.
- We present an inventory of tools into three major sections based on a traditional data analytics pipeline:
  - Collection and storage tools
  - Pre-processing and analytics tools
  - Visualization tools
- We also select a set of ISI papers to show how the tools can be used together to facilitate research.

Collection and Storage Tools

- The collection and storage component of relevant data is the first stage in typical data analytics exercises.

- Data collection aims to:
  - Identify and capture relevant fields of data from a specific source (e.g., web forums, Twitter, etc.)
  - Index and store it in a database or some other format which can be can be retrieved and used for pre-processing and further analytics.

- The collection process comprises three steps to pull from the online sources into the database: **extract**, **transform**, and **load** (ETL).
  - Table 1 summarizes tools to perform such tasks.

Collection Process: ETL

Collection Stage	Description	Category	Tool Name	Notes
Extract	Extracting data from their sources (e.g., websites, API's)	Spidering Tools	Offline Explorer	GUI for scheduling various crawling projects
			cURL	Offers proxy support, user authentication, etc.
			Wget	Recursive download, conversion of links
		Packages for Customized Spiders	HtmlUnit	A headless web browser written in Java
			Serenium	A browser automation library in Python
Transform	Transforming raw data into target data elements	Transformation	Regex	General string pattern matching
			JSoup	Java library for parsing HTML
			BeautifulSoup	Python package for parsing HTML and XML
			urllib	High-level interface for fetching data across the Web
Load	Loading data into data warehouse	Databases	MySQL	Widely used open-source RDBMS
			MS SQL Server	Commercial RDBMS by Microsoft
			Oracle Database	Commercial RDBMS by Oracle
			Apache HBase	Open-source, distributed, NoSQL DBMS on top of Hadoop
			Apache Hive	Open-source data warehouse infrastructure on top of Hadoop
			MongoDB	Open-source NoSQL DBMS. Uses JSON-like documents with schemas
			Apache Lucene	High-performance, full-featured text search engine library in Java

Table 1. Extraction, Transformation and Loading Tools

Pre-Processing and Analytics Tools

- Collected data needs to be pre-processed and transformed (cleaning, normalizing, transforming, tokenizing, etc.) prior to analysis.
  - Often consumes the majority (70-75%) of the time in data analytic projects.
- Past security analytics have used dozens of techniques after pre-processing, ranging from summary statistics to complex algorithms (e.g., deep learning).
- Many common data and text mining algorithms/applications are bundled into single packages (e.g., WEKA, Natural Language Toolkit (NLTK)).
  - Other analytics offered in specialized packages (e.g., hidden Markov models (HMM))
- Tables 2 and 3 summarize various pre-processing and analytical tools.

Pre-Processing and Analytics Tools

Category	Tool Name	Programming Language	Notes
General Data Mining	WEKA	Java, GUI	One-stop tools that cover common pre-processing, classification, and clustering algorithms. RapidMiner and WEKA can be used independently without a specific programming language.
	Scikit-learn	Python	
	RapidMiner	GUI	
	R	R	A widely used programming language and software environment for statistical computing and graphics.
General Text Mining	Natural Language Toolkit (NLTK)	Python	One-stop tools that cover word/sentence tokenization, POS tagging, parsing, chunking, named entity recognition, etc. NLTK has interfaces to call Stanford NLP tools.
	Stanford CoreNLP	Java	
	Apache OpenNLP	Java	
Hidden Markov Models (HMM)	hmmlearn	Python	General HMM package
	NLTK	Python	Specialized in POS tagging
Conditional Random Fields (CRF)	Stanford NER CRF	Java	CRF implementation for named entity recognition (NER)
	CRF++	C++	General CRF package
	NLTK	Python	Specialized in POS tagging
Latent Dirichlet Allocation (LDA)	Mallet	Java	Command line based tool for standard LDA
	Stanford Topic Modelling Toolbox	GUI	GUI based tool that supports LDA, labelled LDA, partially labelled LDA, and calculating perplexity. Can also perform temporal LDA
	Gensim	Python	Perform latent semantic analysis (LSA) and LDA in Python

Table 2. General and Specialized Data and Text Mining Tools

Pre-Processing and Analytics Tools

Category	Tool Name	Programming Language	Notes
Social Network Analysis (SNA)	UCINET	GUI	Licensed software (minimum $40) that can handle medium sized networks (2 millions nodes max)
	Gephi	GUI	Open source GUI based software that can handle larger data sizes than UCINET. Can read directly from databases
	NetworkX	Python	Python based network analysis tools. Can read from a variety of data sources. Allows for significant customization compared to other tools
Ontologies	WordNet	-	English lexical database grouped into synonyms
	SentiWordNet	-	Tagged WordNet with positivity, negativity, and neutrality for opinion mining
Word2vec	Gensim	Python, C	A two-layer neural net that processes text. Outputs a set of vectors: feature vectors for words in that corpus. Turns text into a numerical form for deep nets.
	DL4J	Java, Scala	
Deep Learning	Keras	Python	High-level neural networks library running on top of either TensorFlow or Theano. Recommended for fast experimentation.
	TensorFlow	Python, C++	Two different low-level implementations for deep learning models
	Theano	Python	

Table 3. General and Specialized Data and Text Mining Tools (cont'd)

Visualization Tools

- The final stage in the data often incorporates a visualization component.
- Desktop software (table 4) provide turnkey solutions to manage, connect, pivot data and render predefined types of visualizations in a GUI.
- For better customizability, lightweight toolkits, packages, and online services can be implemented along with analytical scripts (table 5).

Tool Name	Cost	Notes
Microsoft Excel	License required	Excel supports charts, graphs, generated from specified groups of cells. Excel 2010 and later support Pivot Table.
Tableau	Free education license	Generates graph types that can be combined into dashboards and shared over the internet.
ParaView	Free, open-source	Developed to analyze extremely large datasets using distributed memory computing resources.

Table 4. Desktop Visualization Software

Visualization Tools

Category	Tool Name	Programming Language	Notes
General Data Visualization Toolkits	Visualization Toolkit (VTK)	C++, Python, Java	General tools enabling users to customize their visualization components (e.g., point, line, axes, legends, layout, color coding) programmatically.
	OpenFrameworks (OF)	C++	
	Matplotlib	Python	
	ggplot2	R	
	Processing	Java, Python, JS	
	Seaborn	Python	
	pandas	Python	
Word Cloud	Wordle	Online, JS	Word cloud is a graphical representation of word frequencies. It can be used to visualize most frequently used keywords in the corpus.
Geo-map Tools	Mapbox	Online, JS	When location data (e.g. state, zipcode, latitude and longitude) is available, these geo-map tools can help you layout the data onto a map and generate visualizations such as color map, flow maps, etc.
	geoplotlib	Python	
	choroplethr	R	
Network Visualization Tools	Gephi	GUI, Java	Network visualization tools can visualize the relationship between data attributes or different data sources. The built in layout algorithms automatically generate visually pleasing graphs.
	networkx	Python	
	graph-tool	Python	
	igraph	R	
Color Selection (Aesthetic)	Color Brewer 2	Online	These color selection tools helps to improve the aesthetic of the visualization. They also provide safe color selections for web presenting, printing, color-blind cases.
	Palettable	Python	
	RColorBrewer	R	

Table 5. Lightweight Toolkits, Packages, and Online Services

Example ISI Papers

- To show the research context of applying the listed tools, we reviewed over 100 research papers from past ISI conferences and workshops.
  - 56 papers from IEEE ISI 2016
  - 47 from IEEE ISI 2015
  - 8 from FOSINT-SI 2016
  - 10 from ISI-ICDM 2015
- We selected representative papers to show how those tools can be used together to support and facilitate research.
  - References are attached at the end.

Example ISI Papers

Paper	Collection and Storage	Pre-Processing and Analytics	Visualization
Samtani et al. (2016)	Offline Explorer, MySQL, Regex	RapidMiner, Stanford Topic Modelling Toolbox	Tableau, D3.js
Grisham et al. (2016)	Selenium, MySQL	Stanford Topic Modelling Toolbox	-
Benjamin & Chen (2016)	Offline Explorer, MySQL, Regex	Word2vec	-
Benjamin & Chen (2014)	IRC Bots	WEKA	-
Samtani & Chen (2016)	Offline Explorer, MySQL, Regex	Gephi	Gephi
Solaimani et al. (2016)	MongoDB	CoreNLP, WordNet	-
Dobolyi & Abbasi (2016)	PhishTank API, Wget	R	R
Park et al. (2016)	SQLite	Apache OpenNLP	-

Table 6. Example ISI Papers

References

- Benjamin, V., & Chen, H. (2016, September). Identifying language groups within multilingual cybercriminal forums. In *Intelligence and Security Informatics (ISI), 2016 IEEE Conference on* (pp. 205-207). IEEE.
- Dobolyi, D. G., & Abbasi, A. (2016, September). PhishMonger: A free and open source public archive of real-world phishing websites. In *Intelligence and Security Informatics (ISI), 2016 IEEE Conference on* (pp. 31-36). IEEE.
- Grisham, J., Barreras, C., Afarin, C., Patton, M., & Chen, H. (2016, September). Identifying top listers in Alphabay using Latent Dirichlet Allocation. *In Intelligence and Security Informatics (ISI), 2016 IEEE Conference on* (pp. 219-219). IEEE.
- Park, A. J., Beck, B., Fletche, D., Lam, P., & Tsang, H. H. (2016, August). Temporal analysis of radical dark web forum users. *In Advances in Social Networks Analysis and Mining (ASONAM), 2016 IEEE/ACM International Conference on (pp. 880-883).* IEEE.
- Samtani, S., & Chen, H. (2016, September). Using social network analysis to identify key hackers for keylogging tools in hacker forums*. In Intelligence and Security Informatics (ISI), 2016 IEEE Conference on* (pp. 319-321). IEEE.

- Samtani, S., & Chen, H. (2016, September). Using social network analysis to identify key hackers for keylogging tools in hacker forums. *In Intelligence and Security Informatics (ISI), 2016 IEEE Conference on* (pp. 319-321). IEEE.

- Samtani, S., Chinn, K., Larson, C., & Chen, H. (2016, September). AZSecure Hacker Assets Portal: Cyber threat intelligence and malware analysis. *In Intelligence and Security Informatics (ISI), 2016 IEEE Conference on* (pp. 19-24). IEEE.
- Solaimani, M., Salam, S., Mustafa, A. M., Khan, L., Brandt, P. T., & Thuraisingham, B. (2016, September). Near real-time atrocity event coding. In *Intelligence and Security Informatics (ISI), 2016 IEEE Conference on* (pp. 139-144). IEEE.

AZSecure Hacker Assets Portal:
Enhancing Cybersecurity Education

Sagar Samtani, Kory Chinn, Cathy Larson, Hsinchun Chen

Artificial Intelligence Lab, The University of Arizona

March 31, 2017

This material is based upon work supported by the National Science Foundation under Grant No. NSF DUE-1303362 (SFS) and NSF SES-1314631 (SaTC).

AZSecure Hacker Assets Portal Team

Sagar Samtani
-3rd Year Ph.D. and SFS Student
-University of Arizona

Kory Chinn
-Undergraduate Senior
-University of Arizona

Cathy Larson
-Former Associate Director, AI Lab
-University of Arizona

Dr. Hsinchun Chen
-Regents' Professor
-Director, AI Lab
-University of Arizona

Introduction: “Know Your Enemy”

- Recent years have seen a significant increase in cybersecurity education initiatives.
- One novel way to enhance cybersecurity education and bolster future cyber-defenses is to directly study tools disseminated in online hacker communities.
- **Online hacker forums allow hackers to share assets such as malicious tutorials, code, attachments**.
- Spanning regions such as the US and Russia, there are tens of millions of posts in hundreds of forums made by millions of members.
  - Tens of thousands of malicious assets

Introduction – Hacker Asset Examples

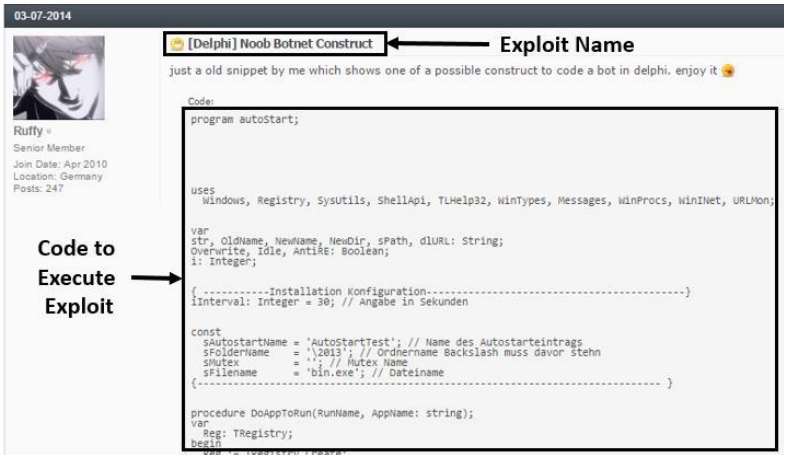


Figure 1. Forum post with source code to create botnets

Figure 2. Forum post with BlackPOS malware attachment

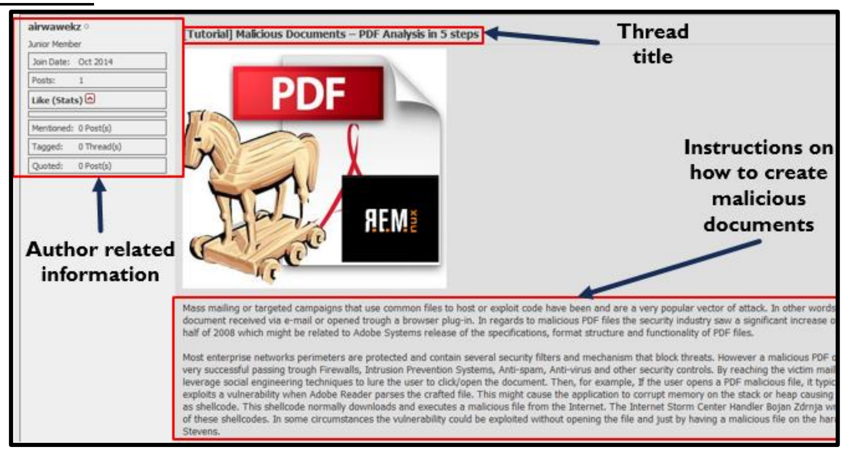


Figure 3. Tutorial on how to create malicious documents

Introduction – AZSecure Hacker Assets Portal Objective

- Given the rich nature of hacker forum data, ***we aim to design a web portal providing hacker forum contents and analysis for cybersecurity education, research, and training purposes.***

- We achieve this goal by:
  - Identifying large English, Russian, and Arabic hacker forums
  - Extracting assets using advanced web crawling approaches
  - Analyzing assets using scalable text and data analytic methods
  - Developing a portal allowing users search, download, and analyze assets

AZSecure Hacker Assets Portal – Data Testbed

- We use a Tor routed web crawler to automatically collect one Arabic, two Russian, and two English forums known for containing malicious assets (Table 1).
- 15,576 code, 14,851 attachments, and 987 tutorials posted between 2/7/05-10/31/16.
- In addition to integrating other forums, we update our collection monthly to continually identify new and emerging assets.

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**

Table 1. Summary of AZSecure Hacker Assets Portal System Data

AZSecure Hacker Assets Portal – Data Mining Approach

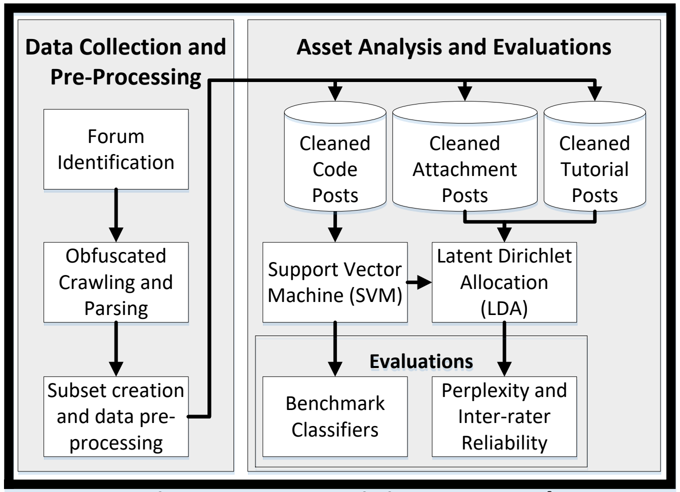


Figure 4. Data Mining Approach

Algorithm	Accuracy	Precision	Recall	F1
SVM	**98.20**	**96.36**	**98.20**	**98.28**
k-Nearest Neighbor	64.00	83.47	64.00	72.24
Naïve Bayes	86.00	88.57	86.00	87.26
Decision Tree	82.60	86.41	82.60	84.42

Table 2. Benchmark Classifier Evaluation Results

- We use two automatic methods to sort assets (Figure 4).
- First, we trained a Support Vector Machine (SVM) with 1,000 code files to classify hacker code into 10 languages.
  - Java, Python, C/C++, HTML, Delphi, VB, SQL, Ruby, and Perl
  - SVM outperformed other classifiers in standard metrics (Table 2)
- We then use Latent Dirichlet Allocation (LDA) to each asset category to identify major themes (e.g., DDoS, Zeus, etc.).
- Six SFS students evaluated accuracy of LDA results and reached a Cronbach's alpha of 0.9393, indicating a high level of consistency.

AZSecure Hacker Assets Portal System Design and Features

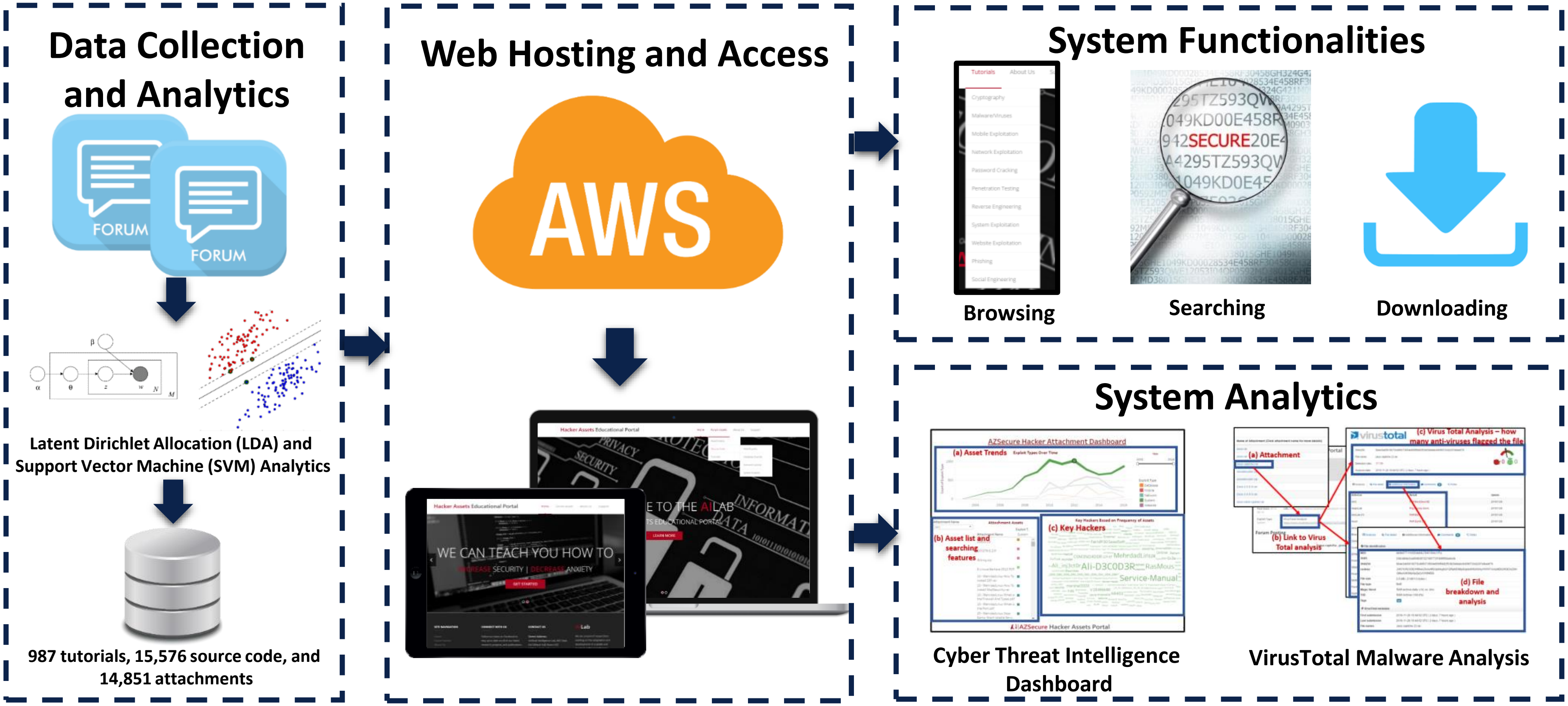


Figure 5. AZSecure Hacker Assets Portal System Design and Features

Tutorial Data Collection Summary

Tutorial Category	Count	Examples of Content
Website Exploitation	348	SQL Injection, XSS attacks
System Exploitation	230	BIOS hacking, rootkit creation, shellcode, spoofing files
Carding	201	Carding, bank hacking
Network Exploitation	112	Nmap scanning, Wireshark, DDoS
Password Cracking	43	Bruteforcing, password cracking approaches
Malware/Viruses	22	Malware analysis, detecting malware
Penetration Testing	13	Metasploit trainings, Google hacking
Mobile Exploitation	8	Android Malware
Cryptography	4	Basics of cryptography
Reverse Engineering	2	Basics of reverse engineering
Social Engineering	2	Social engineering psychology
Phishing	2	Basics of phishing

Table 2. Summary of Tutorial Data Collection Content

- Tutorials can provide the most direct cybersecurity education.
- We currently have 987 tutorials in 11 categories (Table 2).
- Tutorials teach various topics including:
  - Carding
  - SQL injections
  - Password cracking
  - Creating Android malware
  - Phishing
  - DDoS

Code and Attachment Data Collection Summary

Asset Type	Exploit Type	Count	Examples of Exploits
Source Code	System	9,746	Crypters, shellcode, DLL injections, Remote Administration Tools (RATs)
	Website	5,598	Content management system (CMS) exploits, SQL Injections
	Network	232	Bots/botnet/DDoS
Attachments	System	7,935	Zeus Malware, RATs, binders, Crypters, keyloggers
	Website	3,112	Cross-site scripting (XSS), website backdoors, website defacing, phishing
	Network	2,555	Bots/botnet/DDoS, firewall exploits, flooders
	Database	1,039	SQL payloads, dumpers, SQLmap
	Mobile	210	Android dumpers, crackers, malware, and pentests

Table 3. Summary of Source Code and Attachment Collection Content

- Students can use code and attachment assets to understand how tools are created, implemented, and operated.
- Code assets include:
  - Crypters
  - DLL injections
  - DDoS
- Attachments contain exploits such as:
  - Zeus
  - Android malware
  - Remote administration tools
  - Botnets
  - Keyloggers

Cyber Threat Intelligence (CTI) Applications

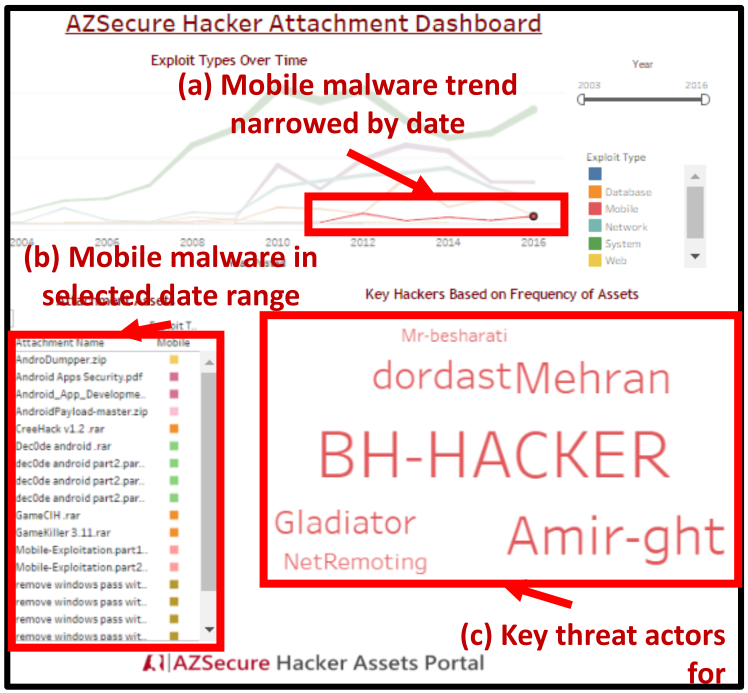


Figure 6. (a) Selecting a specific time range for mobile malware, (b) list of mobile malware in selected range, and (c) key threat actors for selected malware.

- For each asset, we detail which category of cyber-asset it targets (e.g., database, web, etc.) and where, when, who posted the asset in a CTI dashboard (Figure 6).
- Assuming an organization understands their own systems, hacker assets can create proactive CTI inform future cyber-defenses.
- For example, an organization can improve mobile device security given the recent increase in mobile malware.
  - Can also identify key threat actors to monitor

Please access at: http://www.azsecure-hap.com/

OR

Contact Sagar Samtani at sagars@email.arizona.edu

New users will need to enter their name, organization, position, and intended use to gain portal access.

We will then evaluate and confirm portal access.

AZSecure Hacker Underground Economy Collection and Analytics

Weifeng Li, Hsinchun Chen

Artificial Intelligence Lab, The University of Arizona

March 31st, 2017

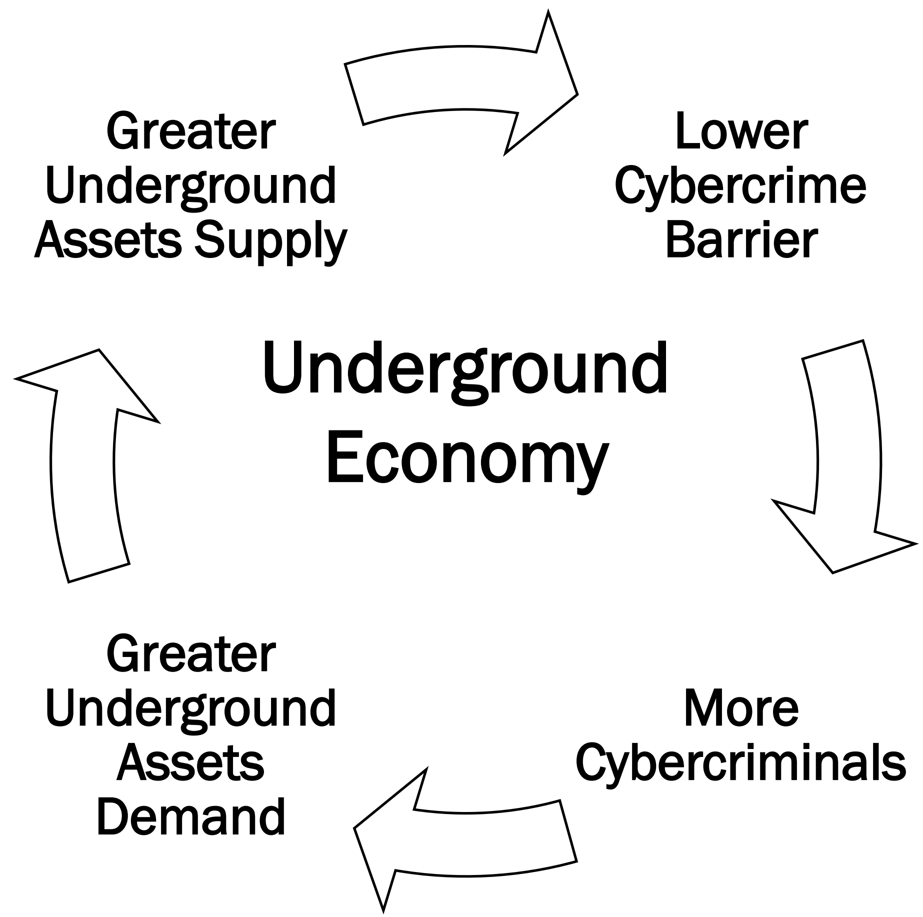
*This work is supported by the National Science Foundation under Grant DUE-1303362 and SES-1314631

Hacker Underground Economy

- International online black markets for hacking services and tools
- Provides comprehensive support for conducting data breach crimes:

  2013: Target;
  2014: Home Depot, Chase;
  2015: Anthem;
  2016: Yahoo

- Common platforms: hacker forums, DarkNet marketplaces, carding shops



UPDATE -- Now of

Devices

Note - This Devices Can Bring in A serious cash flow
Possible and realistic :cool:

Sell skimmer, pinpad+camera

Hello,
Price for GSM skimmer = 2 000 USD
+ pinpad (500 USD) + mini-camera (50$) = 2 550 USD (pm or btc)
Decode wav(audio)=Track 2 , track 1\3
Worldwide delivery via DHL
Escrow accepted

ATM Skimmer

Accessories

#1

...d to serve all your ...eeds with JAMES BOND ...ervices. ...verified since 2012!

...panel
...astics

Features

Orbitall EMV encoder -

EMV encoder

ORBITALL v2.3

ORBITALL v2.3
Corporate ©2016

Configuraçao / Configuration
Trilha / Track :

Seleção / Selection
Credito / Credit [ITAU] [SANTA]
Debito / Debit
Credito 128k / Credit 128k
Ticket / Voucher

Gerar / Generate

Limpar / Clean

Orbitall v2.3 - Corporate ©2016

sendspace.com/f

Good luck !!!

...old in
...atch

...ank Credit/Debit Cards (Plastics)

Collection Summary

Forums

# Forums	# Products	# Users	# Topics	Languages
34	169,009	257,183	414,530	English/Russian/Arabic

DarkNet Marketplaces

# Markets	# Products	# Users	# Reviews	Languages
9	80,590	5,528	690,411	English/Russian/Dutch

Carding Shops

# Shops	# Listings (cards, SSNs)	# Users	Languages
21	1,401,708	N/A	English/Russian

In Total:

# Platforms	# Products	# Participants	Languages
64	1,651,307	262,711	English/Russian/ Arabic/Dutch

* Hacker assets include but are not limited to: malware(encrypter/ransomware, Trojan, exploit), zero-day vulnerabilities, POS/ATM skimmer, stolen credit/ debit card, fake documents (driver's license, SSN), etc.

Sample Stolen Data in Collection

- Dump: magnetic strip information (for fraud purchase in store)

Card Number

Victim

Expiration

B5177220046948089^FARAH TUKAN (Track 1)
^1504101100000000000000149 (Track 2)
5177220046948089=150410110000149 (Track 3)

Ramification:
Fraudulent purchase in store

- CC/CVV: magnetic strip information

Card Number: 4266841209090735
Expire Date: 01/2012
CVV: 131
Cardholder Name: Walter Leger
Address: 4701 Rue Laurent
City: Metairie – LA – 70002

Ramification:
Fraudulent purchase online

- Fullz: magnetic strip information

Driver's License

James|Gayner|28540 Doyle Creek Rd.|Saint Marys|KS|66536|785-437-2803|362-82-4079|k00073521|KS|03-17-1967|JPGayner@yahoo.com|1ps72bn93d

SSN

Ramification:
Fraudulent loan application/tax return

- Health insurance records

Health Insurance

PRIMARY, Cigna Healthcare,U04197556,2461898, SECONDARY, UGA Athletic Dept.,254718352,,650 West Conway Dr,,27y,4/8/1989,,Atlanta,, MARGARET,A,MCWHIRTER,,(404)401-3108,,F,254-71-8352,GA,30327

Ramification:
Fraudulent healthcare claim

Summary of Major Data Breach Services in Collection

Category	Service	Examples	Price
Infrastructure	Malware	POS malware; ATM skimmers	$300~5000
	Phishing	Phishing emails; scam sites	$2.5~100/wk
	Botnets	Hosting relays for stolen cards	$2~60/hr
Data	Payment Cards	Dumps; CC/CVV	$0.1~25
	Identities (Fullz)	Social Security Numbers; driver's license; insurance cards	$1~260
	Credentials	Bank accounts; Paypal accounts	$1~300
Cashing	Forging	Blank credit cards; driver's license template	$40~110
	Change of Billing (COB)	Change of billing address for carders to make purchases	$35~140
	Drop	Location carders can have illicitly purchased goods sent to	~50% Royalty

Table 1. Common Hacking Services and Their Prices

Analytics: Key Seller Identification

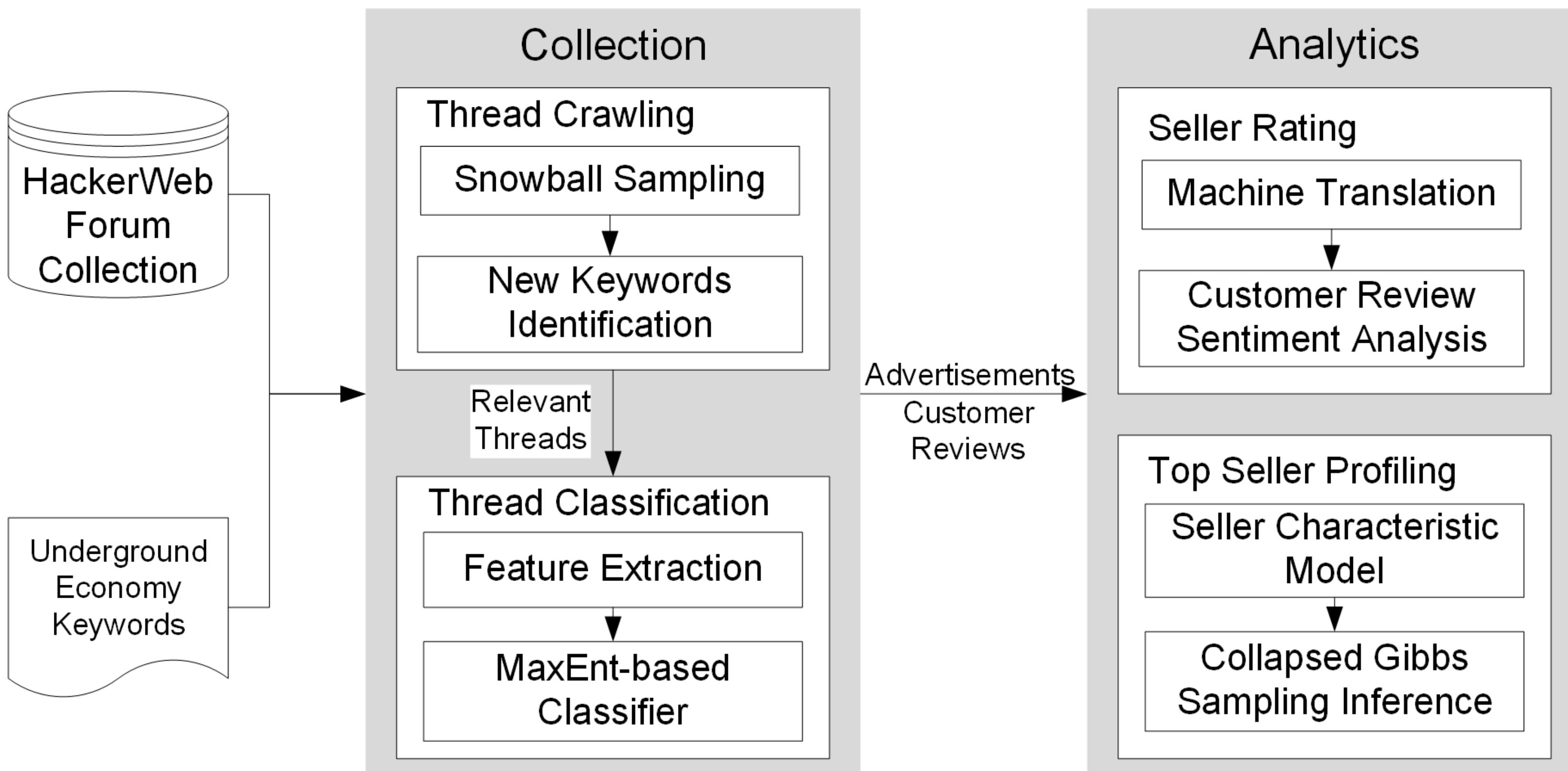
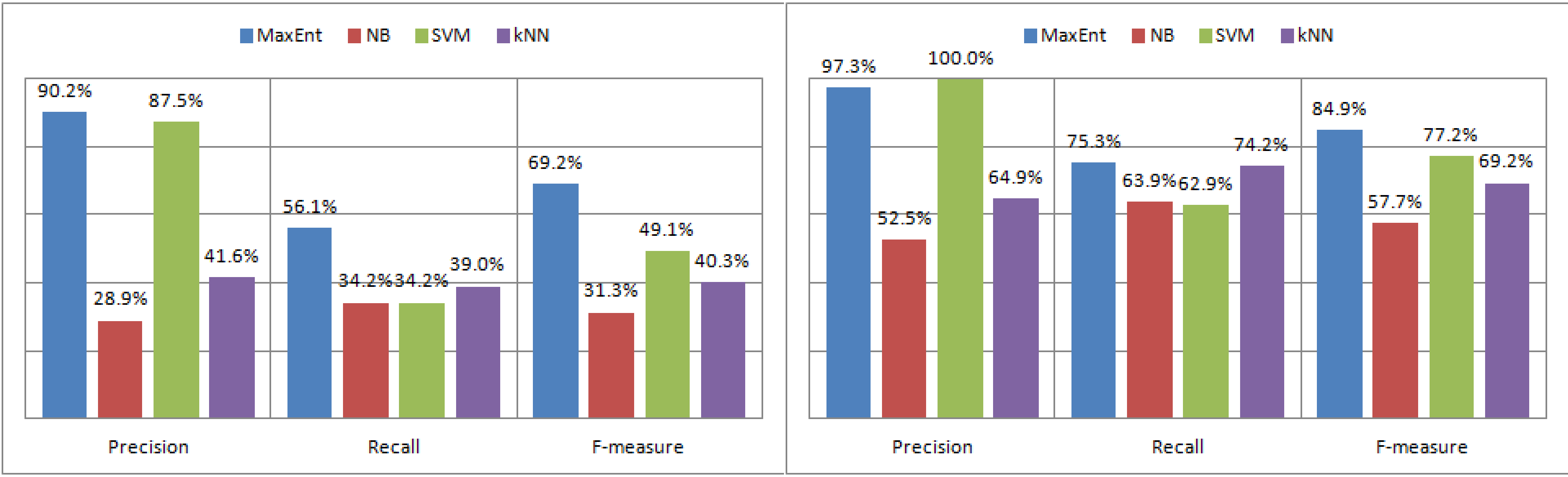
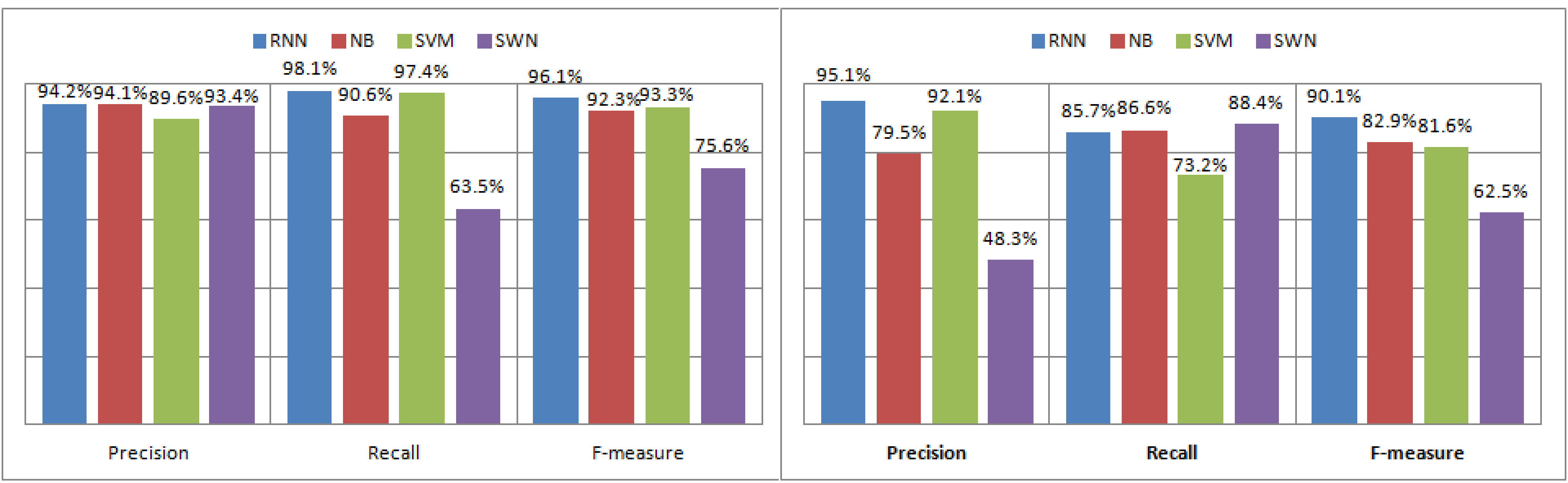


Figure 1. The AZSecure Key Seller Identification Framework



(a) Malware Advertisements (b) Stolen Data Advertisements

Figure 2. Thread Classification Performance



(a) Positive Sentiment (b) Negative Sentiment

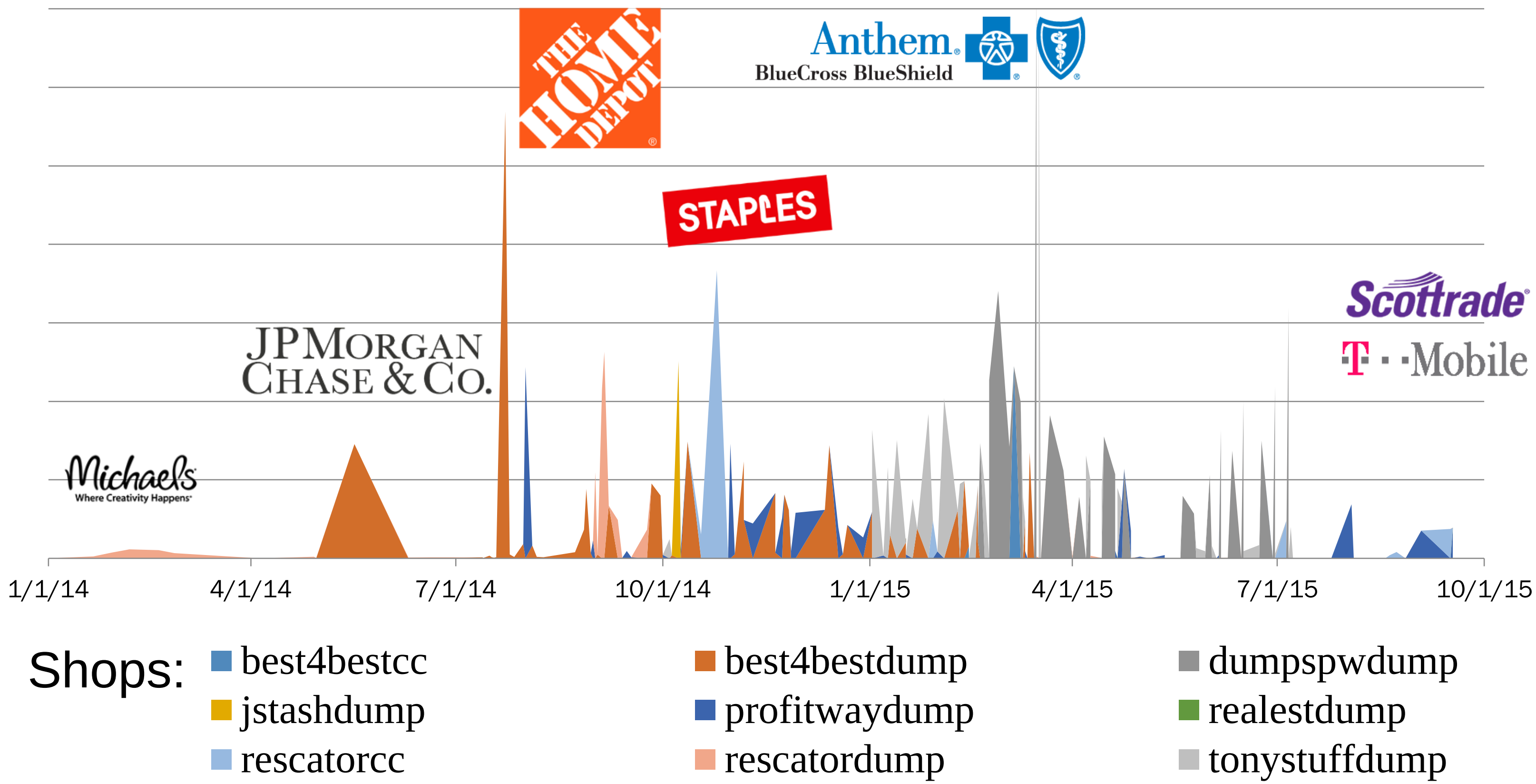
Figure 3. Seller Rating Performance

Who are the key sellers?

	Top 3 Best				Top 3 Worst			
	Infrastructure		Data		Infrastructure		Data	
Rank	Seller	Score	User	Score	User	Score	User	Score
	Antichat							
1	LEOnidUKG	5	inferno[DGT]	3.6	@NoFrag@	1.8	Isis	2.3
2	VISMU	4.5	amorales	3.5	k0lbasa	2	PEPSICOLA	2.3
3	gogol	4	«DEXTER»	3.4	Doktor_radosti	2	sultan128	2.4
	CrdPro							
1	HackingAll	4	**Rescator**	**4.4**	N1K70	2	finu2004	1.3
2	balt	4	Faaxxx	4	1vanu4	2	sonny13	1.3
3	SunSeller	4	ResellerInc	4	MID	2	l33tsu	1.3
	Zloy							
1	PerfectCrypt	5	BigBuyer	4	root	1.5	riodetray	1
2	DiXakMan	4	Buyers11	4	w370w370	1.6	madman	1
3	DjVellf	4	sellcc	4	gorpen	2	jekaa	2

Table 2. Top 3 Best/Worst Malware and Stolen Data Sellers for Each Forum

Key Sellers' Collection v.s. Data Breach Events



- Before the announcements of data breach events, bulks of breached data already appear in key sellers' shops for sale.

How much is each card worth?

Card Features: Brand, Type, Mark, Bank	Price Difference
Visa Electron Card	$21.06
American Express Card	$19.41
World Elite MasterCard for Business	$17.11
Corporate Purchasing Card	$13.15
Base Price	**$7.48**
Bank of America Card	$2.63
Debit Card	$2.00
Credit Card	$1.72
J.P. Morgan Chase Card	–$1.78
Standard Card	–$2.43
Prepaid Card	–$4.72
Visa/MasterCard Classic Card	–$4.77

Table 3. Selected Card Features Affecting Card Price in the Underground Economy

* Results were obtained using standard linear regression model with significance level of 0.001.