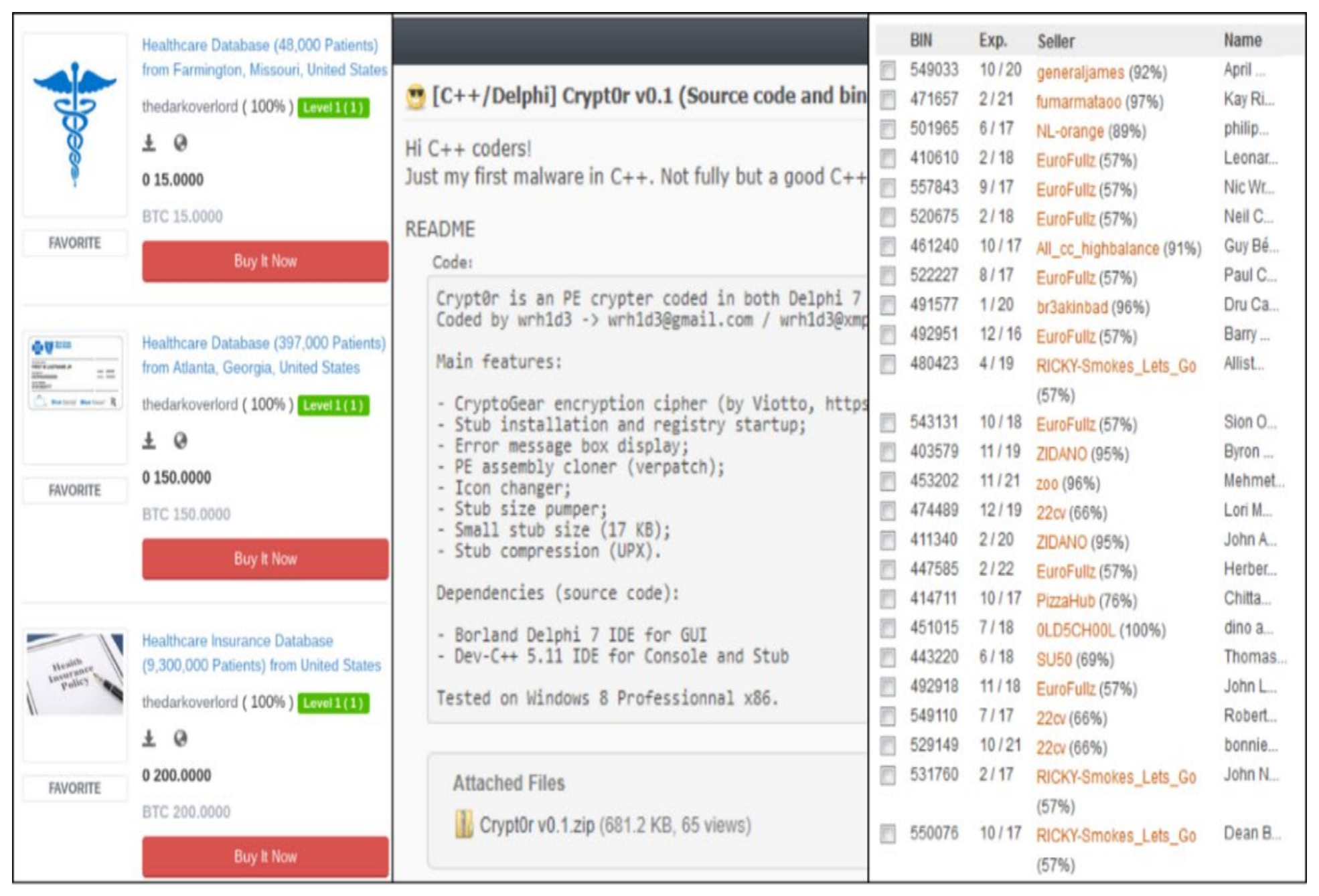
SaTC: CORE: Small: Cybersecurity Big Data Research for Hacker Community: A Topic and Language Modeling Approach

PI: Dr. Hsinchun Chen, Regents' Professor, ACM/IEEE Fellow, University of Arizona, AI Lab Director

Co-PI: Dr. Weifeng Li, University of Georgia

https://eller.arizona.edu/departments-research/centers-labs/artificial-intelligence/research/big-data



Healthcare Database (48,000 Patients) from Farmington, Missouri, United States
thedarkoverlord (100%) Level 1 (1)
0 15.0000
BTC 15.0000
FAVORITE
Buy It Now

Healthcare Database (397,000 Patients) from Atlanta, Georgia, United States
thedarkoverlord (100%) Level 1 (1)
0 150.0000
BTC 150.0000
FAVORITE
Buy It Now

Healthcare Insurance Database (9,300,000 Patients) from United States
thedarkoverlord (100%) Level 1 (1)
0 200.0000
BTC 200.0000
FAVORITE
Buy It Now

[C++/Delphi] Crypt0r v0.1 (Source code and bin

Hi C++ coders!
Just my first malware in C++. Not fully but a good C++

README

Code:

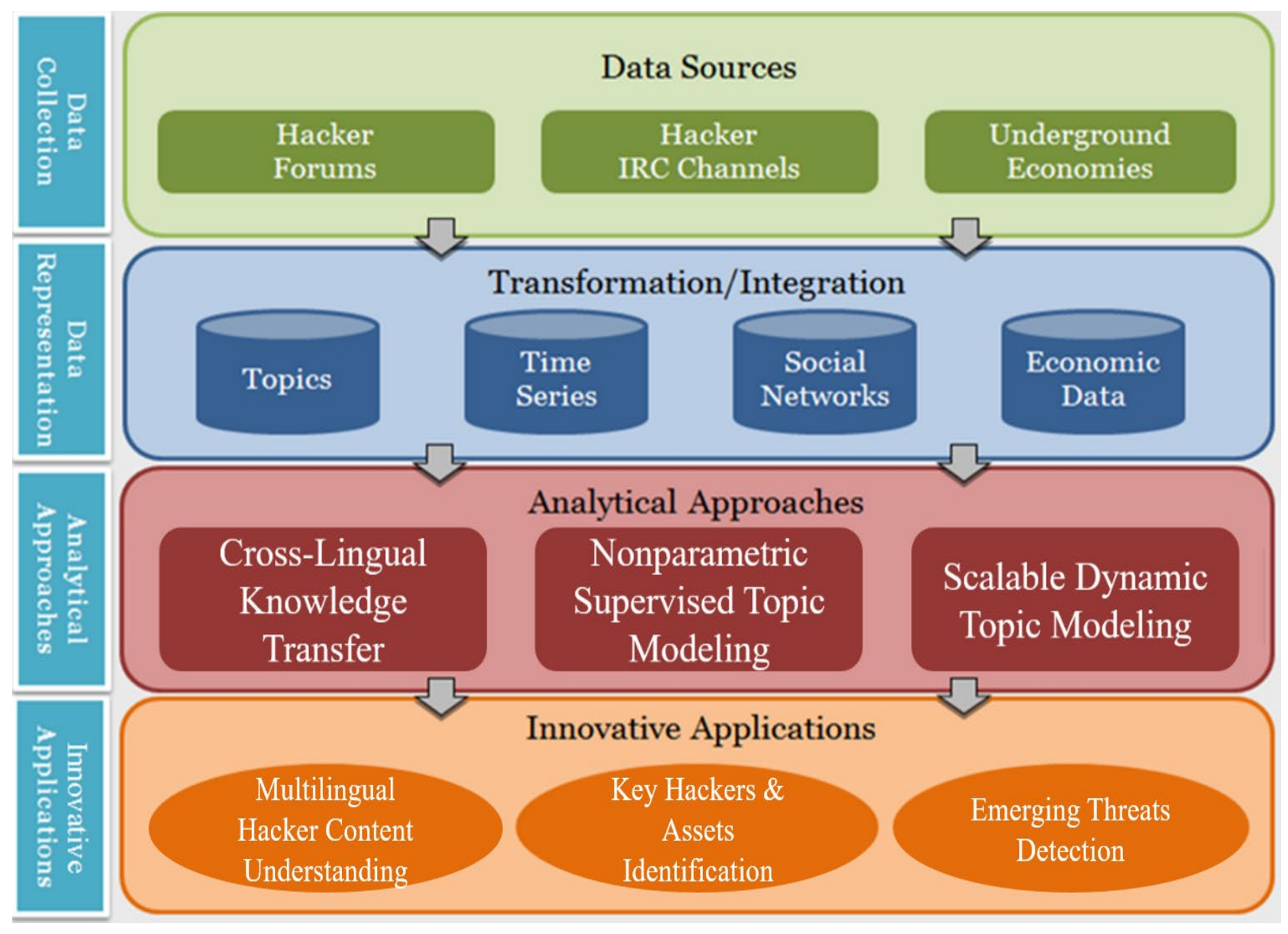
```
Crypt0r is an PE crypter coded in both Delphi 7
Coded by wrh1d3 -> wrh1d3@gmail.com / wrh1d3@xmp

Main features:

- CryptoGear encryption cipher (by Viotto, https
- Stub installation and registry startup;
- Error message box display;
- PE assembly cloner (verpatch);
- Icon changer;
- Stub size pumper;
- Small stub size (17 KB);
- Stub compression (UPX).

Dependencies (source code):

- Borland Delphi 7 IDE for GUI
- Dev-C++ 5.11 IDE for Console and Stub

Tested on Windows 8 Professionnal x86.
```

Attached Files
Crypt0r v0.1.zip (681.2 KB, 65 views)

	BIN	Exp.	Seller	Name
□	549033	10 / 20	generaljames (92%)	April ...
□	471657	2 / 21	fumarmataoo (97%)	Kay Ri...
□	501965	6 / 17	NL-orange (89%)	philip...
□	410610	2 / 18	EuroFullz (57%)	Leonar...
□	557843	9 / 17	EuroFullz (57%)	Nic Wr...
□	520675	2 / 18	EuroFullz (57%)	Neil C...
□	461240	10 / 17	All_cc_highbalance (91%)	Guy Bé...
□	522227	8 / 17	EuroFullz (57%)	Paul C...
□	491577	1 / 20	br3akinbad (96%)	Dru Ca...
□	492951	12 / 16	EuroFullz (57%)	Barry ...
□	480423	4 / 19	RICKY-Smokes_Lets_Go (57%)	Allist...
□	543131	10 / 18	EuroFullz (57%)	Sion O...
□	403579	11 / 19	ZIDANO (95%)	Byron ...
□	453202	11 / 21	zoo (96%)	Mehmet...
□	474489	12 / 19	22cv (66%)	Lori M...
□	411340	2 / 20	ZIDANO (95%)	John A...
□	447585	2 / 22	EuroFullz (57%)	Herber...
□	414711	10 / 17	PizzaHub (76%)	Chitta...
□	451015	7 / 18	0LD5CH00L (100%)	dino a...
□	443220	6 / 18	SU50 (69%)	Thomas...
□	492918	11 / 18	EuroFullz (57%)	John L...
□	549110	7 / 17	22cv (66%)	Robert...
□	529149	10 / 21	22cv (66%)	bonnie...
□	531760	2 / 17	RICKY-Smokes_Lets_Go (57%)	John N...
□	550076	10 / 17	RICKY-Smokes_Lets_Go (57%)	Dean B...

Research Background

- Conventional Cyber Threat Intelligence (CTI) that analyzes historical attack data (e.g., network logs, honeypots, system logins, and IDS/IPS events) suffers two major limitations:
  (1) Focusing on *reactive* intelligence from previous attacks rather than *proactive* intelligence about threats that have the potential to cause damage (e.g., zero-day attacks) but have not yet been used for cyberattacks,
  (2) Ignoring the threat actors responsible for the attacks, missing the full picture of the hacker ecosystem (e.g., communities, specialties, tools, etc.)
- Hacker communities (e.g., forums, Internet-Relay-Chat (IRC) channels, and underground economies) are of particular interest as they allow hackers to share malicious assets such as hacking tools, malware source code, and hacking tutorials with one another.
- However, several technical difficulties in hacker community data collection and analytics exist:
  (1) The massive volume of data collection
  (2) The heterogeneity and covert nature of the data elements and their often not so obvious linkages
  (3) The ability to comprehend common hacker terms and concepts embedded in communities across multiple regions



Key Challenges to be Addressed

This project aims to address two research goals:

1) to advance current capabilities for scalable identification, collection, and analysis of international hacker community contents
2) to make contributions to the cybersecurity community by developing new big data techniques that could enable security researchers to conduct analyses on hacker content and within other related domains.

To address these research goals, this project features four major research projects:

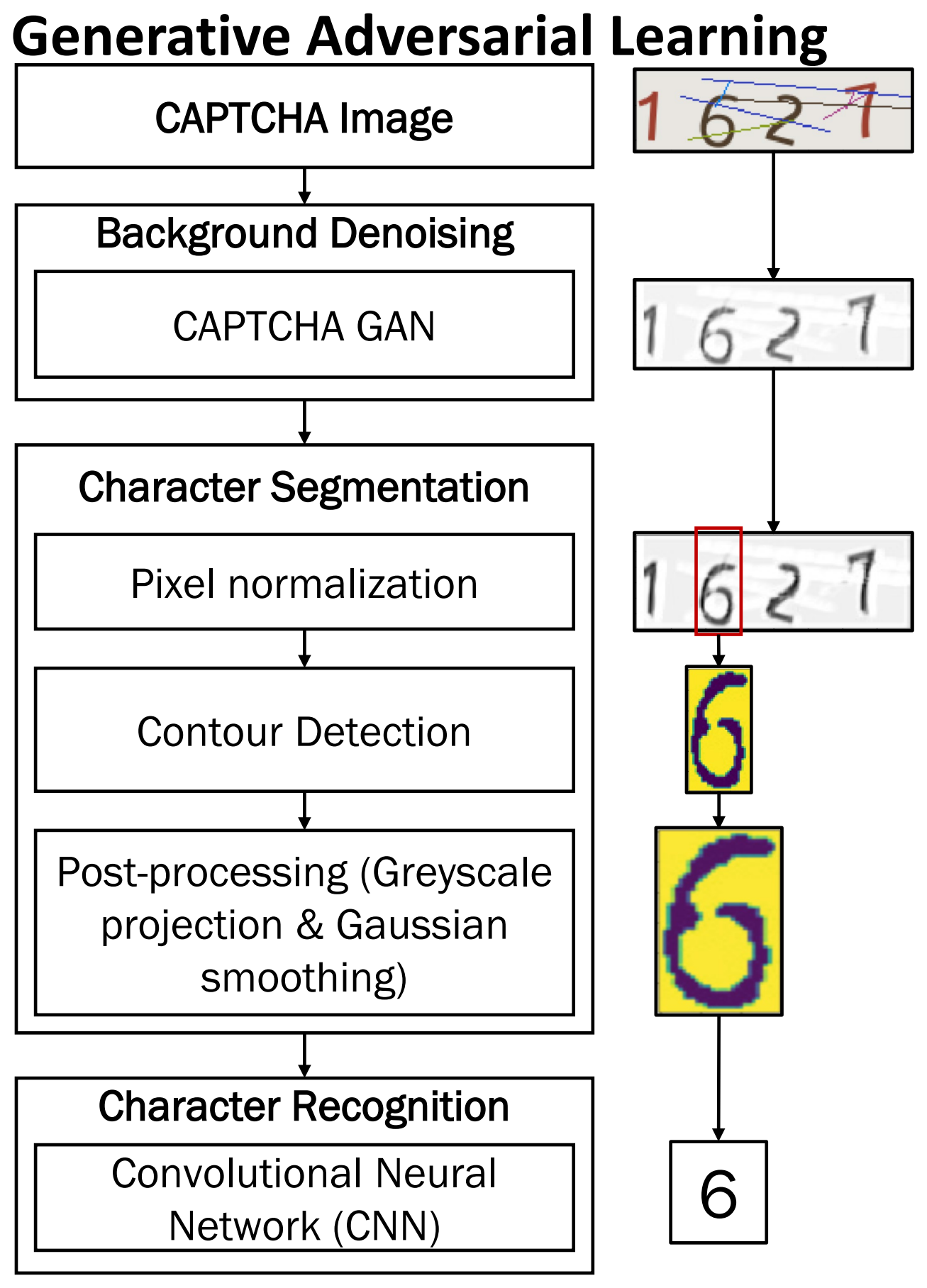
1. to develop a large, comprehensive, and longitudinal testbed of international online hacker community contents, including forums, IRCs, underground economies, and other emerging hacker assets
2. to develop an automated method for understanding multilingual hacker community contents to facilitate multilingual CTI analytics
3. to develop a scalable framework for the identification of key hacker assets from the massive underground economy
4. to develop novel dynamic topic modeling capabilities to detect emerging topics from hacker communities, and
5. to address the emerging threat of leaked PII records circulating the hacker community through developing deep learning techniques for privacy risk assessment and a web portal for the general public to get educated and take preventive measures.

Scientific Impact

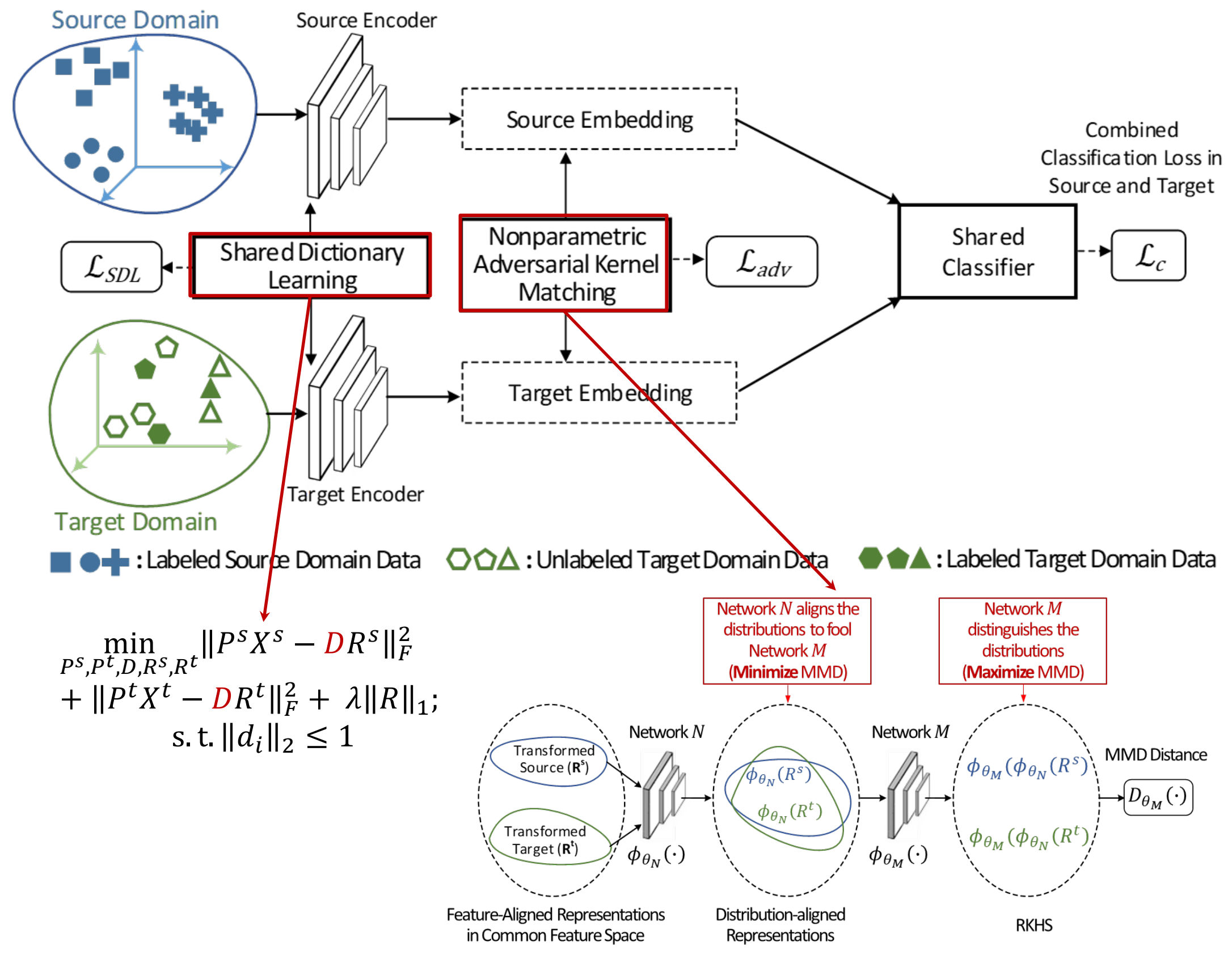
- Methodological contributions in hacker content analysis to assist cybersecurity practitioners in studying hacker communication and developing proactive CTI from critical hacker communities
  - Exploration of hacker languages, topics, assets, and threats across different international hacker communities
- A rich, large-scale, longitudinal, open-source collection of diverse hacker community field data to support timely and data-driven cybersecurity and criminology exploration and hypothesis testing
  - Increasing the number of individuals working on CTI research, and subsequently advancing important and complex hacker community CTI inquiries
- Advancing computational knowledge and capabilities in deep transfer learning, deep generative modeling, supervised topic modeling, dynamic topic modeling, neural variational inference, and numerous other important domains.
- Applicability to other big data problems, especially in the social media context:
  - Community identification and collection procedures: applicable to other studies seeking to collect deep web data, especially from sources that employ anti-crawling mechanisms (common among DarkNet markets.)
  - Developed methodologies: applicable to analyzing multilingual community data in other contexts, especially for other legitimate electronic networks-of-practice where tutorials, source code, and other tools may be shared.

Selected Technical Approaches:

1. Hacker Community Collection: Counteracting Dark Web CAPTCHA with Generative Adversarial Learning



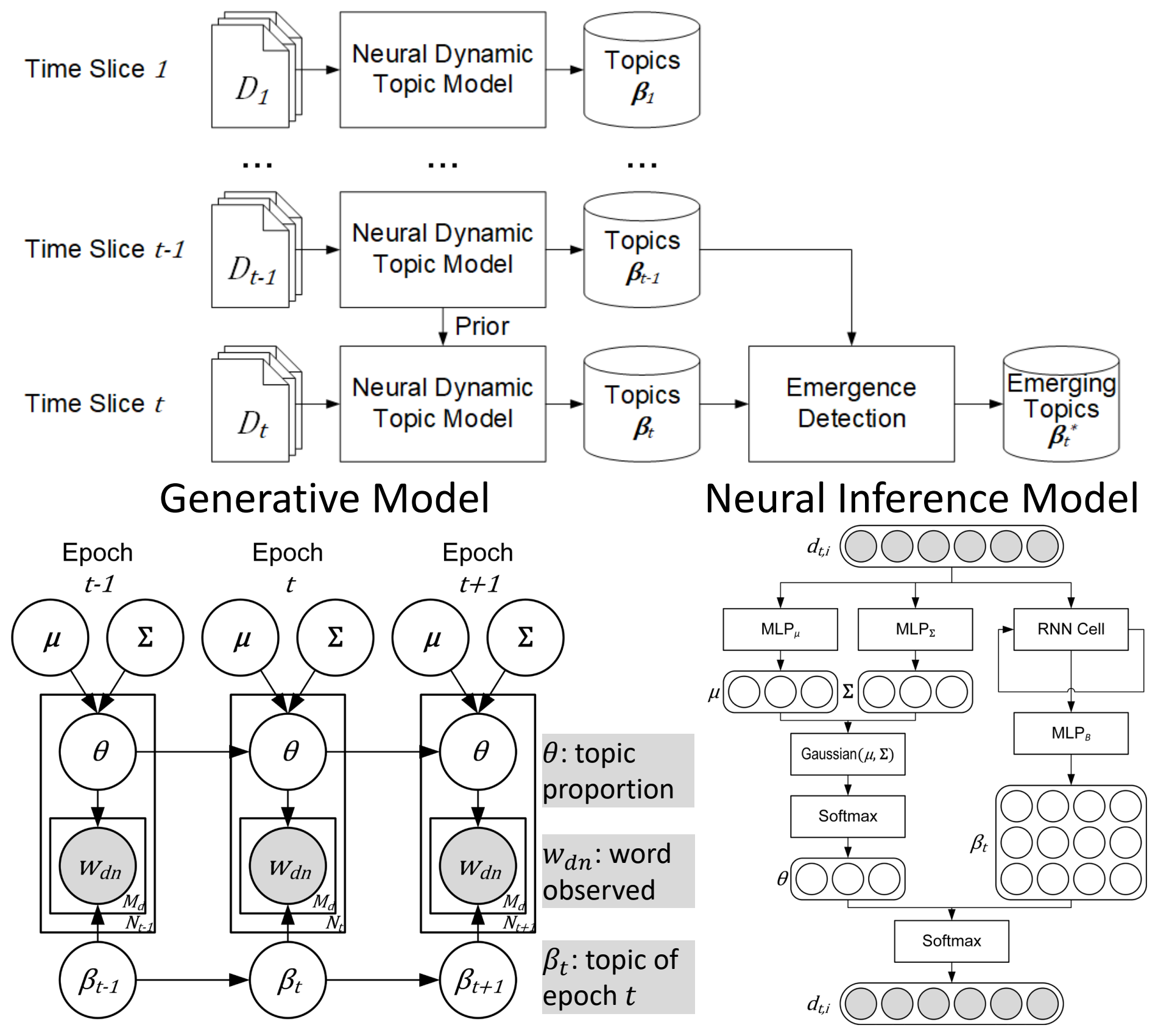
2. Cross-lingual Analytics: Heterogeneous Domain Adaptation with Adversarial Neural Representation Learning



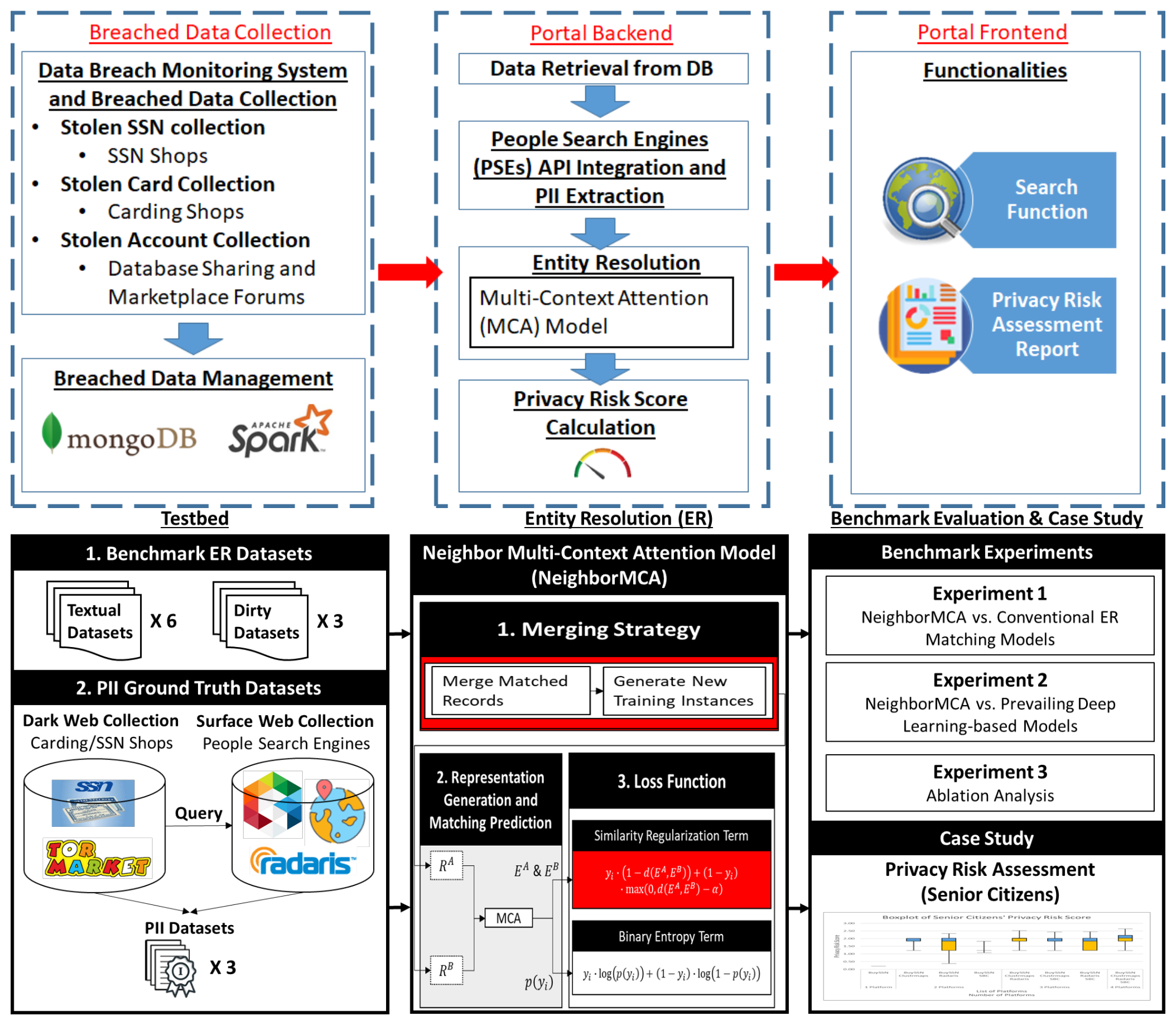
Source Domain
Source Encoder
Source Embedding
Combined Classification Loss in Source and Target
$\mathcal{L}_{SDL}$
Shared Dictionary Learning
Nonparametric Adversarial Kernel Matching
$\mathcal{L}_{adv}$
Shared Classifier
$\mathcal{L}_c$
Target Embedding
Target Encoder
Target Domain
■ ●✚ : Labeled Source Domain Data
⬡⬠△ : Unlabeled Target Domain Data
⬢⬟▲ : Labeled Target Domain Data
Network N aligns the distributions to fool Network M (Minimize MMD)
Network M distinguishes the distributions (Maximize MMD)
Transformed Source (R^s)
Transformed Target (R^t)
Network N
$\phi_{\theta_N}(\cdot)$
$\phi_{\theta_N}(R^s)$
$\phi_{\theta_N}(R^t)$
Network M
$\phi_{\theta_M}(\cdot)$
$\phi_{\theta_M}(\phi_{\theta_N}(R^s))$
$\phi_{\theta_M}(\phi_{\theta_N}(R^t))$
MMD Distance
$D_{\theta_M}(\cdot)$
Feature-Aligned Representations in Common Feature Space
Distribution-aligned Representations
RKHS

$$\min_{P^s,P^t,D,R^s,R^t} \|P^sX^s - DR^s\|_F^2 + \|P^tX^t - DR^t\|_F^2 + \lambda\|R\|_1; \quad \text{s.t.}\ \|d_i\|_2 \le 1$$

3. Emerging Topic/Threat Detection: Dynamic Topic Modeling with Neural Variational Inference



4. Hacker Community Breached Data Analytics: AZSecure Privacy Analytics



Breached Data Collection
Data Breach Monitoring System and Breached Data Collection
Stolen SSN collection
SSN Shops
Stolen Card Collection
Carding Shops
Stolen Account Collection
Database Sharing and Marketplace Forums
Breached Data Management
mongoDB
Spark
Portal Backend
Data Retrieval from DB
People Search Engines (PSEs) API Integration and PII Extraction
Entity Resolution
Multi-Context Attention (MCA) Model
Privacy Risk Score Calculation
Portal Frontend
Functionalities
Search Function
Privacy Risk Assessment Report
Testbed
1. Benchmark ER Datasets
Textual Datasets X 6
Dirty Datasets X 3
2. PII Ground Truth Datasets
Dark Web Collection
Carding/SSN Shops
Surface Web Collection
People Search Engines
Query
radaris
PII Datasets
X 3
Entity Resolution (ER)
Neighbor Multi-Context Attention Model (NeighborMCA)
1. Merging Strategy
Merge Matched Records
Generate New Training Instances
2. Representation Generation and Matching Prediction
R^A
R^B
E^A & E^B
MCA
$p(y_i)$
3. Loss Function
Similarity Regularization Term
$y_i \cdot (1 - d(E^A,E^B)) + (1 - y_i) \cdot \max(0, d(E^A,E^B) - \alpha)$
Binary Entropy Term
$y_i \cdot \log(p(y_i)) + (1 - y_i) \cdot \log(1 - p(y_i))$
Benchmark Evaluation & Case Study
Benchmark Experiments
Experiment 1
NeighborMCA vs. Conventional ER Matching Models
Experiment 2
NeighborMCA vs. Prevailing Deep Learning-based Models
Experiment 3
Ablation Analysis
Case Study
Privacy Risk Assessment (Senior Citizens)
Boxplot of Senior Citizens' Privacy Risk Score

Broader Impact on Society

Practitioner communities

- Government agencies (e.g., US Postal Inspection Service and Canada Revenue Agency)
- Law enforcement (e.g., US Secret Service)
- Criminologists (e.g., The Society for the Policing of Cyberspace (POLCYB))
- Intelligence analysts (e.g., National Cyber-Forensics & Training Alliance (NCFTA))

Dissemination of Collected Data

- https://www.azsecure-data.org/, our sustainable platform for retaining and sharing our selected data collections and analytical approaches (facilitated by our NSF-funded DIBBs project)

Dissemination of Research Findings

- NCFTA and POLCYB: our primary domain advisory and end-user feedback groups
- Periodical knowledge sharing with practitioner community facilitated by NCFTA and POLCYB

Broader Impact on Education and Outreach

Education at University of Arizona

- Top-ranked MS in Cybersecurity program: integrating selected results into the Cyber Warfare class
- AZSecure Scholarship-for-Service (SFS) program: preparing master's and Ph.D. students with cybersecurity research experience as per the independent studies requirement of the SFS curriculum

Education at University of Georgia

- Area of Emphasis on Information Security: integrating research findings from selected projects into the Cyber Threat Intelligence class (~300 students over three years)

Outreach

- IEEE ICDM Deep Learning for Cyber Threat Intelligence (DL-CTI) Workshop
- ACM KDD Workshop on AI-enabled Cybersecurity Analytics
- IEEE TDSC Special Issue on Explainable AI for Cyber Threat Intelligence (XAI-CTI)
- IEEE ICDM Machine Learning for Cybersecurity (MLC) Workshop

Broader Impact and Broader Participation

Student Training

- University of Arizona: six graduate students and three undergraduate students trained in hacker community collection and analysis; contributed to the development of tools and tutorials
- University of Georgia: two undergraduate students trained in data collection

Professional Development:

- Data collection (e.g., DBMS, modern programming languages)
- Infrastructure design (e.g., virtual machine management) and development
- Project management and team membership
- Assisting faculty with the development of practical, hands-on cybersecurity-related activities
- Contributing their knowledge about cutting-edge research to enhance the course curriculum
- Opportunity to interact with faculty and others in the field

Award ID#: CNS-1936370