Curriculum Development for Cyber Threats and Vulnerabilities using Big Data and Secure Analytics

Dr. Latifur Khan (PI), Dr. Bhavani Thuraisingham (Co-PI)

Department of Computer Science, The University of Texas at Dallas, Richardson, Texas, USA
Contact: lkhan@utdallas.edu

1. Problem Statement

- ➢ Develop a syllabus to study cyber threats and vulnerabilities using big data and secure analytics.
- ➢ Develop course materials, instructor notes, interactive videos and lab exercises that will inform and train students in applying learned skills to cyber defense and trusted analytics.
- ➢ Leverage current existing research work in big data, cyber security and data processing using trusted hardware extensions (TEE).

2. Course Materials

Developed:

- Machine Learning and Analytics
- Big Data Analytics
- Secure Big Data Analytics

In progress:

- Cyber Threats and Vulnerability Data Processing
- Actionable Insight from Cyber Threat and Vulnerable Data
- High Level knowledge Extraction and Kill Chain Inference

Machine Learning and Analytics

- ➢ Module focuses on different machine learning algorithms so that students get an understanding of the emerging techniques.
- ➢ Contains three set of PowerPoint slides for the lessons describing text mining, text feature extraction techniques and some efficient classification algorithms such as Support Vector Machines (SVM) and k-Nearest Neighbor Methods (KNN).
- ➢ Contains some lab tasks that give empirical knowledge on different classification and feature extraction processes.

Big Data Analytics

- ➢ Focuses on data mining & machine learning algorithms for analyzing very large amount of data with the help of tools/frameworks such as Hadoop MapReduce and Spark.
- ➢ Contains four modules:
  - Hadoop MapReduce Basics: contains three set of PowerPoint slides for the lessons describing the functionality and mechanisms of MapReduce framework.
  - Hadoop Setup and Programming: contains two set of PowerPoint slides explaining how to setup and program in MapReduce framework and some lab tasks that teach techniques to analyze massive amounts of data using MapReduce.
  - Spark Basics: contains three set of PowerPoint slides for the lessons that discuss Spark, Spark SQL and it's advantage over Hadoop.
  - Spark Setup and Programming: contains two set of PowerPoint slides explaining how to setup and program in Spark framework and some lab tasks that show techniques for processing vast data using Spark.

Secure Big Data Analytics

- ➢ Module focuses on securing big data analytics by using the Intel Software Guard Extension (SGX) as a secure hardware solution.
- ➢ Contains the PowerPoint slides on the application of SGX in securing logging operations, the design of SGX-enabled log server, programs and protocols.
- ➢ Contains lab tasks that provide comprehensive knowledge to effectively run a new system designed to secure system logs using SGX-based log server.

3. Study Cyber Threat Reports and Vulnerabilities

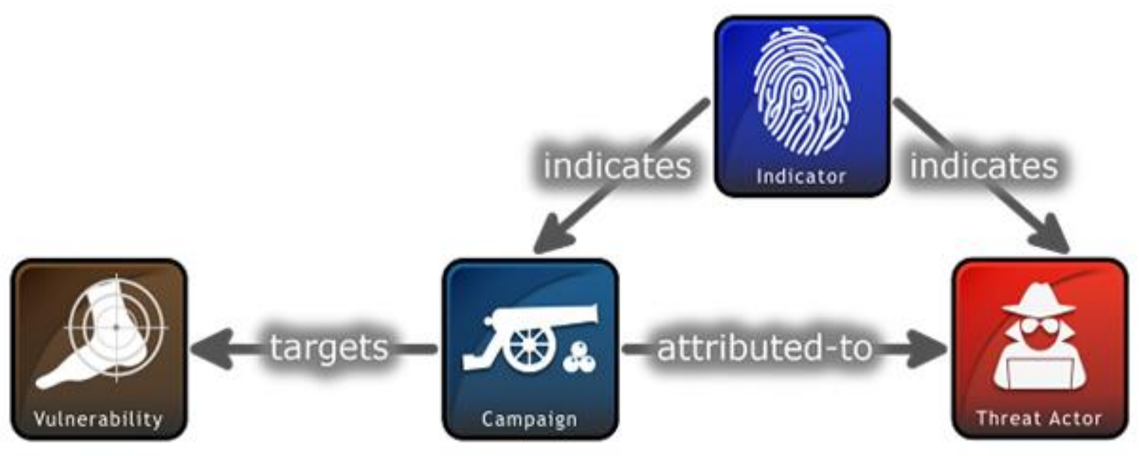
Learning Objectives

- ➢ Allow students to gain actionable insights from cyber threat and vulnerable data repositories.
- ➢ Help students gain hands-on experience using techniques from NLP and Big Data Analytics.
- ➢ Prepare and instruct students in state-of-the-art skills needed to defend against attack techniques and challenges in software security.

Sample Problem for Lab

Challenges:

- ➢ Computer Security Threat reports are in unstructured text format.
- ➢ Large scale generation of threat reports due to increase in computer security breaches.
- ➢ Requires human analyst to interpret reports and extract meaningful information.
- ➢ Limited training data due to limited labeled threat reports.
- ➢ Need for automated system in extracting information and classifying threat reports to tactics and techniques



Problem Setting:

- ➢ Given a threat report, obtain the tactics and techniques used by the attacker as described in the report.
- ➢ After classifying to tactics, students need to identify the techniques used out of the available techniques under the tactics.
- ➢ For Technique classification, follow Top-Down approach: classify to tactics and then classify to techniques.
- ➢ Requires pruning the techniques under the tactics to identify relevant techniques.

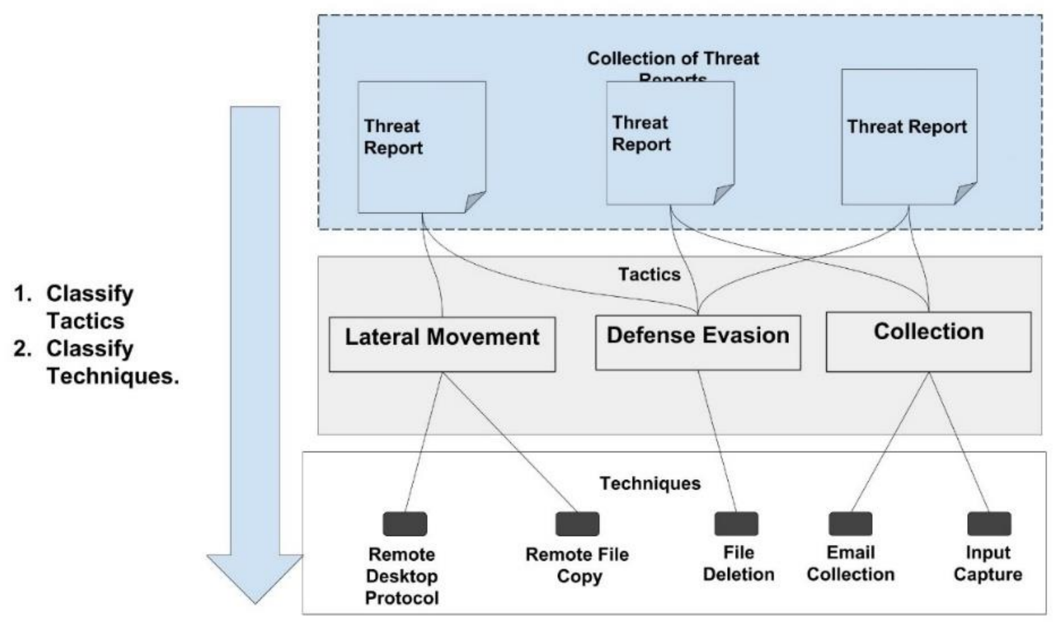
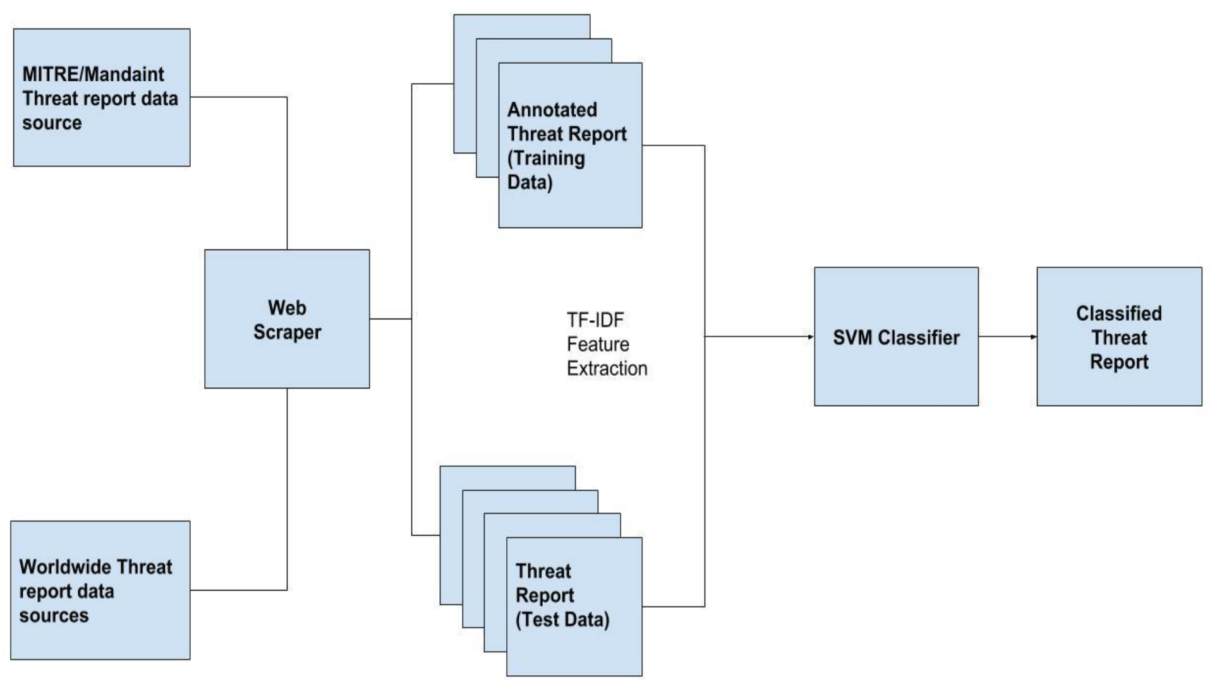


Figure: Technique Classification Top-Down Approach

Proposed Solution

- ➢ Download threat reports from different computer security vendors.
- ➢ Data Cleaning and Filtering to remove noise.
- ➢ Apply natural language processing techniques to extract features from unstructured text.
- ➢ Apply Classification method such as SVM classifier.


MITRE/Mandaint Threat report data source
Worldwide Threat report data sources
Web Scraper
Annotated Threat Report (Training Data)
Threat Report (Test Data)
TF-IDF Feature Extraction
SVM Classifier
Classified Threat Report

Data Set

Dataset	Average File Size	No. of Documents	Average Words per Document	Total Data Size
APT	50KB	488	10,000	1 GB
Symantec Data	5KB	17,600	1,000	147MB

Raw Data Sources:

APT: https://github.com/kbandla/APTnotes

Symantec Data: https://www.symantec.com/security-center/a-z

Packages/Tools

- ❑ Anaconda: Open source platform for python data science and machine learning
- ❑ gensim: Open-source vector space modeling and topic modeling toolkit
- ❑ nltk: Natural Language Toolkit written in python for NLP
- ❑ numpy: Python package for scientific computing
- ❑ scipy: Python package for scientific computing
- ❑ matplotlib: 2D plotting library for python
- ❑ jupyter: : Open source platform for data science and scientific computing that supports many programming languages

4. Materials Delivered So Far

Machine Learning and Analytics Module Outline

Lectures	No. of Slides	Labs	Quiz
1. Text Mining	26	2 Lab tasks	Assessment Quiz containing 13 questions
2. Text Feature Extraction	31	2 Lab tasks	
3. Classification Algorithms	47	1 Lab task	

Big Data Analytics Unit Outline

Module	Lectures		Labs/Quizzes
	Name	No. of Slides	
Hadoop MapReduce Basics	1. Introduction to Hadoop MapReduce	56	Assessment Quiz containing 20 questions
	2. MapReduce Algorithm Design	44	
	3. Processing Relational Data with MapReduce	27	
Hadoop Setup and Programming	1. Hadoop Setup	21	4 Lab tasks
	2. Hadoop Program	33	
Spark Basics	1. Introduction to Spark	52	Assessment Quiz containing 20 questions
	2. Programming with Spark	33	
	3. Spark SQL	16	
Spark Setup and Programming	1. Spark Setup	9	4 Lab tasks
	2. Spark Program	15	

Secure Big Data Analytics Module Outline

Lectures	No. of Slides	Labs	Quiz
1. Installing and Running	N/A	1 Lab task	Assessment Quiz containing 8 questions
2. Understanding Theory	87		
3. Programs and Protocols	27		