Securing systems with Chaos Engineering and Artificial Intelligence

Universidad del Rosario | MACC Matemáticas Aplicadas y Ciencias de la Computación

BSIDES Cybersecurity conference
Bogotá, Colombia, April 22nd ,2023

Martin Bedoya

Leader of Hacking and Appsec, NTT Data
MSc in Applied Mathematics and Computer Science (C),
University of Rosario

mbedoyar@emeal.nttdata.com
martin.bedoya@urosario.edu.co

Sara Palacios Chavarro

Cybersecurity Engineer, NTT Data
Professional in Applied maths and Computer Science,
University of Rosario

sara.palacioschavarro@nttdata.com
sara.palaciosc@urosario.edu.co

Daniel Díaz, PhD

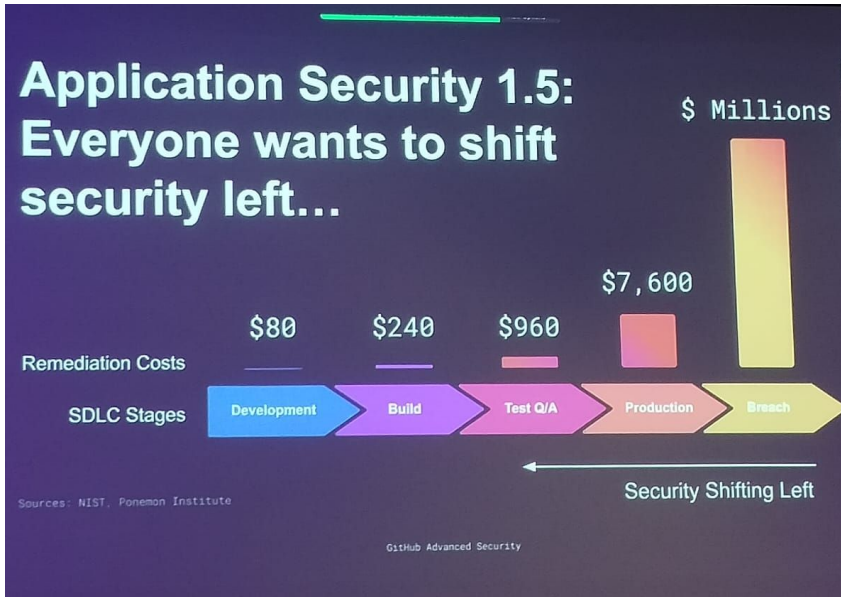
Principal Professor,
School of Engineering, Science and Technology, University of Rosario

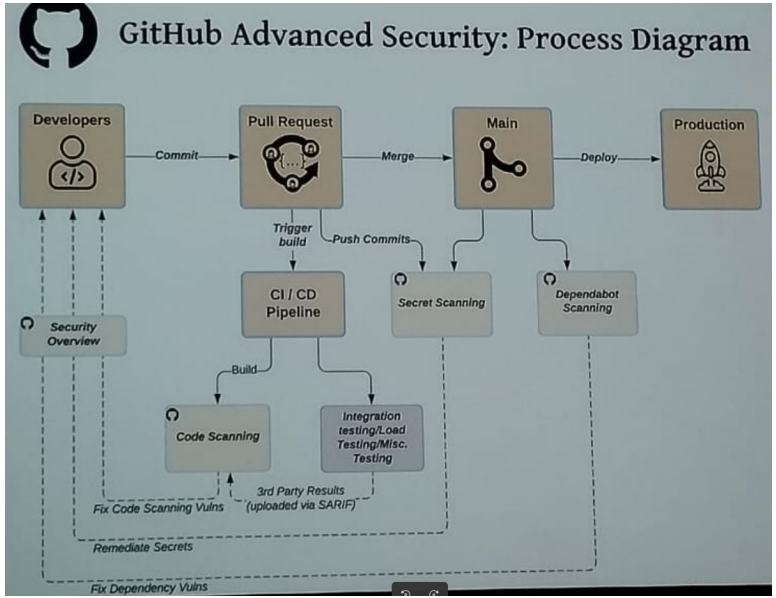
danielo.diaz@urosario.edu.co

The problem

Universidad del Rosario | MACC Matemáticas Aplicadas y Ciencias de la Computación

- Let’s remember the talk “Codificando Seguro” of Michael Cantu



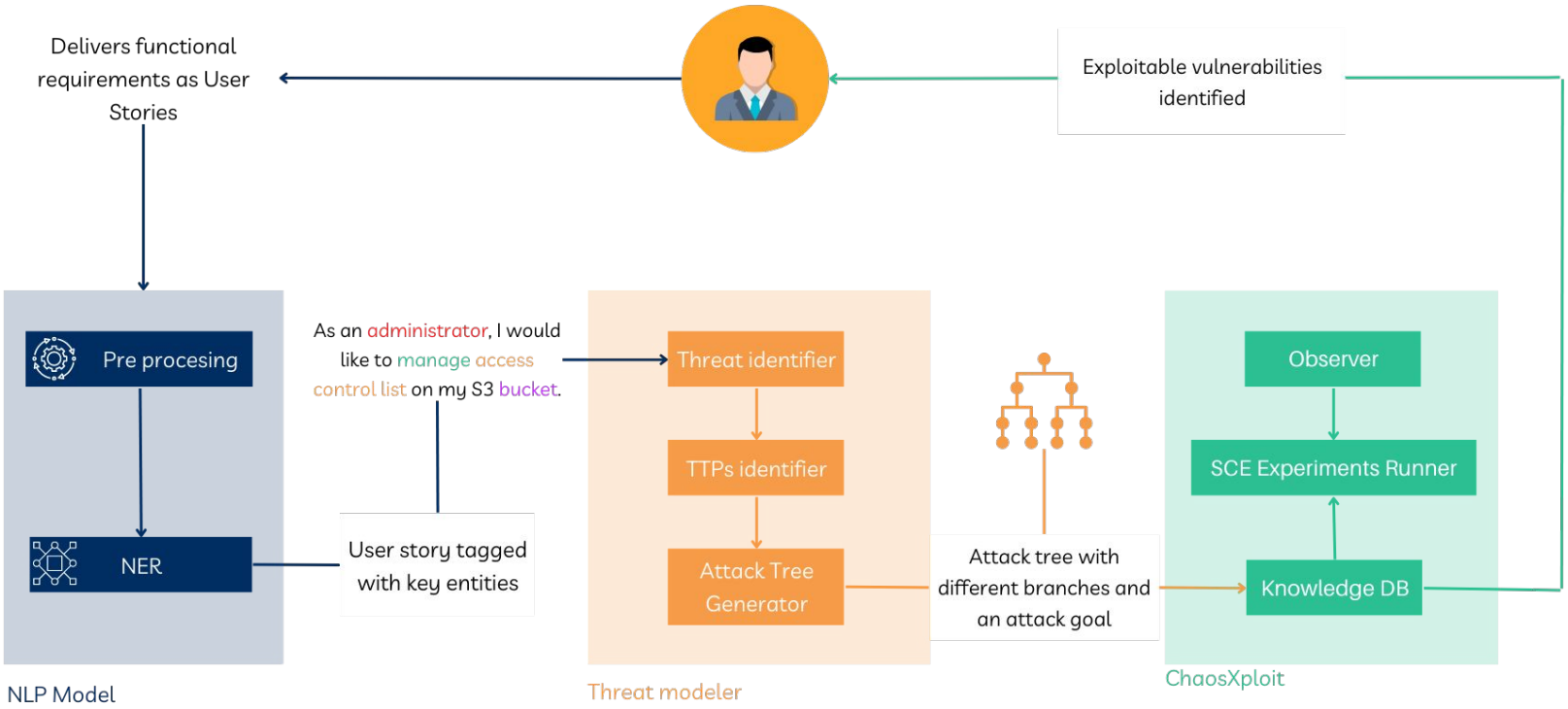


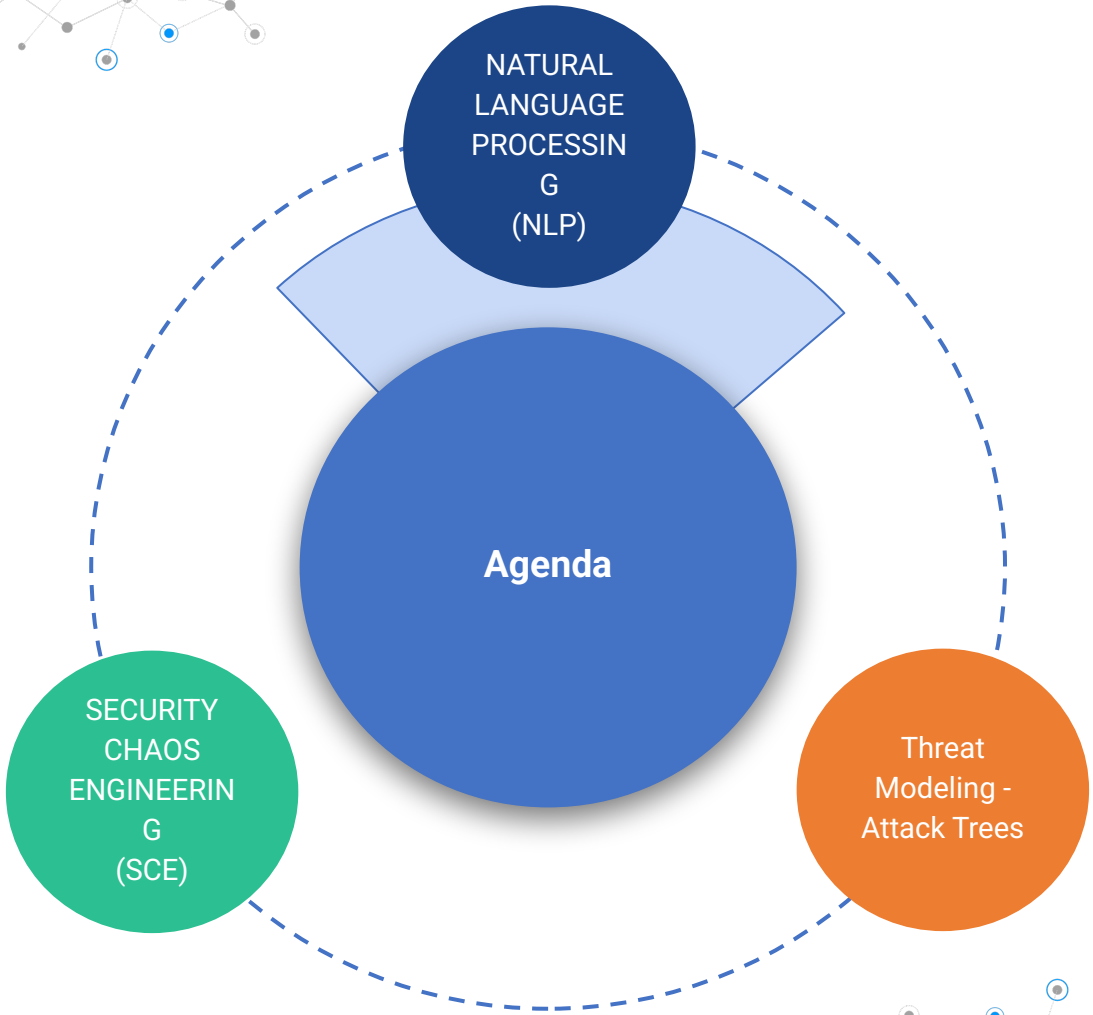
How can we achieve a real secure software development life cycle?

What about of we automatize parts of the SSDL?

What if we consider even the security in the User Stories?

Our Goal


Delivers functional requirements as User Stories
Exploitable vulnerabilities identified
Pre procesing
NER
User story tagged with key entities
As an administrator, I would like to manage access control list on my S3 bucket.
Threat identifier
TTPs identifier
Attack Tree Generator
Attack tree with different branches and an attack goal
Observer
SCE Experiments Runner
Knowledge DB
NLP Model
Threat modeler
ChaosXploit


NATURAL LANGUAGE PROCESSING (NLP)
Agenda
SECURITY CHAOS ENGINEERING (SCE)
Threat Modeling - Attack Trees

An Inspiring Phrase

"If a machine is expected to be infallible, it cannot also be intelligent."

Alan Turing

Natural Language Processing

"A branch of **artificial intelligence** that attempts to solve tasks related to human language, allowing communication between human and computer through natural language or solving different tasks that involve some kind of **preprocessing of text or speech**."

– IBM, 2017

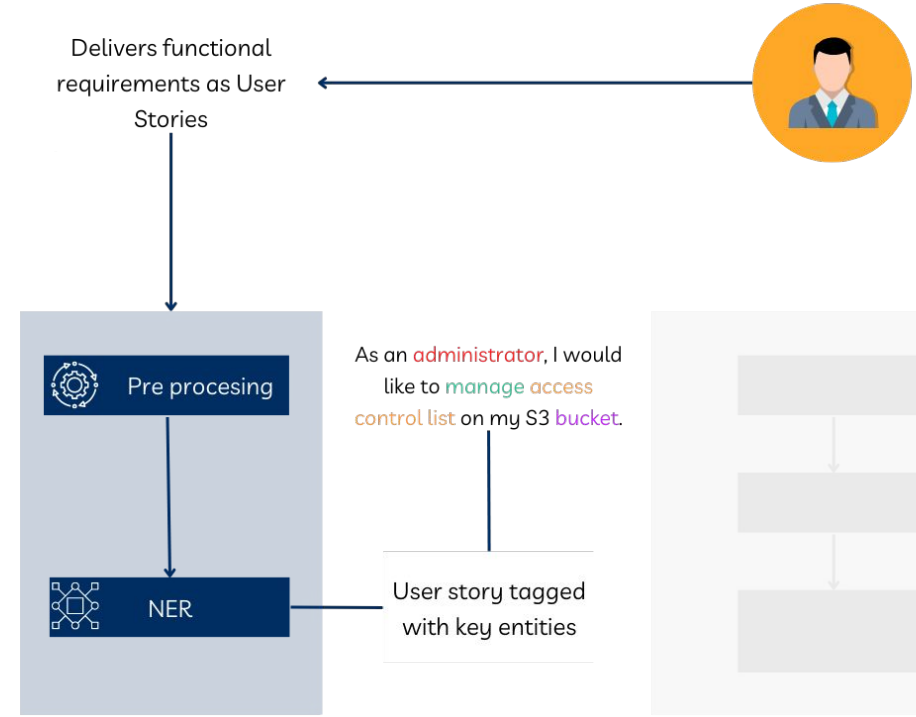
State of the Art of NLP models in cybersecurity

Universidad del Rosario | MACC Matemáticas Aplicadas y Ciencias de la Computación

Title	NLP Models/Algorithms	Challenge Resolve	Functionality
Cyber Security Vulnerability Detection Using Natural Language Processing. IEEE. Kanchan Singh, Sakshi S Grover, Ranjini Kishen Kumar. 2022	BERT	Identification and classification of vulnerable source code	**An alternative for traditional rule-based SAST.**
Natural Language Processing Model for Automatic Analysis of Cybersecurity-Related Documents. MDPI. Tiberiu-Marian Georgescu. 2020	NER, BERT	Cognitive text analysis of documents related to cybersecurity.	**Rapid understanding of cybersecurity items through graphs.**
SecureBERT: A Domain-Specific Language Model for Cybersecurity. Ehsan Aghaei, et al. 2022	BERT	Corpus with a lot of cybersecurity-related words.	**Allows text classification, text generation, sentiment analysis, etc.**
On the power of social networks to analyze threatening trends. IEEE. Ramirez Julian, et al. 2022	K-MEANS, POLARITY	Identification of threats in social networks.	**A way to discover the role of social networks in civic movements**

Are there NLP models available for software security? And even for security in software development?

Natural Language Processing



A Named Entity Recognition Based Approach for Privacy Requirements Engineering

Guntur Budi Herwanto
University of Vienna
Faculty of Computer Science
Vienna, Austria
a11947751@unet.univie.ac.at
Universitas Gadjah Mada
Yogyakarta, Indonesia
gunturbudi@ugm.ac.id

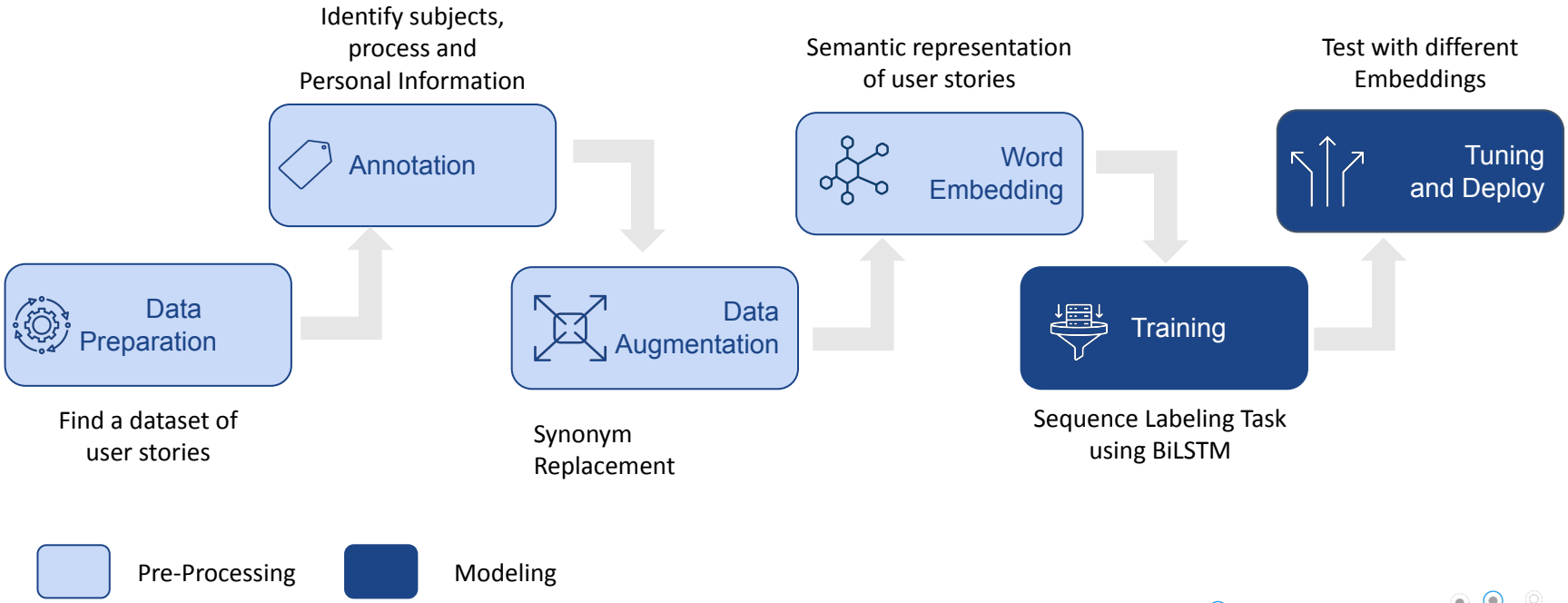
Gerald Quirchmayr
University of Vienna
Faculty of Computer Science
Research Group MIS
Währinger Straße 29
A-1090 Vienna, Austria
gerald.quirchmayr@univie.ac.at

A Min Tjoa
Vienna University of Technology
Institute of
Information and Software Engineering
Favoritenstrasse 9-11
A-1040 Vienna, Austria
a.tjoa@tuwien.ac.at

G. B. Herwanto at al, "A Named Entity Recognition Based Approach for Privacy Requirements Engineering," 2021 IEEE 29th International Requirements Engineering Conference Workshops (REW), Notre Dame, IN, USA, 2021, pp. 406-411, doi: 10.1109/REW53955.2021.00072

- It proposes an automated approach to assist **agile teams** in performing security activities
- It is mainly focused on supporting **threat modeling**
- It is focused on identify **Personally Identifiable Information** used in user story

Proposed Architecture in G. B. Herwanto at al



Data Preparation

As a Consumer, I want to view a data package online, so that I can get a sense of whether this is the dataset I want.
As a publisher, I want to show the world how my published data is, so that that it immediately catches consumer's attention.
As a consumer, I want to view the data package, so that that I can get a sense of whether I want this dataset or not.
As a Publisher, I want to preview a datapackage I have prepared, so that that I can check it works and share the results.
As a Consumer, I want to see how much the data has been downloaded, so that that I can choose most popular in the case when there a
As a Publisher, I want to see real examples of published packages so that I can understand how useful and simple the datapackage fo
As a Consumer, I want to see some example data packages quickly, so that I get a sense of what is on this site and if it is useful
As a Consumer, I want to search data packages, so that that I can find the ones I want.
As a Consumer, I want to search based on description of data package, so that that I can find package which related to some key wor
As a Consumer, I want to download the data package in one file, so that that I don't have to download descriptor and each resource
As a Developer, I want to use data package as a node lib in my project, so that that I can depend on it using my normal dependency
As a Consumer, I want to load a Data Package from R, so that that I can immediately start playing with it.
As a Data Analyst I want to download a data package, so that that I can study it and wrangle with it to infer new data or generate
As a Data Analyst, I want to update previously downloaded data package, so that that I can work with the most recent data.
As a Consumer, I want to download a DataPackage's data one coherent SQLite database, so that that I can get it easily in one form.
As a Data Analyst, I want to compare different versions of some datapackage locally, so that that I can see schema changes clearly

Fabiano Dalpiaz

https://data.mendeley.com/datasets/7zbk8zsd8y/1

Annotation

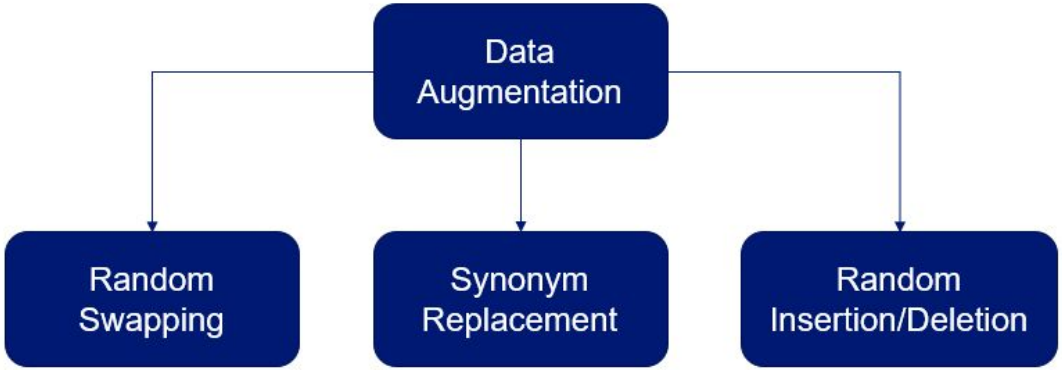
Define three entity categories:

Data Subject ⇨ The Actor
Processing ⇨ An Action
Personal Data ⇨ Privacy

```
16  As O
17  a O
18  camp B-Data Subject
19  administrator, I-Data Subject
20  I O
21  want O
22  to O
23  be O
24  able O
25  to O
26  keep B-Processing
27  camper B-PII
28  records I-PII
```

```
124  As O
125  a O
126  site B-Data Subject
127  member, I-Data Subject
128  I O
129  want O
130  to O
131  mark B-Processing
132  my O
133  profile B-PII
134  as O
135  private O
```

Data Augmentation



Is a technique to create synthetic data

As an administrator, I would like to manage access control list on my S3 bucket.

As an owner, I want to be able to configure permissions on my S3 bucket.

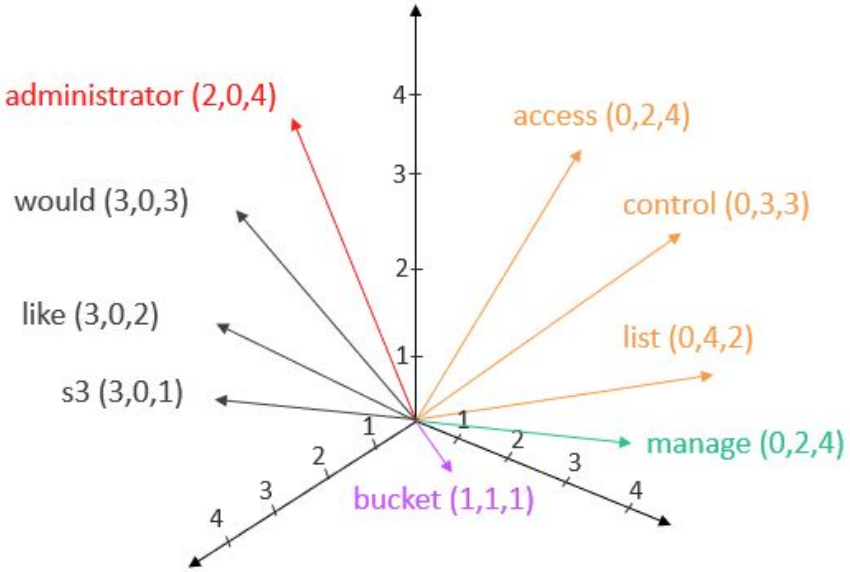
Universidad del Rosario
MACC
Matemáticas Aplicadas y Ciencias de la Computación

Word Embedding

Is a way to transform words into vectors

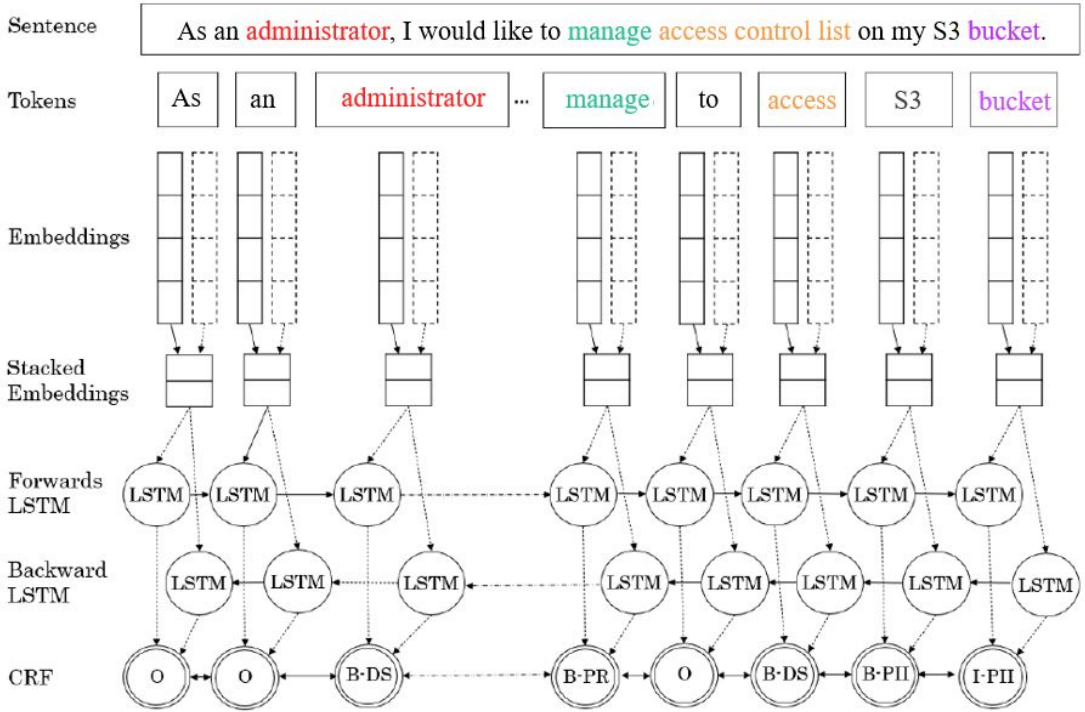
	a	b	c	d	e	f	g
administrator	0.5	0.7	0.2	-0.2	-0.9	0.1	0.3
would	0.4	0.6	0.3	0	-0.7	0.2	0.5
like	0.6	0.4	0	-0.2	-0.8	0.5	0.4
manage	-0.8	0.5	0	-0.3	0.2	-0.4	0.4
access	-0.5	0.1	0.9	0.8	0.7	-0.8	-0.9
control	-0.4	0.3	-0.1	-0.3	0.6	0.8	0.1
list	-0.4	0.2	0.7	0.5	0.4	-0.6	-0.2
S3	0.3	0	0.9	-0.2	0.7	-0.2	-0.3
bucket	0.2	0.2	0.6	0	0.9	-0.3	-0.1

Dimensional Reduction



As an administrator, I would like to manage access control list on my S3 bucket.

Training



Pipeline

Address the problem of model training as a sequence labeling task. The sequence is the word or phrase that has been assigned to a particular entity.

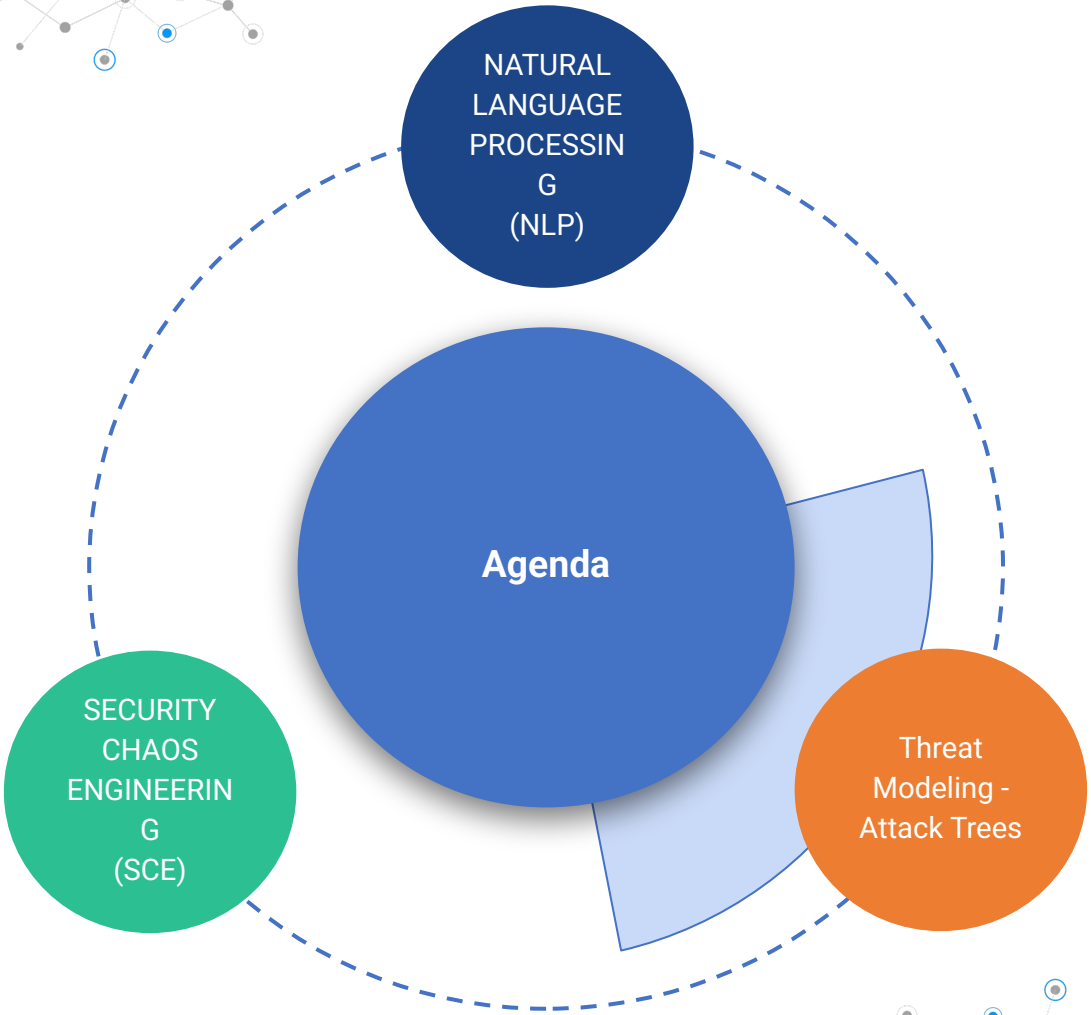
Deploy

```
2023-04-19 23:00:40,606 SequenceTagger predicts: Dictionary with 10 tags: <unk>, O, B-Data, B-Processing, B-PII, I-PII, I-Pr
ocessing, I-Data, <START>, <STOP>
As a developer, I want to configure access controls for an S3 bucket, so that only authorized users can view or modify the o
bjects stored in the bucket.
As a developer, I would like to authenticate myself in an S3 bucket with credentials
As a mobile application developer, I want to use Amazon S3 as a storage backend for user-generated content, such as images o
r videos.
As an administrator, I would like to manage access control list on S3 buckets
As a content creator, I want to be able to quickly retrieve files from an S3 bucket, so that I can incorporate them into my
content creation process.
As an administrator, I want to manage the version of the access policies of the S3 buckets
As a analyst, I want to be able to upload large datasets to an S3 bucket, so that I can analyze the data using AWS tools.
As a website owner, I want to store static website content in an S3 bucket
As a backup administrator, I want to configure lifecycle policies for an S3 bucket, so that information stored are automatic
ally deleted when they are no longer needed.
As a compliance officer, I want to be able to audit and monitor access to an S3 bucket, so that I can ensure that data is be
ing accessed and modified only by authorized users.
As an administrator, I would like to be able to assign permissions on an S3 bucket
As a data scientist, I want to be able to access S3 data through Jupyter notebooks, so that I can easily analyze and visuali
ze the data.
As owner I want to be able to store information on S3 buckets
As a security analyst, I want to configure S3 bucket policies and enable encryption, so that sensitive data is protected fro
m unauthorized access or disclosure.
As a DevOps engineer, I want to automate the deployment of code artifacts to an S3 bucket, so that my team can deploy change
s to the production environment more quickly and reliably.
```

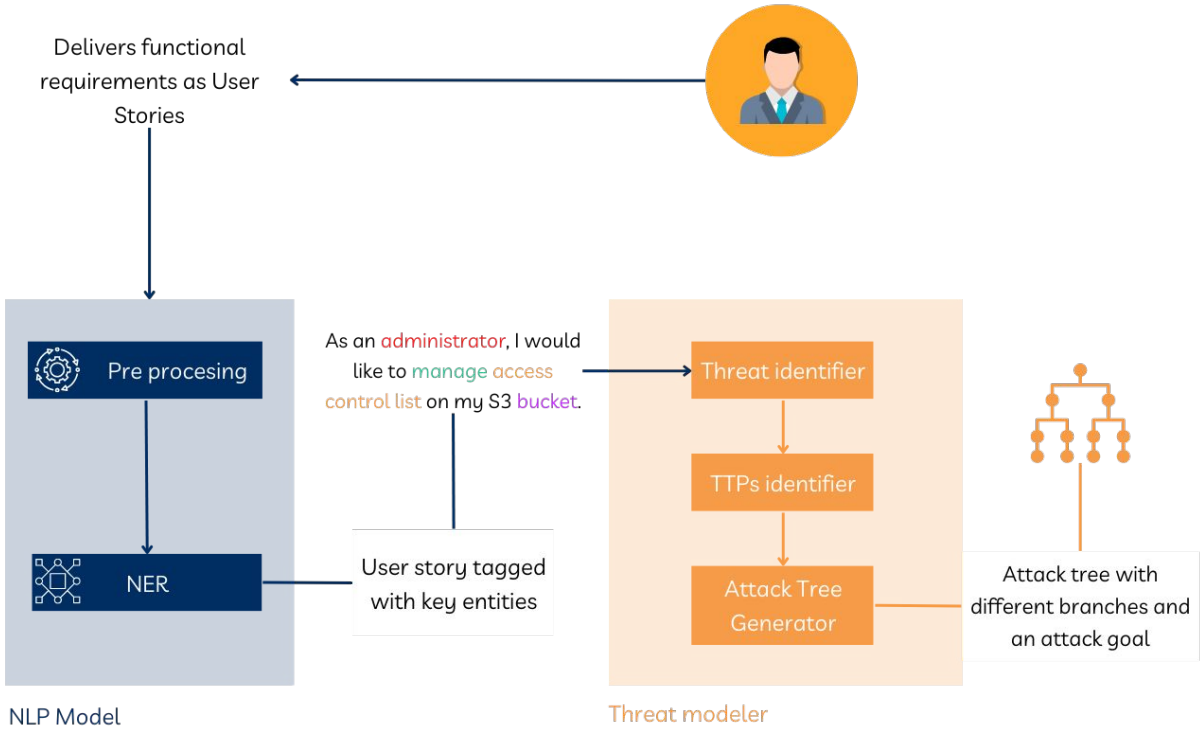
Resume

What we have learn until now?

- A mechanism to identify **actors**, **process** and **personal data** in user stories
- Applications of Natural Language Processing (NLP) to the cybersecurity context


NATURAL LANGUAGE PROCESSING (NLP)
Agenda
SECURITY CHAOS ENGINEERING (SCE)
Threat Modeling - Attack Trees

Threat modeling + Attack Trees



THREAT MODELING USING ATTACK TREES*

Vineet Saini
Acxiom Corporation, Conway AR
501-450-3401
vineetsai@gmail.com

Qiang Duan
University of Central Arkansas, Conway AR
501-450-3308
qduan@uca.edu

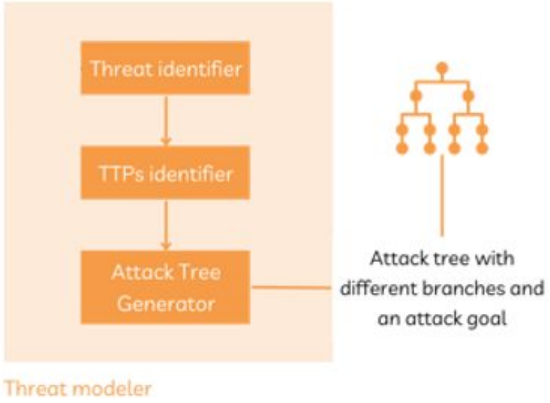
Vamsi Paruchuri
University of Central Arkansas, Conway AR
501-852-8537
vparuchuri@uca.edu

Saini, V., Duan, Q., & Paruchuri, V. (2008). Threat modeling using attack trees. *Journal of Computing Sciences in Colleges*, *23*(4), 124-131.

- Attack Trees provide a formal, methodical way of **describing** the security of systems, based on varying **attacks**.
- If an attack **costs** the perpetrator more than the **benefit**, that attack will most likely not occur.

Next Steps

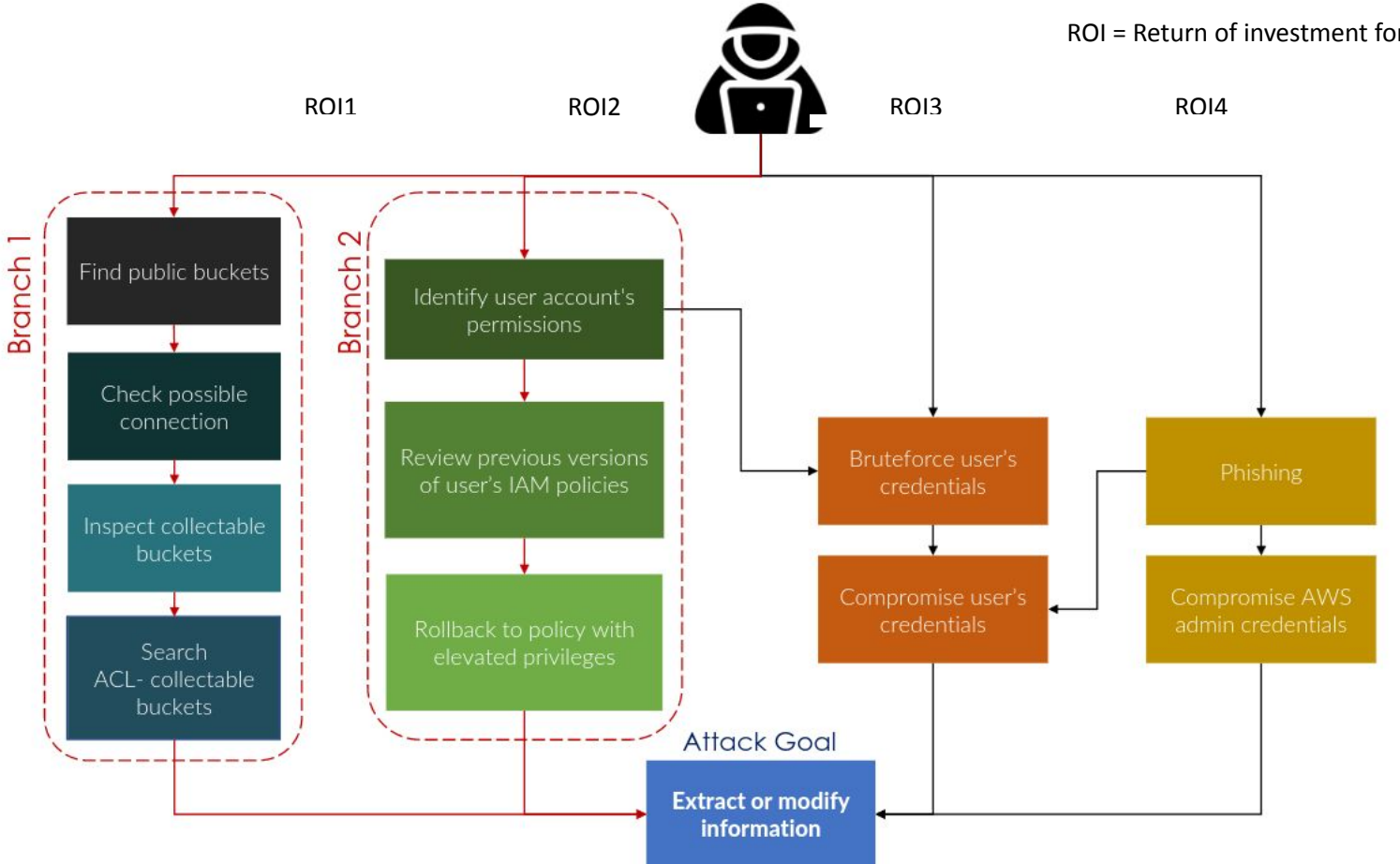
- Develop a threat correlator
- Automated attack tree generator based on **tactics**, **techniques** and **procedures**

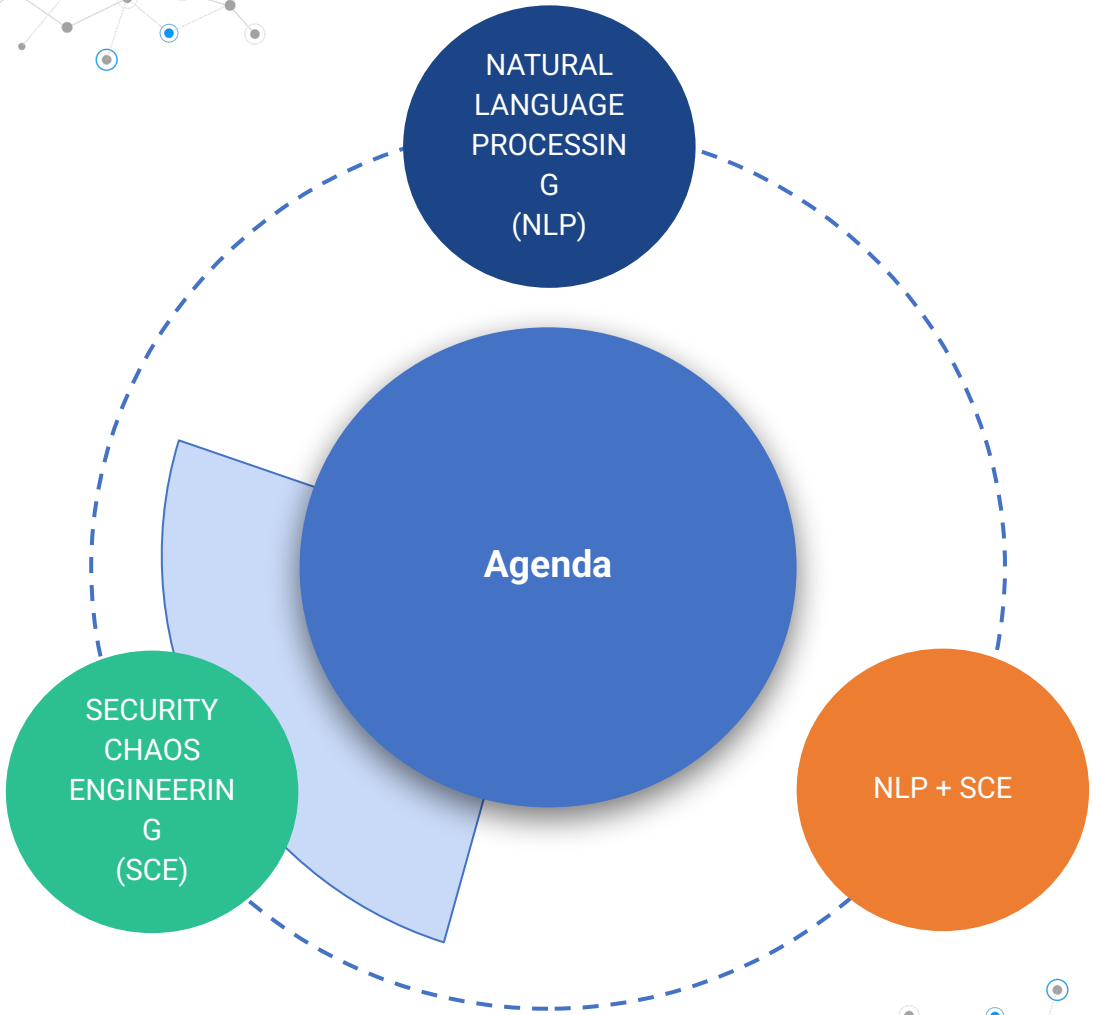


Threat modeling + Attack Trees

Universidad del Rosario | MACC Matemáticas Aplicadas y Ciencias de la Computación

ROI = Return of investment for the attacker




NATURAL LANGUAGE PROCESSIN G (NLP)
Agenda
SECURITY CHAOS ENGINEERIN G (SCE)
NLP + SCE

Chaos Engineering

"Chaos engineering is the practice of intentionally injecting failures into a system in order to improve its resilience." - Nora Jones

Security Chaos Engineering

"The identification of security control failures through proactive experimentation to build confidence in the system's ability to defend against malicious conditions in production."

Security Chaos Engineering, Gaining Confidence in Resilience and Safety at Speed and Scale.
Aaron Rinehart y Kelly Shortridge, Technical Report, O'REILLY, 2020

Security Chaos Engineering

Universidad del Rosario | MACC Matemáticas Aplicadas y Ciencias de la Computación

Methodology applied to SCE experiments

Rosenthal & Jones, 2020

- **Observation** — Establish metrics for further analysis, e.g. **HTTP request's response time**
- **Steady state** — Determines the equilibrium of the system to validate its security, e.g HTTP request's response time **< 10 ms**
- **Hypothesis** — Assumption made based on the system's resources that serves as a basis for determining its security maturity, e.g Assume that the **auto scale-up config** will work when a DoS attack shows up
- **Running the experiment** — Introduces offensive actions to validate or refute the hypothesis e.g. Try to deliberately **exhaust** the system

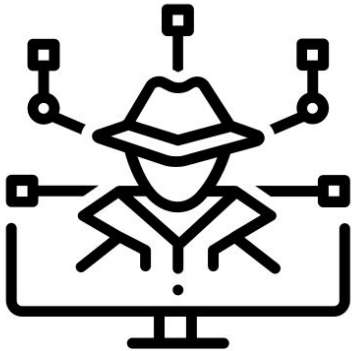
How does SCE complement traditional pentesting exercises?

- Offensive + defensive activities
- Vulnerabilities + Assumptions
- Scientific Method
- Applicable in the Testing + Deployment + Maintenance phases

Our Goal

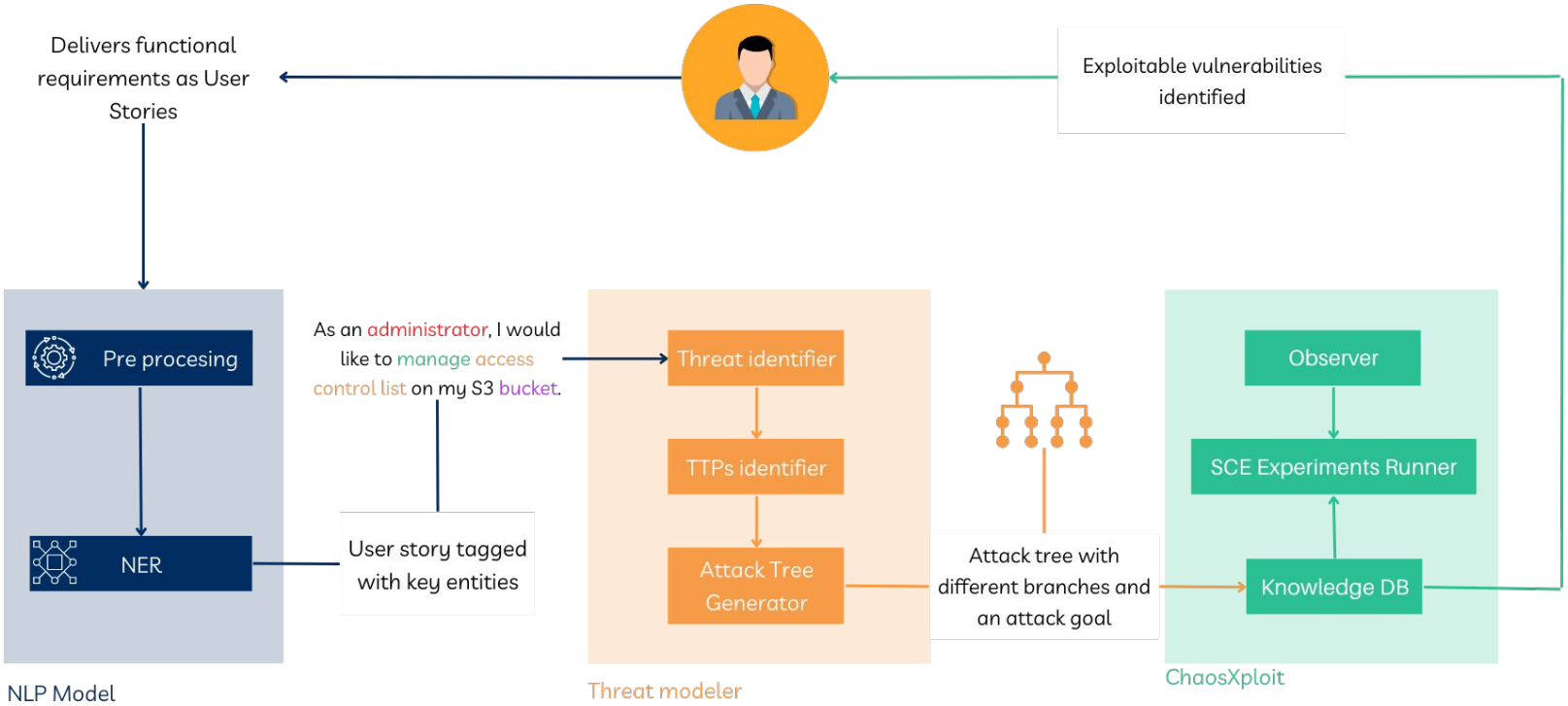
Universidad del Rosario | MACC Matemáticas Aplicadas y Ciencias de la Computación

Design and Implement a Security Chaos Engineering solution that could be used to identify vulnerabilities in software



Natural Language Processing + Security Chaos Engineering

Universidad del Rosario | MACC Matemáticas Aplicadas y Ciencias de la Computación



Our proposal: ChaosXploit

Universidad del Rosario | MACC Matemáticas Aplicadas y Ciencias de la Computación

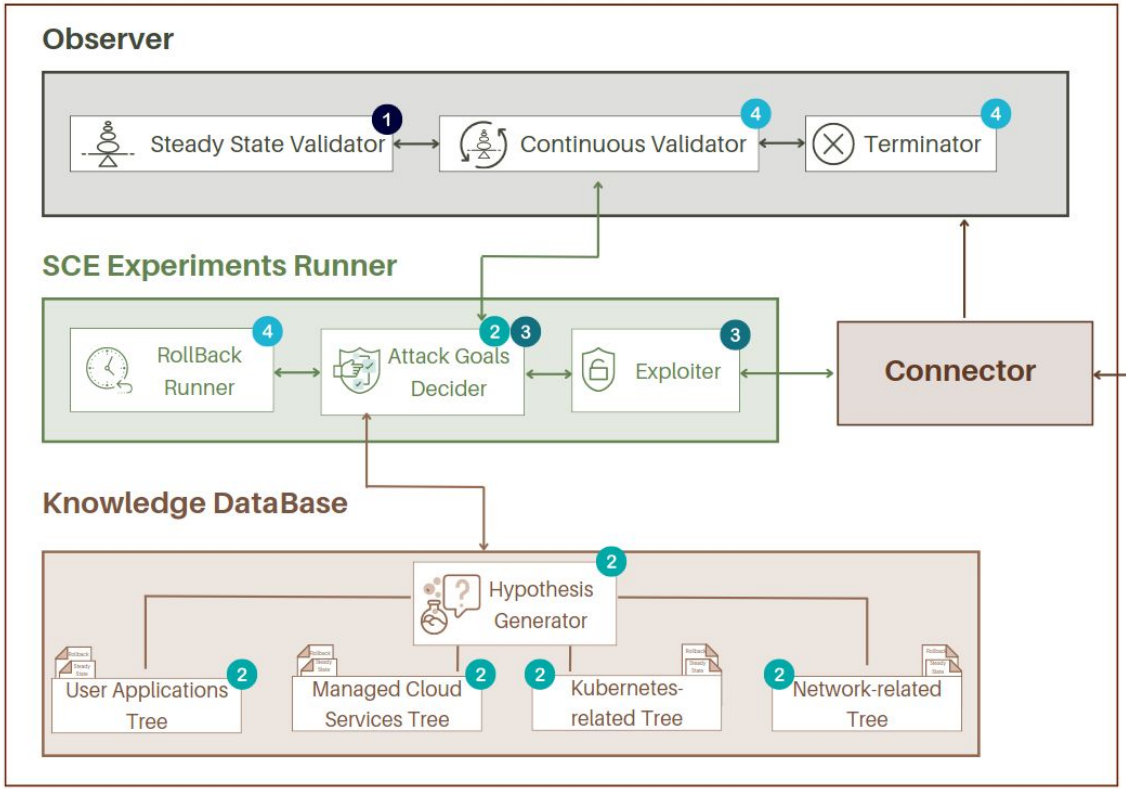
ChaosXploit

It is a framework powered by SCE based on attack trees composed of different modules that support the application of this methodology in different information systems.

Some ChaosXploit's modules are supported by ChaosToolkit

ChaosXploit is publicly available for the cybersecurity community through the repository http://github.com/SaraPalaciosCh/ChaosXploit

Our proposal: ChaosXploit



CE Method

1. Identify Steady State
2. Define the hypothesis
3. Execute the experiment
4. Observability & verification

Results

Observability

Accessibility to private information hosted in public AWS S3 Buckets

Steady state

The buckets have properly configured access controls that prevent unauthorized access.

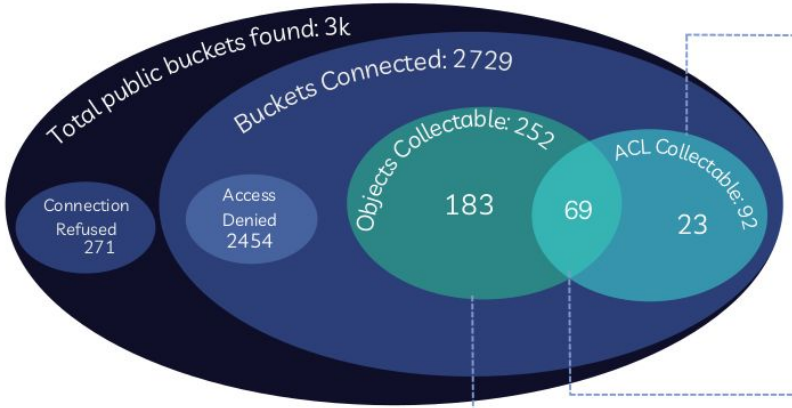
Hypothesis

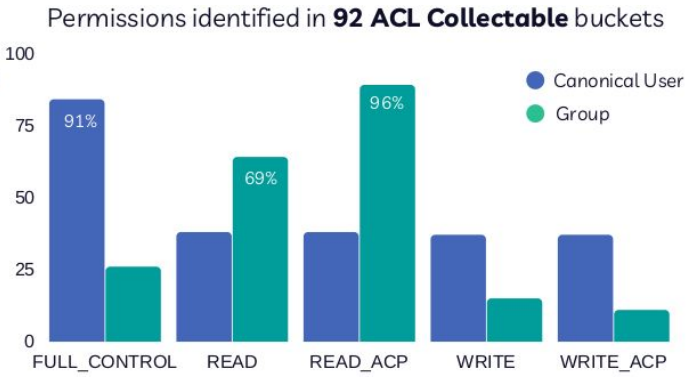
If someone external tries to access the objects stored in the buckets, they will NOT be able to see their content, NOR the permission settings, as they are properly configured to prevent information leaks.

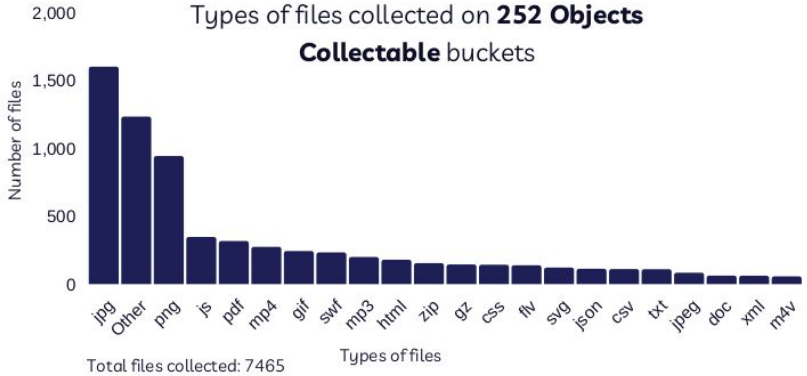
```
(p37) → Runner ./ChaosXploit exploit
------- Executing experiment: Public AWS s3 buckets ------
[2022-06-21 19:45:52 INFO] Validating the experiment's syntax
[2022-06-21 19:45:52 INFO] Experiment looks valid
[2022-06-21 19:45:52 INFO] Running experiment: Public buckets
[2022-06-21 19:45:52 INFO] Steady-state strategy: default
[2022-06-21 19:45:52 INFO] Rollbacks strategy: default
[2022-06-21 19:45:52 INFO] Steady state hypothesis: All the buckets in the list are properly configured
[2022-06-21 19:45:52 INFO] Probe: Are-Buckets-Collectable
Is Steady State validated?: True
Steady State validated
[2022-06-21 19:45:52 INFO] Steady state hypothesis is met!
[2022-06-21 19:45:52 INFO] Playing your experiment's method now...
[2022-06-21 19:45:52 INFO] Action: Collecting-Objects
All tests will be executed in anonymous mode
```

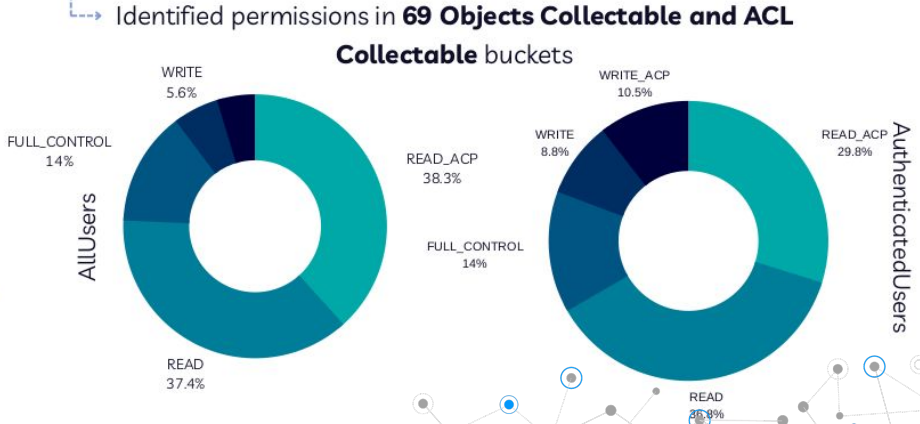
```
Checking bucket brent
Checking bucket chakra
Checking bucket confucius
Permission: FULL_CONTROL found
Permission: FULL_CONTROL found
Permission: FULL_CONTROL found
Bucket 'chakra' collectable: http://s3.us-east-1.amazonaws.com/chakra/Chakra Intro/the chakra system Intro.mp4 !!!
Checking bucket data123
Finding up to 50 files in bucket chakra
Checking bucket gw4
Checking bucket dvr2
Checking bucket enigma
Bucket 'brent' collectable: http://s3.ap-southeast-2.amazonaws.com/brent/Edited Skin/Meika Outside3.m4v !!!
Checking bucket biostat
Checking bucket howell
Finding up to 50 files in bucket brent
[2022-06-21 19:46:14 INFO] Steady state hypothesis: All the buckets in the list are properly configured
[2022-06-21 19:46:14 INFO] Probe: Are-Buckets-Collectable
Is Steady State validated?: False
Failed validation for Collectable Buckets
[2022-06-21 19:46:14 CRITICAL] Steady state probe 'Are-Buckets-Collectable' is not in the given tolerance so failing this experim
[2022-06-21 19:46:14 INFO] Experiment ended with status: deviated
[2022-06-21 19:46:14 INFO] The steady-state has deviated, a weakness may have been discovered
(p37) → Runner
```

Results


Total public buckets found: 3k
Buckets Connected: 2729
Objects Collectable: 252
183
69
ACL Collectable: 92
23
Connection Refused 271
Access Denied 2454


Permissions identified in 92 ACL Collectable buckets
Canonical User
Group
100
75
50
25
0
91%
69%
96%
FULL_CONTROL
READ
READ_ACP
WRITE
WRITE_ACP

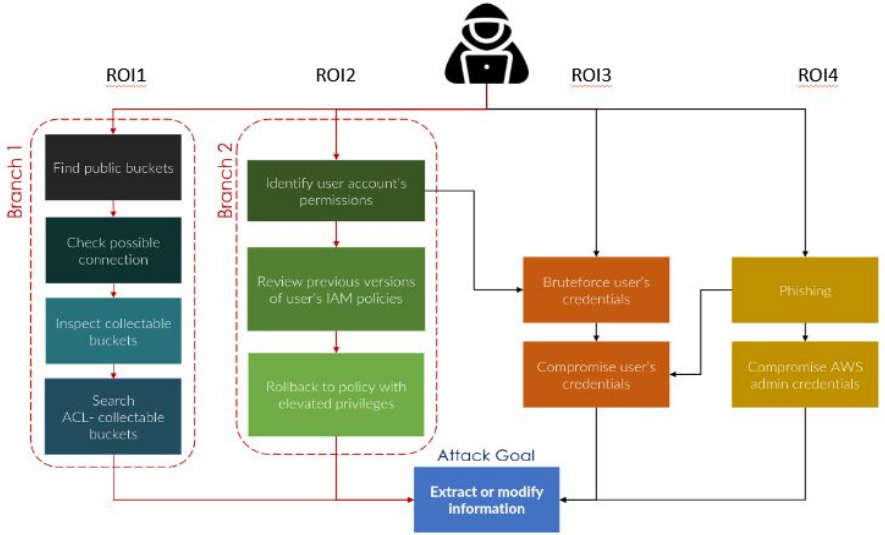

Types of files collected on 252 Objects Collectable buckets
Number of files
2,000
1,500
1,000
500
0
jpg
Other
png
js
pdf
mp4
gif
swf
mp3
html
zip
gz
css
flv
svg
json
csv
txt
jpeg
doc
xml
m4v
Types of files
Total files collected: 7465

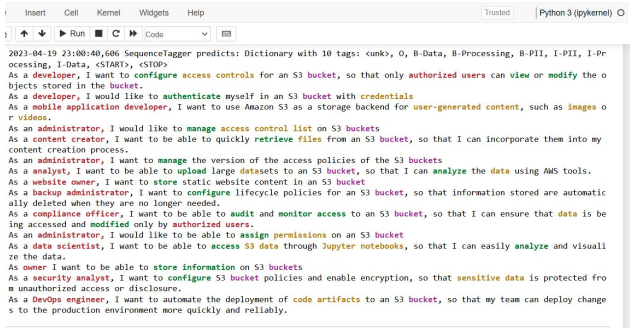

Identified permissions in 69 Objects Collectable and ACL Collectable buckets
AllUsers
WRITE
5.6%
FULL_CONTROL
14%
READ
37.4%
AuthenticatedUsers
WRITE_ACP
10.5%
WRITE
8.8%
READ_ACP
29.8%
READ_ACP
38.3%
FULL_CONTROL
14%
READ
36.8%

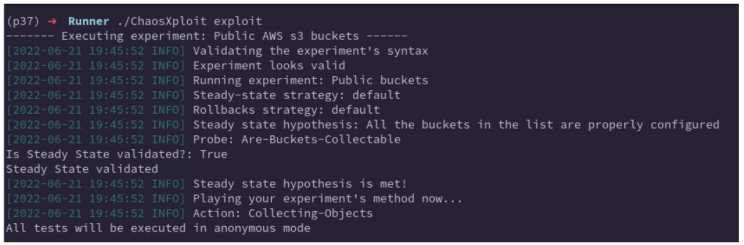
Universidad del Rosario | MACC Matemáticas Aplicadas y Ciencias de la Computación

NLP + SCE

- Information Extraction from user stories
- Correlation of threat
- Attack Trees
- Automated attack tree generator
- Security Test







ChaosXploit

Some thoughts

- SCE is an alternative way for securing systems throughout their life cycle
- SCE offers an automated mechanism for documenting security activities during system's life cycle
- SCE is an interesting way for security testing using observability, steady state and hypothesis
- SCE is a friendly approach to highlight the importance of securing systems.

THANKS

Universidad del Rosario | MACC Matemáticas Aplicadas y Ciencias de la Computación

BSIDES Cybersecurity conference
Bogotá, Colombia, April 22nd ,2023