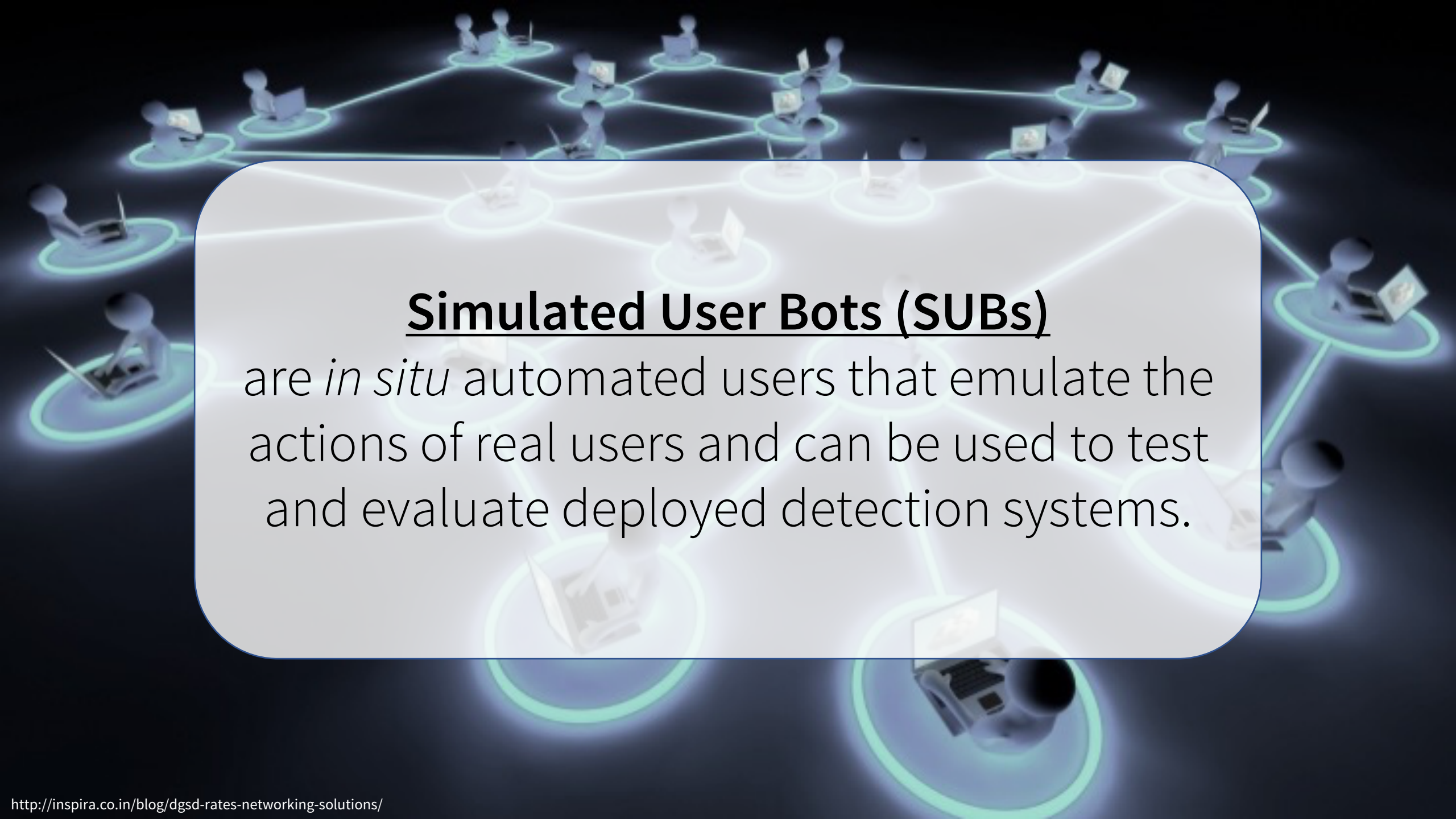
Simulated User Bots: Real Time Testing of Insider Threat Detection Systems

Preetam K. Dutta, Gabe Ryan, Alek Zieba and **Salvatore J. Stolfo**

COLUMBIA UNIVERSITY
IN THE CITY OF NEW YORK

Finding an Insider Threat is an established problem:
But, does the system ***work***?

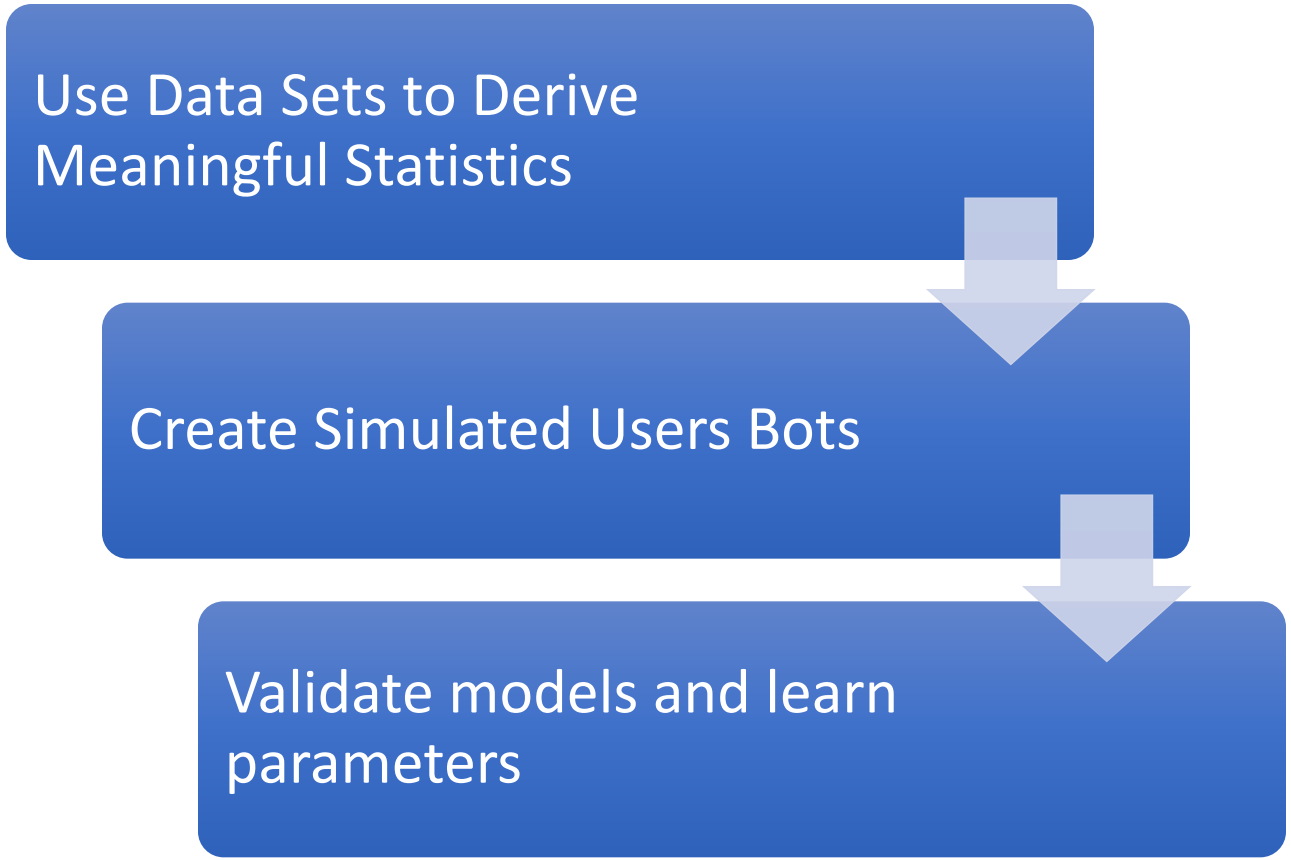


Simulated User Bots (SUBs)

are *in situ* automated users that emulate the actions of real users and can be used to test and evaluate deployed detection systems.

http://inspira.co.in/blog/dgsd-rates-networking-solutions/

Goals


Use Data Sets to Derive Meaningful Statistics
Create Simulated Users Bots
Validate models and learn parameters

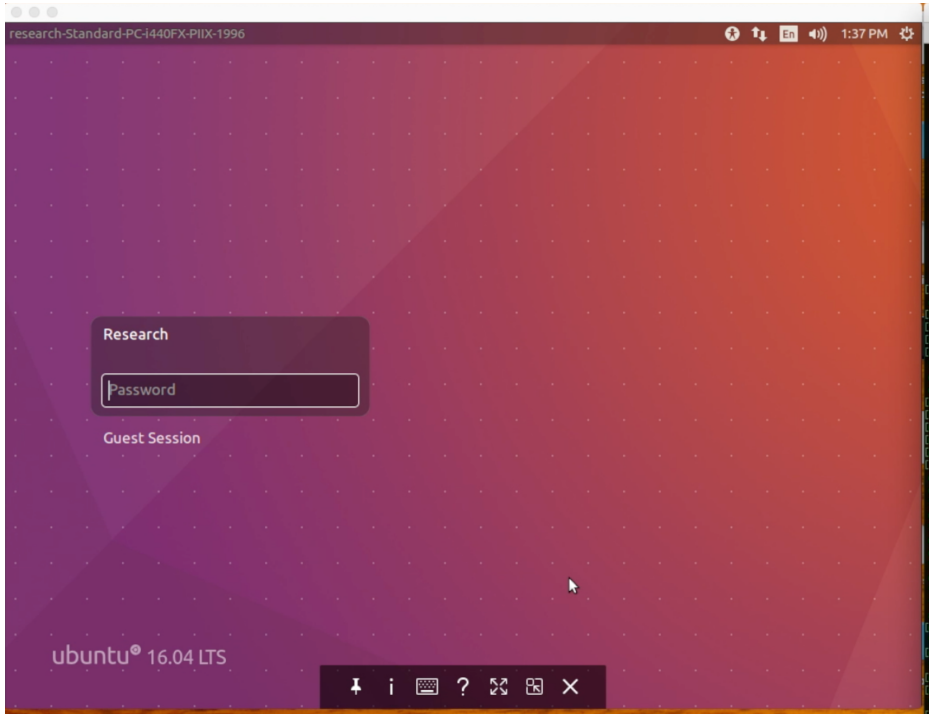
What are the Parts?

- QEMU (Quick Emulator)
  - Open source and emulates a full system, including a processor

- KVM (Kernel-based Virtual Machine)
  - Virtualization infrastructure for Linux kernel that turns it into a hypervisor

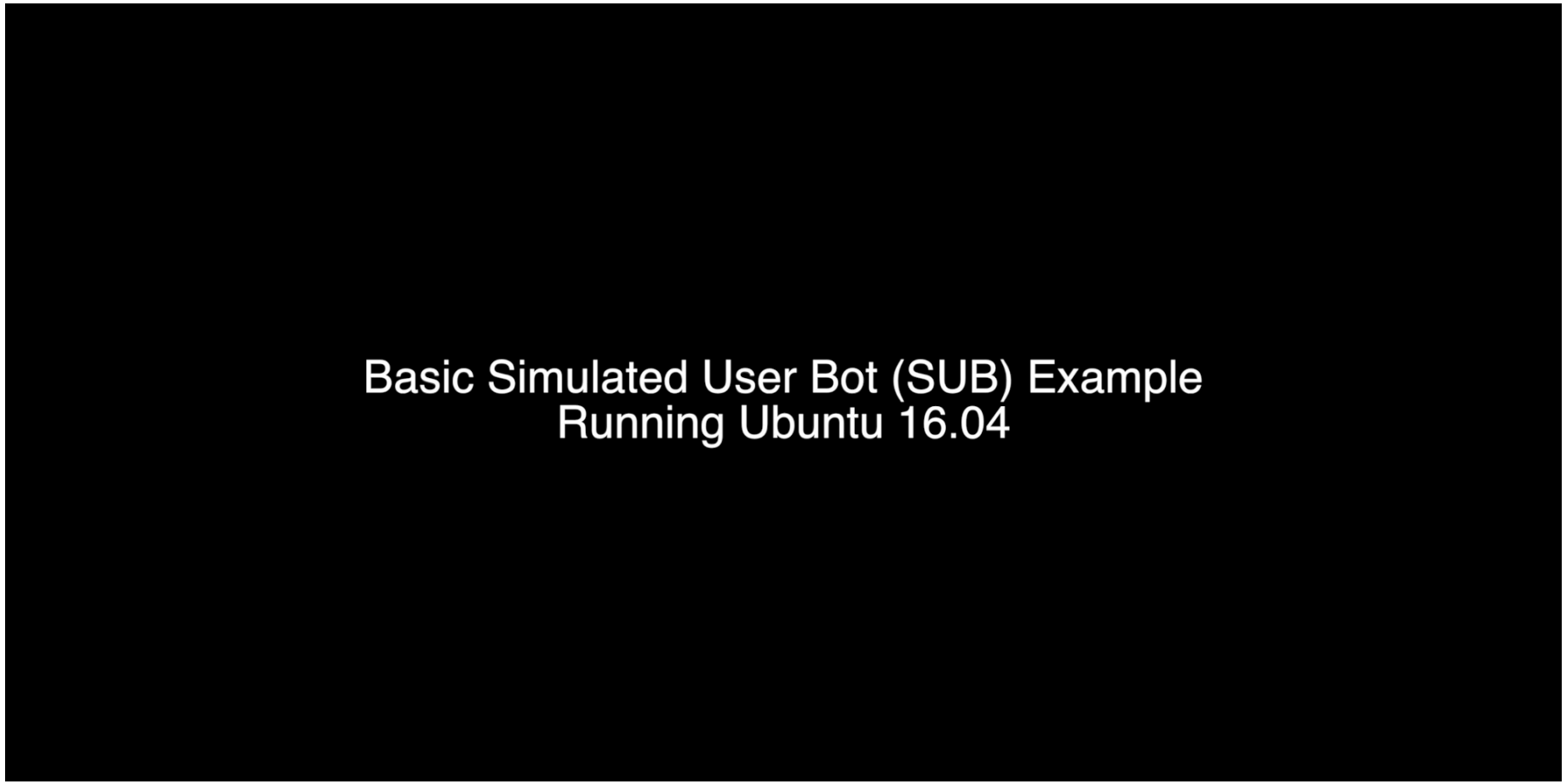
- VNCDoTool (a Python Library)
  - Automate input through VNC connections

What functionality is supported?

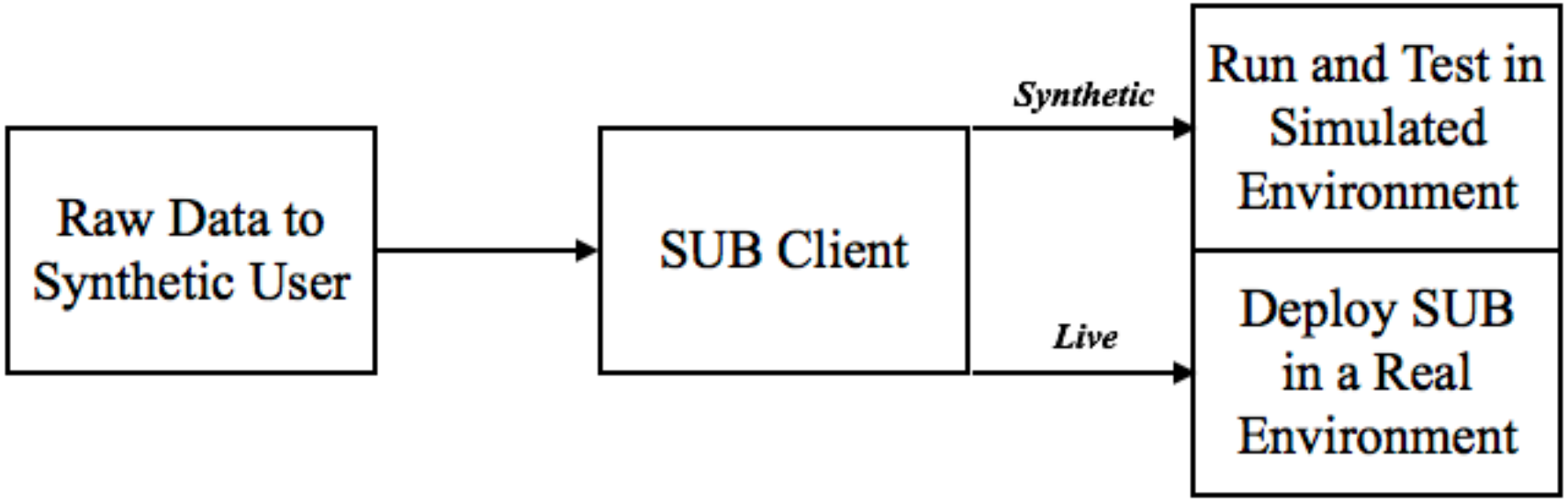


- Opening an app
- Visiting a website
- Performing a web search
- Sending an email
- Copying & pasting
- User Login
- Etc.

What does a SUB look like?



What is the SUB Framework?

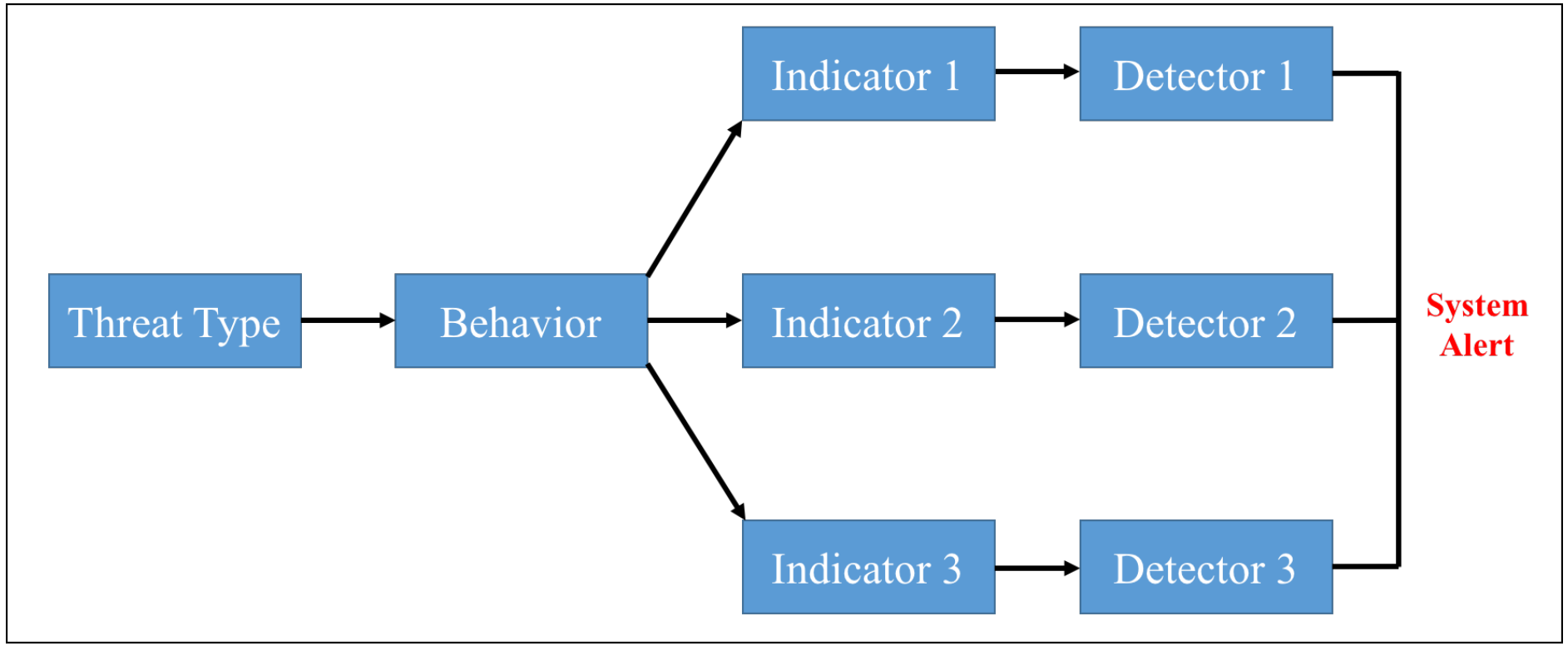


Synthetic Environment: Identify Problem

Threat Type	Behavior	Indicator	Detector
Individuals with abnormal work habits	Uses a work-owned machine outside of normal work hours (i.e., 12AM-7AM EST) at work, at home, or at remote sites	Has anti-virus updates between 12AM and 7AM EST	At least 1 anti-virus log entry for definition updates between 12AM and 7AM EST
		In the top 5% of the frequency distribution of VPN activity between 12AM and 7AM EST	At least 28 VPN log records showing a connection that started between 12AM and 7AM EST
		In the top 5% of the frequency distribution of workstation logins between 12AM and 7AM EST	At least 128 ADDC log entries with timestamps between 12AM and 7AM EST
		In the top 5% of the frequency distribution of email activity between 12AM and 7AM EST	At least 24 email log records for outbound emails sent between 12AM and 7AM EST
		In the top 5% of the frequency distribution of website activity between 12AM and 7AM EST	At least 24,500 proxy log entries with a timestamp between 12AM and 7AM EST
Individuals with abnormally large data transfers	Uses a work-owned machine for abnormally	In the top 5% of the frequency distribution of attempted access to prohibited file sharing websites	At least 2000 proxy log entries of attempted access to prohibited file sharing websites
		In the top 5% of the frequency distribution of large emails sent	At least 3 emails with attachments larger than 5 MB
		In the top 5% of the frequency distribution of VPN sessions that transfer large amounts of data	At least 7 VPN sessions where at least 210 MB are transferred

Adopted from the IARPA SCITE Program

Synthetic Environment: Identify Problem



Adopted from the IARPA SCITE Program

Synthetic Environment: Problem

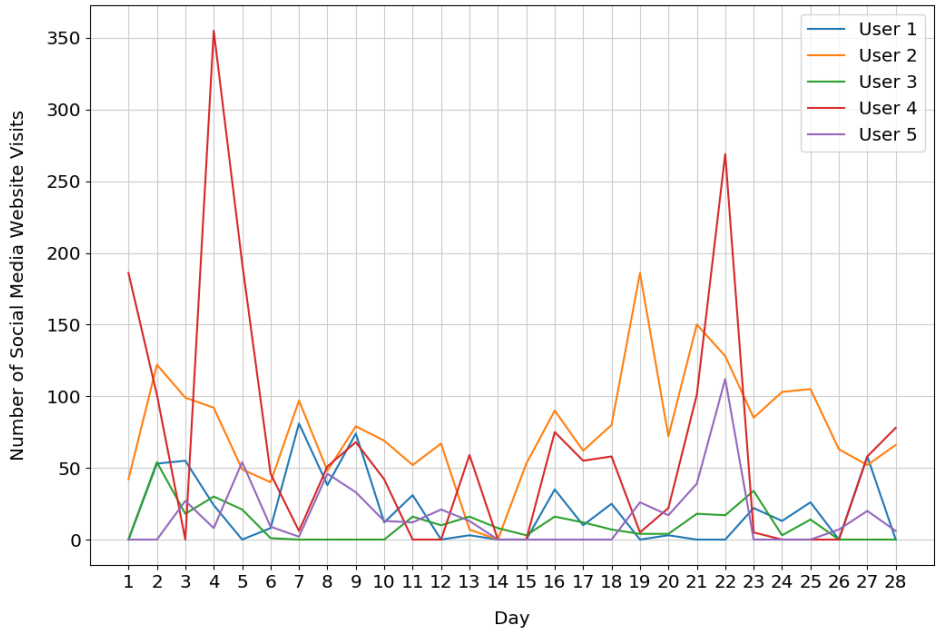
Indicators:

1. In the top 5 percent of the daily frequency average distribution of Google or Bing searches between 5:00:01 PM and 6:59:59 AM EST
2. In the top 5 percent of the daily frequency average distribution of social media website visits between 5:00:01 PM and 6:59:59 AM EST
3. In the top 5 percent of the daily frequency average distribution of actions on files and documents between 5:00:01 PM and 6:59:59 AM EST

Detectors:

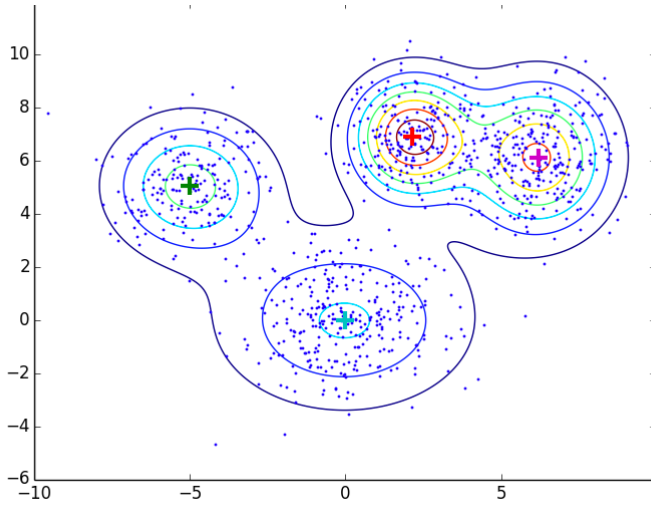
1. At least 13 log entries for a Google or Bing search between 5:00:01 PM and 6:59:59 AM EST.
2. At least 61 log entries for a social media website visit between 5:00:01 PM and 6:59:59 AM EST.
3. At least 90 log entries for actions on files and documents between 5:00:01 PM and 6:59:59 AM EST.

Synthetic Environment: Data

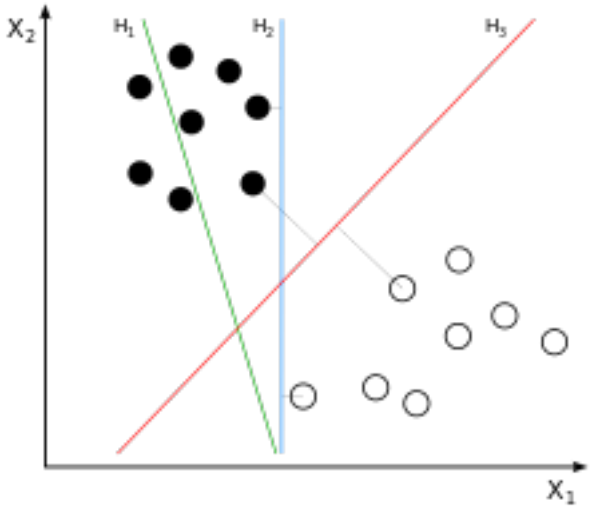


Contains over 60 users (~18 GB of data with roughly 1 million records per user) whose actions performed were monitored.

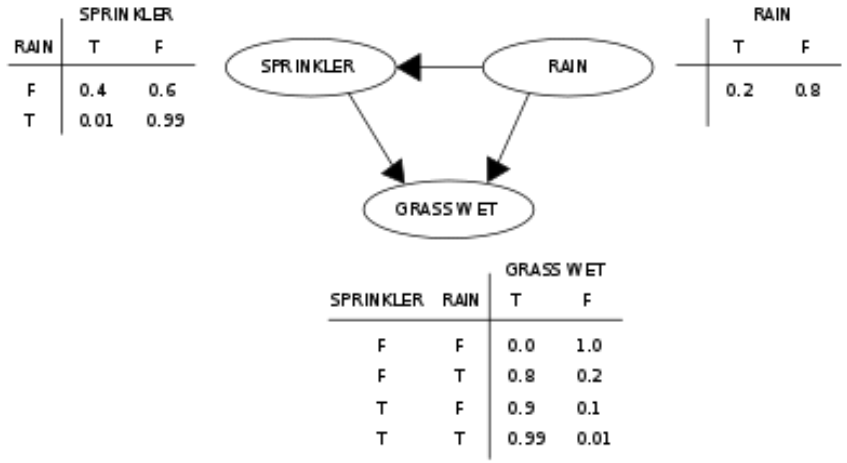
Synthetic Environment: Setup



Gaussian Mixture Model



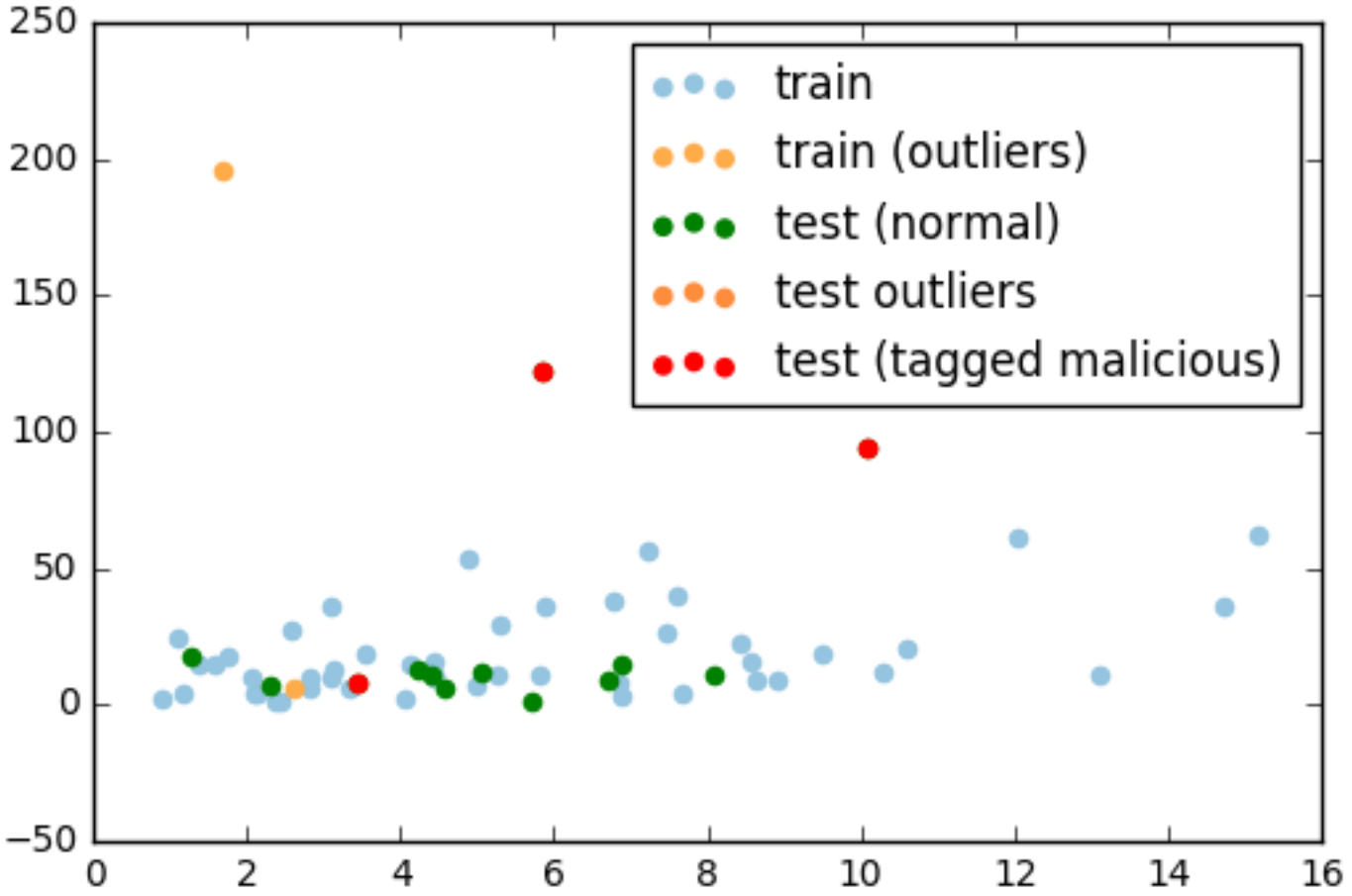
Support Vector Machine



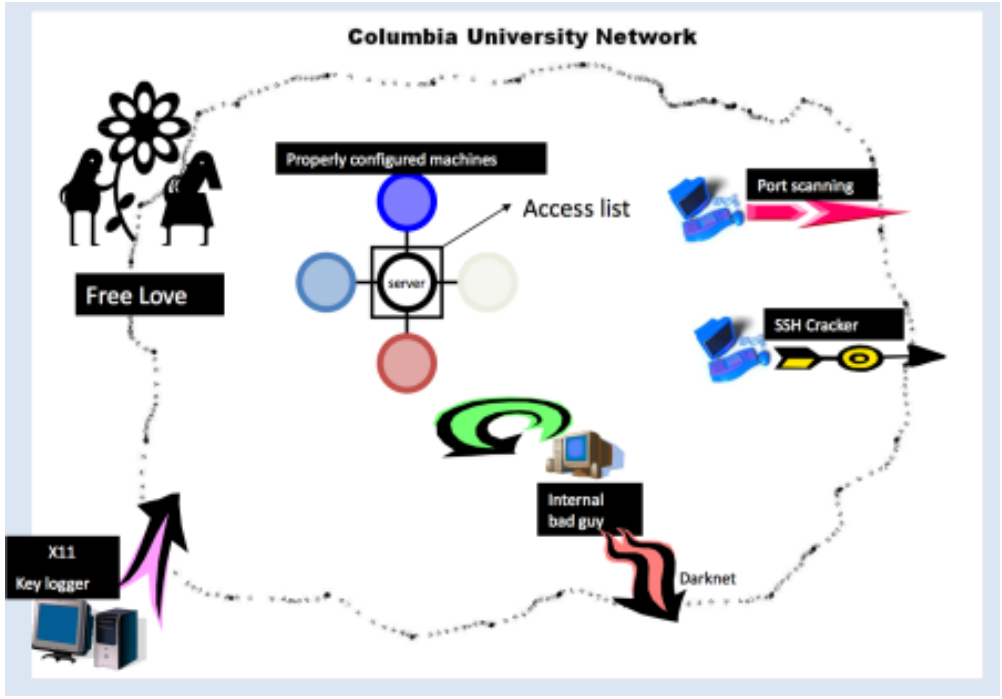
Bayesian Network

http://yulearning.blogspot.com/2014/11/einsteins-most-famous-equation-is-emc2.html
https://en.wikipedia.org/wiki/Support_vector_machine
https://en.wikipedia.org/wiki/Bayesian_network

Synthetic Environment: Results



Live Testing: Columbia University

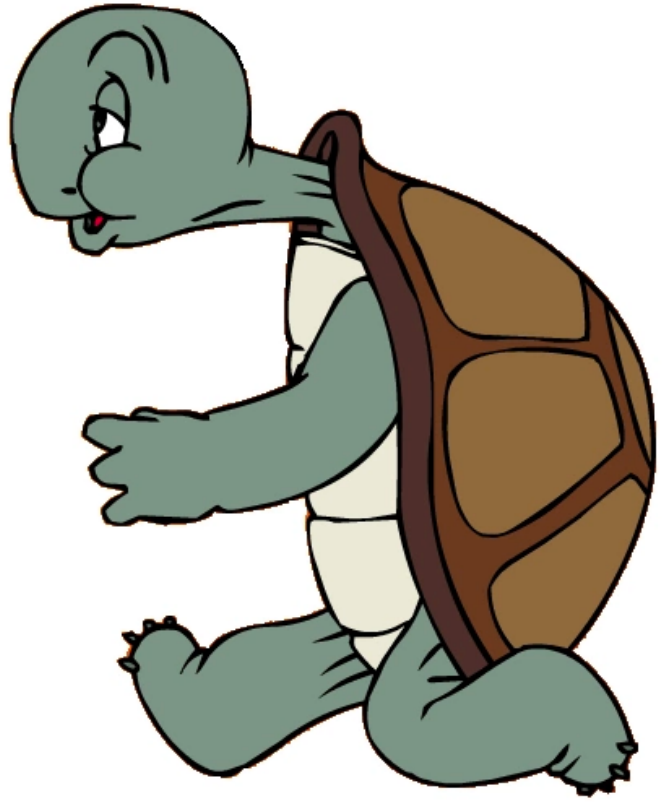


- Decentralized management structure
- Over 100,000 network nodes
- Over 55,000 MAC addresses active on average
- No sniffing traffic or scanning machines permitted
- No university wide firewalls
- Approximately 80,000 active email addresses

Live Testing: Columbia University

Low and Slow Attack

```
for i in {1..2000};
        do ssh -i ~/.ssh/aws-key.pem gryan@67.80.188.110;
        sleep 15m;
done;
```



Conclusions

1. SUBs have the ability to mimic the behaviors & actions of real users with no interference of normal system operations.
2. Successfully able to design experiments to show that we could inject malicious activities such that a normal user appears to act as a malicious user.
3. Validated the risk of low & slow attacks, which underscores the importance of the health and boundaries of an intrusion detection system

Future Work

1. Find a Corporate Environment to test efficacy of SUBs
2. Test new Indicator and Detectors to improve IDS
3. Test the robustness of other technologies such as decoys

COLUMBIA UNIVERSITY
IN THE CITY OF NEW YORK