# Data+ project plaln

## Problem statement and objective

Interested in understanding the types of attacks targeting Duke and other universities?  Work with OIT and the IT Security Office to analyze threat intelligence data to identify trends and patterns of attacks.  Duke blocks an average of 1.5 billion malicious connection attempts/day and is working with other universities to share the attack data.  One untapped area is research into the types of attacks and learning how universities are targeted.  Students will get to work alongside the security and IT professionals in analyzing the data and with the intent to discern patterns.

## Resources

**OIT-project resources and stakeholders**

* Jesse Bowling (jesse.bowling@duke.edu), IT Consultant (IT Security)
* John Haws (john.haws@duke.edu), IT Consultant (Data Science)
* Eric Hope (eric.hope@duke.edu), Sr. IT Analyst
* Gagan Kaur (gk87@duke.edu), Data Analyst

**Data+ team**

* Susan Jacobs (susan.jacobs@duke.edu), Project Manager
* Rijish Ganguly (rg239@duke.edu) – MS in Computer Engineering
* Sibora Seranaj (ss801@duke.edu) – BS in Statistical Science

## Onboarding

Initial meeting

1) Discuss the problem statement – questions to get started

2) Explain the data

3) Introduce the environment – need for additional resources, any training for working with the cluster? intro to threat intelligence? STINGAR Q/A with Jesse for security specific questions

Questions to get started

1) Are there any differences between attacks on our cloud and local honeypots?

2) How do the attacks between different universities vary after normalizing for the number of honeypots deployed by each?

3) Are there patterns in evolution of attacks over time, at indicator level, sensor level? (Trends in time of day, day of week for an attacks)

4) Do we see the indicators returning after being blocked? If yes, is there any change in activity in terms of number of connections over time or usernames/password?

5) Attacks classification based on distribution of number of connections, number of sensors hit, usernames, password tried, urls used

## Deliverables

Expected to evolve throughout the project, we are looking for insights into the following

* Trends in attacks among universities
* Trends at Duke
* Can we classify and predict attacks?

The analysis could be organized into different phases as above through visualizations, reports, presentations, etc.

## Schedule

Duration – 10 Weeks

### Regular Meetings

* Short weekly updates (in-person or hangouts) – to make improvements each week and discuss roadblocks/challenges
* Slack for general questions like clarifications on dataset, or sharing ideas, or maybe even motivation

### Milestones

June 24 (Week 5) – First talk

Focused on presenting insights into how the attacks compare between universities

July 29 (Week 10) – Final presentation

Focused on trends at Duke/attack classification

TBD – Final presentation at OIT (internal)

## Evaluation

Key evaluation metrics

* Making progress every week and staying within the schedule
* Generating critical insights and reports and sharing it effectively with all stakeholders
* Students feel confident in their abilities to tackle real-world analytics projects

## Risks

* New tool training overhead for our current environment (Spark) could delay the project
* Risks associated with working with big data environment, larger run-times for certain crucial queries, connection losses due to VPN
* Data might be unfit for detecting certain patterns in time-series
* Lesser data available for analyzing common attacks