SEAS-8414

Analytical Tools for Cyber Analytics

Survey of analytical tools for analyzing cyber security data with particular attention to the use of data analytics procedures in supporting appropriate cyber security policy decisions.

Dr. M

Welcome to SEAS Online at George Washington University

SEAS-8414 class will begin shortly

- **Audio:** To eliminate background noise, please be sure your audio is muted. To speak, please click the hand icon at the bottom of your screen (**Raise Hand**). When instructor calls on you, click microphone icon to unmute. When you've finished speaking, ***be sure to mute yourself again***.
- **Chat:** Please type your questions in Chat.
- **Recordings:** As part of the educational support for students, we provide downloadable recordings of each class session to be used exclusively by registered students in that particular class for their own private use. Releasing these recordings is strictly prohibited.

Agenda

Week-3: Introduction to endpoint security analytics tools

We will shift the focus from data to endpoint-centric security tools for gwuscc.com. We will cover:

- Endpoint Detection and Response (EDR)
- Endpoint Protection Platform (EPP)
- Extended Detection and Response (XDR)

We will learn about how XDR leverages time-series data from system event logs and leverages regression for predictive analytics. We will tie how data and asset-centric security are essential for an enterprise cybersecurity program.

Class-3

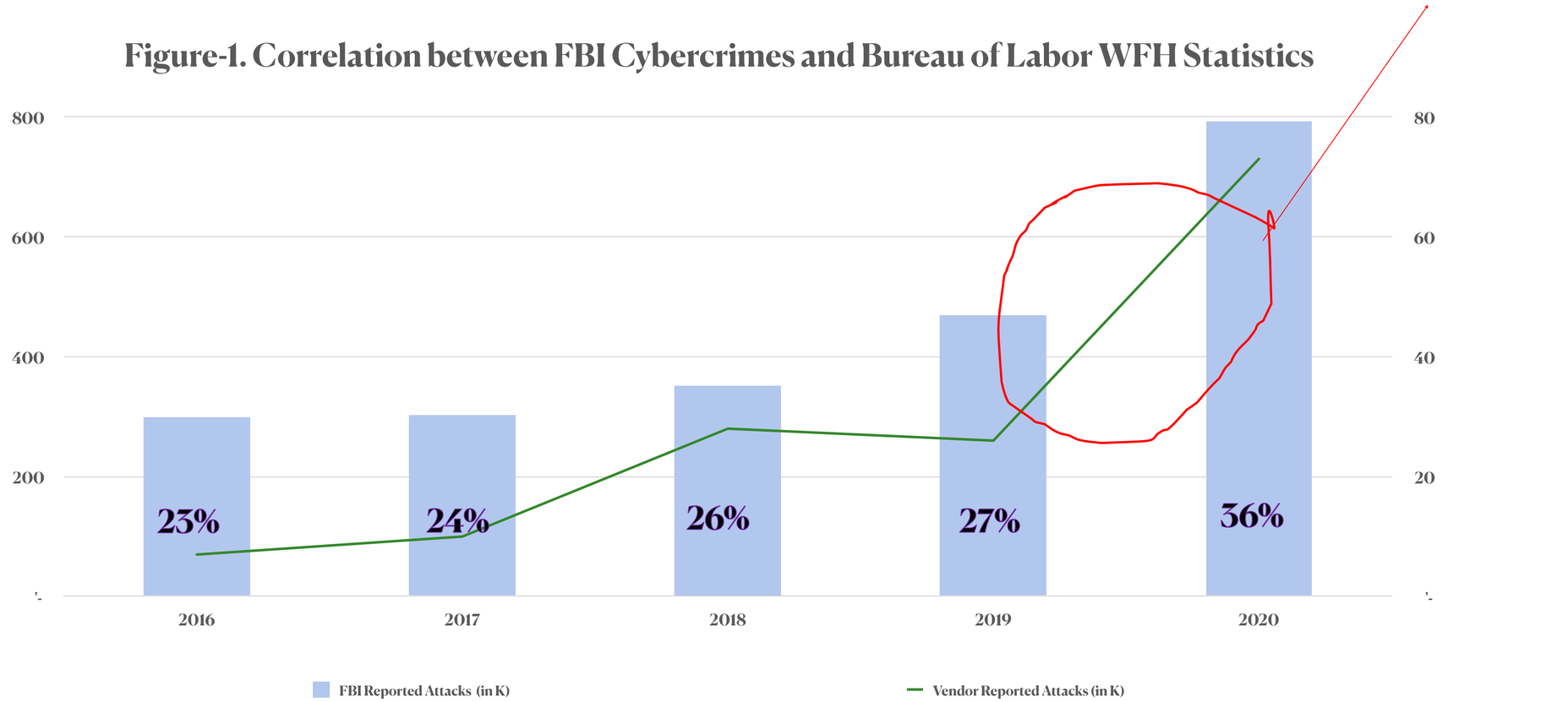
Structure

1. Importance of protecting endpoints
2. Understanding endpoint controls
3. Introduction to research methodology

Problem

Fact

Attacks in 2020 = Attacks in 2018 + 2019



Numbers (in purple) = percentage of employees working from home

So what? FBI claims $13.3B losses in the last 5 years.

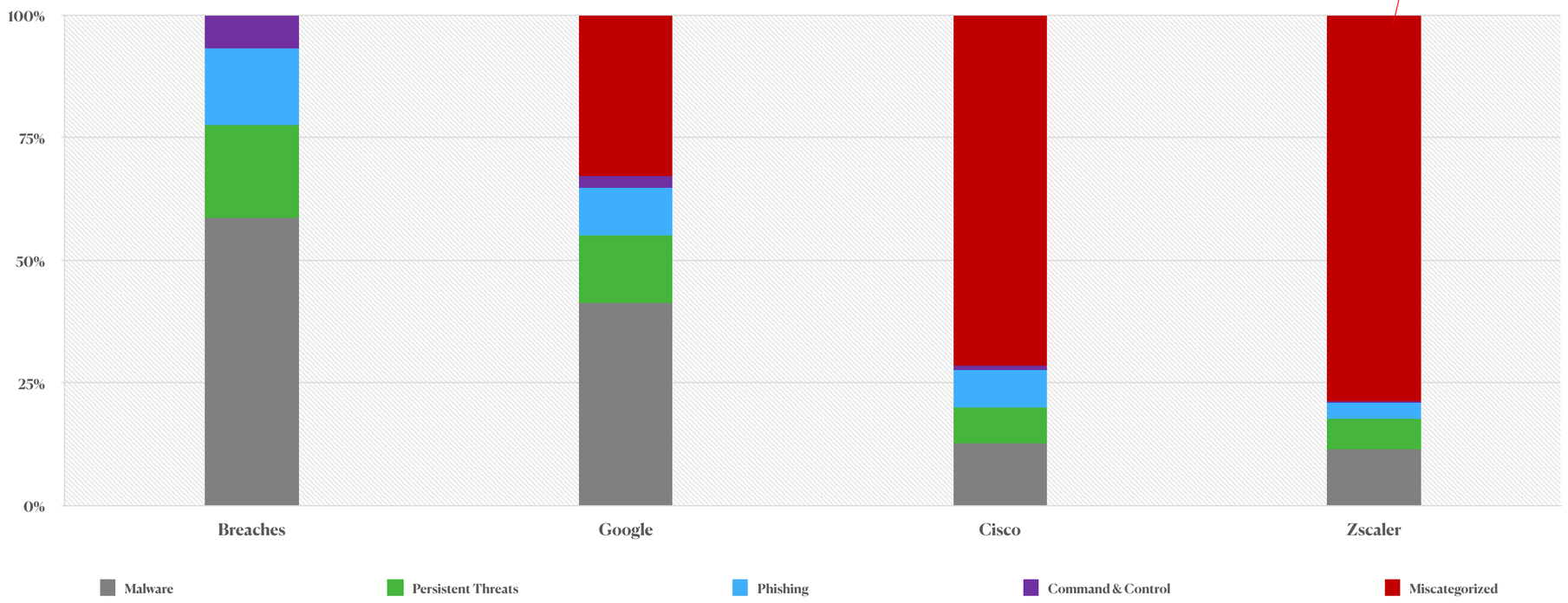
Evidence

Organization	Year	Initial Access	Elevated Access	Expanded Access	Public Access
Target	2013	Phishing	Malware	Botnet	Exfil over Web
Sony	2014	Spear Phishing	Pass the Hash	Command & Control	Exfil over Web
Yahoo	2014	Spear Phishing	Malware	Lateral Movement	Exfil over Web
Anthem	2014	Spear Phishing	Malware	Lateral Movement	Exfil over Web
U.S. OPM	2014	Phishing	Malware	Command & Control	Exfil over Web
RUAG	2015	Watering Hole	Malware	Botnet	Exfil over Web
Tesla	2018	Insider	NA	NA	Exfil over Web
Capital One	2019	Insider	NA	NA	Exfil over Web
G.E.	2020	Insider	NA	NA	Exfil over Web
SolarWinds	2020	Supply Chain	NA	Command & Control	Exfil over Web

Challenge

Red means failure to block a breach

Figure-2. Predicting Data Breaches is a Challenging Problem



So what? Even with cybersecurity controls, it is hard to prevent data breaches

Problem Statement

Problem Statement – Definition

First step in Research Formulation is to address the following questions:

What is the problem that is being solved?

Why is problem important and its solution non-trivial?

Problem statement captures the issue you plan to research and its importance (i.e. “so what”).

Problem Statement structure:

Problem Statement = Issue + “so what”

Issue

- The thing at your profession that is bothering you

”So What”

- What is the consequence of the issue?
- Justify the need
- Why is the issue important?

Problem Statement – Checklist

Clear

It has a clear issue

It has a clear “so what”

It does not have acronyms or technical jargons

It is written so people not familiar with your business are not lost

Concise

It is a single sentence

It is 30 words or less

Specific

It does not use vague terms

Single Issue

It focus on a single issue

Is your issue many issues?

The aim here is to drill down to root cause

Problem Statement

A sharp rise in telecommuting has exposed corporations to an increasing number of data breaches, with each breach costing an average of $X in the U.S.

Thesis Statement – Formulation

Single ***clear*** and ***concise*** sentence

(i.e. 30 words or less)

Clear

- Aim to make your TS easy to understand
- Does not require expertise in a particular field to be understandable
- Avoids acronyms and jargons
- https://expresswriters.com/writing-clear-sentences

Concise

- Long sentences or a paragraph muddles the idea and loses the reader

Uses ***specific*** language

Do not vague terms (e.g. most, best, etc.)

Explicitly states the research product

TS needs to contain a clear deliverable.

For example: Decision Support Tool to help…, Predictive model to ..

A TS can always will be refined as you go deeper into your research.

Thesis Statement - Checklist

It is clear

It states your claim

It has a clear deliverable

It does not use acronyms or vague terms

It does not include technical jargon

It is in layman's terms

Does not require expertise in a field to be understandable

It is concise

It is a single sentence

It is 30 words or less

It is Specific

Uses specific language and not vague terms (e.g. most, best, etc.)

Ties back to your problem statement

Thesis Statement

A predictive model using the domain risk factors to identify malicious communication is required to reduce corporate data breaches.

Research Objectives

Research Objectives are steps you need to take to answers your research questions

Objectives are statements (not questions) of what you intend to do.

Typically start with the infinitive verb.

"to examine" or "to evaluate", "to identify", "to assess", "to measure", "to compare", "to collect", "to analyze"

We typically have several objectives.

Research Objectives – Explanation 1

- **RO1**: To improve the prediction accuracy of data breach controls to block malicious communication given domain risk factors (e.g., popularity, suspicions, citations).
- **RO2**: To improve the prediction accuracy of data breach controls to block malicious communication given employee risk factors (e.g., background, behavior, associations).
- **RO3**: To improve the prediction accuracy of data breach controls to block malicious communication given content risk factors (e.g., PII, sentiment, malware signatures).
- **RO4**: To reduce data breaches by developing a predictive model capable of integrating domain, employee, and content risk factors.

How to improve data breach prediction?

- *santander.com* bank sent low-cash email.
- *santander.com* hack is reported by News.

Date of Breach: Jan 5th, 2021 (Santander Data Breach)

Decide based on information about “domain (where)” from various authoritative sources.

- Dad posted new patent info on *github.com*.
- TV posted the same patent on *github.com*.

Date of Breach: Aug 8th, 2020 (Github Data Breach)

Make a decision based on information about the “user (who)” is browsing the Internet.

- Employee posted PII data to AWS *S3*.
- Employee posted excessive PII data to *S3*.

Date of Breach: July 17th, 2019 (Capital One Data Breach)

Make a decision based on information about “content (what)” to/from the Internet.

Research Objectives – Explanation 2

- **RO1**: To improve the prediction accuracy of data breach controls to block malicious communication given domain risk factors (e.g., popularity, suspicions, citations).

- **RO2**: To improve the prediction accuracy of data breach controls to block malicious communication given employee risk factors (e.g., background, behavior, associations).

- **RO3**: To improve the prediction accuracy of data breach controls to block malicious communication given content risk factors (e.g., PII, sentiment, malware signatures).

- **RO4**: To reduce data breaches by developing a predictive model capable of integrating domain, employee, and content risk factors.

Figure-4: Communication Analogy

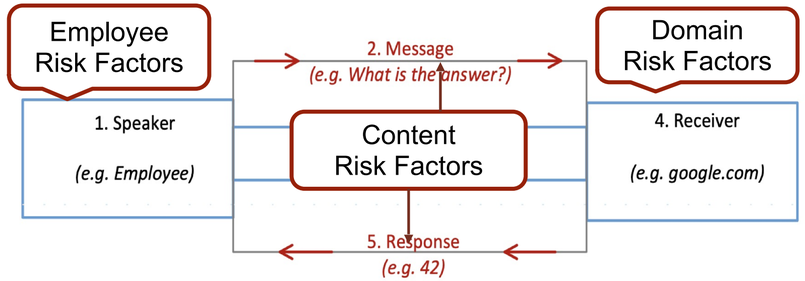
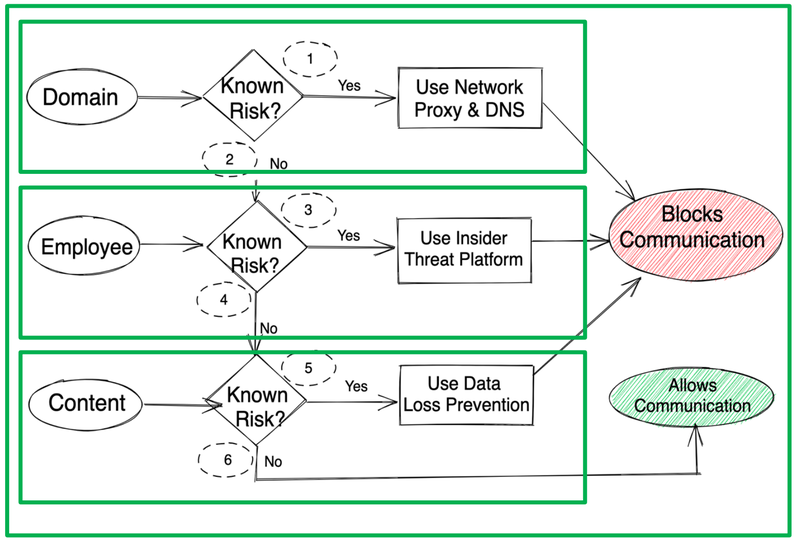


Figure-5: Thesis Statement as Flow Chart



THE GEORGE WASHINGTON UNIVERSITY
WASHINGTON, DC

Research Questions & Hypotheses

Research Questions

- **RQ1**: What is the relationship between domain risk factors (e.g., popularity, suspicions, citations) and data breaches?
- **RQ2**: What is the relationship between employee risk factors (e.g., background, behavior, associations) and data breaches?
- **RQ3**: What is the relationship between content risk factors (e.g., personal identifiable information, sentiment, malware signatures) and data breaches?
- **RQ4**: Can a predictive model be developed using the domain, employee, and content risk factors to reduce data breaches?

Research Hypotheses

- **H1**: Using domain risk factors improves the prediction accuracy of data breaches.
- **H2**: Using employee risk factors improves the prediction accuracy of data breaches.
- **H3**: Using content risk factors improves the prediction accuracy of data breaches.
- **H4**: A predictive model using the domain, employee, and content risk factors reduces data breaches.

Hypotheses

Hypothesis 1 (H1)
H1: Using domain risk factors improves the prediction accuracy of data breaches.
Independent Variables / Input: Domain name. **Attributes/Data sources**: (1) Popularity ranks from data source: Alexa, Majestic, Umbrella, (2) Number of suspicion report attributes from data source: VirusTotal (3) Number of domain registration attributes from data source: WHOIS (4) Probability of a domain being computer-generated from data source: domain generation algorithm (5) Website traffic, citations, and adult content attributes from data source: Alexa web analytics (6) Number of attributes from commercially curated threat intelligence data source: Anomali
Dependent Variables: Binary classification of network communication as benign or malicious
Testable: H_0: Prediction Accuracy (without Domain Risk Factors) < Prediction Accuracy (With Domain Risk Factors)

Hypothesis 2 (H2)
H2: Using employee risk factors improves the prediction accuracy of data breaches.
Independent Variables / Input: Employee name. **Attributes/Data sources**: (1) Number of attributes such as criminal records from data source: Background checks (2) Number of attributes such as aggression from data source: Behavioral indicators (3) Number of attributes such as divorce from data source: Personal indicators (4) Number of attributes such as anti-malware alerts from data source: Endpoint indicators (5) Number of attributes such as correspondence with competitors from data source: Network indicators (6) Number of attributes such as authentication failures from data source: Service indicators
Dependent Variables: Binary classification of network communication as benign or malicious
Testable: H_0: Prediction Accuracy (without Employee Risk Factors) < Prediction Accuracy (With Employee Risk Factors)

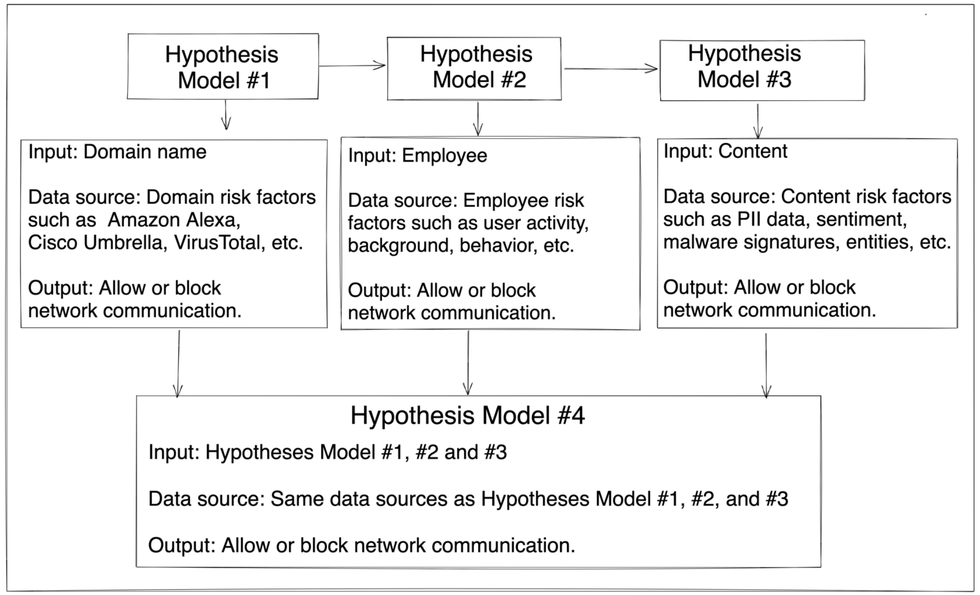
Hypotheses

Hypothesis 3 (H3)
H3: Using content risk factors improves the prediction accuracy of data breaches.
Independent Variables / Input: Network Content **Attributes/Data sources**: (1) Personal identifiable information (PII) indicators from service: AWS Comprehend (2) Key entities match score from service: AWS Comprehend (3) Sentiment Analysis score from service: AWS Comprehend (4) Malware content signature from data source: VirusTotal (5) Intent modeling and match score from service: Python ScaPy (6) Website content filter from data source: ZScaler
Dependent Variables: Binary classification of network communication as benign or malicious
Testable: H_0: Prediction Accuracy (without Content Risk Factors) < Prediction Accuracy (With Content Risk Factors)

Hypothesis 4 (H4)
H4: A predictive model using the domain, employee, and content risk factors reduces data breaches.
Independent Variables / Input: Domain name, Employee name, and Network content **Attributes/Data sources**: (1) Hypothesis 1 output (2) Hypothesis 2 output (3) Hypothesis 3 output
Dependent Variables: Binary classification of network communication as benign or malicious
Testable: H_0: Prediction Accuracy (without domain, employee, and content Risk Factors) < Prediction Accuracy (without domain, employee, and content Risk Factors)

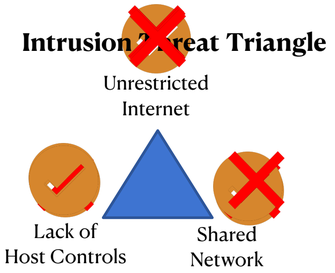
How are Hypotheses related?

- **H1**: Using domain risk factors improves the prediction accuracy of data breaches.
- **H2**: Using employee risk factors improves the prediction accuracy of data breaches.
- **H3**: Using content risk factors improves the prediction accuracy of data breaches.
- **H4**: A predictive model using the domain, employee, and content risk factors reduces data breaches.



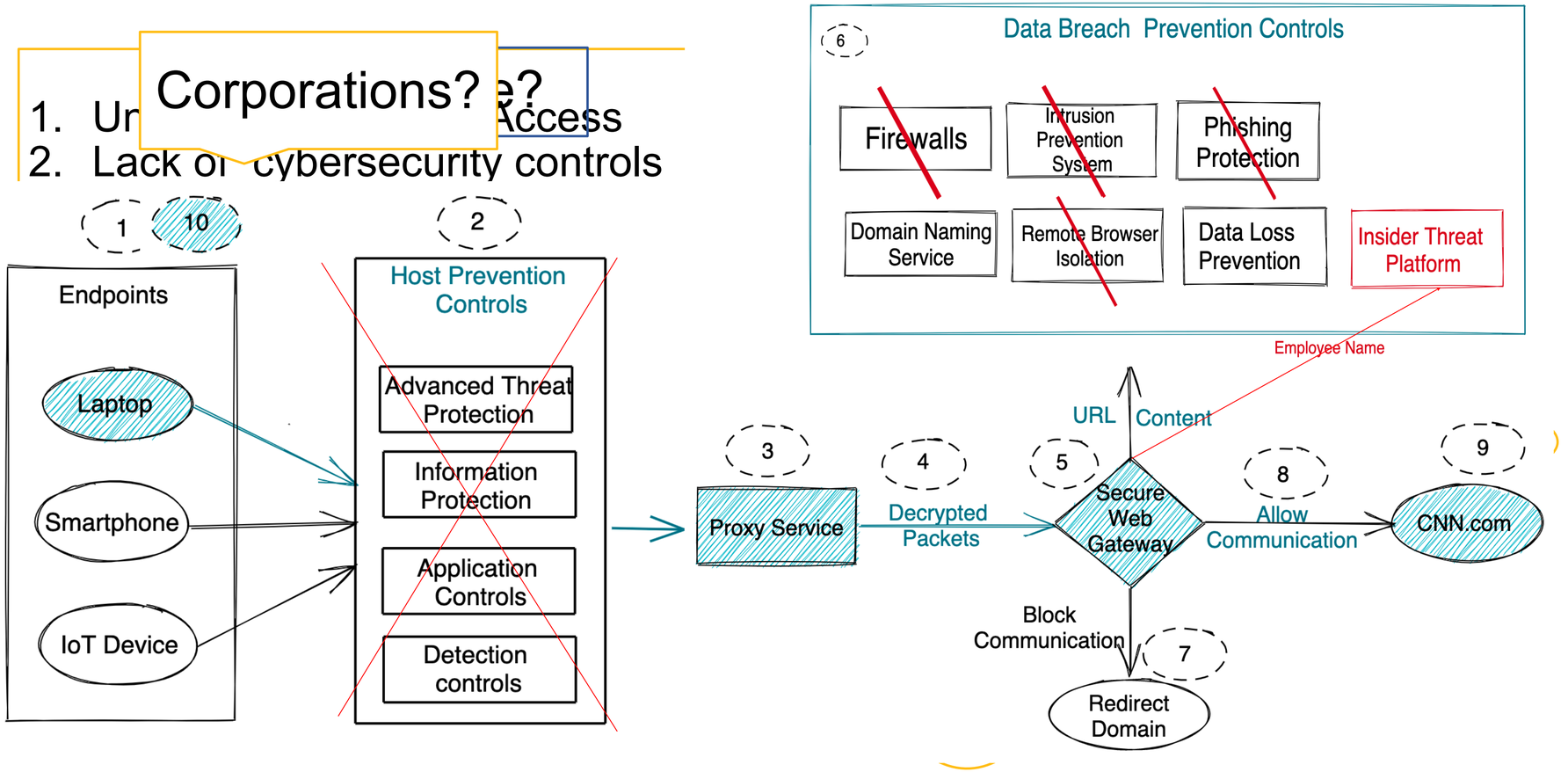
Proposal

- How does browsing a website work?



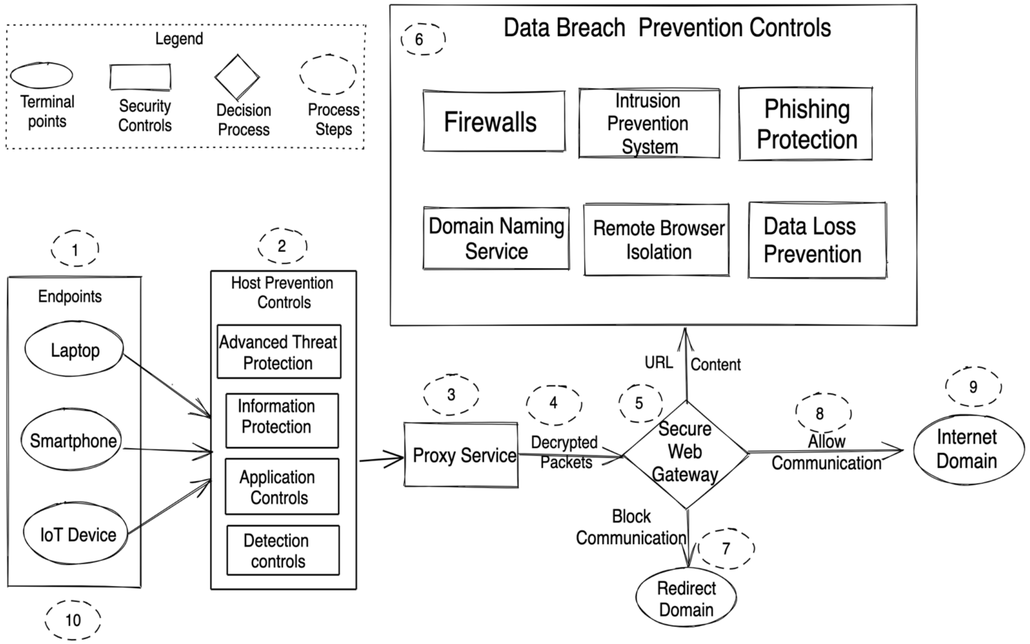
Corporations?

1. Ur Access
2. Lack of cybersecurity controls



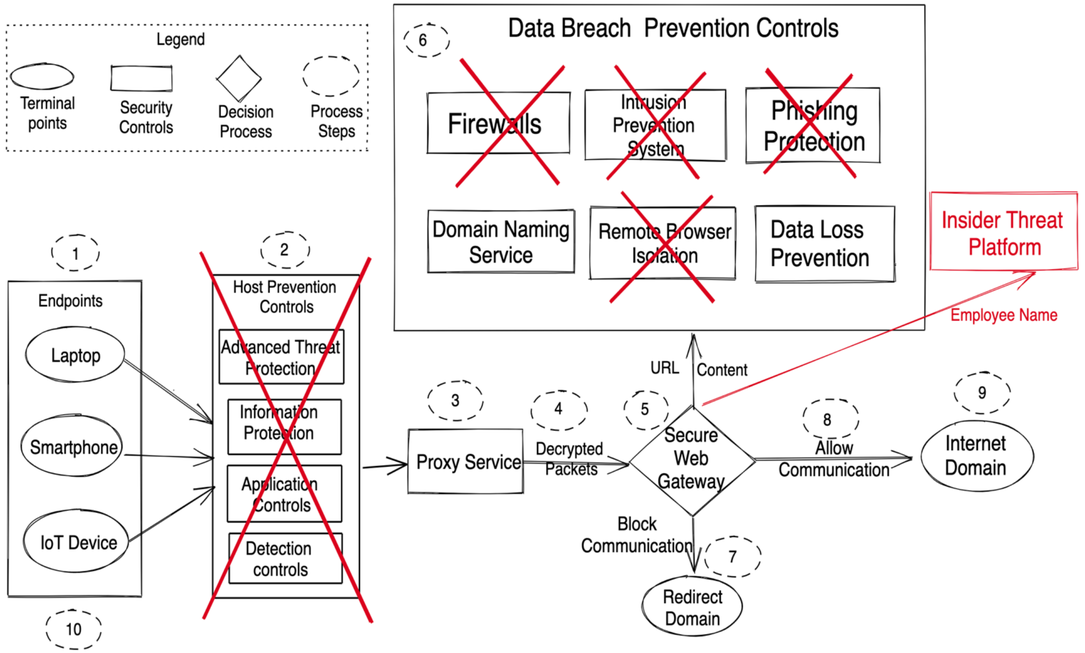
Proposed WFH Architecture

Typical Data Breach Architecture for WFO



✗ Unrealistic for WFH

Proposed Data Breach Architecture for WFH



✓ Feasible for WFH

Literature Review

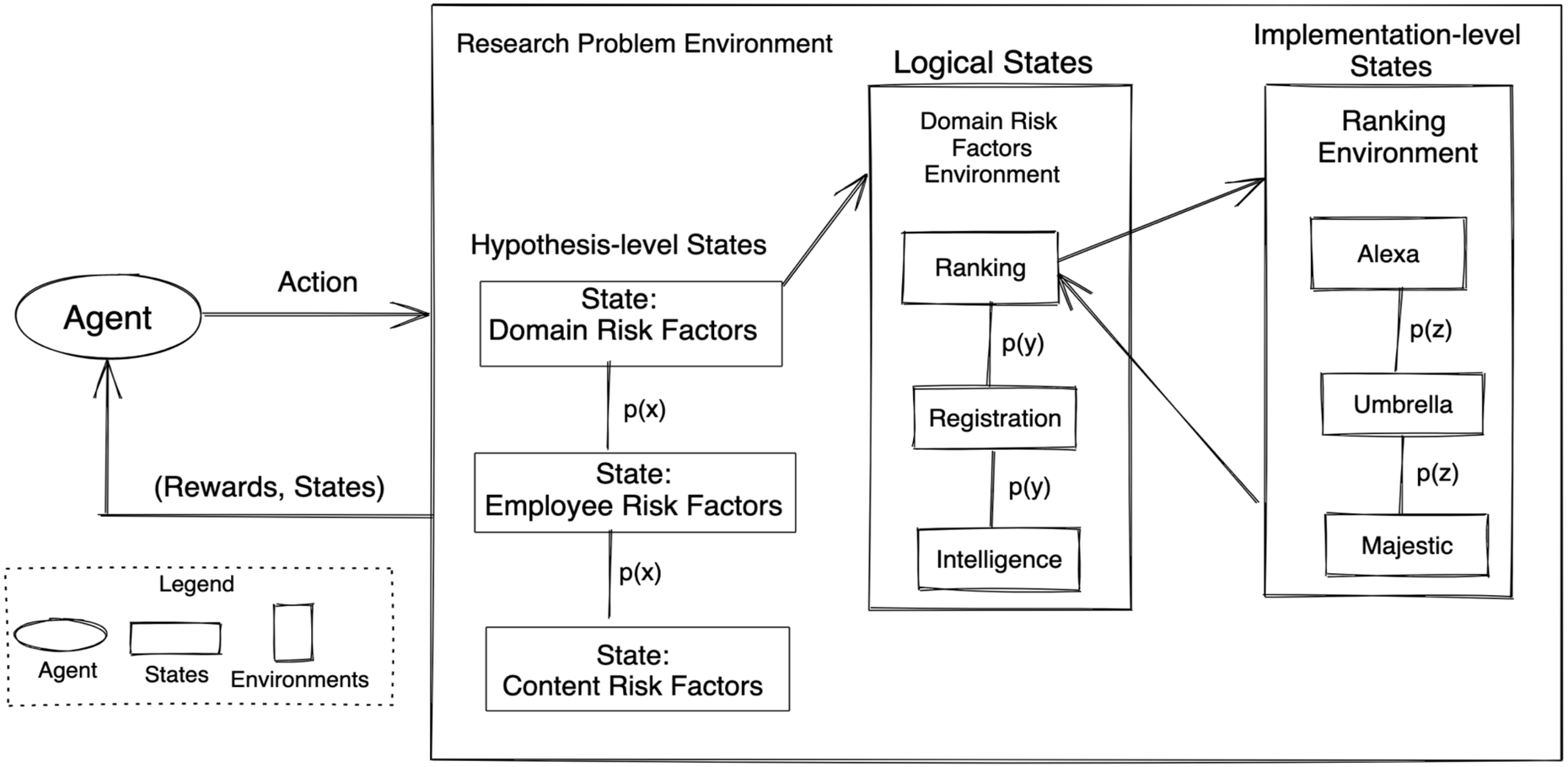
Literature Review

Paper	Contributions	Gaps
Bul'ajoul et al., 2019	• Intrusion Prevention System (IPS) with a signature and anomaly-based models increase data breach prediction accuracy.	• IPS that rely on supervised learning models yield poor accuracy in detecting data breaches with varying exploitation patterns.
Kaur et al., 2017	• Data Leak Prevention (DLP) using Markov decision processes improves accuracy of detecting breaches with small and slow data leaks.	• DLP yields high false positives when small and slow leaks involve changing destinations, common with malware-based breaches.
Duff., 2020	• The endpoint threat protection platform (EPP) malware tactics dataset improves data breach prediction.	• The malware dataset's deceptive tactics (e.g., file deletion) yield a high false-positive rate with EPP.
Sun et al., 2020	• Domain Naming Service (DNS) using graph convolutional network improves data breach domain prediction accuracy.	• DNS with graph convolutional model yields high false positives for newly registered and unpopular domains.
Nicolaou et al., 2020	• Insider threat platform (InTP) using a psychology-inspired anomaly detection model improves data breach prediction.	• InTP using anomaly detection model raises the false positive rate with change in employees work environment.

Research Methodology

Methodology in Action

(Sum of Domain



+ γ (Sum of Cont

Methodology

- What is the summary of input, output, data sources, and methodology?

Hypotheses Model	Research Question	Input Variables	Data Sources	Methodology	Output Variable
Hypothesis-1 Model	RQ1	Domain Name	Domain risk factors	Reinforcement Learning using Q-Learning Algorithm	Binary classification (allow or block network session)
Hypothesis-2 Model	RQ2	Employee Name	Employee risk factors		
Hypothesis-3 Model	RQ3	Network Content	Content risk factors		
Hypothesis-4 Model	RQ4	Domain name, Employee name, Network Content	H1, H2, and H3 models' output		
Base Model	-	Domain Name, Network Content	-	Ensemble Learning (Zscaler's Website)	

How did hypotheses models perform?

1. Performance Metrics for H1 Model

Metric	Benign domain prediction	Malicious domain prediction
Precision	0.54	0.89
Recall	0.73	0.77
Specificity	0.77	0.73
Accuracy	0.76	0.76
F-Score	0.62	0.83
AP Score	0.46	

2. Performance Metrics for H2 Model

Metric	Benign domain prediction	Malicious domain prediction
Precision	0.99	0.84
Recall	0.49	0.99
Specificity	0.99	0.49
Accuracy	0.86	0.86
F-Score	0.66	0.91
AP Score	0.62	

3. Performance Metrics for H3 Model

Metric	Benign domain prediction	Malicious domain prediction
Precision	0.98	0.85
Recall	0.98	0.85
Specificity	0.85	0.98
Accuracy	0.97	0.97
F-Score	0.98	0.85
AP Score	0.98	

4. Performance Metrics for H4 Model

Metric	Benign domain prediction	Malicious domain prediction
Precision	0.98	0.99
Recall	0.98	0.99
Specificity	0.99	0.98
Accuracy	0.99	0.99
F-Score	0.98	0.99
AP Score	0.98	

How model evaluation answers hypotheses?

	Precision	Recall	Specificity	Accuracy	F1-Score	Hypotheses Interpretation
Base Model (ZScaler)	97%	26%	98%	45%	42%	-
H1 Model	89%	77%	73%	76%	83%	H1 is not rejected
H2 Model	84%	99%	49%	86%	91%	H2 is not rejected
H3 Model	85%	85%	98%	97%	85%	H3 is not rejected
H4 Model	99%	99%	98%	99%	99%	H4 is not rejected

H4 Model yielded relatively best performance to reduce data breaches.

What are McNemar's Test Assumptions?

Hypotheses	Statistical Test	Reasoning	What are the assumptions?	How are the assumptions met?
H1, H2, H3, H4	McNemar's Chi-Square Test	The test needs to determine if the difference between paired samples with binary results is statistically significant. <u>How was the test selected?</u> **Question 1:** What is the dependent and independent variable data type used in statistical testing? *Nominal dichotomous (aka binary).* **Question 2**: What is the statistical test objective? *Comparison.* **Question 3:** Are the comparison datasets independent or dependent? *They are dependent (i.e., paired data sample).*	1. Exposure and the output variables type should be binary. 2. Data samples to compare should be paired. 3. Data should be a representative sample of the population.	1. Both exposure and output variable values are binary (i.e., benign or malicious). 2. The same independent variable is evaluated with two different models. 3. All cases for validation are selected using stratified random sampling.

Is validation dataset random?

Domain type	# of domains	Source	Confidence in the domain type	Severity	Expected classification
Advanced persistent threats (APT)	1119	Anomali	> 90%	Very high	Blocked
Command and control (C2)	2590	Anomali	> 90%	Very high	Blocked
Fraud (e.g., spam, sinkhole)	2275	Anomali	> 99%	Very high	Blocked
Malware	2672	Anomali	> 90%	Very high	Blocked
Newly registered	2433	WHOIS	> 99%	Very high	Blocked
Popular	2534	Alexa	> 99%	Not relevant	Allowed
Unpopular	2448	WHOIS	> 90%	Not relevant	Allowed

How to calculate evaluation and validation?

Confusion Matrix

	Actual positive	Actual negative
Predicted positive	True positive (TP)	False positive (FP)
Predicted negative	False negative (FN)	True negative (TN)

McNemar's Contingency Matrix

	Model 2 Correct Predictions	Model 2 Incorrect Predictions
Model 1 Correct Predictions	a	b
Model 1 Incorrect Predictions	c	d

Where:

a = (concordant) Both models **correctly** predicted actual values
b = (discordant) Model 1 predicted **correctly** and Model 2 **incorrectly**
c = (discordant) Model 1 predicted **incorrectly** and Model 2 **correctly**
d = (concordant) Both models **incorrectly** predicted actual values

H0 => P(b) == P(c)

If b+c <= 25, use Binomial CDF i.e., 2 x P(num of "r" heads in "n" tosses)

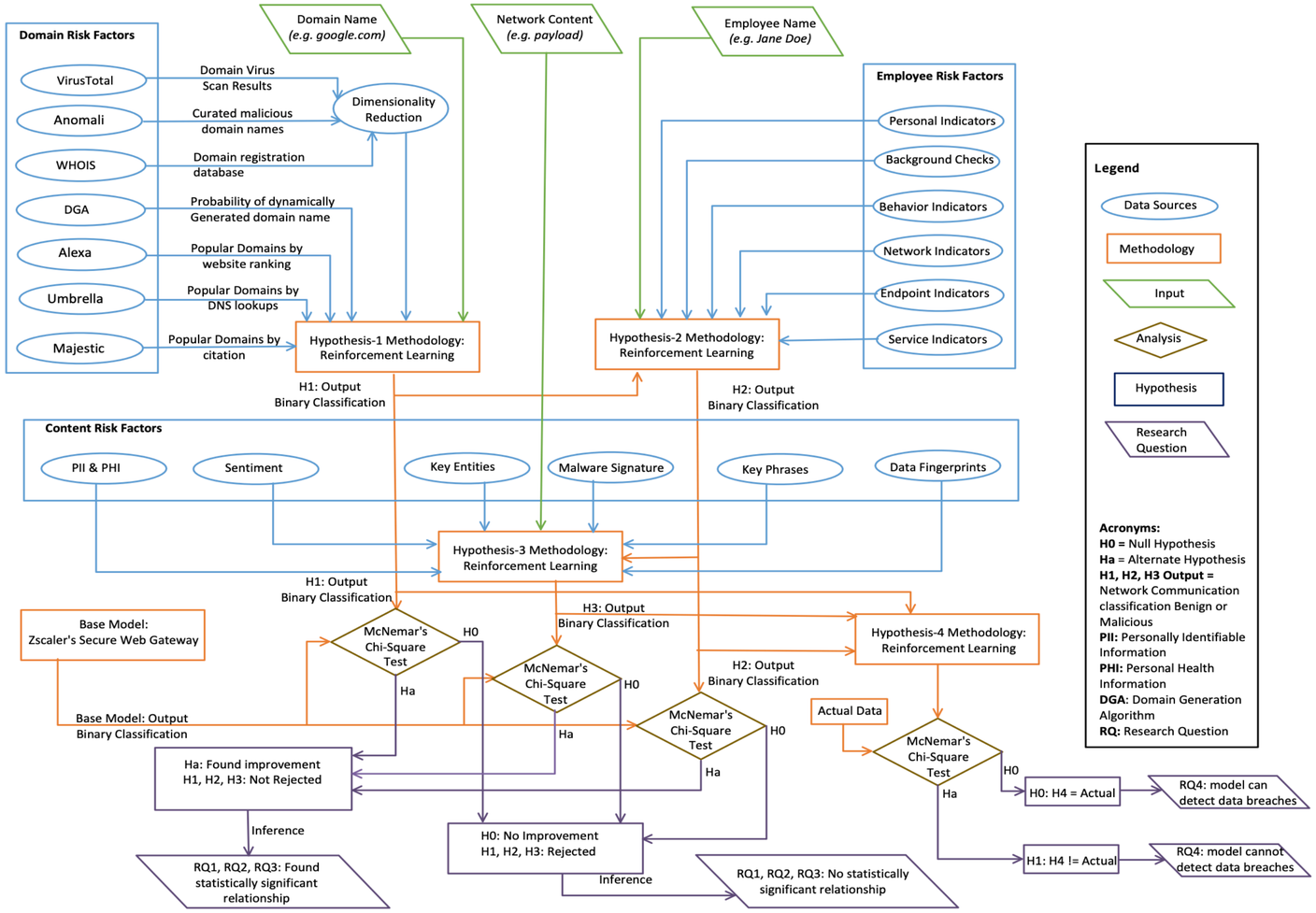
What is the evaluation criteria?

Hypotheses	Evaluation criteria	Ideal results
Base Model (ZScaler) & H1 Model	Malicious domains such as APT, C2, fraud, malware are blocked. Popular and unpopular domains are allowed.	4,982 domains are allowed 13,927 domains are blocked
H2 Model	Malicious domains that are allowed by Hypothesis 1 should be blocked.	4,982 domains are allowed 13,927 domains are blocked
H3 Model	Allowed domains by Hypothesis 1 and 2 that contain Personal Identifiable Information content should be blocked.	5,877 domains are allowed 13,032 domains are blocked
H4 Model	Dataset created with a random combination H1, H2, and H3 data.	2226 domains are allowed 16683 domains are blocked

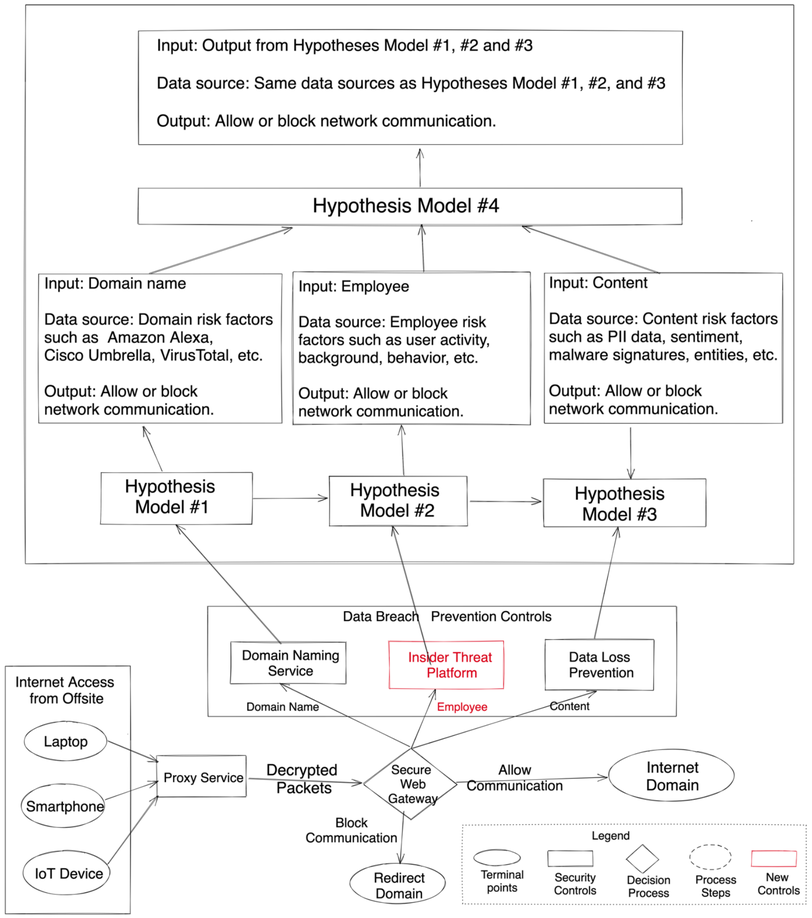
What are the right evaluation metrics?

Performance Metric	Metric suitability for evaluation of imbalanced dataset
Accuracy	Not suitable for heavily imbalanced data. It is easy to get high accuracy by simply classifying all observations as majority class.
AUC-ROC	Not suitable for heavily imbalanced data. The FPR for deeply imbalanced datasets is pulled down due to many TN.
F1-Score	Suitable for heavily imbalanced data because it uses harmonic mean to balance precision and recall
PR-AUC & AP-Score	Suitable for heavily imbalanced data. Because ROC AUC looks at a TPR and FPR, PR AUC looks at positive predictive value and TPR. The precision-recall equation is helpful for imbalanced classes due to the absence of TN in the calculation.
Mathew's Correlation Coefficient (MCC)	Not suitable because the praxis requires comparing base model and proposed models results against expected results. MCC is great when the classifier results need to be evaluated against expected results.

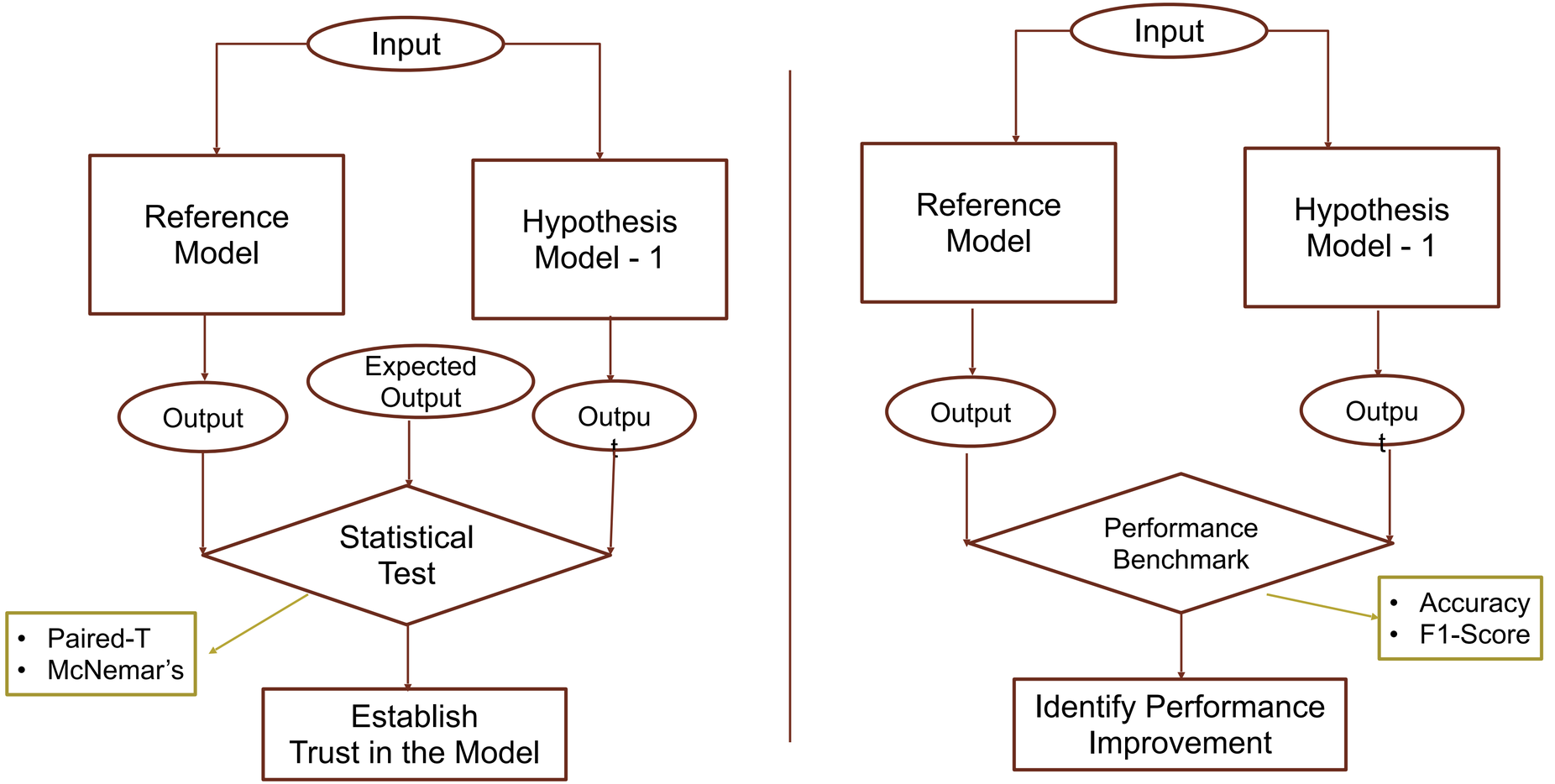
Graphical architecture



How are controls & models related?



What is validation vs. evaluation?



Methodology

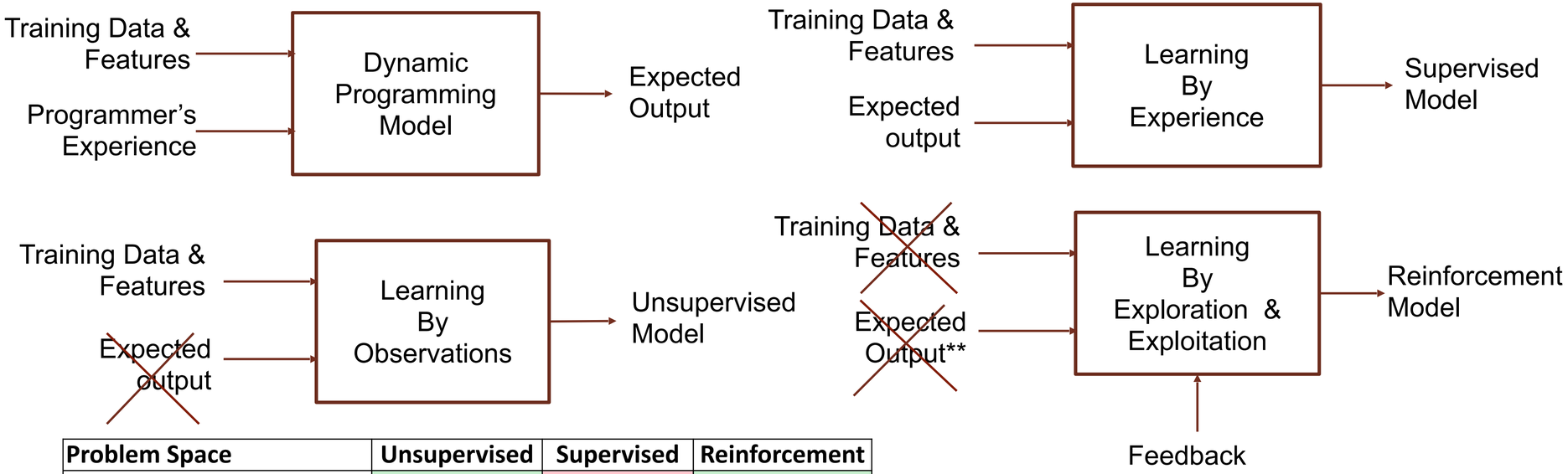
- What are the data sources?

#	Name	Brief Description	R
1	Alexa 1M domains	Alexa generates top 1M websites based on number of visitors	30M
2	Umbrella 1M domains	Umbrella generates top 1M domains based on DNS lookups	30M
3	Majestic1M domains	Majestic generates top 1M domains based on number of citations	30M
4	Anomali & FireEye	Anomali & FireEye are commercial threat intelligence services	700K
5	WHOIS records	WHOIS records indicate when the Internet domains are registered..	700K
6	VirusTotal database	VirusTotal is a Google's online database of threat intelligence	700K

#	Name	Brief Description	R
7	Background Indicators	Criminal, and civil indicators from background checks	5K
8	Personal Indicators	Financial stress and family events from annual tax audits	5K
9	Behavior Indicators	Aggression, excessive activity from co-worker observations	5K
10	Cyber Indicators	Anomalous cyber activity on corporate IT services	10K
11	AWS Services	AWS service returns PII, Sentiment, key Entities	5K
12	Domain category	Zscaler categorizes domain based on web content.	1M

***Legend**:* R = records, M = Millions, K = Thousands; **Color codes**: Grey = Open source; Lavender = Commercial; Yellow = Company

What is Reinforcement Learning?



Problem Space	Unsupervised	Supervised	Reinforcement
Clustering	Yes	No	Yes
Dimensionality reduction	Yes	No	Not optimal
Classification without training	No	Yes	Yes
Constantly	No	Yes	Not optimal
Optimization	No	Not Optimal	Yes
Sequential decision making	No	No	Yes

- **Exploration** helps identify training data.
- **Exploitation** helps identify expected output.

Legend: ** = no labeled output, but the end results is known

Why reinforcement learning?

Hypotheses	Functional Requirements	Technical Requirements	UL	SL	RL
H1: Using domain risk factors improves the prediction accuracy of data breaches.	A sequence of decisions are required to answer if a network connection is benign or malicious.	Support sequential decision making	No	No	Yes
H2: Using employee risk factors improves the prediction accuracy of data breaches.	User's current risk profile and action decides if a network connection benign or malicious, not past actions.	Support Markov Decision Process	No	No	Yes
H3: Using content risk factors improves the prediction accuracy of data breaches.	There is no predefined right or wrong answer for a network connection or a domain to be benign of malicious.	Support for classification without training data	No	No	Yes
H4: A predictive model using the domain, employee, and content risk factors reduces data breaches.	Combined requirements of H1, H2, and H3	Support for H1, H2, and H3 technical requirements	No	No	Yes

Legend: UL = Unsupervised Learning, SL = Supervised Learning, RL = Reinforcement Learning

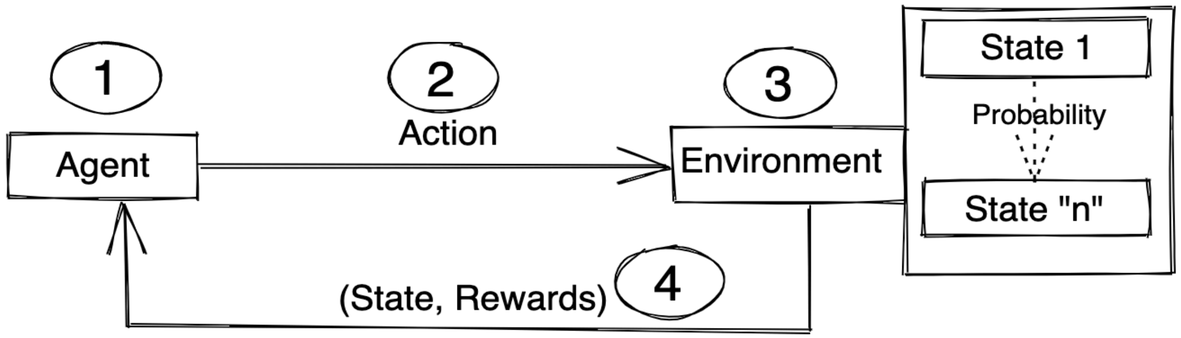
How did RL perform against SL models?

- Which machine learning models do cybersecurity vendors use?
- What is the appropriate machine learning model for H1 model?

Machine Learning Algorithms	Accuracy	Precision	Recall	F1 Score
	Before	Before	Before	Before
Nearest Neighbors	88	86	84	85
Linear SVM	86	91	74	79
Radial Basis SVM	88	86	83	84
Decision Tree	88	85	84	85
Random Forest	86	92	74	78
AdaBoost	88	86	83	84
Neural Networks	88	86	84	85
Q- learning Methodology by Reinforcement Learning (RL)	86	91	74	79

RL is the most appropriate framework when data is missing and/or noisy.

Which problems can be solved by RL?



- Agent = Hypothesis Model
- Action = Allow or Deny access
- Environment = WFH data sources of risk factors
- State = Benign or malicious state from risk factor
- Reward = Feedback from state given an action

What is Q-Learning theory?

- **Reinforcement learning** problem is an undirected graph problem (unlike Bayesian networks that are directed graphs), which requires reaching the terminal state with maximum rewards. Hence, which ever path leads to the highest grand total, that is our solution.

$$G_s = R_s + \gamma R_{s+2} + \ldots + \gamma^{S-1} R_S$$

- Every reinforcement learning problem can be solved using **Markov Decision Process,** which uses current STATE and ACTIONS to predict next state, NOT the previous states. Where every action may return REWARDS.

Reward Calculation Using MDP requires a tuple of (State, Action, Probability of state transition, Reward, Future reward discount)

- **Bellman Optimality Equation** can calculate state**,** but it does not tell you which action to take.

$$V*(s) = \max a \sum_{s'} P(s,\ a,\ s')\ [R(s,\ a,\ s') + \gamma\ .\ V(s')]$$

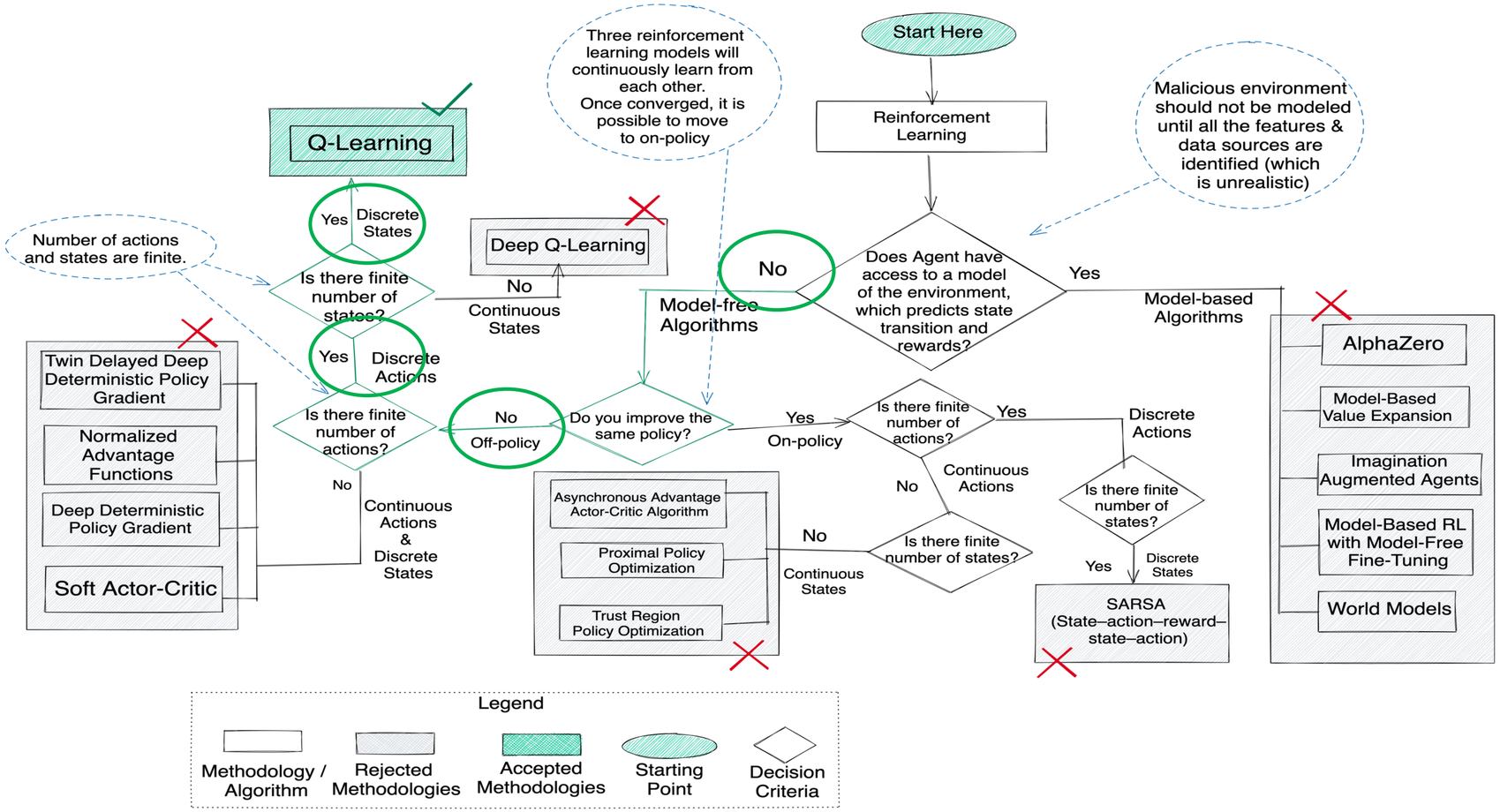
- **Bellman Equation** can calculate **State-Action** and tell you which action to take. Also known as Q-value.

$$Q_{k+1}(s,\ a) \leftarrow \sum_{s'} P(s,\ a,\ s')\ \left[R(s,\ a,\ s') + \gamma\ .\ \max_{a'} Q_k(s',\ a')\right] \text{ for all } (s,\ a)$$

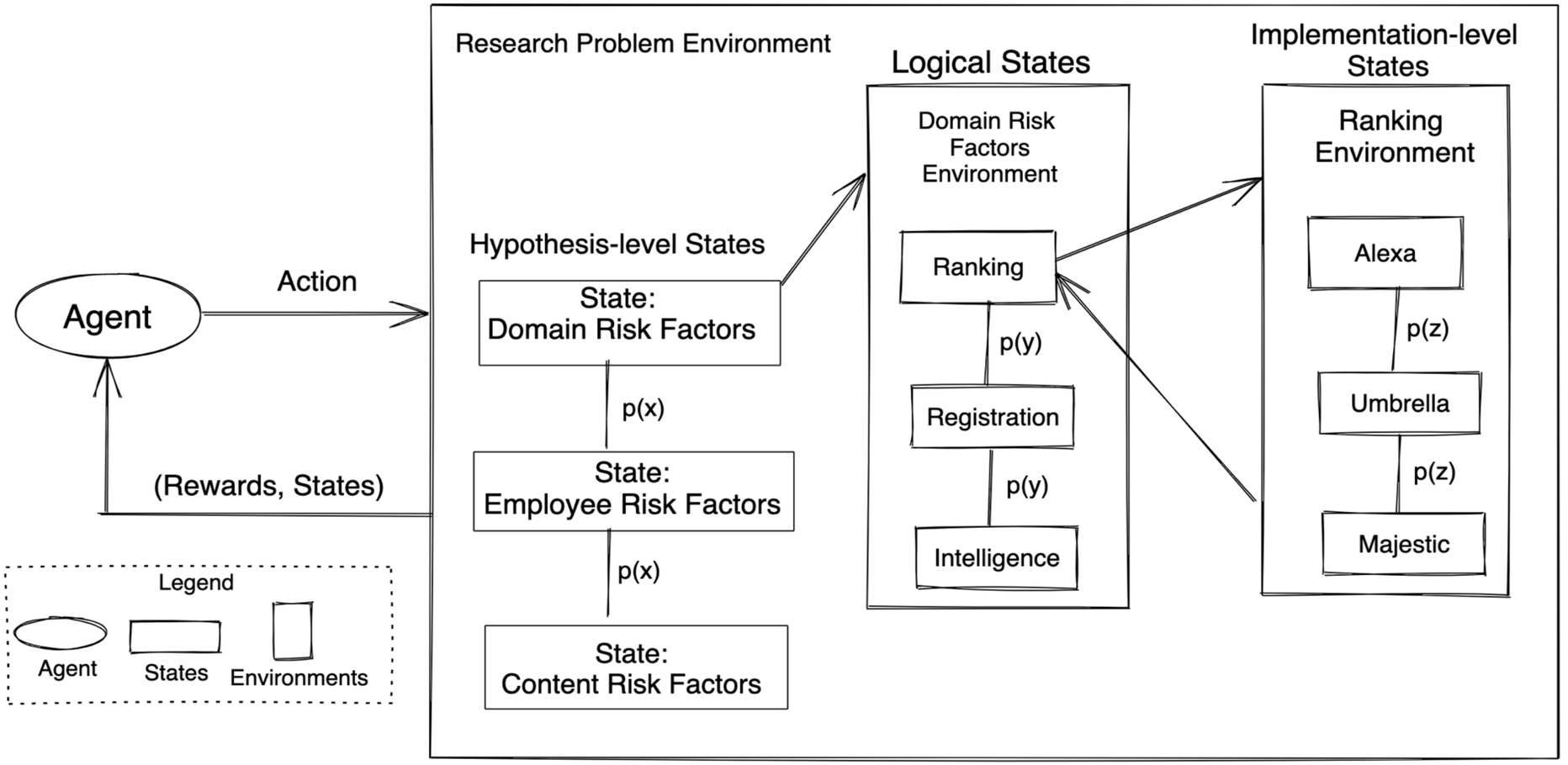
- Now, How to calculate **transition probability between states** $\mathbf{P(s,\ a,\ s')}$?
  - **Episodic learning**: Agent run trials, constantly collecting samples, getting rewards, and thereby evaluating the V or Q functions.
    - **Monte-Carlo**, which run all the possible iterations to calculated expected value.
    - **Temporal-Difference - TD(0),** which only accounts for current value instead of entire chain of sequence.
  - *Monte-Carlo is time consuming to run, but accurate. Temporal-Difference not accurate, but good at estimation.*
- **Q-learning** method uses TD(lambda), which combines Monte-Carlo and TD(0). Instead of running every iteration to generate Q-Table ahead of time. It uses a policy like **Decaying-Epsilon-Greedy**, to keep updating Q-table with the best value from every iteration of decision.

Why Q-learning?

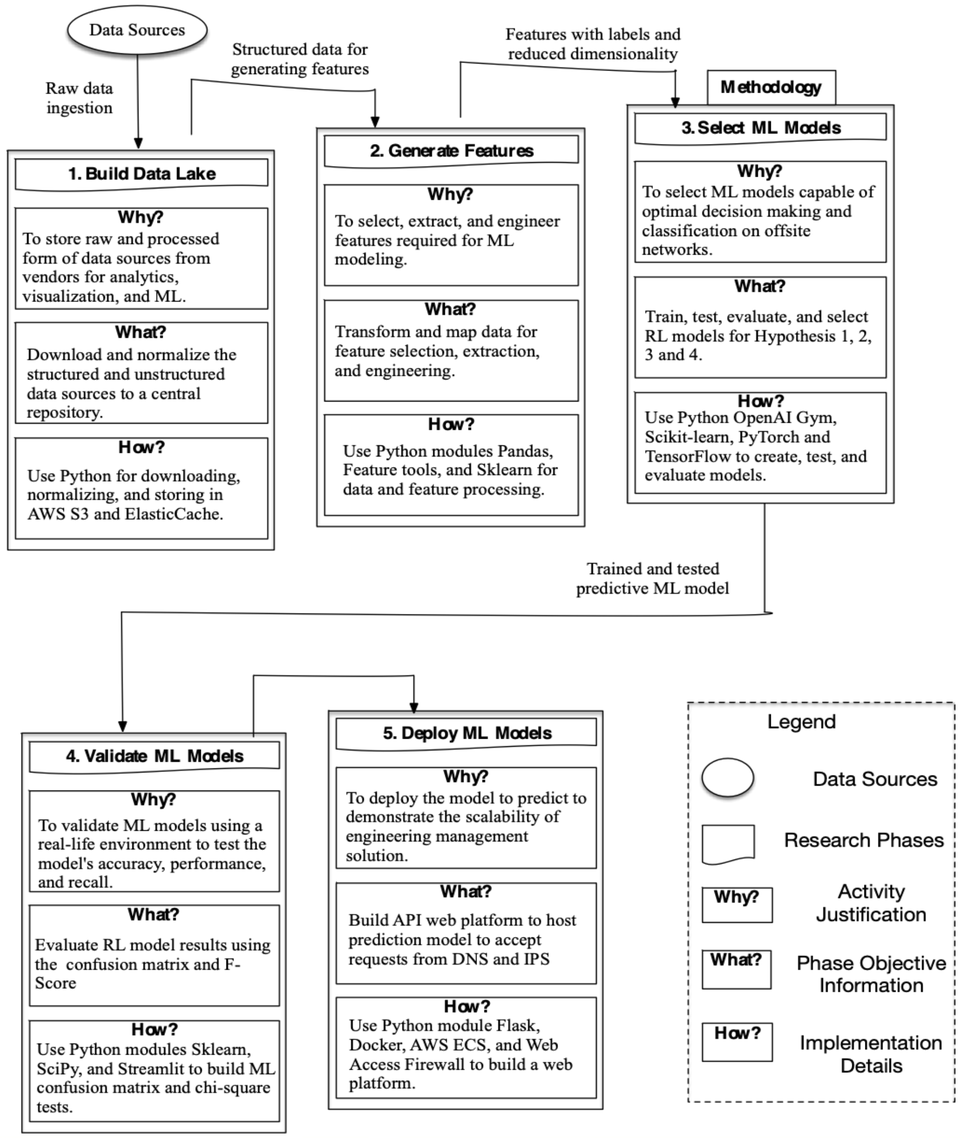
1. Can all possible states of the environment be modeled? No
2. Is probability between the state transition is deterministic? No
3. Are number of actions discrete? Yes
4. Are number of states finite? Yes



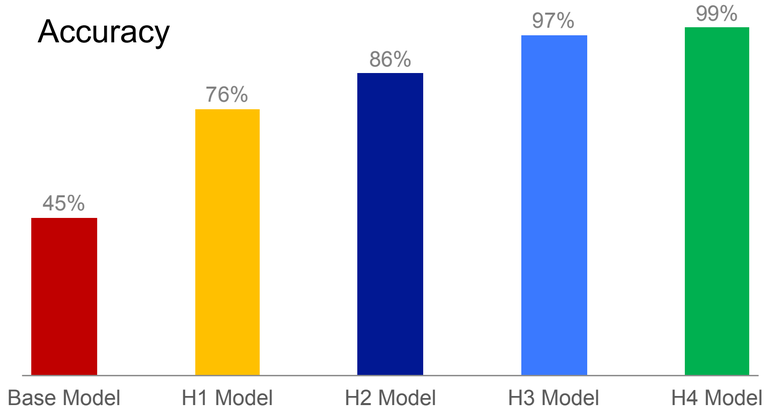
Technical Explanation of Methodology

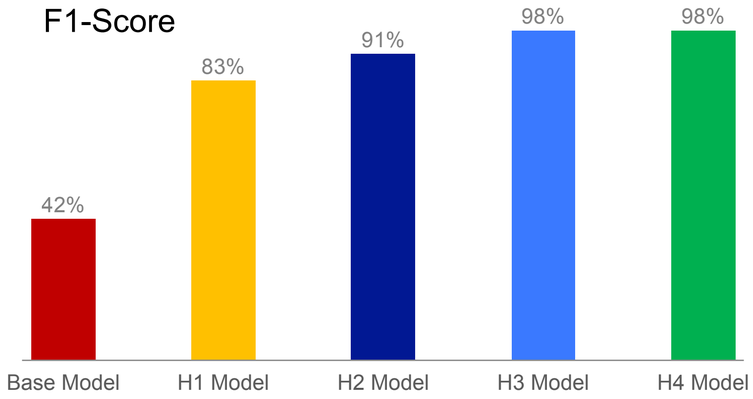


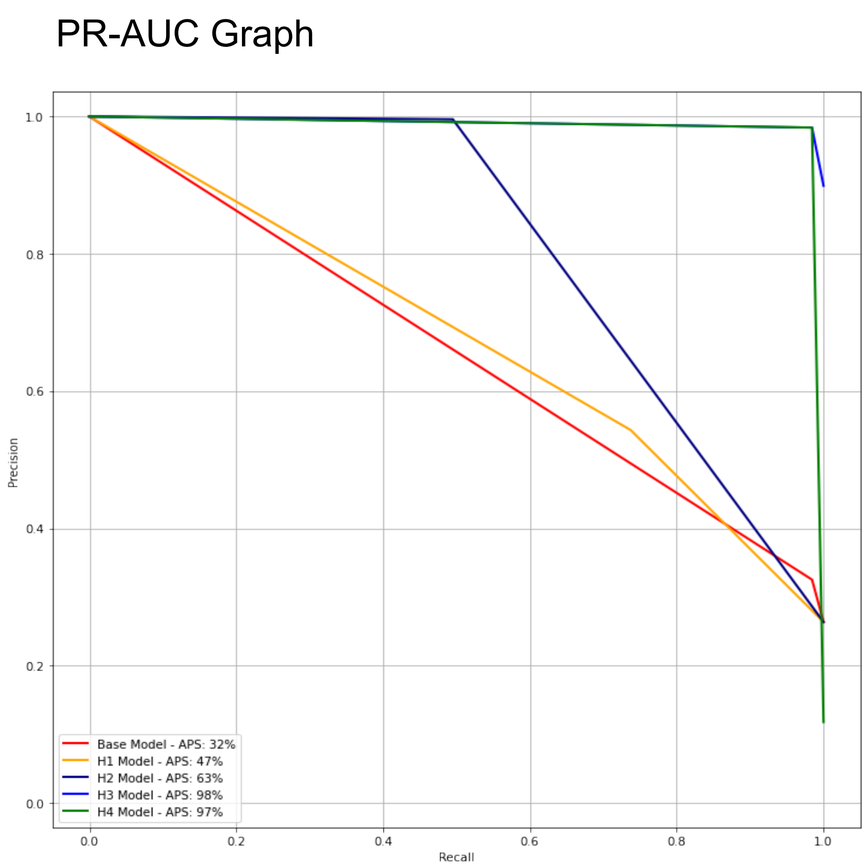
Research Workflow



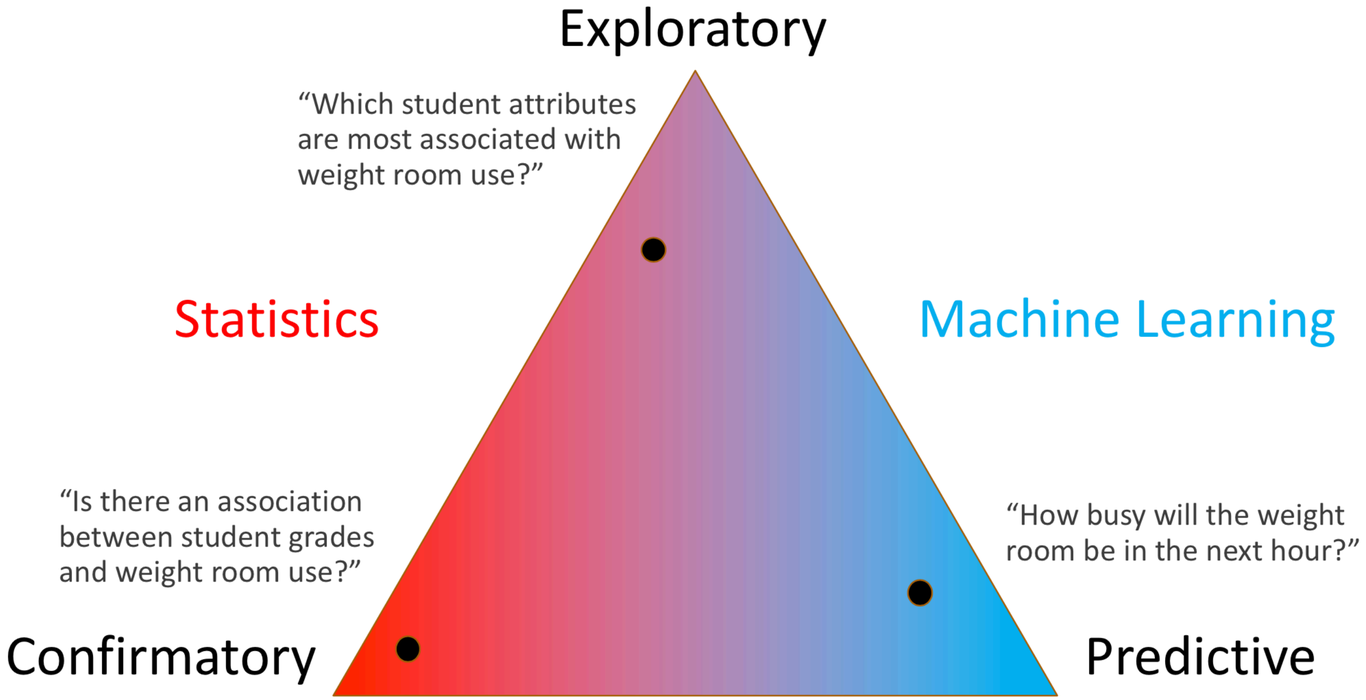
Comparison of Models Performance



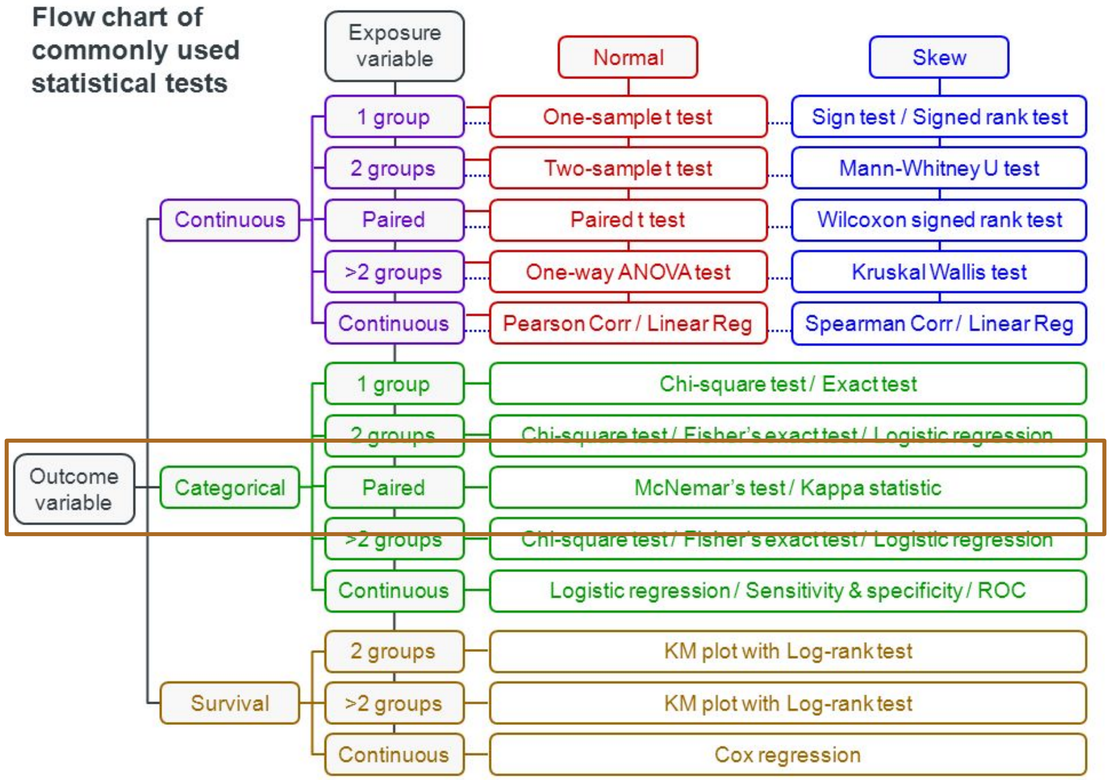


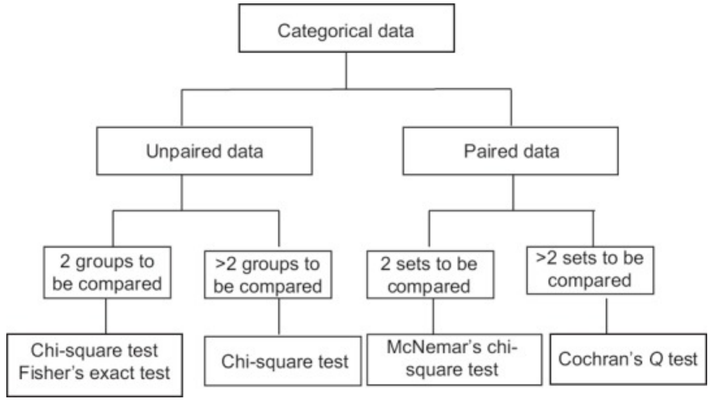


Why do we need to validate a model?



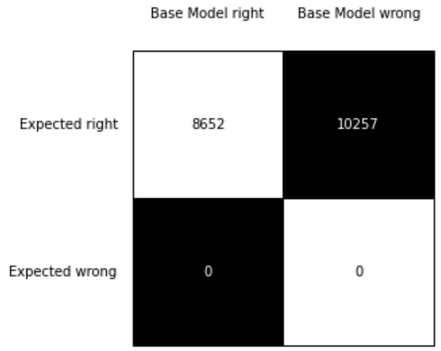
How to validate results?





Choosing Statistical test: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2996580/ & https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4966396/

Result: Base Model Validation



Hypothesis	Comparison	Chi-square Statistic	P-value	Interpretation
H1	Base vs. H1 Model	3824.380252	0.00000	H1 is not rejected
H2	Base vs. H2 Model	4709.031445	0.00000	H2 is not rejected
H3	Base vs. H3 Model	6922.445215	0.00000	H3 is not rejected
H4	Actual vs. H4 Model	0.056338	0.81238	H4 is not rejected
Base	Actual vs. Base Model	10255.000097	0.00000	Base Model is rejected

Result: Cohen-Kappa Evaluation

Cohen's kappa coefficient is a statistic that is used to measure inter-rater reliability for qualitative items. It is generally thought to be a more robust measure than simple percent agreement calculation, as κ takes into account the possibility of the agreement occurring by chance. **However, it CANNOT tell you what is the disagreement about? Is the disagreement due to poor classification of benign or malicious? Without such information, it is hard judge the model's ability to reduce data breaches.**

```
Base vs. h1:  0.18850816152597505
Base vs. h2:  0.06760063352372 48
Base vs. h3:  0.0275192009426308 28
Base vs. h4:  0.060710798021157 53

Actual vs. h1:  0.46239969857792 64
Actual vs. h2:  0.58948753794268 78
Actual vs. h3:  0.84177119486167 72
Actual vs. h4:  0.98192771449251 4
```

Interpretation of Kappa

	Poor	Slight	Fair	Moderate	Substantial	Almost perfect
Kappa	0.0	.20	.40	.60	.80	1.0

Kappa	Agreement
< 0	Less than chance agreement
0.01–0.20	Slight agreement
0.21– 0.40	Fair agreement
0.41–0.60	Moderate agreement
0.61–0.80	Substantial agreement
0.81–0.99	Almost perfect agreement

SL vs. UL vs. RL

	AI Planning	SL	UL	RL	IL
Optimization	X			X	X
Learns from experience		X	X	X	X
Generalization	X	X	X	X	X
Delayed Consequences	X			X	X
Exploration				X	

- SL = Supervised learning; UL = Unsupervised learning; RL = Reinforcement Learning; IL = Imitation Learning
- Imitation learning assumes input demonstrations of good policies
- IL reduces RL to SL. IL + RL is promising area

Results

Organization	Year	Initial Access	Elevated Access	Expanded Access	Public Access
Target	2013	Phishing	Malware	Botnet	Exfil over Web
Sony	2014	Spear Phishing	Pass the Hash	Command & Control	Exfil over Web
Yahoo	2014	Spear Phishing	Malware	Lateral Movement	Exfil over Web
Anthem	2014	Spear Phishing	Malware	Lateral Movement	Exfil over Web
U.S. OPM	2014	Phishing	Malware	Command & Control	Exfil over Web
RUAG	2015	Watering Hole	Malware	Botnet	Exfil over Web
Tesla	2018	Insider	NA	NA	Exfil over Web
Capital One	2019	Insider	NA	NA	Exfil over Web
G.E.	2020	Insider	NA	NA	Exfil over Web
SolarWinds	2020	Supply Chain	NA	Command & Control	Exfil over Web
Colonial Pipeline	2021	Phishing	Ransomware	Command & Control	Exfil over Web

Legend: NA = Not Available (could not find details to simulate); Green = Success; Yellow = Success if Internet download; White = Out of Scope

Results

Summary

Categorical (O)	Categorical (E)	Paired (G)	Comparison (P)

Can you trust the results produced by the model?

Outcome Variable	Exposure Variable	Groups	Purpose	Normality	Statistical Test
		1 group		Not Applicable	Chi-Square Goodness of Fit Test
		2 groups			Fisher's Exact Test (for
Categorical	Categorical	Paired	Comparison		McNemar's Test
		> 2 groups			Independence
Continuous	Continuous	1 group	Comparison	Normal	One Sample T-Test
		2 groups			Two Sample T-Test
		Paired			Paired T-Test
		> 2 groups			One-way ANOVA Test
		1 group		Skew	Signed Rank Test
		2 groups			Mann-Whitney U-Test
		Paired			Wilcox Signed Test
		> 2 groups			Kruskal-Wallis Test
Continuous	Continuous	1 group	Relationship	Normal	Pearson Correlation
		1 group		Skew	Spearman Correlation
Categorical	Continuous	1 group		NA	Logistic Regression
	Categorical	Paired			Cohen's Kappa

Legend	
Outcome Variable	Variable represents the predicted value by statistical model (e.g., predicted height of Eiffel tower is 900 ft)
Exposure Variable	Variable represents the actual value for comparison (e.g., actual height of Eiffel tower is 1063 ft)
Groups	Combination of test subjects and objects.
Purpose	Purpose of confirmatory statistics
Normality	Defines how the data is distributed.
Statistical Test	Names of the statistical tests

Results

- How does model validation answer hypotheses?

	H1 Model Correct	H1 Model Incorrect
Base Model Correct	7095	1557
Base Mode Incorrect	7416	2841

	H2 Model Correct	H2 Model Incorrect
Base Model Correct	6173	2479
Base Mode Incorrect	10210	47

	H3 Model Correct	H3 Model Incorrect
Base Model Correct	4546	1213
Base Mode Incorrect	10040	3110

	H4 Model Correct	H4 Model Incorrect
Actual Correct	16103	34
Actual Incorrect	37	2735

Hypothesis	Comparison	Chi-square Statistic	P-value	Interpretation
H1	Base vs. H1 Model	3824.380252	0.00000	H1 is not rejected
H2	Base vs. H2 Model	4709.031445	0.00000	H2 is not rejected
H3	Base vs. H3 Model	6922.445215	0.00000	H3 is not rejected
H4	Actual vs. H4 Model	0.056338	0.81238	H4 is not rejected

Discussion

- Interpretation of results and hypotheses resolution

Hypothesis (H)	Hypotheses Findings	Research Questions (RQ)	Research Question Findings
H1	Not Rejected	RQ1	Domain risk factors have a statistically significant relationship to improve data breach prediction
H2	Not Rejected	RQ2	Employee risk factors have a statistically significant relationship to improve data breach prediction
H3	Not Rejected	RQ3	Content risk factors have a statistically significant relationship to improve data breach prediction
H4	Not Rejected	RQ4	A predictive model with domain, employee, and content risk factors can reduce data breaches

Conclusions

1. Reinforcement learning models can significantly boost the prediction accuracy than supervised learning models when domain risk factors have noisy or missing data.

2. Employee risk factors are crucial in preventing data breaches when domain risk factors have inconclusive or missing data.

3. Content risk factors are expensive to process in real-time, but they are most reliable in predicting data breaches.

Why is data breach prediction difficult?

1. There is no explicit right or wrong answer.
   - github.com is benign for publishing source code.
   - github.com is malicious for publishing bank details.
2. Requires a sequence of decisions to answer.
   - chase.com is benign as it is your popular local bank.
   - chase.com is malicious as it is reported hacked by WSJ.
3. Current actors makes future decision, not previous actions.
   - cnn.com is benign for a **dad** to write racism remark **yesterday**.
   - cnn.com is malicious for a **child** to write racism remark **today**.

Definitions

Definitions

Backdoor: A secret channel to bypass authentication and authorization to access the asset.

Botnet: A network of devices running malicious programs and communicating over the Internet. They are typically the launchpads of distributed attacks.

Corporate network: A corporate-owned network that is continuously secured, managed, and monitored.

Content risk factors: Content attributes that define the ability and probability of compromise—for example, content type, sensitivity, sentiment, personally identifiable information (PII), and risk levels.

Cyber asset: Valuable information owned or held responsible by the company, such as intellectual property, personally identifiable information, and customer data.

Data breach: Confirmed disclosure of confidential, sensitive, or protected data to an unauthorized environment.

Detection control: A cybersecurity tool or process responsible for identifying a specific event in an environment such as unauthorized access.

Definitions

Domain: The address of a website or a company on the Internet.

Domain risk factors: An Internet domain’s attributes that define the ability and probability of compromise, for example, domain registration date, popularity, and previously reported suspicious activity.

Drive-by download: Occurs when a computer makes an unintentional download (of malware) due to browser, application, or operating system vulnerability.

Eavesdropping: Sniffing for data on unsecured networks such as home, hotel, or guest networks.

Employee risk factors: Employee attributes that define the ability and probability of compromise, for example, skillset, behavior, financial situation, or stress indicators.

Endpoint: A computer device, such as a laptop, desktop, tablet, mobile phone, or virtual machine, at the end of a computer network.

Exfiltration over web: A technique for stealing data using websites through mechanisms such as posting to a blog site, news site, torrent, or dark website.

Firewall: A preventive control responsible for blocking unauthorized and allowing authorized network access.

Forensic control: A cybersecurity tool or process responsible for analyzing a specific event in an environment such as unauthorized access.

Gateway: Hardware or software used on the network to control data flow between discrete networks.

Definitions

Governance: In cybersecurity, governance involves describing policies, processes, and managing risk decisions.

Indicators of attack: Signs leveraged by detective controls to identify an attack in a network—for example, large binary downloads.

Indicators of compromise: Signs leveraged by forensic controls to identify the existence of malicious programs or communication in a network—for example, domain names and malware hash code.

Integrity: In cybersecurity, integrity is a property of information that ensures any unauthorized users did not modify it.

Intrusion detection system: A detection control device responsible for monitoring the network communication to identify malicious activity and policy violations.

Intrusion prevention system: A prevention control device responsible for monitoring the network communication to block malicious activity and policy violations.

Lateral movement: The technique of moving within the corporate network to gain elevated access to corporate assets.

Malicious software: Any software that harms the computer system. Also known as *malware*.

Malicious website: A website that attempts to steal, ransom, gain access to a computer, or install malicious software.

Definitions

Malware: See *Malicious software*.

Nation-state: In cybersecurity, *nation-state* refers to a perpetrator with access to a large amount of money, skills, resources, and an intention to cause harm.

Offsite: Any location outside of a corporate network or campus such as home, café, or hotel.

Onsite: See *Corporate network*.

Phishing: A practice of luring victims to reveal personal information, such as corporate credentials requesting information through an email that looks like it is from a reputable company.

Preventive control: A cybersecurity tool or process responsible for blocking a specific event in an environment such as unauthorized access.

Proxy: A server in the cloud or data center that acts as an intermediary controlling the communications between the endpoint and the Internet.

Ransomware: A malware program that turns the endpoint into an unusable state and demands a fee to disinfect.

Reinforcement learning: A machine learning methodology that learns from trial and error and the past decisions of unlabeled data to make the next decision.

Definitions

Remote work: Typically refers to employees who are always operating from geographically distant locations. Unlike telework, remote work employees do not work on the organization's campus in general.

Security control: Safeguards to detect, avoid, counteract, and minimize security risks.

Security incident: Involves compromising the confidentiality, integrity, or availability of a cyber asset.

Smash-and-grab: An attack in which the perpetrator exfiltrates data indiscriminately.

Supervised learning: A machine learning methodology that learns from labeled data; once it understands the data, it can predict the labels from unlabeled new data.

Telecommute: See *Telework*.

Telework: The ability of an employee to work from any place outside the organization's physical campus, such as a home, hotel, or train. *Telework* refers to employees who are sometimes on campus and sometimes remote.

Threat intelligence: In cybersecurity, threat intelligence is a collection of indicators that reflect the attacker's targets, motives, and behavior.

Unsupervised learning: A machine learning methodology that learns from the patterns in unlabeled data and catalogs them into different categories.

Watering hole: A practice of luring victims to reveal personal information such as corporate credentials by requesting information through a website that looks like it is from a reputable company.