RSAConference2019
San Francisco | March 4–8 | Moscone Center

SESSION ID: IDY-R11

Adversarial Machine Learning against Modern Behavioral Biometrics Systems

Heng Tang
@9he_ta0ng

Ajit Gaddam
@AjitGaddam

“The Password has become kind of a nightmare!”

Prof. Fernando J. Corbato
Creator of password back in 1961 at MIT for the Compatible Time-Sharing System

Image source: wired.com

Data breaches of passwords & by passwords

5M Gmail logins were leaked on September 9th , 2014

Source: time.com. Sep 2014

167M Hacked LinkedIn accounts are on sale

Source: csoonline.com, May 2016

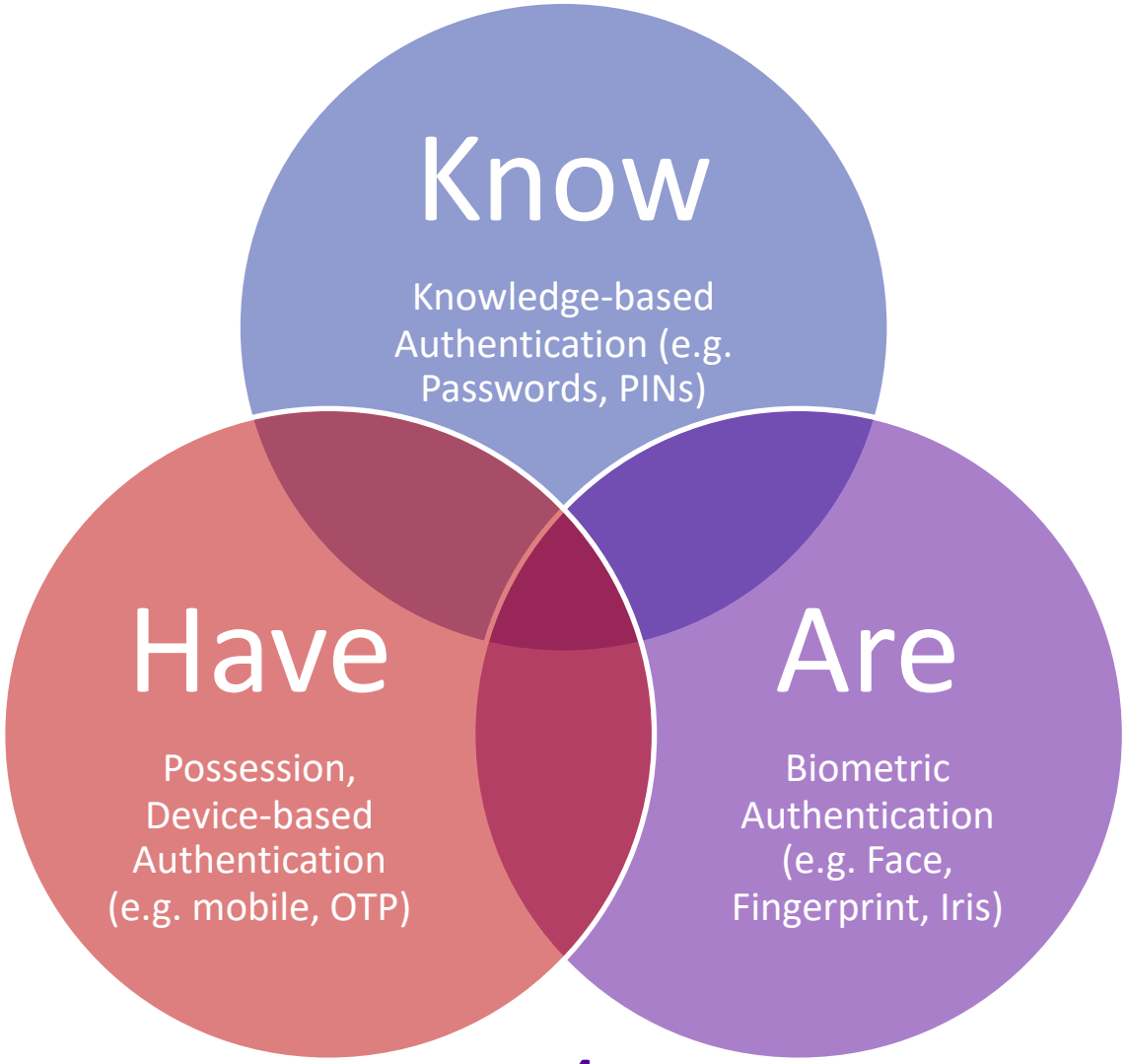
2.3B credentials stolen from 51 organizations in 2017

Shape Credential Spill Report, 2018

"80% of all data breaches resulted from weak or stolen passwords."

Verizon Data Breach report 2017

Multi-Factor Authentication



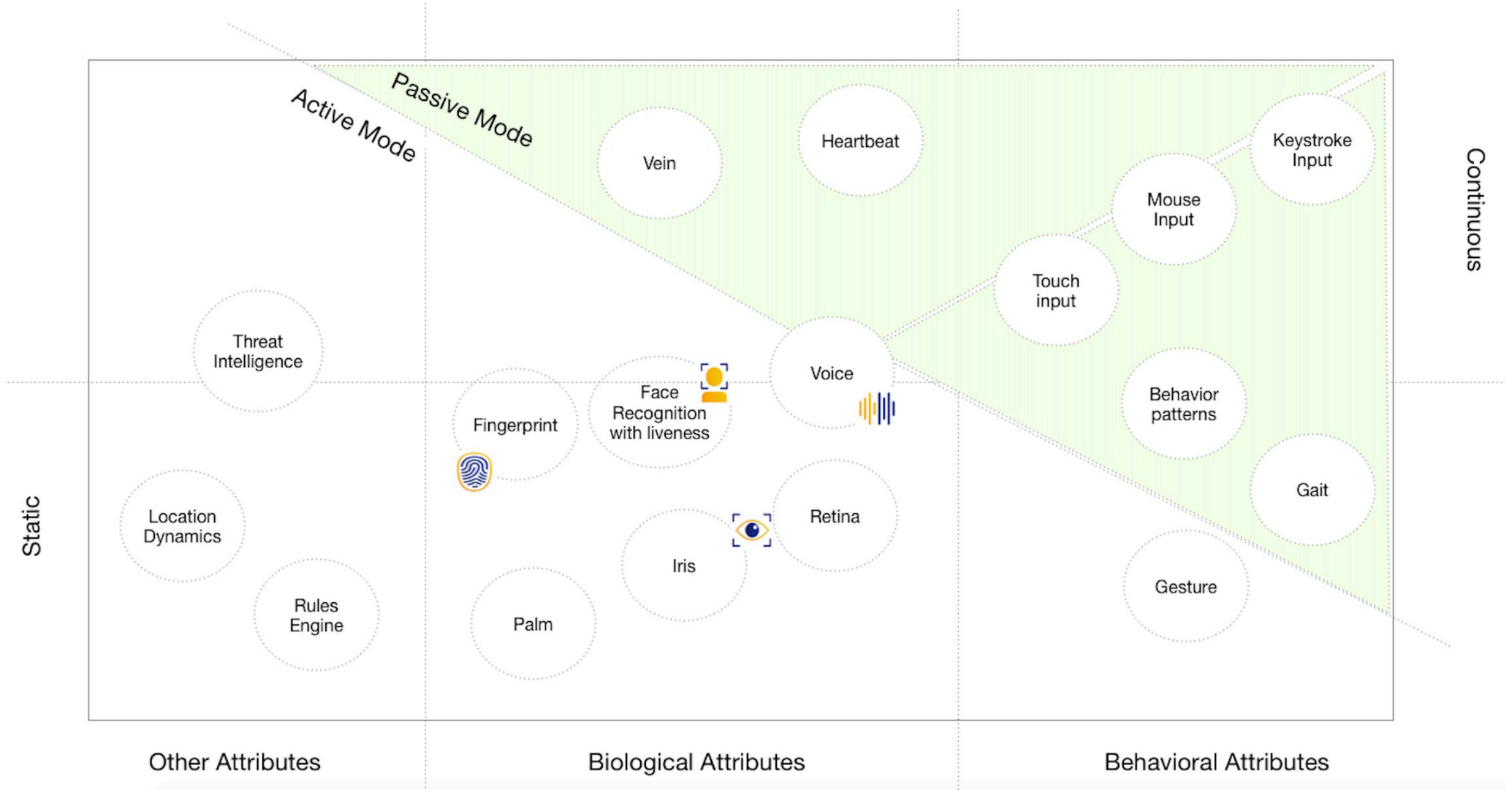
RSAConference2019

Biometric Authentication

Biological & Behavioral Biometrics

Biometric Authentication

Spectrum of Biometric Authentication



Biological vs. Behavioral Biometrics

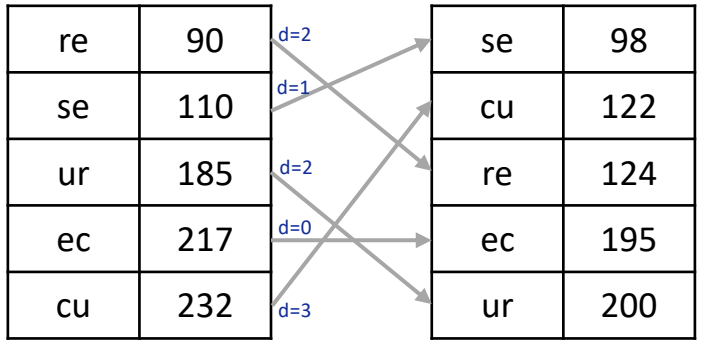
Biological	Behavioral
Static	Dynamic
More “Human Perceivable”	Less intuitive to understand raw data
Persistent	Fluctuated

Behavioral Biometrics – Keystroke Dynamics

Key pairs generated from typing “secure”

re	90
se	110
ur	185
ec	217
cu	232

re	124
se	98
ur	200
ec	195
cu	122



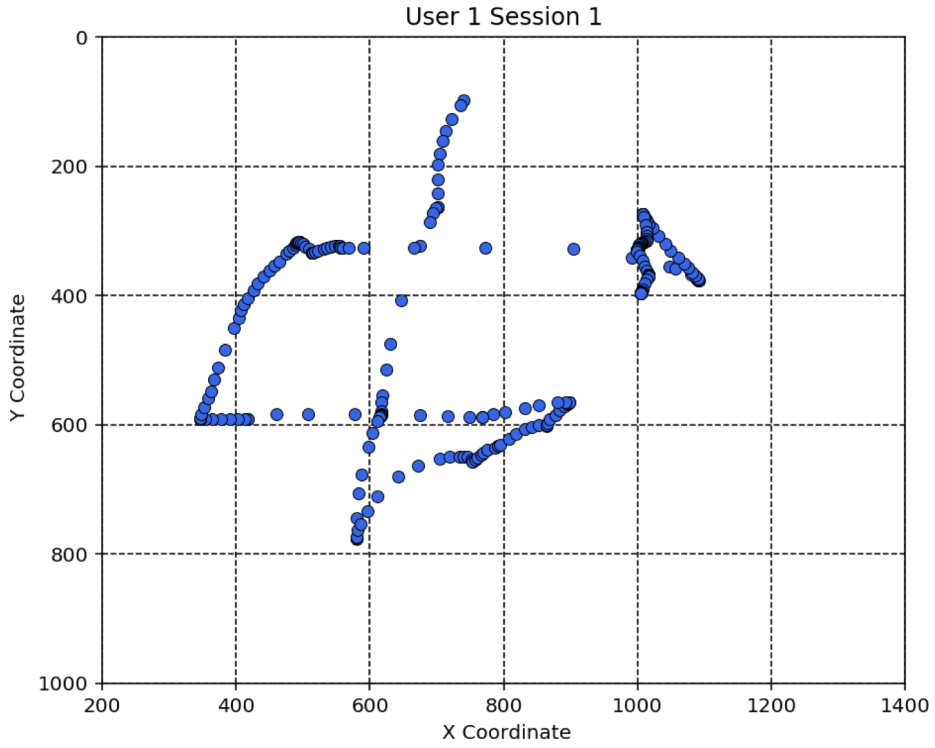
$$AS = \sum_{ng} \frac{M}{N} \; AS_{ng} \quad ; \; AS_{ng} = 1 - \frac{\sum_i sim(ng_i)}{\sum_{ng_i} 1}$$

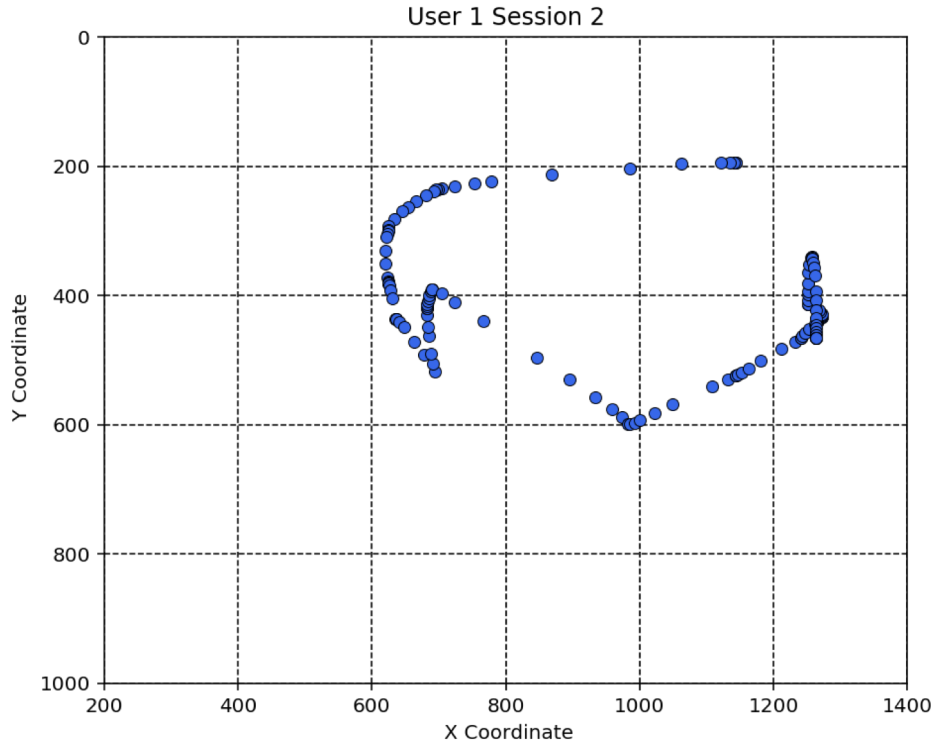
$$sim(ng_i) = \begin{cases} 1, if \; 1 < \dfrac{\max(du_{ji})}{\min(du_{ji})} < t \\ 0, otherwise \end{cases}$$

$$RS = \sum_{ng} \frac{M}{N} \; RS_{ng} \qquad RS_{ng} = \frac{\sum_i d_i}{RS_{ng-max}}$$

*Figures are Illustrative purpose only, not actual data

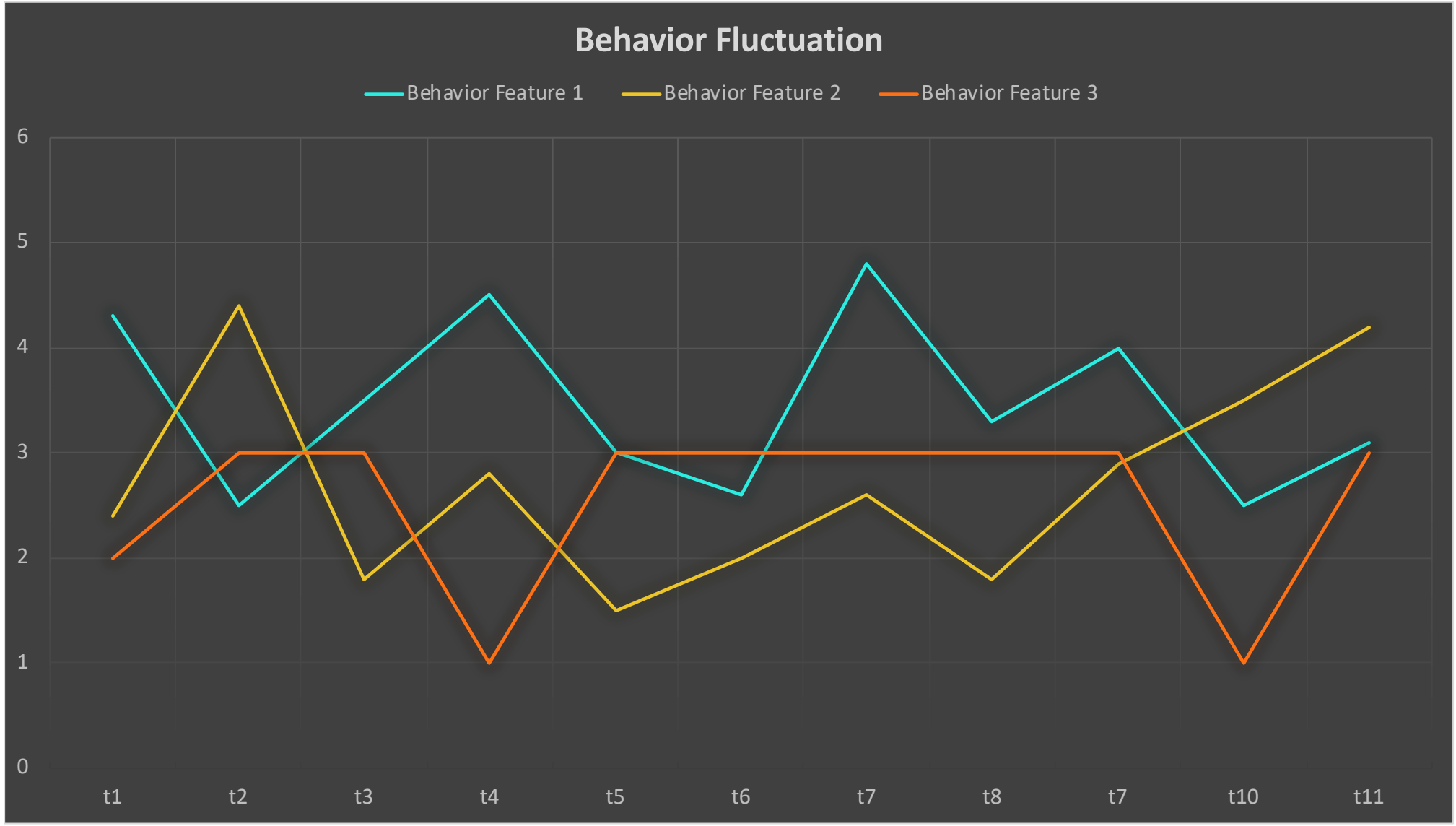
Behavioral Biometrics – Mouse Dynamics





*Figures are Illustrative purpose only, not actual data

Behavioral Biometrics Features



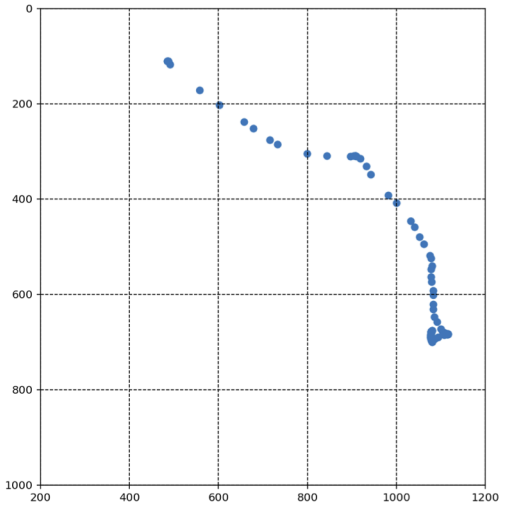
*Figures are Illustrative purpose only, not actual data

RSAConference2019

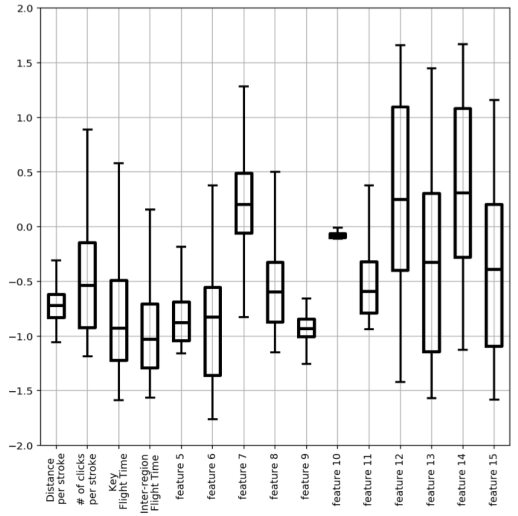
Adversarial Machine Learning

Introduction and methodologies

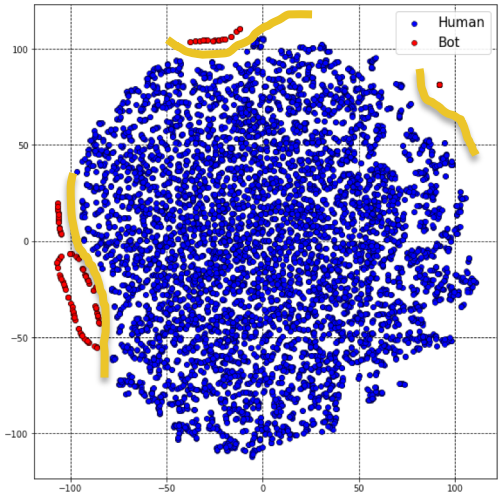
How ML/DL Model Works



Preprocessing /Data Cleaning



Feature Analysis



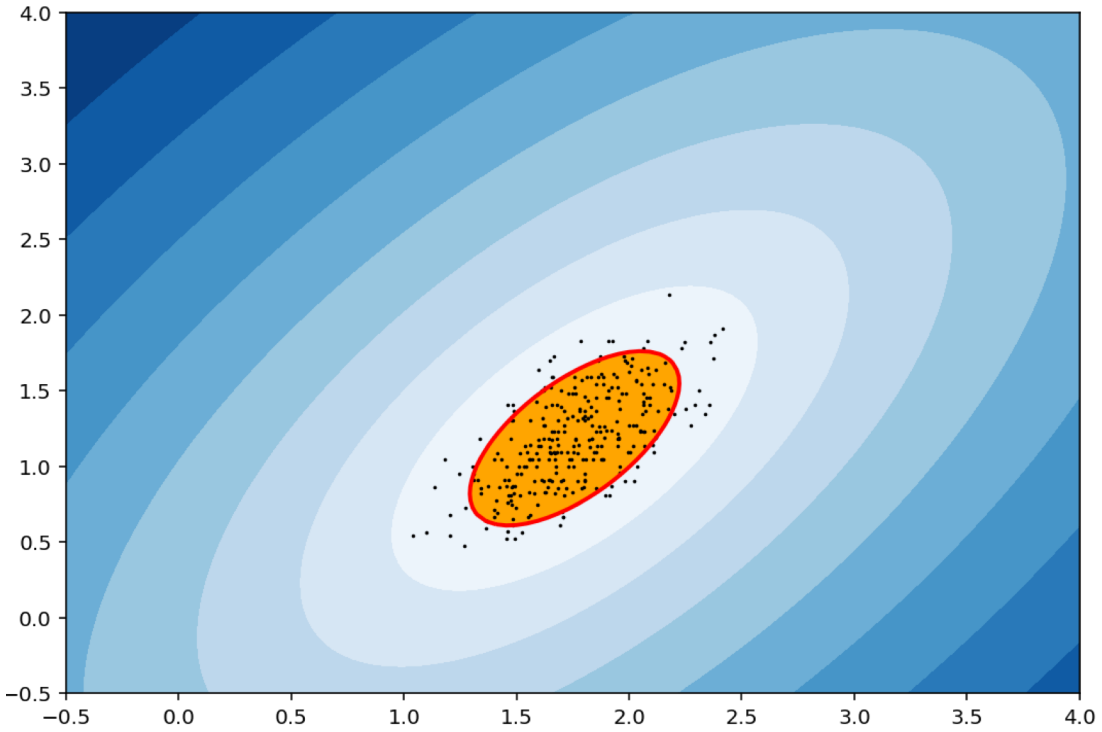
Raw Data	Features	Model
• Images	• Feature Engineering	• Statistic Models
• Measurements	• Implicit Features	• Deep Neural Networks
• Time-series Data	• Embedding	• Non-linear Models

*Figures are Illustrative purpose only, not actual data

Adversarial Machine Learning

Attacker

- Trying to explore the decision boundary of original ML/DL based models
- Generating synthetic samples that can fit into the decision boundary to be classified with expected labels



Adversarial Machine Learning – Definition

$\boldsymbol{x}$: $Original\ Inputs$
$\boldsymbol{y}$: $True\ Labels$
$\boldsymbol{y'}$: $Targeted\ Malicious\ Labels$
$C(\boldsymbol{x})$: $Trained\ Classifiers\ -s.t.\ C(\boldsymbol{x}) = \boldsymbol{y}$
$p(\boldsymbol{x}), p(\boldsymbol{x}, \boldsymbol{y'})$: $Perturbation\ functions$

Targeted Attack

$$\underset{p(\boldsymbol{x},\boldsymbol{y'})}{\arg\min}\ C(\underbrace{\boldsymbol{x} + p(\boldsymbol{x}, \boldsymbol{y'})}) = \boldsymbol{y'}$$

Crafted Inputs

Untargeted Attack

$$\underset{p(\boldsymbol{x})}{\arg\min}\ C\big(\underbrace{\boldsymbol{x} + p(\boldsymbol{x})}\big) \neq \boldsymbol{y}$$

Crafted Inputs

Adversarial Machine Learning – Image Recognition

Image source: pixnio.com

Imperceptible Difference

- $\underset{p(\boldsymbol{x},\boldsymbol{y}')}{\arg\min} C(\boldsymbol{x} + p(\boldsymbol{x}, \boldsymbol{y}')) = \boldsymbol{y}'$
- Metrics to reflect the imperceptibility of the change to observer
  - $L^0\ distance$
  - $L^1\ distance$
  - $L^2\ distance$
  - $L^\infty\ distance$ – Max Norm

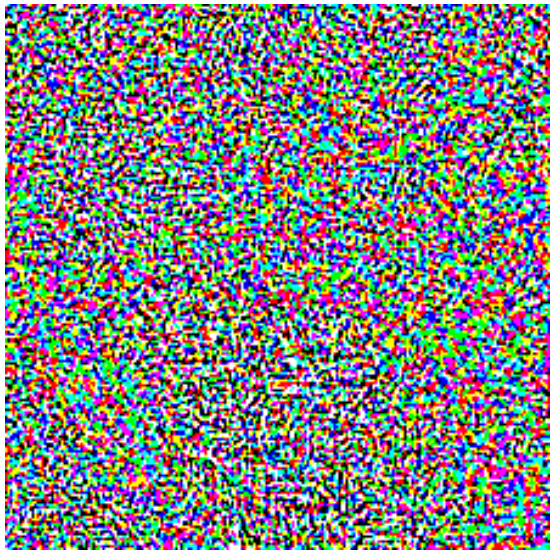
Adversarial Machine Learning – Image Recognition

$\boldsymbol{x}$

Prediction “Gorilla” with High Confidence

+

$p(\boldsymbol{x}, \boldsymbol{y}')$



$\varepsilon \cdot sign(\nabla_x L(\boldsymbol{\theta}, \boldsymbol{x}, \boldsymbol{y}'))$

FGSM[Goodfellow et al., 2014]

$\varepsilon = L^{\infty}\big(p(\boldsymbol{x}, \boldsymbol{y}')\big)$

Weights

=

$\boldsymbol{x} + p(\boldsymbol{x}, \boldsymbol{y}')$

Prediction “Human” with High Confidence

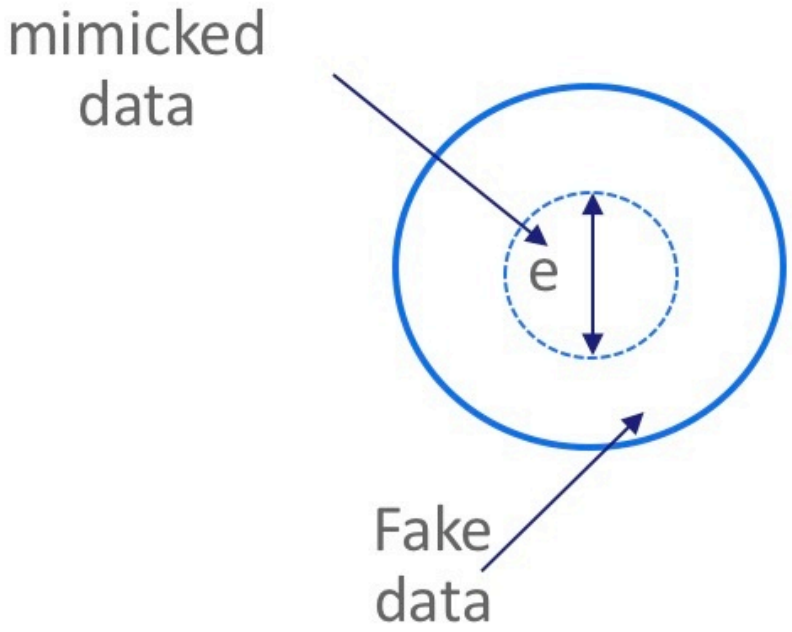
Image source: pixnio.com

Adversarial Machine Learning – Behavioral Biometrics

Targeted Attack

$$\arg\max_{p(\boldsymbol{x},\boldsymbol{y})} C(\boldsymbol{x} + p(\boldsymbol{x},\boldsymbol{y})) = \boldsymbol{y}$$

Crafted Inputs



Impact on Behavioral Biometrics Systems

- Adaptive Systems vs. Static Systems
- Model Poisoning
  - Usability
  - Discredit the underlying system

- Attacks may not always target tweaking ML/DL models
  - Tweak vs. Circumvent
  - Carefully crafted attack vs. random guess

RSAConference2019

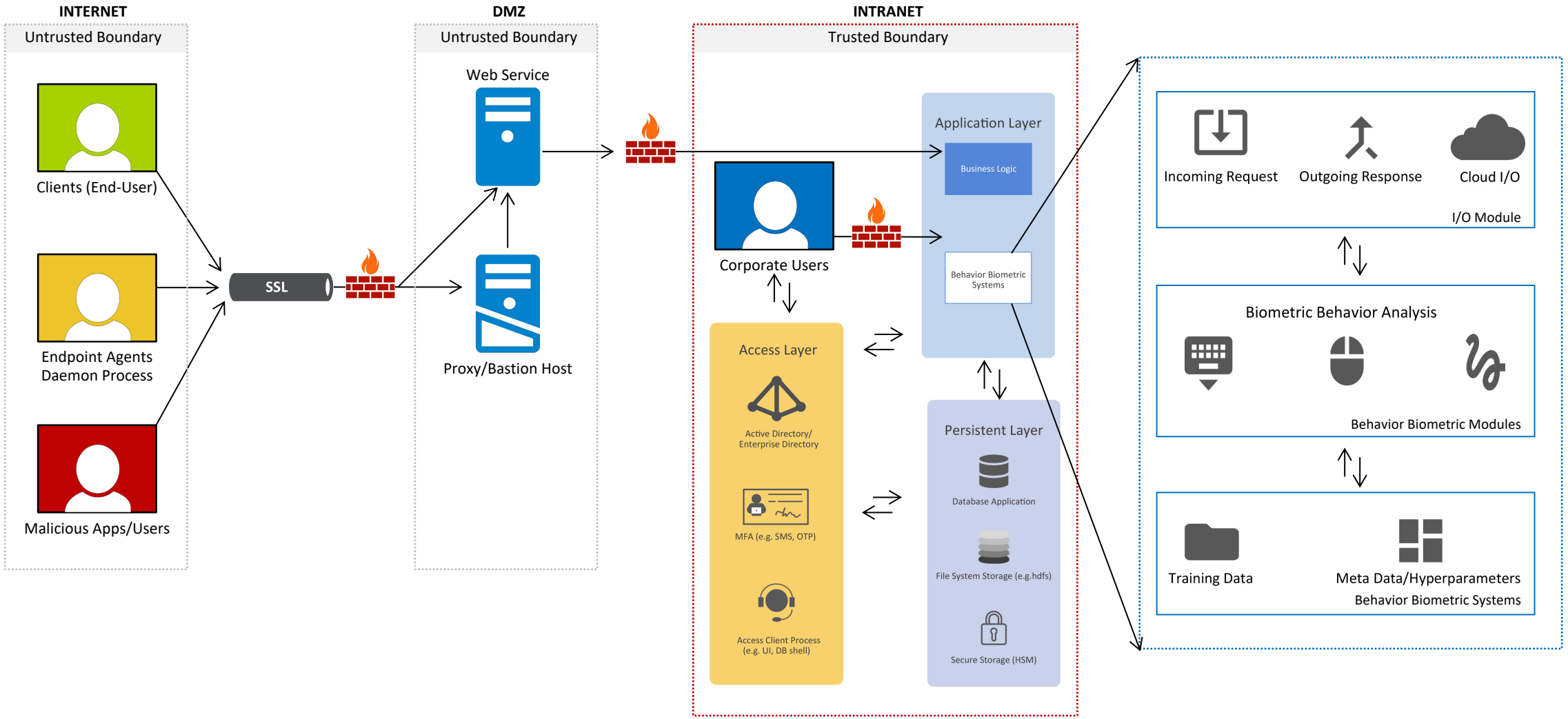
Defense

Threat Modeling and Countermeasures

Different Levels of Knowledge

- **White-box** – Has full knowledge and potentially full access to any component (e.g. Inside attack with domain knowledge)
- **Black-box** – Has limited knowledge and access on targeted system/model
- **Grey-box** – Something in between, more realistic assumption

Threat Model



Threat Model

Attack 1: Session Hijacking
– MTM attack
– Sniffing

Attack 2: Reverse Engineering
– Crafted payload

Attack 3: Malicious External Attackers
– Randomly crafted payload
– Impersonation
– Compromised account
– DDos & Botnet

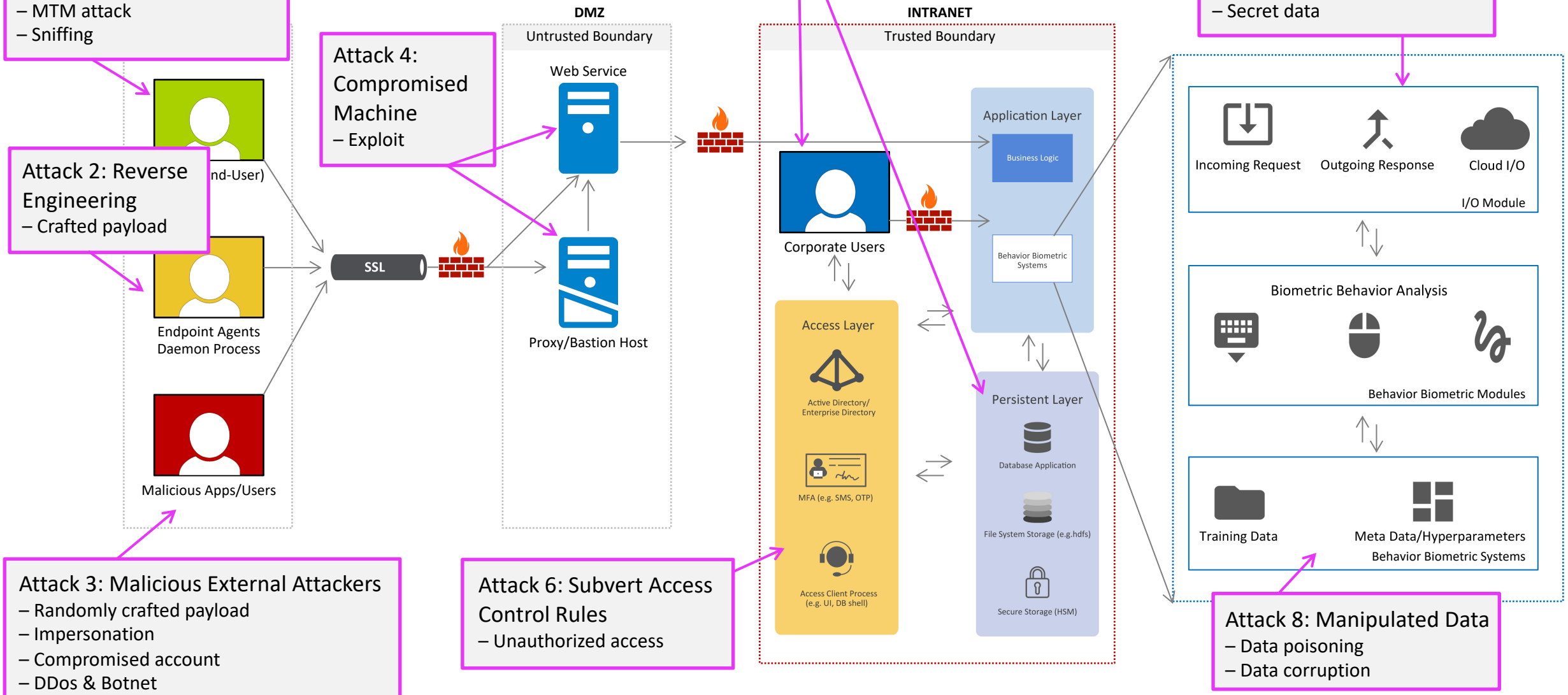
Attack 4: Compromised Machine
– Exploit

Attack 5: Insider Attack
– Abusing access

Attack 6: Subvert Access Control Rules
– Unauthorized access

Attack 7: Data Leakage
– Unencrypted communication
– Secret data

Attack 8: Manipulated Data
– Data poisoning
– Data corruption



Attack Category Groupings

Client side

- Hijacked sessions
- Reverse engineering endpoint agents
- Malicious crafted payload

Server side

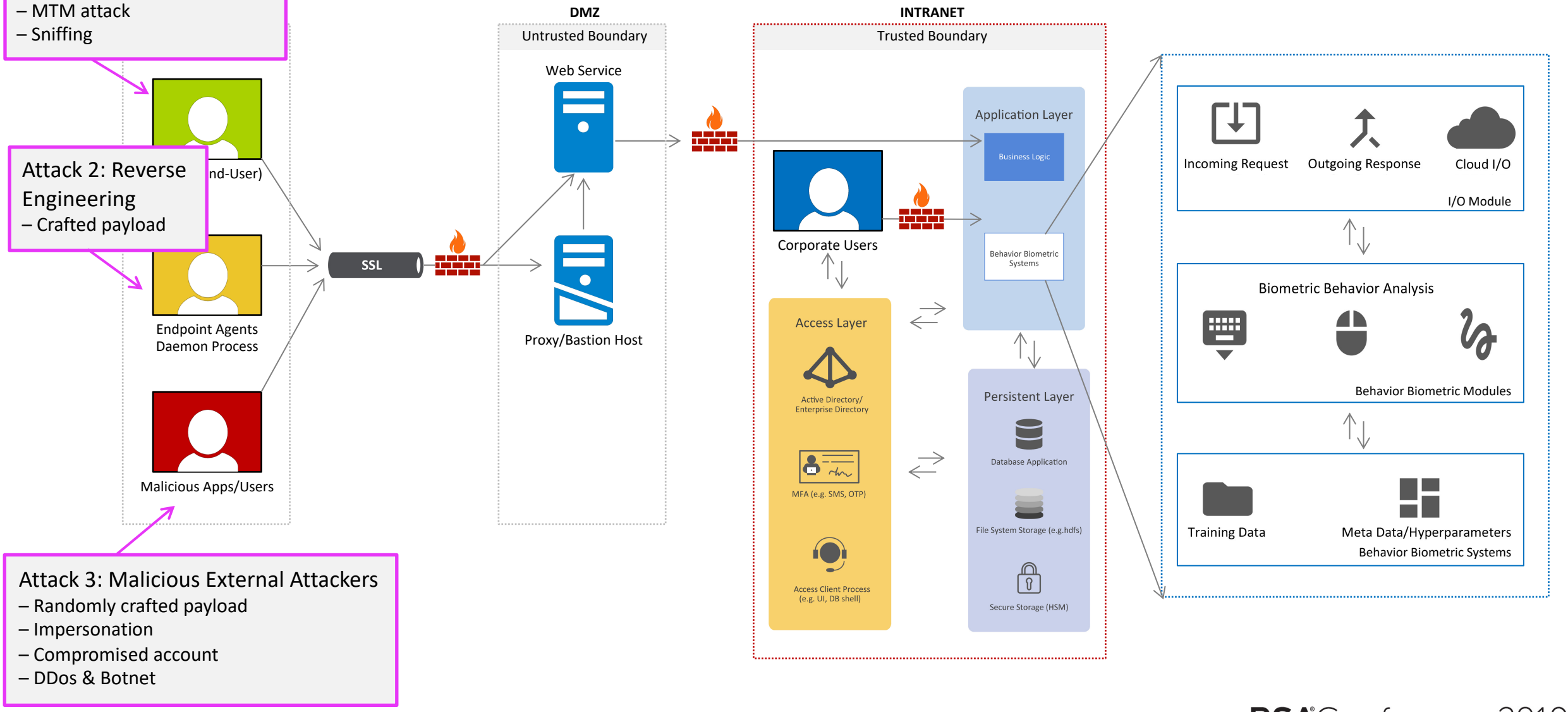
- Compromised machines
- Manipulated data & tweaked models
- Insider attack

Client Side Attacks

Attack 1: Session Hijacking
– MTM attack
– Sniffing

Attack 2: Reverse Engineering
– Crafted payload

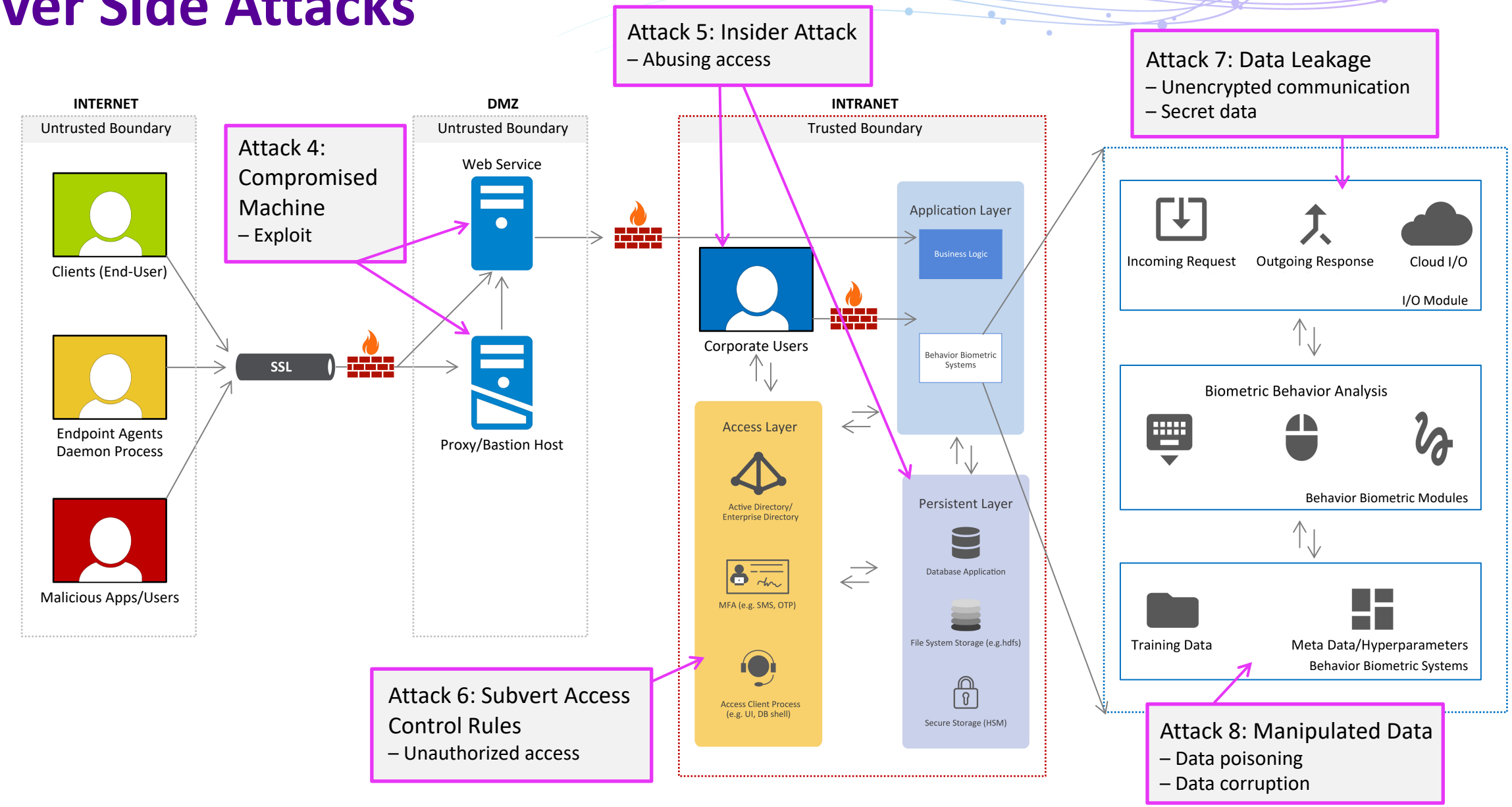
Attack 3: Malicious External Attackers
– Randomly crafted payload
– Impersonation
– Compromised account
– DDos & Botnet



Client Side Attack Protection

- Targeting to manipulate input and perform black-box attacks
- End point protection
  - Binary Hardening
  - Code Obfuscation
  - Transport Level Encryption + Message Level Encryption + Integrity checking
- Never Trust External Input

Server Side Attacks



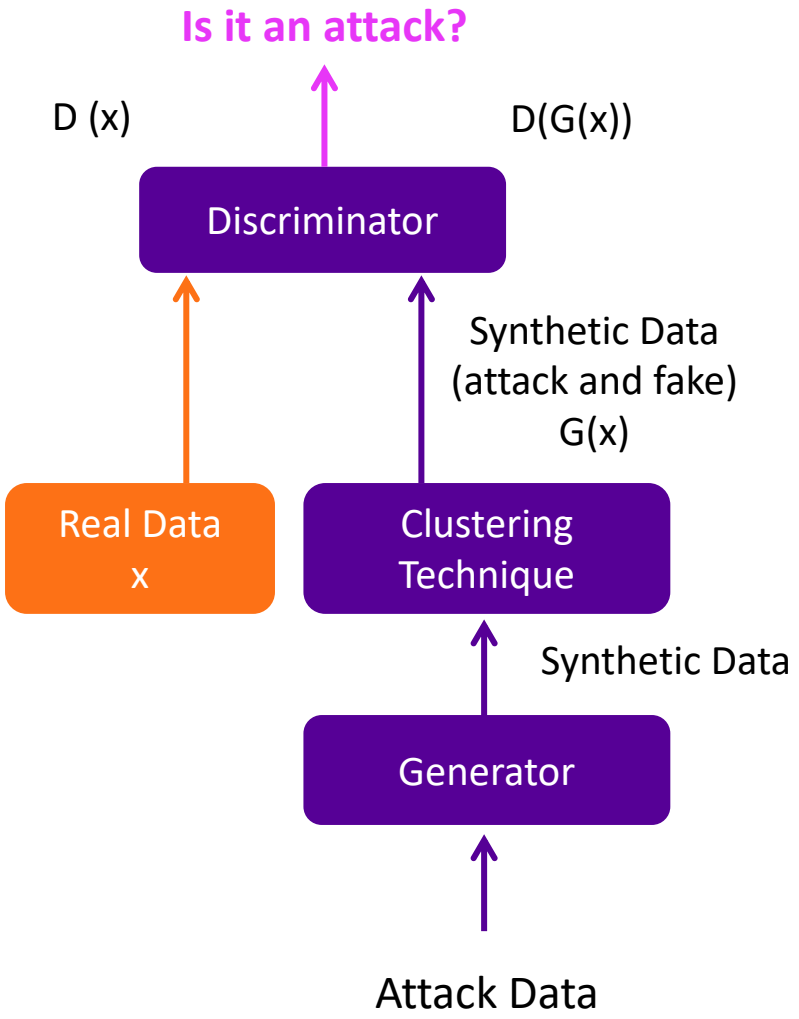
Server Side Attack Protection

- Targeting to manipulate internal data + white-box attacks
- Protection against abused/unauthorized access
  - Encrypting data at rest
  - Access control/monitoring on key infrastructure (e.g. DB/model store)
  - Message Authentication + Firewall

Model Level Attack Protection

Detection Models

- Filtering/Preprocessing the samples by another model
- Trained with attack data and real data to detect attacks

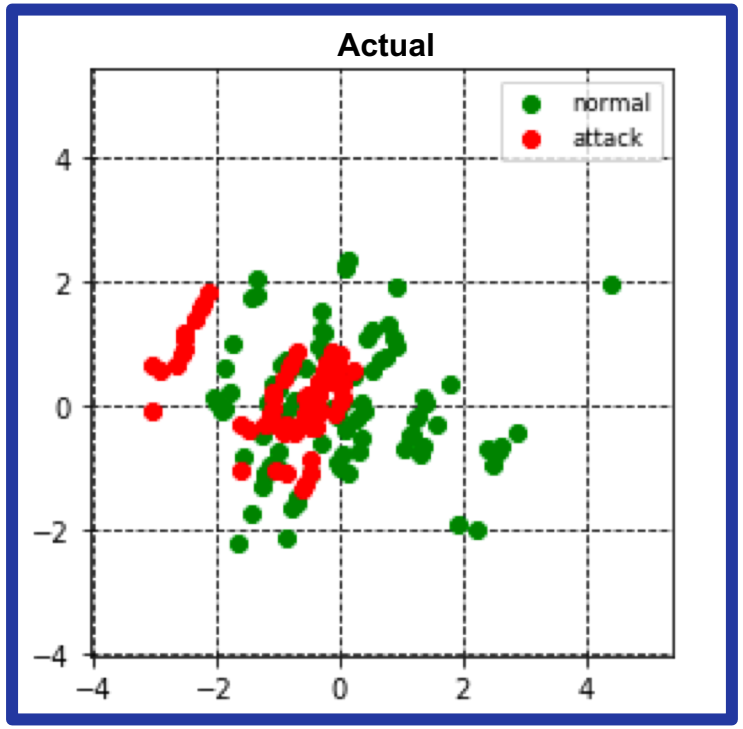


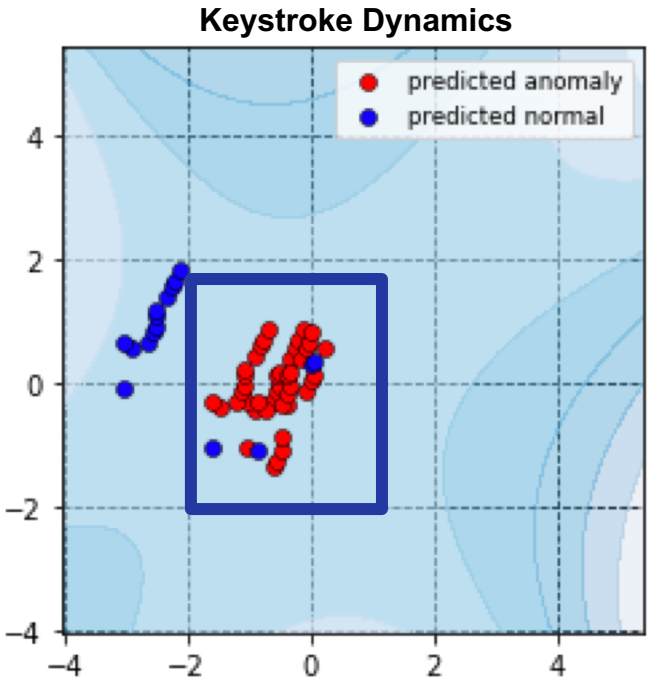
Model Level Attack Protection

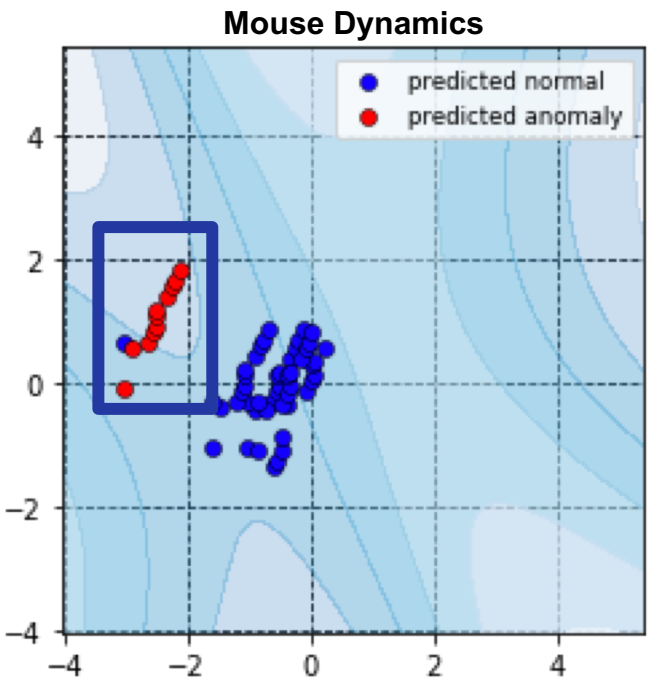
Model Hardening

- Model immutable to crafted input/manipulated data
  - Trained with attack data
  - Additional labels to attack data
- Combination of different models – Ensemble Models
  - Better results (smaller false acceptance/rejection)
  - Better security controls (more entropy)

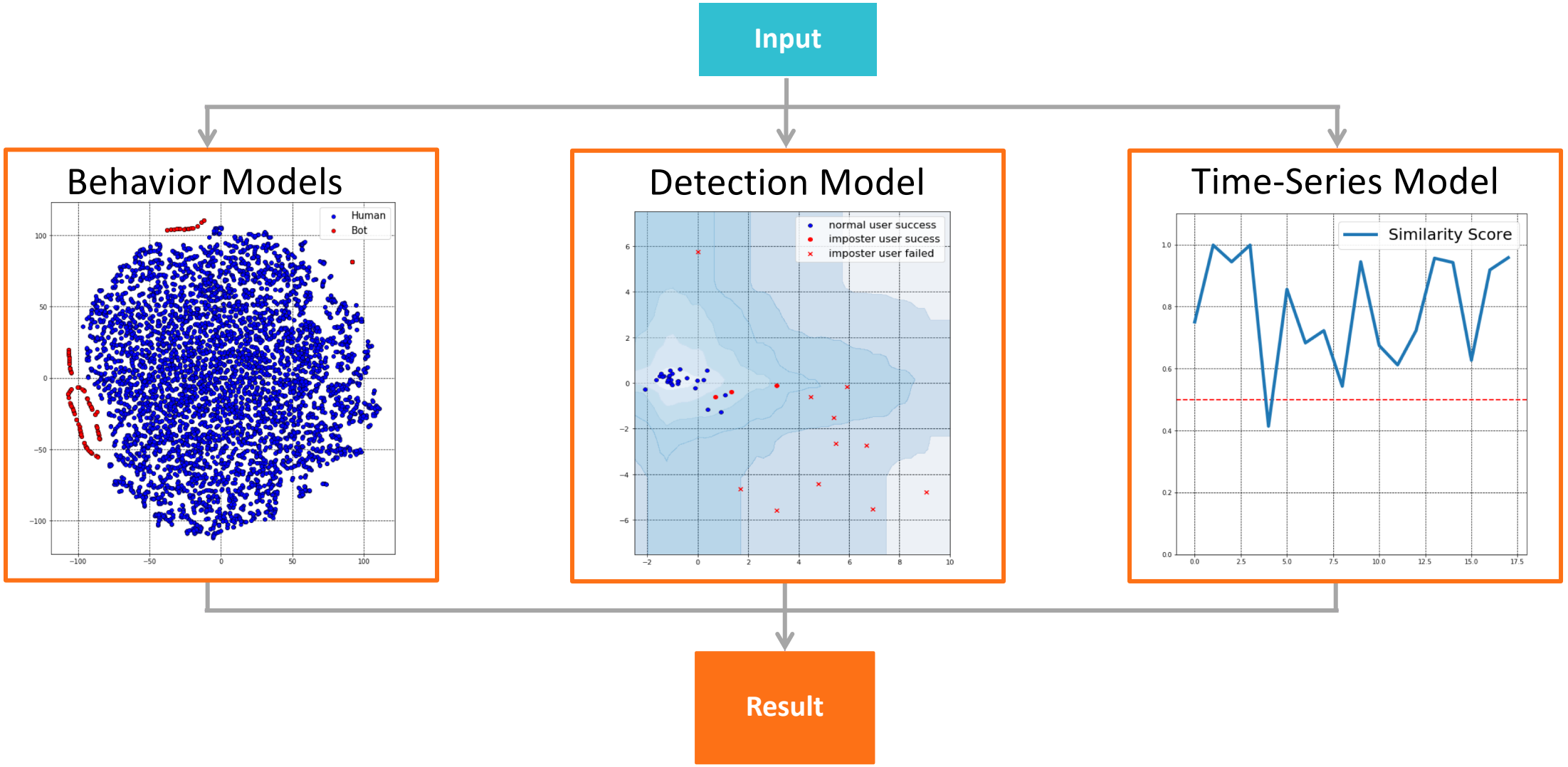
Ensemble models







Overall Architecture for Behavioral Biometrics Systems



Takeaway

Adversarial Machine Learning based attacks against behavioral biometric systems can cause significant impact to this authentication mechanism. Traditional security controls combined with **Machine Learning specific countermeasures** mitigate the impact.

Q & A

References

- Gaddam, Ajit. "Usage of Behavioral Biometric Technologies to Defend Against Bots and Account Takeover Attacks." (2019).
- Goodfellow, Ian J., Jonathon Shlens, and Christian Szegedy. "Explaining and harnessing adversarial examples." *arXiv preprint arXiv:1412.6572* (2014).
- Grosse, Kathrin, et al. "On the (statistical) detection of adversarial examples." *arXiv preprint arXiv:1702.06280*(2017).
- Gunetti, Daniele, and Claudia Picardi. "Keystroke analysis of free text." *ACM Transactions on Information and System Security (TISSEC)* 8.3 (2005): 312-347.
- Lu, Jiajun, et al. "No need to worry about adversarial examples in object detection in autonomous vehicles." *arXiv preprint arXiv:1707.03501* (2017).
- Lee, Hyeungill, Sungyeob Han, and Jungwoo Lee. "Generative adversarial trainer: Defense to adversarial perturbations with gan." *arXiv preprint arXiv:1705.03387*(2017).