We Don’t Need no Bot Infestation:

Machine Learning for Enterprise Security

Falaah Arif Khan,

Dell Cloud Identity, Dell Commerce Platform

Personal: https://falaaharifkhan.github.io/research

Blog: https://thefaladox.wordpress.com/

Twitter: @ArifFalaah

Internal Use - Confidential

To Do:

- ❑ Why is Security a hard problem?
- ❑ Why Machine Learning?
- ❑ Misconceptions about ML for ES?
- ❑ How can you build ‘intelligent’ security? (With demos!)
- ❑ What I’ve learned (so far)

Internal Use - Confidential

Why is Security a Hard Problem?

- ❑ Security is fundamentally asymmetric
- ❑ Need to Build for past, present and future capabilities
- ❑ Specialized, domain expertise required

Silver lining (?): Be Creative!

Internal Use - Confidential

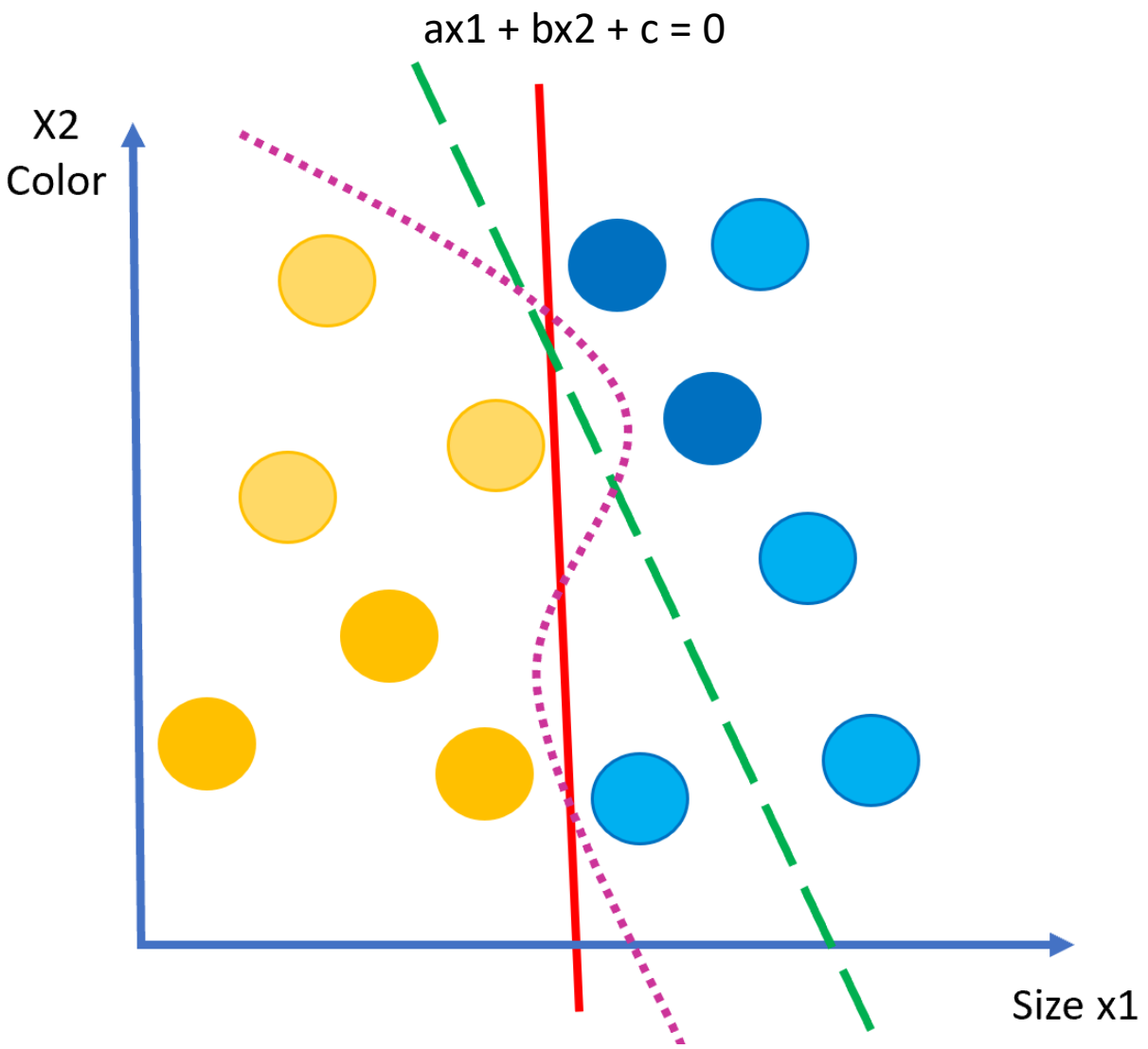
Machine Learning 101

A Quick Review

Internal Use - Confidential

Machine Learning 101 : Supervised Learning

Classify apples vs oranges on the basis of their color and size



Decision Boundary:

h = ax1 + bx2 + c

$$\hat{y} = f(h)$$

$$h = \sum_i w_i x_i$$

$$output = \begin{cases} 0, & \sum_i w_i x_i \leq \theta \\ 1, & \sum_i w_i x_i > \theta \end{cases}$$

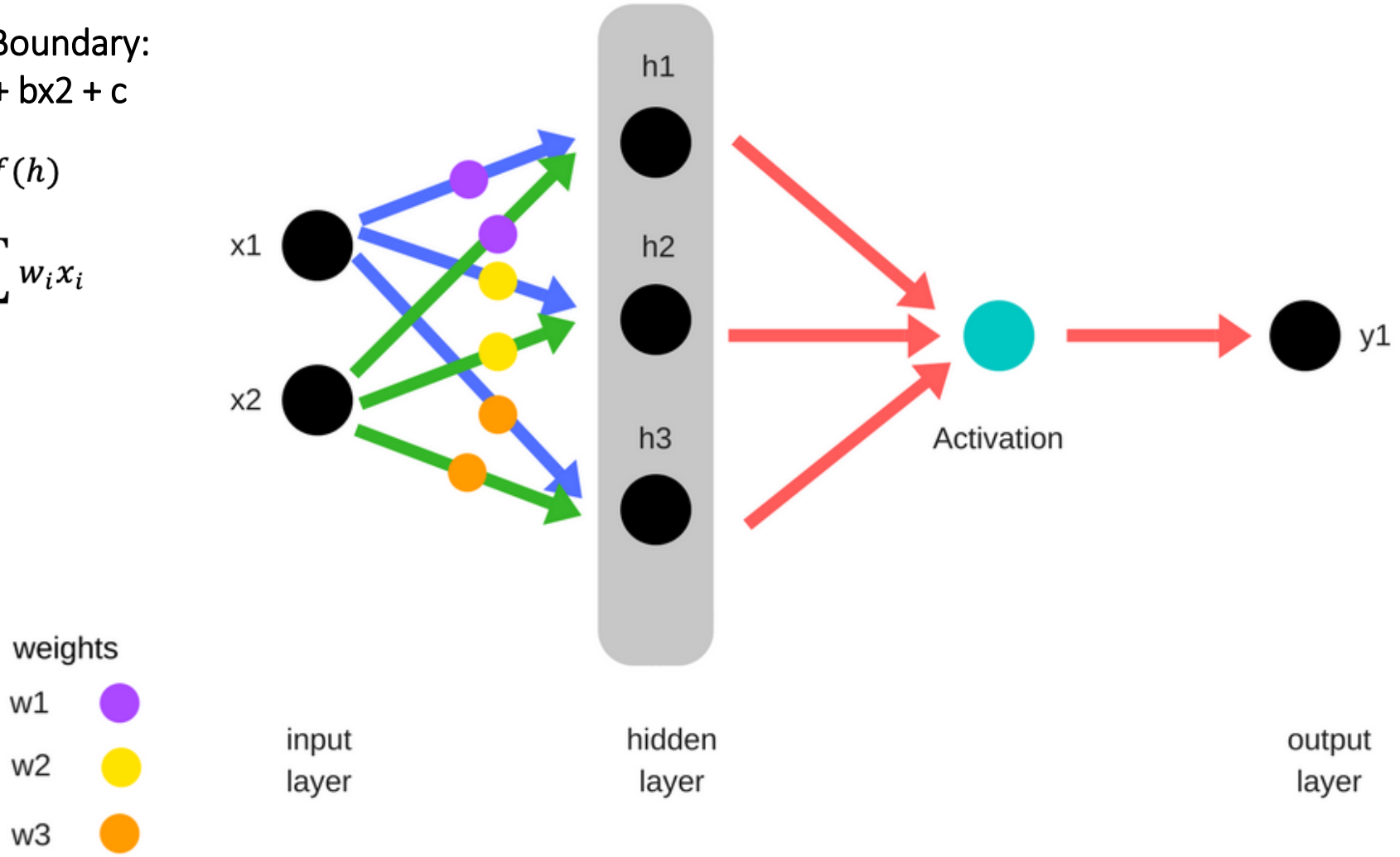
Internal Use - Confidential

Machine Learning 101: Supervised Learning

Decision Boundary:
h = ax1 + bx2 + c

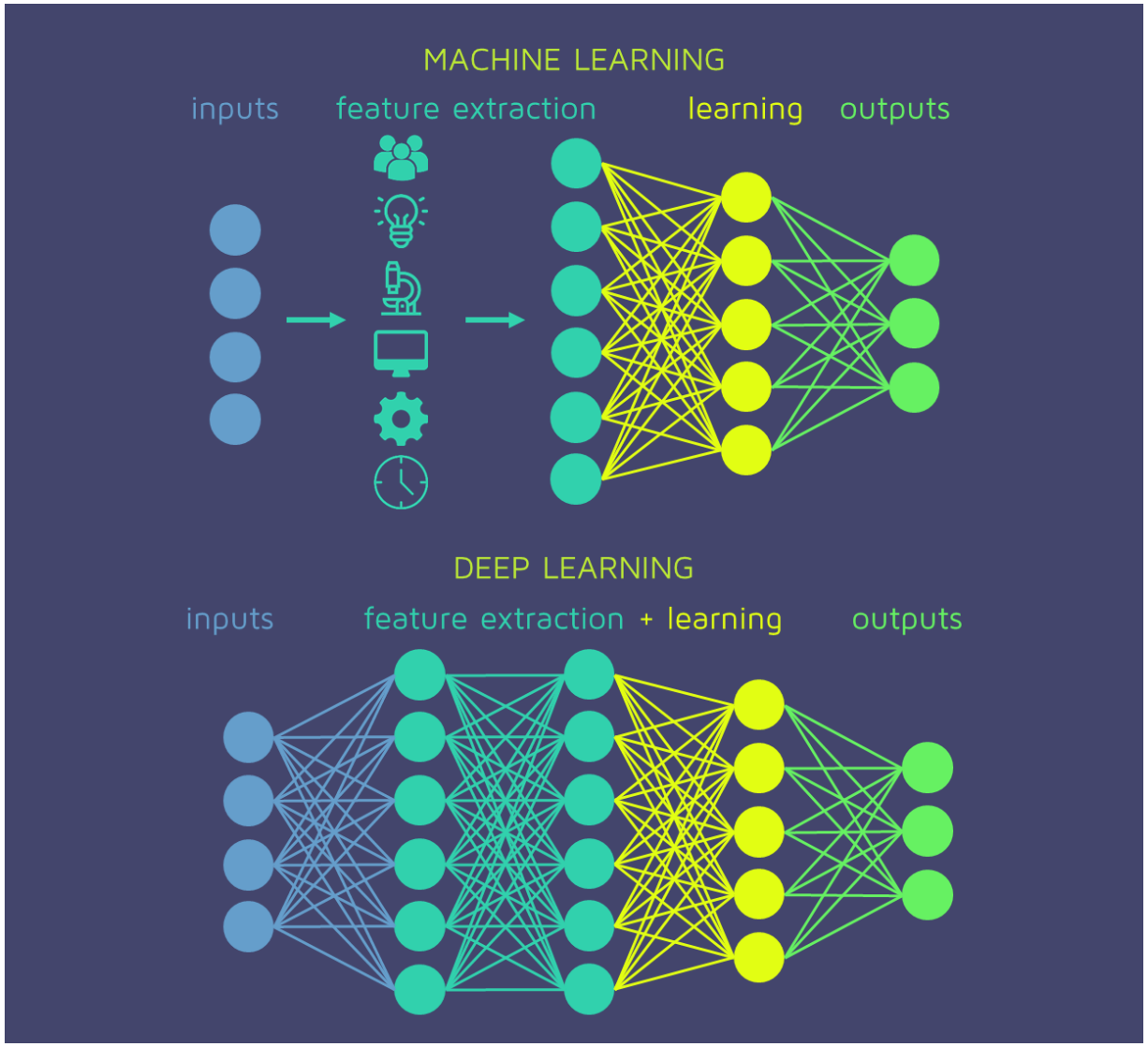
$$\hat{y} = f(h)$$

$$h = \sum_i w_i x_i$$



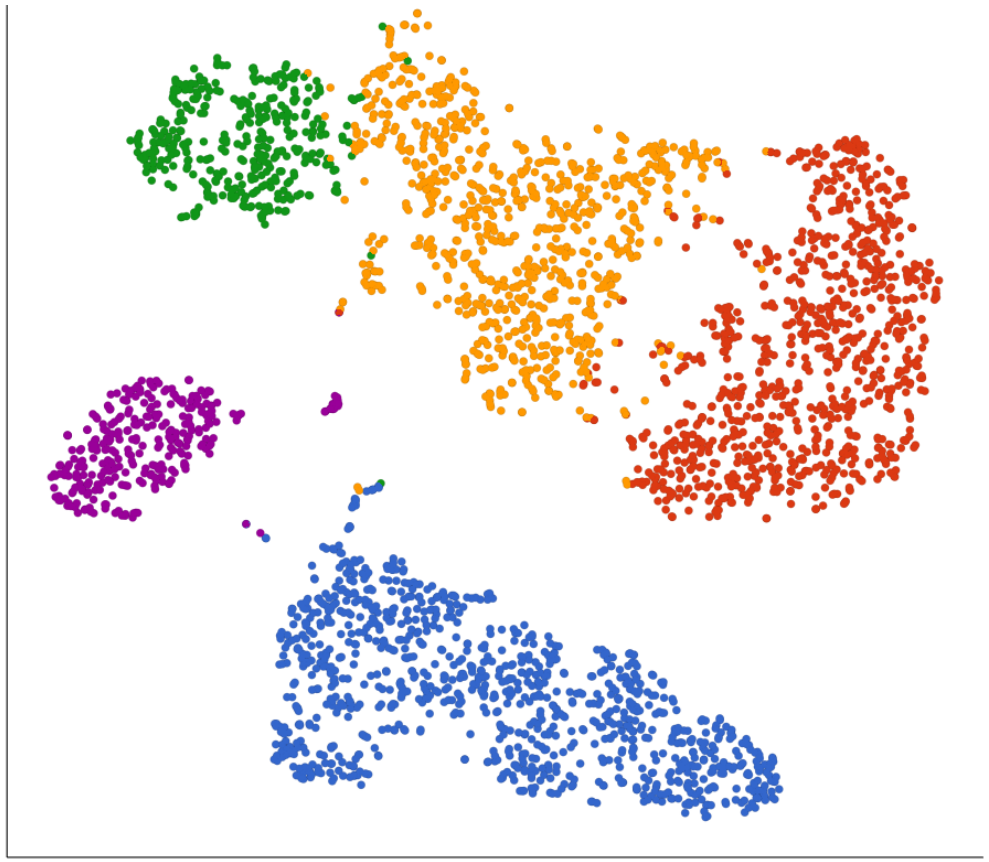
Internal Use - Confidential

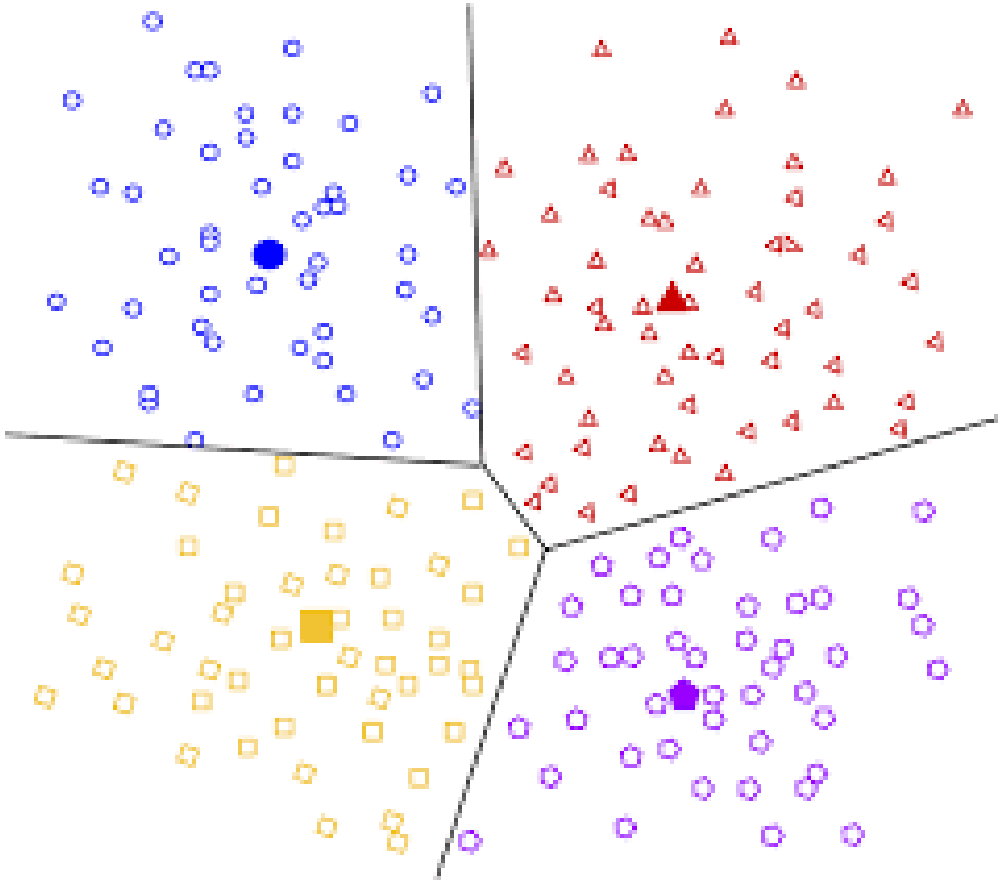
Machine Learning 101: Deep Learning



Internal Use - Confidential

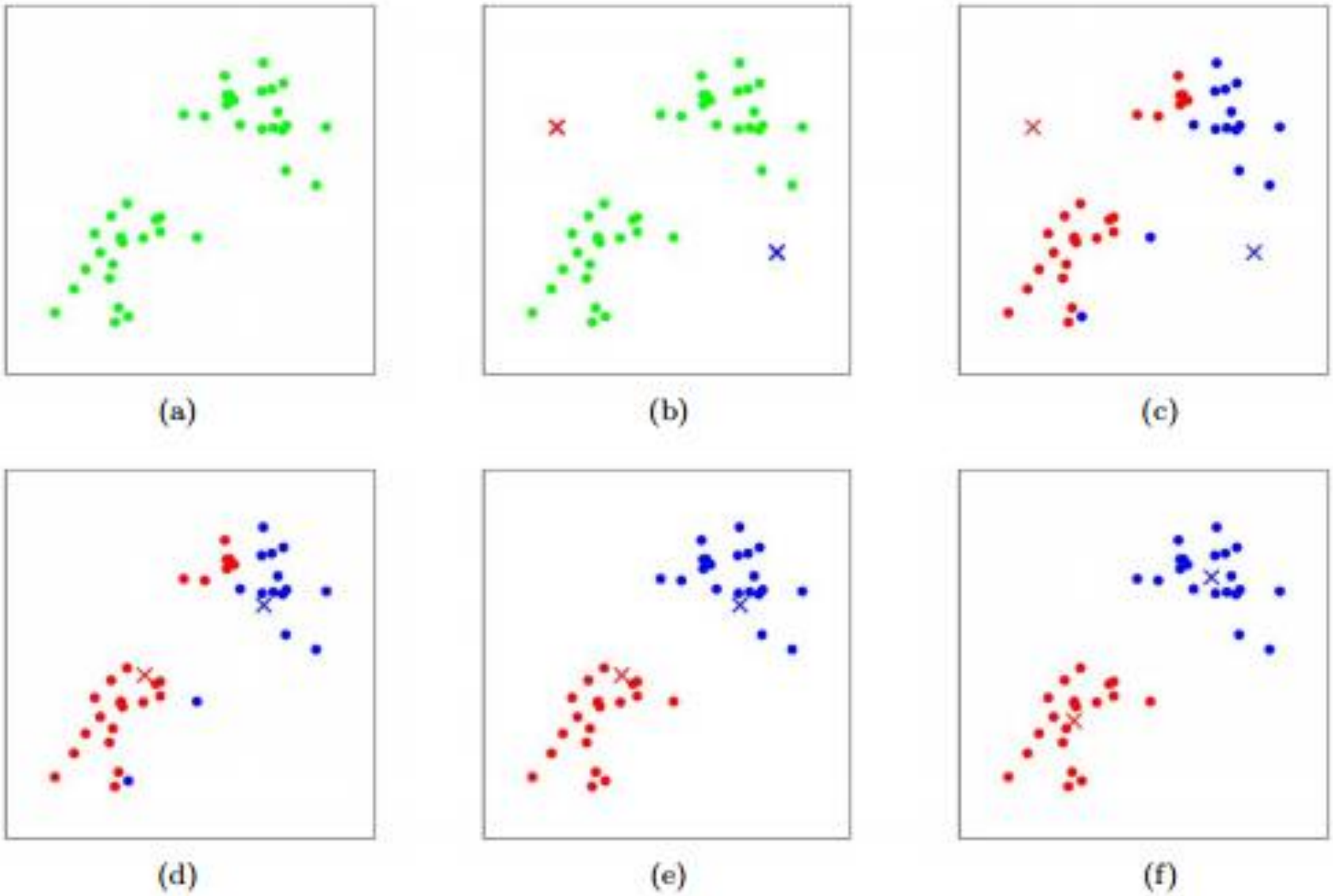
Machine Learning 101: Unsupervised Learning





Internal Use - Confidential

Machine Learning 101: Unsupervised Learning



Internal Use - Confidential

Why Machine Learning?

Security is fundamentally asymmetric

- ML detects patterns in large volumes of data

Need to build for past, present and future capabilities

- Data is the secret sauce. Retrain/tune model to new capabilities

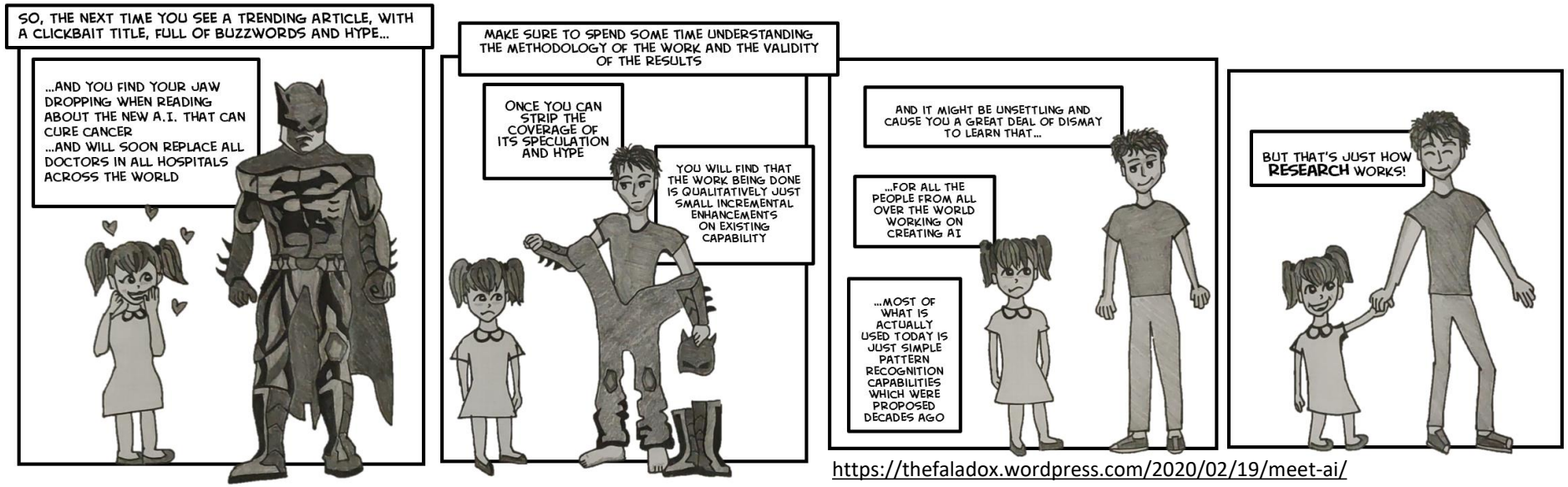
Specialized domain expertise required

- Assist, not replace. Reduce attack perimeter

Internal Use - Confidential

Common Misconceptions

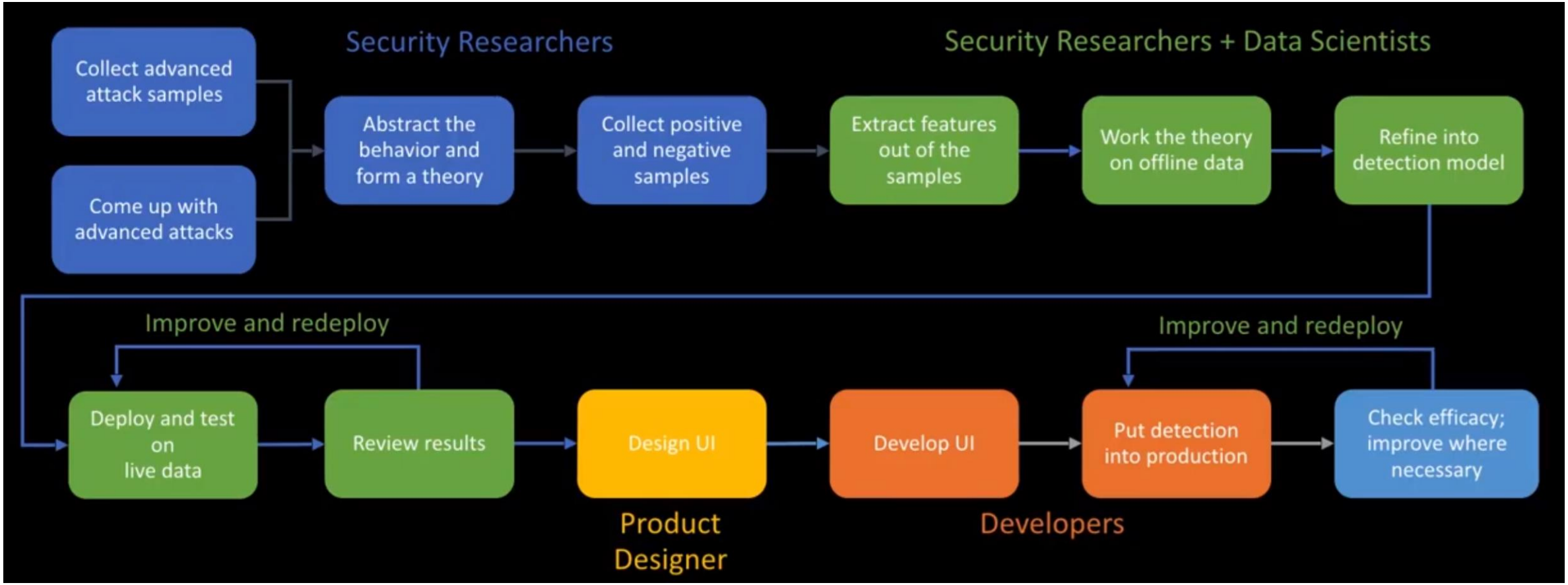
- My ML model is going to help me catch 1000s of attacks
- I can completely automate my product security using ML (deploy and forget)
- If I train long enough, on a large enough dataset, I can build a general security model for my application
- I can replace my L3 team with my ML model



https://thefaladox.wordpress.com/2020/02/19/meet-ai/

Internal Use - Confidential

How to build Intelligent Security



Internal Use - Confidential

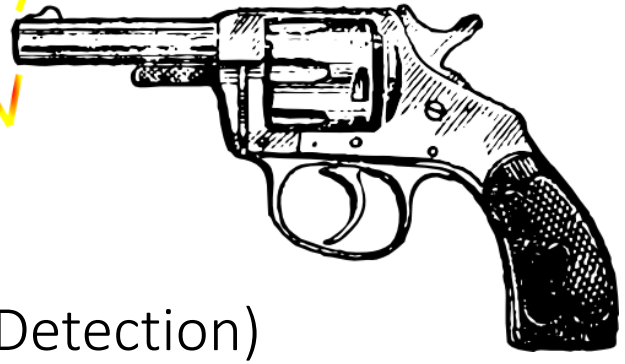
How to build Intelligent Security

1. What is your protected resource? What are all the possible attack boundaries? Who uses your resource?
2. What is the right data to collect? Can I get both positive and negative samples?
3. Train your model
4. Test & refine on data (offline, test env, synthetic)
5. Deploy and monitor; improve/retrain wherever necessary

Internal Use - Confidential

How to build Intelligent Security

Trigger



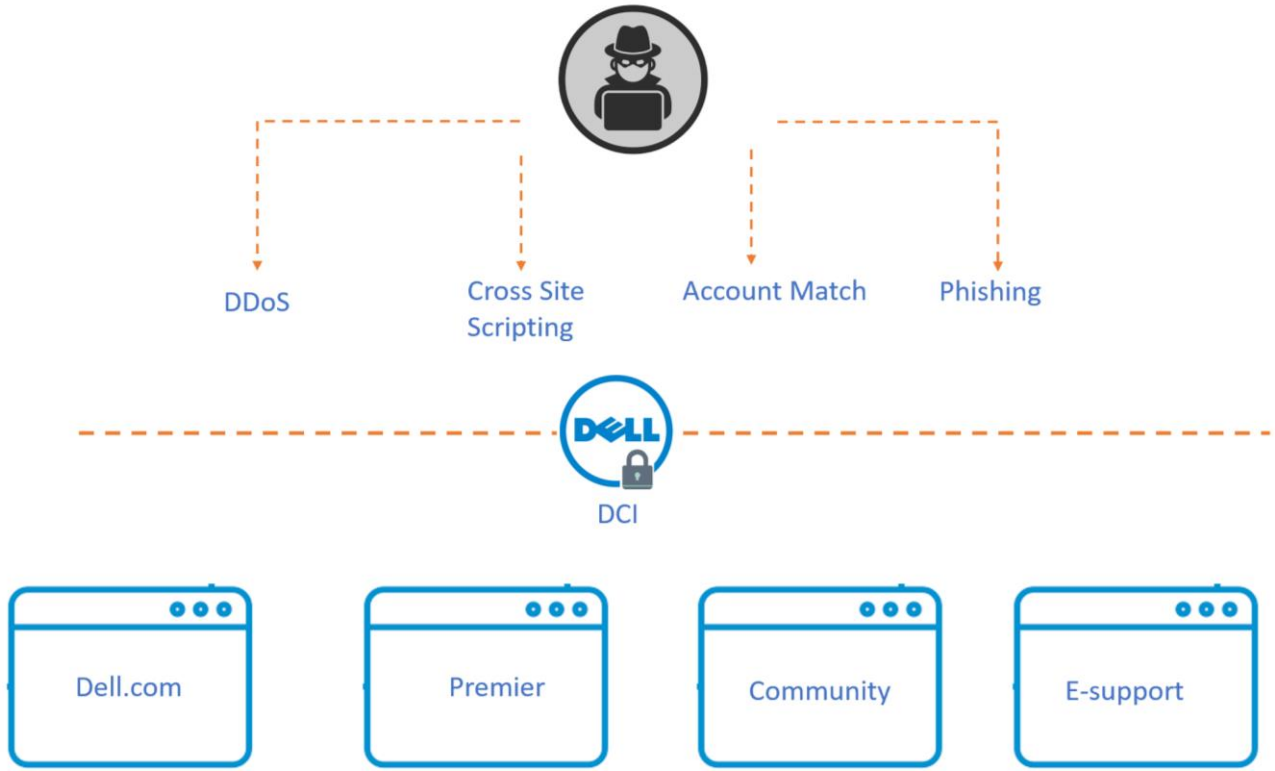
(Anomaly Detection)

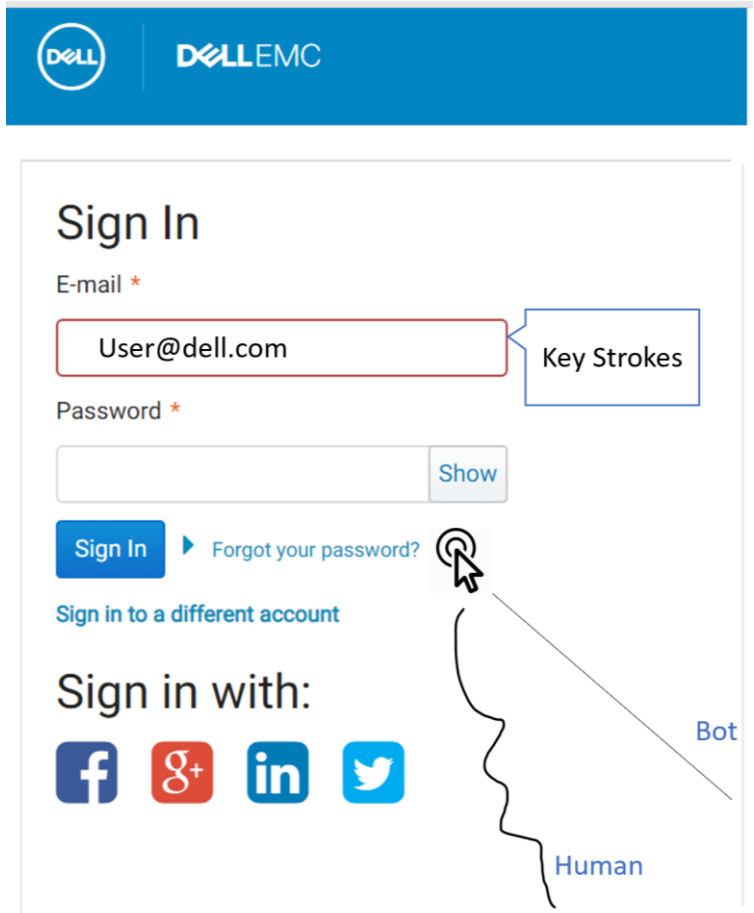
Response (Step-up mechanism)

Analytics (Monitoring)

Internal Use - Confidential

Behavioral Biometrics and Machine Learning to secure Website Logins





Paper: Arif Khan F., Kunhambu S., G K.C. (2019) Behavioral Biometrics and Machine Learning to Secure Website Logins
US Patent: Arif Khan, Falaah, Kunhambu, Sajin and Chakravarthy G, K. Behavioral Biometrics and Machine Learning to secure Website Logins. US Patent 16/257650, filed January 25, 2019

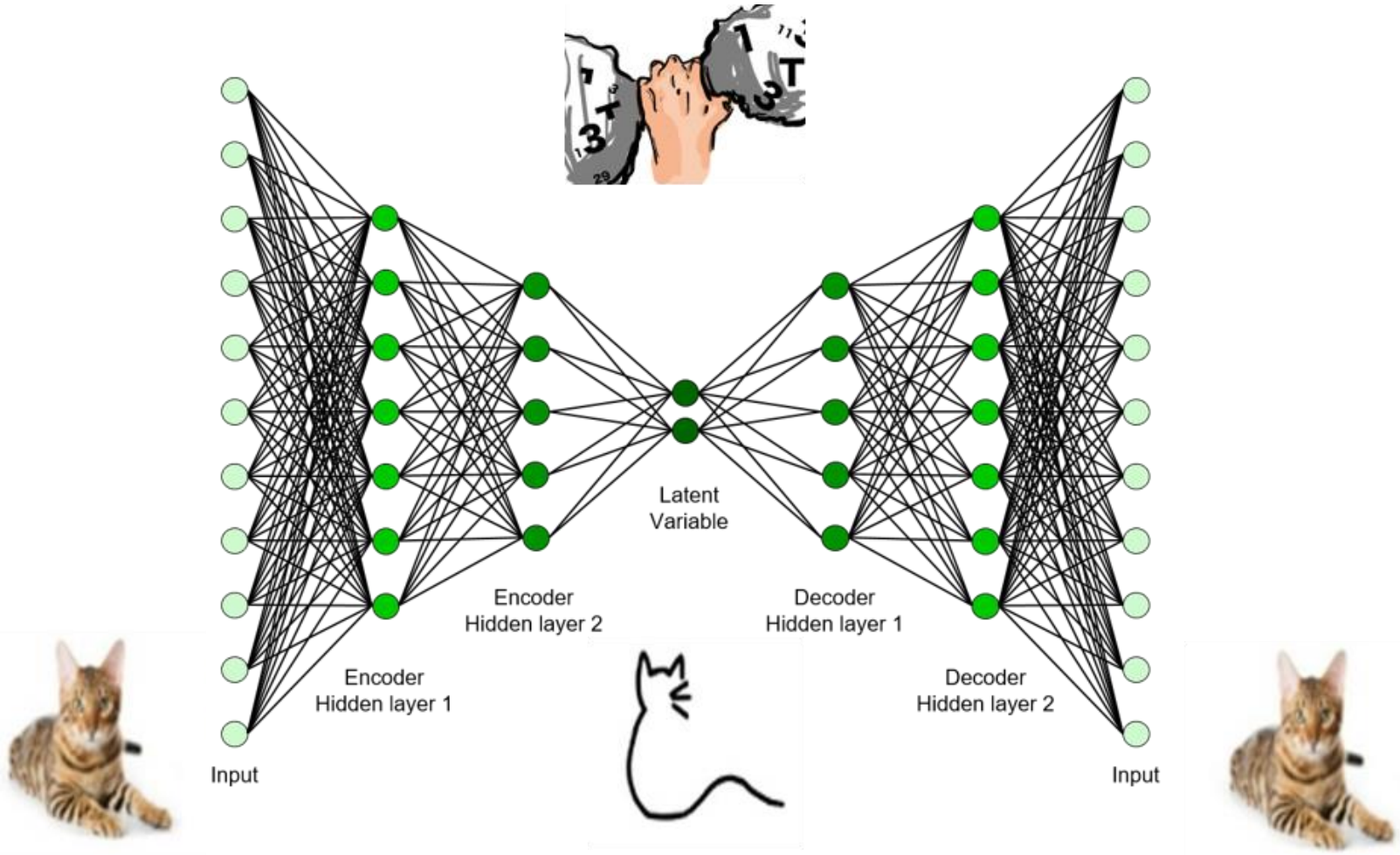
Internal Use - Confidential

How to build Intelligent Security: Anomaly Detection

1. What is your protected resource? Dell.com login
   Who uses your resource? Commercial, premier, partner
2. What is the right data to collect? Behavioral information
   Can I get both positive and negative samples? Generate data for negative samples
3. Train your model
   What are the suitable features? Speed n-grams, mouse directions,timings
   Model Architecture Ensemble for multimodal classification
4. Test & refine on data (offline, test env, synthetic)
   Test on synthetic data
   Validate predictions using clustering
   Monitor predictions in live env
5. Deploy and monitor; improve/retrain wherever necessary

Internal Use - Confidential

Improvements: One-sided model to deal with class imbalance



Auto associative Neural Network

Internal Use - Confidential

How to build Intelligent Security: Monitoring

1. What is your protected resource? Dell Cloud Identity Login
   Who uses your resource? Commercial, Premier, Channel, Internal
2. What is the right data to collect? Time Series, Product logs
3. Train your model
   Model Architecture: Regressor, LSTM RNN
4. Test & refine on data (offline, test env, synthetic)
   Validate forecasted against actual
5. Deploy and monitor; improve/retrain wherever necessary

Internal Use - Confidential

How to build Intelligent Security: Monitoring

Baseline Product Traffic

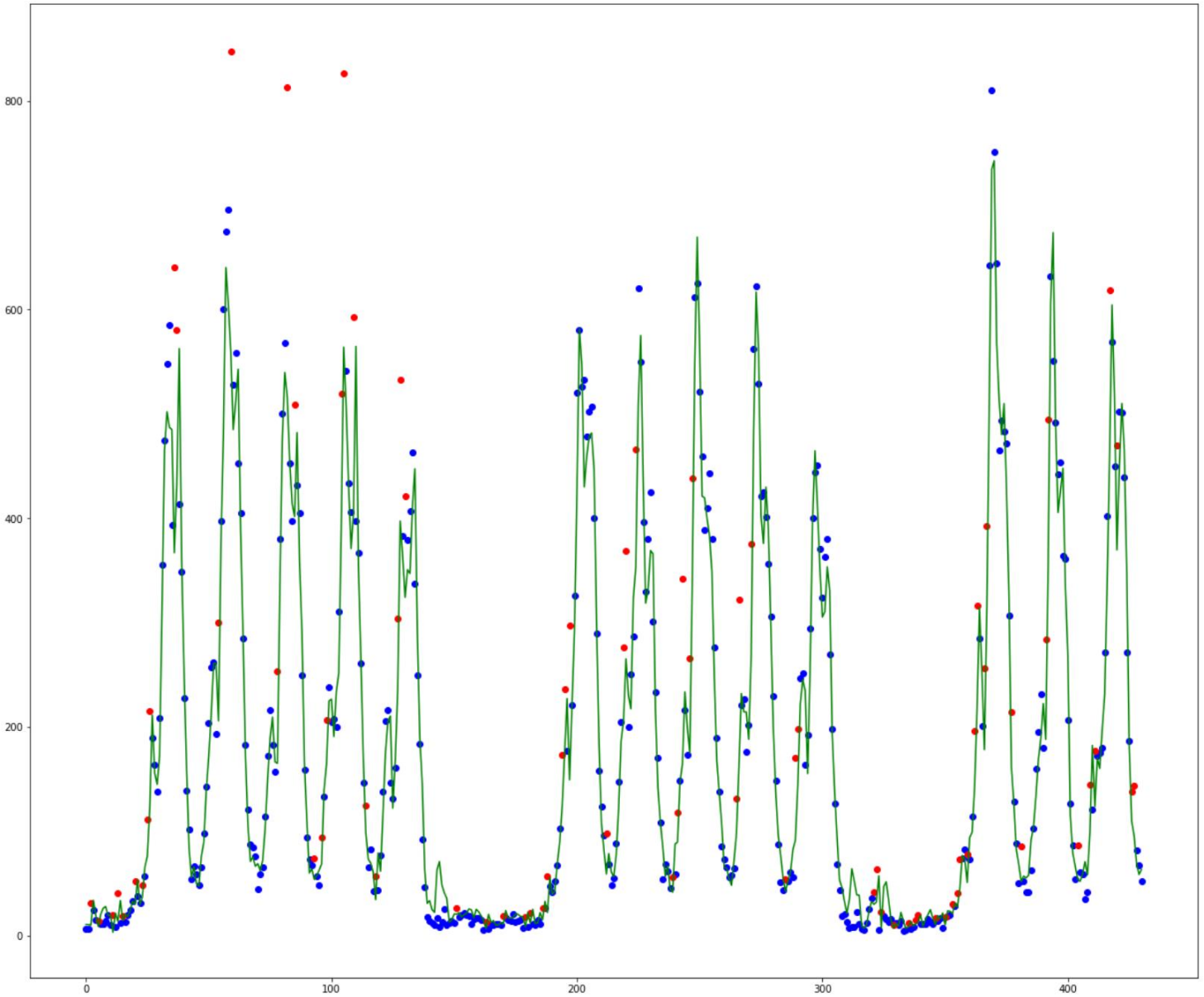
Dell Cloud Identity
per Connection (client, region)
per Activity (login, register, etc)
per Hour

Timestamp	index	ConnectionID	Activity	Attempts	Success	Failure	AccountCount
2018-08-01 00:00:00	0	43283458-e321-4b11-a7b4-50466538db52	Login	32	13	19	24
2018-08-01 00:00:00	1	43283458-e321-4b11-a7b4-50466538db52	CreateAccount	5	2	3	3
2018-08-01 00:00:00	2	43283458-e321-4b11-a7b4-50466538db52	ForgotPassword	8	2	6	7
2018-08-01 00:00:00	3	43283458-e321-4b11-a7b4-50466538db52	ResetPassword	5	3	2	2
2018-08-01 00:00:00	4	98254675-ac6f-47d9-a7e4-4537724f135d	Login	16	8	8	12
2018-08-01 00:00:00	5	98254675-ac6f-47d9-a7e4-4537724f135d	CreateAccount	1	0	1	1
2018-08-01 00:00:00	6	98254675-ac6f-47d9-a7e4-4537724f135d	ForgotPassword	3	2	1	3
2018-08-01 00:00:00	7	98254675-ac6f-47d9-a7e4-4537724f135d	ResetPassword	3	3	0	3
2018-08-01 00:00:00	8	f454c791-0fe0-4adc-ba08-e94f97d20ab9	Login	0	0	0	0
2018-08-01 00:00:00	9	f454c791-0fe0-4adc-ba08-e94f97d20ab9	CreateAccount	1	1	0	0
2018-08-01 00:00:00	10	f454c791-0fe0-4adc-ba08-e94f97d20ab9	ForgotPassword	0	0	0	0
2018-08-01 00:00:00	11	f454c791-0fe0-4adc-ba08-e94f97d20ab9	ResetPassword	0	0	0	0
2018-08-01 00:00:00	12	ffe8cec9-6142-430d-a90a-88bf736701e3	Login	38	31	7	29
2018-08-01 00:00:00	13	ffe8cec9-6142-430d-a90a-88bf736701e3	CreateAccount	1	0	1	1
2018-08-01 00:00:00	14	ffe8cec9-6142-430d-a90a-88bf736701e3	ForgotPassword	2	2	0	0
2018-08-01 00:00:00	15	ffe8cec9-6142-430d-a90a-88bf736701e3	ResetPassword	3	2	1	3
2018-08-01 00:00:00	16	566178c1-e152-4ac5-8539-23efcc474552	Login	6	5	1	4
2018-08-01 00:00:00	17	566178c1-e152-4ac5-8539-23efcc474552	CreateAccount	0	0	0	0
2018-08-01 00:00:00	18	566178c1-e152-4ac5-8539-23efcc474552	ForgotPassword	0	0	0	0
2018-08-01 00:00:00	19	566178c1-e152-4ac5-8539-23efcc474552	ResetPassword	1	1	0	1
2018-08-01 00:00:00	20	75814991-4252-4cc4-a977-670cc21309a7	Login	18	16	2	15
2018-08-01 00:00:00	21	75814991-4252-4cc4-a977-670cc21309a7	CreateAccount	0	0	0	0

Internal Use - Confidential

No of login requests

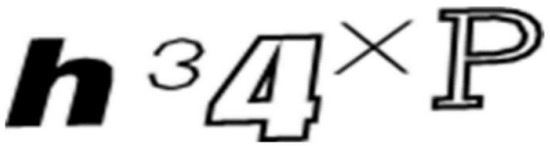
(Threshold: 20%)



Time (hourly)

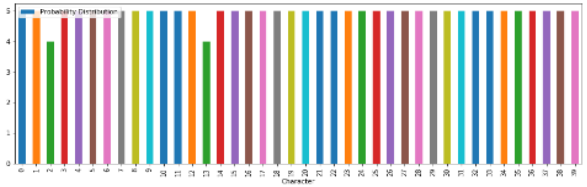
Internal Use - Confidential

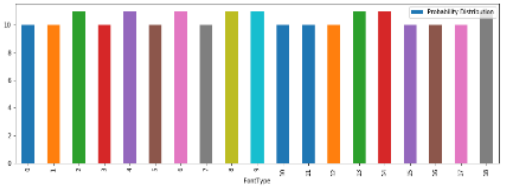
Make CAPTCHAs smart again: A framework to design Completely Automated Reverse Turing tests

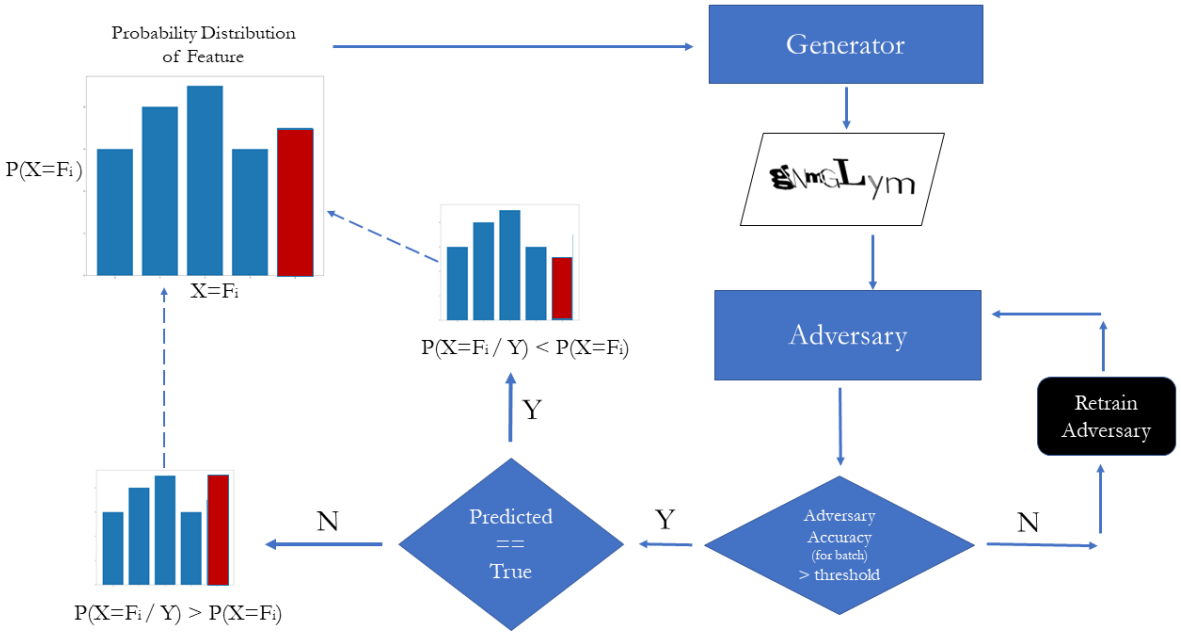


Human Preferences: Solved (not refreshed), solved correctly
Attacker Preferences: Custom deep OCR performance

Character Parameters					
Character	h	3	4	X	P
Font Type	Font 1	Font 7	Font 4	Font 3	Font 9
Font Size	74	62	73	63	77
Hollow/ Solid	Solid	Solid	Hollow	Solid	Hollow
X Coordinate	21	60	93	129	169
Y Coordinate	49	48	54	41	46
Image Parameters					
Skew Points	P1(x1,y1)	P2(x2,y2)	P3(x3,y3)	P4(x4,y4)	







Bayesian Inference: $$p(\theta|\mathbf{D}) = \frac{\mathcal{L}(\mathbf{D}|\theta)\pi(\theta)}{\mathbf{p}(\mathbf{D})}$$

US Patent: Arif Khan, Falaah and Sharma, Hari Surender. Framework to Design Completely Automated Reverse Turing Tests, authorized, in preparation

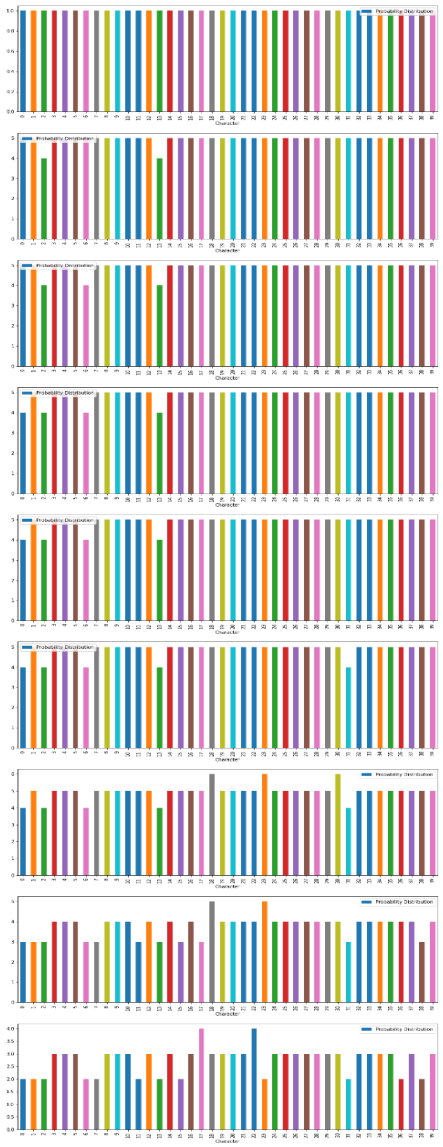
Internal Use - Confidential

How to build Intelligent Security: Step up Mechanism

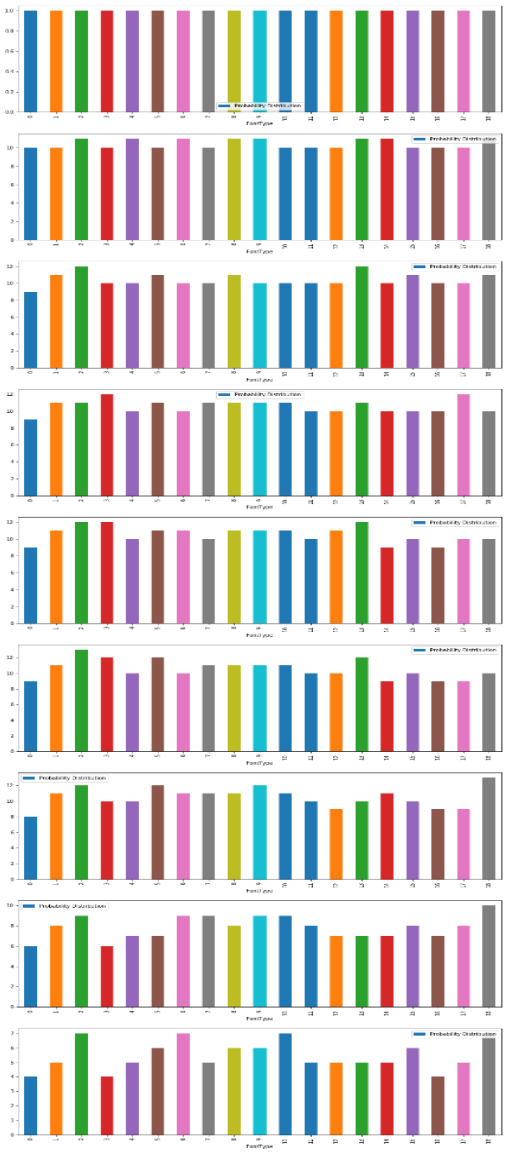
1. What is your protected resource? DCI-Captcha Microservice
   Who uses your resource? Dell logins
2. What is the right data to collect? Challenge parameters, user performance
   Can I get both positive and negative samples? Yes. Collect positive samples from client performance on challenges, collect negative samples by running an OCR on the challenge
3. Train your model
   Model Architecture Human in the loop, Online learning, Bayesian Optimization
4. Test & refine on data (offline, test env, synthetic)
   Test new challenges on OCR
   Validate client performance in test env
   Monitor both OCR and client performance
5. Deploy and monitor; improve/retrain wherever necessary

Internal Use - Confidential

How to build Intelligent Security: Step up Mechanism



Probability Distribution of Characters



Probability Distribution of Fonts

Update iteration	Accuracy of OCR
0	87
1	85.71
2	82
3	82
4	80
5	80
6	75.71
7	72.49
8	70

			Adversary Performance	
User Performance			Correct	Incorrect
	Refreshed		93	520
	Attempted	Correct	884	3300
		Incorrect	231	945

Internal Use - Confidential

Friendly Advice/ Things I've learned!

1. Intelligence ! = Automation
2. Understand your protected resource and attack perimeter
3. Feasibility > Accuracy
4. Occum's Razor is real

Internal Use - Confidential

Thank You!

(Questions)

Internal Use - Confidential