Cyber Attacks & AI Predictions
Artificial Intelligence for Infosec: Actively Learning to Mimic an Analyst

Kalyan Veeramachaneni,
MIT Institute for Data, Systems and Society & PatternEx

Joint work with:

Ignacio Arnaldo Lucas, Alfredo Cuesta-Infante, Vamsi Korrapati, Costas Bassias, Kei Li.

Overview

- Intro self
  - Artificial Intelligence - Research Scientist @ MIT CSAIL
  - InfoSec – Co-Founder @ PatternEx
  - What I have built before ?
  - Why Info sec is different than anything I have worked on?
- Unsupervised learning solutions
  - Why they are not enough?
- How to bring supervision into learning?
  - Challenges and benefits
- Metrics for evaluation

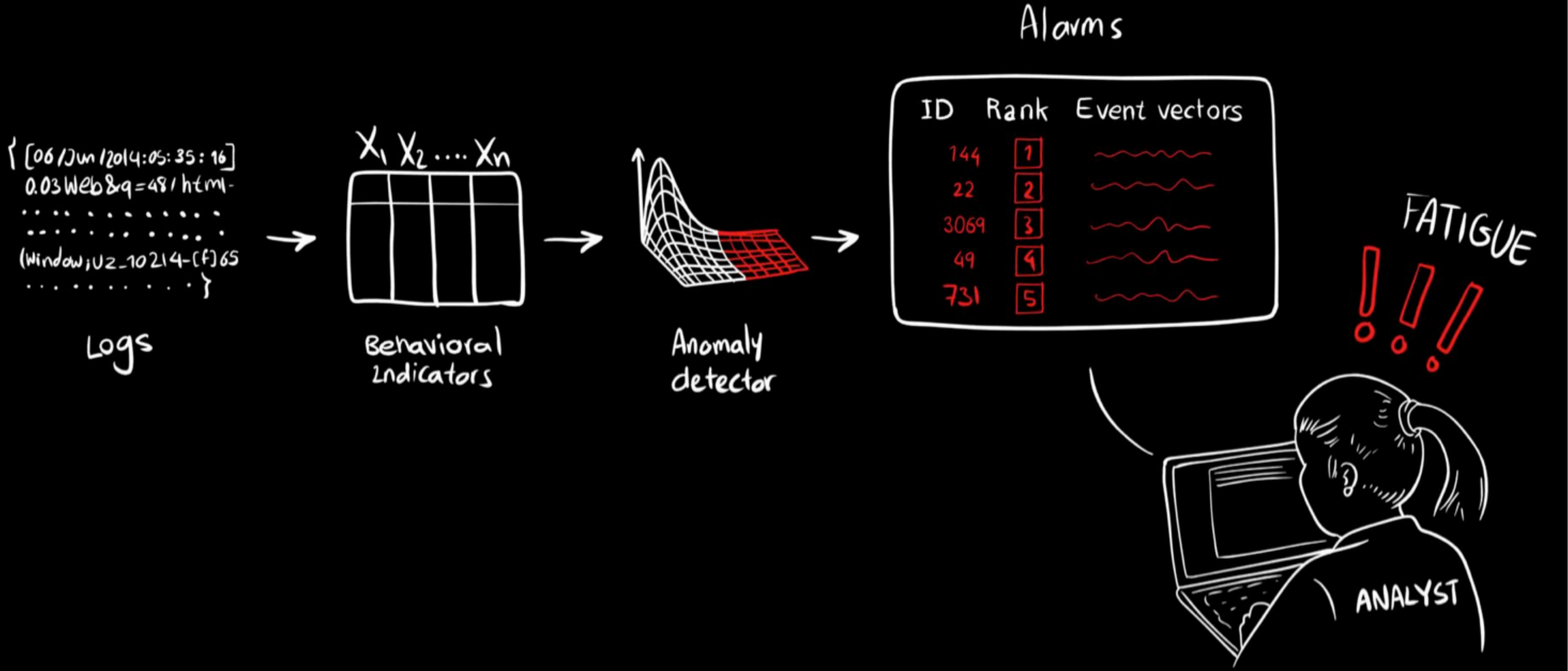
What have I built before ?

- Predict if a patient is not going to show up for the doctors appointment
- Predict what music you might like to listen when driving home
- If you liked this movie, what else would you like?
- In almost all these problems
  - We had data from past to use
  - This past data has occurrences of what we want to predict
  - Stationary – when we find that pattern that predicts, it may not change.

Why info sec is different ?

- When I started in info sec, I asked:
  - If we want to predict attacks, are there past occurrences of those in the data to learn what leads to them ?
    - Answer: No
  - If yes, can I use them to build predictive models and use them? Wouldn't that be helpful?
    - Answer: The models would be irrelevant
- So what do we do?

Unsupervised learning system



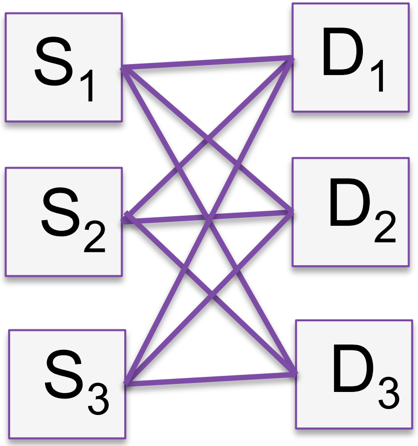
```
{ [06/Jun/2014:05:35:16]
  0.03 Web&q=48/html-
  ........
  ........
  (Window;U2_10214-(f)65
  ........}
```

ID	Rank	Event vectors
144	1	
22	2	
3069	3	
49	4	
731	5	

Why unsupervised learning is not enough?

High outlier score but not malicious

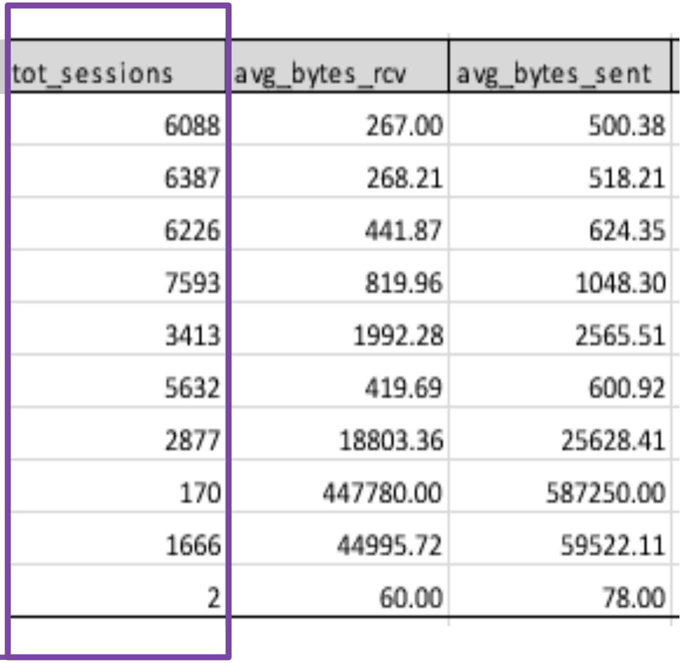
- Three hosts connecting to same 3 destinations
- Three destinations are not partner sites or known
- Connections look programmatic
  - Regular intervals
  - Same #packets in and out
  - Same duration across different hosts
  - Each source connected to all 3 destinations same number of times
  - But different sources had different number for connections
  - Perhaps bot or malware traffic ?



- Once we examined the remote host and looked at the raw data
  - Manually configured NTP systems

Why unsupervised learning is not enough?

Low outlier score but malicious



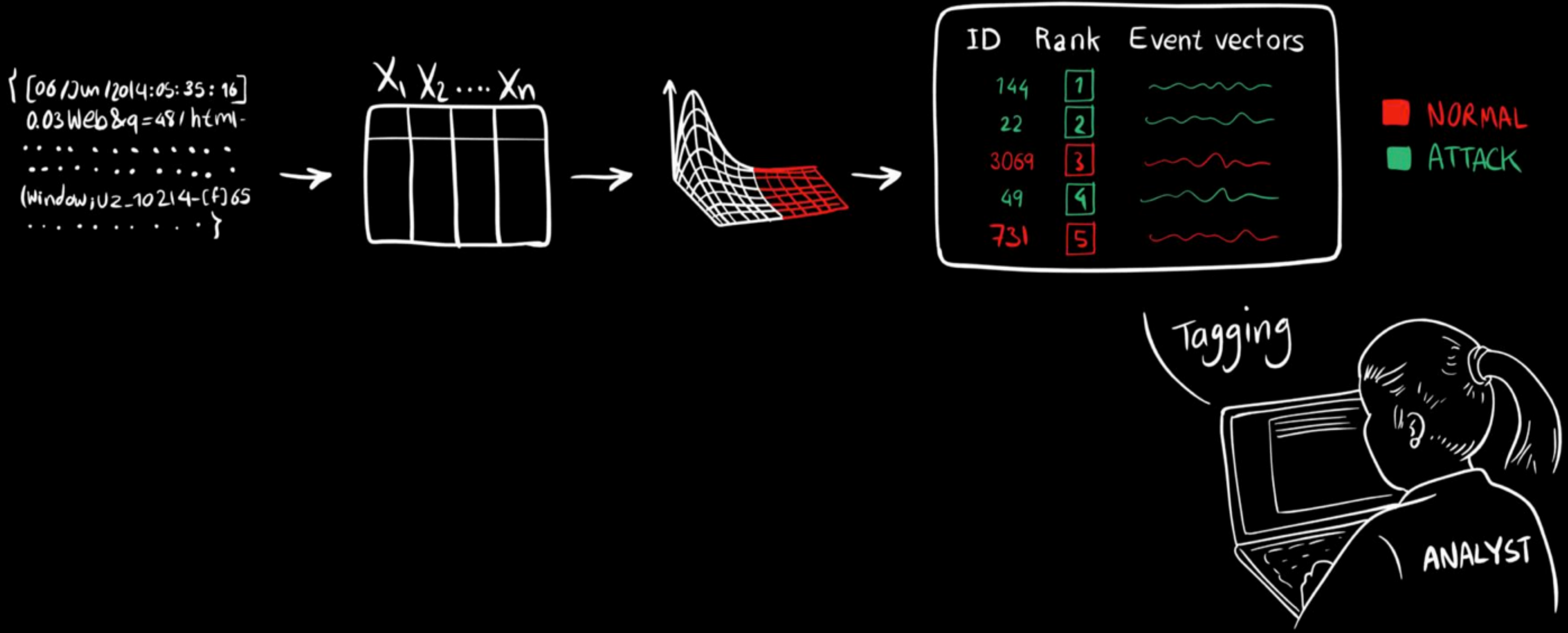
srcip	dstip	resolved	tot_sessions	avg_bytes_rcv	avg_bytes_sent
10.137..x.x			6088	267.00	500.38
10.137..x.x			6387	268.21	518.21
10.137..x.x			6226	441.87	624.35
10.137..x.x			7593	819.96	1048.30
10.137..x.x			3413	1992.28	2565.51
10.137..x.x			5632	419.69	600.92
10.137..x.x			2877	18803.36	25628.41
10.137..x.x			170	447780.00	587250.00
10.137..x.x			1666	44995.72	59522.11
10.137..x.x			2	60.00	78.00

Same source -- random remote destinations
Thousands of sessions
Very small data transfer

What did an analyst provide ?

- Subjective assessment and intuition based on
  - Looking at multiple events simultaneously
  - Collating multiple pieces of information
- Pull together external sources of information

An interactive system with analyst giving input



ID	Rank	Event vectors
144	1	
22	2	
3069	3	
49	4	
731	5	

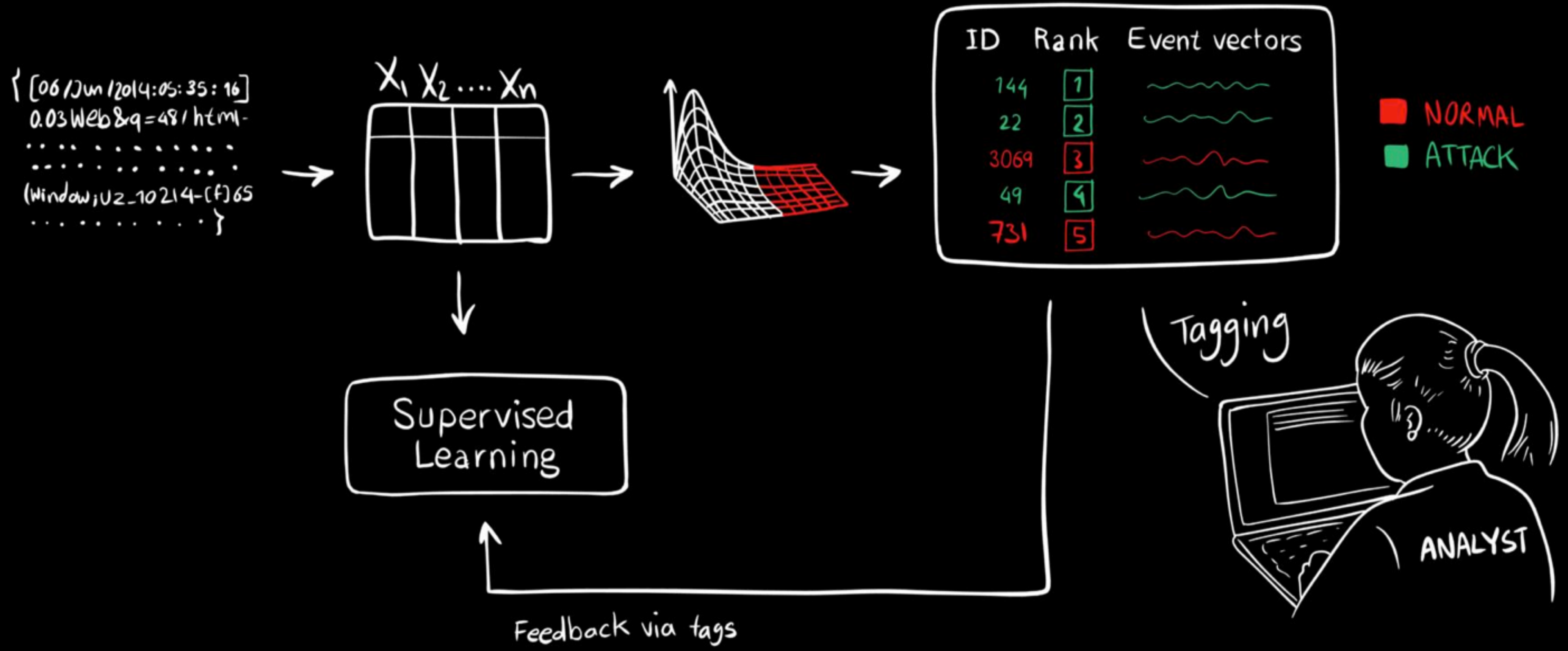
🟥 NORMAL
🟩 ATTACK

What are the challenges ?

In getting human input

- Expert sourcing
  - Not crowd sourcing, or even customer sourcing
- Limited bandwidth
- What information to show?
- How to capture most input?
  - Tags, text, or even write code?

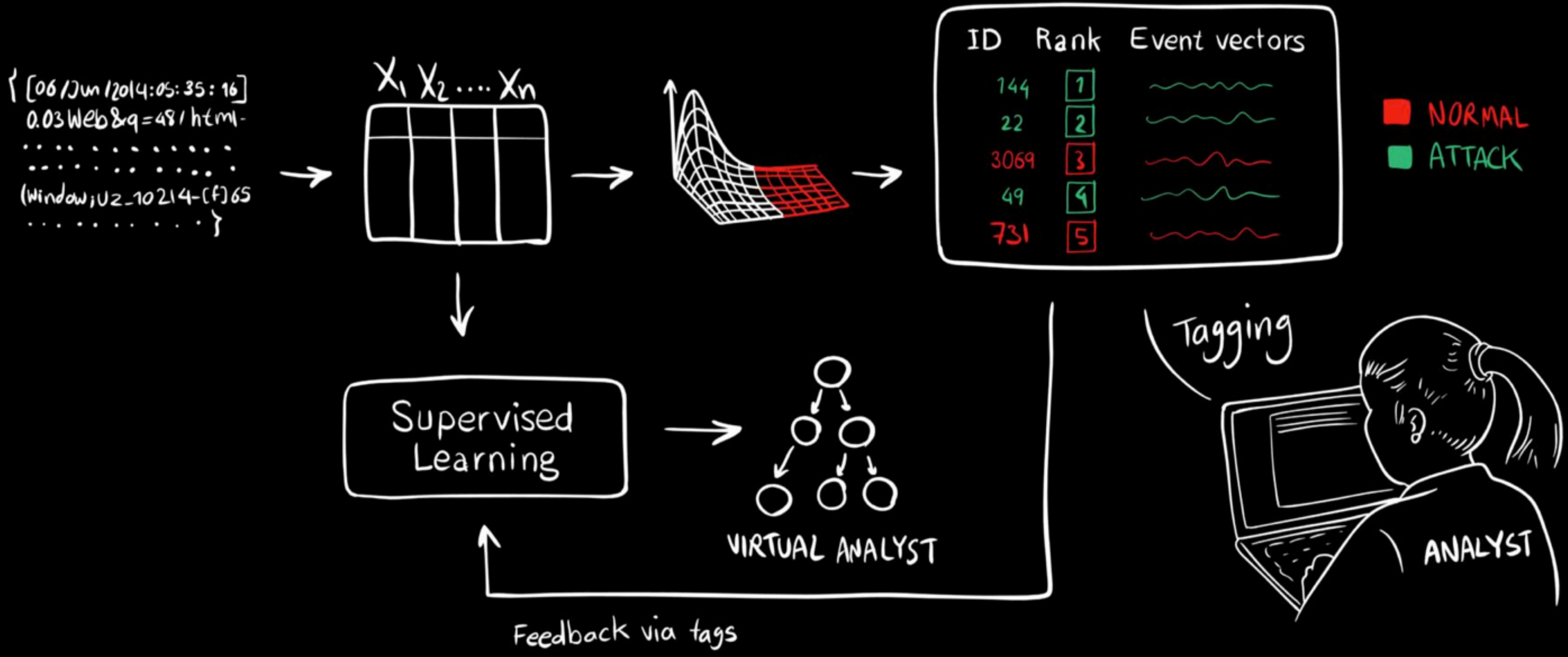
Mimicking an analyst



ID	Rank	Event vectors
144	1	
22	2	
3069	3	
49	4	
731	5	

🟥 NORMAL
🟩 ATTACK

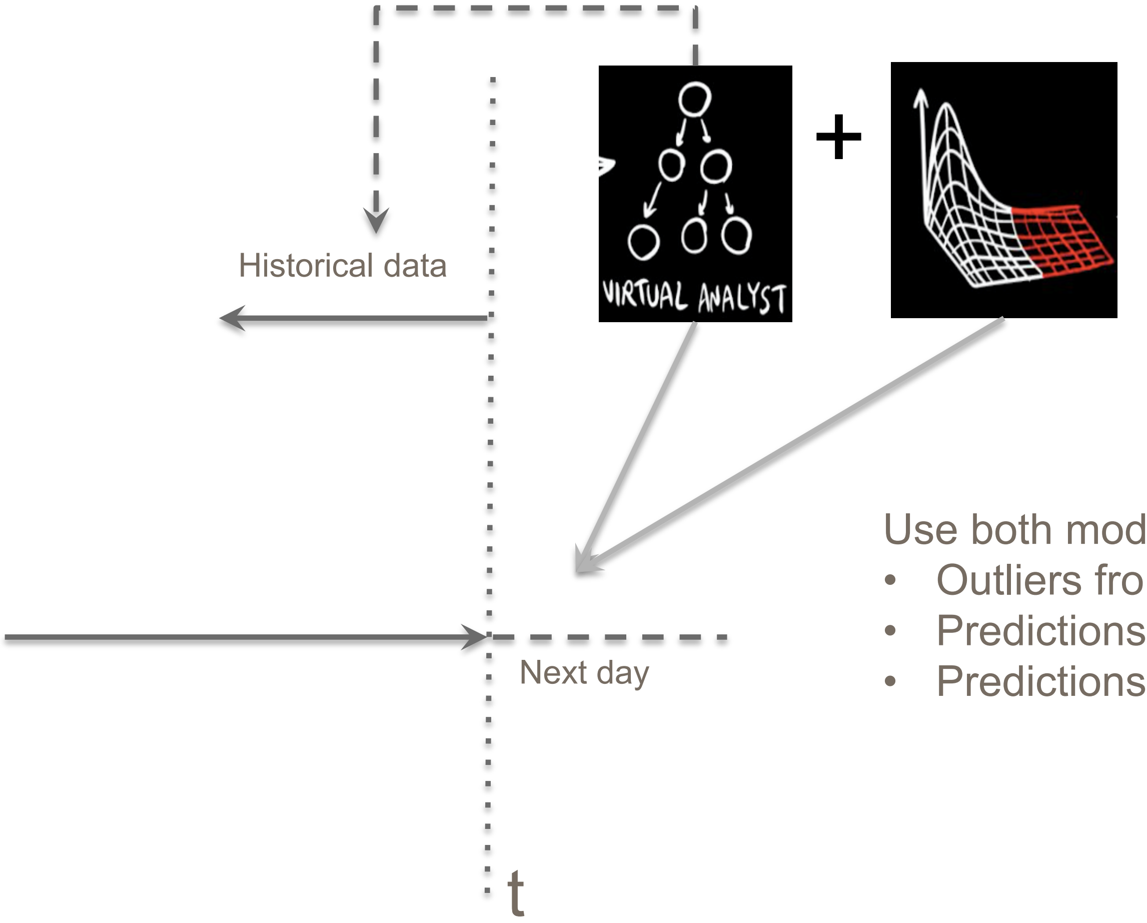
Mimicking an analyst



ID	Rank	Event vectors
144	1	
22	2	
3069	3	
49	4	
731	5	

🟥 NORMAL
🟩 ATTACK

Next day

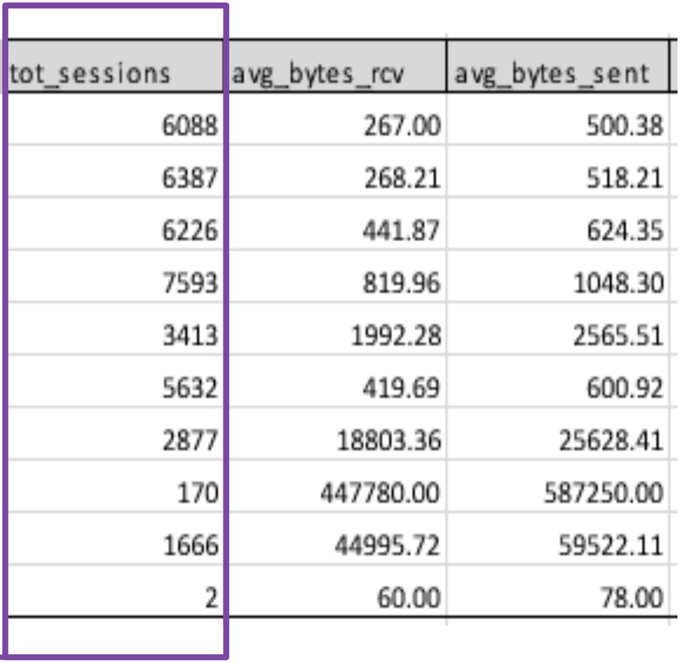


Use both models and show analyst:

- Outliers from day t+1
- Predictions from Virtual Analyst on new data
- Predictions from Virtual Analyst on old data

Going back to our example

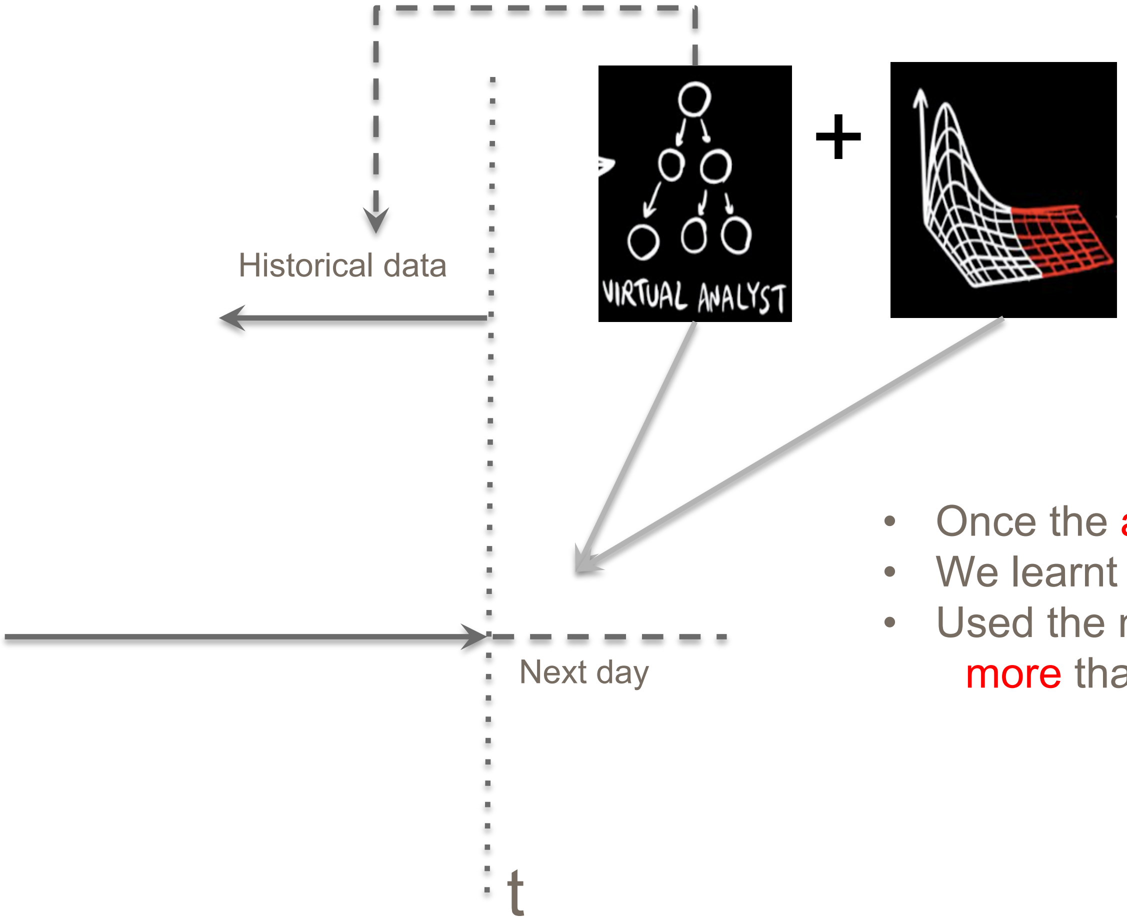
Low outlier score but malicious



srcip	dstip	resolved	tot_sessions	avg_bytes_rcv	avg_bytes_sent
10.137..x.x			6088	267.00	500.38
10.137..x.x			6387	268.21	518.21
10.137..x.x			6226	441.87	624.35
10.137..x.x			7593	819.96	1048.30
10.137..x.x			3413	1992.28	2565.51
10.137..x.x			5632	419.69	600.92
10.137..x.x			2877	18803.36	25628.41
10.137..x.x			170	447780.00	587250.00
10.137..x.x			1666	44995.72	59522.11
10.137..x.x			2	60.00	78.00

Same source -- random remote destinations
Thousands of sessions
Very small data transfer

Using virtual analyst on historical data



Historical data

Next day

t

- Once the analyst tagged 10 low outlier events
- We learnt a virtual analyst
- Used the model on historical data and found 27 more that were low on the outlier scale

What are the challenges ?

In getting human input

Expert sourcing

- Not crowd sourcing, or even customer sourcing

Limited bandwidth

What information to show?

How to capture most input?

- Tags, text, or even write code?

Dynamic learning and updating

Thin label space

- Only 10 or 20 positive labels per day

Deploying and updating on a continuous basis

Peer reviewed research paper

Experimental Setup

Real world data set with reported attacks

- 3.6B log lines
- 70.2M entities
- 318 known attacks

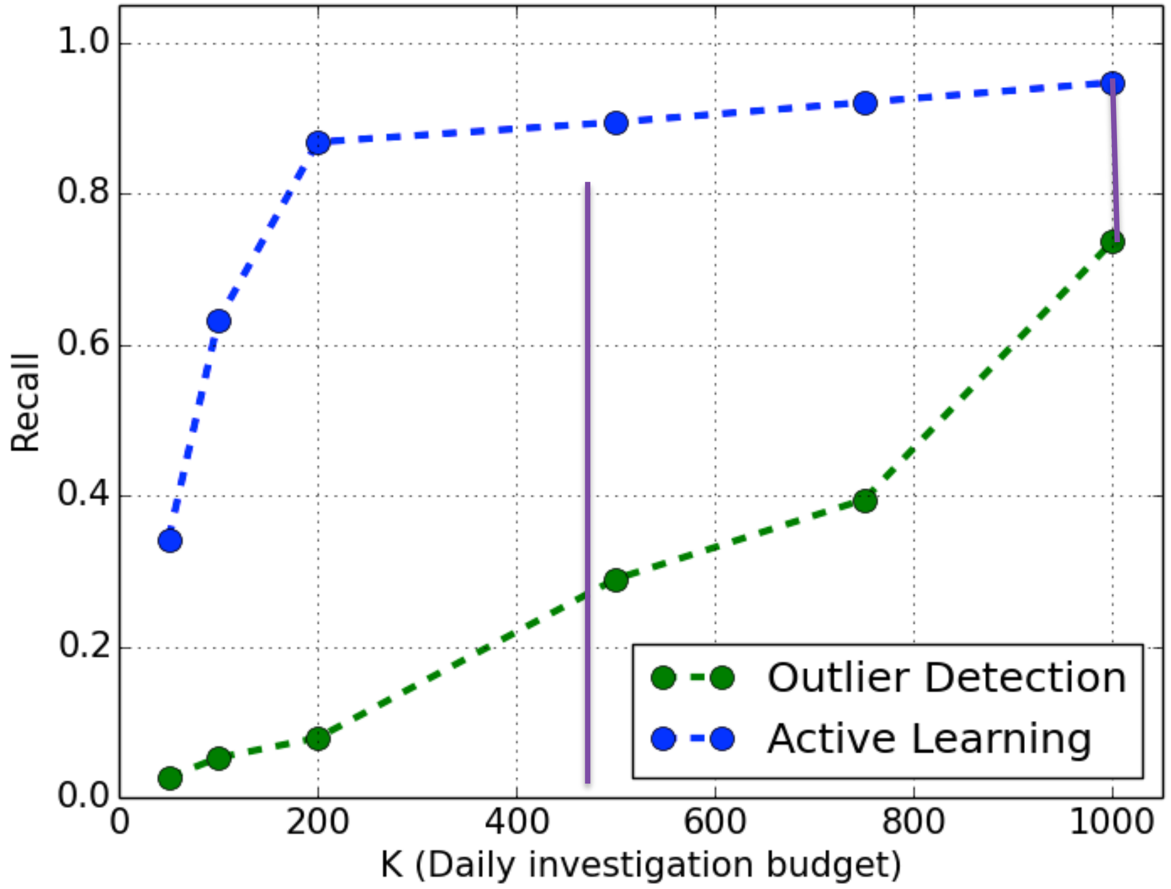
Results

Our system is bootstrapped without labeled data

The detection rate improves over time

Unsupervised-alone approaches captured a tiny fraction of the attacks

Results - Putting virtual analysts to use

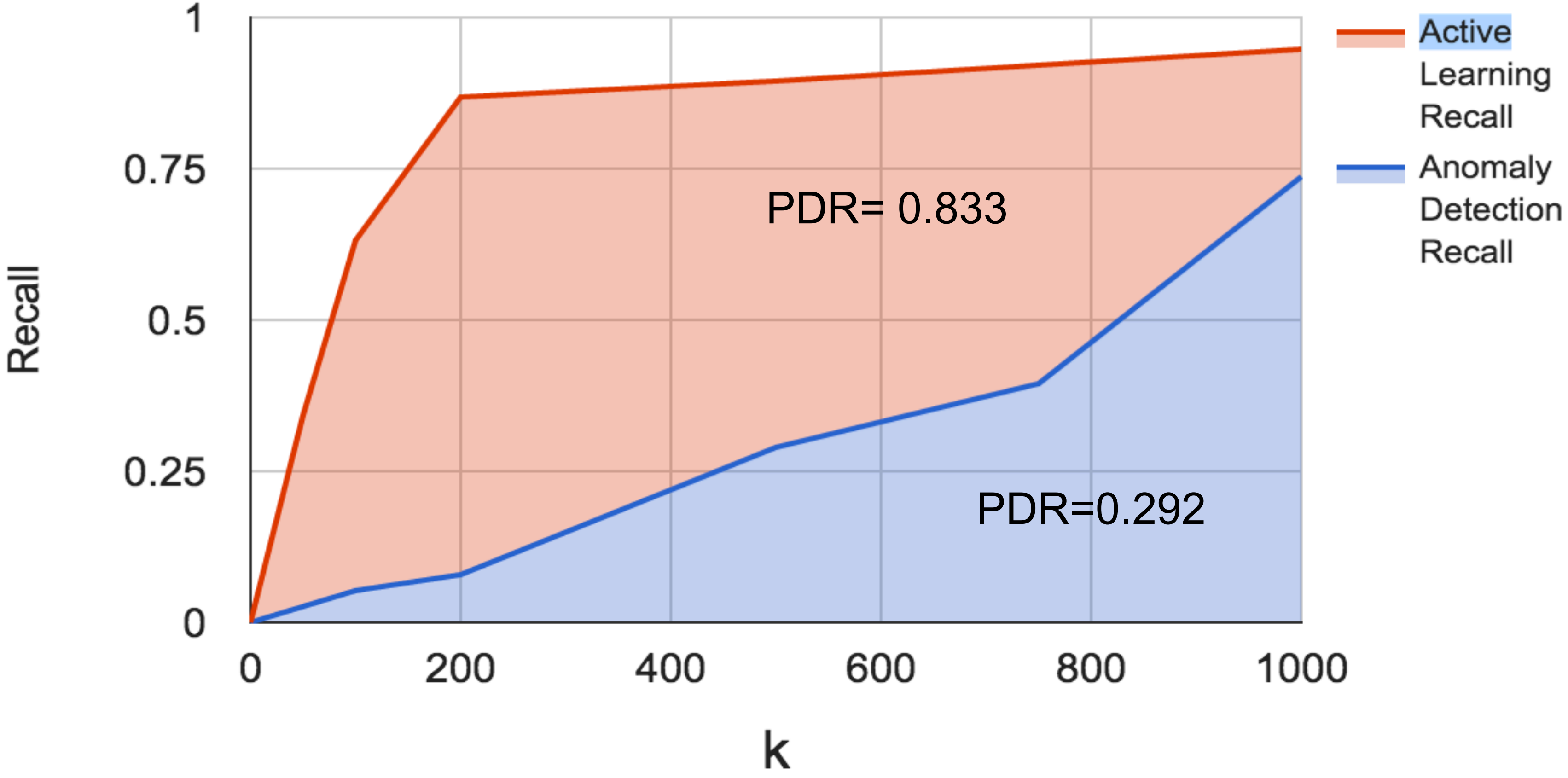


At K=200 Alerts, AI approach achieves 0.85 recall

At K=200, Outlier Detection achieves only 0.15 recall

Measure - Pattern detection ratio

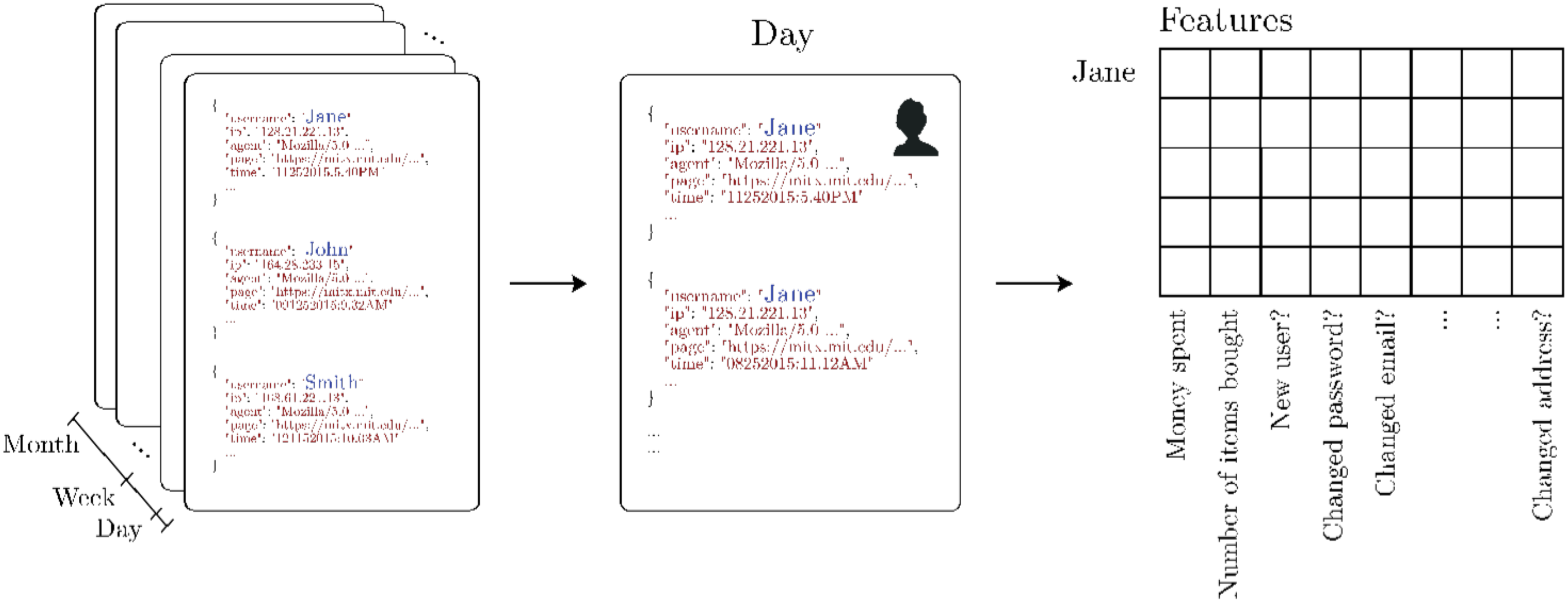
Pattern Detection Ratio – Ratio of AUC to Maximum AUC



What did an analyst provide?

- Subjective assessment and intuition based
  - Look at multiple events simultaneously
  - Collate multiple pieces of information
- Pull together external sources of information
- Analysts are also suggesting ideas for “features” implicitly
  - Distance between the feature vector from the source to all random destinations?

Where do the features come from?



```
{
  'username': 'Jane'
  'ip': '128.21.221.13',
  'agent': 'Mozilla/5.0 ...',
  'page': 'https://mitx.mit.edu/...',
  'time': '11252015:5.40PM'
  ...
}

{
  'username': 'John'
  'ip': '164.28.233.15',
  'agent': 'Mozilla/5.0 ...',
  'page': 'https://mitx.mit.edu/...',
  'time': '09252015:9.32AM'
  ...
}

{
  'username': 'Smith'
  'ip': '168.61.22.13',
  'agent': 'Mozilla/5.0 ...',
  'page': 'https://mitx.mit.edu/...',
  'time': '12112015:10.08AM'
  ...
}
```

```
{
  "username": "Jane"
  "ip": "128.21.221.13",
  "agent": "Mozilla/5.0 ...",
  "page": "https://mitx.mit.edu/...",
  "time": "11252015:5.40PM"
  ...
}

{
  "username": "Jane"
  "ip": "128.21.221.13",
  "agent": "Mozilla/5.0 ...",
  "page": "https://mitx.mit.edu/...",
  "time": "08252015:11.12AM"
  ...
}

...
...
```

Data Scientist vs. Security Analyst

Data Scientist

Data Scientist - features

- Follow one to many relationships
- Sessions → Duration
- Averages, Standard deviations, trends and other mathematical/ statistical functions.

Security Analyst

Security Analyst - features

- Number of unique applications
(HTTP, SSL, Skype, Streaming media, DNS..)
- Number of protocols being used (UDP, TCP, etc).
- Number of times the traffic originates from a reserved port.

Key takeaways

It is essential to build an analyst in-the-loop system to develop a truly adaptive artificial intelligence system

Replicating analysts intuition through models in real time is critical

- So as to stay relevant

Analyst bandwidth is the real metric

- Because you can achieve arbitrarily high true positive rate, if you make them investigate everything
- Or achieve zero false positive rate, if you don’t show anything

What you can do?

False Positives

True Positives

Number of Alerts

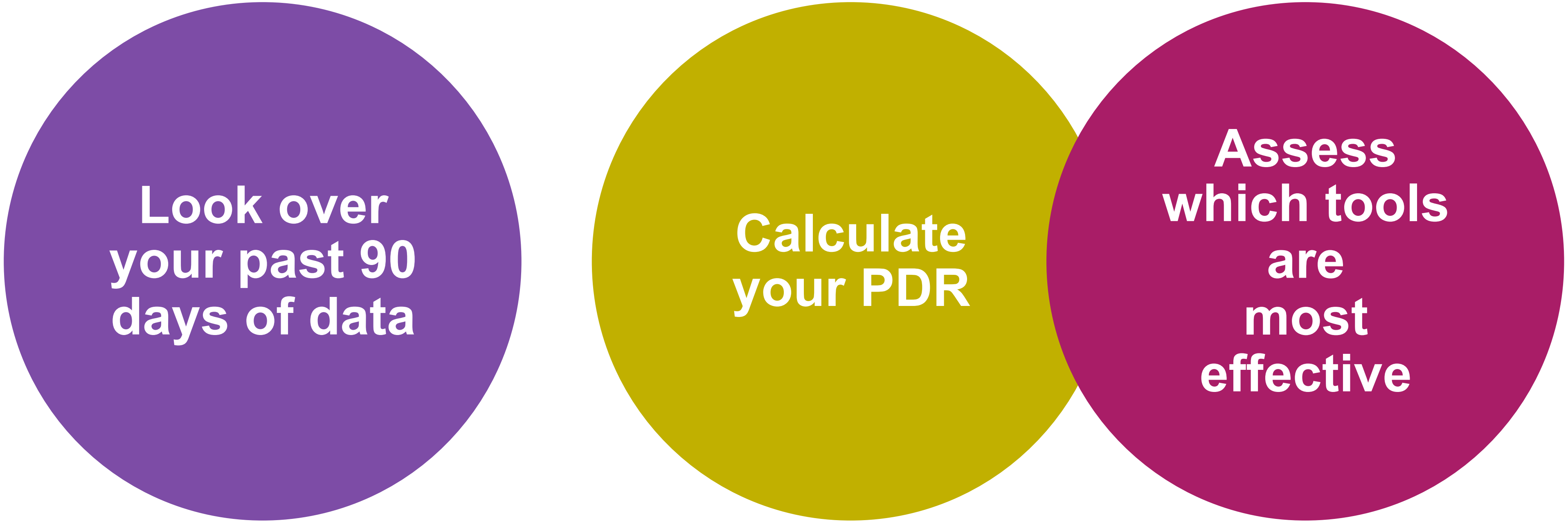
Measure PDR

Maintain PDR for every system that you use for detection and observe how it changes over time

What you can do?

Next week

First three months



Look over your past 90 days of data

Calculate your PDR

Assess which tools are most effective