FICO Decisions

# Cyber Analytics

## Streaming transaction analytics to identify malevolent cyber intrusions

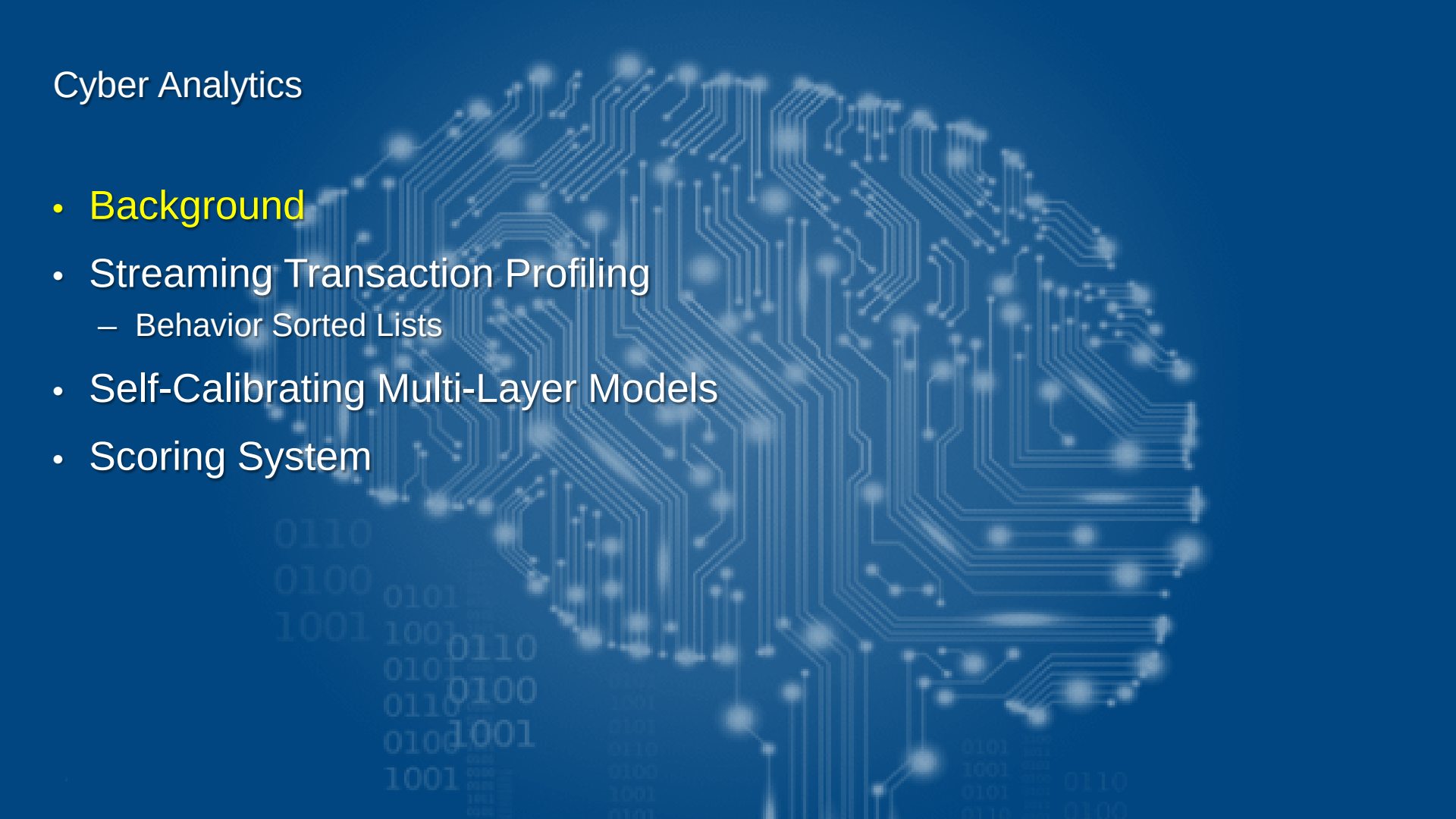

Jehan Athwal
Senior Scientist, FICO

Matthew Kennel
Principal Scientist, FICO

Scott Zoldi
Vice President, FICO

# Cyber Analytics

- Background
- Streaming Transaction Profiling
  - Behavior Sorted Lists
- Self-Calibrating Multi-Layer Models
- Scoring System



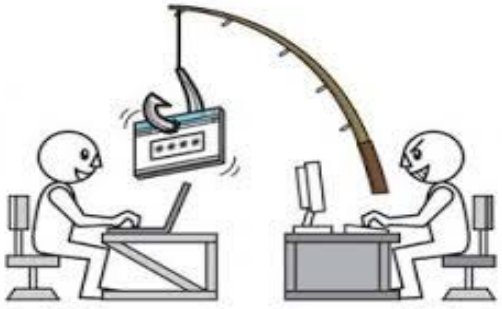
# Advanced Persistent Threat
## Carbanak 2013

- Successful spear-phishing tricked employees to install remote backdoor Carbanak .
- MS Office bugs exploited

95%

"95% of all attacks on enterprise networks are the result of successful spear phishing"

Source: Allan Paller, Director of Research - SANS Institute



Carbanak *mirrored employees screens* back to criminal cyber-lair

- Watched employees work & access internal systems
- Malefactors learnt how to steal expertly, without setting off alarms

# Why did existing systems fail?

- Standard firewall failed
  - communication was initiated from trusted internal machines, not from outside
- Standard Intrusion Detection System failed
  - looking for sites on blacklist or specific connection text signatures
- Virus scanners failed
  - looking for specific bit patterns in software
- Criminal technology is too sophisticated for existing protection tech
  - High value targets justify custom code development, avoid known signatures
  - Polymorphic, adaptive malware now productized in Dark Net

*Do you think you are immune?*

```
From: Your Supervisor/Dean
To: You
Subject: Pls review attachment on new compensation policy ASAP
```

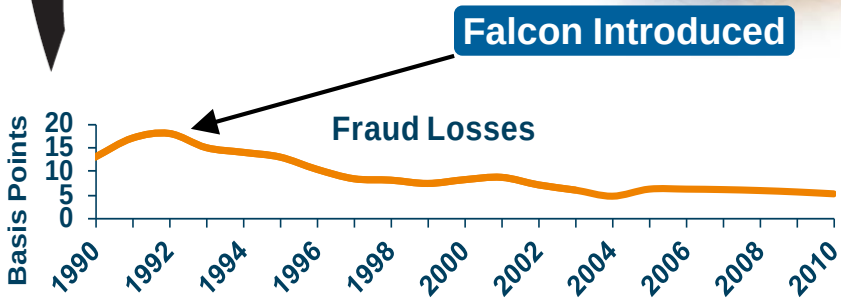
 

# Carbanak Damage

- Exploitation
  - Internal money transfer to compromised accounts
  - External money transfer via SWIFT
  - ATM network manipulated to dispense cash.
  - Up to 100 financial institutions in 30 countries affected, losses ~ $1000 million USD

- Many indicators of suspicious network behavior were present
  - Atypical connections on bank employee computers
  - Uncommon lateral movements within networks (attacker's reconnaissance of network)
  - Exfiltration of large amounts of data (mirroring of employees' screens) to previously unseen servers

# FICO's Falcon protects 65% of world's payment cards




- Supervised neural network models
- Individual cardholder profiles for streaming transactions

Rapidly assimilates data such as

- Purchase frequency
- Amounts
- Time
- Volume
- Merchant category
- Location

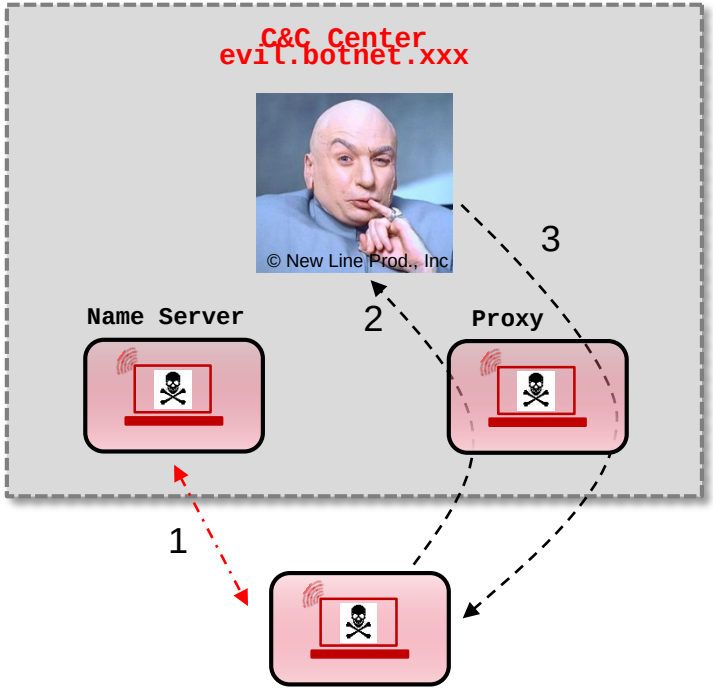
Can we analyze the network streams of client machines, servers, ports, and protocols using similar techniques?

# Difficulties of Cyber Data

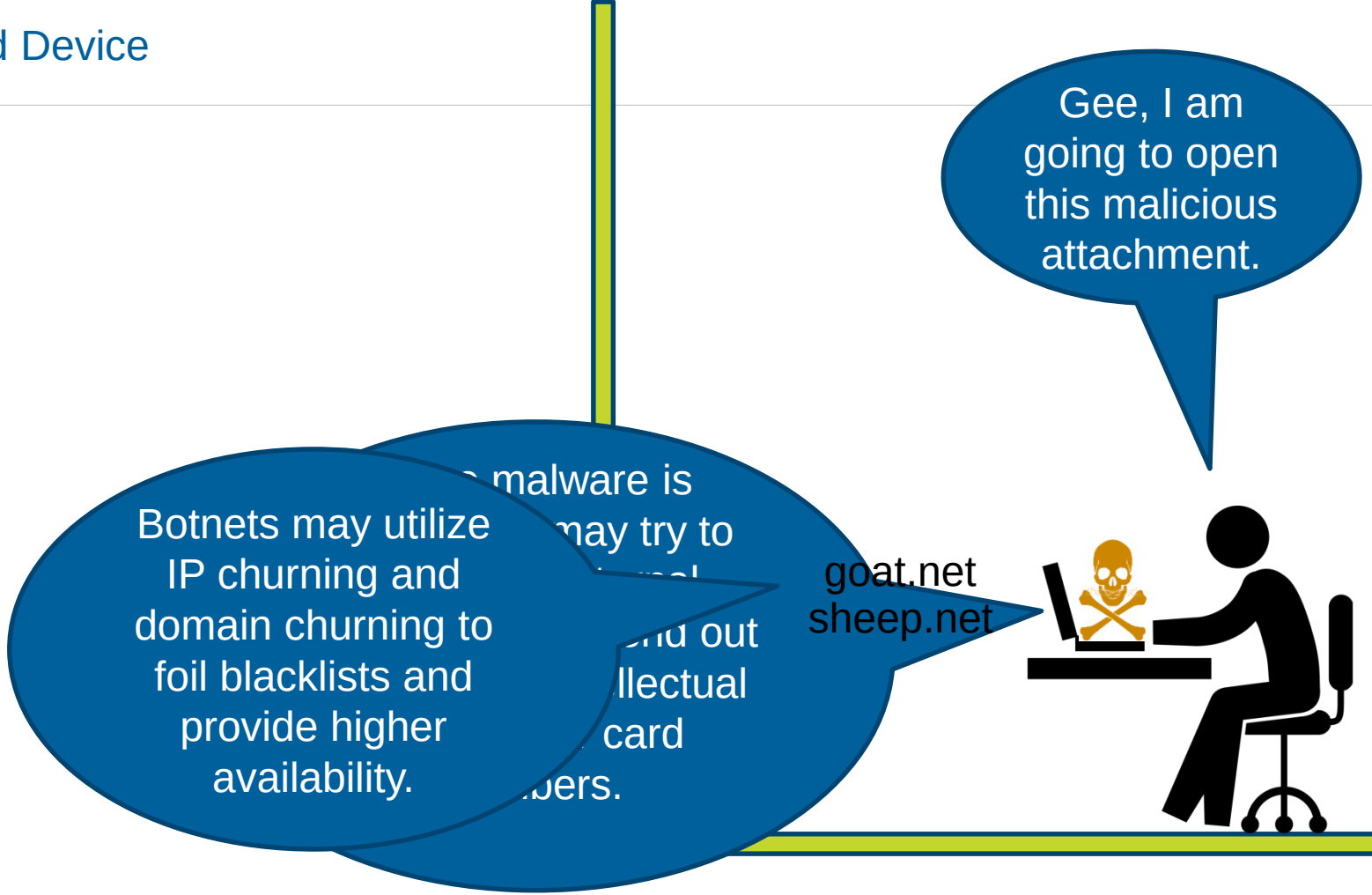
- Common models in finance---e.g. credit risk and card fraud---are supervised
  - Ground-truth labels of good and bad examples available in historical training data
- For cyber attacks, large labeled sets of network data are mythical
  - Details of attacks are secretive
  - Normal network behavior is considered sensitive
  - Data sets rarely shared
  - Data sets are enormous
  - Labeling of attacks at network packet level is exceptionally rare
  - Customers can't validate their own activity
  - Can **you** answer this question?
    - On 2015-08-10 11:32:11.023 was the DNS request from your laptop to "axm23.analytics.net" legitimate or fraudulent?

# Botnets: criminals assimilating computers
## The infected machines inside contact infected machines outside

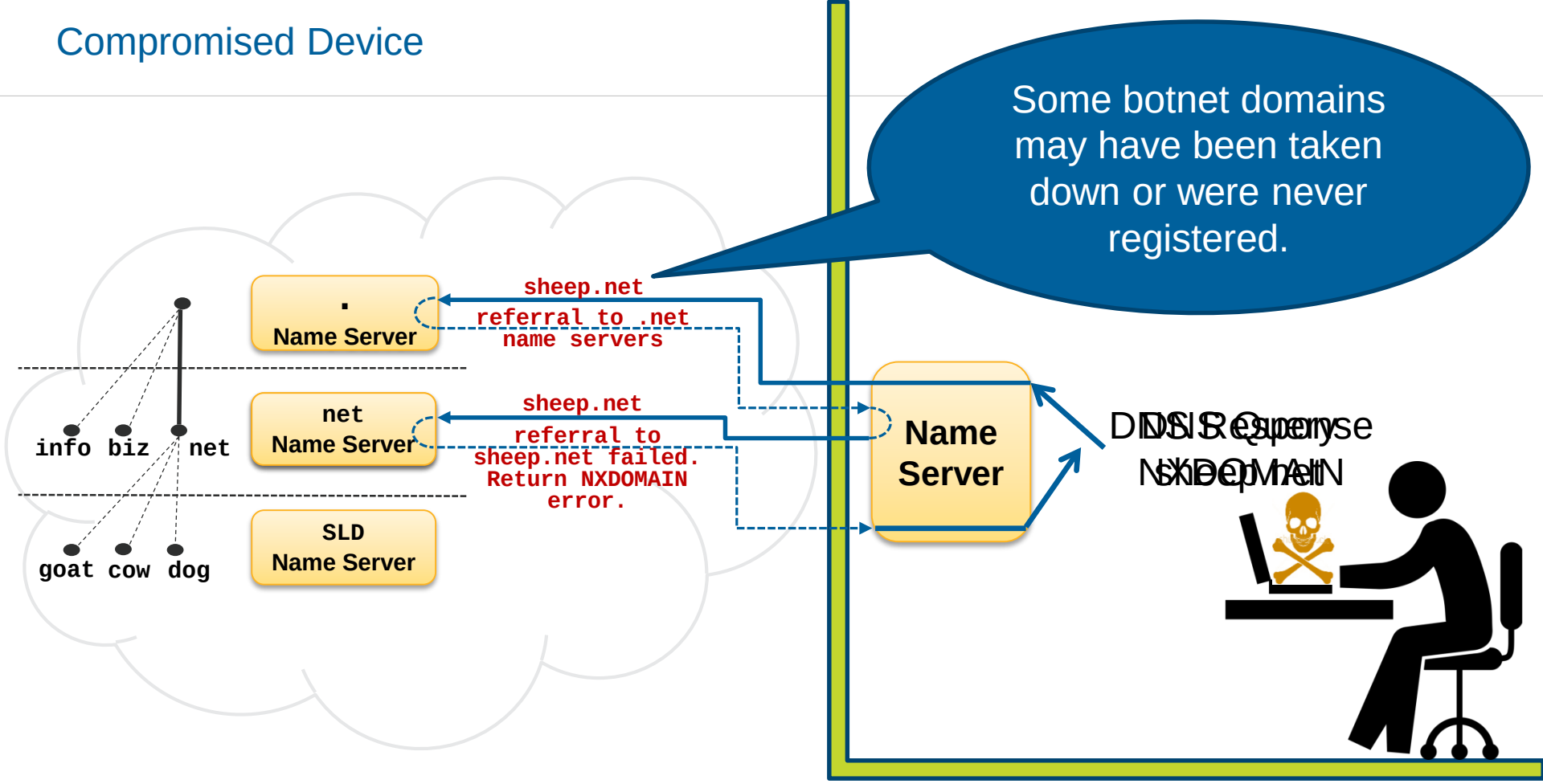
1. Botnet operator infects many computers with malware
2. Compromised machine contacts Command and Control (C&C) server via proxy
   - Beaconing: I am alive now!
     - Proxies in botnet may be offline periodically
   - Once connection is made, transmit to server sensitive data: e.g., key logs, bank account numbers, passwords, etc.
3. C&C updates bot with new instructions and code



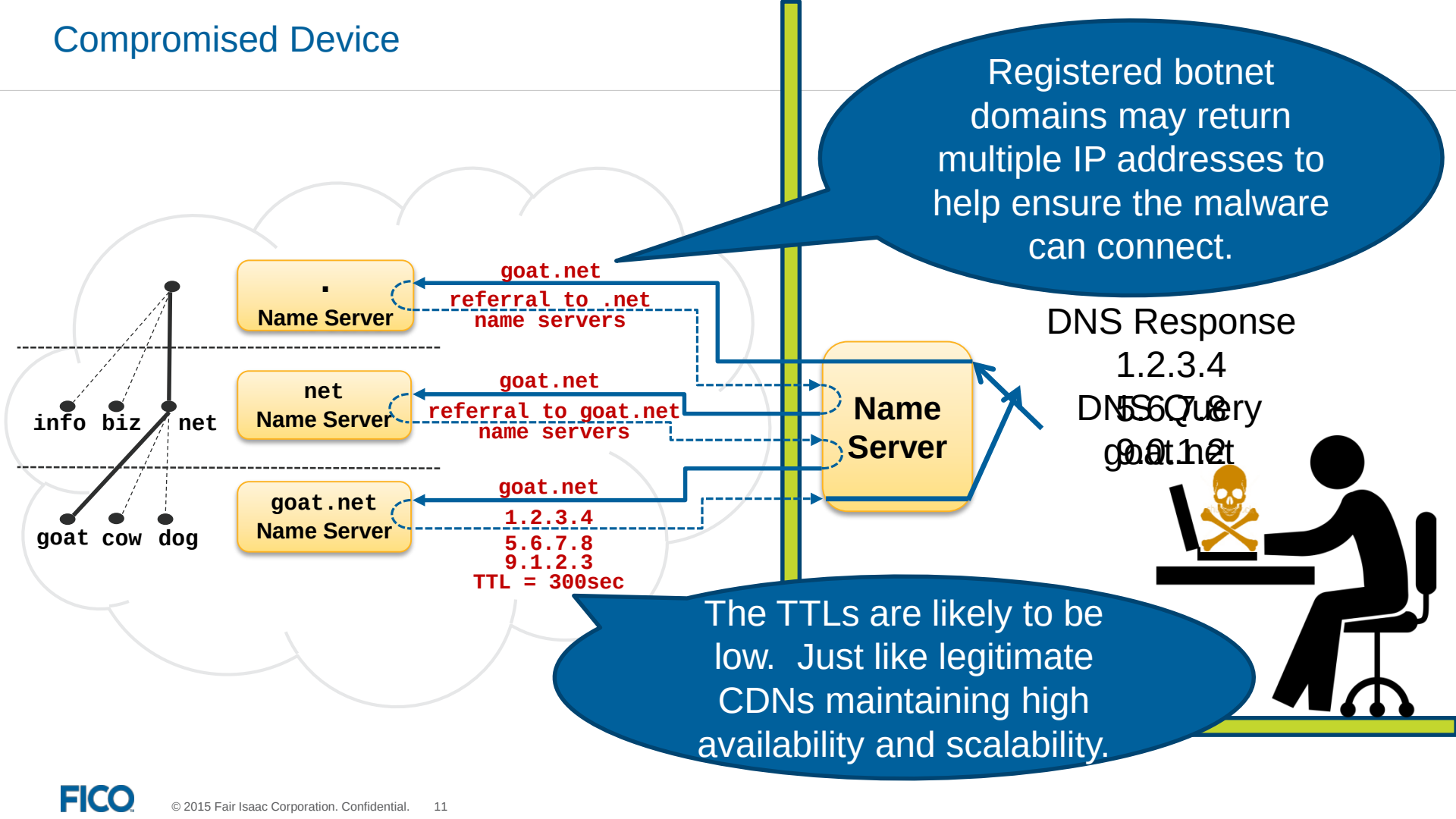
# Compromised Device



# Compromised Device

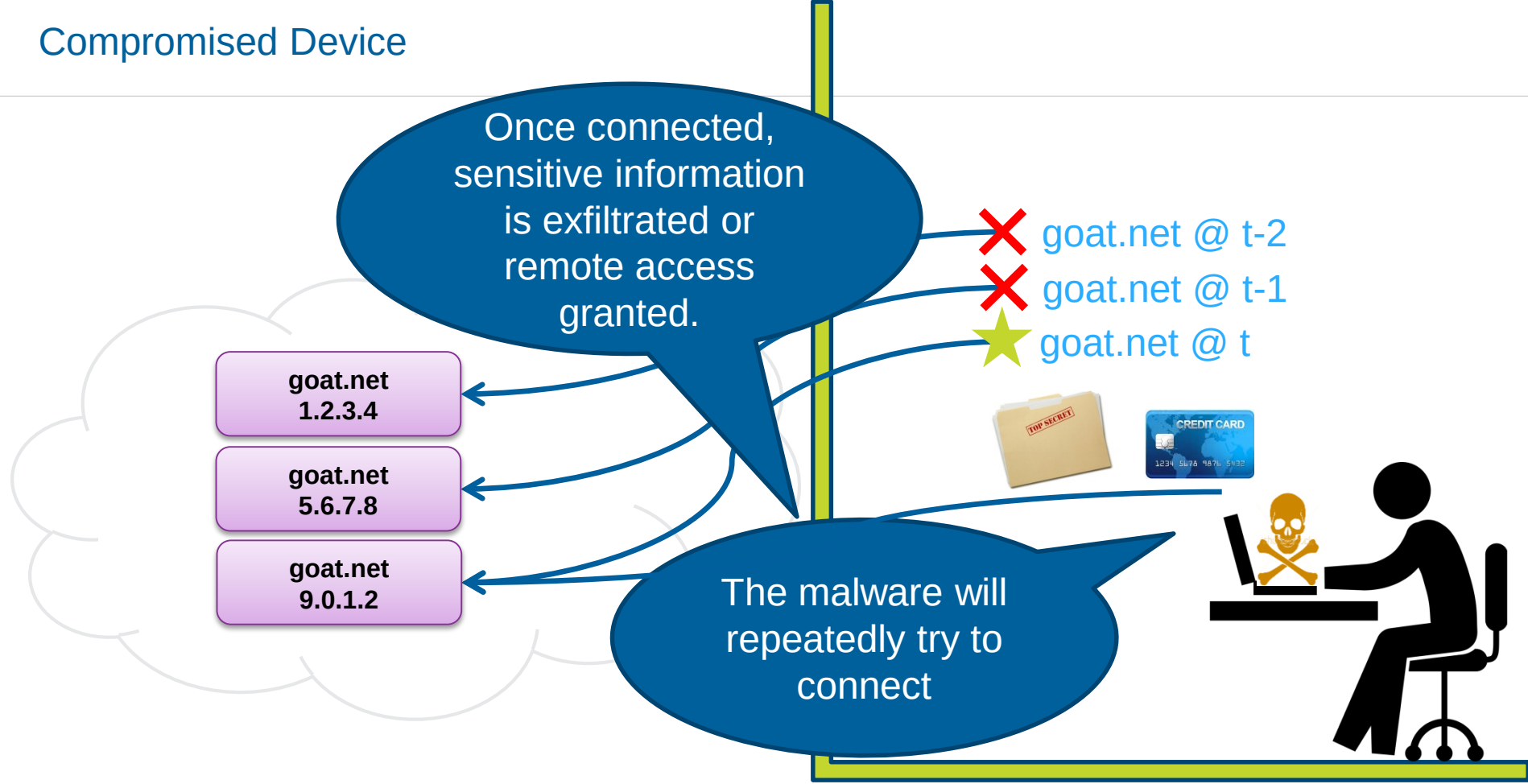


# Compromised Device



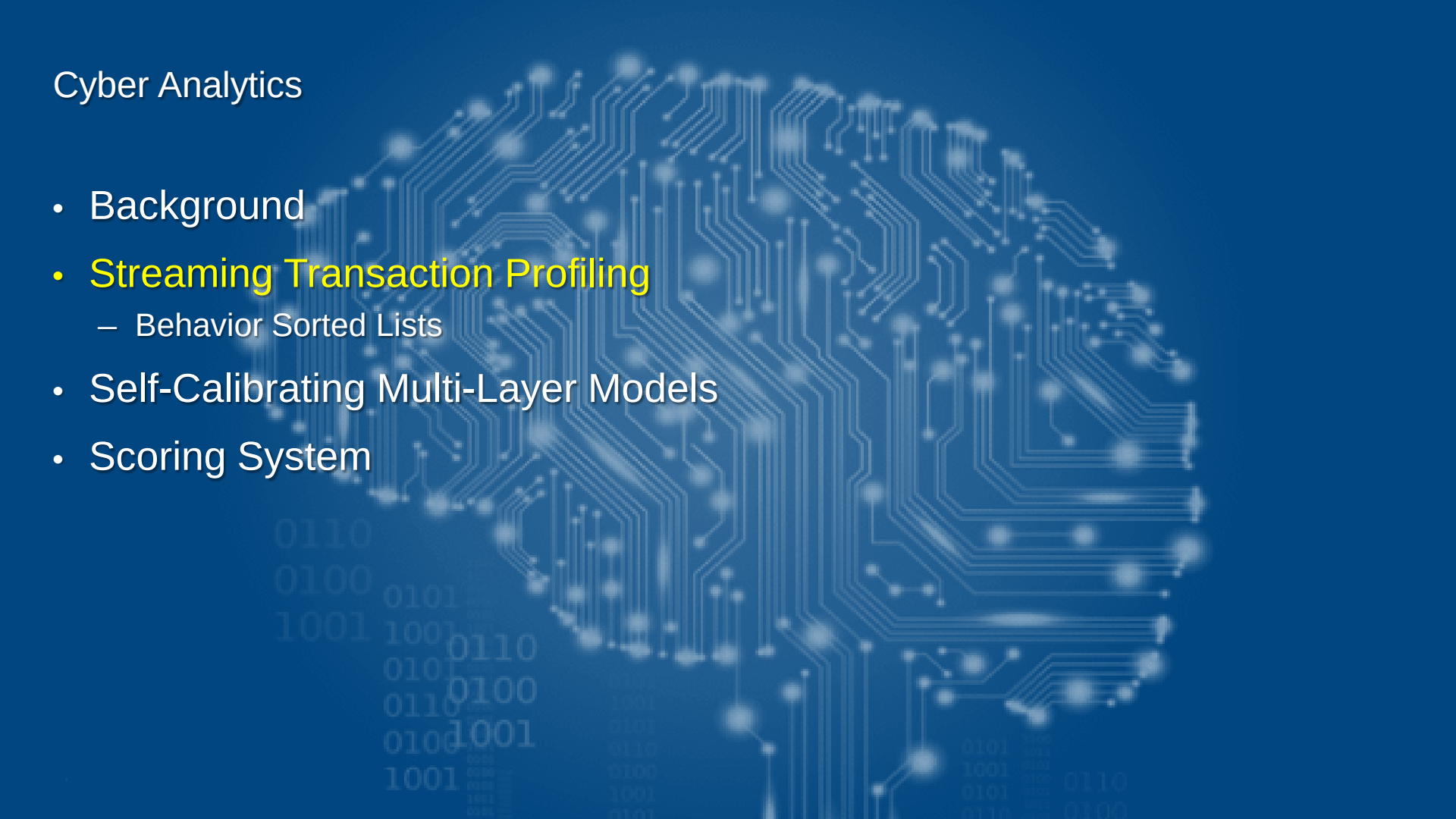

 

# Compromised Device



# Cyber Analytics

- Background
- Streaming Transaction Profiling
  - Behavior Sorted Lists
- Self-Calibrating Multi-Layer Models
- Scoring System



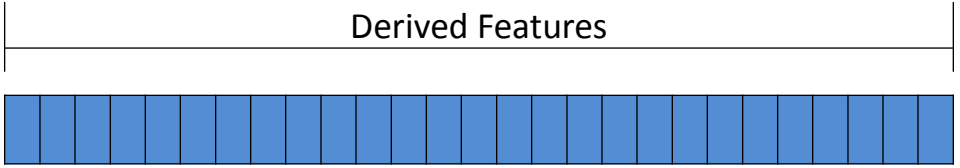
# Boiling Down Transaction Histories

## Event History

```
12:31:05, TCP, amazonaws.com, ...
12:31:05, TCP, ads.heias.com, ...
12:31:05, TCP, ib.adnxs.com, ...
12:31:05, TCP, i.ctnsnet.com, ...
...
...
...
12:42:27, UDP, pandora.com, ...
...
...
...
```

Now

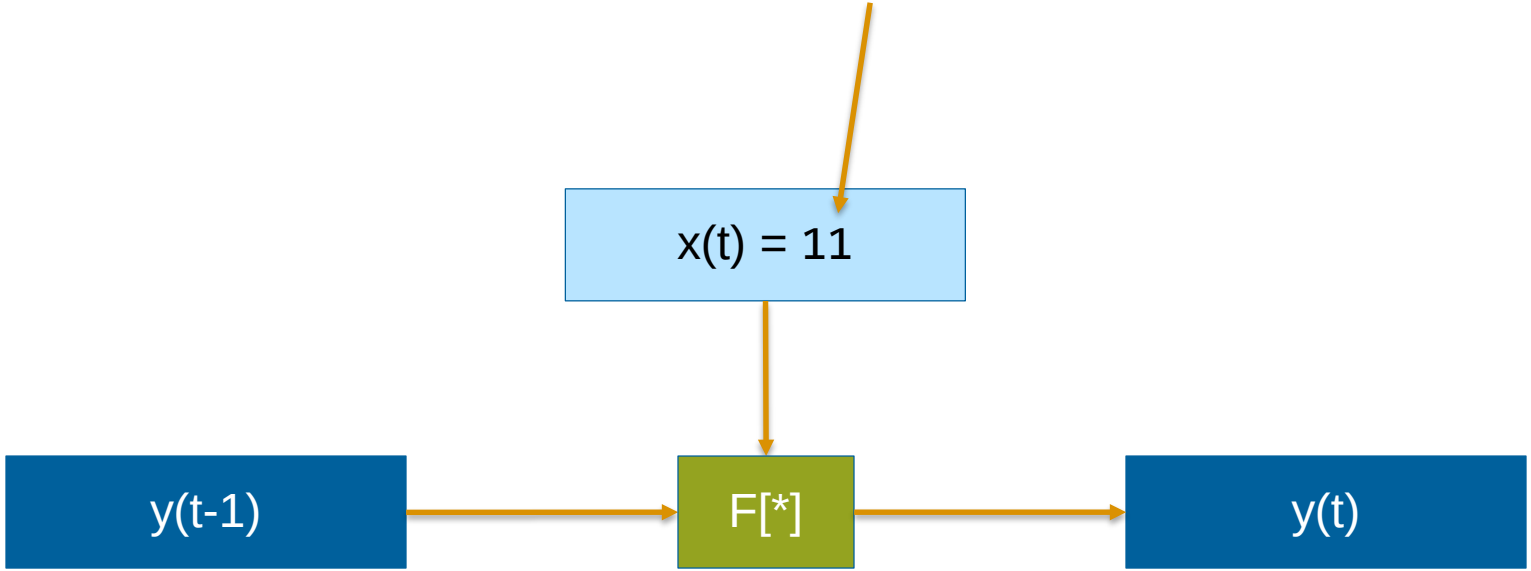
## Profile Storage




- No complex DB queries possible at scale: simple key-value store is challenging enough!
- Transaction Profiles contain recursive feature estimators to summarize derived quantities from raw data
- Decays past information smoothly over events or physical timescales
- Computationally efficient & compact – essential at cyber volumes

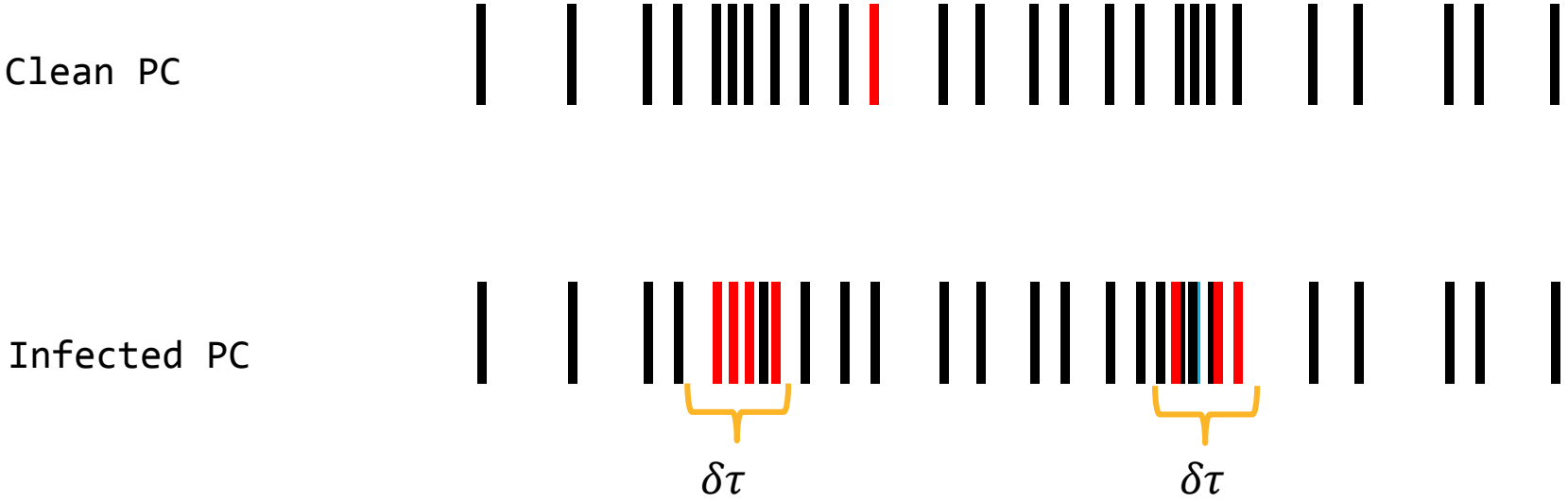
# Example of Profile Variables
## Recursive re-estimation

Incoming streaming data: 1 2 0 1 13 0 11 11 6 0 3 2 0 0 …




- F[*] = 0.9 y(t-1) + x(t)
- F[*] = the Kalman filter:  it landed men on the Moon

# Example of Profile Variables
## DNS lookup failure




Suspicious: Burst of errors in short time window.

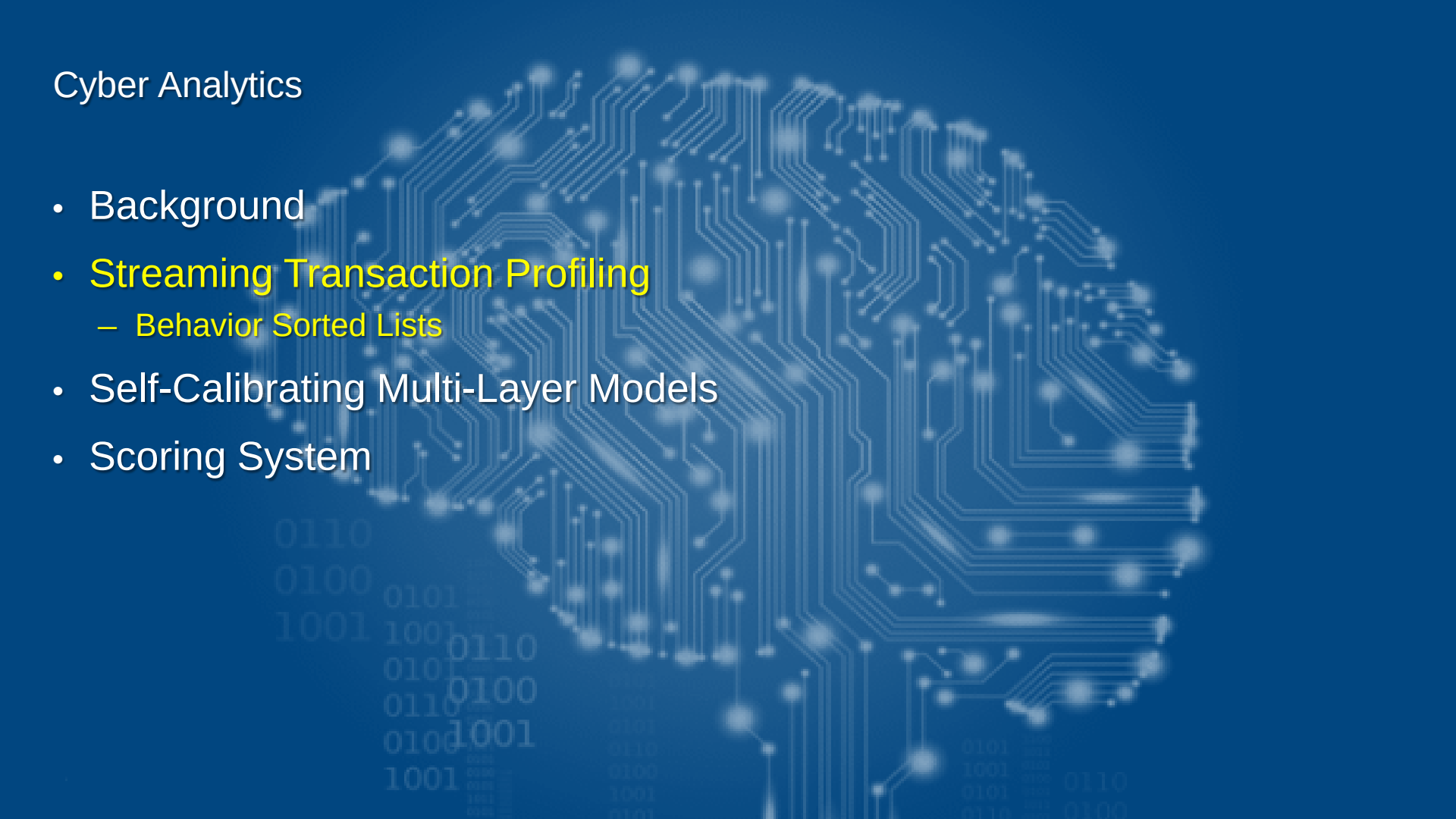
x(t) = $\delta(\ (\tau_i - \ \tau_{i-1}) < \delta\tau_i)$ "is the time between successive errors smaller than $\delta\tau$?"
y(t) = Filter [y(t-1), x(t)]

# Multi Perspective Profiling – Examples
## What is the profile “key”?

- **Requestor IP Profile**
  - I am Joe Q Analysts’s PC: 123.45.67.89
  - Which internal machines do I talk to?
  - Which external networks do I talk to?
  - What network errors do I see?
- **Domain Name Profile**
  - all lookups of abc.def.net
  - Who asked for me?
  - When was I first queried?
  - What time of day?
- **Requestor IP x Domain Name Profile**
  - I am 123.45.67.89 and asked for `www.google.com` in business hours
  - I am 123.45.67.89 and asked for `xxxx123.abc.north_elbonia.net` every night at midnight

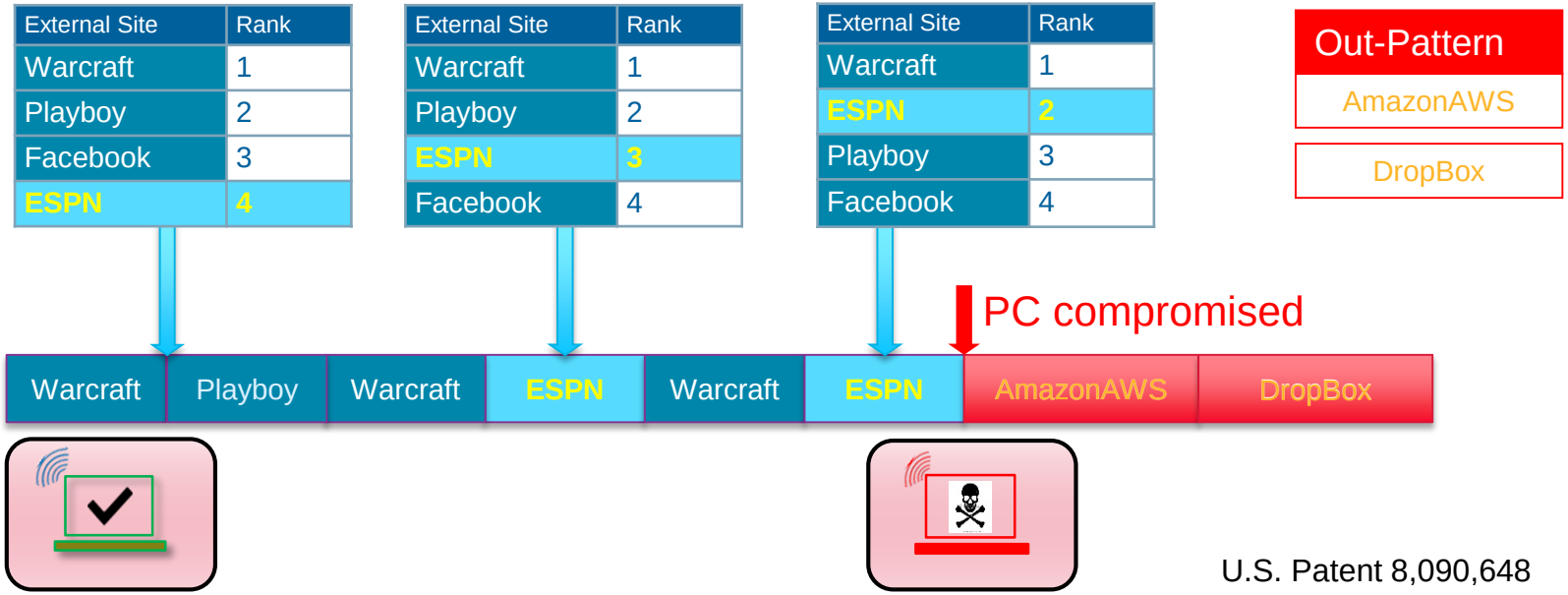
# Cyber Analytics

- Background
- Streaming Transaction Profiling
  - Behavior Sorted Lists
- Self-Calibrating Multi-Layer Models
- Scoring System



# Behavior Sorted List
## Establishing normal behavior patterns

- In-Pattern (top ranks in list) indicates normal usage
- Out-Pattern (non-frequent activity) is more likely to be misuse / malicious
- Dynamically updating adaptive table in streaming for efficiency




U.S. Patent 8,090,648

# Behavior Sorted List

## Frequency and recency

- Analytic algorithm smoothly balances
  - Frequency: “How often do I see ❹”?

Item ❹ is a favorite: high rank

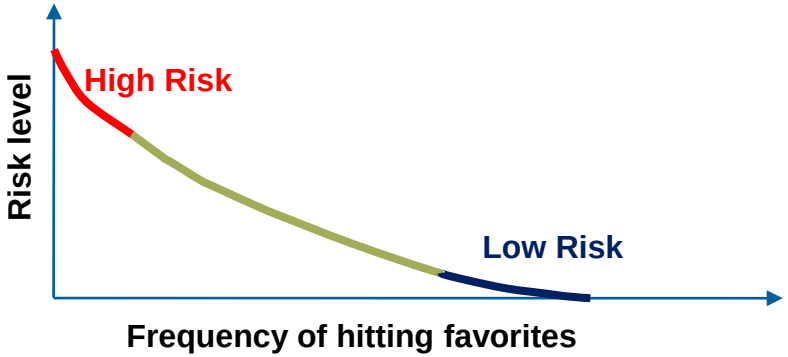
  - Recency: “How recently have I typically seen ❹?”

time →

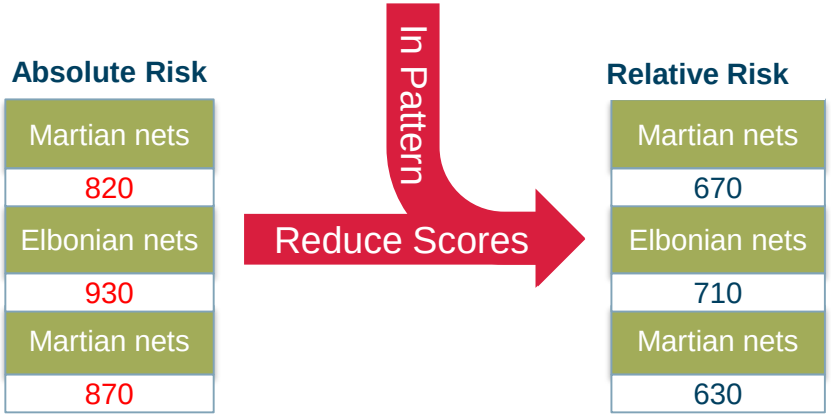
Haven’t seen ❹ as recently, lower rank than previous

# Behavior Sorted List Differentiating Risks

- **In-Pattern** means low risk while **Out-Pattern** means high risk.
- B-LIST helps to reduce scores for normal IP In-Pattern connections even in risky categories.

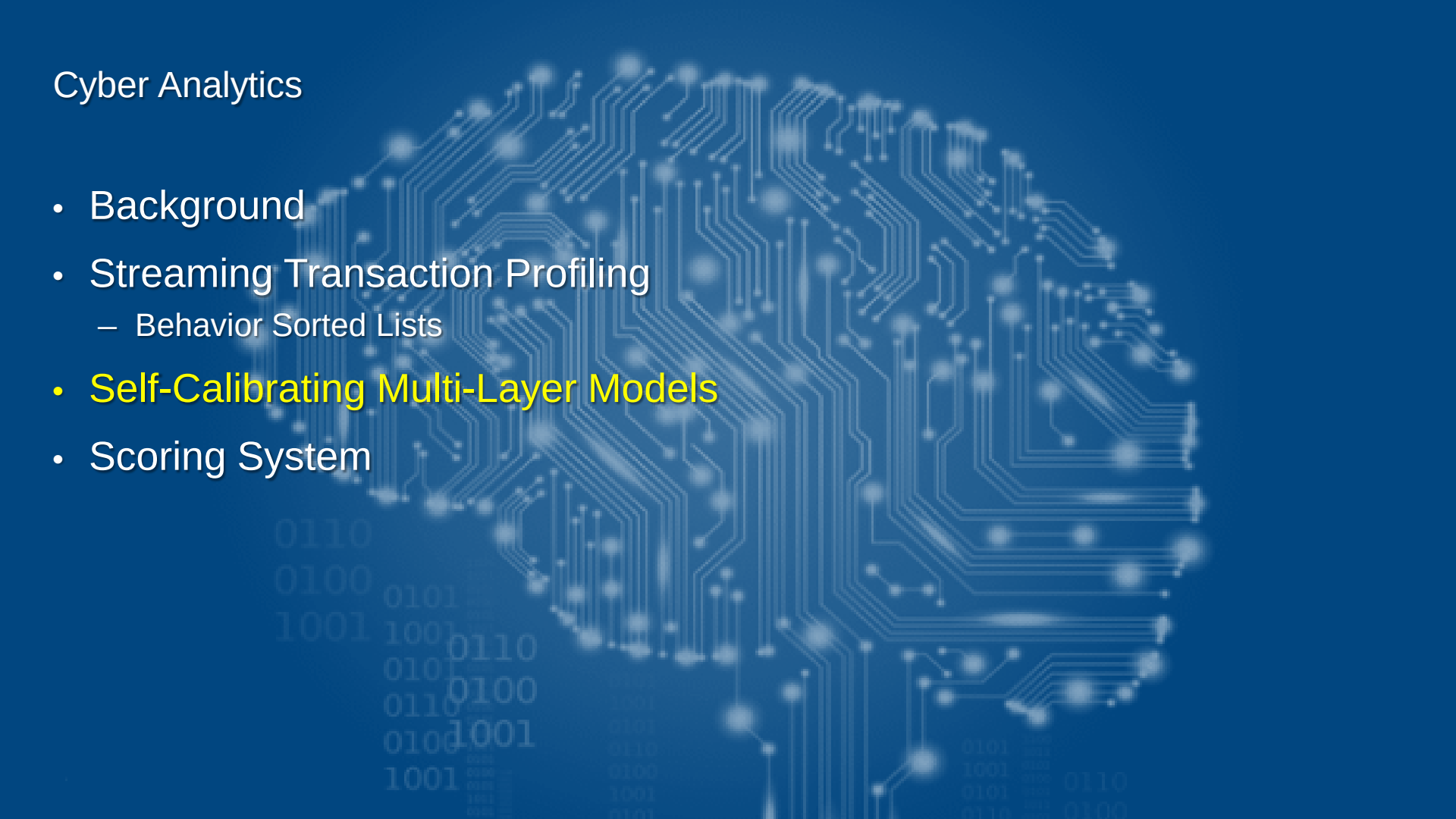



| Frequent networks | Rank |
|---|---|
| Martian nets | 1 |
| Elbonian nets | 2 |



# Cyber Analytics

- Background
- Streaming Transaction Profiling
  - Behavior Sorted Lists
- Self-Calibrating Multi-Layer Models
- Scoring System



# Self Calibrating Models

Data availability issue

- Data export restrictions
- Gathering of data impractical
- Unreliable or no tags/targets

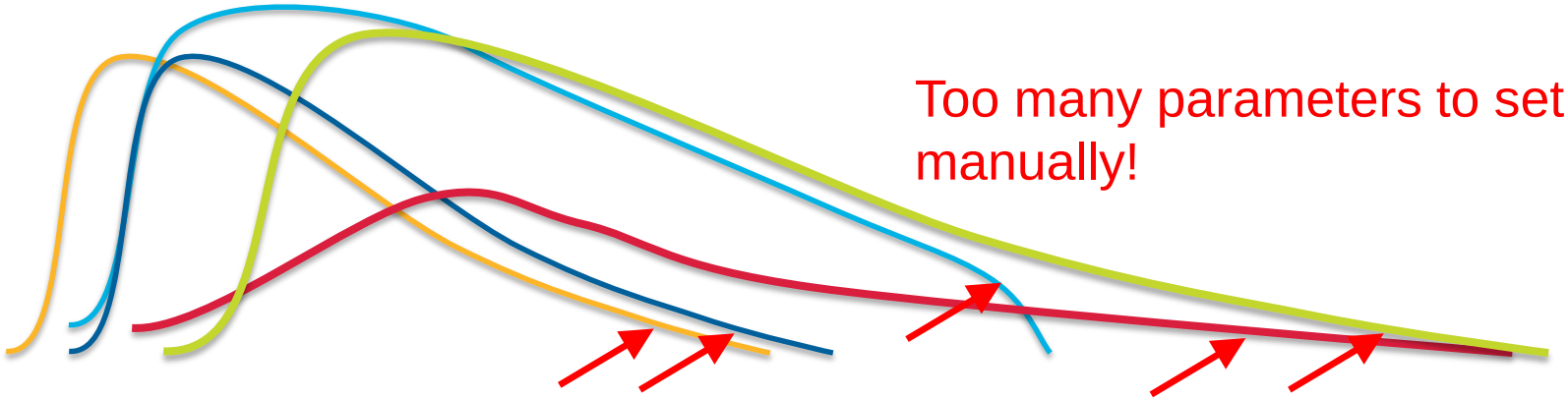
Dynamic Environment

- Real-time adjustment to changes
- Design model features such that higher values indicate riskier behavior
- Entities exhibiting behavior outside norms are usually interesting

**Production Variations may Require Flexible Streaming Analytics**

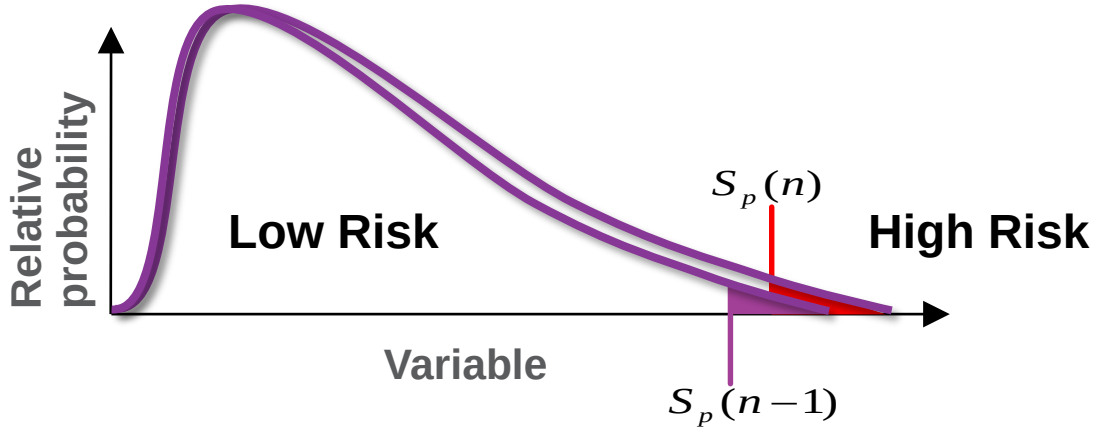
# Self-Calibrating Outlier Models




- FICO technology: streaming self-calibration
  - Update profile estimate of 95% and 99% points **automatically**
  - Very memory & time efficient, no historical dataset storage
  - Adapt to population changes & drifts
  - Use percentiles to form normalized “outlier score” relative for each distribution

U.S. Patent 8,027,439; 8,041,597

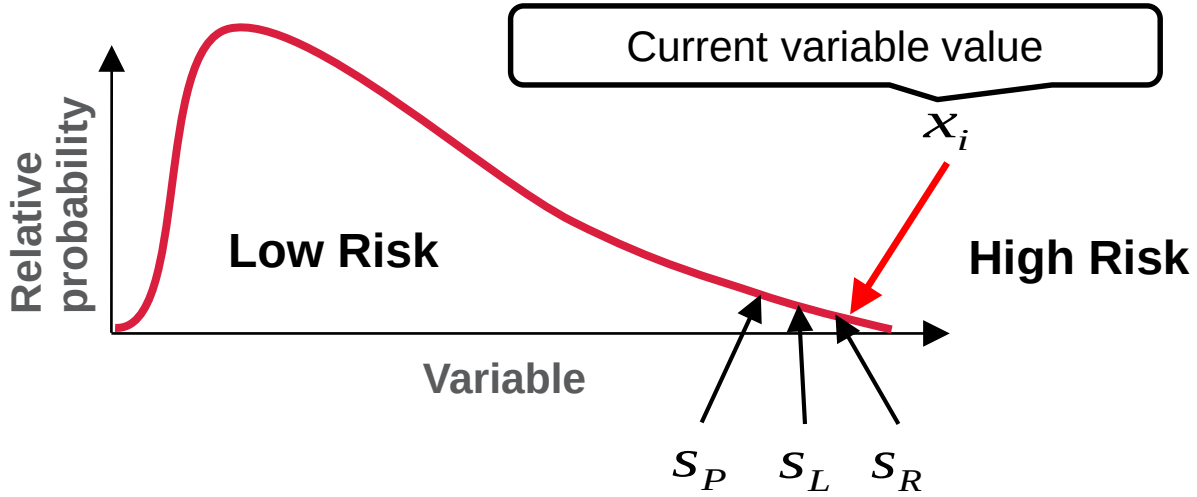
# Self-Calibrating Outlier Analytics
## Quantile estimation on the fly




- Compute quantile “S” on the fly
  - Online iterative estimation observing consecutive observations
  - Each iteration results in a real-time estimate of the quantile
  - Very efficient

U.S. Patent 8,027,439; 8,041,597

# Self-Calibrating Outlier Score




Outlier score for one variable

$$q(x_i \mid s) = \min\left( \max\left( \frac{x_i - s_P}{s_R - s_L}, 0 \right), C \right) \in [0, C]$$

U.S. Patent 8,027,439; 8,041,597

# Self-Calibrating Outlier Analytics
## Combining scores

$$Score = F(k, \sum_{i=1}^{I} w_i q(x_i \mid t, s))$$

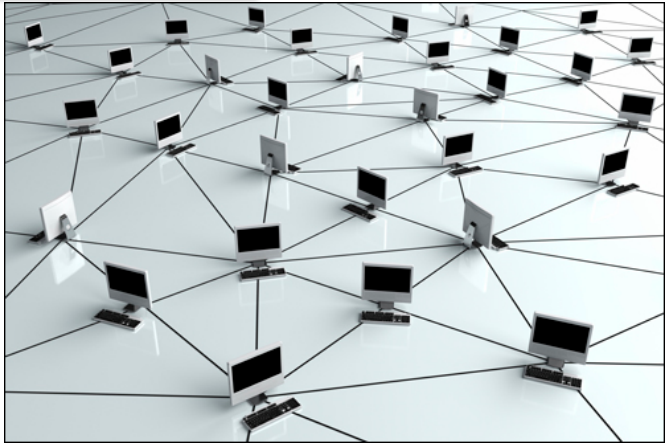
Hidden layer

Quantile estimates of peer group

| | |
|---|---|
| $\{ x_1,\ldots,x_I \}$ | = *features* |
| $w_i$ | = *feature weights* |
| $\{ t_1,\ldots,t_M \}$ | = *peer group selector* |
| $\{ s_1,\ldots,s_J \}$ | = *quantile estimates* |
| $\{ k_1,\ldots,k_N \}$ | = *hidden layer params* |

Weights can be assigned

- Uniformly
- Using Expert Knowledge
- Based on limited data

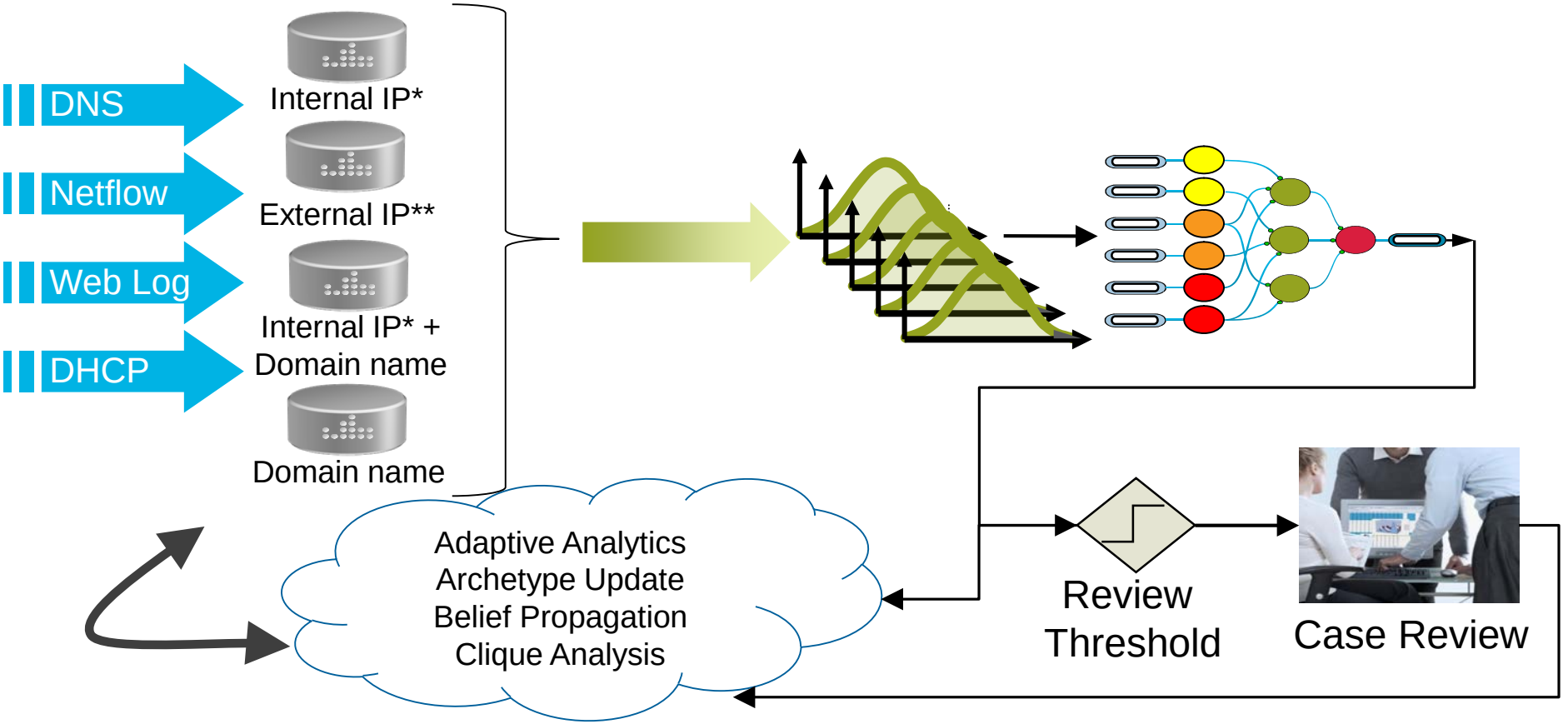


U.S. Patent 8,027,439; 8,041,597

# Cyber Analytics

- Background
- Streaming Transaction Profiling
  - Behavior Sorted Lists
- Self-Calibrating Multi-Layer Models
- Scoring System

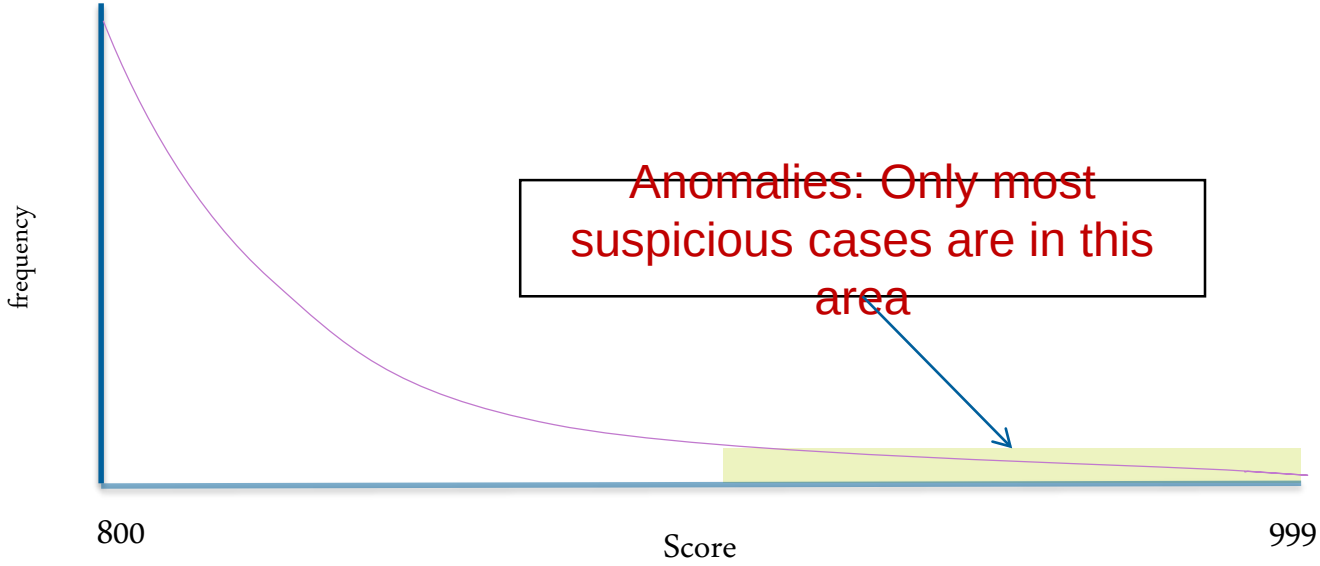
# Cyber Analytics Solution Data flow




U.S. Patent 8,027,439; 8,041,597; 13/367,344; 14/149,598

# Score Distribution

- InfoSec teams can set the number of events to work with high granularity
- Score range is 1-999



# Thank You

Matthew Kennel
+1 858 369 8455
,matthewkennel@fico.com