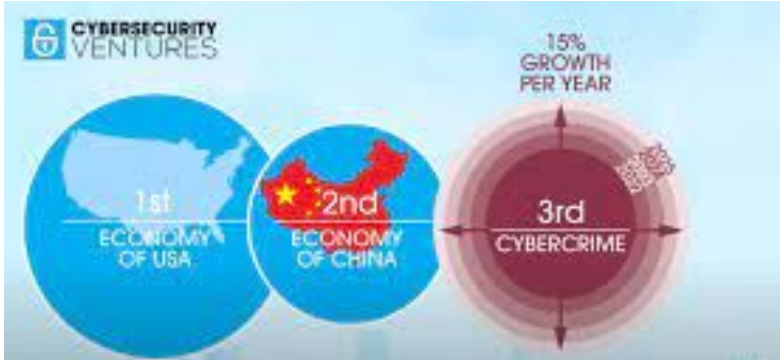
Sapient:
Muhammad Jawaid, Michael Malavé, Brendan Mattina

Staying ahead of cyber attacks.

Cyber Security Today



Cyber crime as the 3rd largest economy after U.S. and China at $6T USD

- Everyday people trust companies to safeguard their data and assets.
- Publicly available cyber defense solutions are **based on data that can be decades old** and cannot keep pace with cyber criminals who are growing more sophisticated.
- **Cyber security is reactive. It must become proactive.**

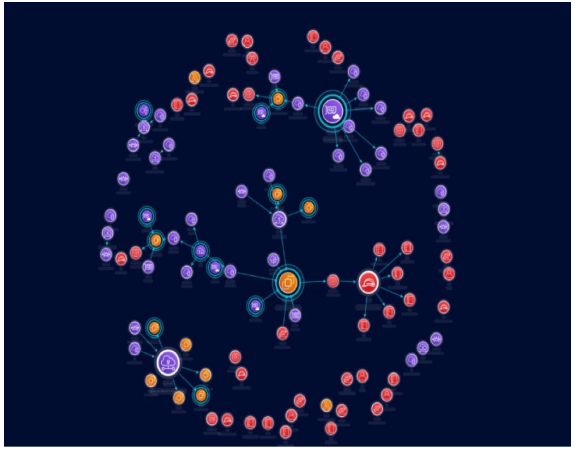
Customer: Cybersecurity Providers

- Companies utilize cyber security providers and security teams.
- Cybersecurity providers enable discovery of attacks in companies systems.
- Additionally, red teams mock threats in systems to allow for new threat paths to be discovered.
- **Red teams* are costly** and not every company can afford them.
- The **costs will continue to rise** hurting the bottom line.
- Cybersecurity providers can **use Sapient to improve detection of cyber attack.**

*Red teams are authorized by a company to emulate cyber attacks in a network.

Value Proposition:

- **Data highly imbalanced** (60K Benign records to 1 Malicious Record)
- We aim to **generate emerging attack paths within systems to enable faster and cheaper threat discovery.**
- This includes:
  - Endpoint based events (eg. mobile devices, servers)
  - **Advanced persistent threats**
    - Credential harvesting
    - Shellcode injection
    - Lateral movement



Porter's 5 Forces: Understand Competitive Landscape

1. **Potential Entrants:** Even though the field requires cutting edge research, large volumes of data, and infrastructure to train models, **there are no moats**. it is important to establish specialization that can demonstrate a ROI.

2. **Suppliers:** Due to the limited options in this space today, **suppliers have leverage** on pricing with respect to buyers looking to protect their data.

3. **Buyers:** Very few companies in this space right now that are specializing in APTs, so buyers have limited leverage to influence pricing. However, over time, the **technology will get commoditized and become cheaper**.

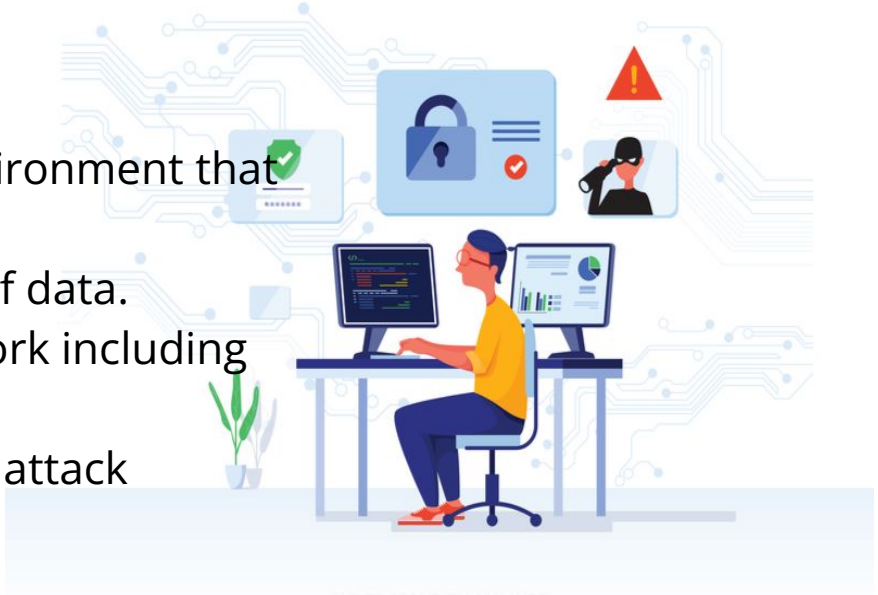
4. **Substitutes:** Identifying advanced persistent threats will become absolutely necessary to mitigate the risk of **financial and reputational risk**.

5. **Complements:** Rather than a winner takes all, the pie will be split with companies **specializing in specific domains** such as identity access management, data leakage, etc.

Target - Cyber Security Analyst

I want to be able to:

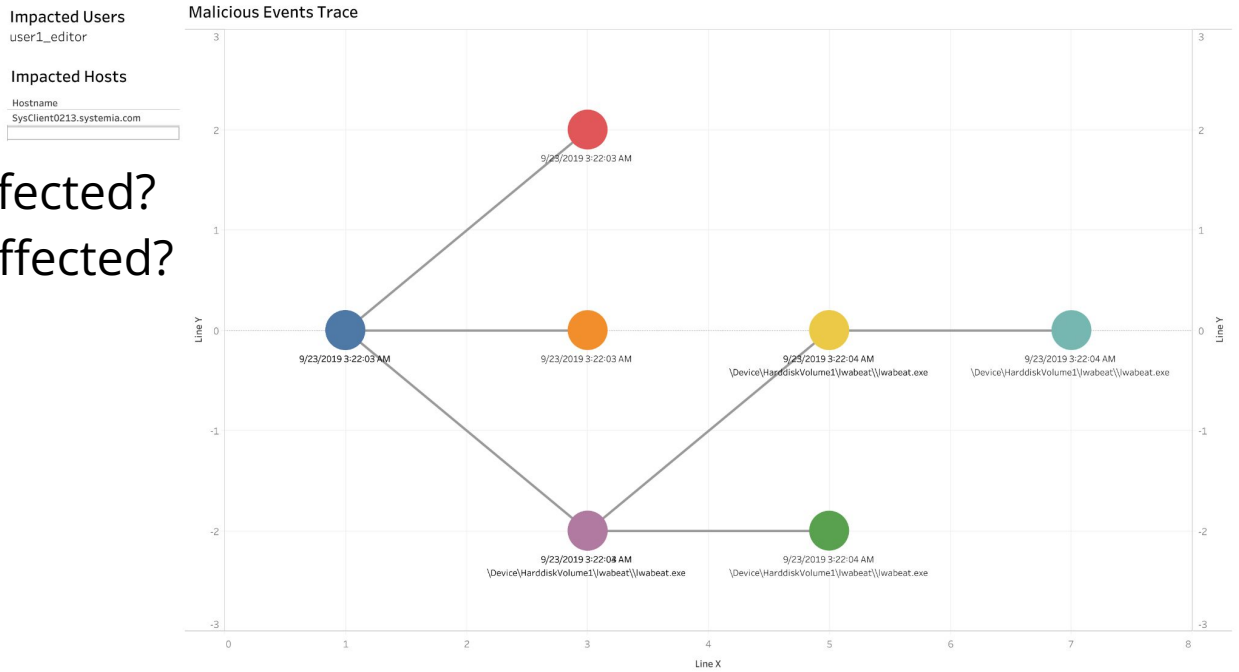
- **Identify true attack patterns** in the environment that I am unaware of.
- **Avoid days** of analyzing billions of rows of data.
- **Visualize the path** of events in the network including important features.
- **Enhance Red Team** operations with new attack patterns to emulate.



Target User Needs

Security Priorities:

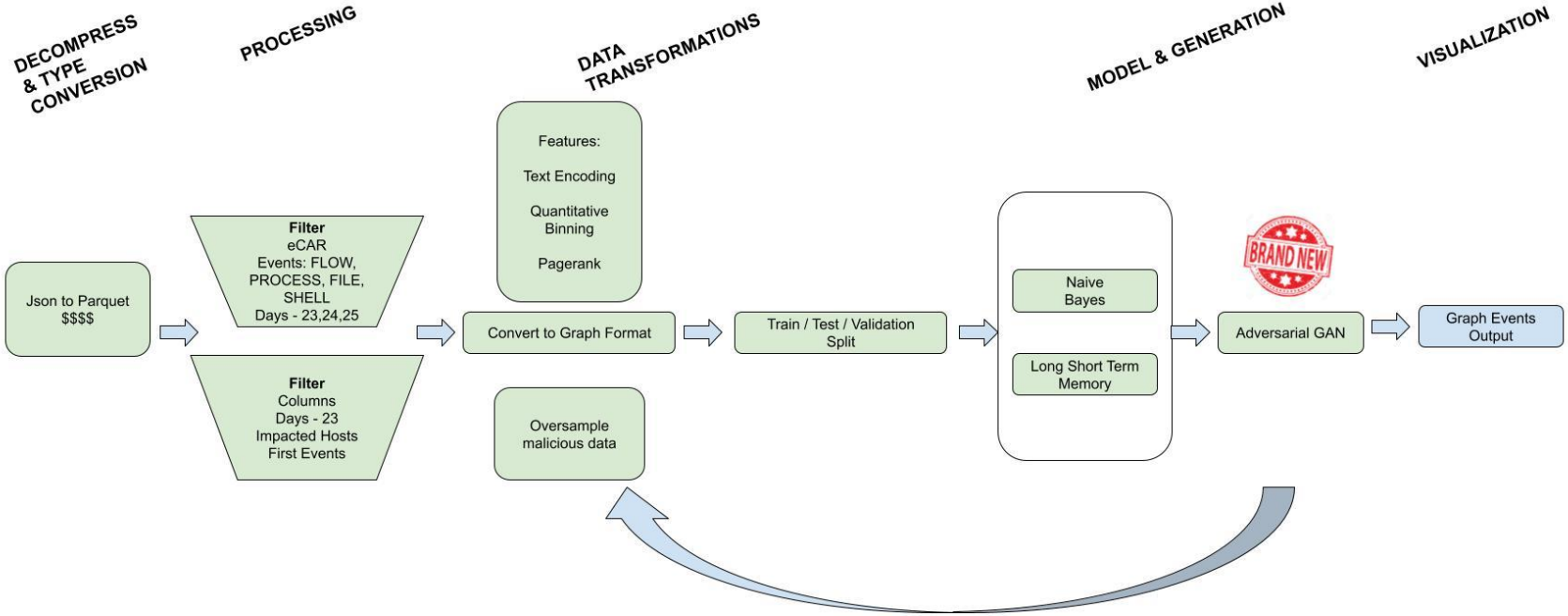
- Which users were affected?
- Which assets were affected?
- How did it spread?



Spotlight - Visualization

Malicious Events Trace

Sapient Architecture

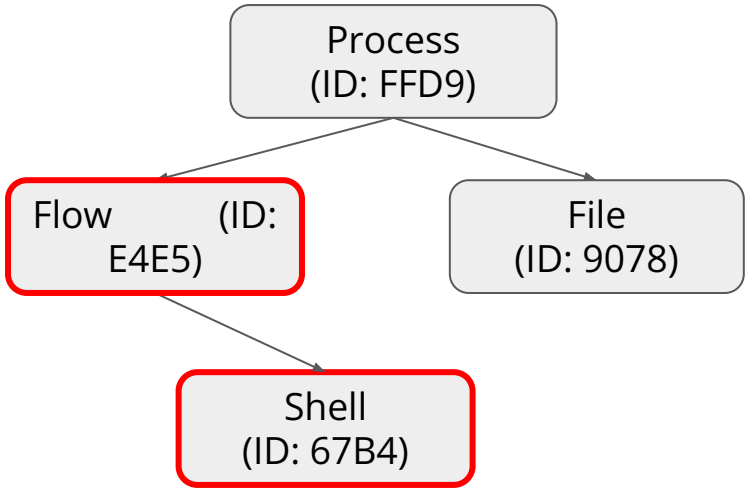


Spotlight - Feature Engineering

Trace*	Object ID	Actor ID	Time Diff (ms)	Object Type	Action	Image Path	Parent Image Path	File Path	Malicious Event
1	FFD9	23EE	None	Process	Create	DllHost.exe	python.exe	.dll	0
1	9078	FFD9	100	File	Read	sample.doc	word.exe	.doc	0
2	E4E5	FFD9	None	Flow	Start	chrome.exe	winlogon.exe	None	0
2	67B4	E4E5	10000	Shell	Open	cmd.exe	cmd.exe	None	1

*An event trace is a chronological collection of related events (parents and children). Think words and sentences.

Creating an endpoint event graph using PySpark and GraphFrames.



Node	Edges	
FFD9	E4E5	87EE
E4E5	67B4	

Spotlight - Event Encoding

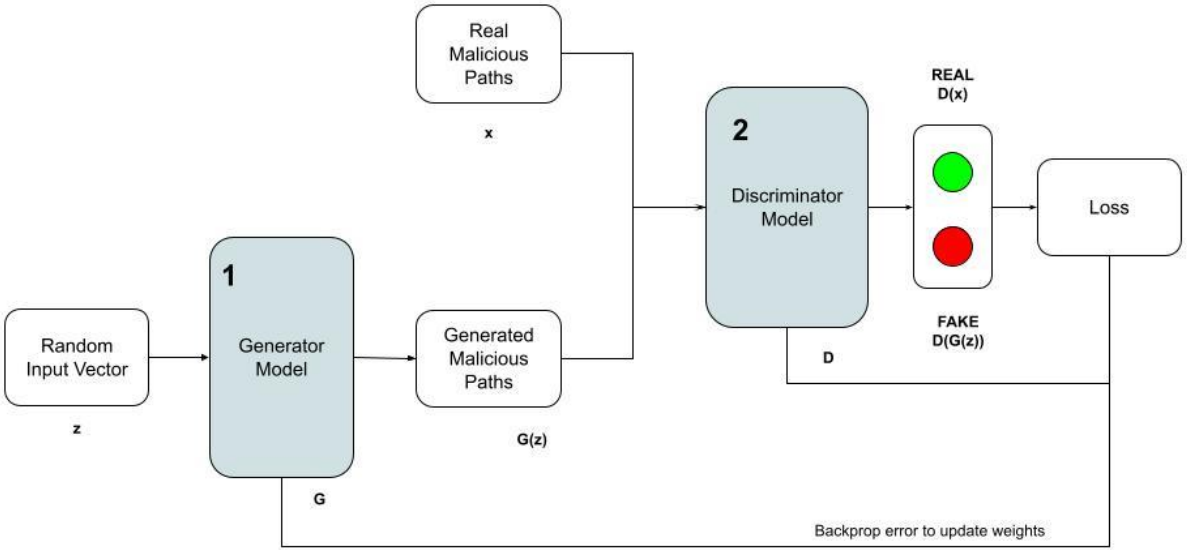
Trace	Event 1	Event 2	Event 3	Event 4	Event 5	Event 6	Malicious Trace
1	[100, 100, 100, 100 …]	[100, 100, 100, 010 …]	[100, 100, 010, 010 …]	[100, 010, 010, 010 …]	[010, 010, 010, 010 …]	[010, 010, 010, 001 …]	1
2	[010, 010, 001, 001 …]	[010, 001, 001, 001 …]	[001, 001, 001, 001 …]	[001, 001, 001, 100 …]	[001, 001, 100, 100 …]	[001, 100, 100, 100 …]	0

Spotlight - Event Encoding (cont.)

Binary String	[100 100 010 001 010 010]					
Field	Time Diff	Object Type	Action	Image Path	Parent Image Path	File Path
Bins/ One Hot Encoding	[0-10m, 10-20ms, 20-30ms]	[File, Process, Shell]	[Create, Read, Delete]	[Chrome, Python, Word]	[Chrome, Python, Word]	[.exe, .doc, None]
Decoded Event	[0-10ms, File, Read, Word, Python, .doc]					

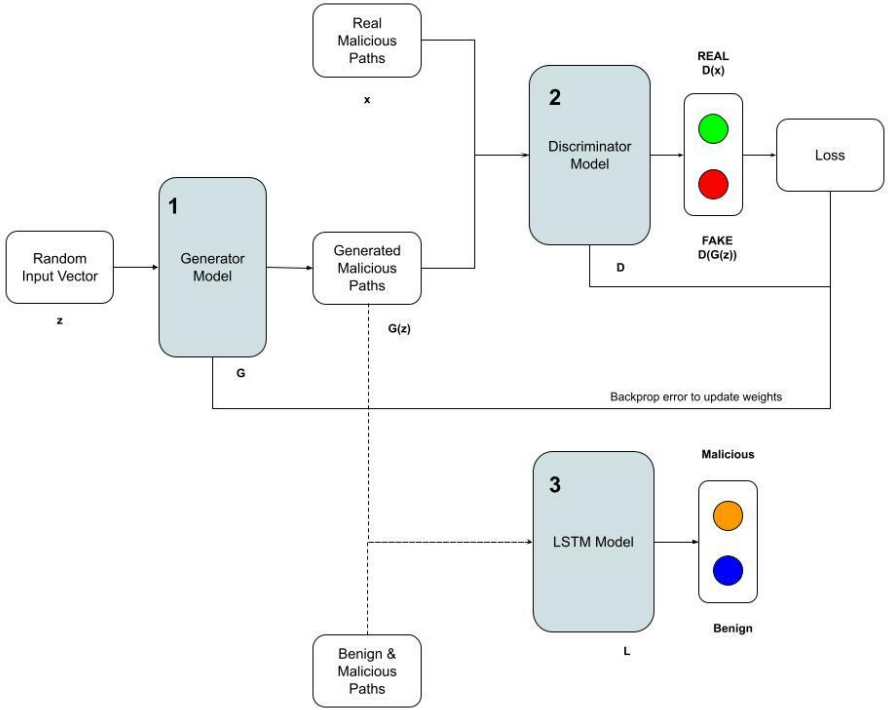
Spotlight - GAN Architecture

We train our generator through a GAN architecture to create samples that look like our real data using randomized input.



Spotlight - GAN Model

We use this trained generator to create over 1000 distinct synthetic events to upsample our malicious data.

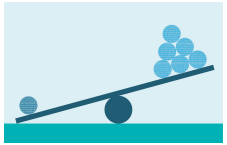


Model Evaluations

Model	KFold	Malicious to Benign Ratio	Train Test Split	Label	Accuracy	Precision	Recall
Naive Bayes	No	1:1	70/30	malicious	0.84	0.84	NA
LSTM	Yes	1:5	80/20	malicious	0.87	0.54	0.41
LSTM + GAN (1*500)	Yes	1:5	80/20	malicious	0.85	0.61	0.24
LSTM + GAN (1.1k)	Yes	1:5	80/20	malicious	0.94	0.01	0.07

Model Challenges

Data volume: 17 Billion records; Used parquet and filtering to improve processing



Data imbalance: Roughly 60k malicious records to 1 benign record

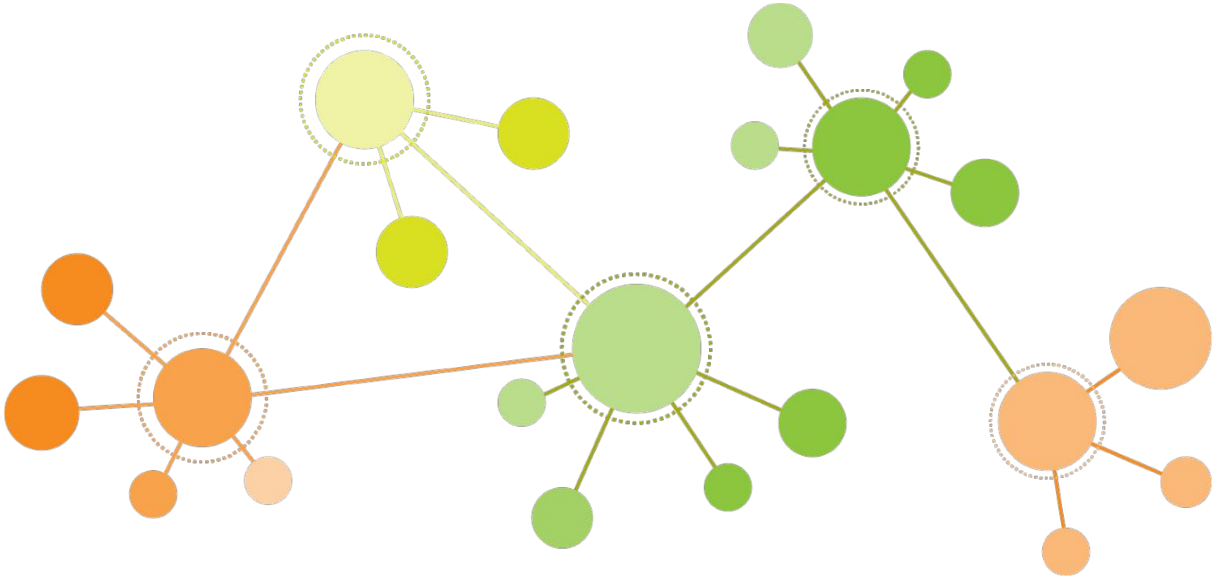
GAN: Leveraged the GAN to upsample our malicious data

Sapient - Roadmap

- Adjust GAN to better evaluate output
- Augment training data with all 3 days
- Create more customized visualizations + metrics of attack patterns

Sapient: Identification of Future Threats

Combining synthetic data, network data, and visualizations to deliver views of potential threats of the future that security teams can take action on today.



Thank you!