Protecting the Protector:

Hardening Machine Learning Defenses Against Adversarial Attacks

Jugal Parikh, Senior Data Scientist
Holly Stewart, Principal Research Manager
Randy Treit, Senior Researcher

In a single day…

2.6 million people

232 country/regions

1.7 million first seen attacks

60% of these attacks were over within the hour

Windows Defender Advanced Threat Protection

Endpoint Protection → **Detection and Remediation**

Windows Defender SmartScreen		Windows Defender Endpoint Protection		Windows Defender Endpoint Detection and Response	
Blocks malicious websites	Blocks low reputation web downloads	Blocks malicious programs and content (scripts, docs, etc.)	Monitors behaviors and terminates bad processes	After execution - Windows Defender ATP monitors for post-breach signals	After execution - Windows Defender Hexadite can reverse damage

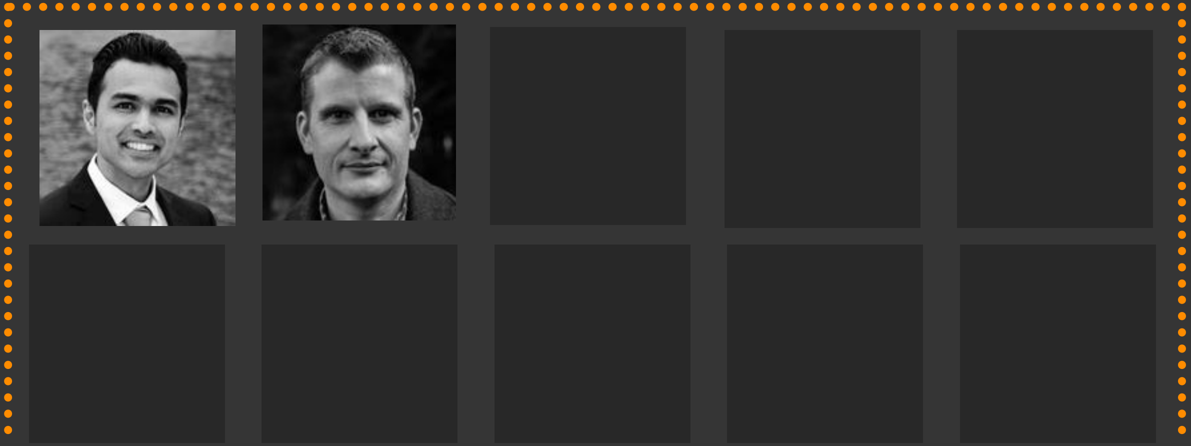

Holly!


Jugal!


Randy!

Windows Defender ATP Research

Threat Predict Research Team

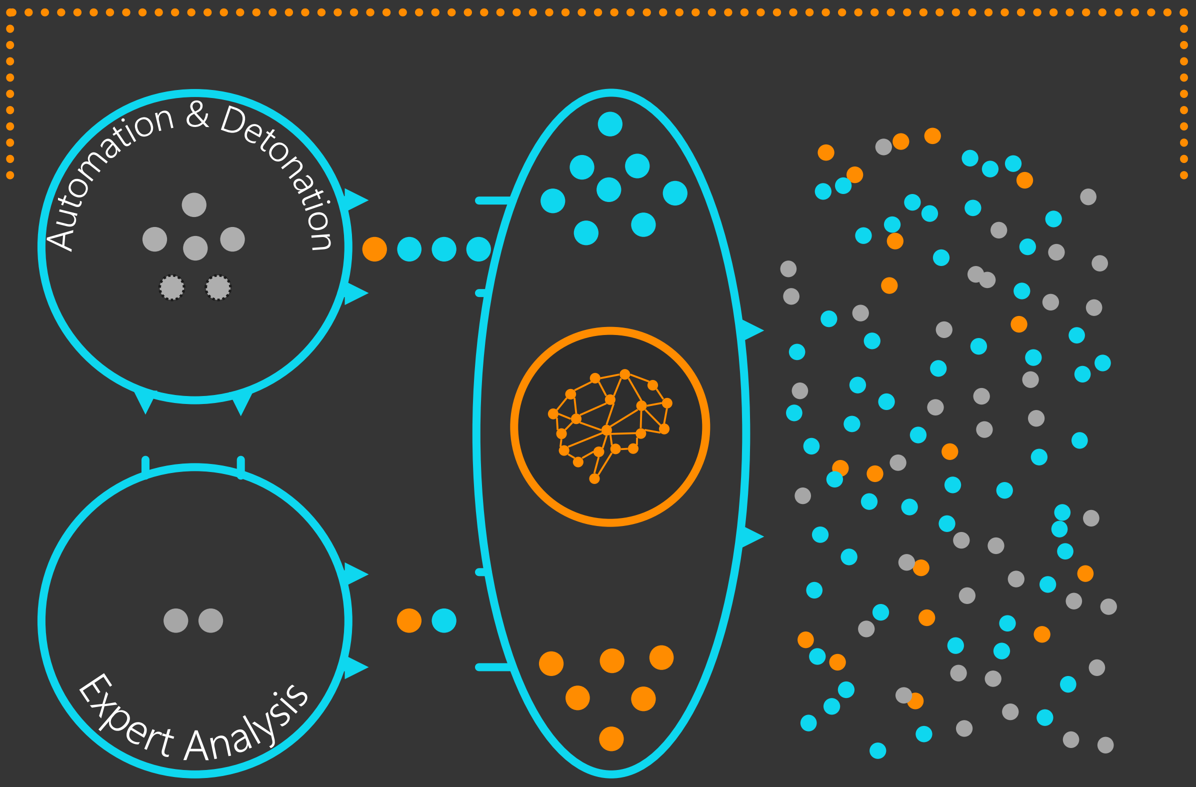


Our focus: Use machine learning to block attacks for Windows Defender ATP

Machine Learning Primer

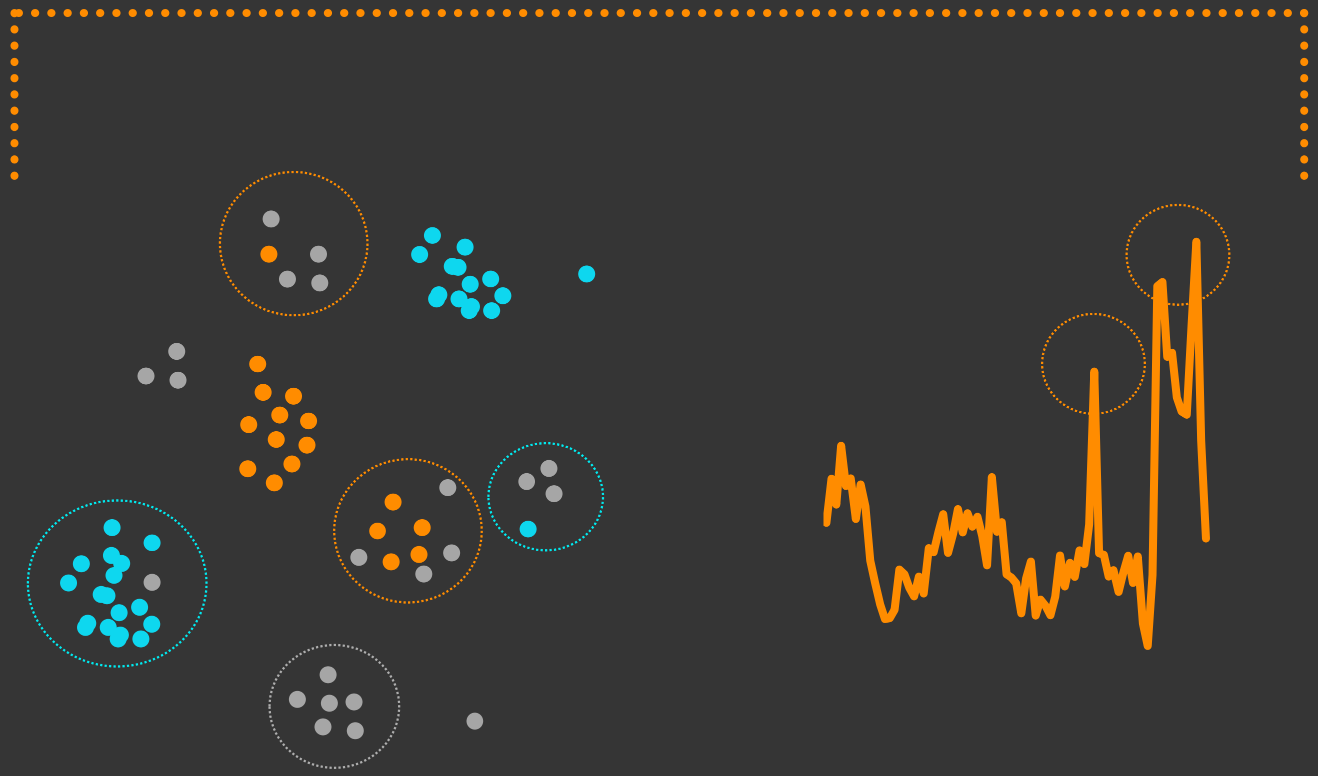
Types of Machine Learning

SUPERVISED



EXPERTS→ LABELS→ ML→ PREDICTIONS

UNSUPERVISED

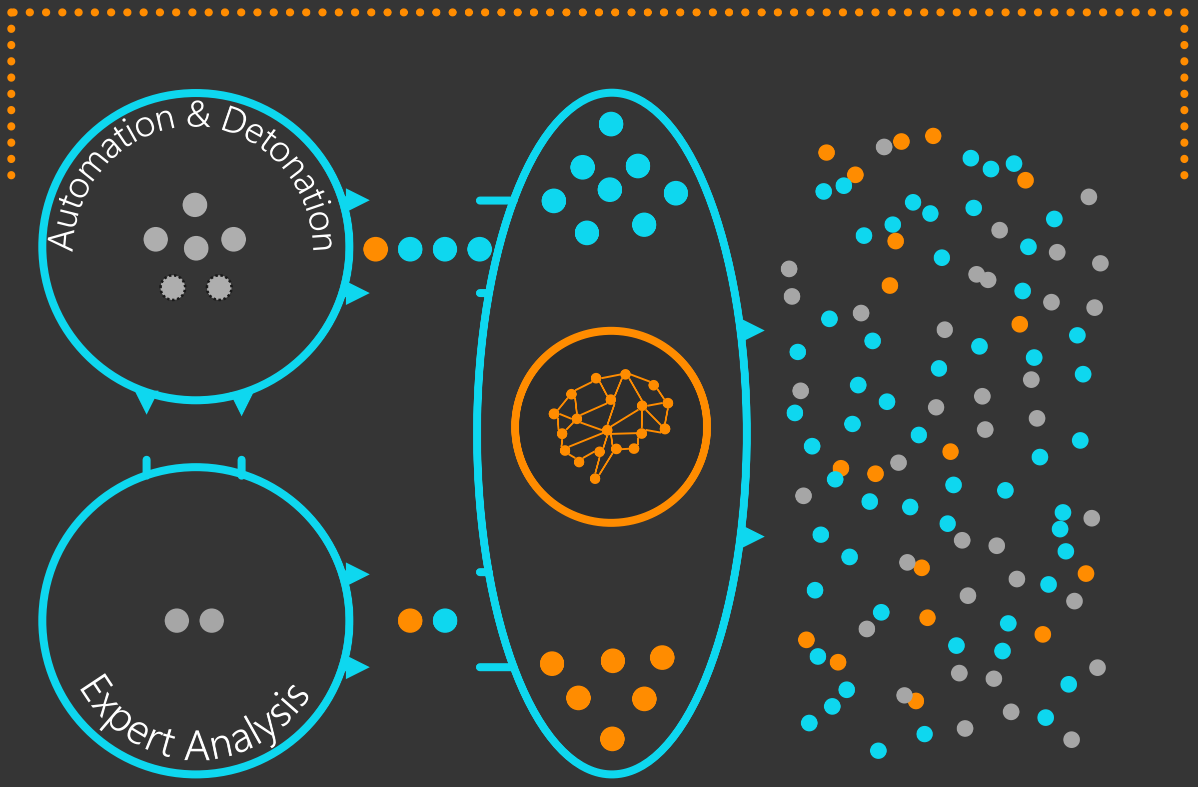


UNKNOWN UNKNOWNS

ANOMALY DETECTION

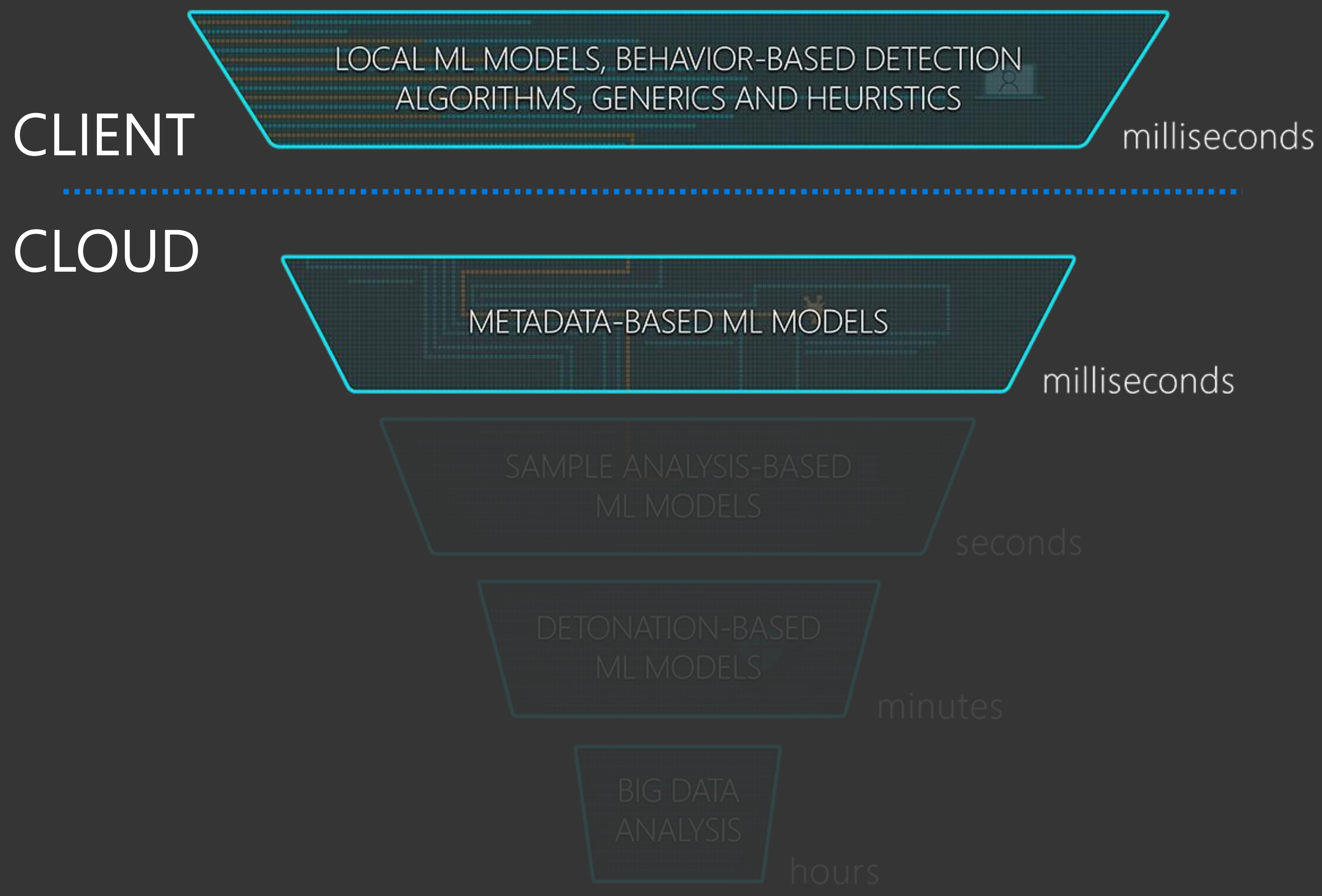
Types of Machine Learning

SUPERVISED

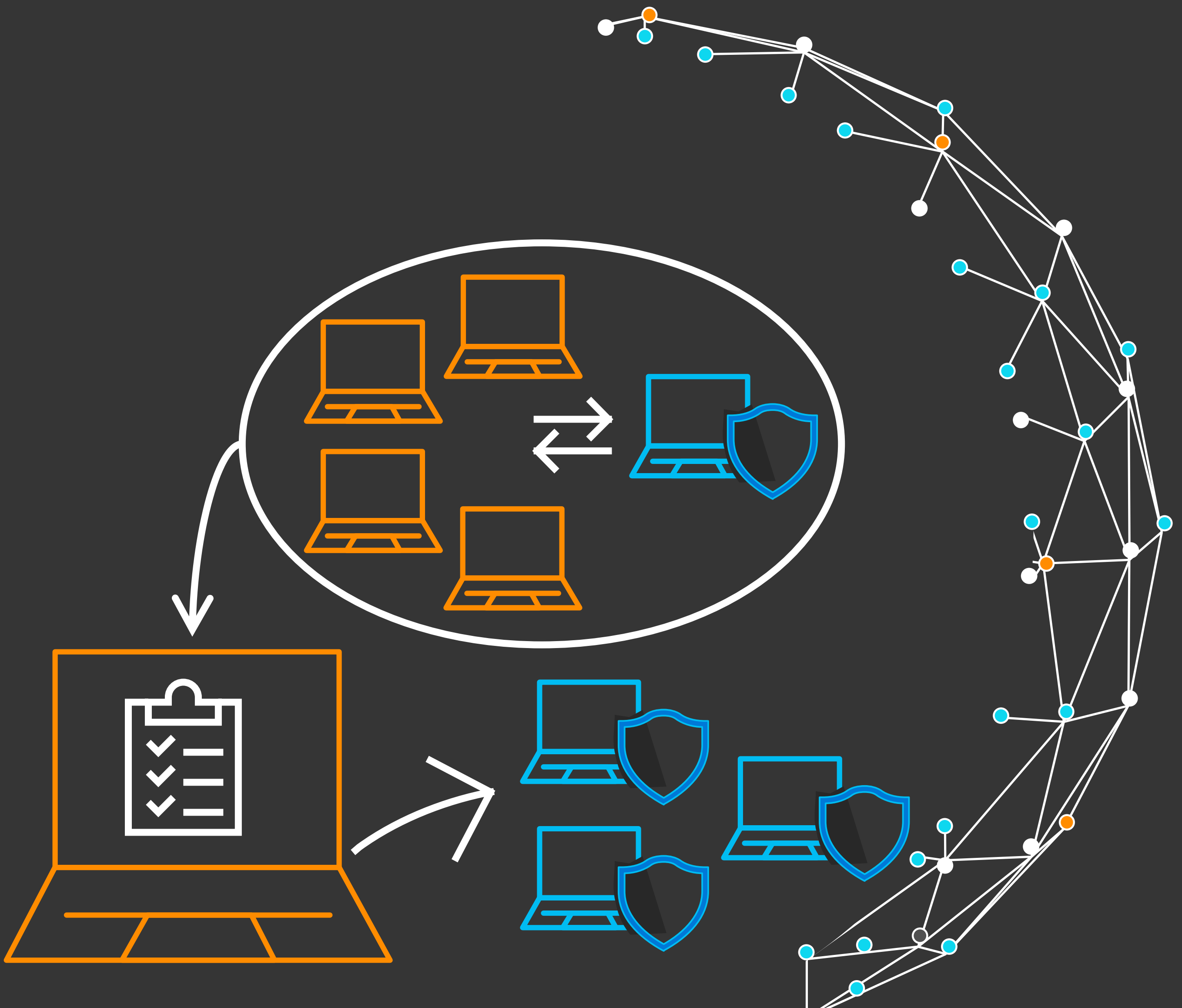


EXPERTS→ LABELS→ ML→ PREDICTIONS

Machine Learning for Endpoint Protection



Client Machine Learning



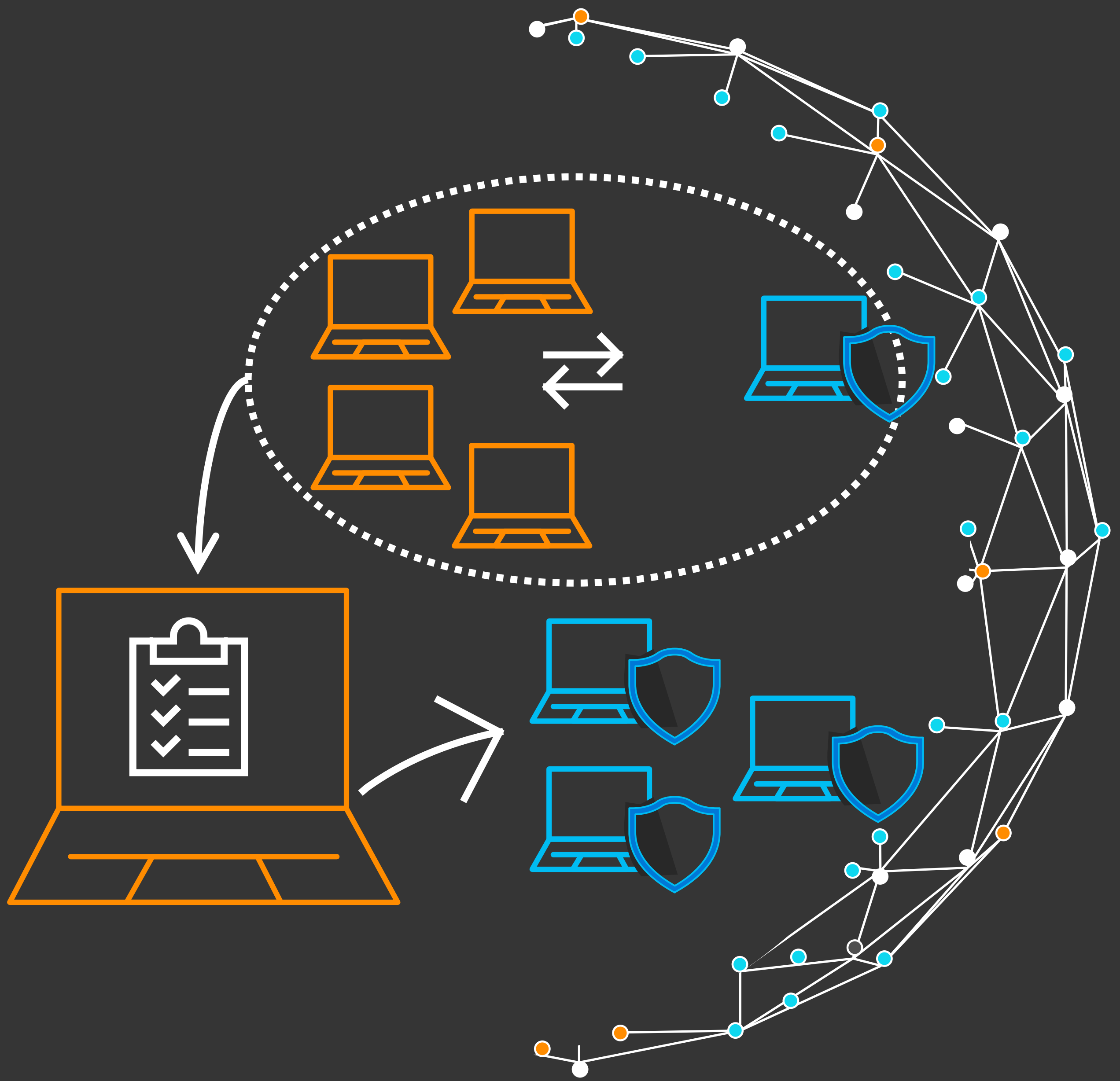
Pro:
Disconnected protection

Con: Silent adversarial brute force attacks

Cloud Machine Learning

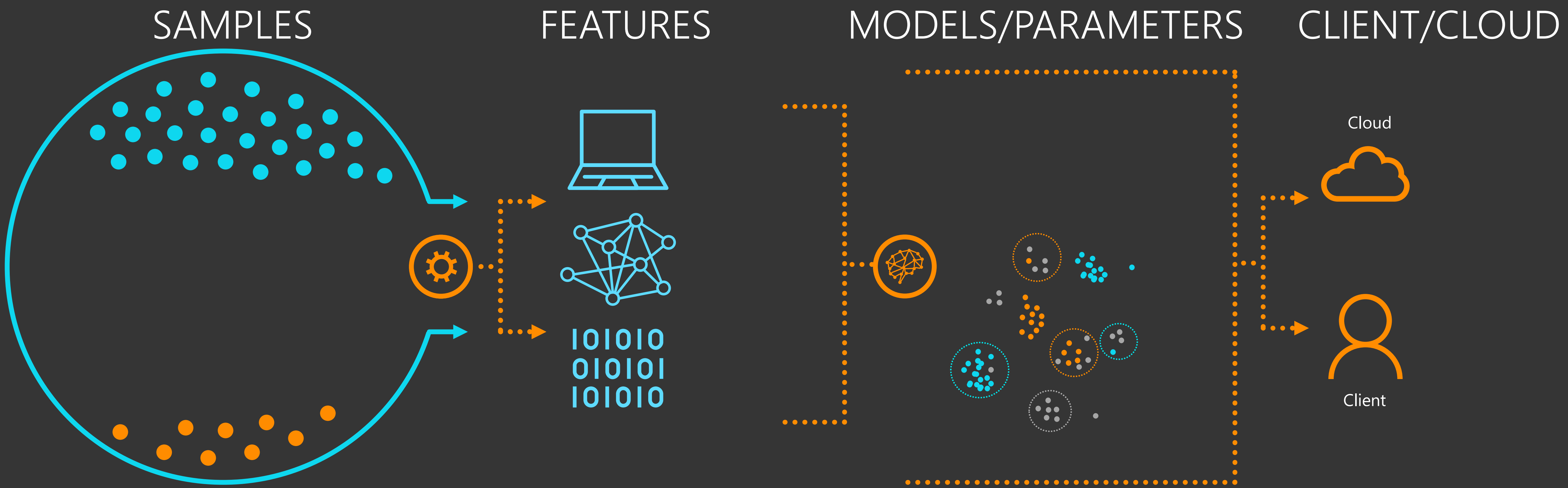
No private brute forcing

Minimal client performance impact



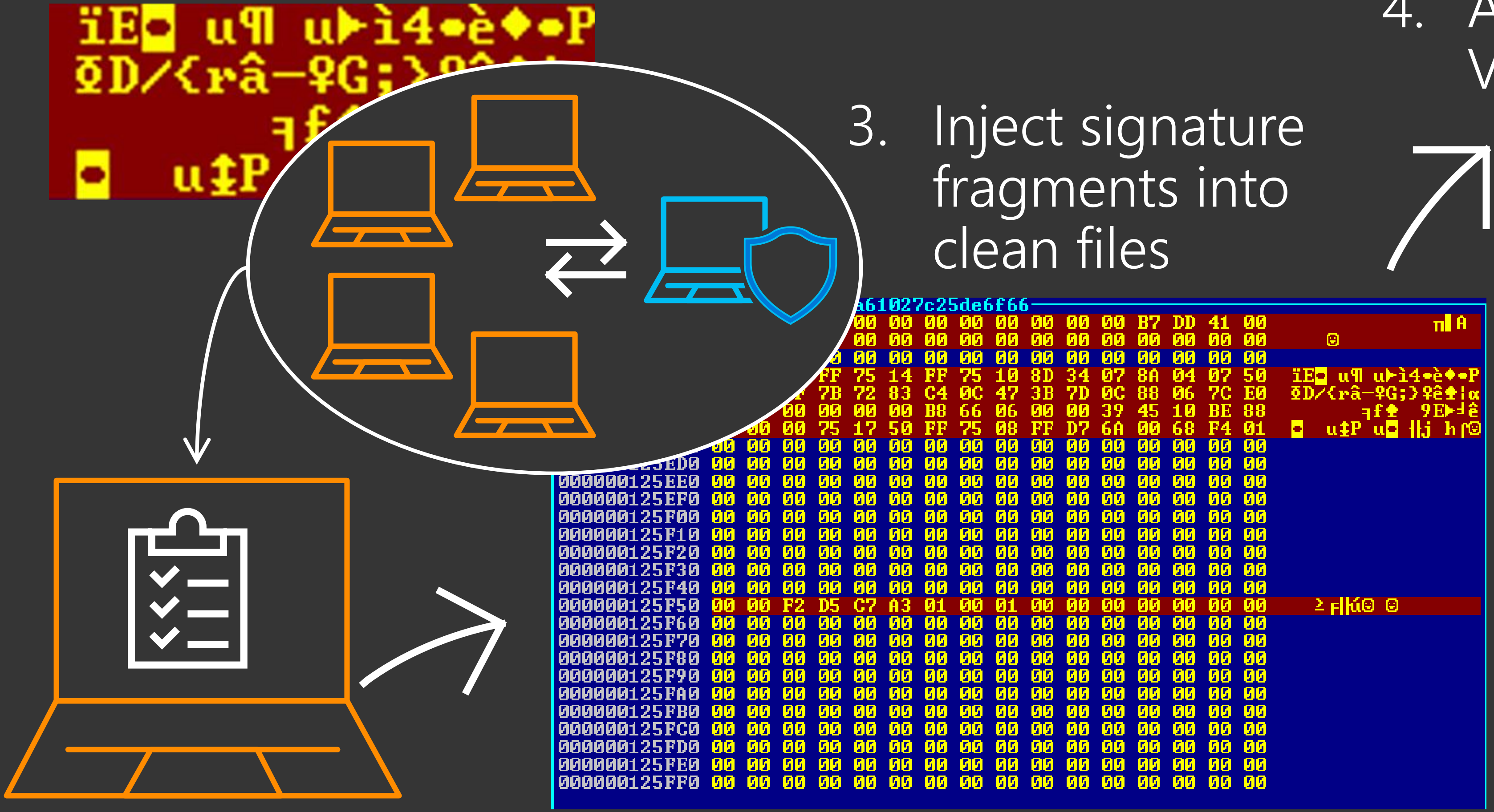
Adversarial ML Examples

Theoretical Attack Vectors: Supervised Model



Specially crafted files→AV industry FPs

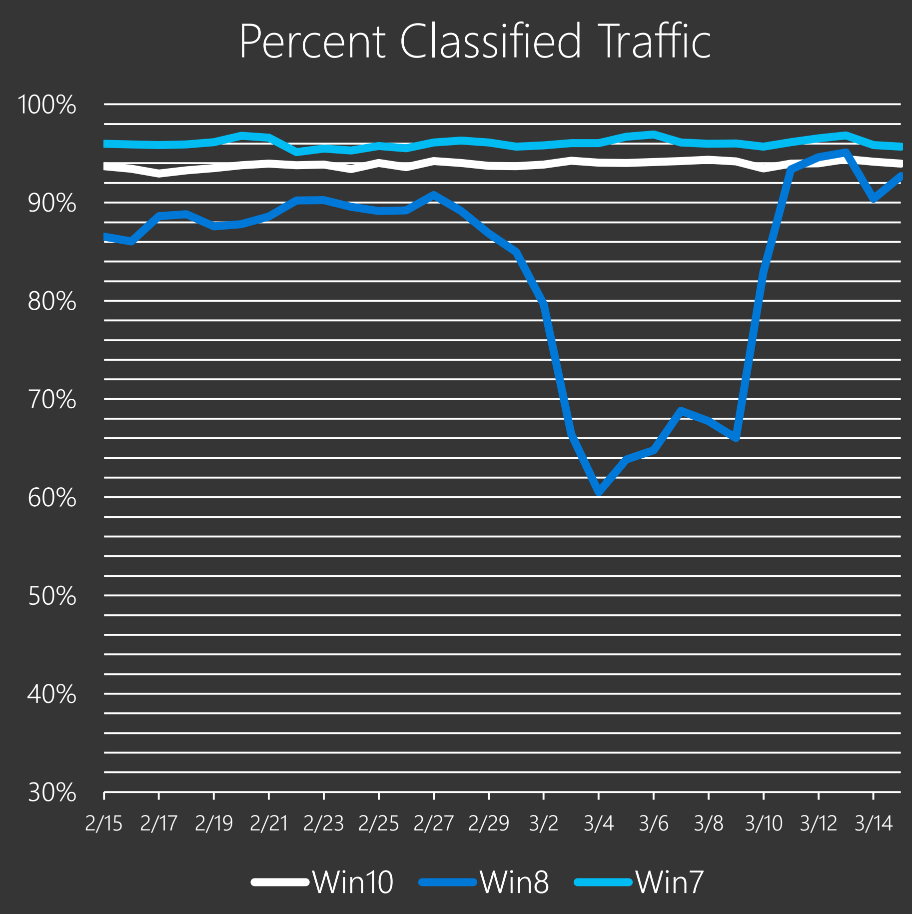
1. Identify signature fragments detected as malicious
2. Identify automated detection techniques
3. Inject signature fragments into clean files
4. Add crafted files to VirusTotal using TOR
5. Automation signs targeted clean files, multiple vendors have FPs



For more details, see [Immunity from antimalware automation attacks] presented at Virus Bulletin 2013

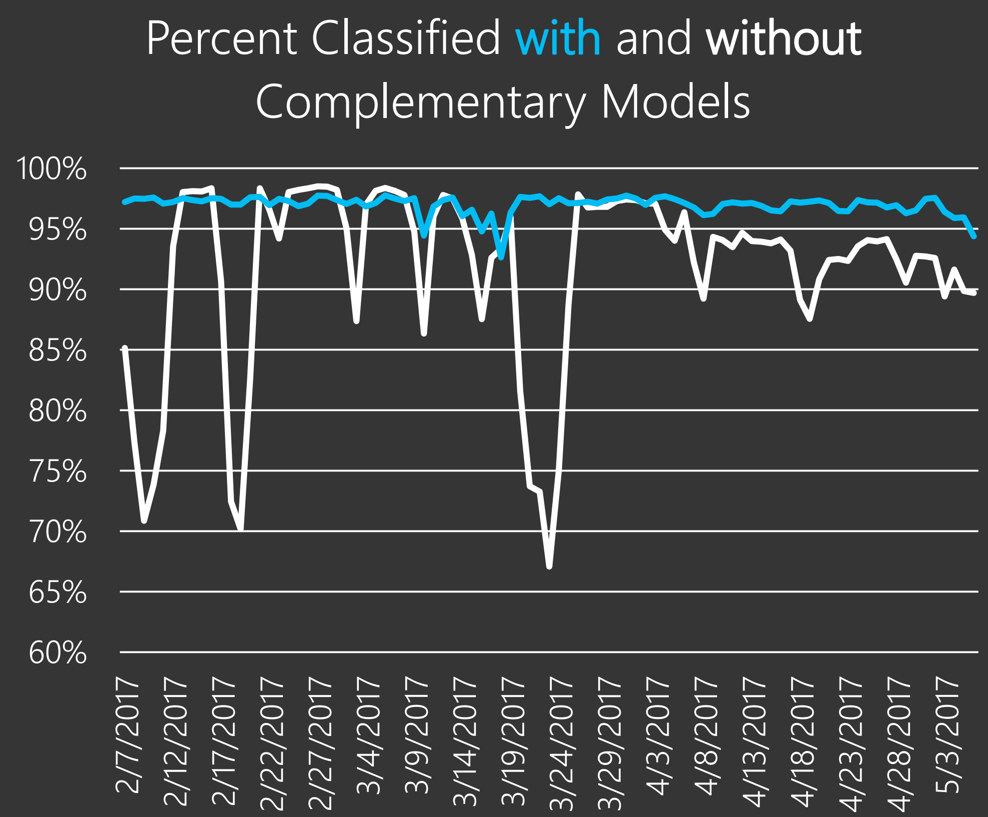
Attacks on Certificate Reputation (Early 2017)

- Synthetic traffic designed to quickly gain reputation on a digital certificate
- Targeted Windows 8
- Originally surfaced as a high percentage of traffic that wasn't classified
- **Low-volume** and **unsigned** file attacks were also identified during investigation



Attacks on Certificate Reputation (cont.)

- Attackers guessed major features (time, traffic, digital certificate)
- Team developed complementary models with additional features that filtered fake traffic out of telemetry
- Combination of models removed attack traffic from training data



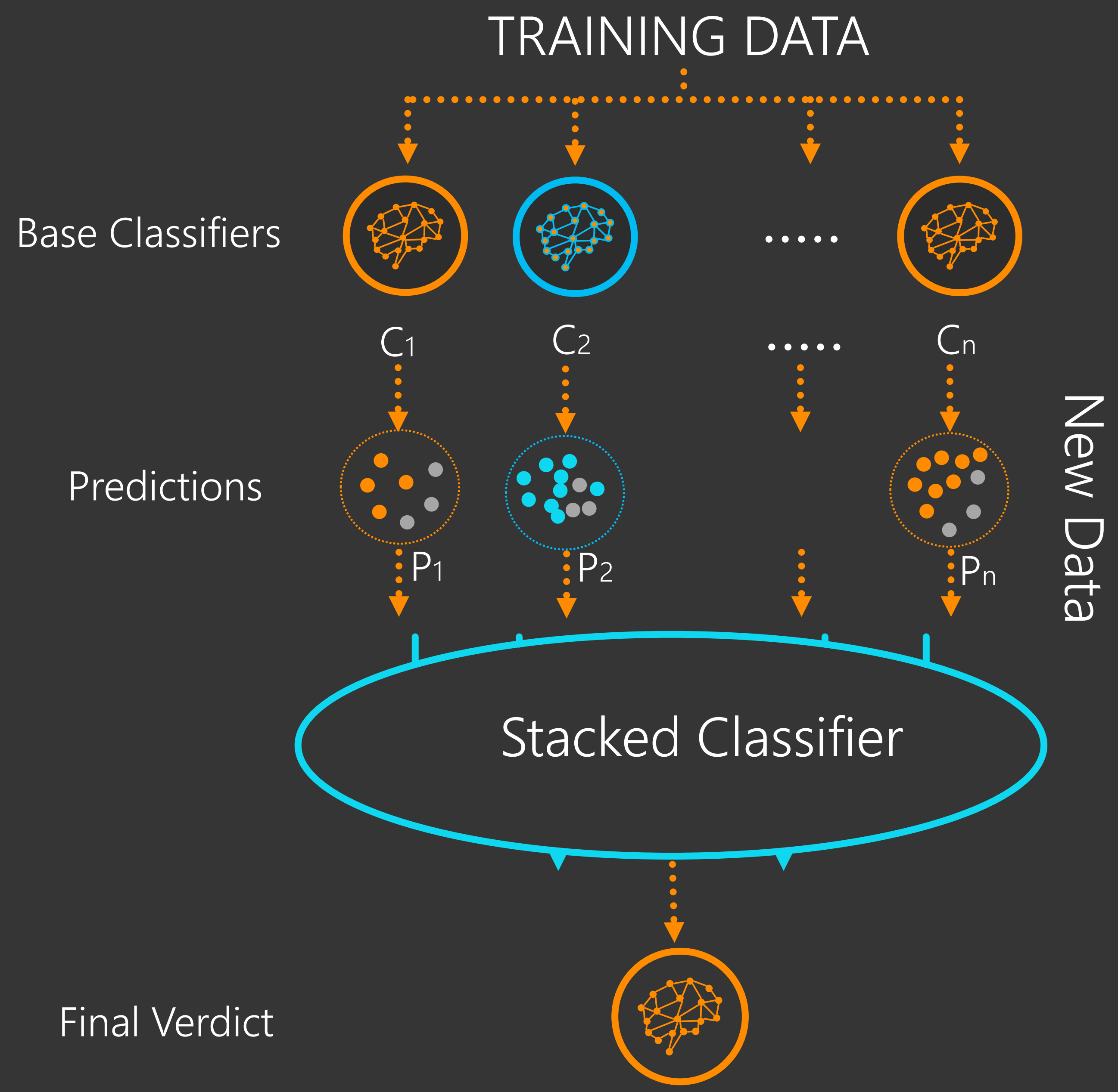
Previous research pointed to ensembles

- Research on adversarial attacks against deep learning classifiers
- Showed that an ensemble of classifiers helped defend against the attacks tested in the paper
- See more at:

  Attack and Defense of Dynamic Analysis-Based, Adversarial Neural Malware Classification Models
  Jack W. Stokes, De Wang, Mady Marinescu, Marc Marino, Brian Bussone
  https://arxiv.org/abs/1712.05919

Ensemble Machine Learning Primer

Overview



Wolpert, David H. "Stacked generalization." Neural networks 5.2 (1992): 241-259.

Challenges

Dealing with:

- Active adversarial
- Volatility/ Covariate Shift
- Noisy environment

Scale:

- Petabytes of threat Intelligence daily

Evaluate:

- ~2.3 Billion global queries everyday

Diversity

Diverse Models

1. Different feature sets
2. Different training algorithms
3. Different training data sets
4. Different optimization settings

Features - Highly dimensional data

Machine attributes

OS version
Processor
Security settings

Researcher Expertise

Static file attributes

10k+ researcher attributes

100k+ static attributes

10k+ behavioral attributes

Behavioral and contextual attributes

Process and installation
ProcessName
ParentProcess
TriggeringSignal
TriggeringFile
Download IP and URL
Parent/child relationships

Behavioral
Connection IP
System changes
API calls
Process injection

Locale
Locale setting
Geographical location

Partial and Fuzzy hashes
ClusterHash
ImportHash
Fuzzy hashes

Full File Content
Header
Footer
Raw file content

File properties
File Geometry
FileSize
FileMetaData
....

Signer info
Issuer
Publisher
Signer

Diverse Set of Classifiers

Client Models

Cloud Models

Full File Content Models

Feature Set	Learner	# of Features
PE Properties	Fast Tree Ensemble	10K+ features
Researcher Expertise	Boosted Tree Ensemble	190K+ features
Behavioral	Boosted Tree Ensemble	6M+ features
Fuzzy Hash 1	Random Forest	512+ features
Fuzzy Hash 2	SDCA	10M+ features
Static, Dynamic and Contextual	Averaged Perceptron	16M+ features
Researcher Expertise, Fuzzy Hash	Averaged Perceptron	12M+ features
File Emulation	DNN	150K+ features
File Detonation	DNN	10M+ features

PE

JavaScript

VBS

PDF

Macro

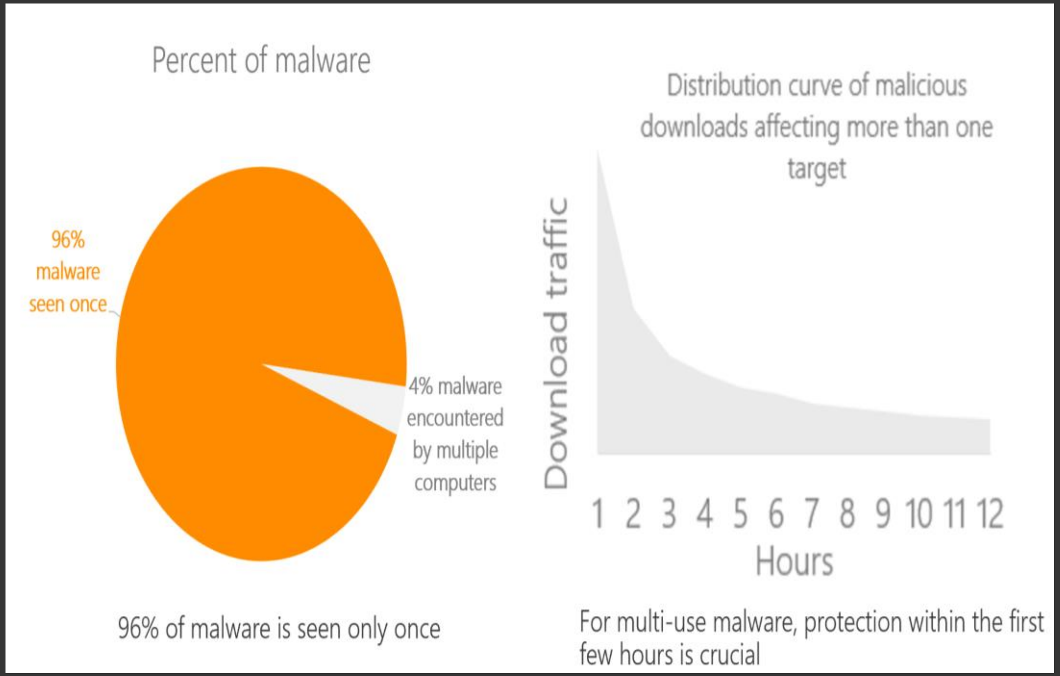
Feature Set   Training Algorithms   Training Data Sets   Optimization Settings

Optimizing for Different Threat Scenarios

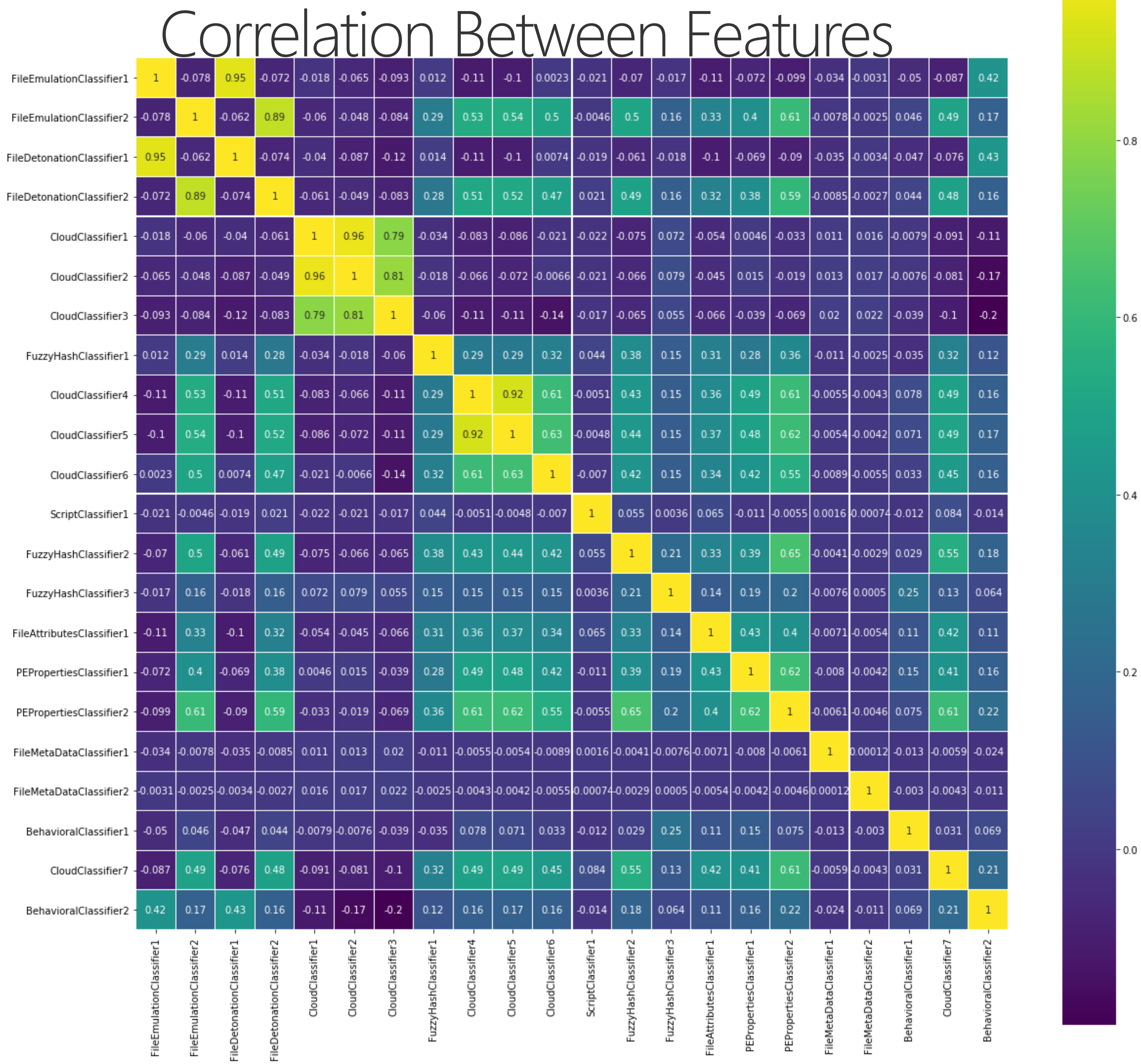
Training Cadence:

Classifiers:

- Malware
- Clean
- PUA
- Enterprise specific
- File Type specific



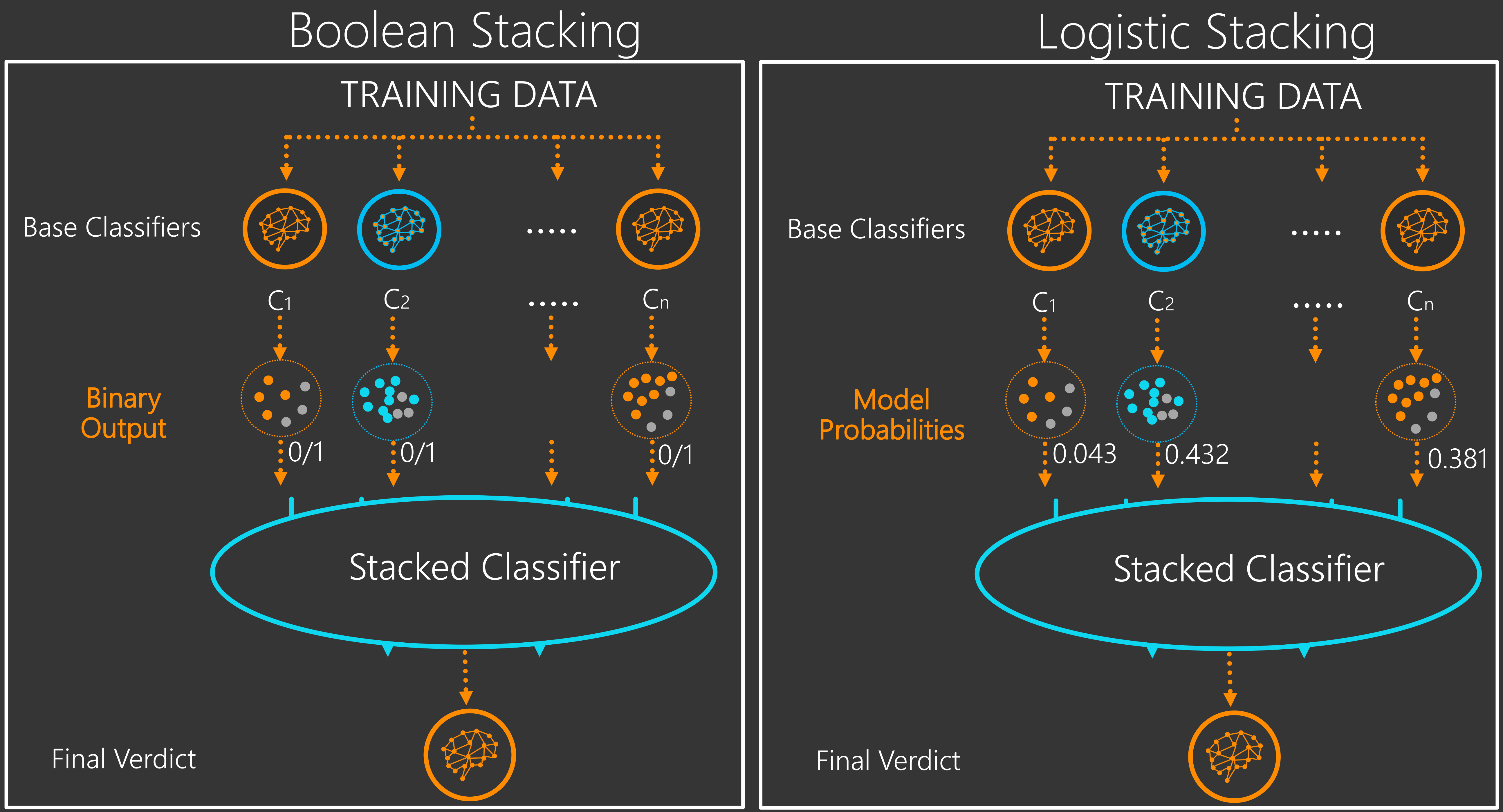
Feature Diversity



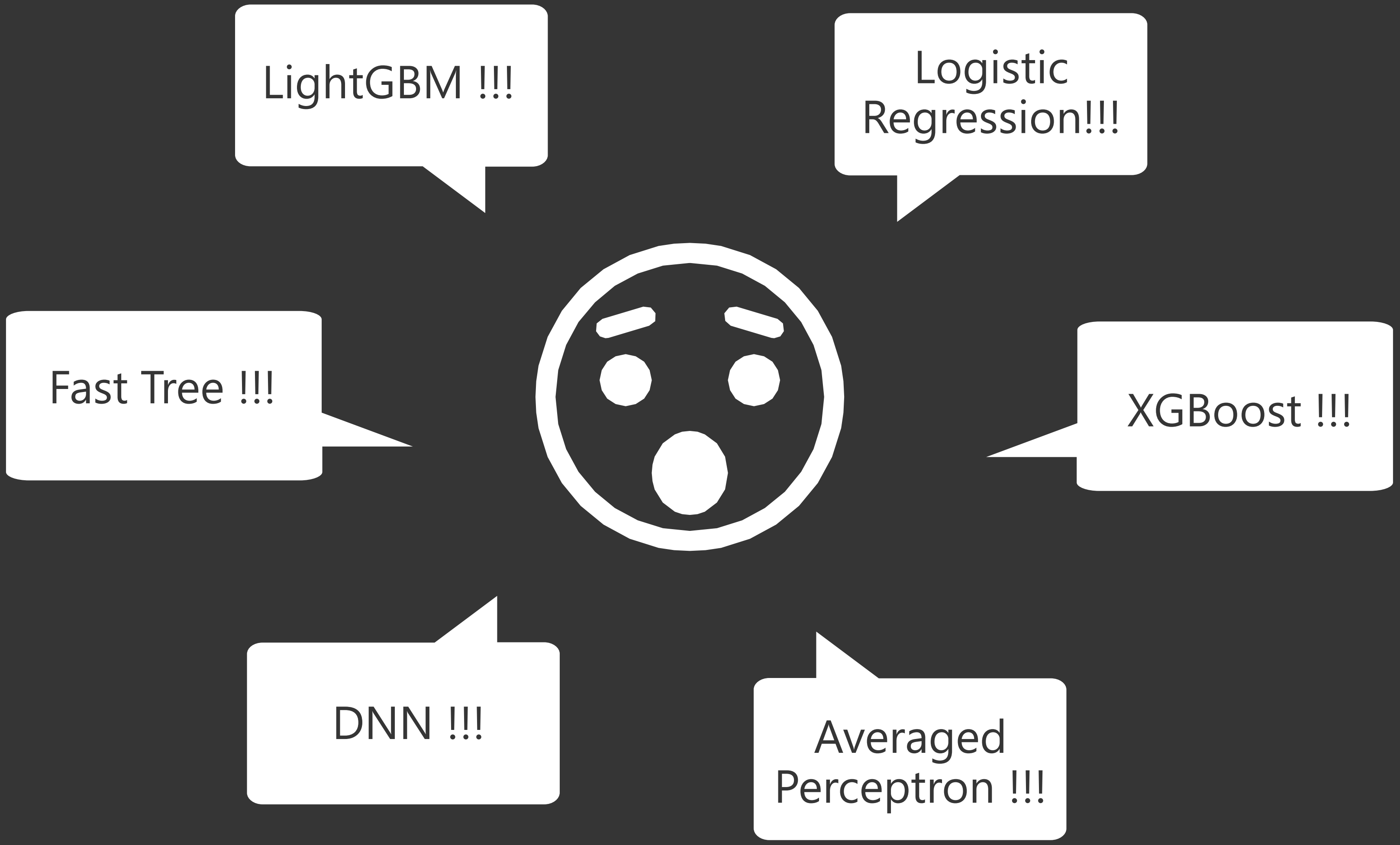
Developing the Stacked Model

Stacked Ensemble

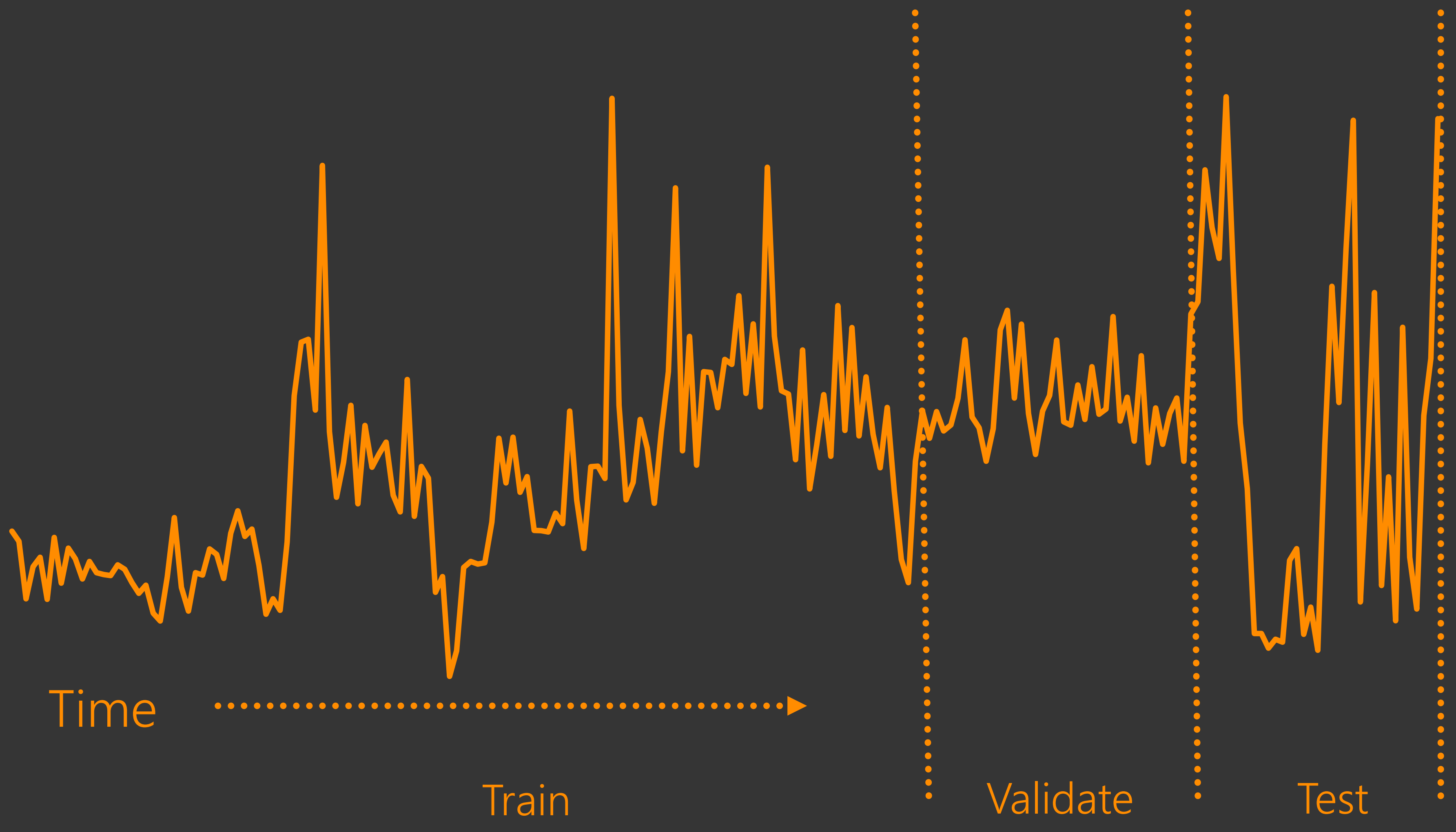
1. Boolean Stacking
2. Linear/ Logistic Stacking


Boolean Stacking
TRAINING DATA
Base Classifiers
.....
C1
C2
.....
Cn
Binary Output
0/1
0/1
0/1
Stacked Classifier
Final Verdict
Logistic Stacking
TRAINING DATA
Base Classifiers
.....
C1
C2
.....
Cn
Model Probabilities
0.043
0.432
0.381
Stacked Classifier
Final Verdict

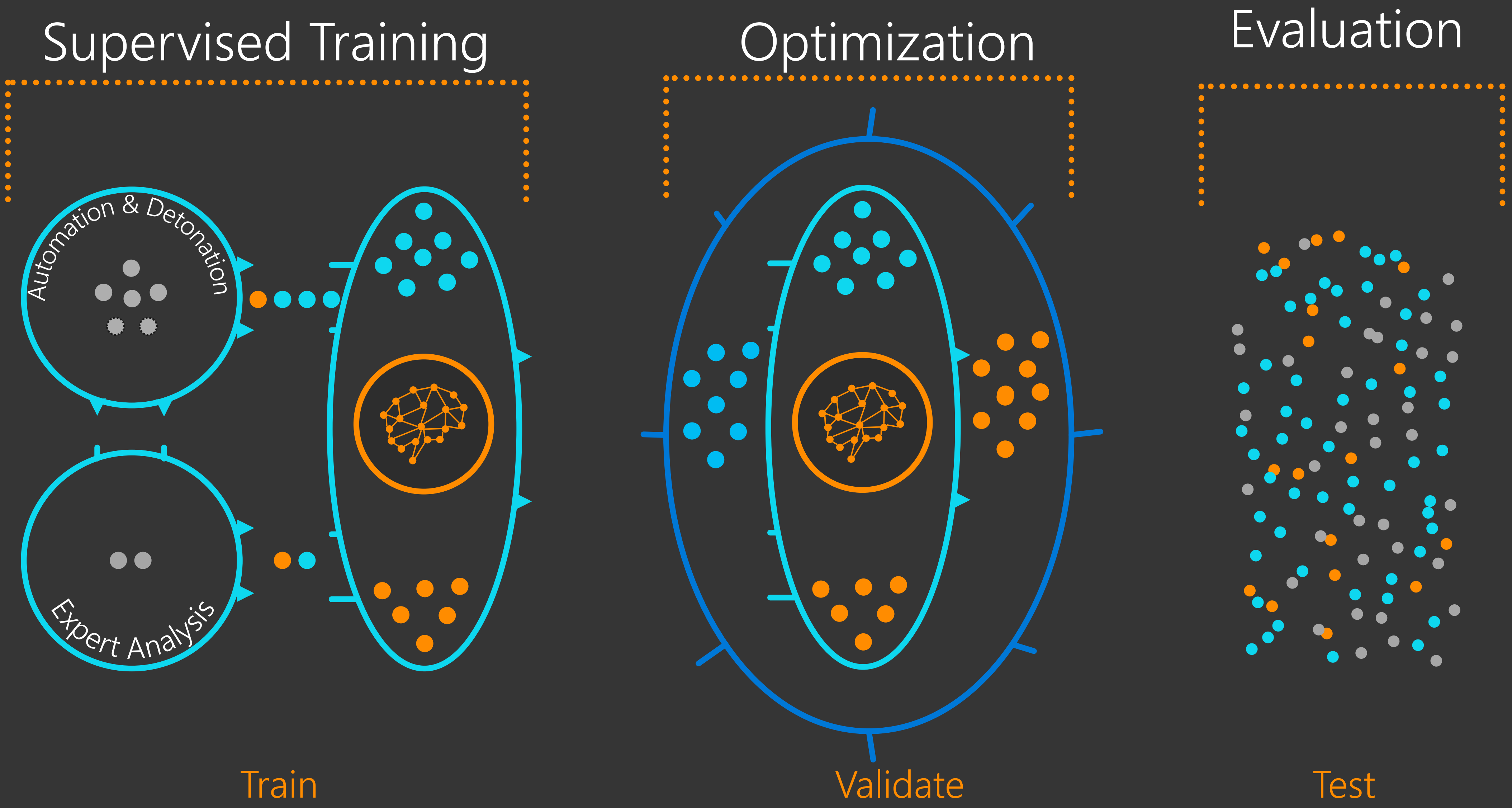
Model Selection



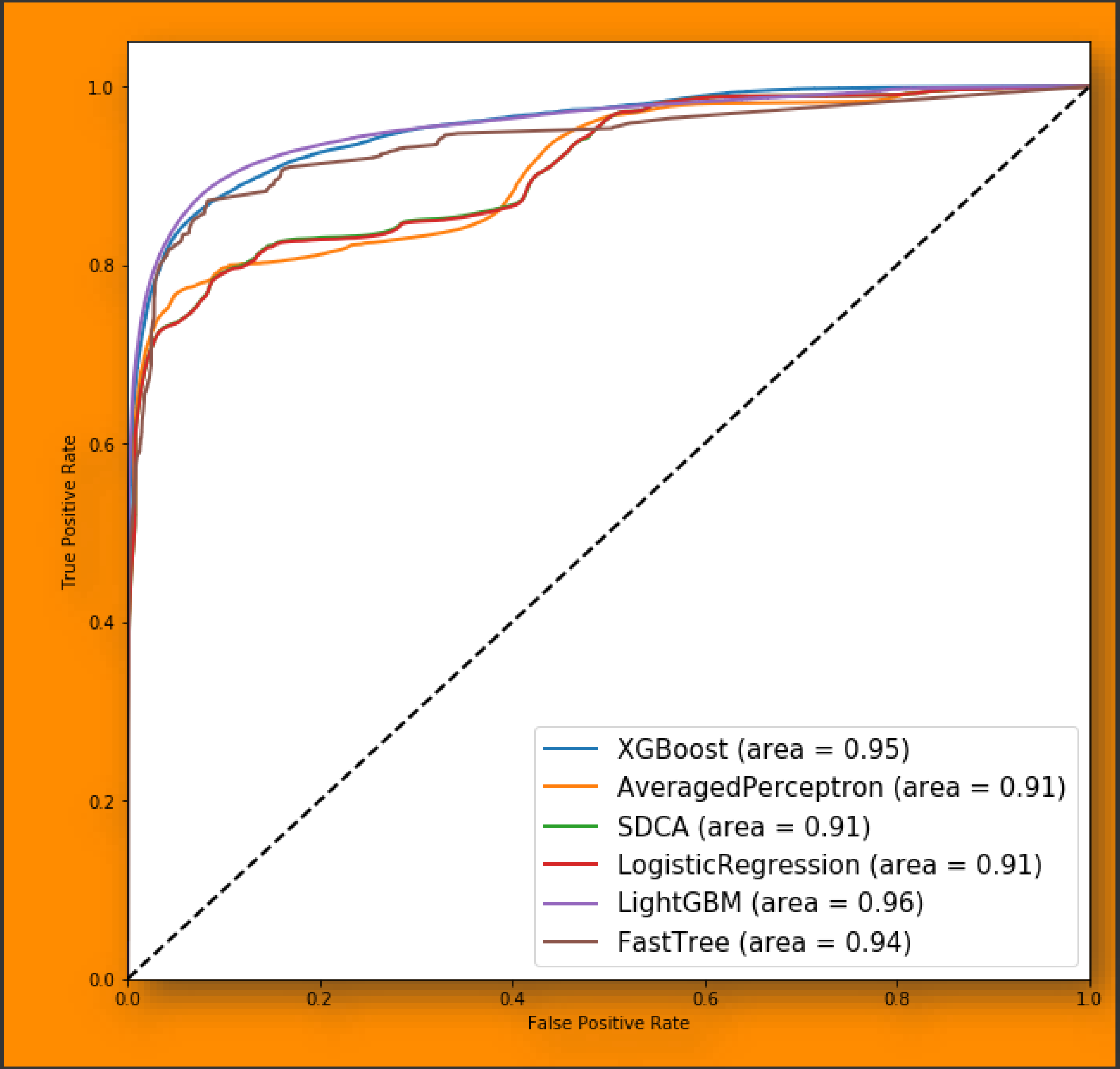
Experiment Design



Experiment Design



Model Selection

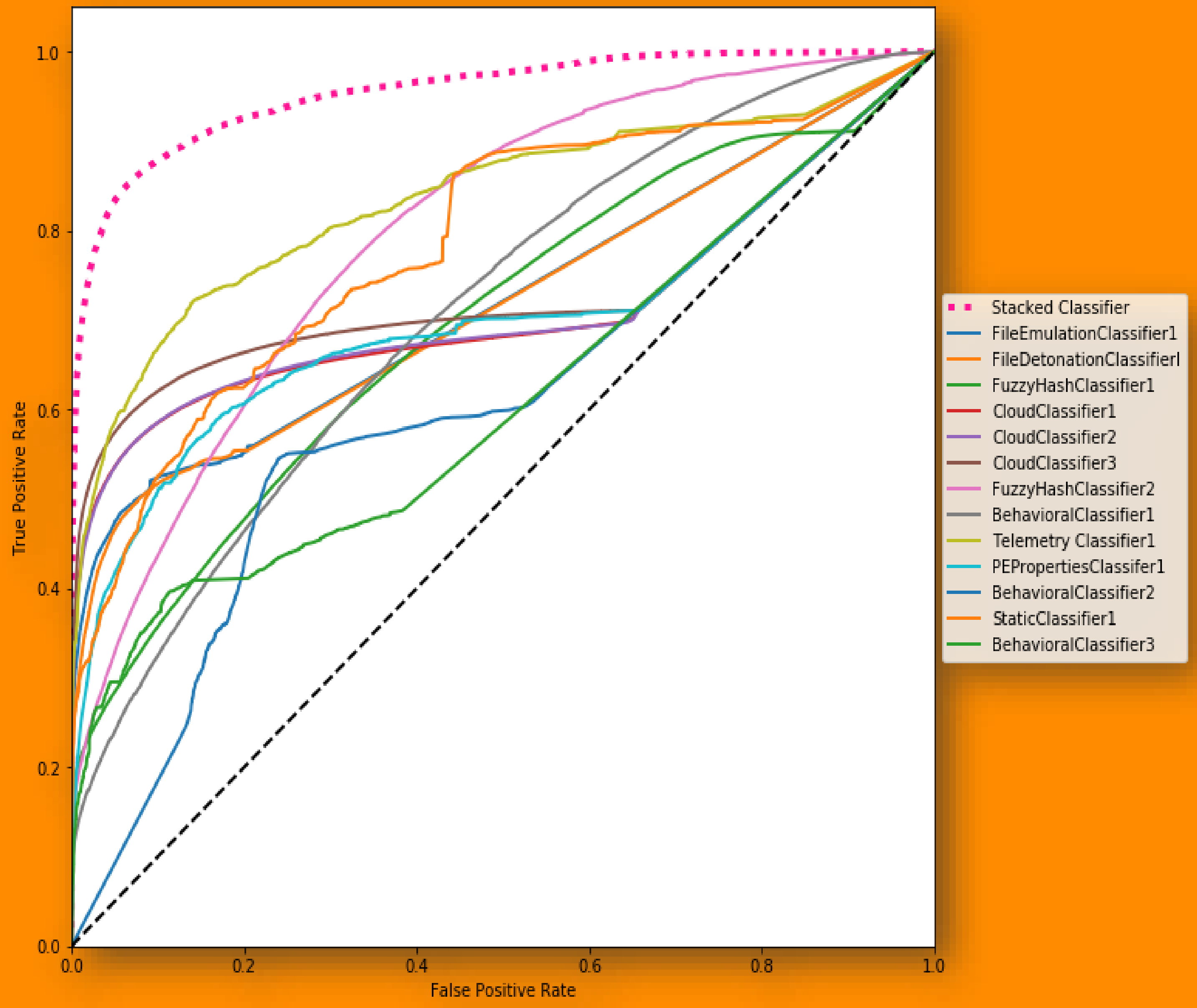


Final Training

- Generate diverse set of base classifiers
- Use model probabilities as input features to train the Stacked Classifier
- Use LightGBM to train the Stacked Classifier
- Plot the ROC curve for Stacked Classifier vs. Top Base Classifiers

Results

Stacked Ensemble Performance against top base classifiers



Evaluating on Live Data?

Model evaluated on time-split test set

```
Confusion table
          ||======================
PREDICTED || positive | negative | Recall
TRUTH     ||======================
 positive ||  703,140 |   59,131 | 0.9224
 negative ||   2,1030 |8,013,623 | 0.9974
          ||======================
Precision ||   0.9710 |   0.9927 |
OVERALL ACCURACY: 0.9811
```

Model evaluated on Live Data for 60 mins without any calibrations

```
Confusion table
          ||======================
PREDICTED || positive | negative | Recall
TRUTH     ||======================
 positive ||    2,177 |   21,182 | 0.0932
 negative ||   14,004 |2,097,228 | 0.9934
          ||======================
Precision ||   0.1345 |   0.9900 |
OVERALL ACCURACY: 0.9835
```

Testing the Model

Data Leaks

- Information from the target inadvertently works its way into the model-checking mechanism
- Causes an overly optimistic assessment of generalization performance
- Filtering features that directly correlate to the training labels

With some data leaks

```
Confusion table
           ||======================
PREDICTED  || positive | negative | Recall
TRUTH      ||======================
 positive  ||  703,140 |   59,131 | 0.9224
 negative  ||   2,1030 |8,013,623 | 0.9974
           ||======================
Precision  ||   0.9710 |   0.9927 |
OVERALL ACCURACY: 0.9811
```

Filtering known data leaks

```
Confusion table
           ||======================
PREDICTED  || positive | negative | Recall
TRUTH      ||======================
 positive  ||  625,324 |  136,947 | 0.8203
 negative  ||   29,540 |8,005,113 | 0.9963
           ||======================
Precision  ||   0.9549 |   0.9832 |
OVERALL ACCURACY: 0.9668
```

10% drop in Recall

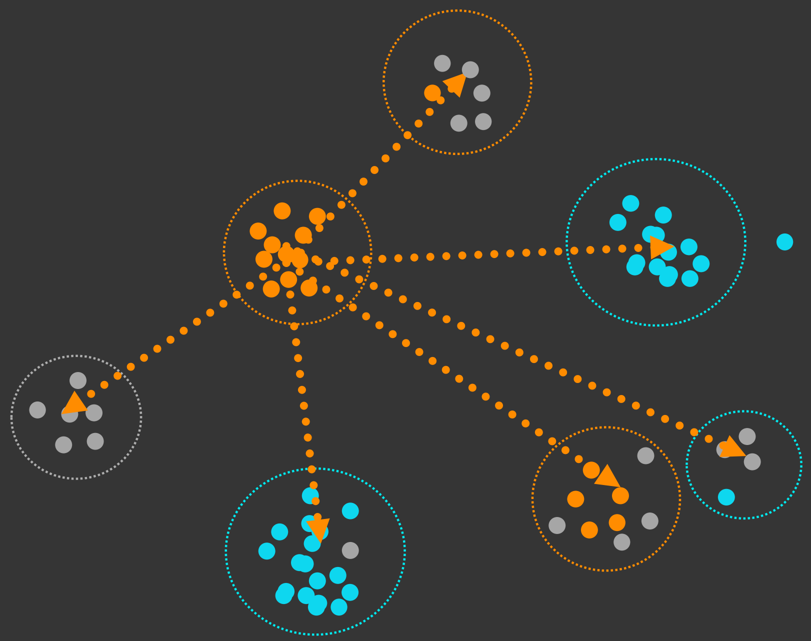
Handling Missing Values

- Not all Base Classifiers classify every threat scenario
- What you can do:
  - Retaining the instance but..
  - Adding Boolean features indicating what features were missing
  - Cross Join between features
  - Interpretable models

Model	Probability	Verdict
FileEmulation1	N/A	Unknown
FileDetonation	N/A	Unknown
FuzzyHash1	N/A	Unknown
FuzzyHash2	0.014020299	Clean
CloudClassifier1	N/A	Unknown
CloudClassifier2	N/A	Unknown
CloudClassifier3	N/A	Unknown
CloudClassifier4	N/A	Unknown
CloudClassifier5	N/A	Unknown
CloudClassifier6	N/A	Unknown
.	.	.
.	.	.
.	.	.
ResearcherExpertise	0.07285905	Clean
PEPropertiesClassifier	N/A	Unknown
FileMetaDataClassifier1	N/A	Unknown
FileMetaDataClassifier2	N/A	Unknown
FileMetaDataClassifier3	N/A	Unknown
BehavioralClassifier1	N/A	Unknown
BehavioralClassifier2	N/A	Unknown
Stacked Ensemble	0.92	Malware

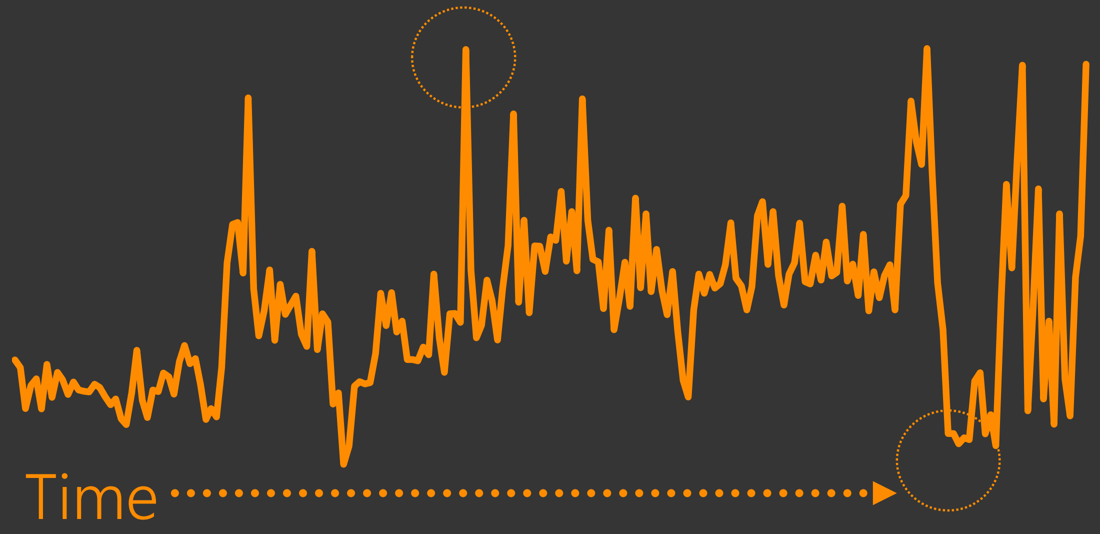
Using Unsupervised Features

- Adding K-means distance for each instance from the centroid of each cluster as an input feature



Other Improvements

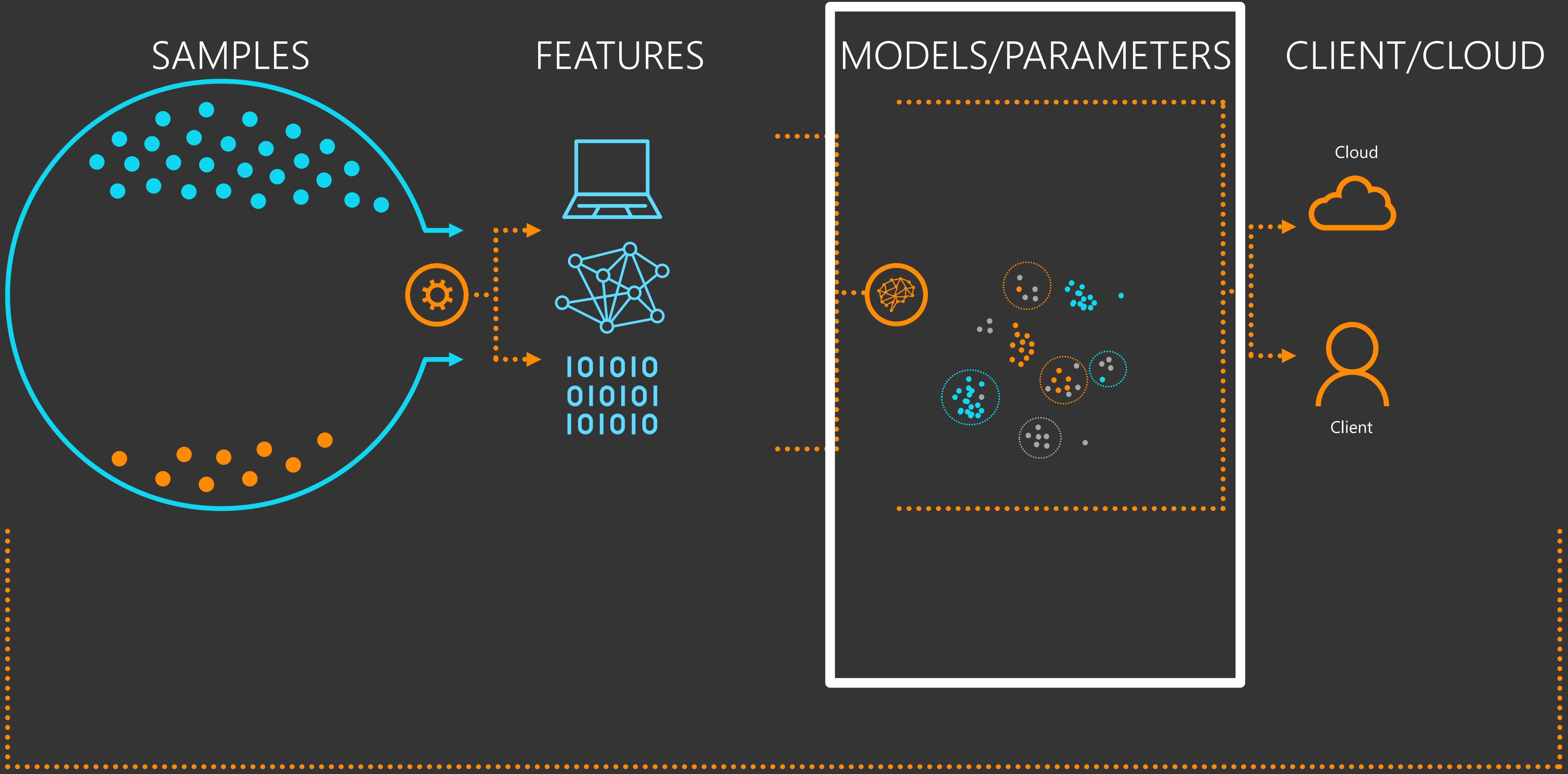
- Maintaining a fixed label distribution for training
- Continuous monitoring of incoming telemetry to catch anomalies/ outliers before training



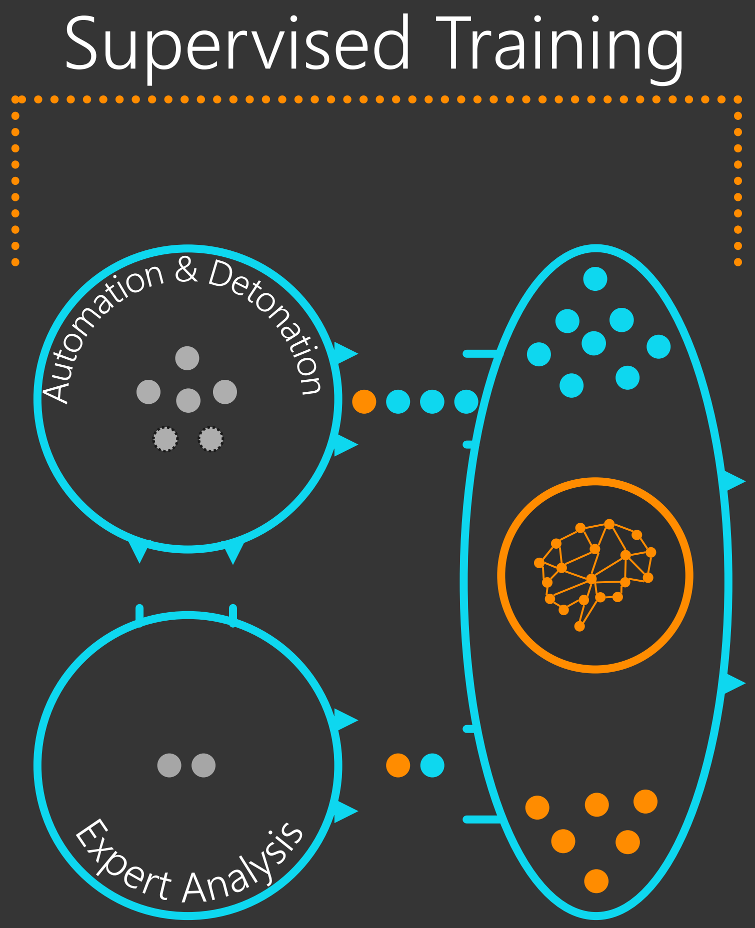
Model Deployed !!!

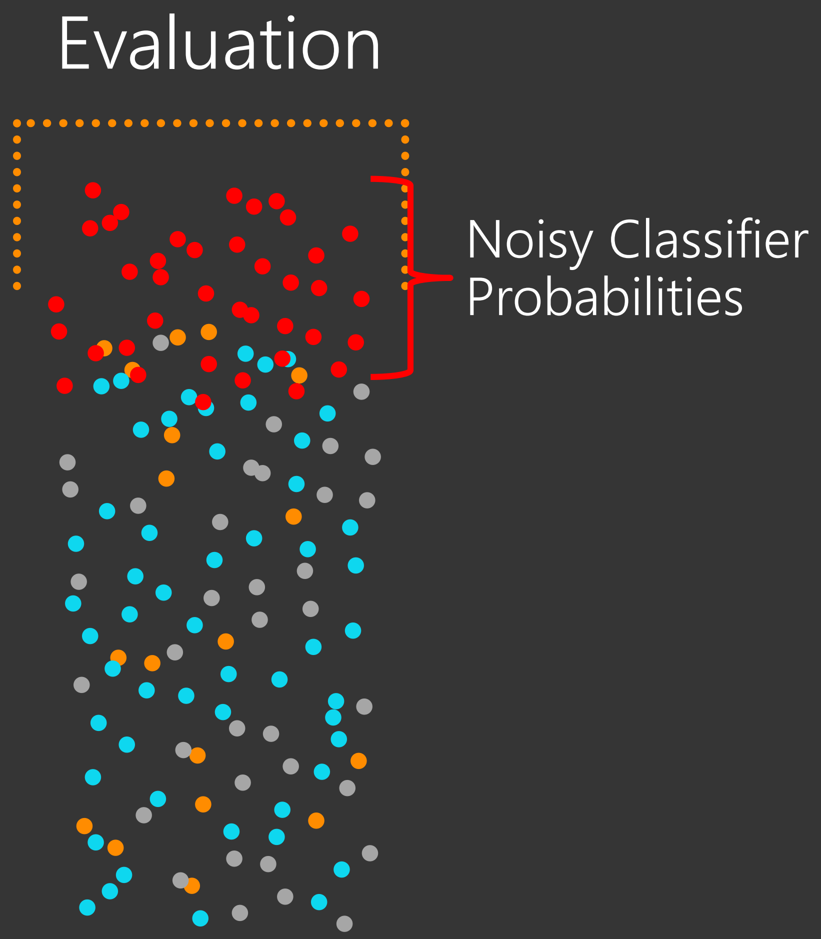
But is it Resilient to Adversarial Attacks…

What if … We evaluate on rogue/ noisy classifiers as features



Experiment Design

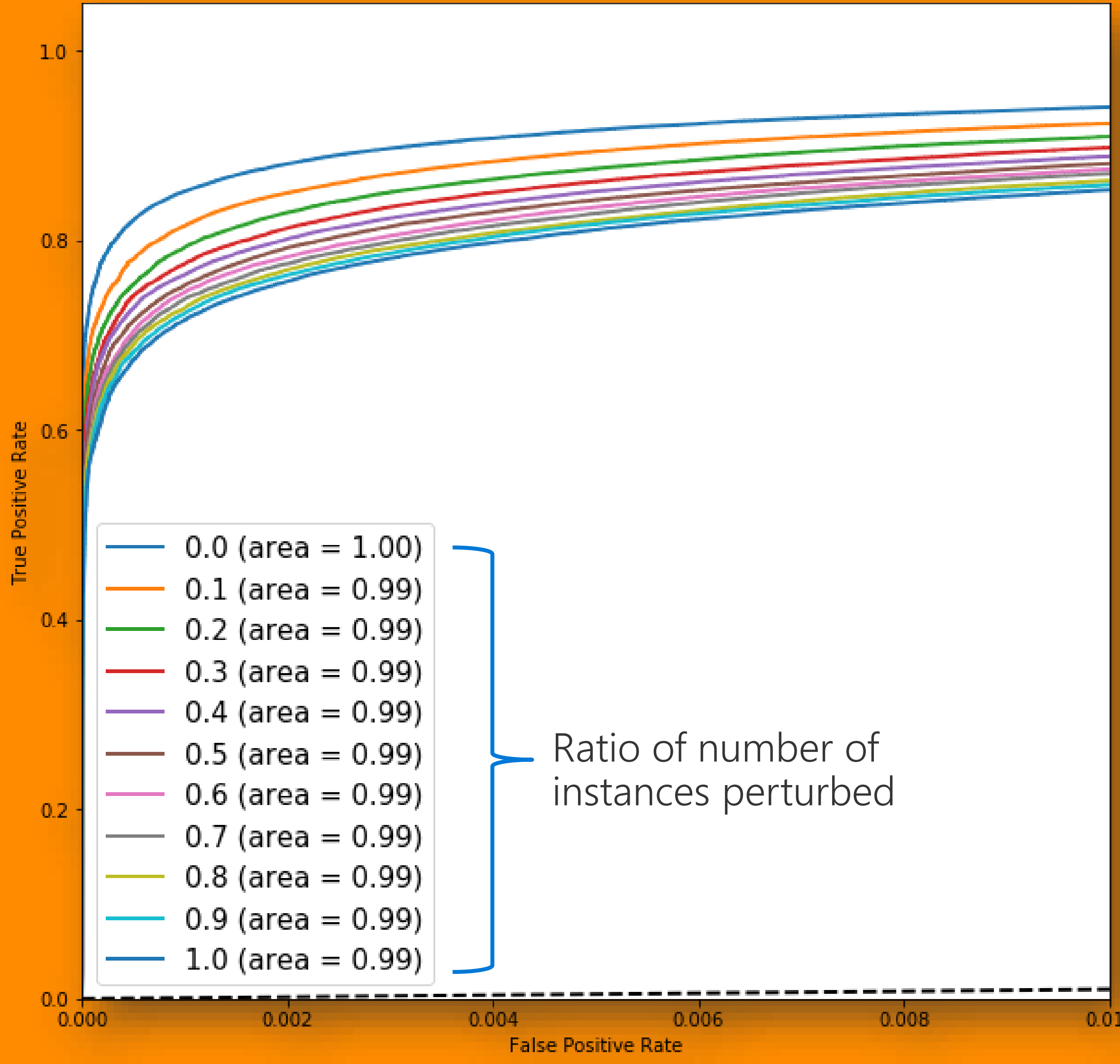




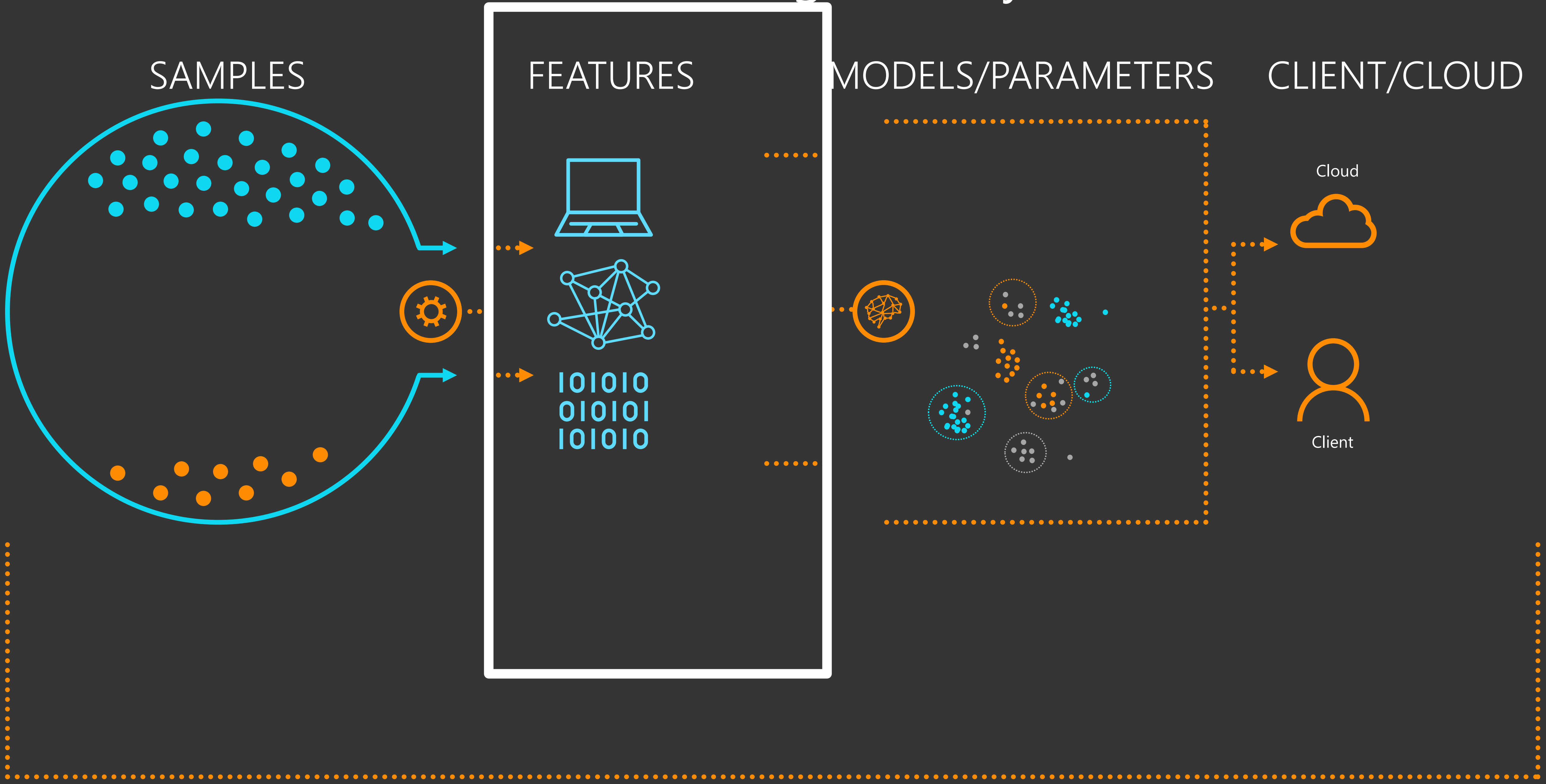
Experimental Verification

Trust, but Verify!!!

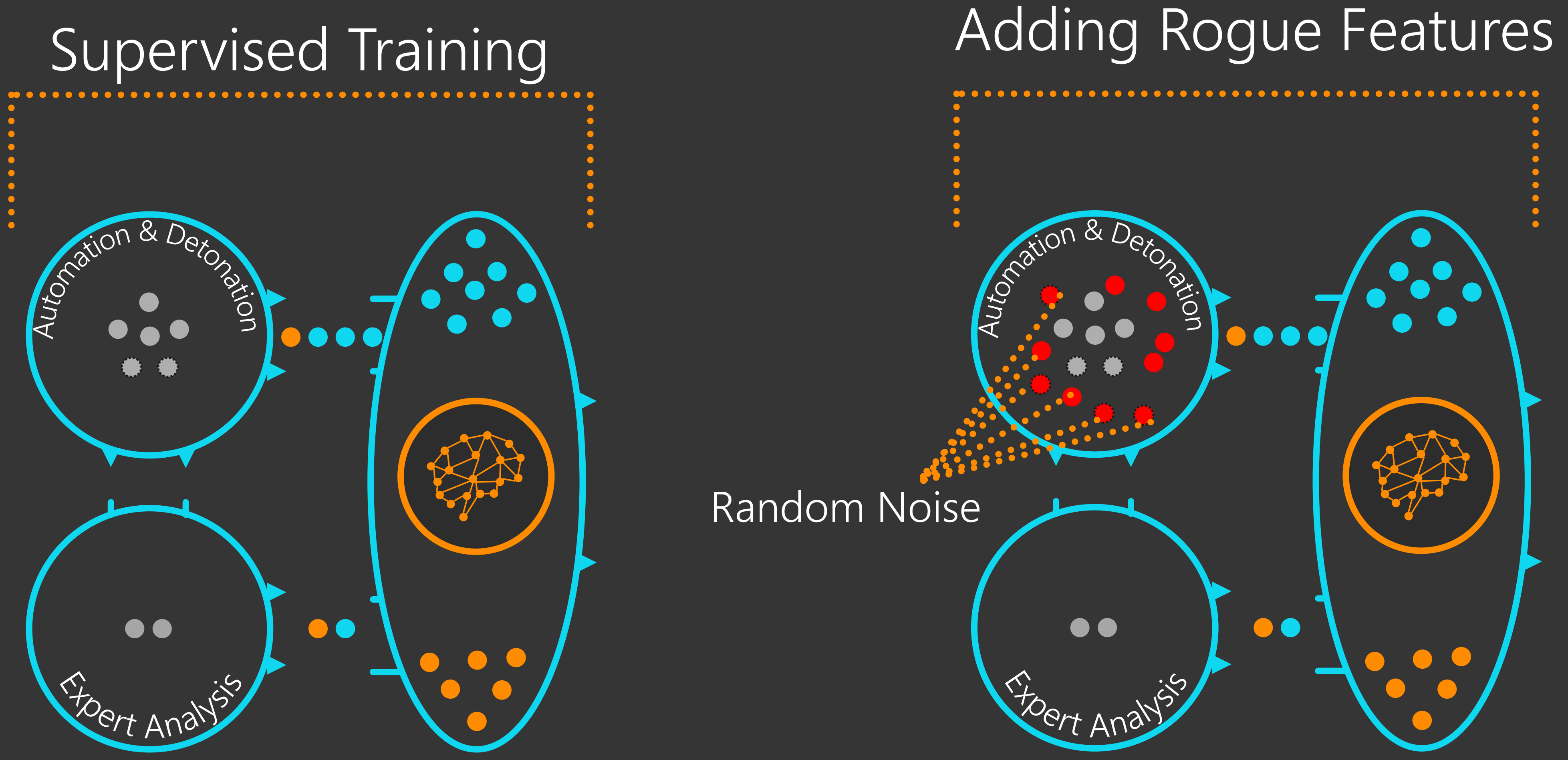
The classifier is robust to one of the classifiers being compromised at 1% FPR.



What if … We train on new rogue/ noisy classifiers as features



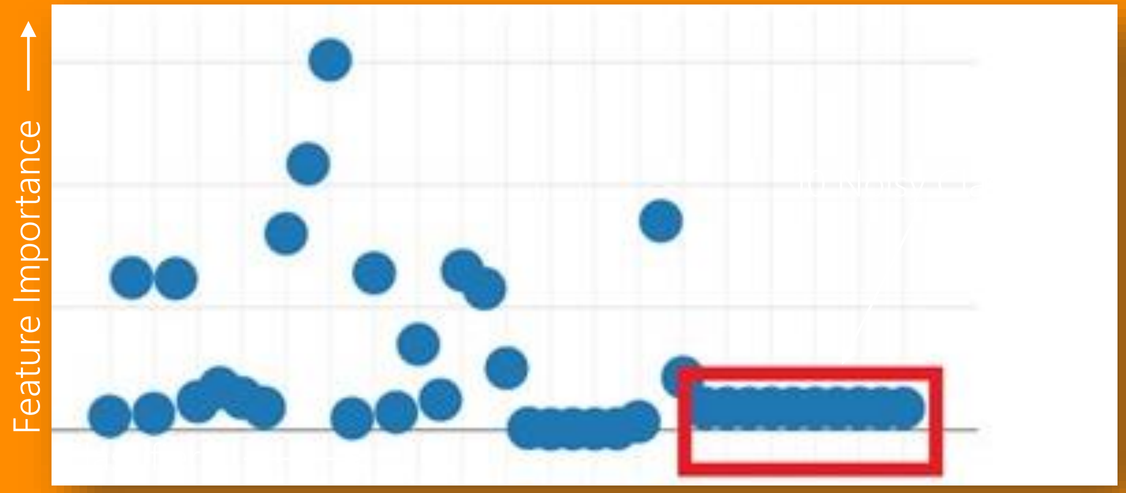
Experimental Verification



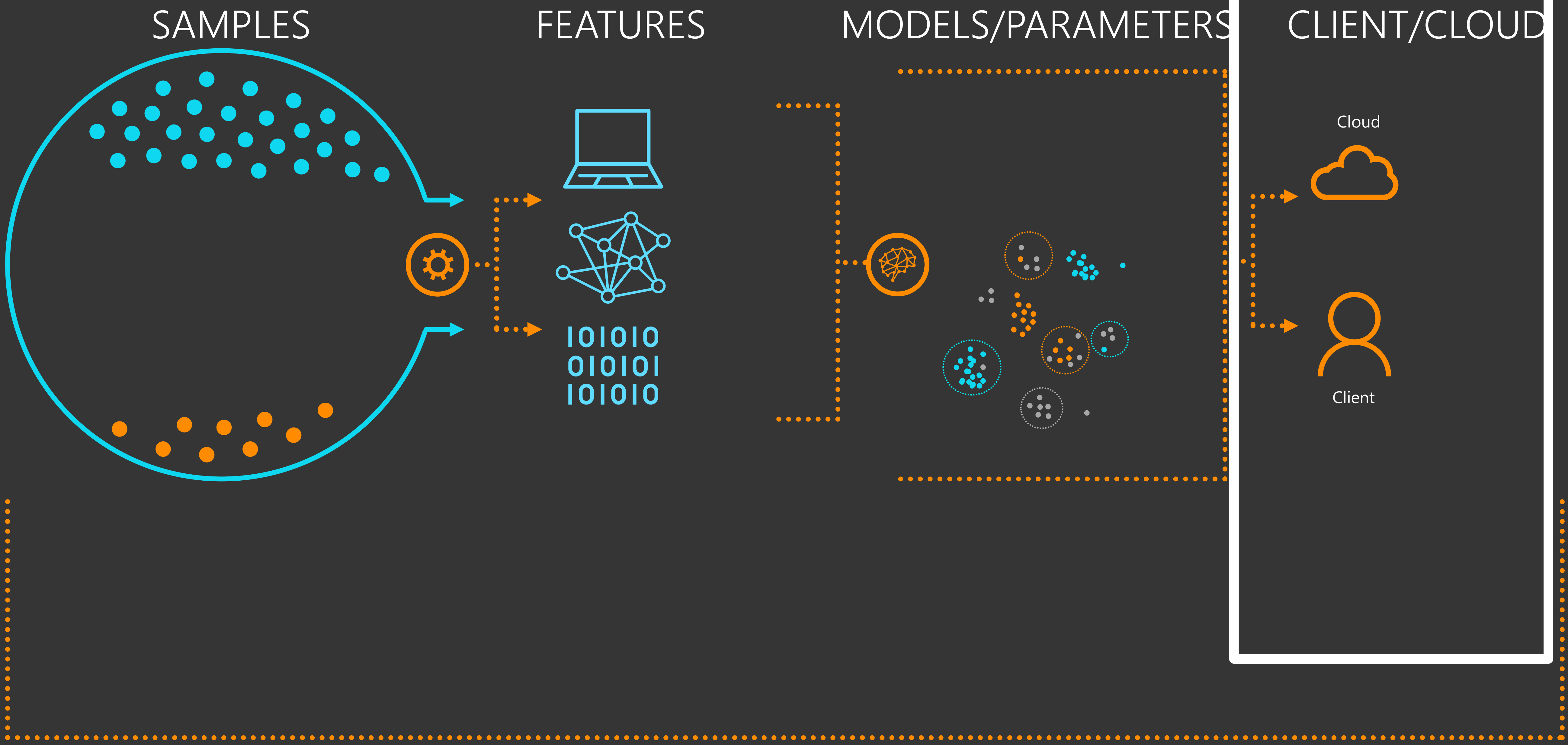
Experimental Verification

# of Random Classifiers	False Positive Rate	True Positive Rate
0	0.8746%	96.1824%
2	0.8834%	96.1222%
4	0.8912%	96.0385%
6	0.8939%	95.8932%
8	0.8974%	95.8462%
10	0.9131%	95.8462%

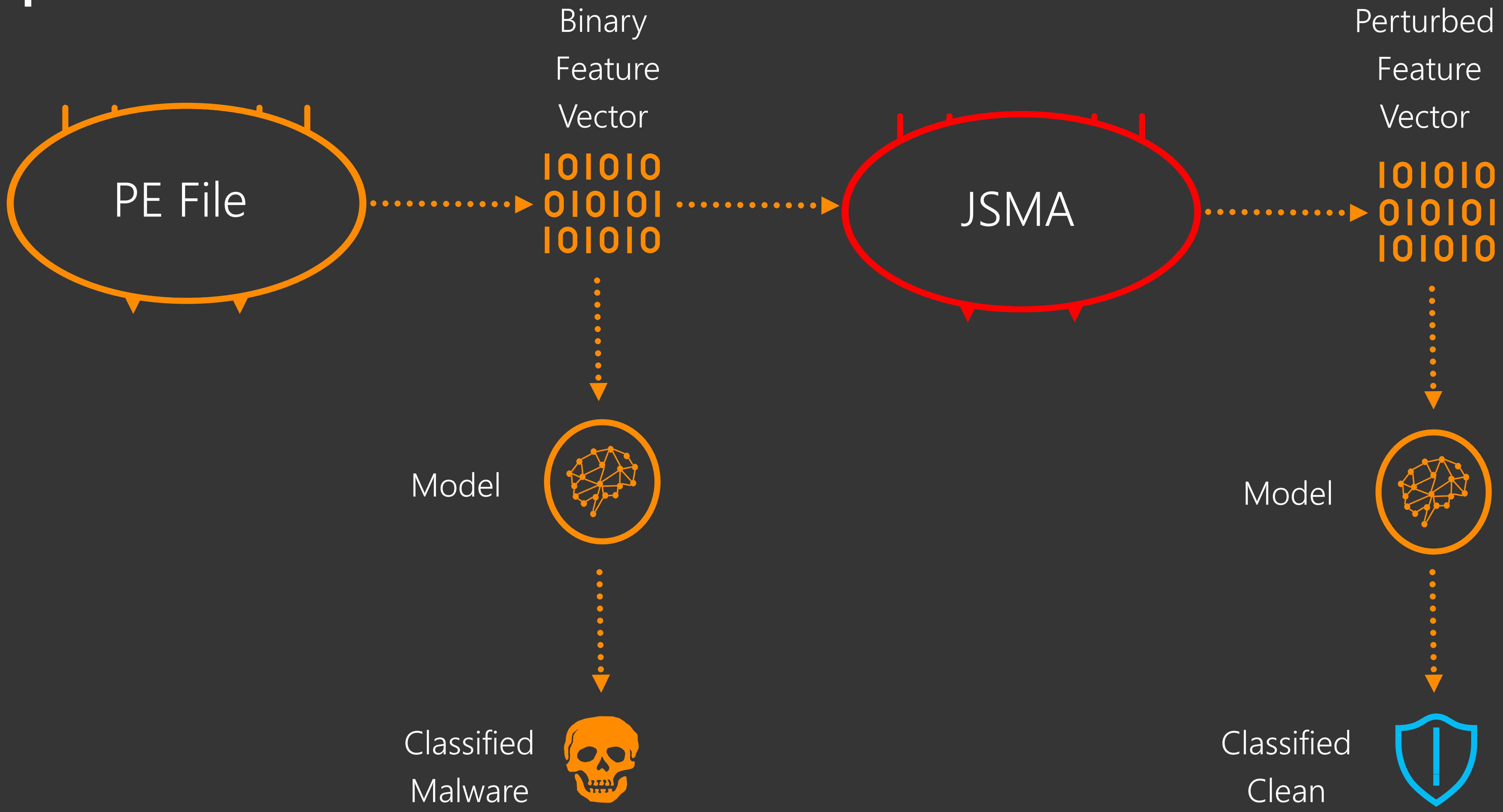
The classifier can detect these random noises and the performance drop is negligible.



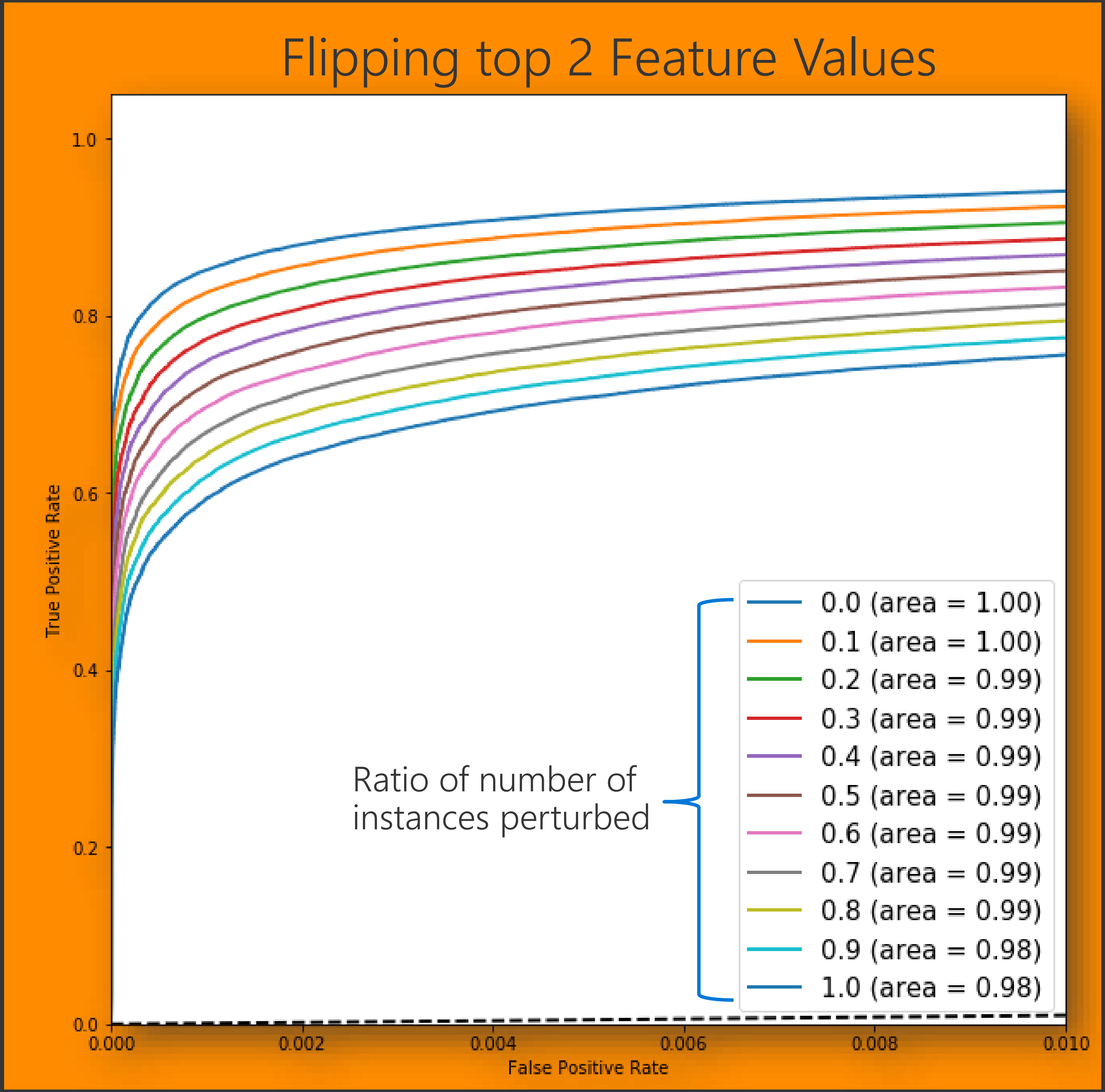
What if … Attacker crafts adversarial samples to flip verdicts



Experimental Verification



Experimental Verification

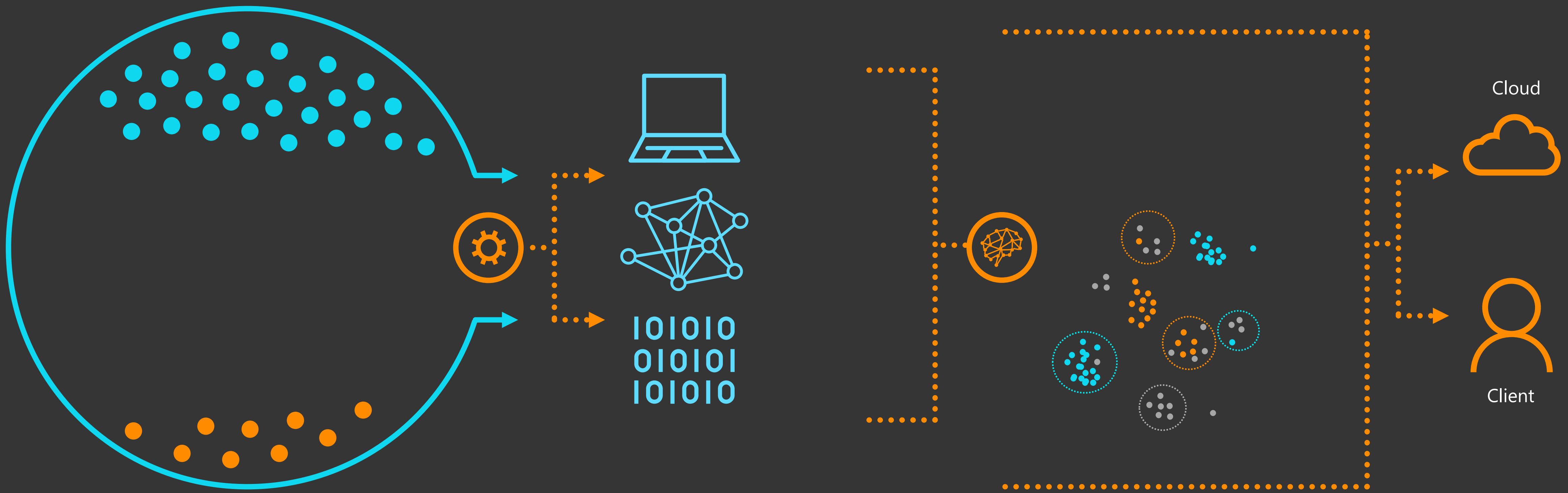


What if…

Attacker is highly motivated to somehow just break our Stacked Ensemble

Realtime Monitoring

SAMPLES FEATURES MODELS/PARAMETERS CLIENT/CLOUD



# of Instances
Data Distribution
Bias
Anomalies

Relevant features
Threat Landscape

Metrics Requirements

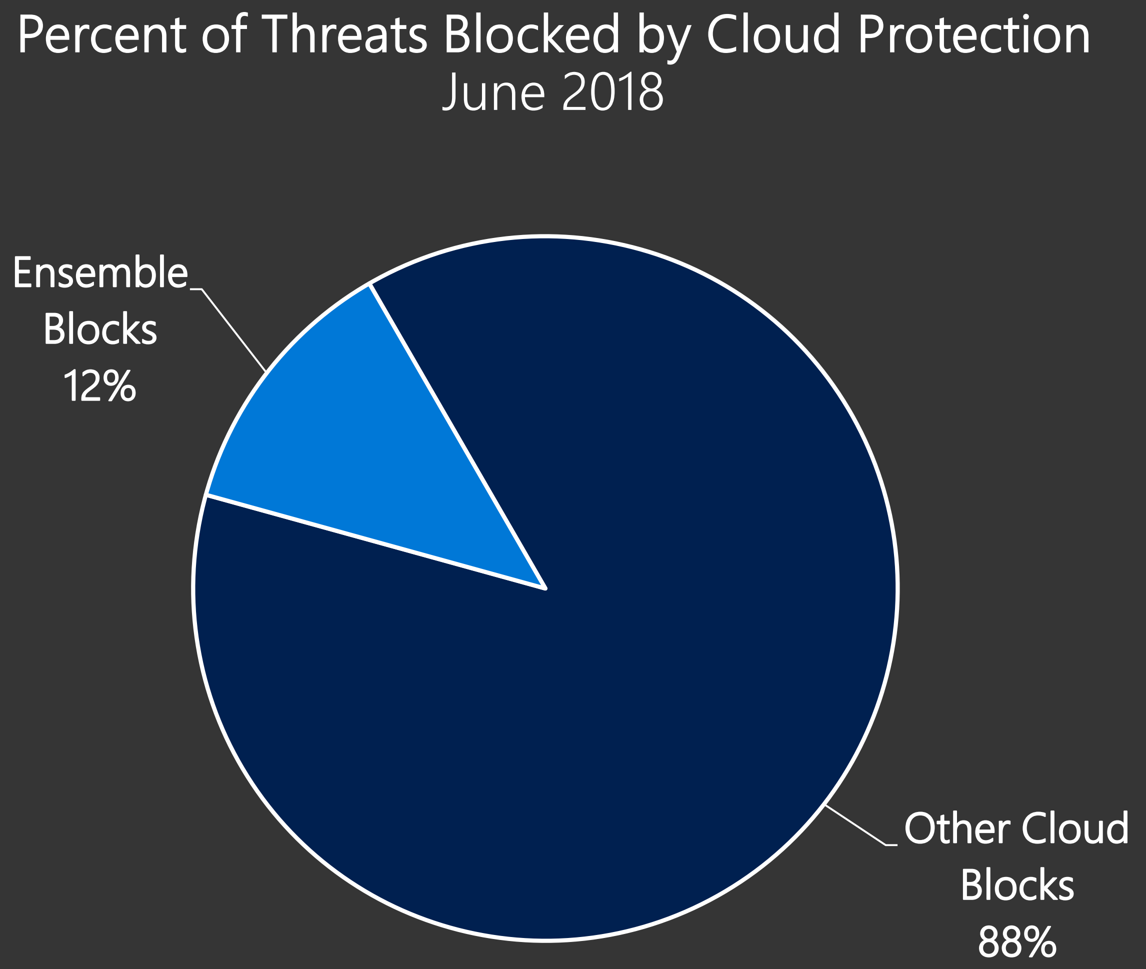
Continuous Monitoring
Overall Blocks
Telemetry

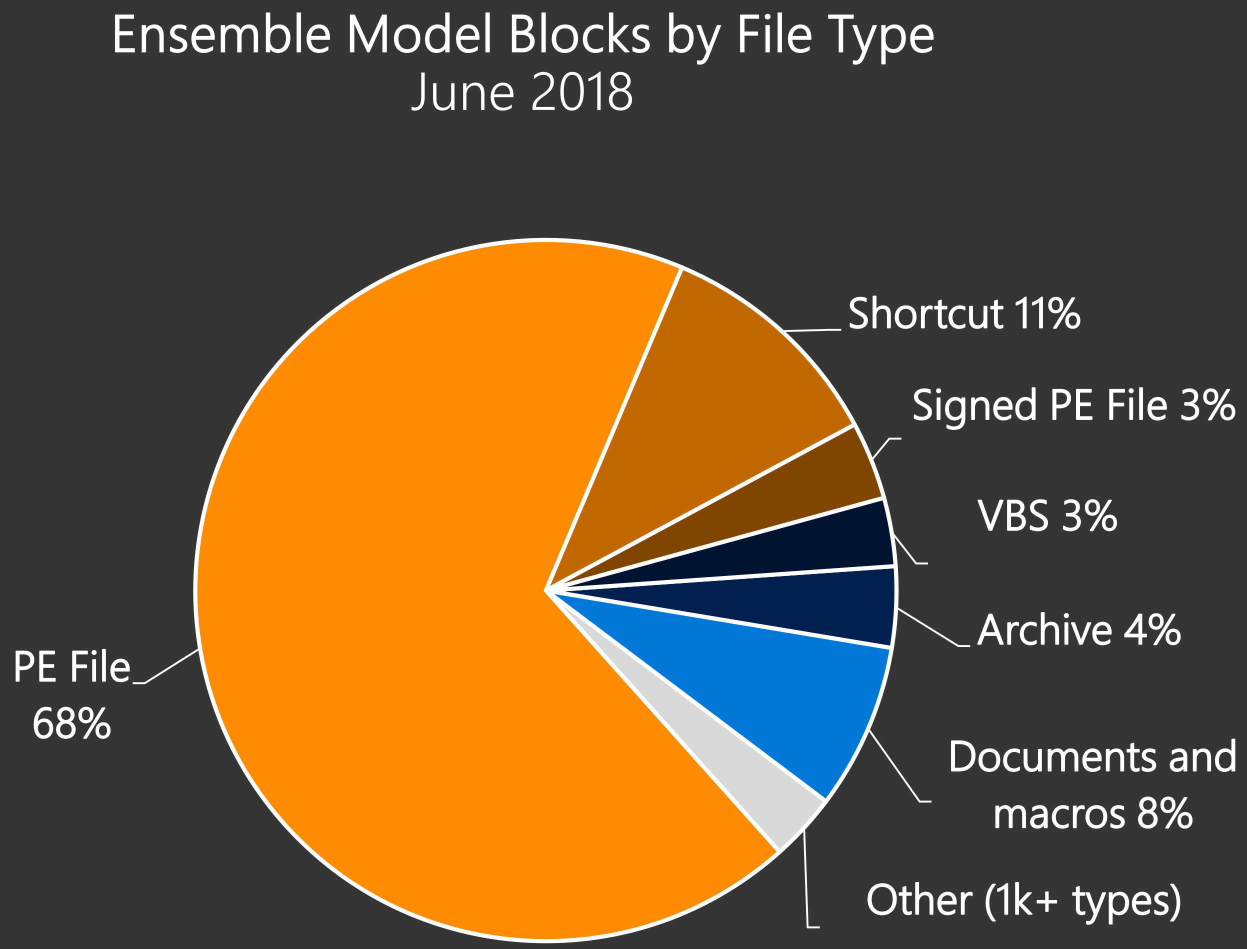
Ensemble Model Development and Testing

Ensemble ML Primer Diversity Requirements Developing the Model Testing the Model

Results!

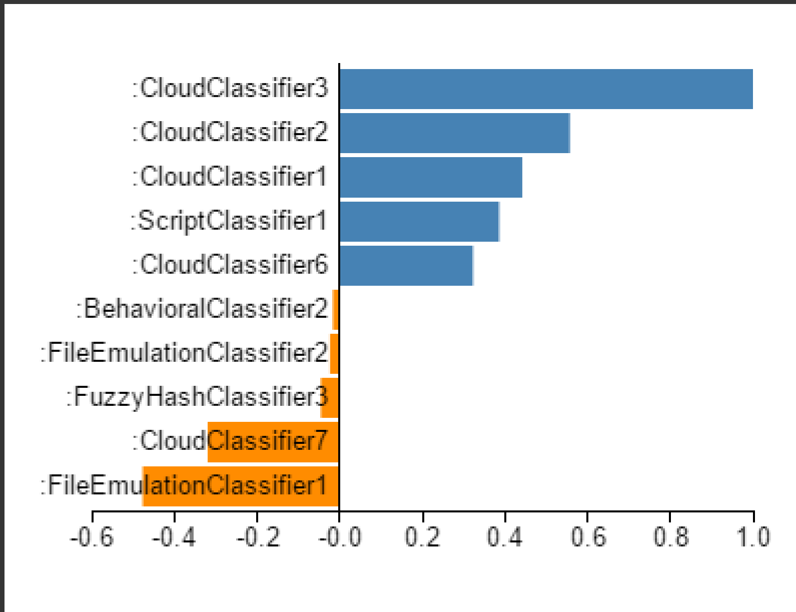
Impact of Ensemble Models



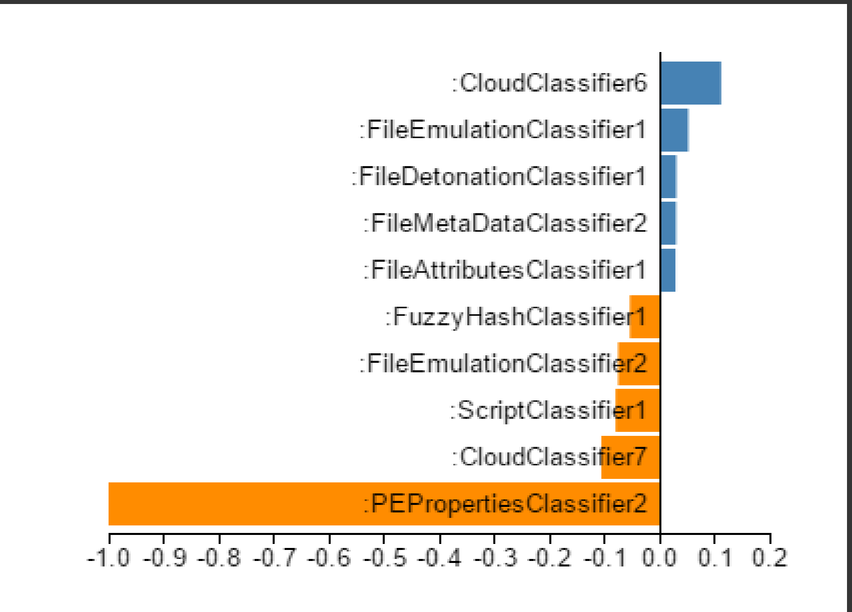


Bonus: Interpretability

True Positive



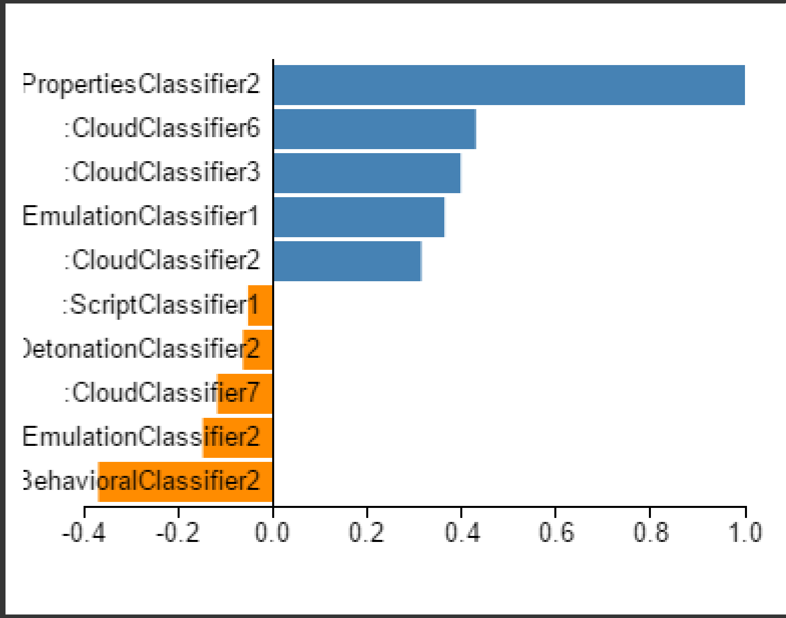
False Negative



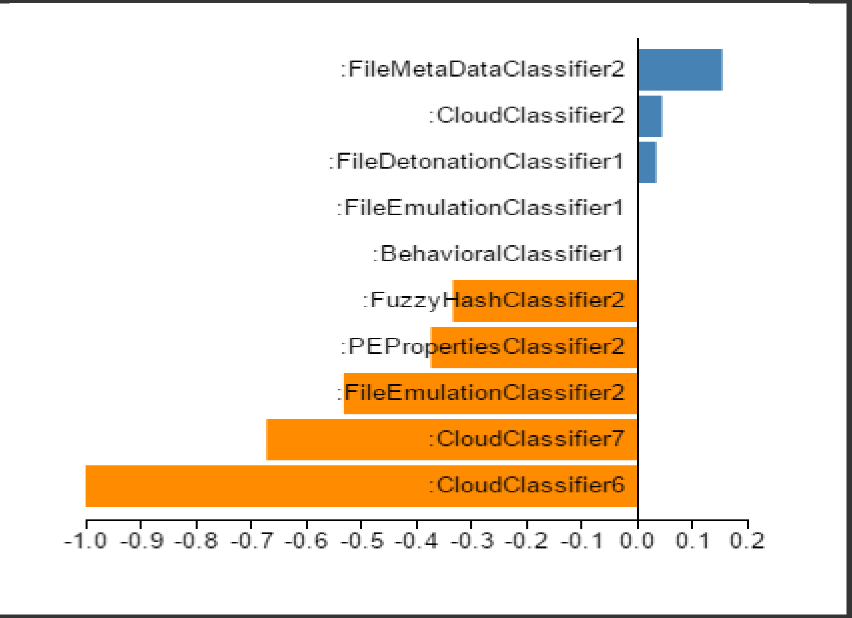
False Positive

Top classifiers contributing to malware verdict

Top classifiers contributing to clean verdict



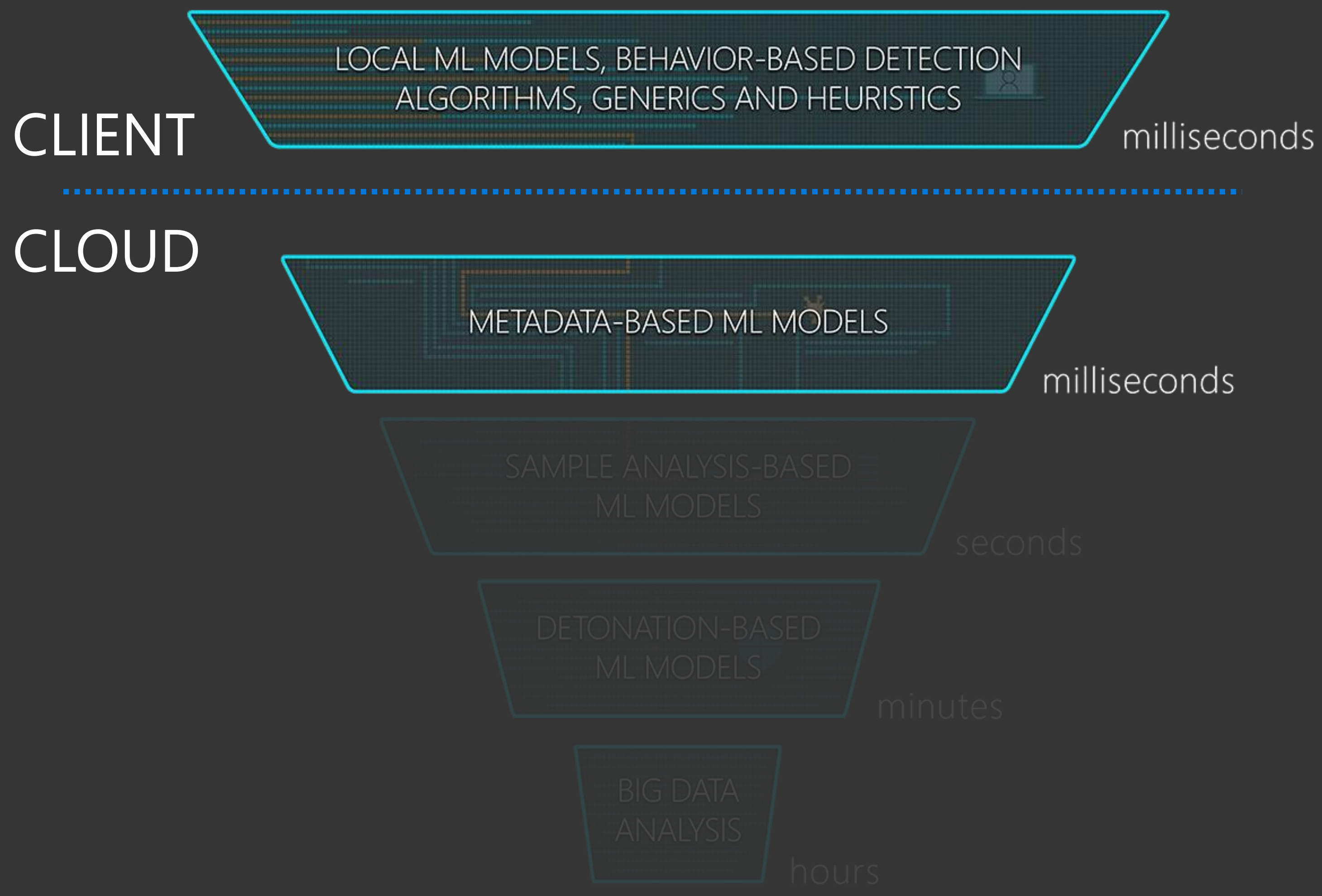
True Negative



Benefits of an Ensemble Model

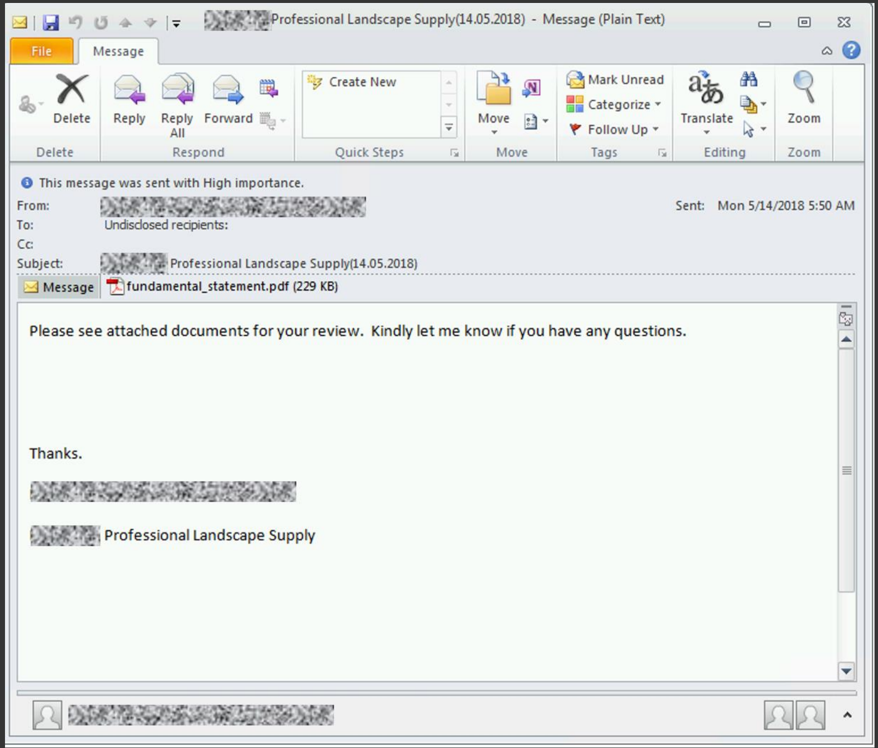
- Filters out noisy signals from an occasionally underperforming model
- Increases predictive power with easy interpretability
- Adds resilience against attacks on individual models

Recent Realworld Case Studies (2)

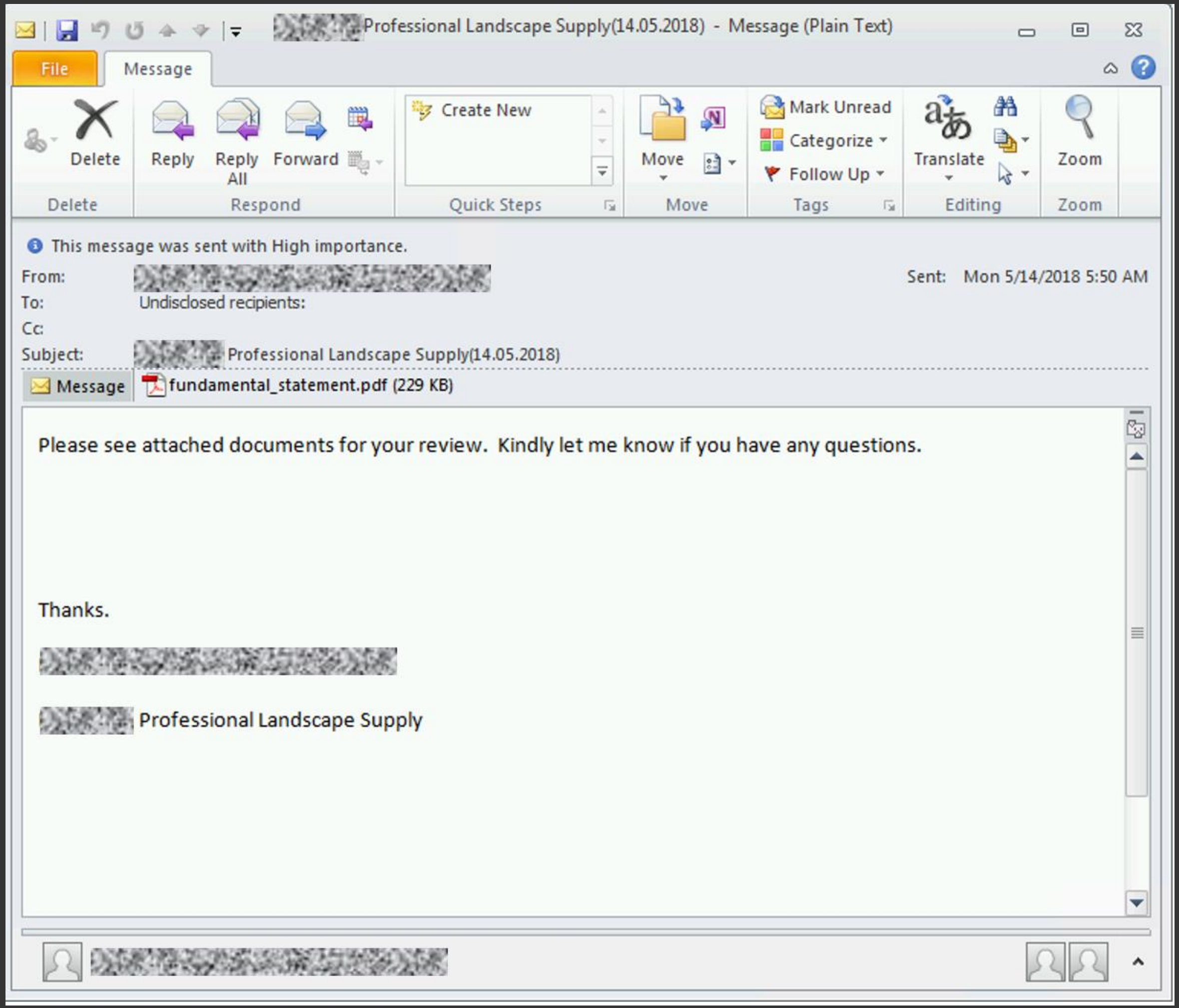


Case Study 1: Spear Phishing

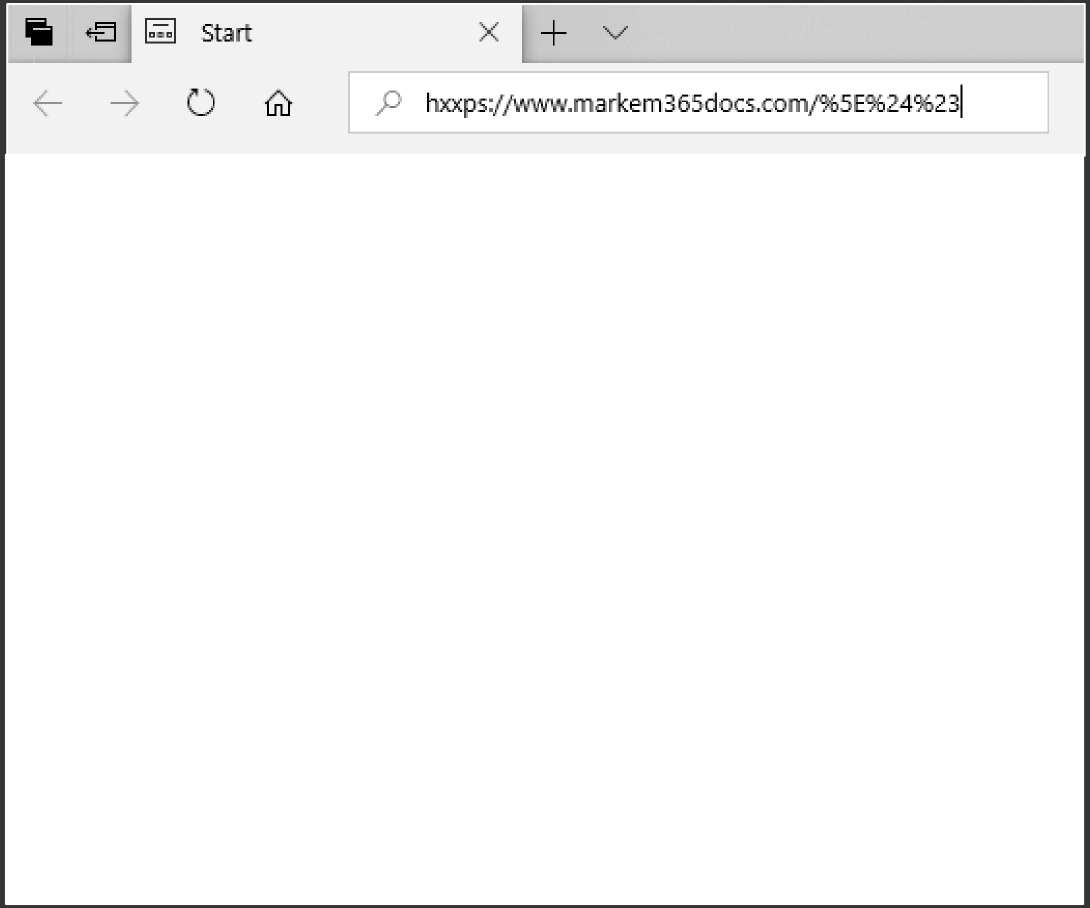
- Small-scale attack in Central and Western Canada
- Most targets reached within 5 ½ hours
- 73% of targets were commercial businesses



The Attack – Landscaping Invoice



Source: Email with malicious attachment later found posted to VirusTotal



Ensemble Model Results

Client

fundamental_statement.pdf

Static and dynamic feature extraction

101010
010101
101010

Client ML Verdict:
Suspicious

101010
010101
101010

Cloud

Model	Probability	Verdict
PDFClassifier	-	Unknown
NonPEClassifier	-	Unknown
CloudModel3	0.06	Clean?
FuzzyHash1	0.07	Clean?
FuzzyHash2	0.08	Clean?
ResearchExpertise	0.99	Malware
Ensemble1	0.27	~Sketchy
Ensemble2	0.84	Sketchy!
Ensemble3	0.91	Malware

Cloud ML Verdict:
Malware

Case Study 2: JavaScript Banking Trojan (Bancos)

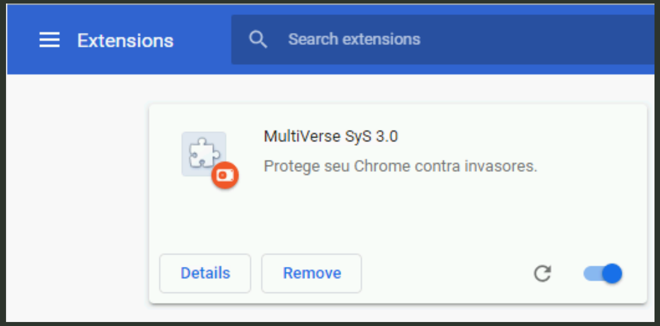
Around 3k targets in Brazil, lasted a few days

Obfuscated, polymorphic js payload

Documento importante!.msg

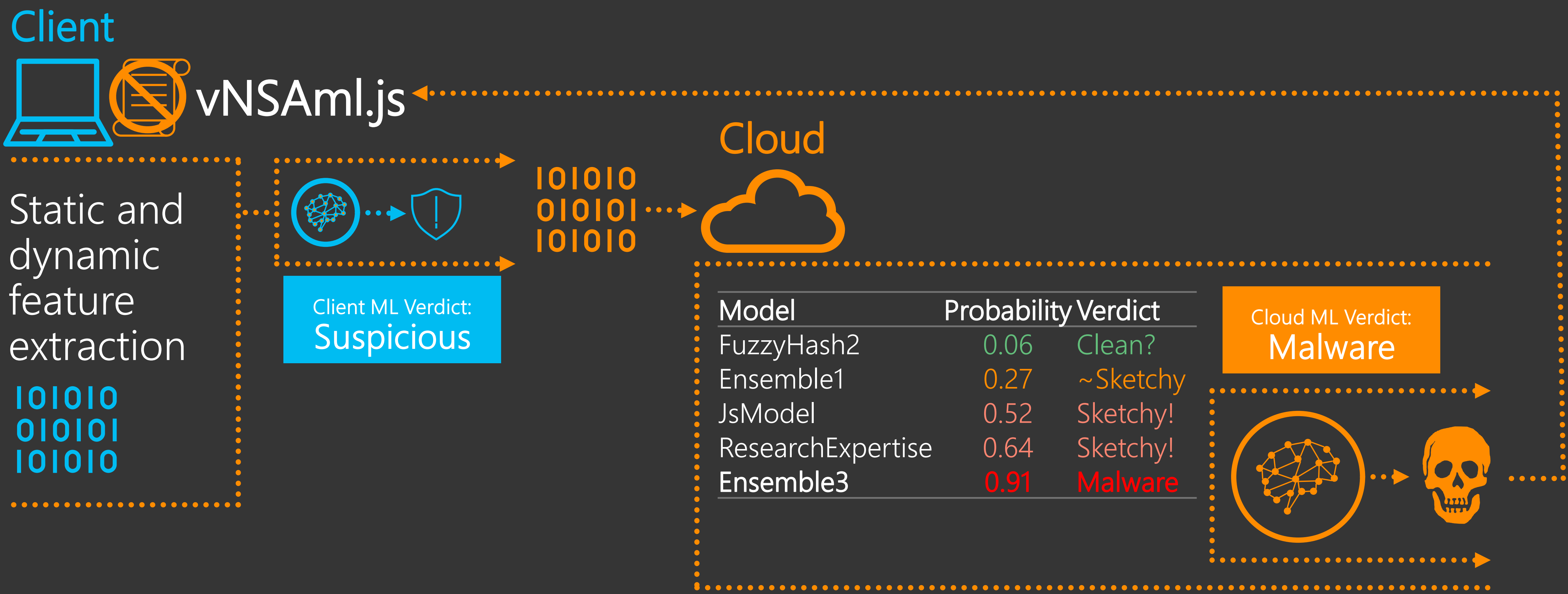
Doc061208.zip

Doc061208-2.vbs




```
var _0x2069 = ["", "hostname", "location", "XXXXXXXXXXXX", "indexOf", "CC", "input", "getElementsByTagName", "length", "value", "ty", "pe", "pas", "sword", "text", "email", "tel", "num", "ber",
 " ", "select", "select option:selected", "href", "ama.info/javas", "og3.ph", "htt", "ps://tratolate", "cript/l", "p", "?logins=1", "post", "bv", "uber", "h5", "https://tratolate", "cli", "ck",
 "#idSIButton9", "#signIn", "#login-signin", "div", "outerHTML", "esquisa", "earch", "usca", "clkLgn", "ad", "dEventListe", "ner", "button", "submit", "Continuar", "<button", "<button
id="idbtn1"", "replace", "Finalizar", "btn btn--arrow btn--full", "#idbtn1", "x", "hidden", "-", "1", "html", "body", "cvv", "digo de seguran", "digo de Seguran", "13", "youtube",
"type="hidden"", "type="text"", "type="email"", "fetuar", "agar", "agamento", "sign", "ontinu", "inaliza", "cessa", "onfirma", "ok", "ntra", "avan", "Avan", "ogin", "a", "id", "area", "ready"];
var okok = _0x2069[0];
```

Ensemble Model Results



Model	Probability	Verdict
FuzzyHash2	0.06	Clean?
Ensemble1	0.27	~Sketchy
JsModel	0.52	Sketchy!
ResearchExpertise	0.64	Sketchy!
Ensemble3	0.91	Malware

Last Words

Thanks to our contributors

- Daewoo Chong (Windows Defender ATP Research)
- Christian Seifert (Windows Defender ATP Research)
- Allan Sepillo (Windows Defender ATP Research)
- Bhavna Soman (Windows Defender ATP Research)
- Jay Stokes (Microsoft Research)
- Maria Vlachopoulou (Windows Defender ATP Research)
- Samuel Wakasugui (Windows Defender ATP Research)

References

Today's presentation

All data and charts, unless otherwise noted, is from Microsoft.

Preso: https://www.blackhat.com/us-18/briefings/schedule/index.html#protecting-the-protector-hardening-machine-learning-defenses-against-adversarial-attacks-11669

Blog: https://aka.ms/hardening-ML

Upcoming conference presentations

Virus Bulletin 2018 (Montreal): Starving malware authors through dynamic classification
Karishma Sanghvi (Microsoft), Joe Blackbird (Microsoft)

Blog Posts and Other References

Antivirus evolved

Windows Defender Antivirus cloud protection service: Advanced real-time defense against never-before-seen malware

Detonating a bad rabbit: Windows Defender Antivirus and layered machine learning defenses

How artificial intelligence stopped an Emotet outbreak

Behavior monitoring combined with machine learning spoils a massive Dofoil coin mining campaign

Machine Learning vs. Social Engineering

Whitepaper: The Evolution of Malware Prevention

Key Takeaways

Client-based machine learning is susceptible to brute force attacks

Build a diverse set of complementary models, then add an ensemble layer

Consider the various vectors of attack, identify most likely vectors, then test them

After you deploy, ensure you have monitors to alert on potential tampering

Microsoft

Thank you!

adversarialml@microsoft.com

PS We're hiring Data Scientists, Researchers, Hunters, Security Engineers – come talk to us!