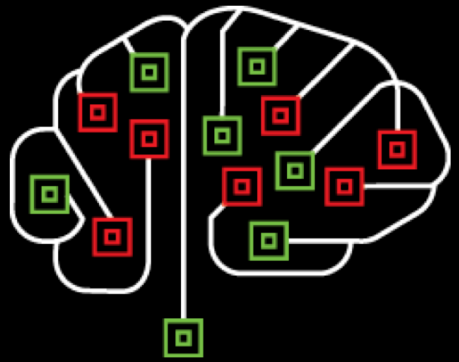
MACHINE LEARNING IN SECURITY

John Brock, Data Scientist

INTRODUCTION

- Data scientist at Cylance
- I build machine learning models to detect malware

WHAT IS MACHINE LEARNING?

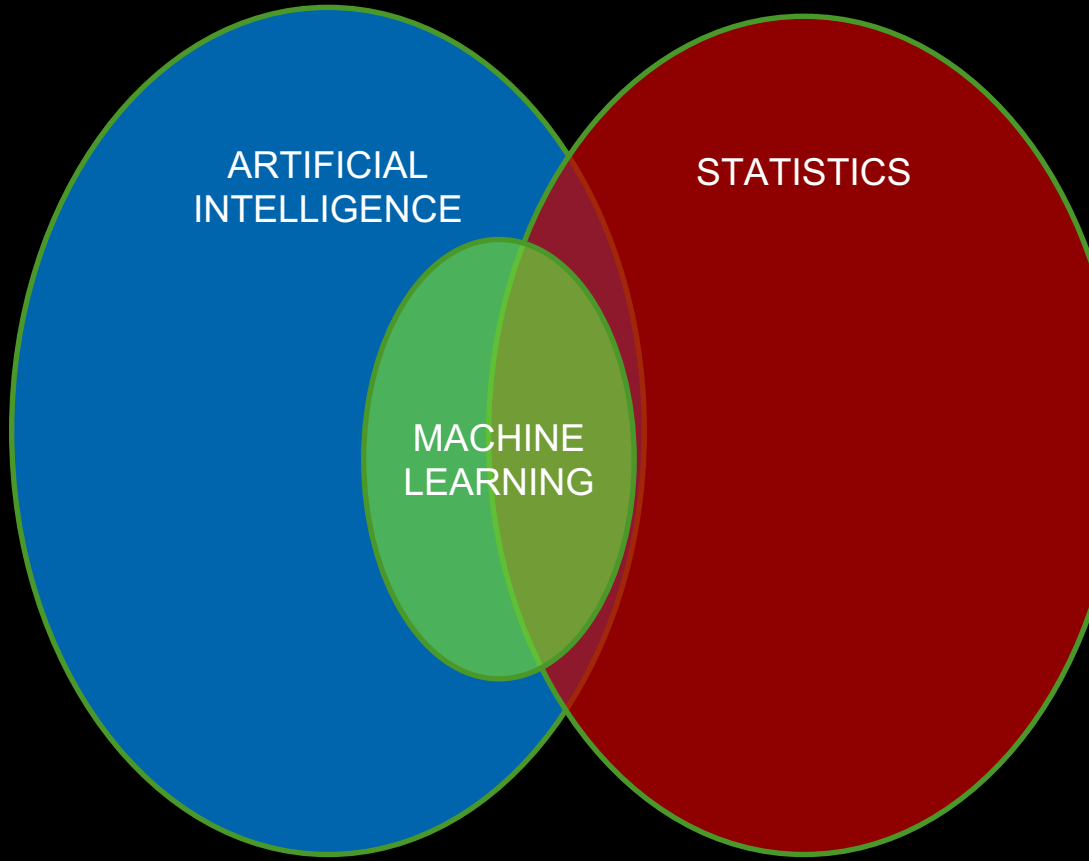


WHAT IS MACHINE LEARNING?

- Subfield of computer science and AI, strongly related to statistics.
- Computers “learn” to perform a task.
- Don’t be told what to do; find out how to do it from the data.
- Not rule-based.
- Good at solving “fuzzy” problems.

AI vs. ML vs. STATS?

Boundaries are fuzzy, but Venn diagram arguably looks like this.



TRADITIONAL AI

1. Human tells AI:

$$penguin \Rightarrow bird$$

$$parrot \Rightarrow bird$$

~~$$hasWings \Rightarrow canFly$$~~

$$bird \Rightarrow hasWings$$

~~$$hasWings \wedge !penguin \Rightarrow canFly$$~~

$$hasWings \wedge !penguin \wedge !(parrot \wedge wingsClipped) \Rightarrow canFly$$

2. AI derives:

~~$$penguin \Rightarrow canFly$$~~

$$parrot \Rightarrow canFly$$

MACHINE LEARNING APPROACH

1. Human shows the computer a bunch of descriptive examples of birds, and whether or not each can fly:

species	Wings clipped?	Can fly?
parrot	no	yes
penguin	no	no
parrot	yes	no
parrot	yes	yes
penguin	no	no
parakeet	no	yes

2. AI learns to predict whether a bird can fly based on its species and whether its wings are clipped.

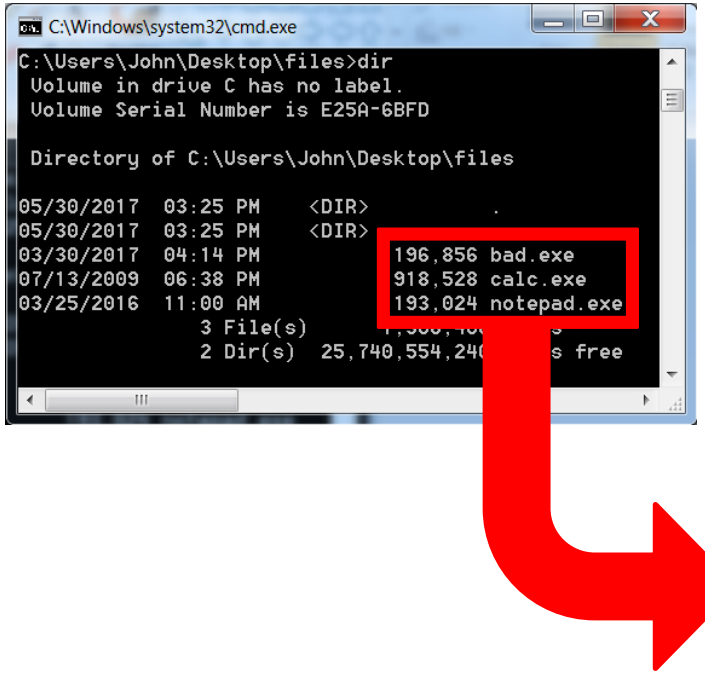
MACHINE LEARNING IN SECURITY

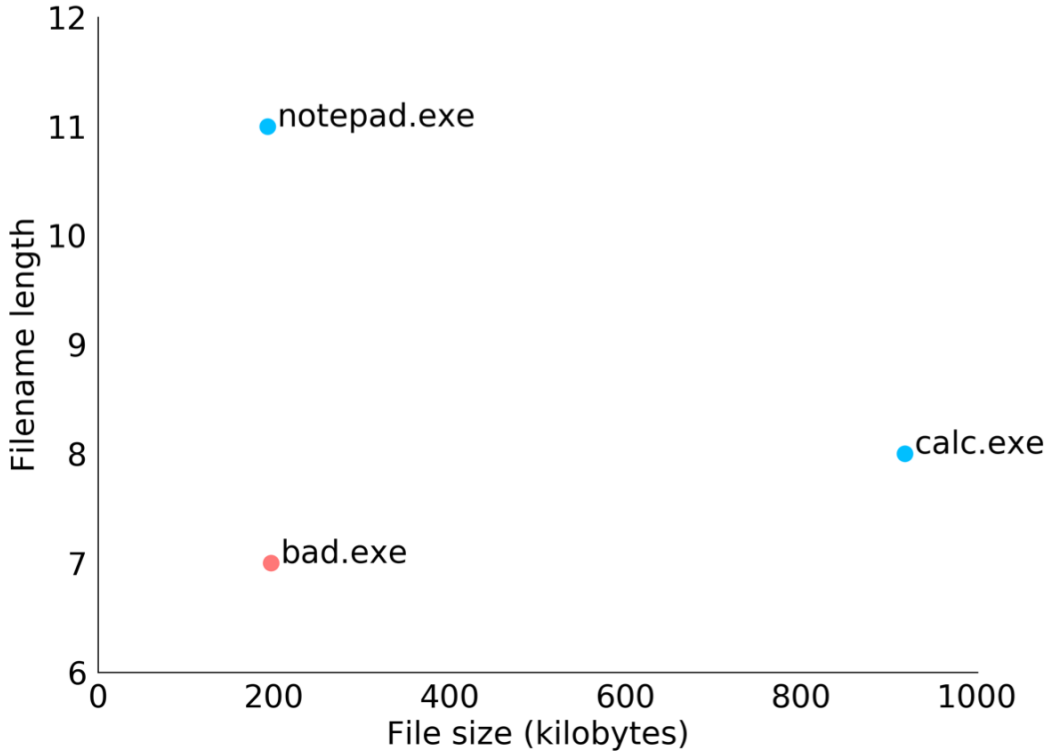
CYBERSECURITY: WHY ML?

- Networks growing in complexity, attack surfaces are increasing, e.g., IoT.
- Not enough experienced security professionals.
- Humans are slow.
- Massive amounts of data
  - Malware
  - Network traffic
  - Process behavior
  - User behavior

MALWARE DETECTION

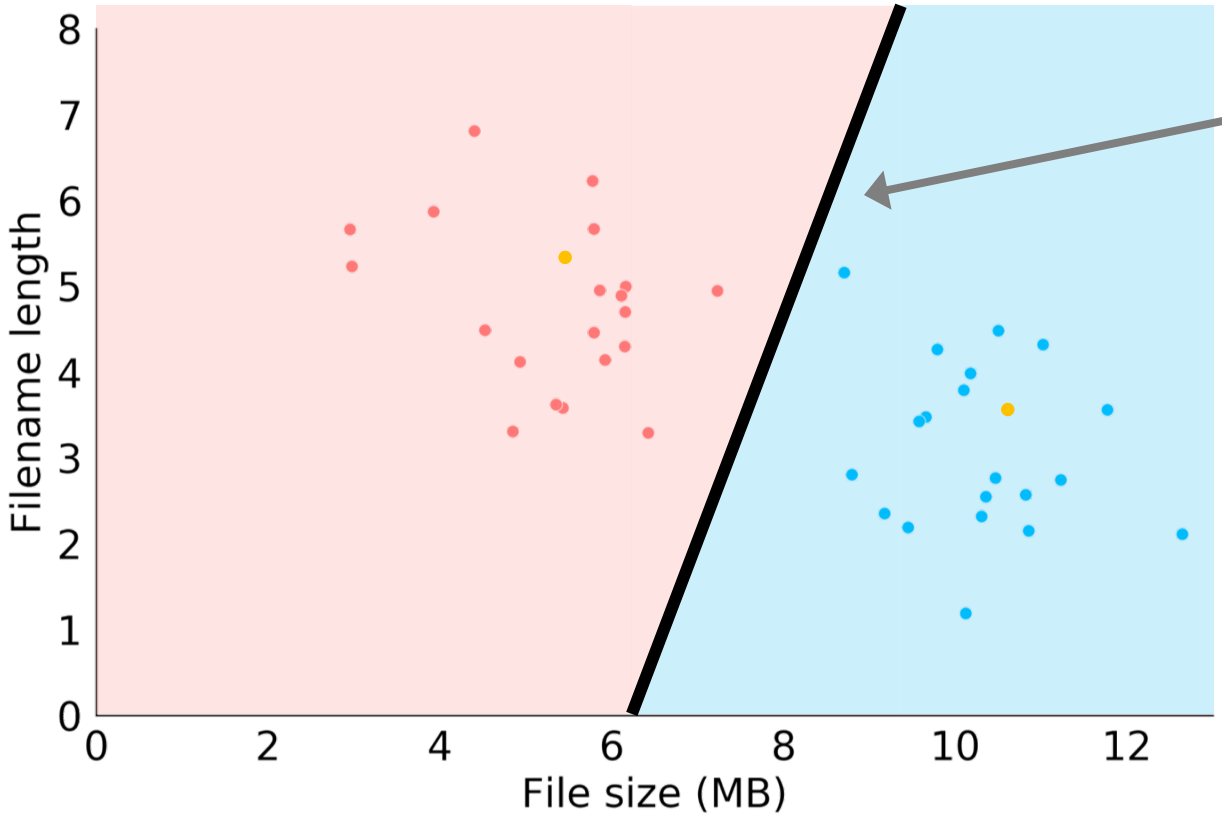
MALWARE DETECTION AS A GEOMETRY PROBLEM





CYLANCE

MALWARE DETECTION AS A GEOMETRY PROBLEM



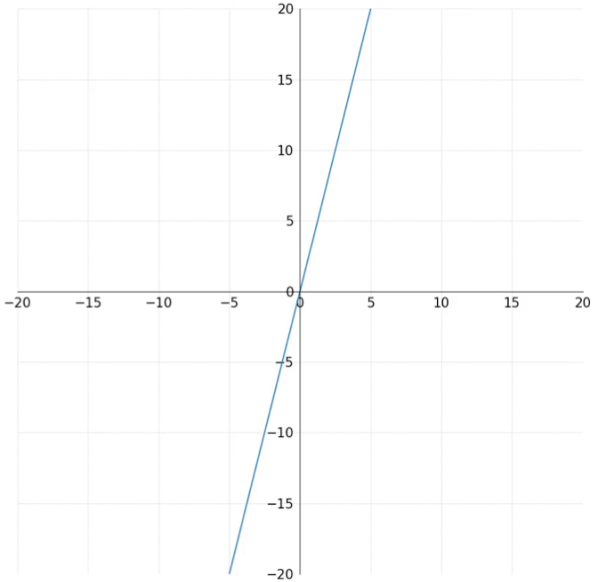
“linear decision boundary”

EQUATION OF A LINE

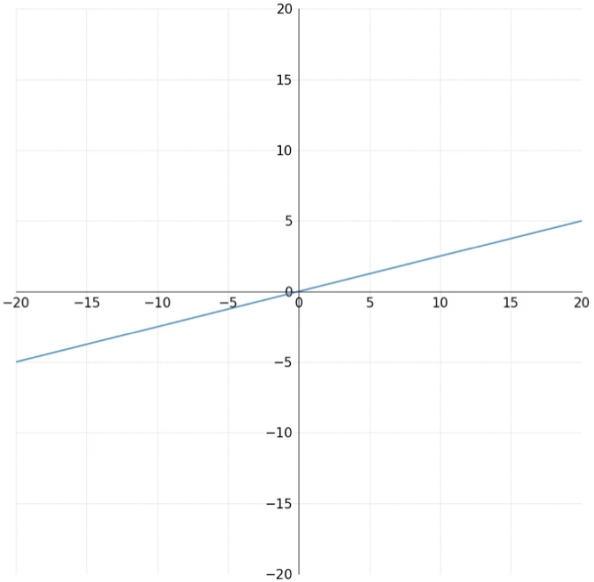
“standard form”: $Ax + By = C$

$Ax + By - C = 0$

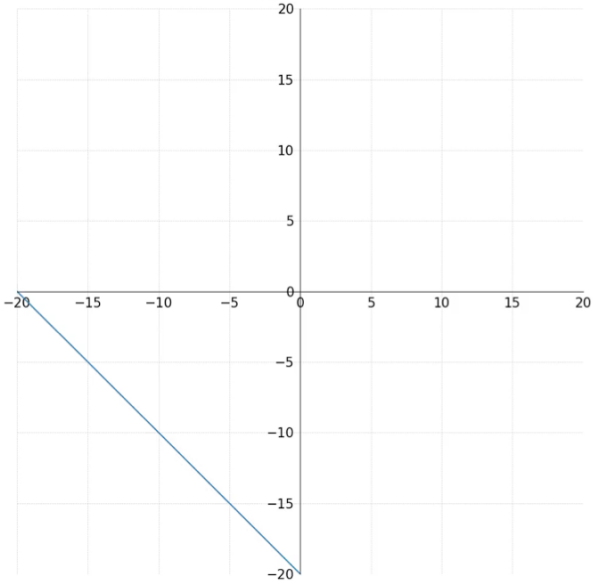
-20.0x + 5y - 0 = 0



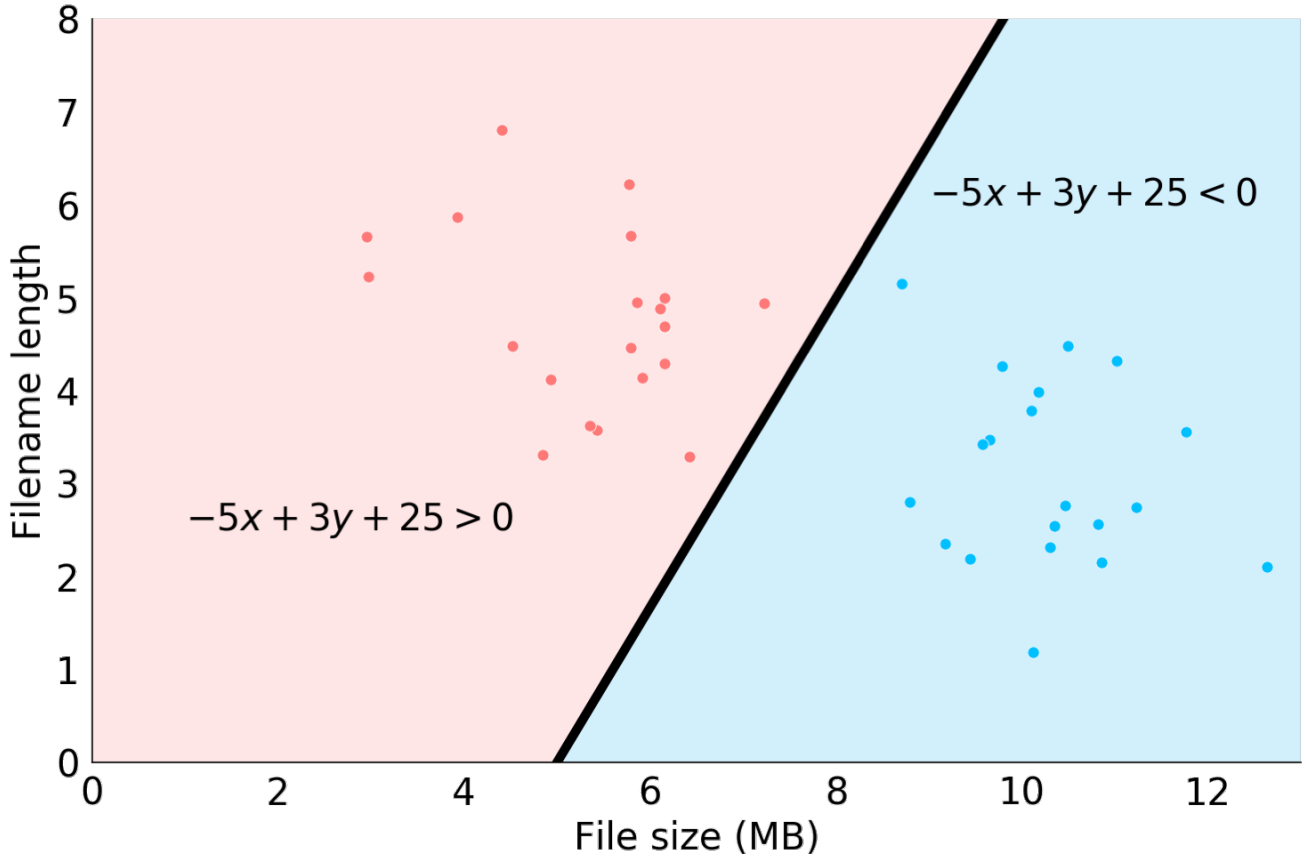
5x - 20.0y - 0 = 0



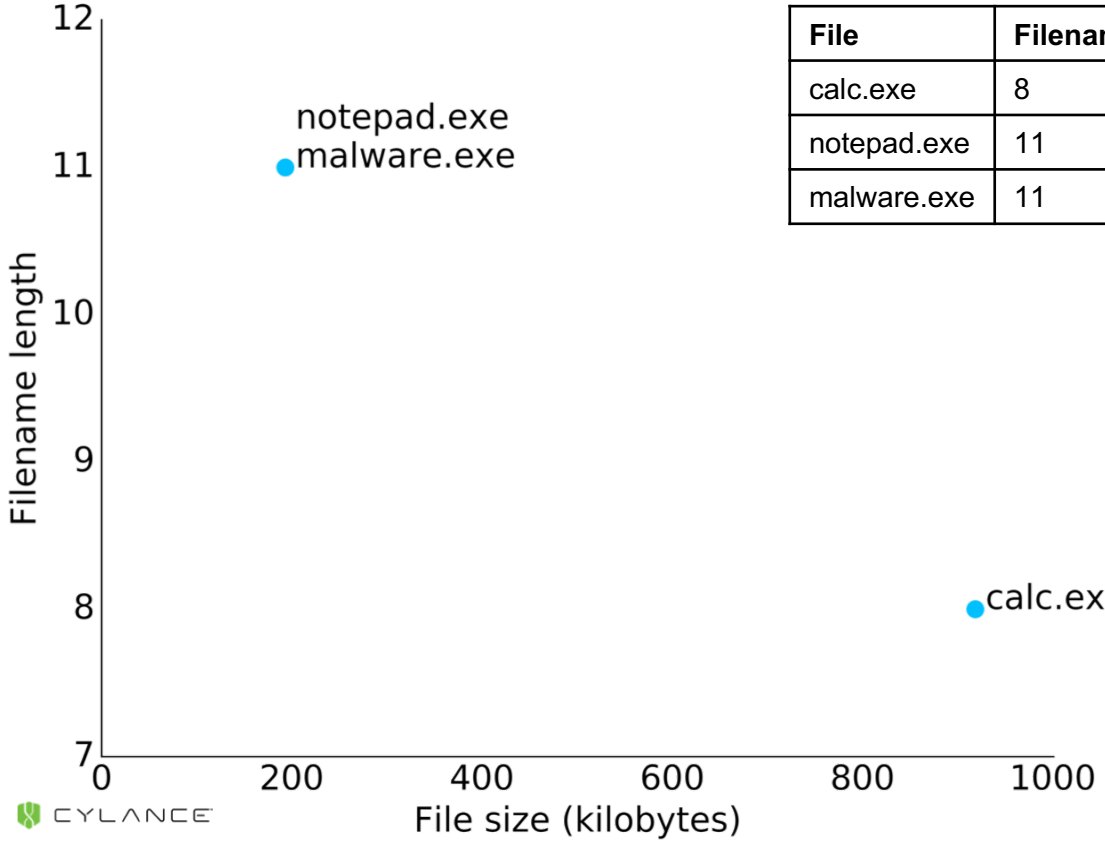
1x + 1y + 20.0 = 0



WHICH SIDE OF THE LINE IS THE POINT ON?



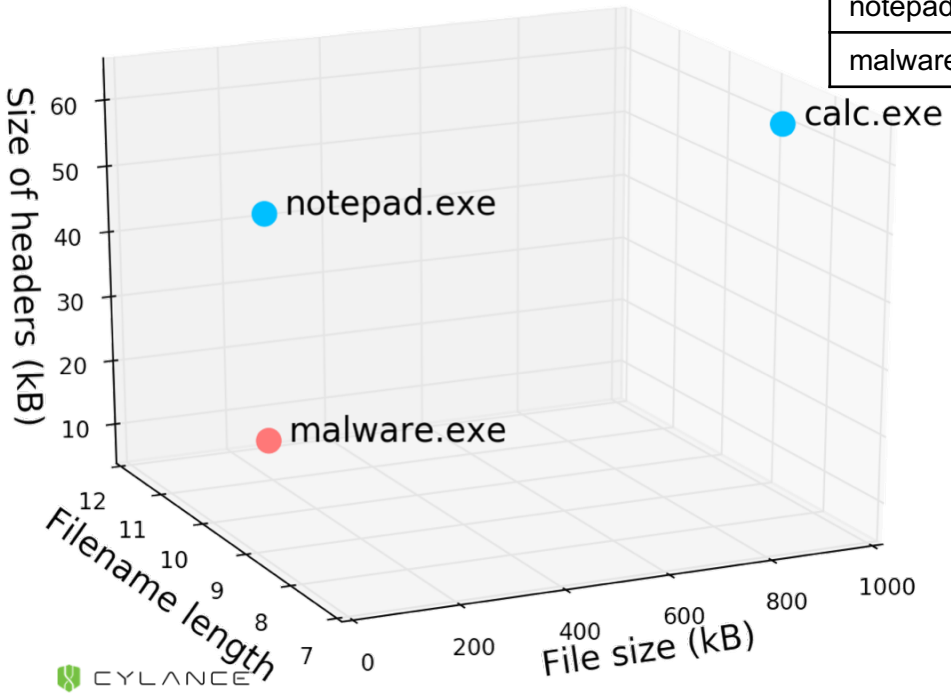
TWO FEATURES PROBABLY AREN'T ENOUGH



File	Filename length	File size (kB)
calc.exe	8	918.528
notepad.exe	11	193.024
malware.exe	11	193.024

ADDITIONAL FEATURES CAN HELP

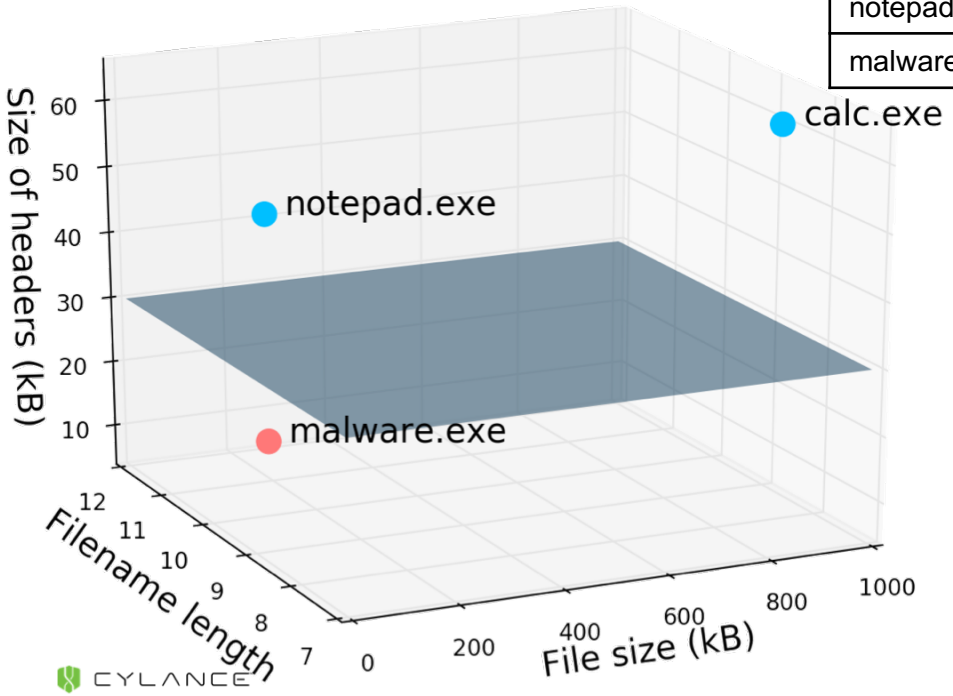
File	Filename length	File size (kB)	Size of headers (kB)
calc.exe	8	918.528	63
notepad.exe	11	193.024	45
malware.exe	11	193.024	10



CYLANCE

ADDITIONAL FEATURES CAN HELP

File	Filename length	File size (kB)	Size of headers (kB)
calc.exe	8	918.528	63
notepad.exe	11	193.024	45
malware.exe	11	193.024	10

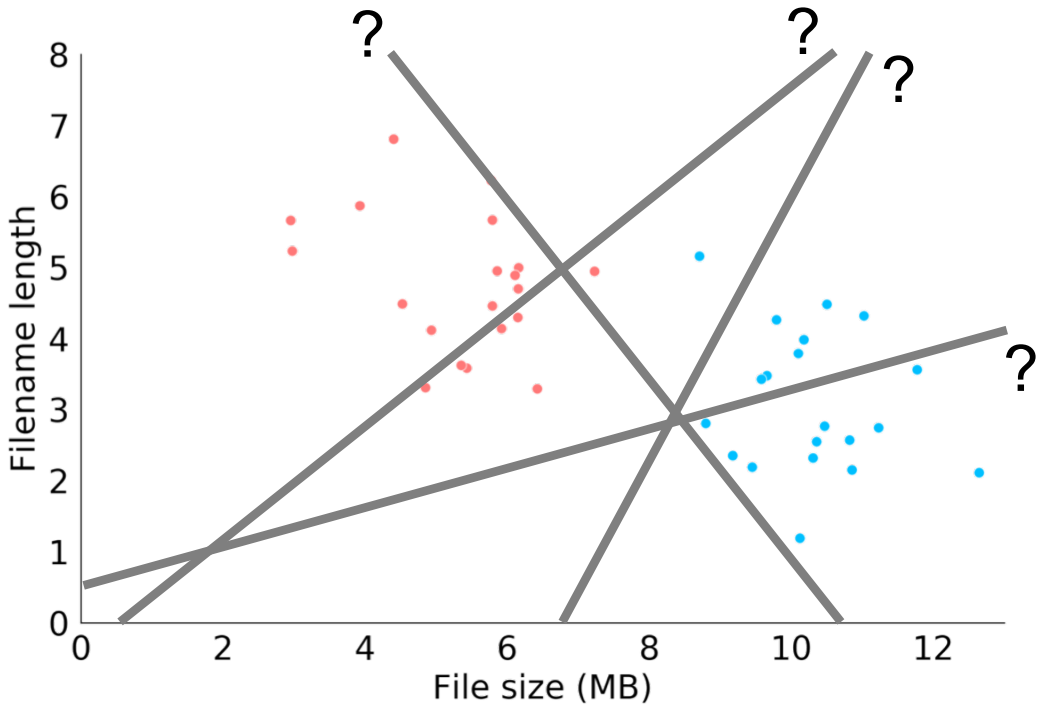


$$Ax + By + Cz - D = 0$$

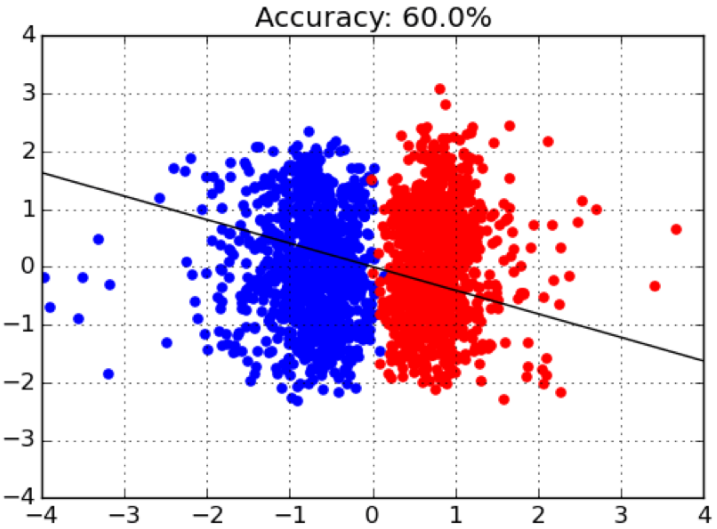
AUTOMATICALLY FINDING THE DECISION BOUNDARY

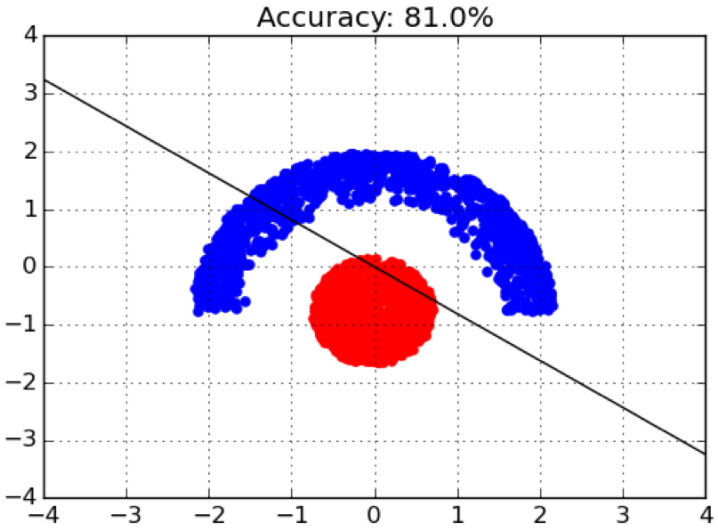
$Ax + By - C = 0$

A = ?
B = ?
C = ?



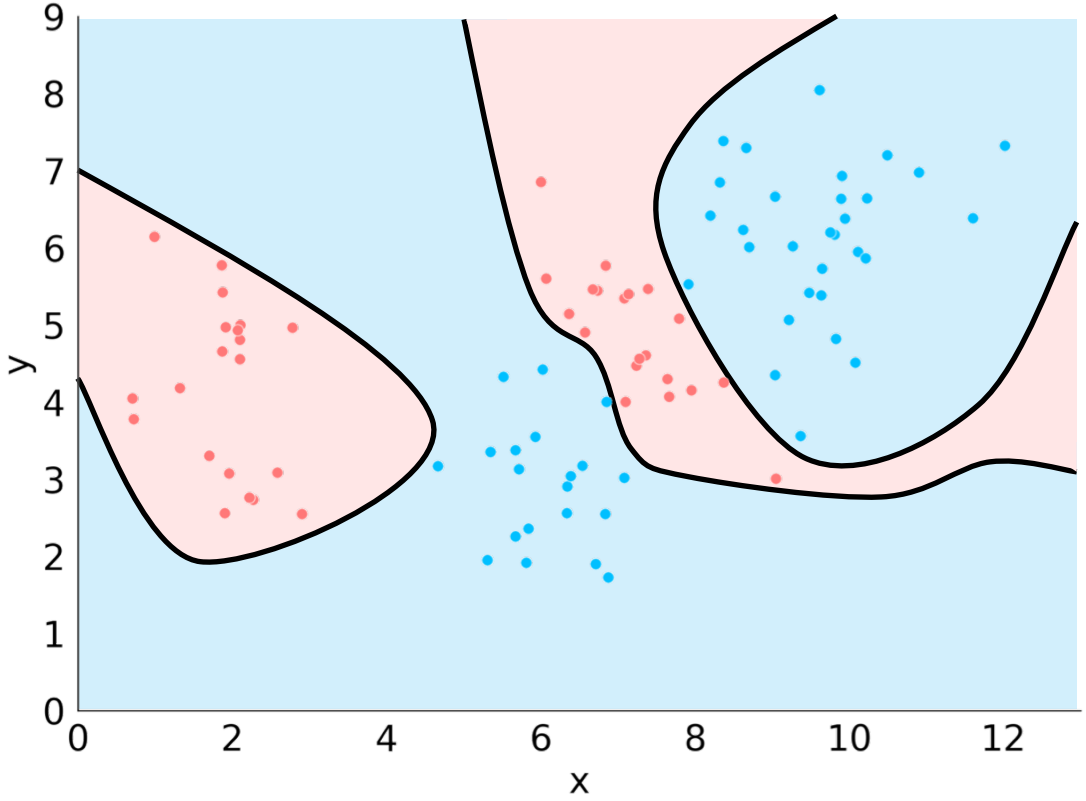
LINEAR CLASSIFIER LIMITATIONS





Credit: Andrew Davis

NON-LINEAR DECISION BOUNDARIES



MALWARE DETECTION

Static features - extract “static” information about the file

- File size?
- Strings in the file?
- File header metadata?
- What DLLs does it use? What functions from each DLL does it call?
- Are parts of it encrypted? Compressed?

Dynamic features - watch the program as it executes

- What API functions does the sample call?
- Is it talking to the network?
- Is it reading/writing/removing files or registry entries?
- Is it opening crypto libraries?

MALWARE DETECTION

Static Analysis:

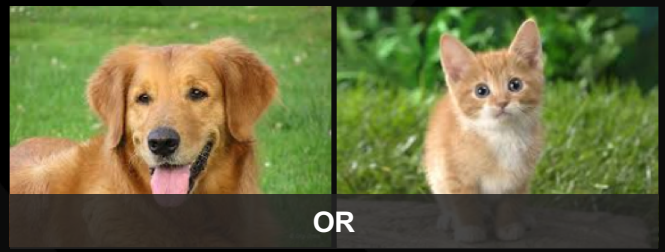
- No execution, no damage (unless a false negative)
- Very fast
- Easy to miss obscured information (compressed/encrypted information)

Dynamic analysis:

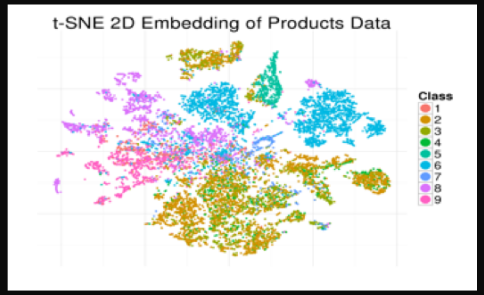
- Executes, system may be compromised by the time ML detects
- Slower - the sample must run
- Malicious samples can “run out the clock” on dynamic analysis
- Samples can “short-circuit” if they detect they are being watched

FILE SIMILARITY

THE TWO TYPES OF ML



VS.



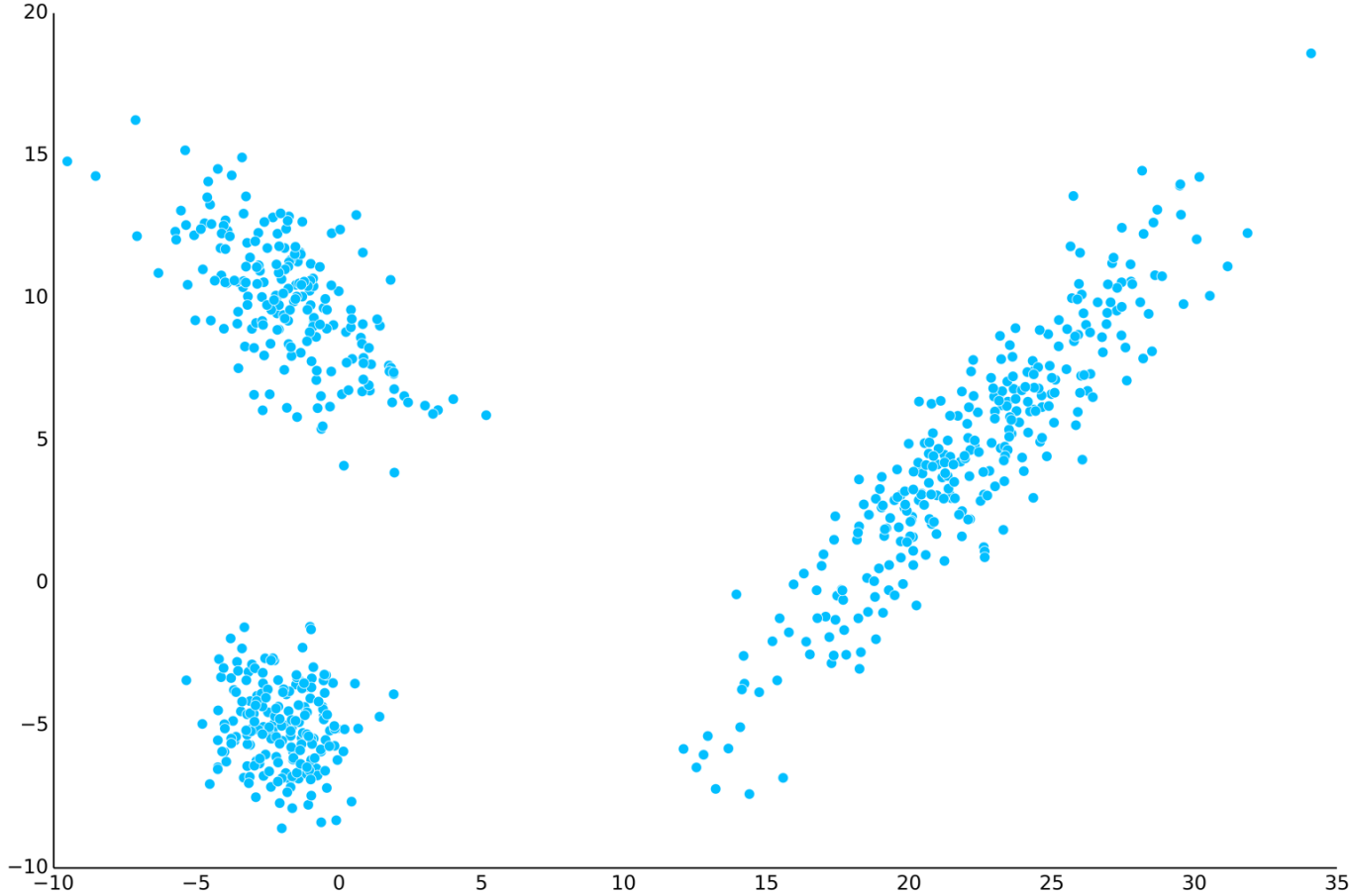
Supervised ML

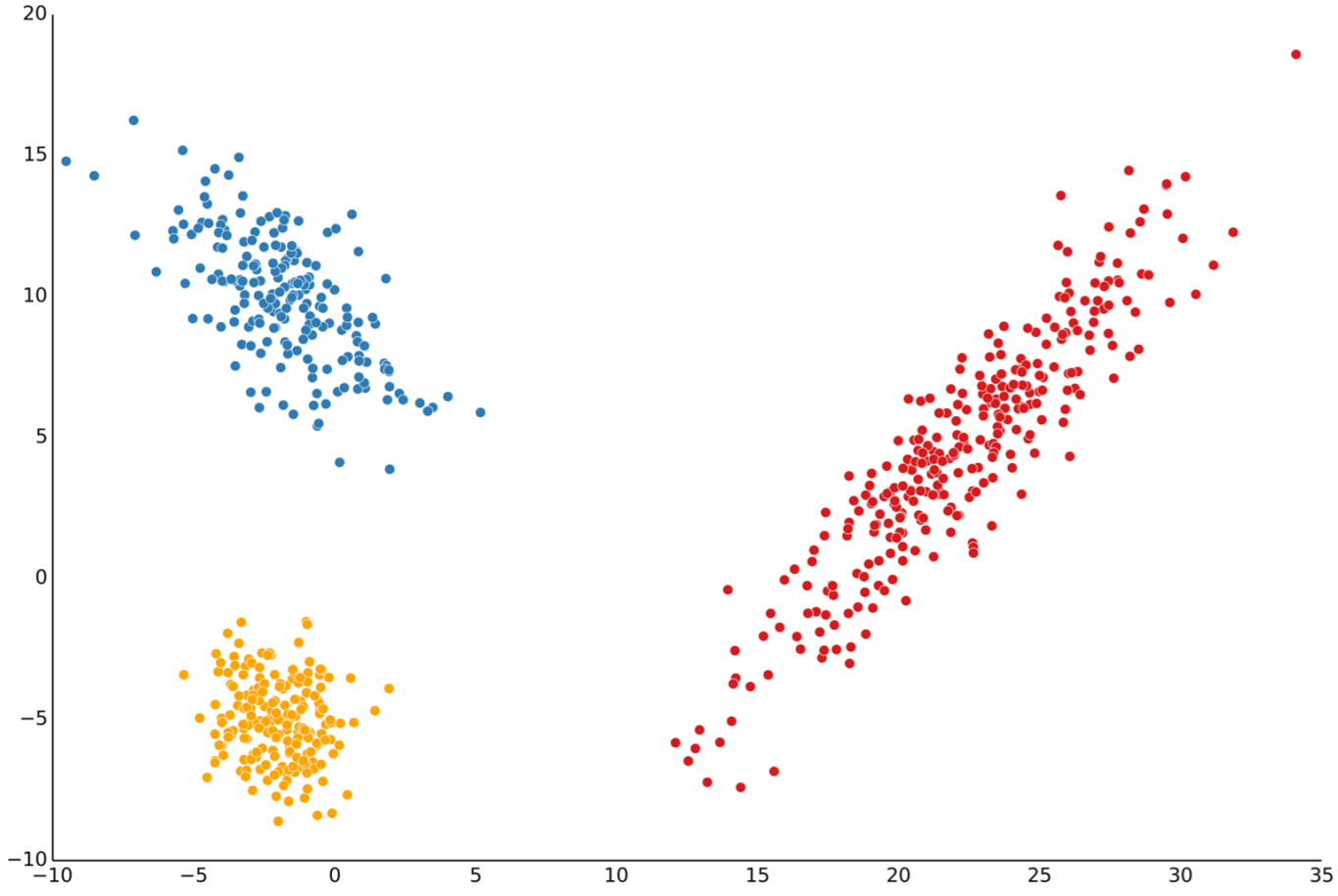
- Used when ground-truth information is plentiful
- Used to classify different samples into different distinct categories
- Judging the quality of the classifier is easy -- what is the accuracy?

Unsupervised ML

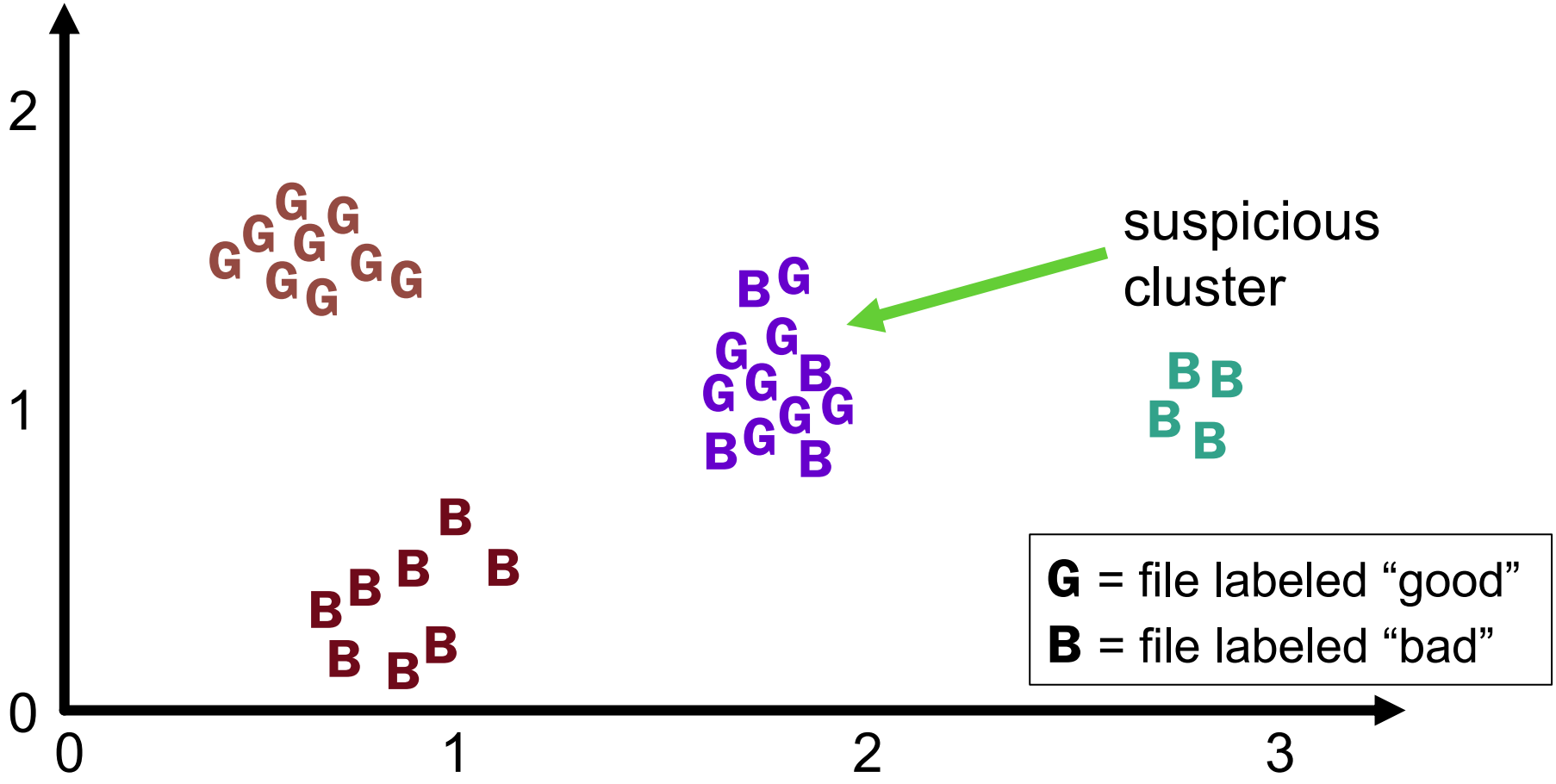
- Use with little/no ground truth
- How similar are samples are to each other?
- More careful and qualitative human analysis to check the resulting model

*Thanks to Ben Hammer at Kaggle


20
15
10
5
0
−5
−10
−10
−5
0
5
10
15
20
25
30
35


20
15
10
5
0
−5
−10
−10
−5
0
5
10
15
20
25
30
35

Finding Mislabeled Files



ANOMALY DETECTION

ANOMALY DETECTION

- How to tell if something anomalous is in your network?
  - Data exfiltration
  - Atypical logins
- Collecting large amounts of data is easy
- Observing anomalous events is rare, so datasets are relatively small
- How to evaluate?
- Best fit: unsupervised learning

CHALLENGES

CYBERSECURITY & ML: CHALLENGES

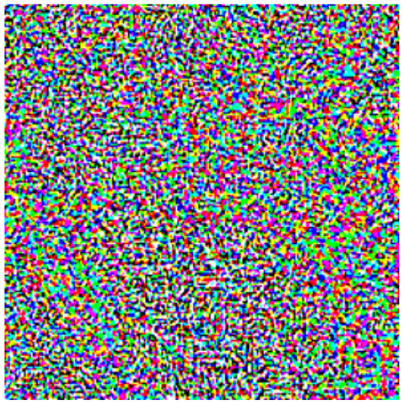
- Cybersecurity data is often:
  - Big (terabytes to petabytes).
  - Extremely high-dimensional (thousands to millions).
  - Extremely sparse.
  - Features are typically discrete or mix of discrete and continuous.
- Domain is highly adversarial.
- Human labeling is time-consuming.
- Interpretability is often highly valued.
- Maintaining privacy is crucial.

CYBERSECURITY & ML: ADVERSARIAL ATTACKS

$\boldsymbol{x}$

"panda"
57.7% confidence

$+ .007 \times$



$\text{sign}(\nabla_{\boldsymbol{x}} J(\boldsymbol{\theta}, \boldsymbol{x}, y))$

"nematode"
8.2% confidence

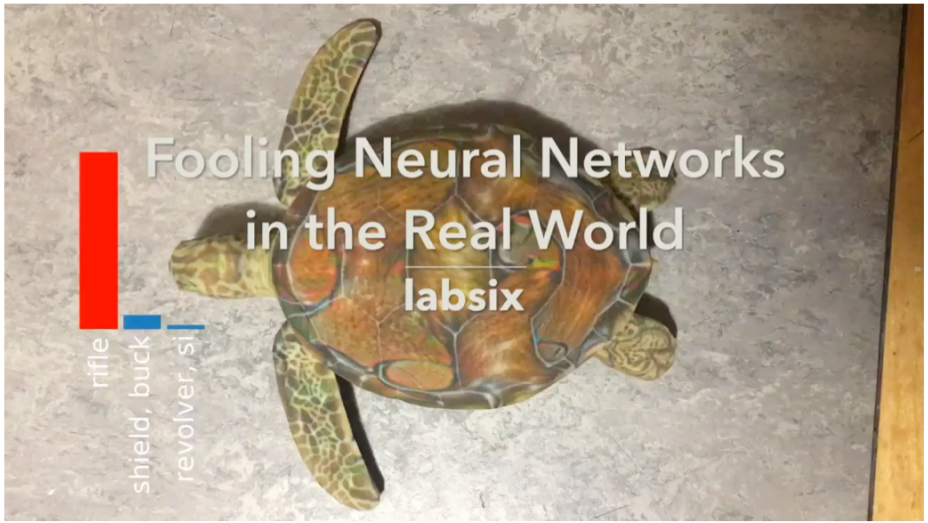
$=$

$\boldsymbol{x} + \epsilon\text{sign}(\nabla_{\boldsymbol{x}} J(\boldsymbol{\theta}, \boldsymbol{x}, y))$

"gibbon"
99.3 % confidence

Goodfellow, et al., Ian J., Jonathon Shlens, and Christian Szegedy. "Explaining and harnessing adversarial examples." https://arxiv.org/abs/1412.6572

CYBERSECURITY & ML: ADVERSARIAL ATTACKS



Anish Athalye, Logan Engstrom, Andrew Ilyas, and Kevin Kwok
http://www.labsix.org/physical-objects-that-fool-neural-nets/

CYBERSECURITY & ML: CHALLENGES

- Cybersecurity data is often:
  - Big (terabytes to petabytes).
  - Extremely high-dimensional (thousands to millions).
  - Extremely sparse.
  - Features are typically discrete or mix of discrete and continuous.
- Domain is highly adversarial.
- Human labeling is time-consuming.
- Interpretability is often highly valued.
- Maintaining privacy is crucial.

FURTHER READING

- “Introduction to Artificial Intelligence for Security Professionals” by Cylance Inc. Free PDF at: http://defense.ballastsecurity.net/static/IntroductionToArtificialIntelligenceForSecurityProfessionals_Cylance.pdf
- MLSec Project https://www.mlsecproject.org/
- Long list of resources: https://github.com/wtsxDev/Machine-Learning-for-Cyber-Security
- For background on computer security: “Hacking Exposed 7: Network Security Secrets and Solutions” by Stuart McClure, Joel Scambray, and George Kurtz

QUESTIONS

AND

ANSWERS