black hat USA 2016

APPLIED MACHINE LEARNING FOR DATA EXFIL AND OTHER FUN TOPICS

Matt Wolff, Chief Data Scientist
Brian Wallace, Security Researcher
Xuan Zhao, Data Scientist

CYLANCE

GOALS OF THIS TALK:
APPLIED MACHINE LEARNING

- Identify suitable problems for ML approaches
- Demonstrate by example how to apply ML
- Help jumpstart additional research in the Security + ML space

WHO WE ARE/CYLANCE

- Endpoint security company built around the capabilities of artificial intelligence
- Protecting millions of enterprise endpoints
- Founded in 2012, $177 mm raised
- Booth #1124

Matt Wolff
Chief Data Scientist

Brian Wallace
Senior Security Researcher

Xuan Zhao
Data Scientist

TALK OVERVIEW

- Machine Learning Introduction
- NMAP Clustering
  - Feature Spaces
  - Distances
  - Clustering
- Botnet Panel Identification
  - Classification
  - Feature Reduction
- Obfuscating Data with Markov Chains

MACHINE LEARNING OVERVIEW

- Machine learning techniques are data driven
- Available data should be able to solve the problem in a meaningful way
- Approaches exist for dealing with raw data, as well as labeled or annotated data

MACHINE LEARNING OVERVIEW

- Given some data, different types of machine learning can be applied
- Clustering is useful for finding similarity across dataset to uncover trends or other insight
- With labeled data, classification can be useful to build predictive models

MACHINE LEARNING OVERVIEW

- Often, raw data has to be transformed in some way to be used by machine learning algorithms
- Typical process is to extract features from data, and turn those features into vectors
- Vectors are then fed into ML algorithms for training or other purposes

MACHINE LEARNING OVERVIEW

- Recommended resources
  - Scikit-learn.org
  - Python
- Source code for all tools in this talk available on the Cylance public git repo
  - https://github.com/CylanceSPEAR
- Should be able to pull talk source and start modifying as needed to suit data driven problems in your own organization or research group.

TOOL – NMAP CLUSTERING

- NMAP is a popular port scanning tool
- Produces large amount of data per IP address
- Scans over large number of IPs can be difficult to make sense of
- NMAP Clustering is a tool which clusters (groups) IPs based on their open ports, services, etc

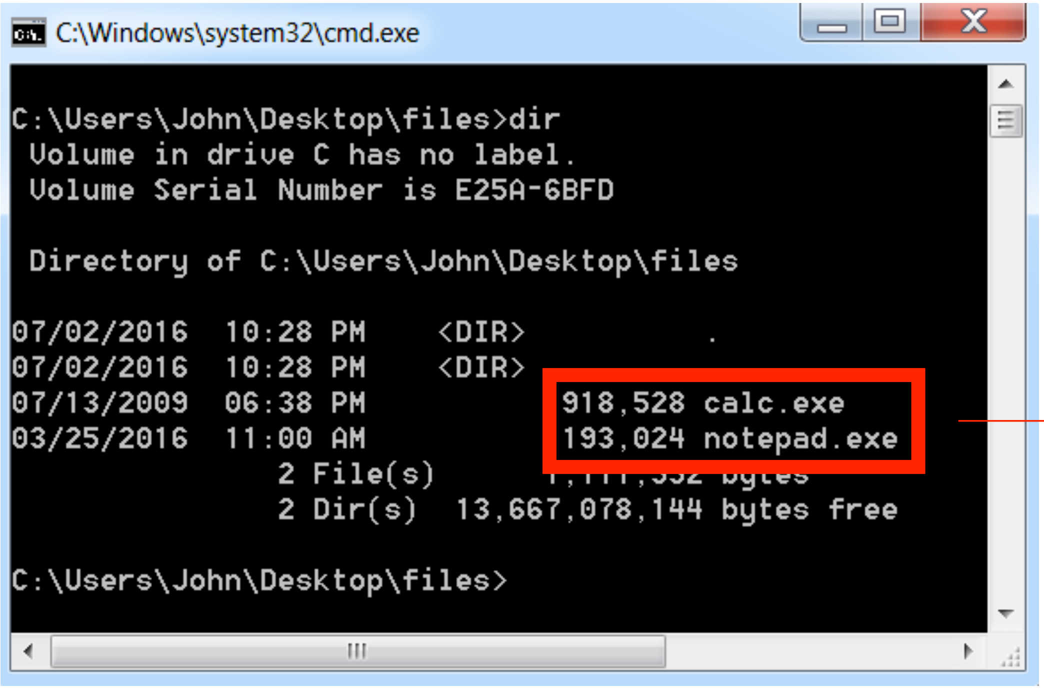
FEATURES

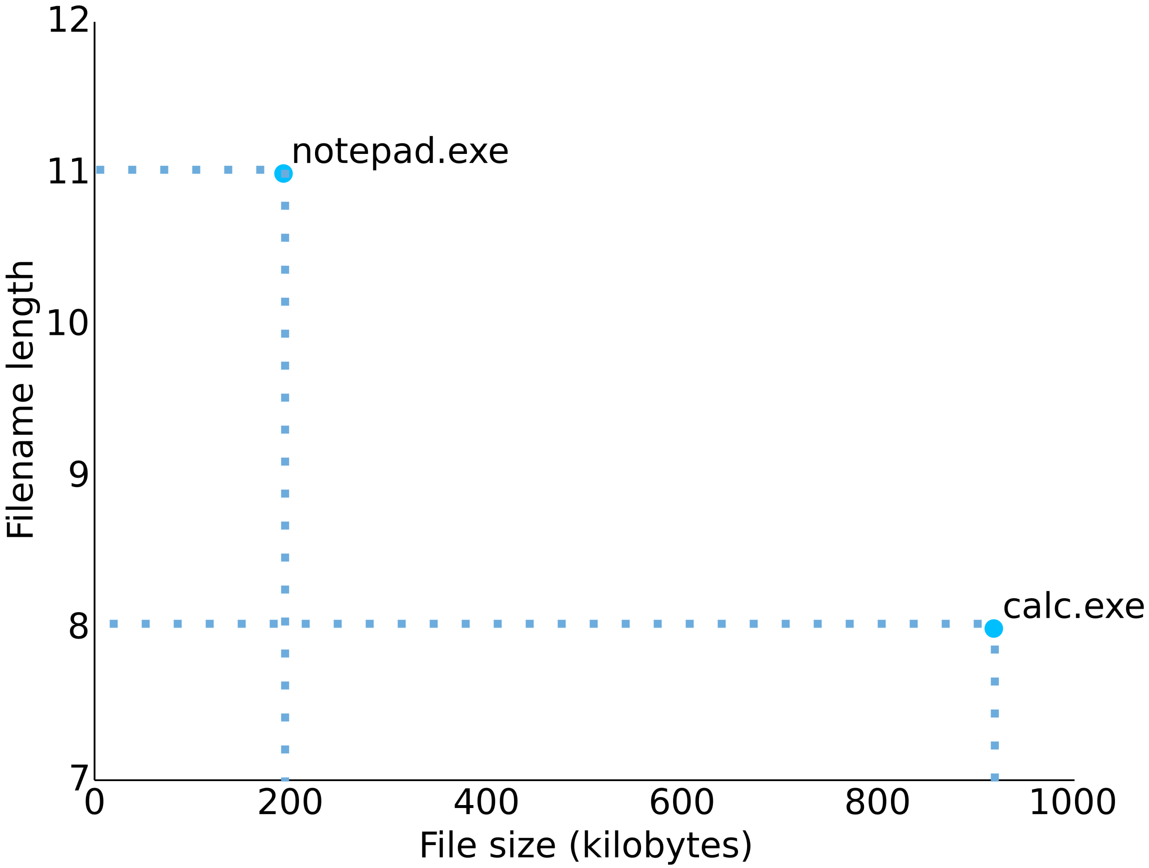
- Features are informative, discriminative information that can describe a sample/observation/phenomenon/etc.
- Feature extraction is pivotal to the machine learning pipeline
- Our features are based on NMAP output
- Each port is a feature, each service on each port is a feature, each version of each service on each port is a feature, etc
- Script output included in features (including website titles, public keys, etc)

VECTORS

- Numerical representation of a sample (IP in NMAP case)
- Array of values which represent all features from one sample
- Vectors can be thought of as points in high dimensional space
- Each feature is a dimension, the value of the feature in the vector is the position in that dimension
- If we have only two features, it is really easy to visualize

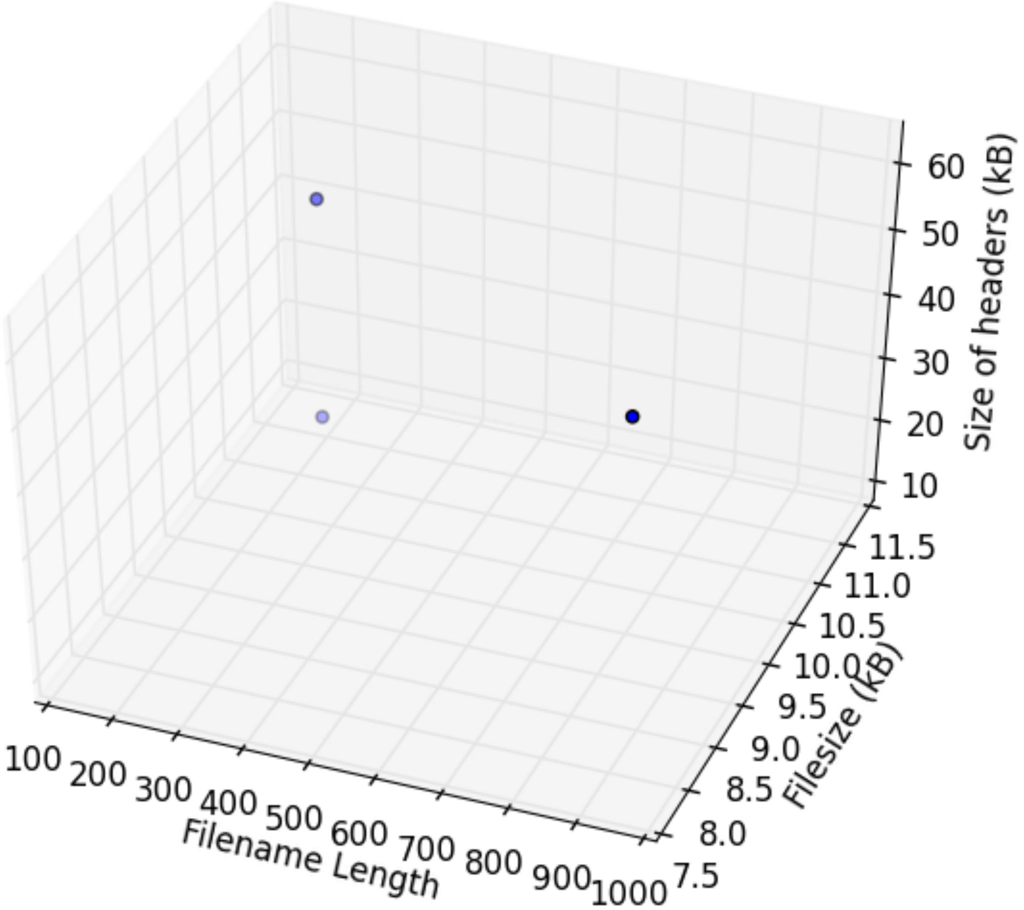
VECTORS – 2D


C:\Windows\system32\cmd.exe
C:\Users\John\Desktop\files>dir
Volume in drive C has no label.
Volume Serial Number is E25A-6BFD
Directory of C:\Users\John\Desktop\files
07/02/2016 10:28 PM <DIR> .
07/02/2016 10:28 PM <DIR>
07/13/2009 06:38 PM 918,528 calc.exe
03/25/2016 11:00 AM 193,024 notepad.exe
2 File(s)
2 Dir(s) 13,667,078,144 bytes free
C:\Users\John\Desktop\files>

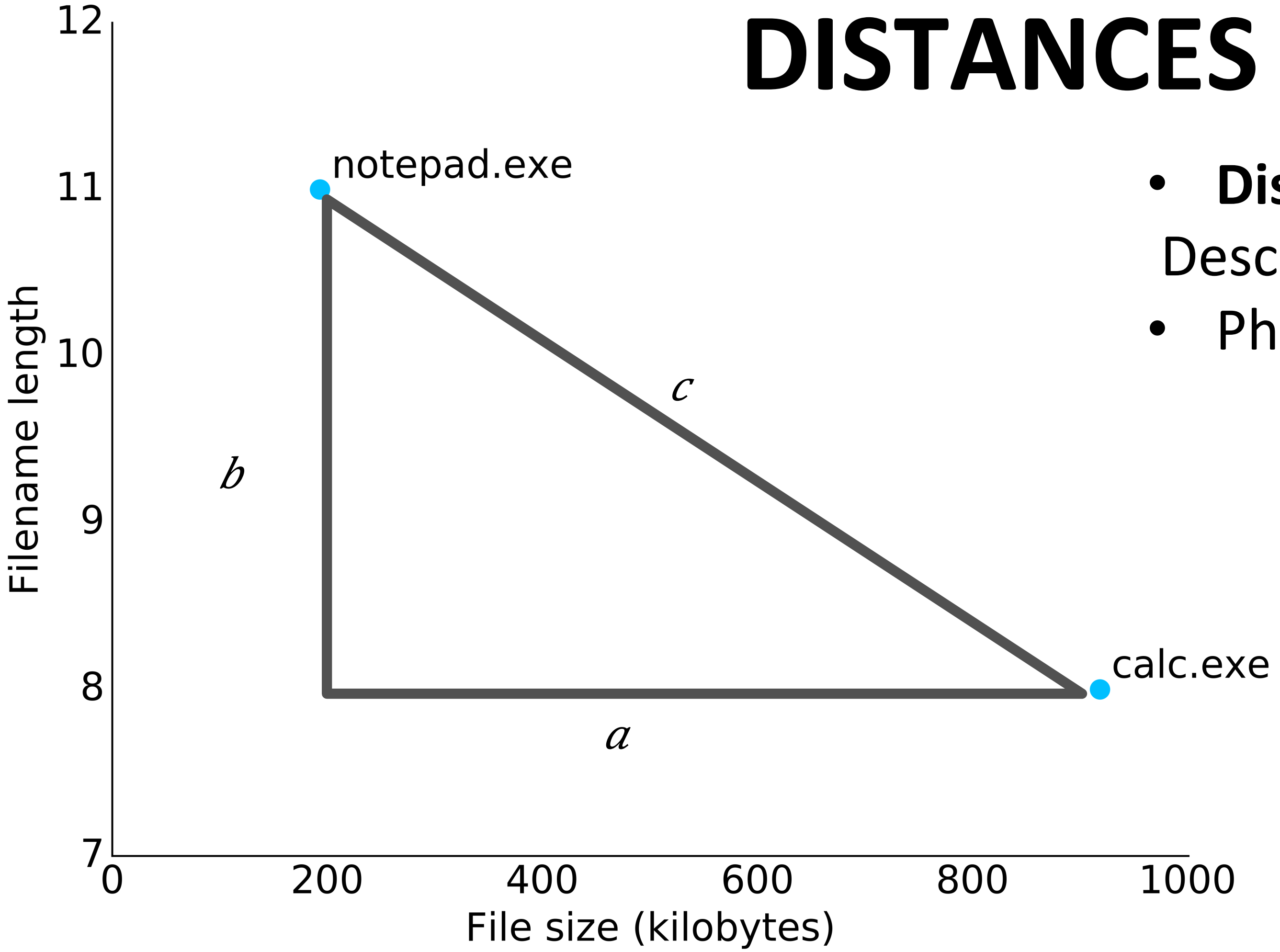

notepad.exe
calc.exe
Filename length
12
11
10
9
8
7
0
200
400
600
800
1000
File size (kilobytes)

VECTORS – 3D

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



DISTANCES



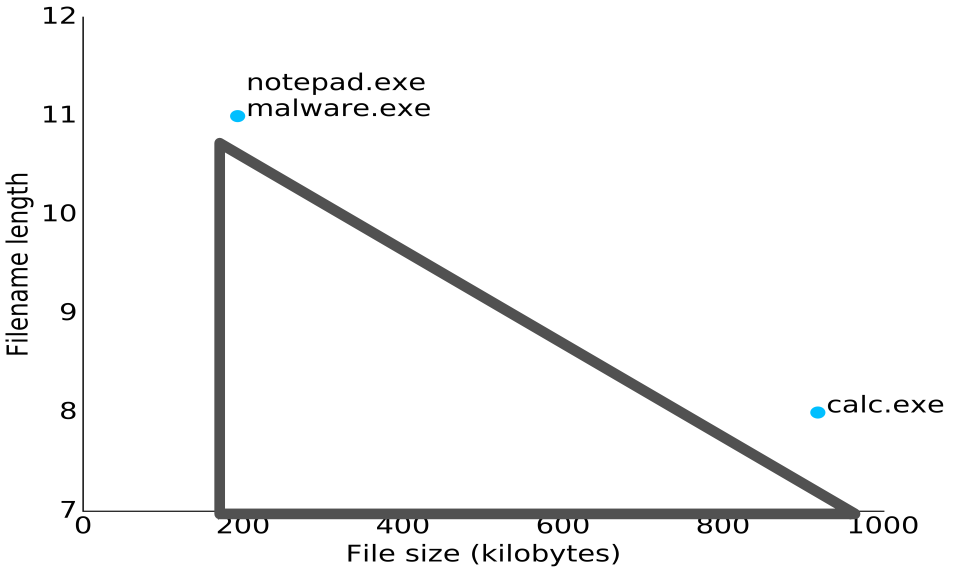
- **Distance:**
  Describe the discrepancy between two points
- Physical distance between two points:

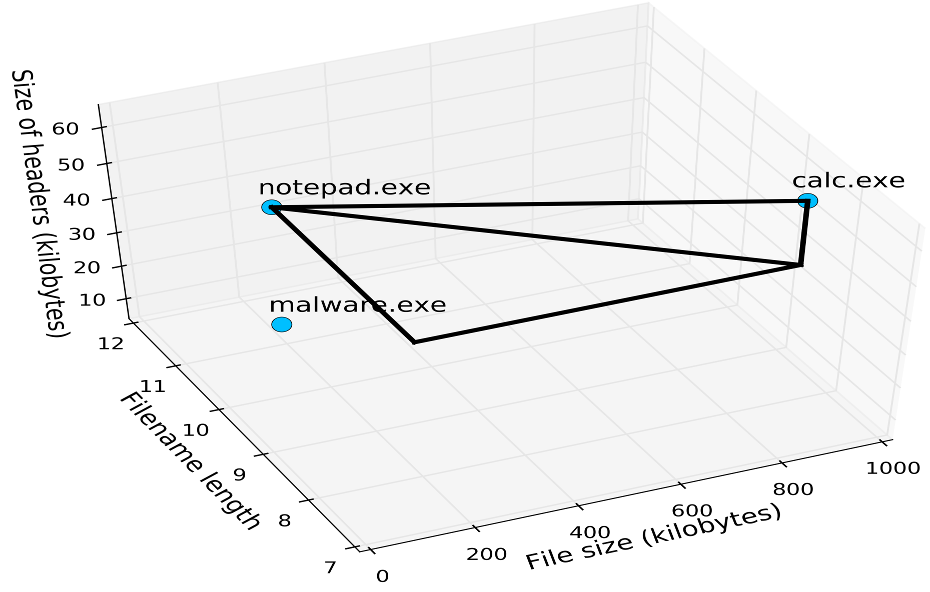
Pythagorean's theorem:

$$a^2 + b^2 = c^2$$

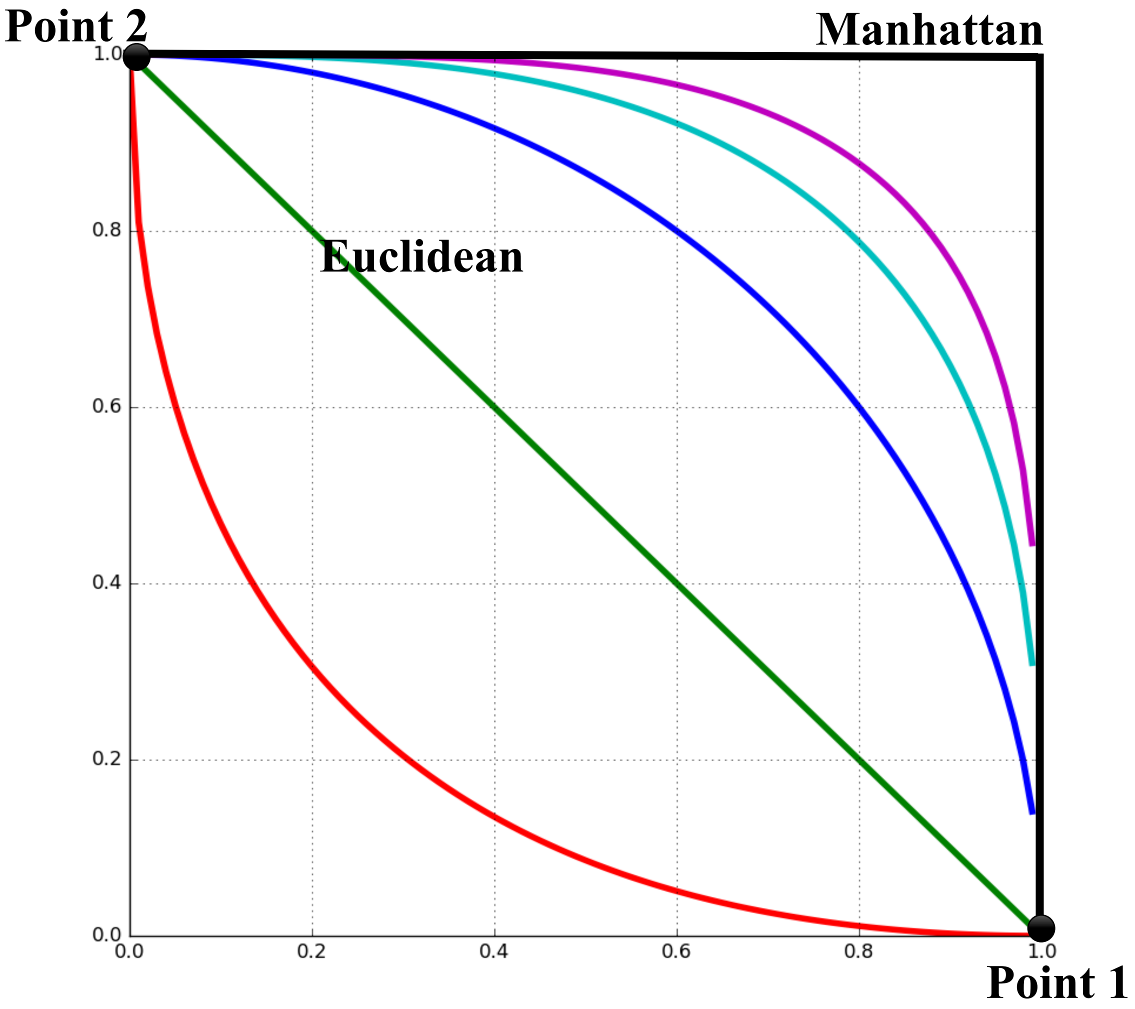
File	Filename Length	Filesize (kB)	Size of headers (kb)
calc.exe	8	918.528	63
notepad.exe	11	193.024	45
malware.exe	11	193.024	10

2D





3D



DISTANCES

Multiple Distance Metrics –

As long as an operation satisfy certain mathematical criteria, it can be used as a distance metric

- Euclidean Distance: $\sqrt{(a^2 + b^2)}$
- Manhatttan Distance: $|a| + |b|$
- Other Distances
  - L_p Norms: $(a^p + b^p)^{1/p}$
  - cosine Distance: $\frac{\mathbf{a} \cdot \mathbf{b}}{\|\mathbf{a}\| \|\mathbf{b}\|}$

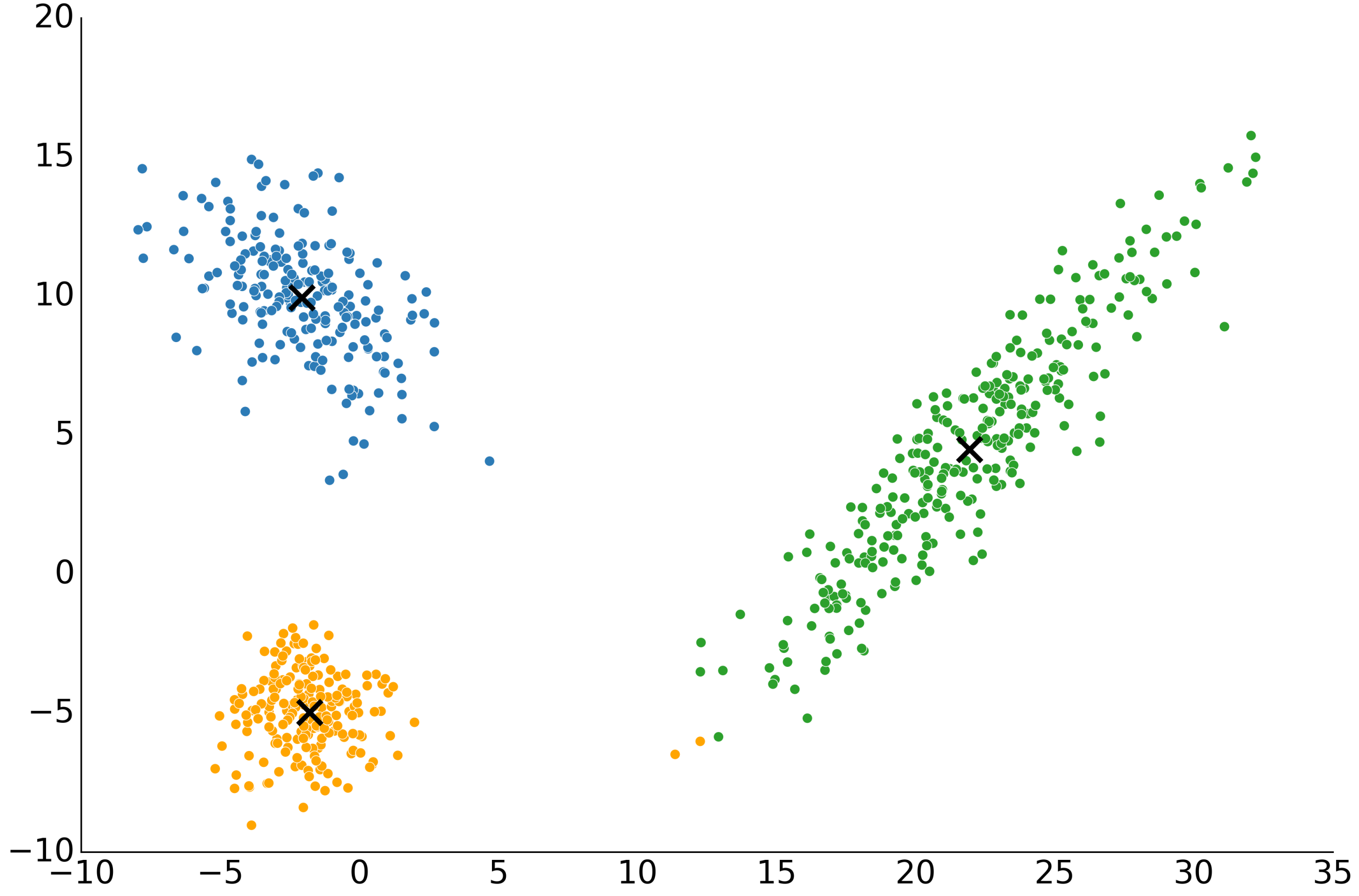
CLUSTERING

- With a way to measure distance, we can group items by how close they are, aka clustering
- Clusters are distinct groups of samples (IP) which have been grouped together
- Clustering is generally unsupervised learning
- Different algorithms with different configurations group these samples in different ways

k-Means

- Clustering algorithm
- User supplies *k* (destinated number of clusters)
- All samples are assigned to random clusters
- Center of each cluster is computed by taking mean (average) of all samples in cluster
- Samples are then assigned to the cluster whose center they are closest to
- Centers are recomputed, algorithm loops until no samples change clusters

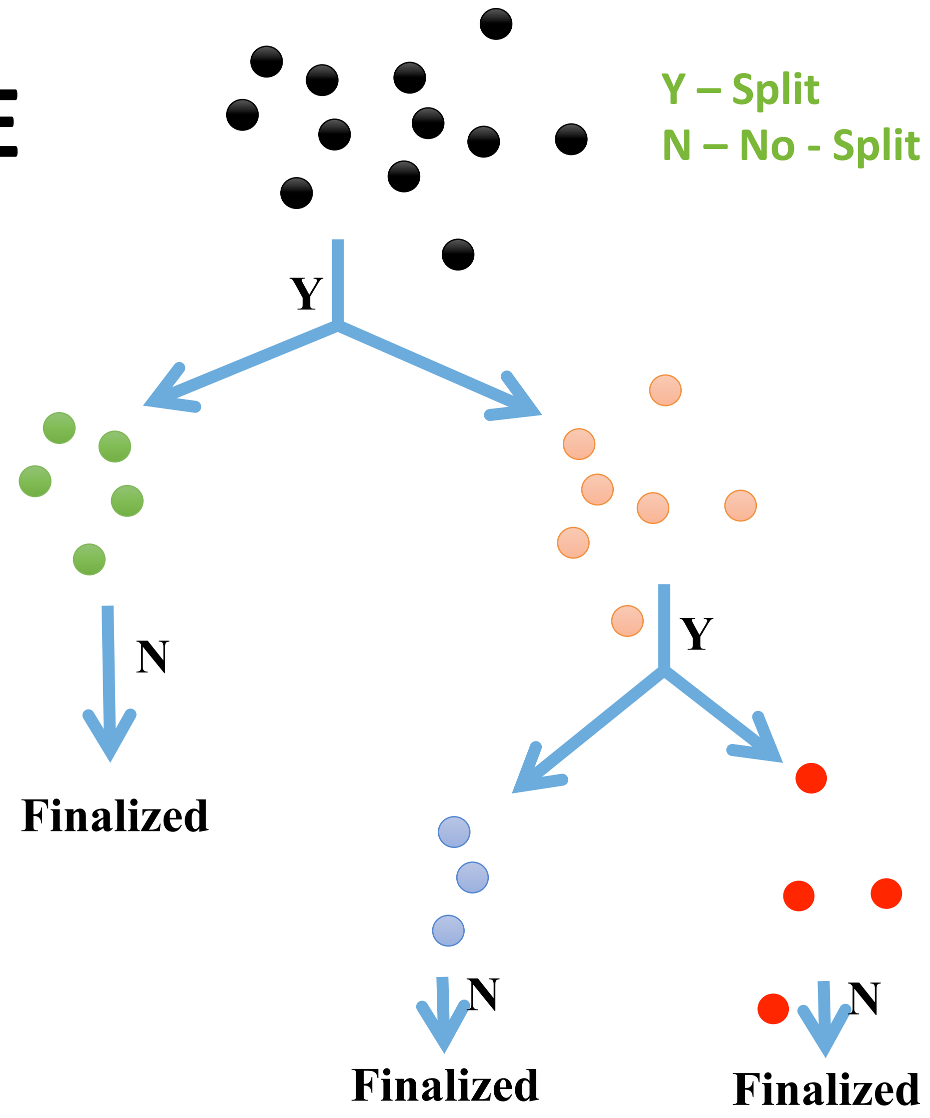
k-Means



NMAP CLUSTERING – MANUAL/AUTOMATIC

- Manual allows you to supply your own clustering parameters
- Automatic tries many different methods with theoretically-found optimal parameters and picks what it determines to be the best
- Demo with manual strategy
- Demo with automatic strategy

NMAP CLUSTERING - INTERACTIVE



- Incorporate the User's decision into the clustering process.
- The Clustering result will be customized according to customer's preference in this way
- Process (will show with a demo):
    1. User decide whether the cluster needs to be split or not:
    2. If yes, then split using divisive clustering
    3. If no, finalize this cluster
    4. Recursively split until users are satisfied with all the clusters

TOOL – ID PANEL

- Botnet panels (command and control websites) can be difficult to identify
  - Need previous knowledge of the botnet panels
  - Often modified to avoid detection or vanity
  - Many are based off others, so distinguishing can be difficult
- We can train a model to identify if we are looking at a bot panel and which one it is, with a small number of requests
- Minimizing the number of requests to classify improves stealth and rate of classification

CLASSIFICATION

- This is a classification problem
- Classification answers "Is it what we are looking for?"
- Classification is generally supervised learning
- Supervised learning requires training samples to have labels
- Classification methods range from simple to highly complex

ID PANEL FEATURES

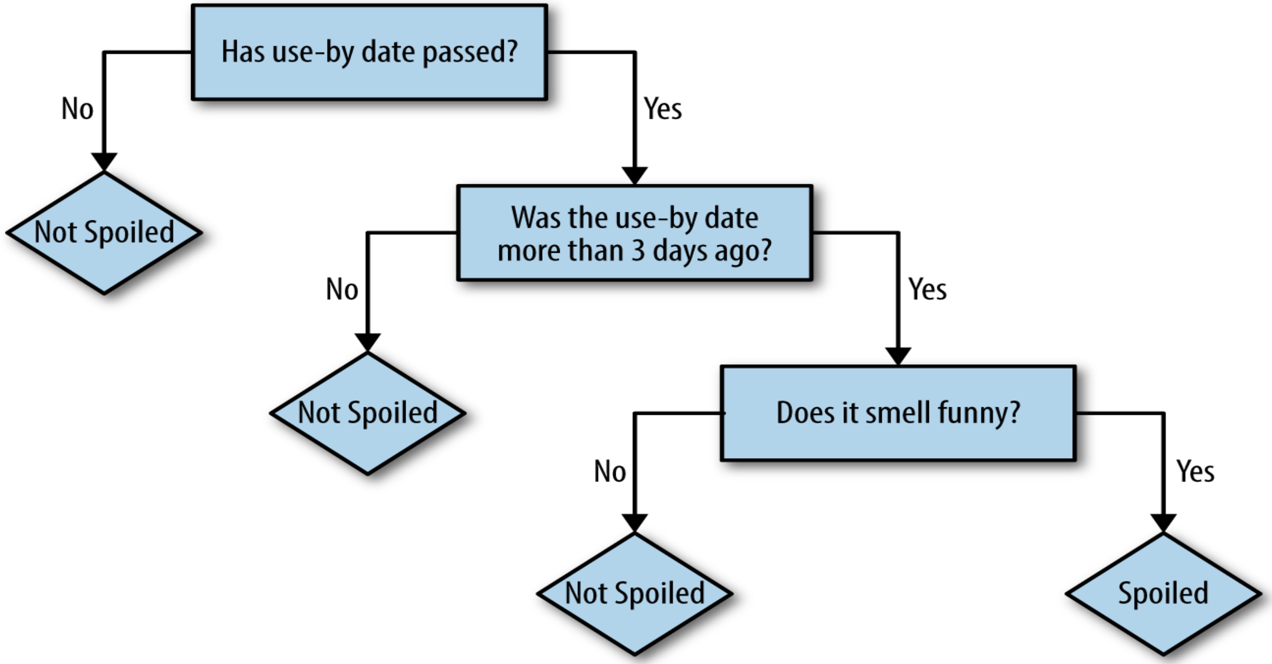
- Botnet panels are similar to normal websites
- Contain various file types, often edited
- HTTP response codes + content comparison
- Encoding content as features difficult
- ssDeep provides a continuous value by comparing content

COLLECTING DATA

- Collection of known botnet panels
- Request all known paths for all known botnet panel types
- Store HTTP status code and ssDeep of content
- Collection of sites that are not botnet panels needed as well

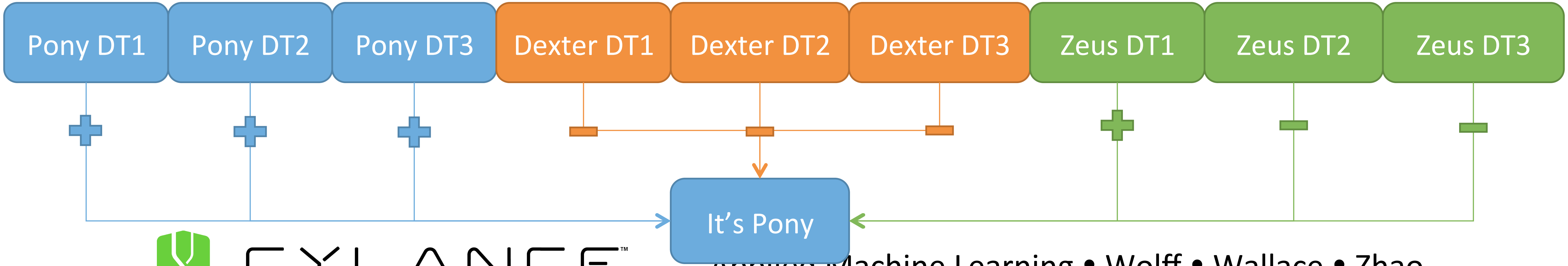
DECISION TREES

- Decision trees are simple classifiers
- Splits the dataset one feature at a time until decision is confident
- Results in a tree of queries where the results produce a decision
- Simple to train



ENSEMBLE OF DECISION TREES

- A single decision tree may be over focused on training data
- Can alleviate this problem by building multiple Decision Trees for each label
- Combining the results allows each Decision Tree to vote
- Partial answers may still be of interest to the user
- Ensembles can obtain better predictive performance



ID PANEL DEMO – COMMAND LINE

- Quick way to check if a website directory contains a botnet panel
- Easy to batch searching of multiple websites/directories
- Easy to grep results

ID PANEL DEMO – CHROME EXTENSION

- Every website directory visited is tested
- Results are stored in browser
- ssDeep ported from C to Javascript
- https://github.com/kripken/emscripten
- Extension available in Chrome extension store (free, of course)

TOOL – MARKOV OBFUSCATE

- Data exfiltration from a network often requires avoiding an outbound firewall
- Deep packet inspection looks to block anything undesirable
- Easy to encrypt data, but its also easy to drop information that can’t be read
- We can make our data look like something else entirely

MARKOV CHAIN

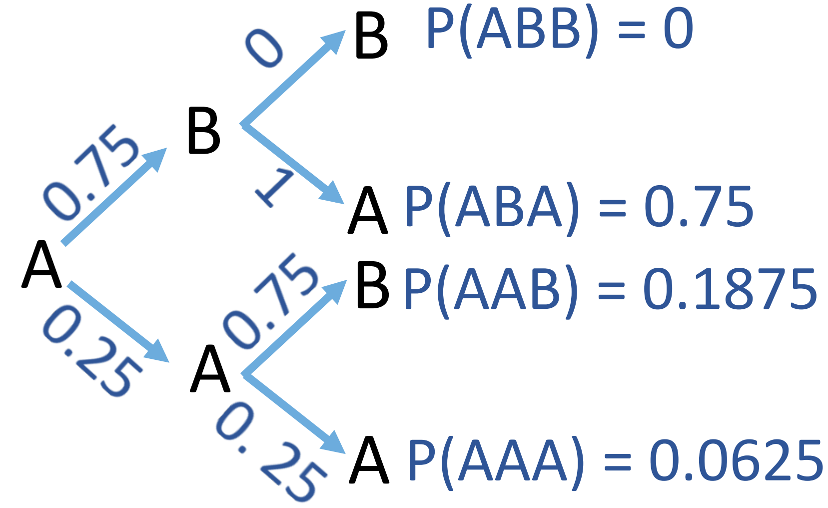
- Simple machine learning method for characterizing sequence data
- Learns the transition pattern from a state to another based on how likely a state comes after another state in the training data
- We can create sequences with transition patterns that are learned from the data it was trained on

Training Data

AAB
BAB
ABA

Transition Matrix

	A	B
A	1 (25%)	3 (75%)
B	2 (100%)	0 (0%)

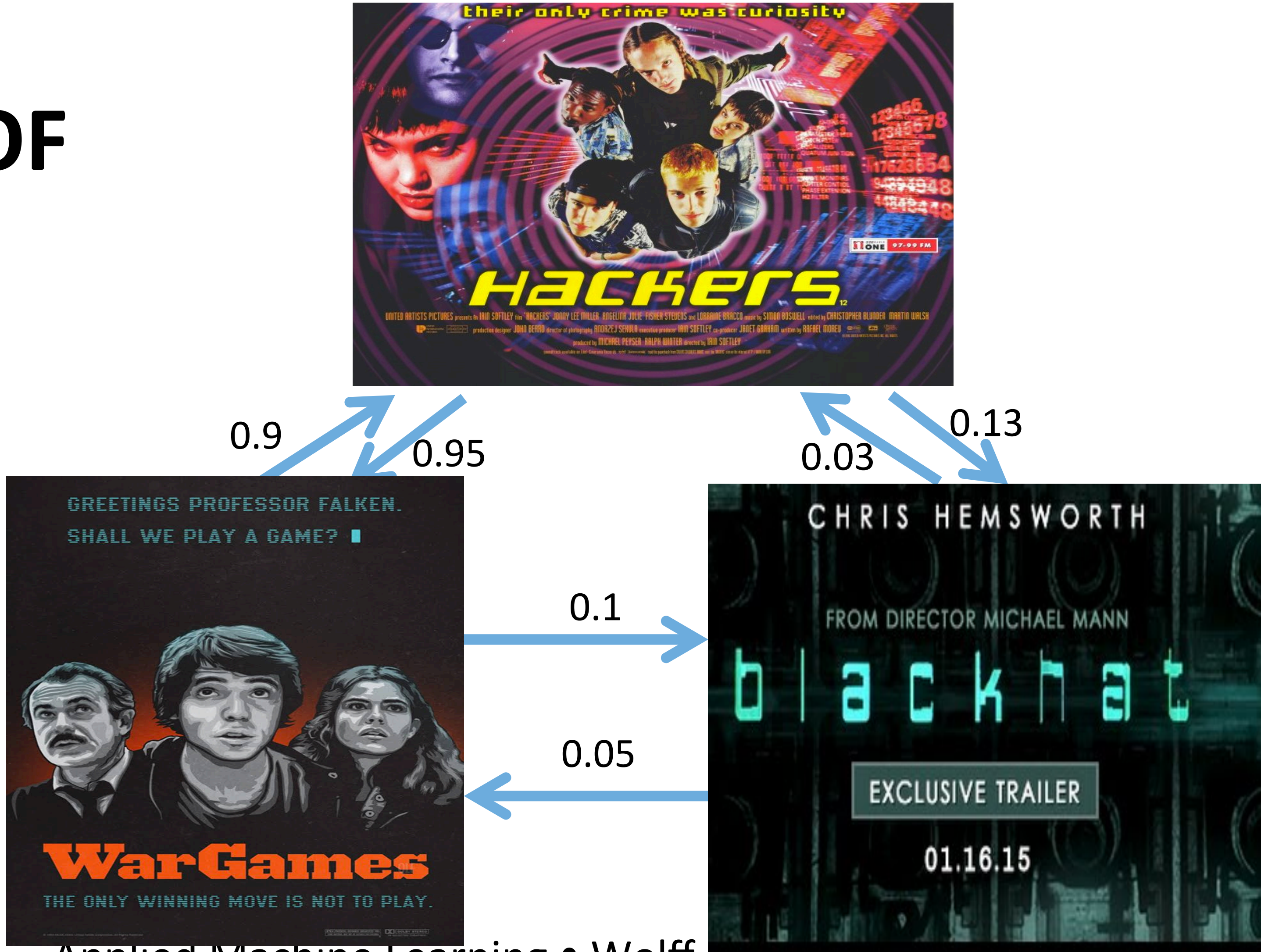


POPULAR USE CASES OF MARKOV CHAINS

Weather Prediction

	Sunny	Rainy
Sunny	0.9999	0.0001
Rainy	0.9	0.1

Recommendation



ENCODING DATA WITH A MARKOV CHAIN

- Given a transition matrix, we can sort items by how likely they are to follow our current item
- If we choose the 5^{th} most likely item, we can identify it's the 5^{th} most likely with a model trained on the same data
- This encodes the number 5 in the transition from our first item to our second item

MARKOV OBFUSCATE - ENCODING

- Train our model with a book
- Observing transitions from word to word
- Generate data based on transition probabilities
- Demo

MARKOV OBFUSCATE - WRAPPING

- Simple to transfer our data through a pipeline that looks like normal HTTP traffic
- Looks like a user posting to their blog
- Demo

MARKOV OBFUSCATE – HAVING FUN

- Train our models on Taylor Swift lyrics
- Train a Markov Model based on Taylor Swift songs
- Play the generate lyrics through festival with tones/beats learned from songs
- First live “Tylance Swift” concert, demo

WRAPPING UP

- Any problem where there is a significant amount of data generated could benefit from a machine learning approach
- Lots of great online resource to help anyone get started
- Having labeled or annotated data makes more ML approached viable compared to unlabeled data

QUESTIONS?

- Email: machinelearning@cylance.com
- Stop by booth #1124
- Career opportunities: https://www.cylance.com/cylance-careers