CMSC 449/691
Malware Analysis

Lecture <flexible>

Malware Data Science

Revised April 7, 2025

Malware Data Science

- Hundreds of thousands of unique, new malware samples daily
- Impossible for human analysts to investigate every file
- Need to rely on automation and data science!
  - Can we automate common analysis tasks?
  - Can we enhance threat hunting, malware classification, etc.?

Public Malware Datasets

- **EMBER2018 Dataset** - PE metadata extracted from ~500,000 malicious files and ~500,000 benign files
  - https://github.com/elastic/ember
- **SOREL-20M Dataset** - PE Metadata from ~10M benign files and ~10M disarmed malware samples from 11 categories
  - https://github.com/sophos/SOREL-20M
- **MOTIF Dataset** - 3,095 malicious PE files from 454 malware family families with ground truth labels
  - https://github.com/boozallen/MOTIF/

Automating Malware Analysis

Malware Analysis Pipeline

- Most large malware analysis shops have a pipeline which automatically processes malware
  - Newly-ingested files
  - Re-processing older files
- Basic static analysis
- Basic dynamic analysis
- Malware signatures (YARA, Snort, AV)

Automating Basic Static Analysis

- Extract file metadata
  - Python libraries such as pefile, lief for extracting PE metadata
- Compute similarity-preserving hashes
  - SSDEEP
  - TLSH
  - LZJD / BWMD
- Compute metadata hashes
  - pehash
  - imphash

Side Note: Automating Disassembly

- Can automate many advanced static analysis tasks
  - Too slow to apply to every malware sample though
- Capstone library for Python is an excellent linear disassembler
  - Pefile can also do some of this
- Most modern disassemblers/decompilers (including IDA Pro, Ghidra, Binary Ninja) support plugins
  - Can use these for automating many advanced static analysis tasks

Automating Basic Dynamic Analysis

- Automated sandboxes such as Cuckoo and DRAKVUF can be self-hosted and used for identifying malware behavior

- Generate report about the malware's actions:
  - Process tree
  - Created files
  - Network traffic
  - Configuration changes

Malware Signatures

- Many malware analysis shops have a collection of in-house antivirus products
- May also maintain other large collections of signatures:
  - YARA rules - based on file contents
  - Snort rules - based on network traffic
- The yara-python library lets YARA be used programmatically

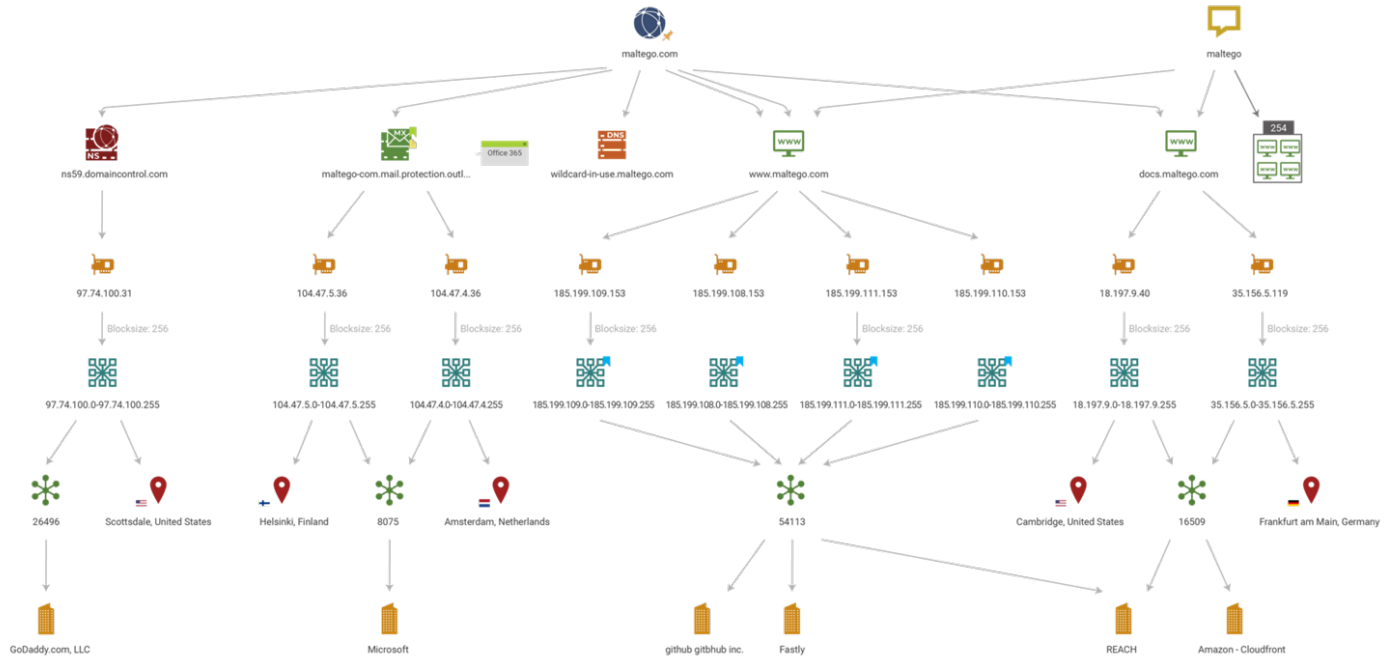
Malware Data Science Problems

Threat Hunting

- Goal is to identify malware of interest, usually related to a set of known IOCs
  - Often investigating a malware family, campaign, or threat group
- File similarity metrics, metadata hashing
- Identifying files which contact the same IPs / domain names
- Lots of nearest-neighbor lookup research in this area!

Threat Hunting with Maltego

- Threat hunting tool which can be used to pivot from known IOCs to related ones
- Generates a graph showing how IOCs are related



Malware Featurization

- A **feature vector** is the standard input for most machine learning algorithms (nearest-neighbor lookup, clustering, classification, outlier detection etc.)

- Essentially a list of numbers which somehow describe the attributes of a data point
  - `< 10, -2, 3, 7, 0 >`

- Each number in the feature vector describes a specific attribute of the data point

Malware Featurization

- How do we best represent malware as a feature vector?

- **EMBER vector** - Based on metadata from PE files
  - Features include byte frequency, metadata from PE headers, strings, imports, resources, etc.
  - Vector contains 2,351 features

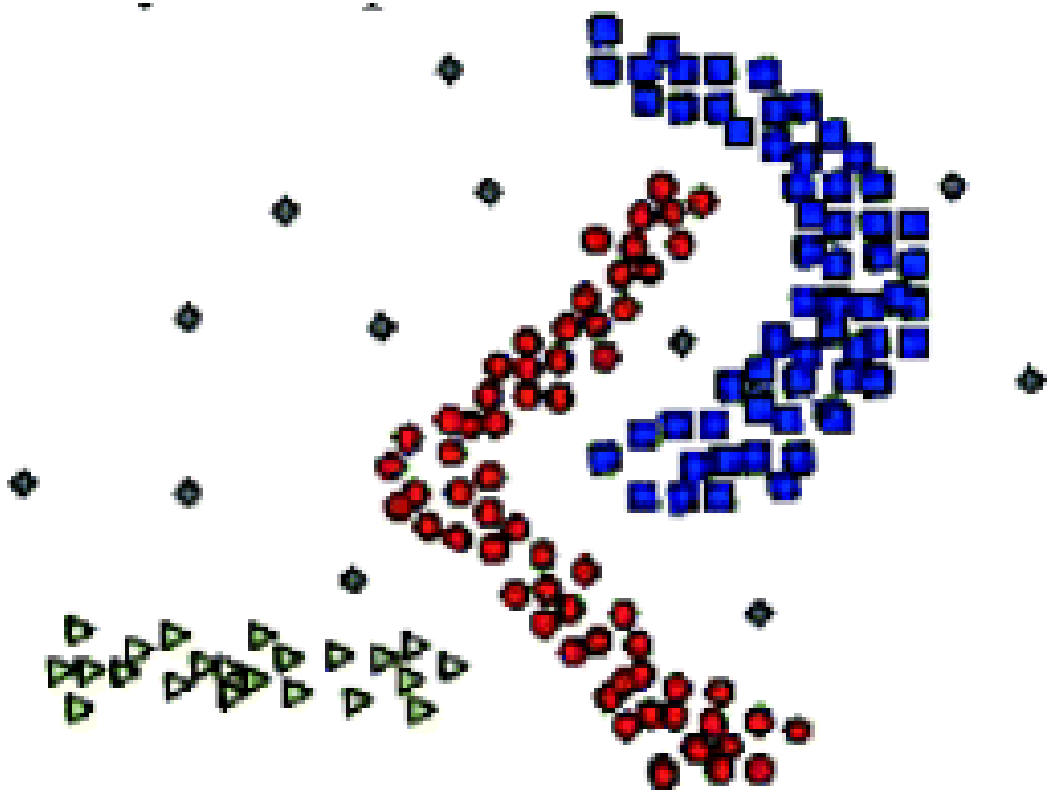
- **BWMD vector** - Based on the Burrows-Wheeler Transform
  - Converts any sequence of bytes into a fixed-length vector
  - Vector contains 65,536 features

Malware Clustering

- Clustering: Identify groups of similar malware samples
- Usually need to convert malware into feature vectors first!
  - Or have a method for computing similarity between two files
- Many different clustering algorithms, depending on your situation and goals
  - Density-based and hierarchical clustering algorithms work best
  - But may run slowly on large datasets

Density-Based Clustering

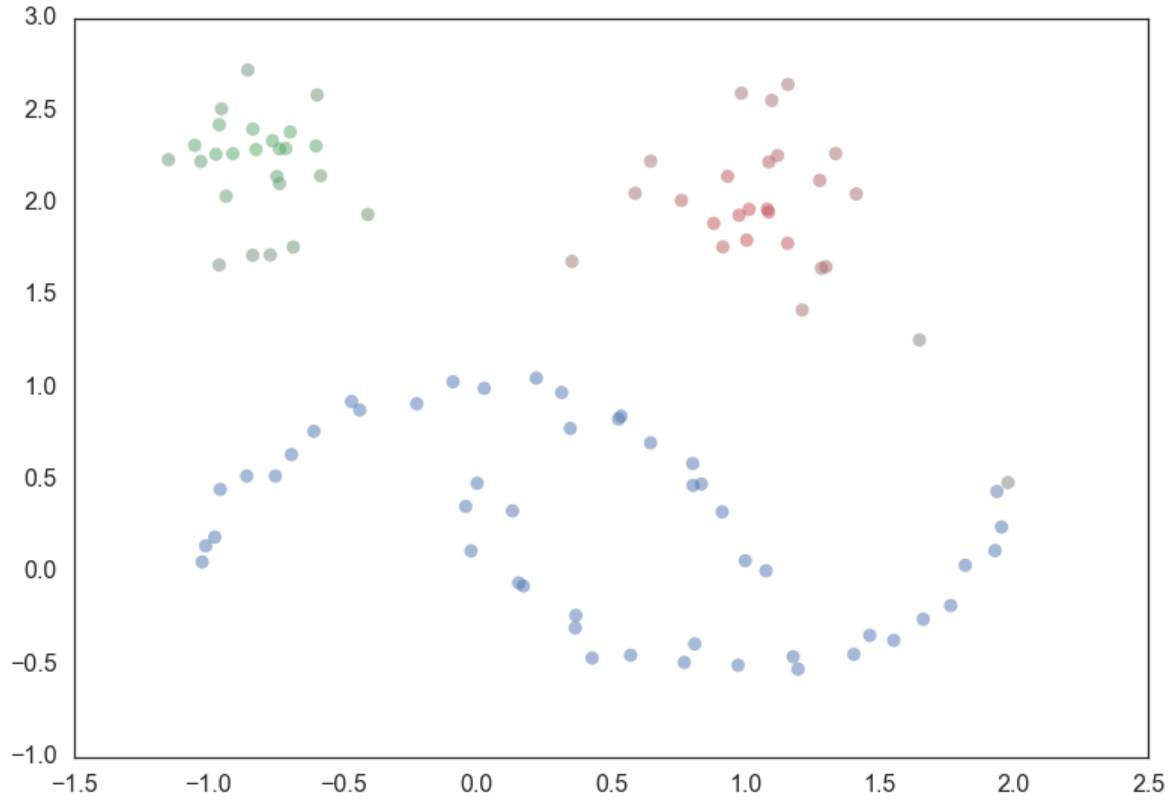
- Assume that malware data tends to form lots of small, dense clusters (each cluster being a malware family)
- Dense groups and their neigbors become a cluster
- Algorithms such as DBSCAN and OPTICS often work well!



Hierarchical Clustering

- Idea is to form a “hierarchy” of data points, iteratively grouping the nest most similar points

- Algorithms such as Hierarchical Agglomerative Clustering (HAC) and HDBSCAN are good!

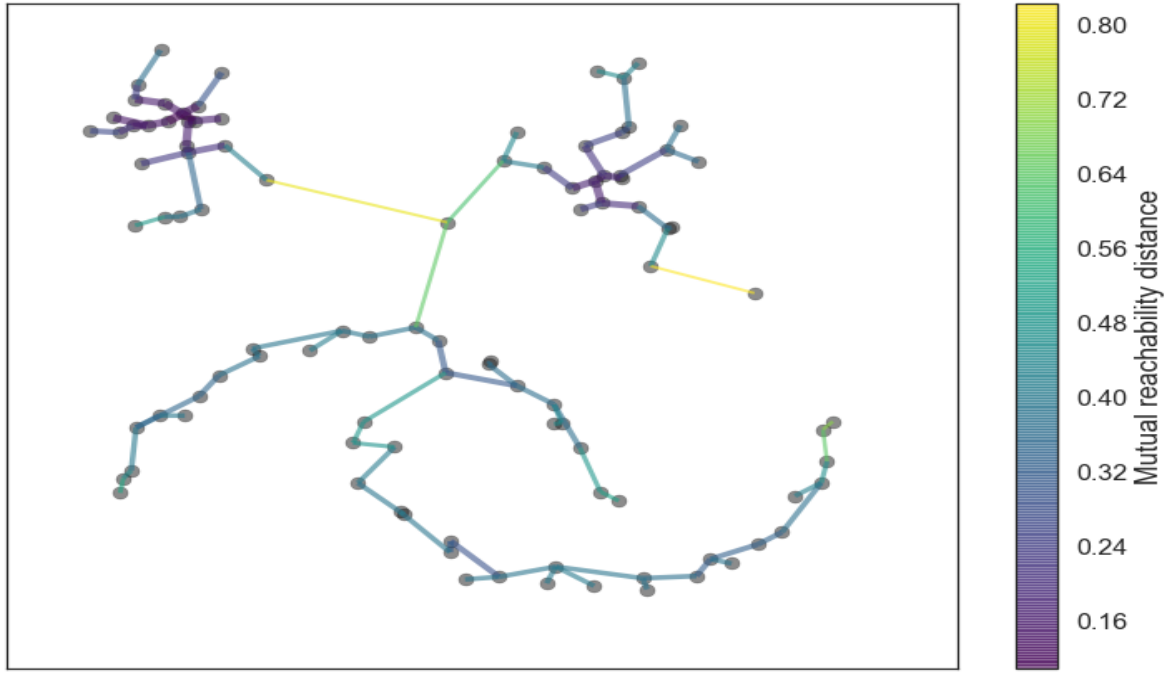


https://hdbscan.readthedocs.io/en/latest/how_hdbscan_works.html

Hierarchical Clustering

- Idea is to form a “hierarchy” of data points, iteratively grouping the nest most similar points

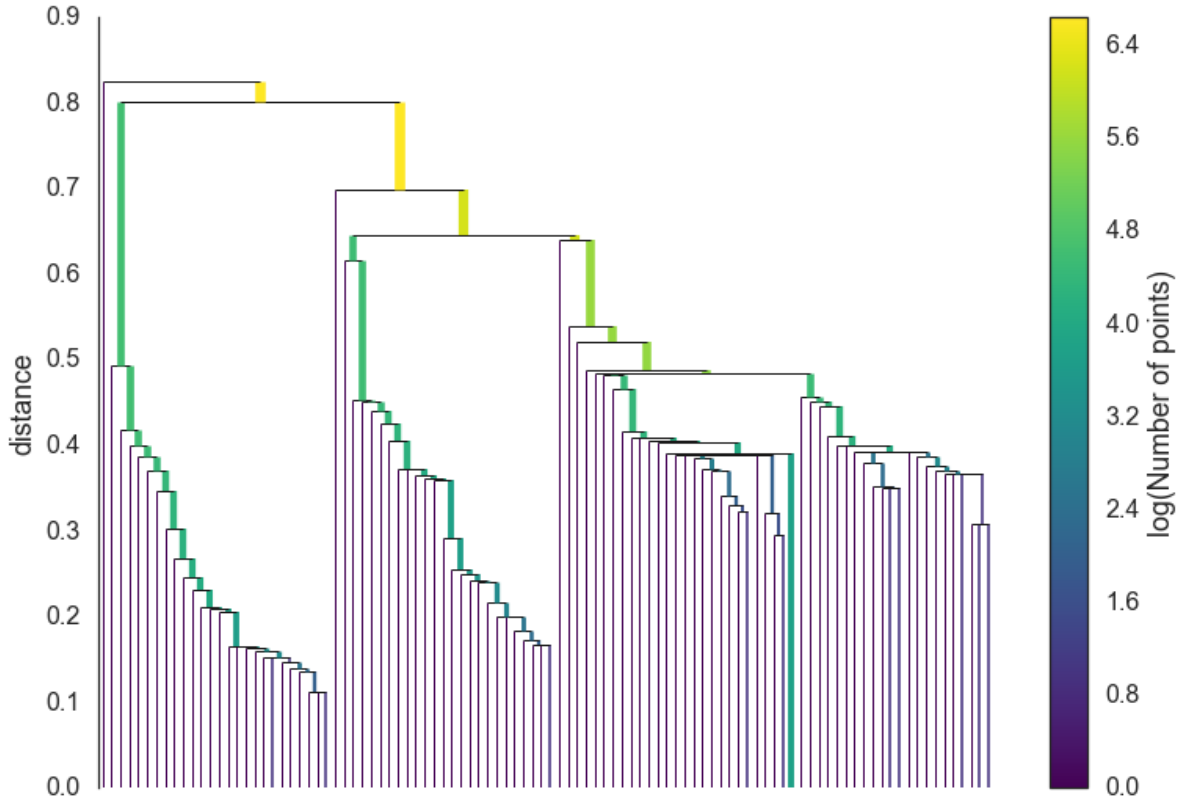
- Algorithms such as Hierarchical Agglomerative Clustering (HAC) and HDBSCAN are good!



https://hdbscan.readthedocs.io/en/latest/how_hdbscan_works.html

Hierarchical Clustering

- Idea is to form a “hierarchy” of data points, iteratively grouping the nest most similar points
- Hierarchical Clustering can be used to produce a dendrogram

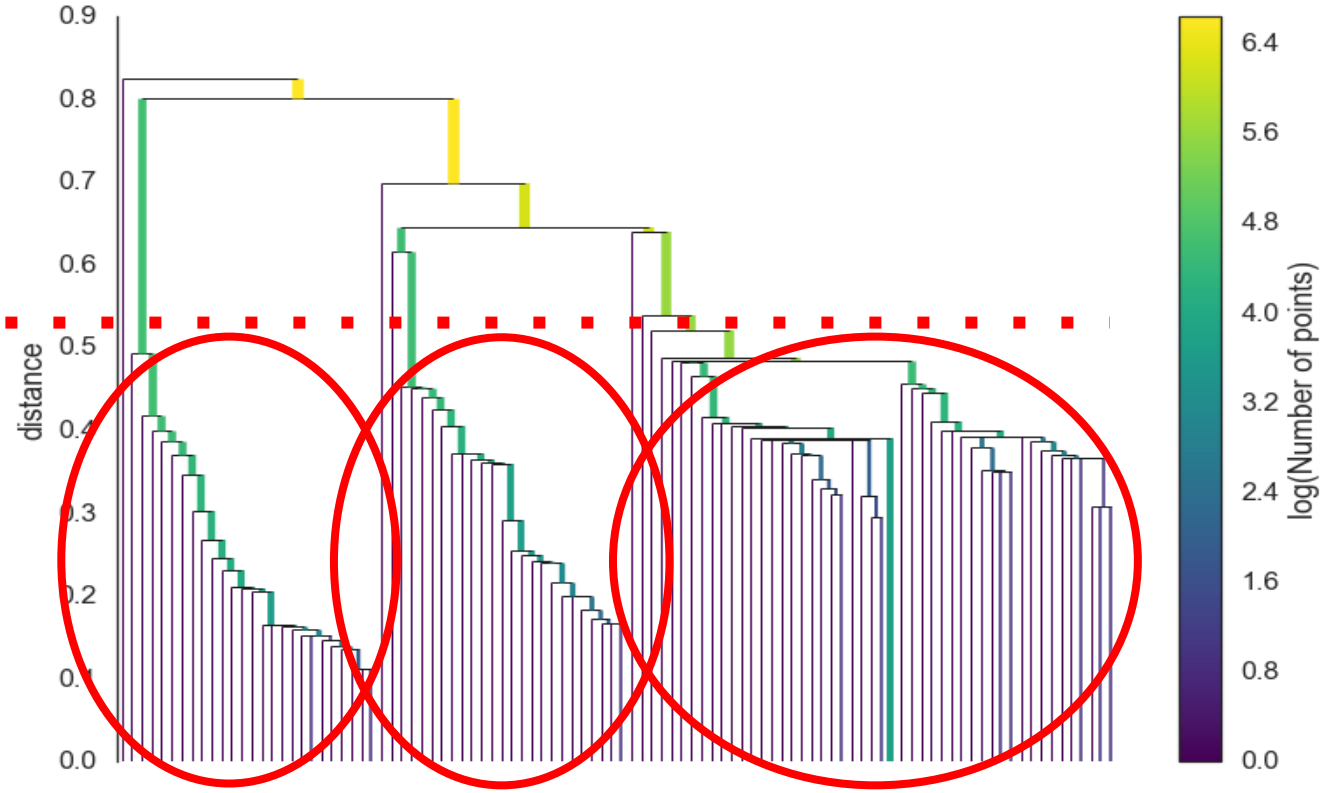


https://hdbscan.readthedocs.io/en/latest/how_hdbscan_works.html

Hierarchical Clustering

- Idea is to form a “hierarchy” of data points, iteratively grouping the nest most similar points

- Algorithms such as Hierarchical Clustering can be agglomerative (bottom-up) or divisive (top-down)



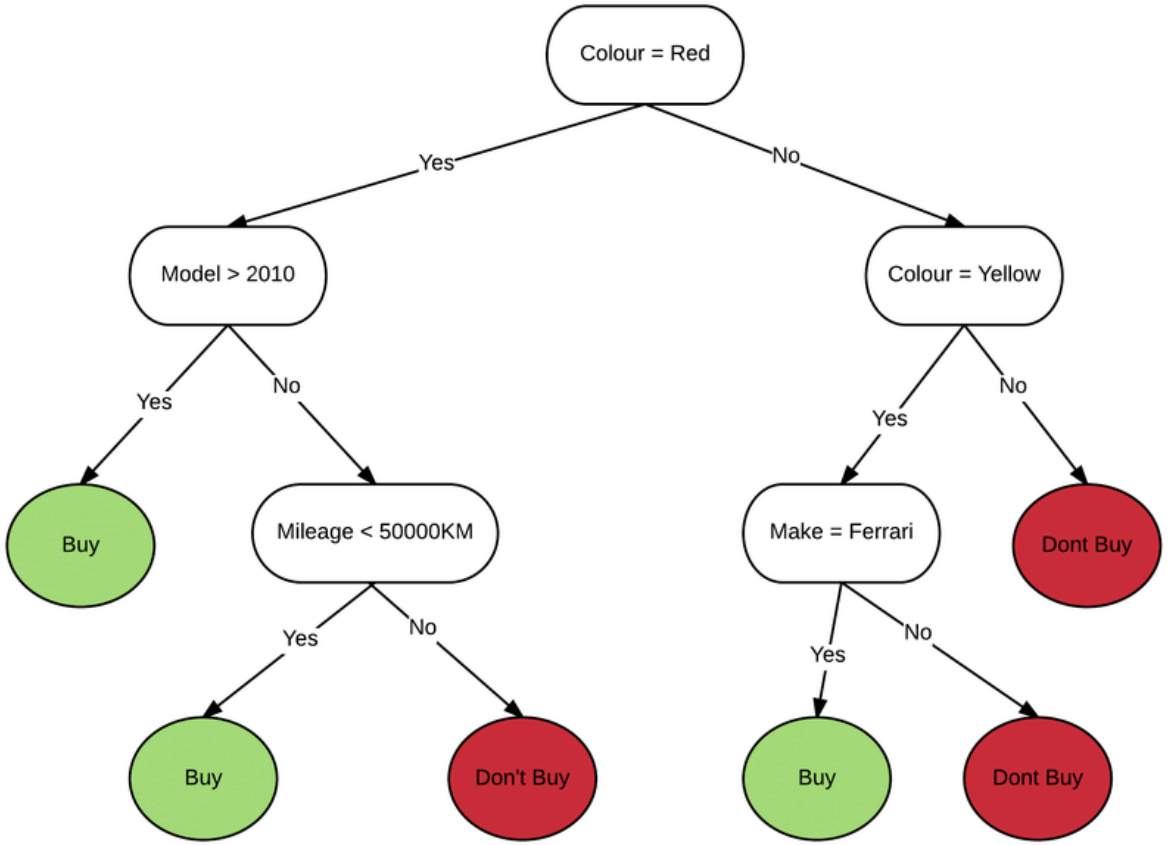
Malware Classification

- **Classification:** Task of assigning a data point to a class
- **Malware Detection:** Classify file as benign/malicious
- **Category Classification:** Classify file by behavior category (e.g. ransomware, keylogger, etc.)
- **Family classification:** Classify file by malware family

Decision Tree Classifiers

- Decision Tree - simple classification algorithm that maps combinations of features to a class outcome

- In the figure, the result for red cars newer than 2010 is the “buy” class



Decision Tree-Based Ensemble Classifiers

- **Ensemble:** Powerful machine learning technique where a collection of classifiers (often decision trees) vote on a class

- Algorithms like Random Forests, XGBoost, and LightGBM are extremely accurate for malware detection
  - Usually trained on EMBER feature vectors

Deep Learning-Based Malware Classifiers

- Lots of research using neural networks and deep learning for malware classification

- MalConv and MalConv2 are leading models
  - Treat malware as a large sequence of bytes
  - Apply 1-D convolution to extract spatial relationships from data
  - Features are learned through convolution!