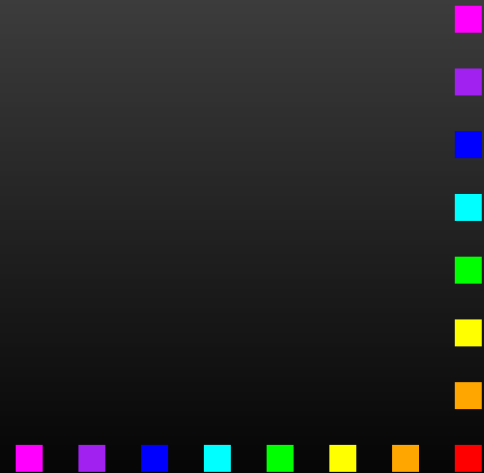


# Data Mining for Network Intrusion Detection

S Terry Brugger

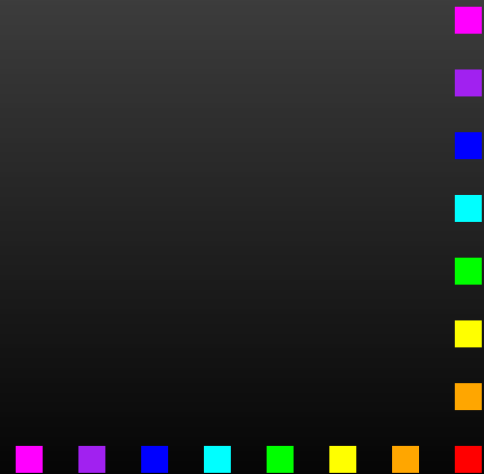
UC Davis

Department of Computer Science



# Overview

- This is important for defense in depth
- Much work has been done in the area, but no solution yet
- I will investigate an ensemble approach as a possible solution



# We need to detect intrusions

- Can't stop intrusions, so need to mitigate them
- Can mitigate (stop the attackers) when they're detected, or take other corrective action (improving defenses)
- Part of defense in depth

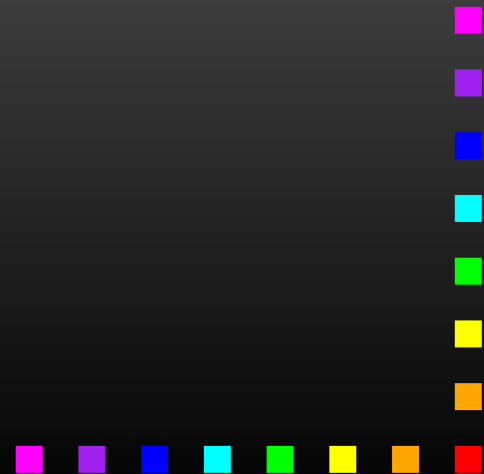
# Current IDSs are not sufficient

- Only detect known attacks
- Can't detect insider attacks (privilege abuse)
- Don't have a holistic picture of the network to detect multi-step attacks over a long time period
- Data for detection is available, but sysadmin resources are limited

# The solution is Data Mining

Data Mining: The process of extracting useful and previously unnoticed models or patterns from large data stores.

(Also called “sensemaking”.)



# Data mining should be done in an Offline environment

- Last line of defense – used in concert with real-time systems
- Allows system to be queried post hoc
- More complete session information
- Data mining techniques are expensive (even with mitigation through cost-based models [Lee])



# More reasons to work in an Offline environment

- Periodic batch processing provides trade-off between timeliness and efficiency
- Allows for holistic picture, grouping related activity
- Harder to attack IDS via denial of service

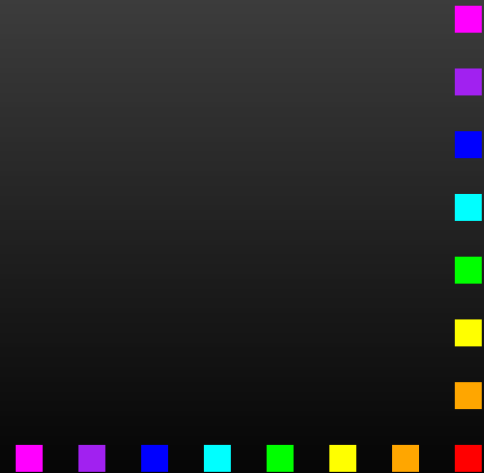


# We mine network connection records

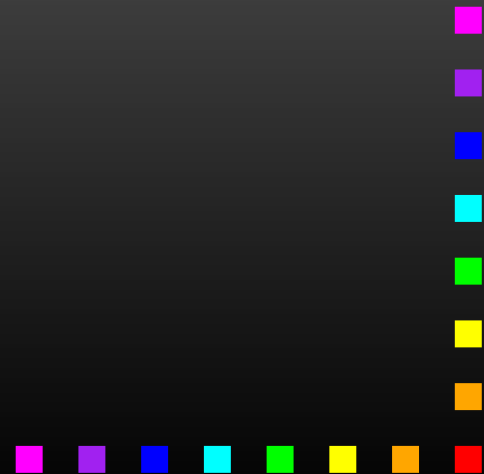
- Readily available
- Efficient (good size to information ratio)
- Easy for data mining methods to operate on
- Avoids privacy issues and encryption of data streams

# How network connection records break down

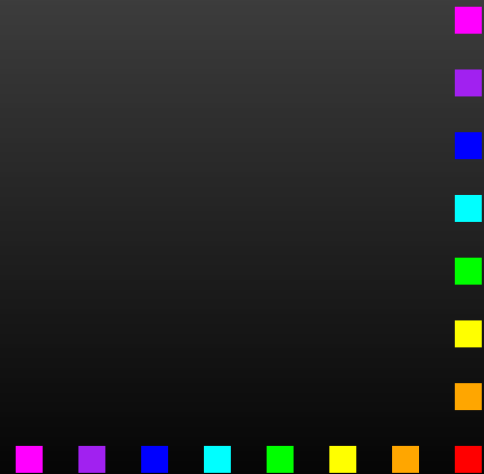
- Intrinsic attributes
  - Essential attributes
    - Axis and reference attributes
  - Secondary attributes
- Calculated attributes



# Place-holder for intrinsic attributes table



# Place-holder for calculated attributes table



# Additional useful information

- Calendar schema
- Normalization (pseudo-Bayes estimators, probability given other values)
- Compression (UDP, ICMP, source net, information gain)
- Selection using genetic algorithm [Helmer]



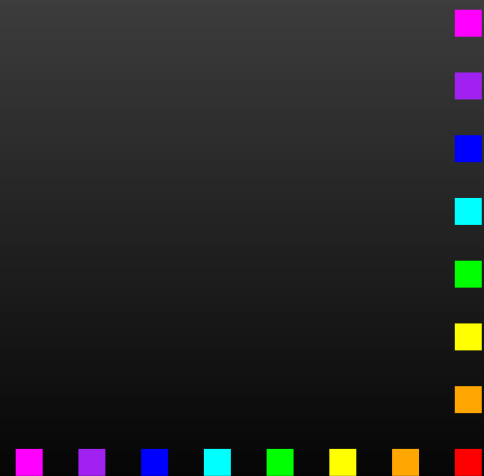
# Many datasets, none great

- Information Exploration Shootout (IES)
- Internet Traffic Archive [LBL]
- Security Suite 16 [InfoWorld]
- DARPA Off-line Intrusion Detection Evaluation
  - 1998
  - KDD-Cup
  - 1999



# What's so bad with the IDEval?

- McHugh identified numerous procedural problems
  - Unrealistic data rates
  - Failure to show relation to real traffic



# More problems with IDEval?

- Mahoney & Chan found problems with the data
  - Some fields like TTL predictable
  - Allowed naive methods to achieve high detection rates
  - Correctable by mixing with real traffic
- Despite all this, DARPA dataset still the standard



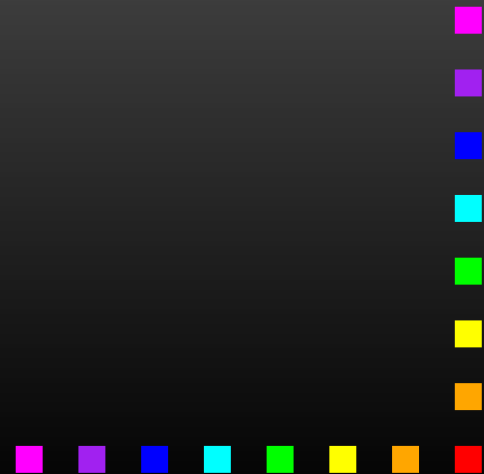
# Some desired features in existing systems

- Information Security Officer's Assistant (ISOA) [Winkler] and Distributed Intrusion Detection System (DIDS) [Snapp] did data fusion and multi-sensor correlation
- SRI work: IDES, NIDES, EMERALD provide more published research in this area



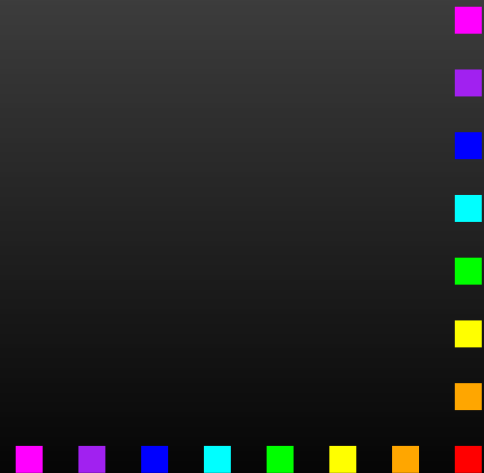
# Commercial offerings to correlate alarms

- RealSecure SiteProtector
- Symantec ManHunt
- nSecure nPatrol
- Cisco IDS
- Network Flight Recorder (NFR)



# Commercial offerings for audit trail integrity

- Computer Associates' eTrust Intrusion Detection Log View
- NetSecure Log

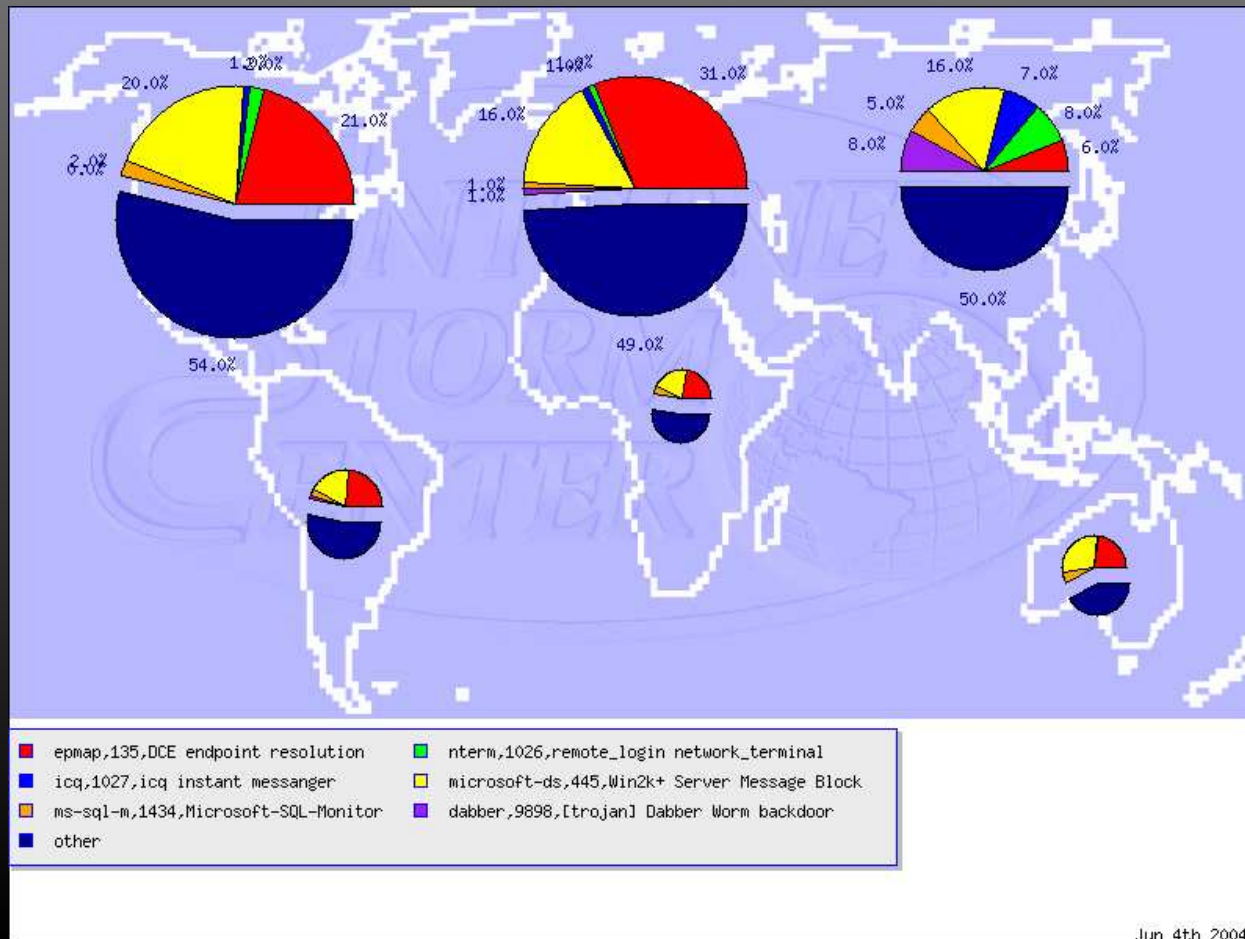


# Some functionality offered by services

- SANS Internet Storm Center
- dShield (Independent Storm Center Analysis and Coordination Center)
- myNetWatchman
- Security Focus DeepSight Analyzer
- Managed service available from Counterpane, ISS, and Symantec



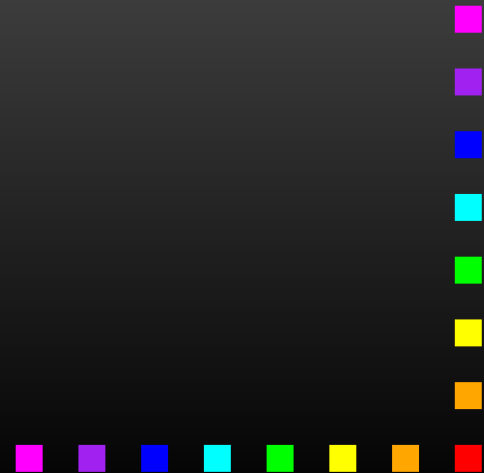
# Internet Storm Center



Jun 4th 2004

# Two major data mining approaches

- Statistical (top-down)
- Machine learning (bottom-up)



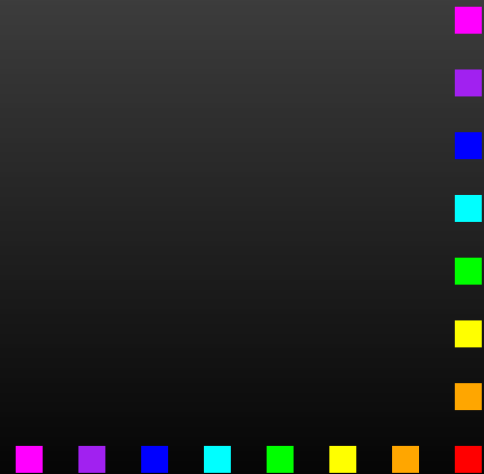
# Statistical techniques

- Probability of record given correlated probability of individual fields [SRI]
- Probability of record given Bayes network of conditional probabilities [Staniford]
- Probability of value not seen in training given alphabet size and time since last anomaly [Mahoney]
- Decision trees (ID3) [Sinclair]



# Many types of machine learning

- Classification
- Clustering
- Support Vector Machines [Eskin, Mukkamala]
- Others



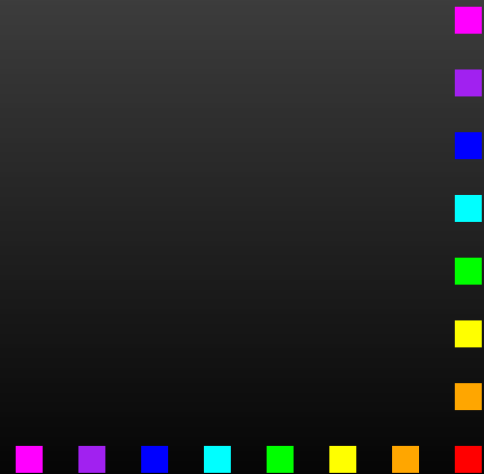
# Classification approaches

- Inductive rule [Lee, Helmer, Warrender]
- Genetic algorithms  
[Neri, Sinclair, Dasgupta, Crosbie, Chittur]
- Fuzzy rules [Dickerson, Luo]
- Neural nets [Giacinto, Ghosh, Ryan, Endler]
- Immunological [Hofmeyr, Dasgupta, Fan]



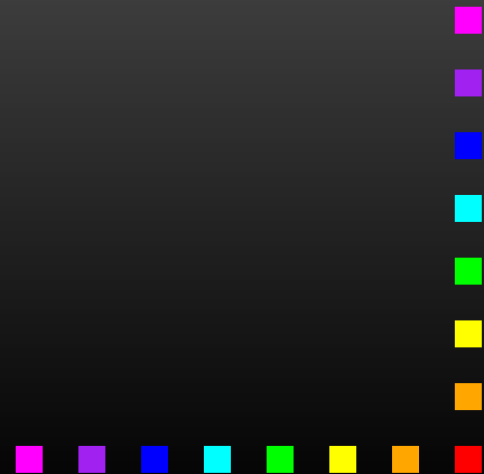
# Clustering approaches

- Fixed width, k-nearest neighbor [Portnoy, Eskin, Chan]
- k-means [Bloedorn]
- Learning Vector Quantization [Marin]
- Simulated annealing [Staniford]



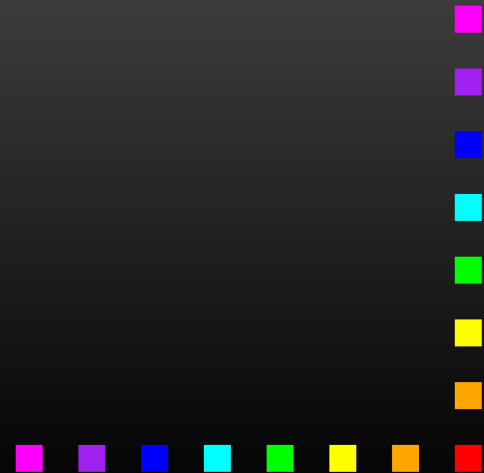
# More Clustering approaches

- Approximate Distance Clustering & AKMDE [Marchette]
- Dynamic Clustering [Sequeira]
- Parzen-window [Yeung]
- Instance-based learner [Lane]



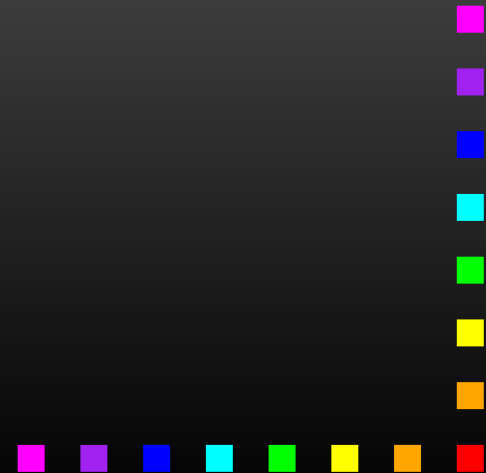
# Other approaches

- Colored Petri nets [Kumar]
- Graphs [Staniford, Tolle]
- Markov models [Lane, Warrender]



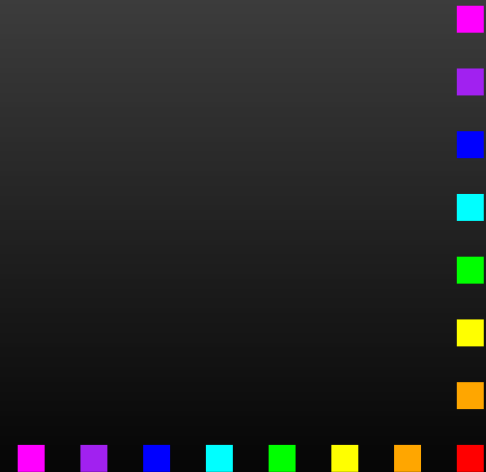
# Other proposed methods

Proposed by	Method
[Denning]	operational model
[Denning]	mean and standard deviation
[Denning]	multivariate model
[Denning]	Markov process model
[Kumar]	generalized Markov chain
[Denning]	time series model
[Frank,Endler]	Recurrent neural network



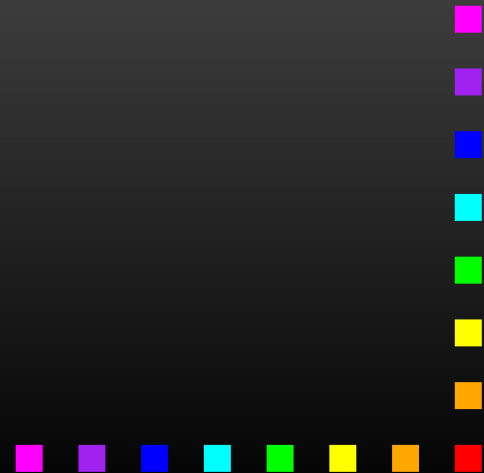
# More proposed methods

[Chan,Prodromidis]	C4.5, ID3, CART, WPEBLS
[Bass]	Dempster-Shafer method
[Bass]	Generalized EPT
[Lane]	Spectral analysis
[Lane]	Principle component analysis
[Lane]	Linear regression
[Lane]	Linear predictive coding
[Lane]	$(\gamma, \epsilon)$ -similarity



# Research to date has provided progress, no solution

- Most data mining methods for ID are good at detecting particular types of malicious activity
- False positive rates are high (base-rate fallacy [Axelsson])



# Better performance through Ensemble techniques

(Also called meta-learning or multi-strategy learning)

“It is well known in the machine learning literature that appropriate combination of a number of weak classifiers can yield a highly accurate global classifier.” [Lane]



# More support for Ensemble techniques

Neri notes “that combining classifiers learned by different learning methods, such as hill-climbing and genetic evolution, can produce higher classification performances because of the different knowledge captured by complementary search methods.”



# Ensemble techniques important for ID

“In reality there are many different types of intrusions, and different detectors are needed to detect them.” [Axelsson]

“Combining evidence from multiple base classifiers . . . is likely to improve the effectiveness in detecting intrusions.” [Lee]



# Some work has been done with Ensemble techniques

- Manually built covariance matrix in [N]IDES to use multiple classifiers
- Crosbie's autonomous agents and Staniford's SPICE also do basic correlation of statistical classifiers
- Lee, Fan, et al. proposed use for incorporating classifiers trained on new data and aging out old classifiers



# More prior work with Ensemble techniques

- Lee, Fan, et al. also used cost-based meta-classifiers
- ADAM uses multiple classifiers for filtering [Barbará]
- Giacinto et al. use ensemble of neural nets trained on different feature sets

# Outstanding questions

1. For baseline purposes, what is the accuracy of a contemporary NID on the DARPA dataset?
2. Ideal number of states for a Hidden Markov Model, and what parameters influence this value?
3. Ideal feature sets for different data mining techniques?
4. Should connectionless protocols like UDP and ICMP, be compressed to a single connection (as in TCP)?
5. Separate training sets for classifiers and meta-classifiers?



# More Outstanding questions

6. What is the accuracy of ensemble based offline NID employing numerous, different, techniques?
7. How much data is required in order to properly train a data-mining based IDS?
8. How dependent is data mining performance on training on same network as it's used?
9. Should hosts and / or services be grouped together for usage profiles?



# More Outstanding questions

10. Other forms of data compression to improve accuracy?
11. Predictive capabilities of an offline network intrusion detection system?
12. How much will the incorporation additional data sources improve performance?
13. Better accuracy by considering state of hosts with connection as transition operator?
14. Does the ideal time window,  $w$ , depend on the current state of a host?

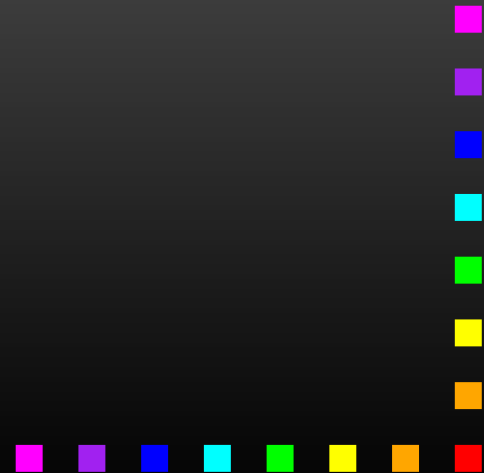


# Big questions

15. What similarities or differences exist in the traffic characteristics between different types of networks that impact the performance characteristics of a network intrusion detector?
16. What is the user acceptability level of false alarms?
17. How much can false alarms be reduced through the use of user feedback, and learning algorithms or classifier retraining?

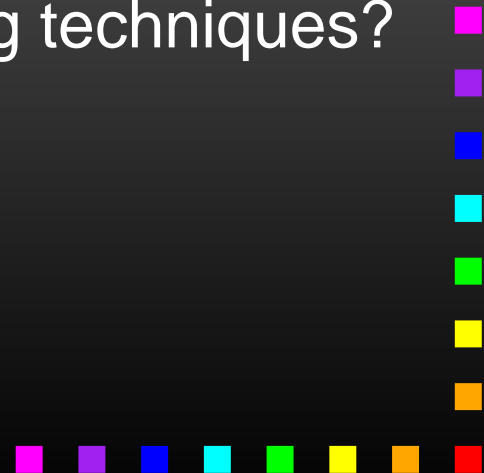


# What am I going to do about it?



# Answer the first six

1. For baseline purposes, what is the accuracy of a contemporary NID on the DARPA dataset?
2. Ideal number of states for a Hidden Markov Model, and what parameters influence this value?
3. Ideal feature sets for different data mining techniques?



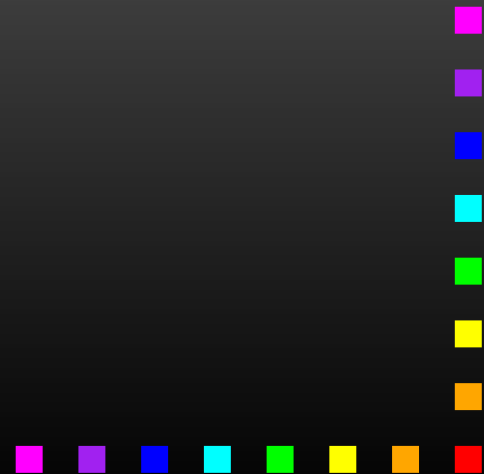
# Goal

4. Should connectionless protocols like UDP and ICMP, be compressed to a single connection (as in TCP)?
5. Separate training sets for classifiers and meta-classifiers?
6. What is the accuracy of ensemble based offline NID employing numerous, different, techniques?

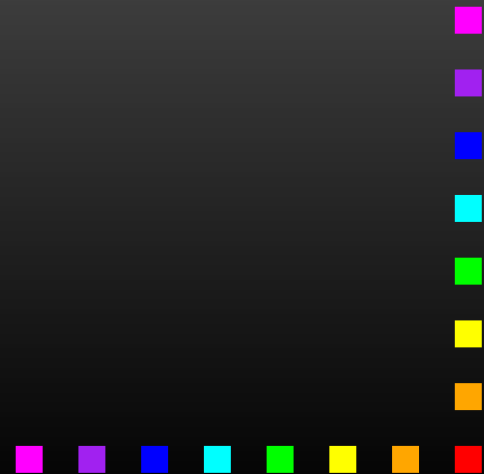


# Approach

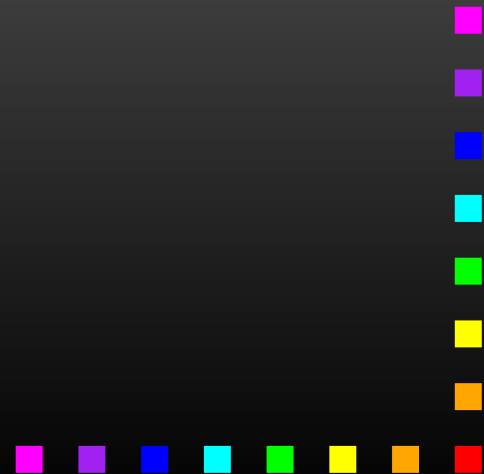
- Datasets
  - 1998, 1999 DARPA TCP data
  - 1998 DARPA mixed with real data
- Baseline Snort
- Connection Mining (tcpreduce)



# Place-holder for Connection Table Creation

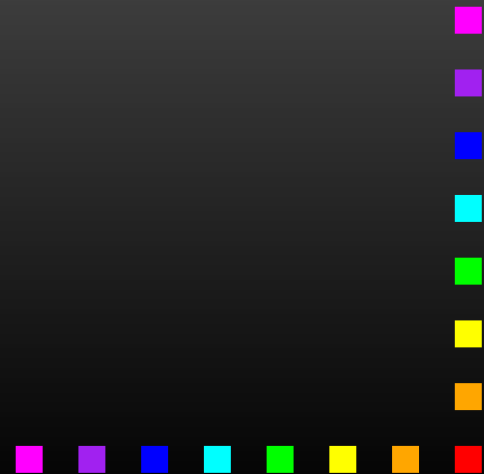


# Place-holder for Results Table Creation



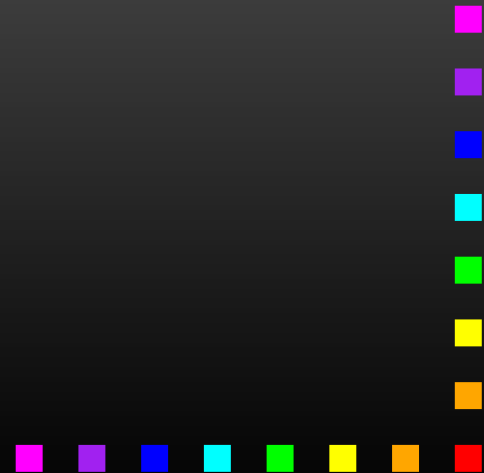
# Base method creation

- Training mode
- Classification mode



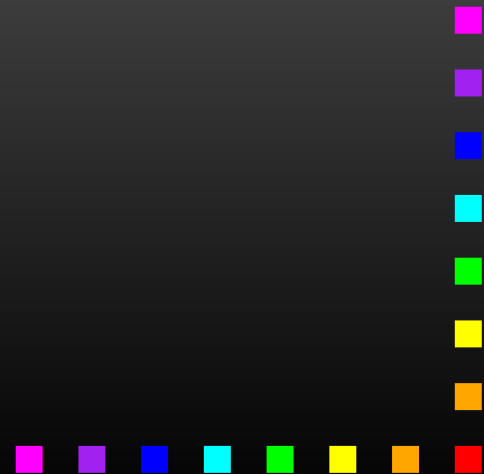
# Anomaly detection methods

- Bayes network
- Non-self bit-vectors
- Hidden Markov Model



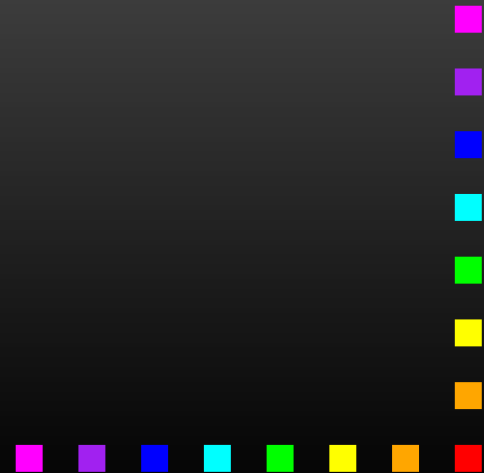
# Classification methods

- Decision tree
- Associative rules
- Neural network
- Elman network
- Genetic algorithm
- Clustering algorithm
- Support Vector Machine



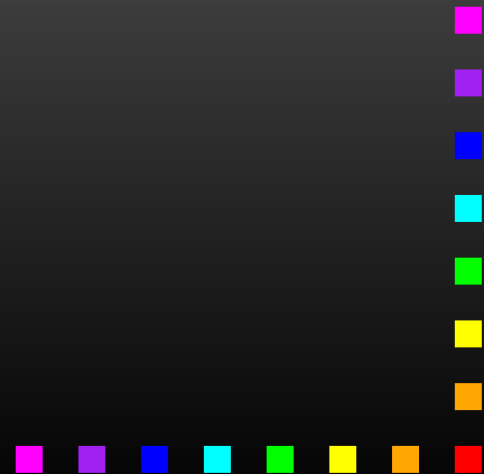
# Further approach

- Ideal parameter determination
- Ideal feature set determination
- Base classifier analysis
- Training information population



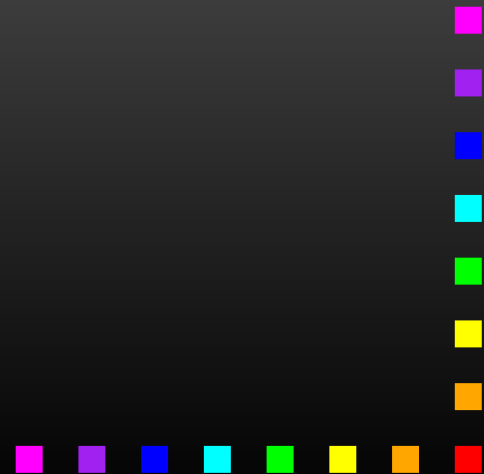
# About Meta-classifiers

1. Total anomaly from individual
2. Total probe from individual
3. Total DoS from individual
4. Total R2L from individual
5. Total U2L from individual
6. Total threat from individual
7. Total threat from totals



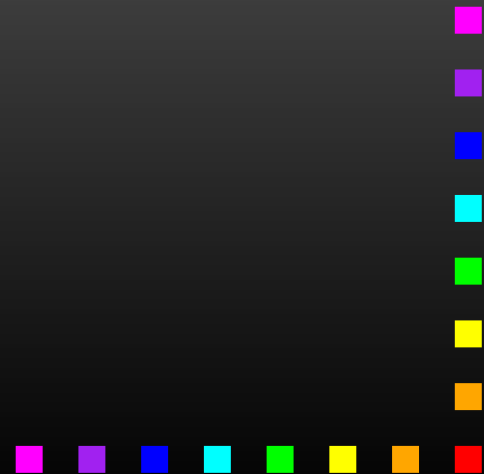
# Meta-classifiers

- Naive-Bayes
- Decision tree
- Associative rules
- Neural network
- Genetic algorithm
- Support Vector Machine



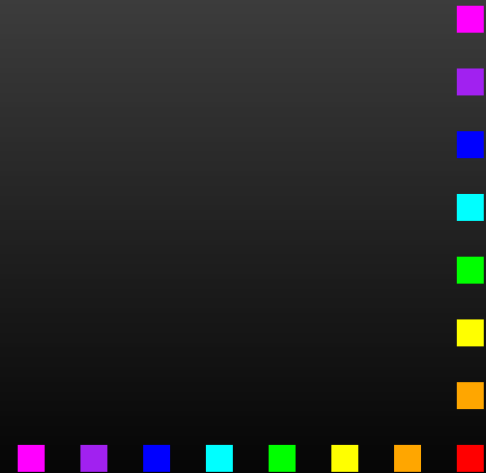
# Final approach

- Ideal meta-classifier training
- Ideal performance testing
- Analysis and writeup



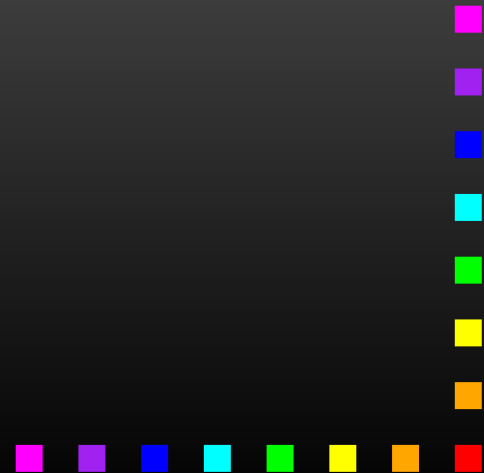
# Timeline

Step	Finish date
Mixed dataset generated	7 October 2004
Baseline	1 November 2004
Connection mining	15 October 2004
Table creation	15 August 2004
Base classifiers	1 April 2005
HMM parameter estimation	15 April 2005
Ideal feature set determination	1 July 2005



# A little more time

Base classifier analysis	15 August 2005
Training information population	1 September 2005
Meta-classifiers	22 October 2005
Ideal meta-classifier training	15 December 2005
Ideal performance testing	1 February 2006



# Completion of dissertation

1 May 2006

