



Clustering and Classification for Cyber Crime

Jesse Kornblum

Outline

- Introduction
- Similarity
- Fuzzy Hashing
- Features
- Distance Measures
- Feature Selection
- Clustering
- Classification
- Questions

Introduction



- Kyrus
 - Winston Wolf of computer security
- Computer Forensics Research Guru
 - md5deep/hashdeep
 - fuzzy hashing (ssdeep)
 - foremost

Introduction

- Analyzing an infinite number of programs
 - Only five minutes per sample
- Which of these programs are similar to each other?
- Which of these programs fit into existing categories?
 - Variant of {Zeus|Spybot|Blackhole}
 - Written by Evil D. Hacker?
 - Related to the last intrusion?

Similarity

- What does it mean for two things to be similar?

Similarity

- Depends on:
 - The kind of things be compared
 - How they're being compared

Example



Example

- Both live in Washington DC
- Both like a good hamburger
- Both are dog people

- Conclusion: Similar

- President Obama is much taller
- Presenter does not have gray hair
- Work in different career fields

- Conclusion: Not similar

Current Tools

- Cryptographic Hashing
 - Exact match
 - e.g. MD5
- Fuzzy Hashing
 - Similar ones and zeros
- Ad hoc analysis
- Reverse Engineering

Fuzzy Hashing



A 7 b F d r t 8 5 d N 4 o P 5 k



A 7 b F 9 r t 8 5 d N 4 o P 5 k

Fuzzy Hashing

- Compare signatures:
 - A 7 b F d r t 8 5 d N 4 o P 5 k
 - A 7 b F 9 r t 8 5 d N 4 o P 5 k

Fuzzy Hashing



A 7 b F d r t 8 5 d N 4 o P 5 k



A j b F 9 2 b 5 @ N q o P Y k

- Better algorithm for matching similar files
- Developed by Dr. Vassil Roussev, University of New Orleans
- Like ssdeep, ignores file type data
- Variable sized hashes
 - About 3% of input size
- Handles data reordering
- Matches more files than ssdeep
 - There is not, technically, “better”
- Code, paper, roadmap:
 - <http://roussev.net/sdhash/>

sdhash



Similar Programs

- Similarity depends on:
 - The kind of things be compared
 - How they're being compared
- What makes programs similar?

Similar Programs

- Do the same thing
- Have the same look and feel
- Connect to the same servers
- Written by the same person
- Used in the same intrusion
- Run on the same platform

Features

- What features can we extract from a program?

Features

- Signed code?
- Which APIs are called
- How often APIs are called
- Order in which APIs are called
- Entropy
- DLLs used
- Percentage of code coverage
- Magic strings
- N-grams of instructions
- Control-flow graph
- IP addresses accessed
- ...



Image courtesy of Flickr user doctor_keats and used under Create Commons license.

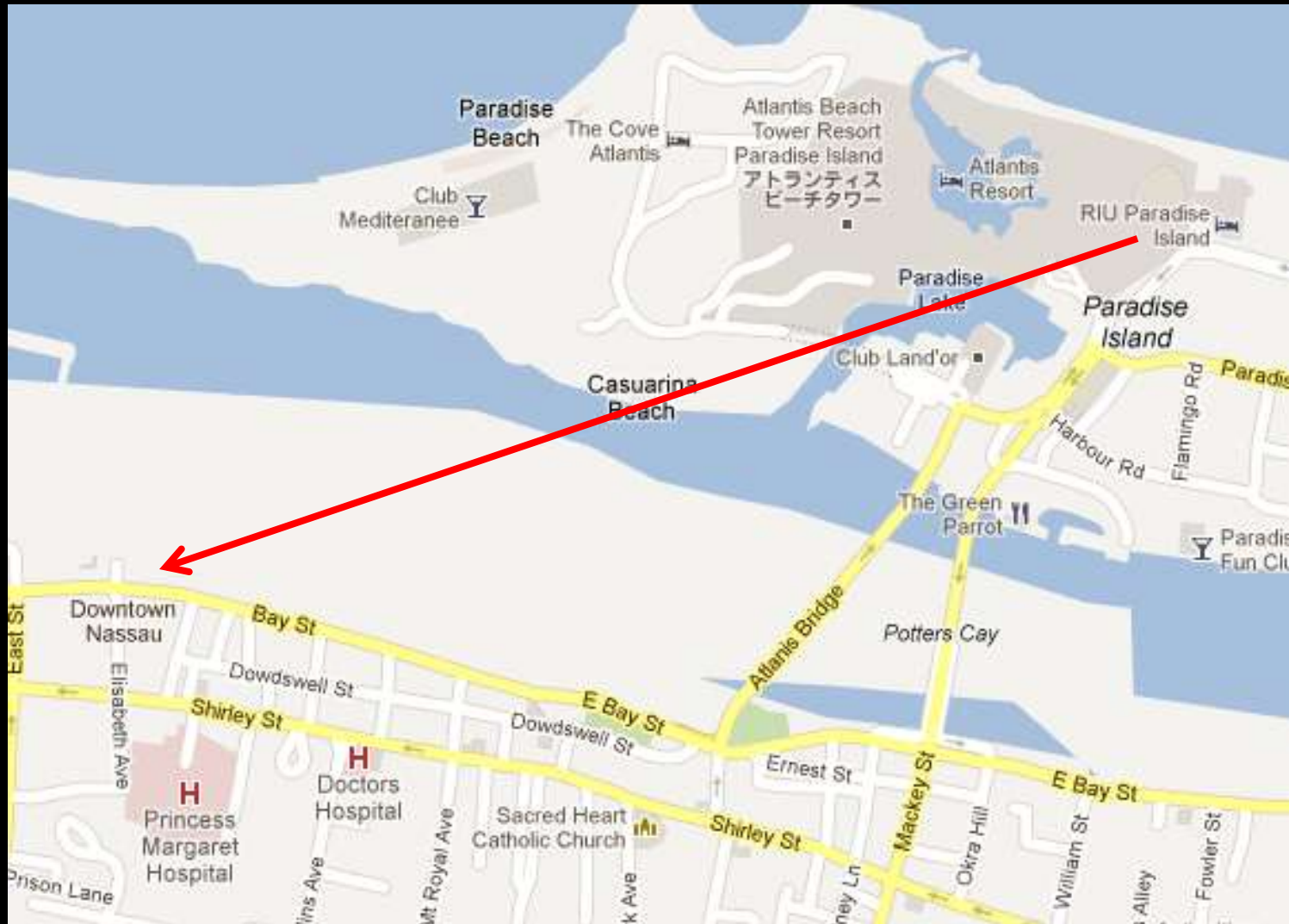
Distance Measures

- Clustering
 - Group of inputs which are close to each other
- Closeness depends on distance measure
- What it sounds like
 - How far apart are the input programs?
 - As measured by our features
- Alternatively, how similar are they?

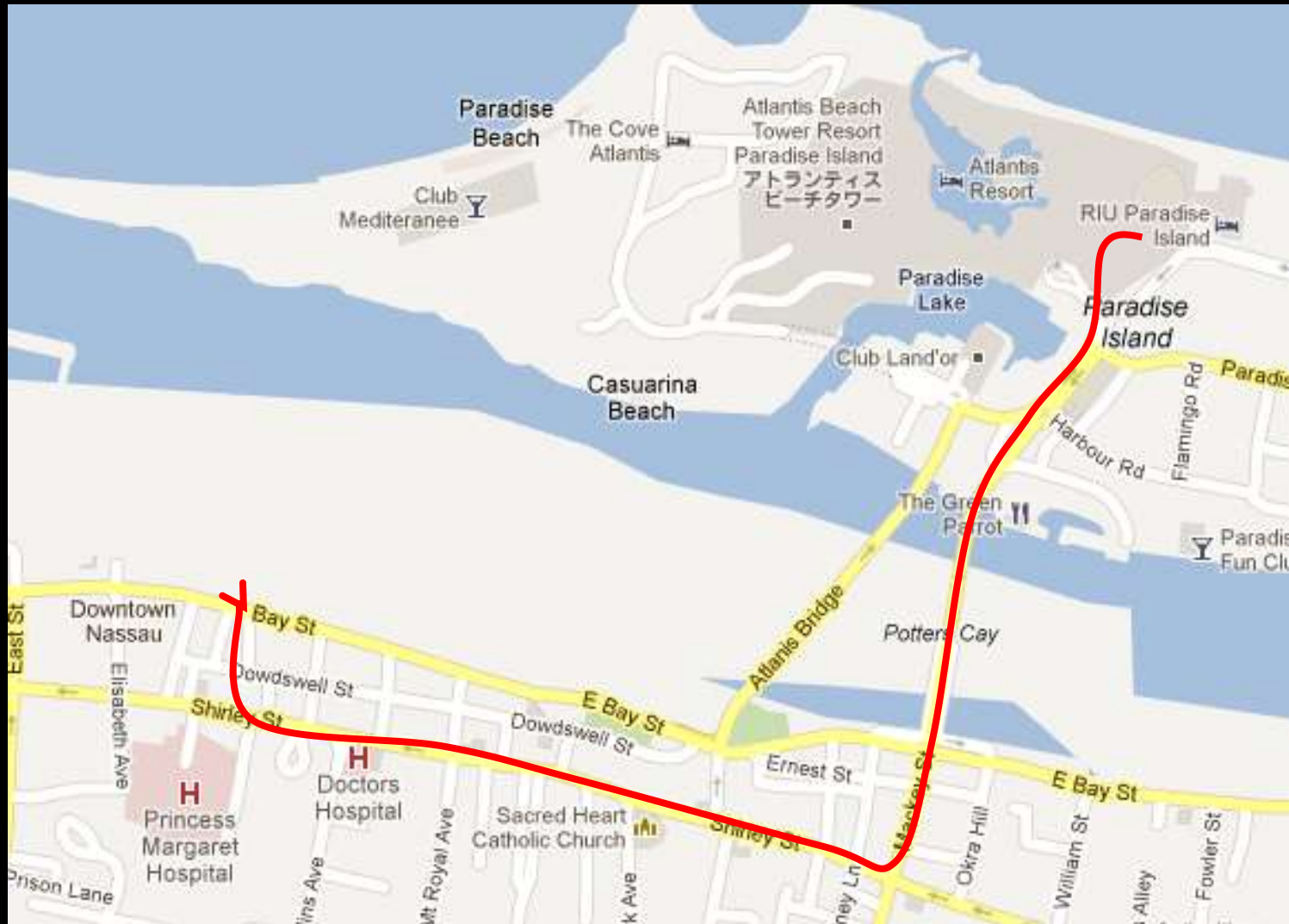
Distance Measures



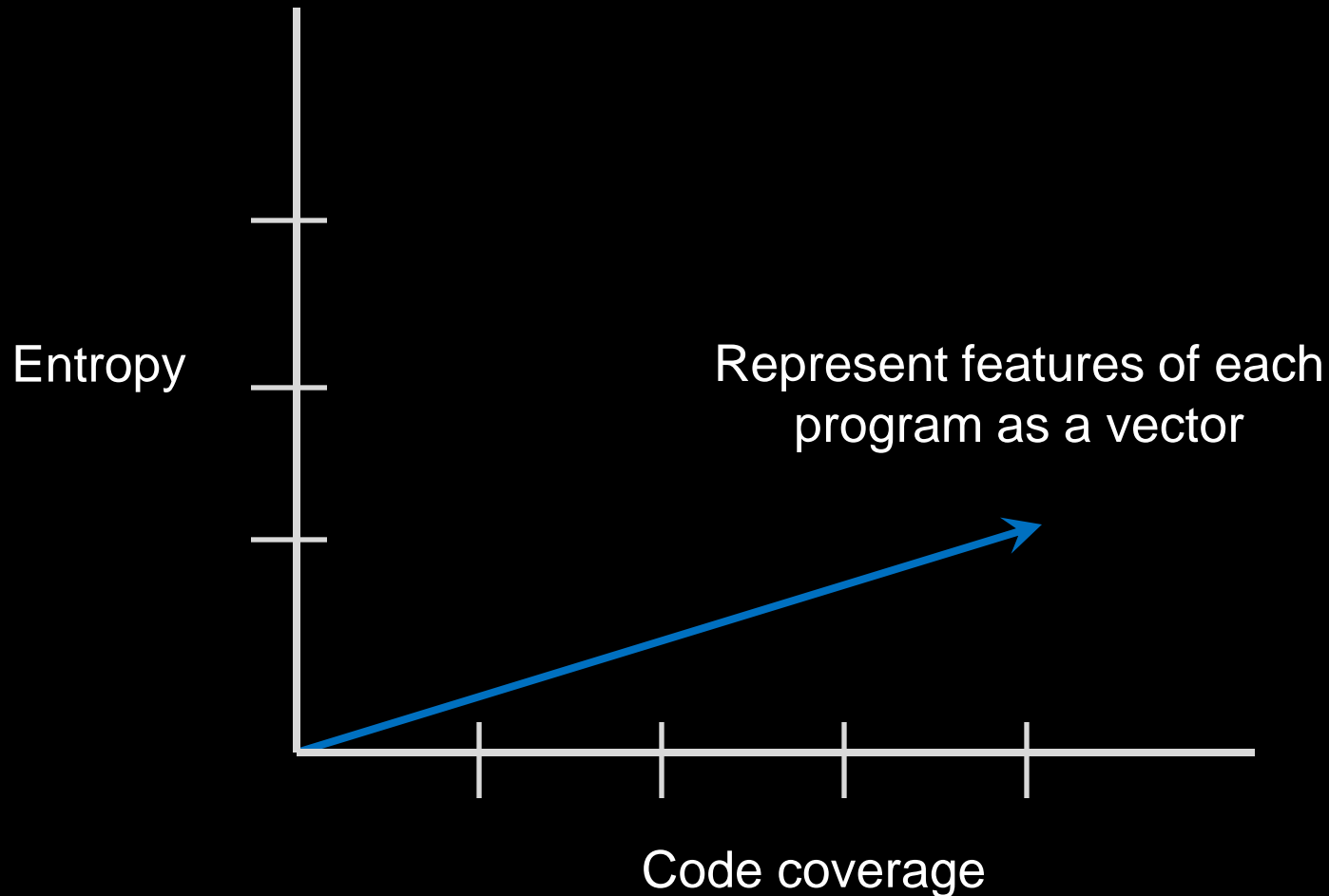
Distance Measures



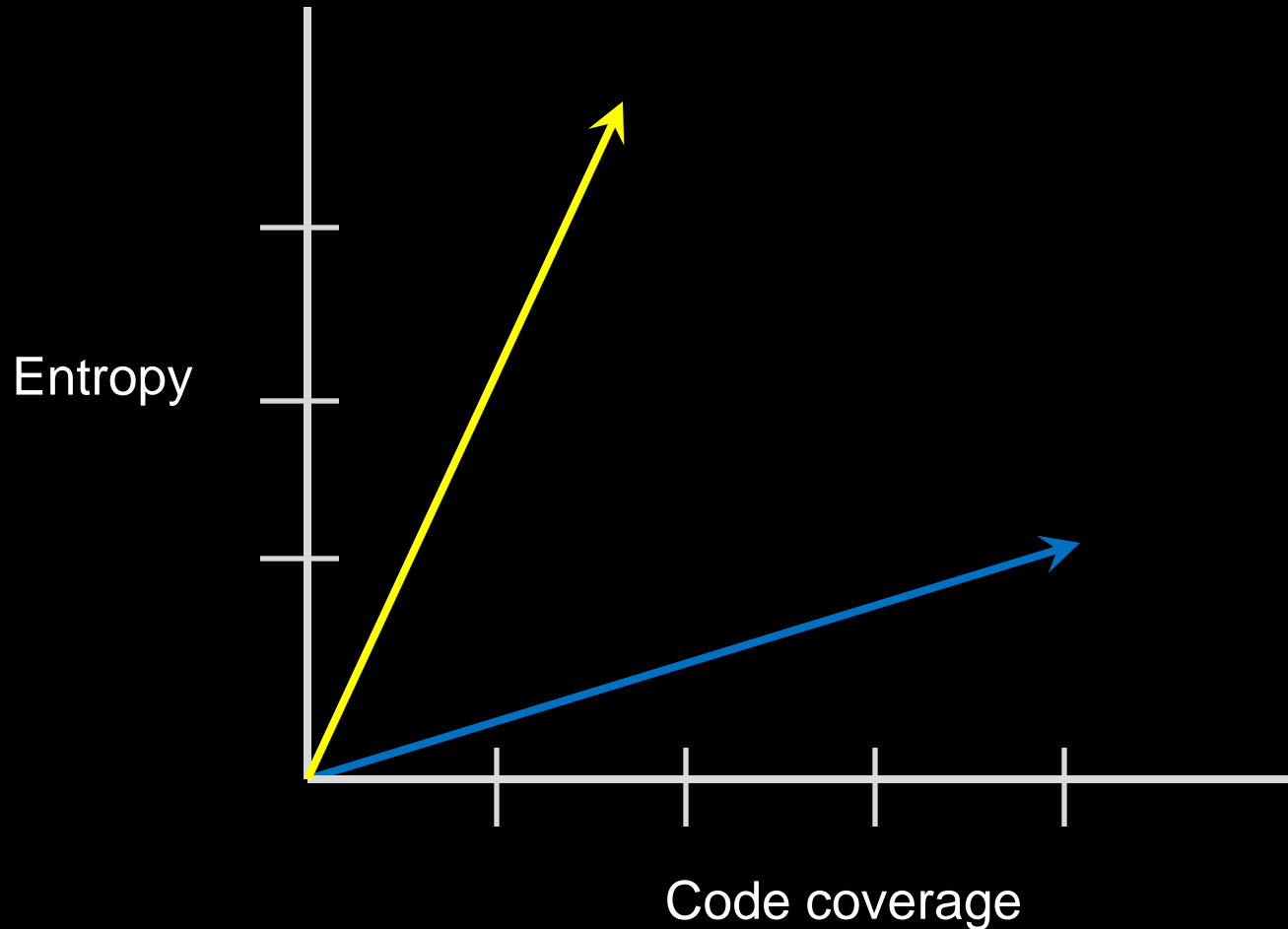
Distance Measures



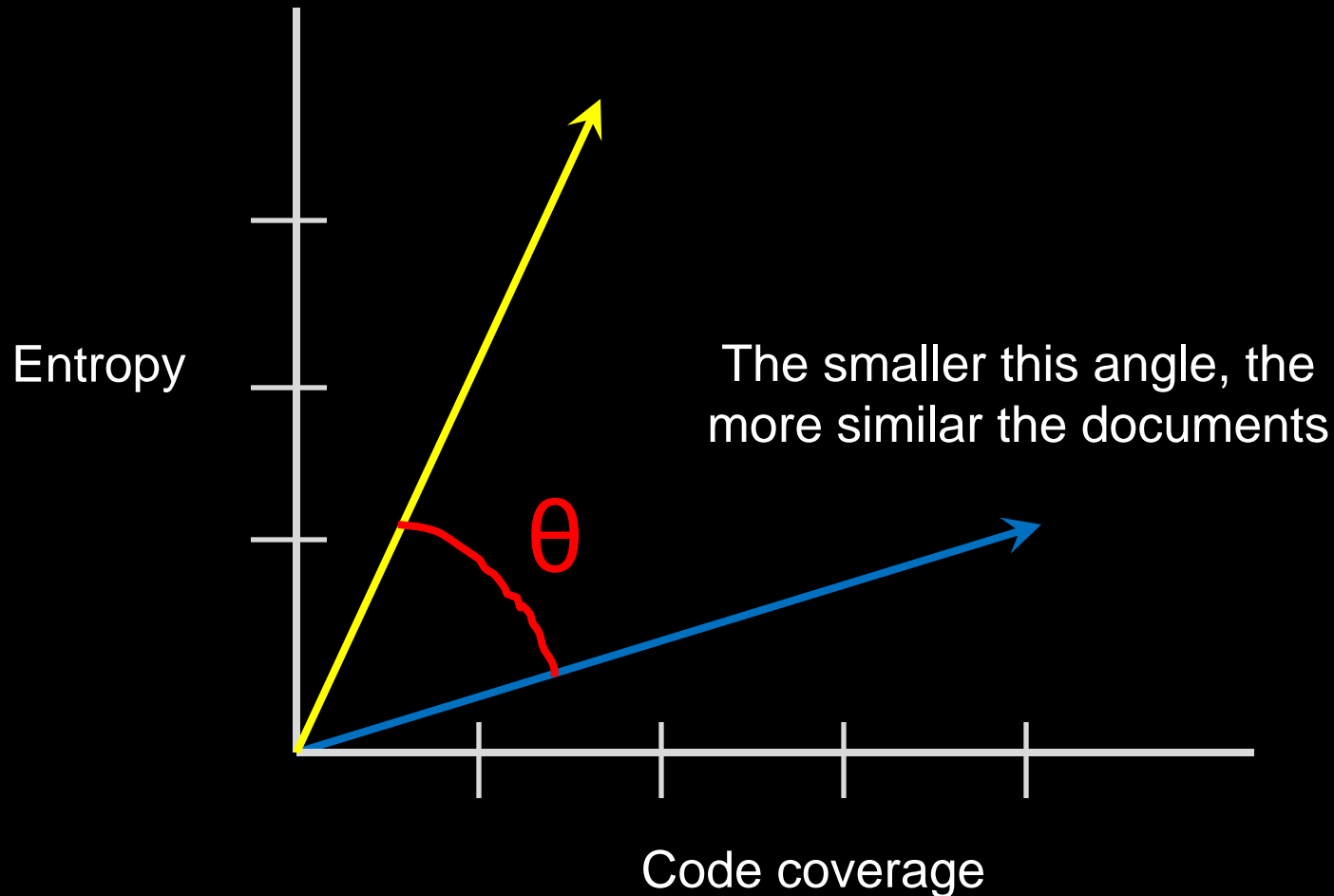
Distance Measures



Distance Measures



Cosine Similarity



Feature Selection

- The Curse of Dimensionality
 - So many dimensions (features) that comparisons become too time consuming or too complex
- No problem
- Select the “important” features
 - (Insert mathy stuff here)
- Example:
 - Presence of crypto constants
 - Depends on context

Comparisons

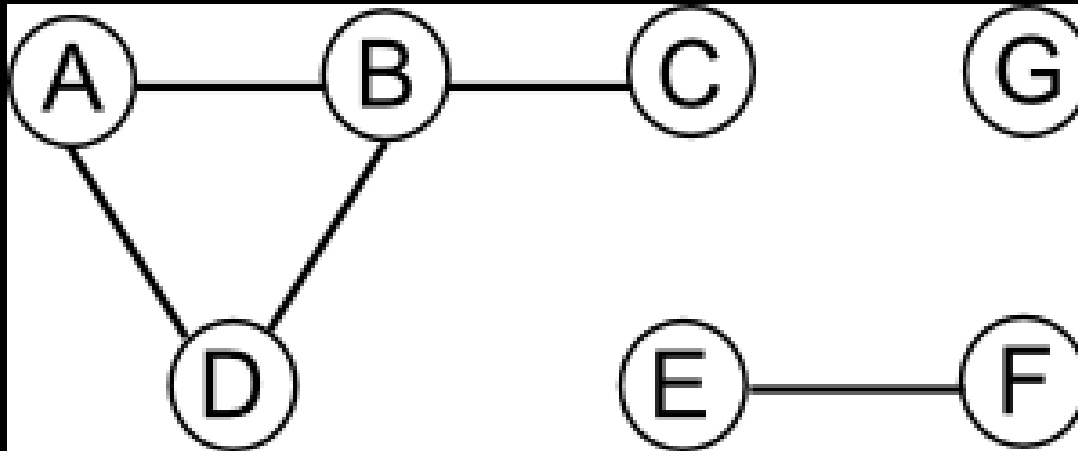
- Can find programs similar to any query
- Similar to a kind of fuzzy hashing
 - “Signature” is the set of selected features

Clustering

- Can find clusters of similar programs
 - Unsupervised machine learning
 - Artificial intelligence
 - There are many algorithms
- Start with pile of programs
- Press “go”
- End up with clusters of similar programs
- Example:
 - Programs A, B, C, D, E, F, and G

Exclusive Clusters

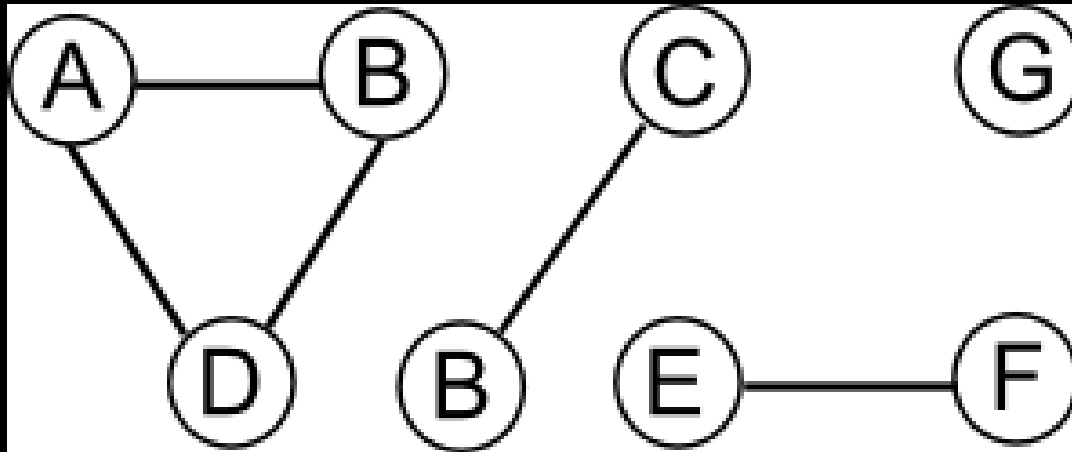
- Each program belongs to at most one cluster



- Not all programs in a cluster are similar to each other
- Some programs are not similar to any others
 - Unique programs

Non-Exclusive Clusters

- Each program can belong to any number of clusters



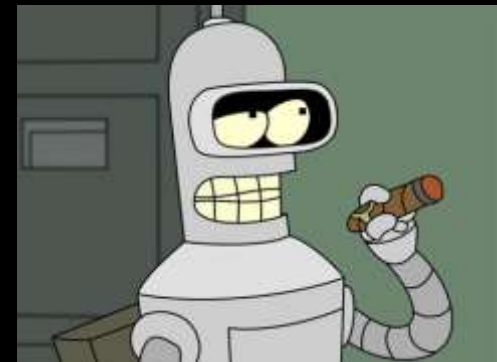
- Every program in a cluster is similar to the others

Clustering vs. Classification

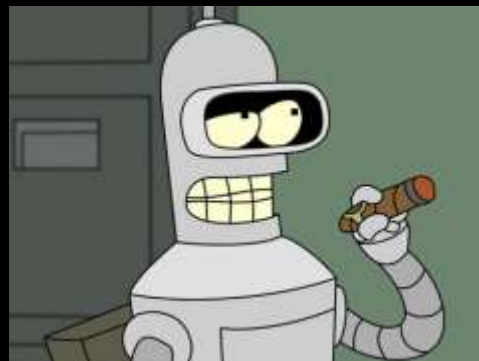
Clustering

Classification

What Makes Things Similar?



Which Things are Similar?



Classification

- Also known as:
 - Predictive Coding
 - Assisted Machine Learning
- Choose all programs which belong in my group
 - Zeus variant or Not Zeus variant
 - Written by Jesse, Written by Evil Hacker, Unknown

Classification

- User must create a set of training data
- Must identify some documents for each possible outcome
- The more the better

Classification

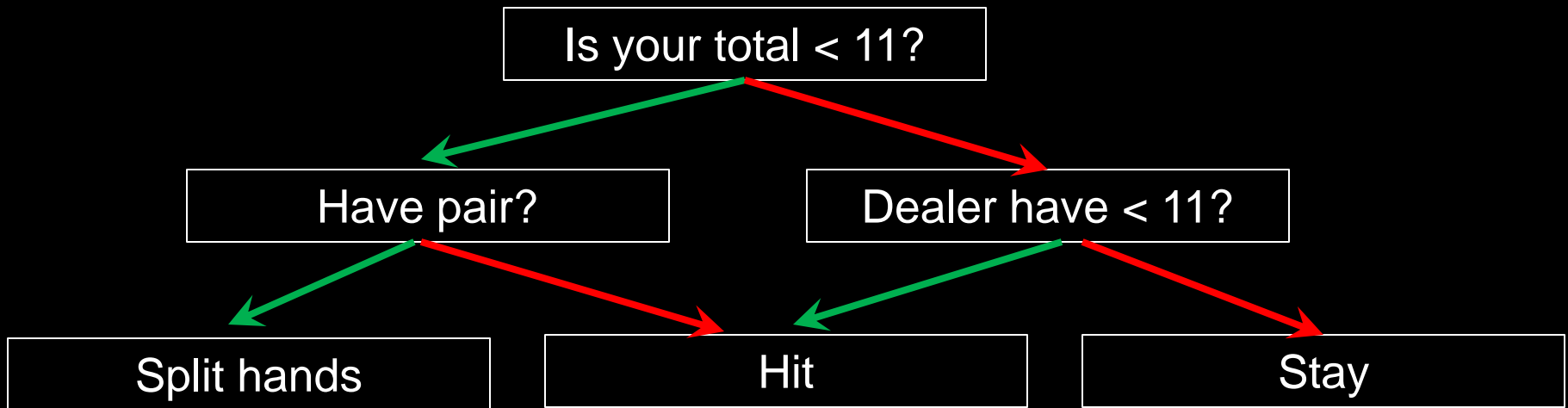
- Artificial intelligence is just math
- There are many algorithms:
 - Naïve Bayesian classifier
 - K-Nearest Neighbor
 - Locality Sensitive Hashing
 - Decision Trees
 - Neural Networks
 - Hidden Markov Models
- See Wikipedia article on Classification (machine learning)

Naïve Bayesian Classifier

- Also used for spam detector
 - Also a classification problem
- $P(B \text{ given } A) = (P(B) * P(A \text{ given } B)) / P(A)$
- Email contains features (words):
- $P(\text{spam given features}) = P(\text{spam}) * P(\text{features given spam}) / P(\text{feat})$
- $P(\text{notspam given feat}) = P(\text{notspam}) * P(\text{features given not}) / P(\text{feat})$
- Which probability is greater?

Decision Tree

- Build a flowchart of questions on the features
- Each question should divide the data equally
- Blackjack example:



Decision Tree

- Quick to classify, but slow to construct
- What questions are best at which point in the tree?
- [Insert mathy stuff here]
- You could make a career out of efficient decision tree generation
 - And people do

Classifier Performance

- There are several measures of classifier performance
- Precision and Recall
- Receiver operating characteristic
 - Aka ROC curve
- Confusion matrix

Precision and Recall

- Precision measures false positives
 - $P = TP / (TP + FP)$
- Recall measures false negatives
 - $R = TP / (TP + FN)$
- Both are on a scale from zero to one
 - One being perfect

Classifier Performance

- If you're not happy with the performance, you can:
 - Add more training values
 - (easy)
 - Change feature selection
 - (moderate)
 - Change features
 - (difficult)
 - Change algorithms
 - (PITA)

Classification Packages

- All of these are Free and Open Source:
 - Weka
 - Apache Mahout
 - Malheur
 - LibSVM

- Which is the best?

Classification Systems

- Academia
 - “Solved problem”
- eDiscovery
 - They love this stuff
 - Predictive coding
 - Features are n-grams of text
- For you?
 - Some assembly required
 - Your Agency puts it together

Conclusion

- Analyzing an infinite number of programs
 - Only five minutes per sample
 - **Computer time is cheap**
- Which of these programs are similar to each other?
 - **Build clusters of programs**
- Which of these programs fit into existing categories?
 - Variant of {Zeus|Spybot|Blackhole}
 - Written by Evil D. Hacker?
 - Related to the last intrusion?
 - **Build classifiers for these categories**

Outline

- Introduction
- Similarity
- Features
- Distance Measures
- Feature Selection
- Clustering
- Classification
- Questions

Questions?



Jesse Kornblum
jesse.kornblum@kyrus-tech.com