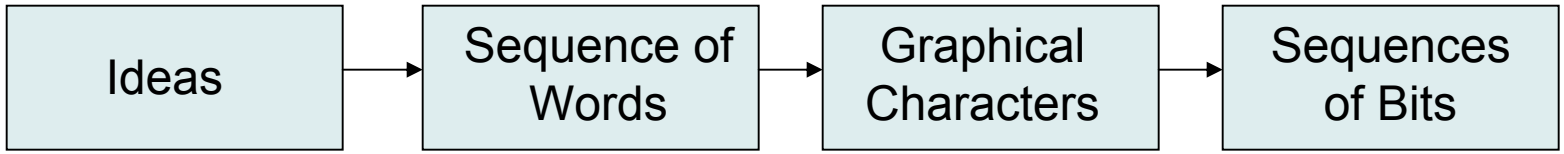
Mining Unstructured Text at Gigabyte per Second Speeds

Alan Ratner

Aim

- Identify language and topic of streaming web documents
  - blogs, newsgroup posts
- In the presence of noise
  - bytes of HTML & JavaScript >> bytes of content → everything looks like English
- At network speeds
  - this year at 2.5, 10, & 40 Gb/s
  - next year at 100 Gb/s

The Nature of Text



Topic

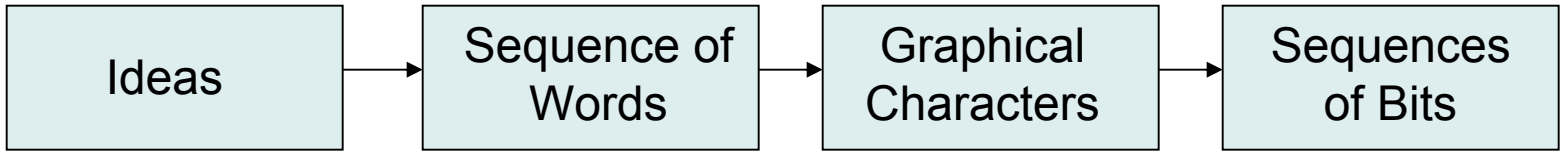
Language

Script

native & transliterated

Encoding

The Nature of Text



	Topic	Language	Script native & transliterated	Encoding
Variations:	>>10^6	4,000	29+	ASCII, WCP, ISO, UTF, language-specific

6 Fundamental Questions in Mining Text

- **Language**
  1. Detection: Is there language?
  2. Identification: What language is it?
  3. Clustering: Is this language like any other I've seen?
- **Topic**
  4. Labeling: How can I describe the topic?
  5. Identification: Is it on a specific topic of interest?
  6. Clustering: Is this topic like any other I've seen?

Agenda

- Computing Platforms ◄
- Language Detection
- Language Identification
- Topic Identification

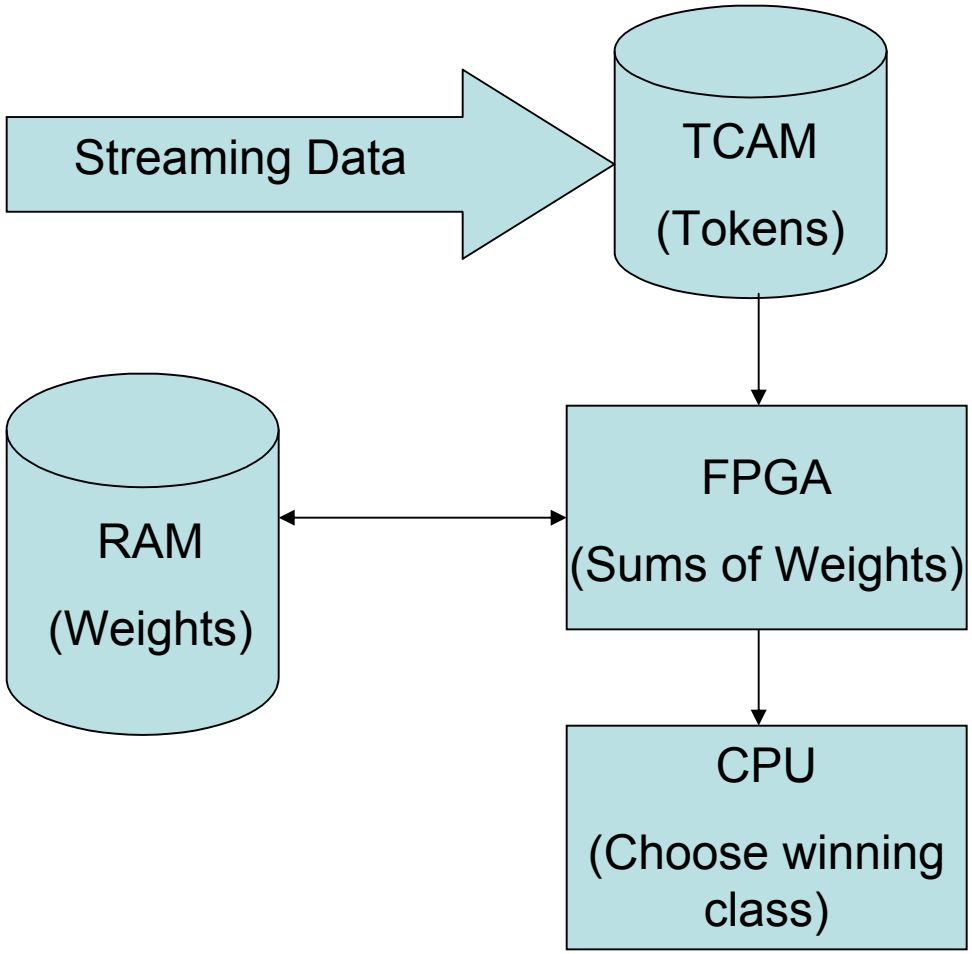
High Speed Processing

- Processing time
  - O(nanosecond per byte)
  - O(microsecond per document)
- Platforms
  - FPGA*
  - CAM*
  - NPU
  - GPU
  - N-core CPU (N >> 1)

Content Addressable Memory

- If string in memory then return address
- CAM is typically ternary (base 3)
  - specified bits/bytes can be ignored
  - ignoring the 6^{th} bit in many encodings makes the comparison case-insensitive
- **10^{13}** 18-byte string compares per second

Architecture



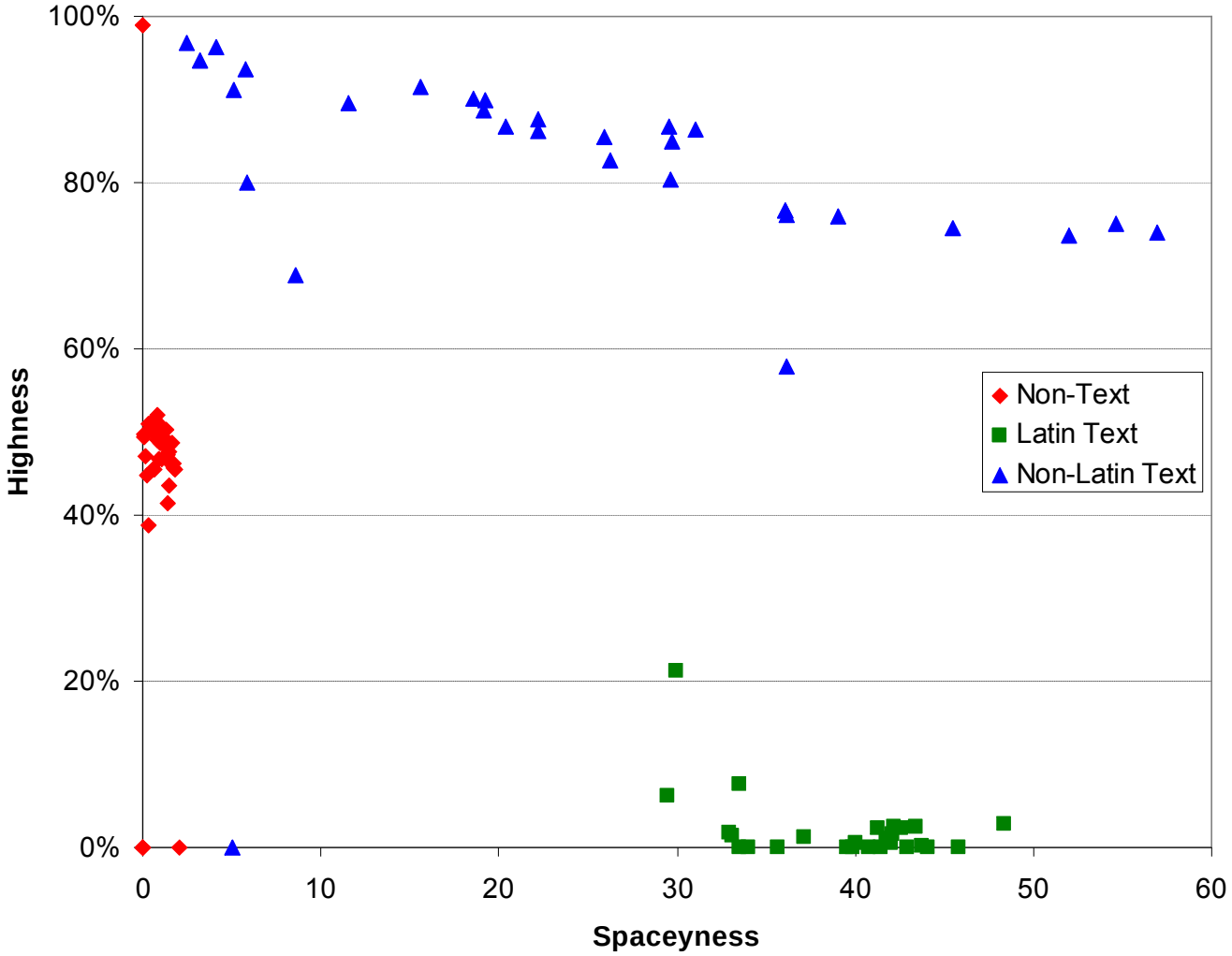
Agenda

- Computing Platforms
- Language Detection ◄
- Language Identification
- Topic Identification

Does a Given File (or Streaming Data) Contain Language?

- Language ≡ vocabulary AND grammar
- Shannon showed we can correctly guess the next character in English text using a 27-letter alphabet 69% of the time
- Text is typically more predictable than well-compressed non-text (zip, audio, video, images)

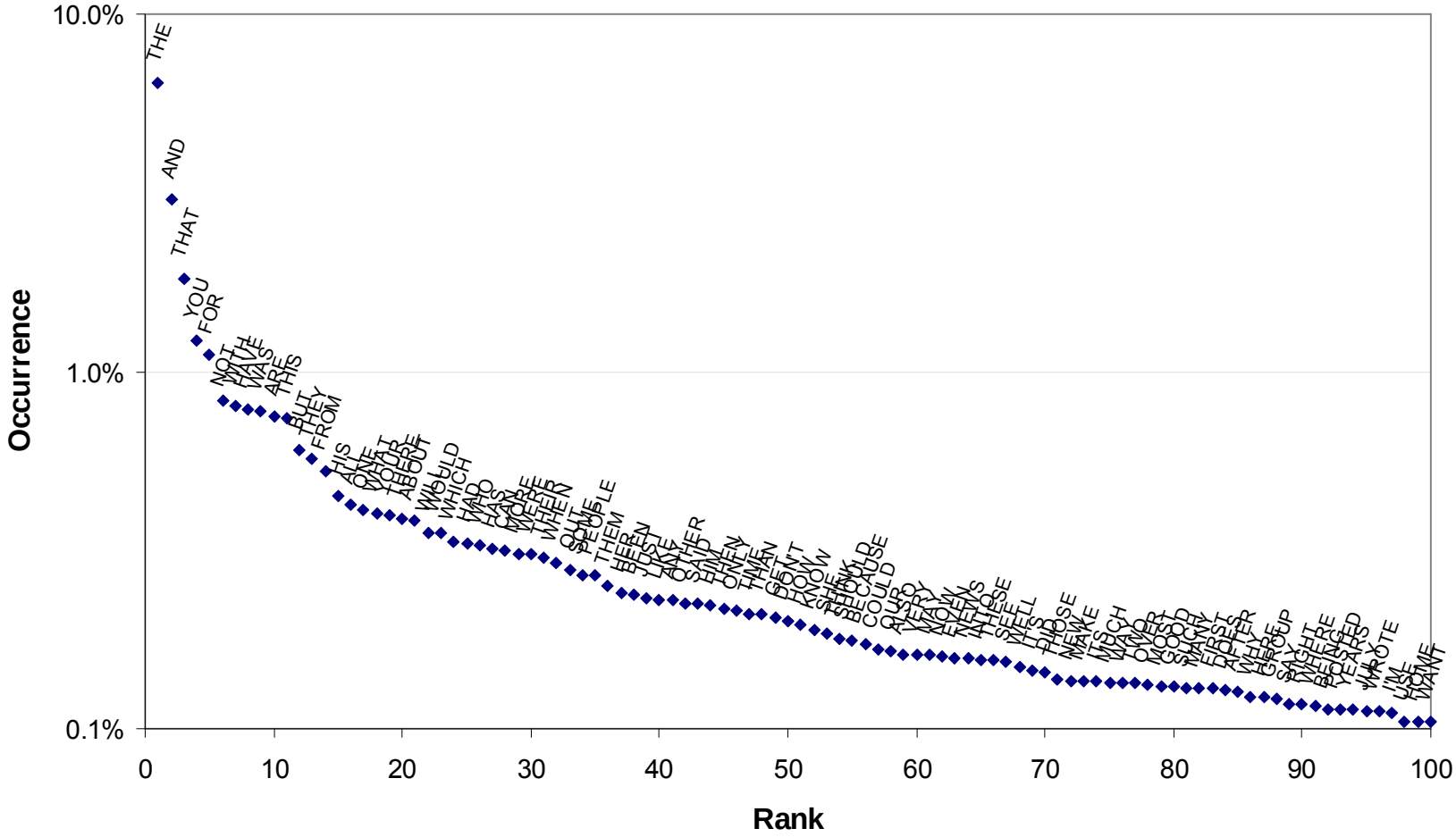
Spaceyness vs. Highness
(on noiseless text)



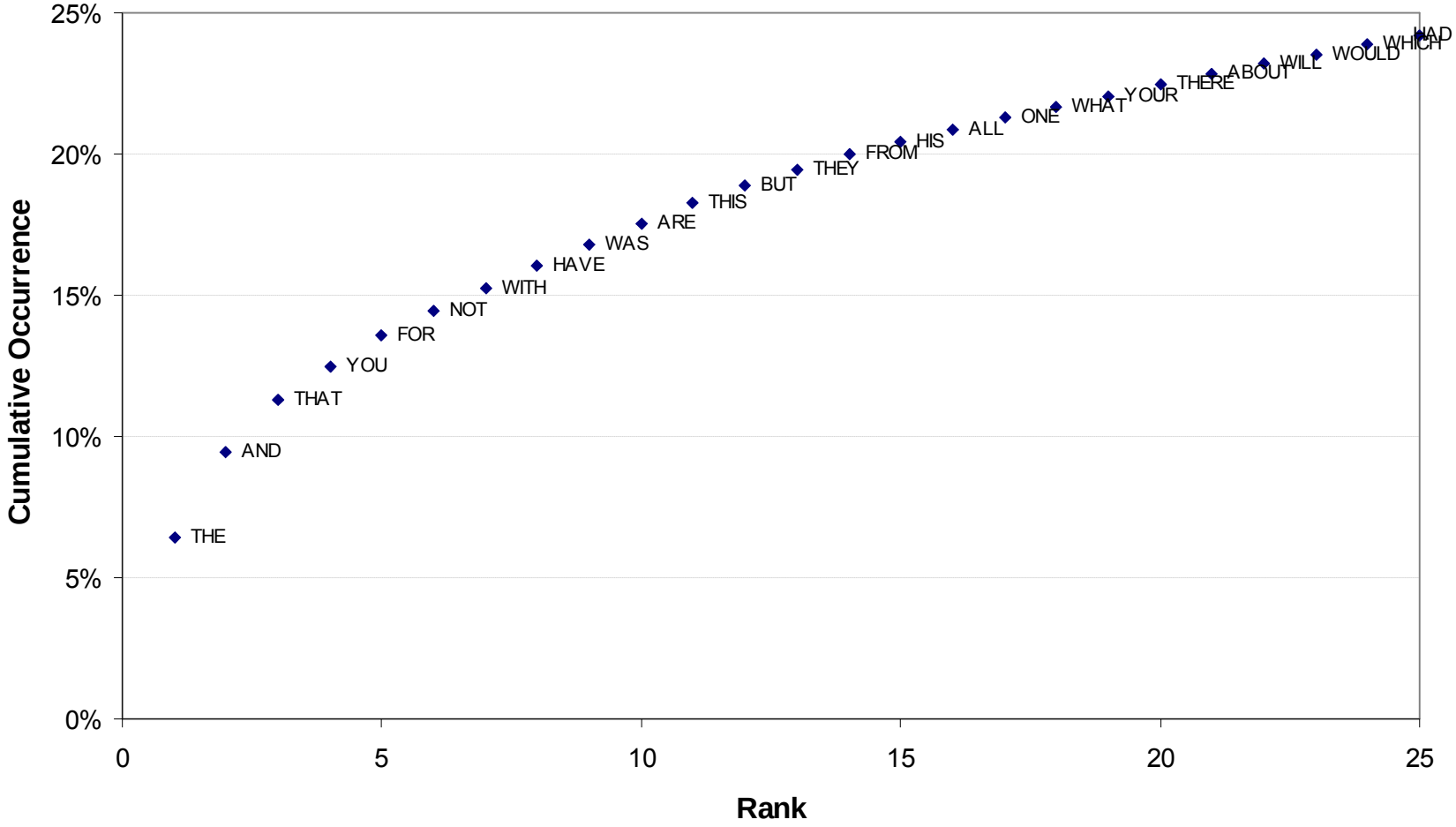
Agenda

- Computing Platforms
- Language Detection
- Language Identification ◄
- Topic Identification

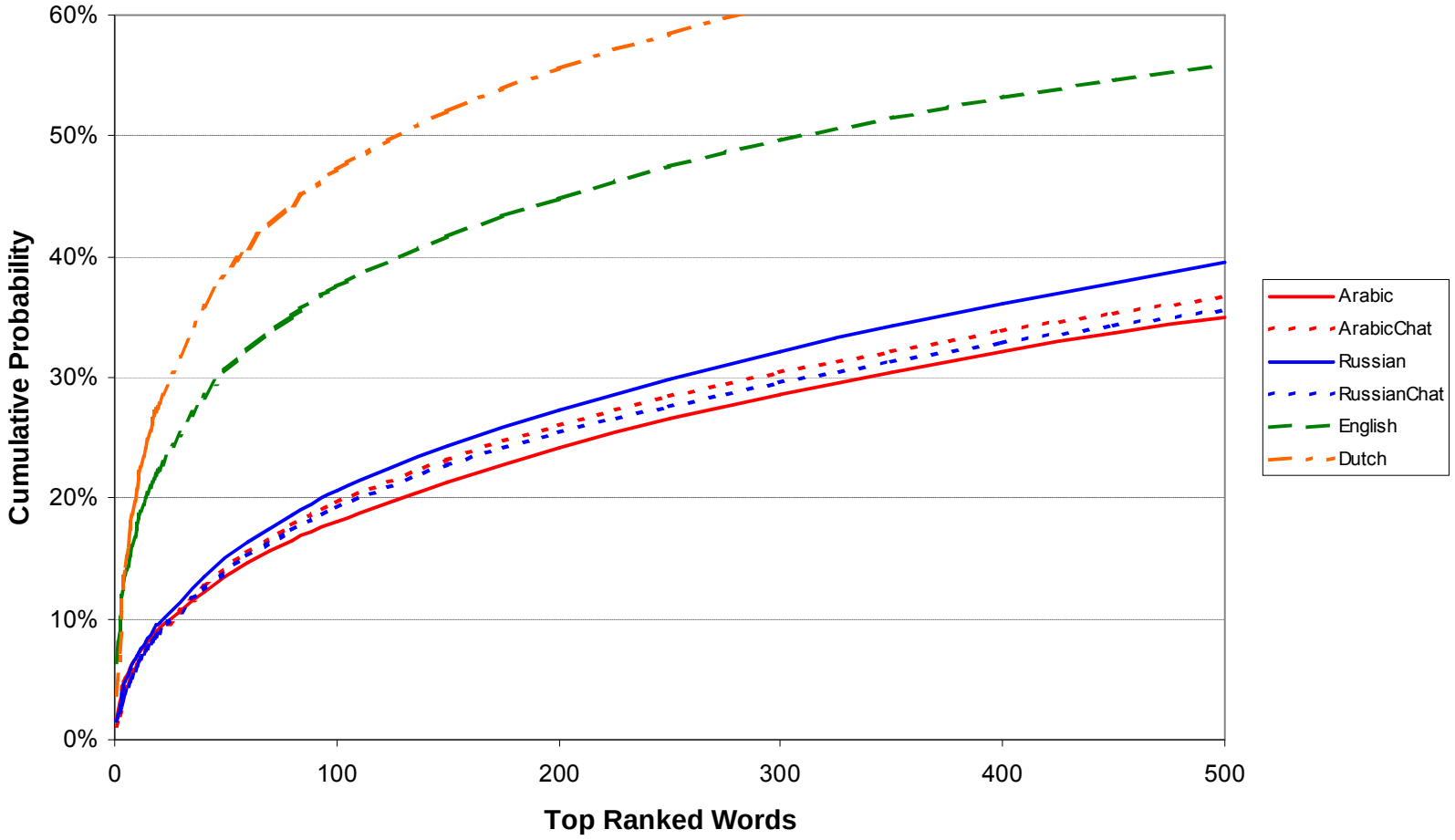
English Word Occurrence
(>=3 letters)



English Cumulative Word Occurrence



Comparative Cumulative Word Occurrence



Selecting Patterns for Language Id

- Spaced languages
  - Common words ≥ 3 characters with spaces before & after
  - Avoid
    - HTML keywords
    - JavaScript keywords
    - International words (proper nouns)
- Unspaced languages
  - Common character pairs

Agenda

- Computing Platforms
- Language Detection
- Language Identification
- Topic Identification ◄

Topic Id & Search

- How do you find the 10-100 relevant docs to satisfy a user's need for information (out of 10^{12} candidates)?
- Search engines force keyword search onto users
- What users generally want is topic id
  - "Show me docs on topic X"
  - "Show me docs that look like this and not like this"
- False negatives OK
  - User doesn't know what you have missed (unlike anti-spam)
- Precision needs to be high
  - Don't annoy the user!
  - Extremely low false positive rate (0.0000001%)

Algorithms for Topic ID

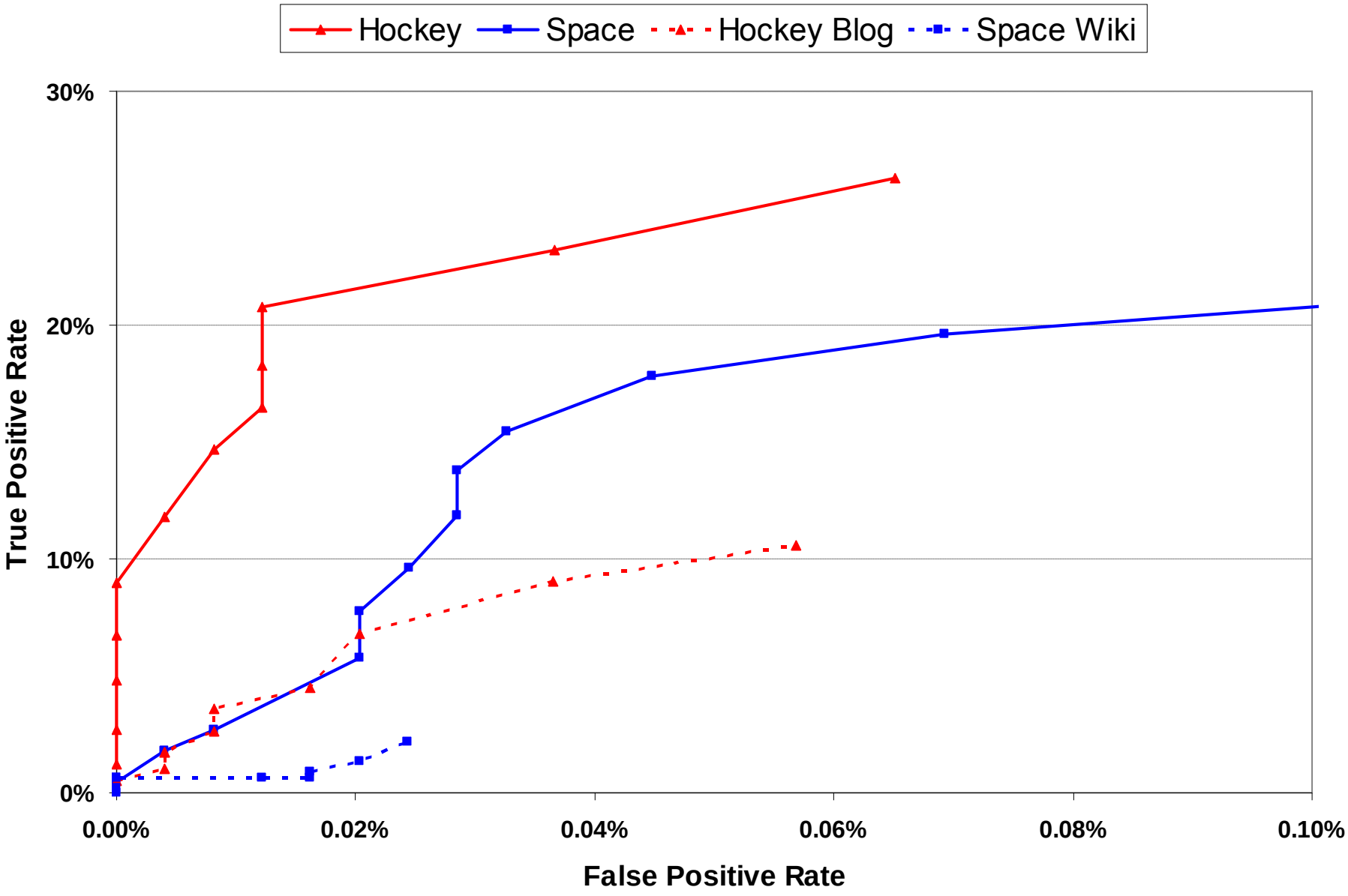
- Bayes
- Markov
- Markov Orthogonal Sparse Bi-word
- Hyperspace
- Correlative
- Entropy
- Term Frequency * Inverse Doc Frequency
- Minimum Description Length
- Morphological
- Centroid-based
- Logistic Regression (similar to SVM & single-layer NN)
- …
- Most of these use either additive or multiplicative weights of detected tokens

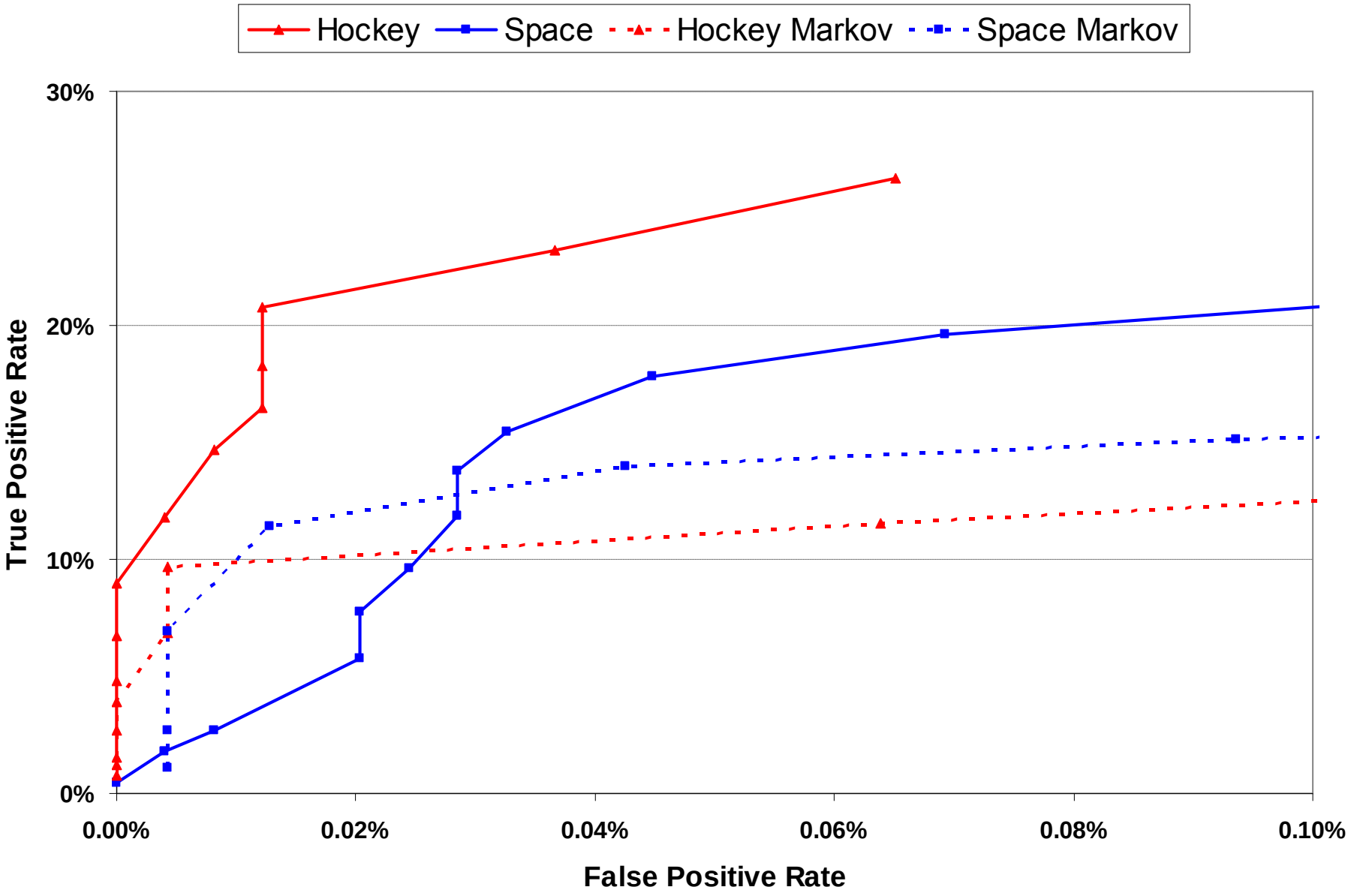
Topic ID Implementation

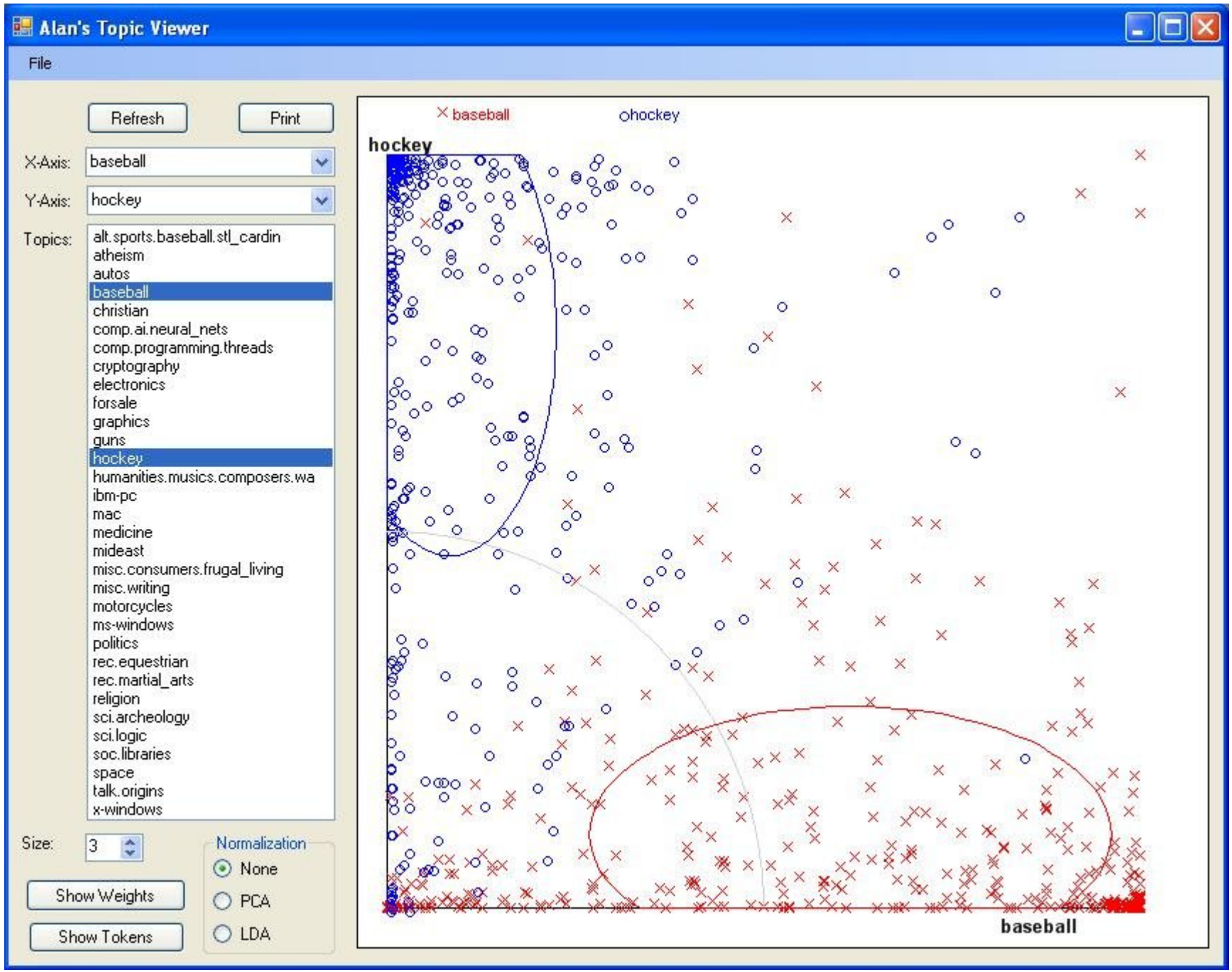
1. Select on-topic & off-topic docs for training
2. Select tokens (typically 1K on-topic/10K off-topic)
3. Compute additive weights that minimize false positives with acceptable recall

Topic Id Evaluation

- Consistent labeling of informal documents by topic is difficult
- Need millions of labeled docs to cover major languages and wide variety of topics
- Used 30K newsgroup posts on 32 topics
  - Topic assumed to be name of newsgroup
  - Considerable overlap in topics, especially those on sports and religion


Hockey
Space
Hockey Blog
Space Wiki
30%
20%
10%
0%
True Positive Rate
0.00%
0.02%
0.04%
0.06%
0.08%
0.10%
False Positive Rate


Hockey
Space
Hockey Markov
Space Markov
30%
20%
10%
0%
True Positive Rate
0.00%
0.02%
0.04%
0.06%
0.08%
0.10%
False Positive Rate


Alan's Topic Viewer
File
Refresh
Print
X-Axis:
baseball
Y-Axis:
hockey
Topics:
alt.sports.baseball.stl_cardin
atheism
autos
baseball
christian
comp.ai.neural_nets
comp.programming.threads
cryptography
electronics
forsale
graphics
guns
hockey
humanities.musics.composers.wa
ibm-pc
mac
medicine
mideast
misc.consumers.frugal_living
misc.writing
motorcycles
ms-windows
politics
rec.equestrian
rec.martial_arts
religion
sci.archeology
sci.logic
soc.libraries
space
talk.origins
x-windows
Size:
3
Normalization
None
PCA
LDA
Show Weights
Show Tokens
baseball
hockey
hockey
baseball

Conclusions

1. Language detection solved on clean documents but unsolved on noisy web documents
2. Language identification works well at gigabyte per second speeds
3. Topic identification should work well at gigabyte per second speeds; further testing needed on larger data sets
4. Topic clustering needs further work