15-492 / 11-682:
Introduction to IR, NLP, MT, and Speech

Text Categorization

Jamie Callan
Carnegie Mellon University
callan@cs.cmu.edu

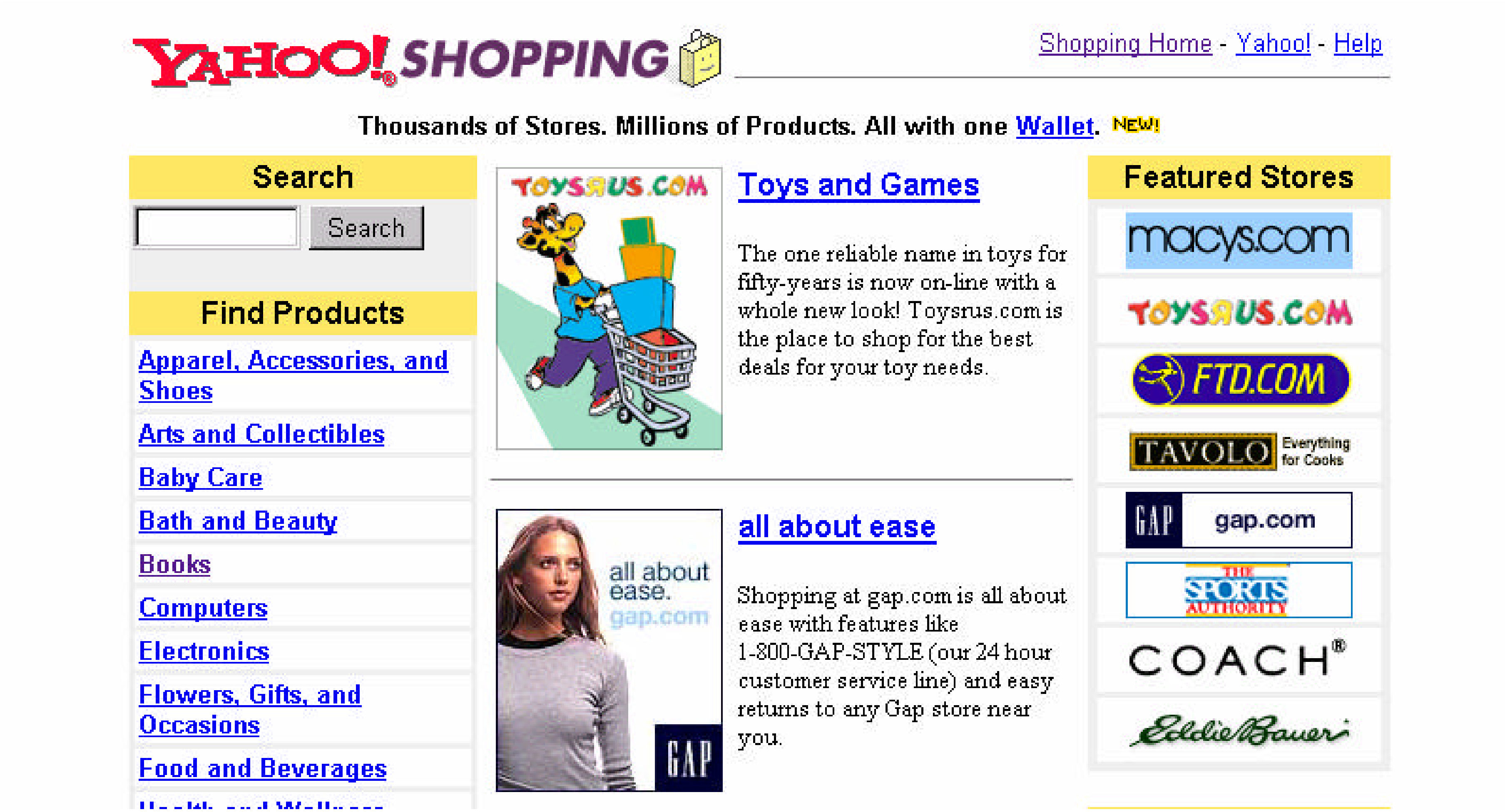
Outline

- **Introduction to text categorization**
- **Manual categorization**
- **Automatic categorization**
  - Algorithms
  - Training data

Automatic Text Categorization: Introduction

- **Categorization:** Assigning labels to objects
  - One label per object
  - Multiple labels per object
- **Automatic text categorization:** Labels assigned by computer
  - Lower cost
  - Greater consistency
  - Maybe greater accuracy (or, maybe not)
- **Class labels are equivalent to controlled vocabulary terms**
  - A form of metadata
- **Text categorization is an “old” research area**
  - Renewed interest due to growing use of electronic documents

Document Classification



Document Classification

YAHOO! SHOPPING

Shopping Home - Yahoo! - Help

Search Result: Found **335** products in **90** stores for "**MP3 players**" Shopping Home

Merchants: 3 matching 'MP3 players':

- MP3's from i2Go.com: eGo - Interactive portable digital audio (**MP3**) player.
- HyCD Store @ Yahoo!: CD Recording software supports **MP3** encoder and **MP3** Player.
- Frontier Labs Online Store: Portable audio digital **MP3** devices and **players**.

Categories: 1 matching 'MP3 players':

- Electronics > Portable Audio > **MP3 Players**

Products: Found 335 products in 90 stores matching 'MP3 players'. Showing stores 1 - 20:

sort listing by: relevance | increasing price | decreasing price

Playstation MP3 Player TOP SERVICE from Core Computer / Core Concepts Inc - (1 match) **$41.00**

Playstation **MP3** Player Brand new device for 1999 / 2000: This device is capable of playing **MP3** CD files on your playstation as well as use cheat codes (from the Gameshark) and play import / CDR backups games. Due to the compression ratio of **MP3** files, you can fit over 100 songs on 1 CD and still...

Text Categorization Examples

- Topic names to newswire publications
- LCSH codes to library materials
- MeSH codes to medical publications
- MeSH codes to Medline queries
- ICD9 codes to patient discharge summaries
- Patent classes to patent applications
- Priority classes to email messages
- Pornography probabilities to Web pages
- Yahoo! subject categories to Web pages
- Individuals to customer support email
- Advertising categories to prospective customers

Approaches to Automatic Text Categorization

- **Classifier:** A process that assigns one or more labels to objects
- **Manual classifier:** A person creates a classifier manually
  - Examples: Email filters
  - Usually rule-based
  - Classifier usually easy for humans to understand
- **Automatic classifier:** A machine learning algorithm creates the classifier
  - Requires a set of documents classified manually (training data)
  - Many algorithms
    - » Rule-based, decision tree, nearest neighbor, EM, Ripper, …
  - Classifier often difficult for humans to understand

Automatic Text Categorization: Rules Created Manually

Manual rules are usually based on intuition and experience

- **Advantages:** Leverage human abilities, meet expectations
- **Disadvantage:** Human classifiers may not provide good recall

V1: machine AND learning

V2: (machine AND learning) OR (neural AND networks) OR
(decision AND tree)

V3: (machine AND learning) OR (neural AND networks) OR
(decision AND tree) AND C4.5 OR Ripper OR EG OR EM

V4: (machine AND learning) OR (neural AND networks) OR
(decision AND tree) AND C4.5 OR (Ripper AND NOT Jack) OR
(EG AND algorithm AND gradient) OR (EM AND NOT printing)

Automatic Text Categorization: Rules Created Manually

- **Human classifiers are based on all of a person's experiences**
  - Manual classifiers are often not corpus-specific
  - Too much effort on patterns that probably won't occur
  - Not enough effort on patterns that make sense only within that corpus
- **Example:** The task is to identify news stories about terrorist events
  - People think of words such as "bomb" and "kill"
  - Those words also occur in stories about wars
  - "broken windows" is highly correlated with terrorist events
- **The human tendency to produce classifiers that "make sense" causes them to miss effective corpus-specific language patterns**

Cost of Manual Text Categorization

- **Yahoo!**
  - 200 (?) people manually labeling Web pages
  - Using a hierarchy of 500,000 categories
- **MEDLINE (National Library of Medicine)**
  - $2 million/year for manual indexing of journal articles
  - Using MeSH headings (18,000 categories)
- **Mayo Clinic**
  - $1.4 million annually for coding patient-record events
  - Using the International Classification of Diseases (ICD) for billing insurance companies
- **U.S. Census Bureau decennial census (1990, 22 million responses)**
  - 232 industry categories and 504 occupation categories
  - $15 million if done completely manually

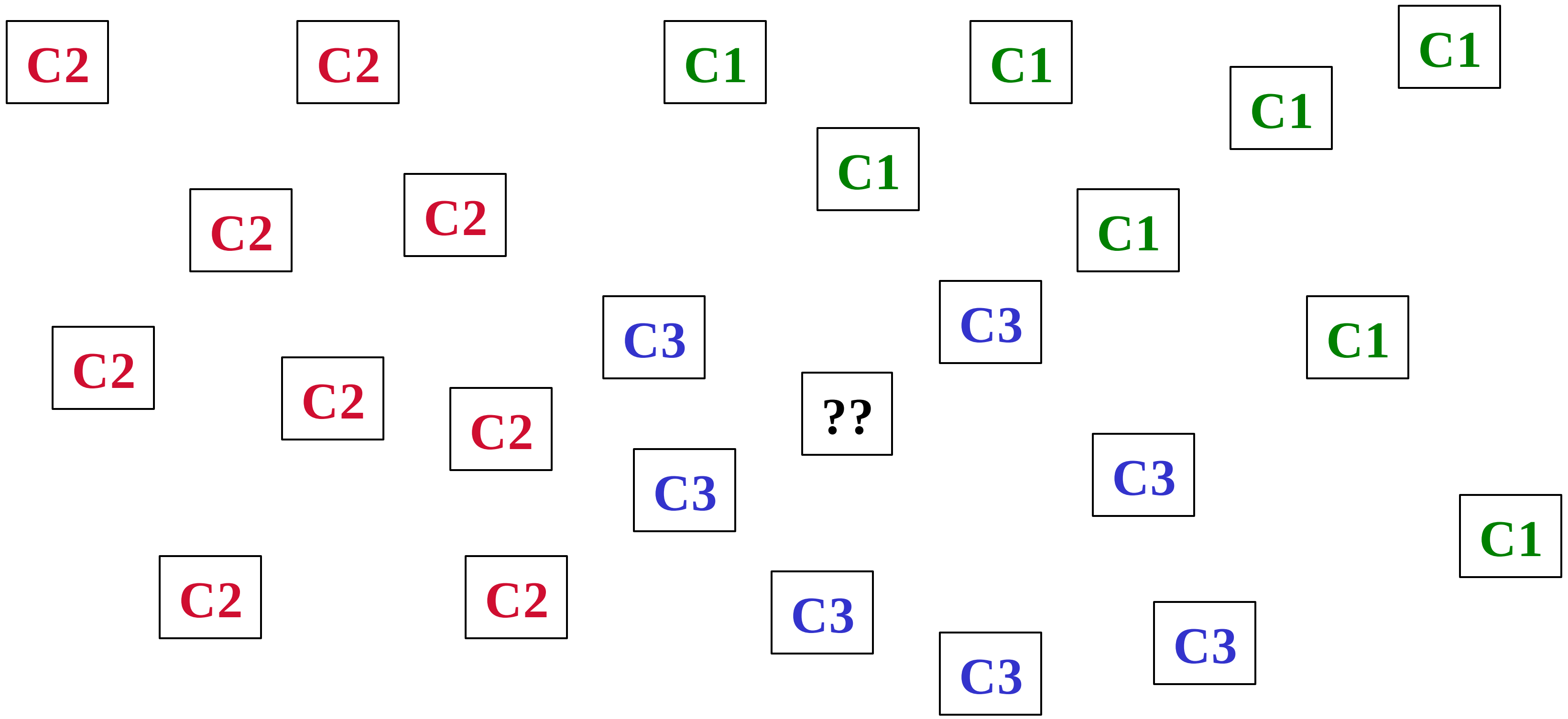
(Yang, 2001)

Automatic Text Categorization: Classifiers Created Automatically

- **A set of training data is provided to a machine learning algorithm**
  – A set of representative objects, with labels
  – The larger the set, the better (usually)
  – The algorithm searches for patterns correlated with each label
  – Patterns are used to create a classifier
- **Good training data is crucial**
  – The labels must be assigned accurately and consistently
  – The objects must be described accurately and consistently
- **How should a text document be described?**
  – By the words it contains
  – By any known metadata (e.g., author, publisher, …)

Nearest Neighbor

To classify a new object, find the most similar object in a training set. Assign the new object the same label(s).



Nearest Neighbor

To classify a new object, find the most similar object in a training set. Assign the new object the same label(s).

- This obviously works well if there is an exact match
- It usually works well if there is a close match
- Generalization: Use the *k* most similar neighbors (KNN)
  - *k*-NN is usually more robust than nearest neighbor (*k*=1)

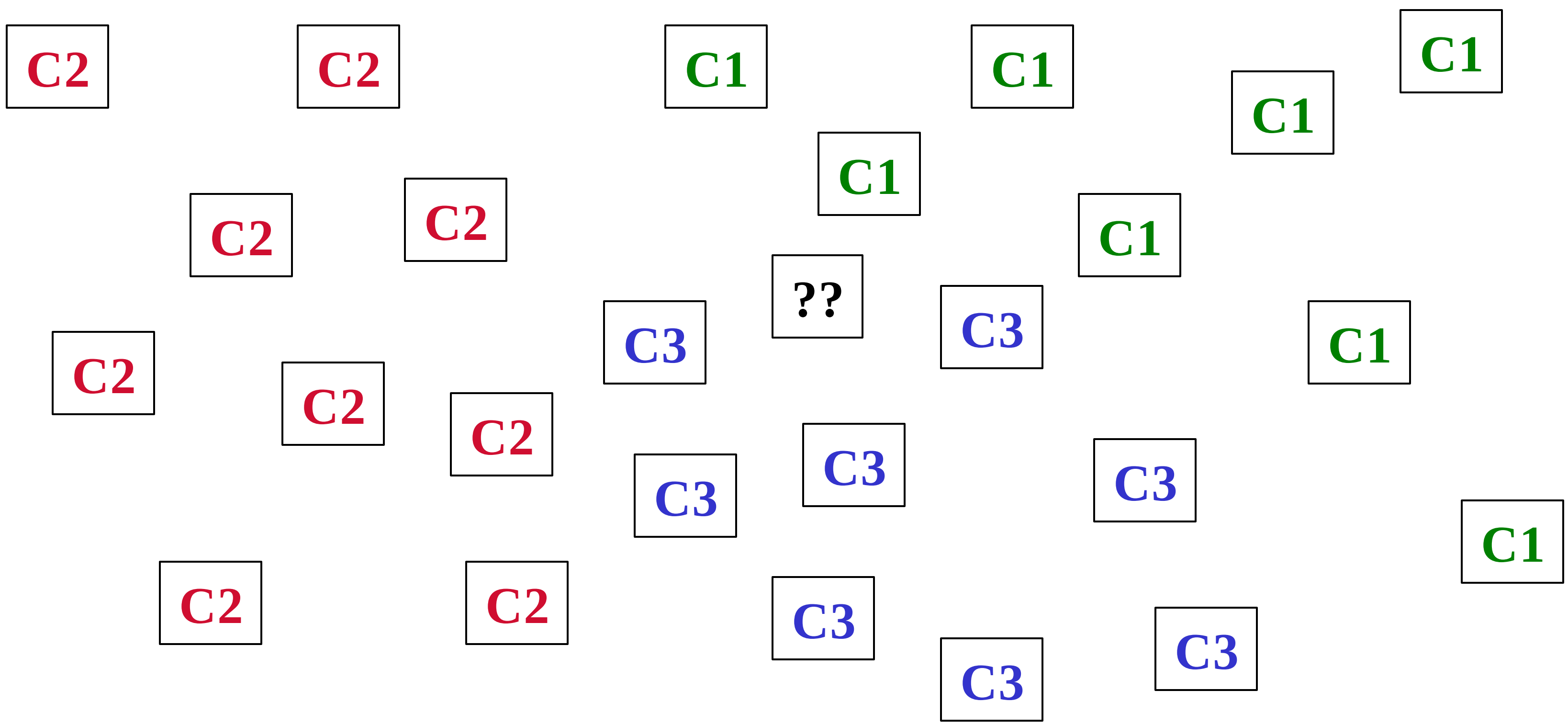
k-Nearest Neighbor (KNN) in an IR Environment

- **Represent each training document as a vector of term weights**
  - E.g., tf.idf
- **Treat new document as a query vector**
- **Retrieve the top k documents**
  - E.g., using cosine similarity as a distance function
- **Score each category associated with any returned document**
  - Returned documents define the neighborhood
- **Apply thresholds to convert scores into yes/no decisions**

(Yang, 2001)

k-Nearest Neighbor

To classify a new object, find the *k* most similar objects in a training set. Assign the new object the same label(s).



k-Nearest Neighbor (KNN): Distance Function

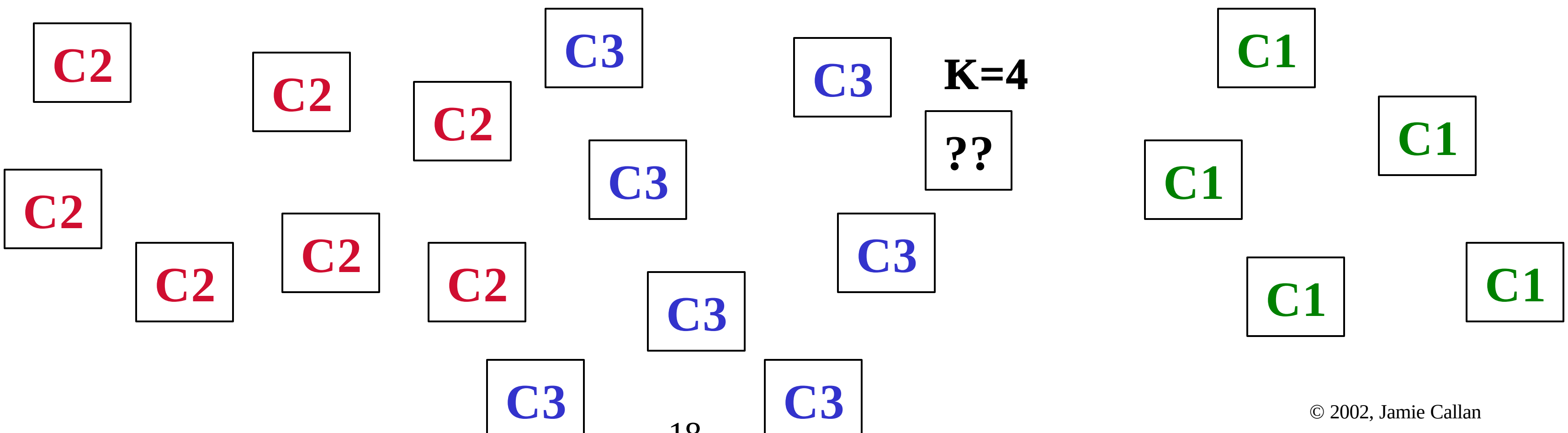
- **It is important to select a good distance function**
  - Often it is not obvious what distance function to use
    - » Selected empirically, tuned empirically
- **For text data, the cosine similarity metric is often effective**
  - Represent each document as a word vector
    - » Scalar values indicates the "weight" of each word
  - Similarity is inversely related to the cosine of the angle between the vectors

$$\cos(\vec{x}, \vec{y}) = \frac{\vec{x} \cdot \vec{y}}{|\vec{x}||\vec{y}|} = \frac{\sum_{i=1}^{n} x_i y_i}{\sqrt{\sum_{i=1}^{n} x_i^{2}} \sqrt{\sum_{i=1}^{n} y_i^{2}}}$$

k-Nearest Neighbor (KNN)

What if the neighbors have different labels?

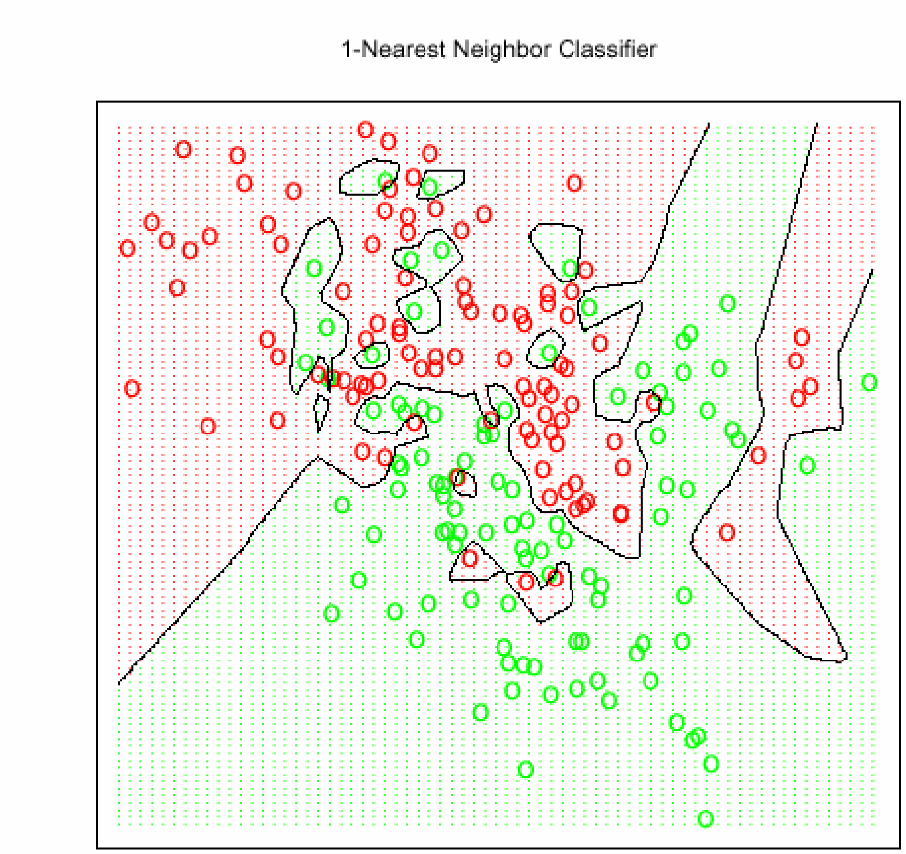
- **Intersection:** Assign only labels that all k neighbors share
- **Union:** Assign any label assigned to any of the k neighbors
- **Voting:** Assign any label assigned to at least t neighbors, t ≤ k
- **The effect of a neighbor may be weighted by its distance**
  - Distant neighbors have less influence than near neighbors



k-Nearest Neighbor (KNN): Choosing k

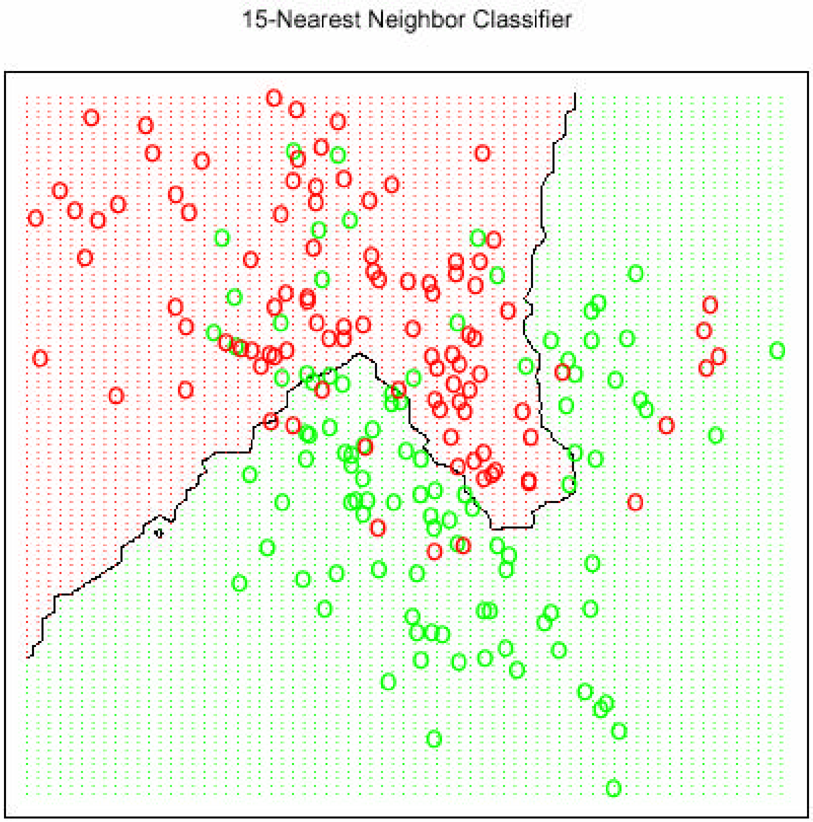
- **k is usually determined empirically, e.g., by cross-validation**
- **Hold out a subset of training data as validation set**
  – Don't use for training or testing
- **For all reasonable values of k**
  – Train on training data
  – Evaluate on validation data
- **Select value of k that gives best value**
- **Test on testing data**

Low bias, high variance for K = 1



Source : Elements of Statistical Learning (2001): Hastie, Tibshirani, Friedman

Higher bias, lower variance with higher K



Source : Elements of Statistical Learning (2001): Hastie, Tibshirani, Friedman

k-Nearest Neighbor (KNN): Summary

- **KNN is a relatively simple algorithm to implement**
  - Need a distance function
  - Need a label selection method
  - Need a k
- **KNN can be computationally expensive**
  - O(number of training items)
  - The distance calculation is O(number of dimensions)
  - Usually, one dimension per database vocabulary word
- **KNN can be very effective**
  - If training set is large, error rate approaches twice Bayes error rate
    - » Bayes error rate is optimal error rate if distribution is known

Other Learning Algorithms

There are many categorization algorithms

- Perceptron
- Widrow-Hoff
- Decision trees
- Support Vector Machines (SVM)
- Naïve Bayes
- Neural networks
- Maximum Entropy Modeling
- : : : :

Automatic Categorization: Datasets

- **Reuters collection (Modified Apte Split)**
  - 12,902 Reuters newswire documents from 1987
  - 9,603 training articles, 3,299 test articles
  - Articles categorized into more than 100 topics
    - “mergers and acquisitions”, “interest rates”, “earnings”, ….
- **Oregon Health Sciences University Medical database (OHSUMED)**
  - 348,566 Medline medical journal articles (1987-1991)
  - 106 queries
  - Articles categorized with MeSH codes

Automatic Categorization: What Makes it Hard?

- **Many similar categories**
- **Categories with small classes**
- **Hierarchical categorization**
- **Monothetic vs Polythetic categories:**
  - Human categories tend to be monothetic
    - » Every object shares one or more traits
    - » Monotheism is often conceptual, not vocabulary-based
    - » Example: Every document is about cancer
  - Machine learning categories are often polythetic
    - » Objects share a set of traits, but no trait is common to all
    - » Example: Documents contain words correlated with cancer

Automatic Categorization: State of the Art

Task	Computers	Humans
Essay grading (e.g., GMAT)	96-97%	95%
Medical (OHSUMED, MESH)	50-60%	?
Medical (ICD9)	45-60%	?
Newswire (Reuters)	80-90%	?
Yahoo! Science categories	60-70%	?
Web pages	80-90%	?
Internet newsgroups	80-90%	?
TREC relevance assessments	?	70%

Automatic Categorization: Assessment

- **Humans are not perfect**
  – but human error-rate is often ignored
- **Computers are not perfect**
  – but computer error-rate is often discussed
- **Cost factors encourage greater use of automatic categorization**
  – automatic categorization in relatively easy domains
  – the 80/20 rule applies in some domains (80% automatic, …)
  – human-assisted categorization
- **Current algorithms appear reasonably accurate**
  – significant research activity, considerable progress

Automatic categorization is practical

For More Information

- Y. Yang and X. Liu. "A re-examination of text categorization methods." In *Proceedings of ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR),* pp 42-49. 1999.

  http://www.cs.cmu.edu/~yiming/publications.html

- Y. Yang and J.O. Pedersen. "A comparative study on feature selection in text categorization." In *Proceedings of the Fourteenth International Conference on Machine Learning (ICML'97),* 1997.

  http://www.cs.cmu.edu/~yiming/publications.html