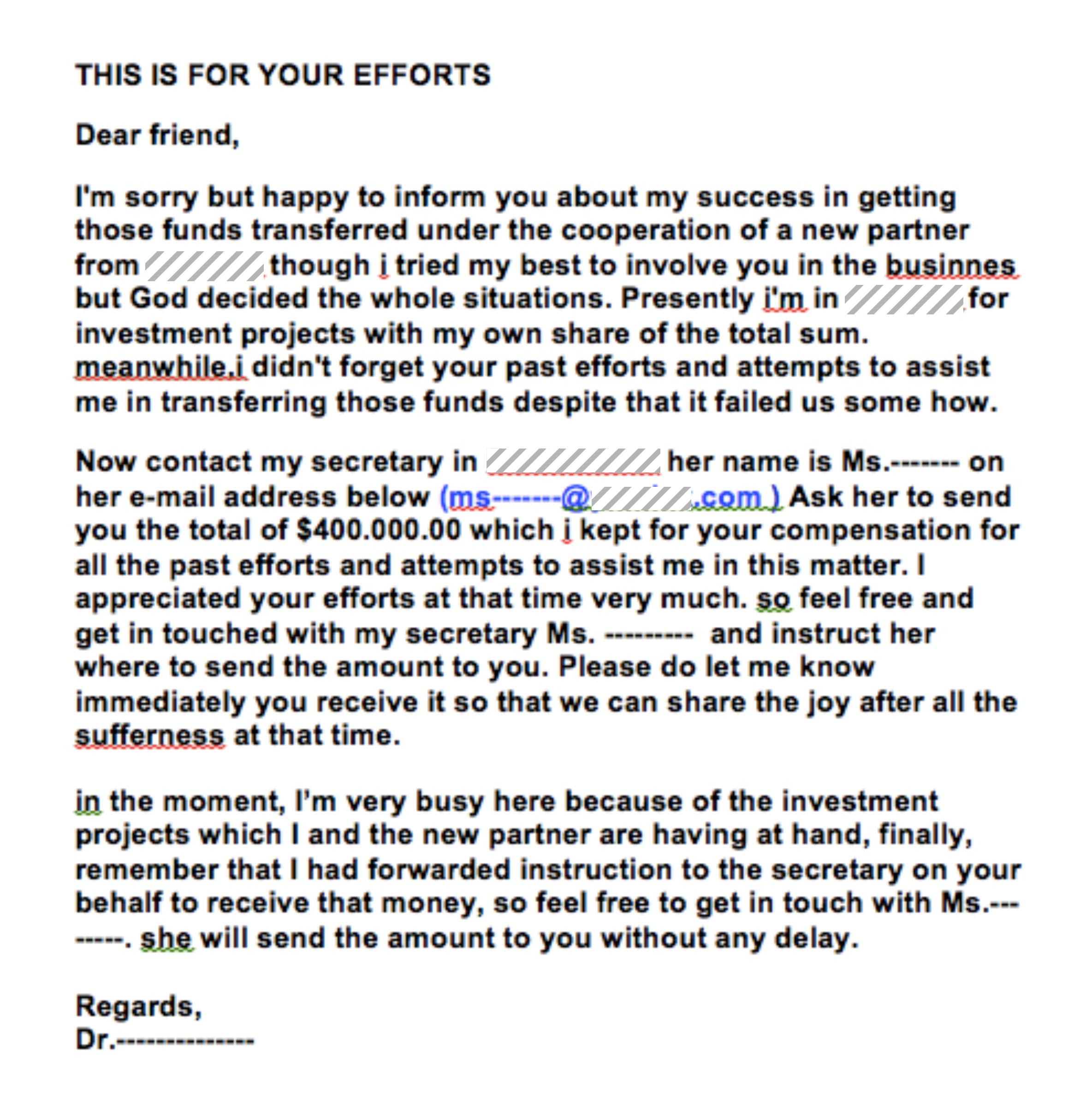
CS6200
Information Retrieval

Nada Naji

najin@ccs.neu.edu

College of Computer and Information Science
Northeastern University



THIS IS FOR YOUR EFFORTS

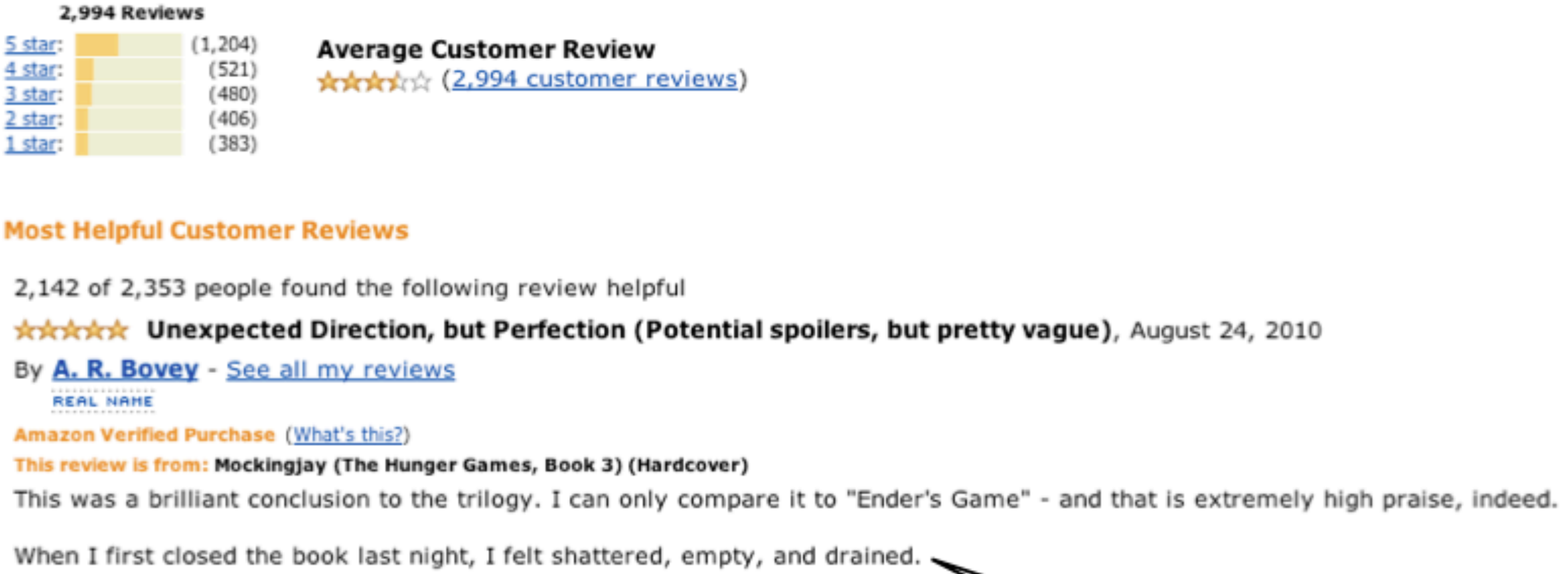
Dear friend,

I'm sorry but happy to inform you about my success in getting those funds transferred under the cooperation of a new partner from [redacted] though i tried my best to involve you in the businnes but God decided the whole situations. Presently i'm in [redacted] for investment projects with my own share of the total sum. meanwhile.i didn't forget your past efforts and attempts to assist me in transferring those funds despite that it failed us some how.

Now contact my secretary in [redacted] her name is Ms.------- on her e-mail address below (ms-------@[redacted].com) Ask her to send you the total of $400.000.00 which i kept for your compensation for all the past efforts and attempts to assist me in this matter. I appreciated your efforts at that time very much. so feel free and get in touched with my secretary Ms. --------- and instruct her where to send the amount to you. Please do let me know immediately you receive it so that we can share the joy after all the sufferness at that time.

in the moment, I'm very busy here because of the investment projects which I and the new partner are having at hand, finally, remember that I had forwarded instruction to the secretary on your behalf to receive that money, so feel free to get in touch with Ms.--------. she will send the amount to you without any delay.

Regards,
Dr.--------------



2,994 Reviews

5 star:	(1,204)
4 star:	(521)
3 star:	(480)
2 star:	(406)
1 star:	(383)

Average Customer Review

(2,994 customer reviews)

Most Helpful Customer Reviews

2,142 of 2,353 people found the following review helpful

★★★★★ **Unexpected Direction, but Perfection (Potential spoilers, but pretty vague)**, August 24, 2010

By **A. R. Bovey** - See all my reviews

REAL NAME

Amazon Verified Purchase (What's this?)

This review is from: Mockingjay (The Hunger Games, Book 3) (Hardcover)

This was a brilliant conclusion to the trilogy. I can only compare it to "Ender's Game" - and that is extremely high praise, indeed.

When I first closed the book last night, I felt shattered, empty, and drained.

Maybe not so good if found in a camera review

Classification and Clustering

- Classification and clustering are classical pattern recognition / machine learning problems
- Classification
  - Asks “what class does this item belong to?”
  - *Supervised learning* task
- Clustering
  - Asks “how can I group this set of items?”
  - *Unsupervised learning* task
- Items can be documents, queries, emails, entities, images, etc.
- Useful for a wide variety of search engine tasks

Classification

- Classification is the task of automatically applying labels to items
- Useful for many search-related tasks
  - Spam detection
  - Sentiment classification
  - Online advertising
- Two common approaches
  - Probabilistic
  - Geometric

How to Classify?

- How do humans classify items?
- For example, suppose you had to classify the “healthiness” of a food
  - Identify set of *features* indicative of health
    - fat, cholesterol, sugar, sodium, etc.
  - *Extract* features from foods
    - Read nutritional facts, chemical analysis, etc.
  - *Combine evidence* from the features into a hypothesis
    - Add health features together to get “healthiness factor”
  - Finally, *classify* the item based on the evidence
    - If “healthiness factor” is above a certain value, then deem it healthy

Ontologies

- Ontology is a labeling or categorization scheme
- Examples
  - Binary (spam, not spam)
  - Multi-valued (red, green, blue)
  - Hierarchical (news/local/sports)
- Different classification tasks require different ontologies

Standing queries

- The path from IR to text classification:
  - You have an information need to monitor, say:
    - Unrest in the Niger delta region
  - You want to rerun an appropriate query periodically to find new news items on this topic
  - You will be sent new documents that are found
    - I.e., it's not ranking but classification (relevant vs. not relevant)
- Such queries are called **standing queries**
  - Long used by "information professionals"
  - A modern mass instantiation is **Google Alerts**
- Standing queries are (hand-written) text classifiers

Naïve Bayes Classifier

- Probabilistic classifier based on Bayes' rule:

$$\begin{aligned} P(C|D) &= \frac{P(D|C)P(C)}{P(D)} \\ &= \frac{P(D|C)P(C)}{\sum_{c \in \mathcal{C}} P(D|C=c)P(C=c)} \end{aligned}$$

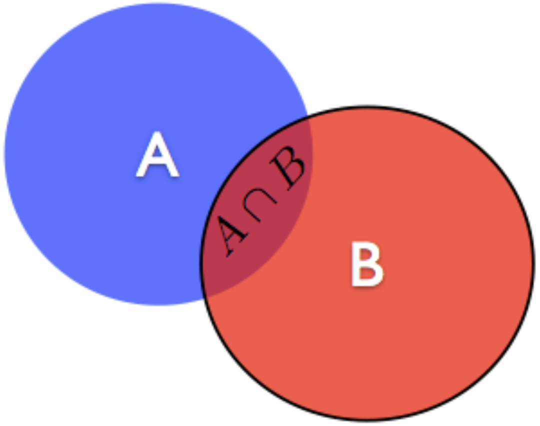
- *C* is a random variable corresponding to the class
- *D* is a random variable corresponding to the input (e.g. document)

Probability 101: Random Variables

- Random variables are non-deterministic
  - Can be discrete (finite number of outcomes) or continues
  - Model uncertainty in a variable
- P(*X* = *x*) means “the probability that random variable X takes on value x”
- Example:
  - Let X be the outcome of a coin toss
  - P(X = heads) = P(X = tails) = 0.5
- Example: $Y = 5 - 2X$
  - If *X* is random, then *Y* is random
  - If *X* is deterministic then *Y* is also deterministic
    - *Note:* “Deterministic” just means P(X = x) = 1.0!

Conditional Probability

$$P(A \mid B) = \frac{P(A, B)}{P(B)}$$



$$P(A, B) = P(B)P(A \mid B) = P(A)P(B \mid A)$$

$$\begin{aligned} P(A_1, A_2, \ldots, A_n) &= P(A_1)P(A_2 \mid A_1)P(A_3 \mid A_1, A_2) \\ &\quad \cdots P(A_n \mid A_1, \ldots, A_{n-1}) \end{aligned}$$

Chain rule

Naïve Bayes Classifier

- Documents are classified according to:

$$\begin{aligned}\text{Class}(d) &= \arg\max_{c\in\mathcal{C}} P(c|d) \\ &= \arg\max_{c\in\mathcal{C}} \frac{P(d|c)P(c)}{\sum_{c\in C} P(d|c)P(c)}\end{aligned}$$

- Must estimate P(*d* | *c*) and P(*c*)
  - P(*c*) is the probability of observing class *c*
  - P(*d* | *c*) is the probability that document d is observed given the class is known to be *c*

- What we want:

  $p(☺ \mid w_1, w_2, \ldots, w_n) > p(☹ \mid w_1, w_2, \ldots, w_n)$?

- What we know how to build:
  - A language model for each class
    - $p(w_1, w_2, \ldots, w_n \mid ☺)$

Estimating *P*(*c*)

- P(*c*) is the probability of observing class *c*
- Estimated as the proportion of training documents in class *c*:

$$P(c) = \frac{N_c}{N}$$

- N_c is the number of training documents in class *c*
- *N* is the total number of training documents

Estimating P(*d* | *c*)

- P(*d* | *c*) is the probability that document *d* is observed given the class is known to be *c*
- Estimate depends on the *event space* used to represent the documents
- What is an event space?
  - The set of all possible outcomes for a given random variable
  - For a coin toss random variable the event space is *S* = {heads, tails}

SpamAssassin Features:

- Basic (Naïve) Bayes spam probability
- Mentions: Generic Viagra
- Regex: millions of (dollar) ((dollar) NN,NNN,NNN.NN)
- Phrase: impress ... girl
- Phrase: ‘Prestigious Non-Accredited Universities’
- From: starts with many numbers
- Subject is all capitals
- HTML has a low ratio of text to image area

Recall: Vector Space Representation

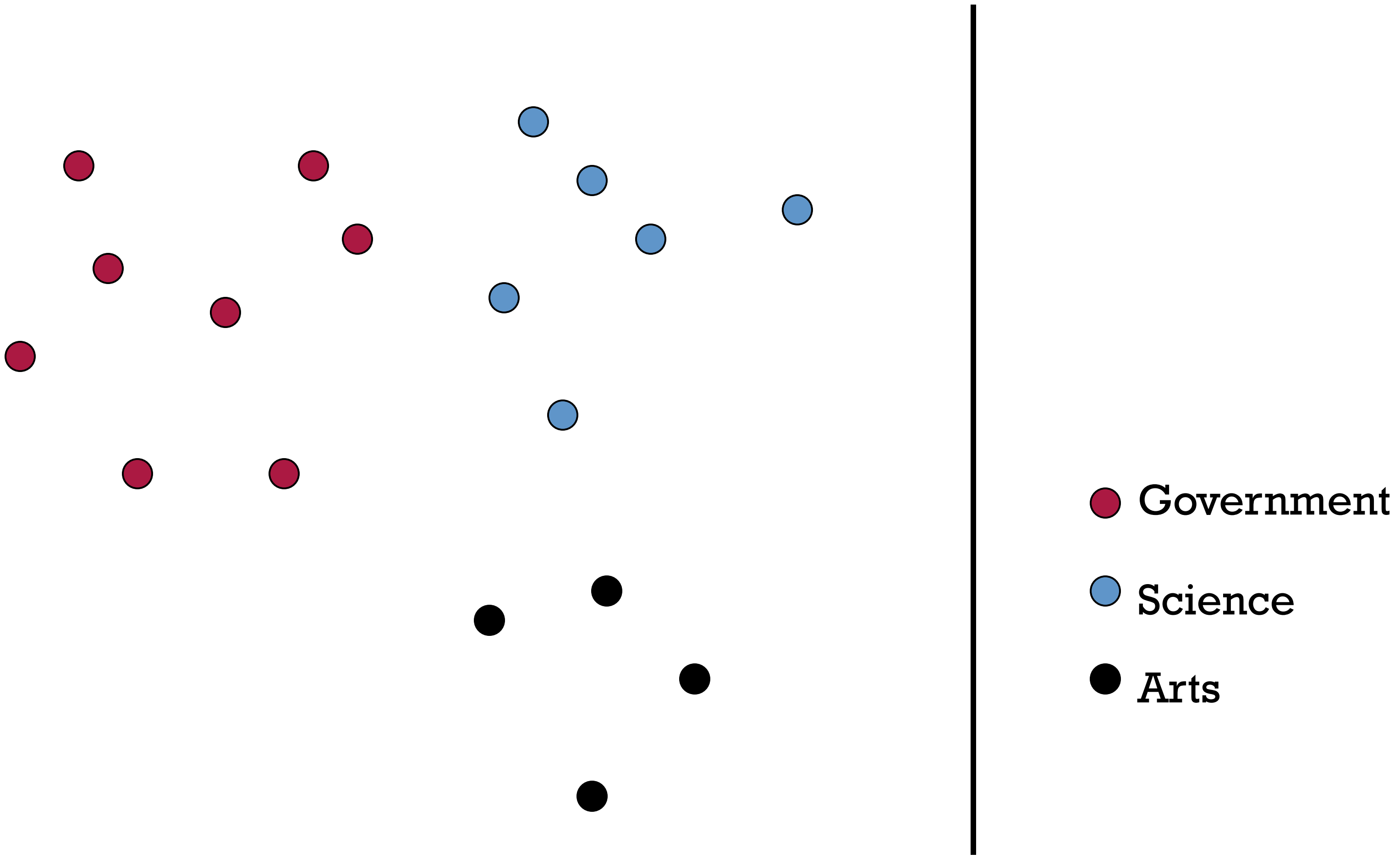
- Each document is a vector, one component for each term (= word).
- Normally normalize vectors to unit length.
- High-dimensional vector space:
  - Terms are axes
  - 10,000+ dimensions, or even 100,000+
  - Docs are vectors in this space

- How can we do classification in this space?

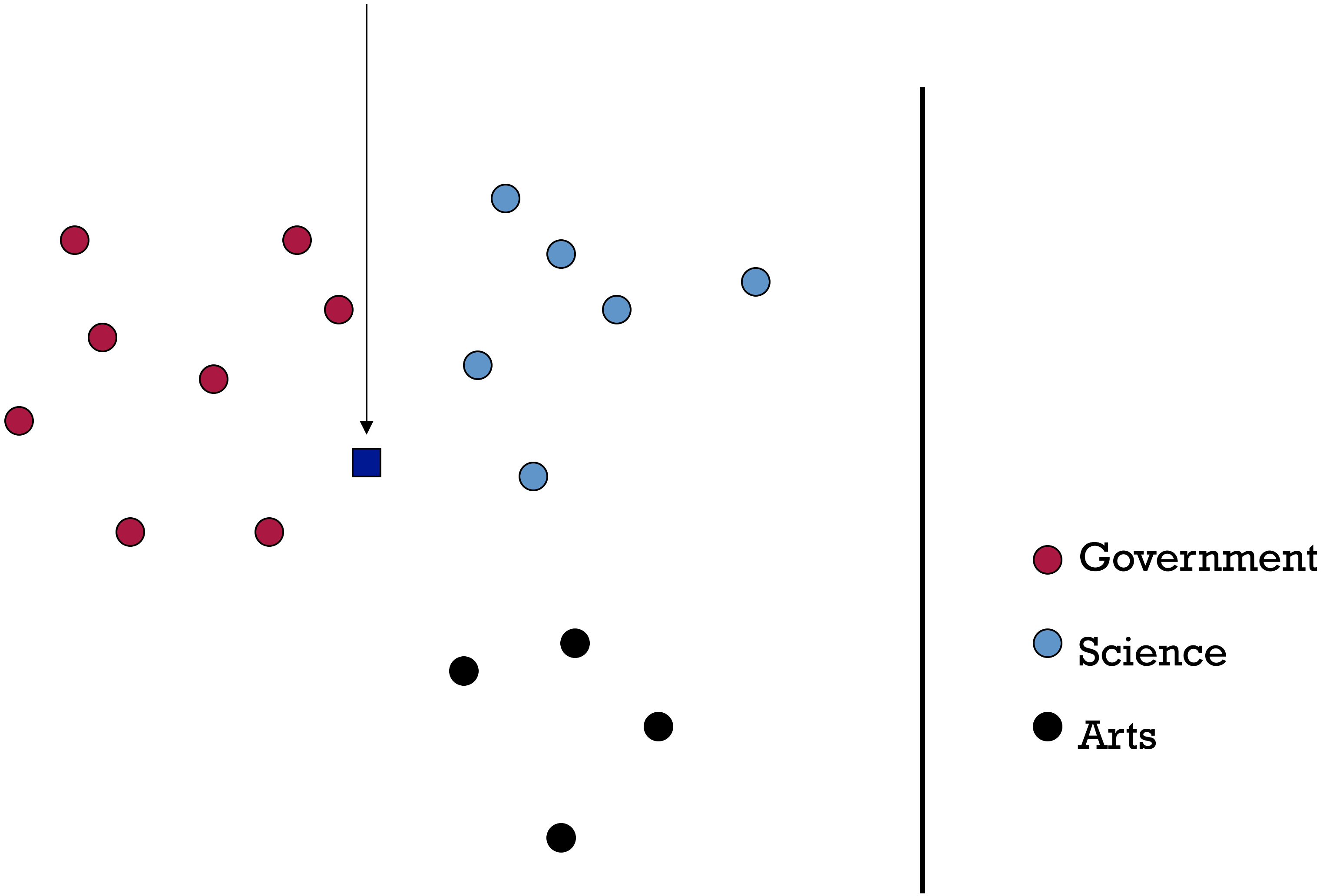
Classification Using Vector Spaces

- As before, the training set is a set of documents, each labeled with its class (e.g., topic)
- In vector space classification, this set corresponds to a labeled set of points (or, equivalently, vectors) in the vector space
- Premise 1: Documents in the same class form a contiguous region of space
- Premise 2: Documents from different classes don't overlap (much)
- We define surfaces to delineate classes in the space

Documents in a Vector Space



Test Document of what class?



Definition of centroid

$$\vec{\mu}(c) = \frac{1}{|D_c|} \sum_{d \in D_c} \vec{v}(d)$$

- Where D_c is the set of all documents that belong to class *c* and *v*(*d*) is the vector space representation of *d.*

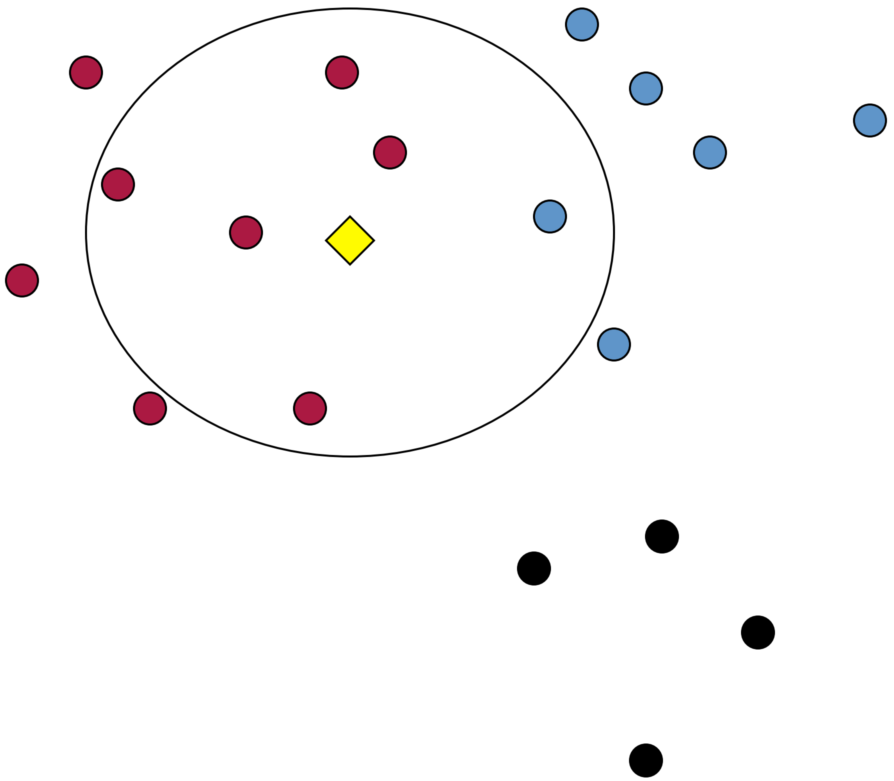
- *Note that centroid will in general* not *be a unit vector even when the inputs are unit vectors.*

k Nearest Neighbor Classification

- kNN = *k* Nearest Neighbor

- To classify a document *d*:
- Define *k*-neighborhood as the *k* nearest neighbors of *d*
- Pick the majority class label in the *k*-neighborhood

Example: k=6 (6NN)



P(science|◇)?

🔴 Government

🔵 Science

⚫ Arts

Nearest-Neighbor Learning

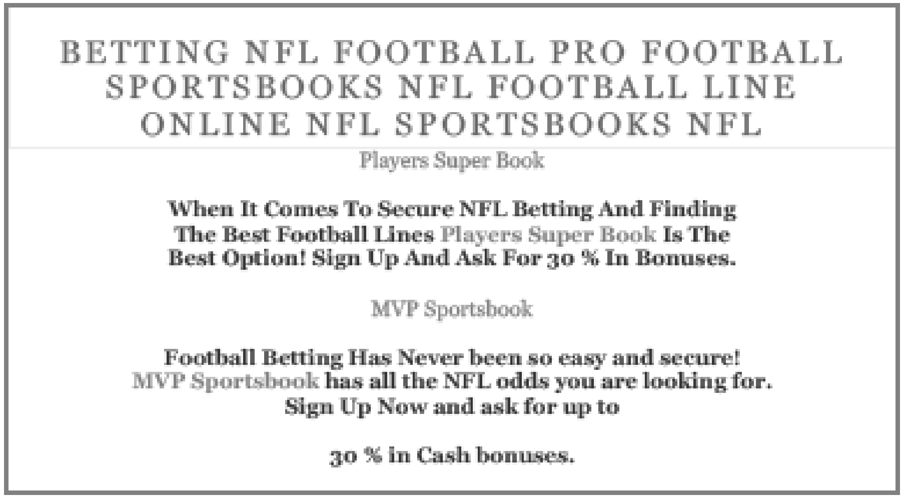
- Learning: just store the labeled training examples *D*
- Testing instance *x (under 1NN)*:
  - Compute similarity between *x* and all examples in *D*.
  - Assign *x* the category of the most similar example in *D*.
- Does not compute anything beyond storing the examples
- Also called:
  - Case-based learning
  - Memory-based learning
  - Lazy learning
- Rationale of kNN: contiguity hypothesis

Feature Selection

- Document classifiers can have a very large number of features
  - Not all features are useful
  - Excessive features can increase computational cost of training and testing
- *Feature selection* methods reduce the number of features by choosing the most useful features

Spam Example

Website:



BETTING NFL FOOTBALL PRO FOOTBALL SPORTSBOOKS NFL FOOTBALL LINE ONLINE NFL SPORTSBOOKS NFL

Players Super Book

When It Comes To Secure NFL Betting And Finding The Best Football Lines Players Super Book Is The Best Option! Sign Up And Ask For 30 % In Bonuses.

MVP Sportsbook

Football Betting Has Never been so easy and secure! MVP Sportsbook has all the NFL odds you are looking for. Sign Up Now and ask for up to

30 % in Cash bonuses.

Term spam:

pro football sportsbooks nfl football line online nfl sportsbooks nfl football gambling odds online pro nfl betting pro nfl gambling online nfl football spreads offshore football gambling online nfl gamblibg spreads online football gambling line online nfl betting nfl sportsbook online online nfl betting spreads betting nfl football online online football wagering online gambling online gambling football online nfl football betting odds offshore football sportsbook online nfl football gambling …

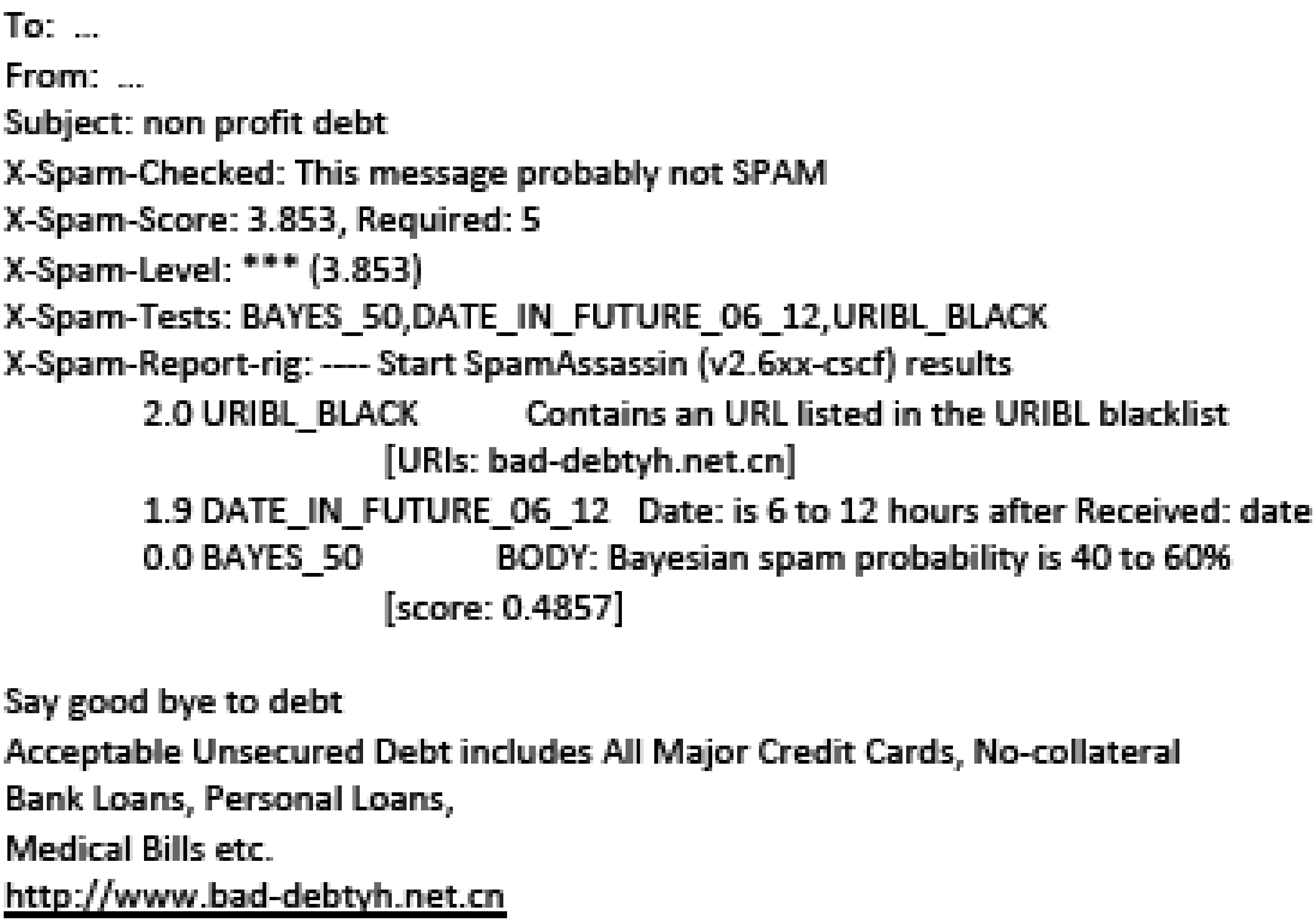
Link spam:

MVP Sportsbook Football Gambling Beverly Hills Football Sportsbook
Players SB Football Wagering Popular Poker Football Odds
Virtual Bookmaker Football Lines V Wager Football Spreads
Bogarts Casino Football Point Spreads Gecko Casino Online Football Betting
Jackpot Hour Online Football Gambling MVP Casino Online Football Wagering
Toucan Casino NFL Betting Popular Poker NFL Gambling
All Tracks NFL Wagering Bet Jockey NFL Odds
Live Horse Betting NFL Lines MVP Racebook NFL Point Spreads
Popular Poker NFL Spreads Bogarts Poker NFL Sportsbook …

Spam Detection

- Useful features
  - Unigrams
  - Formatting (invisible text, flashing, etc.)
  - Misspellings
  - IP address
- Different features are useful for different spam detection tasks
- Email and web page spam are by far the most widely studied, well understood, and easily detected types of spam

Example Spam Assassin Output



```
To:  …
From:  …
Subject: non profit debt
X-Spam-Checked: This message probably not SPAM
X-Spam-Score: 3.853, Required: 5
X-Spam-Level: *** (3.853)
X-Spam-Tests: BAYES_50,DATE_IN_FUTURE_06_12,URIBL_BLACK
X-Spam-Report-rig: ---- Start SpamAssassin (v2.6xx-cscf) results
            2.0 URIBL_BLACK          Contains an URL listed in the URIBL blacklist
                                     [URIs: bad-debtyh.net.cn]
            1.9 DATE_IN_FUTURE_06_12  Date: is 6 to 12 hours after Received: date
            0.0 BAYES_50             BODY: Bayesian spam probability is 40 to 60%
                                     [score: 0.4857]

Say good bye to debt
Acceptable Unsecured Debt includes All Major Credit Cards, No-collateral
Bank Loans, Personal Loans,
Medical Bills etc.
http://www.bad-debtyh.net.cn
```

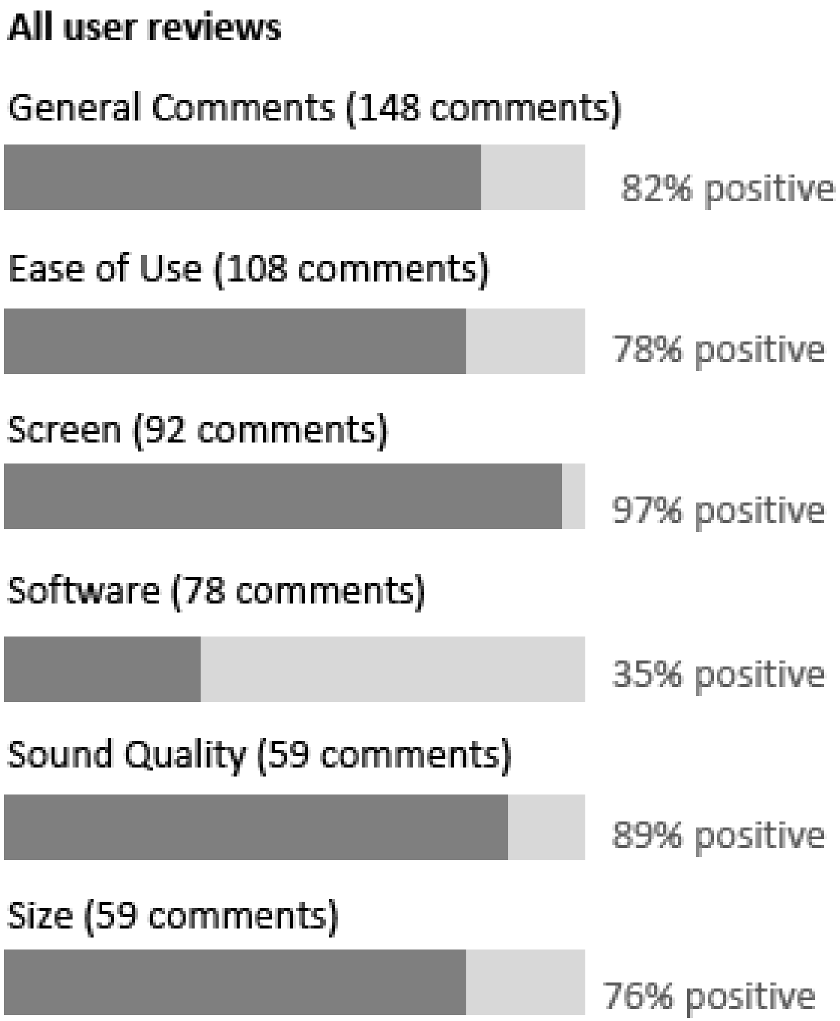
Sentiment

- Blogs, online reviews, and forum posts are often opinionated
- Sentiment classification attempts to automatically identify the polarity of the opinion
  - Negative opinion
  - Neutral opinion
  - Positive opinion
- Sometimes the strength of the opinion is also important
  - “Two stars” vs. “four stars”
  - Weakly negative vs. strongly negative

Classifying Sentiment

- Useful features
  - Unigrams
  - Bigrams
  - Part of speech tags
  - Adjectives
- SVMs with unigram features have been shown to be outperform hand built rules

Sentiment Classification Example



All user reviews

General Comments (148 comments)

82% positive

Ease of Use (108 comments)

78% positive

Screen (92 comments)

97% positive

Software (78 comments)

35% positive

Sound Quality (59 comments)

89% positive

Size (59 comments)

76% positive

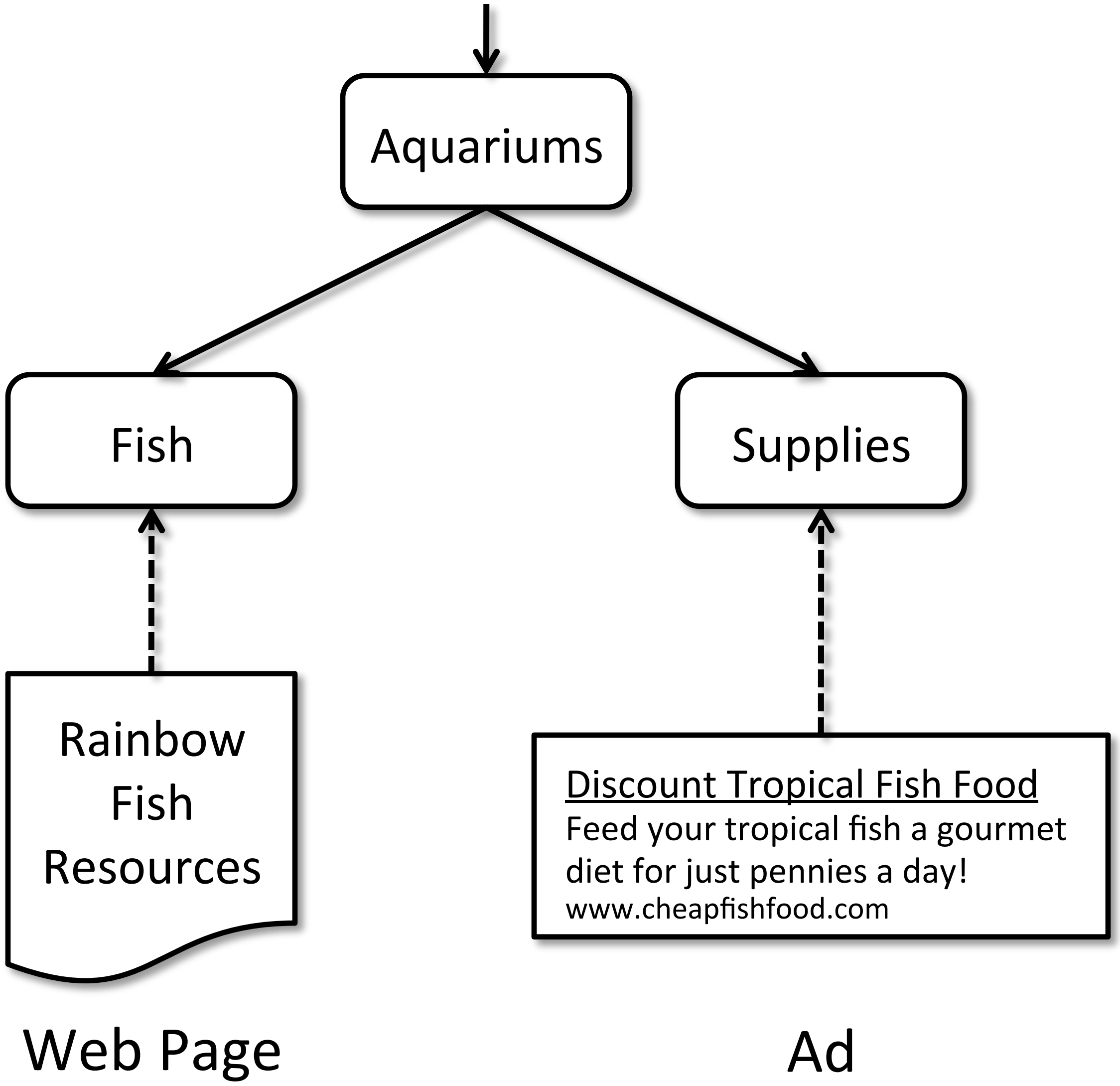
Classifying Online Ads

- Unlike traditional search, online advertising goes beyond "topical relevance"
- A user searching for 'tropical fish' may also be interested in pet stores, local aquariums, or even scuba diving lessons
- These are semantically related, but not topically relevant!
- We can bridge the semantic gap by classifying ads and queries according to a semantic hierarchy

Semantic Classification

- Semantic hierarchy ontology
  - Example: Pets / Aquariums / Supplies
- Training data
  - Large number of queries and ads are manually classified into the hierarchy
- Nearest neighbor classification has been shown to be effective for this task
- Hierarchical structure of classes can be used to improve classification accuracy

Semantic Classification



Aquariums

Fish

Supplies

Rainbow Fish Resources

Web Page

Discount Tropical Fish Food
Feed your tropical fish a gourmet diet for just pennies a day!
www.cheapfishfood.com

Ad

Clustering

- A set of unsupervised algorithms that attempt to find latent structure in a set of items
- Goal is to identify groups (clusters) of similar items
- Suppose I gave you the shape, color, vitamin C content, and price of various fruits and asked you to cluster them
  - What criteria would you use?
  - How would you define similarity?
- Clustering is very sensitive to how items are represented and how similarity is defined!

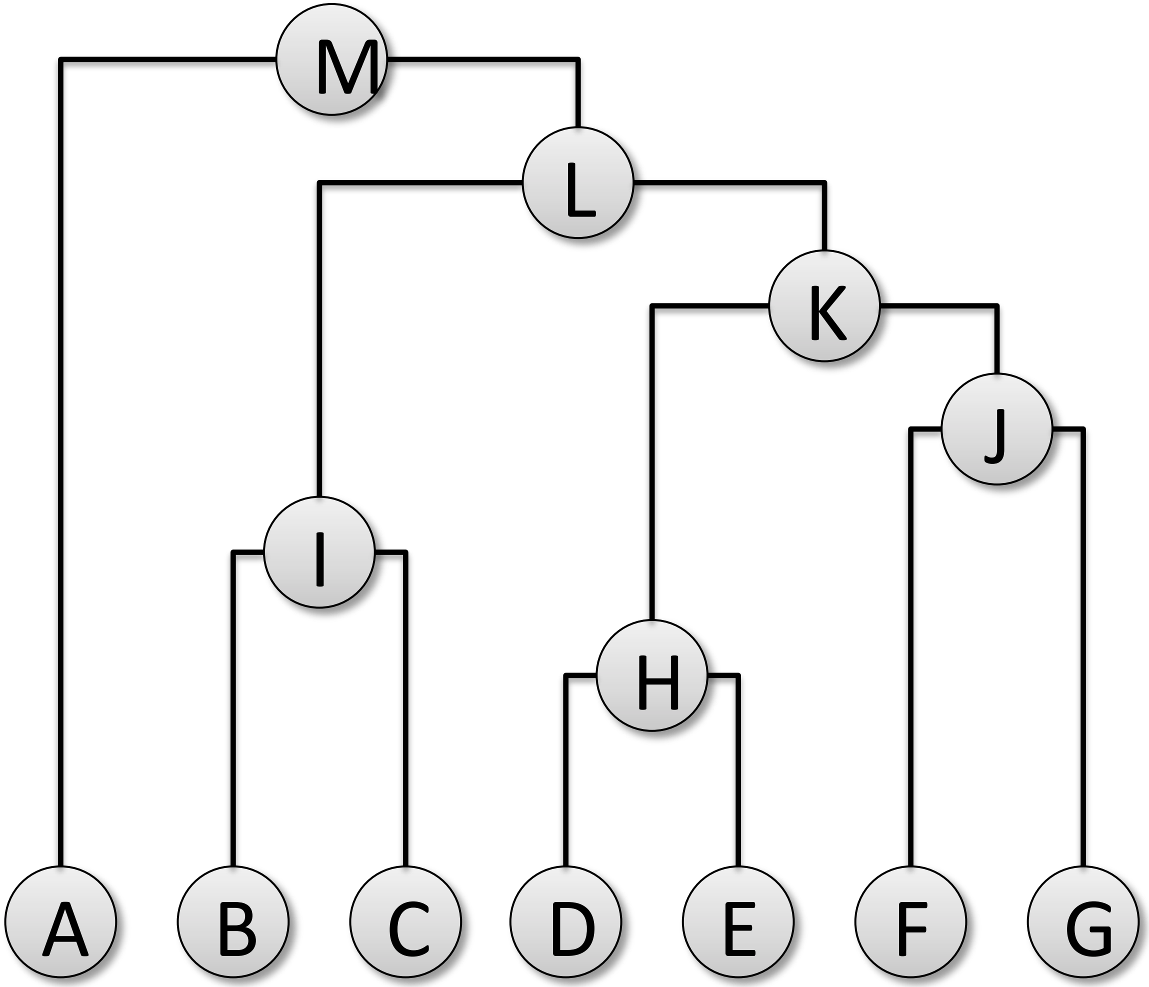
Clustering

- General outline of clustering algorithms
  1. Decide how items will be represented (e.g., feature vectors)
  2. Define similarity measure between pairs or groups of items (e.g., cosine similarity)
  3. Determine what makes a “good” clustering
  4. Iteratively construct clusters that are increasingly “good”
  5. Stop after a local/global optimum clustering is found
- Steps 3 and 4 differ the most across algorithms

Hierarchical Clustering

- Constructs a hierarchy of clusters
  - The top level of the hierarchy consists of a single cluster with all items in it
  - The bottom level of the hierarchy consists of *N* (# items) singleton clusters
- Two types of hierarchical clustering
  - Divisive (“top down”)
  - Agglomerative (“bottom up”)
- Hierarchy can be visualized as a *dendogram*

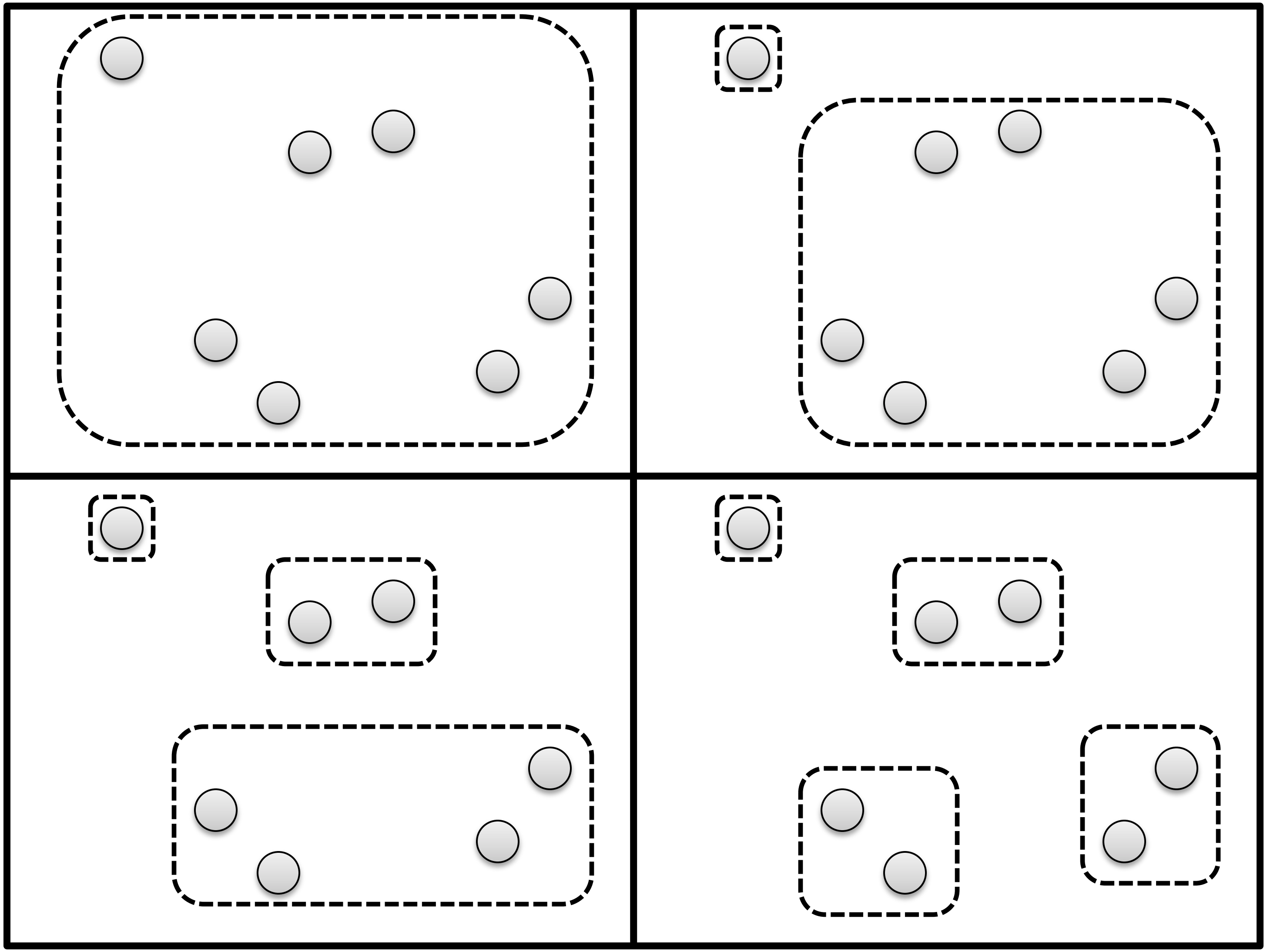
Example Dendrogram



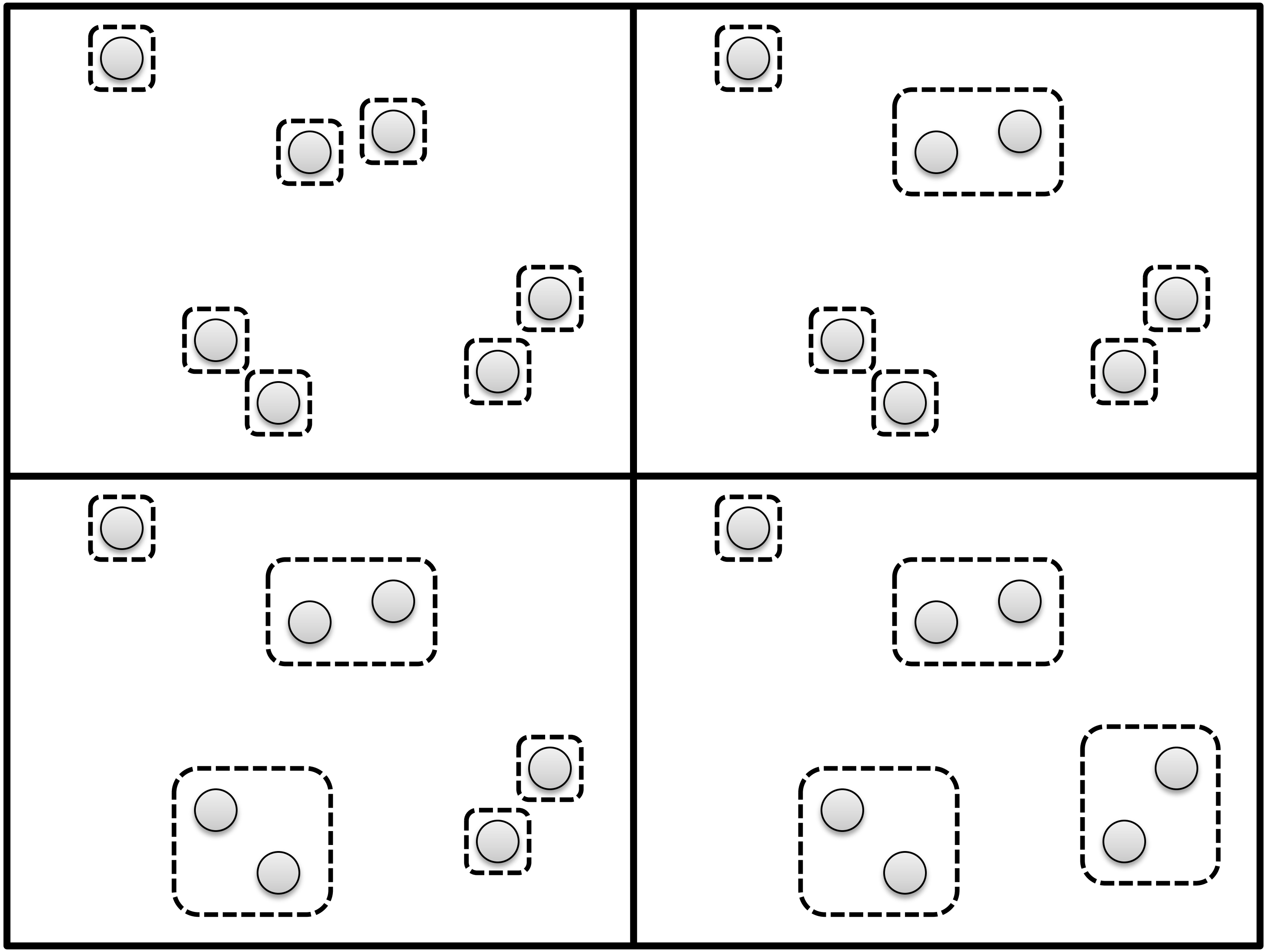
Divisive and Agglomerative Hierarchical Clustering

- Divisive
  - Start with a single cluster consisting of all of the items
  - Until only singleton clusters exist…
    - **Divide** an existing cluster into two new clusters
- Agglomerative
  - Start with *N* (# items) singleton clusters
  - Until a single cluster exists…
    - **Combine** two existing cluster into a new cluster
- How do we know how to divide or combined clusters?
  - Define a division or combination cost
  - Perform the division or combination with the lowest cost

Divisive Hierarchical Clustering



Agglomerative Hierarchical Clustering



Clustering Costs

- Single linkage

$$COST(C_i, C_j) = \min\{dist(X_i, X_j) | X_i \in C_i, X_j \in C_j\}$$

- Complete linkage

$$COST(C_i, C_j) = \max\{dist(X_i, X_j) | X_i \in C_i, X_j \in C_j\}$$

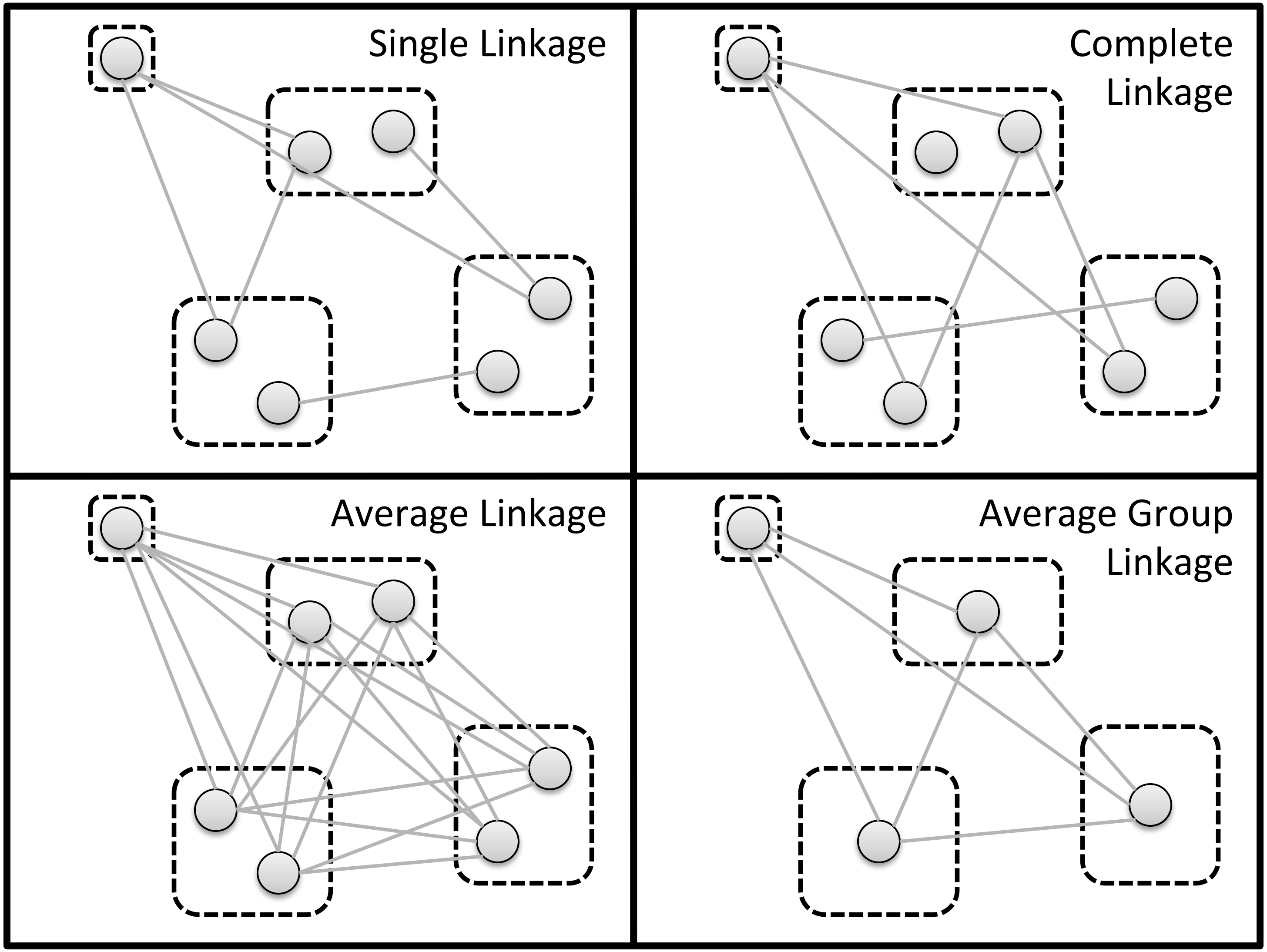
- Average linkage

$$COST(C_i, C_j) = \frac{\sum_{X_i \in C_i, X_j \in C_j} dist(X_i, X_j)}{|C_i||C_j|}$$

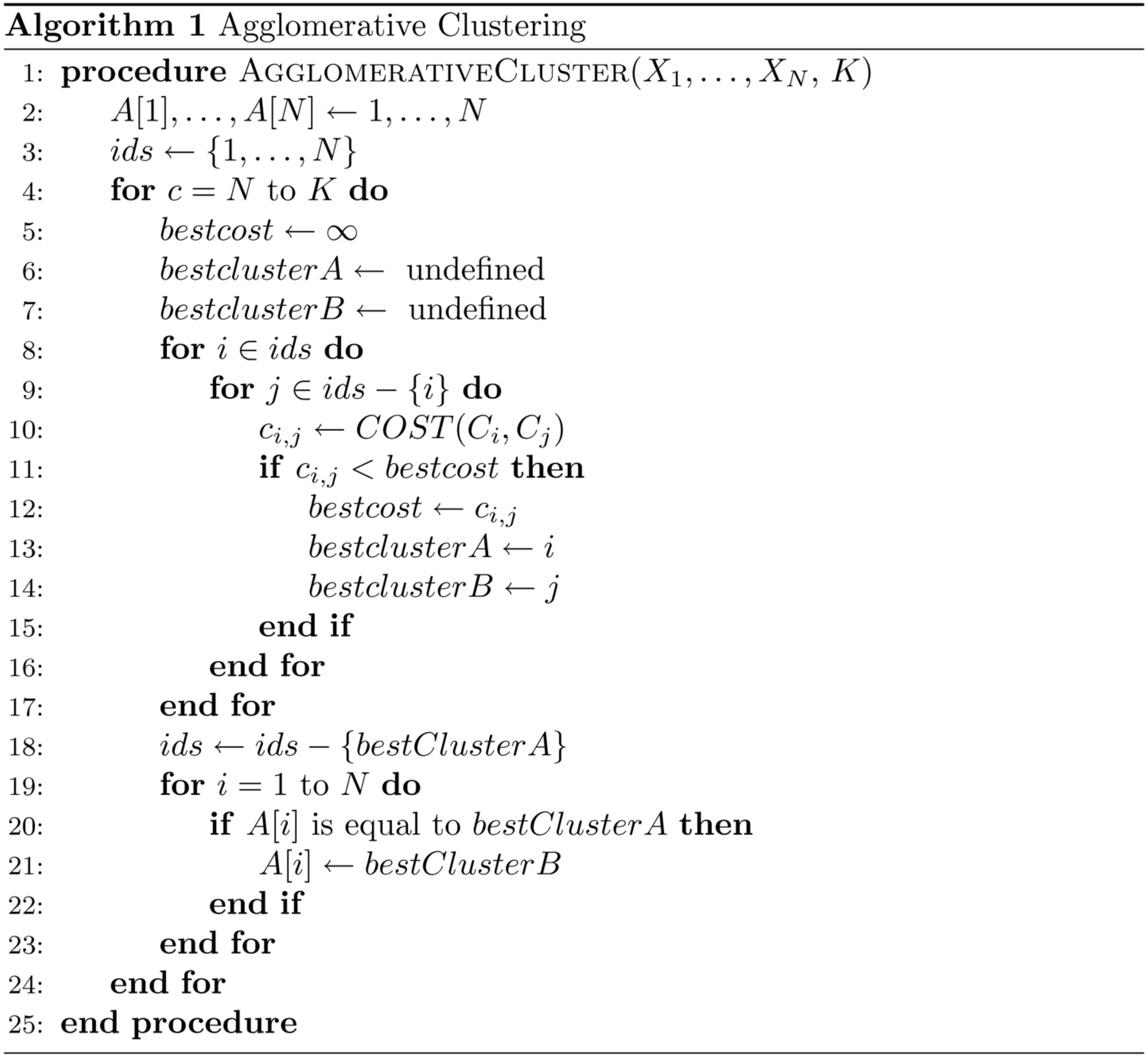
- Average group linkage

$$COST(C_i, C_j) = dist(\mu_{C_i}, \mu_{C_j})$$

Clustering Strategies



Agglomerative Clustering Algorithm



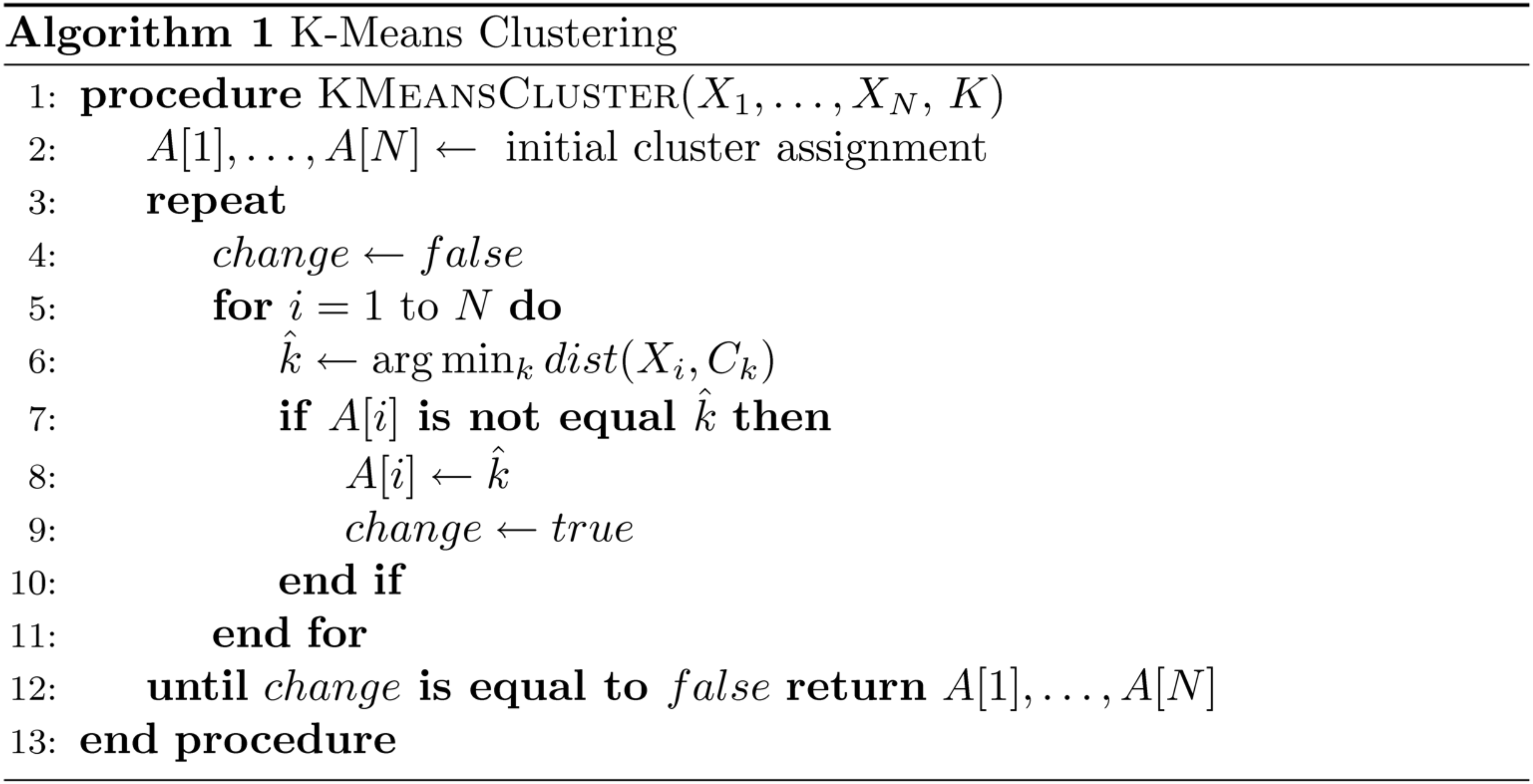
Algorithm 1 Agglomerative Clustering

```
procedure AGGLOMERATIVECLUSTER($X_1, \ldots, X_N$, $K$)
    $A[1], \ldots, A[N] \leftarrow 1, \ldots, N$
    $ids \leftarrow \{1, \ldots, N\}$
    for $c = N$ to $K$ do
        $bestcost \leftarrow \infty$
        $bestclusterA \leftarrow$ undefined
        $bestclusterB \leftarrow$ undefined
        for $i \in ids$ do
            for $j \in ids - \{i\}$ do
                $c_{i,j} \leftarrow COST(C_i, C_j)$
                if $c_{i,j} < bestcost$ then
                    $bestcost \leftarrow c_{i,j}$
                    $bestclusterA \leftarrow i$
                    $bestclusterB \leftarrow j$
                end if
            end for
        end for
        $ids \leftarrow ids - \{bestClusterA\}$
        for $i = 1$ to $N$ do
            if $A[i]$ is equal to $bestClusterA$ then
                $A[i] \leftarrow bestClusterB$
            end if
        end for
    end for
end procedure
```

K-Means Clustering

- Hierarchical clustering constructs a hierarchy of clusters
- K-means always maintains exactly *K* clusters
  - Clusters represented as centroids (“center of mass”)
- Basic algorithm:
  - Step 0: Choose *K* cluster centroids
  - Step 1: Assign points to closet centroid
  - Step 2: Recompute cluster centroids
  - Step 3: Goto 1
- Tends to converge quickly
- Can be sensitive to choice of initial centroids
- Must choose *K*!

K-Means Clustering Algorithm



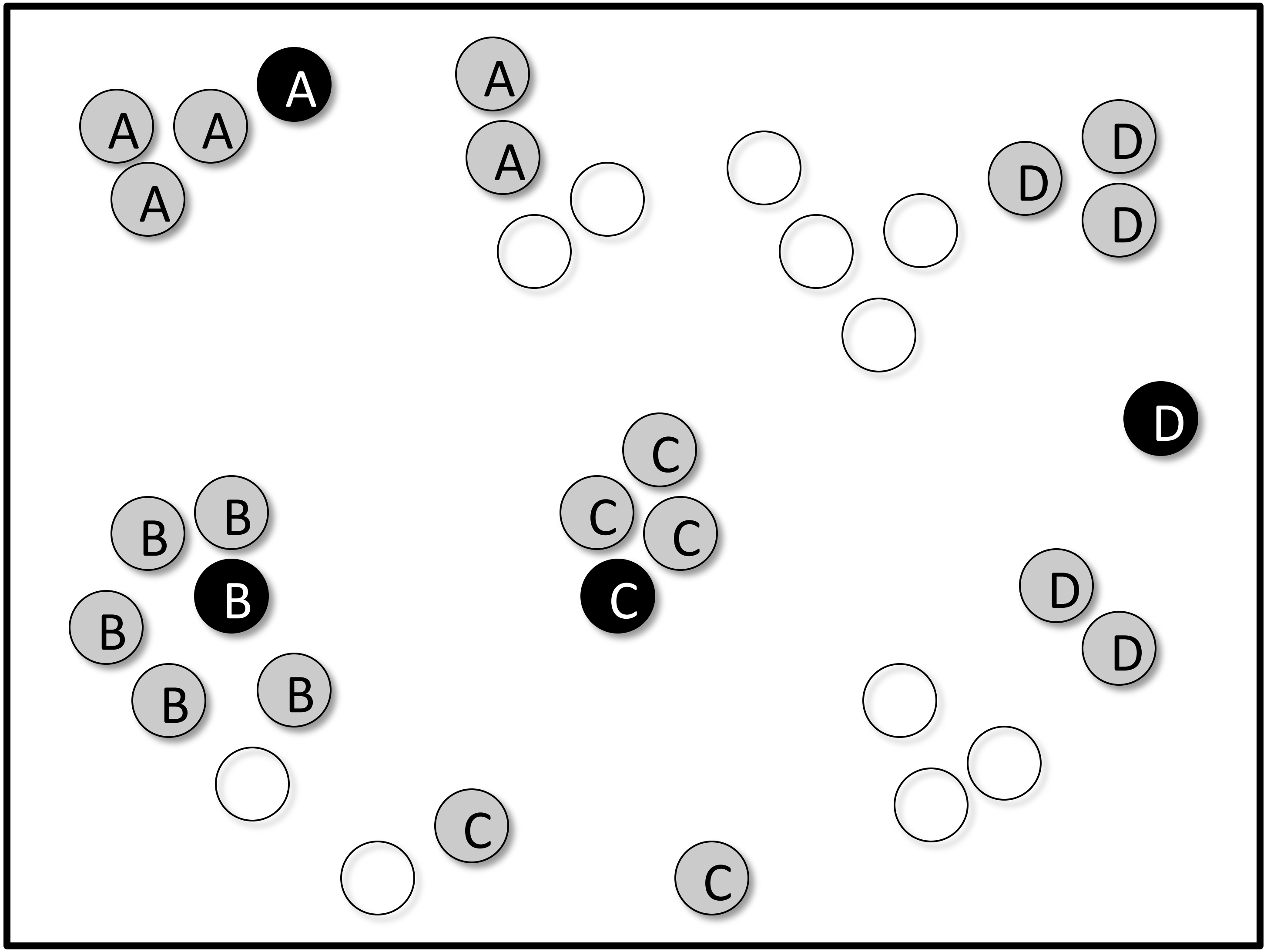
Algorithm 1 K-Means Clustering

```
procedure KMeansCluster($X_1, \ldots, X_N, K$)
    $A[1], \ldots, A[N] \leftarrow$ initial cluster assignment
    repeat
        $change \leftarrow false$
        for $i = 1$ to $N$ do
            $\hat{k} \leftarrow \arg\min_k dist(X_i, C_k)$
            if $A[i]$ is not equal $\hat{k}$ then
                $A[i] \leftarrow \hat{k}$
                $change \leftarrow true$
            end if
        end for
    until $change$ is equal to $false$ return $A[1], \ldots, A[N]$
end procedure
```

K-Nearest Neighbor Clustering

- Hierarchical and K-Means clustering partition items into clusters
  - Every item is in exactly one cluster
- K-Nearest neighbor clustering forms one cluster per item
  - The cluster for item *j* consists of *j* and *j*'s *K* nearest neighbors
  - Clusters now overlap

5-Nearest Neighbor Clustering



Evaluating Clustering

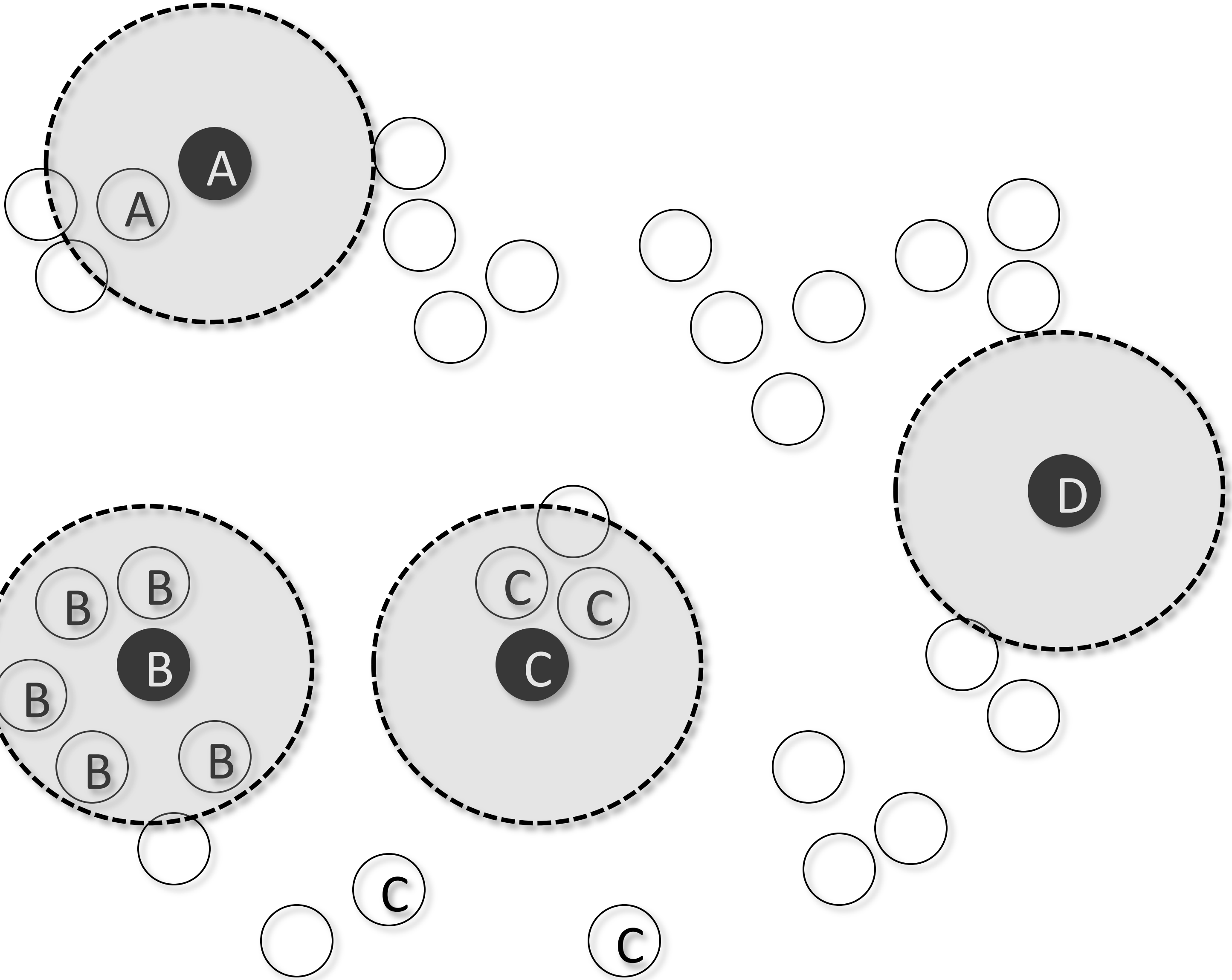
- Evaluating clustering is challenging, since it is an ***unsupervised*** learning task
- If labels exist, can use standard IR metrics, such as precision and recall
- If not, then can use measures such as “cluster precision”, which is defined as:

$$ClusterPrecision = \frac{\sum_{i=1}^{K} |\mathrm{MaxClass}(C_i)|}{N}$$

How to Choose K?

- K-means and K-nearest neighbor clustering require us to choose *K*, the number of clusters
- No theoretically appealing way of choosing *K*
- Depends on the application and data
- Can use hierarchical clustering and choose the best level of the hierarchy to use
- Can use adaptive *K* for K-nearest neighbor clustering
  - Define a ‘ball’ around each item
- Difficult problem with no clear solution

Adaptive Nearest Neighbor Clustering



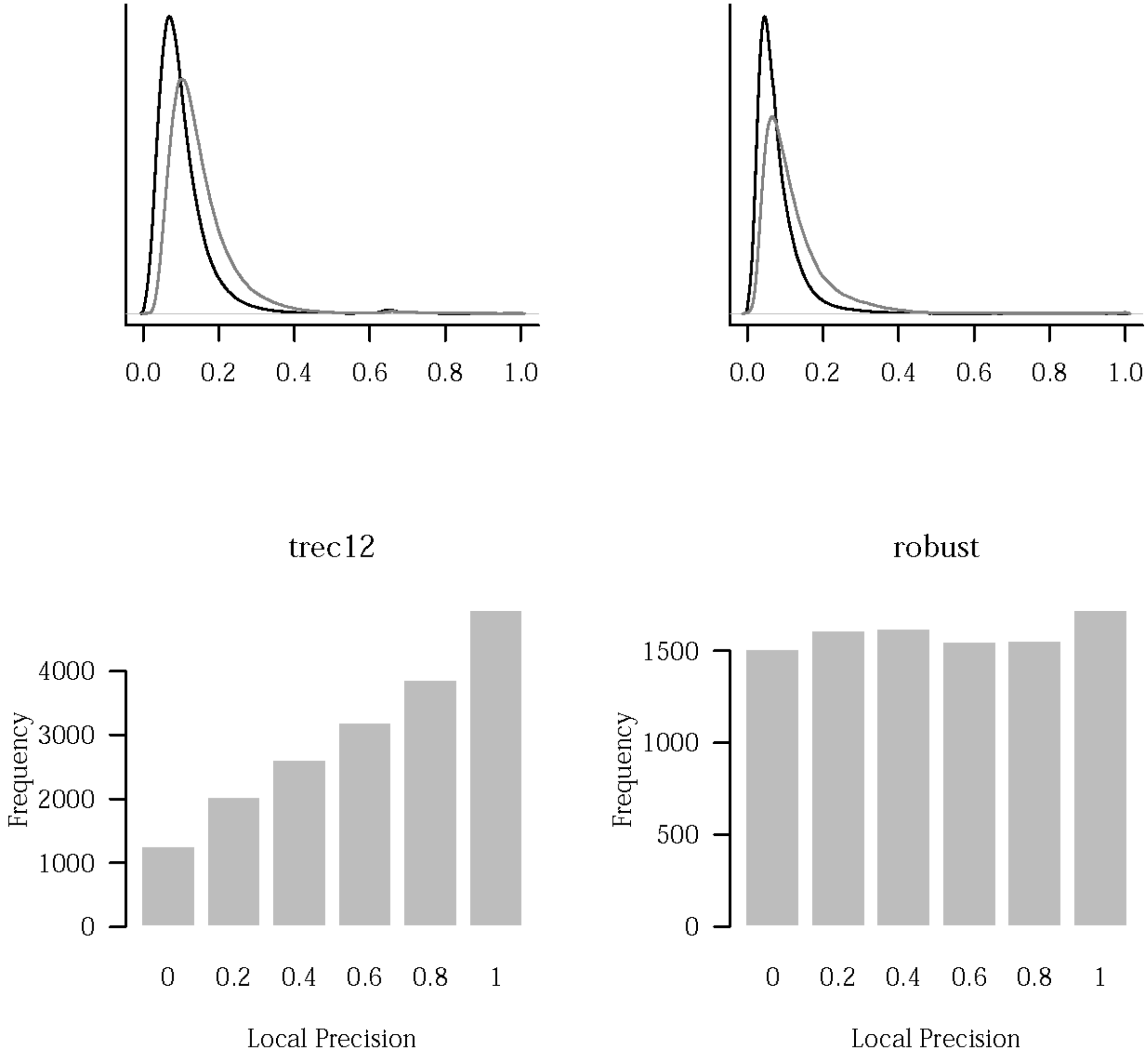
Clustering and Search

- Cluster hypothesis
  - "Closely associated documents tend to be relevant to the same requests" – van Rijsbergen '79
- Tends to hold in practice, but not always
- Two retrieval modeling options
  - Retrieve clusters according to P(Q | C_i)
  - Smooth documents using K-NN clusters:

$$P(w|D) = (1 - \lambda - \delta)\frac{f_{w,D}}{|D|} + \delta \sum_{C_j} \frac{f_{w,C_j}}{|C_j|} P(D|C_j) + \lambda \frac{f_{w,Coll}}{|Coll|}$$

- Smoothing approach more effective

Testing the Cluster Hypothesis


trec12
robust
Frequency
Local Precision
Frequency
Local Precision