Information Storage and Retrieval

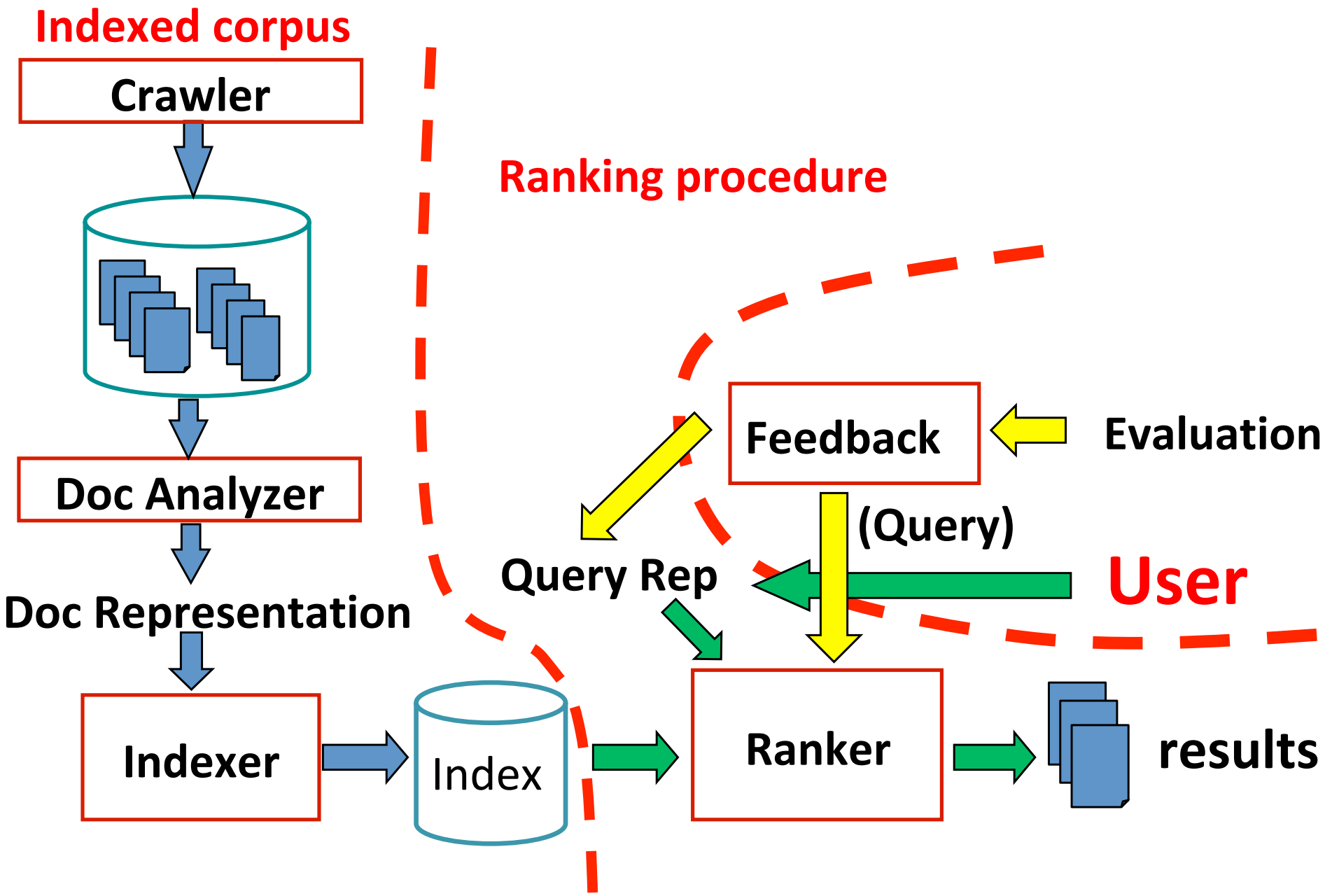
CSCE 670
Texas A&M University
Department of Computer Science & Engineering
Instructor: Prof. James Caverlee

Search Engine Basics + Learning Basics
23/25 January 2018

Some slides from https://nlp.stanford.edu/IR-book/, from Stanford CS 221, and from Hongning Wang

This week

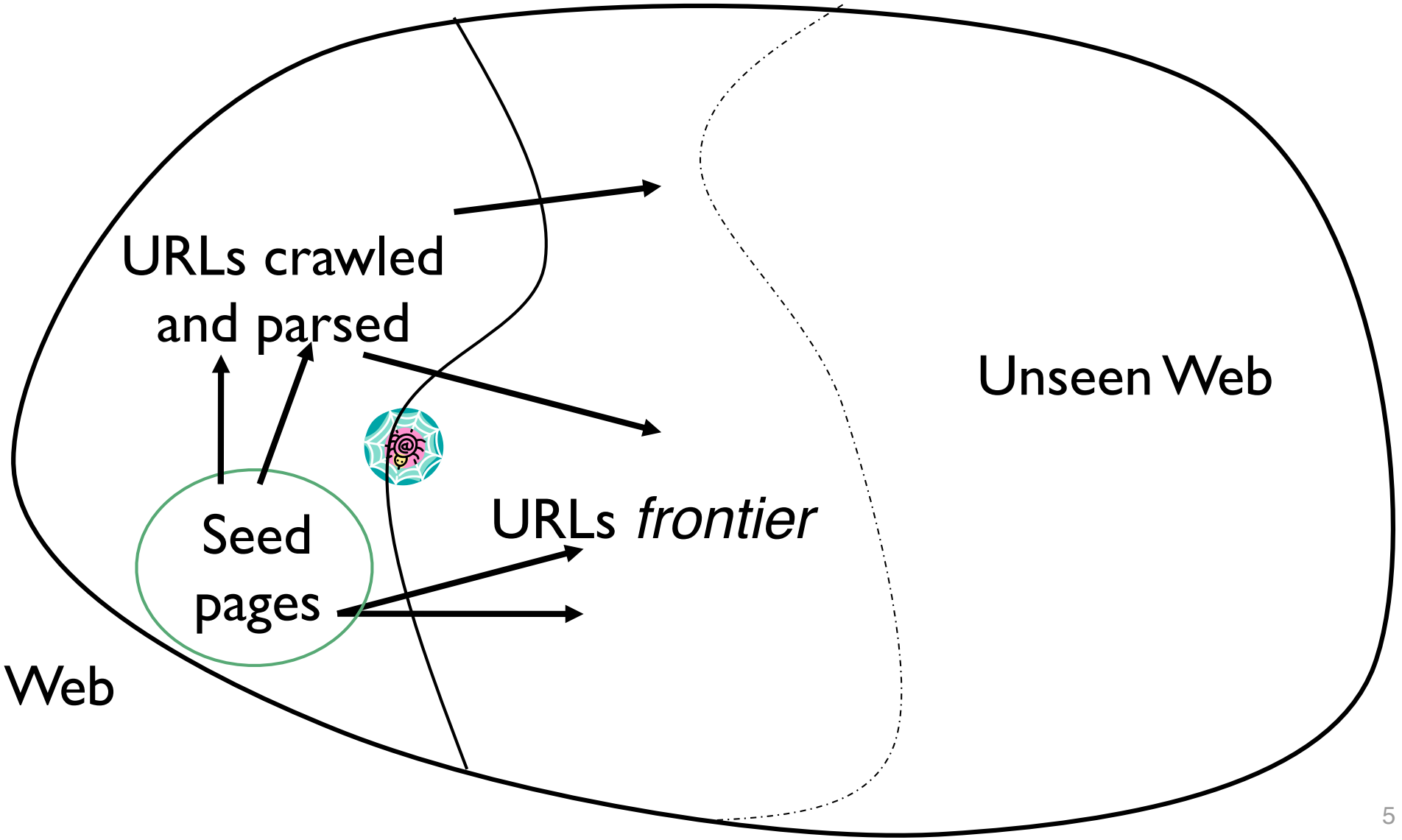
- Today: Search engine basics
  - Overall architecture
  - Crawling
  - Parsing
  - Indexing
- Thursday:
  - Learning as a key facet of search
  - Foundations of ML


Indexed corpus
Crawler
Doc Analyzer
Doc Representation
Indexer
Index
Ranking procedure
Feedback
Evaluation
(Query)
Query Rep
User
Ranker
results

Web crawler

- An automatic program that systematically browses the web for the purpose of Web content indexing and updating
  - Synonyms: spider, robot, bot

Crawling picture



How does it work

- In pseudo code

```
Def Crawler(entry_point) {
    URL_list = [entry_point]
    while (len(URL_list)>0) {
        URL = URL_list.pop();
        if (isVisited(URL) or !isLegal(URL) or !checkRobotsTxt(URL))
            continue;
        HTML = URL.open();
        for (anchor in HTML.listOfAnchors()) {
            URL_list .append(anchor);
        }
        setVisited(URL);
        insertToIndex(HTML);
    }
}
```

Which page to visit next?

Is it visited already?
Or shall we visit it

Is the access granted?

Simple picture – complications

- Web crawling isn't feasible with one machine
  - All of the above steps distributed
- Malicious pages
  - Spam pages
  - Spider traps – incl dynamically generated
- Even non-malicious pages pose challenges
  - Latency/bandwidth to remote servers vary
  - Webmasters' stipulations
    - How "deep" should you crawl a site's URL hierarchy?
  - Site mirrors and duplicate pages
- Politeness – don't hit a server too often

Robots.txt

- Protocol for giving spiders ("robots") limited access to a website, originally from 1994
  - www.robotstxt.org/wc/norobots.html
- Website announces its request on what can(not) be crawled
  - For a server, create a file `/robots.txt`
  - This file specifies access restrictions

Robots.txt example

- No robot should visit any URL starting with "/yoursite/temp/", except the robot called "searchengine":

```
User-agent: *
Disallow: /yoursite/temp/

User-agent: searchengine
Disallow:
```

Visiting strategy

- Breadth first
  - Uniformly explore from the entry page
  - Memorize all nodes on the previous level
  - As shown in pseudo code
- Depth first
  - Explore the web by branch
  - Biased crawling given the web is not a tree structure
- Focused crawling
  - Prioritize the new links by predefined strategies

OK, now we have a collection of web pages …

SECTIONS SEARCH SUBSCRIBE NOW LOG IN

ENGLISH 中文 (CHINESE) ESPAÑOL

The New York Times

Monday, January 15, 2018 | Today's Paper | Video | 59°F | S. & P. 500 +0.67% ↑

World U.S. Politics N.Y. Business Opinion Tech Science Health Sports Arts Style Food Travel Magazine T Magazine Real Estate ALL

'I'm Not a Racist,' Trump Says, as DACA Hopes Dim

By THOMAS KAPLAN, NOAH WEILAND and MICHAEL D. SHEAR

- President Trump joined two Republican senators on Sunday in disputing that he made a derogatory comment during a meeting on immigration last week.
- The outcry over his vulgar remarks overshadowed key issues facing the capital,

Dang Van Phuoc/Associated Press

If We Had Cellphone Alerts in 1968

s. Protests. What would 1968 have looked

Processing request...

Opinion

To Make Prisons 'Safer,' Some Are Banning . . . Books

By TARIRO MZEZEWA

Denying people the right to read has a sordid history.

EDITORIAL

Some Bright Hopes for New York's Schools

The city has trained a generation of highly capable administrators, and some would be excellent candidates for the top job.

- Blow: Trump Is a Racist.

Donald Trump's Racism: The Definitive List

By DAVID LEONHARDT and IAN PRASAD PHILBRICK

The media uses euphemisms when describing Trump's comments about race. Here's the truth: Donald Trump is a racist.

Guess Who's Coming to 'Peanuts'

By DAVID KAMP

The introduction, 50 years ago, of a black character into the Schulz comic strip was a major social statement

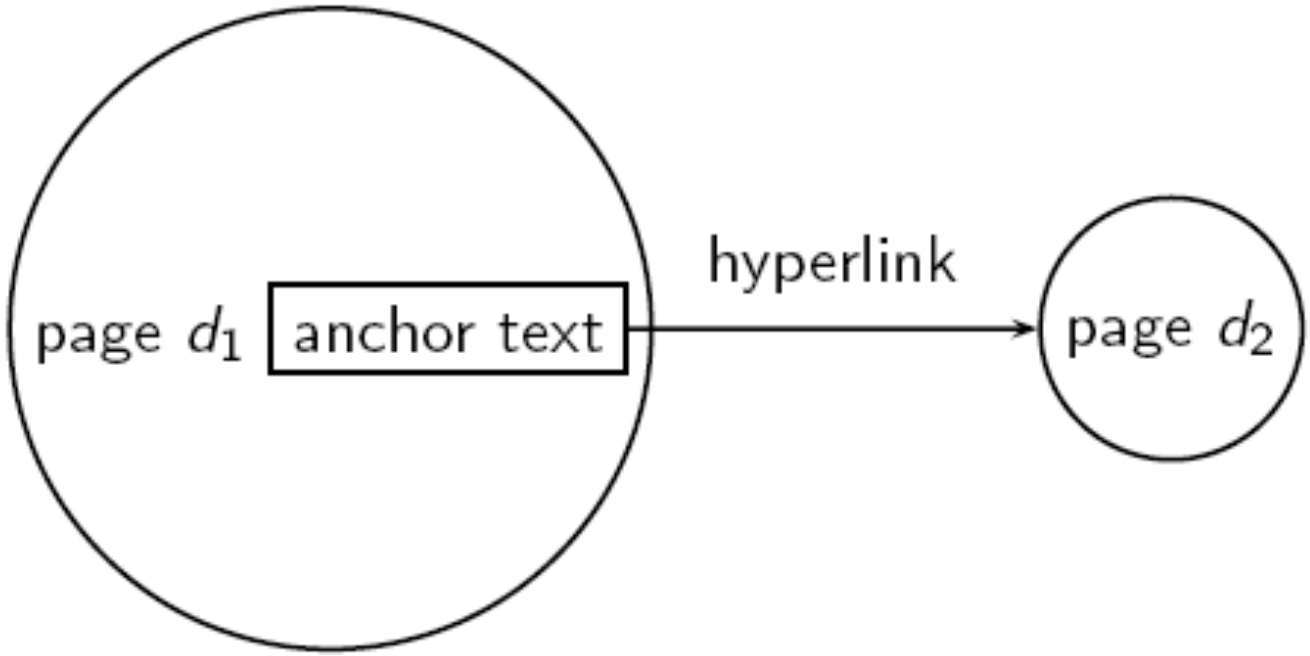
… but this is what we have

```
<!DOCTYPE html>
<!--[if (gt IE 9)|!(IE)]> <!--> <html lang="en" class="no-js edition-domestic app-homepage"  itemscope xmlns:og="http://opengraphprotocol.org/schema/"> <!--<![endif]-->
<!--[if IE 9]> <html lang="en" class="no-js ie9 lt-ie10 edition-domestic app-homepage" xmlns:og="http://opengraphprotocol.org/schema/"> <![endif]-->
<!--[if IE 8]> <html lang="en" class="no-js ie8 lt-ie10 lt-ie9 edition-domestic app-homepage" xmlns:og="http://opengraphprotocol.org/schema/"> <![endif]-->
<!--[if (lt IE 8)]> <html lang="en" class="no-js lt-ie10 lt-ie9 lt-ie8 edition-domestic app-homepage" xmlns:og="http://opengraphprotocol.org/schema/"> <![endif]-->
<head>
    <title>The New York Times - Breaking News, World News & Multimedia</title>
        <meta http-equiv="X-UA-Compatible" content="IE=edge,chrome=1" />
<link rel="shortcut icon" href="https://static01.nyt.com/favicon.ico" />
<link rel="apple-touch-icon-precomposed" sizes="144×144" href="https://static01.nyt.com/images/icons/ios-ipad-144x144.png" />
<link rel="apple-touch-icon-precomposed" sizes="114×114" href="https://static01.nyt.com/images/icons/ios-iphone-114x144.png" />
<link rel="apple-touch-icon-precomposed" href="https://static01.nyt.com/images/icons/ios-default-homescreen-57x57.png" />
<meta name="sourceApp" content="nyt-v5" />
<meta id="applicationName" name="applicationName" content="homepage" />
<meta id="foundation-build-id" name="foundation-build-id" content="" />
<link rel="canonical" href="https://www.nytimes.com" />
<link rel="alternate" type="application/rss+xml" title="RSS" href="http://www.nytimes.com/services/xml/rss/nyt/HomePage.xml" />
<link rel="alternate" media="only screen and (max-width: 640px)" href="http://mobile.nytimes.com" />
<link rel="alternate" media="handheld" href="http://mobile.nytimes.com" />
<meta name="robots" content="noarchive,noodp,noydir" />
<meta name="description" content="The New York Times: Find breaking news, multimedia, reviews & opinion on Washington, business, sports, movies, travel, books, jobs, education, real estate,
cars & more at nytimes.com." />
<meta name="CG" content="Homepage" />
<meta name="SCG" content="" />
<meta name="PT" content="Homepage" />
<meta name="PST" content="" />
<meta name="application-name" content="The New York Times" />
<meta name="msapplication-starturl" content="https://www.nytimes.com" />
<meta name="msapplication-task" content="name=Search;action-uri=http://query.nytimes.com/search/sitesearch?src=iepin;icon-uri=https://static01.nyt.com/images/icons/search.ico" />
<meta name="msapplication-task" content="name=Most Popular;action-uri=http://www.nytimes.com/gst/mostpopular.html?src=iepin;icon-uri=https://static01.nyt.com/images/icons/mostpopular.ico" />
<meta name="msapplication-task" content="name=Video;action-uri=http://video.nytimes.com/?src=iepin;icon-uri=https://static01.nyt.com/images/icons/video.ico" />
<meta name="msapplication-task" content="name=Homepage;action-uri=https://www.nytimes.com?src=iepin&amp;adxnnl=1;icon-uri=https://static01.nyt.com/images/icons/homepage.ico" />
<meta property="og:url" content="https://www.nytimes.com" />
<meta property="og:type" content="website" />
<meta property="og:title" content="Breaking News, World News & Multimedia" />
<meta property="og:description" content="The New York Times: Find breaking news, multimedia, reviews & opinion on Washington, business, sports, movies, travel, books, jobs, education, real
estate, cars & more at nytimes.com." />
<meta property="og:image" content="https://static01.nyt.com/images/icons/t_logo_291_black.png" />
<meta property="fb:app_id" content="9869919170" />
<meta name="apple-itunes-app" content="app-id=357066198, affiliate-data=at=10lIEQ&ct=Web%20iPad%20Smart%20App%20Banner&pt=13036" />
<meta name="keywords" content="United States Politics and Government,Perdue, David A Jr,Cotton, Tom,Trump, Donald J,Durbin, Richard J,Immigration and Emigration,Blacks,Race and
Ethnicity,Ministers (Protestant),United States Politics and Government,African Methodist Episcopal Church,Immigration and Emigration,Discrimination,Trump, Donald J,We Shall Overcome
(Song),United States Politics and Government,Immigration and Emigration,Executive Orders and Memorandums,Courts and the Judiciary,Deferred Action for Childhood Arrivals,Deferred Action for
Childhood Arrivals,Immigration and Emigration,Decisions and Verdicts,Citizenship and Immigration Services (US),Trump, Donald J,United States Politics and Government,United States Defense and
Military Forces,United States International Relations,United States Special Operations Command,United States Air Force,Joint Chiefs of Staff,United States Army,Flournoy, Michele A,Kim Jong-
un,Mattis, James N,Milley, Mark A,Tillerson, Rex W,North Korea,Nuclear Weapons,False Alarms,United States International Relations,USSR (Former Soviet Union),Reagan, Ronald Wilson,Cold War
Era,Politics and Government,False Alarms,Hawaii,Ige, David Y.,Colleen Hanabusa,Elections, Governors,Hawaii Tourism Authority,Federal Communications Commission,Hawaii Emergency Management
Agency,Gabbard, Tulsi (1981- ),Hawaii,Wireless Communications,Disasters and Emergencies,Cellular Telephones,Federal Communications Commission,Federal Emergency Management
Agency,Iran,Demonstrations, Protests and Riots,Khamenei, Ali,Rouhani, Hassan,Politics and Government,Landslides and Mudslides,Wildfires,Montecito (Calif),California,Evacuations and
Evacuees,Landslides and Mudslides,Roads and Traffic,California,Highway 101,Santa Barbara (Calif),Montecito (Calif),Football,New Orleans Saints,Minnesota Vikings,Diggs, Stefan (1994-
),Demonstrations, Protests and Riots,Vietnam War,Blacks,Presidential Election of 1968,Nineteen Hundred Sixties,Police Brutality, Misconduct and Shootings,Civil Rights Movement (1954-68),King,
Martin Luther Jr Kennedy, Robert Francis Nixon, Richard Milhous Work-Life Balance Telecommuting Freelancing, Self-Employment and Independent Contracting Cholesterol Heart Statins (Cholesterol-
```

Analyze text + links

- How to extract text? links?
- How to represent?
- Do we keep everything? Skip some parts?
- How to deal with duplicates?

Anchor text



- **Assumption: The anchor text describes the target page**
  - We use anchor text somewhat loosely here: the text surrounding the hyperlink. Example: "You can find cheap cars here."

[document text only] vs. [document text + anchor text]

- Searching on [document text + anchor text] is often more effective than searching on [document text only].
- Example: Query ***IBM***
  - Matches IBM's copyright page
  - Matches many spam pages
  - Matches IBM wikipedia article
  - May not match IBM home page! (if IBM home page is mostly graphical)
- Searching on anchor text is better for the query IBM.
- **Represent each page by all the anchor text pointing to it.**
- In this representation, the page with the most occurrences of IBM is www.ibm.com.

Full text indexing

- Bag-of-Words representation
  - Doc1: Information retrieval is helpful for everyone.
  - Doc2: Helpful information is retrieved for you.

	information	retrieval	retrieved	is	helpful	for	you	everyone
Doc1	1	1	0	1	1	1	0	1
Doc2	1	0	1	1	1	1	1	0

Word-document adjacency matrix

Full text indexing

- Bag-of-Words representation
  - Assumption: word is independent from each other
  - Pros: simple
  - Cons: grammar and order are missing
  - ***The most frequently used document representation***
    - ***Image, speech, gene sequence***

Full text indexing

- Improved Bag-of-Words representation
  - N-grams: a contiguous sequence of n items from a given sequence of text
    - E.g., Information retrieval is helpful for everyone
    - Bigrams: ‘information_retrieval’, ‘retrieval_is’, ‘is_helpful’, ‘helpful_for’, ‘for_everyone’
  - Pros: capture local dependency and order
  - Cons: purely statistical view, increase vocabulary size $O(V^N)$

Full text indexing

- Index document with all the occurring word
  - Pros
    - Preserve all information in the text (hopefully)
    - Fully automatic
  - Cons
    - Vocabulary gap: cars v.s., car
    - Large storage: e.g., in N-grams $O(V^N)$
  - Solution
    - Construct controlled vocabulary

Zipf's law

- The *i*th most frequent term has frequency proportional to *1/i*.
- cf is the collection frequency: the number of occurrences of the term in the collection
- A few words occur very often
- Many words are infrequent
- Zipf, 1902-1950: linguistic prof at Harvard

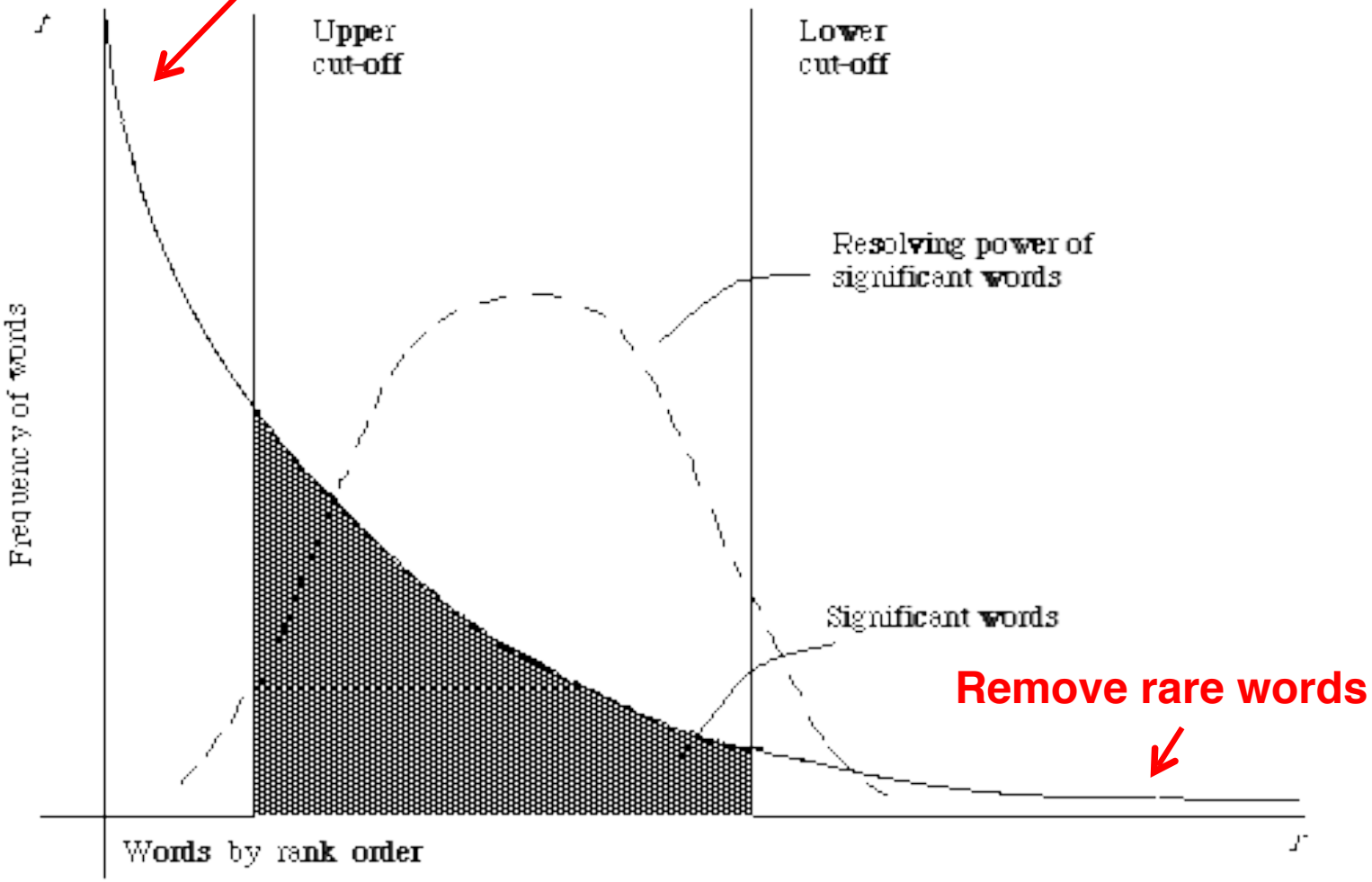
$$\mathrm{cf}_i \propto \frac{1}{i}$$

Zipf's law tells us

- Head words may take large portion of occurrence, but they are semantically meaningless
  - E.g., the, a, an, we, do, to
- Tail words take major portion of vocabulary, but they rarely occur in documents
  - E.g., dextrosinistral
- The rest is most representative
  - To be included in the controlled vocabulary

Automatic text indexing

Remove non-informative words



Remove rare words

Figure 2.1. A plot of the hyperbolic curve relating f, the frequency of occurrence and r, the rank order (Adapted from Schultz[44] page 120)

Stopwords

- Useless words for query/document analysis
  - Not all words are informative
  - Remove such words to reduce vocabulary size
  - No universal definition
  - Risk: break the original meaning and structure of text
    - E.g., this is not a good option -> option

      to be or not to be -> null

Normalization

- Convert different forms of a word to normalized form in the vocabulary
  - U.S.A -> USA, St. Louis -> Saint Louis
- Solution
  - Rule-based
    - Delete periods and hyphens
    - All in lower case
  - Dictionary-based
    - Construct equivalent class
      - Car -> “automobile, vehicle”
      - Mobile phone -> “cellphone”

Stemming

- Reduce inflected or derived words to their root form
  - Plurals, adverbs, inflected word forms
    - E.g., ladies -> lady, referring -> refer, forgotten -> forget
  - Bridge the vocabulary gap
  - Risk: lose precise meaning of the word
    - E.g., lay -> lie (a false statement? or be in a horizontal position?)
  - Solutions (for English)
    - Porter stemmer: pattern of vowel-consonant sequence
    - Krovetz Stemmer: morphological rules

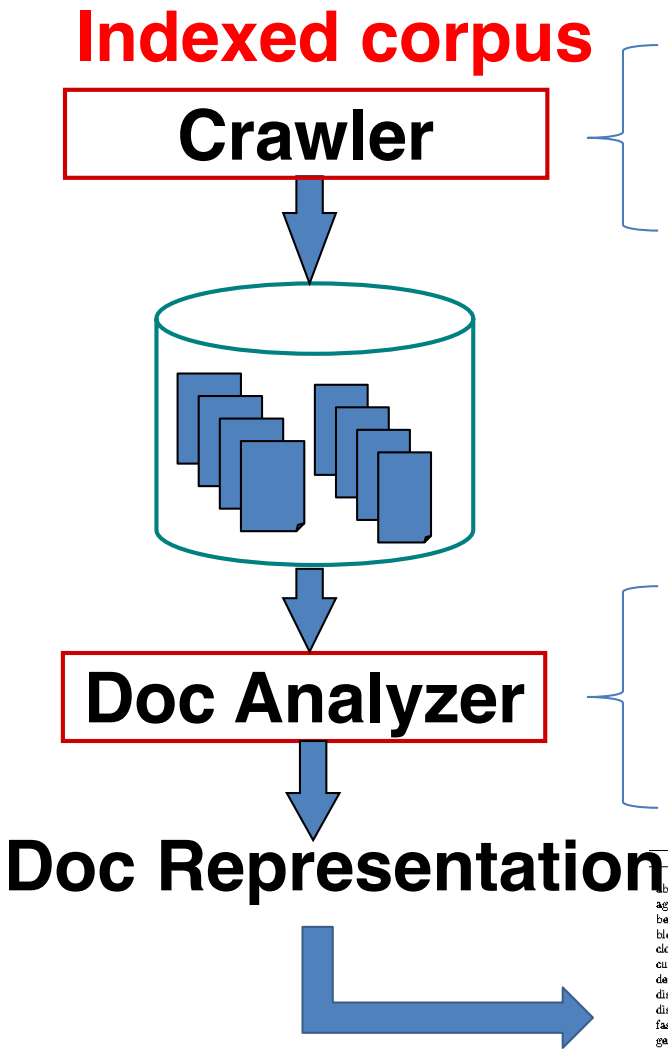
Pair up … and …

Create a set of rules to tokenize this paragraph:

The Texas A&M Aggies, buoyed by their victory over South Carolina, moved up 12 spots to No. 9 in the AP Top 25 after the opening weekend of college football. The top four in the rankings -- Florida State, Alabama, Oregon and Oklahoma -- are unchanged, but the No. 1 Seminoles and No. 2 Crimson Tide lost some support in the first poll of the regular season after close victories against heavy underdogs. Texas A&M began the post-Johnny Manziel era with a 52-28 victory at South Carolina. The loss dropped the Gamecocks from No. 9 to No. 21.

What are the tokens emitted by your approach?

Abstraction of search engine architecture



1. Visiting strategy
2. Avoid duplicated visit
3. Re-visit policy

1. HTML parsing
2. Tokenization
3. Stemming/normalization
4. Stopword/controlled vocabulary filter

Terms	Documents													
	M1	M2	M3	M4	M5	M6	M7	M8	M9	M10	M11	M12	M13	M14
abnormalities	0	0	0	0	0	0	0	1	0	1	0	0	0	0
age	1	0	0	0	0	0	0	0	0	0	0	1	0	0
behavior	0	0	0	0	1	1	0	0	0	0	0	0	0	0
blood	0	0	0	0	0	0	0	1	0	0	1	0	0	0
close	0	0	0	0	0	0	1	0	0	0	1	0	0	0
culture	1	1	0	0	0	0	0	1	1	0	0	0	0	0
depressed	1	0	1	1	1	0	0	0	0	0	0	0	0	0
discharge	1	1	0	0	0	1	0	0	0	0	0	0	0	0
disease	0	0	0	0	0	0	0	0	1	0	1	0	0	0
fast	0	0	0	0	0	0	0	0	0	1	0	1	1	1
generation	0	0	0	0	0	0	0	0	1	0	0	0	1	0
oestrogen	0	0	1	1	0	0	0	0	0	0	0	0	0	0
patients	1	1	0	1	0	0	0	1	0	0	0	0	0	0
pressure	0	0	0	0	0	0	0	0	0	0	1	0	0	1
rats	0	0	0	0	0	0	0	0	0	0	0	0	1	1
respect	0	0	0	0	0	0	0	1	0	0	0	1	0	0
rise	0	0	0	1	0	0	0	0	0	0	0	0	0	1
study	1	0	1	0	0	0	0	0	1	0	0	0	0	0

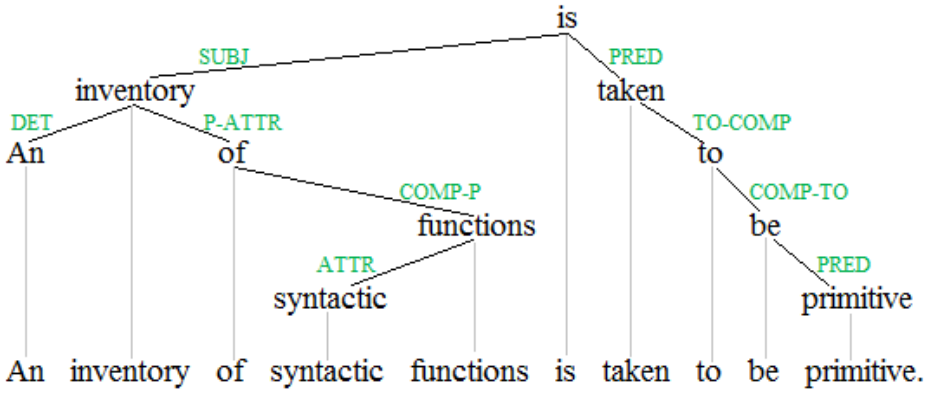
BagOfWord representation!

Automatic text indexing

- In modern search engine

Query: "to be or not to be"

  - **No** stemming or stopword removal, since computation and storage are no longer the major concern
  - More advanced NLP techniques are applied
    - Named entity recognition
      - E.g., people, location and organization
    - Dependency parsing



Alternative index structures

How can we improve on the basic index?

- Need a better index than simple <term: docs>
  - **Skip pointers:** faster postings merges
  - **Positional index:** Phrase queries and Proximity queries
  - **Permuterm index:** Wildcard queries
  - **k-gram index:** Wildcard queries and spell correction

Shifting gears: Learning is fundamental to modern search engines

Baron Schwartz 

@xaprb

Follow

When you're fundraising, it's AI
When you're hiring, it's ML
When you're implementing, it's linear regression
When you're debugging, it's printf()

9:52 PM - 14 Nov 2017

4,574 Retweets **10,348** Likes

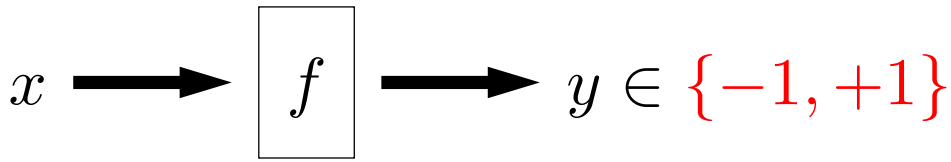
73 4.6K 10K

Learning

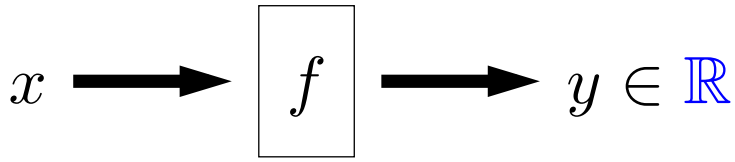
- **Regression**: trying to predict a real value
- **Binary classification:** trying to predict a simple yes/no response
- **Multiclass classification:** trying to put an example into one of a number of classes
- **Ranking:** trying to put a set of objects in order of relevance

Types of prediction tasks

Binary classification (e.g., email $\Rightarrow$ spam/not spam):

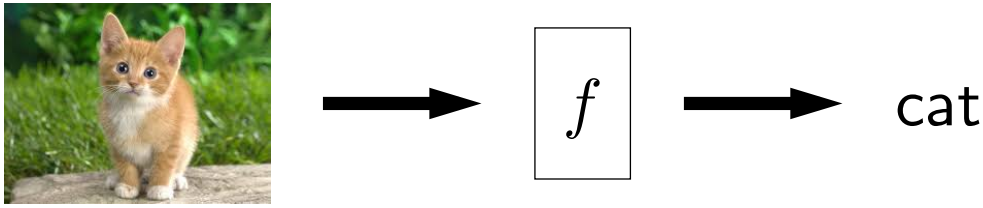


Regression (e.g., location, year $\Rightarrow$ housing price):

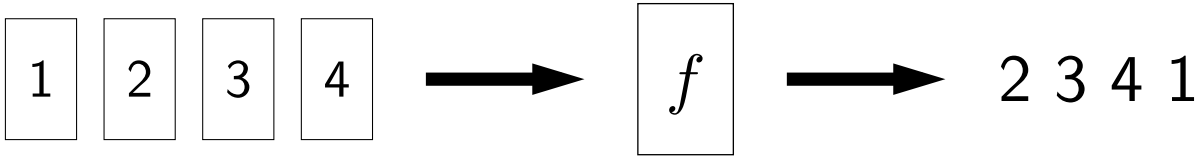


Types of prediction tasks

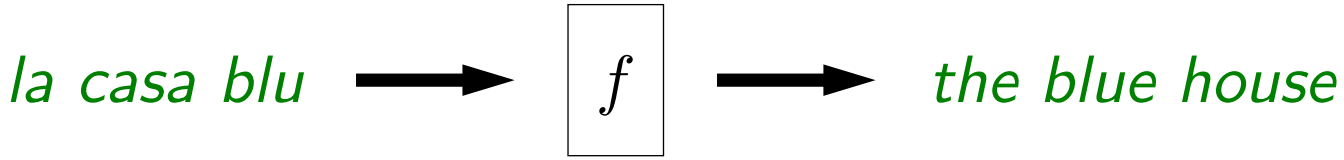
Multiclass classification: y is a category



Ranking: y is a permutation



Structured prediction: y is an object which is built from parts



90° Earthquuuuuuuaaakesss!!!

the ONION
America's Finest News Source

Grandmother Classifies 79% Of Everything A Shame

NEWS · Family · Local · ISSUE 47·47 ISSUE 45·29 · Jul 18, 2009

SANDUSKY, OH—According to those close to Gertrude Wharton, the grandmother of nine declares 79 percent of everything she witnesses, experiences, or hears about from friends to be "a shame."

"No matter what happens, her response is always, 'That's a shame,'" said Wharton's son Kevin, 46. "From the recent passing of her friend Lillian to the fact that her coupon for chicken bouillon cubes expired last week, I can't have a conversation with her without being told something is a shame. Is this really how she

Formal definition of Text Classification: Training

Given:

- A document space X
  - Documents are represented in this space – typically some type of high-dimensional space.
- A fixed set of classes $C = \{c_1, c_2, \ldots, c_J\}$
  - The classes are human-defined for the needs of an application (e.g., relevant vs. nonrelevant).
- A training set D of labeled documents with each labeled document $\langle d, c \rangle \in X \times C$

Using a learning method or learning algorithm, we then wish to

learn a classifier Υ that maps documents to classes:

$$\Upsilon : X \rightarrow C$$

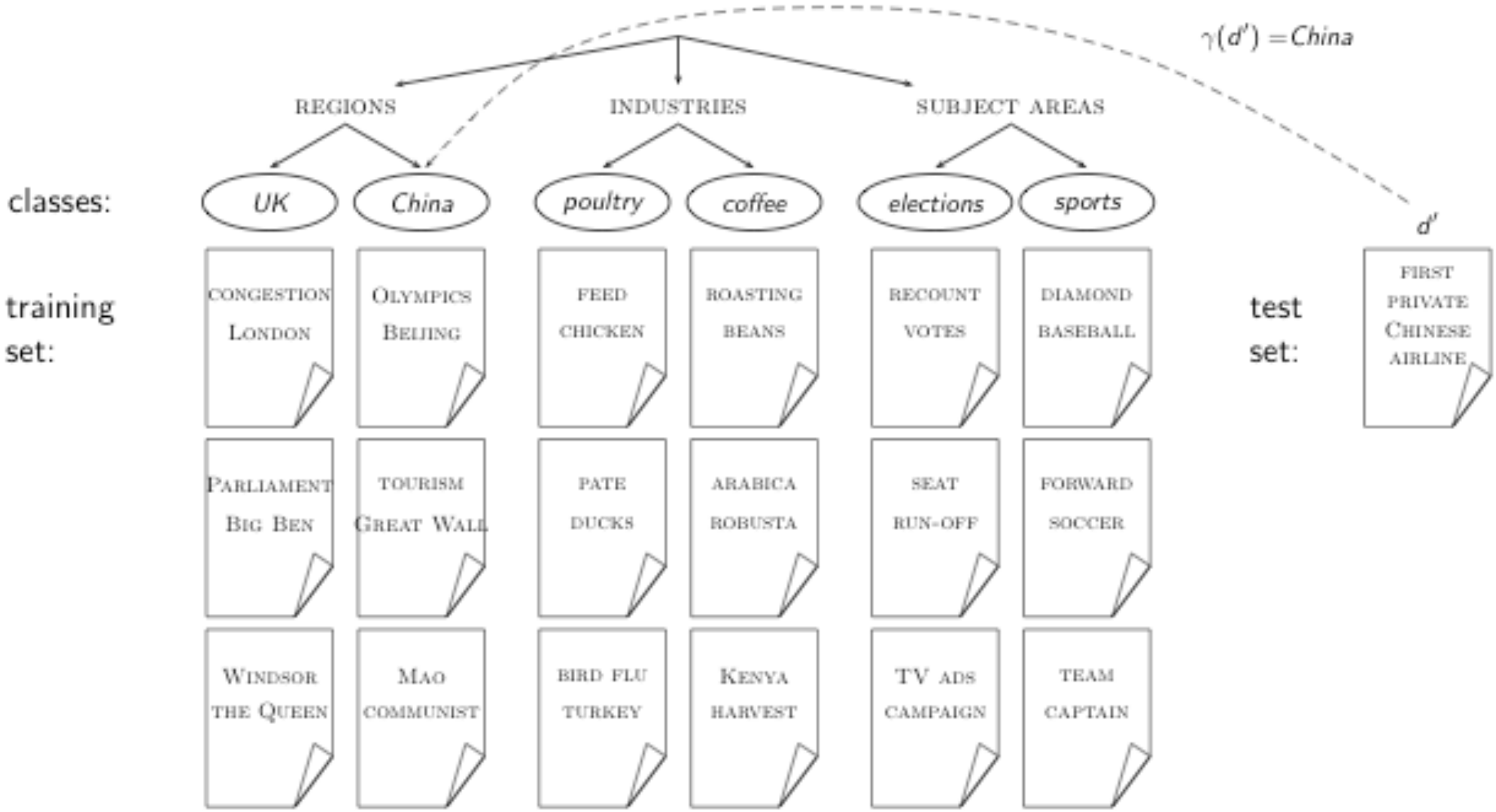
Formal definition of Text Classification: Application/Testing

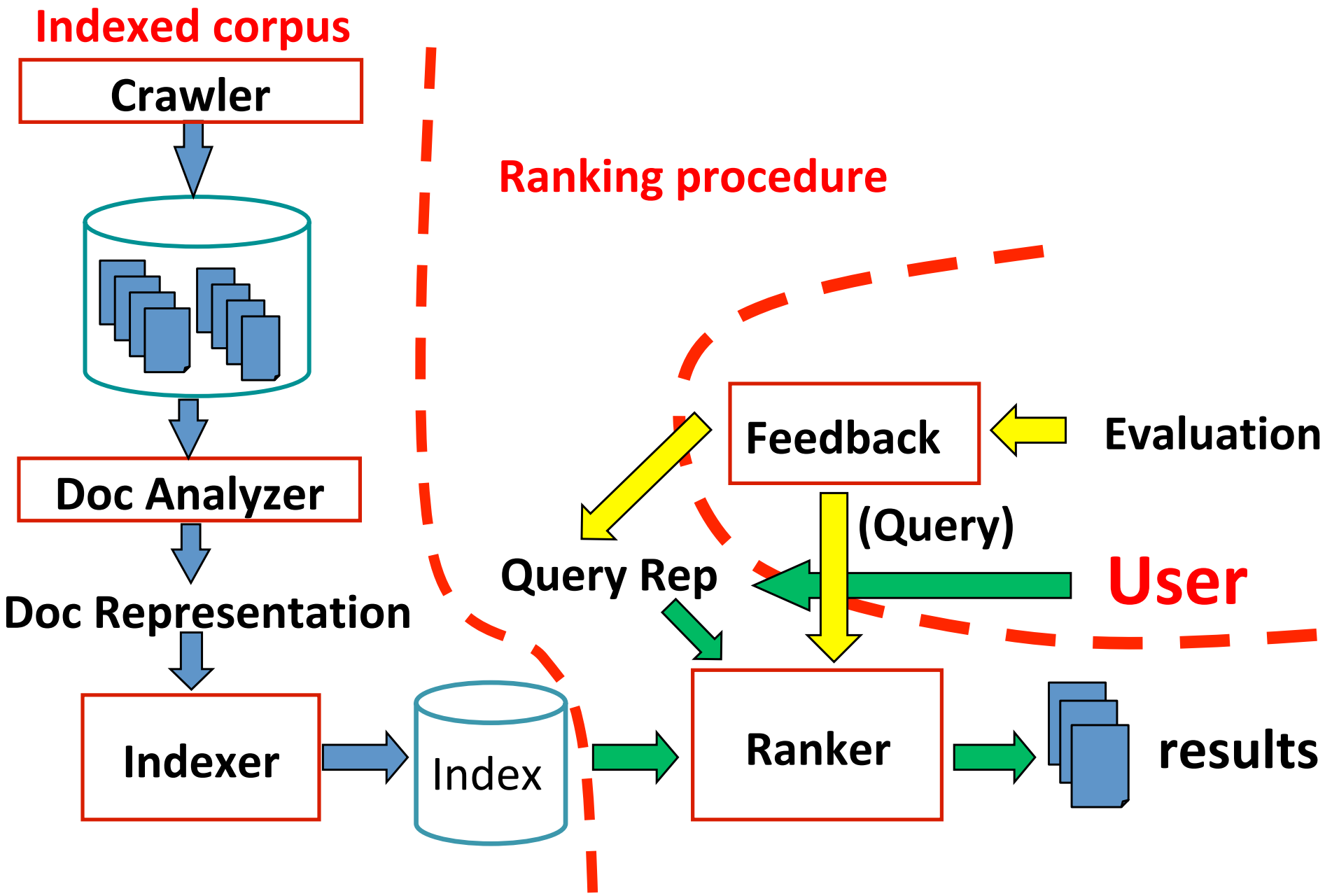
Given: a description $d \in X$ of a document

Determine: $\Upsilon(d) \in C$,

that is, the class that is most appropriate for d

Example: Topic classification




Indexed corpus
Crawler
Doc Analyzer
Doc Representation
Indexer
Index
Ranking procedure
Feedback
Evaluation
(Query)
Query Rep
User
Ranker
results

Examples of how search engines use classification

- Crawling —> Classify pages as news or not —> impacts crawling frequency
- Indexing —> Classify pages into tiers — for improving search quality — so “high class”, middle class, landfill
- Classify users (personalization) —> TAMU or not, Harvard or not, location, country (but that’s easy!), language (also probably easy), my interests (i want to buy something or not)
- Craigslist —> classify users as browsing (not buying), ready to buy
- More generally, query intent (buy or not buy)
- Queries — return a map? or an infobox? or what?who /what /where?
- Ranking —> Classify document-query pairs as Perfect, Good, etc. (that’s really hard, and something we would like to do)
- Web pages —> topic (sports? or computers? or what?), malicious or not, often updated or not, adult content or not, language of the page, location of the page, static vs dynamic page,

Classification methods: 1. Manual

- Manual classification was used by Yahoo in the beginning of the web. Also: ODP, PubMed
- Very accurate if job is done by experts
- Consistent when the problem size and team is small
- Scaling manual classification is difficult and expensive.
- → We need automatic methods for classification.

Classification methods: 2. Rule-based

- There are IDE-type development environments for writing very complex rules efficiently. (e.g., Verity)
- Often: Boolean combinations (as in Google Alerts)
- Accuracy is very high if a rule has been carefully refined over time by a subject expert.
- Building and maintaining rule-based classification systems is cumbersome and expensive.

A Verity topic (a complex classification rule)

comment line
top-level topic
topic definition modifiers
subtopic topic
evidence topic
topic definition modifier
evidence topic
topic definition modifier
evidence topic
topic definition modifier
subtopic
subtopic
subtopic

```
# Beginning of art topic definition
art ACCRUE
    /author = "fsmith"
    /date   = "30-Dec-01"
    /annotation = "Topic created
                   by fsmith"
* 0.70 performing-arts ACCRUE
** 0.50 WORD
    /wordtext = ballet
** 0.50 STEM
    /wordtext = dance
** 0.50 WORD
    /wordtext = opera
** 0.30 WORD
    /wordtext = symphony
* 0.70 visual-arts ACCRUE
** 0.50 WORD
    /wordtext = painting
** 0.50 WORD
    /wordtext = sculpture
* 0.70 film ACCRUE
** 0.50 STEM
    /wordtext = film
** 0.50 motion-picture PHRASE
*** 1.00 WORD
    /wordtext = motion
*** 1.00 WORD
    /wordtext = picture
** 0.50 STEM
    /wordtext = movie
* 0.50 video ACCRUE
** 0.50 STEM
    /wordtext = video
** 0.50 STEM
    /wordtext = vcr
# End of art topic
```

Classification methods: 3. Statistical/ Probabilistic

- This was our definition of the classification problem – text classification as a learning problem
- (i) Supervised learning of the classification function Υ and

  (ii) its application to classifying new documents

- No free lunch: requires hand-classified training data
- But this manual classification can be done by non-experts.

Vector Space Classification

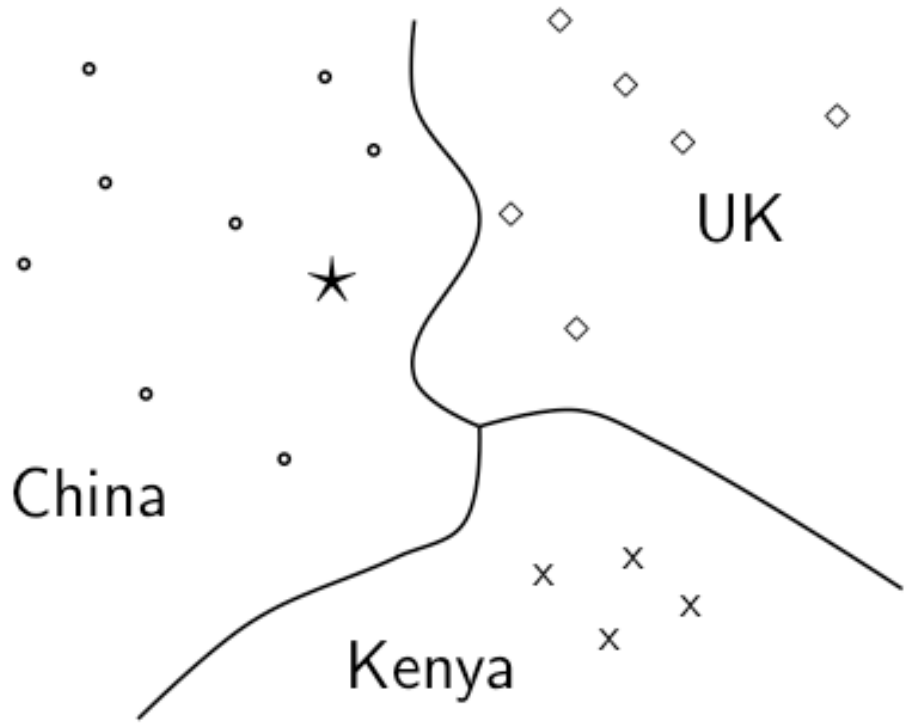
Vector space representation

- Each document is a vector, one component for each term.
- Terms are axes.
- High dimensionality: 100,000s of dimensions
- Normalize vectors (documents) to unit length
- How can we do classification in this space?

Vector space classification

- As before, the training set is a set of documents, each labeled with its class.
- In vector space classification, this set corresponds to a labeled set of points or vectors in the vector space.
- Premise 1: Documents in the same class form a contiguous region.
- Premise 2: Documents from different classes don't overlap.
- We define lines, surfaces, hypersurfaces to divide regions.

Classes in the vector space



Should the document ⋆ be assigned to China, UK or Kenya?
Find separators between the classes
Based on these separators: ⋆ should be assigned to China
How do we find separators that do a good job at classifying new documents like ⋆? – Main topic of today

Rocchio

Rocchio classification: Basic idea

- Compute a centroid for each class
  - The centroid is the average of all documents in the class.
- Assign each test document to the class of its closest centroid.

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 $\vec{v}(d)$ is the vector space representation of *d*.

Rocchio algorithm

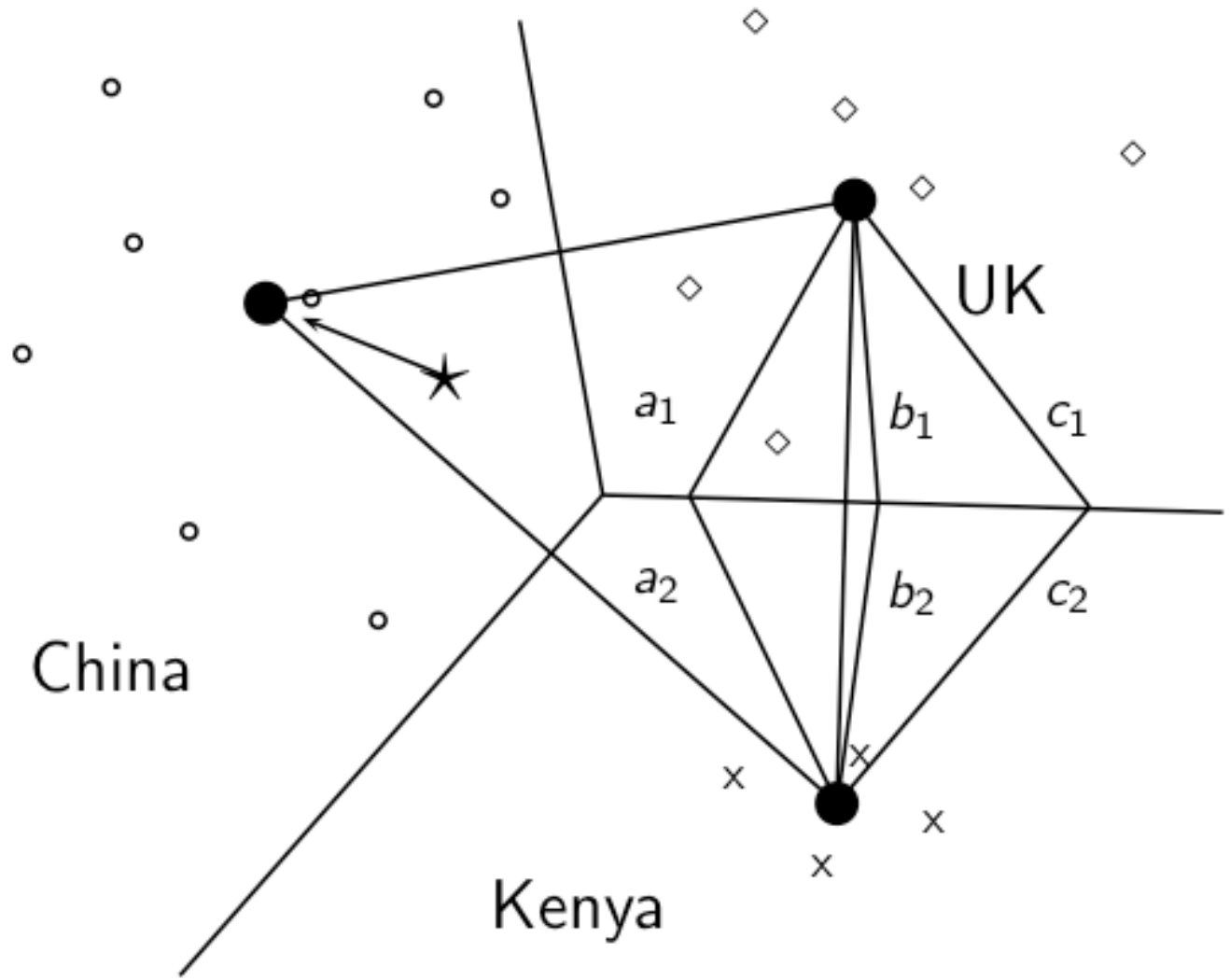
TRAINROCCHIO($\mathbb{C}, \mathbb{D}$)

1 **for each** $c_j \in \mathbb{C}$
2 **do** $D_j \leftarrow \{d : \langle d, c_j \rangle \in \mathbb{D}\}$
3 $\quad \vec{\mu}_j \leftarrow \frac{1}{|D_j|} \sum_{d \in D_j} \vec{v}(d)$
4 **return** $\{\vec{\mu}_1, \ldots, \vec{\mu}_J\}$

APPLYROCCHIO($\{\vec{\mu}_1, \ldots, \vec{\mu}_J\}, d$)

1 **return** $\arg\min_j |\vec{\mu}_j - \vec{v}(d)|$

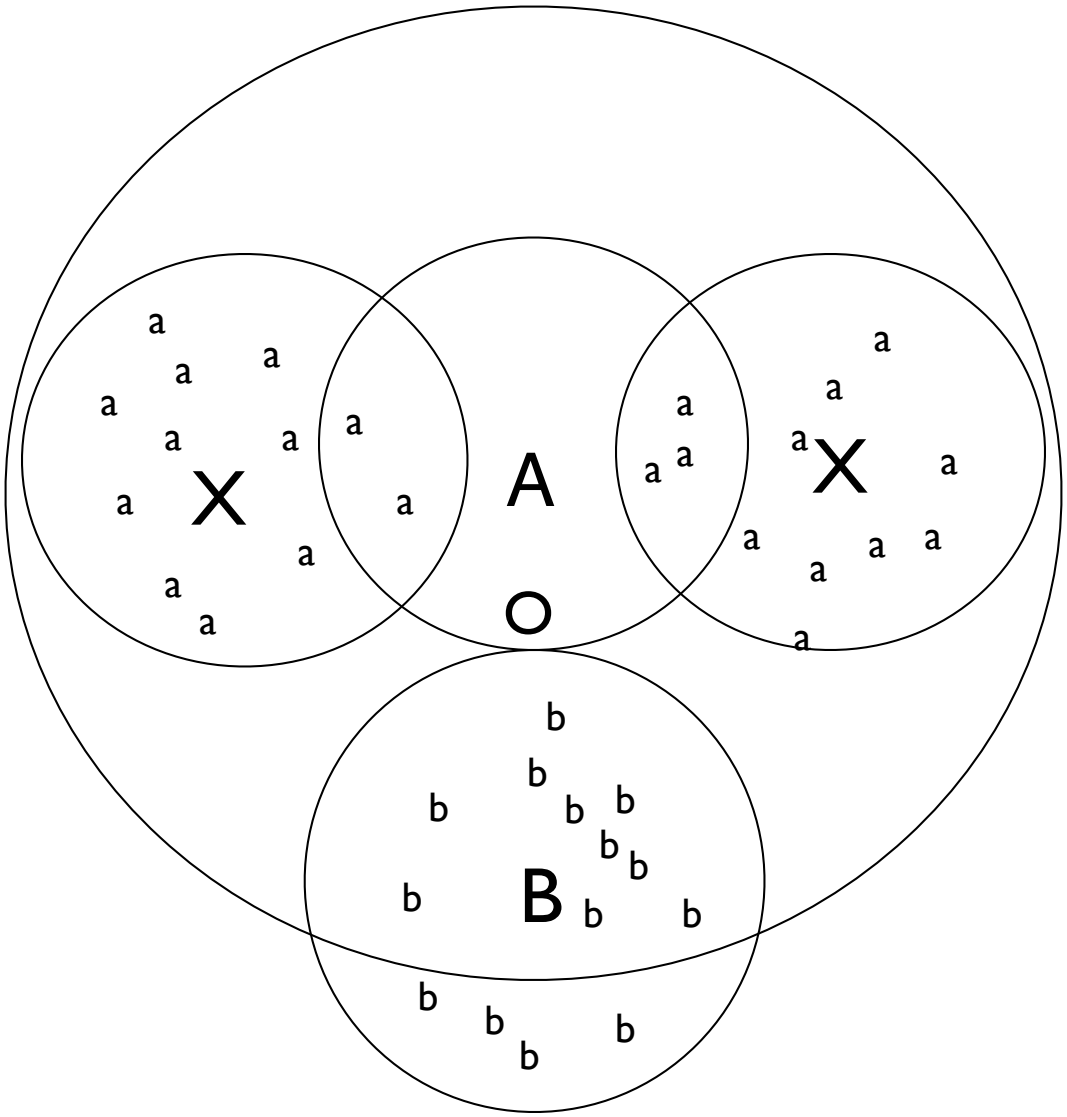
Rocchio illustrated : a1 = a2, b1 = b2, c1 = c2



Rocchio properties

- Rocchio forms a simple representation for each class: the centroid
  - We can interpret the centroid as the prototype of the class.
- Classification is based on similarity to / distance from centroid/prototype.
- Does not guarantee that classifications are consistent with the training data!

Rocchio cannot handle nonconvex, multimodal classes



Exercise: Why is Rocchio not expected to do well for the classification task a vs. b here?

- A is centroid of the a's, B is centroid of the b's.
- The point o is closer to A than to B.
- But o is a better fit for the b class.
- A is a multimodal class with two prototypes.
- But in Rocchio we only have one prototype.

kNN

kNN classification

- kNN classification is another vector space classification method.
- It also is very simple and easy to implement.
- kNN is more accurate (in most cases) than Naive Bayes and Rocchio.
- If you need to get a pretty accurate classifier up and running in a short time . . .
- . . . and you don’t care about efficiency that much . . .
- . . . use kNN.

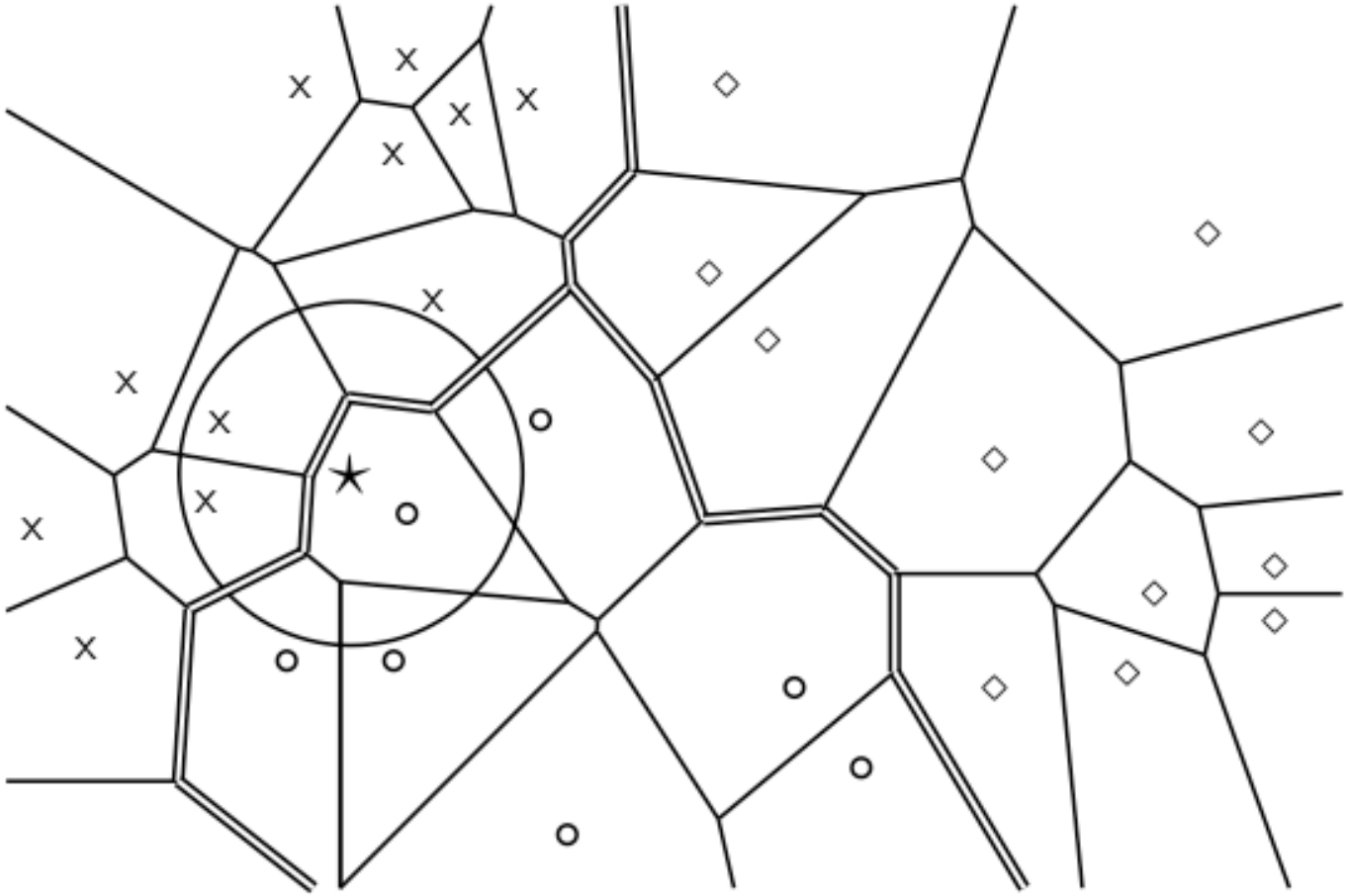
kNN classification

- kNN = *k* nearest neighbors
- kNN classification rule for $k = 1$ (1NN): Assign each test document to the class of its nearest neighbor in the training set.
- 1NN is not very robust – one document can be mislabeled or atypical.
- kNN classification rule for $k > 1$ (kNN): Assign each test document to the majority class of its *k* nearest neighbors in the training set.
- Rationale of kNN: contiguity hypothesis
  - We expect a test document d to have the same label as the training documents located in the local region surrounding *d*.

Probabilistic kNN

- Probabilistic version of kNN: $P(c|d)$ = fraction of k neighbors of d that are in c
- kNN classification rule for probabilistic kNN: Assign d to class c with highest $P(c|d)$

Probabilistic kNN



1NN, 3NN classification decision for star?

kNN algorithm

TRAIN-KNN($\mathbb{C}, \mathbb{D}$)
1 $\mathbb{D}' \leftarrow$ PREPROCESS($\mathbb{D}$)
2 $k \leftarrow$ SELECT-K($\mathbb{C}, \mathbb{D}'$)
3 **return** $\mathbb{D}', k$

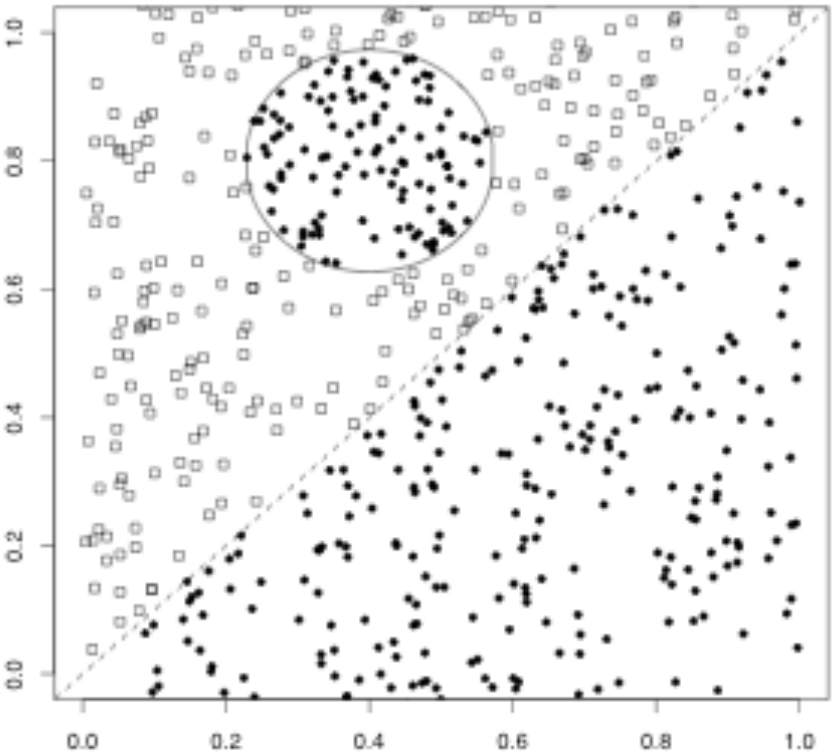
APPLY-KNN($\mathbb{D}', k, d$)
1 $S_k \leftarrow$ COMPUTENEARESTNEIGHBORS($\mathbb{D}', k, d$)
2 **for each** $c_j \in \mathbb{C}(\mathbb{D}')$
3 **do** $p_j \leftarrow |S_k \cap c_j|/k$
4 **return** $\arg\max_j p_j$

Exercise

How is star classified by:

(i) 1-NN (ii) 3-NN (iii) 9-NN (iv) 15-NN (v) Rocchio?

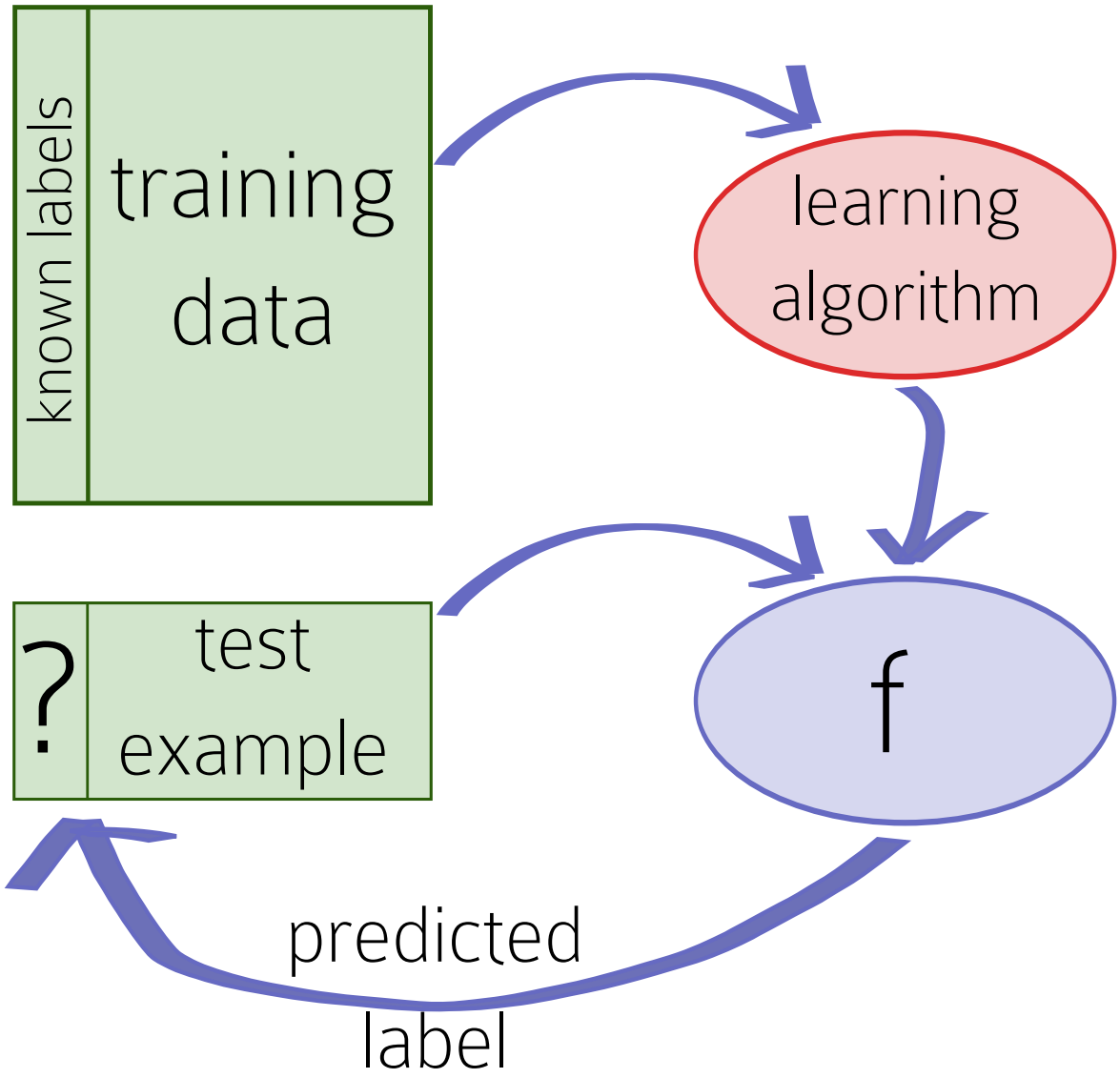
A nonlinear problem

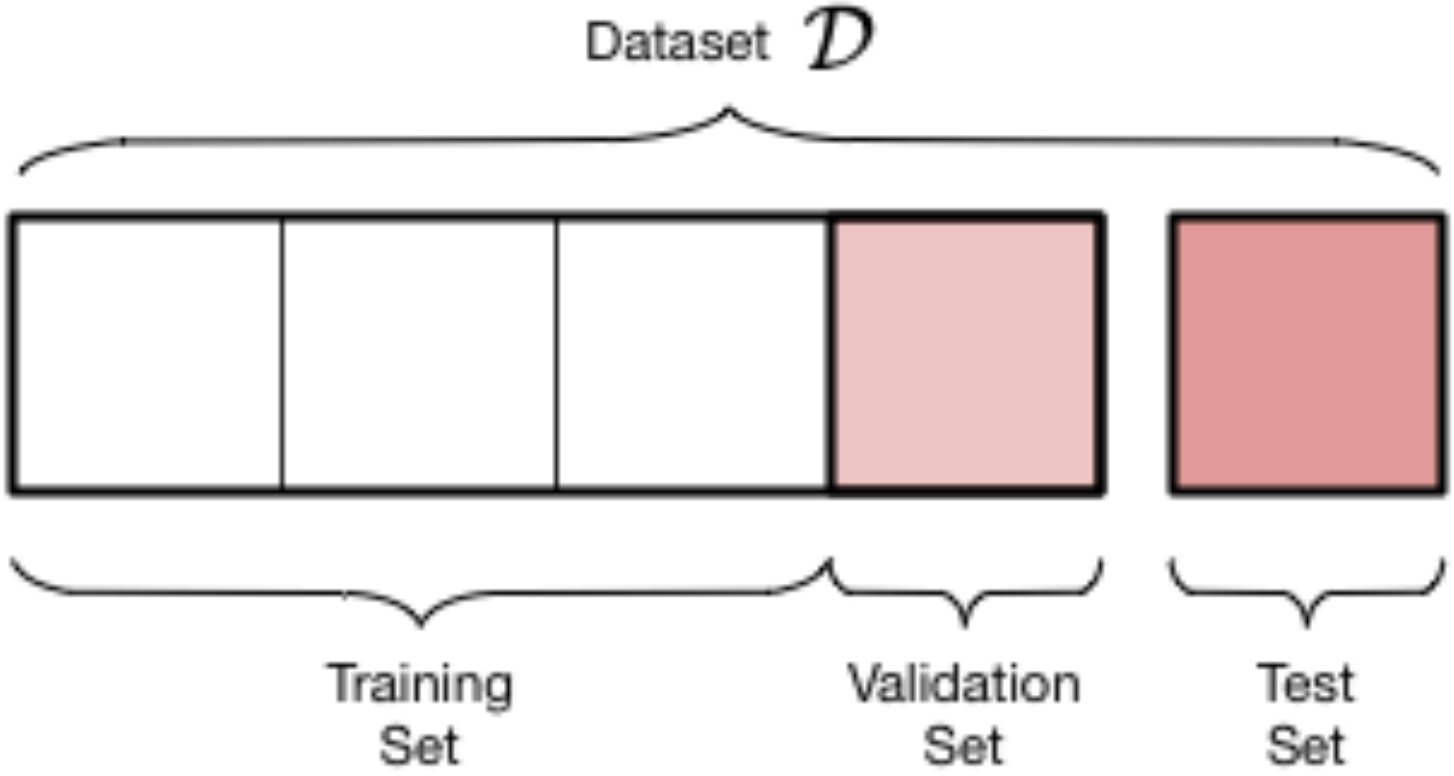


- Linear classifier like Rocchio does badly on this task.
- kNN will do well (assuming enough training data)

kNN: Discussion

- No training necessary
  - But linear preprocessing of documents is as expensive as training Naive Bayes.
  - We always preprocess the training set, so in reality training time of kNN is linear.
- kNN is very accurate if training set is large.
- Optimality result: asymptotically zero error if Bayes rate is zero.
- But kNN can be very inaccurate if training set is small.


known labels
training data
learning algorithm
?
test example
f
predicted label


Dataset $\mathcal{D}$
Training Set
Validation Set
Test Set

Framework