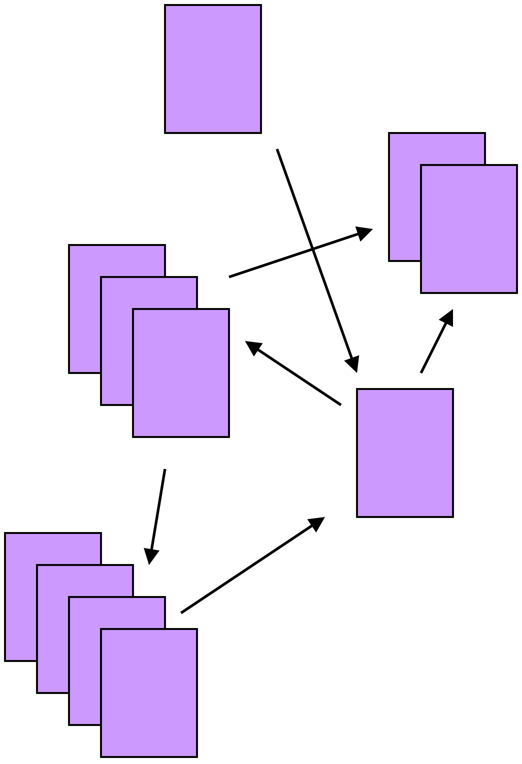
Introduction to

Information Retrieval

Ch 19 Web Search Basics

The Web document collection



The Web

- No design/co-ordination
- Distributed content creation, linking, democratization of publishing
- Content includes truth, lies, obsolete information, contradictions …
- Unstructured (text, html, …), semi-structured (XML, annotated photos), structured (Databases)…
- Scale much larger than previous text collections … but corporate records are catching up
- Growth – slowed down from initial “volume doubling every few months” but still expanding
- Content can be *dynamically generated*

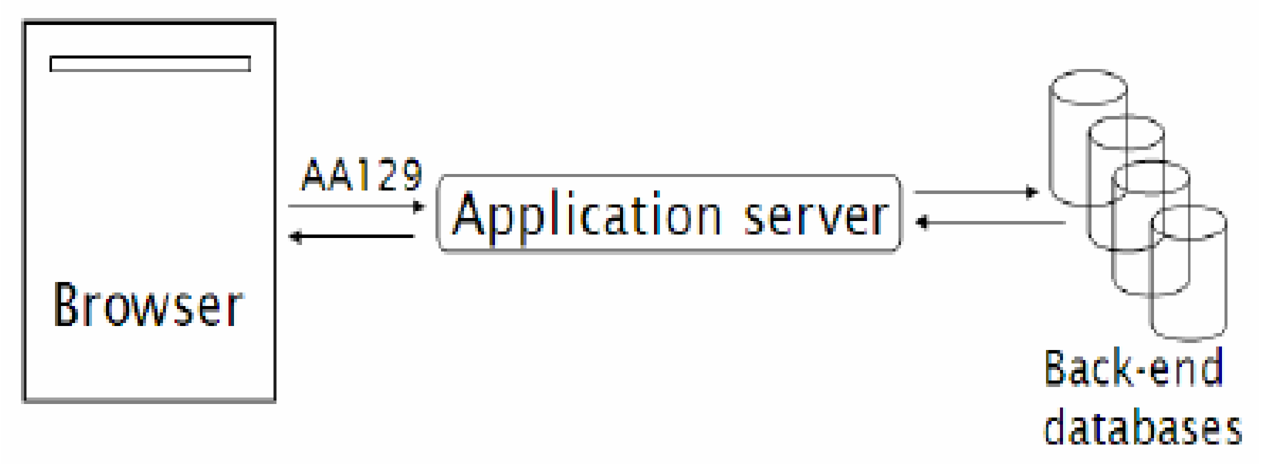
Web

- Web usage has shown tremendous growth to the point where it now claims a good fraction of people as participants, by relying on a simple, open client-server design:

(1) the server communicates with the client via a protocol (the *http* or hypertext transfer protocol) HTTP that is lightweight and simple, asynchronously carrying a variety of payloads (text, images and – over time – richer media such as audio and video files) encoded in a simple markup language called *HTML* (for hypertext markup language);

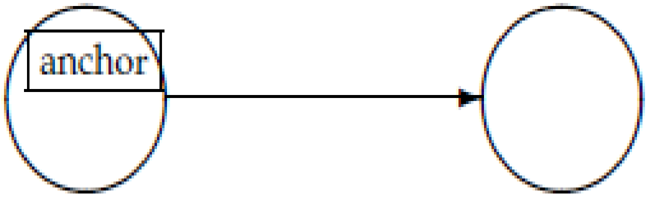
(2) The client – generally a *browser*, an application within a graphical user environment

- `http://www.stanford.edu/home/atoz/contact.html`.
- The string `www.stanford.edu` is known as the *domain*
- `/home/atoz/contact.html` is a path in this hierarchy with a file `contact.html` that contains the information to be returned by the web server at `www.stanford.edu` in response to this request.



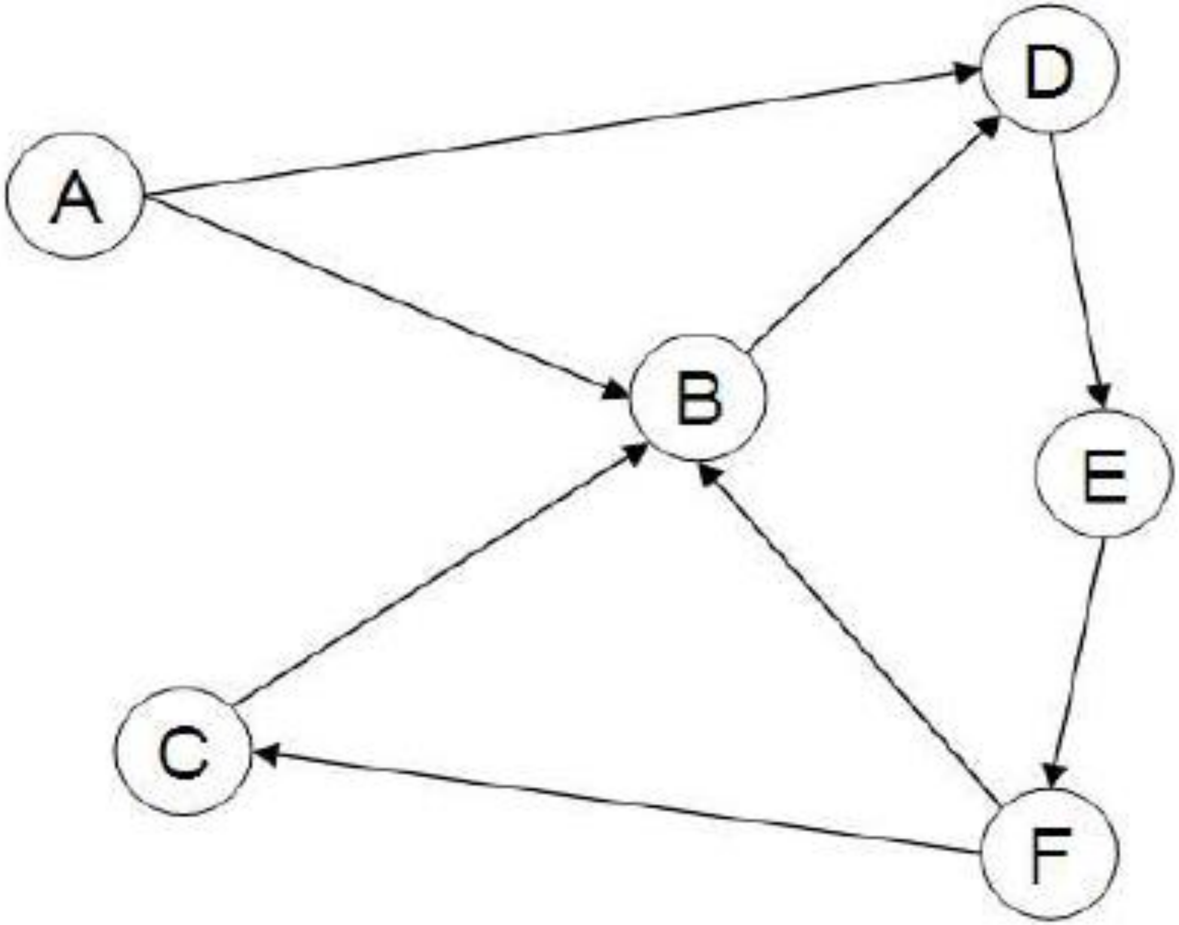
▶ **Figure 19.1** A dynamically generated web page. The browser sends a request for flight information on flight AA129 to the web application, that fetches the information from back-end databases then creates a dynamic web page that it returns to the browser.

Web Graph



▶ **Figure 19.2** Two nodes of the web graph joined by a link.

Figure 19.2 shows two nodes A and B from the web graph, each corresponding
to a web page, with a hyperlink from A to B. We refer to the set of all such nodes and directed edges as the web graph.
Figure 19.2 also shows that there is some text surrounding the origin of the hyperlink on page A. This text is generally encapsulated in the `href` attribute of the <a> (for anchor) tag that encodes the hyperlink in the HTML code of page A, and is referred to as *anchor text*.



► **Figure 19.3** A sample small web graph. In this example we have six pages labeled A-F. Page B has in-degree 3 and out-degree 1. This example graph is not strongly connected: there is no path from any of pages B-F to page A.

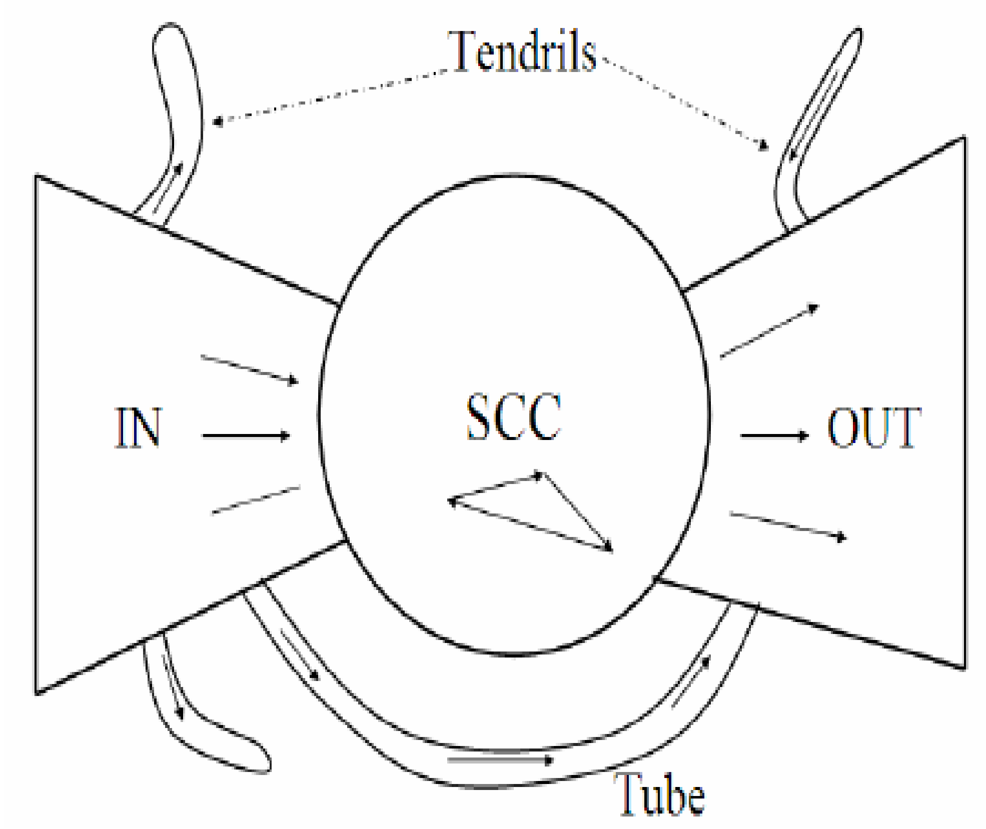
POWER LAW

- Distribution of the number of links into a web page does not follow the Poisson distribution. Rather, this distribution is widely reported to be a *power law*, in which the total number of web pages with in-degree i is proportional to $1/i^{\alpha}$;

  the value of α typically reported by studies is 2.1.

Bowtie structure of the Web

The directed graph connecting web pages has a *bowtie* shape



▶ Figure 19.4 The bowtie structure of the Web. Here we show one tube and three tendrils.

SCC- Strongly connected Core.

- There are three major categories of web pages that are sometimes referred to as IN, OUT and SCC. Most web pages fall into one of these three sets.
- The remaining pages form into *tubes* that are small sets of pages outside SCC that lead directly from IN to OUT, and *tendrils* that either lead nowhere from IN, or from nowhere to OUT

SPAM

- Early in the history of web search, web search engines were an important means for connecting advertisers to prospective buyers.
- A user searching for Apartments is not merely seeking news but instead likely to be seeking to purchase such a property.
- Sellers of such property and their agents, therefore, have a strong motive to create web pages that rank highly on this query.
- In a search engine whose scoring was based on term frequencies, a web page with numerous repetitions of apartments would rank highly.
- This is similar to the use of company names beginning with a long string of A's to be listed early in a yellow pages category.
- This led to the first generation of *spam*, which (in the context of web search) is the manipulation of web page content for the purpose of appearing high up in search results for selected keywords (Eg Godreg Apartments).
- To avoid irritating users with these repetitions, sophisticated *spammers* resorted to such tricks as rendering these repeated terms in the same color as the background.

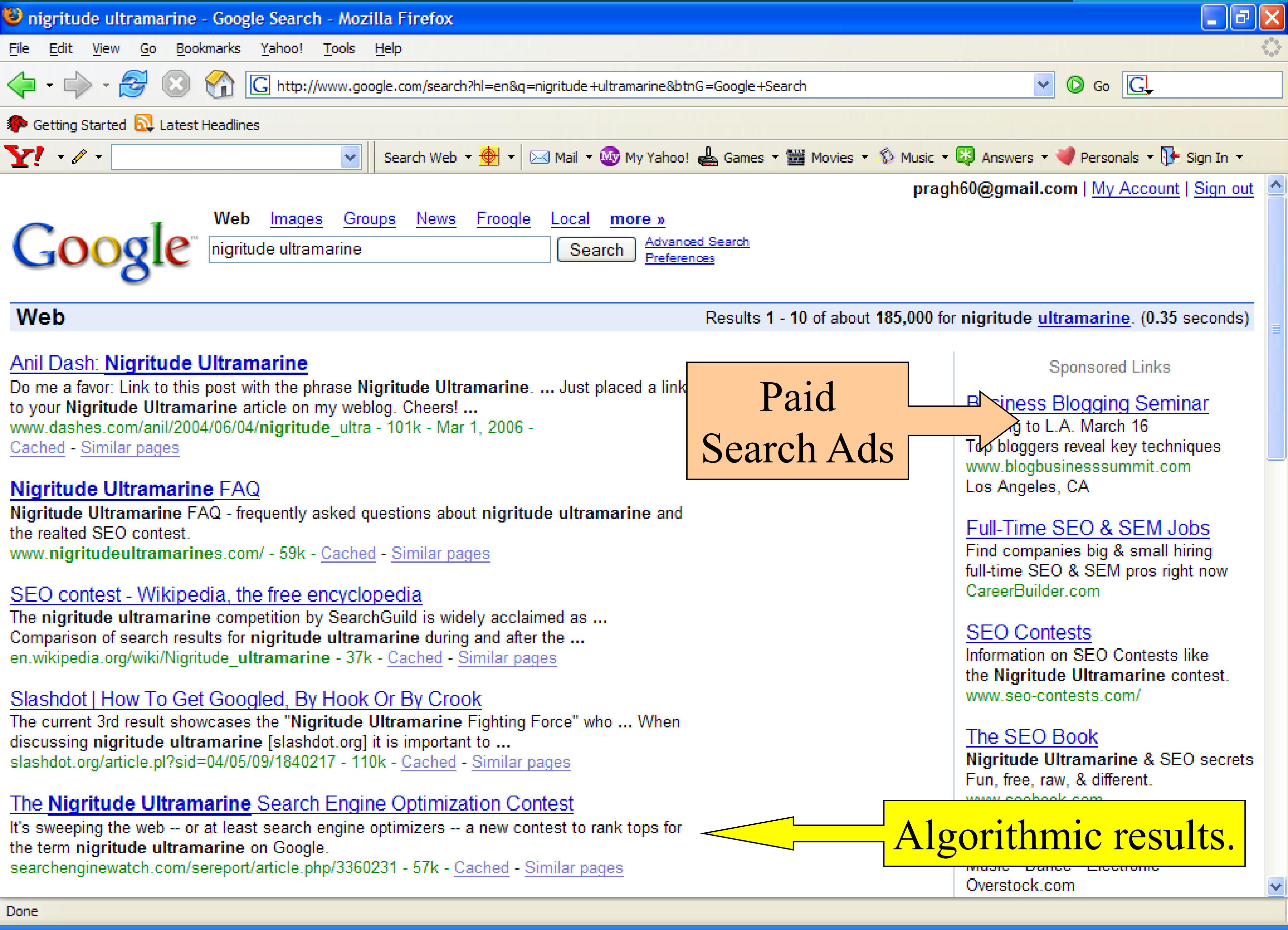
- PAID INCLUSION : In many search engines, it is possible to pay to have one's web page included in the search engine's index – a model known as *paid inclusion*.
- Search engines soon became sophisticated enough in their spam detection to screen out a large number of repetitions of particular keywords.
- Spammers responded with a richer set of spam techniques

Advertising as the economic model

- Early in the history of the Web, companies used graphical banner advertisements on web pages at popular websites (news and entertainment sites such as MSN, America Online, Yahoo! and CNN). The primary purpose of these advertisements was *branding*.
- Advertisements were priced on a *cost per mil* (*CPM*) basis: the CPM cost to the company of having its banner advertisement displayed 1000 times.
- Advertisements were later priced by the number of times it was *clicked on* by the user. This pricing model is known as the *cost per click* (*CPC*) model. This is a transaction-oriented advertising.
- CPC billing model – clicks could be metered and monitored by the website and billed to the advertiser

- **CPM Definition:**
- CPM = Cost per “mille,” or 1,000 impressions. A $2 CPM means you pay $2 for every 1,000 impressions your ad receives.
- **CPC Definition:**
- CPC = Cost per click. If your campaign generated 1,000 clicks at a $2 CPC, you would pay $2,000.
- GOTO Model
- The pioneer in this direction of CPC was a company named Goto
- GOTO was not in the traditional sense, a search engine but a just a search interface.
- For every query term *q* it accepted *bids* from companies who wanted their web page shown on the query *q*. In response to the query *q*, Goto would return the pages of all advertisers who bid for *q*, ordered by their bids.
- Furthermore, when the user clicked on one of the returned results, the corresponding advertiser would make a payment to Goto

- A user typing the query *q* into Goto’s search interface was actively expressing an interest and intent related to the query *q*.
- Goto only got compensated when a user actually expressed interest in an advertisement – as evinced by the user clicking the advertisement.
- It was a mechanism by which advertisers could get connected to consumers.
- This style of search engine came to be known as *sponsored search* or *search advertising*.
- Given these two kinds of search engines – the “pure” search engines such as Google and Yahoo, versus the sponsored search engines – the logical next step was to combine them into a single user experience.
- Current search engines follow this model: they provide pure search results (generally known as *algorithmic search* results) as the primary response to a user’s search, together with sponsored search results displayed separately and distinctively to the right of the algorithmic results.


nigritude ultramarine - Google Search - Mozilla Firefox
File Edit View Go Bookmarks Yahoo! Tools Help
http://www.google.com/search?hl=en&q=nigritude+ultramarine&btnG=Google+Search
Go
Getting Started
Latest Headlines
Search Web
Mail
My Yahoo!
Games
Movies
Music
Answers
Personals
Sign In
Done

pragh60@gmail.com | My Account | Sign out

Web Images Groups News Froogle Local **more »**

nigritude ultramarine | Search | Advanced Search | Preferences

Web Results **1** - **10** of about **185,000** for **nigritude ultramarine**. (**0.35** seconds)

Anil Dash: **Nigritude Ultramarine**
Do me a favor: Link to this post with the phrase **Nigritude Ultramarine**. ... Just placed a link to your **Nigritude Ultramarine** article on my weblog. Cheers! ...
www.dashes.com/anil/2004/06/04/**nigritude**_ultra - 101k - Mar 1, 2006 -
Cached - Similar pages

Nigritude Ultramarine FAQ
Nigritude Ultramarine FAQ - frequently asked questions about **nigritude ultramarine** and the realted SEO contest.
www.**nigritudeultramarine**s.com/ - 59k - Cached - Similar pages

SEO contest - Wikipedia, the free encyclopedia
The **nigritude ultramarine** competition by SearchGuild is widely acclaimed as ...
Comparison of search results for **nigritude ultramarine** during and after the ...
en.wikipedia.org/wiki/Nigritude_**ultramarine** - 37k - Cached - Similar pages

Slashdot | How To Get Googled, By Hook Or By Crook
The current 3rd result showcases the "**Nigritude Ultramarine** Fighting Force" who ... When discussing **nigritude ultramarine** [slashdot.org] it is important to ...
slashdot.org/article.pl?sid=04/05/09/1840217 - 110k - Cached - Similar pages

The **Nigritude Ultramarine** Search Engine Optimization Contest
It's sweeping the web -- or at least search engine optimizers -- a new contest to rank tops for the term **nigritude ultramarine** on Google.
searchenginewatch.com/sereport/article.php/3360231 - 57k - Cached - Similar pages

Sponsored Links

B iness Blogging Seminar
g to L.A. March 16
Top bloggers reveal key techniques
www.blogbusinesssummit.com
Los Angeles, CA

Full-Time SEO & SEM Jobs
Find companies big & small hiring full-time SEO & SEM pros right now
CareerBuilder.com

SEO Contests
Information on SEO Contests like the **Nigritude Ultramarine** contest.
www.seo-contests.com/

The SEO Book
Nigritude Ultramarine & SEO secrets
Fun, free, raw, & different.
www.seobook.com

Music - Dance - Electronic
Overstock.com

Paid Search Ads

Algorithmic results.

- For advertisers, understanding how search engines do this ranking and how to allocate marketing campaign budgets to different keywords and to different sponsored search engines has become a profession known as *search engine marketing* (SEM).
- CLICK SPAM
- It refers to clicks on sponsored search results that are not from bona fide search users.
- For instance, a devious advertiser may attempt to exhaust the advertising budget of a competitor by clicking repeatedly (through the use of a robotic click generator) on that competitor's sponsored search advertisements.
- Search engines face the challenge of identifying click spam, to avoid charging their advertiser clients for such clicks.

The search user experience

- It is crucial that we understand the users of web search.
- Traditional information retrieval users were typically professionals with at least some training in the art of phrasing queries over a well-authored collection whose style and structure they understood well.
- In contrast, web search users tend to not know (or care) about the heterogeneity of web content, the syntax of query languages and the art of phrasing queries.
- A range of studies has concluded that the average number of keywords in a web search is somewhere between 2 and 3.
- Syntax operators (Boolean connectives, wildcards, etc.) are seldom used.
- It is clear that the more user traffic a web search engine can attract, the more revenue it stands to earn from sponsored search.

How do search engines differentiate themselves and grow their traffic? Eg. Google

- Google identified two principles that helped it grow at the expense of its competitors:

(1) A focus on relevance, specifically precision rather than recall in the first few results

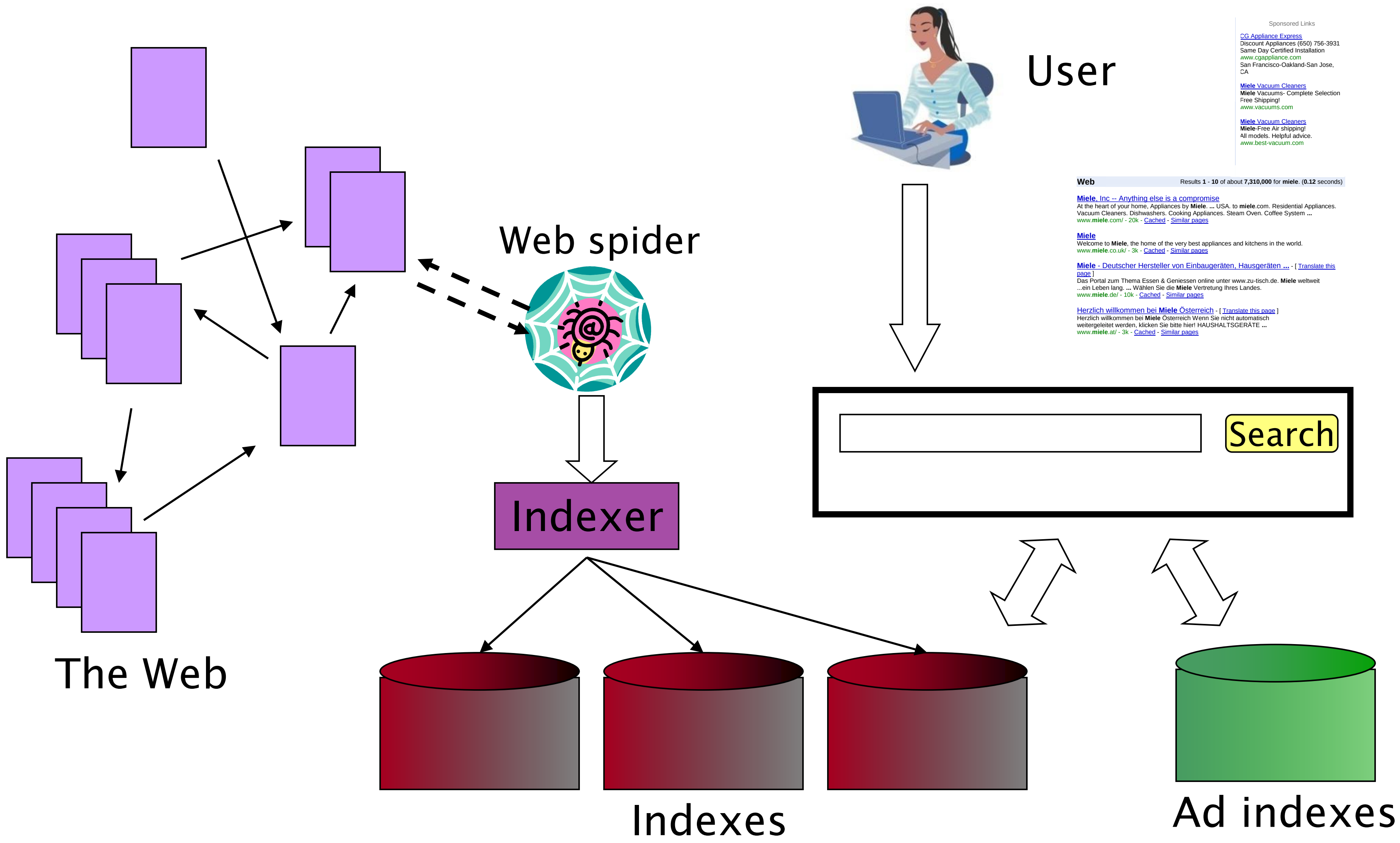
(2) a user experience that is lightweight, meaning that both the search query page and the search results page are uncluttered and almost entirely textual, with very few graphical elements.

- The effect of the first was simply to save users time in locating the information they sought.
- The effect of the second is to provide a user experience that is extremely responsive, or at any rate not bottlenecked by the time to load the search query or results page.

User Query Needs

- There appear to be three broad categories into which common web search queries can be grouped:
- **Informational** (~40% / 65%) Low hemoglobin
  - seek general information on a broad topic
  - want to learn about something
  - single web page may not contain all the information
- **Navigational** – (~25% / 15%) United Airlines Seattle weather
  - want to go to that page
  - seek the website or home page of a single entity that the user has in mind
- **Transactional** – (~35% / 20%) Car rental Brasil
- want to (web-mediated)
  - Access a service
  - Downloads
  - Shopping
- The category not only governs the algorithmic search results, but the suitability of the query for sponsored search results

Web search basics



User

Web spider

Indexer

Search

The Web

Indexes

Ad indexes

Sponsored Links

CG Appliance Express
Discount Appliances (650) 756-3931
Same Day Certified Installation
www.cgappliance.com
San Francisco-Oakland-San Jose, CA

Miele Vacuum Cleaners
Miele Vacuums- Complete Selection
Free Shipping!
www.vacuums.com

Miele Vacuum Cleaners
Miele-Free Air shipping!
All models. Helpful advice.
www.best-vacuum.com

Web Results **1** - **10** of about **7,310,000** for **miele**. (**0.12** seconds)

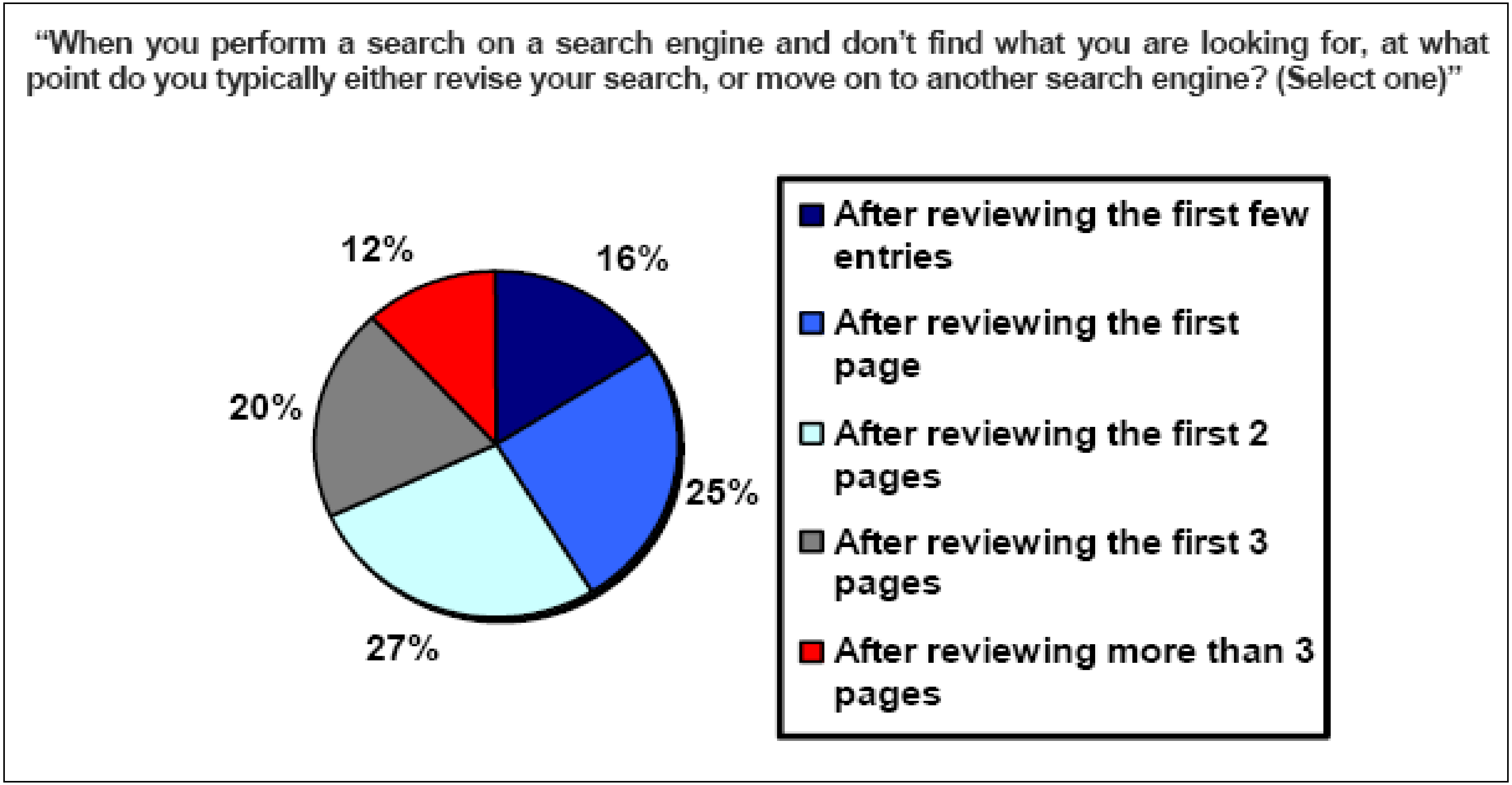
Miele, Inc -- Anything else is a compromise
At the heart of your home, Appliances by **Miele**. **...** USA. to **miele**.com. Residential Appliances. Vacuum Cleaners. Dishwashers. Cooking Appliances. Steam Oven. Coffee System **...**
www.**miele**.com/ - 20k - Cached - Similar pages

Miele
Welcome to **Miele**, the home of the very best appliances and kitchens in the world.
www.**miele**.co.uk/ - 3k - Cached - Similar pages

Miele - Deutscher Hersteller von Einbaugeräten, Hausgeräten **...** - [Translate this page]
Das Portal zum Thema Essen & Geniessen online unter www.zu-tisch.de. **Miele** weltweit ...ein Leben lang. **...** Wählen Sie die **Miele** Vertretung Ihres Landes.
www.**miele**.de/ - 10k - Cached - Similar pages

Herzlich willkommen bei **Miele** Österreich - [Translate this page]
Herzlich willkommen bei **Miele** Österreich Wenn Sie nicht automatisch weitergeleitet werden, klicken Sie bitte hier! HAUSHALTSGERÄTE **...**
www.**miele**.at/ - 3k - Cached - Similar pages

How far do people look for results?



"When you perform a search on a search engine and don't find what you are looking for, at what point do you typically either revise your search, or move on to another search engine? (Select one)"

(Source: iprospect.com WhitePaper_2006_SearchEngineUserBehavior.pdf)

Users' empirical evaluation of results

- Quality of pages varies widely
  - Relevance is not enough
  - Other desirable qualities (non IR!!)
    - Content: Trustworthy, diverse, non-duplicated, well maintained
    - Web readability: display correctly & fast
    - No annoyances: pop-ups, etc
- Precision vs. recall
  - On the web, recall seldom matters
- What matters
  - Precision at 1? Precision within top-K?
  - Comprehensiveness – must be able to deal with obscure queries
    - Recall matters when the number of matches is very small
- User perceptions may be unscientific, but are significant over a large aggregate

Users' empirical evaluation of engines

- Relevance and validity of results
- UI – Simple, no clutter, error tolerant
- Trust – Results are objective (unbiased, balanced observation)
- Coverage of topics for polysemic queries
- Pre/Post process tools provided
  - Mitigate user errors (auto spell check, search assist,…)
  - Explicit: Search within results, more like this, refine ...
  - Anticipative: related searches, instant searches
    - Impact on stemming, spell-check, etc
  - Web addresses typed in the search box

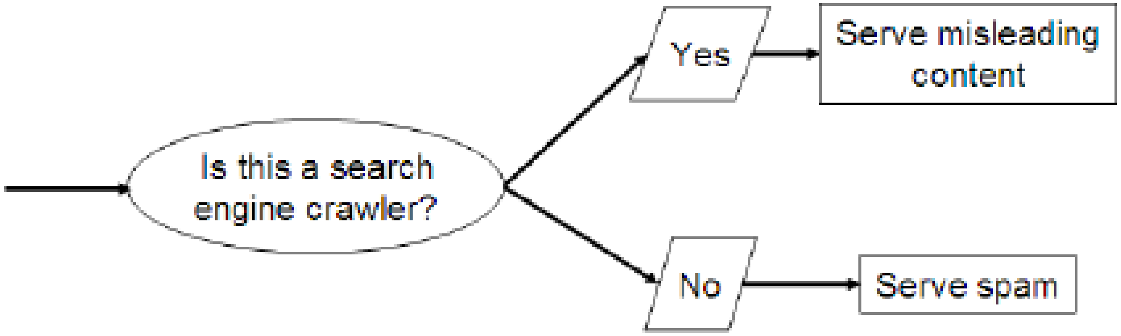
SPAM :
Simplest forms: Keyword Stuffing

- First generation engines relied heavily on *tf/idf*
  - The top-ranked pages for the query **maui resort** were the ones containing the most **maui**’s and **resort**’s
- SEOs -- dense repetitions of chosen terms
  - e.g., **maui resort maui resort maui resort**
  - Often, the repetitions would be in the same color as the background of the web page
    - Repeated terms got indexed by crawlers
    - But not visible to humans on browsers

Pure word density cannot be trusted as an IR signal

Cloaking

- Here, the spammer’s web server returns different pages depending on whether the http request comes from a web search engine’s crawler or from a human user’s browser.
- The web page will be indexed by the search engine under misleading keywords.
- When the user searches for these keywords and elects to view the page, he receives a web page that has altogether different content than that indexed by the search engine.
- Here user and Google bot get different results

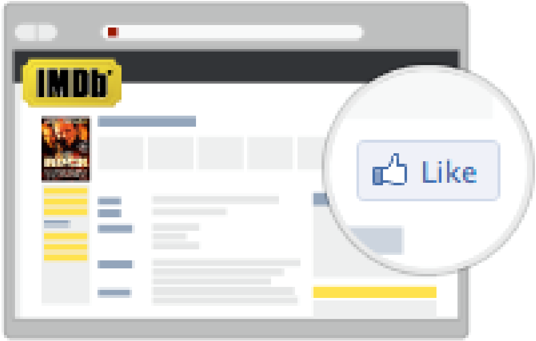


► **Figure 19.5** Cloaking as used by spammers.

More spam techniques

- **Doorway pages**
  - Pages optimized for a single keyword that re-direct to the real target page
- **Link spamming**
  - Mutual admiration societies, hidden links, awards
  - *Domain flooding:* numerous domains that point or re-direct to a target page
- **Robots**
  - Fake query stream – rank checking programs
    - "Curve-fit" ranking programs of search engines
  - Millions of submissions via Add-Url

The war against spam



- Quality signals - Prefer authoritative pages based on:
  - Votes from authors (linkage signals)
  - Votes from users (usage signals)
- Policing of URL submissions
  - Anti robot test
- Limits on meta-keywords
- Robust link analysis
  - Ignore statistically implausible linkage (or text)
  - Use link analysis to detect spammers (guilt by association)

- Spam recognition by machine learning
  - Training set based on known spam
- Family friendly filters
  - Linguistic analysis, general classification techniques, etc.
  - For images: flesh tone detectors, source text analysis, etc.
- Editorial intervention
  - Blacklists
  - Top queries audited
  - Complaints addressed
  - Suspect pattern detection

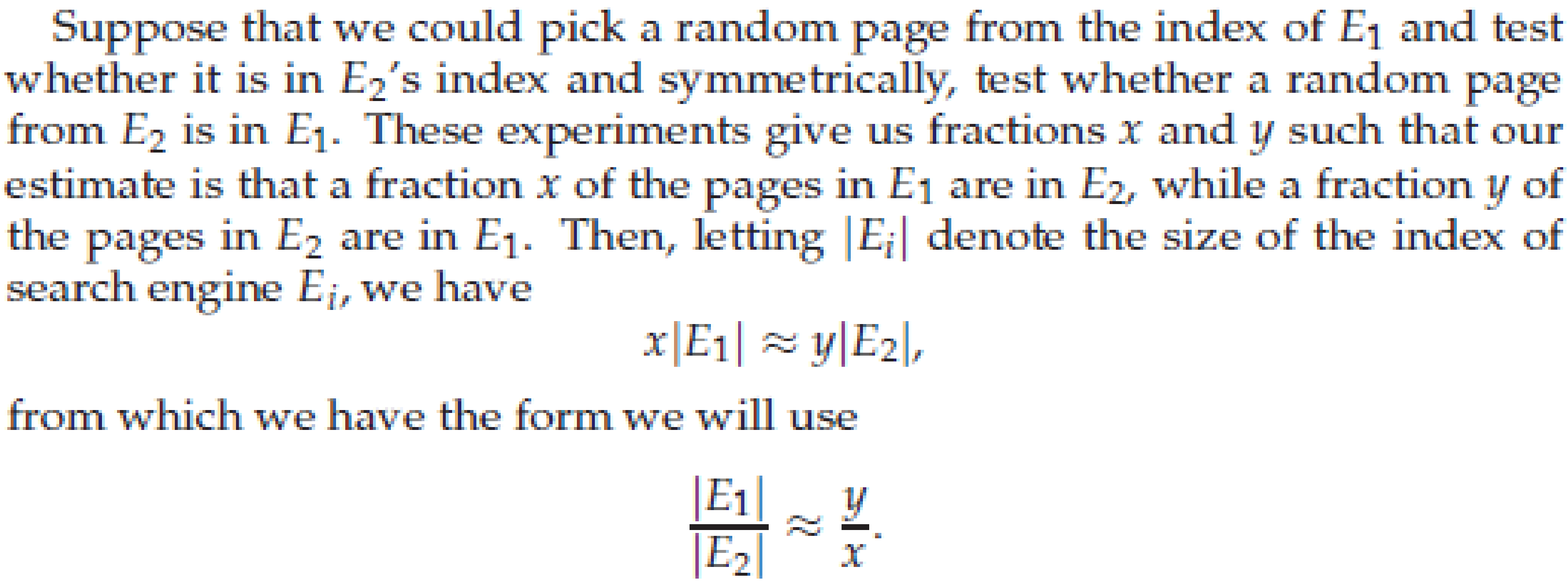
SEO (Search Engine Optimization)

- Given that spamming is inherently an economically motivated activity, there has sprung around it an industry of *Search Engine Optimizers* or SEOs to provide consultancy services for clients who seek to have their web pages rank highly on selected keywords.
- SEO is the practice of increasing the quantity and quality of traffic to your website through organic search engine results (unpaid)
- Web search engines frown on this business of attempting to decipher and adapt to their proprietary ranking techniques and indeed announce policies on forms of SEO behavior they do not tolerate.
- Inevitably, the parrying between such SEOs (who gradually infer features of each web search engine's ranking methods) and the web search engines (who adapt in response) is an unending struggle;
- Indeed, the research sub-area of *adversarial information retrieval* has sprung up around this battle.
- To combat spammers who manipulate the text of their web pages is the exploitation of the link structure of the Web – a technique known as *link analysis*.
- Spammers now invest considerable effort in subverting it – this is known as *link spam*.

Index size and estimation

- Given two search engines, what are the relative sizes of their indexes?
- Issues
  - In response to queries a search engine can return web pages whose contents it has not (fully or even partially) indexed.
  - In some cases, a search engine is aware of a page *p* that is *linked to* by pages it has indexed, but has not indexed *p* itself.
  - Search engines generally organize their indexes in various tiers and partitions , not all of which are examined on every search
- Thus, search engine indexes include multiple classes of indexed pages, so that there is no single measure of index size.
- These issues not withstanding, a number of techniques have been devised for crude estimates of the ratio of the index sizes of two search engines, *E*1 and *E*2.

- The basic hypothesis underlying these techniques is that each search engine indexes a fraction of the Web chosen independently and uniformly at random.
- But assumptions are questionable
  - first, that there is a finite size for the Web from which each search engine chooses a subset, and
  - second, that each engine chooses an independent, uniformly chosen subset.
- A classical estimation technique known as the *capture recapture method* is used.



Suppose that we could pick a random page from the index of E_1 and test whether it is in E_2's index and symmetrically, test whether a random page from E_2 is in E_1. These experiments give us fractions x and y such that our estimate is that a fraction x of the pages in E_1 are in E_2, while a fraction y of the pages in E_2 are in E_1. Then, letting $|E_i|$ denote the size of the index of search engine E_i, we have

$$x|E_1| \approx y|E_2|,$$

from which we have the form we will use

$$\frac{|E_1|}{|E_2|} \approx \frac{y}{x}.$$

Sampling Phase

- *Random searches:*
  - Begin with a search log of web searches; send a random search from this log to *E*1 and a random page from the results.
  - trap all search queries going out of a work group (say scientists in a research center)
  - This approach has a number of issues, including the bias from the types of searches made by the work group.
  - Also, a random document from the results of such a random search to *E*1 is not the same as a random document from *E*1.
- *Random IP addresses:*
  - Generate random IP addresses
  - Find a web server at the given address If there's one
  - Collect all pages from server
  - From this, choose a page at random
  - Issue - many hosts may share one IP - or not accept http requests from the host where the experiment is conducted.

- *Random walks:*
- View the Web as a directed graph (if strongly connected directed graph)
- Build a random walk on this graph
- This walk would converge to a steady state distribution
- This method, too has a number of biases.
  - First, the Web is not strongly connected so that, even with various corrective rules, it is difficult to argue that we can reach a steady state distribution starting from any page.
  - Second, the time it takes for the random walk to settle into this steady state is unknown and could exceed the length of the experiment.

Clearly each of these approaches is far from perfect.

- Random queries:
- The idea is to pick a page (almost) uniformly at random from a search engine's index by posing a random query to it.
- It should be clear that picking a set of random terms from (say) Webster's dictionary is not a good way of implementing this idea.
  - There are a great many terms in web documents that do not occur in a standard dictionary such as Webster's
- To address the problem of vocabulary terms not in a standard dictionary, we begin by creating a sample web dictionary. This could be done by crawling a limited portion of the Web, or by crawling a manually-assembled representative subset of the Web such as Google or Yahoo!

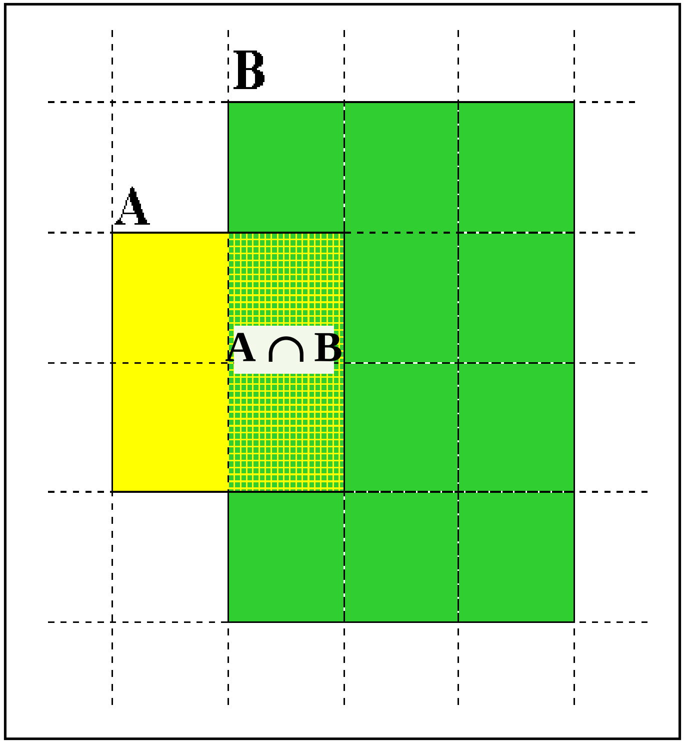
Testing Phase

- Operationally, the process is as follows:
- We use a random conjunctive query on *E*1 and pick from the top 100 returned results a page *p* at random.
- We then test *p* for presence in *E*2 by choosing 6-8 low-frequency terms in *p* and using them in a conjunctive query for *E*2.
- We can improve the estimate by repeating the experiment a large number of times.
- Both the sampling process and the testing process have a number of issues.

1. Our sample is biased towards longer documents.
2. Picking from the top 100 results of *E*1 induces a bias from the ranking algorithm of *E*1.
3. During the Testing phase, a number of additional biases are introduced: for instance, *E*2 may not handle 8-word conjunctive queries properly.
4. Either *E*1 or *E*2 may refuse to respond to the test queries, treating them as robotic spam rather than as bona fide queries.
5. There could be operational problems like connection time-outs.

There is no perfect solution

Relative Size from Overlap Given two engines A and B



Sample URLs randomly from A

Check if contained in B and vice versa

$$A \cap B = (1/2) * \text{Size } A$$

$$A \cap B = (1/6) * \text{Size } B$$

$$(1/2) * \text{Size } A = (1/6) * \text{Size } B$$

$$\therefore \text{Size } A / \text{Size } B = (1/6)/(1/2) = 1/3$$

Each test involves: (i) Sampling (ii) Checking

Duplicate documents

- The web is full of duplicated content
- Strict duplicate detection = exact match
  - Not as common
- But many, many cases of near duplicates
  - E.g., Last modified date the only difference between two copies of a page

Eg, Near-duplicate videos

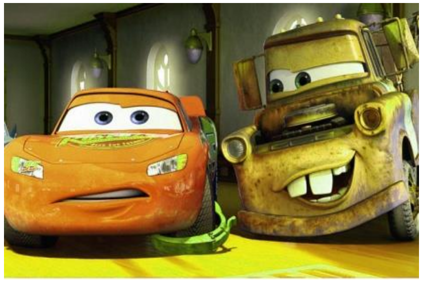
< Original Video>

Contrast

Brightne

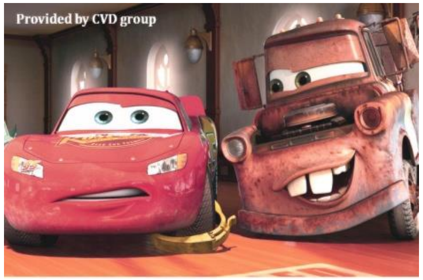
Crop

Color Enhancement



Color

TV Size



Multi-editing

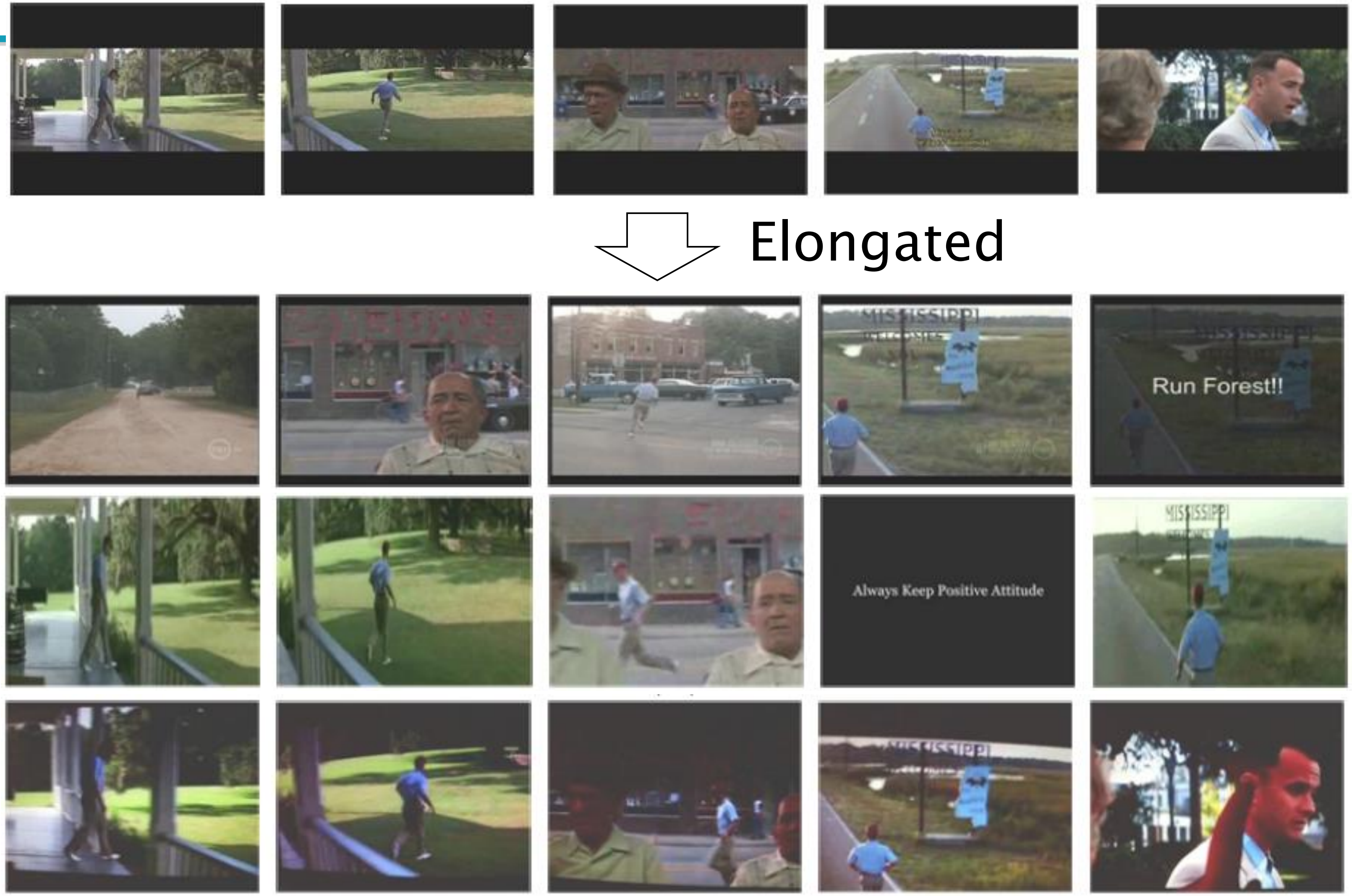
Low resolution

Noise/Blur

Provided by CVD group

Small Logo

Eg, Near-duplicate videos



Original video

Elongated

Copied video

Duplicate/Near-Duplicate Detection

- *Duplication*: Exact match can be detected with fingerprints
- *Near-Duplication*: Approximate match
    - Compute syntactic similarity with an edit-distance measure
    - Use similarity threshold to detect near-duplicates
        - E.g., Similarity > 80% => Documents are “near duplicates”

Computing Similarity - Shingling

- Features:
  - Segments of a document (natural or artificial breakpoints)
  - Shingles (Word N-Grams) - a solution to the problem of detecting near-duplicate web pages. *Shingling*- Given a positive integer *k* and a sequence of terms in a document *d*, define the *k*-shingles of *d* to be the set of all consecutive sequences of *k* terms in *d*.
  - ***a rose is a rose is a rose***

    a_rose_is_a

    rose_is_a_rose

    is_a_rose_is

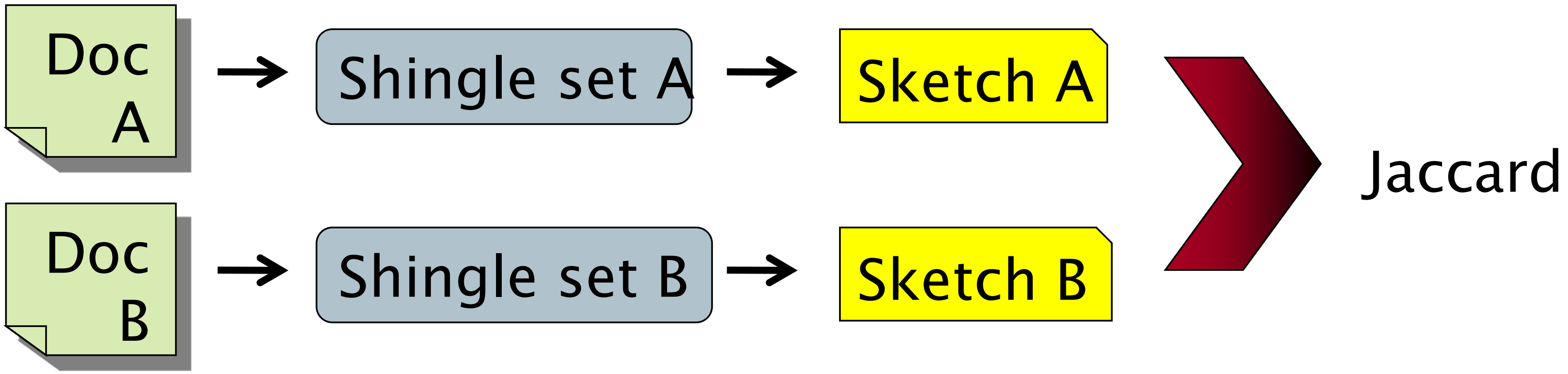
    a_rose_is_a

    my rose is a rose is yours
- Similarity Measure between two docs (= sets of shingles)
  - Set intersection
  - Specifically (Size_of_Intersection / Size_of_Union)

$$\mathrm{Jaccard}(C_i, C_j) = \frac{|C_i \cap C_j|}{|C_i \cup C_j|}$$

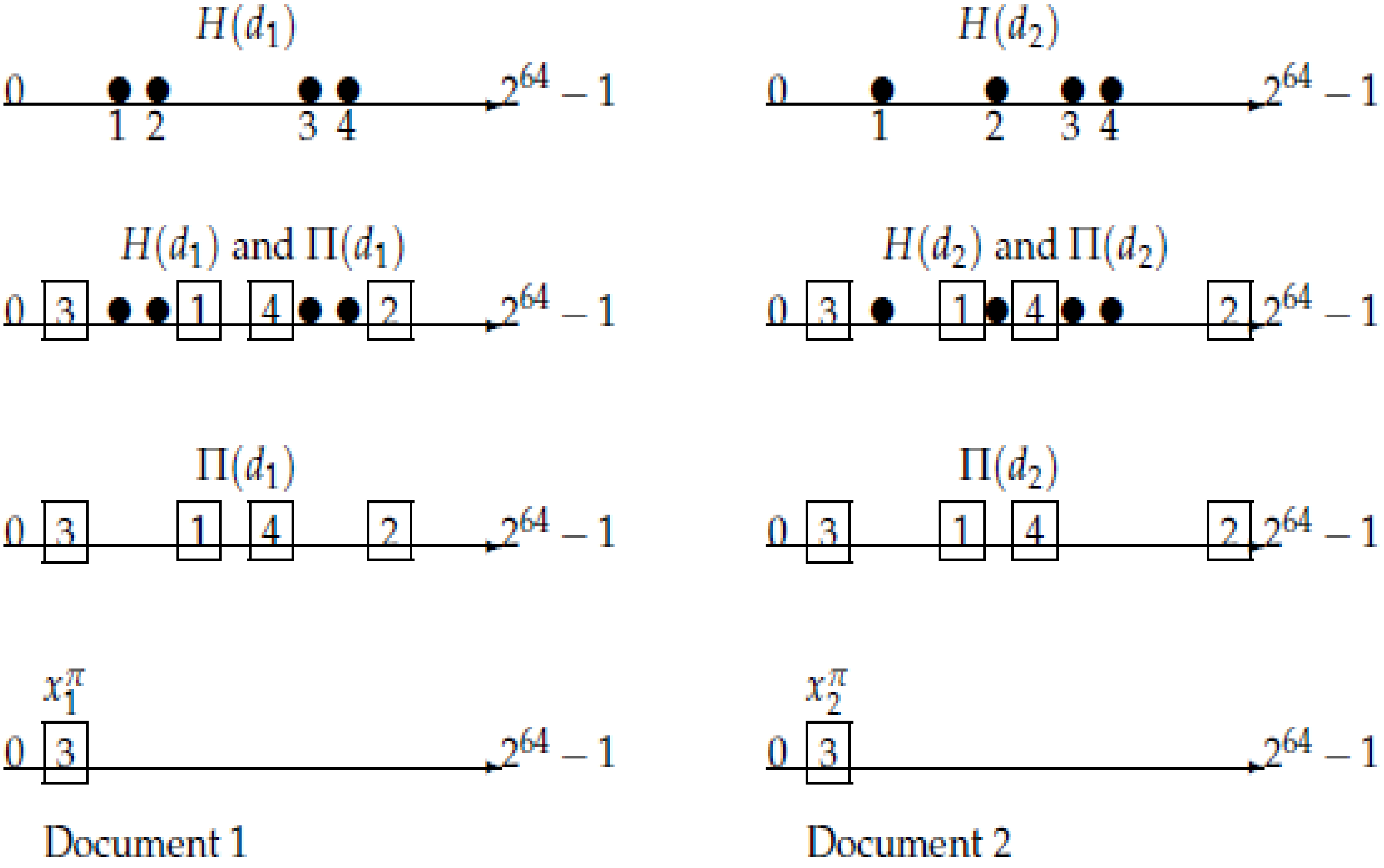
Shingles + Set Intersection

- Issue: Computing exact set intersection of shingles between all pairs of documents is expensive
  - Solution → Approximate using a cleverly chosen subset of shingles from each (called a *sketch*)
- Estimate (size_of_intersection / size_of_union) based on a short sketch

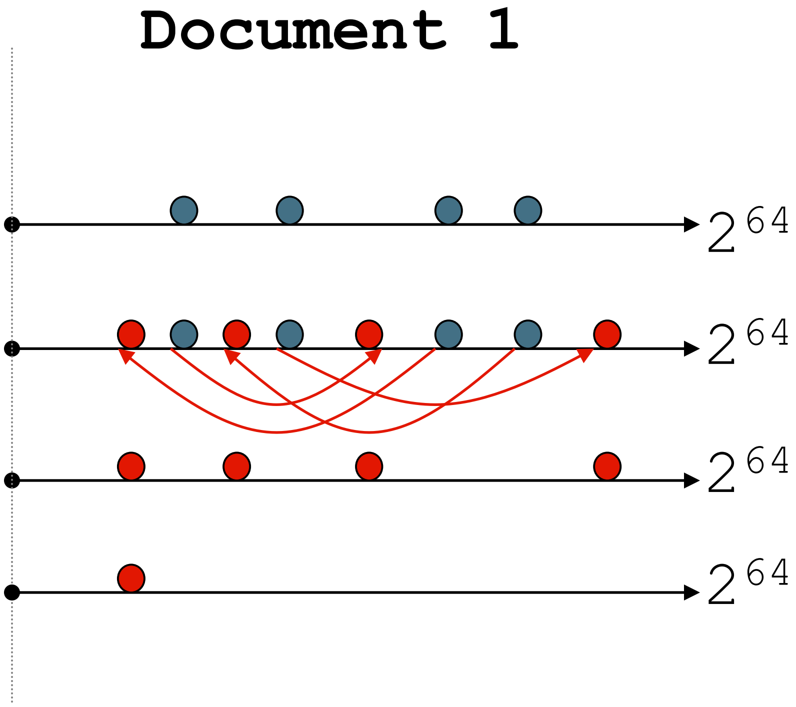


Sketch of a document

- Create a “sketch vector” (of size ~200) for each document
  - Documents that share ≥ *t* (say 80%) corresponding vector elements are near duplicates
  - For doc *D*, sketch$_D$[*i*] is as follows:
    - Let f map all shingles in the universe to 0..2^m (e.g., f = fingerprinting)
    - Let π_i be a *random permutation* on 0..2^m
    - Pick MIN $\{\pi_i(f(s))\}$ over all shingles *s* in *D*


$H(d_1)$
0
1 2 3 4
$2^{64}-1$
$H(d_1)$ and $\Pi(d_1)$
0
3 1 4 2
$2^{64}-1$
$\Pi(d_1)$
0
3 1 4 2
$2^{64}-1$
x_1^π
0
3
$2^{64}-1$
Document 1
$H(d_2)$
0
1 2 3 4
$2^{64}-1$
$H(d_2)$ and $\Pi(d_2)$
0
3 1 4 2
$2^{64}-1$
$\Pi(d_2)$
0
3 1 4 2
$2^{64}-1$
x_2^π
0
3
$2^{64}-1$
Document 2

Computing Sketch[i] for Doc1

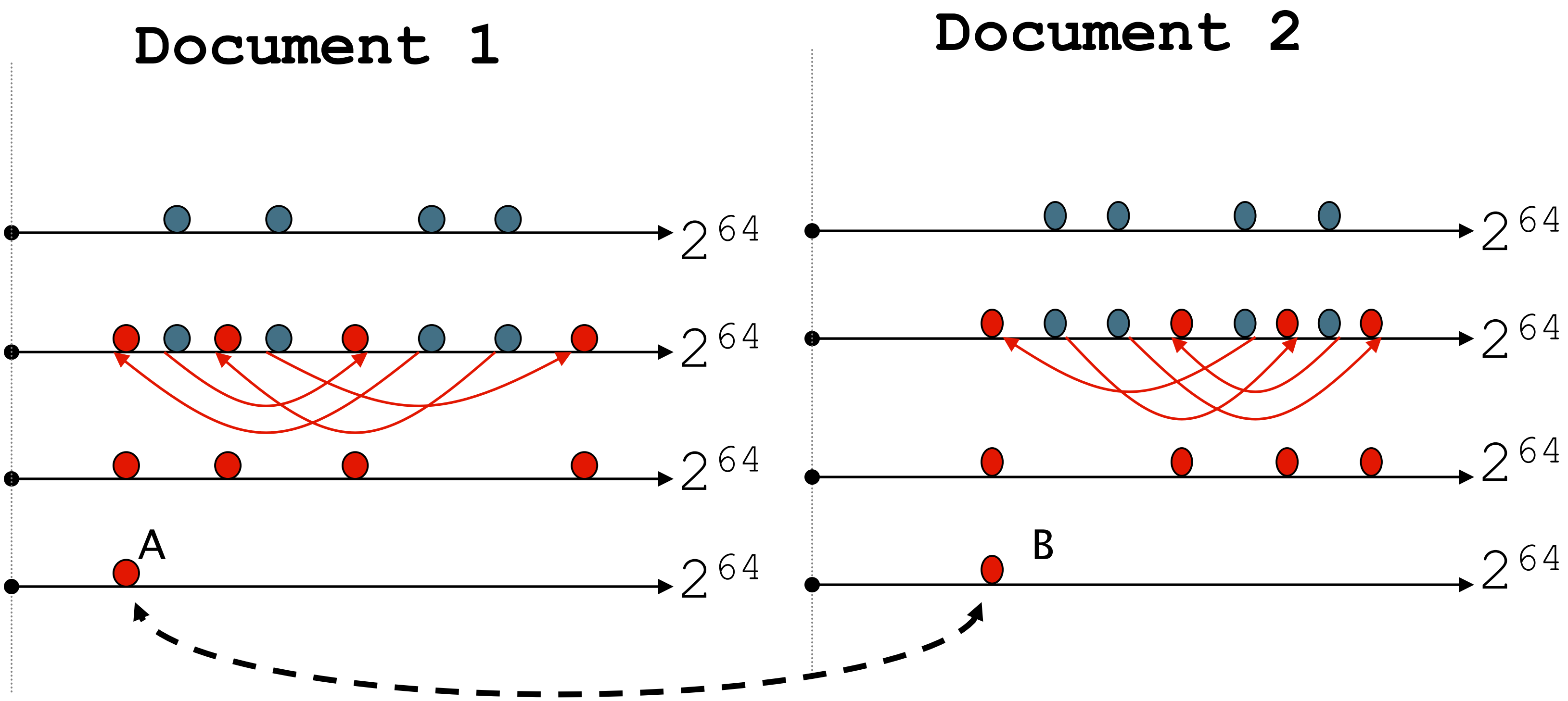


Start with 64-bit *f*(shingles)

Permute on the number line with π_i

Pick the min value

Test if Doc1.Sketch[i] = Doc2.Sketch[i]



Are these equal?

Test for 200 random permutations: $\pi_1, \pi_2,\ldots \pi_{200}$

Sketch Eg (|shingle|=4, |U|=5, |sketch|=2)

- ***a rose is a rose is a rose***

  a_rose_is_a

  rose_is_a_rose

  is_a_rose_is

  a_rose_is_a

- ***my rose is a rose is yours***

  my_rose_is_a

  rose_is_a_rose

  is_a_rose_is

  a_rose_is_yours

Min-Hash Technique:

Theory behind the “Sketch vector” idea

Set Similarity of sets C_i , C_j

$$\mathrm{Jaccard}(C_i, C_j) = \frac{|C_i \cap C_j|}{|C_i \cup C_j|}$$

- View sets as columns of a matrix A; one row for each element in the universe. $a_{ij} = 1$ indicates presence of item i in set j
- Example

C_1	**C_2**
0	1
1	0
1	1
0	0
1	1
0	1

Jaccard$(\mathbf{C_1},\mathbf{C_2}) = \mathbf{2/5 = 0.4}$

Jaccard coefficcient

- Jaccard coefficcient computes the similarity between sets.

$$Jaccard(C_i, C_j) = \frac{|C_i \cap C_j|}{|C_i \cup C_j|}$$

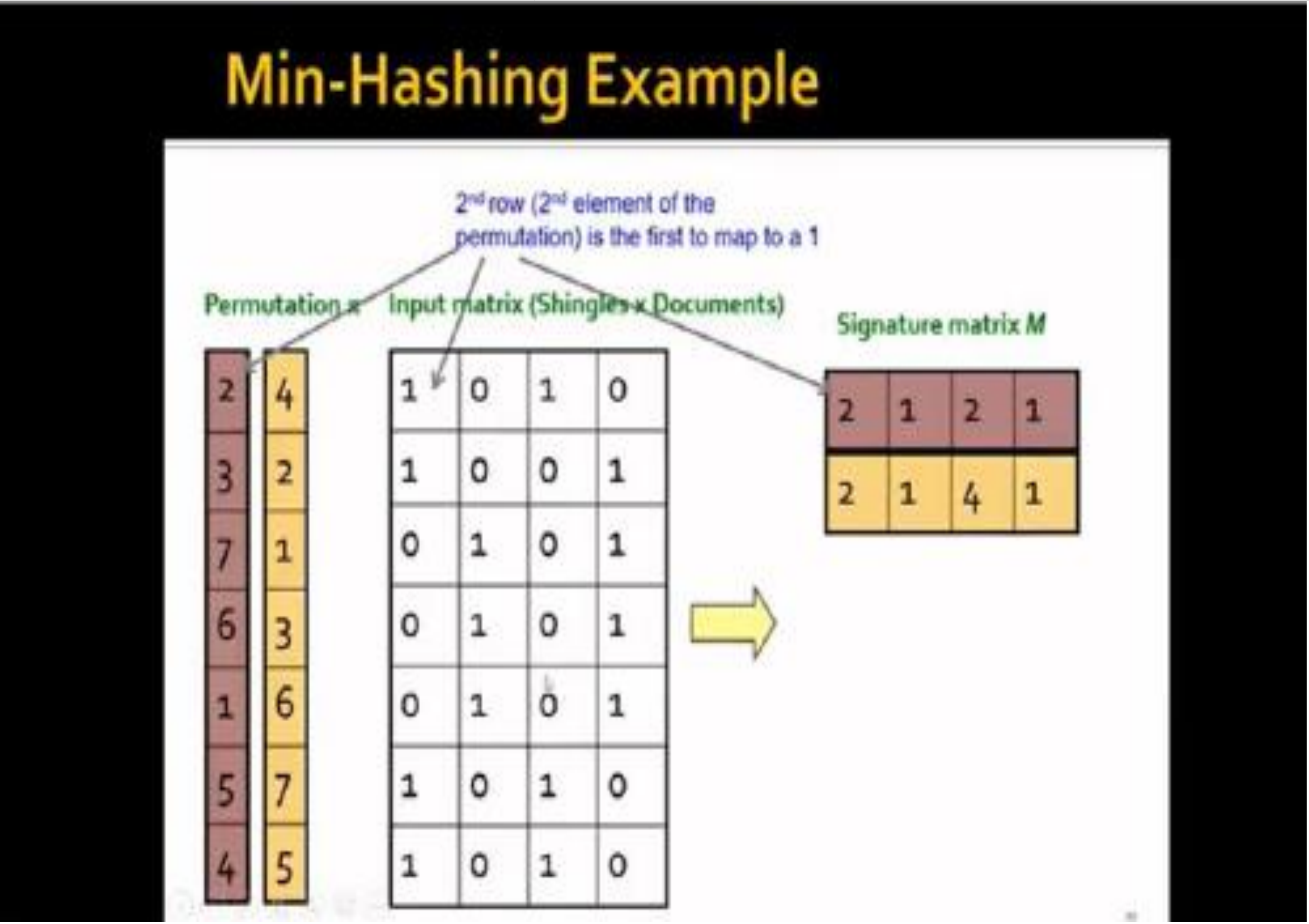
- View sets as columns of a matrix A:
  - one row for each shingle in the universe
  - one column for each document
  - a_{ij} = 1 indicates presence of shingle *i* in document *j*

- Example: $Jaccard(C_1, C_2) = \frac{3}{6}$

Documents

Shingles

1	1	1	0
1	1	0	1
0	1	0	1
0	0	0	1
1	0	0	1
1	1	1	0
1	0	1	0



Min-Hashing Example

2nd row (2nd element of the permutation) is the first to map to a 1

Permutation π

2	4
3	2
7	1
6	3
1	6
5	7
4	5

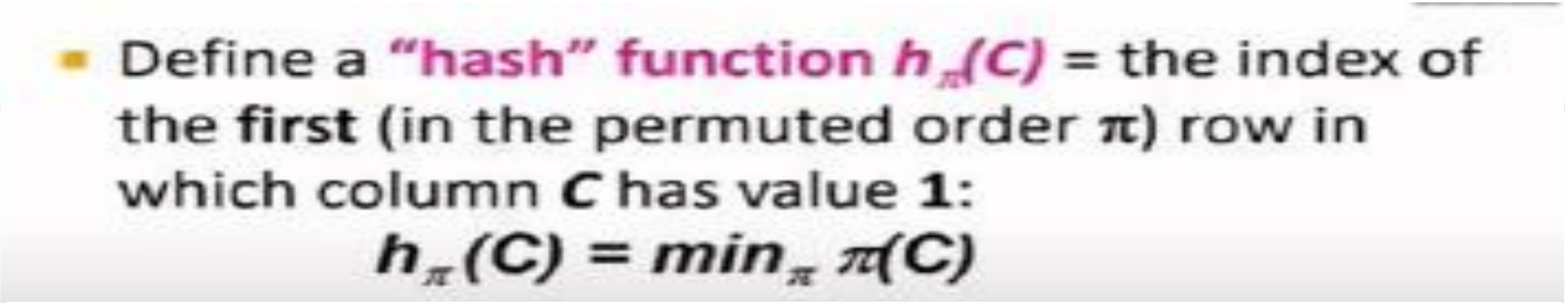
Input matrix (Shingles x Documents)

1	0	1	0
1	0	0	1
0	1	0	1
0	1	0	1
0	1	0	1
1	0	1	0
1	0	1	0

Signature matrix *M*

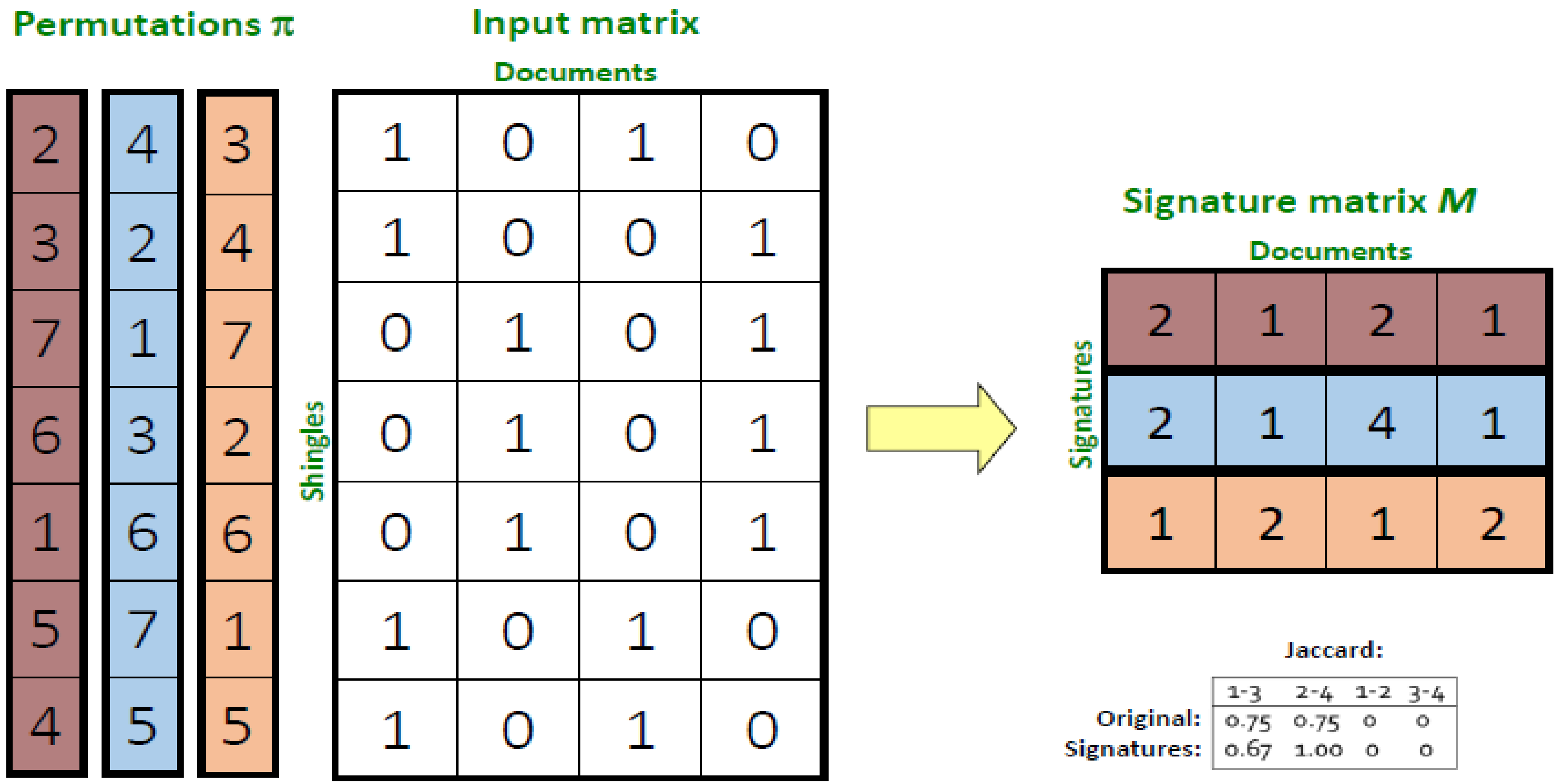
2	1	2	1
2	1	4	1

Min-Hashing example



- Define a **"hash" function $h_\pi(C)$** = the index of the **first** (in the permuted order π) row in which column C has value **1**:

$$h_\pi(C) = min_\pi \pi(C)$$



Permutations π

2	4	3
3	2	4
7	1	7
6	3	2
1	6	6
5	7	1
4	5	5

Input matrix

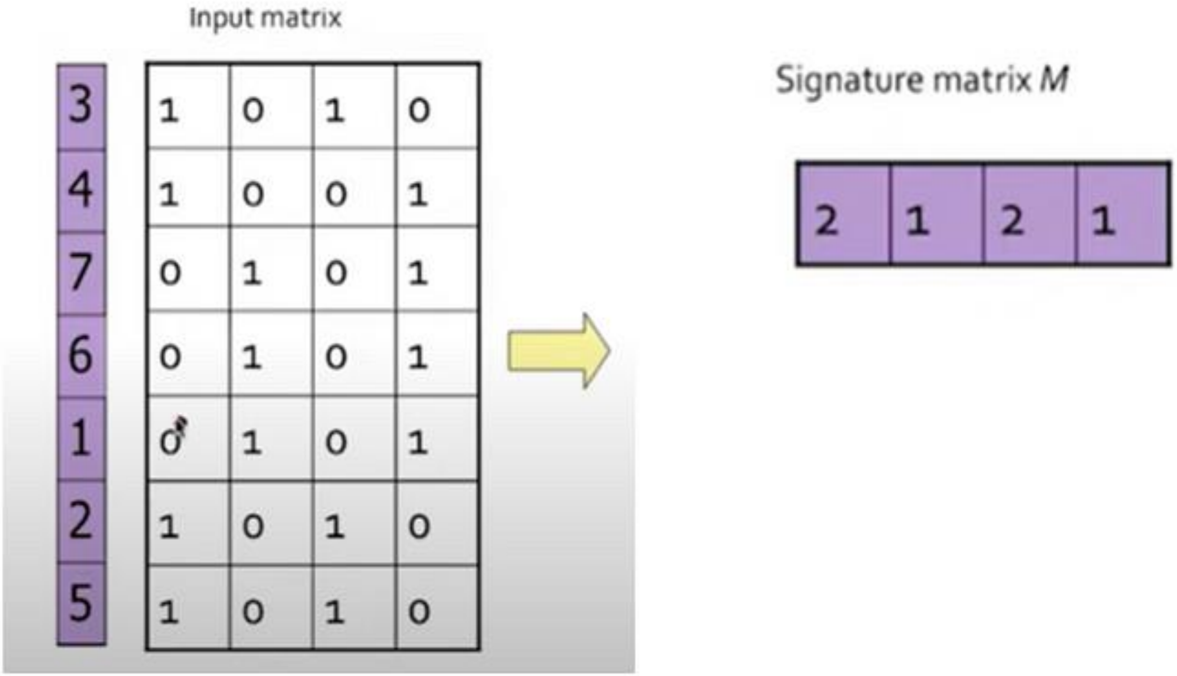
$$\begin{array}{c|cccc} & \multicolumn{4}{c}{\text{Documents}} \\ \hline \text{Shingles} & 1 & 0 & 1 & 0 \\ & 1 & 0 & 0 & 1 \\ & 0 & 1 & 0 & 1 \\ & 0 & 1 & 0 & 1 \\ & 0 & 1 & 0 & 1 \\ & 1 & 0 & 1 & 0 \\ & 1 & 0 & 1 & 0 \end{array}$$

Signature matrix *M*

$$\begin{array}{c|cccc} & \multicolumn{4}{c}{\text{Documents}} \\ \hline \text{Signatures} & 2 & 1 & 2 & 1 \\ & 2 & 1 & 4 & 1 \\ & 1 & 2 & 1 & 2 \end{array}$$

Jaccard:

	1-3	2-4	1-2	3-4
Original:	0.75	0.75	0	0
Signatures:	0.67	1.00	0	0



Input matrix

3	1	0	1	0
4	1	0	0	1
7	0	1	0	1
6	0	1	0	1
1	0	1	0	1
2	1	0	1	0
5	1	0	1	0

Signature matrix *M*

2	1	2	1

- So, for for C1 , row with π = 1 has 0, row with π = 2 has 1. Thus, the minimum indexed row, with a 1 is 2.
- Thus, for ***C1***, the corresponding entry in signature matrix is ***2***.
- Similarly, for ***C2 , C3*** and ***C4*** , we have ***1, 2*** and ***1*** respectively.