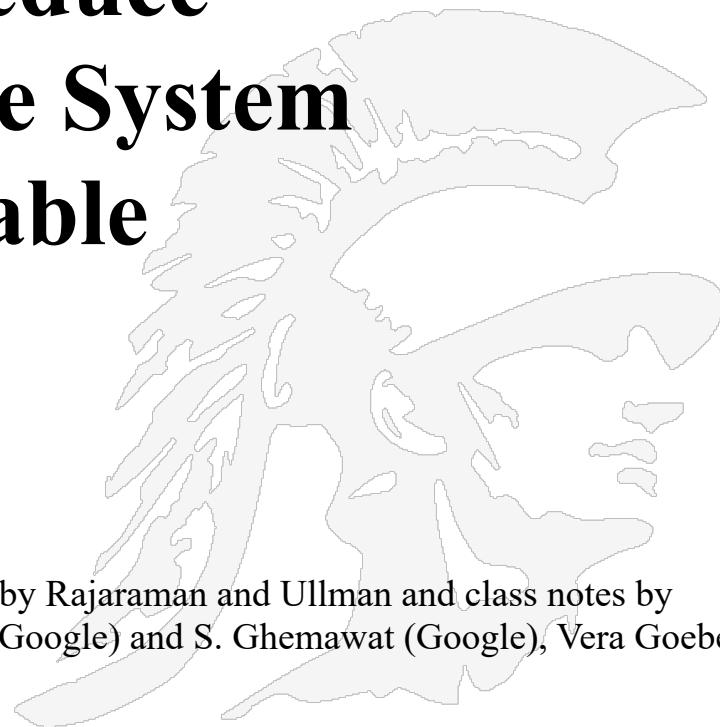


Google Building Blocks:

MapReduce

Google File System

BigTable



These slides borrow material from Mining Massive Datasets by Rajaraman and Ullman and class notes by Rajaraman (Stanford) and Weld (U. Washington), D. Weld (Google) and S. Ghemawat (Google), Vera Goebel (Univ. of Oslo)

Google Specialized Software Systems

- Google has built several major software systems for their internal processing

1. MapReduce - an easy way to write and run large-scale jobs on clusters of machines

- generate production index data more quickly
- perform ad-hoc experiments rapidly
- Dean & Ghemawat. *MapReduce: Simplified Data Processing on Large Clusters*, OSDI, 2004

2. GFS (Google File System) a large-scale distributed file system

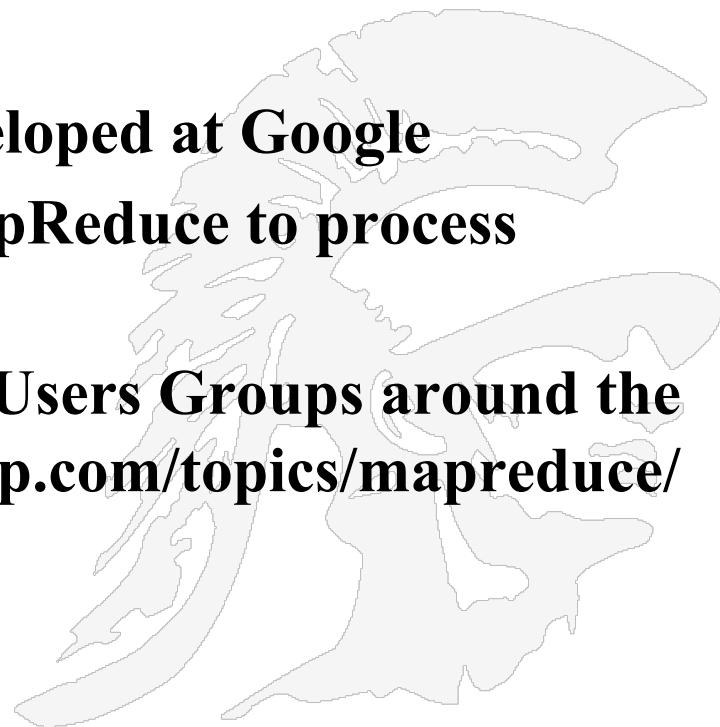
- Ghemawat, Gobioff, & Leung. *Google File System*, SOSP 2003

3. BigTable - a semi-structured storage system

- online, efficient access to per-document information at any time
- multiple processes can update per-doc info asynchronously
- critical for updating documents in minutes instead of hours
- Chang, Dean, Ghemawat, Hsieh, Wallach, Burrows, Chandra, Fikes, & Gruber. *Bigtable: A Distributed Storage System for Structured Data*, OSDI 2006

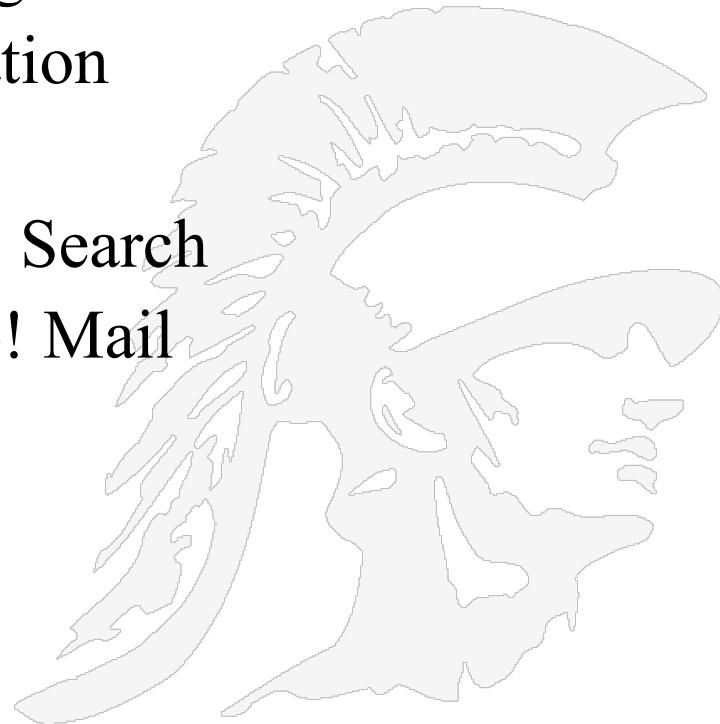
Introduction to MapReduce

- MapReduce is a methodology for exploiting parallelism in computing clouds (racks of interconnected processors)
- It has become a common way to analyze very large amounts of data
- MapReduce was initially developed at Google
- In 2004 Google was using MapReduce to process 100TB/day of data
- Today there are MapReduce Users Groups around the world, see <https://www.meetup.com/topics/mapreduce/>



How is MapReduce Used by Search Engines?

- **At Google:**
 - Building Google's Search Index
 - Article clustering for Google News
 - Statistical machine translation
- **At Yahoo!:**
 - Index building for Yahoo! Search
 - Spam detection for Yahoo! Mail
- **At Facebook:**
 - Data mining
 - Ad optimization
 - Spam detection



Motivation Beyond Search Engines

- Modern Internet applications have created a need to manage immense amounts of data quickly.
- In many of these applications, the data is extremely regular, and there is ample opportunity to exploit parallelism.

• Some examples

1. Dish network collecting every click of the remote

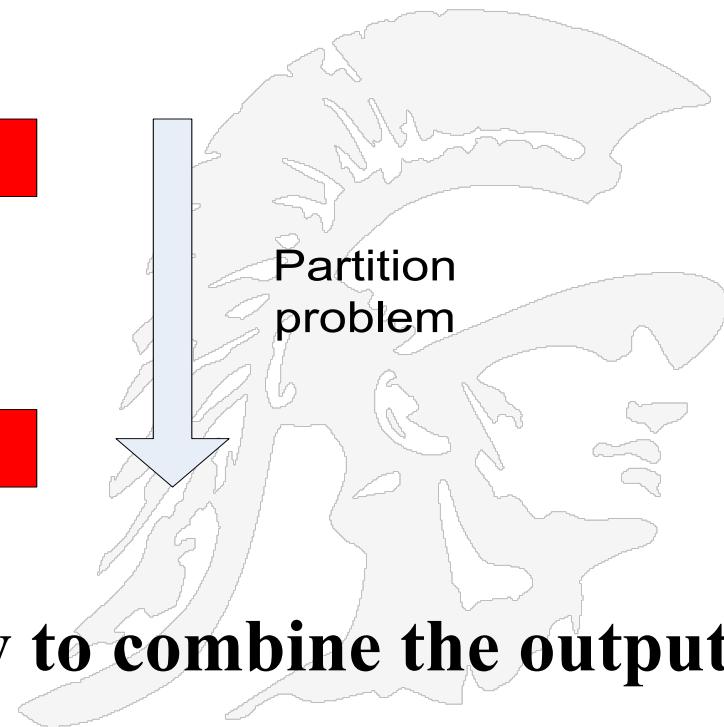
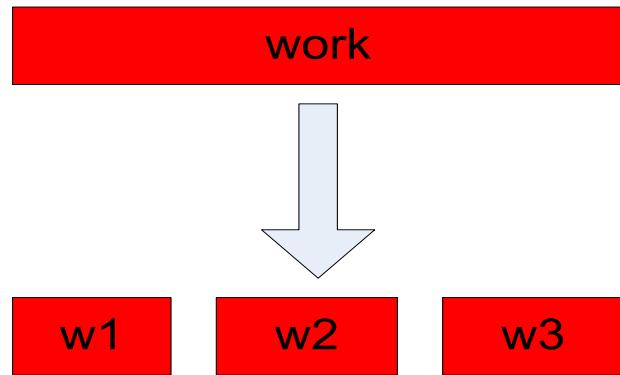
- Dish network supplies TV reception via satellite; they collect data on their set top box and send it back to headquarters

2. Tesla collecting every usage of the car

- Tesla's are connected to the cellular network; the car reports back all of its actions to Tesla

Why Parallelization is Hard

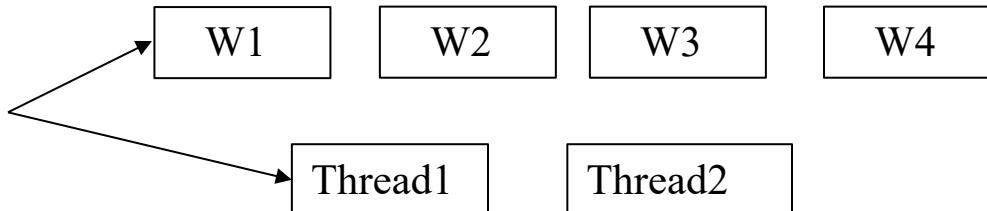
- Parallelization is “easy” if processing can be cleanly split into n units:



- And there is an easy way to combine the outputs

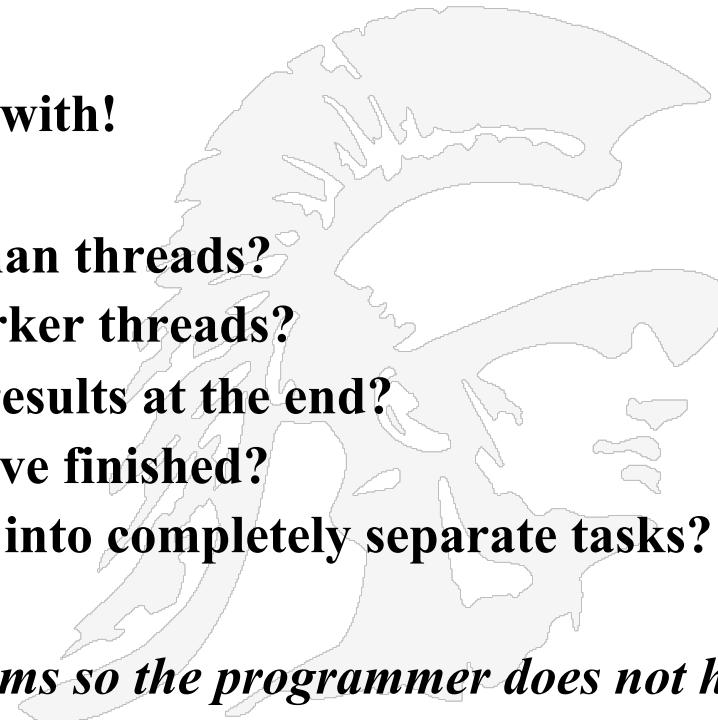
Why Parallelization is Hard

we would like to have as many threads as there are work units, but this may not be the case

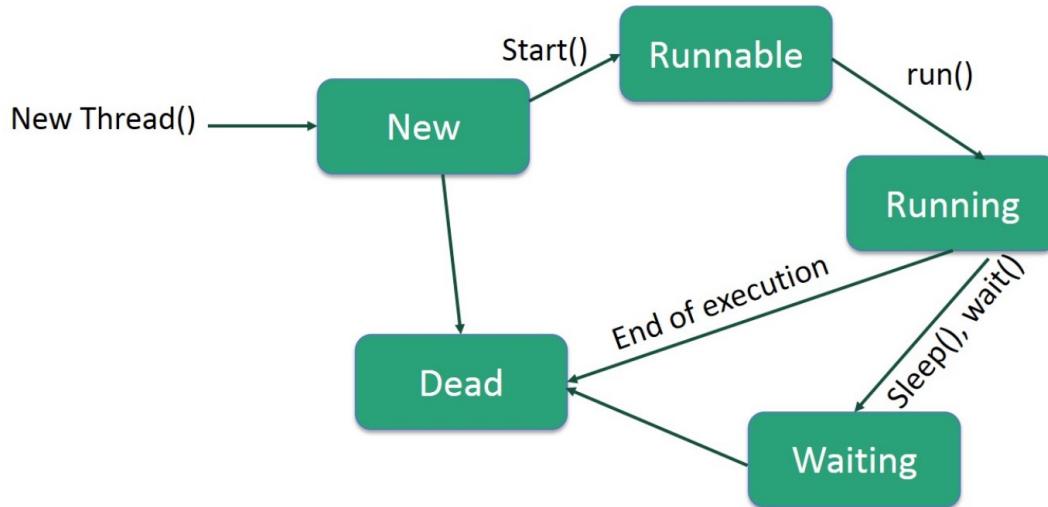


So there are complicated issues to deal with!

- What if we have more work units than threads?
 - How do we assign work units to worker threads?
 - How do we aggregate/combine the results at the end?
 - How do we know all the workers have finished?
 - What if the work cannot be divided into completely separate tasks?
-
- *MapReduce solves all of these problems so the programmer does not have to deal with them*



Programming with Multiple Threads Poses Challenges



Life Cycle of a Thread

Thread 1:

```
void foo() {
```

```
    x++;
```

```
    y = x;
```

```
}
```

Thread 2:

```
void bar() {
```

```
    y++;
```

```
    x++;
```

```
}
```

If the initial state is $x = 6$, $y = 0$, what are the final values of x and y after the threads finish running? Possible solutions include: (8,8) and (8,7)

Multithreaded = Unpredictability

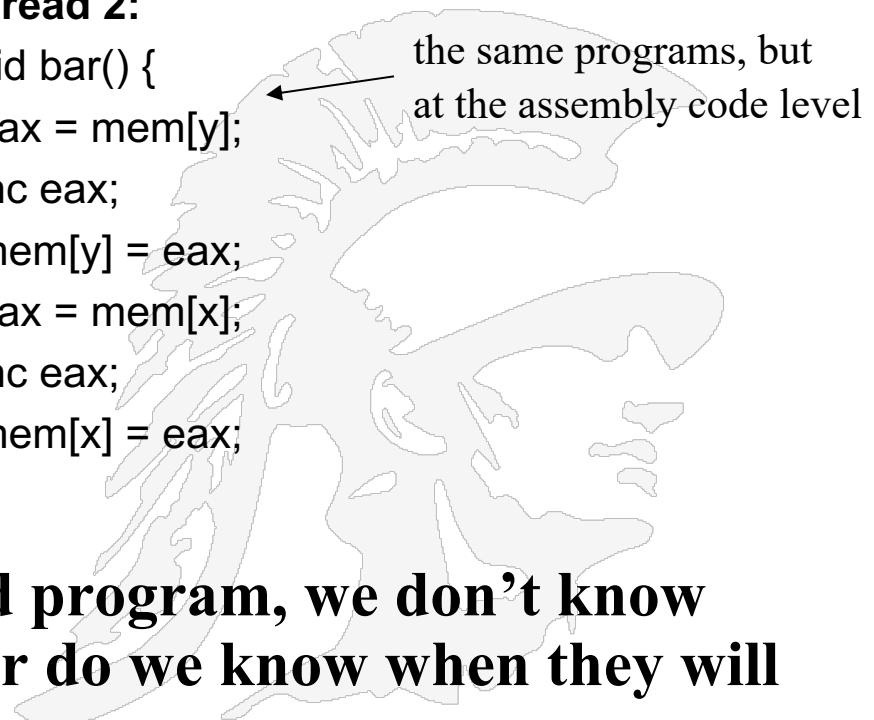
- Many things that look like “one step” operations actually take several steps under the hood:

Thread 1:

```
void foo() {  
    eax = mem[x];  
    inc eax;  
    mem[x] = eax;  
    ebx = mem[x];  
    mem[y] = ebx;  
}
```

Thread 2:

```
void bar() {  
    eax = mem[y];  
    inc eax;  
    mem[y] = eax;  
    eax = mem[x];  
    inc eax;  
    mem[x] = eax;  
}
```



- When we run a multithreaded program, we don't know what order threads run in, nor do we know when they will interrupt one another.

The “corrected” example

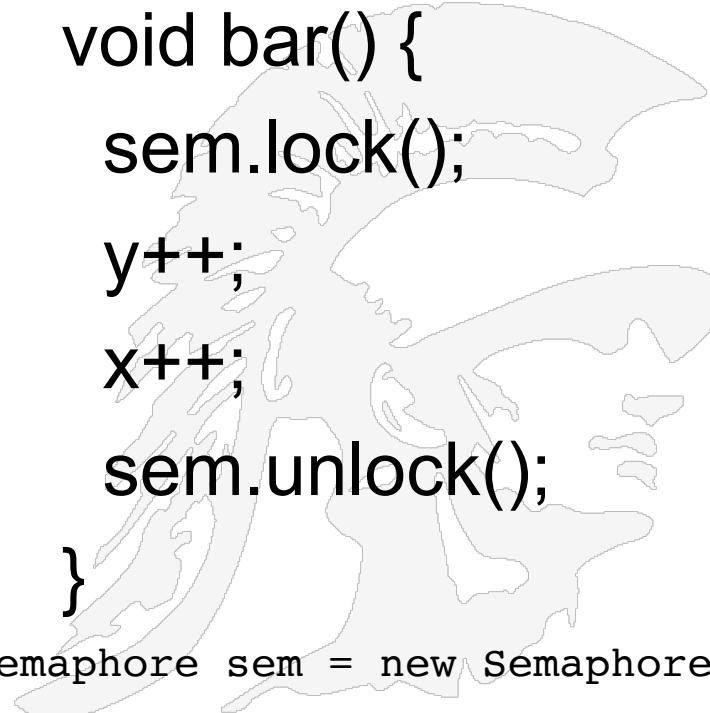
Thread 1:

```
void foo() {  
    sem.lock();  
  
    x++;  
  
    y = x;  
  
    sem.unlock();  
}
```

The global variable `sem`, as defined here `Semaphore sem = new Semaphore();` guards access to `x` & `y`

Thread 2:

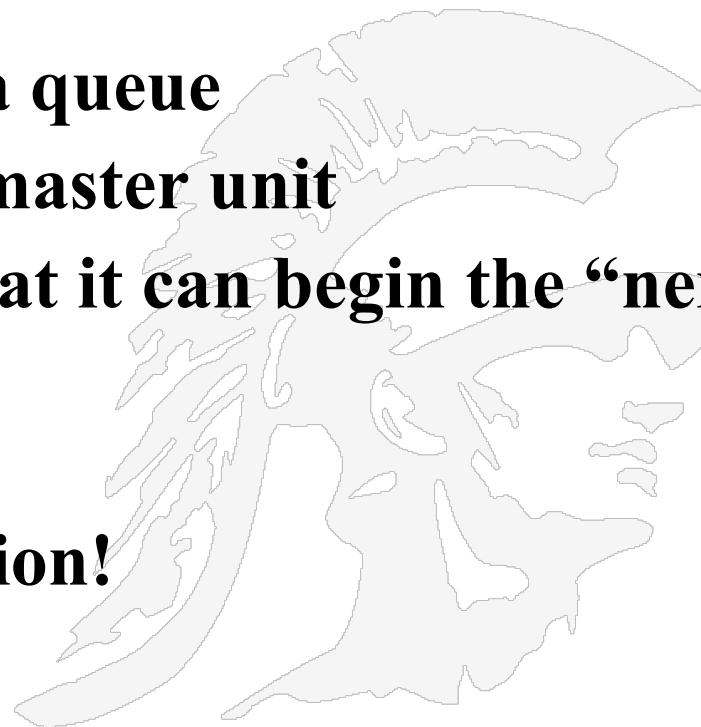
```
void bar() {  
    sem.lock();  
  
    y++;  
  
    x++;  
  
    sem.unlock();  
}
```



Unpredictability on Many Levels

This applies to more than just low level operations:

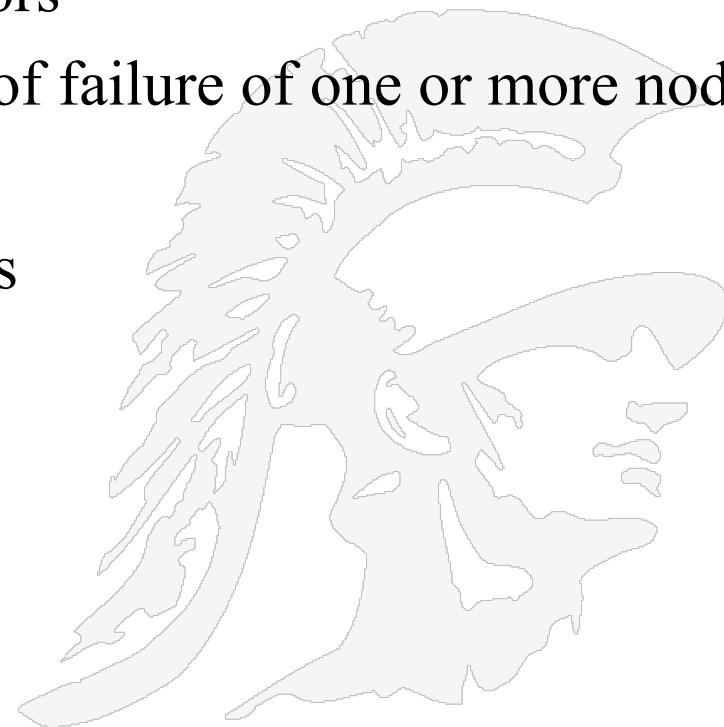
- Pulling work units from a queue
- Reporting work back to master unit
- Telling another thread that it can begin the “next phase” of processing



... All require synchronization!

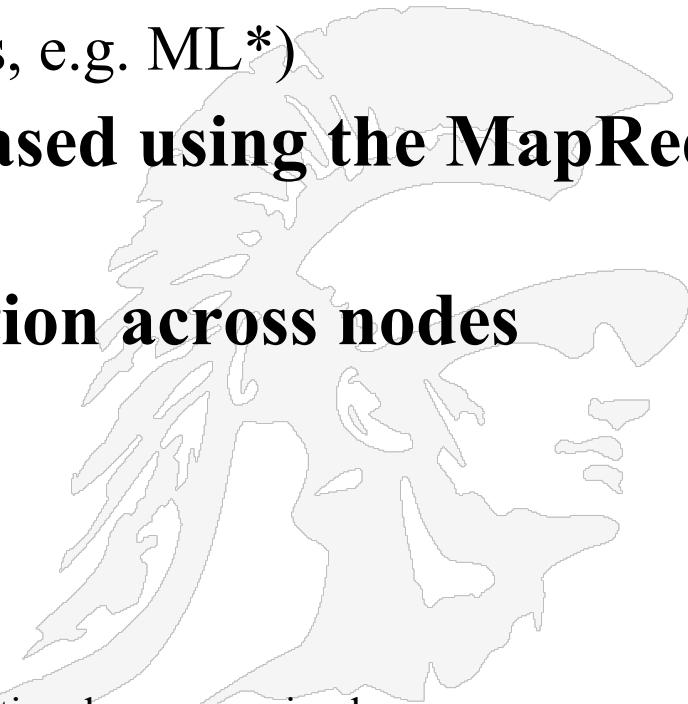
How MapReduce Solves the Parallelization Problems

- So MapReduce provides
 - Automatic parallelization of code across multiple threads and across multiple processors
 - Fault tolerance in the event of failure of one or more nodes
 - I/O scheduling
 - Monitoring & Status updates



Map/Reduce - Beginnings

- **Map/Reduce**
 - Is a programming model borrowed from the programming language Lisp
 - (and other functional languages, e.g. ML*)
- **Many problems can be phrased using the MapReduce paradigm**
- **Easy to distribute computation across nodes**
- **Nice retry/failure semantics**



*ML programming language is a general purpose functional programming language, see
[https://en.wikipedia.org/wiki/ML_\(programming_language\)](https://en.wikipedia.org/wiki/ML_(programming_language))

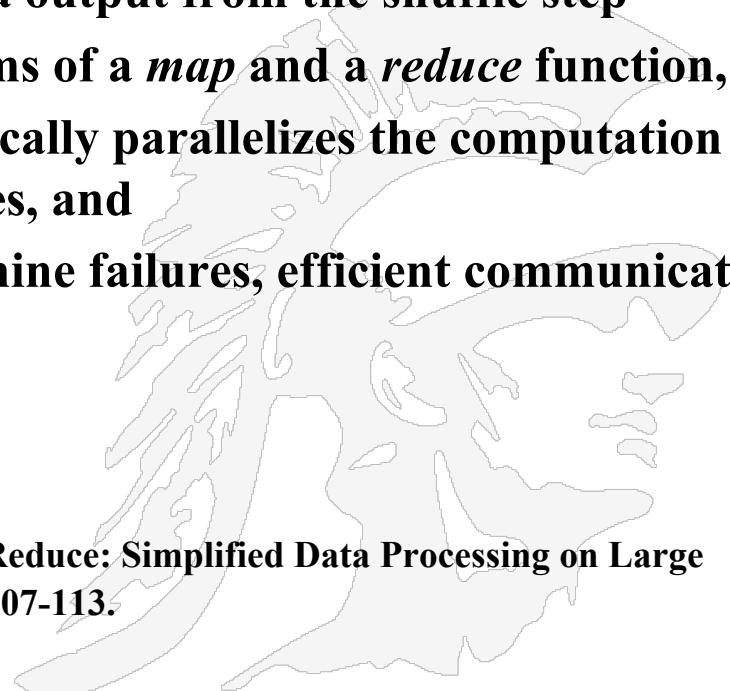
Map and Reduce Functions in LISP (Scheme is a dialect of Lisp)

- **(map *f* *list* [*list*₂ *list*₃ ...])** General formulation
- Specific example
 - (map square '(1 2 3 4))
 - (1 4 9 16)
- **(reduce *f* *id* *list*)**
- Specific example
 - (reduce + 0 '(1 4 9 16))
 - (+ 16 (+ 9 (+ 4 (+ 1 0))))
 - 30



What is MapReduce?

- MapReduce is a programming model that generically works this way:
 - A *map function* extracts some intelligence from raw data
 - A *shuffle step* organizes the resulting output
 - A *reduce function* aggregates the data output from the shuffle step
 - Users specify the computation in terms of a *map* and a *reduce* function,
 - Underlying runtime system automatically parallelizes the computation across large-scale clusters of machines, and
 - Underlying system also handles machine failures, efficient communications, and performance issues.
- Reference: Dean, J. and Ghemawat, S. 2008 “MapReduce: Simplified Data Processing on Large Clusters”, *Communication of ACM* 51, 1 (Jan. 2008), 107-113.



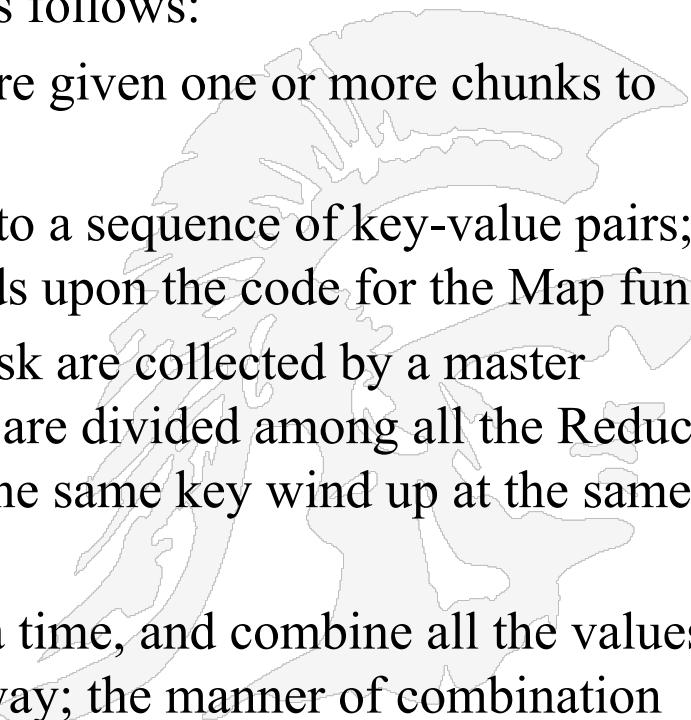
The Map/Reduce Paradigm

1. A large number of records are broken into segments
2. **Map:** extracts something of interest from each segment
3. **Group:** sorts the intermediate results from each segment (sometimes called **shuffle**)
4. **Reduce:** aggregates intermediate results
5. Generate final output

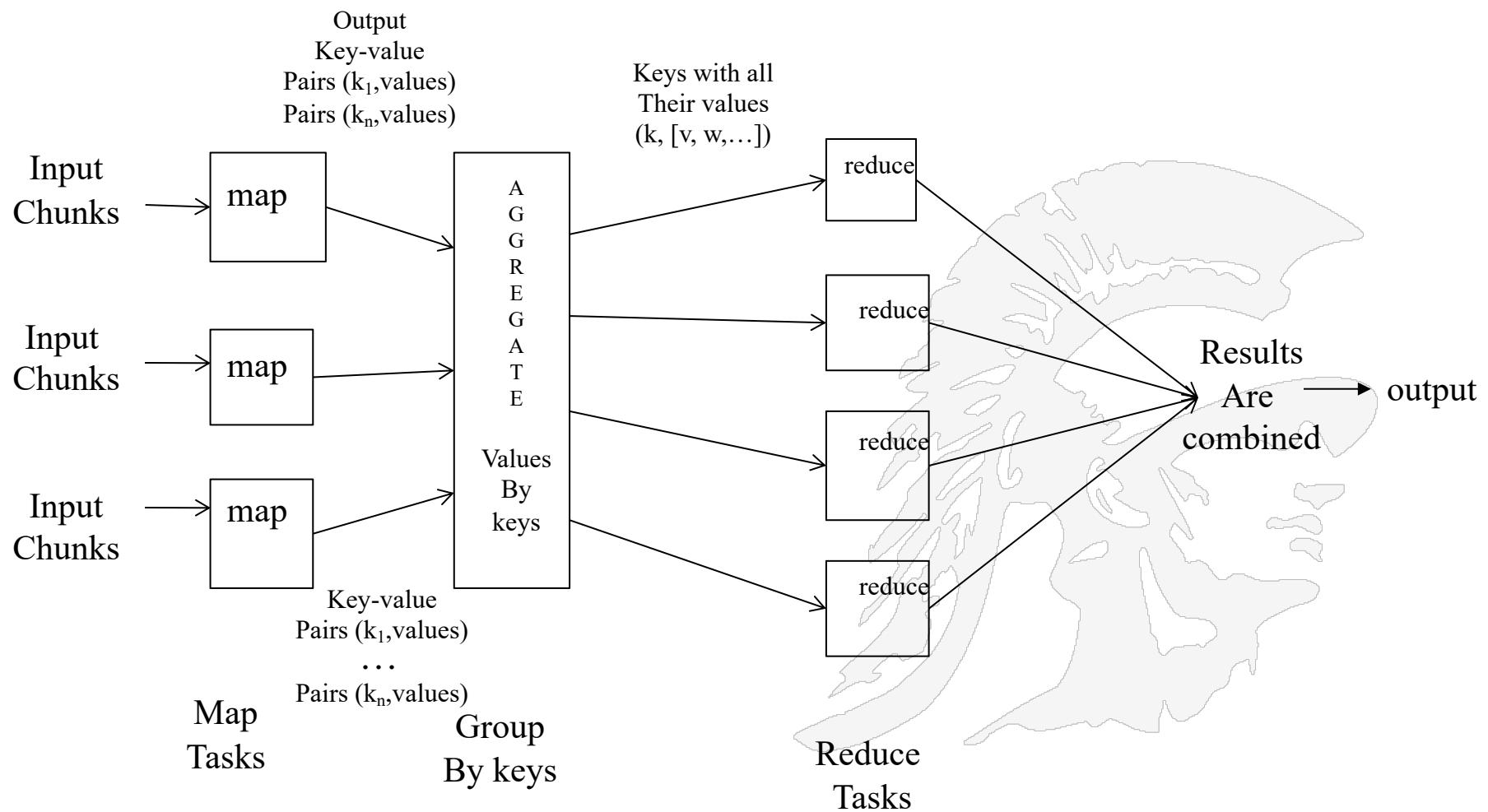
Key idea: to re-phrase problems in such a way that the input can be divided into parts and operated on in parallel and the results combined to produce a solution to the original problem

The Map&Reduce Routines

- Using map-reduce one must write 2 functions called *Map* and *Reduce*
- The system manages the parallel execution and coordination of tasks; it is all done automatically
- A map-reduce computation proceeds as follows:
 1. Some number of map tasks each are given one or more chunks to process
 2. These map tasks turn the chunk into a sequence of key-value pairs; the way the pairs are produced depends upon the code for the Map function
 3. Key-value pairs from each Map task are collected by a master controller and sorted by key; keys are divided among all the Reduce tasks, so all key-value pairs with the same key wind up at the same Reduce task
 4. Reduce tasks work on one key at a time, and combine all the values associated with that key in some way; the manner of combination depends upon the Reduce code

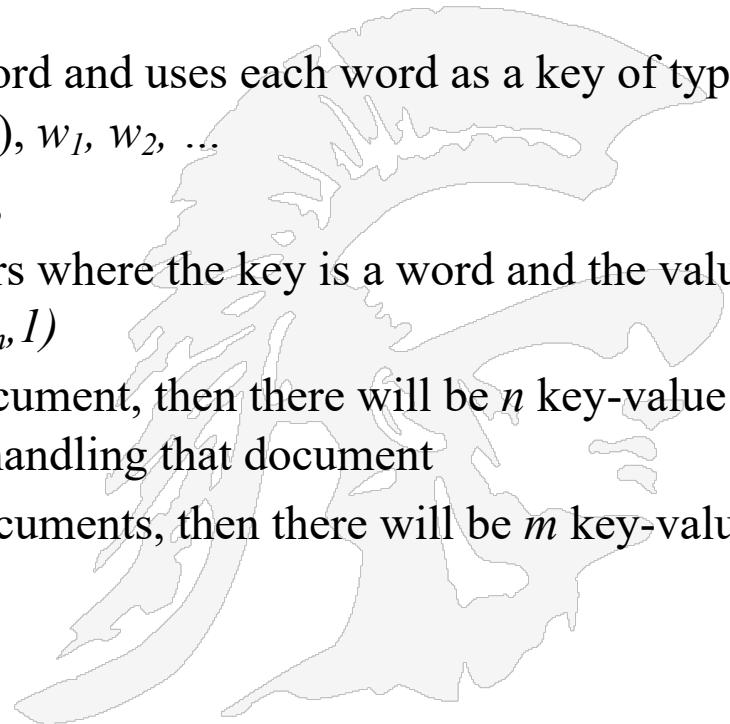


Schematic of a Map-Reduce Computation



A MapReduce Example – Counting Word Occurrences

- Counting the number of occurrences for each word in a collection of documents
- The input file is a repository of documents
- Each document is an element passed to a separate processor
- The Map function
 - Parses the document, extracts each word and uses each word as a key of type String (the words obtained by parsing), w_1, w_2, \dots
 - For each word it assigns an integer, 1;
 - Each processor outputs key-value pairs where the key is a word and the value is always 1, namely $(w_1, 1), (w_2, 1), \dots, (w_n, 1)$
- If a word w appears n times in a single document, then there will be n key-value pairs $(w, 1)$ in the output of the processor handling that document
- If a word w appears m times among all documents, then there will be m key-value pairs $(w, 1)$ in the output



Count Word Occurrences Pseudo-Code

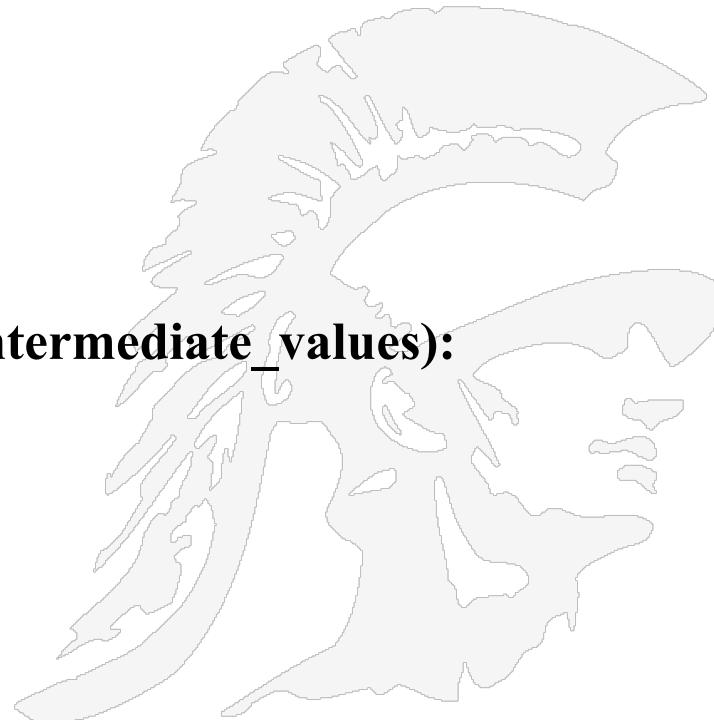
- The code below is similar to what a programmer would write to process multiple documents on a cluster of machines using map/reduce

Map(String input_key, String input_value):

```
// input_key: document name  
// input_value: document contents  
for each word w in input_value:  
    EmitIntermediate(w, "1");
```

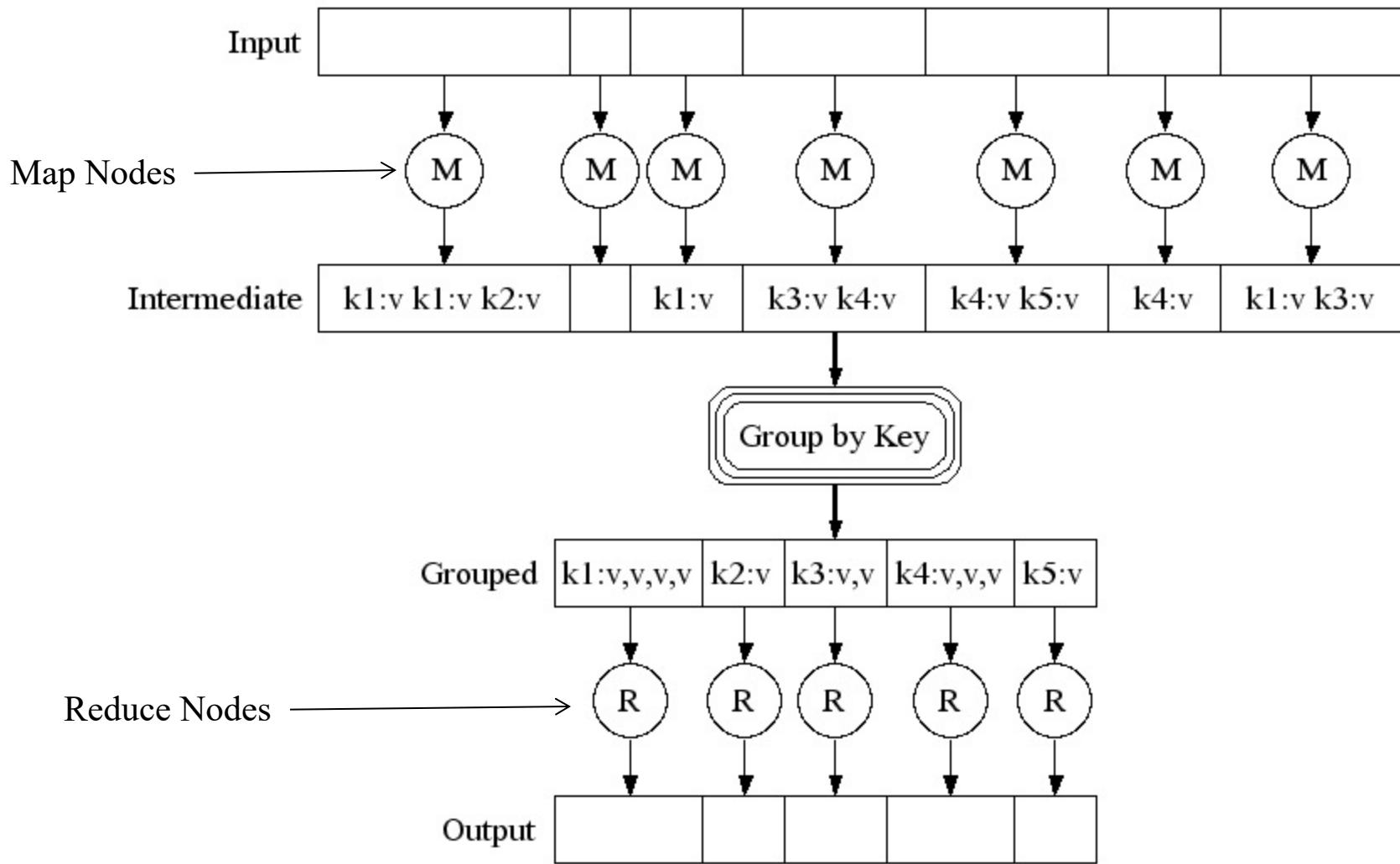
reduce(String output_key, Iterator intermediate_values):

```
// output_key: a word  
// output_values: a list of counts  
int result = 0;  
for each v in intermediate_values:  
    result += ParseInt(v);  
Emit(AsString(result));
```

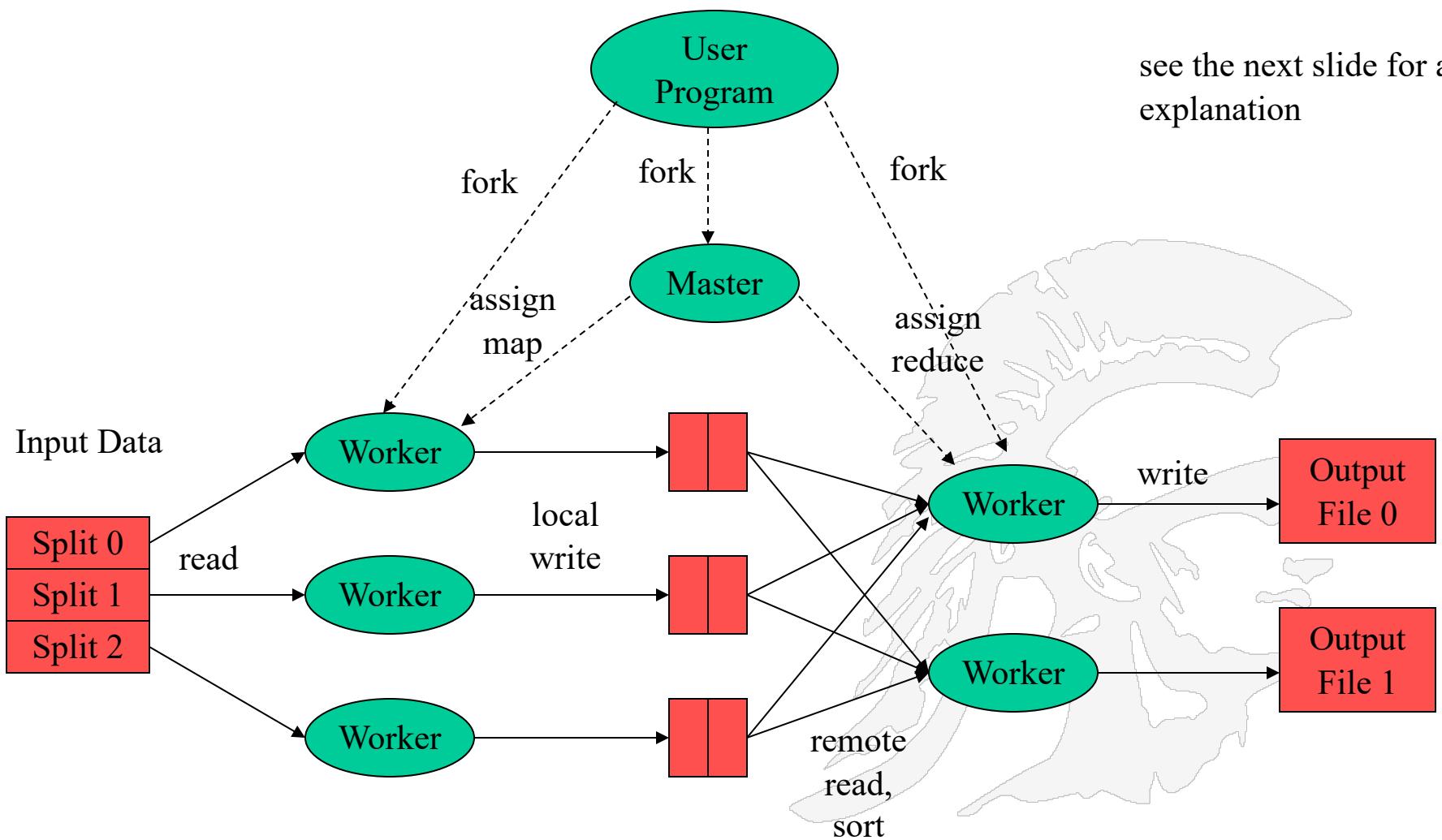




Word Count Execution An Alternate Graphical View



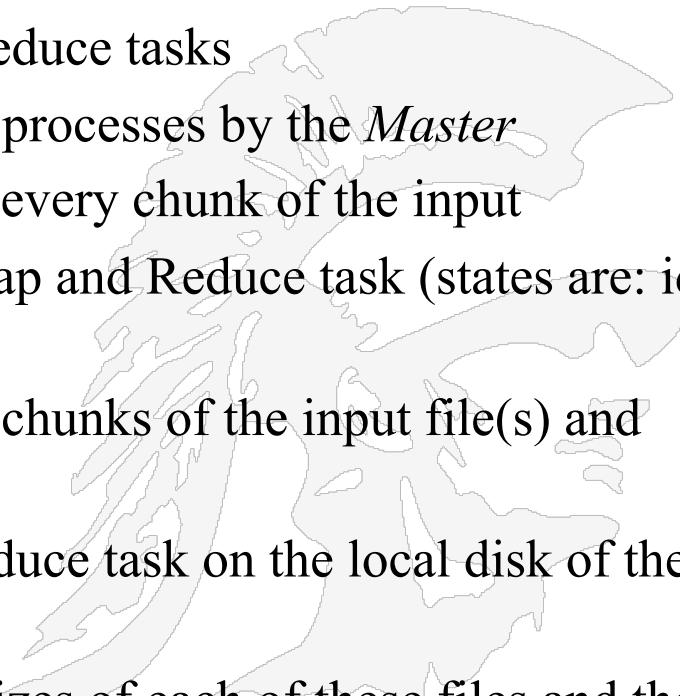
Distributed Execution Overview



see the next slide for an explanation

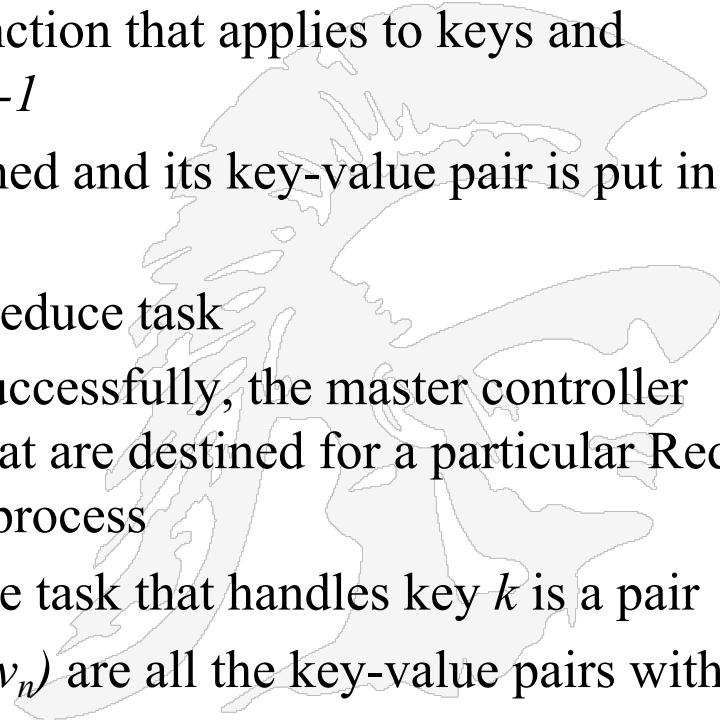
Parallel Execution of Map-Reduce - Looking Under the Hood

- The user program forks a *Master* controller process and some number of *Worker* processes at different compute nodes;
- A *Worker* handles either Map tasks or Reduce tasks, but not both
- The *Master* must
 - Create some number of Map and Reduce tasks
 - These tasks are assigned to *Worker* processes by the *Master*
 - Typically there is one Map task for every chunk of the input
 - Keeps track of the status of each Map and Reduce task (states are: idle, executing on a Worker, completed)
- Each Map task is assigned one or more chunks of the input file(s) and executes on it the code
- The Map task creates a file for each Reduce task on the local disk of the Worker that executes the Map task
- The *Master* is told of the location and sizes of each of these files and the Reduce task for which each is destined



Looking Under the Hood at the Reduce Task

- There is a *master controller* process that knows how many Reduce tasks there will be, say r
- The user defines r
- The master controller picks a hash function that applies to keys and produces a bucket number from 0 to $r-1$
- Each key output by a Map task is hashed and its key-value pair is put in one of r local files
 - Each file will be processed by a Reduce task
- After all Map tasks have completed successfully, the master controller merges the file from each Map task that are destined for a particular Reduce task and feeds the merged file to that process
- For each key k , the input to the Reduce task that handles key k is a pair $(k, [v_1, \dots, v_n])$ where $(k, v_1), (k, v_2), \dots, (k, v_n)$ are all the key-value pairs with key k coming from all the Map tasks



Explanation of the Reduce Task

- The *Reduce function* is written to take pairs consisting of a key and a list of associated values, and combines them in some way
- The *Reduce function* output is a sequence of key-value pairs consisting of each input key k paired with the combined value
- Outputs from all Reduce tasks are merged into a single file
- *Reduce function* adds up all the values and outputs a sequence of (w,m) pairs where w is a word that appears at least once in the documents and m is the total number of occurrences
- The *Reduce function* is generally associative and commutative implying values can be combined in any order yielding the same result

Fault Tolerance in MapReduce

1. If a task crashes:

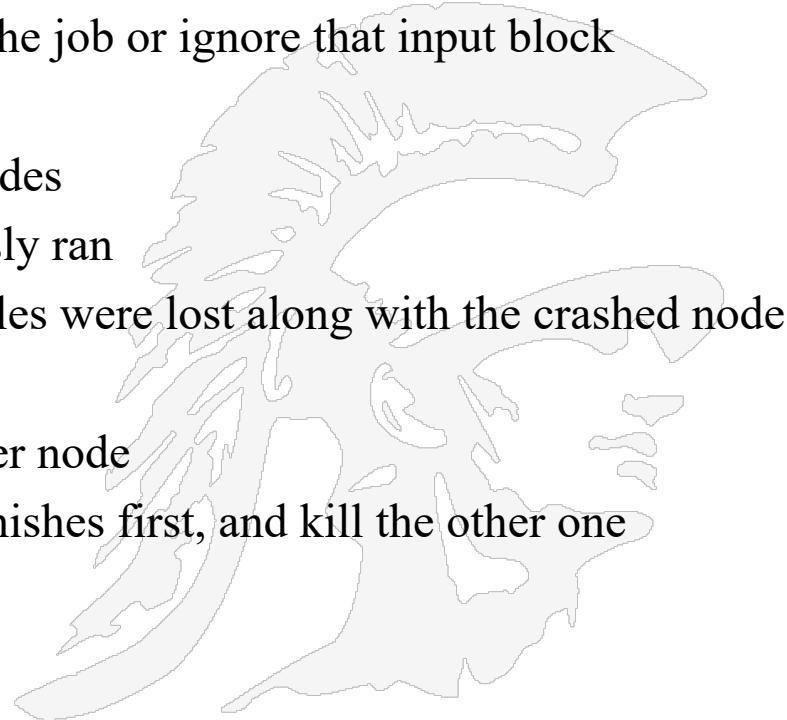
- Retry on another node
 - OK for a map because it had no dependencies
 - OK for reduce because map outputs are on disk
- If the same task repeatedly fails, fail the job or ignore that input block

2. If a node crashes:

- Relaunch its current tasks on other nodes
- Relaunch any maps the node previously ran
 - Necessary because their output files were lost along with the crashed node

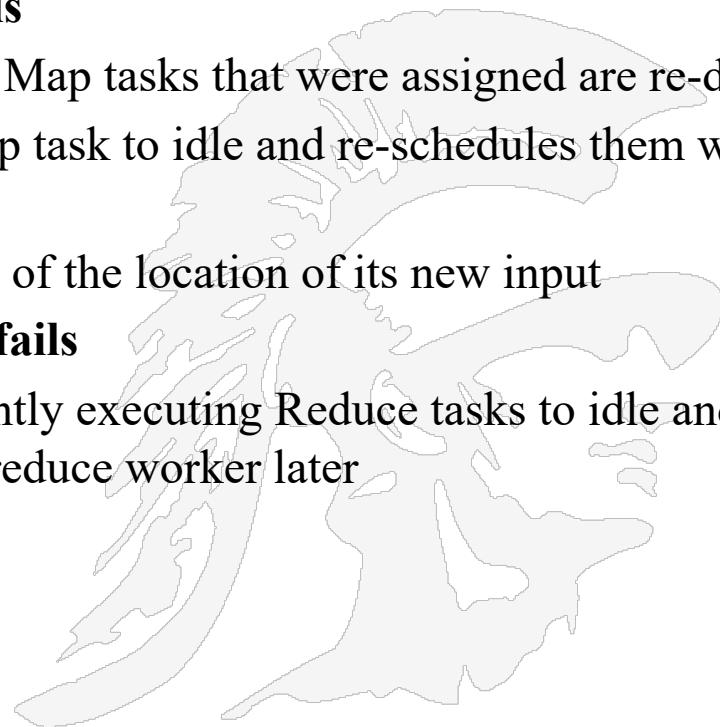
3. If a task is going slowly (straggler):

- Launch second copy of task on another node
- Take the output of whichever copy finishes first, and kill the other one



Coping with Node Failure

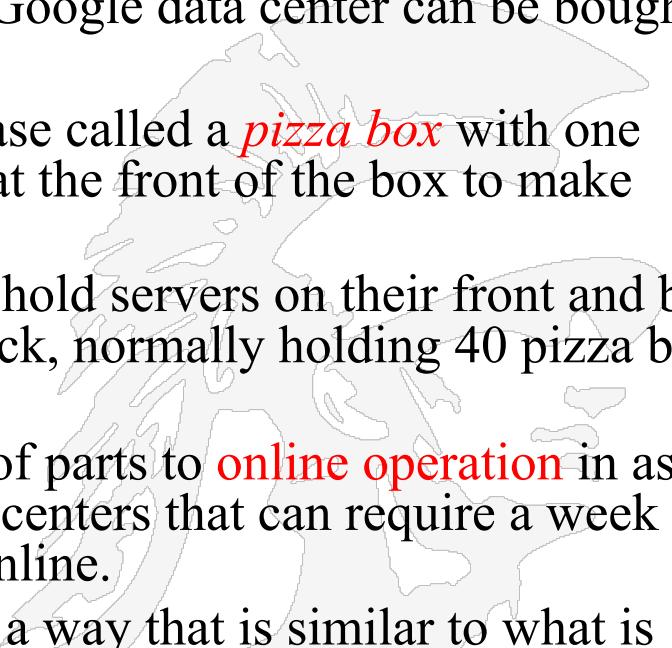
- Worst case: **the compute node where the Master is executing fails**
 - **Result:** the entire map-reduce job must be restarted
- Other failures are less severe and are handled by the Master
- **The compute node of a Map worker fails**
 - This is detected by the Master and all Map tasks that were assigned are re-done
 - The Master sets the status of each Map task to idle and re-schedules them when a worker becomes available
 - The Master informs each Reduce task of the location of its new input
- **The compute node of a Reduce worker fails**
 - The Master sets the status of its currently executing Reduce tasks to idle and they will be re-scheduled on another reduce worker later



Typical Data Center Cluster

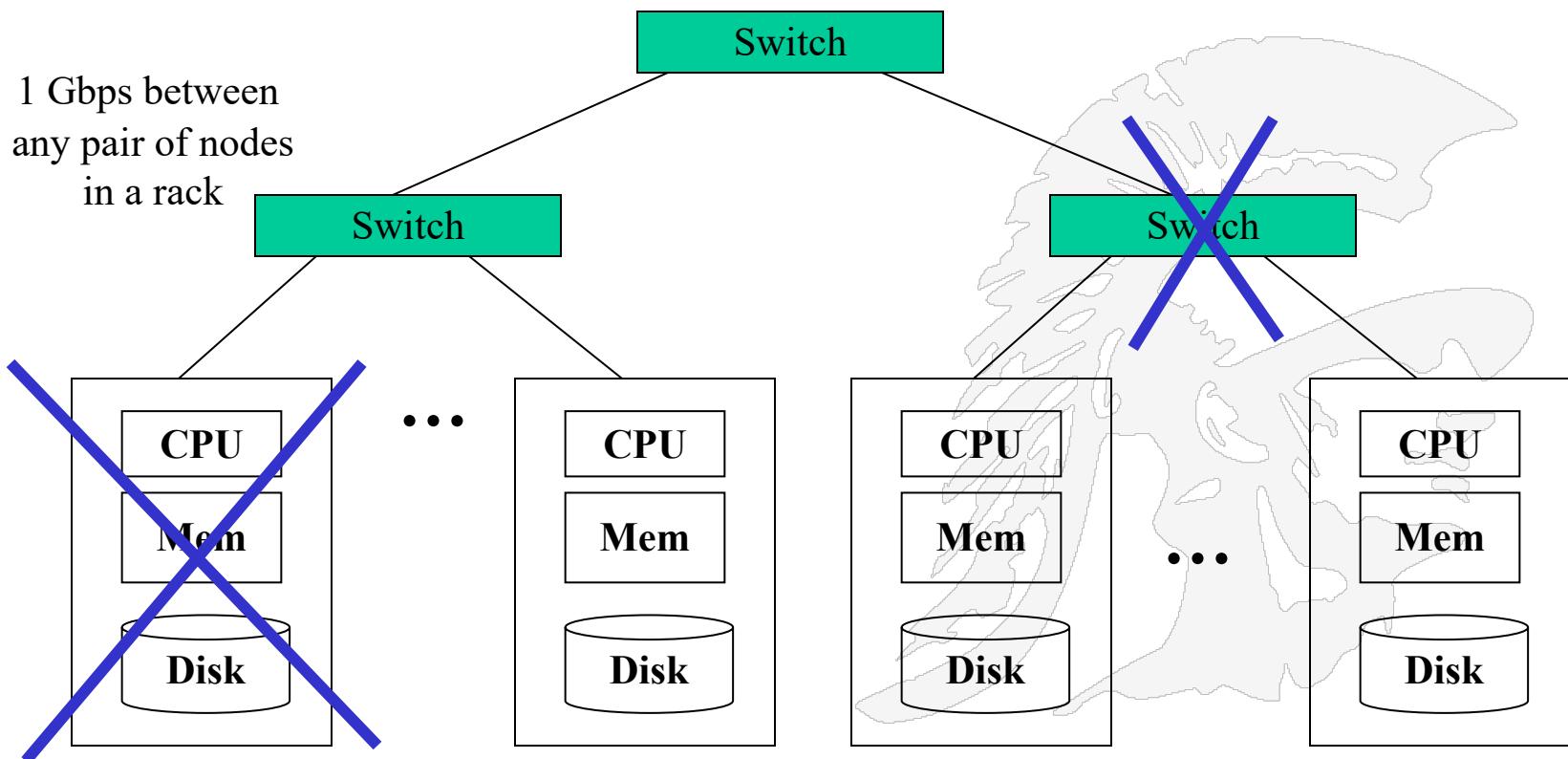


Characteristics of a Google DataCenter

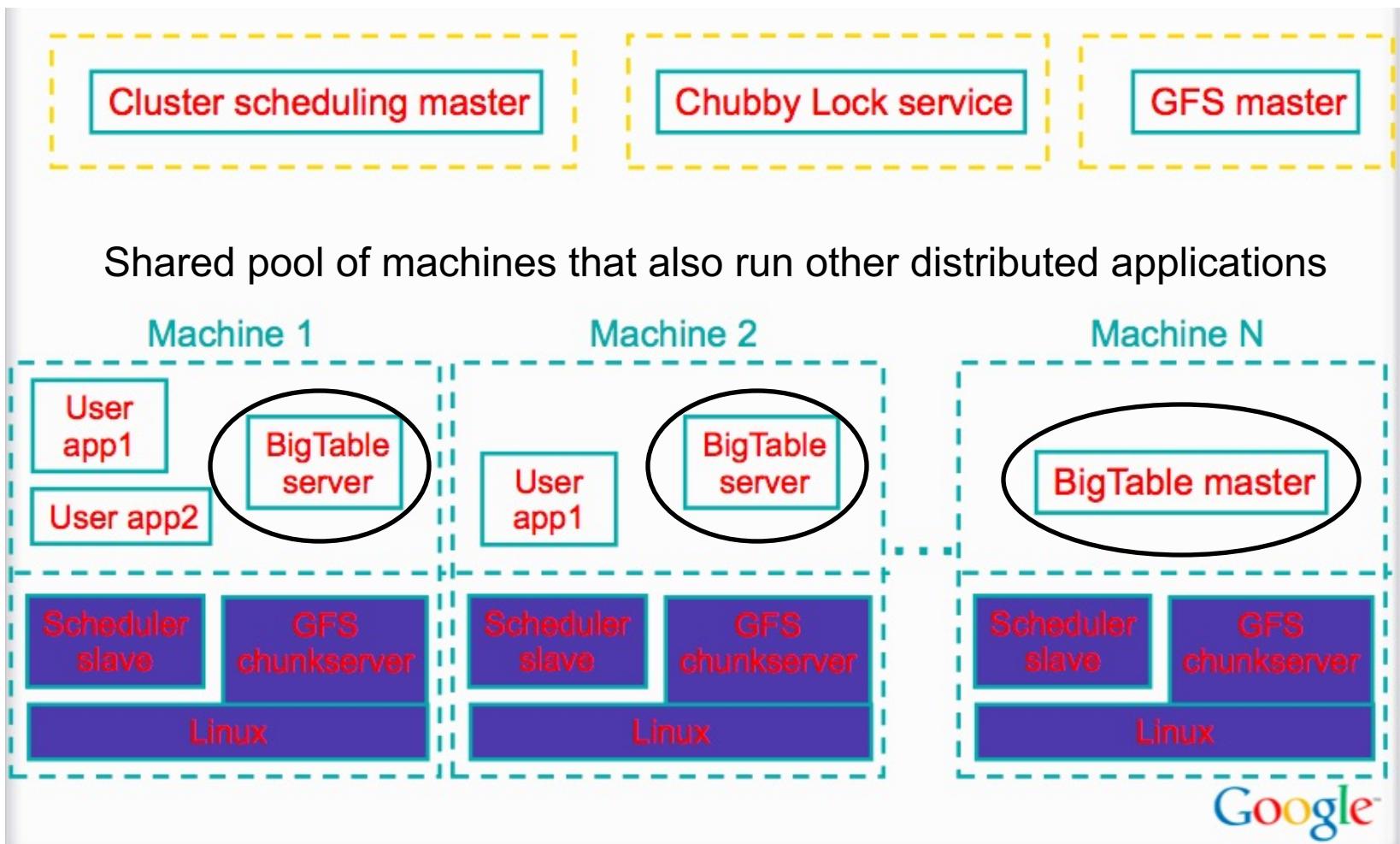
- 
1. **Google data centers** (approx. two dozen): They come online and automatically, under the direction of the Google File System, start getting work from other data centers. These facilities, sometimes filled with 10,000 or more Google computers, find one another and configure themselves with minimal human intervention.
 2. **Standard desktop PCs**: The hardware in a Google data center can be bought at a local computer store.
 3. Each Google server comes in a standard case called a *pizza box* with one important change: the plugs and ports are at the front of the box to make access faster and easier.
 4. **Google racks** are assembled for Google to hold servers on their front and back sides. This effectively allows a standard rack, normally holding 40 pizza box servers, to hold 80 servers.
 5. A Google data center can go from a stack of parts to **online operation** in as little as **72 hours**, unlike more typical data centers that can require a week or even a month to get additional resources online.
 6. Each server, rack and data center works in a way that is similar to what is called "**plug and play**." Like a mouse plugged into the USB port on a laptop, Google's network of data centers knows when more resources have been connected. These resources, **for the most part**, go into operation without ²⁹human intervention.

Typical Cluster Architecture

- Each rack of cpu's contains between 16-64 nodes
- Nodes within a single rack are connected by gigabyte Ethernet
- Each rack is connected to another rack by a switch with speeds of 2-10 Gbps
- Individual cpu's can fail; switches between racks can fail
2-10 Gbps backbone between racks

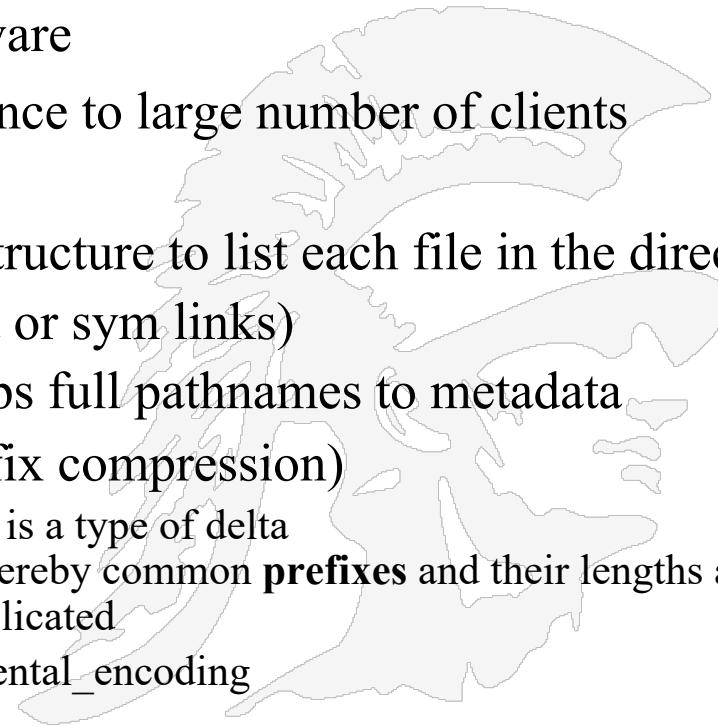


Typical Cluster Machine Configuration



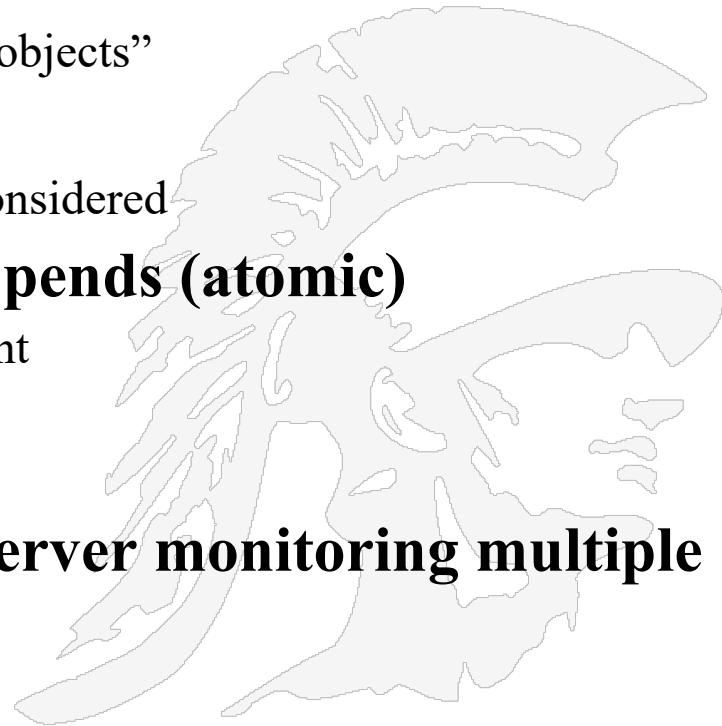
Google File System General Goals

- A scalable, distributed file system for large distributed data-intensive applications
 - Provides fault tolerance
 - Runs on cheap, commodity hardware
 - Delivers high aggregate performance to large number of clients
- GFS: not your typical file system
 - Lacks typical per-directory data structure to list each file in the directory
 - Does not support aliases (i.e. hard or sym links)
 - Namespace: lookup table that maps full pathnames to metadata
 - Lookup table fits in memory (prefix compression)
 - Also known as incremental encoding is a type of delta encoding **compression** algorithm whereby common **prefixes** and their lengths are recorded so that they need not be duplicated.
 - https://en.wikipedia.org/wiki/Incremental_encoding



The Google File System Design Assumptions –

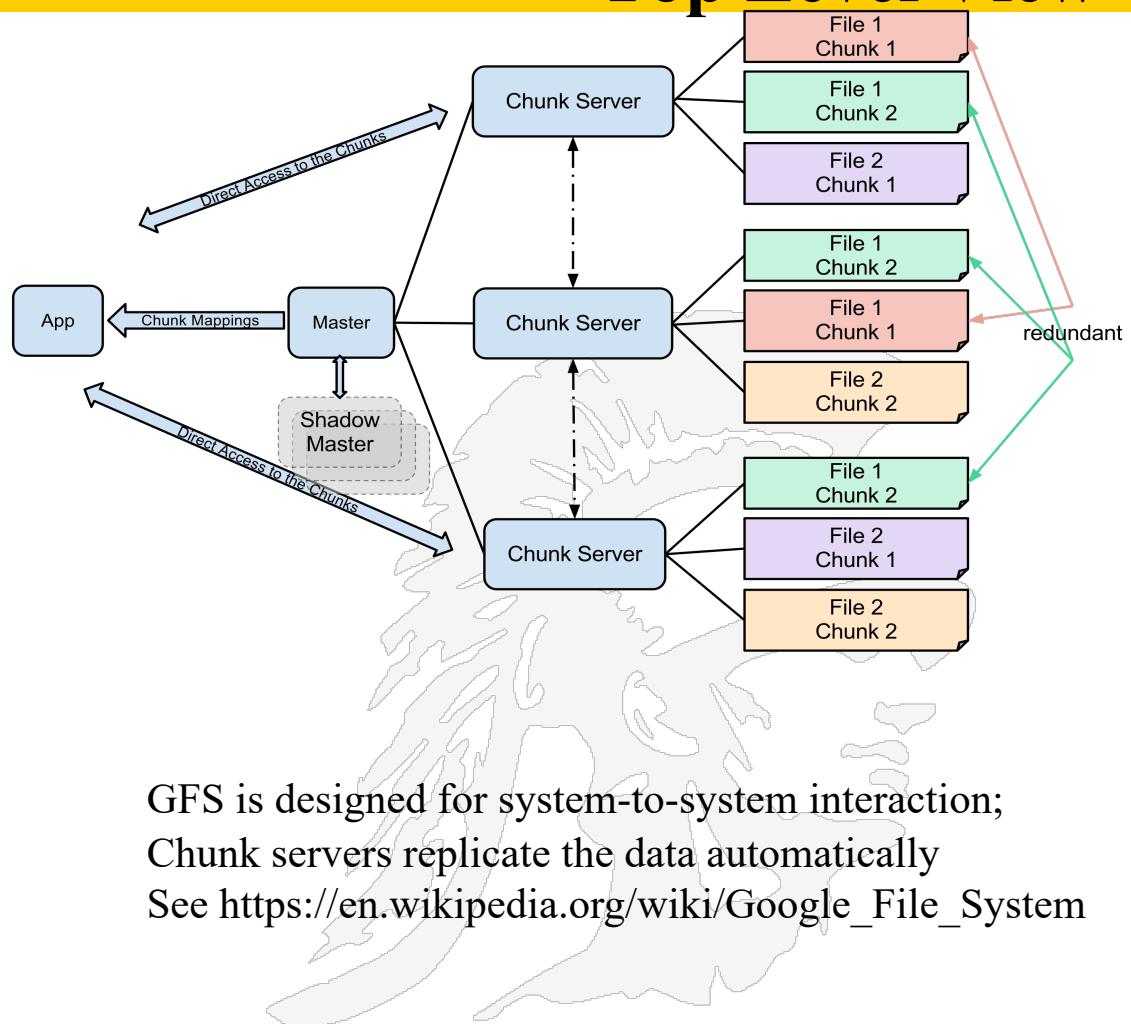
- **Files will be HUGE**
- **Multi-gigabyte files are common**
 - Not practical to have ~8 billion files
 - Each file contains many “application objects”
- **Multi-terabyte datasets**
 - I/O operations, block sizes must be considered
- **Most file modifications are appends (atomic)**
 - Random writes practically non-existent
 - Once written... sequential reads
 - Caching not terribly important
- **there will be a single master server monitoring multiple chunk servers**



Google File System

Top Level View

- **Google File System (GFS)**, is a proprietary distributed file system for efficient, reliable access to data using large clusters of commodity hardware
- Files are divided into fixed-size *chunks* of 64 megabytes, similar to clusters or sectors in regular file systems, which are only extremely rarely overwritten, or shrunk; files are usually appended to or read



GFS is designed for system-to-system interaction;
 Chunk servers replicate the data automatically
 See https://en.wikipedia.org/wiki/Google_File_System

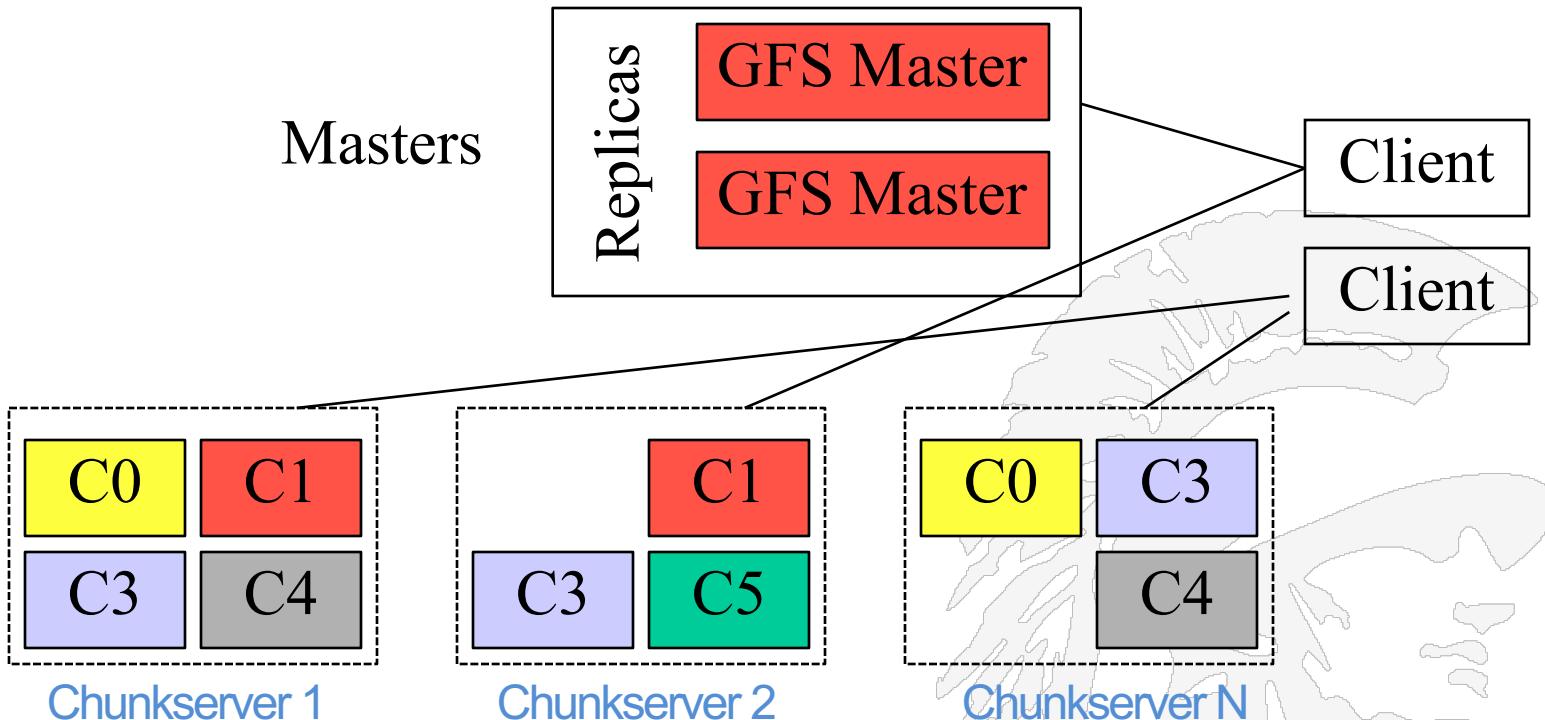
Master Server and Chunk Servers

- **Master Server** holds all metadata:
 - Namespace (directory hierarchy)
 - Accesscontrol information (per-file)
 - Mapping from files to chunks
 - Current locations of chunks (chunkservers)
- Delegates **consistency** management
- **Garbage collects** orphaned chunks
- Migrates chunks between chunkservers

• **Chunk Server**

- Stores 64 MB file chunks on **local disk** using standard **Linux filesystem**, each with version number and checksum
- Read/write requests specify **chunk handle** and **byte range**
- Chunks replicated on configurable number of chunkservers (default: 3)
- **No caching of file data** (beyond standard Linux buffer cache)

Google File System (GFS)



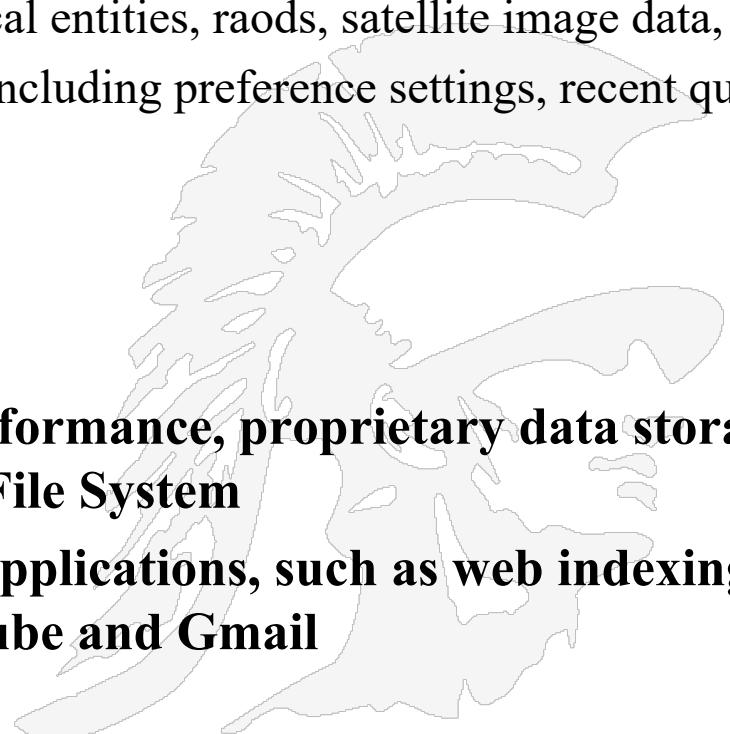
- Master manages metadata
- Data transfers happen directly between clients/chunkservers
- Files broken into chunks (typically 64 MB)
- Chunks triplicated across three machines for safety

GFS: Major Aspects

- **Append vs. Rewrite**
 - GFS is optimized for appended files rather than rewrites. That's because clients within Google rarely need to overwrite files -- they add data onto the end of files instead. While it's still possible to overwrite data on a file in the GFS, the system doesn't handle those processes very efficiently
- **Which Replica Does GFS use?**
 - The GFS separates replicas into two categories: **primary replicas** and **secondary replicas**. A primary replica is the chunk that a chunkserver sends to a client. Secondary replicas serve as backups on other chunkservers.
 - The master server decides which chunks will act as primary or secondary. If the client makes changes to the data in the chunk, then the master server lets the chunkservers with secondary replicas know they have to copy the new chunk off the primary chunkserver to stay current.
- **What About Big Files?**
 - If a client creates a write request that affects multiple chunks of a particularly large file, the GFS breaks the overall write request up into an individual request for each chunk. The rest of the process is the same as a normal write request.
- **Heartbeats and Handshakes**
 - The GFS components give system updates through electronic messages called **heartbeats** and **handshakes**. These short messages allow the master server to stay current with each chunkserver's status.

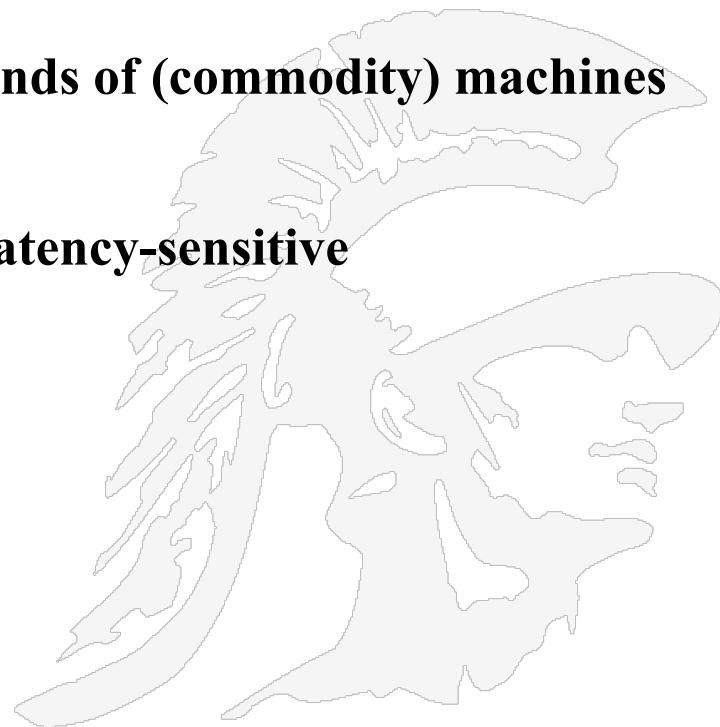
Google File System vs. BigTable

- **GFS provides raw data storage**
- **But Google needs a system for handling:**
 - Trillions of URLs
 - Geographic locations such as physical entities, roads, satellite image data, etc
 - Per user data for billions of people including preference settings, recent queries and searches
 - And it must be capable of
 - storing semi-structured data
 - Reliable, scalable, etc
- **Bigtable is a compressed, high performance, proprietary data storage system built on top of the Google File System**
- **It is used by a number of Google applications, such as web indexing, MapReduce, Google Maps, YouTube and Gmail**



Motivation and Design Goal

- **Distributed Storage System for Structured Data**
 - **Scalability**
 - Petabytes of data on Thousands of (commodity) machines
 - **Wide Applicability**
 - Throughput-oriented and Latency-sensitive
 - **High Performance**
 - **High Availability**

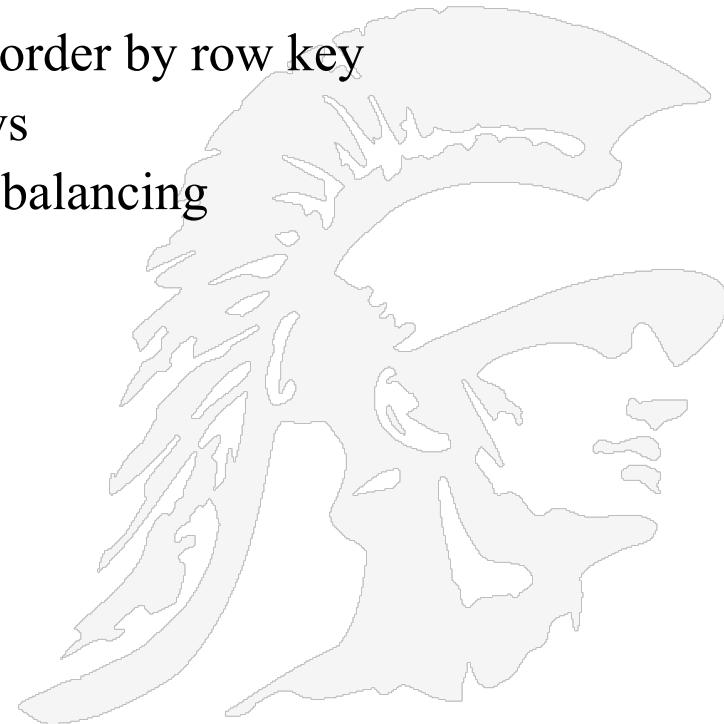


- Not a Full Relational Data Model
- Provides a simple data model
 - Supports dynamic control over data layout
 - Allows clients to reason about the locality properties
- A Table in Bigtable is a:
 - Sparse
 - Distributed
 - Persistent
 - Multidimensional
 - Sorted map

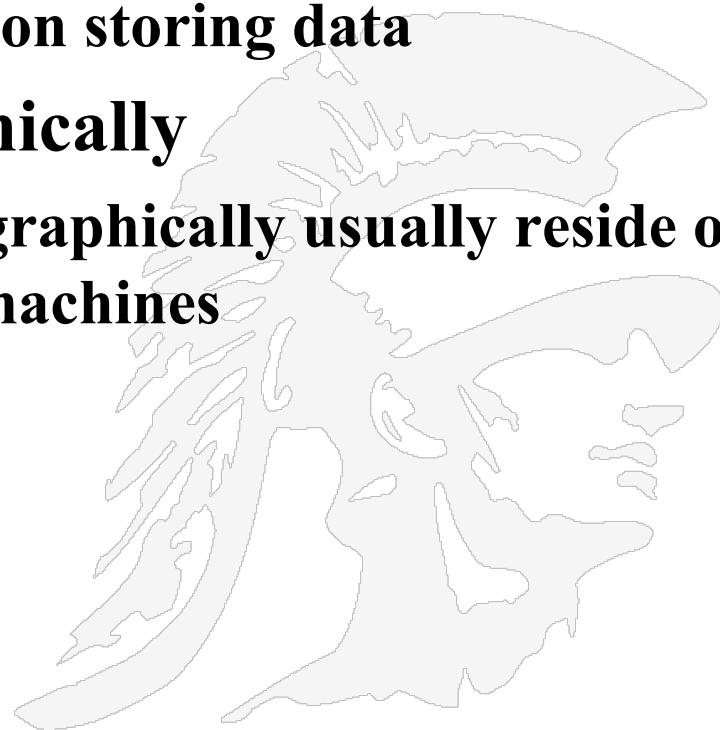


Data Model

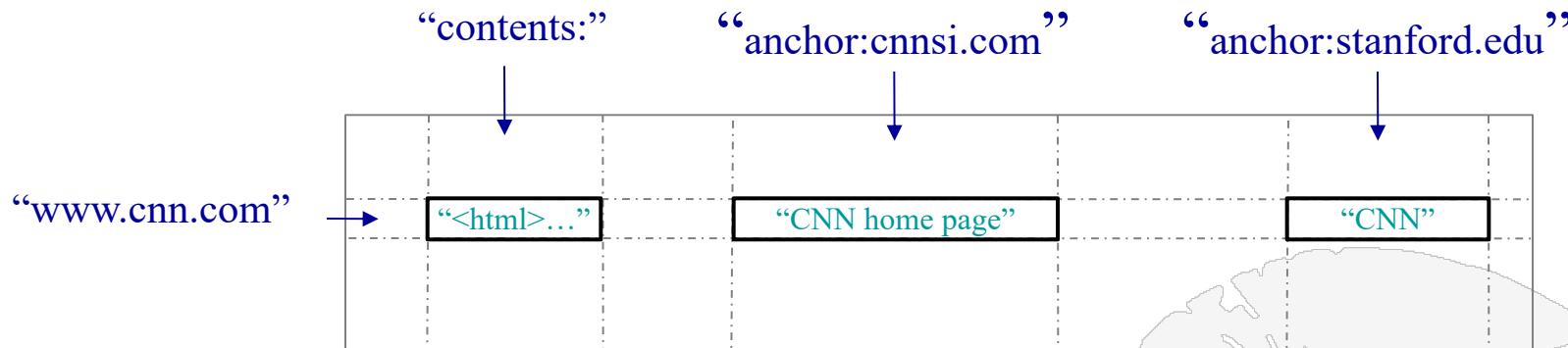
- **Data is indexed using row and column names**
- **Data is treated as uninterpreted strings**
 - $(\text{row:string}, \text{column:string}, \text{time:int64}) \rightarrow \text{string}$
- **Rows**
 - Data maintained in lexicographic order by row key
 - Tablet: rows with consecutive keys
 - Units of distribution and load balancing
- **Columns**
 - Column families
- **Cells**
- **Timestamps**



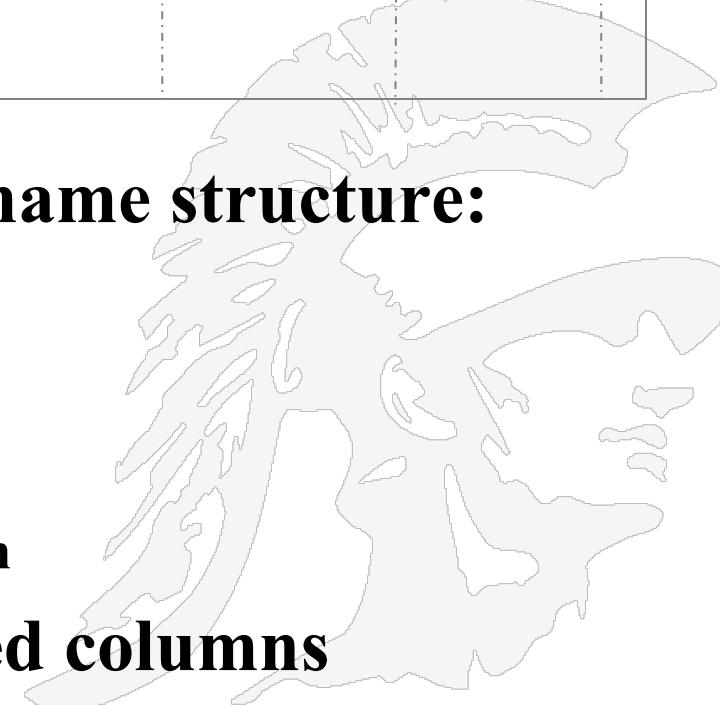
- Name is an arbitrary string
 - Access to data in a row is atomic
 - Row creation is implicit upon storing data
- Rows ordered lexicographically
 - Rows close together lexicographically usually reside on one or a small number of machines



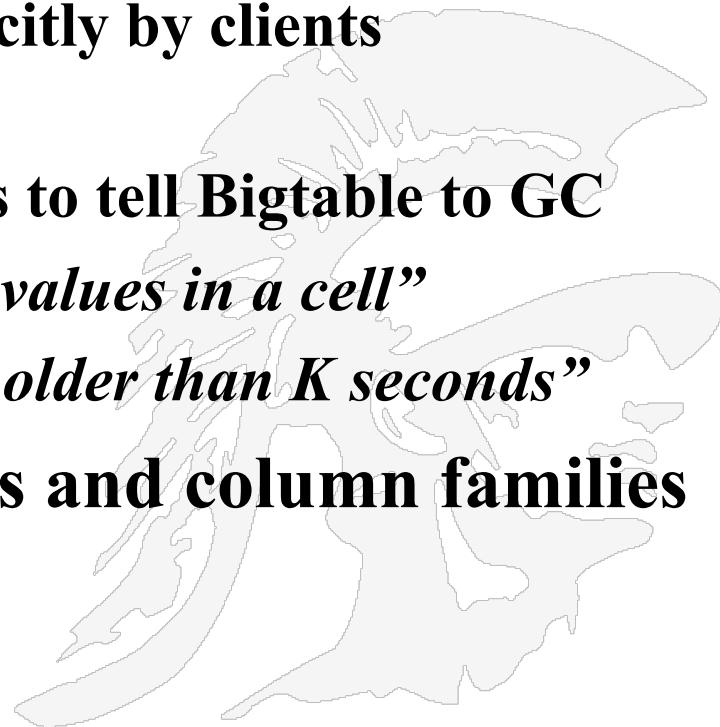
Columns



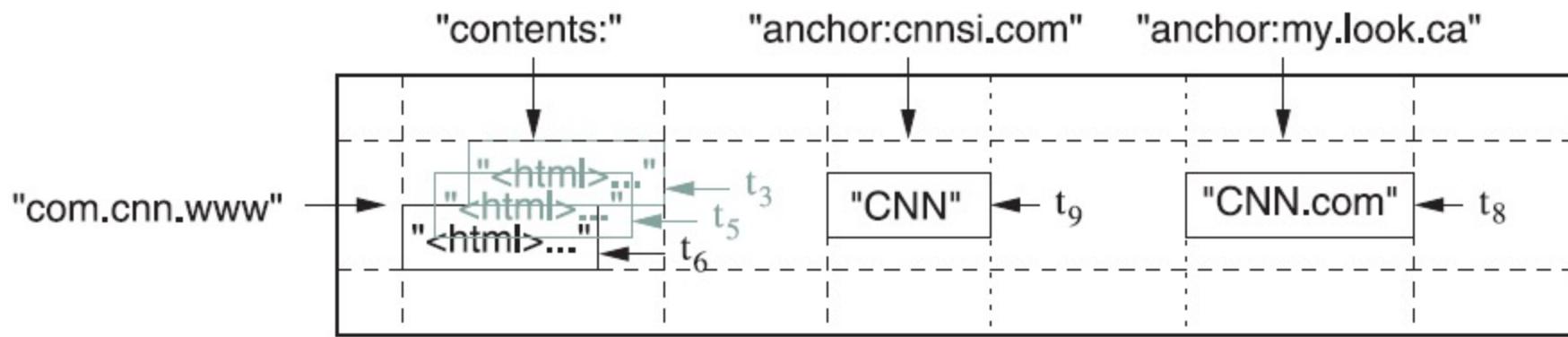
- **Columns have two-level name structure:**
 - **family:optional_qualifier**
- **Column family**
 - Unit of access control
 - Has associated type information
- **Qualifier gives unbounded columns**
 - Additional level of indexing, if desired



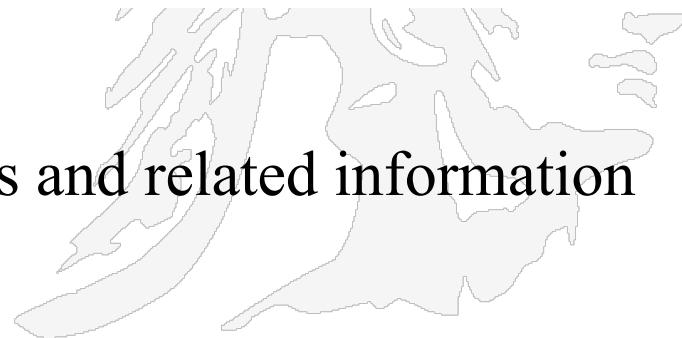
- Used to store different versions of data in a cell
 - New writes default to current time, but timestamps for writes can also be set explicitly by clients
- Garbage Collection
 - Per-column-family settings to tell Bigtable to GC
 - “*Only retain most recent K values in a cell*”
 - “*Keep values until they are older than K seconds*”
- API: Create / delete tables and column families



Data Model – WebTable Example (1 of 7)



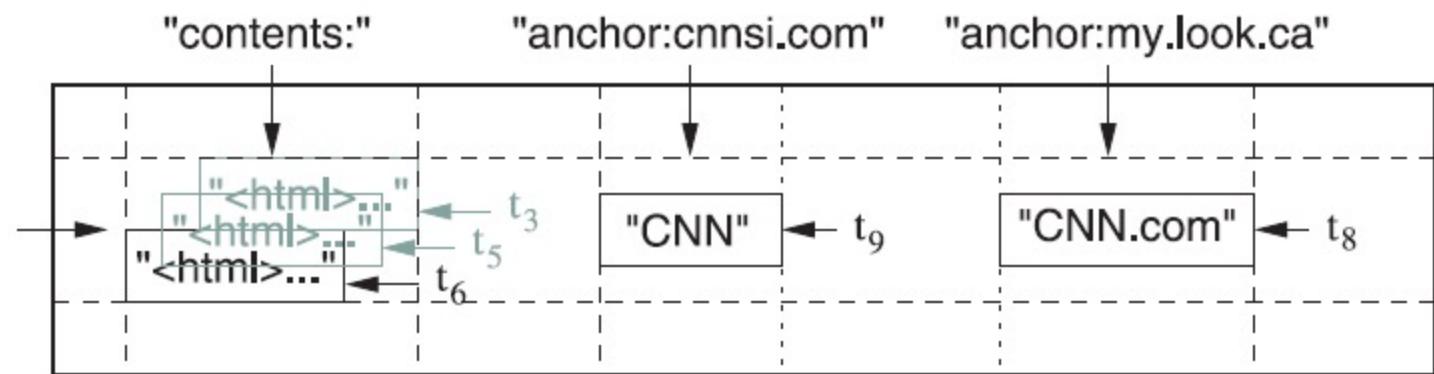
A large collection of web pages and related information



Data Model – WebTable Example (2)

Row Key

"com.cnn.www"

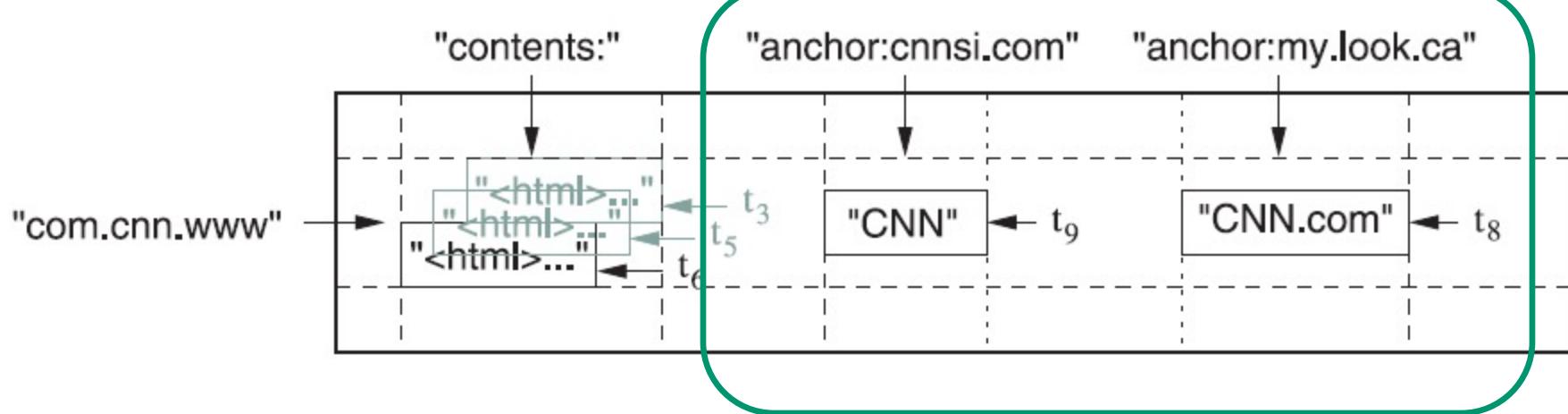


Tablet - Group of rows with consecutive keys.

Unit of Distribution

Bigtable maintains data in lexicographic order by row key

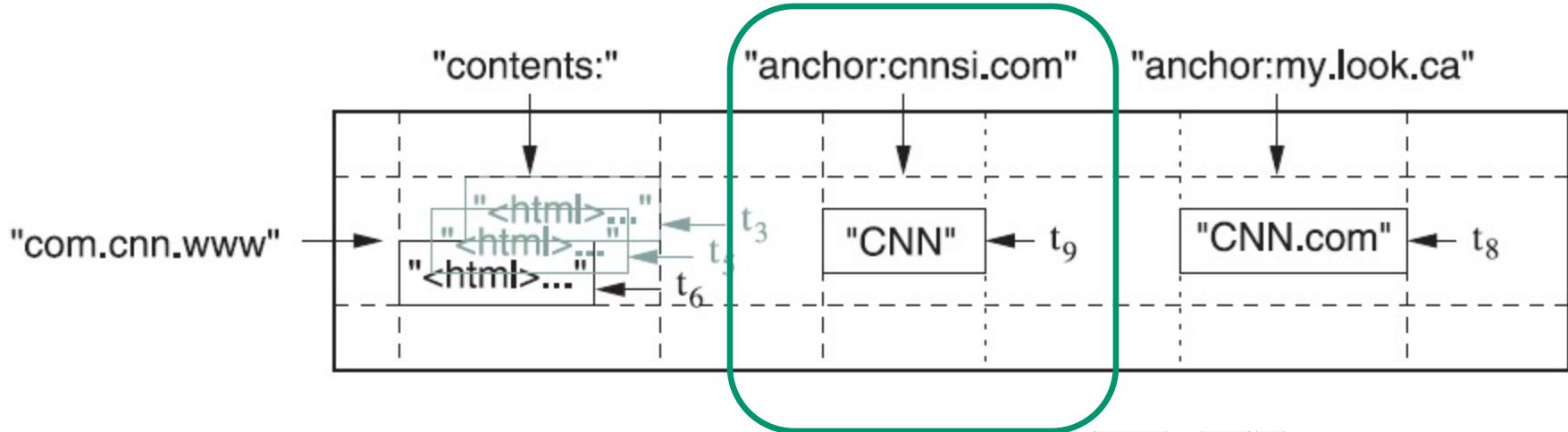
Data Model – WebTable Example (3)



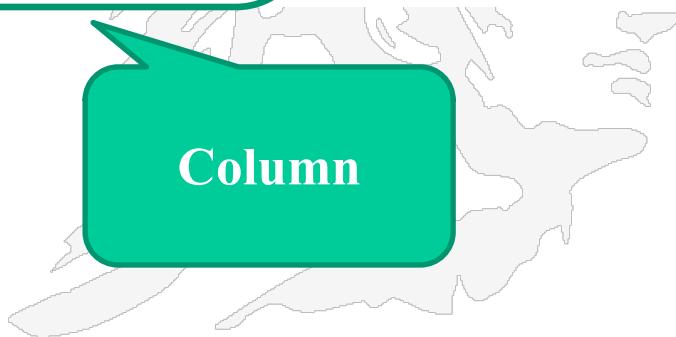
Column family is the unit of access control

Column
Family

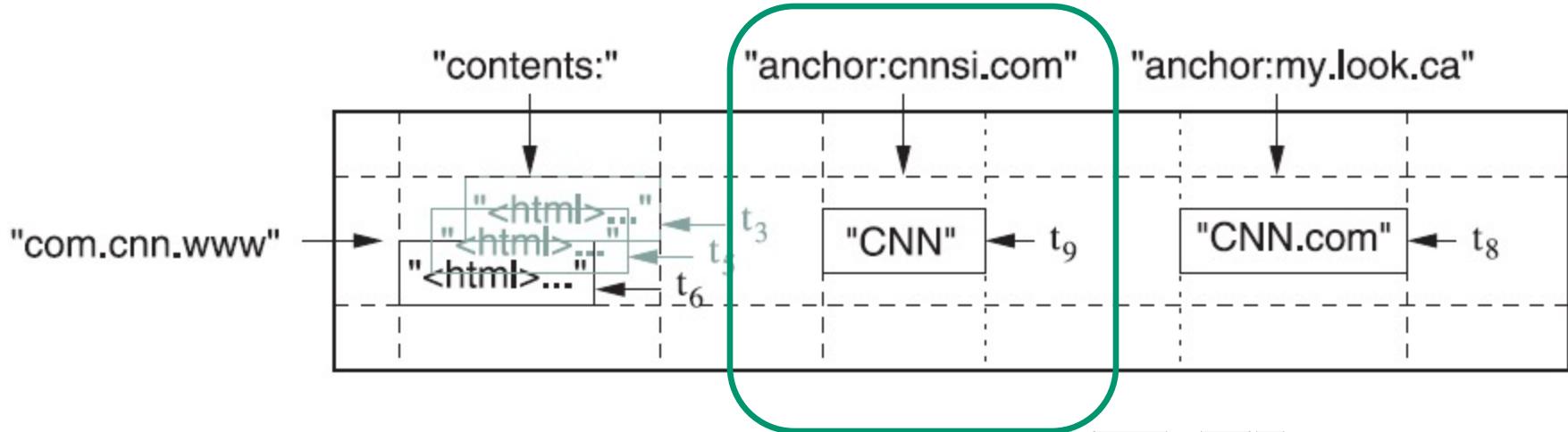
Data Model – WebTable Example (4)



Column key is specified by
“Column family:qualifier”



Data Model – WebTable Example (5)

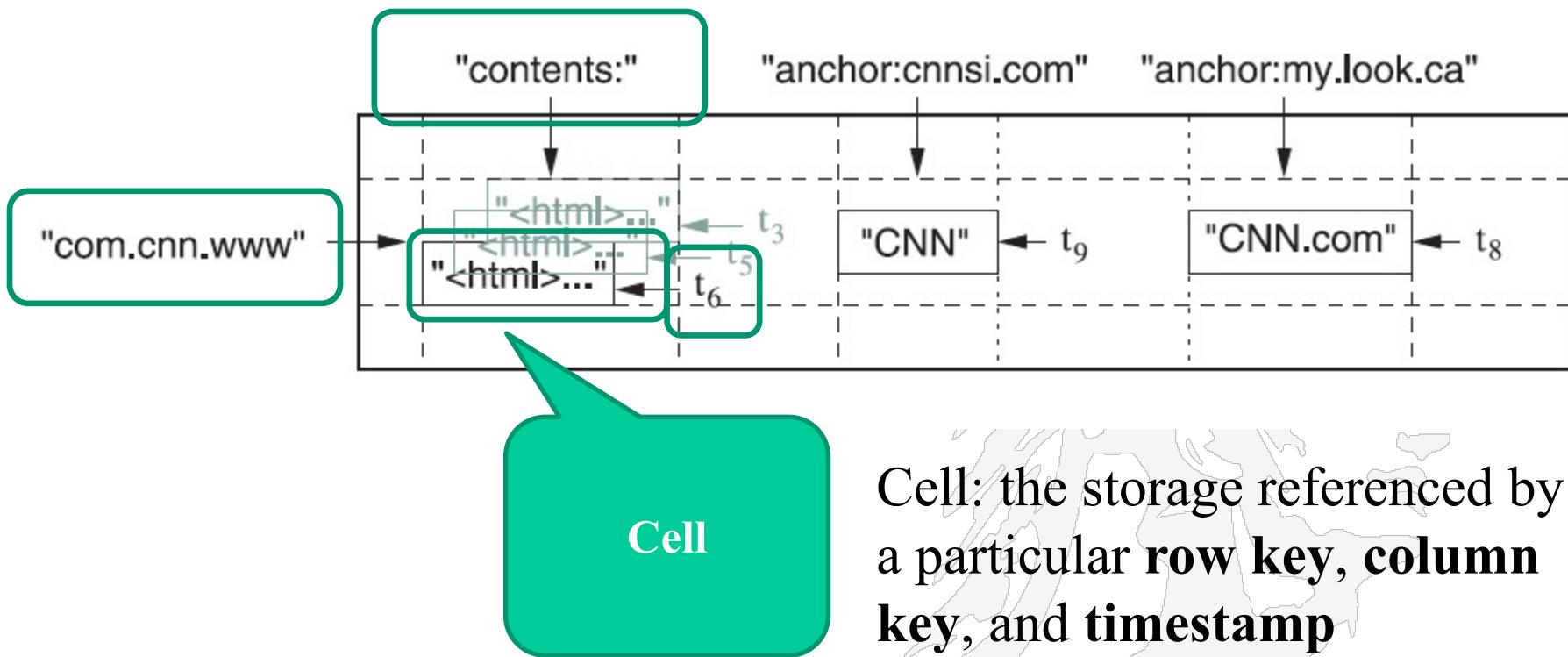


You can add a column in a column family if the column family was created



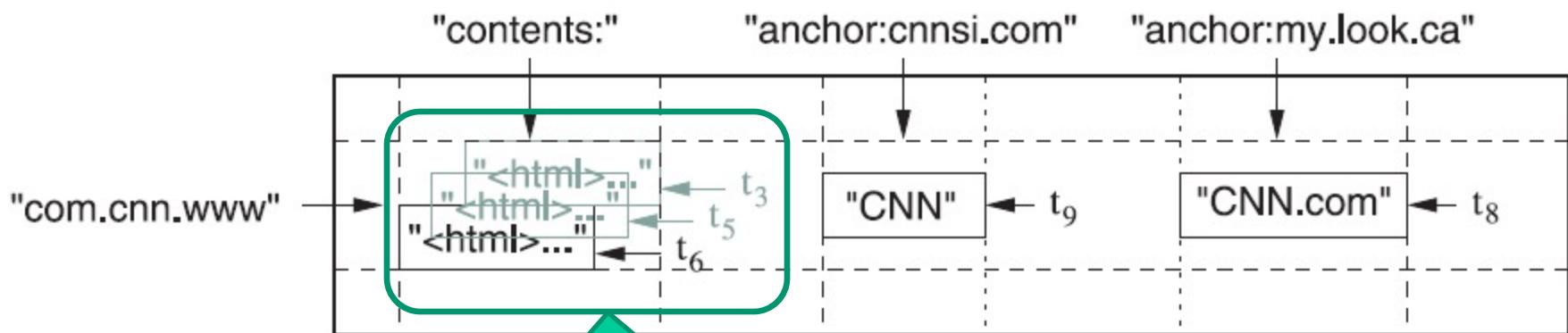
Column

Data Model – WebTable Example (6)



Cell: the storage referenced by a particular **row key**, **column key**, and **timestamp**

Data Model – WebTable Example (7)



Different cells in a table
can contain multiple
versions indexed by
timestamp





BigTable is Used on many Real Applications

Project name	Table size (TB)	Compression ratio	# Cells (billions)	# Column Families	# Locality Groups	% in memory	Latency-sensitive?
<i>Crawl</i>	800	11%	1000	16	8	0%	No
<i>Crawl</i>	50	33%	200	2	2	0%	No
<i>Google Analytics</i>	20	29%	10	1	1	0%	Yes
<i>Google Analytics</i>	200	14%	80	1	1	0%	Yes
<i>Google Base</i>	2	31%	10	29	3	15%	Yes
<i>Google Earth</i>	0.5	64%	8	7	2	33%	Yes
<i>Google Earth</i>	70	–	9	8	3	0%	No
<i>Orkut</i>	9	–	0.9	8	5	1%	Yes
<i>Personalized Search</i>	4	47%	6	93	11	5%	Yes

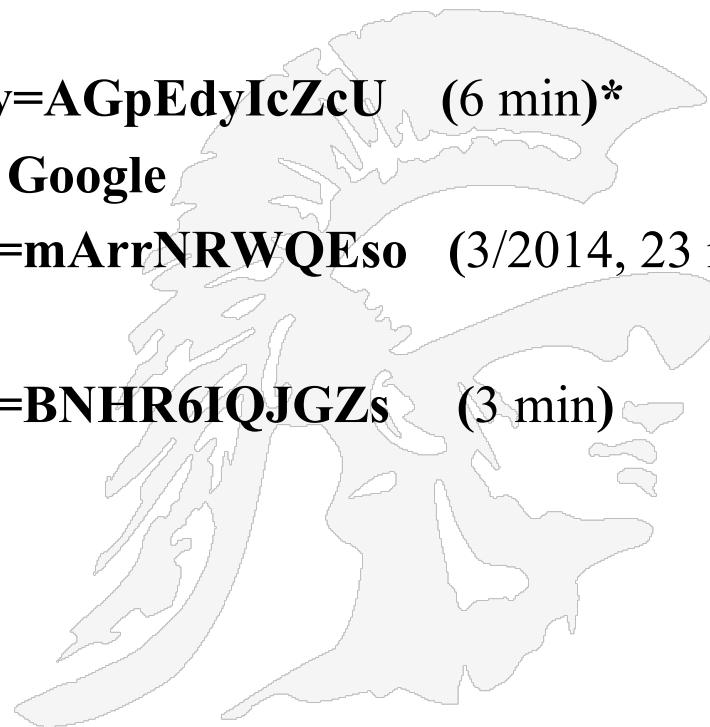


References

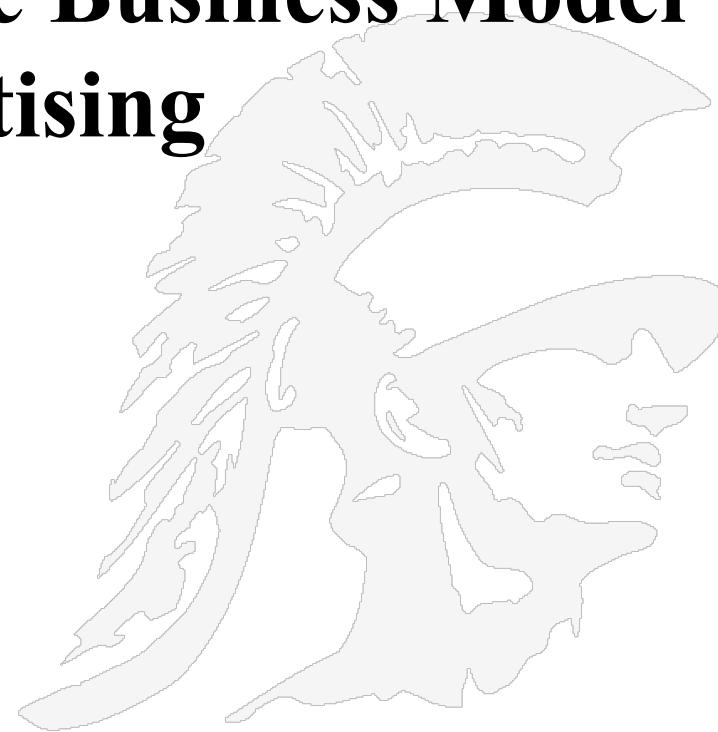
- **Google Videos on map/reduce**
<https://www.youtube.com/watch?v=yjPBkvYh-ss> (Lecture 1, 46 min)
<https://www.youtube.com/watch?v=-vD6PUdf3Js> (Lecture 2, 52 min)
- **Wikipedia**, <http://en.wikipedia.org/wiki/MapReduce>
- *Data-Intensive Text Processing with MapReduce*, Jimmy Lin and Chris Dyer, Morgan & Claypool Synthesis Lectures on Human Language Technologies, 2010
<http://www.umiacs.umd.edu/~jimmylin/MapReduce-book-final.pdf>
- **Hadoop** is an open source implementation of MapReduce
<http://hadoop.apache.org/>
- **MapReduce: Simplified Data Processing on Large Clusters**, by Jeffrey Dean and Sanjay Ghemawat, <http://research.google.com/archive/mapreduce.html>

Other Useful Videos

- **How does Google search work?**
 - <https://www.youtube.com/watch?v=KyCYyoGusqs> (7 min)
- **How does Google decide when to display multiple results from the same website**
 - <https://www.youtube.com/watch?v=AGpEdyIcZcU> (6 min)*
- **Larry Page of the future directions of Google**
 - <http://www.youtube.com/watch?v=mArrNRWQEso> (3/2014, 23 min)
- **How Search Works by Matt Cutts**
 - <http://www.youtube.com/watch?v=BNHR6IQJGZs> (3 min)

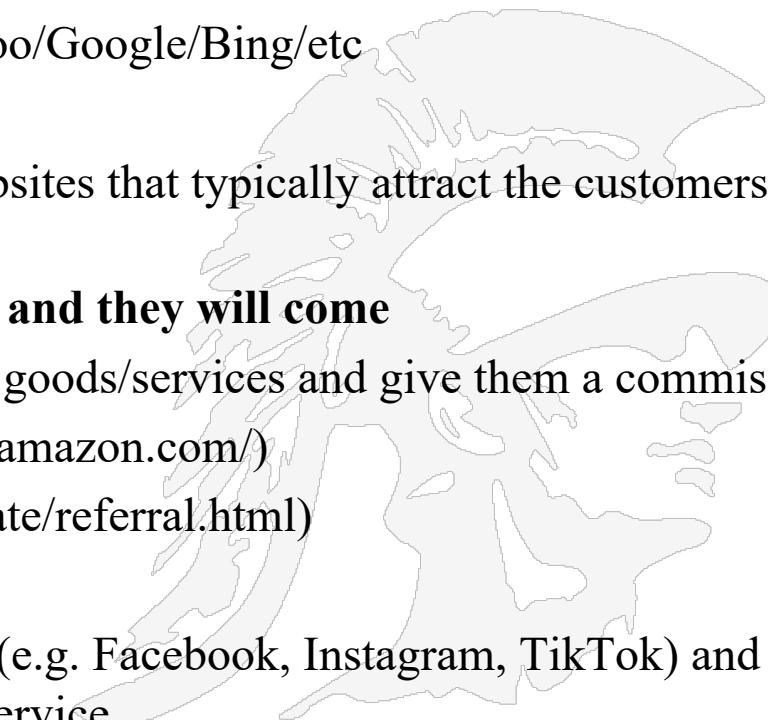


The Search Engine Business Model Advertising



Types of Online Advertising

- **Banner Advertising**
 - The earliest form of online ads
 - People have developed “banner blindness”
- **Pay-per-click Advertising**
 - Introduced by search engines Yahoo/Google/Bing/etc
- **Website Advertising**
 - Place ads on blogs/newsletters/websites that typically attract the customers you are aiming at
- **Affiliate Marketing/build a platform and they will come**
 - Let third party sites advertise your goods/services and give them a commission
 - Amazon (<https://affiliate-program.amazon.com/>)
 - Ebay (<http://pages.ebay.com/affiliate/referral.html>)
- **Social Media Marketing**
 - The use of social media platforms (e.g. Facebook, Instagram, TikTok) and websites to promote a product or service



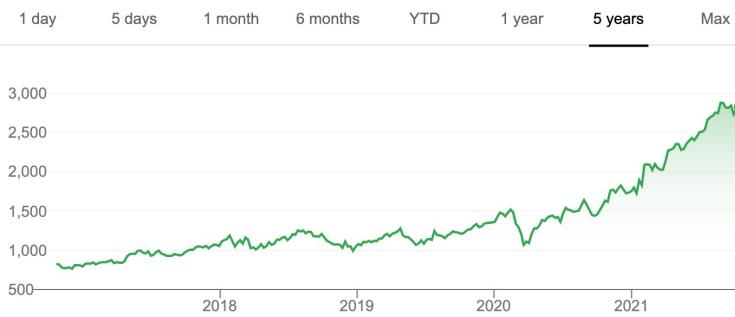
How Big is Search Engine Advertising

Market Summary > Alphabet Inc Class A

NASDAQ: GOOGL

2,827.36 USD +2,003.30 (243.10%) ↑ past 5 years

Closed: Oct 15, 6:58 PM EDT · Disclaimer
After hours 2,827.99 +0.63 (0.02%)

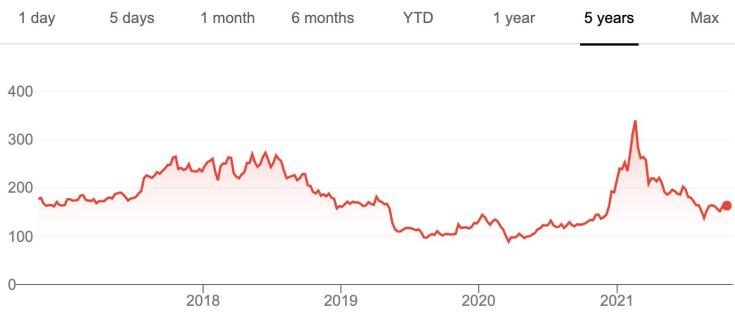


Market Summary > Baidu Inc

NASDAQ: BIDU

163.63 USD -13.13 (-7.43%) ↓ past 5 years

Closed: Oct 15, 7:59 PM EDT · Disclaimer
After hours 163.64 +0.0100 (0.0061%)



Google/Alphabet earned more than \$257 billion in 2021

Alibaba earned \$109 billion in 2021

Yahoo earned \$4.6 billion in 2019

(Verizon has purchased Yahoo for \$4.5 billion in 2015)

Bing, a division of Microsoft earned \$7.7 billion in 2021

Baidu earned \$5.9 billion in 2021

Google and Baidu Stock History



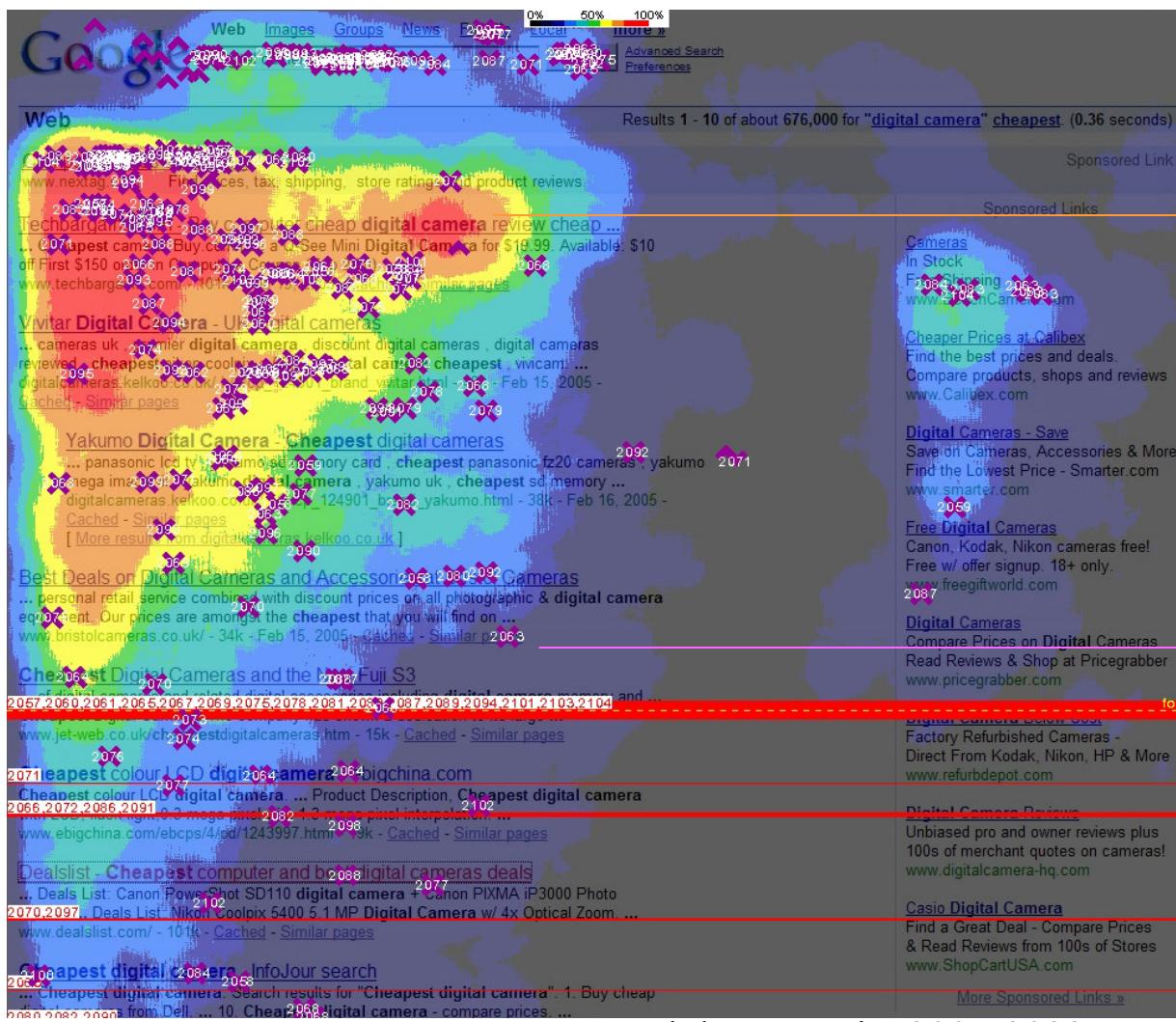
Positional Bias is Very Strong

A google result page showing where people looked and where people clicked

Eye pupil hotspots

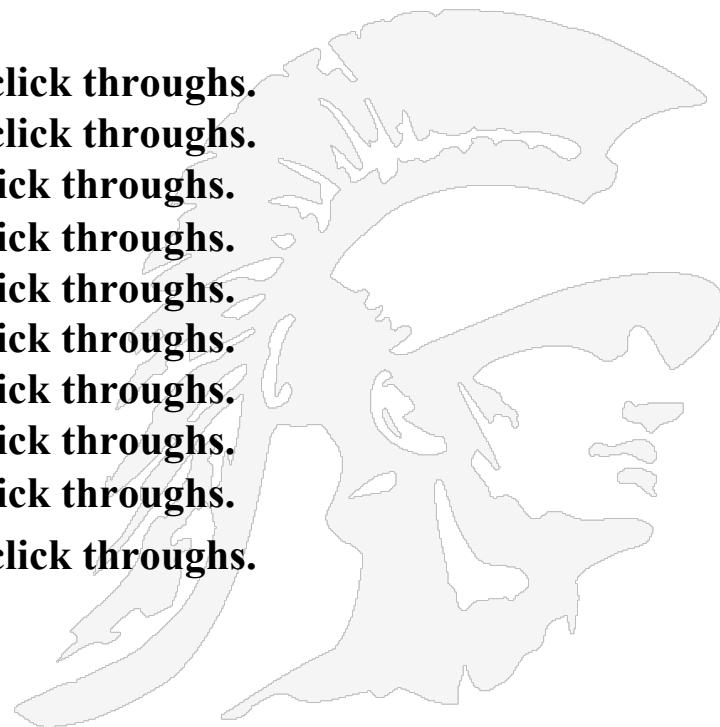
See
<http://www.entertop.net/>

X marks
the clicks



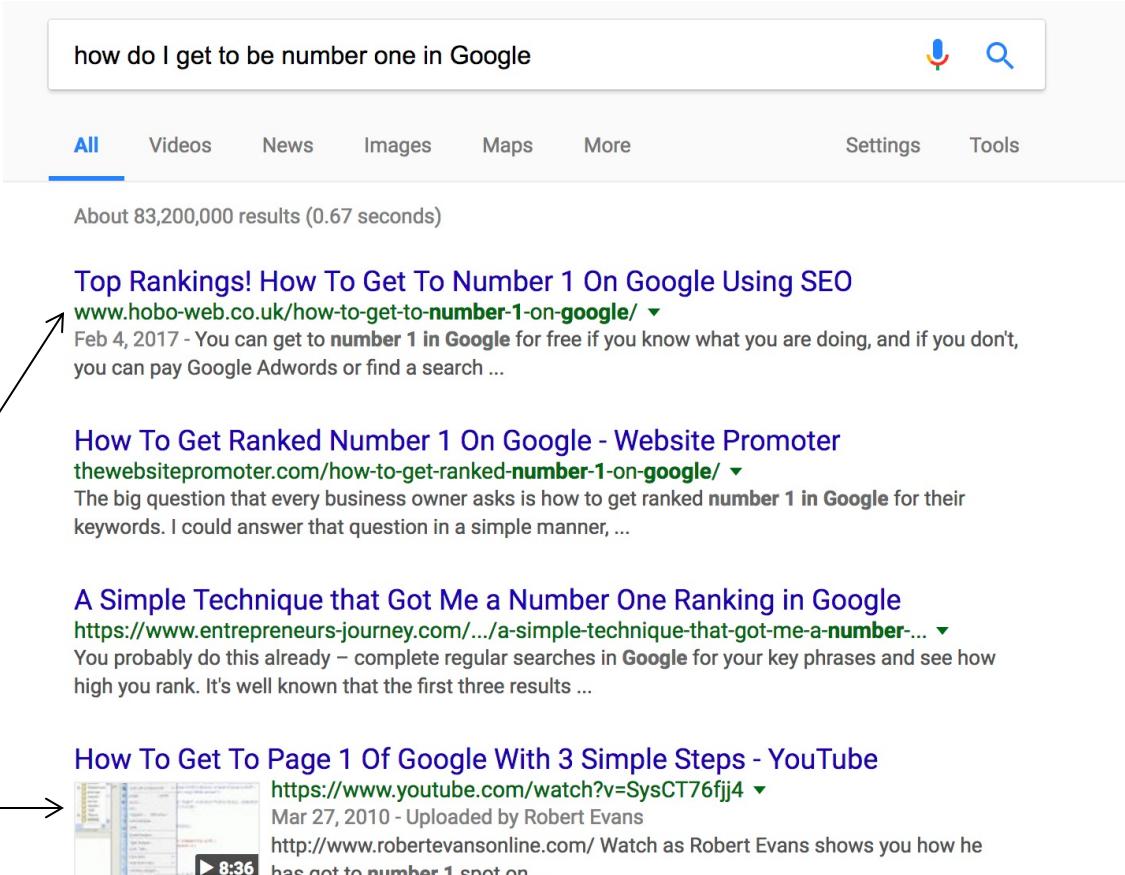
Estimating the Value of Rank in a Sponsored List

- A survey worked by examining click data from AOL, HitWise and Overture yielding statistics about how often a user clicked on the first search result, second search result, etc.
- Total Searches: 9,038,794
- Total Clicks: 4,926,623
- Ranking Number 1 receives 42.1 percent of click throughs.
Ranking Number 2 receives 11.9 percent of click throughs.
Ranking Number 3 receives 8.5 percent of click throughs.
Ranking Number 4 receives 6.1 percent of click throughs.
Ranking Number 5 receives 4.9 percent of click throughs.
Ranking Number 6 receives 4.1 percent of click throughs.
Ranking Number 7 receives 3.4 percent of click throughs.
Ranking Number 8 receives 3.0 percent of click throughs.
Ranking Number 9 receives 2.8 percent of click throughs.
- Ranking Number 10 receives 3.0 percent of click throughs.



THE MILLION DOLLAR QUESTION

Q. How do I get to number one in Google search results?



how do I get to be number one in Google

All Videos News Images Maps More Settings Tools

About 83,200,000 results (0.67 seconds)

Top Rankings! How To Get To Number 1 On Google Using SEO
www.hobo-web.co.uk/how-to-get-to-number-1-on-google/ ▾
Feb 4, 2017 - You can get to number 1 in Google for free if you know what you are doing, and if you don't, you can pay Google Adwords or find a search ...

How To Get Ranked Number 1 On Google - Website Promoter
thewebsitepromoter.com/how-to-get-ranked-number-1-on-google/ ▾
The big question that every business owner asks is how to get ranked number 1 in Google for their keywords. I could answer that question in a simple manner, ...

A Simple Technique that Got Me a Number One Ranking in Google
<https://www.entrepreneurs-journey.com/.../a-simple-technique-that-got-me-a-number-...> ▾
You probably do this already – complete regular searches in Google for your key phrases and see how high you rank. It's well known that the first three results ...

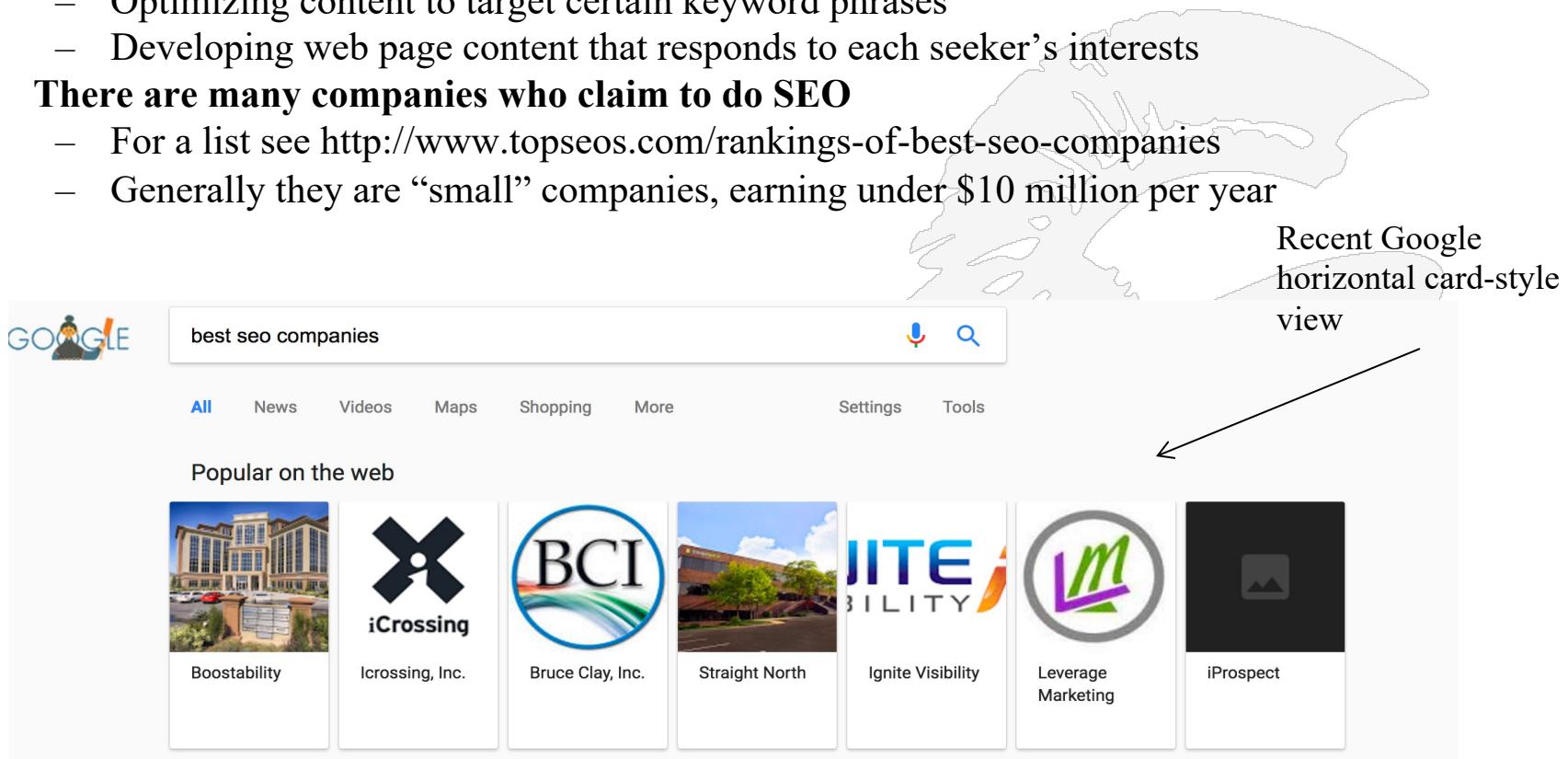
How To Get To Page 1 Of Google With 3 Simple Steps - YouTube
<https://www.youtube.com/watch?v=SysCT76fjj4> ▾
Mar 27, 2010 - Uploaded by Robert Evans
http://www.robertevansonline.com/ Watch as Robert Evans shows you how he has got to number 1 spot on ...

Google's organic search results are produced by a proprietary algorithm that is often being changed, so there are never any guarantees that someone will continue to be highly ranked

But there are lots of people ready to give you advice

Search Engine Optimization

- There is an industry of companies whose focus is to develop and refine a company's online presence;
 - They are called Search Engine Optimizers
- **Search Engine Optimization (SEO) involves:**
 - Making pages show up higher in search engine's organic results
 - Optimizing content to target certain keyword phrases
 - Developing web page content that responds to each seeker's interests
- **There are many companies who claim to do SEO**
 - For a list see <http://www.topseos.com/rankings-of-best-seo-companies>
 - Generally they are “small” companies, earning under \$10 million per year



A screenshot of a Google search results page for the query "best seo companies". The results are displayed in a horizontal card-style view, which is highlighted by a callout arrow pointing from the text "Recent Google horizontal card-style view" to the right side of the search results. The cards feature logos and names of various SEO companies: Boostability, iCrossing, Bruce Clay, Inc., Straight North, Ignite Visibility, Leverage Marketing, and iProspect.

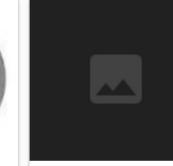
Recent Google horizontal card-style view

GOOGLE

best seo companies

All News Videos Maps Shopping More Settings Tools

Popular on the web

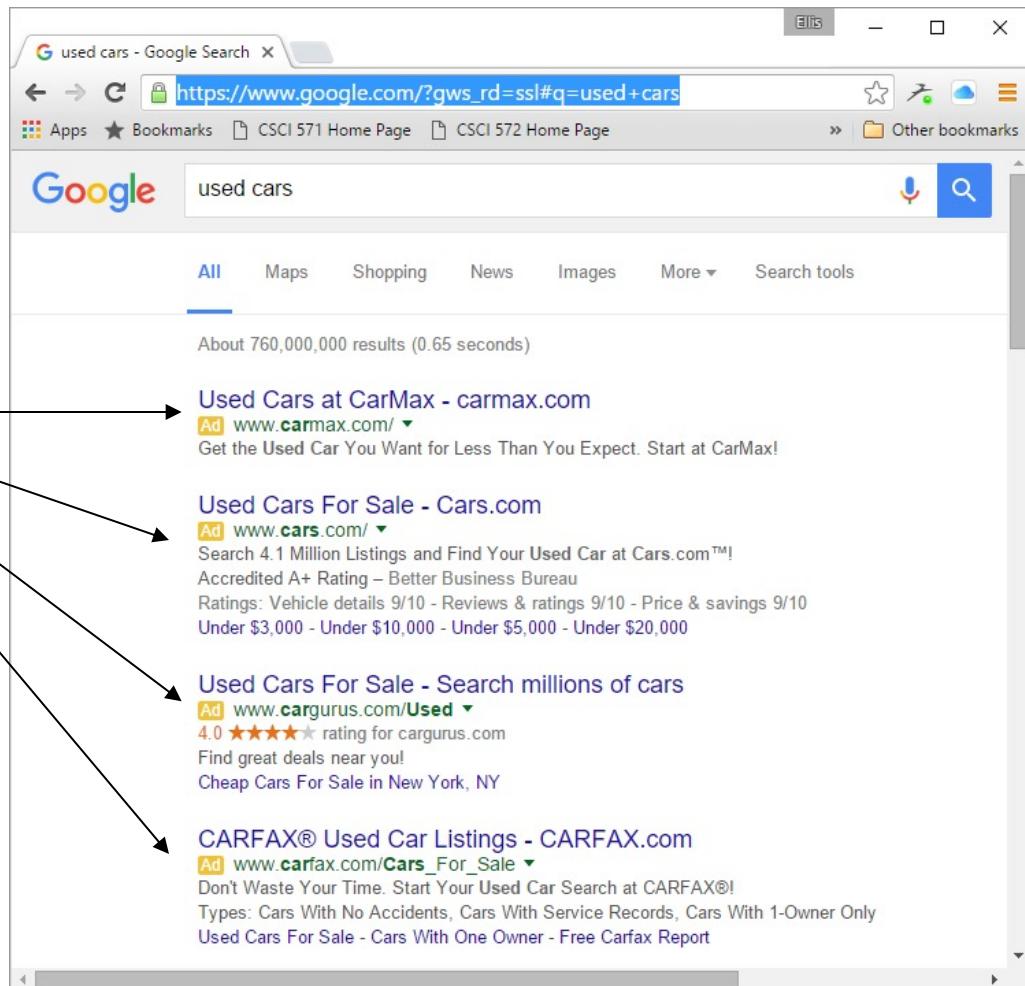
 Boostability	 iCrossing, Inc.	 Bruce Clay, Inc.	 Straight North	 Ignite Visibility	 Leverage Marketing	 iProspect
---	--	---	--	--	---	--

THE MILLION DOLLAR ANSWER

Q. How do I get to number one in Google?
A. Use paid search

Paid search is one answer; Google and many other search engines use the pay-per-click (PPC) model of advertising;

Pay enough for a click and you can pretty much guarantee to get a top spot on the list of ads for your chosen keywords



The screenshot shows a Google search results page for the query "used cars". The results are filtered to show only paid ads (Ads). There are four visible ads:

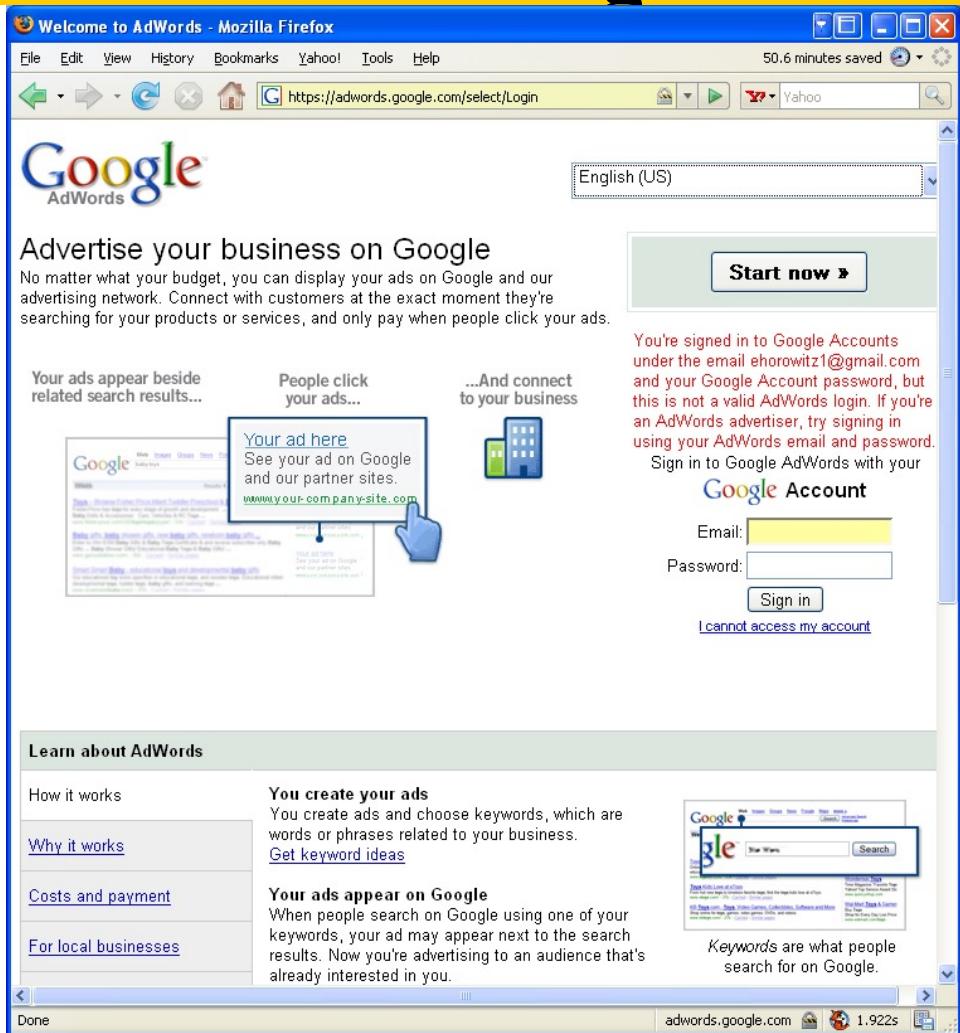
- Used Cars at CarMax - carmax.com**
Ad www.carmax.com/
Get the Used Car You Want for Less Than You Expect. Start at CarMax!
- Used Cars For Sale - Cars.com**
Ad www.cars.com/
Search 4.1 Million Listings and Find Your Used Car at Cars.com™!
Accredited A+ Rating – Better Business Bureau
Ratings: Vehicle details 9/10 - Reviews & ratings 9/10 - Price & savings 9/10
Under \$3,000 - Under \$10,000 - Under \$5,000 - Under \$20,000
- Used Cars For Sale - Search millions of cars**
Ad www.cargurus.com/Used
4.0 ★★★★☆ rating for cargurus.com
Find great deals near you!
Cheap Cars For Sale in New York, NY
- CARFAX® Used Car Listings - CARFAX.com**
Ad www.carfax.com/Cars_For_Sale
Don't Waste Your Time. Start Your Used Car Search at CARFAX®!
Types: Cars With No Accidents, Cars With Service Records, Cars With 1-Owner Only
Used Cars For Sale - Cars With One Owner - Free Carfax Report

Renamed as Google Ads
 Google's program for accepting pay-per-click ads.

Its home page is
<https://ads.google.com>

For every keyword phrase there is an auction where bidders agree to pay a certain amount to Google if their ad is clicked on;

Lets Take a Close Look at Google Adwords



The screenshot shows the Google AdWords login page. The browser title bar says "Welcome to AdWords - Mozilla Firefox". The address bar shows "https://adwords.google.com/select/Login". The main content area features the Google logo and the heading "Advertise your business on Google". It explains that users can display ads on Google and its network, connect with customers, and only pay when ads are clicked. A central diagram illustrates the process: "Your ads appear beside related search results..." leads to "People click your ads...", which then connects to "...And connect to your business". Below this, a "Start now" button is visible. To the right, a message states that the user is signed in to Google Accounts under a specific email, but it's not a valid AdWords login. It encourages the user to sign in using their AdWords email and password. There are fields for "Email" and "Password", a "Sign in" button, and a link for "I cannot access my account". At the bottom, there's a "Learn about AdWords" section with links for "How it works", "Why it works", "Costs and payment", and "For local businesses". The right side also shows a snippet of a Google search results page with a highlighted "gle" entry.

Campaigns Begin By Choosing Keywords

Google Keyword Estimator

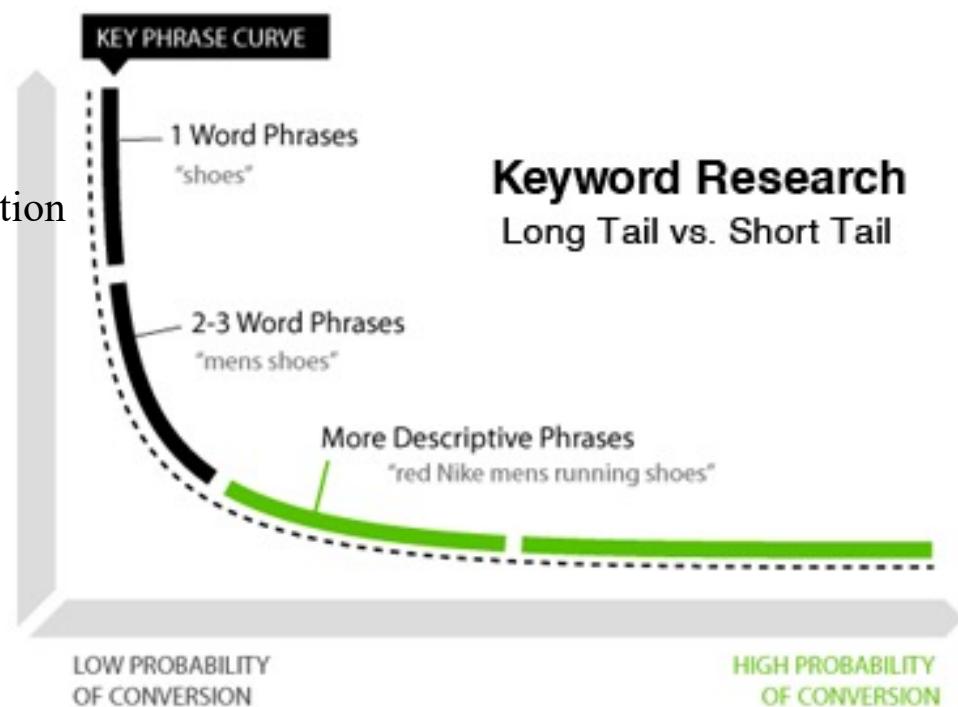
For specific keywords Google provides the approximate cost per click and the resulting position in the list, and other stats

Maximum CPC:	Daily budget:	Get New Estimates			
Keywords ▾	Search Volume	Estimated Avg. CPC	Estimated Ad Positions	Estimated Clicks / Day	Estimated Cost / Day
Search Total		\$5.23 - \$7.34	1 - 3	270 - 348	\$1,450 - \$2,660
crime scene detective class	<div style="width: 10%;">10%</div>		Not enough data to give estimates. [?]		
criminal justice	<div style="width: 20%;">20%</div>	\$4.85 - \$6.84	1 - 3	221 - 277	\$1,080 - \$1,900
criminal justice associates degree	<div style="width: 5%;">5%</div>	\$9.58 - \$12.76	1 - 3	0 - 1	\$0 - \$20
criminal justice career	<div style="width: 5%;">5%</div>	\$3.53 - \$4.41	1 - 3	2 - 3	\$8 - \$20
criminal justice career training	<div style="width: 10%;">10%</div>		Not enough data to give estimates. [?]		
criminal justice classes	<div style="width: 5%;">5%</div>	\$8.72 - \$10.90	1 - 3	0 - 1	\$0 - \$20
criminal justice college	<div style="width: 10%;">10%</div>	\$4.76 - \$5.95	1 - 3	10 - 14	\$50 - \$90
criminal justice course	<div style="width: 5%;">5%</div>	\$10.63 - \$15.94	1 - 3	0 - 1	\$0 - \$20
criminal justice coursework	<div style="width: 10%;">10%</div>		Not enough data to give estimates. [?]		
criminal justice degree	<div style="width: 10%;">10%</div>	\$13.76 - \$20.64	1 - 3	16 - 21	\$230 - \$440
criminal justice online class	<div style="width: 10%;">10%</div>		Not enough data to give estimates. [?]		
criminal justice program	<div style="width: 10%;">10%</div>	\$7.34 - \$9.92	1 - 3	2 - 4	\$20 - \$40
criminal justice school	<div style="width: 10%;">10%</div>	\$7.37 - \$9.31	1 - 3	4 - 6	\$30 - \$60
detective training	<div style="width: 10%;">10%</div>	\$3.58 - \$4.48	1 - 3	2 - 3	\$8 - \$20
law enforcement career	<div style="width: 10%;">10%</div>	\$4.40 - \$5.49	1 - 3	3 - 4	\$20 - \$30
law enforcement program	<div style="width: 10%;">10%</div>	\$3.29 - \$4.12	1 - 3	0 - 1	\$0 - \$5
law enforcement training	<div style="width: 10%;">10%</div>	\$3.44 - \$4.30	1 - 3	10 - 12	\$40 - \$80

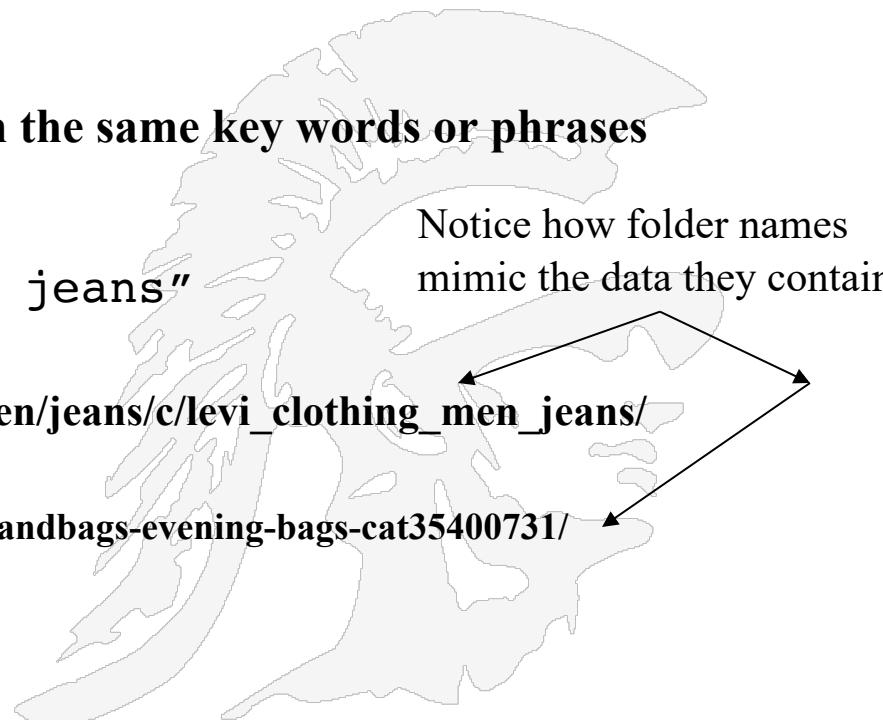
Estimates for these keywords are based on clickthrough rates for current advertisers. Some of the keywords above are subject to review by Google and may not trigger your ads until they are approved. Please note that your traffic estimates assume your keywords are approved.

Long Tailed Keywords can be Valuable

- Long-tail keywords are search queries made up of three-four word phrases that are very specific to a product, good, or service that's being sold.
 - Long-tail keywords are the phrases search engine users are generally more likely to type in when they're closer to purchasing an item.
- The conversion rate for long-tail keywords is approximately **2.5 times higher than it is for head (shorter) keywords.**
- long-tail keywords that present *less* competition also offer *lower cost-per-click prices* since few marketers are targeting them



Your Keyword Phrases Should be Mapped on Your Website

- Use key phrases in the content on your page
- Develop metadata that includes key phrases in:
 - TITLE tags
 - Meta Description and Keyword tags
 - ALT tags
- Name directories, files and images with the same key words or phrases
- Query: "Levi's men's clothing jeans" produces the URL
https://www.levi.com/US/en_US/clothing/men/jeans/c/levi_clothing_men_jeans/
- Query: "Ladies handbags" 
<https://www.neimanmarcus.com/c/handbags-all-handbags-evening-bags-cat35400731/>

Notice how folder names mimic the data they contain

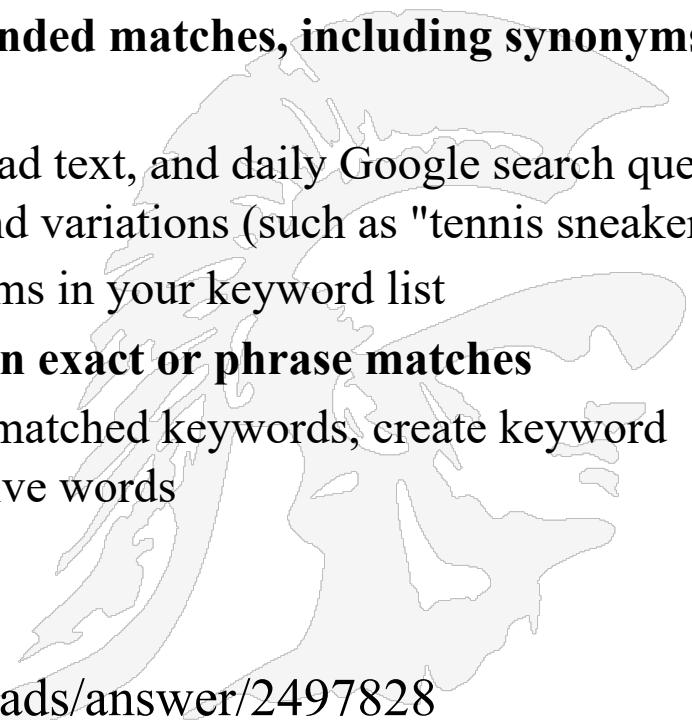
Advertisers Designate Keyword Matching Rules

- The *advertiser* will specify the type of matching to be done against his keyword phrases
 - Four types of keyword matching options can help you refine your ad targeting
 1. Broad Match
 2. Exact Match
 3. Phrase Match
 4. Negative Keyword



Keyword: Broad Match

- **A broad match is the default option**
 - When you include keyword phrases like "tennis shoes" in your keyword list, your ads will appear when users search for "tennis" and "shoes", in any order (and possibly along with other terms)
- **Broad Match ads may also show for expanded matches, including synonyms and plurals**
 - Google will analyze your keyword list, ad text, and daily Google search queries, and show your ads for relevant terms and variations (such as "tennis sneakers")
 - Even if you didn't include these terms in your keyword list
- **Broad matches are often less targeted than exact or phrase matches**
 - If you decide to run your ads on broad-matched keywords, create keyword phrases containing at least two descriptive words



<https://support.google.com/google-ads/answer/2497828>

Keyword: Exact Match

- The search query must exactly match your keyword
 - Originally Exact Match meant that "tennis shoes" would only match a user request for "tennis shoes" and not for "red tennis shoes," even though the second query contains your keywords
 - However, Google now includes rewording and reordering for exact match keywords
 - Exact match will now ignore function words (in, to), conjunctions (for, but), articles (a, the) and other words that don't impact the intent of the query
 - <https://support.google.com/google-ads/answer/7478529>

Keyword	Query	Why it matched
jobs in united states nike shoes women	jobs in the united states nike shoes for women	Function words added
parks in san diego paint for a deck	parks san diego paint for deck	Function words removed
news from today bahamas cruise from miami	news for today miami to bahamas cruise	Function words changed

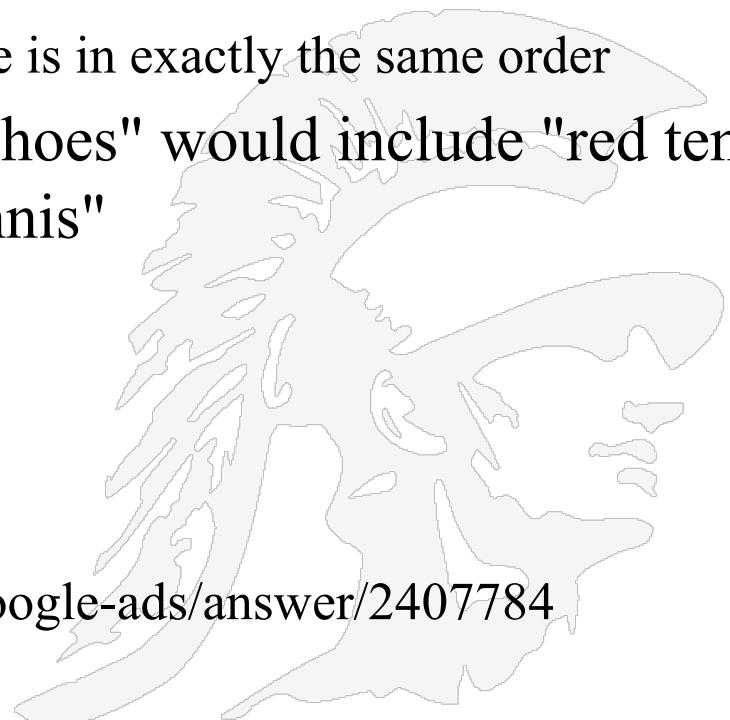


Keyword	Query	Why it matched
running shoes men's dress shirt	shoes running dress shirt men's	Words reordered

<https://support.google.com/google-ads/answer/2497825>

Keyword: Phrase Match

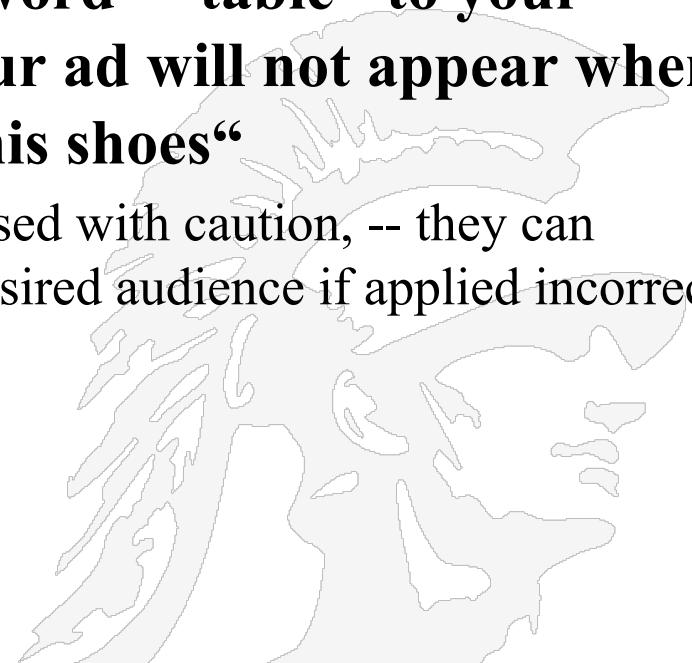
- Your ad appears when users search on the exact phrase
 - AND when their search contains additional terms
 - As long as the keyword phrase is in exactly the same order
 - A phrase match for "tennis shoes" would include "red tennis shoes" but not "shoes for tennis"



<https://support.google.com/google-ads/answer/2407784>

Keyword: Negative Keyword

- **Negative keywords allow you to eliminate searches that you know are not related to your message**
 - If you add the negative keyword "**–table**" to your keyword "**tennis shoes,**" your ad will not appear when a user searches on "**table tennis shoes**"
 - Negative keywords should be used with caution, -- they can eliminate a large portion of a desired audience if applied incorrectly



<https://support.google.com/google-ads/answer/2453972>

A Sample Google Ads Campaign Screen



Google AdWords: Campaign Summary - Mozilla Firefox

All Campaigns

+ Create a new campaign : keyword-targeted | site-targeted

Jan 29, 2003 to Feb 7, 2007

Campaign Name	Current Status	Current Budget [?]	Clicks ▾	Impr.	CTR	Avg. CPC	Cost
Campaign #2	Deleted	[\$10.00 / day]	747	87,551	0.85%	\$0.44	\$326.34
Campaign #1	Deleted	[\$5.00 / day]	328	9,893	3.31%	\$0.53	\$174.94
Total - all 2 campaigns	-	\$0.00 / day active campaigns	1,075	97,444	1.10%	\$0.47	\$501.28

Find: money

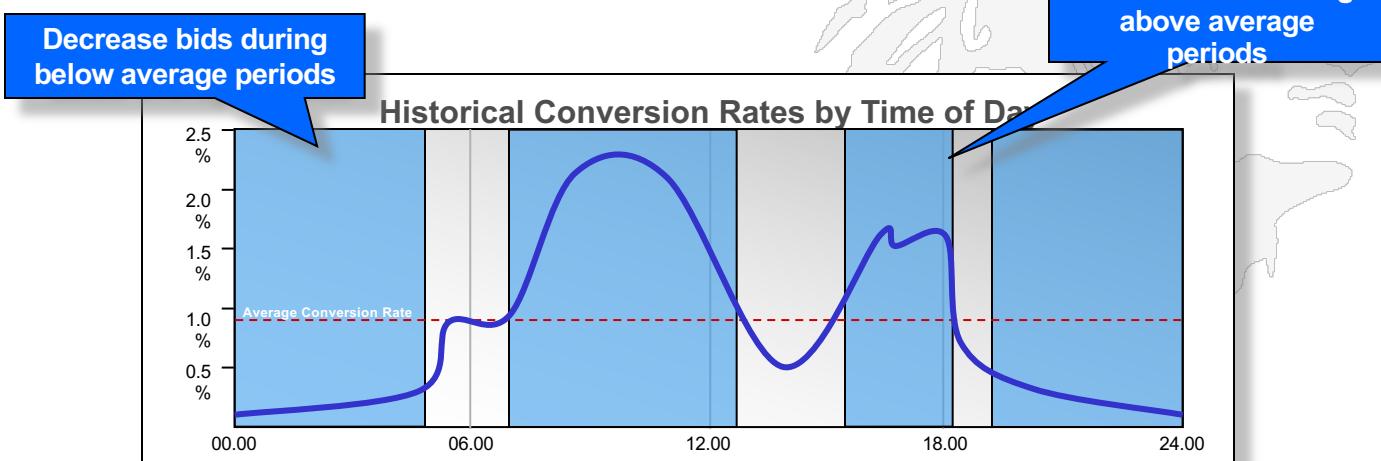
- **Ads advertisers create advertising campaigns**
 - There are two campaigns defined above
- **Each campaign has a set of keyword and associated ads, and includes**
 - a budget, recorded clicks, recorded impressions, click thru rate, average cost/click, and total cost of the campaign

Capabilities of Search Engine Ad Servers

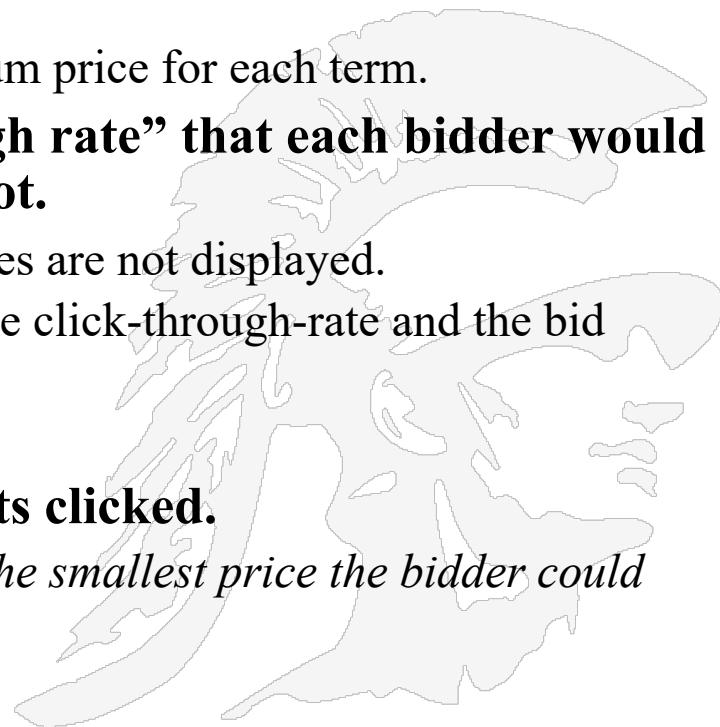
- **The typical common functionality of ad servers includes:**
 - Uploading the creative (*creative* is the term used to describe the ad that will appear)
 - Maintaining business rules for placing ads
 - Targeting ads to different users or content
 - Optimizing appearance of a set of creatives based upon results (choosing the most effective ad)
 - Reporting impressions, clicks, post-click activities, and interaction metrics.
- **Advanced functionality may include:**
 - *Frequency capping* creatives so users only see messages a limited amount of time.
 - *Sequencing* creatives so users see messages in a specific order (sometimes known as surround sessions).
 - Excluding competitive creatives so users do not see competitors' ads directly next to one another.
 - Displaying creatives so an advertiser can own 100% of the inventory on a page (sometimes known as roadblocks).
 - Targeting creatives to users based on their previous behavior (behavioral marketing or behavioral targeting).

Search Engine Ad Servers are Versatile

- Advertisers with accounts on Google's Ads can define a set of criteria for placement of their ad; typical criteria might include rules such as
 - Only display my ad
 - from 9:00AM-5:00PM EST
 - once/day
 - if the viewer is located in the United States
 - dayparting* is a technique that involves increasing your bids during times when conversion rates are typically above average, and decreasing them when rates are typically below average



- **Each bidder specifies (i) search terms that trigger its bid and (ii) the amount to bid for each search term.**
 - Bidders may also establish an overall ad budget and limits for each kind of bid.
 - Google may set a reserve or minimum price for each term.
- **Google estimates the “click-through rate” that each bidder would have if it were listed in the first spot.**
 - Ads with very low click-through rates are not displayed.
 - Google ranks bids by multiplying the click-through-rate and the bid amount.
 - Ads are displayed in rank order.
- **Google is paid only when an ad gets clicked.**
 - In that case, the price it receives is *the smallest price the bidder could have bid to get its ranking*.

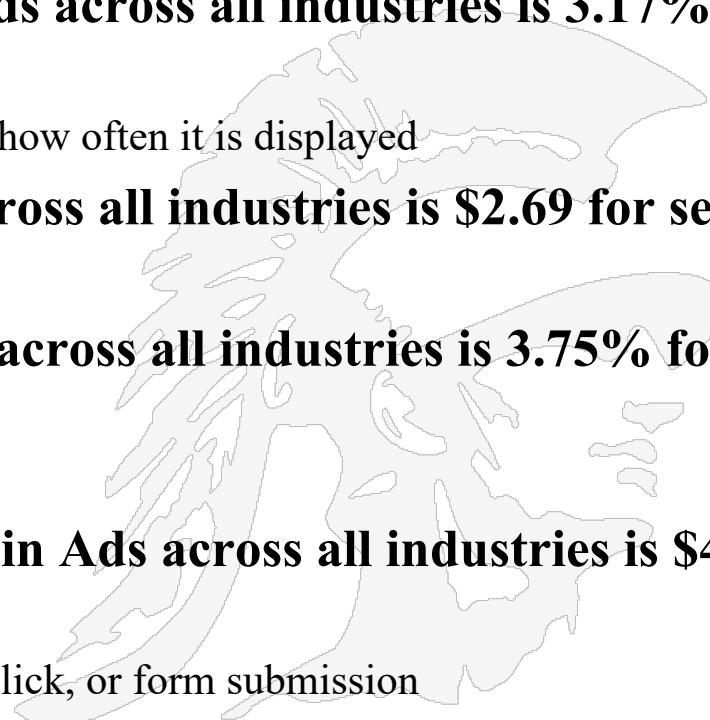


How does AdWords bidding work?

- The actual position of your ad is determined by your ad rank (Maximum Bid multiplied by the Quality Score).
- The highest ad rank gets the 1st ad position.
- Your actual CPC (cost per click) will be determined by the ad rank of the next highest ad below you
- **Exception:** when you are the only bidder or the lowest bid in the Ads auction; then you pay your maximum bid per click!
- Ads bidding heavily penalizes advertisers who bid with low quality scores. Conversely, those with high Quality Scores get higher ad ranks and lower CPC
- The average cost per click on Ads varies by keyword and industry, but is roughly \$2.00 on the search network and \$0.58 on the display network

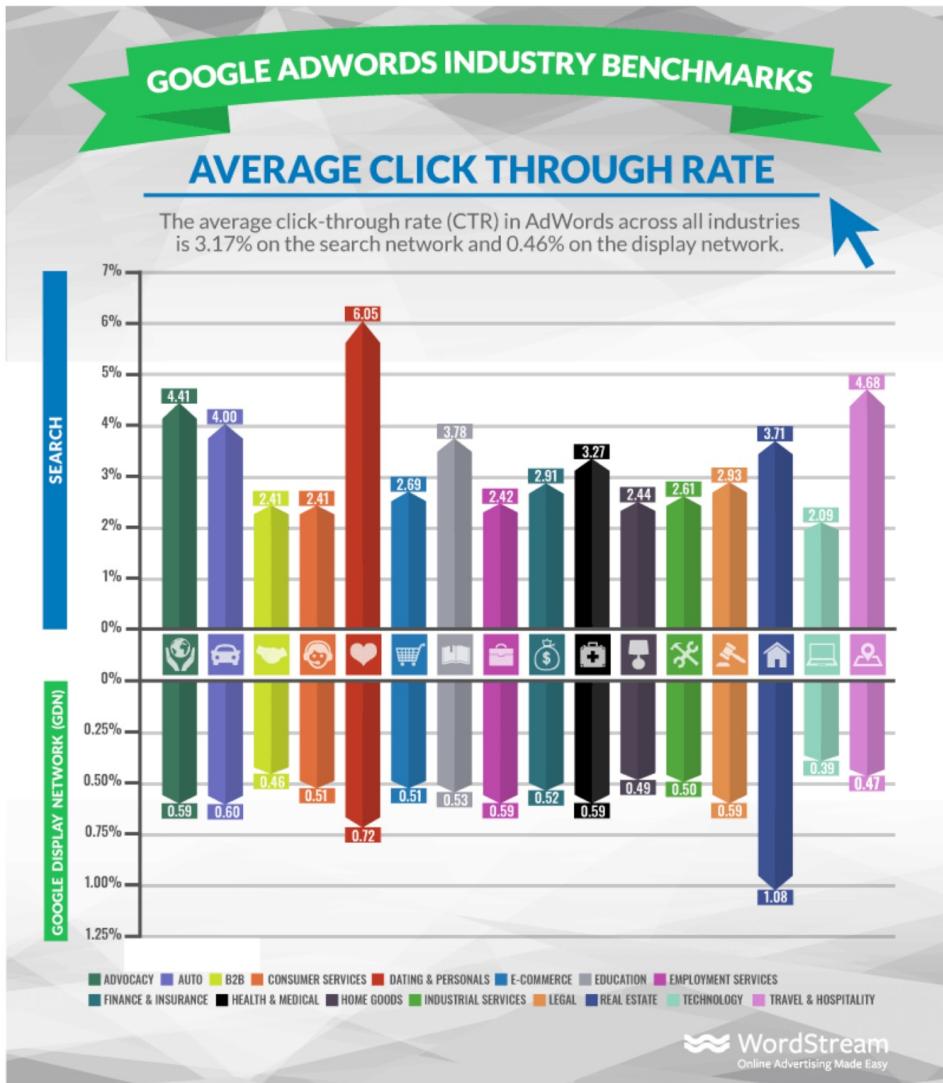
Some Facts About Online Advertising

- <https://www.wordstream.com/blog/ws/2016/02/29/google-adwords-industry-benchmarks>
- A study of over 14,000 US based clients in multiple markets
- The average *click-through rate* in Ads across all industries is 3.17% for search and 0.46% for display ads
 - How often people click on your ad versus how often it is displayed
- The average *cost per click* in Ads across all industries is \$2.69 for search and \$0.63 for display
- The average *conversion rate* in Ads across all industries is 3.75% for search and 0.77% for display
 - How often a click on an ad leads to a sale
- The average *Cost Per Action (CPA)* in Ads across all industries is \$48.96 for search and \$75.51 for display
 - Refers to a specific action such as a sale, click, or form submission



GDN refers to the Google Display Network where ads are placed on 3rd party websites

Average Click Thru Rate by Industry



Bars in order:

- Advocacy
- Auto
- B2B
- Consumer services
- Dating and personals
- E-commerce
- Education
- Employment services
- Finance & insurance
- Health & medical
- Home goods
- Industrial services
- Legal
- Real estate
- Technology
- Travel & hospitality

Second Price Auction Using Only Bid Amount

- In a **First-Price Auction** if your bid wins, you pay exactly what you bid. This maximizes revenue potential for the seller.
- A **Second-Price Auction** is a digital buying model where if your bid wins, you pay \$0.01 above the **second** highest bid in the **auction**.
 - In this type of **auction**, it is in your best interest to bid the highest amount you are willing to pay, knowing that often you will end up paying less than that amount.

Bid2: (\$9)

Bid3: (\$6)

Bid1: (\$5)

- **Bidder 2 is ranked 1st**
 - Pays $\$6+1\text{¢}=\6.01
- **Bidder 3 is ranked 2nd**
 - Pays $\$5+1\text{¢}=\5.01

Let $P(C)$ be the Probability
of the Ad Being Clicked

Bid2: \$9 $P(C)=0.1$

Bid3: \$6 $P(C)=0.1$

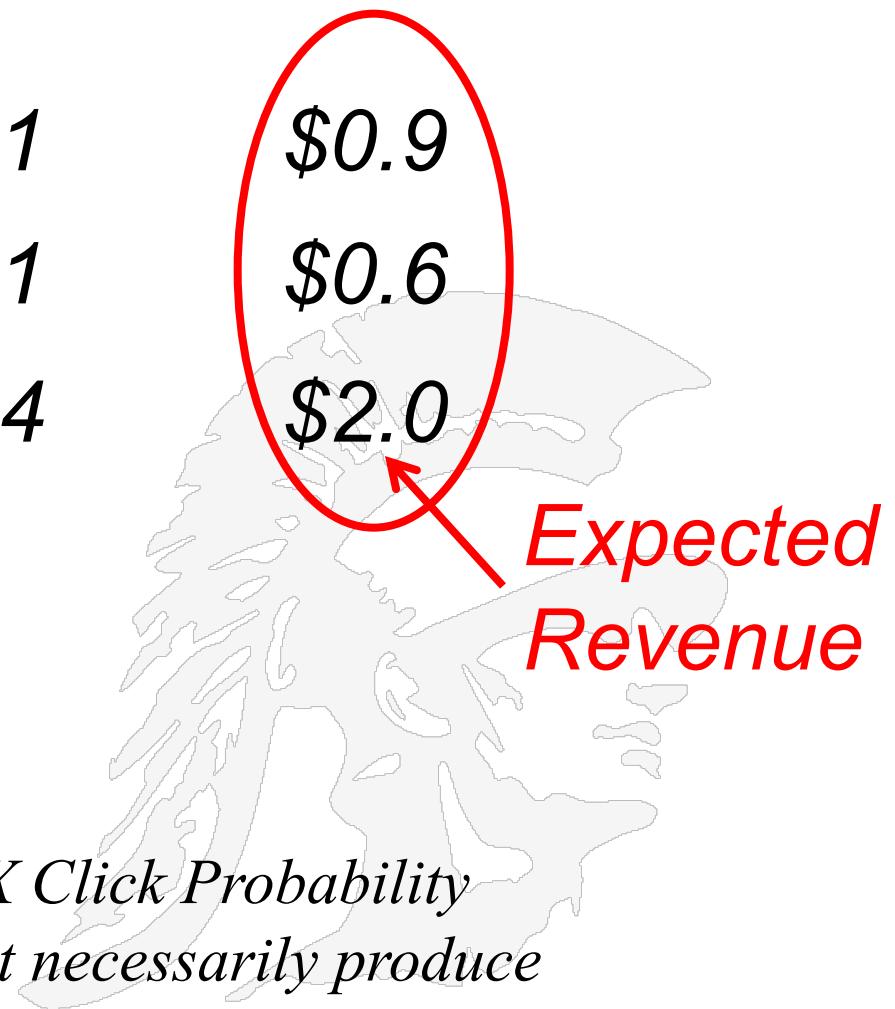
Bid1: \$5 $P(C)=0.4$

$P(c)$ can depend upon:

- Historical click performance of the ad
- Landing page quality
- Relevance to the user
- User click through rates

Expected Revenue = Bid amount X Click Probability

So the ad with highest bid does not necessarily produce the most revenue



Google Ranks Ads by the Product of Bid Amount*Click Probability

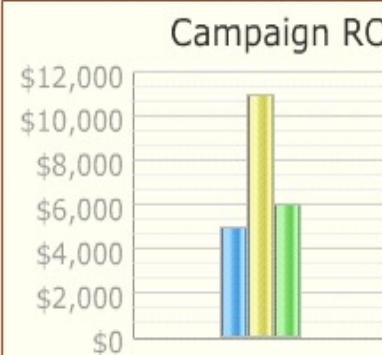
Bid1: \$5 $P(C)=0.4$ is placed 1st

Bid2: \$9 $P(C)=0.1$ is placed 2nd

Bid3: \$6 $P(C)=0.1$ is placed 3rd

Ad Rank= Bid X Click Probability

Return On Investment: What is Each Conversion Worth?



Total Campaign Cost: **\$5,000.00**

of Responders: **5000**

of Buyers: **110**

Revenue Generated: **\$11,000.00**

Profit: **\$6,000.00**

Cost per Responder: **\$1.00**

Cost per Buyer: **\$45.45**

note: double-click cells below to enter data

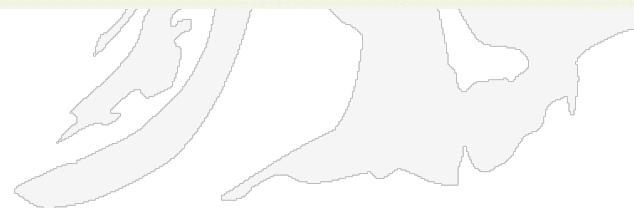
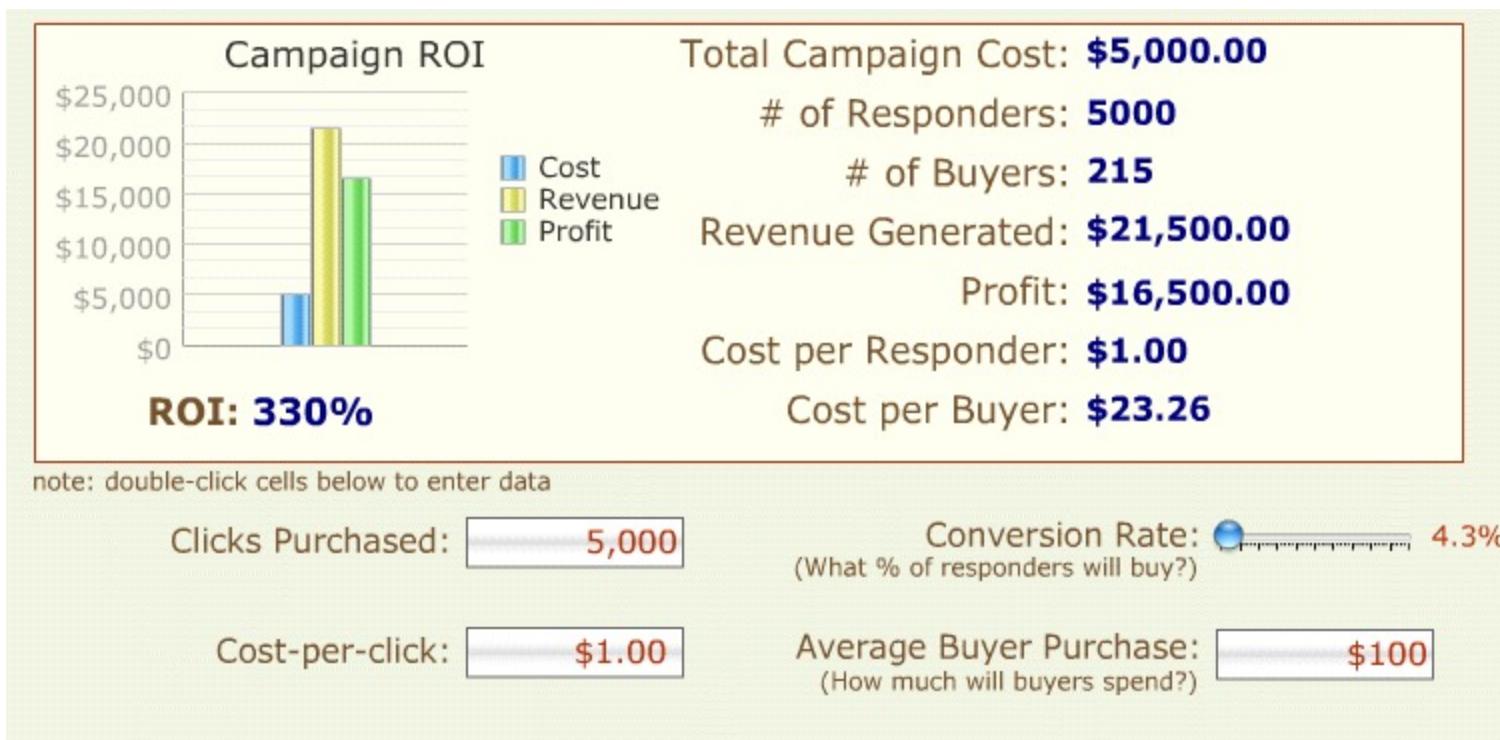
Clicks Purchased:

Conversion Rate:
(What % of responders will buy?)

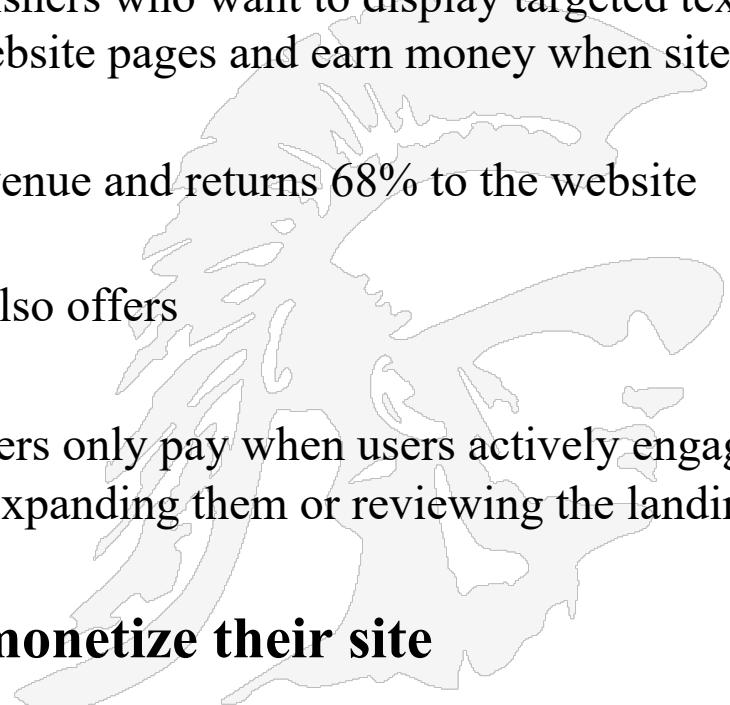
Cost-per-click:

Average Buyer Purchase:
(How much will buyers spend?)

Improved Conversion Rate

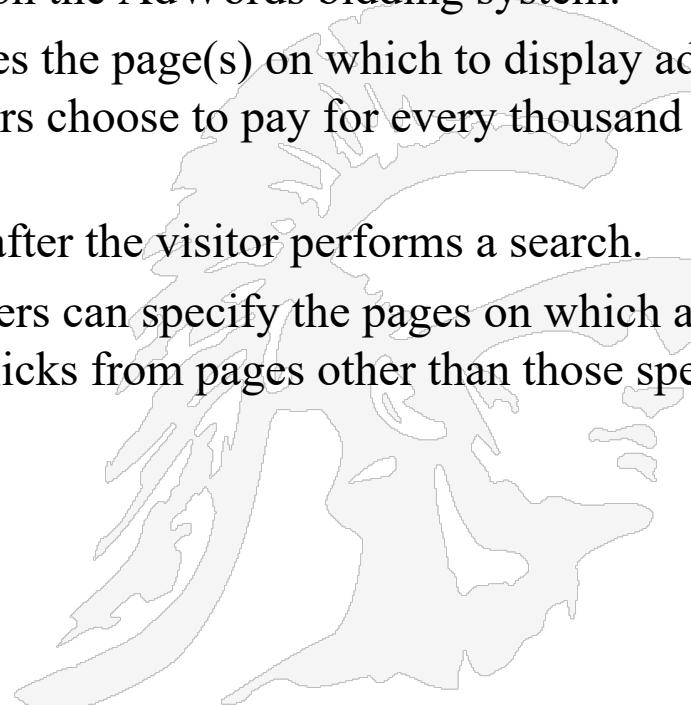


- AdSense from Google is a service for placing Google ads on third party web pages
- Google purchased the content matching technology from Oingo, a small Santa Monica-based search engine in 2003 for \$102 million
- The program is designed for website publishers who want to display targeted text, video or image advertisements on their website pages and earn money when site visitors view or click the ads
- Typically Google keeps 32% of the ad revenue and returns 68% to the website developer
- In addition to cost-per-click ads, Google also offers
 - *Cost per Thousand* displays, CPM
 - *Cost per Engagement*, where advertisers only pay when users actively engage with ads, e.g. hovering over them or expanding them or reviewing the landing page for more than a few seconds
- **Typically blogs use AdSense to monetize their site**



How AdSense Works

1. The webmaster inserts the AdSense JavaScript code into a webpage.
2. Each time this page is visited, the JavaScript code uses inlined JSON to display content fetched from Google's servers.
3. Google's servers use a cache of the page to determine a set of high-value keywords. Ads are served for those keywords based on the AdWords bidding system.
4. For site-targeted ads, the advertiser chooses the page(s) on which to display ads, and pays based on CPM, or the price advertisers choose to pay for every thousand advertisements displayed.
5. Search ads are added to the list of results after the visitor performs a search.
6. To protect against fraud, AdSense customers can specify the pages on which ads should be shown. AdSense then ignores clicks from pages other than those specified.



Google's AdSense program claims to place ads on third-party websites, where the ads are **relevant** to the site's content;

Each time a visitor visits a page with an AdSense tag, a piece of JavaScript writes an iframe tag, whose src attribute includes the URL of the page. Google's servers use a cache of the page to determine a set of high-value keywords.

Special Reports

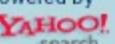
Make phone calls through your high-speed internet connection.
glingo
 The life of broadband
[SIGN UP NOW!](#)

SERVICES

[Video](#)
[E-mail Newsletters](#)
[Your E-mail Alerts](#)
[RSS](#)
[CNNtoGO](#)
[TV Commercials](#)
[Contact Us](#)

SEARCH

Web  

[Search](#) Powered by 

Changing stories and fuzzy details on specific dates are consistent with the way children who are victims of sexual abuse behave, an expert testified Monday at the Michael Jackson trial.

But under cross-examination, Anthony Urquiza acknowledged that his expertise did not extend to false allegations of sexual abuse, which he had not studied.

The testimony came as prosecutors in Jackson's child molestation trial tried to shore up earlier contradictory statements made by the pop star's accuser.

Jackson arrived late again Monday, walking with assistance into the Santa Maria, California, courthouse. Jackson offered a weak wave to supporters as he arrived, then made his way inside.

Santa Barbara County Superior Court Judge Rodney Melville, who threatened to jail Jackson after a similar episode on

AdSense Ads on Websites are Not Always Relevant



Michael Jackson is escorted out of the courtroom shortly after arriving.

Image: 

advertiser links [what's this?](#)

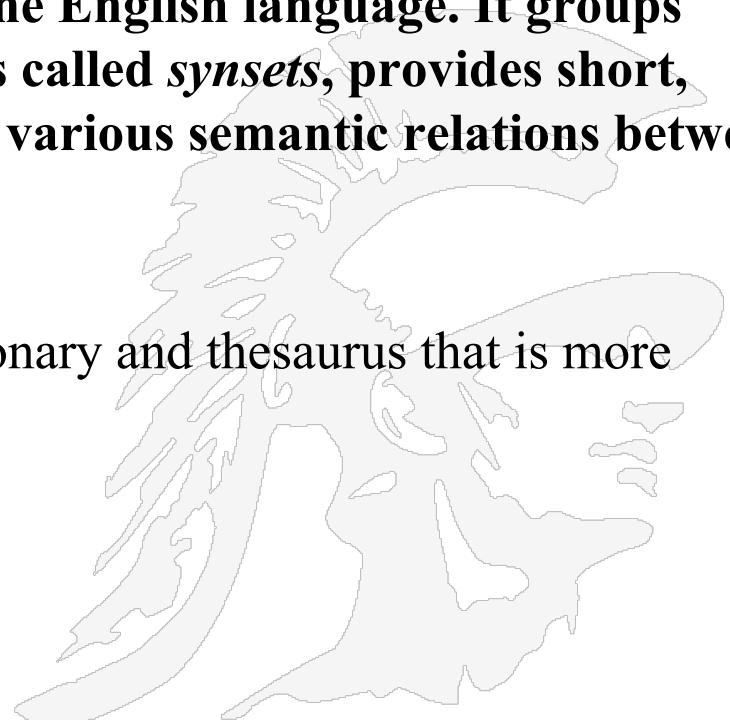
[MyCashNow - \\$100 - \\$1,000 Overnight](#)
 Payday Loan Cash goes in your account overnight. Very low fees. Fast decisions....
www.mycashnow.com

[Refinance Rates Hit Record Lows](#)
 Get \$150,000 loan for \$625 per month.

Above, an article about Michael Jackson includes an irrelevant ad about payday loans

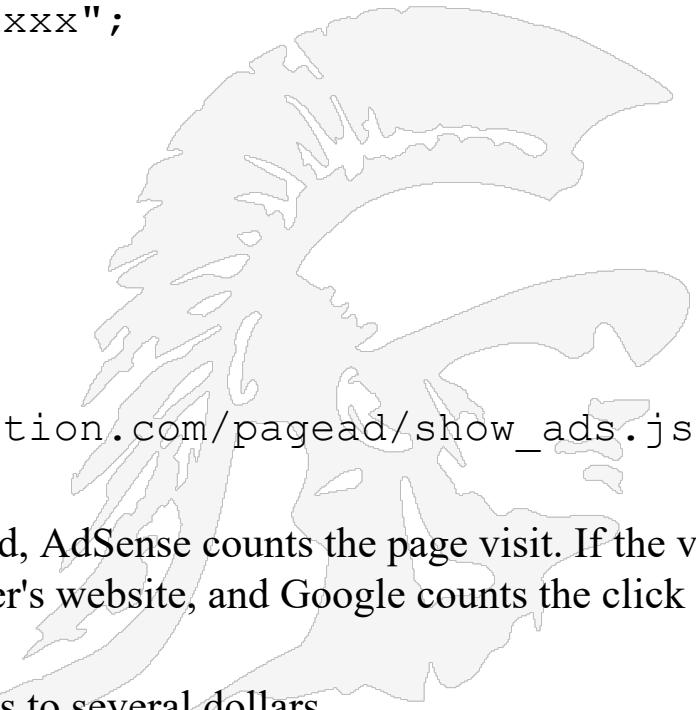
Adsense Content Matching Originally Based on WordNet

- AdSense technology is based upon a database of word meanings initially developed at Princeton, called WordNet
 - <https://wordnet.princeton.edu/>
- WordNet is a semantic lexicon for the English language. It groups English words into sets of synonyms called *synsets*, provides short, general definitions, and records the various semantic relations between these synonym sets.
 - The purpose is twofold:
 1. to produce a combination of dictionary and thesaurus that is more intuitively usable, and
 2. to support automatic text analysis.



AdSense Code

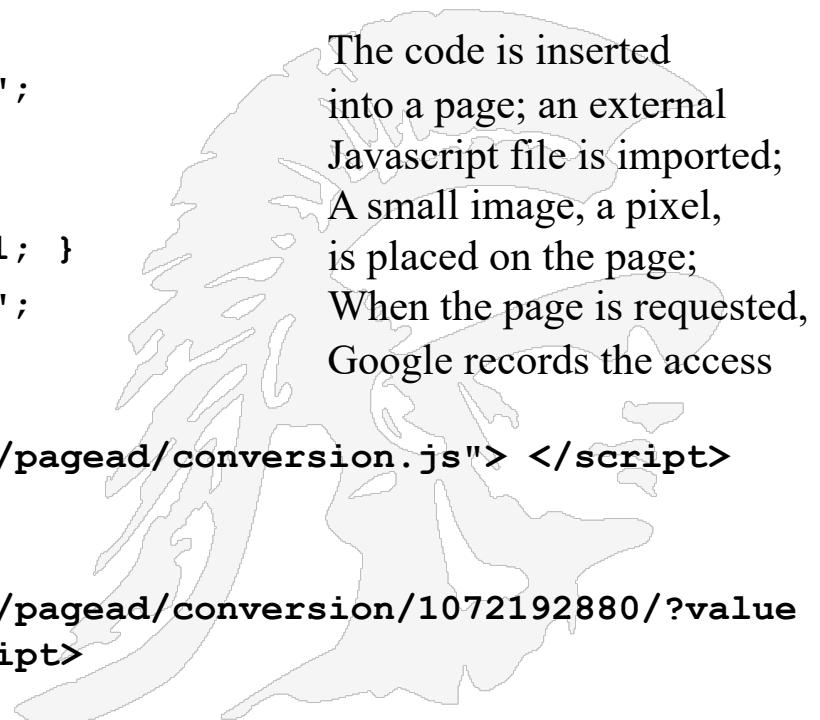
- Google AdSense creates the code, which the publisher copies and pastes into a web page or blog post at the point where it is to appear.
- This is an example of AdSense code for a 728-by-90-pixel ad. Google assigns the numbers following "google_ad_client" and "google_ad_slot."
- ```
<script type="text/javascript"><!--
google_ad_client = "xxxxxxxxxxxxxxxxxx";
/* 728x90, created 10/28/08 */
google_ad_slot = "xxxxxxxxxx";
google_ad_width = 728;
google_ad_height = 90;
//-->
</script>
<script type="text/javascript"
src="http://pagead2.googlesyndication.com/pagead/show_ads.js">
</script>
```
- When a webpage containing this code is displayed, AdSense counts the page visit. If the visitor clicks on the ad, s/he is redirected to the advertiser's website, and Google counts the click and credits the AdSense publisher's account
- Payments for clicks on ads range from a few cents to several dollars.



# Sample Google Conversion Code for Tracking Purchases/Sales

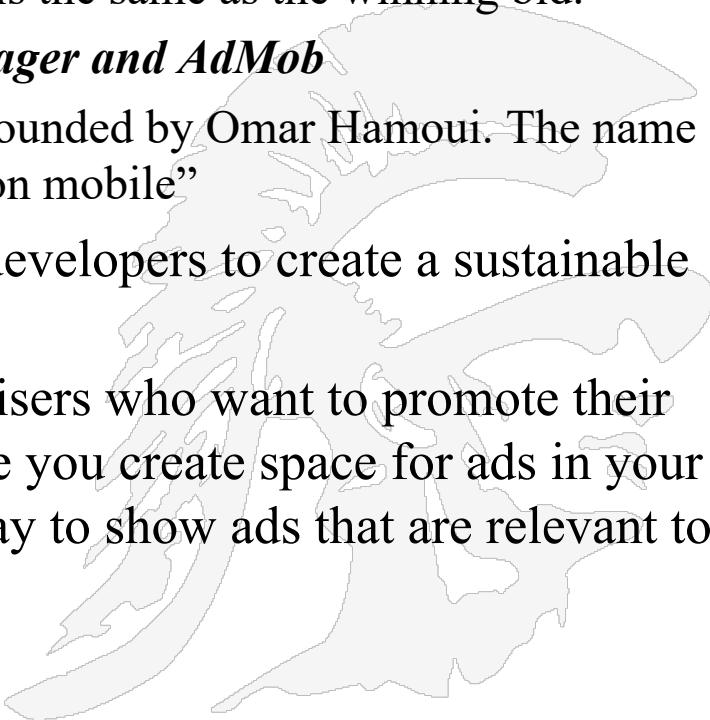
- Google/Yahoo offer tracking pixels
- Tracking pixels are small, typically transparent images on a web page that have special names which permit the loading of the web page to be tracked by a web server.

```
<!-- Google Code for purchase Conversion Page -->
<script language="JavaScript" type="text/javascript">
var google_conversion_id = 1072192880;
var google_conversion_language = "en_US";
var google_conversion_format = "1";
var google_conversion_color = "666666";
if (1) { var google_conversion_value = 1; }
var google_conversion_label = "purchase";
</script>
<script language="JavaScript"
src="http://www.googleadservices.com/pagead/conversion.js"> </script>
<noscript>
 </noscript>
```



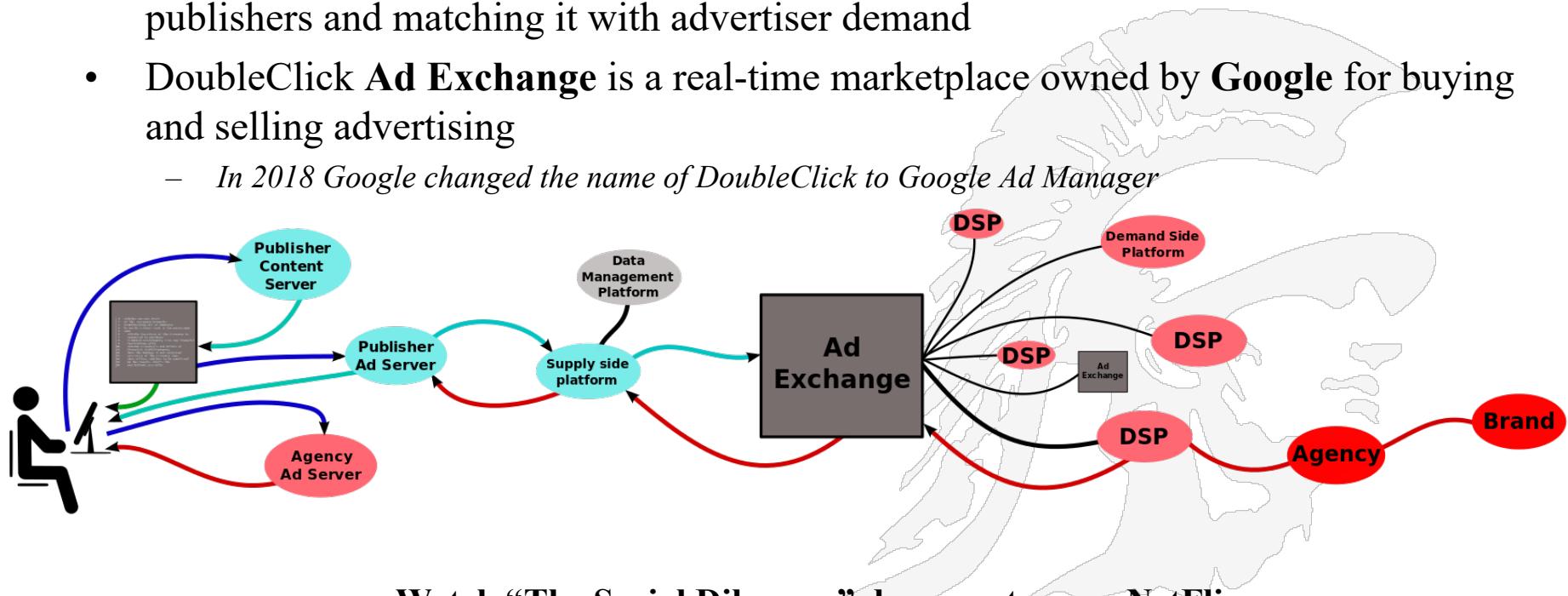
# Google AdSense, AdMob Moves to a First-Price Auction Model

- **First-price vs. second-price auctions**
  - In a second-price auction, the final price paid by the winner is determined by the second-highest bid.
  - In a first-price auction, the final price is the same as the winning bid.
- ***This change applies to AdSense, Ad Manager and AdMob***
- AdMob is a mobile advertising company founded by Omar Hamoui. The name AdMob is a portmanteau for "advertising on mobile"
- Showing ads to app users allows app developers to create a sustainable source of revenue
- Ads are created and paid for by advertisers who want to promote their products or services to app users. Once you create space for ads in your app, AdMob works with advertisers who pay to show ads that are relevant to your users



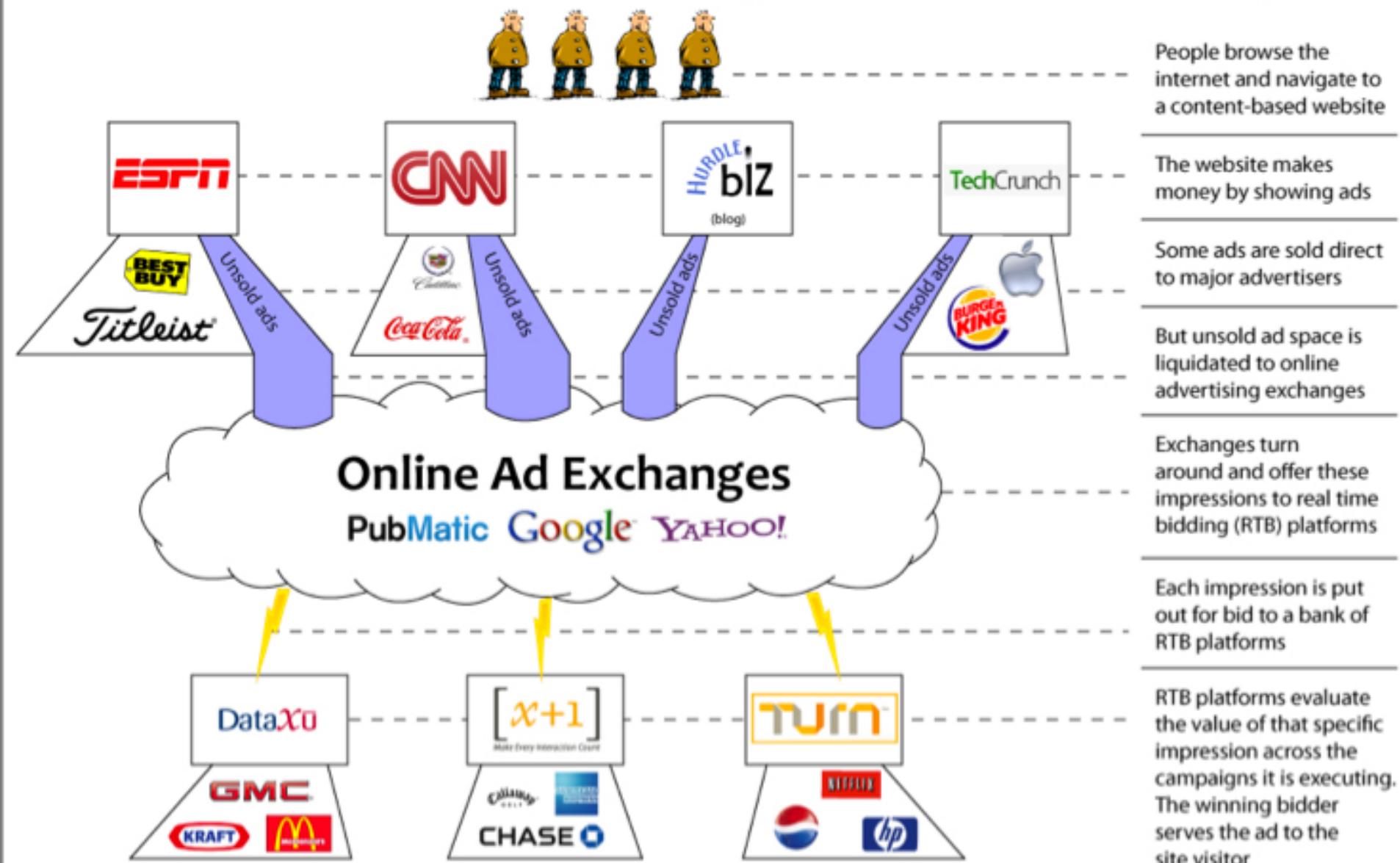
# Ad Exchanges and DoubleClick

- An **ad exchange** is a technology platform that facilitates the buying and selling of media advertising inventory from multiple **ad** networks.
- Prices for the inventory are determined through technology-driven bidding
- The key function of an **ad network** is aggregation of **ad** space supply from publishers and matching it with advertiser demand
- DoubleClick **Ad Exchange** is a real-time marketplace owned by **Google** for buying and selling advertising
  - In 2018 Google changed the name of DoubleClick to Google Ad Manager



Watch “The Social Dilemma” documentary on Netflix  
<https://www.netflix.com/title/81254224>

# The Real-Time Bidding Paradigm in Online Advertising



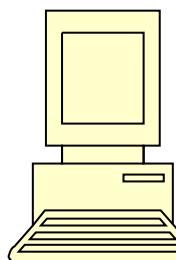
# Today's advertising model

Publishers

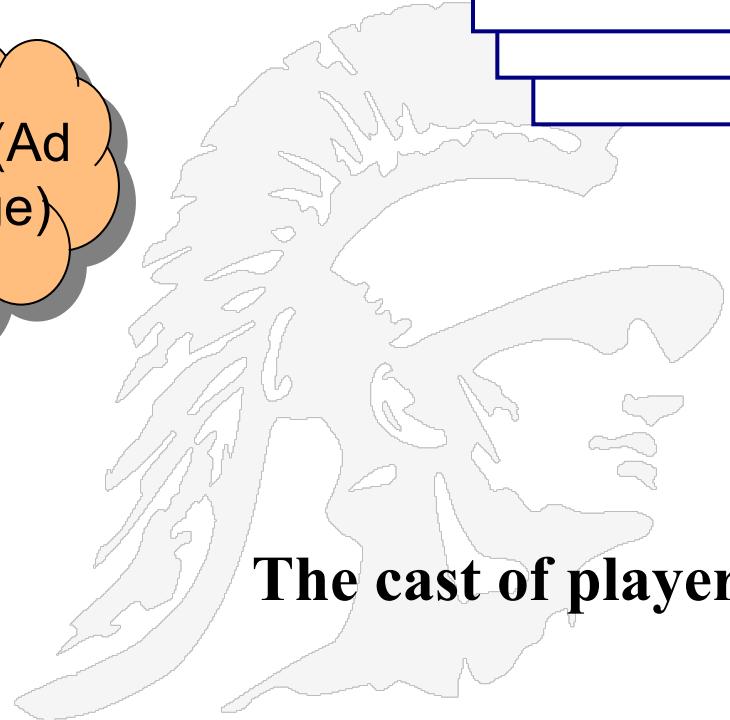
Trackers

Advertisers

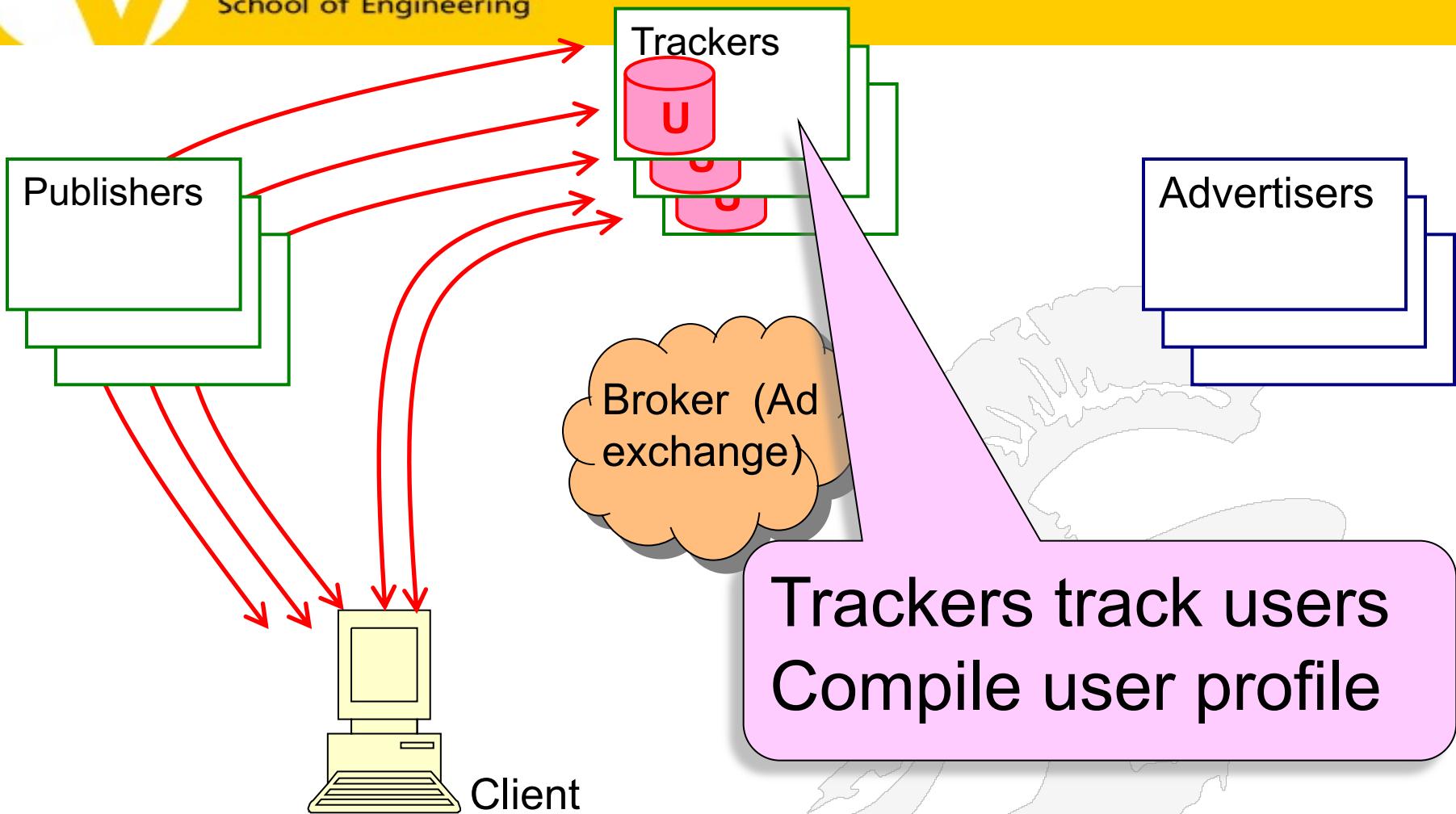
Broker (Ad  
exchange)

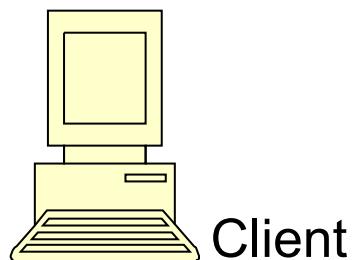
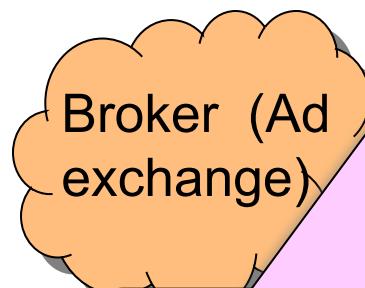
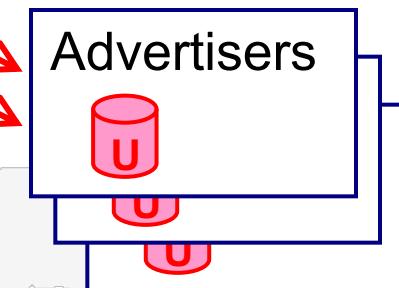
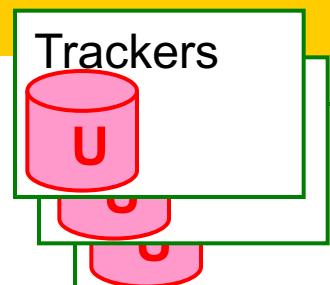
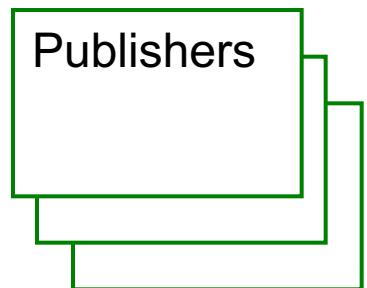


Client

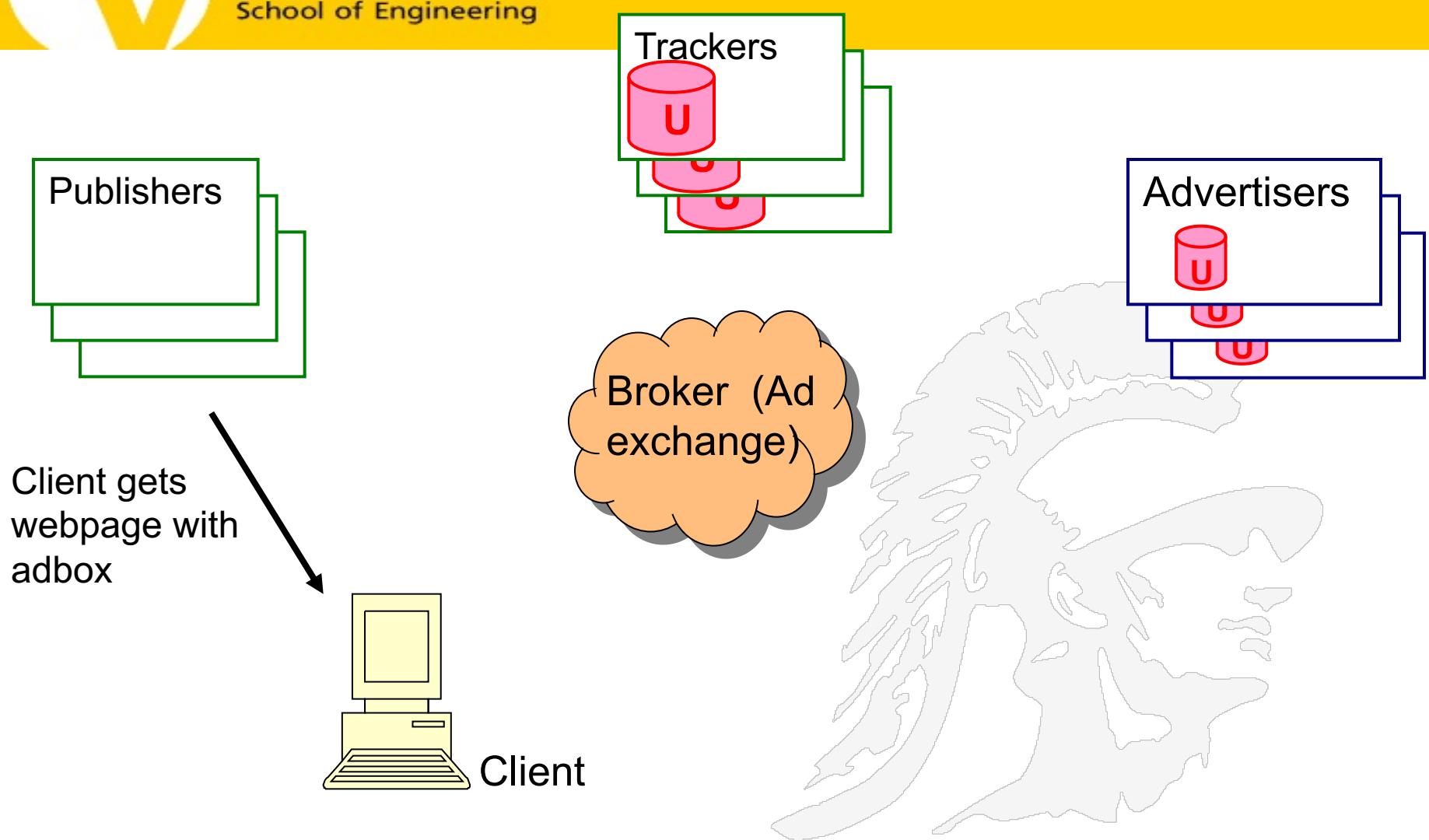


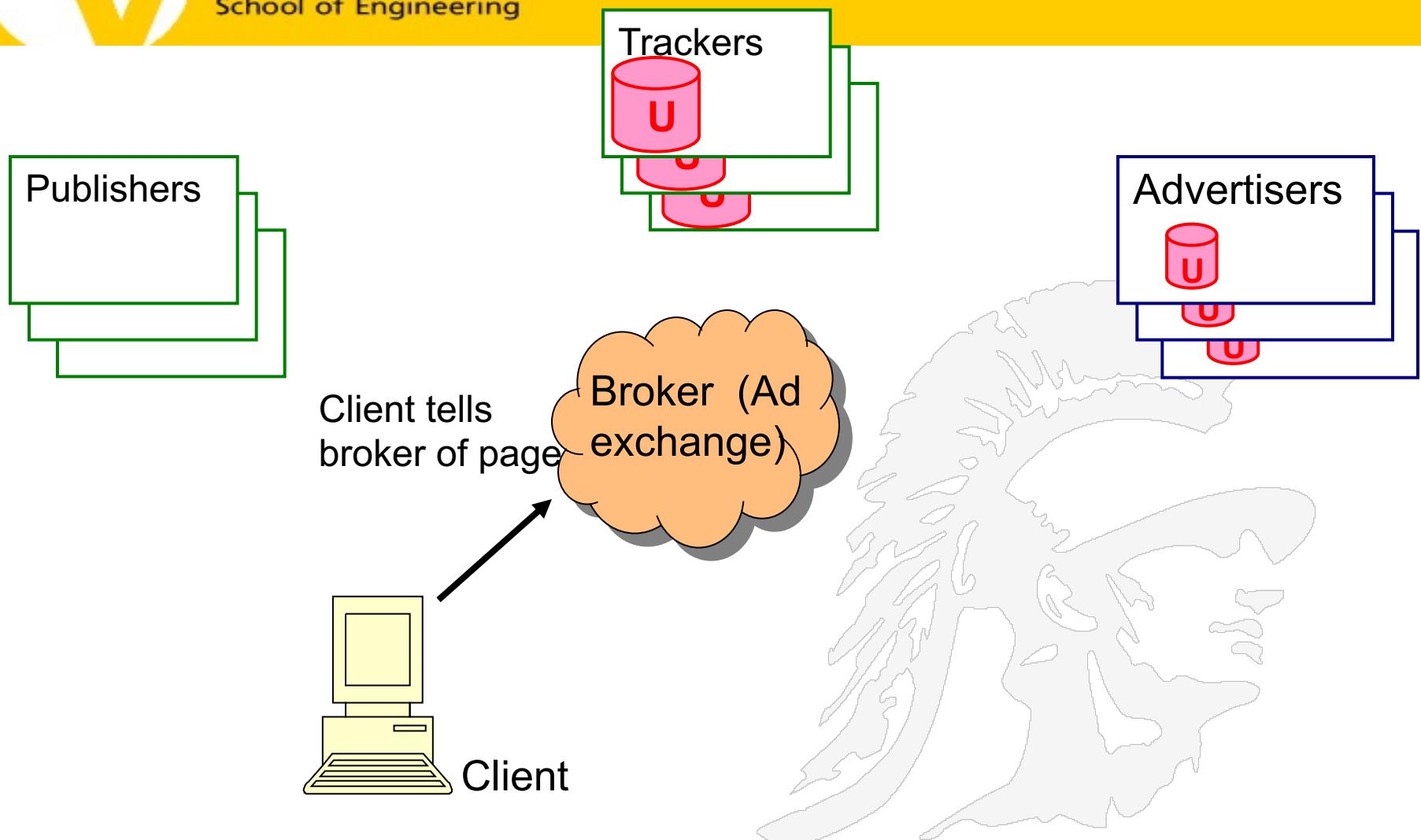
## The cast of players





Trackers may share profiles with advertisers?





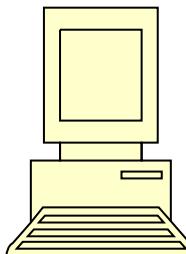
Publishers

Trackers

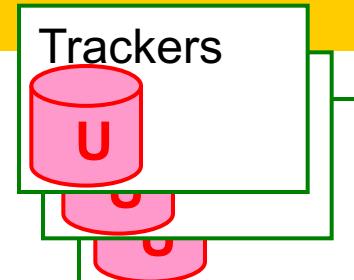
Advertisers

Broker (Ad  
exchange)

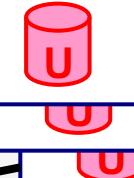
Broker launches auction  
(for given user visiting  
given webpage ....)  
Also does clickfraud etc.



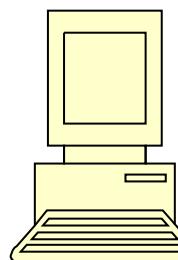
Publishers



Advertisers

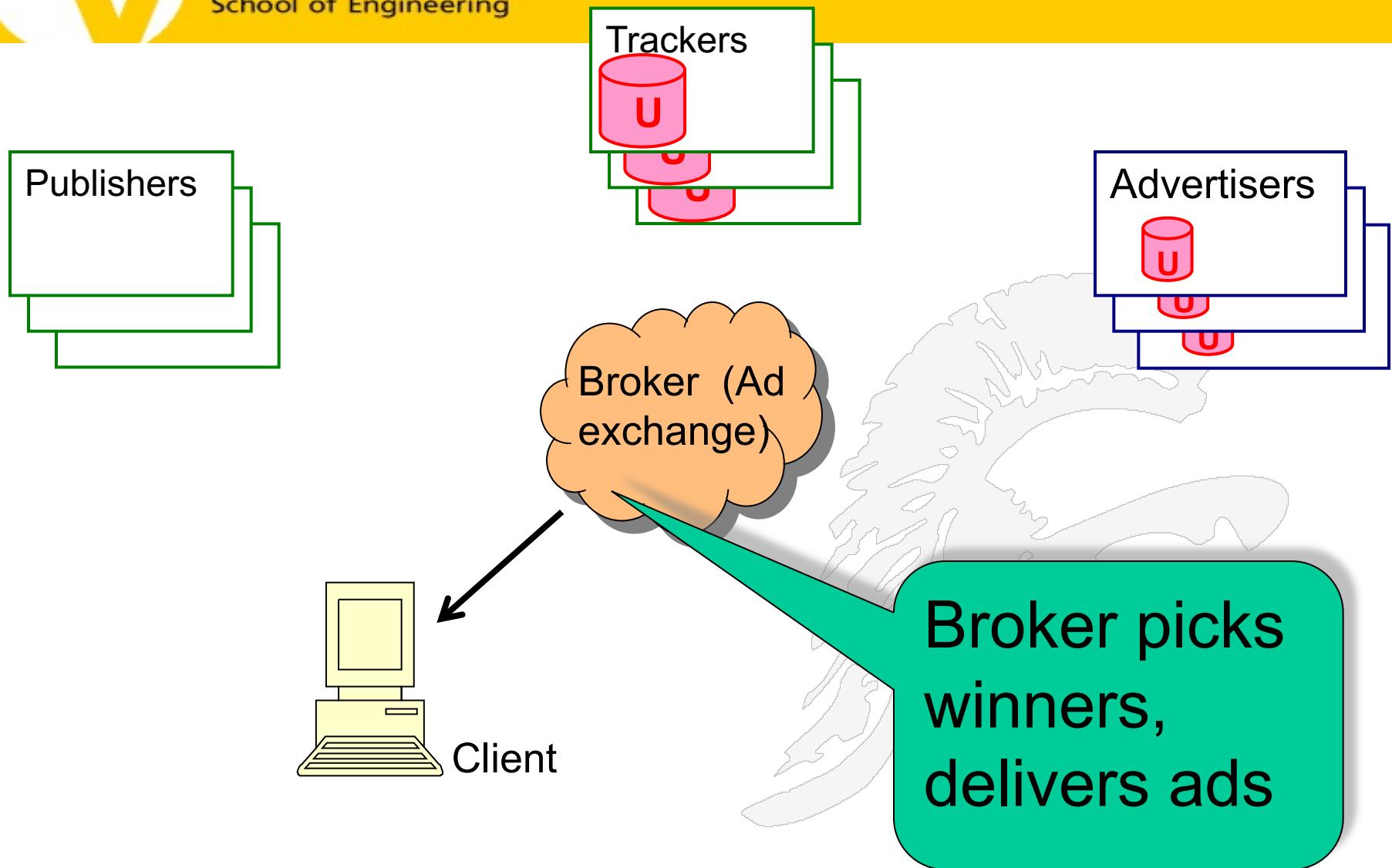


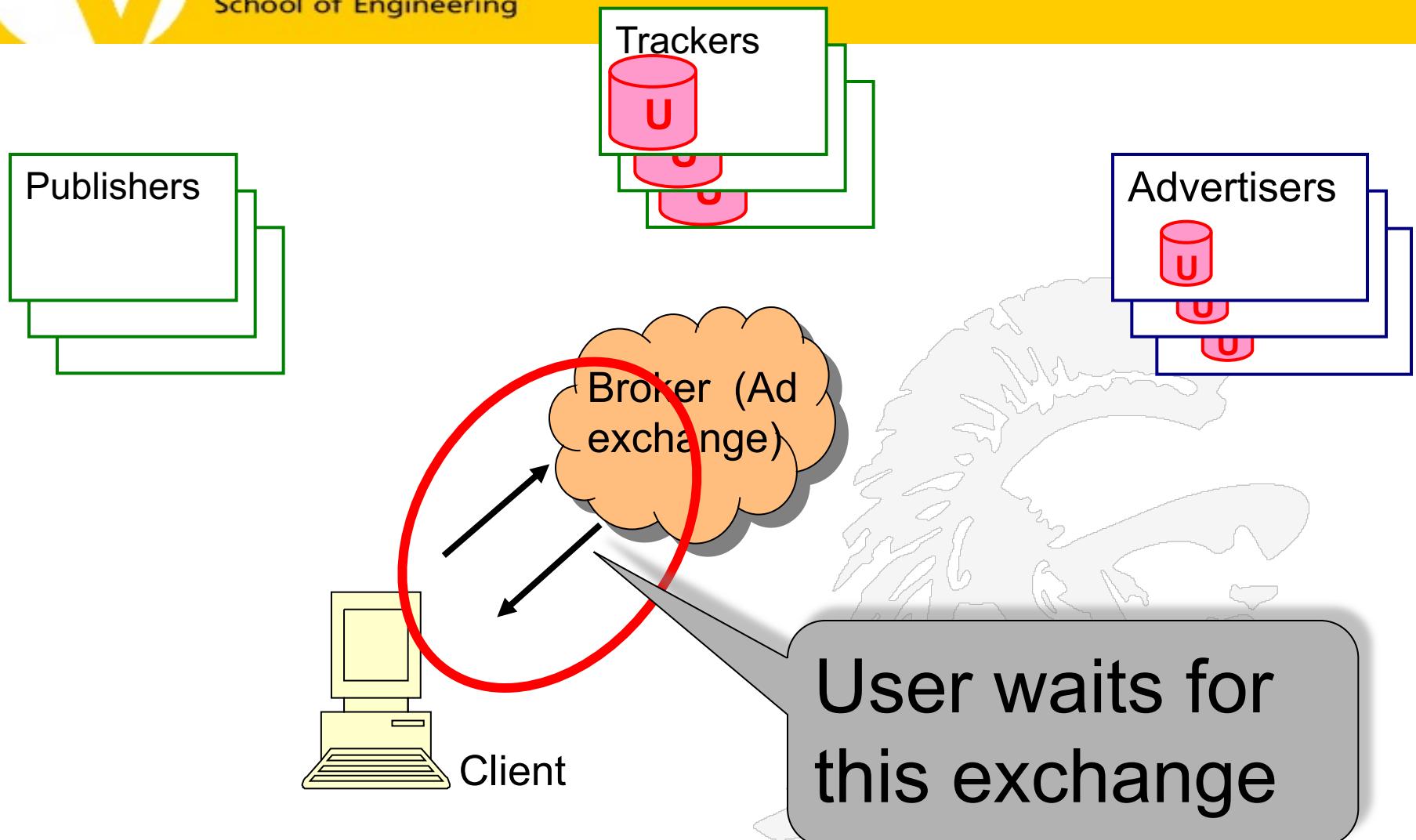
Broker (Ad exchange)

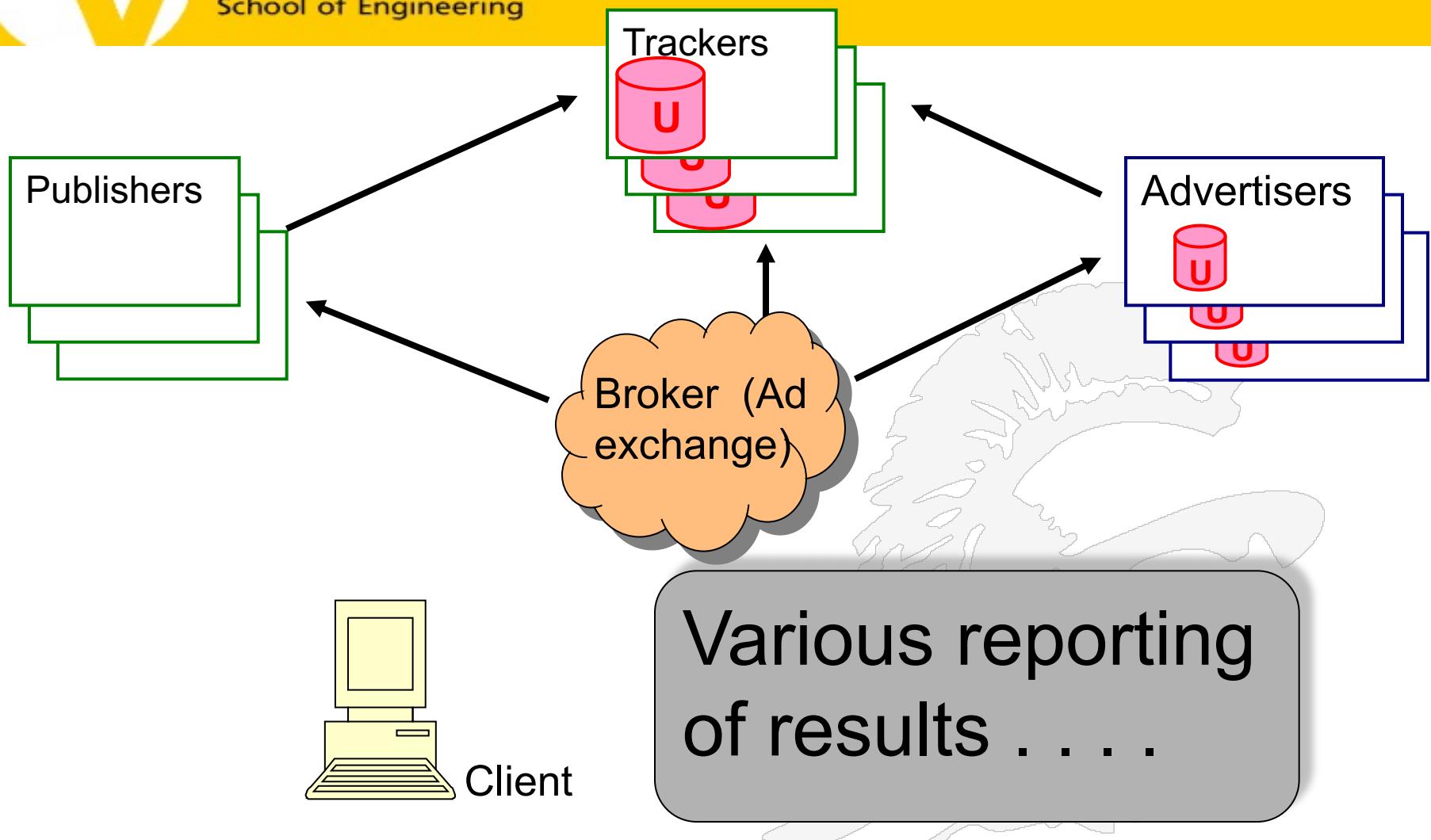


Client

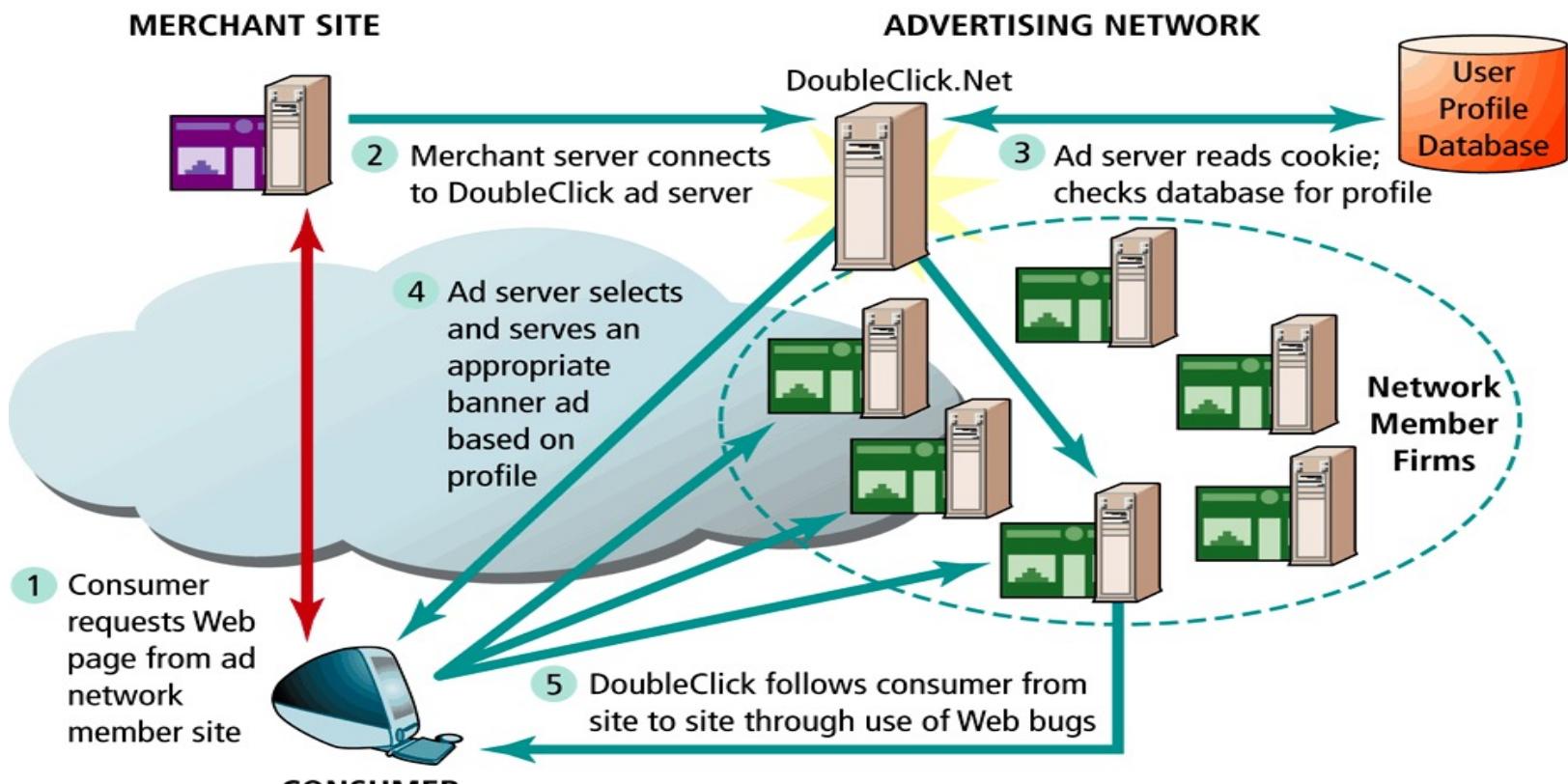
Advertisers  
present bids  
and ads



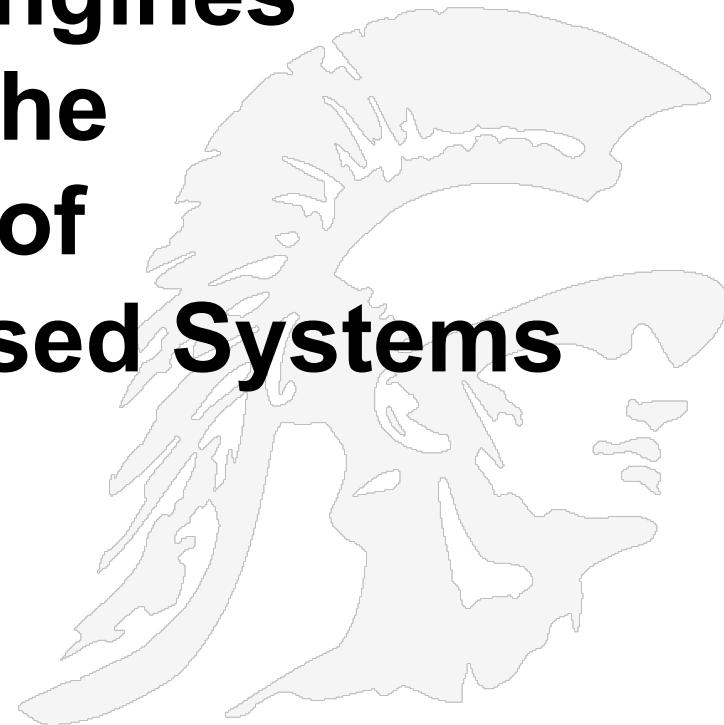




# How an Advertising Network such as DoubleClick Works

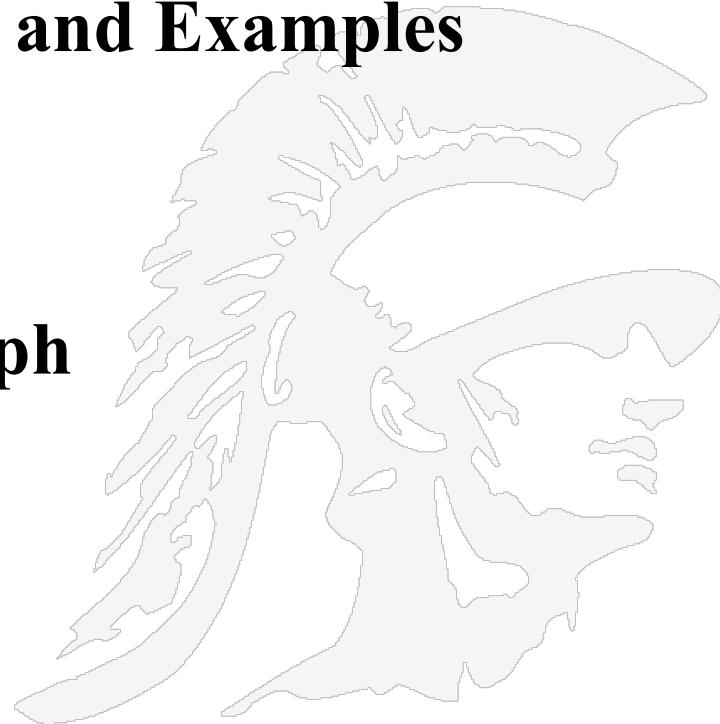


# Search Engines and the Use of Knowledge-Based Systems



# Outline

- **Basic definitions: Taxonomy, Ontology, Knowledgebase**
- **Knowledgebase Internals and Examples**
- **WordNet**
- **Wikipedia**
- **Google's Knowledge Graph**



# What is a Taxonomy

## What is a taxonomy

$\tauαξις, \tauaxis$   
 "arrangement"



$\nuουος, \nuomos$   
 "law"

A taxonomy is a classification or categorization of a complex system.

Football team  
is a  
Real Madrid C.F.



e.g. the ACM Computing Classification System, <https://dl.acm.org/ccs>  
 the Mathematics Subject Classification

Article Talk Read Edit View history Search

**Football team**

From Wikipedia, the free encyclopedia

**Football team** is the collective name given to a group of players selected together in the various team sports known as football. Such teams could be selected to play in a match against an opposing team, to represent a football club, group, state or nation, an All-star team or even selected as a hypothetical team (such as a Dream Team or Team of the Century) and never play an actual match.

There are several varieties of football, notably Association football, Gridiron football, Australian rules football, Gaelic football, rugby league, and rugby union. The number of players selected for each team within

Article Talk Read View source View history Search

**Real Madrid C.F.**

From Wikipedia, the free encyclopedia

"Real Madrid" redirects here. For the basketball team, see Real Madrid Baloncesto. For the football club in South Africa, see Real Madrid (South Africa). For other uses, see Real Madrid (disambiguation).

**Real Madrid Club de Fútbol** (in Spanish: "Re al Madrid" /fʊtbol/; Royal Madrid Football Club) is a community known as the "Madrid", or simply as a professional football club based in Madrid, Spain.

Founded in 1902 as Madrid Football Club, the team has traditionally worn a white kit since the word *real* is Spanish for royal and was bestowed to the club by King Alfonso XIII in 1920 together with the royal crown in the emblem. The team has played its home matches in the 85,444-seater Santiago Bernabéu Stadium in downtown Madrid since 1947. One of the most successful football clubs, Real Madrid's members (sociedad) have owned and operated the club since its inception.

The club is the world's richest football club in terms of revenue, with an annual turnover of €513 million, and the most valuable sports team, worth €2.4 billion (US\$3.3 billion) PEIFER.

Article Talk Read View source View history Search

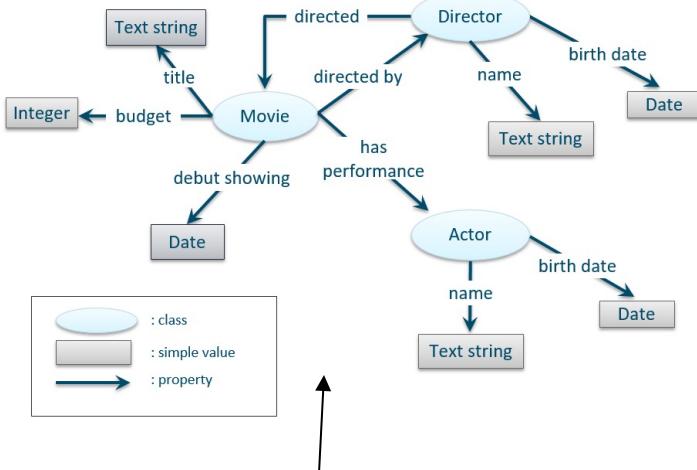
**Real Madrid**

Real Madrid C.F.

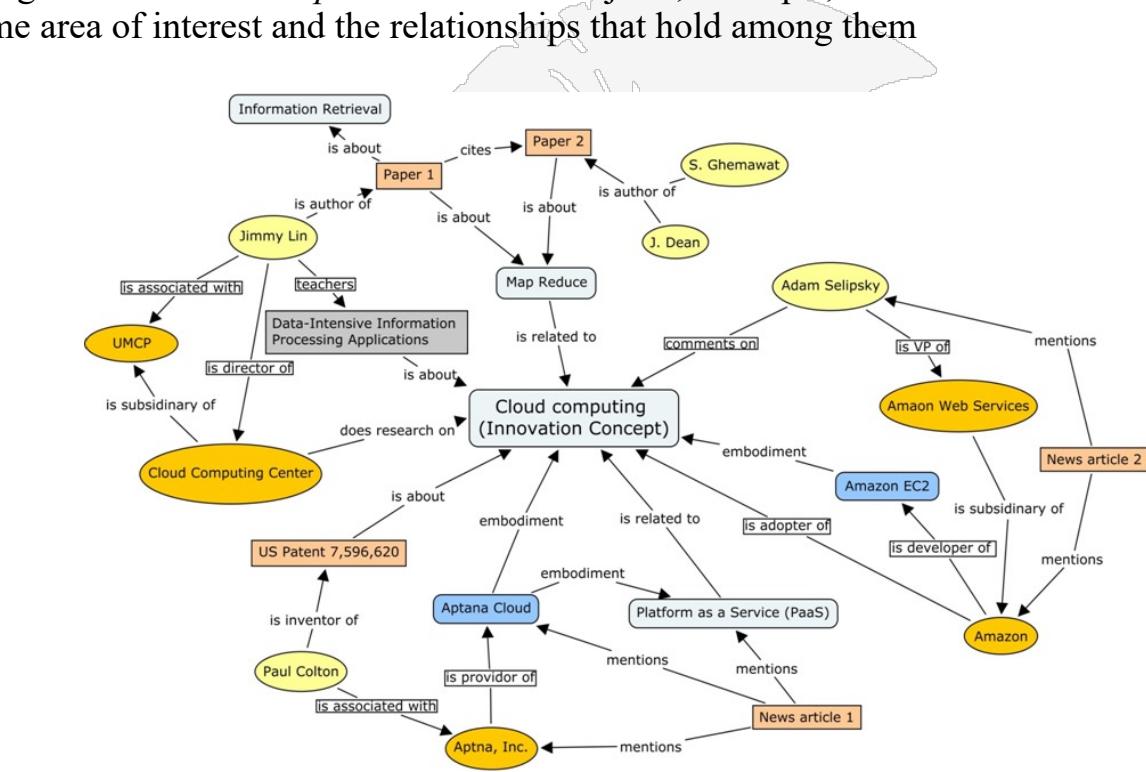
# What is an Ontology

- **3 definitions**

1. a set of concepts and categories in a subject area or domain that shows their properties and the relations between them, or
2. a description (like a formal specification of a program) of the concepts and relationships that can exist for an agent or a community of agents, or
3. a body of formally represented knowledge based on a *conceptualization*: the objects, concepts, and other entities that are assumed to exist in some area of interest and the relationships that hold among them

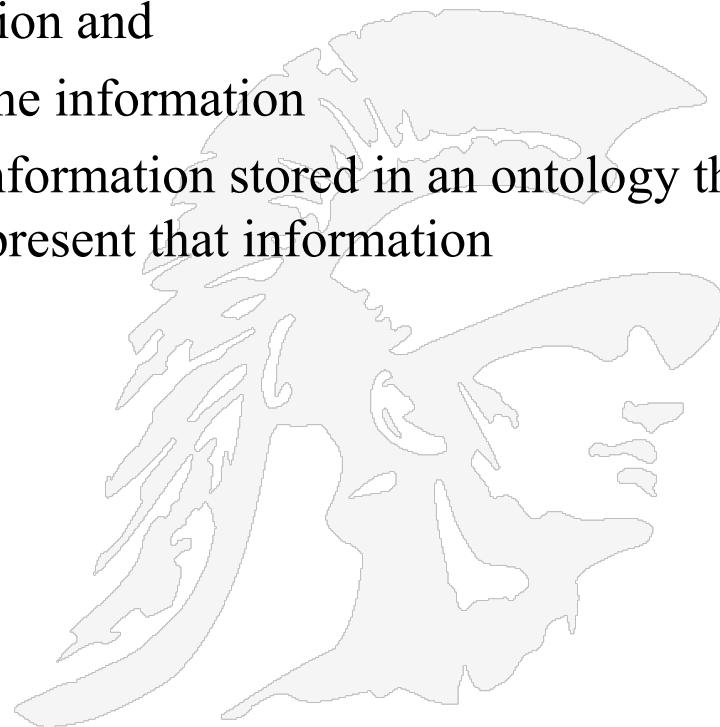


Brief ontology of movies



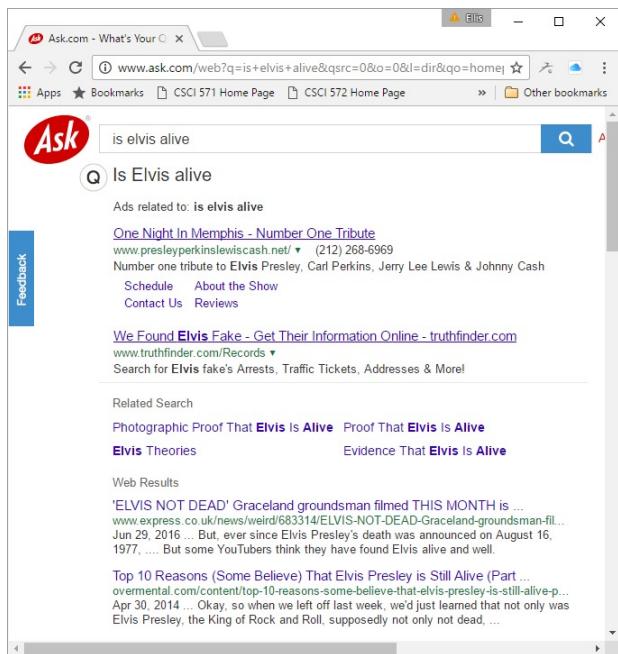
# What is a Knowledgebase

- A **knowledgebase (KB)** is a *technology* used to store and retrieve complex structured and unstructured information as stored in an ontology
  - Two components:
  - 1. a way of *representing* information and
  - 2. a method for *reasoning* about the information
- A **knowledgebase** is a collection of information stored in an ontology that includes software used to author and present that information



# USC Viterbi

School of Engineering



Ask.com - What's Your  X

[www.ask.com/web?q=is+elvis+alive&qsrc=0&o=0&l=dir&qo=home](http://www.ask.com/web?q=is+elvis+alive&qsrc=0&o=0&l=dir&qo=home)

Ads related to: is elvis alive

[One Night In Memphis - Number One Tribute](#)  
www.presleyperkinslewiscash.net/ (212) 268-6969

Number one tribute to Elvis Presley, Carl Perkins, Jerry Lee Lewis & Johnny Cash

[Schedule](#) [About the Show](#)  
[Contact Us](#) [Reviews](#)

We Found [Elvis Fake - Get Their Information Online](#) - truthfinder.com  
www.truthfinder.com/Records

Search for Elvis fake's Arrests, Traffic Tickets, Addresses & More!

Related Search

[Photographic Proof That Elvis Is Alive](#) [Proof That Elvis Is Alive](#)  
[Elvis Theories](#) [Evidence That Elvis Is Alive](#)

Web Results

'ELVIS NOT DEAD' Graceland groundsman filmed THIS MONTH is ...  
www.express.co.uk/news/weird/683314/ELVIS-NOT-DEAD-Graceland-groundsman-fil... Jun 29, 2016 ... But ever since Elvis Presley's death was announced on August 16, 1977, ... But some YouTubers think they have found Elvis alive and well.

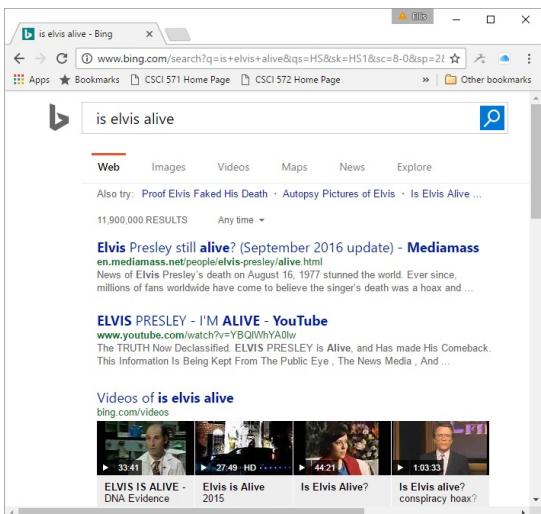
Top 10 Reasons (Some Believe) That Elvis Presley is Still Alive (Part ... overmental.com/content/top-10-reasons-some-believe-that-elvis-presley-is-still-alive-p... Apr 30, 2014 ... Okay, so when we left off last week, we'd just learned that not only was Elvis Presley, the King of Rock and Roll, supposedly not only not dead, ...

Feedback

Ask.com – Mostly Yes

**2 Yes'es and 2 No's;  
text matching alone is insufficient  
we need to “understand” the query**

Copyright Ellis Horowitz 2011-2022



Bing.com - is elvis alive X

[www.bing.com/search?q=is+elvis+alive&qs=HS&sk=HS1&sc=8-0&sp=21](http://www.bing.com/search?q=is+elvis+alive&qs=HS&sk=HS1&sc=8-0&sp=21)

Also try: [Proof Elvis Faked His Death](#) · [Autopsy Pictures of Elvis](#) · [Is Elvis Alive ...](#)

11,900,000 RESULTS Any time →

**Elvis Presley still alive? (September 2016 update) - Mediabass**  
en.mediabass.net/people/elvis-presley/alive.html

News of Elvis Presley's death on August 16, 1977 stunned the world. Ever since, millions of fans worldwide have come to believe the singer's death was a hoax and ...

**ELVIS PRESLEY - I'M ALIVE - YouTube**  
www.youtube.com/watch?v=tBOMWYAl0lw

The TRUTH Now Declassified! ELVIS PRESLEY Is Alive, and Has made His Comeback. This Information Is Being Kept From The Public Eye, The News Media, , And ...

**Videos of is elvis alive**  
bing.com/videos

ELVIS IS ALIVE - DNA Evidence 2015 | Elvis is Alive | Is Elvis Alive? | Is Elvis alive? conspiracy hoax?

Bing.com – Yes/Maybe



WolframAlpha computational knowledge engine X

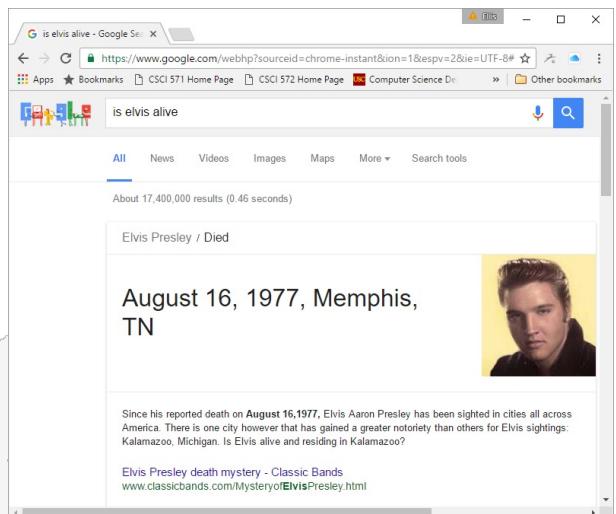
<https://www.wolframalpha.com/input/?i=is+elvis+alive>

Input interpretation: **Elvis Presley alive?**

Result: **No**

Powered by the WOLFRAM LANGUAGE

# Is Elvis Alive



Google.com - is elvis alive X

<https://www.google.com/webhp?sourceid=chrome-instant&ion=1&espv=2&ie=UTF-8#q=is+elvis+alive>

About 17,400,000 results (0.46 seconds)

Elvis Presley / Died

August 16, 1977, Memphis, TN



Since his reported death on August 16, 1977, Elvis Aaron Presley has been sighted in cities all across America. There is one city however that has gained a greater notoriety than others for Elvis sightings: Kalamazoo, Michigan. Is Elvis alive and residing in Kalamazoo?

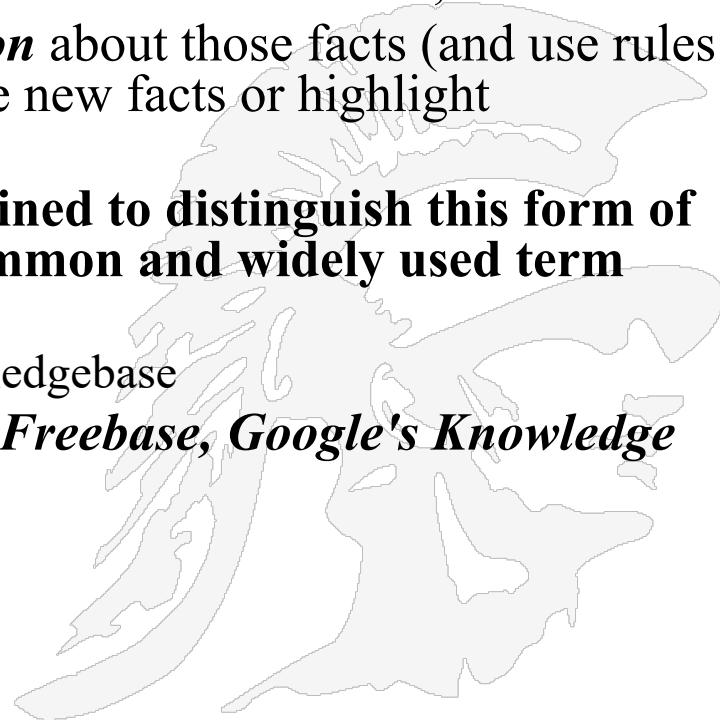
Elvis Presley death mystery - Classic Bands  
www.classicbands.com/MysteryofElvisPresley.html

Google.com - No

Wolfram Alpha - No

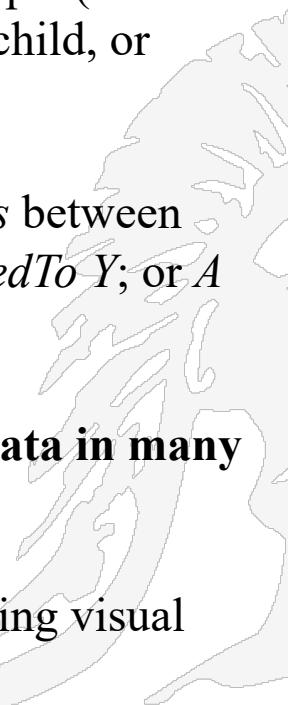
# KnowledgeBases-Developed by AI Community

- To move away from just using keyword matching, search engines borrowed techniques developed by AI researchers
- A knowledge-based system consists of two elements:
  1. a *knowledgebase* that *represents* facts about the world, and
  2. an *inference engine* that can *reason* about those facts (and use rules and other forms of logic to deduce new facts or highlight inconsistencies)
- The term "knowledgebase" was coined to distinguish this form of knowledge store from the more common and widely used term *database*
  - a relational database is NOT a knowledgebase
- *Early examples of knowledgebases: Freebase, Google's Knowledge Graph, Apple's Siri, IBM's Watson*



## Search Engines Use Knowledgebases to Enhance the Display of Results

- The representation of knowledge in a knowledgebase is an *object model*
  - Includes classes, subclasses and instances
- A **taxonomy** is usually only a hierarchy of concepts (i.e. the *only relation* between the concepts is parent/child, or subClass/superClass, or broader/narrower)
- In a **knowledgebase**, *arbitrary complex relations* between concepts can be expressed as well, e.g. ( $X \text{ marriedTo } Y$ ; or  $A \text{ worksFor } B$ ; or  $C \text{ locatedIn } D$ , etc )
- **Search engines utilize this linked, structured data in many ways, such as**
  - Providing direct answers to queries
  - enhanced displays in many varieties of engaging visual formats, e.g. see query “Picasso” in Google



Pablo Picasso

Spanish painter

Pablo Ruiz Picasso was a Spanish painter, sculptor, printmaker, ceramicist and theatre designer who spent most of his adult life in France. [Wikipedia](#)

**Born:** October 25, 1881, Málaga, Spain

**Died:** April 8, 1973, Mougins, France

**On view:** [The Museum of Modern Art](#), [The Art Institute of Chicago](#), MORE

**Periods:** Cubism, Surrealism, Expressionism, Post-Impressionism, MORE

**Full name:** Pablo Diego José Francisco de Paula Juan Nepomuceno María de los Remedios Cipriano de la Santísima Trinidad Ruiz y Picasso

**Spouse:** [Jacqueline Roque](#) (m. 1961–1973), [Olga Khokhlova](#) (m. 1918–1955)

### Artworks



Guernica



Les



The

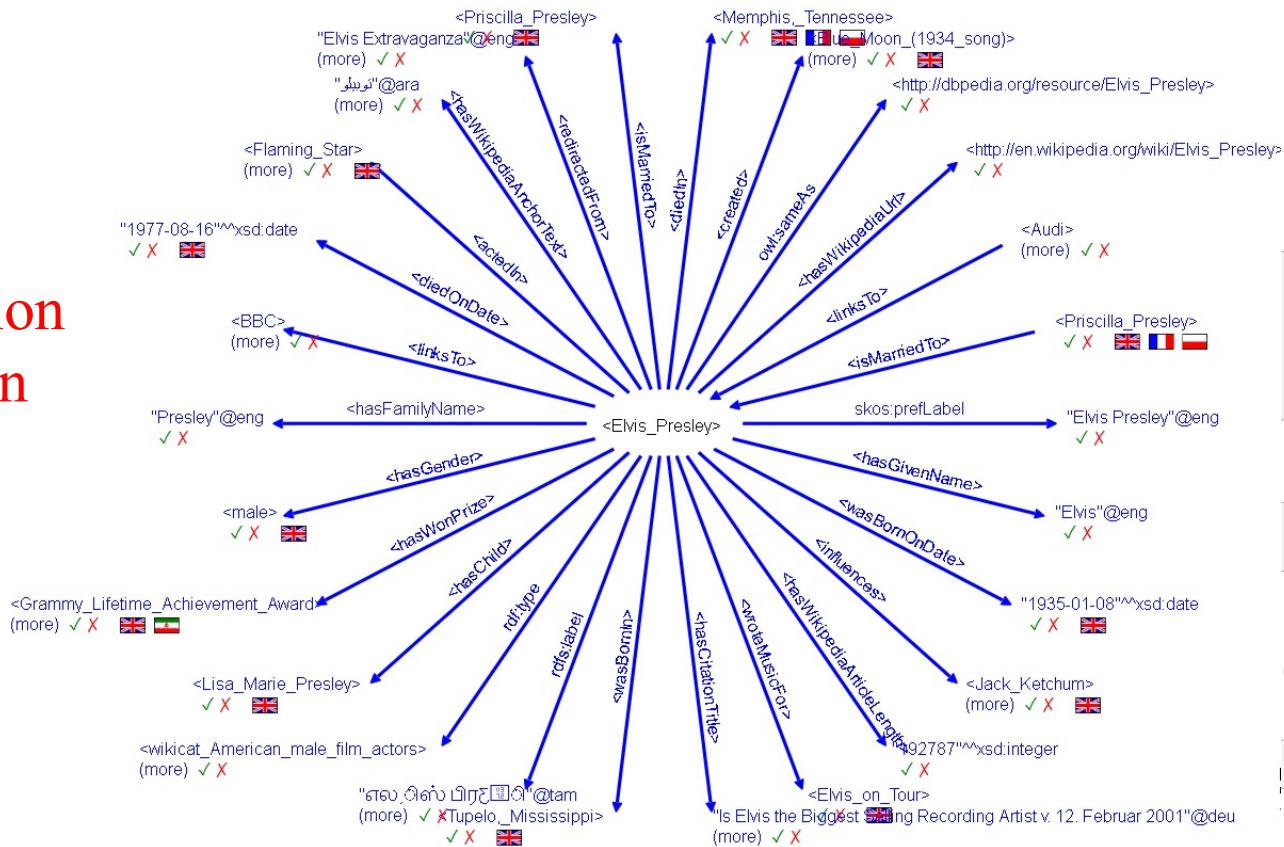
View 25+ more



The Old

# Close Up: A KnowledgeBase for Elvis

# Information Extraction



"Elvis Presley, the first and greatest American rock-and roll star, died yesterday at the age of 42."

# Content analytics

# Types of Knowledge For a KnowledgeBase

Elvis Presley type American singer  
Elvis Presley type Baritone  
American singer subclassOf singer  
Elvis Presley sang All Shook Up  
Elvis Presley bornIn Tupelo  
id11: Elvis Presley marriedTo Priscilla Presley  
id11 validDuring [1967, 1977]  
Elvis Presley „has twin brother“ Jesse Garon  
Elvis Presley „possibly has origin“ Cherokee  
Elvis Presley knownAs „The King of R&R“

**taxonomic knowledge**

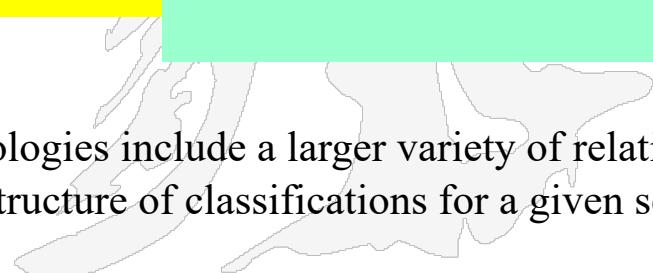
**factual knowledge**

**temporal knowledge**

**emerging knowledge**

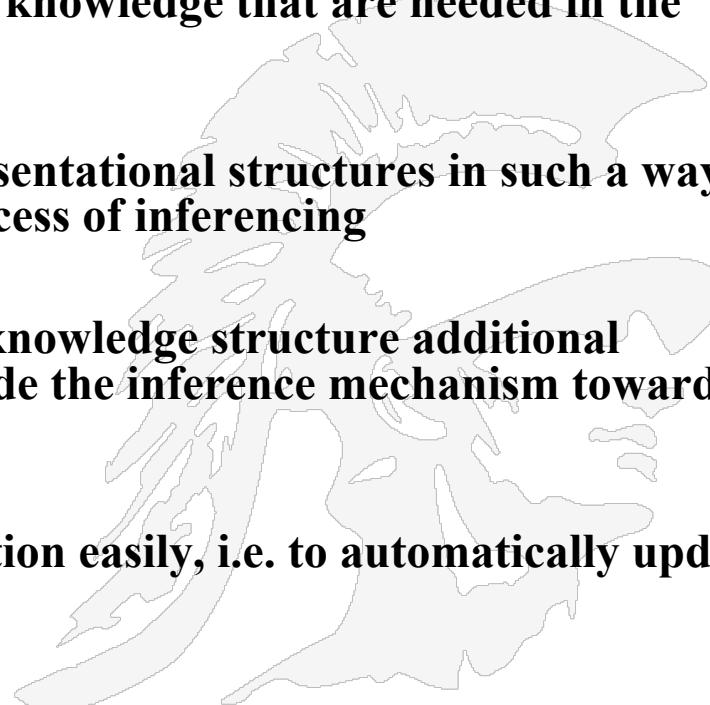
**terminological  
knowledge**

Taxonomies are narrower than ontologies since ontologies include a larger variety of relation types.  
Mathematically, a hierarchical **taxonomy** is a tree structure of classifications for a given set of objects  
An ontology is a directed, labeled, cyclic graph.



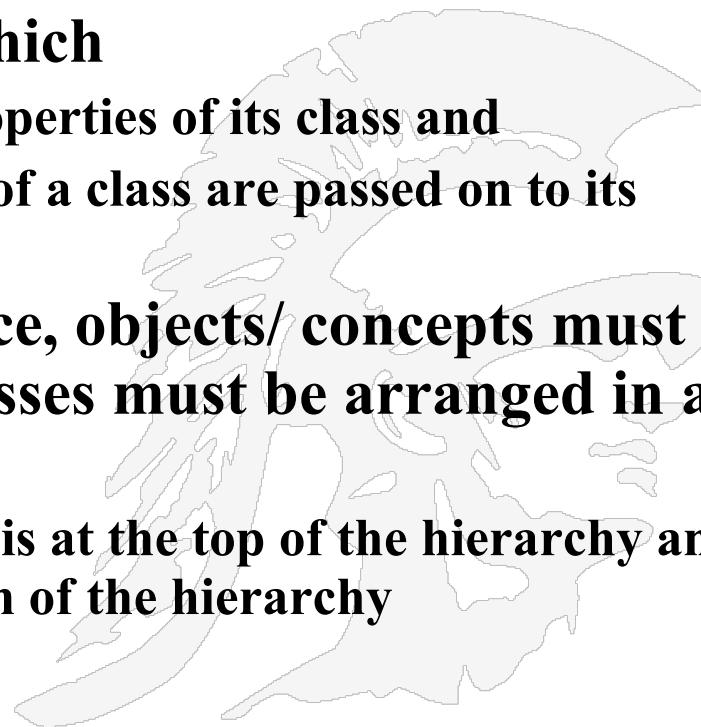
# Properties of A Good Knowledge Representation system

- Whatever knowledge representation scheme has been chosen to represent a particular domain it should exhibit the following four properties
- *Representation adequacy*
  - The ability to represent all kinds to knowledge that are needed in the domain
- *Inferential ability*
  - The ability to manipulate the representational structures in such a way as to derive new structures by the process of inferencing
- *Inferential efficiency*
  - The ability to incorporate into the knowledge structure additional information that can be used to guide the inference mechanism towards the most promising path
- *Acquisitional efficiency*
  - The ability to acquire new information easily, i.e. to automatically update the structure with new knowledge

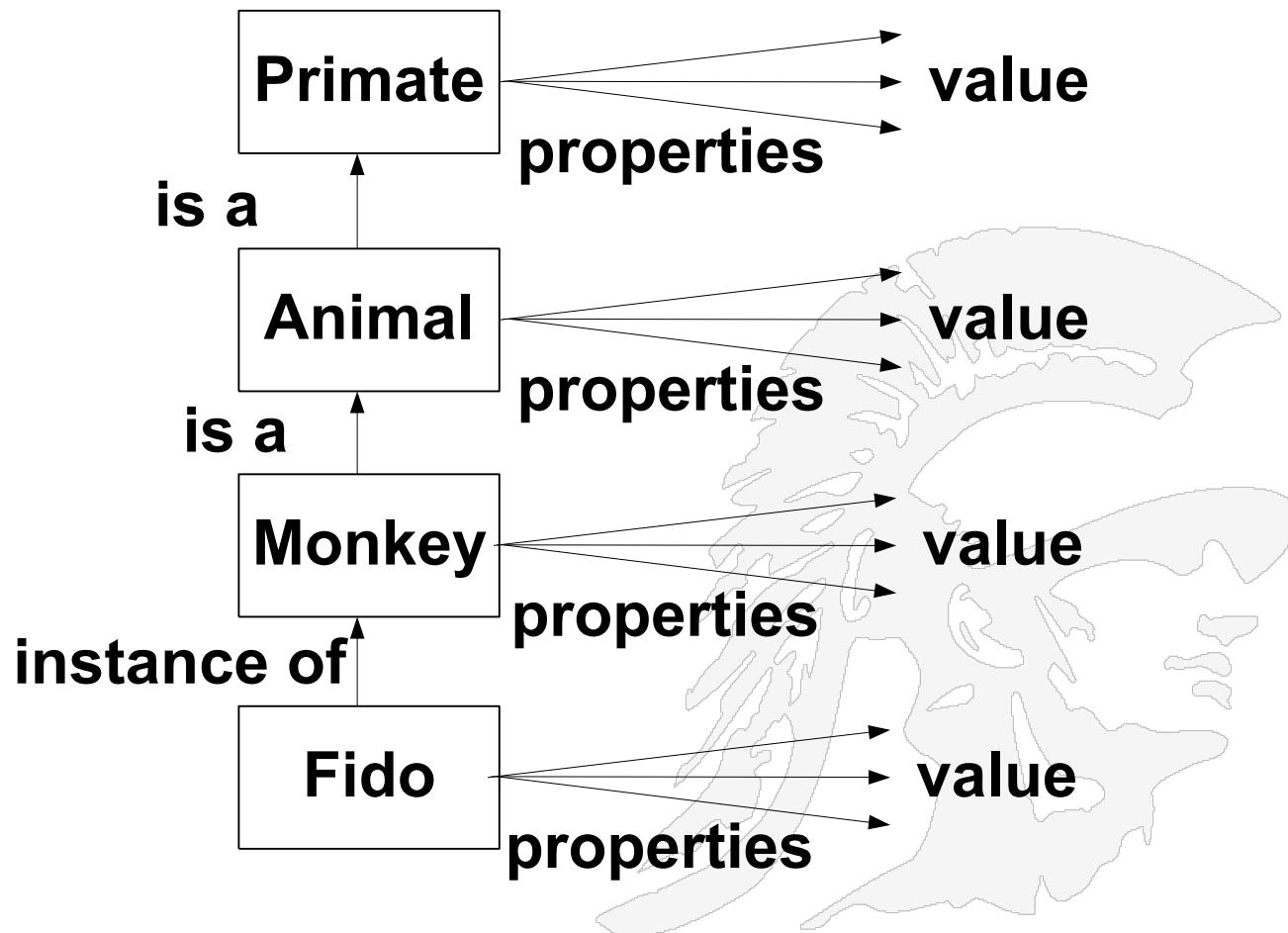


# Inheritance

- An important feature of knowledge representation is its organization into *class hierarchies*. Classes can be based on the properties of objects/concepts
- Inheritance is a relation by which
  - 1. an individual assumes the properties of its class and
  - 2. determines which properties of a class are passed on to its subclass
- In order to support inheritance, objects/ concepts must be organized into classes and classes must be arranged in a generalized hierarchy
  - the most generic object/concept is at the top of the hierarchy and the most specific is at the bottom of the hierarchy

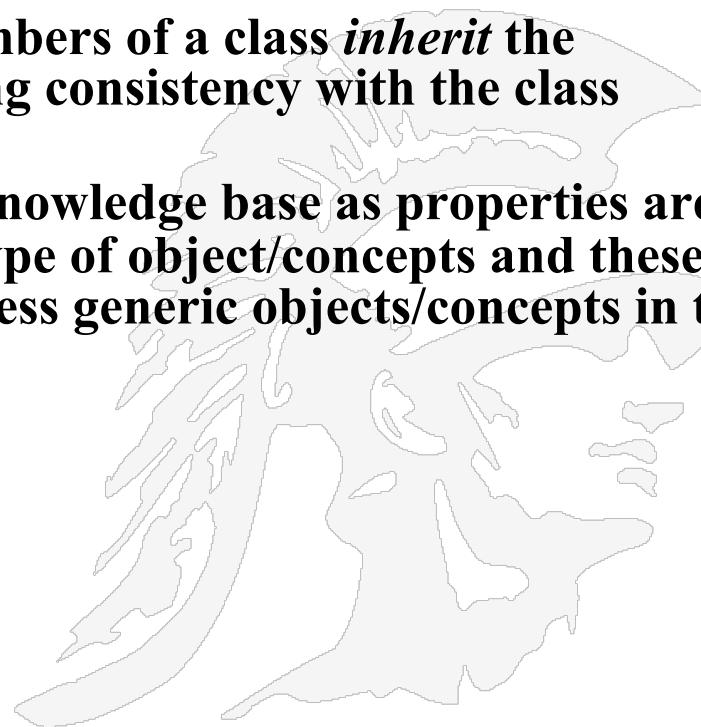


# Inheritance Example



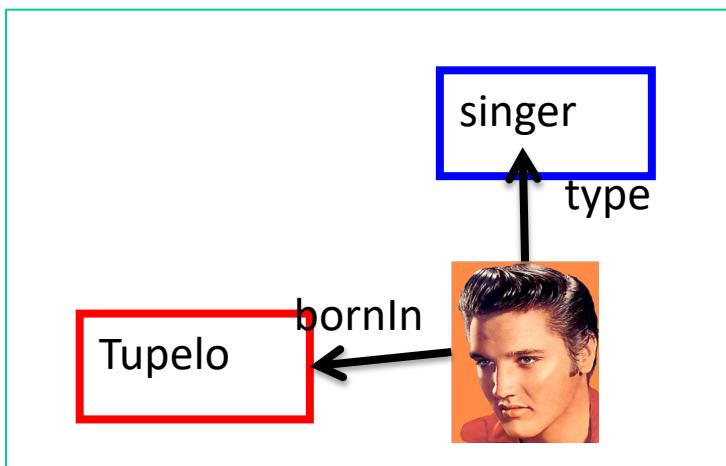
# Some Advantages of Inheritance

- Inheritance provides a *natural* mechanism for representing taxonomically structured knowledge
- Inheritance provides an *economical* means of expressing properties common to a class of objects/concepts
- Inheritance guarantees that all members of a class *inherit* the appropriate properties thus ensuring consistency with the class definition
- Inheritance *reduces the size* of the knowledge base as properties are defined once for the most general type of object/concepts and these properties are then shared by other less generic objects/concepts in the type hierarchy.



- Resource Description Format (RDF) is a W3C spec used for creating ontologies;
  - <https://www.w3.org/RDF/>
  - Sometimes "RDF Ontology" and "KnowledgeBase (KB)" are used synonymously.

### Graph notation:



### Different Notations for a KnowledgeBase

#### Triple notation:

Subject	Predicate	Object
Elvis	type	singer
Elvis	bornIn	Tupelo
...	...	...

#### Logical notation:

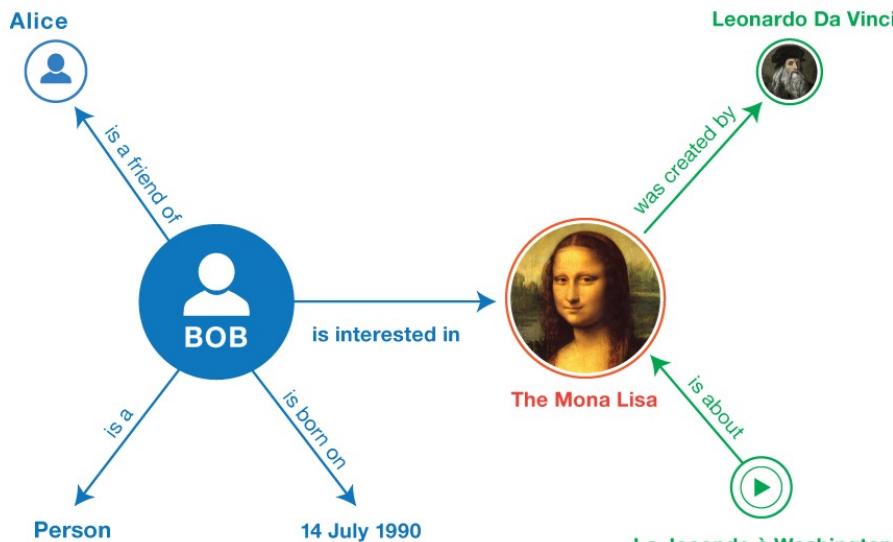
`type(Elvis, singer)`  
`bornIn(Elvis, Tupelo)`

# RDF Data Model

- RDF allows us to make statements about resources. The format of these statements is: <subject> <predicate> <object>

- Some examples

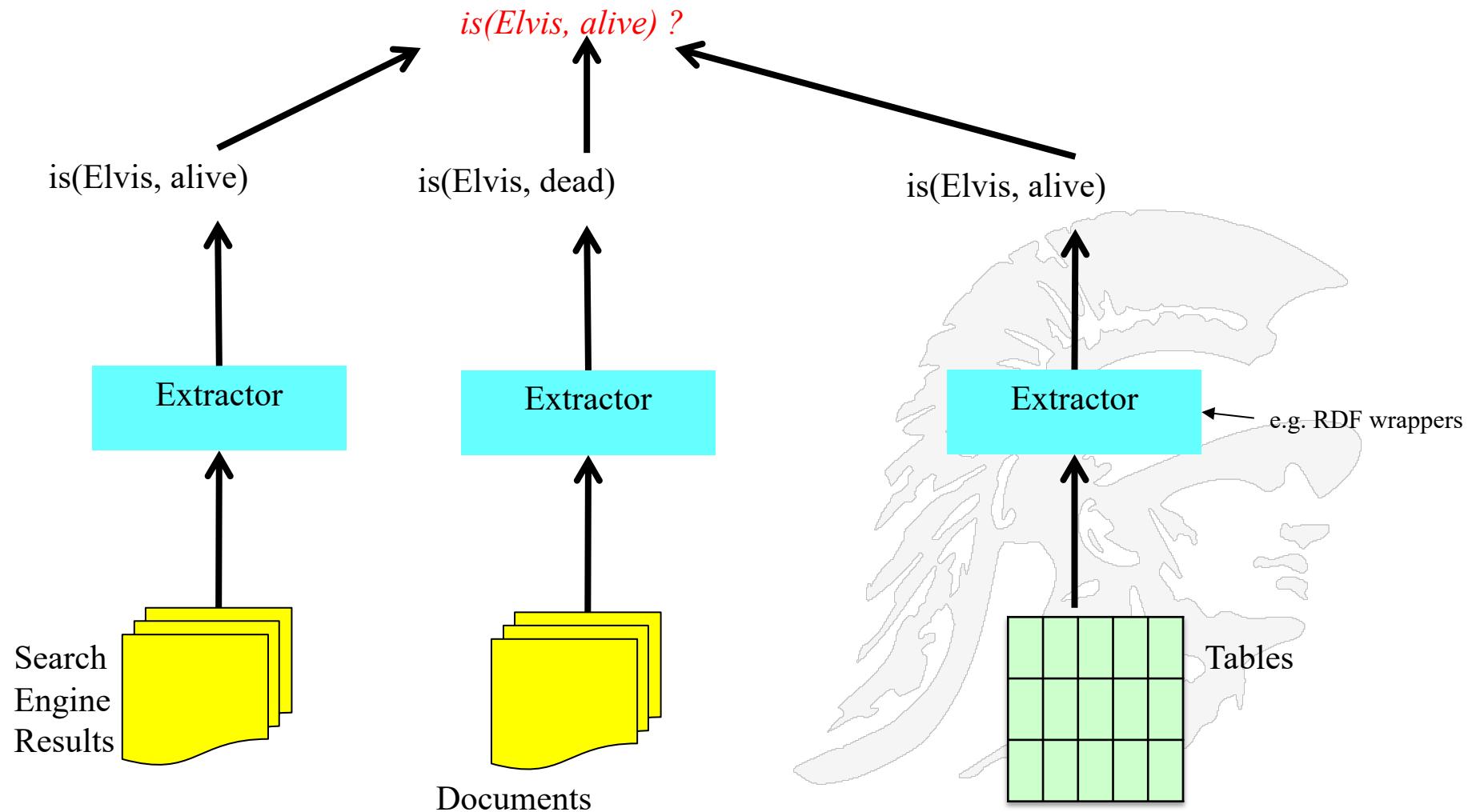
<Bob> <is a> <person>  
 <Bob> <is a friend of> <Alice>  
 <Bob> <is born on> <the 4th of July 1990>  
 <Bob> <is interested in> <the Mona Lisa>  
 <the Mona Lisa> <was created by> <Leonardo da Vinci>  
 <the video 'La Joconde à Washington'> <is about> <the Mona Lisa>



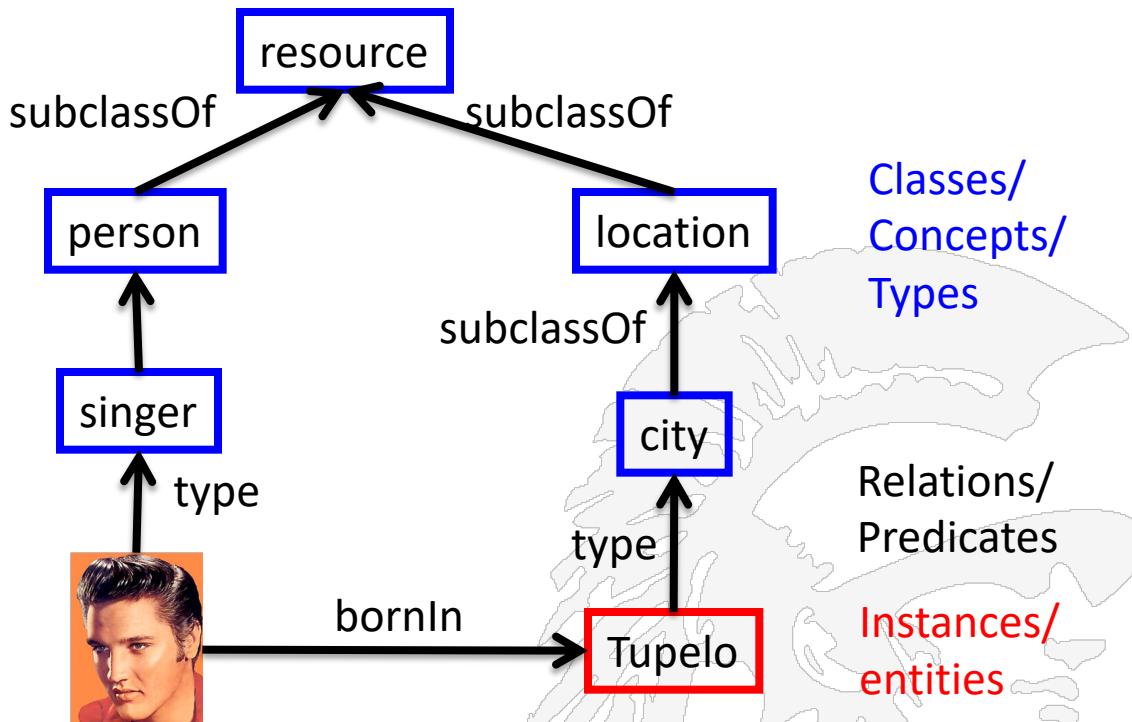
We can visualize triples as a connected **graph**. Graphs consists of nodes and arcs. The subjects and objects of the triples make up the nodes in the graph; the predicates form the arcs

One query language for making inferences on these graphs is SPARQL

## To Answer a Question Knowledgebases Need to Combine Information From *Multiple Sources*



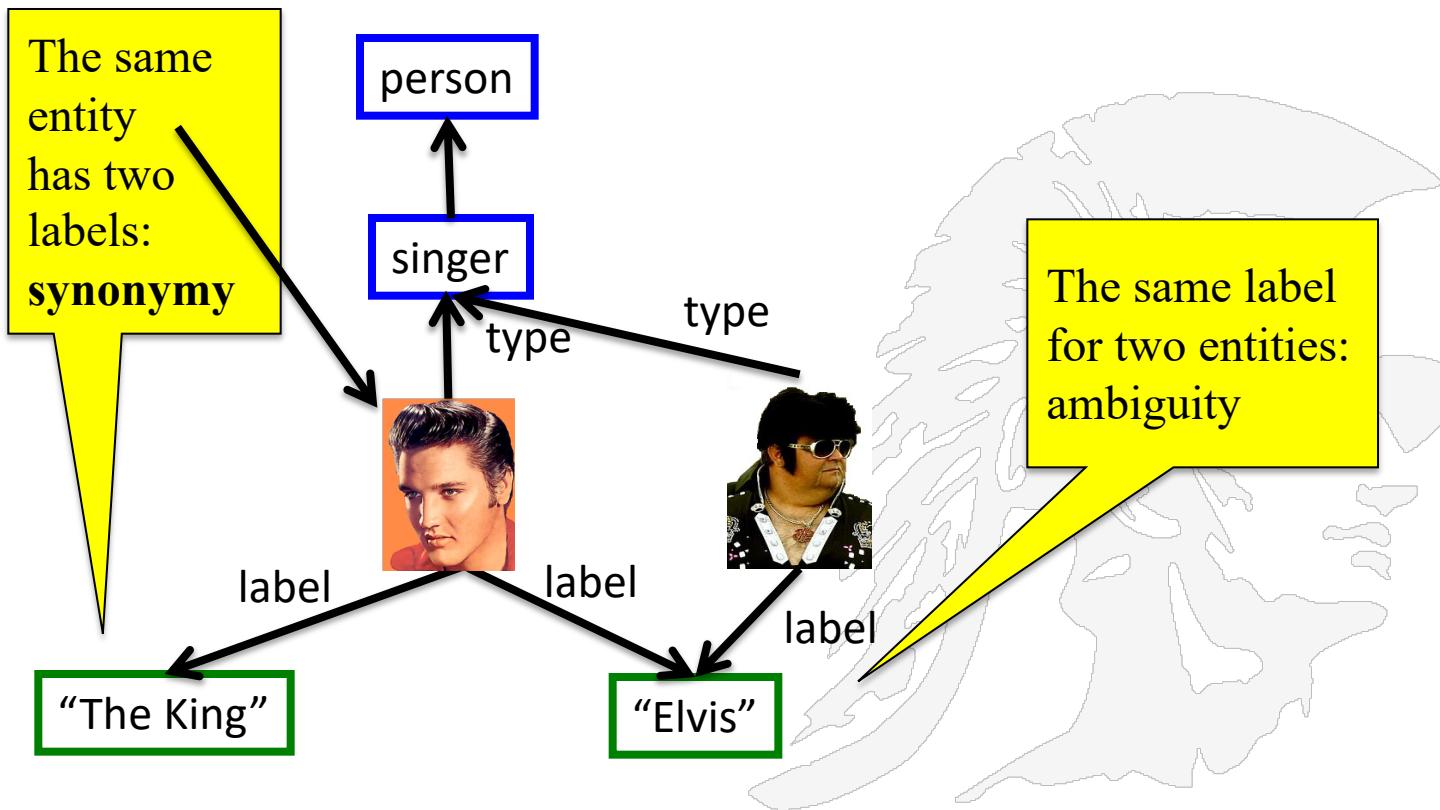
# KnowledgeBases Can Be Represented as Labeled MultiGraphs



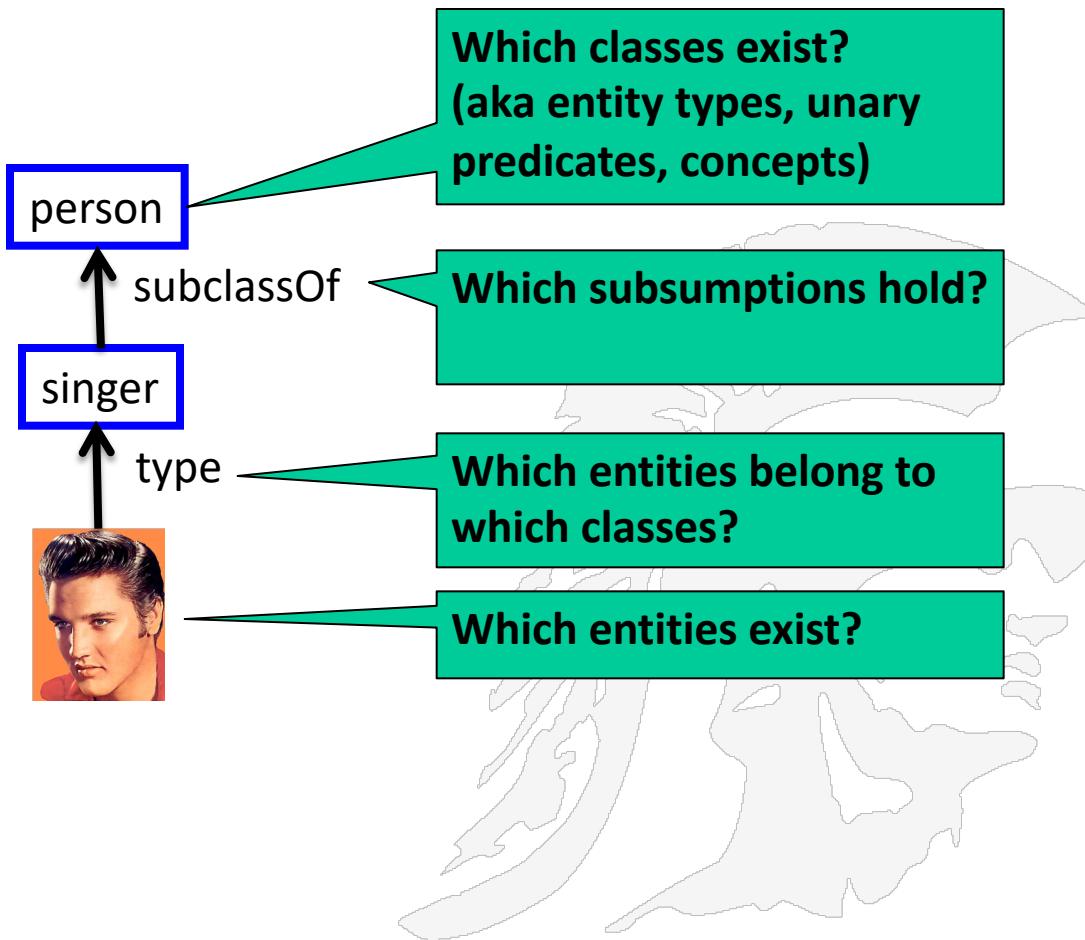
A knowledgebase can be seen as a directed labeled multigraph, where the nodes are entities and the edges relations.

A **multigraph** is a graph which is permitted to have multiple edges that have the same end nodes. Two vertices may be connected by more than one edge

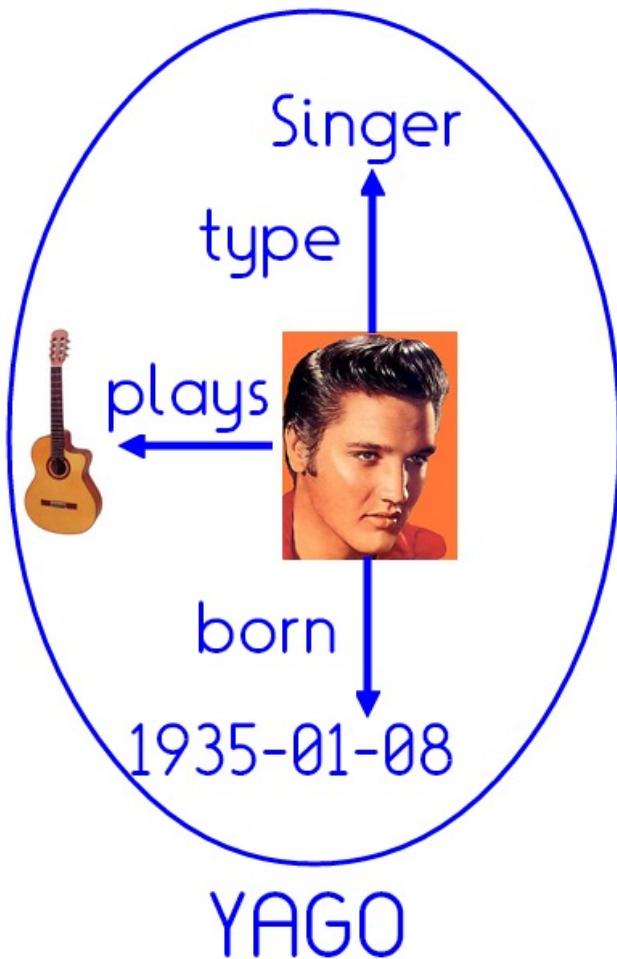
# A Single Entity Can Have Different Labels



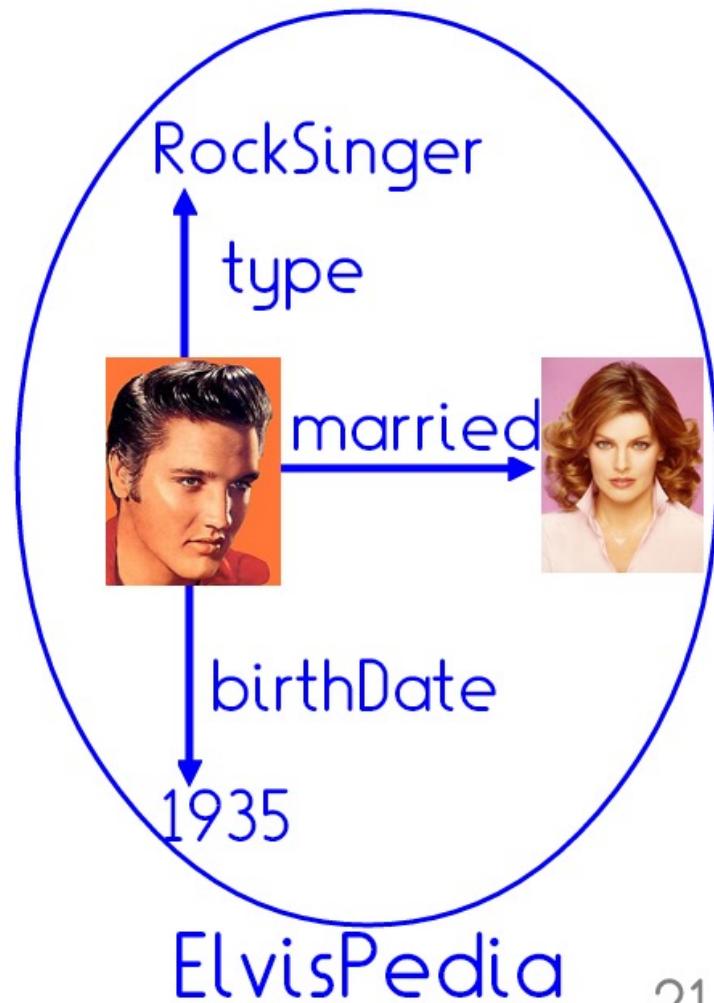
## To Build a Knowledgebase One Must Find Classes and Instances



# Two Knowledgebases With Complementary Information



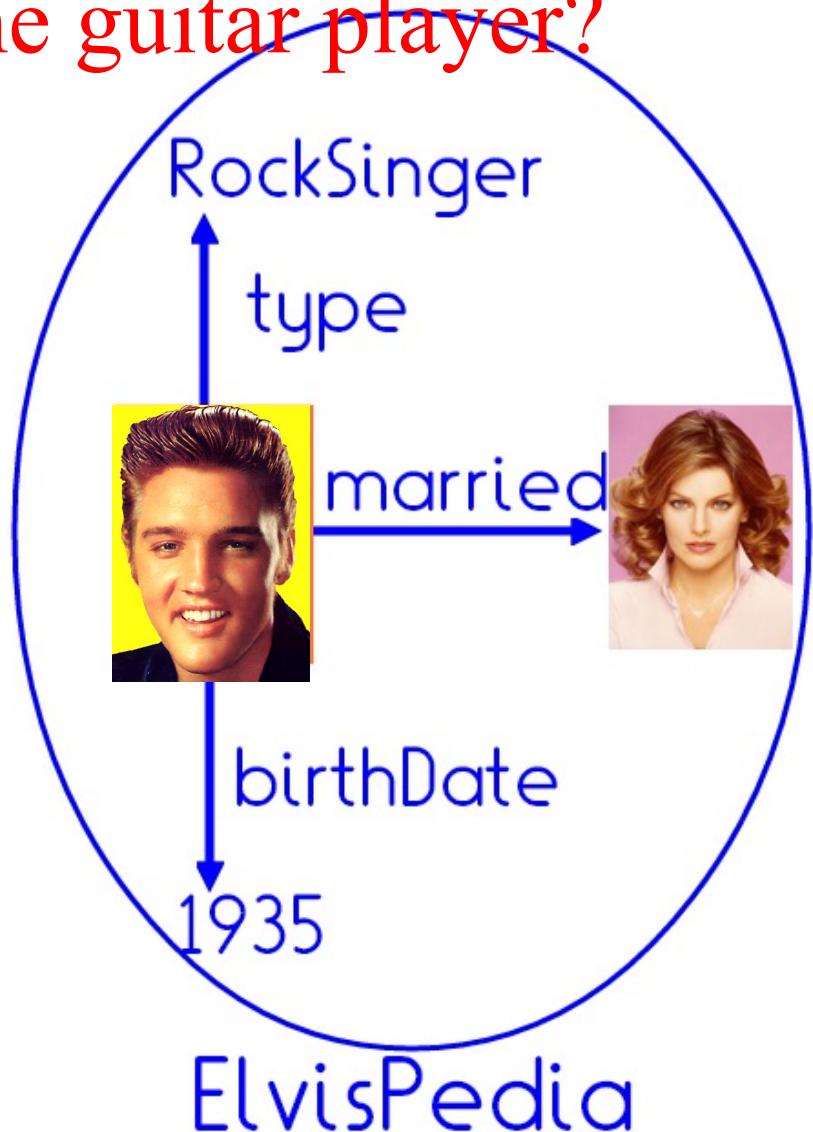
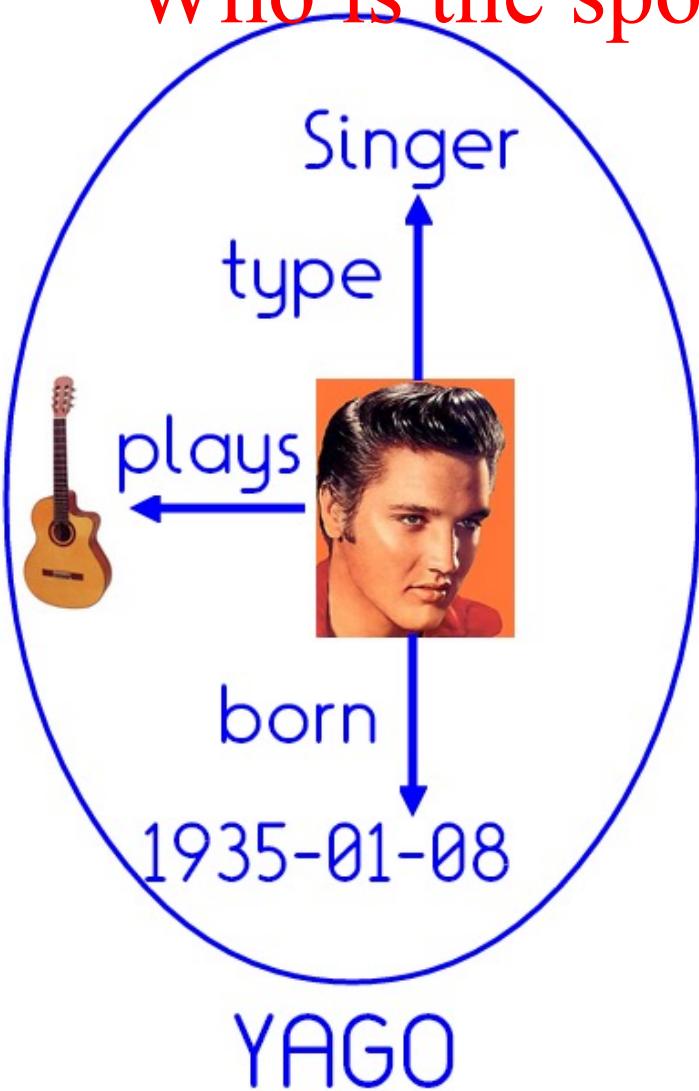
See [https://en.wikipedia.org/wiki/YAGO\\_\(database\)](https://en.wikipedia.org/wiki/YAGO_(database))



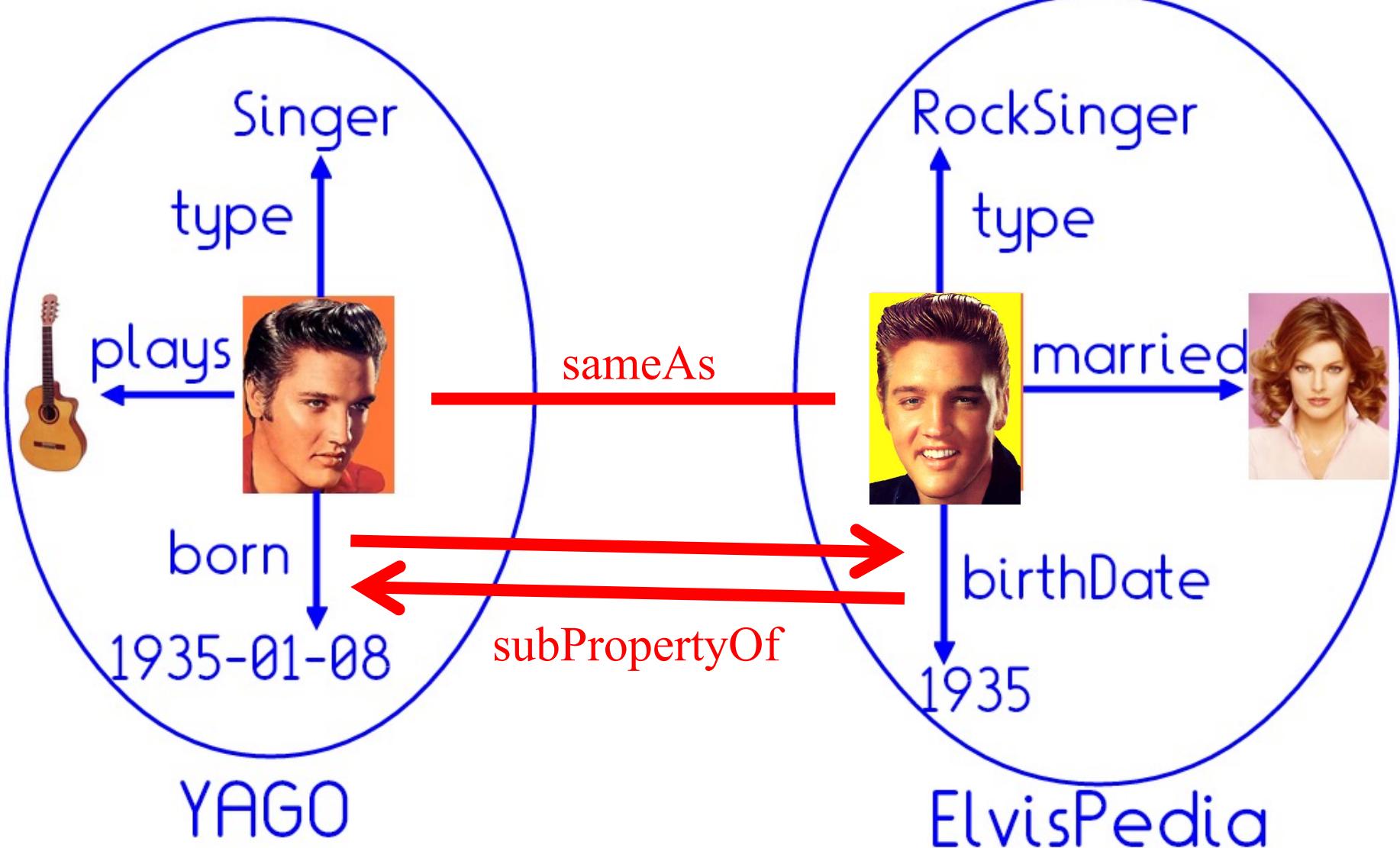
See <https://theelvispedia.com/>

## A Knowledgebase Must Work Across Multiple Ontologies

Who is the spouse of the guitar player?



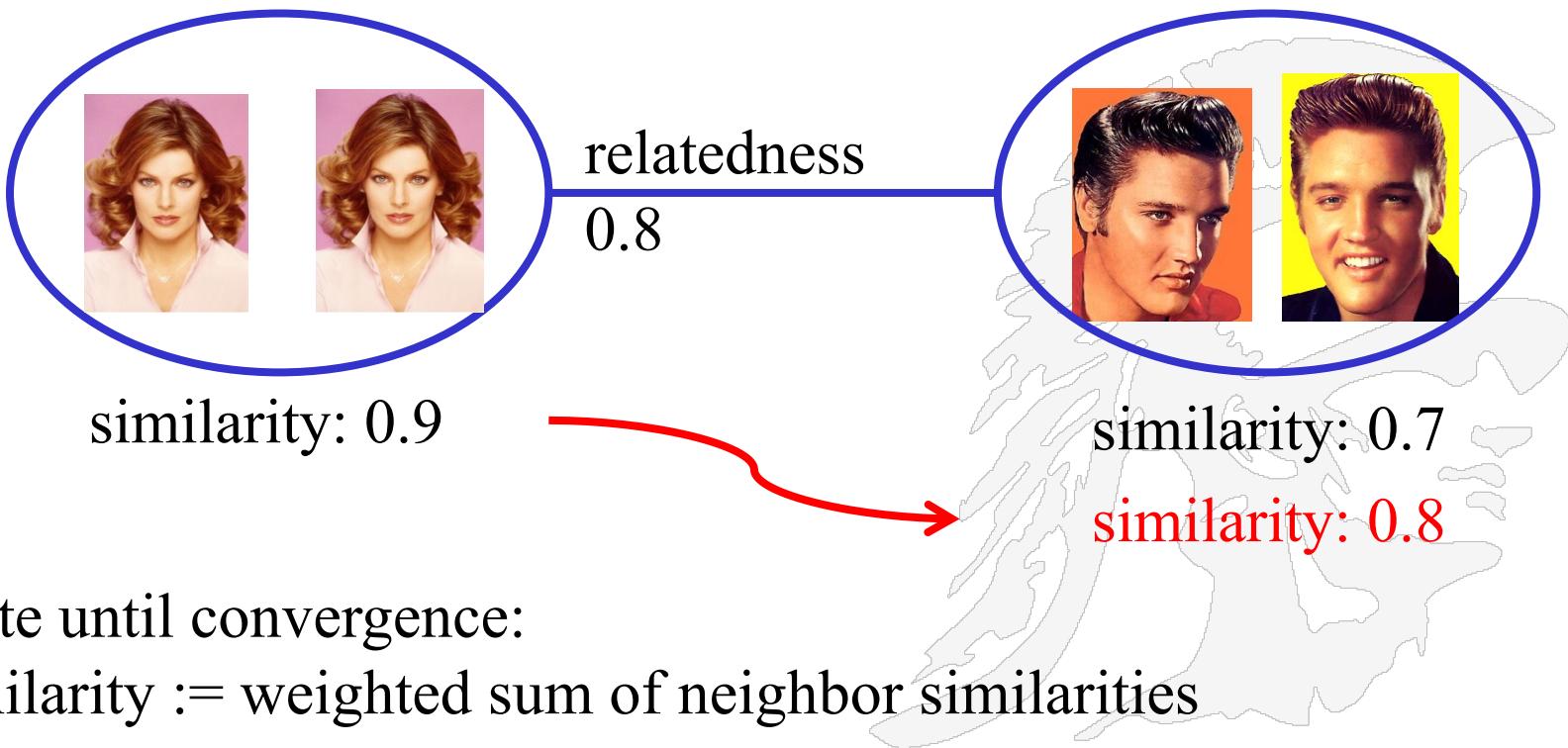
# We Need to Match Entities, Classes and Relations



# Combining Elements From Different Knowledgebases Means Matching Entities

Build a graph:

- nodes: pairs of entities, weighted with similarity
- edges: weighted with degree of relatedness



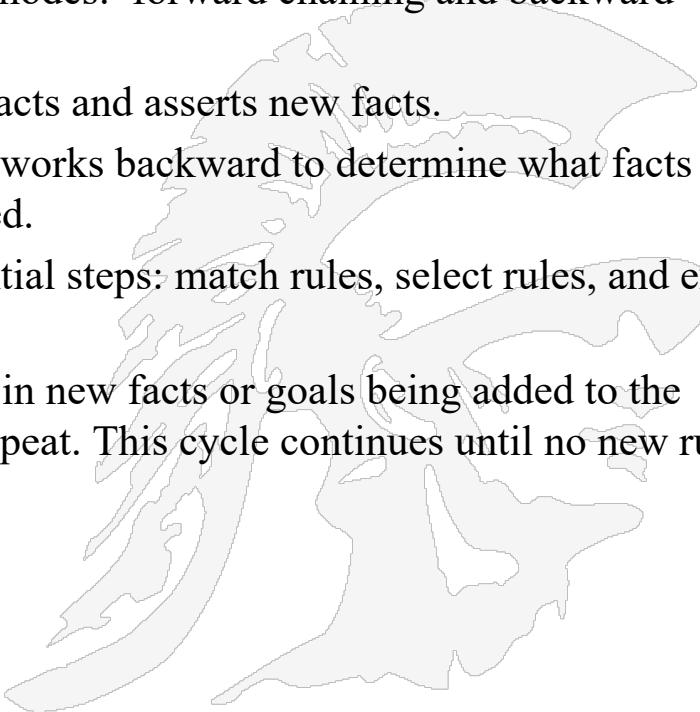
Iterate until convergence:

similarity := weighted sum of neighbor similarities

many variants (belief propagation, label propagation, etc.)

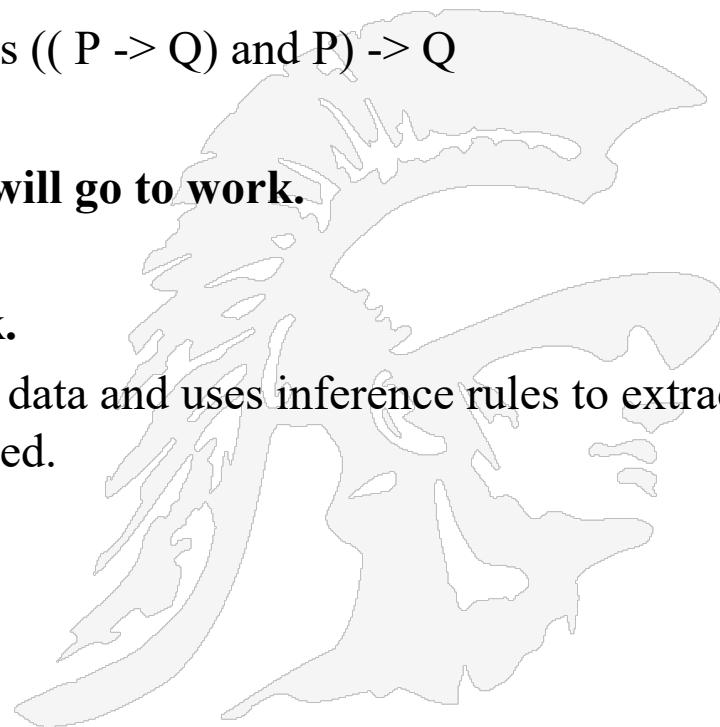
# Inferencing on KnowledgeBases

- An **inference engine** is a component of a system that applies logical rules to a knowledgebase to deduce new information
- This process is ongoing as each new fact in the knowledgebase can trigger additional rules in the inference engine.
- Inference engines work primarily in one of two modes: forward chaining and backward chaining
  - **Forward chaining** starts with the known facts and asserts new facts.
  - **Backward chaining** starts with goals, and works backward to determine what facts must be asserted so that the goals can be achieved.
- An inference engine cycles through three sequential steps: match rules, select rules, and execute rules
- The execution of the rules will sometimes result in new facts or goals being added to the knowledgebase which will trigger the cycle to repeat. This cycle continues until no new rules can be matched
- Search engines typically use forward chaining



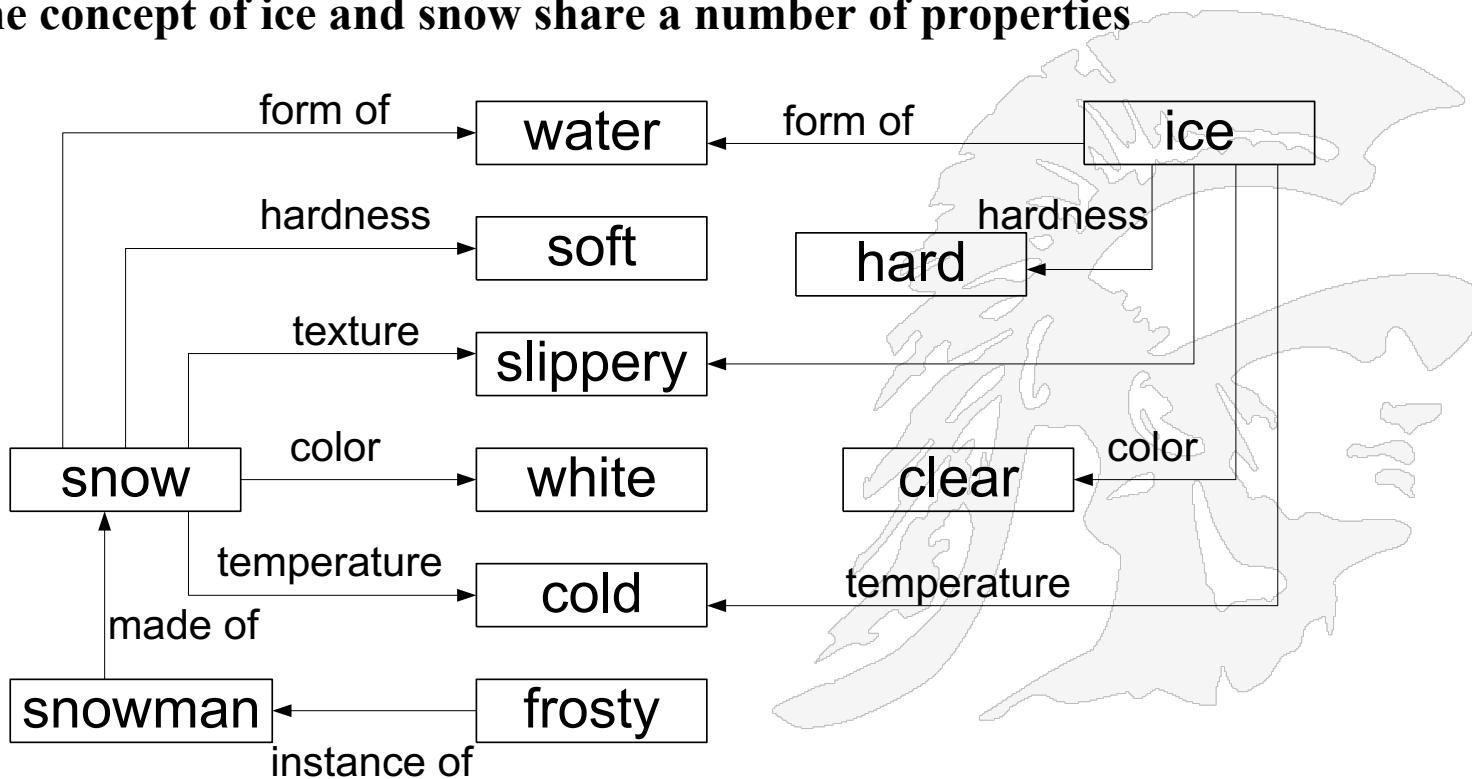
# Forward Chaining

- **Forward chaining** is the repeated application of modus ponens
- In propositional logic, ***modus ponens*** is the rule
  - “ $P$  implies  $Q$ ” and “ $P$ ” are both asserted to be true, so therefore  $Q$  must be true.”
  - Sometimes modus ponens is written as  $(( P \rightarrow Q) \text{ and } P) \rightarrow Q$
  - For Example
    - **If today is Tuesday, then John will go to work.**
    - **Today is Tuesday.**
    - **Therefore, John will go to work.**
- Forward chaining starts with the available data and uses inference rules to extract more data until a goal or endpoint is reached.

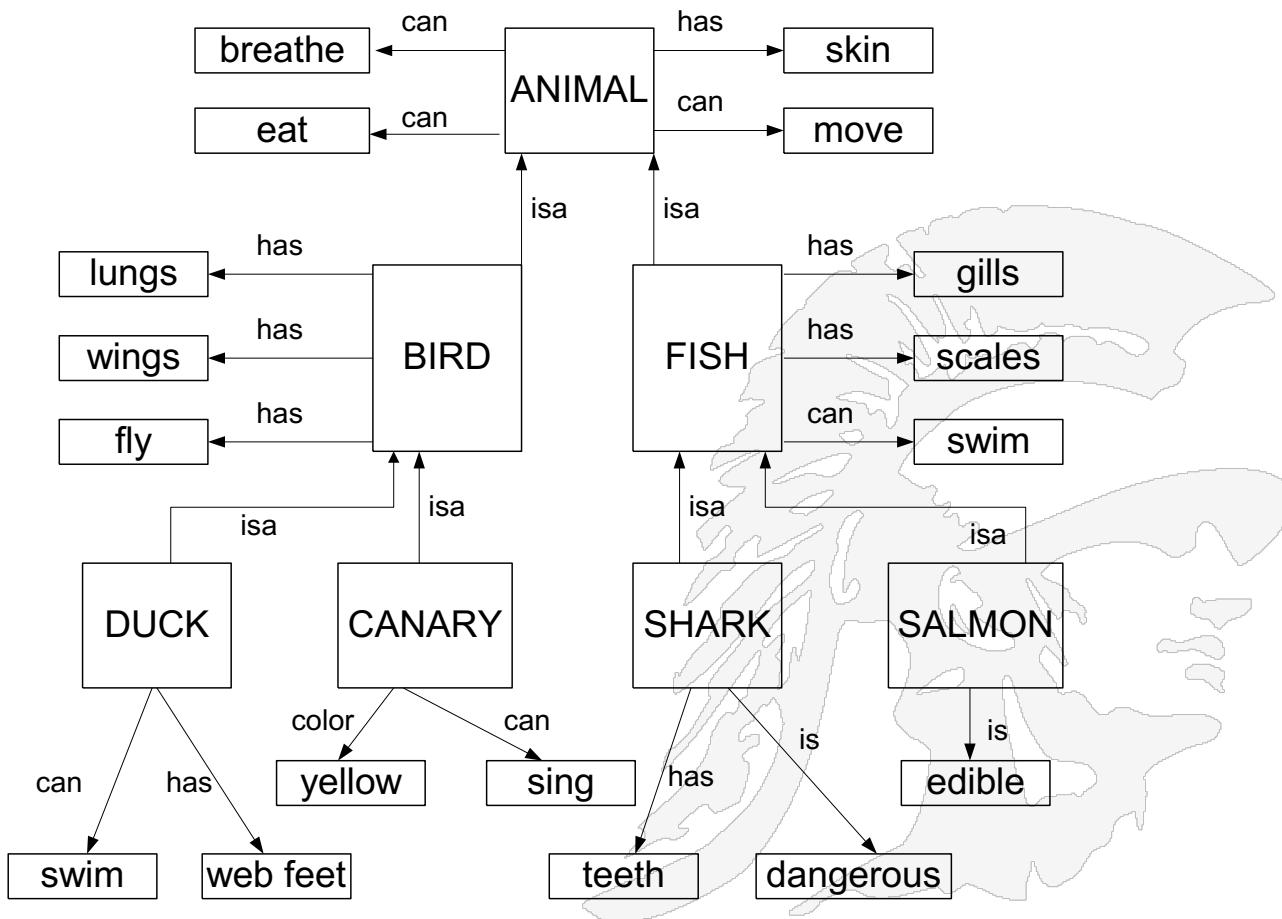


# Semantic Network

- The term semantic network is often used as a synonym for a knowledgebase
- Here is a semantic network that defined the properties of snow and ice
- The concept snowman inherits all the properties of snow
- The concept of ice and snow share a number of properties

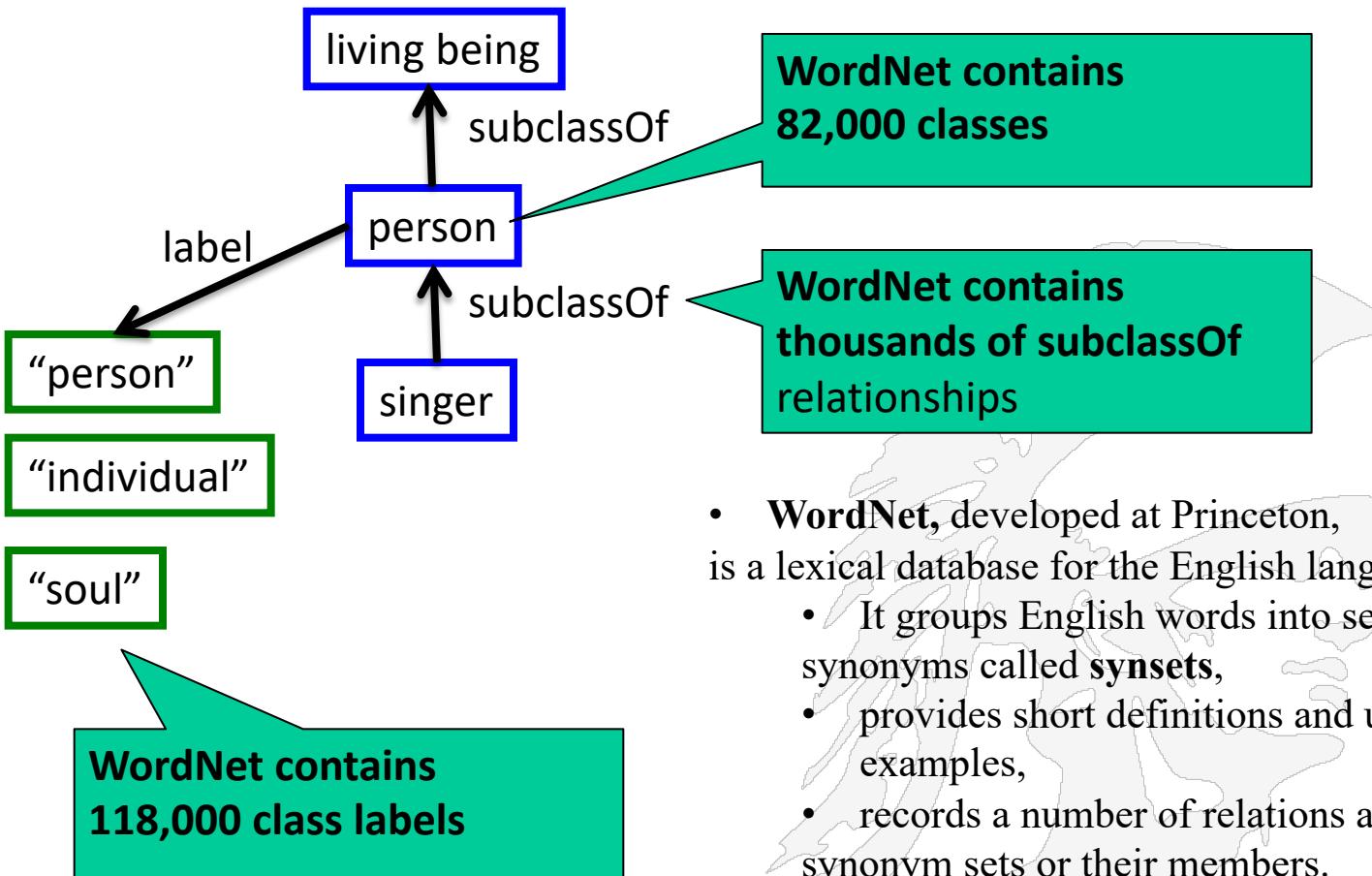


# The is-a Relationship in a Semantic Networks



# WordNet





- WordNet, developed at Princeton, is a lexical database for the English language.
  - It groups English words into sets of synonyms called **synsets**,
  - provides short definitions and usage examples,
  - records a number of relations among these synonym sets or their members.

Lexical means text-only

## WordNet Example: Superclass of Person

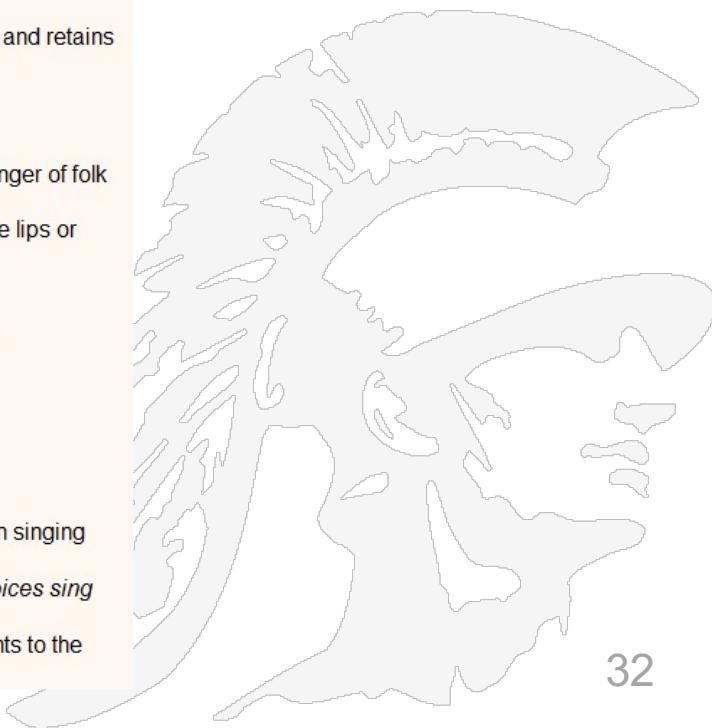
- S: (n) person, individual, someone, somebody, mortal, soul (a human being)  
*"there was too much for one person to do"*
  - direct hyponym / full hyponym
  - part meronym
  - member holonym
  - direct hypernym / inherited hypernym / sister term
    - S: (n) organism, being (a living thing that has (or can develop) the ability to act or function independently)
    - S: (n) living thing, animate thing (a living (or once living) entity)
      - S: (n) whole, unit (an assemblage of parts that is regarded as a single entity) *"how big is that part compared to the whole?"*; *"the team is a unit"*
      - S: (n) object, physical object (a tangible and visible entity; an entity that can cast a shadow) *"it was full of rackets, balls and other objects"*
      - S: (n) physical entity (an entity that has physical existence)
      - S: (n) entity (that which is perceived or known or inferred to have its own distinct existence (living or nonliving))

Note the terms:

- **hyponym**
  - More specific
- **Holonym**
  - Denoting the whole
- **Hypernym**
  - A broad or superordinate

## WordNet Example: Subclass of Singer

- S: (n) singer, vocalist, vocalizer, vocaliser (a person who sings)
  - direct hyponym / full hyponym
    - S: (n) alto (a singer whose voice lies in the alto clef)
    - S: (n) baritone, barytone (a male singer)
    - S: (n) bass, basso (an adult male singer with the lowest voice)
    - S: (n) canary (a female singer)
    - S: (n) caroler, caroller (a singer of carols)
    - S: (n) castrato (a male singer who was castrated before puberty and retains a soprano or alto voice)
    - S: (n) chorister (a singer in a choir)
    - S: (n) contralto (a woman singer having a contralto voice)
    - S: (n) crooner, balladeer (a singer of popular ballads)
    - S: (n) folk singer, jongleur, minstrel, poet-singer, troubadour (a singer of folk songs)
    - S: (n) hummer (a singer who produces a tune without opening the lips or forming words)
    - S: (n) lieder singer (a singer of lieder)
    - S: (n) madrigalist (a singer of madrigals)
    - S: (n) opera star, operatic star (singer of lead role in an opera)
    - S: (n) rapper (someone who performs rap music)
    - S: (n) rock star (a famous singer of rock music)
    - S: (n) songster (a person who sings)
    - S: (n) soprano (a female singer)
    - S: (n) tenor (an adult male with a tenor voice)
    - S: (n) thrush (a woman who sings popular songs)
    - S: (n) torch singer (a singer (usually a woman) who specializes in singing torch songs)
    - S: (n) voice ((metonymy) a singer) "he wanted to hear trained voices sing it"
    - S: (n) warbler (a singer; usually a singer who adds embellishments to the song)



## WordNet Example: Instances But Very Few

- S: (n) [singer](#), [vocalist](#), [vocalizer](#), [vocaliser](#) (a person who sings)
  - [direct hyponym](#) / [full hyponym](#)
  - [has instance](#)
  - S: (n) [Bailey](#), [Pearl Bailey](#), [Pearl Mae Bailey](#) (United States singer (1918-1990))
  - S: (n) [Cash](#), [Johnny Cash](#), [John Cash](#) (United States country music singer and songwriter (1932-2003))
  - S: (n) [Chevalier](#), [Maurice Chevalier](#) (French actor and cabaret singer (1888-1972))
  - S: (n) [Dietrich](#), [Marlene Dietrich](#), [Maria Magdalene von Losch](#) (United States film actress (born in Germany) who made many films with Josef von Sternberg and later was a successful cabaret star (1901-1992))
  - S: (n) [Dylan](#), [Bob Dylan](#) (United States songwriter noted for his protest songs (born in 1941))
  - S: (n) [Fitzgerald](#), [Ella Fitzgerald](#) (United States scat singer (1917-1996))
  - S: (n) [Garland](#), [Judy Garland](#) (United States singer and film actress (1922-1969))
  - S: (n) [Horne](#), [Lena Horne](#), [Lena Calhoun Horne](#) (United States singer and actress (born in 1917))
  - S: (n) [Iglesias](#), [Julio Iglesias](#) (Spanish singer noted for his ballads and love songs (born in 1943))
  - S: (n) [Jackson](#), [Mahalia Jackson](#) (United States singer who did much to popularize gospel music (1911-1972))
  - S: (n) [Jackson](#), [Michael Jackson](#), [Michael Joe Jackson](#) (United States singer who began singing with his four brothers and later became a highly successful star during the 1980s (born in 1958))

only 32 singers !?

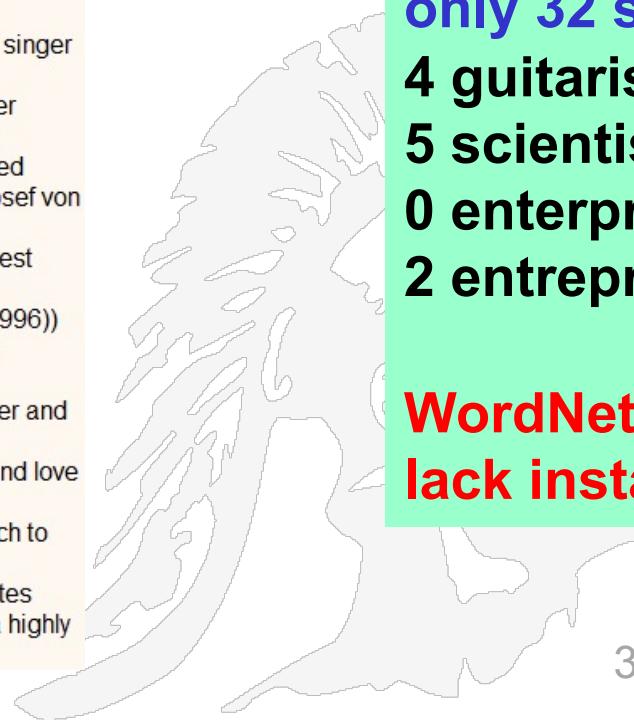
4 guitarists

5 scientists

0 enterprises

2 entrepreneurs

WordNet classes  
lack instances ✎



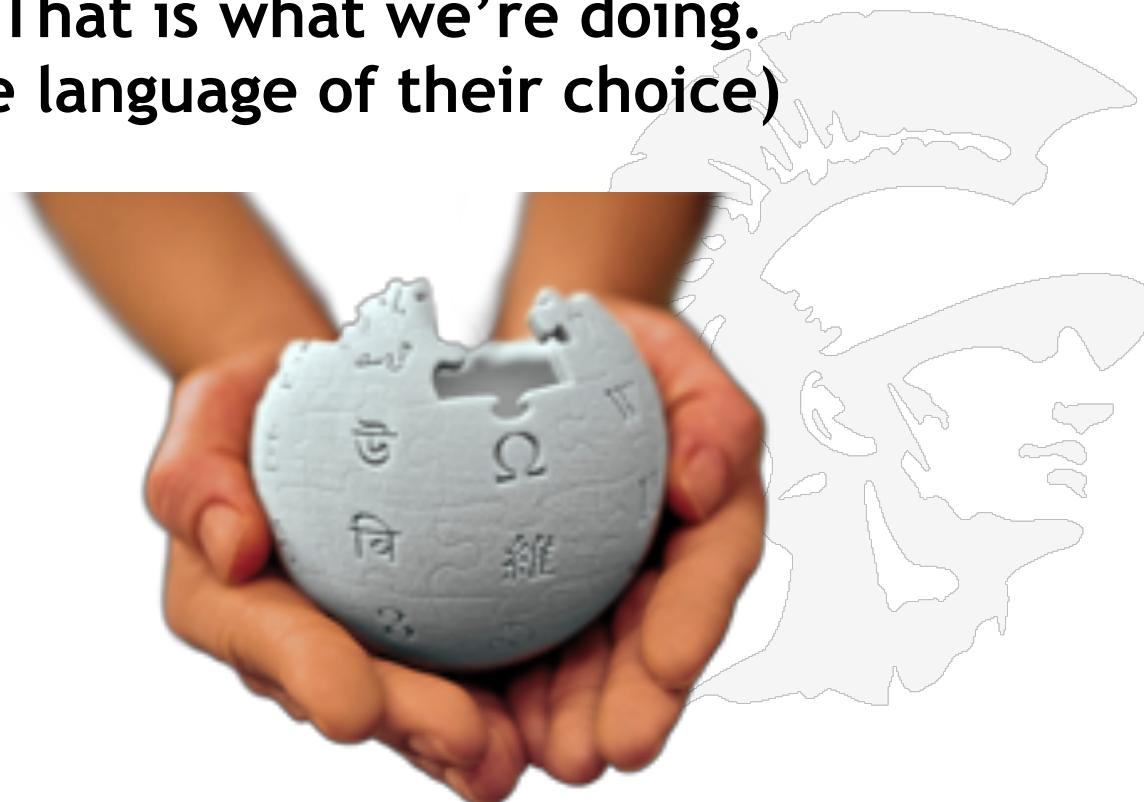
# Wikipedia



## Wikipedia: Transformation from Database to KnowledgeBase

### Wikipedia's Original Mission Statement

“Imagine a world in which every person on the planet shares in the sum of all human knowledge. That is what we’re doing. (for free, in the language of their choice)

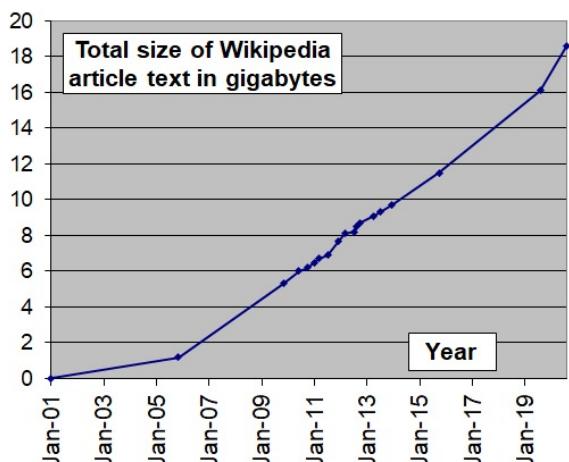


# Wikipedia's Scale

- As of April, 2021 Wikipedia's database when dumped takes up 100GBs compressed, 10TBs uncompressed
- Total wiki pages: 51,000,000+
- Total English articles: 6.1 million
- Unique visitors per month: 500 million
- Monthly mobile page views: 3.7 billion



Jimmy Wales, Founder





## 1. Encyclopaedia

- Notable topics
  - No original research (**NOR**)

## **2. Neutral point of view (NPOV)**

- #### – Verifiability (referencing)

### 3. Free content

- Anyone can edit
  - No copyright infringements

#### 4. Be civil

## 5. No firm rules

# Wikipedia's Five Pillars (5P)

The following screenshot illustrates the five pillars of Wikipedia:

**Pillar 1: Wikipedia is an encyclopedia.**

It incorporates elements of general and specialized [encyclopedias](#), [almanacs](#), and [gazetteers](#). [Wikipedia is not a soapbox, an advertising platform, a vanity press, an experiment in anarchy or democracy, an indiscriminate collection of information, or a web directory](#). It is not a [dictionary](#), a [newspaper](#), or a collection of [source documents](#); that kind of content should be contributed instead to the [Wikimedia sister projects](#).

**Pillar 2: Wikipedia is written from a neutral point of view.**

We strive for articles that document and explain the major [points of view](#) in a balanced and impartial manner. We avoid advocacy and we characterize information and issues rather than debate them. In some areas there may be just one well-recognized point of view; in other areas we describe multiple points of view, presenting each accurately and in context, and not presenting any point of view as "the truth" or "the best view". All [articles](#) must strive for [verifiable accuracy](#): unreferenced material may be removed, so [please provide references](#). Editors' [personal experiences, interpretations, or opinions](#) do not belong here. That means citing [verifiable, authoritative sources](#), especially on controversial topics and when the subject is a [living person](#).

**Pillar 3: Wikipedia is free content that anyone can edit, use, modify, and distribute.**

Respect [copyright laws](#), and do not [plagiarize](#) sources. [Non-free content](#) is allowed under [fair use](#), but strive to find free alternatives to any media or content that you wish to add to Wikipedia. Since all your contributions are [freely licensed to the public](#), no editor [owns any article](#); all of your contributions can and will be mercilessly edited and redistributed.

**Pillar 4: Editors should interact with each other in a respectful and civil manner.**

Respect and be polite to your fellow [Wikipedians](#), even when you disagree. Apply Wikipedia [etiquette](#), and avoid [personal attacks](#). Find [consensus](#), avoid [edit wars](#), and remember that there are 4,143,499 articles on the English Wikipedia to work on and discuss. Act in good faith, and [never disrupt Wikipedia to illustrate a point](#). Be open and [welcoming](#), and [assume good faith](#) on the part of others. When conflict arises, discuss details on the [talk page](#), and follow [dispute resolution](#).

**Pillar 5: Wikipedia does not have firm rules.**

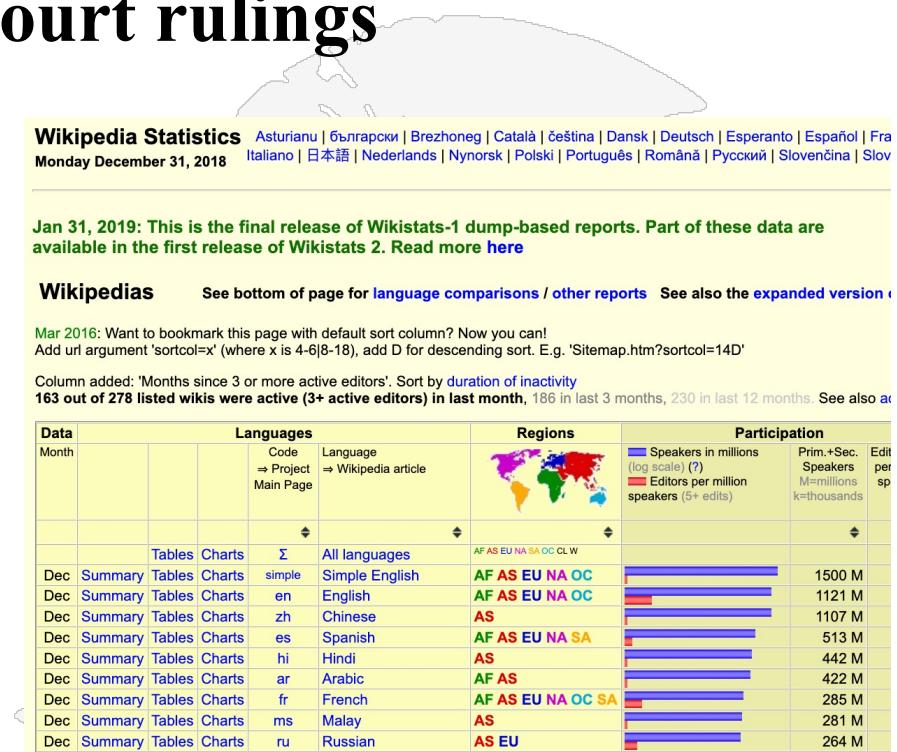
Rules in Wikipedia are not carved in stone, as their wording and interpretation are likely to change over time. The principles and spirit of Wikipedia's rules matter more than their literal wording, and sometimes improving Wikipedia requires making an exception to a rule. [Be bold](#) (but not [reckless](#)) in updating articles and [do not worry about making mistakes](#). [Prior versions of pages are saved](#), so any mistakes can be corrected.

# Wikipedia Statistics

- Among top 10 most visited websites
- 70% of traffic is from search engines
- Cited in hundreds of U.S. court rulings

<https://stats.wikimedia.org/EN/Sitemap.htm>  
and

<https://stats.wikimedia.org/#/all-projects>





## Steve Jobs

From Wikipedia, the free encyclopedia

*For the biography, see [Steve Jobs \(biography\)](#).*

**Steven Paul Jobs** (/dʒɒbz/; February 24, 1955 – October 5, 2011)<sup>[4][5]</sup> was an American businessman and inventor widely recognized as a charismatic pioneer of the personal computer revolution.<sup>[6][7]</sup> He was co-founder, chairman, and chief executive officer of Apple Inc. Jobs also co-founded and served as chief executive of Pixar Animation Studios; he became a member of the board of directors of The Walt Disney Company in 2006, following the acquisition of Pixar by Disney.

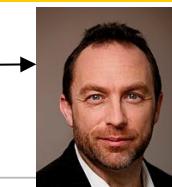
In the late 1970s, Apple co-founder Steve Wozniak engineered one of the first commercially successful lines of personal computers, the Apple II series. Jobs directed its aesthetic design and marketing along with A.C. "Mike" Markkula, Jr. and others. In the early 1980s, Jobs was among the first to see the commercial potential of Xerox PARC's mouse-driven graphical user interface, which led to the creation of the Apple Lisa (engineered by Ken Rothmuller and John Couch) and, one year later, creation of Apple employee Jef Raskin's Macintosh.

After losing a power struggle with the board of directors in 1985, Jobs left Apple and founded NeXT, a computer platform development company specializing in the higher-education and business markets. NeXT was eventually acquired by Apple in 1996, which brought Jobs back to the company he co-founded, and provided Apple with the NeXTSTEP codebase, from which the Mac OS X was developed.<sup>[8]</sup> Jobs was named Apple advisor in 1996, interim CEO in 1997, and CEO from 2000 until his resignation. He oversaw the development of the iMac, iTunes, iPod, iPhone, and iPad and the company's Apple Retail Stores.<sup>[9]</sup> In 1986, he acquired the computer graphics division of Lucasfilm Ltd, which was spun off as Pixar Animation Studios.<sup>[10]</sup> He was credited in *Toy Story* (1995) as an executive producer. He remained CEO and majority shareholder at 50.1 percent until its acquisition by The Walt Disney Company in 2006,<sup>[11]</sup> making Jobs Disney's largest individual shareholder at seven percent and a member of Disney's Board of Directors.<sup>[12][13]</sup>

In 2003, Jobs was diagnosed with a pancreas neuroendocrine tumor. Though it was initially treated, he reported a hormone imbalance, underwent a liver transplant in 2009, and appeared progressively thinner as his health declined.<sup>[14]</sup> On medical leave for most of 2011, Jobs resigned as Apple CEO in August that year and was elected Chairman of the Board. On October 5, 2011, Jobs died of respiratory arrest related to his metastatic tumor. He

# Wikipedia is a Rich Source of Instances

Wikipedia founders



Jimmy  
Wales



Larry  
Sanger

Steve Jobs



Jobs holding a white iPhone 4 at Worldwide Developers Conference 2010

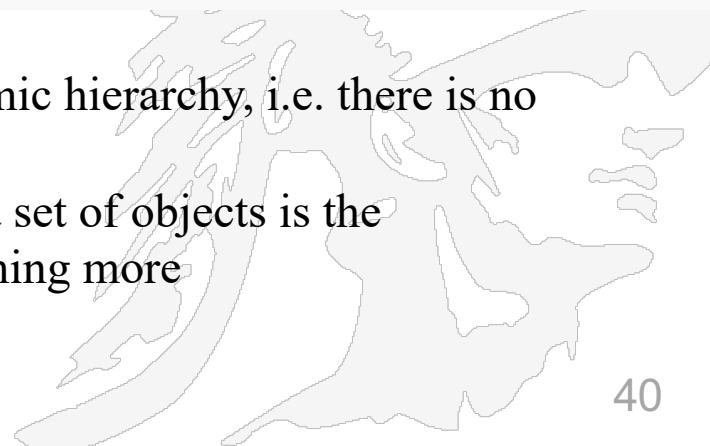
<b>Born</b>	Steven Paul Jobs February 24, 1955 <sup>[1][2]</sup> San Francisco, California, U.S. <sup>[1][2]</sup>
<b>Died</b>	October 5, 2011 (aged 56) <sup>[2]</sup> Palo Alto, California, U.S.
<b>Nationality</b>	American
<b>Alma mater</b>	Reed College (dropped out)

## Wikipedia's Categories Also Contain Classes

Categories: Steve Jobs | 1955 births | 2011 deaths | American adoptees | American billionaires  
| American chief executives | American computer businesspeople | American industrial designers  
| American inventors | American people of German descent | American people of Swiss descent  
| American people of Syrian descent | American technology company founders | American Zen Buddhists  
| Apple Inc. | Apple Inc. employees | Businesspeople from California | Businesspeople in software  
| Cancer deaths in California | Computer designers | Computer pioneers | Deaths from pancreatic cancer  
| Disney people | Internet pioneers | National Medal of Technology recipients | NeXT  
| Organ transplant recipients | People from the San Francisco Bay Area | Pescetarians  
| Reed College alumni

But categories do not form a taxonomic hierarchy, i.e. there is no ISA hierarchy

An isa hierarchy only specifies that a set of objects is the subclasses of another object, but nothing more



# Structure of a Wikipedia Page

- **Types of links**
  - **Article links**
    - links from one article to another of the same language;
  - **Category links**
    - links from an article to special “Category” pages;
  - **Interlingual links**
    - links from an article to a presumably equivalent, article in another language;
- **Types of special pages**
  - **Redirect pages**
    - short pages which often provide equivalent names for an entity
  - **Disambiguation pages**
    - a page with little content that links to multiple similarly named articles.
- **Infoboxes, templates, list pages, wikipedia commons, ...**

**WIKIPEDIA**  
The Free Encyclopedia

Main page  
Contents  
Featured content  
Current events  
Random article  
Donate to Wikipedia  
Wikipedia store

Interaction  
Help  
About Wikipedia  
Community portal  
Recent changes  
Contact page

Tools  
What links here  
Related changes  
Upload file  
Special pages  
Permanent link  
Page information  
Wikidata item  
Edit this page

Print/export  
Create a book  
Download as PDF  
Printable version

In other projects  
Wikimedia Commons  
Wikiquote

Languages  
Afrikaans  
Alemannisch  
ՀՊԵՐ  
Ænglisc  
العربية  
Aragonés  
Asturianu  
★ Azərbaycanca  
تۆرکجه  
ଓଡ଼ିଆ  
Bân-lâm-gú  
Башҡортса  
Беларуская  
Беларуская  
(тарашкевіца)

Elvis Presley

From Wikipedia, the free encyclopedia  
(Redirected from Elvis presley)

“Elvis” redirects here. For other uses, see [Elvis \(disambiguation\)](#).

**Elvis Aaron Presley**<sup>[1]</sup> January 8, 1935 – August 16, 1977 was an American rock musician and actor. Regarded as one of the most significant cultural icons of the 20th century, he is often referred to as “the King of Rock and Roll”, or simply, “the King”.

Presley was born in Tupelo, Mississippi, as a twinless twin—his brother was stillborn. When he was 13 years old, he and his family relocated to Memphis, Tennessee. His music career began there in 1954, when he recorded a song with producer Sam Phillips at Sun Records. Accompanied by guitarist Scotty Moore and bassist Bill Black, Presley was an early popularizer of rockabilly, an up-tempo, backbeat-driven fusion of country music and rhythm and blues. RCA Victor acquired his contract in a deal arranged by Colonel Tom Parker, Presley's manager since 1954. His first RCA single, “Heartbreak Hotel”, was released in January 1956 and became a number-one hit in the United States. He was regarded as the leading figure of rock and roll after a series of successful network television appearances and chart-topping records. His energetic interpretation of songs and sexual provocative performance style, combined with a singularly potent mix of influences across color lines that coincided with the dawn of the Civil Rights Movement, made him enormously popular—also controversial!

In November 1956, he made his film debut in *Love Me Tender*. In 1958, he was drafted into military service. He resumed his recording career two years later, producing some of his most commercially successful work before devoting much of the 1960s to making Hollywood films and their accompanying soundtracks, albums, most of which were critically panned. In 1968, following a seven-year break from performances, he returned to the stage in the acclaimed televised comeback special *Elvis*, which led to an extended Las Vegas concert residency and a string of highly profitable tours. In 1973, Presley was featured in the first globally broadcast concert via satellite, *Aloha from Hawaii*. Several years of prescription drug abuse severely damaged his health, and he died in 1977 at the age of 42.

Presley is one of the most celebrated and influential musicians of the 20th century. Commercially successful in many genres, including pop, blues and gospel, he was the best-selling solo artist in the history of recorded music<sup>[2][3][4]</sup> with estimated record sales of around 600 million units worldwide.<sup>[5]</sup> He won three Grammys, also receiving the Grammy Lifetime Achievement Award at age 36, and has been inducted into multiple music halls of fame.

Contents [hide]

1 Life and career  
1.1 1935–53: Early years  
1.2 1953–55: First recordings  
1.3 1955–58: Commercial breakthrough and controversy  
1.4 1958–60: Military service and mother's death  
1.5 1960–67: Focus on films  
1.6 1968–73: Comeback  
1.7 1973–77: Health deterioration and death

• Elvis Presley@ at the Internet Movie Database  
• Elvis Presley@ at the TCM Movie Database  
• Elvis Presley@ at the CMT Movie Database

• Elvis Presley Enterprises@ official site of the Elvis Presley brand  
• Elvis Presley@ Discogs  
• Elvis The Music@ official record label site

• Elvis Presley@ at AllMovie  
• Elvis Presley@ at AllMusic  
• Elvis Presley@ at Wikipedia

• Elvis Presley@ on official site of the Elvis Presley Australia site

• Elvis Presley@ on officially sanctioned Elvis Australia site

• Elvis Presley@ at DMOZ

Elvis Presley

Country Music Hall of Fame 1990s

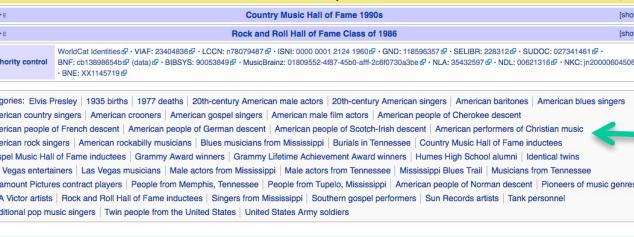
Rock and Roll Hall of Fame Class of 1996

WorldCat identifier# - VIAF: 23424830# - LCCN: n76079487# - ISBN: 0000 0001 2124 1960# - GND: 1165920# - SELIBR: 23831# - SLUB: 027341461# - BNF: cb33998546# (data) - BIBSYS: 3005384# - MusicBrainz: 01090532-487-4520-aff-2c6f7303a3b# - NLA: 35432597# - NDL: 00521316# - NKC: jr20000604506# - BNE: XX1145719#

Categories: Elvis Presley | 1935 births | 1977 deaths | 20th-century American male actors | 20th-century American singers | American baritones | American blues singers | American country singers | American crooners | American gospel singers | American male film actors | American people of Cherokee descent | American people of French descent | American people of German descent | American people of Scotch-Irish descent | American performers of Christian music | American rock singers | American rockabilly musicians | Blues musicians from Mississippi | Burials in Tennessee | Country Music Hall of Fame inductees | Gospel Music Hall of Fame inductees | Grammy Award winners | Grammy Lifetime Achievement Award winners | Humes High School alumni | Identical twins | Las Vegas entertainers | Las Vegas musicians | Male actors from Mississippi | Male actors from Tennessee | Mississippi Blues Trail | Musicians from Tennessee | Paramount Pictures contract players | People from Memphis, Tennessee | People from Tupelo, Mississippi | American people of Norman descent | Pioneers of music genres | RCA Victor artists | Rock and Roll Hall of Fame inductees | Singers from Mississippi | Southern gospel performers | Sun Records artists | Tank personnel | Traditional pop music singers | Twin people from the United States | United States Army soldiers

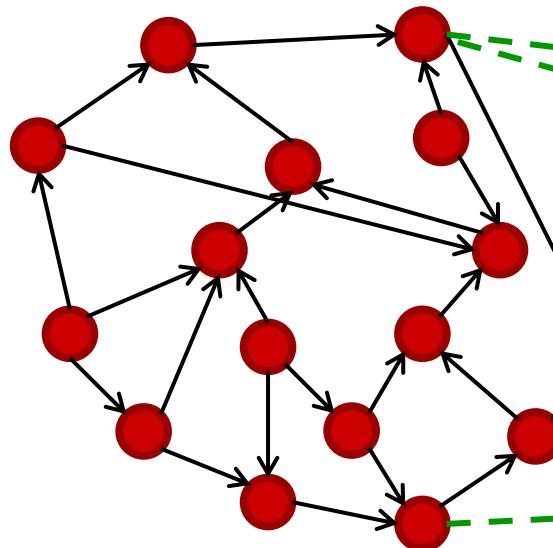
This page was last modified on 24 September 2016, at 13:03.


 infobox

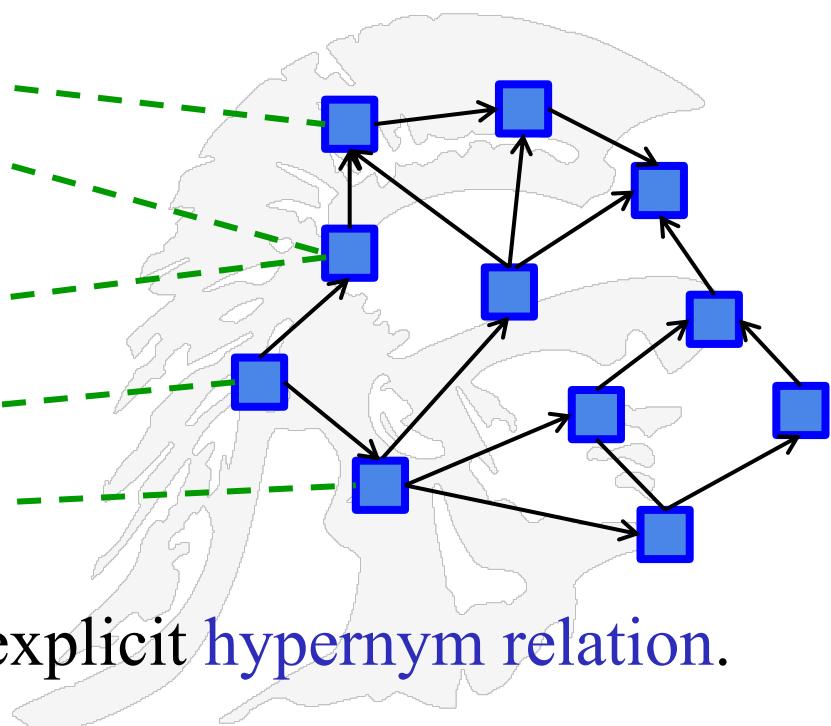

 Category links

# Combining The Wikipedia Classes and Instances

Article pages  
~4M



Category pages  
~700K



Two noisy graphs with no explicit hypernym relation.

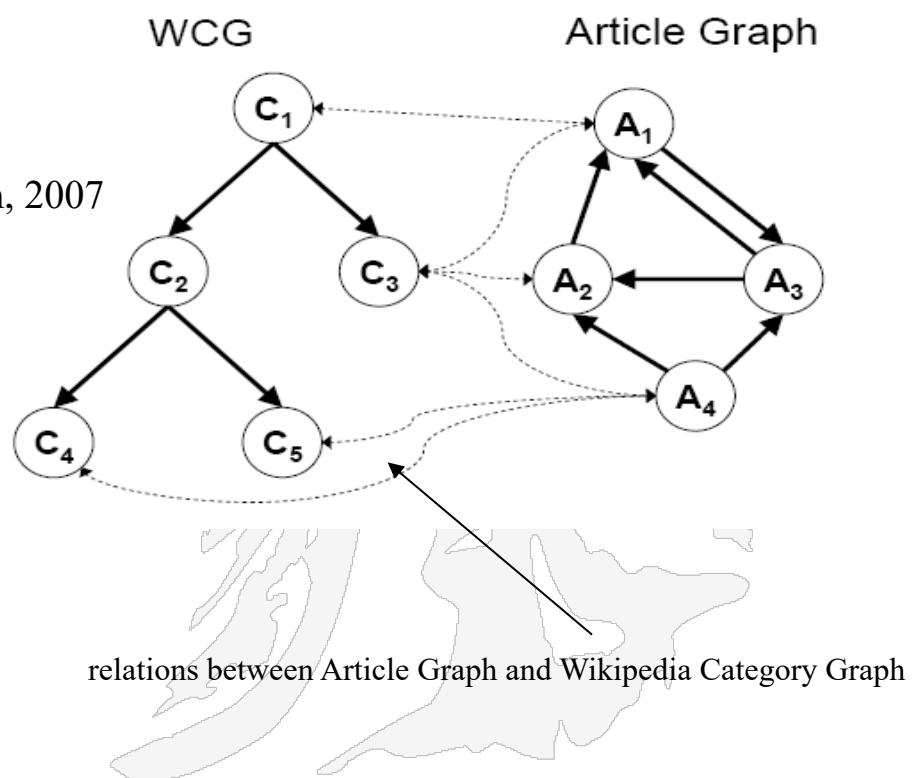
# Organization of Wikipedia

Wikipedia **articles** form a network of semantically related terms, while the **categories** are organized in a taxonomy-like structure called Wikipedia Category Graph (WCG)

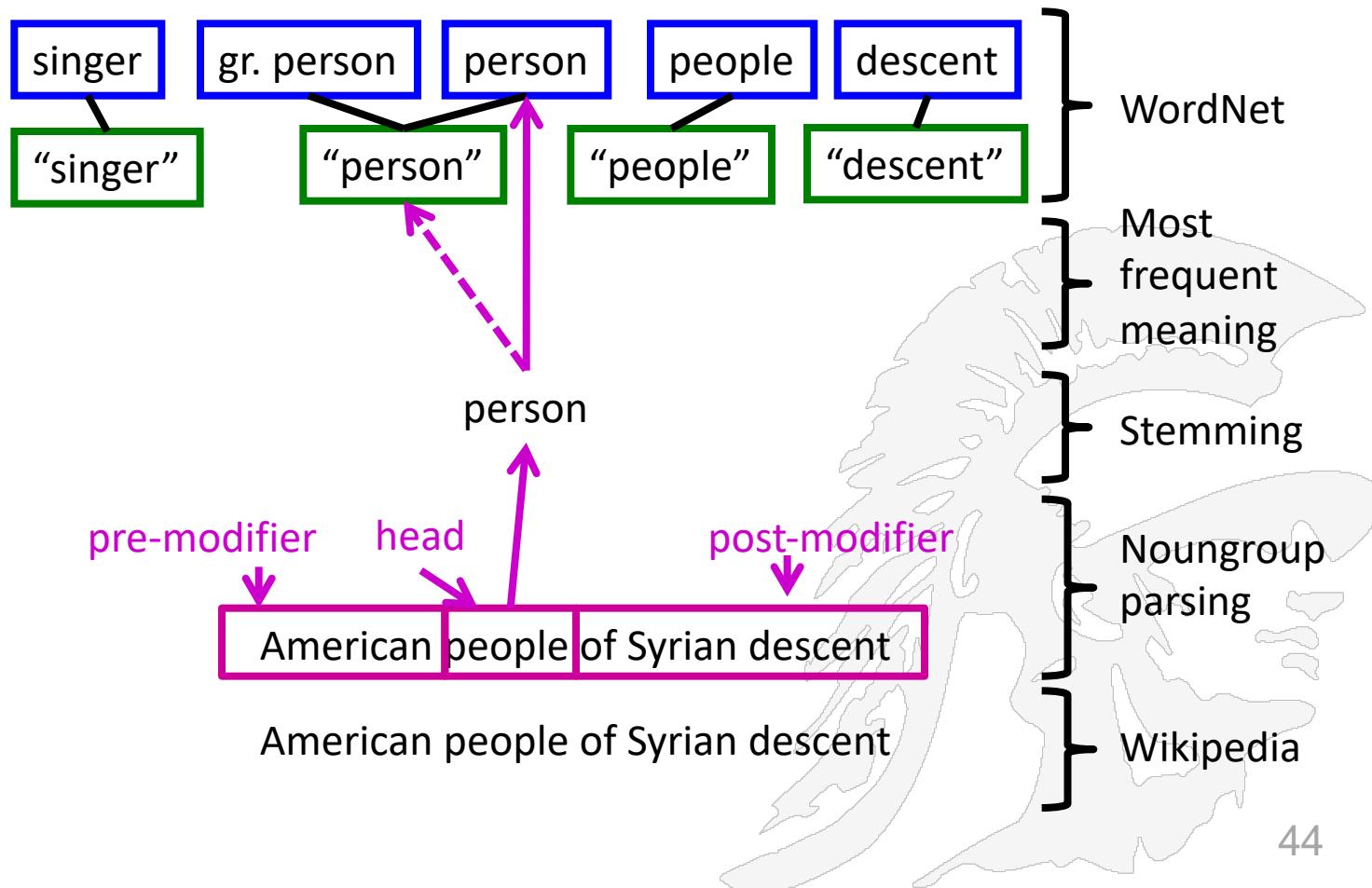
See the article by Torsten Zesch and Iryna Gurevych, 2007

Wikipedia Article Graph, WAG

Wikipedia Category Graph, WCG



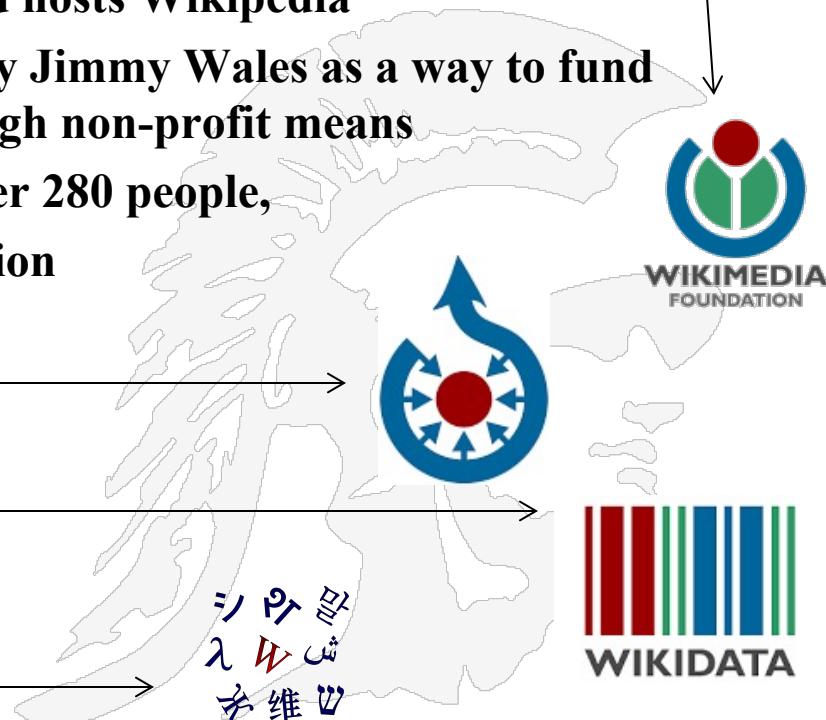
# Combining Wikipedia Named Entities to WordNet Synsets



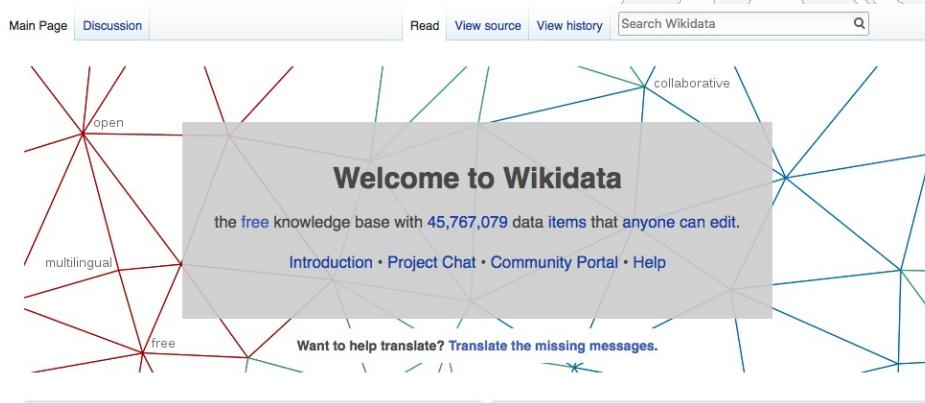
See “*Mapping WordNet synsets to Wikipedia*” Fernando & Stevenson, and  
“*Mapping WordNet Instances to Wikipedia*”, McCrae

# Wikimedia Foundation

- **Wikimedia Foundation, Inc. (WMF)** is an American non-profit and charitable organization headquartered in San Francisco.
- It owns the internet domain names and hosts Wikipedia
- The foundation was founded in 2003 by Jimmy Wales as a way to fund Wikipedia and its sister projects through non-profit means
- As of 2015, the foundation employs over 280 people, with annual revenues in excess of \$75 million
- Related projects to Wikipedia:
  - Commons for multimedia,
  - Wiktionary as free dictionary, and
  - Wikidata for structured data.

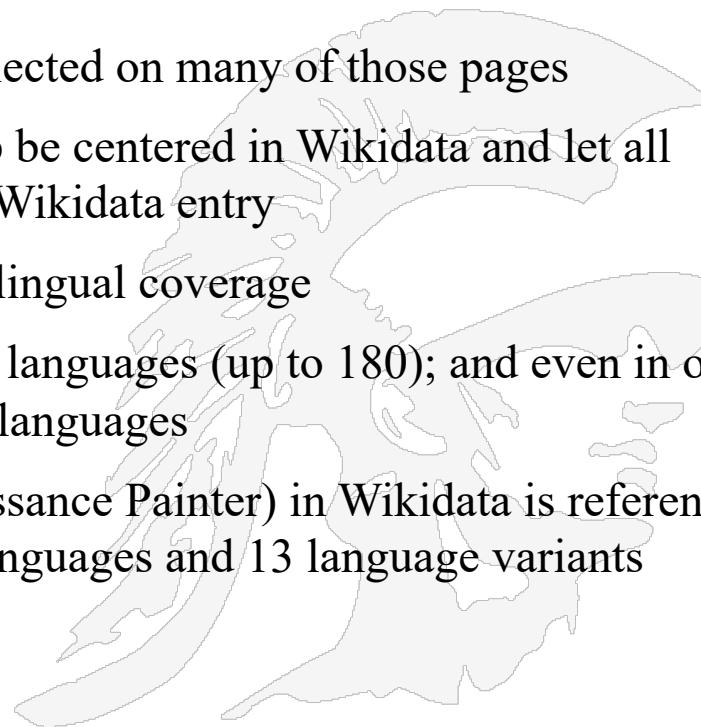


- WikiData is an effort to convert the Wikipedia data into a knowledgebase
- WikiData aims to create a free RDF-like KB about the world that can be read/edited by humans & machines
- Wikidata clients use the repository, e.g. to populate Web pages or Wikipedia infoboxes
- WikiData increases the quality and lowers the maintenance costs of Wikipedia and related projects



# Wikidata Multilingual Coverage

- The challenge: Wikipedia has many named entities that occur in numerous articles
  - E.g Ex-President Obama or President Trump are mentioned in over 100 articles
  - If one of them dies, this must be reflected on many of those pages
- Solution: Let the entry for Obama/Trump be centered in Wikidata and let all references to Obama/Trump point to the Wikidata entry
- Another aspect of Wikidata is their multilingual coverage
  - Popular entities are present in many languages (up to 180); and even in one Wikipedia page there may be many languages
  - E.g. Lucas Cranach (German Renaissance Painter) in Wikidata is referenced in 57 language tags, representing 44 languages and 13 language variants



# Wikipedia Page and WikiData Page for Douglas Adams

Article Talk Read Edit View history Search Wikipedia Q

## Douglas Adams

From Wikipedia, the free encyclopedia  
(Redirected from [Douglas Adams](#))

For other people named Douglas Adams, see [Douglas Adams \(disambiguation\)](#).

Douglas Noel Adams (11 March 1952 – 11 May 2001) was an English author, scriptwriter, essayist, humorist, satirist and dramatist.

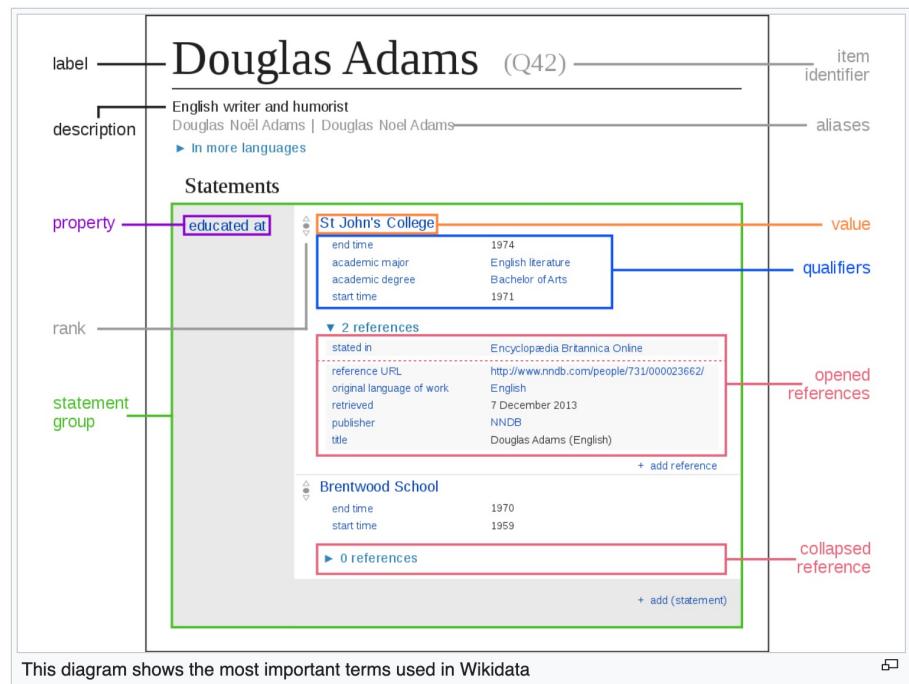
Adams was author of *The Hitchhiker's Guide to the Galaxy*, which originated in 1978 as a BBC radio comedy before developing into a "trilogy" of five books that sold more than 15 million copies in his lifetime and generated a television series, several stage plays, comics, a computer game, and in 2005 a feature film. Adams's contribution to UK radio is commemorated in [The Radio Academy's Hall of Fame](#).<sup>[1]</sup>

Adams also wrote *Dirk Gently's Holistic Detective Agency* (1987) and *The Long Dark Tea-Time of the Soul* (1988), and co-wrote *The Meaning of Life* (1983), *The Deeper Meaning of Life* (1990), *Last Chance to See* (1990), and three stories for the television series *Doctor Who*; he also served as script editor for the show's seventeenth season in 1979. A posthumous collection of his works, including an unfinished novel, was published as *The Salmon of Doubt* in 2002.

Adams was an advocate for environmentalism and conservation, a lover of fast cars,<sup>[2]</sup> technological innovation and the Apple Macintosh, and a self-proclaimed radical atheist.

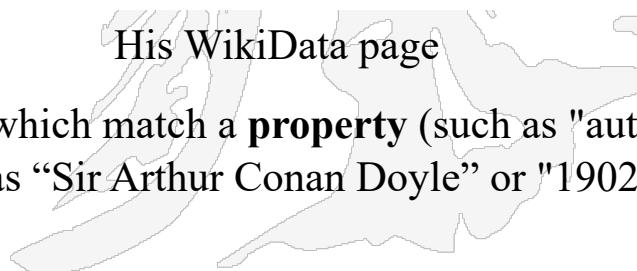
Contents [hide]

- 1 Early life
  - 1.1 Education
- 2 Career
  - 2.1 Writing
    - 2.1.1 *Doctor Who*
    - 2.1.2 *The Hitchhiker's Guide to the Galaxy*
    - 2.1.3 *Dirk Gently* series
  - 2.2 Music
    - 2.2.1 Pink Floyd
  - 2.3 Computer games and projects
- 3 Personal beliefs and activism



His Wikipedia page

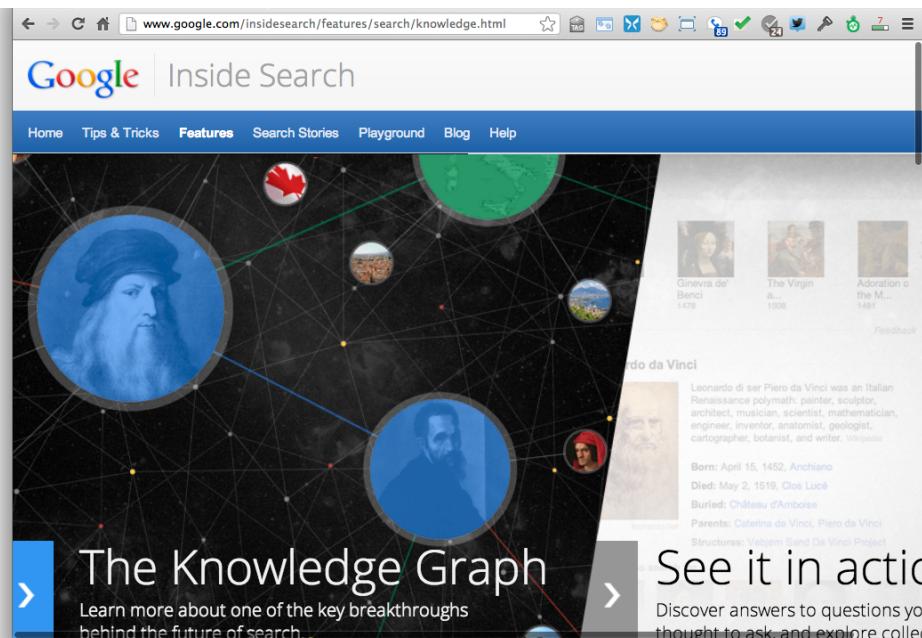
Statements in Wikidata consist of key-value pairs, which match a **property** (such as "author", or "publication date") with one or more **values** (such as "Sir Arthur Conan Doyle" or "1902").



# Google's Knowledge Graph



# Google Knowledge Graph



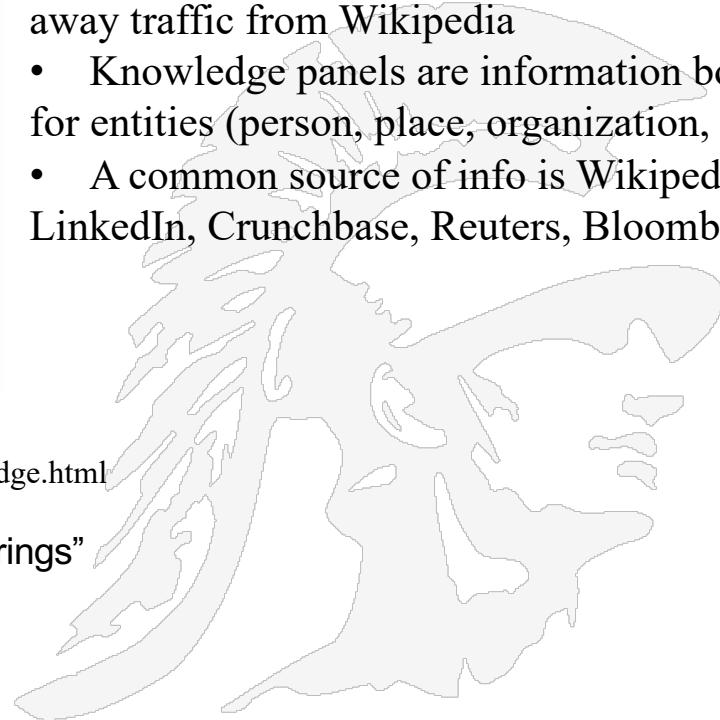
The screenshot shows the Google Inside Search interface for the Knowledge Graph. At the top, there's a navigation bar with links for Home, Tips & Tricks, Features, Search Stories, Playground, Blog, and Help. Below the navigation is a large, dark-themed Knowledge Graph visualization. It features several circular nodes, each containing a portrait of a historical figure: Leonardo da Vinci, another man with a beard, and a woman. These nodes are interconnected by a network of thin green lines, representing relationships. To the right of the graph, there are three smaller rectangular boxes showing snippets of information: "Ginevra de' Benci 1474", "The Virgin a... 1485", and "Adoration o... 1481". Below the graph, a prominent blue banner reads "The Knowledge Graph" and "Learn more about one of the key breakthroughs behind the future of search.". Another banner below it says "See it in action" and "Discover answers to questions you thought to ask, and explore colle...".

<https://www.google.com/intl/es419/insidesearch/features/search/knowledge.html>

Google's slogan for the knowledge graph: “things, not strings”

<https://www.youtube.com/watch?v=mmQl6VGvX-c>

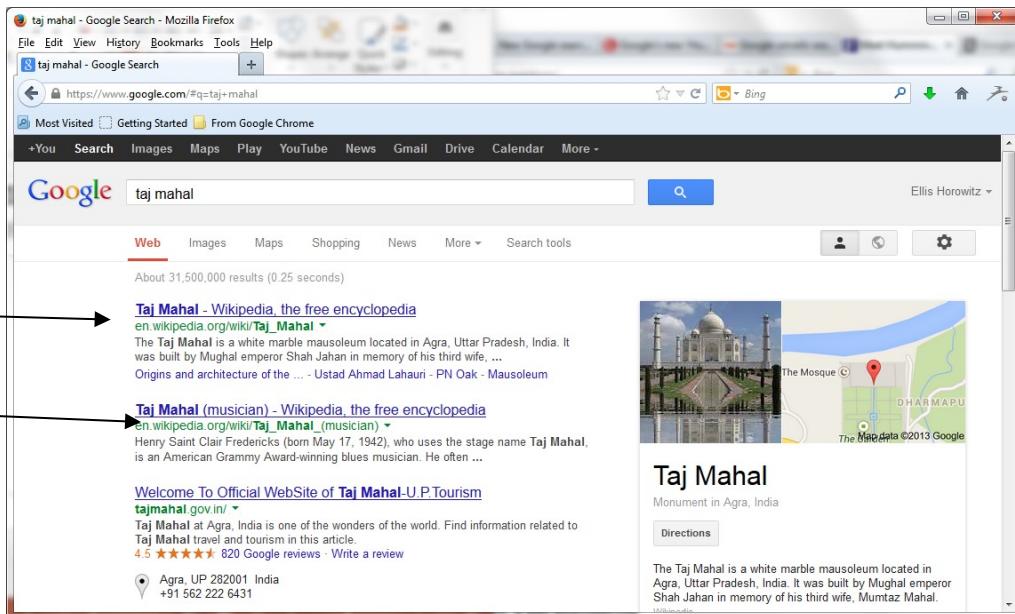
- Introduced in 2012
- Powered in part by Freebase
- knowledge graph was accused of taking away traffic from Wikipedia
- Knowledge panels are information boxes for entities (person, place, organization, event, etc)
- A common source of info is Wikipedia, LinkedIn, Crunchbase, Reuters, Bloomberg



# Knowledge Graph Enhances Google Search in 3 main ways (1):

mausoleum

musician



The screenshot shows a Google search results page for the query "taj mahal". The results are filtered by "Web". The first result is a link to the Wikipedia page for the Taj Mahal mausoleum, with a snippet describing it as a white marble mausoleum built by Mughal emperor Shah Jahan. The second result is a link to the Wikipedia page for the musician Taj Mahal, with a snippet describing him as an American Grammy Award-winning blues musician. To the right of the search results, there is a thumbnail image of the Taj Mahal and a map showing its location in Agra, India.

**1. To improve the variety of search results,**  
Google uses the knowledge graph to locate  
alternate interpretations of query terms,

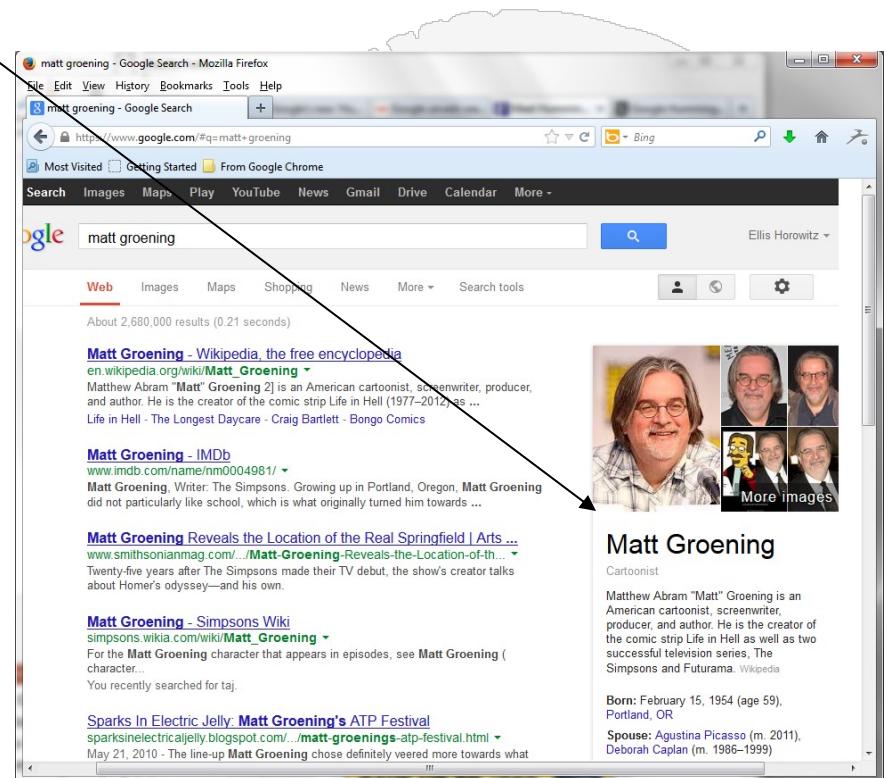
**Here it offers two of them with the same  
name e.g.  
"taj mahal" - the mausoleum or musician**



## Knowledge Graph Enhances Google Search in 3 main ways (2):

- 2. To provide deeper and broader results, typically in an info box**  
 e.g. person entities include relations such as age, birthplace, marital status, children, education, etc.,  
 here is a sample result for Matt Groening

- creator of The Simpsons
- **Go Deeper**
  - his photo
  - when he was born
  - his spouse
  - his parents
  - why he is famous
- **Go Broader**
  - other people related to Groening



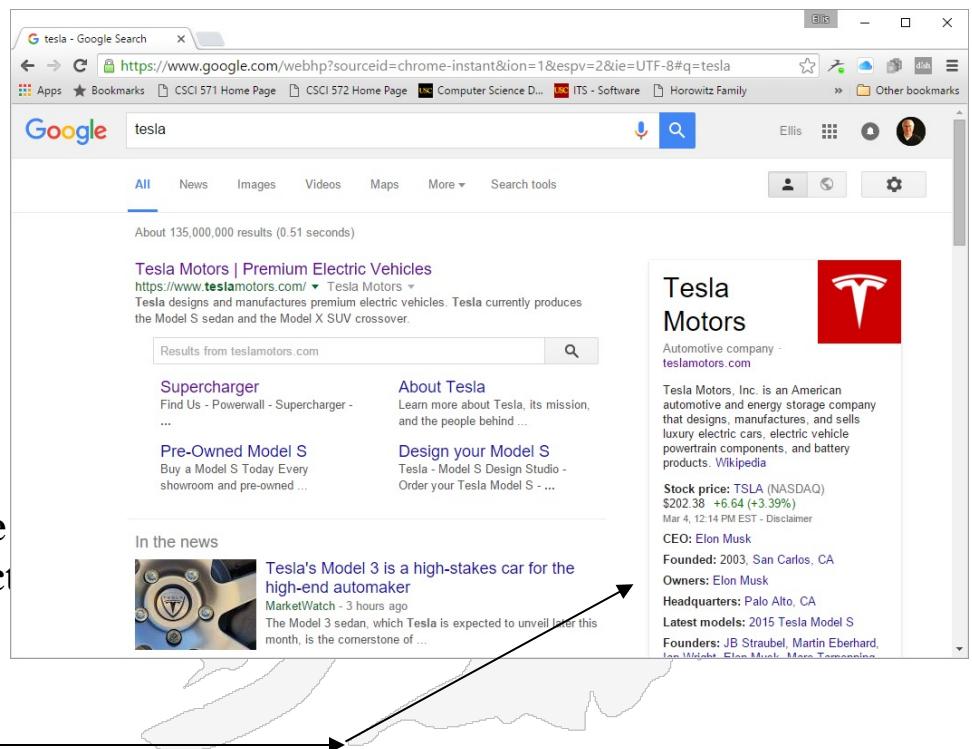
# Knowledge Graph Enhances Google Search in 3 main ways: (3)

### 3. To provide the best summary

the knowledge graph exploits the relationships among the entities  
 e.g. the query “Tesla”

The knowledge graph allows Google to summarize relevant content around that topic, including key facts you’re likely to need for that particular thing. E.g.

**Tesla Motors, Inc. is an American automotive and energy storage company that designs, manufactures, and sells luxury electric cars, electric vehicle powertrain components, and battery products**

The screenshot shows a Google search results page for the query "tesla". The top result is a Knowledge Graph summary card for Tesla Motors. The card includes the company's logo, a brief description ("Tesla Motors | Premium Electric Vehicles"), and links to "Supercharger", "About Tesla", "Pre-Owned Model S", and "Design your Model S". Below the card, there is a news snippet about Tesla's Model 3. The right side of the screen displays detailed information about Tesla, including its stock price (\$202.38), CEO (Elon Musk), and latest models (2015 Tesla Model S).

# Google Knowledge Graph and Wikipedia

- For most search results first sentences come from Wikipedia

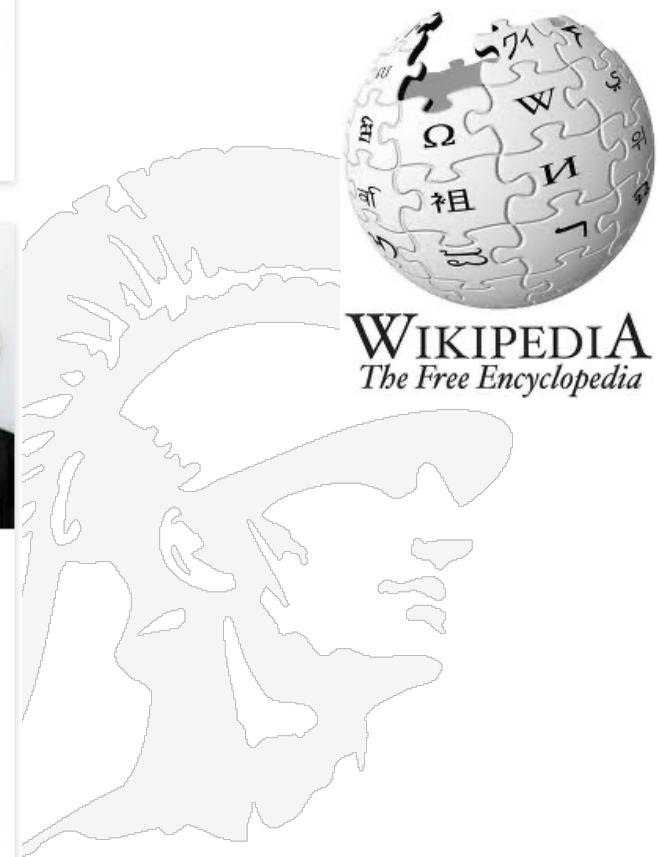


A small screenshot of a Facebook profile. It shows a photo of a man with short grey hair smiling, his name "Dieter Fensel", his location "Innsbruck", and a green "Friends" button.



**Dieter Fensel**

Dieter Fensel is a researcher in the field of formal languages and the semantic web. He is University Professor at the University of Innsbruck, where he directs the Semantic Technologies Institute ...  
[Wikipedia](#)



**WIKIPEDIA**  
*The Free Encyclopedia*



**People also search for**

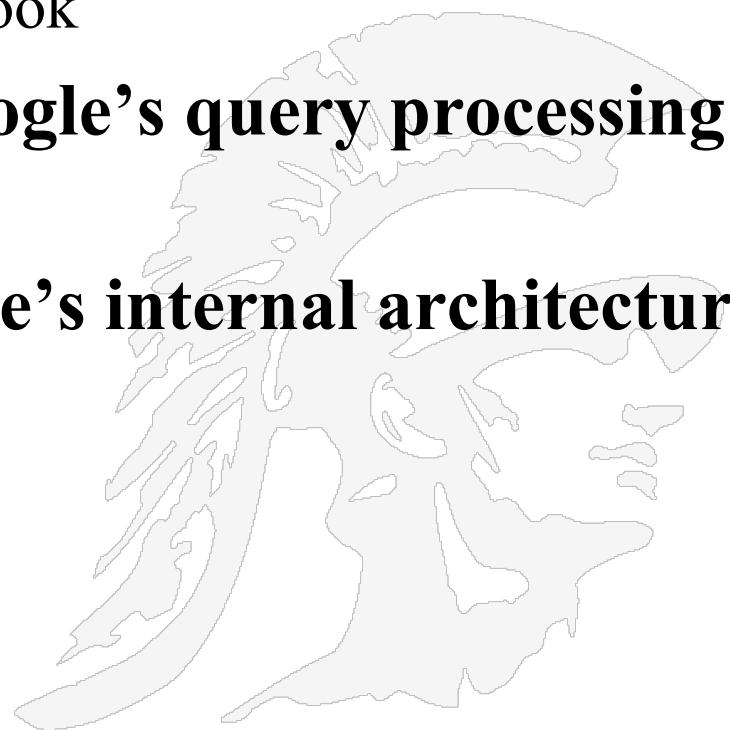
 Rudi Studer	 Frank van Harmelen	 James Hendler	 Ian Horrocks	 Deborah McGuinness
----------------------------------------------------------------------------------------------------	--------------------------------------------------------------------------------------------------------------	---------------------------------------------------------------------------------------------------------	---------------------------------------------------------------------------------------------------------	----------------------------------------------------------------------------------------------------------------

# Query Processing



# 3 Parts to Today's Lecture

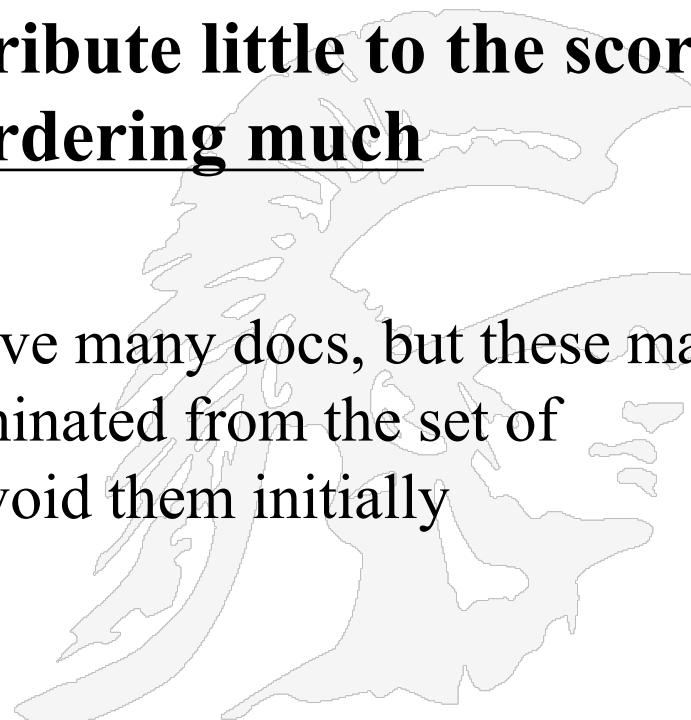
- 1. Restructuring the inverted index to speed up processing**
  - See Chapter 7 of our textbook
- 2. Reverse engineering Google's query processing algorithm**
- 3. A close up look at Google's internal architecture**



# Speeding Up Indexed Retrieval

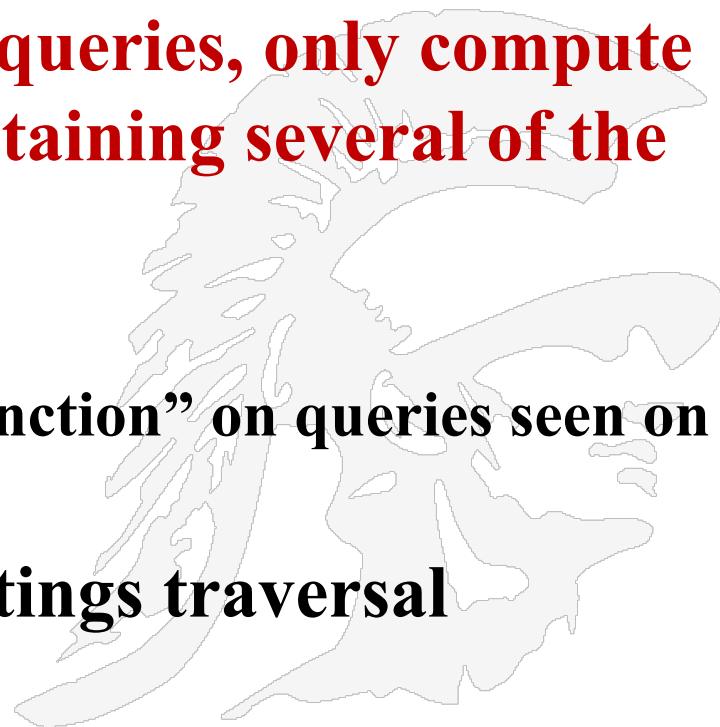
- **User has a task and formulates it as a query**
- **The search engine's task is to**
  1. **Minimally return documents that contain the query terms**
    - Use inverted index and cosine similarity to identify matching documents
    - Rank the results to try and identify the K top scoring documents and return those
  2. **Determine what the user is actually trying to accomplish, even though the query may be (at best) vaguely stated**
    - Use knowledge graph, user location, profile, etc to create the most likely responses
- **The following slides contain heuristics that can be applied to speed up step 1 of the process**

- For a query such as *catcher in the rye*
- Only accumulate (cosine) scores for *catcher* and *rye*
- Intuition: *in* and *the* contribute little to the scores and so don't alter rank-ordering much
- Benefit:
  - Postings of low-idf terms have many docs, but these many docs will eventually get eliminated from the set of contenders, so it is best to avoid them initially



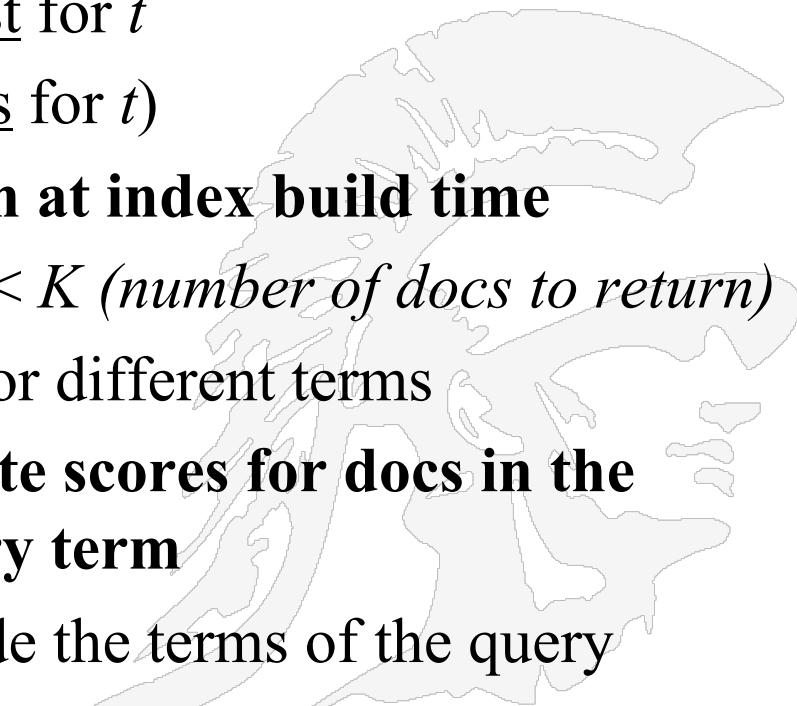
## Consider Only Docs Containing Several Query Terms

- In theory, any doc with at least one query term is a candidate for the output list
- However, for multi-term queries, only compute cosine scores for docs containing several of the query terms
  - Say, at least 3 out of 4
  - This imposes a “soft conjunction” on queries seen on web search engines
- Easy to implement in postings traversal



## Strategy 3: Introduce Champion Lists Heuristic

- Pre-compute for each dictionary term  $t$ , the  $r$  docs of highest weight (tf-idf) in  $t$ 's postings
  - Call this the champion list for  $t$
  - (aka fancy list or top docs for  $t$ )
- Note that  $r$  has to be chosen at index build time
  - Thus, it's possible that  $r < K$  (*number of docs to return*)
  - The value of  $r$  can vary for different terms
- At query time, only compute scores for docs in the champion list of some query term
  - champion lists that include the terms of the query
  - Pick the  $K$  top-scoring docs from among these

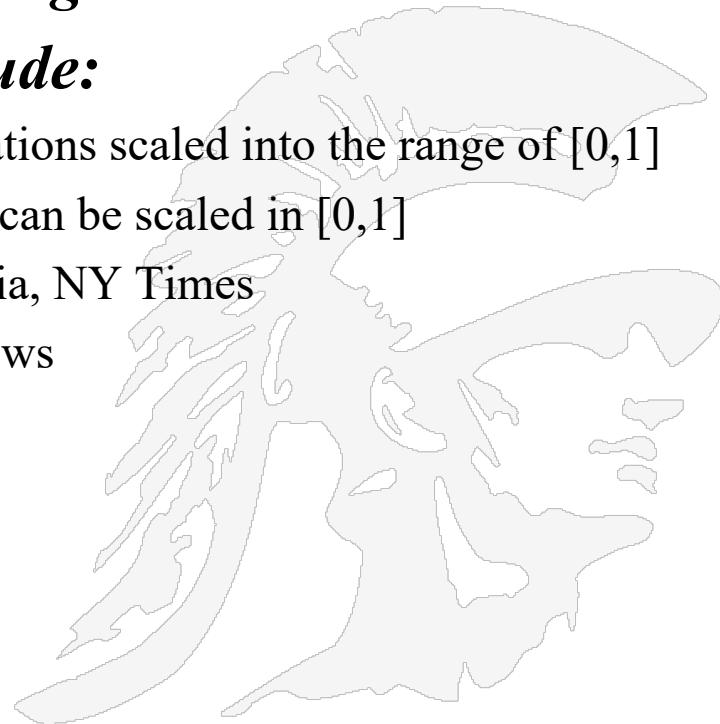


# Static Quality Scores Heuristic

- We want top-ranking documents to be both *relevant* and *authoritative*
- *Relevance* is being modeled by cosine scores
- *Authority* is typically a query-independent property of a document
- Examples of authority signals
  - Wikipedia result among other websites
  - Articles in curated newspapers
  - A paper/webpage with many citations, or equivalently
  - A web page with high PageRank

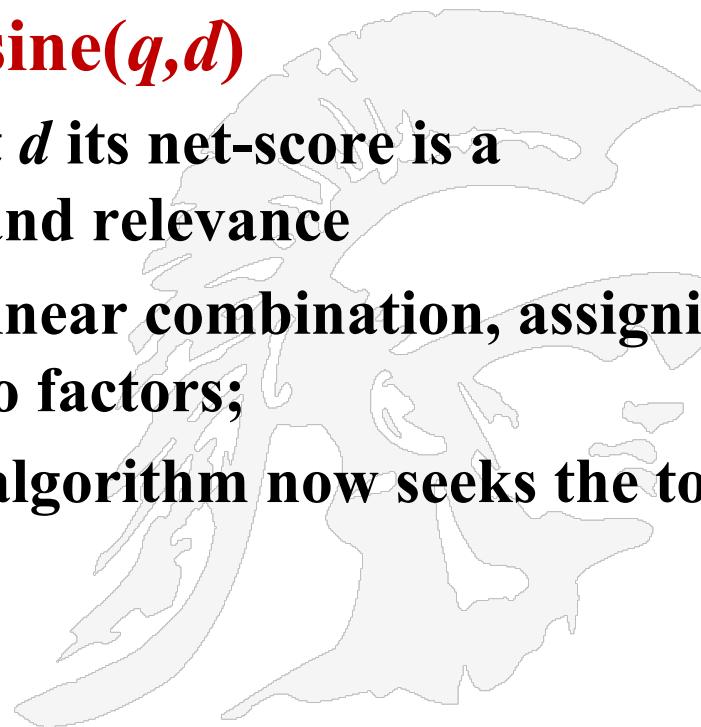
## Strategy 4: Introduce an Authority Measure

- Assign to each document  $d$  a query-independent quality score in  $[0,1]$
- Denote this by  $g(d)$ ,  $g$  stands for goodness
- *Authority measures might include:*
  - Documents with a high number of citations scaled into the range of  $[0,1]$
  - Documents with high PageRank, also can be scaled in  $[0,1]$
  - Heavily curated content, e.g. Wikipedia, NY Times
  - Documents with many favorable reviews



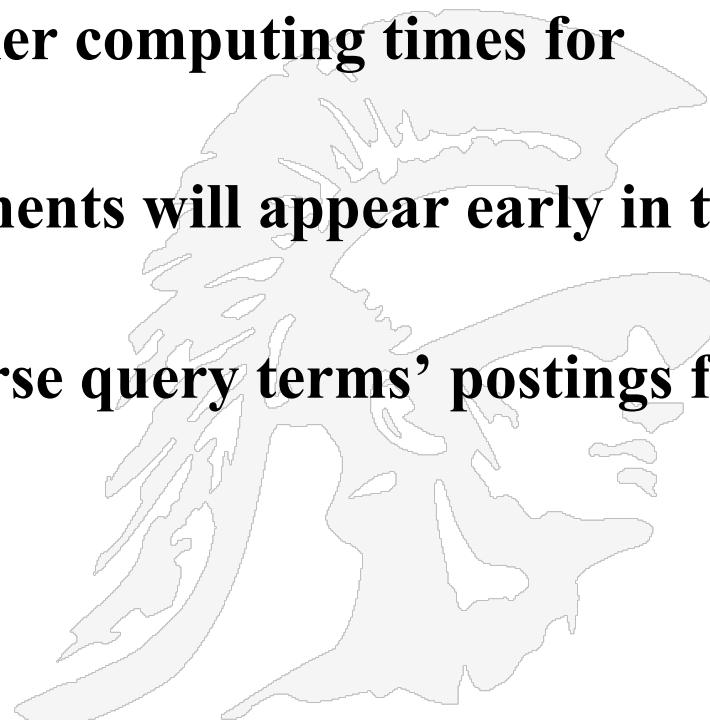
## Combine Relevance and Authority

- Consider a simple total score combining cosine relevance and authority
- $\text{net-score}(q,d) = g(d) + \cosine(q,d)$ 
  - For query  $q$  and document  $d$  its net-score is a combination of authority and relevance
  - We could use some other linear combination, assigning different weights to the two factors;
  - In processing a query the algorithm now seeks the top  $K$  docs by net-score



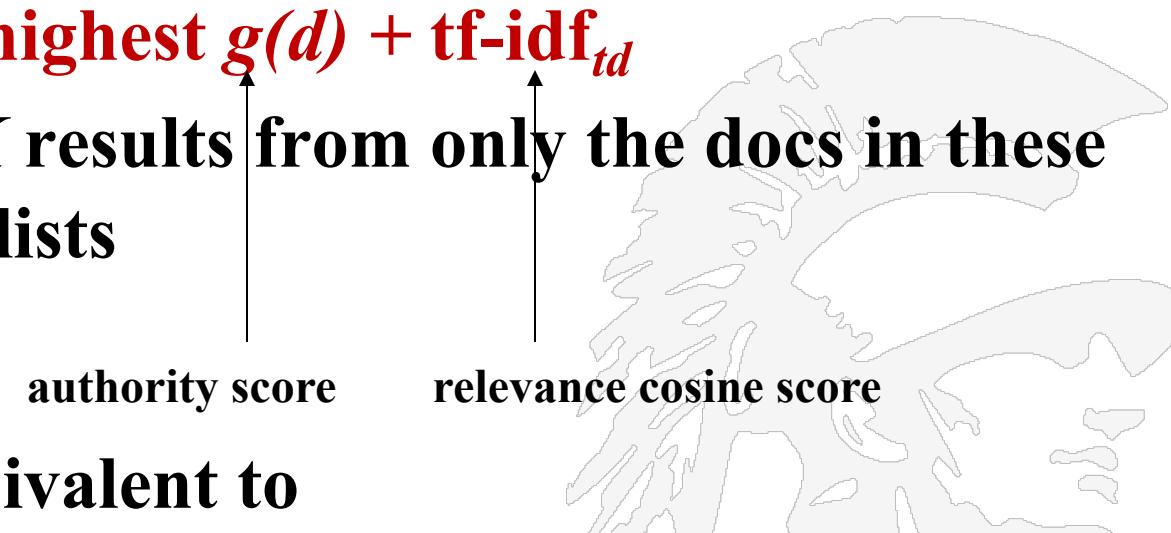
## Strategy 5: Reorganize the Inverted List

- So far we assumed that all documents were ordered by docID, even those on the champion lists
- Instead order all postings by  $g(d)$  the authority measure
- This does not change the earlier computing times for merging
- The most authoritative documents will appear early in the postings list
- Thus, can concurrently traverse query terms' postings for
  - Postings intersection
  - Cosine score computation



# Computing Net Score

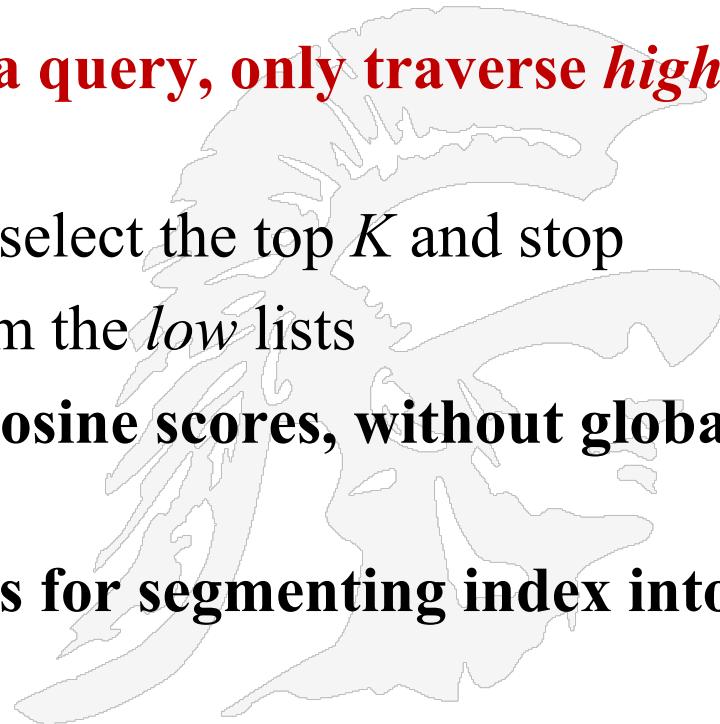
- Combine champion lists with  $g(d)$ -ordering
- Maintain for each term a champion list of the  $r$  docs with highest  $g(d) + \text{tf-idf}_{td}$
- Seek top- $K$  results from only the docs in these champion lists
- This is equivalent to



$$\text{net-score}(q, d) = g(d) + \frac{\vec{V}(q) \cdot \vec{V}(d)}{|\vec{V}(q)| |\vec{V}(d)|}.$$

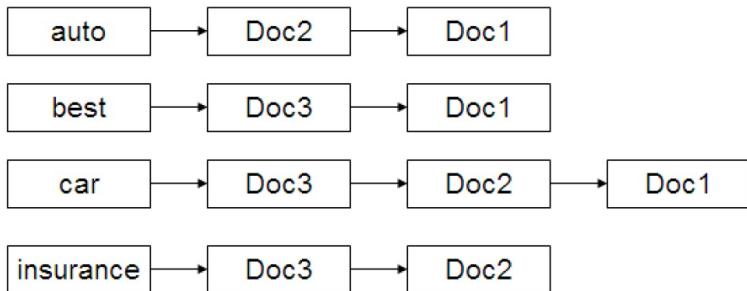
## Strategy 6: High and Low Lists Heuristic

- For each term, maintain two postings lists called *high* and *low*
  - Think of *high* as the champion list
- When traversing postings on a query, only traverse *high* lists first
  - If we get more than  $K$  docs, select the top  $K$  and stop
  - Else proceed to get docs from the *low* lists
- Can be used even for simple cosine scores, without global quality  $g(d)$
- This assumes we have a means for segmenting index into two tiers



	Doc1	Doc2	Doc3
car	27	4	24
auto	3	33	0
insurance	0	33	29
best	14	0	17

4 documents with term frequencies

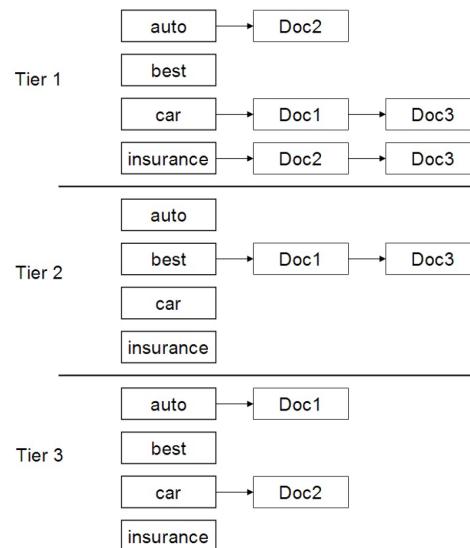


Static quality-ordered index;  
 Assume doc1, doc2, doc3 have  
 quality scores  $g(1)=0.25$ ,  
 $g(2)=0.5$ ,  $g(3)=1$

## Example of a Tiered Inverted Index; A Generalization of Champion Lists

term	$df_t$	$idf_t$
car	18,165	1.65
auto	6723	2.08
insurance	19,241	1.62
best	25,235	1.5

Inverse document frequencies



Tiered index, threshold of 20 for tier 1, 10 for tier 2  
 if tier 1 doesn't provide enough results, try tier 2, etc

# Recap: How to Compute Cosine Score

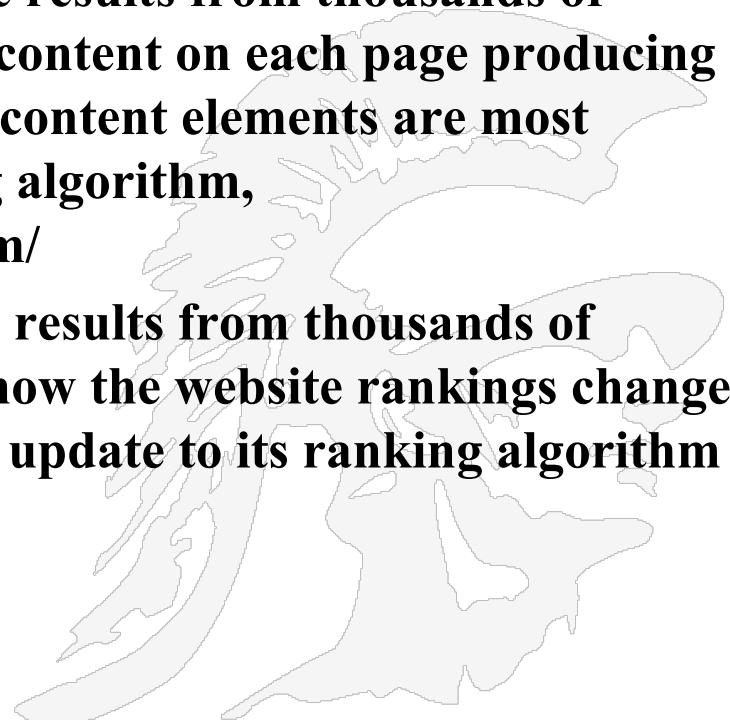
$$\cos(\vec{q}, \vec{d}) = \frac{\vec{q} \bullet \vec{d}}{\|\vec{q}\| \|\vec{d}\|} = \frac{\vec{q}}{\|\vec{q}\|} \bullet \frac{\vec{d}}{\|\vec{d}\|} = \frac{\sum_{i=1}^{|V|} q_i d_i}{\sqrt{\sum_{i=1}^{|V|} q_i^2} \sqrt{\sum_{i=1}^{|V|} d_i^2}}$$

$\cos(\vec{q}, \vec{d})$  is the cosine similarity of  $\vec{q}$  and  $\vec{d}$  ... or, equivalently, the cosine of the angle between  $\vec{q}$  and  $\vec{d}$ .

The algorithm for computing cosine scores can be found in Figure 6.14 of our textbook

## Part 2: Google's Query Processing Algorithm

- Now let's switch gears and look at the problem of reverse engineering Google's query processing (ranking) algorithm
- There are two main companies trying to do this:
  1. *Searchmetrics* which tracks the results from thousands of keywords while analyzing the content on each page producing a ranking that determines what content elements are most important in Google's ranking algorithm,  
<https://www.searchmetrics.com/>
  2. *Moz.com* which also tracks the results from thousands of keywords and then measures how the website rankings changed whenever Google performs an update to its ranking algorithm  
<https://moz.com/>



## Searchmetrics Tracks Ranking Factors Valued by Google's Algorithm

### Reverse engineering the Google ranking algorithm

Download at:

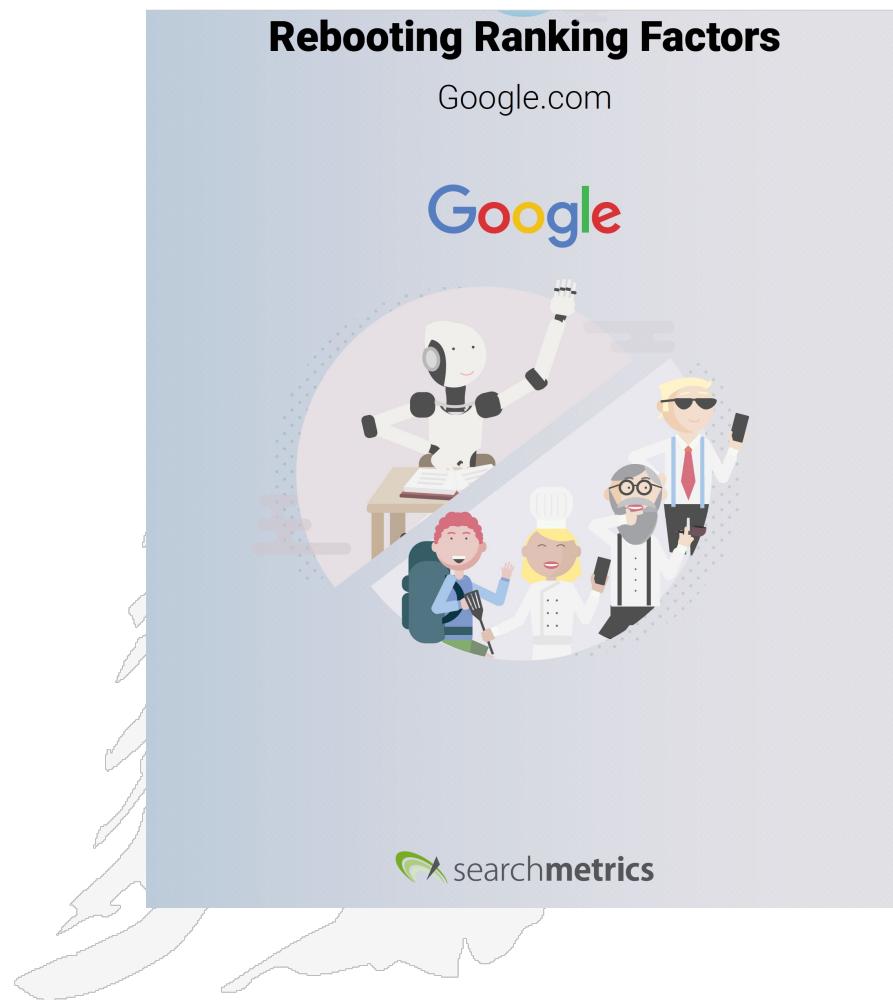
<http://csci572.com/papers/Searchmetrics.pdf>

### Why search is important commercially



### Rebooting Ranking Factors

Google.com



# General Ranking Factors

Here are the major categories and some factors

## Content

Overall relevance  
Word count



## User Signals

CTR  
Bounce rate



## Technical

HTTPS  
Presence of H1/H2



## User Experience

Font size  
number of internal/external links



## Social Signals

Pinterest  
Facebook/Tweet



## Backlinks

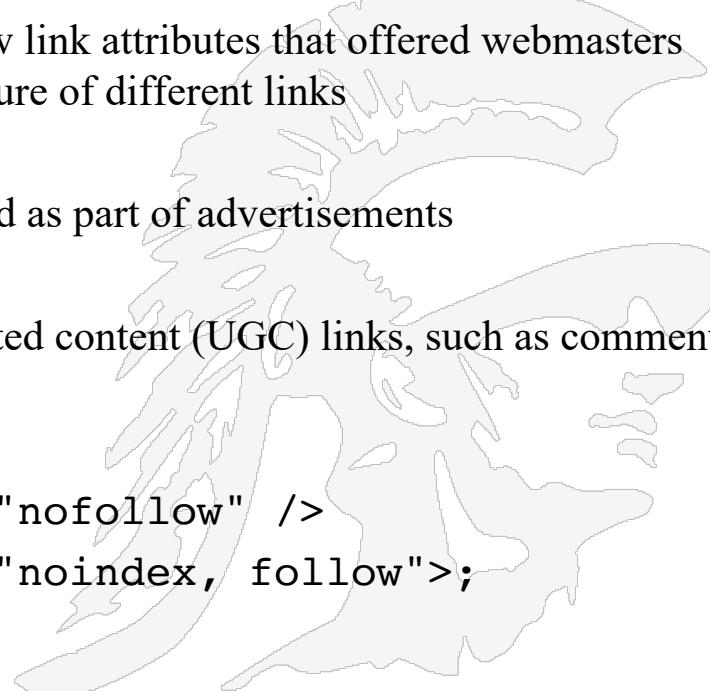
number of No-follow

Copyright Ell



# Note: What Are Nofollow Links

- The nofollow tag tells search engines to ignore that link
- Google introduced the rel="nofollow" option in 2005 for bloggers that were struggling with people using comment spam to try and build links in the hope of ranking for specific keywords
- Examples
  - `<a href="https://example.com" rel="nofollow">don't follow this link</a>`
- In September 2019, Google announced two new link attributes that offered webmasters additional ways to help Google identify the nature of different links
  - `rel = "sponsored"`
  - identify links on your site that were created as part of advertisements
  - `rel = "ugc"`
  - Google recommends marking user-generated content (UGC) links, such as comments and forum posts, as UGC
- Other relevant attributes
  - `<meta name="robots" content="nofollow" />`
  - `<meta name="robots" content="noindex, follow">;`



The analysis shows that the content relevance, decreases as the position in the search results drops.

The highest content relevance scores were found among the results for positions 3 to 6.

Thereafter, the landing pages on subsequent positions show lower relevance scores

# Content Factors

Overall Content Relevance  
- disregarding the search term itself -



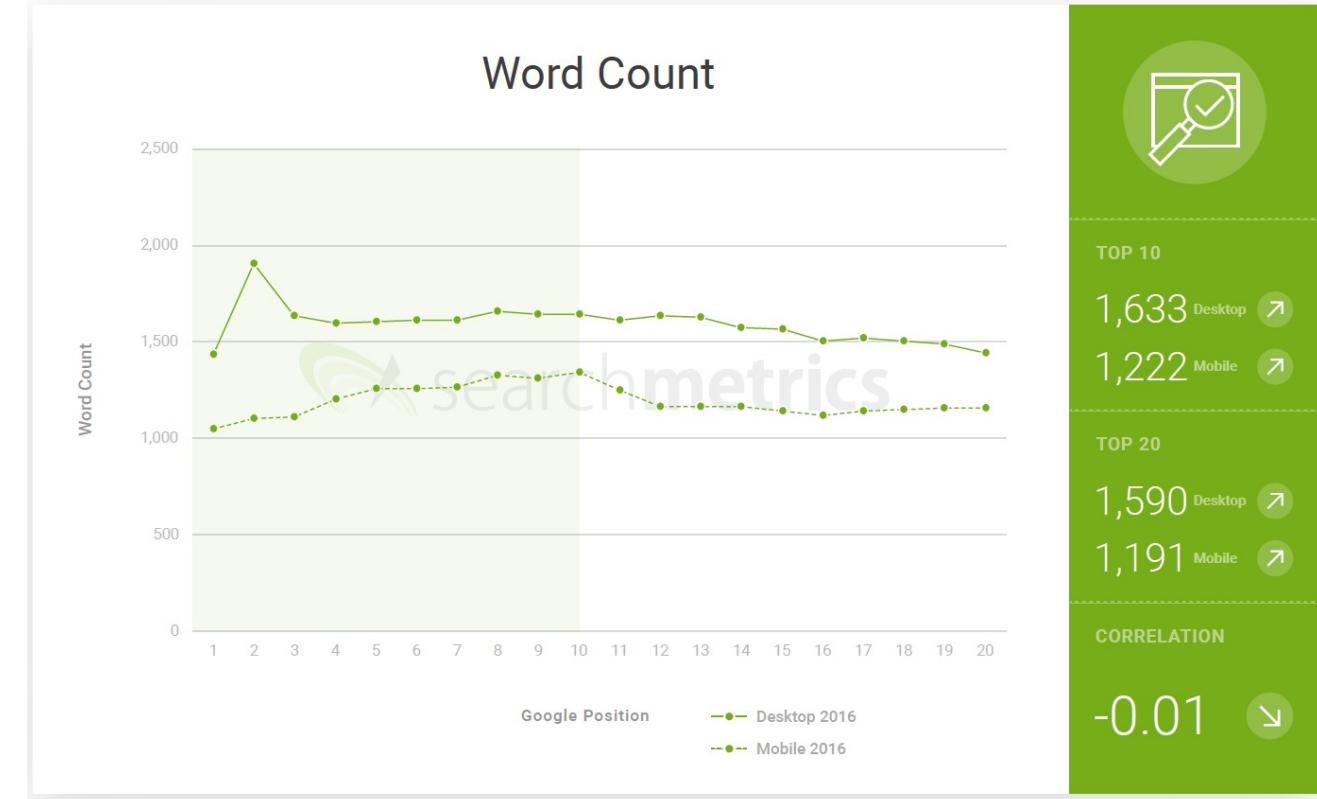
CORRELATION

0.04



# Word Count

- The word count of a landing page ranked among the top positions
- Pages rank well under the condition that the content is not simply long, but also relevant,

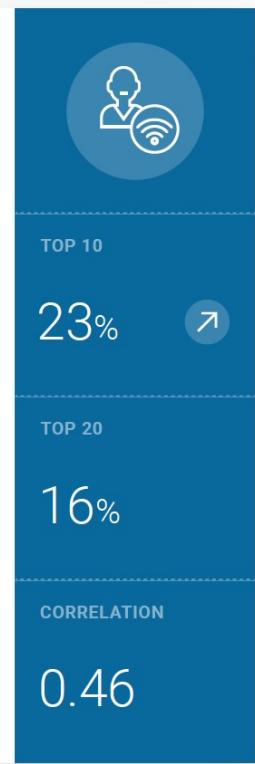
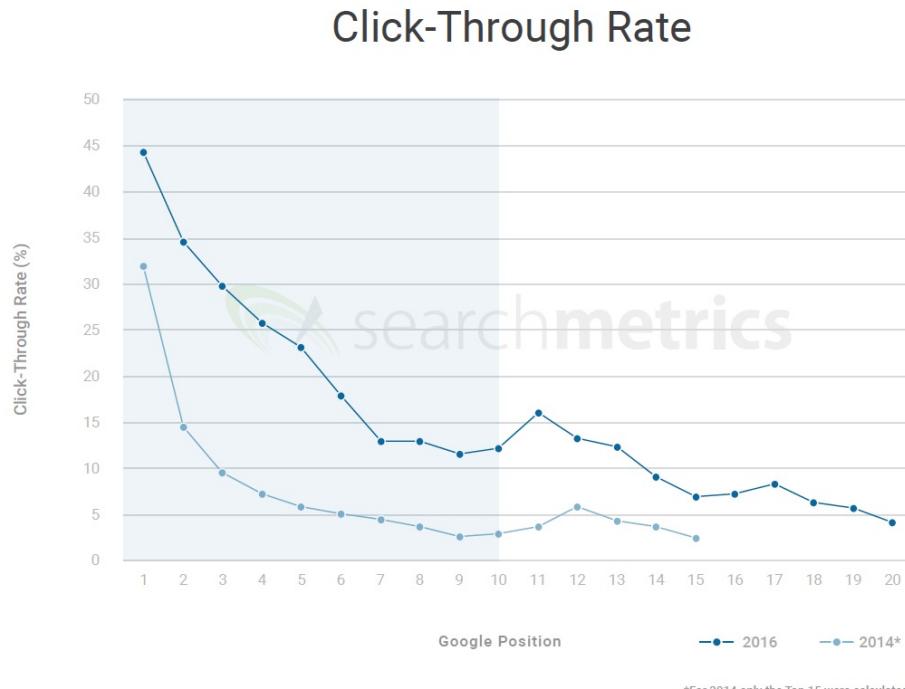


The Click-Through Rate measures the average percentage of users who click on the result at each position on the SERP\*.

Keywords in position 1 have an average CTR of 44%, the rate dropping to 30% for position 3.

The click rate for landing pages at the top of the second results page is higher than for results at the bottom of page 1.

# Click-Through Rate

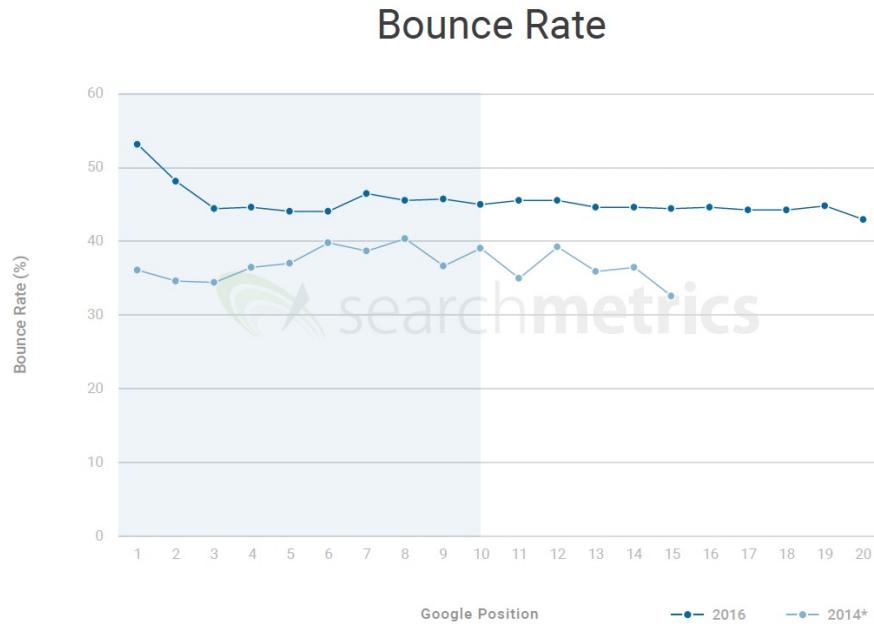


\*SERP: Search Engine Results Page

# Bounce Rate

The Bounce Rate measures the percentage of users who only click on the URL from Google's search results, without visiting any other URLs at that domain, and then return back to the SERP\*.

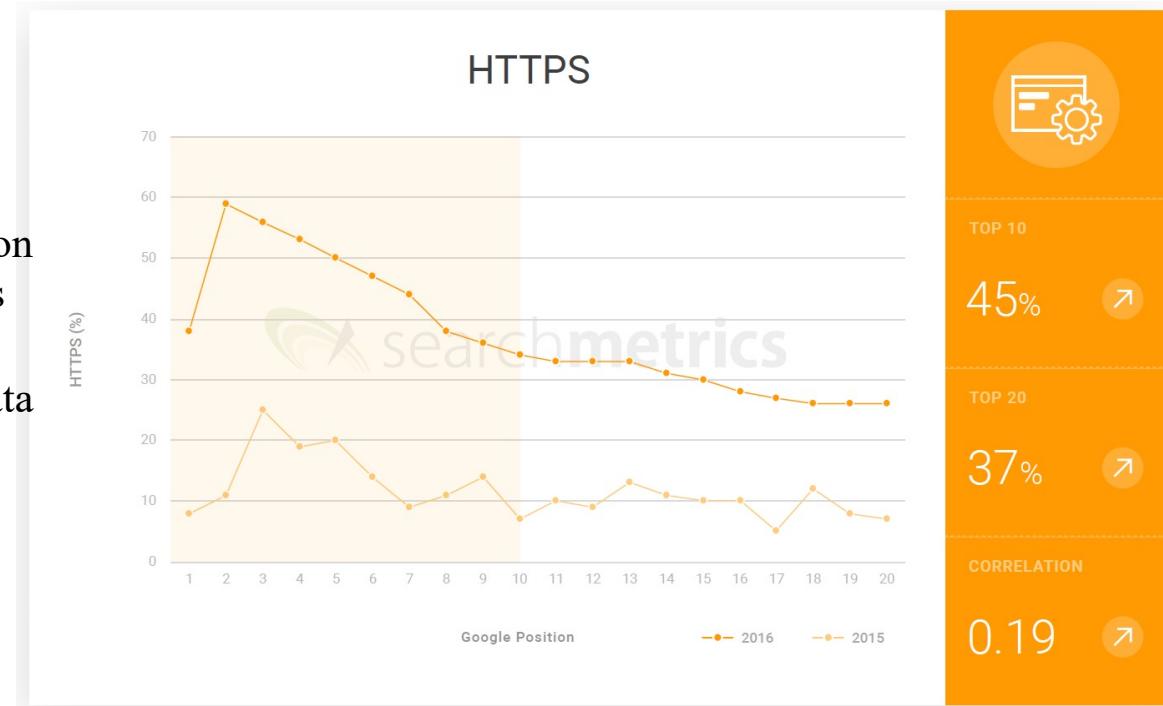
These are single-page sessions where the user leaves the site without interacting with the page.



\*SERP: Search Engine Result Page

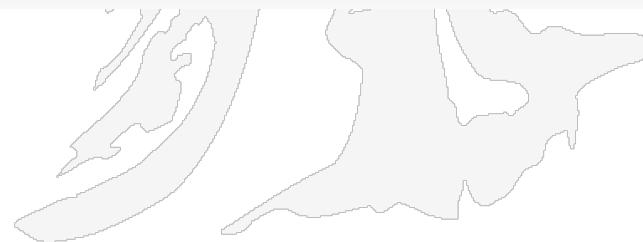
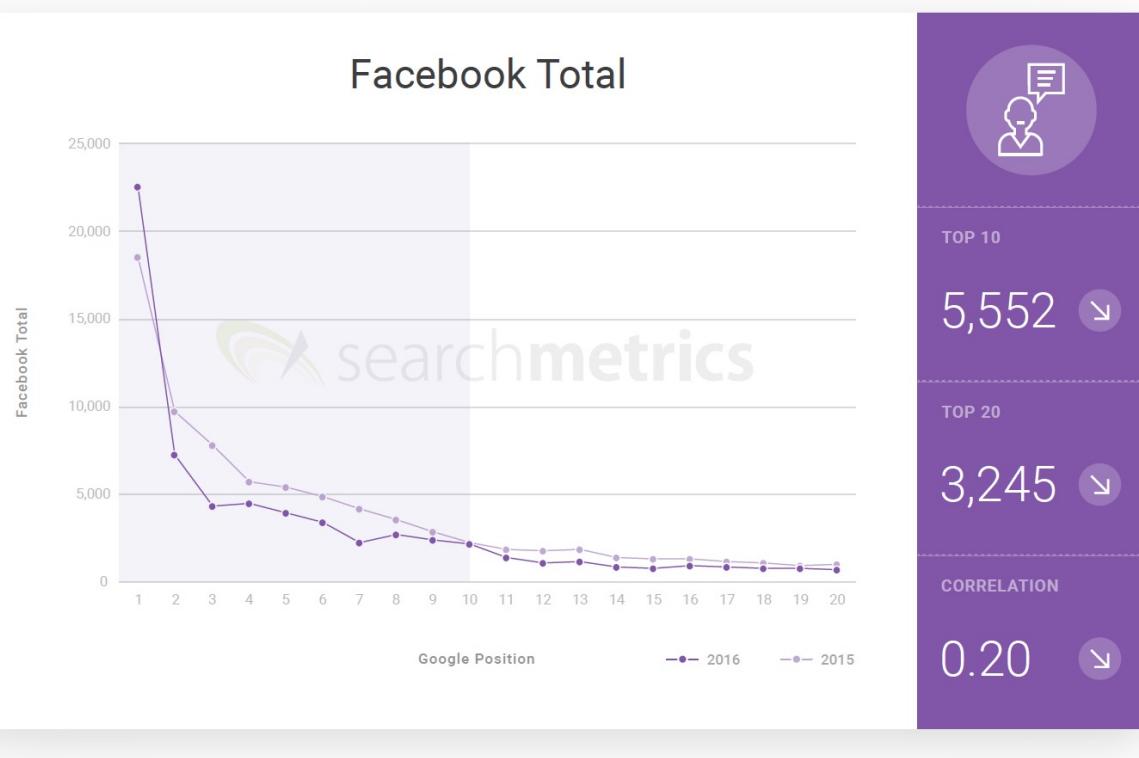
- Page encryption using HTTPS is rising.
- In 2017 only 12% of pages relied on data transfer via HTTP. Today, this has more than tripled, with over a third of websites encrypting the data traffic on their pages
- In 2017 Google announced that pages that have not switched to HTTPS would be marked as “unsafe” in its Chrome browser.

# Technical Factors



# Social Signals

- The correlation between social signals and ranking position is extremely high
- Facebook remains the social network with by far the highest level of user interactions.
- Facebook, compared with the other social networks, shows relatively high signals across the first search results page



All top 100 websites have a mobile-friendly version; they use either a mobile sub-domain or responsive design;

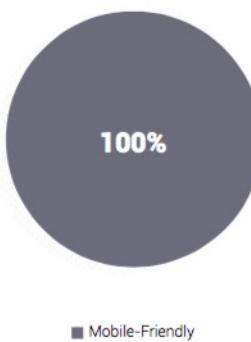
Separate mobile websites are diminishing in popularity, but e.g. try Sephora.com on desktop and mobile. Last time I looked they were still using m.sephora.com

Over a fifth of websites outside of the top 100 offer no mobile-friendly solution

#### Mobile-friendly websites

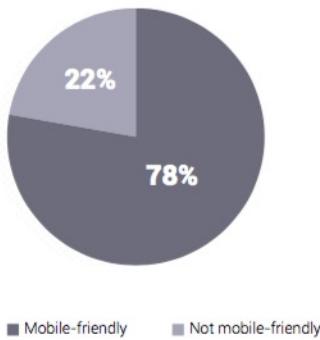
The following graphics show the frequency of websites with mobile-friendly solutions amongst the top 100 domains by SEO visibility.

Top 100 Domains Google US



That's right. All 100 of the top 100 have a mobile-friendly solution. These include the use of a mobile sub-domain, dynamic serving, responsive design and/or mobile apps.

Mobile-friendliness: Sample of smaller domains



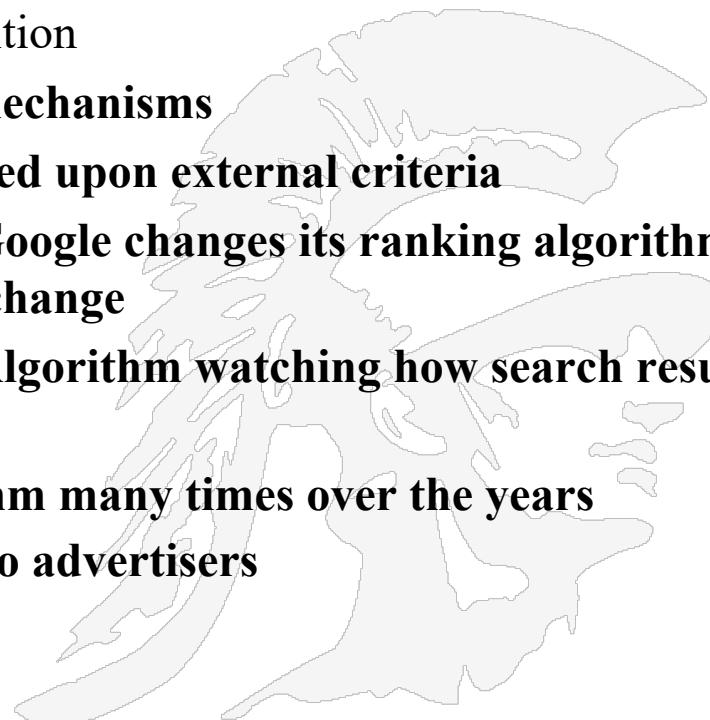
Over a fifth of websites outside the top 100, based on a sample of smaller domains, offer no mobile-friendly solution to smartphone users. The upcoming shift to a mobile-first index will have a negative impact on such websites, should they fail to react and implement mobile-friendly solutions.

## Correlation is Not Necessarily Causation

- See the full Searchmetrics report at
  - <http://csci572.com/papers/Searchmetrics.pdf>
- **WARNING: SEO studies always make it clear that their findings may not actually define the way the Google search result algorithm actually works**
  - Correlations are not synonymous with causal relationships
- **There are many examples of illusory correlations which are referred to as “logical fallacy”; here are two examples**
  1. the co-appearance of phenomena like a higher number of observed storks and the higher birthrate in certain areas,
  2. the relationship between sales of ice-cream and increased incidence of sunburn in the summer.
- **These examples show a (illusory) correlation, not a causal relationship.**
- **For many more funny ones go to: <http://tylervigen.com/spurious-correlations>**

# Another Way to Reverse Engineer Google's Query Processing Algorithm

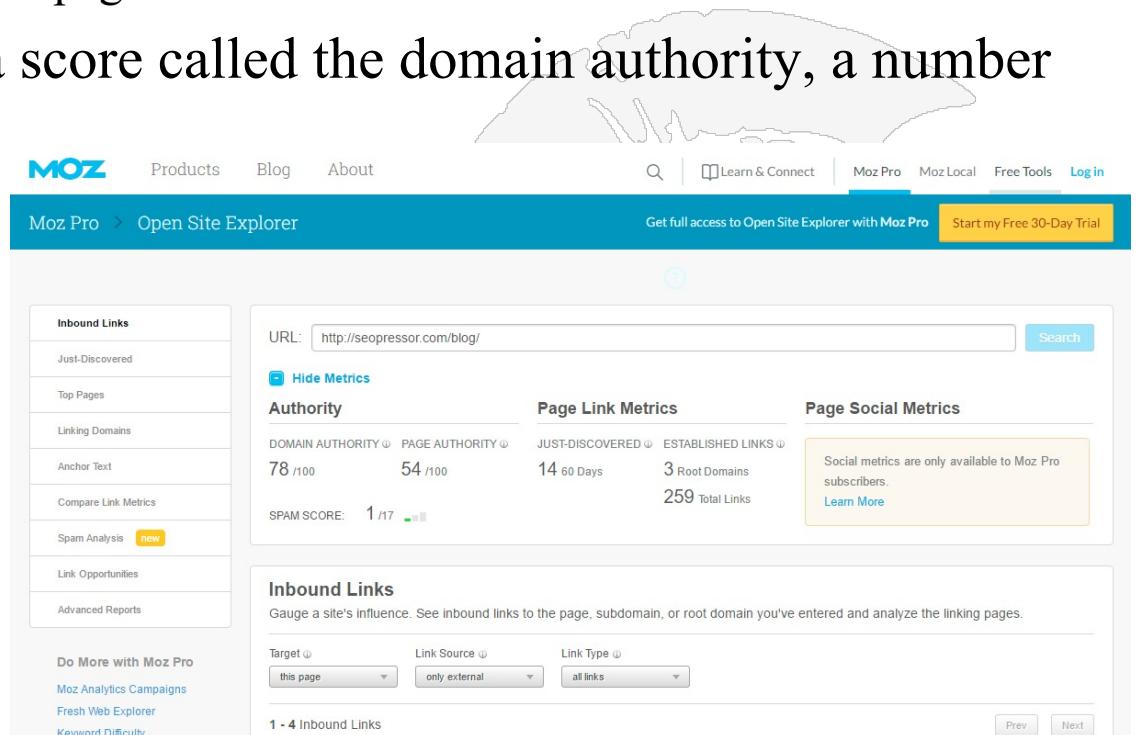
- Moz.com is an SEO company that sells its software as a service (SaaS); capabilities offered include:
  - Keyword research
  - Improving your ranking
  - Comparing your site with the competition
- As part of its service MOZ offers two mechanisms
  - *MozRank* scores your web page based upon external criteria
  - *MozCast* keeps track of whenever Google changes its ranking algorithm, see <https://moz.com/google-algorithm-change>
- Moz.com Monitors Google's Ranking Algorithm watching how search results are affected
- Google has changed its ranking algorithm many times over the years
- This algorithm is especially important to advertisers



- MozRank is a logarithmically scaled 10-point measurement of website linking authority or popularity of a given web page (<https://moz.com/help/link-explorer>)
  - It could be viewed as analogous to PageRank
  - See 4 minute video on the page
- MozRank is based on a score called the domain authority, a number between 1 and 100

Criteria include:

- Number of links to your site
- Quality of sites you link to
- Number of trusted sites linked to
- Quality of your content
- Social signals referencing your site



The screenshot shows the Moz Open Site Explorer interface. At the top, there's a navigation bar with 'MOZ' logo, 'Products', 'Blog', 'About', a search bar, and links for 'Learn & Connect', 'Moz Pro', 'Moz Local', 'Free Tools', and 'Log in'. A banner at the top right encourages upgrading to 'Moz Pro' with a 'Start my Free 30-Day Trial' button.

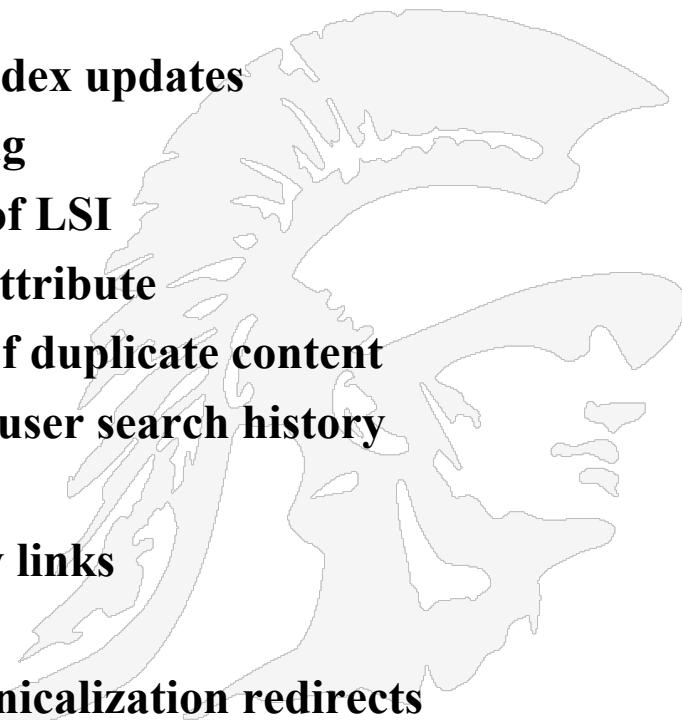
The main content area has a sidebar titled 'Inbound Links' containing links like 'Just-Discovered', 'Top Pages', 'Linking Domains', 'Anchor Text', 'Compare Link Metrics', 'Spam Analysis', 'Link Opportunities', and 'Advanced Reports'. The main panel displays domain authority metrics for the URL <http://seopressor.com/blog/>. The metrics shown are:

Authority	Page Link Metrics	Page Social Metrics
DOMAIN AUTHORITY 78 /100	JUST-DISCOVERED 14 /60 Days	ESTABLISHED LINKS 3 Root Domains
PAGE AUTHORITY 54 /100	ESTABLISHED LINKS 259 Total Links	Social metrics are only available to Moz Pro subscribers. <a href="#">Learn More</a>
SPAM SCORE: 1 /17		

Below this, there's a section titled 'Inbound Links' with a sub-section 'Target' set to 'this page', 'Link Source' set to 'only external', and 'Link Type' set to 'all links'. It shows '1 - 4 Inbound Links'.

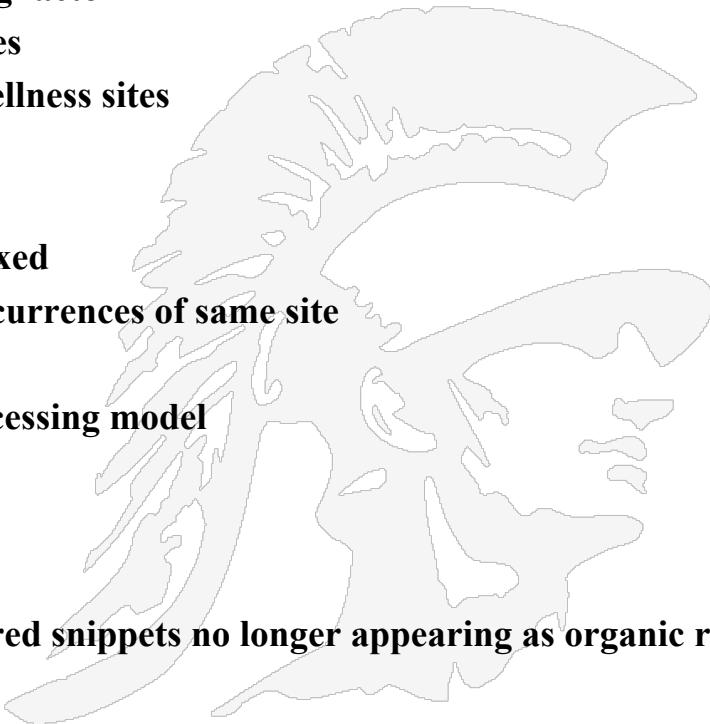
## Confirmed Google Updates to Their Ranking Algorithm – Early Days

- <https://moz.com/google-algorithm-change>
- Select any date to view the changes
  
- 2003 Feb, Boston, index refresh
- 2003, July, Fritz, switch to incremental index updates
- 2003, Nov, Florida, block keyword stuffing
- 2004, Feb, Brandy, index expansion, use of LSI
- 2005, Jan, Nofollow, introduce nofollow attribute
- 2005, May, Bourbon, improved treating of duplicate content
- 2005, June, Personalized search, keeping user search history
- 2005, June, XML sitemaps
- 2005, Oct, Jagger, eliminating low quality links
- 2005, Oct, Google local maps,
- 2005, Dec BigDaddy, handling URL canonicalization redirects



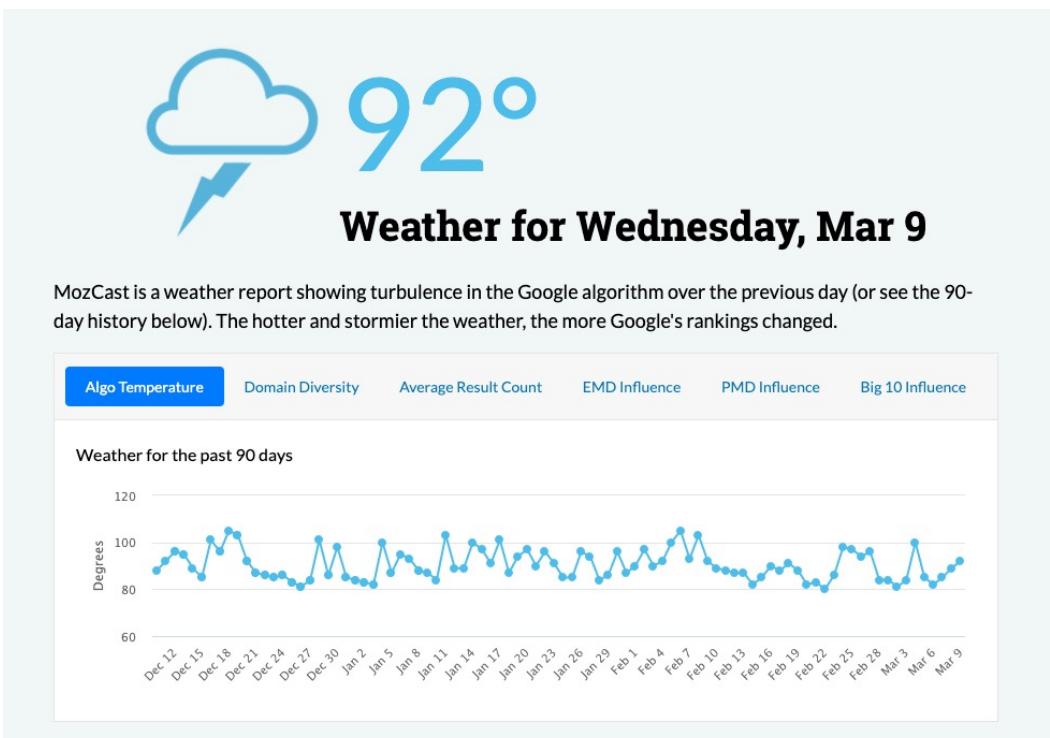
## Confirmed Google Updates to their Ranking Algorithm - Recent

- 2018, Mar, Mobile-First Index Roll-out
- 2018, Apr, Unnamed Core Update, no information given
- 2018, May, Snippet Length drop, rolled back snippet length to 150-160 characters
- 2018, Jun, Video Carousels, new display format
- 2018, Jul, Mobile speed update, page speed is a ranking factor
- 2018, Jul, chrome security warning on non-HTTPS sites
- 2018, Aug, Medic Core update, affecting health and wellness sites
  
- 2019, Mar, core update
- 2019, Apr, May, Jun, Deindexing Bug identified and fixed
- 2019, Jun, Site Diversity Update, reducing multiple occurrences of same site
- 2019, Sep, core Update, no information provided
- 2019, Oct, BERT update, BERT natural language processing model
- 2019, Dec, International rollout of BERT
  
- 2020, Jan, Core update
- 2020, Jan Featured Snippet De-duping, URLs in featured snippets no longer appearing as organic results
- 2020, May Core update
- 2020, Aug, Google Glitch, broken and then quickly fixed
- 2020, Sep, Oct, bug fixes



- Every 24 hours, Moz tracks a hand-picked set of 1,000 keywords and grab the top 10 Google organic results.
  - Keywords were deliberately chosen to avoid obvious local intent, are distributed evenly across 5 "bins" by query volume, and are tracked at roughly the same time every day from the same location
- Each day, they take the current top 10 and compare it to the previous day's top 10 (for any given keyword) and calculate a rate of change or "delta".
- This is done across all 1,000 keywords; they express the result as a “temperature in Farenheit; an average day is about 70° F.
  - <https://moz.com/mozcast>

# Moz.com Tracks Google Algorithm Updates



# The Google Architecture

See Google's Website  
on how search works at  
<http://www.google.com/insidesearch/howsearchworks/thestory/>



Much of these notes are based upon Keith Erikson's CSE497 and C. Lee Giles from Penn State IST 441 and Jeff Dean's Slides on Google

**2001**, adds “did you mean”

**2002**, handles synonyms

**2004**, added news & stock quotes

**2005**, added Autocomplete

**2006**, added video, weather, flights

**2007**, added movie times & patents

**2008**, Google search mobile app

**2009**, voice search

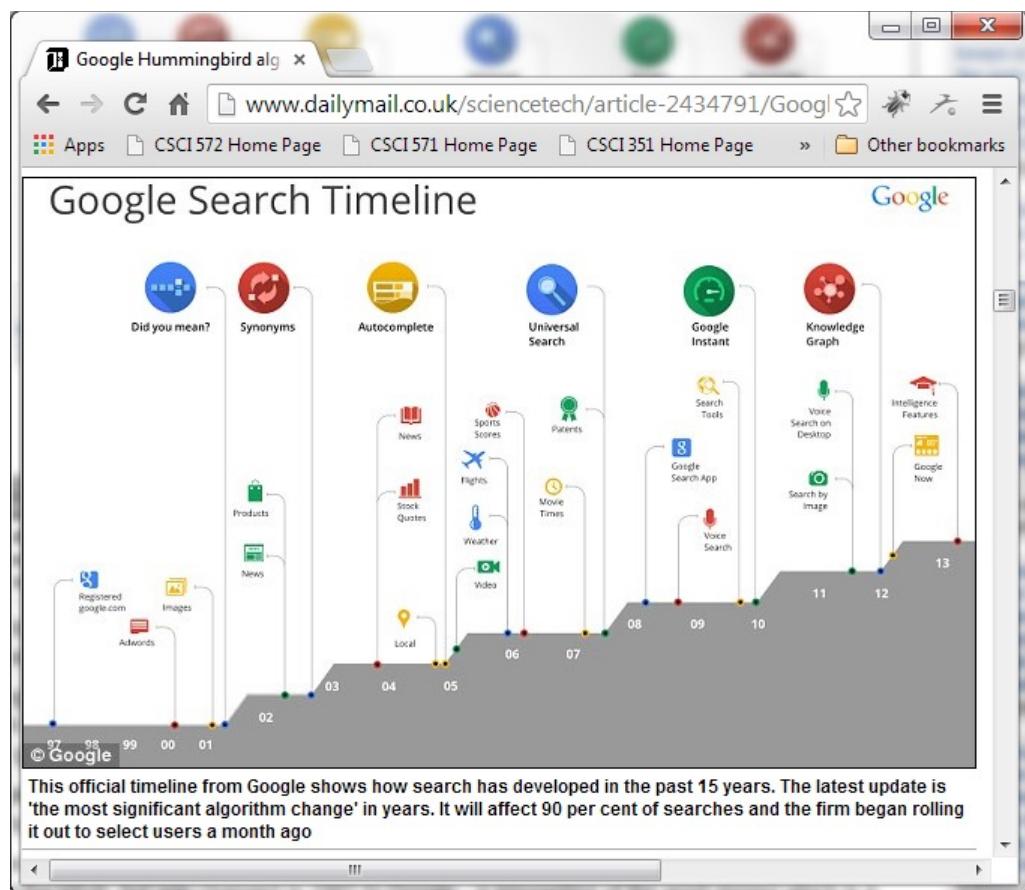
**2010**, Google Instant

**2011**, added image search

**2012**, added knowledge graph

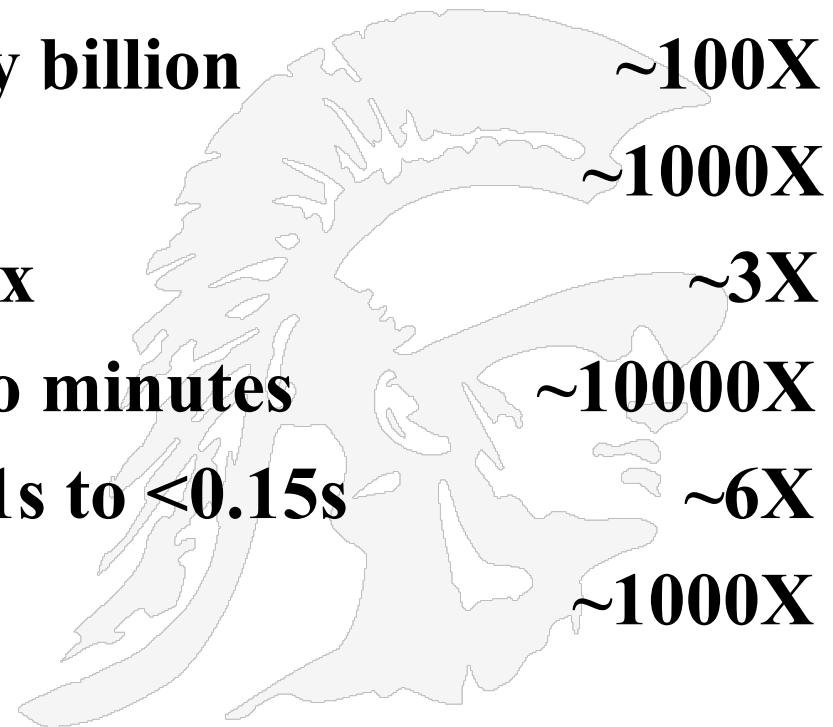
**2013**, use of carousels for display

## How Google Search Has Changed Over the Years

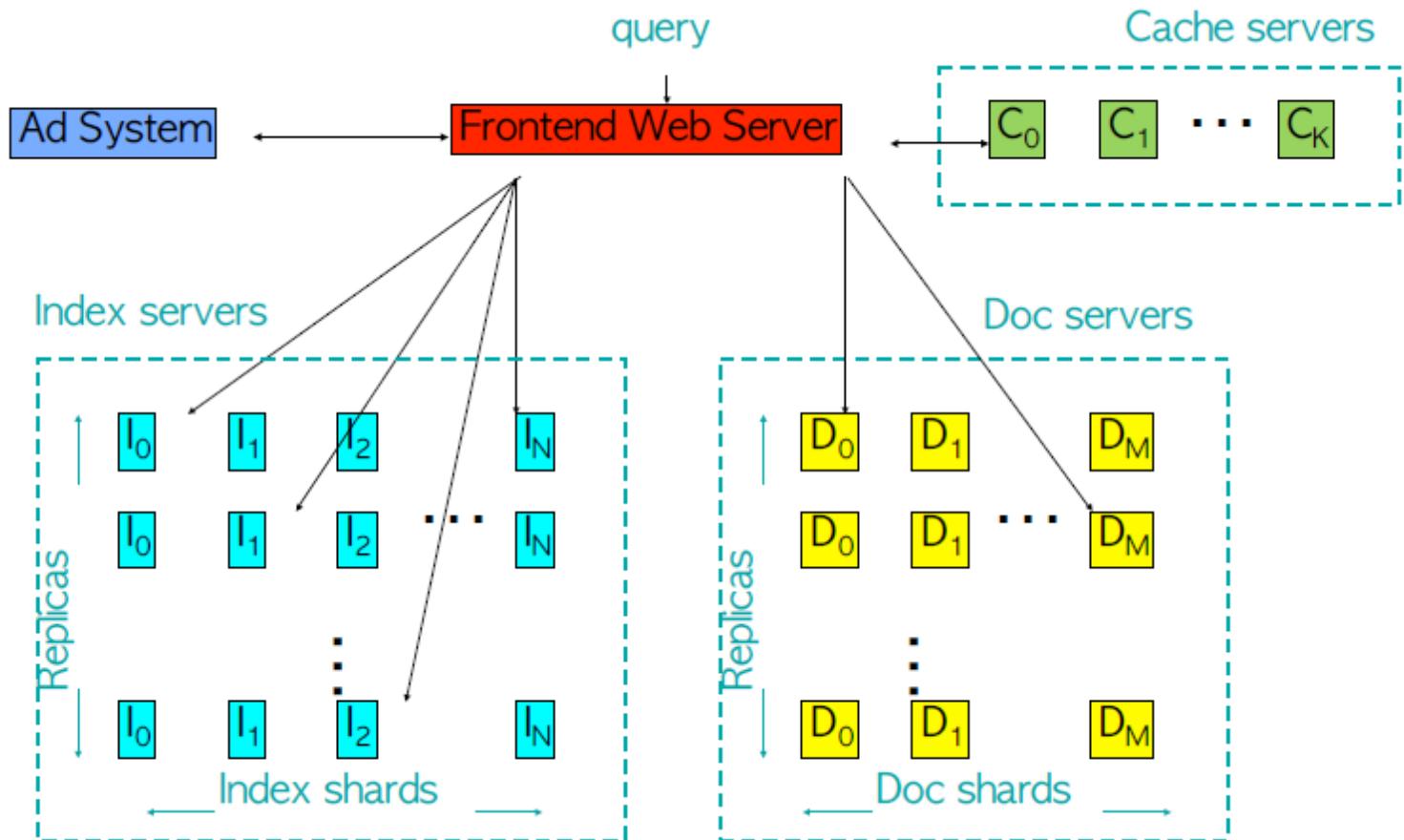


# Information Retrieval Challenges for Google from 1999 to 2019

- According to Jeff Dean, Google fellow how things have changed:
  - #doc: ~70 million to many billion
  - # queries processed/day
  - size of the document index
  - update latency: months to minutes
  - average query latency: <1s to <0.15s
  - more machines



# Google Serving System circa 1999



A database **shard** is a horizontal partition of data in a database or search engine. Each individual partition is referred to as a **shard**. Each **shard** is held on a separate database server instance, to spread the load.

**Logical Entities**

- URL Server
- Crawler – across multiple machines
- Store Server
- Repository – all web pages
- Indexer – parses documents
- URL Resolver – converts relative URLs
- Barrels – contain words in documents
- Sorter – takes barrels sorted by document and re Sorts by word
- Lexicon – word/phrase index

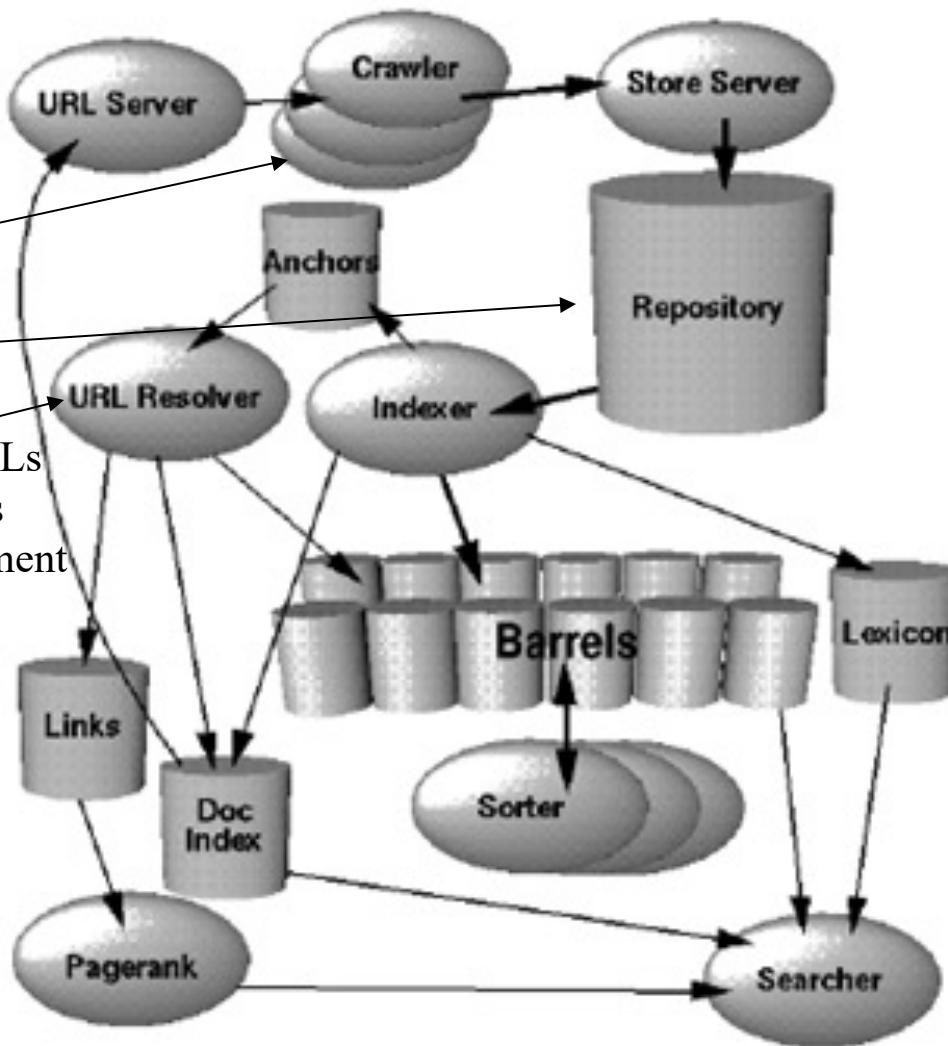


Diagram from

*"The Anatomy of a Large-Scale Hypertextual Web Search Engine"*

<http://infolab.stanford.edu/~backrub/google.html>

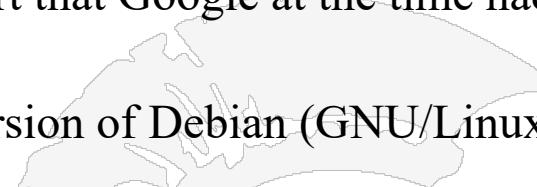
Figure 1. High Level Google Architecture

# Google Query Processing Basic Steps

1. Parse the query
2. Convert words into wordIDs using the lexicon
3. Select the barrels that contain documents which match the wordIDs
4. Scan through the document list until there is a document that matches all of the search terms
5. Compute the rank of that document for the query (using PageRank as one component)
6. Repeat step 4 until no documents are found and we've examined all of the barrels
7. Sort the set of returned documents that have been matched by document rank and return the top k.

# Google Architecture Today

- **Google data centers** combine large amounts of digital storage (mainly hard drives and SSDs), compute nodes organized in aisles of racks, internal and external networking, environmental controls and operations software (especially as concerns load balancing and fault tolerance).
- Google data centers estimated in a July 2016 report that Google at the time had 2.5 million servers.
- As of 2014, Google used a heavily customized version of Debian (GNU/Linux).
- Google ranks as the third largest ISP
- Google data Centers in the U.S. →

**United States:**

- 1.Berkeley County, South Carolina
- 2.Council Bluffs, Iowa
- 3.Douglas County, Georgia
- 4.Jackson County, Alabama<sup>(3)</sup>
- 5.Lenoir, North Carolina
- 6.Montgomery County, Tennessee
- 7.Pryor Creek, Oklahoma at MidAmerica Industrial Park
- 8.The Dalles, Oregon

The Dalles data center is a \$600 million complex built in 2006 and is approximately the size of two American football fields, with cooling towers four stories high.

The site was chosen to take advantage of inexpensive hydroelectric power, and to tap into the region's large surplus of fiber optic cable

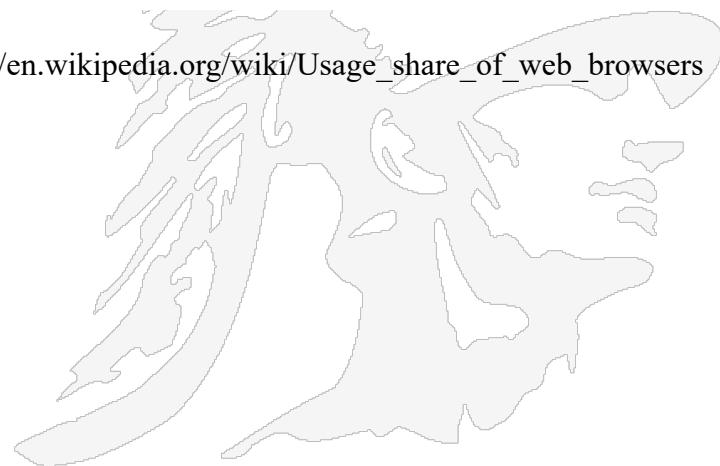


- Google has seven services claiming over a billion monthly active users
  1. Google Search,
  2. Google Maps,
  3. YouTube,
  4. Android,
  5. Gmail,
  6. Play Store, and
  7. Google Chrome
- there are now over 2.65 billion Chrome browsers in active use as of 2022
  - <https://backlinko.com/chrome-users/>
  - Google search is the default on Chrome; Google is the default search engine for Firefox, and Bing is the default for Microsoft's Edge and Internet Explorer

# Google is Much More Than Search

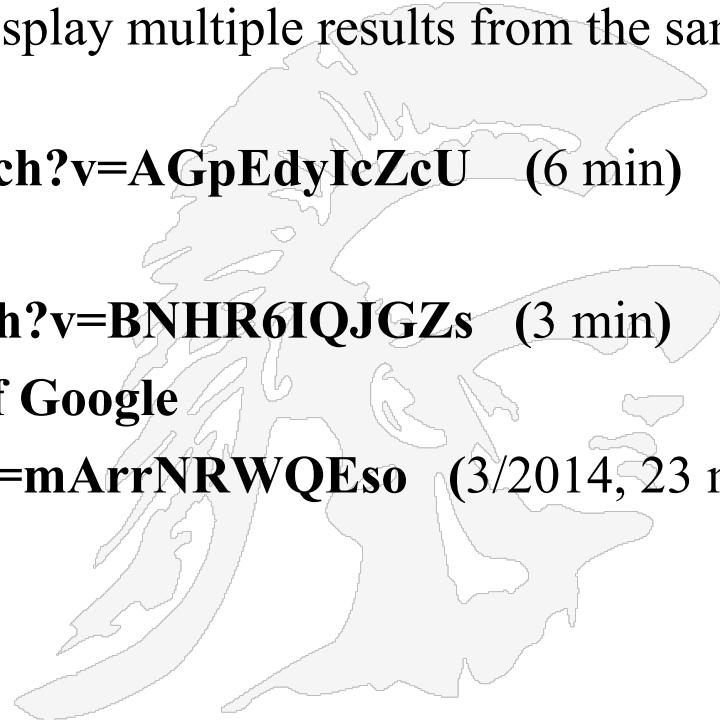
Browser	Usage share of all browsers		
	StatCounter <sup>[15]</sup> October 2021	NetMarketShare <sup>[16]</sup> October 2021	Wikimedia <sup>[17]</sup> October 2021
Chrome	64.67%	66.64%	52.5%
Safari	19.06%	13.92%	23.9%
Edge	4.10%	4.55%	3.0%
Firefox	3.66%	2.18%	4.4%
Samsung Internet	2.81%	3.04%	2.2%
Opera	2.36%	3.02%	1.0%
Others	3.34%	6.65%	13.0%

[https://en.wikipedia.org/wiki/Usage\\_share\\_of\\_web\\_browsers](https://en.wikipedia.org/wiki/Usage_share_of_web_browsers)

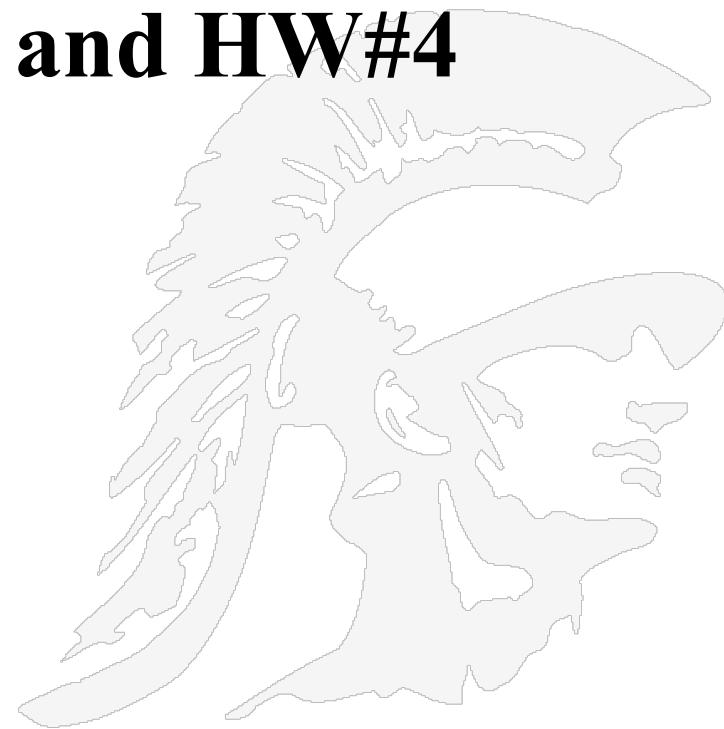


# Other Useful Videos

- **Matt Cutts videos**
  - How does Google search work?
    - <https://www.youtube.com/watch?v=KyCYyoGusqs> (7 min)
  - How does Google decide when to display multiple results from the same website
    - <https://www.youtube.com/watch?v=AGpEdyIcZcU> (6 min)
  - How Search Works
    - <http://www.youtube.com/watch?v=BNHR6IQJGZs> (3 min)
- **Larry Page on the future directions of Google**
  - <http://www.youtube.com/watch?v=mArrNRWQEso> (3/2014, 23 min)

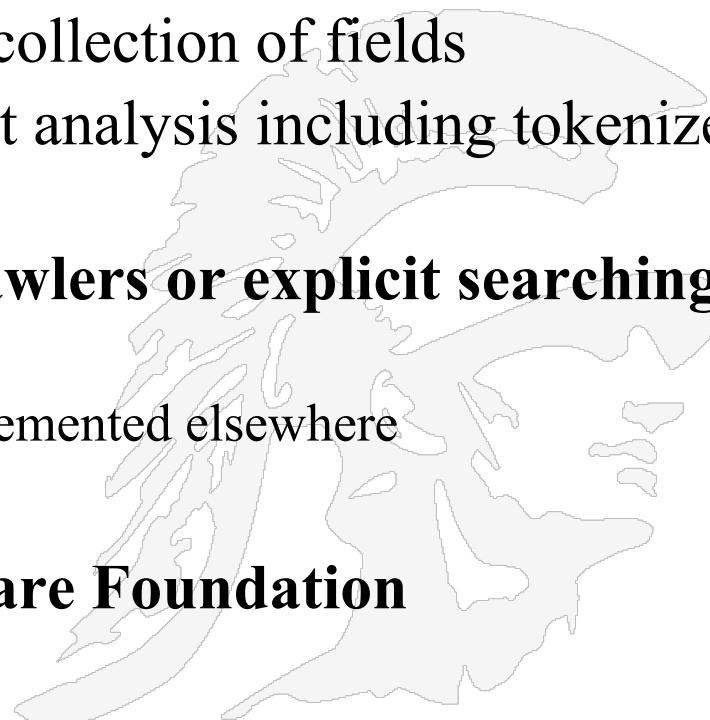


# Lucene, Solr and HW#4



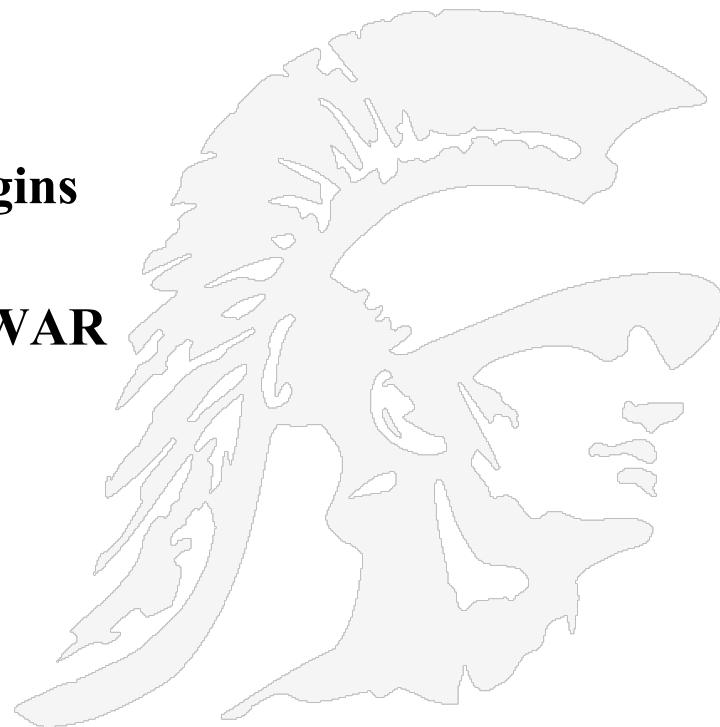
# What is Lucene

- **High performance, scalable, full-text search library**
- **Basically one adds documents to Lucene which creates an inverted index**
  - A document is defined as a collection of fields
  - Lucene includes flexible text analysis including tokenizers and filters
- **Lucene does NOT include crawlers or explicit searching of documents**
  - Searching the index must be implemented elsewhere
- **100% Java, no dependencies**
- **Offered by the Apache Software Foundation**

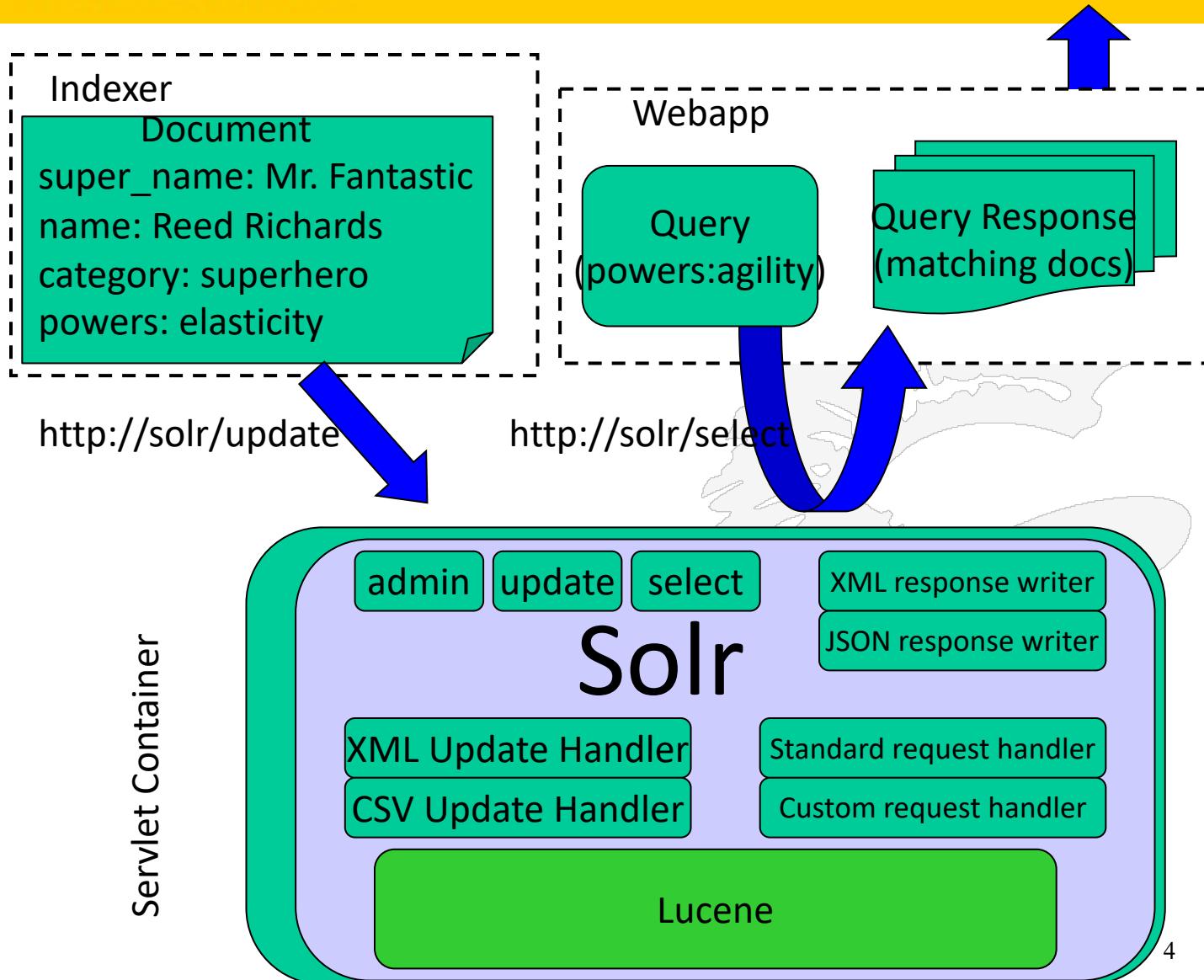


# What is Solr

- A full text search server based on Lucene
- Communicates with Lucene via XML/HTTP, JSON Interfaces
- Flexible data schema to define types and fields
- Output includes highlighting matches
- Configurable advanced caching
- Index Replication
- Extensible Open Architecture, Plugins
- Web Administration Interface
- Written in Java5, deployable as a WAR



# Top-Level Overview of Lucene/Solr





The screenshot shows the Apache Lucene homepage. At the top, there's a large green banner with the Apache Lucene logo. Below it, a dark banner states: "Apache Lucene set the standard for search and indexing performance. Lucene is the search core of both Apache Solr™ and Elasticsearch™." A "DOWNLOAD" button is prominently displayed. To the left, a section titled "Welcome to Apache Lucene" provides an overview of the project. On the right, there's a "Projects" sidebar listing "Lucene Core (Java)", "PyLucene", and "Open Relevance (Discontinued)".

## Solr Downloads

Official releases are usually created when the [developers](#) feel there are sufficient changes, improvements and bug fixes to warrant a release. Due to the voluntary nature of Solr, no releases are scheduled in advance.

### Solr 8.11.1

Solr 8.11.1 is the most recent Apache Solr release.

- Source release: [solr-8.11.1-src.tgz \[PGP\] \[SHA512\]](#)
- Binary releases: [solr-8.11.1.tgz \[PGP\] \[SHA512\]](#) / [solr-8.11.1.zip \[PGP\] \[SHA512\]](#)
- Docker: [solr:8.11.1](#)
- [Change log](#)

### Solr 7.7.3

Solr 7.7.3 is the last release in the 7.x series.

# Downloading Solr

Lucene  
home page



# Lucene Internals – Positional Inverted Index

**Document 1**

The bright blue butterfly hangs on the breeze.

**Document 2**

It's best to forget the great sky and to retire from every wind.

**Document 3**

Under blue sky, in bright sunlight, one need not search around.

Query  
 Q: "blue sky"

Inverted index

ID	Term	Document : position
1	best	2 : 3
2	blue	1: 3, 3 : 2
3	bright	1 : 2, 3 : 5
4	butterfly	1 : 4
5	breeze	1 : 8
6	forget	2 : 5
7	great	2 : 7
8	hangs	1 : 5
9	needs	3 : 8
10	retire	2 : 11
11	search	3 : 10
12	sky	2 : 8, 3 : 3
13	wind	2 : 14

Match on sequential terms

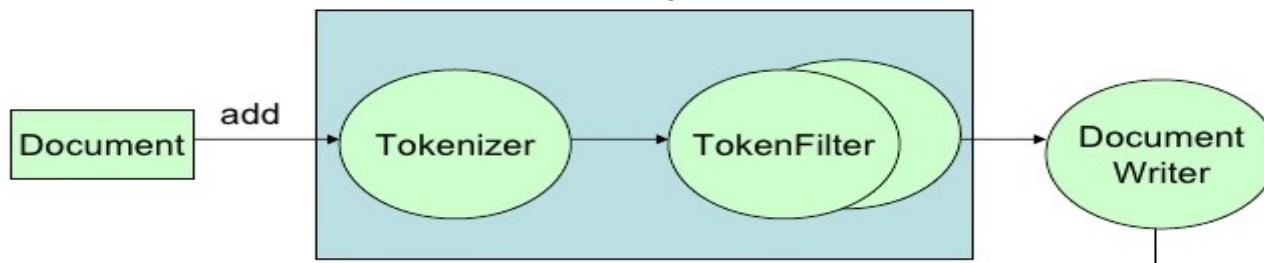
blue - 3 : 2  
 sky - 3 : 3

Search object

Document reference	Relevance
3	100%

*takes documents, parses them into terms and builds a positional inverted index;*

# Lucene Indexing Pipeline Analyzer



- **Analyzer : create tokens using a Tokenizer and/or applying Filters (Token Filters)**
  - Splits words at punctuation characters, removing punctuation. However, a dot that's not followed by whitespace is considered part of a token.
  - Splits words at hyphens, unless there's a number in the token, in which case the whole token is interpreted as a product number and is not split.
  - Recognizes email addresses and internet hostnames as one token.

## Two Lucene Scoring Concepts

### TF – IDF & cosine similarity

### Lucene scores using a combination of TF-IDF and vector closeness

$$w_{x,y} = tf_{x,y} \times \log \left( \frac{N}{df_x} \right)$$

**TF-IDF**

Term  $x$  within document  $y$

$tf_{x,y}$  = frequency of  $x$  in  $y$

$df_x$  = number of documents containing  $x$

$N$  = total number of documents

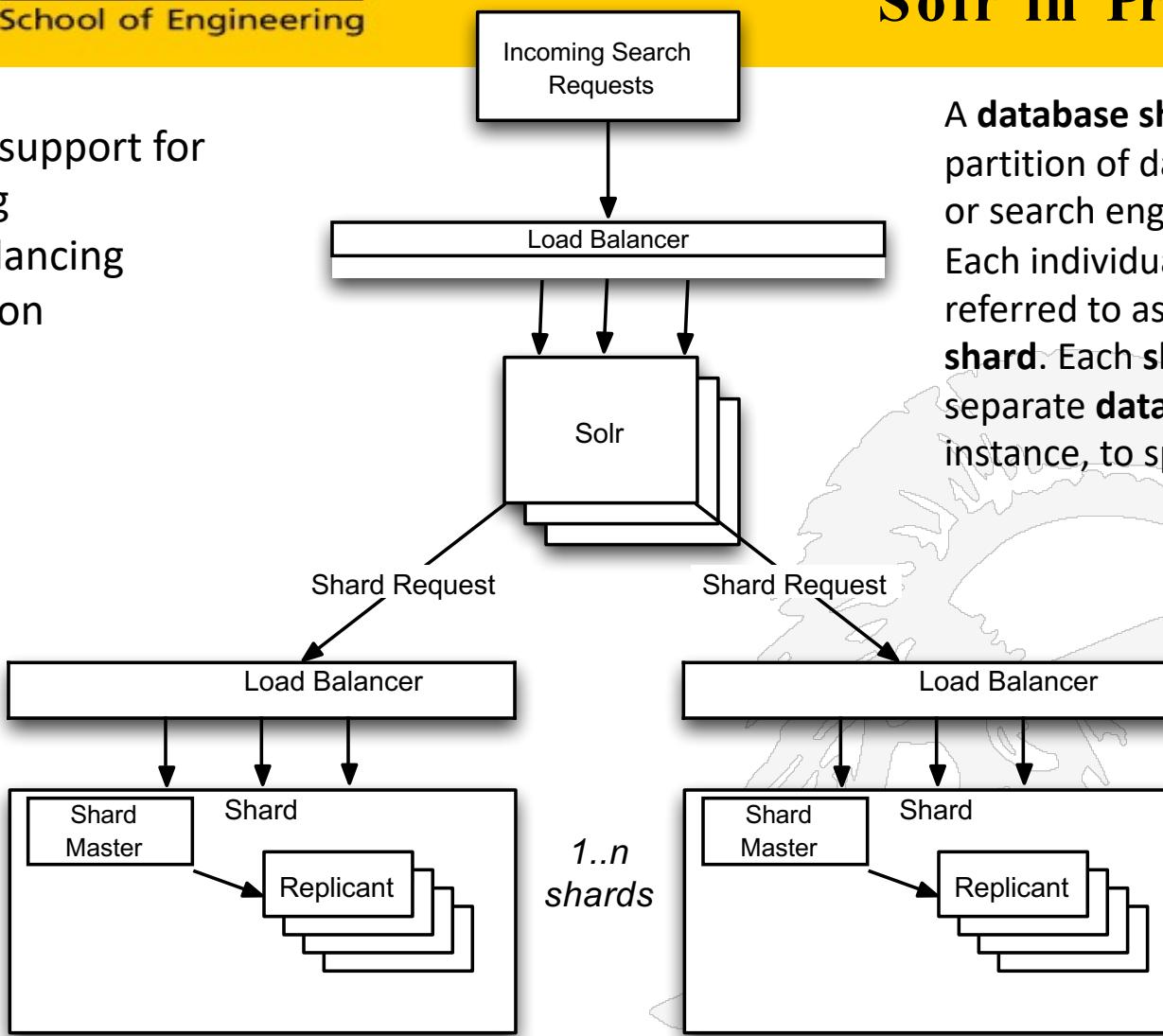
**TF - IDF = Term Frequency X Inverse Document Frequency**

cosine-similarity(query\_vector, document\_vector) =  $V(q) * V(d) / |V(q)| * |V(d)|$   
 where  $V(q) * V(d)$  is the dot product of the weighted vectors and  $|V(q)|$ ,  $|V(d)|$  are the Euclidean norms of the vectors (square root of the sum of squares)

for details see

[https://lucene.apache.org/core/4\\_0\\_0/core/org/apache/lucene/search/similarities/TFIDFSimilarity.html](https://lucene.apache.org/core/4_0_0/core/org/apache/lucene/search/similarities/TFIDFSimilarity.html)

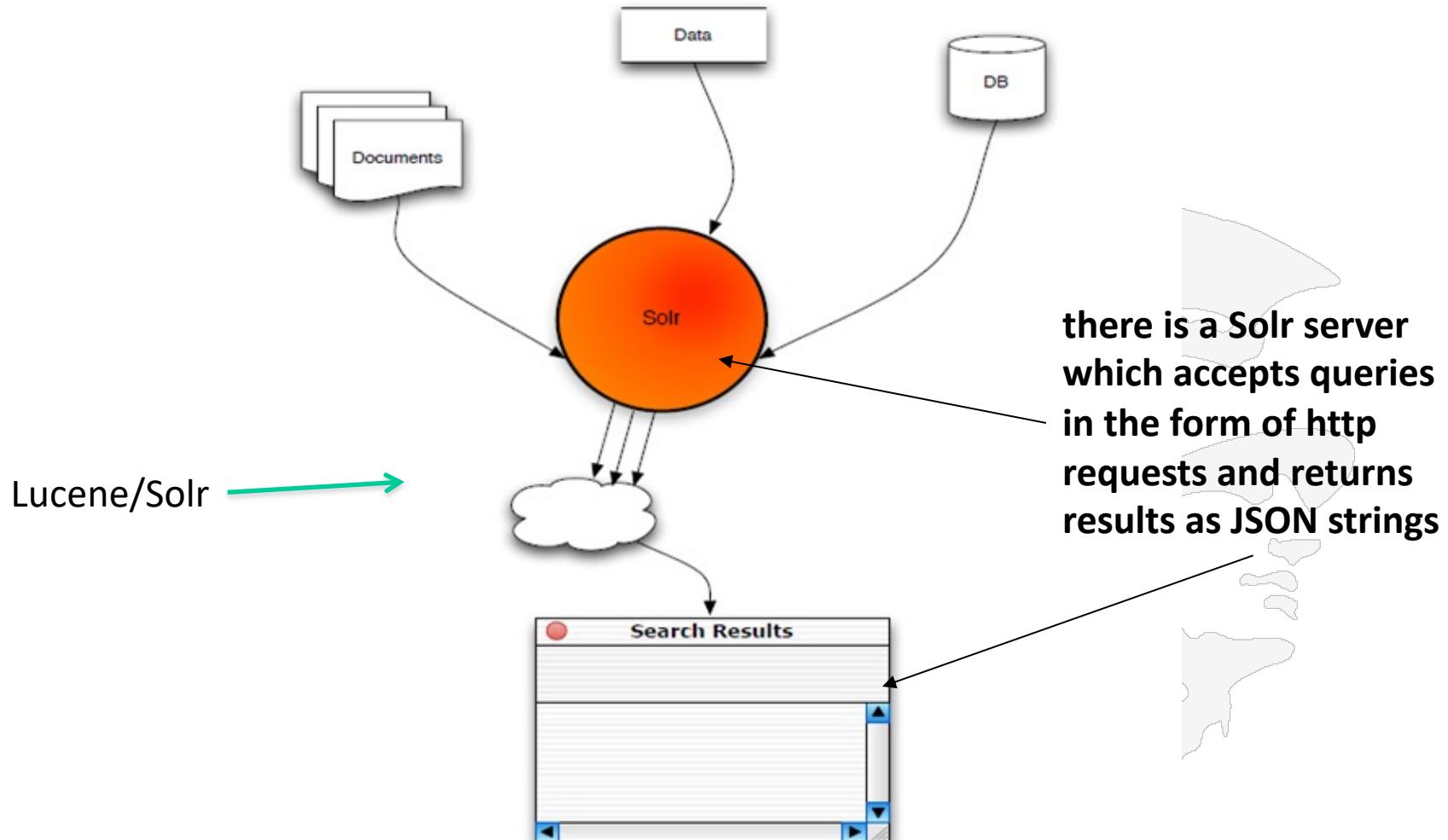
Includes support for  
-Sharding  
-Load balancing  
-replication



## Solr in Production

A **database shard** is a horizontal partition of data in a **database** or search engine. Each individual partition is referred to as a **shard**. Each **shard** is held on a separate **database server** instance, to spread the load.

# High Level Overview



# Solr Request and Result

`http://localhost:8983/solr/select?q=video&start=0&rows=2&fl=name,price`

```
<response><responseHeader><status>0</status>
<QTime>1</QTime></responseHeader>
<result numFound="16173" start="0">
 <doc>
 <str name="name">Apple 60 GB iPod with Video</str>
 <float name="price">399.0</float>
 </doc>
 <doc>
 <str name="name">ASUS Extreme N7800GTX/2DHTV</str>
 <float name="price">479.95</float>
 </doc>
</result>
</response>
```

- status, always 0 in a successful response
- QTime, the server-side query time in milliseconds
- numFound, the total number of documents matching the query
- start, the offset into the ordered list of results
- field types in <doc> include str, boolean, int, long, float, double, date, lst, arr
  - lst is a named list <lst><int name="foo">33</int><int name="bar">42</int></lst>
  - arr is an array <arr><int>33</int><int>42</int></arr>
- multivalued fields are returned in an <arr> element.

# How to Start Solr

**Complete installation instructions can be found at**

[https://solr.apache.org/guide/8\\_11/installing-solr.html](https://solr.apache.org/guide/8_11/installing-solr.html)

Once it is installed:

## 1. Start Solr

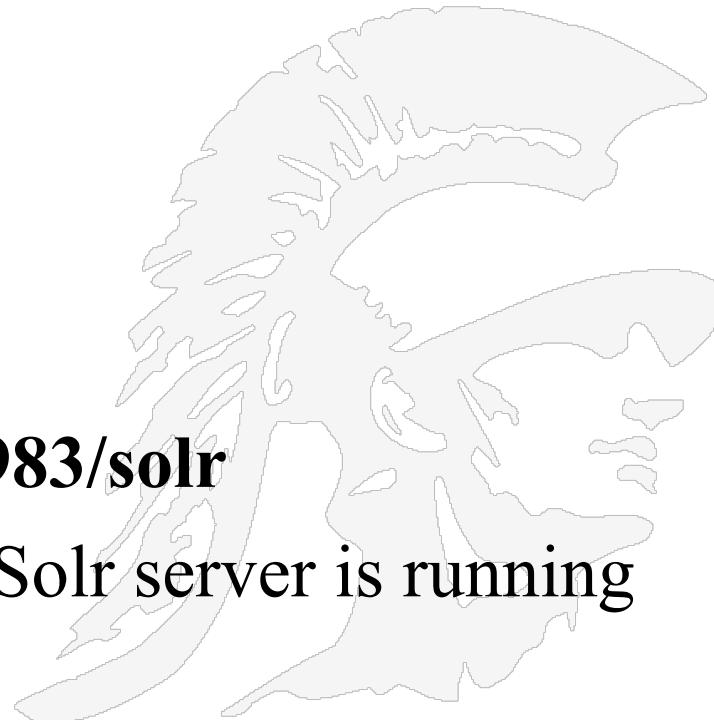
```
java -jar start.jar
```

## 2. Index your XML data

```
java -jar post.jar *.xml
```

## 3. Search <http://localhost:8983/solr>

localhost indicates the Solr server is running locally on port 8983



## Querying Data and XML Response

- HTTP GET or POST with parameters are used to specify queries
- E.g. here are 4 sample queries, some with various parameters

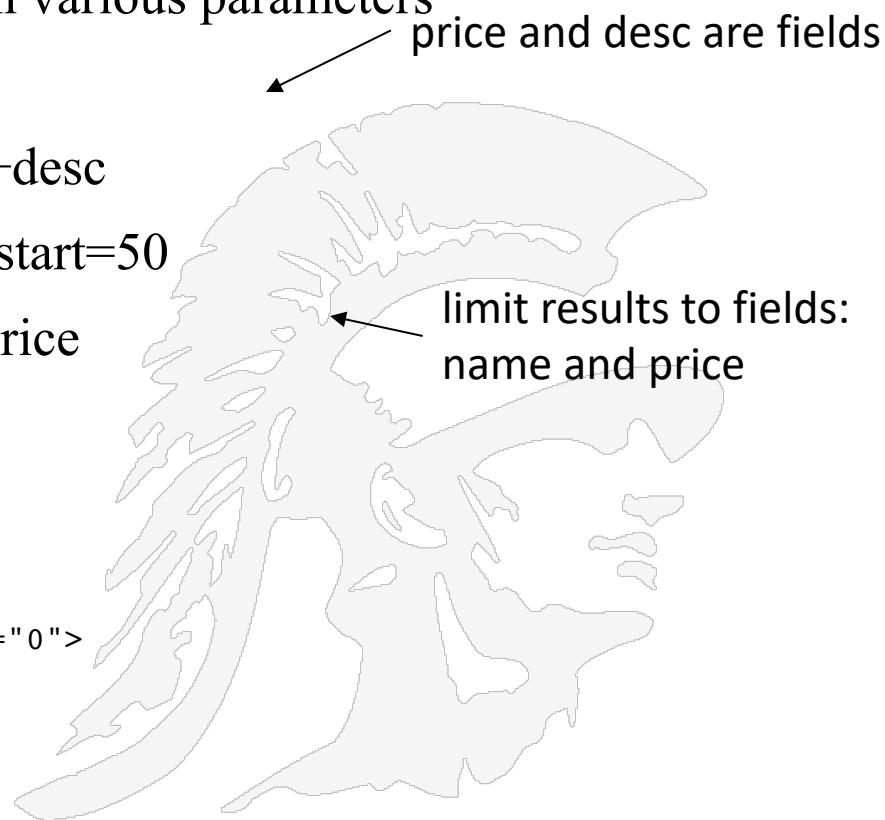
`http://solr/select?q=electronics`

`http://solr/select?q=electronics&sort=price+desc`

`http://solr/select?q=electronics&rows=50&start=50`

`http://solr/select?q=electronics&fl=name+price`

```
<response>
 <lst name="responseHeader">
 <int name="status">0</int>
 <int name="QTime">1</int>
 </lst>
 <result name="response" numFound="14" start="0">
 <doc>
 <arr name="cat">
 <str>electronics</str>
 <str>connector</str>
 </arr>
 ...
 ...
</result>
</response>
```



# Querying Data: Results

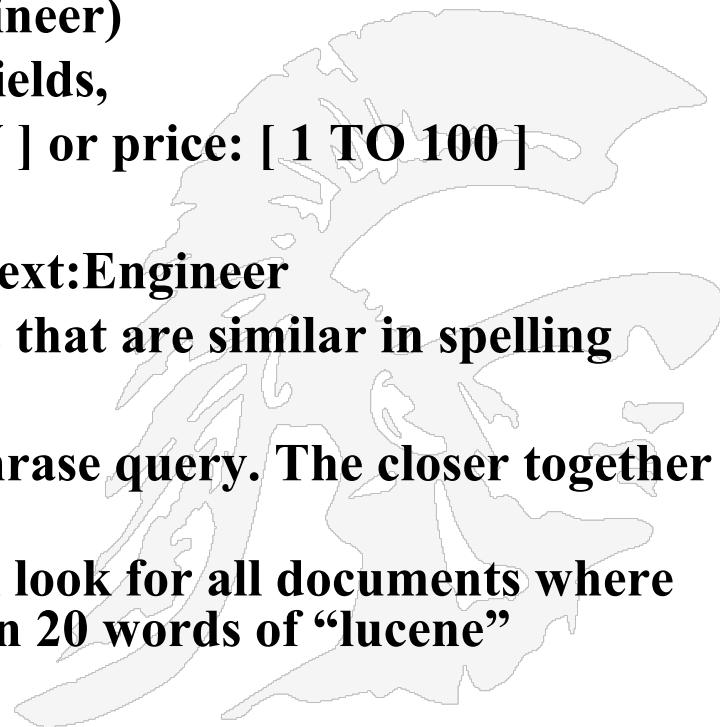
- The standard response format is JSON

```
{
 "responseHeader":{
 "zkConnected":true,
 "status":0,
 "QTime":8,
 "params":{
 "q":"foundation"}},
 "response":{ "numFound":4, "start":0, "maxScore":2.7879646, "docs": [
 {
 "id": "0553293354",
 "cat": ["book"],
 "name": "Foundation",
 "price": 7.99,
 "price_c": "7.99,USD",
 "inStock": true,
 "author": "Isaac Asimov",
 "author_s": "Isaac Asimov",
 "series_t": "Foundation Novels",
 "sequence_i": 1,
 "genre_s": "scifi",
 "_version_": 1574100232473411586,
 "price_c___l_ns": 799}]
 }}
```



## Query Types Supported by Solr

- Single and multi-term queries
  - ex fieldname:value or title: software engineer
- +, -, AND, OR, NOT operators are supported
  - ex. title: (software AND engineer)
- Range queries on date or numeric fields,
  - ex: timestamp: [ \* TO NOW ] or price: [ 1 TO 100 ]
- Boost queries:
  - e.g. title:Engineer ^1.5 OR text:Engineer
- Fuzzy search : is a search for words that are similar in spelling
  - e.g. roam~0.8 => noam
- Proximity Search : with a sloppy phrase query. The closer together the two terms appear, higher the score.
  - ex “apache lucene”~20 : will look for all documents where “apache” word occurs within 20 words of “lucene”



- **Search Engine**
  - Yandex.ru, DuckDuckGo.com
- **Newspaper**
  - Guardian.co.uk
- **Music/Movies**
  - Apple.com, Netflix.com
- **Events**
  - Stubhub.com, Eventbrite.com
- **Cloud Log Management**
  - Loggly.com
- **Others**
  - Whitehouse.gov
- **Jobs**
  - Indeed.com, Simplyhired.com, Naukri.com
- **Auto**
  - AOL.com
- **Travel**
  - Cleartrip.com
- **Social Network**
  - Twitter.com, LinkedIn.com, mylife.com

## Solr/Lucene is Used WorldWide

### Apache Solr Popularity

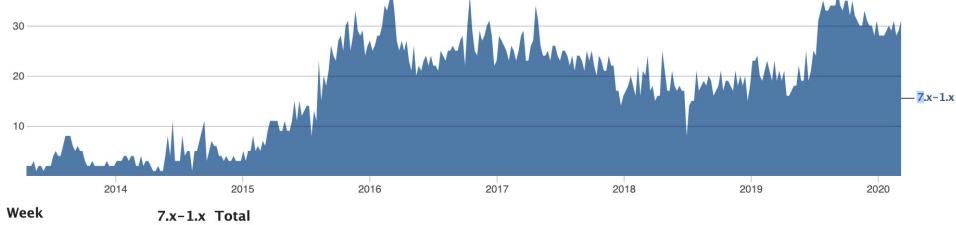
#### Usage statistics for Apache Solr Popularity

This page provides information about the usage of the *Apache Solr Popularity* project, including summaries across all versions and details for each release. For each week beginning on the given date the figures show the number of sites that reported they are using a given version of the project.

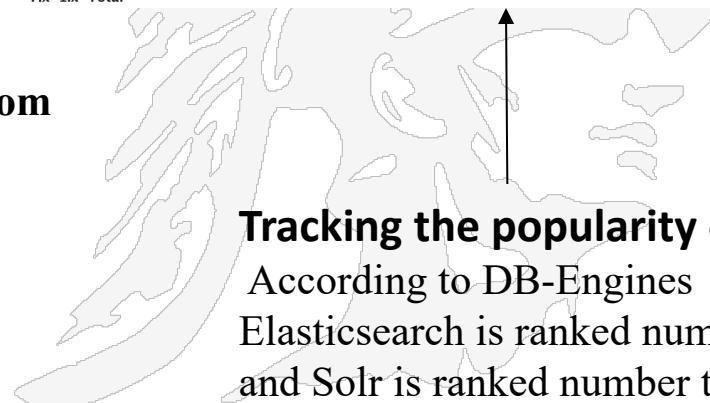
These statistics are incomplete; only Drupal websites using the [Update Status](#) module are included in the data. This module has been included with the download of Drupal since version 6.x so the data does not include older sites. [Read more information about how these statistics are calculated.](#)

[Apache Solr Popularity project page](#)  
[Usage statistics for all projects](#)

#### Weekly project usage



Week      7.x-1.x Total

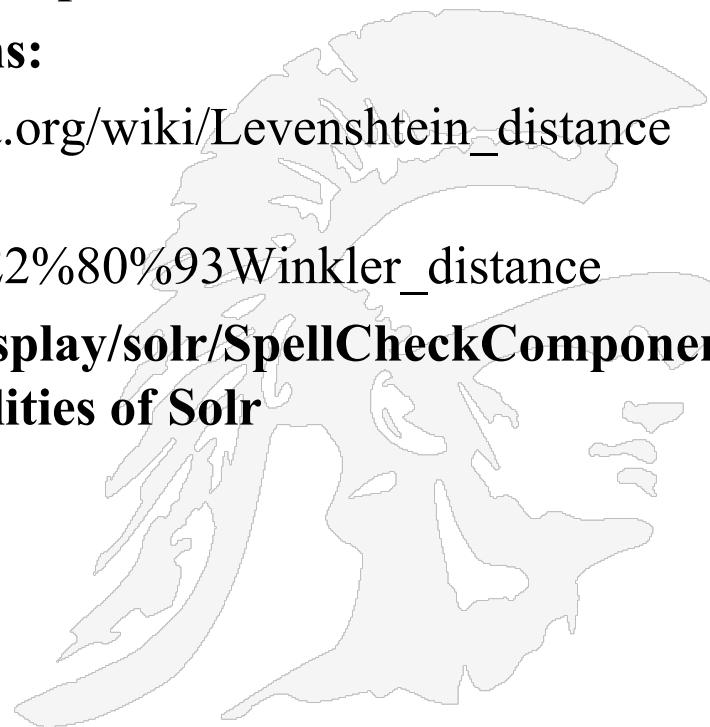


**Tracking the popularity of Solr**  
 According to DB-Engines  
 Elasticsearch is ranked number one,  
 and Solr is ranked number three

# Solr Includes Spell Checking

(you will implement this for homework #5)

- Not enabled by default, see example config to wire it in
  - [https://lucene.apache.org/solr/guide/8\\_11/spell-checking.html](https://lucene.apache.org/solr/guide/8_11/spell-checking.html)
- Uses file or index-based dictionaries for spell correction
- Supports pluggable distance algorithms:
  - Levenstein alg: [https://en.wikipedia.org/wiki/Levenshtein\\_distance](https://en.wikipedia.org/wiki/Levenshtein_distance)
  - JaroWinkler alg: ,  
[https://en.wikipedia.org/wiki/Jaro%E2%80%93Winkler\\_distance](https://en.wikipedia.org/wiki/Jaro%E2%80%93Winkler_distance)
- <https://cwiki.apache.org/confluence/display/solr/SpellCheckComponent> is a full discussion of the spellchecking abilities of Solr



# Solr Includes Autosuggestion (you will implement this for homework #5)

Enter your keywords:

Did you mean: teaching

teach	17
teachers	2
teacher	1
teach book	15
teach world	11
teach wide	11
teach teaching	9
teach computer	9

Find dinn|

dinner

dinner restaurant

dinner and drinks

dinner cruise

dinner and dancing

dinner date

dinner theater

dinner show

dinner buffet

dinner and live jazz



[https://lucene.apache.org/solr/guide/8\\_11/suggester.html](https://lucene.apache.org/solr/guide/8_11/suggester.html)

# HW#4 Exercise

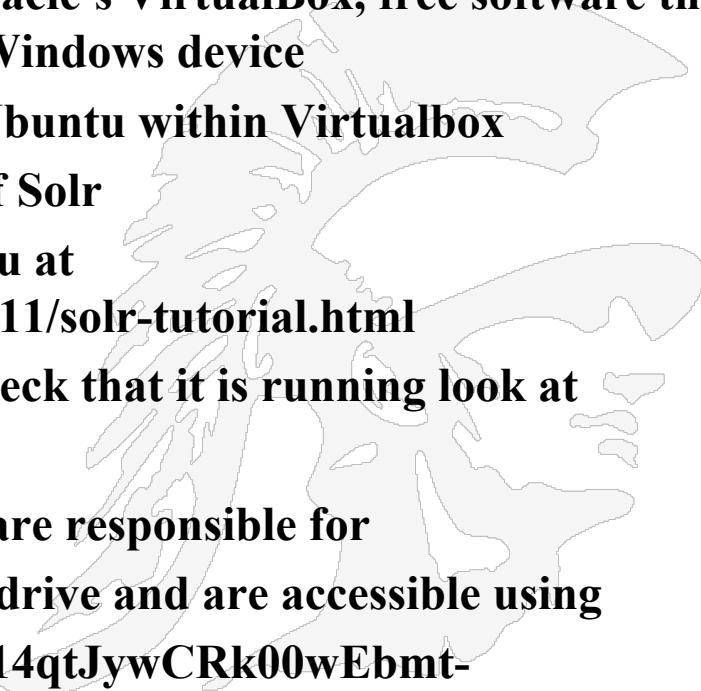
**There are several components to this homework; you will possibly not need all of them;**

- 1. Installing Ubuntu on Windows**
- 2. Installing Solr**
- 3. Using Solr to Index a web site**
- 4. The actual exercise - comparing ranking algorithms**
- 5. What to submit**
- 6. Grading guidelines**



# What you will need to do SUMMARY (1 of 2)

- 1. If you are using Windows, you will have to install Ubuntu on your machine**
  - if you have a Mac, skip to step 2
  - Solr requires a Unix platform
  - first you must download and install Oracle's VirtualBox, free software that will permit you to install Ubuntu on your Windows device
  - second you will download and install Ubuntu within Virtualbox
- 2. Download and install the current release of Solr**
  - there is a Quick Start Guide to help you at  
[https://lucene.apache.org/solr/guide/8\\_11/solr-tutorial.html](https://lucene.apache.org/solr/guide/8_11/solr-tutorial.html)
  - the Solr server should be started; to check that it is running look at
  - <http://localhost:8983/solr>
- 3. Download the reference news website you are responsible for**
  - 3 news websites are located on Google drive and are accessible using
  - <https://drive.google.com/drive/folders/14qtJywCRk00wEbmt-Pv3efyPIBqwcoZr>

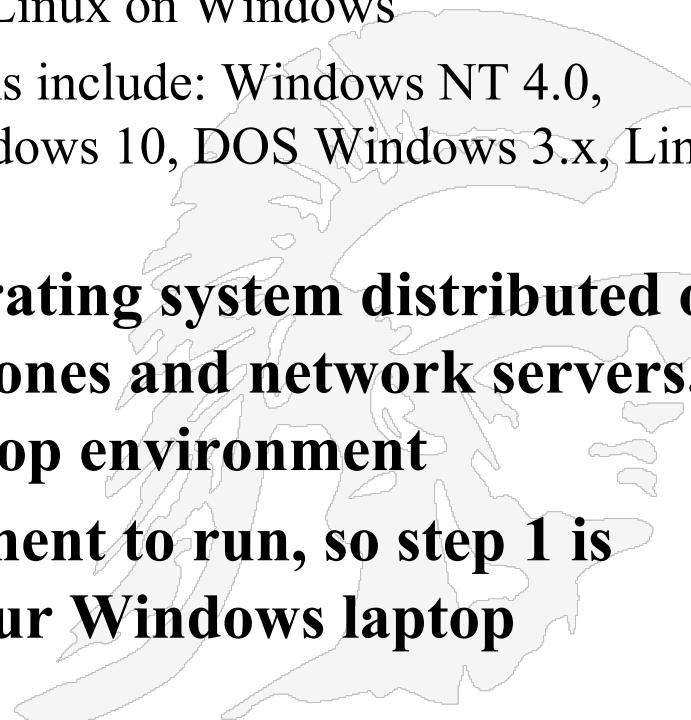


## What you will need to do SUMMARY (2 of 2)

4. Import the news website web pages you are responsible for into Solr
  - There are only three new websites, use last digits of your USC ID to determine which one you are responsible for
5. Using your laptop install or use an existing web server to create a website
  - Macs (I believe) have Apache pre-installed
  - in your website create a web page that looks like a Google query box
  - in your website write a program that takes a query and sends it to Solr
  - in your website write a program that accepts the results from Solr and displays them as a web page
6. Run the set of 9 queries at the end of SolrExercise.pdf
  - save the top ten results for each of the 9 queries
7. Use the NetworkX library, the downloaded web pages and Jsoup library to create a network graph of the downloaded web pages
  - use the PageRank function from NetworkX to create a file, pagerankFile.txt that contains individual page ranks for every downloaded web page
  - install the pagerankFile in Solr and direct Solr to use it on queries
8. Run the set of 9 queries at the end of SolrExercise.pdf
  - save the top ten results for each of the queries
9. compare the two sets of results and describe any overlap

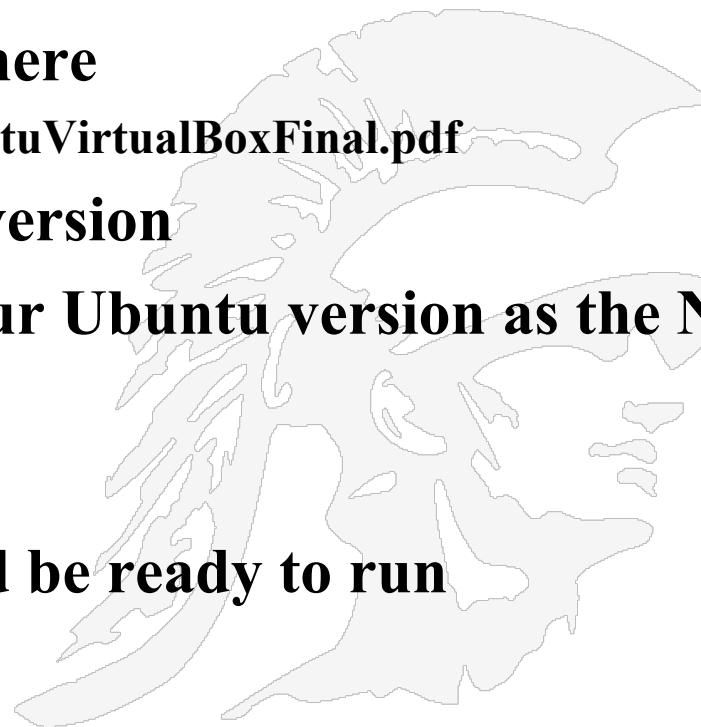
## Step 1: Ubuntu with VirtualBox

- **VirtualBox is an open source, freely available Windows application (it also runs on other platforms) that lets you run multiple operating systems on your single machine**
  - E.g. run Windows on a Mac, run Linux on Windows
  - Major supported operating systems include: Windows NT 4.0, Windows 2000, Windows 8, Windows 10, DOS Windows 3.x, Linux, Solaris, FreeBSD, OpenBSD
- **Ubuntu is a Linux-based operating system distributed on personal computers, smartphones and network servers. It uses Unity as its default desktop environment**
- **Solr requires a Unix environment to run, so step 1 is required if you plan to use your Windows laptop**

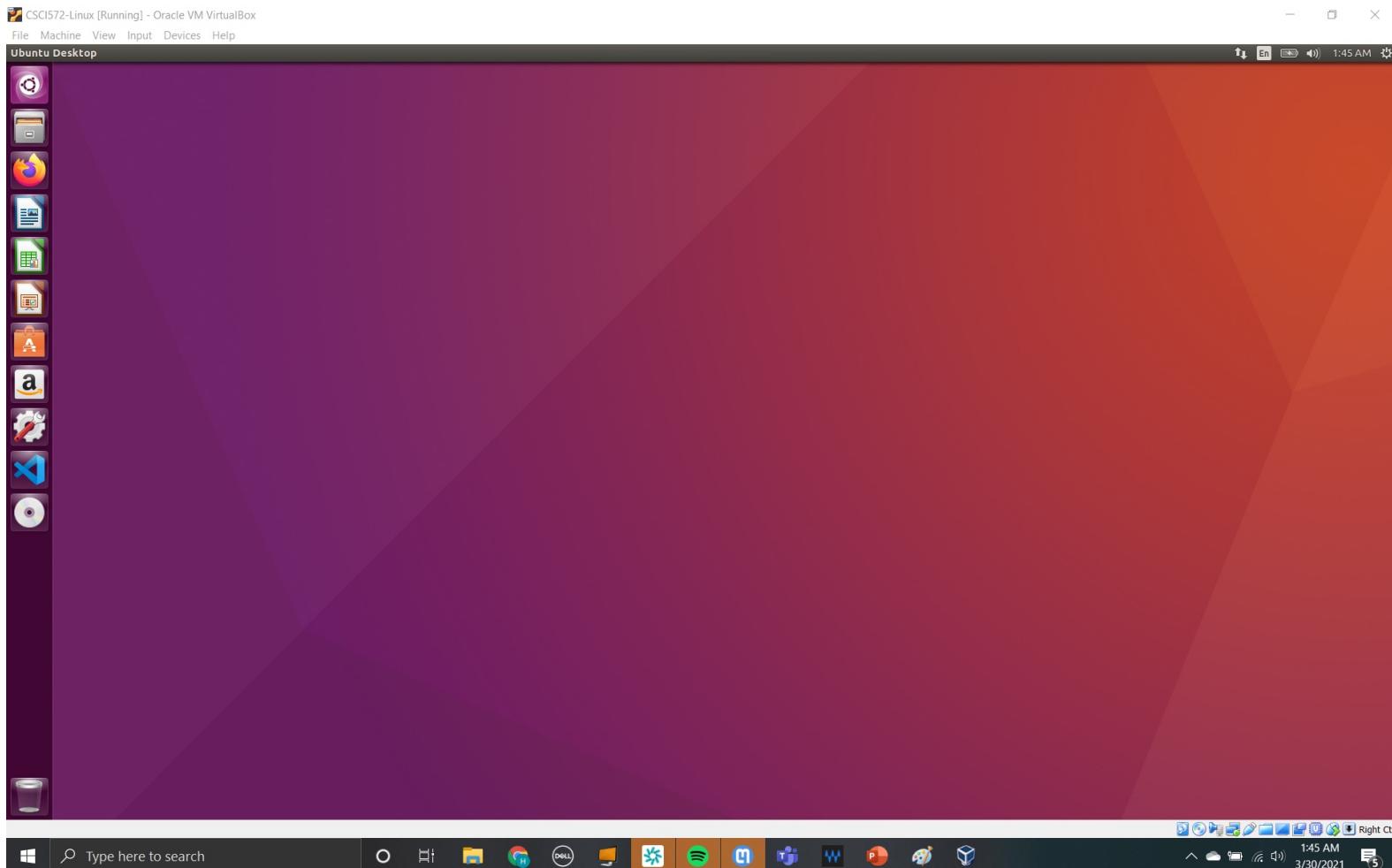


## Step 1: Setting Up Ubuntu with VirtualBox

- 1. Download the free version of VirtualBox for Windows machines**
  - Instructions can be found here**  
<http://csci572.com/2022Spring/hw4/UbuntuVirtualBoxFinal.pdf>
- 2. Download the Ubuntu 64-bit version**
- 3. Run VirtualBox and select your Ubuntu version as the New Application**
- 4. Set various parameters**
- 5. Install Ubuntu and you should be ready to run**

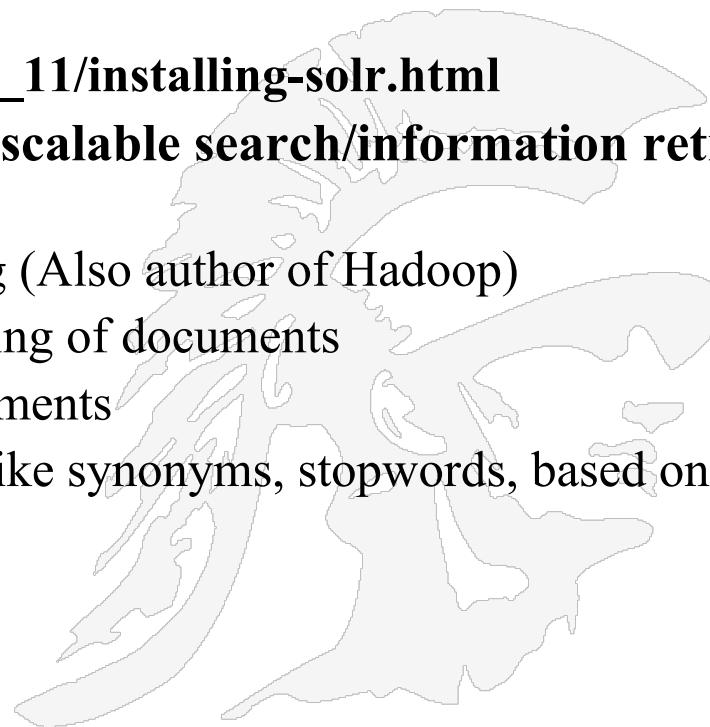


# Your Ubuntu Desktop



## Step 2: Installing Solr

- Solr is an open source enterprise search server based on the Lucene Java search library
- Instructions for downloading and installing Solr can be found here
  - [https://solr.apache.org/guide/8\\_11/installing-solr.html](https://solr.apache.org/guide/8_11/installing-solr.html)
- Lucene is a fast, high performance, scalable search/information retrieval library
  - Initially developed by Doug Cutting (Also author of Hadoop)
  - it provides for Indexing and Searching of documents
  - produces an Inverted Index of documents
  - Provides advanced Search options like synonyms, stopwords, based on similarity, proximity.



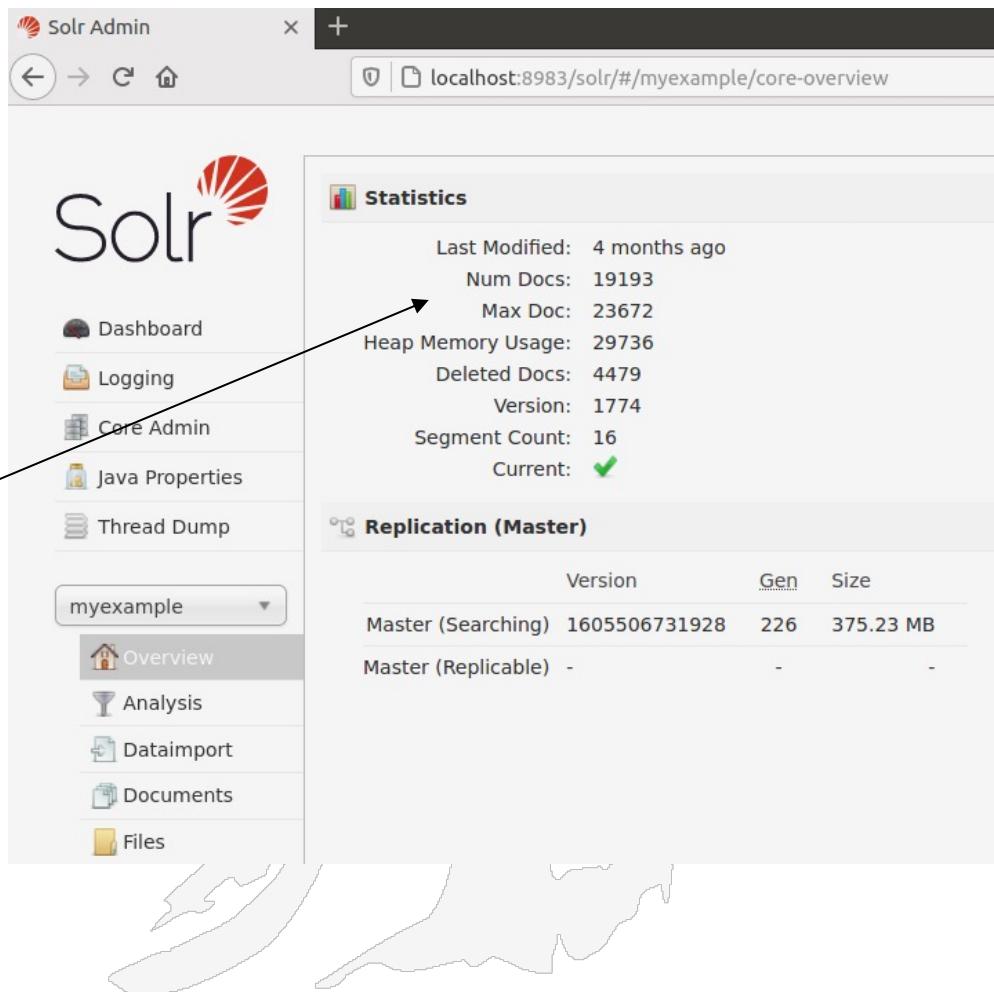
# Downloading Your Data Set of Web Pages

- We have created a reference set of web pages for all news websites and placed them on Google drive
- Below is a URL that will give you access to these data sets
  - Each data set contains three files
- The folders are:
  - NYTIMES [01 ~ 30]
  - FOXNEWS [31 ~ 60]
  - LATIMES [61 ~ 00]
- Download only the folder you are responsible for
  - Check your USC ID against the list in the exercise
- here is the URL again
  - <https://drive.google.com/drive/folders/14qtJywCRk00wEbmt-Pv3efyPIBqwcoZr>



## Step 3: Use Solr to Index a Web Site

- start the Solr server
- start a new Solr core
- Use Tika to import your saved files
- Use the Solr interface to check that the files have been properly indexed
- Note the URL:  
**localhost:8983/solr/#/myexample**
- 19193 docs successfully indexed



The screenshot shows the Solr Admin interface for the 'myexample' core. The left sidebar lists 'Dashboard', 'Logging', 'Core Admin', 'Java Properties', and 'Thread Dump'. Below that is a dropdown menu set to 'myexample' with options 'Overview' (which is selected), 'Analysis', 'Dataimport', 'Documents', and 'Files'. The main content area has two tabs: 'Statistics' and 'Replication (Master)'. The 'Statistics' tab displays various metrics:

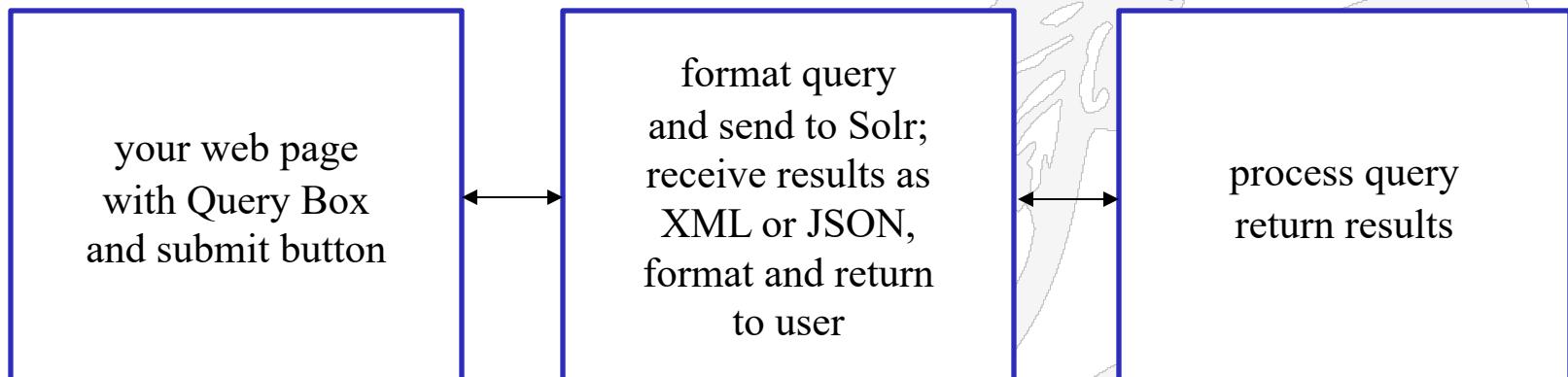
Last Modified	4 months ago
Num Docs	19193
Max Doc	23672
Heap Memory Usage	29736
Deleted Docs	4479
Version	1774
Segment Count	16
Current	✓

The 'Replication (Master)' tab shows the master node's status:

Version	Gen	Size
Master (Searching)	1605506731928	226 375.23 MB
Master (Replicable)	-	-

## Step 4: Comparing Search Engine Ranking Algorithms

1. You should download the reference files for the news website you are responsible for;
2. You should install Solr as described previously;
3. Take the web pages from the reference files and index them in Solr, as described in earlier slides
4. Build a front end to Solr that permits a visitor to enter a query and get matching results
5. Solr will return the results in JSON format; your server needs to take the results and format them for the user



# Apache Solr Client APIs

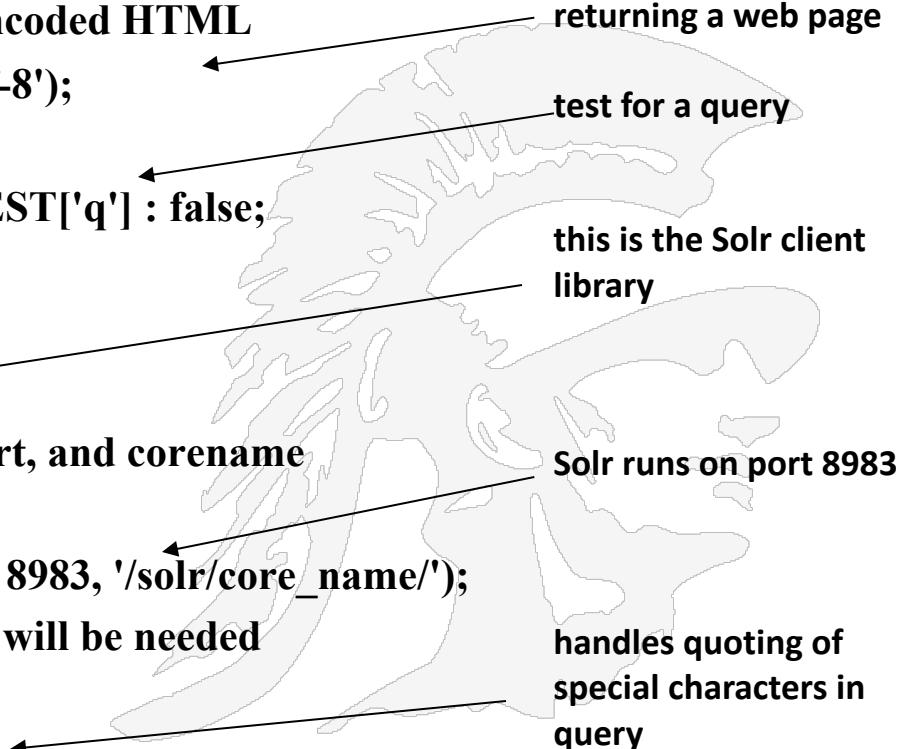
- **http is the protocol to be used between client applications and Solr**
- **Clients use Solr's five basic operations: query, index, delete, commit, and optimize**
- **JavaScript is the standard client API and there is no package to be installed**
  - http requests can be sent using XMLHttpRequest
  - Solr will respond with JSON output
- **there is an output format specifically for Python**
  - See [https://lucene.apache.org/solr/guide/8\\_11/using-python.html](https://lucene.apache.org/solr/guide/8_11/using-python.html)
- **Full list of Solr API clients can be found here**
  - [https://lucene.apache.org/solr/guide/8\\_11/client-api-lineup.html](https://lucene.apache.org/solr/guide/8_11/client-api-lineup.html)



## Step 4

Here is a PHP client that accepts input from the user in a HTML form, and sends the request to the Solr server. After the Solr server processes the query, it returns the results which is parsed by the PHP program and formatted for display

```
<?php
// make sure browsers see this page as utf-8 encoded HTML
header('Content-Type: text/html; charset=utf-8');
$limit = 10;
$query = isset($_REQUEST['q']) ? $_REQUEST['q'] : false;
$results = false;
if ($query)
{ require_once('Apache/Solr/Service.php');
// create a new solr service instance - host, port, and corename
// path (all defaults in this example)
$solr = new Apache_Solr_Service('localhost', 8983, '/solr/core_name/');
// if magic quotes is enabled then stripslashes will be needed
if (get_magic_quotes_gpc() == 1)
{ $query = stripslashes($query); }
```

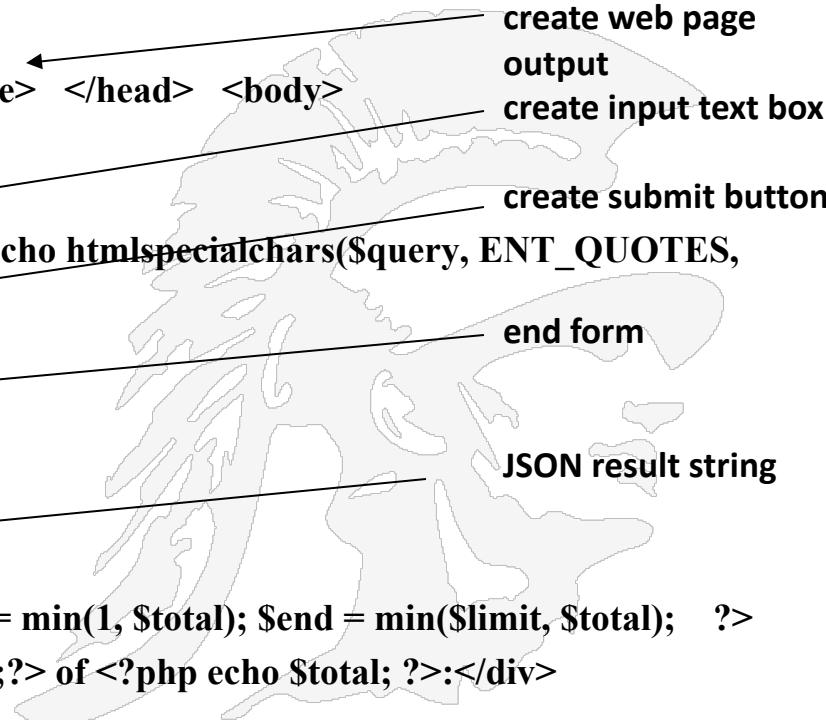


- returning a web page
- test for a query
- this is the Solr client library
- Solr runs on port 8983
- handles quoting of special characters in query

## PhP Program (2 of 3)

```
try
{ $results = $solr->search($query, 0, $limit); } ← send query to Solr
catch (Exception $e) catch any exception
{ die("<html><head><title>SEARCH EXCEPTION</title><body><pre>{$e-
>__toString()}</pre></body></html>"); } }
?>
<html> <head> <title>PHP Solr Client Example</title> </head> <body>
<form accept-charset="utf-8" method="get">
<label for="q">Search:</label>
<input id="q" name="q" type="text" value="<?php echo htmlspecialchars($query, ENT_QUOTES,
'utf-8'); ?>"/>
<input type="submit"/>
</form>
<?php
// display results
if ($results)
{ $Total = (int) $results->response->numFound; $start = min(1, $Total); $end = min($limit, $Total); ?>
<div>Results <?php echo $start; ?> - <?php echo $end;?> of <?php echo $total; ?>:</div>

```



send query to Solr  
catch any exception

create web page  
output  
create input text box  
create submit button

end form

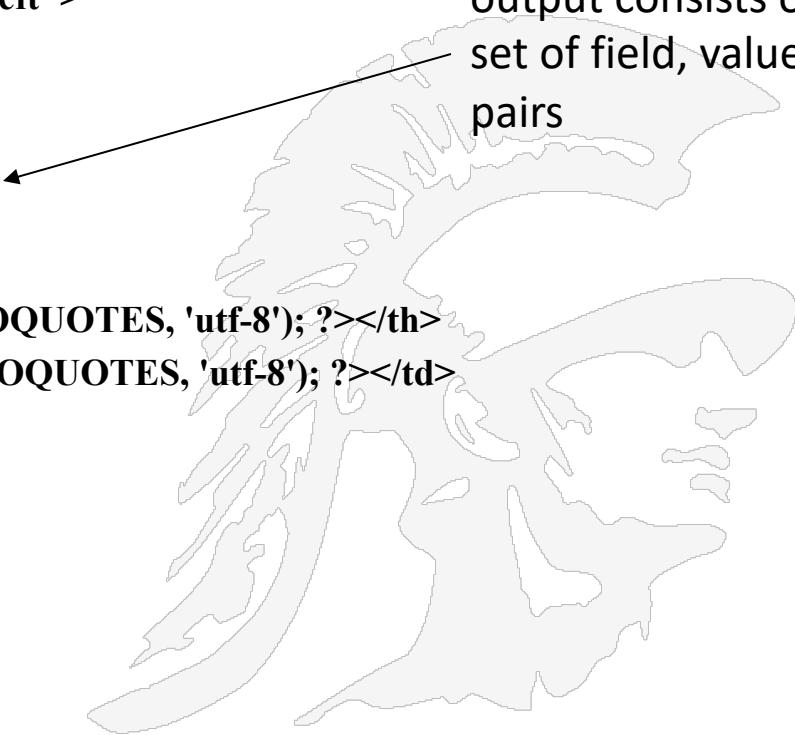
JSON result string

## PhP Program (3 of 3)

```
<?php
// iterate result documents
foreach ($results->response->docs as $doc)
{ ?>
<table style="border: 1px solid black; text-align: left">
<?php
// iterate document fields / values
foreach ($doc as $field => $value)
{ ?>
<tr>
<th><?php echo htmlspecialchars($field, ENT_NOQUOTES, 'utf-8'); ?></th>
<td><?php echo htmlspecialchars($value, ENT_NOQUOTES, 'utf-8'); ?></td>
</tr>
<?php }
?> </table>
<?php } ?>

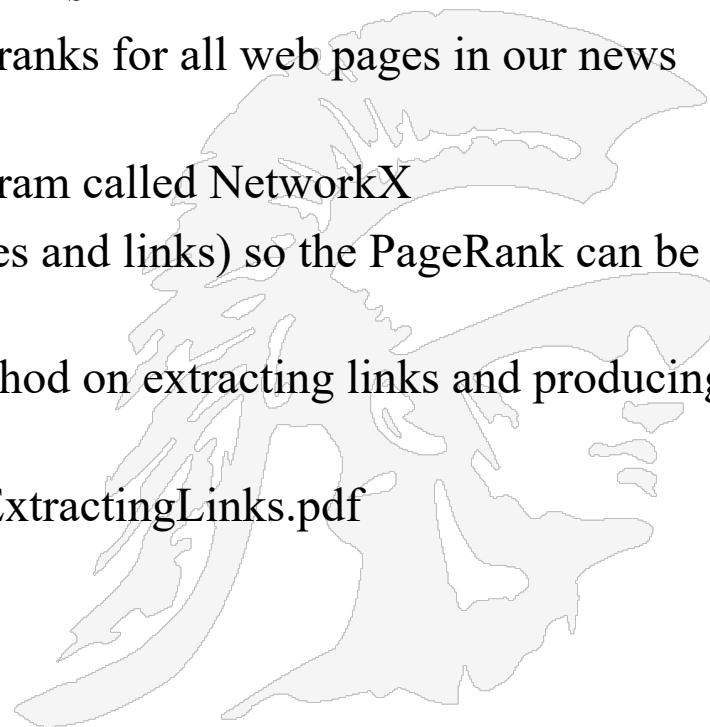
<?php
} ?></body> </html>
```

output consists of a set of field, value pairs

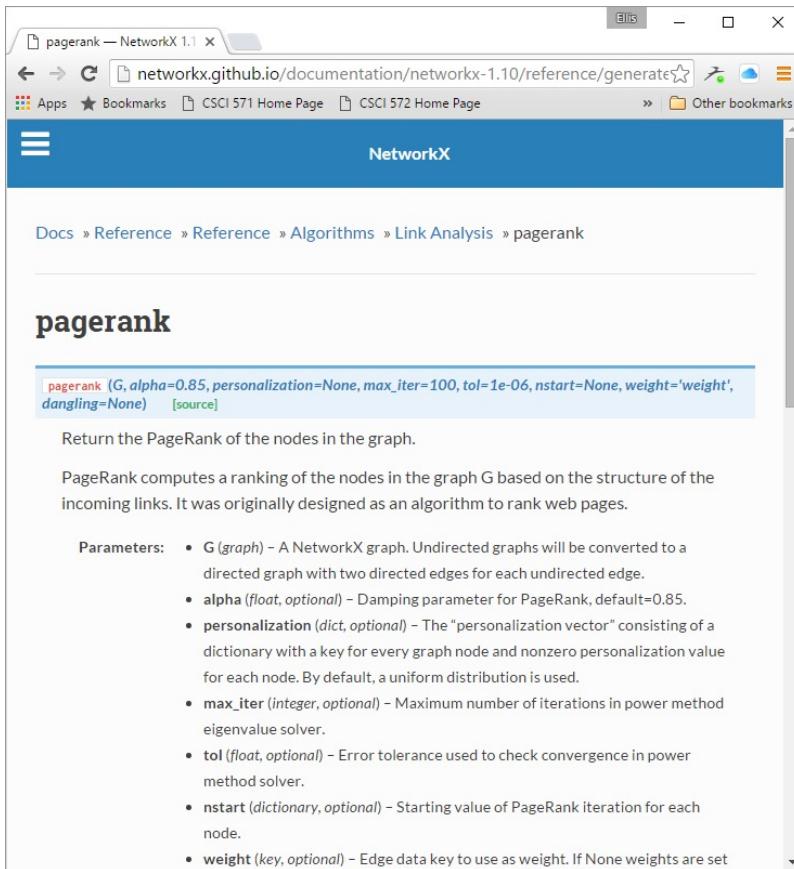


# Comparing Ranking Algorithms

- we have already seen the built-in Solr ranking method
  - Based on Lucene
- Solr permits alternative ranking algorithms
  - to use PageRank we must create pageranks for all web pages in our news website
  - to do this we use an open source program called NetworkX
  - we need to create the web graph (pages and links) so the PageRank can be determined
  - See the following document for a method on extracting links and producing the NetworkX file:
  - <http://csci572.com/2022Spring/hw4/ExtractingLinks.pdf>



# [http://networkx.github.io/documentation/networkx-1.10/reference/generated/networkx.algorithms.link\\_analysis.pagerank\\_alg.pagerank.html](http://networkx.github.io/documentation/networkx-1.10/reference/generated/networkx.algorithms.link_analysis.pagerank_alg.pagerank.html)



The screenshot shows a web browser window titled "pagerank — NetworkX 1.1". The address bar contains the URL "networkx.github.io/documentation/networkx-1.10/reference/generated/networkx.algorithms.link\_analysis.pagerank\_alg.pagerank.html". The page content is the documentation for the `pagerank` function. It includes the function signature `pagerank(G, alpha=0.85, personalization=None, max_iter=100, tol=1e-06, nstart=None, weight='weight', dangling=None)`, a brief description stating it returns PageRank values, and a detailed description explaining PageRank computation based on graph structure. Below this is a "Parameters" section listing various optional parameters with their descriptions.

**Parameters:**

- `G (graph)` – A NetworkX graph. Undirected graphs will be converted to a directed graph with two directed edges for each undirected edge.
- `alpha (float, optional)` – Damping parameter for PageRank, default=0.85.
- `personalization (dict, optional)` – The “personalization vector” consisting of a dictionary with a key for every graph node and nonzero personalization value for each node. By default, a uniform distribution is used.
- `max_iter (integer, optional)` – Maximum number of iterations in power method eigenvalue solver.
- `tol (float, optional)` – Error tolerance used to check convergence in power method solver.
- `nstart (dictionary, optional)` – Starting value of PageRank iteration for each node.
- `weight (key, optional)` – Edge data key to use as weight. If None weights are set

- You are going to use an open source PageRank algorithm, networkx, located at URL above;
- networkx is a Python program
- You need to create a graph that the PageRank algorithm can work on

## Important Parameters:

- a NetworkX graph
- a damping parameter (e.g. 0.85)
- maximum number of iterations
- error tolerance
- starting Page Rank value of nodes
- (see expansion on next slide)

[http://networkx.github.io/documentation/networkx-1.10/\\_modules/networkx/algorithms/link\\_analysis/pagerank\\_alg.html#pagerank](http://networkx.github.io/documentation/networkx-1.10/_modules/networkx/algorithms/link_analysis/pagerank_alg.html#pagerank)

```
def pagerank(G, alpha=0.85, personalization=None,
 max_iter=100, tol=1.0e-6, nstart=None, weight='weight', dangling=None):
```

Return the PageRank of the nodes in the graph.

#### Parameters

**G : graph:** A NetworkX graph.

**alpha : float, optional** Damping parameter for PageRank, *default=0.85*.

**personalization:** dict, optional; By *default*, a uniform distribution is used.

**max\_iter :** integer, optional Maximum number of iterations in power method eigenvalue solver.

**tol :** float, optional; Error tolerance used to check convergence in power method solver.

**nstart :** dictionary, optional Starting value of PageRank iteration for each node.

**weight :** key, optional; If None weights are set to 1.

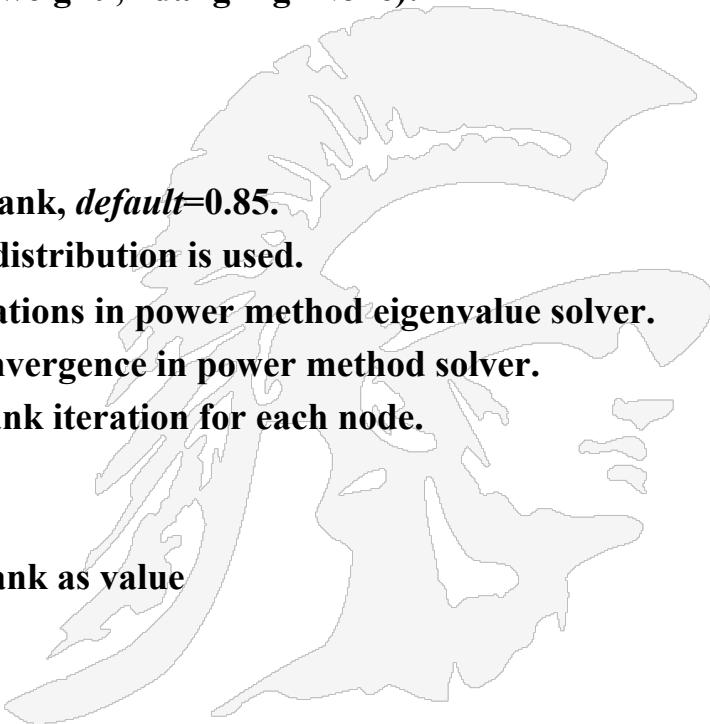
#### Returns

**pagerank :** dictionary; Dictionary of nodes with PageRank as value

#### Examples

```
>>> G = nx.DiGraph(nx.path_graph(4))
```

```
>>> pr = nx.pagerank(G, alpha=0.9)
```



# Some NetworkX Operations

- Create an empty graph with no nodes and no edges.

```
>>> import networkx as nx
```

```
>>> G=nx.Graph()
```

- add one node at a time

```
>>>G.add_node(1)
```

- add a list of nodes

```
>>>G.add_nodes_from([2,3])
```

- add one edge at a time

```
>>>G.add_edge(1,2)
```

```
>>>e=(2,3)
```

```
>>>G.add_edge(*e) #unpack edge tuple*
```

- add a list of edges

```
>>>G.add_edges_from([(1,2),(1,3)])
```



# Final Steps

- **Input to the PageRank algorithm a file containing every document ID and associated with each ID, the IDs that are pointed to by links within the document ID**
- **Output from the PageRank algorithm is a file containing every document ID and its associated PageRank**
- **place this file in solr-8.11.1/server/solr/core\_name, call the file external\_pageRankFile.txt**
- **add the PageRank field to the schema.xml file**

```
<fieldType name="external" keyField="id" defVal="0" class="solr.ExternalFileField" />
<field name="pageRankFile" type="external" stored="false" indexed="false" />
```

# The Queries

- Once both ranking algorithms are working you should input the queries below and compare the results

## Nine Queries

Cannes

Congress

Democrats

Patriot Movement

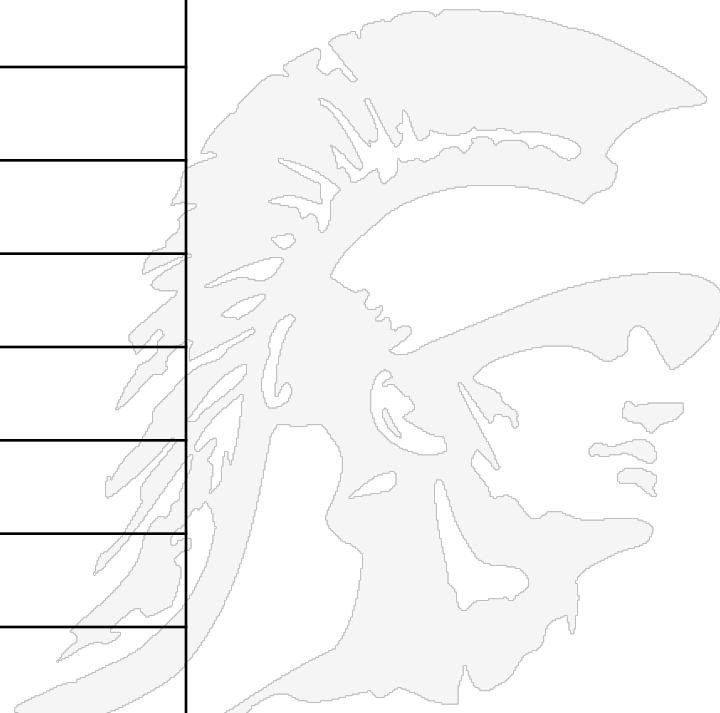
Republicans

Senate

Olympics 2020

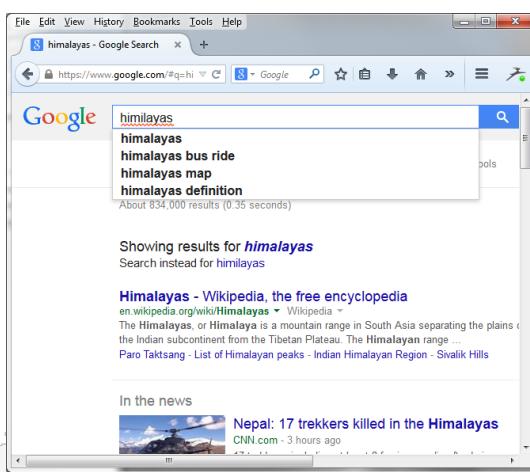
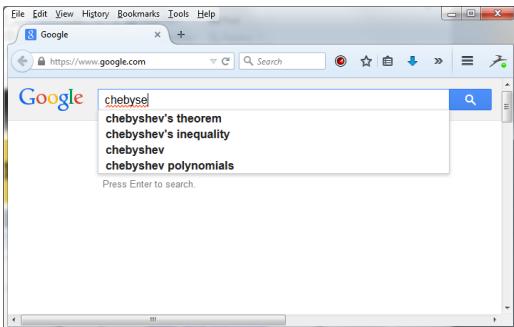
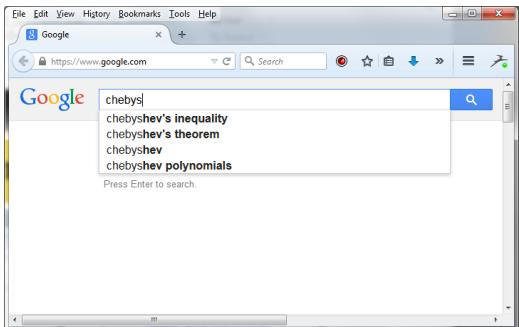
Stock

Virus

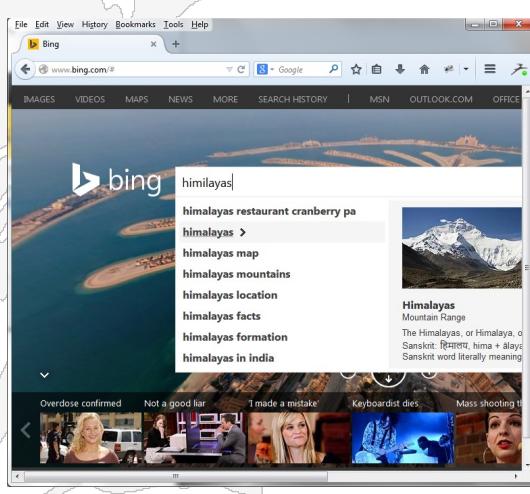
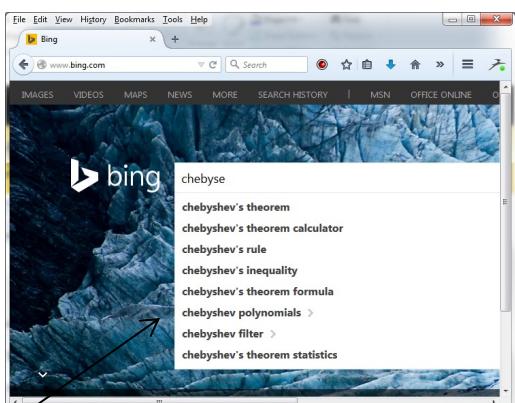
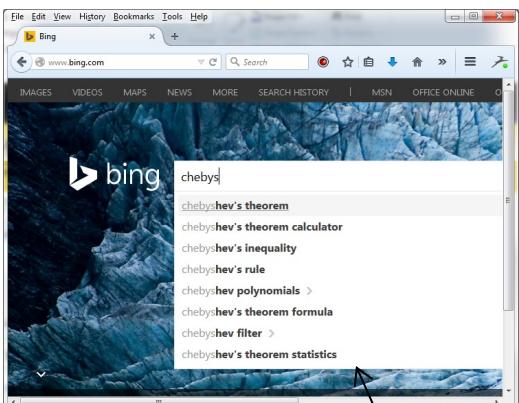


# Spell Checking and Correction

# Some Google/Bing Examples



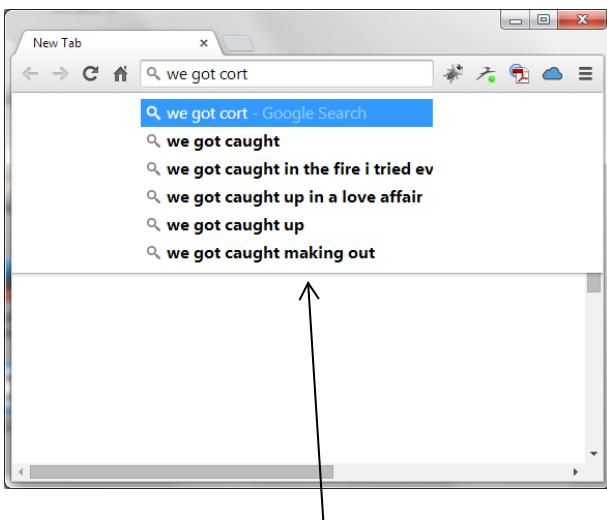
Russian mathematician, notice red underline appears as soon as the first incorrect character is typed



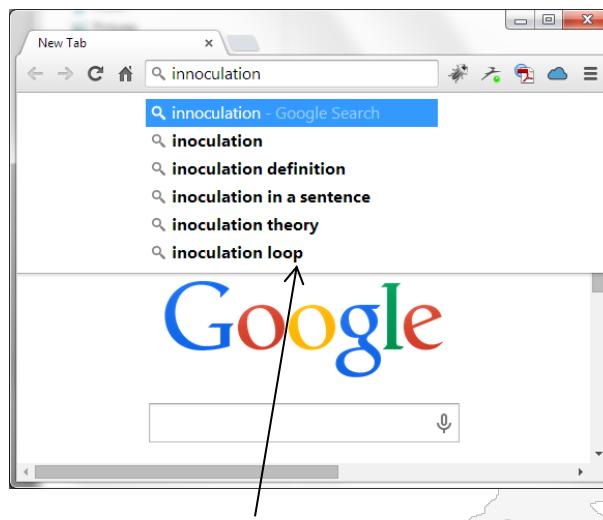
Bing also combines autocomplete with spelling correction but there is no red underline

Himilayas

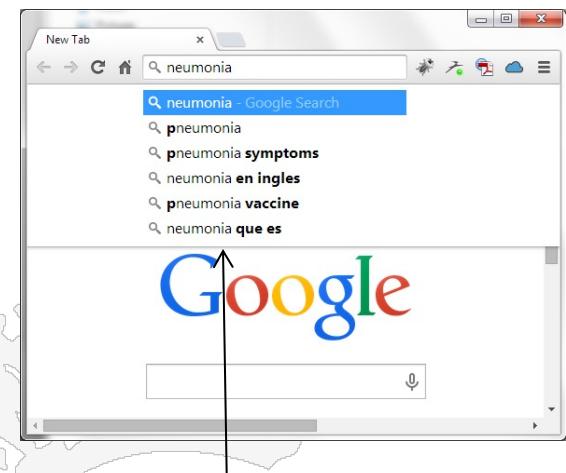
# More Examples



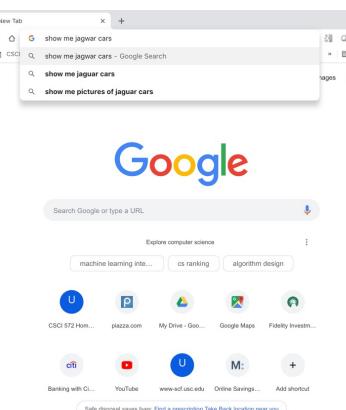
easy for people, but harder for computers  
to correct, **likely use of n-grams**



easy for a computer to correct  
**likely use of a database of words**



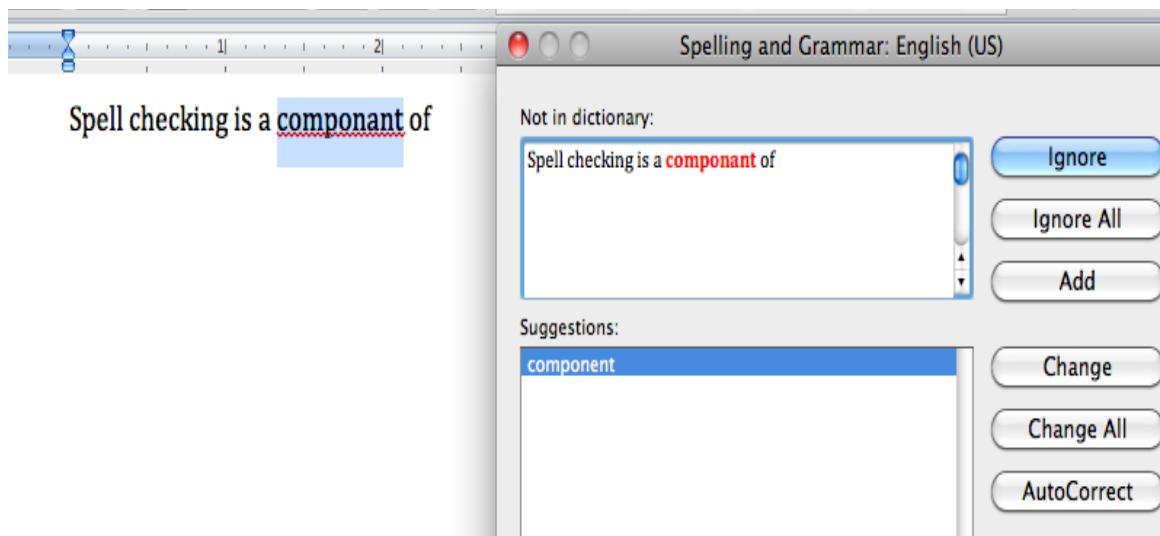
computer needs to both  
identify the error and correct  
the misspelling



Google combines spelling correction  
with the **most likely terms**  
as it comes up with “cars” in  
autocomplete for the query “jagwa”  
(misspelled) but  
leaves the user’s misspelling for a  
while

# Spelling Correction is Done in Many Places

## 1. Word processing



Word processing is the classic application for spelling correction

Word and PowerPoint have mode to auto-correct

- set as the default
- the spell dictionary can be modified

Typing on a virtual keyboard can be doubly difficult (for seniors)

## 2. Smartphone input



# Rates of Spelling Errors

Error rates vary depending upon the application

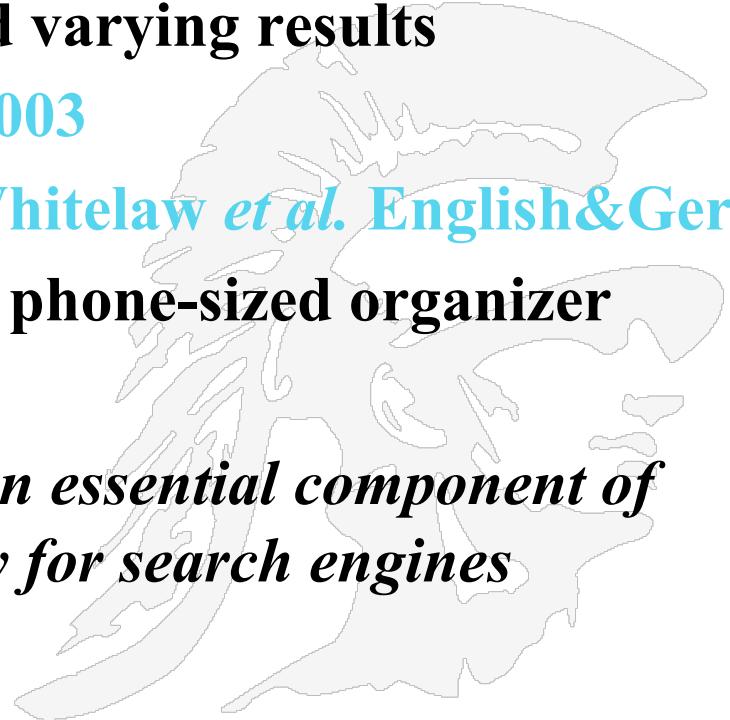
- **Typing is very error prone, and especially difficult on smartphones**
- **Different studies have produced varying results**

**26%:** Web queries [Wang et al. 2003](#)

**13%:** Retyping, no backspace: [Whitelaw et al. English&German](#)

**7%:** Words corrected retyping on phone-sized organizer

*So seamless spelling correction is an essential component of information retrieval and especially for search engines*



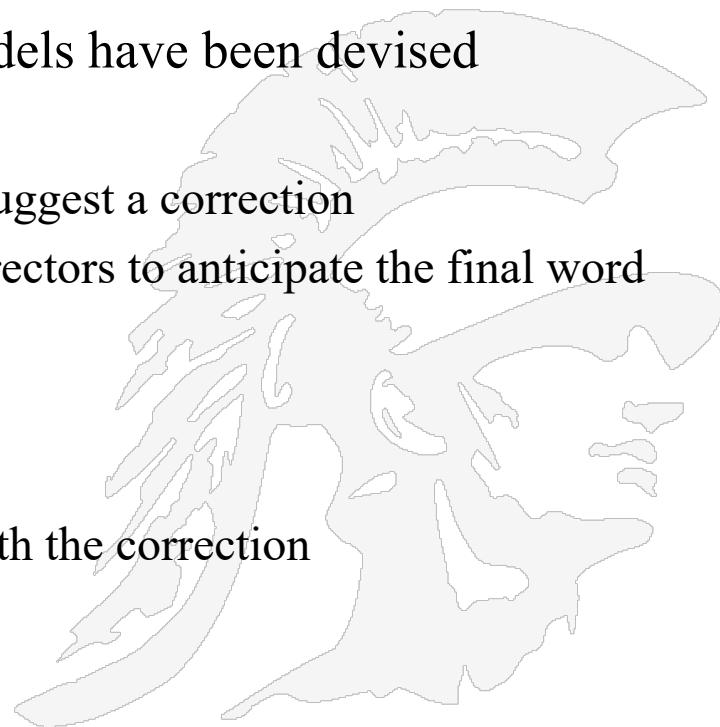
# The Two Main Spelling Tasks

## 1. Spelling Error *Detection*

- Obviously we need a big dictionary and the ability to search it quickly
- Using context may be necessary
  - To do this spelling error models have been devised

## 2. Spelling Error *Correction*

- Web search engines **always** try to suggest a correction
- Autocomplete requires spelling correctors to anticipate the final word
  - Fast response time is required
- The two major techniques are
  1. edit distance algorithms or
  2. n-gram matching to come up with the correction



# Three Types of Spelling Errors

## 1. Non-word errors

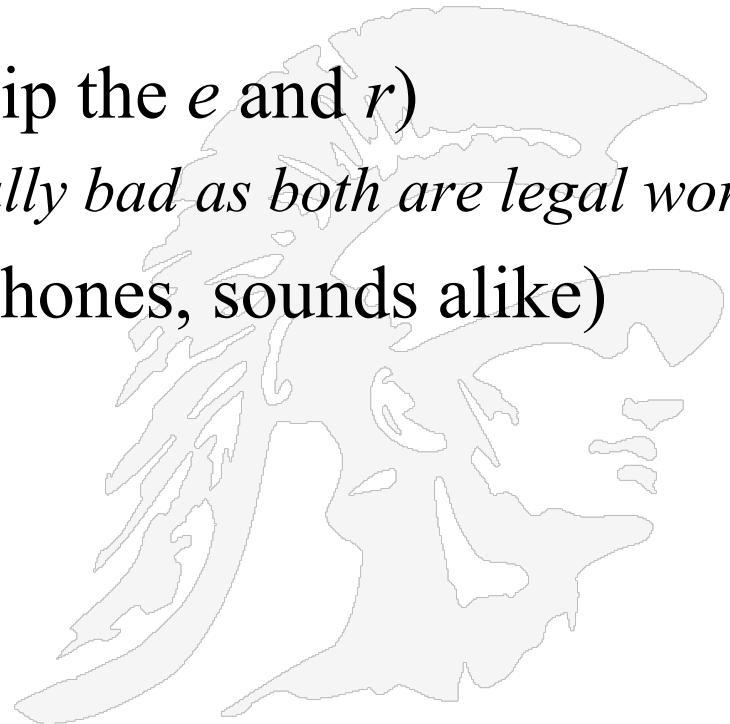
- *graffe* → *giraffe*

## 2. Typographical errors (flip the *e* and *r*)

- *three* → *there*      (*especially bad as both are legal words*)

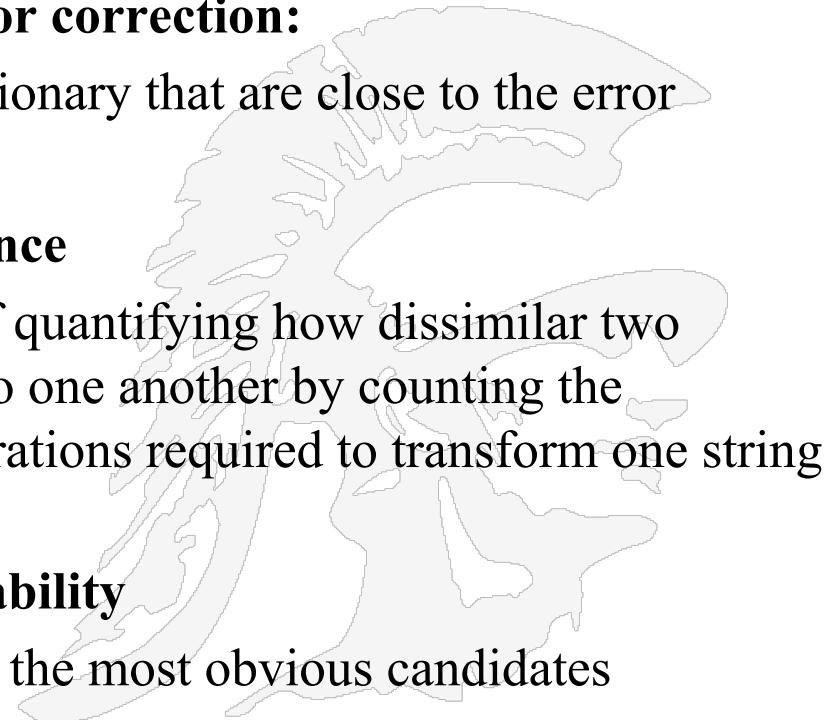
## 3. Cognitive errors (homophones, sounds alike)

- *piece* → *peace*,
- *too* → *two*
- *your* → *you're*



# Non-Word Spelling Errors

- **Non-word spelling error detection:**
  - Any word not in a *dictionary* is presumed to be an error
  - The larger the dictionary the better
- **Approach to non-word spelling error correction:**
  - Generate candidates from the dictionary that are close to the error
  - **How do we do this?**
    - **Shortest weighted edit distance**
      - **Edit distance** is a way of quantifying how dissimilar two strings (e.g., words) are to one another by counting the minimum number of operations required to transform one string into the other
    - **Highest noisy channel probability**
      - use probabilities to select the most obvious candidates



# Causes of Misspellings

Cause	Misspelling	Correction
typing quickly	exxit mispell	exit misspell
keyboard adjacency	importamt	important
inconsistent rules	conceive concierge	conceive concierge
ambiguous word breaking	silver light	silverlight
new words	kinnect	kinect

According to Cucerzan and Brill, **more than 10% of search engine queries are misspelled**  
*"Spelling Correction as an iterative process that exploits the collective knowledge of web users"*  
<http://csci572.com/papers/Cucerzan.pdf> (advocates using query logs to guess the correct spelling)



# How Many Error Variations of a Word May Actually Appear

<http://www.netpaths.net/blog/britney-spears-spelling-variations/>

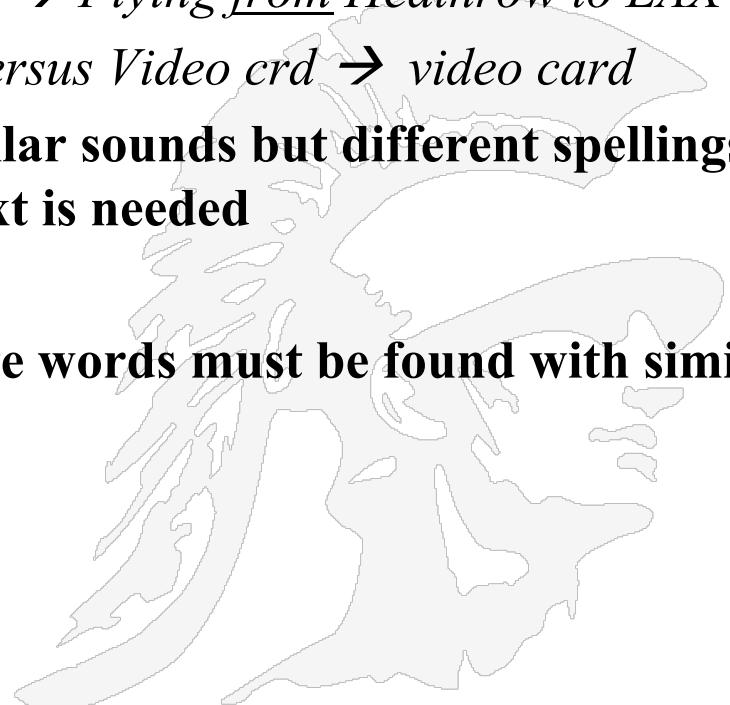
Google's list of spelling errors for "Britney Spears" includes a count of how many different users spelled her name in various ways, e.g.

488,941 people used the correct spelling "britney spears" while 40,134 used "britnay spears"

there are 593 different variations  
that actually occurred

# Spelling Errors Needing Context

- Some misspellings require context to disambiguate
  1. consider whether the surrounding words “make sense” for your candidate set, e.g.
    - *Flying form Heathrow to LAX* → *Flying from Heathrow to LAX*
    - *Power crd* → *power cord* versus *Video crd* → *video card*
  2. For candidate words with similar sounds but different spellings and different meanings, context is needed
    - e.g. *there, their*
  3. To resolve the above, candidate words must be found with similar *pronunciations*
    - use the Soundex algorithm



# More Challenges for Identifying Spelling Errors

- Some additional challenges

- 4. Allow for insertion of a space or hyphen

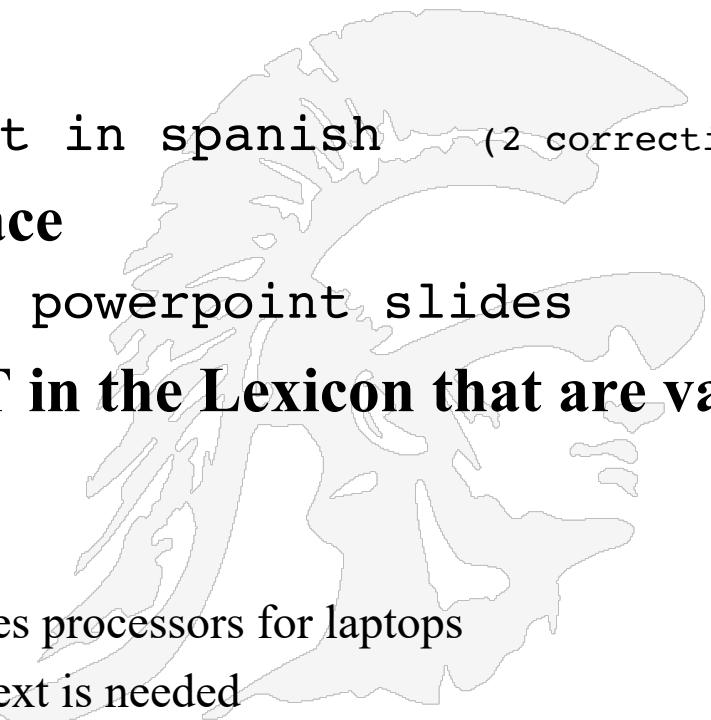
- thisidea → this idea
    - inlaw → in-law
    - chat inspanich → chat in spanish (2 corrections)

- 5. Allow for deletion of a space

- power point slides → powerpoint slides

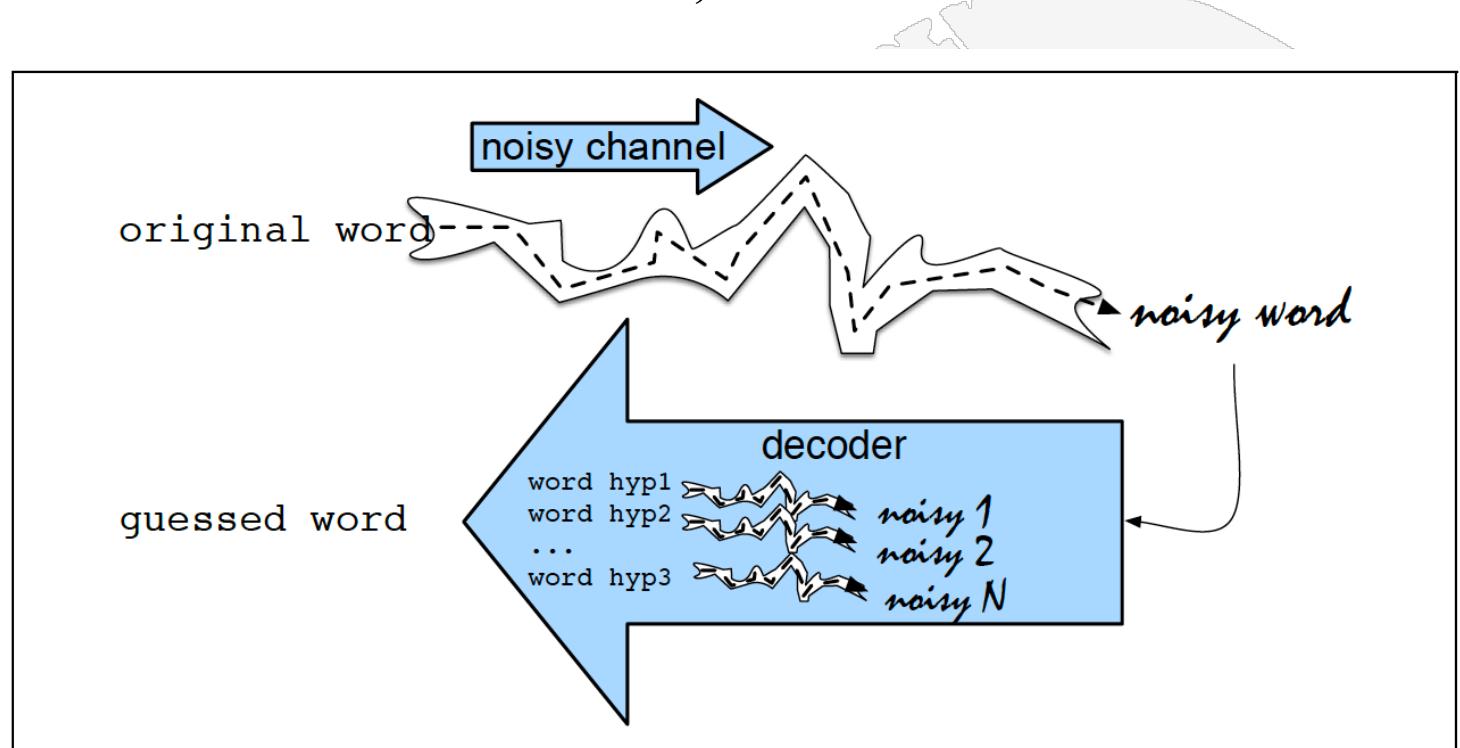
- 6. Watch out for words NOT in the Lexicon that are valid, e.g.

- amd processors
      - AMD is a company that makes processors for laptops
      - Another example where context is needed



# The Noisy Channel Model

- This model suggests treating the misspelled word as if a correctly spelled word has been distorted by being passed through a noisy communication channel
- Noise in this case refers to substitutions, insertions or deletions of letters



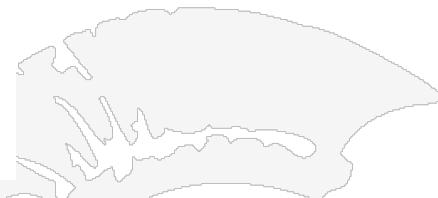
# Bayesian Inference Implies We Can Use Previous Combinations to Predict the Correct Word

- We see an observation  $x$  (a misspelled word) and our job is to find the correct word  $w$  that generated this misspelled word
- Out of all possible words in the vocabulary  $V$  we want to find the word  $w$  such that  $P(w|x)$  is highest. We use the hat notation  $\hat{w}$  to mean “our estimate of the correct word”

$$\hat{w} = \operatorname{argmax}_{w \in V} P(w|x)$$

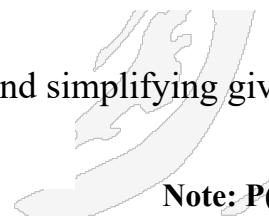
- Out of all words in the vocabulary, we want the particular word that maximizes the right-hand side  $P(w|x)$
- *Bayes Rule:*

$$P(a|b) = \frac{P(b|a)P(a)}{P(b)}$$



The probability of  $a$  given  $b$  is equal to the probability of  $b$  given  $a$ , times the probability of  $a$ , divided by the probability of  $b$ .

implies  $\hat{w} = \operatorname{argmax}_{w \in V} \frac{P(x|w)P(w)}{P(x)}$  and simplifying gives  $\hat{w} = \operatorname{argmax}_{w \in V} P(x|w)P(w)$

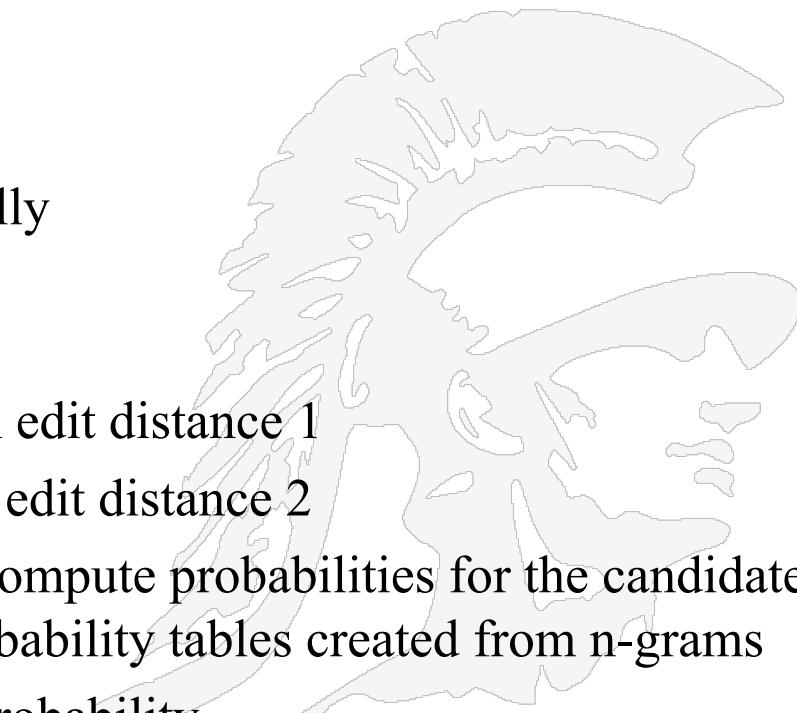


Note:  $P(x|w)$  is a constant so we can ignore it

**Conclusion:** we need a table of probabilities

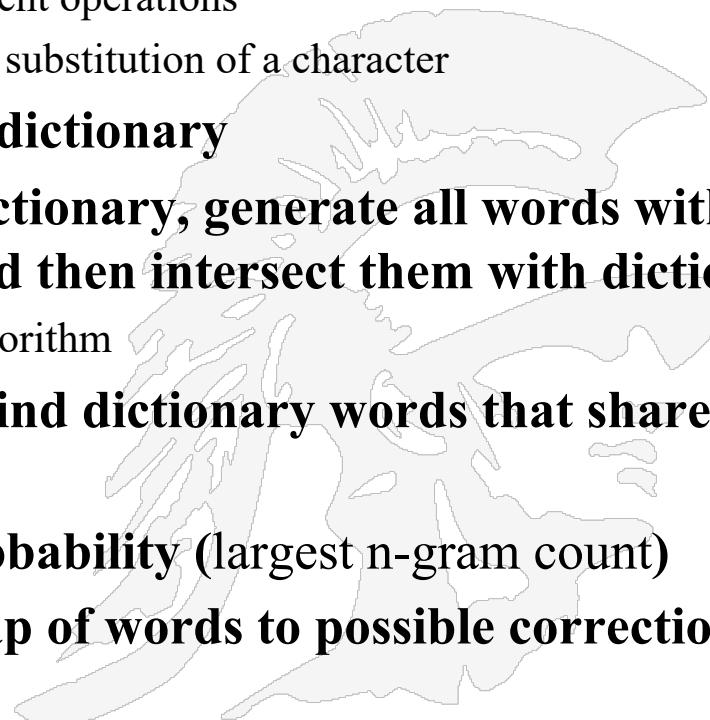
# The Basic Spelling Correction Algorithm

1. **Initial step:** Create a dictionary and encode it for fast retrieval
2. When a query is submitted, the spell checker examines each word and for words not in the dictionary looks for possible character edits, namely
  - insertions,
  - deletions,
  - substitutions, and occasionally
  - transpositions
  - **Observation:**
    - 80% of errors are within edit distance 1
    - Almost all errors within edit distance 2
3. Take the output of step 2 and compute probabilities for the candidates using previously identified probability tables created from n-grams
4. Select the result with highest probability



# The Basic Spelling Correction Algorithm Refined

- **Edit distance** is a way of quantifying how dissimilar two strings (e.g. words) are to one another by counting the minimum number of operations required to transform one string into the other
    - different algorithms assume slightly different operations
    - e.g., Levenshtein uses: removal, insertion, substitution of a character
1. **Check each query term against the dictionary**
  2. **For each term NOT found in the dictionary, generate all words within edit distance  $\leq k$  (e.g.,  $k = 1$  or  $2$ ) and then intersect them with dictionary**
    - Compute them fast with a Levenshtein algorithm
  3. **Use a character  $n$ -gram index and find dictionary words that share “most”  $n$ -grams with word**
  4. **Select the word with the highest probability (largest n-gram count)**
  5. **For speed, have a pre-computed map of words to possible corrections**



# Use Edit Distance To Produce Candidate Corrections

Input	Candidate Correction	Correct Letter	Error Letter	correction Type
acress	actress	t	–	insertion
acress	cress	–	a	deletion
acress	caress	ca	ac	transposition
acress	access	c	r	substitution
acress	across	o	e	substitution
acress	acres	–	s	deletion

Six words within 1 of acress

Context check is necessary to choose the appropriate word

(try this yourself in Google/Bing)

For the word "acress" there are six dictionary words all within edit distance 1

# Now Apply Probabilities

- We now need to compute the prior probability of each occurrence
- We can do this using unigrams, bigrams, trigrams, etc
- Using the Corpus of Contemporary English, 404,253,213 words we get the following
- *Across* is the most likely choice, followed by

word	Frequency of word	$P(w)$
actress	9,321	.0000230573
cress	220	.0000005442
caress	686	.0000016969
access	37,038	.0000916207
→ across	120,844	.0002989314
acres	12,874	.0000318463

For fun try:  
I need a cress to . . .  
I love the a cress . . .  
a kiss and a a cress . . .

"across" is the  
most likely  
correction →

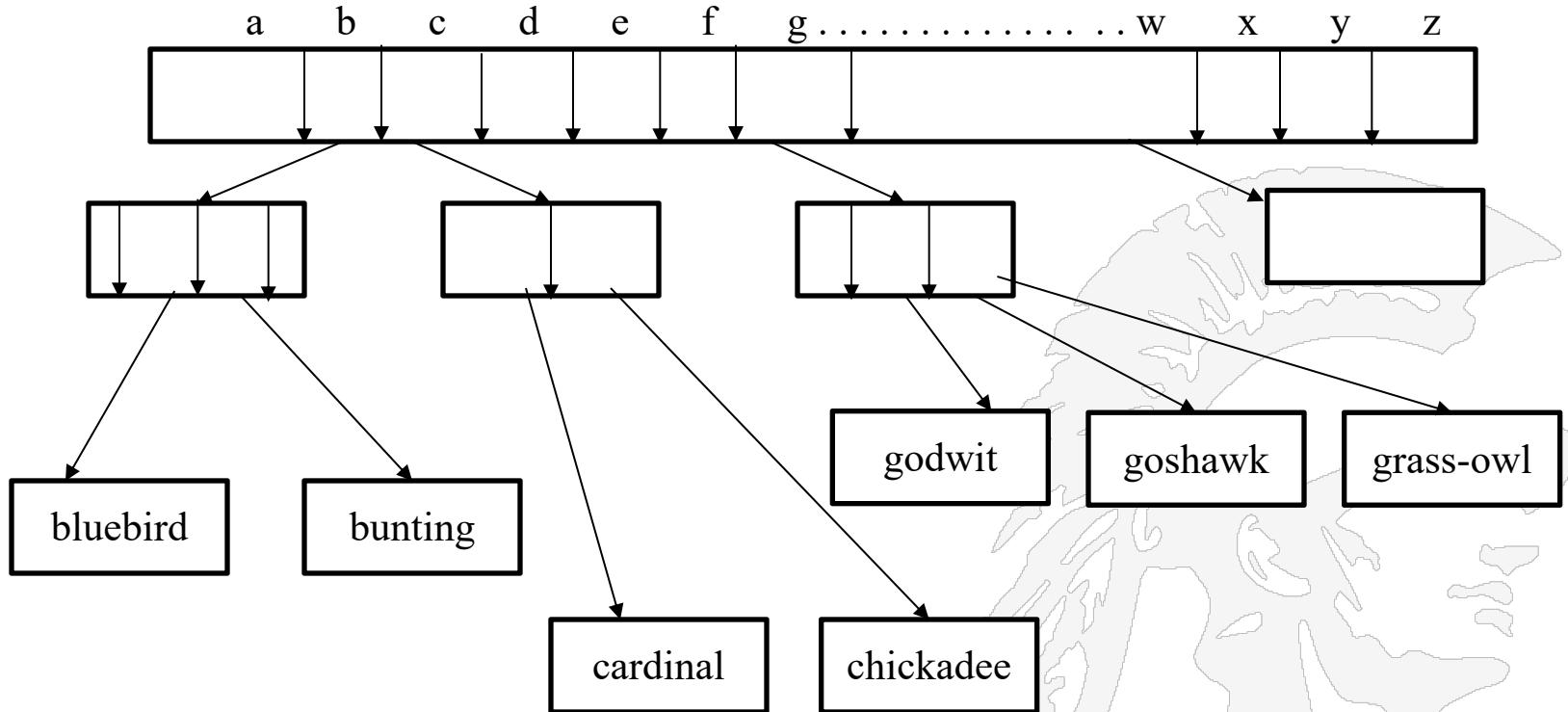
# The Spelling Correction Dictionary and Autocomplete

- The search for corrections is carried out from left-to-right
  - At each point, a partial hypothesis is expanded with every character which could follow the partial hypothesis and lead to one of the known words (the user input is always allowed as an output hypothesis).
  - Thus the branching factor controls the amount of time required to search for spelling corrections.
- The terms of the lexicon must be stored in a data structure that affords efficient *prefix matching*
  - Often a trie data structure is used

# The Spelling Correction Dictionary

## Example of a Trie

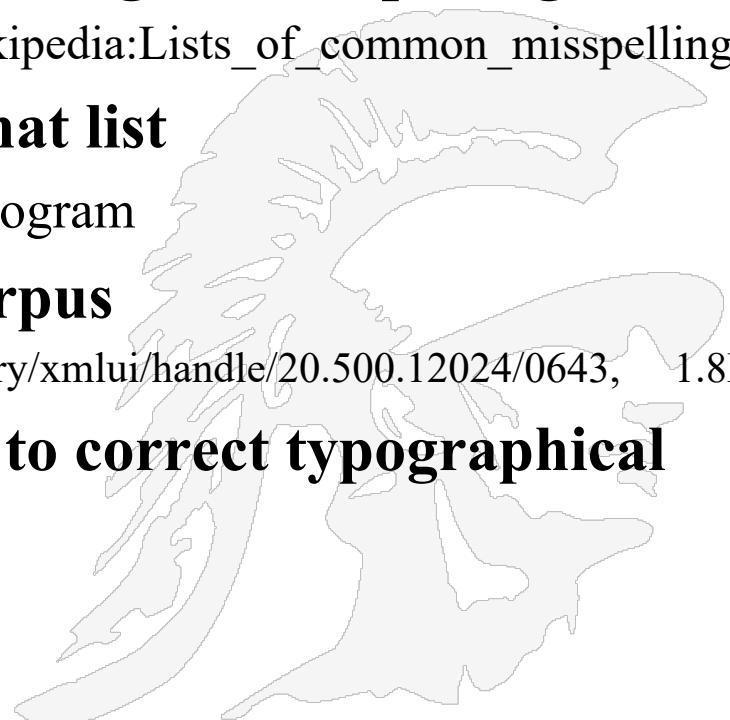
- a prefix tree (sometimes called a trie from the word retrieval) is a tree of degree  $\geq 2$  in which the branching at any level is determined by a portion of the key



branch nodes take you down the tree to element nodes; At any stage one is pointing at all keyword matches that contain the same prefix; Computing time for retrieval is  $O(m)$  where  $m$  is the length of the string, at the expense of increased storage

# The Spelling Correction Dictionary Error Test Sets

- To enhance a lexicon one can include a table of common misspellings
- there are many possible spelling error test sets, e.g.
  - Wikipedia's list of common English misspelling
    - [https://en.wikipedia.org/wiki/Wikipedia:Lists\\_of\\_common\\_misspellings](https://en.wikipedia.org/wiki/Wikipedia:Lists_of_common_misspellings)
  - Aspell filtered version of that list
    - <http://aspell.net/> is the spell program
  - Birkbeck spelling error corpus
    - <https://ota.bodleian.ox.ac.uk/repository/xmlui/handle/20.500.12024/0643>, 1.8MBs
- These sets are primarily used to correct typographical errors



# Using N-Grams For Spelling Correction

- An ***n*-gram model** is a type of probabilistic language model for predicting the next item in a sequence
- Two benefits of *n*-gram models (and algorithms that use them) are simplicity and scalability – with larger *n*, a model can store more context with a well-understood space–time tradeoff, enabling small experiments to scale up efficiently
- **Sample of 3-gram sequences**
- ceramics collectables fine (130)
- ceramics collected by (52)
- ceramics collectible pottery (50)
- ceramics collectibles cooking (45)

**Sample of 4-gram sequences and # of times appeared**

serve as the incoming	(92)
serve as the incubator	(99)
serve as the independent	(794) ←
serve as the index	(223)
serve as the indication	(72)
serve as the indicator	(120) ↑

a query such as "serve as the indapendant" would match the above

# Google's N-Gram Data

- Google has collected and uses a great deal of N-gram data
- Google is using the Linguistics Data Consortium to distribute more than one trillion words they have extracted from public web pages
- Below is a statistical summary of the data they are distributing

File sizes: approx. 24 GB compressed (gzip'ed) text files

Number of tokens: 1,024,908,267,229

Number of sentences: 95,119,665,584

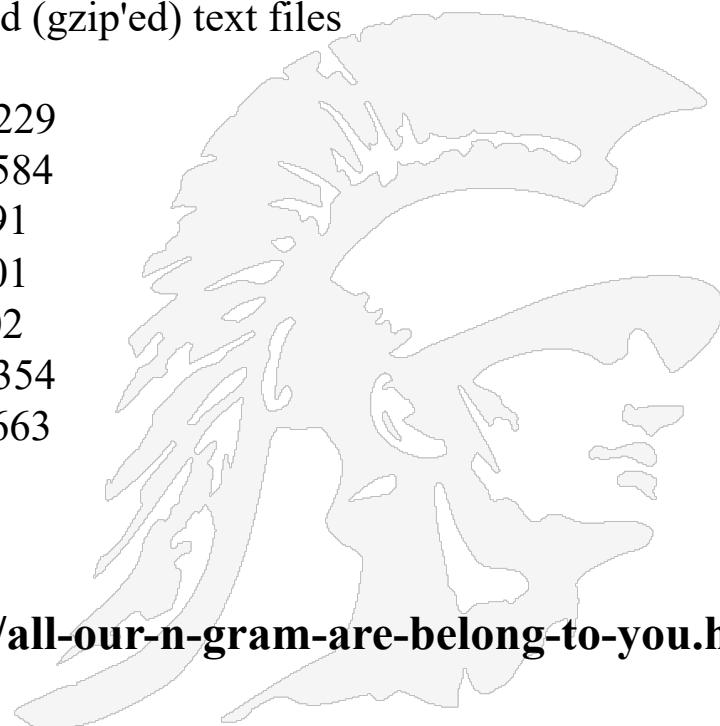
Number of unigrams: 13,588,391

Number of bigrams: 314,843,401

Number of trigrams: 977,069,902

Number of fourgrams: 1,313,818,354

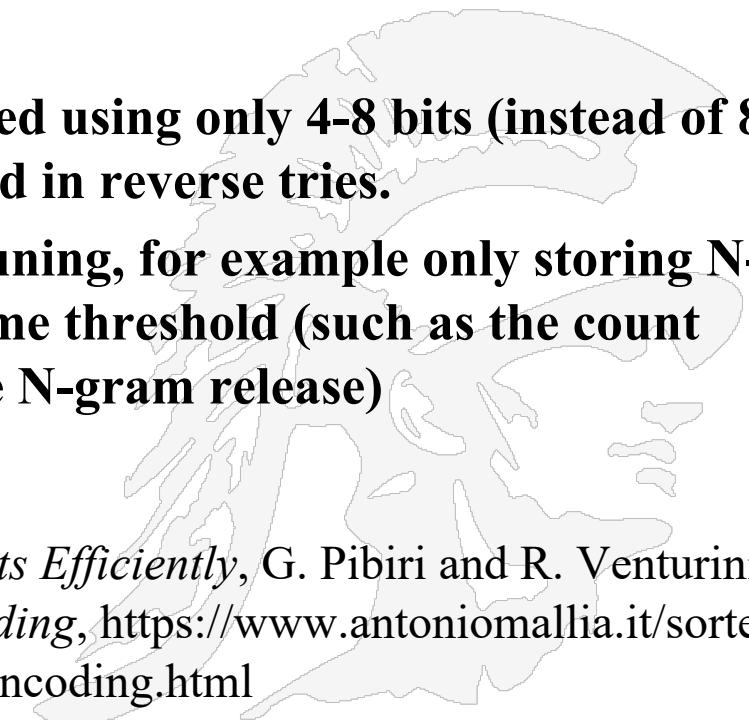
Number of fivegrams: 1,176,470,663



<http://googleresearch.blogspot.com/2006/08/all-our-n-gram-are-belong-to-you.html>

# Handling N-Gram Data

- Efficiency considerations are important when building language models that use such large sets of N-grams.
- Rather than store each word as a string, it is generally represented in memory as a 64-bit hash number, with the words themselves stored on disk.
- Probabilities are generally quantized using only 4-8 bits (instead of 8-byte floats), and N-grams are stored in reverse tries.
- N-grams can also be shrunk by pruning, for example only storing N-grams with counts greater than some threshold (such as the count threshold of 40 used for the Google N-gram release)
- See *Handling Massive N-Gram Datasets Efficiently*, G. Pibiri and R. Venturini which improves upon *Elias-Fano encoding*, <https://www.antoniomallia.it/sorted-integers-compression-with-elias-fano-encoding.html>



# Applying the N-Gram Model to Spelling Correction

- Let's say you are using 4-grams to calculate the probability of a next word in text.
  - You have "this is a very" followed by "sunny".
  - But assume "sunny" does not occur in the n-gram set that includes "this is a very"
  - *Implication:* so for the 4-gram model "sunny" has probability 0,
- Back-off means you go back to a  $n-1$ -gram level to calculate the probabilities when you encounter a word with probability = 0.
  - So in the above you will use a 3-gram model to calculate the probability of "sunny" in the context "is a very".
- Whenever you go back one level you multiply the odds by an empirically derived number, in this case 0.4. So if sunny exists in the 3-gram model the probability would be
$$0.4 * P("sunny" | "is a very")$$

This leads to an algorithm for handling n-grams to do spelling correction, called  
*The Stupid Back-off Algorithm*

1. If a higher-order N-gram has a zero count, we simply backoff to a lower order N-gram, weighed by a fixed (context-independent) weight
  2. The backoff terminates in the unigram
- See Wikipedia's description of the Back-off Model

# A Complete Spelling Correction Program



# Peter Norvig's Spelling Corrector written in Python

```

import re, collections
def words(text): return re.findall('[a-z]+', text.lower())
def train(features):
 model = collections.defaultdict(lambda: 1)
 for f in features:
 model[f] += 1
 return model
NWORDS = train(words(file('big.txt').read()))
alphabet = 'abcdefghijklmnopqrstuvwxyz'
def edits1(word):
 splits = [(word[:i], word[i:]) for i in range(len(word) + 1)]
 deletes = [a + b[1:] for a, b in splits if b]
 transposes = [a + b[1] + b[0] + b[2:] for a, b in splits if len(b)>1]
 replaces = [a + c + b[1:] for a, b in splits for c in alphabet if b]
 inserts = [a + c + b for a, b in splits for c in alphabet]
 return set(deletes + transposes + replaces + inserts)

def known_edits2(word):
 return set(e2 for e1 in edits1(word) for e2 in edits1(e1) if e2 in NWORDS)

def known(words): return set(w for w in words if w in NWORDS)

def correct(word):
 candidates = known([word]) or known(edits1(word)) or known_edits2(word) or [word]
 return max(candidates, key=NWORDS.get)

```

Correct takes a word  
and returns a likely  
correct form of the word:

```

correct('speling')
Spelling
correct('korrechter')
corrector

```

<http://norvig.com/spell-correct.html>

# How the Program Works

- The algorithm and its description can be found at
- <http://norvig.com/spell-correct.html>
- The file `big.txt` contains a million words
  - It includes text of books from Project Gutenberg, Wiktionary, British National Corpus
- Extract the individual words (function `words` converts everything to lower case)
- *The* and *the* are the same, but “don’t” is seen as “don” and “t”
- The program counts how many times each word occurs using function `train`
- `NWORDS[w]` holds a count of how many times the word `w` has been seen
- For words that are not in our set we set their occurrence to a default, non-zero value of 1 using the Python hash table statement `collections.defaultdict`
- Edit distance is the number of changes it would take to turn one word into another
- An edit can be one of: {deletion, transposition, alteration, insertion}
- The 7 lines that end in `return set(deletes + transposes + replaces + inserts)` is a function that returns a set of all words `c` that are one edit away from `w`
  - This can be a large set
- The literature on spelling correction claims that 80-95% of spelling errors are an edit distance of 1 from the target

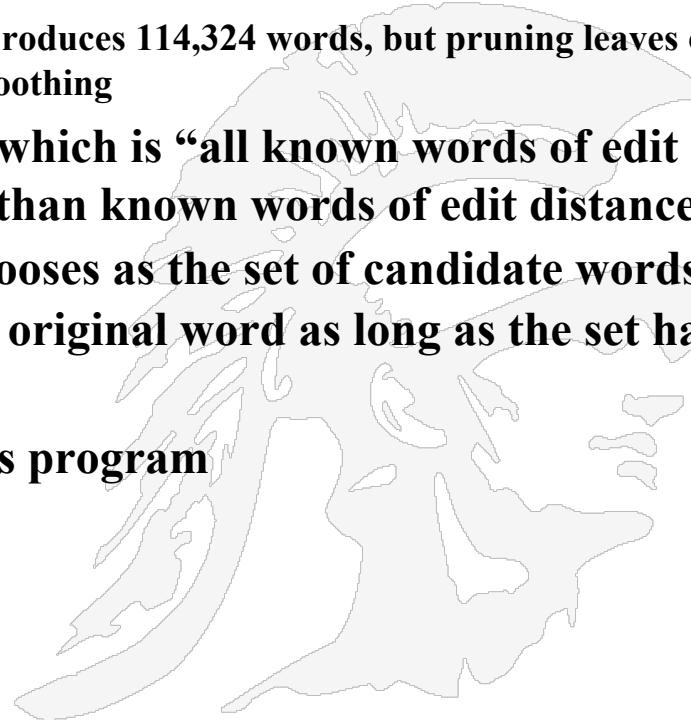
## How the Program Works (cont'd)

- To compute edit distance 2, just apply edits1 to all the results of edits1

```
Def edit2(word) :
```

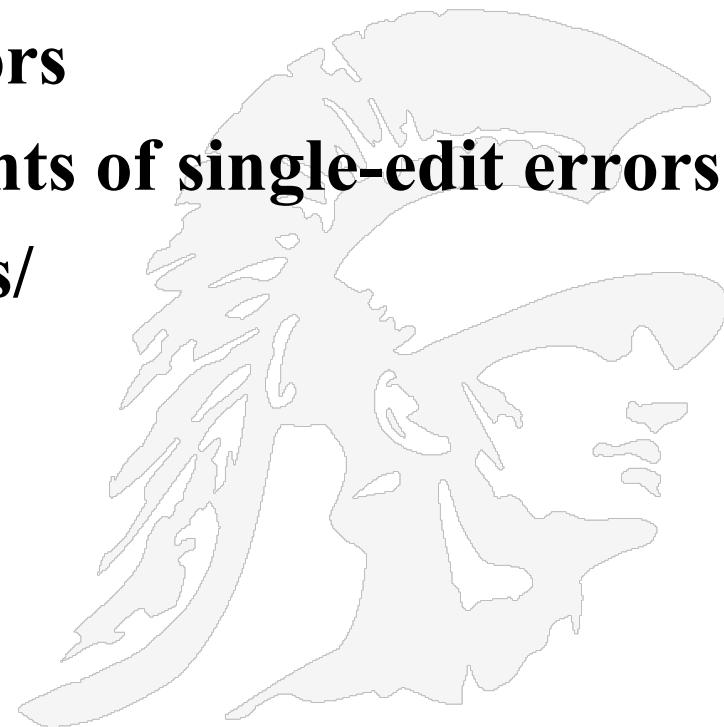
```
 return set(e2 for e1 in edits1(word) for e2 in edits1(e1))
```

- The function known\_edits2 removes words that are not in our original set and greatly reduces the size of edits2
  - E.g. all edit 2 combinations for “something” produces 114,324 words, but pruning leaves only 4 words: smoothing, seething, something, and soothing
- Now he assumes his probability model, which is “all known words of edit distance 1 are infinitely more probable than known words of edit distance 2
- He defines function correct which chooses as the set of candidate words the set with the shortest edit distance to the original word as long as the set has some known words
- See his web page for an evaluation of his program



# Natural Language Corpus Data

- Peter Norvig has provided a web page that is full of useful data for a spelling corrector
  - <http://norvig.com/spell-correct.html>
- Peter Norvig's list of errors
- Peter Norvig's list of counts of single-edit errors
- <http://norvig.com/ngrams/>



# Some References

- *How Difficult is it to Develop a Perfect Spell-checker? A Cross-linguistic Analysis through Complex Network Approach*, Monojit Choudhury<sup>1</sup>, Markose Thomas<sup>2</sup>, Animesh Mukherjee<sup>1</sup>, Anupam Basu<sup>1</sup>, and Niloy Ganguly<sup>1</sup>

<http://citeseerx.ist.psu.edu/viewdoc/download;jsessionid=52A3B869596656C9DA285DCE83A0339F?doi=10.1.1.146.4390&rep=rep1&type=pdf>

- *Using the web for language independent spellchecking and autocorrection* by C. Whitelaw et al Proc. 2009 Conf. on Empirical Methods in Natural Language Processing, pp890-899

[http://static.googleusercontent.com/external\\_content/untrusted\\_dlcp/research.google.com/en/us/pubs/archive/36180.pdf](http://static.googleusercontent.com/external_content/untrusted_dlcp/research.google.com/en/us/pubs/archive/36180.pdf)

**Spell Checking by Computer, by Roger Mitton,**

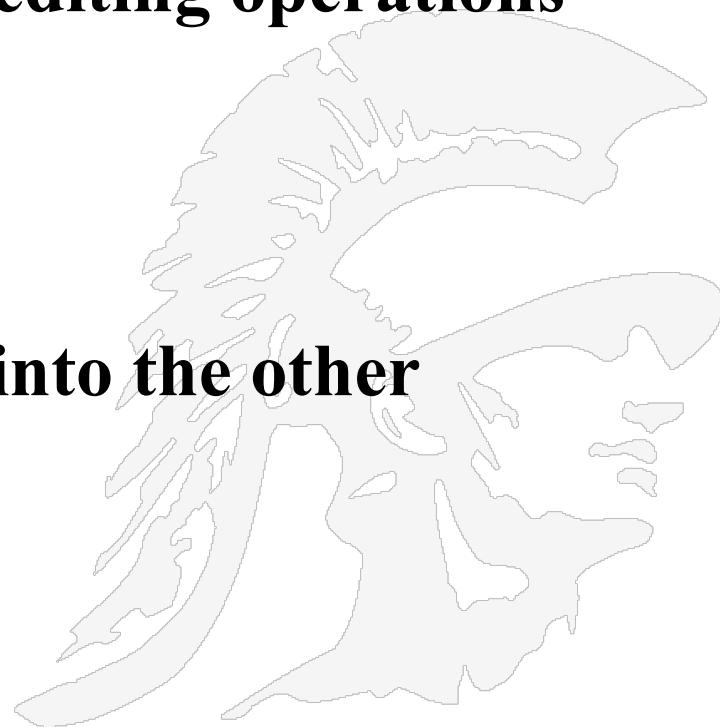
<http://www.dcs.bbk.ac.uk/~roger/spellchecking.html>

# Edit Distance & Levenshtein Algorithm



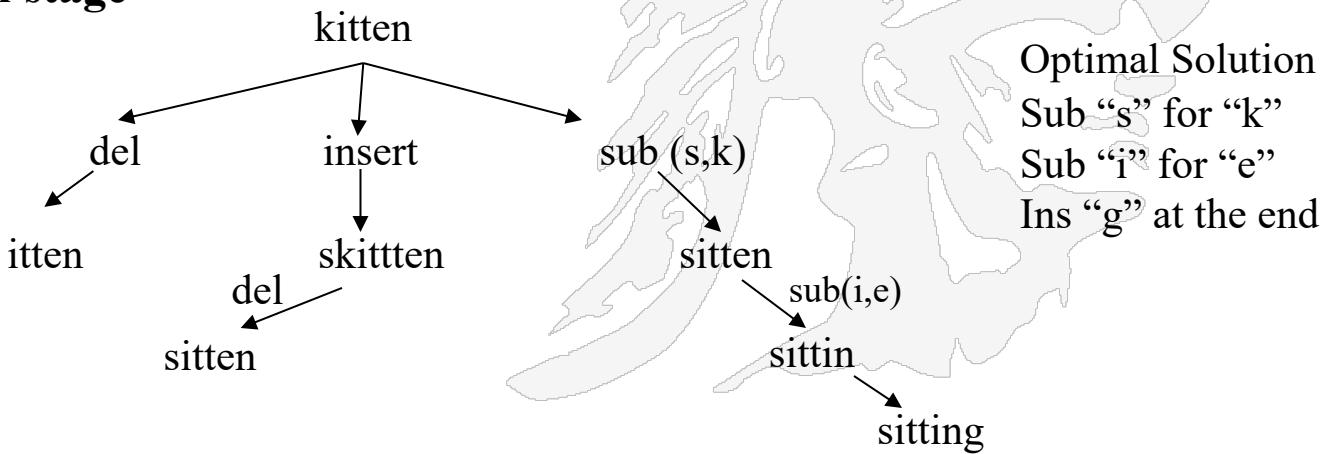
# Edit Distance

- the minimum edit distance between two strings is the minimum number of editing operations
    - insertion
    - deletion
    - substitution
- needed to transform one into the other**



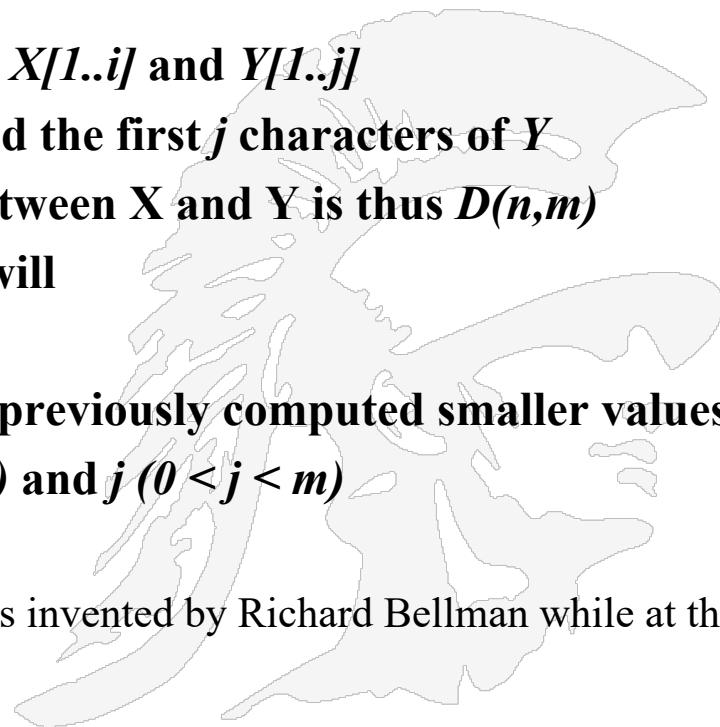
# How to Find the Minimum Edit Distance

- Searching for a path (sequence of edits) from the start string to the final string
  - initial state: word we're transforming (e.g. kitten)
  - operators: insert, delete, substitute
  - goal state: the word we're trying to get to (e.g. sitting)
  - path cost: what we want to minimize, the number of edits
- If we blindly generate all possible paths in an effort to produce the goal state, our algorithm will take exponentially long
- But we realize that we needn't do that, we may only follow the path that is optimal at each stage



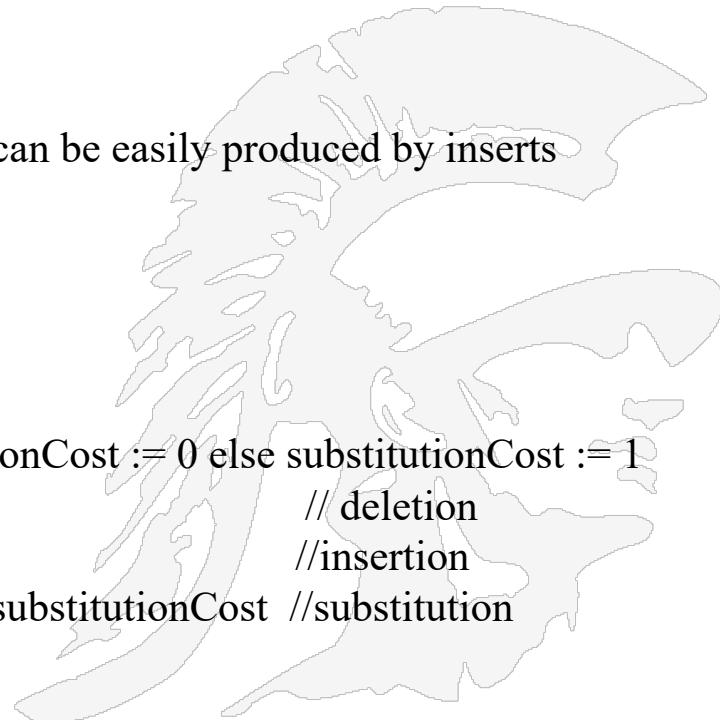
# Dynamic Programming for Minimum Edit Distance

- Dynamic programming: a technique for solving optimization problems by combining solutions to sub-optimal problems bottom-up
- Given two strings:  $X$  of length  $n$  and  $Y$  of length  $m$ , define
  - $d(i, j)$  as
    - the minimum edit distance between  $X[1..i]$  and  $Y[1..j]$ 
      - i.e. the first  $i$  characters of  $X$  and the first  $j$  characters of  $Y$
    - Then the minimum edit distance between  $X$  and  $Y$  is thus  $D(n,m)$
  - The dynamic programming algorithm will
    - compute  $D(i,j)$  for small  $i, j$
    - and compute larger  $D(i,j)$  based on previously computed smaller values
    - i.e. compute  $D(i,j)$  for all  $i$  ( $0 < i < n$ ) and  $j$  ( $0 < j < m$ )
  - Note: the dynamic programming technique was invented by Richard Bellman while at the Rand Corp, later a professor at USC



# Pseudocode Implementation of Levenshtein Distance

```
function LevenshteinDistance(char s[1..m], char t[1..n]):
 //for all i and j, d[i,j] will hold the Levenshtein distance between
 //the first i characters of s and the first j characters of t
 declare int d[0..m, 0..n]
 //Set each element in d to zero
 for i from 1 to m, j from 1 to n: d[i,j] := 0
 Starting with empty character source and target can be easily produced by inserts
 for i from 1 to m: d[i,0] := i
 for j from 1 to n: d[0,j] := j
 //main loop
 for j from 1 to n:
 for i from 1 to m:
 if s[i] = t[j] then substitutionCost := 0 else substitutionCost := 1
 d[i,j] := min (d[i-1,j] + 1,
 d[i,j-1] + 1,
 d[i-1, j-1] + substitutionCost) //deletion
 //insertion
 //substitution
)
return d[m,n]
```



# Levenshtein Example

- Consider the strings: sitting, kitten
- Here is the initial matrix

	#	K	I	T	T	E	N
#	0	1	2	3	4	5	6
S	1						
I	2						
T	3						
T	4						
I	5						
N	6						
G	7						

# is the null string  
Looking at the first column, to go from # to S requires 1 insert;  
to go from # to SI requires 2 inserts; to go from # to SIT requires 3 inserts;  
Similarly for the first row, to go from # to K requires 1 insert, etc.

# Levenshtein Allows For Insertion, Deletion and Substitution

- To get the value in the 1,1 position use the formula:

$$d(i, j) = \min \{ d(i-1, j) + 1, d(i, j-1) + 1, d(i-1, j-1) + 1 \text{ if chars not equal, 0 otherwise} \}$$

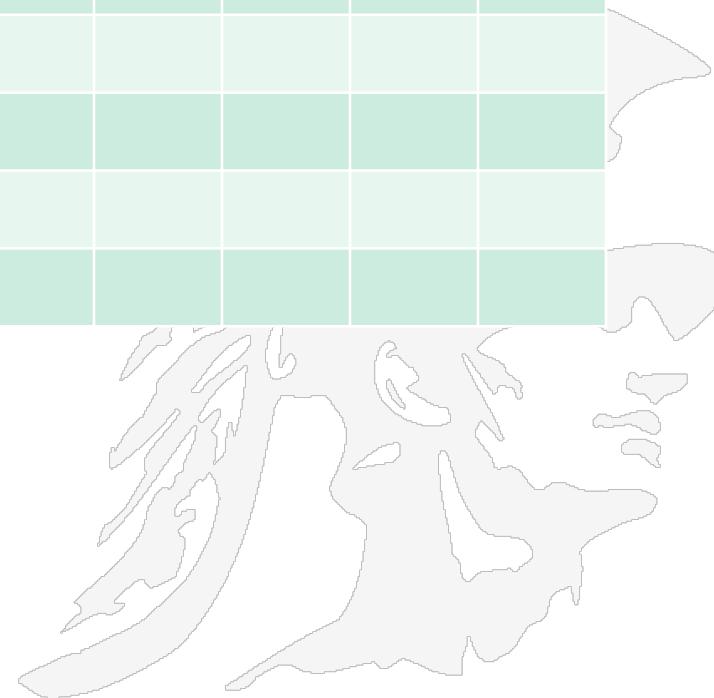
We can fill in the second row and the second column observing that at each step there is a single deletion followed by insertions

	#	K	I	T	T	E	N
#	0	1	2	3	4	5	6
S	1	1	2	3	4	5	6
I	2	2	1	2	3	4	5
T	3	3	2	1	2	3	4
T	4	4	3	2	1	2	3
I	5	5	4	3	2	2	3
N	6	6	5	4	3	3	2
G	7	7	6	5	4	4	3



# Levenshtein Example 2

	#	S	A	T	U	R	D	A	Y
#	0	1	2	3	4	5	6	7	8
S	1	0	1	2	3	4	5	6	7
U	2	1	1	2	2				
N	3								
D	4								
A	5								
Y	6								



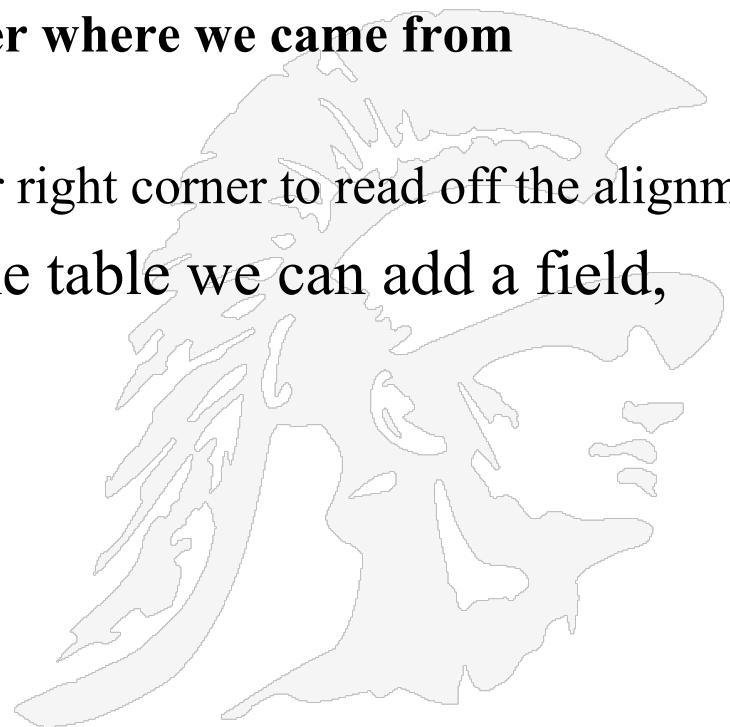
# Performance

- Time:  $O(nm)$
- Space:  $O(nm)$
- Backtrace:  $O(n + m)$



# Computing Alignments

- **Edit distance isn't sufficient**
  - we often need to align each character of the two strings to each other
- **we do this by keeping a "backtrace"**
- **every time we enter a cell, remember where we came from**
- **so when we reach the end,**
  - trace back the path from the upper right corner to read off the alignment
- For example for each field of the table we can add a field,
- $\text{ptr}(i, j)$  whose value is either
  - LEFT, for insertion
  - DOWN, for deletion
  - DIAG, for substitution



# Weighted Edit Distance

- why would we add weights to the computation?
  - spell correction: some letters are more likely to be mistyped than others
- a **confusion matrix** is a specific table layout that allows visualization of the performance of an algorithm; each column of the matrix represents the instances in a predicted class while each row represents the instances in an actual class (or vice-versa)

10/102 3 - 4 - Weighted Minimum Edit Distance - Stanford NLP - Professor Dan Jurafsky & Chris Manning

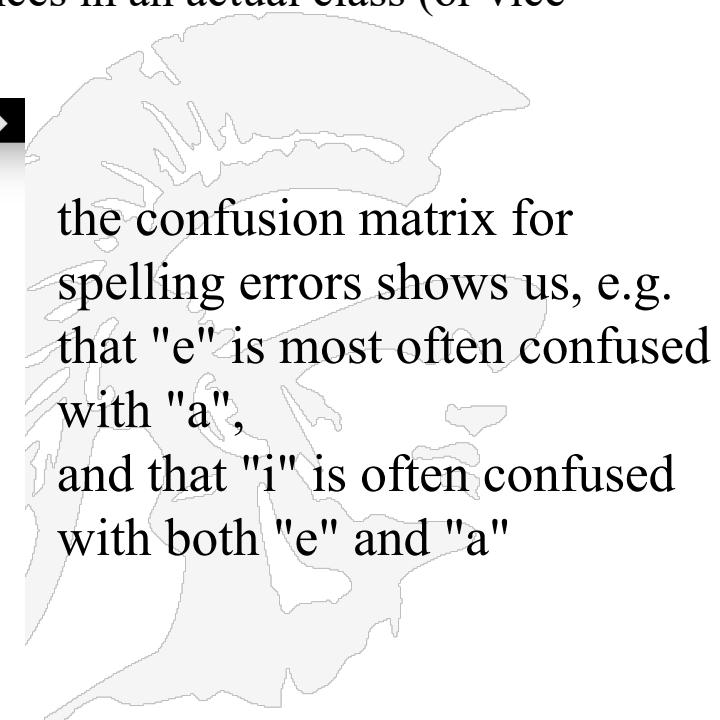


Dan Jurafsky



## Confusion matrix for spelling errors

X	sub[X, Y] = Substitution of X (incorrect) for Y (correct)																									
	Y (correct)																									
	a	b	c	d	e	f	g	h	i	j	k	l	m	n	o	p	q	r	s	t	u	v	w	x	y	z
a	0	0	7	1	342	0	0	2	118	0	1	0	0	3	76	0	0	1	35	9	9	0	1	0	5	0
b	0	0	9	9	2	2	3	1	0	0	0	5	11	5	0	10	0	0	2	1	0	0	8	0	0	0
c	6	5	0	16	0	9	5	0	0	0	1	0	7	9	1	10	2	5	39	40	1	3	7	1	1	0
d	1	10	13	0	12	0	5	5	0	0	2	3	7	3	0	1	0	43	30	22	0	0	4	0	2	0
e	388	0	3	11	0	2	2	0	89	0	0	3	0	5	93	0	0	14	12	6	15	0	1	0	18	0
f	0	15	0	3	1	0	5	2	0	0	0	3	4	1	0	0	0	6	4	12	0	0	2	0	0	0
g	4	1	11	11	9	2	0	0	0	1	1	3	0	0	2	1	3	5	13	21	0	0	1	0	3	0
h	1	8	0	3	0	0	0	0	0	0	2	0	12	14	2	3	0	3	1	11	0	0	2	0	0	0
i	103	0	0	0	146	0	1	0	0	0	6	0	0	49	0	0	0	2	1	47	0	2	1	15	0	0
j	0	1	1	9	0	0	1	0	0	0	0	2	1	0	0	0	0	5	0	0	0	0	0	0	0	0
k	1	2	8	4	1	1	2	5	0	0	0	5	0	2	0	0	0	6	0	0	0	4	0	0	3	0
l	2	10	1	4	0	4	5	6	13	0	1	0	0	14	2	5	0	11	10	2	0	0	0	0	0	0
m	1	3	7	8	0	2	0	6	0	0	4	4	0	180	0	6	0	0	9	15	13	3	2	2	3	0
n	2	7	6	5	3	0	1	19	1	0	4	35	78	0	0	7	0	28	5	7	0	0	1	2	0	2
o	91	1	1	3	116	0	0	25	0	2	0	0	0	14	0	2	4	14	39	0	0	0	18	0	0	0
p	0	11	1	2	0	6	5	0	2	9	0	2	7	6	15	0	0	1	3	6	0	4	1	0	0	0
q	0	0	1	0	0	0	27	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
r	0	14	0	30	12	2	2	8	2	0	5	8	4	20	1	14	0	0	12	22	4	0	0	1	0	0
s	11	8	27	33	35	4	0	1	0	1	0	27	0	6	1	7	0	14	0	15	0	0	5	3	20	1
t	3	4	9	42	7	5	19	5	0	1	0	14	9	5	5	6	0	11	37	0	0	2	19	0	7	6
u	20	0	0	0	44	0	0	0	64	0	0	0	0	2	43	0	0	4	0	0	0	2	0	8	0	0
v	0	0	7	0	0	3	0	0	0	0	1	0	0	1	0	0	0	8	3	0	0	0	0	0	0	0
w	2	2	1	0	1	0	0	2	0	0	1	0	0	0	7	0	6	3	3	1	0	0	0	0	0	0
x	0	0	0	2	0	0	0	0	0	0	0	0	0	0	0	0	0	9	0	0	0	0	0	0	0	0
y	0	0	2	0	15	0	1	7	15	0	0	0	2	0	6	1	0	7	36	8	5	0	0	1	0	0
z	0	0	0	7	0	0	0	0	0	7	5	0	0	0	0	2	21	3	0	0	0	0	3	0	0	



# Snippets Normal, Featured and Rich



# Snippets in Google Search

In 1998, Google introduced *Snippet*, a short description of, or excerpt from, a website which appears in Google search results. Snippets are created automatically based on the site's content and the query terms

key query terms are highlighted:

number

x-

y-intercepts

quadratic

functions

notice that

**have** and **may** are also  
in bold

The screenshot shows a Google search results page. The search bar at the top contains the query: "what is the number of x- and y-intercepts that quadratic functions may have". Below the search bar, there are two tabs: "Web" and "Results 1 - 10 of about".  
  
The first search result is a snippet from a page titled "Pre-Calculus Advanced >> Quadratic Functions >> Intercepts, Zeros ...". The snippet includes the following text:  
A quadratic function will have at most two x-intercepts. ... Notice that this corresponds to the number of solutions a quadratic equation can have (2, 1 or 0) . ... As with y-intercepts, it may sometimes be difficult to read the ...  
The URL for this result is [www.wsd1.org/waec/math/Pre-Calculus%20Advanced/Quadratic%20Functions/Intercepts/interintro.htm](http://www.wsd1.org/waec/math/Pre-Calculus%20Advanced/Quadratic%20Functions/Intercepts/interintro.htm). It has 9k views, a "Cached" link, and a "Similar pages" link.  
  
The second search result is a snippet from "Yahoo! Canada Answers - What is the number of x-and y-intercepts ...". The snippet includes the following text:  
quadratic functions have exactly 1 y-intercept and no more than 2 x- ... The highest power of x, shows the maximum number of x-intercepts it 'could' have. ...  
The URL for this result is [answers.yahoo.ca/question/index?qid=20080428215000AAU30GI](http://answers.yahoo.ca/question/index?qid=20080428215000AAU30GI). It has 38k views, a "Cached" link, and a "Similar pages" link.  
  
The third search result is a snippet from "Quadratic Functions(General Form)". The snippet includes the following text:  
27 Nov 2007 ... You may change the values of coefficient a, b and c and observe the graphs obtained. ... When you graph a quadratic function, the graph will either have a maximum ... The x intercepts of the graph of a quadratic function f given by ... Use the applet window to check the y intercept for the quadratic ...

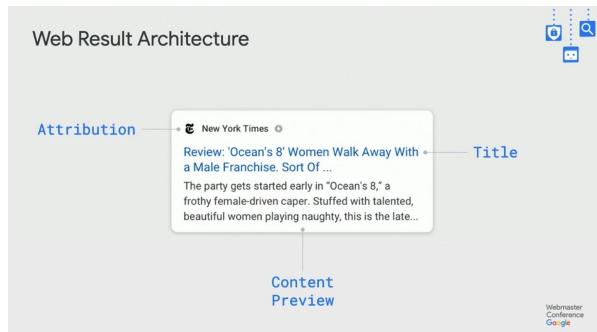
- if the snippet begins with ellipses ( . . . . ) that indicates the snippet was excerpted from a larger body of text and text preceding the ellipses was omitted
- when ellipses follow at the end of the snippet, the snippet was truncated
- the maximum length of a snippet is 156 characters
  - As we saw earlier, Google has played with the size
- Google uses the meta description (if there is one) as the default for a snippet
- if there is an Open Directory Project listing for a website, Google uses its meta description over the meta description in the web page
  - <http://www.dmoz.org/>
  - The Open Directory Project that uses human editors to organize websites closed as of March, 2017

# Google Snippet Lifecycle Changes

## Classic snippet

## Adding images in 2016

## Adding videos in 2018



### Images in Previews

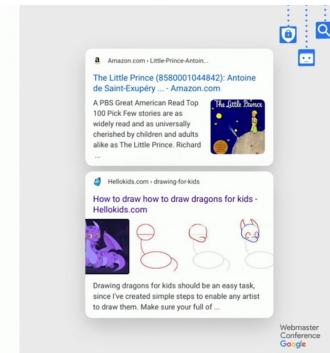
Images are relevant to the query

Placement on right

- Images are secondary to title/snippet

Galleries support contentful pages

Users visit a greater diversity of sites

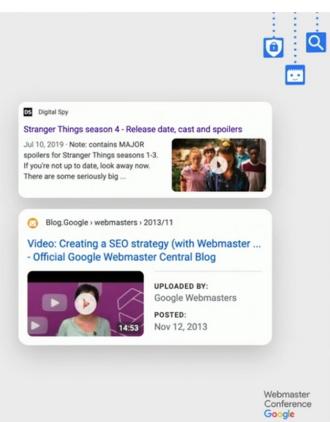


### Video in Previews

Video is relevant to the query

Preview indicates if video is dominant or supportive

Video metadata informs user experience



## Adding Sitelinks to snippets

### Sitelinks in Previews

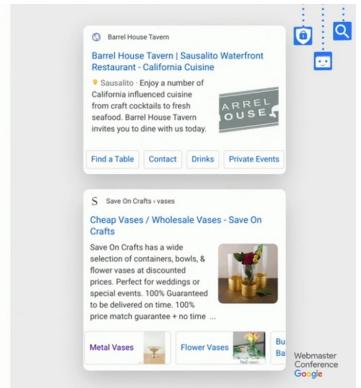
Links are relevant to query

Links extracted algorithmically

- Menus, Site-structure
- Drives traffic into a diverse set of sites

Sitelink-images help users

- Pithy links are better understood



## Adding Entity Facts

### Entity Facts in Previews

Relevance to needs around the entity

Facts extracted algorithmically

- Tables, Lists



## Adding Tables and Lists

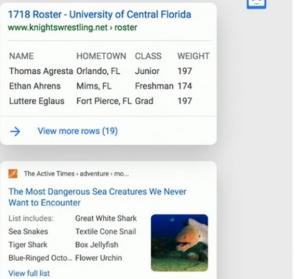
<https://youtu.be/ezLO7yC4aFo>, an 8 minute video  
Discussing Titles, Snippets and Result Previews

### Tables & Lists as Previews

Pages with dominant Tables/Lists

Helps users contrast content

Structure and position on the page guides the preview



- **Snippets are computed at query time**
  - They vary depending upon the query
  - the content that ends up in the text snippet can come from anywhere on your page. First sentence, last sentence, footer, call out box
- **If Google determines your site is a discussion forum, in gray text they put out**
  - "[number] posts – [number] authors – Last post: [some date]"

TESLA DISCUSSION FORUM autopilot accident

All News Videos Images Shopping More Settings Tools

About 1,670,000 results (0.45 seconds)

NTSB Wants Information on Tesla Autopilot Accident | Tesla Motors Club  
<https://teslamotorsclub.com> › General Forum › Autonomous Vehicles

Jan 23, 2018 - 20 posts - 15 authors

A Model S using Autopilot crashed into a firetruck near Los Angeles on Monday prompting inquiry from the U.S. National Transportation Safety Board, according to a report from Bloomberg. The Tesla driver was reportedly traveling at 65 mph when he rear-ended the truck. There were no injuries in the crash.

Autopilot worked for me today and saved an accident 20 posts Dec 12, 2016  
AutoPilot Crash today-Tesla response less than stellar? 20 posts Nov 7, 2016  
Tesla in Pasadena Accident: Driver Fled 20 posts Oct 7, 2016  
My friend's model X crashed using AP yesterday 20 posts Jul 10, 2016

More results from teslamotorsclub.com

- **If Google determines your site is a scholarly article, in gray text they put out**
  - "by J. Smith – 2010" or "by J. Smith – Cited by 1 – Related articles"

# Snippets Can Vary for a Single Site Depending Upon the Query

A screenshot of a Google search results page for the query "what cholesterol levels mean". The search bar at the top contains the query. Below it, the "All" tab is selected, along with other options like News, Images, Shopping, Videos, More, Settings, and Tools. A message indicates "About 828,000 results (0.58 seconds)". The first result is a snippet from the Cleveland Clinic's website, which includes a meta description: "HDL (high-density lipoprotein) cholesterol is also called "good" cholesterol. HDL protects against heart disease by taking the bad cholesterol out of your blood and keeping it from building up in your arteries. Your HDL cholesterol number is: Low (and considered a risk factor) if it is less than 40." Below this is a "People also ask" section with several questions: "What is a normal cholesterol level?", "What is the normal range for cholesterol levels?", "What is a healthy cholesterol level UK?", and "What foods are high in cholesterol?". Arrows point from the text "A long snippet, and a PAA" to the main snippet and the "People also ask" section respectively.

Result for the query  
“what cholesterol levels mean”

A long snippet, and a  
PAA

Google uses the meta description

A screenshot of a Google search results page for the query "cholesterol cleveland clinic". The search bar at the top contains the query. Below it, the "All" tab is selected, along with other options like News, Images, Shopping, Maps, More, Settings, and Tools. A message indicates "About 763,000 results (0.51 seconds)". The first result is a snippet from the Cleveland Clinic's website, which includes a meta description: "Understanding Your Cholesterol Numbers | Cleveland Clinic". Below this is another snippet from the same website, titled "Cholesterol-Lowering: Heart-Healthy Strategies - Cleveland Clinic", with a description: "Cholesterol is a soft, fat-like, waxy substance only found in animal products. Too much cholesterol leads to a build-up of fatty materials and debris (called plaque) on the walls of the arteries supplying blood to the heart and other organs. Some cholesterol is needed by the body. In fact, cholesterol plays a role in normal body ...". An arrow points from the text "A different query ‘cholesterol Cleveland Clinic’ produces the same first result but a different snippet" to the second snippet.

A different query  
“cholesterol Cleveland Clinic”  
produces the same first result  
but a different snippet

- **Automatic summarization** by computer is a traditional subject of *information retrieval*
- Automatic summarization is also part of *machine learning* and *data mining*
- Document summarization tries to create a representative summary or abstract of the entire document, by finding the most informative sentences
- There are two general approaches to automatic summarization:
  - extraction *Extractive methods work by selecting a subset of existing words, phrases, or sentences in the original text to form the summary*
  - abstraction *abstractive methods build an internal semantic representation and then use natural language generation techniques to create a summary that is closer to what a human might express*
  - *Research to date has focused primarily on extractive methods, which are appropriate for documents, images, and videos*

- Featured snippets are Google's attempt to answer the query right on the search results page.
  - *Introduced in 2016, Google wants to give the user an immediate answer so they don't have to search the actual results.*
  - *Featured snippets **show up above the #1 ranked spot**, and typically appear above the fold.*
  - *Google pulls snippet answers from pages that rank on Page 1 of the results for that query (spots #1 through #10) - but the page that wins the featured snippet isn't necessarily the #1 result. Google picks the excerpt from the page that best answers the query in a simple, concise format .*

## 1. Paragraph featured snippet

Marketing automation refers to the software that exists with the goal of automating marketing actions. Many marketing departments have to automate repetitive tasks such as emails, social media, and other website actions. The technology of marketing automation makes these tasks easier.

What is Marketing Automation? - HubSpot  
<https://www.hubspot.com/marketing-automation-information>

[About this result](#) [Feedback](#)

## 2. List featured snippet

### 14 of the Best College Websites

- University of Maryland. ...
- University of Notre Dame. ...
- Bucknell University. ...
- University of Chicago. ...
- University of Michigan. ...
- Rhode Island School of Design. ...
- George Washington University. ...
- Middlebury College.

[More items...](#)

### 14 of the Best College Websites (And Why They're So Awesome)

<https://blog.hubspot.com/marketing/best-college-websites>

[About this result](#) [Feedback](#)

## 3. Table featured snippet

Google aviation jobs

All News Shopping Maps Videos More Search tools

About 47,600,000 results (0.37 seconds)

Latest 15 Job Listings		
Date	ID #	Job Title
7/12/2016	4187	Project Engineer / AOD
7/1/2016	4185	Propulsion Engineer
6/29/2016	4184	A & P Technician
6/29/2016	4183	Accessory Shop Technician
27 more rows, 1 more column		

[Aviation Jobs](#) | [Aviation Job Seekers](#) | [Aviation Careers](#) | [Aviation ...](#)

[jobs.aviationweek.com/](http://jobs.aviationweek.com/) Aviation Week & Space Technology ▾

[About this result](#) [Feedback](#)

- Becoming a featured snippet can be achieved by simple on-page adjustments that very clearly define the topic to users
  - One of the goals of the featured snippet is to fuel voice search
    - *Create your text so it would answer a query clearly if read back on voice search?*
1. Look for a place in your content to add a “What Is [Keyword]” heading tag.
    - *This sends clear signals to Google that the following text could be used for the featured snippet*
  2. Use the “is statement” e.g.
    - *“Agile methodology is a type of project management process, mainly used for software development...”*
  3. Define the topic in 2 or 3 sentences
  4. Match the featured snippet format: paragraph, bulleted or numbered list, table
  5. Don’t use first person, e.g. “Our avocados have many health benefits . . . ”
- For more details see
  - <https://searchengineland.com/featured-snippets-the-9-rules-of-optimization-342627>

# TOO LONG DIDN'T READ (TLDR)

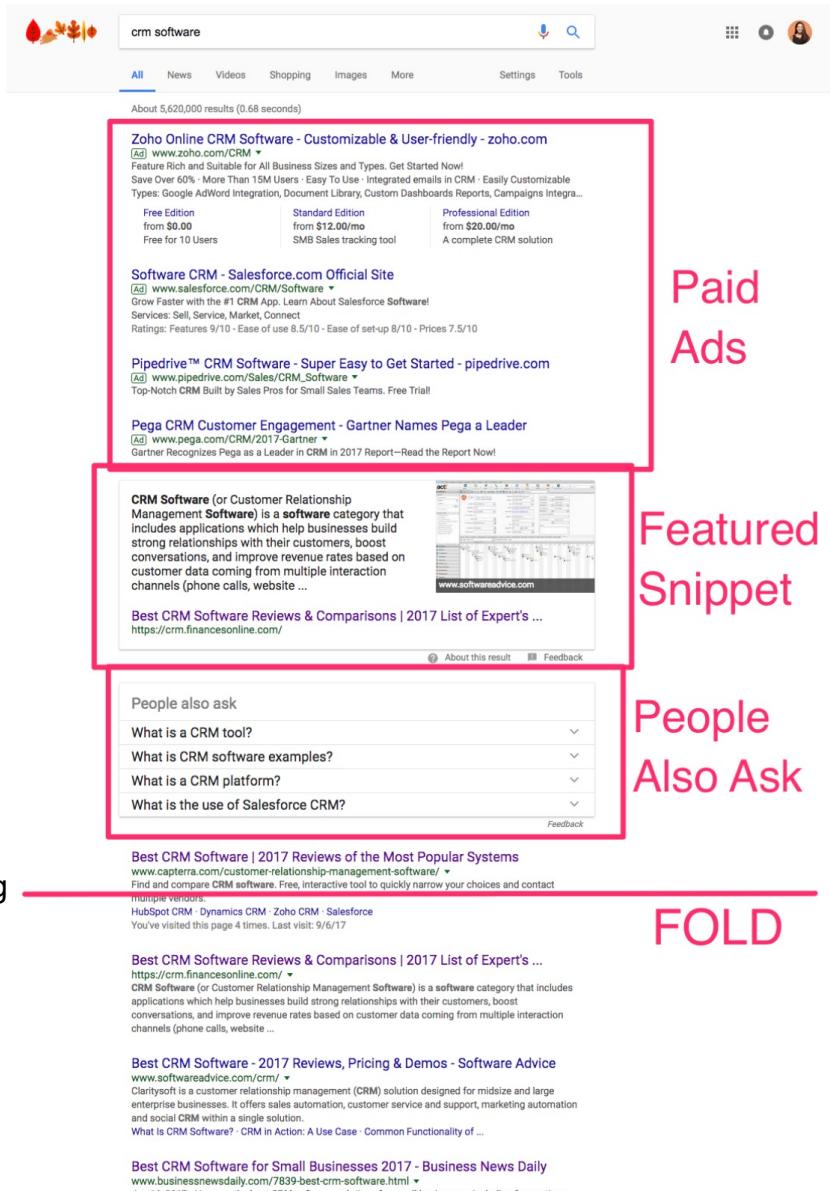
## Internet slang

For the query “CRM software”, which is a very popular query term, above the fold there are:

- 4 paid ads
- A featured snippet paragraph
- People Also Ask
- A portion of the #1 ranking organic result

- **Conclusion:** it is harder than ever to be found in the organic search results

Above the **fold** refers to a **search engine results page** ranking on the first **page** that is visible without having to scroll down



# Extracting a Snippet is Not Always Easy

A screenshot of a Google search results page. The query "tesla announces quarterly" is entered in the search bar. The top result is a snippet from the New York Times article titled "Apple and Tesla to Report Earnings". The snippet includes a photo of Elon Musk and the text: "Elon Musk, chief executive of Tesla Motors, which will report its third-quarter earnings on Wednesday. ... On Thursday, the Commerce Department will announce data on ...".

query: "Tesla reports financial results"

one search result and its snippet

A screenshot of the New York Times article "Apple and Tesla to Report Earnings". The headline is "Apple and Tesla to Report Earnings". Below it, a sub-headline reads "A loss is likely for Tesla, despite popularity.". The main text discusses Tesla's financial performance, mentioning deliveries rose 70 percent and the introduction of the Model 3. It also notes that investors will listen closely to Elon Musk and details on a pending merger with SolarCity.

portions of the article used to create the snippet; note how long the article is; "financial results" equates to "earnings"

A screenshot of the New York Times article "Apple and Tesla to Report Earnings". A snippet from the article is shown, reading: "Economists predict rise in orders for durable goods. On Thursday, the **Commerce** Department will announce data on durable goods orders in September. Economists are predicting that durable goods demand will be up by 0.7 percentage point, with much of that gain resulting from a jump in aircraft orders. The less volatile core capital goods category is thought to have declined by 0.1 percentage point, on continuing caution among businesses in terms of new investment, especially in the industrial sector. *Nelson D. Schwartz*".

Growth is expected to improve.

On Friday, the **Commerce** Department will release its initial estimate of economic growth in the third quarter. After lackluster gains in the first half of 2016 — 0.8 percentage point in the first quarter and 1.4 percent in the second — economists think the economy expanded at an annual rate of 2.5 percent in the July-to-September period.

# Extracting a Snippet is Not Always Easy Nor Obvious

A screenshot of a Google search results page. The query "google introduces cloud" is entered in the search bar. The top result is a link to "Cloud computing - A simple introduction - Explain that Stuff" from [www.explainthatstuff.com/cloud-computing-introduction.html](http://www.explainthatstuff.com/cloud-computing-introduction.html). The snippet below the title reads: "Aug 13, 2016 - An easy-to-understand introduction to cloud computing. ... When you sit at your PC and type a query into Google, the computer on your desk ...".

query: "cloud computing"

one search result  
and its snippet

A screenshot of the "Cloud computing" article from [www.explainthatstuff.com/cloud-computing-introduction.html](http://www.explainthatstuff.com/cloud-computing-introduction.html). The page features a large image of server racks, a title "Cloud computing", and a short bio by Chris Woodford. The main content starts with: "History has a funny way of repeating itself, or so they say. But it may come as some surprise to find this old cliché applies just as much to the history of computers as to wars, revolutions, and kings and queens. For the last three decades, one trend in computing has been loud and clear: big, centralized, mainframe systems have been "out"; personalized, power-to-the-people, do-it-yourself PCs have been "in." Before personal computers took off in the early 1980s, if your company needed sales or payroll figures calculating in a hurry, you'd most likely have bought in "data-processing" services from another company, with its own expensive computer systems, that specialized in number crunching, these days, you can do the job just as easily on your desktop with off-the-shelf software. Or can you? In a striking throwback to the 1970s, many companies are finding, once again, that buying in computer services makes more business sense than do-it-yourself. This new trend is called **cloud computing** and, not surprisingly, it's linked to the Internet's inexorable rise. What is cloud computing? How does it work? Let's take a closer look!"

"An easy-to-understand introduction"  
**occurs nowhere in the article**  
**It is in the meta-description**  
"sit at your PC" occurs lower in the article

A screenshot of the same "Cloud computing" article with a cursor pointing to the search bar at the top of the page. The search bar contains the text "Type a query into Google". Below the search bar, the first few lines of the article are visible: "Simple examples of cloud computing".

Most of us use cloud computing all day long without realizing it. When you **sit at your PC** and **type a query into Google**, the computer on your desk isn't playing much part in finding the answers you need: it's no more than a messenger. The words you type are

# How Does Google Generate Snippets?

One way to find out is to go to [patents.google.com](http://patents.google.com) and search for all patents with the term “snippets” assigned to Google

Many are patent applications still being reviewed by the patent office

Some are already awarded

The screenshot shows a Google Patents search interface. The search bar contains 'snippets' with a 'Synonym' button. Below it, there's a 'Search Fields' section with dropdowns for 'Assignee' set to 'google' and 'Priority date' set to 'Before priority YYYY-MM-DD'. A 'MORE' button is also present. The search results page displays several patent documents:

- G06F17/30861?**  
Retrieval from the Internet, e.g. browsers  
Generating snippets for prominent users for information retrieval queries ...  
Application WO201405576A3 • Bogdan DOROHONCEANU • Google Inc.  
Priority 2012-10-04 • Filed 2013-10-03 • Published 2014-07-10
- Expanded snippets** A system provides a list of search results, where one of the ...  
Application WO2007115079A3 • Paul Fontes • Google Inc.  
Priority 2006-03-31 • Filed 2007-03-29 • Published 2007-11-22
- Expanded snippets** A system provides a list of search results, where one of the search results includes a snippet from a corresponding search result document. The system receives selection of the snippet and provides an expanded snippet based on the selection of the ...
- Variable length snippet generation** A method and system are disclosed that ...  
Application WO2006001920A1 • Paul Buchheit • Google Inc.  
Priority 2004-06-09 • Filed 2005-05-10 • Published 2006-01-05
- Local Search Using Address Completion** A local search server receives ...  
Application US20080065694A1 • Jiang Qian • Google Inc.  
Priority 2006-09-08 • Filed 2007-05-22 • Published 2008-03-13
- Document search engine including highlighting of confident results** A search ...  
Application US20110029518A1 • Simon Tong • Google Inc.  
Priority 2003-06-10 • Filed 2010-10-08 • Published 2011-02-03
- System and method for personalized snippet generation** Snippets of text ...  
Grant US8631006B1 • Taher H. Haveliwala • Google Inc.  
Priority 2005-04-14 • Filed 2005-04-14 Granted 2014-01-14

At the bottom of the search results, there are links for 'About', 'Send Feedback', 'Terms', and 'Privacy Policy'.

# Lets take a closer look US Patent 8,145,617

## Title:

*Generation of document snippets based on queries and search results*

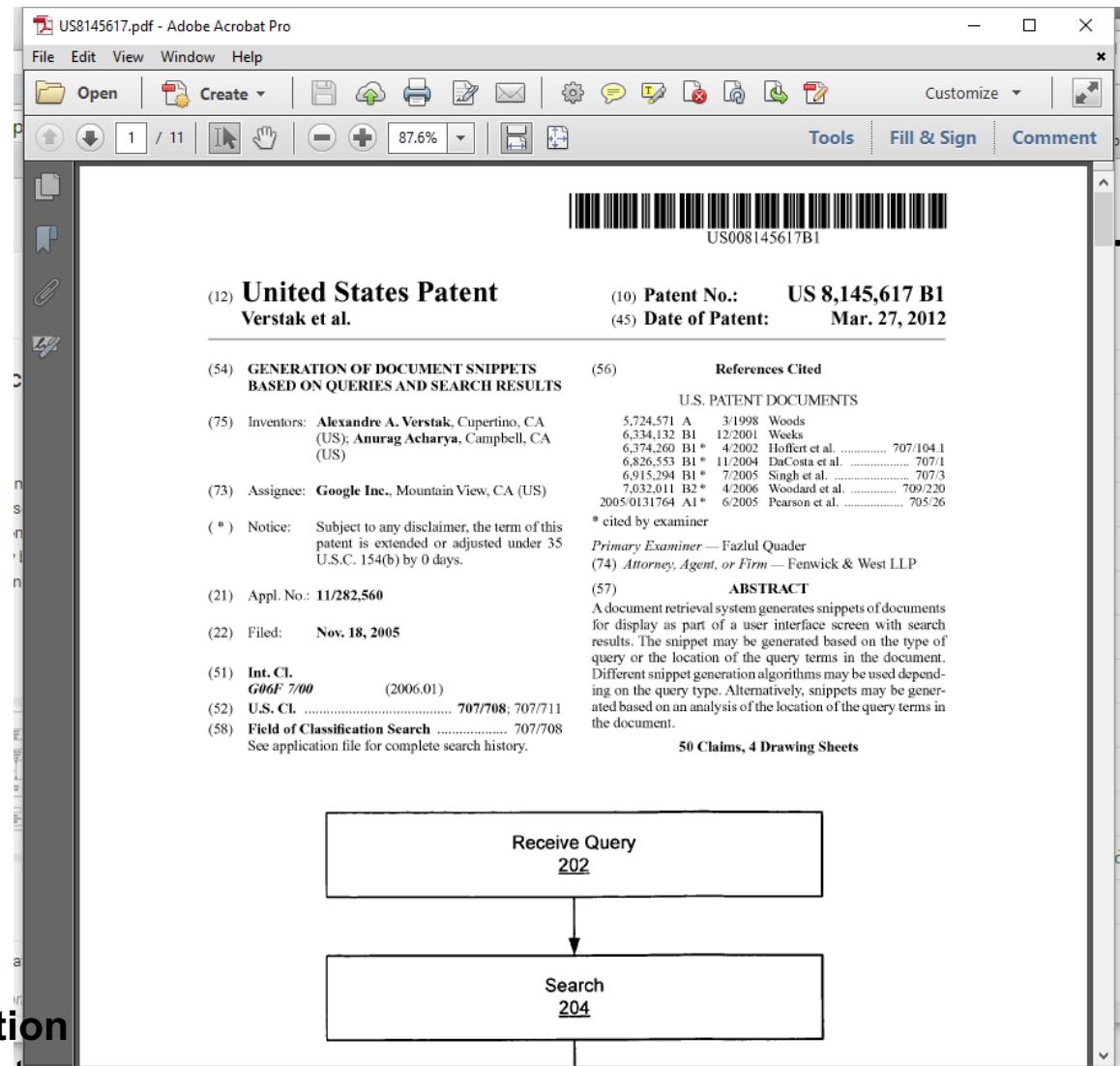
**filed:** 2005

**awarded:** 2012

## Abstract

A document retrieval system generates snippets of documents for display as part of a user interface screen with search results. The snippet may be generated *based on the type of query or the location of the query terms in the document.*

Different snippet generation algorithms may be used depending on the query type. Alternatively, snippets may be generated **based on an analysis of the location of the query terms in the document**



- **The algorithm**
  1. Identify the paragraphs that include the query terms
  2. Score the paragraphs as described below determining the paragraph with the highest score
  3. *Return the phrase in that paragraph that includes the query terms*

## Quoting from the Detailed Description

- The snippet algorithm selects a paragraph that is near the **beginning** of the document if there is an abstract, executive summary, or long introduction. The **end** of the document is used when there is a conclusion or summarization at the end
- **Scoring includes:**  
paragraphs shorter than threshold score 0;  
k-th paragraph from the start gets a score of

$k\text{-positionFactor} + \max(\text{actual paragraph length}, \text{maxParagraphLen})$

***The paragraph with the highest score is selected for the snippet***

- **Use the meta description**
- **Location based rules**
  - Examine the length of the paragraphs that include the query terms and its distance from either the beginning or end of the document
- **Language dependent rules**
  - The number of bold or italicized words in the paragraph
- Some paragraphs (in part or full) that include all of the query terms might rank poorly as choices for snippets for other reasons. These paragraphs might end up with a score of 0 because they:
  - *Are shorter than a certain threshold*
  - *Are mostly punctuation, or have punctuation above a certain threshold*
  - *Contain italicized or bold words above a certain threshold*
  - *Are too far from the start or the end of a page*

*System and Method for  
Personalized Snippet Generation***Filed:** April 14, 2005**Awarded:** Jan. 14, 2014**Abstract:**

*Snippets of text are generated based in part on a user's profile. An item, such as a document, is examined to identify terms related to the user's profile. A term profile for an identified term is compared to a user's profile. The more closely related the identified term is to the user's profile, the higher a similarity score will be. Alternatively, terms found in a document may have a user profile score which may be obtained by looking the term up in the user's profile. Terms having high profile similarity scores or high user profile scores are used in identifying snippets which may be relevant to a user. The high scoring terms may be added to search terms and provided to a snippet generator*



US008631006B1

(12) **United States Patent**  
Haveliwala et al.(10) **Patent No.:** US 8,631,006 B1  
(45) **Date of Patent:** Jan. 14, 2014(54) **SYSTEM AND METHOD FOR  
PERSONALIZED SNIPPET GENERATION**(75) Inventors: **Taher H. Haveliwala**, Mountain View, CA (US); **Sepandar D. Kamvar**, San Francisco, CA (US)(73) Assignee: **Google Inc.**, Mountain View, CA (US)

(\*) Notice: Subject to any disclaimer, the term of this patent is extended or adjusted under 35 U.S.C. 154(b) by 1650 days.

(21) Appl. No.: 11/107,490

(22) Filed: Apr. 14, 2005

(51) **Int. Cl.**  
*G06F 7/00* (2006.01)  
*G06F 17/30* (2006.01)(52) **U.S. Cl.**  
USPC ..... 707/732; 707/722; 707/723(58) **Field of Classification Search**  
USPC ..... 707/3, 5, 722, 723, 732  
See application file for complete search history.(56) **References Cited**

## U.S. PATENT DOCUMENTS

6,144,944 A	11/2000	Kurtzman, II et al.
6,275,820 B1	8/2001	Navin-Chandra et al.
6,701,310 B1 *	3/2004	Sugiura et al. .... 707/5

7,092,901	B2 *	8/2006	Davis et al. ....	705/26
7,165,091	B2 *	1/2007	Lunenfeld ....	709/203
7,418,447	B2 *	8/2008	Caldwell et al. ....	707/100
2003/0009440	A1 *	1/2003	Inaba et al. ....	707/1
2004/0034652	A1 *	2/2004	Hofmann et al. ....	707/102
2004/0236721	A1 *	11/2004	Pollack et al. ....	707/2
2004/0267723	A1	12/2004	Bharat	
2005/0240580	A1 *	10/2005	Zamir et al. ....	707/4
2006/0074883	A1 *	4/2006	Teevan et al. ....	707/3
2006/0112079	A1 *	5/2006	Holt et al. ....	707/3
2006/0248059	A1 *	11/2006	Chi et al. ....	707/3

\* cited by examiner

Primary Examiner — Apu Mofiz

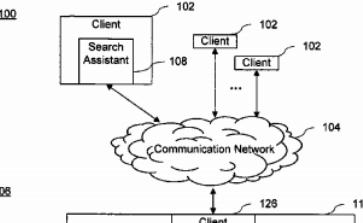
Assistant Examiner — Jared Bibbee

(74) Attorney, Agent, or Firm — Morgan, Lewis &amp; Bockius LLP

**ABSTRACT**

Snippets of text provided are generated based in part on a user's profile. An item, such as a document, is examined to identify terms related to the user's profile. A term profile for an identified term is compared to a user's profile. The more closely related the identified term is to the user's profile, the higher a similarity score will be. Alternatively, terms found in a document may have a user profile score which may be obtained by looking the term up in the user's profile. Terms having high profile similarity scores or high user profile scores are used in identifying snippets which may be relevant to a user. The high scoring terms may be added to search terms and provided to a snippet generator.

13 Claims, 6 Drawing Sheets



# Featured Snippets Results in Google Web Search: An Exploratory Study – Strzelecki, Rutecka

**Table 1.** Type of featured snippet.

featured type	frequency	percentage
paragraph	114465	70,05%
list	46509	28,46%
table	2438	1,49%

Paragraph snippets are the overwhelming type

**Table 2.** Ranking position for featured snippet

position	frequency	percentage
0	485	0,30%
1	79867	48,87%
2	30618	18,74%
3	20878	12,78%
4	14469	8,85%
5	9582	5,86%
6	2860	1,75%
7	1909	1,17%
8	1319	0,81%
9	860	0,53%
10	554	0,34%

Position 1, the second position on the SERP is most common

<https://www.nwsdigital.com/Blog/What-is-the-Zero-Position#>

**Table 3.** Other snippets displayed along with featured snippet

params	frequency	percentage
image thumbs	102934	62,99%
site links	41348	25,30%
brand	24214	14,82%
wiki	18675	11,43%
ads	3148	1,93%
name	2850	1,74%
map	1807	1,11%
city	1062	0,65%
news	107	0,07%

# Google's People Also Ask (PAA) Feature Introduced in 2015 for desktop and mobile

A screenshot of a Google search results page. The search bar at the top contains the query "how does google generate snippets". Below the search bar are navigation links for "All", "Videos", "News", "Images", "Shopping", "More", "Settings", and "Tools". A status message indicates "About 3,830,000 results (0.49 seconds)". The first result is a sponsored link titled "Rich Snippets & Star Ratings | Display Reviews & Increase CTR" from www.yotpo.com. The second result is a link titled "Anatomy Of A Google Snippet - Search Engine Land" from https://searchengineland.com. The third result is a link titled "How Google Might Generate Snippets for Search Results" from www.seobythesea.com. The fourth result is a link titled "Google's Featured Snippets: How to Get Your Content to Appear" from contentmarketinginstitute.com.

About 3,830,000 results (0.49 seconds)

## Rich Snippets & Star Ratings | Display Reviews & Increase CTR

Ad www.yotpo.com/Rich\_Snippet ▾ (646) 655-8389

Get high quality traffic from your paid & organic search results. Get a Demo!

## Anatomy Of A Google Snippet - Search Engine Land

https://searchengineland.com/anatomy-of-a-google-snippet-38357 ▾

Mar 18, 2010 - This is a crucially important detail: snippets are determined query-time; in other words, they vary depending on the keyword being searched on, ... price, size, style, manufacturer) would be gathered together — since it would otherwise be unlikely that a Google-generated snippet would capture all of this ...

## How Google Might Generate Snippets for Search Results

www.seobythesea.com › Search Engine Optimization (SEO) ▾

Feb 25, 2013 - A document retrieval system generates snippets of documents for display as part of a user interface screen with search results. The snippet may be generated based on the type of query or the location of the query terms in the document. Different snippet generation algorithms may be used depending on the query type.

You've visited this page 3 times. Last visit: 10/25/17

## Google's Featured Snippets: How to Get Your Content to Appear

contentmarketinginstitute.com/2017/03/google-featured-snippets/ ▾

Mar 27, 2017 - The primary thing you need to understand about featured snippets is that they do not bypass Google's complex ranking system. They are among the topic organic results for a query. Because of this, it's critical to keep all the standard ranking requirements in mind as you create featured content. This means ...

### People also ask

How does Google select featured snippets?

What is a snippet in Google?

What is the use of snippets SEO?

What is snippet generation?

Feedback

### PAA result for query “how does google generate snippets

In one study, the “People Also Ask” box appeared on 364 keywords out of 1,788, 20%.

### People also ask

#### How does Google select featured snippets?

Here are a few simple steps I've used to create content that ranks in the snippets.

1. Create content specifically to answer questions. Provide in-depth answers. ...
2. Know the questions your readers are asking. ...
3. Create truly high-quality content. ...
4. Work to provide the best answer. ...
5. Use question-and-answer pages.

#### Google's Featured Snippets: How to Get Your Content to Appear

contentmarketinginstitute.com/2017/03/google-featured-snippets/

Search for: How does Google select featured snippets?

#### What is a snippet in Google?

Rich Snippets is the term used to describe structured data markup that site operators can add to their existing HTML, which in turn allow search engines to better understand what information is contained on each web page.

#### A Beginner's Guide to Rich Snippets | Unamo Blog

https://unamo.com/blog/seo/beginners-guide-rich-snippets

Search for: What is a snippet in Google?

What is the use of snippets SEO?

What is snippet generation?

What is a featured snippet?

What is a snippet of a song?

What is the code snippet?

Feedback

### Expansion of People Also Ask

# People Also Ask (PAA) is Growing Fast

- The “People Also Ask” box is a Google universal SERP result that answers questions related to the searcher’s initial query.
- It is a cousin of the featured snippet
- Each PAA box contains anywhere from one to four related questions which expand to reveal answers that Google has pulled from other websites
- The site’s URL appears below each answer, along with a “Search for” link, which guides the user to a Google SERP of the PAA question.



Use of PAAs are growing faster than snippets according to  
<https://moz.com/blog/infinite-people-also-ask-boxes>  
Copyright Ellis Horowitz 2012 - 2022

- In 2009, Google announced *Rich Snippets*, a mechanism *for website developers* to include information that Google's results algorithm will *display as a snippet*
- The mechanism calls for *embedding structured data in web pages* with the objective of displaying the structured data to a user in a visually outstanding way.
- Rich Snippets give users a convenient summary information about their search results at a glance.

For example,

the results for Club

Deluxe includes  
internal data such as:

address

hours

directions

club deluxe san francisco

Search

About 277,000 results (0.20 seconds)

Advanced search

Club Deluxe - Pizza & Jazz Club

Place page

1511 Haight Street  
San Francisco, CA 94117-2912  
(415) 552-6949  
Public transit: Cole St & Carl St  
[Get directions](#) - Is this accurate?

Open Weekdays 4pm-2am; Weekends 2pm-2am

29 reviews - [Write a review](#)



A map of the Haight-Ashbury neighborhood in San Francisco, showing the intersection of Haight Street and Page Street. A red marker indicates the location of Club Deluxe. Labels on the map include Hayes St, Fell St, Panhandle, Oak St, K St, Cole St, Clayton St, Ashby St, Haight St, Red Victorian, Booksmith, and Buena Vista Park. Arrows point from the text labels 'address', 'hours', and 'directions' to the corresponding information on the right side of the search result.

[Club Deluxe - Haight-Ashbury - San Francisco, CA](#) 

 2.5 stars - Price range: \$\$

214 Reviews of Club Deluxe "This is like my own little hidden GEM in the Haight. I swear I had walked by this place hundreds of times before I finally ..."

[www.yelp.com/biz/club-deluxe-san-francisco](http://www.yelp.com/biz/club-deluxe-san-francisco) - 8 hours ago - Cached - Similar

# Rich Snippets Examples: People Snippets



pravir gupta

About 13,400 results (0.28 seconds)

Search

Advanced search

Everything

More

Show search tools

[Pravir Gupta | Facebook](#) ☆

Friends: Sam Tyagi, Geeta Shroff, Siddarth Jain, Shradha Balakrishnan, Richa Kumar  
**Pravir Gupta** is on Facebook. Join Facebook to connect with **Pravir Gupta** and others you may know. Facebook gives people the power to share and makes the ...  
[www.facebook.com/pravigupta](http://www.facebook.com/pravigupta) - Cached

[Home \(pravir\)](#) ☆

Pravir Gupta. ... attachment removed by Pravir Gupta. edited by Pravir Gupta ... created by Pravir Gupta. Home. created by Pravir Gupta ...  
[pravigupta.com/](http://pravigupta.com/) - Cached - Similar

[Pravir Gupta - Knol: a unit of knowledge](#) ☆

Pravir Gupta. Verify Name. Agra, India. Public activity feed. Sort by: ... byPravir Gupta. We are continuously looking at enabling sites. ...  
[knol.google.com/k/pravir-gupta/-/3philmrwubhfj/0](http://knol.google.com/k/pravir-gupta/-/3philmrwubhfj/0) - Cached - Similar

[The Journey is the Reward - a knol by Pravir Gupta](#) ☆

Jul 20, 2009 ... Debut novel by Anil Kumar Gupta which was published in July 2009.  
[knol.google.com/k/pravir-gupta/the-journey-is-the-reward/.../4](http://knol.google.com/k/pravir-gupta/the-journey-is-the-reward/.../4) - Cached

[+ Show more results from knol.google.com](#)

[Pravir Gupta - Senior Software Engineer | LinkedIn](#) ☆

San Francisco Bay Area - Senior Software Engineer  
View Pravir Gupta's (87 connections) professional profile on LinkedIn. LinkedIn is the world's largest business network, helping professionals like Pravir ...  
[www.linkedin.com/pub/pravir-gupta/2/180/a70](http://www.linkedin.com/pub/pravir-gupta/2/180/a70) - Cached - Similar

[Pravir Gupta - Directory | LinkedIn](#) ☆

View the profiles of professionals named Pravir Gupta on LinkedIn. There are 2 professionals named Pravir Gupta who use LinkedIn to exchange information, ...  
[www.linkedin.com/pub/dir/Pravir/Gupta/](http://www.linkedin.com/pub/dir/Pravir/Gupta/)

[Pravir Gupta, Google Inc, Mountain View, CA | Spoke](#) ☆

Pravir Gupta, Google Inc of Google Inc's information - including email, business address, business phone, biography, title, company, jobs and associations, ...  
[www.spoke.com/info/p90ookh/PravirGupta](http://www.spoke.com/info/p90ookh/PravirGupta) - Cached

here the snippets describe the pages containing the information about the individual:  
Facebook,  
LinkedIn,  
Google



# Rich Snippets Examples: Events



fillmore events

Search

About 1,610,000 results (0.21 seconds)

[Advanced search](#)

Everything

More

All results

Timeline

More search tools

[The Fillmore Concert Tickets, Schedule, Seating Chart | Official ...](#)

Get email alerts and never miss your favorite **events** at The Fillmore. Please enter your e-mail address. That is not a valid e-mail address format. ...  
[www.thefillmore.com/](http://www.thefillmore.com/) - Cached - Similar

[The Fillmore San Francisco - The Fillmore Schedule | Eventful](#)

View The Fillmore's upcoming **event** schedule and profile - San Francisco, CA. The Fillmore, also known as Fillmore Auditorium, is located in San ...

[Carolina Chocolate Drops](#) Thu, Jun 24

[Josh Ritter & the Royal City Band](#) Thu, Jun 24

[Robert Earl Keen](#) Sat, Jun 26

[eventful.com](http://eventful.com) > San Francisco venues - Cached - Similar

[Fillmore Events: Events in Fillmore, California](#)

**Fillmore Events** Directory. Includes listings for Events in Fillmore, California.  
[www.californiacoast-worldweb.com/Fillmore/Events/](http://www.californiacoast-worldweb.com/Fillmore/Events/) - Cached - Similar

[San Francisco The Fillmore Events, Shows & Things to do - SF Gate](#)

Find 48 San Francisco The **Fillmore events** and show tickets and more on Zvents. Popular The **Fillmore Events** are Salsa Festival on the Fillmore, Fillmore Jazz ...  
[events.sfgate.com/san-francisco-ca/events/the+fillmore](http://events.sfgate.com/san-francisco-ca/events/the+fillmore) - Cached

[New York Fillmore Events Events, Shows & Things to do - NY Daily News](#)

Find 29 New York **Fillmore Events** events and show tickets and more on Zvents. Popular **Fillmore Events** Events are On Fillmore Plus Rachel Grimes, ...  
[events.nydailynews.com/new-york-ny/events/fillmore+events](http://events.nydailynews.com/new-york-ny/events/fillmore+events) - Cached

[Charlotte Charlotte Fillmore Events, Shows & Things to do - The ...](#)

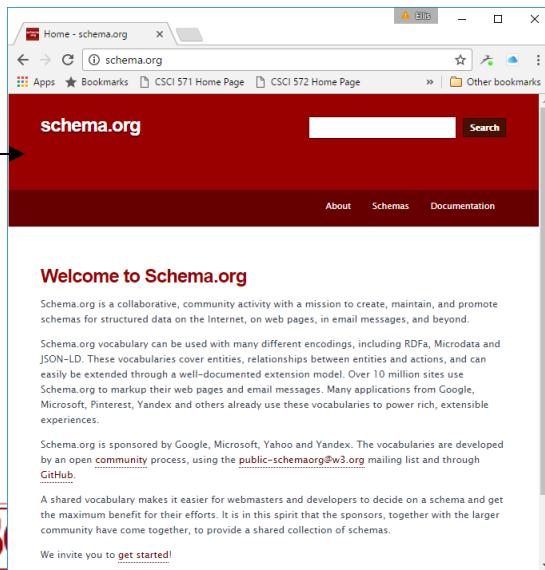
Find 8 Charlotte Charlotte **Fillmore events** and show tickets and more on Zvents. Popular Charlotte **Fillmore Events** are Smashing Pumpkins, Adam Lambert with ...  
[events.charlotteobserver.com/charlotte-nc/events/charlotte+fillmore](http://events.charlotteobserver.com/charlotte-nc/events/charlotte+fillmore) - Cached

the Filmore theatre can highlight future concerts by regularly updating their webpage with the latest rich snippet information

## Benefits of Rich Snippets in Google Search ...

- **Webmasters:** Provides webmasters the ability to add useful information to their web search result snippets to help Google make sense of their bits.
- **Purpose** Provides more information to a user about the content that exists on page so they can decide which result is more relevant for their query.
- Two good reasons for using rich snippets
  - 1. Additional traffic to a webpage** With extra information people tend to rely more on a particular search result with linked data, thus an increasing number of impressions noted on sites with Rich Snippets.
  - 2. Higher Click Through Rate** An increasing number of higher click-through rate for pages with Rich Snippets was experienced as shown in a paper by *Kavi Goel, Pravir Gupta*
    - <http://www.dataversity.net/google-yahoo-and-bing-announce-schema-org/>
- **Easy to add** simple lines of Markup to existing HTML, no affect to visual appearance of the webpage.

- In June, 2011 Google, Yahoo, and Bing agree on a single standard
- They establish the website schema.org which defines the mechanism for creating rich snippets
- They decide to standardize on microdata format
- <https://developers.google.com/structured-data/rich-snippets/>



## Google, Yahoo! and Bing Announce Schema.org

By Eric Franzon / June 2, 2011 / 0 Comments



[Revised and re-posted at 4:03pm EST]



In a collaborative effort reminiscent of sitemaps.org, Google, Yahoo! and Bing have announced the launch of schema.org. Perhaps the most significant aspect of this announcement is the particular standard they have focused on: namely, microdata.

In the [Google announcement](#), Kavi Goel and Pravir Gupta of Google's search team say, "Historically, we've supported three different standards for structured data markup: microdata, microformats, and RDFa. We've decided to focus on just one format for schema.org to create a simpler story for webmasters and to improve consistency across search engines relying on the data."

From the [Yahoo! announcement](#) comes this: "Today's announcement offers tremendous opportunity for growth. In addition to consolidating the schemas for the vocabularies we already support, there are schemas for more than a hundred newly created categories including movies, music, organizations, TV shows, products, places and more. We will continue to expand these categories by listening to feedback from the community and will continue publishing new schemas on a regular basis. Don't worry if your site has already added RDFa or microformats currently supported by our Enhanced Displays program, that site will still appear with an Enhanced Display on Yahoo! – no changes required."

And [Bing](#) has this to add: "At Bing we understand the significant investment required to implement markup, and feel strongly that by partnering with Google and Yahoo! on standard schemas webmasters can be more efficient with the time they invest... Bing accepts a wide variety of markup formats today (Open Graph, microformat, etc.) for features like Tiles and will continue to do so, but by standardizing on schema.org we are looking to simplify the markup choices for webmasters and amplify the value the receive in return."

The schema.org site "provides a collection of schemas, i.e., html tags, that webmasters can use to markup their pages in ways recognized by major search providers. Search engines...rely on this markup to improve the display of search results, making it easier for people to find the right web pages."

# Rich Snippet Technology Definitions

- Google suggests using the microdata formalism for snippets
- <http://www.whatwg.org/specs/web-apps/current-work/multipage/microdata.html>

A screenshot of a web browser window displaying the HTML Standard specification. The title bar says "HTML Standard". The address bar shows the URL "https://html.spec.whatwg.org/multipage/microdata.html". The main content area is titled "HTML" and "Living Standard — Last Updated 21 October 2016". Below this, there is a "Microdata" section with several sub-sections listed: "Introduction", "Overview", "The basic syntax", "Typed items", "Global identifiers for items", "Selecting names when defining vocabularies", "Encoding microdata", "The microdata model", and "Items". At the bottom of the page, there is a link "File an issue about the selected text".

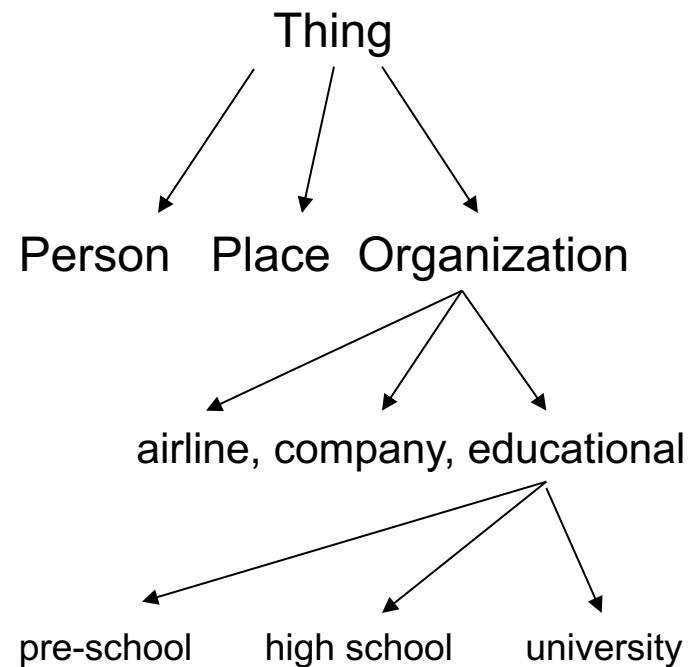
<https://www.w3.org/TR/microdata/>

Now goes to  
<https://html.spec.whatwg.org/multipage/>

A screenshot of a web browser window displaying the W3C Working Group Note for HTML Microdata. The title bar says "W3C Working Group Note". The address bar shows the URL "https://www.w3.org/TR/microdata/". The main content area is titled "HTML Microdata" and "W3C Working Group Note 29 October 2013". It provides links to various versions of the specification, including "This Version", "Latest Published Version", "Latest Editor's Draft", "Previous Versions", and "Editor".

- Two other formalisms for creating rich snippets have been suggested:
- RDFa (Resource Description Framework – in Attributes)  
<http://en.wikipedia.org/wiki/RDFa>
- Microformat Encoding  
<http://en.wikipedia.org/wiki/Microformat>

- Schema.org defines an object hierarchy
- The most general item type is Thing with properties: name, description, url, and image
  - Person, Place and Organization are types of Things
  - More specific items inherit the properties of their parent
- Some commonly used types include:
  - Creative works: book, movie, music recording, recipe, TV Series
  - Embedded object: image, video
  - Event
  - Organization
  - Person
  - Place, LocalBusiness, e.g. Restaurant
  - Product, Offer, Aggregate Offer
  - Review, AggregateRating



## Entities supported by Google Rich Snippets as of now....

- Software applications
- Breadcrumbs
  - a breadcrumb trail on a page indicates the page's position in the site hierarchy. A user can navigate all the way up in the site hierarchy, one level at a time, by starting from the last breadcrumb in the breadcrumb trail
  - for example, [Books](#) > [Authors](#) > [Ann Leckie](#) > [Ancillary Justice](#)
- Events
- Music
- Businesses and Organizations
- People
- Products
- Recipes
- Review Ratings
- Reviews: should include: item being reviewed, reviewer rating, date
- Videos: Facebook Share

- **Microformats** use only existing HTML, e.g. the *class* attribute in HTML tags (often `<span>` or `<div>`) to assign brief and descriptive names to entities and their properties
- **Microdata** extends HTML5 by introducing new attributes like `itemprop`
- **Microformat Example**

```
<div class="vcard">

 <strong class="fn">Bob Smith
 Senior editor at ACME Reviews

 200 Main St
 Desertville, AZ
 12345

</div>
```

microformat class attributes in this example include vcard, photo, title, org, adr, locality, etc

# A MicroData Example: A Web Page About the Movie Avatar

- **To begin, identify the section of the page that is "about" the movie Avatar. To do this, add the itemscope element to the HTML tag that encloses information about the item, and you can specify the type of item using the itemtype attribute like this:**

```
<div itemscope itemtype="http://schema.org/Movie">
 <h1>Avatar</h1>
 Director: James Cameron (born August 16, 1954)
 Science fiction
 Trailer
</div>
```

- **By adding itemscope, you are specifying that the HTML contained in the <div>...</div> block is about a particular item.**

- The **itemprop** attribute is used to label properties of a movie such as actors, director, ratings.
- For example, to identify the director of a movie, add **itemprop="director"** to the element enclosing the director's name. (There's a full list of all the properties you can associate with a movie at <http://schema.org/Movie>.)

```
<div itemscope itemtype ="http://schema.org/Movie">
<h1 itemprop= "name">Avatar</h1>
Director: James Cameron (born
 August 16, 1954)
Science fiction
Trailer
</div>
```

# Partial List of Movie Properties (Schema.org/Movie)

**Thing > CreativeWork > Movie**

A movie.

Property	Expected Type	Description
<b>Properties from Thing</b>		
<code>description</code>	Text	A short description of the item.
<code>image</code>	URL	URL of an image of the item.
<code>name</code>	Text	The name of the item.
<code>url</code>	URL	URL of the item.
<b>Properties from CreativeWork</b>		
<code>about</code>	Thing	The subject matter of the content.
<code>accountablePerson</code>	Person	Specifies the Person that is legally accountable for the CreativeWork.
<code>aggregateRating</code>	AggregateRating	The overall rating, based on a collection of reviews or ratings, of the item.
<code>alternativeHeadline</code>	Text	A secondary title of the CreativeWork.
<code>associatedMedia</code>	MediaObject	The media objects that encode this creative work. This property is a synonym for encodings.
<code>audio</code>	AudioObject	An embedded audio object.
<code>author</code>	Person or Organization	The author of this content. Please note that author is special in that HTML 5 provides a special mechanism for indicating authorship via the rel tag. That is equivalent to this and may be used interchangeably.
<code>award</code>	Text	An award won by this person or for this creative work.
<code>awards</code>	Text	Awards won by this person or for this creative work. (legacy spelling; see singular form, award)
<code>comment</code>	UserComments	Comments, typically from users, on this CreativeWork.
<code>contentLocation</code>	Place	The location of the content.
<code>contentRating</code>	Text	Official rating of a piece of content—for example, 'MPAA PG-13'.
<code>contributor</code>	Person or Organization	A secondary contributor to the CreativeWork.
<code>copyrightHolder</code>	Person or Organization	The party holding the legal copyright to the CreativeWork.

**Movie - schema.org - Mozilla Firefox**

<code>copyrightYear</code>	Number	The year during which the claimed copyright for the CreativeWork was first asserted.
<code>creator</code>	Person or Organization	The creator/author of this CreativeWork or UserComments. This is the same as the Author property for CreativeWork.
<code>dateCreated</code>	Date	The date on which the CreativeWork was created.
<code>dateModified</code>	Date	The date on which the CreativeWork was most recently modified.
<code>datePublished</code>	Date	Date of first broadcast/publication.
<code>discussionUrl</code>	URL	A link to the page containing the comments of the CreativeWork.
<code>editor</code>	Person	Specifies the Person who edited the CreativeWork.
<code>encoding</code>	MediaObject	A media object that encode this CreativeWork.
<code>encodings</code>	MediaObject	The media objects that encode this creative work (legacy spelling; see singular form, encoding).
<code>genre</code>	Text	Genre of the creative work
<code>headline</code>	Text	Headline of the article
<code>inLanguage</code>	Text	The language of the content. please use one of the language codes from the <a href="#">IETF BCP 47 standard</a> .
<code>interactionCount</code>	Text	A count of a specific user interactions with this item—for example, 20 UserLikes, 5 UserComments, or 300 UserDownloads. The user interaction type should be one of the sub types of <a href="#">UserInteraction</a> .
<code>isFamilyFriendly</code>	Boolean	Indicates whether this content is family friendly.
<code>keywords</code>	Text	The keywords/tags used to describe this content.
<code>mentions</code>	Thing	Indicates that the CreativeWork contains a reference to, but is not necessarily about a concept.
<code>offers</code>	Offer	An offer to sell this item—for example, an offer to sell a product, the DVD of a movie, or tickets to an event.
<code>provider</code>	Person or Organization	Specifies the Person or Organization that distributed the CreativeWork.
<code>publisher</code>	Organization	The publisher of the creative work.
<code>publishingPrinciples</code>	URL	Link to page describing the editorial principles of the organization primarily responsible for the creation of the

# MicroData Markup for “Pirates of the Caribbean”

```
<div itemscope itemtype="http://schema.org/Movie">
<h1 itemprop="name">Pirates of the Caribbean: On Stranger Tides (2011)</h1>
Jack Sparrow and Barbossa embark on a quest to find the elusive fountain of
youth, only to discover that Blackbeard and his daughter are after it too.
Director: <div itemprop="director" itemscope itemtype="http://schema.org/Person">
 Rob Marshall </div>
Writers:
<div itemprop="author" itemscope itemtype="http://schema.org/Person">
 Ted Elliott </div>
<div itemprop="author" itemscope itemtype="http://schema.org/Person">
 Terry Rossio </div> , and 7 more credits
Stars:
<div itemprop="actor" itemscope itemtype="http://schema.org/Person">
 Johnny Depp, </div>
<div itemprop="actor" itemscope itemtype="http://schema.org/Person">
 Penelope Cruz, </div>
<div itemprop="actor" itemscope itemtype="http://schema.org/Person">
 Ian McShane </div>
<div itemprop="aggregateRating" itemscope itemtype="http://schema.org/AggregateRating">
 8/10 stars from
 200 users.
Reviews: 50. </div> </div>
```

Includes  
Movie name  
Description  
Director  
Author  
Actors  
rating

## More Examples: Clarifying Hard to Understand Content

- **The <time> element has attributes: dates, times and durations:**

- `<time datetime="2022-04-01">04/01/20</time>`
- `<time datetime=2022-05-08T19:30">May 8, 7:30pm</time>`
- `<time itemprop="cookTime" datetime=PT1H30M">1 ½ hrs</time>`

- **Here is markup for a concert on May 8, 2022**

```
<div itemscope itemtype="http://schema.org/Event">
 <div itemprop="name">Spinal Tap</div>
 One of the loudest bands ever reunites for an unforgettable
 two-day show.
```

Event date:

```
<time itemprop="startDate"
 datetime="2022-05-08T19:30">May 8, 7:30pm</time>
</div>
```

- **Here is markup for an enumeration**

```
<div itemscope itemtype="http://schema.org/Offer">
 Blend-O-Matic
 $19.95
 Available today!</div>
```

This screenshot shows the left panel of the MERKLE Schema Generator. It features a sidebar with various SEO tools: robots.txt Tester, Sitemap Generator, Fetch & Render, Pre-rendering Tester, Mobile SEO, International SEO, Schema Generator, SERP Simulator, Local Search Tool, and Docs. The main area displays a "Schema Markup Generator (JSON-LD)" interface with a dropdown menu asking "Which Schema.org markup would you like to create?". Below this is a preview section showing "208 COMMENTS".

This screenshot shows the "Event" schema generator interface. It includes fields for Name, Event's description, and Image URL. A JSON-LD code snippet is displayed: 

```
<script type="application/ld+json">
{
 "@context": "https://schema.org",
 "@type": "Event",
 "name": "",
 "startDate": ""
}
</script>
```

 There are also dropdowns for Start date, End date, Start time (e...), End time (e...), Event Status, Attendance Mode, Performer (@type), and Performer's name. At the bottom, there are buttons for "+ ADD TICKET T" and "Current...".

For checking your site see  
<https://search.google.com/test/rich-results>

input

a web interface tool for  
creating a rich snippet

The screenshot shows the Google Rich Results Test interface. At the top, it asks "Does your page support rich results?". Below that, there's a URL input field containing "www.usc.edu". Underneath the URL, there's a dropdown menu set to "Googlebot smartphone" and a yellow "TEST URL" button. The main area displays the URL "http://www.usc.edu/" followed by a "Test results" section. This section indicates "No loading issues" and shows a green checkmark icon.

## Test results

This is a detailed view of the test results for www.usc.edu. It shows the URL "http://www.usc.edu/" and a "Test results" section. The section states "No loading issues" and includes a green checkmark icon. Below this, it says "Tested on: Apr 1, 2021 at 11:04 AM" and "Page is eligible for rich results". A note below says "All structured data on the page can generate rich results." and a "VIEW RENDERED HTML" link is provided.

## Detected items

Sitelinks searchbox	
<input checked="" type="checkbox"/>	Unnamed item
type	WebSite
id	<a href="https://www.usc.edu/#website">https://www.usc.edu/#website</a>
url	<a href="https://www.usc.edu/">https://www.usc.edu/</a>
name	University of Southern California

Results for www.usc.edu

This screenshot shows the Google Rich Results Test interface for the URL "http://www.cs.usc.edu". The top bar shows the URL and a "Test results" section indicating "No loading issues". Below this, it says "Tested on: Apr 1, 2021 at 11:07 AM" and "Page is eligible for rich results". A green checkmark icon is present. The "Detected items" section shows a breadcrumb structure:

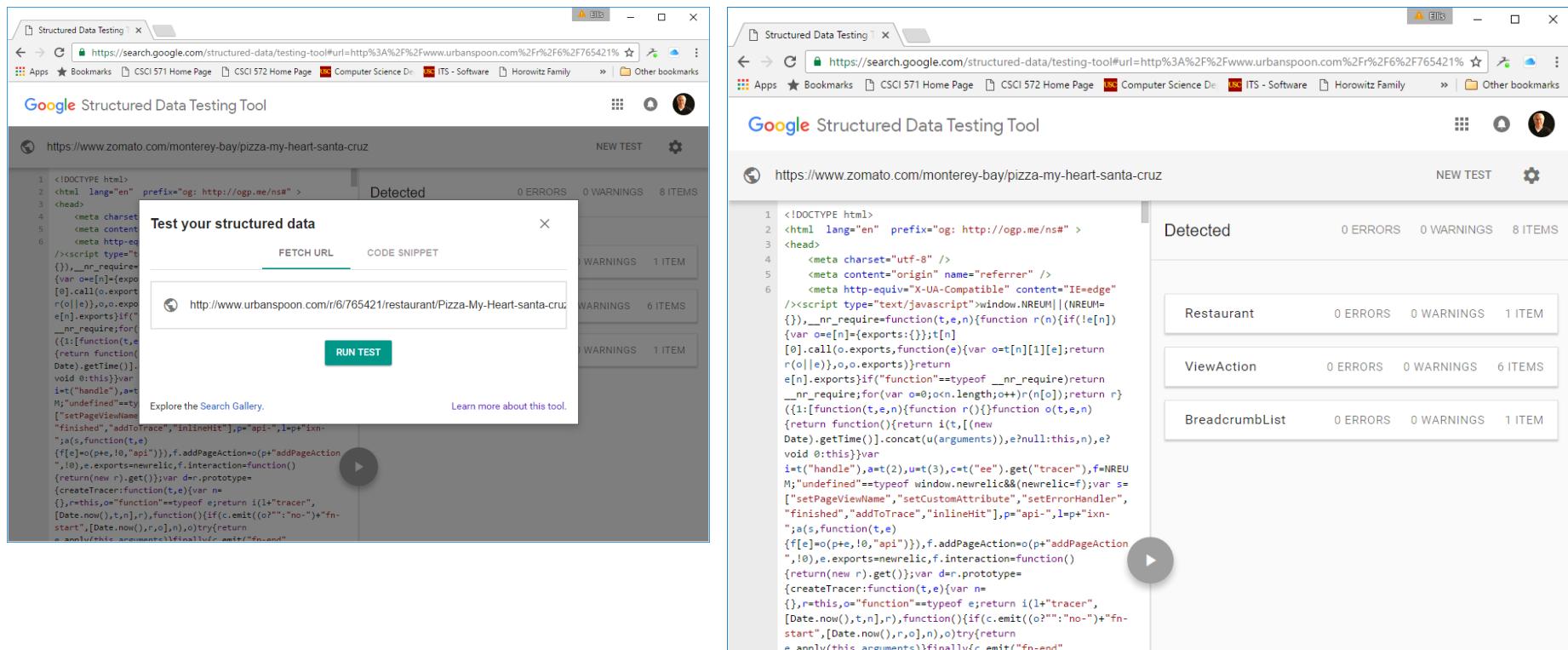
Breadcrumbs	
<input checked="" type="checkbox"/>	Unnamed item
type	BreadcrumbList
id	<a href="https://www.cs.usc.edu/#breadcrumb">https://www.cs.usc.edu/#breadcrumb</a>
itemListElement	
type	ListItem

An arrow points from the "List Item" entry in the breadcrumb table to the "ListItem" entry in the breadcrumb table above it.

Results for www.cs.usc.edu



# Google's Structured Testing Tool



<https://search.google.com/structured-data/testing-tool>

Google has created a tool for examining a web site with microformat data and indicating if there are any errors, e.g.

<http://www.urbanspoon.com/r/6/765421/restaurant/Pizza-My-Heart-santa-cruz>

# Google's Rich Snippets Testing Tool

<https://developers.google.com/structured-data/testing-tool/>

The screenshot shows the Google Structured Data Testing Tool interface. On the left, there is a code editor window containing the following HTML code:

```
1 <!-- Add this code to person page: -->
2 <div itemscope itemprop="author"
3 itemtype="http://schema.org/Person">
4 <meta itemprop="description" content="Dr. Ellis Horowitz is
5 currently Professor of Computer Science and Electrical Engineering
6 at the University of Southern California.
7 The company designed and developed UNIX application
8 software.">
9 Ellis
10 Horowitz
11 <span itemprop="member" itemscope
12 itemtype="http://schema.org/Organization"
13 style="display:block;">University of Southern California
14 <span itemprop="jobTitle"
15 style="display:block;">Professor
16 <span itemprop="email"
17 style="display:block;">ehorowitz1@gmail.com
18 <div itemprop="address" itemscope
19 itemtype="http://schema.org/PostalAddress">90089
21 </div>
```

Below the code editor is a red error icon with the number "1". To the right of the code editor is a "VALIDATE" button.

The main area on the right is titled "Results - Filter by use case". It shows a single result for a "Person" with a count of "(1)". There is one error indicated by a red exclamation mark. The result details are as follows:

Person
<b>description:</b> Dr. Ellis Horowitz is currently Professor of Computer Science and Electrical Engineering at the University of Southern California. The company designed and developed UNIX application software.
<b>name:</b> Ellis Horowitz
<b>email:</b> ehorowitz1@gmail.com
<b>jobTitle:</b> Professor
<b>member [Organization]:</b> University of Southern California
<b>address [PostalAddress]:</b>
<b>postalCode:</b> 90089

Below the results, there is a "Custom Search Result Filters" section.

At the bottom of the page, there are navigation links: Connect, Programs, Developer Consoles, and Explore. The USC logo is also present at the bottom left.

- The robots meta tag is added to an HTML page's <head>; here are some new tags:
  - "nosnippet"  
This is an existing option to specify that you don't want any textual snippet shown for this page.
  - "max-snippet:[number]"  
New! Specify a maximum text-length, in characters, of a snippet for your page.
  - "max-video-preview:[number]"  
New! Specify a maximum duration in seconds of an animated video preview.
  - "max-image-preview:[setting]"  
New! Specify a maximum size of image preview to be shown for images on this page, using either "none", "standard", or "large".
- They can be combined, for example:

```
<meta name="robots" content="max-snippet:50, max-image-preview:large">
```

- A new way to help limit which part of a page is eligible to be shown as a snippet is the "data-nosnippet" HTML attribute on span, div, and section elements.
  - With this, you can prevent that part of an HTML page from being shown within the textual snippet on the page.
- For example:
- `<p><span data-nosnippet>Harry Houdini</span> is undoubtedly the most famous magician ever to live.</p>`
- **To opt out of featured snippets**
- The [nosnippet tag](#) blocks all snippets (featured snippets and regular snippets) for the tagged page.
- Text marked by the [data-nosnippet tag](#) won't appear in featured snippets (or regular snippets either).
- If both nosnippet and data-nosnippet appear in a page, nosnippet takes priority, and snippets won't be shown for the page.

- **Snippets can be divided into five categories**

1. **Regular snippets**, displayed in organic search results
2. **Rich snippets** come from structured data dictionary schema.org including RDFa, Microdata or JSON
3. **Google News**, created automatically from news feeds to Google
4. **Entity types**, come from the KnowledgeGraph, are constructed object and concepts including people, movies, places, events, books, etc
5. **Features snippets**, determine that a page contains a likely answer to the user's question; the snippet is displayed. In four different forms: paragraph, table, ordered list, unordered list

- **QAPage** focuses on a specific question and its answer(s)
- **<https://schema.org/QAPage>**
- **Question**, a specific question from a user seeking answers online or collected in a FAQ document
- **<https://schema.org/QAPage>**
- **HowTo**, instructions that explain how to achieve a results by performing a sequence of steps
- **<https://schema.org/HowTo>**
- **Here is an article on infinite PAAs, <https://moz.com/blog/infinite-people-also-ask-boxes>**
- **Matt Cutts Discusses Snippets**
  - **<https://www.youtube.com/watch?v=vS1Mw1Adrk0>**
  - **<https://www.youtube.com/watch?v=NlJiLDn9-38>**
- Three useful webpages on rich snippets:

<http://support.google.com/webmasters/bin/answer.py?hl=en&answer=99170&topic=21997&ctx=topic>

<http://support.google.com/webmasters/bin/answer.py?hl=en&answer=1093493>

- **Web Duplication**
- **<https://www.youtube.com/watch?v=LsfIlviyDvE>**
- **Rendering**
- **<https://www.youtube.com/watch?v=rq8sFkl0KnI>**
- **Googlebot and Web Hosting (9 min)**
- **<https://www.youtube.com/watch?v=JvYh1oe5Zx0>**
- **Titles, Snippets and Result Previews**
- **<https://youtu.be/ezL07yC4aFo>**

# Clustering



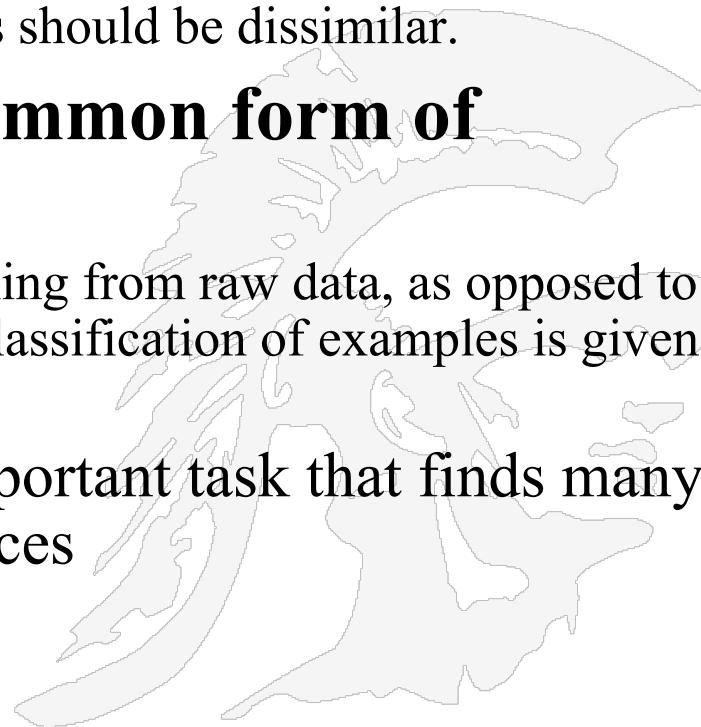
# Today's Topic: Clustering

- Document clustering
  - Motivations
  - Document representations
  - Success criteria
- Clustering algorithms
  - Partitional
  - Hierarchical



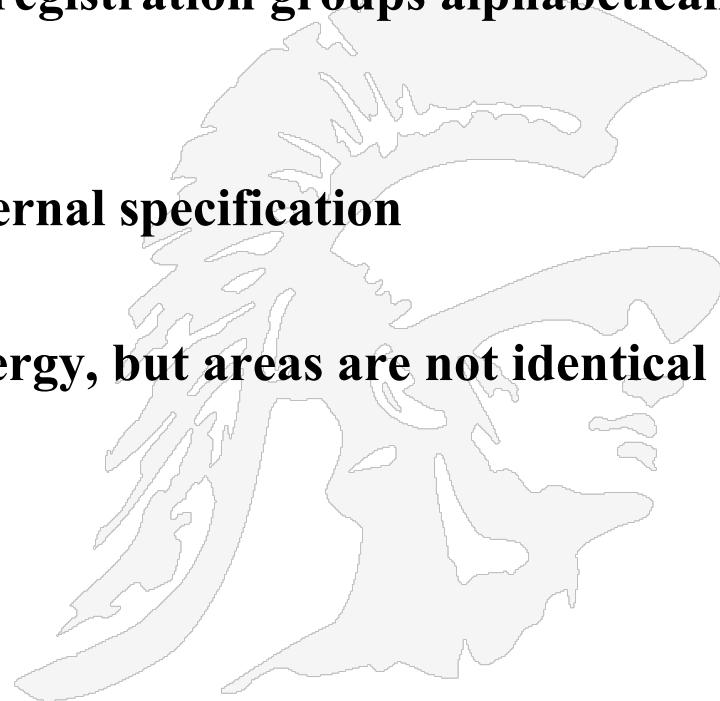
# What is Clustering?

- **Clustering: the process of grouping a set of objects into classes of similar objects**
  - Documents within a cluster should be similar.
  - Documents from different clusters should be dissimilar.
- **Clustering is the most common form of *unsupervised learning***
  - Unsupervised learning = learning from raw data, as opposed to supervised learning where a classification of examples is given *a priori*
- Clustering is a common and important task that finds many applications in IR and other places



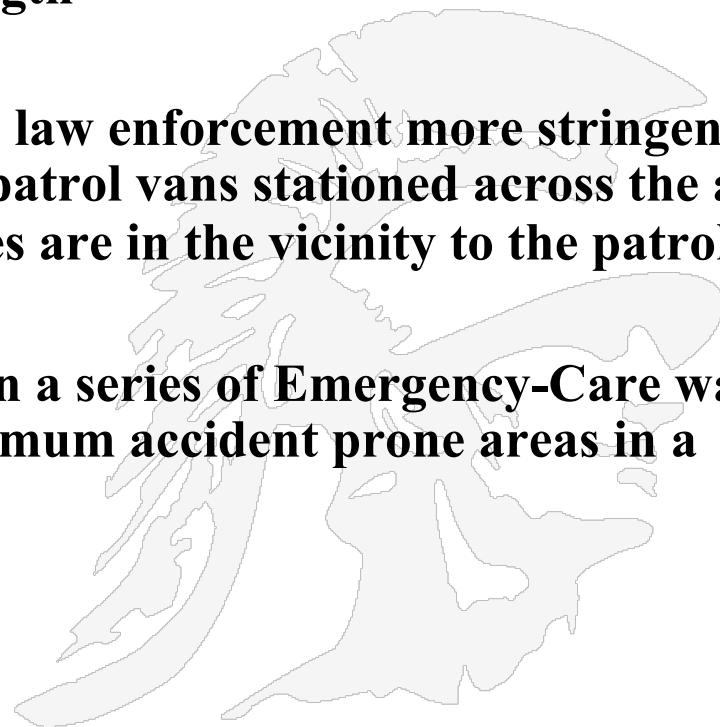
# What is NOT Clustering

- **Supervised classification**
  - Have class label information
- **Simple segmentation**
  - Dividing students into different registration groups alphabetically, by last name
- **Results of a query**
  - Groupings are a result of an external specification
- **Graph partitioning**
  - Some mutual relevance and synergy, but areas are not identical



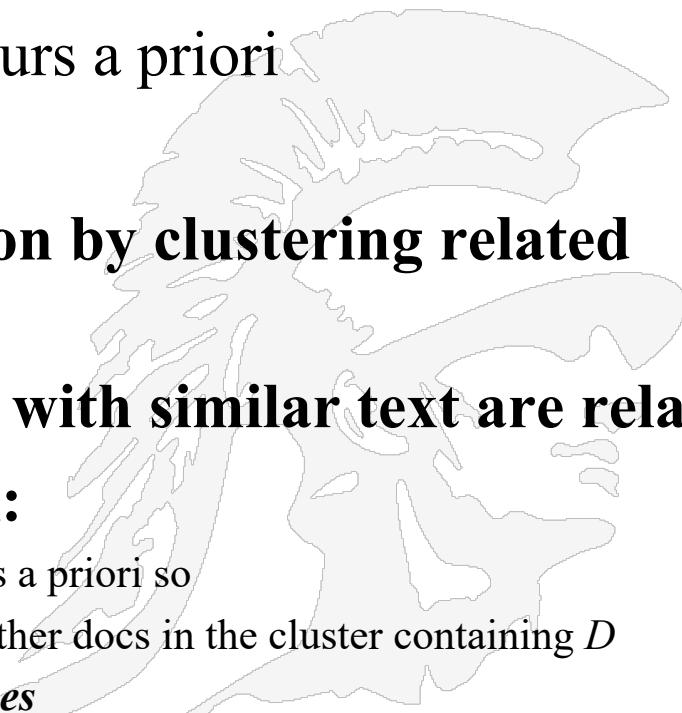
# Applications of Clustering 3 Scenarios

1. A telephone company needs to establish its network by putting its towers in a particular region it has acquired. The location of putting these towers can be found by using a clustering algorithm so that all its users receive optimum signal strength
  
2. The Miami DEA wants to make its law enforcement more stringent and hence have decided to make their patrol vans stationed across the area so that the areas of high crime rates are in the vicinity to the patrol vans
  
3. A hospital care chain wants to open a series of Emergency-Care wards, keeping in mind the factor of maximum accident prone areas in a region



# Why Search Engines Cluster Documents

1. **For improving recall in search applications**
  - Better search results; similar documents are grouped
2. **For speeding up vector space retrieval**
  - Faster search if clustering occurs *a priori*
3. **Cleaner user interface**
4. **Automatic thesaurus generation by clustering related terms**
  - ***Cluster hypothesis* - Documents with similar text are related**
  - **Ergo, to improve search recall:**
    - In theory we could cluster docs in our corpus *a priori* so when a query matches a doc  $D$ , we also return other docs in the cluster containing  $D$
    - ***This strategy doesn't work for search engines***





https://www.google.com/#q=cars  
 Apps CSCI 572 Home Page CSCI 571 Home Page CSCI 351 Home Page Ellis Horowitz' Home Computer Sci

### How GM Beat Tesla to the First True Mass-Market Electric Car

[www.wired.com/2016/01/gm-electric-car-chevy-bolt-mary-barra/](http://www.wired.com/2016/01/gm-electric-car-chevy-bolt-mary-barra/)  
 In short, the electric car business has taken the form of an old-fashioned race for a prize—a race in very soft sand. There's no Moore's law for batteries, which are ...

### The Dream Life of Driverless Cars - The New York Times

[www.nytimes.com/2015/11/15/.../the-dream-life-of-driverless-cars.html](http://www.nytimes.com/2015/11/15/.../the-dream-life-of-driverless-cars.html)  
 What they hoped to scan was not just the shape of the city streets but the inner life of the autonomous cars that may soon come to dominate ...

### Hidden Obstacles for Google's Self-Driving Cars

[https://www.technologyreview.com/.../hidden-obstacles-for-googles-self-dri...](http://www.technologyreview.com/.../hidden-obstacles-for-googles-self-dri...)  
 Would you buy a self-driving car that couldn't drive itself in 99 percent of the country? Or that knew nearly nothing about parking, couldn't be ...

### New & Used Car Reviews & Ratings - Consumer Reports

[www.consumerreports.org/cro/cars/index.htm](http://www.consumerreports.org/cro/cars/index.htm) ▾ Consumer Reports ▾  
 Provides car reviews, automobile safety information, car buying guidance.

#### Searches related to cars

autotrader carmax  
 cars for sale cars 2 full movie  
 used cars cars 2  
 cars 2006 cars for sale by owner

Goooooooooooooogle >  
 1 2 3 4 5 6 7 8 9 10 Next

bbs - Bing https://www.bing.com/search?q=cars&go=Submit&qs=n&form=QBLH&pq=cars&sc=9-4&sp=-1&sk=&cvid=a4703723

Images of cars  
[bing.com/images](#)

AutoTrader.com - Official Site  
[www.autotrader.com](#)  
 Find used cars and new cars for sale at Autotrader. With millions of cars, finding your next new car or used car and the car reviews and information you're looking ...

Local results for cars near los angeles california 90272 u...  
 Bing Local

Cars With Class  
[carsclassic.com](#)  
 ★★★★ 5 Yelp reviews

Certified Cars  
[www.certifiedcars.com](#)

Major Motor Cars Inc  
[www.majormotors.com](#)  
 ★★★★ 62 Yelp reviews

1 1115 Wilshire Blvd, Santa Monica, CA 90401 (310) 656-3444  
 2 1011 Swardmore Ave 2, Los Angeles, CA 90272 (888) 304-1622  
 3 2935 Santa Monica Blvd, Santa Monica, CA 90404 (310) 829-1100

Y cars - Yahoo Search Results https://search.yahoo.com/search;\_ylt=A86.ItHuazFVgY0AhI6bvZx4?p=cars&toggel=1&cop=mss&ei=UTF-8&fr=yfp

Home Mail Search News Sports Finance Weather Games Answers Screen Flickr Mobile More Sign In Mail

YAHOO! cars Search See more ads for: cars all\_electric\_cars used\_cars\_sale hybrid\_cars rental\_cars

Ads related to cars

**Cars.com™ Official Site**  
[www.Cars.com](#) Ad  
 Search 4.1 Million Listings and Find Your Used Car at Cars.com™!  
 Cars.com: New or Used Listings, Reviews, Advice, Service Info

**Under \$10,000**  
 Looking for a Used Car under \$10k?  
 Find a Great Deal at Cars.com Today

**Under \$5,000**  
 Find & Compare Used Car Inventory  
 Under \$5,000 at Cars.com Now.

**Under \$3,000**  
 Limited Budget? Find Affordable  
 Used Cars Around You at Cars.com!

**Under \$20,000**  
 Find an Incredible Vehicle for  
 Under \$20,000 By Shopping Online!

**Official Mazda USA Site**  
[www.MazdaUSA.com](#) Ad  
 See the entire lineup of new Mazda cars. Search Mazda Dealer Inventory

**Car pricing info - Wondering what to pay for a new car.**  
[truecar.com/incentives](#) Ad  
 4.5 ★★★★☆ rating for truecar.com  
 Wondering what to pay for a new car. See what others paid with TrueCar.  
 Brands: Acura, Alfa Romeo, Aston Martin, Audi, Bentley, Buick and more

**Cars**  
[www.Ford.com/Ford\\_Fusion](#)  
 Discover the Smart & Efficient Performance of the 2015 Ford Fusion

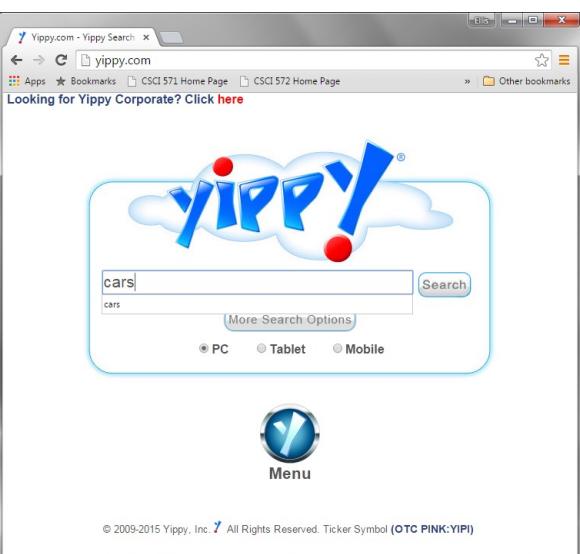
**Cars | ebayclassifieds.com**  
[www.ebayclassifieds.com](#) Ad

Google related searches  
 Yahoo does some clustering via  
 alternate queries

Bing does a little better

# yippy.com Search Engine

- Yippy (formerly Clusty) is a metasearch engine developed by Vivísimo which emphasizes clusters of results.



initial screen  
with query "cars"

This screenshot shows the search results for 'cars'. On the left, a sidebar lists various clusters: 'All Results (548)', 'Sale (51)', 'Reviews (35)', 'Attack (38)', 'Dealer (48)', 'Photographs (63)', 'Prices (23)', 'Car crash (24)', 'Rental (28)', 'Bomb, Killed (21)', 'Classic (27)', 'Trucks, SUV (18)', 'Car Club (27)', 'Slide Show (11)', 'Surprise, Accident (11)', 'Ride (12)', 'Caused (11)', 'Automakers (11)', 'Fire (12)', 'Insurance (10)', 'Police car (8)', and 'Self-driving (10)'. The main content area displays search results for cars, including links to Autotrader, Edmunds, eBay, and Discovery Channel.

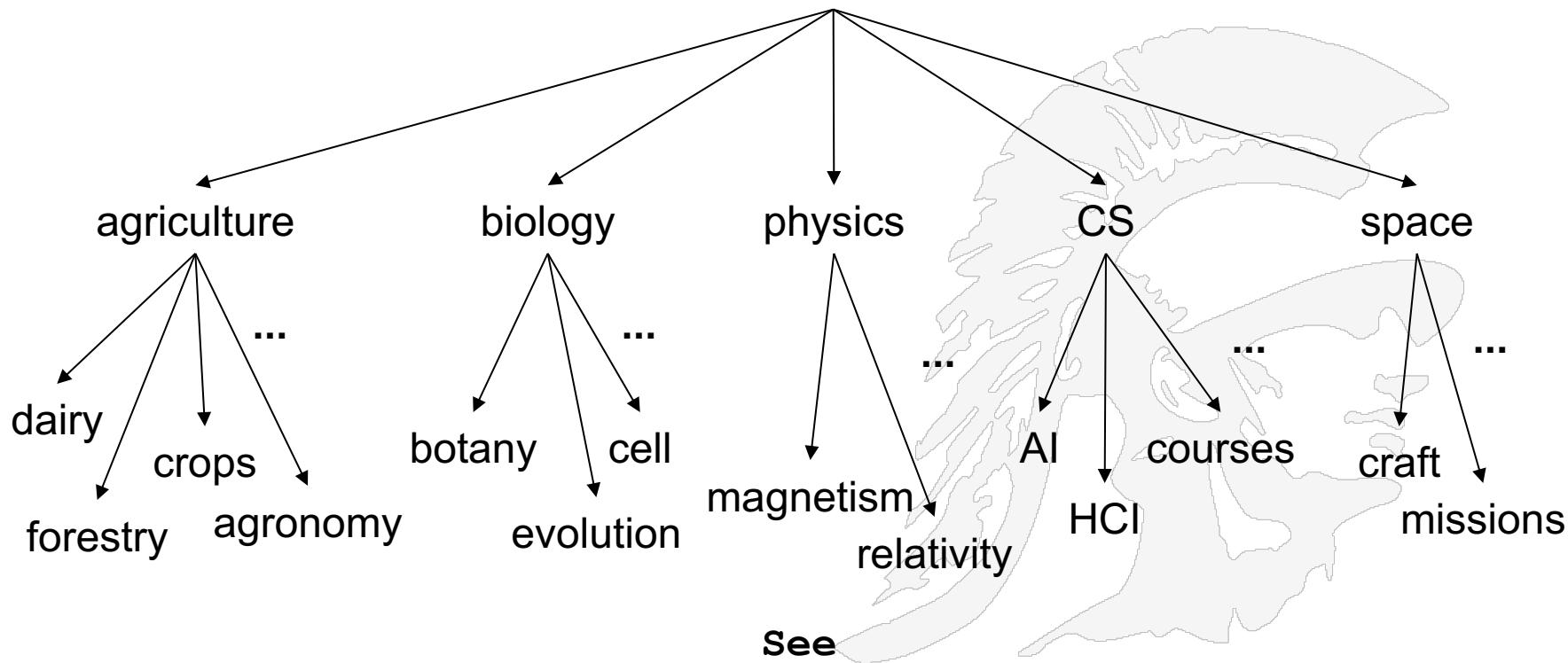
clustered results appear  
on the left column: e.g.  
sale  
reviews  
dealers  
rentals

This screenshot shows a deeper level of clustering. The sidebar now includes 'Clouds', 'Sources', 'Sites', and 'Time' buttons. The 'All Results (548)' section has expanded into more specific categories: 'Sale (51)', 'Car Sales (9)', 'Reviews (5)', 'Buying And Selling (6)', 'Prices (5)', 'Car Dealers (5)', 'Trucks, SUV and other vehicles (4)', 'eBay, Electronics, Cars, Fashion, Collectibles, Coupons And More (2)', 'Ride-Sharing (2)', 'Certified Pre-Owned (2)', 'North America (2)', 'Canadian, Automakers (2)', 'Classic Muscle Cars (2)', 'Other Topics (17)', 'Reviews (35)', 'Attack (28)', 'Jerusalem (18)', 'Car Bomb (4)', 'Prosecutor, Egypt (2)', and 'Leader, Car attacked (2)'. The main content area shows news articles about car attacks and protests.

multiple level clusters:  
car dealers  
trucks  
ebay

## Yahoo's Name Derives from Yet Another *Hierarchical Officious Oracle*

**Yahoo! Hierarchy *isn't* clustering but *is* the kind of output you want from clustering – a taxonomy**



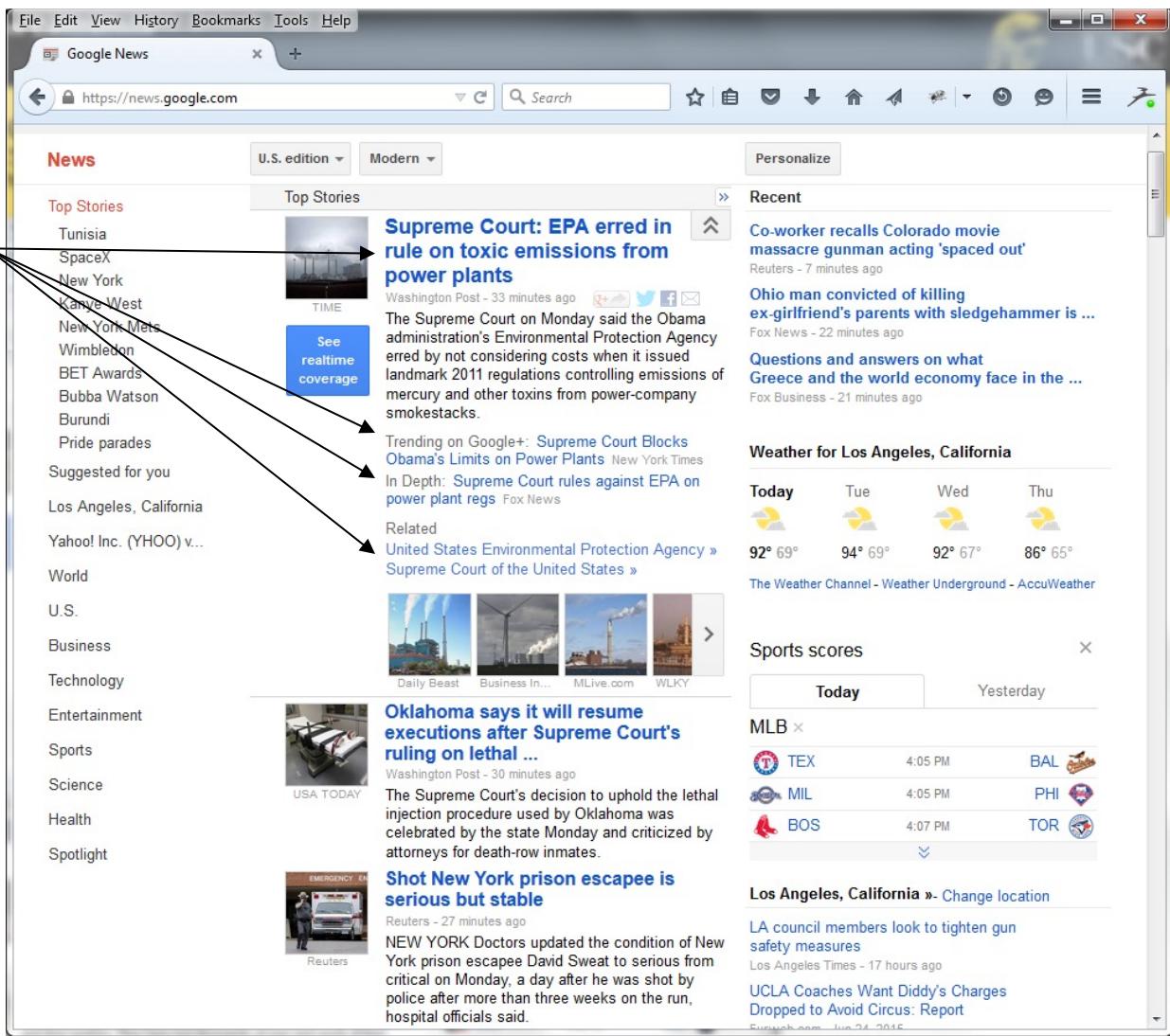
<https://searchengineland.com/yahoo-directory-close-204370>

# Google News: Automatic Clustering Gives an Effective News Presentation Metaphor

recent Supreme court decisions clustered together

Typical newspaper clusters:  
 World, US, Business,  
 Technology, Sports, etc

These clusters must be constantly re-computed to make sure the latest news is included



The screenshot shows the Google News interface. On the left, a sidebar lists 'Top Stories' including Tunisia, SpaceX, New York, Kanye West, New York Mets, Wimbledon, BET Awards, Bubba Watson, Burundi, and Pride parades. Below this is a 'Suggested for you' section for Los Angeles, California, and a 'World' section. A large bracket on the left groups these under the heading 'Typical newspaper clusters: World, US, Business, Technology, Sports, etc'. Another bracket on the right groups the 'Top Stories' and 'Suggested for you' sections under the heading 'recent Supreme court decisions clustered together'. The main content area displays a 'Top Stories' section with a large image of a power plant and the headline 'Supreme Court: EPA erred in rule on toxic emissions from power plants'. Below this are 'Trending on Google+', 'In Depth', 'Related' stories, and a weather forecast for Los Angeles. At the bottom, there's a 'Sports scores' section and a 'Los Angeles, California' news feed.

# Clustering Examples from Google RSS Feeds

Two examples of Google feeds  
There is a main article, some text  
and beneath that related or  
clustered articles

## Tesla's Model 3 Market Opportunity Is Bigger Than You Think

Motley Fool

Tesla's (NASDAQ:TSLA) forthcoming Model 3 will be unveiled next month, go into production in late 2017, cost about \$35,000 before incentives, and ...



### Tesla Signs Lease for 40K-SF Red Hook Dealership - Commercial Observer

Advertising enters the equation for Tesla Motors - Seeking Alpha

Tesla Motors Finally Gets Its Paws on Tesla.com - Inverse

Full Coverage



## There's one new Tesla car that nobody is talking about

Businessinsider India

These two **Teslas** are all anyone has been talking about lately - especially Wall Street analysts who want to figure out which way **Tesla's** extremely ...



### VIDEO: Tesla Drag Race! Model S vs. Model X In STUNNING Showdown - AutoSpies.com

Tesla Model S & Model X Comparison (Price, Range, Acceleration) After Removal Of 85 kWh Version - InsideEVs

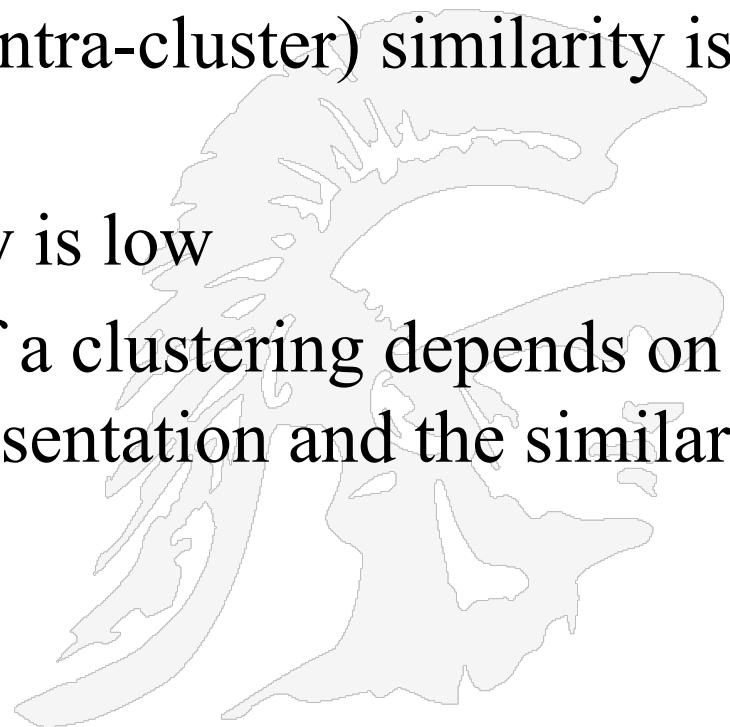
Tesla Will Begin Taking Preorders on Its Make-or-Break Vehicle - GreatNews

Full Coverage



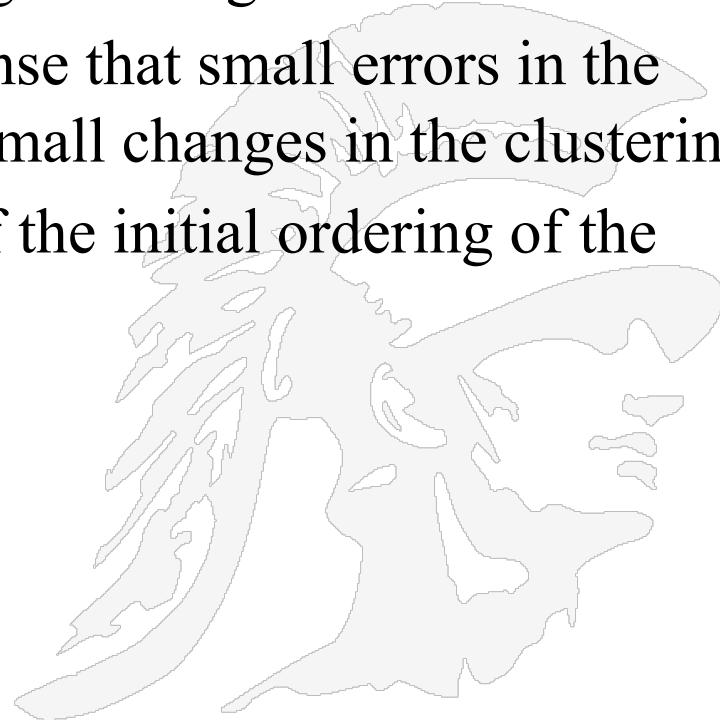
# What Is A Good Clustering?

- Internal criterion: A good clustering will produce high quality clusters in which:
  - the intra-class (that is, intra-cluster) similarity is high
  - the inter-class similarity is low
  - The measured quality of a clustering depends on both the document representation and the similarity measure used



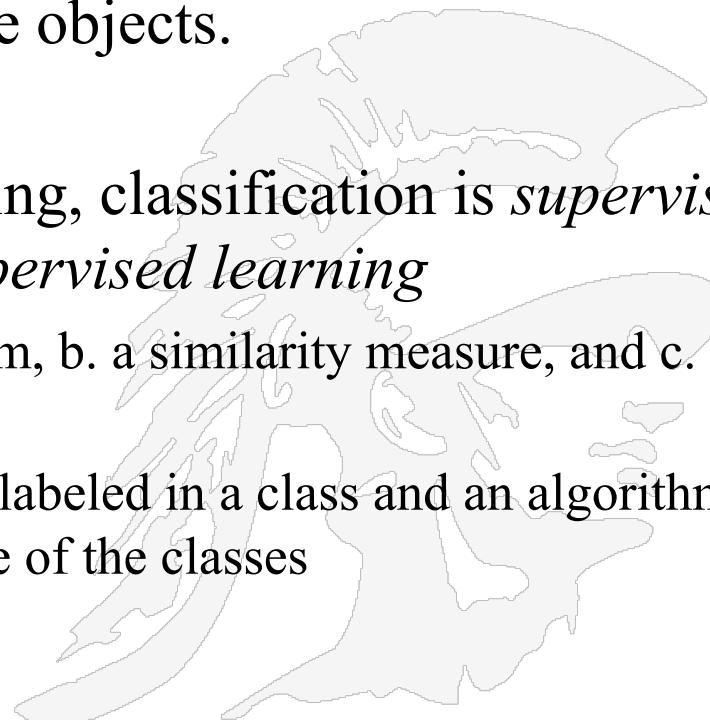
## Three Criteria of Adequacy for Clustering Methods

1. The method produces a clustering which is **unlikely to be altered drastically** when further objects are incorporated
  - i.e. it is stable even under significant growth
2. The method is **stable** in the sense that small errors in the description of objects lead to small changes in the clustering
3. The method is **independent** of the initial ordering of the objects



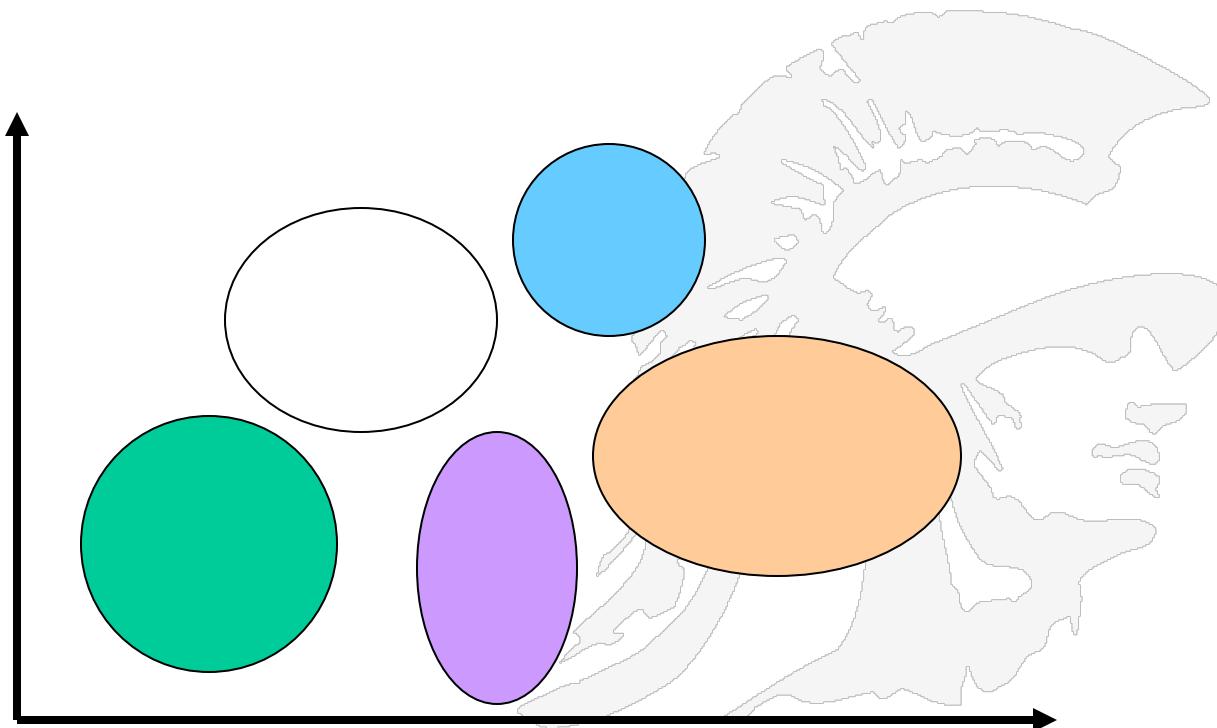
# Classification is Different from Clustering

- In general, in **classification** you have a set of predefined classes and want to know which class a new object belongs to.
- **Clustering** tries to group a set of objects and find whether there is *some* relationship between the objects.
  - Clustering *precedes* classification
- In the context of machine learning, classification is *supervised learning* and clustering is *unsupervised learning*
  - **Clustering** requires a. an algorithm, b. a similarity measure, and c. a number of clusters
  - **classification** has each document labeled in a class and an algorithm that assigns new documents to one of the classes



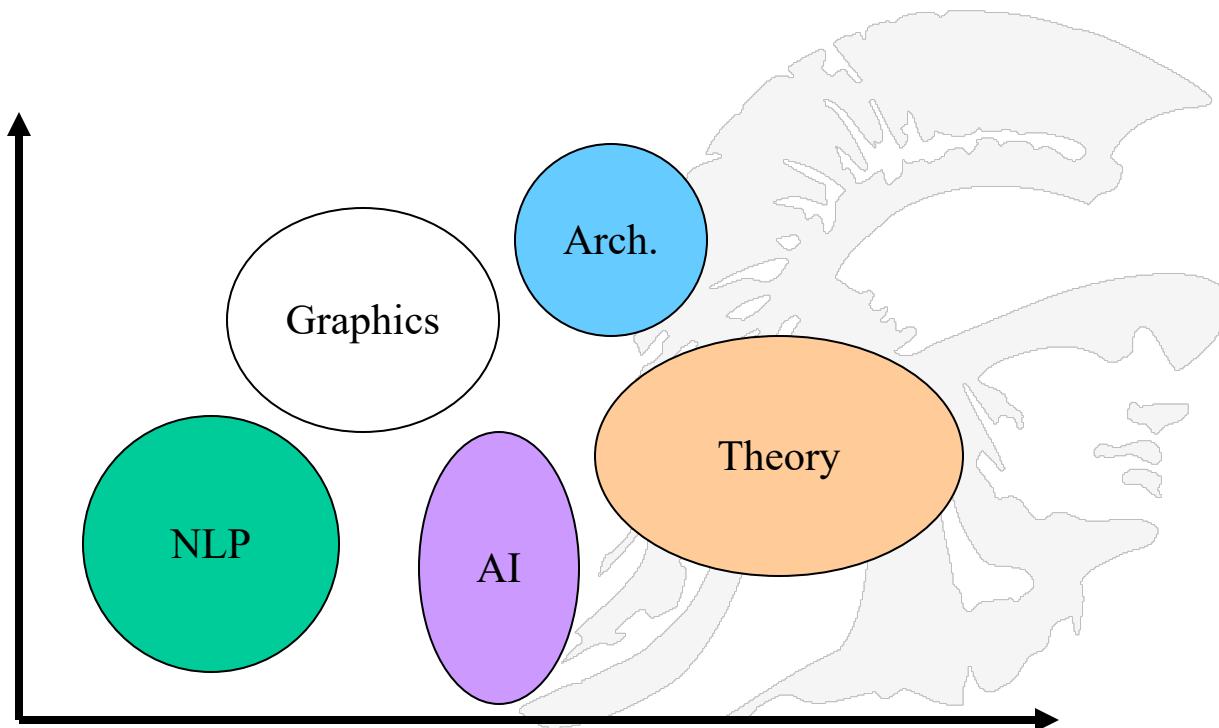
# Begin with Clustering

- Step 1: Given a large set of computer science documents, first we cluster them using some algorithm (to be presented)



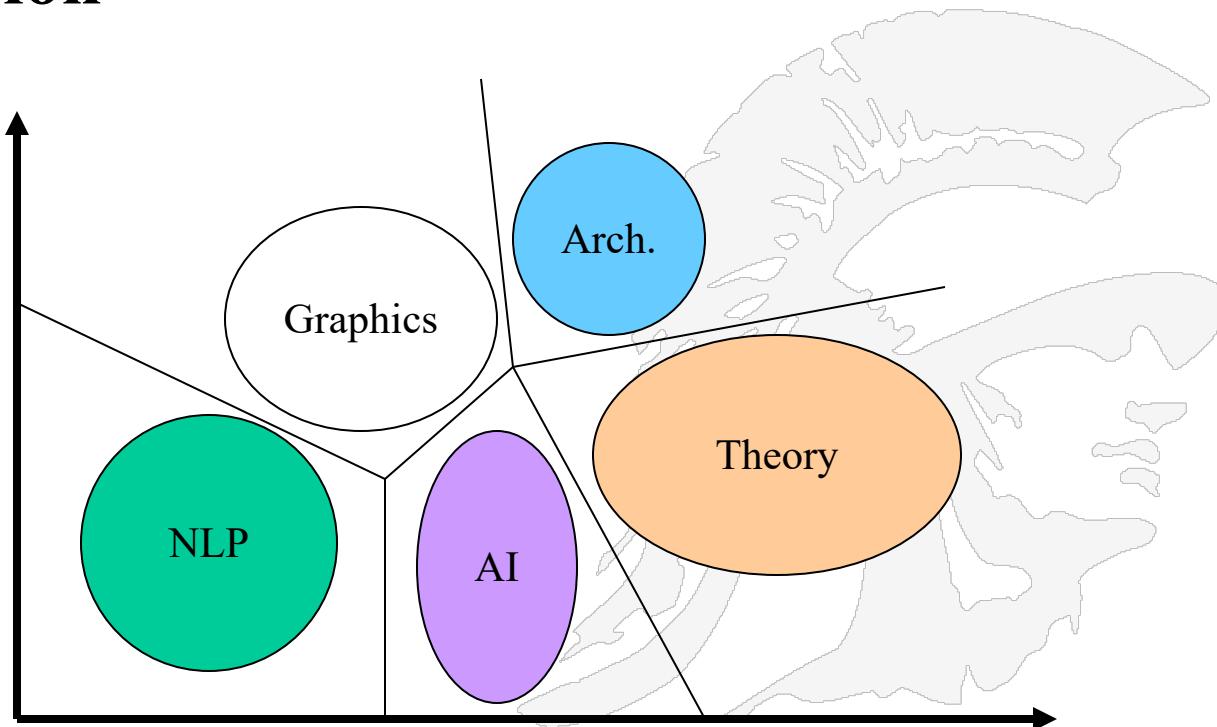
# Then We Name the Clusters

- Step 2: we label the clusters
  - choosing a popular name from each document cluster



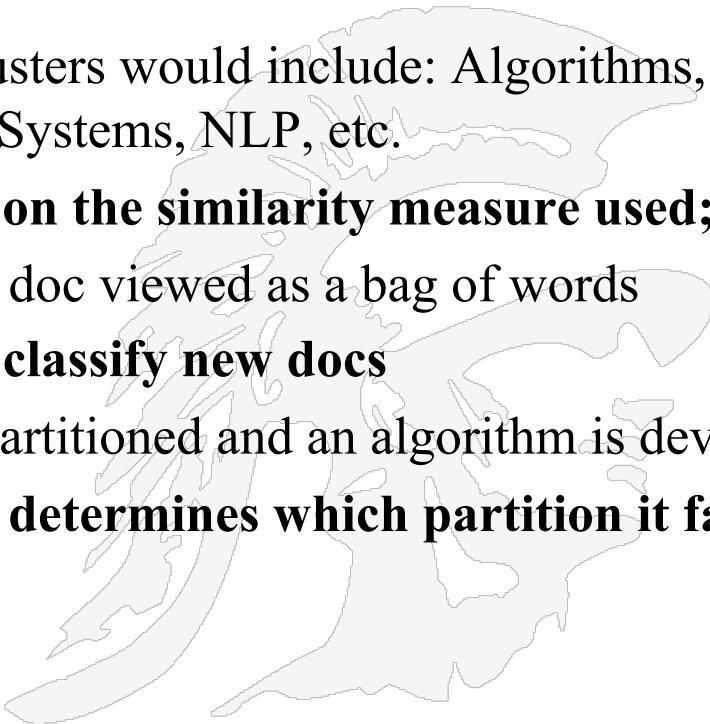
## Still Clustering: Determine Decision Boundaries

- Step 3: we compute boundaries for the clusters that can be used as new documents appear; i.e. classification



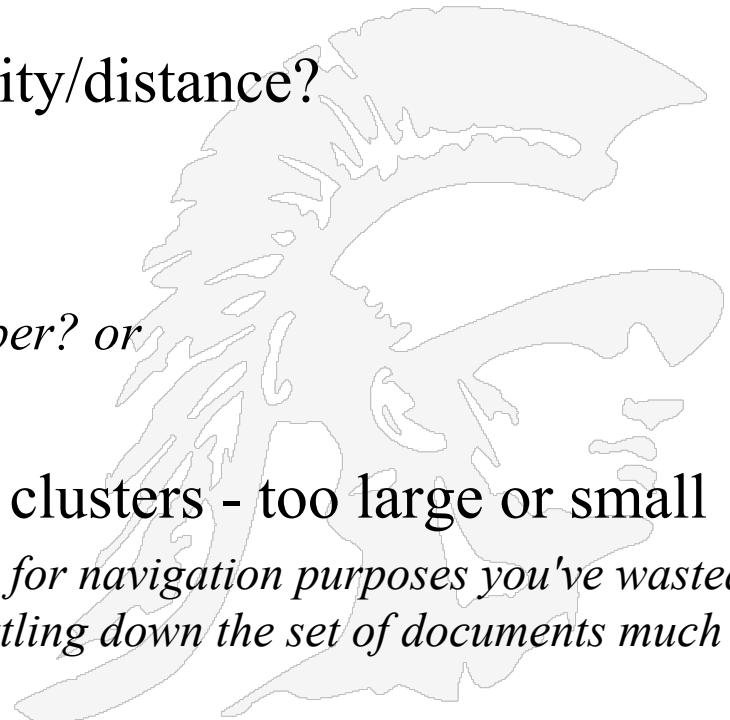
# Classification Requires Initial Clusters and Boundaries

- **Definition:** *Supervised Learning*, inferring a function from labeled training data
- 1. **The documents in each cluster define the “training” docs for each category**
  - E.g. in computer science named clusters would include: Algorithms, Theory, AI, Databases, Operating Systems, NLP, etc.
- 2. **Documents are in a cluster based upon the similarity measure used;**
  - generally a vector space with each doc viewed as a bag of words
- 3. **A classifier is an algorithm that will classify new docs**
  - Essentially, the decision space is partitioned and an algorithm is devised
- 4. **Given a new doc, the new algorithm determines which partition it falls into**



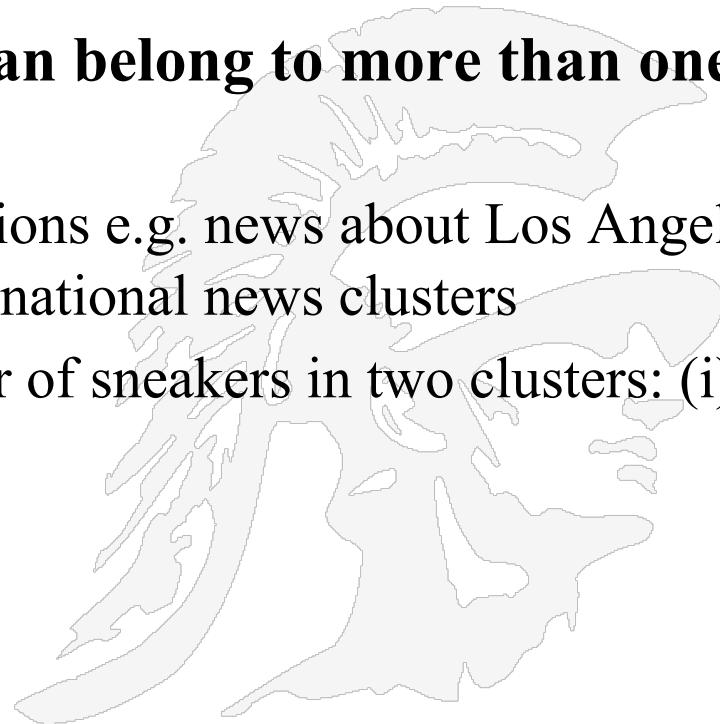
## Now Let's Return to the Earlier Problem: Clustering

- **Questions to consider when clustering**
  - How do we represent the document?
    - *Usually as a vector space*
  - How do we compute similarity/distance?
    - *Using cosine similarity*
  - How many clusters?
    - *will it be a fixed a priori number? or*
    - *completely data driven?*
  - Be careful to avoid “trivial” clusters - too large or small
    - *If a cluster is too large, then for navigation purposes you've wasted an extra user click without whittling down the set of documents much*



# Issue: Hard vs. Soft Clustering

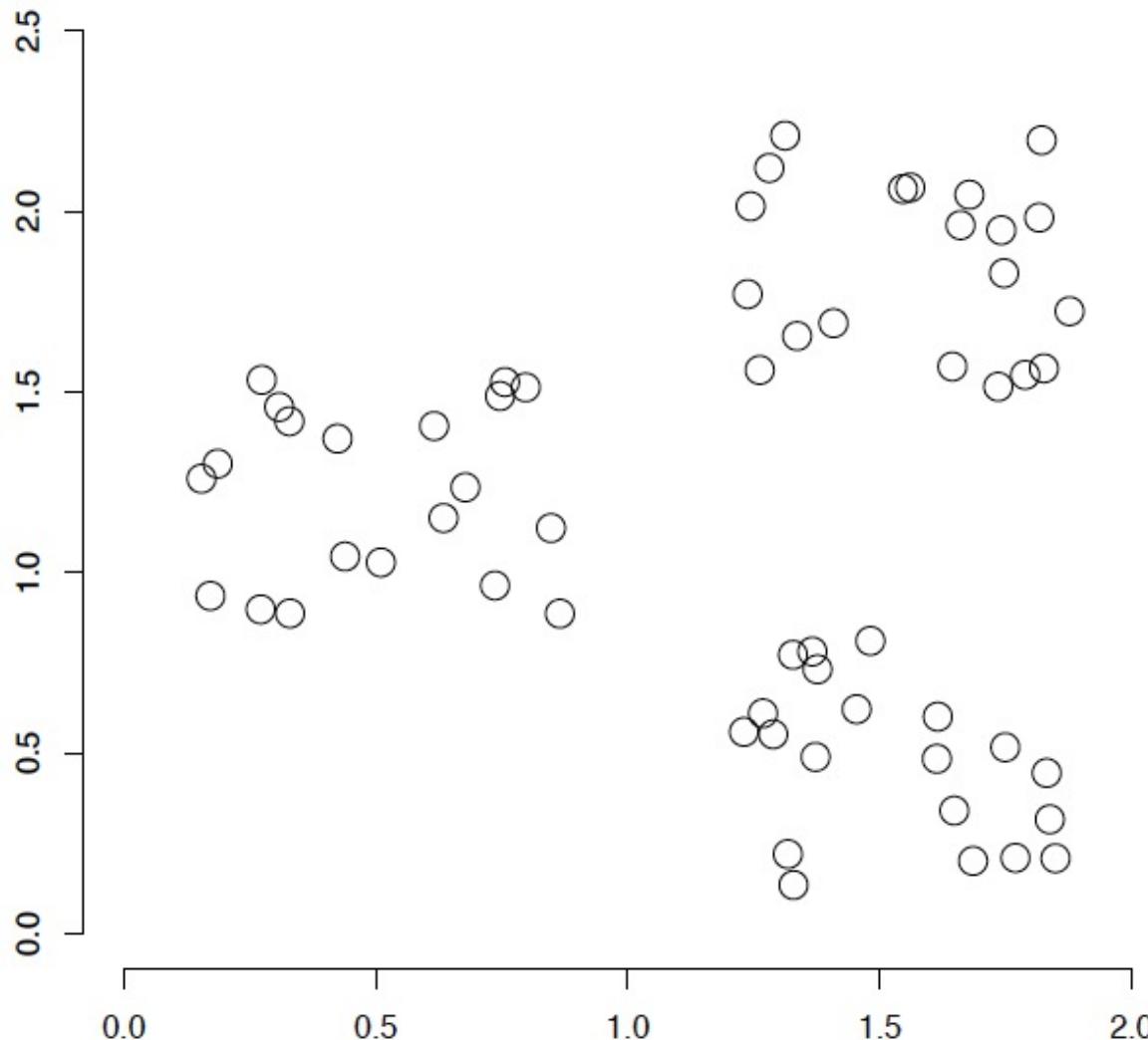
- ***Hard clustering:*** Each document belongs to exactly one cluster
  - More common and easier to do
- ***Soft clustering:*** A document can belong to more than one cluster.
  - Makes sense for some applications e.g. news about Los Angeles might be included in local and national news clusters
  - E.g. you may want to put a pair of sneakers in two clusters: (i) sports apparel and (ii) shoes



# What Definition of Similarity/Distance Will Be Used

- Once again we will treat documents as vectors
  - Cosine similarity (seen before many times)
    - Cosine similarity is a measure of similarity between two vectors of an inner product space that measures the cosine of the angle between them. Range from 0 (dissimilar) to 1 (exactly similar)
  - Most clustering implementations use cosine similarity
  - Euclidean distance is a close alternative that is also popular

# A Data Set with Clear Cluster Structure

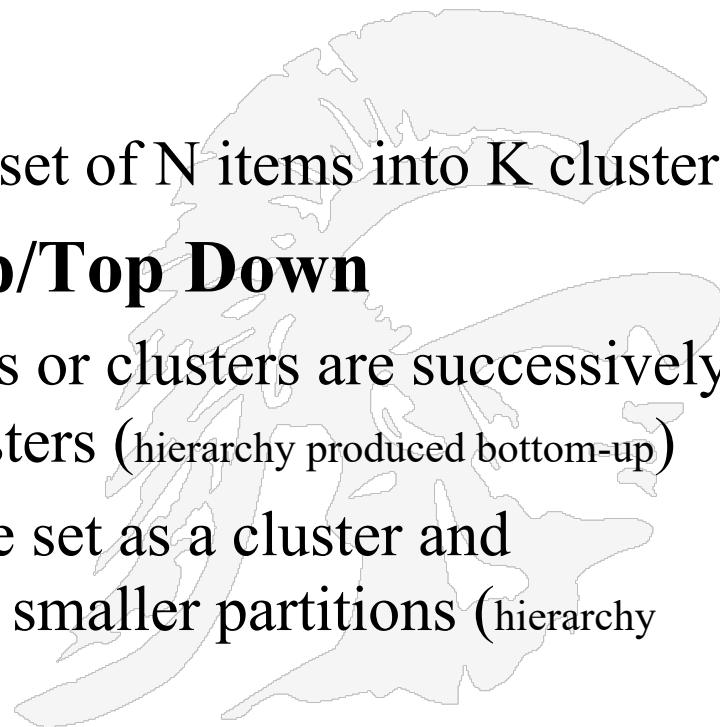


Circles represent documents as N-vectors

- How would you design an algorithm for finding the three clusters in this case?
- Hint: use a distance measure

# Clustering Algorithms

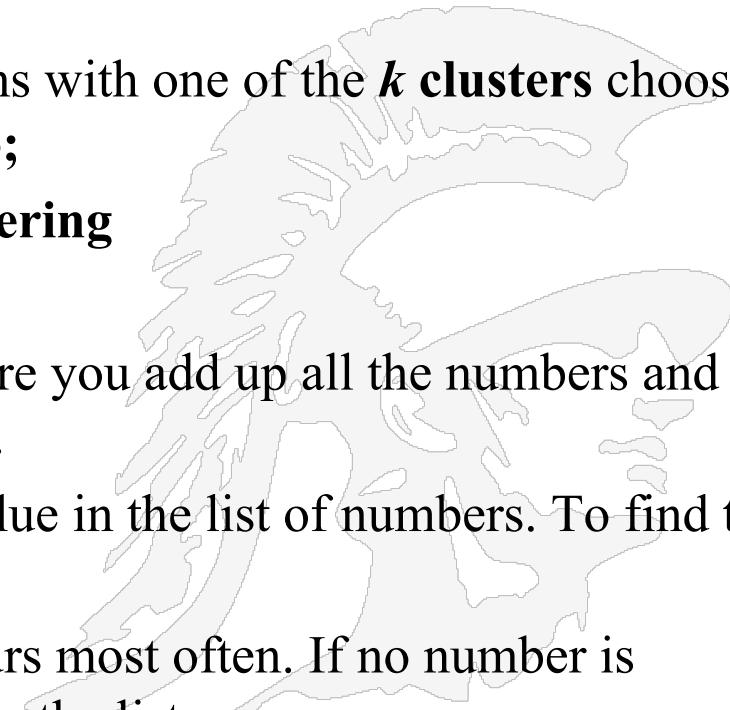
- **Two general methodologies**
  - Partitioning Based Algorithms
  - Hierarchical Algorithms
- **Partitioning Based**
  - Choose K and then divide a set of N items into K clusters
- **Hierarchical – Bottom Up/Top Down**
  - **agglomerative**: pairs of items or clusters are successively linked to produce larger clusters (hierarchy produced bottom-up)
  - **divisive**: start with the whole set as a cluster and successively divide sets into smaller partitions (hierarchy produced top-down)



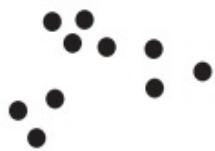
# A Partitioning Algorithm

## K-Means Clustering Algorithm

- **Clustering algorithm strategy**
  - Choose  $k$  random data items out of the  $n$  items; call these items the *means*; they designate the prototype or name of the cluster
  - **Refine it iteratively**
    - Associate each of the  $n-k$  items with one of the  **$k$  clusters** choosing the **cluster** that it is nearest to;
    - **This is called  $K$ -means clustering**
- **Recall**
  - The "**mean**" is the "average" where you add up all the numbers and then divide by the number of numbers.
  - The "**median**" is the "middle" value in the list of numbers. To find the median, you may have to sort
  - The "**mode**" is the value that occurs most often. If no number is repeated, then there is no mode for the list



# Different Ways of Clustering the Same Set of Points



(a) Original points.



(b) Two clusters.



(c) Four clusters.



(d) Six clusters.

*K-means clustering critically depends upon the value of k*



# "Optimal" K-Means Clustering

- The **optimal  $k$ -means clustering problem** calls for finding cluster centers that minimize the intra-class variance, i.e. the sum of squared distances from each data point being clustered to its cluster center;
- **The problem stated formally:**
  - Given a finite set  $S$  where each element is a vector of length  $d$ , find a subset  $T$  of size  $k$  that minimizes the sum of squares of the distances between elements in  $S$  and their closest element in  $T$
- Finding an exact solution to the  $k$ -means problem for arbitrary input has been shown to be **NP-hard**
- **NP-hardness** (non-deterministic polynomial-time **hard**), in computational complexity theory, is a class of problems that are, informally, "at least as **hard** as the hardest problems in **NP**".
- finding a polynomial algorithm to solve any NP-hard problem would give polynomial algorithms for all the problems in NP, which is unlikely

# K-Means Clustering Algorithm

## Mathematical Formulation

*(stated mathematically)*

Given an initial set of  $k$  means  $m_1^{(1)}, \dots, m_k^{(1)}$ , the algorithm proceeds by alternating between two steps:

1

**Assignment step:** Assign each observation to the cluster whose mean yields the least within-cluster sum of squares. Since the sum of squares is the squared **Euclidean distance**, this is intuitively the "nearest" mean

$$S_i^{(t)} = \{x_p : \|x_p - m_i^{(t)}\|^2 \leq \|x_p - m_j^{(t)}\|^2 \forall j, 1 \leq j \leq k\},$$

where each  $x_p$  is assigned to exactly one  $S_i^{(t)}$ , even if it could be assigned to two or more of them.

**Update step:** Calculate the new means to be the **centroids** of the observations in the new clusters.

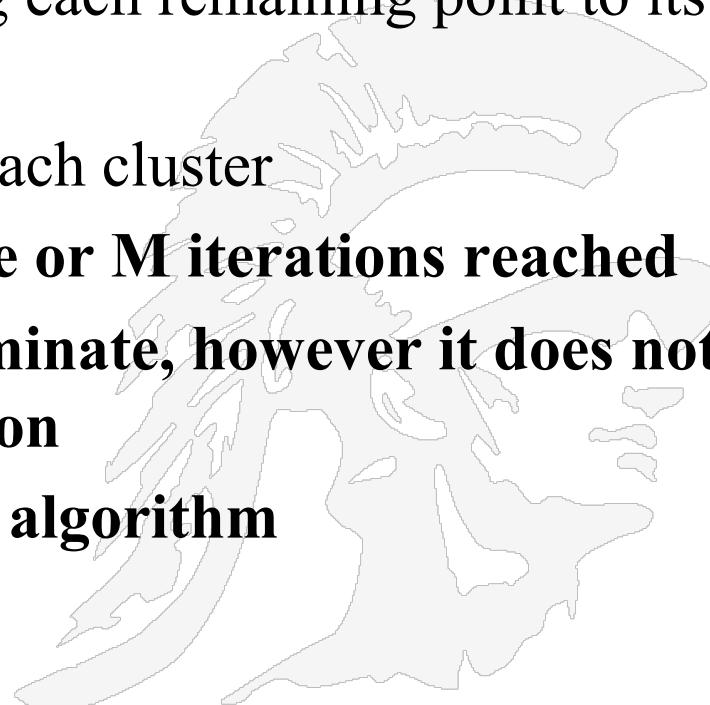
$$m_i^{(t+1)} = \frac{1}{|S_i^{(t)}|} \sum_{x_j \in S_i^{(t)}} x_j$$

- The algorithm has converged when the assignments no longer change.
- The algorithm will converge to a (local) optimum.
- There is no guarantee that the global optimum is found using this algorithm.



# An Approximation Clustering Algorithm

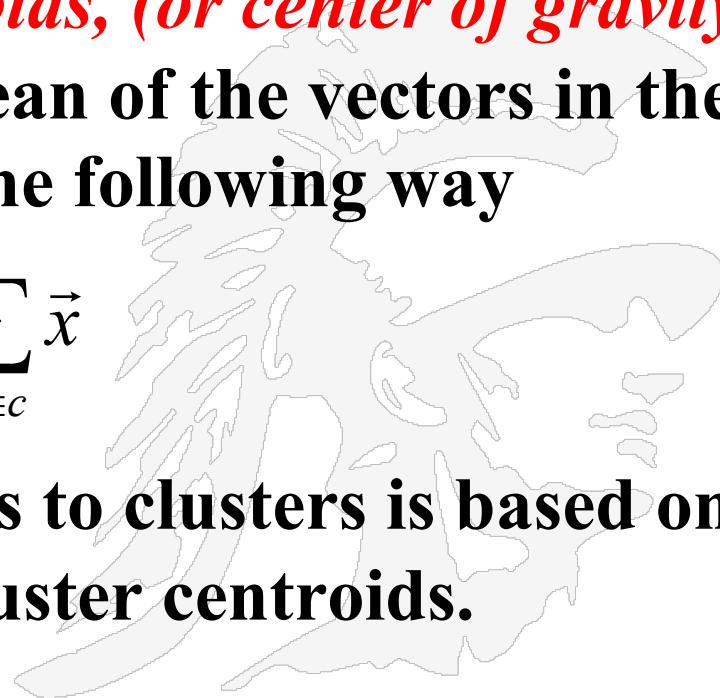
1. **Select K points as initial centroids**
2. **repeat**
  - form K clusters by assigning each remaining point to its closest centroid
  - re-compute the centroid of each cluster
3. **until centroids do not change or M iterations reached**
  - **the algorithm will always terminate, however it does not always find the optimal solution**
  - **this is an example of a greedy algorithm**



## K-Means Depends on Centroids

- Assumes instances are real-valued vectors
  - Let  $\vec{x}$  represent the vectors in a cluster  $c$
- Then we define the *centroids, (or center of gravity)*, of the cluster to be the mean of the vectors in the cluster; we write this in the following way

$$\vec{\mu}(c) = \frac{1}{|c|} \sum_{\vec{x} \in c} \vec{x}$$



- Reassignment of instances to clusters is based on distance to the current cluster centroids.

# There are Several Possible Distance Metrics

- Euclidean distance ( $L_2$  norm):

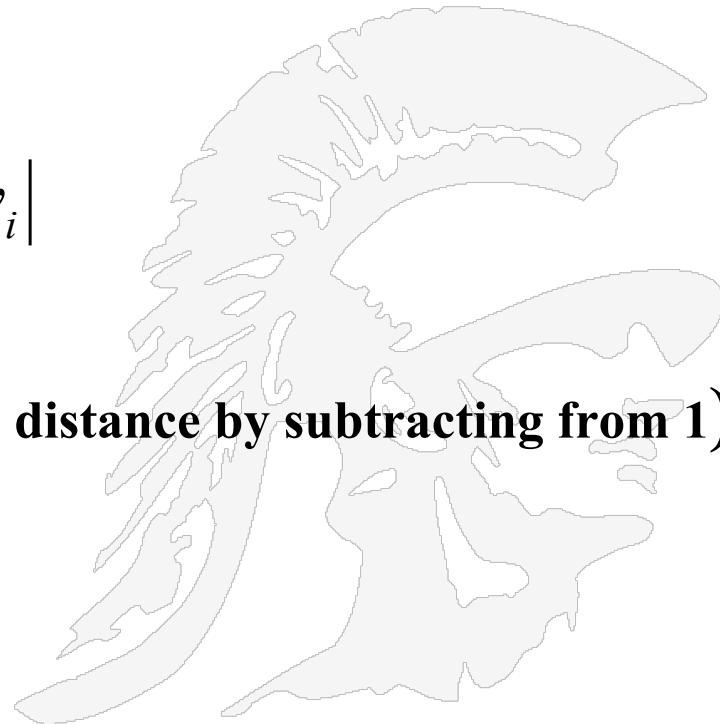
$$L_2(\vec{x}, \vec{y}) = \sqrt{\sum_{i=1}^m (x_i - y_i)^2}$$

- $L_1$  norm:

$$L_1(\vec{x}, \vec{y}) = \sum_{i=1}^m |x_i - y_i|$$

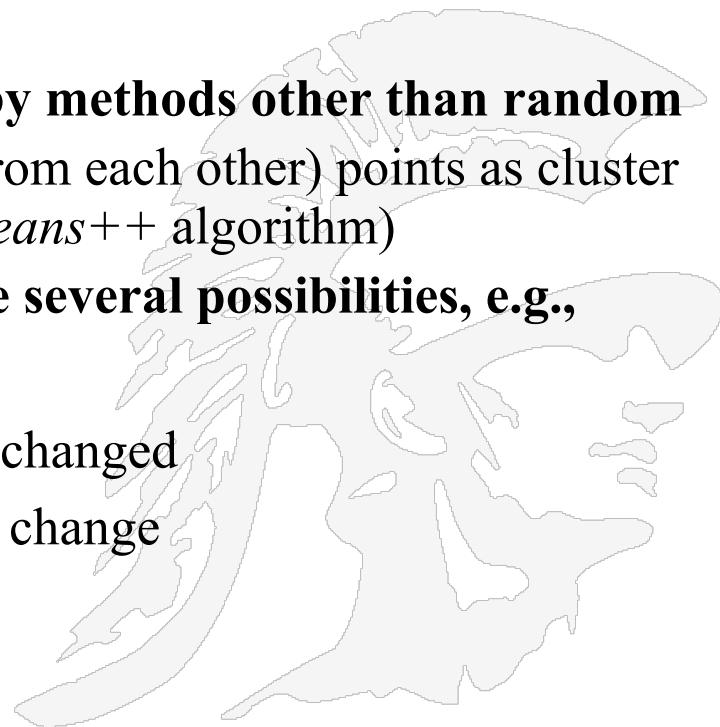
- Cosine Similarity (transform to a distance by subtracting from 1):

$$1 - \frac{\vec{x} \cdot \vec{y}}{|\vec{x}| \cdot |\vec{y}|}$$



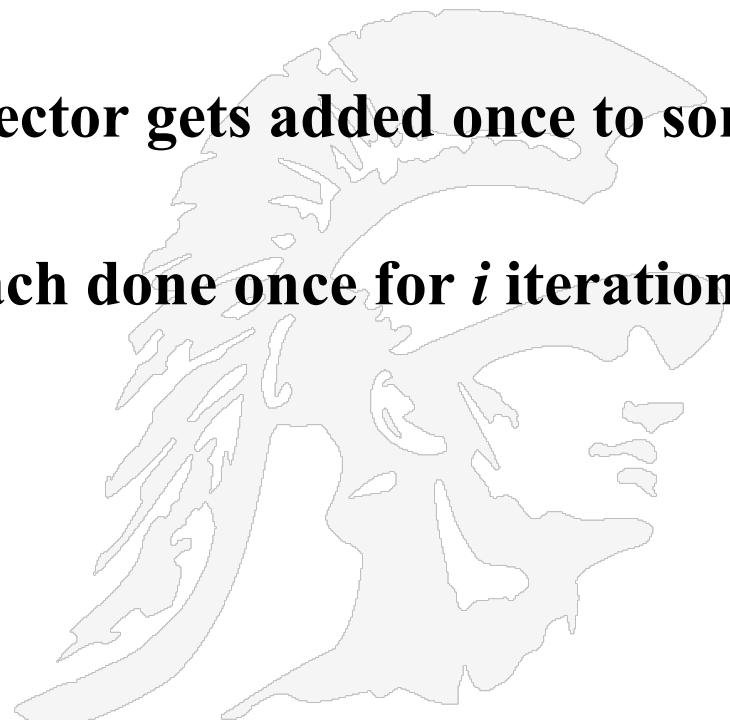
# Some Adjustments to the Algorithm

- How to pick the initial cluster means points
  - Try multiple runs
    - Choose different random points as the cluster means and see which yields the best result
  - Select the original set of means by methods other than random
    - E.g., pick the most distant (from each other) points as cluster centers (this is called the *k-means++* algorithm)
- For termination conditions there are several possibilities, e.g.,
  - After a fixed number of iterations
  - When the document partition is unchanged
  - When the centroid positions don't change



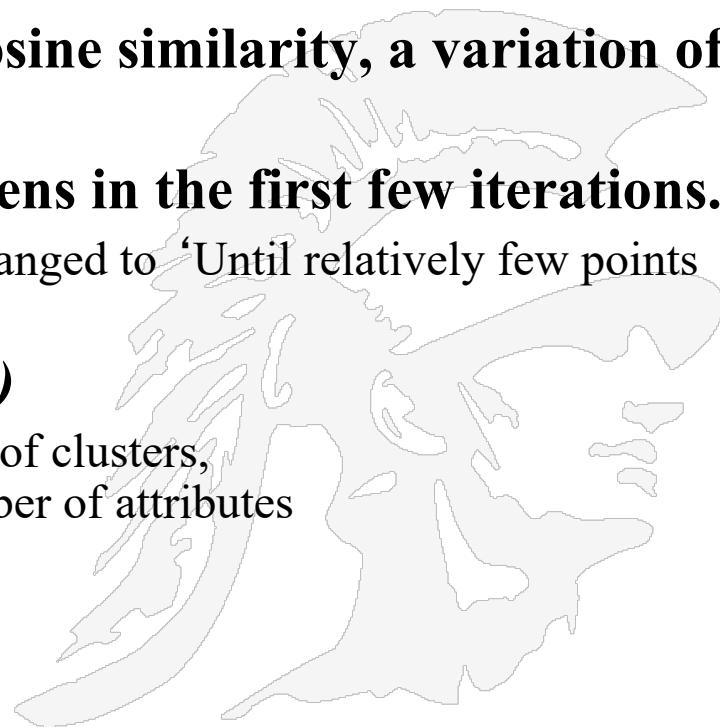
# Time Complexity

- Computing distance between two vectors is  $O(m)$  where  $m$  is the dimensionality of the vectors
- Re-assigning  $n$  vectors to  $k$  clusters:  $O(kn)$  distance computations, or  $O(knm)$
- Computing centroids: Each vector gets added once to some centroid:  $O(nm)$
- Assume these two steps are each done once for  $i$  iterations:  $O(iknm)$
- Note:
  - $m$  is the size of the vector
  - $n$  is the number of vectors (items)
  - $k$  is the number of clusters
  - $i$  depends upon convergence



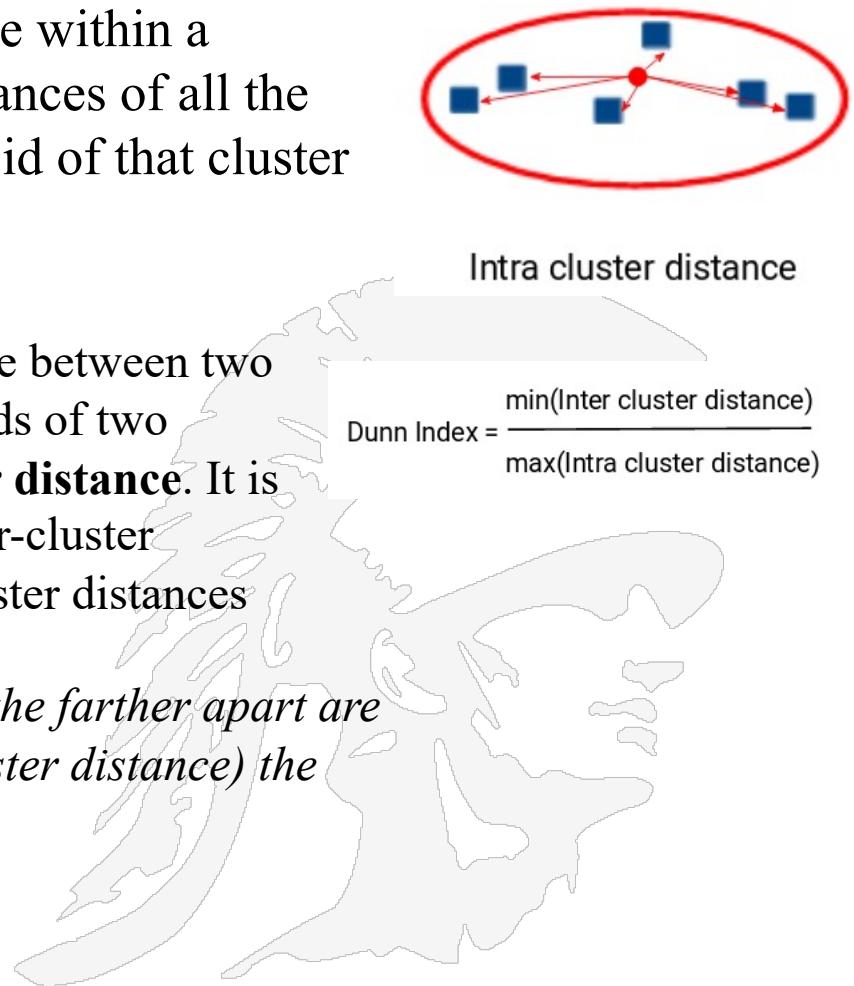
# K-means Clustering – Summary Details

- **Initial centroids are often chosen randomly**
  - Clusters produced vary from one run to another
- **The centroid is (typically) the mean of the points in the cluster**
- **‘Closeness’ is measured by cosine similarity, a variation of Euclidean distance**
- **Most of the convergence happens in the first few iterations.**
  - Often the stopping condition is changed to ‘Until relatively few points change clusters
- **Complexity is  $O(i * k * n * m)$** 
  - $n$  = number of points,  $k$  = number of clusters,  
 $i$  = number of iterations,  $m$  = number of attributes



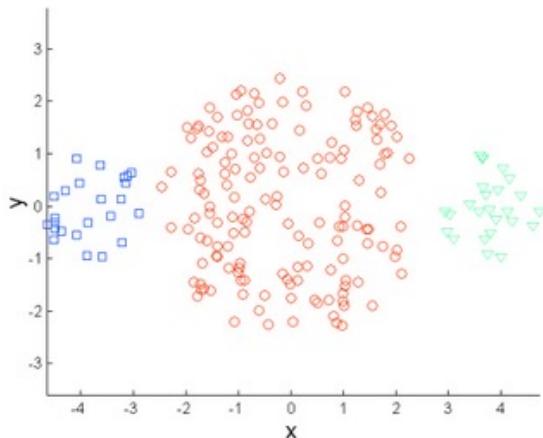
# Additional Evaluation Metrics for K-means Clustering

- ***inertia*** evaluates how far the points are within a cluster, or specifically the sum of distances of all the points within a cluster from the centroid of that cluster
- **Dunn Index** takes into account the distance between two clusters. This distance between the centroids of two different clusters is known as **inter-cluster distance**. It is computed as the ratio of the minimum inter-cluster distance and the maximum of the intra-cluster distances
- *The larger the min(inter-cluster distance) the farther apart are the clusters; the smaller the max(intra-cluster distance) the more compact are the clusters*

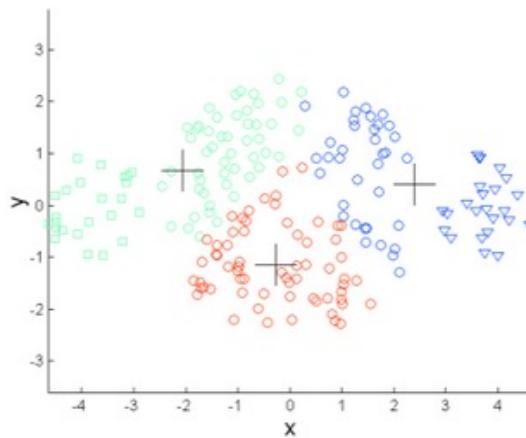


## Difficulties with K-Means Clustering

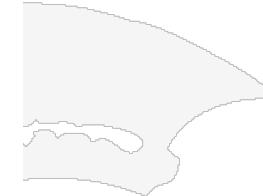
When cluster sizes are very different in size, points in the larger section can be mis-clustered



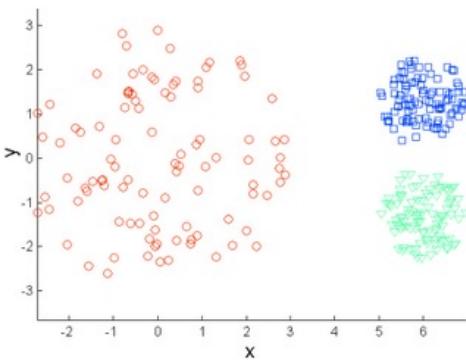
Original Points



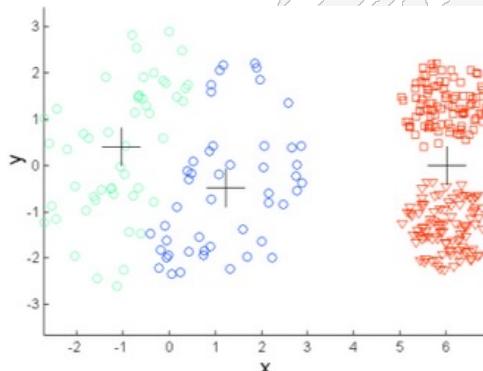
K-means ( $k = 3$ )



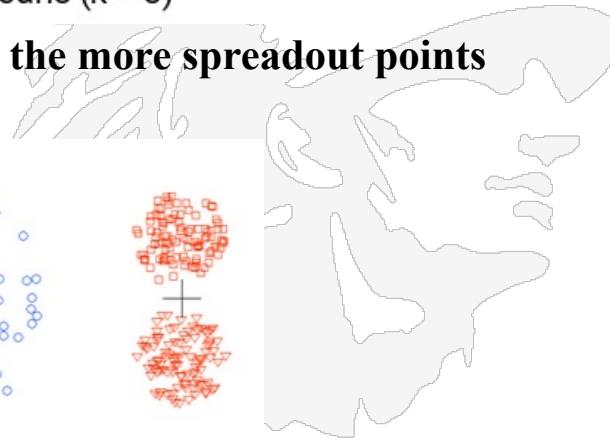
When the densities of the original points are different the more spreadout points can be mis-clustered



Original Points

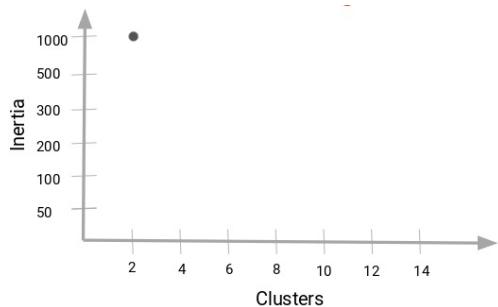


K-means ( $k = 3$ )

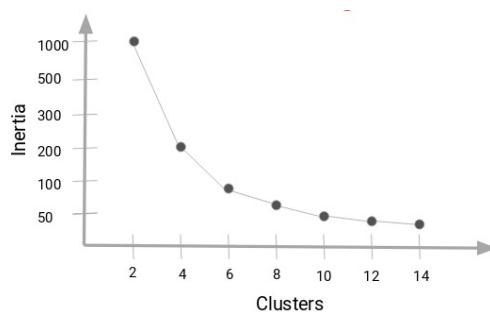


# How to Choose the Right Number of Clusters

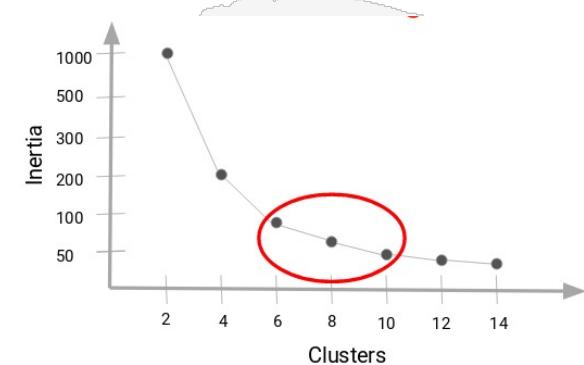
- Plot a graph, also known as an elbow curve, where the x-axis will represent the number of clusters and the y-axis will be an evaluation metric.
- Let's say we use inertia



Train your model on 2 clusters  
Compute and plot the inertia



Train the model on  
successively higher clusters

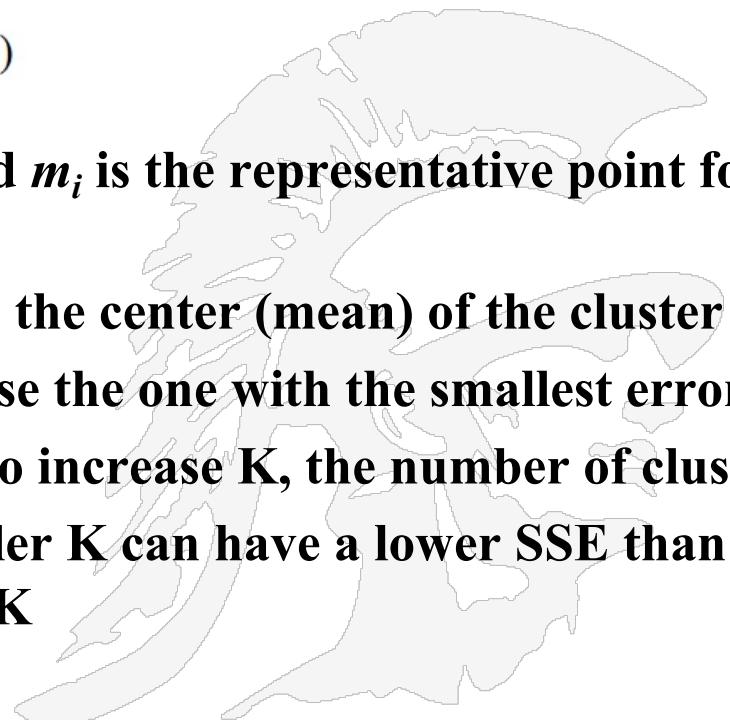


Any number of clusters  
between 6 and 10 will work

*the cluster value where this decrease in inertia value becomes constant can be chosen as the right cluster value for your data.*

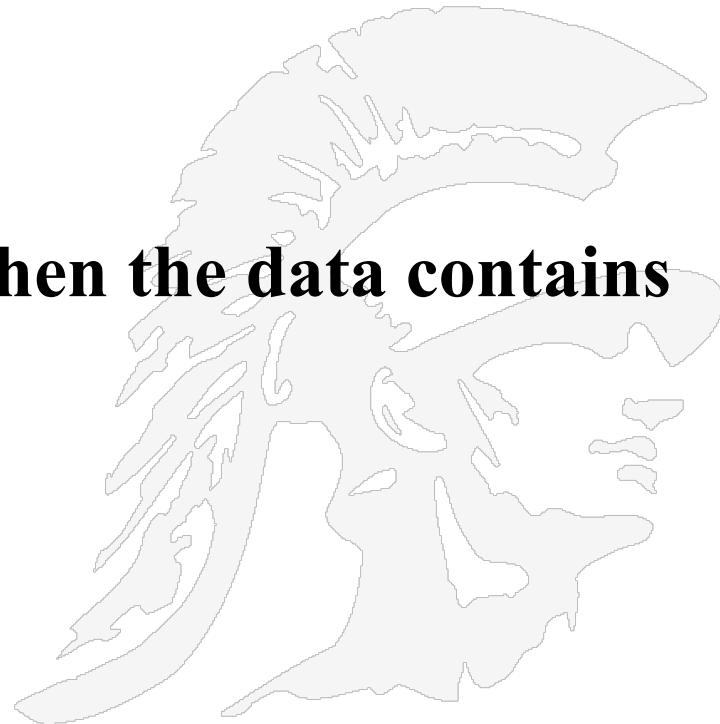
# Evaluating K-Means Clusters

- Most common measure is Sum of Squared Error (SSE)
  - For each point, the error is the distance to the nearest cluster
  - To get SSE, we square these errors and sum them.
- $x$  is a data point in cluster  $C_i$  and  $m_i$  is the representative point for cluster  $C_i$
- can show that  $m_i$  corresponds to the center (mean) of the cluster
- Given two clusters, we can choose the one with the smallest error
- One easy way to reduce SSE is to increase K, the number of clusters
  - A good clustering with smaller K can have a lower SSE than a poor clustering with higher K



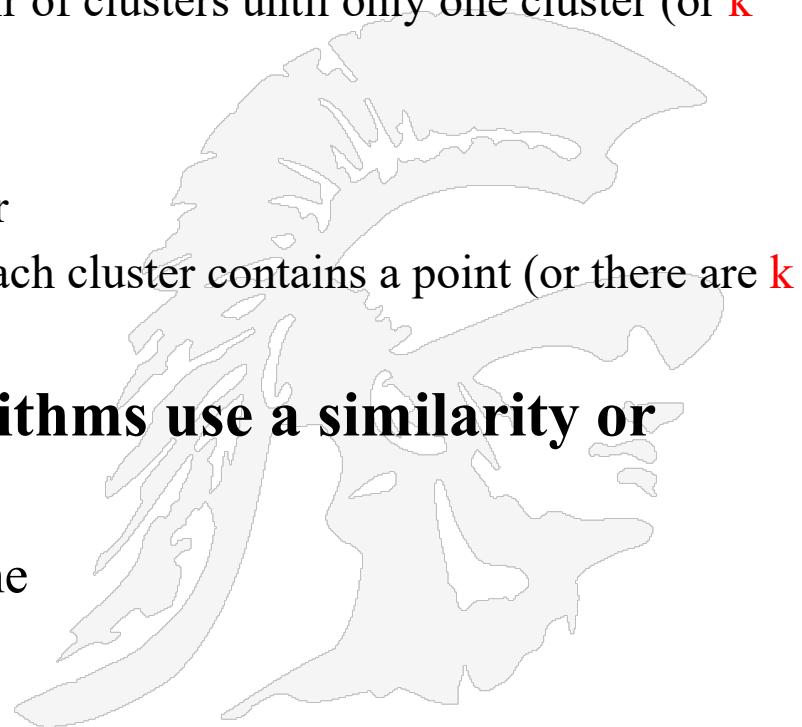
# Limitations of K-Means

- K-means has problems when clusters are of differing
  - Sizes
  - Densities
  - Non-globular shapes
- K-means has problems when the data contains outliers



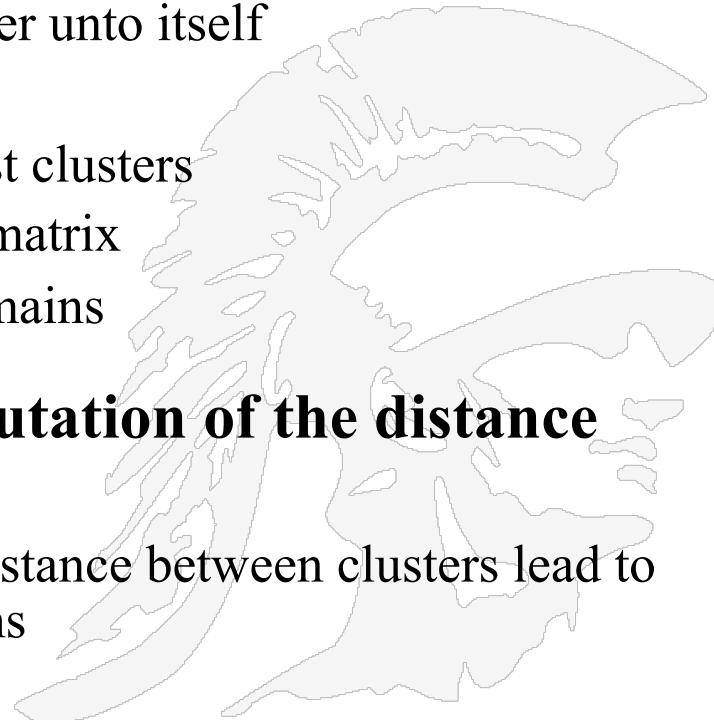
# Hierarchical Clustering Algorithms

- Two main types of hierarchical clustering
  - **Agglomerative:**
    - Start with the points as individual clusters
    - At each step, merge the closest pair of clusters until only one cluster (or  $k$  clusters) left (bottom-up)
  - **Divisive:**
    - Start with one, all-inclusive cluster
    - At each step, split a cluster until each cluster contains a point (or there are  $k$  clusters), (top-down)
- Traditional hierarchical algorithms use a similarity or distance matrix
  - Merge or split one cluster at a time



## Agglomerative Clustering Algorithm - a Bottom Up Approach

- **Basic Agglomerative Clustering Algorithm**
  1. Compute the distance matrix between the input data points (i.e. the distance between all pairs of points)
  2. Let each data point be a cluster unto itself
  3. Repeat
    4. Merge the two closest clusters
    5. Update the distance matrix
  6. Until only a single cluster remains
- **Key operation is the computation of the distance between two clusters**
  - Different definitions of the distance between clusters lead to somewhat different algorithms



# How Can We Compute the Distance Between Two Clusters

- As before, the **Centroid** of a cluster is the component-wise average of the vectors in a cluster, which is itself a vector
- Example, the Centroid of (1,2,3); (4,5,6); (7,2,6); is (4,3,5)
- **4 possible ways to compute the distance between two clusters**

## 1. Center of Gravity

- Compute the distance between the two centroids of the cluster

## 2. Average Link

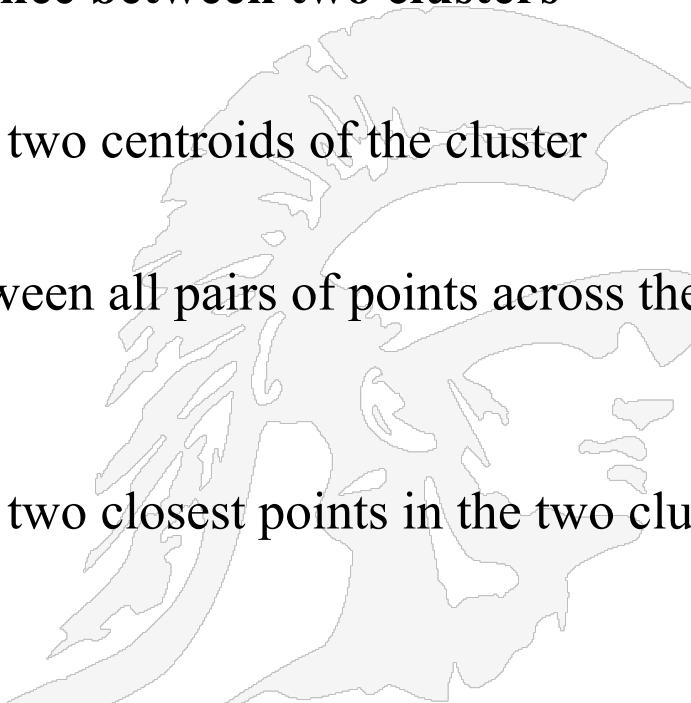
- Compute the average distance between all pairs of points across the two clusters

## 3. Single Link

- Compute the distance between the two closest points in the two clusters, i.e. the most cosine similar

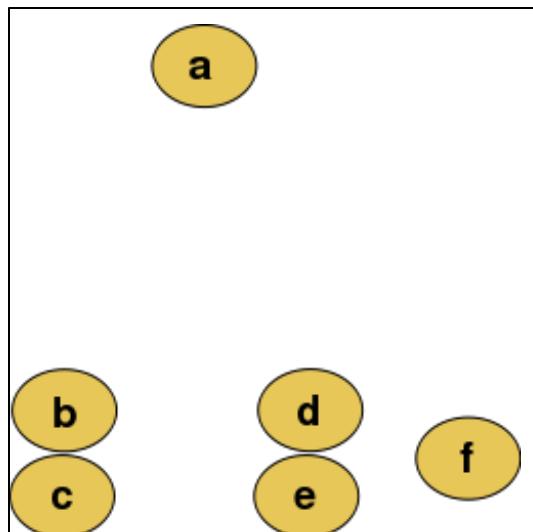
## 4. Complete Link

- Compute the distance between the furthest points in the two clusters, i.e. the least cosine similar

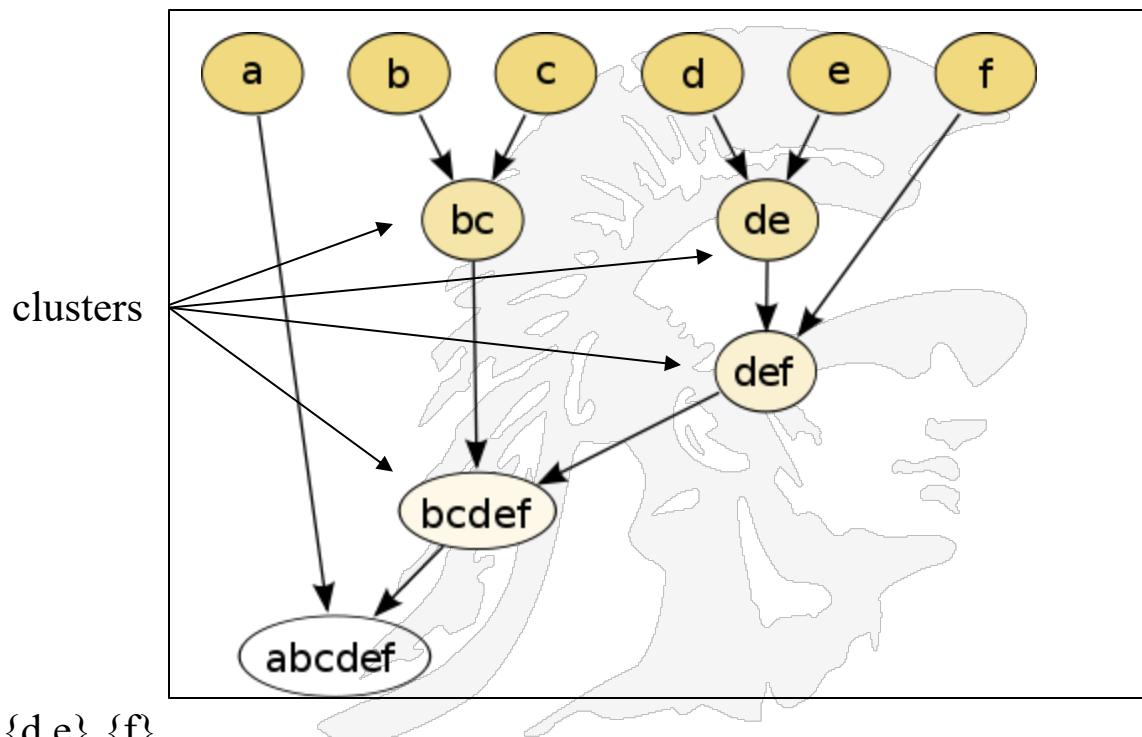


# A Dendrogram is Used to Display Clusters

- A **dendrogram** is a tree diagram frequently used to illustrate the arrangement of the clusters produced by hierarchical clustering



original input

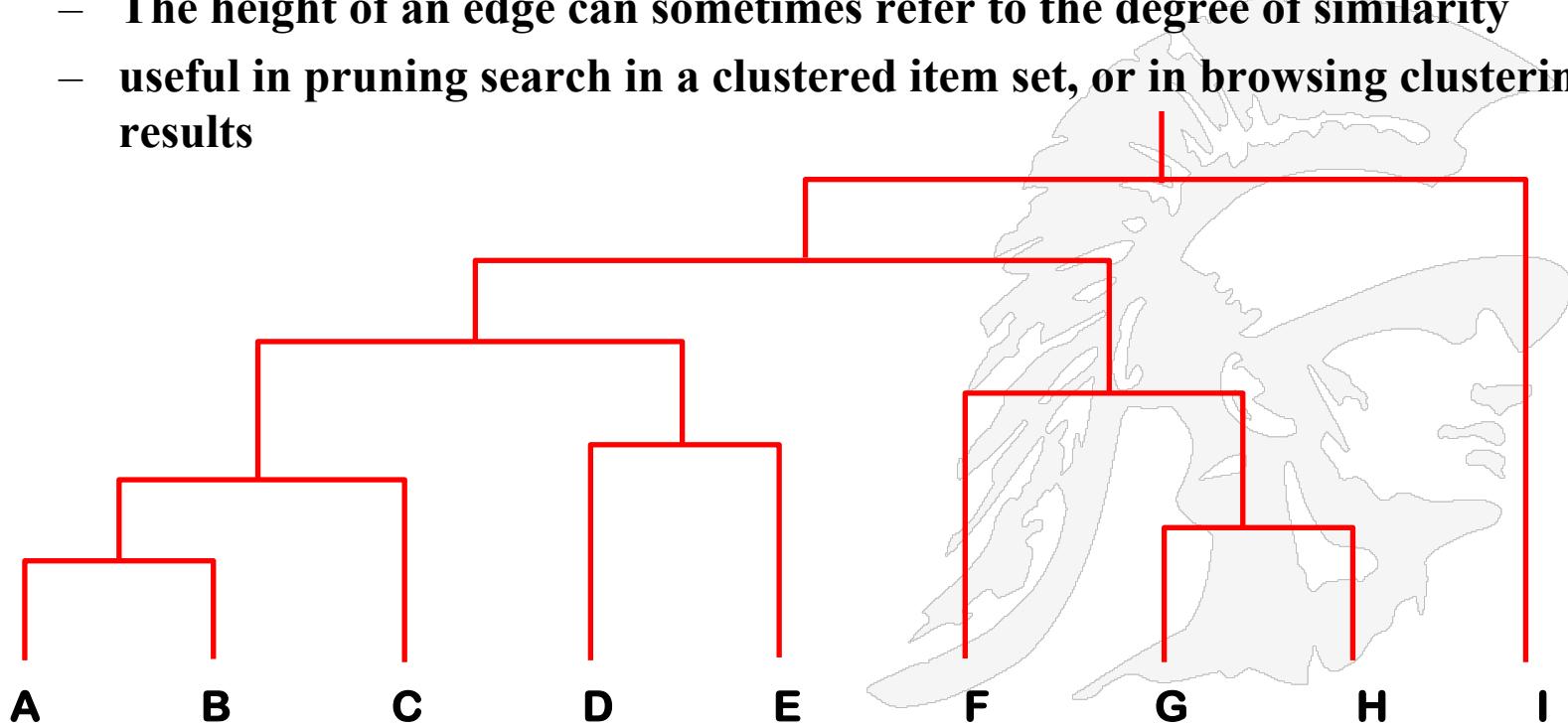


second row clusters are: {a}, {b c}, {d e} {f}  
third row clusters are: {a}, {b c} {d e f}

corresponding dendrogram

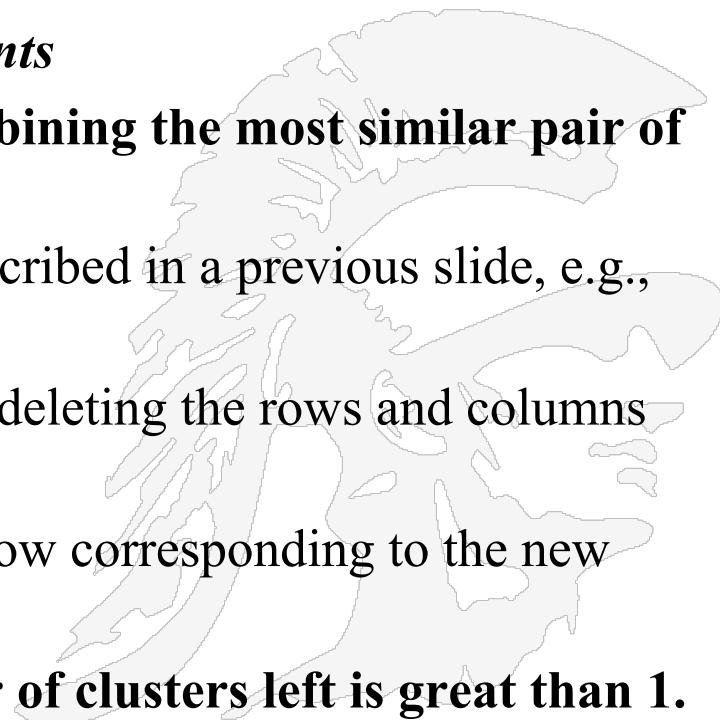
# Hierarchical Agglomerative Clustering

- HAC starts with unclustered data and performs successive pairwise joins among items (or previous clusters) to form larger ones
  - this results in a hierarchy of clusters which can be viewed as a **dendrogram**
  - Dendograms are usually drawn as shown below
  - The height of an edge can sometimes refer to the degree of similarity
  - useful in pruning search in a clustered item set, or in browsing clustering results



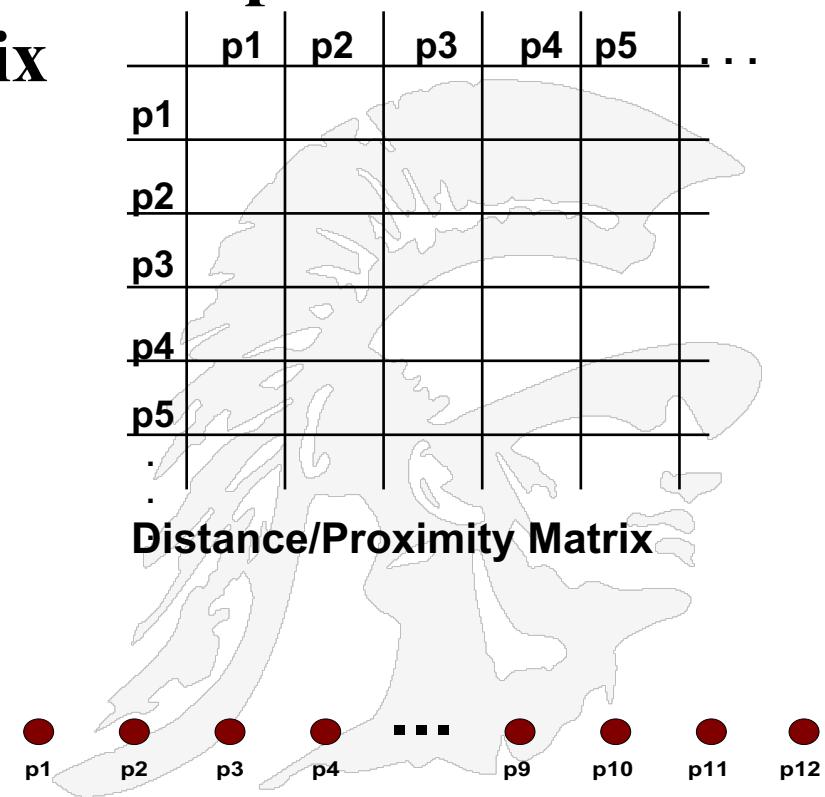
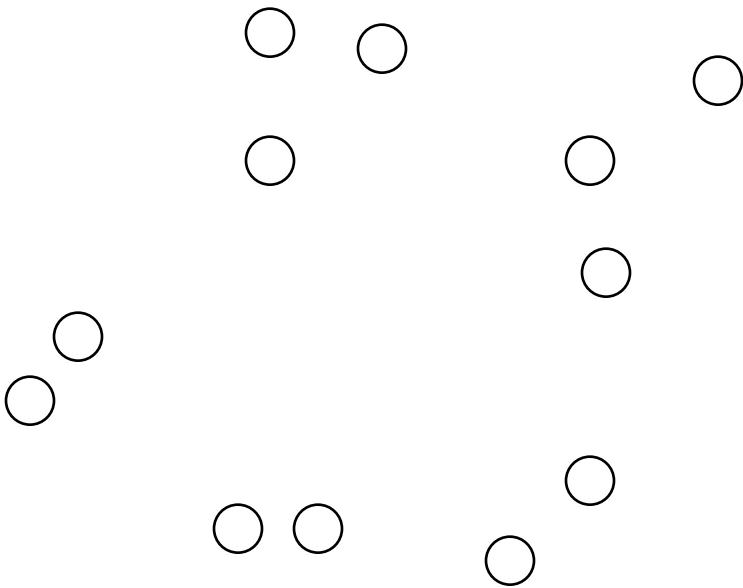
# Hierarchical Agglomerative Clustering

- **Basic procedure**
  - 1. Place each of N documents into a class of its own.
  - 2. Compute all pairwise document-document similarity coefficients
    - *Total of  $N(N-1)/2$  coefficients*
  - 3. **Form a new cluster by combining the most similar pair of current clusters  $i$  and  $j$** 
    - (use one of the methods described in a previous slide, e.g., complete link, etc.);
    - update similarity matrix by deleting the rows and columns corresponding to  $i$  and  $j$ ;
    - calculate the entries in the row corresponding to the new cluster  $i+j$ .
  - 4. **Repeat step 3 if the number of clusters left is great than 1.**



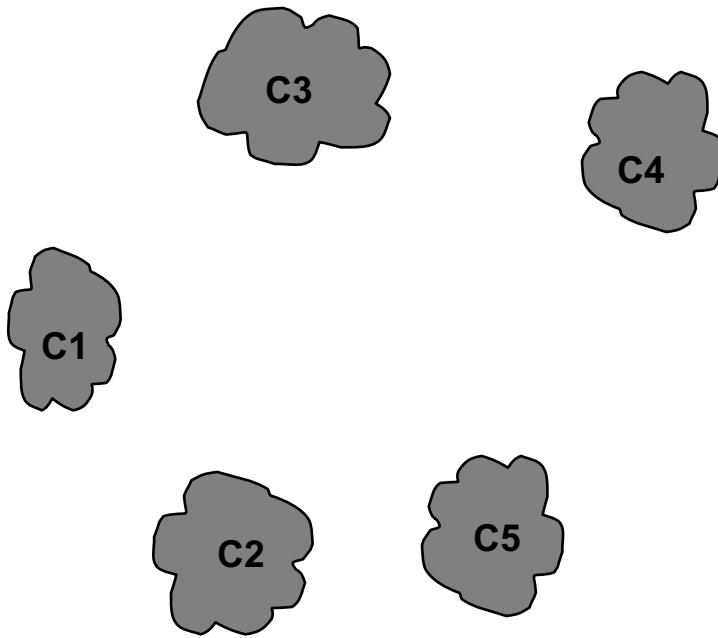
## Example of HAC Input/ Initial setting

- Start with clusters of individual points and a distance/proximity matrix



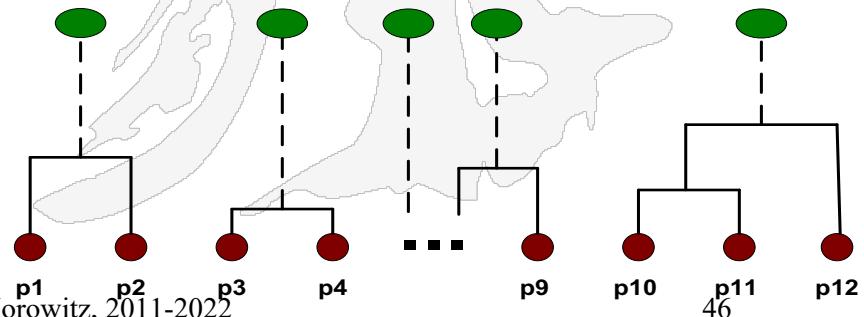
# Intermediate State

- After some merging steps, we have some clusters



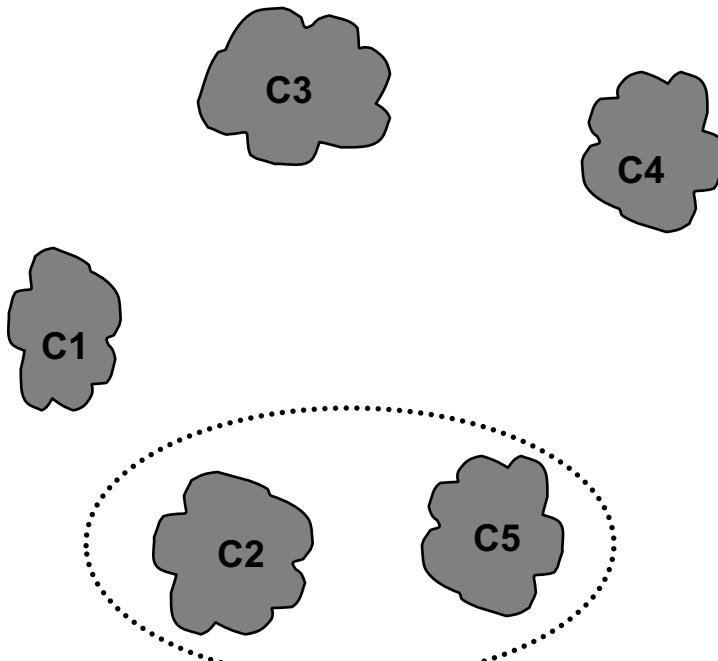
	C1	C2	C3	C4	C5
C1					
C2					
C3					
C4					
C5					

Distance/Proximity Matrix (above)  
Dendrogram (below)



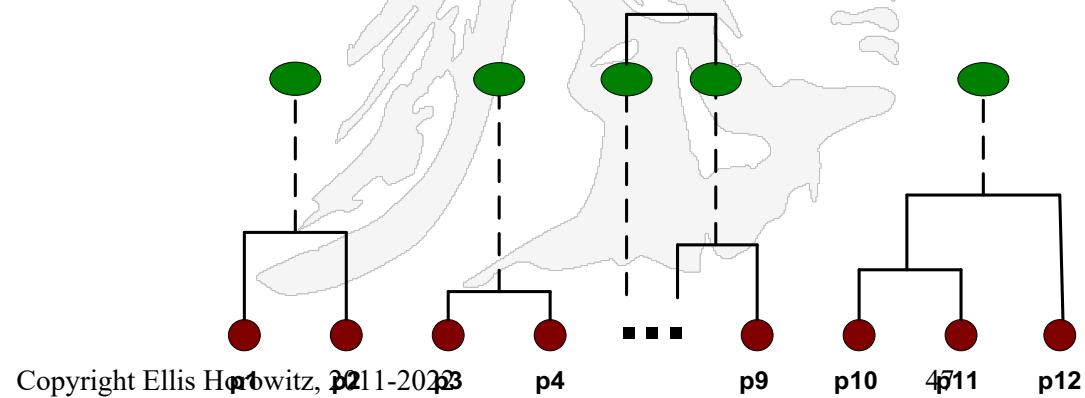
# Intermediate State

- Merge the two closest clusters (C2 and C5) and update the distance matrix.



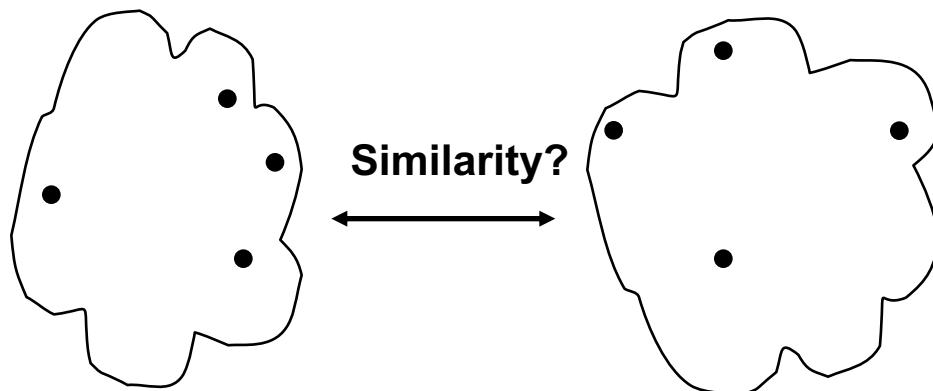
	C1	C2	C3	C4	C5
C1					
C2					
C3					
C4					
C5					

Distance/Proximity Matrix



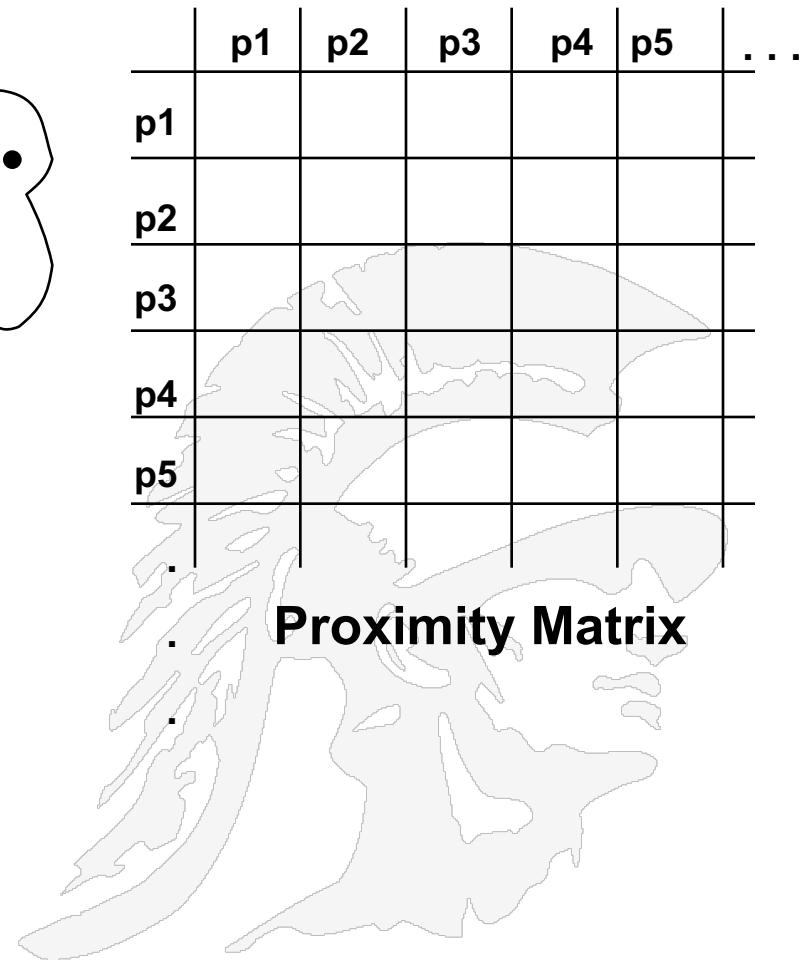
# How to Define Inter-Cluster Similarity

1

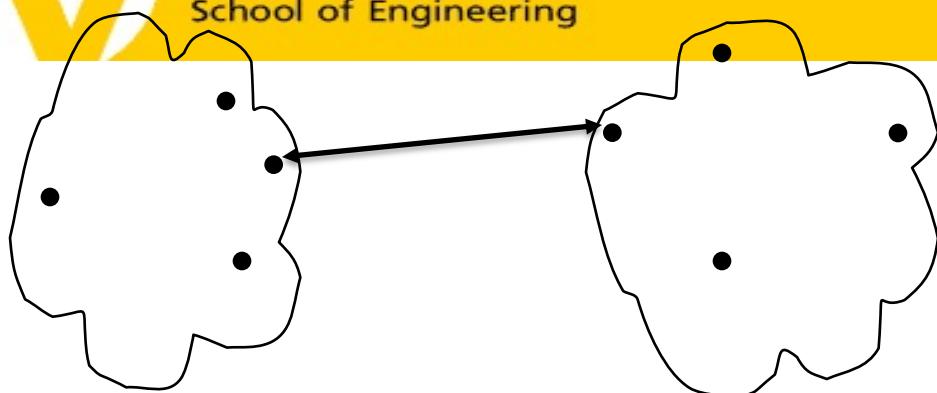


- MIN
- MAX
- Group Average
- Distance Between Centroids

**Look back at slide 38**

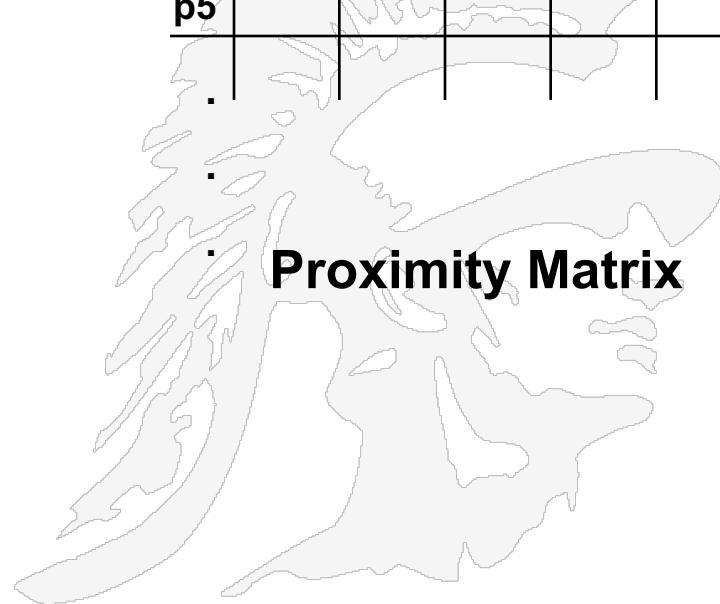


## How to Define Inter-Cluster Similarity

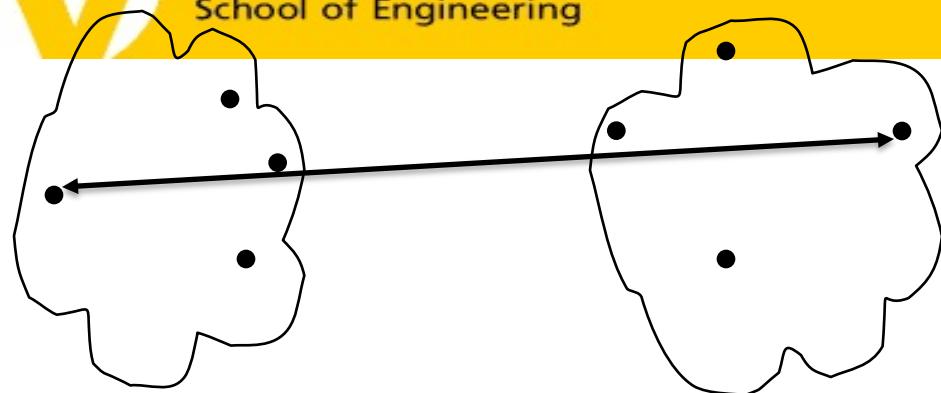


- MIN
- MAX
- Group Average
- Distance Between Centroids

	p1	p2	p3	p4	p5	...
p1						
p2						
p3						
p4						
p5						



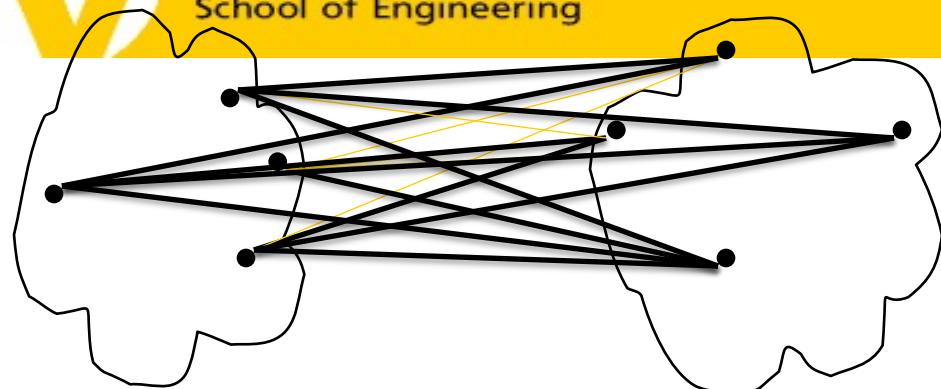
## How to Define Inter-Cluster Similarity



- MIN
- MAX
- Group Average
- Distance Between Centroids

	p1	p2	p3	p4	p5	...
p1						
p2						
p3						
p4						
p5						

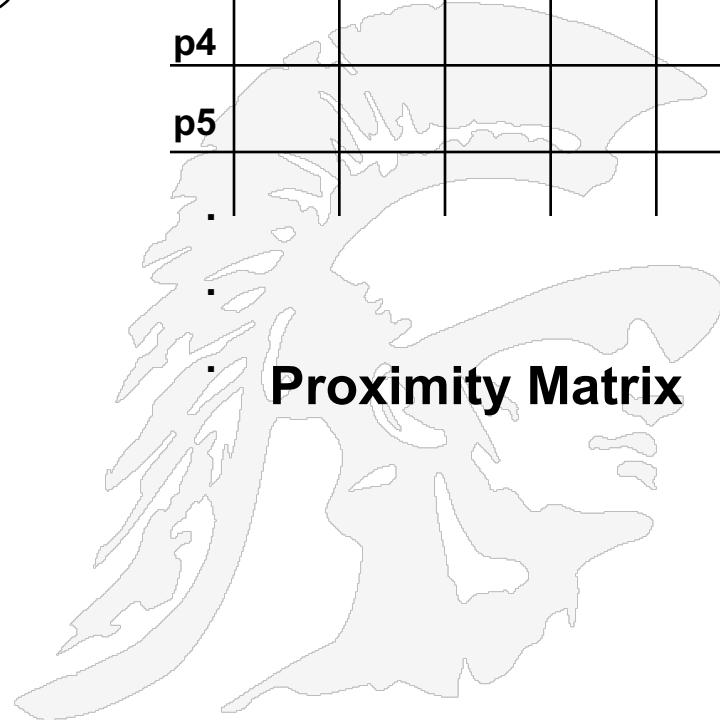




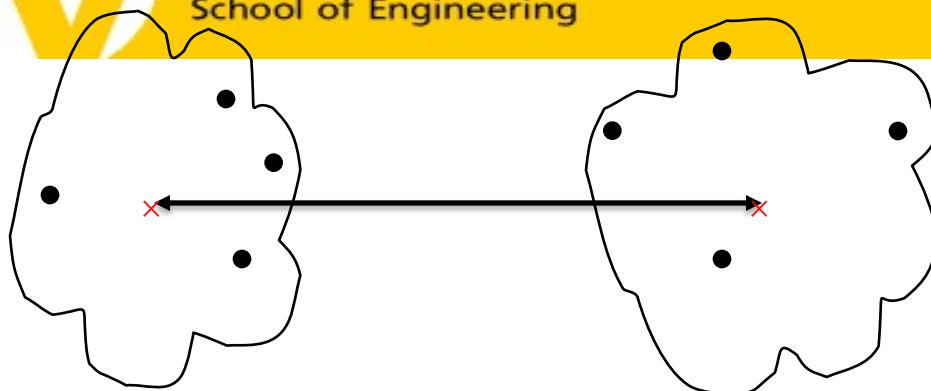
- MIN
- MAX
- Group Average
- Distance Between Centroids

## How to Define Inter-Cluster Similarity

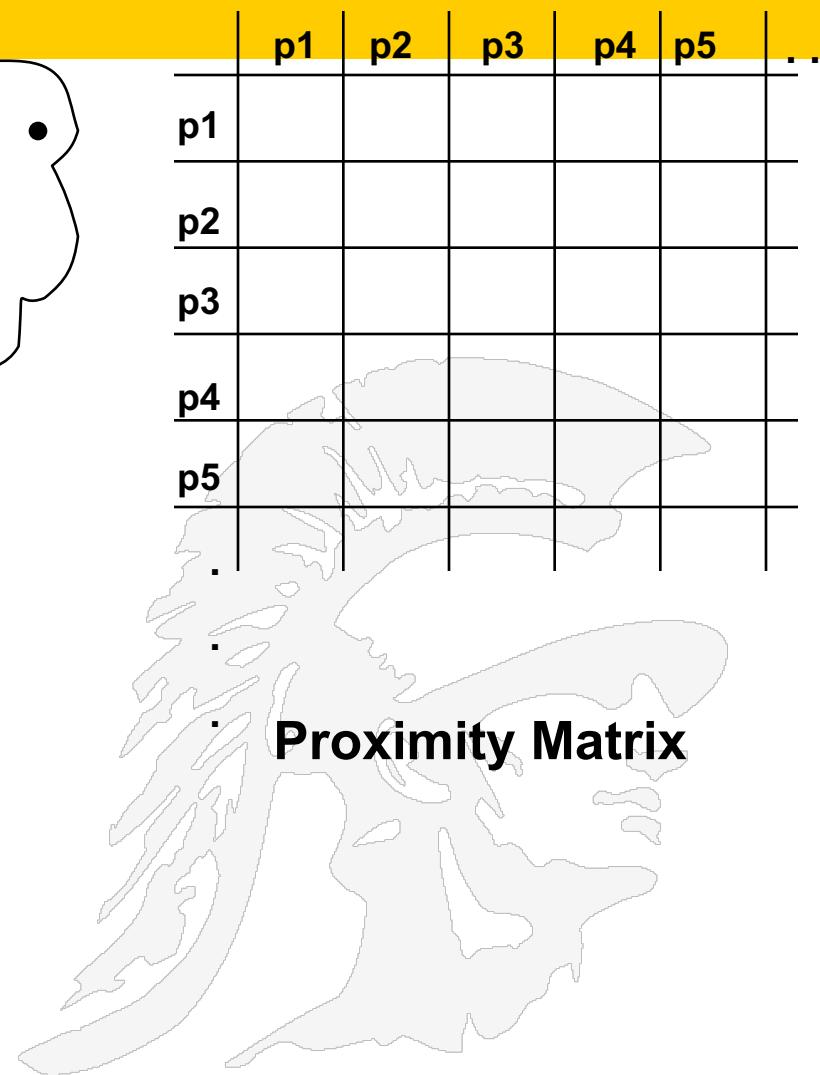
	p1	p2	p3	p4	p5	...
p1						



# How to Define Inter-Cluster Similarity



- MIN
- MAX
- Group Average
- Distance Between Centroids



# General Hierarchical Agglomerative Clustering Algorithm and Complexity

1. Compute similarity between all pairs of documents

  
 $O(N^2)$

2. Do  $N - 1$  times

1. Find closest pair of documents/clusters to merge







Naïve:  $O(N^2)$  Priority Queue:  $O(N)$  Single link:  $O(N)$

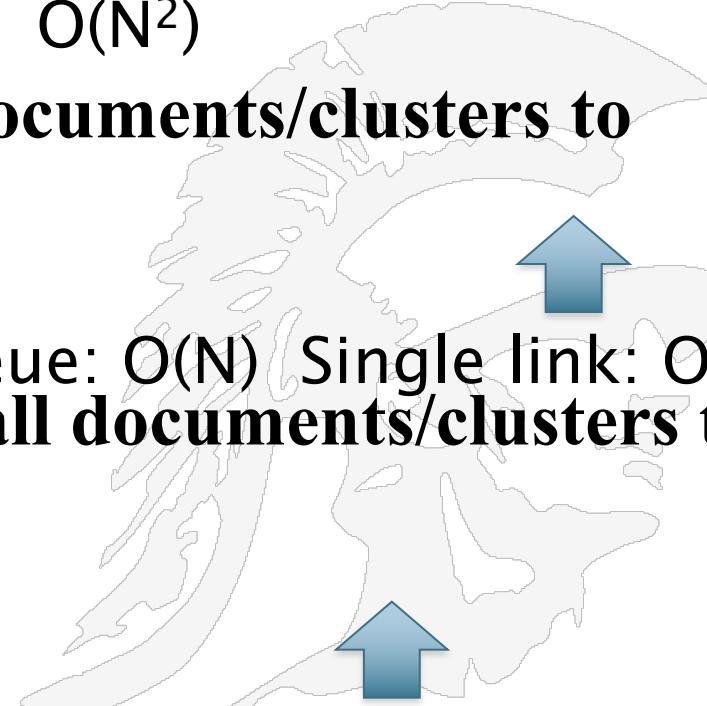
1. Update similarity of all documents/clusters to new cluster



Naïve:  
 $O(N)$



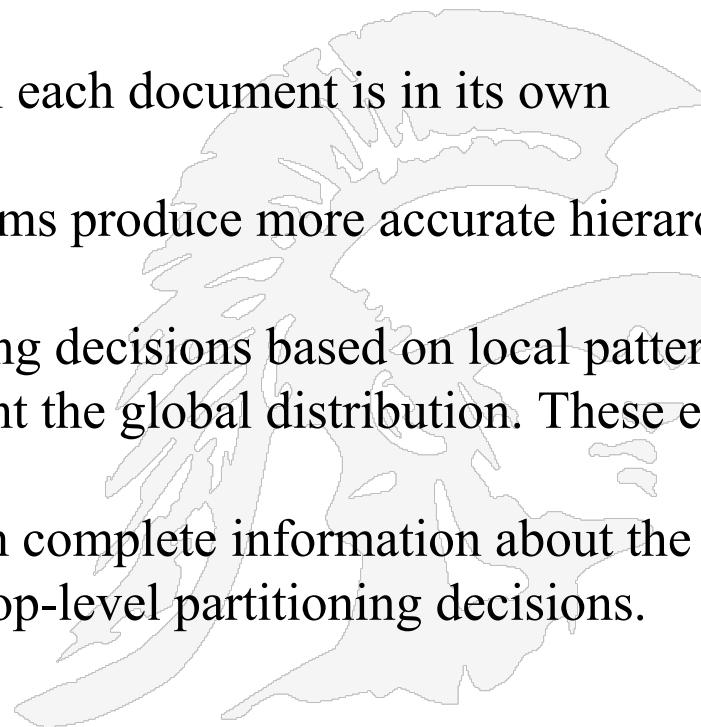
Priority Queue:  $O(N \log N)$



Single link:  
 $O(N)$

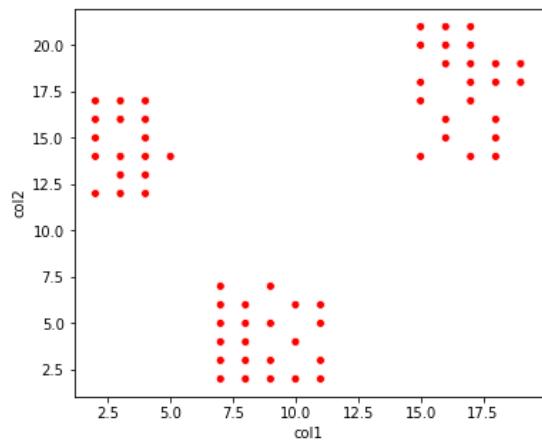
# Divisive Clustering Algorithm

1. Start at the top with all documents in one cluster.
  2. The cluster is split using a partitioning clustering algorithm.
    - Use the k-means clustering algorithm, which is linear in computing time whereas HAC (hierarchical agglomerative clustering) algorithms are quadratic
  3. Apply the procedure recursively until each document is in its own singleton cluster
- Studies show that the divisive algorithms produce more accurate hierarchies than bottom up
    - Bottom-up methods make clustering decisions based on local patterns without initially taking into account the global distribution. These early decisions cannot be undone.
    - Top-down clustering benefits from complete information about the global distribution when making top-level partitioning decisions.

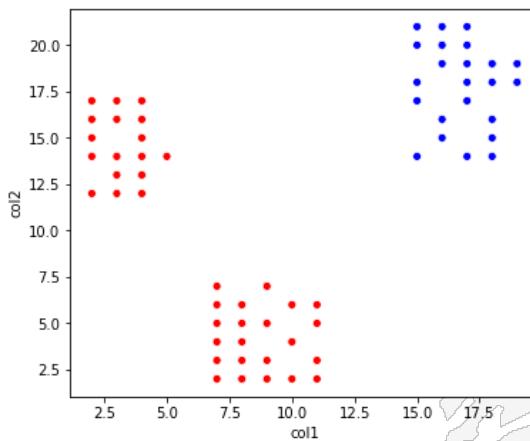


# Divisive Clustering Example

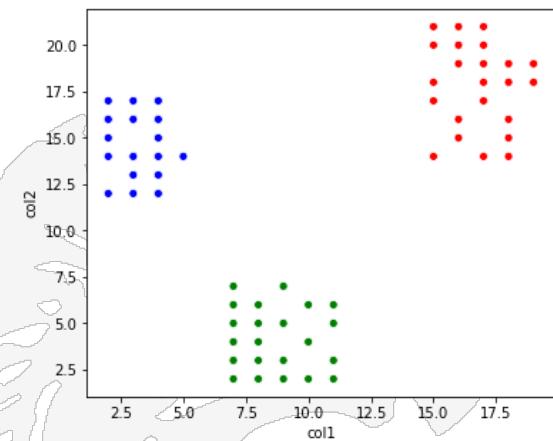
1. Initially, all points in the dataset belong to one single cluster.
2. Partition the cluster into two least similar cluster
3. Proceed recursively to form new clusters until the desired number of clusters is obtained.



All points in one cluster



Two clusters (blue/red)



Three clusters (blue/red/green)

- At this point the sum of inertia within each of the three clusters is smaller than the previous two examples of two clusters and one cluster
- subsequent splitting will only divide points within the existing three clusters

# How to Label Clusters

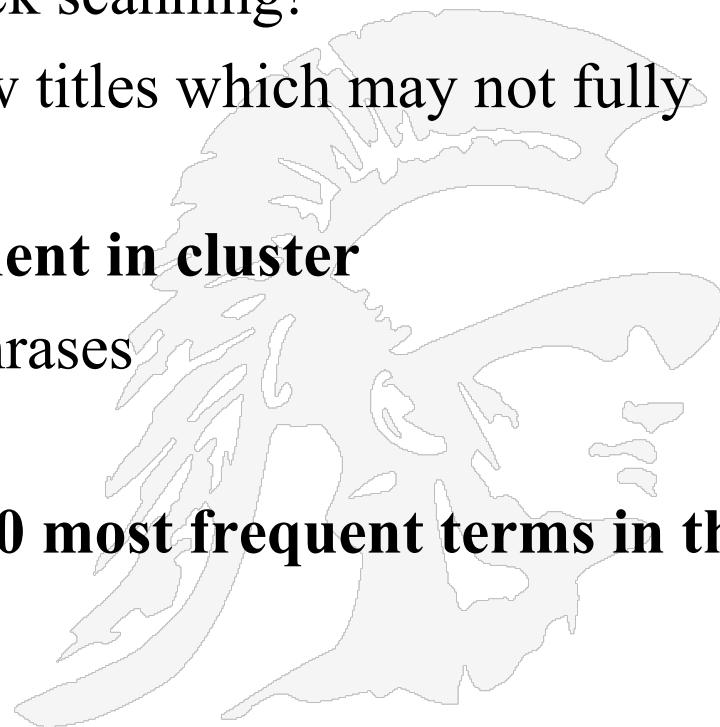
## Two Approaches

### 1. Show titles of typical documents

- Titles are easy to scan
- Authors create them for quick scanning!
- But you can only show a few titles which may not fully represent cluster

### 2. Show words/phrases prominent in cluster

- Use distinguishing words/phrases
- But harder to scan
- **Common heuristics - list 5-10 most frequent terms in the centroid vector**
  - Drop stop-words;

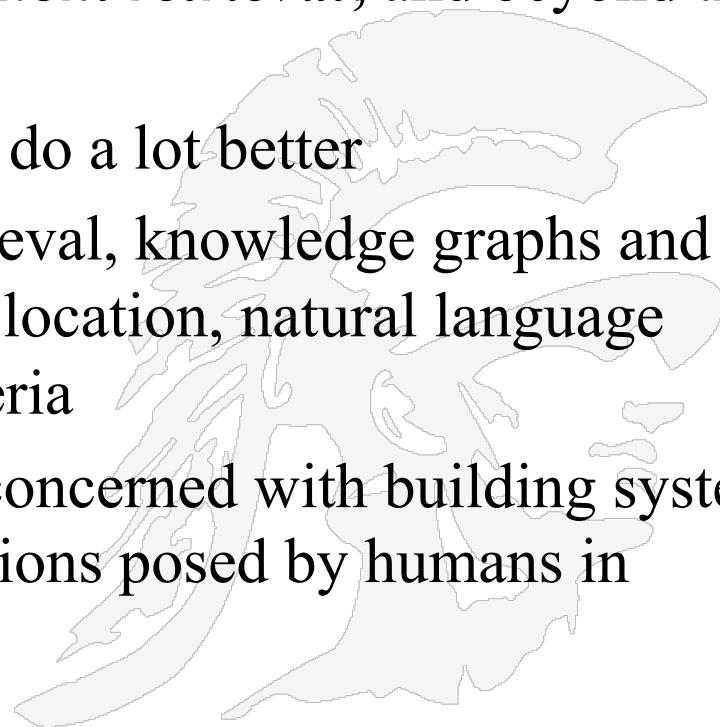


# Search Engine Question Answering



# Information Retrieval v. Question Answering

- The name “**information retrieval**” is standard, but as traditionally practiced, it’s not really right
- In the past all we got was ***document retrieval***, and beyond that the job is up to us
  - Modern search engines now do a lot better
- They combine information retrieval, knowledge graphs and inferencing, past query history, location, natural language processing and many other criteria
- **Question Answering (QA)** is concerned with building systems that automatically answer questions posed by humans in a natural language



People *want* to ask questions...

## Examples from Ask.com query log

how much should I weigh

what does my name mean

how to get pregnant

where can I find pictures of hairstyles

who is the richest man in the world

what is the meaning of life

why is the sky blue

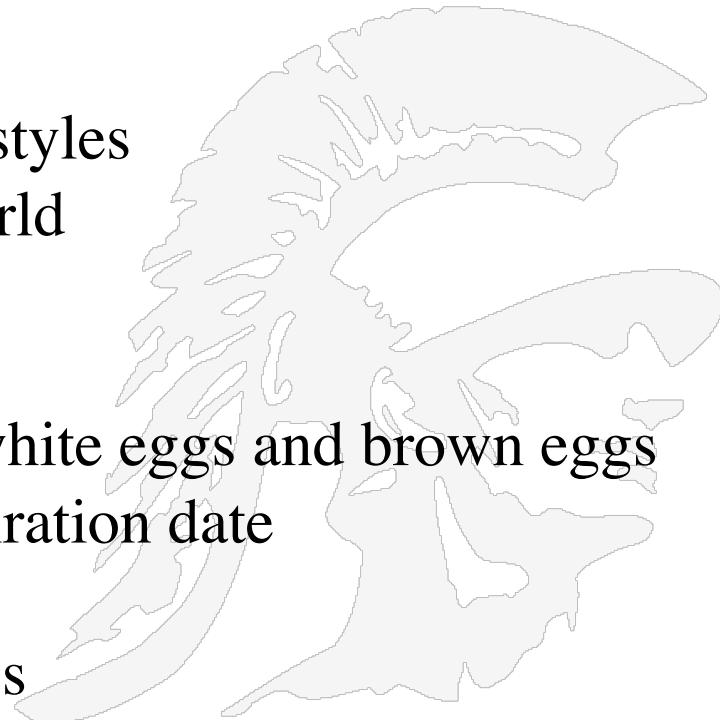
what is the difference between white eggs and brown eggs

can you drink milk after the expiration date

what is true love

what is the jonas brothers address

**Around 10-20% of query logs are questions such as these**



Question: *Who was the prime minister of Australia during the Great Depression?*  
 Answer: *James Scullin (Labor) 1929–31*

Google Search: Who was the prime minister of Australia during the Great Depression? - Microsoft Internet Explorer

Address: +of+Australia+during+the+Great+Depression%3F&btnG=Google+Search

Google Search Site News

Advanced Search Preferences Language Tools Search Tips

Who was the prime minister of Australia Google Search

The following words are very common and were not included in your search: Who was the of the. [details]

Web Images Groups Directory News

Searched the web for **Who was the prime minister of Australia during the Great Depression?**. Results 1 - 1

Asking a question? Try out [Google Answers](#).

### From Poor Boy to Prime Minister

... how did he come to lead **Australia during** World War ... April 1939 Menz takes over as **Prime Minister** after the death of Lyons; Sept 3 1939 **Australia** declares war ...  
[john.curtin.edu.au/manofpeace/boytopm.html](http://john.curtin.edu.au/manofpeace/boytopm.html) - 23k - Mar 1, 2003 - [Cached](#) - [Similar pages](#)

### Activity: Banning of the Communist Party in World War II

... The **Great Depression** had brought enormous suffering to workers ... by the 'Prime Minister' and His ... the Communist Party in **Australia during** ...  
[john.curtin.edu.au/letters/activities/communism.html](http://john.curtin.edu.au/letters/activities/communism.html) - 8k - [Cached](#) - [Similar pages](#)  
[\[ More results from john.curtin.edu.au \]](#)

### Prime Ministers of Australia - Chifley

... defying the federal United **Australia** Party government ... Second World War until led du the 1930s. ... He became **Prime Minister** following Curtin's death, succeeding ...  
[www.nma.gov.au/primereministers/3.htm](http://www.nma.gov.au/primereministers/3.htm) - 30k - [Cached](#) - [Similar pages](#)

Page about Curtin (WW II Labor Prime Minister)  
 (Can deduce answer)

Page about Curtin (WW II Labor Prime Minister)  
 (Lacks answer)

Page about Chifley (Labor Prime Minister)  
 (Can deduce answer)

How Google used to respond to questions

Question: *Who was the prime minister of Australia during the Great Depression?*  
Answer: *James Scullin (Labor) 1929–31*

## Google's result today



who was the prime minister of a... x +

← → C ⌂ 🔍 google.com/search?q=who+was+the+prime+minister+of+australia+during+the+great+depression&rlz=1C5C... 🔍 ⌂ 🌐

\_apps\_ CSCI 572 Home P... USC Computer Science... ITS - Software Masters Student... University of Sout... Other Bookmarks

Google who was the prime minister of australia during the great depression

All Images News Videos Shopping More Settings Tools

About 9,390,000 results (0.72 seconds)

**James Scullin** became the new prime minister and **Bruce** lost his own seat of Flinders, the first sitting Australian prime minister to do so. However, on 24 October 1929, one week after Labor took power, the US stock market crashed.

www.nma.gov.au › defining-moments › resources › great-depression › Great Depression | National Museum of Australia

About Featured Snippets Feedback

en.wikipedia.org › wiki › Great\_Depression\_in\_Australia › Great Depression in Australia - Wikipedia

Australia suffered badly during the period of the Great Depression of the 1930s. ... The conservative Prime Minister of Australia, Stanley Bruce, wished to ... 1929–1935: Scullin and ... · Varying experiences of ... · Legacy of the Great ...

People also ask

Who was the prime minister of Australia during ww2? ▾

What areas of Australia were most affected by the Great Depression? ▾

Did the Great Depression affect Australia? ▾

Who was hit the hardest during the Great Depression? ▾

Feedback

# Google has Improved Its Ability to Answer Many Questions

how old is mariah carey - Google

google.com/search?q=how+old+is+mariah+carey&rlz=1C5CHFA\_enUS728US728&oq=how+ol...

CSCI 572 Home P... Piazza Spring2022 CSCI572\_Spring2... DEN D2L Page USC USC Schedule of...

Other Bookmark:

Google how old is mariah carey

All Images News Books Videos More Tools

About 127,000,000 results (0.73 seconds)

Mariah Carey / Age

**53 years**

March 27, 1969

People also search for

Nick Cannon 41 years Jennifer Lopez 52 years Beyoncé 40 years

Feedback

People also ask :

What is Mariah Carey's net worth 2021?

Why does Mariah Carey touch her ear?

Is Mariah Carey richer than Nick Cannon?

Who is Mariah Carey husband?

Feedback

**Mariah Carey**  
American singer-songwriter

Available on

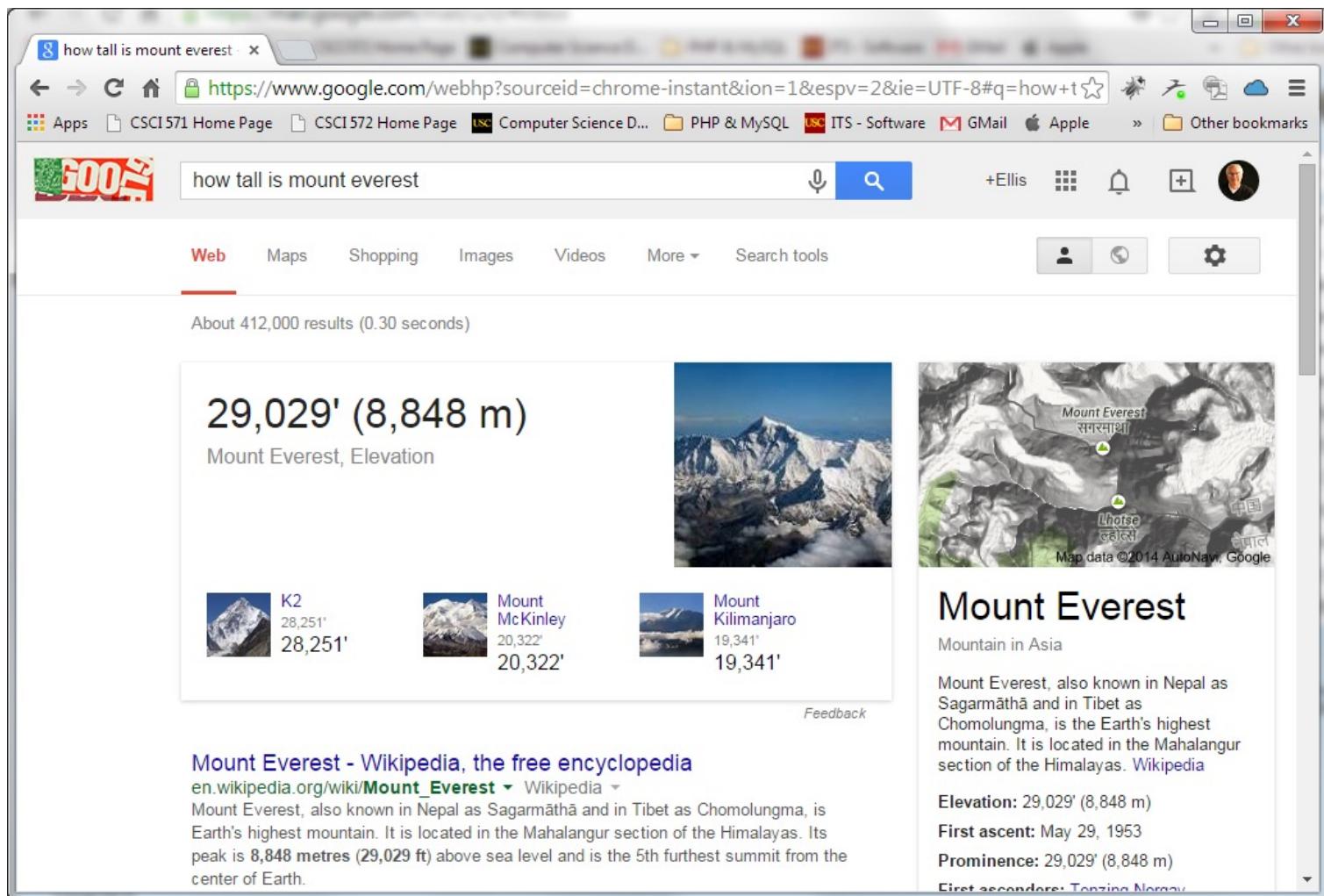
YouTube Spotify Apple Music

More music services

Mariah Carey is an American singer, actress, and record producer. Referred to as the "Songbird Supreme" and the "Queen of Pop", she is noted for her five-octave vocal range, melismatic singing style, and signature whistle register. Carey rose to fame with her eponymous debut album. [Wiki...](#)

Born: March 27, 1969 (age 53 years)  
Children: Moroccan Scott Cannon  
Spouse: Nick Cannon (m. 2008–2014); Mottola (m. 1993–1998)

# Some Questions are Easily Answered



Google search results for "how tall is mount everest".

Search bar: how tall is mount everest

Results:

- 29,029' (8,848 m)**  
Mount Everest, Elevation
-  K2  
28,251'  
28,251'
-  Mount McKinley  
20,322'  
20,322'
-  Mount Kilimanjaro  
19,341'  
19,341'

Feedback

**Mount Everest - Wikipedia, the free encyclopedia**  
[en.wikipedia.org/wiki/Mount\\_Everest](https://en.wikipedia.org/wiki/Mount_Everest) ▾ Wikipedia ▾  
Mount Everest, also known in Nepal as Sagarmāthā and in Tibet as Chomolungma, is Earth's highest mountain. It is located in the Mahalangur section of the Himalayas. Its peak is 8,848 metres (29,029 ft) above sea level and is the 5th furthest summit from the center of Earth.

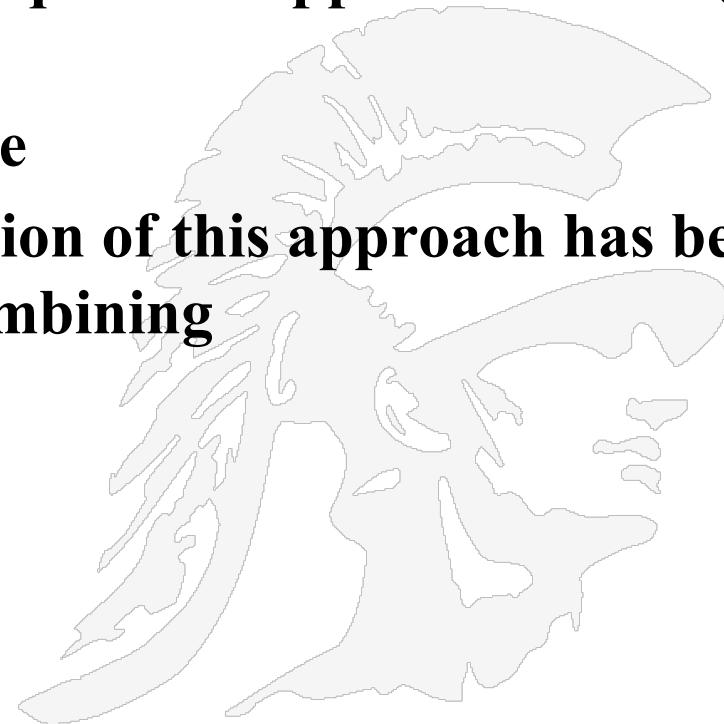
**Mount Everest**  
Mountain in Asia

Mount Everest, also known in Nepal as Sagarmāthā and in Tibet as Chomolungma, is the Earth's highest mountain. It is located in the Mahalangur section of the Himalayas. [Wikipedia](#)

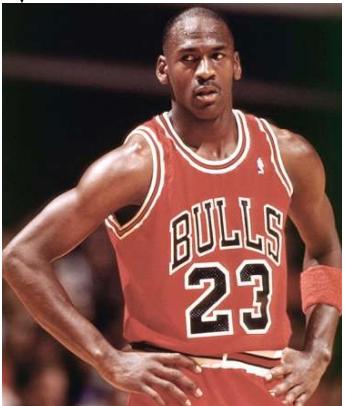
**Elevation:** 29,029' (8,848 m)  
**First ascent:** May 29, 1953  
**Prominence:** 29,029' (8,848 m)  
**First ascender:** Tenzing Norgay

## The Original Google Approach

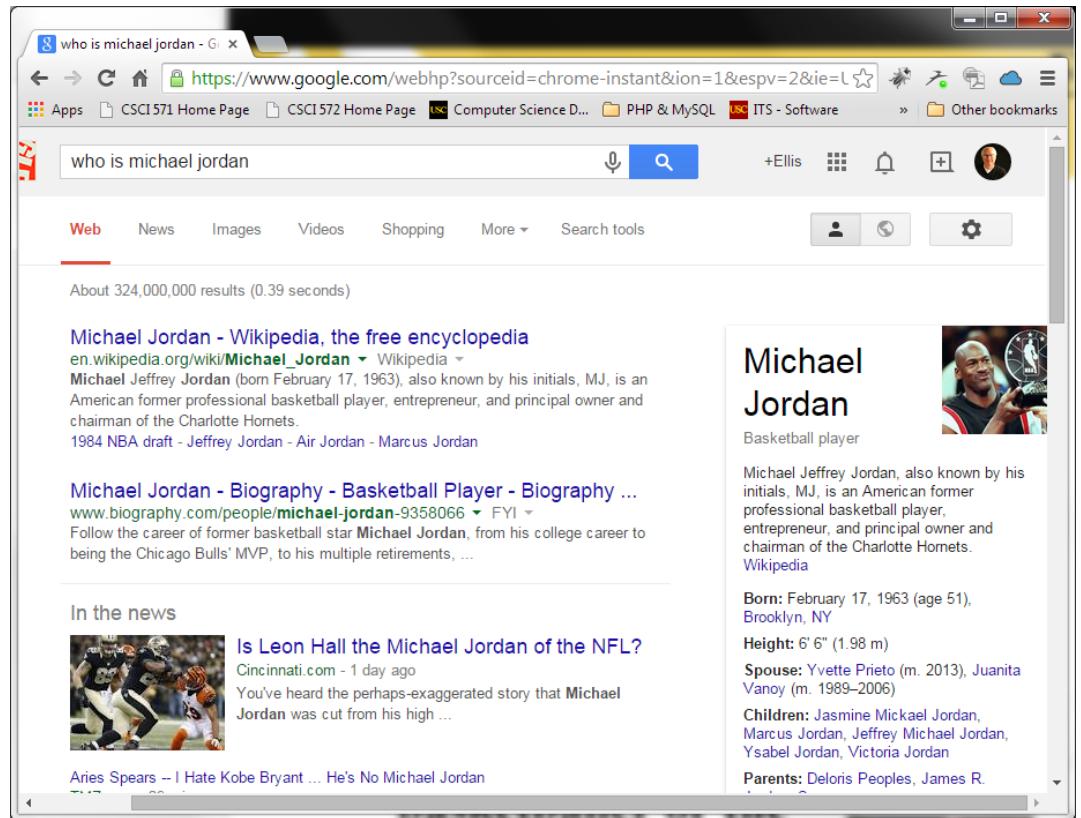
- Take the question and try to find it as a string on the web
- Return the next sentence on that web page as the answer
- Works brilliantly if this exact question appears as a FAQ question, etc.
- Works poorly most of the time
- But a more sophisticated version of this approach has been introduced in recent years combining
  - Knowledge graph
  - N-grams
  - WordNet
  - NLP techniques



- Who is Michael Jordan?
  - Michael Jordan the basketball player or the Machine Learning guy?
- Key requirement is that entities get identified and disambiguated



# Many Questions Pose Semantic Difficulties



Google search results for "who is michael jordan":

Search term: who is michael jordan

About 324,000,000 results (0.39 seconds)

Web

[Michael Jordan - Wikipedia, the free encyclopedia](https://en.wikipedia.org/wiki/Michael_Jordan)  
en.wikipedia.org/wiki/Michael\_Jordan ▾ Wikipedia ▾  
Michael Jeffrey Jordan (born February 17, 1963), also known by his initials, MJ, is an American former professional basketball player, entrepreneur, and principal owner and chairman of the Charlotte Hornets.  
1984 NBA draft - Jeffrey Jordan - Air Jordan - Marcus Jordan

[Michael Jordan - Biography - Basketball Player - Biography ...](https://www.biography.com/people/michael-jordan-9358066)  
www.biography.com/people/michael-jordan-9358066 ▾ FYI ▾  
Follow the career of former basketball star Michael Jordan, from his college career to being the Chicago Bulls' MVP, to his multiple retirements, ...

In the news

 [Is Leon Hall the Michael Jordan of the NFL?](#)  
Cincinnati.com - 1 day ago  
You've heard the perhaps-exaggerated story that Michael Jordan was cut from his high ...

[Aries Spears -- I Hate Kobe Bryant ... He's No Michael Jordan](#)

Michael Jordan

Basketball player

Michael Jeffrey Jordan, also known by his initials, MJ, is an American former professional basketball player, entrepreneur, and principal owner and chairman of the Charlotte Hornets.

Wikipedia

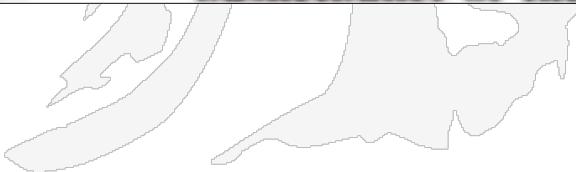
Born: February 17, 1963 (age 51). Brooklyn, NY

Height: 6' 6" (1.98 m)

Spouse: Yvette Prieto (m. 2013), Juanita Vanoy (m. 1989–2006)

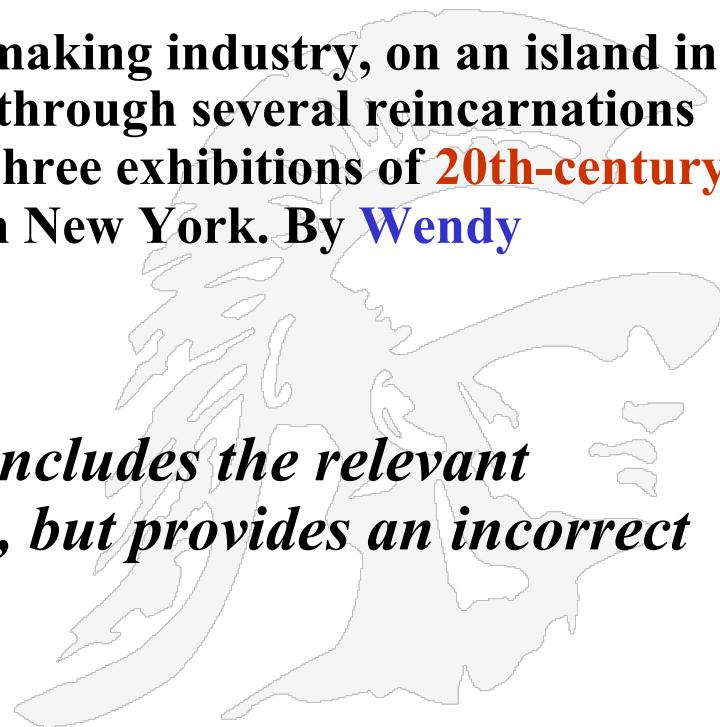
Children: Jasmine Mickael Jordan, Marcus Jordan, Jeffrey Michael Jordan, Ysabel Jordan, Victoria Jordan

Parents: Deloris Peoples, James R.



## Why Natural Language Processing is Required

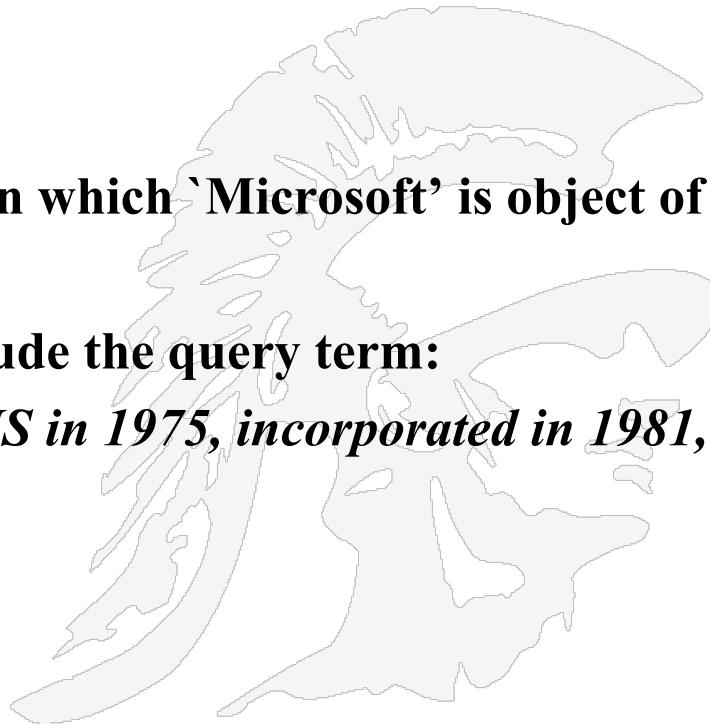
- **Question:** “When was Wendy’s founded?”
- **Passage candidate:**
  - “The renowned Murano glassmaking industry, on an island in the Venetian lagoon, has gone through several reincarnations since it was **founded** in 1291. Three exhibitions of **20th-century** Murano glass are coming up in New York. By **Wendy Moonan**.”
- **Answer:** **20<sup>th</sup> Century**
- *the candidate passage below includes the relevant keywords (Wendy's, founded), but provides an incorrect answer*



# More NLP Challenges

## Predicate-Argument Structure

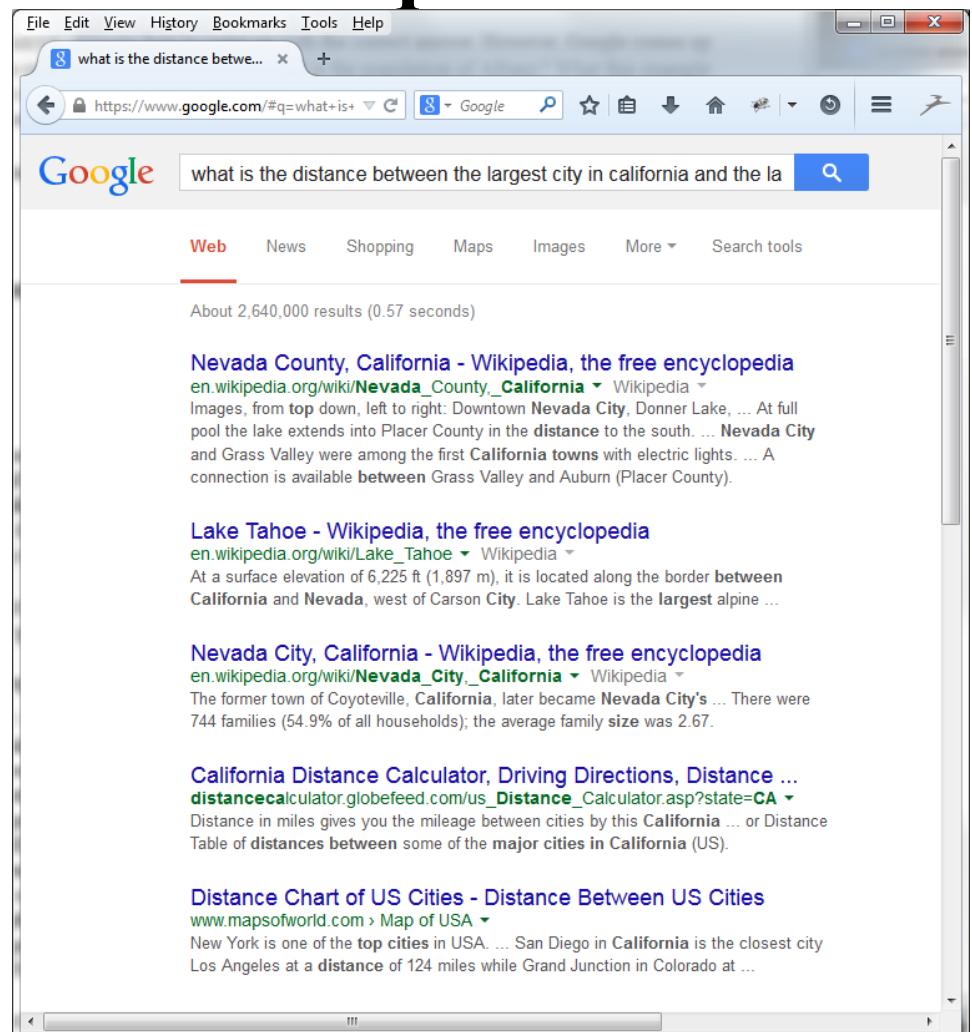
- Q336: *When was Microsoft established?*
- Difficult because Microsoft tends to establish lots of things...  
*Microsoft plans to establish manufacturing partnerships in Brazil and Mexico in May.*
- Need to be able to detect sentences in which 'Microsoft' is object of 'establish' or close synonym.
- A correct result might *not* even include the query term:  
*Microsoft Corp was founded in the US in 1975, incorporated in 1981, and established in the UK in 1982.*



**What is the distance between the largest city in California and the largest city in Nevada?**

Google does poorly on this query, misinterpreting Nevada as Nevada County, California

ps: Try the query in Google today to see if they have improved their answer

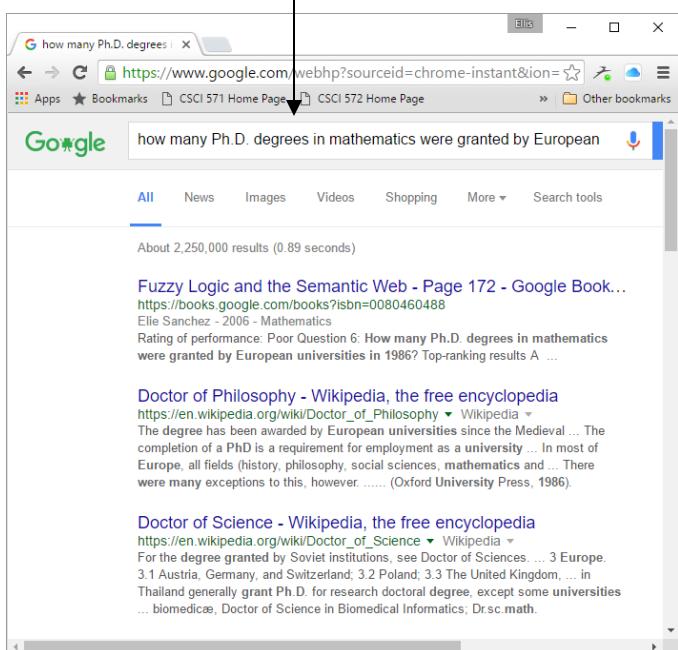


File Edit View History Bookmarks Tools Help  
what is the distance betwe... +  
<https://www.google.com/#q=what+is+the+distance+between+the+largest+city+in+california+and+the+largest+city+in+nevada> Google what is the distance between the largest city in california and the la ...  
Web News Shopping Maps Images More Search tools  
About 2,640,000 results (0.57 seconds)  
**Nevada County, California - Wikipedia, the free encyclopedia**  
[en.wikipedia.org/wiki/Nevada\\_County,\\_California](http://en.wikipedia.org/wiki/Nevada_County,_California) Wikipedia  
Images, from top down, left to right: Downtown Nevada City, Donner Lake, ... At full pool the lake extends into Placer County in the distance to the south. ... Nevada City and Grass Valley were among the first California towns with electric lights. ... A connection is available between Grass Valley and Auburn (Placer County).  
**Lake Tahoe - Wikipedia, the free encyclopedia**  
[en.wikipedia.org/wiki/Lake\\_Tahoe](http://en.wikipedia.org/wiki/Lake_Tahoe) Wikipedia  
At a surface elevation of 6,225 ft (1,897 m), it is located along the border between California and Nevada, west of Carson City. Lake Tahoe is the largest alpine ...  
**Nevada City, California - Wikipedia, the free encyclopedia**  
[en.wikipedia.org/wiki/Nevada\\_City,\\_California](http://en.wikipedia.org/wiki/Nevada_City,_California) Wikipedia  
The former town of Coyoteville, California, later became Nevada City's ... There were 744 families (54.9% of all households); the average family size was 2.67.  
**California Distance Calculator, Driving Directions, Distance ...**  
[distancecalculator.globefeed.com/us\\_Distance\\_Calculator.asp?state=CA](http://distancecalculator.globefeed.com/us_Distance_Calculator.asp?state=CA)  
Distance in miles gives you the mileage between cities by this California ... or Distance Table of distances between some of the major cities in California (US).  
**Distance Chart of US Cities - Distance Between US Cities**  
[www.mapsofworld.com > Map of USA](http://www.mapsofworld.com/Map_of_USA)  
New York is one of the top cities in USA. ... San Diego in California is the closest city Los Angeles at a distance of 124 miles while Grand Junction in Colorado at ...

how many Ph.D. degrees in mathematics were granted by European universities in 1986?

All results are irrelevant

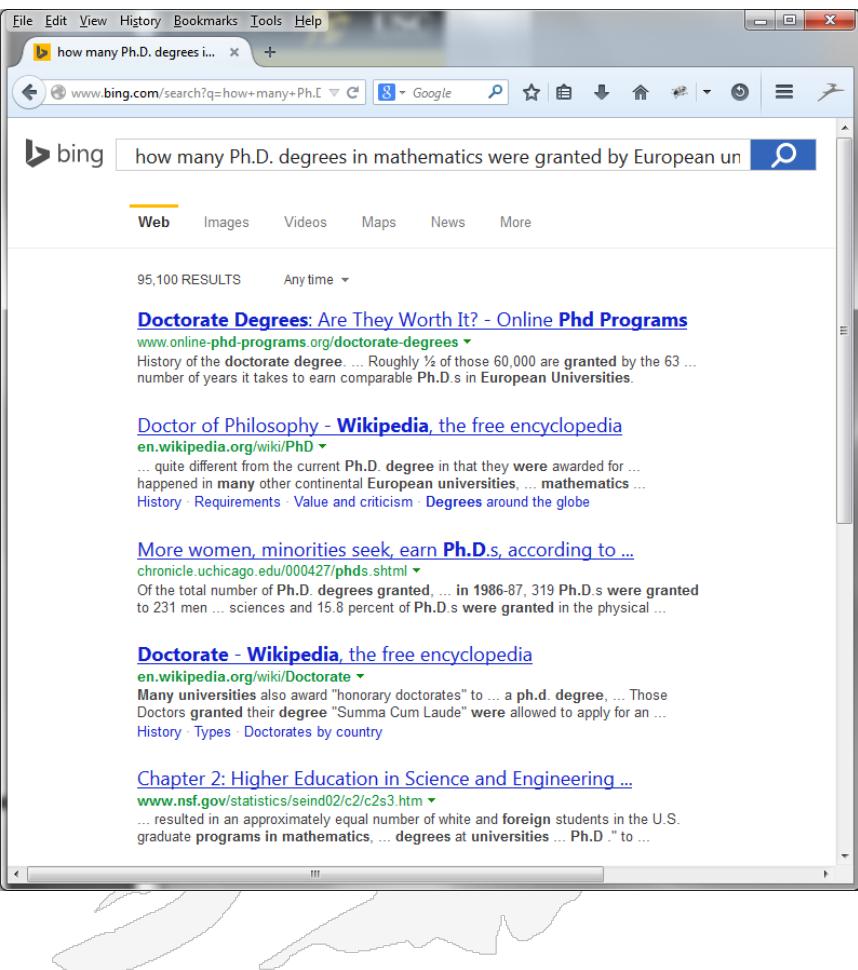
a more recent result; still no relevant links



The screenshot shows a Google search results page with the query "how many Ph.D. degrees in mathematics were granted by European universities in 1986?". The results are as follows:

- Fuzzy Logic and the Semantic Web - Page 172 - Google Book...**  
<https://books.google.com/books?id=0080460488>  
 Elie Sanchez - 2006 - Mathematics  
 Rating of performance: Poor Question 6: How many Ph.D. degrees in mathematics were granted by European universities in 1986? Top-ranking results A ...
- Doctor of Philosophy - Wikipedia, the free encyclopedia**  
[https://en.wikipedia.org/wiki/Doctor\\_of\\_Philosophy](https://en.wikipedia.org/wiki/Doctor_of_Philosophy) ▾ Wikipedia  
 The degree has been awarded by European universities since the Medieval ... The completion of a PhD is a requirement for employment as a university ... In most of Europe, all fields (history, philosophy, social sciences, mathematics and ... There were many exceptions to this, however ..... (Oxford University Press, 1986).
- Doctor of Science - Wikipedia, the free encyclopedia**  
[https://en.wikipedia.org/wiki/Doctor\\_of\\_Science](https://en.wikipedia.org/wiki/Doctor_of_Science) ▾ Wikipedia  
 For the degree granted by Soviet institutions, see Doctor of Sciences. ... 3 Europe. 3.1 Austria, Germany, and Switzerland; 3.2 Poland; 3.3 The United Kingdom, ... in Thailand generally grant Ph.D. for research doctoral degree, except some universities ... biomedicæ, Doctor of Science in Biomedical Informatics, Dr.sc.math.

# In Some Cases the Data May Not Exist

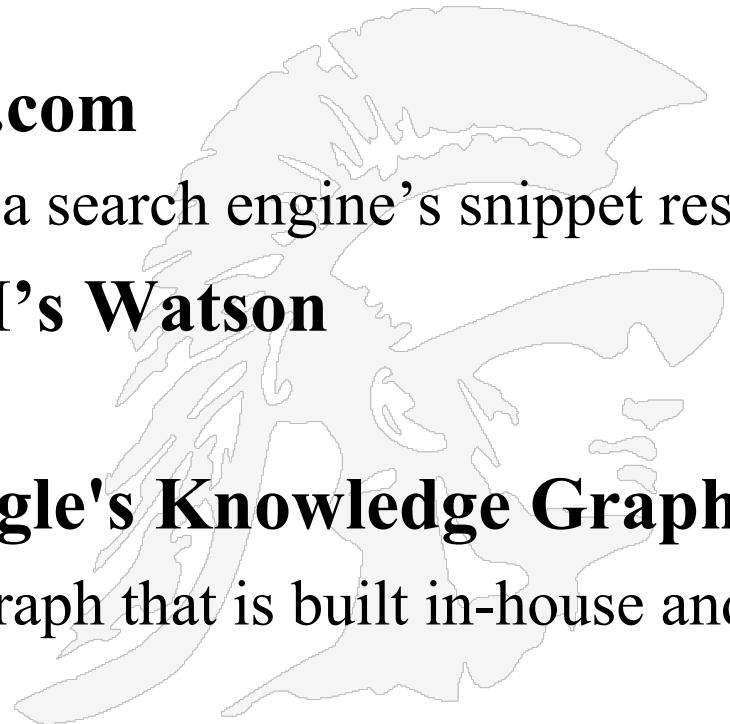


The screenshot shows a Bing search results page with the query "how many Ph.D. degrees in mathematics were granted by European un". The results are as follows:

- Doctorate Degrees: Are They Worth It? - Online Phd Programs**  
[www.online-phd-programs.org/doctorate-degrees](http://www.online-phd-programs.org/doctorate-degrees) ▾  
 History of the doctorate degree. ... Roughly ½ of those 60,000 are granted by the 63 ... number of years it takes to earn comparable Ph.D.s in European Universities.
- Doctor of Philosophy - Wikipedia, the free encyclopedia**  
[en.wikipedia.org/wiki/PhD](http://en.wikipedia.org/wiki/PhD) ▾  
 ... quite different from the current Ph.D. degree in that they were awarded for ... happened in many other continental European universities, ... mathematics ... History · Requirements · Value and criticism · Degrees around the globe
- More women, minorities seek, earn Ph.D.s, according to ...**  
[chronicle.uchicago.edu/00427/phds.shtml](http://chronicle.uchicago.edu/00427/phds.shtml) ▾  
 Of the total number of Ph.D. degrees granted ... in 1986-87, 319 Ph.D.s were granted to 231 men ... sciences and 15.8 percent of Ph.D.s were granted in the physical ...
- Doctorate - Wikipedia, the free encyclopedia**  
[en.wikipedia.org/wiki/Doctorate](http://en.wikipedia.org/wiki/Doctorate) ▾  
 Many universities also award "honorary doctorates" to ... a ph.d. degree, ... Those Doctors granted their degree "Summa Cum Laude" were allowed to apply for an ... History · Types · Doctorates by country
- Chapter 2: Higher Education in Science and Engineering ...**  
[www.nsf.gov/statistics/seind02/c2/c2s3.htm](http://www.nsf.gov/statistics/seind02/c2/c2s3.htm) ▾  
 ... resulted in an approximately equal number of white and foreign students in the U.S. graduate programs in mathematics, ... degrees at universities ... Ph.D." to ...

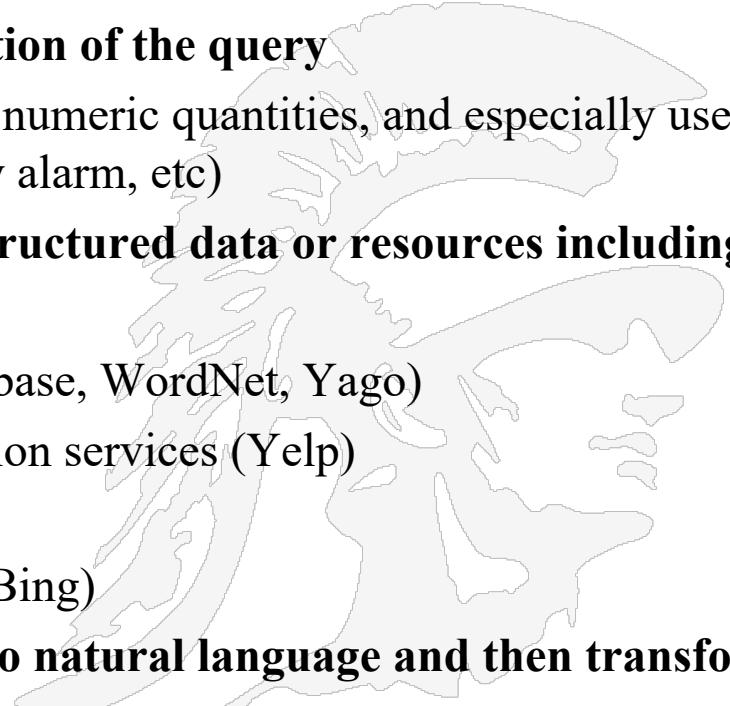
# Some Popular Products Designed for Question/Answering

- **Approach 1: used by Siri**
  - map to known entities and use existing databases over the internet
- **Approach 2: used by Ask.com**
  - detect question type and use a search engine's snippet results
- **Approach 3: used by IBM's Watson**
  - combine approaches 1 and 2
- **Approach 4: used by Google's Knowledge Graph**
  - use an entity - relationship graph that is built in-house and infer the answer



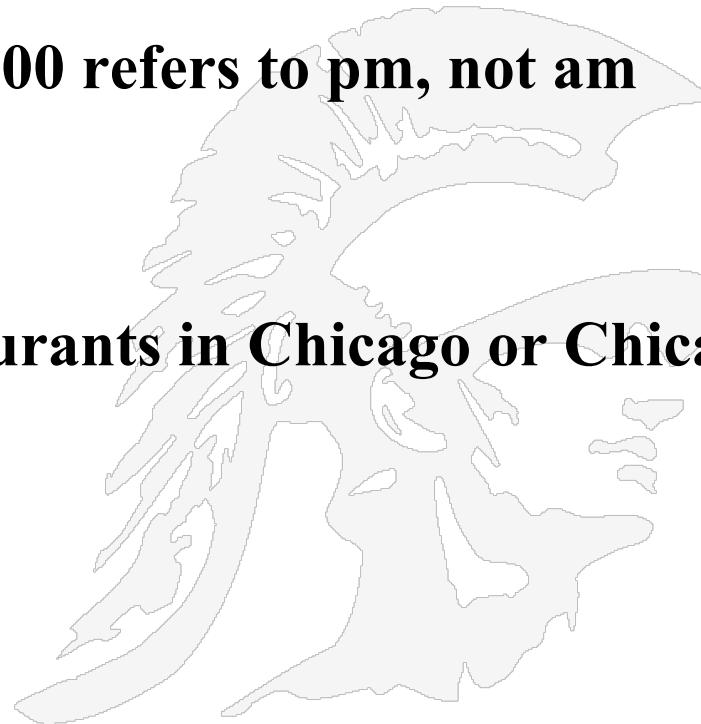
## Approach used by Siri: Knowledge-Based Approach

- Siri was begun as a DARPA project called CALO/PAL (Personalized Assistant that Learns)
  1. First your voice query is put through a recognizer and a language model and Siri comes up with an interpretation of what was said
  2. Second Siri builds a semantic representation of the query
    - Extract times, dates, locations, entities, numeric quantities, and especially user actions (e.g. schedule a meeting, set my alarm, etc)
  3. Siri maps from this semantics to query structured data or resources including:
    - Geospatial databases
    - Ontologies (Wikipedia infoboxes, Freebase, WordNet, Yago)
    - Restaurant review sources and reservation services (Yelp)
    - Scientific databases (Wolfram Alpha)
    - Conventional search engines (Google, Bing)
  4. Siri then transforms the output above into natural language and then transforms the text back to speech



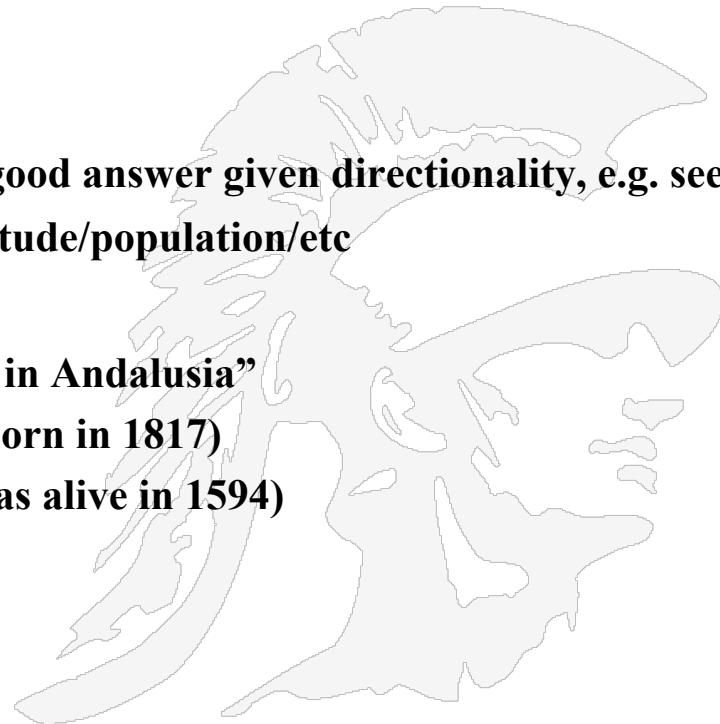
## Context and Conversation in Virtual Assistants like Siri

- Coreference helps resolve ambiguities
- U: “book a table at Il Fornaio at 7:00 with my mom”
- U: “also send her an email reminder”
- “her” refers to “my mom”; 7:00 refers to pm, not am
- Clarification questions:
- U: “chicago pizza”
- S: “Did you mean pizza restaurants in Chicago or Chicago-style pizza?”



# CANDIDATE ANSWER SCORING IN IBM WATSON

- Each candidate answer gets scores from > 50 components
- From unstructured text, semi-structured text, triple stores
- Logical form (parse) match between question and candidate
- Passage source reliability
- Geospatial location
  - Denver is “southwest of Montana” is a good answer given directionality, e.g. see
  - geonames.org which gives latitude/longitude/population/etc
- Temporal relationships
  - “In 1594 he took a job as a tax collector in Andalusia”
  - Candidates: Thoreau is a bad answer (born in 1817)
  - Candidates: Cervantes is possible (he was alive in 1594)
- Taxonomic classification



# AskJeeves (now Ask.com)

- Earlier AskJeeves.com was well-known as a search engine specializing in Questions/Answers
- Though it still exists, it performs far weaker than sites such as Google

how old is mariah carey

Ad - <https://music.amazon.com/>

**Mariah Carey on Amazon Music Unlimited**

★★★★★ rating for amazon.com

Stream ad-free music, podcasts, artist live-streams and more! Try now. Starts at \$7.99/month after. New subscribers only. Terms apply. Try it free. Any song, anywhere. The HD difference. Alexa Voice Controls. Prime Member Discounts. Unlimited Skips. Experience spatial audio. Styles: Hip-Hop, Rock.

**All Hits Playlist**  
The Biggest Songs in the World. Updated Fridays. Stream Now

**Free Music Streaming**  
No credit card needed. Try Amazon Music Free

**Rock Arena Playlist**  
Play It Loud! Updated Fridays Curated by Amazon's Music Experts

**Pop Culture Playlist**  
The Ultimate Stop for Today's Pop Curated by Amazon's Music Experts.

Ad - <https://www.costumes.com/>

**The Mariah Carey Holiday Collection**

We stock the latest costume collections. High quality & realistic. Costumes from the latest films, classic films, superheroes, horror & loads of kids' outfits. Multiple payment options. Track your order. Sign up for offers. View size charts.

New & Popular: Hair & Wigs | Kids & Cartoons | Comics & Superheroes | Horror Collections

PRODUCT ADS FROM 

How old is Mariah Carey  
Snapshot taken 04/2022

how tall is mount everest

1-10 of 40,900,000 results

**Mount Everest Height - Find It Here**

Search for Mount Everest Height. Smart Results Here. BestDiscoveries. Search Everything You Need. Save Time & Get Quick Results. Better Results. Find Right Now. Useful Info. Find More. Multi Search.

**The Best Nepal Tour Operator - Heaven Nepal Adventure - Local Tour Operator**

Everest Base Camp Trail Itinerary is our standard itinerary. Heaven Nepal Adventure | Trekking Agency in Nepal.

**Climb High Himalaya - Everest Base Camp Treks**

Climb High Himalaya Offers Trekking to Everest, Annapurna and everywhere in Nepal. Trekking in Nepal for the Adventurer - Trek in the Himalayas.

**Nepal Mustang**      **Reviews**  
**Barunse Expedition**      **Lata Rajen Thapa**  
**Services**      **Reservation**

Ad - <https://www.walmart.com/biography-&-memoirs>

**Biography & Memoirs**

★★★★★ rating for walmart.com

Save On Biography & Memoirs. Free Shipping Site to Store. Free In-store pickup. Best selling books. Top brands - low prices. Free shipping over \$50. Highlights: App Available, Store Directory Available.

6433 Fallbrook Ave, West Hills, CA

How tall is mount Everest  
Snapshot taken 04/2022

# Question Types: Many Questions Fall into Distinct Categories

Who	Person, Organization
When	Date, Year
Where	Location
In What	Location
How many	Number

# 3 Main Phases for Question/Answering

## 1. QUESTION PROCESSING

- Detect question type (who, what, when, where, etc)
- Identify important entities and formulate queries to send to a search engine

## 2. PASSAGE RETRIEVAL

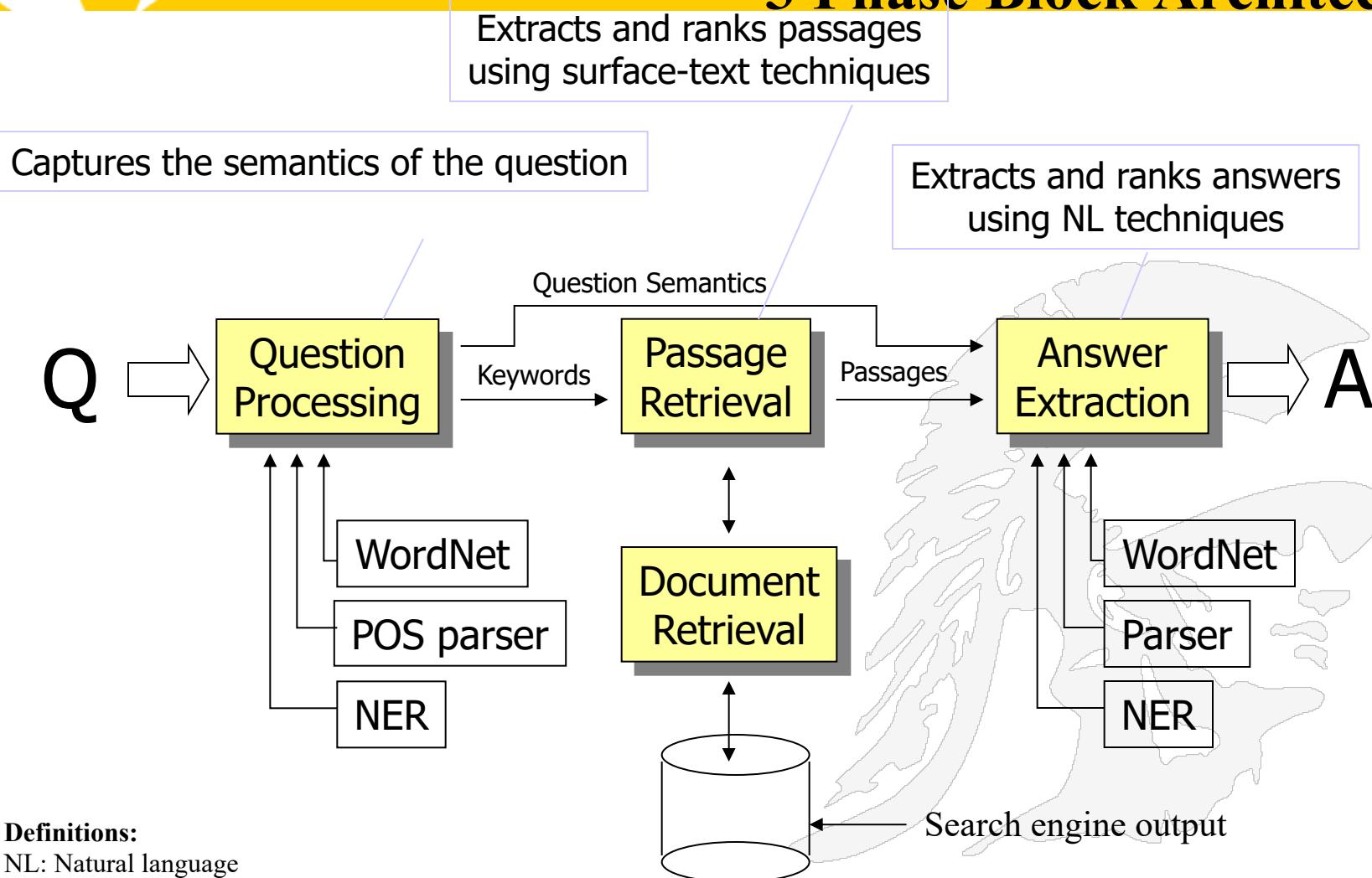
- Retrieve ranked documents (snippets only)
- Break into suitable passages and match against entities

## 3. ANSWER PROCESSING

- Extract candidate answers (as named entities)
- Rank candidates
  - **using evidence from relations in the text and external sources**

# Question Answering

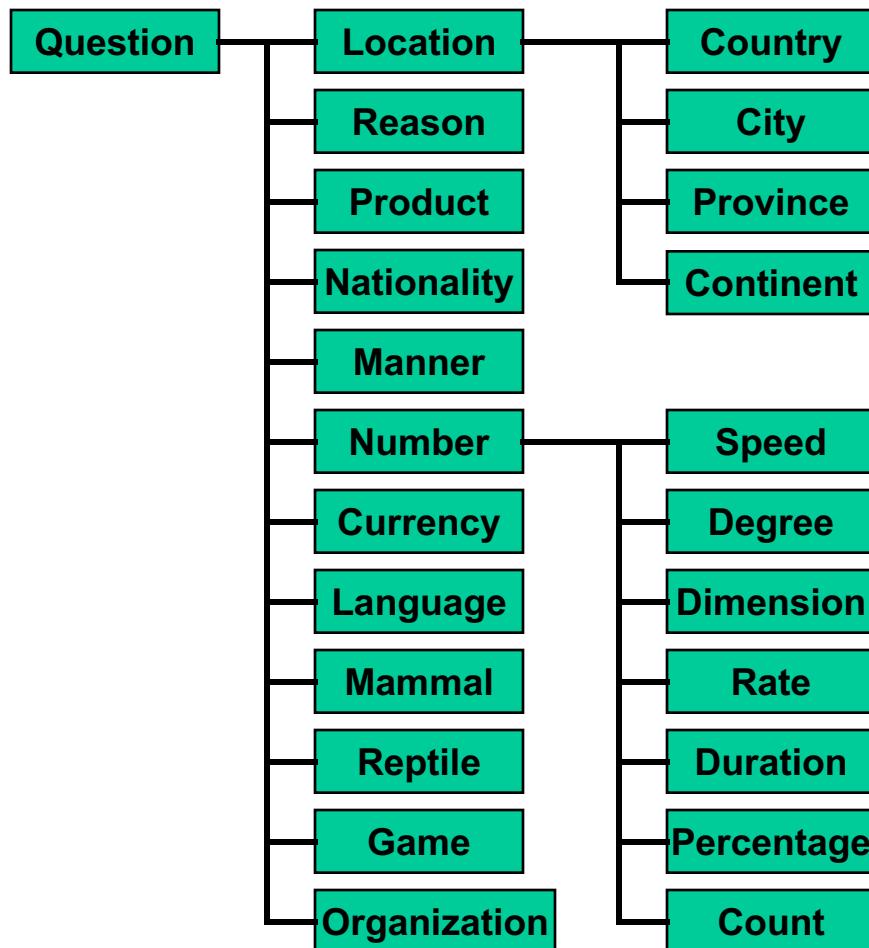
## 3 Phase Block Architecture

**Definitions:**

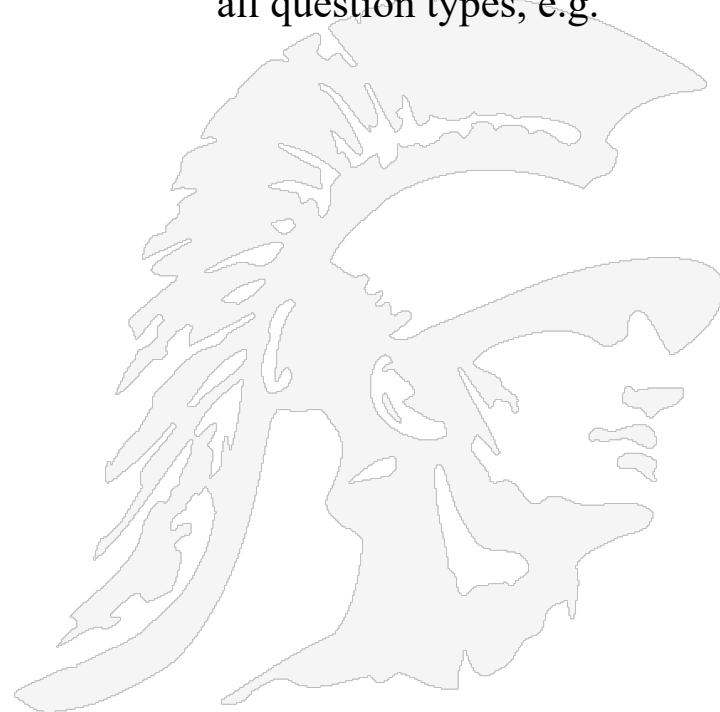
NL: Natural language

NER: Named Entity Recognition

# Question Taxonomy



Researchers have tried to organize all question types, e.g.



## Factoid Questions

- Who, where, when, how many
- The answers fall into a limited and somewhat predictable set of categories, see e.g. to the right
- This set is from a paper by Li and Roth, 2005, *Learning Question Classifiers*
- Their major categories include
  - Abbreviation
  - Description
  - Entity
  - Human
  - Location
  - Numeric
- Questions can be labeled at one or two levels, e.g. either as Numeric or as Numeric:date or as Location or Location:mountain
- This approach argues for a template applied to the query

# However Question Taxonomies Can Get Very Large

Tag	Example
ABBREVIATION	
abb	What's the abbreviation for limited partnership?
exp	What does the "c" stand for in the equation E=mc <sup>2</sup> ?
DESCRIPTION	
definition	What are tannins?
description	What are the words to the Canadian National anthem?
manner	How can you get rust stains out of clothing?
reason	What caused the Titanic to sink ?
ENTITY	
animal	What are the names of Odin's ravens?
body	What part of your body contains the corpus callosum?
color	What colors make up a rainbow ?
creative	In what book can I find the story of Aladdin?
currency	What currency is used in China?
disease/medicine	What does Salk vaccine prevent?
event	What war involved the battle of Chapultepec?
food	What kind of nuts are used in marzipan?
instrument	What instrument does Max Roach play?
lang	What's the official language of Algeria?
letter	What letter appears on the cold-water tap in Spain?
other	What is the name of King Arthur's sword?
plant	What are some fragrant white climbing roses?
product	What is the fastest computer?
religion	What religion has the most members?
sport	What was the name of the ball game played by the Mayans?
substance	What fuel do airplanes use?
symbol	What is the chemical symbol for nitrogen?
technique	What is the best way to remove wallpaper?
term	How do you say " Grandma " in Irish?
vehicle	What was the name of Captain Bligh's ship?
word	What's the singular of dice?

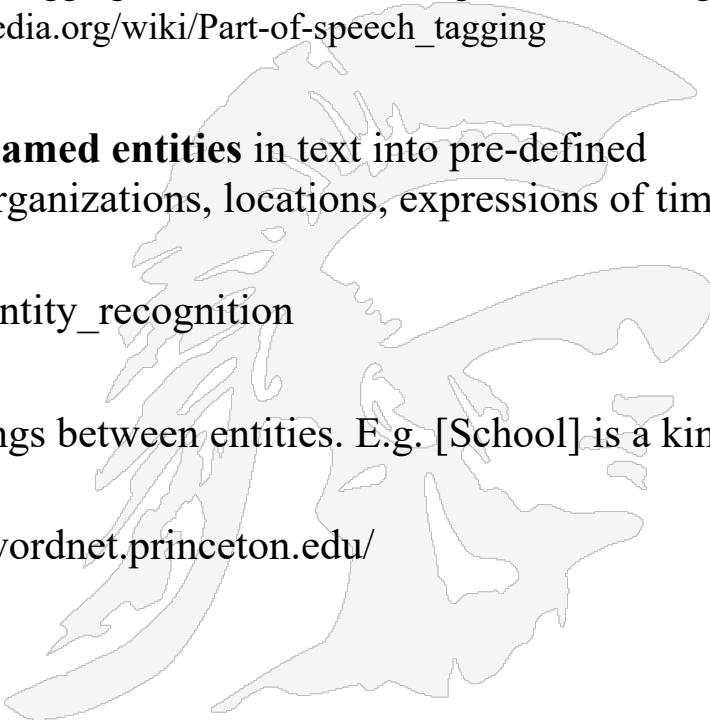


# More Question Types and Examples

HUMAN	
description	Who was Confucius?
group	What are the major companies that are part of Dow Jones?
ind	Who was the first Russian astronaut to do a spacewalk?
title	What was Queen Victoria's title regarding India?
LOCATION	
city	What's the oldest capital city in the Americas?
country	What country borders the most others?
mountain	What is the highest peak in Africa?
other	What river runs through Liverpool?
state	What states do not have state income tax?
NUMERIC	
code	What is the telephone number for the University of Colorado?
count	About how many soldiers died in World War II?
date	What is the date of Boxing Day?
distance	How long was Mao's 1930s Long March?
money	How much did a McDonald's hamburger cost in 1963?
order	Where does Shanghai rank among world cities in population?
other	What is the population of Mexico?
period	What was the average life expectancy during the Stone Age?
percent	What fraction of a beaver's life is spent swimming?
speed	What is the speed of the Mississippi River?
temp	How fast must a spacecraft travel to escape Earth's gravity?
size	What is the size of Argentina?
weight	How many pounds are there in a stone?

# Some General Capabilities for Question-Answering Systems

- **Part-of-Speech Tagging**
  - a piece of software that reads text in some language and assigns **parts of speech** to each word, such as noun, verb, adjective, etc.
  - Markov Models are now the standard method for part-of-speech assignment
  - Some current major algorithms for part-of-speech tagging include the Viterbi algorithm, Brill tagger, and Baum-Welch algorithm, see [https://en.wikipedia.org/wiki/Part-of-speech\\_tagging](https://en.wikipedia.org/wiki/Part-of-speech_tagging)
- **Named Entity Extraction**
  - Software that seeks to locate and classify **named entities** in text into pre-defined categories such as the **names** of persons, organizations, locations, expressions of times, quantities ...
  - See [https://en.wikipedia.org/wiki/Named-entity\\_recognition](https://en.wikipedia.org/wiki/Named-entity_recognition)
- **Determining Semantic Relations**
  - **semantic relations** are concepts or meanings between entities. E.g. [School] is a kind of [educational institution]
    - Opportunity to use WordNet, <https://wordnet.princeton.edu/>
- **Dictionaries/Thesauri**



# Question Processing Tool

## Part-of-Speech Recognizer

WP:Wh-pronoun (who/what/when/where)

VBD:Verb, past tense

DT:Determiner

JJ:Adjective

NNP:proper noun, singular

VB:verb

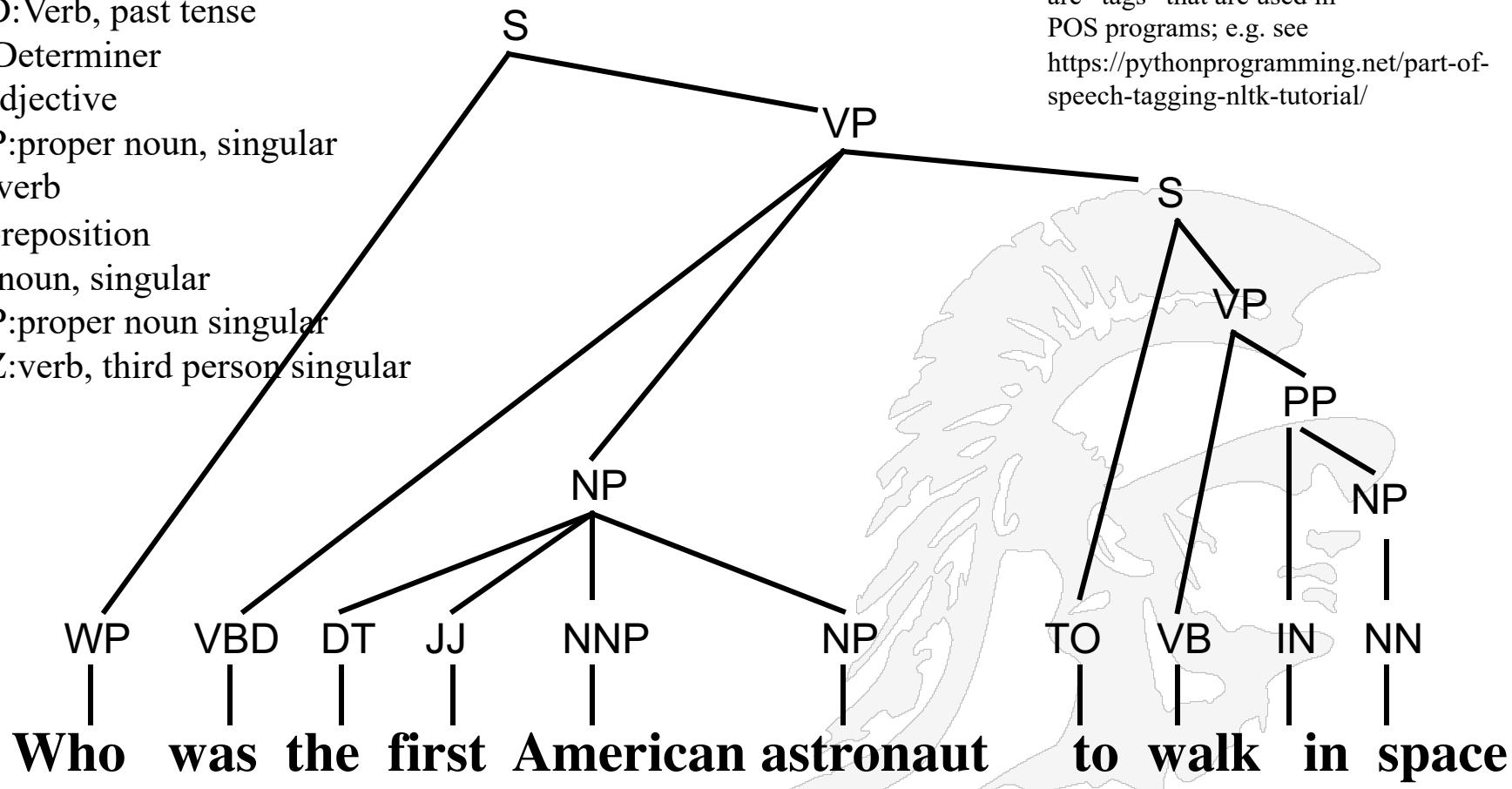
IN:preposition

NN:noun, singular

NNP:proper noun singular

VBZ:verb, third person singular

The two-letter abbreviations are “tags” that are used in POS programs; e.g. see  
<https://pythonprogramming.net/part-of-speech-tagging-nltk-tutorial/>



# Question Processing Tool

## Named Entity Recognizer Example

contentSkip to site indexPoliticsSubscribeLog InSubscribeLog InToday's PaperAdvertisementSupported ORG byF.B.I. Agent Peter Strzok PERSON ,

Who Criticized Trump PERSON in Texts, Is FiredImagePeter Strzok, a top F.B.I. GPE counterintelligence agent who was taken off the special counsel investigation after his disparaging texts about President Trump PERSON were uncovered, was fired. CreditT.J. Kirkpatrick PERSON for The New York TimesBy Adam Goldman ORG and Michael S. SchmidtAug PERSON . 13 CARDINAL , 2018WASHINGTON CARDINAL — Peter Strzok PERSON , the F.B.I. GPE senior counterintelligence agent who disparaged President Trump PERSON in inflammatory text messages and helped oversee the Hillary Clinton PERSON email and Russia GPE investigations, has been fired for violating bureau policies, Mr. Strzok PERSON 's lawyer said Monday DATE .Mr. Trump and his allies seized on the texts — exchanged during the 2016 DATE campaign with a former F.B.I. GPE lawyer, Lisa Page — in PERSON assailing the Russia GPE investigation as an illegitimate "witch hunt." Mr. Strzok PERSON , who rose over 20 years DATE at the F.B.I. GPE to become one of its most experienced counterintelligence agents, was a key figure in the early months DATE of the inquiry.Along with writing the texts, Mr. Strzok PERSON was accused of sending a highly sensitive search warrant to his personal email account.The F.B.I. GPE had been under immense political pressure by Mr. Trump PERSON to dismiss Mr. Strzok PERSON , who was removed last summer DATE from the staff of the special counsel, Robert S. Mueller III PERSON . The president has repeatedly denounced Mr. Strzok PERSON in posts on

- The process of recognizing information units like names, including persons, organizations, location names, and numeric expressions including time, date, money and percent expressions from unstructured text.
- This is an example of supervised learning as training sets are first created
- See <https://nlp.stanford.edu/software/> for a Java program and explanation

## NER Example and Its Translation Jeopardy Example

The Jeopardy query:

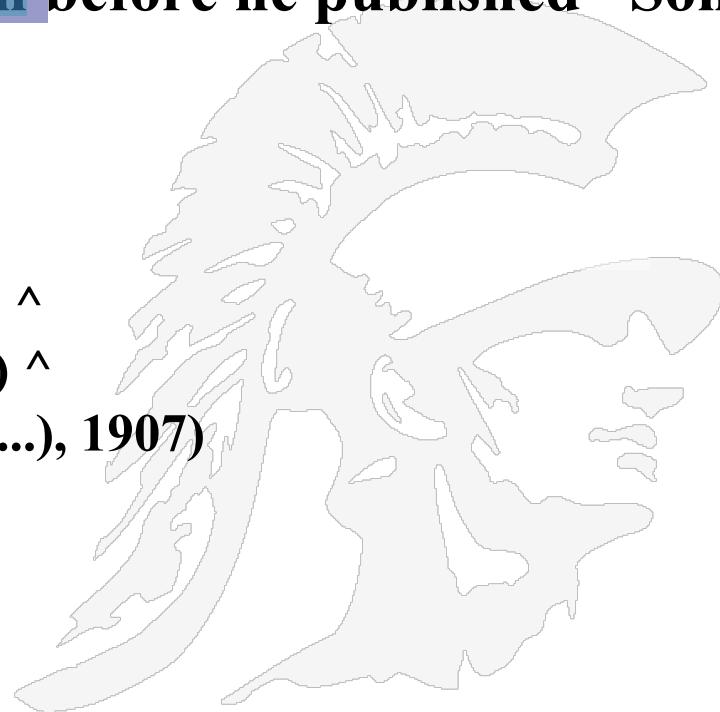
Category: Poets and Poetry: *GEO*

He was a bank clerk in the Yukon before he published “Songs of a Sourdough” in 1907.

*YEAR*

Produces the logic formula:

authorof(focus, “Songs of a Sourdough”) ^  
publish (e1, he, “Songs of a Sourdough”) ^  
in (e2, e1, 1907) ^ temporallink(publish(...), 1907)



# Extracting Candidate Answers from Triple Stores

- Once we extract a relation from the question, e.g.  
**... he published “Songs of a sourdough”**  
**(author-of ?x “Songs of a sourdough”)**
- Many information sources support querying via a triple store
  - Wikipedia infoboxes, DBpedia, FreeBase, etc.
  - author-of(“Songs of a Sourdough”, “Robert Service”)



The screenshot shows a Wikipedia page for "Songs of a Sourdough". The page includes a sidebar with links like Main page, Contents, and Random article. The main content discusses the book's publication history and its connection to the Klondike Gold Rush. A large watermark of a map of North America is visible across the page.

WIKIPEDIA  
The Free Encyclopedia

Main page  
Contents  
Current events  
Random article  
About Wikipedia  
Contact us  
Donate  
Contribute  
Help  
Learn to edit  
Community portal  
Recent changes  
Upload file  
Tools  
What links here

Article Talk Read Edit View history Search Wikipedia

## Songs of a Sourdough

From Wikipedia, the free encyclopedia

For the 1916 film, see *The Spell of the Yukon (film)*.

**Songs of a Sourdough** is a book of poetry published in 1907 by Robert W. Service. In the United States, the book was published under the title *The Spell of the Yukon and Other Verses*. The book is well known for its verse about the Klondike Gold Rush in the Yukon a decade earlier, particularly the long, humorous ballads, "The Shooting of Dan McGrew" and "The Cremation of Sam McGee."

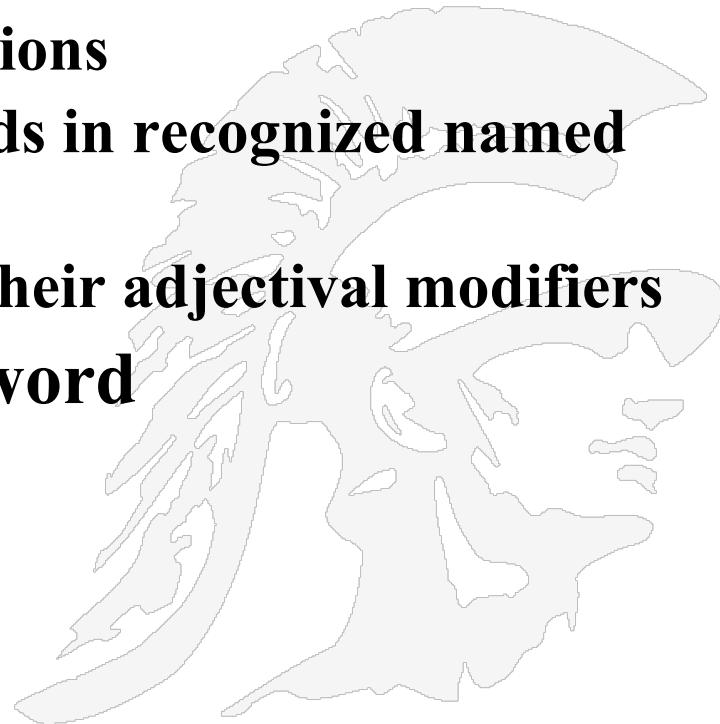
*Songs of a Sourdough* has sold more than three million copies.<sup>[1]</sup>

Contents [hide]

- 1 History
- 2 Contents
- 3 References
- 4 External links

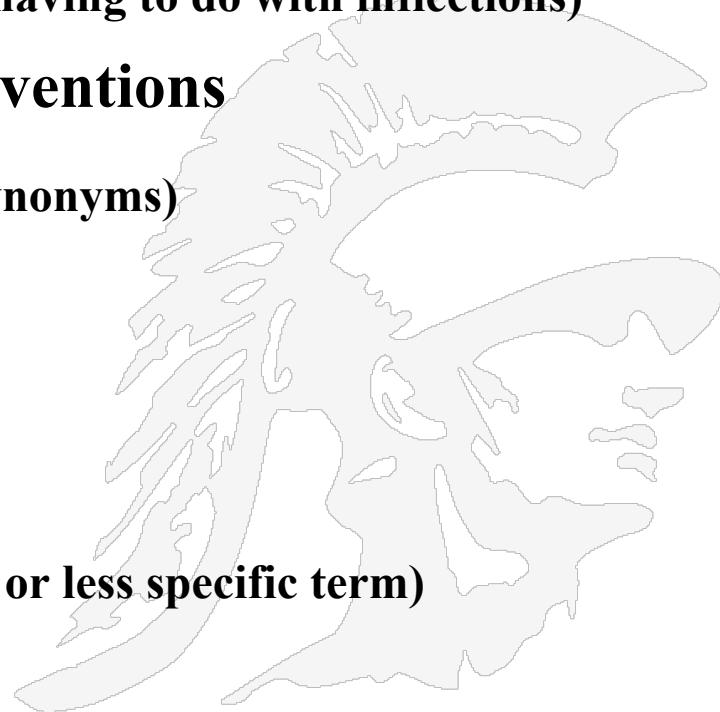
# General Keyword Selection Algorithm

1. Use the part-of-speech recognizer to identify all
  - nouns
  - verbs
  - non-stopwords in quotations
  - NNP (proper noun) words in recognized named entities
  - complex nominals with their adjectival modifiers
2. Select the answer type word



# Expanding the Keyword Set Using Variants

- There are 3 distinct ways to expand the keyword set determined by the keyword selection algorithm
- Morphological variants (having to do with inflections)
  - invented → inventor → inventions
- Lexical variants (similar to synonyms)
  - killer → assassin
  - far → distance
- Semantic variants
  - like → prefer (a more specific or less specific term)



## How to Incorporate Lexical Variants Using Hypernins and Hyponims

**Question:** When was the internal combustion engine invented?

**Answer:** The first internal combustion engine was built in 1867.

*Lexical chains:*

- (1) invent:v#1 → HYPERNIM → create by mental\_act:v#1 →  
HYPERNIM → create:v#1 → HYPONIM → build:v#1

**Question:** How many chromosomes does a human zygote have?

**Answer:** 46 chromosomes lie in the nucleus of every normal human cell.

*Lexical chains:*

- (1) zygote:n#1 → HYPERNIM → cell:n#1 → HAS.PART →  
nucleus:n#1

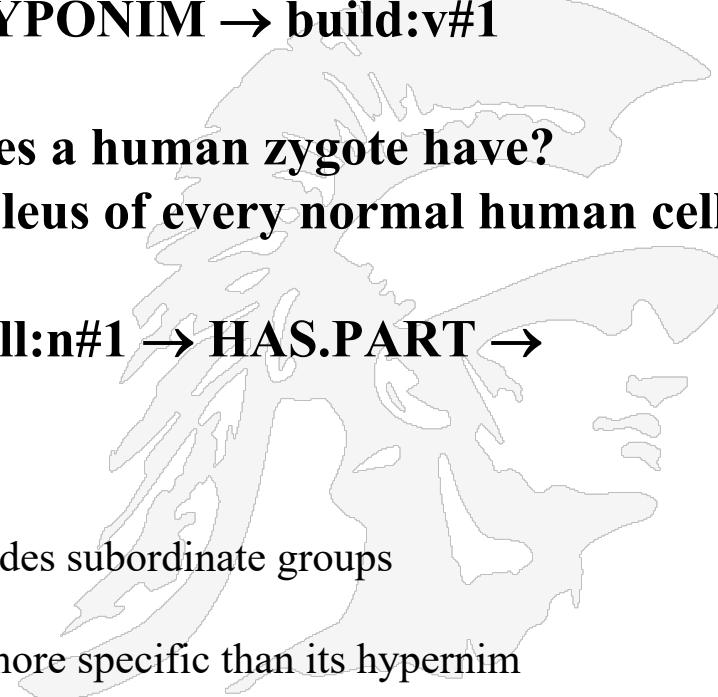
*WordNet provides hypernims and hyponims*

**Hypernym** is a superordinate grouping which includes subordinate groups

e.g. a musical instrument is a hypernym of guitar;

**Hyponim** is a word or phrase whose semantics is more specific than its hypernym

e.g. purple is a hyponym of color

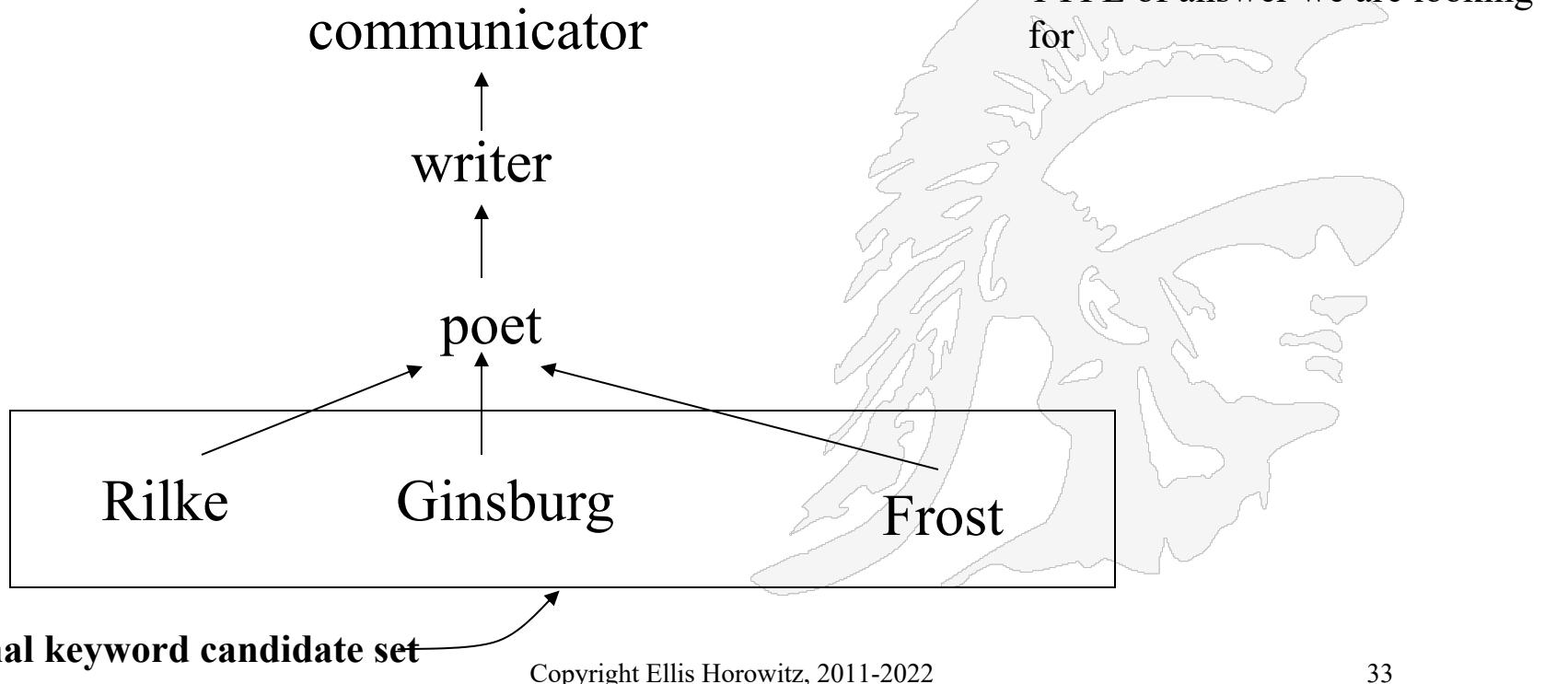


## Use WordNet for Type Identification

We have already seen the use of WordNet, a lexical database of English nouns, verbs, adjectives, adverbs

**“What 20<sup>th</sup> century poet wrote Howl?”**

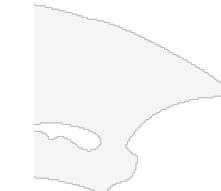
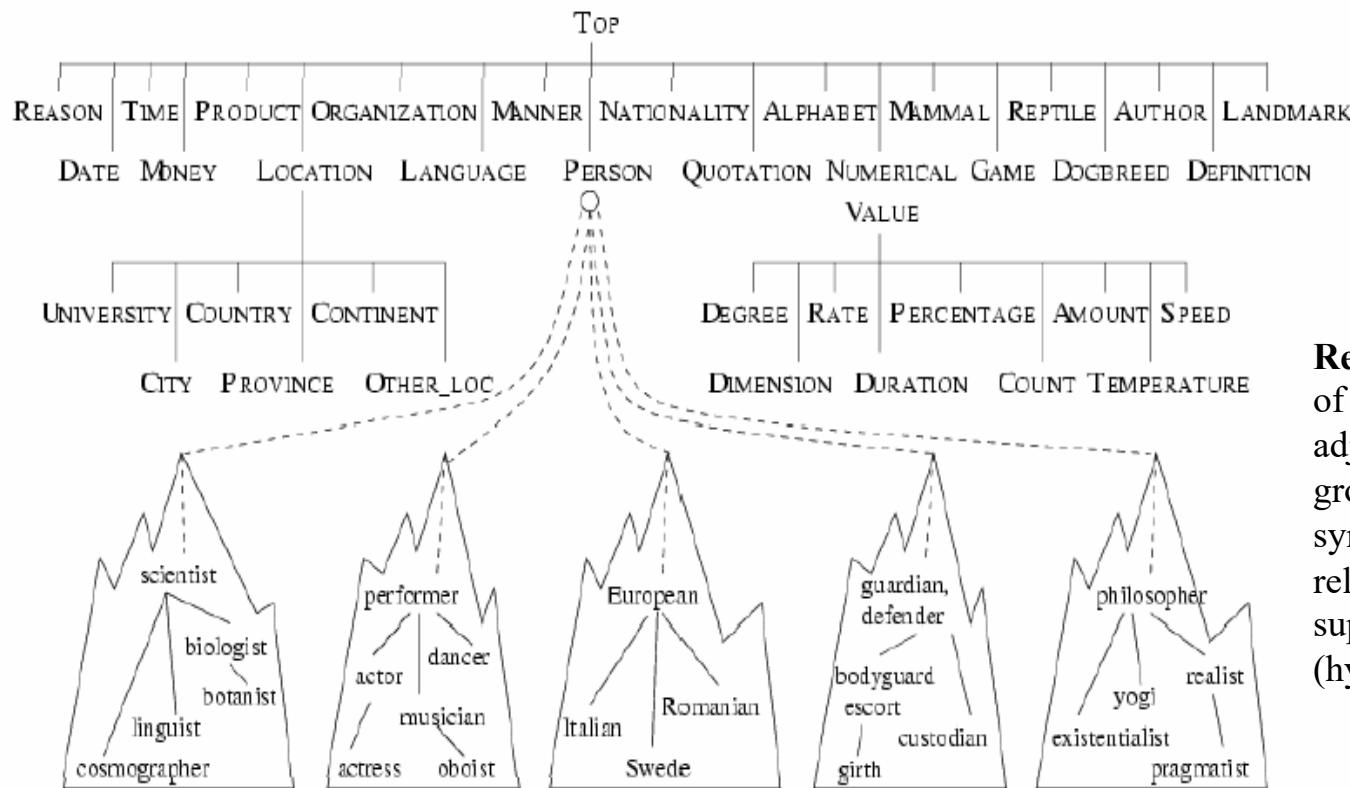
WordNet permits refinement  
of poet to specific instances





# Answer Type Taxonomy

- Use WordNet to merge named entities with the WordNet hierarchy

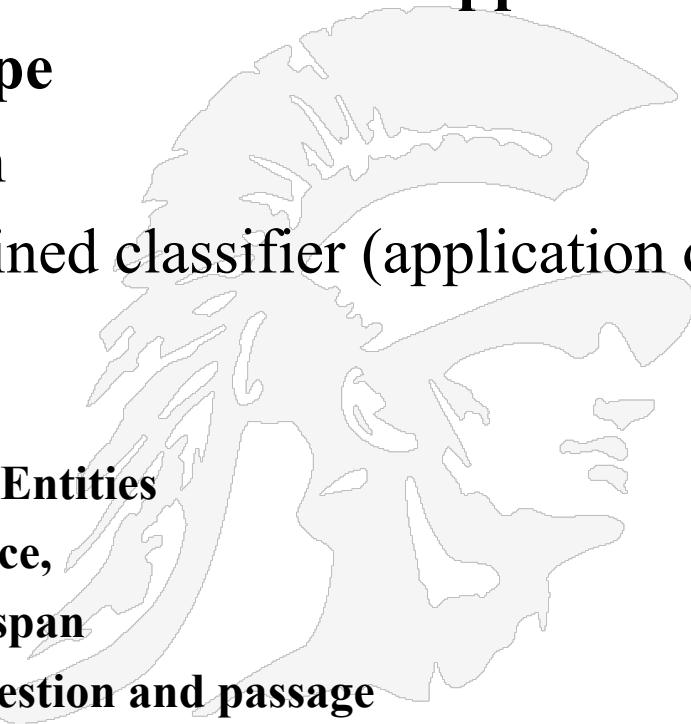


**Recall:** WordNet is a database of English nouns, verbs, adjectives and adverbs grouped into sets of synonyms (synsets) and relations showing super/subordinate relations (hyperonymy/hyponymy)

If you know the answer should be a person  
 WordNet helps determine what sort of person

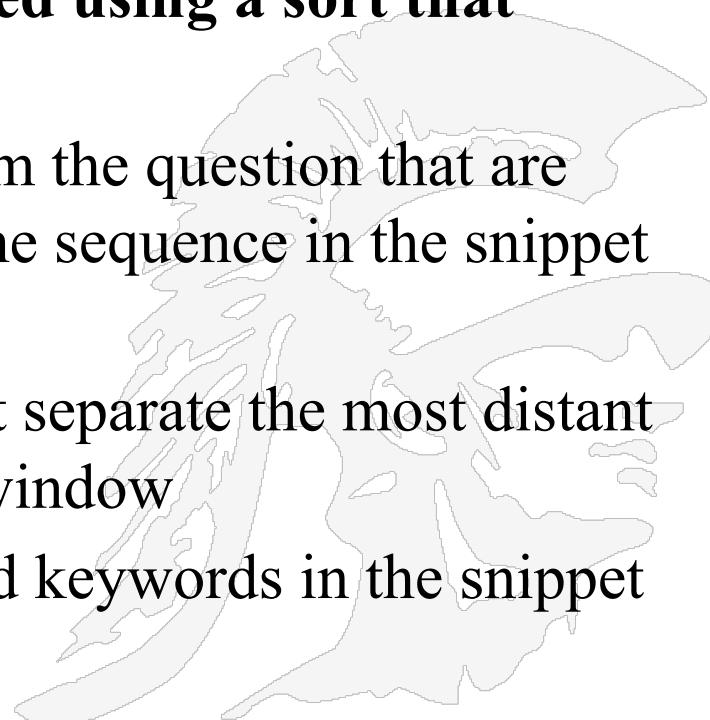
## Part 2: Passage Retrieval

- Once we have formulated queries using tools like NER, POS, variant expansion and WordNet results
- Send queries to a search engine and retrieve snippet results
- Filter the results for correct type
  - use answer type classification
  - Rank passages based on a trained classifier (application of machine learning)
    - Features:
      - Question keywords, Named Entities
      - Longest overlapping sequence,
      - Shortest keyword-covering span
      - N-gram overlap between question and passage



# Passage Scoring Method

- **Focus on the snippets that are returned, the answers must be extracted from them**
- **Passage ordering is performed using a sort that involves three scores:**
  1. The number of words from the question that are recognized and in the same sequence in the snippet window
  2. The number of words that separate the most distant keywords in the snippet window
  3. The number of unmatched keywords in the snippet window



# Ranking Candidate Answers

Q066: Name the first private citizen to fly in space.

- Answer type: Person
- Text passage:

“Among them was Christa McAuliffe, the first private citizen to fly in space. Karen Allen, best known for her starring role in “Raiders of the Lost Ark”, plays McAuliffe. Brian Kerwin is featured as shuttle pilot Mike Smith...”

## Scoring

There are five words from the question “the first private citizen space”

The answer is adjacent to “the first private citizen. . . ”

There are no unmatched keywords in “the first private citizen. . . ”



# Ranking Candidate Answers

Q066: Name the first private citizen to fly in space.

■ Answer type: Person

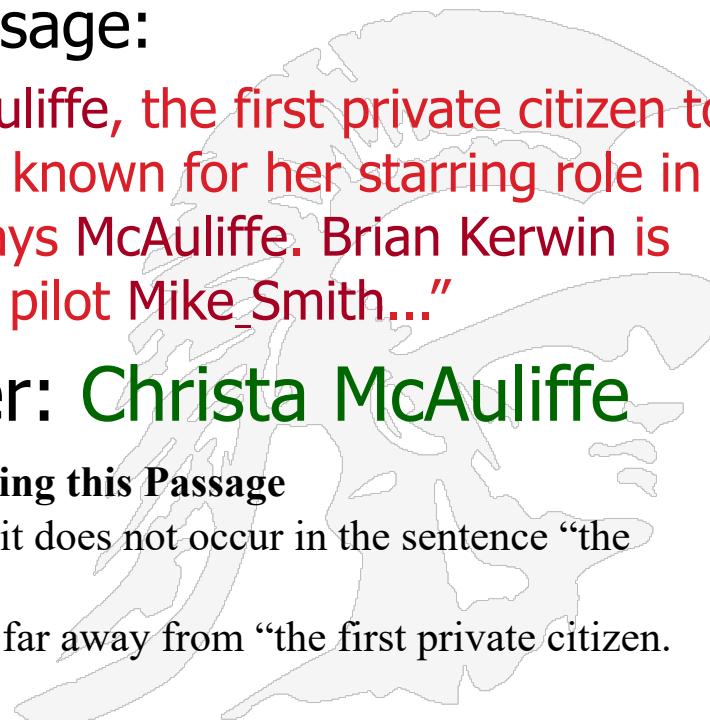
■ Text passage:

“Among them was Christa McAuliffe, the first private citizen to fly in space. Karen Allen, best known for her starring role in “Raiders of the Lost Ark”, plays McAuliffe. Brian Kerwin is featured as shuttle pilot Mike\_Smith...”

■ Best candidate answer: Christa McAuliffe

## Comments on Scoring this Passage

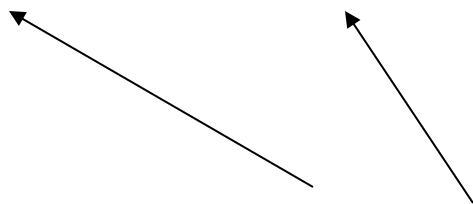
- Karen Allen is rejected as an answer as it does not occur in the sentence “the first private citizen. . . ”
- Brian Kerwin is rejected as the name is far away from “the first private citizen. . . ”



## Local Alignment Example (1 of 7)

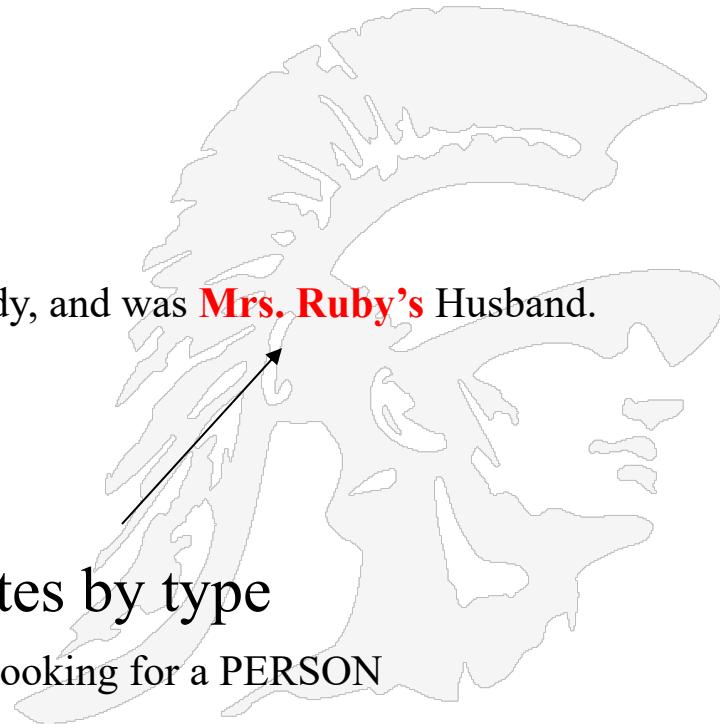
Who shot Kennedy?

Jack assassinated **Oswald**, the man who shot Kennedy, and was **Mrs. Ruby's Husband**.



Three Potential Candidates by type

WHO indicates we are looking for a PERSON



## Local Alignment Example (2 of 7)

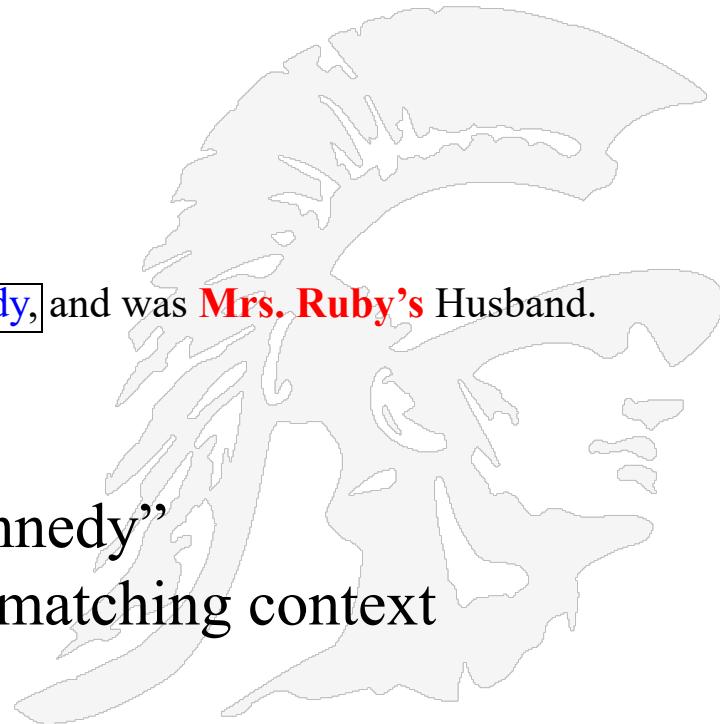
Question

Head  
word

→ *Who shot Kennedy?*

Jack assassinated Oswald, the man who **shot Kennedy**, and was Mrs. Ruby's Husband.

“shot Kennedy”  
gives us a verb and matching context



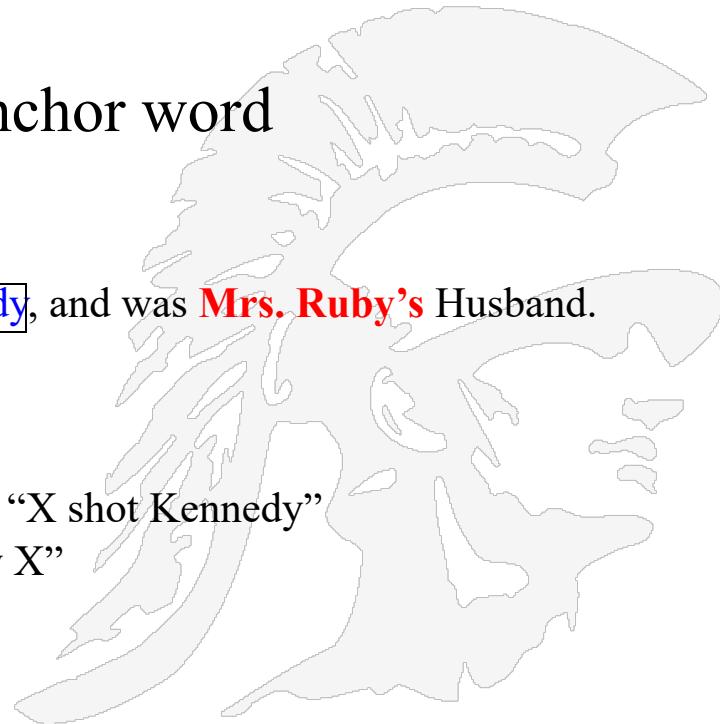
## Local Alignment Example (3 of 7)

*Who shot Kennedy?*

Anchor word

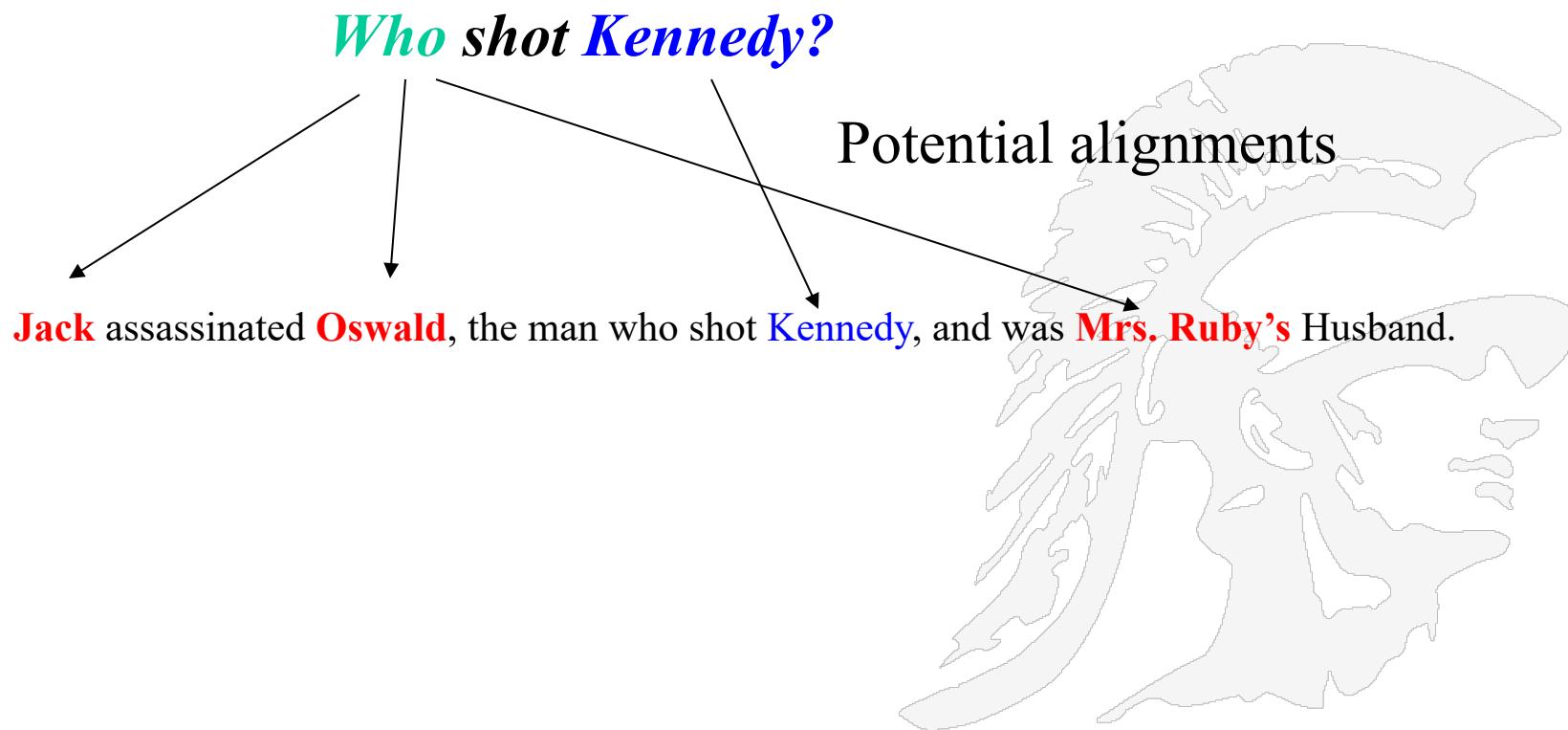
Jack assassinated Oswald, the man who shot **Kennedy**, and was **Mrs. Ruby's Husband**.

Look for phrases such as “X shot Kennedy”  
or “Kennedy was shot by X”



## Local Alignment (4 of 7)

In principle it can be anyone of the three people identified



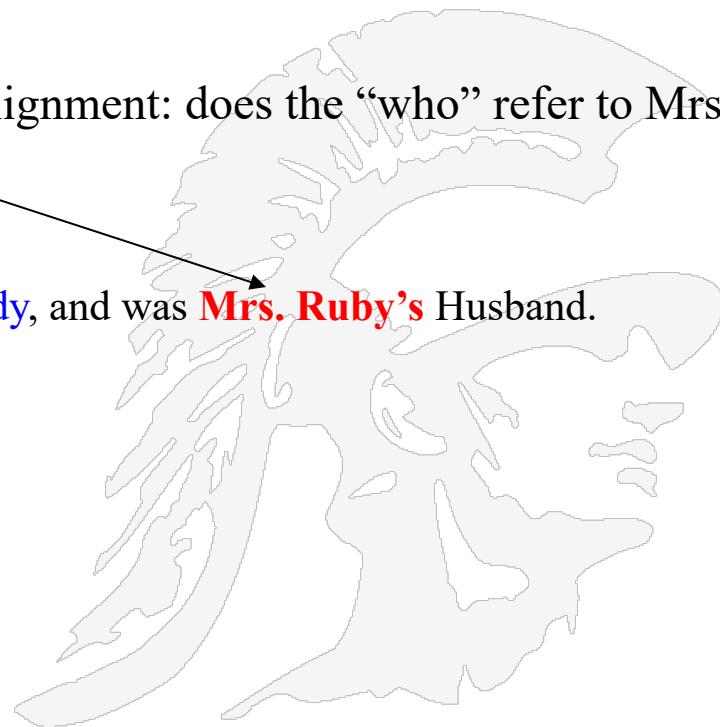
## Local Alignment Example (5 of 7)

*Who shot Kennedy?*

One Alignment: does the “who” refer to Mrs. Ruby?

Jack assassinated **Oswald**, the man who shot **Kennedy**, and was **Mrs. Ruby’s Husband**.

Three Alignment Features :



## Local Alignment Example (6 of 7)

1  
↔

*Who shot Kennedy?*

The distance between the question head word “who” and the anchor word Kennedy is 1

Jack assassinated **Oswald**, the man who shot **Kennedy**, and was **Mrs. Ruby’s Husband**.

One Alignment : does the “who” refer to Mrs. Ruby?  
The distance from Kennedy to Mrs. Ruby

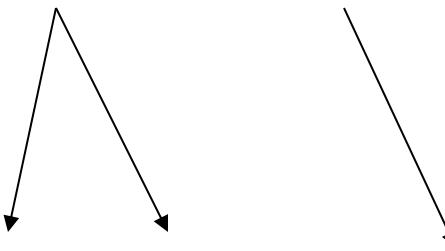
2  
↔

Three Alignment Features :

1. Distance between Question Head word (“who”) and the Anchor word (“Kennedy”) in the sentence

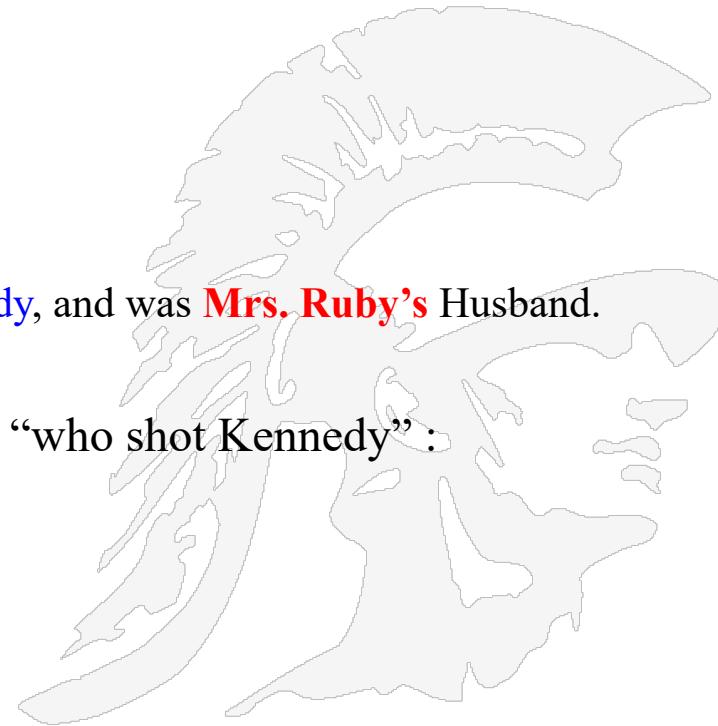
## Local Alignment Example (7 of 7)

*Who shot Kennedy?*



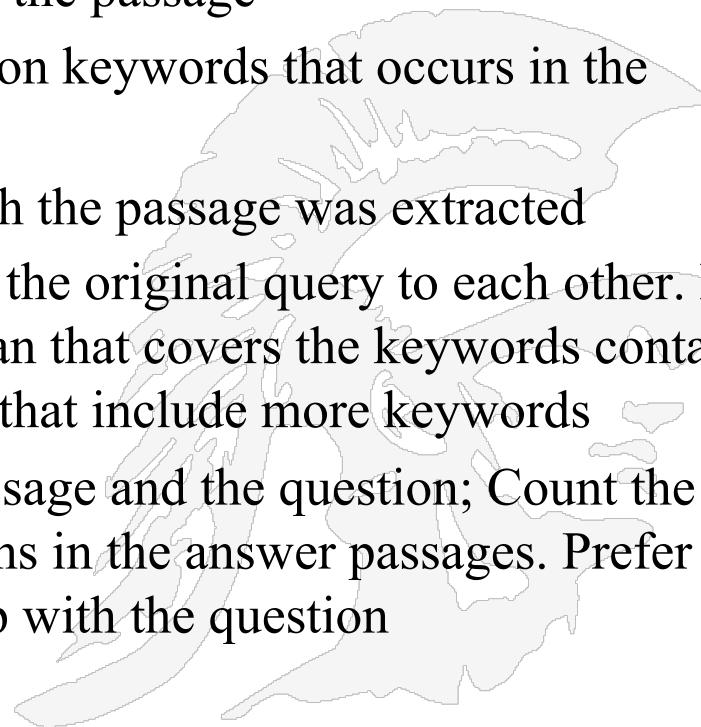
Jack assassinated **Oswald**, the man who shot **Kennedy**, and was **Mrs. Ruby's** Husband.

Oswald is properly aligned with “who shot Kennedy”:



# A Refined Ranking Scheme

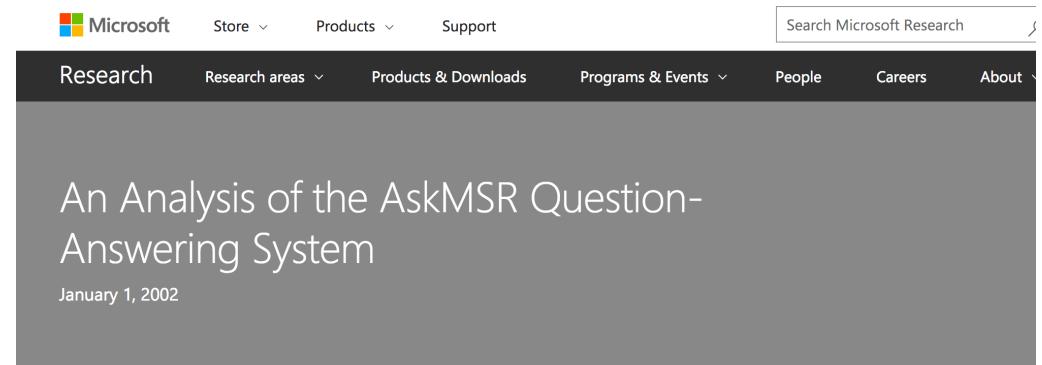
- **Refining the Passage Scoring Method, we can use supervised machine learning to rank the candidate passages according to six criteria**
  1. The number of named entities of the right type in the passage
  2. The number of question keywords in the passage
  3. The longest exact sequence of question keywords that occurs in the passage
  4. The rank of the document from which the passage was extracted
  5. The proximity of the keywords from the original query to each other. For each passage identify the shortest span that covers the keywords contained in that passage. Prefer smaller spans that include more keywords
  6. The N-gram overlap between the passage and the question; Count the N-grams in the question and the N-grams in the answer passages. Prefer the passages with higher N-gram overlap with the question



- AskMSR is a question answering system developed at Microsoft
- Rather than doing sophisticated linguistic analyses it relies upon information scattered around the web
- AskMSR system stressed how much could be achieved by very simple methods

[http://research.microsoft.com/en-us/um/people/sdumais/EMNLP\\_Final.pdf](http://research.microsoft.com/en-us/um/people/sdumais/EMNLP_Final.pdf)

# Microsoft's AskMSR Answering System



The screenshot shows the Microsoft Research homepage with a search bar. Below it, a navigation bar includes links for Research, Research areas, Products & Downloads, Programs & Events, People, Careers, and About. A large, dark grey rectangular area contains the title "An Analysis of the AskMSR Question-Answering System" and the date "January 1, 2002".

[Download PDF](#)

BibTex

#### Authors

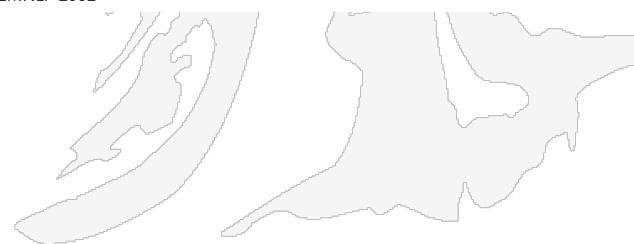
Eric Brill  
*Susan Dumais*  
 Michele Banko

#### Published In

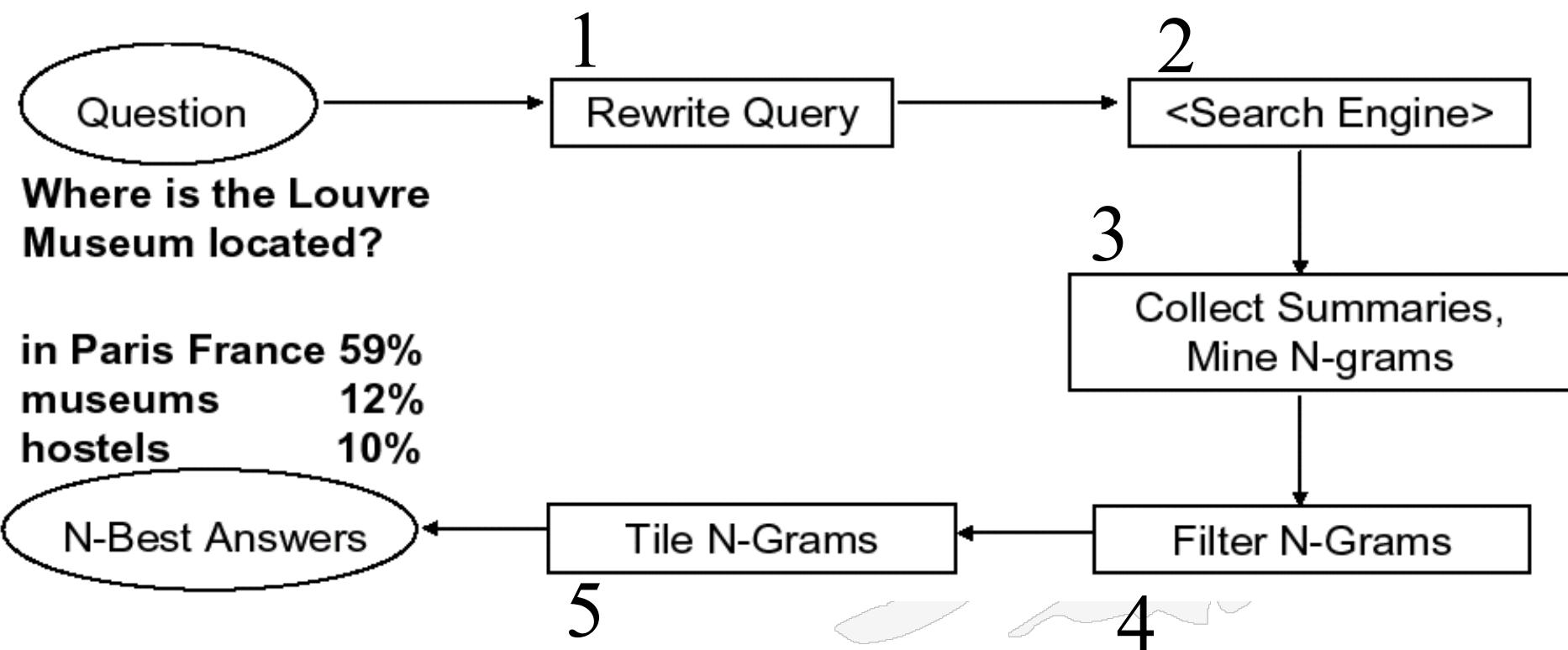
Proceedings of EMNLP 2002

[Abstract](#) [Related Info](#)

We describe the architecture of the AskMSR question answering system and systematically evaluate contributions of different system components to accuracy. The system differs from most question answering systems in its dependency on data redundancy rather than sophisticated linguistic analyses of either questions or candidate answers. Because a wrong answer is often worse than no answer, we also explore strategies for predicting when the question answering system is likely to give an incorrect answer.



# AskMSR: Details



# AskMSR:

## Step 1:Query Rewriting

- Classify question into categories
  - Who is/was/are/were...?
  - When is/did/will/are/were ...?
  - Where is/are/were ...?

### a. Category-specific transformation rules

eg “For Where questions, move ‘is’ to all possible locations”

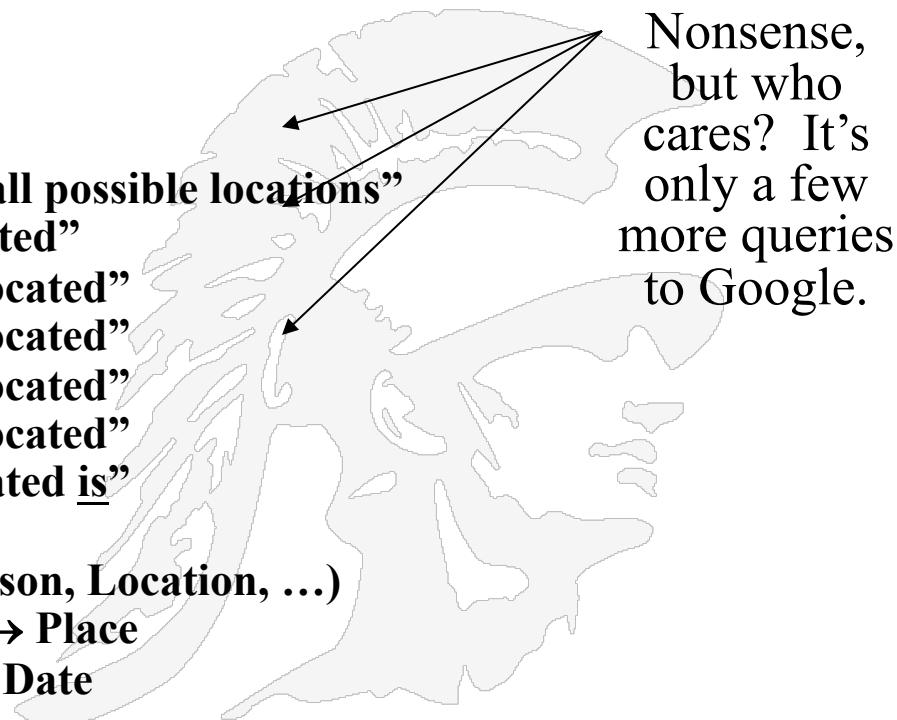
“Where is the Louvre Museum located”

- “is the Louvre Museum located”
- “the is Louvre Museum located”
- “the Louvre is Museum located”
- “the Louvre Museum is located”
- “the Louvre Museum located is”

### b. Expected answer “Datatype” (eg, Date, Person, Location, ...)

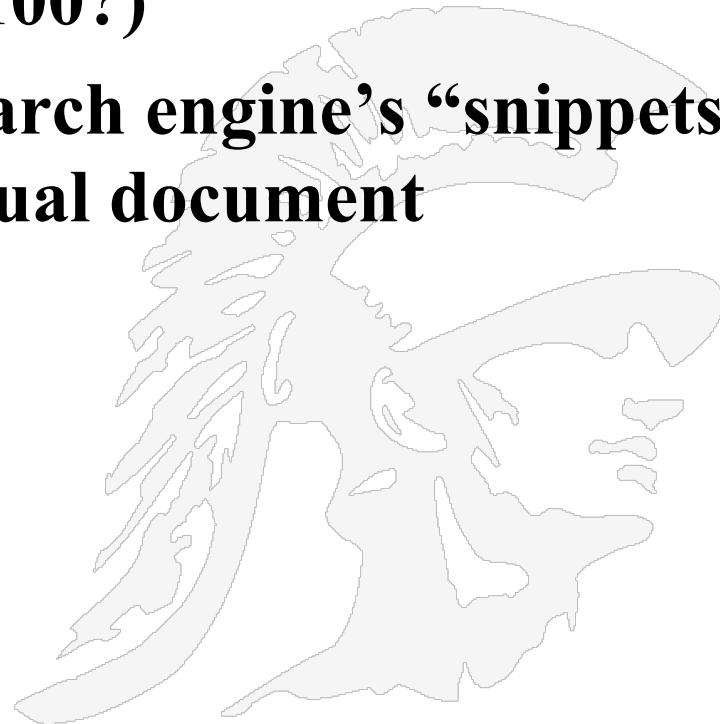
Where is the Louvre Museum located → Place

When was the French Revolution? → Date



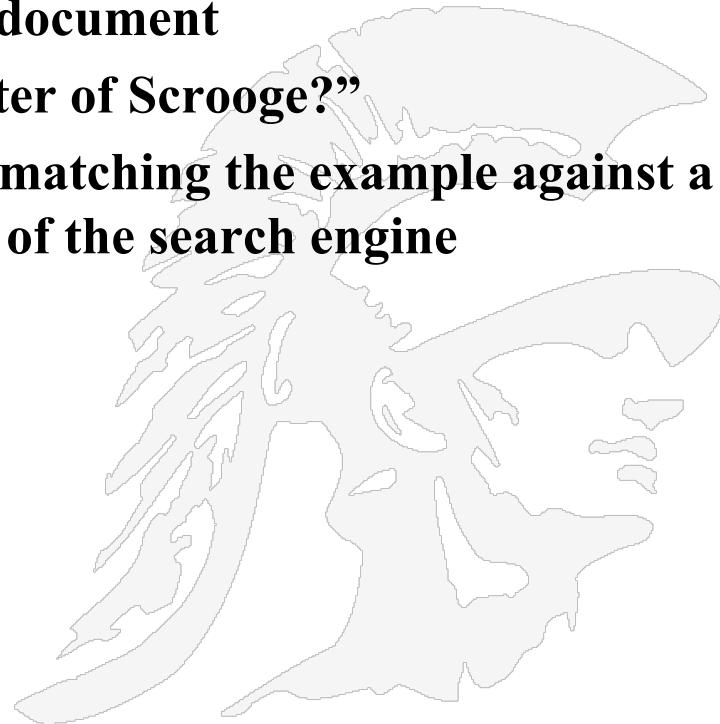
## AskMSR: Step 2: Query Search Engine

- Send all rewrites to a Web search engine
- Retrieve top N answers (100?)
- For speed, rely just on search engine's "snippets", not the full text of the actual document



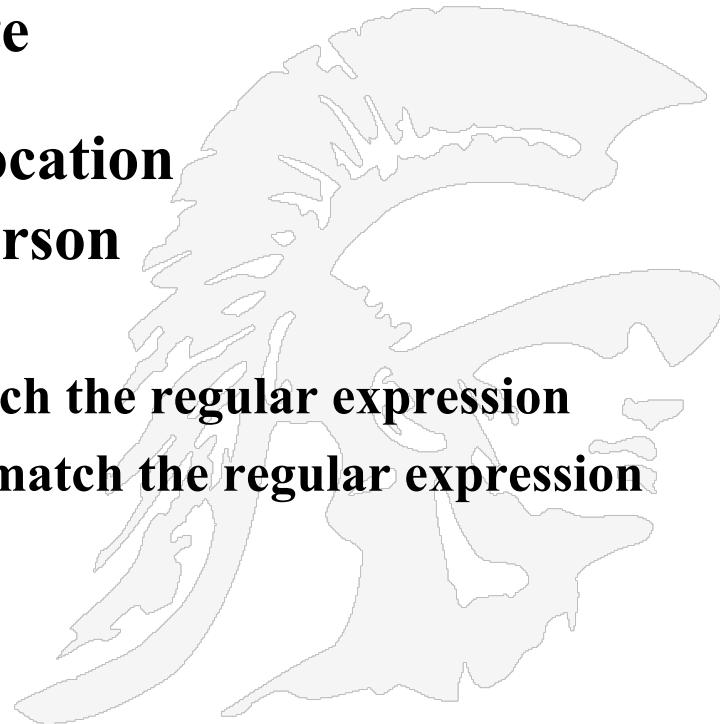
## AskMSR: Step 3: Mining N-Grams

- **Simple:** Enumerate all N-grams ( $N=1,2,3$  say) in all retrieved snippets
  - Use hash table and other data structures to make this efficient
- **Weight of an n-gram:** occurrence count, each weighted by “reliability” (weight) of rewrite that fetched the document
- **Example:** “Who created the character of Scrooge?”
- **Below are the weights produced by matching the example against a set of N-grams in the N-gram database of the search engine**
  - Dickens - 117
  - Christmas Carol - 78
  - Charles Dickens - 75
  - Disney - 72
  - Carl Banks - 54
  - A Christmas - 41
  - Christmas Carol - 45
  - Uncle - 31



AskMSR:  
**Step 4: Filtering N-Grams**

- Each question type is associated with one or more “data-type filters” = regular expression
- When... → Date
- Where... → Location
- What ... → Person
- Who ... →
- Boost score of n-grams that do match the regular expression
- Lower score of n-grams that don’t match the regular expression



AskMSR:  
**5: Tiling the Answers****Scores**

20

15

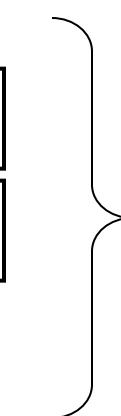
10

Charles Dickens

Dickens

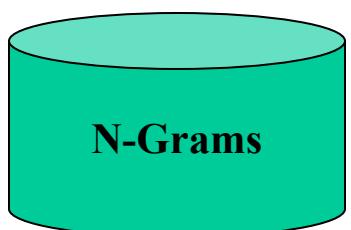
Mr Charles

Score 45

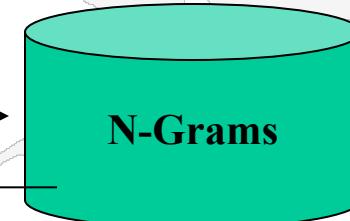
merged, discard  
old n-grams

Mr Charles Dickens

tile highest-scoring n-gram

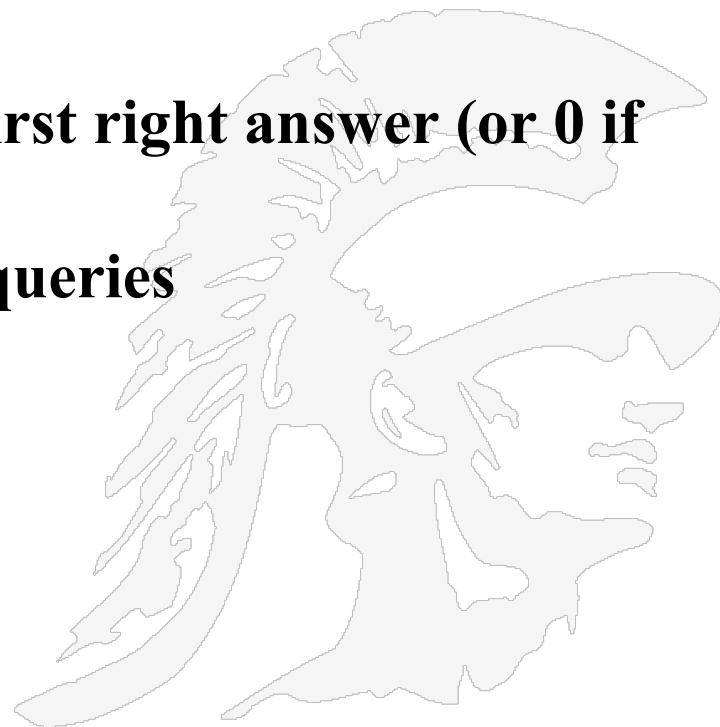


Repeat, until no more overlap



## Common Evaluation Metric

- Accuracy (does answer match gold-labeled answer?)
- Mean Reciprocal Rank
  - For each query return a ranked list of M candidate answers
  - Its score is 1/Rank of the first right answer (or 0 if no answers are correct)
  - Take the mean over all N queries
  - $$\text{MRR} = \sum_{i=1}^{i=N} (1/\text{rank}_i) / N$$



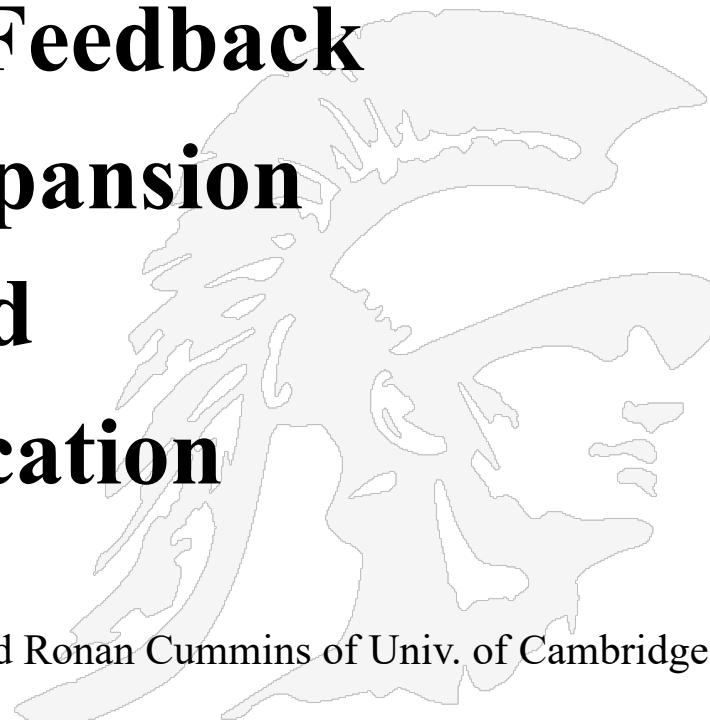
Introduction to

# Relevance Feedback

## Query Expansion

And

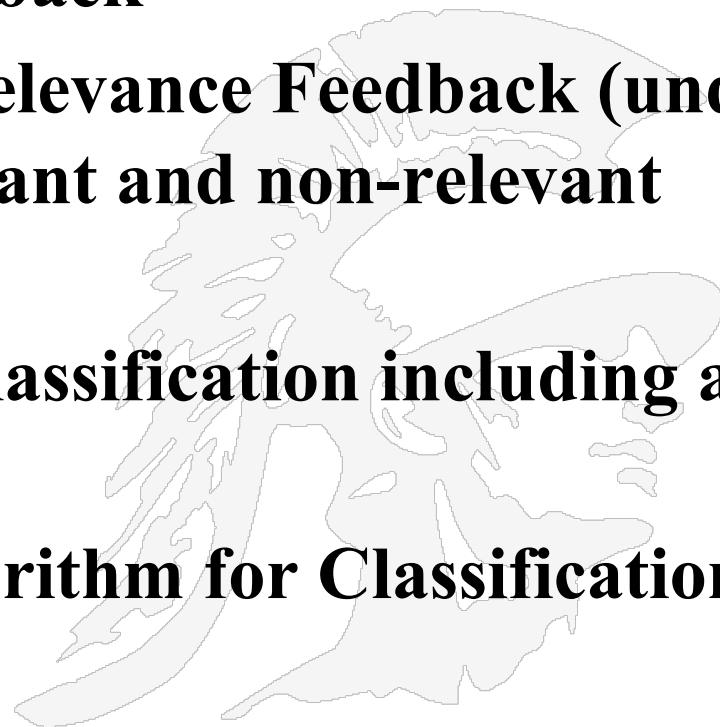
## Classification



Some slides are from Helen Yannakoudakis and Ronan Cummins of Univ. of Cambridge

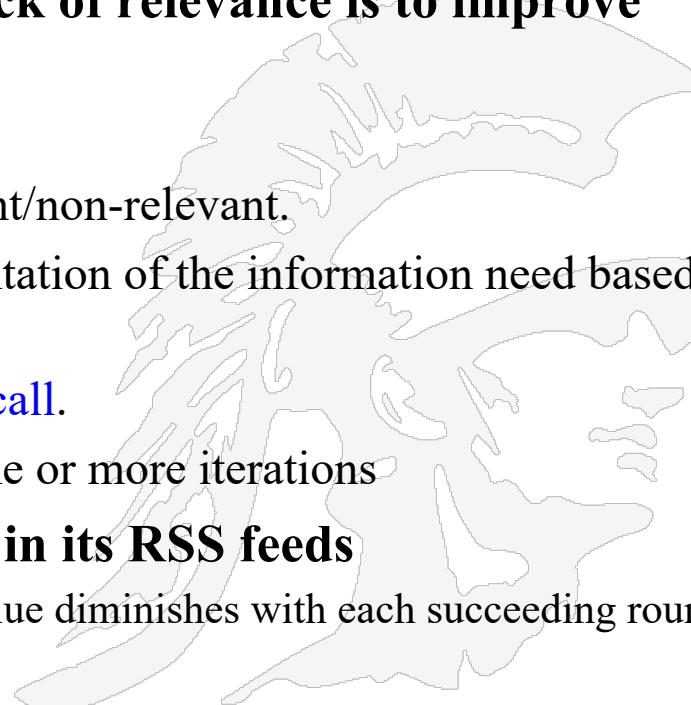
# Outline

- Examples of Relevance Feedback
- Query Expansion examples and techniques for producing relevance feedback
- Rocchio Algorithm for Relevance Feedback (under ideal conditions, i.e. relevant and non-relevant documents are known)
- Rocchio Algorithm for Classification including an online version
- K-Nearest Neighbor Algorithm for Classification



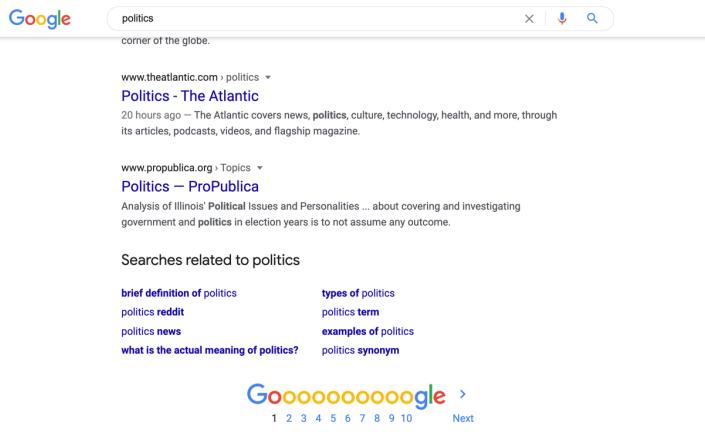
# Relevance Feedback

- **Idea:** The idea is to involve the user in the retrieval process by having the user give feedback about which results are most relevant
- The idealized process of user feedback of relevance is to improve recall
  - User issues a (short, simple) query
  - The user marks some results as relevant/non-relevant.
  - The system computes a better representation of the information need based on feedback.
  - New results have (hopefully) better recall.
  - Relevance feedback can go through one or more iterations
- **E.g. Google uses relevance feedback in its RSS feeds**
  - One round is generally sufficient, as the value diminishes with each succeeding round of relevance feedback



# Examples where Relevance Feedback Would be Useful

“politics”  
unrefined  
query



Google search results for "politics":

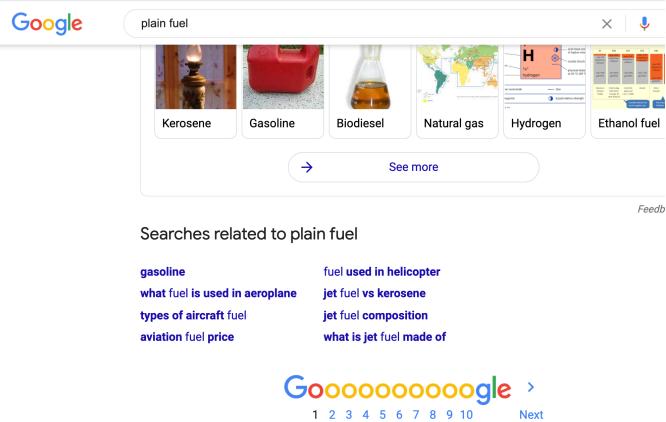
- [www.theatlantic.com - politics](http://www.theatlantic.com/politics/) Politics - The Atlantic 20 hours ago – The Atlantic covers news, politics, culture, technology, health, and more, through its articles, podcasts, videos, and flagship magazine.
- [www.propublica.org/Topics](http://www.propublica.org/Topics) Politics — ProPublica Analysis of Illinois' Political Issues and Personalities ... about covering and investigating government and politics in election years is to not assume any outcome.

Searches related to politics

<a href="#">brief definition of politics</a>	<a href="#">types of politics</a>
<a href="#">politics reddit</a>	<a href="#">politics term</a>
<a href="#">politics news</a>	<a href="#">examples of politics</a>
<a href="#">what is the actual meaning of politics?</a>	<a href="#">politics synonym</a>

Goooooooooooooogle >  
1 2 3 4 5 6 7 8 9 10 Next

“plain fuel”  
Two interpretations are possible



Google search results for "plain fuel":

Images shown: Kerosene, Gasoline, Biodiesel, Natural gas, Hydrogen, Ethanol fuel.

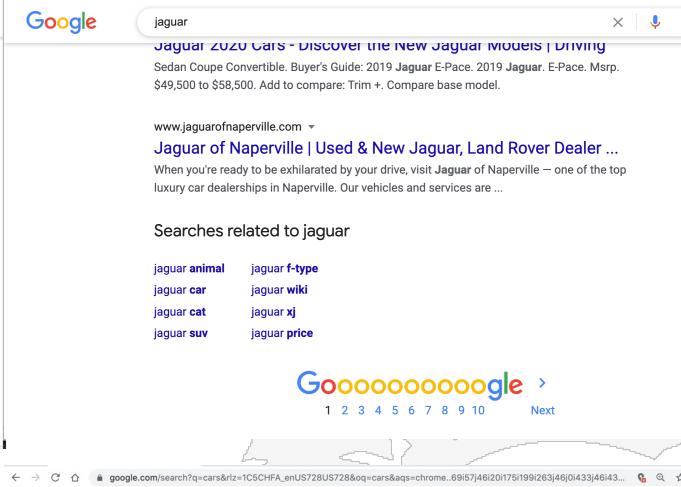
Searches related to plain fuel

<a href="#">gasoline</a>	<a href="#">fuel used in helicopter</a>
<a href="#">what fuel is used in aeroplane</a>	<a href="#">jet fuel vs kerosene</a>
<a href="#">types of aircraft fuel</a>	<a href="#">jet fuel composition</a>
<a href="#">aviation fuel price</a>	<a href="#">what is jet fuel made of</a>

Goooooooooooooogle >  
1 2 3 4 5 6 7 8 9 10 Next

- Relevance feedback and query expansion aim to overcome these problems.

“jaguar”  
ambiguous



Google search results for "jaguar":

Images shown: Sedan Coupe Convertible. Buyer's Guide: 2019 Jaguar E-Pace, 2019 Jaguar. E-Pace. Msrp. \$49,500 to \$58,500. Add to compare: Trim +. Compare base model.

Jaguar of Naperville | Used & New Jaguar, Land Rover Dealer ... When you're ready to be exhilarated by your drive, visit Jaguar of Naperville – one of the top luxury car dealerships in Naperville. Our vehicles and services are ...

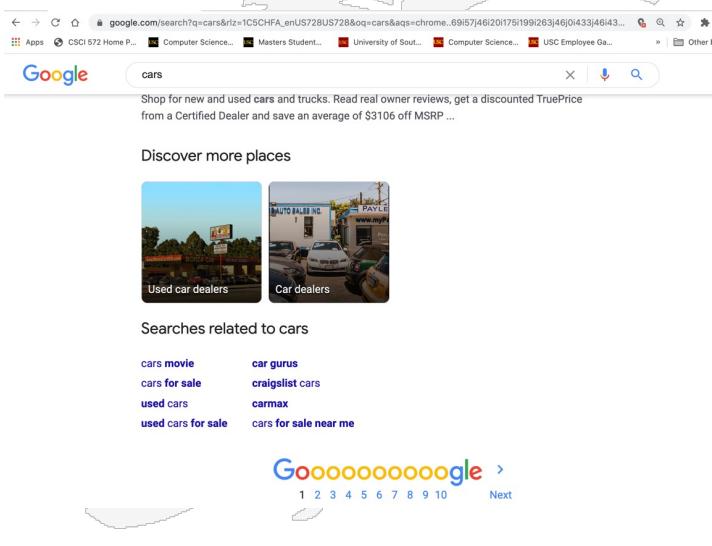
Searches related to jaguar

<a href="#">jaguar animal</a>	<a href="#">jaguar f-type</a>
<a href="#">jaguar car</a>	<a href="#">jaguar wiki</a>
<a href="#">jaguar cat</a>	<a href="#">jaguar xj</a>
<a href="#">jaguar suv</a>	<a href="#">jaguar price</a>

Goooooooooooooogle >  
1 2 3 4 5 6 7 8 9 10 Next



“cars”  
unrefined



Google search results for "cars":

Images shown: Used car dealers, Car dealers.

Discover more places

Searches related to cars

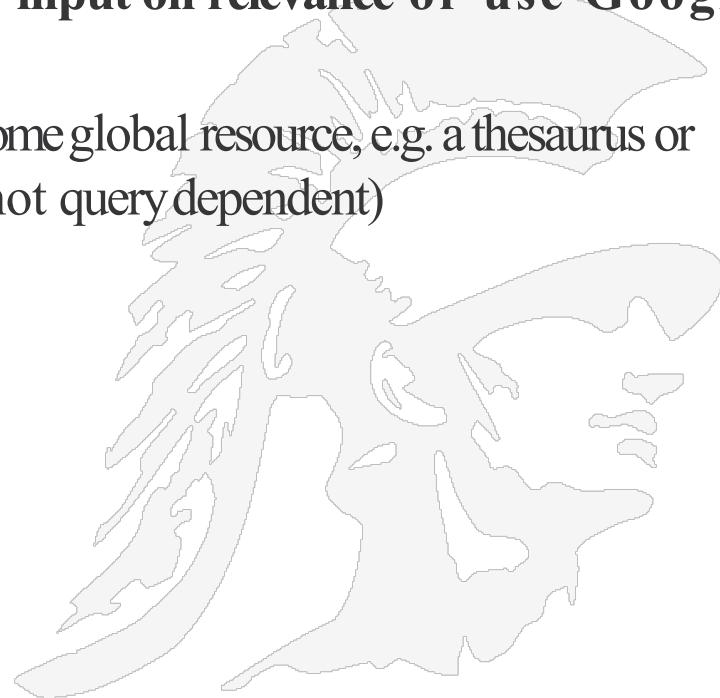
<a href="#">cars movie</a>	<a href="#">car gurus</a>
<a href="#">cars for sale</a>	<a href="#">craigslist cars</a>
<a href="#">used cars</a>	<a href="#">carmax</a>
<a href="#">used cars for sale</a>	<a href="#">cars for sale near me</a>

Goooooooooooooogle >  
1 2 3 4 5 6 7 8 9 10 Next



# Relevance Feedback will Improve Recall

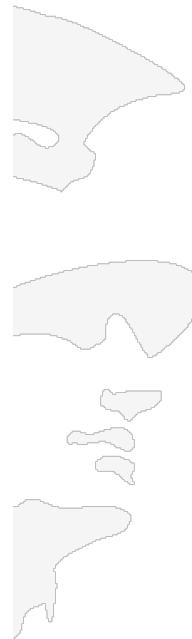
- Methods for tackling this problem split into two classes
  - 1. Local methods: adjust a query relative to the documents returned (query-time analysis on a portion of documents)
    - Main local method: either get user input on relevance or use Google's approach of “Did you mean?”
  - 2. Global methods: adjust query based on some global resource, e.g. a thesaurus or knowledge graph (i.e., a resource that is not query dependent)



# Google Uses Query Expansion in its Search Results

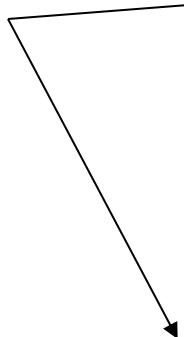
- Google uses six techniques for expanding a query to make it more meaningful

Google's Query Expansion	Explanation and observations
Word stemming	The term searched is reduced to a root or 'stem'- words of the same stem can be ranked for the search containing only one of them.
Acronyms	An abbreviation / acronym / initialism searched is automatically resolved to the full version. If there are several possible variants of one and the same acronym, Google will do its best to <b>mention all variations on the first page</b> in the following order: (1) most popular and hence most probable one (2) all the rest.
Misspellings	With misspelled words searched, Google will suggest the correct variant as well as list sites that use that correct variant. Note: based on common webmasters' experience, <b>Google uses some other (unclear) algorithm for typos</b> : sites ranked for the misspelled word (using it nowhere on the page or in the text of backlinks) may have low authority and rank nowhere if the correct spelling is searched.
Synonyms	Sometimes Google includes related words in the search results. Very often the query is extended this way when it is <b>obvious the phrase was used improperly</b> . The synonym substituting the part of the phrase is <b>not bolded</b> as a rule.
Translations	In some instances, Google seems to translate search keywords into other languages and return results from that language.
Ignored words	Occasionally some ("insignificant") words appear to have been dropped completely from the query.

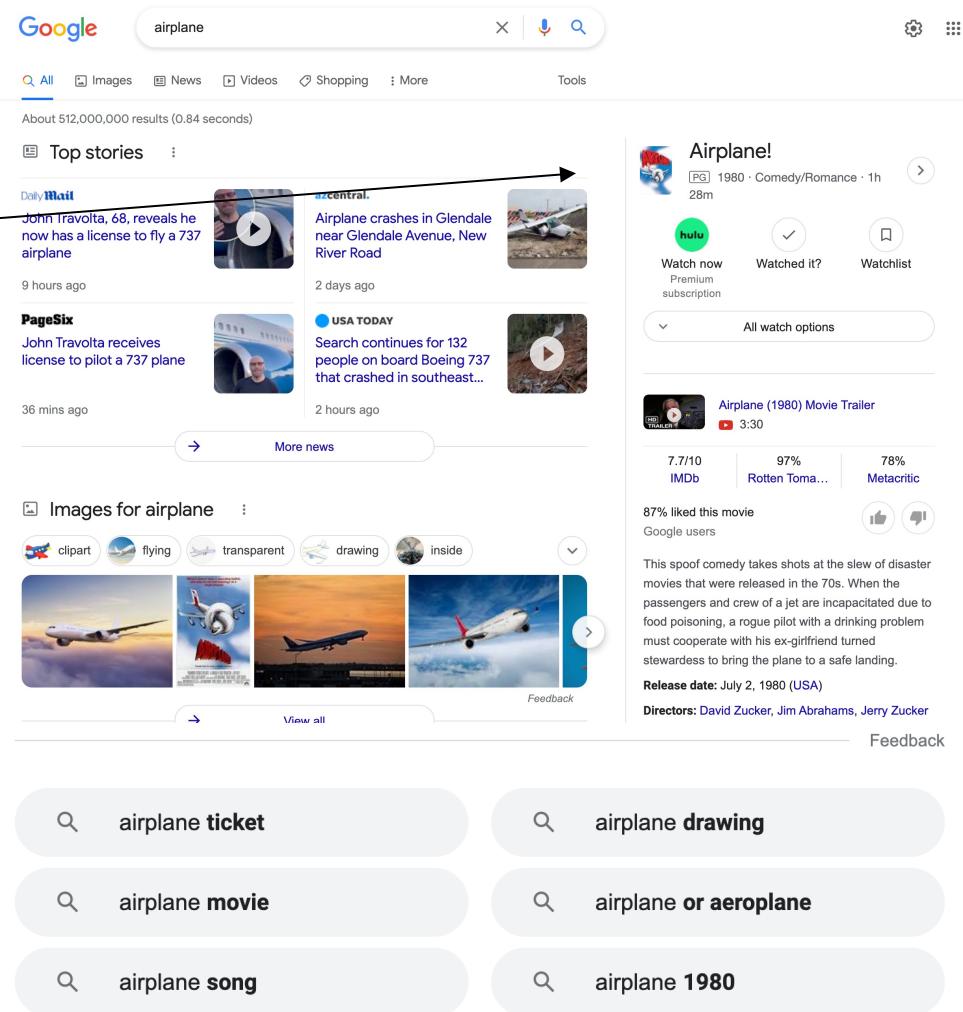


# Query Expansion Example 1

- Query expansion is used to increase precision and recall.
- The query “airplane” primarily produces a set of movie results



- query expansion which Google calls Feedback will expand the query making use of the knowledge graph

The screenshot shows a Google search results page for the query "airplane". The top section displays "Top stories" with news articles from Daily Mail, PageSix, and USA TODAY. One story from Daily Mail discusses John Travolta's license to fly a 737 airplane. Another story from USA TODAY discusses a Boeing 737 crash. To the right of the news, there is a "Feedback" sidebar for the movie "Airplane! (1980)". The sidebar includes a thumbnail, rating (PG), release year (1980), genres (Comedy/Romance), runtime (1h 28m), and links to Hulu, Watch now, Watched it?, and Watchlist. Below the news, there is an "Images for airplane" section with several thumbnail images of various aircraft. At the bottom of the page, there are six suggested search terms: "airplane ticket", "airplane drawing", "airplane movie", "airplane or aeroplane", "airplane song", and "airplane 1980".

# Query Expansion: Example 2

Google

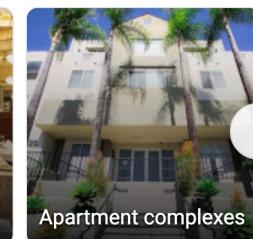
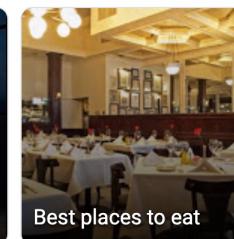
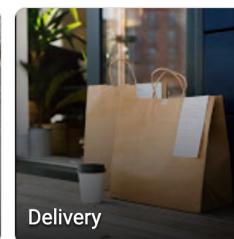
palm

[X](#) | [Microphone](#) [Search](#)

- The query is “palm”
- The carousel of choices includes restaurants, hotels all named “palm”
- The related searches include references to palm trees, palm pdas and palm hotels

[www.verizon.com › Connected Devices › Palm](#) ▾[Palm Companion Phone with NumberShare - Small Size ...](#)Sync your **Palm** with any IOS or Android device to ensure you stay in touch while on the go.About the size of a credit card, **Palm** packs a huge punch. Unleash ... Rating: 5 · 10 reviews · \$14.58 · In stock

Discover more places



Searches related to palm

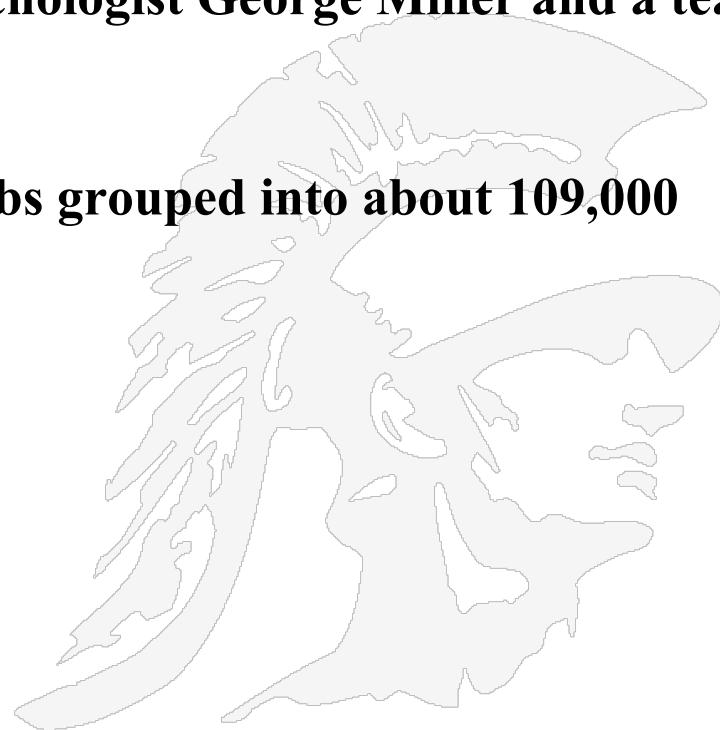
palm tree	palm, inc
palm hand	palm companion
palm phone	palm phone 2
palm device	palm pda

Gooooooooogle &gt;

1 2 3 4 5 6 7 8 9 10 Next

## Once Again We Look at WordNet to Implement Query Expansion

- One way to perform query expansion is to use WordNet
- A more detailed database of semantic relationships between English words.
- Developed by famous cognitive psychologist George Miller and a team at Princeton University.
- About 144,000 English words.
- Nouns, adjectives, verbs, and adverbs grouped into about 109,000 synonym sets called *synsets*.

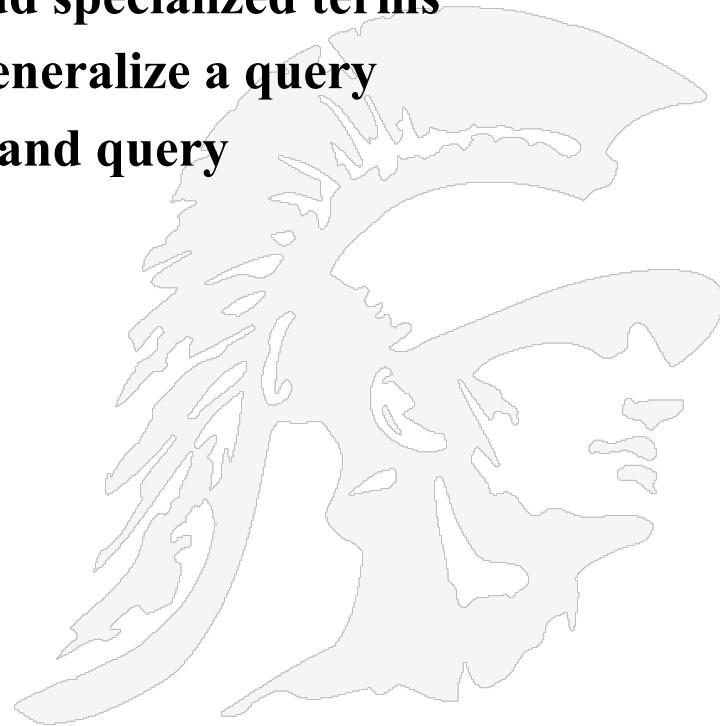


# WordNet Synset Relationships

- **Antonym:** front → back (opposite)
- **Attribute:** benevolence → good (noun to adjective)
- **Pertainym:** alphabetical → alphabet (relating to; adjective to noun)
- **Similar:** unquestioning → absolute (resembling)
- **Cause:** kill → die
- **Entailment:** breathe → inhale (a necessary part)
- **Holonym:** chapter → text (part to whole)
- **Meronym:** computer → cpu (whole to part)
- **Hyponym:** plant → tree (specialization)
- **Hypernym:** apple → fruit (generalization)

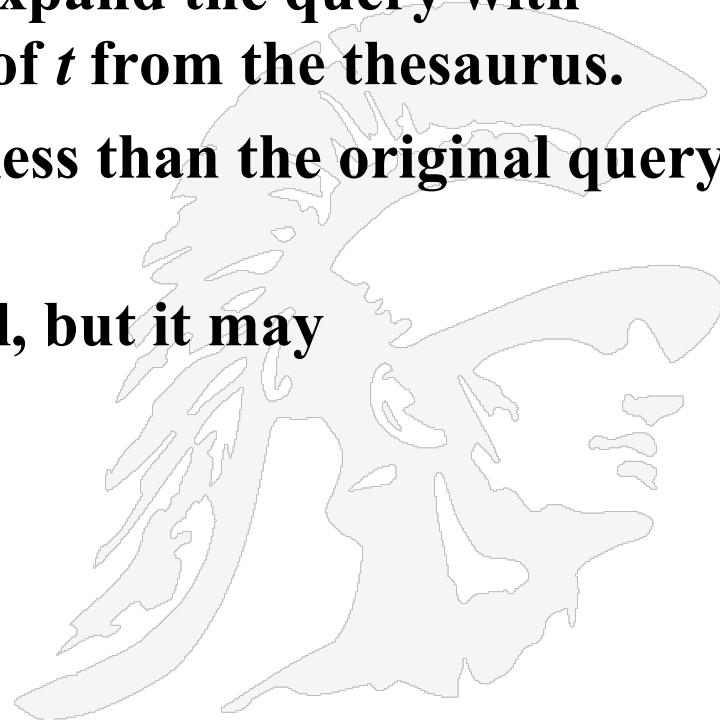
# Query Expansion Using WordNet

- **Form an expanded query by:**
  1. Adding WordNet synonyms in the same synset
  2. Adding WordNet hyponyms to add specialized terms
  3. Adding WordNet hypernyms to generalize a query
  4. Adding other related terms to expand query



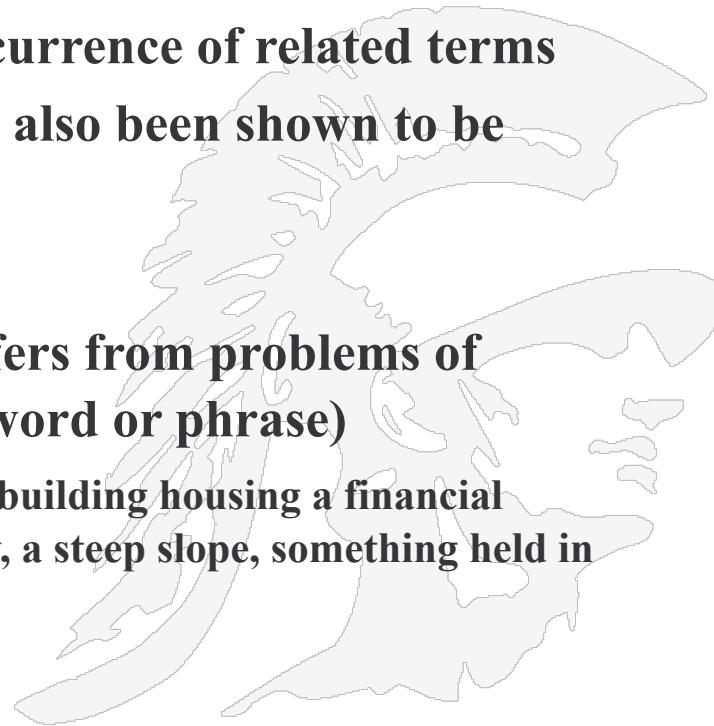
# Query Expansion Using Thesaurus

- Another way to do query expansion is to make use of a thesaurus
- For each term,  $t$ , in a query, expand the query with synonyms and related words of  $t$  from the thesaurus.
- One can weight added terms less than the original query terms.
- This generally increases recall, but it may significantly decrease precision



# Query Expansion Summary of Points

- Query Expansion is transparent in that it allows the user to see (select) expansion terms.
- Use WordNet to identify synonyms and related terms
- Use a thesaurus to develop a co-occurrence of related terms
- Query log mining approaches have also been shown to be useful
- However, query expansion still suffers from problems of polysemy (multiple meanings of a word or phrase)
  - E.g. “bank” is a financial institution, a building housing a financial institution, the process of saving money, a steep slope, something held in reserve, “you can rely (bank) on me”



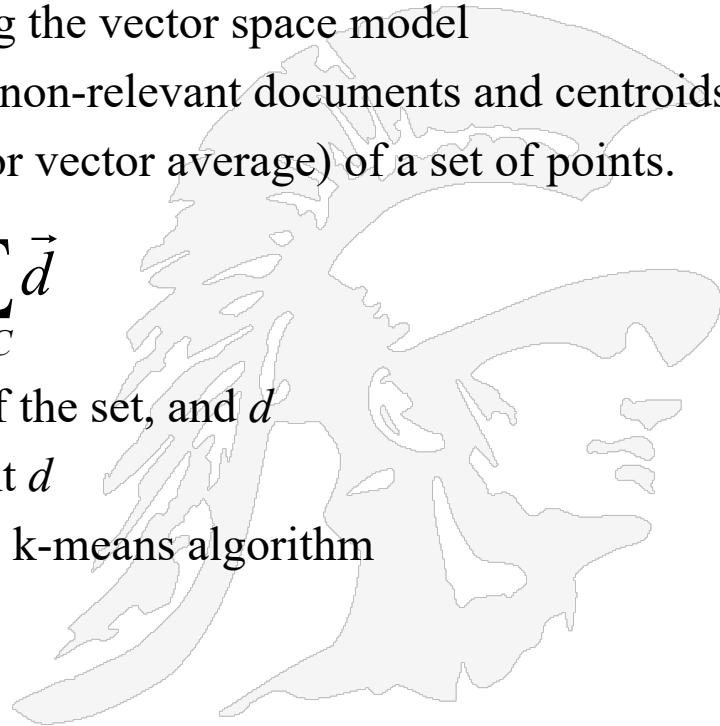
## Rocchio Algorithm: Basics

- The Rocchio algorithm is a method of relevance feedback
- It was initially developed by the SMART Information Retrieval System in 1960-1964.
- It assumes documents are represented using the vector space model
- The algorithm uses the notions of relevant/non-relevant documents and centroids
- Recall: the centroid is the center of mass (or vector average) of a set of points.
- *Definition:* Centroid

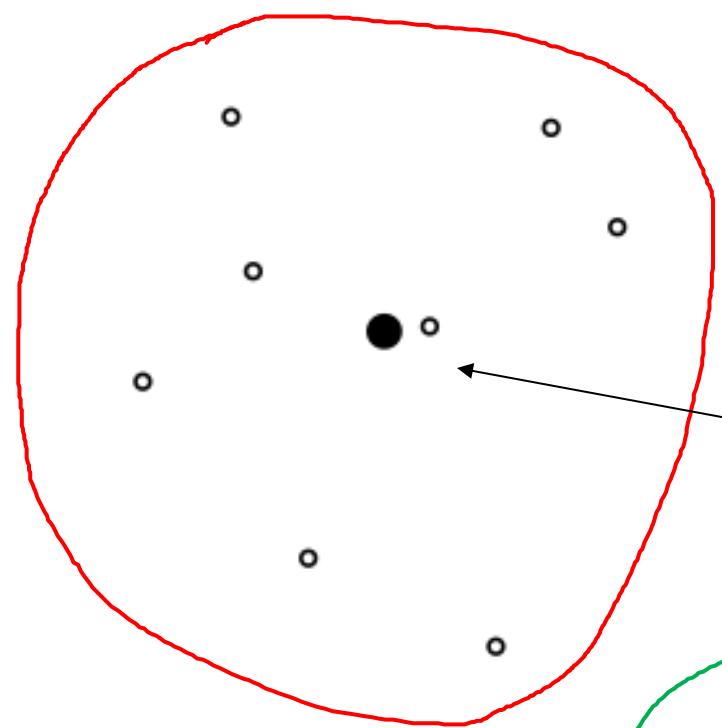
$$\vec{\mu}(C) = \frac{1}{|C|} \sum_{d \in C} \vec{d}$$

where  $C$  is a set of documents,  $|C|$  is the size of the set, and  $\vec{d}$  is the normalized vector representing document  $d$

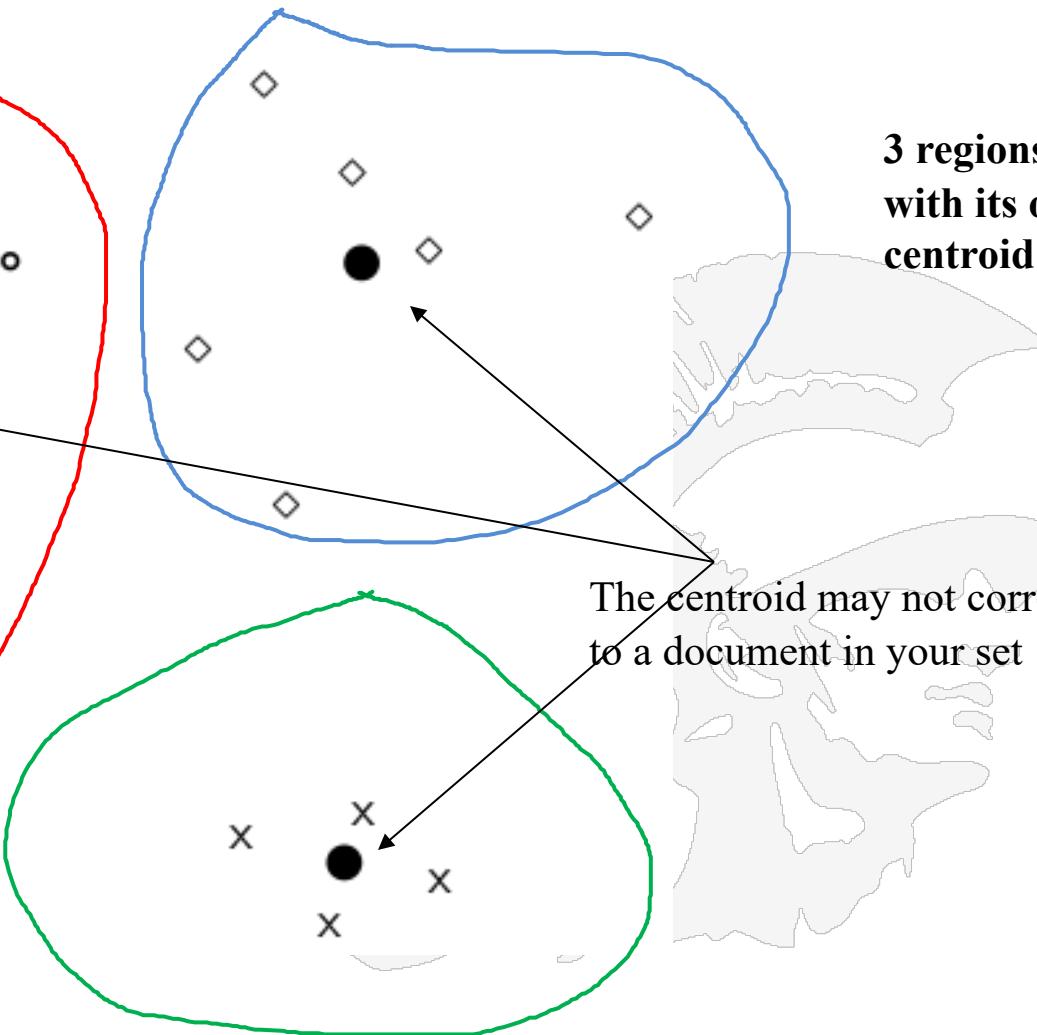
- **Note:** We have seen centroids before in the k-means algorithm



# Centroid Example



The boundary between two classes in Rocchio classification is the set of points with equal distance from the two centroids



3 regions each with its own centroid

The centroid may not correspond to a document in your set

## Rocchio Algorithm Derivation

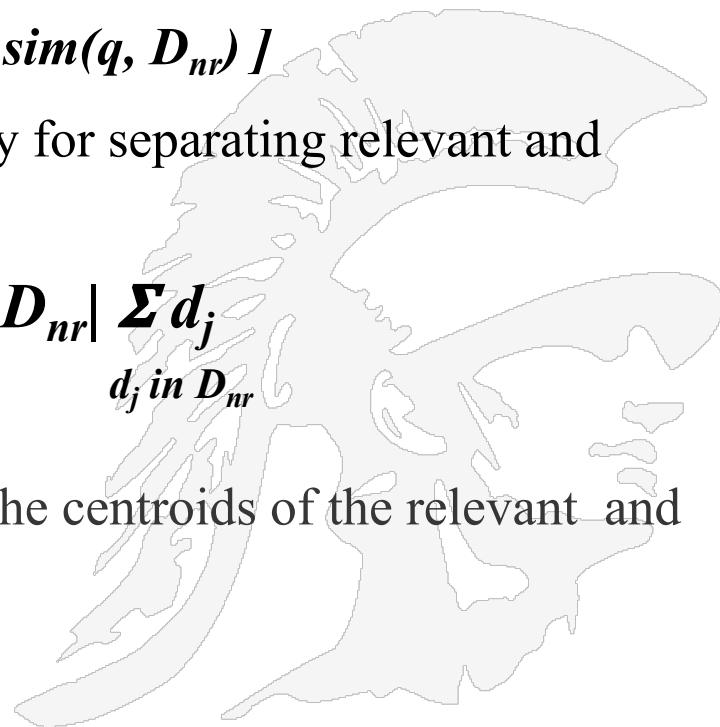
Assuming someone has identified the set of relevant ( $D_r$ ) and non-relevant ( $D_{nr}$ ) documents, the algorithm aims to find the query  $q$  that maximizes similarity with the set of relevant documents  $D_r$  while minimizing similarity with the set of non-relevant documents  $D_{nr}$ :

$$q_{opt} = \arg \max [sim(q, D_r) - sim(q, D_{nr})]$$

Under cosine similarity, the optimal query for separating relevant and non-relevant documents is:

$$q_{opt} = \frac{1}{|D_r|} \sum_{d_j \in D_r} d_j - \frac{1}{|D_{nr}|} \sum_{d_j \in D_{nr}} d_j$$

which is the vector difference between the centroids of the relevant and non-relevant documents.



# Rocchio Algorithm for Relevance Feedback - in Practice

- In practice, however, we usually do not know the full set of relevant and non-relevant sets.
- For example, a user might only label a few documents as relevant / non-relevant.
- Therefore, in practice Rocchio is often parameterised as follows:

$$\vec{q}_m = \alpha \vec{q} + \frac{\beta}{|D_r|} \sum_{\forall \vec{d}_j \in D_r} \vec{d}_j - \frac{\gamma}{|D_n|} \sum_{\forall \vec{d}_j \in D_n} \vec{d}_j$$

where  $q$  is the original query vector;  $D_r$  and  $D_n$  are the sets of known relevant and non-relevant documents.

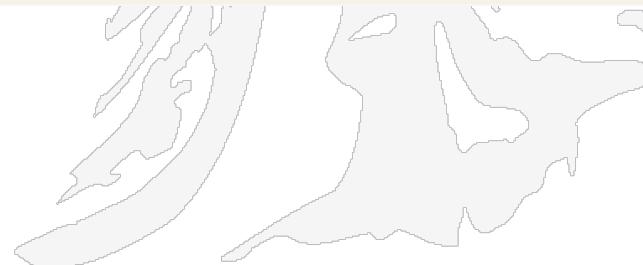
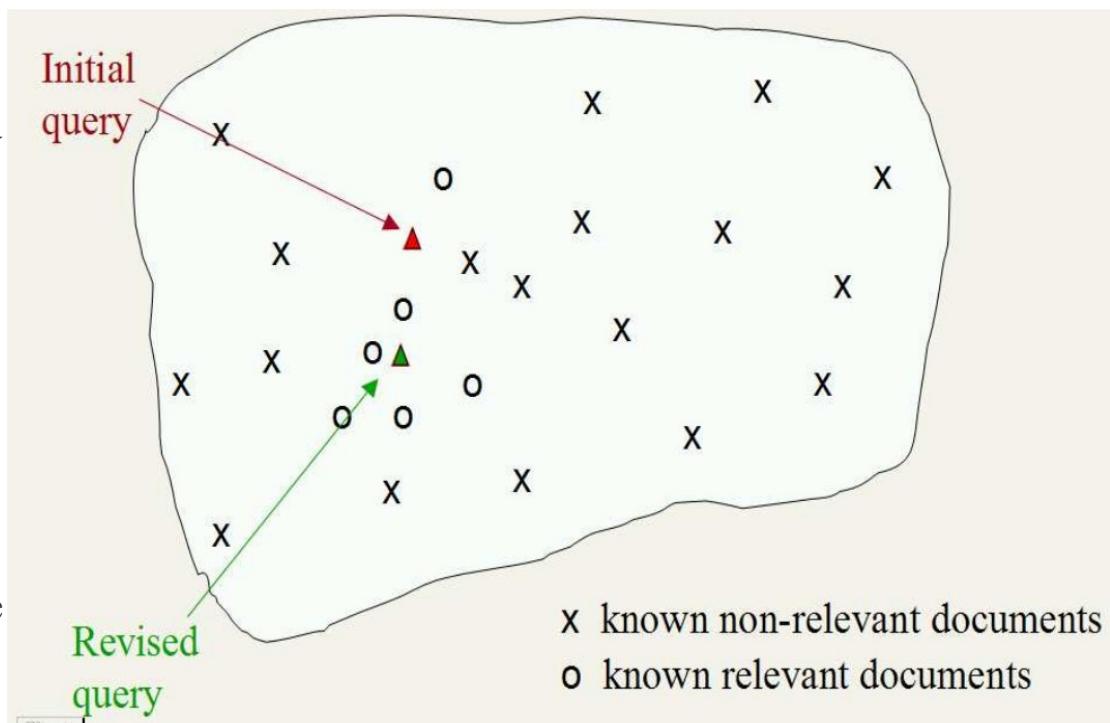
$\alpha$ ,  $\beta$ , and  $\gamma$  are weight parameters attached to each component.

Reasonable values are  $\alpha = 1.0$ ,  $\beta = 0.75$ ,  $\gamma = 0.15$

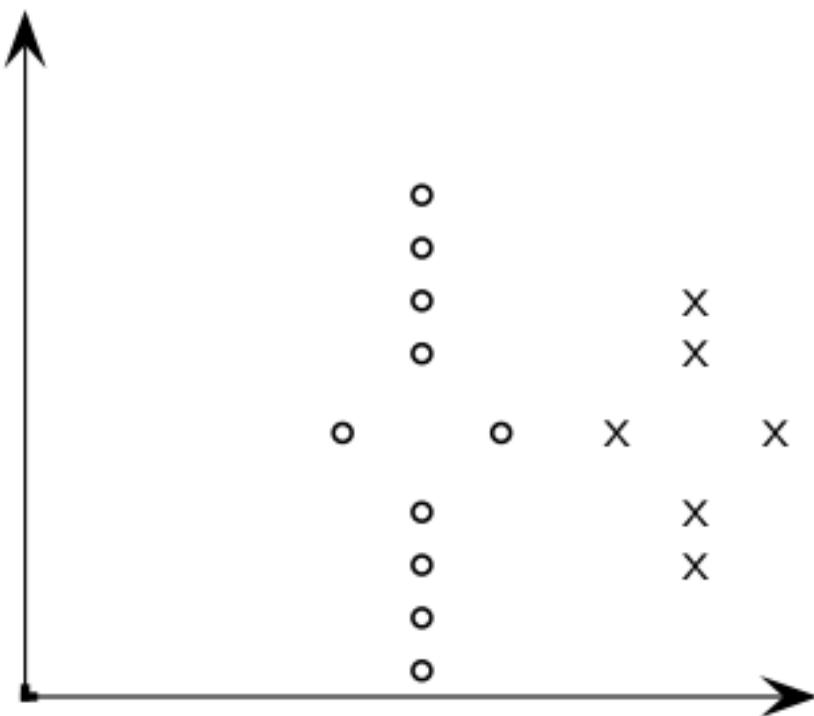
• Note: if the final value of  $q_m$  has negative term weights, set those to 0.

- Represent query and documents as weighted vectors (e.g., tf-idf).
- Use Rocchio formula to compute new query vector (given some known relevant / non-relevant documents).
- Calculate cosine similarity between new query vector and the documents.
- Rocchio has been shown useful for increasing both precision and recall because it contains aspects of positive and negative feedback.
- Positive feedback is much more valuable than negative (i.e., indications of what *is* relevant) so typically systems set  $\gamma < \beta$  or even  $\gamma = 0$ .

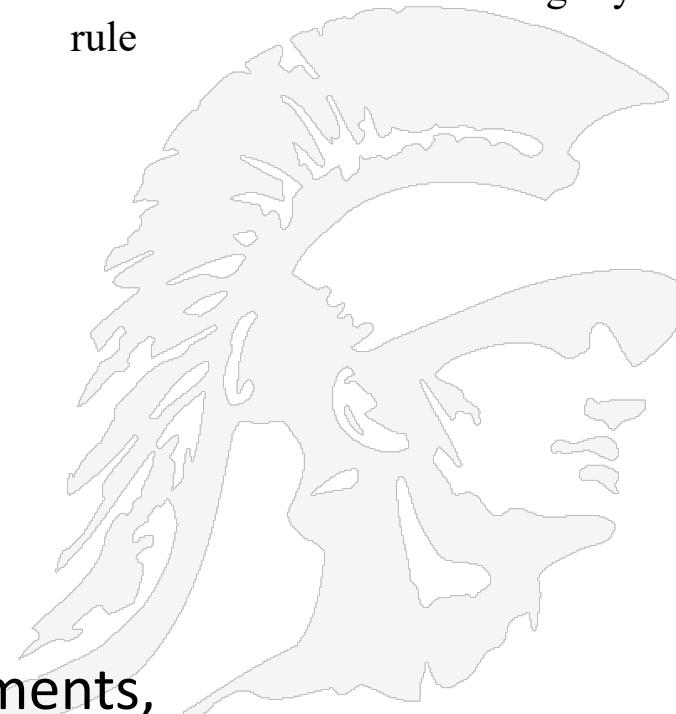
# Rocchio in Practice



# 2D Rocchio Example

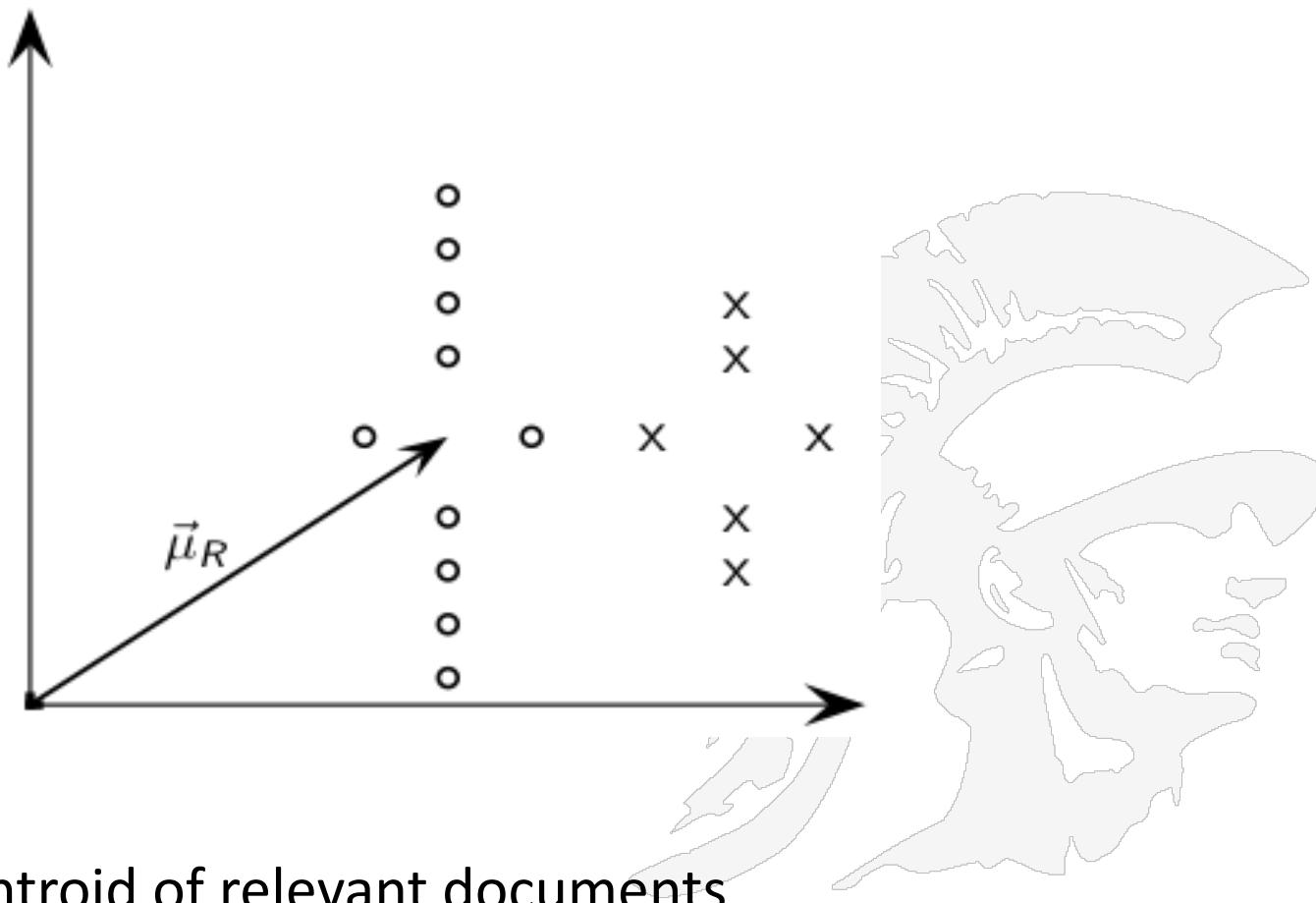


- For 2D examples the relevant set is generally much smaller than the non-relevant set;
- As a result we need a slightly modified rule

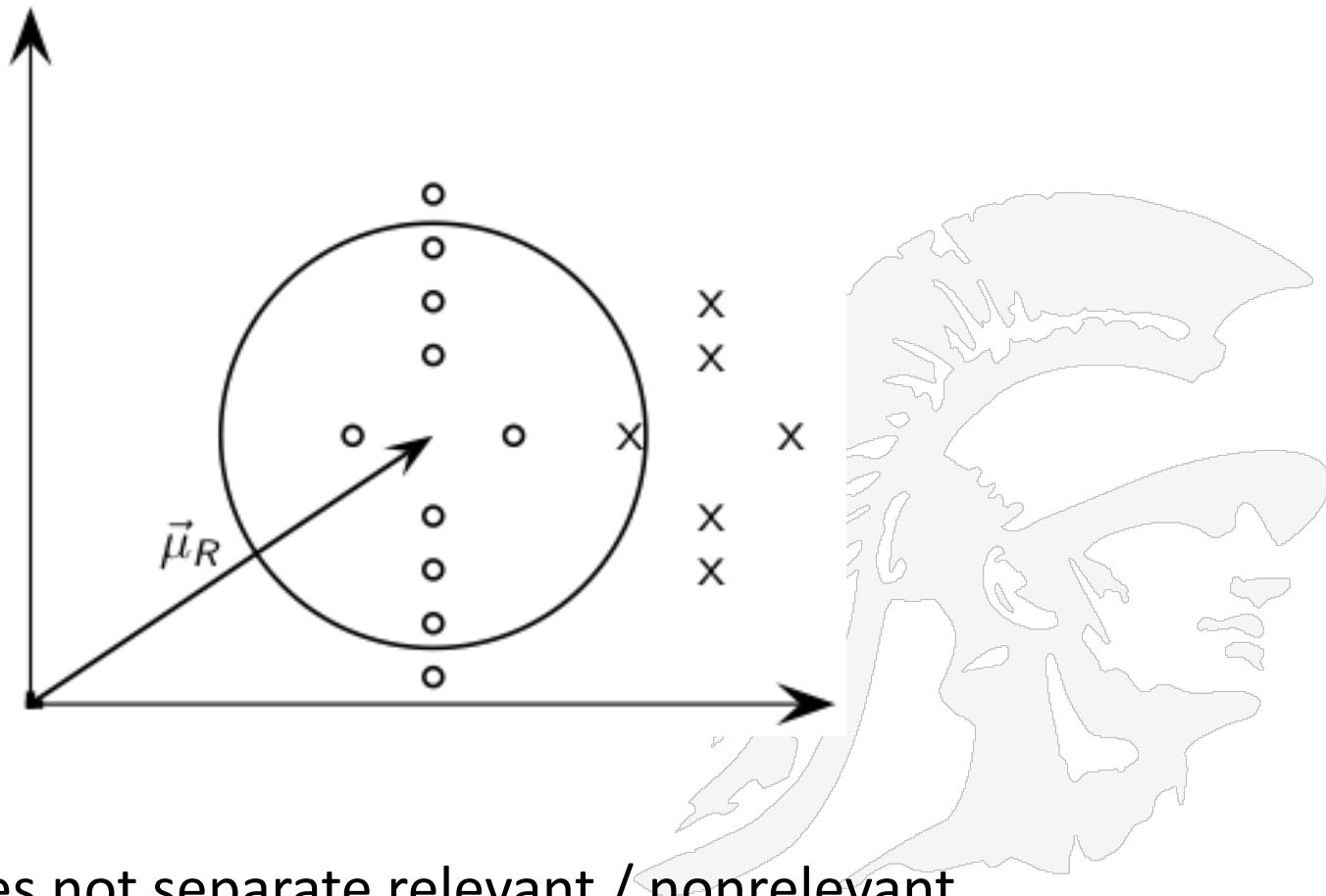


Let circles represent relevant documents,  
Let Xs represent nonrelevant documents

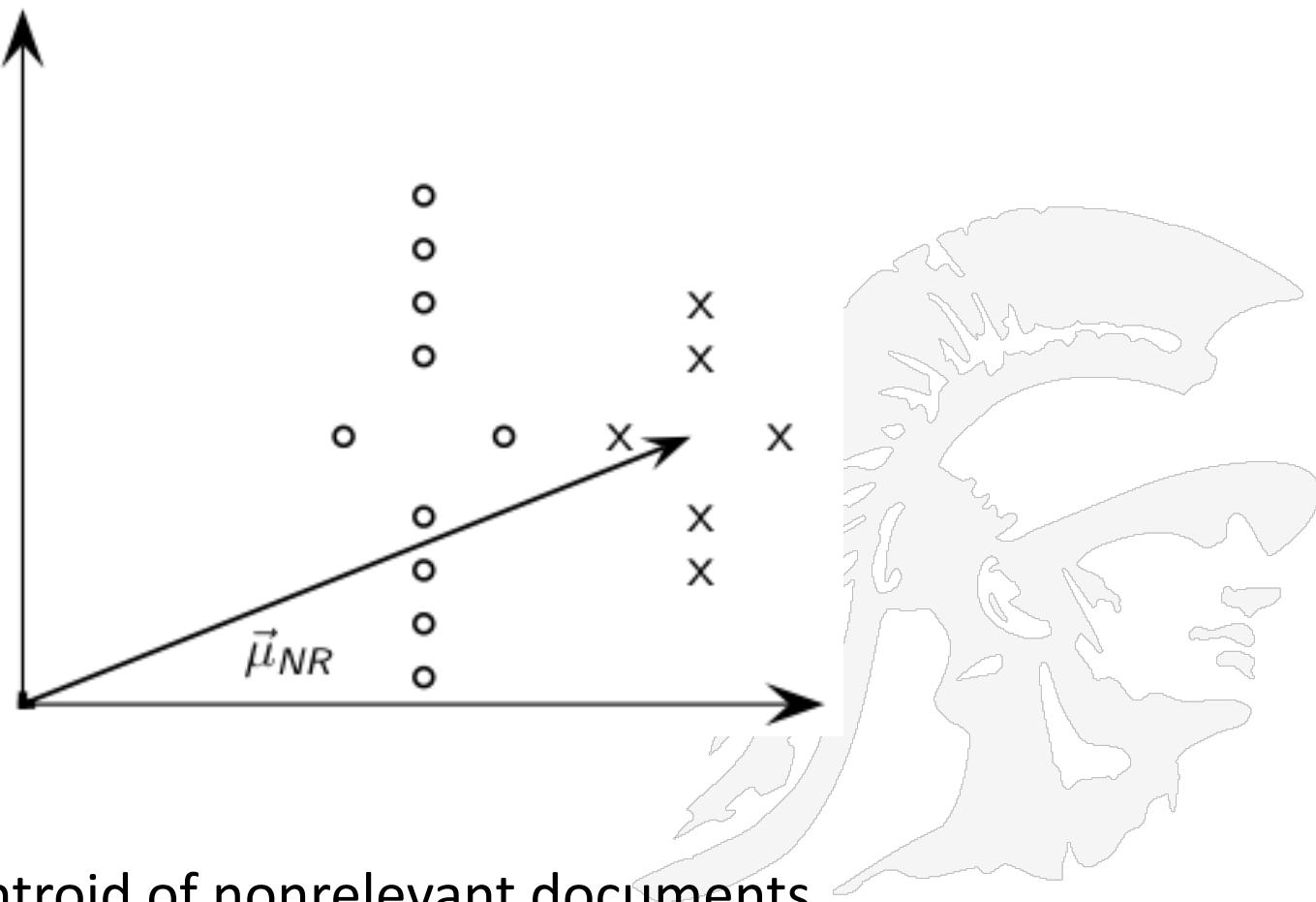
## 2D Rocchio Illustrated (1 of 9)



## 2D Rocchio Illustrated (2 of 9)

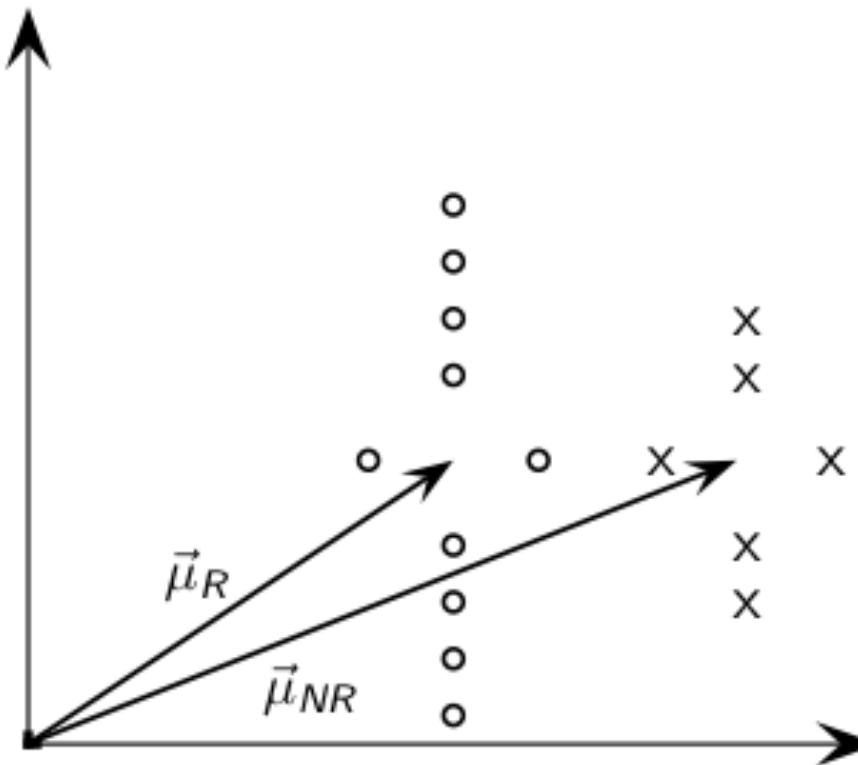


## 2D Rocchio Illustrated (3 of 9)

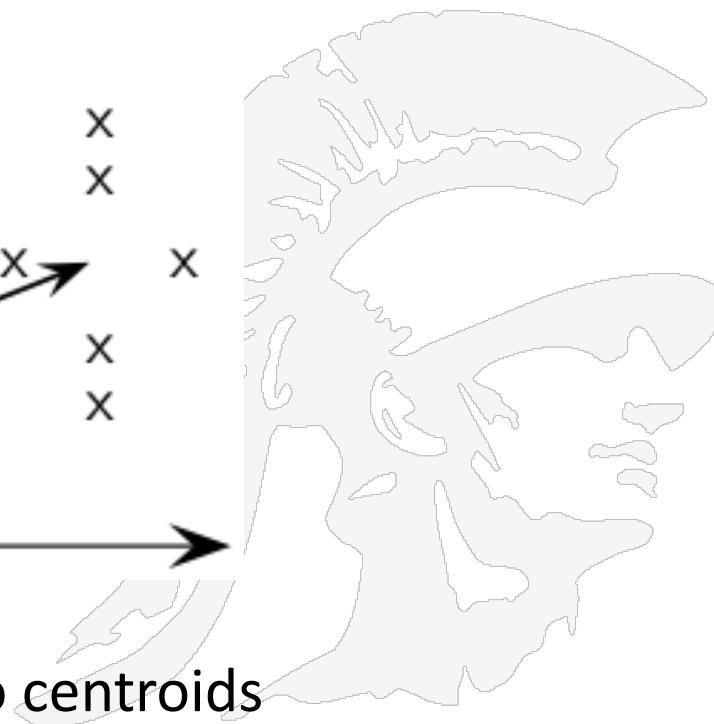


$\vec{\mu}_{NR}$ : centroid of nonrelevant documents.

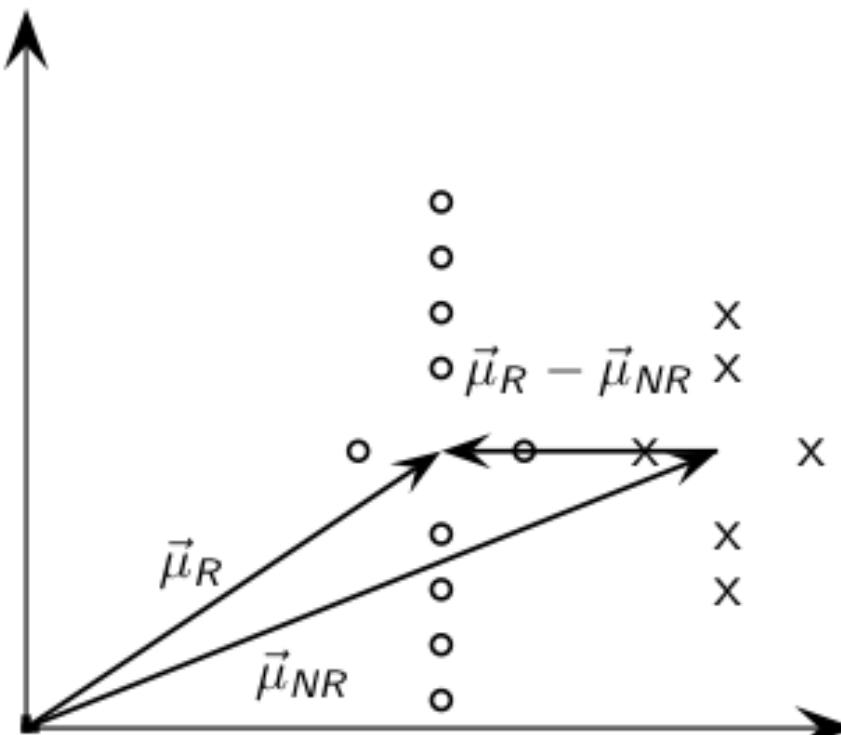
## 2D Rocchio Illustrated (4 of 9)



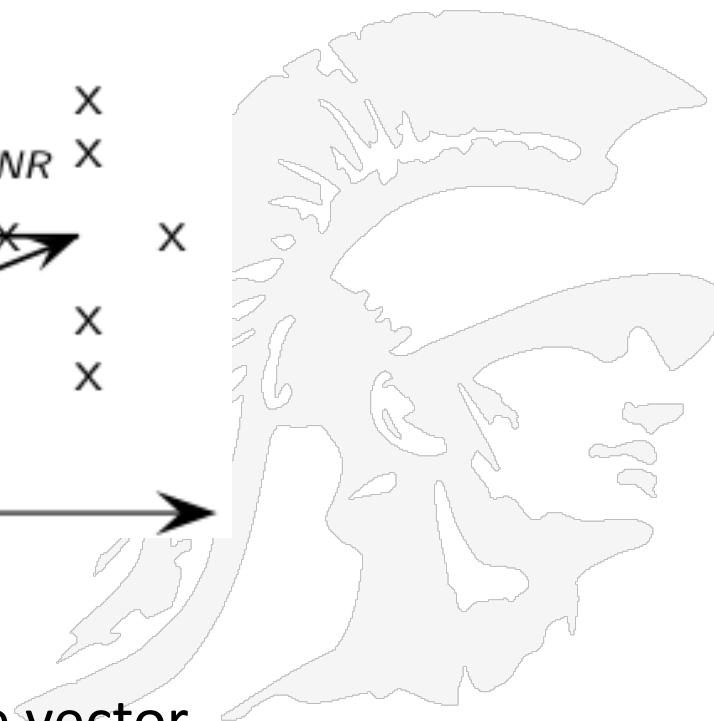
Consider the two centroids



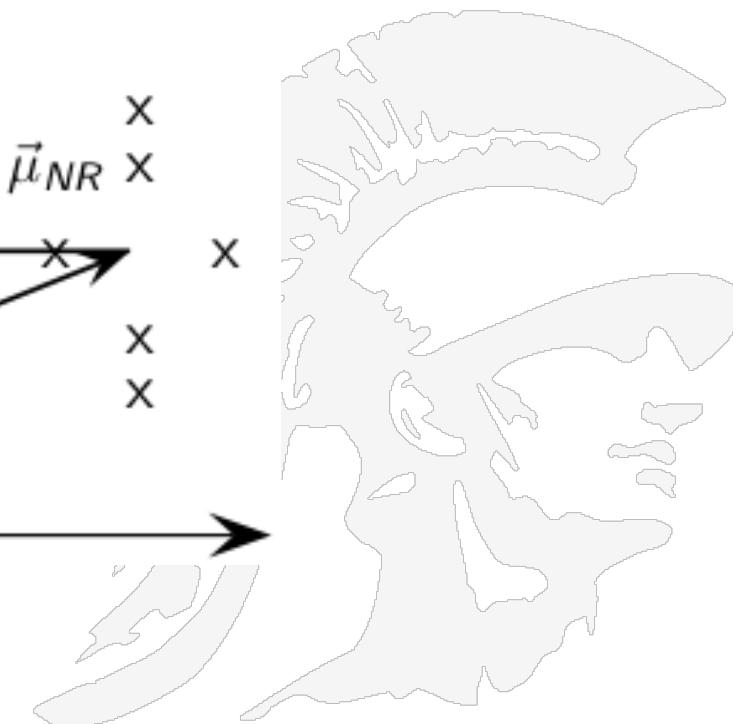
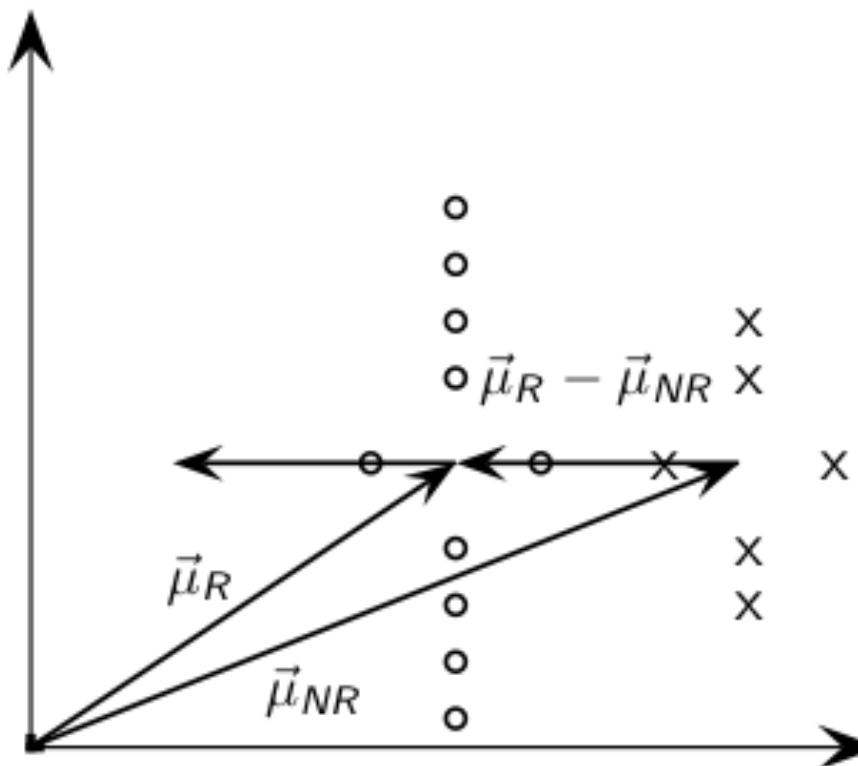
## 2D Rocchio Illustrated(5 of 9)



$\vec{\mu}_R - \vec{\mu}_{NR}$ : centroid difference vector

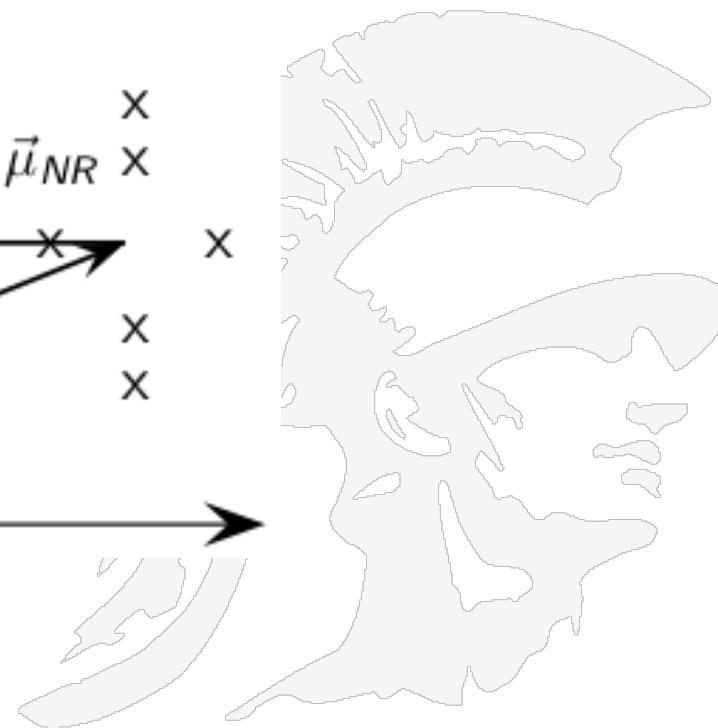
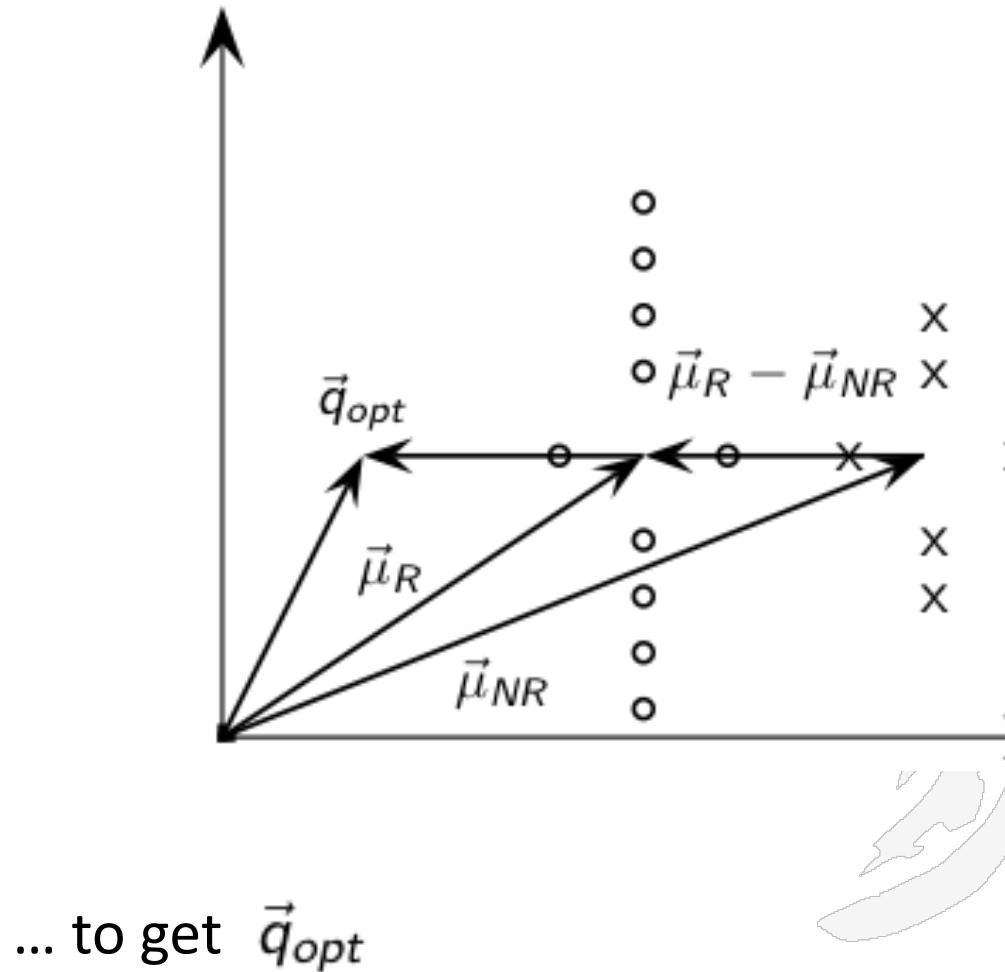


## 2D Rocchio Illustrated(6 of 9)

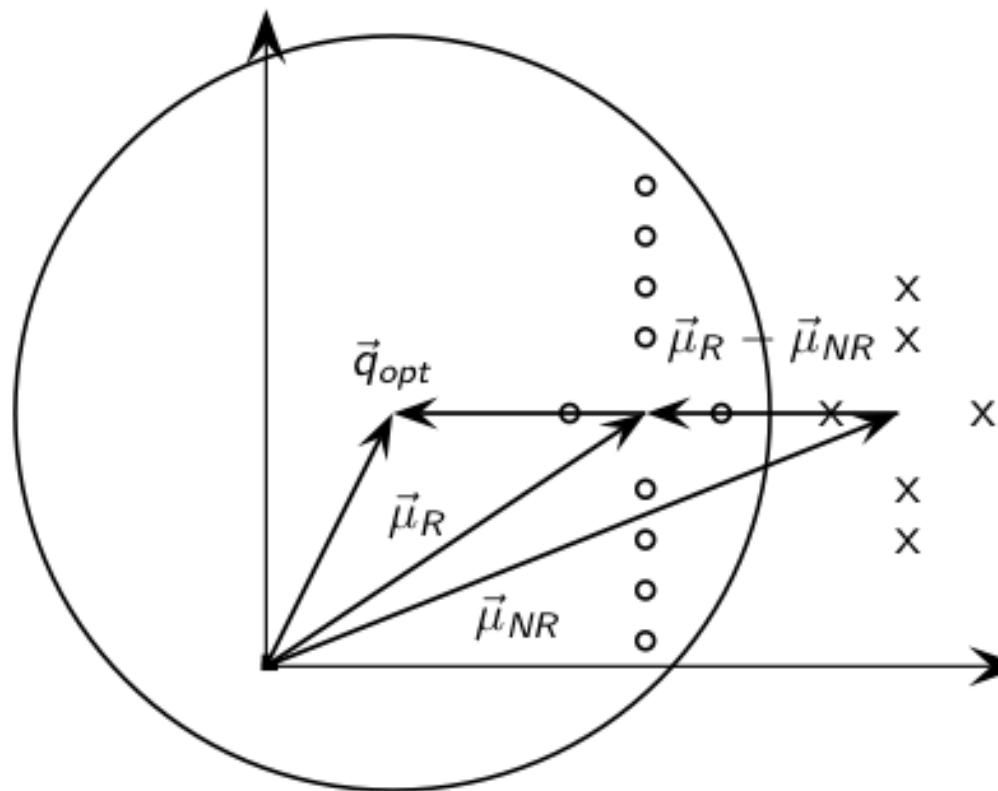


Add difference vector to  $\vec{\mu}_R$  ...

## 2D Rocchio Illustrated(7 of 9)



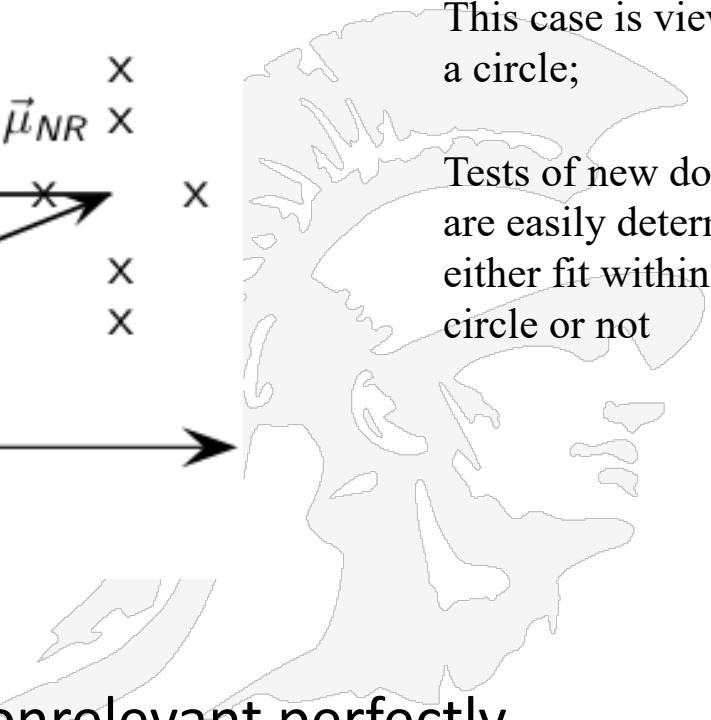
## 2D Rocchio Illustrated(8 of 9)



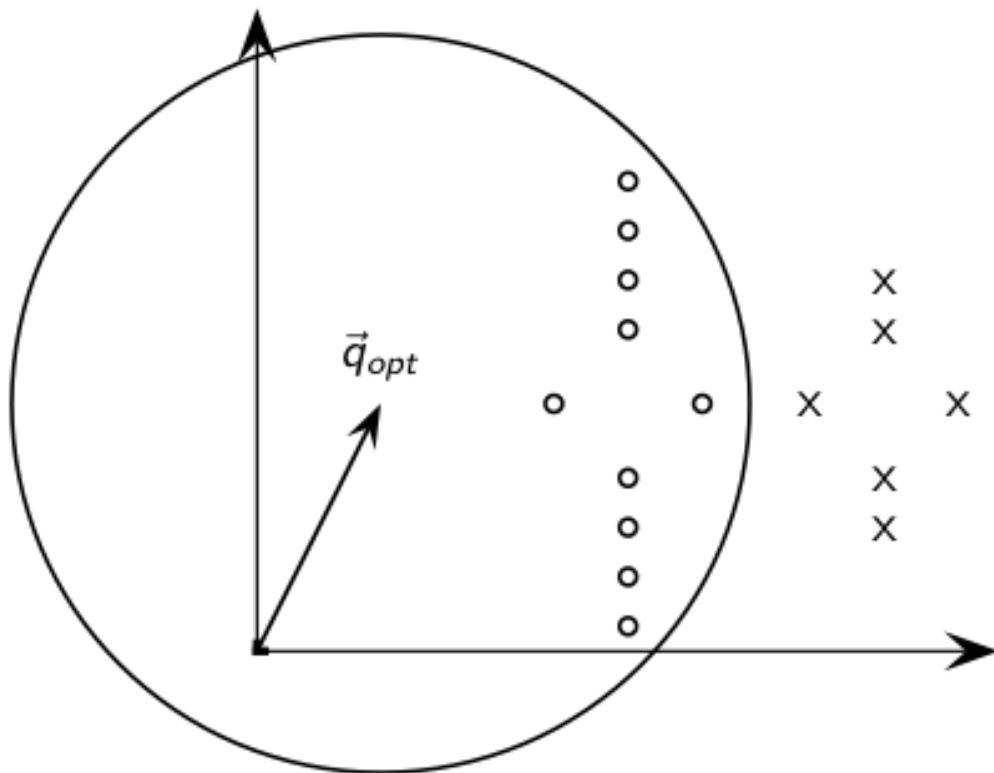
$\vec{q}_{opt}$  now separates relevant / nonrelevant perfectly.

Note that the boundary computed during the Rocchio algorithm in This case is viewed as a circle;

Tests of new documents are easily determined to either fit within the circle or not



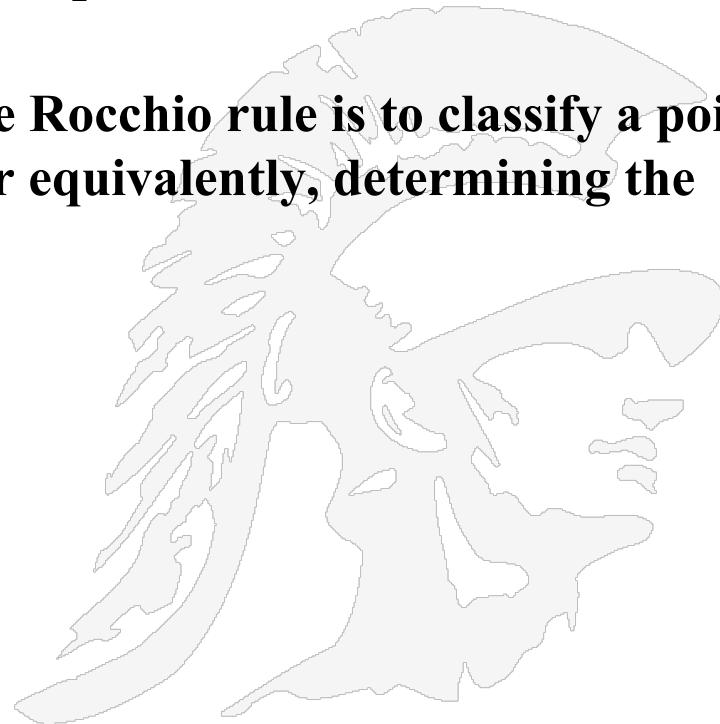
## 2D Rocchio Illustrated(9 of 9)



$\vec{q}_{opt}$  separates relevant / nonrelevant perfectly.

# Rocchio Algorithm Used for Classification

- More typically, the boundary determination in Rocchio is not a circle, but a hyperplane
- Given two centroids of two classes of documents, the boundary between the two classes is the set of points with equal distance from the two centroids
- Once the boundary is determined, the Rocchio rule is to classify a point according to the region it falls into, or equivalently, determining the centroid that the point is closest to



# Rocchio Classification Algorithm

*TrainRocchio( $C, D$ )*

**For each**  $c_j$  in  $C$

**do**  $D_j = \{d: \langle d, c_j \rangle \text{ is in } D\}$

$u_j = (1/|D_j|) * \sum d_j \text{ for } d_j \text{ in } D$

**return**  $\{u_1, \dots, u_j\}$

See Figure 14.4 in our textbook

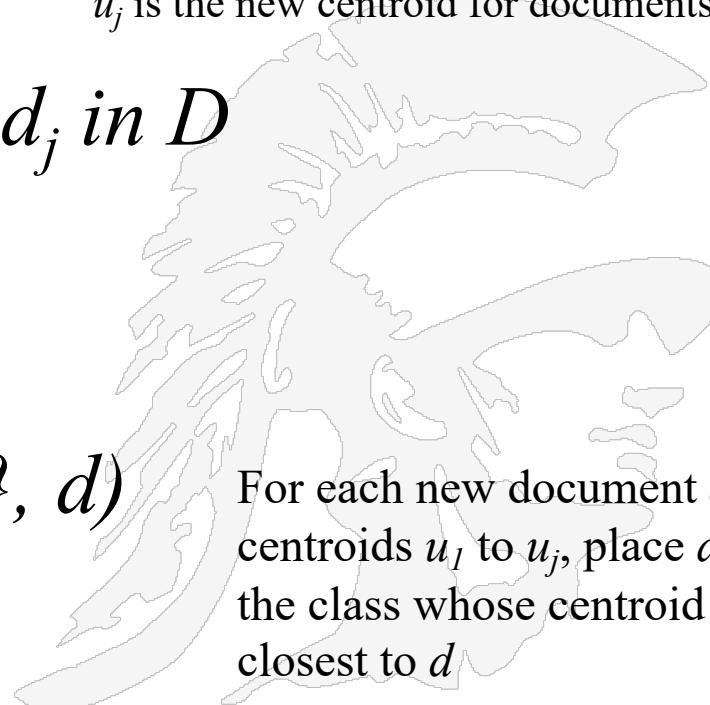
$D$  is the set of documents

$C$  is the set of classes determined earlier

$D_j$  is computed using Euclidean distance

and includes the set of documents in class  $c_j$

$u_j$  is the new centroid for documents in  $D_j$



*ApplyRocchio( $\{u_1, \dots, u_j\}, d$ )*

**return**  $\arg \min |u_j - d|$

For each new document  $d$  and centroids  $u_1$  to  $u_j$ , place  $d$  in the class whose centroid is the closest to  $d$

# Computing Time for Rocchio

- Mode
- Training
- testing

## Complexity

$$O(|D|L_{ave} + |C| |V|) \quad \leftarrow \text{Add all docs to their centroid, And compute the centroid}$$

$$O(L_a + |C| M_a) = O(|C| M_a) \leftarrow \# \text{centroids} * \# \text{types}$$

- $L_{ave}$  is the average number of tokens per document (*this is small*)
- $L_a$  and  $M_a$  are the numbers of tokens and types respectively in the test document
- The time to compute the Euclidean distance between the class centroid and a document is  $O(|C|M_a)$  (*the number of centroids times number of tokens*)
- Adding all documents to their respective (unnormalized) centroid is  $\Theta(|D|L_{ave})$  (as opposed to  $\Theta(|D||V|)$ ) since we need only consider non-zero entries.
- Dividing each vector sum by the size of its class to compute the centroid is  $\Theta(|V|)$ .

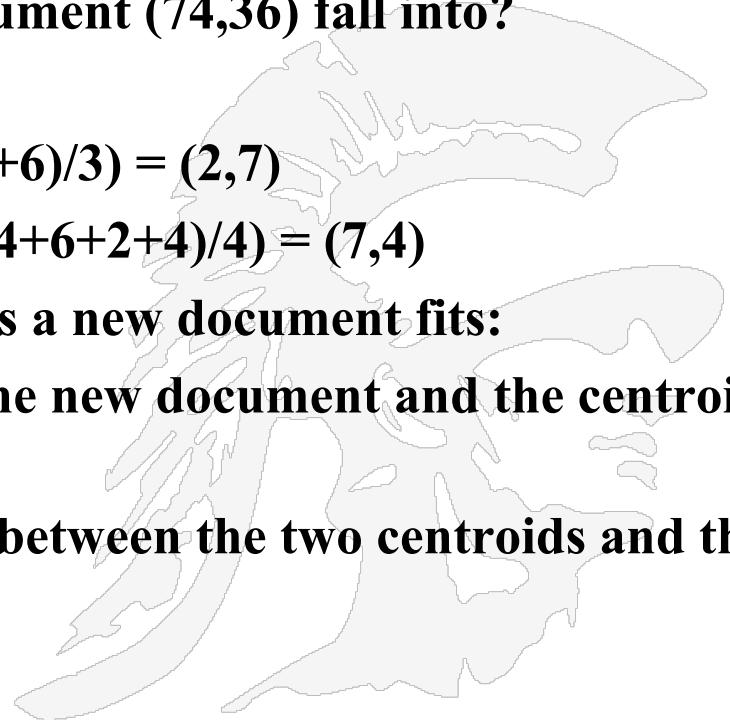
See Table 14.2 in our textbook

**Conclusion:** Overall training time is linear in the size of the collection

# Rocchio Example

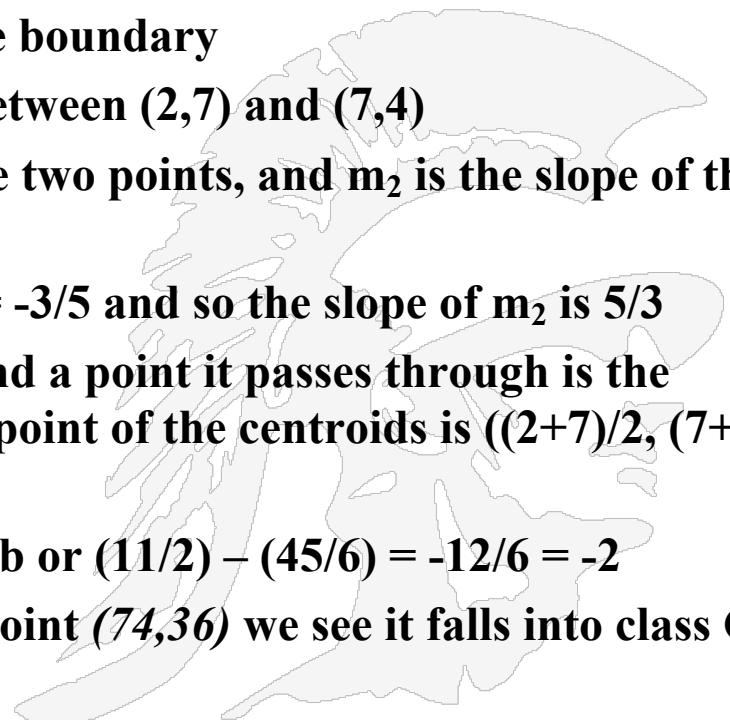
## Simple Case

- Given two classes, C1 and C2 and the document vectors are
- C1: (2,7), (3,8), (1,6) (3 documents)
- C2: (6,4), (8,6), (5,2), (9,4) (4 documents)
- Which class, C1 or C2 does the document (74,36) fall into?
- Training phase of Rocchio
  - centroid of C1,  $((2+3+1)/3, (7+8+6)/3) = (2,7)$
  - centroid of C2 =  $((6+8+5+9)/4, (4+6+2+4)/4) = (7,4)$
- Actual phase to find into which class a new document fits:
  - compute the distance between the new document and the centroids or alternatively
  - draw the perpendicular divisor between the two centroids and then see where the new vector falls



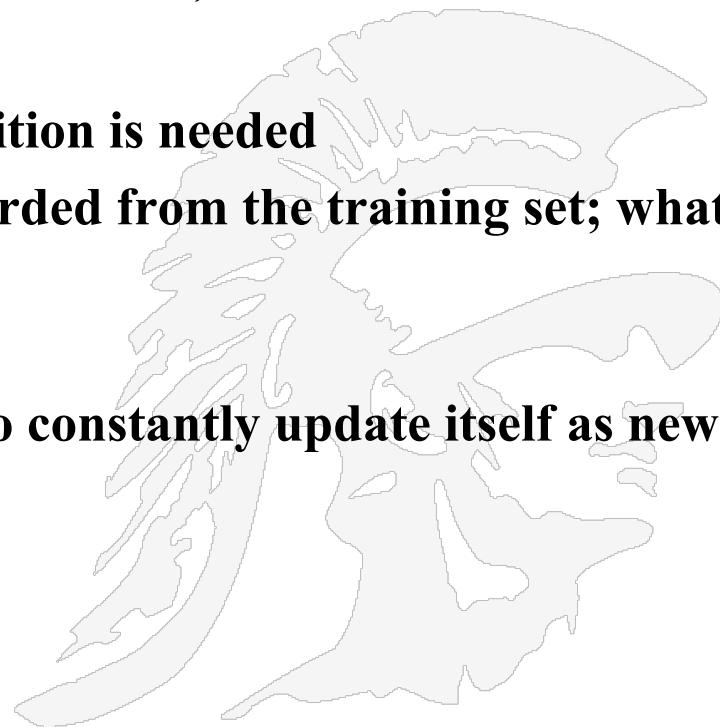
# Rocchio Example Simple Case

- To draw the perpendicular divisor between the two centroids and then see where the new vector falls recall:
  - The decision boundary will be a straight line, the perpendicular bisector of the two centroids;
  - The equation of the line defines the boundary
- Compute the perpendicular bisector between (2,7) and (7,4)
- if  $m_1$  is the slope of the line between the two points, and  $m_2$  is the slope of the perpendicular bisector, then  $m_1 m_2 = -1$
- Slope  $m_1 = (y_2 - y_1)/(x_2 - x_1) = (7 - 4)/(2 - 7) = -3/5$  and so the slope of  $m_2$  is  $5/3$
- The line is  $y = mx + b$  where  $m = 5/3$  and a point it passes through is the midpoint of the two centroids; the midpoint of the centroids is  $((2+7)/2, (7+4)/2) = (9/2, 11/2)$
- So the value of  $b$  is  $11/2 = (5/3)*(9/2) + b$  or  $(11/2) - (45/6) = -12/6 = -2$
- $y = (5/3)x - 2$ ; returning to the original point  $(74, 36)$  we see it falls into class C2



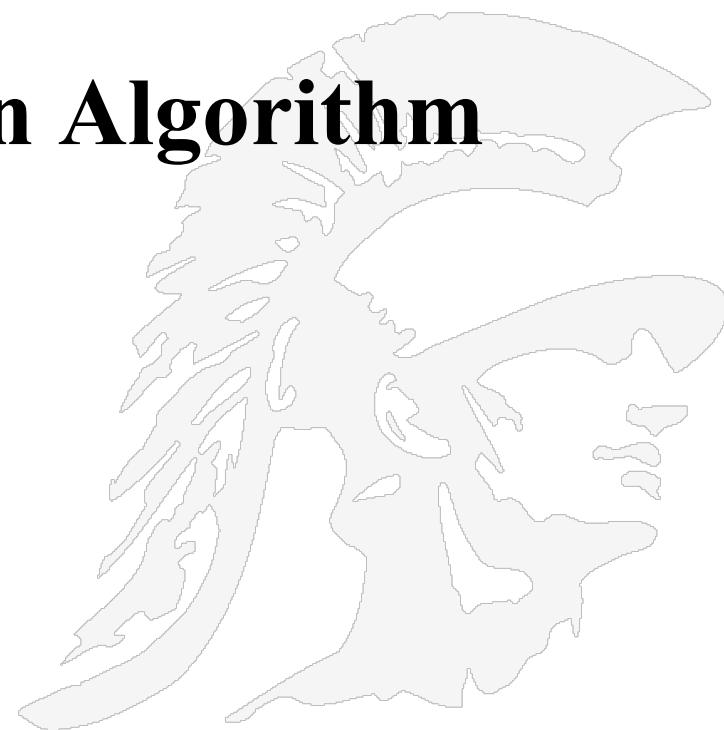
# Rocchio Extension

- Suppose we have new documents arriving, how can we update the centroids in constant time?
  - Suppose we had  $N$  documents with centroid  $(x,y)$  and a new document with coordinates  $(a, b)$  arrives; then the new centroid is:
  - $( (Nx + a)/(N+1), (Ny+b)/(N+1) )$
  - One multiplication and one addition is needed
- Now suppose one document is discarded from the training set; what is the new centroid
  - $( (Nx-a)/(N-1), (Ny-b)/(N-1) )$
- The online Rocchio algorithm has to constantly update itself as new documents are added and deleted



# $kNN$ - $k$ Nearest Neighbor Method

## Classification Algorithm



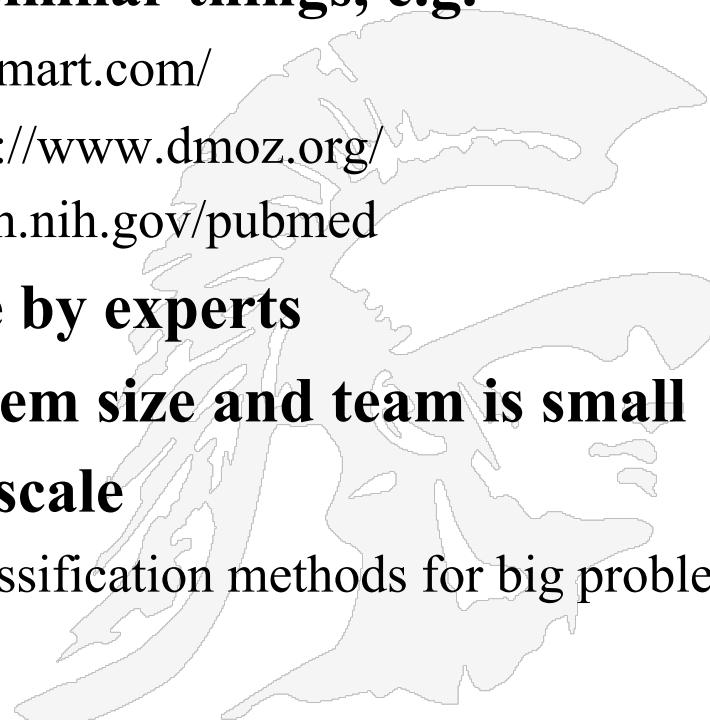
# Classification is Different from Clustering

- In general, in **classification** you have a set of predefined classes and want to know which class a new object (document) belongs to.
- Remember, **Clustering** tries to group a set of objects and find whether there is *some* relationship between the objects.
  - we already saw two algorithms for clustering, K-Means Algorithm and Agglomerative Clustering algorithm
- In the context of machine learning, classification is *supervised learning* and clustering is *unsupervised learning*
  - **Clustering** requires a. an algorithm, b. a similarity measure, and c. a number of clusters
  - **classification** has each document labeled in a class and an algorithm that assigns documents to one of the classes

# Classification Methods

- **Manual classification**

- Used by the original Yahoo! Directory
- **Other search engines did similar things, e.g.**
  - Looksmart, <http://www.looksmart.com/>
  - Open Directory Project, <https://www.dmoz.org/>
  - PubMed, <http://www.ncbi.nlm.nih.gov/pubmed>
- **Accurate when job is done by experts**
- **Consistent when the problem size and team is small**
- **Difficult and expensive to scale**
  - Means we need automatic classification methods for big problems

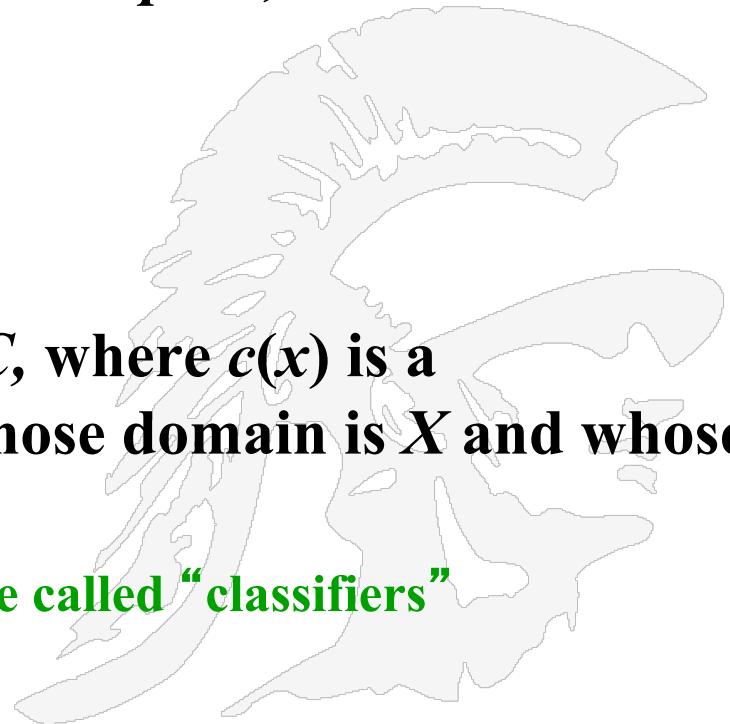


# The Problem Statement for Classification

- Given two things:
  1. A description of an instance,  $x \in X$ , where  $X$  is the *instance language* or *instance space*, and
  2. A fixed set of categories:

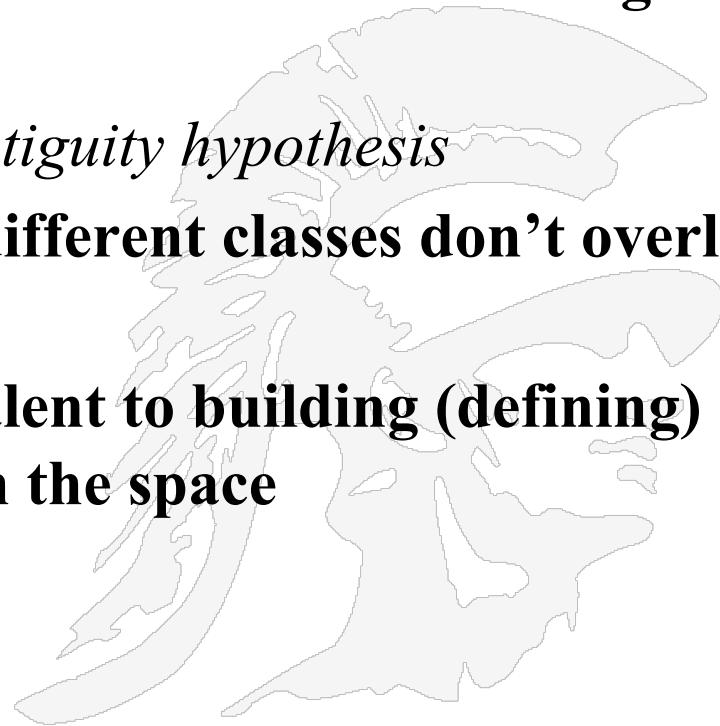
$$C = \{c_1, c_2, \dots, c_n\}$$

- Determine:
  - The category of  $x$ :  $c(x) \in C$ , where  $c(x)$  is a *categorization function* whose domain is  $X$  and whose range is  $C$ .
    - Functions that categorize are called “classifiers”



# Classification Using Vector Spaces

- In vector space classification, the training set corresponds to a labeled set of document vectors
- Premise 1: Documents in the same class form a contiguous region of space
  - This is referred to as the *contiguity hypothesis*
- Premise 2: Documents from different classes don't overlap (much)
- Learning a classifier is equivalent to building (defining) surfaces to delineate classes in the space



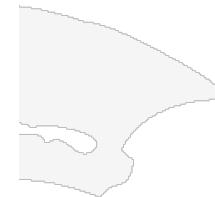
# Ways to Measure Distance

*For normalized vectors Euclidean distance and cosine similarity correspond*

## Distance functions

**Euclidean**

$$\sqrt{\sum_{i=1}^k (x_i - y_i)^2}$$

**Manhattan**

$$\sum_{i=1}^k |x_i - y_i|$$

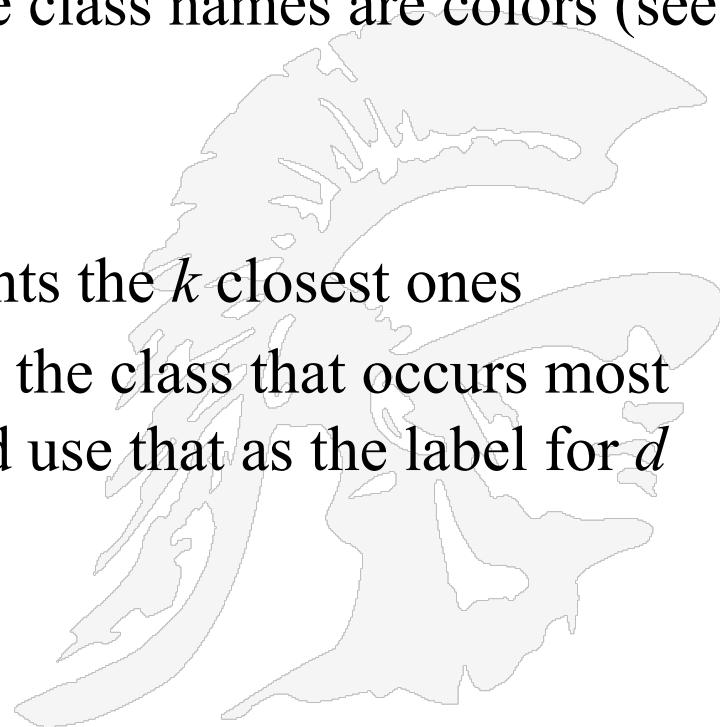
**Minkowski**

$$\left( \sum_{i=1}^k (|x_i - y_i|)^q \right)^{1/q}$$



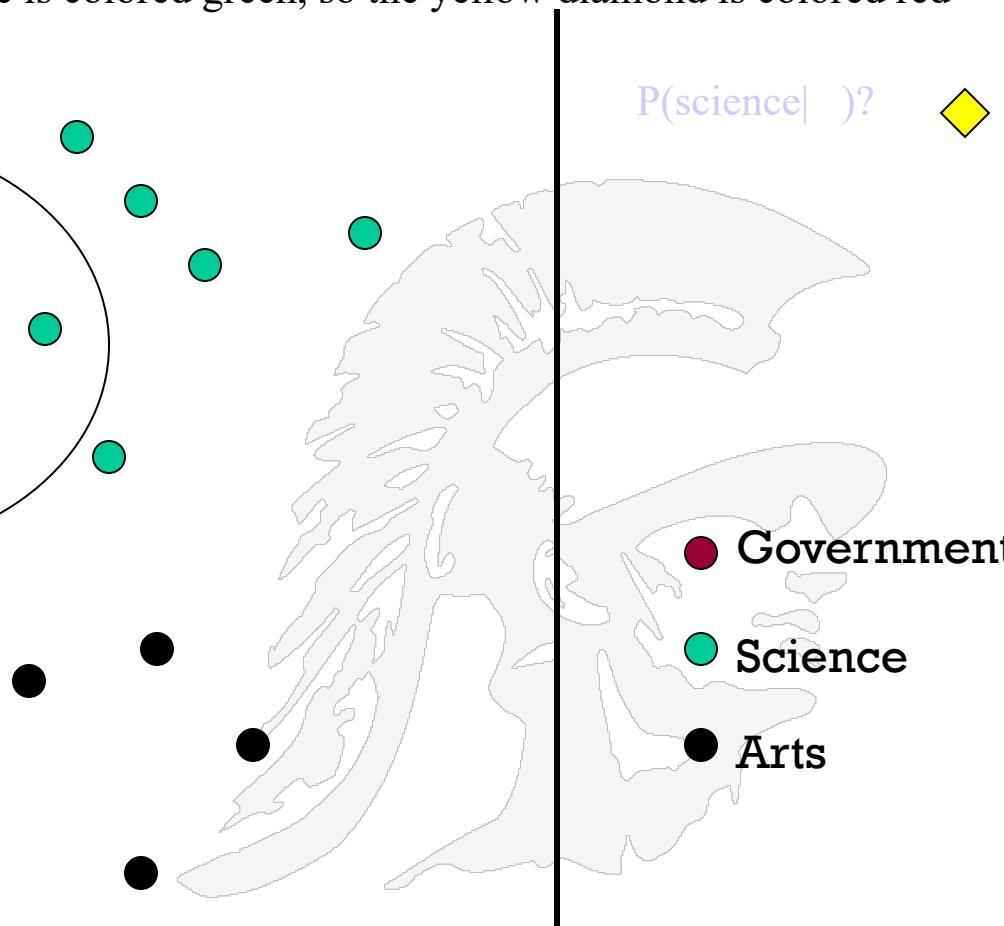
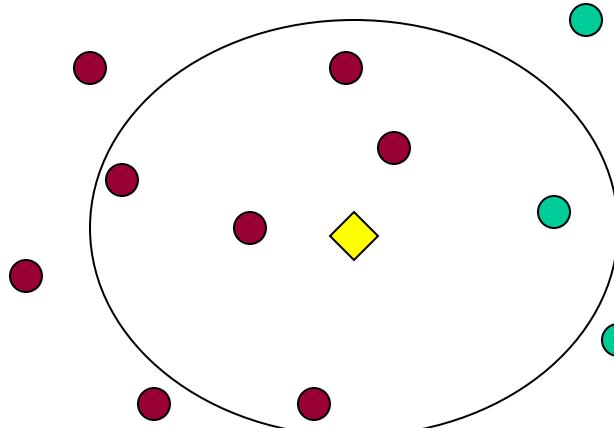
# **$k$ Nearest Neighbor Classification Algorithm**

- Initially we assume we have a set of  $N$  documents that have already been classified
  - the WDM videos assume the class names are colors (see the Schedule of Lectures)
- To classify a document  $d$ 
  - locate among the  $N$  documents the  $k$  closest ones
  - from these  $k$  neighbors, pick the class that occurs most often, the majority class, and use that as the label for  $d$



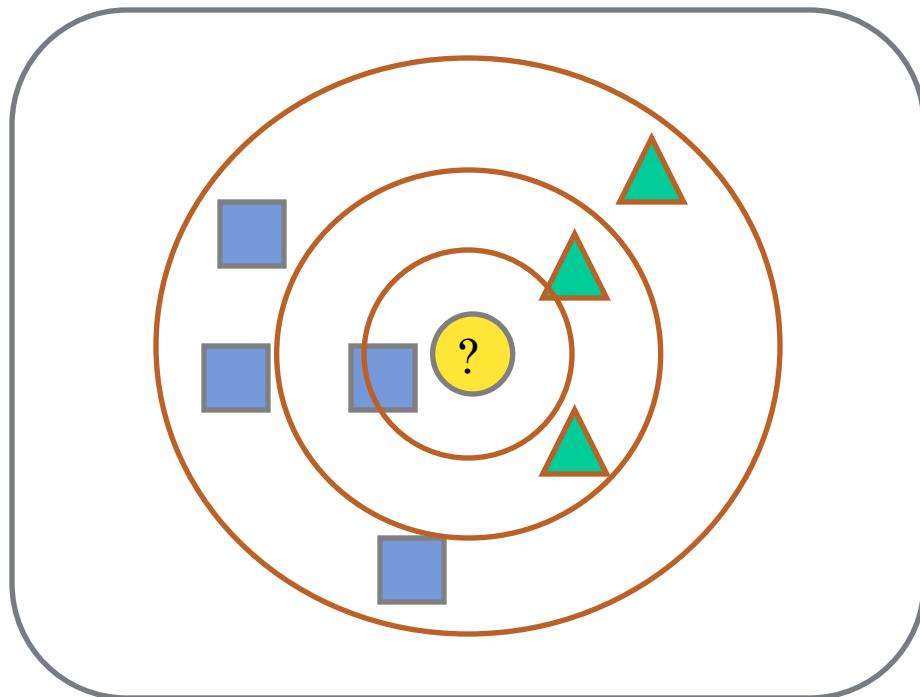
# Example: $k=6$ ( $6NN$ )

5 neighbors are colored red, one is colored green, so the yellow diamond is colored red

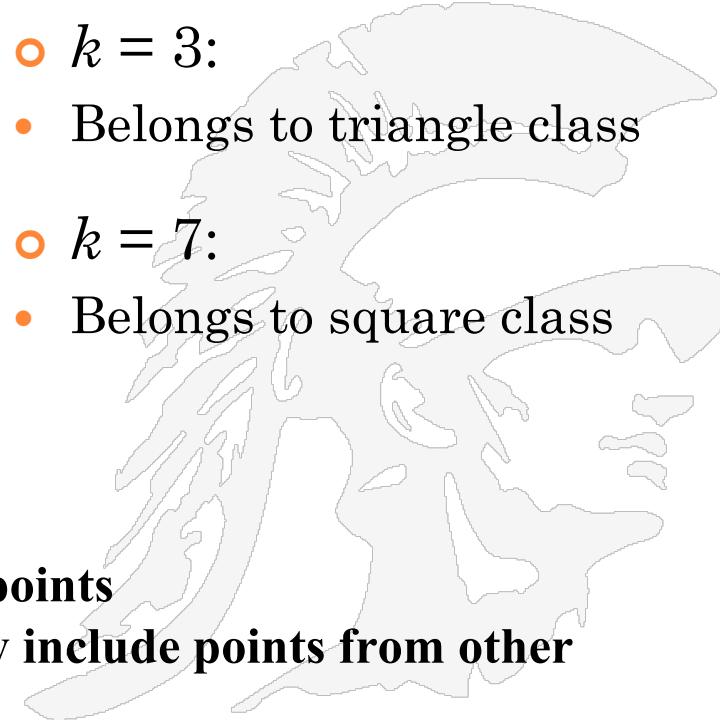


When  $k=1$ , the document is assigned  
to its nearest neighbor

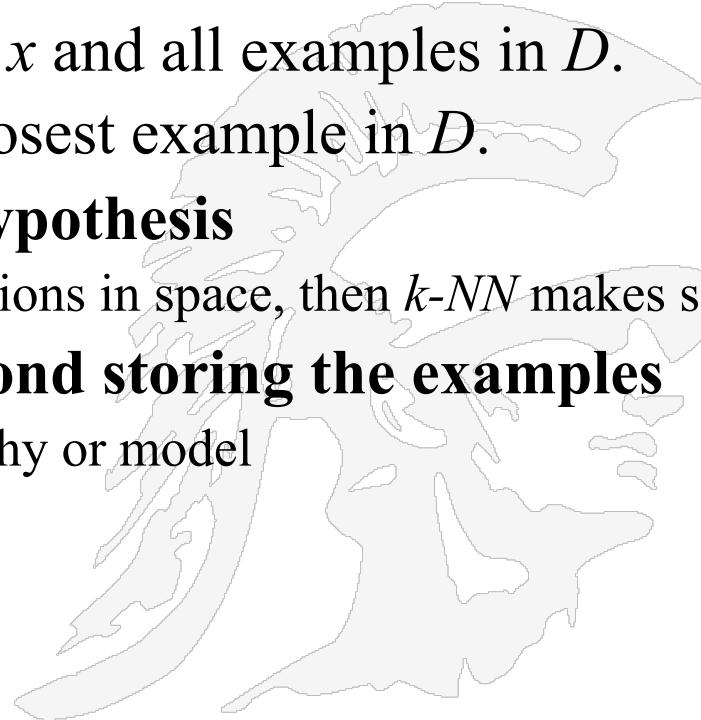
# K-Nearest Neighbor Another Example



- **Choosing the value of  $k$ :**
  - If  $k$  is too small, sensitive to noise points
  - If  $k$  is too large, neighborhood may include points from other classes
  - Choose an odd value for  $k$ , to eliminate ties

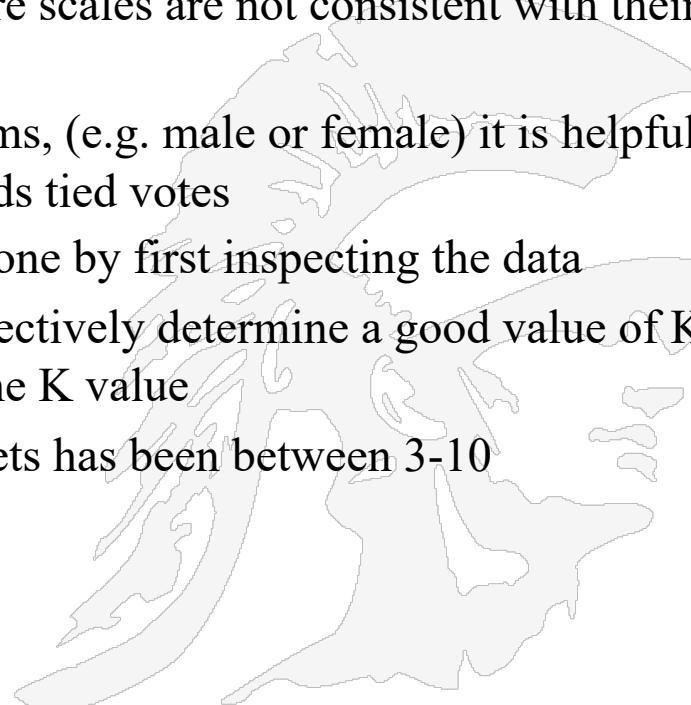


- **Learning:** there is no learning step; just store the labeled training examples  $D$
- **Testing instance  $x$  (*under 1-NN*):**
  - Compute the distance between  $x$  and all examples in  $D$ .
  - Assign  $x$  the category of the closest example in  $D$ .
- **Rationale of  $k$ -NN: contiguity hypothesis**
  - if documents do form contiguous regions in space, then  $k$ -NN makes sense
- **Does not compute anything beyond storing the examples**
  - we are NOT determining any hierarchy or model
- **$K$ -NN has also been called:**
  - Case-based learning
  - Memory-based learning
  - Lazy learning



# Choice of $K$

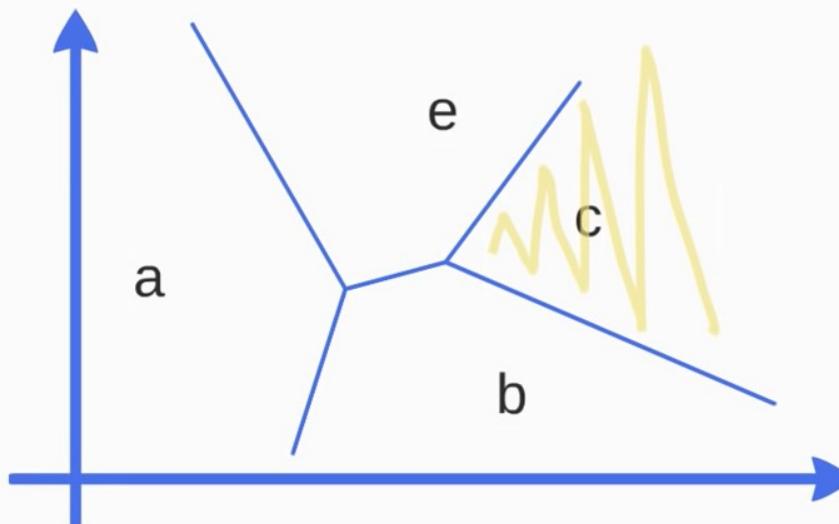
- The best choice of  $k$  depends upon the data;
  - generally, larger values of  $k$  reduce the effect of noise on the classification, but make boundaries between classes less distinct.
- The accuracy of the  $k$ -NN algorithm can be severely degraded by the presence of noisy or irrelevant features, or if the feature scales are not consistent with their importance
- In binary (two class) classification problems, (e.g. male or female) it is helpful to choose  $k$  to be an odd number as this avoids tied votes
- Choosing the optimal value for  $k$  is best done by first inspecting the data
- Cross-validation is another way to retrospectively determine a good value of  $K$  by using an independent dataset to validate the  $K$  value
- Historically, the optimal  $K$  for most datasets has been between 3-10



# Voronoi Diagram

For the k-Nearest Neighbor Algorithm,  $k = 1$  is a special case

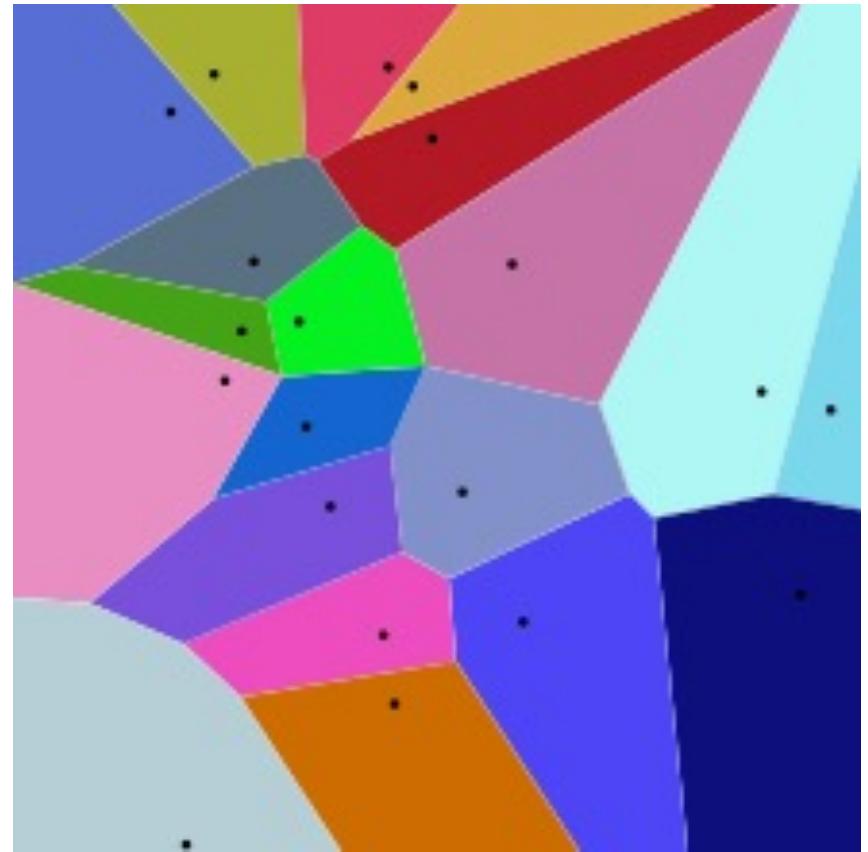
When  $k = 1$ , each training vector defines a region in space, defining a *Voronoi* partition of the space



$$R_i = \{x : d(x, x_i) < d(x, x_j), i \neq j\}$$

# When $k=1$ – A Special Case

- A **Voronoi diagram** is a partitioning of a plane into regions based on distance to points in a specific subset of the plane
- Decision boundaries in  $1\text{-NN}$  are concatenated segments of a Voronoi tessellation (e.g. polygons)
- The set of points (called class labels) is specified beforehand
- For each class label there is a corresponding region consisting of all points closer to that class label than to any other. These regions are called Voronoi cells



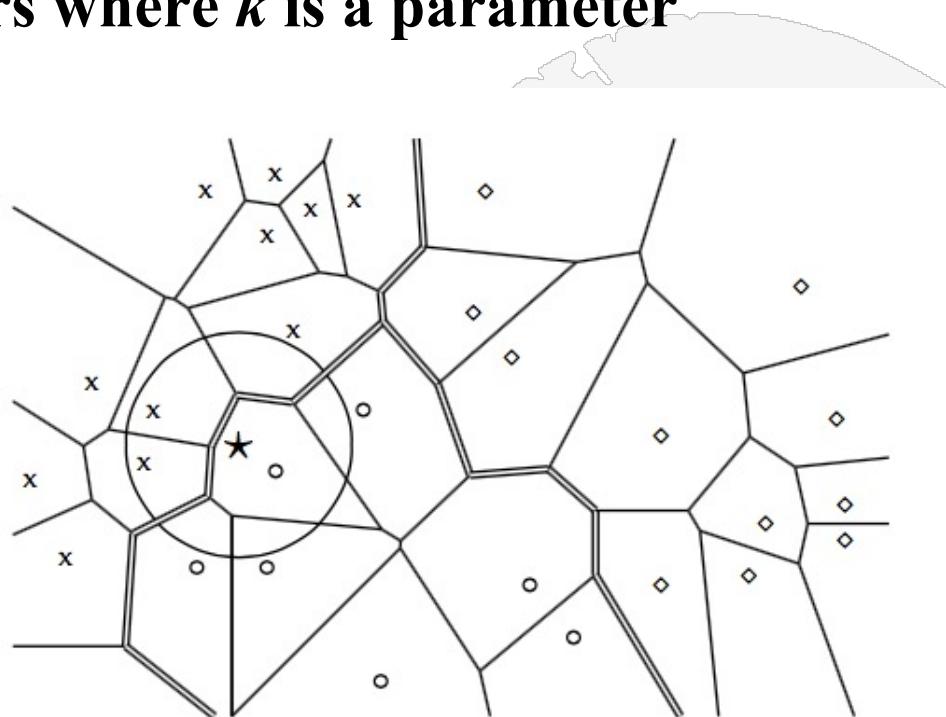
20 points (class labels) and their Voroni regions;  
Line segments are all points equidistant to three  
or more regions

## $K=1$ Nearest Neighbor Regions are Polygons

- For  $1\text{-NN}$  we assign each document to the class of its closest neighbor
- For  $k\text{-NN}$  we assign each document to the majority class of its  $k$  closest neighbors where  $k$  is a parameter

The two classes are: X and circle, and the star document is falling into the circle area;  
Double lines define the regions in space where documents are similar;  
think of each region as defining a cellphone tower

**$K\text{-NN}$  is an example of a non-linear classifier; (Rocchio is a linear classifier)**



- Training a  $kNN$  classifier simply consists of determining  $k$  and pre-processing the documents
- If we preselect  $k$  and do not pre-process, then  $kNN$  requires no training at all
- It makes more sense to pre-process training documents once as part of the training phase rather than repeatedly every time we classify a new test document



### **$kNN$ with preprocessing of training set**

training  $\Theta(|D|L_{ave})$

testing  $\Theta(L_a + |D|M_{ave}M_a) = \Theta(|D|M_{ave}M_a)$

### **$kNN$ without preprocessing of training set**

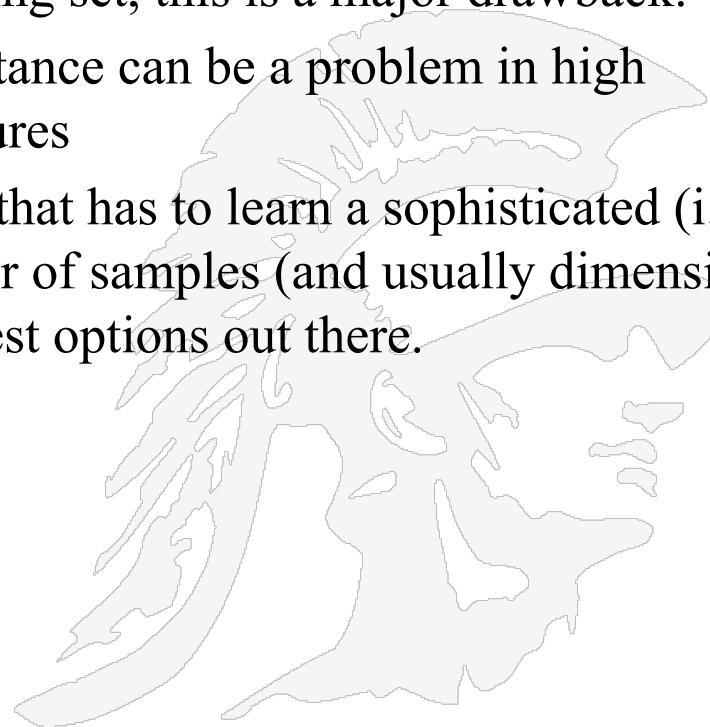
training  $\Theta(1)$

testing  $\Theta(L_a + |D|L_{ave}M_a) = \Theta(|D|L_{ave}M_a)$

**Recall from previous slide:**  $L_{ave}$  is the average number of tokens per document  
 $L_a$  and  $M_a$  are the numbers of tokens and types respectively in the test document

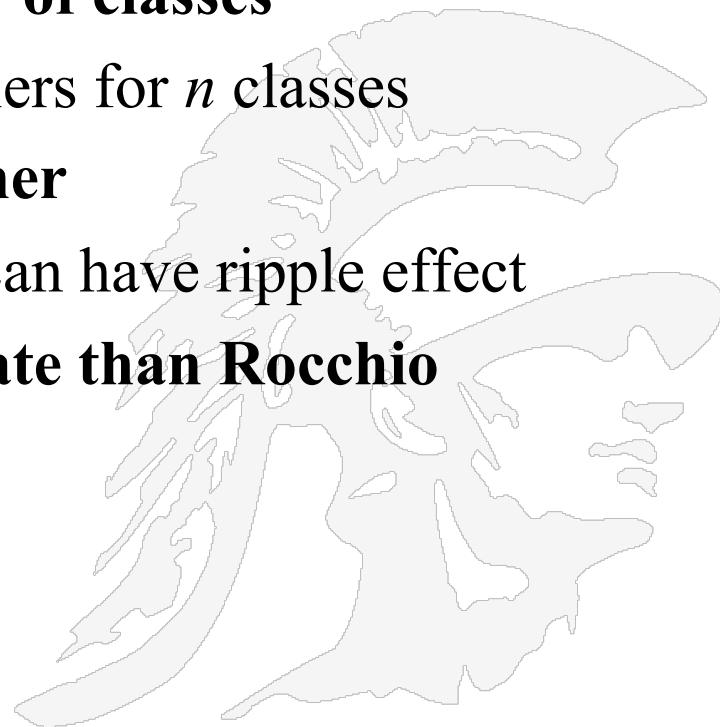
# Observations on $k$ -NN

- The K-Nearest-Neighbor model has two major drawbacks.
    1. **performance.**
      - you have to load all of your training data and calculate distances to all training samples; Over a big training set, this is a major drawback.
    2. **distance metric.** Using Euclidean distance can be a problem in high dimensions as well as with noisy features
- **To summarize**, if you have a system that has to learn a sophisticated (i.e. nonlinear) pattern with a small number of samples (and usually dimensions),  $k$ -NN models are usually one of the best options out there.



# K-NN: Final Points

- **No feature selection necessary**
- **No training necessary**
- **Scales well with large number of classes**
  - Don't need to train  $n$  classifiers for  $n$  classes
- **Classes can influence each other**
  - Small changes to one class can have ripple effect
- **In most cases it's more accurate than Rocchio**



# Algorithm Comparison

K-Means	K-Nearest Neighbors
Clustering algorithm	Classification Algorithm
Uses distance from data points to k-centroids to cluster data into k-groups.	Calculates k nearest data points from data point X. Uses these points to determine which class X belongs to
Centroids are not necessarily data points.	“Centroid” is the point X to be classified.
Updates centroid on each pass by calculations over all data in a class.	Data point to be classified remains the same.
Must iterate over data until center point doesn't move.	Only requires k distance calculations.



# Recommendation Systems



**Collaborative Filtering &  
Content-Based Recommendations**

# Recommendations are Found Everywhere

Amazon.com: The Age of Spiritual Machines: When Computers Exceed Human Intelligence (9780140282023): Ray Kurzweil: Books - Mozilla Firefox

File Edit View History Bookmarks Tools Help  
[www.amazon.com/Age-Spiritual-Machines-Computers-Intel](#) Freecorder Customized Web S Home

Amazon.com: The Age of Spiritual Machines:... +

Hello, Ellis Horowitz. We have [recommendations](#) for you. ([Not Ellis?](#))  
[Ellis's Amazon.com](#) |  Today's Deals | [Gifts & Wish Lists](#) | [Gift Cards](#)

amazon.com

Shop All Departments Search Books Books Advanced Search Browse Subjects New Releases Best Sellers The New York Times® Best Sellers

The Age of Spiritual Machines and over one million other books are available for Ama

**Click to LOOK INSIDE!**

**THE AGE OF SPIRITUAL MACHINES**  
 WHEN COMPUTERS EXCEED HUMAN INTELLIGENCE  
 RAY KURZWEIL

**The Age of Spiritual Machines: When Computers Exceed Human Intelligence**  
 [Paperback]  
 Ray Kurzweil (Author)

★★★★★ (183)  
 List Price: \$18.00  
 Price: \$12.24  
 Shipping:  
 You Save: \$5.76 (3)  
**In Stock.**  
 Ships from and sold  
 Want it delivered next 45 hours and 0 checkout. [Details](#)  
 44 new from \$9.20

Amazon.com: The Age of Spiritual Machines: When Computers Exceed Human Intelligence (9780140282023): Ray Kurzweil: Books - Mozilla Firefox

File Edit View History Bookmarks Tools Help  
[www.amazon.com/Age-Spiritual-Machines-Computers-Intelligence/dp/0140282025](#) Freecorder Customized Web S Home

Amazon.com: The Age of Spiritual Machines:... +

Customers Who Bought This Item Also Bought

The Singularity Is Near: When Humans Transcend Bi... by Ray Kurzweil  
 ★★★★★ (187) \$12.78

In the Garden of Beasts: Love, Terror, and an Amer... by Erik Larson  
 ★★★★★ (636) \$15.60

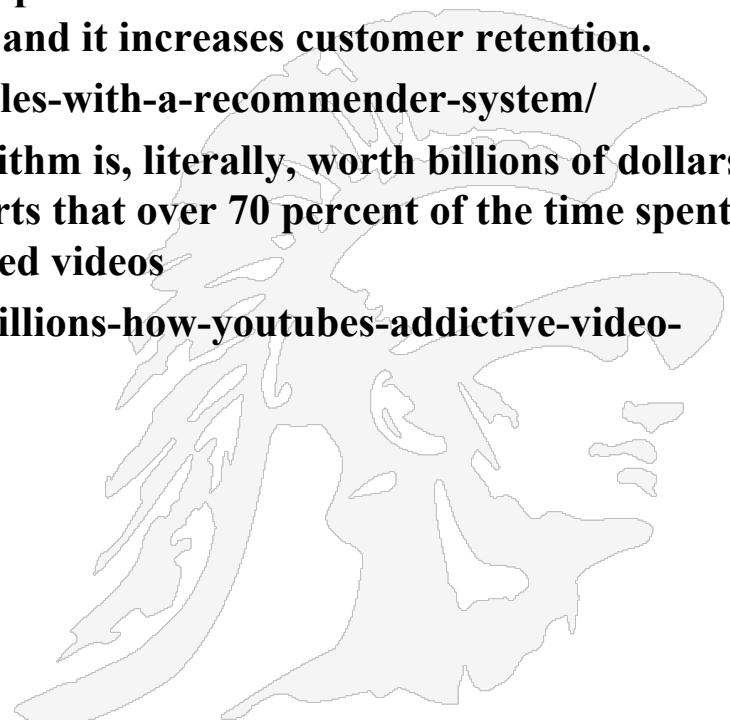
In the Plex: How Google Thinks, Works, and Shapes... by Steven Levy  
 ★★★★★ (55) \$25.95

The Shadow Party: How George Soros, Hillary Cli... by David Horowitz  
 ★★★★★ (85) \$10.87

Secret Weapon: How Economic Terrorism Brought... by Kevin D. Freeman  
 ★★★★★ (33) \$18.45

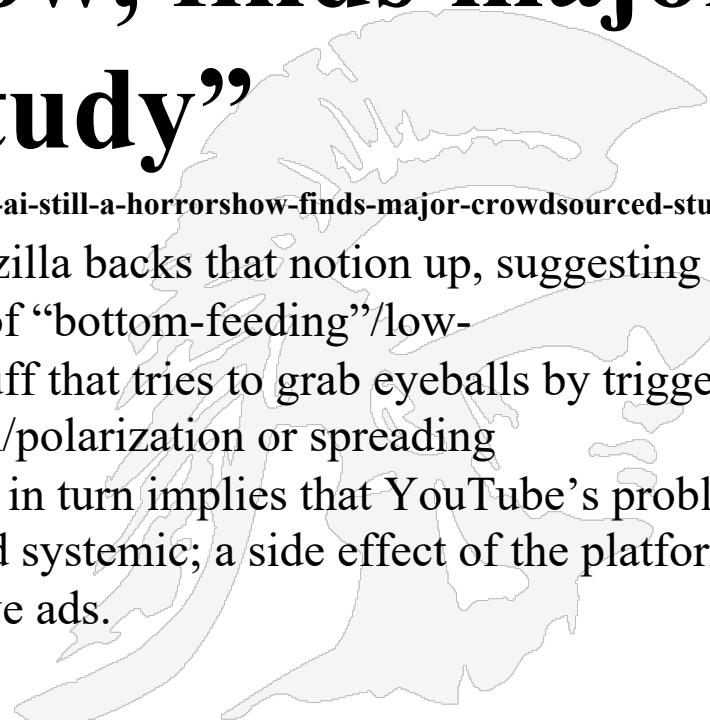
# Value of Recommendation Systems

- **Recommendation Systems have been shown to substantially increase sales at on-line stores**
  - Recommendation engines can significantly increase revenue, improve CTRs and conversions. It also contributes to the improvement of factors more difficult to measure, such as customer satisfaction, and it increases customer retention.
  - <https://neoteric.eu/blog/how-to-boost-sales-with-a-recommender-system/>
  - “YouTube’s video recommending algorithm is, literally, worth billions of dollars:” the company’s CPO, Neal Mohan, reports that over 70 percent of the time spent on YouTube is spent watching recommended videos
  - <https://faun.pub/the-algorithm-worth-billions-how-youtubes-addictive-video-recommender-works-d75646dac6a3>



- “YouTube’s recommender AI still a horror show, finds major crowdsourced study”

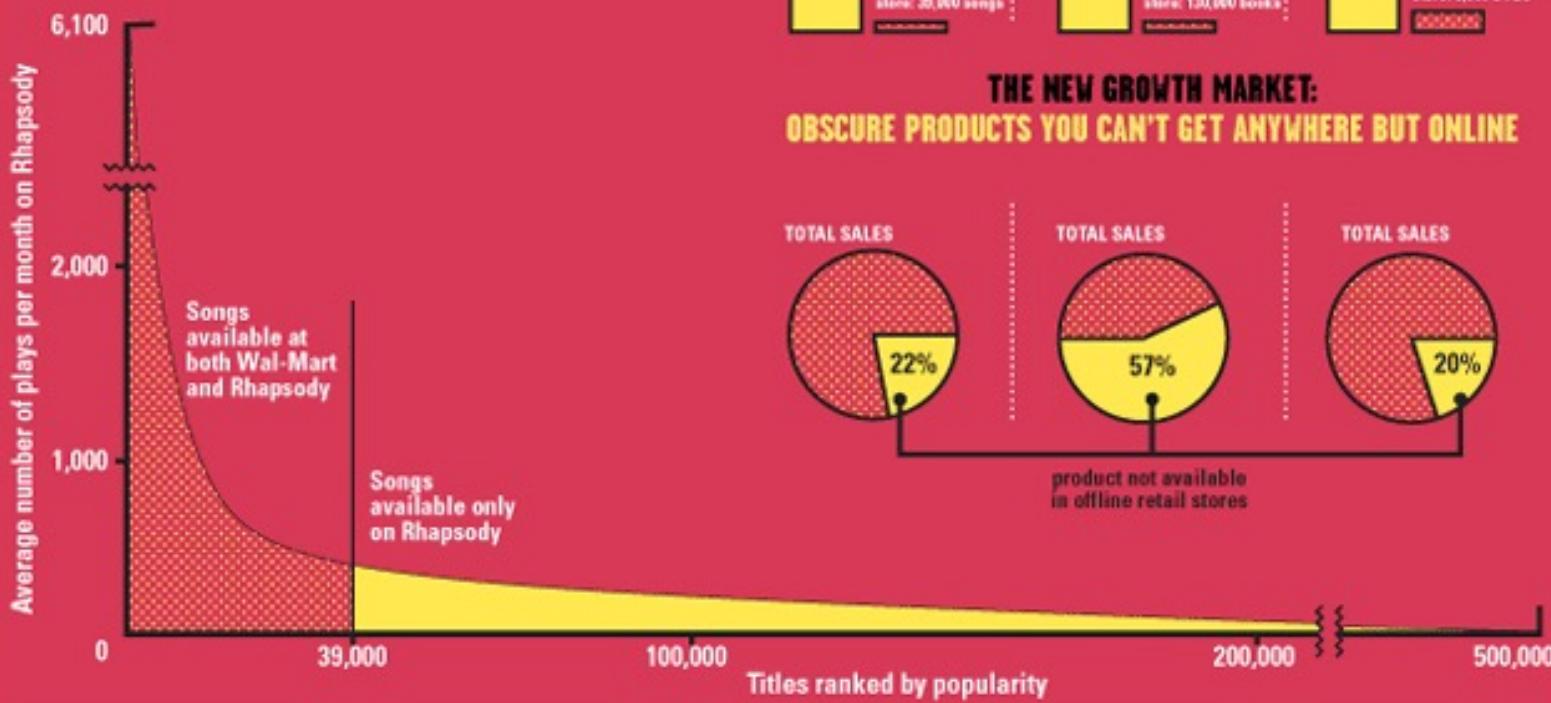
- <https://techcrunch.com/2021/07/07/youtubes-recommender-ai-still-a-horrorshow-finds-major-crowdsourced-study/>
- New research published July 2021 by Mozilla backs that notion up, suggesting YouTube’s AI continues to puff up piles of “bottom-feeding”/low-grade/divisive/disinforming content — stuff that tries to grab eyeballs by triggering people’s sense of outrage, sewing division/polarization or spreading baseless/harmful disinformation — which in turn implies that YouTube’s problem with recommending terrible stuff is indeed systemic; a side effect of the platform’s rapacious appetite to harvest views to serve ads.



# Why Recommendation Systems Are Needed

## ANATOMY OF THE LONG TAIL

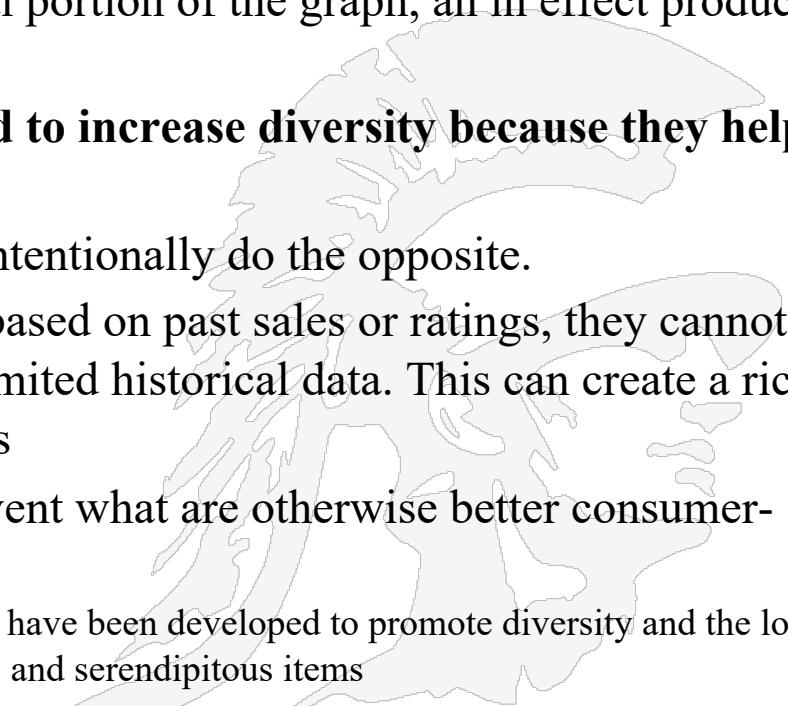
Online services carry far more inventory than traditional retailers. Rhapsody, for example, offers 19 times as many songs as Wal-Mart's stock of 39,000 tunes. The appetite for Rhapsody's more obscure tunes (charted below in yellow) makes up the so-called Long Tail. Meanwhile, even as consumers flock to mainstream books, music, and films (right), there is real demand for niche fare found only online.



Sources: Erik Brynjolfsson and Jeffrey H. Hu, MIT, and Michael Smith, Carnegie Mellon; Barnes & Noble; Netflix; ReelNetworks

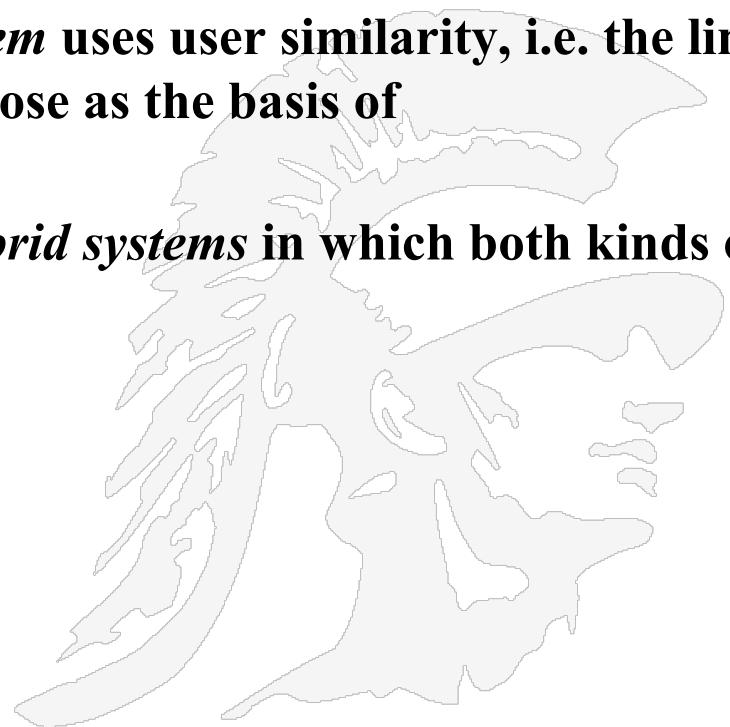
# Scarcity versus Abundance

- Since online systems maintain large quantities of goods, systems that provide recommendations serve an important purpose
  - In some cases items sold from the long tail, (i.e. those not particularly popular) can cumulatively outweigh the initial portion of the graph, an in effect produce the majority of sales
- Recommendation systems are expected to increase diversity because they help us discover new products.
  - However, some algorithms may unintentionally do the opposite.
  - Because they recommend products based on past sales or ratings, they cannot usually recommend products with limited historical data. This can create a rich-get-richer effect for popular products
  - This bias toward popularity can prevent what are otherwise better consumer-product matches.
    - Several collaborative filtering algorithms have been developed to promote diversity and the long tail by recommending novel, unexpected, and serendipitous items
  - See [https://en.wikipedia.org/wiki/Collaborative\\_filtering](https://en.wikipedia.org/wiki/Collaborative_filtering)



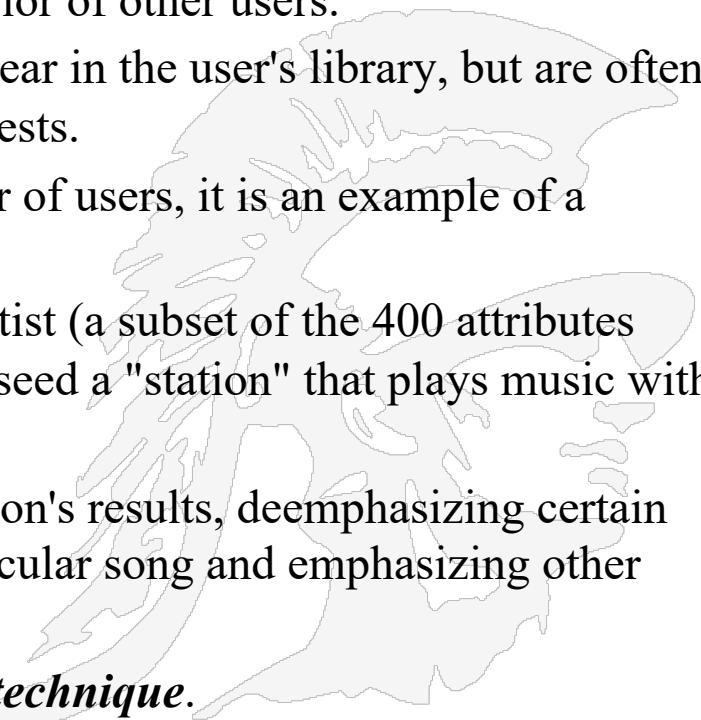
## Two Types of Recommendation Systems

- A *recommendation system* is any system which provides a recommendation/prediction/opinion to a user on items
- 1. A classic *content-based filtering system* uses item similarity/clustering to recommend items like ones you like
- 2. A classic *collaborative filtering system* uses user similarity, i.e. the links between users and the item they chose as the basis of recommendations
- Commonly many companies use *hybrid systems* in which both kinds of techniques are employed



## Difference between Collaborative and Content-based Filtering-An Example

- *Here are two early systems that recommended music*
- *Last.fm* creates a "station" of recommended songs by observing what bands and individual tracks **the user** has listened to on a regular basis (*user similarity*) and comparing those against the listening behavior of other users.
  - Last.fm will play tracks that do not appear in the user's library, but are often played by other users with similar interests.
  - As this approach leverages the behavior of users, it is an example of a **collaborative filtering technique**.
- *Pandora* uses the **properties of a song** or artist (a subset of the 400 attributes provided by the Music Genome Project) to seed a "station" that plays music with similar properties (*item similarity*).
  - User feedback is used to refine the station's results, deemphasizing certain attributes when a user "dislikes" a particular song and emphasizing other attributes when a user "likes" a song.
  - This is an example of a **content-based technique**.



# An Example Restaurant Recommendations

- Suppose we have a list of all Los Angeles restaurants
  - with  $\uparrow$  and  $\downarrow$  ratings for *some*
  - as provided by USC students
- Which restaurant(s) should I recommend to you?

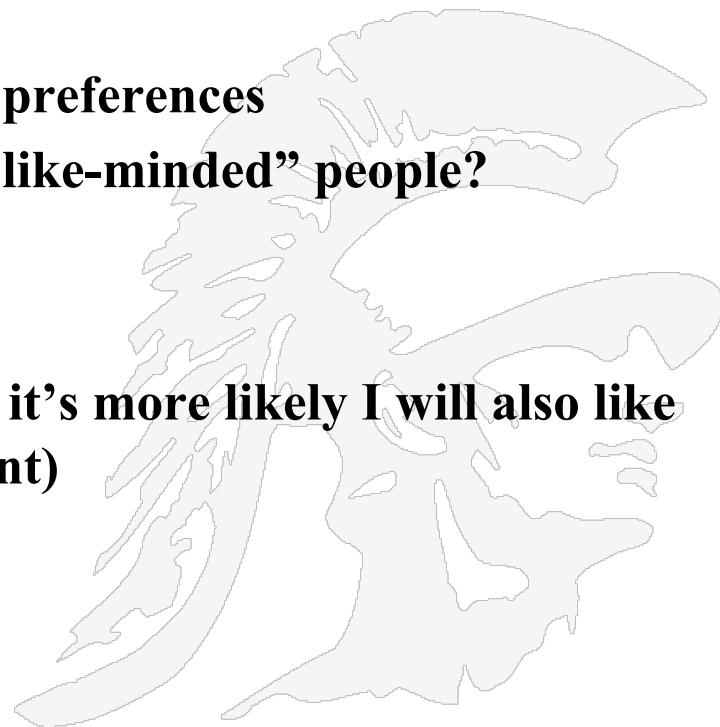


# Input

Alice	Il Fornaio	Yes
Bob	Ming's	No
Cindy	Straits Café	No
Dave	Ming's	Yes
Alice	Straits Café	No
Estie	Zao	Yes
Cindy	Zao	No
Dave	Brahma Bull	No
Dave	Zao	Yes
Estie	Ming's	Yes
Fred	Brahma Bull	No
Alice	Mango Café	No
Fred	Ramona's	No
Dave	Homma's	Yes
Bob	Higashi West	Yes
Estie	Straits Café	Yes

# Algorithm 0

- **Strategy:** Recommend to you the most popular restaurants
  - say # positive votes minus # negative votes
- **But this ignores**
  - your culinary preferences
  - judgments of those with similar preferences
- **How can we exploit the wisdom of “like-minded” people?**
- **Basic assumption**
  - Preferences are not random
  - **Assumption:** if I like Il Fornaio, it's more likely I will also like Cenzo (another Italian restaurant)



# Cast the Input as a Matrix

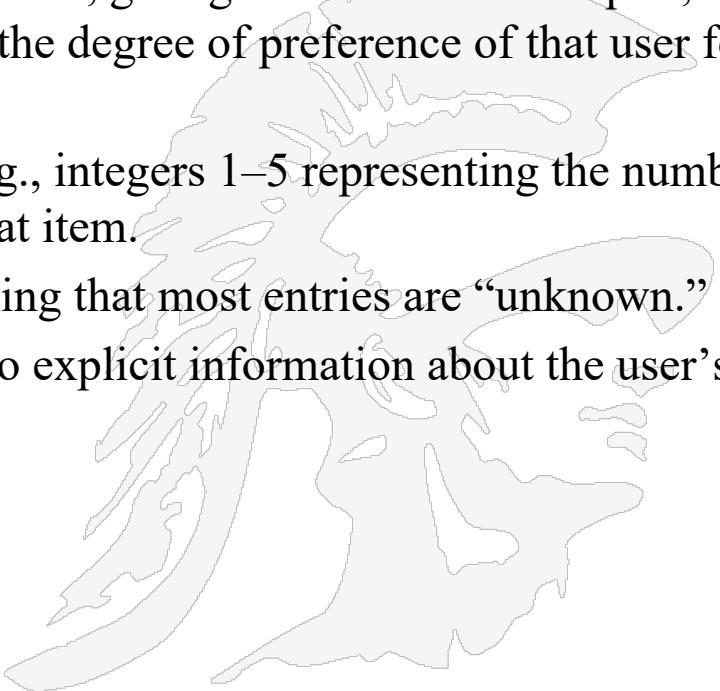
	Brahma Bull	Higashi West	Mango	Il Fornaio	Zao	Ming's	Ramona's	Straits	Homma's
Alice		Yes	No	Yes				No	
Bob		Yes				No		No	
Cindy				Yes	No			No	
Dave	No			No	Yes	Yes			Yes
Estie				No	Yes	Yes		Yes	
Fred	No						No		

Called a *utility matrix*

Each row represents an individual and each column a restaurant  
 In this example, entries are either yes/no;  
 In the more general case they can be any value

# The Utility Matrix

- In a recommendation-system application there are two classes of entities, which we shall refer to as *users* and *items*
  - Users have preferences for certain items, and these preferences must be teased out of the data.
- The data itself is represented as a *utility matrix*, giving for each user-item pair, a value that represents what is known about the degree of preference of that user for that item.
- *Values might come from an ordered set*, e.g., integers 1–5 representing the number of stars that the user gave as a rating for that item.
- We assume that the matrix is *sparse*, meaning that most entries are “unknown.”
- An unknown rating implies that we have no explicit information about the user’s preference for the item.



# Now That We Have a Matrix

	Brahma Bull	Higashi West	Mango	Il Fornaio	Zao	Ming's	Ramona's	Straits	Homma's
Alice		1	-1	1				-1	
Bob		1				-1		-1	
Cindy				1	-1			-1	
Dave	-1			-1	1	1			1
Estie				-1	1	1		1	
Fred	-1						-1		

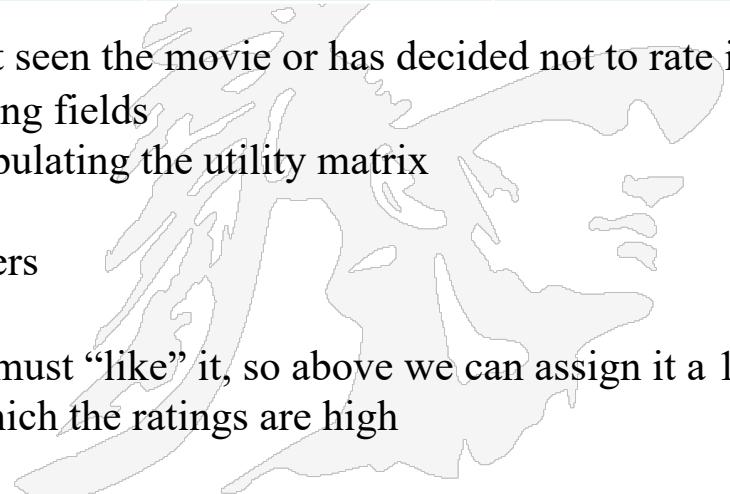
View all other entries as zeros for now.

- To compute the similarity between individual's preference vectors we can use inner products as a good place to start, e.g.
  - Dave has similarity 3 with Estie,
    - e.g.  $(-1,0,0,-1,1,1,0,0,1)$  and  $(0,0,0,-1,1,1,0,1,0)$
    - (i.e. there are three matching values of either 1 or -1)
  - but -2 with Cindy  $(0,0,0,1,-1,0,0,-1,0)$  (a zero value doesn't count).
- Perhaps recommend Straits Cafe to Dave and Il Fornaio to Bob, etc.

# Another Utility Matrix Example - Movies

	Avatar	LOTR	MATRIX	PIRATES
ALICE	1		0.2	
BOB		0.5		0.3
CAROL	0.2		1	
DAVID				0.4

- The blank spaces indicate either the user has not seen the movie or has decided not to rate it
- Main issue: how to fill in the values in the missing fields
- In general there are two basic techniques for populating the utility matrix
  - Ask users to rate items  
E.g. movies, online stores from purchasers
  - Make inferences from user behaviors  
Assumption: Users who watch a movie must “like” it, so above we can assign it a 1;  
We are mostly interested in fields for which the ratings are high



## Recommending documents can be Viewed as a Form of Recommendation System

- If the items we are considering are documents, then the profile will be the set of “important” words in the document
- How do we pick important words
  - We use the TF-IDF formulation seen earlier

Profile of a document  $d_j$  is the vector of weights  $w_{i,j}$   $\underline{Content(d_j)} = (w_{1,j}, \dots, w_{k,j})$ .

$$w_{i,j} = TF_{i,j} \times IDF_i \quad TF_{i,j} = \frac{f_{i,j}}{\max_z f_{z,j}}, \quad IDF_i = \log \frac{N}{n_i}.$$

- **TF : Term Frequency, IDF : Inverse Document Frequency**
- **N : Number of the documents**
- **$n_i$  : How many times an element is seen in all of the documents**
- **$f_{i,j}$  : Number of times an element is seen in the document  $d_j$**

# Example 1

## Boolean Utility Matrix

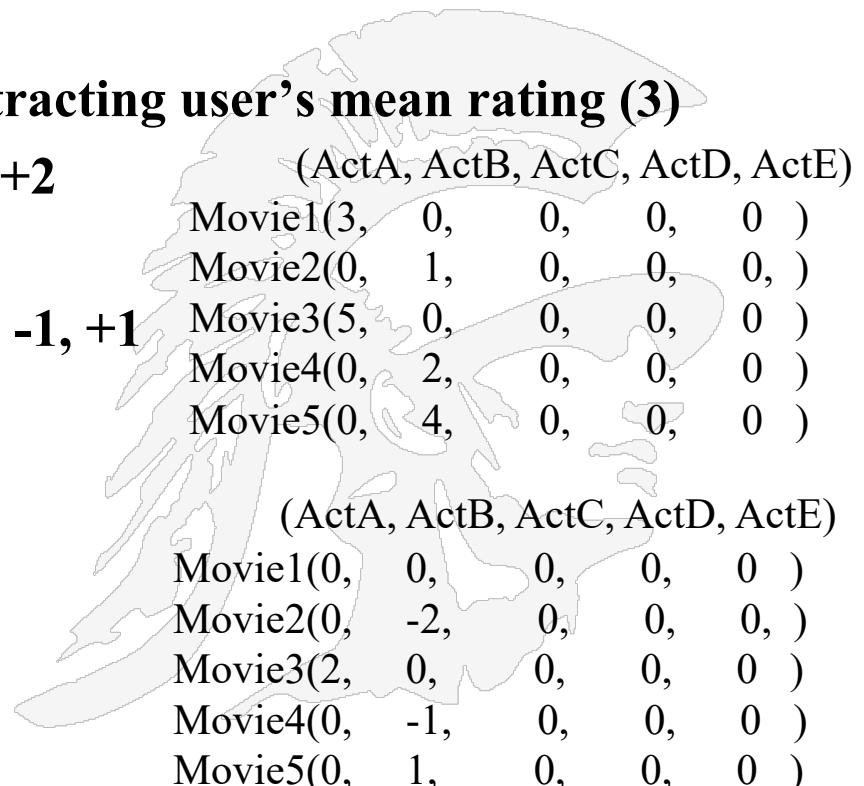
- Items are movies, only feature is Actor
  - Item profile: vector with  $0$  or  $1$  for each actor
- Suppose user  $X$  has watched 5 movies
  - 2 movies featuring actor A (movies 1 and 3)
  - 3 movies featuring actor B (movies 2, 4, and 5)
- User profile = mean of item profiles
  - Feature A's weight =  $2/5 = 0.4$
  - Feature B's weight =  $3/5 = 0.6$

	(ActA, ActB, ActC, ActD, ActE)
Movie1(1,	0, 0, 0, 0, 0 )
Movie2(0,	1, 0, 0, 0, 0 )
Movie3(1,	0, 0, 0, 0, 0 )
Movie4(0,	1, 0, 0, 0, 0 )
Movie5(0,	1, 0, 0, 0, 0 )

$$\text{ActA's weight} = \text{Sum(ActA)}/5$$

## Example 2 Star Ratings

- Same example, 1-5 star ratings
  - Actor A's movies rated 3 and 5
  - Actor B's movies rated 1, 2, and 4 (note: 1 and 2 are negative ratings)
- Useful step: normalize ratings by subtracting user's mean rating (3)
  - Actor A's normalized ratings = 0, +2
    - Profile weight =  $(0 + 2)/2 = 1$
  - Actor B's normalized ratings = -2, -1, +1
    - Profile weight =  $-2/3$



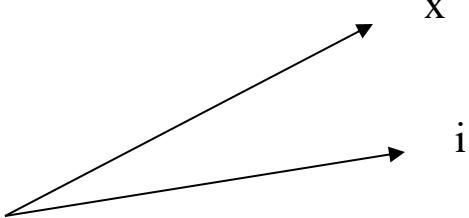
	ActA	ActB	ActC	ActD	ActE
Movie1	(3,	0,	0,	0,	0 )
Movie2	(0,	1,	0,	0,	0 , )
Movie3	(5,	0,	0,	0,	0 )
Movie4	(0,	2,	0,	0,	0 )
Movie5	(0,	4,	0,	0,	0 )

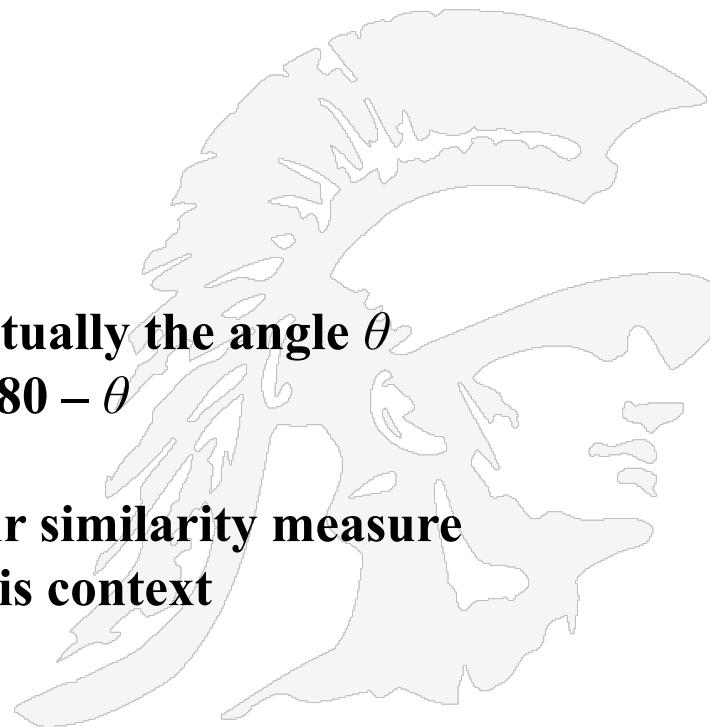
	ActA	ActB	ActC	ActD	ActE
Movie1	(0,	0,	0,	0,	0 )
Movie2	(0,	-2,	0,	0,	0 , )
Movie3	(2,	0,	0,	0,	0 )
Movie4	(0,	-1,	0,	0,	0 )
Movie5	(0,	1,	0,	0,	0 )

# Making Predictions

- Given user profile  $x$  (movies he/she watched) and item profile  $i$  (*movies with actor profiles*)
- Estimate the similarity of  $U(x,i) = \cos(\theta) = (x \cdot i) / (|x| |i|)$

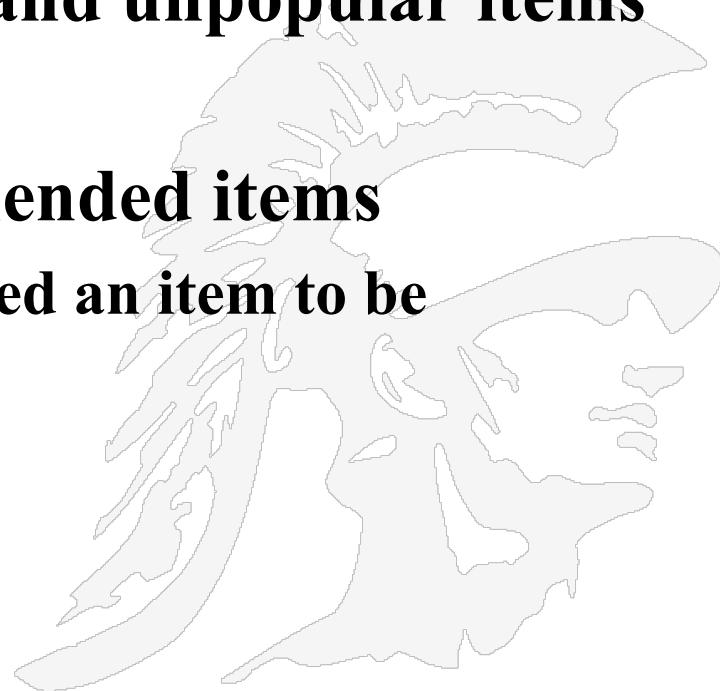


- Technically the cosine distance is actually the angle  $\theta$  and the cosine similarity is the angle  $180 - \theta$
- For convenience we use  $\cos(\theta)$  as our similarity measure and call it the “cosine similarity” in this context



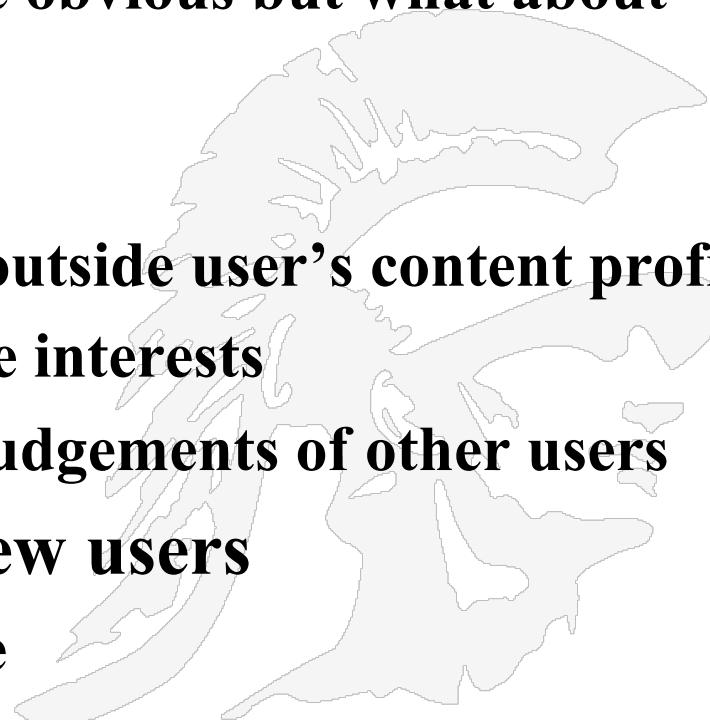
## Pros: Content-based Approach

- No need for data on other users
- Able to recommend to users with unique tastes
- Able to recommend new and unpopular items
  - No first rater problem
- Explanations for recommended items
  - Content features that caused an item to be recommended



## Cons: Content-Based Approach

- **Finding the appropriate features is not always obvious**
  - E.g. movie features may be obvious but what about images and music
- **Overspecialization**
  - Never recommends items outside user's content profile
  - People might have multiple interests
  - Unable to exploit quality judgements of other users
- **Cold-start problem for new users**
  - How to build a user profile

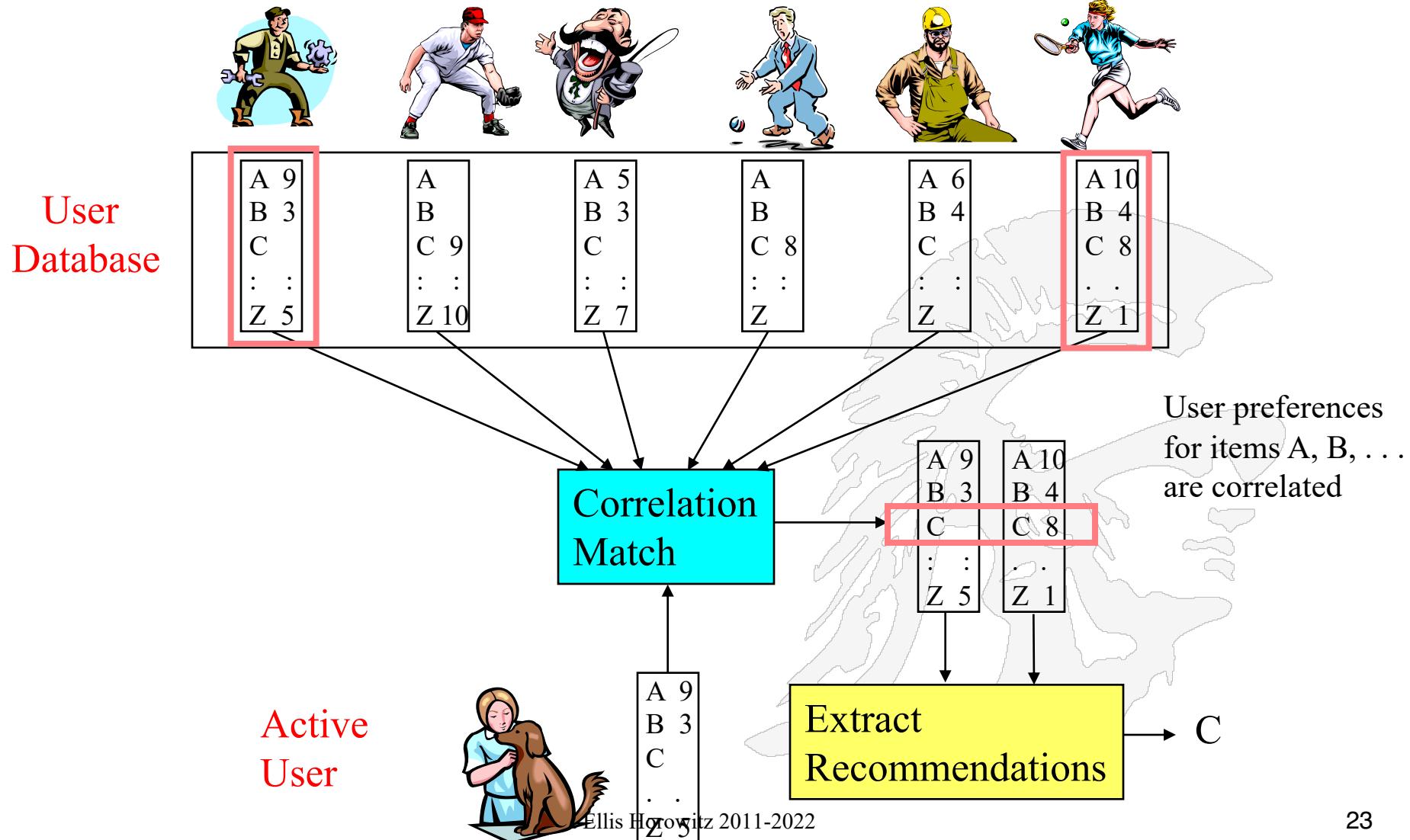


# Let's Switch Focus to Collaborative Filtering

- Collaborative filtering is a method of making automatic predictions (filtering) about the interests of a user by collecting preferences or taste information from many other users (collaborating).
  - The underlying assumption of the collaborative filtering approach is that if a person *A* has the same opinion as a person *B* on an issue, *A* is more likely to have *B*'s opinion on a different issue than that of a randomly chosen person



# Collaborative Filtering



## Similar Users and Jaccard Similarity

movies

	HP1	HP2	HP3	TW	SW1	SW2	SW3	
A	4			5	1			HP:Harry Potter
B	5	5	4					SW:Star Wars
C				2	4	5		TW:Twilight
D		3					3	

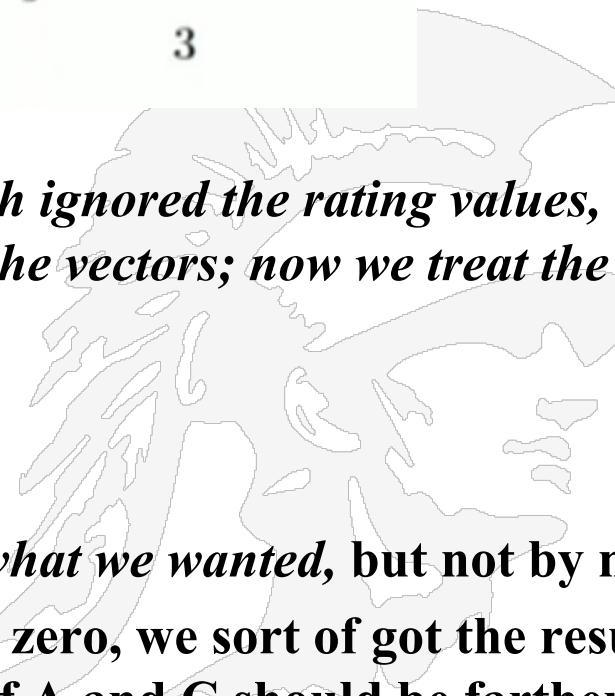
4 users

- Consider users  $x$  and  $y$  with rating vectors  $rx$  and  $ry$ 
  - The rating vector of user  $B$  is  $(5,5,4,0,0,0,0)$
- We need a similarity metric  $sim(x,y)$  between rating vectors
- The metric should capture the intuition that  $sim(A,B) > sim(A,C)$ 
  - $A$  and  $B$  both liked HP1, but  $A$  and  $C$  had very different opinions about TW and SW1
- Recall  $sim(A,B) = |r_a \text{ intersect } r_b| / |r_a \text{ union } r_b|$  try Jaccard Similarity
- $Sim (A,B) = 1/5$ ;  $sim(A,C) = 2/4$ . since A&B rated only one movie in common
  - But using Jaccard Similarity we get a result we don't want, namely
  - $Sim(A,B) < Sim (A,C)$
- Problem: ignores rating values

# Cosine Similarity

	HP1	HP2	HP3	TW	SW1	SW2	SW3
A	4			5	1		
B	5	5	4				
C				2	4	5	
D		3					3

- Instead of using Jaccard similarity, which ignored the rating values, let's use cosine similarity, the angle between the vectors; now we treat the unknown values as zero
  - $\text{sim}(A, B) = \cos(r_a, r_b)$
- $\text{sim}(A, B) = 0.38$ ,  $\text{sim}(A, C) = 0.32$ 
  - Now  $\text{sim}(A, B) > \text{sim}(A, C)$ , which is what we wanted, but not by much
- Problem: by treating missing ratings as zero, we sort of got the result we wanted, but actually the similarity of A and C should be farther apart than what we computed; using zero was not a great idea



# Centered Cosine Captures User Preferences

Solution: Normalize ratings by subtracting row mean

	HP1	HP2	HP3	TW	SW1	SW2	SW3
A	4			5	1		
B	5	5	4				
C				2	4	5	
D		3					3

	HP1	HP2	HP3	TW	SW1	SW2	SW3
A	2/3			5/3	-7/3		
B	1/3	1/3	-2/3				
C				-5/3	1/3	4/3	
D		0					0

	HP1	HP2	HP3	TW	SW1	SW2	SW3
A	2/3			5/3	-7/3		
B	1/3	1/3	-2/3				
C				-5/3	1/3	4/3	
D		0					0

- The average rating for A is  $10/3$ ;
- The average rating for B is  $14/3$ ;
- The solution is to subtract the average rating for each user's score; e.g. user A and HP1 we get  $4 - (10/3) = 2/3$

The resulting matrix

$$\sim sim(A,B) = \cos(r_a, r_b) = 0.09; \\ sim(A,C) = -0.56$$

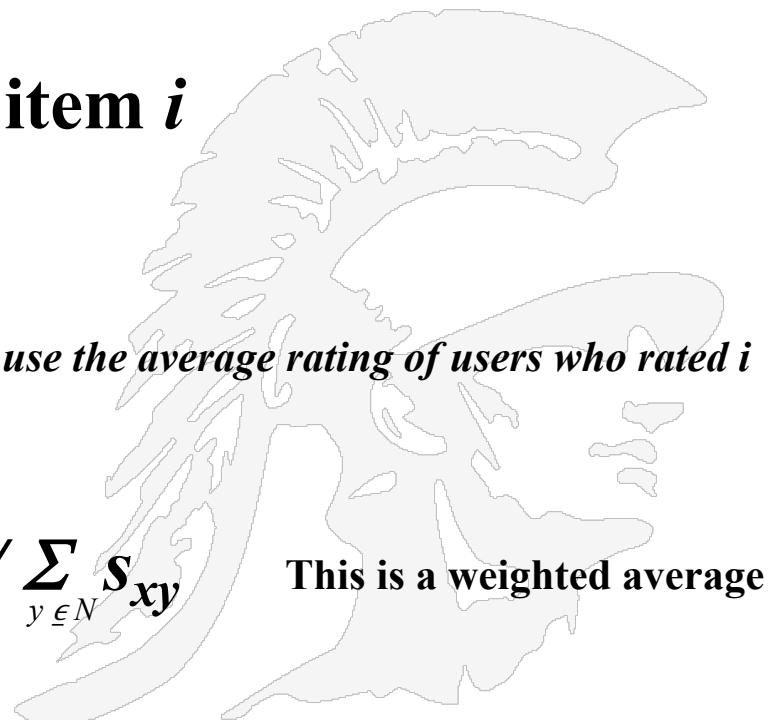
Result: A and C are very DISsimilar

- $sim(A,B) > sim(A,C)$
- Captures intuition better
  - Missing ratings treated as average
  - Handles "tough raters" and "easy raters"

Note: Summing the rows for any user gives zero, so positive ratings means they liked the movie  
 Another name for centered cosine is Pearson Correlation

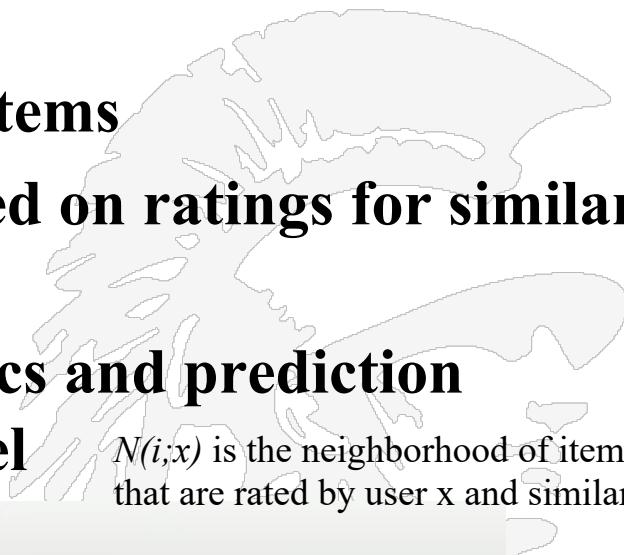
## Making Rating Predictions for a User

- Let  $r_x$  be the vector of user  $x$ 's ratings
  - Let  $N$  be the set of  $K$  users most similar to  $x$  who have also rated item  $i$
  - Prediction for user  $x$  and item  $i$
- 
- Option 1:  $r_{xi} = 1/k \sum_{y \in N} (r_{yi})$  *use the average rating of users who rated i*
  - Option 2:  $r_{xi} = \sum_{y \in N} (s_{xy} r_{yi}) / \sum_{y \in N} s_{xy}$   
where  $s_{xy} = sim(x,y)$



# Item-Item Collaborative Filtering

- So far: we have used user-user collaborative filtering
- Another view: item-item
  - For item  $i$ , find other similar items
  - Estimate rating for item  $i$  based on ratings for similar items
  - Can use same similarity metrics and prediction functions as in user-user model



$N(i;x)$  is the neighborhood of items that are rated by user  $x$  and similar to item  $i$

$$r_{xi} = \frac{\sum_{j \in N(i;x)} s_{ij} \cdot r_{xj}}{\sum_{j \in N(i;x)} s_{ij}}$$

$s_{ij}$ ... similarity of items  $i$  and  $j$

$r_{xj}$ ... rating of user  $x$  on item  $j$

$N(i;x)$ ... set items rated by  $x$  similar to  $i$

# Let's Do An Example

## Item – Item CF ( $|N| = 2$ )

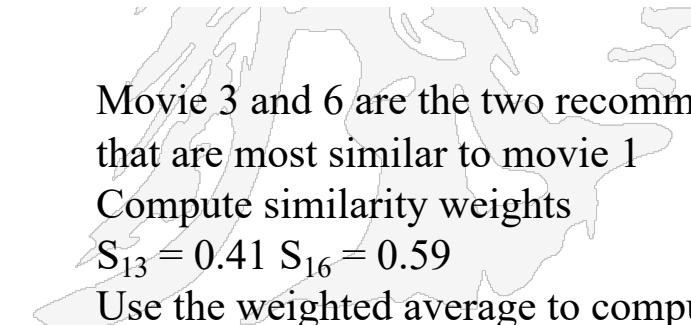
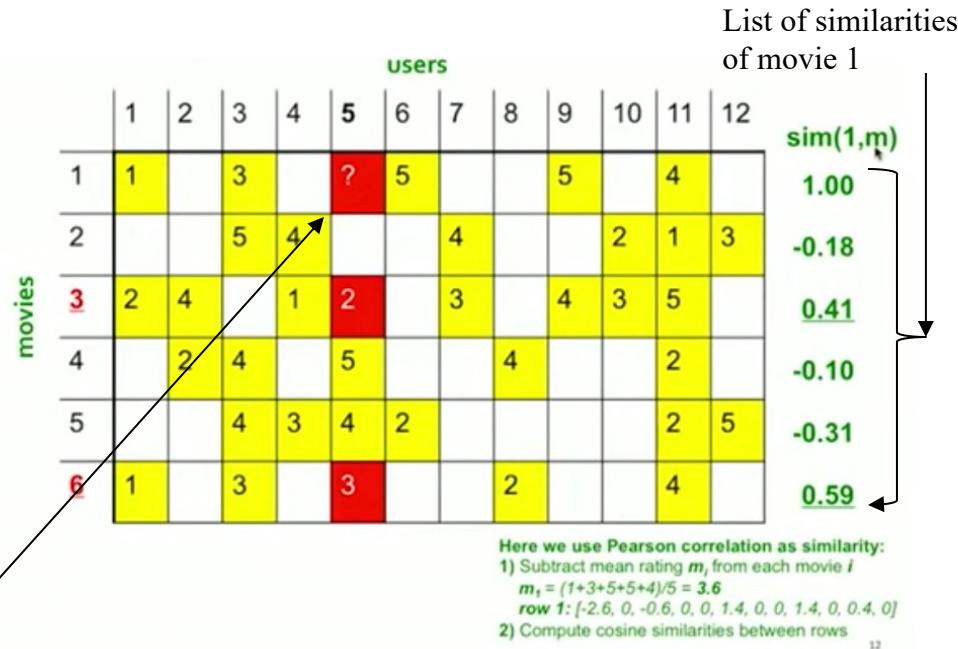
	users											
	1	2	3	4	5	6	7	8	9	10	11	12
movies	1	1		3			5			5		4
2			5	4	*		4			2	1	3
3	2	4		1	2		3		4	3	5	
4		2	4		5			4			2	
5			4	3	4	2				2	5	
6	1			3				2			4	

  - unknown rating        - rating between 1 to 5

Goal: Estimate rating of movie 1 by user 5

$N$  is 2, looking at the two nearest neighbors

Conclusion: user 5 will like movie 1: 2.6



Movie 3 and 6 are the two recommendations that are most similar to movie 1

Compute similarity weights

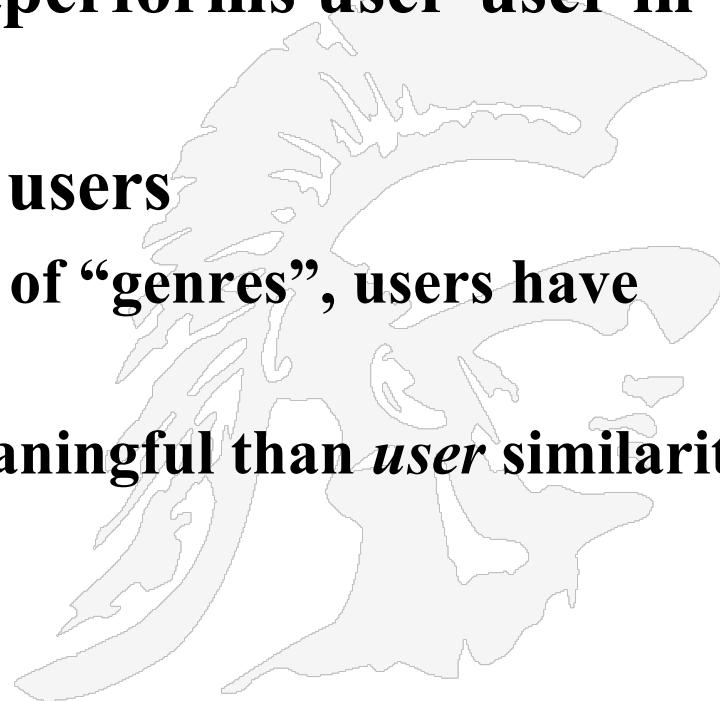
$$S_{13} = 0.41 \quad S_{16} = 0.59$$

Use the weighted average to compute

$$R_{15} = (0.41*2 + 0.59*3)/(0.41+0.59) = 2.6$$

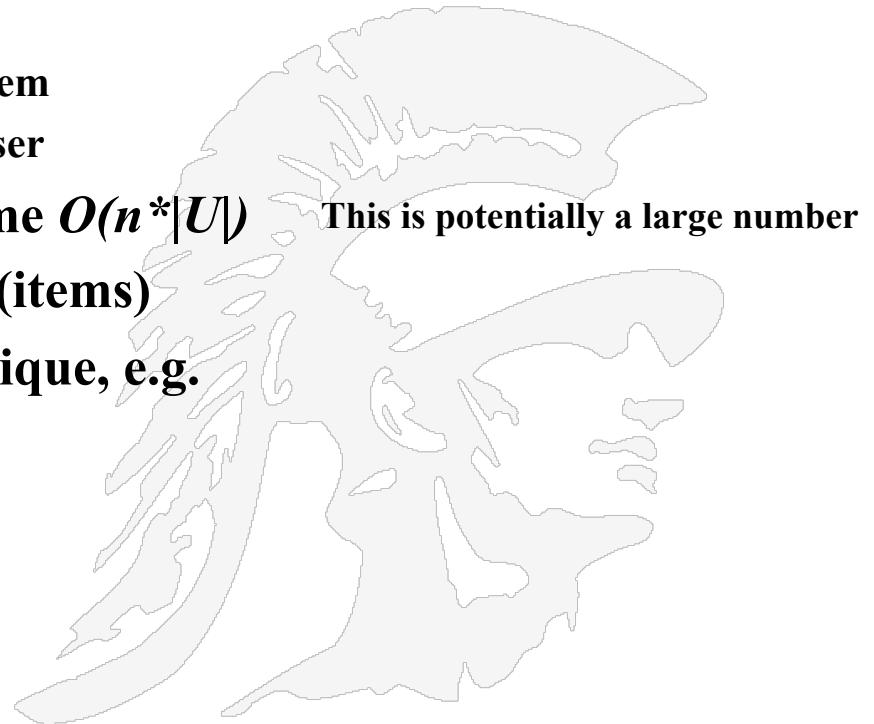
# Item-Item vs. User-User

- In theory user-user and item-item are dual approaches
- In practice, item-item outperforms user-user in many use cases
- Items are “simpler” than users
  - Items belong to a small set of “genres”, users have varied tastes
  - *Item* similarity is more meaningful than *user* similarity



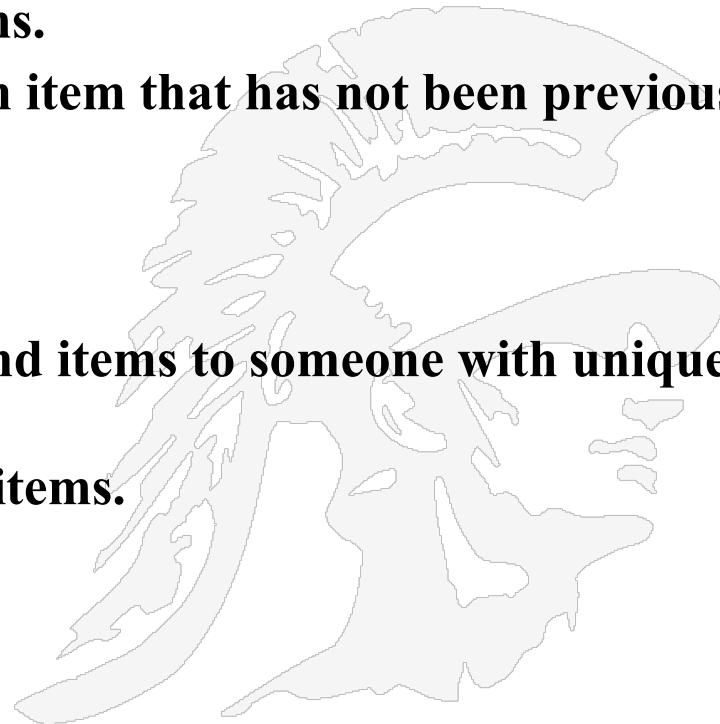
# Collaborative filtering: Algorithm Complexity

- Expensive step is finding  $k$  most similar users (or items):  $O(|U|)$ 
  - $|U|$  = size of utility matrix
- Too expensive to do at runtime
  - Could pre-compute
    - The set of similar items for every item
    - The set of similar users for every user
  - Naïve pre-computation takes time  $O(n^*|U|)$  This is potentially a large number
    - Where  $n$  = number of users (items)
  - But we can use a previous technique, e.g.
    - Clustering



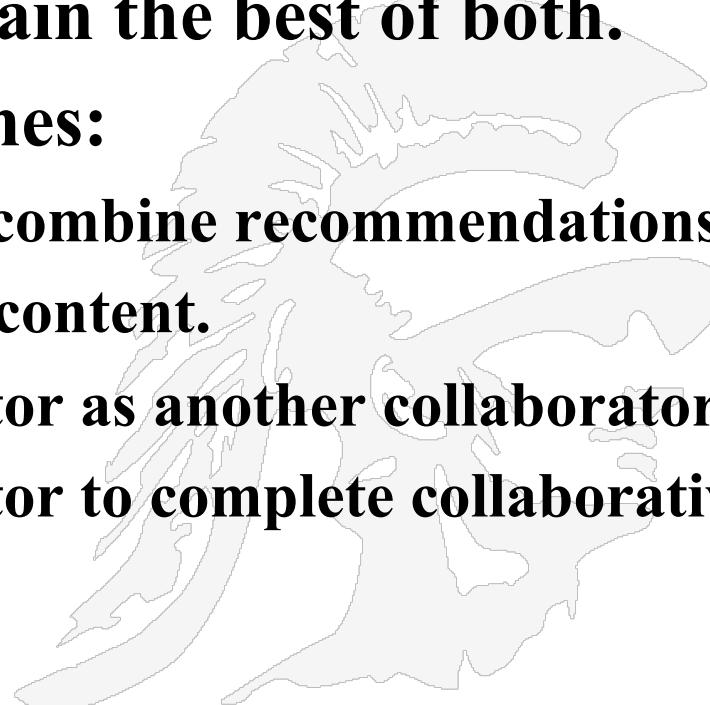
# Problems with Collaborative Filtering

- **Cold Start:** There needs to be enough other users already in the system to find a match.
- **Sparsity:** If there are many items to be recommended, even if there are many users, the user/ratings matrix is sparse, and it is hard to find users that have rated the same items.
- **First Rater:** Cannot recommend an item that has not been previously rated.
  - New items
  - Esoteric items
- **Popularity Bias:** Cannot recommend items to someone with unique tastes.
  - Tends to recommend popular items.



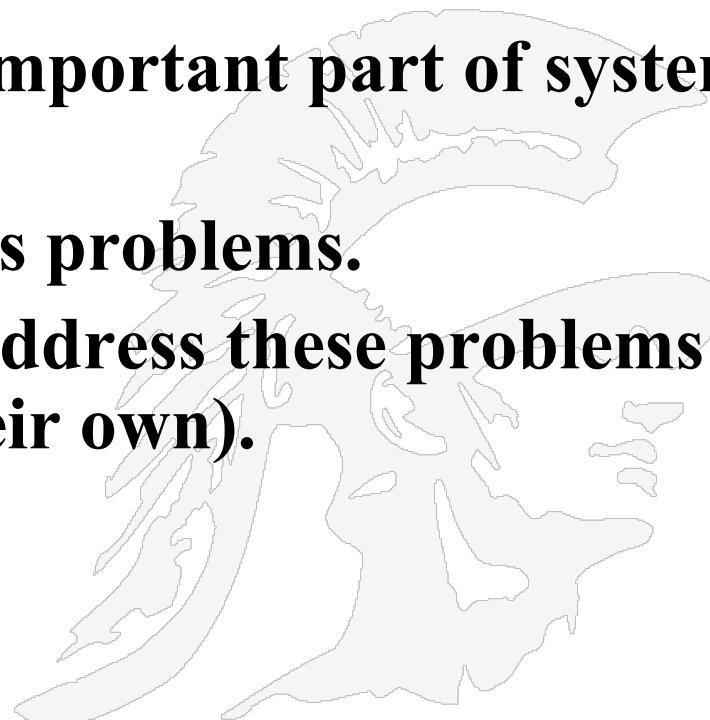
# Combining Content and Collaboration

- **Content-based and collaborative methods have complementary strengths and weaknesses.**
- **Combine methods to obtain the best of both.**
- **Various hybrid approaches:**
  - Apply both methods and combine recommendations.
  - Use collaborative data as content.
  - Use content-based predictor as another collaborator.
  - Use content-based predictor to complete collaborative data.



# Conclusions

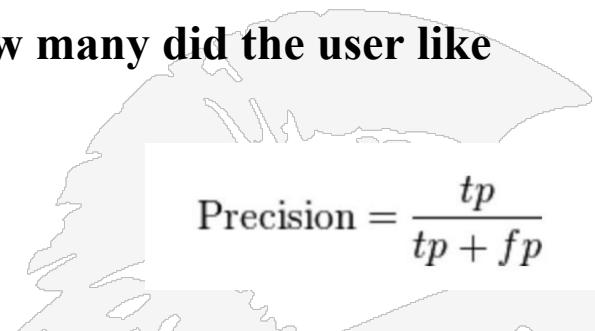
- Recomending and personalization are important approaches to combating information overload.
- Machine Learning is an important part of systems for these tasks.
- Collaborative filtering has problems.
- Content-based methods address these problems (but have problems of their own).
- Integrating both is best.



# Evaluation Metrics for Recommendation Engines

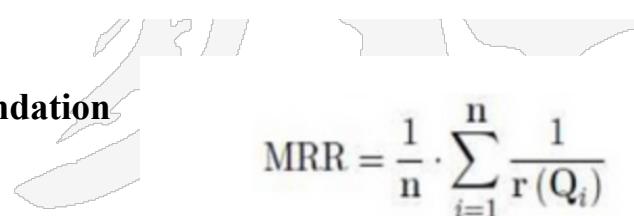
- ***Recall*** – the proportion of items that a user like that were recommended
  - $t_p$  = number of recommended items
  - $f_n$  = the remaining items
- ***Precision*** – out of all recommended items, how many did the user like
  - $t_p$  number of recommended items
  - $t_p + f_p$  the total items recommended
- ***Root Mean Squared Error (RMSE)***
  - Measures error in the predicted rating
- ***Mean Reciprocal Rank***
  - The larger the MRR, the better the recommendation

$$\text{Recall} = \frac{tp}{tp + fn}$$



$$\text{Precision} = \frac{tp}{tp + fp}$$

$$RMSE = \sqrt{\frac{\sum_{i=1}^N (Predicted_i - Actual_i)^2}{N}}$$



$$MRR = \frac{1}{n} \cdot \sum_{i=1}^n \frac{1}{r(Q_i)}$$

# Web Search Engines Image Search



Some slides come from ImageNet:A large-scale hierarchical image database  
By Deng, Dong, Socher, Li-Jia Li, Kai Li and Li Fei-Fei

# Search Engines Expand their Search to Images and Videos

- **Image Search**
  - Google and Bing and others
  - Search primarily based on
    - tags (FlickR, FaceBook)
    - surrounding text
    - using image features
- **Video Search Presenters**
  - Google and YouTube are the leaders
  - other video search engines include:
    - Bing videos <http://www.bing.com/videos/>
    - AOL video, <https://www.aol.com/video/>
    - eHow, <https://www.ehow.com/>
    - MeFeedia, iPhone/Android app



MediaMagic Video Production

# Image and Video Searching

## There is a Lot of It on the Web

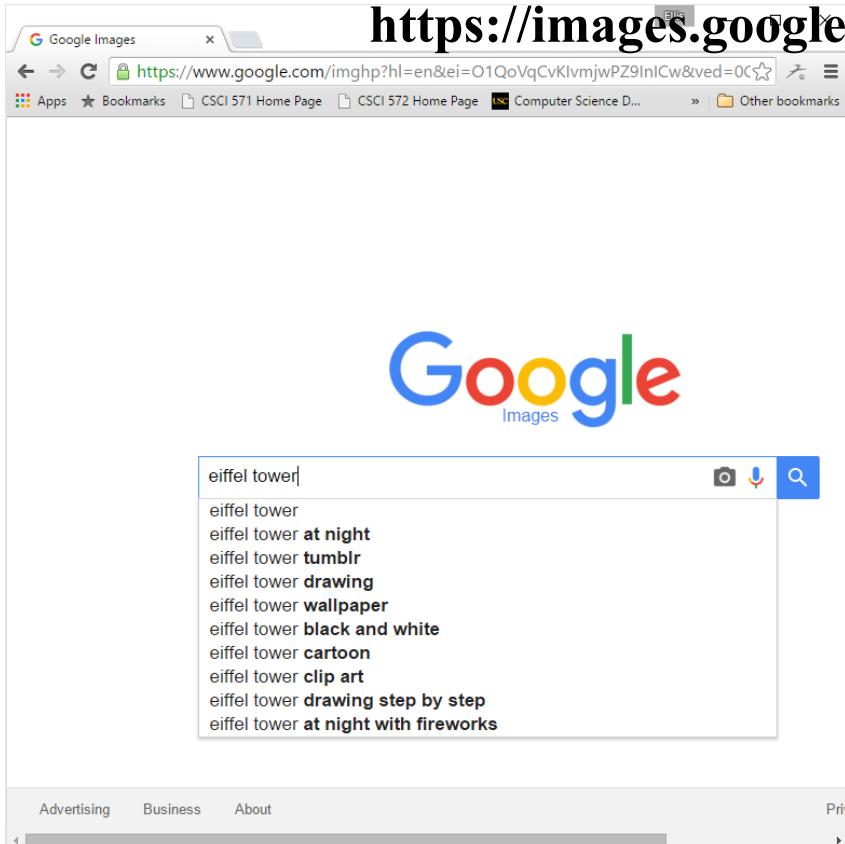
- “Images are returned for 19% of search queries on Google. There are over 600 million visual searches on Pinterest each month. Image-based Pinterest Ads have an 8.5% conversion rate, and Pinterest is projected to clear \$1 billion a year in ad revenue by 2020”
  - <https://www.socialmediatoday.com/news/why-visual-search-will-be-one-of-the-biggest-digital-marketing-trends-of-20/545999/>

### Top 100 most searched keywords in the world

#	Keyword	Total Searches	Organic Traffic	Paid Traffic	Biggest Winner
1	youtube	2.945B	2.899B	45.79M	<a href="https://www.youtube.com">youtube.com</a>
2	facebook	1.904B	1.844B	60.01M	<a href="https://www.facebook.com">facebook.com</a>
3	whatsapp web	1.850B	1.850B	94.22K	<a href="https://www.whatsapp.com">whatsapp.com</a>
4	netflix	983.08M	983.04M	34.60K	<a href="https://www.netflix.com">netflix.com</a>
5	roblox	767.70M	759.14M	8.56M	<a href="https://www.roblox.com">roblox.com</a>
6	amazon	304.23M	224.72M	79.51M	<a href="https://www.amazon.com">amazon.com</a>

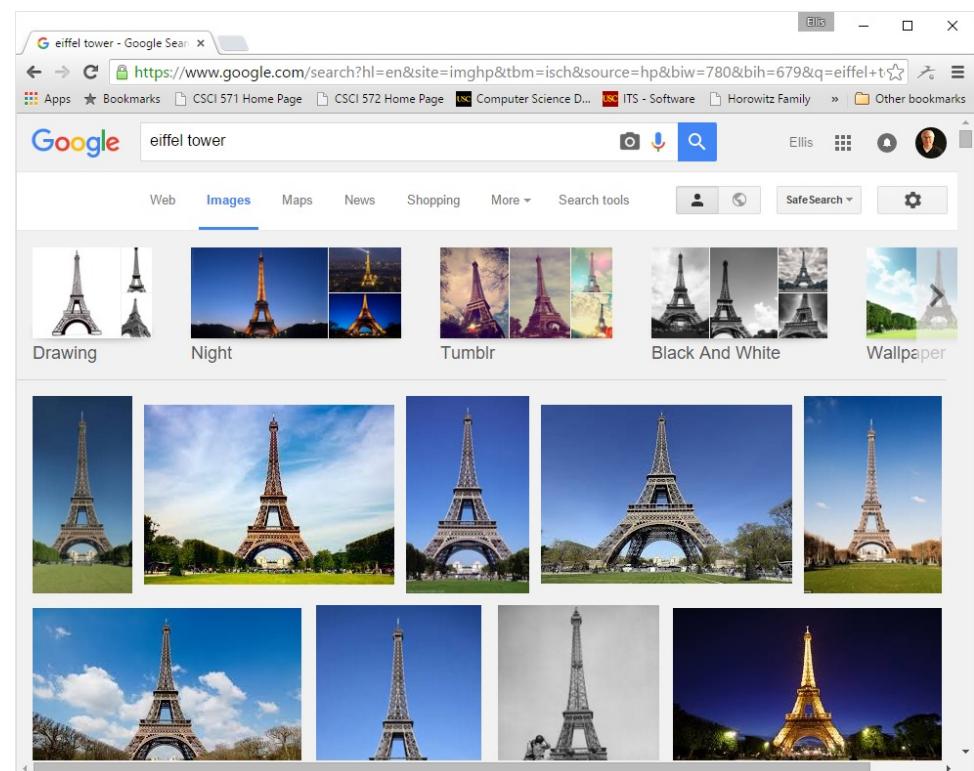
# Using Google Image Search

Google has a distinct page for image searching  
<https://images.google.com/>



A screenshot of a web browser showing the Google Images homepage. The search bar at the top contains the query "eiffel tower". Below the search bar, a dropdown menu shows autocomplete suggestions: "eiffel tower", "eiffel tower at night", "eiffel tower tumblr", "eiffel tower drawing", "eiffel tower wallpaper", "eiffel tower black and white", "eiffel tower cartoon", "eiffel tower clip art", "eiffel tower drawing step by step", and "eiffel tower at night with fireworks". At the bottom of the page, there are links for "Advertising", "Business", and "About".

Query plus autocomplete options



A screenshot of a web browser showing Google search results for the query "eiffel tower". The search bar at the top contains "eiffel tower". The results are displayed under the "Images" tab. The first row shows five image thumbnails: "Drawing", "Night", "Tumblr", "Black And White", and "Wallpaper". Below this, there are two rows of five larger image thumbnails each, showing various photographs of the Eiffel Tower from different angles and times of day.

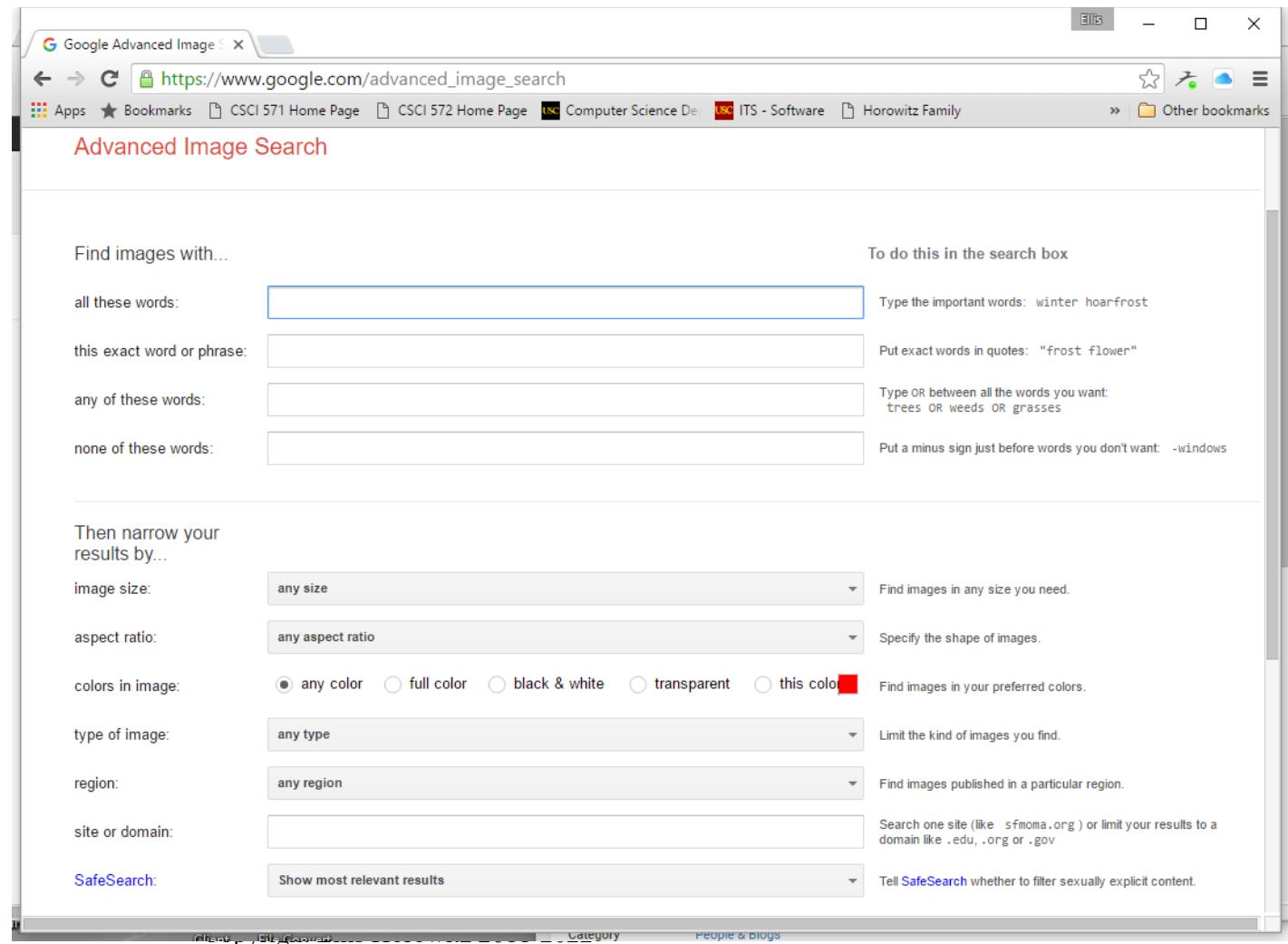
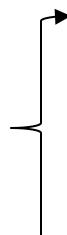
Search results

# Google Advanced Image Search

text-based  
query criteria



image-based  
query criteria



Find images with...

all these words:

this exact word or phrase:

any of these words:

none of these words:

To do this in the search box

Type the important words: winter hoarfrost

Put exact words in quotes: "frost flower"

Type OR between all the words you want: trees OR weeds OR grasses

Put a minus sign just before words you don't want: -windows

Then narrow your results by...

image size:

aspect ratio:

colors in image:

type of image:

region:

site or domain:

SafeSearch:

Find images in any size you need.

Specify the shape of images.

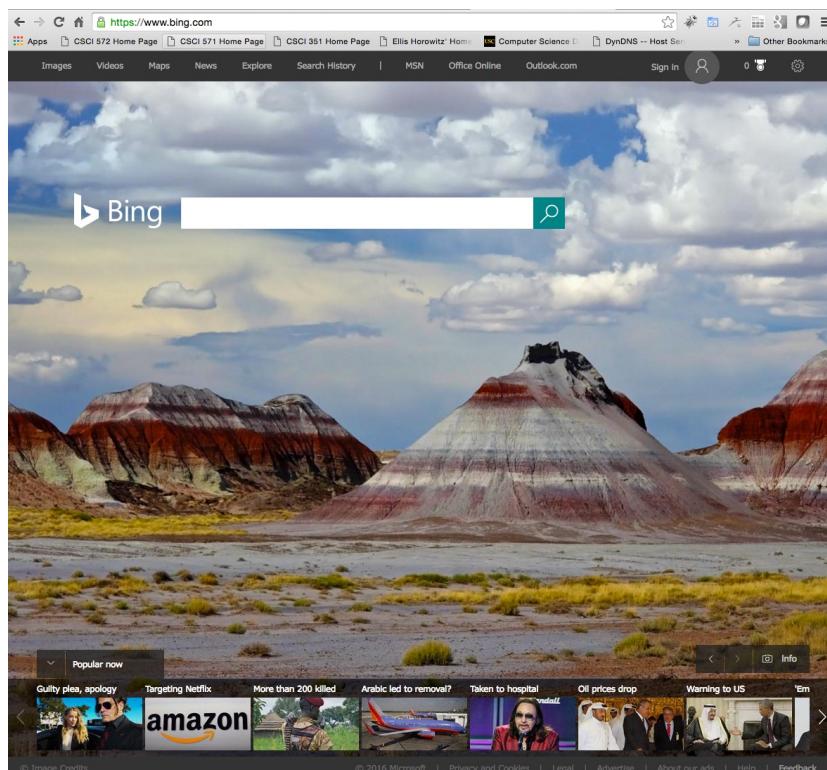
Find images in your preferred colors.

Limit the kind of images you find.

Find images published in a particular region.

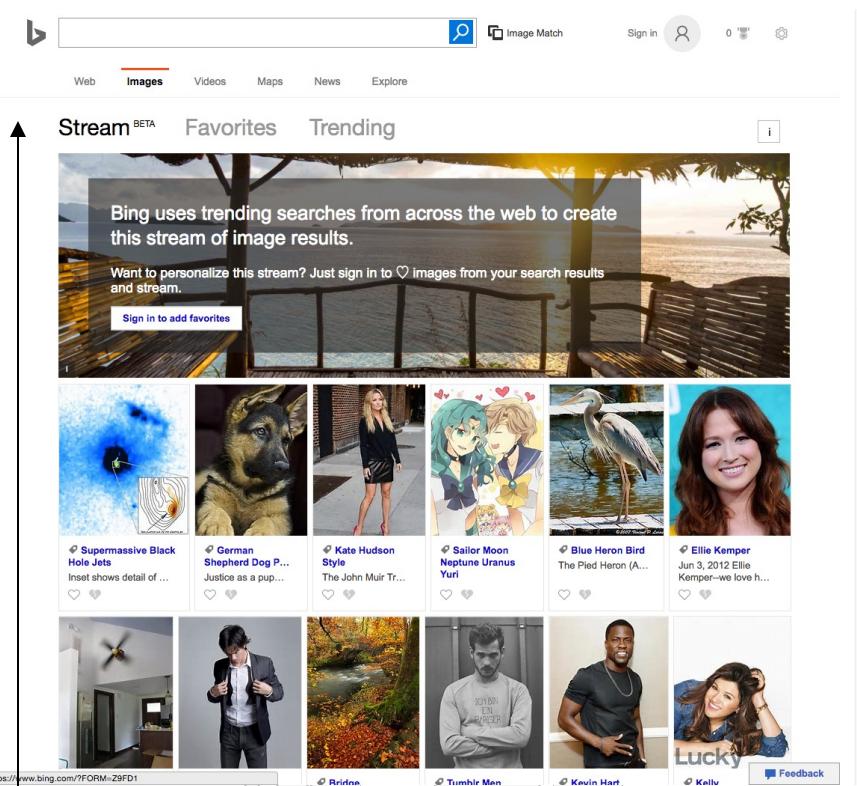
Search one site (like sfmoma.org) or limit your results to a domain like .edu, .org or .gov

Tell SafeSearch whether to filter sexually explicit content.



Bing Initial Page with Options:  
Images, Videos, Maps, News,  
Explore, Search History

# Bing Image Search

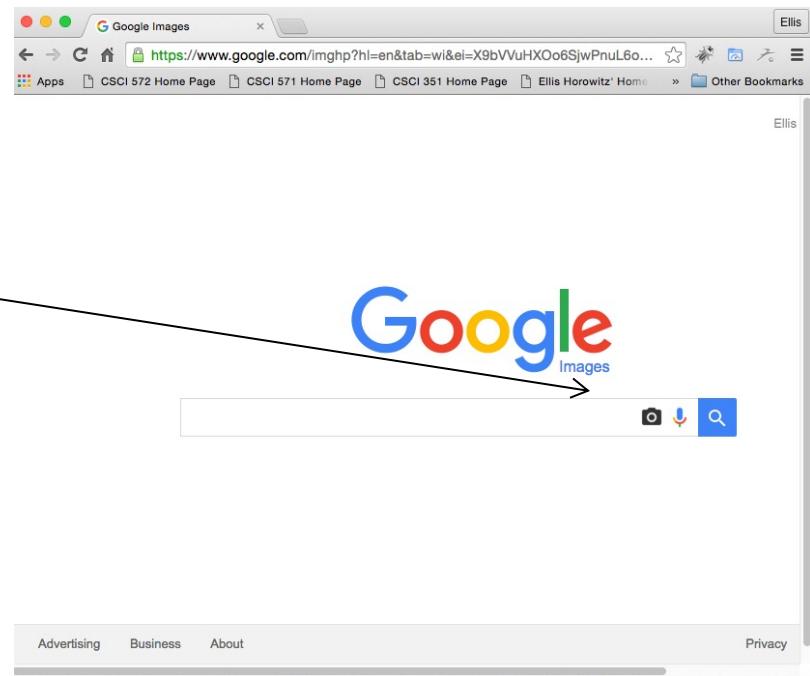


A screenshot of the Bing Images search results page. The top navigation bar shows "Web" and "Images" (which is highlighted). Below the navigation are tabs for "Stream BETA", "Favorites", and "Trending". The main content area displays a grid of image thumbnails. One thumbnail is highlighted with a yellow box and text: "Bing uses trending searches from across the web to create this stream of image results. Want to personalize this stream? Just sign in to ❤ images from your search results and stream." Below this, there are several other image thumbnails with captions and user interaction options like hearts and comments.

Bing Initial Images Page;  
Options are: Stream,  
Favorites, Trending

# Using Google Image Search by Similarity

- **Search By Image** is available now at [images.google.com](https://images.google.com) or via the “Images” tab in the right-side menu on Google.com. You should see a small camera icon on the far right side of the search bar.
- There are several ways to access it:
  1. Drag and drop an image on the search bar
  2. Click the camera icon to upload an image from your computer
  3. Paste the URL of a photo on the web into the search bar
  4. Use the Chrome or Firefox extensions that add a search option to your computer’s contextual menu (right click)



# Google Image Similarity Match Example

- Google will try to match an image that you drag into the search box
- The original image
- The search results

Google Michelle.jpg toque

About 6 results (1.55 seconds)

 Image size:  
1296 × 1590  
No other sizes of this image found.

Possible related search: [toque](#)

<https://en.wikipedia.org/wiki/Toque>  
**Toque - Wikipedia**  
A toque blanche (French for 'white hat'), often shortened to **toque**, is a tall, round, pleated, starched white hat worn by chefs. ... The **toque** most likely ...

<https://www.merriam-webster.com/dictionary/toque>  
**Toque Definition & Meaning - Merriam-Webster**  
Definition of **toque** · 1 : a woman's small hat without a brim made in any of various soft close-fitting shapes · 2 : **tuque** · 3 : a tall brimless hat worn by a chef.

 Visually similar images



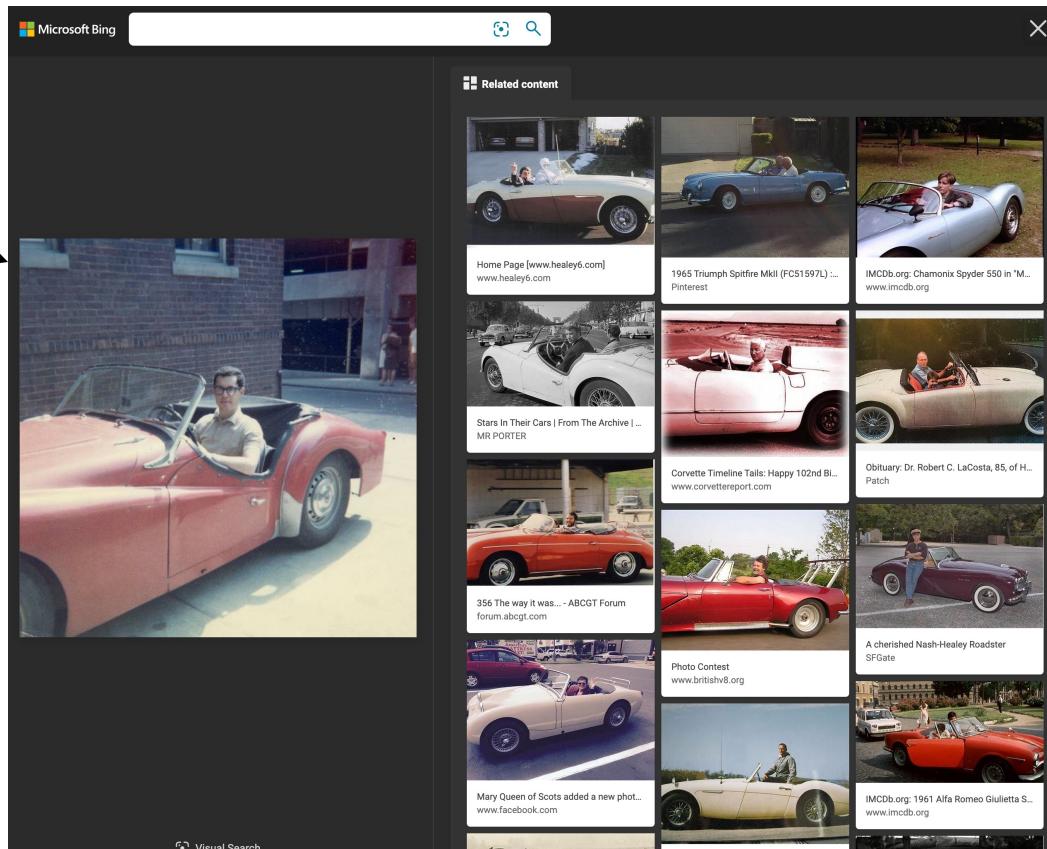
**Toque**  
Cap

A toque is a type of hat with a narrow brim or no brim at all. Toques were popular from the 13th to the 16th century in Europe, especially France. The mode was revived in the 1930s. [Wikipedia](#)

# Bing Image Similarity Match Example

Original photo

Matching photos



Some Useful Links Describing Bing Image Search

<https://blogs.bing.com/search-quality-insights/May-2018/Internet-Scale-Deep-Learning-for-Bing-Image-Search>

<http://searchengineland.com/bing-image-search-redesigned-to-add-more-image-details-to-the-results-218276>

# List of Image Search Engines

- **TinEye** is a reverse image search engine
- It primarily uses image identification rather than keywords, metadata
- Upon submitting an image, TinEye creates a "unique and compact digital signature or fingerprint" of the image and matches it with other indexed images
- This procedure is able to match even heavily edited versions of the submitted image, but will not usually return similar images in the results.

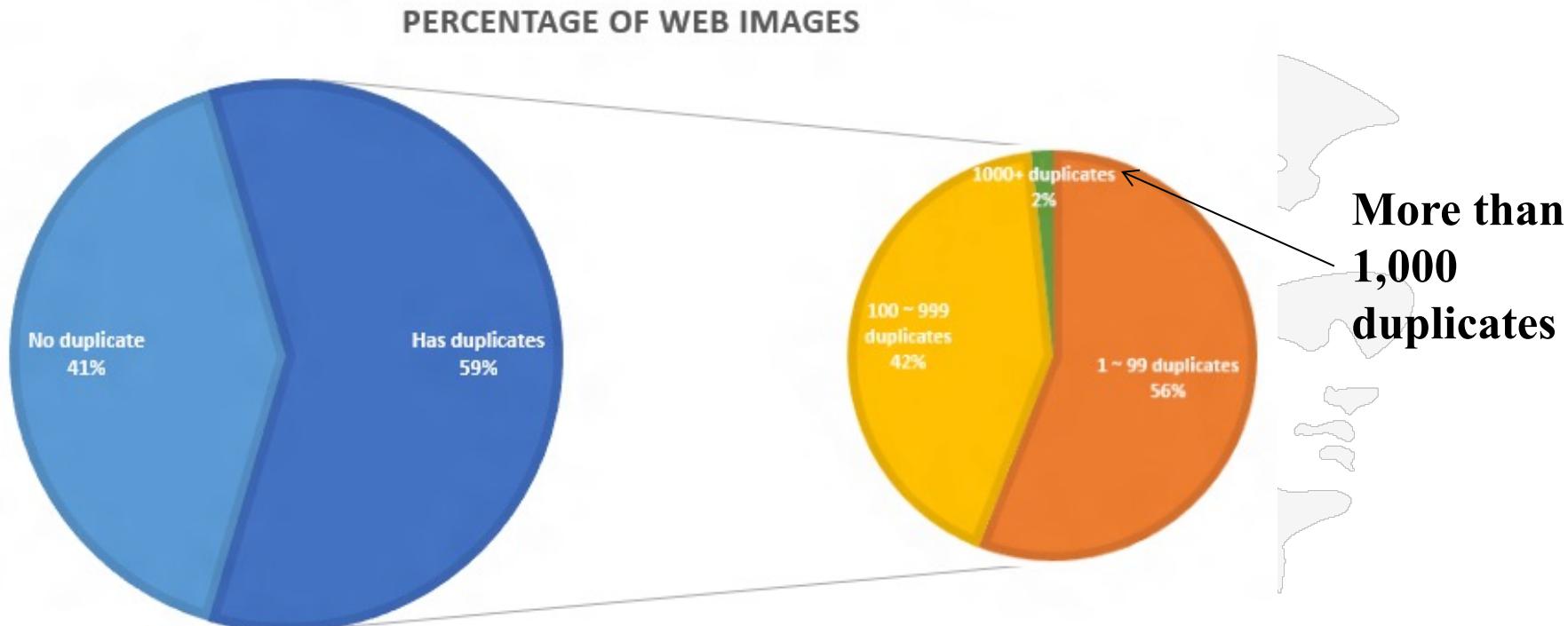
## List of Image Search Engine

	Reverse Search	Price
Bing	No	Free
<a href="#">Yahoo</a>		
<a href="#">Altavista</a>		
(All 3 uses Bing algorithm)		
Ask.com	No	Free
(Uses Google algorithm)		
Google Image	Yes	Free
	<a href="#">Drag and Drop Tutorial</a>	
Corbis	No	Buy sell Images, some are royalty free
<a href="#">Imnese</a>	No	Free
<a href="#">TinEye</a>	Yes	Free
<a href="#">GazoPa</a>	No	For Business
<a href="#">PicSearch</a>	No	Free

<https://www.searchenginejournal.com/best-image-search-engines/299963/#close>

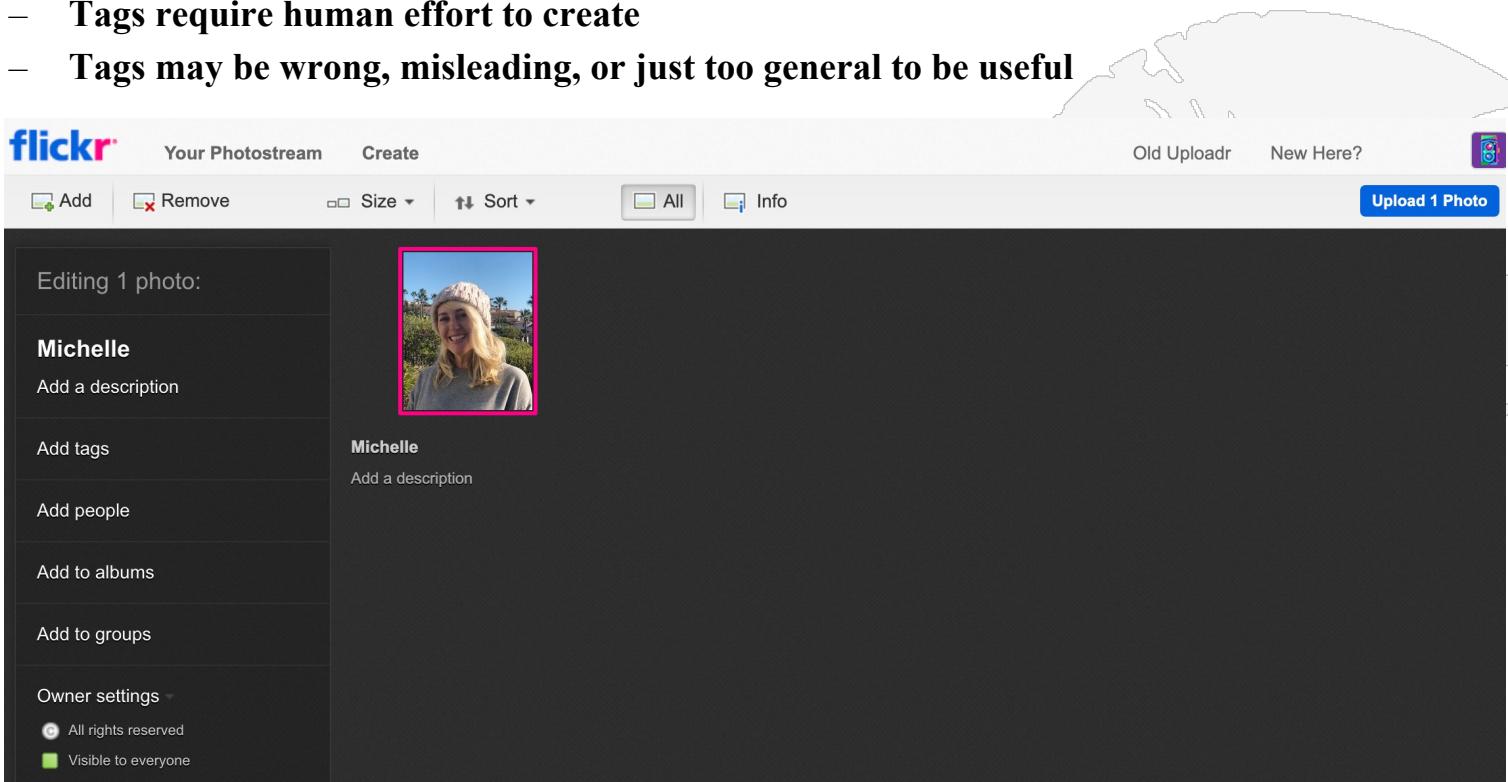
# Bing on Image Duplication

- Bing Data on Image Duplication across the Web-there is a lot of it!
- 59% have at least one duplicate; many images have many more duplicates



# How Search Engines do Image Indexing – First Use Tags

- **Search over tags associated with images**
  - Users manually add tags to images
  - Find images with tags that match the query keyword
  - Flickr, Facebook, and many other image hosting sites ask users to tag their uploaded images
- **Limitations**
  - Tags require human effort to create
  - Tags may be wrong, misleading, or just too general to be useful



## How Search Engines do Image Indexing – Next Use Surrounding Text

- Use text associated with images for indexing
  - Search web for images, <img src=...> and then use:
    - Text in URL for image filename
    - Text in HTML on page
- Example: Google Image Search for “Sunset” gives
  - Sunset at Rocky Point in Australia
  - Sunset Beach, Oahu
  - Frank Smiles at Sunset
- A single keyword like “Sunset” produces a diverse set of results, no surprise



Sunset at Rocky Point



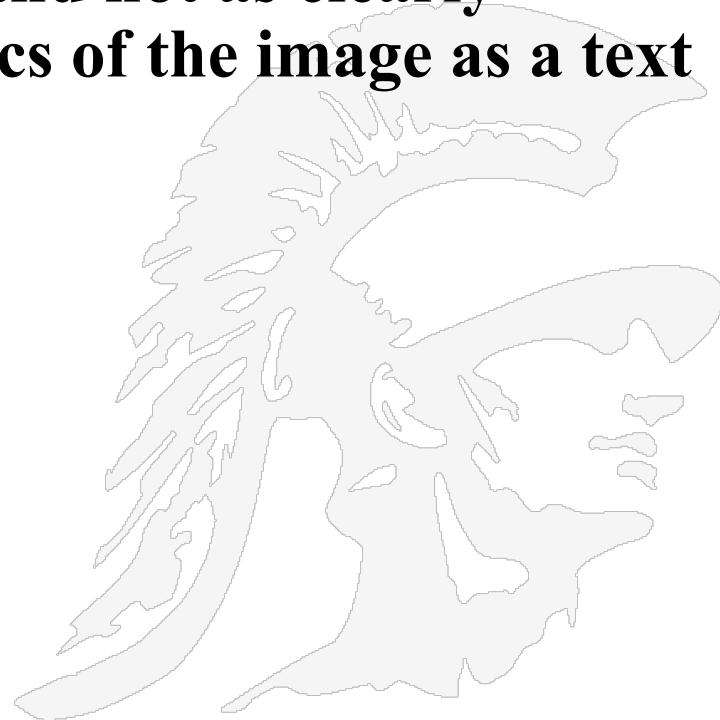
Frank Smiles  
at Sunset



Sunset Beach, Oahu

## How Search Engines do Image Indexing - Finally Use Feature Extraction

- Feature extraction from images is far more difficult than identifying surrounding text
- Features may be low-level and not as clearly associated with the semantics of the image as a text description



# Feature Extraction

## 3 Types of Image Features

- Typical examined features are those related to color, texture, and shape

### 1. Primitive features

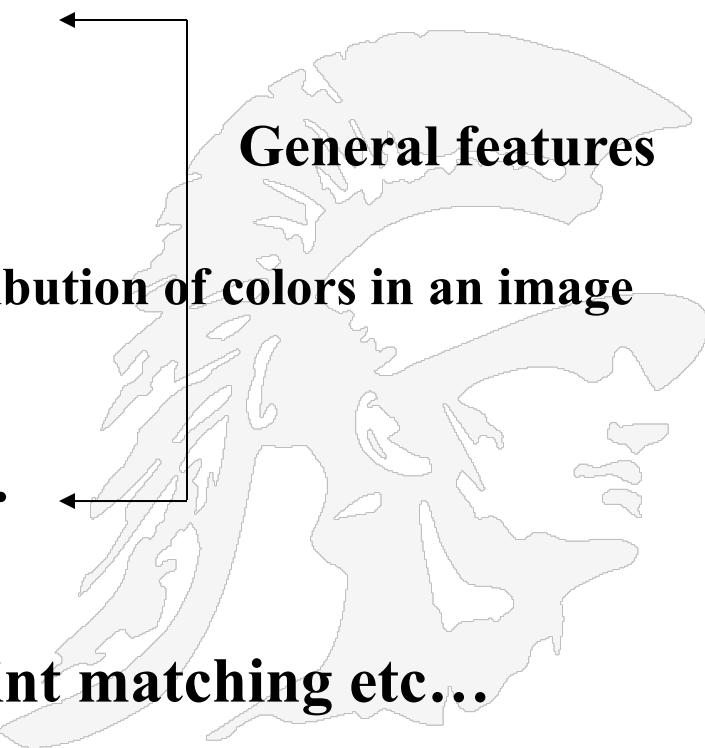
- Mean color (RGB)
- Color Histogram
  - A representation of the distribution of colors in an image

### 2. Semantic features:

- Color Layout, texture etc...

### 3. Domain specific features

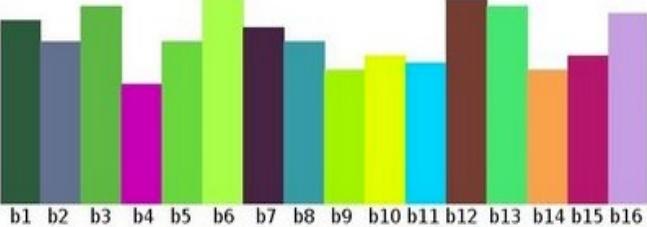
- Face recognition, fingerprint matching etc...



# Feature Extraction

## Color Histograms

- Histograms are collected counts of data organized into a set of predefined bins
- A color histogram measures the intensity of the color of every pixel
- To the right is a matrix containing the intensity of an image, a value between 0 and 255
- We can divide the values in the range [0, 255] into 16 bins: [0,15]U[16,31]U... U[240,255] and keep count of the number of pixels that fall in each bin; we get

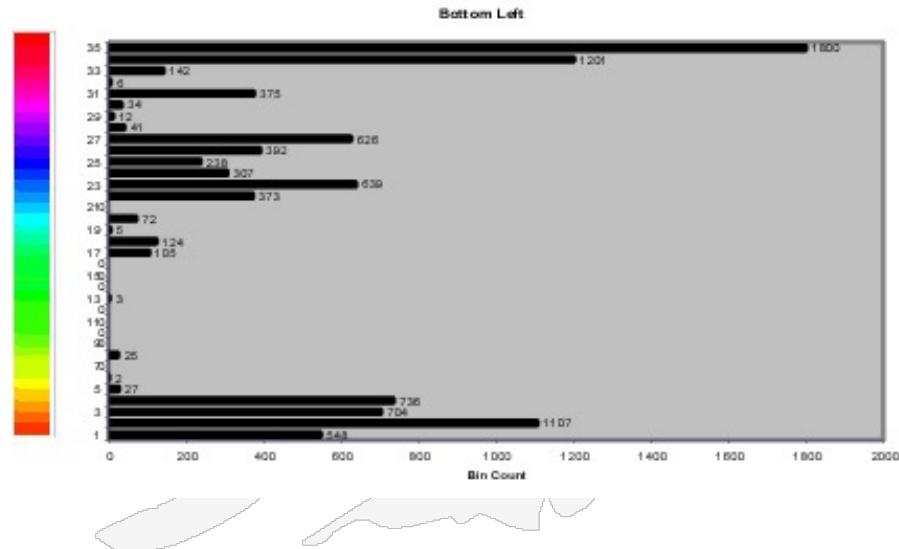


254	143	203	176	109	229	177	220	192	9	229	142	138	64	0	63	26	8	86	82
27	68	231	75	141	107	149	210	13	239	141	35	68	242	110	208	244	0	33	88
54	42	17	215	230	254	47	41	98	180	55	253	235	47	122	208	76	110	152	100
9	186	192	71	104	193	88	171	37	233	18	147	174	1	143	211	176	188	192	68
179	20	238	182	190	132	41	248	22	134	83	133	110	254	176	238	168	234	51	204
232	25	0	183	174	129	61	30	110	189	0	173	197	183	153	43	22	87	68	118
235	35	151	165	129	81	239	170	195	94	38	21	67	101	58	37	196	149	52	154
155	242	54	0	104	109	189	47	130	254	225	156	31	181	121	15	128	35	252	205
223	114	79	129	147	6	201	68	89	107	58	44	253	84	38	1	62	5	231	218
55	188	237	188	80	101	131	241	68	133	124	151	111	28	190	4	240	78	117	145
152	155	229	78	90	217	219	105	116	77	38	49	2	9	214	181	205	116	135	33
182	94	176	198	20	149	57	223	232	113	32	45	177	15	31	179	100	119	208	81
224	118	124	172	75	29	69	180	187	195	41	44	8	170	158	101	131	31	28	112
238	83	38	7	83	69	173	183	98	237	67	227	18	218	248	237	75	192	201	146
88	195	224	207	140	22	31	118	234	34	162	116	23	47	68	242	169	152	110	248
140	37	101	230	246	145	122	64	27	58	229	1	225	143	91	100	98	90	40	195
251	4	178	139	121	95	97	174	249	162	77	115	223	186	182	82	63	252	83	198
179	180	223	230	87	182	148	78	176	19	17	4	184	176	183	102	83	81	132	206
173	137	185	242	181	181	214	49	74	238	197	37	98	102	15	217	148	8	102	188
85	9	17	222	18	210	70	21	78	241	184	216	93	93	208	102	153	212	119	47

# Feature Extraction

## Similar Color Content

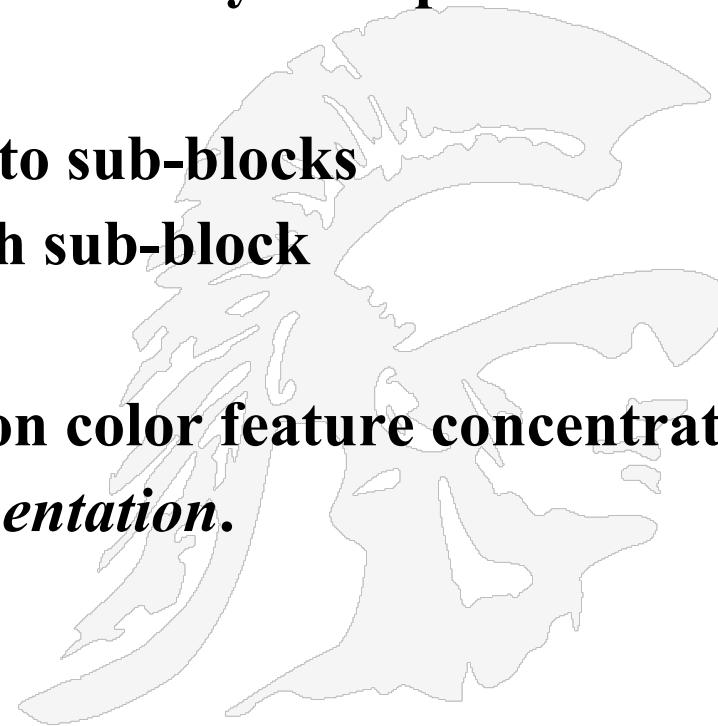
- A **correlogram** for an image is a table indexed by color pairs, where the  $d$ -th entry for row  $(i,j)$  specifies the probability of finding a pixel of color  $j$  at a distance  $d$  from a pixel of color  $i$  in this image. Here  $d$  is chosen from a set of distance values  $D$ .
- An **autocorrelogram** captures *spatial correlation between identical colors only*. This information is a subset of the correlogram and consists of rows of the form  $(i,j)$  only.



Correlograms answer questions like: is the data random  
is one area of the image related to another?

# Feature Extraction Color Layout

- **Need for Color Layout**
  - *Global* color features give too many false positives
- **How color layout works:**
  - Divide the whole image into sub-blocks
  - Extract features from each sub-block
- **One step further**
  - Divide into regions based on color feature concentration
  - This process is called *segmentation*.



# Example: Color layout

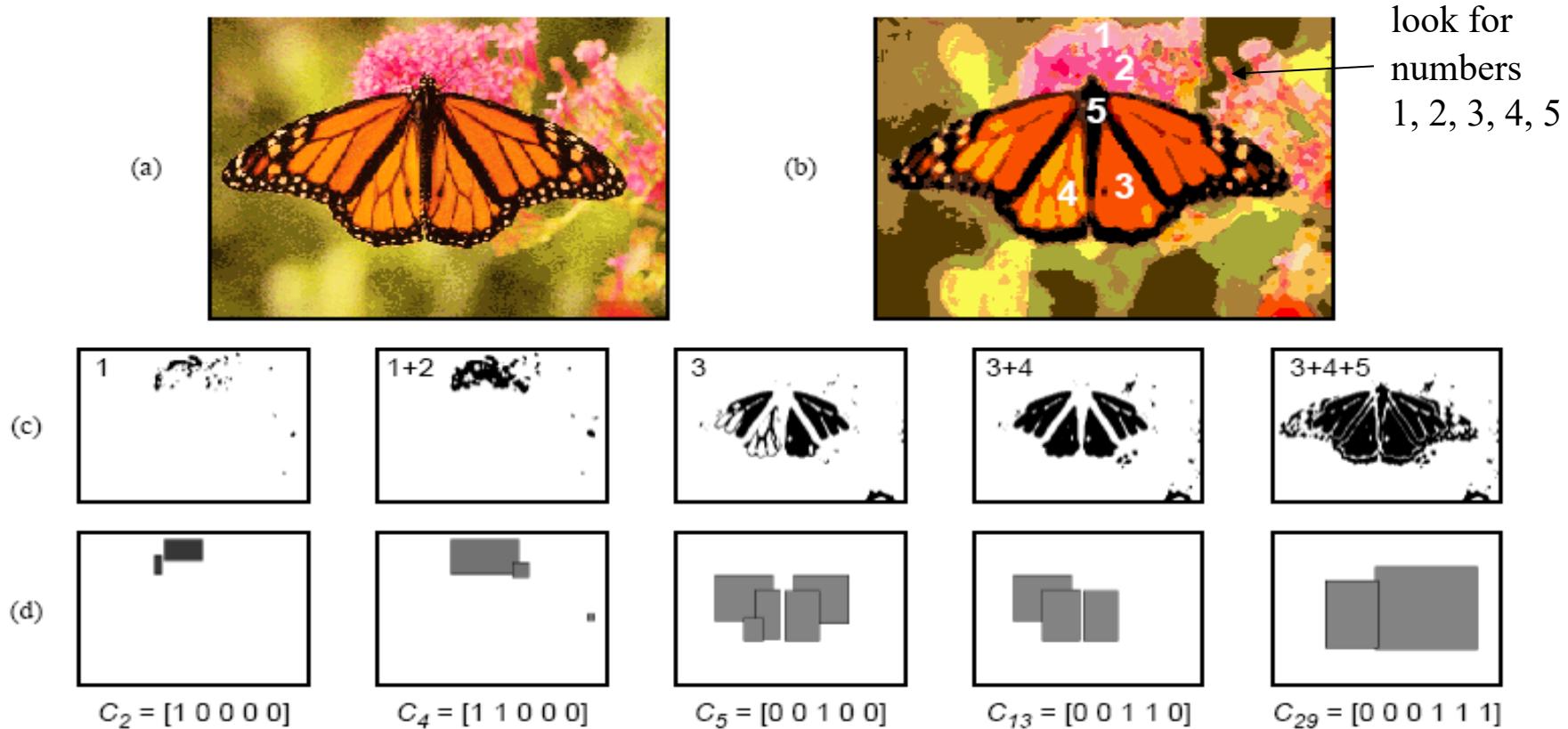


FIGURE 2. (a) *Butterfly* color image, (b) processed color image with 30 colors, (c) pixels from image (b) belonging to color set  $C_i$ , (d) minimum bounding rectangles (MBRs) for extracted regions used to index the image collection.

\*\* Image adapted from Smith and Chang : Single Color Extraction and Image Query

Copyright Eric Horowitz, 2011-2022

## Example: Texture and Shape

- **Texture – an innate property of all surfaces**
  - e.g. clouds, trees, bricks, hair etc...
  - Refers to visual patterns of homogeneity
  - Texture is spatial arrangement of gray levels in the image
- **Shape features describe the form of object boundaries and edges**
- **Examples:**



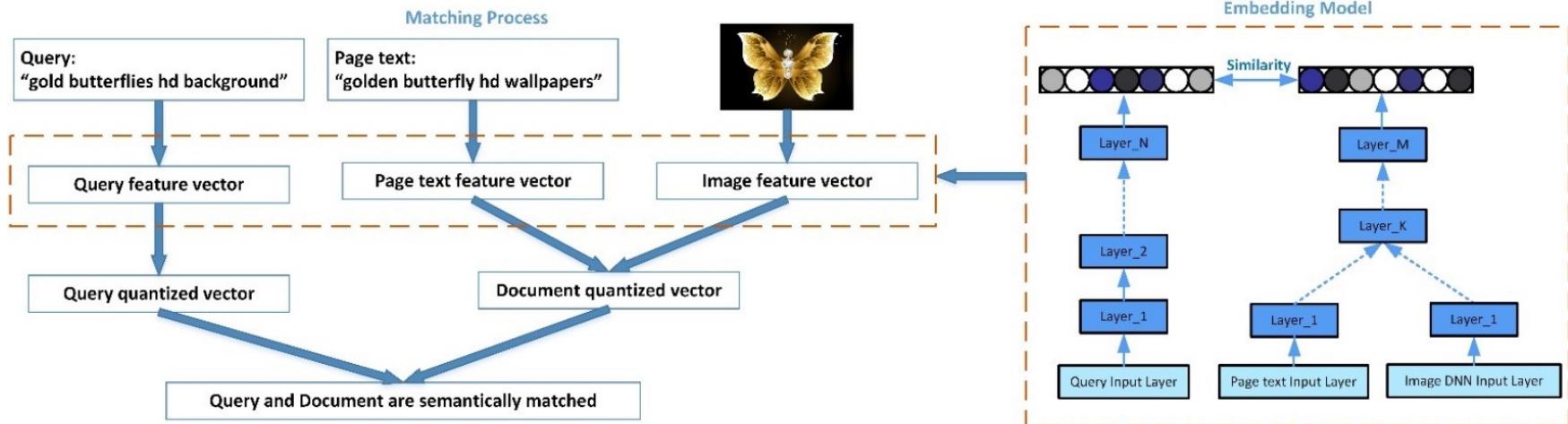
# How Bing does Deep Learning of Images

- Deep learning maps an image to a vector through a process called *image embedding*
- Similarly, natural language processing techniques let us read the words of a query and represent it as a vector through a process called *text embedding*
- A model is first trained using deep learning techniques to map (embed) a query and an image to a high dimensional vector space in such a way that the corresponding vectors are similar if the image is semantically relevant to the query, and further apart otherwise
- Top level image matching algorithm
  1. A matching stage - to select candidate images from a huge index of images
  2. Multiple ranking stages - to use computationally expensive methods to score each candidate image independently and rank all images
  3. Multiple set ranking stages - to re-rank previous candidates lists considering information from the entire candidate set, not just independent images



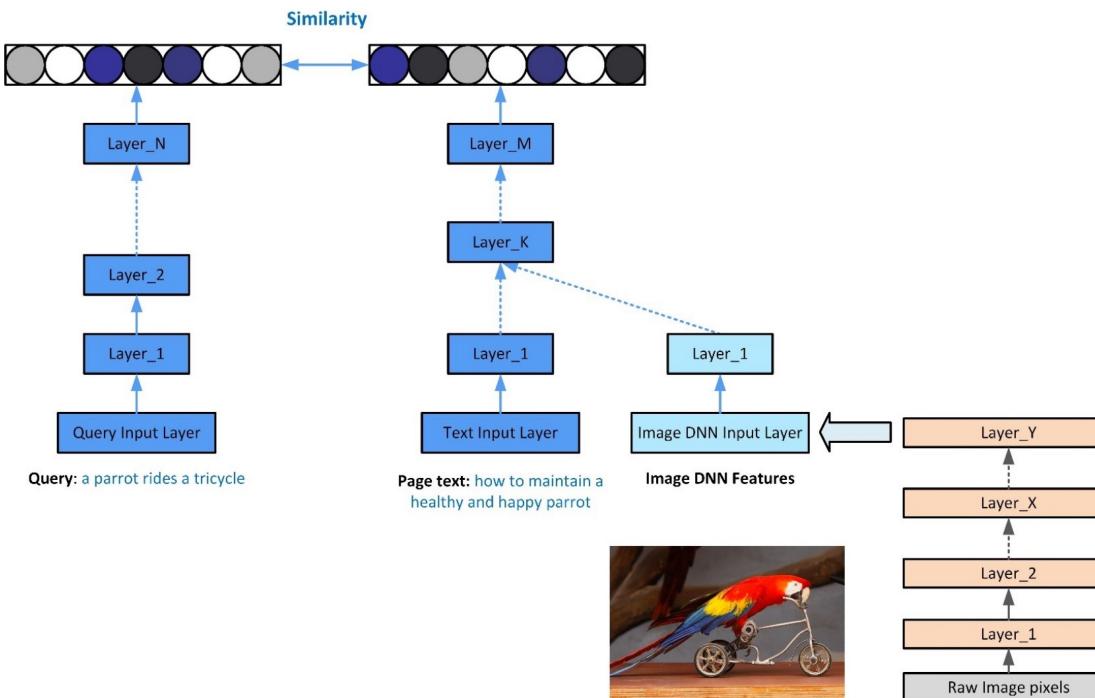
# Bing Deep Learning in Matching

- The matching stage selects a candidate set of images from billions of images
- During matching, it is essential to obtain a result set with *high recall* and moderate precision.
- query, image and page text are embedded by a deep learning model into embedding vectors. Then a vector search algorithm is used to scan a billions-scale index to find the N-best document vectors for a query vector.
- below: the left part shows that even though the page text ‘golden butterfly hd wallpapers’ is not exactly the same as the query ‘gold butterflies hd background’, we are able to retrieve the relevant image based on query embedding, page text embedding and image embedding, which are generated from the deep learning model as shown in the right diagram.



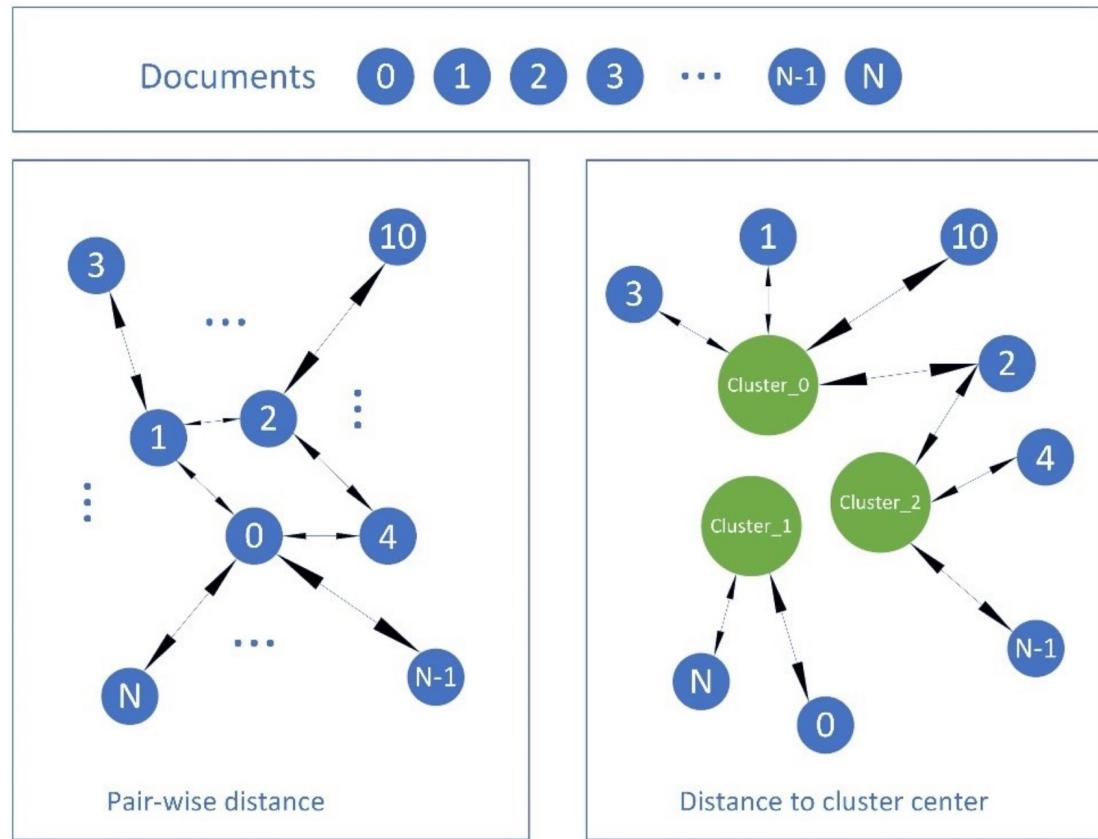
# Bing Deep Learning in Ranking

- Now rank all the candidate images according to their relevance to the query
- Pruning of images has occurred - reuse the previously computed embedding vectors but do more exact calculations of semantic distance between the vectors.
- in the diagram below, the page contains the text ‘how to maintain a healthy and happy parrot’ however the system will recognize that the image matches the query ‘a parrot rides a tricycle’ because it is looking inside the image using image embedding.



# Bing Deep Learning in Set Ranking

- deep learning is used to embed the image and page text into vectors and then compute pair-wise distances between all images in the candidate list.
- compute higher-order features from these pairwise distances and use them to re-rank the candidate list.
- For example, the diagram shows one such approach where the distance of each image from the cluster centroids is estimated from the vectors of the top-N images.



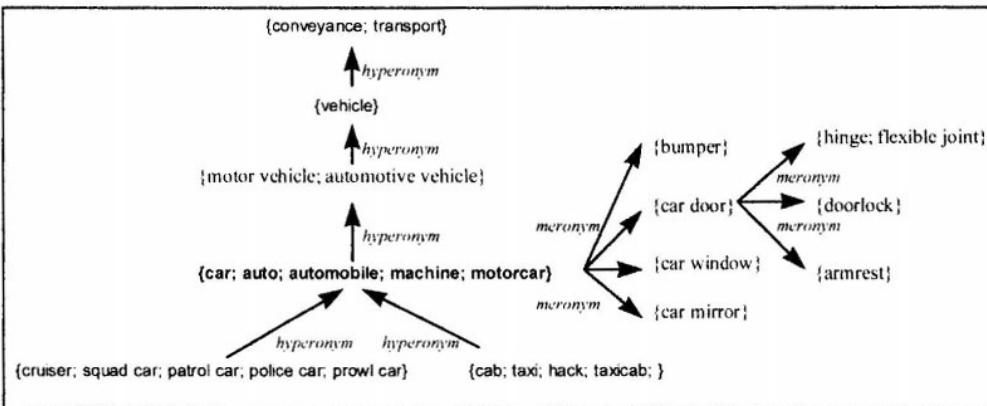
# ImageNet: A Large-Scale Hierarchical Image Database

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li and Li Fei-Fei Dept. of  
Computer Science, Princeton University, USA



# Recall WordNet

- We discussed WordNet in the lecture on Question/Answering;
- WordNet is a large lexical database of English.
  - Nouns, verbs, adjectives and adverbs are grouped into sets of cognitive synonyms (synsets), each expressing a distinct concept.
- WordNet as an ontology



**WordNet Search - 3.1**  
[WordNet home page](#) - [Glossary](#) - [Help](#)

Word to search for:

Display Options:

Key: "S:" = Show Synset (semantic) relations, "W:" = Show Word (lexical) relations  
 Display options for sense: (gloss) "an example sentence"

**Noun**

- S: (n) [unidentified flying object](#), [UFO](#), [flying saucer](#) (an (apparently) flying object whose nature is unknown; especially those considered to have extraterrestrial origins)
  - [direct hypernym](#) / [inherited hypernym](#) / [sister term](#)
  - S: (n) [apparition](#), [phantom](#), [phantasm](#), [phantasma](#), [fantasm](#), [shadow](#) (something existing in perception only) "a *ghostly apparition at midnight*"

- **ImageNet** is an image database organized according to the *WordNet hierarchy*, in which each node of the hierarchy is depicted by hundreds and thousands of images.
- Currently there are an average of over five hundred images per node
- 14,197,122 images covering 21,841 synsets indexed



14,197,122 images, 21841 synsets indexed

[Home](#) [Download](#) [Challenges](#) [About](#)
Not logged in. [Login](#) | [Signup](#)

### An Update to the ImageNet Website and Dataset

March 11, 2021

We are proud to see ImageNet's wide adoption going beyond what was originally envisioned. However, the decade-old website was burdened by growing download requests. To serve the community better, we have redesigned the [website](#) and upgraded its hardware. The new website is simpler; we removed tangential or outdated functions to focus on the core use case—enabling users to [download the data](#), including the full ImageNet dataset and the [ImageNet Large Scale Visual Recognition Challenge \(ILSVRC\)](#).

Meanwhile, the computer vision community has progressed, and so has ImageNet. The dataset was created to benchmark object recognition—at a time when it barely worked. The problem then was how to collect labeled images at a sufficiently large scale to be able to train complex models in laboratories. Today, computer vision is in real-world systems impacting people's Internet experience and daily lives. An emerging problem now is how to make sure computer vision is fair and preserves people's privacy. We are continually evolving ImageNet to address these emerging needs.

In a [FAT 2020 paper](#), we filtered 2,702 synsets in the "person" subtree that may cause problematic behaviors of the model. We have updated the full ImageNet data on the website to remove these synsets. The update does not affect the 1,000 categories in ILSVRC.

In a [more recent paper](#), we investigate privacy issues in ILSVRC. 997 out of 1000 categories in ILSVRC are not people categories; nevertheless, many incidental people are in the images, whose privacy is a concern. We first annotated faces in the images and then constructed a face-blurred version of ILSVRC. Experiments show that one can use the face-blurred version for benchmarking object recognition and for transfer learning with only marginal loss of accuracy. We release our [face annotations](#) to facilitate further research on privacy-aware visual recognition.

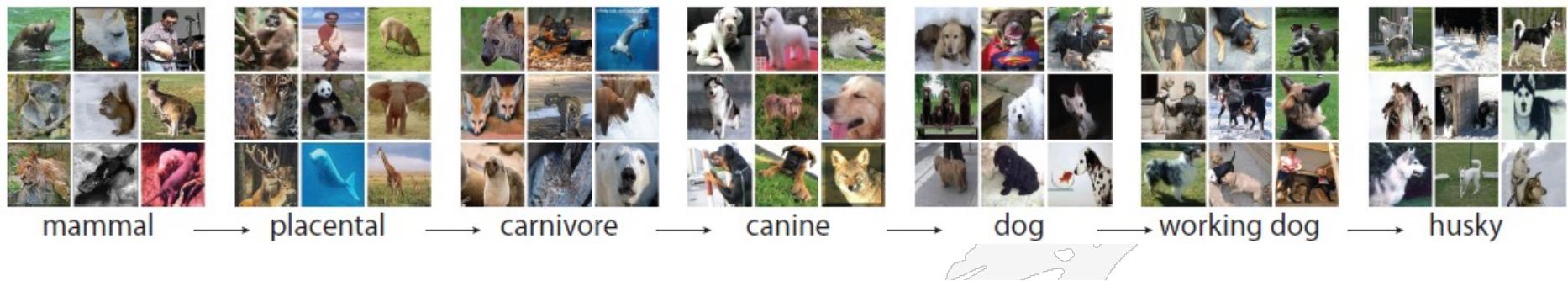
Team members working on these new improvements: [Kaiyu Yang](#) (Princeton), [Jacqueline Yau](#) (Stanford), [Li Fei-Fei](#) (Stanford), [Jia Deng](#) (Princeton), [Olga Russakovsky](#) (Princeton).

© 2020 Stanford Vision Lab, Stanford University, Princeton University [imagenet.help.desk@gmail.com](#) Copyright infringement



# ImageNet is a Knowledge Ontology

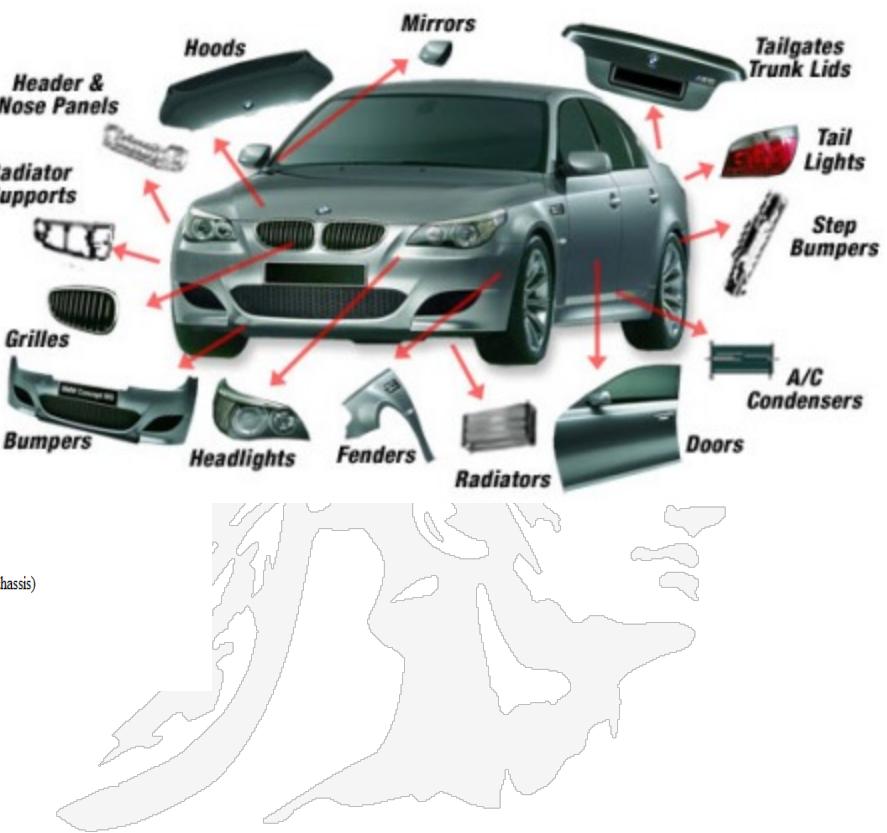
- Below is the taxonomy for a Husky dog from WordNet and the corresponding ImageNet taxonomy of images



- S: (n) [Eskimo dog](#), [husky](#) (breed of heavy-coated Arctic sled dog)
  - [direct hypernym](#) / [inherited hypernym](#) / [sister term](#)
  - S: (n) [working dog](#) (any of several breeds of usually large powerful dogs bred to work as draft animals and guard and guide dogs)
  - S: (n) [dog](#), [domestic dog](#), [Canis familiaris](#) (a member of the genus Canis (probably descended from the common wolf) that has been domesticated by man since prehistoric times; occurs in many breeds) "the dog barked all night"
    - S: (n) [canine](#), [canid](#) (any of various fissiped mammals with nonretractile claws and typically long muzzles)
    - S: (n) [carnivore](#) (a terrestrial or aquatic flesh-eating mammal) "terrestrial carnivores have four or five clawed digits on each limb"
    - S: (n) [placental](#), [placental mammal](#), [eutherian](#), [eutherian mammal](#) (mammals having a placenta; all mammals except monotremes and marsupials)
    - S: (n) [mammal](#), [mammalian](#) (any warm-blooded vertebrate having the skin more or less covered with hair; young are born alive except for the small subclass of monotremes and nourished with milk)
      - S: (n) [vertebrate](#), [craniate](#) (animals having a bony or cartilaginous skeleton with a segmented spinal column and a large brain enclosed in a skull or cranium)
      - S: (n) [chordate](#) (any animal of the phylum Chordata having a notochord or spinal column)
        - S: (n) [animal](#), [animate being](#), [beast](#), [brute](#), [creature](#), [fauna](#) (a living organism characterized by voluntary movement)
        - S: (n) [organism](#), [being](#) (a living thing that has (or can develop) the ability to act or function independently)
        - S: (n) [living thing](#), [animate thing](#) (a living (or once living) entity)
          - S: (n) [whole](#), [unit](#) (an assemblage of parts that is regarded as a single entity) "how big is that part compared to the whole?", "the team is a unit"
            - S: (n) [object](#), [physical object](#) (a tangible and visible entity; an entity that can cast a shadow) "it was full of rackets, balls and other objects"
            - S: (n) [physical entity](#) (an entity that has physical existence)
          - S: (n) [entity](#) (that which is perceived or known or inferred to have its own distinct existence (living or nonliving))

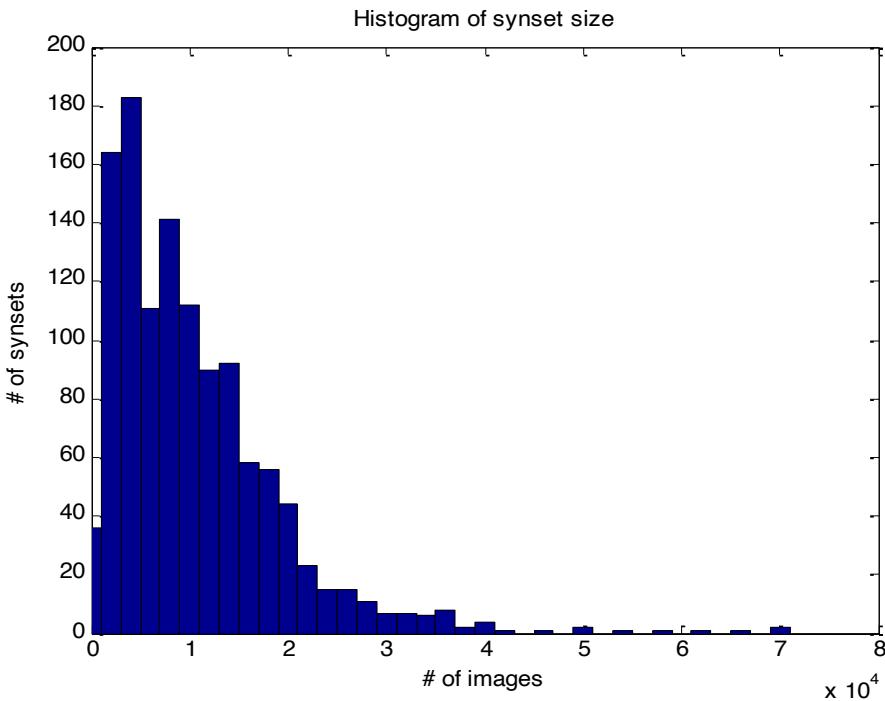
# ImageNet is a Knowledge Ontology

- [S: \(n\) car, auto, automobile, machine, motorcar](#) (a motor vehicle with four wheels; usually propelled by an internal combustion engine) "he needs a car to get to work"
  - [direct hyponym / full hyponym](#)
  - [part meronym](#)
    - [S: \(n\) accelerator, accelerator pedal, gas pedal, gas, throttle, gun](#) (a pedal that controls the throttle valve) "he stepped on the gas"
    - [S: \(n\) air bag](#) (a safety restraint in an automobile; the bag inflates on collision and prevents the driver or passenger from being thrown forward)
    - [S: \(n\) auto accessory](#) (an accessory for an automobile)
    - [S: \(n\) automobile engine](#) (the engine that propels an automobile)
    - [S: \(n\) automobile horn, car horn, motor horn, horn, hooter](#) (a device on an automobile for making a warning noise)
    - [S: \(n\) buffer, fender](#) (a cushion-like device that reduces shock due to an impact)
    - [S: \(n\) bumper](#) (a mechanical device consisting of bars at either end of a vehicle to absorb shock and prevent serious damage)
    - [S: \(n\) car door](#) (the door of a car)
    - [S: \(n\) car mirror](#) (a mirror that the driver of a car can use)
    - [S: \(n\) car seat](#) (a seat in a car)
    - [S: \(n\) car window](#) (a window in a car)
    - [S: \(n\) fender, wing](#) (a barrier that surrounds the wheels of a vehicle to block splashing water or mud) "in Britain they call a fender a wing"
    - [S: \(n\) first gear, first, low gear, low](#) (the lowest forward gear ratio in the gear box of a motor vehicle; used to start a car moving)
    - [S: \(n\) floorboard](#) (the floor of an automobile)
    - [S: \(n\) gasoline engine, petrol engine](#) (an internal-combustion engine that burns gasoline; most automobiles are driven by gasoline engines)
    - [S: \(n\) glove compartment](#) (compartment on the dashboard of a car)
    - [S: \(n\) grill, radiator grille](#) (grating that admits cooling air to car's radiator)
    - [S: \(n\) high gear, high](#) (a forward gear with a gear ratio that gives the greatest vehicle velocity for a given engine speed)
    - [S: \(n\) hood, bonnet, cowl, cowling](#) (protective covering consisting of a metal part that covers the engine) "there are powerful engines under the hoods or cowling in order to repair the plane's engine"
    - [S: \(n\) luggage compartment, automobile trunk, trunk](#) (compartment in an automobile that carries luggage or shopping or tools) "he put his golf bag in the
    - [S: \(n\) rear window](#) (car window that allows vision out of the back of the car)
    - [S: \(n\) reverse, reverse gear](#) (the gears by which the motion of a machine can be reversed)
    - [S: \(n\) roof](#) (protective covering on top of a motor vehicle)
    - [S: \(n\) running board](#) (a narrow footboard serving as a step beneath the doors of some old cars)
    - [S: \(n\) stabilizer bar, anti-sway bar](#) (a rigid metal bar between the front suspensions and between the rear suspensions of cars and trucks; serves to stabilize the chassis)
    - [S: \(n\) sunroof, sunshine-roof](#) (automobile roof having a sliding or raisable panel) "sunshine-roof is a British term for 'sunroof'"
    - [S: \(n\) tail fin, tailfin, fin](#) (one of a pair of decorations projecting above the rear fenders of an automobile)
    - [S: \(n\) third gear, third](#) (the third from the lowest forward ratio gear in the gear box of a motor vehicle) "you shouldn't try to start in third gear"
    - [S: \(n\) window](#) (a transparent opening in a vehicle that allow vision out of the sides or back; usually is capable of being opened)



## Some Synsets Have Very Few Images While Some Have Many

- “Mammal” subtree ( 1180 synsets )
- Average # of images per synset: 10.5K



Most populated	Least populated
Humankind (118.5k)	Algeripithecus minutus (90)
Kitty, kitty-cat ( 69k)	Striped muishond (107)
Cattle, cows ( 65k)	Mylodonitid (127)
Pooch, doggie ( 62k)	Greater pichiciego (128)
Cougar, puma ( 57k)	Damaraland mole rat (188)
Frog, toad ( 53k )	Western pipistrel (196)
Hack, jade, nag (50k)	Muishond (215)



# Constructing ImageNet

## ► 2-step process

Step 1 :  
Collect candidate  
images Via the Internet



Step 2 :  
Clean up the candidate  
Images by humans



## Step 1: Collect Candidate Images from the Internet

- ▶ For each synset, the queries are the set of WordNet synonyms
- ▶ Accuracy of Internet Image search results: 10 %
  - For 500-1000 clean images, needs 10K images
- ▶ Query expansion
  - Synonyms: German police dog, German shepherd dog
  - Appending words form ancestors: sheepdog, dog
- ▶ Multiple Languages
  - Italian, Dutch, Spanish, Chinese e.g. 德国牧羊犬, pastore tedesco
- ▶ More engines: Yahoo!, flickr, Google
- ▶ Parallel downloading



## Step 2: Clean Up the Candidate Images by Humans

- ▶ Rely on humans to verify each candidate image collected for a given synset
- ▶ Amazon Mechanical Turk (AMT)
  - used for labeling vision data
  - 300 images: 0.02 dollar
  - 14,197,122 images: 946 dollars
  - 10 repetition: 9460 dollars
  - Jul 2008 -Apr 2010: 11 million images
- ▶ Present the users with a set of candidate images and the definition of the target synset
- ▶ let users select the best match ones



# A Task on AMT

(amazon mechanical turk)

Blank Instructions Unsure? Look up in Wikipedia Google [ Additional input ] No good photos? Have expertise? comments? Click here!

First time workers please click here for instructions.

Click on the photos that contain the object or depict the concept of: **cow**: mature female of mammals of which the male is called 'bull' [PLEASE READ DEFINITION CAREFULLY].

Pick as many as possible. PHOTOS ONLY, NO PAINTINGS, DRAWINGS, etc. It's OK to have other objects, multiple instances, occlusion or text in the image.

Do not use back or forward button of your browser. OCCASIONALLY THERE MIGHT BE ADULT OR DISTURBING CONTENT.



Below are the photos you have selected FROM THIS PAGE ONLY | they will be saved when you navigate to other pages | Click to deselect.

what's this? select all deselect all < page 1 of 5 > Submit PREVIEW MODE. TO WORK ON THIS HIT, ACCEPT IT FIRST.

Blank Instructions Unsure? Look up in Wikipedia Google [ Additional input ] No good photos? Have expertise? comments? Click here!

First time workers please click here for instructions.

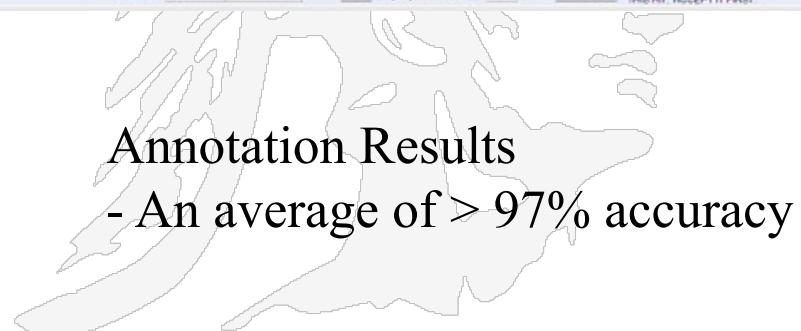
Click on the photos that contain the object or depict the concept of: **lion, king of beasts, Panthera leo**: large gregarious predatory feline of Africa pick as many as possible. PHOTOS ONLY, NO PAINTINGS, DRAWINGS, etc. It's OK to have other objects, multiple instances, occlusion or text in the image. Do not use back or forward button of your browser. OCCASIONALLY THERE MIGHT BE ADULT OR DISTURBING CONTENT.



Below are the photos you have selected FROM THIS PAGE ONLY | they will be saved when you navigate to other pages | Click to deselect.

what's this? select all deselect all < page 1 of 4 > Submit PREVIEW MODE. TO WORK ON THIS HIT, ACCEPT IT FIRST.

Workers do annotation on AMT  
 -Multiple annotations for each images



# Ensure Accuracy

## ► Users Enhancement

- Provide wiki and Google links for definitions
- Make sure workers read the definition
  - Definition quiz
- Allow more feedback. E.g. “unimaginable synset” expert opinion



Main Instructions Unsure? Look up in Wikipedia Google [Additional input]  
You can support Wikipedia by making a tax-deductible donation  
[article](#) [discussion](#) [edit this page](#) [history](#)

**Delta**  
From Wikipedia, the free encyclopedia

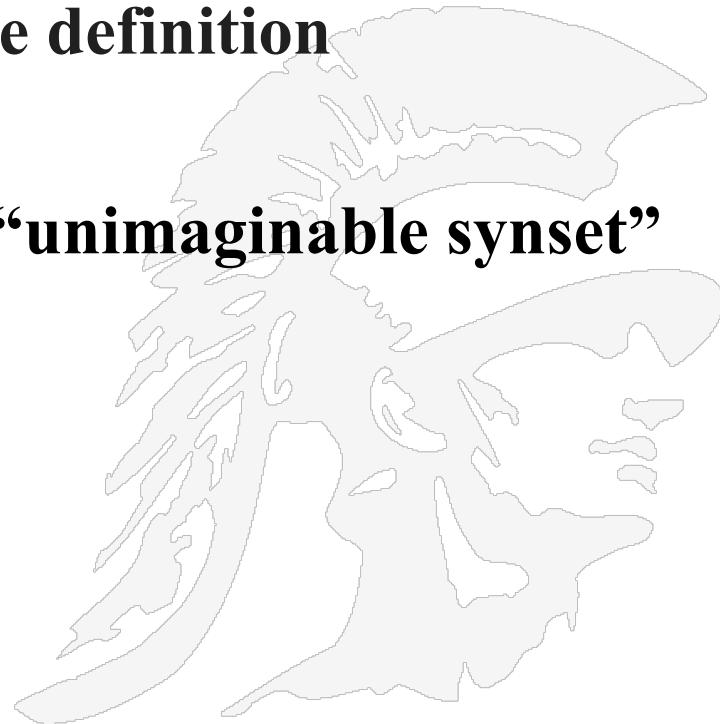
Delta commonly refers to:

- **Delta (letter)**, Δ or δ in the Greek alphabet, also usi
- **River delta**, a landform at the mouth of a river

Delta may also refer to:

**Places**

The page shows a sidebar with navigation links: Main page, Contents, Featured content, Current events, Random article.



# Ensure Accuracy

- Human users make mistakes
- Not all users follow the instructions
- Users do not always agree with each other
  - Subtle or confusing synsets, e.g. Burmese cat
- Quality Control System



User 1	Y	Y	Y
User 2	N	Y	Y
User 3	N	Y	Y
User 4	Y	N	Y
User 5	Y	Y	Y
User 6	N	N	Y

#Y	#N	Conf Cat	Conf BCat
0	1	0.07	0.23
1	0	0.85	0.69
1	1	0.46	0.49
2	0	0.97	0.83
0	2	0.02	0.12
3	0	0.99	0.90
2	1	0.85	0.68

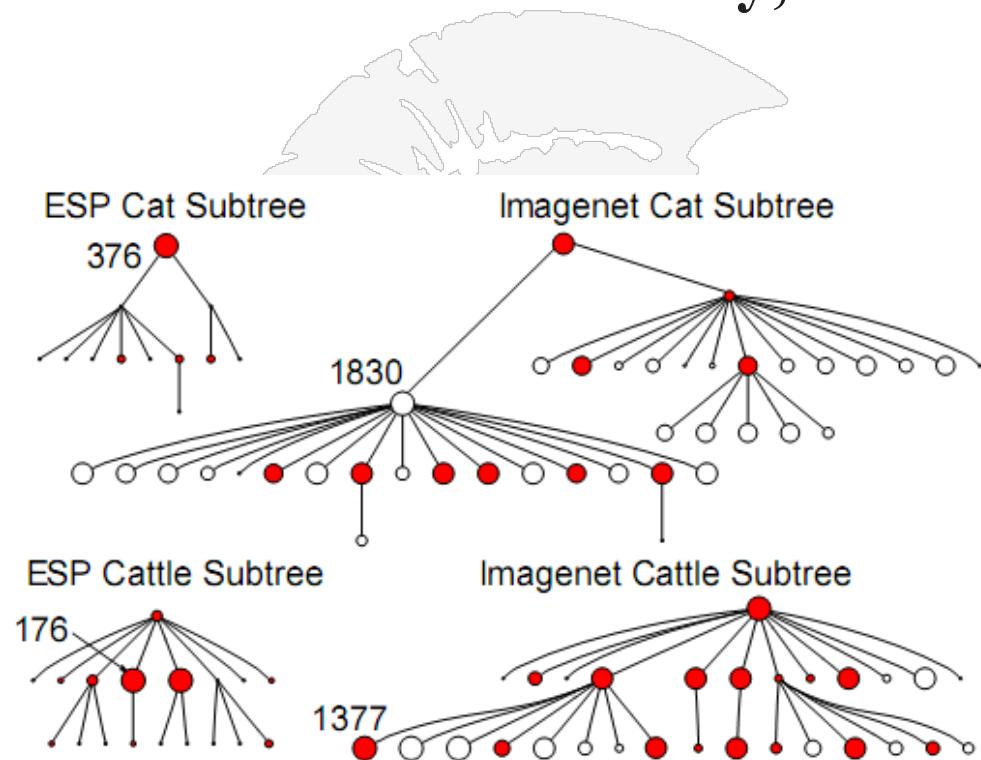
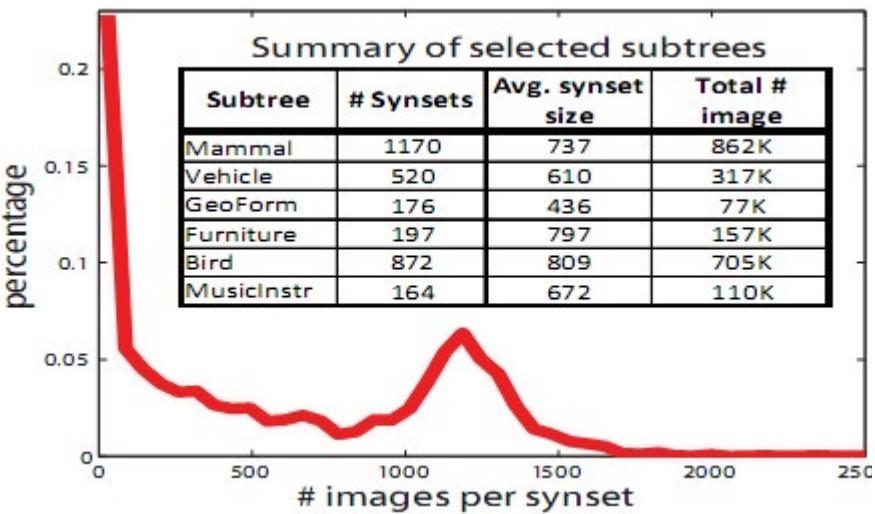
# Quality Control System

- ▶ randomly sample an initial subset of image to users
  - Have multiple users independently label same image
- ▶ obtain a confidence score table, indicating the probability of an image being a good image given the user votes
  - Different categories requires different levels of consensus
- ▶ Proceed until a pre-determined confidence score threshold reached



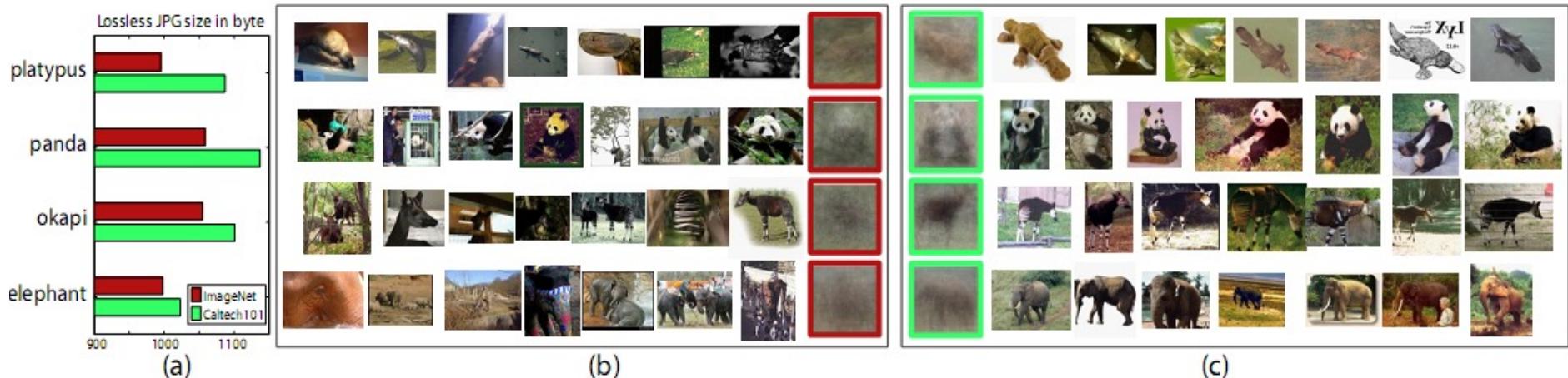
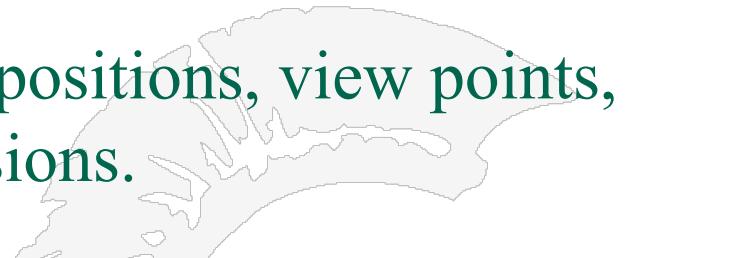
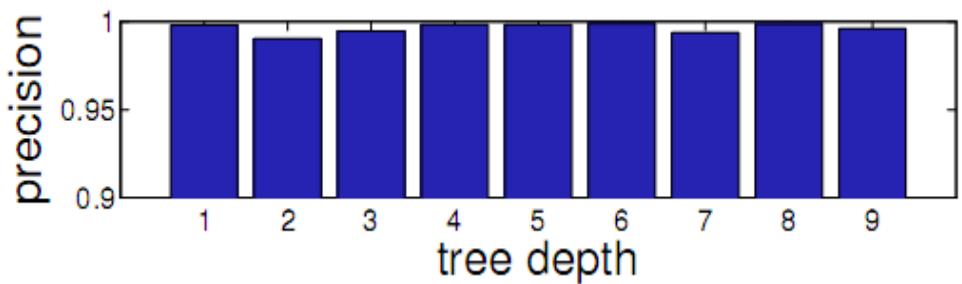
# Properties of ImageNet

- ▶ **Scale:** 12 subtrees, 3,2 million images, 5247 categories
- ▶ **Hierarchy:** densely populated semantic hierarchy, based on WordNet



# Properties of ImageNet

- ▶ **Accuracy:** clean dataset at all level
- ▶ **Diversity:** variable appearances, positions, view points, poses, background clutter, occlusions.



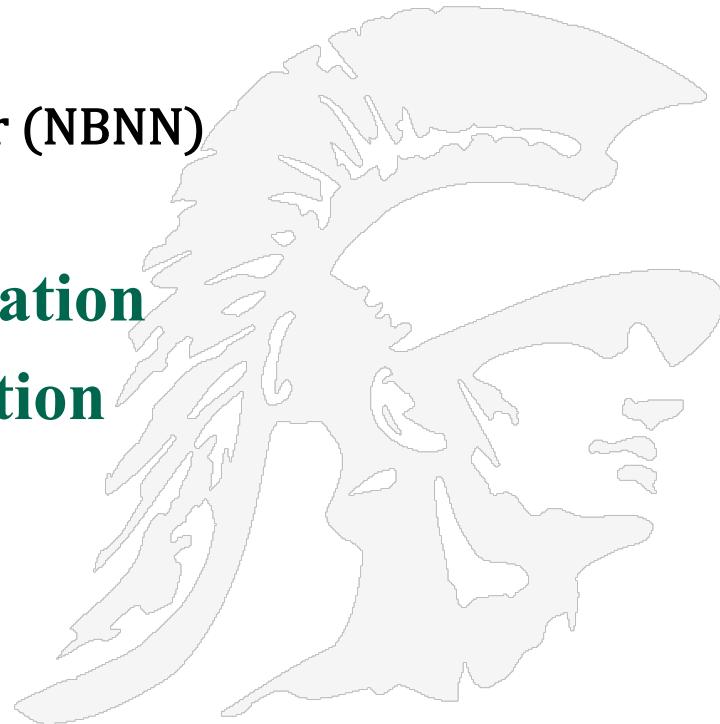
# ImageNet Applications

## ► Non-parametric Object Recognition

1. NN-voting + noisy ImageNet
2. NN-voting + clean ImageNet
3. Naive Bayesian Nearest Neighbor (NBNN)
4. NBNN-100

## ► Tree Based Image Classification

## ► Automatic Object Localization



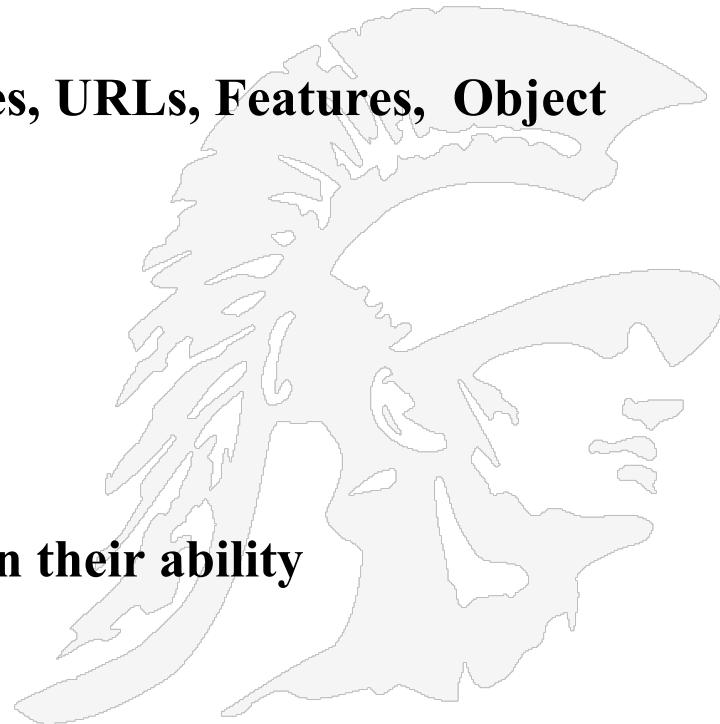
# Pros and Cons

## ► Pros

1. **Crowdsourcing**
2. **Benchmarking**
3. **Open: Download Original Images, URLs, Features, Object Attributes, API**

## ► Cons

1. **Improve algorithm: PageRank**
2. **AMT: hierarchical users based on their ability**
3. **Only one tag per image**



# Summary by the ImageNet Developers

- **ImageNet is intended to serve as**
  - A dataset
  - A knowledge ontology
- **Construction of large-scale image dataset is a relatively new research area**
  - Crowdsourcing might be the future of many such tasks
  - see an ImageNet developer's TED talk,  
<https://www.youtube.com/watch?v=40riCqvRoMs>
- **Benchmarking: what does classifying 10k+ image categories tell us?**
  - Computation matters
  - Size matters
  - Density matters
  - Hierarchy matters

