

← 1/40 → *** 10:08:01

Knowledge graphs

USC's work: <https://usc-isi-i2.github.io/home>

..

University of Southern California  USC

 USC **Viterbi**
School of Engineering

Outline

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



Copyright Ellis Horowitz, 2011-2022

2

••

University of Southern California  USC

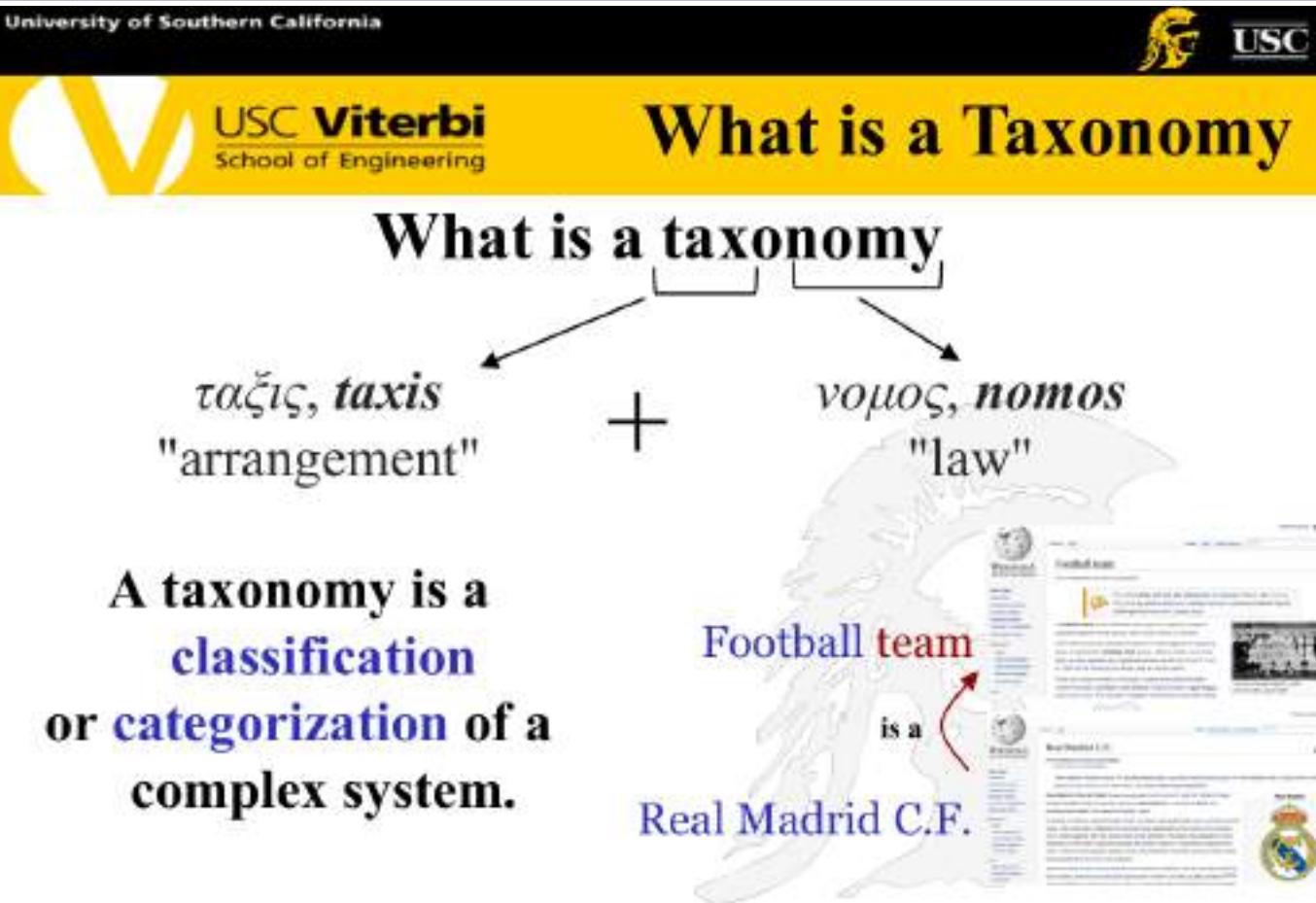
What is a Taxonomy

What is a taxonomy

$\tau\alpha\xi\iota\xi, \text{taxis}$
"arrangement" + $\nuομο\xi, \text{nomos}$
"law"

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

Football team
is a
Real Madrid C.F.



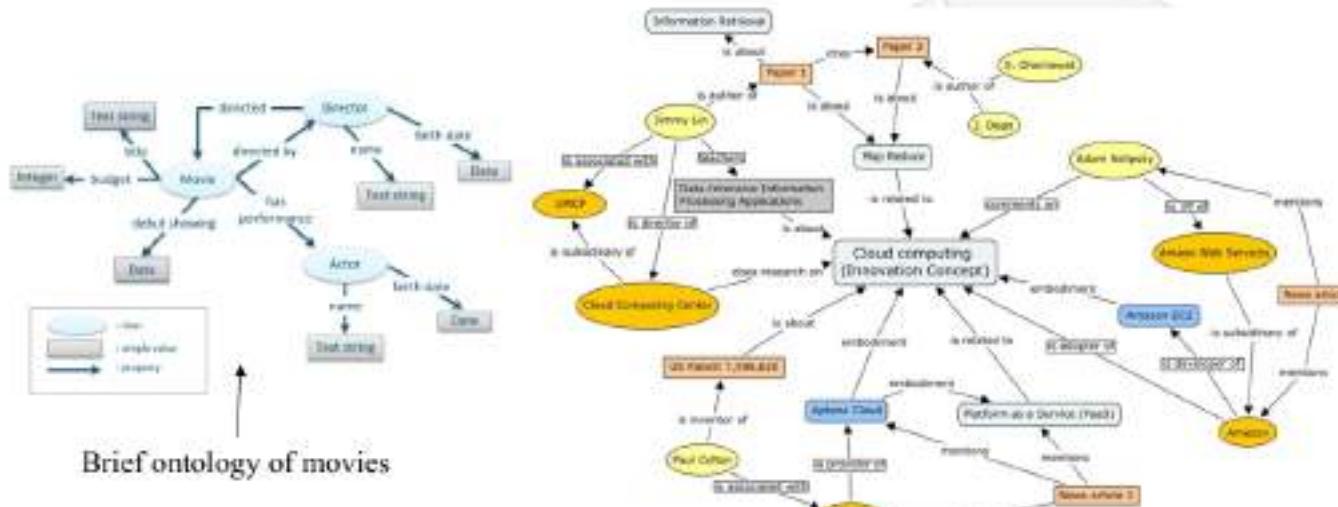
e.g. the ACM Computing Classification System, <https://dl.acm.org/ecs>
the Mathematics Subject Classification

•

University of Southern California  USC

What is an Ontology

- a set of concepts and categories in a subject area or domain that shows their properties and the relations between them, or
- 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
- 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

Cloud computing (Innovation Concept)

Copyright Ellis Horowitz, 2011-2022

4

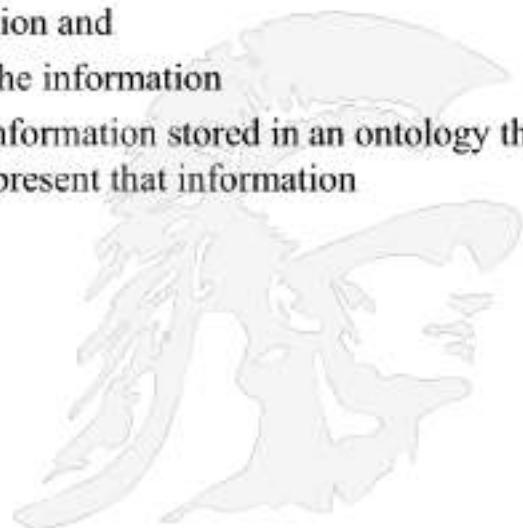
••

University of Southern California  USC

 USC **Viterbi**
School of Engineering

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

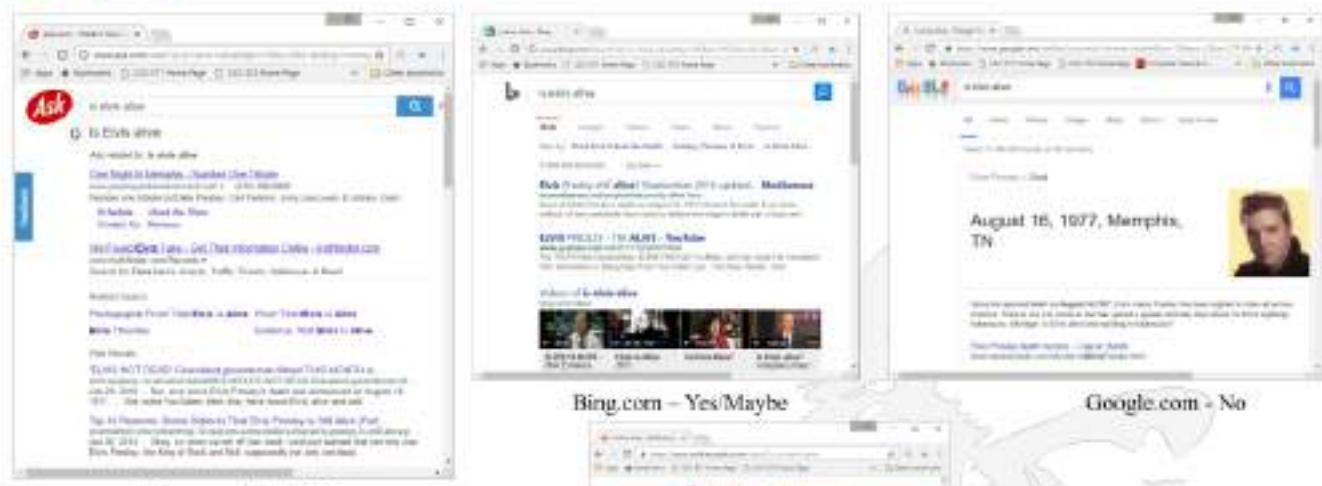


Copyright Ellis Horowitz, 2011-2022

5

University of Southern California 

Is Elvis Alive



- Ask.com – Mostly Yes**: Shows a snippet from a news article stating "Elvis Presley died on August 16, 1977, in Memphis, Tennessee. He was 47 years old at the time of his death. His death was ruled a heart attack. He had been suffering from heart problems for several years." Below it says "Top 10 Reasons Why People Think Elvis Presley Is Still Alive".
- Bing.com – Yes/Maybe**: Shows a snippet from a news article stating "Elvis Presley died on August 16, 1977, in Memphis, Tennessee. He was 47 years old at the time of his death. His death was ruled a heart attack. He had been suffering from heart problems for several years." Below it says "Top 10 Reasons Why People Think Elvis Presley Is Still Alive".
- Google.com – No**: Shows a snippet from a news article stating "Elvis Presley died on August 16, 1977, in Memphis, Tennessee. He was 47 years old at the time of his death. His death was ruled a heart attack. He had been suffering from heart problems for several years." Below it says "Top 10 Reasons Why People Think Elvis Presley Is Still Alive".
- Wolfram Alpha - No**: Shows a snippet from a news article stating "Elvis Presley died on August 16, 1977, in Memphis, Tennessee. He was 47 years old at the time of his death. His death was ruled a heart attack. He had been suffering from heart problems for several years." Below it says "Top 10 Reasons Why People Think Elvis Presley Is Still Alive".

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

Copyright Ellis Horowitz 2011-2022

••

University of Southern California  USC

 USC **Viterbi**
School of Engineering

KnowledgeBases-Developed by AI Community

- **To move away from just using keyword matching, search engines borrowed techniques developed by AI researchers**
- A *knowledgebase* (KB) is a technology used to store complex structured and unstructured information used by a computer system.
- **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***
 - Is a relational database a knowledgebase?
- **Examples of knowledgebases: Freebase, Google's Knowledge Graph, Apple's Siri, IBM's Watson**

Copyright Ellis Horowitz, 2011-2022

• •

University of Southern California  USC  USC

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](#)

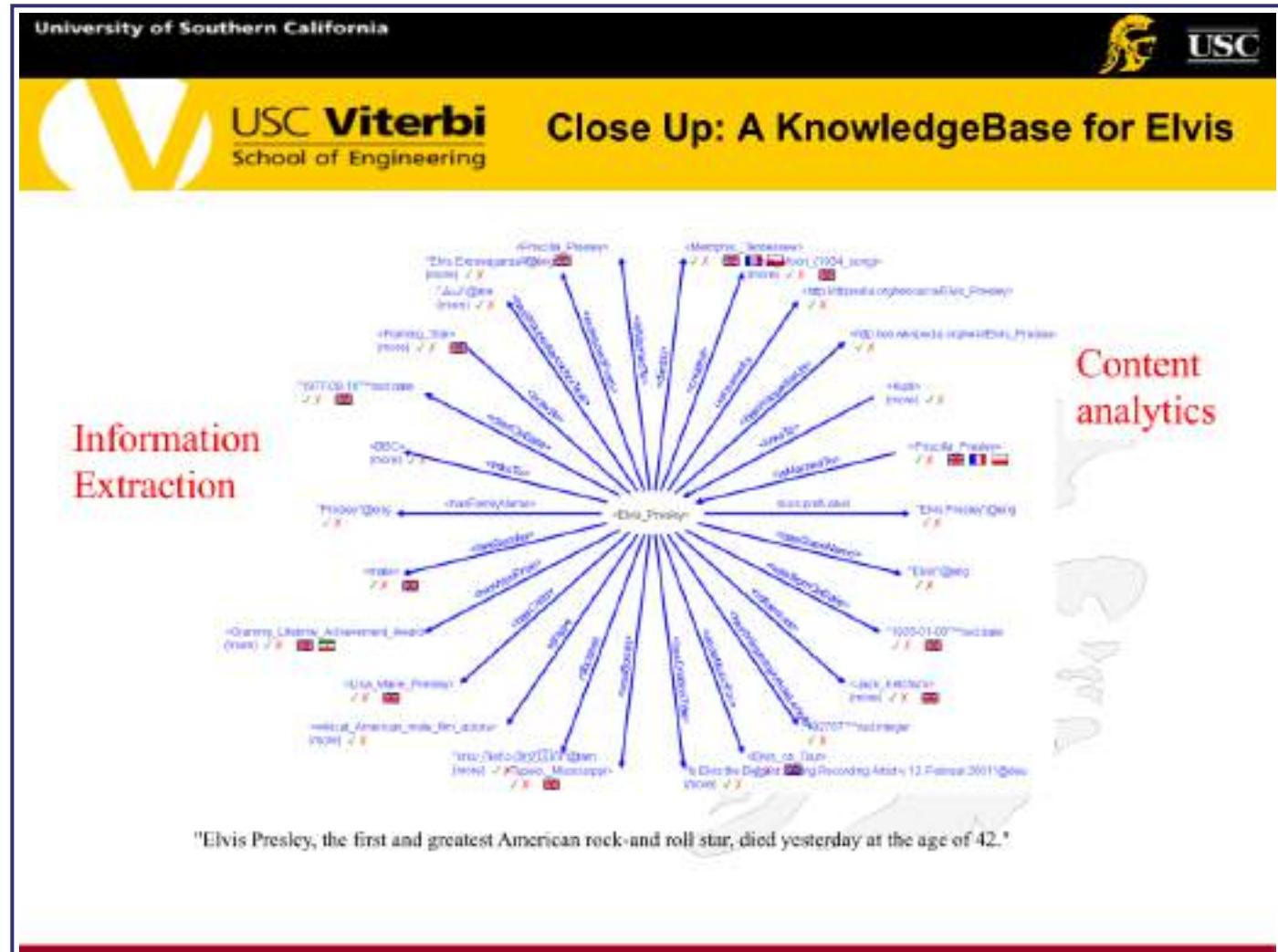
Periods: Cubism, Surrealism, Expressionism, Post-Impressionism. [MORE](#)

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. 1918–1973), Dora Maar (m. 1936–1955)

Artworks [View Collection](#)



Copyright Ellis Horowitz, 2011-2022



• •

University of Southern California  USC

 USC Viterbi
School of Engineering

Types of Knowledge For a KnowledgeBase

<p>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“</p>	<p>taxonomic knowledge factual knowledge temporal knowledge emerging knowledge terminological knowledge</p>
---	--

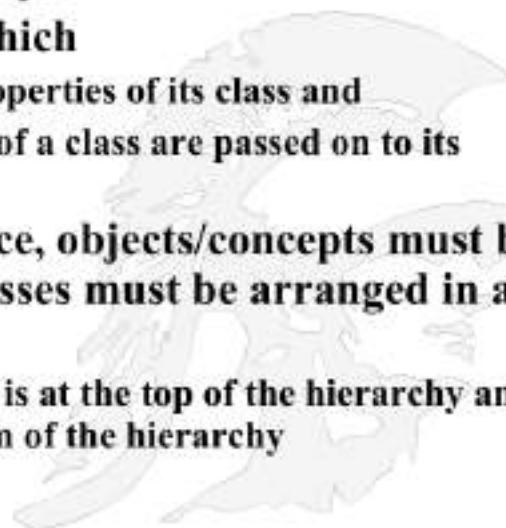
Taxonomies are narrower than ontologies since ontologies include a larger variety of relation types.
Mathematically, a hierarchical **taxonomy** is a tree structure of classifications for a given set of objects
An ontology is a directed, labeled, cyclic graph.

Copyright Ellis Horowitz 2011-2022

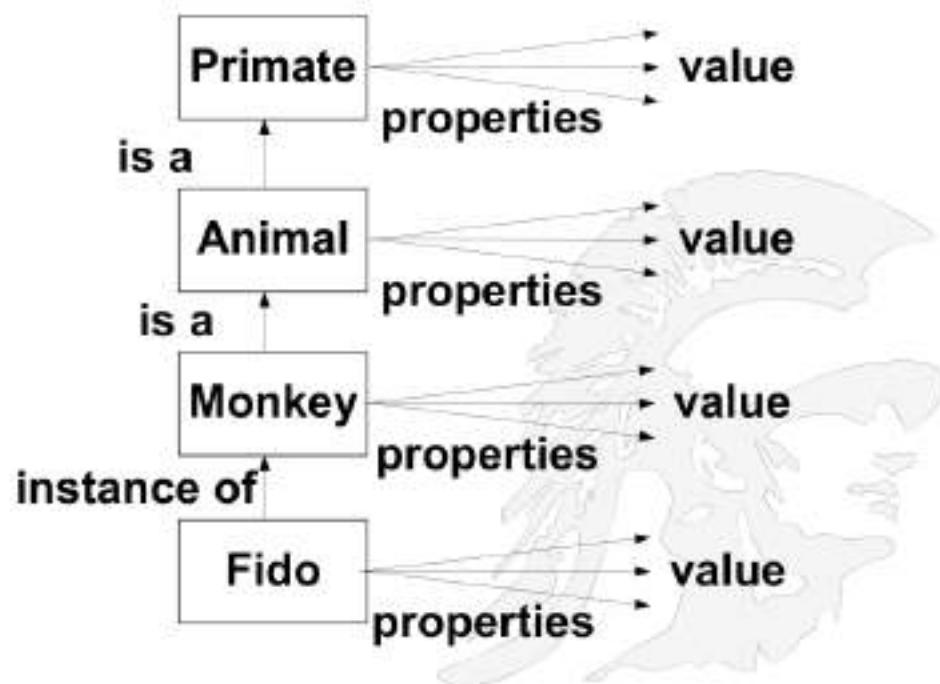
••

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



Copyright Ellis Horowitz, 2011-2022

13

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.



Copyright Ellis Horowitz, 2011-2022

14

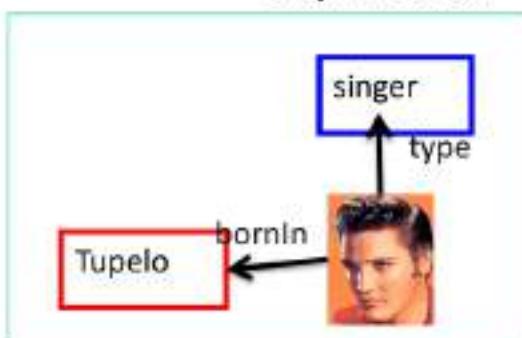
• •

University of Southern California  USC  USC

Different Notations for a KnowledgeBase

- 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:



```
graph TD; Elvis[Elvis] -- type --> Singer[singer]; Elvis -- bornIn --> Tupelo[Tupelo]
```

Triple notation:

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

Logical notation:

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

•

University of Southern California  USC 

RDF Data Model

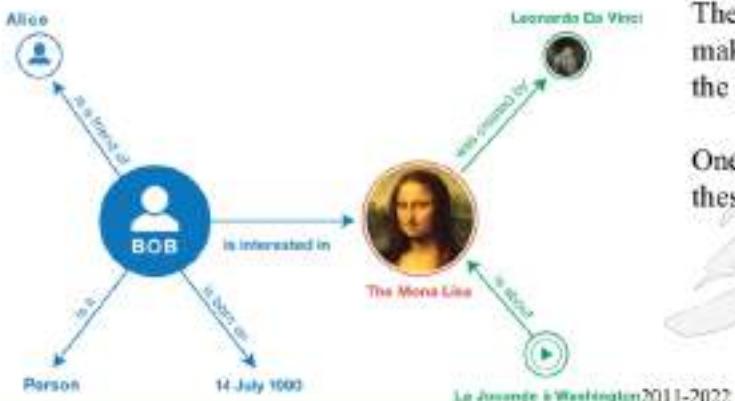
- RDF allows us to make statements about resources.

The format of these statements is:

```
<subject> <predicate> <object>
```

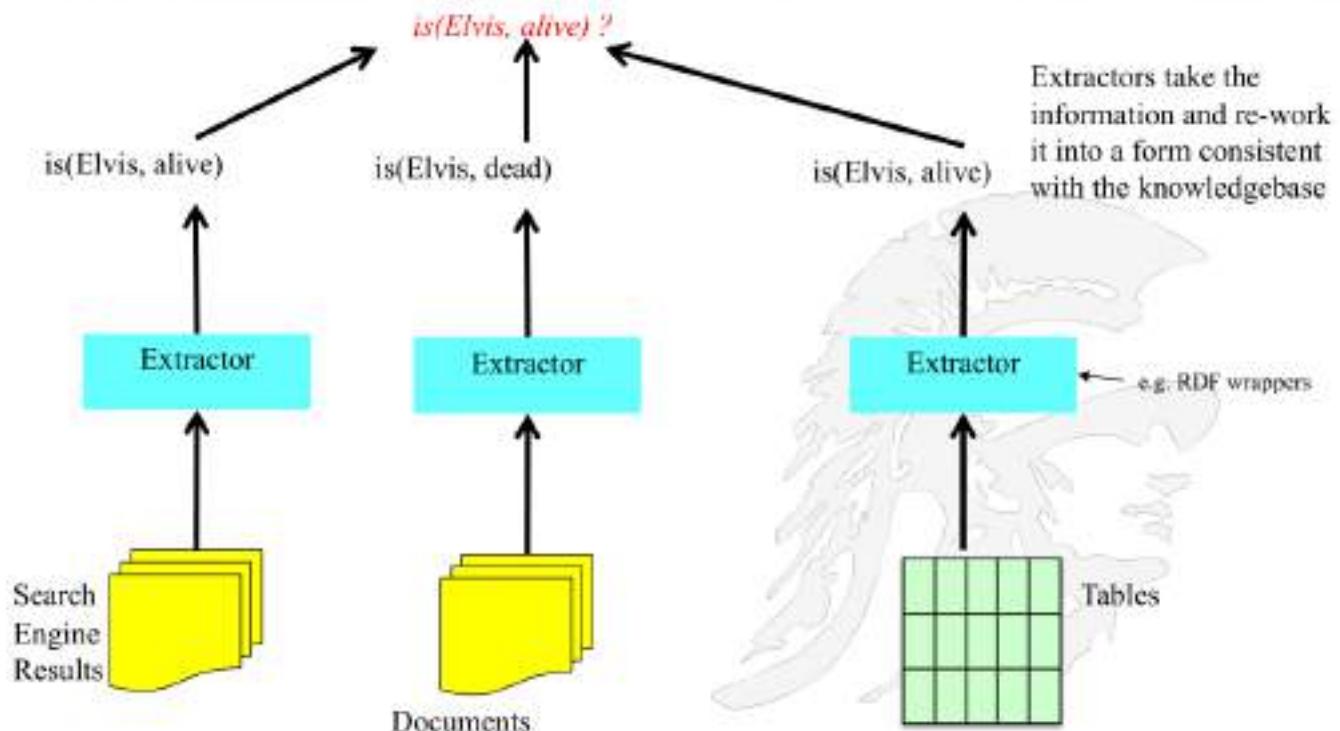
- Some examples

```
<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



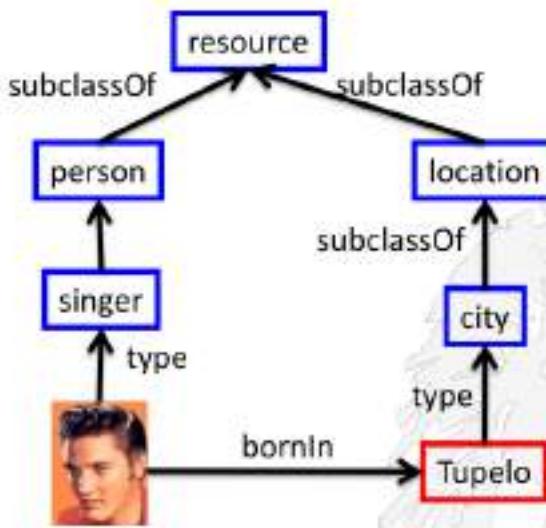
Copyright Ellis Horowitz, 2011-2022

17

••

University of Southern California  USC  USC

KnowledgeBases are Can Be Represented as Labeled MultiGraphs

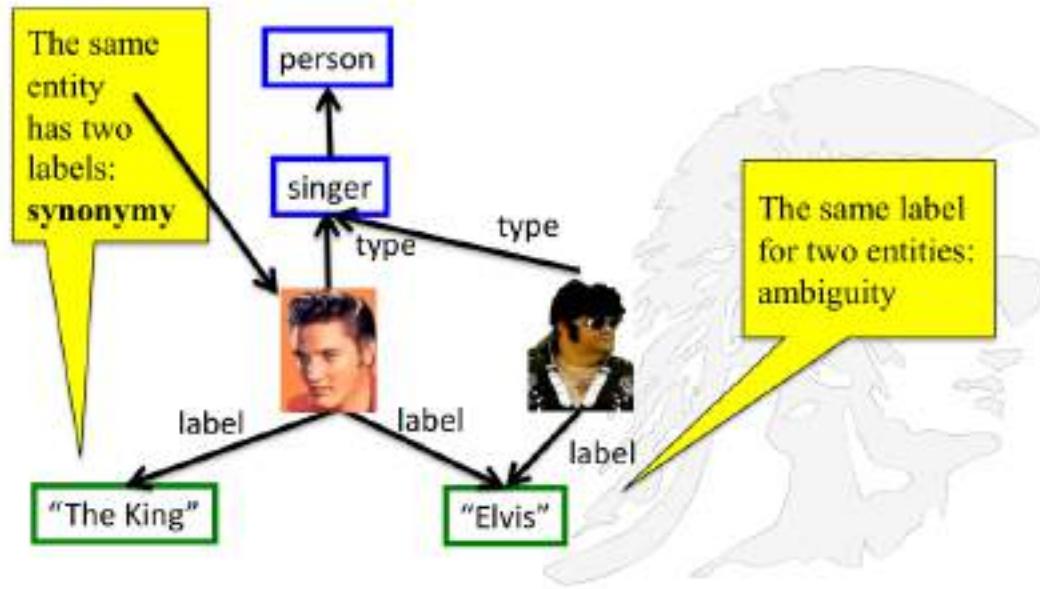


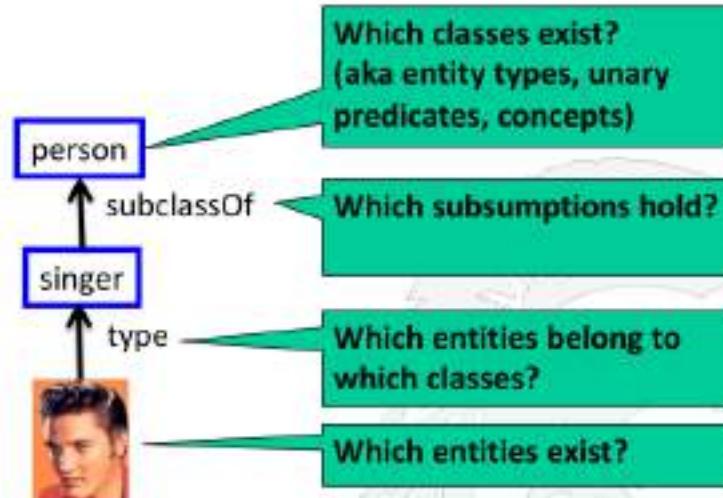
The diagram illustrates a knowledgebase as a directed labeled multigraph. It features three main categories of nodes: Classes/Concepts/Types (blue boxes), Relations/Predicates (black arrows), and Instances/entities (orange/red boxes). The 'resource' node (blue box) is the root. It has two children: 'person' and 'location' (both blue boxes). 'person' has one child: 'singer' (blue box). 'singer' has one child: a portrait of a man (orange box). 'location' has one child: 'city' (blue box). 'city' has one child: 'Tupelo' (red box). There are also two type edges: one from 'singer' to the man's portrait, and another from 'city' to 'Tupelo'. A relation edge connects the man's portrait to 'Tupelo' with the label 'bornIn'.

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

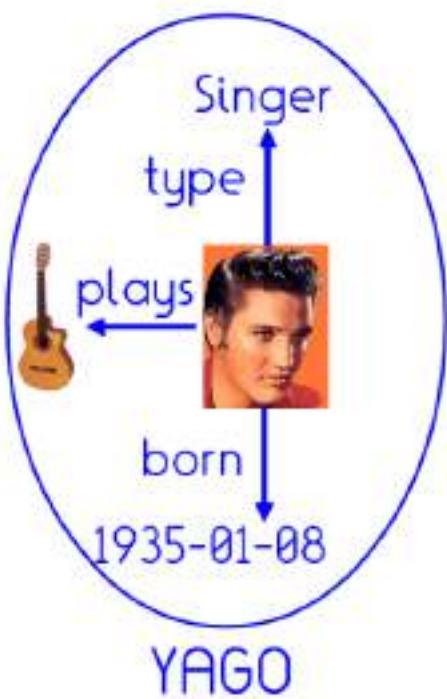
A **multigraph** is a graph which is permitted to have multiple edges that have the same end nodes. Two vertices may be connected by more than one edge

A Single Entity Can Have Different Labels

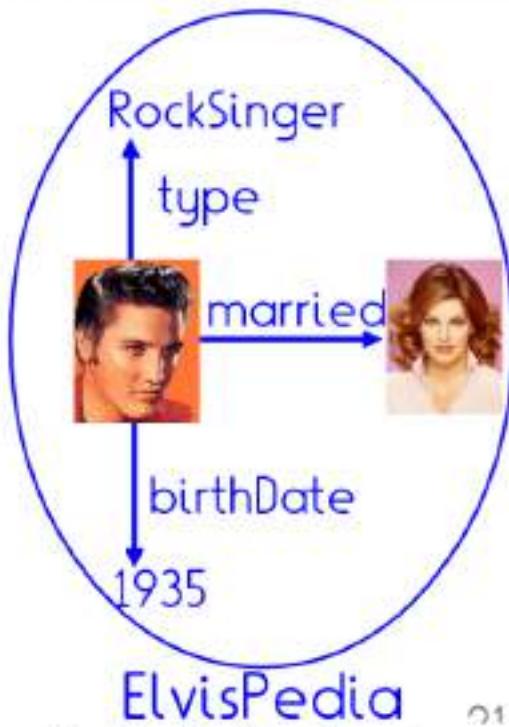


To Build a Knowledgebase
One Must Find Classes and Instances

Two Knowledgebases With Complementary Information

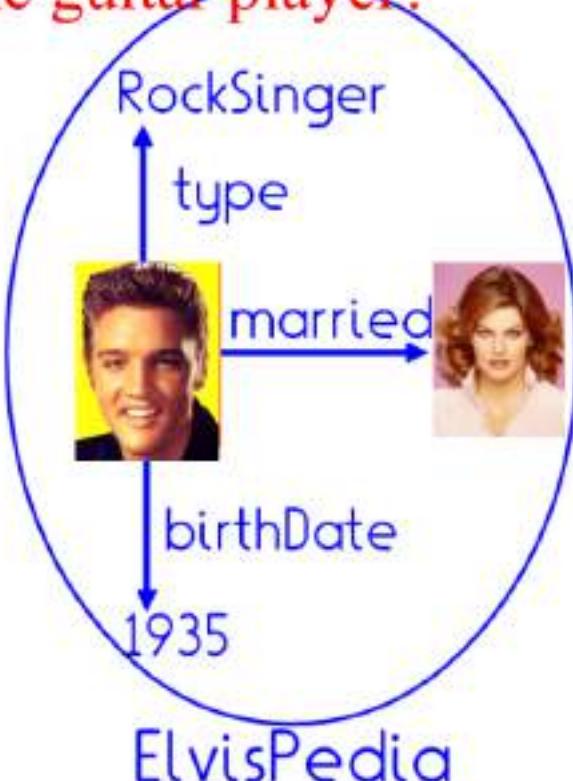
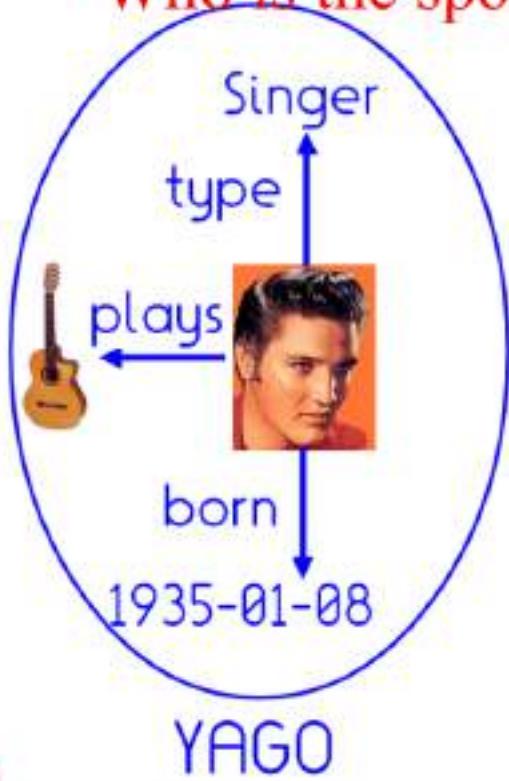


See [https://en.wikipedia.org/wiki/YAGO_\(database\)](https://en.wikipedia.org/wiki/YAGO_(database))

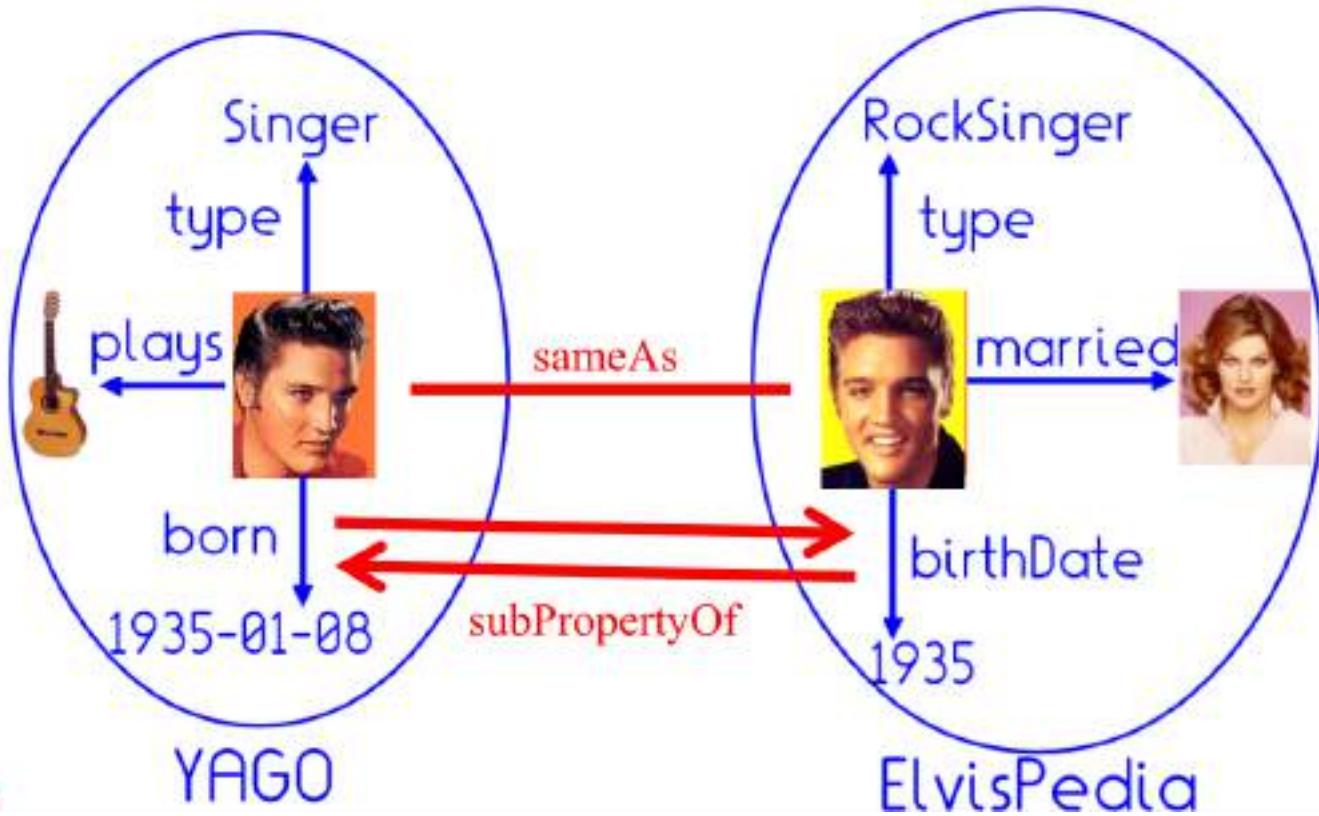


See <https://theelvispedia.com/>

Who is the spouse of the guitar player?



We Need to Match Entities, Classes and Relations

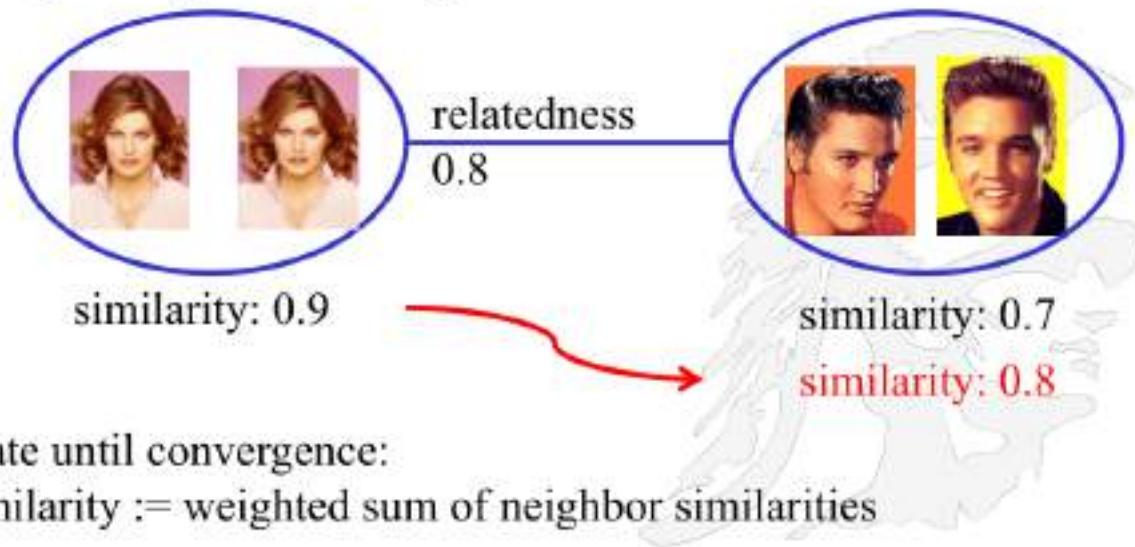


Combining Elements From Different Knowledgebases Means Matching Entities

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.)

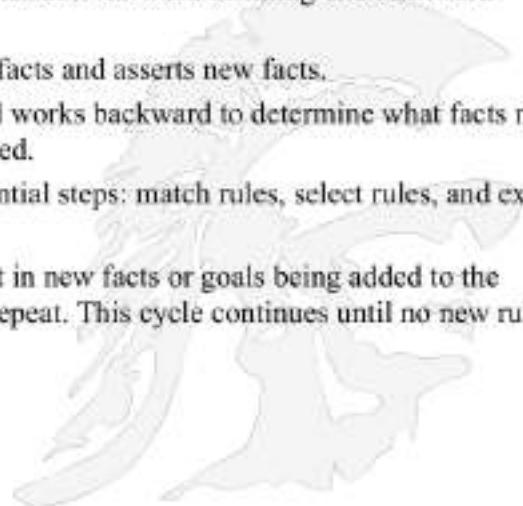
• •

University of Southern California  USC

 USC **Viterbi**
School of Engineering

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

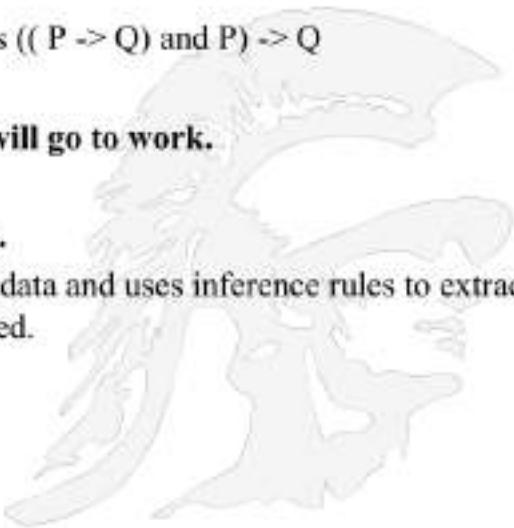


Copyright Ellis Horowitz, 2011-2022

25

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.



Copyright Ellis Horowitz, 2011-2022

26

Here are some sample binary relations with their type signature, e.g.

hasAdvisor: Person × Person

graduatedAt: Person × University

bornOn: Person × Date

Here are instances of the above binary relations

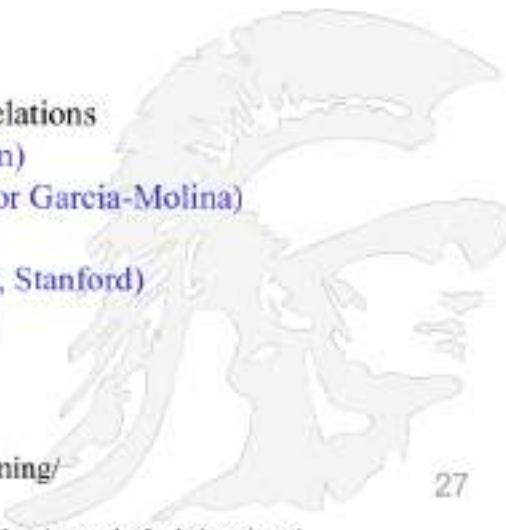
hasAdvisor (JimGray, MikeHarrison)

hasAdvisor (Susan Davidson, Hector Garcia-Molina)

graduatedAt (JimGray, Berkeley)

graduatedAt (HectorGarcia-Molina, Stanford)

bornOn (JohnLennon, 9-Oct-1940)



27

For more examples see

<https://www.tutorialandexample.com/forward-chaining/>

And

<https://www.javatpoint.com/forward-chaining-and-backward-chaining-in-ai>

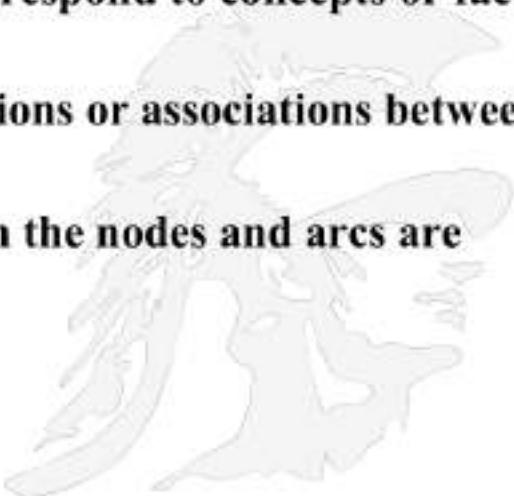
••

University of Southern California  USC

 USC **Viterbi**
School of Engineering

Semantic Network

- **A semantic network is a knowledge representation scheme that represents knowledge as a graph.**
 - The nodes of the graph correspond to concepts or facts and
 - the arcs correspond to relations or associations between concepts.
 - In a semantic network both the nodes and arcs are labeled

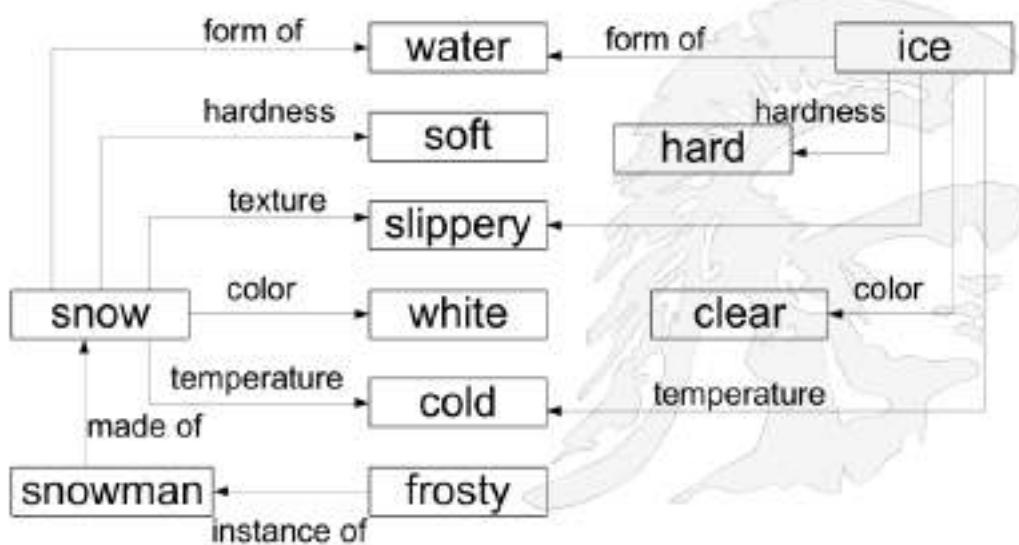


Copyright Ellis Horowitz, 2011-2022

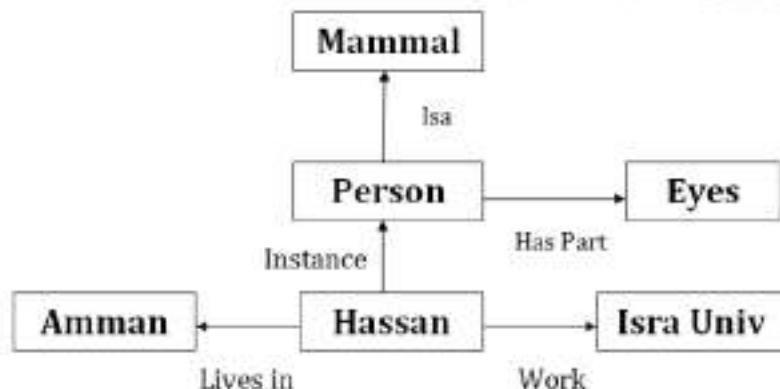
29

Semantic Network

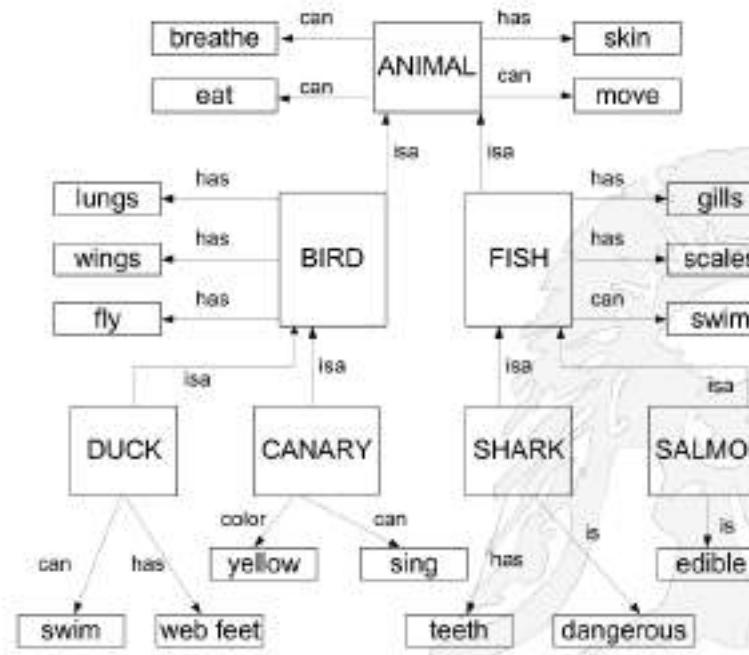
- A semantic network that defines the properties of snow and ice
- The concept snowman inherits all the properties of snow
- The concept of ice and snow share a number of properties



- The most important link in a semantic network is the is-a link
- A semantic network can organize knowledge in a hierarchy by using the is-a link such that the lower nodes inherit the properties of the higher nodes



The is-a relationship in a Semantic Network



Copyright Ellis Horowitz, 2011-2022

32

• •



The logo for the University of Southern California (USC) Viterbi School of Engineering. It features a yellow banner across the top with the text "University of Southern California" on the left and the USC Trojan logo with the letters "USC" on the right. Below the banner, the "V" logo of the Viterbi School of Engineering is displayed, followed by the text "USC Viterbi" and "School of Engineering".

WordNet



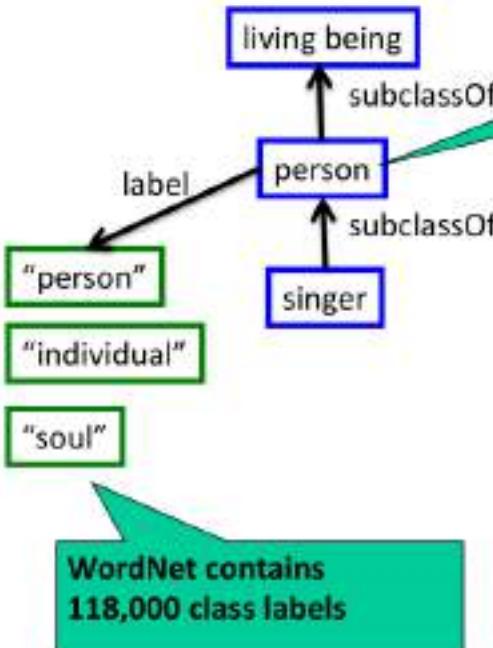
A faint watermark of a Trojan head wearing a helmet, positioned in the lower right quadrant of the slide.

Copyright Ellis Horowitz 2011-2022

• •

University of Southern California 

WordNet is a Lexical DataBase with many Classes, Subclasses, and Superclasses



```
graph TD; living_being["living being"] -- subclassOf --> person["person"]; person -- subclassOf --> singer["singer"]; person -- label --> person_label["\"person\""]; person -- label --> individual_label["\"individual\""]; person -- label --> soul_label["\"soul\""]
```

WordNet contains 82,000 classes

WordNet contains thousands of subclassOf relationships

WordNet contains 118,000 class labels

- 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

• •

University of Southern California  USC

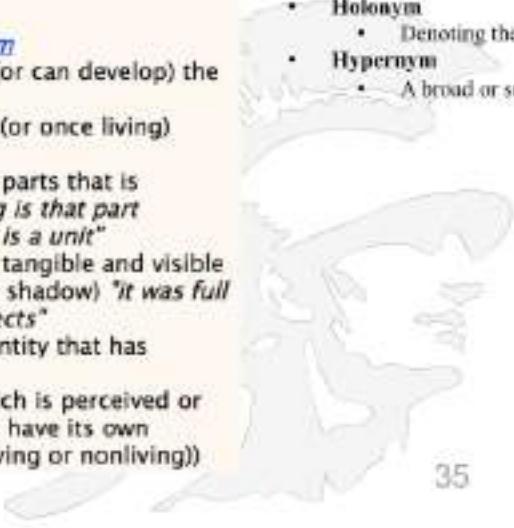
 USC Viterbi
School of Engineering

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



35



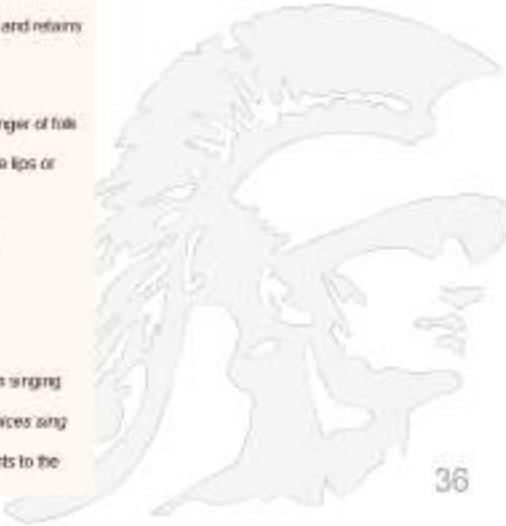
University of Southern California



USC Viterbi WordNet Example: Subclass of Singer

School of Engineering

- E. (n) **singer, vocalist, vocalizer, vocaliser** (a person who sings)
 - direct hyponym / full hyponym
 - S. (m) **alto** (a singer whose voice lies in the alto clef)
 - S. (m) **bassoon, bassooner** (a male singer)
 - S. (m) **bass, basso** (an adult male singer with the lowest voice)
 - S. (m) **canary** (a female singer)
 - S. (m) **carol, caroler** (a singer of carols)
 - S. (m) **castrato** (a male singer who was castrated before puberty and retains a soprano or alto voice)
 - S. (m) **choirie** (a singer in a choir)
 - S. (m) **contralto** (a woman singer having a contralto voice)
 - S. (m) **crooner, balladeer** (a singer of popular ballads)
 - S. (m) **folk singer, folksinger, minstrel, post-singer, troubadour** (a singer of folk songs)
 - S. (m) **hummer** (a singer who produces a tune without opening the lips or forming words)
 - S. (m) **loller singer** (a singer of lodes)
 - S. (m) **madrigalist** (a singer of madrigals)
 - S. (m) **opera star, operatic star** (singer of lead role in an opera)
 - S. (m) **rapper** (someone who performs rap music)
 - S. (m) **rock star** (a famous singer of rock music)
 - S. (m) **sonatist** (a person who sings)
 - S. (m) **soprano** (a female singer)
 - S. (m) **tenor** (an adult male with a tenor voice)
 - S. (m) **thrush** (a woman who sings popular songs)
 - S. (m) **torch singer** (a singer (usually a woman) who specializes in singing torch songs)
 - S. (m) **voces** ((metonymy) a singer) "He wanted to hear trained voices sing it"
 - S. (m) **writer** (a singer; usually a singer who adds embellishments to the song)



36

• •

University of Southern California  USC

WordNet Example: Instances But Very Few

• S: (m) [singer](#), [vocalist](#), [vocalizer](#), [vocals](#) (a person who sings)

- [direct hyperonym](#) / [full hyperonym](#)
- [has instance](#)
 - S: (m) [Bailey, Pearl Bailey](#), [Pearl Mae Bailey](#) (United States singer (1918-1990))
 - S: (m) [Cash, Johnny Cash](#), [John Cash](#) (United States country music singer and songwriter (1932-2003))
 - S: (m) [Chevalier, Maurice Chevalier](#) (French actor and cabaret-singer (1888-1972))
 - S: (m) [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: (m) [Dylan, Bob Dylan](#) (United States songwriter noted for his protest songs (born in 1941))
 - S: (m) [Fitzgerald, Ella Fitzgerald](#) (United States scat singer (1917-1996))
 - S: (m) [Garland, Judy Garland](#) (United States singer and film actress (1922-1969))
 - S: (m) [Home, Lena Home](#), [Lena Calhoun Home](#) (United States singer and actress (born in 1917))
 - S: (m) [Iglesias, Julio Iglesias](#) (Spanish singer noted for his ballads and love songs (born in 1943))
 - S: (m) [Jackson, Mahalia Jackson](#) (United States singer who did much to popularize gospel music (1911-1972))
 - S: (m) [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 ✕

37

• •

University of Southern California

USC Viterbi
School of Engineering

Wikipedia

Copyright Ellis Horowitz 2011-2022

• •

University of Southern California 

 USC **Viterbi**
School of Engineering

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)



••

University of Southern California  USC

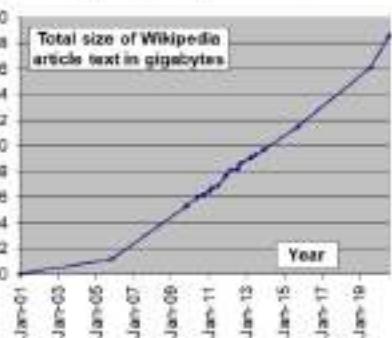
USC Viterbi
School of Engineering

Wikipedia's Scale

- As of April, 2019 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



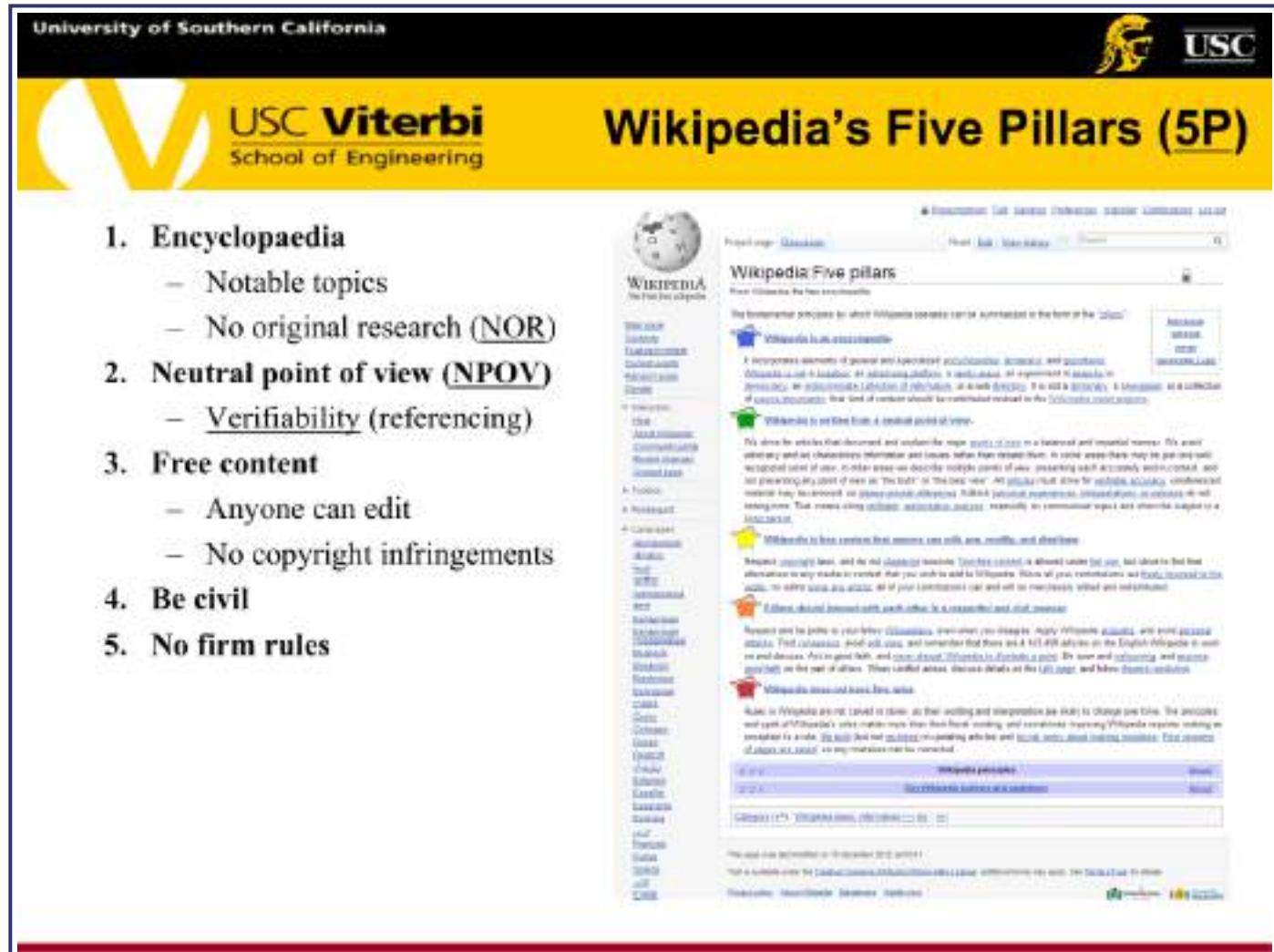
Jimmy Wales, Founder



Total size of Wikipedia article text in gigabytes

Year	Total size of Wikipedia article text in gigabytes
Jan-01	0.5
Jan-03	0.8
Jan-05	1.2
Jan-07	2.5
Jan-09	4.5
Jan-11	7.5
Jan-13	11.5
Jan-15	15.5
Jan-17	19.5
Jan-19	23.5

By Mikael Häggström



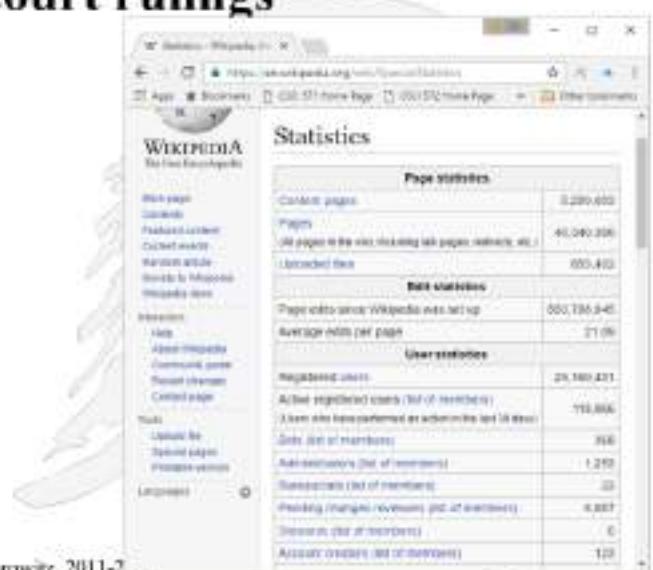
••

University of Southern California  USC 

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>



Page statistics	
Committed pages	3,289,452
Pages	46,249,336
All pages in the sites (including talk pages, redirects, etc.)	655,452
Unprotected pages	

Edit statistics	
Page edits since Wikipedia was set up	680,736,446
Average edits per page	31.0%

User statistics	
Registered users	29,169,471
Active registered users (last 7 days, excluding bot accounts)	719,866
Last user to register	2006-07-10 12:00:00 UTC
Users (not of members)	366
Administrators (not of members)	1,259
Administrators (not of members)	13
Wikipedia (not of members)	4,827
Wikinews (not of members)	6
Wikibooks (not of members)	123

Copyright Ellis Horowitz, 2011-2012

• •

University of Southern California  **USC**

USC Viterbi
School of Engineering

Wikipedia is a Rich Source of Instances

Steve Jobs

From Wikipedia, the free encyclopedia

For the biography, see Steve Jobs (biography).

Steven Paul Jobs ([Vegan](#); February 24, 1955 – October 5, 2011)^[10] was an American business magnate and inventor widely recognized as a charismatic pioneer of the personal computer revolution.^[10] 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 2005, 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" Markula 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 Rothmiller 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.^[11] Jobs was named Apple advisor in 1998, interim CEO in 1997, and CEO from 2000 until his resignation. He oversaw the development of the iMac, iMacs, iPod, iPhone, and iPad and the company's Apple Retail Stores.^[12] In 1996, he acquired the computer graphics division of Lucasfilm Ltd., which was spun off as Pixar Animation Studios.^[13] He was credited in Toy Story (1995) as an executive producer. He remained CEO and majority shareholder at 58.1 percent until its acquisition by The Walt Disney Company in 2005^[14], making Jobs Disney's largest individual shareholder at seven percent and a member of Disney's Board of Directors.^{[15][16]}

In 2003, Jobs was diagnosed with a pancreatic 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.^[17] 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 founders → 

Jimmy Wales **Larry Sanger**

Steve Jobs



Jobs holding a white iPhone 4 at Worldwide Developers Conference 2010

Birth	Steven Paul Jobs February 24, 1955 ^[10]
Death	October 5, 2011 (aged 56) ^[11] Palo Alto, California, U.S.
Nationality	American
Alma mater	Reed College (undergraduate)

43

• •

University of Southern California  USC

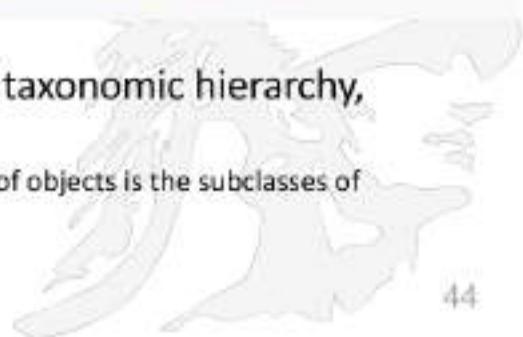
 USC Viterbi
School of Engineering

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



44

• •

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;

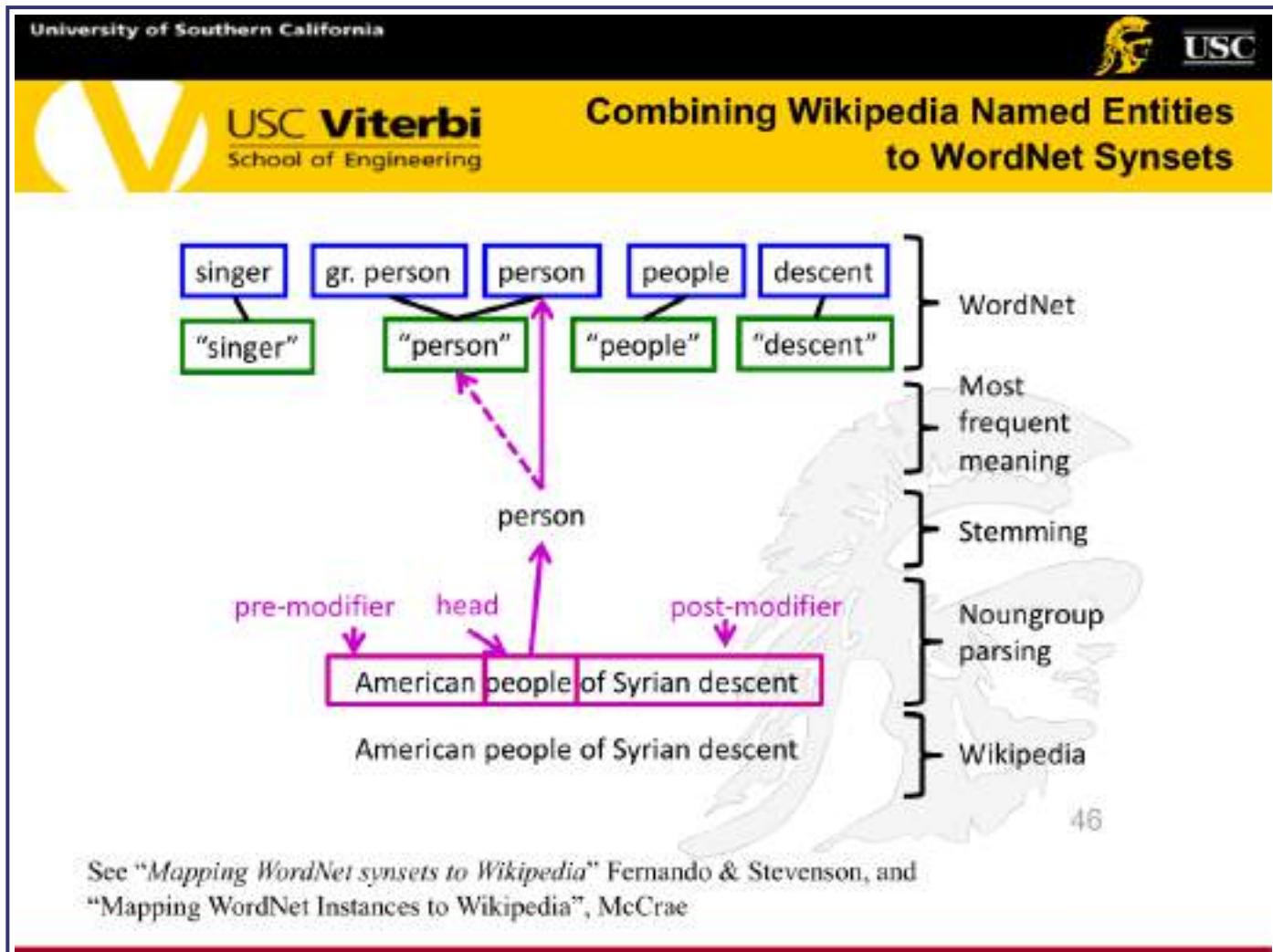
Types of special pages

- *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, ...

Copyright Ellis Horowitz, 2011-2022

• •



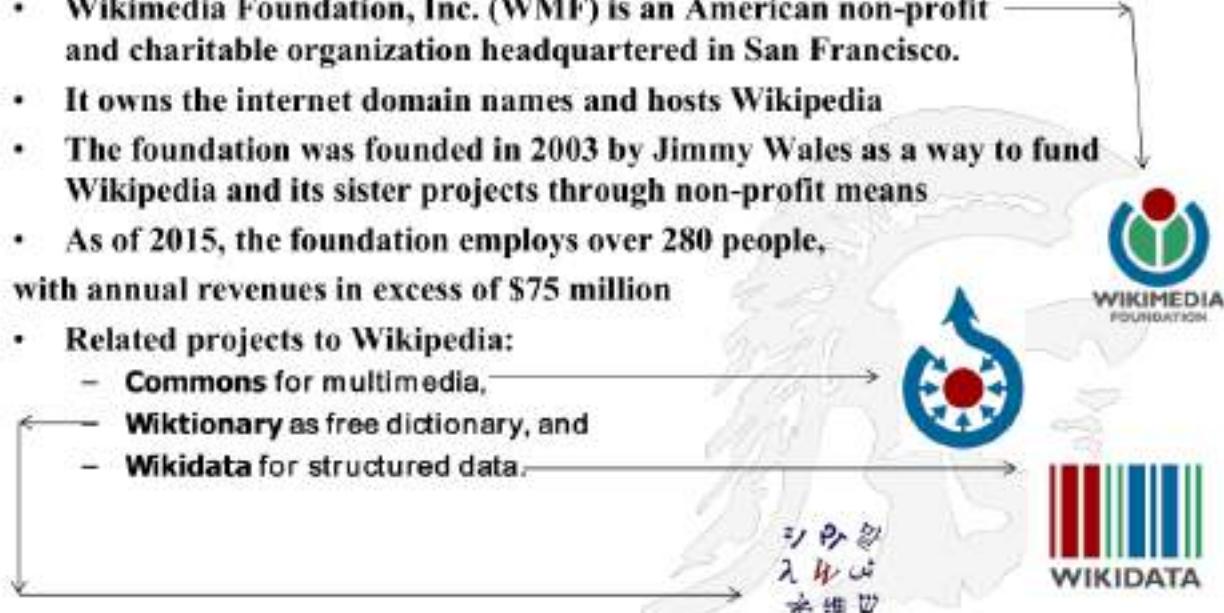
••

University of Southern California  USC

 USC **Viterbi**
School of Engineering

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.



Wikimedia Foundation

Wikipedia

Wiktionary
The free dictionary

Wikidata

Copyright Ellis Horowitz, 2011-2022

47

••



University of Southern California
USC Viterbi
School of Engineering

WikiData

- **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**



Welcome to Wikidata

The Free Knowledge Base with 10,787,279 data items that anyone can edit.

Individual · Project Chat · Community Portal · Help

Want to help Wikidata? Fix some missing messages!

48

• •



University of Southern California
USC Viterbi
School of Engineering

Wikidata Multilingual Coverage

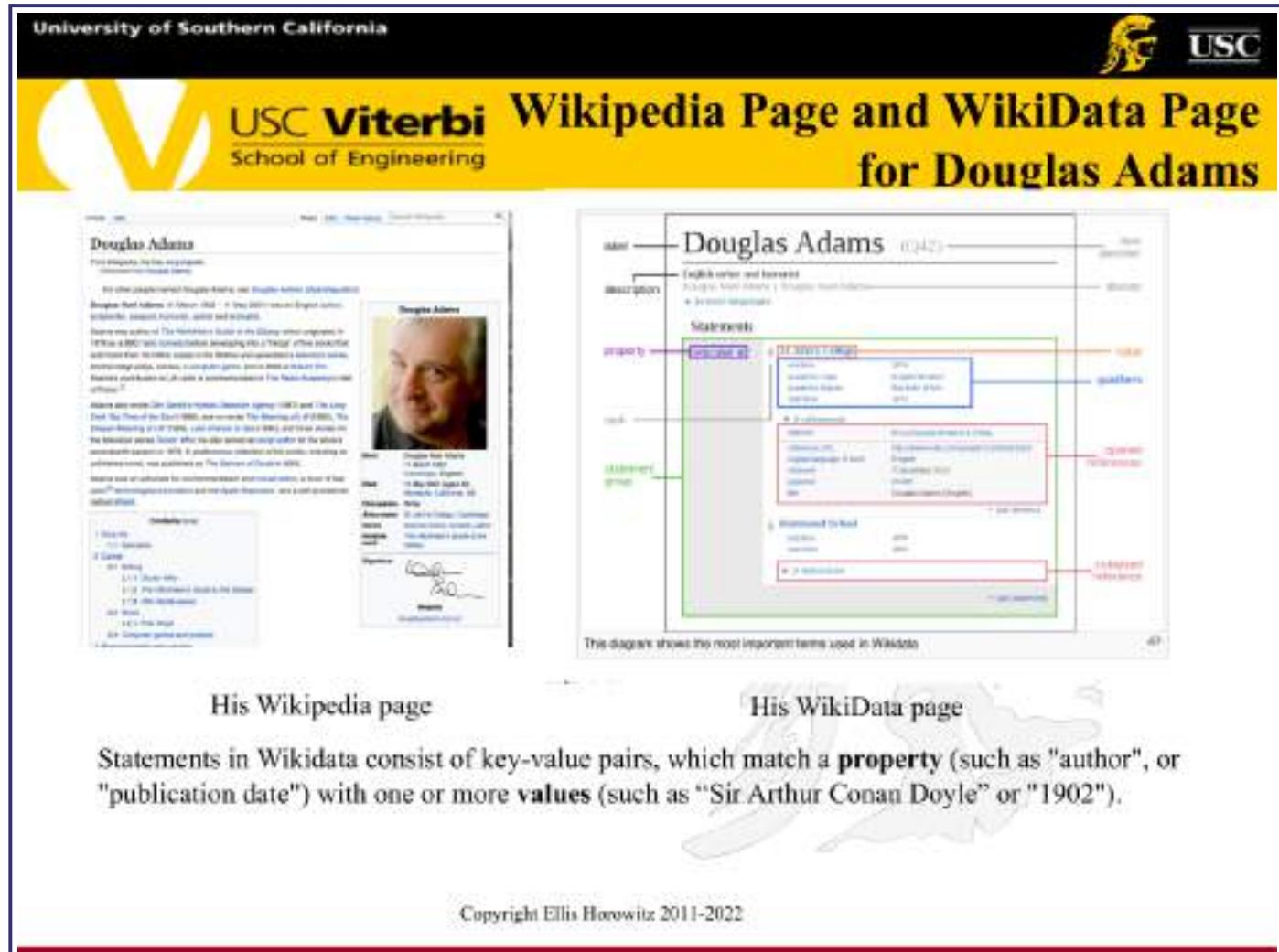
→ The challenge: Wikipedia has many named entities that occur in numerous articles

- 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



• •

University of Southern California  USC

 USC **Viterbi**
School of Engineering

Google's Knowledge Graph



Copyright Ellis Horowitz 2011-2022

• •



University of Southern California 

Google Knowledge Graph

Google Inside Search

The Knowledge Graph

See it in action

- Introduced in 2012 with the Hummingbird update, see <https://searchengineland.com/google-hummingbird-172816>
- Powered in part by Firebase
- KnowledgeGraph 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

<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=mmQI6VGvX-c>

Copyright Ellis Horowitz, 2011-2022

52

• •

University of Southern California  **USC**

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



mausoleum

musician

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

Copyright Ellis Horowitz, 2011-2022

53

• •

University of Southern California  USC

USC Viterbi
School of Engineering

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



Copyright Ellis Horowitz 2011-2022

• •

University of Southern California  USC

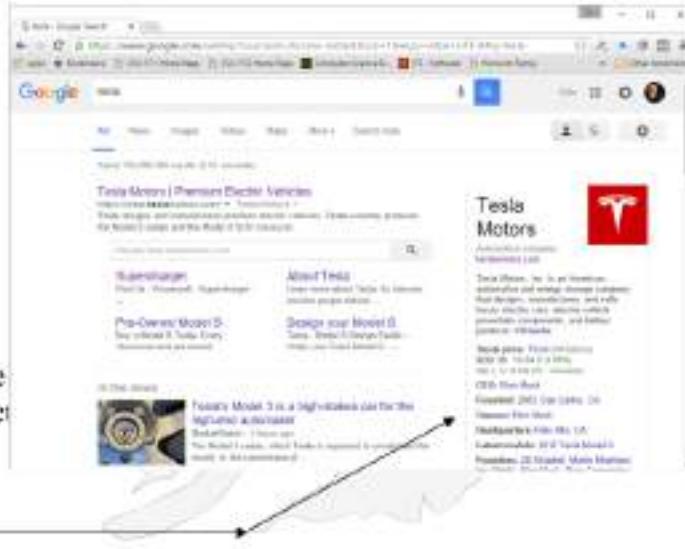
USC Viterbi
School of Engineering

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

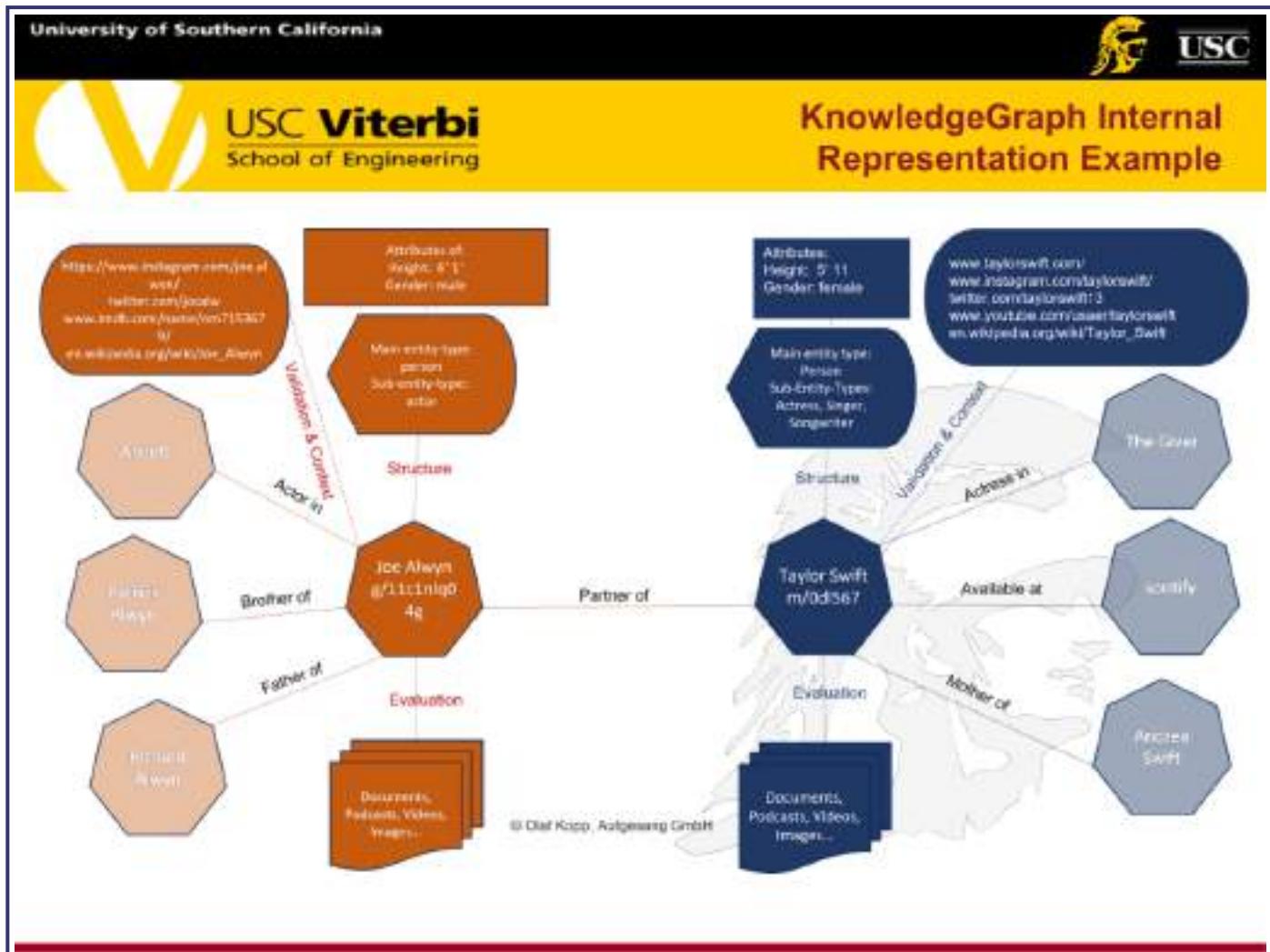
3. To provide the best summary
 the knowledge graph exploits the relationships among the entities
 e.g. the query "Tesla"

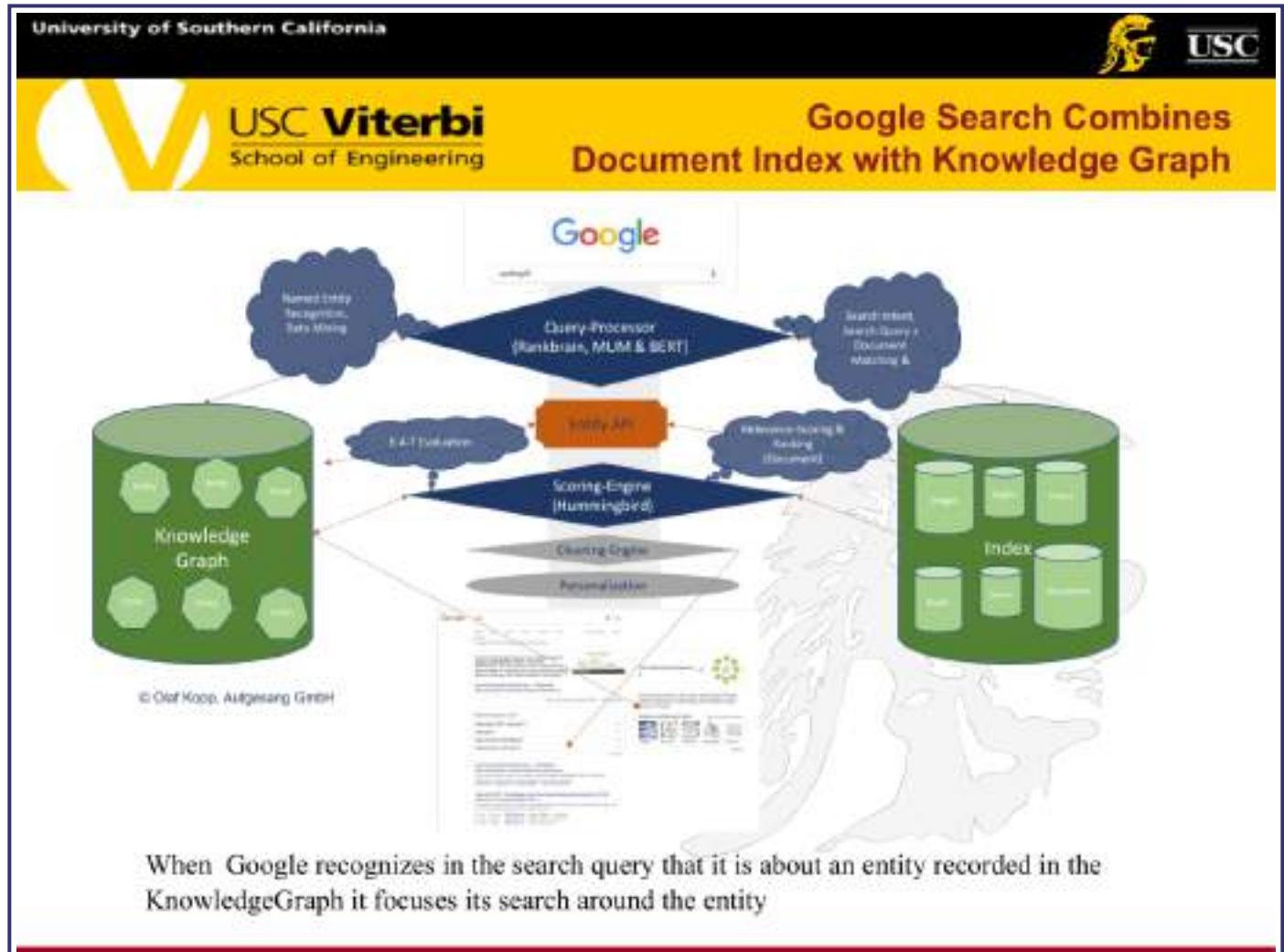
The knowledge graph allows Google to summarize relevant content around that topic, including key facts you're likely to need for that particular thing. E.g.

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.

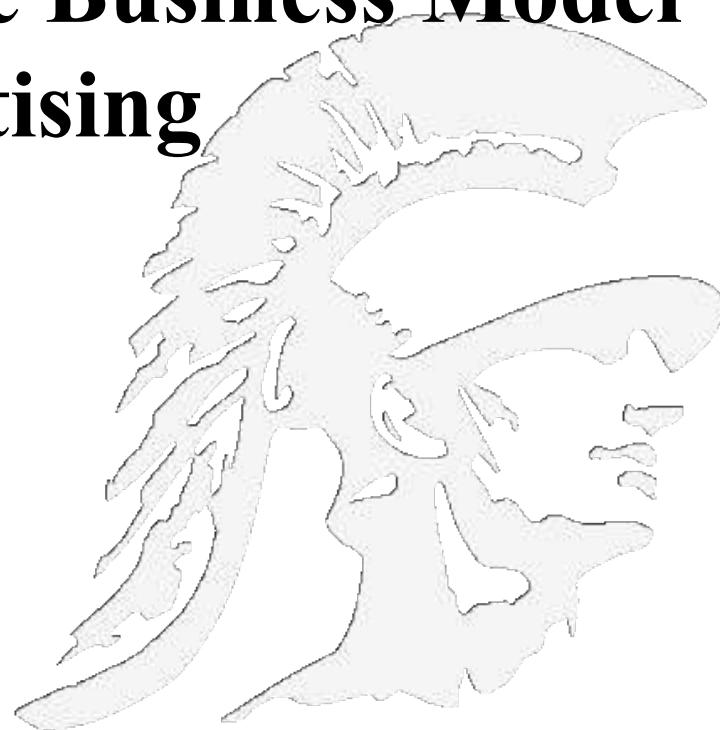


Copyright Ellis Horowitz 2011-2022



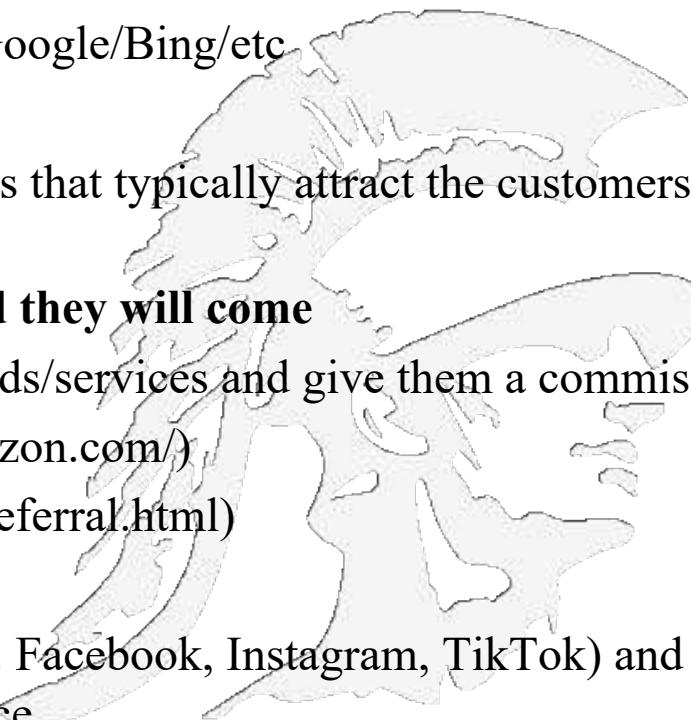


The Search Engine Business Model Advertising



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

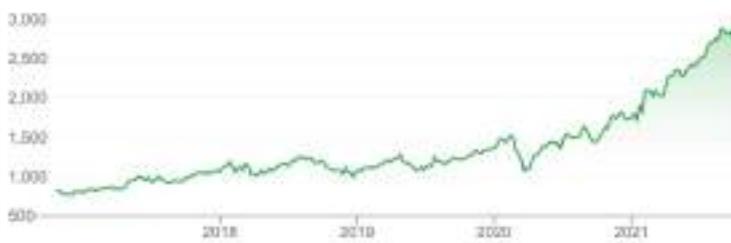
Market Summary > Alphabet Inc Class A

NASDAQ: GOOGL

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

Closed: Oct 15, 4:58 PM EDT · Disclaimer
After hours 2,027.39 +0.63 (0.032%)

1 day 5 days 1 month 6 months YTD 1 year **5 years** Max



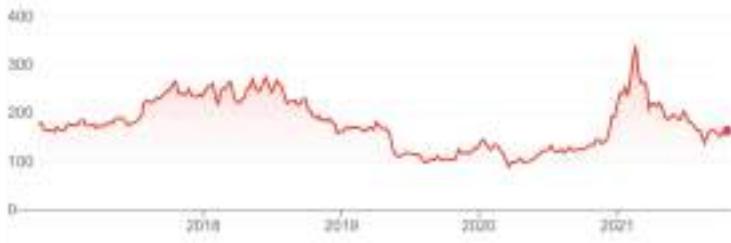
Market Summary > Baidu Inc

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 (0.0041%)

1 day 5 days 1 month 6 months YTD 1 year **5 years** Max



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

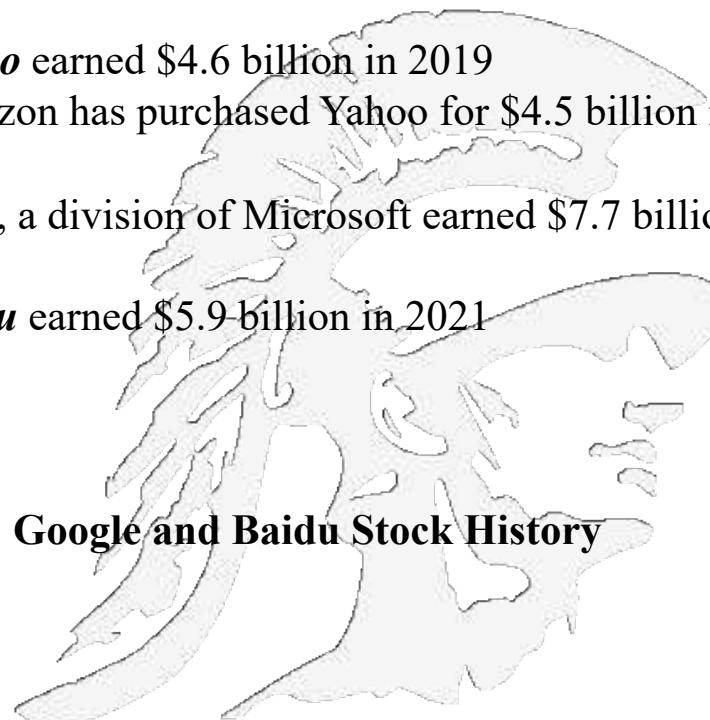
Alibaba earned \$109 billion in 2021

Yahoo earned \$4.6 billion in 2019
(Verizon has purchased Yahoo for \$4.5 billion in 2015)

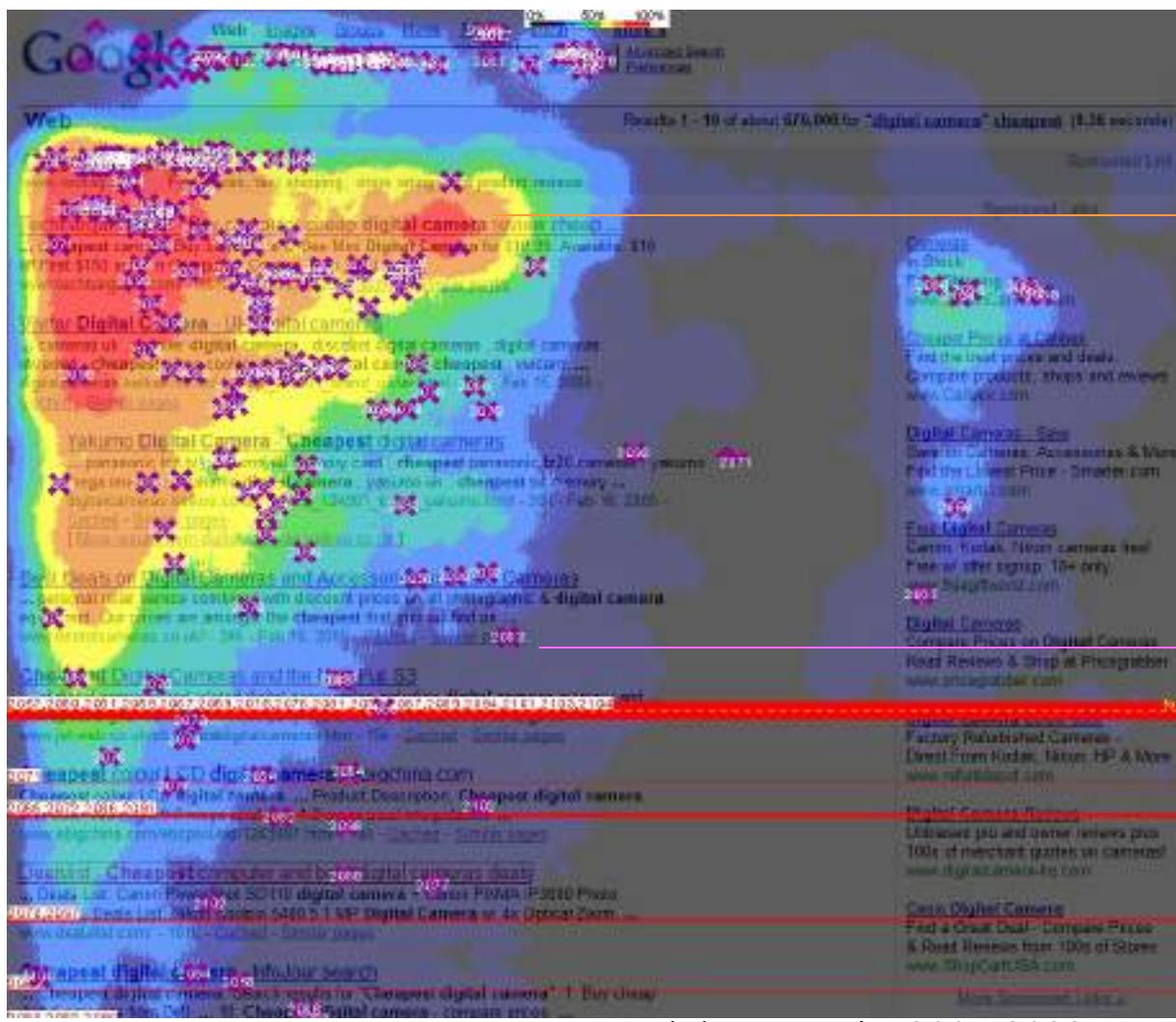
Bing, a division of Microsoft earned \$7.7 billion in 2021

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

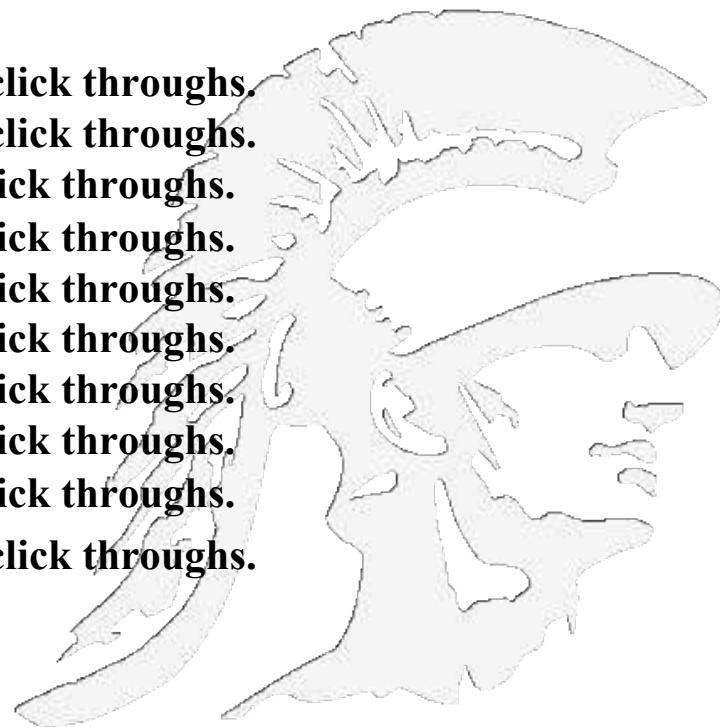
Eye pupil
hotspots

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

X marks
the clicks

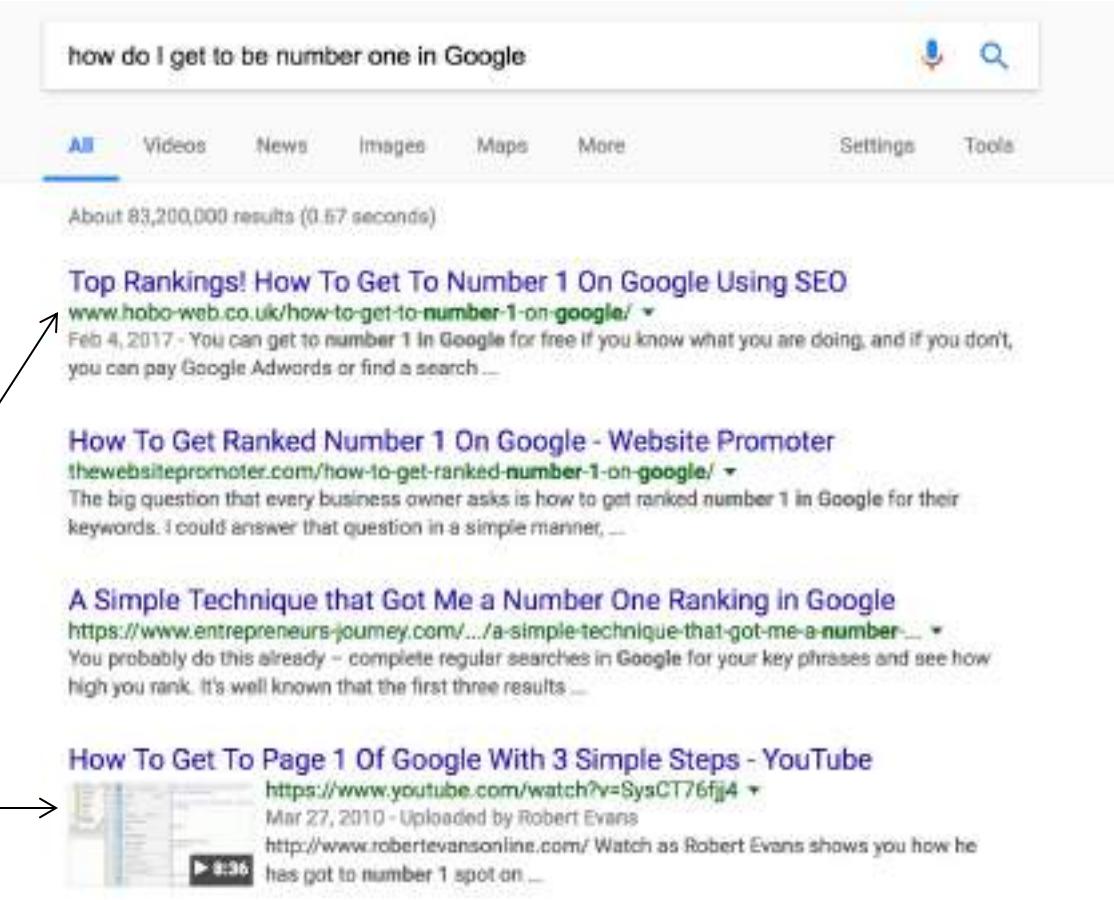
Estimating the Value of Rank in a Sponsored List

- 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 (0.57 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
thewebsitemarketing.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 ...

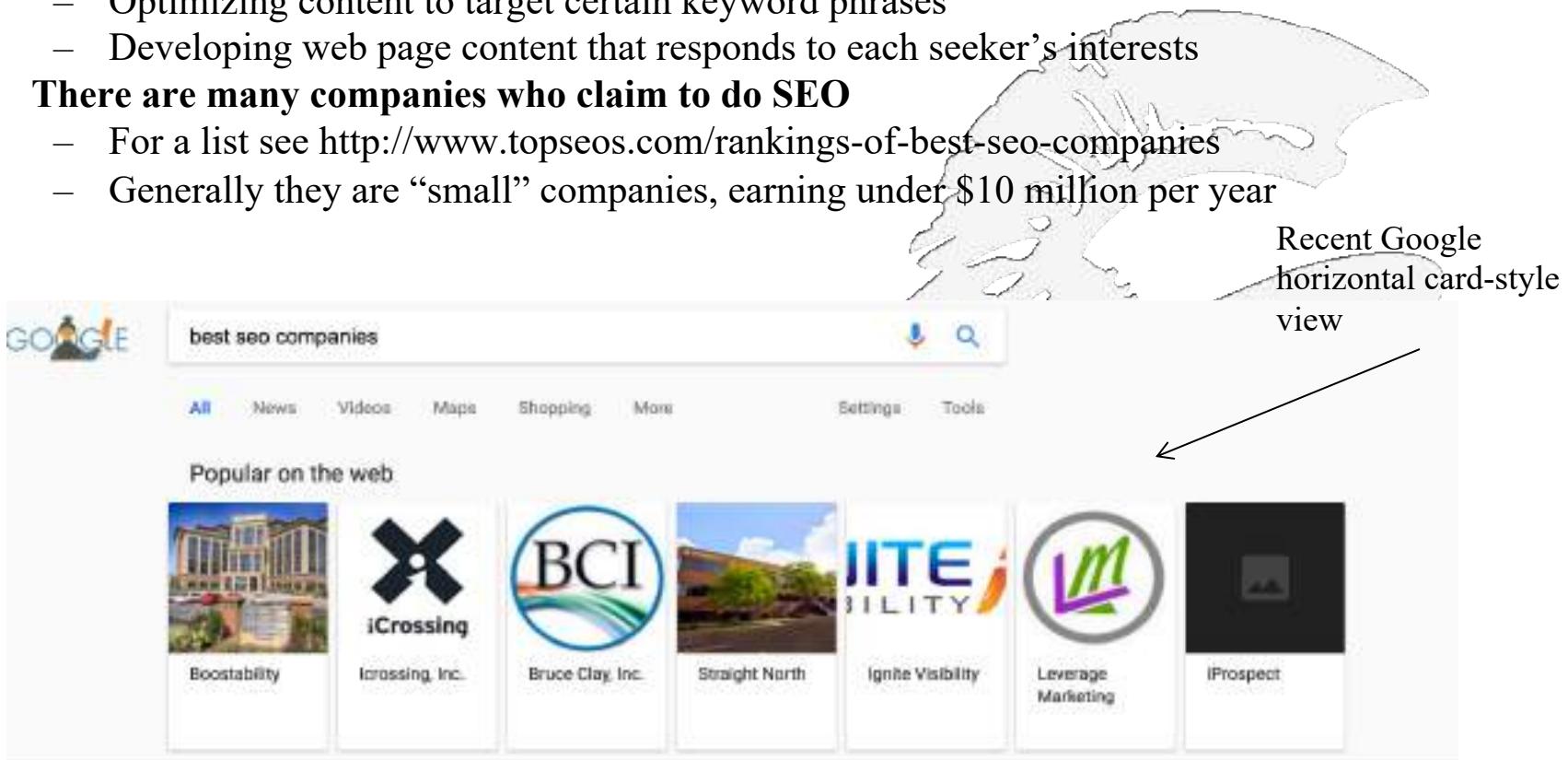
How To Get To Page 1 Of Google With 3 Simple Steps - YouTube
<https://www.youtube.com/watch?v=SysCT76fj4> ▾
Mar 27, 2010 - Uploaded by Robert Evans
<http://www.robertevansonline.com/> Watch as Robert Evans shows you how he has got to number 1 spot on ...

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 page features the classic Google header with the logo, a search bar containing the query, and standard navigation links for All, News, Videos, Maps, Shopping, More, Settings, and Tools. Below the header, a section titled "Popular on the web" displays seven cards, each representing a different SEO company. From left to right, the cards are: Boostability (with a building image), iCrossing (with a stylized 'X' logo), Bruce Clay, Inc. (with a blue circular logo), Straight North (with a building image), Ignite Visibility (with a colorful logo), Leverage Marketing (with a green 'LM' logo), and iProspect (with a dark square logo). An arrow points from the text "Recent Google horizontal card-style view" to the "Popular on the web" section of the search results.

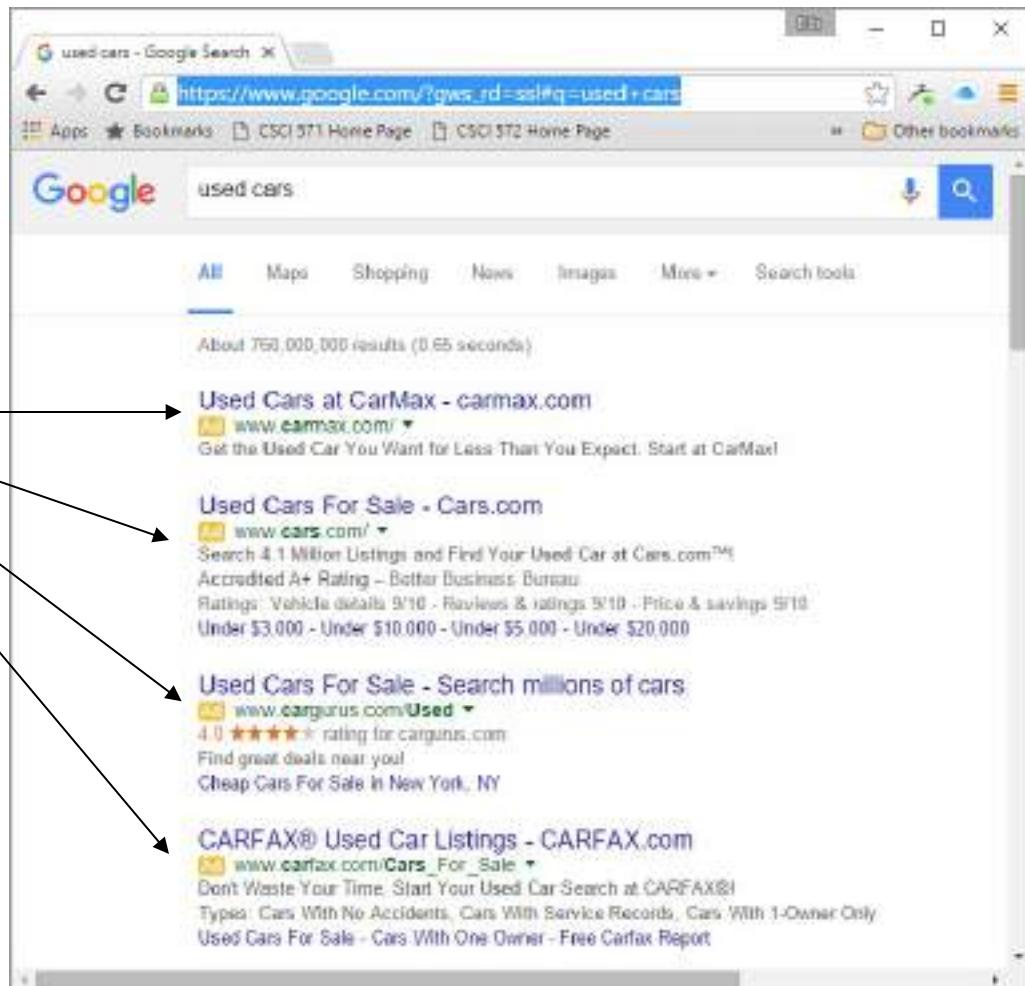
Recent Google horizontal card-style view

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



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



Campaigns Begin By Choosing Keywords

Google Keyword Estimator

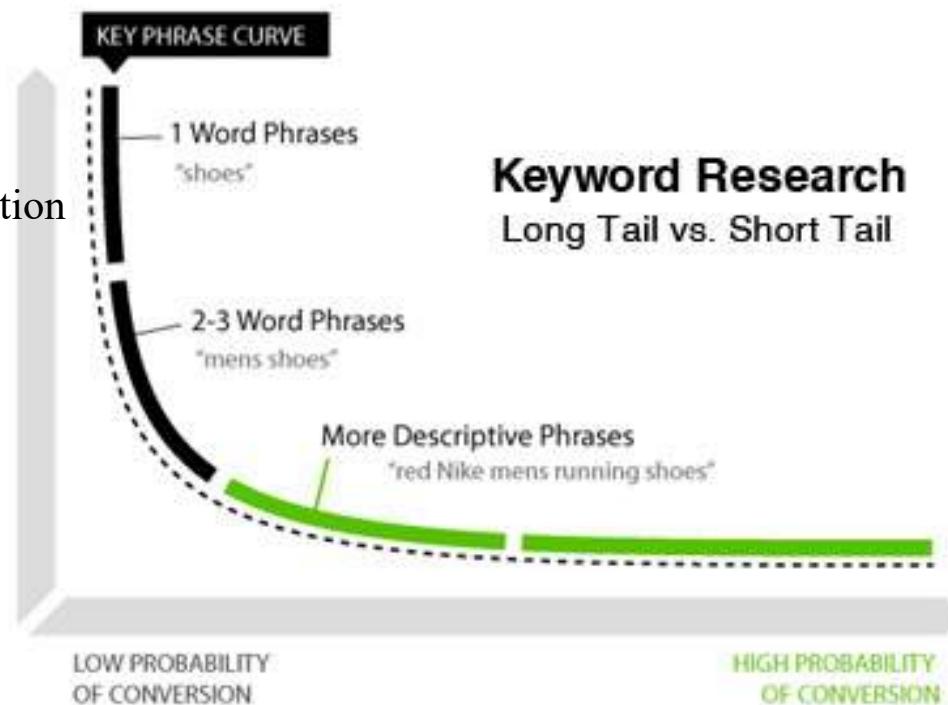
For specific keywords Google provides the approximate cost per click and the resulting position in the list, and other stats

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			Not enough data to give estimates. [?]		
criminal justice		\$4.85 - \$6.84	1 - 3	221 - 277	\$1,080 - \$1,900
criminal justice associates degree		\$9.58 - \$12.76	1 - 3	0 - 1	\$0 - \$20
criminal justice career		\$3.63 - \$4.41	1 - 3	2 - 3	\$8 - \$60
criminal justice career training			Not enough data to give estimates. [?]		
criminal justice classes		\$8.72 - \$10.90	1 - 3	0 - 1	\$0 - \$20
criminal justice college		\$4.76 - \$6.95	1 - 3	10 - 14	\$50 - \$90
criminal justice course		\$10.63 - \$15.94	1 - 3	0 - 1	\$0 - \$20
criminal justice coursework			Not enough data to give estimates. [?]		
criminal justice degree		\$13.76 - \$20.64	1 - 3	16 - 21	\$230 - \$440
criminal justice online class			Not enough data to give estimates. [?]		
criminal justice program		\$7.34 - \$9.92	1 - 3	2 - 4	\$20 - \$40
criminal justice school		\$7.37 - \$9.31	1 - 3	4 - 6	\$30 - \$60
detective training		\$3.58 - \$4.48	1 - 3	2 - 3	\$8 - \$20
law enforcement career		\$4.40 - \$5.49	1 - 3	3 - 4	\$20 - \$30
law enforcement program		\$3.29 - \$4.12	1 - 3	0 - 1	\$8 - \$15
law enforcement training		\$3.44 - \$4.30	1 - 3	10 - 12	\$40 - \$60

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.

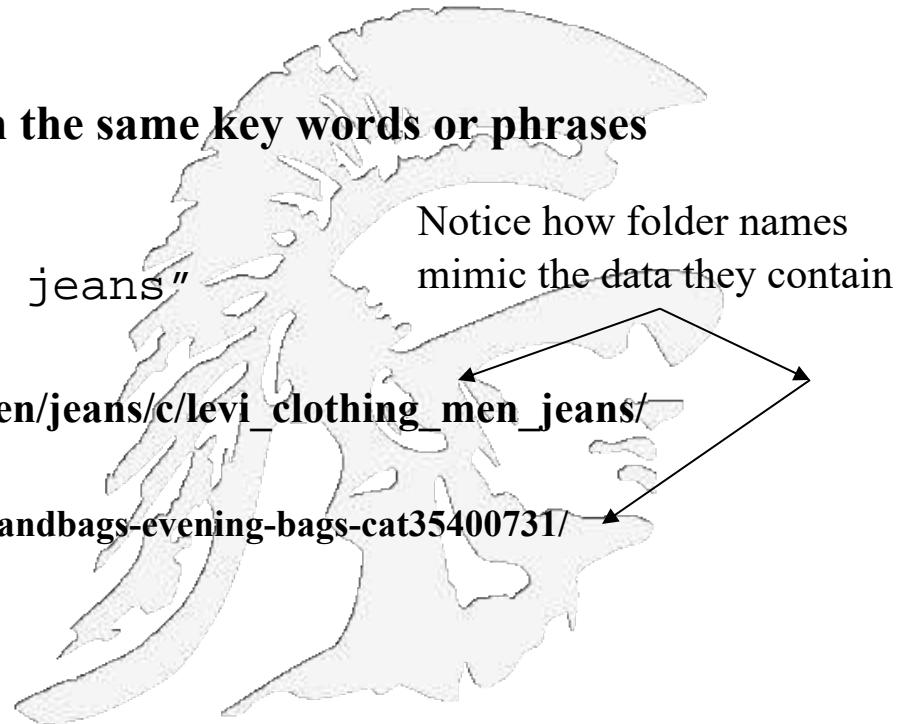
Long Tailed Keywords can be Valuable

- 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/>



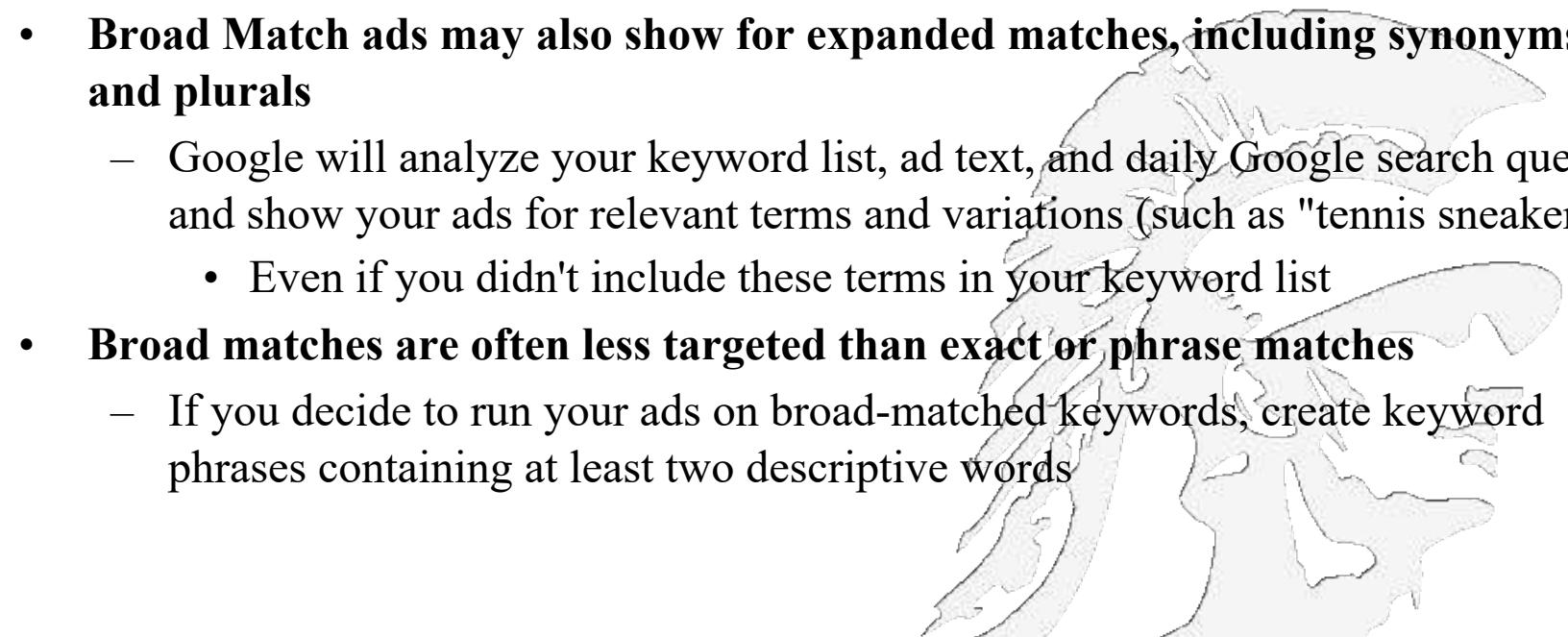
Advertisers Designate Keyword Matching Rules

- 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

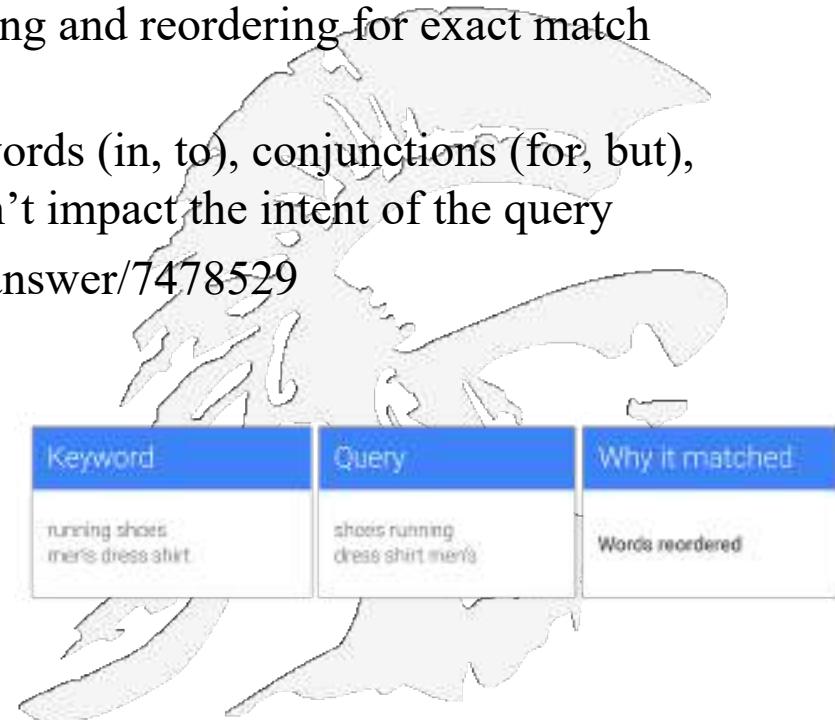


<https://support.google.com/google-ads/answer/2497828>

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>

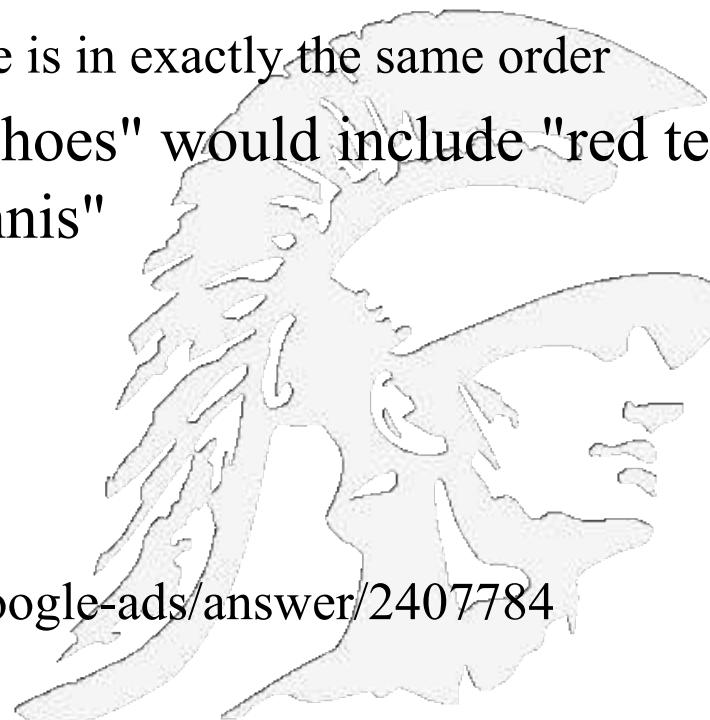
Keyword	Query	Why it matched
jobs in united states nike shoes women	jobs in the united states nike shoes for women	Function words added
parks in san diego paint for a deck	parks san diego paint for deck	Function words removed
news from today bahamas cruise from miami	news for today miami to bahamas cruise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



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

Keyword: Phrase Match

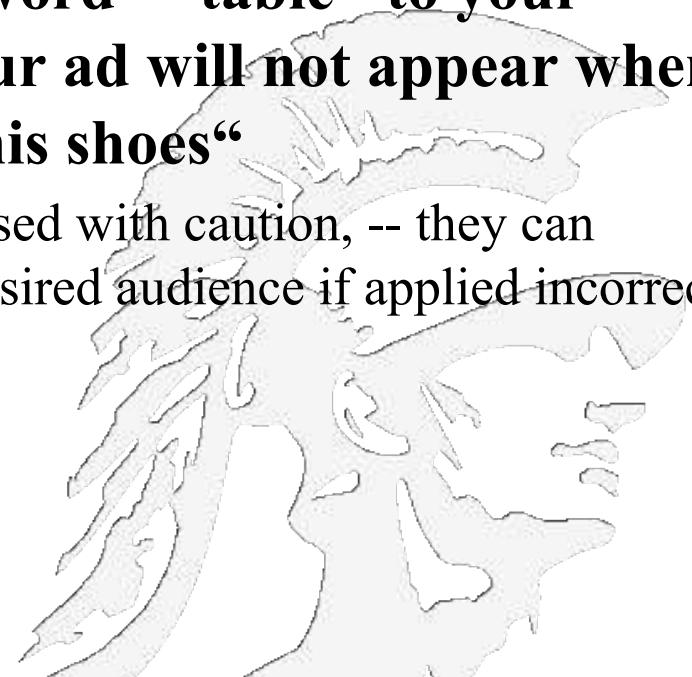
- Your ad appears when users search on the exact phrase
 - AND when their search contains additional terms
 - As long as the keyword phrase is in exactly the same order
 - A phrase match for "tennis shoes" would include "red tennis shoes" but not "shoes for tennis"



<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

File Edit View History Bookmarks Yahoo! Tools Help

51.5 minutes saved

All Campaigns

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

Jan 29, 2007 to Feb 7, 2007

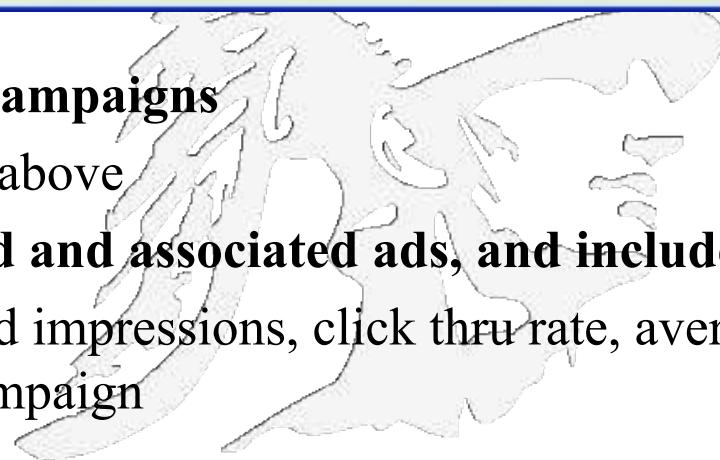
all time

Campaign Name	Current Status	Current Budget [?]	Clicks ▾	Imp.	CTR	Avg. CPC	Cost
Campaign #2	Deleted	\$10.00 / day	747	87,551	0.85%	\$0.44	\$36.34
Campaign #1	Deleted	\$5.00 / day	328	9,893	3.31%	\$0.53	\$174.94
Total - all 2 campaigns		\$10.00 / day (active campaign)	1,075	97,444	1.10%	\$0.47	\$591.28

Find: money

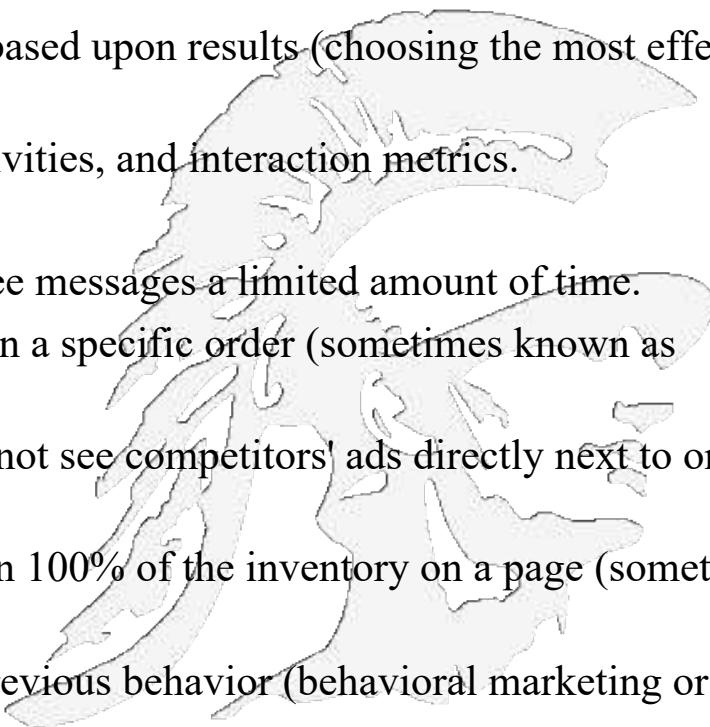
Next Previous Highlight all Match case

- **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



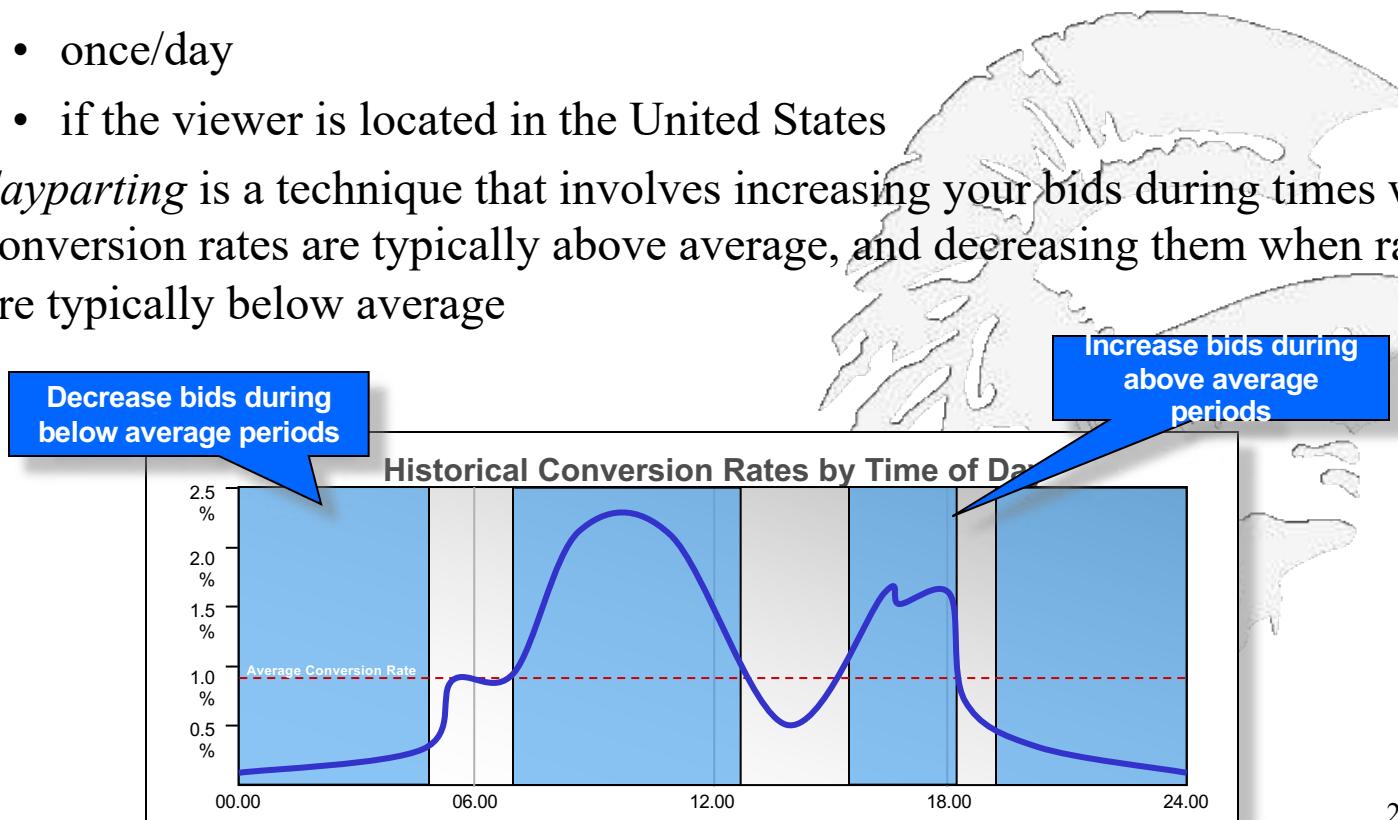
Capabilities of Search Engine Ad Servers

- **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).

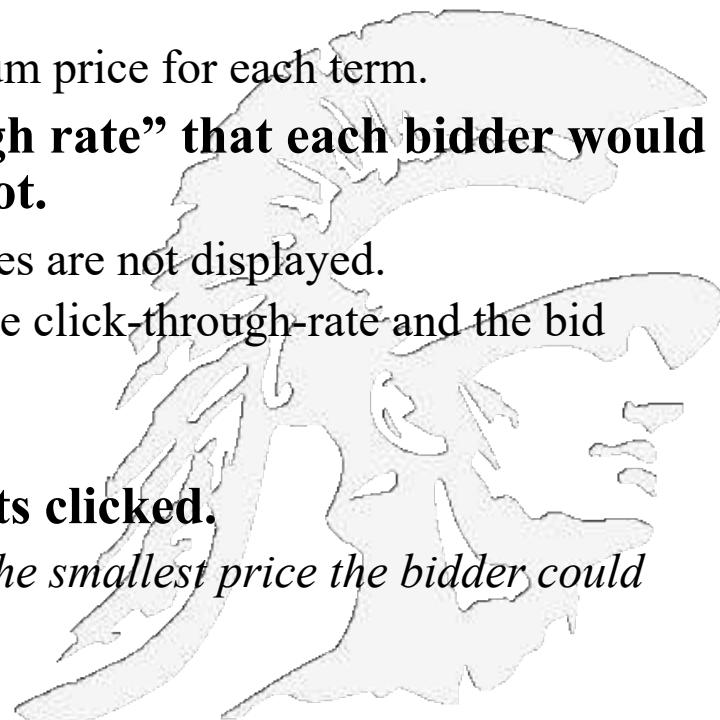


Search Engine Ad Servers are Versatile

- 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

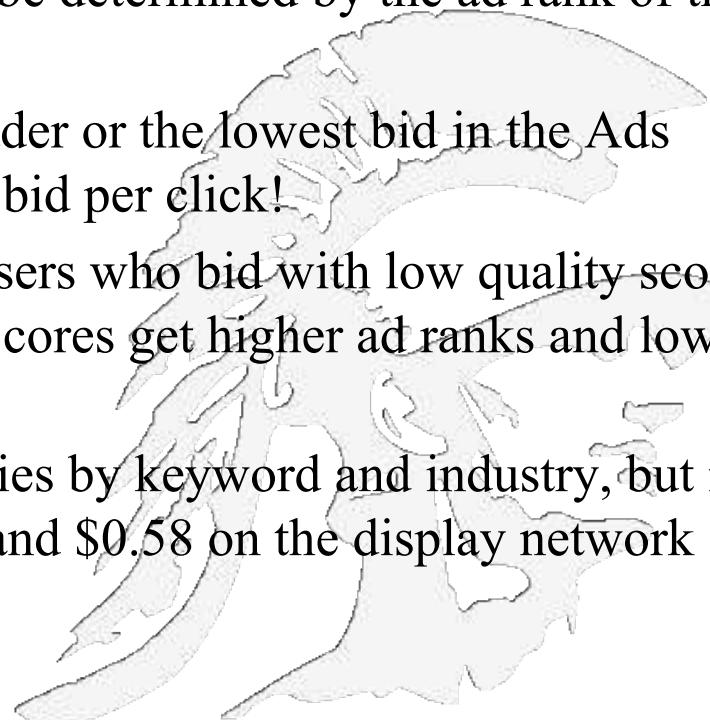


- **Each bidder specifies (i) search terms that trigger its bid and (ii) the amount to bid for each search term.**
 - Bidders may also establish an overall ad budget and limits for each kind of bid.
 - Google may set a reserve or minimum price for each term.
- **Google estimates the “click-through rate” 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*.



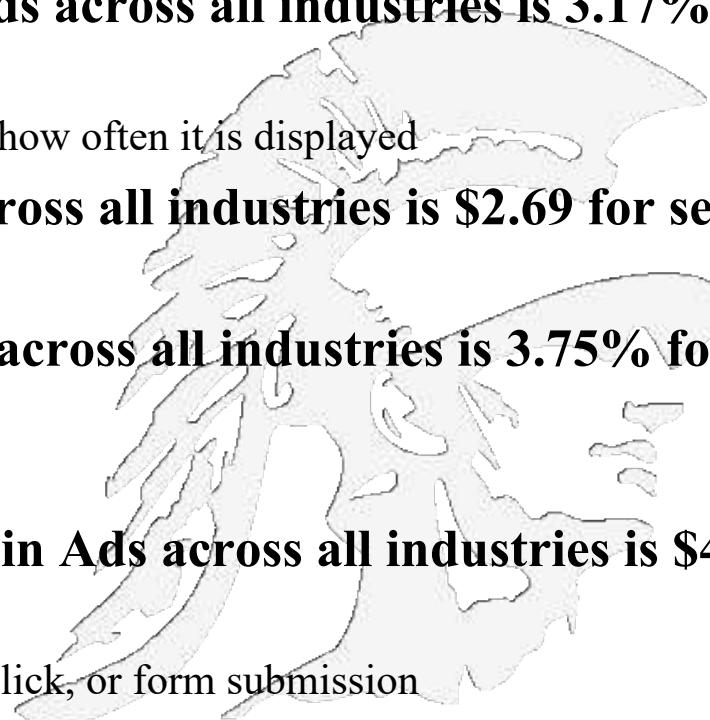
How does AdWords bidding work?

- 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



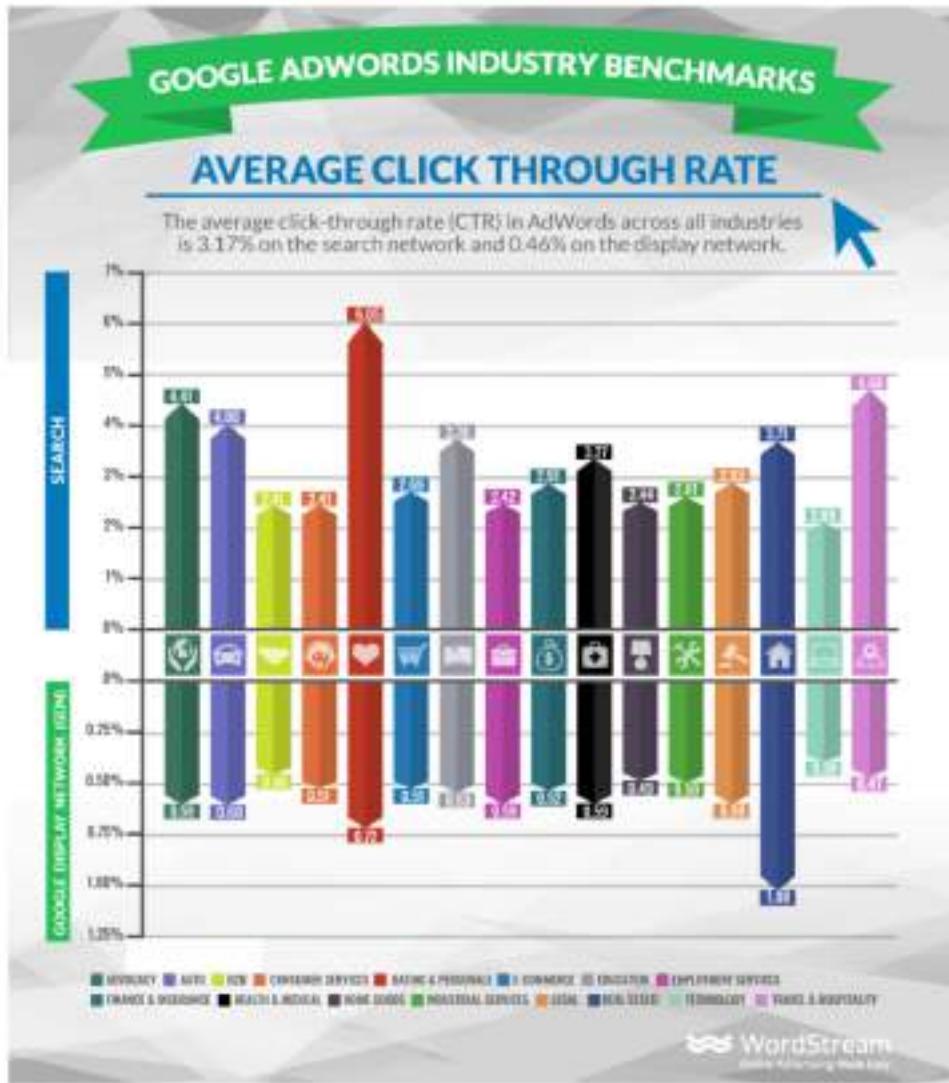
Some Facts About Online Advertising

- <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 $P(C)=0.1$

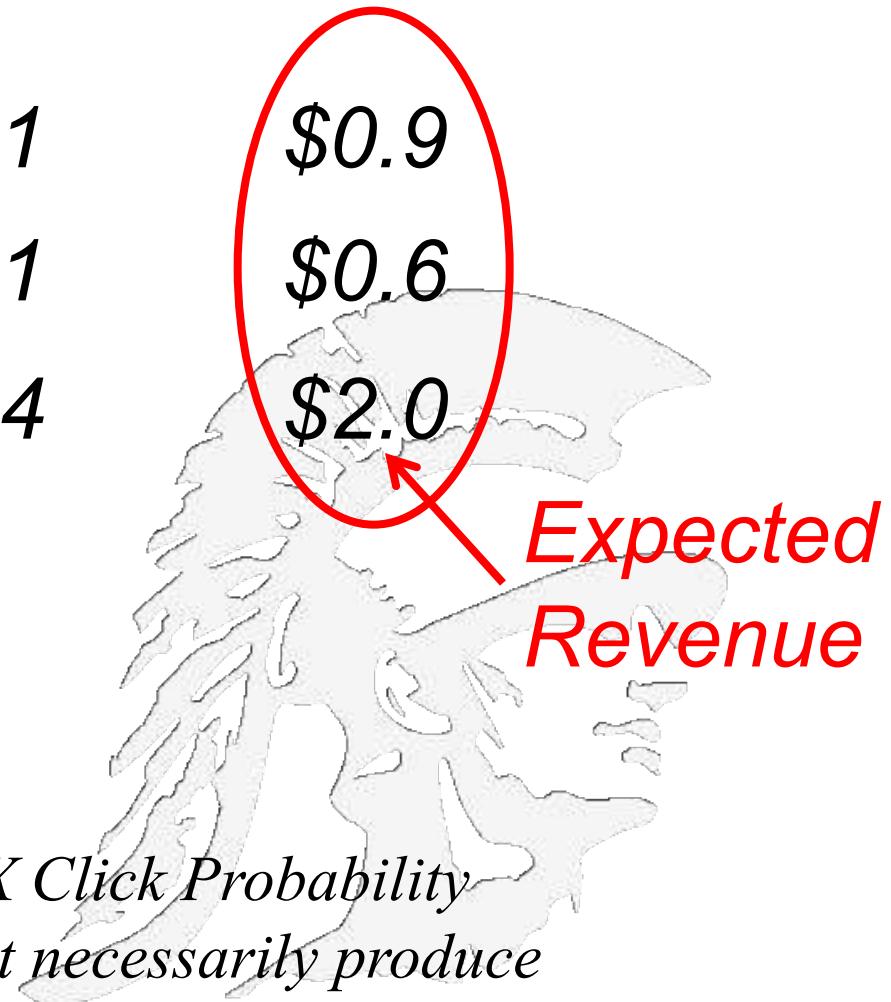
Bid3: \$6 $P(C)=0.1$

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



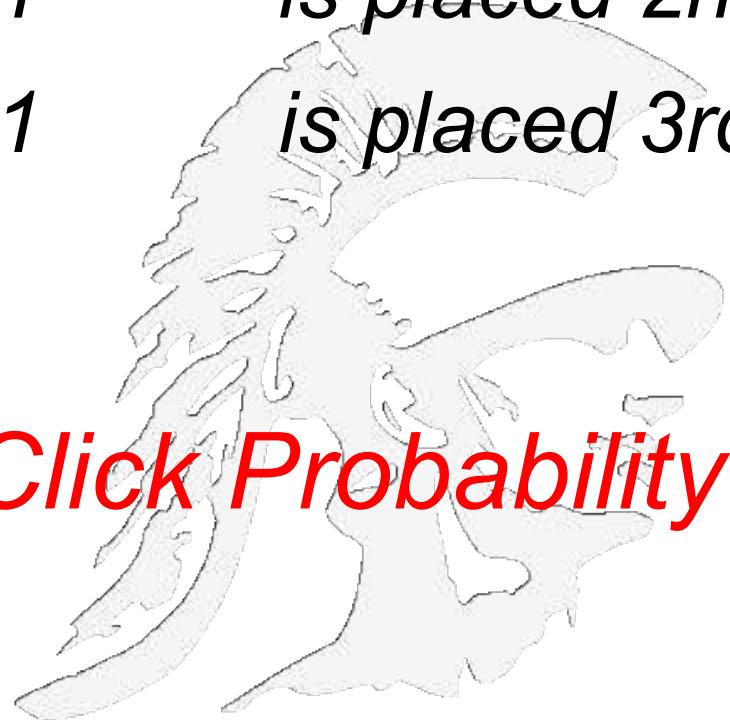
Google Ranks Ads by the Product of Bid Amount*Click Probability

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?)

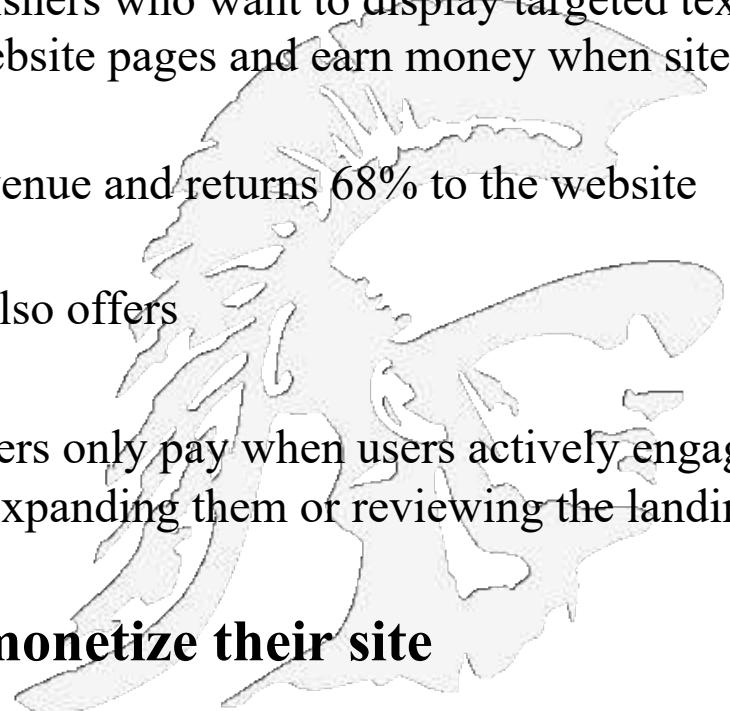
Cost-per-click:

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

Improved Conversion Rate

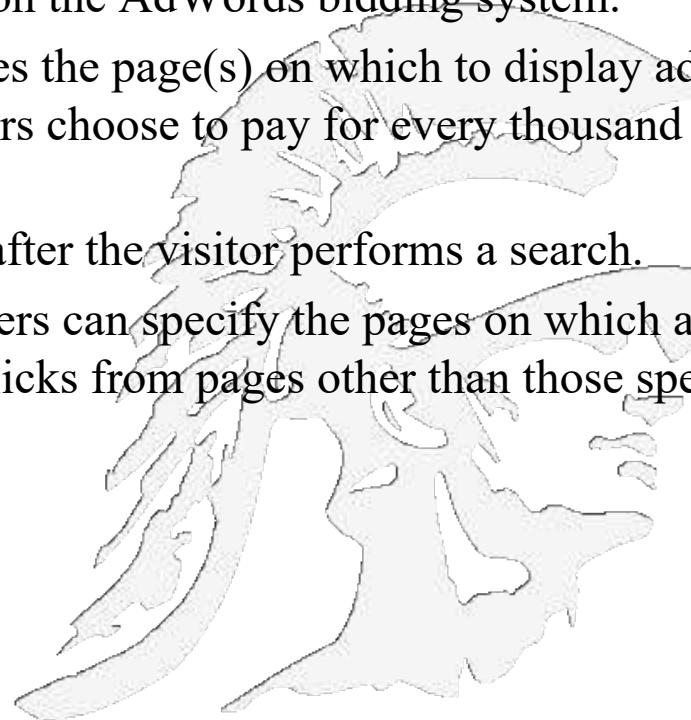


- AdSense from Google is a service for placing Google ads on third party web pages
- Google purchased the content matching technology from Oingo, a small Santa Monica-based search engine in 2003 for \$102 million
- The program is designed for website publishers who want to display targeted text, video 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**



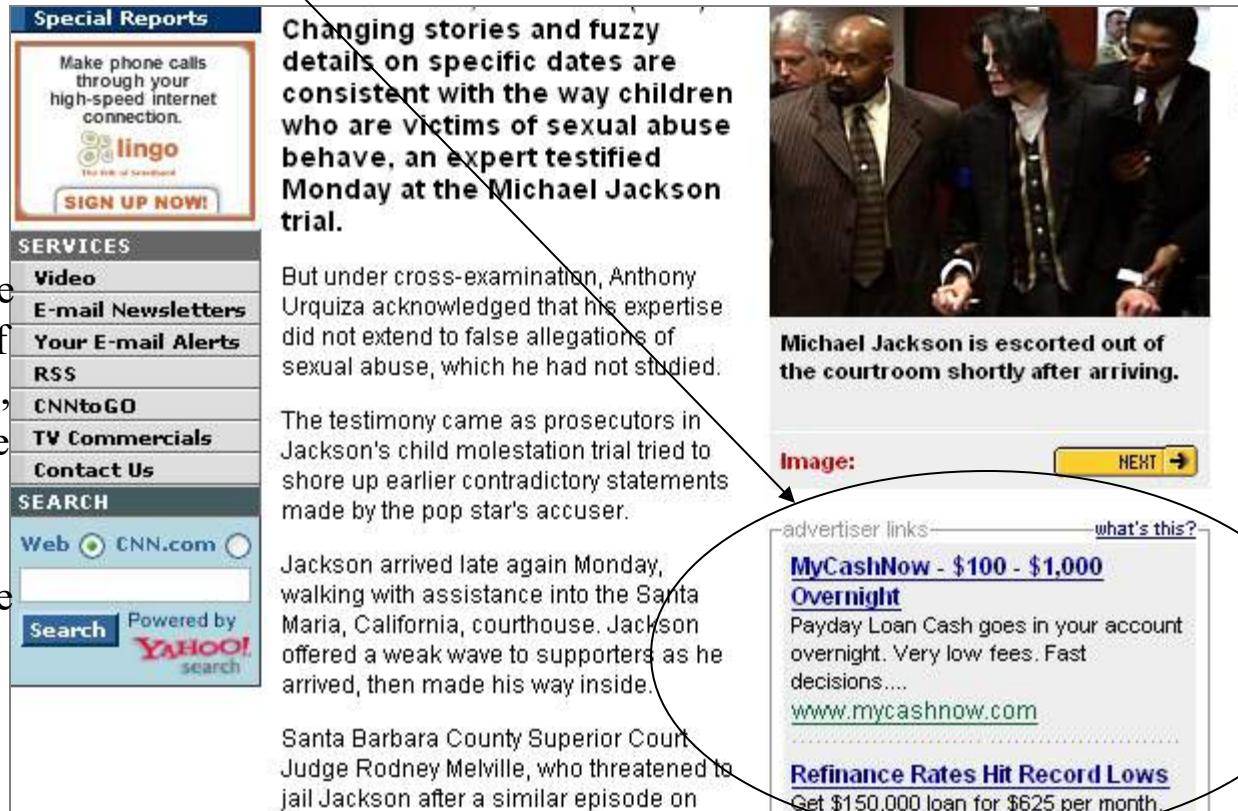
How AdSense Works

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.
 iLingo
[SIGN UP NOW!](#)

SERVICES

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

SEARCH

Web  CNN.com

[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

Image: 

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

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

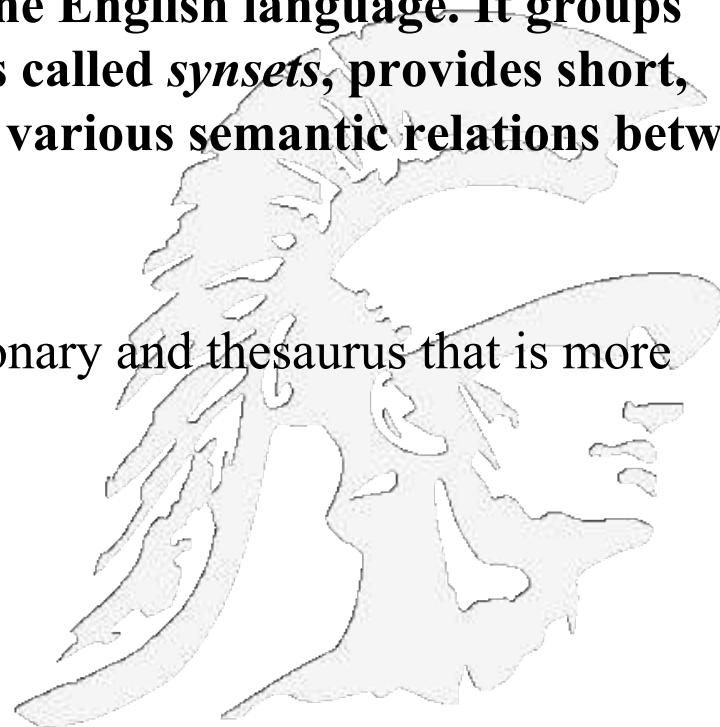
[MyCashNow - \\$100 - \\$1,000 Overnight](#)
 Payday Loan Cash goes in your account overnight. Very low fees. Fast decisions....
www.mycashnow.com

[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

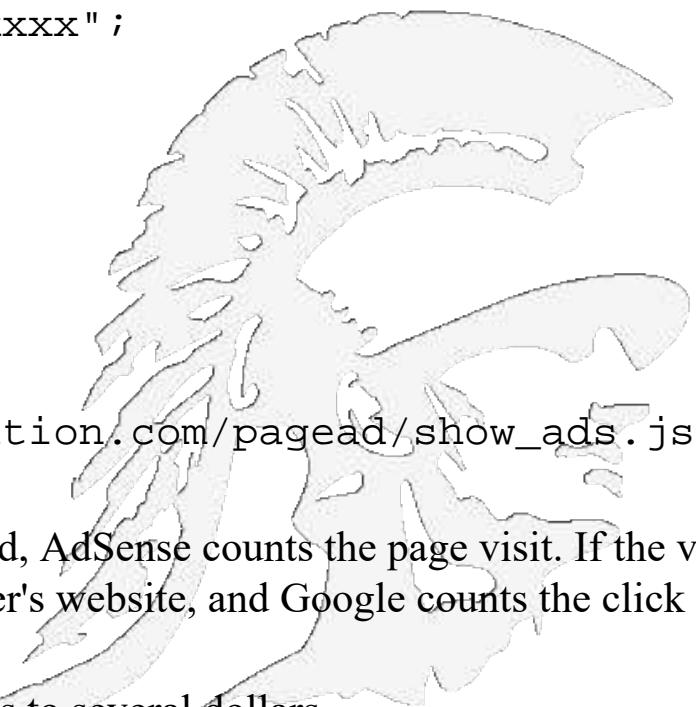
Adsense Content Matching Originally Based on WordNet

- 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

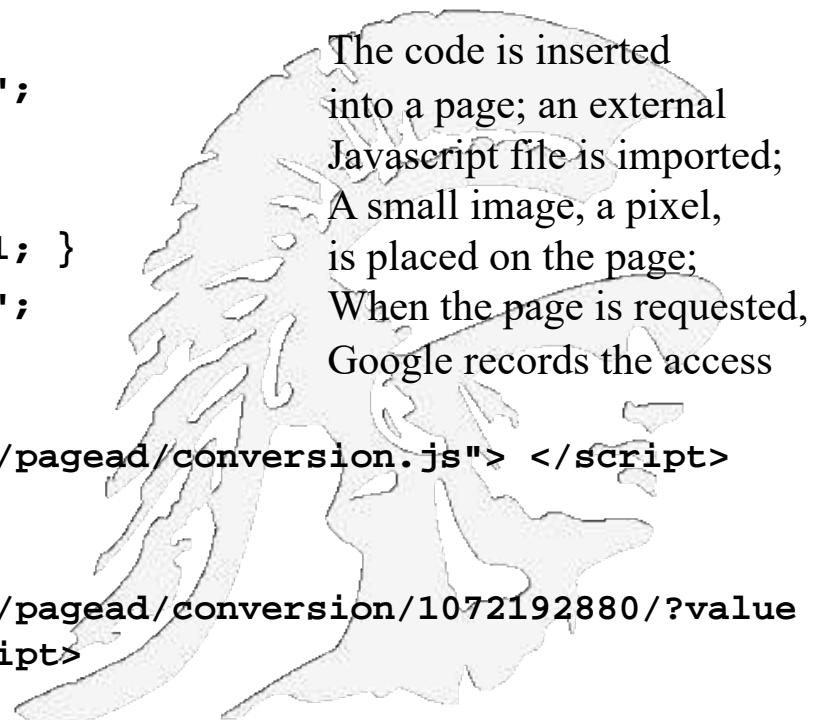
- Google AdSense creates the code, which the publisher copies and pastes into a web page or blog post at the point where it is to appear.
- This is an example of AdSense code for a 728-by-90-pixel ad. Google assigns the numbers following "google_ad_client" and "google_ad_slot."
- ```
<script type="text/javascript"><!--
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

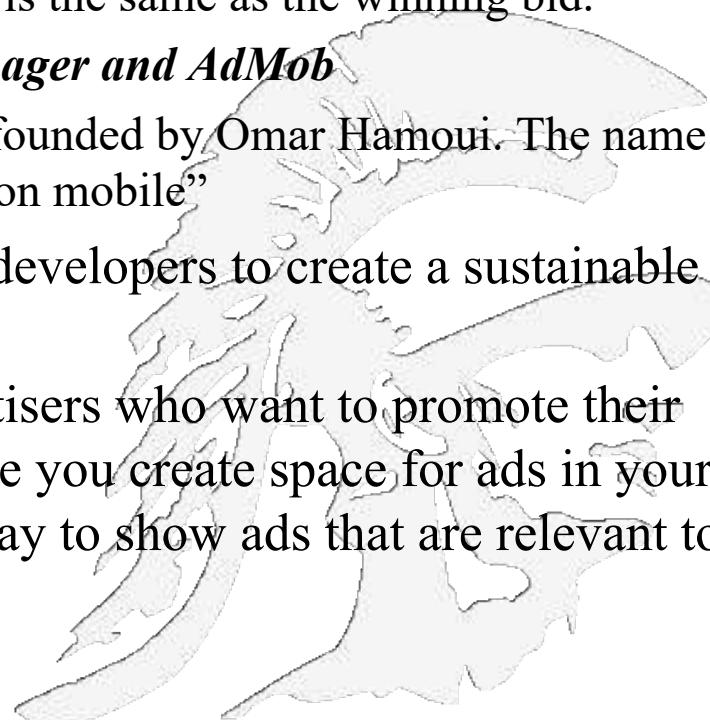
- Google/Yahoo offer tracking pixels
- Tracking pixels are small, typically transparent images on a web page that have special names which permit the loading of the web page to be tracked by a web server.

```
<!-- 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>
```



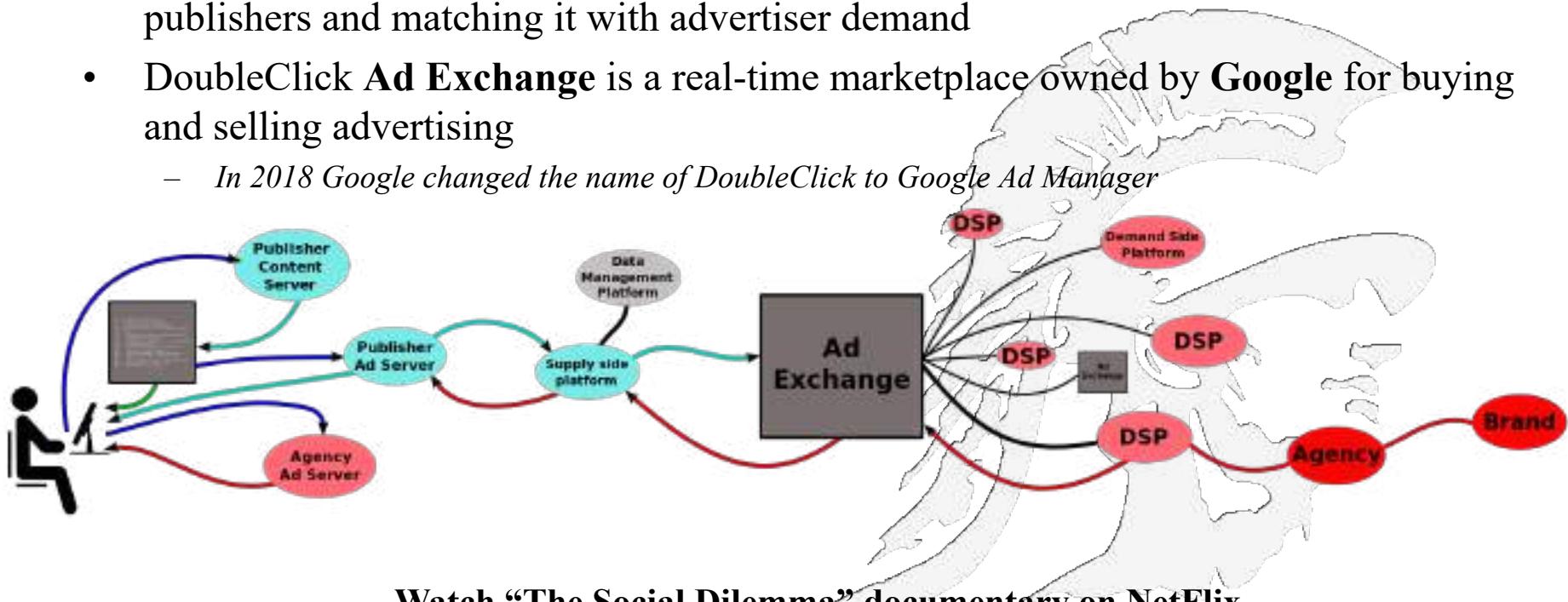
# Google AdSense, AdMob Moves to a First-Price Auction Model

- **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



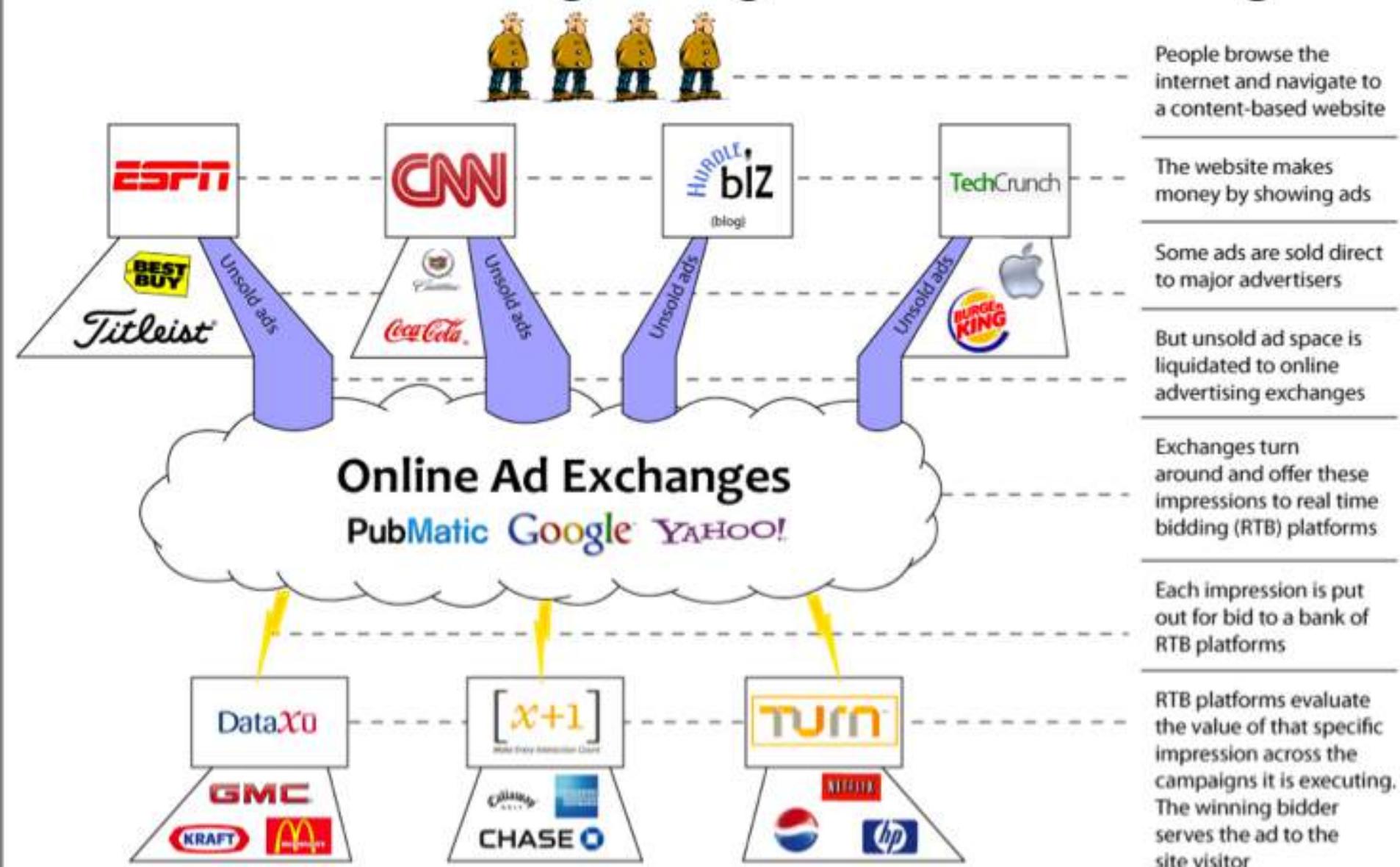
# Ad Exchanges and DoubleClick

- 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



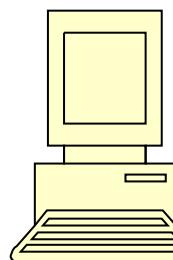
# Today's advertising model

Publishers

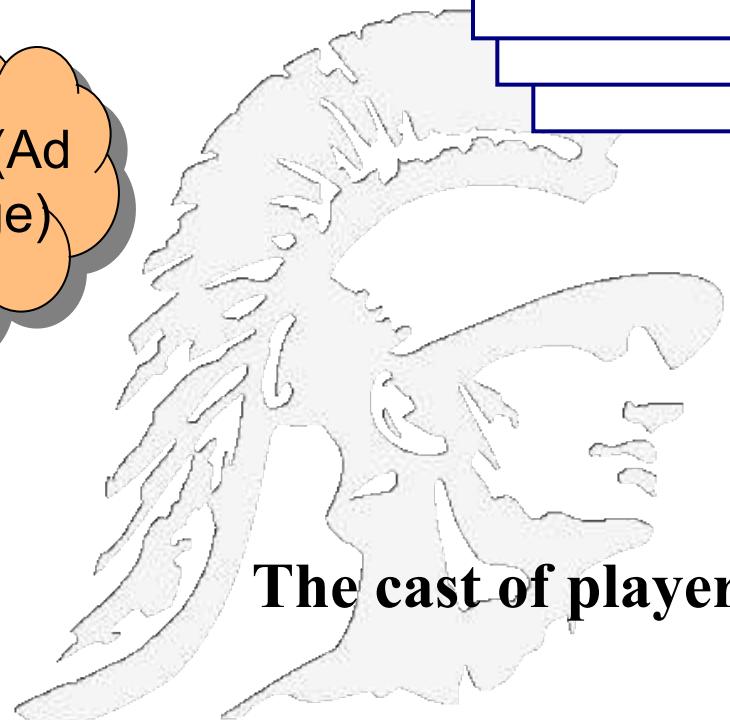
Trackers

Advertisers

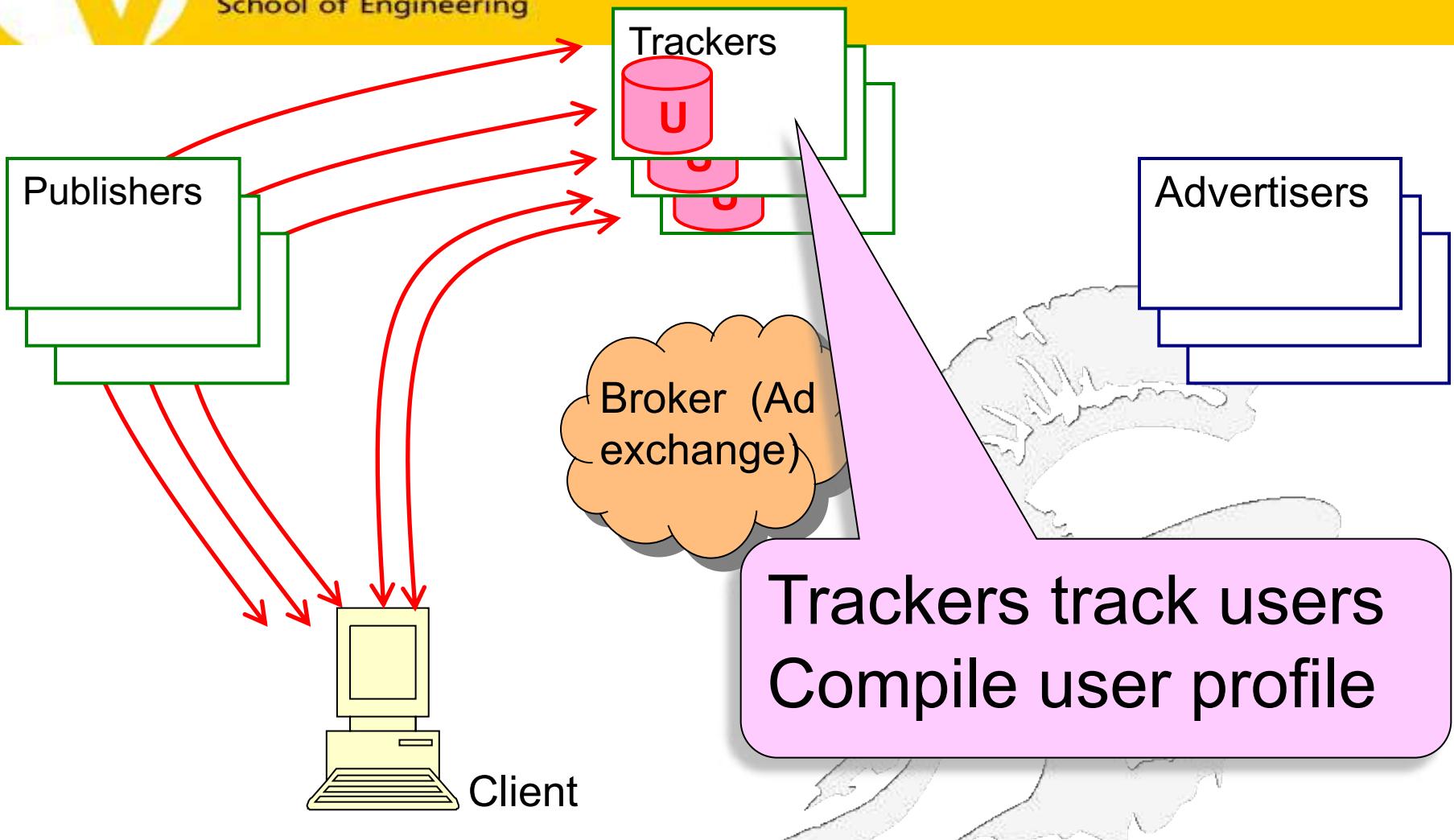
Broker (Ad  
exchange)

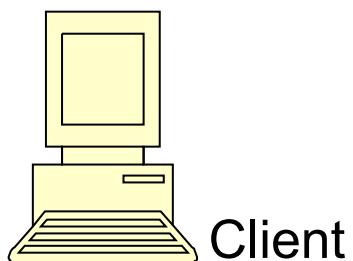
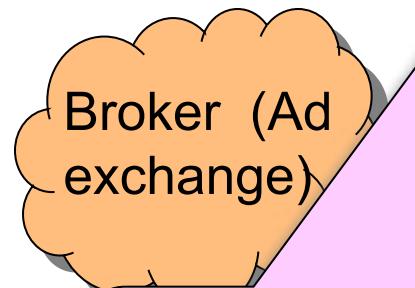
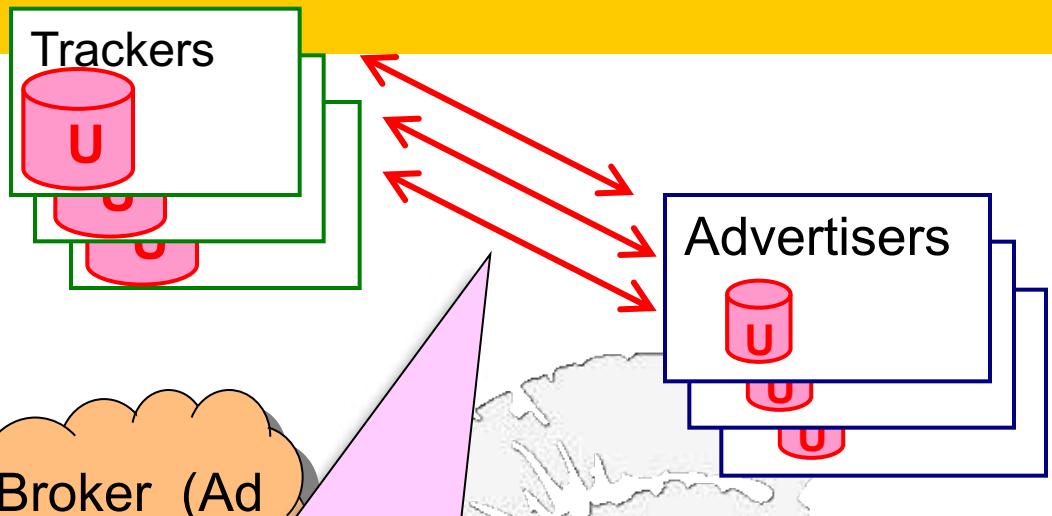
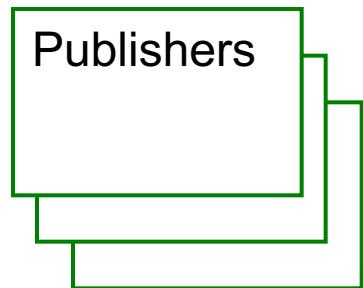


Client

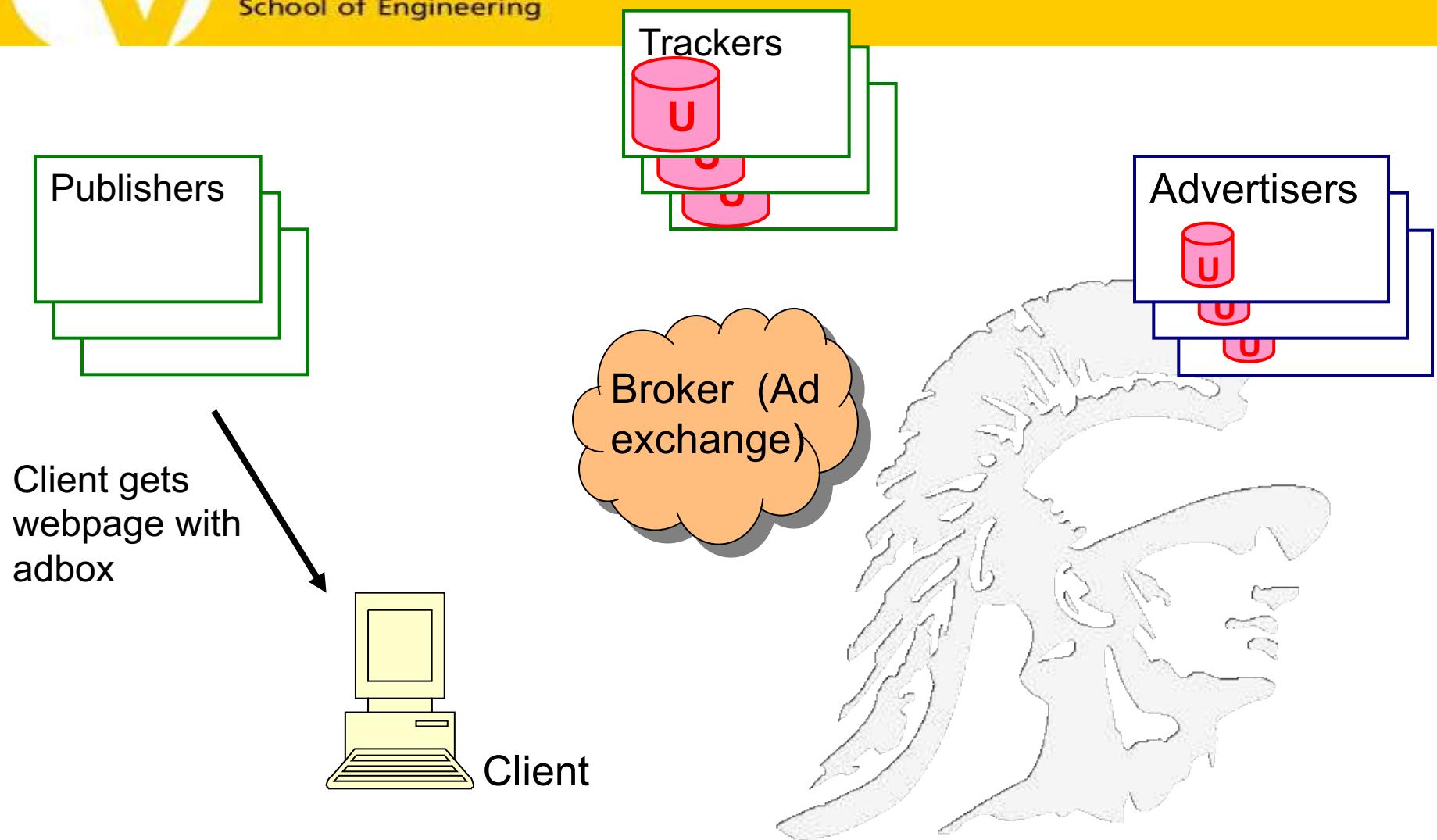


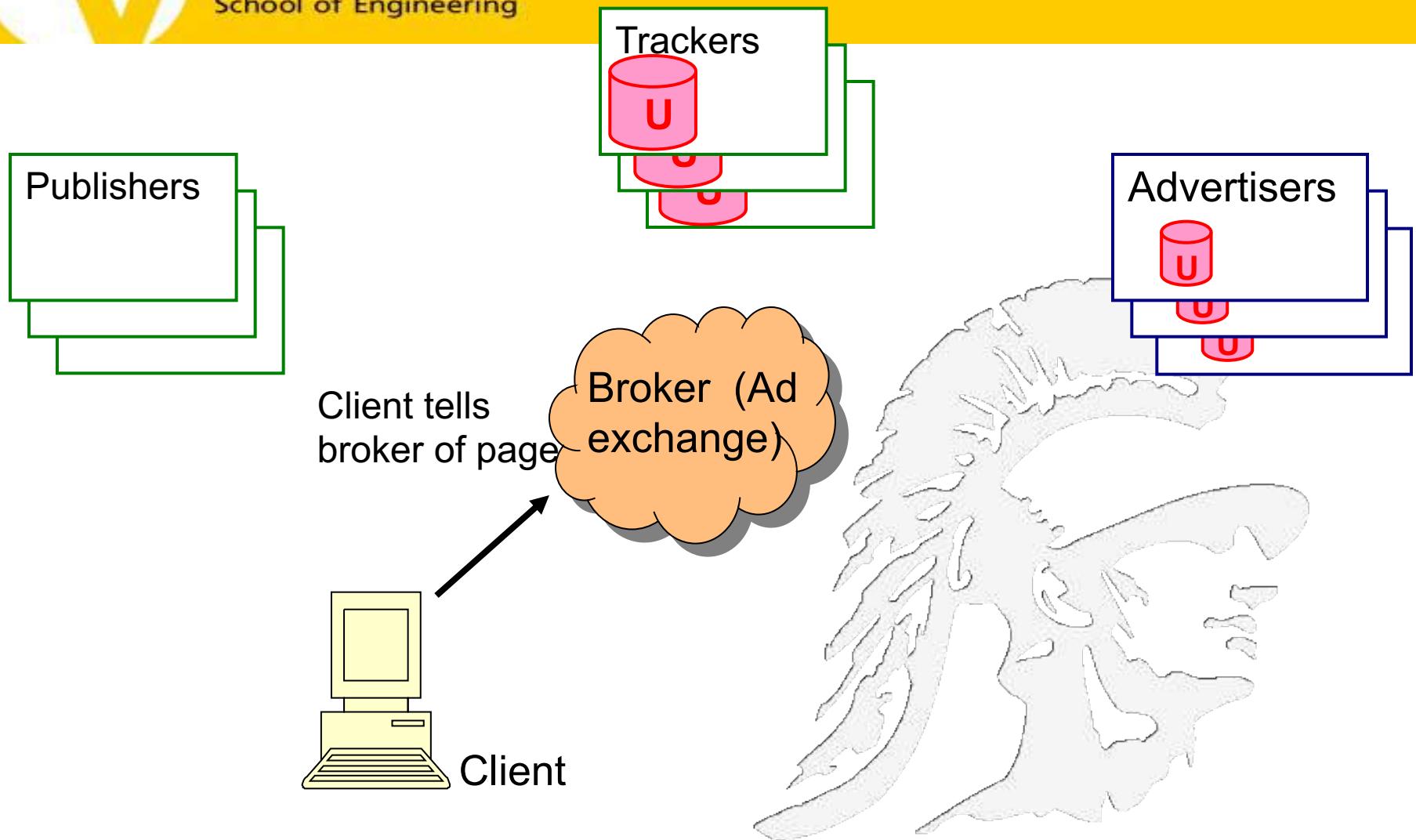
## The cast of players

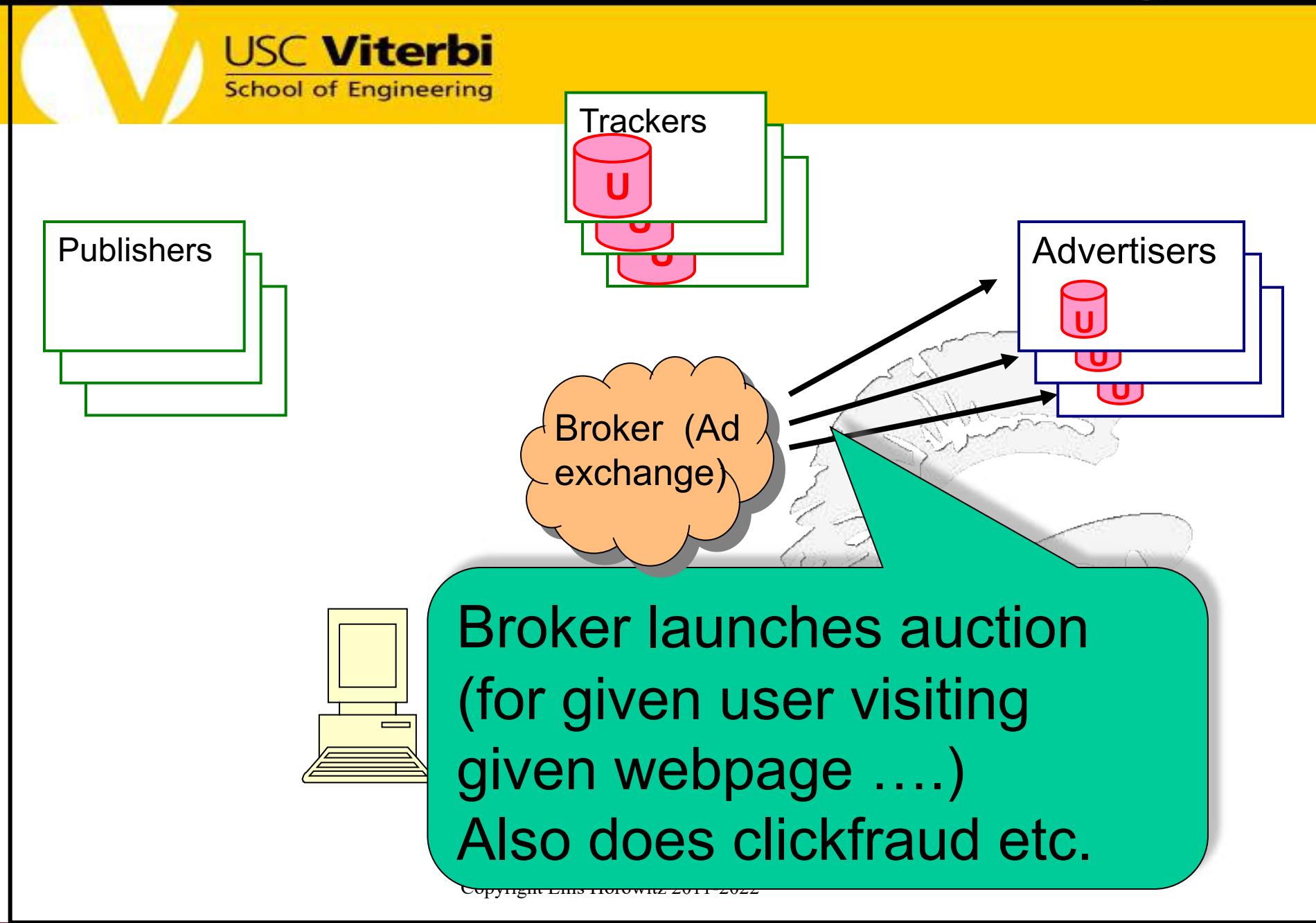




Trackers may share  
profiles with advertisers?





A diagram illustrating the advertising ecosystem. It features four main components: 'Publishers' (green border), 'Advertisers' (blue border), 'Trackers' (pink cylinder icon), and a 'Broker (Ad exchange)' (orange cloud icon). Arrows show the flow of data from Publishers to Trackers, from Trackers to Advertisers, and from Publishers to the Broker. The Broker is shown launching an auction, indicated by a green speech bubble containing text. A computer monitor at the bottom left displays a small version of the diagram.

Publishers

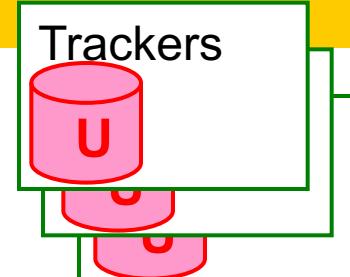
Trackers

Advertisers

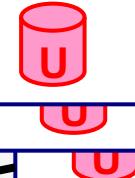
Broker (Ad  
exchange)

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

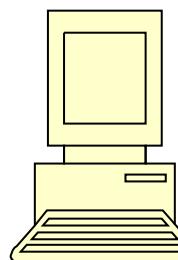
Publishers



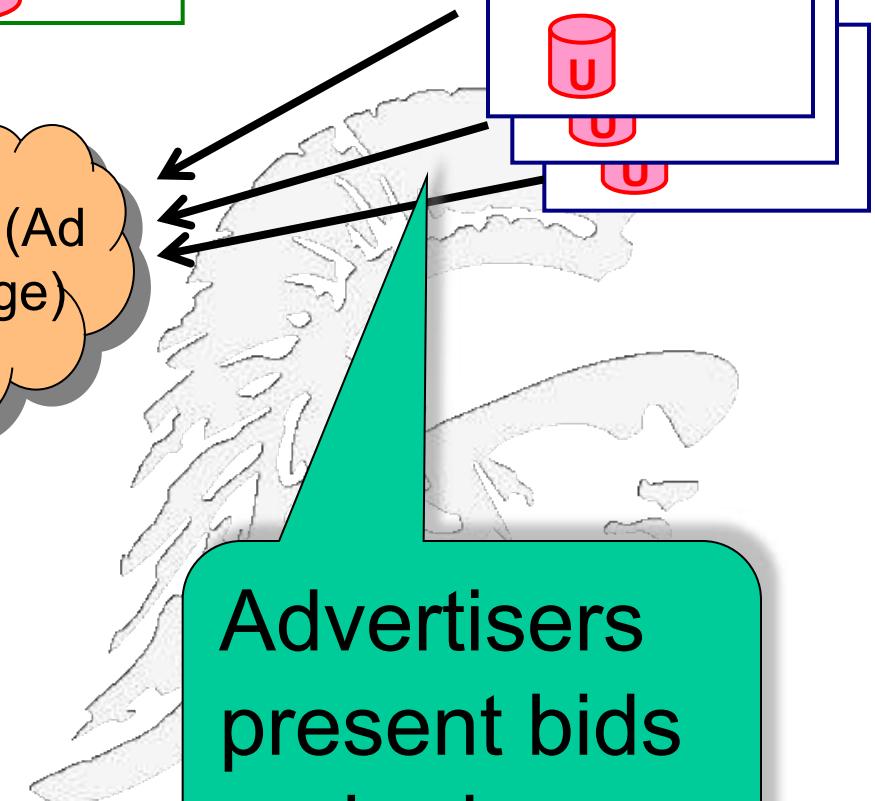
Advertisers



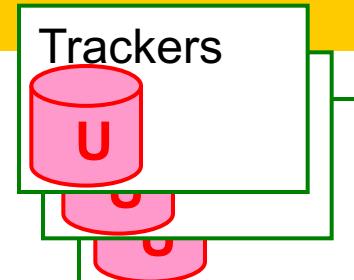
Broker (Ad exchange)



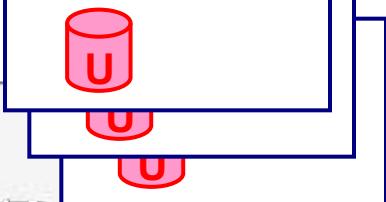
Client



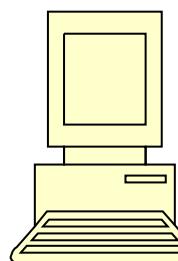
Publishers



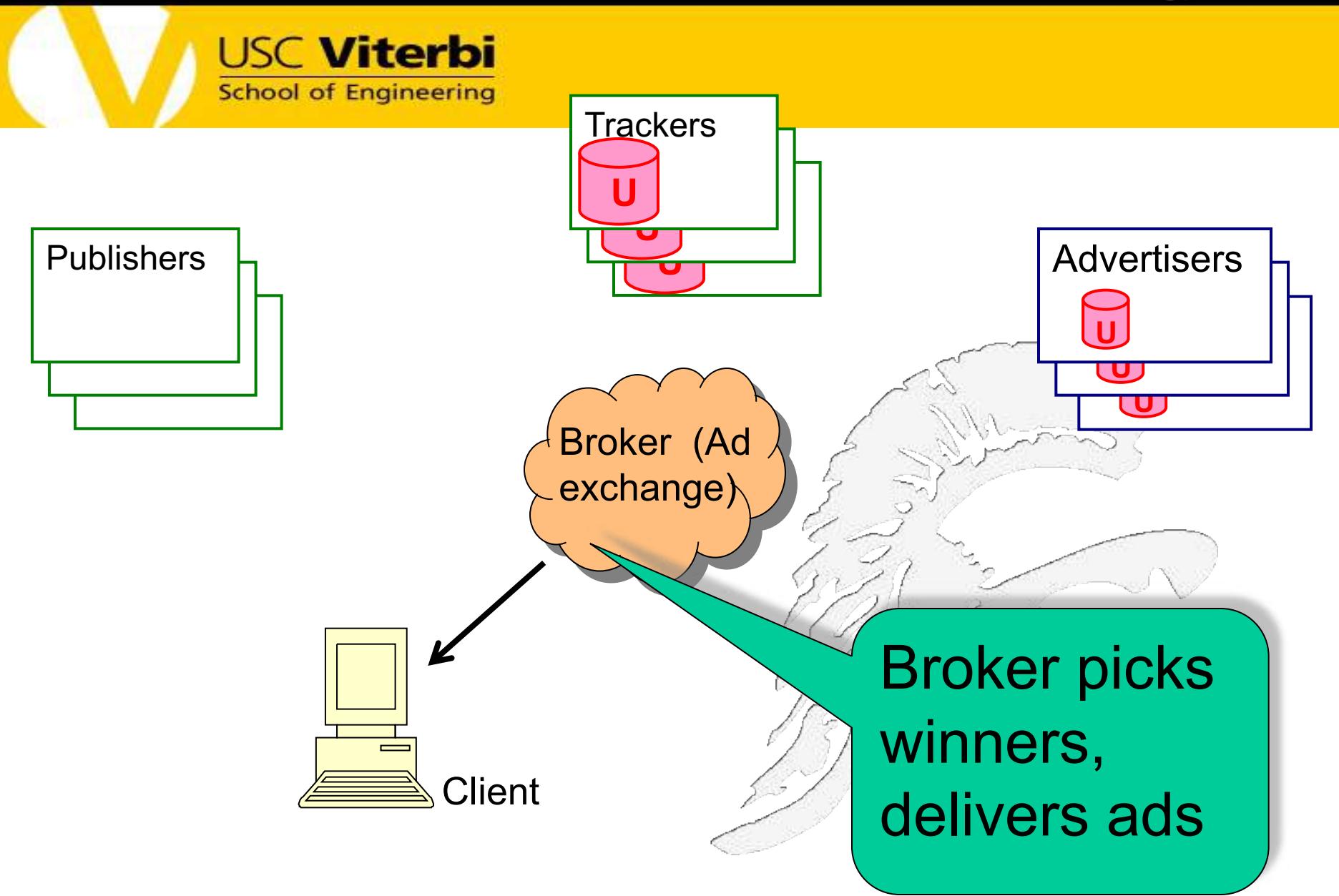
Advertisers



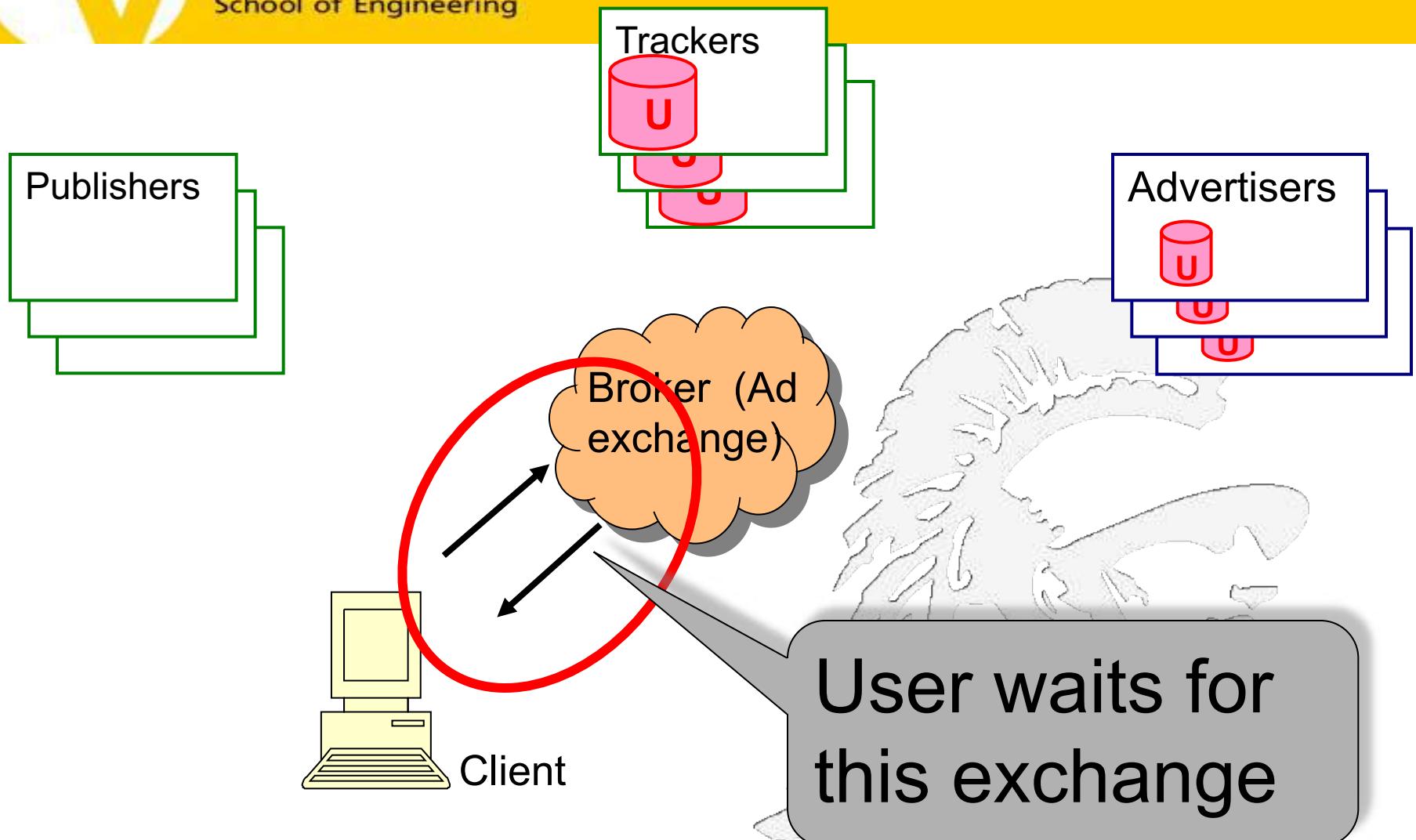
Broker (Ad exchange)

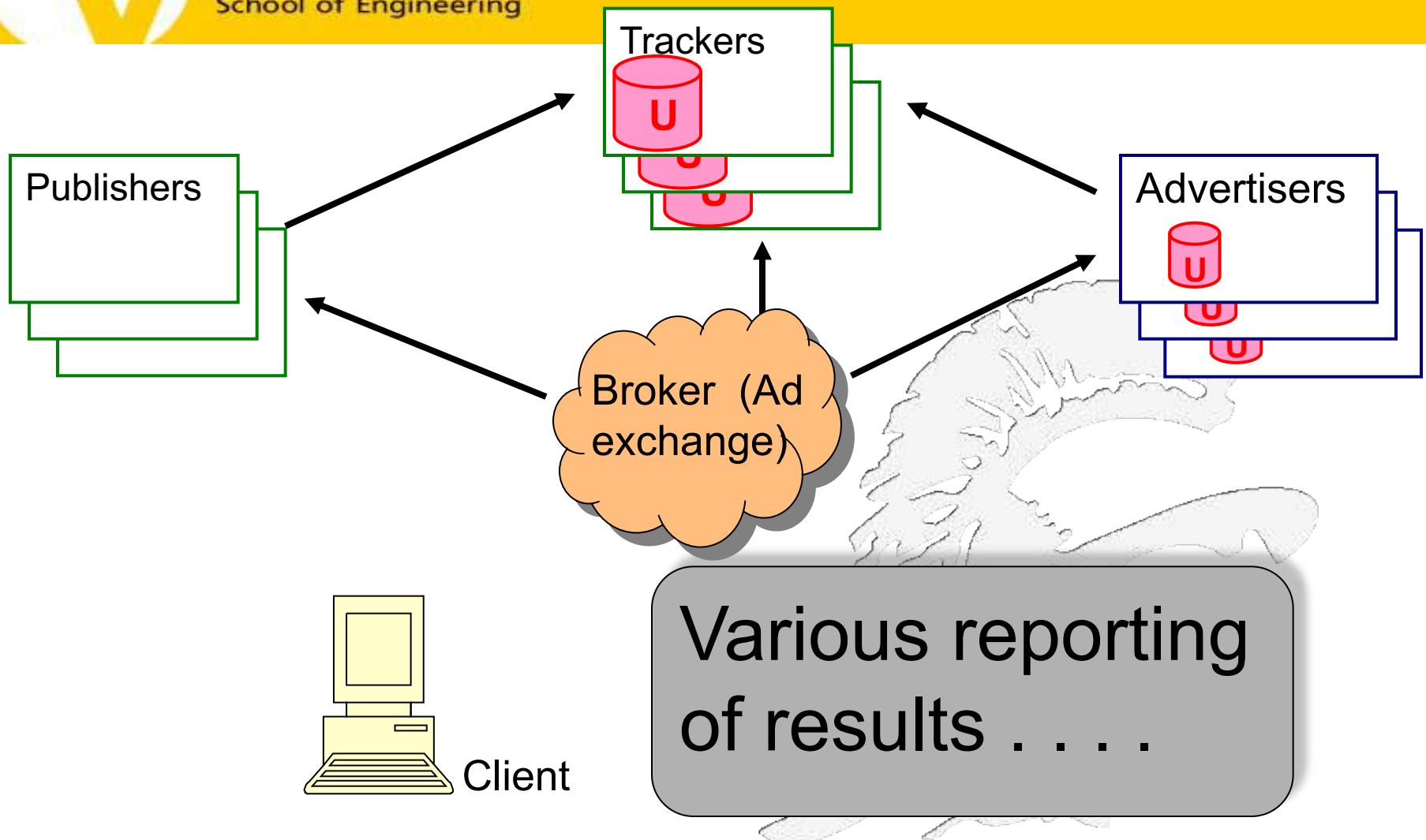


Client

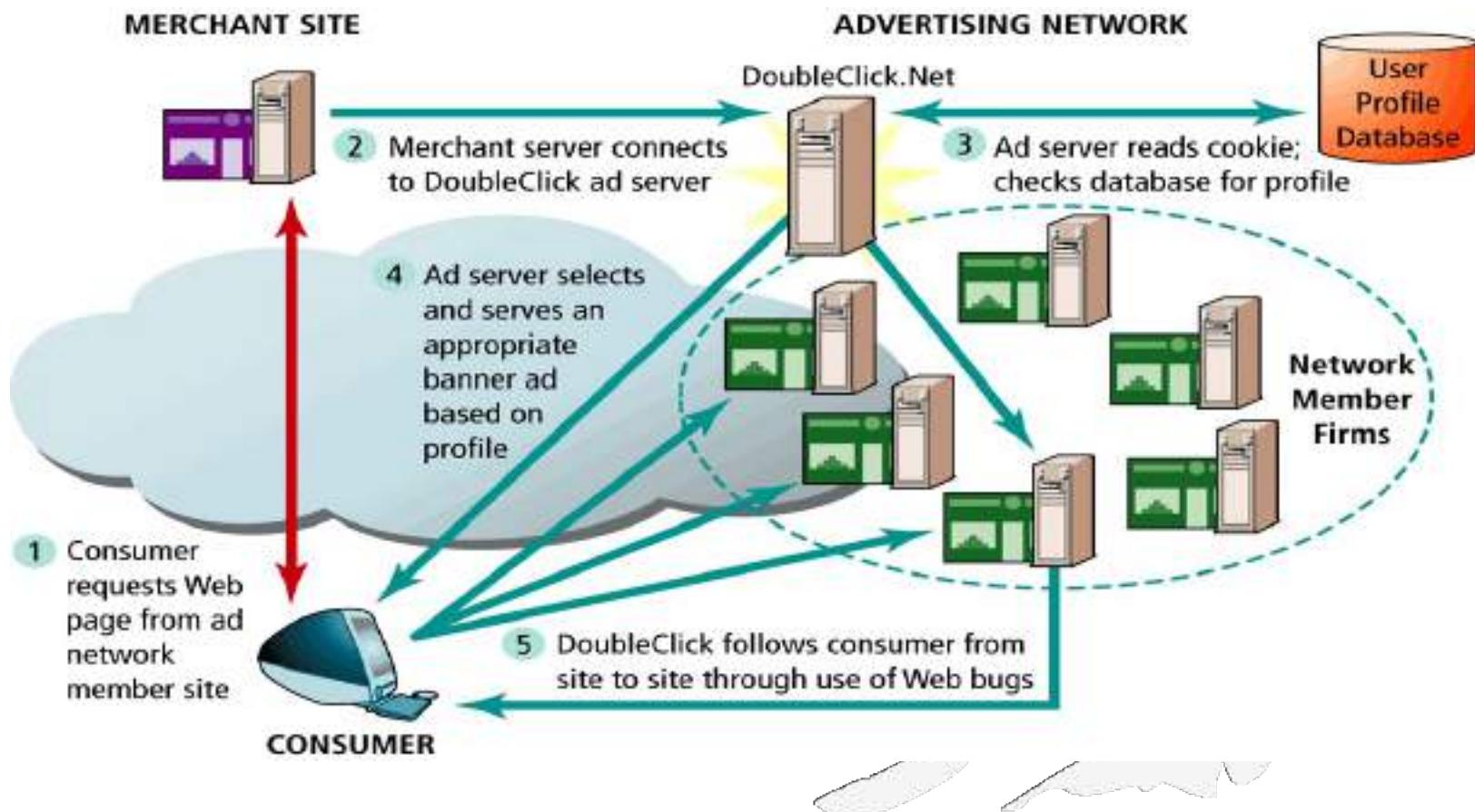


Broker picks  
winners,  
delivers ads





# How an Advertising Network such as DoubleClick Works



← 1/37 → \*\*\* 09:23:31

# PageRank

[You voted for me and who else?  
How popular are you?]

••

University of Southern California  USC

 USC **Viterbi**  
School of Engineering

## Overview

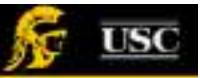
- **Background on Citation Analysis**
- **Google's PageRank Algorithm**
- **Simplified Explanation of PageRank**
- **Examples of how PageRank algorithm works**
- **Observations about PageRank algorithm**
- **Importance of PageRank**
- **Kleinberg's HITS Algorithm**



Copyright Ellis Horowitz, 2011-2022

2

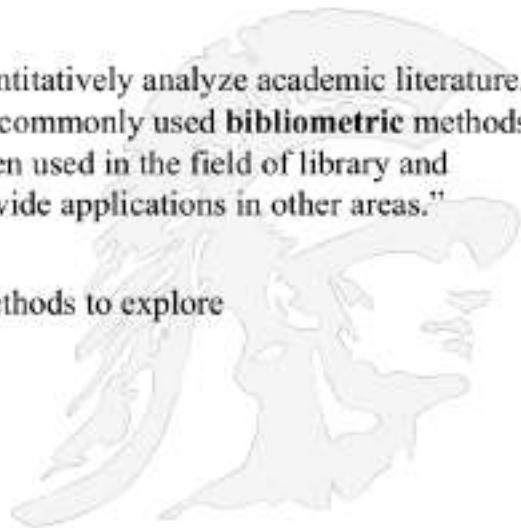
• •

University of Southern California 

 USC **Viterbi**  
School of Engineering

## History of Link Analysis

- **Bibliometrics has been active since at least the 1960's**
- A definition from Wikipedia:
  - "Bibliometrics is a set of methods to quantitatively analyze academic literature. Citation analysis and content analysis are commonly used **bibliometric** methods. While **bibliometric** methods are most often used in the field of library and information science, **bibliometrics** have wide applications in other areas."
- Many research fields use **bibliometric** methods to explore
  - the impact of their field,
  - the impact of a set of researchers, or
  - the impact of a particular paper.



Copyright Ellis Horowitz, 2011-2022

3



# Bibliometrics

- One common technique of **Bibliometrics** is *citation analysis*
- Citation analysis is the examination of the frequency, patterns, and graphs of citations in articles and books.
- citation analysis can observe links to other works or other researchers.
- **Bibliographic coupling:** two papers that cite many of the same papers
- **Co-citation:** two papers that were cited by many of the same papers
- **Impact factor (of a journal):** frequency with which the average article in a journal has been cited in a particular year or period

<http://citeseerx.ist.psu.edu/stats/citations>

Top Ten Most Cited Articles in CS Literature

CiteSeer<sup>x</sup> is a search engine for academic papers



## Most Cited Computer Science Citations

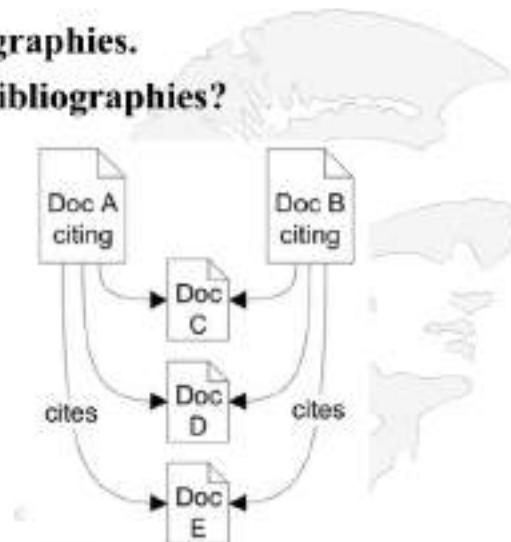
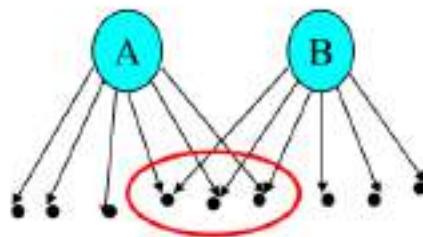
This list is generated from measurements in the CiteSeer <sup>x</sup> database as of March 19, 2013. This list is automatically generated and may contain errors. The last index and citation counts may differ from those currently in the CiteSeer <sup>x</sup> database. This list database is continuously updated.	
<a href="#">View list</a>	
1.	M. McConney, D. S. Johnson Comments and counterexamples: A critique of the Theory of MPCommunication
2.	J. S. Barhenstock, G. F. Franklin, T. M. Apostoloski Minimum-Hanging-in-Laboratory-Wireless: Vol. 1, Oxford University Press, 1994
3.	S. N. Watanabe Statistical Theory: 1968
4.	R. M. Gray, J. A. Thomas Elements of Information Theory: Basics in Telecommunications, 1990
5.	S. U. T. Lamont Challenge of distributed problems during the existence of the head of Informationage
6.	J. T. Tufts, G. E. Carlson, R. L. Wilson, J. Wierwille Introduction to Optimization, 1980
7.	F. K. P. Dargatzka, H. H. Lutz, D. G. Rubin Maximum likelihood from incomplete data via EM algorithm
8.	E. O. Neffert Statistical Algorithms in Speech, Communications and Machine Learning, 1988
9.	R. J. Pfeifer Principles of Programming in Intelligent Systems, 1988
10.	C. E. Shannon, W. Weaver The Mathematical Theory of Communication, 1949

Copyright Ellis Horowitz 2011-2022



# Bibliographic Coupling

- Measure of similarity of documents introduced by Kessler of MIT in 1963.
- The bibliographic coupling of two documents  $A$  and  $B$  is the number of documents cited by both  $A$  and  $B$ .
- Size of the intersection of their bibliographies.
- Maybe want to normalize by size of bibliographies?



Copyright Ellis Horowitz, 2011-2022

5



## Journal Impact Factor

- Developed by Garfield in 1972 to measure the importance (quality, influence) of scientific journals.
- Measure of how often papers in the journal are cited by other scientists.
- Computed and published annually by the Institute for Scientific Information (ISI).
  - It is now owned by Thomson Reuters
- The *impact factor* of a journal  $J$  in year  $Y$  is the average number of citations (from indexed documents published in year  $Y$ ) to a paper published in  $J$  in year  $Y-1$  or  $Y-2$ .
- Does not account for the quality of the citing article
- <https://www.scimagojr.com/journalrank.php?area=1700&category=1702>



## Top Journals for Computer Science

<http://www.guide2research.com/journals/>

### Overall Computer Science

## Top Journals for Computer Science and Electronics

Journal Title	Impact Factor
 IEEE Transactions on Image Processing	9.220
 IEEE Transactions on Fuzzy Systems	8.701
 IEEE Transactions on Signal Processing	8.671
 IEEE Transactions on Industrial Electronics	8.303
 Elsevier Science Direct Books	8.130
 Springer Books	8.093
 Wiley-Blackwell Books	8.077
 IEEE Transactions on Pattern Analysis and Machine Intelligence	8.000
 IEEE Transactions on Evolutionary Computation	8.000
 Elsevier Books	8.000
 Oxford University Press Books	8.000

Software Engineering

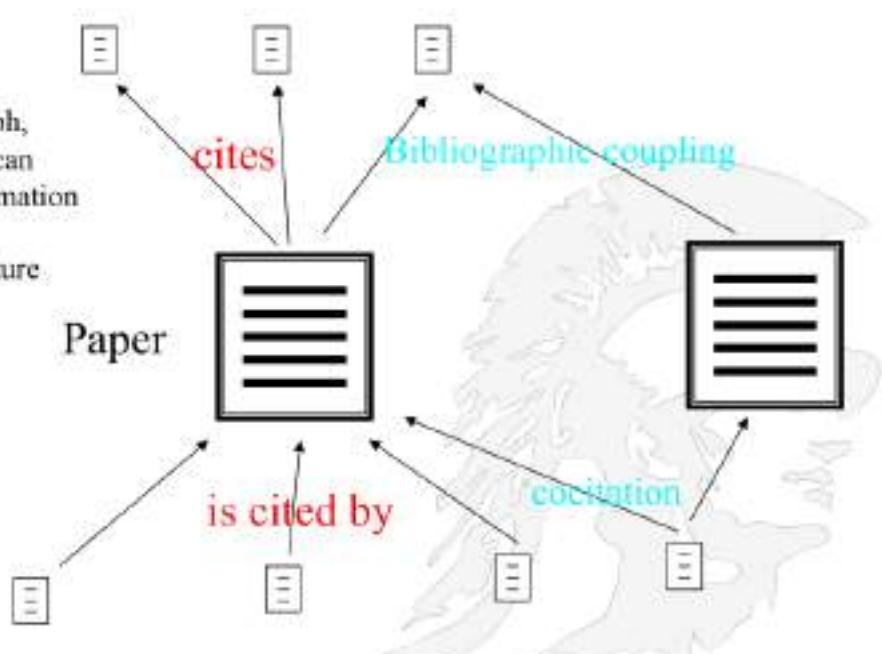
Top Journals for Computer Science and Electronics			
 Scopus	 Impact Factor	Scopus Impact Factor	Scopus Impact Factor
 CrossMark	 CrossMark	CrossMark	CrossMark
 Wiley	 Wiley	Wiley Interdisciplinary Information & Technology	2.512
 IEEE	 IEEE Transactions on Reliability	IEEE Transactions on Reliability	1.287
 Springer	 Springer	Reliability and Risk Analysis: Theory and Application	2.086
 Elsevier	 Elsevier	Information Systems	1.832
 IET	 IET	Reliability Engineering & Safety	1.672
 IEEE	 IEEE	IEEE Transactions on Dependable and Secure Computing	1.582
 Springer	 Springer	Reliability and Risk Analysis: Theory and Application	1.582
 IET	 IET	Reliability of Computer and System Components	1.582
 Springer	 Springer	Information and Software Technology	1.582
 IEEE	 IEEE	IEEE Transactions on Reliability Engineering	1.516
 IET	 IET	Reliability Engineering & Safety	1.516

Copyright Ellis Horowitz 2011-2022

••

University of Southern California  USC 

**Citation Graph**



The diagram illustrates a citation graph. At the center is a large square icon representing a "Paper", containing four horizontal lines. Surrounding this central node are several smaller square icons, each representing a "Paper". Arrows point from the central node to some of the surrounding nodes, labeled "cites" in red text. Other arrows point from the surrounding nodes back towards the central node, labeled "Bibliographic coupling" in cyan text. A third set of arrows points from the surrounding nodes to other peripheral nodes, labeled "is cited by" in red text, and "cocitation" in cyan text.

The structure of this graph, independent of content, can provide interesting information about the similarity of documents and the structure of information

Note that academic citations nearly always refer to the author's earlier work.

Copyright Ellis Horowitz 2011-2022

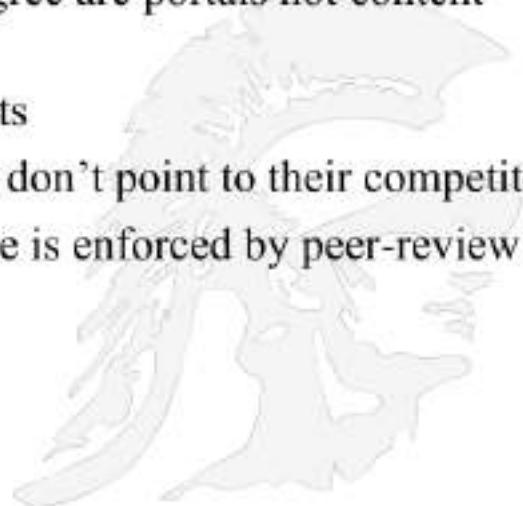
••

University of Southern California 

 USC **Viterbi**  
School of Engineering

## Citations vs. Web Links

- **Web links are a bit different than citations:**
  - Many links are navigational
  - Many pages with high in-degree are portals not content providers
  - Not all links are endorsements
  - Company websites normally don't point to their competitors
  - Citations to relevant literature is enforced by peer-review



Copyright Ellis Horowitz, 2011-2022

9

••

University of Southern California 

 USC **Viterbi**  
School of Engineering

## What is PageRank?

- PageRank is a **web link analysis algorithm** introduced by Google
- PageRank was developed at Stanford University by Google founders **Larry Page and Sergey Brin**
  - The paper describing PageRank was co-authored by Rajeev Motwani and Terry Winograd
- PageRank is a “**vote**”, by all the other pages on the Web, about how important a page is.
- A link to a page counts as a vote of support.
- If there’s no link there’s no support (but it’s an abstention from voting rather than a vote against the page).
- PageRank says nothing about the content or size of a page, the language it’s written in, or the text used in the anchor of a link!
- Looked at another way, PageRank is a **probability distribution** used to represent the likelihood that a person randomly clicking on links will arrive at any particular page

Copyright Ellis Horowitz, 2011-2022

10

• •



[www.freepatentsonline.net/us6285999.pdf](http://www.freepatentsonline.net/us6285999.pdf)

SEARCH THIS SITE Main Page - Home Faculty Resources Contact Us Search Google Search - Index (2)

**United States Patent**  
Page

**101 United States Patent**  
**102 Patent No.: US 6,285,999 B1**  
**103 Date of Patent: Sep. 4, 2001**

**104 METHOD FOR NODE RANKING IN A LINKED DATABASE**

**105 Inventor:** Lawrence Page, Mountain, CA (USA)  
**106 Assignee:** The Board of Trustees of the Leland Stanford Junior University, Mountain, CA (USA)

**107 Status:** Subject to any disclaimer, the term of this patent is extended or adjusted under 35 U.S.C. 154(b) by 9 days.

**108 Appl. No.:** 08/984,827  
**109 Filed:** Jun. 9, 1998

**110 Related U.S. Application Data**  
**111 Provisional application No. 60/018,100, filed on Jan. 10, 1998.**

**112 Int'l Cl.:** G06F 17/30  
**113 U.S. Cl.:** 885K, 307.7, 307.80  
**114 Field of Search:** 307.00, 5, 7, 907/51, 1-3, 18, 184, 301, 305/400, 307/35, 228, 230, 231

**115 References Cited**  
**116 Patent Documents**

JP2000-110000	1999-06-01	Cohen et al.	1001000
JP09-100000	1999-01-01	Sankar	1001000
US5805963	1998-09-01	Chung et al.	1001000
US5202500	1993-04-13	Feldman	1001000
US5202504	1993-04-13	Spagno et al.	1001000
US5004407	1991-04-03	Adamic et al.	1001000
US4904559	1990-02-20	Israel et al.	1001000

**117 OTHER PUBLICATIONS**

K. Jacobo Cordeiro et al., "Web Query Processing and Visualizing the Web through Connectivity", Computer Networks and ISDN Systems 29 (1997) pp. 1257-1267.  
Wong et al, "Practicing in Web Wide Web", IEEE 7796, pp. 28-32.  
Ritter et al, "Statistics, Probability and Database Organization", Electrical Abstract, 1996, pp. 172-276.

**118 Primary Examiner—Peter Black**  
**119 Assistant Examiner—Ursula Lai**  
**120 Attorneys, Agents, or Firms—Haus & Sander LLP**

**121 ABSTRACT**

A method assigns a ranking ratio to nodes in a linked database, such as one database of documents containing obvious, that would work well in any other hypermedia database. The rank assigned to a document is calculated from the ratio of documents citing it. In addition, the role of a document is calculated from a measure representing the probability that a browser, following the database will accidentally jump to the document. The method is particularly useful in enhancing the performance of search engine results for hypermedia databases, such as the world wide web, where documents have a large variation in quality.

25 Claims, 3 Drawing Sheets

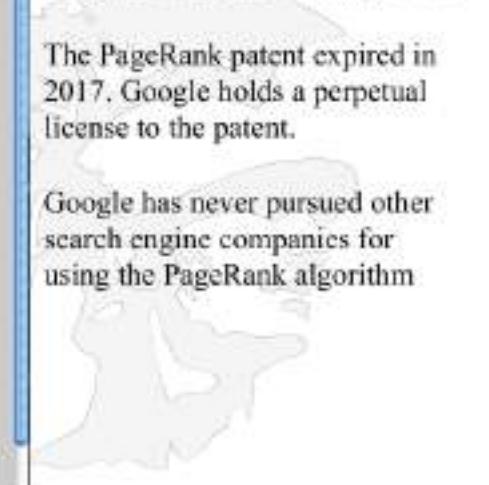
Copyright © 2001 EPO/USPTO. All rights reserved.

## Page Rank Patented

A copy of the front page of the patent of the PageRank algorithm; Larry Page is credited as the inventor; the patent was awarded to Stanford University; the patent was filed January 1998

The PageRank patent expired in 2017. Google holds a perpetual license to the patent.

Google has never pursued other search engine companies for using the PageRank algorithm



<https://bytes.usc.edu/cs572/s23-sear-chhh/lectures/PR/index.html>

11/41

• •

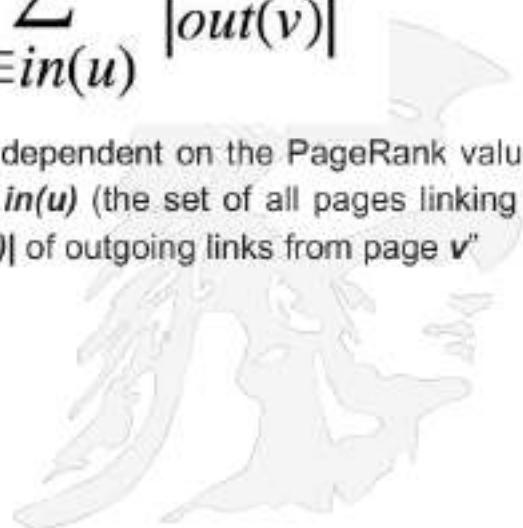
University of Southern California 

 USC **Viterbi**  
School of Engineering

## Initial PageRank Formulation

$$PR(u) = \sum_{v \in in(u)} \frac{PR(v)}{|out(v)|}$$

→ "the PageRank value for a page  $u$  is dependent on the PageRank values for each page  $v$  contained in the set  $in(u)$  (the set of all pages linking to page  $u$ ), divided by the number  $|out(v)|$  of outgoing links from page  $v$ "



Copyright Ellis Horowitz, 2011-2022

12

••

University of Southern California  USC

 USC **Viterbi**  
School of Engineering

## Steps for Simplified Algorithm



- Iteration 0:** Initialize all ranks to be  $1/(\text{number of total pages})$ .
- Iteration 1:** For each page  $u$ , update  $u$ 's rank to be the sum of each incoming page  $v$ 's rank from the previous iteration, divided by the number total number of links from page  $v$ .

Copyright Ellis Horowitz, 2011-2022

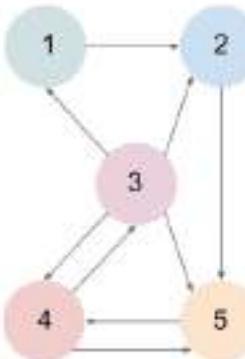
13

• •

University of Southern California 

**USC Viterbi The Simplified PageRank Algorithm**

## Example 1



	Iteration 0	Iteration 1
P <sub>1</sub>	1/5	1/20
P <sub>2</sub>	1/5	5/20
P <sub>3</sub>	1/5	1/10
P <sub>4</sub>	1/5	5/20
P <sub>5</sub>	1/5	7/20

1. Iteration 0: Initialize all pages to have rank  $\frac{1}{5}$ .  
 2. Iteration 1:  
 3. P<sub>1</sub>: has 1 link from P<sub>3</sub>, and P<sub>3</sub> has 4 outbound links, so we take the rank of P<sub>3</sub> from iteration 0 and divide it by 4, which results in rank  $(\frac{1}{5})/4 = 1/20$  for P<sub>1</sub>  

$$\text{PR}(P_1) = (\frac{1}{5})/4 = 1/20$$
  
 4. P<sub>2</sub>: has 2 links from P<sub>1</sub> and P<sub>3</sub>, P<sub>1</sub> has 1 outbound link and P<sub>3</sub> has 4 outbound links, so we take (the rank of P<sub>1</sub> from iteration 0 and divide it by 1) and add that to (the rank of P<sub>3</sub> from iteration 0 and divided that by 4) to get  $\frac{1}{5} + 1/20 = 5/20$  for P<sub>2</sub>  

$$\text{PR}(P_2) = \frac{1}{5} + (\frac{1}{5})/4 = 5/20$$

Copyright Ellis Horowitz, 2011-2022

14

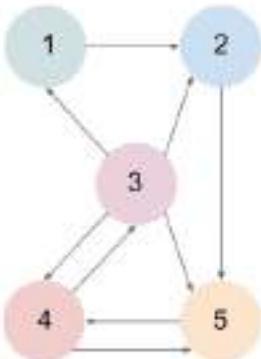
• •

University of Southern California 

**USC Viterbi**  
School of Engineering

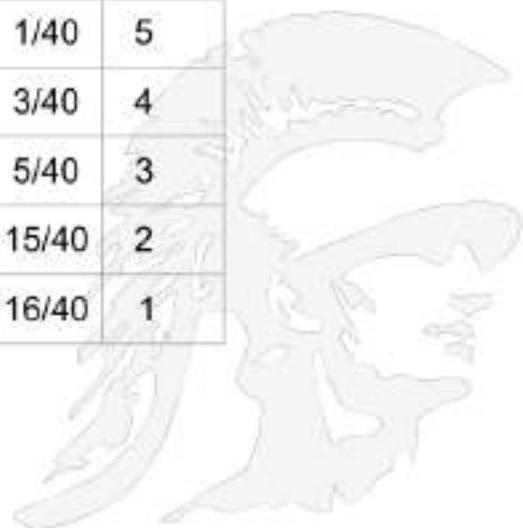
## The Simplified PageRank Algorithm

### Example 1: After 2 iterations



	Iteration 0	Iteration 1	Iteration 2	Final Ranking
P <sub>1</sub>	1/5	1/20	1/40	5
P <sub>2</sub>	1/5	5/20	3/40	4
P <sub>3</sub>	1/5	1/10	5/40	3
P <sub>4</sub>	1/5	5/20	15/40	2
P <sub>5</sub>	1/5	7/20	16/40	1

$PR(P_5) = \frac{1}{5} + \frac{1}{5} * \frac{1}{4} + \frac{1}{5} * \frac{1}{2} = \frac{7}{20}$



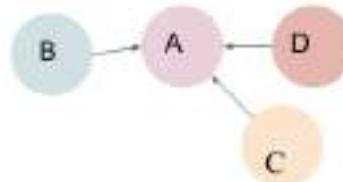
Copyright Ellis Horowitz, 2011-2022

15

• •

## Another Example

### Example 2



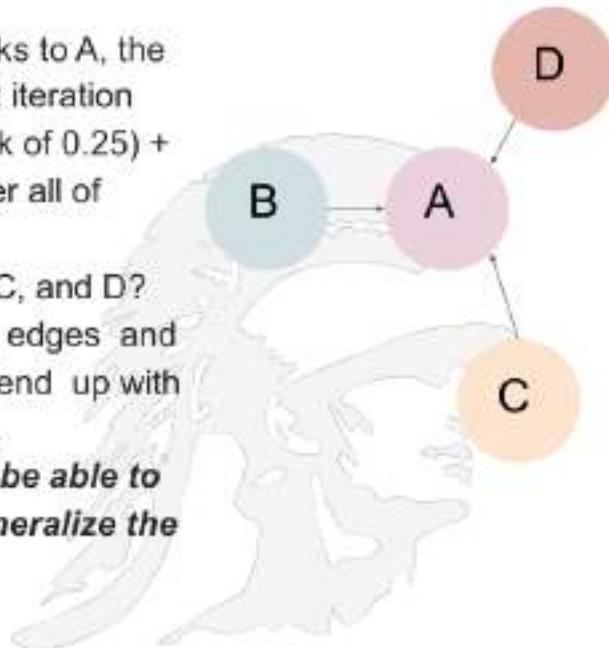
- Say we have four pages: **A, B, C** and **D**.
- Links from a page to itself are ignored (there aren't any in this example).
- Multiple links from one page to another are treated as a single link.
- In this example, every page would start out with a rank of 0.25

••

## Another Example

### Example 2

- Since B, C, and D all have outbound links to A, the PageRank of A will be **0.75** upon the first iteration
  - (B with rank of 0.25) + (C with rank of 0.25) + (D with rank of 0.25) would transfer all of those ranks to A
- But wait! What about ranks of pages B,C, and D? Because B, C, and D have no incoming edges and they give all their rank to A, they will all end up with a rank of 0. This doesn't add up to 1 . . .
  - ***So the simplified algorithm needs to be able to handle border cases, so we must generalize the PageRank algorithm!***



• •

University of Southern California  USC

**Complete PageRank Algorithm**

- Quoting from the original Google paper, PageRank is defined like this:

*"We assume page A has pages T1...Tn which point to it (i.e., are citations). The parameter d is a damping factor which can be set between 0 and 1. We usually set d to 0.85. There are more details about d in the next section. Also C(A) is defined as the number of links going out of page A."*\*<sup>8</sup>

The PageRank of a page A is given as follows:

$$PR(A) = (1-d) + d (PR(T1)/C(T1) + \dots + PR(Tn)/C(Tn))$$
- Note:
  - That the PageRanks form a probability distribution over web pages, so the sum of all web pages' PageRanks will be one.

\*<sup>8</sup>The Anatomy of a Large-Scale Hypertextual Web Search Engine by Brin and Page, <http://infolab.stanford.edu/pub/papers/google.pdf>

Copyright Ellis Horowitz, 2011-2022

18

• •

University of Southern California 

 USC **Viterbi**  
School of Engineering

## Explanation

$$PR(A) = (1-d) + d (PR(T1)/C(T1) + \dots + PR(Tn)/C(Tn))$$

- *PR(A)* is PageRank of Page A (one we want to calculate)
- *PR(T<sub>1</sub>)* is the PageRank of Site T<sub>1</sub> pointing to page A
- *C(T<sub>1</sub>)* is the number of outgoing links of page T<sub>1</sub>
- *PR(T<sub>n</sub>)/C(T<sub>n</sub>)* : If page A has a backlink from page “T<sub>n</sub>” the share of the vote page A will get is “PR(T<sub>n</sub>)/C(T<sub>n</sub>)”
- *d(...)* : All these fractions of votes are added together but, to stop the other pages having too much influence, this total vote is “damped down” by multiplying it by the factor “d”, (0.85)
- *(1-d)* : Since “sum of all web pages’ PageRanks will be one”. Even if the *d(...)* is 0 then the page will get a small PR of 0.15

Copyright Ellis Horowitz, 2011-2022

19

••

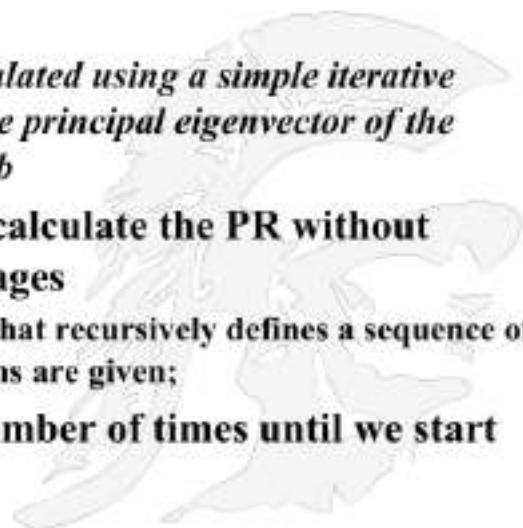
University of Southern California  USC

 USC Viterbi  
School of Engineering

## How PageRank is Calculated

- PR of each page depends on PR of other pages which are pointing to it. But we don't know PR of a given page until the PR of other pages is calculated and so on...
- From the Google paper:

*PageRank or  $PR(A)$  can be calculated using a simple iterative algorithm, and corresponds to the principal eigenvector of the normalized link matrix of the web*
- What this means is that we can calculate the PR without knowing the final PR of other pages
  - Recurrence Relation: an equation that recursively defines a sequence of values once one or more initial terms are given;
- We calculate PR iteratively a number of times until we start converging to the same value.



Copyright Ellis Horowitz, 2011-2022

20

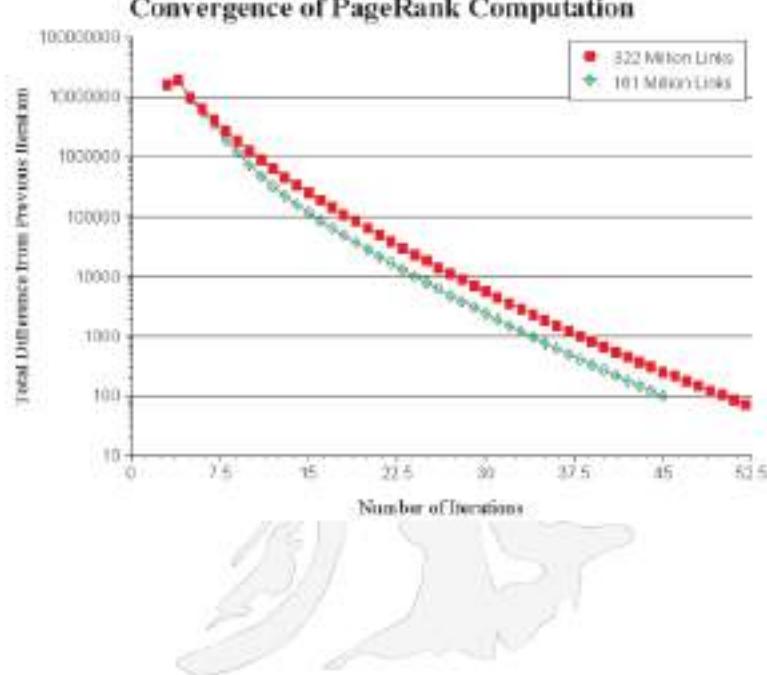
• •

University of Southern California  USC 

## How Fast Does the PageRank Algorithm Converge

- Early experiments on Google used 322 million links
- PR (322 Million Links): 52 iterations
- PR (161 Million Links): 45 iterations
- Number of iterations required for convergence is empirically (but not formally derived)  $O(\log n)$  (where  $n$  is the number of links)
- Hence the calculation is quite efficient

Convergence of PageRank Computation



Number of Iterations	Total Difference (322 Million Links)	Total Difference (161 Million Links)
0	2,000,000	2,000,000
10	100,000	100,000
20	10,000	10,000
30	1,000	1,000
40	100	100
50	10	10

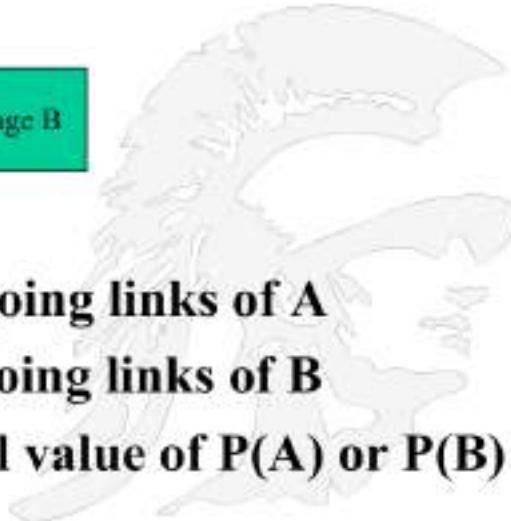
Copyright Ellis Horowitz, 2011-2022

21

••



- Consider 2 pages: Page A and Page B pointing to each other.



- $C(A) = 1$ , number of outgoing links of A
- $C(B) = 1$ , number of outgoing links of B
- What should be the initial value of  $P(A)$  or  $P(B)$  ?

••

University of Southern California 

 USC **Viterbi**  
School of Engineering

## Guess 1:

- Suppose the initial values are :
  - $P(A) = 1$  and  $P(B) = 1$  and  $d = 0.85$ ; then
  - $PR(A) = (1 - d) + d * (PR(B)/1)$
  - $PR(B) = (1 - d) + d * (PR(A)/1)$
- i.e.
- $PR(A) = 0.15 + 0.85 * 1$   
 $= 1$
- $PR(B) = 0.15 + 0.85 * 1$   
 $= 1$
- In one iteration we are done
- Let's try another set of initial values.



Copyright Ellis Horowitz, 2011-2022

• •

University of Southern California 

 USC **Viterbi**  
School of Engineering

## Guess 2 With 3 Iterations

- Initial Values :  $P(A) = 0$ ,  $P(B) = 0$  and  $d = 0.85$
- $PR(A) = (1 - d) + d(PR(B)/1)$
- $PR(B) = (1 - d) + d(PR(A)/1)$
- $PR(A) = 0.15 + 0.85 * 0 = 0.15$
- $PR(B) = 0.15 + 0.85 * 0.15 = 0.2775$

Iterating again we get:

- $PR(A) = 0.15 + 0.85 * 0.2775 = 0.385875$
- $PR(B) = 0.15 + 0.85 * 0.385875 = 0.47799375$

And iterating again

- $PR(A) = 0.15 + 0.85 * 0.47799375 = 0.5562946875$
- $PR(B) = 0.15 + 0.85 * 0.5562946875 = 0.62285048437$



Copyright Ellis Horowitz, 2011-2022

24

••

## Guess 2: Continued...

- After 20 iterations...
- $\text{PR(A)} = 0.99$
- $\text{PR(B)} = 0.99$
- Both approaching to 1.

	A	B	C	D
1	C(A)		1	
2	C(B)		1	
3				
4	Iterations		PR(A)	PR(B)
5				
6	0	0		0
7	1	0.15		0.2775
8	2	0.305875	0.47799375	
9	3	0.566294688	0.622860484	
10	4	0.679422912	0.727509475	
11	5	0.768383054	0.803125596	
12	6	0.8126656756	0.857756243	
13	7	0.879094506	0.91723033	
14	8	0.912645781	0.925748914	
15	9	0.9368886577	0.94635359	
16	10	0.964400552	0.961240469	
17	11	0.967054399	0.971996239	
18	12	0.976196803	0.979767283	
19	13	0.98200219	0.985381862	
20	14	0.987574582	0.989438395	
21	15	0.991022636	0.99236924	
22	16	0.993513854	0.994486776	
23	17	0.99531376	0.996016696	
24	18	0.996614191	0.997122063	
25	19	0.997553753	0.99792069	
26	20	0.998232587	0.998497699	
27				

• •

University of Southern California  USC

**USC Viterbi**  
School of Engineering

## Guess 3:

- Initial Values :  $P(A) = 40$  and  $P(B) = 40$   
 $d = 0.85$
- $PR(A) = (1 - d) + d(PR(B)/1)$
- $PR(B) = (1 - d) + d(PR(A)/1)$
- Notice decreasing value of PR
- The page rank is again approaching to 1.
- So it doesn't matter where you start your guess, once the PageRank calculations have settled down, the "normalized probability distribution" (the average PageRank for these two pages) will be 1.0

	A	B	C	D
1	C(A)		1	
2	C(B)		1	
3				
4	Iterations	PR(A)		PR(B)
5				
6	0	0		40
7	1	34.15		29.1775
8	2	24.950875		21.36824375
9	3	18.30450719		15.70883111
10	4	13.60250644		11.62713048
11	5	10.03306091		8.678101769
12	6	7.526306504		6.547420528
13	7	5.716314249		5.000017112
14	8	4.406814545		3.895792363
15	9	3.461423509		3.092209982
16	10	2.778378485		2.511621712
17	11	2.284878455		2.092146687
18	12	1.928324684		1.789075981
19	13	1.670714584		1.570107397
20	14	1.484591287		1.411902594
21	15	1.350117205		1.297599624
22	16	1.262959681		1.216015728
23	17	1.1827633969		1.156348864
24	18	1.132046534		1.112238654
25	19	1.095403621		1.081093078
26	20	1.068929116		1.059599749
27				

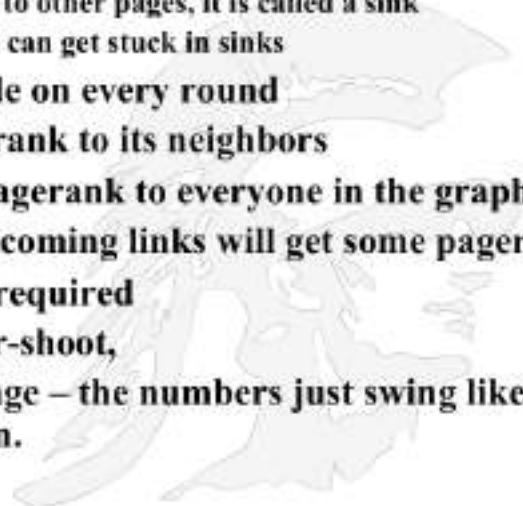
• •

University of Southern California 

**USC Viterbi**  
School of Engineering

## PageRank Convergence?

- Some observations about the damping factor
- The damping factor value and its effect:
  - For certain graphs the simple update rule can cause pagerank to accumulate and get stuck in certain parts of the graph
    - E.g. if a page has no outgoing links to other pages, it is called a sink
    - The simplified pagerank algorithm can get stuck in sinks
  - This is fixed by having each node on every round
    - Give a  $d$  fraction of its pagerank to its neighbors
    - Give a  $(1-d)$  fraction of its pagerank to everyone in the graph
    - As a result, pages with no incoming links will get some pagerank
  - If too high, more iterations are required
  - If too low, you get repeated over-shoot,
  - Both above and below the average – the numbers just swing like pendulum and never settle down.



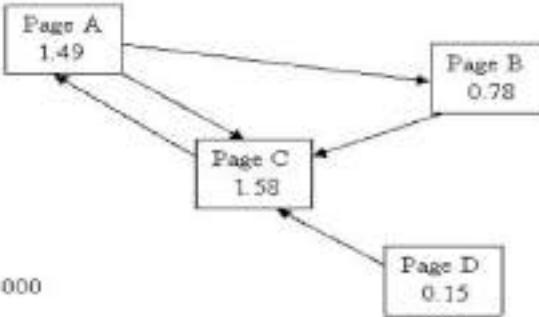
Copyright Ellis Horowitz, 2011-2022

27

••

University of Southern California  USC Viterbi School of Engineering

## Examples of Relative PageRanks 1:



The diagram illustrates a network of four pages (A, B, C, D) connected by directed links. Page A has a PageRank of 1.49 and points to Page B and Page C. Page B has a PageRank of 0.78 and points to Page C. Page C has a PageRank of 1.58 and points to Page D. Page D has a PageRank of 0.15. Below the diagram, the text "Average PR: 1.000" is displayed.

- **Observations:**
  - Every page has at least a PR of 0.15 to start out
  - Page D has no votes but still it has a PR of 0.15
  - It is believed that Google undergoes a post-spidering phase whereby any pages that have no incoming links at all are ignored wrt PageRank
- Examples on the following pages are taken from <http://www.sirgroeane.net/google-page-rank/>

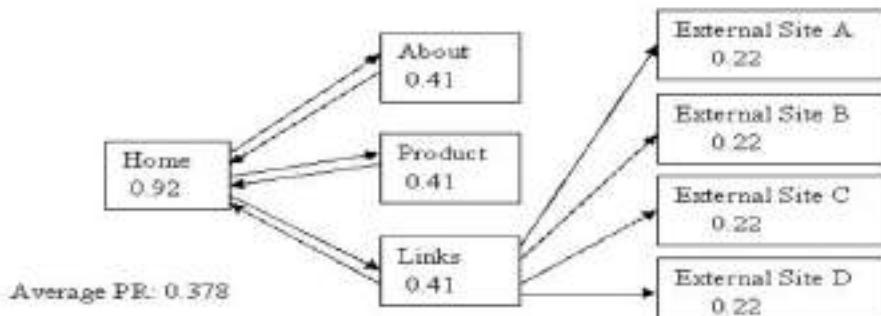
32

• •

University of Southern California  USC

USC Viterbi School of Engineering

### Example 2: Simple hierarchy with Some Outgoing Links



Average PR: 0.378

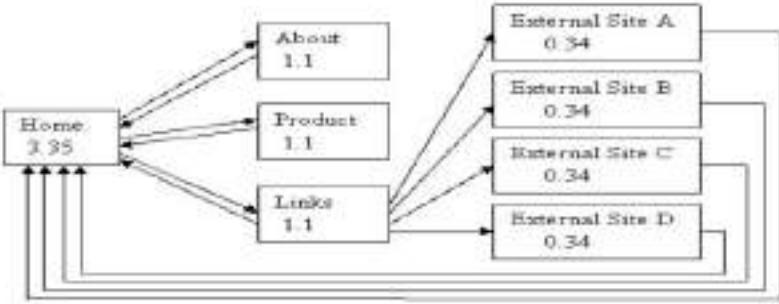
- Observations:**
  - Home has the most PR
  - But average PR is 0.378
    - “External site” pages are not voting for anyone.
  - Links within your own site can have a significant effect on PageRank.

33

• •

University of Southern California  USC

**Example 3: Linking External Sites Back into our Home Page**



The diagram shows a network of pages. A central 'Home' page with a PR of 3.95 links to three internal pages: 'About' (PR 1.1), 'Product' (PR 1.1), and 'Links' (PR 1.1). These three pages link to four external sites: 'External Site A' (PR 0.34), 'External Site B' (PR 0.34), 'External Site C' (PR 0.34), and 'External Site D' (PR 0.34). Arrows indicate the direction of links between pages.

Average PR: 1.000

- **Observations:**
  - Comparing this example with the previous one, now the PageRank of the Home Page has increased significantly.
  - Moreover, the average PageRank is better because of all the external sites linking back to the home page.



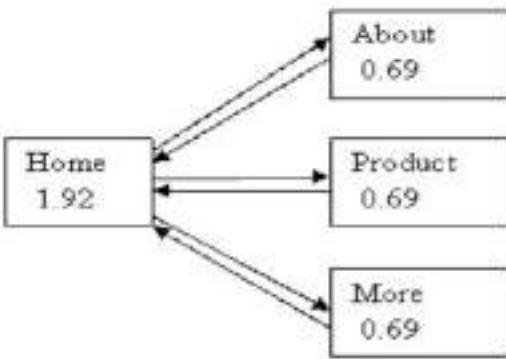
34

••

University of Southern California 

 USC **Viterbi**  
School of Engineering

## Example 4: Simple Hierarchy



The diagram illustrates a simple hierarchy of four pages: Home, About, Product, and More. The Home page has a PageRank of 1.92. It links to three child pages: About, Product, and More, each of which has a PageRank of 0.69. Arrows indicate the direction of links from the Home page to its children.

- **Observations:**
  - Home Page has PageRank of 2.5 times the page rank of its child pages.
  - A hierarchy structure concentrates votes and PageRank into one page.

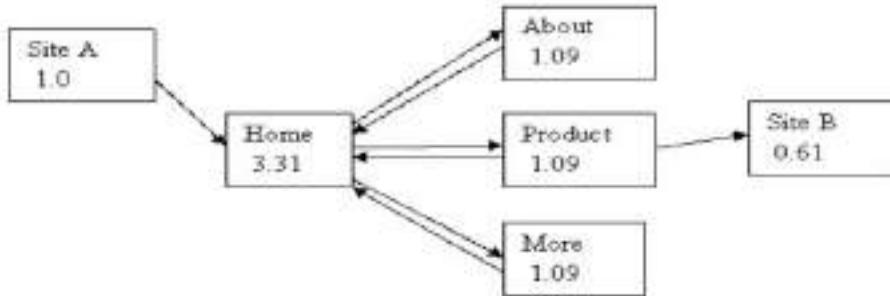
35

• •

University of Southern California 

 USC **Viterbi**  
School of Engineering

### Example 5: Hierarchical – But with One Link In and One Out



```
graph LR; SiteA[Site A
1.0] --> Home[Home
3.31]; SiteA --> About[About
1.09]; SiteA --> Product[Product
1.09]; SiteA --> More[More
1.09]; SiteB[Site B
0.61] --> Product;
```

- **Observations:**
  - The PageRank of Home page has increased from 1.92 (Previous Example) to 3.31
  - Site A contributed 0.85 PR to Home page and the raised PageRank in the “About”, “Product” and “More” pages has had a lovely “feedback” effect, pushing up the home page’s PageRank even further!
  - A well structured site will amplify the effect of any contributed PR.

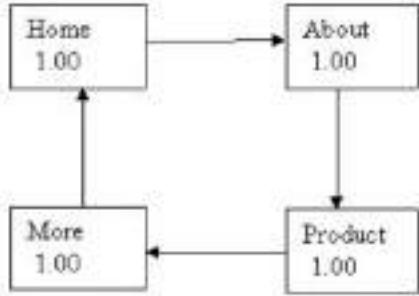
36

• •

University of Southern California 

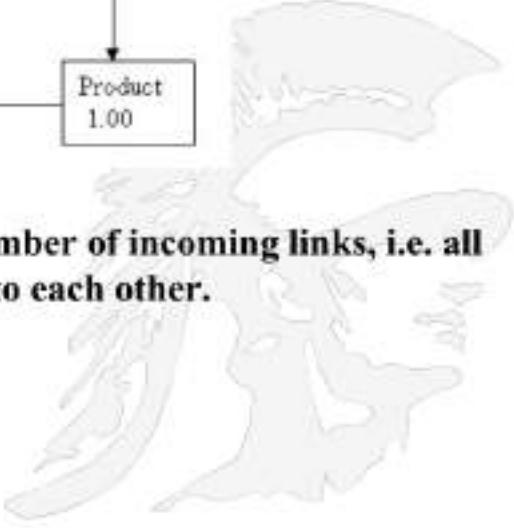
 USC **Viterbi**  
School of Engineering

## Example 6: Looping



```
graph TD; Home[Home
1.00] --> About[About
1.00]; About --> Product[Product
1.00]; Product --> More[More
1.00]; More --> Home;
```

- **Observations:**
  - All the pages have the same number of incoming links, i.e. all pages are of equal importance to each other.
  - Each page has PR of 1.0
  - Average PR is 1.0

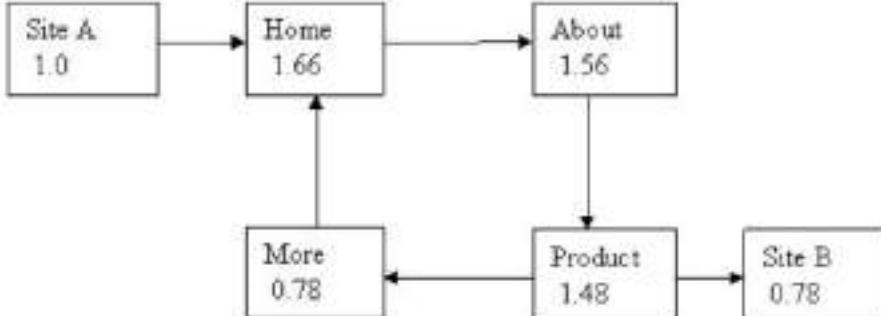


37

••

University of Southern California  USC

### Example 7: Looping – But with a Link in and a Link Out



```
graph LR; SA[Site A
1.0] --> H[Home
1.66]; H --> A[About
1.56]; A --> M[More
0.78]; M --> P[Product
1.48]; P --> SB[Site B
0.78]; SB --> M; P --> H
```

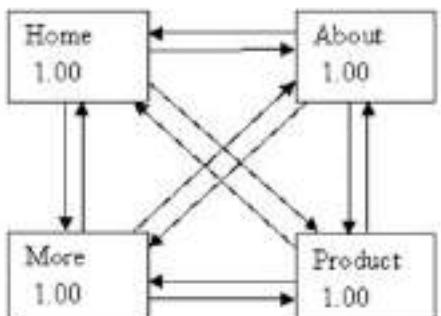
- **Observations:**
  - PageRank of our home page has gone up a little, but what's happened to the “More” page? Its PageRank has gone down
  - Now the PageRank value of the external Site B is equal to the “More” page.
  - The vote of the “Product” page has been split evenly between “More” page and the external site B.
  - This is good for Site B for otherwise its PageRank would be 0.15

38

• •

University of Southern California 

**Example 8: Extensive Interlinking or Fully Meshed**



```
graph TD; Home[Home
1.00] <--> About[About
1.00]; Home <--> More[More
1.00]; Home <--> Product[Product
1.00]; About <--> More; About <--> Product; More <--> Product;
```

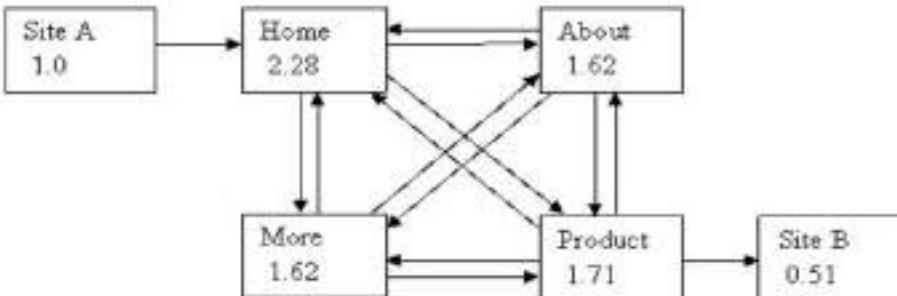
- **Observations:**
  - All pages are of equal importance to each other so the resulting PageRank is no different than what was gotten in Example 6

39

••

University of Southern California 

**Example 9: Fully Meshed – But with One Vote in and One Vote Out**



The diagram illustrates a fully meshed network of pages. There are six nodes: Site A (1.0), Home (2.28), About (1.62), More (1.62), Product (1.71), and Site B (0.51). Every node has a bidirectional arrow connecting it to every other node. The initial PageRank values are: Site A (1.0), Home (2.28), About (1.62), More (1.62), Product (1.71), and Site B (0.51). The arrows represent the flow of votes from one page to another.

- **Observations:**
  - “Product” page has kept three quarters of its vote within our site unlike example 9 where it was giving away fully half of it’s vote to the external site!
  - Keeping just this small extra fraction of the vote within our site has had a very nice effect on the Home Page too.

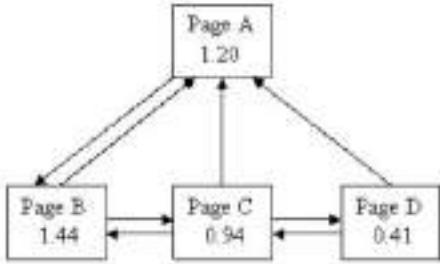
40

••

University of Southern California 

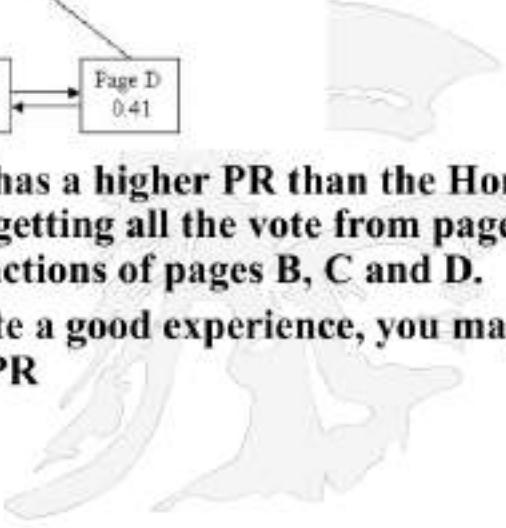
 USC **Viterbi**  
School of Engineering

**Example 10: Previous ... Next ... Documentation Page Layout**



```
graph TD; A[Page A
1.20] --> B[Page B
1.44]; A --> C[Page C
0.94]; A --> D[Page D
0.41]; B <--> C
```

- **The first page of the document has a higher PR than the Home Page! This is because page B is getting all the vote from page A, but page A is only getting fractions of pages B, C and D.**
- **In order to give users of your site a good experience, you may have to take a hit against your PR**

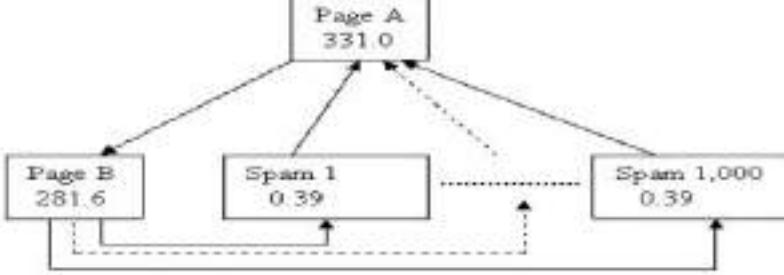


41

• •

University of Southern California  USC

**Example 11: Getting Higher PR the Wrong Way!**



- **Observations:** Average PR: 1.000
  - 1000 incoming links and only one outgoing link
  - It doesn't matter how many pages you have in your site, your average PR will always be 1.0 at best.
  - But a hierarchical layout can strongly concentrate votes, and therefore the PR, into the home page!
  - This is a technique used by some disreputable sites

A **link farm** is set of web pages created with the sole aim of linking to a target page, in an attempt to improve that page's search engine ranking. 42

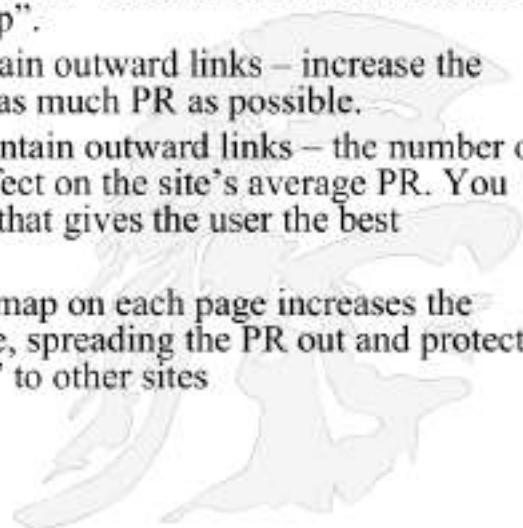
• •

University of Southern California  USC

 USC **Viterbi**  
School of Engineering

## Some Suggestions Based on What We Have Seen in Examples.

- **Suggestions for improving your page rank**
  - Increasing the internal links in your site can minimize the damage to your PR when you give away votes by linking to external sites.
  - If a particular page is highly important – use a hierarchical structure with the important page at the “top”.
  - Where a group of pages may contain outward links – increase the number of internal links to retain as much PR as possible.
  - Where a group of pages do not contain outward links – the number of internal links in the site has no effect on the site’s average PR. You might as well use a link structure that gives the user the best navigational experience.
  - **Use Site Maps:** Linking to a site map on each page increases the number of internal links in the site, spreading the PR out and protecting you against your vote “donations” to other sites



Copyright Ellis Horowitz, 2011-2022

43

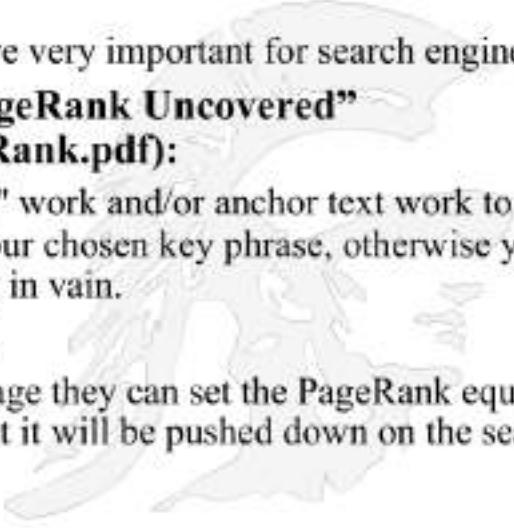
••

University of Southern California 

 USC **Viterbi**  
School of Engineering

## Importance of PageRank

- **PageRank is just one factor Google uses to determine a page's relevance. It assumes that people will link to your page only if they think the page is good. But this is not always true.**
- **Content is still the king!!!**
  - Anchor, body, title tags etc. still are very important for search engines
- **From Chris Ridings' Paper, "PageRank Uncovered"** (<http://www.voelspriet2.nl/PageRank.pdf>):
  - You must do enough "on the page" work and/or anchor text work to get into that subset of top pages for your chosen key phrase, otherwise your high PageRank will be completely in vain.
- **PageRank is a multiplier factor.**
  - If Google wants to penalize any page they can set the PageRank equal to a small number, even 0. As a result it will be pushed down on the search results page.



Copyright Ellis Horowitz, 2011-2022

44

..

You can play with PR here:

[https://bytes.usc.edu/~saty/tools/xem/run.html?  
x=PR](https://bytes.usc.edu/~saty/tools/xem/run.html?x=PR)

1/44

\*\*\* 09:24:41

# **Snippets**

## **[plain, featured, 'rich']**

• •

## Snippets in Google Search

USC Viterbi  
School of Engineering

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

Google what is the number of x- and y-intercepts that quadratic functions may have

Web Results 1 - 10 of about

**Pre-Calculus Advanced >> Quadratic Functions >> Intercepts, Zeroses ...** [+] [X]  
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 ...  
[www.wad1.org/wad1/math/Pre-Calculus%20Advanced/Quadratic%20Functions/Intercepts/Intro.htm](http://www.wad1.org/wad1/math/Pre-Calculus%20Advanced/Quadratic%20Functions/Intercepts/Intro.htm) - 9k - Cached - Similar pages - ↗

**Yahoo! Canada Answers - What is the number of x- and y-intercepts ...** [+] [X]  
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. ...  
[answers.yahoo.ca/question/index?qid=20080428215000AAUJ0G1](http://answers.yahoo.ca/question/index?qid=20080428215000AAUJ0G1) - 38k - Cached - Similar pages - ↗

**Quadratic Functions(General Form)** [+] [X]  
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 is given by ... Use the applet window to check the y intercept for the quadratic ...

USC

• •

## Some Elementary Facts About Google Snippets

USC Viterbi  
School of Engineering

- 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

USC

••

## Google Snippet Lifecycle Changes

USC Viterbi  
School of Engineering

### Classic snippet



### Adding images in 2016

Images are relevant to the query  
Placement on right  
Images are downloadable or shareable  
Galleries support content-rich pages  
Users visit a greater diversity of sites



### Adding videos in 2018

Video is relevant to the query  
Preview indicates if video is dominant or supportive  
Video metadata informs user experience



**USC**

••

## Google Snippet Lifecycle Changes

**USC Viterbi**  
School of Engineering

### Adding Sitelinks to snippets

**Sitelinks In Previews**

Links are relevant to query  
Links extracted algorithmically

- Sitelinks, Site-structure
- Drives traffic from a diverse set of sites

Sitelink+images helps users  
- Why 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

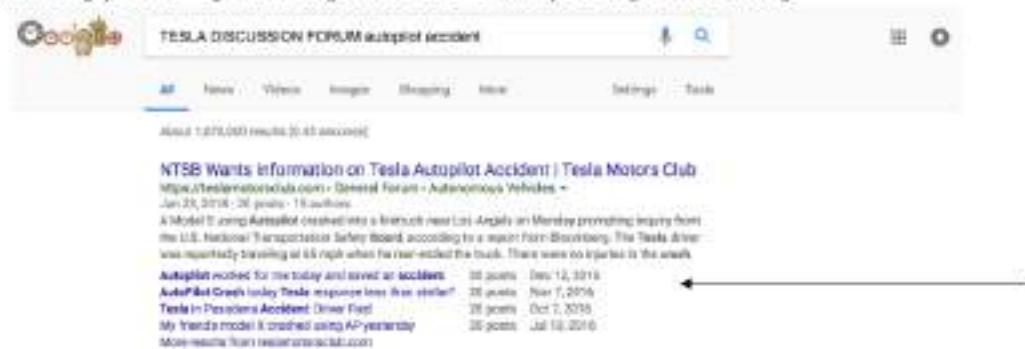




Copyright Ellis Horo

## More Facts on Snippets

- **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]"



- **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**

**Result for the query "what cholesterol levels mean"**

**A long snippet, and a PAA**

**Google uses the meta description**

**A different query "cholesterol Cleveland Clinic" produces the same first result but a different snippet**

The top screenshot shows the search results for "what cholesterol levels mean". It includes a snippet from the Cleveland Clinic website: "HDL (high-density lipoprotein) cholesterol is also called "good" cholesterol. This protein carries most of the cholesterol in your blood. The cholesterol carried out in your bloodstream in HDL cholesterol is used to repair arteries. Your total cholesterol numbers fall into one of three categories: high (normal) if it is less than 200 mg/dL; intermediate if it is between 200 and 240 mg/dL; and high (abnormal) if it is more than 240 mg/dL." Below the snippet is a People Also Ask box with questions like "What is a normal cholesterol level?" and "What is the normal range for cholesterol levels?".

The bottom screenshot shows the search results for "cholesterol Cleveland Clinic". The top result is from the Cleveland Clinic website with the title "Understanding Your Cholesterol Numbers | Cleveland Clinic". The snippet reads: "Understanding Your Cholesterol Numbers | Cleveland Clinic https://my.clevelandclinic.org/health/.../11162-cholesterol-numbers-what-do-they-mean... When trying to interpret your cholesterol numbers, terms like HDL and LDL can be very confusing. Your total blood cholesterol is a measure of the cholesterol components LDL (low-density lipoprotein) cholesterol, HDL (high-density lipoprotein) cholesterol, and VLDL (very low density lipoprotein).

**USC**

Copyright Ellis Horowitz 2012 - 2022

••

## Snippets are an Instance of Summarization

USC Viterbi  
School of Engineering

- **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



Copyright Ellis Horowitz 2012-2022

••

## Featured Snippets

USC Viterbi  
School of Engineering

- 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 .*



• •

## Three Types of *Featured Snippets*

**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 search results for "aviation jobs". The results page shows a table with 15 job listings. The table has three columns: Job Title, Company, and Job Details. The first few rows show titles like "Project Engineer - NEX", "Project Engineer", "Aviation Express", and "A & P Services".

Latest 15 Job Listings		
Job Title	Company	Job Details
Project Engineer - NEX	aerotek.com	Project Engineer - NEX
Project Engineer	aviationjobs.com	Project Engineer
Aviation Express	aviationexpress.com	A & P Services
A & P Services	aviationexpress.com	Necessary Shop Technican
Necessary Shop Technican	aviationexpress.com	

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

For a fourth type see: <https://www.semrush.com/blog/featured-snippets/>

**USC**

••

## Modifying Your Page to Produce a *Featured Snippet*

USC Viterbi  
School of Engineering

- 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>

USC

• •

## TOO LONG DIDN'T READ (TLDR) Internet slang

**USC Viterbi**  
School of Engineering

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



**USC**

Copyright E

••

## Extracting a Snippet is Not Always Easy

USC Viterbi  
School of Engineering

The image shows three separate Google search results for the query "tesla announces quarter".

- Top Result:** A snippet from the New York Times titled "Apple and Tesla to Report Earnings" is shown. It includes a photo of Elon Musk and a short summary: "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...".
- Middle Result:** A snippet from the New York Times titled "A loss is likely for Tesla, despite popularity." is shown. It includes a photo of Elon Musk and a short summary: "Tesla Motors' finances will come into focus on Wednesday when it reports its earnings for the third quarter. The company previously disclosed that deliveries rose 20 percent to 24,500 cars in the July-to-September period. But a loss is likely because the company is spending heavily to introduce a new car, the Model 3, and to start up its gigantic Gigafactory in Nevada, where it will produce batteries. Investors will listen closely to Tesla's chief executive, Elon Musk, and any details he offers on the pending merger with...".
- Bottom Result:** A snippet from the New York Times titled "Economists predict rise in orders for durable goods." is shown. It includes a photo of a chart and a short summary: "In this issue, 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 low volatile core capital goods category is thought to have declined by 0.1 percentage point, or maintained near several quarters in terms of new investment, especially in the industrial sector. Nilsen D. Solberg."

**Annotations:**

- A callout points to the top result with the text: "query: 'Tesla reports financial results'".
- A callout points to the middle result with the text: "one search result and its snippet".
- A callout points to the bottom result with the text: "portions of the article used to create the snippet; note how long the article is; 'financial results' equates to 'earnings'".
- A callout points to the bottom result with the text: "Ellis".

• •

## Extracting a Snippet is Not Always Easy Nor Obvious

**USC Viterbi**  
School of Engineering

query: "cloud computing"

one search result and its snippet

"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

**EXPLAINTHATSTUFF!**

Cloud computing - A simple introduction - Explain that Stuff

www.explainthatstuff.com/cloud-computing-introduction.html

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...

**Cloud computing - A simple introduction**

www.explainthatstuff.com/cloud-computing-introduction.html

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...

Copyright Ellis Horowitz, 2014

USC

••

## How Does Google Generate Snippets?

One way to find out is to go to [patents.google.com](https://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

[Abstract](https://patents.google.com/patent/US20120137111A1) (1,744 results, ordered by relevance) > [grouped by classifier] >

Applicant: Google Inc. > Assignee: Google Inc.

Priority: 2010-05-04 > Filed: 2010-10-08 > Published: 2012-05-17

Generating snippets for patent users for information retrieval queries...  
Abstract: This invention relates to generating snippets for patent users for information retrieval queries. Information retrieval queries, also referred to as search queries, typically include retrieving target query data. The target query data identifying one or more target queries and one or more sets of subject query terms, each set of subject query terms being associated with a respective...

Excerpted snippets: A system provides a list of search results, where one of the...

Applicant: Google Inc. > Assignee: Google Inc.

Priority: 2008-03-01 > Filed: 2008-03-01 > Published: 2009-07-22

Patented 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...

Invisible length snippets generation: A method and system are disclosed that...

Applicant: Google Inc. > Assignee: Google Inc.

Priority: 2004-06-09 > Filed: 2004-06-09 > Published: 2005-03-15

A method and system are disclosed that provide a variable length snippet when querying a database in response to a search request. Under conditions where the search query matches a document with a high degree of certainty, a shorter snippet is provided when the document does not match the...

Local Search Using Address Completion: A local search server receives...

Applicant: Google Inc. > Assignee: Google Inc.

Priority: 2003-05-01 > Filed: 2003-05-02 > Published: 2003-09-19

The system and computer program product for the computer a search engine interface module for displaying snippets related to documents listed by documents terms and returning alternatives about the business, and a snippet analysis module for grouping the snippets to determine the relevance...

Document search engine including highlighting of confident results: A search...

Applicant: Google Inc. > Assignee: Google Inc.

Priority: 2002-09-04 > Filed: 2002-09-04 > Published: 2003-07-16

One method of improving the user's confidence associated with a particular link is to allow display of "snippets" of text with the link. Ideally, the snippets of text should encompass the portion of the text in proximity, the snippets are typically shown front end of the document addressed by the link, although not...

System and method for personalized snippet generation: Snippets of text...

Applicant: Google Inc. > Assignee: Google Inc.

Priority: 2002-04-14 > Filed: 2002-04-14 > Published: 2003-07-16

Snippets of text provided are generated based in part on user profiles. An item such as a document is associated with identity information stored in the user's profile. A query profile is selected...

About | Help | Feedback | Terms | Privacy Policy

Copyright Ellis Horowitz 2012 - 2022

••

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

**USC Viterbi**  
School of Engineering

### Title:

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

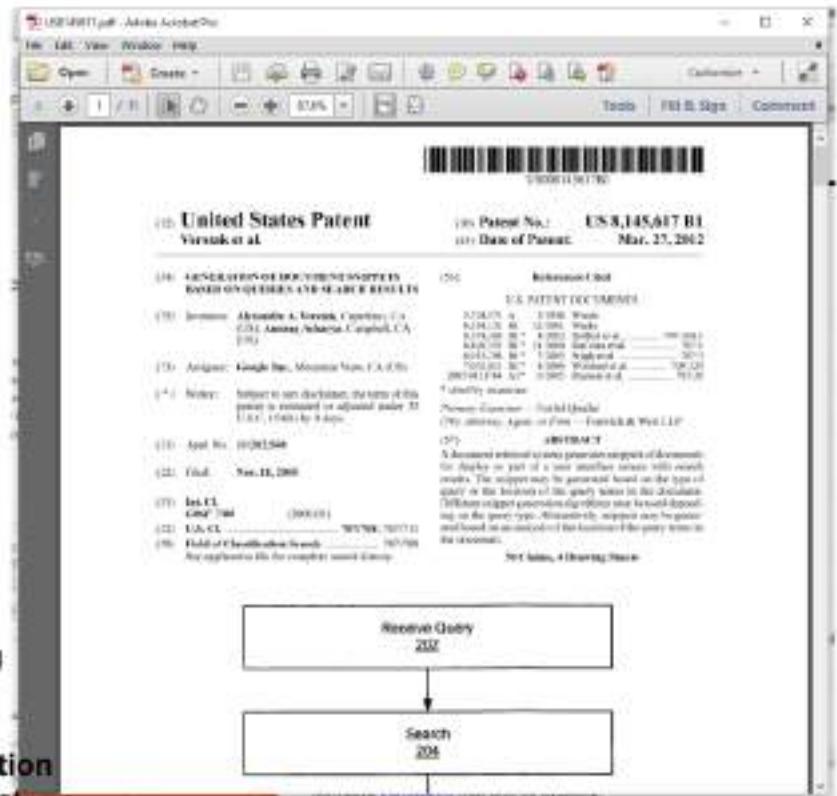
**filed:** 2005

**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**.



**USC**

Copyright Ellis Horowitz 2012 - 2022

• •

## Some Guidelines for Snippet Generation

USC Viterbi  
School of Engineering

- **Location Based Rules**

- based on the location of the query terms in page. A paragraph or a portion of a paragraph might be chosen as search results snippets based on the length and distance of the paragraph from the start or end of pages.
- Every paragraph that includes the query terms is given a score based on the length of the paragraph and the distance of the paragraph from a predetermined location in the document, such as the beginning or the end of the document.
  - documents that include abstracts, executive summaries or comprehensive introductions are identified and used to create a snippet
  - Similarly, the ends of pages can be used if they include a conclusion or summarization

- **Language Dependent Rules**

- How much of the paragraph are punctuation characters
- Whether the paragraph ends with punctuation or proposition
- Whether any of the words in the paragraph is overly long
- The number of bold or italicized words in the paragraph

- **Rejection rules**

- 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

USC

••

## US Patent 8,145,617 Defines an Algorithm for Snippet Generation

USC Viterbi  
School of Engineering

- **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***



Copyright Ellis Horowitz 2012 - 2022

US Patent 8,631,006  
Snippets can be based on a User's Profile

**USC Viterbi**  
School of Engineering

## *System and Method for Personalized Snippet Generation*

Filed: April 14, 2005

Awarded: Jan. 14, 2014

## Abstract

*S*nippets 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.



US2007003451

United States Patent		Patent No.:	US 8,631,806 B1
		Date of Patent:	Jun. 14, 2014
(43) Inventor:	Jesse H. Hwang, Mountain View, CA (US); Siparad D. Karmakar, San Francisco, CA (US)	(10) Application No.:	13/812,981
(11) Filed:	2013-03-04	(21) Art Unit:	3200
(22) Prior Art Date:	2008-04-04	(31) Examiner:	Harris et al.
(33) Publication:	2013-07-11	(43) Assignee:	Google Inc., Mountain View, CA (US)
(45) Related:	Subject to any disclaimer, the term of this patent is extended or adjusted under 35 U.S.C. 154(b).	* cited by examiner	

[View all posts by jessica@jessicabrown.com](#)

Primary consumer - Agro Biodiversity

Journal Economy – Issued without  
regarding to page – 35

1740, January 1898, or New — Morgan, Lewis & Bostwick  
LLP

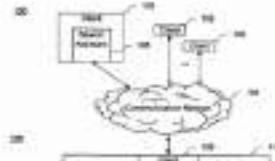
111

卷之三

卷之三

Supports of *not* associated are generated based on an analysis of the user's profile. An *issue*, such as an *incorrect* or *possibly incomplete* location related to the user's profile, is a *non-profile* support. The score of an *ideal* location is compared to a *user's* profile. The measure of *possibly incorrect* locations is based on the *user's* profile, that is, *highlights* a *clarification* answer will be *highlights*. Another *ideal* location is a *document* answer. This is a *profile* score which may be also obtained by looking for *issues* up to the user's profile. Towards having high profile similarity scores or high user profile similarity scores, more *associations* are identified supporting which may be relevant in a test. The *highlighting* terms may be extended to search terms and provided to a *multiple* *passenger*.

116 *American Ethnologist*



Copyright Ellis Horowitz 2012 - 2022

USC

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

**USC Viterbi**  
School of Engineering

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	78867	48,87%
2	30618	18,74%
3	26878	12,78%
4	14469	8,85%
5	9582	5,86%
6	2860	1,75%
7	1909	1,17%
8	1119	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



How does Google generate snippets

All Videos News Images Shopping Maps Settings Tools

About 0.000,000 results (0.01 seconds)

**Rich Snippets & Star Ratings | Display Reviews & Increase CTR**  
[www.bytes.com/RichSnippet](#) + (44) 659 6591  
 Get high-quality traffic from your past & regular search results. Get a boost!

**Anatomy Of A Google Snippet - Search Engine Land**  
[www.searchengineland.com/anatomy-of-a-google-snippet-90327](#) +  
 (11/16/2011) This is a crucial important detail: snippets are determined query-based in other words, they vary depending on the keyword being searched (e.g., price, size, name). Don't just try to be gathered together – ensure it would otherwise be unlikely that a Google-generated snippet could capture all of this.

**How Google Might Generate Snippets for Search Results**  
[www.bytes.com/search-engine-optimization/8900](#) +  
 (4/6/2013) A document retrieval system generates snippets of documents for display as part of a user interface screen with search results. The snippet may not generate 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're invited! (no rating) Last visit: 10/20/17

**Google's Featured Snippets: How to Get Your Content to Appear**  
[contentmarketinginstitute.com/2017/03/google-featured-snippets/](#) +  
 (Mar 21, 2017) The primary thing you need to understand about featured snippets is that they do not appear in Google's complex ranking system. They are showing the top organic results for a query. Because of this, it's critical to keep the standard ranking requirements in mind as you create featured content. (no rating)

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**  
[http://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](#)

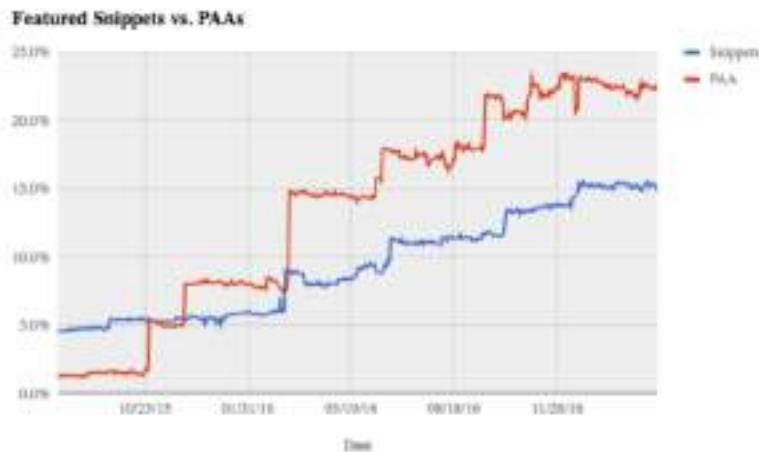
**PAA result for query "how does google generate snippets"**

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

Copyright Ellis Horowitz 2012 - 2022

## 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

• •

## Rich Snippets

**USC Viterbi**  
School of Engineering

- 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  Advanced search

About 277,000 results (0.20 seconds)

**Club Deluxe - Pizza & Jazz Club**  
[Place info](#)

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  
[22 reviews - Write a review](#)

**Club Deluxe - Haight-Ashbury - San Francisco, CA** 

Price range: \$\$  
214 Reviews of Club Deluxe: "This is like my own little hidden GEM in the Height. 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

**USC**

••

## Rich Snippets Examples: People Snippets





Search

About 13,400 results (0.26 seconds) Advanced search

Everything More

**Pravir Gupta | Facebook** 

Friends: Sam Tyagi, Genta Shresth, Sibashis Jain, Shreetha Balasubramanian, 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/pravirgupta](http://www.facebook.com/pravirgupta) · Cached · Similar

**Home (pravir)** 

Pravir Gupta ... attachment removed by Pravir Gupta · edited by Pravir Gupta ... shared by Pravir Gupta · Home · created by Pravir Gupta ...  
[pravirgupta.com](http://pravirgupta.com) · Cached · Similar

**Pravir Gupta - Knol, a unit of knowledge** 

Pravir Gupta · Verify Name: Agra, India · Public activity level: Set by: ... by Pravir Gupta · View · are continuously looking at enabling sites ...  
[knol.google.com/k/pravir-gupta/controversy#0](http://knol.google.com/k/pravir-gupta/controversy#0) · Cached · Similar

**The Journey is the Reward - a knol by Pravir Gupta** 

Jul 29, 2009 ... Debut novel by Anil Kumar Gupta which was published in July 2009.  
[knol.google.com/k/pravir-gupta/the-journey-is-the-reward#0](http://knol.google.com/k/pravir-gupta/the-journey-is-the-reward#0) · Cached

**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/21/8/a78](http://www.linkedin.com/pub/pravir-gupta/21/8/a78) · 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/search/dir?keywords=Pravir+Gupta](http://www.linkedin.com/search/dir?keywords=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/links/PravirGupta](http://www.spoke.com/info/links/PravirGupta) · Cached

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











Rich Snippets Examples:  
Events

**USC Viterbi**  
School of Engineering

Google

filmore events

Search

About 1,610,000 results (0.21 seconds) Advanced search

[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

**Carina Chocolate Drops** Thu, Jun 24  
**Josh Ritter & the Royal City Band** Thu, Jun 24  
**Robert Earl Keen** Sat, Jun 26

[Fillmore Events - Events in Fillmore, California](#)  
Fillmore Events Directory. Includes listings for Events in Fillmore, California.  
[www.californiacast-worldwide.com/Fillmore/Events/](http://www.californiacast-worldwide.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, 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 Games, ...  
[events.nydailynews.com/new-york-city/events/fillmore+events](http://events.nydailynews.com/new-york-city/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

• •

## Advantages of Rich Snippets



### 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.



• •

## A Joint Effort by Google, Yahoo! And Bing

**USC Viterbi**  
School of Engineering

- 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 Pfeiffer | June 1, 2011 | 2 Comments



(Visited and last updated on May 2011)

schema.org

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

In the Google announcement, Kevin Systrom and Perry Cohen of Google's search team say: "Historically, you've supported three different standards for structured data: microformat, microdata, and RDFa. We're decided to focus on just one format for schema.org to create a simpler way for webmasters and to improve consistency across search engines relying on the data."

From the Viterbi announcement comes this: "Today's announcement offers tremendous opportunity for growth. In addition to consolidating the schema.org for microdata 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 schema.org in regular time. Don't worry if your site has already added RDFa or microformats currently supported by our Enhanced Snippet program, that site will 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 schema.org, and feel strongly that by partnering with Google and Yahoo! we maintain schema.org's momentum to move efficiently with the time their invest... Bing accepts a wide variety of markup formats today (Open Graph, microformats, etc.) for features like Tiles and will continue to do so... but by standardizing on schema.org, we are looking at simplifying the markup choices for webmasters and simplifying the value they receive in return."

The schema.org site "provides a collection of schemas, i.e., data tags, that webmasters can use to markup their pages to be recognized by major search providers. Search engines...only use this markup to improve the quality of search results; making it easier for people to find the right web pages."

•

# Rich Snippet Technology Definitions

**USC Viterbi**  
School of Engineering

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



<https://www.w3.org/TR/microdata/>  
Now goes to  
<https://html.spec.whatwg.org/multipage/>



- 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>

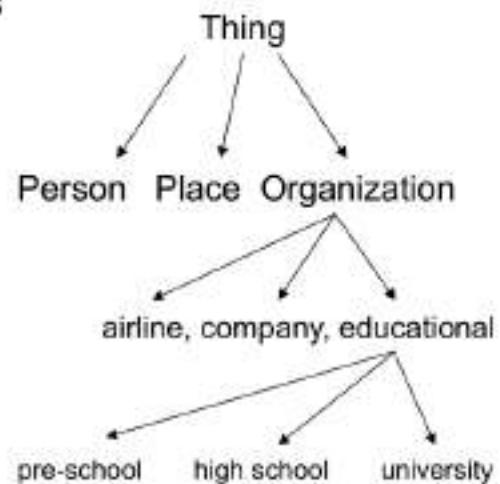
**USC**

••

## Schema.org Vocabulary

USC Viterbi  
School of Engineering

- 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



USC

• •

## Entities in *Rich Snippet Encodings*



### 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



• •

## Rich Snippets



- **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

USC Viterbi  
School of Engineering

- **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 itemprop 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.**

USC

• •

## Avatar Example Continued



- 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**

Property	Supported Type	Description
<code>description</code>	Text	A short description of the item.
<code>image</code>	URL	URL of an image of this 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>arrivedAt</code>	Place	Specifies the Place that is usually accompanied by the CreativeWork.
<code>appropriateFor</code>	AgeGroup	The overall rating, based on a collection of reviews or ratings, of the item.
<code>alternateName</code>	Text	An auxiliary title of the CreativeWork.
<code>arrivesAt</code>	Place	The meeting place that receives this creative work. This property is a connector for <code>arrivedAt</code> .
<code>author</code>	Person	The author of the content. Please note that author is special in the HTML. It provides a special mechanism for indicating authority via the rel tag. This is equivalent to <code>creator</code> and may be used interchangeably.
<code>content</code>	Text	Or, for a person, by the person to whom the content such as American by the person to whom the content such as George Washington, George Bush, yourself.
<code>contentUrl</code>	LinkContainer	Comments, typically from users, on the CreativeWork.
<code>contributor</code>	Person	The location of the content.
<code>copyrightHolder</code>	Text	Official name of a group or person—for example, WPA, FBI, CIA.
<code>director</code>	Person	A director/producer of the CreativeWork.
<code>producer</code>	Person	The party holding the legal copyright to the CreativeWork.

**Movie > schema.org/Movie**

Property	Number	Description
<code>copyrightYear</code>	Integer	The year during which the claimed copyright for this CreativeWork was first asserted.
<code>director</code>	Person or Organization	The creator/producer of the CreativeWork, or its Creators. This is the same as the <code>Author</code> property for CreativeWork.
<code>duration</code>	Date	The date on which the CreativeWork was released.
<code>endDate</code>	Date	The date on which the CreativeWork was most recently updated.
<code>isAccessibleForFree</code>	Date	Date of first free availability.
<code>isContentOf</code>	URL	A link to the page comprising the contents of the CreativeWork.
<code>isReviewedBy</code>	Review	Link to the Person who edited the CreativeWork.
<code>isSimilarTo</code>	MediaObject	A media object that is similar to this content such as a song, reading, or image here, according.
<code>language</code>	MediaObject	The media object that contains this content such as a song, reading, or image here, according.
<code>mainEntity</code>	Text	Headline of the article.
<code>inLanguage</code>	Text	The language of the content. Please use one of the language codes from the IETF BCP 47 specification.
<code>interactionCount</code>	Text	A count of a specific user interaction with the item—for example, 42 likes, 1000 shares, 500 comments, or 2000 dislikes/likes. The user interaction must be one of the subtypes of <code>UserAction</code> .
<code>isFamilyFriendly</code>	Boolean	Indicates whether this item is family friendly.
<code>isInLanguage</code>	Text	The language(s) used to describe this item.
<code>isPartOf</code>	Thing	Indicates that the CreativeWork, item, or reference to it is an interview with a person.
<code>isPublished</code>	Offer	An offer to sell the item—for example, an offer to sell a product, the DMO of a movie, or tickets to an event.
<code>isReviewedBy</code>	Review	Specifies the Person or Organization that reviewed the CreativeWork.
<code>isStereotype</code>	Organization	The publisher of the creative work.
<code>isTranslatedVersion</code>	URL	Link to a page describing the editorial principles of the organization primarily responsible for the creation of the item.



Copyright Ellis Horowitz 2012 - 2022

## MicroData Markup for "Pirates of the Caribbean"

USC Viterbi  
School of Engineering

```
<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
<h1>	Movie name
<span>	Description
<div>	Director
<span>	Author
</div>	Actors
</div>	rating

USC

• •

## More Examples: Clarifying Hard to Understand Content



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

- `<time datetime="2022-04-01">04/01/22</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.
```

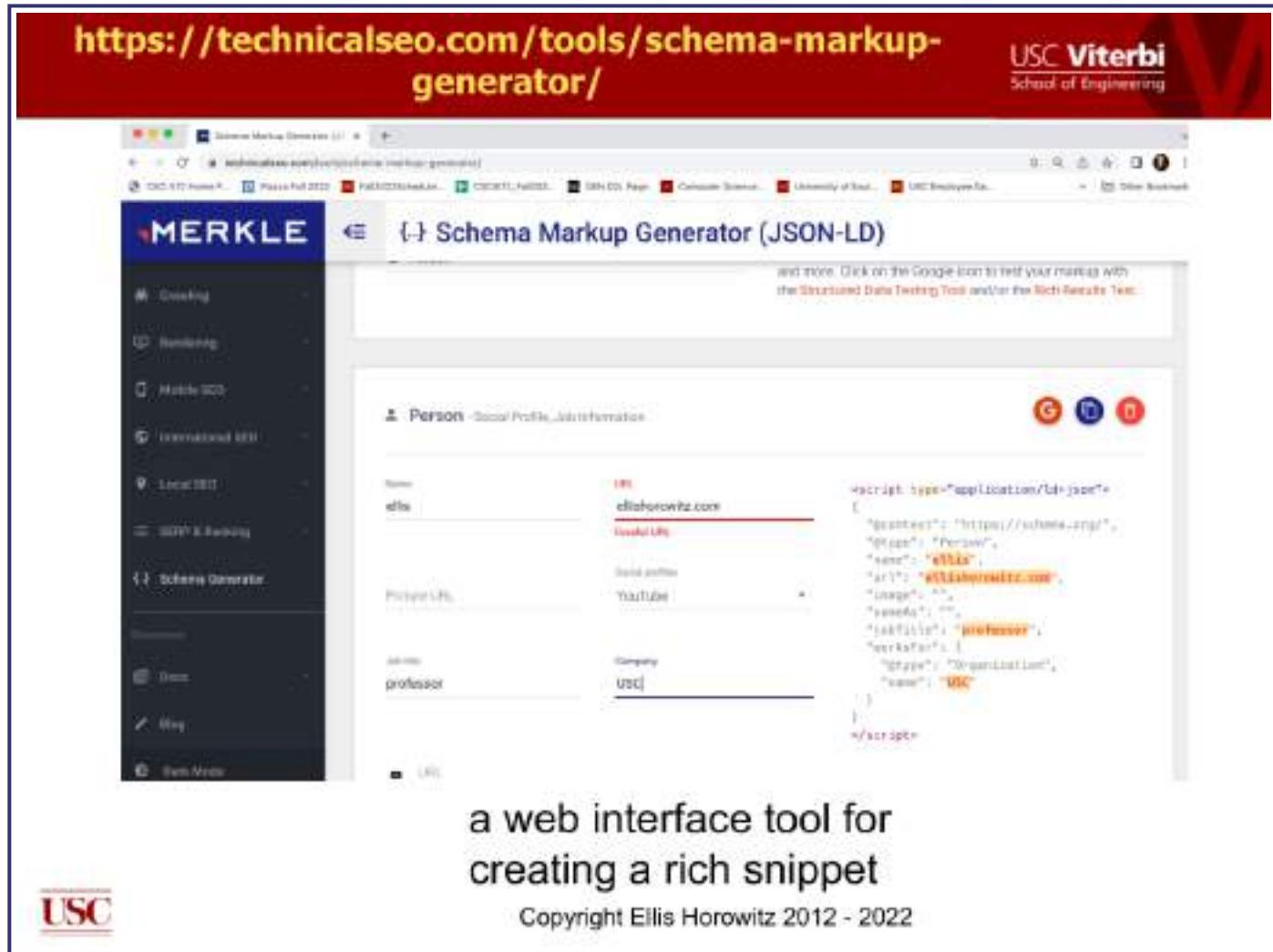
Event date:

```
<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>
```

**USC**



The screenshot shows a web browser window with two tabs open. The left tab displays a test results page for [www.usc.edu](https://www.usc.edu), indicating it is eligible for rich results with a green checkmark. The right tab displays a test results page for [www.cs.usc.edu](https://www.cs.usc.edu), also indicating it is eligible for rich results with a green checkmark. Both pages show detailed structured data results, such as address, phone number, and email.

https://search.google.com/test/rich-results

USC Viterbi  
School of Engineering

Rich Results Test

Does your page support rich results?

Test results

Page is eligible for rich results

Page is eligible for rich results

Results for www.usc.edu

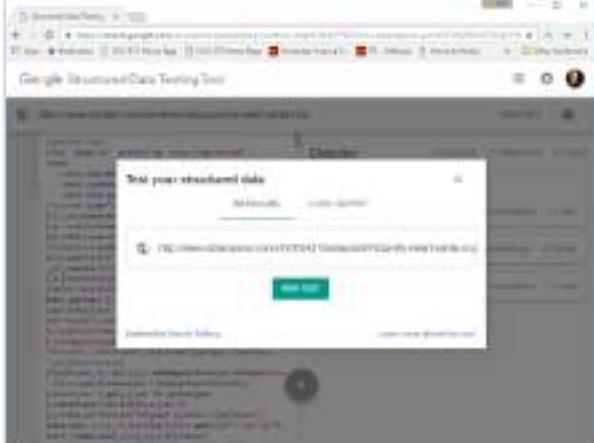
Results for www.cs.usc.edu

right Ellis Horowitz 2012 - 2022

• •

## Google's Structured Testing Tool

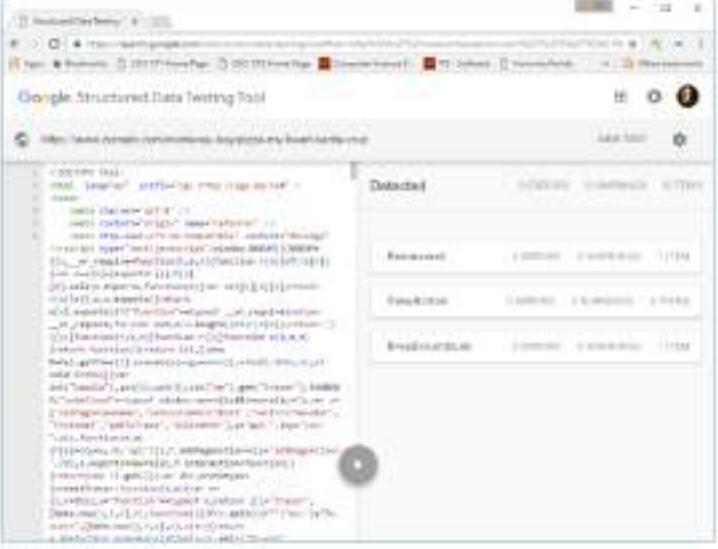




<https://search.google.com/structured-data/testing-tool>

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>



Detected	0/0 errors	0 warnings	0/0 errors
Resource	0 errors/0 errors	0 warnings/0 warnings	0/0 errors
Open API	0 errors/0 errors	0 warnings/0 warnings	0/0 errors
Breakdown	0 errors/0 errors	0 warnings/0 warnings	0/0 errors

```

[{"type": "error", "msg": "Missing structured data for the page.", "loc": "https://www.urbanspoon.com/r/6/765421/restaurant/Pizza-My-Heart-santa-cruz"}, {"type": "warning", "msg": "The page contains multiple structured data items for the same type (e.g., multiple \u0022Organization\u0022 items).", "loc": "https://www.urbanspoon.com/r/6/765421/restaurant/Pizza-My-Heart-santa-cruz"}]

```


Copyright Ellis Horowitz 2012 - 2022

• •

## Google's Rich Snippets Testing Tool

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

USC Viterbi  
School of Engineering

The screenshot shows the Google Rich Snippets Testing Tool interface. On the left, there is a code editor window displaying an HTML snippet with various schema.org tags. On the right, there is a results panel titled "Person (1)" which displays the extracted information from the code. The results include:

- Description:** 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.
- Name:** Dr. Ellis Horowitz
- Email:** ellis@viterbi.usc.edu
- Job Title:** Professor
- Member [Organization]:** University of Southern California
- Address (PostalAddress):** 3600 La Cienega Blvd, Los Angeles, CA 90009, USA

At the bottom of the results panel, there is a link to "Custom Search Result Filters".

• •

## Google Introduces New Tags for Snippet Control(1)

USC Viterbi  
School of Engineering

- 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">
```

USC

••

## Google Introduces New Tags for Snippet Control(2)

USC Viterbi  
School of Engineering

- 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.

USC

• •

## Summary

USC Viterbi  
School of Engineering

- **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

USC

• •

## Extending Schema.org to handle PAA

USC Viterbi  
School of Engineering

- **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=NJjiLDn9-38>
- 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>

USC

← 1/32 → \*\*\* 09:25:11

# Query processing

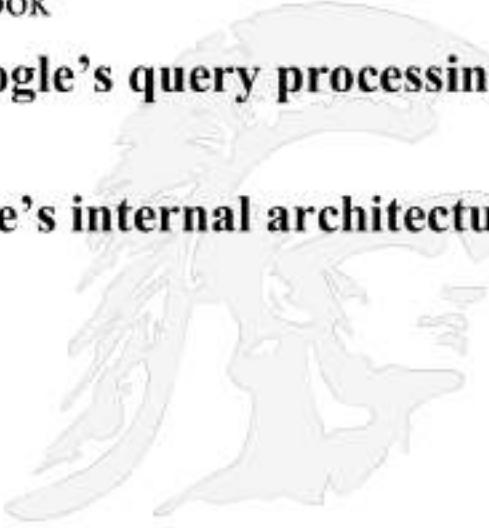
## how does the search engine respond?

••

University of Southern California 

 USC **Viterbi**  
School of Engineering **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**



Copyright Ellis Horowitz, 2011-2022 2

••

University of Southern California  USC

**USC Viterbi**  
School of Engineering

## 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
    - Try to 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



Copyright Ellis Horowitz, 2011-2022

3

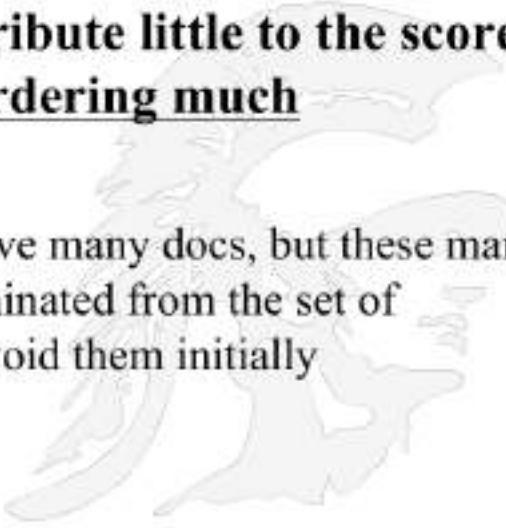
••

University of Southern California  USC

 USC **Viterbi**  
School of Engineering

**Strategy 1:**  
**Consider Only Query Terms with High-idf Scores**

- 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



Copyright Ellis Horowitz, 2011-2022

4

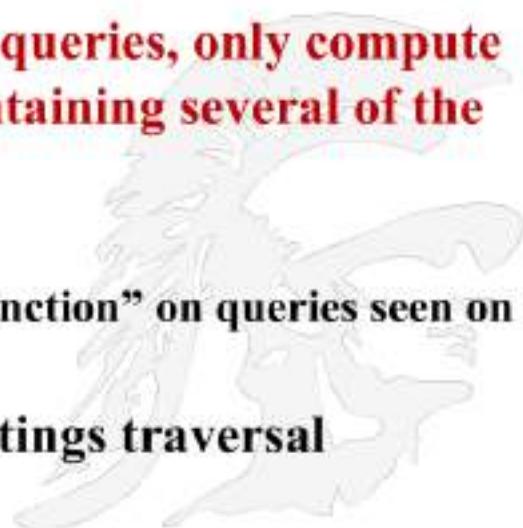
••

University of Southern California 

 USC **Viterbi**  
School of Engineering

Strategy 2:  
**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



Copyright Ellis Horowitz, 2011-2022

5

••

University of Southern California 

 USC **Viterbi**  
School of Engineering

## 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

Copyright Ellis Horowitz, 2011-2022

6

••

University of Southern California 

 USC **Viterbi**  
School of Engineering

## 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 among websites
  - Articles in curated newspapers
  - A paper/webpage with many citations, or equivalently
  - A web page with high PageRank

Copyright Ellis Horowitz, 2011-2022

7

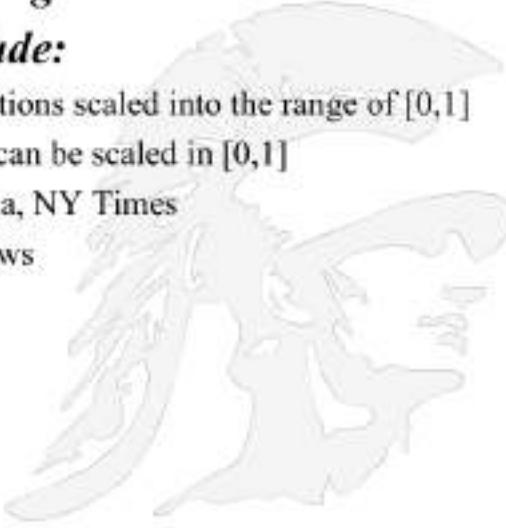
••

University of Southern California 

 USC **Viterbi**  
School of Engineering

## 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



Copyright Ellis Horowitz, 2011-2022

8

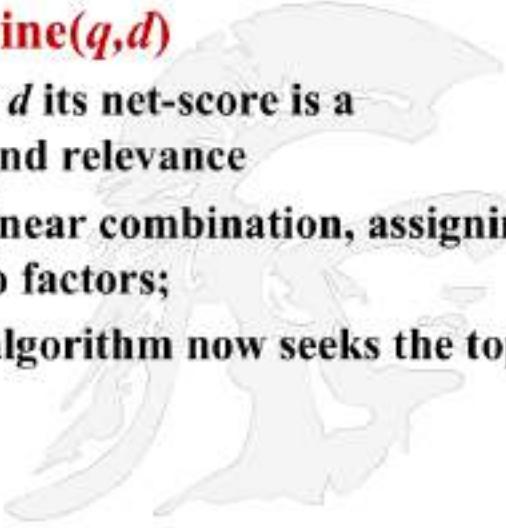
••

University of Southern California 

 USC **Viterbi**  
School of Engineering

Combine Relevance and Authority

- Consider a simple total score combining cosine relevance and authority
- **net-score( $q, d$ ) =  $g(d) + \text{cosine}(q, d)$** 
  - 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



Copyright Ellis Horowitz, 2011-2022

9

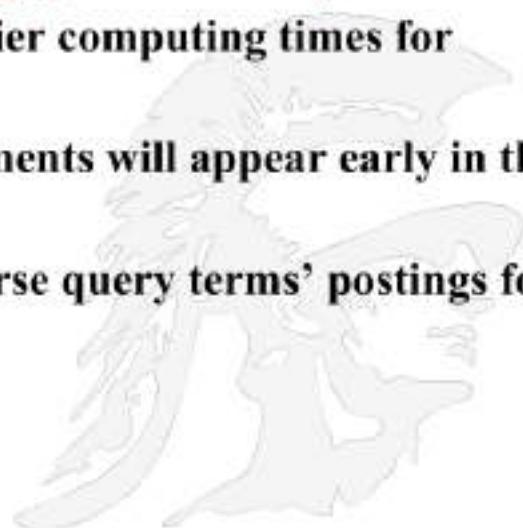
••

University of Southern California 

 USC **Viterbi**  
School of Engineering

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



Copyright Ellis Horowitz, 2011-2022

10

••

University of Southern California 

 USC Viterbi  
School of Engineering

## 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

↑ authority score      ↑ relevance cosine score

- This is equivalent to

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

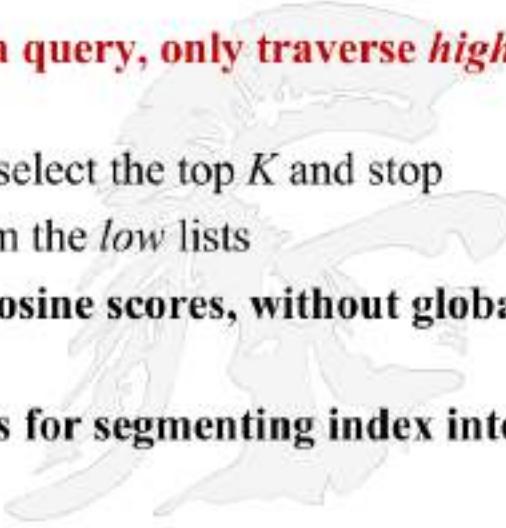
Copyright Ellis Horowitz, 2011-2022

11

••

## 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



••

University of Southern California 

 USC **Viterbi**  
School of Engineering

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
  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



Copyright Ellis Horowitz, 2011-2022

15

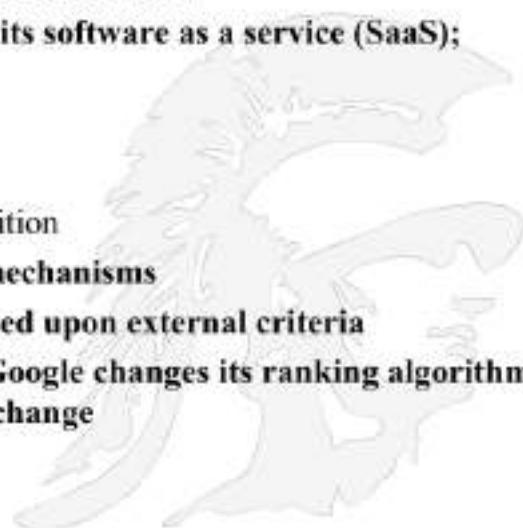
This is Searchmetrics' doc that lists the various metrics ('ranking factors') they use, to characterize Google's page ranking algorithm.

••

University of Southern California  USC  USC

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

- 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
- 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>



Copyright Ellis Horowitz, 2011-2022

27

This page at moz.com lists every change in Google's SERP (results page) influencing algorithms, going back to the year 2000.

••

University of Southern California  

## MozRank

- 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



Copyright Ellis Horowitz, 2011-2022 28

Here is Moz's link explorer page.

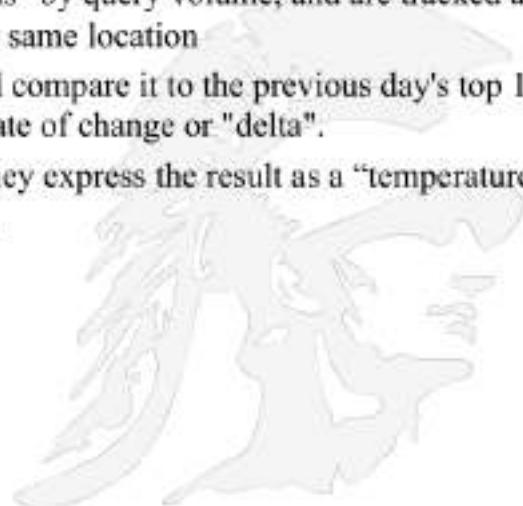
• •

University of Southern California 

**USC Viterbi**  
School of Engineering

## Moz.com Tracks Google Algorithm Updates

- MozCast is a statistical technique designed to highlight the affects of Google modifying their ranking algorithm
- 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 Fahrenheit; an average day is about 70° F.



Copyright Ellis Horowitz, 2011-2022

31

••

University of Southern California 

 USC **Viterbi**  
School of Engineering

# 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

Copyright Ellis Horowitz 2011-2022

This is the 'how search works' page.

••

University of Southern California  USC 

## How Google Search Has Changed Over the Years

**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**

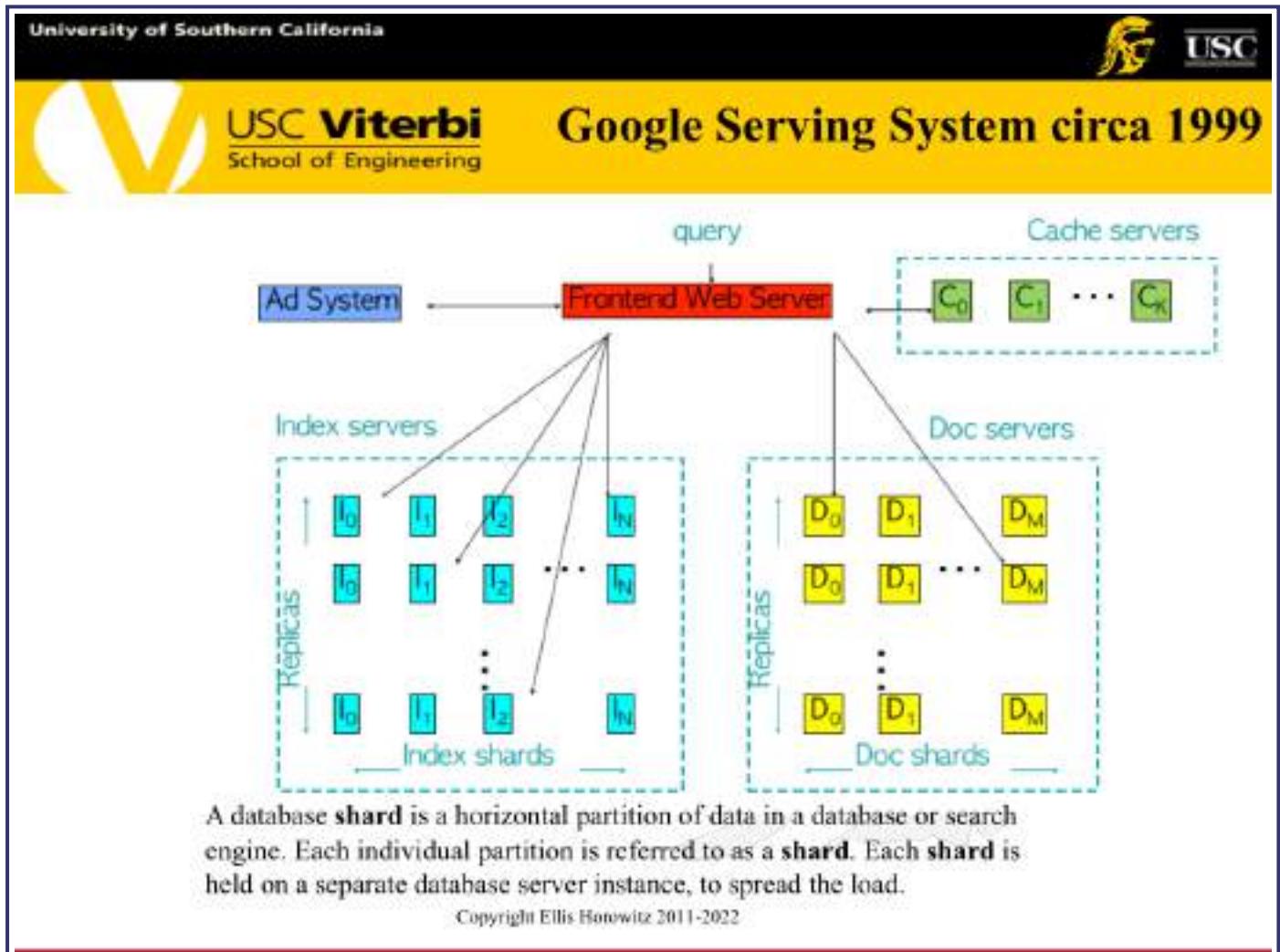
**2013, use of carousels for display**

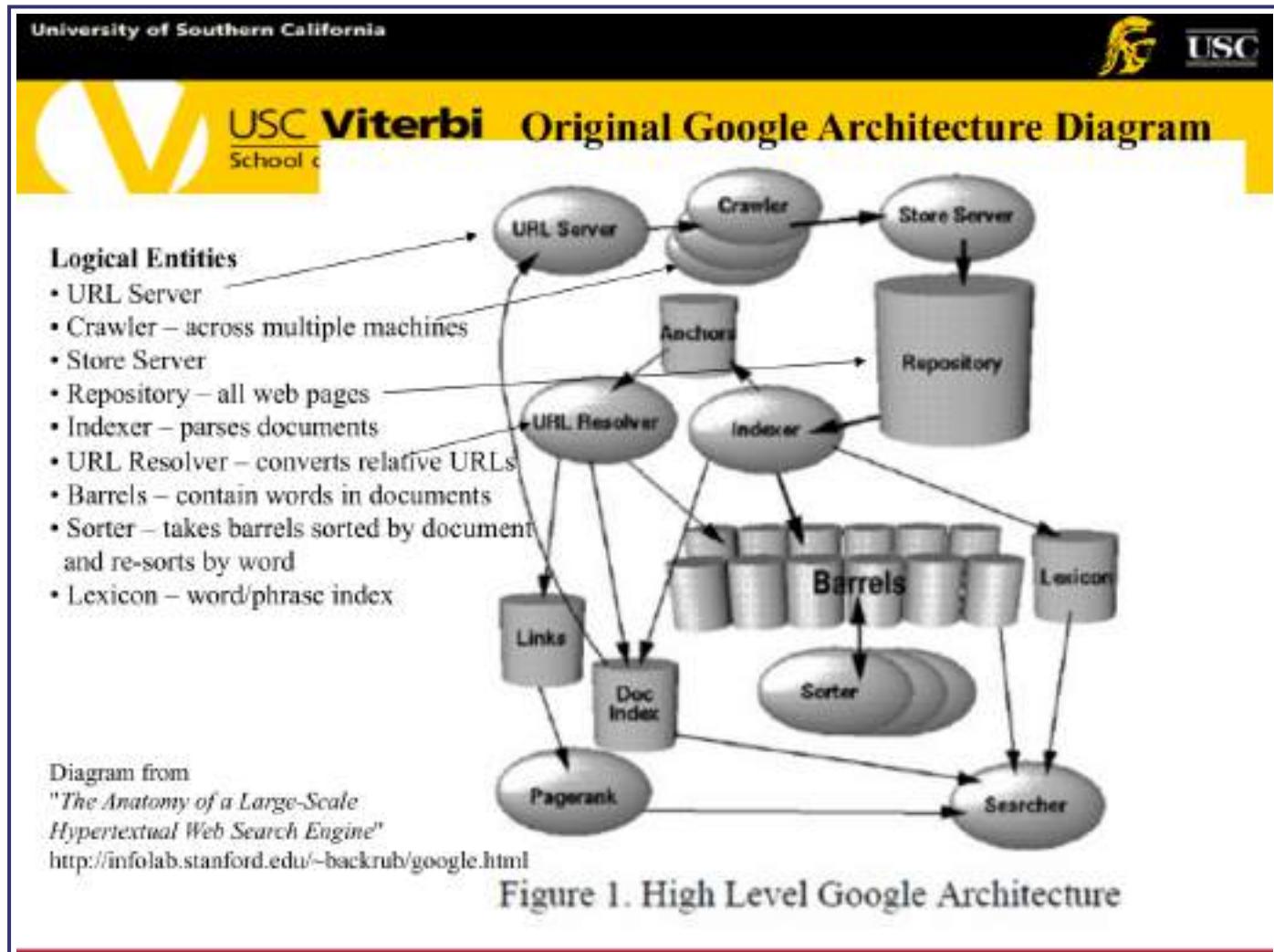


This official timeline from Google shows how search has developed in the past 15 years. The latest update is the most significant algorithm change in years. It will affect 80 per cent of searches and the firm began rolling it out to select users a month ago.

Copyright Ellis Horowitz 2011-2022

••



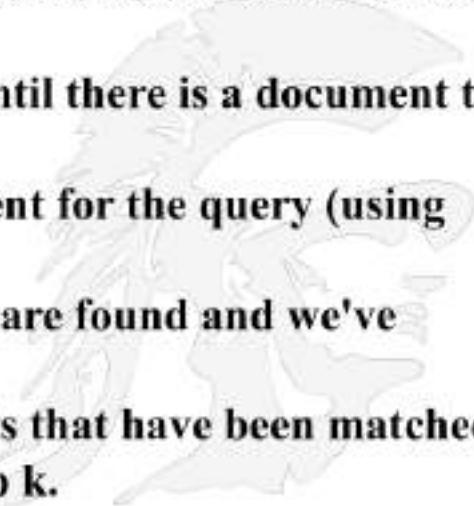


••

University of Southern California 

 USC **Viterbi**  
School of Engineering

## Google's Early 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.**

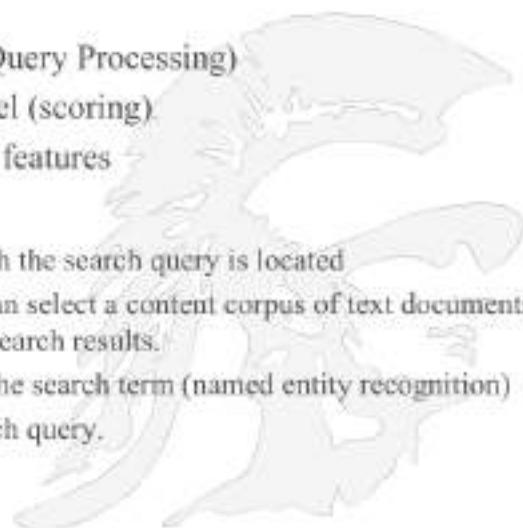
••

University of Southern California  USC

 USC **Viterbi**  
School of Engineering

## Modern Query Processing Methodology

- Google (and others) have now moved query processing far beyond keyword matching
- In *semantic* information retrieval systems, entities play a central role in several tasks.
  1. Understanding the search query (Search Query Processing)
  2. Relevance determination at document level (scoring)
  3. Compilation of search results and SERP\* features
- **Important steps**
  1. Identification of the thematic ontology in which the search query is located
    - If the thematic context is clear, Google can select a content corpus of text documents, videos, images ... as potentially suitable search results.
  2. Identification of entities and their meaning in the search term (named entity recognition)
  3. Understanding the semantic meaning of a search query.
  4. Identification of the search intent
  5. Semantic annotation of the search query
  6. Refinement of the search term
- \*Search engine results page



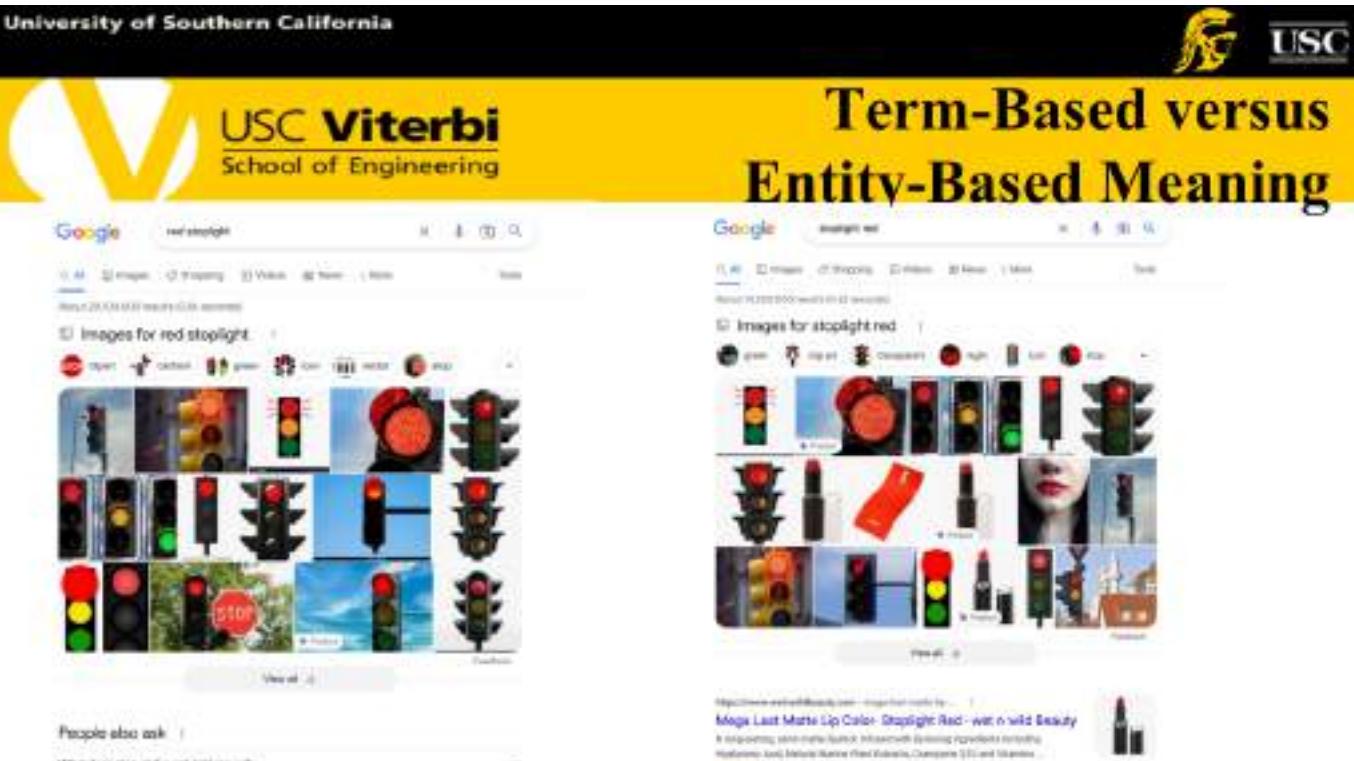
Copyright Ellis Horowitz, 2011-2022

38

••

University of Southern California 

# Term-Based versus Entity-Based Meaning



An entity based search recognizes the different context based on the different arrangement; “Stoplight red” is an entity, not to be confused as separate terms “red” and “stoplight”

Copyright Ellis Horowitz 2011-2022

••

University of Southern California  USC  USC

## Google's Query Processing Elements

1. Rankbrain/MUM/Bert are used to identify entities  
 2. Knowledge Graph: an ontology  
 3. Named Entity Recognition  
 4. Document Index: inverted index of terms  
 5. Relevance Scoring  
 6. Duplicate removal  
 7. Personalization



© Olaf Kopp, Aufgesang GmbH

Copyright Ellis Horowitz, 2011-2022

40

• •

University of Southern California  USC  USC

## Using the KnowledgeGraph to Identify Entities

- The biggest challenge for Google with regard to semantic search is identifying and extracting entities, their attributes and other information from data sources such as websites.
- The information is mostly not structured and not error-free.
- The current Knowledge Graph is largely based on the structured content from Wikidata and the semistructured data from Wikipedia or Wikimedia.

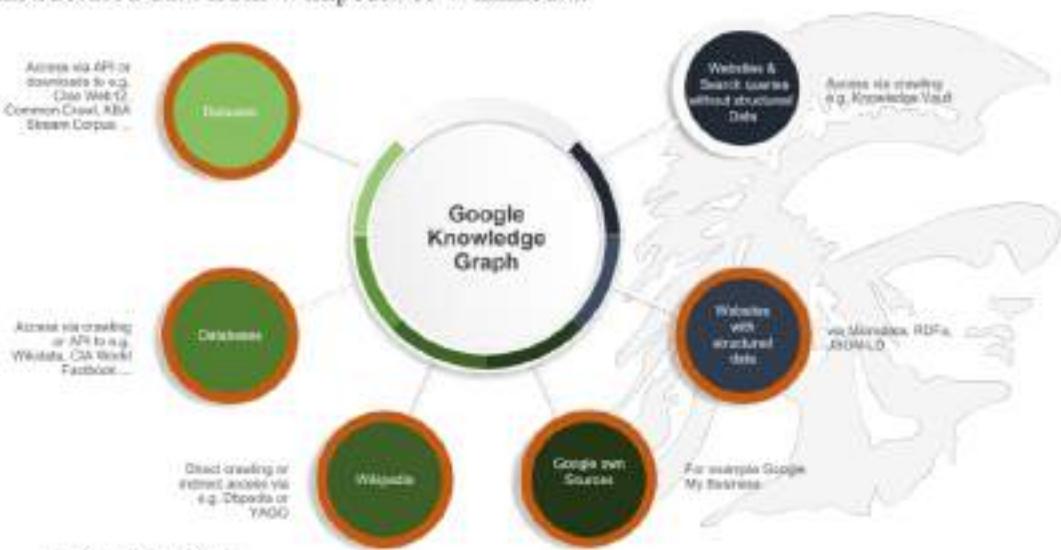


Diagram illustrating the Google Knowledge Graph's data sources:

- Database**: Access via API or direct access e.g. DBpedia, YAGO, Wikipedia, CIA World Factbook...
- Database**: Access via crawling or API to e.g. Wikidata, CIA World Factbook...
- Wikidata**: Direct crawling or indirect access via e.g. DBpedia or YAGO
- Google own Sources**: For example Google My Business
- Websites with structured Data**: via Microdata, RDFa, JSON-LD
- Websites & Search sources without structured Data**: Access via crawling e.g. Knowledge Vault

Copyright Ellis Horowitz, 2011-2022

41

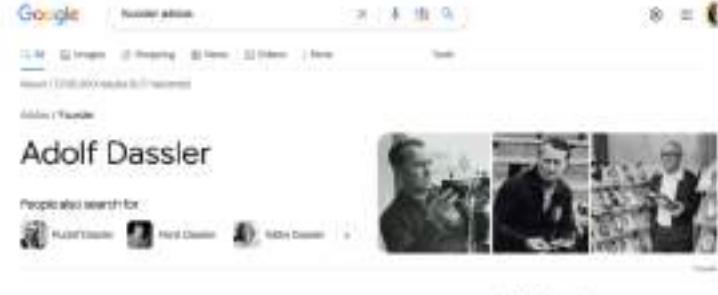
••

University of Southern California  USC 

## Entity Recognition in the Knowledge Graph: Two Examples



**Query: ceo vw**  
No where is the person's name mentioned  
Entities are: "vw" and "boss"



**Query: founder adidas**  
Entities: "Adidas", "founder"



Copyright Ellis Horowitz 2011-2022

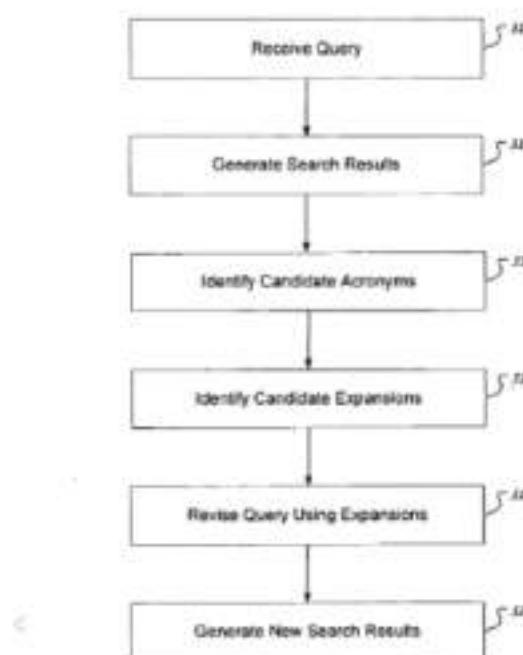
••

University of Southern California  USC

**USC Viterbi**  
School of Engineering

## RankBrain An Entity-Based Processor

- RankBrain is a deep learning-based algorithm that is used after the selection of an initial subset of search results
- Introduced in 2015 as part of their Hummingbird algorithm update
  - RankBrain **maps keywords into entities** which are then looked for in the Knowledge Graph;
  - Words surrounding the entity are considered the context of the query
  - Google claims that RankBrain is the third most important factor in their ranking algorithm (links/words being numbers 1 and 2)



```
graph TD; A[Receive Query] --> B[Generate Search Results]; B --> C[Identify Candidate Acronyms]; C --> D[Identify Candidate Expansions]; D --> E[Revise Query Using Expansions]; E --> F[Generate New Search Results];
```

The flowchart illustrates the RankBrain process. It starts with 'Receive Query', followed by 'Generate Search Results'. Then it branches into two parallel paths: 'Identify Candidate Acronyms' and 'Identify Candidate Expansions'. Both of these paths lead to 'Revise Query Using Expansions', which then leads to 'Generate New Search Results'. Each step is accompanied by a small bracketed number indicating its rank: 1, 2, 3, 4, 5, and 6 respectively.

• •

University of Southern California  USC

 USC **Viterbi**  
School of Engineering

## RankBrain in Its Simplest Form

- 1. Google receives a query for something it's never seen before (like a new movie title or a phrase connecting two topics, like "which country has the best cars")
- 2. Google assigns the entity a unique identifier, like 9202a8c04000641f80000000000006567.
- 3. Google determines the entity's relatedness to other entities, then assigns it a value.
- 4. Google determines the entity's notability, then assigns it a value.
- 5. Google determines the entity's contribution, then assigns it a value.
- 6. Google evaluates any awards the entity received, then assigns a value.
- 7. Each value is weighted according to the entity's query type. For example, in the case of the best car example, Google may prioritize relatedness and awards for particular brands, and return the result as a carousel of options rather than a single webpage.

- The strength of RankBrain is its **ability to** handle novel queries. In addition RankBrain is a framework for continual learning of entities.

Copyright Ellis Horowitz, 2011-2022

46

• •

University of Southern California 

## USC Viterbi School of Engineering

What is RankBrain observing exactly?  
It's paying very close attention to **how you interact with the search results**. Specifically, it's looking at:

- \* Organic Click-Through-Rate
- \* Dwell Time
  - \* Dwell Time is the amount of time that a Google searcher spends on a page from the search results before returning back to the SERPs.
- \* Bounce Rate
  - Bounce Rate is defined as the percentage of visitors that leave a webpage without taking an action, such as clicking on a link, filling out a form, or making a purchase.
- \* Pogo Sticking
  - Pogo sticking is when a search engine users visits several different search results in order to find a result that satisfies their search query
- \* These are known as user experience signals (UX signals).

### How RankBrain (Probably) Uses UX Signals



```
graph TD; A[User Types Keyword Into Google] --> B[RankBrain Turns Keyword Into Concepts]; B --> C[Shows a Set of Results For That Concept]; C --> D{Did a Page Satisfy the User?}; D -- NO --> E[Try Another Page Next Time]; D -- YES --> F[Uprank That Page]
```

The flowchart illustrates the process RankBrain follows when a user types a keyword into Google:

- User Types Keyword Into Google
- RankBrain Turns Keyword Into Concepts
- Shows a Set of Results For That Concept
- A decision diamond asks: Did a Page Satisfy the User?
- If NO, the process leads to Try Another Page Next Time.
- If YES, the process leads to Uprank That Page.

Copyright Ellis Horowitz 2011-2022

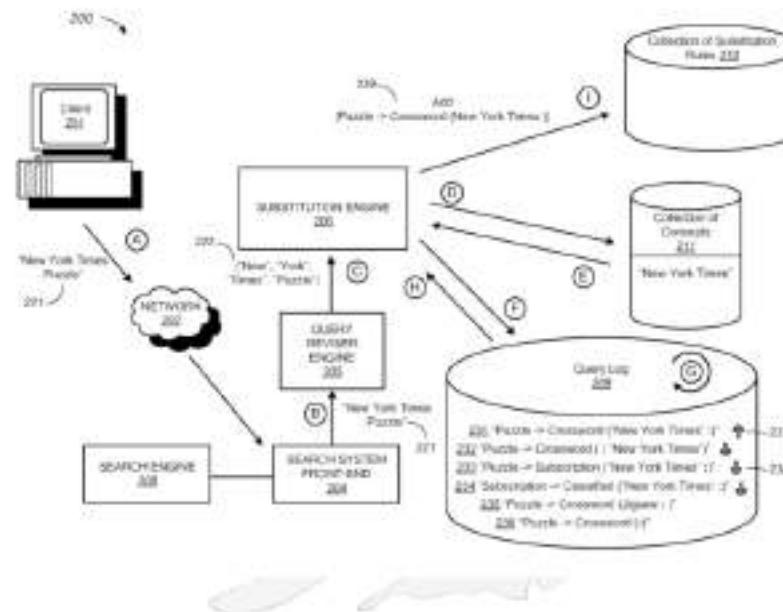
• •

University of Southern California  USC  USC

## RankBrain Patent

### Using concepts as contexts for query term substitutions

- a method includes receiving a query that includes at least three sequential query terms; determining that the sequential query terms represent a concept; and in response to determining that the sequential query terms represent a concept, collecting query term substitution data for one or more query terms that occur in queries that include the concept.



Copyright Ellis Horowitz, 2011-2022

48

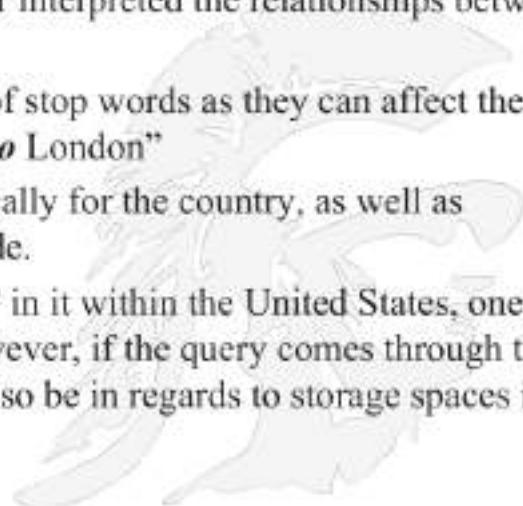
••

University of Southern California 

 USC **Viterbi**  
School of Engineering

## RankBrain Offline

- When offline, RankBrain is given
  - batches of past searches and learns by matching search results
  - Newspaper articles and learns to associate items within a news article
- Studies showed how RankBrain better interpreted the relationships between words.
  - RankBrain includes the analysis of stop words as they can affect the meaning of a query, e.g. "flights **to** London"
  - RankBrain learns phrases specifically for the country, as well as language, in which a query is made.
  - E.g. a query with the word "boot" in it within the United States, one will get information on footwear. However, if the query comes through the UK, then the information could also be in regards to storage spaces in cars



Copyright Ellis Horowitz, 2011-2022

50

•

University of Southern California  USC

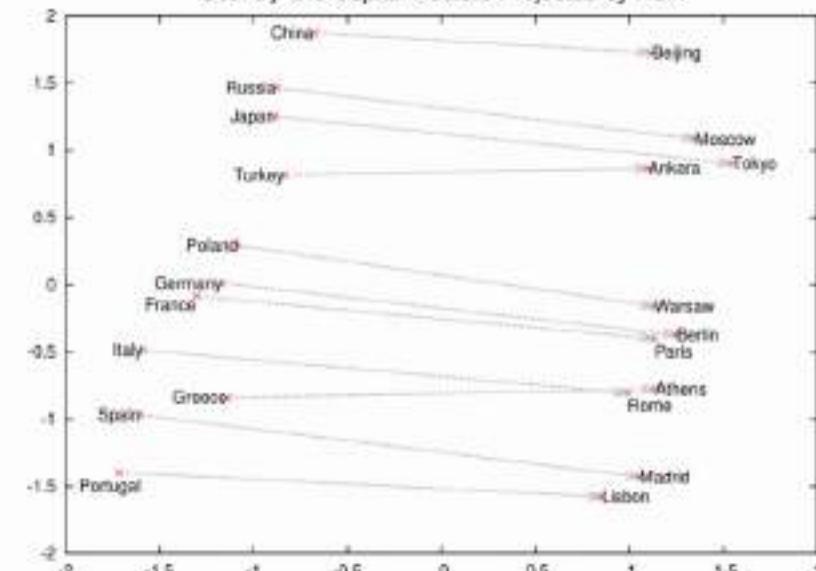
**USC Viterbi**  
School of Engineering

## RankBrain Learns the Concept of Capital Cities

- Using offline sources RankBrain is able to identify these associations

The figure illustrates the ability of the model to automatically organize concepts and learn implicitly the relationship between them, as during the training no supervised info was input about what a capital city means

Country and Capital Vectors Projected by PCA



Country	Capital
China	Beijing
Russia	Moscow
Japan	Tokyo
Turkey	Ankara
Poland	Warsaw
Germany	Berlin
France	Paris
Italy	Athens
Greece	Rome
Spain	Madrid
Portugal	Lisbon

Copyright Ellis Horowitz, 2011-2022

51

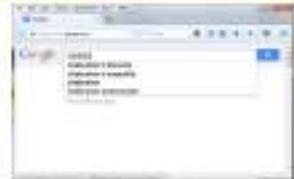
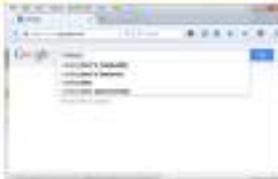
← 1/28 → \*\*\* Objektiv

# Auto correction/completion

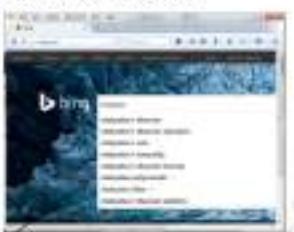
..

University of Southern California  USC

## 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

Copyright Ellis Horowitz 2011-2022

Himalayus

• •

University of Southern California  USC

## More Examples



The figure shows three screenshots of web browsers illustrating different approaches to auto-correction:

- Left screenshot:** Shows a search bar with the query "we get caught". Below it is a dropdown menu with suggestions: "we get caught", "we get caught in the fine线条", "we get caught up in a fire office", "we get caught up", and "we get caught making out". An arrow points from the text "easy for people, but harder for computers to correct, likely use of n-grams" to this screenshot.
- Middle screenshot:** Shows a search bar with the query "jagwa". Below it is a dropdown menu with suggestions: "jagwah", "jagwah definition", "jagwah in sentence", "jagwah theory", and "jagwah map". An arrow points from the text "easy for a computer to correct, likely use of a database of words" to this screenshot.
- Right screenshot:** Shows a search bar with the query "jagwa". Below it is a dropdown menu with suggestions: "jagwah", "jagwah definition", "jagwah in sentence", "jagwah theory", and "jagwah map". An arrow points from the text "computer needs to both identify the error and correct the misspelling" to this screenshot.

Copyright Ellis Horowitz 2011-2022

••

University of Southern California  
USC Viterbi  
School of Engineering

## 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)

Copyright Ellis Horowitz, 2011-2022

**2. Smartphone input**

..

University of Southern California  USC

**USC Viterbi**  
School of Engineering

## 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*

Copyright Ellis Horowitz, 2011-2022

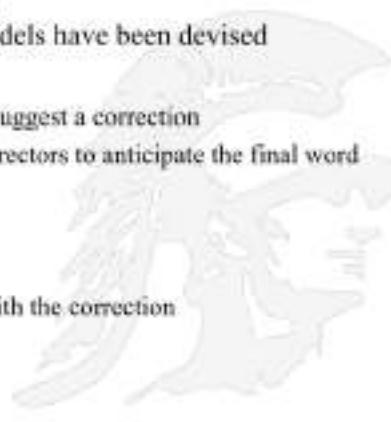
5

..

University of Southern California  USC

**USC Viterbi**  
School of Engineering

## 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

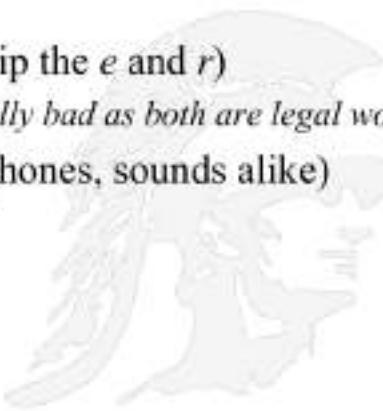
Copyright Ellis Horowitz, 2011-2022

4

..

University of Southern California  USC

**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*

Copyright Ellis Horowitz, 2011-2022

7

..

University of Southern California  USC

## 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

Copyright Ellis Horowitz, 2011-2022

•

University of Southern California USC

**USC Viterbi**  
School of Engineering

## 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)

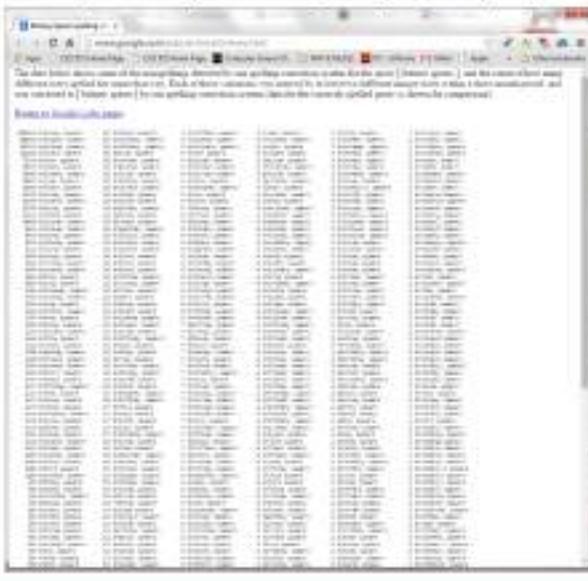
Copyright Ellis Horowitz, 2011-2022

••

University of Southern California  USC

**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.

..

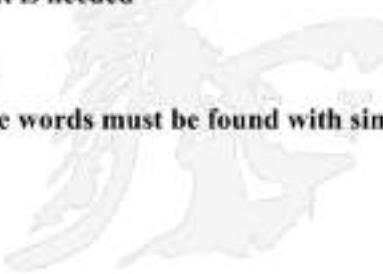
University of Southern California

 USC Viterbi  
School of Engineering

 USC

## 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*
    - *N-grams are most useful here*
  3. To resolve the above, candidate words must be found with similar pronunciations
    - use the Soundex algorithm



Copyright Ellis Horowitz, 2011-2022

11

..

University of Southern California  USC

**USC Viterbi**  
School of Engineering

## 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

Copyright Ellis Horowitz, 2011-2022

12

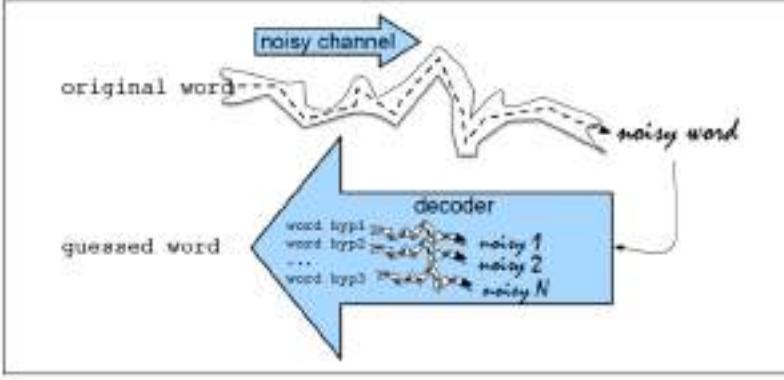
..

University of Southern California  USC

**USC Viterbi**  
School of Engineering

## 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



original word

noisy channel

noisy word

decoder

guessed word

word hyp1  
word hyp2  
...  
word hypN

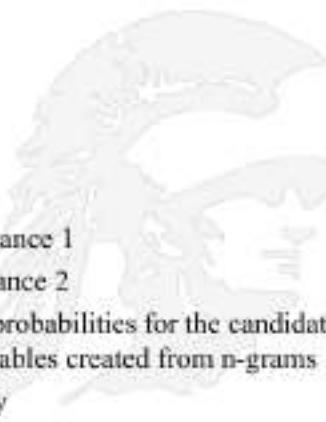
noisy 1  
noisy 2  
noisy N

..

University of Southern California  USC

**USC Viterbi**  
School of Engineering

## 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
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

Copyright Ellis Horowitz, 2011-2022

15

..

University of Southern California  USC

**USC Viterbi**  
School of Engineering

## 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**

Copyright Ellis Horowitz, 2011-2022

16

••

University of Southern California  USC

## Use Edit Distance To Produce Candidate Corrections

Input	Candidate Correction	Correct Letter	Error Letter	Correction Type	
acress	actress	t	-	insertion	Six words within 1 of acress Context check is necessary to choose the appropriate word (try this yourself in Google/Bing)
acress	cress	-	a	deletion	
acress	caress	ca	ac	transposition	
acress	access	c	r	substitution	
acress	across	o	e	substitution	
acress	acres	-	s	deletion	

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

Copyright Ellis Horowitz, 2011-2022

17

..

University of Southern California  USC

**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 across to ...  
I love the across ...  
a kiss and a across ...

"across" is the  
most likely  
correction —→ across

..

University of Southern California  USC

**USC Viterbi**  
School of Engineering

## 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

Copyright Ellis Horowitz, 2011-2022

19

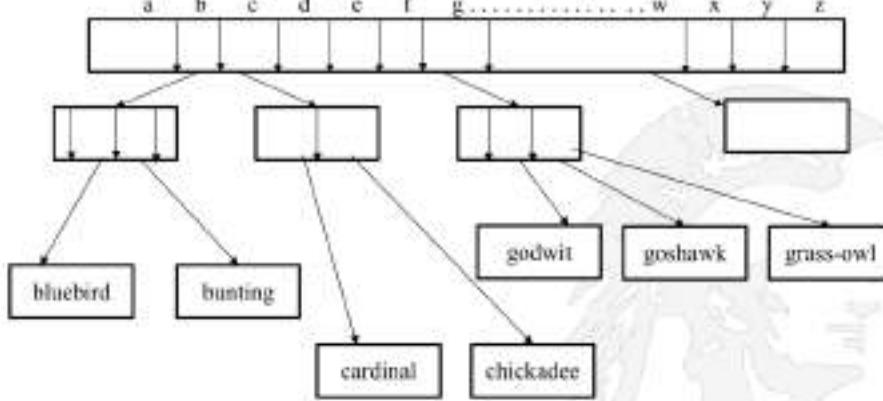
..

University of Southern California  USC

**USC Viterbi**  
School of Engineering

## 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



The diagram illustrates a Trie structure for storing bird names. At the top, a horizontal row of boxes represents the alphabet from 'a' to 'z'. Below this, several levels of boxes represent the structure of the Trie. The first level contains four boxes, each pointing to a second-level node. The second level contains three boxes, each pointing to a third-level node. The third level contains three boxes, each pointing to an element node labeled with a bird name. The labels are: bluebird, bunting, cardinal, chickadee, godwit, goshawk, and grass-owl.

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

Copyright Ellis Horowitz, 2011-2022

20

..

University of Southern California  USC

**USC Viterbi**  
School of Engineering

## 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 misspellings
    - [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

Copyright Ellis Horowitz, 2011-2022

21

•

University of Southern California  USC



**USC Viterbi**  
School of Engineering

## 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

Copyright Ellis Horowitz, 2011-2022

22

••

University of Southern California  USC

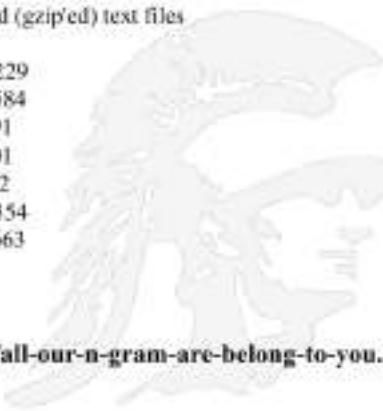
**USC Viterbi**  
School of Engineering

## 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: 95,119,065,584  
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>

Copyright Ellis Horowitz 2011-2022

..

University of Southern California  USC

**USC Viterbi**  
School of Engineering

## 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>

Copyright Ellis Horowitz, 2011-2022

31

..

University of Southern California

 USC Viterbi  
School of Engineering



# Edit Distance & Levenshtein Algorithm

Copyright Ellis Horowitz 2011-2018

..

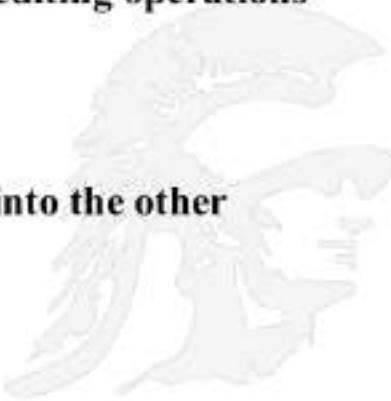
University of Southern California  USC

**USC Viterbi**  
School of Engineering

## 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**



Copyright Ellis Horowitz, 2011-2022

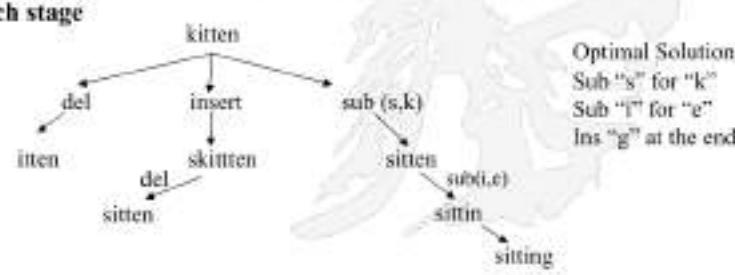
33

..

University of Southern California  USC

## 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



Optimal Solution  
Sub "s" for "k"  
Sub "t" for "e"  
Ins "g" at the end

Copyright Ellis Horowitz, 2011-2022

14

..

University of Southern California  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, // deletion
 d[i,j-1] + 1, // insertion
 d[i-1,j-1] + substitutionCost) //substitution
 return d[m,n]

```

Copyright Ellis Horowitz 2011-2022

You can learn more about the Levenshtein algorithm [here](#).

And, you can run the algorithm [here](#), to understand how the edit distance is incrementally computed; this is a minimal version.

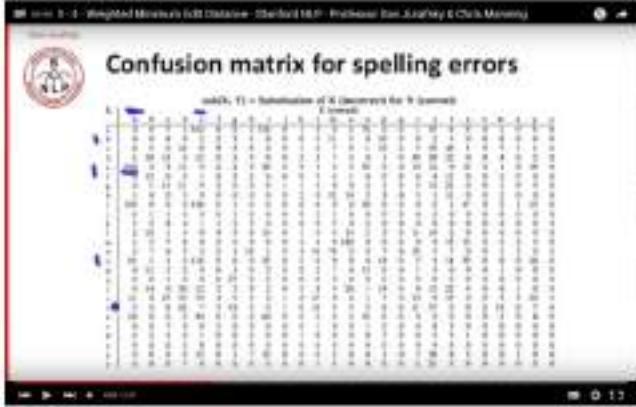
This is an example of how the Levenshtein distance is used in practice.

•

University of Southern CaliforniaUSC Viterbi  
School of Engineering

## 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)



the confusion matrix for spelling errors shows us, e.g., that "e" is most often confused with "a", and that "i" is often confused with both "e" and "a"

1/60

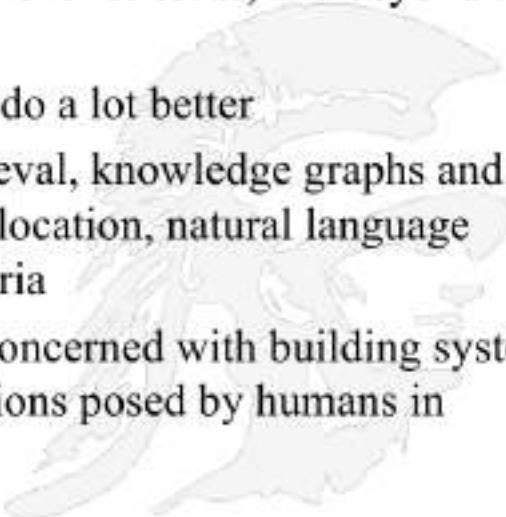
\*\*\* 09:27:48

# 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



••

University of Southern California  USC

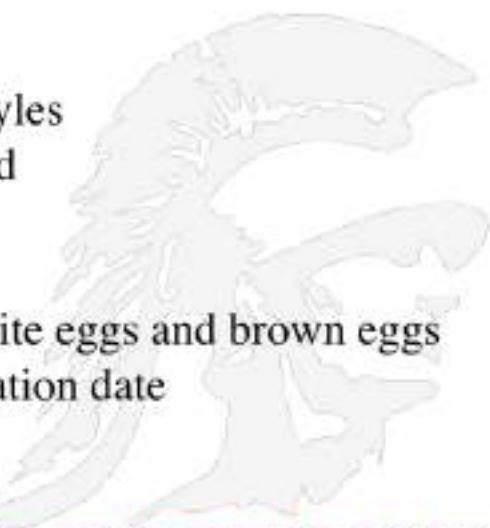
**People want to ask questions...**

**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**

Copyright Ellis Horowitz 2011-2022



• •

University of Southern California  USC

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: http://v.viterbi.usc.edu/Australia/during-the-Great-Depression%20during=Google+Search

Google - Great Depression

How Google used to respond to questions

**From Poor Boy to Prime Minister**  
... how did he come to lead **Australia during** World War ... April 1939 Men... takes over as  
**Prime Minister** after the death of Lyons; Sept 3 1939 **Australia** declares war ...  
john.curtin.edu.au/manofpoorboytopm.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 ... the Prime Minister and His ... the Communist Party in Australia during  
john.curtin.edu.au/letters/activities/comunism.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 ... died du... the 1930s. ... He became **Prime Minister** following Curtin's death, succeeding  
www.nrima.gov.au/prime ministers3.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)



• •

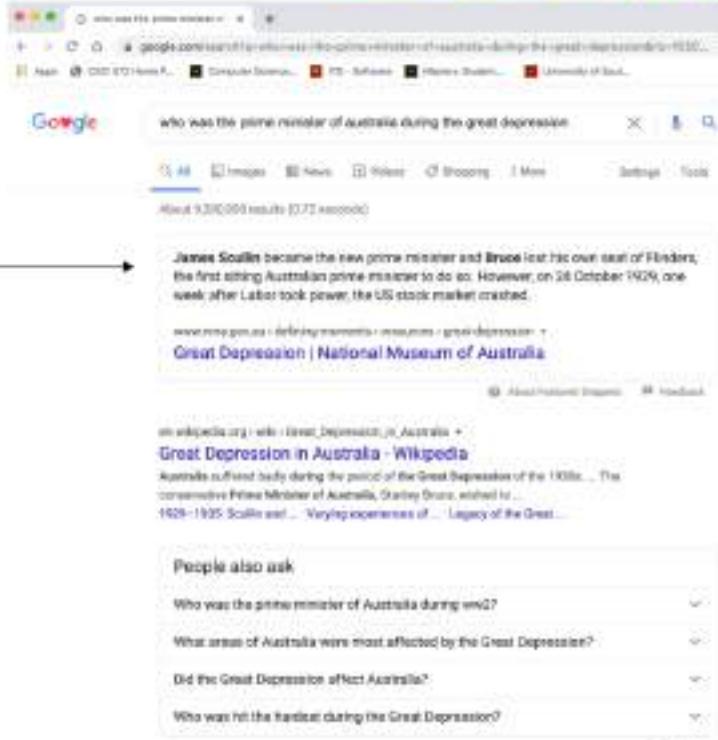
University of Southern California  USC

**USC Viterbi**  
School of Engineering

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 australia during the great depression

All Images News Videos Shopping More Settings Tools

About 9,300,000 results (0.7 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 26 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 Present Images Feedback

en.wikipedia.org › wiki › Great\_depression\_in\_Australia + Great Depression in Australia - Wikipedia

Australians 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... Very brief histories 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?

• •

University of Southern California  USC

## USC Viterbi School of Engineering

# Google has Improved Its Ability to Answer Many Questions

Google how old is mariah carey

About 37,000,000 results (0.73 seconds)

Mariah Carey / Age  
53 years  
March 27, 1969



People also search for

 Nick Cannon 41 years	 Jennifer Lopez 42 years	 Beyoncé 40 years
---------------------------------------------------------------------------------------------------------	------------------------------------------------------------------------------------------------------------	-----------------------------------------------------------------------------------------------------

Feedback

People also ask:

- What is Mariah Carey's net worth 2023? 
- 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
-  Beats
-  Apple Music
-  More music services

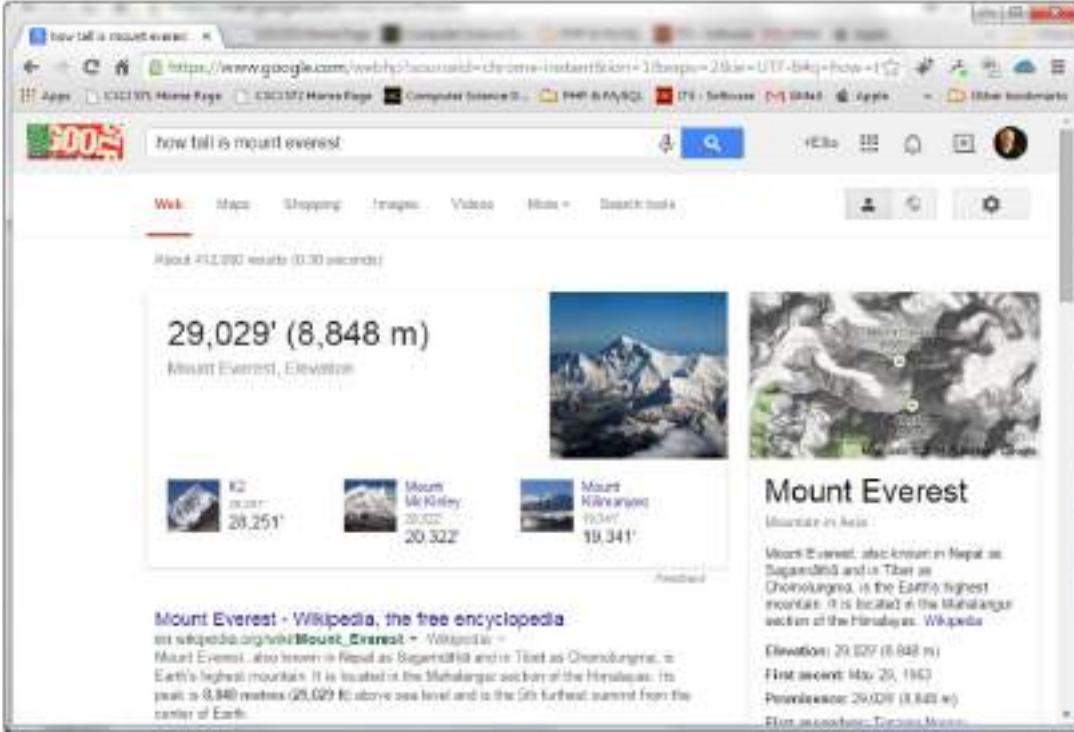
Mariah Carey is an American singer, actress, and record producer. Her "Singing Superstar" and the "Oscar" she is noted for her five-octave, melismatic singing style, and big whistle register. Carey rose to fame with her self-titled debut album, *Mariah Carey*, which was released in 1990. She has sold over 200 million records worldwide.

Born: March 27, 1969 (age 53)   
Children: Monroe Scott Dame  
Spouse: Nick Cannon (m. 2008–present) 

• •

University of Southern California  USC

**Some Questions are Easily Answered**

  
A screenshot of a Google search results page. The search query is "how tall is mount everest". The top result shows the elevation "29,029' (8,848 m)" and "Mount Everest, Elevation". Below this, there are three small thumbnail images with their respective elevations: K2 (28,251'), Mount McKinley (20,322'), and Mount Rainier (10,341'). To the right, there is a large image of Mount Everest and a detailed map showing its location. A link to the "Mount Everest - Wikipedia, the free encyclopedia" is provided, along with a snippet of text about the mountain's height and location.

Copyright Ellis Horowitz 2011-2022

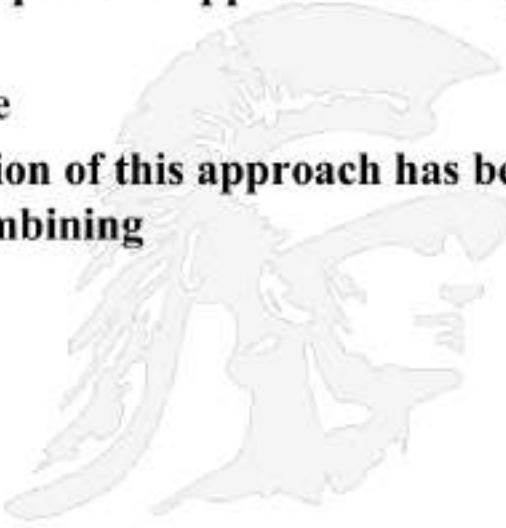
••

University of Southern California  USC

 USC **Viterbi**  
School of Engineering

## 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



Copyright Ellis Horowitz, 2011-2022

---

• •

University of Southern California

**USC Viterbi**  
School of Engineering

USC

## Many Questions Pose Semantic Difficulties

- Who is Michael Jordan?
- Michael Jordan the basketball player or the Machine Learning guy?

- Key requirement is that entities get identified and disambiguated

Google search results for "who is Michael Jordan":

**Michael Jordan - Wikipedia, the free encyclopedia**  
 Michael Jeffrey Jordan (born February 17, 1963) and known by his initials, MJ, is an American former professional basketball player, entrepreneur, and present owner and chairman of the Charlotte Hornets.

**Michael Jordan - Biography - Basketball Player - Biography.com**  
 Michael Jeffrey Jordan, also known by his initials, MJ, is an American former professional basketball player, entrepreneur, and point ball owner and chairman of the Charlotte Hornets.

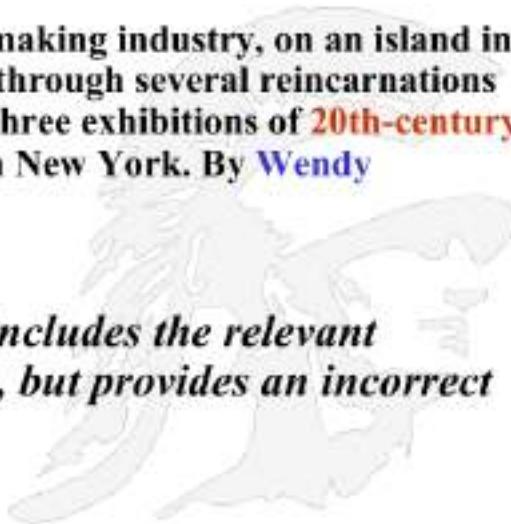
**Is Leon Hall the Michael Jordan of the NFL?**  
 Michael Jeffrey Jordan (born February 17, 1963) and known by his initials, MJ, is an American former professional basketball player, entrepreneur, and point ball owner and chairman of the Charlotte Hornets.

Copyright Ellis Horowitz, 2011-2022

9

••

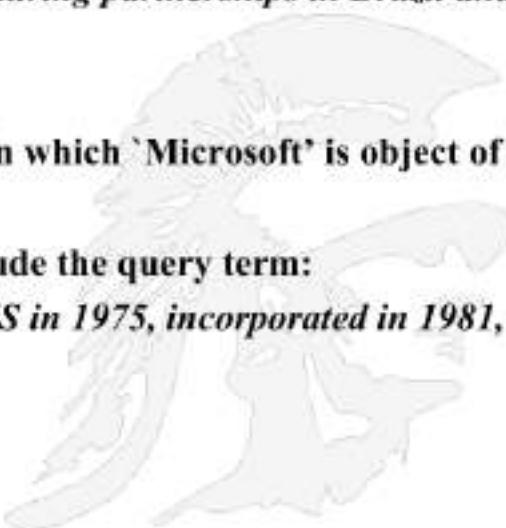
- **Question:** “When was Wendy’s founded?”
- **Passage candidate:**
  - “The renowned Murano glassmaking industry, on an island in the Venetian lagoon, has gone through several reincarnations since it was **founded** in 1291. Three exhibitions of **20th-century** Murano glass are coming up in New York. 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.*



NLP: Natural language processing

Copyright Ellis Horowitz, 2011-2022

11

••

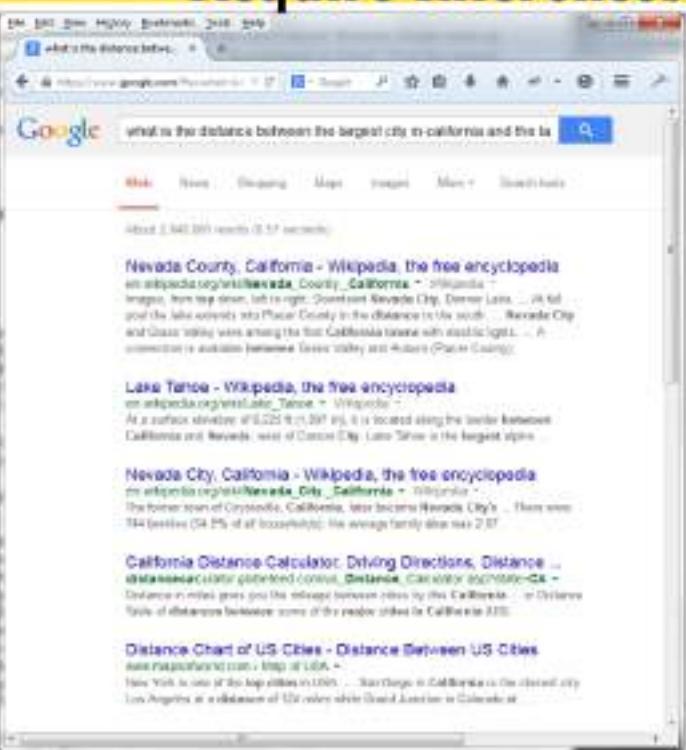
University of Southern California  USC

## Some Questions Require Inferences

**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.



Copyright Ellis Horowitz 2011-2022

• •

**In Some Cases  
the Data May Not Exist**

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

Google search results for "how many Ph.D. degrees in mathematics were granted by European universities in 1986?"

Bing search results for "how many Ph.D. degrees in mathematics were granted by European universities in 1986?"

Chapter 2: Higher Education in Science and Engineering

Ellis Horowitz 2011-2022

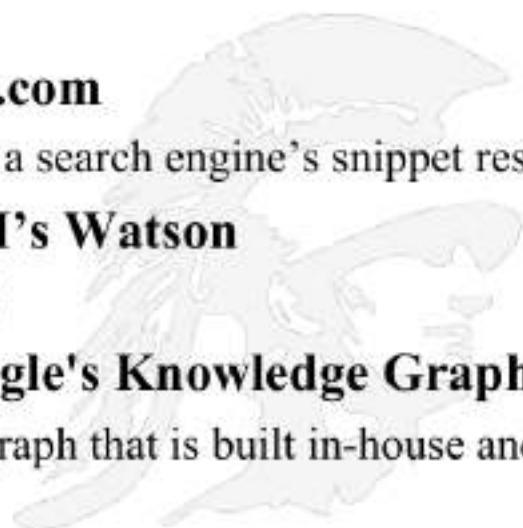
••

University of Southern California  USC

 USC **Viterbi**  
School of Engineering

## 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



Copyright Ellis Horowitz, 2011-2022

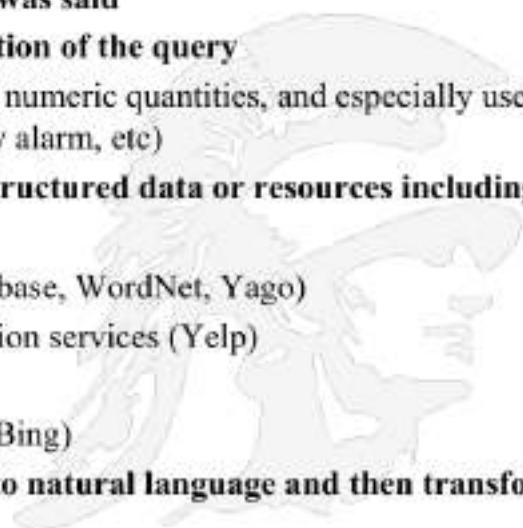
14

• •

University of Southern California  USC

**USC Viterbi**  
School of Engineering

## 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

••

- **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?”



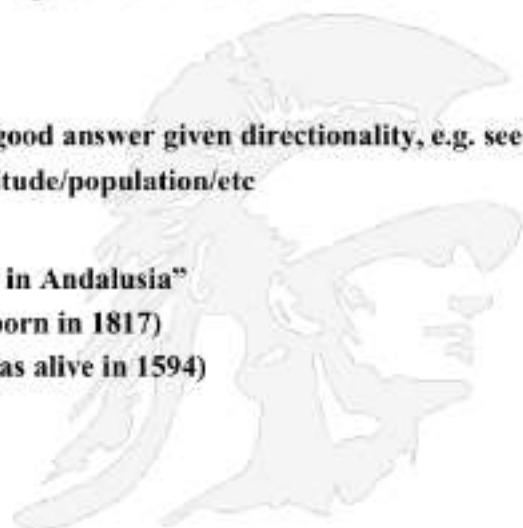
••

University of Southern California  USC

 USC **Viterbi**  
School of Engineering

## 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**



Copyright Ellis Horowitz, 2011-2022

17

• •

University of Southern California 
V USC Viterbi School of Engineering

**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

Snapshot taken 04/2022

How tall is mount Everest

Snapshot taken 04/2022

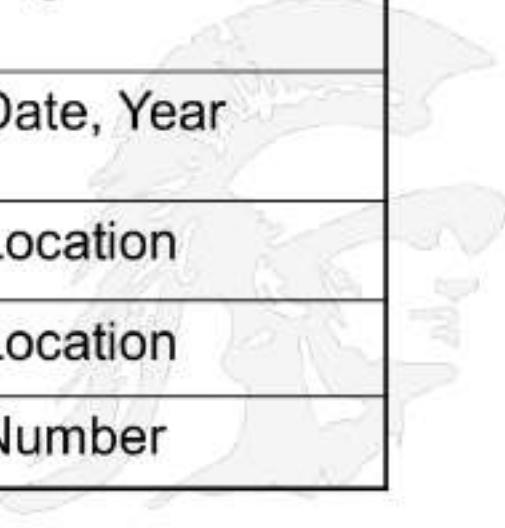
Copyright Ellis Horowitz 2011-2022

••

University of Southern California  USC

**Question Types: Many Questions Fall into Distinct Categories**

Who	Person, Organization
When	Date, Year
Where	Location
In What	Location
How many	Number



Copyright Ellis Horowitz, 2011-2022

19

••

## 3 Main Phases for Question/Answering

### 1. QUESTION PROCESSING

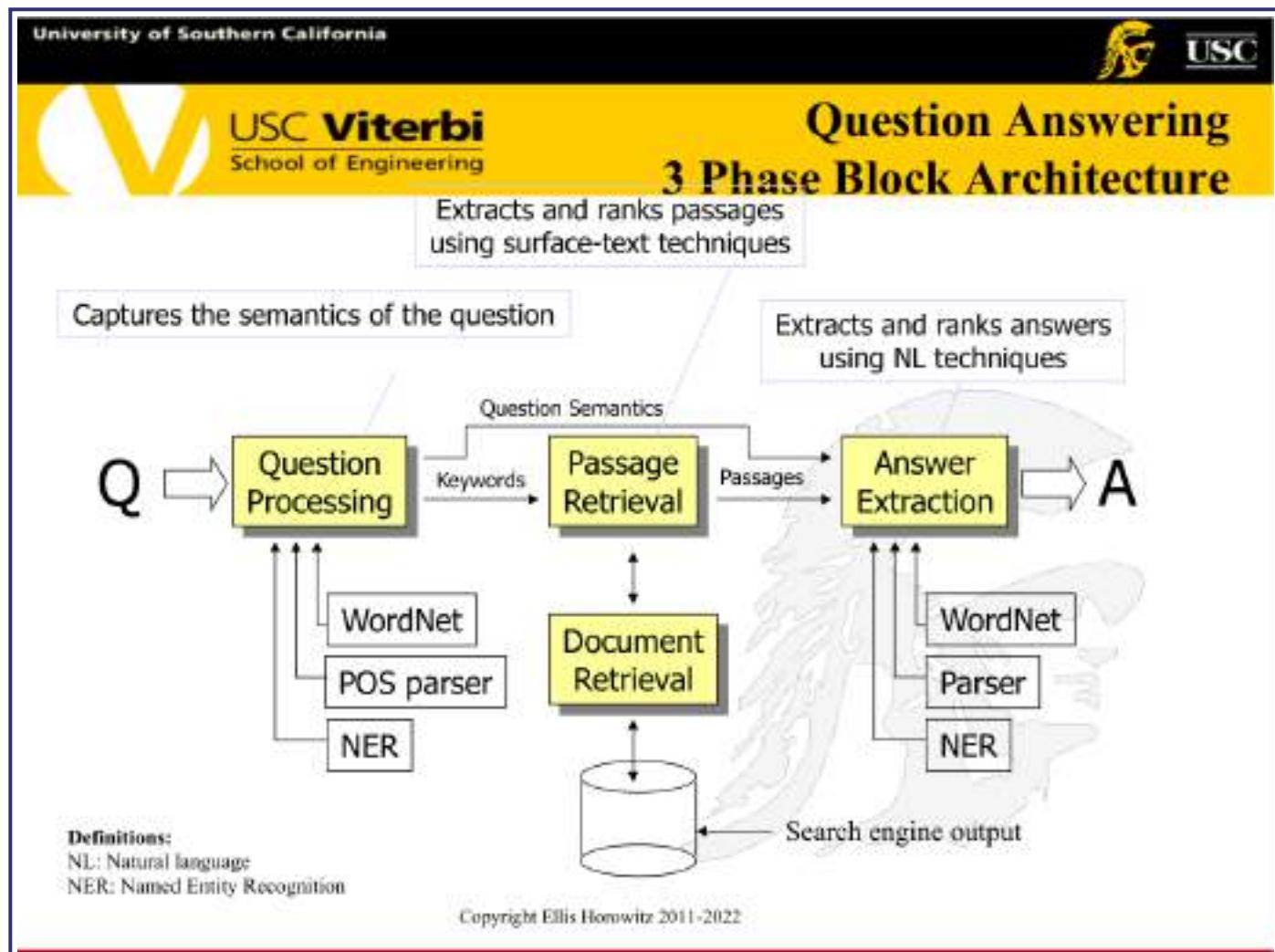
- Detect question type (who, what, when, where, etc)
- Identify important entities and formulate queries to send to a search engine

### 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



• •

University of Southern California  USC

**USC Viterbi**  
School of Engineering

## Question Taxonomy



Question

- Location
  - Country
  - City
  - Province
  - Continent
- Reason
- Product
- Nationality
- Manner
- Number
  - Speed
  - Degree
  - Dimension
  - Rate
  - Duration
  - Percentage
  - Count
- Currency
- Language
- Mammal
- Reptile
- Game
- Organization

Researchers have tried to organize all question types, e.g.

Copyright Ellis Horowitz, 2011-2022

22

• •

University of Southern California



**USC Viterbi**  
School of Engineering



**USC**

## However Question Taxonomies Can Get Very Large

Tag	Example
ABBREVIATION	What's the abbreviation for limited partnership? What does the "e" stand for in the equation E=mc <sup>2</sup> ?
DESCRIPTION	What are tannins? What are the words to the Canadian National anthem? How can you get rid stains out of clothing? What caused the Titanic to sink?
ENTITY	What are the names of Odin's ravens? What part of your body contains the corpus callosum? What colors make up a rainbow? In what book can I find the story of Aladdin? What currency is used in China? What does Salk vaccine prevent? What was involved the battle of Cloudburst? What kind of arts are used in mosaics? What instrument does Max Roach play? What's the official language of Algeria? What letter appears on the cold-water tap in Spain? What is the name of King Arthur's sword? What are some fragrant white climbing roses? What is the fastest computer? What religion has the most members? What was the name of the ball game played by the Mayans? What fuel do airplanes use? What is the classical symbol for mercury? What is the best way to remove wallpaper? How do you say "Grandma" in Irish? What was the name of Captain Bligh's ship? What's the singular of dice?



Copyright Ellis Horowitz, 2011-2022

23

• •

University of Southern California  USC

## 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?

Copyright Ellis Horowitz, 2011-2022

24

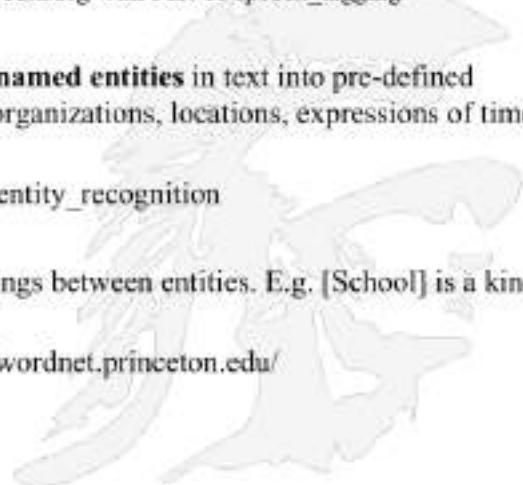
• •

University of Southern California  USC

 USC **Viterbi**  
School of Engineering

## 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**



Copyright Ellis Horowitz, 2011-2022

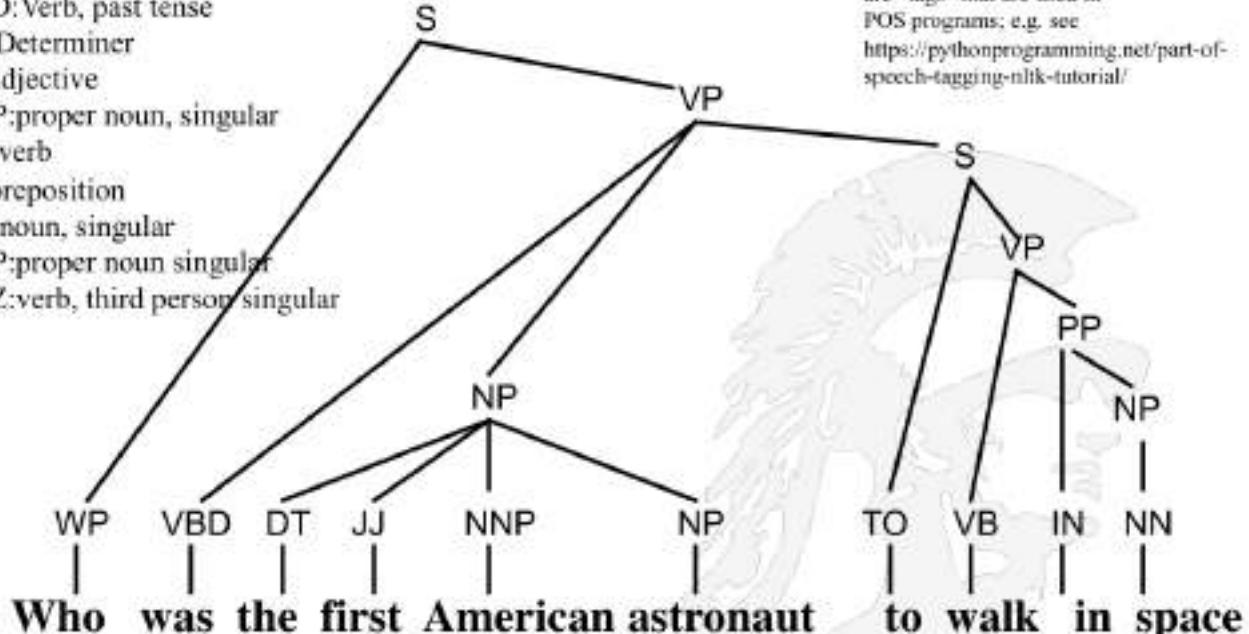
25

••

University of Southern California  USC

**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  
 VBD: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/>

Copyright Ellis Horowitz 2011-2022

• •

University of Southern California  USC

## Question Processing Tool Named Entity Recognizer Example

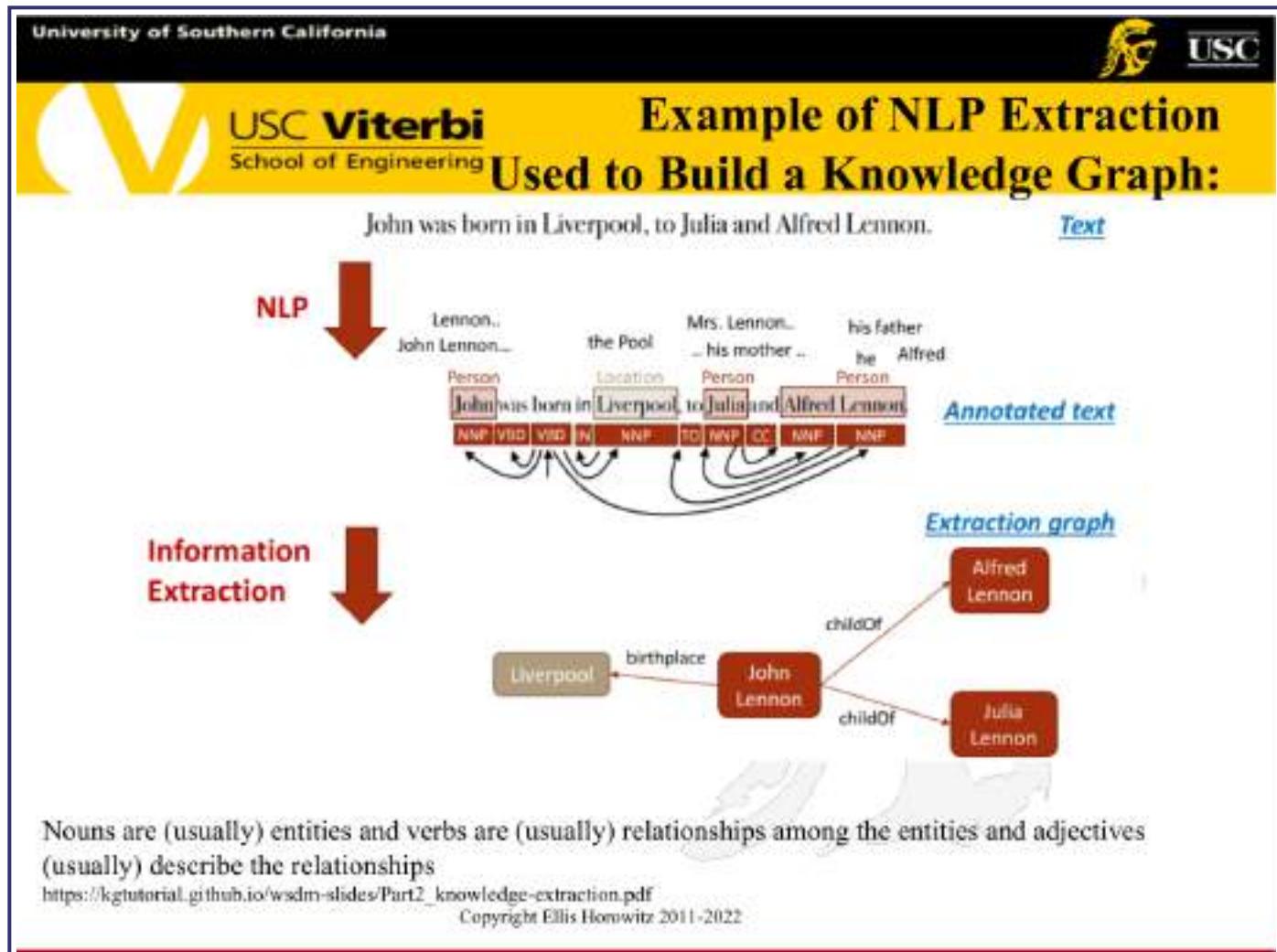
contentSkip to site indexPolicsSubscribelog InSubscribeLog InToday's Paper/AdvertisementSupported ORG byF.B.I. Agent Peter Strzok PERSON

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. CreditJ. 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

Copyright Ellis Horowitz 2011-2022

••



• •

The Jeopardy query:

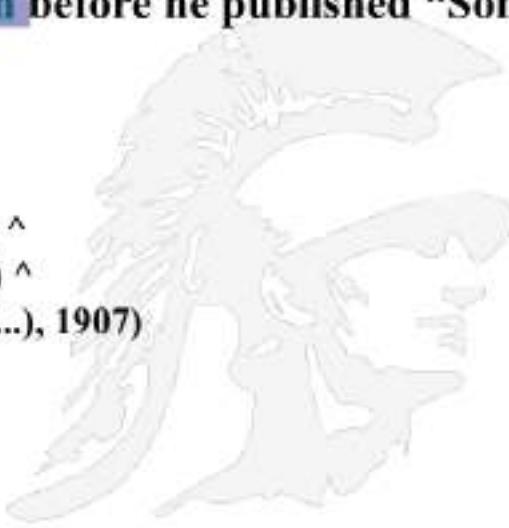
Category: Poets and Poetry: *GEO*

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)
```



• •

University of Southern California 

## 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”)



*Songs of a Sourdough*

From Wikipedia, the free encyclopedia

For the 1914 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*.

This book is well known for its verse about the Klondike Gold Rush in the Yukon a decade earlier, particularly the long, famous ballads, "The Shooting of Dan McGrew" and "The Death of Tom McRelley".

*Songs of a Sourdough* has sold more than three million copies.<sup>1</sup>

**Categories:**

1. Poetry
2. Comedy
3. Fiction
4. Romance

Copyright Ellis Horowitz, 2007 - [www.cs.usc.edu/~horowitz/](http://www.cs.usc.edu/~horowitz/)

••

**1. Use the part-of-speech recognizer to identify all**

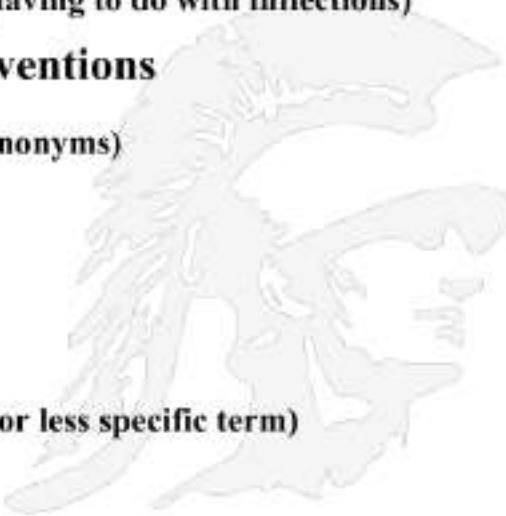
- nouns
- verbs
- non-stopwords in quotations
- NNP (proper noun) words in recognized named entities
- complex nominals with their adjectival modifiers

**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)



• •

University of Southern California  USC

**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;

**Hyponym** is a word or phrase whose semantics is more specific than its hypernym  
e.g. purple is a hyponym of color

Copyright Ellis Horowitz, 2011-2022

33

••

University of Southern California 

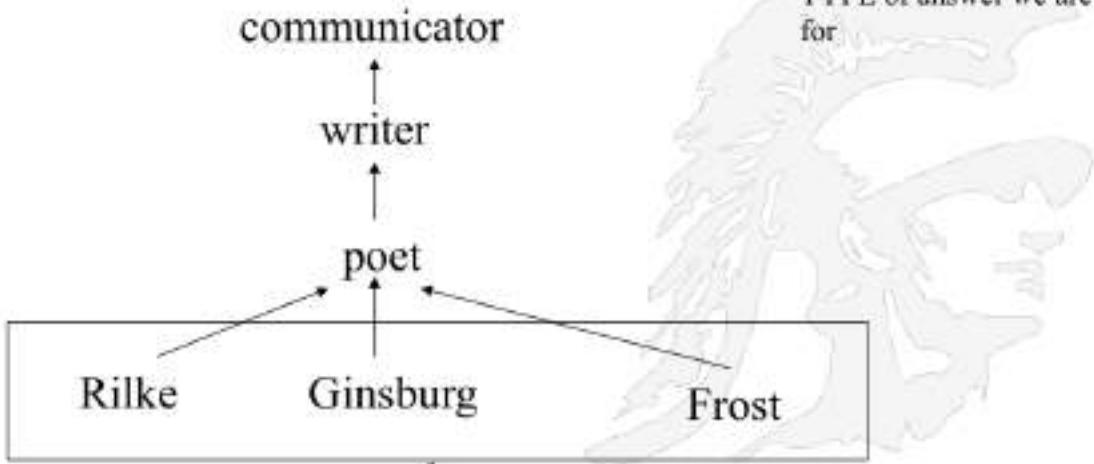
## Use WordNet for Type Identification

We have already seen the use of WordNet, a lexical database of English nouns, verbs, adjectives, adverbs

WordNet permits refinement of poet to specific instances

“What 20<sup>th</sup> century poet wrote Howl?”

WordNet also helps determine the TYPE of answer we are looking for



communicator  
writer  
poet  
Rilke      Ginsburg      Frost

Original keyword candidate set

Copyright Ellis Horowitz, 2011-2022

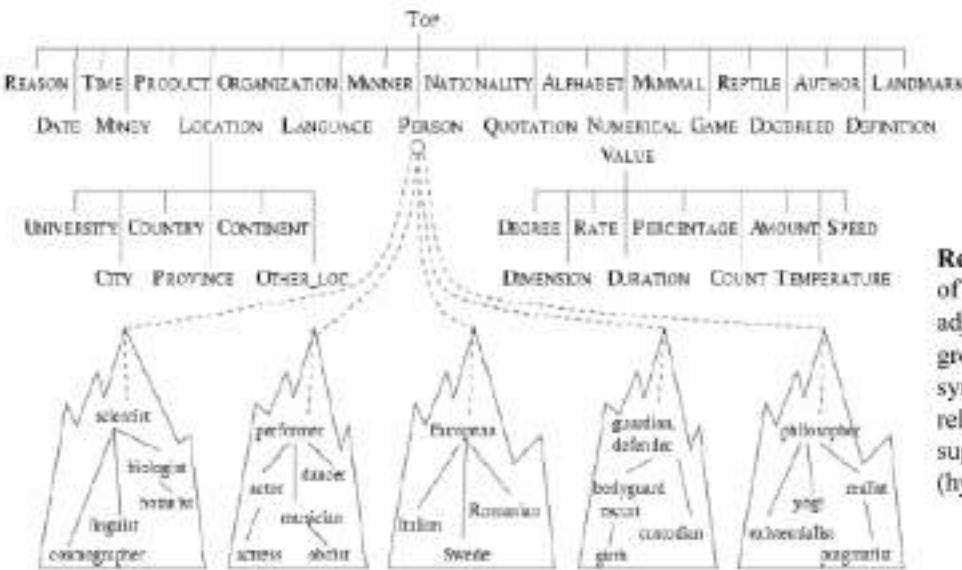
34

•

University of Southern California 

## 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

Copyright Ellis Horowitz, 2011-2022

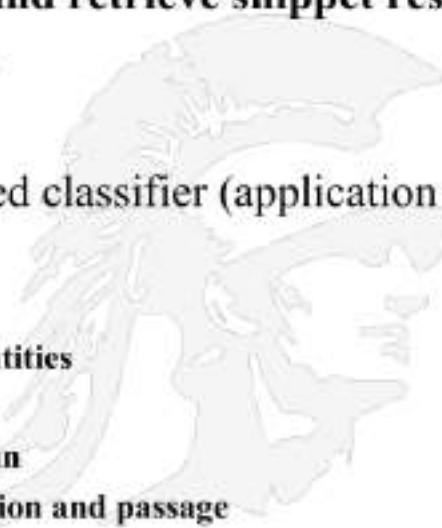
35

••



## 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

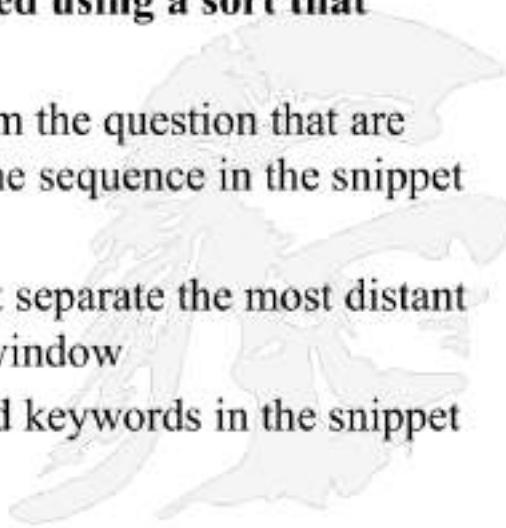


..



## 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



••

University of Southern California  USC

**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 . . ."



Copyright Ellis Horowitz, 2011-2022

38

••

University of Southern California  USC

**Ranking Candidate Answers**

**Q066: Name the first private citizen to fly in space.**

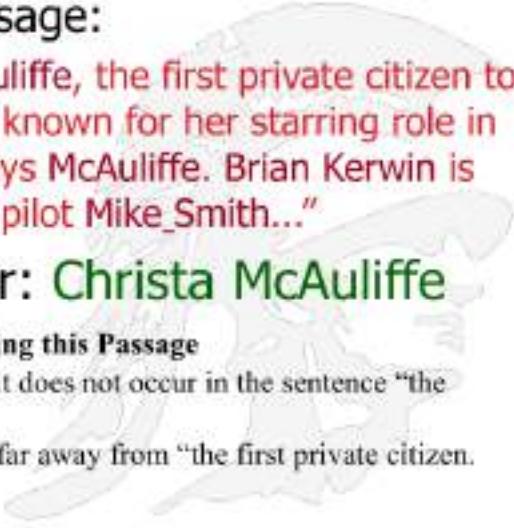
- Answer type: Person
  - Text passage:

"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."



Copyright Ellis Horowitz, 2011-2022

39

• •

University of Southern California



## Local Alignment Example (1 of 7)

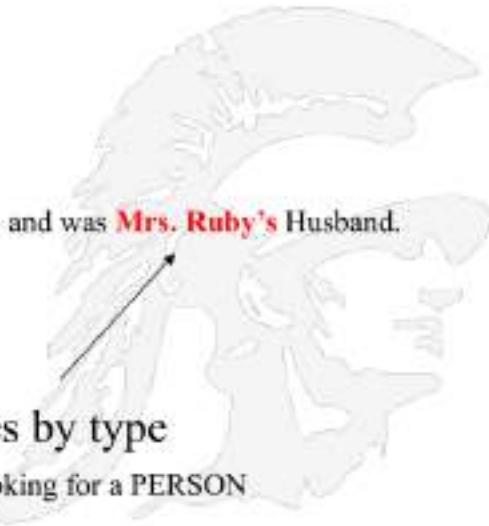
Who shot Kennedy?

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



Copyright Ellis Horowitz, 2011-2022

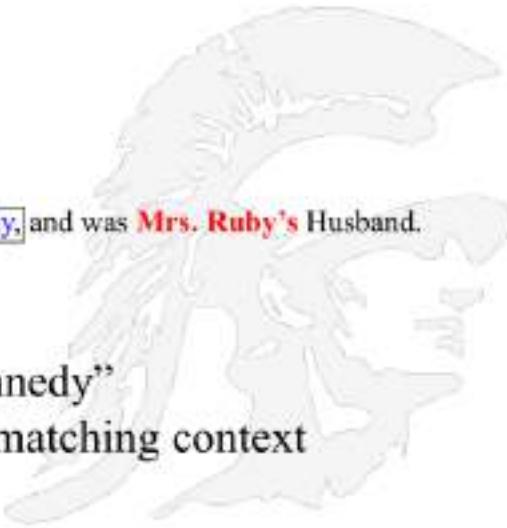
40

• •



## 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

• •

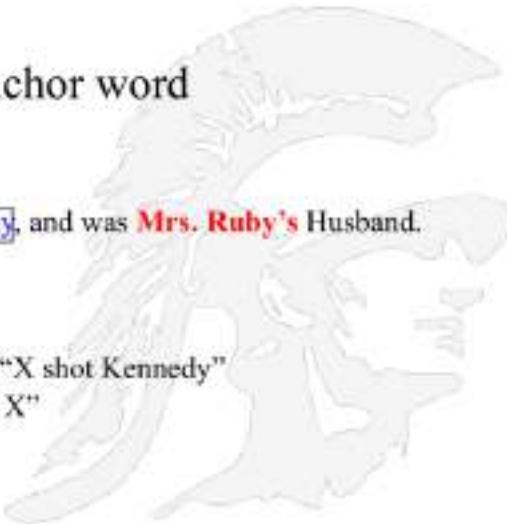


## Local Alignment Example (3 of 7)

*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"

• •

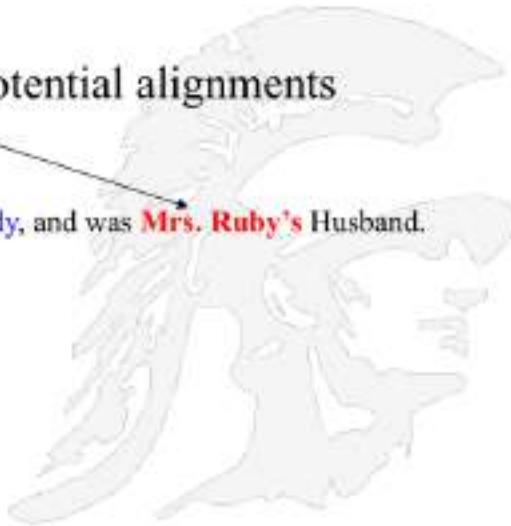


In principle it can be anyone of the three people identified

### *Who shot Kennedy?*

Potential alignments

Jack assassinated Oswald, the man who shot Kennedy, and was Mrs. Ruby's Husband.



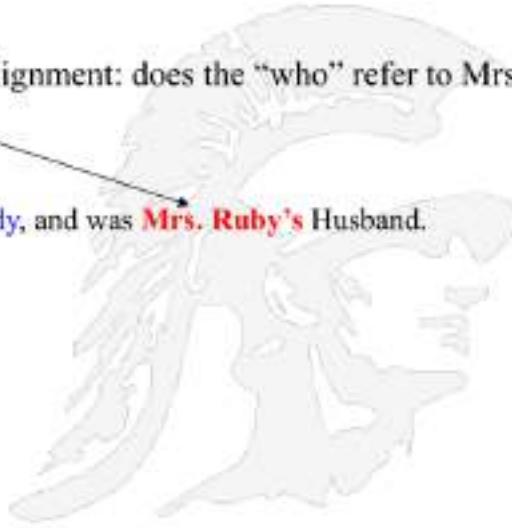
• •



### *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 :

• •

University of Southern California  USC

**Local Alignment Example (6 of 7)**

**Who shot Kennedy?**

The distance between the question head word "who" and the anchor word Kennedy is 1

One Alignment : does the "who" refer to Mrs. Ruby?  
The distance from Kennedy to Mrs. Ruby

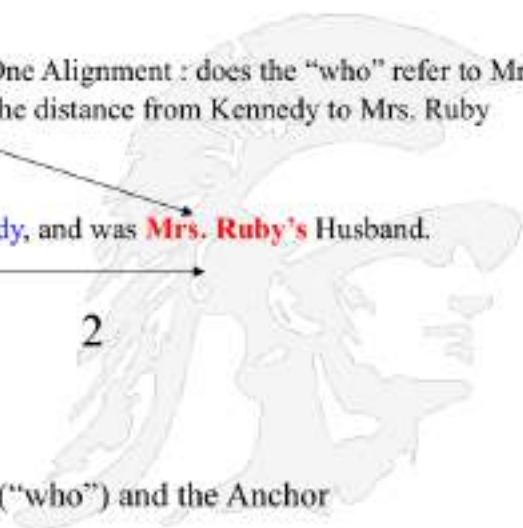
Jack assassinated Oswald, the man who shot Kennedy, and was Mrs. Ruby's Husband.

Three Alignment Features :

1. Distance between Question Head word ("who") and the Anchor word ("Kennedy") in the sentence

Copyright Ellis Horowitz, 2011-2022

45



• •



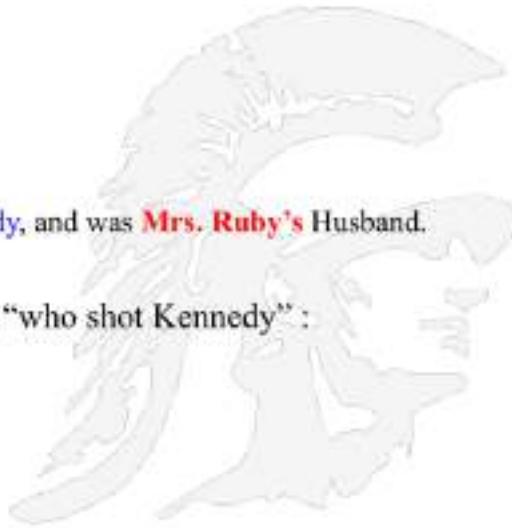
## Local Alignment Example (7 of 7)

**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

• •

## What is BERT

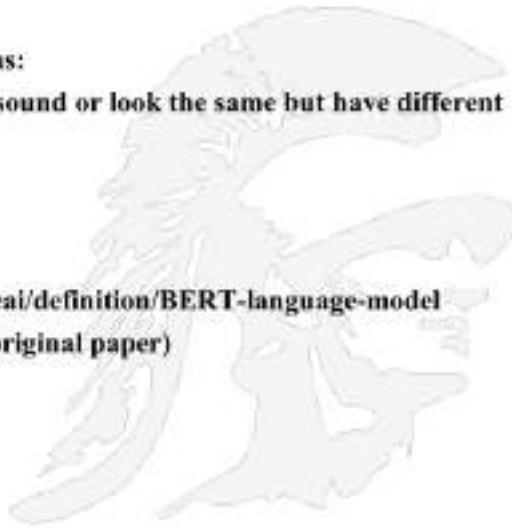
- *Bidirectional Encoder Representations from Transformers*
- In 2018, Google introduced and open-sourced BERT
  - BERT is designed to help computers understand the meaning of ambiguous language in text by using surrounding text to establish context
  - It achieved strong results on problems such as sentiment analysis, semantic role labeling, sentence classification and the disambiguation of polysemous words, or words with multiple meanings
- In October 2019, Google announced that they would begin applying BERT to their United States based production search algorithms.
- BERT was pre-trained using only an unlabeled, plain text corpus (namely the entirety of the English Wikipedia).
- It continues to learn unsupervised from the unlabeled text and improve even as it's being used in practical applications
- the BERT Github repository is available on [github, git clone https://github.com/google-research/bert.git](https://github.com/google-research/bert.git)

• •



## What is BERT used for

- Sequence-to-sequence based language generation tasks such as:
  - Question answering
  - Abstract summarization
  - Sentence prediction
  - Conversational response generation
- Natural language understanding tasks such as:
  - Polysemy and Coreference (words that sound or look the same but have different meanings) resolution
  - Word sense disambiguation
  - Natural language inference
  - Sentiment classification
- <https://www.techtarget.com/searchenterpriseai/definition/BERT-language-model>
- <https://csci572.com/papers/BERT.pdf> (the original paper)

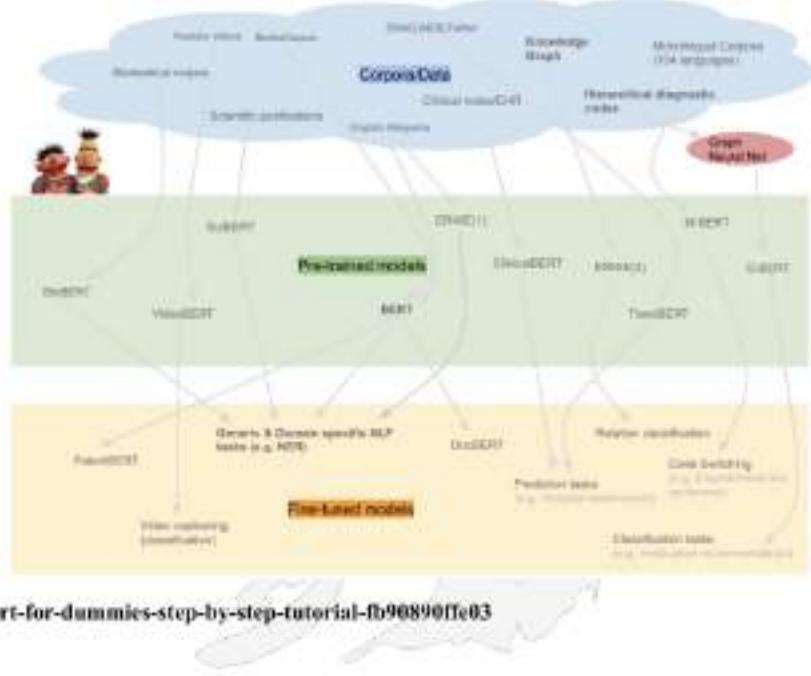


• •

University of Southern California  USC

## BERT Has Many Pre-Trained Models

- BERT encoders have larger feedforward networks (768 and 1024 nodes in Base and Large respectively) and more attention heads (12 and 16 respectively).
- BERT was trained on Wikipedia and other datasets
- To the right you can see a diagram of additional variants of BERT pre-trained on specialized corpora



<https://towardsdatascience.com/bert-for-dummies-step-by-step-tutorial-fb90890ffe03>

Copyright Ellis Horowitz, 2011-2022

50

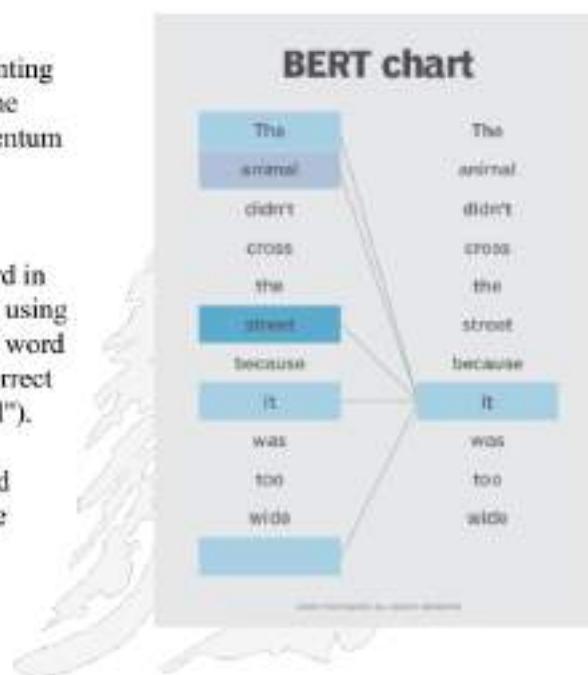
• •

University of Southern California  USC

**How BERT works**

- BERT functions by reading bidirectionally, accounting for the effect of all other words in a sentence on the focus word and eliminating the left-to-right momentum that biases words towards a certain meaning as a sentence progresses.
- At the right BERT is determining which prior word in the sentence the word "it" is referring to, and then using its attention mechanism to weigh the options. The word with the highest calculated score is deemed the correct association (i.e., "it" refers to "street", not "animal").
- If this phrase was a search query, the results would reflect this subtler, more precise understanding the BERT reached.

**BERT chart**



Copyright Ellis Horowitz 2011-2022

• •



## BERT for Question-Answering

- question answering is just a prediction task
- on receiving a question as input, the goal of the application is to identify the right answer from some corpus
- given a question and a context paragraph, the model predicts a start and an end token from the paragraph that most likely answers the question.

- Input Question:

Where do water droplets collide with ice crystals to form precipitation?

- Input Paragraph:

... Precipitation forms as smaller droplets coalesce via collision with other rain drops or ice crystals within a cloud. ...

- Output Answer:

within a cloud



Copyright Ellis Horowitz, 2011-2022

52

• •

University of Southern California  USC 

## Microsoft's AskMSR Answering System

**AskMSR is a question answering system developed at Microsoft**

- 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)



Download PDF

Abstract · Related Info

Authors

Eric Bell · Susan Dumais · Michele Bernick

Published In

Proceedings of EMNLP 2002

An Analysis of the AskMSR Question-Answering System  
January 1, 2002

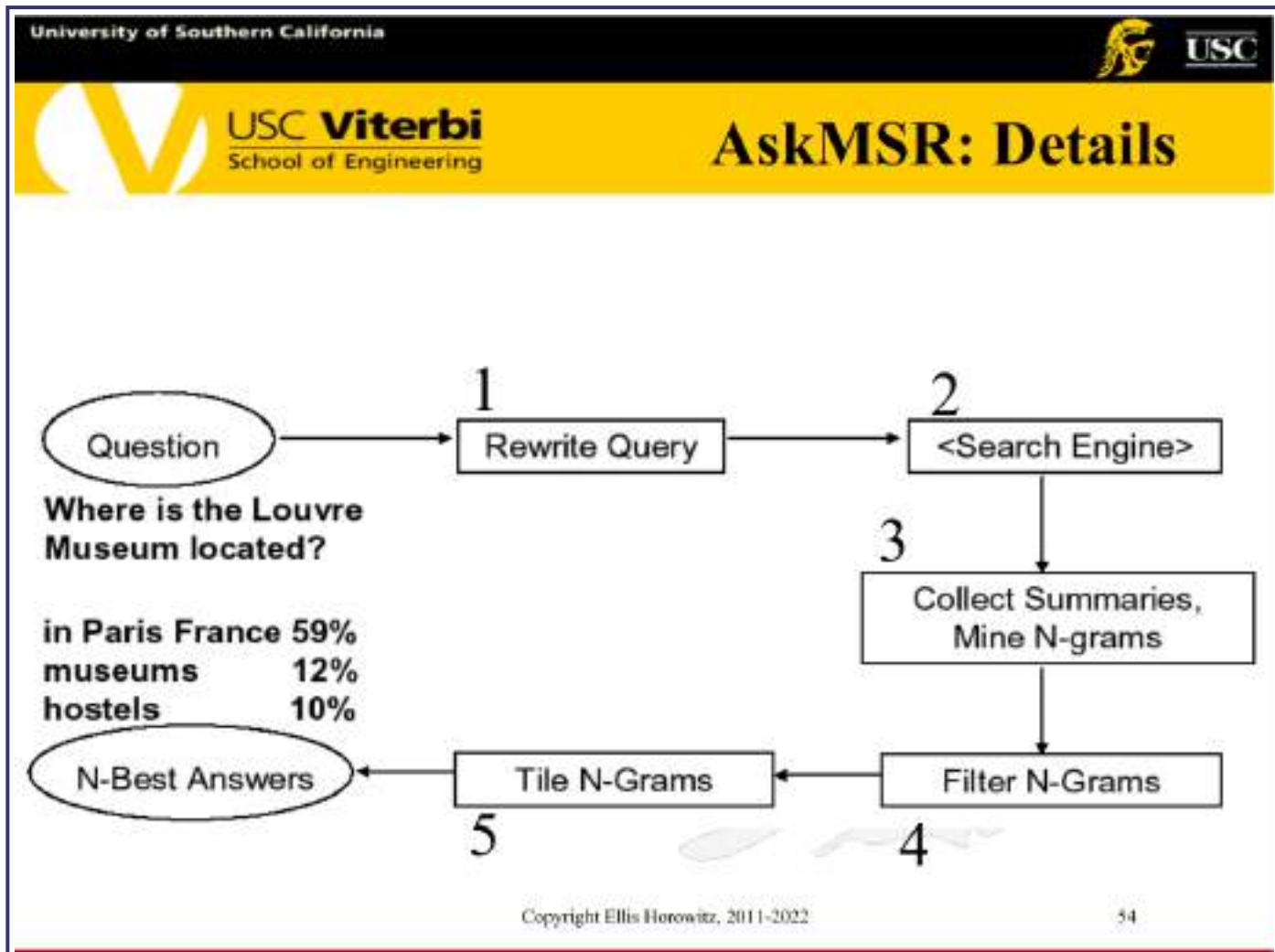
Eric Bell, Susan Dumais, Michele Bernick

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.

Copyright Ellis Horowitz, 2011-2022

53

••



••

University of Southern California  USC

**AskMSR:**  
**Step 1:Query Rewriting**

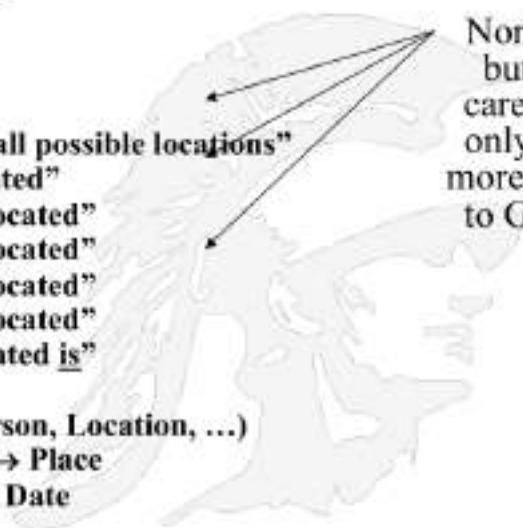
**• Classify question into categories**

- **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



Nonsense, but who cares? It's only a few more queries to Google.

Copyright Ellis Horowitz, 2011-2022

55

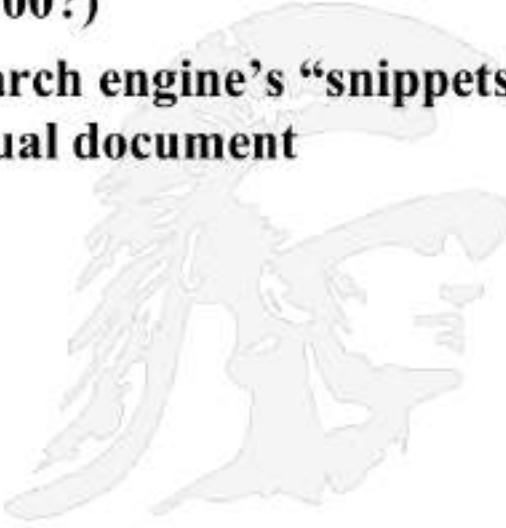
••

University of Southern California  USC

 USC **Viterbi**  
School of Engineering

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**



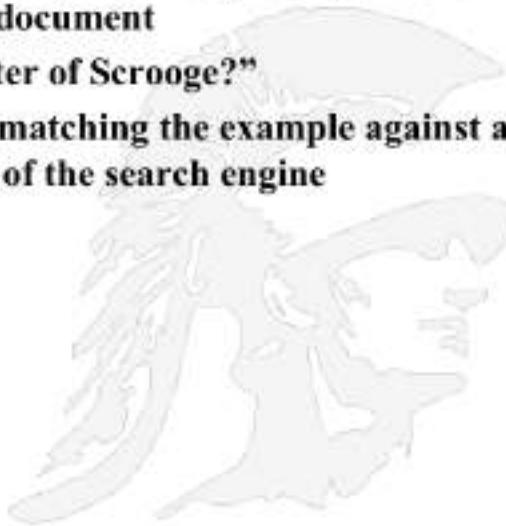
Copyright Ellis Horowitz, 2011-2022

56

••

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



••

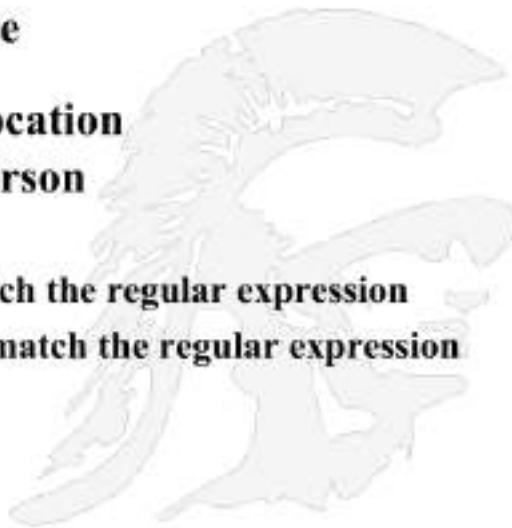
University of Southern California



USC

USC Viterbi  
School of EngineeringAskMSR:  
**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 ... → Person
- 
- Boost score of n-grams that do match the regular expression
  - Lower score of n-grams that don’t match the regular expression



Copyright Ellis Horowitz, 2011-2022

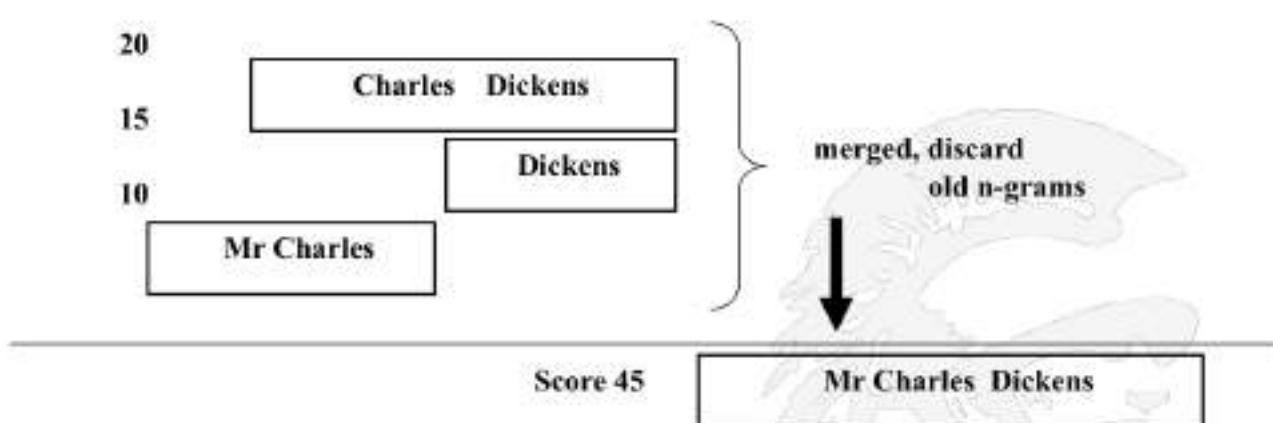
58

• •

University of Southern California  USC

**AskMSR:**  
**5: Tiling the Answers**

**Scores**



20	Charles Dickens
15	Dickens
10	Mr Charles

Score 45      Mr Charles Dickens

tile highest-scoring n-gram

N-Grams      N-Grams

Repeat, until no more overlap

Copyright Ellis Horowitz, 2011-2022

59

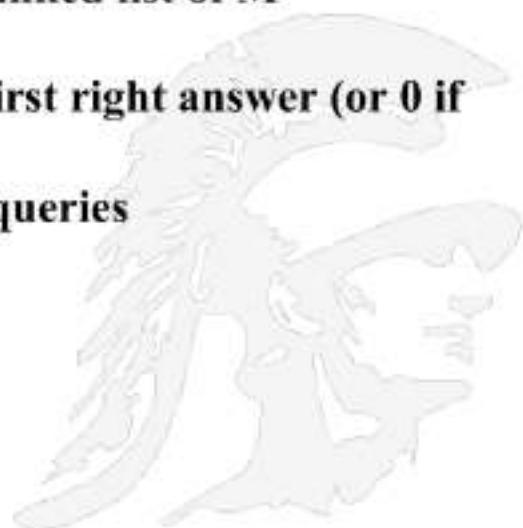
••

University of Southern California  USC

 USC **Viterbi**  
School of Engineering

## 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
  - $$\text{MRR} = \frac{1}{N} \sum_{i=1}^{i=N} \frac{1}{\text{rank}_i}$$



1/26

\*\*\* 09:28:15

# Clustering

## (for classification)

• •



# Clustering



Copyright Ellis Horowitz 2011 -2022

••



USC **Viterbi**  
School of Engineering

# Today's Topic: Clustering

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



Copyright Ellis Horowitz, 2011-2022

2

..



## 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

• •

## Related Searches are a Form of Clustering

**Google related searches**

**Yahoo does some clustering via alternate queries**

**Bing does a little better**

Copyright Ellis Horowitz, 2011-2022

••

**USC Viterbi School of Engineering**

# yippy.com Search Engine

- Yippy (formerly Clusty) is a metasearch engine developed by Vivisimo which emphasizes clusters of results.

initial screen with query "cars"

clustered results appear on the left column: e.g.  
sale  
reviews  
dealers  
rentals

multiple level clusters:  
car dealers  
trucks  
ebay

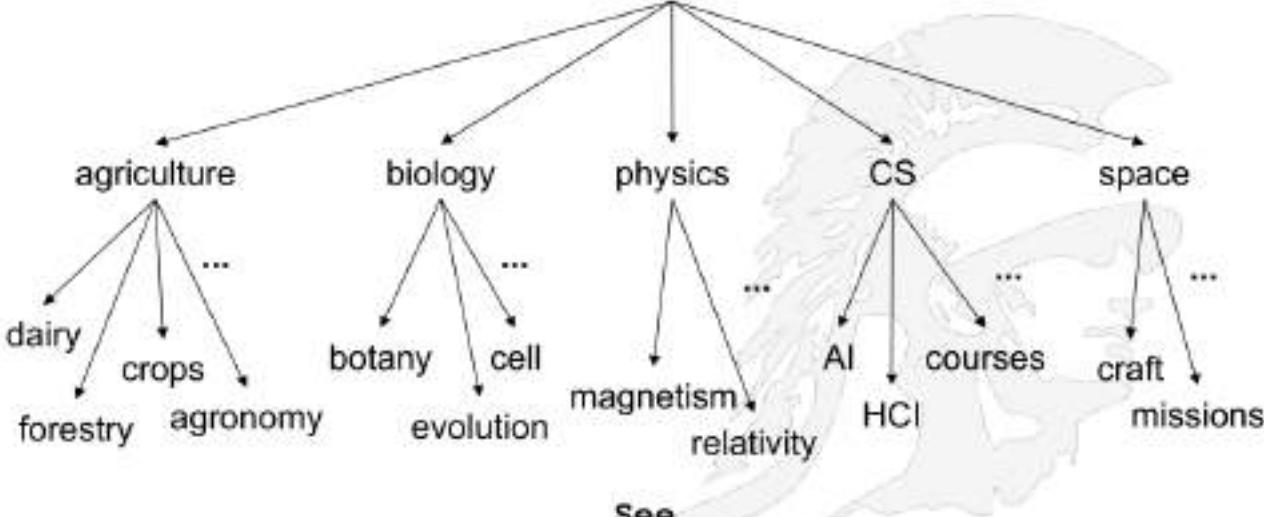
Copyright Ellis Horowitz, 2011-2022

••

 **USC Viterbi**  
School of Engineering

## 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**



```
graph TD; Root[] --> Agriculture[agriculture]; Root --> Biology[biology]; Root --> Physics[physics]; Root --> CS[CS]; Root --> Space[space]; Agriculture --> Dairy[dairy]; Agriculture --> Crops[crops]; Agriculture --> Forestry[forestry]; Agriculture --> Agronomy[agronomy]; Biology --> Botany[botany]; Biology --> Evolution[evolution]; Biology --> Cell[cell]; Physics --> Magnetism[magnetism]; Physics --> Relativity[relativity]; CS --> AI[AI]; CS --> HCI[HCI]; Space --> Craft[craft]; Space --> Missions[missions];
```

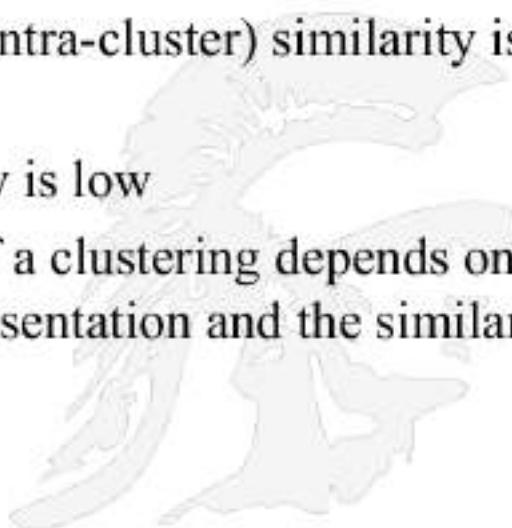
See <https://searchengineland.com/yahoo-directory-close-204370>

Copyright Ellis Horowitz 2011-2022

..



- **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

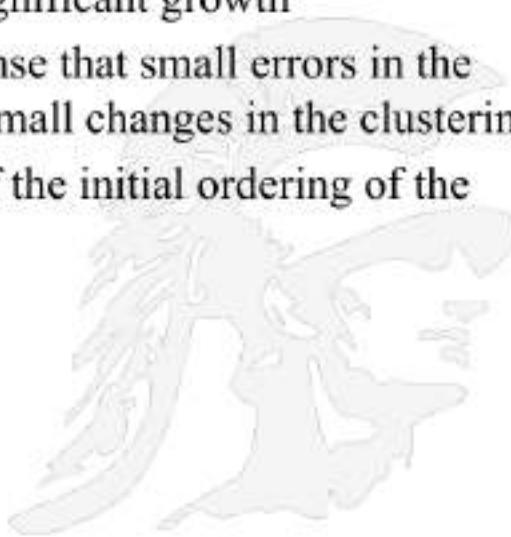


••



## 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

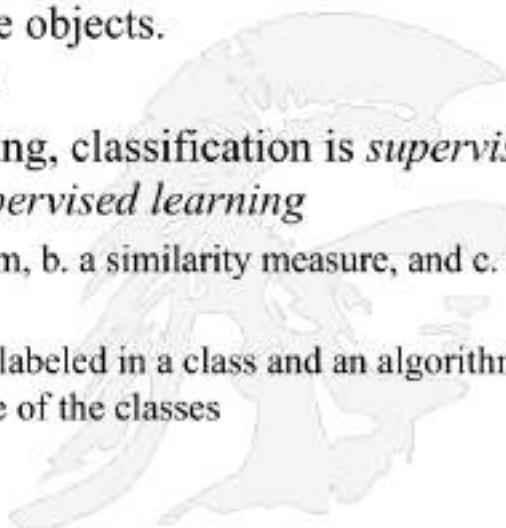


••



## 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)**

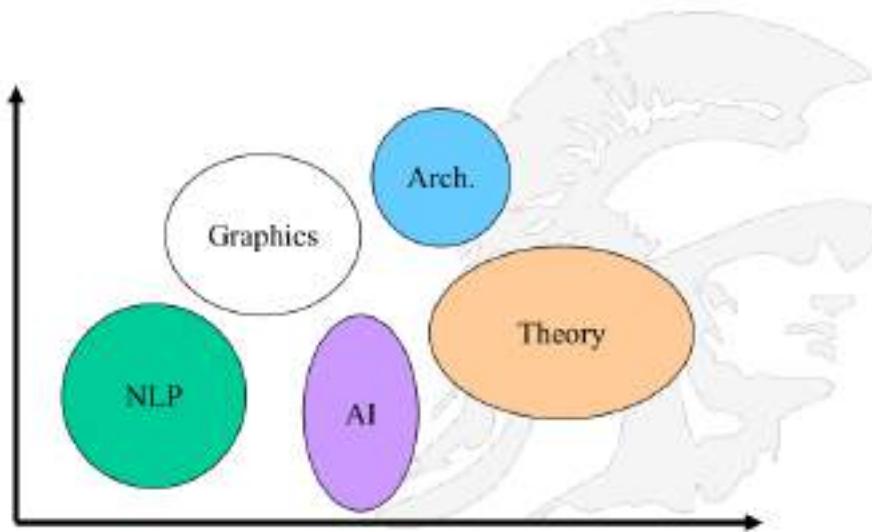


Copyright Ellis Horowitz, 2011-2022

15



- Step 2: we label the clusters
  - choosing a popular name from each document cluster



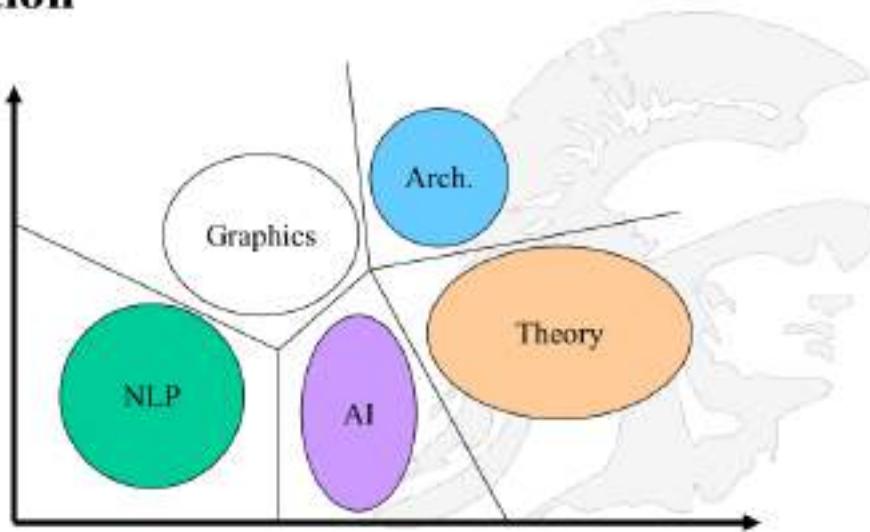
Copyright Ellis Horowitz, 2011-2022

16



## Still Clustering: Determine Decision Boundaries

- Step 3: we compute boundaries for the clusters that can be used as new documents appear; i.e. classification



Copyright Ellis Horowitz, 2011-2022

17

••



## Classification Requires Initial Clusters and Boundaries

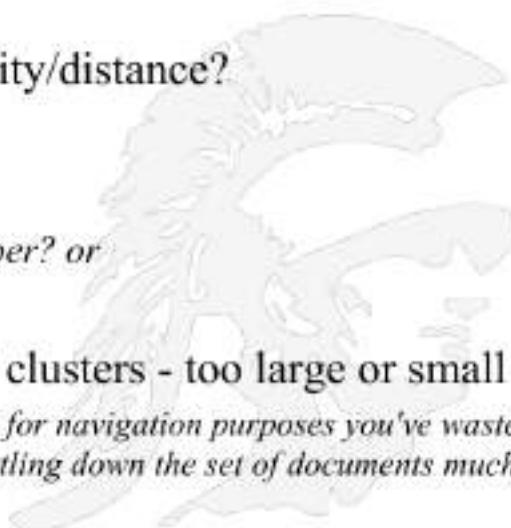
- **Definition:** *Supervised Learning*, inferring a function from labeled training data
- 1. **The documents in each cluster define the “training” docs for each category**
  - E.g. in computer science named clusters would include: Algorithms, Theory, AI, Databases, Operating Systems, NLP, etc.
- 2. **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*

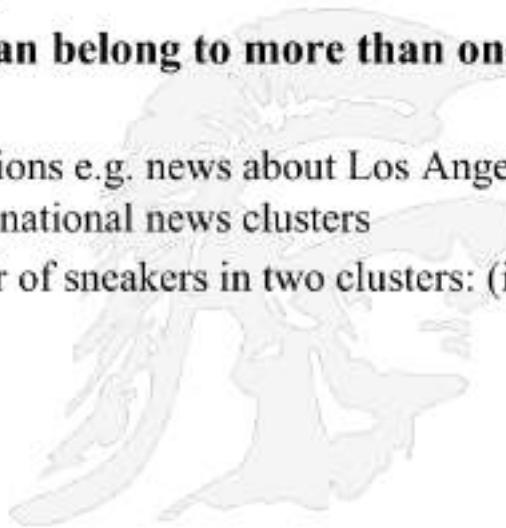


••



## 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



Copyright Ellis Horowitz, 2011-2022

20

••



## 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)

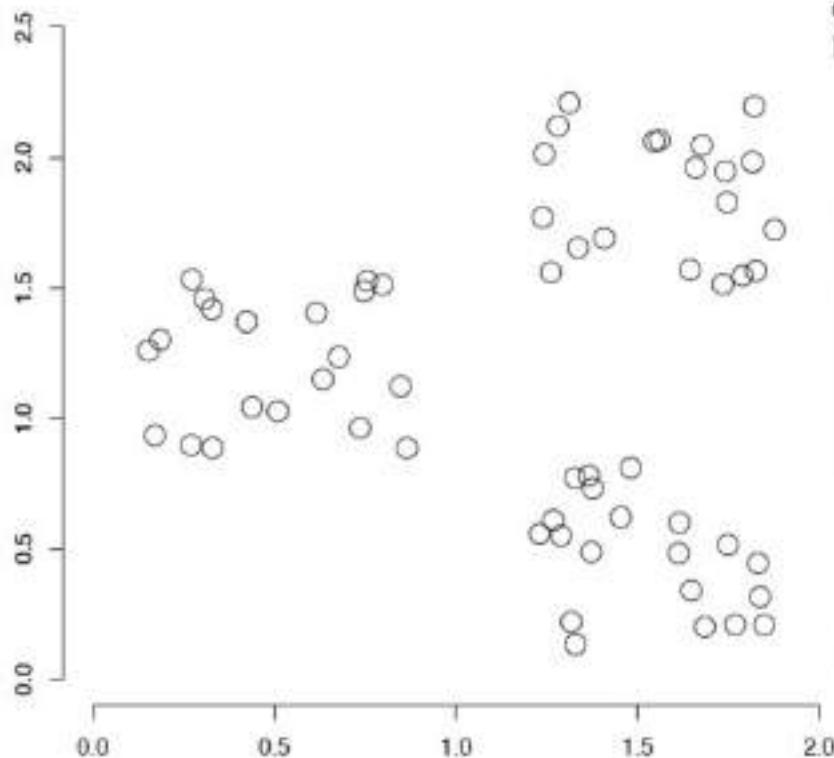
$$\text{similarity} = \cos(\theta) = \frac{A \cdot B}{\|A\| \|B\|} = \frac{\sum_{i=1}^n A_i \times B_i}{\sqrt{\sum_{i=1}^n (A_i)^2} \times \sqrt{\sum_{i=1}^n (B_i)^2}}$$

- 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**

22

••



USC **Viterbi**  
School of Engineering

## 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)

Copyright Ellis Horowitz, 2011-2022

23

••



## A Partitioning Algorithm K-Means Clustering Algorithm

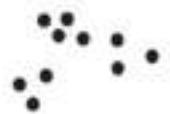
- **Clustering algorithm strategy**
  - Choose  $k$  random data items out of the  $n$  items; call these items the **means**; they designate the prototype or name of the cluster
  - **Refine it iteratively**
    - Associate each of the  $n-k$  items with one of the  **$k$  clusters** choosing the **cluster** 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

Copyright Ellis Horowitz, 2011-2022

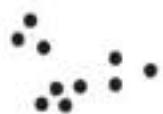
24



## 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*



Copyright Ellis Horowitz 2011-2022

• •



## K-Means Clustering Algorithm Mathematical Formulation

(stated mathematically)

Given an initial set of  $k$  means  $m_1^{(0)}, \dots, m_k^{(0)}$ , the algorithm proceeds by alternating between two steps:

**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  $S^{(t)}$ , 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.

$$m_i^{(t+1)} = \frac{1}{|S_i^{(t)}|} \sum_{x_j \in S_i^{(t)}} x_j$$

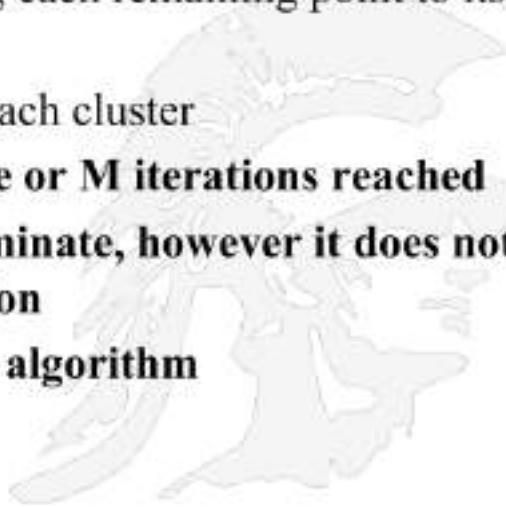
- The algorithm has converged when the assignments no longer change.
- The algorithm will converge to a (local) optimum.
- There is no guarantee that the global optimum is found using this algorithm.



Copyright Ellis Horowitz 2011-2022



- 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

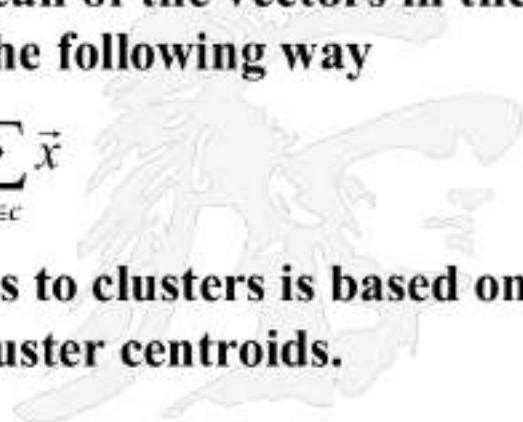




- Assumes instances are real-valued vectors
  - Let  $\vec{x}$  represent the vectors in a cluster  $c$
- Then we define the *centroids, (or center of gravity),* of the cluster to be the mean of the vectors in the cluster; we write this in the following way

$$\bar{\mu}(c) = \frac{1}{|c|} \sum_{\vec{x} \in c} \vec{x}$$

- Reassignment of instances to clusters is based on distance to the current cluster centroids.



Copyright Ellis Horowitz, 2011-2022

29

Here is a demo of k-means clustering; this is a copy where you can modify the code (eg. alter the number of clusters, data points, etc.).

••



## 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):

$$1 - \frac{\vec{x} \cdot \vec{y}}{|\vec{x}| \cdot |\vec{y}|}$$

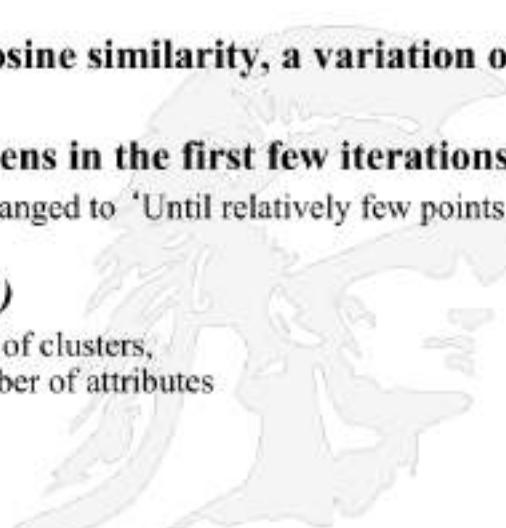


••



## 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

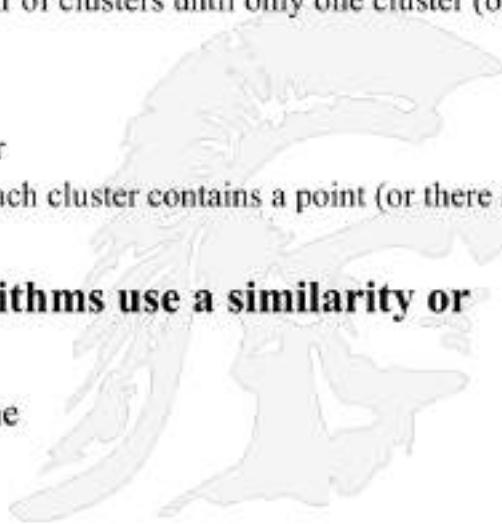


••



## 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



••



## 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

Copyright Ellis Horowitz, 2011-2022

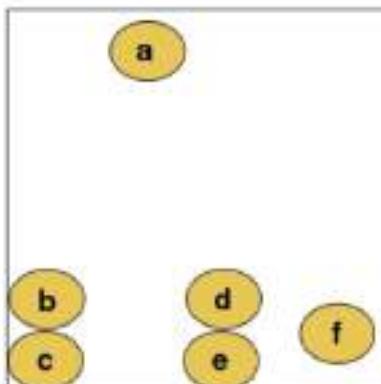
41

••

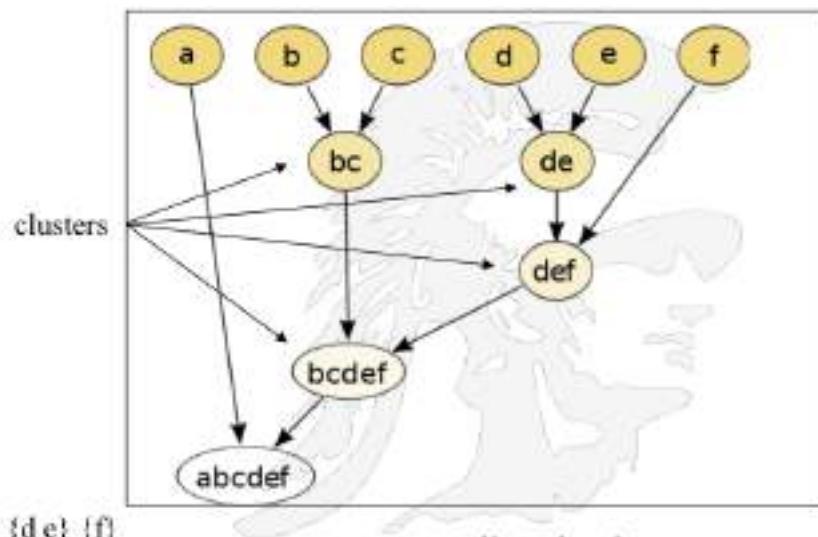


## A Dendrogram is Used to Display Clusters

- A **dendrogram** is a tree diagram frequently used to illustrate the arrangement of the clusters produced by hierarchical clustering



original input



corresponding dendrogram

second row clusters are: {a}, {b c}, {d e}, {f}  
third row clusters are: {a}, {b c}, {d e f}

Copyright Ellis Horowitz, 2011-2022

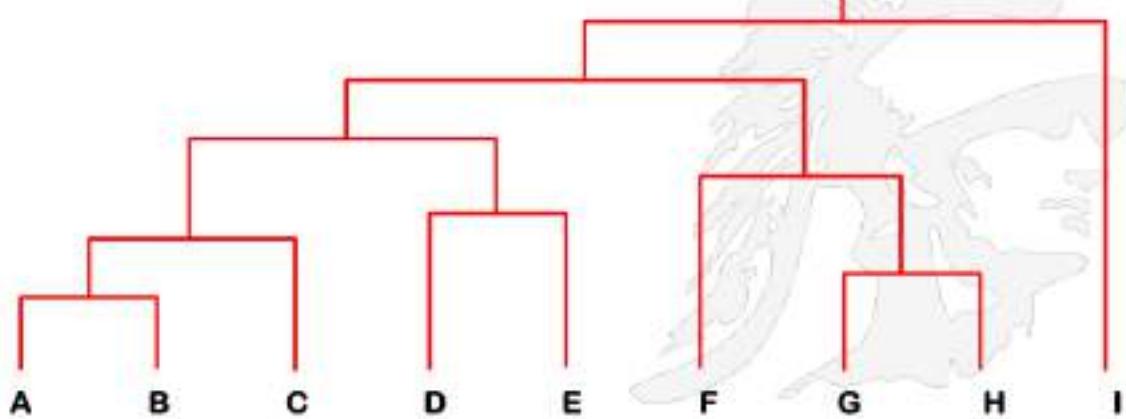
42

••



## 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



Copyright Ellis Horowitz, 2011-2022

43

••



## 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.

1/31

\*\*\* 09:28:29

# How to classify a document?

••



## 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

$$\bar{\mu}(C) = \frac{1}{|C|} \sum_{d \in C} \vec{d}$$

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

Copyright Ellis Horowitz 2011-2022

••

The slide features three diagrams on a yellow background. The top right diagram is titled "Centroid Example". The bottom left diagram contains the text: "The boundary between two classes in Rocchio classification is the set of points with equal distance from the two centroids". The middle diagram shows three regions (red, blue, green) each with its own centroid (black dot). The right diagram shows a single region with a centroid (black dot) and text stating: "3 regions each with its own centroid" and "The centroid may not correspond to a document in your set".

USC Viterbi  
School of Engineering

## 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:

$$q_{opt} = \frac{1/|D_r| \sum_{d_j \text{ in } D_r} d_j - 1/|D_{nr}| \sum_{d_j \text{ in } D_{nr}} d_j}{\sqrt{|D_r| + |D_{nr}|}}$$

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  $\vec{q}$  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  $\alpha = 1.0$ ,  $\beta = 0.75$ ,  $\gamma = 0.15$

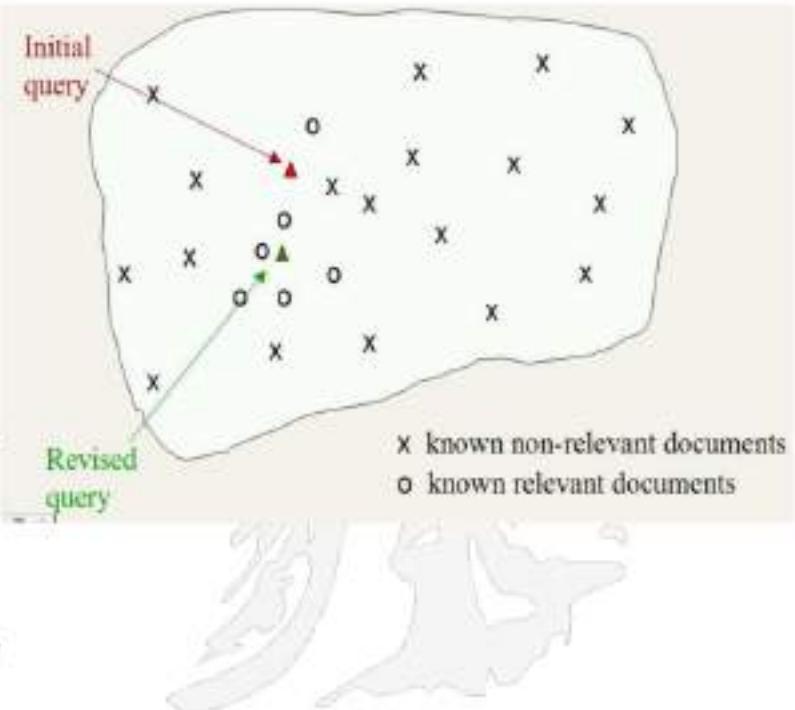
• Note: if the final value of  $\vec{q}_m$  has negative term weights, set those to 0.

••



## Rocchio in Practice

- 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  $\gamma = 0$ .



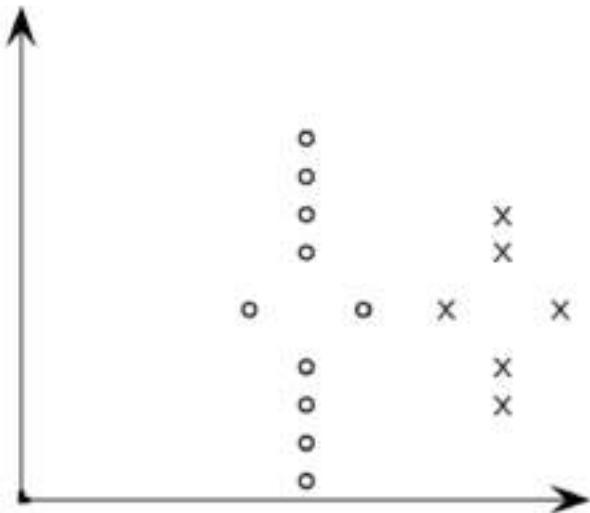
Copyright Ellis Horowitz, 2011-2022

18

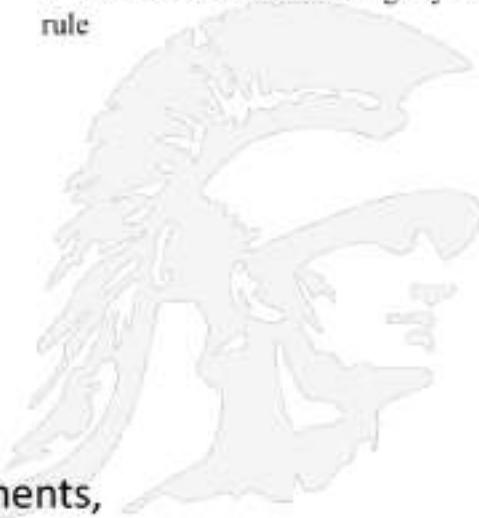
••



## 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

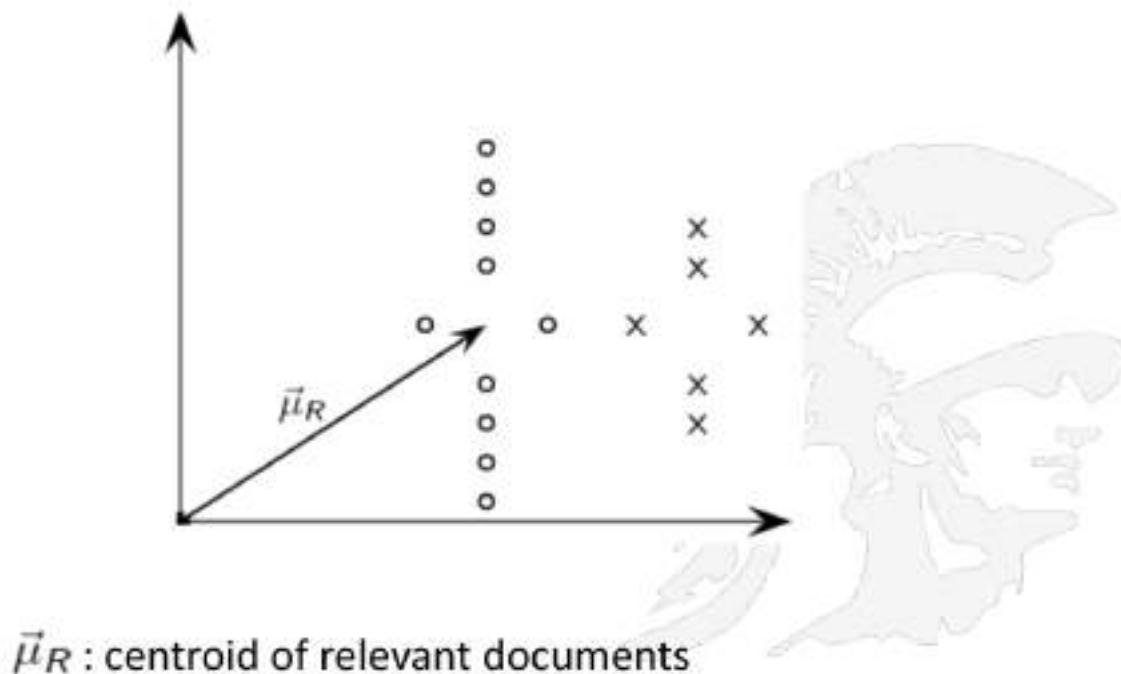


Let circles represent relevant documents,  
Let Xs represent nonrelevant documents

• •



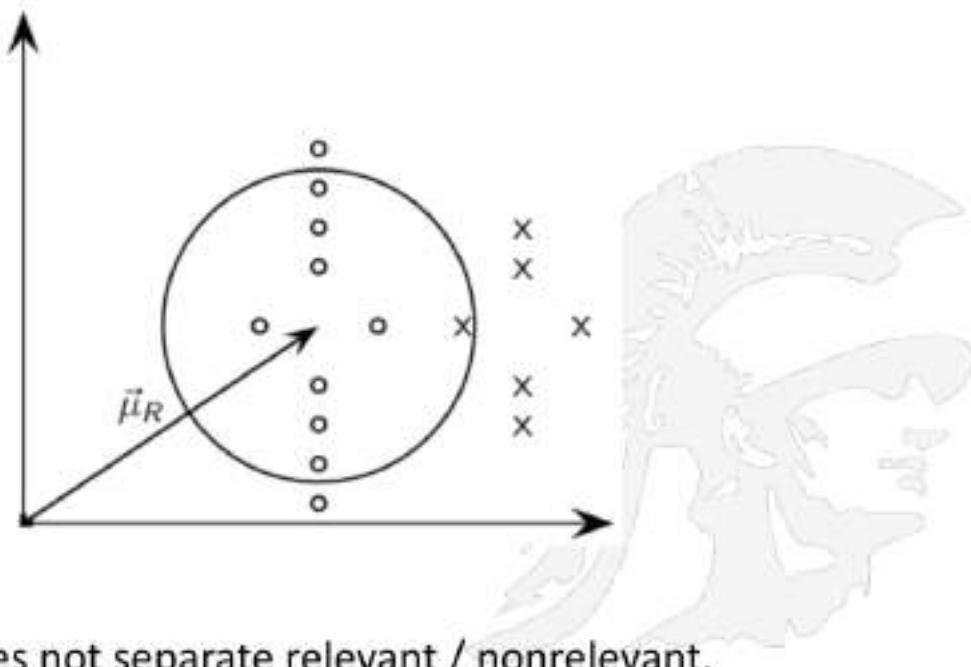
## 2D Rocchio Illustrated (1 of 9)



$\vec{\mu}_R$  : centroid of relevant documents

20

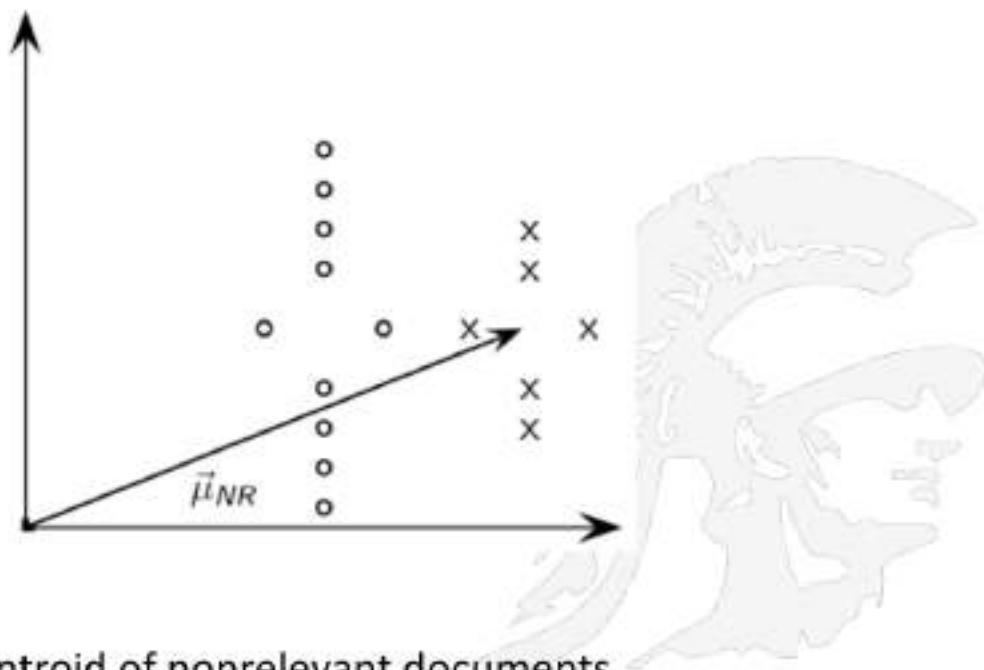
• •

**2D Rocchio Illustrated (2 of 9)**

••



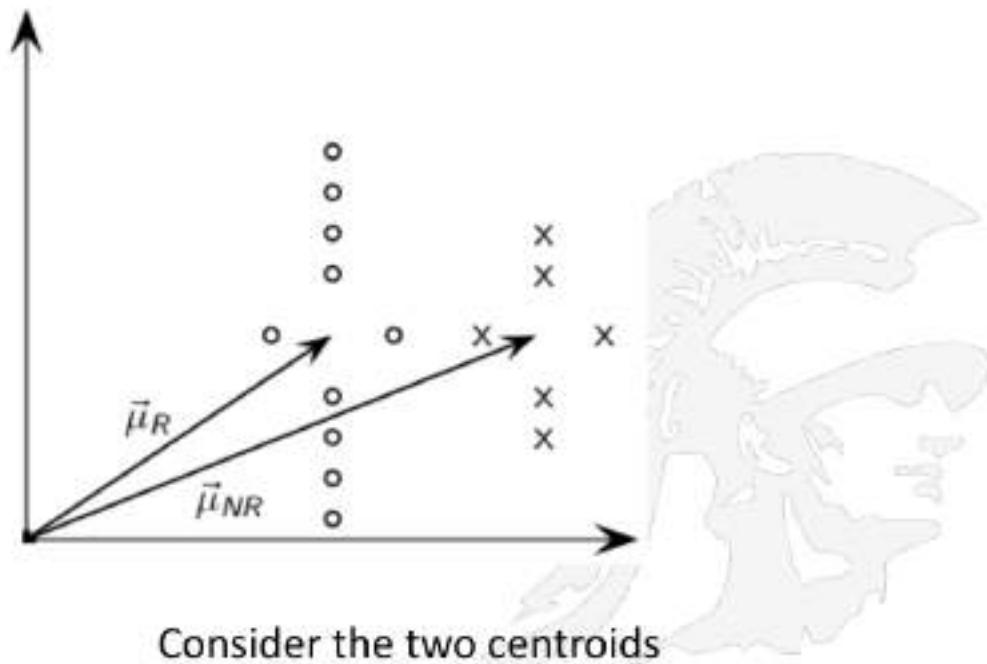
## 2D Rocchio Illustrated (3 of 9)



• •

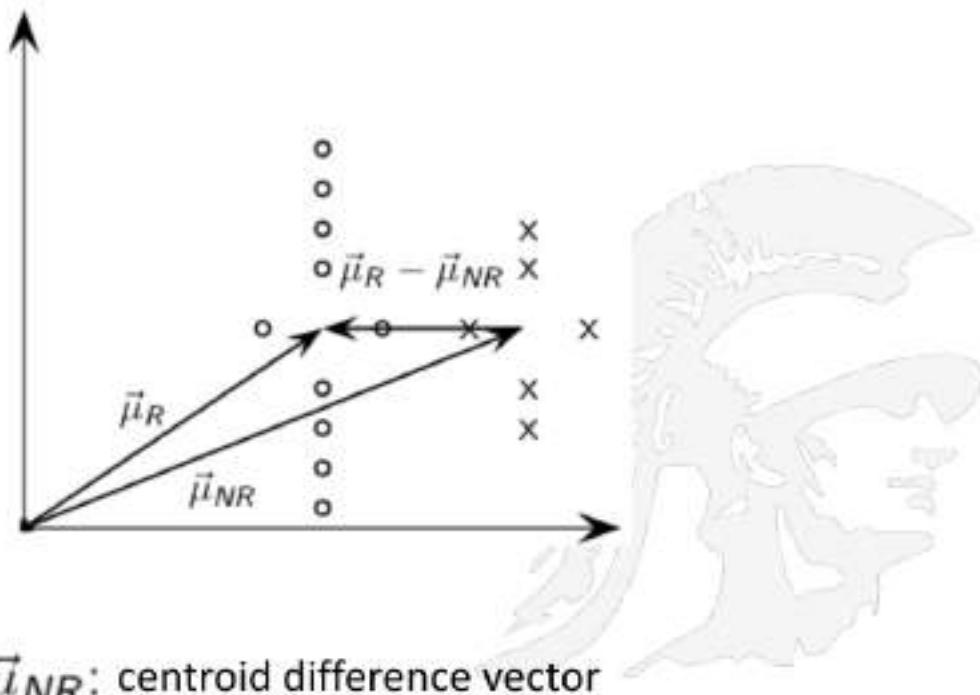


## 2D Rocchio Illustrated (4 of 9)



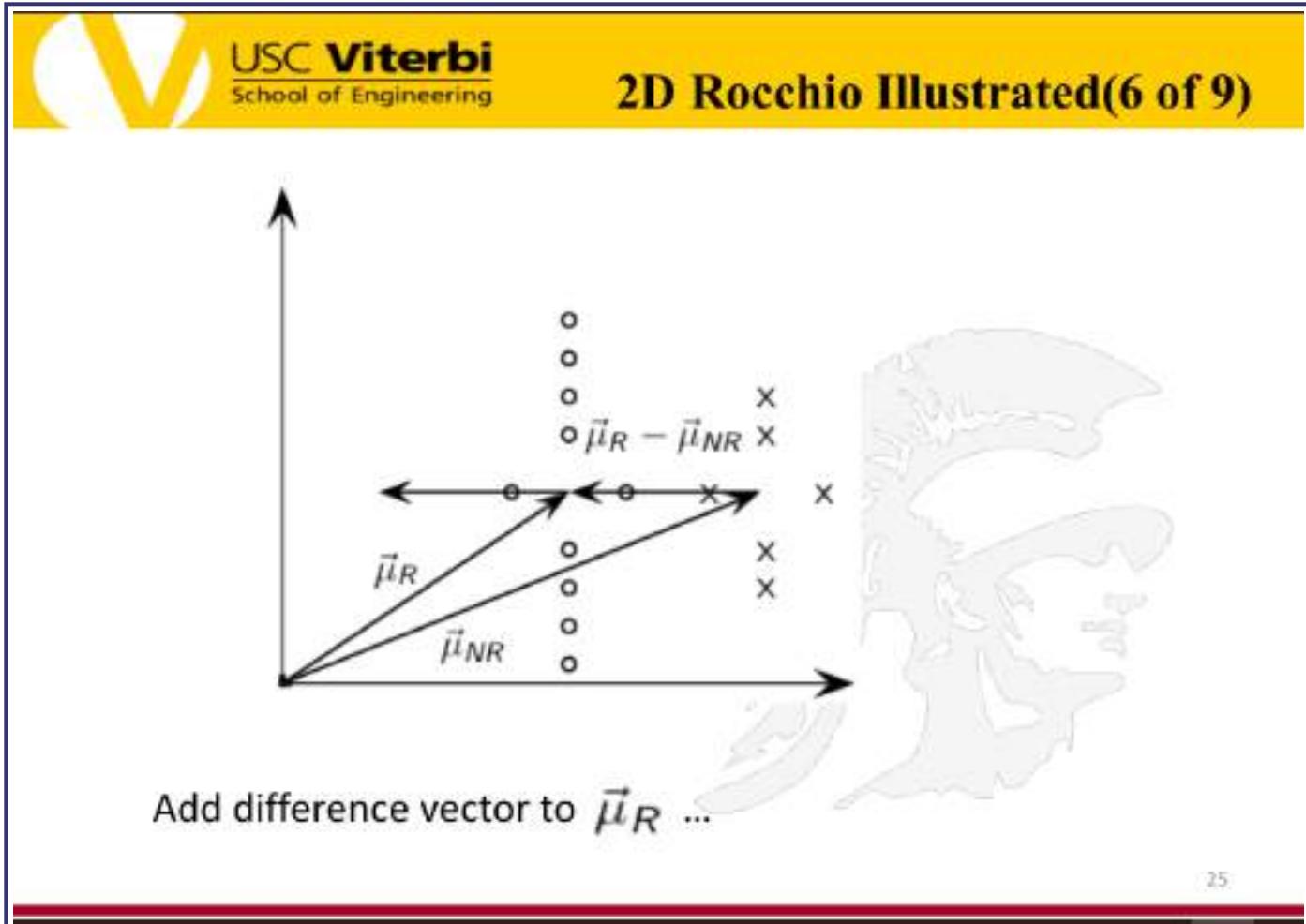
23

••

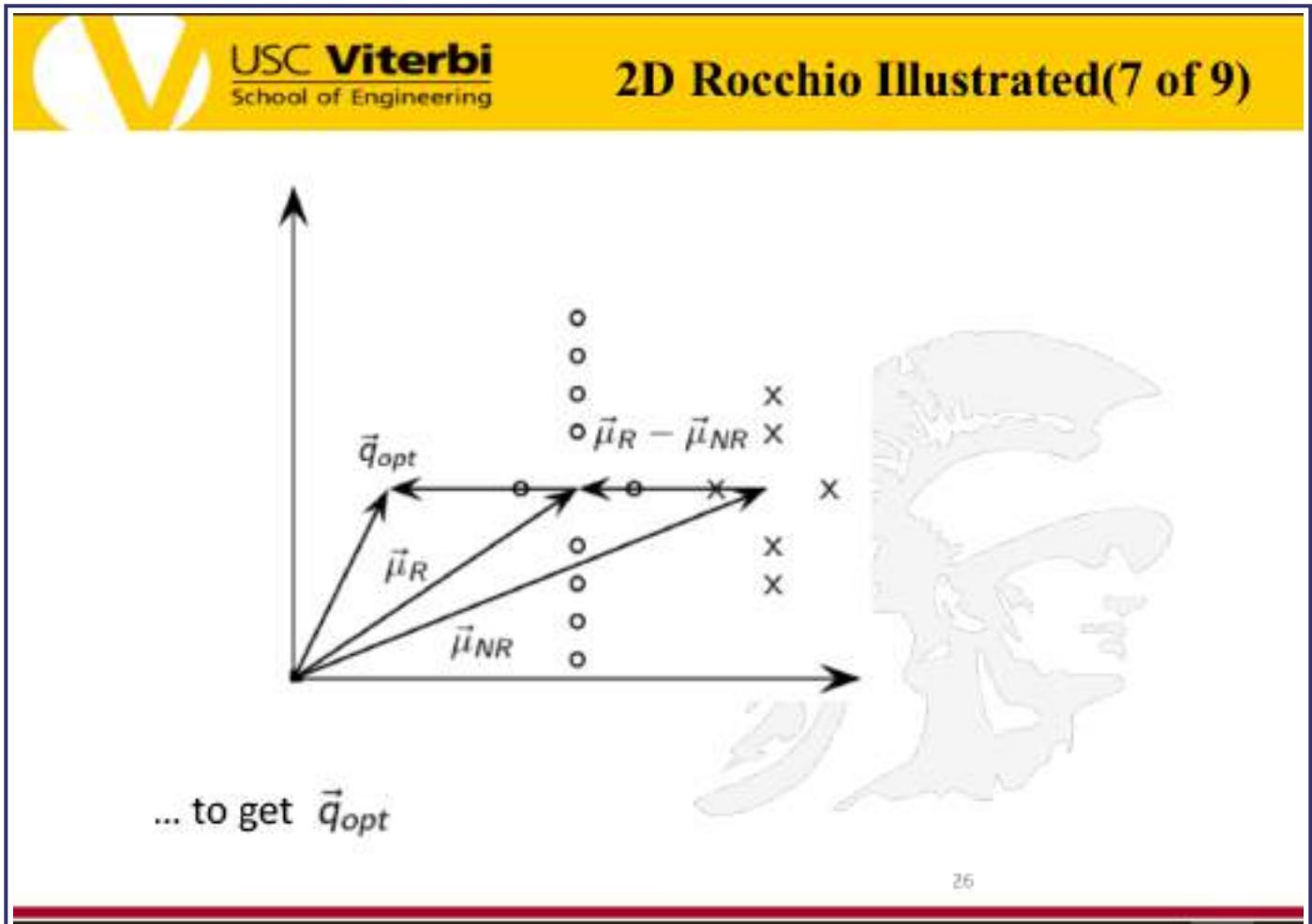
**2D Rocchio Illustrated(5 of 9)**

24

• •



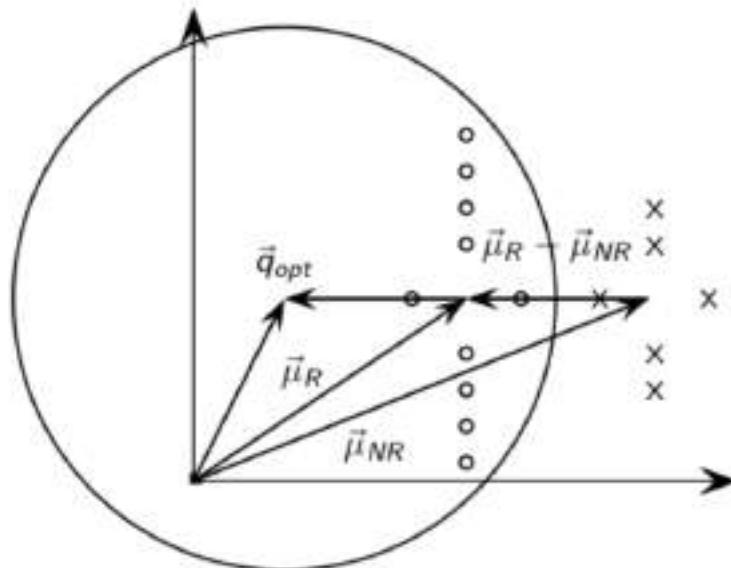
• •



••



## 2D Rocchio Illustrated(8 of 9)



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

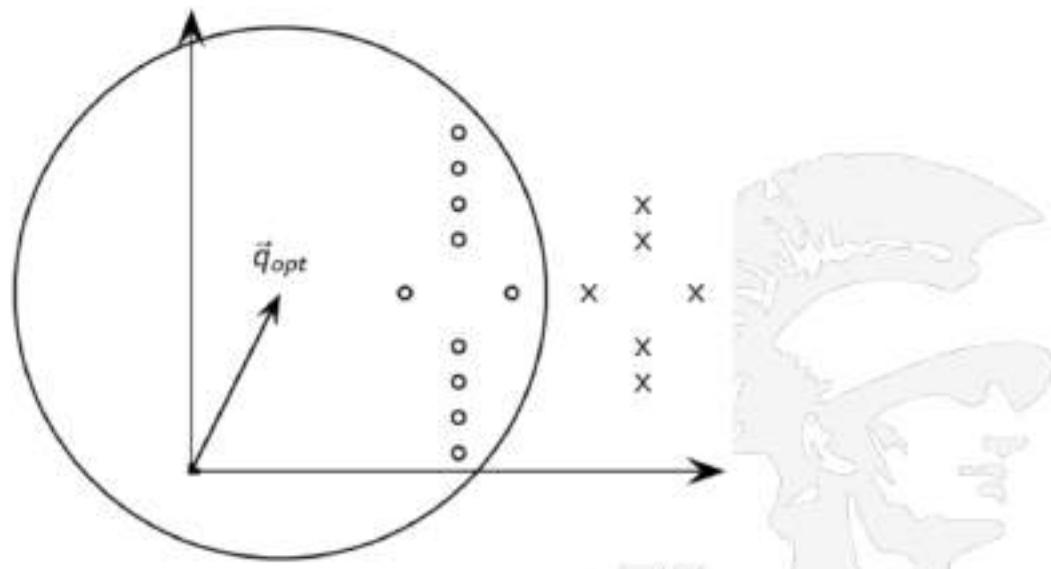
$\vec{q}_{opt}$  now separates relevant / nonrelevant perfectly.

27

••



## 2D Rocchio Illustrated(9 of 9)



$\vec{q}_{opt}$  separates relevant / nonrelevant perfectly.

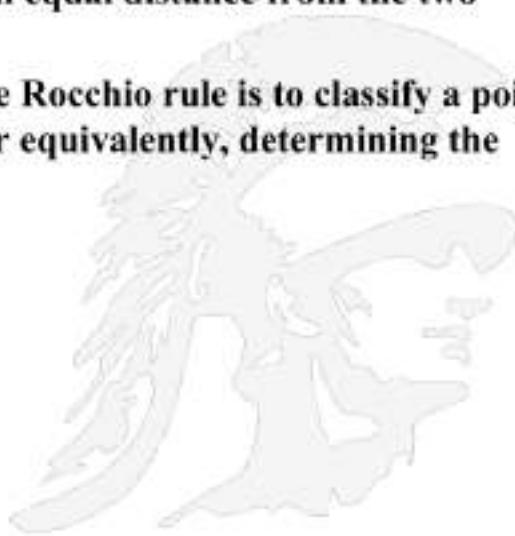
28

••



## 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



Copyright Ellis Horowitz, 2011-2022

29

••



## 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

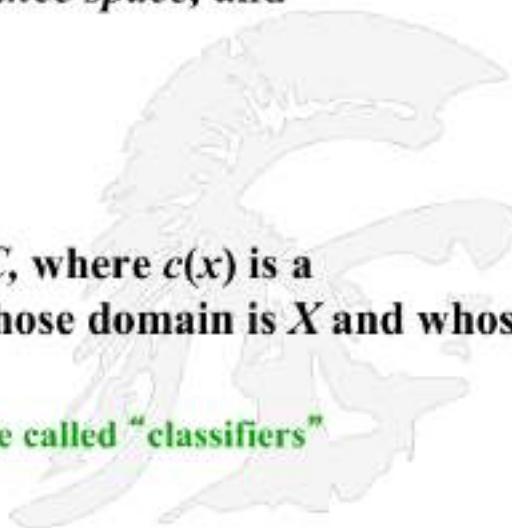
- **Manual classification**
  - Used by the original Yahoo! Directory
  - **Other search engines did similar things, e.g.**
    - Looksmart, <http://www.looksmart.com/>
    - Open Directory Project, <https://www.dmoz.org/>
    - PubMed, <http://www.ncbi.nlm.nih.gov/pubmed>
  - **Accurate when job is 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

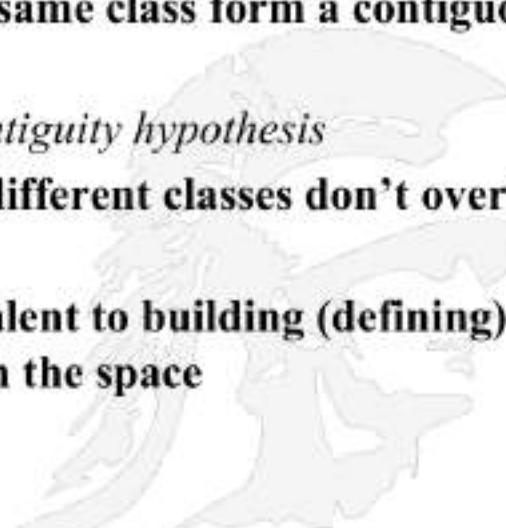
- Given two things:
  1. A description of an instance,  $x \in X$ , where  $X$  is the *instance language* or *instance space*, and
  2. A fixed set of categories:  
$$C = \{c_1, c_2, \dots, c_n\}$$
- Determine:
  - The category of  $x$ :  $c(x) \in C$ , where  $c(x)$  is a *categorization function* whose domain is  $X$  and whose range is  $C$ .
    - Functions that categorize are called “classifiers”



••



- 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

$$\left( \sum_{i=1}^k (|x_i - y_i|)^q \right)^{1/q}$$



Copyright Ellis Horowitz, 2011-2022

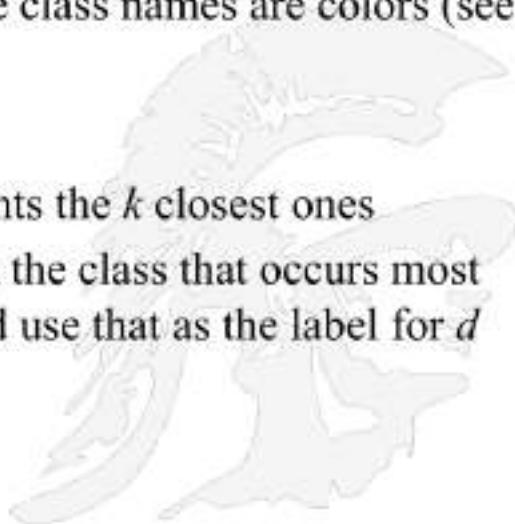
40

••



## **$k$ 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  $d$

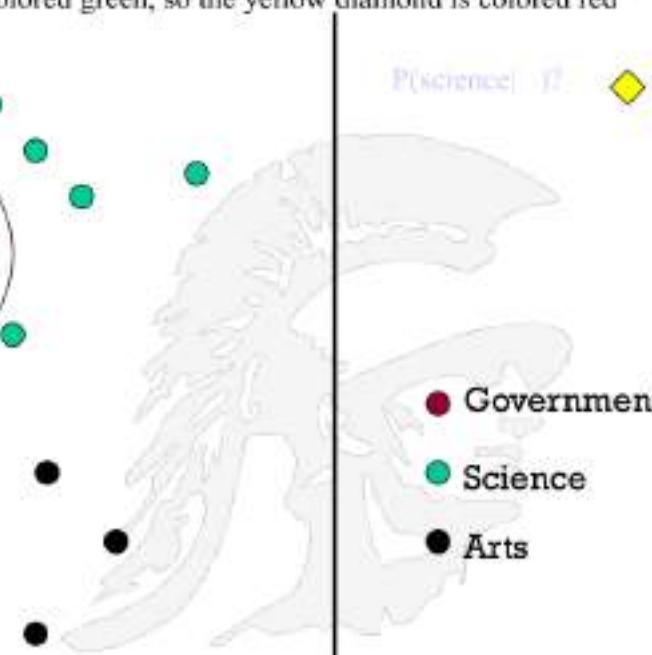
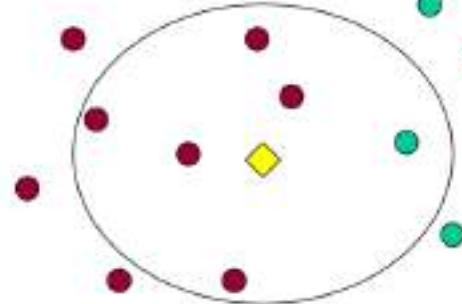


••



## Example: $k=6$ (6NN)

5 neighbors are colored red, one is colored green, so the yellow diamond is colored red



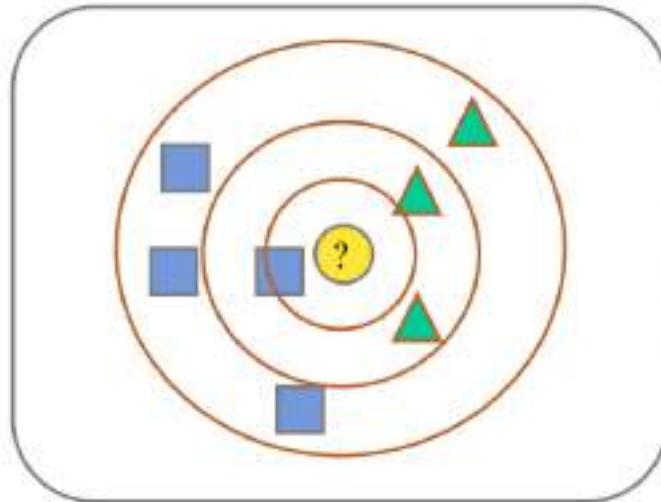
When  $k=1$ , the document is assigned  
to its nearest neighbor

Copyright Ellis Horowitz 2011-2022

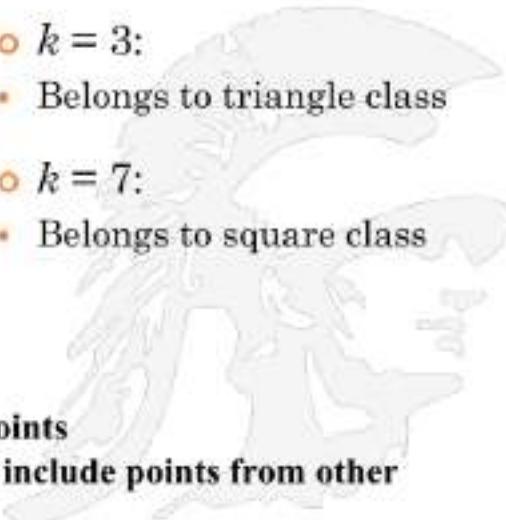
••



## K-Nearest Neighbor Another Example



- $k = 1$ :
  - Belongs to square class
- $k = 3$ :
  - Belongs to triangle class
- $k = 7$ :
  - Belongs to square class



- **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

Copyright Ellis Horowitz, 2011-2022

43

••



- **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

Copyright Ellis Horowitz, 2011-2022

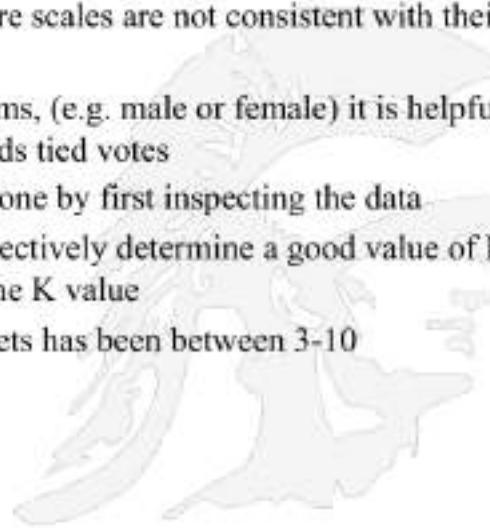
44

••



## 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



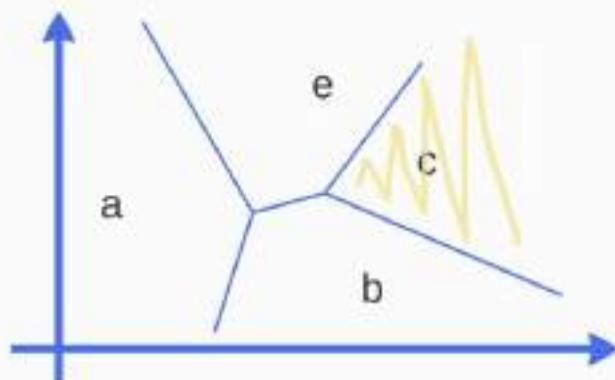
••



## 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



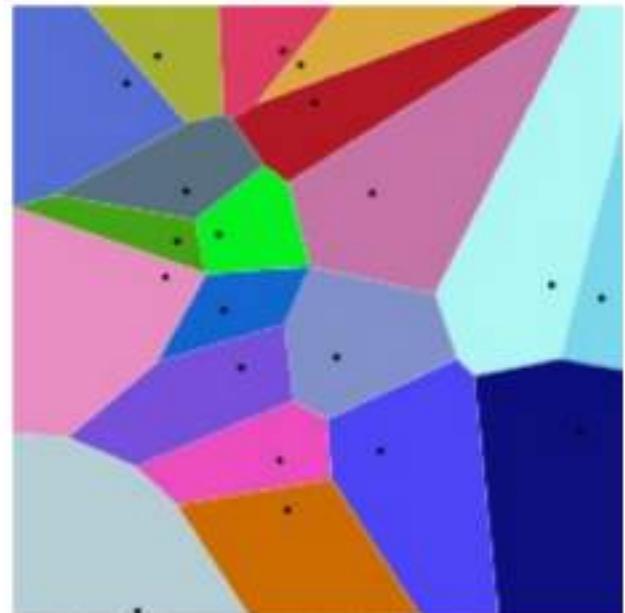
$$R_i = \{x : d(x, x_i) < d(x, x_j), i \neq j\}$$

••



## 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**

Copyright Ellis Horowitz, 2017-2022

47

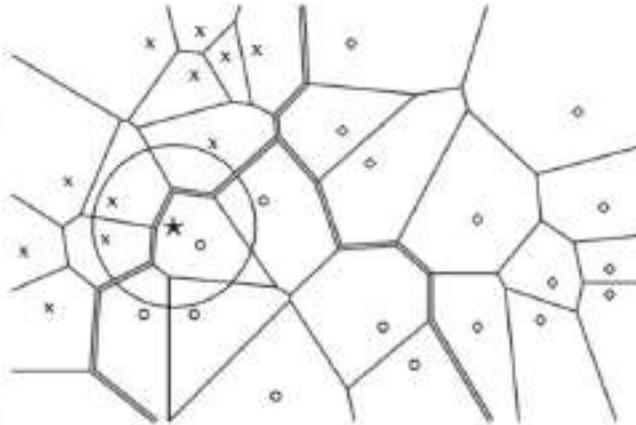
••

 **$K=1$  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)**



Copyright Ellis Horowitz, 2011-2022

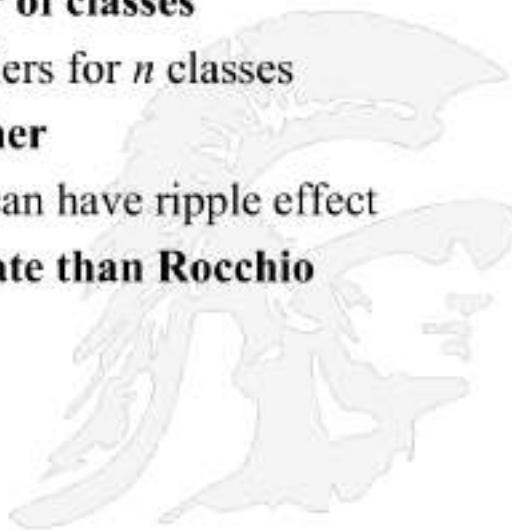
48

..



## 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**



Copyright Ellis Horowitz, 2011-2022

51

← 1/21 → \*\*\* 09:28:48

# Recommendation systems

**"what else to know about?"**

..



USC **Viterbi**  
School of Engineering

# Recommendation Systems

## Collaborative Filtering & Content-Based Recommendations



Copyright Ellis Horowitz 2011-2022

•



USC **Viterbi**  
School of Engineering

## 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)

Copyright Ellis Horowitz, 2011-2022

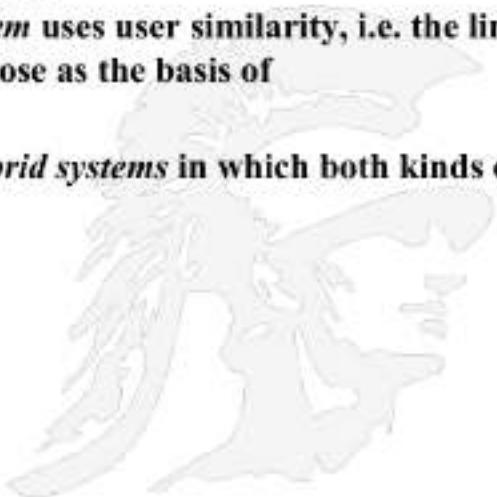
6

..



## 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



Copyright Ellis Horowitz, 2011-2022

7

••



## 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**.

• •



USC **Viterbi**  
School of Engineering

## 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

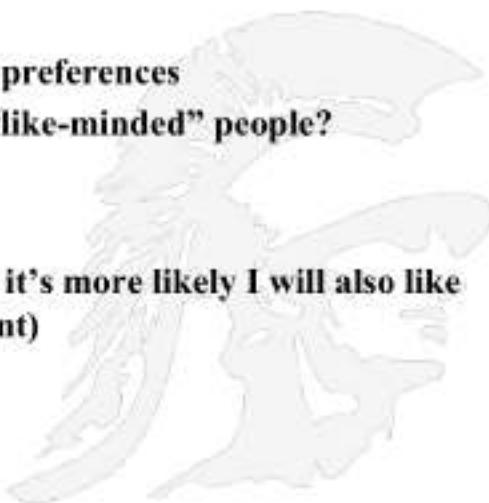
Copyright Ellis Horowitz 2011-2022

..



## 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)



Copyright Ellis Horowitz, 2011-2022

11

••



## 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

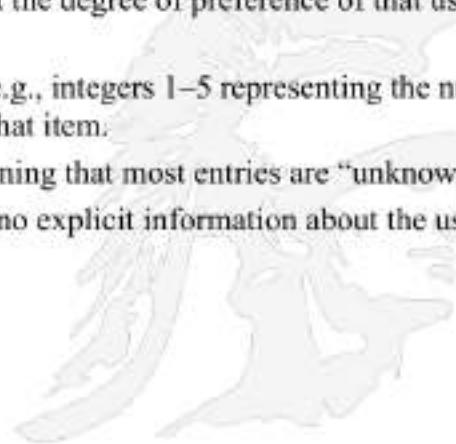


Copyright Ellis Horowitz 2011-2022



# 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.



Copyright Ellis Horowitz, 2011-2022

13

•

 **USC Viterbi**  
School of Engineering

## 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.

Copyright Ellis Horowitz 2011-2022

••



## 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  $w_{ij}$ .  $Content(d_j) = (w_{1j}, \dots, w_{kj})$ .

$$w_{ij} = TF_{i,j} \times IDF_i \quad TF_{i,j} = \frac{f_{i,j}}{\max_z f_{z,j}}, \quad IDF_i = \log \frac{N}{n_i}.$$

- **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
- **$f_{ij}$**  : 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 )
Movie2	(0, 1, 0, 0, 0 )
Movie3	(1, 0, 0, 0, 0 )
Movie4	(0, 1, 0, 0, 0 )
Movie5	(0, 1, 0, 0, 0 )

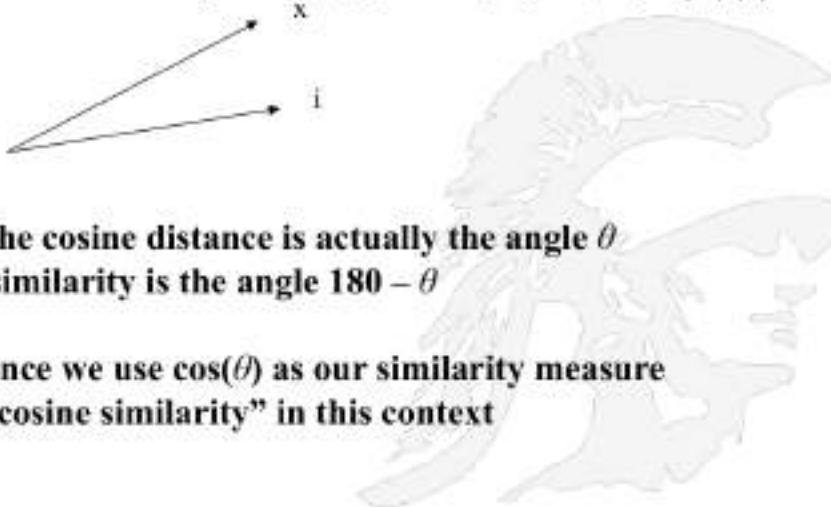
$$\text{ActA's weight} = \text{Sum(ActA)}/5$$

..



## Making Predictions

- Given user profile  $x$  (movies he/she watched) and item profile  $i$  (*movies with actor profiles*)
- Estimate the similarity of  $U(x,i) = \cos(\theta) = (x,i)/(|x| |i|)$



- 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

Copyright Ellis Horowitz, 2011-2022

19

..



## 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

Copyright Ellis Horowitz, 2011-2022

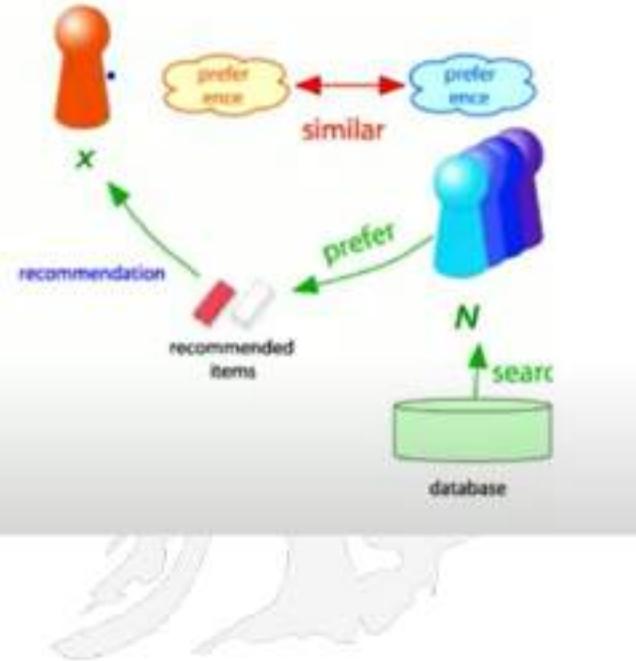
21

..

 USC **Viterbi**  
School of Engineering

## 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



The diagram illustrates the collaborative filtering process. It features two stylized human figures, one red and one blue, representing users. Above them are two thought bubbles: one yellow containing 'prefer once' and one blue containing 'prefer once'. A double-headed arrow between the bubbles is labeled 'similar'. Below the users are two small rectangular items, one red and one white, labeled 'recommended items'. A green arrow points from the red item to the red user, labeled 'recommendation'. Another green arrow points from the red item to the blue user, labeled 'prefer'. To the right of the users is a green cylinder labeled 'database'. An upward-pointing arrow from the database is labeled 'N' and 'search'.

Copyright Ellis Horowitz, 2011-2022

22

••

 USC **Viterbi**  
School of Engineering

### 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  $r_x$  and  $r_y$ 
  - The rating vector of user B is  $(5, 5, 4, 0, 0, 0, 0)$
- We need a similarity metric  $\text{sim}(x,y)$  between rating vectors
- The metric should capture the intuition that  $\text{sim}(A,B) > \text{sim}(A,C)$ 
  - $A$  and  $B$  both liked HP1, but  $A$  and  $C$  had very different opinions about TW and SW1
- Recall  $\text{sim}(A,B) = |r_a \text{ intersect } r_b| / |r_a \text{ union } r_b|$  try Jaccard Similarity
- $\text{Sim}(A,B) = 1/5$ ;  $\text{sim}(A,C) = 2/4$ . since A&B rated only one movie in common
  - But using Jaccard Similarity we get a result we don't want, namely
  - $\text{Sim}(A,B) < \text{Sim}(A,C)$
- Problem: ignores rating values

Copyright Ellis Horowitz, 2011-2022

24

..



## 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

Copyright Ellis Horowitz, 2011-2022

25

••

**USC Viterbi**  
School of Engineering

## 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
- $\text{sim}(A,B) = \cos(r_A r_B) = 0.09$ ;  
 $\text{sim}(A,C) = -0.56$

Result: A and C are very DISsimilar

- $\text{sim}(A,B) > \text{sim}(A,C)$
- 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 **Rearson Correlation**

26

..



## 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  $i$

$$r_{ij} = \frac{\sum_{j \in N(i;x)} s_{ij} \cdot r_{xj}}{\sum_{j \in N(i;x)} s_{ij}}$$

$s_{ij}$ ... 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  $i$

Copyright Ellis Horowitz, 2011-2022

28

••

**Let's Do An Example**  
**Item – Item CF ( $|N|=2$ )**

USC Viterbi  
School of Engineering

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

Copyright Ellis Horowitz, 2011-2022

29

Movie 3 and 6 are the two recommendations that are most similar to movie 1

Compute similarity weights

$S_{13} = 0.41$   $S_{16} = 0.59$

Use the weighted average to compute

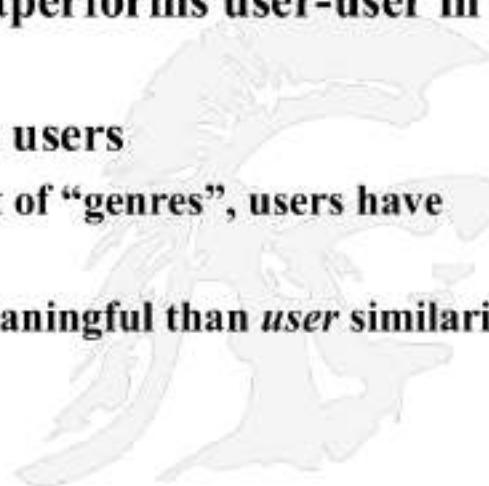
$R_{15} = (0.41 * 2 + 0.59 * 3) / (0.41 + 0.59) = 2.6$

..



## 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



1/10  \*\*\*[←](#)[→](#)

# Assorted topics

## Part 1

# IR – keeps evolving...

As with everything else, the field of information retrieval keeps changing – there are newer forms of data, newer algorithms, cross-pollination from other fields...

In this lecture and the next, we take a BRIEF look at a variety of developments. 'BRIEF' because each topic (eg. vector DBs) is a detailed area, we simply can't cover all of them in any detail. In other words, you can look into what interests you, later (after the course ends).

The following topics are grouped where possible, but are otherwise in no particular order.

# Image understanding and search

ImageNet is a massive DB, of human-annotated images.

It can serve as a vehicle to drive ML algorithms (eg. CNN, VT):

<https://www.youtube.com/watch?v=4OriCqvRoMs>

You can learn more here.

# Code search

GitHub does code search this way: <https://github.blog/2023-02-06-the-technology-behind-githubs-new-code-search/>

# LBS/PS

Location Based Search (LBS), ie. Proximity Search (PS), is about using location data where the query originates, and using that to return responses.

Here is an intro: <https://www.businessnewsdaily.com/5386-location-based-services.html>

<https://www.ncbi.nlm.nih.gov/pmc/articles/PMC8399432/> presents an efficient way to do LBS.

LBS can be exploited, to find (track) a user precisely:

<https://www.cs.uic.edu/~polakis/tr/proximity-tr.pdf>

# Similarity search using vector DBs

Pinecone is a NEW type of DB called 'vector DB', which embeds into vector (XYZ...) orthogonal space, the items (data) we want to search for (using 'cosine similarity'). Here is a primer.

We can also use vector DBs as 'infinite memory', for use with generative QA sessions.

# LDA, for topic modeling

Given a document, how to predict its dominant topic(s)? Here is one approach.

## ANN (Faiss)

Approximate Nearest Neighbor calcs are much faster than exact ones. Here is more.

# Task-specific fine-tuning

'Fine-tuning' is the concept of adapting a general-purpose model, for a specific purpose. The following is an account of GPT-2 was fine-tuned to generate scientific paper abstracts:

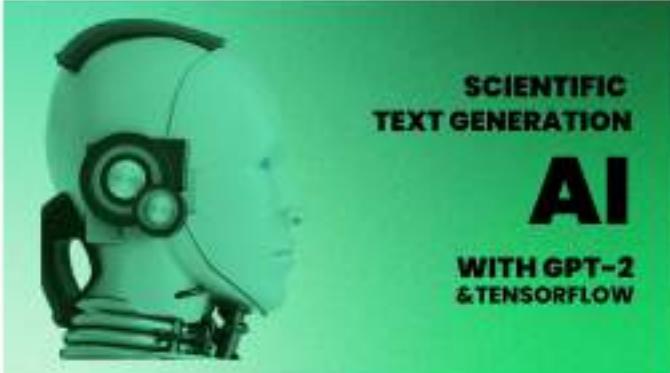
Search Medium

Published in Towards AI

Edoardo Bianchi · Nov 9, 2022 · 10 minutes · Member-only · Listen

## I Fine-Tuned GPT-2 on 110K Scientific Papers. Here's The Result

Content writing by AI is common, but is it possible for an AI to write technical essays?



Scientific text generation with AI. Image by the author.

Artificial agents are widely used nowadays and are able to achieve superhuman performance in multiple tasks. Text generation is one of the emerging applications of AI and is used in several scenarios. Freeform text generation, Q&A, and abstractive summarization are only some of them.

To investigate whether an AI could write technical essays, I trained a casual language model on about 100K machine learning papers.

What is the quality of the result? What are the limitations of the proposed approach? Is it possible to get GPT-2 to write a full paper? These are the questions that I will try to answer.

**Introduction**

The Generative Pre-Trained Transformer (GPT) 2 is an artificial intelligence developed by OpenAI in 2019 and allows for several purposes: text summarization, translation, question-answering, and text generation. GPT-2 is pre-trained on a large English data corpus, furthermore, can be fine-tuned for a specific task.

In this article, I will use the Huggingface Distilled-GPT2 (DistilGPT2) model. DistilGPT2 has 82 million parameters and was developed by knowledge distillation, moreover is lighter and faster than GPT-2.

**1. Importing Tools**

I started by importing all the required tools and libraries.

```
1 from transformers import AutoModelForSeq2SeqLM, AutoTokenizer, AdamW, get_linear_schedule_with_warmup, create_optimizer
2 from transformers import DistilBertForTextGeneration
3 import tensorflow as tf
```

Edoardo Bianchi  
663 Followers  
Naturally Curious Data Scientist | I write about AI, Data Science, and Machine Learning | To Become a member: <https://medium.com/@bianchi/membership>

Follow

More from Medium:

- Adarsh · Artificial Intelligence · Creating Your Own ChatGPT: A Guide to Fine-Tuning LLMs with LoRA
- Noah Klimberg-Sakai · How to fine-tune a GPT-3 model using Python with your own data for improved...
- Noah Klimberg-Sakai · How to fine-tune a GPT-3 model using Python with your own data for improved...
- Somnath ... in JavaScript in Plain English · Coding Won't Exist in 5 Years. This Is Why
- The Latest News in data-driven E... · Secret prompt that ChatGPT loves, with Proofs
- Noah Klimberg-Sakai · How to fine-tune a GPT-3 model using Python with your own data for improved...
- Somnath ... in JavaScript in Plain English · Coding Won't Exist in 5 Years. This Is Why
- The Latest News in data-driven E... · Secret prompt that ChatGPT loves, with Proofs
- Noah Klimberg-Sakai · How to fine-tune a GPT-3 model using Python with your own data for improved...
- Somnath ... in JavaScript in Plain English · Coding Won't Exist in 5 Years. This Is Why
- The Latest News in data-driven E... · Secret prompt that ChatGPT loves, with Proofs

# Task-specific embedding

Multiple NLP tasks can be handled by a system, if it embeds a document in a way that's related to the task: <https://pub.towardsai.net/paper-review-instructor-one-embedder-any-task-6a846b0d3ba>

## 2. Importing the baseline model and tokenizer

Then, I used `TFAutoModelForCausalLM` and `AutoTokenizer` to automatically load the correct model based on a specific checkpoint. A checkpoint contains the weights of a pre-trained model.

In this case, I imported the DistilGPT-2 checkpoint. I also set the end-of-sequence token as a padding token.

```
1 tokenizer = AutoTokenizer.from_pretrained("distilgpt2")
2 tokenizer.pad_token = tokenizer.eos_token
3 model = TFAutoModelForCausalLM.from_pretrained("distilgpt2", pad_token_id=tokenizer.eos_token_id)
```

## 3. Importing Data

The dataset for the fine-tuning operation is available on the Huggingface Hub, and it's a subset of a bigger dataset hosted on Kaggle.

The original dataset, published by Cornell University, contains titles and abstracts of 1.7M+ scientific papers belonging to the STEM category. The subset hosted on the Huggingface Hub contains information on around 100K papers pertaining to the machine learning category.

I decided to fine-tune DistilGPT-2 on abstracts only. I started by loading the dataset from the Huggingface Hub.

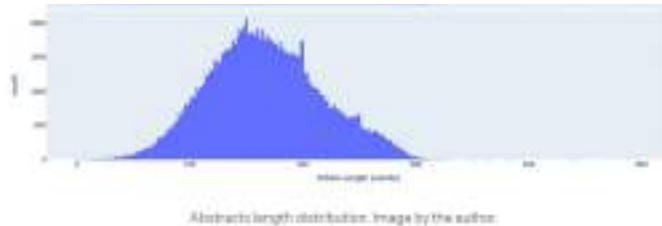
```
1 data = load_dataset("clmlab/mlm-sciv-papers", split="train")
2 data
```

The dataset consists of 117392 rows and has 4 columns (two of them are useless).

```
1 datasets()
2 datasets["train": {"inputs": "#", "outputs": "1", "title": "Title", "abstract": "#", "category": "Category"}, "train": 117392]
```

After this step, I decided to visualize the length distribution of the abstracts (in terms of words) with a histogram.

```
1 abstracts = [len(x.split()) for x in data["abstract"]]
2 plt.hist(abstracts, bins=400, marginal="log", labels={"value": "Abstract length (words)"})
```



Most of the abstracts are between about 100 and 250 words in length, and only a few are over 300 words. In particular: mode=150, mean=167, and median=164.

In addition to giving information about the dataset, the histogram allowed me to determine the maximum length of the inputs to be fed to the model.

I decided to set the maximum input length to 300 tokens: abstracts longer than this will be truncated. This is because all inputs must be padded to the same length, and long sequences of text greatly increase the training time.

#### 4. Split into Train and Validation Set

Next, I split the dataset into train and validation sets with `train_test_split()`. It is also possible to specify the partition sizes with the `test_size` parameter.

`train_test_split()` returns a dictionary of datasets, formerly a `datasetDict`. While it is possible to work with a `datasetDict`, I prefer to use two separate datasets: `train` and `val`.

```
1 data = data.train_test_split(shuffle=True, seed=420, test_size=0.2)
2
3 train = data['train']
4 val = data['test']
```

#### 5. Tokenize Data with HF Tokenizer

To tokenize the data I defined a generic tokenization function, and then I applied this function to all the samples by using `map()`. Inside the tokenization function, I used the tokenizer imported in the beginning.

The tokenizer has some important parameters to set:

1. `column` to tokenize. In this case "abstract".
2. `padding`. In this case = "max\_length" to pad a sequence to the maximum length specified by the `max_length` parameter.
3. `truncation`. If true, truncates sequences longer than the maximum length, specified by the `max_length` parameter.
4. `max_length`. Specifies the maximum length of a sequence.

Please note that by default the `map()` method sends batches of 1000 samples.

```
1 # The tokenization function
2 def tokenize(data):
3 tokenized = data['abstract'].map(lambda x: tokenizer(x, padding='max_length', truncation=True, max_length=300))
4 return tokenized
5
6 # Apply the tokenization to both train and drop all the columns except the tokenization result.
7 train_token = train.map(tokenization, batched=True, remove_columns=['title', 'abstract', 'news'])
8 val_token = val.map(tokenization, batched=True, remove_columns=['title', 'abstract', 'news'])
```

Norah Klinberg Salai  
How to fine-tune a GPT-3 model using Python with your own data for improved...

Somnath ... In JavaScript in Plat...  
Coding Won't Exist in 5 Years. This is Why

The Latest No... in data-driven E...  
Secret prompt that ChatGPT loves, with Proofs

Help Status Writers Shop Contact Privacy Terms About  
Feedback

Norah Klinberg Salai  
How to fine-tune a GPT-3 model using Python with your own data for improved...

Somnath ... In JavaScript in Plat...  
Coding Won't Exist in 5 Years. This is Why

The Latest No... in data-driven E...  
Secret prompt that ChatGPT loves, with Proofs

Help Status Writers Shop Contact Privacy Terms About  
Feedback

Norah Klinberg Salai  
How to fine-tune a GPT-3 model using Python with your own data for improved...

Somnath ... In JavaScript in Plat...  
Coding Won't Exist in 5 Years. This is Why

The Latest No... in data-driven E...  
Secret prompt that ChatGPT loves, with Proofs

Help Status Writers Shop Contact Privacy Terms About  
Feedback

Norah Klinberg Salai  
How to fine-tune a GPT-3 model using Python with your own data for improved...

Somnath ... In JavaScript in Plat...  
Coding Won't Exist in 5 Years. This is Why

The Latest No... in data-driven E...  
Secret prompt that ChatGPT loves, with Proofs

Help Status Writers Shop Contact Privacy Terms About  
Feedback

Norah Klinberg Salai  
How to fine-tune a GPT-3 model using Python with your own data for improved...

Somnath ... In JavaScript in Plat...  
Coding Won't Exist in 5 Years. This is Why

The Latest No... in data-driven E...  
Secret prompt that ChatGPT loves, with Proofs

Help Status Writers Shop Contact Privacy Terms About  
Feedback

Norah Klinberg Salai  
How to fine-tune a GPT-3 model using Python with your own data for improved...

## 6. Adding Labels to Train and Validation Sets

In Casual Language Modeling, the labels are the input tokens (`input_ids`) right-shifted. This operation is automatically done by the Huggingface transformer, thus I created a `labels` column in the datasets with a copy of the tokens (`input_ids`).

After this operation, the train and validation sets had three columns: `input_ids` and `attention_mask` from the tokenization process, and `Labels` from the `create_labels` process.

```

 1 # create 'label1' as a copy of 'input_label'
 2 def create_label(label):
 3 new_label = Label("label1") - copy("input_label").copy()
 4 return new_label
 5
 6 # add the 'label1' column using 'new_label'
 7 # Create a trade item representing 'label1', batched=True, but process=False
 8 new_label = col_label.create_label("label1", batched=True, but_process=False)

```

## 7. Converting Train and Validation Sets to TF Datasets

Next, I converted the datasets to `tf.data.Dataset`, that Keras can understand natively; for this purpose I used `model.fit_generator(tf_dataset)`.

With respect to the `dataset.to_tf_dataset()` method, `Model.from_tf_dataset()` can automatically determine which column names to use as input and provides a default data collate.

Note that I only shuffled the train data. After some experiments, I found that the optimal batch size = 16.

```

1 train_set = model.prepare_tf_dataset(
2 batch_size=10,
3 shuffle=True)
4
5 validation_set = model.prepare_tf_dataset(
6 batch_size=10,
7 shuffle=False)
8
9
```

## 8. Compiling, Fitting, and Evaluating the Model

Before fitting the model, I set up a learning rate scheduler and an optimizer. I used the `ExponentialDecay` scheduler from Keras and the `AdamWeightDecay` optimizer from Huggingface.

Learning rate decay is a technique to reduce the learning rate over time. With exponential decay, the learning rate is reduced exponentially.

```
 3 # Setting up the learning rate schedules
 4 lr_schedules = tf.keras.optimizers.schedules.ExponentialDecay(
 5 initial_learning_rate=0.001,
 6 decay_steps=4000,
 7 decay_rate=0.95)
```



```

1 # Initialize PyTorch
2
3 # + Uncomment this learning rate
4 # optimizer = AdamW(model.parameters(), learning_rate=lr_schedule, weight_decay_ratio=0.01)
5
6 # FineTuningModel.fit(model, train_dataloader, validation_dataloader, epochs=1, lr_scheduler=lr_schedule)

```

Next, I compiled the model. Transformers models generally compute loss internally, and there is no need to specify a loss parameter. For language modeling, the selected loss is cross-entropy.

```

1 model.compile(optimizer=optimizer)
2 model.compile()
3
4 # FineTuningModel.compile(model=model)
5
6 total_params = sum(p.numel() for p in model.parameters())
7 trainable_params = sum(p.numel() for p in model.parameters() if p.requires_grad)
8
9 print(f"Total parameters: {total_params}, Trainable parameters: {trainable_params}")
10
11 # FineTuningModel.compile()

```

At this point, I set up a callback to the Huggingface Hub to save the fine-tuned model.

```

1 # This cell is optional
2 from transformers import AutoModelWithHead
3
4 model_name = "OPT-2"
5 push_to_hub_model_id = f"{model_name}-finetuned-sapex"
6
7 push_to_hub_CALLBACK = PushToHubCallback(
8 output_dir=f"hf_{push_to_hub_model_id}",
9 token=push_to_hub_token,
10 hub_model_id=push_to_hub_model_id,
11 hub_token="your HF token"
12)
13
14 # FineTuningModel.push_to_hub_callback(push_to_hub_CALLBACK)

```

I also set up a callback to Tensorboard.

```

1 # This cell is optional
2 from transformers import AutoModelWithHead
3
4 tensorboard_CALLBACK = TensorBoard(log_dir='./tensorboard',
5 update_freq='1000',
6 histogram_freq=1,
7 profile_batch="2,10")
8
9 # FineTuningModel.push_to_hub_CALLBACK(tensorboard_CALLBACK)
10
11 callbacks = [push_to_hub_CALLBACK, Tensorboard_CALLBACK]
12
13 # FineTuningModel.push_to_hub_callback(callbacks)

```

Finally, I fitted the model by calling the `fit()` method. I specified the train and validation sets and the number of epochs.

```

1 # Fit with callbacks
2 model.fit(train_set, validation_data=validation_set, epochs=1, callbacks=callbacks)
3
4 # FineTuningModel.evaluate(validation_dataloader)

```

After the training step, I evaluated the model and got its cross-entropy loss on the validation set.

The Latest News in data-driven AI  
Secret prompt that ChatGPT loves, with Proofs

Help Status About Blog Contact Privacy Terms About Help in speech

Norah Klinberg Sakai  
How to fine-tune a OPT-3 model using Python with your own data for improved...

Somnath... In JavaScript in Plat...  
Coding Won't Exist In 5 Years, This Is Why

The Latest News in data-driven AI  
Secret prompt that ChatGPT loves, with Proofs

Help Status About Blog Contact Privacy Terms About Help in speech

Norah Klinberg Sakai  
How to fine-tune a OPT-3 model using Python with your own data for improved...

Somnath... In JavaScript in Plat...  
Coding Won't Exist In 5 Years, This Is Why

The Latest News in data-driven AI  
Secret prompt that ChatGPT loves, with Proofs

Help Status About Blog Contact Privacy Terms About Help in speech

Norah Klinberg Sakai  
How to fine-tune a OPT-3 model using Python with your own data for improved...

Somnath... In JavaScript in Plat...  
Coding Won't Exist In 5 Years, This Is Why

The Latest News in data-driven AI  
Secret prompt that ChatGPT loves, with Proofs

Help Status About Blog Contact Privacy Terms About Help in speech

Norah Klinberg Sakai  
How to fine-tune a OPT-3 model using Python with your own data for improved...

Somnath... In JavaScript in Plat...  
Coding Won't Exist In 5 Years, This Is Why

The Latest News in data-driven AI  
Secret prompt that ChatGPT loves, with Proofs

Help Status About Blog Contact Privacy Terms About Help in speech

Norah Klinberg Sakai  
How to fine-tune a OPT-3 model using Python with your own data for improved...

Somnath... In JavaScript in Plat...  
Coding Won't Exist In 5 Years, This Is Why

The Latest News in data-driven AI  
Secret prompt that ChatGPT loves, with Proofs

```
1 eval_loss = model.evaluate(validation_set)
```

ChatGPT pipeline linked with ❤️ by GitHub

[View raw](#)

Loss=2.2371. Generally, the quality of a language model is measured in 'perplexity'. To convert cross-entropy to perplexity, I simply raised e to the power of the cross-entropy loss.

```
1 perplexity = math.exp(eval_loss) * 100
```

ChatGPT pipeline linked with ❤️ by GitHub

[View raw](#)

In this case, perplexity=9.37.

## 9. Generating Text Using a Pipeline

At this point, I leveraged the `pipeline` functionality provided by Huggingface to see the model in action.

I set up a text-generation pipeline and specified the fine-tuned model, the tokenizer, and the framework to use. `max_new_tokens` allows specifying the maximum number of tokens (words) to generate in addition to the initial prompt provided.

```
1 # Setting up the pipeline
2 text_pipeline = pipeline(
3 "text-generation",
4 model=model,
5 tokenizer=tokenizer,
6 framework="pt",
7 max_new_tokens=100
8)
```

ChatGPT pipeline linked with ❤️ by GitHub

[View raw](#)

Two lines of code are enough to generate text with a pipeline:

```
1 text_pipeline = TextPipeline()
2 text_pipeline(text_sentence)
```

ChatGPT pipeline linked with ❤️ by GitHub

[View raw](#)

The `pipeline` is not the only way to use a model: it is possible to manually tokenize the prompt, generate new tokens, and decode the tokens to natural language. Here's an example:

```
1 input_ids = tokenizer.encode(text_sentence, return_tensors="pt")
2 output = model.generate(input_ids, max_length=50)
3 tokenizer.decode(output[0], skip_special_tokens=True)
```

ChatGPT pipeline linked with ❤️ by GitHub

[View raw](#)

## 12. Results Analysis

After fine-tuning the model, I wanted to understand what the model has learned and how the generated text is influenced by the fact that paper abstracts were used for training.

First, I generated a sample text by using "the role of recommender systems" as a prompt. This is the output generated by the model:

"the role of recommender systems in the real-world is still largely to be demonstrated by the lack of data and the need for data. Hence,

Norah Klinberg Salal  
How to fine-tune a GPT-3  
model using Python with your  
own data for improved...



Somnath... In JavaScript in Plat...  
Coding Won't Exist in 5 Years.  
This Is Why



The Latest No... In data-driven E...  
Secret prompt that ChatGPT  
loves, with Proofs



Norah Klinberg Salal  
How to fine-tune a GPT-3  
model using Python with your  
own data for improved...



Somnath... In JavaScript in Plat...  
Coding Won't Exist in 5 Years.  
This Is Why



The Latest No... In data-driven E...  
Secret prompt that ChatGPT  
loves, with Proofs



Norah Klinberg Salal  
How to fine-tune a GPT-3  
model using Python with your  
own data for improved...



Somnath... In JavaScript in Plat...  
Coding Won't Exist in 5 Years.  
This Is Why



The Latest No... In data-driven E...  
Secret prompt that ChatGPT  
loves, with Proofs



Norah Klinberg Salal  
How to fine-tune a GPT-3  
model using Python with your  
own data for improved...



Somnath... In JavaScript in Plat...  
Coding Won't Exist in 5 Years.  
This Is Why



The Latest No... In data-driven E...  
Secret prompt that ChatGPT  
loves, with Proofs



Norah Klinberg Salal  
How to fine-tune a GPT-3  
model using Python with your  
own data for improved...



Somnath... In JavaScript in Plat...  
Coding Won't Exist in 5 Years.  
This Is Why



The Latest No... In data-driven E...  
Secret prompt that ChatGPT  
loves, with Proofs



Norah Klinberg Salal  
How to fine-tune a GPT-3  
model using Python with your  
own data for improved...



For many recommendation systems such as Amazon or Spotify, it is necessary to provide a user knowledge of the content that has been clicked during the recommendation and provide a user knowledge of the user preferences. The previous works attempt to exploit data related to items they have clicked during an appropriate time frame. But little attention has been paid to the problem of item classification where a suitable time frame is available for user prediction. In this paper, we propose a multi-task learning approach to address the problem of item classification. For each task, we apply the contextual cues introduced by the user, and then learn to predict the user's purchased items' interests. Since the contexts of user preferences, we consider the feature that the user's preference (the time-frame) is present at the time of recommendation. In particular, we propose an alternative method for attribute-aware learning that utilizes the contextual cues in the sequence and the user's preferences to learn a classifier that classifies the user according to the contextual cues. This is done by maximizing the mutual information between the user's rating and the content-aware prediction task. The experimental results show that our model achieves better accuracy than the existing state-of-the-art methods, achieving up to 33.6% more accuracy on real-world recommendation tasks compared to the state-of-the-art methods. Our source code is available at <https://github.com/2-medleyFashion/median.github.io>.

This result sounded somehow copied & pasted from one of the existing abstracts, but after a check with some anti-plagiarism solutions, I realized that it is 100% unique.

During learning, the model captured common features of the abstracts and learned how to replicate them while still generating fresh text. Interestingly, the model used scientific language and common expressions: *The previous works... In this paper... We propose... The experimental result...*

The model also learned that sometimes a repository is added to the abstract: in this example, the text generated contains an URL to a GitHub repository. The URL and the repository don't exist, thus have been generated by the model (and not copied).

As a second experiment, I generated a sample text by using "clustering" as a prompt. In this case, the prompt consisted of only one word, so the text generation is not driven by additional context. This is the output:

"clustering can be used to extract clusters from data points. However, in many real-world scenarios, data points often appear in non-Euclidean relaxations, which allows different clusters to be discovered simultaneously without the need for specific optimization. In this paper, we consider a class of applications where clustering methods are applied. A common example is that in image selection problems, we show to the optimizer that the expected improvement will be obtained by minimizing the average performance of the clustering method. Our main contribution is a characterization of optimization problems with respect to clustering methods. Specifically, we present a generalization of the optimal transport method for which our spectral bounds in terms of the maximum number of clusters are derived. We apply this algorithm to classification of human brain tumors, showing that it is computationally efficient. We show that based on our analysis of a particular type of tumor, our analysis leads to faster convergence to the optimal solution."

Also in this case the text is reported to be unique and there are common expressions: *In this paper... Our main contribution... We present...*

For the third experiment, I used "transformers for text generation". Here's the result:

"transformers for text generation are becoming increasingly popular in computer vision, and so are the prevalent datasets that contain high-dimensional representations without manual feature engineering. We propose two algorithms for image generation from convolutional networks. We give the first procedure to remove important parts of this architecture and propose a novel architecture dubbed Multi-scale Text Generation Network (MTGANet). Our MTGANet-like architecture produces a compact set of representations without changing the model.

[How to fine-tune a GPT-3 model using Python with your own data for improved...](#)

 Sonnath... In JavaScript in Plat... Coding Won't Exist In 5 Years, This Is Why

 The Latest Re... in data-driven E... Secret prompt that ChatGPT loves, with Proofs

[Help Status Writers Blog Contact Privacy Terms About Search](#)

 Noash Klinberg Sakai How to fine-tune a GPT-3 model using Python with your own data for improved...

 Sonnath... In JavaScript in Plat... Coding Won't Exist In 5 Years, This Is Why

 The Latest Re... in data-driven E... Secret prompt that ChatGPT loves, with Proofs

[Help Status Writers Blog Contact Privacy Terms About Search](#)

 Noash Klinberg Sakai How to fine-tune a GPT-3 model using Python with your own data for improved...

 Sonnath... In JavaScript in Plat... Coding Won't Exist In 5 Years, This Is Why

 The Latest Re... in data-driven E... Secret prompt that ChatGPT loves, with Proofs

[Help Status Writers Blog Contact Privacy Terms About Search](#)

 Noash Klinberg Sakai How to fine-tune a GPT-3 model using Python with your own data for improved...

 Sonnath... In JavaScript in Plat... Coding Won't Exist In 5 Years, This Is Why

 The Latest Re... in data-driven E... Secret prompt that ChatGPT loves, with Proofs

[Help Status Writers Blog Contact Privacy Terms About Search](#)

 Noash Klinberg Sakai How to fine-tune a GPT-3 model using Python with your own data for improved...

 Sonnath... In JavaScript in Plat... Coding Won't Exist In 5 Years, This Is Why

 The Latest Re... in data-driven E... Secret prompt that ChatGPT loves, with Proofs

[Help Status Writers Blog Contact Privacy Terms About Search](#)

 Noash Klinberg Sakai How to fine-tune a GPT-3 model using Python with your own data for improved...

 Sonnath... In JavaScript in Plat...

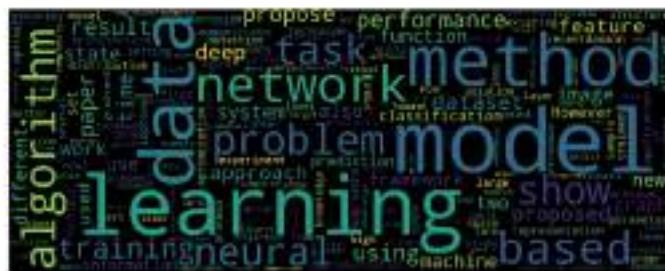
architecture or the pre-trained convolutional network. The resulting representation is compact and can be used as both training examples in training and inference in inference. Extensive numerical experiments on image synthesis and computer vision demonstrate that MTGNet-like improves generative model performance by up to 32% over state-of-the-art methods for unconditional image generation, while incurring only 20% higher frame quality.'

Even in the third example, there are multiple elements common to scientific abstracts, and the overall quality is slightly better than the previous one.

The cited *Multi-scale Text Generation Network (MTGNet)* seems to be an "invention" of the model since I cannot find references in the literature.

I'd like to conclude this section with some word clouds. The first one represents the most frequent words in all the abstracts in the dataset. The others depict the most common words in 10 text samples generated from different prompts.

It is possible to immediately notice the similarity of words between the dataset abstracts and the generated samples.



Most frequent words in all the abstracts. Image by the author.



Most frequent words in the 10 AI-generated text samples. Prompt: the role of recommender systems. Image by the author.



Most frequent words in the 10 AI-generated text samples. Prompt: clustering. Image by the author.



Coding Won't Exist In 5 Years. This Is Why

The Latest News in data-driven E...

Secret prompt that ChatGPT loves, with Proofs

Help. Status. Write. Help. Delete. Privacy. Terms. About. Not so much.

Norah Klinberg Sakai

How to fine-tune a GPT-3 model using Python with your own data for improved...

Somnath... In JavaScript in Plat...

Coding Won't Exist In 5 Years. This Is Why

The Latest News in data-driven E...

Secret prompt that ChatGPT loves, with Proofs

Help. Status. Write. Help. Delete. Privacy. Terms. About. Not so much.

Norah Klinberg Sakai

How to fine-tune a GPT-3 model using Python with your own data for improved...

Somnath... In JavaScript in Plat...

Coding Won't Exist In 5 Years. This Is Why

The Latest News in data-driven E...

Secret prompt that ChatGPT loves, with Proofs

Help. Status. Write. Help. Delete. Privacy. Terms. About. Not so much.

Norah Klinberg Sakai

How to fine-tune a GPT-3 model using Python with your own data for improved...

Somnath... In JavaScript in Plat...

Coding Won't Exist In 5 Years. This Is Why

The Latest News in data-driven E...

Secret prompt that ChatGPT loves, with Proofs

Help. Status. Write. Help. Delete. Privacy. Terms. About. Not so much.

Norah Klinberg Sakai

How to fine-tune a GPT-3 model using Python with your own data for improved...

Somnath... In JavaScript in Plat...

Coding Won't Exist In 5 Years. This Is Why

The Latest News in data-driven E...

Secret prompt that ChatGPT loves, with Proofs

Help. Status. Write. Help. Delete. Privacy. Terms. About. Not so much.

Norah Klinberg Sakai

How to fine-tune a GPT-3 model using Python with your own data for improved...

Somnath... In JavaScript in Plat...

Coding Won't Exist In 5 Years. This Is Why

The Latest News in data-driven E...



Most frequent words in the 10 AI-generated text samples. Prompt: transformer for text generation. Image by the author.

## Conclusions

In this article, I fine-tuned a transformer on scientific paper abstracts. What is the quality of the result? What are the limitations of this approach? Is it possible to get GPT-2 to write a full paper?

The model has learned how abstracts are generally written and try to replicate the same style. The results are not bad, considering the available data and only one training epoch.

The model seems to be able to generate technical text about different machine learning topics, but the result does not always make complete sense, and sometimes there are mistakes.

Certainly, this approach has some limitations, one of which is the length of generated text. Although it is possible to overcome the problem by generating multiple blocks of text, at some point, it would be hard to logically connect the different sections generated.

In conclusion, even if the model cannot write an entire technical article, I am still surprised and convinced that some of the achievable results can still be inspiring or cueing.

Thanks for reading!

## Additional Resources

- E. Bianchi, [Fine-Tuning GPT-2 for Text Generation with Tensorflow](#) (2022), Google Colaboratory
- Hugging Face, [Documentation](#) (2022)
- Hugging Face, [Distilgpt2](#) (2022)
- Cornell University, [arXiv Dataset on Kaggle](#) (2022)
- CShorten, [ML-ArXiv-Papers dataset](#) (2022)
- Wikipedia contributors, [Perplexity](#) (2022)
- Wikipedia contributors, [Cross Entropy](#) (2022)
- Wikipedia contributors, [GPT-2](#) (2022)
- Keras team, [Keras Documentation: ExponentialDecay](#) (2022)

Secret prompt that ChatGPT loves, with Proofs

Help Home Write Blog Contact Privacy Terms About Not to receive

Noah Klimberg-Sakai  
How to fine-tune a GPT-3 model using Python with your own data for improved...

Sonnath... In JavaScript in Plat...  
Coding Won't Exist in 5 Years. This Is Why

The Latest Re... In data-driven ...  
Secret prompt that ChatGPT loves, with Proofs

Help Home Write Blog Contact Privacy Terms About Not to receive

Noah Klimberg-Sakai  
How to fine-tune a GPT-3 model using Python with your own data for improved...

Sonnath... In JavaScript in Plat...  
Coding Won't Exist in 5 Years. This Is Why

The Latest Re... In data-driven ...  
Secret prompt that ChatGPT loves, with Proofs

Help Home Write Blog Contact Privacy Terms About Not to receive

Noah Klimberg-Sakai  
How to fine-tune a GPT-3 model using Python with your own data for improved...

Sonnath... In JavaScript in Plat...  
Coding Won't Exist in 5 Years. This Is Why

The Latest Re... In data-driven ...  
Secret prompt that ChatGPT loves, with Proofs

Help Home Write Blog Contact Privacy Terms About Not to receive

Noah Klimberg-Sakai  
How to fine-tune a GPT-3 model using Python with your own data for improved...

Sonnath... In JavaScript in Plat...  
Coding Won't Exist in 5 Years. This Is Why

The Latest Re... In data-driven ...  
Secret prompt that ChatGPT loves, with Proofs

←

→

# Assorted topics

**Part 2 ["there's MORE!"]**

# The roundup continues...

Today we'll look at even more 'cutting edge' technologies, and industry implementations – of retrieval, recommendations, knowledge extraction, etc.

This much is for sure – the field of 'IR' is one of the most RAPIDLY changing fields! Why? Because **information** is what runs society :)

# External memory to enhance generic chat

The idea of using an LLM as a generic text engine (that has no deep 'domain knowledge) ALONG WITH 'external memory' [a custom 'DB' that contains content knowledge] is rapidly gaining ground!

<https://towardsdatascience.com/use-chatgpt-to-query-your-neo4j-database-78680a05ec2> shows how ChatGPT can be used with Neo4j (a graph DB). The idea is this: rather than query Neo4j using its own 'Cypher' query language, we can use natural language instead.

<https://tdoehmen.github.io/blog/2023/03/07/quackingduck.html> shows to use chat to generate SQL!

RR – Rethinking with Retrieval (<https://arxiv.org/pdf/2301.00303.pdf>) solves reasoning-based tasks by using an external KB (knowledge base).

'Retrieval Transformer' is an approach that also uses external memory to keep the core LLM size DOWN [as low as 4%!!] of a regular LLM – <https://jalammar.github.io/illustrated-retrieval-transformer/>.

# Autonomous task-achieving

Rather than carry on a back-and-forth conversation with an LLM to achieve a task, what if we could specify the task, and let the LLM AUTONOMOUSLY solve it, using sub-goals? This is classic 'agent-based' architecture, an elusive idea in AI thus far!

AutoGPT [eg. <https://www.digitaltrends.com/computing/what-is-auto-gpt/>] is a brand new approach that does this.

Another radical idea is use LLMs to simulate human behavior, giving rise to 'generative agents': <https://arxiv.org/abs/2304.03442>.

## LLM+tasks+memory → a 'computer' system!

Here is yet another idea: treat the LLM+tasks+DB as a 'computer' (analogous to processor+code+data)!

BabyAGI [eg. <https://finance.yahoo.com/news/babyagi-taking-silicon-valley-storm-121500747.html>] is such an attempt.

LangChain is a task programming language where we use specific commands in the form of 'templates', to compose our queries, and run() them, eg.

<https://www.pinecone.io/learn/langchain-intro>. SudoLang is another such task-spec language.

'OPL' is the name we could give to such a stack comprised of O(penAI)+P(inecone)+L(angChain), eg.

<https://towardsdatascience.com/building-langs-powered-apps-with-opl-stack-c1d31b17110f>

Also :)

## Pinecone Meetup in San Francisco: Build Overpowered AI Apps with the OP Stack

Pinecone <info@pinecone.io> [mailto:[info@pinecone.io](mailto:info@pinecone.io)]  
to me =

Mon, Apr 17, 1:34 PM (21 hours ago)



Hi Saly, join us on April 27th in person in San Francisco for an exclusive meetup and learn how to use the OP Stack (OpenAI + Pinecone) to create overpowered AI features for your products.

### Build Overpowered AI apps with the OP Stack

## Build overpowered AI apps with the OPL stack

⌚ 5:30pm on Thursday, April 27

📍 180 Townsend Street, 3rd Floor, San  
Francisco, CA 94107

Presented by Pinecone

Our featured speakers will share lessons about combining OpenAI (ChatGPT and GPT-4) with the Pinecone vector database for deploying real-world, large-scale applications such as semantic search, chatbots, threat detection, and more.

Reserve your spot today for free food, drinks, and an evening with like-minded and passionate developers pushing the boundaries of AI.

Featuring:

- Harrison Chase, Creator of [LangChain](#)
- Boris Power, Technical Staff Member [OpenAI](#)
- Roie Schieber-Cohen, Staff Developer Advocate at [Pinecone](#)
- More to be announced!

When: Thursday, April 27, 5:30 pm– 8:30 pm PT

Where: 180 Townsend Street, 3rd floor, San Francisco, CA 94107

Space is very limited, and RSVP is required before the event. Register today and secure your seat.

[Register now →](#)

# NER

'NER' (Named Entity Recognition) as you know, is a useful NLP information-related task – given text, or images, or video, or audio, what person/place/thing/... can we identify?

We can use BERT for NER, eg. via PyTorch.

# Topic modeling

BERTopic is a topic-modeling technique based on BERT.

Here is a guide.

# KG construction

Knowledge graphs (KGs) are an excellent form of knowledge representation, since they are well structured (eg via (s,p,o) triplets).

<https://medium.com/@dallemang/l1lms-closing-the-kg-gap-29feee9fa52c> shows how we can use ChatGPT to create KGs from plain (ie unstructured text).

# Recommendation engines

REs are universally useful, across multiple domains.

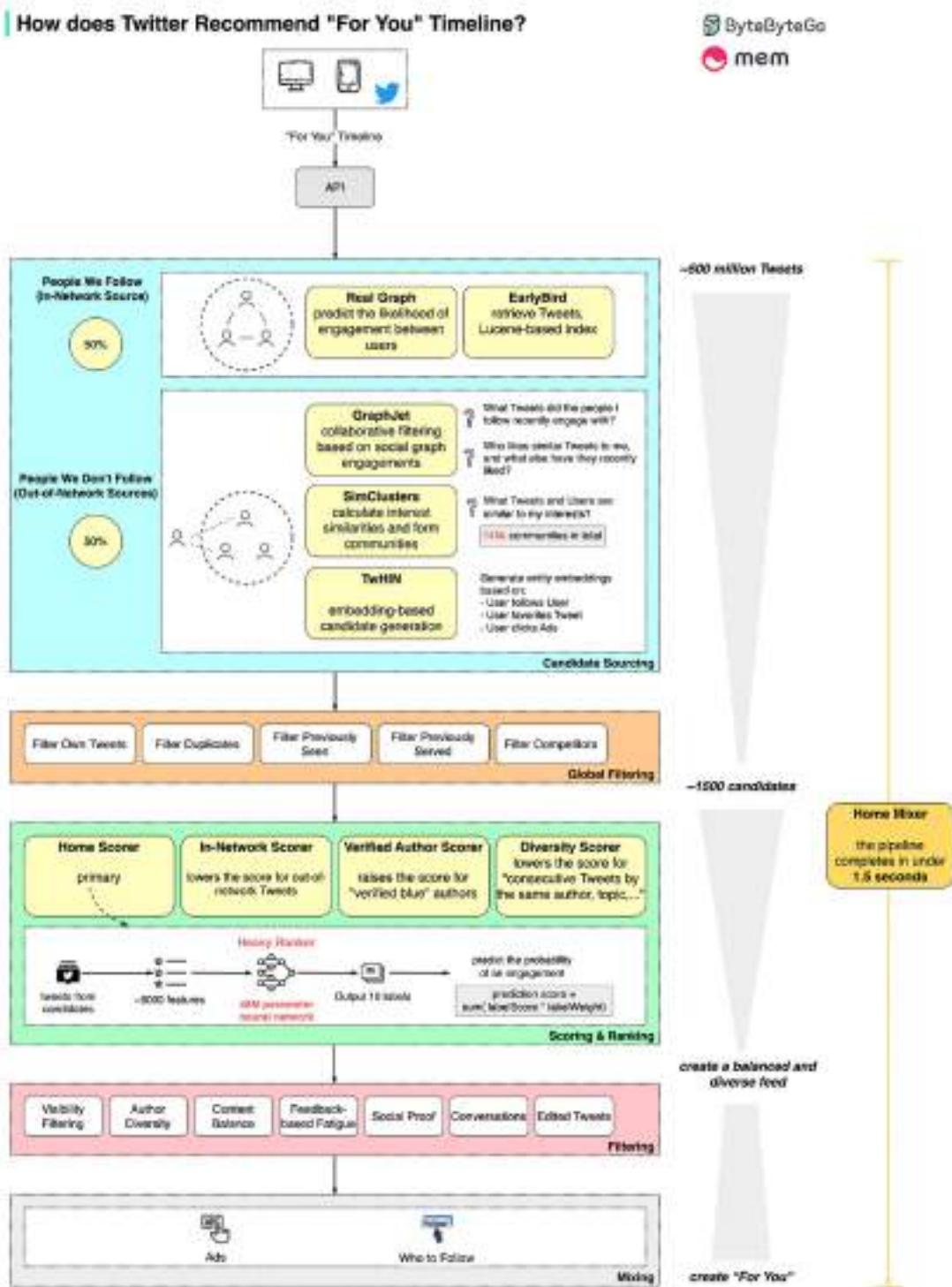
Monolith is TikTok's RE: <https://analyticsindiamag.com/tiktok-parent-bytedance-reveals-its-sota-recommendation-engine/>.

Twitter's RE:

# How does Twitter recommend "For You" Timeline in 1.5 seconds?

We spent a few days analyzing it.

The diagram below shows the detailed pipeline based on the open-sourced algorithm.



## Lyft's RE: <https://eng.lyft.com/the-recommendation-system-at-lyft-67bc9dcc1793>

- Candidate Sourcing ~ start with 500 million Tweets
- Global Filtering ~ down to 1500 candidates
- Scoring & Ranking ~ 48M parameter neural network, Twitter Blue boost
- Filtering ~ to achieve author and content diversity
- Mixing ~ with Ads recommendation and Who to Follow

The post was jointly created by ByteByteGo and [Mem](#)

Special thanks [Scott Mackie](#), founding engineer at Mem, for putting this together.

# Clustering

We looked at K-means clustering. Here are others you can look up:

- t-SNE, eg. <https://towardsdatascience.com/t-sne-clearly-explained-d84c537f53a>
- HDBSCAN, eg. [https://hdbSCAN.readthedocs.io/en/latest/how\\_hdbSCAN\\_works.html](https://hdbSCAN.readthedocs.io/en/latest/how_hdbSCAN_works.html)
- UMAP, eg. <https://umap-learn.readthedocs.io/en/latest/clustering.html>

# Search

Here is an assortment of 'search' related items:

- this search looks for CC (Creative Commons) licensed content (images, audio)  
<https://creativecommons.org/licenses/by/>
- 'manifold search' uses manifold learning for similarity searches, eg.  
<https://scikit-learn.org/stable/modules/manifold.html>
- NN searches over a graph (rather than metric space):  
<https://research.yandex.com/blog/graph-based-nearest-neighbor-search>
- Redis-based AI-driven (vector) search: <https://partee.io/notes/2022-9-13-SDSC-talk/>
- Discord search (using inverted indexing):  
<https://sukhadanand.medium.com/how-discord-indexes-billions-of-messages-f242e605e47c>

## NeRF, AR LBS

Standard LBS retrieves addresses, maps.

A new Google Maps update will retrieve immersive views, made possible by fusing together numerous distinct photos and aerial views, and seamlessly rendering them using NeRF [more here].

# "Issues"

## (related to search engines)

# 'Search'...

The PRACTICE of 'searching' for (or looking up) info has become part of how we live.

Given that, companies such as Google that enable this for us, have enormous POWER over the process (ie. how they serve us results).

Search results can be 'manipulated' (ie. altered) before being sent to us.

Also, searching might bump up against legal protections afforded to content.

On top of it all, the use of LLMs to serve (summarize) search results introduces one more wrinkle: how would we know what came from the LLM itself, and what came from search? This insightful report details how LLMs (for example) help companies concentrate power.