Search Engine
Question Answering

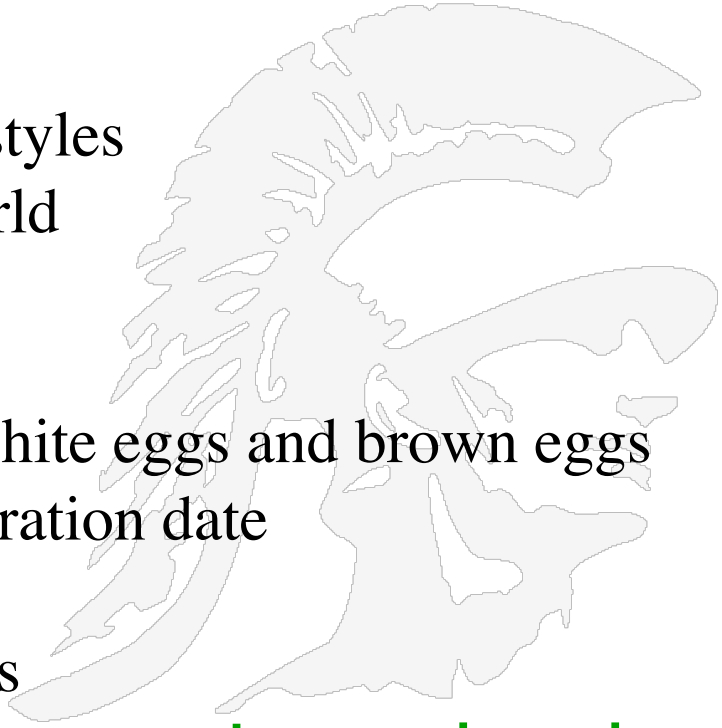
Information Retrieval v. Question Answering

- The name **"information retrieval"** is standard, but as traditionally practiced, it's not really right
- In the past all we got was ***document retrieval***, and beyond that the job is up to us
  - Modern search engines now do a lot better
- They combine information retrieval, knowledge graphs and inferencing, past query history, location, natural language processing and many other criteria
- **Question Answering** (**QA**) is concerned with building systems that automatically answer questions posed by humans in a natural language

People *want* to ask questions…

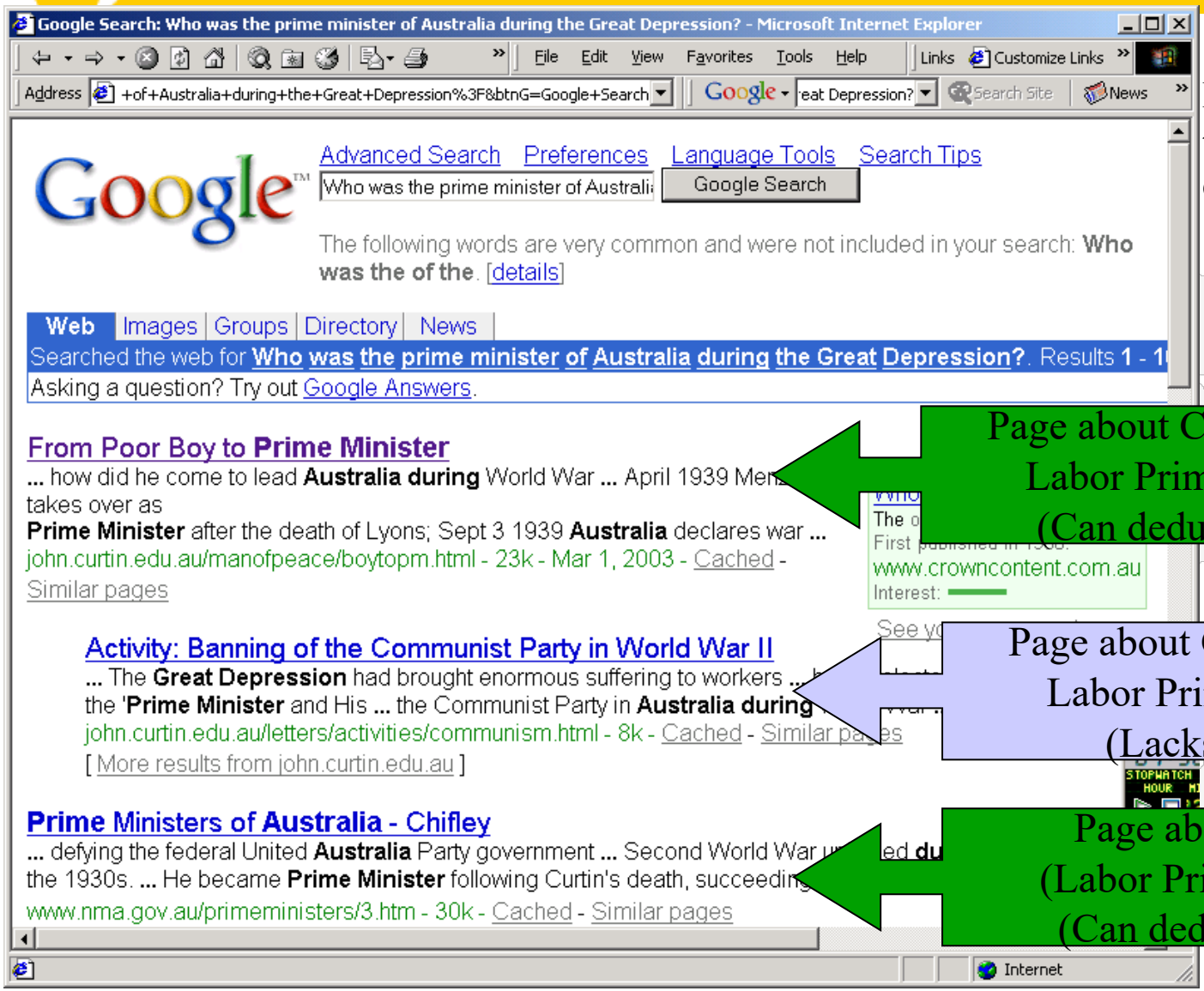
Examples from Ask.com query log

how much should I weigh
what does my name mean
how to get pregnant
where can I find pictures of hairstyles
who is the richest man in the world
what is the meaning of life
why is the sky blue
what is the difference between white eggs and brown eggs
can you drink milk after the expiration date
what is true love
what is the jonas brothers address



Around 10-20% of query logs are questions such as these

Question: *Who was the prime minister of Australia during the Great Depression?*
Answer: *James Scullin (Labor) 1929–31*



How Google used to respond to questions

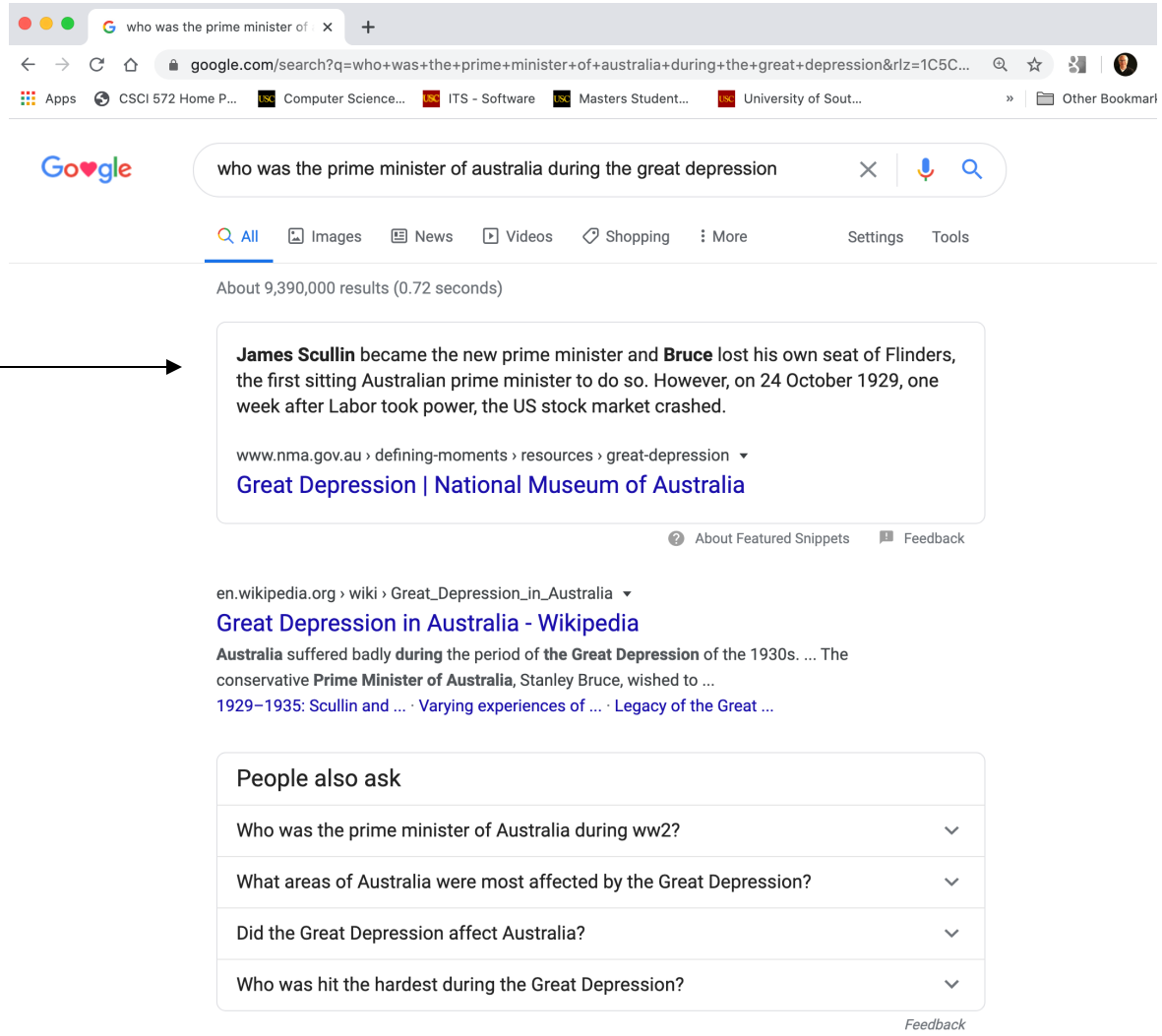
Page about Curtin (WW II Labor Prime Minister) (Can deduce answer)

Page about Curtin (WW II Labor Prime Minister) (Lacks answer)

Page about Chifley (Labor Prime Minister) (Can deduce answer)

Question: *Who was the prime minister of Australia during the Great Depression?*
Answer: *James Scullin (Labor) 1929–31*

Google's result today



Google has Improved Its Ability to Answer Many Questions

how old is mariah carey - Goog

google.com/search?q=how+old+is+mariah+carey&rlz=1C5CHFA_enUS728US728&oq=how+ol...

CSCI 572 Home P... Piazza Spring2022 CSCI572_Spring2... DEN D2L Page USC Schedule of... Other Bookmarks

Google how old is mariah carey

All Images News Books Videos More Tools

About 127,000,000 results (0.73 seconds)

Mariah Carey / Age

53 years

March 27, 1969

People also search for


Nick Cannon
41 years


Jennifer Lopez
52 years


Beyoncé
40 years

Feedback

People also ask

What is Mariah Carey's net worth 2021?

Why does Mariah Carey touch her ear?

Is Mariah Carey richer than Nick Cannon?

Who is Mariah Carey husband?

Feedback

Mariah Carey

American singer-songwriter

Available on

YouTube

Spotify

Apple Music

More music services

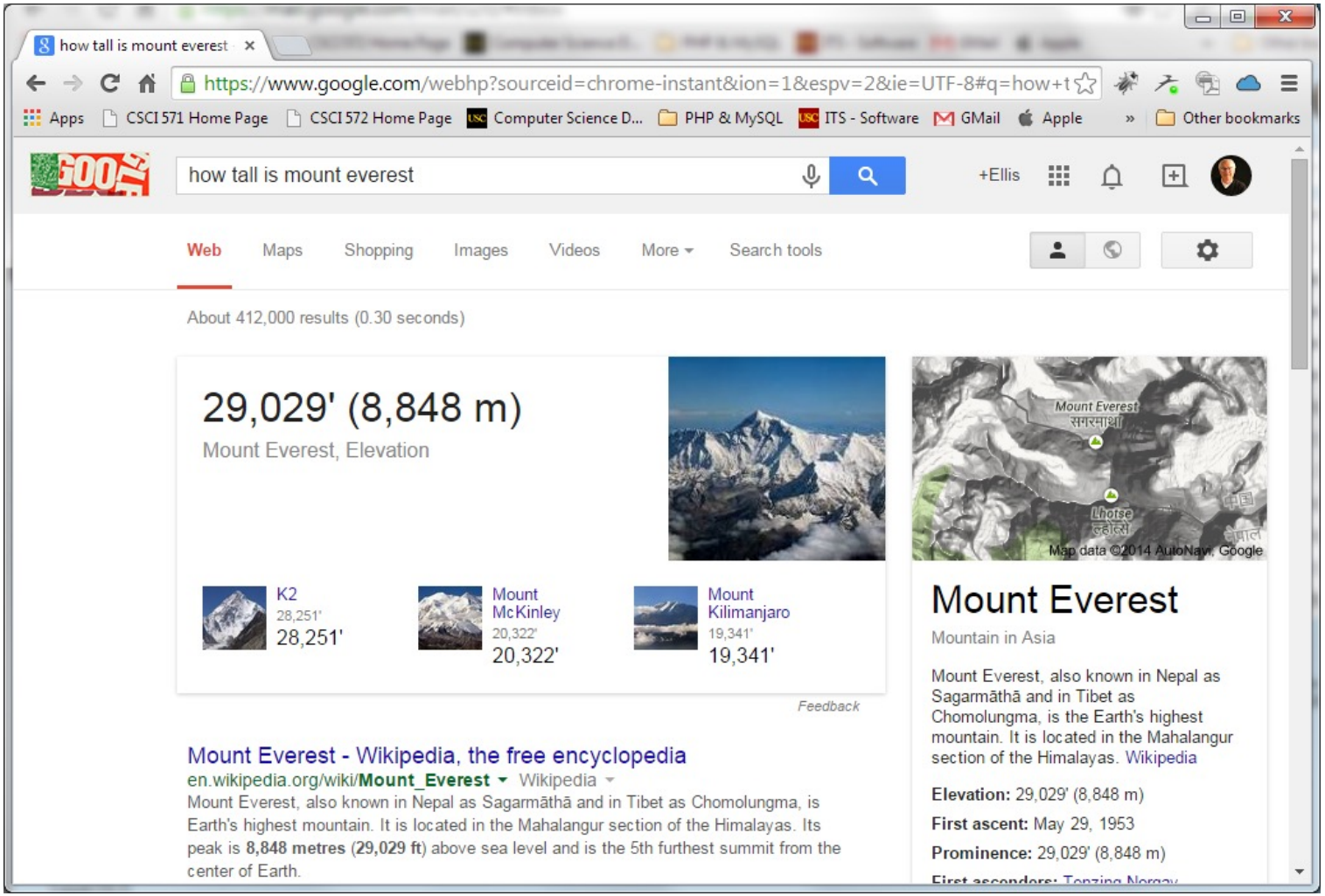
Mariah Carey is an American sir
actress, and record producer. Re
"Songbird Supreme" and the "Qu
she is noted for her five-octave v
melismatic singing style, and sig
whistle register. Carey rose to fa
eponymous debut album. Wikipe

Born: March 27, 1969 (age 53 y

Children: Moroccan Scott Canno

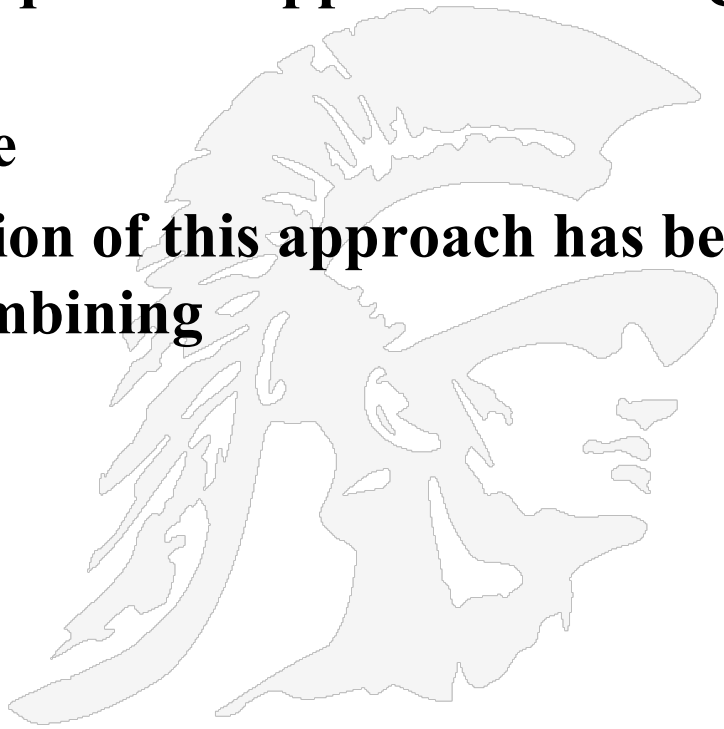
Spouse: Nick Cannon (m. 2008–
Mottola (m. 1993–1998)

Some Questions are Easily Answered



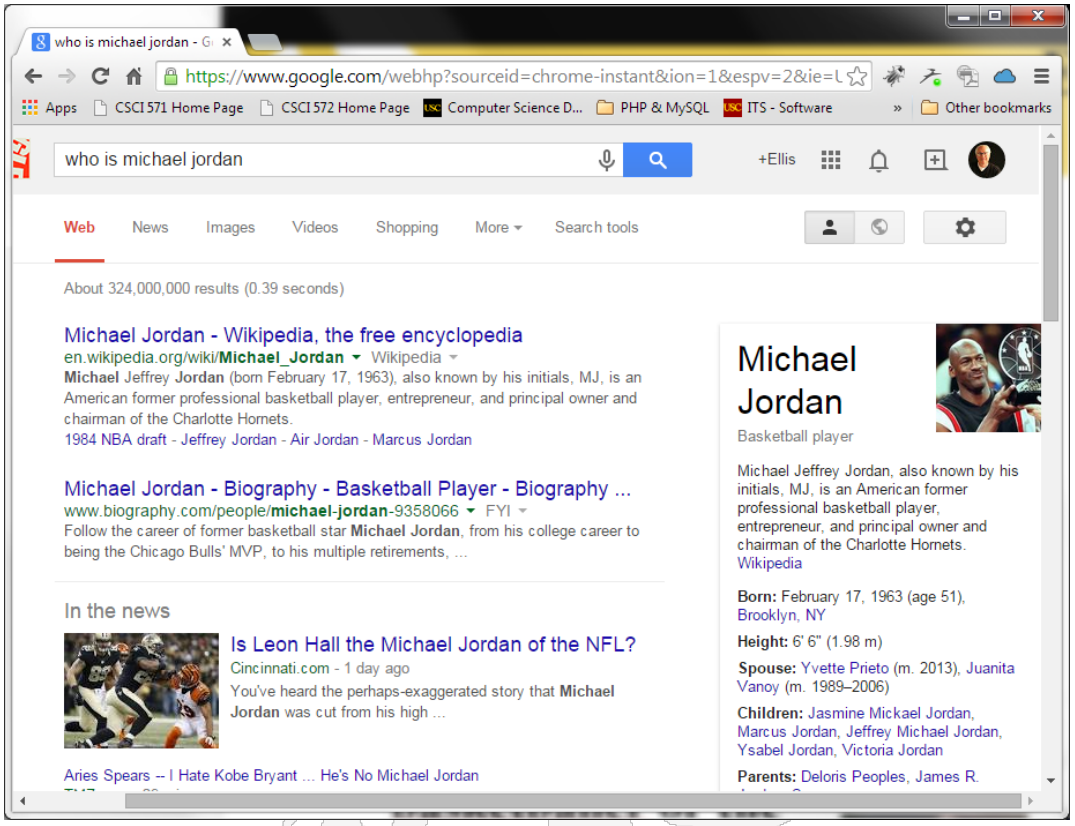
The Original Google Approach

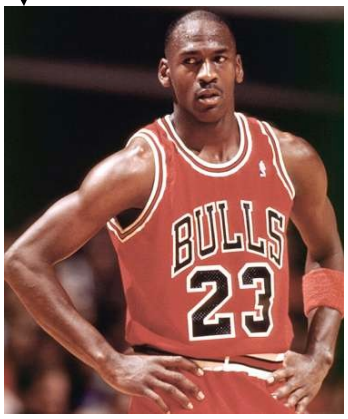
- **Take the question and try to find it as a string on the web**
- **Return the next sentence on that web page as the answer**
- **Works brilliantly if this exact question appears as a FAQ question, etc.**
- **Works poorly most of the time**
- **But a more sophisticated version of this approach has been introduced in recent years combining**
  - Knowledge graph
  - N-grams
  - WordNet
  - NLP techniques



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





Why Natural Language Processing is Required

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

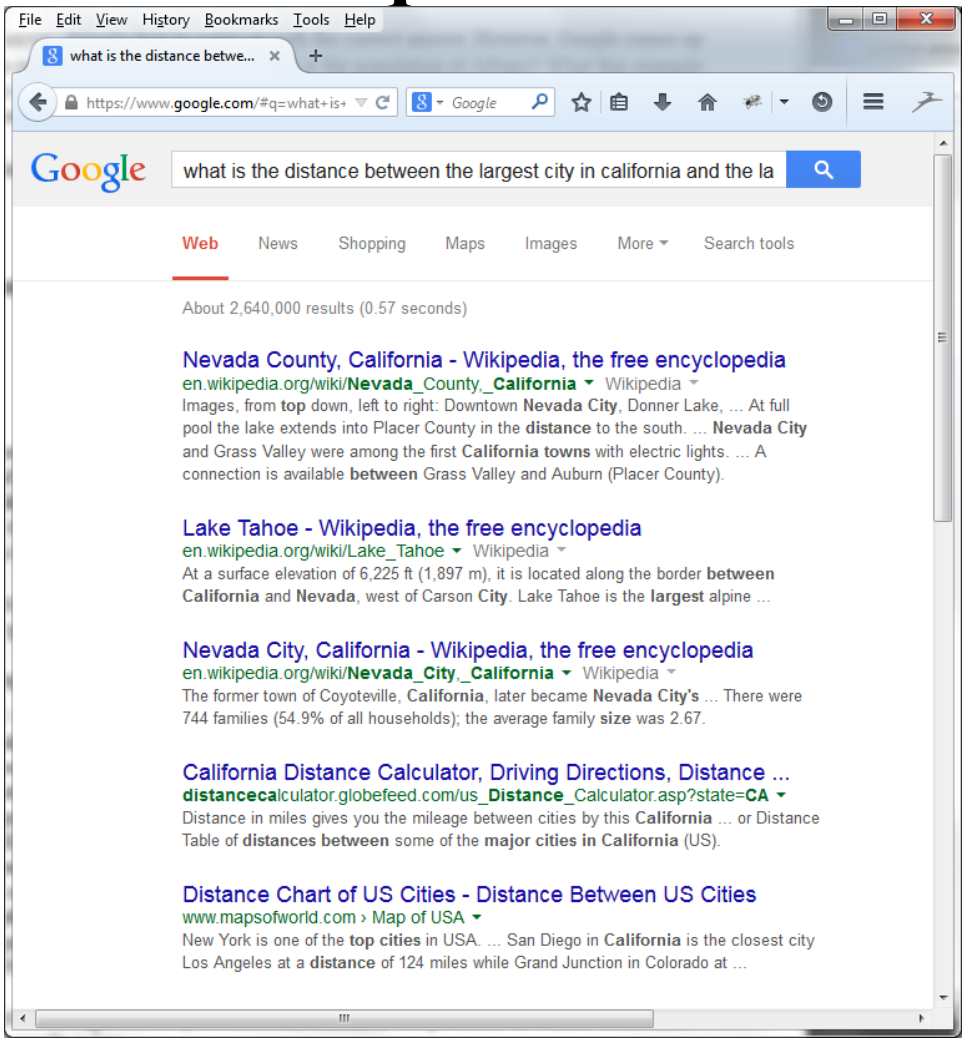
NLP: Natural language processing

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

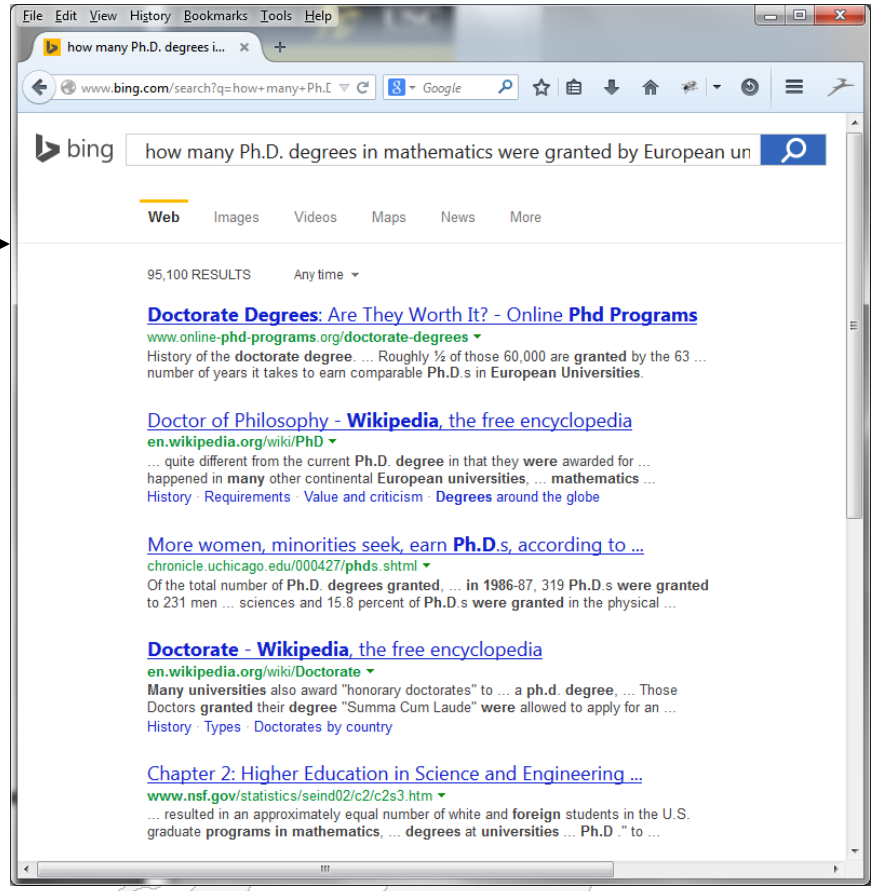
ps: Try the query in Google today to see if they have improved their answer



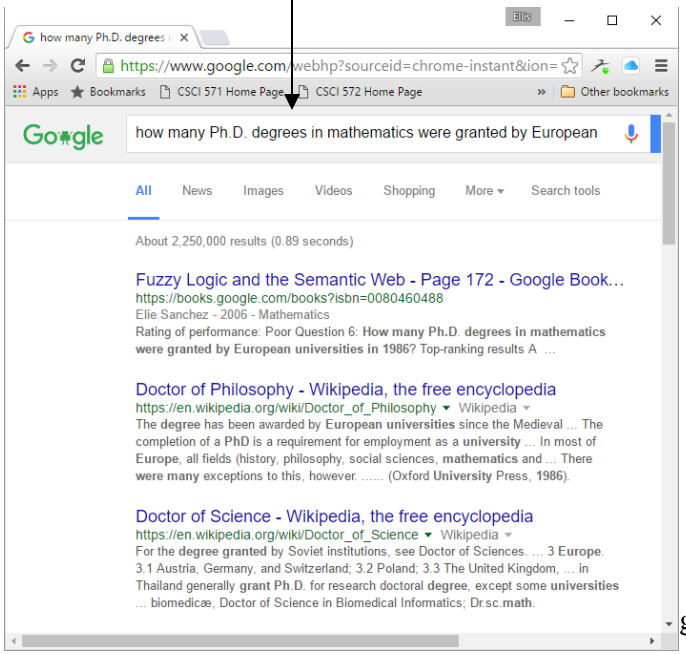
In Some Cases the Data May Not Exist

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



Some Popular Products Designed for Question/Answering

- **Approach 1: used by Siri**
  - map to known entities and use existing databases over the internet
- **Approach 2: used by Ask.com**
  - detect question type and use a search engine's snippet results
- **Approach 3: used by IBM's Watson**
  - combine approaches 1 and 2
- **Approach 4: used by Google's Knowledge Graph**
  - use an entity - relationship graph that is built in-house and infer the answer

Approach used by Siri: Knowledge-Based Approach

- **Siri was begun as a DARPA project called CALO/PAL (Personalized Assistant that Learns)**

1. **First your voice query is put through a recognizer and a language model and Siri comes up with an interpretation of what was said**
2. **Second Siri builds a semantic representation of the query**
   - Extract times, dates, locations, entities, numeric quantities, and especially user actions (e.g. schedule a meeting, set my alarm, etc)
3. **Siri maps from this semantics to query structured data or resources including:**
   - Geospatial databases
   - Ontologies (Wikipedia infoboxes, Freebase, WordNet, Yago)
   - Restaurant review sources and reservation services (Yelp)
   - Scientific databases (Wolfram Alpha)
   - Conventional search engines (Google, Bing)
4. **Siri then transforms the output above into natural language and then transforms the text back to speech**

Context and Conversation in Virtual Assistants like Siri

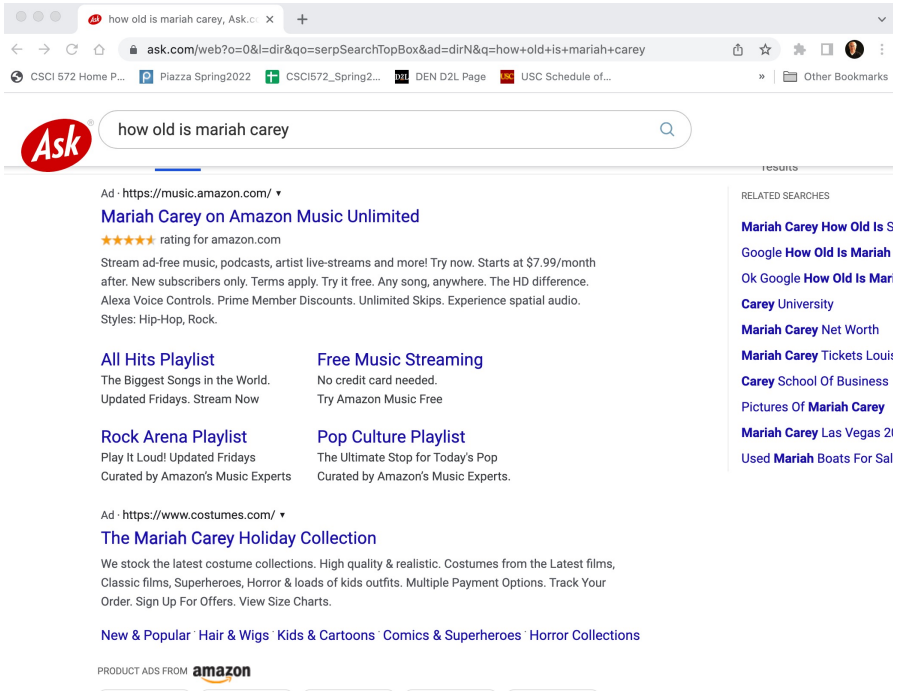
- Coreference helps resolve ambiguities
- U: “book a table at Il Fornaio at 7:00 with my mom”
- U: “also send her an email reminder”
- “her” refers to “my mom”; 7:00 refers to pm, not am
- Clarification questions:
- U: “chicago pizza”
- S: “Did you mean pizza restaurants in Chicago or Chicago-style pizza?”

IBM's WATSON System

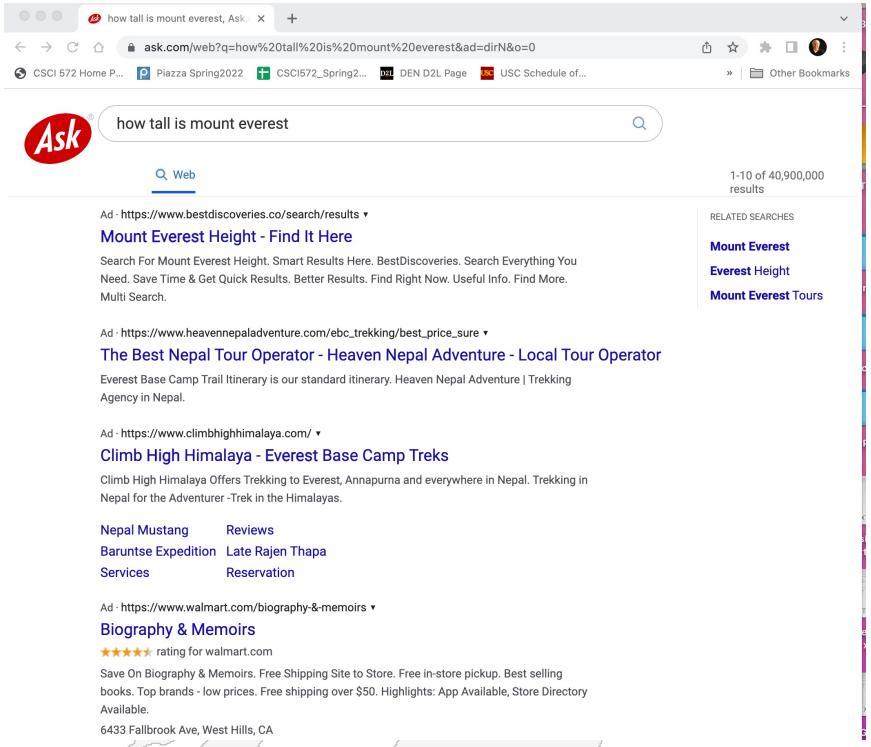
- Watson was created as a question answering (QA) computing system that IBM built to apply advanced natural language processing, information retrieval, knowledge representation, automated reasoning, and machine learning technologies to the field of open domain question answering
- Watson won a contest on the program Jeopardy against the best human winners
  - https://www.nytimes.com/2011/02/17/science/17jeopardy-watson.html
- However, it has failed to meet earlier expectations, see
  - https://www.nytimes.com/2021/07/16/technology/what-happened-ibm-watson.html
- The Allen Institute for Artificial Intelligence compared Watson's performance on standard natural language tasks like identifying persons, places and the sentiment of a sentence with the A.I. services offered by the big tech cloud providers — Amazon, Microsoft and Google. Watson did as well as, and sometimes better than, the big three
- IBM's Watson in 3 minutes,
  - https://www.techrepublic.com/article/ibm-watson-the-smart-persons-guide/

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

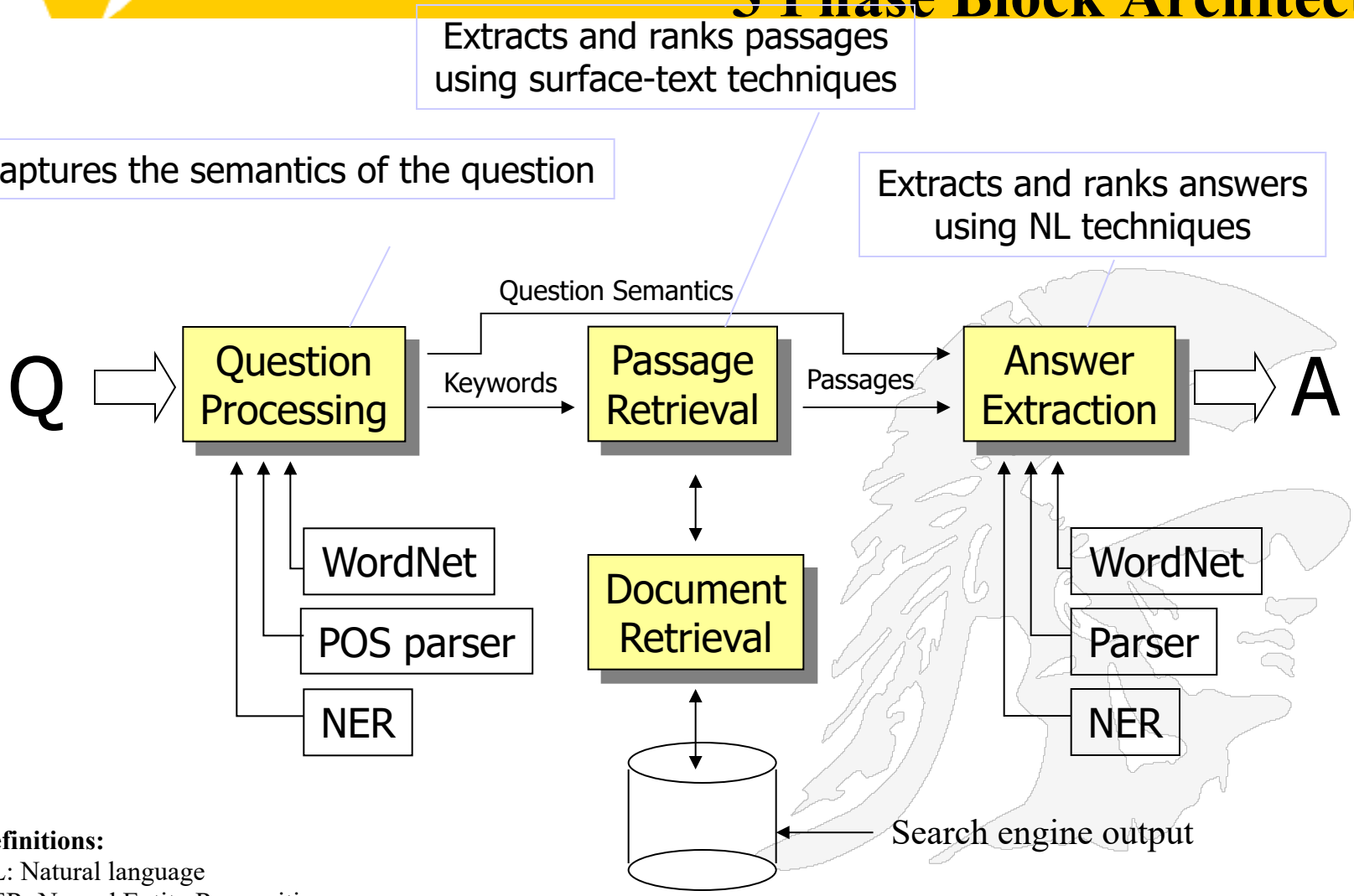
Question Types: Many Questions Fall into Distinct Categories

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

3 Main Phases for Question/Answering

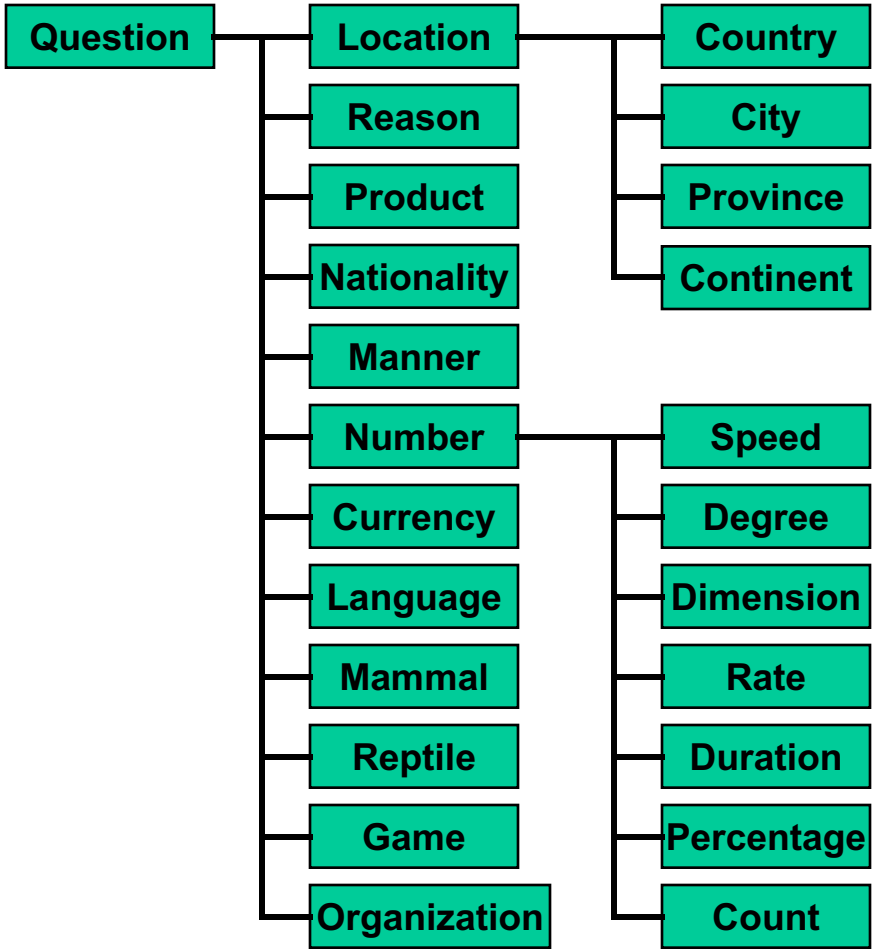
1. **QUESTION PROCESSING**
   - Detect question type (who, what, when, where, etc)
   - Identify important entities and formulate queries to send to a search engine
2. **PASSAGE RETRIEVAL**
   - Retrieve ranked documents (snippets only)
   - Break into suitable passages and match against entities
3. **ANSWER PROCESSING**
   - Extract candidate answers (as named entities)
   - Rank candidates
     - **using evidence from relations in the text and external sources**

Question Answering
3 Phase Block Architecture

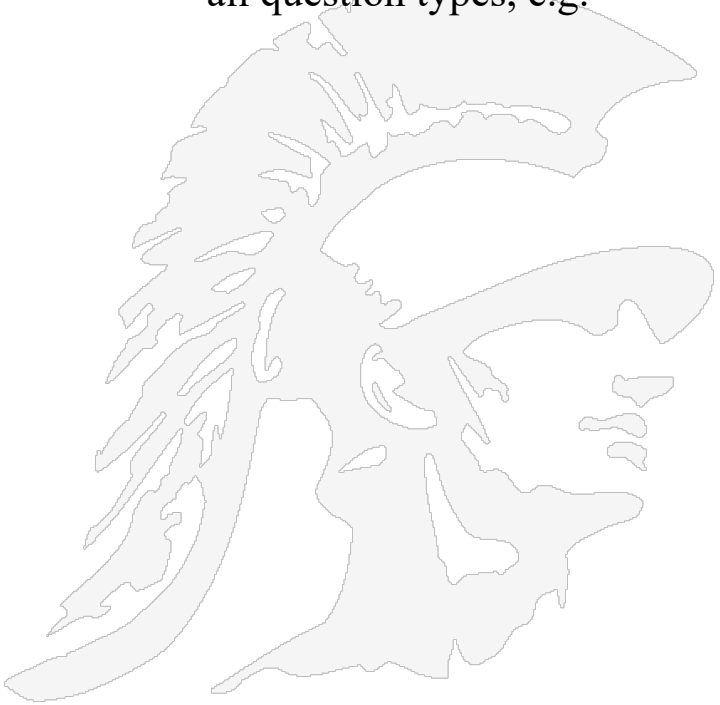


Definitions:
NL: Natural language
NER: Named Entity Recognition

Question Taxonomy



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



However Question Taxonomies Can Get Very Large

- **Factoid Questions**
  - Who, where, when, how many
  - The answers fall into a limited and somewhat predictable set of categories, see e.g. to the right
- This set is from a paper by Li and Roth, 2005, *Learning Question Classifiers*
- Their major categories include
  - Abbreviation
  - Description
  - Entity
  - Human
  - Location
  - Numeric
- Questions can be labeled at one or two levels, e.g. either as Numeric or as Numeric:date or as Location or Location:mountain
- This approach argues for a template applied to the query

Tag	Example
ABBREVIATION	
abb	What's the abbreviation for limited partnership?
exp	What does the "c" stand for in the equation E=mc2?
DESCRIPTION	
definition	What are tannins?
description	What are the words to the Canadian National anthem?
manner	How can you get rust stains out of clothing?
reason	What caused the Titanic to sink ?
ENTITY	
animal	What are the names of Odin's ravens?
body	What part of your body contains the corpus callosum?
color	What colors make up a rainbow ?
creative	In what book can I find the story of Aladdin?
currency	What currency is used in China?
disease/medicine	What does Salk vaccine prevent?
event	What war involved the battle of Chapultepec?
food	What kind of nuts are used in marzipan?
instrument	What instrument does Max Roach play?
lang	What's the official language of Algeria?
letter	What letter appears on the cold-water tap in Spain?
other	What is the name of King Arthur's sword?
plant	What are some fragrant white climbing roses?
product	What is the fastest computer?
religion	What religion has the most members?
sport	What was the name of the ball game played by the Mayans?
substance	What fuel do airplanes use?
symbol	What is the chemical symbol for nitrogen?
technique	What is the best way to remove wallpaper?
term	How do you say " Grandma " in Irish?
vehicle	What was the name of Captain Bligh's ship?
word	What's the singular of dice?

More Question Types and Examples

HUMAN	
description	Who was Confucius?
group	What are the major companies that are part of Dow Jones?
ind	Who was the first Russian astronaut to do a spacewalk?
title	What was Queen Victoria's title regarding India?
LOCATION	
city	What's the oldest capital city in the Americas?
country	What country borders the most others?
mountain	What is the highest peak in Africa?
other	What river runs through Liverpool?
state	What states do not have state income tax?
NUMERIC	
code	What is the telephone number for the University of Colorado?
count	About how many soldiers died in World War II?
date	What is the date of Boxing Day?
distance	How long was Mao's 1930s Long March?
money	How much did a McDonald's hamburger cost in 1963?
order	Where does Shanghai rank among world cities in population?
other	What is the population of Mexico?
period	What was the average life expectancy during the Stone Age?
percent	What fraction of a beaver's life is spent swimming?
speed	What is the speed of the Mississippi River?
temp	How fast must a spacecraft travel to escape Earth's gravity?
size	What is the size of Argentina?
weight	How many pounds are there in a stone?

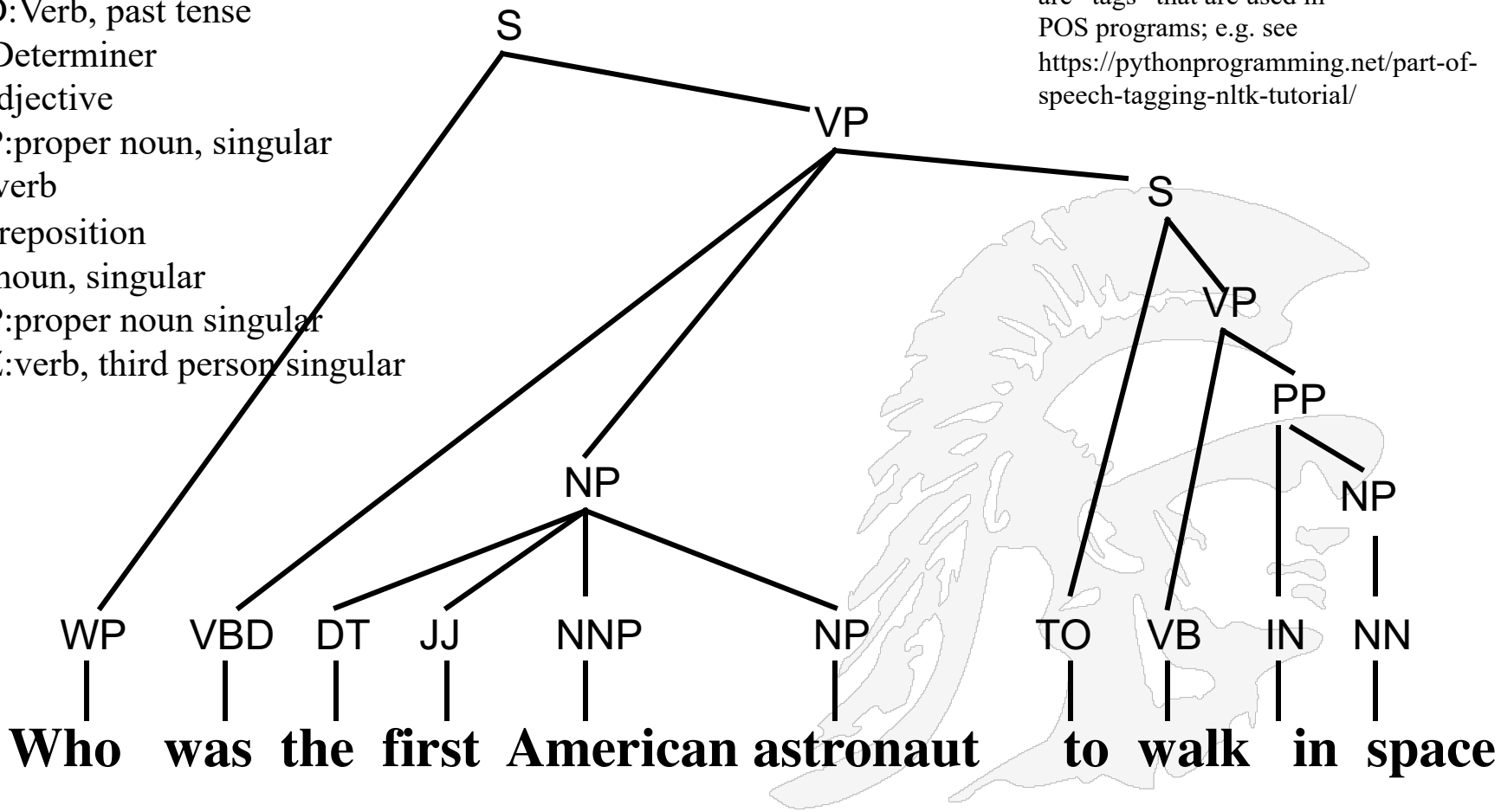
Some General Capabilities for Question-Answering Systems

- **Part-of-Speech Tagging**
  - a piece of software that reads text in some language and assigns **parts** of **speech** to each word, such as noun, verb, adjective, etc.
  - Markov Models are now the standard method for part-of-speech assignment
  - Some current major algorithms for part-of-speech tagging include the Viterbi algorithm, Brill tagger, and Baum-Welch algorithm, see https://en.wikipedia.org/wiki/Part-of-speech_tagging
- **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
- **Determining Semantic Relations**
  - **semantic relations are** concepts or meanings between entities. E.g. [School] is a kind of [educational institution]
    - Opportunity to use WordNet, https://wordnet.princeton.edu/
- **Dictionaries/Thesauri**

Question Processing Tool
Part-of-Speech Recognizer

WP:Wh-pronoun (who/what/when/where)
VBD:Verb, past tense
DT:Determiner
JJ:Adjective
NNP:proper noun, singular
VB:verb
IN:preposition
NN:noun, singular
NNP:proper noun singular
VBZ:verb, third person singular

The two-letter abbreviations are "tags" that are used in POS programs; e.g. see https://pythonprogramming.net/part-of-speech-tagging-nltk-tutorial/

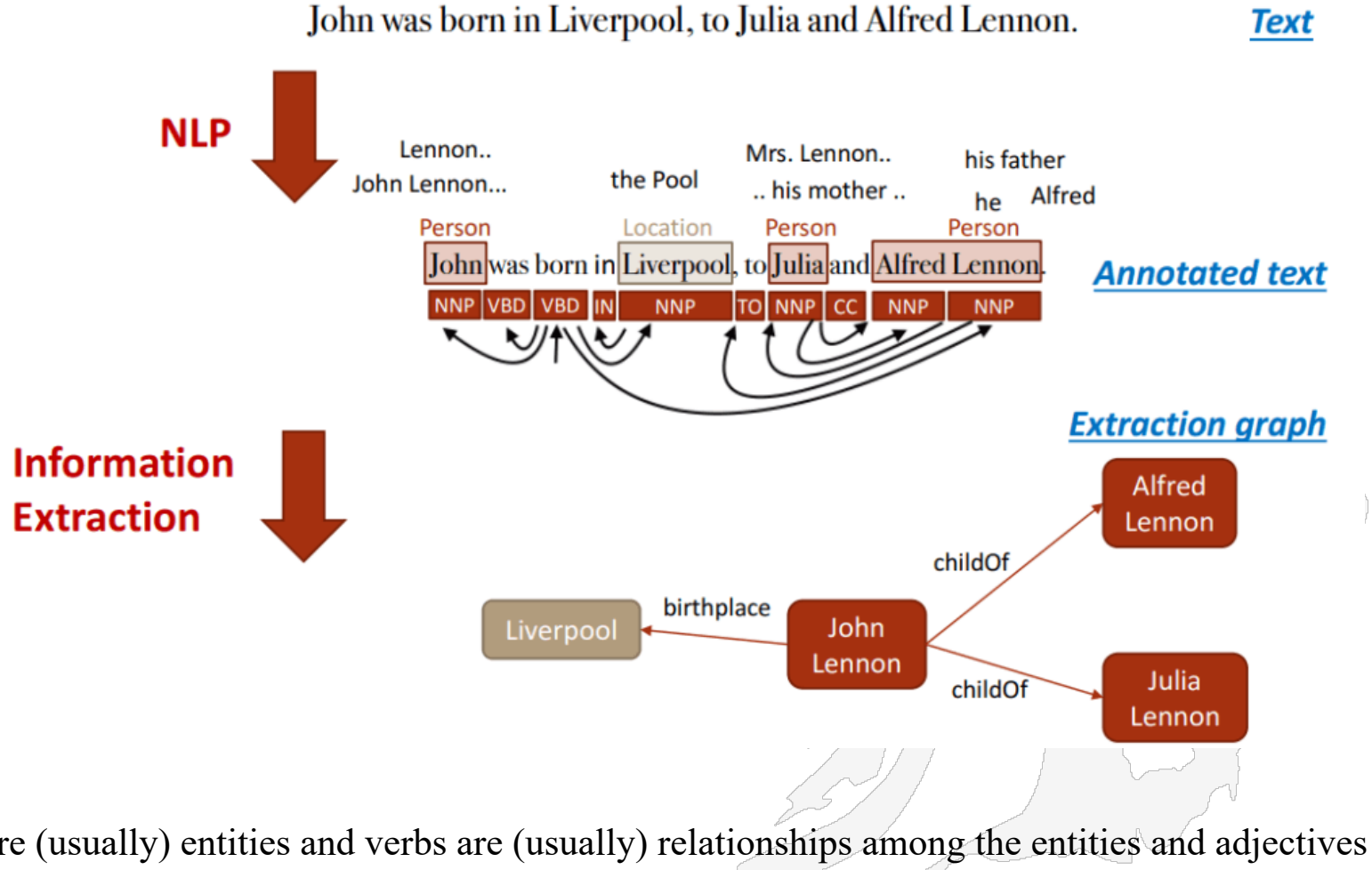


Question Processing Tool
Named Entity Recognizer Example

contentSkip to site indexPoliticsSubscribeLog InSubscribeLog InToday's PaperAdvertisementSupported ORG byF.B.I. Agent Peter Strzok PERSON , Who Criticized Trump PERSON in Texts, Is FiredImagePeter Strzok, a top F.B.I. GPE counterintelligence agent who was taken off the special counsel investigation after his disparaging texts about President Trump PERSON were uncovered, was fired. CreditT.J. Kirkpatrick PERSON for The New York TimesBy Adam Goldman ORG and Michael S. SchmidtAug PERSON . 13 CARDINAL , 2018WASHINGTON CARDINAL — Peter Strzok PERSON , the F.B.I. GPE senior counterintelligence agent who disparaged President Trump PERSON in inflammatory text messages and helped oversee the Hillary Clinton PERSON email and Russia GPE investigations, has been fired for violating bureau policies, Mr. Strzok PERSON 's lawyer said Monday DATE .Mr. Trump and his allies seized on the texts — exchanged during the 2016 DATE campaign with a former F.B.I. GPE lawyer, Lisa Page — in PERSON assailing the Russia GPE investigation as an illegitimate "witch hunt." Mr. Strzok PERSON , who rose over 20 years DATE at the F.B.I. GPE to become one of its most experienced counterintelligence agents, was a key figure in the early months DATE of the inquiry.Along with writing the texts, Mr. Strzok PERSON was accused of sending a highly sensitive search warrant to his personal email account.The F.B.I. GPE had been under immense political pressure by Mr. Trump PERSON to dismiss Mr. Strzok PERSON , who was removed last summer DATE from the staff of the special counsel, Robert S. Mueller III PERSON . The president has repeatedly denounced Mr. Strzok PERSON in posts on

- The process of recognizing information units like names, including persons, organizations, location names, and numeric expressions including time, date, money and percent expressions from unstructured text.
- This is an example of supervised learning as training sets are first created
- See https://nlp.stanford.edu/software/ for a Java program and explanation

Example of NLP Extraction Used to Build a Knowledge Graph:



Nouns are (usually) entities and verbs are (usually) relationships among the entities and adjectives (usually) describe the relationships

https://kgtutorial.github.io/wsdm-slides/Part2_knowledge-extraction.pdf

NER Example and Its Translation
Jeopardy Example

The Jeopardy query:

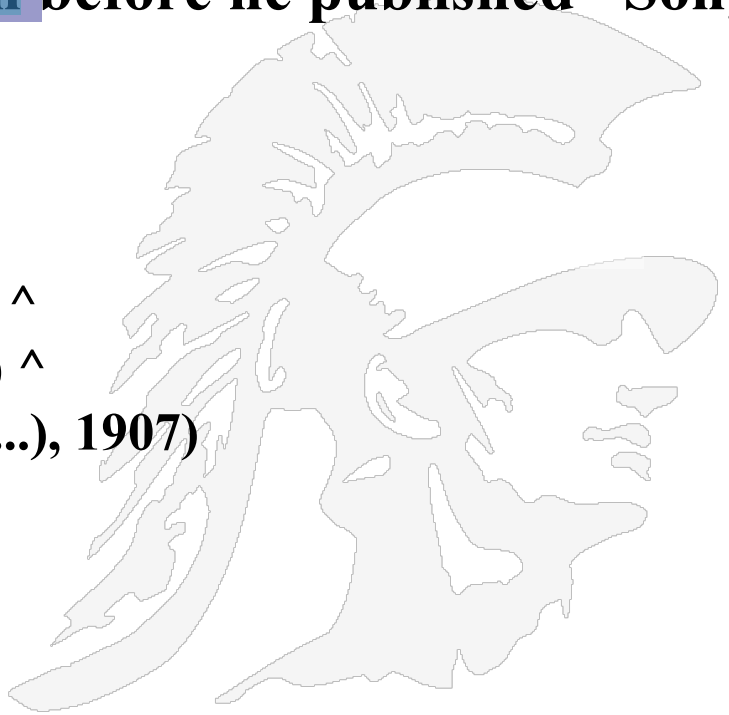
Category: Poets and Poetry: ***GEO***

He was a bank clerk in the Yukon before he published "Songs of a Sourdough" in 1907.

YEAR

Produces the logic formula:

authorof(focus,"Songs of a Sourdough") ^
publish (e1, he, "Songs of a Sourdough") ^
in (e2, e1, 1907) ^ temporallink(publish(...), 1907)



Extracting Candidate Answers from Triple Stores

- **Once we extract a relation from the question, e.g.**

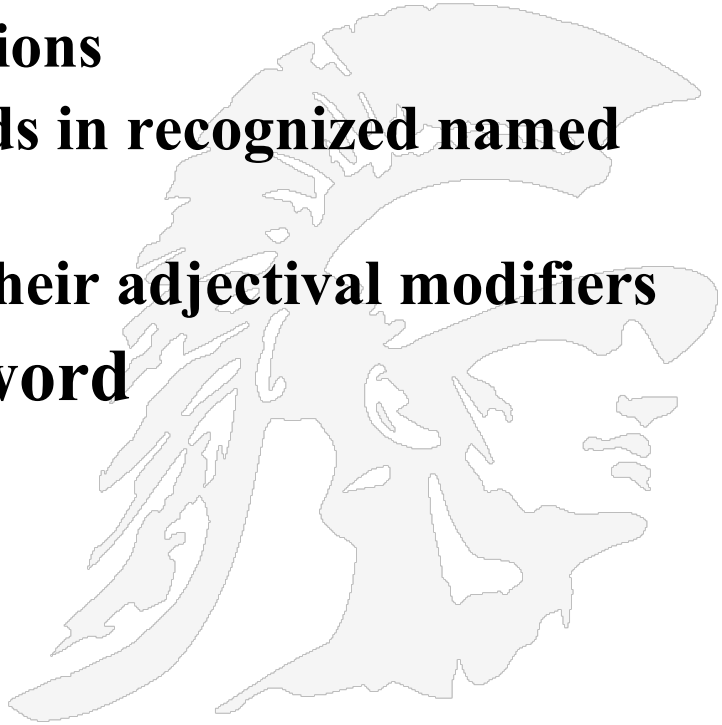
  … he published "Songs of a sourdough"

  `(author-of ?x "Songs of a sourdough")`

- **Many information sources support querying via a triple store**
  - **Wikipedia infoboxes, DBpedia, FreeBase, etc.**
  - **author-of("Songs of a Sourdough", "Robert Service")**

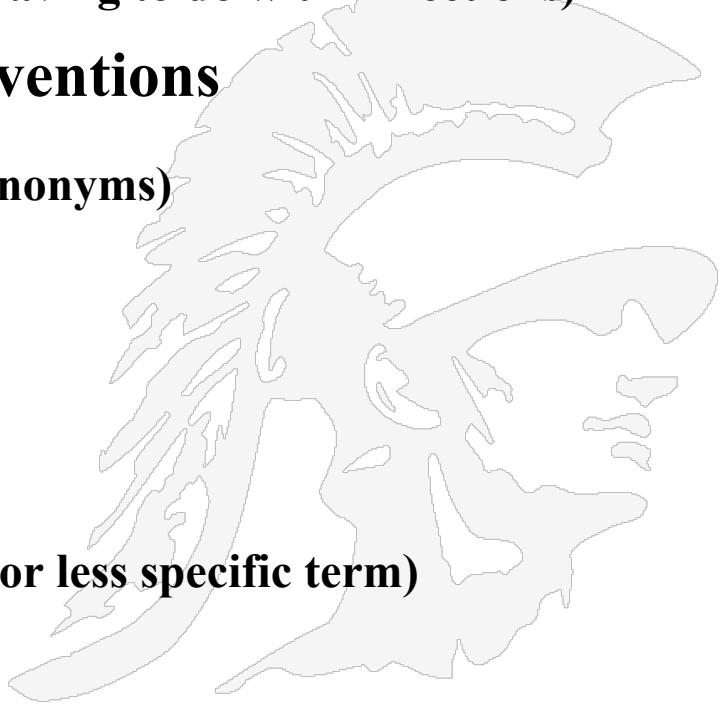
General Keyword Selection Algorithm

1. Use the part-of-speech recognizer to identify all
   - nouns
   - verbs
   - non-stopwords in quotations
   - NNP (proper noun) words in recognized named entities
   - complex nominals with their adjectival modifiers
2. Select the answer type word



Expanding the Keyword Set Using Variants

- There are 3 distinct ways to expand the keyword set determined by the keyword selection algorithm
- Morphological variants (having to do with inflections)
  - invented → inventor → inventions
- Lexical variants (similar to synonyms)
  - killer → assassin
  - far → distance
- Semantic variants
  - like → prefer (a more specific or less specific term)



How to Incorporate Lexical Variants Using Hypernims 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
Hypernim is a superordinate grouping which includes subordinate groups
e.g. a musical instrument is a hypernim of guitar;
Hyponim is a word or phrase whose semantics is more specific than its hypernim
e.g. purple is a hyponim of color

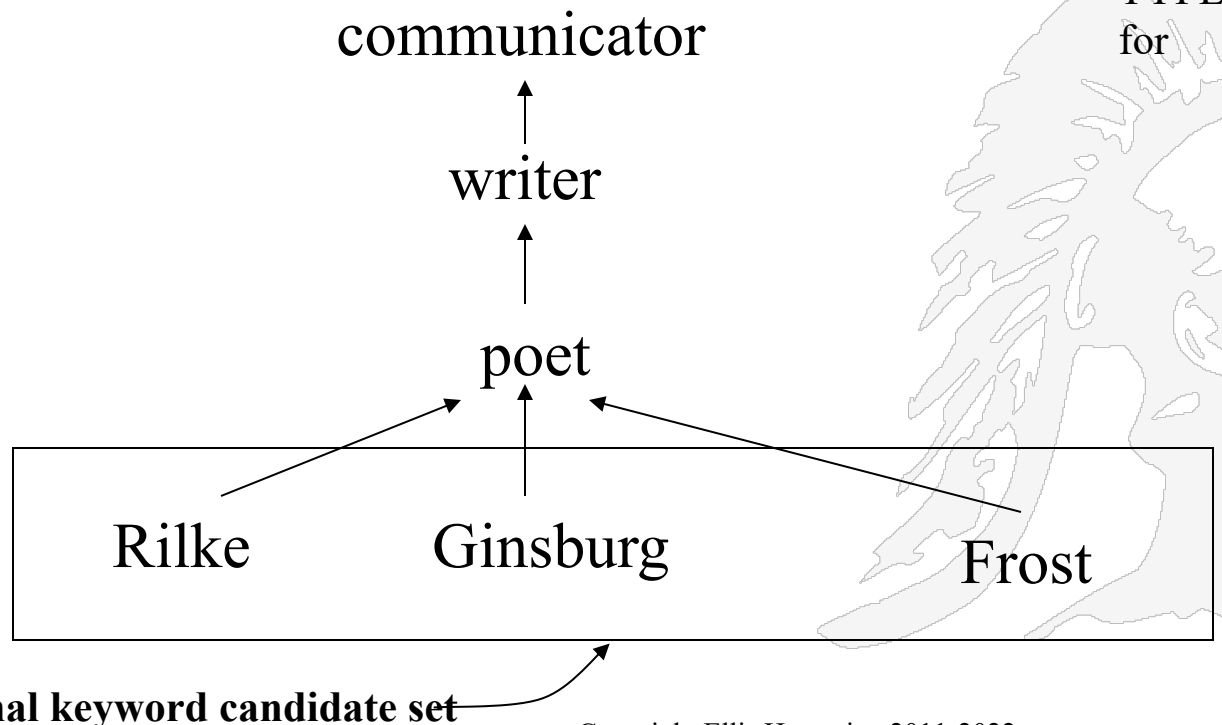
Use WordNet for Type Identification

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

"What 20th century poet wrote Howl?"

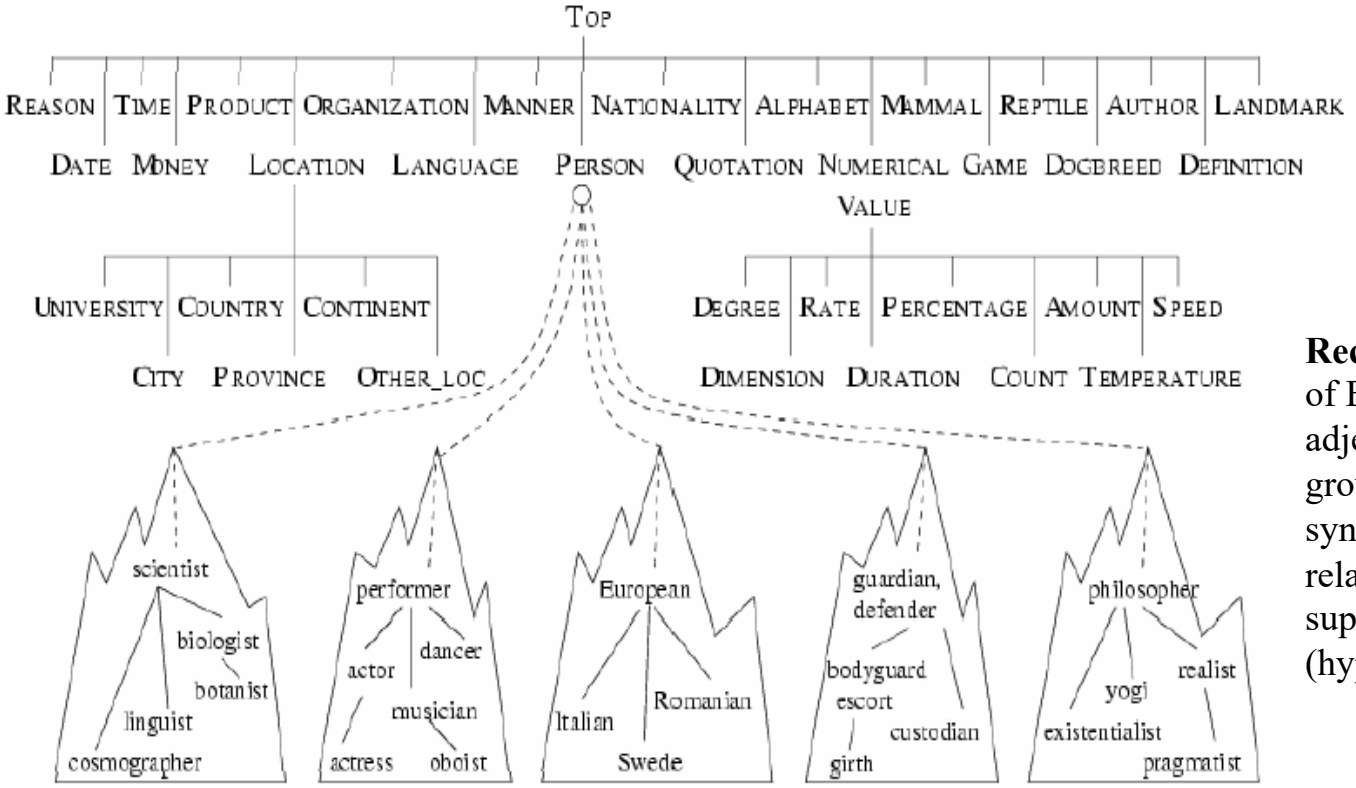
WordNet permits refinement of poet to specific instances

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



Answer Type Taxonomy

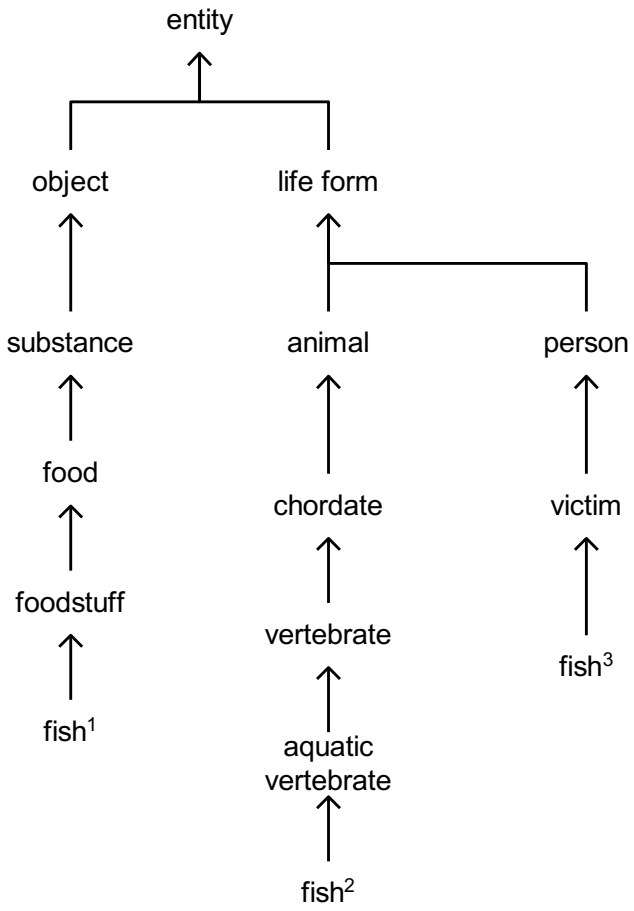
- **Use WordNet to merge named entities with the WordNet hierarchy**

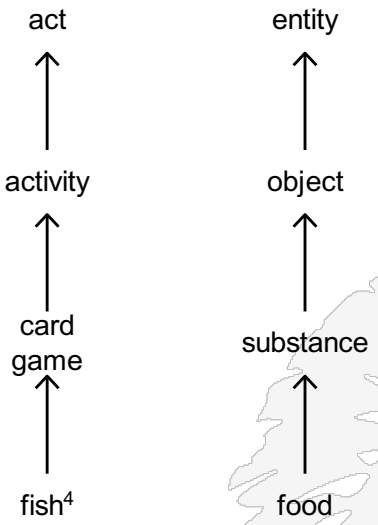


Recall:WordNet is a database of English nouns, verbs, adjectives and adverbs grouped into sets of synonyms (synsets) and relations showing super/subordinate relations (hyperonymy/hyponymy)

If you know the answer should be a person WordNet helps determine what sort of person

Another WordNet Example (1 of 2)





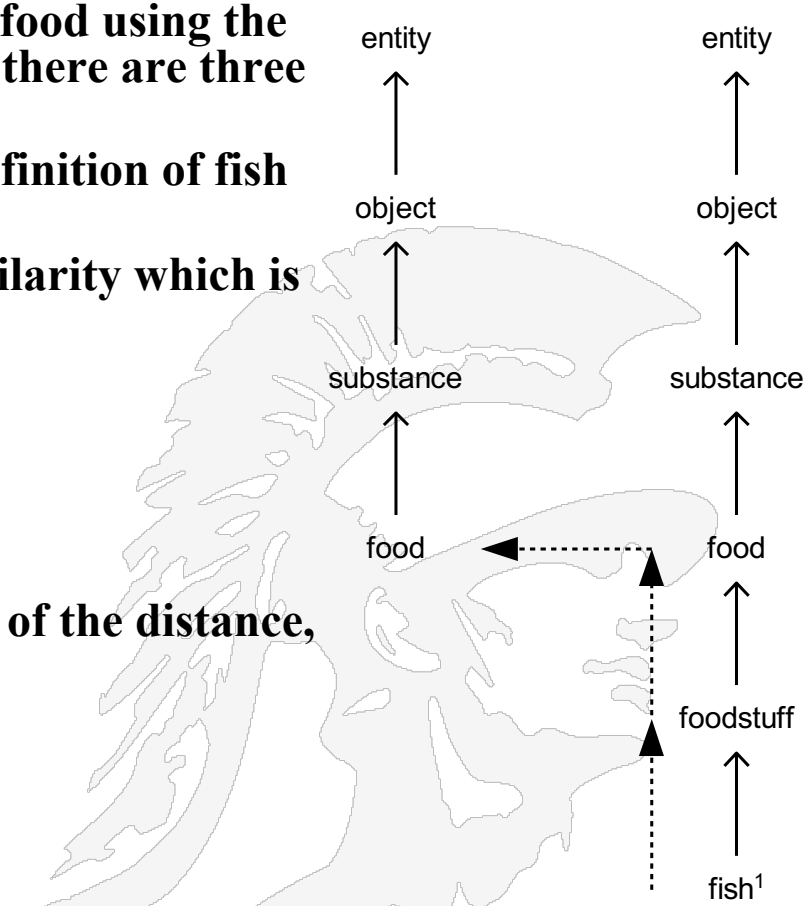
This uses the hypernym (… is a kind of …) relationships in WordNet to construct a path between two entities. For example these hypernym relationships for *fish* and *food* are present in WordNet

Another WordNet Example (2 of 2)

- work out all the paths between fish and food using the generated hypernym trees. It turns out there are three distinct paths.
- The shortest path is between the first definition of fish and the definition of food, as shown.
- One measure is Leacock-Chodorow similarity which is defined as:

$$\text{Semantic Similarity} = -\ln\left(\frac{d}{32}\right)$$

- Which would give a score of 2.37.
- Alternatively one can use the reciprocal of the distance, i.e. 1/3.



Part 2: Passage Retrieval

- **Once we have formulated queries using tools like NER, POS, variant expansion and WordNet results**
- **Send queries to a search engine and retrieve snippet results**
- **Filter the results for correct type**
  - use answer type classification
  - Rank passages based on a trained classifier (application of machine learning)
    - **Features:**
      - **Question keywords, Named Entities**
      - **Longest overlapping sequence,**
      - **Shortest keyword-covering span**
      - **N-gram overlap between question and passage**

Passage Scoring Method

- **Focus on the snippets that are returned, the answers must be extracted from them**
- **Passage ordering is performed using a sort that involves three scores:**
  1. The number of words from the question that are recognized and in the same sequence in the snippet window
  2. The number of words that separate the most distant keywords in the snippet window
  3. The number of unmatched keywords in the snippet window

Ranking Candidate Answers

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

- Answer type: Person
- Text passage:

"Among them was Christa McAuliffe, the first private citizen to fly in space. Karen Allen, best known for her starring role in "Raiders of the Lost Ark", plays McAuliffe. Brian Kerwin is featured as shuttle pilot Mike Smith..."

Scoring

There are five words from the question "the first private citizen space"
The answer is adjacent to "the first private citizen. . . "
There are no unmatched keywords in "the first private citizen. . . "

Ranking Candidate Answers

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

- Answer type: Person
- Text passage:

"Among them was Christa McAuliffe, the first private citizen to fly in space. Karen Allen, best known for her starring role in "Raiders of the Lost Ark", plays McAuliffe. Brian Kerwin is featured as shuttle pilot Mike Smith..."

- Best candidate answer: Christa McAuliffe

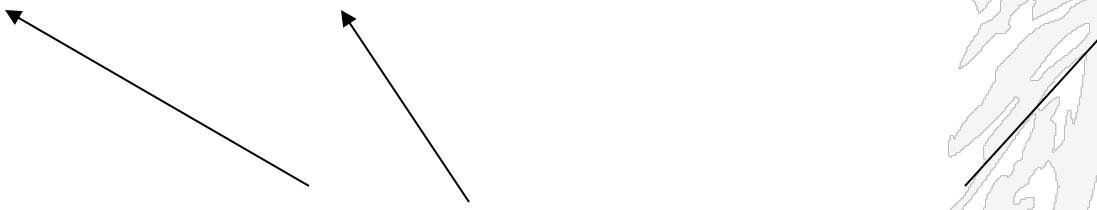
Comments on Scoring this Passage

- Karen Allen is rejected as an answer as it does not occur in the sentence "the first private citizen. . . "
- Brian Kerwin is rejected as the name is far away from "the first private citizen. . . "

Local Alignment Example (1 of 7)

Who shot Kennedy?

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



Three Potential Candidates by type

WHO indicates we are looking for a PERSON

Local Alignment Example (2 of 7)

Question Head word → ***Who shot Kennedy?***

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

"shot Kennedy"
gives us a verb and matching context

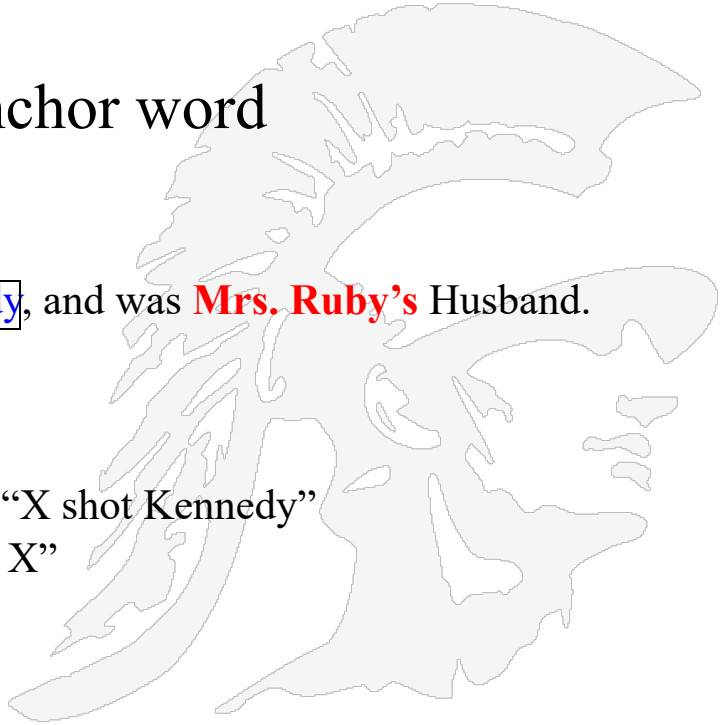
Local Alignment Example (3 of 7)

Who shot Kennedy?

Anchor word

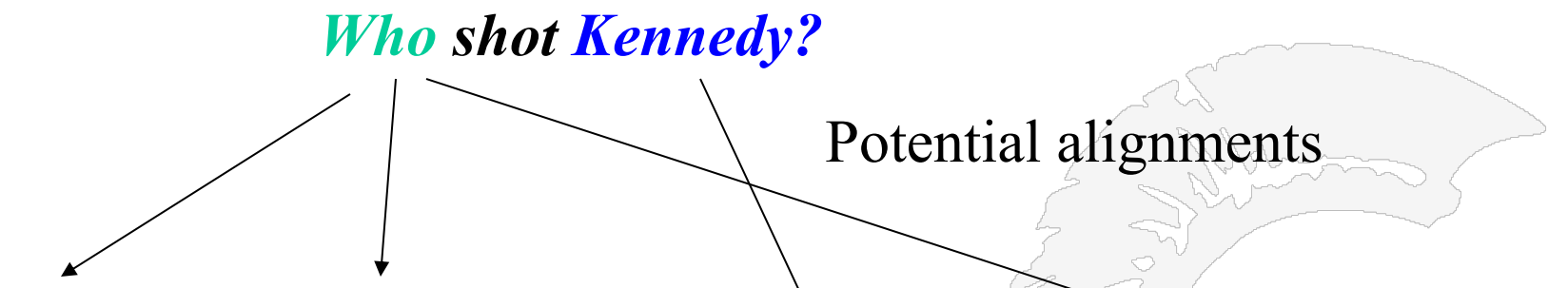
Jack assassinated **Oswald**, the man who shot Kennedy, and was **Mrs. Ruby's** Husband.

Look for phrases such as "X shot Kennedy"
or "Kennedy was shot by X"

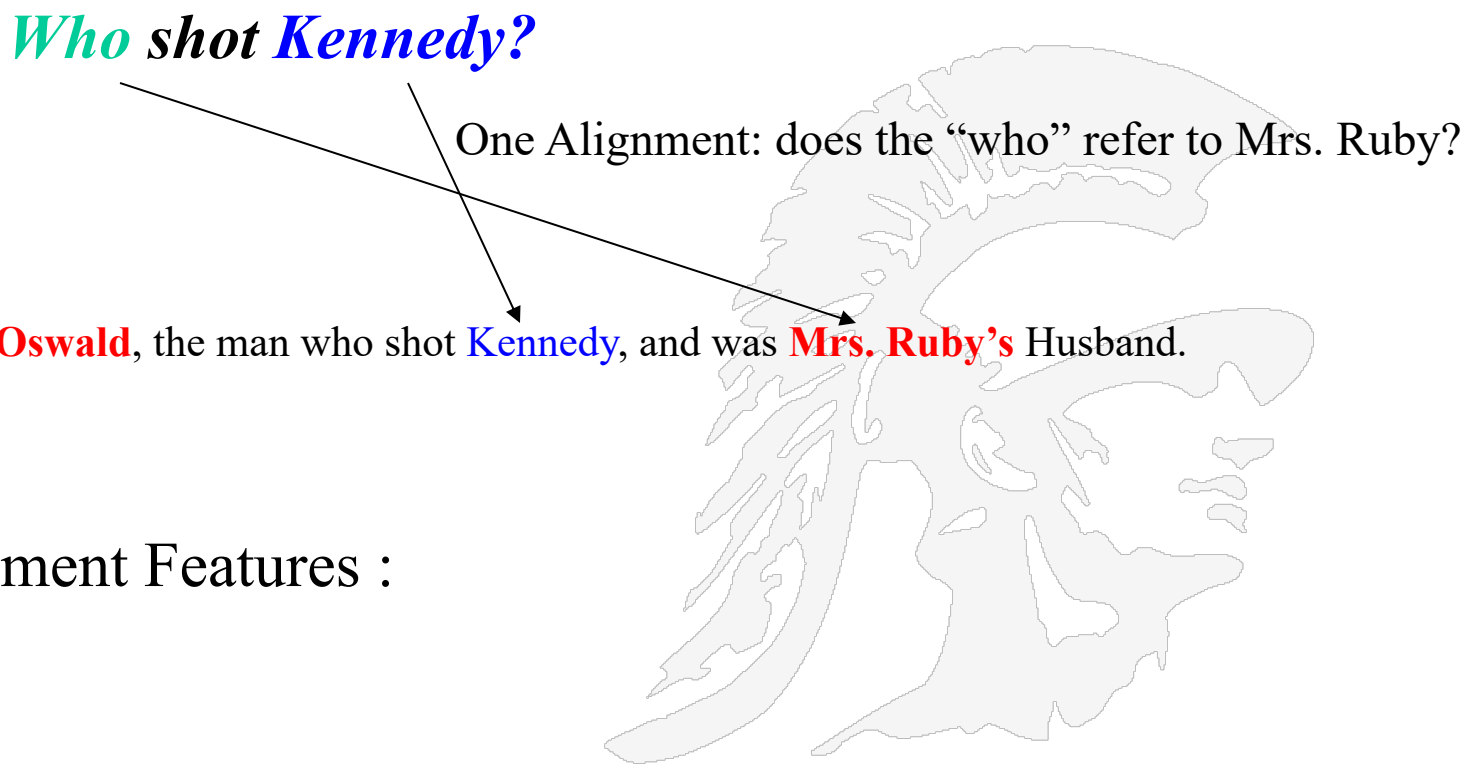


Local Alignment (4 of 7)

In principle it can be anyone of the three people identified

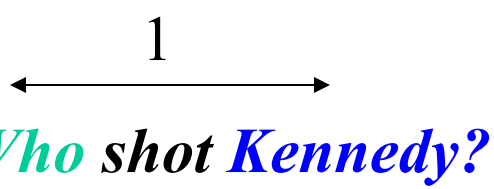


Local Alignment Example (5 of 7)



Three Alignment Features :

Local Alignment Example (6 of 7)



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

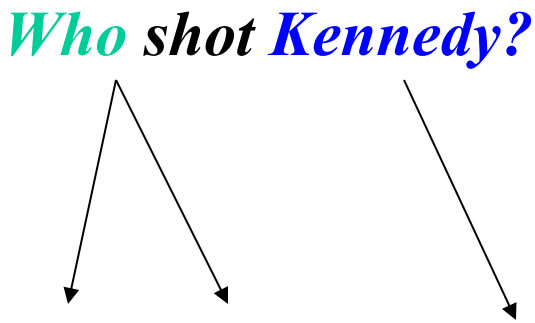
Jack assassinated **Oswald**, the man who shot Kennedy, and was **Mrs. Ruby's** Husband.

2

Three Alignment Features :

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

Local Alignment Example (7 of 7)



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**

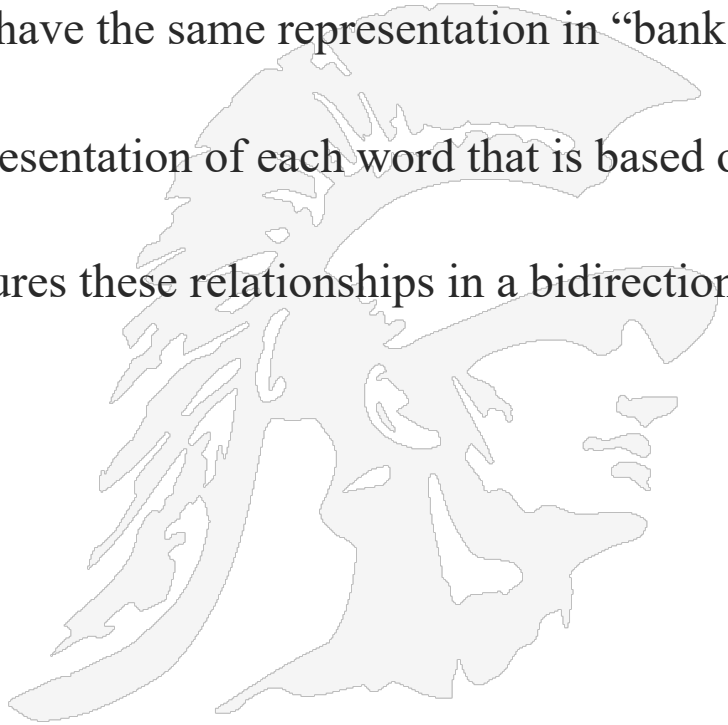
1. The number of named entities of the right type in the passage
2. The number of question keywords in the passage
3. The longest exact sequence of question keywords that occurs in the passage
4. The rank of the document from which the passage was extracted
5. The proximity of the keywords from the original 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

- ***B*idirectional *E*ncoder *R*epresentations from *T*ransformers**
- **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 its being used in practical applications**
- **the BERT Github repository is available on github,**
  git clone https://github.com/google-research/bert.git

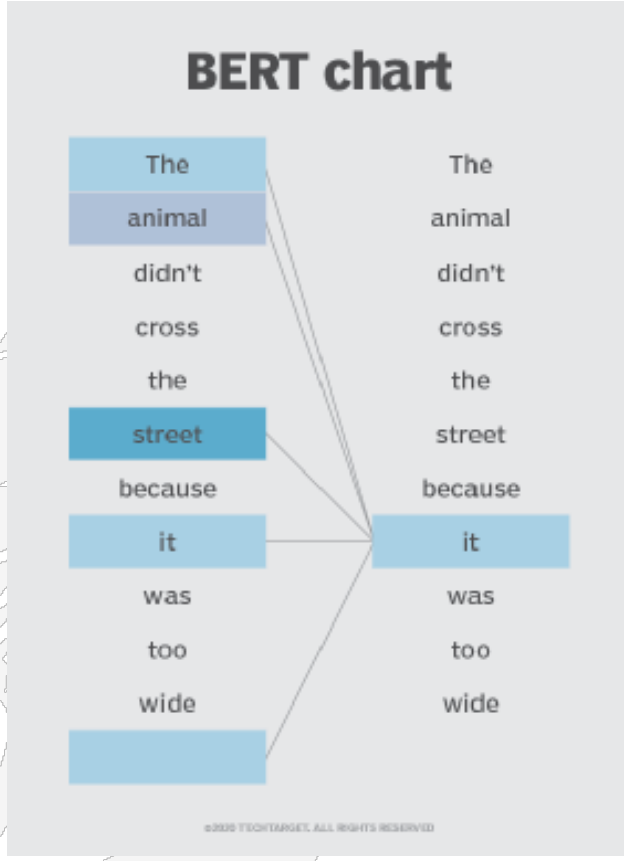
Why do we need BERT

- Proper language representation is key for general-purpose language understanding by machines.
- ***Context-free models*** such as word2vec or GloVe generate a single word embedding representation for each word in the vocabulary.
  - For example, the word "bank" would have the same representation in "bank deposit" and in "riverbank".
- ***Contextual models*** instead generate a representation of each word that is based on the other words in the sentence.
  - BERT is a contextual model that captures these relationships in a bidirectional way.



How BERT works

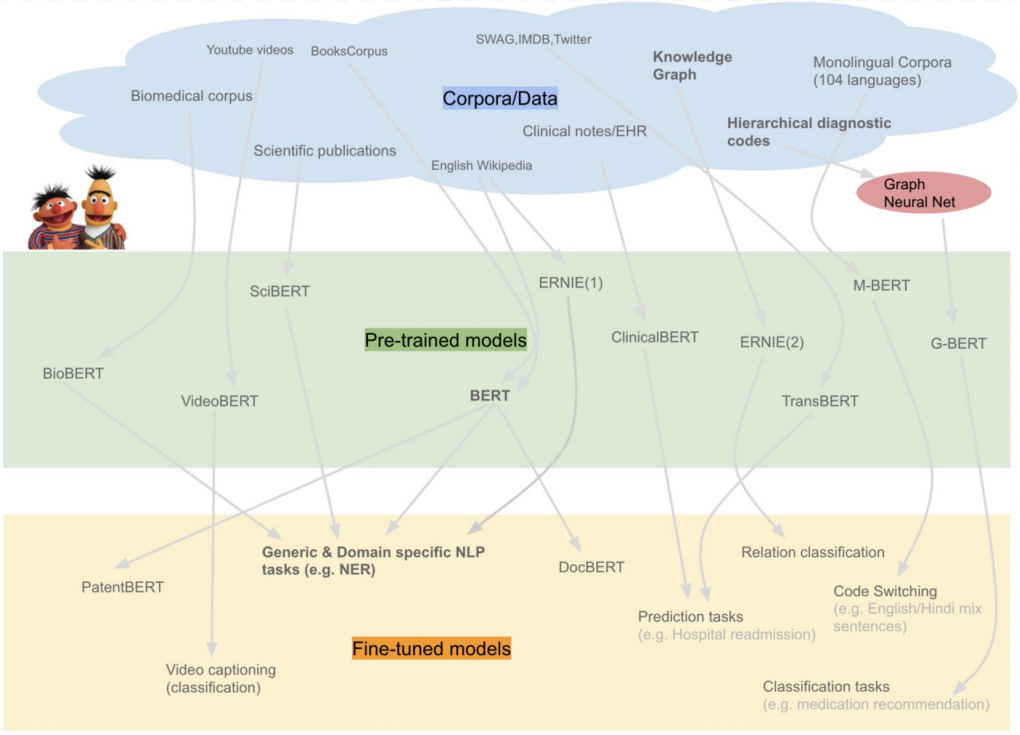
- BERT functions by reading ***bidirectionally***, accounting for the effect of all other words in a sentence on the focus word and eliminating the left-to-right momentum that biases words towards a certain meaning as a sentence progresses.
- At the right BERT is determining which prior word in the sentence 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.



https://www.techtarget.com/searchenterpriseai/definition/BERT-language-model
https://csci572.com/papers/BERT.pdf (the original paper)

BERT Has Many Pre-Trained Models

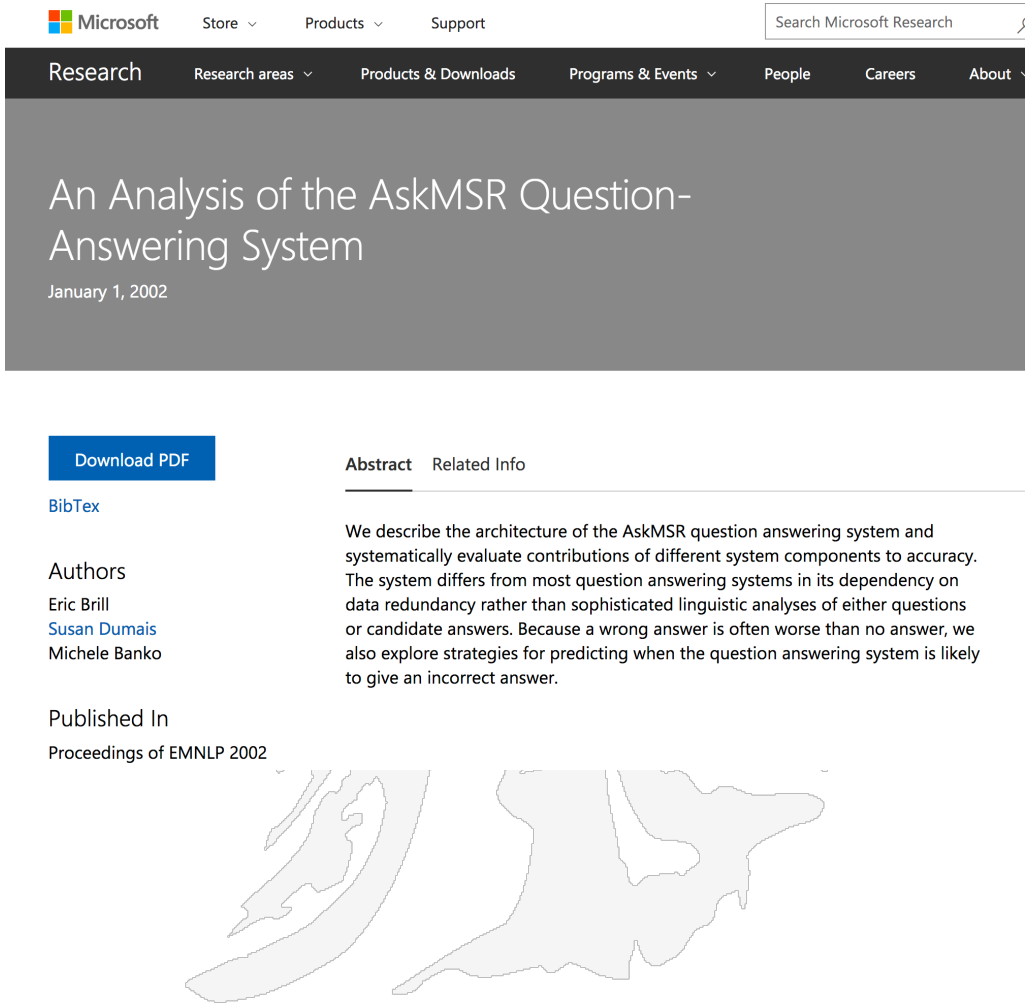
- BERT was trained on Wikipedia and other datasets
- To the right you can see a diagram of additional variants of BERT pre-trained on specialized corpora
- Here are 4 examples of BERT on pre-trained data:
- patentBERT - a BERT model fine-tuned to perform patent classification.
- docBERT - a BERT model fine-tuned for document classification.
- bioBERT - a pre-trained biomedical language representation model for biomedical text mining.
- VideoBERT - a joint visual-linguistic model for process unsupervised learning of an abundance of unlabeled data on Youtube.



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

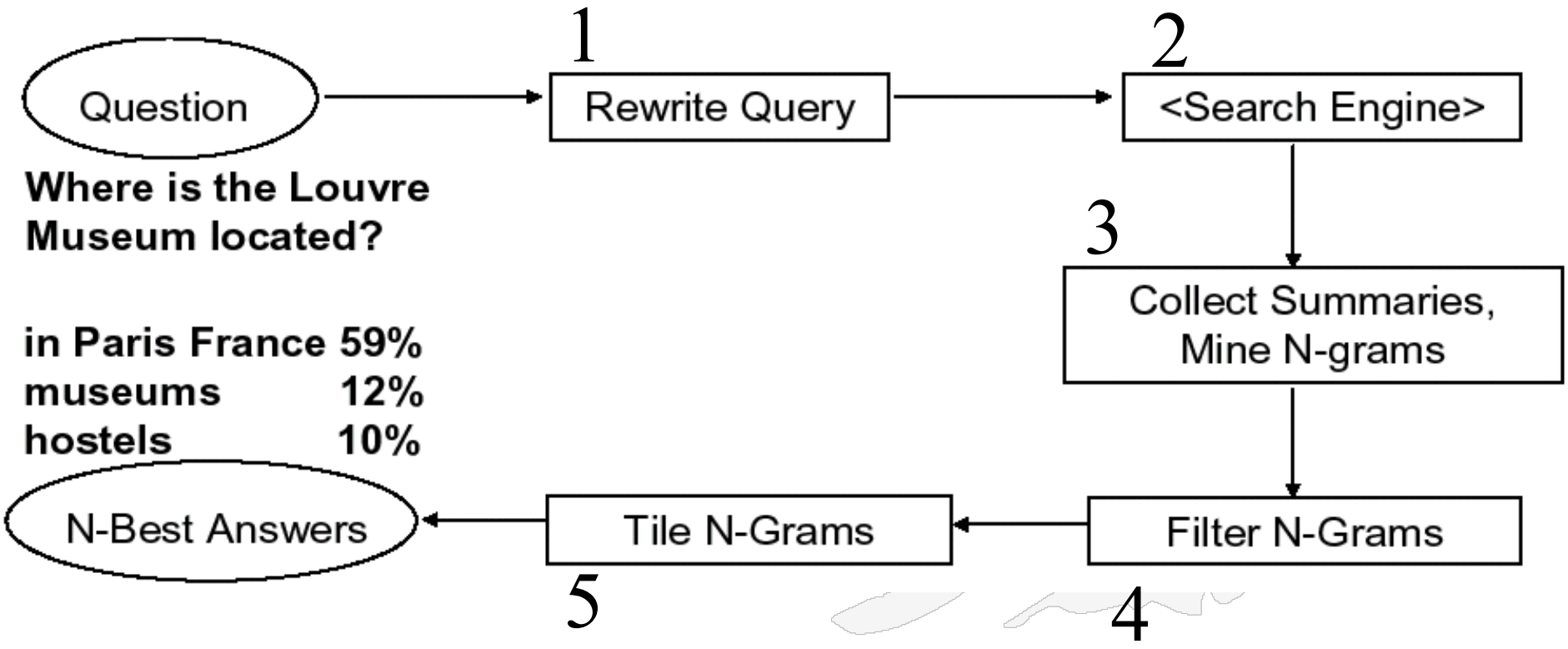
Microsoft's AskMSR Answering System

- 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

AskMSR: Details



AskMSR: Step 1:Query Rewriting

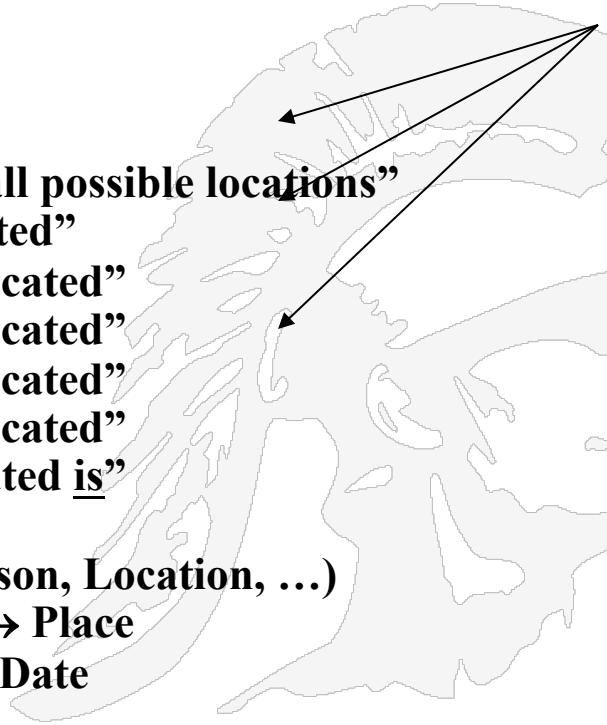
- **Classify question into categories**
  - **Who is/was/are/were…?**
  - **When is/did/will/are/were …?**
  - **Where is/are/were …?**

a. Category-specific transformation rules

eg "For Where questions, move 'is' to all possible locations"

"Where is the Louvre Museum located"

- → **"is the Louvre Museum located"**
- → **"the is Louvre Museum located"**
- → **"the Louvre is Museum located"**
- → **"the Louvre Museum is located"**
- → **"the Louvre Museum located is"**



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

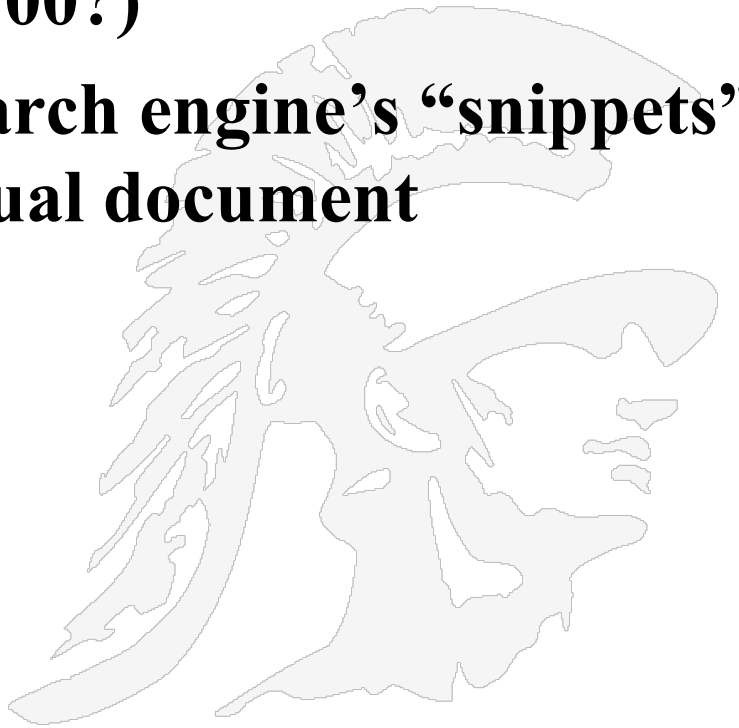
b. Expected answer "Datatype" (eg, Date, Person, Location, …)

Where is the Louvre Museum located → Place

When was the French Revolution? → Date

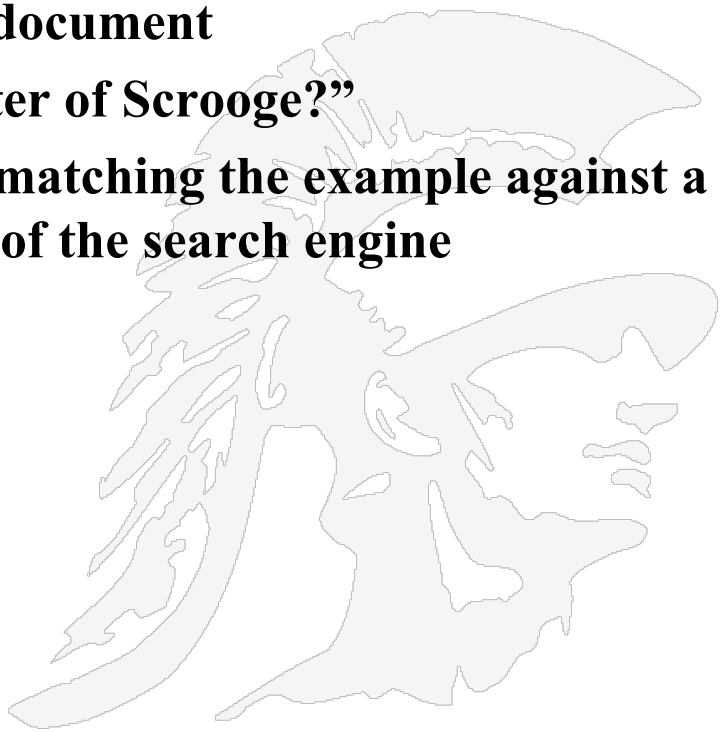
AskMSR: Step 2: Query Search Engine

- Send all rewrites to a Web search engine
- Retrieve top N answers (100?)
- For speed, rely just on search engine's "snippets", not the full text of the actual document



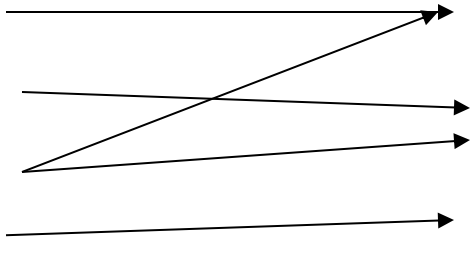
AskMSR: Step 3:Mining N-Grams

- **Simple: Enumerate all N-grams (N=1,2,3 say) in all retrieved snippets**
  - Use hash table and other data structures to make this efficient
- **Weight of an n-gram: occurrence count, each weighted by "reliability" (weight) of rewrite that fetched the document**
- **Example: "Who created the character of Scrooge?"**
- **Below are the weights produced by matching the example against a set of N-grams in the N-gram database of the search engine**
  - Dickens - 117
  - Christmas Carol - 78
  - Charles Dickens - 75
  - Disney - 72
  - Carl Banks - 54
  - A Christmas - 41
  - Christmas Carol - 45
  - Uncle - 31



AskMSR: Step 4: Filtering N-Grams

- Each question type is associated with one or more "data-type filters" = regular expression
- When…
- Where…
- What …
- Who …



Date

Location

Person

- Boost score of n-grams that do match the regular expression
- Lower score of n-grams that don't match the regular expression

AskMSR: 5: Tiling the Answers