Part II

A Brief History of Open-Domain Question Answering

Natural language understanding: early QA systems

- Question-answering machine [Simmons, 1965]
  - General-purpose language processors that communicate with users in natural language (e.g., English)
  - Deal with statements and/or questions

http://csunplugged.org/turing-test

Simmons, 1965. Answering English Questions by Computer: a Survey

Categories of (early) QA systems

- **List-structured database systems**
  - Organizing knowledge (e.g., kinship) in list DB
- **Graphic database systems**
  - Map text and graphic data (e.g., pictures, diagrams) to the same logical representations
- **Text-based systems**
  - Matching questions and text in a corpus to find answers
- **Logical inference systems**
  - Textual entailment, answering science text book questions & algebra word problems

Simmons, 1965. Answering English Questions by Computer: a Survey

Baseball [Green et al., 1963]

Q: How many games did the Yankees play in July?

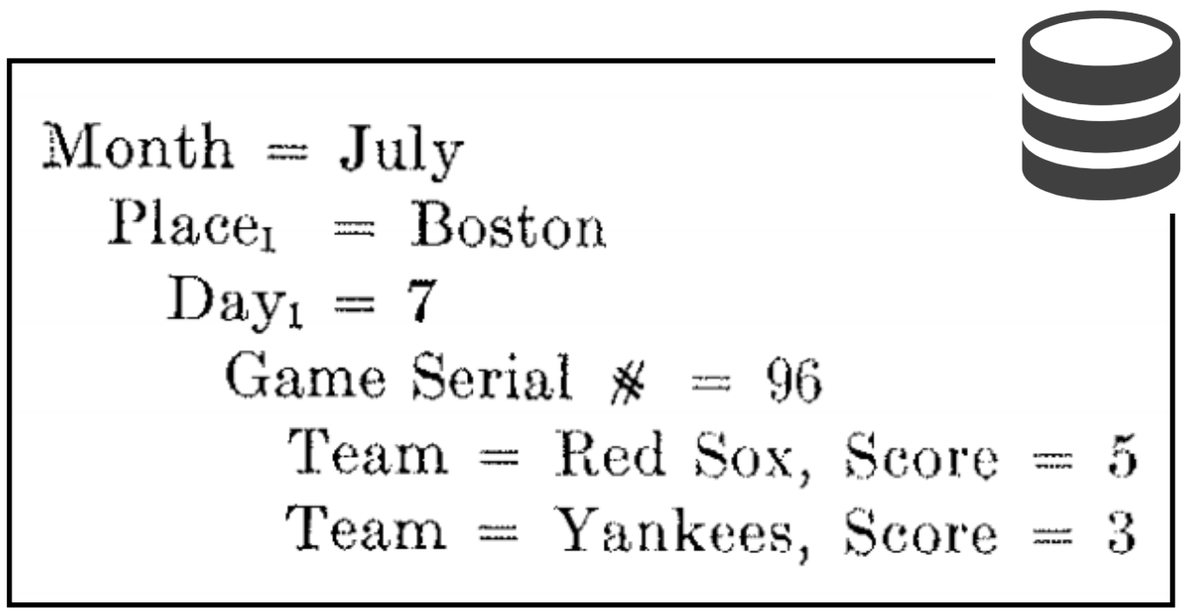
- Step 1: Simple dictionary-based syntactic analysis

  (How many games) did (the Yankees) play (in (July))?

- Step 2: Semantic analysis that builds "spec"

  "Who" ⟶ ("team" = ?)

  Conditions (e.g., "winning", "how many") ⟶ routines



```
Month = July
   Place1  = Boston
     Day1  = 7
        Game Serial #  = 96
           Team = Red Sox, Score = 5
           Team = Yankees, Score = 3
```

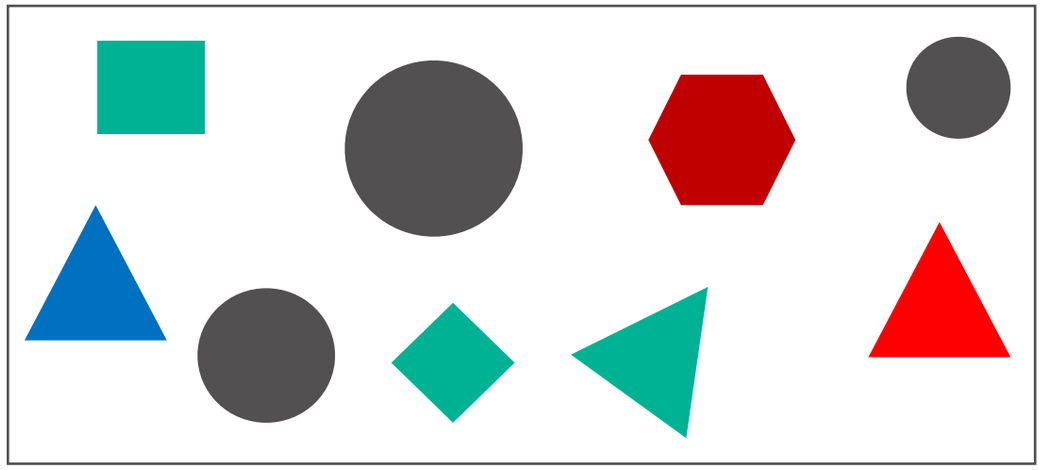
- Step 3: Execution

Green et al.,1963. Baseball: An Automatic Question-Answerer

The Picture Language Machine [Krisch, 1964]

Is the statement true?

"All circles are black circles."



Both pictures and text are translated into logical language:

- Circle(a), Black(a), Bigger(a, b), Between(a, b, c)
- $(\forall x)[\text{Circle}(x) \supset (\exists y)[\text{Circle}(y) \wedge \text{Black}(y) \wedge (x = y)]]$

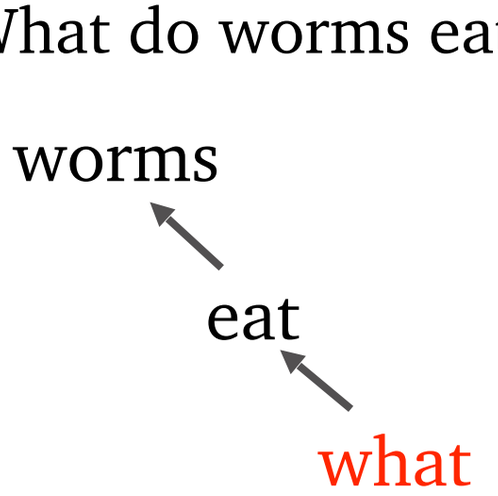
Krisch, 1964. Computer Interpretation of English Text and Picture Patterns

Protosynthex [Simmons+ 1963]

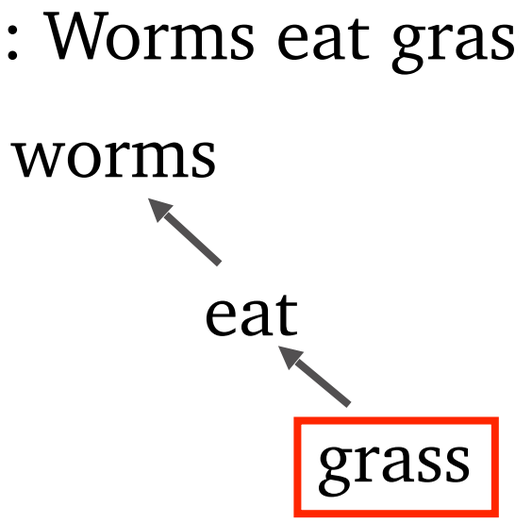
Answer questions from an Encyclopedia

- Matching questions & text in dependency logic [Hays, 1962]

Q: What do worms eat?

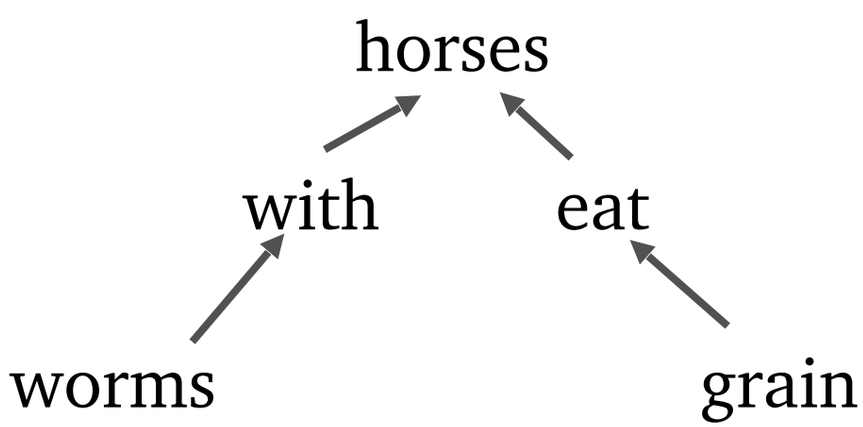


A1: Worms eat grass



Complete Agreement

A2: Horses with worms eat grain



Partial Agreement

Student [Bobrow, 1964]

The first algebra problem solver

- Translate a set of English statements to mathematical equations

> *If the number of customers Tom gets is twice the square of 20% of the number of advertisements he runs, and the number of advertisements is 45, then what is the number of customers Tom gets?*

- Step 1: Simplify text and annotate operators

  “twice” →“two times”, “the square of” →“square”

  Tag operators like “plus”, “percent”, “times”

- Step 2: Heuristics to break problem into simple sentences

- Step 3: Mapping sentences to equations

  Rules based on dictionary of words and numbers

Bobrow, 1964. Natural Language Input for a Computer Problem Solving System

Lessons from old QA systems

Limited success

- Small & limited domains and scopes
  - Often work only on well-controlled, specialized subset of English
- Not data-driven (e.g., machine learning approaches)
  - Mostly rule-based, potentially brittle
- Lacks rigorous evaluation

Open questions [Simmons, 1965]

- Meaning representation & the need of formal languages
- Syntactic and semantic disambiguation
- Combine partial answers from various sources

Simmons, 1965. Answering English Questions by Computer: a Survey

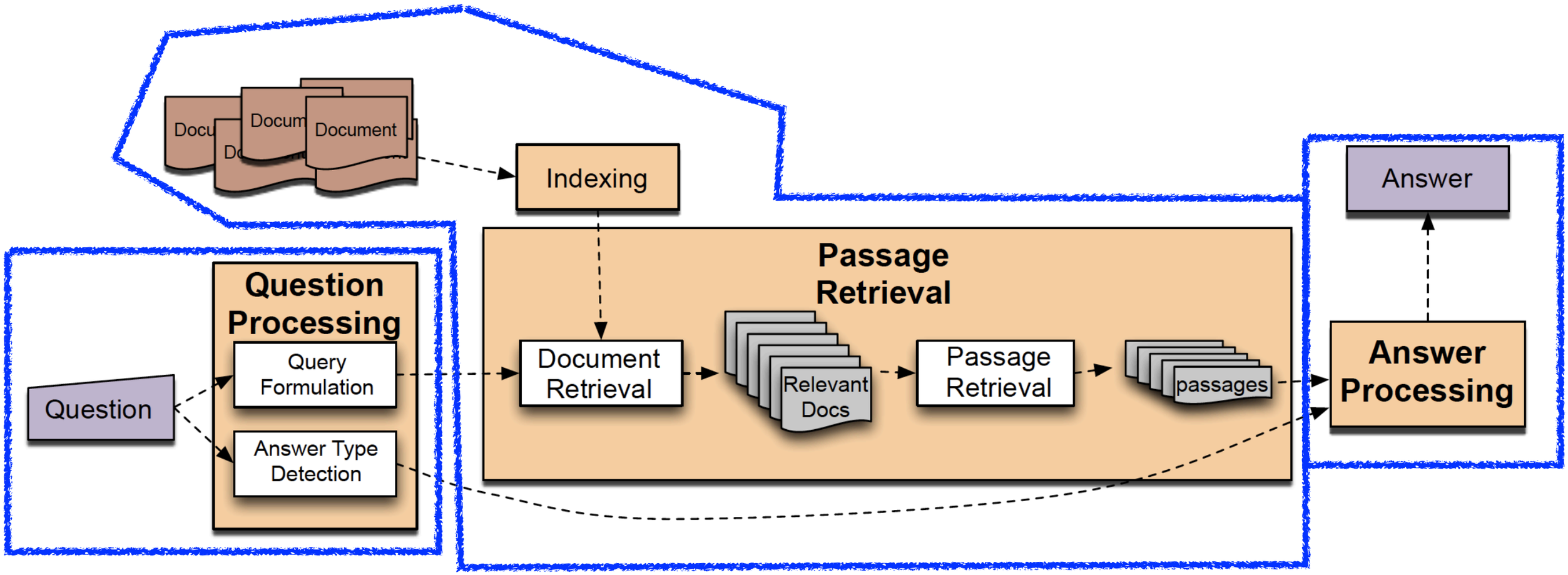
Text Retrieval Conference (TREC)
QA Tracks (1999 - 2007)

https://trec.nist.gov/

- Originating from the IR community as the next version of search
  - Relevant documents → short answer with support

- Shared tasks & competitions
  - Corpus: newswire (AP, WSJ, LA Times, etc.); 979k articles, 3GB
  - Test set: 500 questions from Excite, Encarta, MSNSearch, AskJeeves
  - Human judges decide the correctness of the answers from QA systems

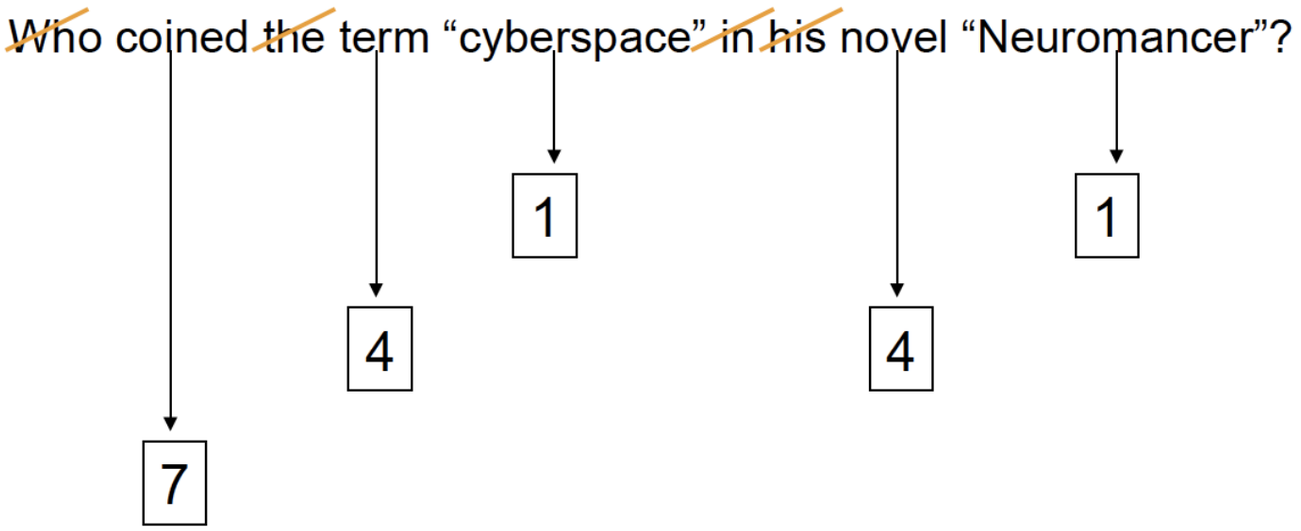
Typical pipeline of TREC-QA systems

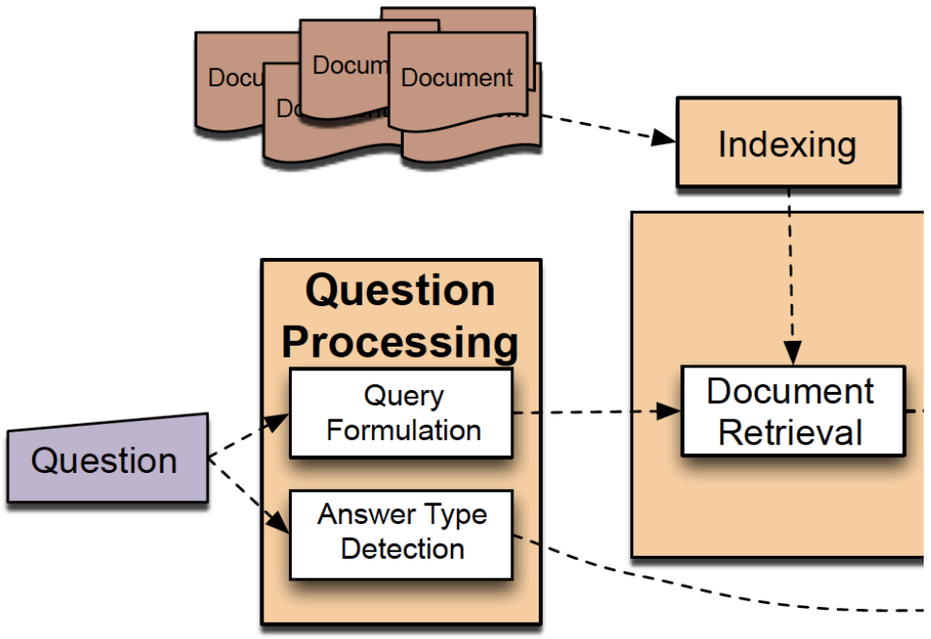


Jurafsky and Martin. Speech and Language Processing. 3rd edition.

Query formulation

Choosing keywords from the question:



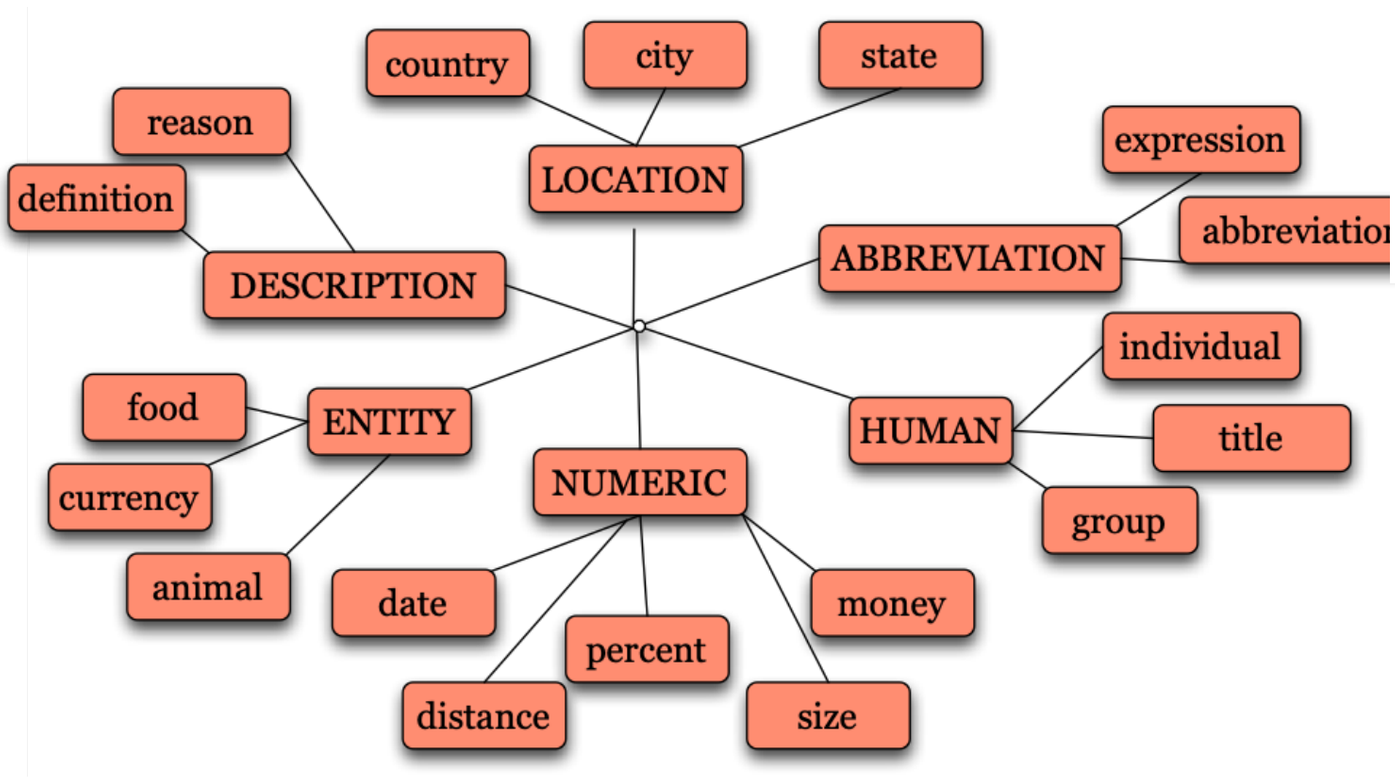


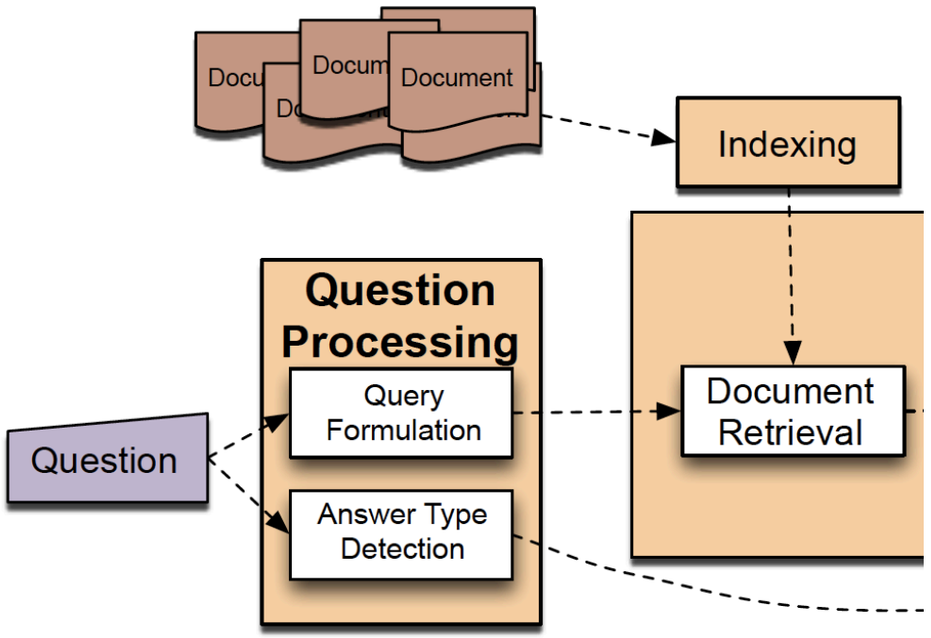
Example from Mihai Surdeanu

Answer type detection

Answer type taxonomy [Li & Roth, 2002]

- 6 coarse classes, 50 fine classes





Li and Roth, 2002. Learning Question Classifiers

Example questions

Answer type: entity

Type	Questions
ENTY:animal	What was the first domesticated bird?
ENTY:letter	What's the second-most-used vowel in English?
ENTY:food	What rum is so "mixable" it is a one-brand bar?
ENTY:color	What's the only color Johnny Cash wears on stage?
ENTY:product	Which two products use a tiger as their symbol?
ENTY:religion	In what religion was Isis the nature goddess?
ENTY:other	What does a spermologer collect?

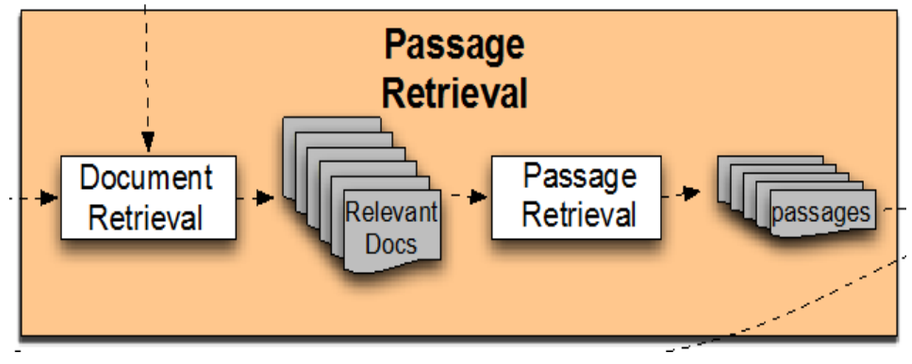
Example questions

Answer type: human, location, numeric

Type	Questions
HUM:ind	What crooner joined The Andrews Sisters for Pistol Packin Mama?
HUM:gr	Who has won the most Super Bowls?
LOC:city	What city did the Flintstones live in?
LOC:other	What stadium do the Miami Dolphins play their home games in?
LOC:state	What U.S. state lived under six flags?
NUM:date	When was Ozzy Osbourne born?
NUM:count	How many people in the world speak French?

Passage retrieval

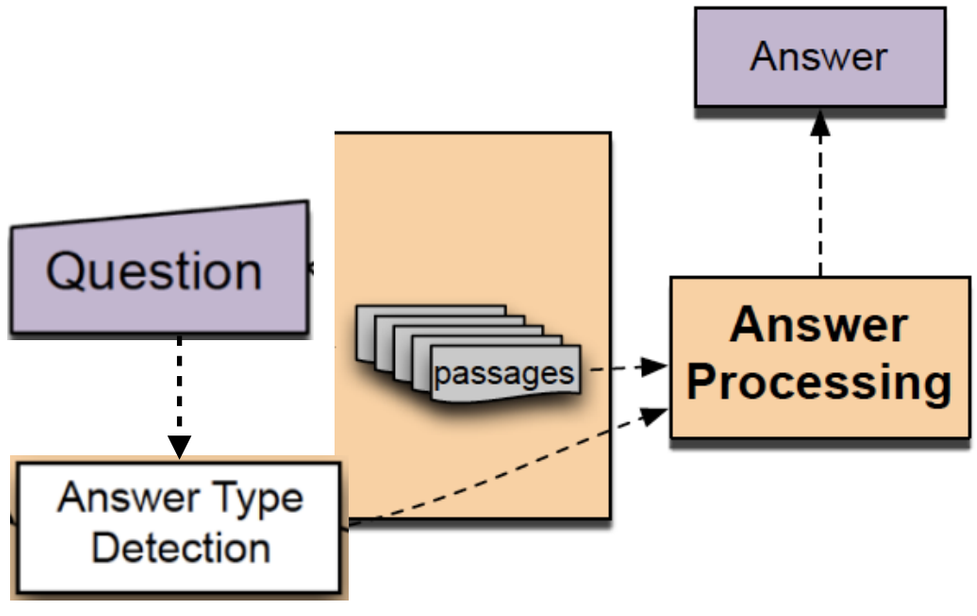
- Document retrieval via standard information retrieval methods
- Retrieved documents segmented into shorter units as paragraphs
  - Only a small chunk of text is assumed relevant
  - Answer extraction/processing is more computation intensive
- Passage ranking/selection
  - Linear ML models based on hand-crafted features
  - Example features
    - Number of Named Entities of the right type in passage
    - Number of query words in passage
    - Rank of the document containing passage



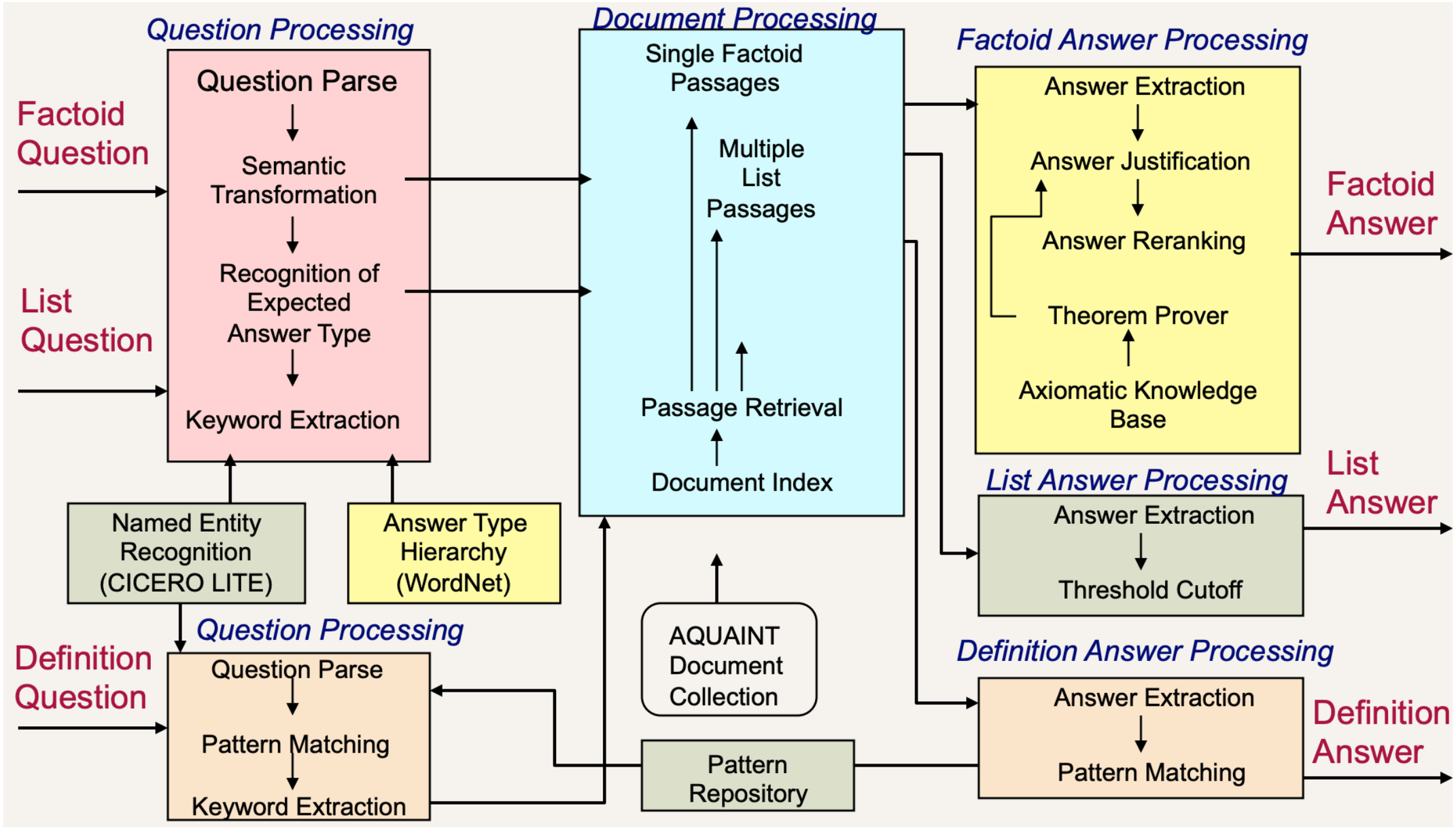
Answer extraction/processing

- Typically another classifier with hand-crafted features plus heuristics
- Run an answer-type named-entity tagger on the passages
  - Each answer type requires a named-entity tagger that detects it
  - If answer type is CITY, tagger has to tag CITY
  - Return the string with the right type

What is the closest city to Mount Rushmore? (City)
Mount Rushmore is located in the southwest corner of South Dakota, just east of the Wyoming border. The closest airport is in Rapid City, which is about 33 miles northeast of Mount Rushmore. The nearest hotels are in Keystone, SD, which is just outside the entrance to Mount Rushmore.



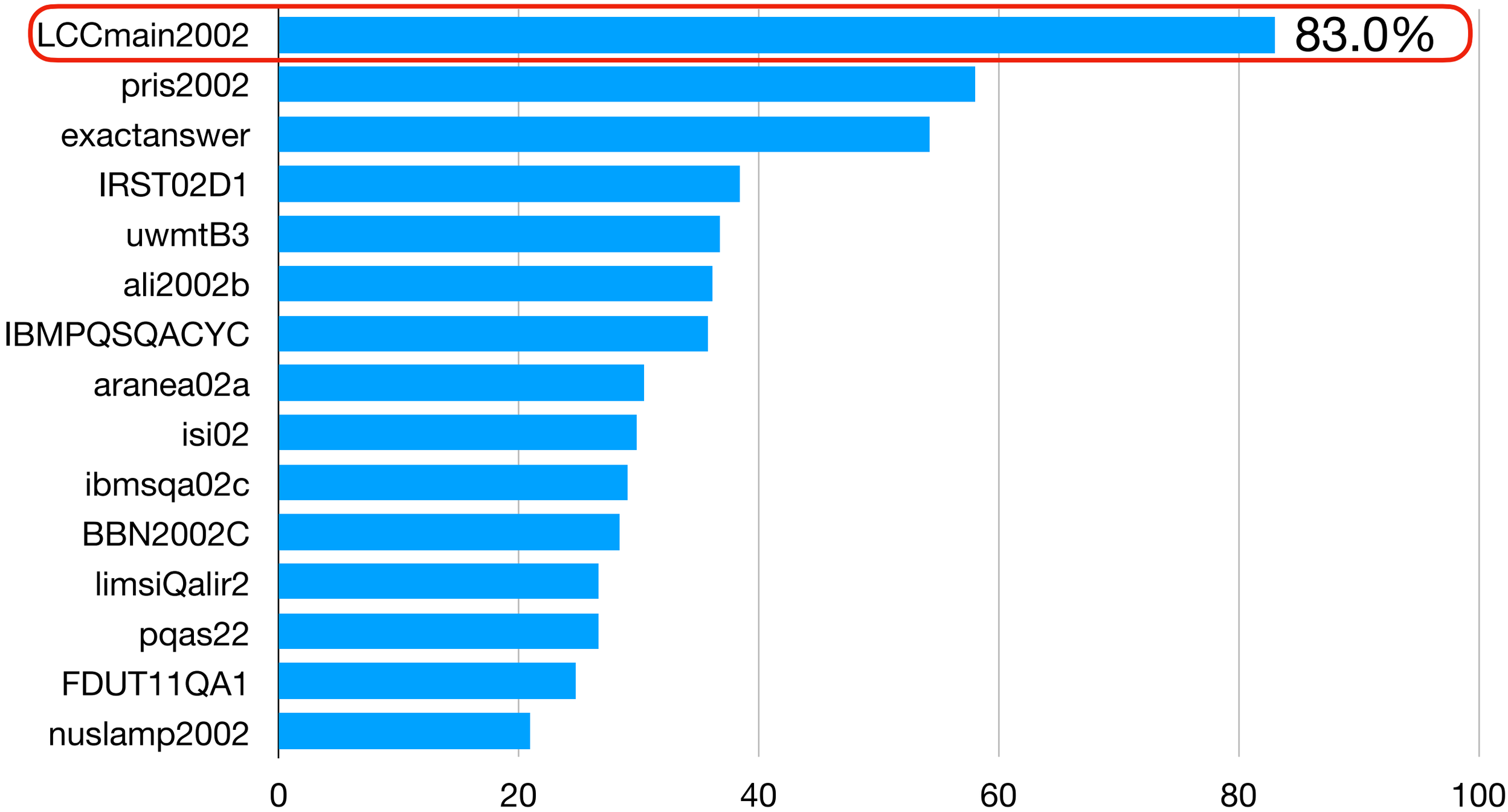
Top TREC-QA system (circa 2003): LCC



Harabagiu and Moldovan, 2003. Question Answering.

https://web.stanford.edu/class/archive/cs/cs224n/cs224n.1086/handouts/cs224n-QA-2008-1up.pdf

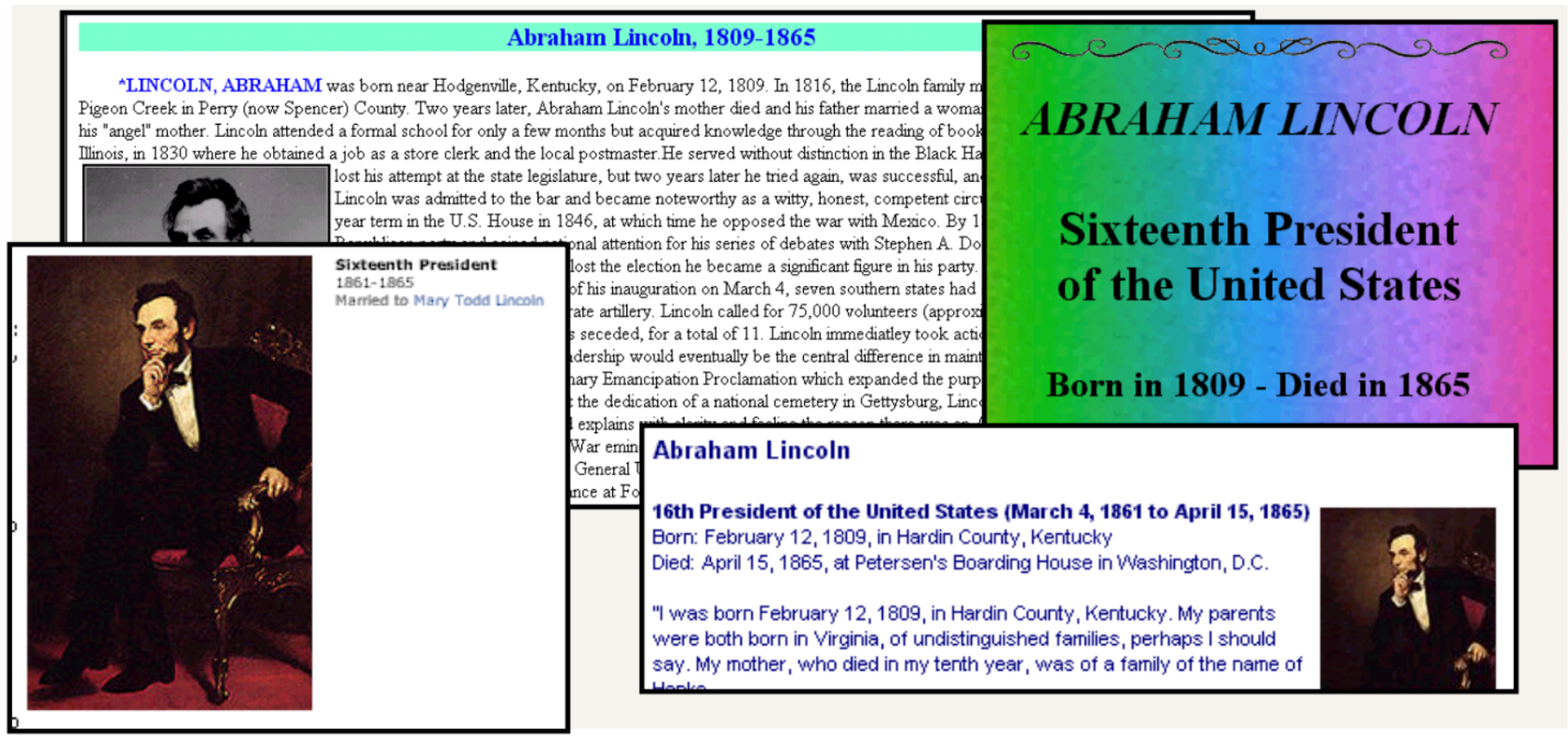
TREC-2002 QA main track results



https://trec.nist.gov/pubs/trec11/papers/QA11.pdf

AskMSR: data-intensive QA

- *"In what year did Abraham Lincoln die?"*
- Leverage "Web Redundancy"
  - Many mentions of the fact on the Web
  - Use patterns to find the "easy" ones



Brill et al., 2002. An Analysis of the AskMSR Question-Answering System
https://web.stanford.edu/class/archive/cs/cs224n/cs224n.1086/handouts/cs224n-QA-2008-1up.pdf

AskMSR: data-intensive QA

- Query patterns of *"In what year did Abraham Lincoln die?"*

  "Abraham Lincoln died in XXXX"

  "Abraham Lincoln (YYYY -- XXXX)"
- Use the most frequent *n*-gram in the documents as answer

- **Observations**
  - Search engine as language model
  - Pattern formulation is no longer needed [Tsai et al., 2015]
  - Works great for head questions, but poorly on tails
  - Cannot provide support evidence (e.g., excerpt, snippet) reliably

Brill et al., 2002. An Analysis of the AskMSR Question-Answering System

Tsai et al., 2015. Web-based Question Answering: Revisiting AskMSR

IBM Watson

The DeepQA Project

IBM Watson defeated two of Jeopardy's greatest champions in 2011.

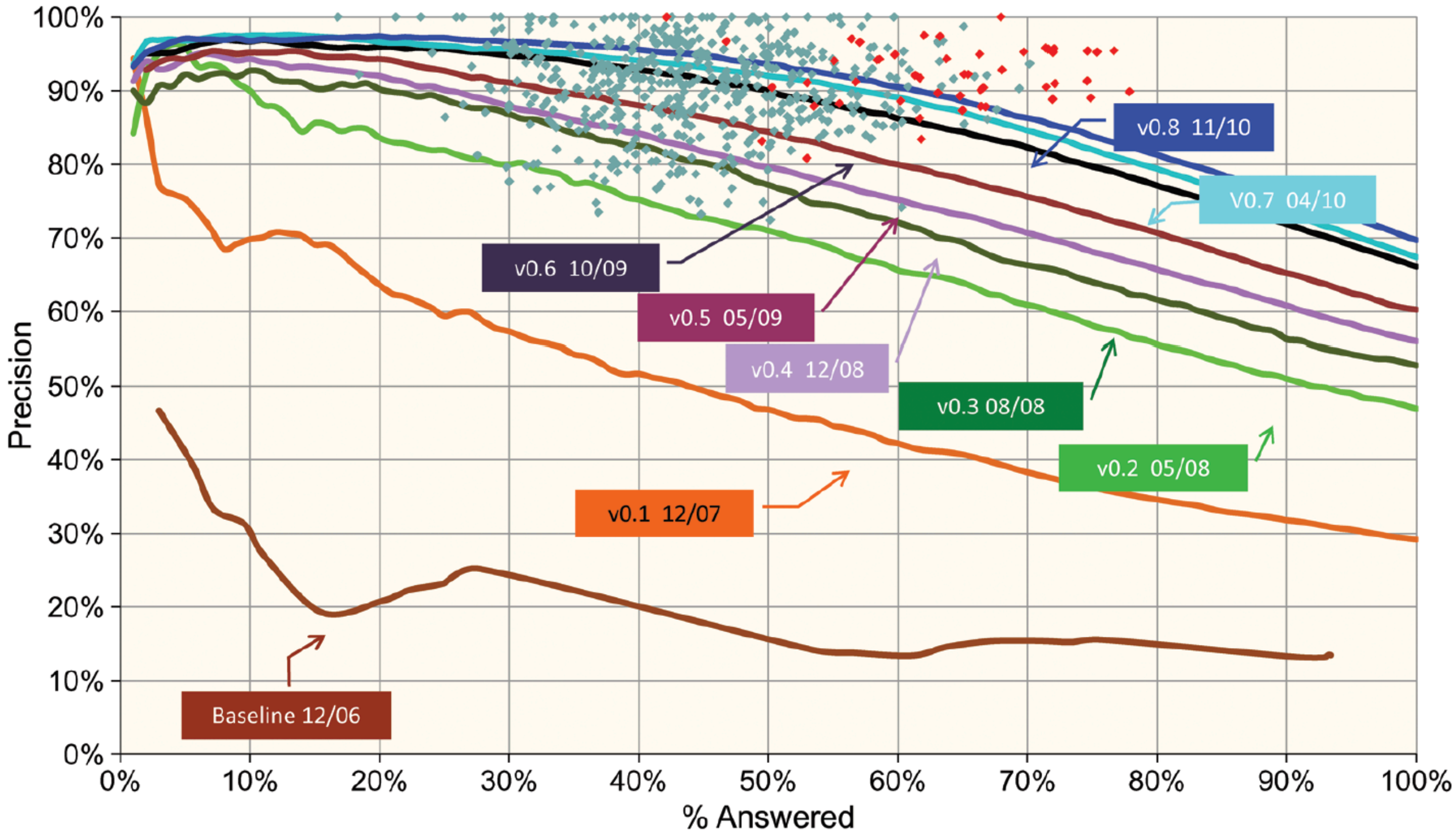
What is Jeopardy?

- Jeopardy! is an American TV quiz show (1964 - present)
  - Question: clues in the form of answers
  - Answer: phrase in the form of question

> Category: Michigan Mania
> Clue: In 1894 C.W. Post created his warm cereal drink Possum in this Michigan city
> Answer: Where is Battle Creek?

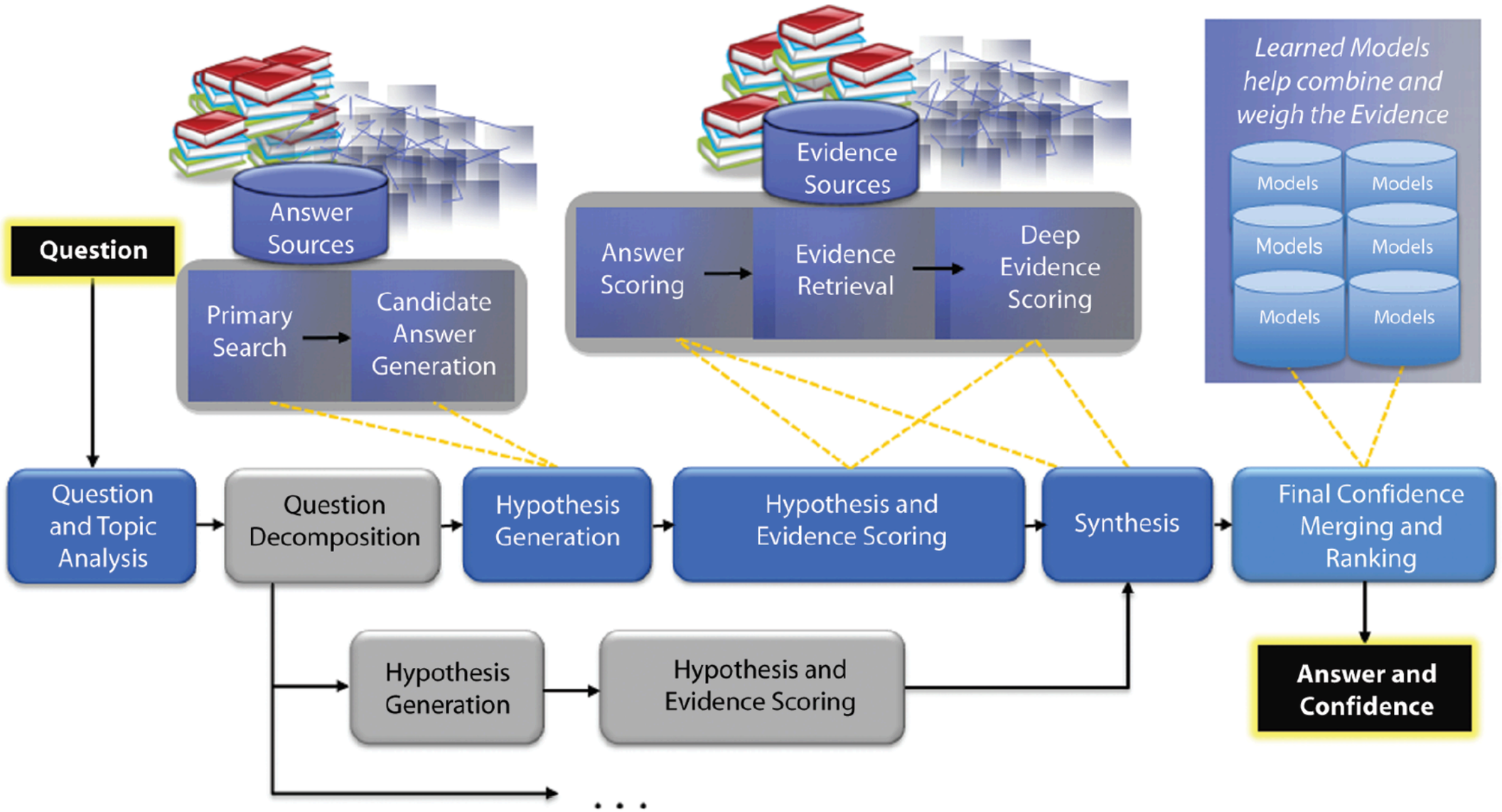
- Jeopardy! questions are "trivia" or "test" questions
  - Typically long and with detailed information
  - Question askers know the answer (not information seeking)

Progress: June-2007 to Nov-2011



Ferrucci, 2012. Introduction to “This is Watson”

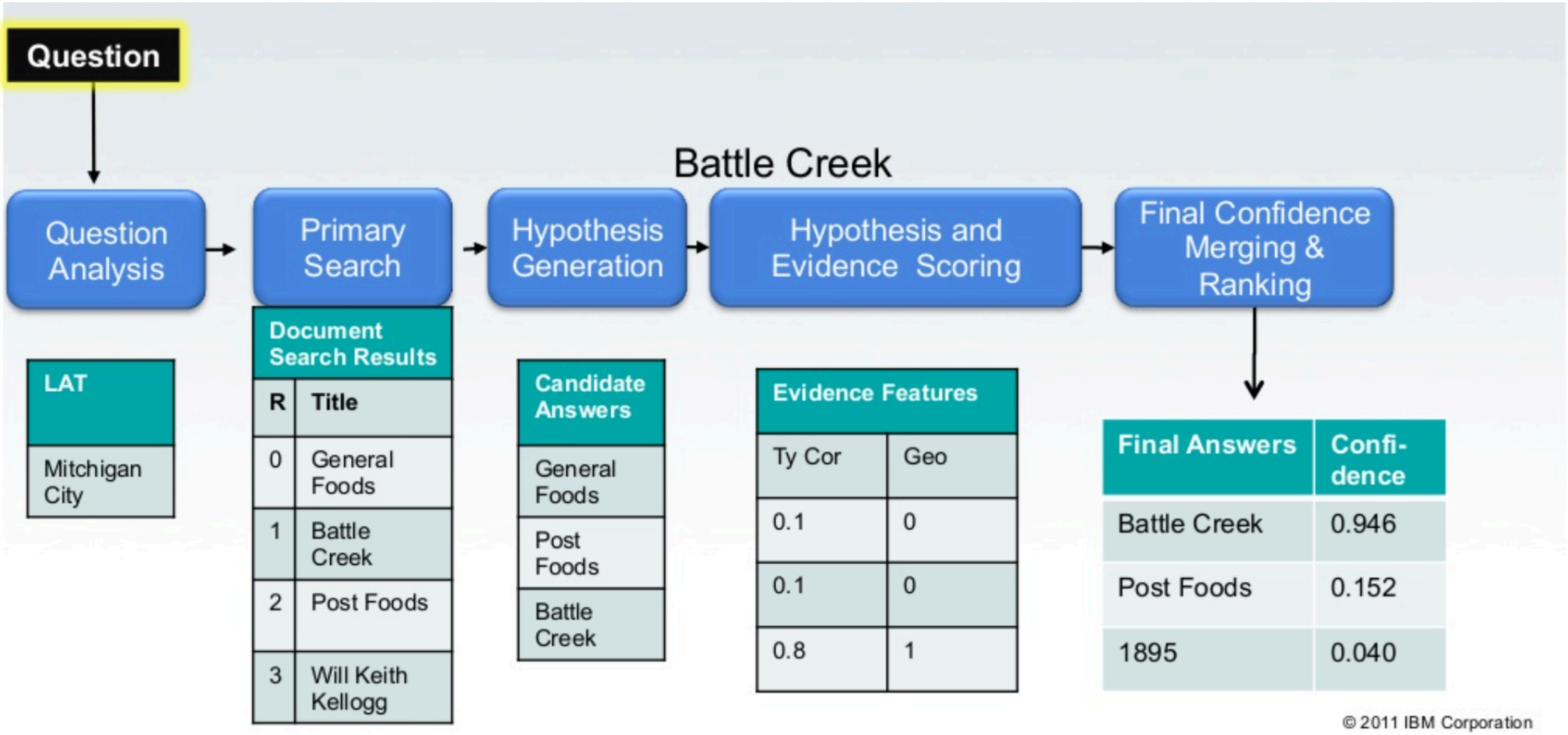
IBM Watson DeepQA architecture



Ferrucci, 2012. Introduction to “This is Watson”

“Minimal” DeepQA pipeline

Category: Michigan Mania
Clue: In 1894 C.W. Post created his warm cereal drink Possum in this Michigan city
Answer: Where is Battle Creek?



Ferrucci, 2012. Introduction to “This is Watson”

Success of IBM Watson DeepQA

Possible reasons

- Large-scale team work with strong engineering support
- Jeopardy! questions may in fact be easier to answer
- Wikipedia is an important resource for trivia questions

Implications

- Perceived as an important milestone of AI
- Rekindle the research interest in QA

Recent developments 2013+

- Trend: Macro-reading ⟶ Micro-reading
- General problem setting
  - Given a question and a "context", answer the question using the "context"
  - Two different goals
    - Test machine's intelligence AI (machine reading comprehension)
    - Fulfill user's information need (answer extraction/processing stage)

- Research directions guided by development of new tasks/datasets
- Rapid progress made by new deep learning models

Machine Comprehension Test

[Richardson et al., 2013]

Timmy liked to play games and play sports but more than anything he liked to collect things. He collected bottle caps. He collected sea shells. He collected baseball cards. He has collected baseball cards the longest. He likes to collect the thing that he has collected the longest the most. He once thought about collecting stamps but never did. His most expensive collection was not his favorite collection. Timmy spent the most money on his bottle cap collection.

1) Timmy liked to do which of these things the most?
   A) Collect things
   B) Collect stamps
   C) Play games
   D) Play sports

2) Which is Timmy's most expensive collection?
   A) Stamps
   B) Baseball Cards
   C) Bottle Cap
   D) Sea Shells

3) Which item did Timmy not collect?
   A) Bottle caps
   B) Baseball cards
   C) Stamps
   D) Sea shells

4) Which item did Timmy like to collect the most?
   A) Stamps
   B) Baseball cards
   C) Bottle caps
   D) Sea shells

Richardson et al., 2013. MCTest: A Challenge Dataset for the Open-Domain Machine Comprehension of Text

Stanford Question Answering Dataset (SQuAD)

[Rajpurkar et al., 2016]

In meteorology, precipitation is any product of the condensation of atmospheric water vapor that falls under **gravity**. The main forms of precipitation include drizzle, rain, sleet, snow, **graupel** and hail... Precipitation forms as smaller droplets coalesce via collision with other rain drops or ice crystals **within a cloud**. Short, intense periods of rain in scattered locations are called "showers".

What causes precipitation to fall?
gravity

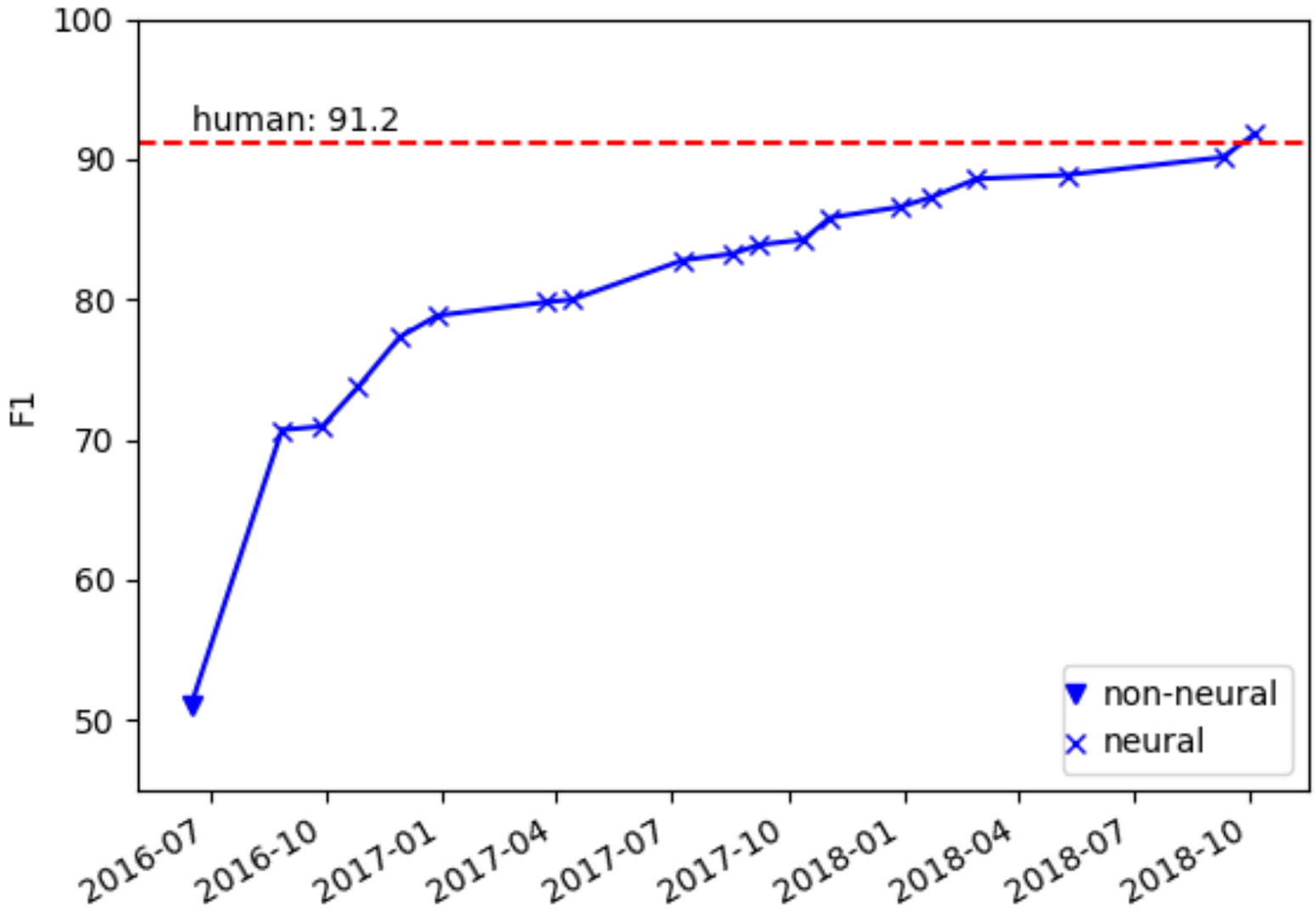
What is another main form of precipitation besides drizzle, rain, snow, sleet and hail?
graupel

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

- (passage, question, answer) triples
- Passage is from Wikipedia, question is crowd-sourced
- Answer must be a span of text in the passage (aka. "extractive question answering")
- SQuAD 1.1: 100k answerable questions, SQuAD 2.0: another 50k unanswerable questions

Rajpurkar et al., 2016. SQuAD: 100,000+ Questions for Machine Comprehension of Text

Stanford Question Answering Dataset (SQuAD)



Rajpurkar et al., 2016. SQuAD: 100,000+ Questions for Machine Comprehension of Text

WikiQA

Given a factoid question, find the sentence in the candidate set that

- Contains the answer
- Can sufficiently support the answer

> Q: Who won the best actor Oscar in 1973?
> S1: Jack Lemmon was awarded the Best Actor Oscar for Save the Tiger (1973).
> S2: Academy award winner Kevin Spacey said that Jack Lemmon is remembered as always making time for others.

Properties

- Real questions (from Bing query logs)
- Candidate sentences from Wikipedia description paragraphs
- Including questions without no answers
  - Answer Triggering: whether answer exists in contexts

Yang et al., 2015. WikiQA: A Challenge Dataset for Open-Domain Question Answering

And many more…

- Reading comprehension
  - RACE [Lai et al., 2017], DuoRC [Saha et al., 2018]
- Fill-in-the-blank questions
  - DeepMind Q&A Dataset [Hermann et al., 2015], Facebook Children Stories [Hill et al., 2016]
- Reasoning challenges
  - Facebook bAbI [Weston et al., 2015], AI2 ARC [Clark et al., 2018], Multi-RC [Khashabi et al., 2018]
- Multi-turn questions
  - SQA [Iyyer et al., 2017], QuAC [Choi et al., 2018], CoQA [Reddy et al., 2019]
- Multi-hop questions
  - HotpotQA [Yang et al., 2018], OBQA [Mihaylov et al., 2018], QASC [Khot et al., 2020]

Summary & Extended Reading

- Early QA Systems
  - Simmons, 1965. Answering English Questions by Computer: a Survey
- TREC Open-domain Question Answering
  - Prager, 2007. Open-Domain Question-Answering
  - Lin, 2007. An Exploration of the Principles Underlying Redundancy-Based Factoid Question Answering
- IBM Watson: The Deep QA project
  - Ferrucci et al., 2010. Building Watson: An Overview of the DeepQA Project
  - Ferrucci, 2012. This is Watson (IBM Journal of Research and Development)
  - Boytsov, 2018. Demystifying IBM Watson
- Recent developments 2013+