The Watson System

— Introduction, Overview and Discussion —

Christoph Benzmüller

Freie Universität Berlin, Germany

June 25, 2013

Chess champions: IBM's Deep Blue vs Kasparov

[Source: wikipedia.de, License: CC-BY]

Deep Blue, a computer similar to this one defeated chess world champion Garry Kasparov in May 1997. It is the first computer to win a match against a world champion. (126 million positions per second).

[Source: afflictor.com]

Chess champions: IBM's Deep Blue vs Kasparov

The 1996 match

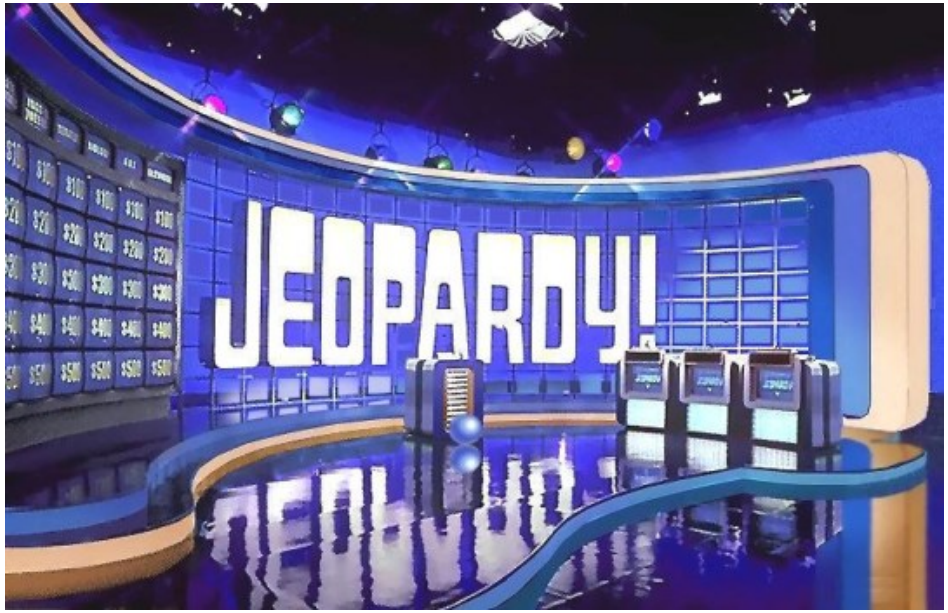
Game #	White	Black	Result	Comment
1	**Deep Blue**	Kasparov	1–0	
2	**Kasparov**	Deep Blue	1–0	
3	Deep Blue	Kasparov	½–½	Draw by mutual agreement
4	Kasparov	Deep Blue	½–½	Draw by mutual agreement
5	Deep Blue	**Kasparov**	0–1	Kasparov offered a draw after the 23rd move.
6	**Kasparov**	Deep Blue	1–0	
Result: **Kasparov–Deep Blue: 4–2**				

The 1997 rematch

Game #	White	Black	Result	Comment
1	**Kasparov**	Deep Blue	1–0	
2	**Deep Blue**	Kasparov	1–0	
3	Kasparov	Deep Blue	½–½	Draw by mutual agreement
4	Deep Blue	Kasparov	½–½	Draw by mutual agreement
5	Kasparov	Deep Blue	½–½	Draw by mutual agreement
6	**Deep Blue**	Kasparov	1–0	
Result: **Deep Blue–Kasparov: 3½–2½**				

[Source: wikipedia.org]

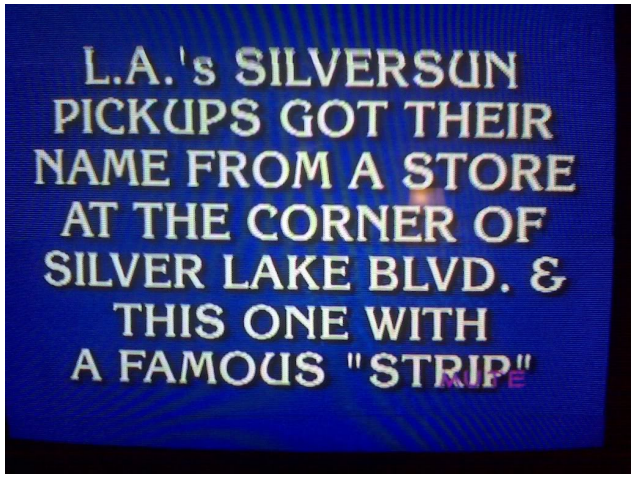
This lecture: Let's play Jeopardy!



[Source: stackoverflow.com]

THE "GU"	CODE NAMES	NAME THE NAMESPACE	HELLO WORLD	MOVIES	BEFORE & AFTER
$200	$200	$200	$200	$200	$200
$400	$400	$400	$400	$400	$400
$600	$600	$600	$600	$600	$600
$800	$800	$800	$800	$800	$800
$1000	$1000	$1000	$1000	$1000	$1000

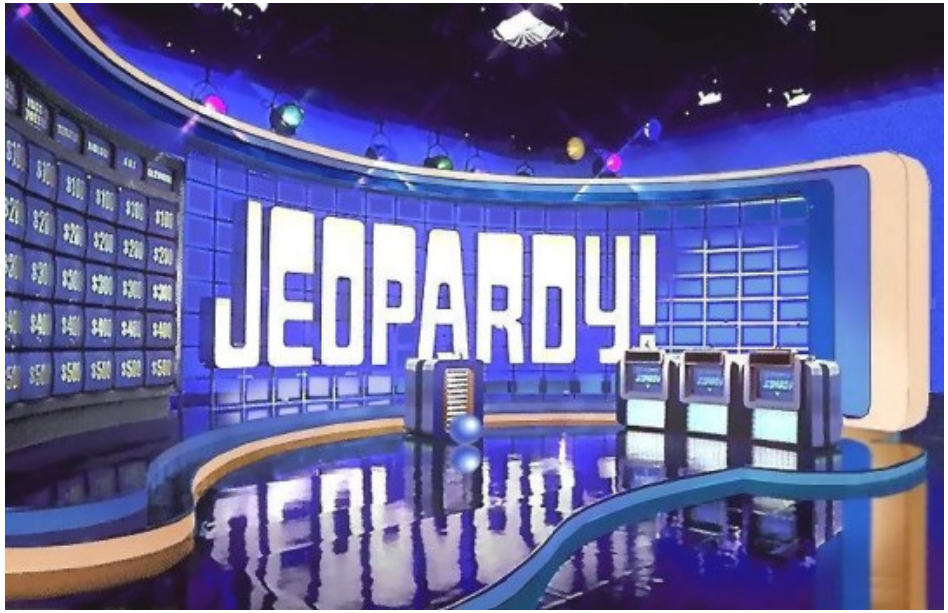
[Source: channel9.msdn.com]



[Source: blogs.laweekly.com]

- famous U.S. quiz show (since 1964)
- answers are presented in natural language
- different categories provide hints
- matching questions are to be provided by contestants

This lecture: Let's play Jeopardy!

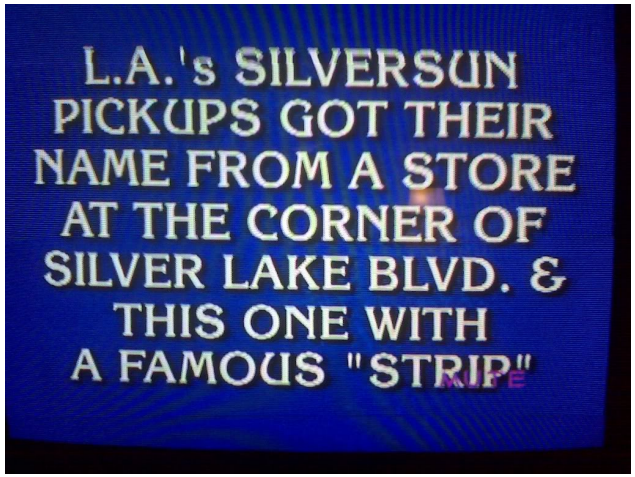


[Source: stackoverflow.com]

- famous U.S. quiz show (since 1964)
- answers are presented in natural language
- different categories provide hints
- matching questions are to be provided by contestants

THE "GU"	CODE NAMES	NAME THE NAMESPACE	HELLO WORLD	MOVIES	BEFORE & AFTER
$200	$200	$200	$200	$200	$200
$400	$400	$400	$400	$400	$400
$600	$600	$600	$600	$600	$600
$800	$800	$800	$800	$800	$800
$1000	$1000	$1000	$1000	$1000	$1000

[Source: channel9.msdn.com]



[Source: blogs.laweekly.com]

Answer:
What is Sunset Boulevard?

Jeopardy! — How does it work?

- **Jeopardy!**
  - six categories / five clues each (incrementally valued)
  - question → contestants buzz in (if confident)
  - correct answer → money added & select next question
  - incorrect answer → money subtracted, others may answer
- **Double Jeopardy!**
  - like above, but values are doubled
- **Daily Double**
  - one in each round above (Jeopardy! and Double Jeopardy!)
  - contestants wager (between 5 dollars and their total score)
- **Final Jeopardy!**
  - barriers between players (they can't see each other anymore)
  - single question
  - all three contestant write answer down
  - beforehand they wager (between 5 dollars and their total score)

Jeopardy! — How does it work?

- **Jeopardy!**
  - six categories / five clues each (incrementally valued)
  - question → contestants buzz in (if confident)
  - correct answer → money added & select next question
  - incorrect answer → money subtracted, others may answer
- **Double Jeopardy!**
  - like above, but values are doubled
- **Daily Double**
  - one in each round above (Jeopardy! and Double Jeopardy!)
  - contestants wager (between 5 dollars and their total score)
- **Final Jeopardy!**
  - barriers between players (they can't see each other anymore)
  - single question
  - all three contestant write answer down
  - beforehand they wager (between 5 dollars and their total score)

Jeopardy! — How does it work?

- **Jeopardy!**
  - six categories / five clues each (incrementally valued)
  - question → contestants buzz in (if confident)
  - correct answer → money added & select next question
  - incorrect answer → money subtracted, others may answer
- **Double Jeopardy!**
  - like above, but values are doubled
- **Daily Double**
  - one in each round above (Jeopardy! and Double Jeopardy!)
  - contestants wager (between 5 dollars and their total score)
- **Final Jeopardy!**
  - barriers between players (they can't see each other anymore)
  - single question
  - all three contestant write answer down
  - beforehand they wager (between 5 dollars and their total score)

Jeopardy! — How does it work?

- **Jeopardy!**
  - six categories / five clues each (incrementally valued)
  - question → contestants buzz in (if confident)
  - correct answer → money added & select next question
  - incorrect answer → money subtracted, others may answer
- **Double Jeopardy!**
  - like above, but values are doubled
- **Daily Double**
  - one in each round above (Jeopardy! and Double Jeopardy!)
  - contestants wager (between 5 dollars and their total score)
- **Final Jeopardy!**
  - barriers between players (they can't see each other anymore)
  - single question
  - all three contestant write answer down
  - beforehand they wager (between 5 dollars and their total score)

Jeopardy! — Humans or Machines?

Ken Jennings and Brad Rutter

- Ken Jennings: 74 wins in a row; $3,172,700 price money
- Brad Rutter: winner of Jeopardy! ultimate tournament of champions (2005); $3,470,102 price money

IBM's Watson

- open-domain question-answering (DeepQA) program
- named after Thomas J. Watson the founder of IBM
- developed since 2007; team of ˜25 members
- project leader D.A. Ferruci

Jan 14, 2011 (broadcasted on TV on Jan 14-16, 2011)

- one million dollar game
- received major public interest in the U.S. media

Jeopardy! — Humans or Machines?

Ken Jennings and Brad Rutter

- Ken Jennings: 74 wins in a row; $3,172,700 price money
- Brad Rutter: winner of Jeopardy! ultimate tournament of champions (2005); $3,470,102 price money

IBM's Watson

- open-domain question-answering (DeepQA) program
- named after Thomas J. Watson the founder of IBM
- developed since 2007; team of ˜25 members
- project leader D.A. Ferruci

Jan 14, 2011 (broadcasted on TV on Jan 14-16, 2011)

- one million dollar game
- received major public interest in the U.S. media

Jeopardy! — Humans or Machines?

Ken Jennings and Brad Rutter

- Ken Jennings: 74 wins in a row; $3,172,700 price money
- Brad Rutter: winner of Jeopardy! ultimate tournament of champions (2005); $3,470,102 price money

IBM's Watson

- open-domain question-answering (DeepQA) program
- named after Thomas J. Watson the founder of IBM
- developed since 2007; team of ˜25 members
- project leader D.A. Ferruci

Jan 14, 2011 (broadcasted on TV on Jan 14-16, 2011)

- one million dollar game
- received major public interest in the U.S. media

Jeopardy! — Introduction Movie

Let's watch a short introduction movie: watson.mpeg

- approx. 10 min
- thanks to Julian Röder

Here is another movie showing Watson in action:
http://www.youtube.com/watch?v=o6oS64Bpx0g

Here is a presentation by D.A. Ferruci
http://www.youtube.com/watch?v=UBM5JRyaoXw

And here is an article on Watson by Prof. Rojas:
http://www.heise.de/tp/artikel/36/36578/1.html

Further Reading

The slides for this lecture have been prepared from

- Building Watson: An Overview of the DeepQA Project. AI Magazine, Vol.31, No.3, 2010
- This is Watson. Journal of Research and Development, Vol.56, No.3/4, 2012

This Is Watson

See also http://www.christoph-benzmueller.de/2012-Watson

Jeopardy! — Examples

Some standard examples

Category: General Science
Clue: When hit by electrons, a phosphor gives off electromagnetic energy in this form.
Answer: Light (or Photons)

Category: Lincoln Blogs
Clue: Secretary Chase just submitted this to me for the third time; guess what, pal. This time I'm accepting it.
Answer: his resignation

Category: Head North
Clue: They're the two states you could be reentering if you're crossing Florida's northern border.
Answer: Georgia and Alabama

[Source: Ferrucci et al., Building Watson, AI Magazine, Vol 31, No 3]

Examples requiring decomposition

Category: Diplomatic Relations
Clue: Of the four countries in the world that the United States does not have diplomatic relations with, the one that's farthest north.
Inner subclue: The four countries in the world that the United States does not have diplomatic relations with (Bhutan, Cuba, Iran, North Korea).
Outer subclue: Of Bhutan, Cuba, Iran, and North Korea, the one that's farthest north.
Answer: North Korea

[Source: Ferrucci et al., Building Watson, AI Magazine, Vol 31, No 3]

Jeopardy! — Examples

Puzzles

Category: Before and After Goes to the Movies
Clue: Film of a typical day in the life of the Beatles, which includes running from bloodthirsty zombie fans in a Romero classic.
Subclue 2: Film of a typical day in the life of the Beatles.
Answer 1: (*A Hard Day's Night*)
Subclue 2: Running from bloodthirsty zombie fans in a Romero classic.
Answer 2: (*Night of the Living Dead*)
Answer: *A Hard Day's Night of the Living Dead*

Category: Rhyme Time
Clue: It's where Pele stores his ball.
Subclue 1: Pele ball (soccer)
Subclue 2: where store (cabinet, drawer, locker, and so on)
Answer: soccer locker

[Source: Ferrucci et al., Building Watson, AI Magazine, Vol 31, No 3]

Jeopardy! — Examples

Other types of questions

- **multiple choice**
  BUSY AS A BEAVER: *Of 1, 5, or 15, the rough maximum number of minutes a beaver can hold its breath underwater.*

Other types of questions

- **multiple choice**
  BUSY AS A BEAVER: *Of 1, 5, or 15, the rough maximum number of minutes a beaver can hold its breath underwater.*
  (Answer: 15)

Jeopardy! — Examples

Other types of questions

- **multiple choice**
  BUSY AS A BEAVER: *Of 1, 5, or 15, the rough maximum number of minutes a beaver can hold its breath underwater.*
  (Answer: 15)
- **constraint bearing**
  ARE YOU A FOOD "E"?: *Escoffier says to leave them in their shells & soak them in a mixture of water, vinegar, salt, and flour.*

Jeopardy! — Examples

Other types of questions

- **multiple choice**
  BUSY AS A BEAVER: *Of 1, 5, or 15, the rough maximum number of minutes a beaver can hold its breath underwater.*
  (Answer: 15)
- **constraint bearing**
  ARE YOU A FOOD "E"?: *Escoffier says to leave them in their shells & soak them in a mixture of water, vinegar, salt, and flour.*
  (Answer: Escargots)

Jeopardy! — Examples

Other types of questions

- **multiple choice**
  BUSY AS A BEAVER: *Of 1, 5, or 15, the rough maximum number of minutes a beaver can hold its breath underwater.*
  (Answer: 15)
- **constraint bearing**
  ARE YOU A FOOD "E"?: *Escoffier says to leave them in their shells & soak them in a mixture of water, vinegar, salt, and flour.*
  (Answer: Escargots)
- **lexical constraint, decomposable, fill-in-the blank**
  ONLY ONE VOWEL: *Proverbially, you can be "flying" this or be this "and dry".*

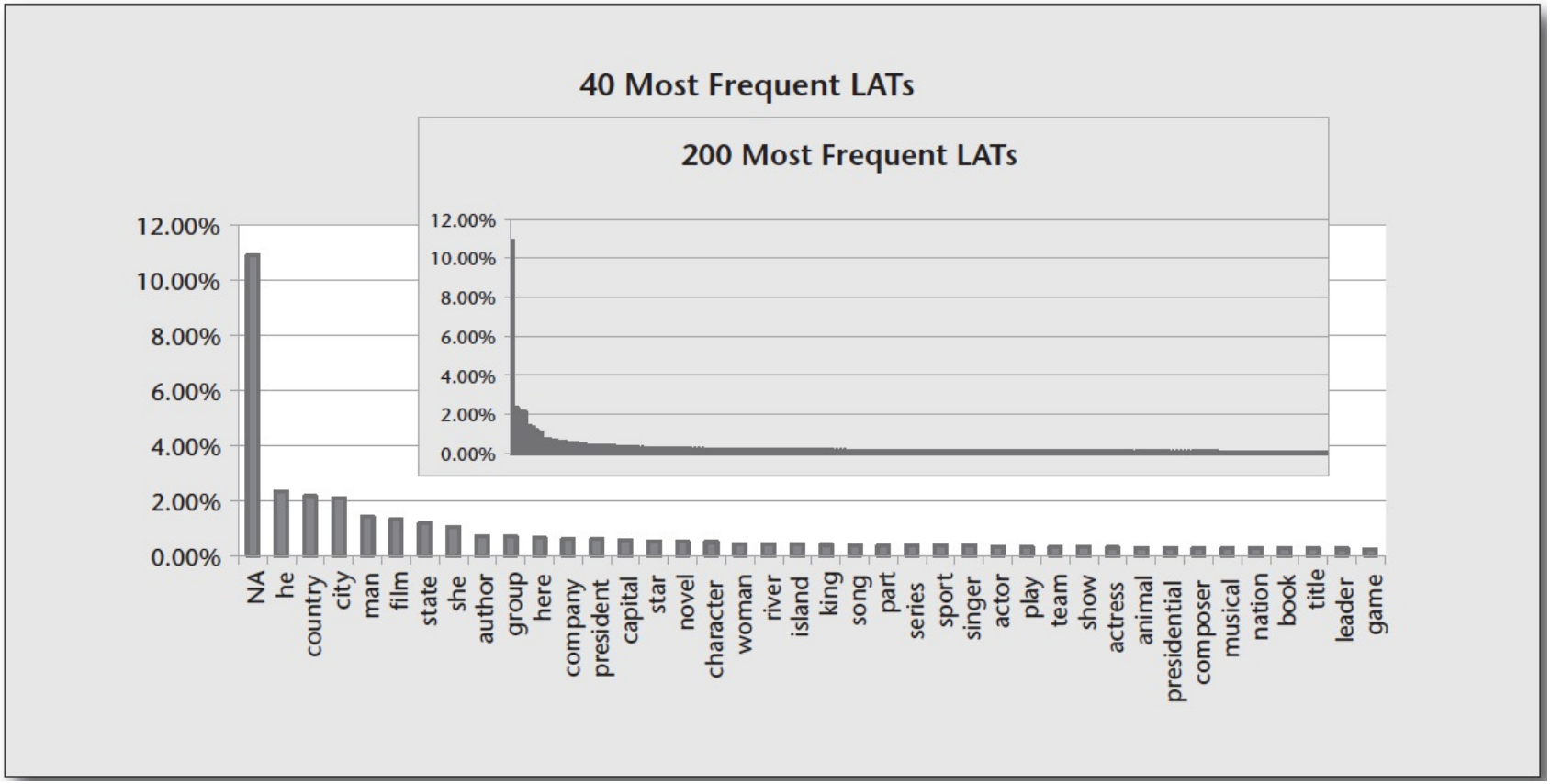
Jeopardy! — Examples

Other types of questions

- **multiple choice**
  BUSY AS A BEAVER: *Of 1, 5, or 15, the rough maximum number of minutes a beaver can hold its breath underwater.*
  (Answer: 15)
- **constraint bearing**
  ARE YOU A FOOD "E"?: *Escoffier says to leave them in their shells & soak them in a mixture of water, vinegar, salt, and flour.*
  (Answer: Escargots)
- **lexical constraint, decomposable, fill-in-the blank**
  ONLY ONE VOWEL: *Proverbially, you can be "flying" this or be this "and dry".*
  (Answer: high)

Jeopardy! — Lexical Answer Types (LAT)

LAT: a word in the clue that indicates the type of the answer, independent of assigning semantics to that word.



[Source: Ferrucci et al., Building Watson, AI Magazine, Vol 31, No 3]

Important hint for searching and verifying answer candidates!

Jeopardy! — How good are the human champs?

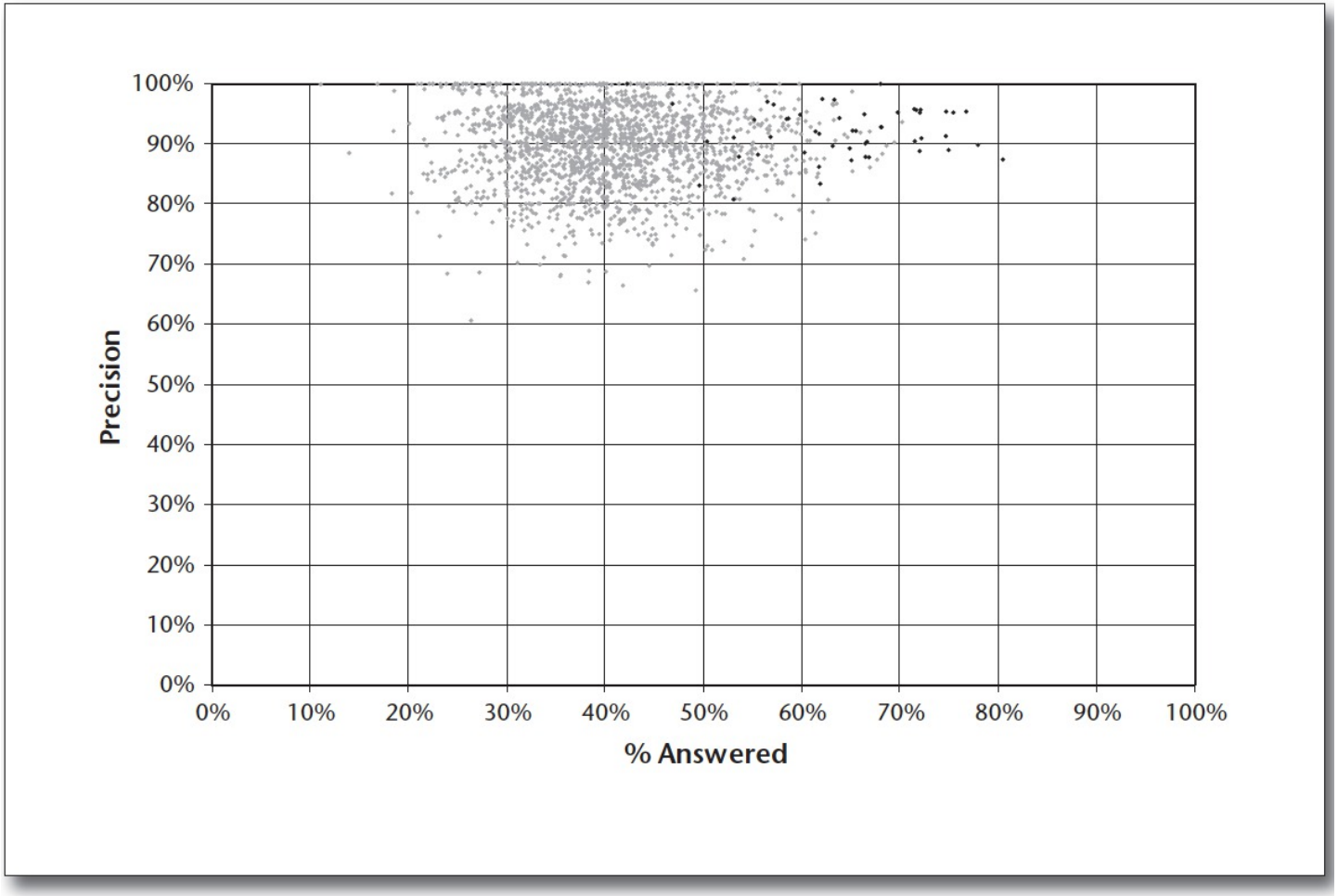
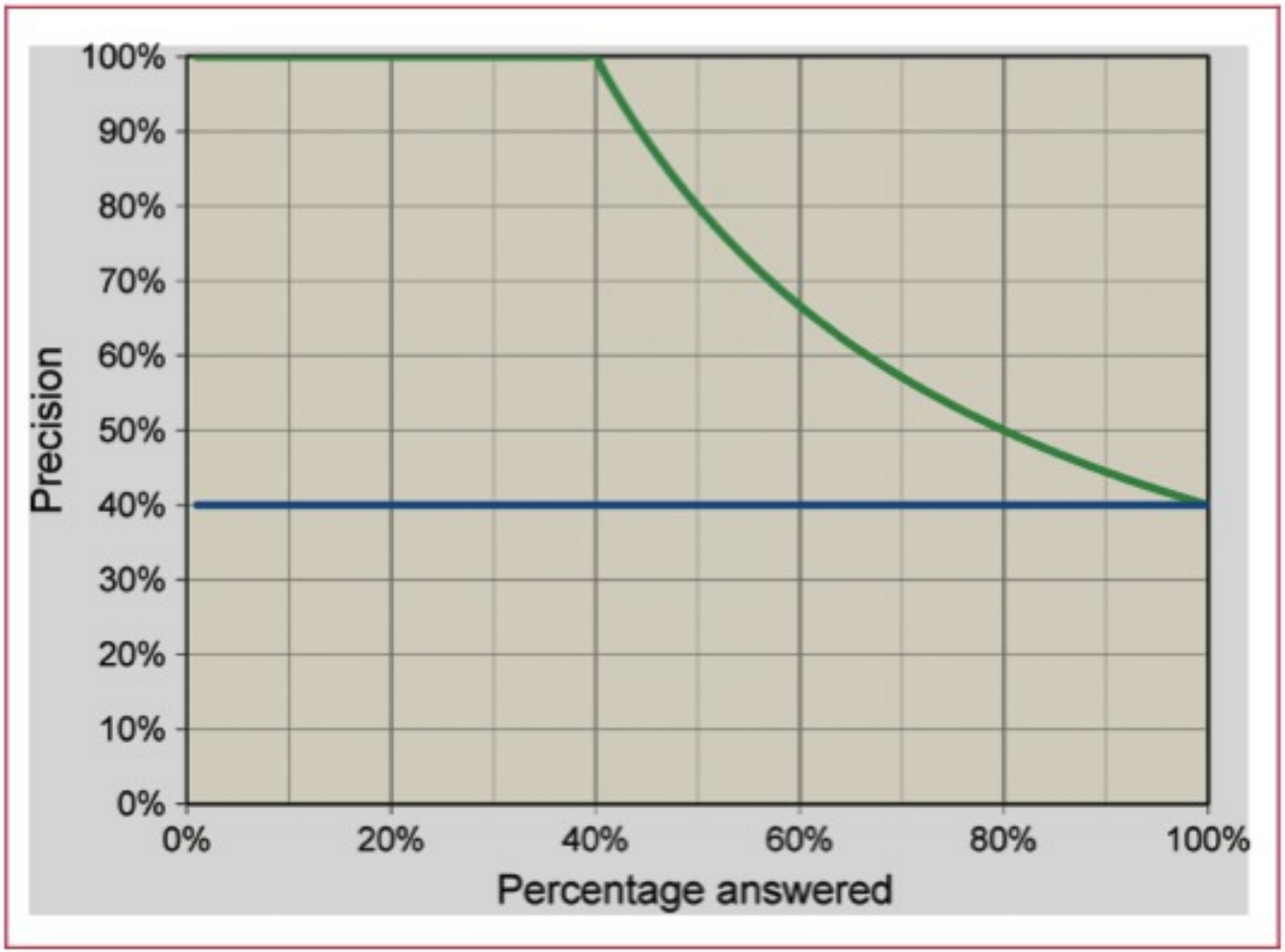


Figure 3. Champion Human Performance at Jeopardy.

[Source: Ferrucci et al., Building Watson, AI Magazine, Vol 31, No 3]

e.g. rightmost dot – a Ken Jennings game: 81% of questions answered; 88% of these with a correct answer
(Precision@70% became an important measure)

Jeopardy! — Importance of Confidence Estimation



[Source: Ferrucci et al., Building Watson, AI Magazine, Vol 31, No 3]

— perfect confidence estimation (40% accuracy assumed)

— no confidence estimation (40% accuracy assumed)

Jeopardy! — What do you think are the challenges?

Discussion:

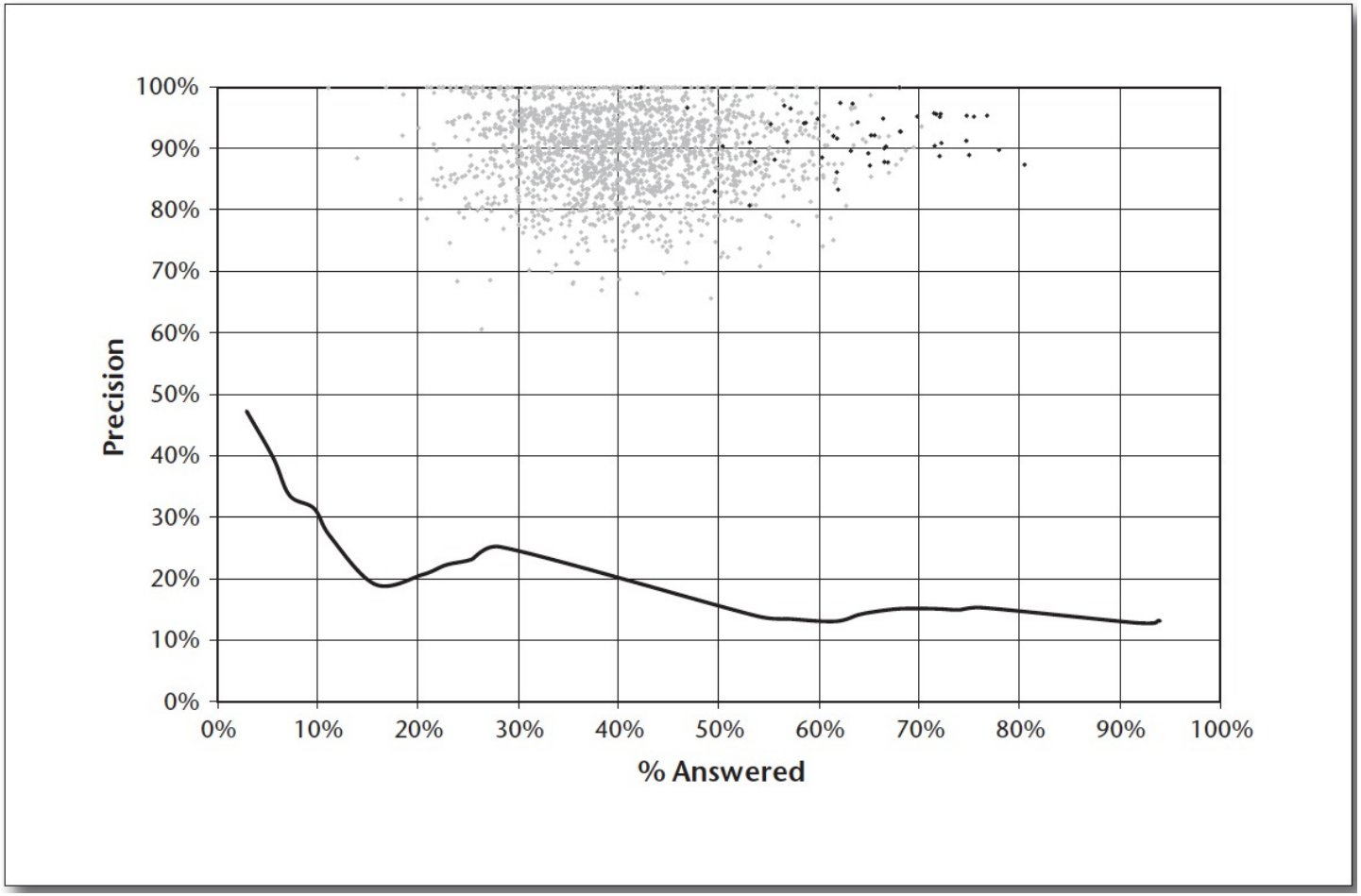
What skills are required for playing Jeopardy!

Jeopardy! — What do you think are the challenges?

Relevant skills include

- question-answering
  - natural language processing (NLP)
  - information retrieval (IR)
  - knowledge representation and reasoning (KR&R)
  - machine learning (ML)
  - human computer interfaces (HCIs)
  - ...
- other important aspects
  - speed
  - confidence estimation
  - clue selection
  - betting strategy

Jeopardy! — How good were machines in 2007?



[Source: Ferrucci et al., Building Watson, AI Magazine, Vol 31, No 3]

Performance of PIQUANT baseline system (developed by 4-person team for 6 years prior to taking on the Jeopardy Challenge).

The overarching principles of DeepQA & Watson

Massive parallelism

- Exploit massive parallelism in the consideration of multiple interpretations and hypotheses.

Many experts

- Facilitate the integration, application, and contextual evaluation of a wide range of loosely coupled probabilistic question and content analytics.

Pervasive confidence estimation

- No component commits to an answer; all components produce features and associated confidences, scoring different question and content interpretations. An underlying confidence-processing substrate learns how to stack and combine the scores.

Integrate shallow and deep knowledge

- Balance the use of strict semantics and shallow semantics, leveraging many loosely formed ontologies.

[Source: Ferrucci et al., Building Watson, AI Magazine, Vol 31, No 3]

The overarching principles of DeepQA & Watson

Massive parallelism

- Exploit massive parallelism in the consideration of multiple interpretations and hypotheses.

Many experts

- Facilitate the integration, application, and contextual evaluation of a wide range of loosely coupled probabilistic question and content analytics.

Pervasive confidence estimation

- No component commits to an answer; all components produce features and associated confidences, scoring different question and content interpretations. An underlying confidence-processing substrate learns how to stack and combine the scores.

Integrate shallow and deep knowledge

- Balance the use of strict semantics and shallow semantics, leveraging many loosely formed ontologies.

[Source: Ferrucci et al., Building Watson, AI Magazine, Vol 31, No 3]

The overarching principles of DeepQA & Watson

Massive parallelism

- Exploit massive parallelism in the consideration of multiple interpretations and hypotheses.

Many experts

- Facilitate the integration, application, and contextual evaluation of a wide range of loosely coupled probabilistic question and content analytics.

Pervasive confidence estimation

- No component commits to an answer; all components produce features and associated confidences, scoring different question and content interpretations. An underlying confidence-processing substrate learns how to stack and combine the scores.

Integrate shallow and deep knowledge

- Balance the use of strict semantics and shallow semantics, leveraging many loosely formed ontologies.

[Source: Ferrucci et al., Building Watson, AI Magazine, Vol 31, No 3]

The overarching principles of DeepQA & Watson

Massive parallelism

- Exploit massive parallelism in the consideration of multiple interpretations and hypotheses.

Many experts

- Facilitate the integration, application, and contextual evaluation of a wide range of loosely coupled probabilistic question and content analytics.

Pervasive confidence estimation

- No component commits to an answer; all components produce features and associated confidences, scoring different question and content interpretations. An underlying confidence-processing substrate learns how to stack and combine the scores.

Integrate shallow and deep knowledge

- Balance the use of strict semantics and shallow semantics, leveraging many loosely formed ontologies.

[Source: Ferrucci et al., Building Watson, AI Magazine, Vol 31, No 3]

Watson's performance development (2007-2011)

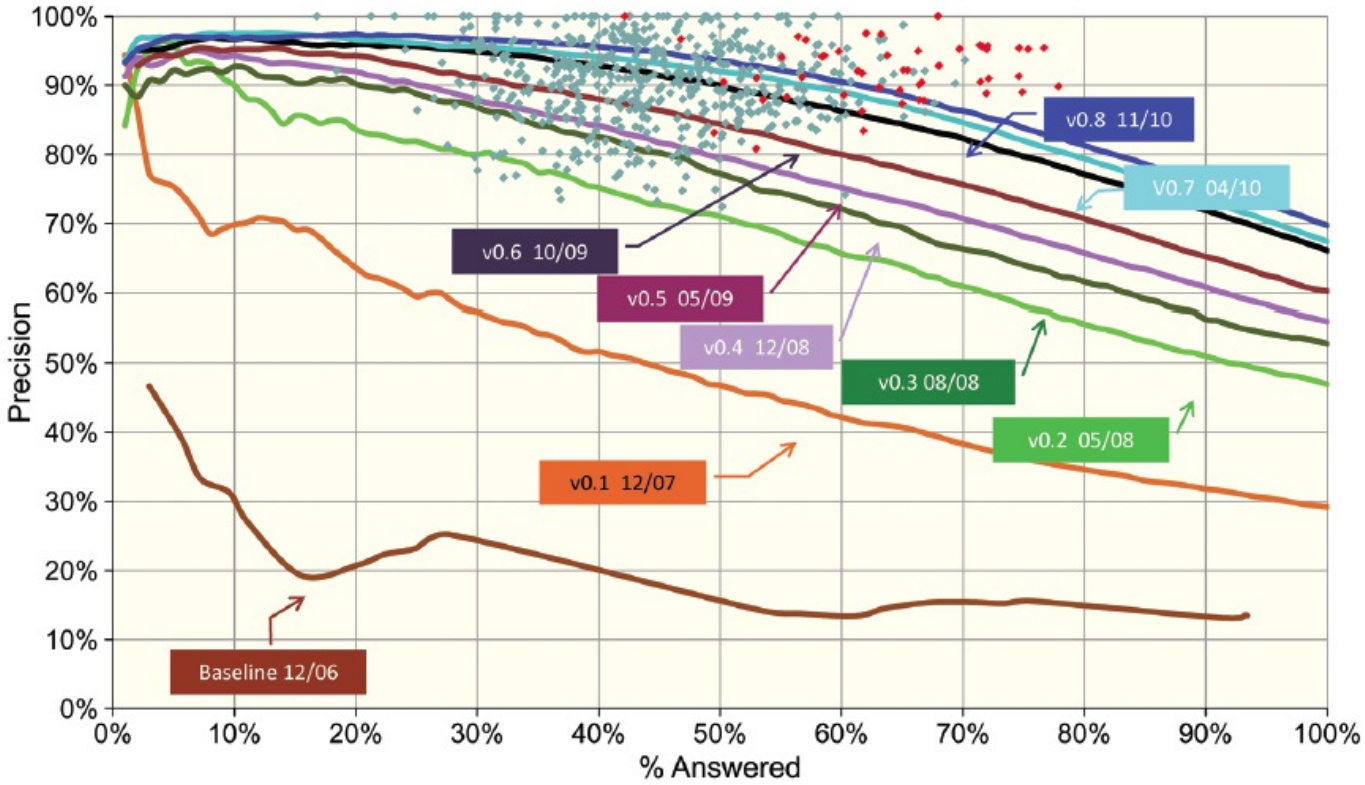


Figure 2

Incremental progress in answering precision on the Jeopardy! challenge: June 2007 to November 2011.

[Source: Ferrucci et al., This is Watson, IBM J. RES. & DEV. VOL. 56 NO. 3/4 PAPER 1]

Watson's system architecture

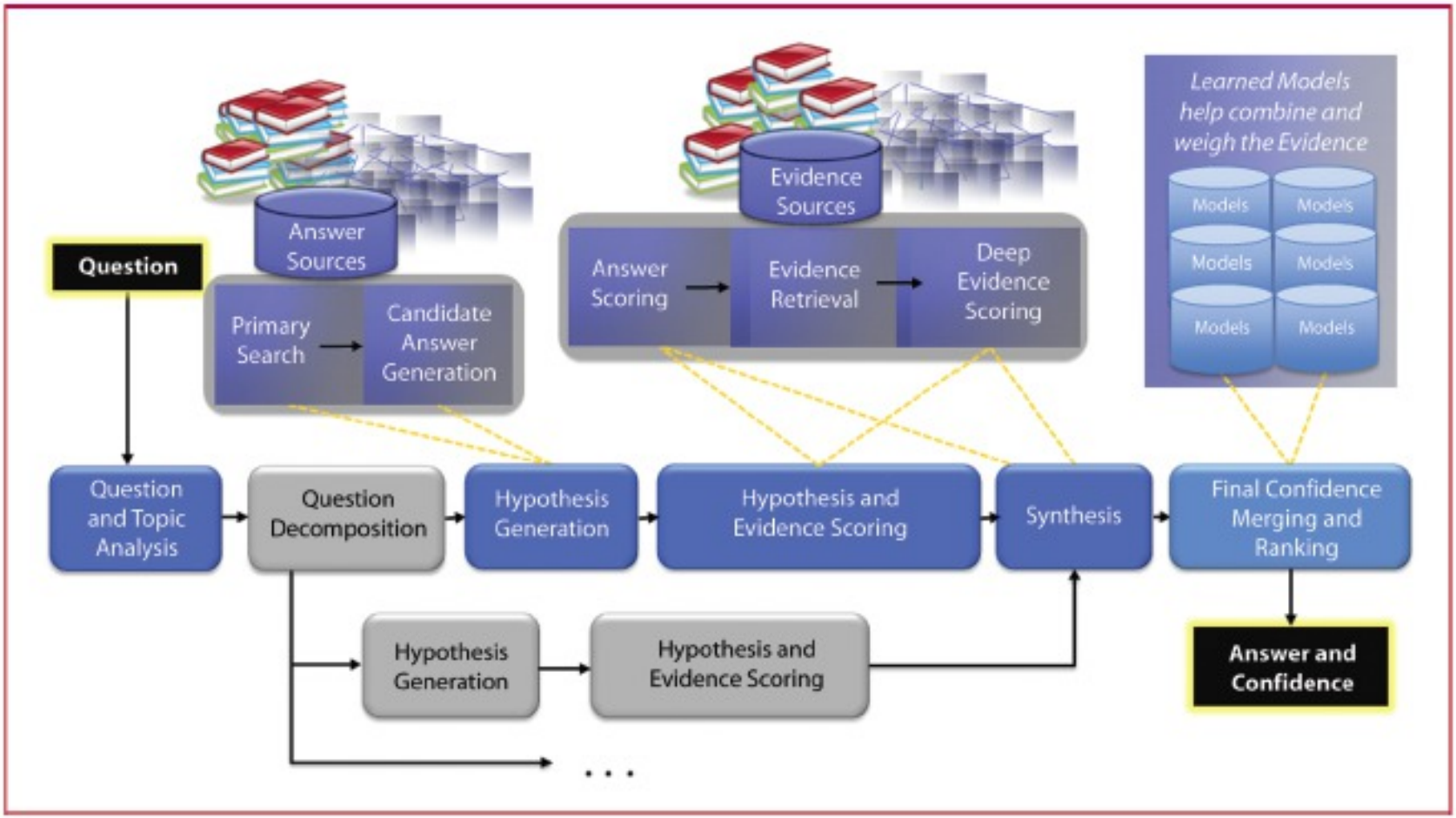


Figure 1

DeepQA architecture.

[Source: Ferrucci et al., This is Watson, IBM J. RES. & DEV. VOL. 56 NO. 3/4 PAPER 1]

Watson's base platform: UIMA

Unstructured Information Management Architecture (UIMA)

- built from 2001 to 2006
- software architecture and framework providing a common platform for integrating diverse collections of text (or speech and image) analytics
- independent of algorithmic approach, programming language, or underlying domain model
- supports cooperating software programs
- in 2006, IBM contributed UIMA to Apache

Open Advancement of QA (OAQA) initiative

- directly engage researchers to replicate, reuse, contribute research results
- foster rapid advancement of the state of the art in QA

Overall development strategy for Watson:

- start with baseline system
- analyze former Jeopardy games, use them as test corpus
- apply the **AdaptWatson** approach
  - run system over and over
  - approx. 8000 documented experiments
  - analyze errors and try to learn from them
  - adapt, replace, add, remove, improve system components

Watson's components: Knowledge Base/Corpus

Content acquisition (manual and automatic steps)

- initial analysis of example questions
  - lead to a selection of a baseline corpus of 3.5 million Wikipedia articles
- iterative process (**AdaptWatson**)
  - error analysis
  - source acquisition: new content
  - source transformation: extract information from sources (as a whole or in part), represent it in a form that the system can use
  - source expansion: increase the coverage by adding new information, includes lexical and syntactic variations
- different kinds of sources
  - encyclopedias, dictionaries, thesauri, newswire articles, literary works, . . . (unstructured)
  - taxonomies and ontologies such as DBpedia, Wordnet, Yago, Cyc, . . . (semi-structured/structured)

Watson's components: Knowledge Base/Corpus

Content acquisition (manual and automatic steps)

- initial analysis of example questions
  - lead to a selection of a baseline corpus of 3.5 million Wikipedia articles
- iterative process (**AdaptWatson**)
  - error analysis
  - source acquisition: new content
  - source transformation: extract information from sources (as a whole or in part), represent it in a form that the system can use
  - source expansion: increase the coverage by adding new information, includes lexical and syntactic variations
- different kinds of sources
  - encyclopedias, dictionaries, thesauri, newswire articles, literary works, . . . (unstructured)
  - taxonomies and ontologies such as DBpedia, Wordnet, Yago, Cyc, . . . (semi-structured/structured)

Watson's components: Knowledge Base/Corpus

Content acquisition (manual and automatic steps)

- initial analysis of example questions
  - lead to a selection of a baseline corpus of 3.5 million Wikipedia articles
- iterative process (**AdaptWatson**)
  - error analysis
  - source acquisition: new content
  - source transformation: extract information from sources (as a whole or in part), represent it in a form that the system can use
  - source expansion: increase the coverage by adding new information, includes lexical and syntactic variations
- different kinds of sources
  - encyclopedias, dictionaries, thesauri, newswire articles, literary works, . . . (unstructured)
  - taxonomies and ontologies such as DBpedia, Wordnet, Yago, Cyc, . . . (semi-structured/structured)

Watson — Question Analysis

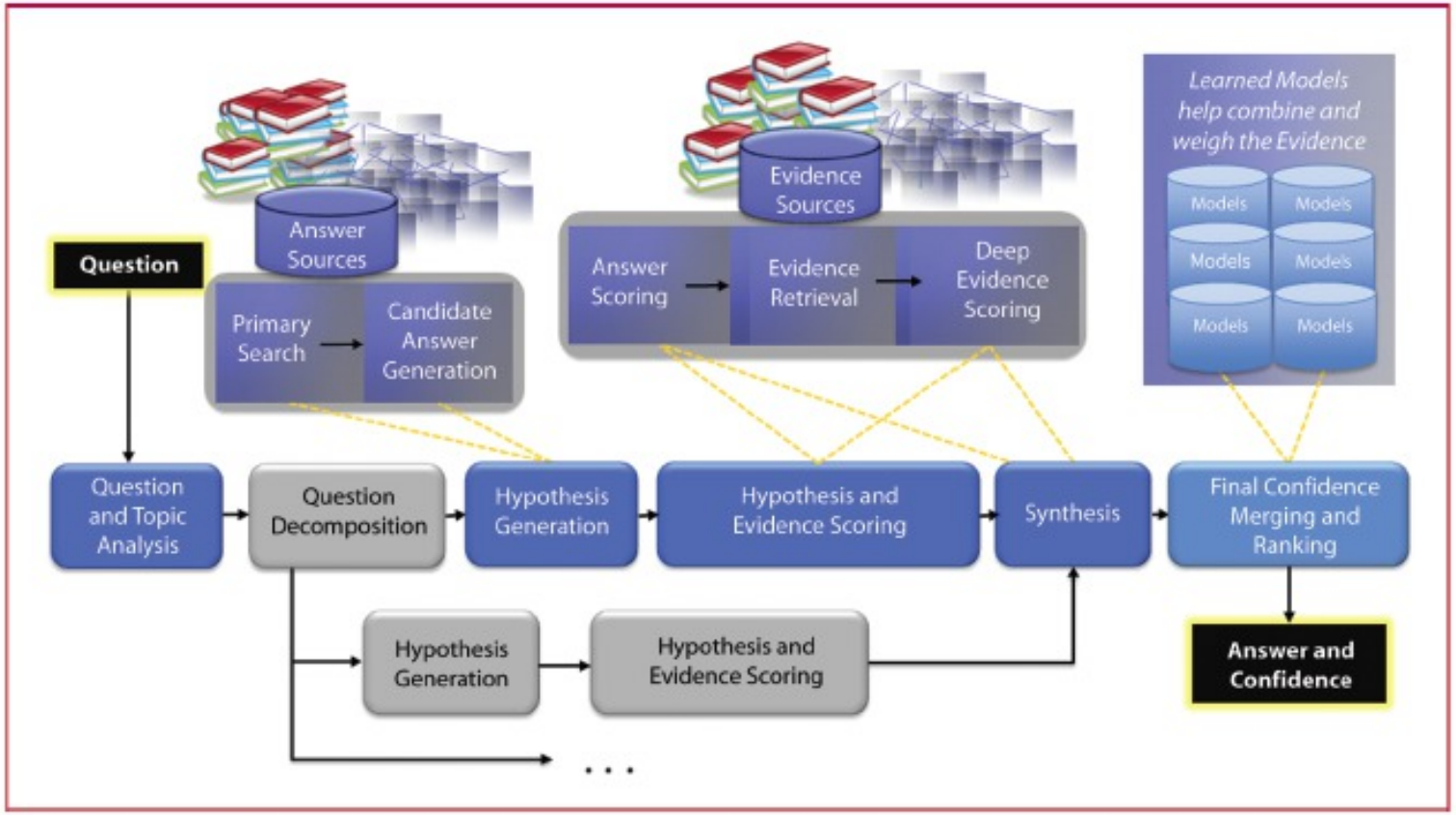


Figure 1

DeepQA architecture.

[Source: Ferrucci et al., This is Watson, IBM J. RES. & DEV. VOL. 56 NO. 3/4 PAPER 1]

Question Analysis:

Try to understand question and determine a processing strategy

Predominantly rule based approach (e.g. using Prolog):

- Question classification
  - identify type: standard, puzzle, multiple-choice, maths, ...
  - identify puns, constraints, components, subclues, ...
- Focus and LAT detection
- Relation detection
- Decomposition

Predominantly rule based approach (e.g. using Prolog):

- Question classification
- Focus and LAT detection
  - THEATRE: *A new play based on this Sir Arthur Conan Doyle canine* **classic** *opened on the London stage in 2007.*
  - POETS & POETRY: **He** *was a bank clerk in the Yukon before he published "Songs of a Sourdough" in 1907.*
  - foci: underlined text
  - foci headwords: bold text
  - LATs in this question: he, poet, clerk
  - different kinds of clues require different detection mechanisms
- Relation detection
- Decomposition

Predominantly rule based approach (e.g. using Prolog):

- Question classification
- Focus and LAT detection
- Relation detection
  - *They're the two states you could be reentering if you're crossing Florida's northern border.*
  - contained relation: borders(Florida,?x,north)
  - such relations are used by various components of Watson
- Decomposition

Predominantly rule based approach (e.g. using Prolog):

- Question classification
- Focus and LAT detection
- Relation detection
- Decomposition
  - break question into subquestions (if applicable)
  - BEFORE & AFTER: *The "Jerry Maguire" star who automatically maintains your vehicle's speed.*
    (Answer: Tom Cruise control)
  - FICTIONAL ANIMALS: *The name of this character, introduced in 1894, comes from the Hindi for "bear".*
    (Answer: Baloo)
  - answer processing (e.g. combination or synthesis)

Watson — Hypothesis Generation

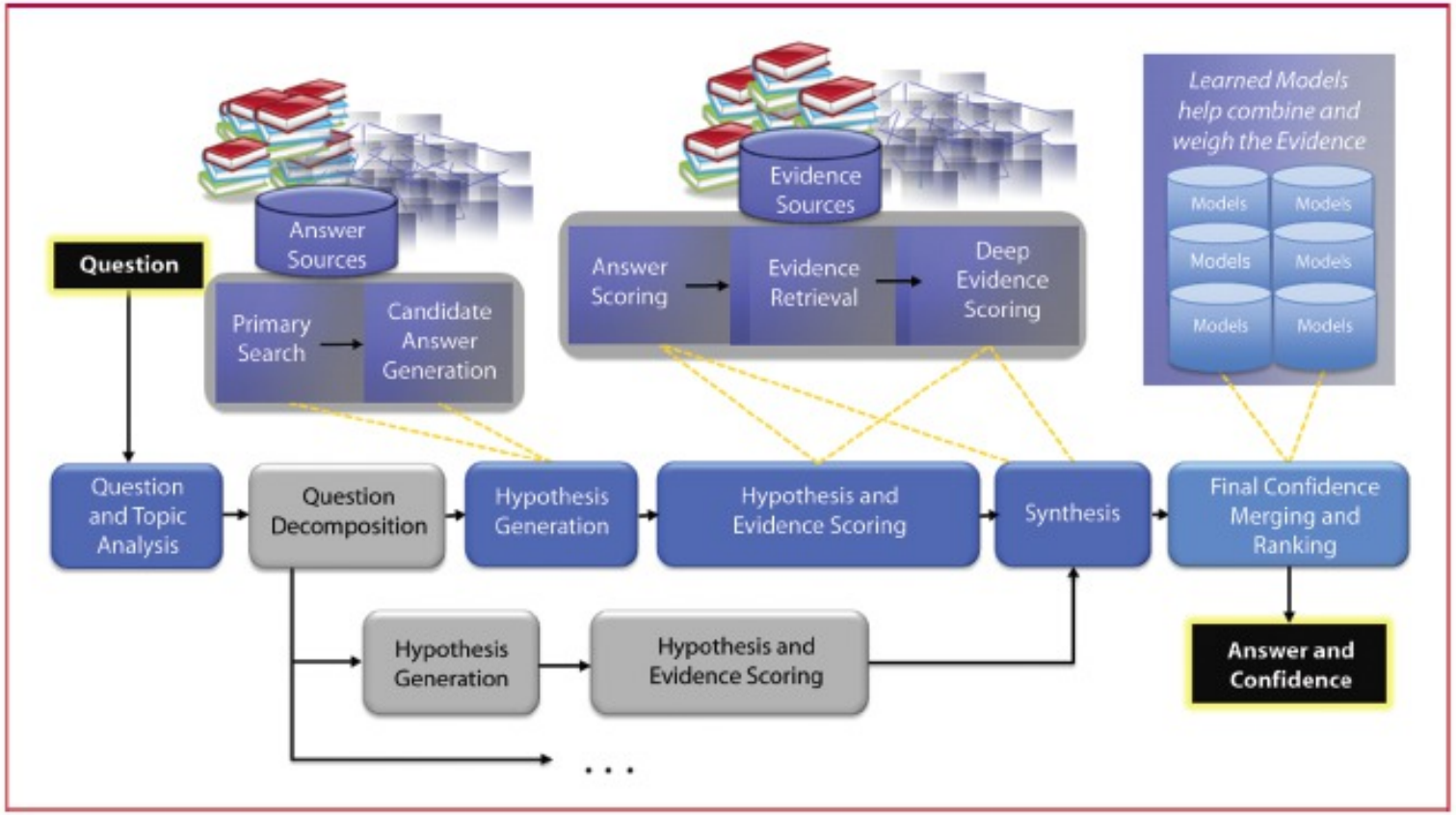


Figure 1

DeepQA architecture.

[Source: Ferrucci et al., This is Watson, IBM J. RES. & DEV. VOL. 56 NO. 3/4 PAPER 1]

Hypothesis Generation:

Use the results of analysis step to generate candidate answers

Hypothesis Generation has two phases

- **Primary Search (distinguish from 'Evidence Gathering')**
  - goal: find as much as possible answer-bearing content
  - uses multiple text search engines, knowledge base search on triple stores, other techniques
  - for 85% of the questions the correct answer is amongst top 250 ranked candidates
- **Candidate Answer Generation**

Hypothesis Generation has two phases

- **Primary Search (distinguish from 'Evidence Gathering')**
  - goal: find as much as possible answer-bearing content
  - uses multiple text search engines, knowledge base search on triple stores, other techniques
  - for 85% of the questions the correct answer is amongst top 250 ranked candidates
- **Candidate Answer Generation**
  - results from Primary Search are processed to generate candidate answers
  - examples: use title of Wikipedia document, apply named entity detection to text passages, simple answer extraction from triple stores
  - Watson typically generates several hundred candidate answers
  - important: at this stage the correct answer needs to be among the generated candidates (otherwise no success)

Watson — Hypothesis and Evidence Scoring

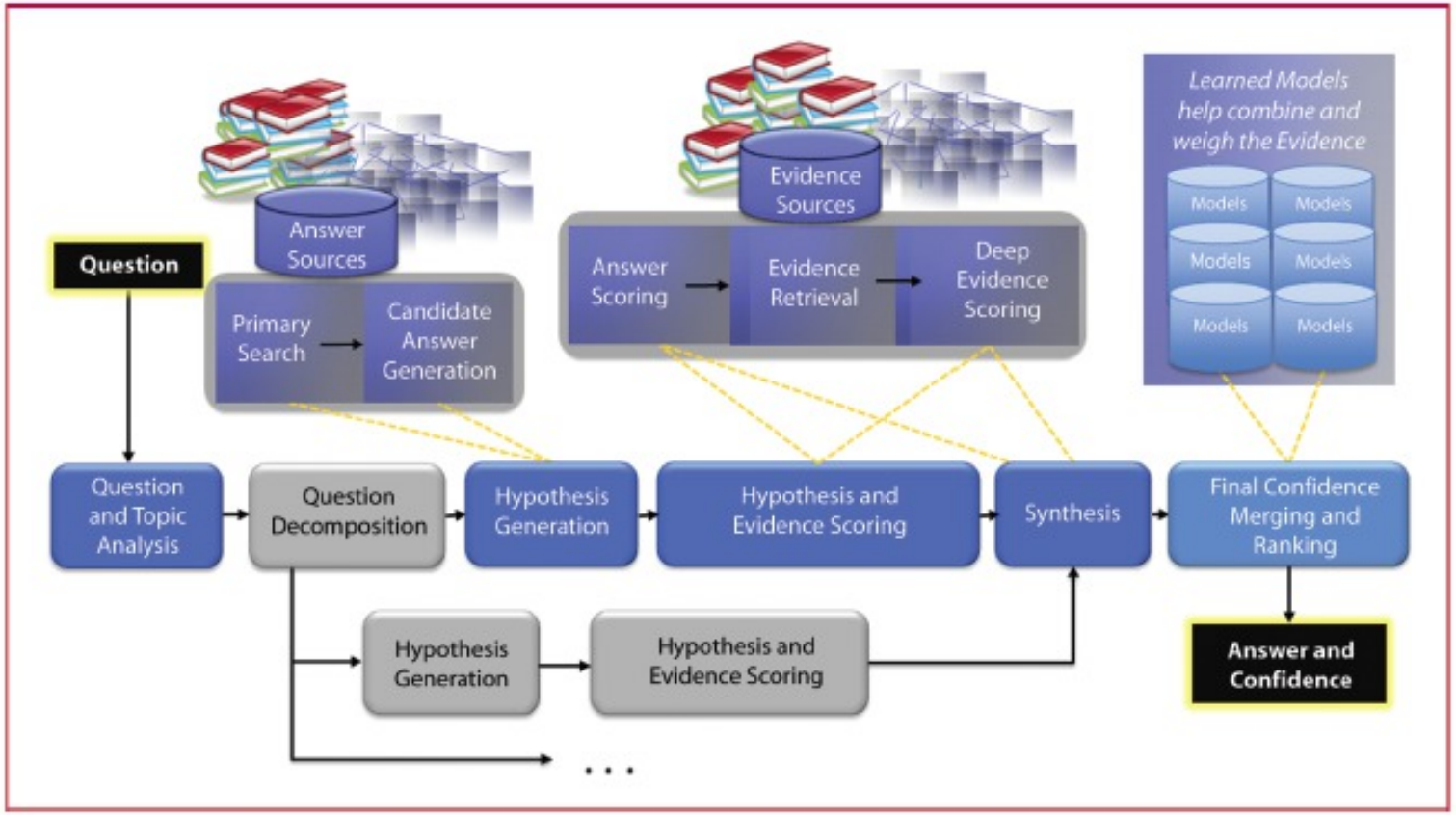


Figure 1

DeepQA architecture.

[Source: Ferrucci et al., This is Watson, IBM J. RES. & DEV. VOL. 56 NO. 3/4 PAPER 1]

Hypothesis and Evidence Scoring:

Filter candidate answers, evaluate and score the remaining ones

Hypothesis and Evidence Scoring

- **Soft Filtering (lightweight)**
  - reduces the candidate set to about 100 entries (parameterizable)
  - examples: compute likelihood of a candidate answer to be an instance of the LAT(s)
  - result is a soft filtering score for each candidate answer
  - candidates above a threshold are passed on for further scoring
- **Resource Intensive Hypothesis and Evidence Scoring**

Hypothesis and Evidence Scoring

- **Soft Filtering (lightweight)**
- **Resource Intensive Hypothesis and Evidence Scoring**
  - Evidence Retrieval:
    - search for additional supporting evidence
    - e.g. passage search: candidate answer is added to the primary search query, this retrieves passages that contain the candidate answer used in the context of the original question terms
    - e.g. search in triple stores
    - retrieved supporting evidence is routed to the deep evidence scoring components
  - Scoring

Hypothesis and Evidence Scoring

- **Soft Filtering (lightweight)**
- **Resource Intensive Hypothesis and Evidence Scoring**
  - Evidence Retrieval:
  - Scoring
    - here the main "deep content analysis" is done: determine degree of certainty that retrieved evidence supports the candidate answers
    - many different, very heterogeneous, components (software agents, exchanging information)
    - many different scoring components (probabilities, counting, etc., in un- & semistructured text, and triple stores)
    - employ spatial and temporal relationships, taxonomic classification, lexical and semantic relations, ...
    - no dominance!
    - individual scores are combined into an overall evidence profile
    - evidence dimensions include: taxonomic, geospatial (location), temporal, source reliability, gender, name consistency,
    - result: **feature vector**

Watson — Synthesis

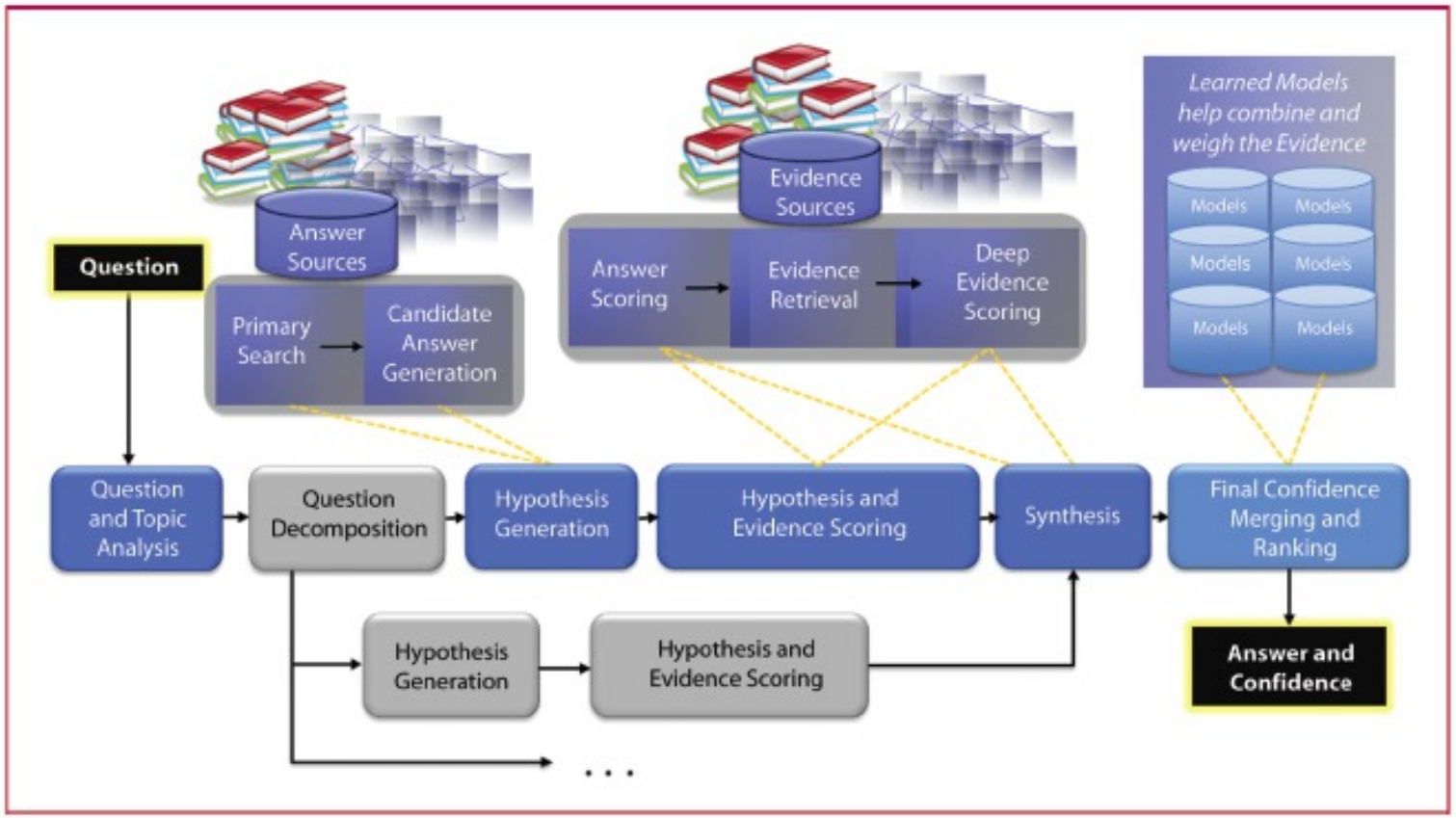


Figure 1

DeepQA architecture.

[Source: Ferrucci et al., This is Watson, IBM J. RES. & DEV. VOL. 56 NO. 3/4 PAPER 1]

Synthesis:

- merge different values for one feature
- combine results for decomposed problems

Watson — Final Confidence Merging and Ranking

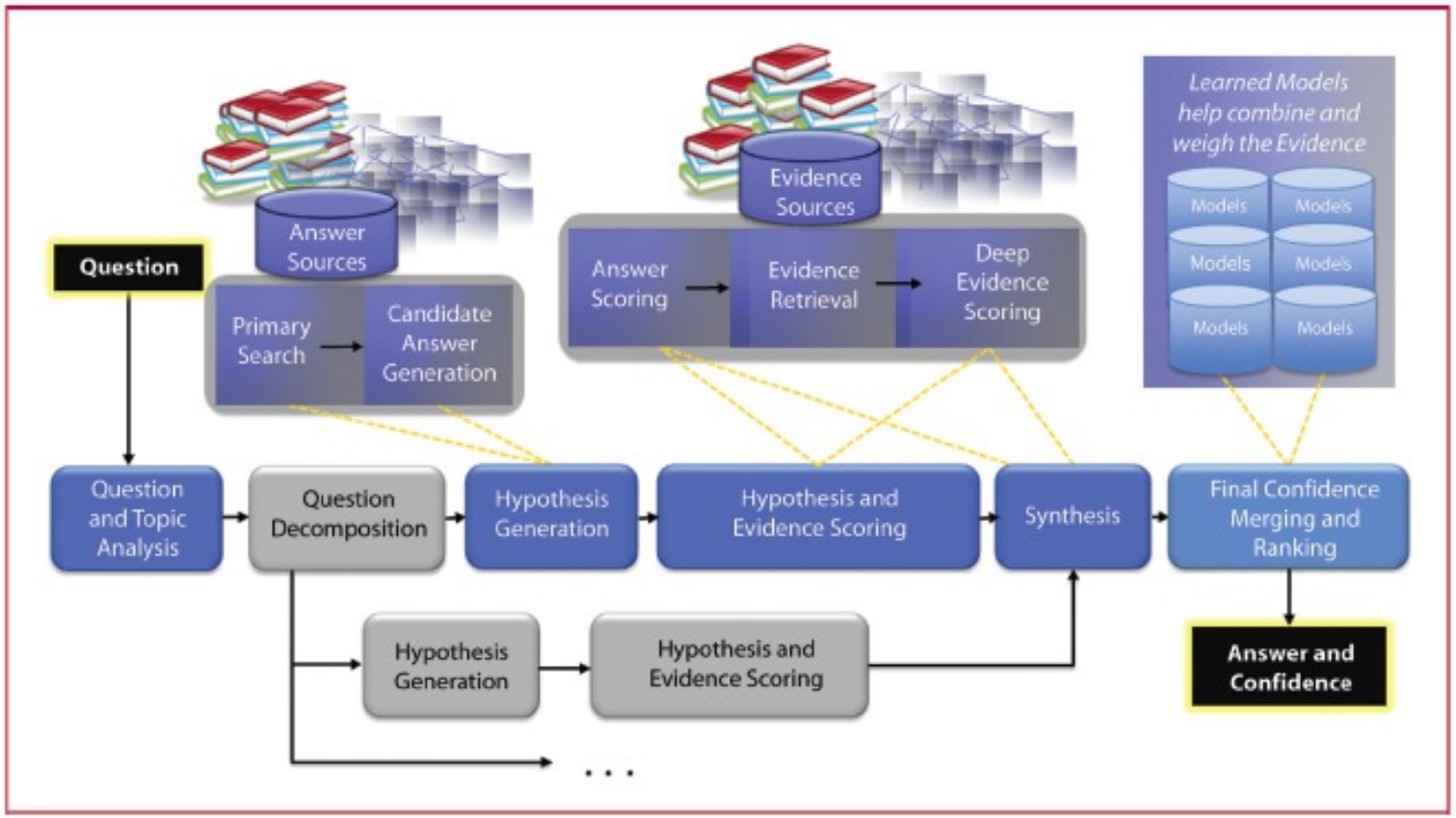


Figure 1

DeepQA architecture.

[Source: Ferrucci et al., This is Watson, IBM J. RES. & DEV. VOL. 56 NO. 3/4 PAPER 1]

Final Confidence Merging and Ranking:

- evaluate hundreds of hypotheses (hundreds of thousands scores)
- identify best-supported answer, estimate confidence in correctness

Final Confidence Merging and Ranking

- **Answer Merging**
  - candidate answers may be equivalent despite different surface forms
  - example: Abraham Lincoln and Honest Abe
  - scores need to be combined
- **Ranking and Confidence Estimation**

Final Confidence Merging and Ranking

- **Answer Merging**
- **Ranking and Confidence Estimation**
  - machine-learning approach
    (system trained with questions with known answers)
  - different phases, hierarchically structured
  - Watson's meta-learner uses multiple trained models to handle different question classes
  - e.g. certain scores that may be crucial to identifying the correct answer for a factoid question may not be as useful on puzzle questions

Watson — Final Confidence Merging and Ranking

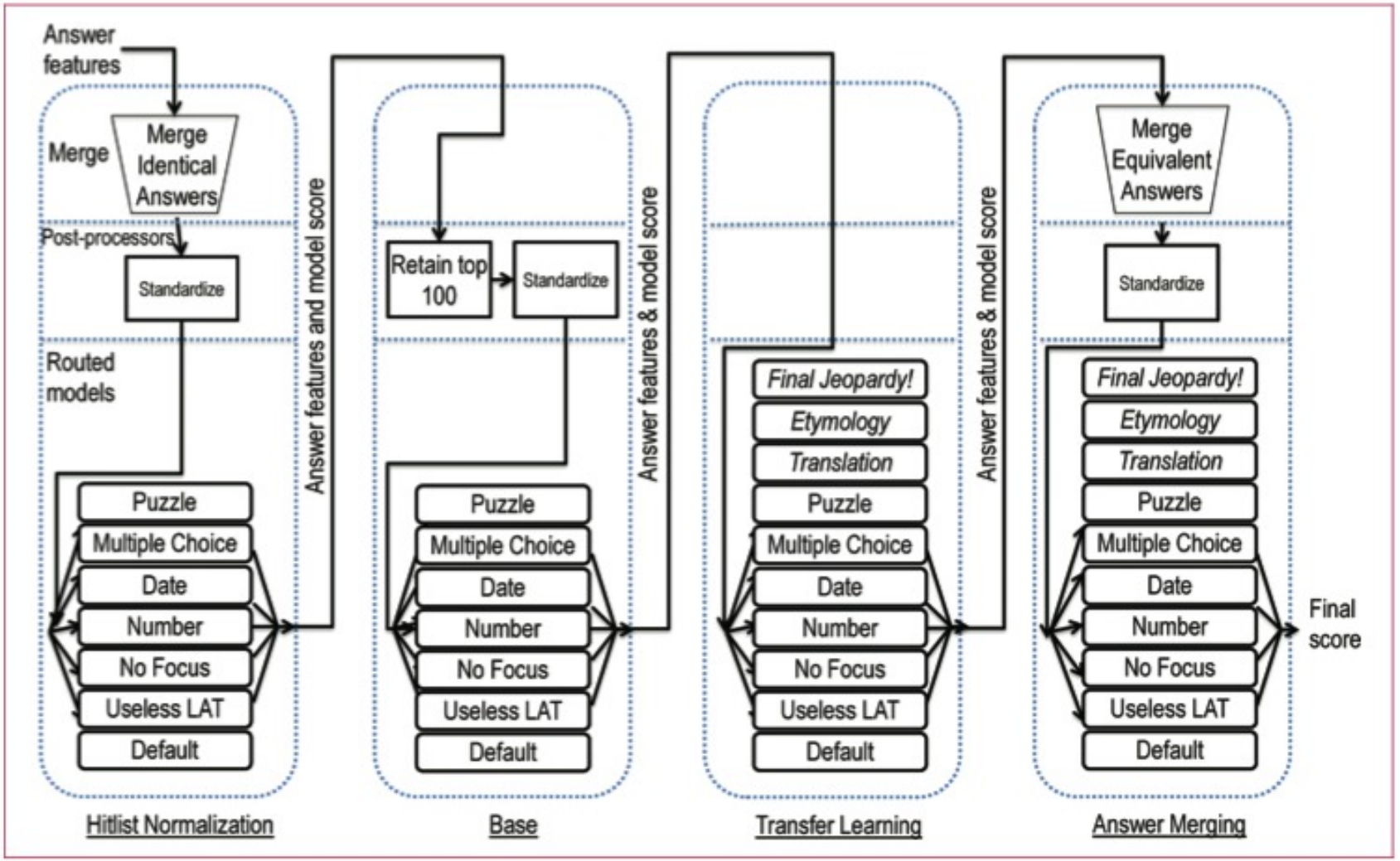


Figure 1

First four phases of merging and ranking in DeepQA.

[Source: Ferrucci et al., This is Watson, IBM J. RES. & DEV. VOL. 56 NO. 3/4 PAPER 14]

Watson — Now Watson can bet on its answer!

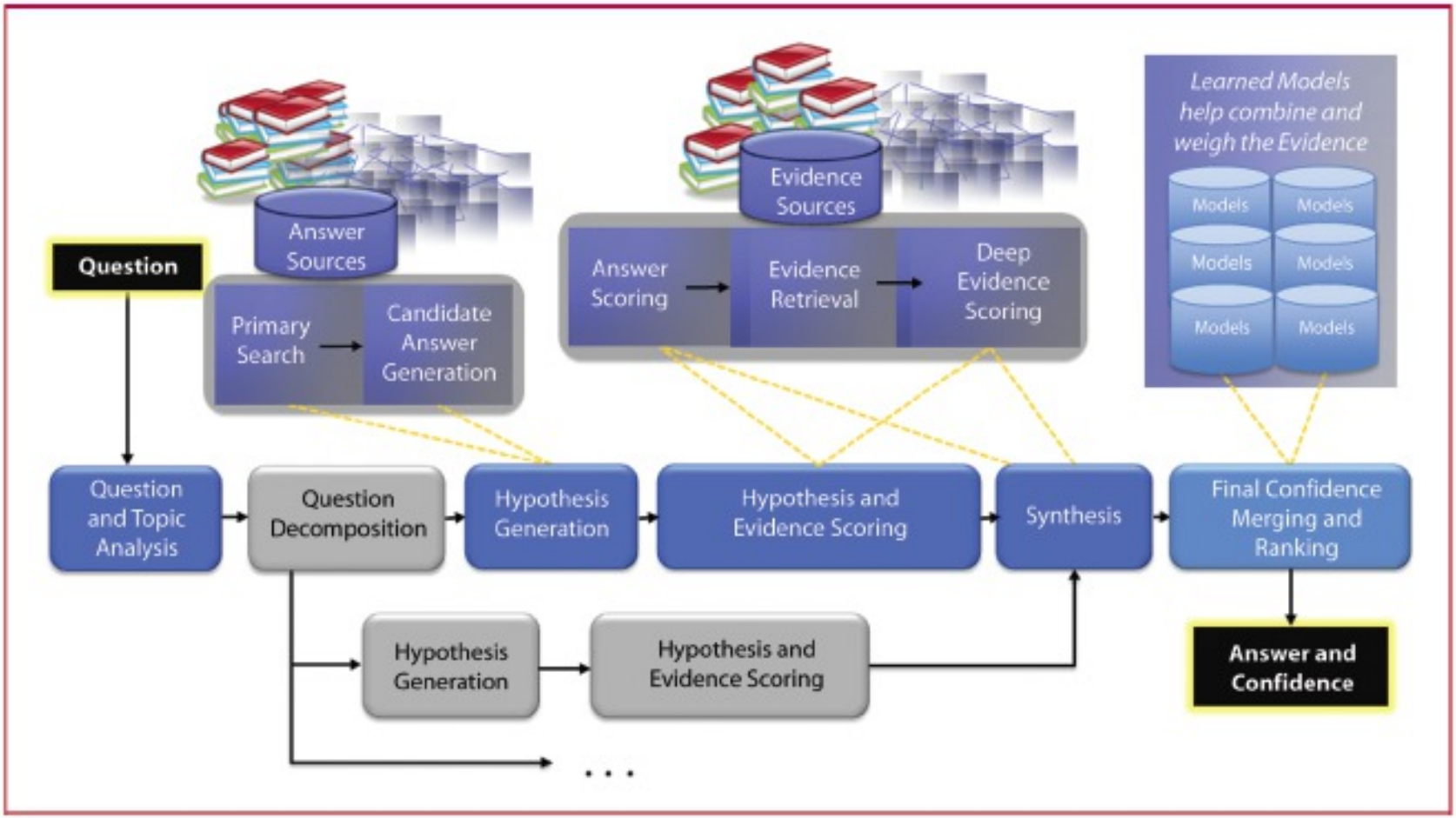


Figure 1

DeepQA architecture.

[Source: Ferrucci et al., This is Watson, IBM J. RES. & DEV. VOL. 56 NO. 3/4 PAPER 1]

Watson — Betting Strategy

Strategic game play is required

- When to attempt an answer (buzz in)?
- What squares to select?
- Wagering on Daily Doubles
- Wagering in Final Jeopardy

Different techniques:
simulation (millions of simulated matches were played), game playing, Bayesian inference, machine-learning, Monte Carlo methods, ...

Watson — needs to be very fast!

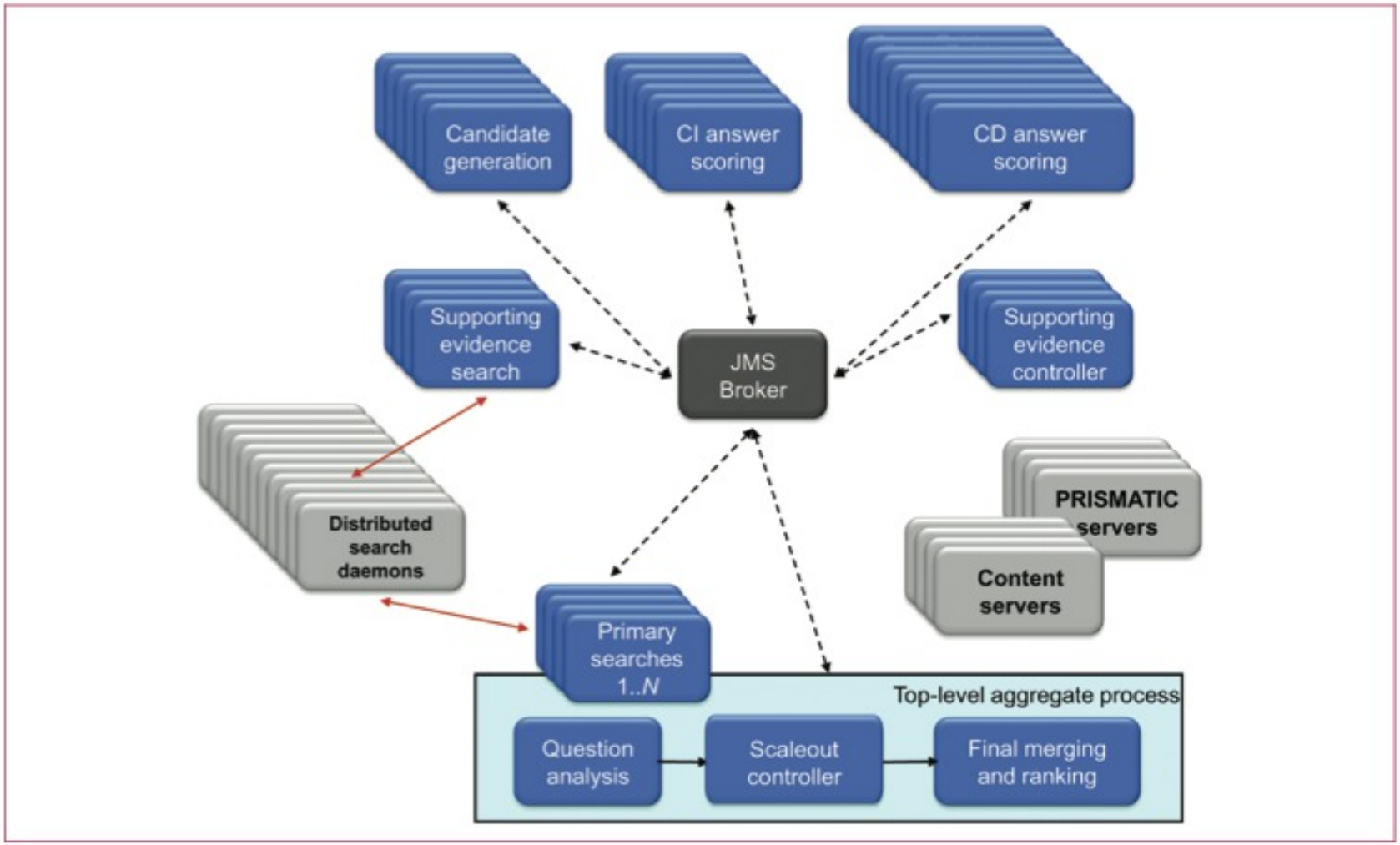


Figure 2

Major components in the Jeopardy! system.

[Source: Ferrucci et al., This is Watson, IBM J. RES. & DEV. VOL. 56 NO. 3/4 PAPER 15]

Watson — Possible Application Directions

Medicine:

- assist physicians in diagnosing and treatment of patients

Big Data:

- assist CEO's in decision making
- assist in controlling the financial sector
- ...

What else?

- well, just think about Prism and Tempora

Watson — Possible Application Directions

Medicine:

- assist physicians in diagnosing and treatment of patients

Big Data:

- assist CEO's in decision making
- assist in controlling the financial sector
- ...

What else?

- well, just think about Prism and Tempora

Watson — Possible Application Directions

Medicine:

- assist physicians in diagnosing and treatment of patients

Big Data:

- assist CEO's in decision making
- assist in controlling the financial sector
- ...

What else?

- well, just think about Prism and Tempora

Related Systems

The following systems are related
(their focus is more on structured, formal knowledge sources):

- Wolfram Alpha: http://www.wolframalpha.com
- Evi (formerly TrueKnowledge): http://www.evi.com

Related Systems

Discussion

Is Watson intelligent?