CSCI5070 Advanced Topics in Social Computing

QA and Deep QA

Irwin King

The Chinese University of Hong Kong

king@cse.cuhk.edu.hk

Outline

- Question Answering
  - Background
  - Traditional QA
    - General Overview
    - Knowledge Mining
    - Knowledge Annotation
  - An Example: Factoid Question Answering
- DeepQA
  - Architecture
  - Examples

QUESTION ANSWERING

Background

- Question Answering (QA) systems
  - Increasingly popular.
  - Why?
    - General Search Engine
      - A list of documents or Web pages.
    - Question Answering System
      - Deliver users short, succinct answers.
      - Intuitive information access.
      - Just the right information.

Examples

START's reply

===> What's the largest city in Florida?

Florida

Largest Cities in Florida: Jacksonville (672,971); Miami (358,548); Tampa (280,015); St. Petersburg (238,629); Hialeah (188,004): Orlando (164,693).

Source: WorldBook

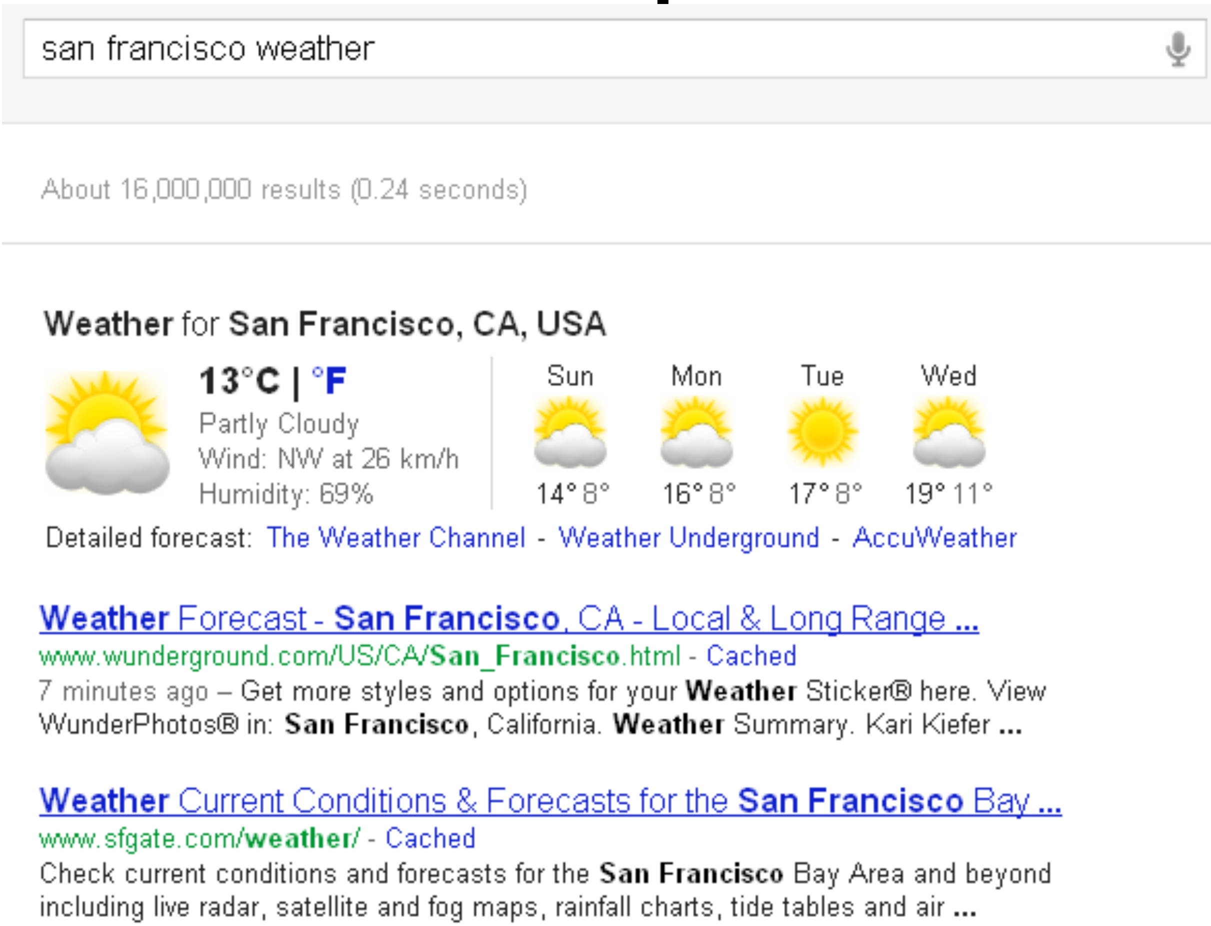
Florida

Largest city: Jacksonville, Miami, Tampa, Saint Petersburg, Hialeah, Orlando, Fort Lauderdale, Tallahassee, Hollywood, Pembroke Pines

Source: 50States.com

START: Natural Language Question Answering System
http://start.csail.mit.edu/

Examples



Google OneBox:
http://googlesystem.blogspot.com/2006/07/google-onebox-results.html

Traditional QA: General Overview

- Two Axes of Exploration
  - Nature of the information.
    - What type of information is the system utilizing to answer natural language questions?
      - Structured Knowledge (Databases)
      - Unstructured Knowledge (Free text)
  - Nature of the technique.
    - How linguistic sophisticated are the techniques employed to answer natural language questions?
      - Linguistically Sophisticated (e.g. syntactic parsing)
      - Linguistically Uninformed (e.g. n-gram generation)

Traditional QA: General Overview

- Short history:
  - Natural language interfaces to databases.
  - Blocks world.
  - Plans and scripts.
  - Modern question answering systems.

Traditional QA: General Overview

- Natural language interfaces to databases (NLIDBs)
  - Map natural languages to SQL queries correctly.
  - Nontechnical people access information.
- Examples
  - BASEBALL – baseball statistics
    - Who did the Red Sox lose to on July 5?
  - LUNAR-analysis of lunar rocks
    - What is the average concentration of aluminum in high alkali rocks?
  - LIFER-personnel statistics
    - How many professors are there in computer science department?
- Limitations
  - Domain knowledge was embedded implicitly in the parser.
  - Users were presumed to have some knowledge of underlying data tables.

Traditional QA: General Overview

- Blocks World
  - Interaction with a robotic arm in a world filled with colored blocks.
- Example
  - Person: Pick up a big red block.
    Computer: OK.
    Person: Grasp the pyramid.
    Computer: I don't understand which pyramid you mean.
    Person (changing their mind): Find a block which is taller than the one you are holding and put it into the box.
    Computer: By "it", I assume you mean the block which is taller than the one I am holding.
    Computer: OK.
    Person: What does the box contain?
    Computer: The blue pyramid and the blue block.
    Person: What is the pyramid supported by?
    Computer: The box.
    Person: How many blocks are not in the box?
    Computer: Four of them.
    Person: Is at least one of them narrower than the one which I told you to pick up?
    Computer: Yes, the red cube.
    etc.

Traditional QA: General Overview

- Plans and Scripts
  - QUALM
    - Application of scripts and plans for story comprehension.
    - Very restrictive domain.
  - Unix Consultant
    - Allow users to interact with UNIX.
    - e.g. "How do I delete a file?"
    - Paradigm not suitable for general purpose question answering

Traditional QA: General Overview

- Before the Web:
  - Limited audience.
  - Knowledge had to be hand-coded and specially prepared.
- START:
  - The first QA system for the World Wide Web.
  - Online and continuous operating since 1993.
  - Engages in "virtual collaboration" by utilizing knowledge freely available on the Web.
  - http://www.ai.mit.edu/projects/infolab

Traditional QA: General Overview

- START:

START's reply

===> where is UC Berkeley

University of California Berkeley

Address:

110 Sproul Hall
Berkeley, CA 94720-5800

Source: U.S.News

- Go back to the START dialog window.

Traditional QA: General Overview

- Recent QA systems are based on information retrieval and information extraction:
  - Large-scale evaluations began with the TREC QA tracks.

- TREC-8 QA Track [Voorhees and Tice 1999,2000b]
  - 200 questions: backformulations of the corpus
  - Systems could return up to five answers

    answer = [answer string, docid]
  - Two test conditions: 50-byte or 250-byte answer strings
  - MRR scoring metric
- TREC-9 QA Track [Voorhees and Tice 2000a]
  - 693 questions: from search engine logs
  - Systems could return up to five answers

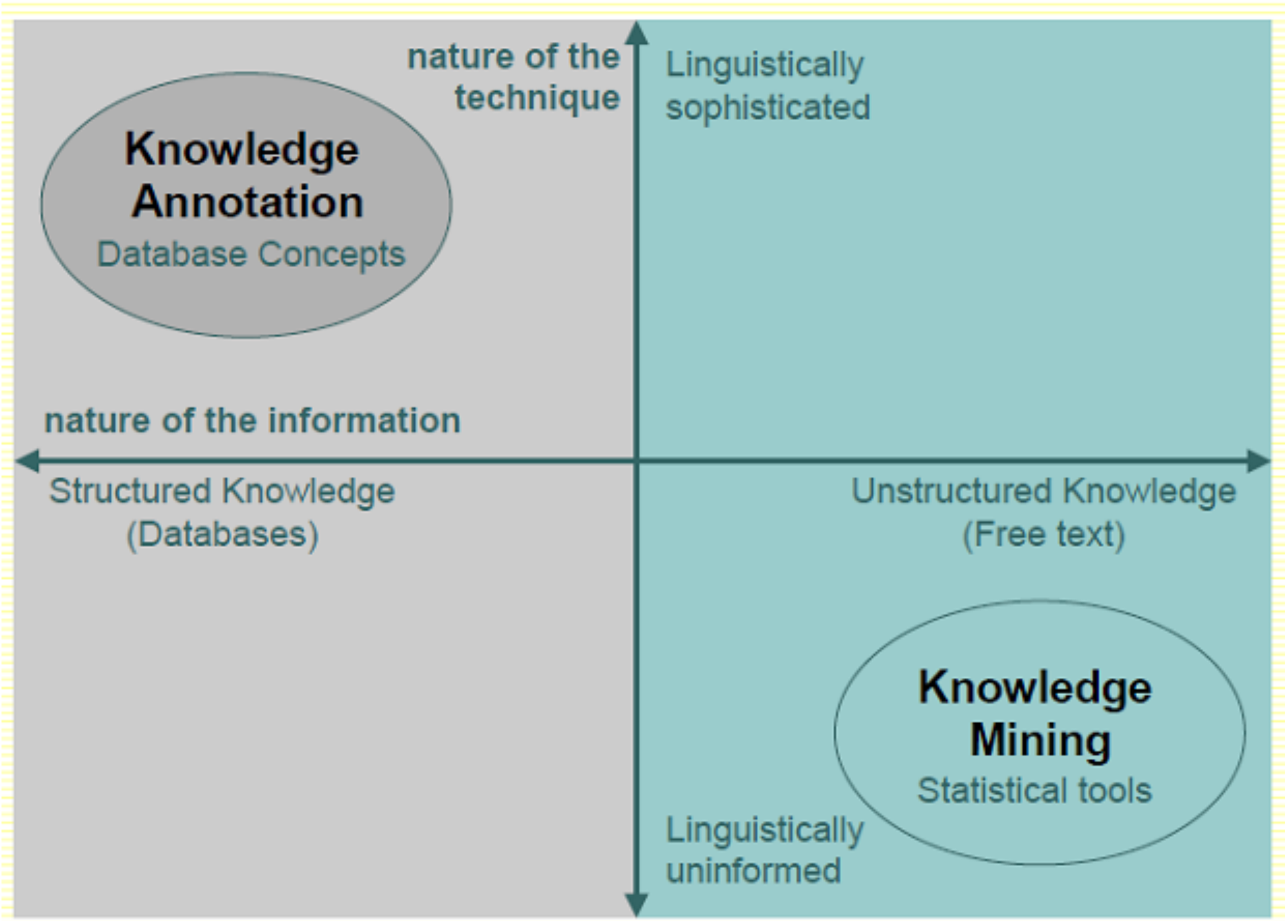
    answer = [answer string, docid]
  - Two test conditions: 50-byte or 250-byte answer strings
  - MRR scoring metric
- TREC 2001 QA Track [Voorhees 2001,2002a]
  - 500 questions: from search engine logs
  - Systems could return up to five answers

    answer = [answer string, docid]
  - 50-byte answers only
  - Approximately a quarter of the questions were definition questions (unintentional)
- TREC 2002 QA Track [Voorhees 2002b]
  - 500 questions: from search engine logs
  - Each system could only return one answer per question

    answer = [exact answer string, docid]
  - All answers were sorted by decreasing confidence
  - Introduction of "exact answers" and CWS metric

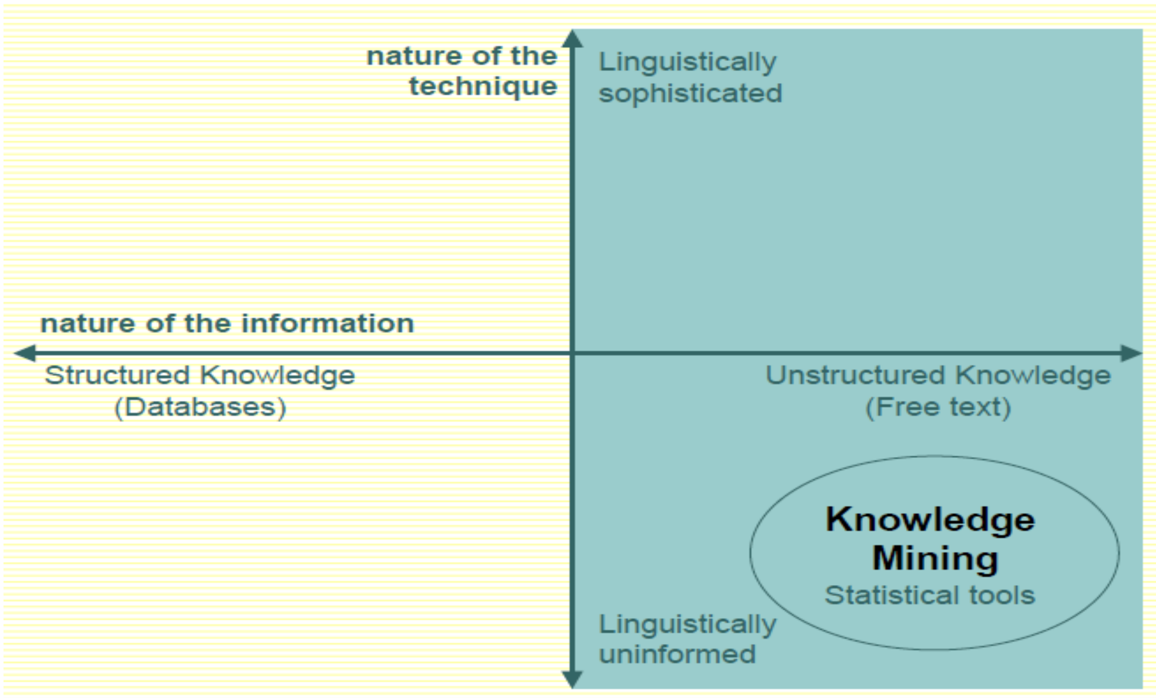
Traditional QA: General Overview

- Two Techniques for traditional QA



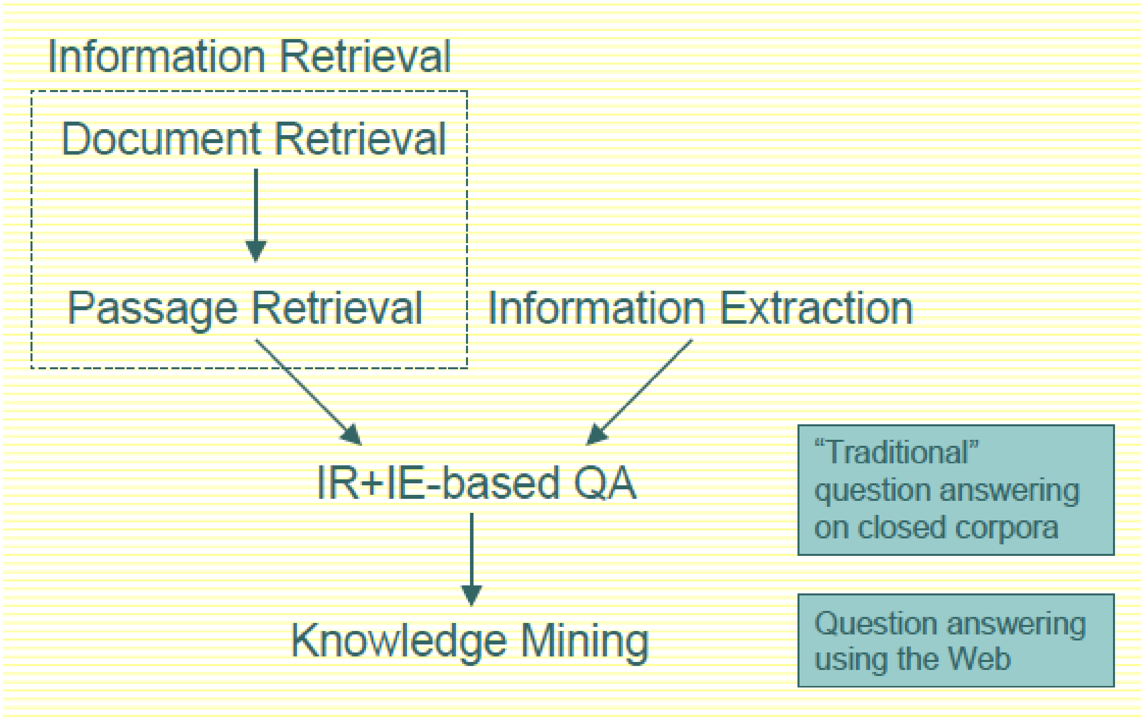
Traditional QA: Knowledge Mining

- Definition
  - Techniques that effectively employ unstructured text on the Web for QA.



Traditional QA: Knowledge Mining

- Framework



Traditional QA: Knowledge Mining

- Ways of using the Web
  - Use the Web as the primary corpus of information.
    - Project answers onto another corpus for verification purpose.
  - Combine use of the Web with other corpora.
    - Employ Web data to supplement a primary corpus.
    - Use the Web only for some questions.
    - Combine Web and non-Web answers.

Traditional QA: Knowledge Mining

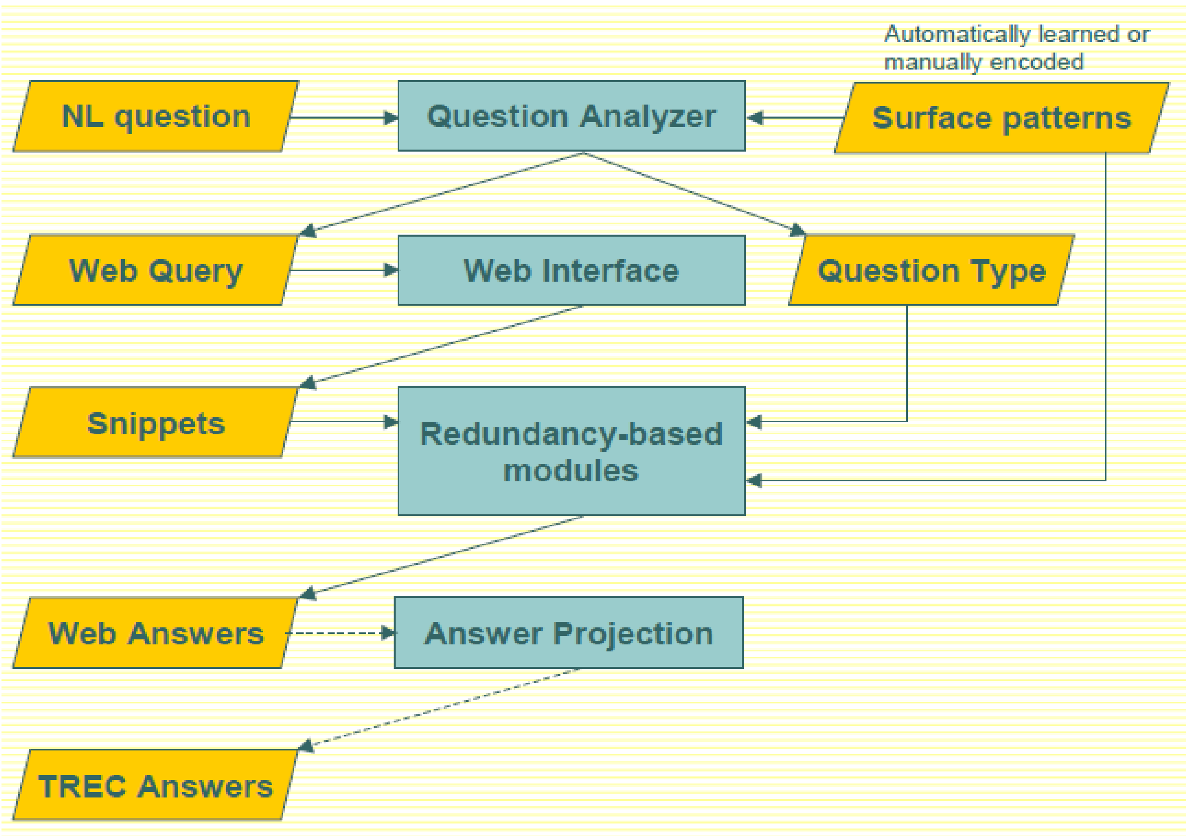
- Issues about Web data redundancy
  - Surrogate for sophisticated NLP.
    - E.g.:
      - Q: Who killed Abraham Lincoln?
      - A1: John Wilkes Booth killed Abraham Lincoln.
      - A2: John Wilkes Booth altered history with a bullet. He will forever be known as a man who ended Abraham Lincoln's life.
  - Overcome poor document quality.
    - E.g.:
      - Q: What is the furthest planet in the Solar System?
      - A1: Blah Pluto blah blah Planet X blah blah
      - A2: Blah Planet X blah blah Pluto

Traditional QA: Knowledge Mining

- Finding Answers
  - Match answers using surface patterns
    - Regular expression.
    - Bypass linguistically sophisticated techniques.
  - Rely on statistics and data redundancy
    - Many occurrence of the answers.
    - Develop techniques for filtering, sorting large number of candidates.

Traditional QA: Knowledge Mining

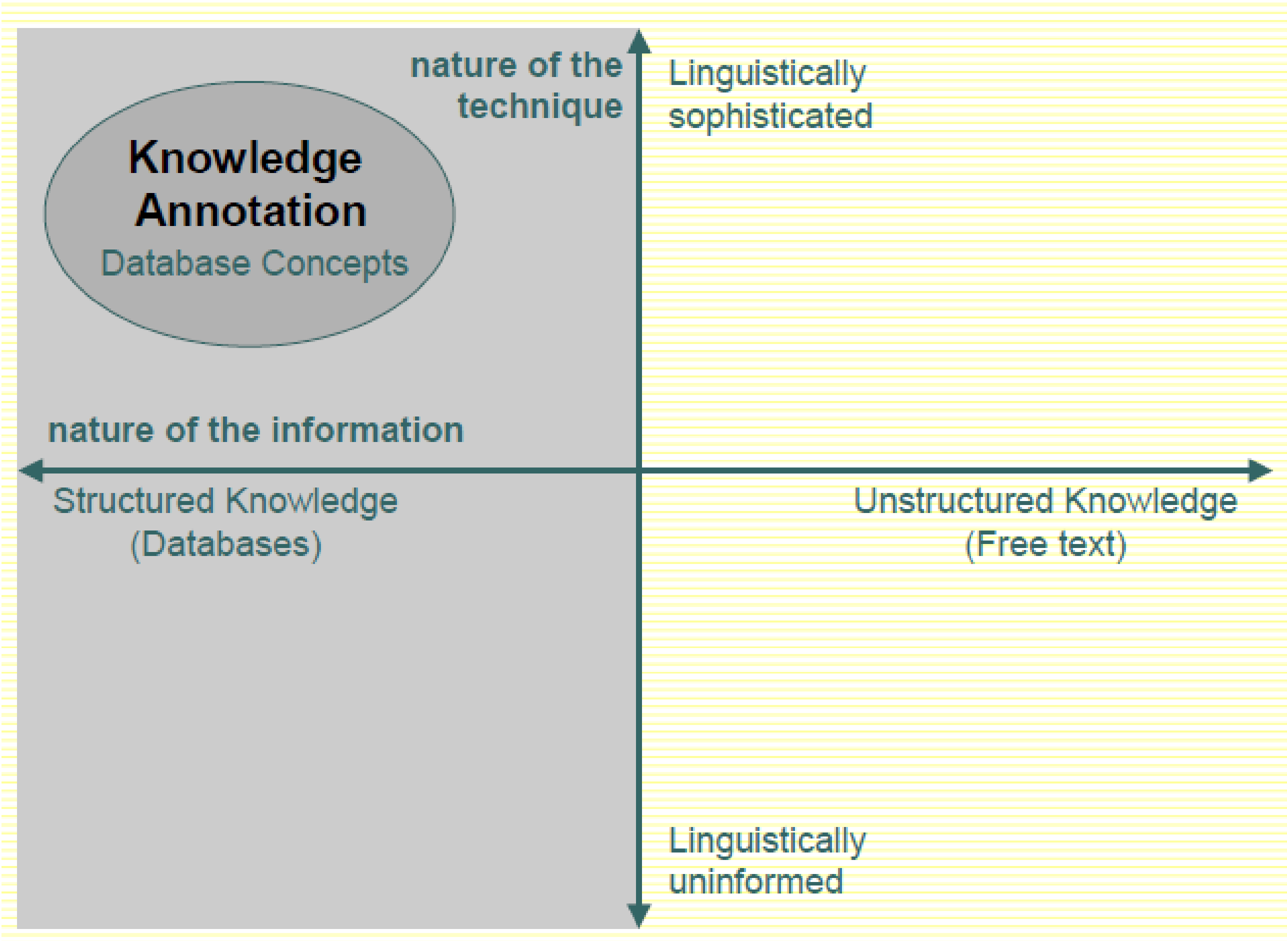
- Detailed System Architecture



Traditional QA: Knowledge Mining

- Use the Web to Perform Answer Validation
  - Compute a continuous function that takes both the question and answer as input
  - f(question,answer)=x
  - if x>threshold, then answer is valid, otherwise, answer is invalid.
- Answer validation functions
  - Pointwise Mutual Information (PMI)
  - Maximum Likelihood Ratio (MLHR)
  - Corrected Conditional Probability (CCP)

Traditional QA: Knowledge Annotation



Traditional QA: Knowledge Annotation

- Approach
  - Structured or semistructured resources on the Web.
  - Organize them to provide convenient methods for access.
  - Annotate these resources with metadata.
  - Connect these annotated resources with natural language to provide question answering capabilities.

Traditional QA: Knowledge Annotation

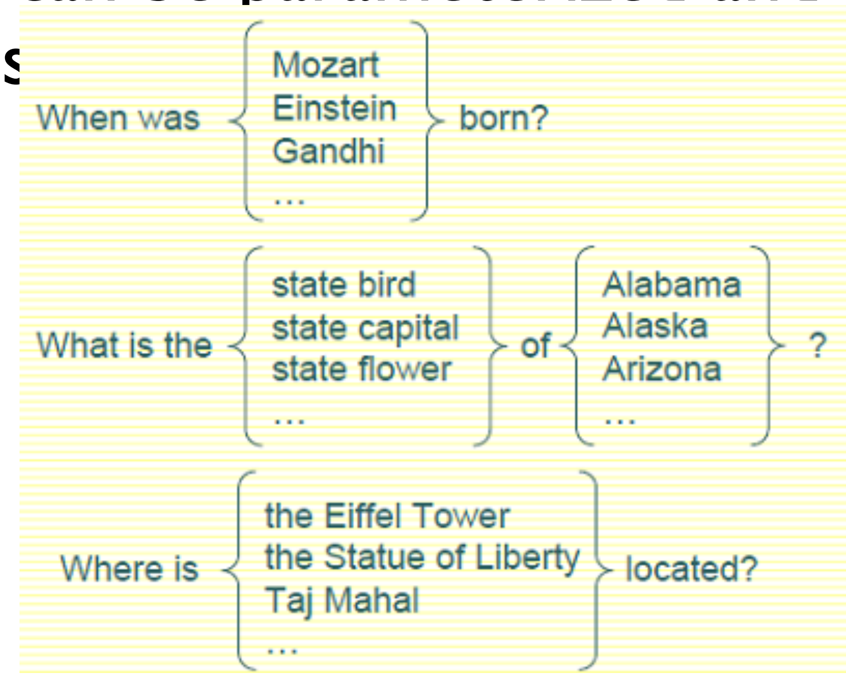
- Why we need knowledge annotation
  - The Web contains many databases.
  - "Hidden" or "deep" Web.
  - Improve question-answering quality.

Traditional QA: Knowledge Annotation

- Examples
  - Internet Movie Database (IMDB)
    - Content: cast, crew and other movie-related information.
  - CIA World Factbook
    - Content: geographic, political, demographic, and economic information.
  - Biography.com
    - Content: short biographies of famous people.
  - Wikipedia
    - Content: all kinds of human knowledge.

Traditional QA: Knowledge Annotation

- Zipf's Law of QA
  – A few "question types" account for a large portion of all question instances.
  – Similar questions can be parameterized and grouped into ques



Traditional QA: Knowledge Annotation

- Ways to access structured and semistructured Web resources
  - Wrap
    - screen scraping.
    - Provide programmatic access to Web resources. (API)
    - Retrieve results dynamically.
  - Slurp
    - "Vacuum" out information from Web resources.
    - Restructure information in a local database.

Traditional QA: Knowledge Annotation

- Wrap
  - Advantages
    - Information is up-to-date.
    - Dynamic information is easy to access.
  - Disadvantages
    - Queries are limited in expressiveness.
    - Reliability issue.
    - Wrapper maintenance: sources change layout.

Traditional QA: Knowledge Annotation

- Slurp
  - Advantages
    - Queries can be arbitrarily expressive.
    - Information is always available.
  - Disadvantages
    - Stale data problem. Original source is updated?
    - Dynamic data problem. Stock data?
    - Resource limitation. Too much data to store locally.

Traditional QA: Knowledge Annotation

- Examples of Knowledge Annotation Systems:
  - START
  - AskJeeves
  - FAQ Finder (U. Chicago)
  - Aranea (MIT)
  - KSP (IBM)
  - Early Answering (U. Waterloo)

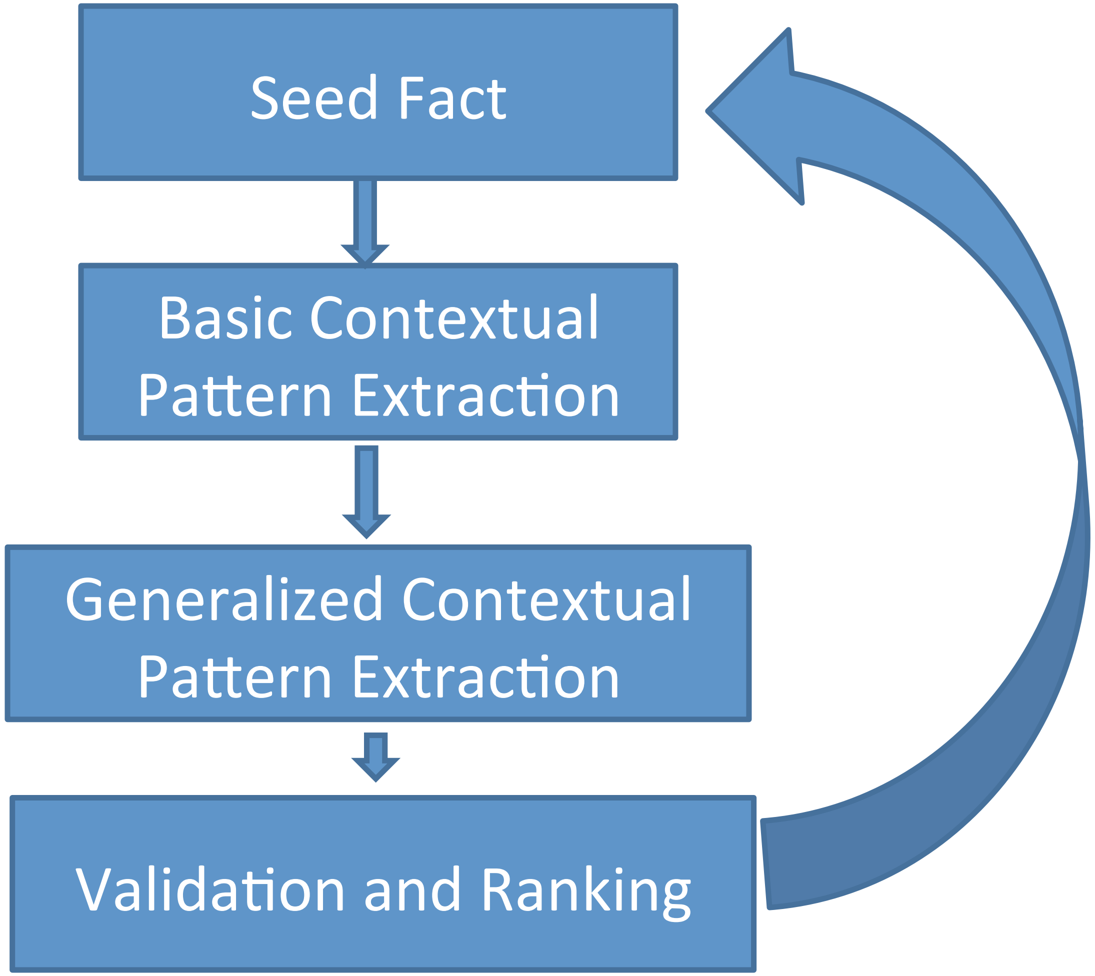
An Example: Factoid Question Answering

- Factoid questions
  - Questions that have a short answer (typically a noun phrase or a simple verb phrase) (Agichtein, Cucerzan and Brill, 2005)
    - How tall is mount Everest?
    - How old was Leonardo da Vinci when he painted Mona Lisa?
  - Fact extraction
    - Gathering collections of reliable fact tables
      - Person-BornIn-Year
      - City-CapitalOf-Country

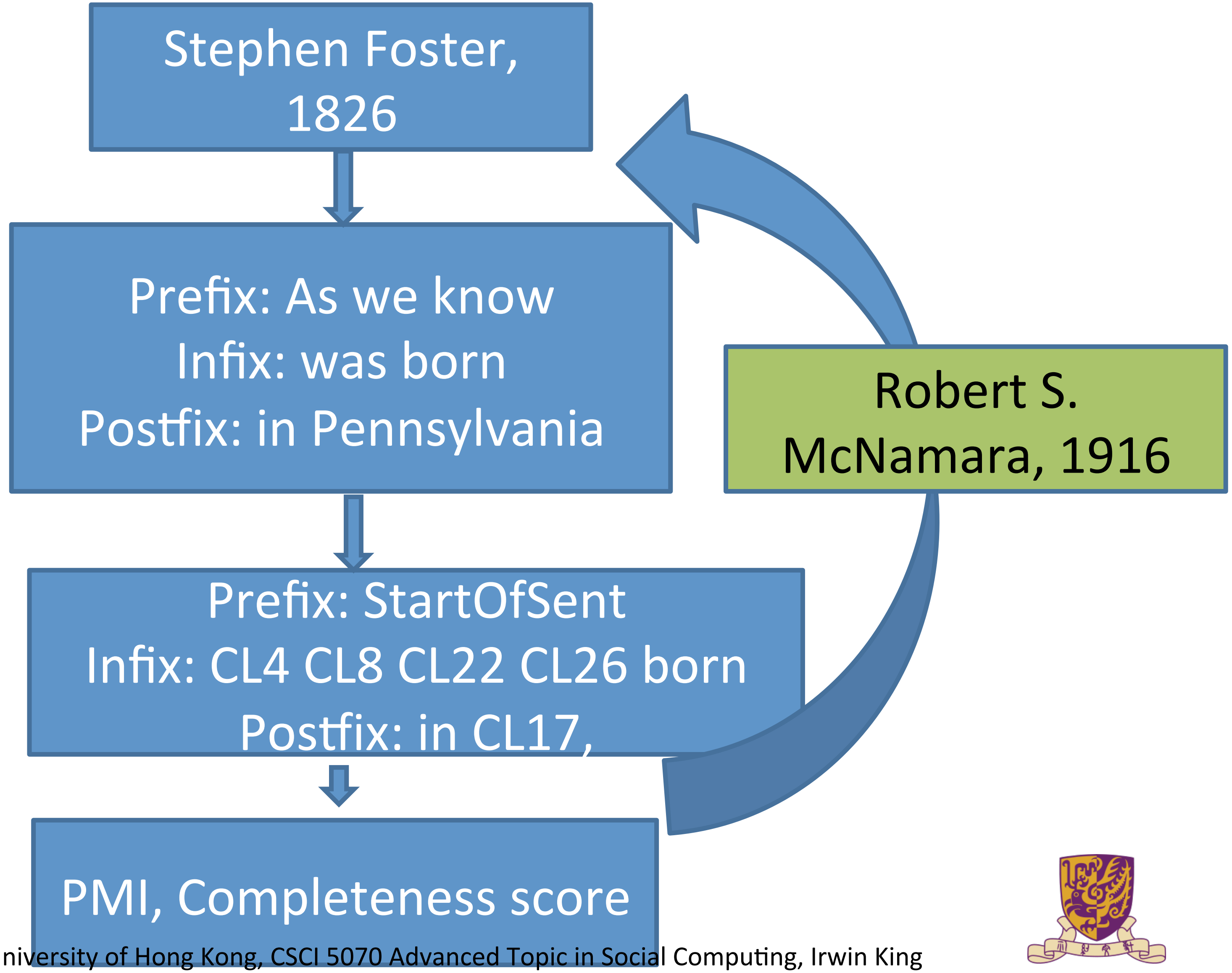
Fact Extraction

- Motivation
  - Web as a decentralized repository of human knowledge remains largely untapped during Web search due to the inherent difficulty of representing and extracting knowledge from noisy natural-language text
  - Access to binary relations among named entities enable new search paradigms
  - Pasca, Lin, Bigham, Lifchits and Jain, 2006

Fact Extraction



Fact Extraction



Fact Extraction

- Seed Fact
  - Pair of phrases in a "hidden" relation
  - (Vincenzo Bellini, 1801), Person-BornIn-Year
  - (Athens, Greece), City-CapitalOf-Country

Fact Extraction

- Basic Contextual Extraction Pattern
  - Each occurrence of the two sides of the fact within the same sentence produces a basic contextual extraction pattern
    - (Prefix, Infix, Postfix)
    - Prefix and postfix are contiguous sequences of a fixed number of terms, the immediate left of the first matched phrase, the immediate right of the second matched phrase
    - Infix contains all terms between two matched phrases

Fact Extraction

- Basic Contextual Extraction Pattern
  - Fact: (Athens, Greece)
  - Prefix: […take students to], Infix: [, the capital of], Postfix: [and home to…]
  - Algorithm
    - Modified trie, both phrases of the seed facts are loaded into the trie, each sentence is then matched onto the trie
    - Parallelized, using MapReduce (Dean and Ghemawat, 2004)

Fact Extraction

- Generalized Contextual Pattern Extraction
  - Terms in prefix, infix and postfix in each basic pattern are replaced with their corresponding classes of distributed similar words
  - Extraction the set of distributionally similar words (Lin 1998)
  - The set of generalized patterns is smaller than the set of basic patterns, but with higher coverage

Fact Extraction

Prefix	Infix	Postfix
CL3 00th :	's Birthday (	) . EndOfSent
StartOfSent	CL4 CL8 CL22 CL26 born	in CL17 ,
Memorial CL47 in	(b. CL3 0 ,	, d. CL3
among CL6 ...	CL4 born on 00 CL3	in CL10 ,
CL8 child :	CL4 born 00 CL3	in Lewisburg ,
CL4 written by	who CL4 born CL3 00 ,	, in Oak

CL3 = {March, October, April, Mar, Aug., February, Jul, Nov., ...}
CL4 = {is, was, has, does, could}
CL6 = {You, Lawmakers, Everyone, Nobody, Participants, ...}
CL8 = {a, the, an, each, such, another, this, three, four, its, most, ...}
CL10 = {Pennsylvania, Denver, Oxford, Marquette, Hartford, ...}
CL17 = {Tipperary, Rennes, Piacenza, Osasuna, Dublin, Crewe, ...}
CL22 = {Brazilian, Chinese, Japanese, Italian, Pakistani, Latin, ...}
CL26 = {entrepreneur, illustrator, artist, writer, sculptor, chef, ...}
.....
CL47 = {Tribute, Homage}
.....

Examples of generalized patterns acquired during the extraction of Person-BornIn-Year facts. A digit is represented by a 0.

Fact Extraction

- Validation and Ranking
  - Candidate facts, similarity scores that aggregate individual word-to-word similarity scores of the component words relative to the seed facts
  - A linear combination of features
  - PMI-inspired score, (Turney 2001)
  - Completeness score, demote candidate facts if any of their two sides are likely to be incomplete
    - E.g. Mary Lou vs. Mary Lou Retton, John F. vs. John F. Kennedy

Fact Extraction

- Scalable
  - Person-BornIn-Year
  - 10 seed facts
  - Expands the initial seed set of 10 facts to 100,000 facts (after iteration 1) and then to one million facts (after iteration 2)

DEEP QA

The *Jeopardy!* Challenge

Broad/Open Domain

Complex Language

High Precision

Accurate Confidence

High Speed

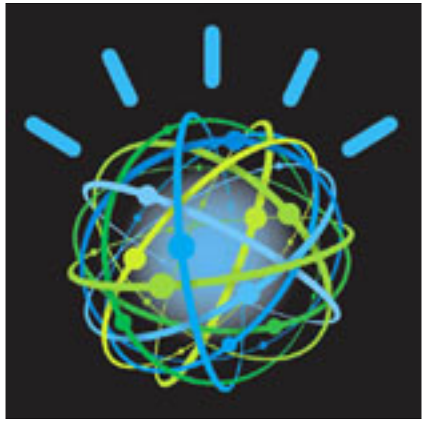
$200
If you're standing, it's the direction you should look to check out the wainscoting.

$800
In cell division, mitosis splits the nucleus & cytokinesis splits this liquid *cushioning* the nucleus

$1000
Of the 4 countries in the world that the U.S. does not have diplomatic relations with, the one that's farthest north

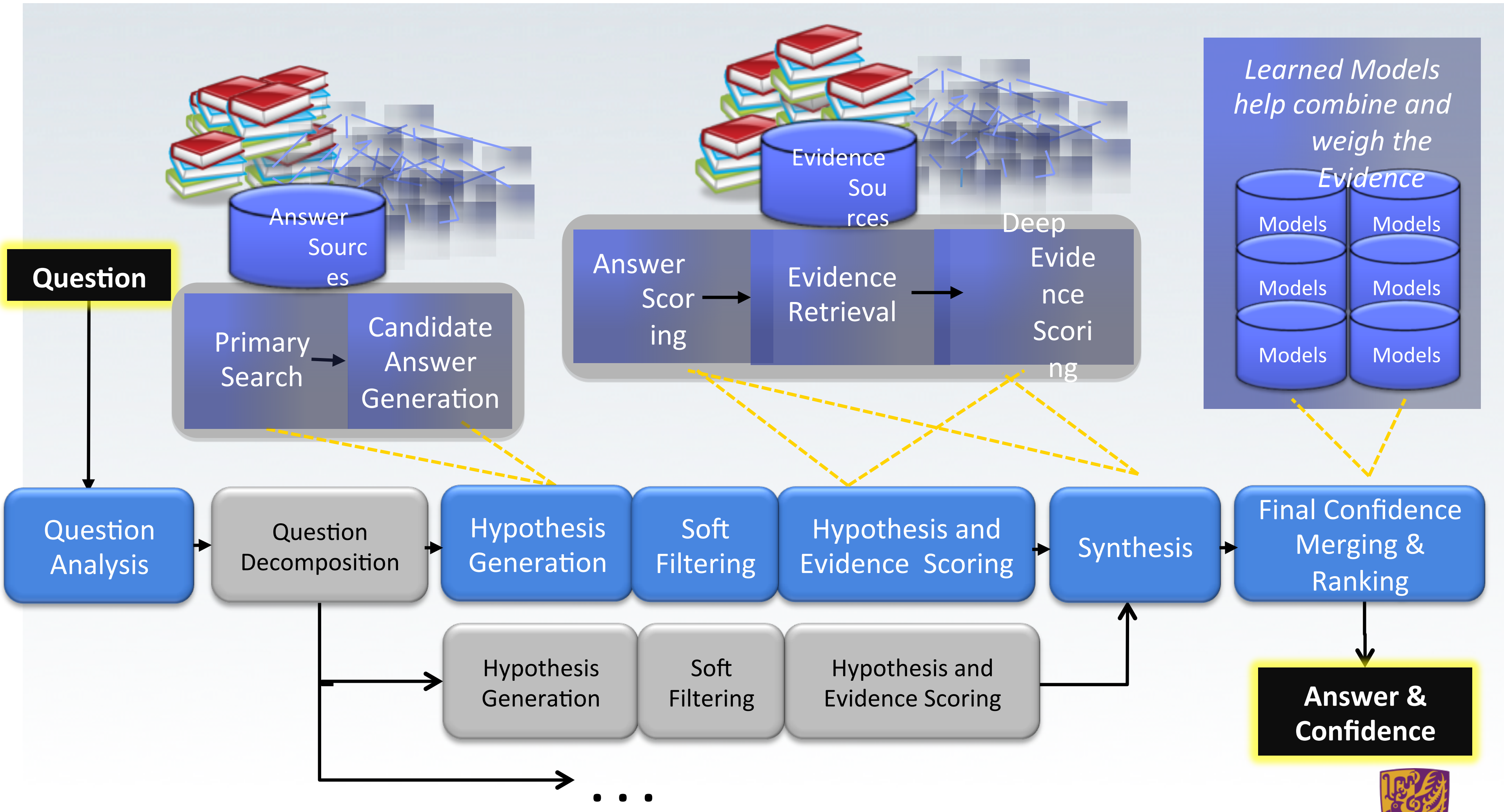
DeepQA

- A massively parallel probabilistic evidence-based architecture
- Four principles
  - Massive parallelism
  - Many experts
  - Pervasive confidence estimation
  - Integrate shallow and deep knowledge



- Watson
  - The implementation of DeepQA by a research team in IBM
  - Beat human on the quiz show *Jeopardy* in 2011

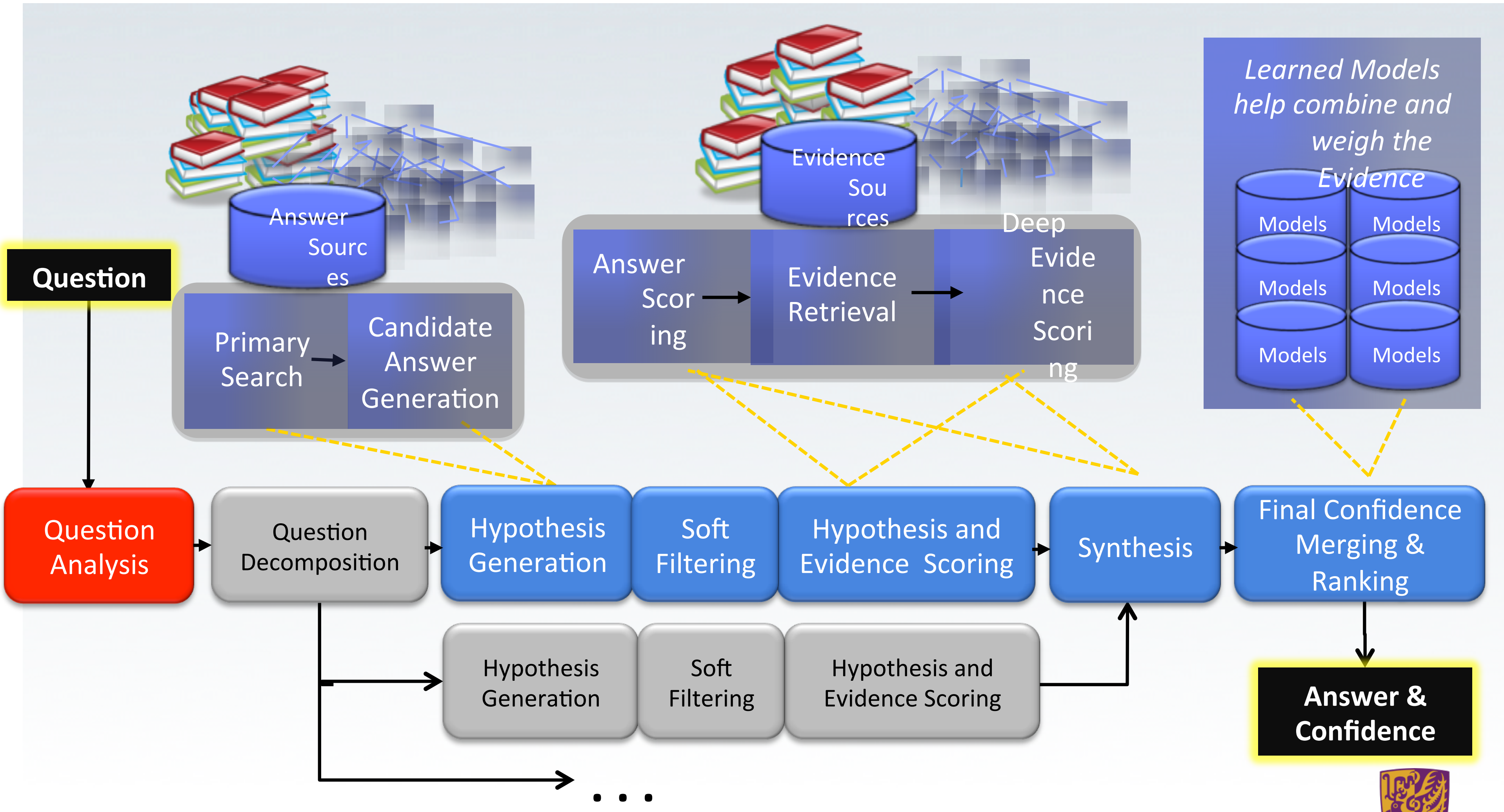
High-Level Architecture



Content Acquisition

- Identify and gather the content to use for the answer and evidence sources
- Sources
  - Encyclopedias, Dictionaries, Thesauri, Newswire articles, etc.
- Corpus expansion
  - Identify seed documents and retrieve related documents from the web
  - Extract self-contained text nuggets
  - Score the nuggets based on their informativeness to the seed documents
  - Merge the most informative nuggets into the corpus

High-Level Architecture



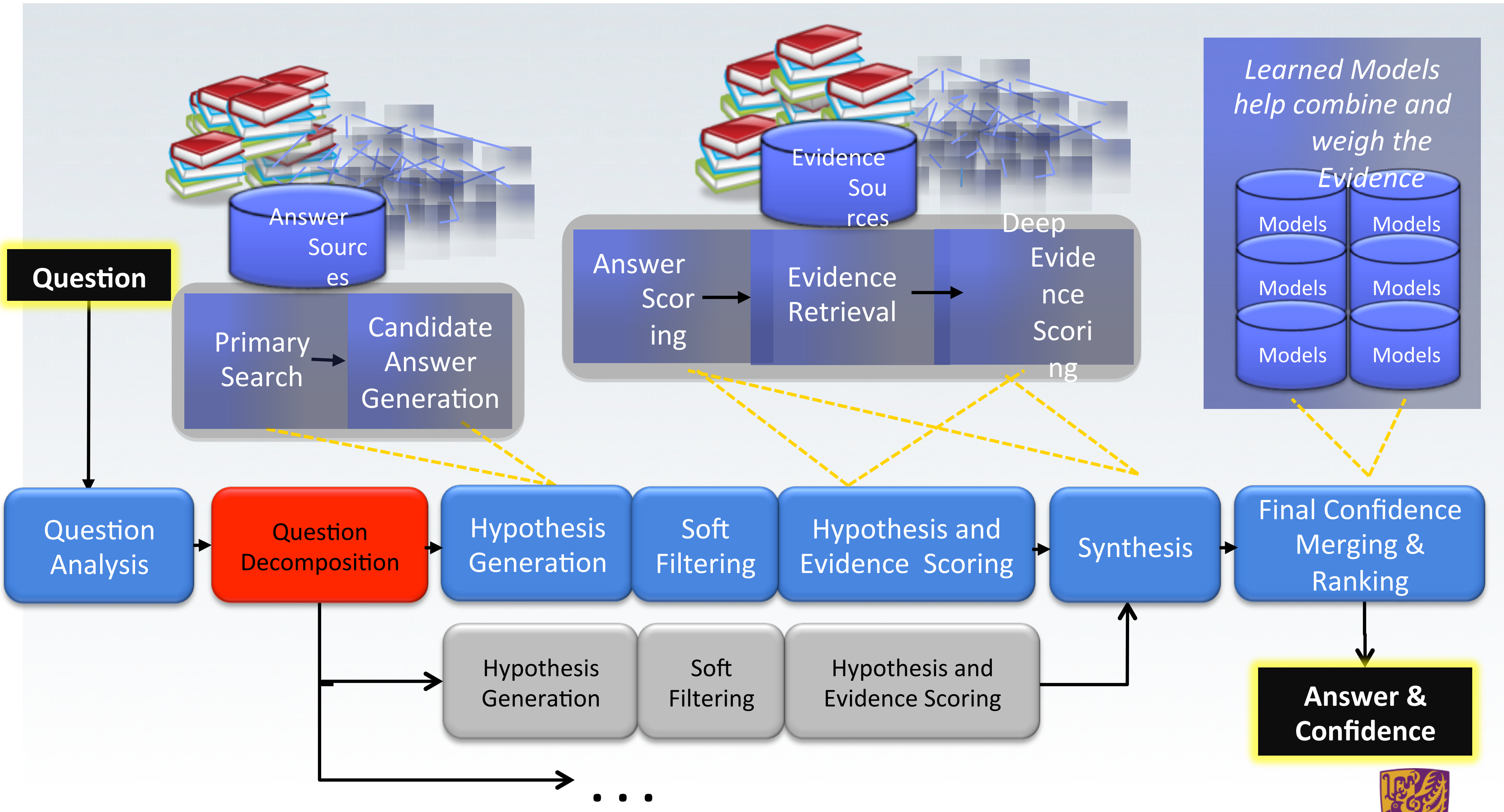
Question Analysis

- Question classification
  - Identify puns, constraints, definition components, entire subclues within questions
  - Question type: puzzle, math, definition, etc.
- Lexical Answer Type (LAT) detection
  - LAT: a word or noun phrase that specifies the type of the answer without any attempt to understand its semantics.
    - E.g., the LAT of the clue "Invented in the 1500s to speed up the game, this maneuver involves two pieces of the same color" is "maneuver".

Question Analysis (cont.)

- Focus detection
  - The focus is the part of the question that, if replaced by the answer, makes the question a stand-alone statement
    - The focus of "When hit by electrons, a phosphor gives off electromagnetic energy in this form" is "this form"
- Relation detection
  - syntactic subject-verb-object predicates or semantic relationships between entities
    - "They're the two states you could be reentering if you're crossing Florida's northern border"
    - Relation: borders(Florida,?x,north)

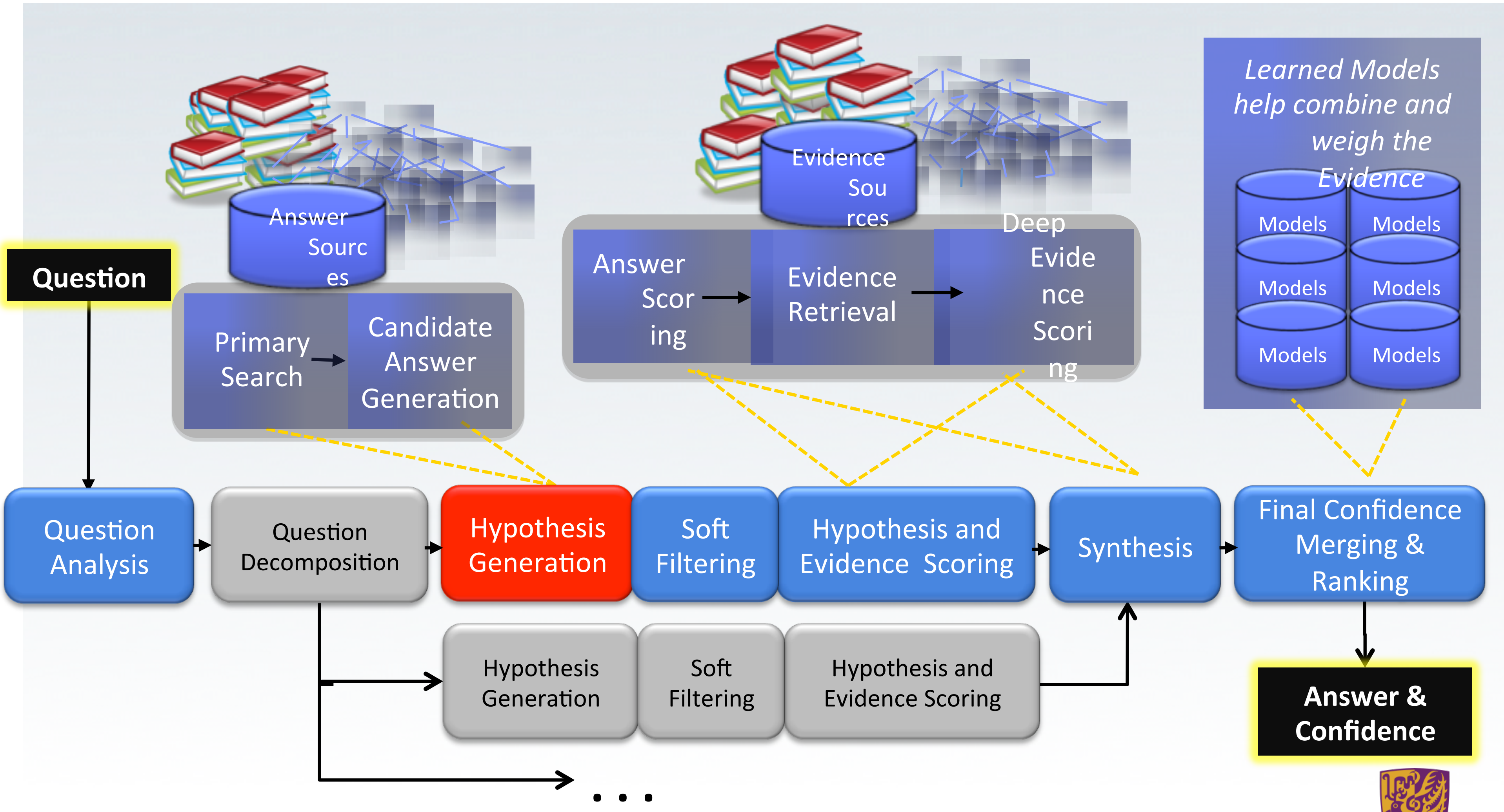
High-Level Architecture



Decomposition

- Help to determine the answer and improve the overall answer confidence
  - Whether questions should be decomposed
  - how best to break them up to sub-questions
    - Rule-based deep parsing
    - statistical classification methods

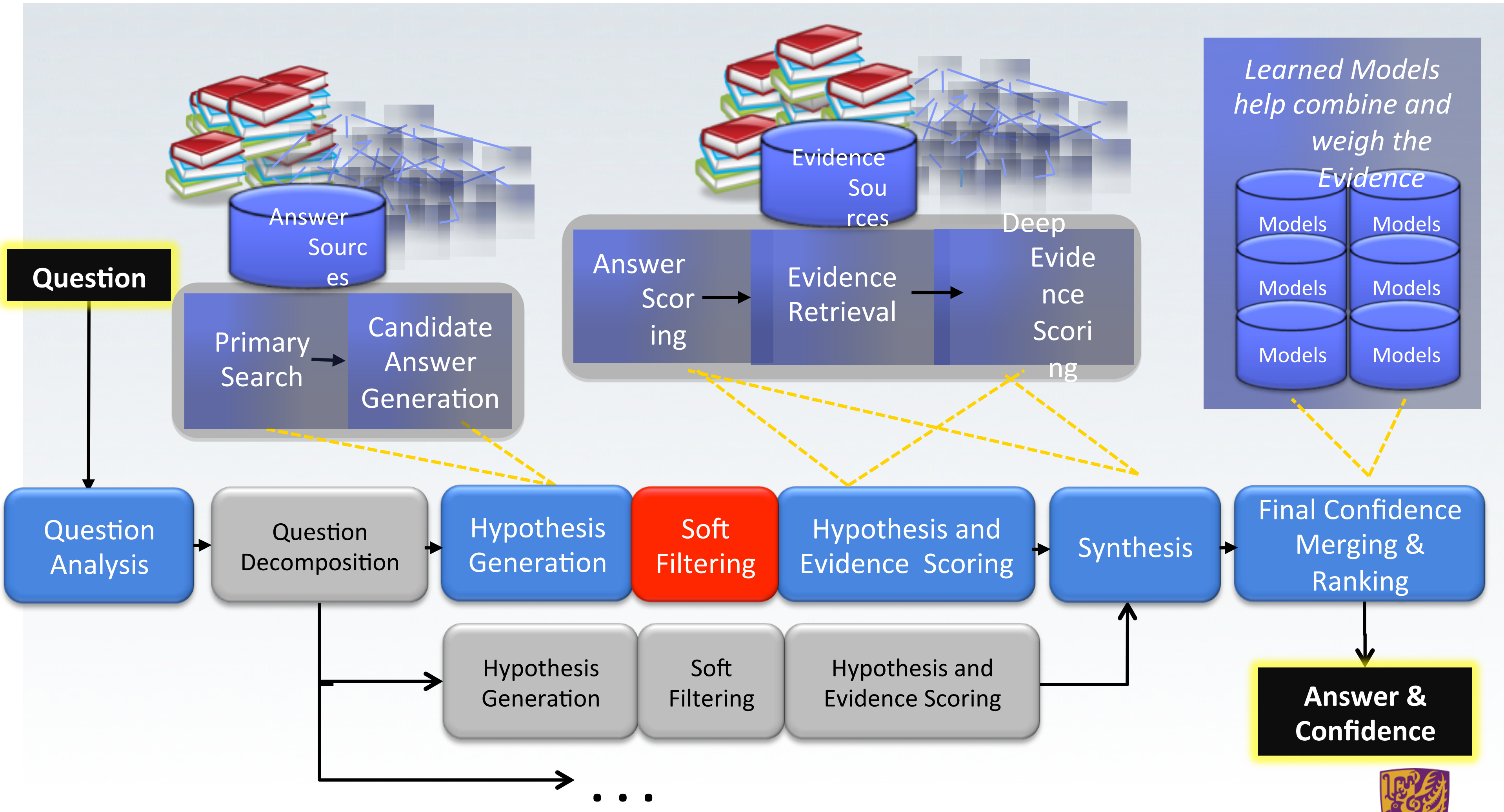
High-Level Architecture



Hypotheses Generation

- Each candidate answer plugged back into the question is considered a hypothesis
  - **Primary Search**
    - Find as much potentially answer-bearing content as possible
    - Operative goal: about 85% percent binary recall for top 250 candidates
    - Techniques used: text search engines (Indri and Lucene), document search, passage search, knowledge base search using SPARQL on triple stores, generation of multiple search queries for a single question, etc.
  - **Candidate Answer Generation**
    - Document search results from "title-oriented" resources: title is extracted as a candidate answer
    - Passage search results: detailed analysis of the passage text
    - Several hundred candidate answers are generated
    - Favor recall over precision

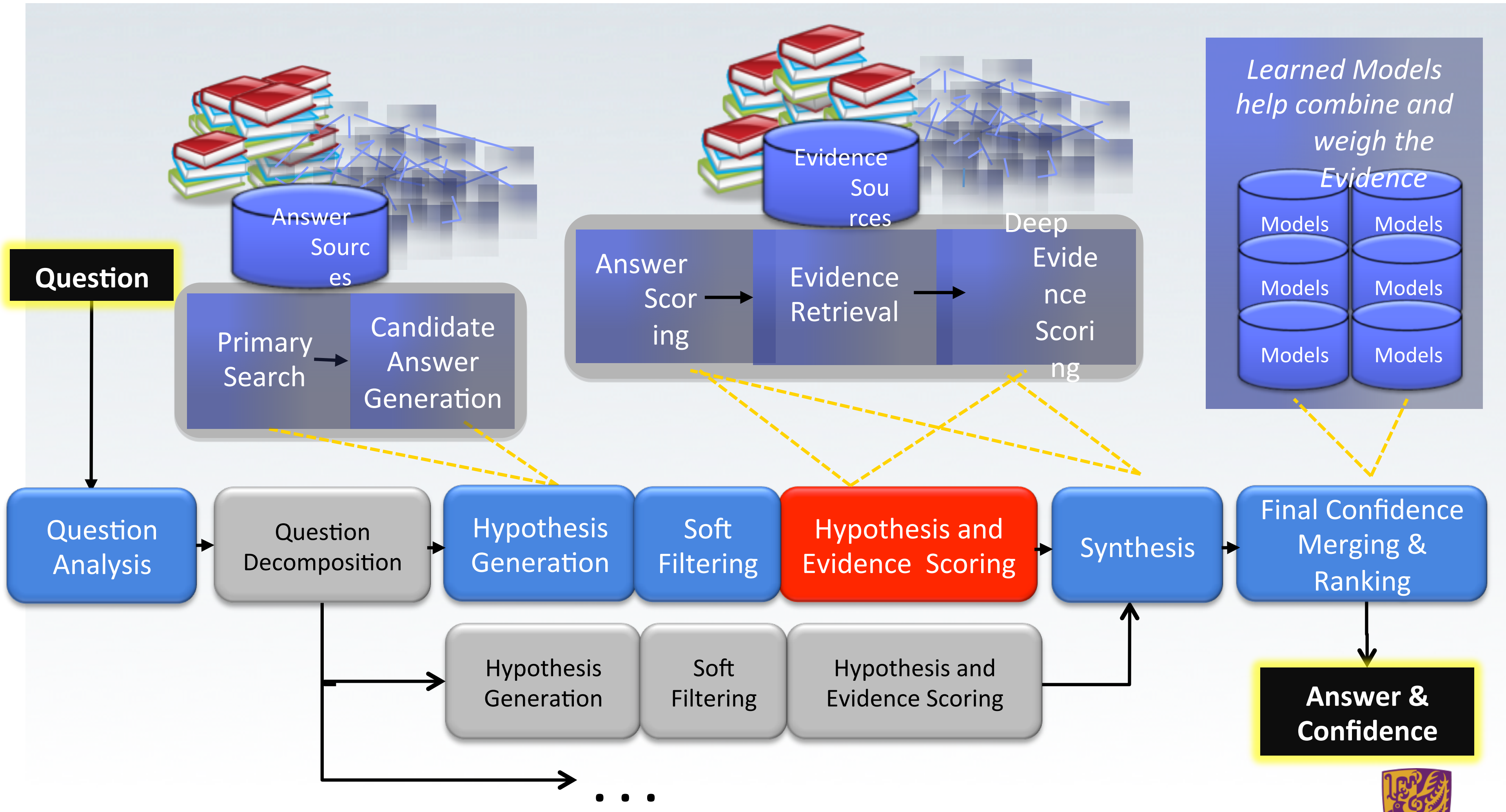
High-Level Architecture



Soft Filtering

- Conduct lightweight scoring algorithms to prune the initial candidates down to smaller set of candidates
  - Less resource intensive scoring models (learning algorithms)
  - For example, the likelihood of a candidate answer being an instance of the LAT

High-Level Architecture

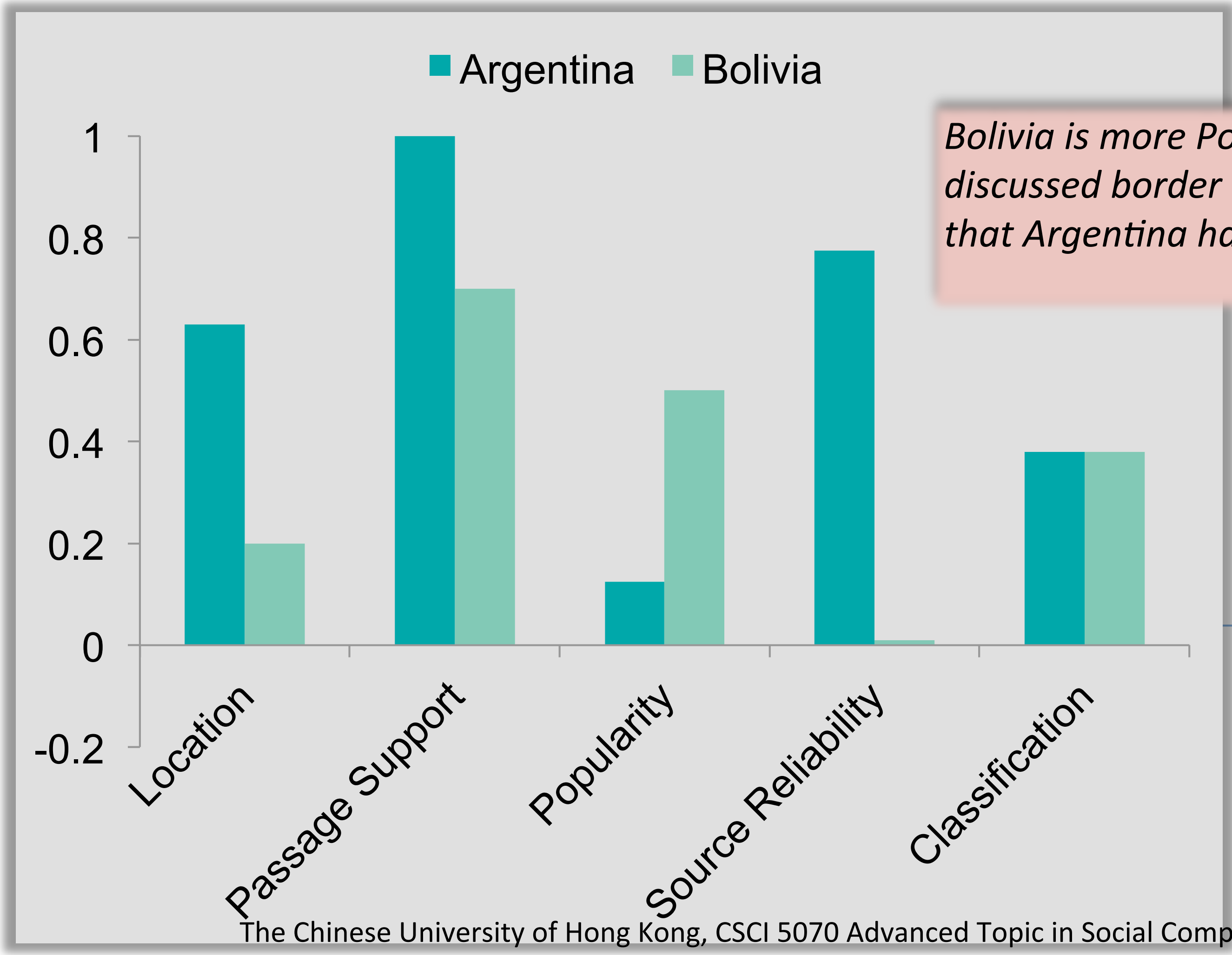


Hypothesis and Evidence Scoring

- Evidence Retrieval
  - Gather additional supporting evidence for each candidate answer
  - E.g., passage search with the candidate answer added to the primary search query
- Scoring
  - Determine the degree of certainty that retrieved evidence supports the candidate answers
  - Various scorers
    - the degree of match between a passage's predicate-argument structure and the question (Smith &Watrman, 1981)
    - deep semantic relationships (Lenat, 1995)
    - passage source reliability
    - geospatial location
    - temporal relationships
    - popularity
    - …

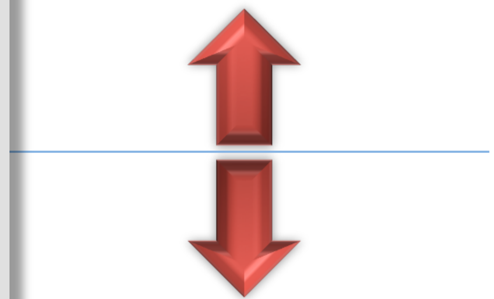
Example

Clue: Chile shares its longest land border with this country.



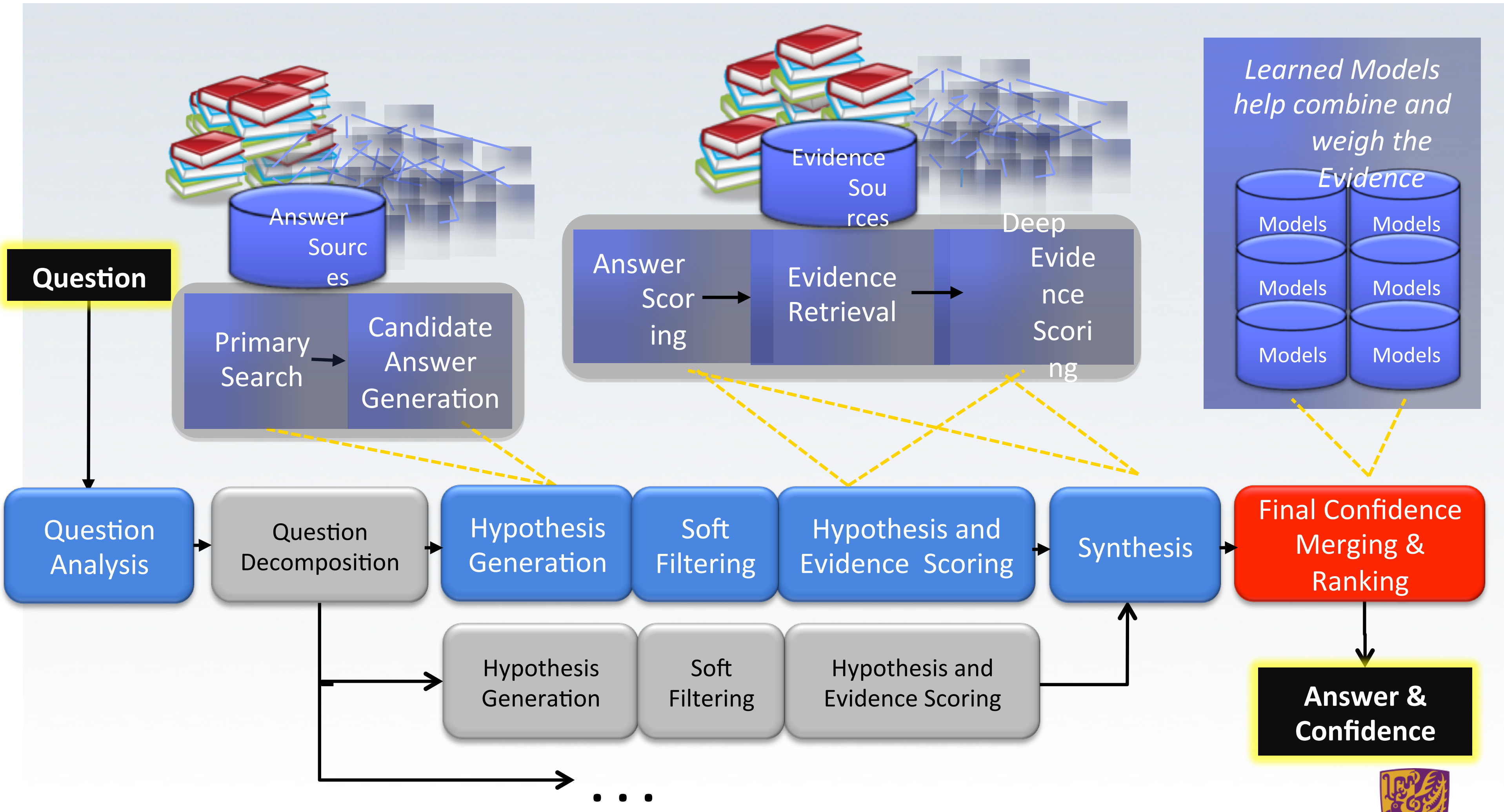
Bolivia is more Popular due to a commonly discussed border dispute. But Watson learns that Argentina has better evidence.

Positive Evidence



Negative Evidence

High-Level Architecture



Answer Merging

- Multiple candidate answers may be equivalent despite very different surface forms
  - Abraham Lincoln and Honest Abe
- Custom merging per feature to combine scores
- Algorithms
  - matching
  - normalization
  - co-reference resolution

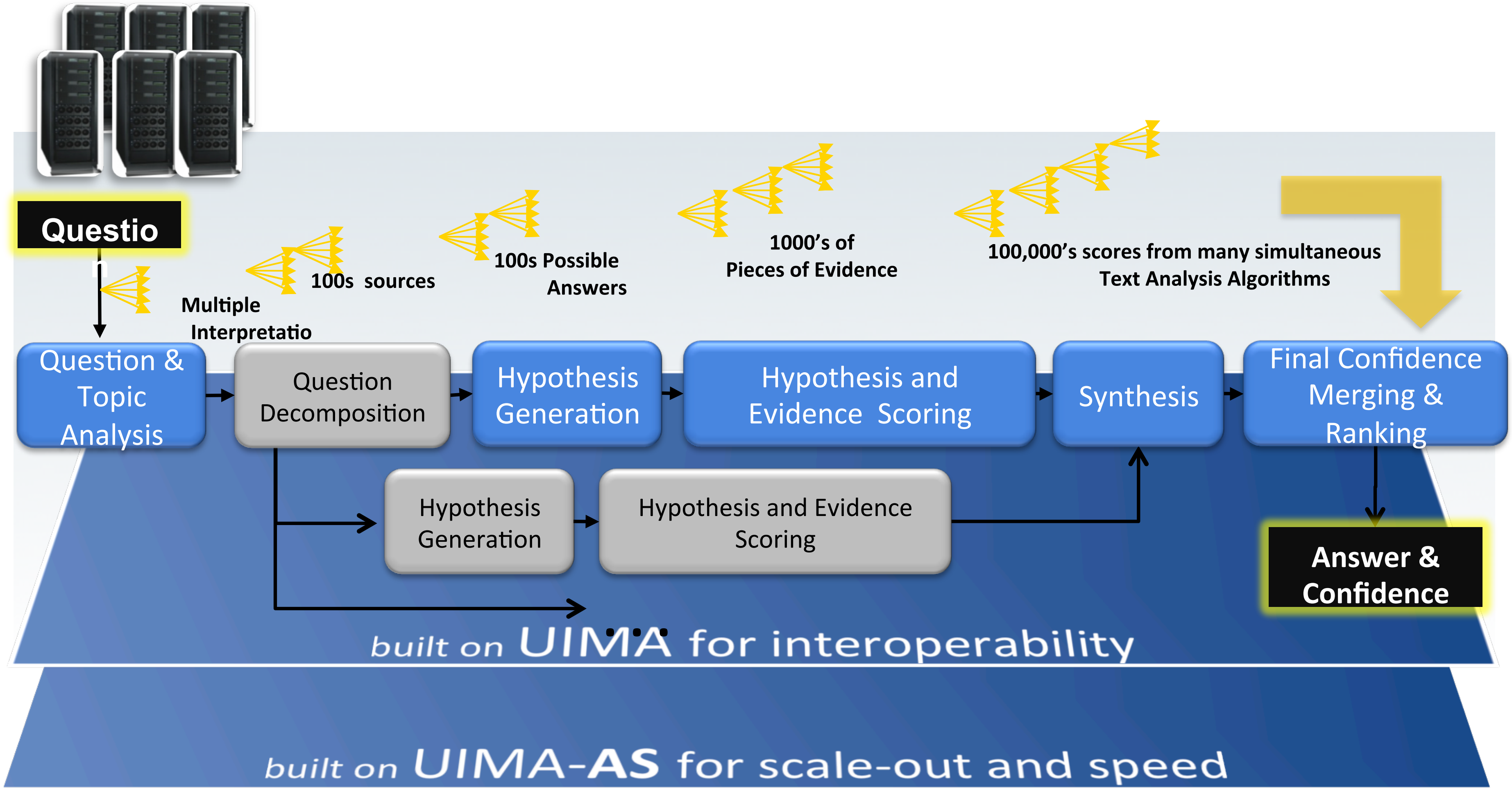
Ranking and Confidence Estimation

- A ranking model is trained over a set of training questions with known answers
  - Multiple trained models are used to handle different question classes
- Confidence estimation
  - Confidence-weighted learning techniques (Dredze et al., 2008)

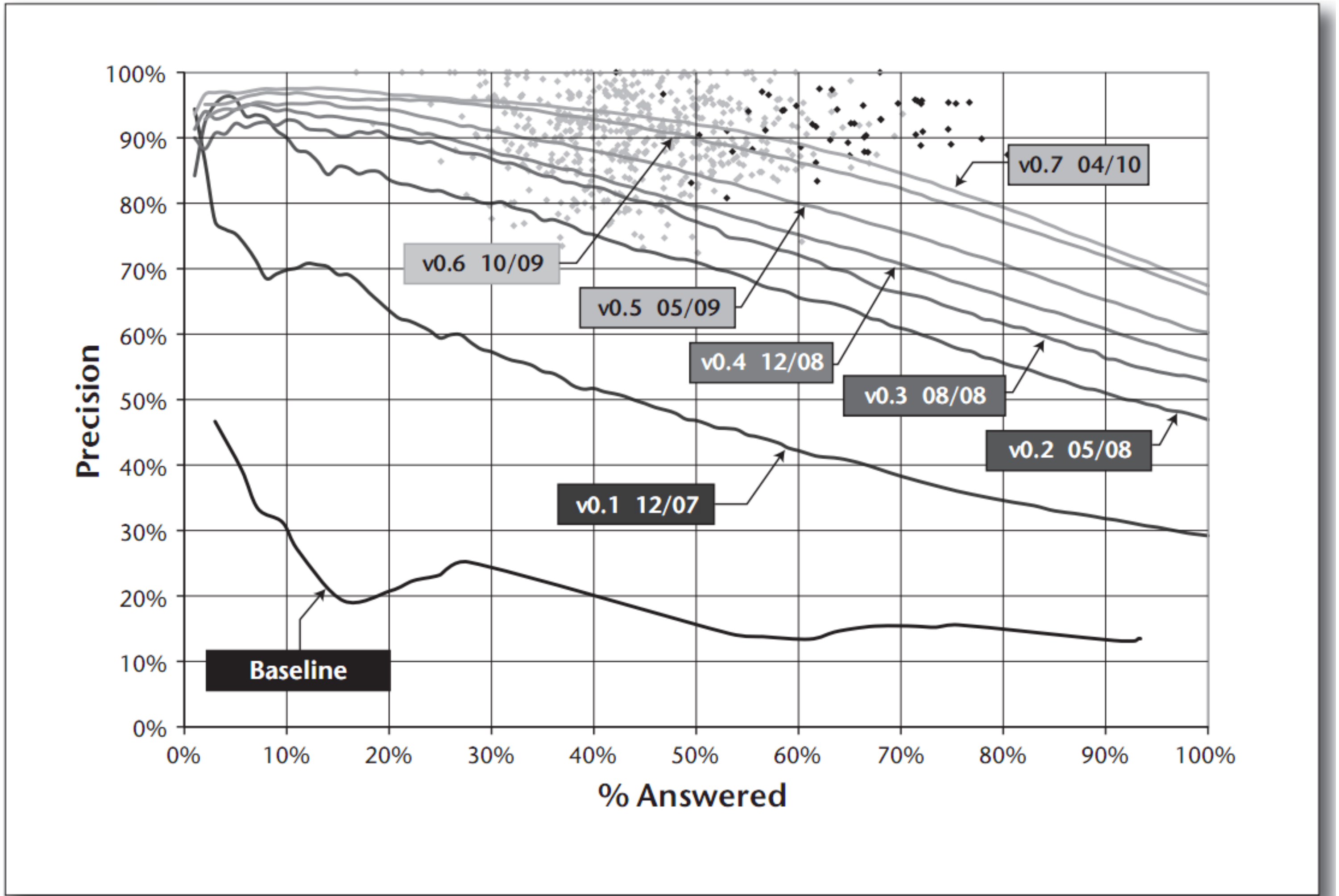
Speed and Scaleout

- Apache UIMA
  - A framework implementation of the Unstructured Information Management Architecture (Ferrucci & Lally , 2004)
  - Support interoperability and scaleout of text and multi-modal analysis applications
  - All of the components in DeepQA are implemented as UIMA annotators

One Jeopardy! question can take **2 hours on a single 2.6Ghz Core**
Optimized & Scaled out on 2880-Core IBM workload optimized POWER7 HPC using UIMA-AS, *Watson* answers in 2-6 seconds.

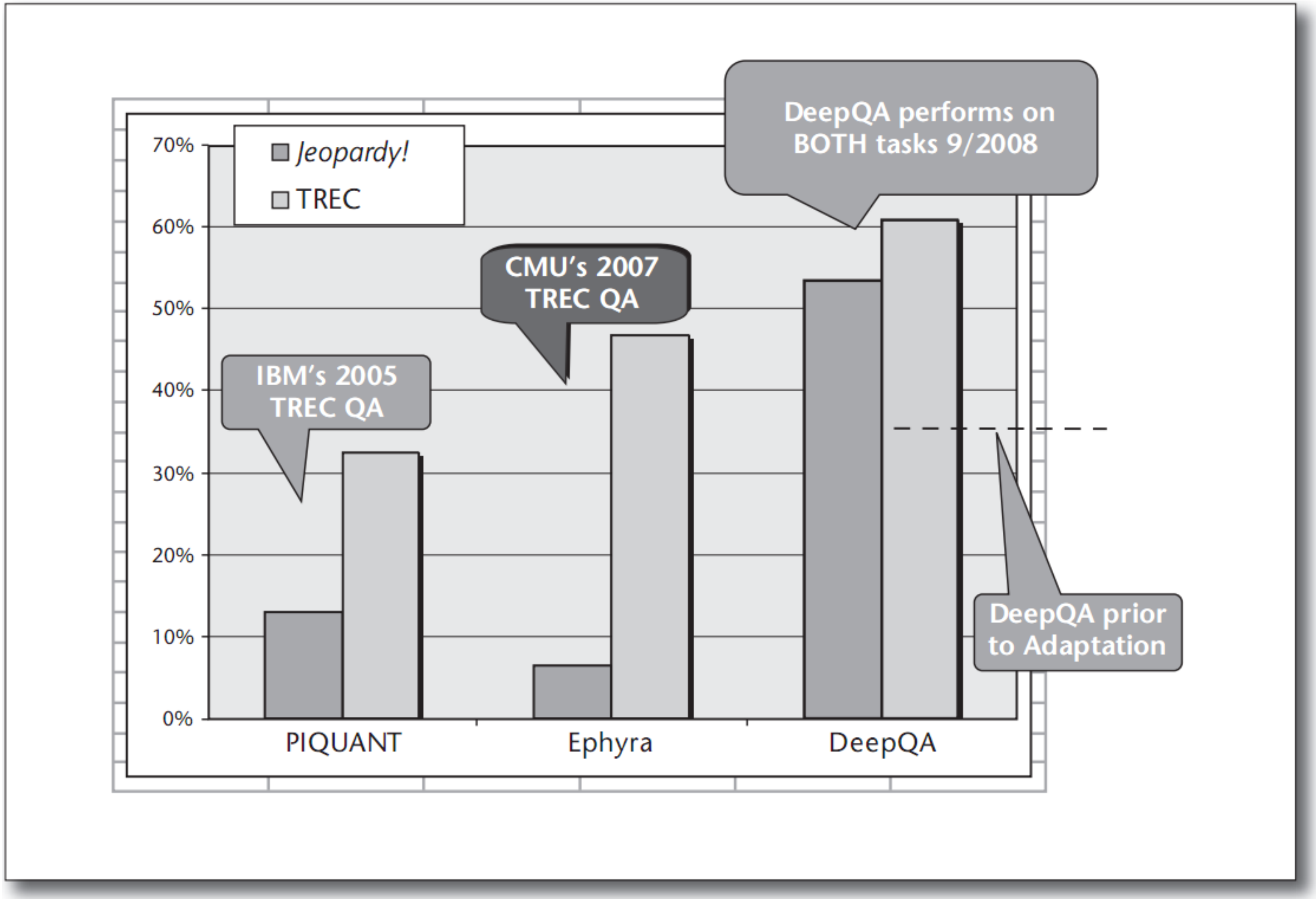


Performance



Watson's Precision and Confidence Progress as of the Fourth Quarter 2009

Performance (cont.)



Accuracy on Jeopardy! *and TREC*

More Resources

- Links
  - DeepQA homepage
  - Watson homepage
  - About Watson on Jeopardy.com
- Videos
  - Building Watson – A Brief Overview of the DeepQA Project
  - IBM "Watson" System to Challenge Humans at Jeopardy!

References

- Eugene Agichtein, Silviu Cucerzan, and Eric Brill. 2005. Analysis of factoid questions for effective relation extraction. In Proc. of SIGIR, 2005.
- Marius Pasca, Dekang Lin, Jeffrey Bigham, Andrei Lifchits, and Alpa Jain. Organizing and Searching the World Wide Web of Facts - Step One: the One-Million Fact Extraction Challenge, In Proc. of AAAI, 2006.
- David Ferrucci, Eric Brown, Jennifer Chu-Carroll, James Fan, David Gondek, Aditya A. Kalyanpur, Adam Lally, J. William Murdock, Eric Nyberg, John Prager, Nico Schlaefer, and Chris Welty. Building Watson: An overview of the DeepQA project. AI Magazine, 31(3):59–79, 2010.
- D Ferrucci, E Nyberg, J Allan, K Barker, E Brown, J Chu-Carroll, A Ciccolo, P Duboue, J Fan, D Gondek, E Hovy, B Katz, A Lally, M McCord, P Morarescu, JW Murdock, B Porter, J Prager, T Strzalkowski, C Welty, W Zadrozny. Towards the Open Advancement of Question Answering Systems. IBM Research Report. RC24789 (W0904-093), 2009.