Natural Language for Communication

Watson Overview

Our Study Path Forward for "Natural Language for Communication"

- Groundwork:
  - Review of probability: Ch. 13
  - Probabilistic reasoning over time: Ch. 15.1-15.3
  - Language models: Ch. 22.1
- Natural language for communication: Ch. 23

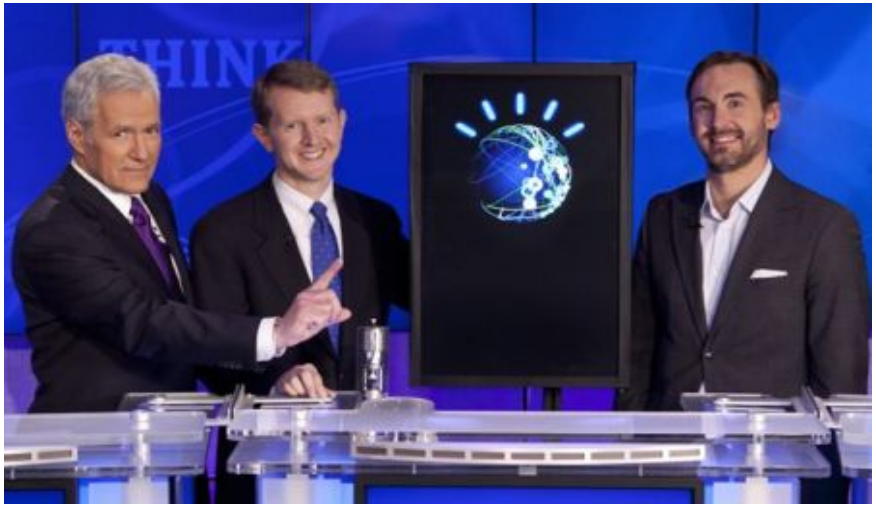
IBM Watson

Slides from Watson Team, Presenter: Joel Farrell, IBM

Reference publication: David Ferrucci, et al, “Building Watson: An Overview of the DeepQA Project.” AI Magazine, 31, 3 (Fall 2010), 59-79.

© 2009 IBM Corporation

IBM’s Watson…



Why Jeopardy?

The game of *Jeopardy!* makes great demands on its players – from the range of topical knowledge covered to the nuances in language employed in the clues. The question IBM had for itself was “is it possible to build a computer system that could process big data and come up with sensible answers in seconds—so well that it could compete with human opponents?”

IBM Watson's project started 2007

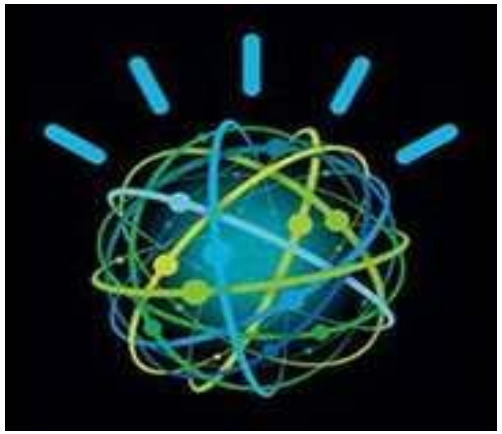
- Project started in 2007, lead David Ferrucci
- Initial goal: create a system able to process natural language & extract knowledge faster than any other computer or human

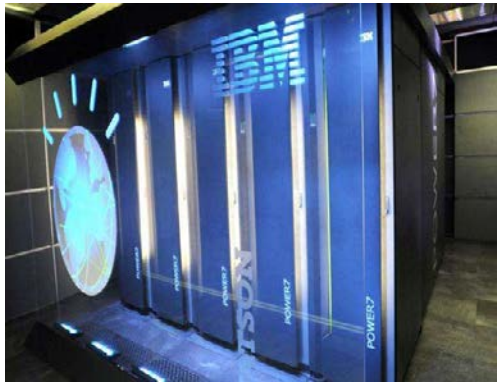
- Jeopardy! was chosen because it's a huge challenge for a computer to find the questions to such "human" answers under time pressure

- Watson was NOT online!

- Watson weighs the probability of his answer being right – doesn't ring the buzzer if he's not confident enough

- Which questions Watson got wrong almost as interesting as which he got right!





Watson – a Workload Optimized System

- 90 x IBM Power 750[1] servers
- 2880 POWER7 cores
- POWER7 3.55 GHz chip
- 500 GB per sec on-chip bandwidth
- 10 Gb Ethernet network
- 15 Terabytes of memory
- 20 Terabytes of disk, clustered
- Can operate at 80 Teraflops
- Runs IBM DeepQA software
- Scales out with and searches vast amounts of unstructured information with UIMA & Hadoop open source components
- Linux provides a scalable, open platform, optimized to exploit POWER7 performance
- 10 racks include servers, networking, shared disk system, cluster controllers

Linux

POWER7™ BUILT ON Power™

[1] Note that the Power 750 featuring POWER7 is a commercially available server that runs AIX, IBM i and Linux and has been in market since Feb 2010

© 2009 IBM Corporation

This means Watson…

- Operates at 80 teraflops. The human brain is estimated to have a processing power of 100 teraflops (100 trillion operations per second).
- Has the equivalent *in memory* (RAM) that the Library of Congress adds in books and media over a 4 month period
- Can process 200 million times more instructions *per second* than the Space Shuttle's computers.
- Parses within *3 seconds* the equivalent of the number of books on a 700 yard long book shelf…and pick out the relevant information, and create an answer.

A Grand Challenge Opportunity

- Capture the imagination
  - The Next *Deep Blue*
- Engage the scientific community
  - Envision new ways for computers to impact society & science
  - Drive important and measurable scientific advances
- Be Relevant to Important Problems
  - Enable better, faster decision making over unstructured and structured content
  - Business Intelligence, Knowledge Discovery and Management, Government, Compliance, Publishing, Legal, Healthcare, Business Integrity, Customer Relationship Management, Web Self-Service, Product Support, etc.

© 2009 IBM Corporation

Real Language is Real Hard

- Chess
  - A finite, mathematically well-defined search space
  - Limited number of moves and states
  - Grounded in **explicit**, **unambiguous** mathematical rules

- Human Language
  - Ambiguous, contextual and implicit
  - Grounded only in **human cognition**
  - Seemingly infinite number of ways to express the same meaning

© 2009 IBM Corporation

What Computers Find Easier (and Hard)

ln((12,546,798 * π) ^ 2) / 34,567.46 = **0.00885**

Select *Payment* where *Owner*="David Jones" and *Type(Product)*="Laptop",

Owner	Serial Number
David Jones	45322190-AK

Serial Number	Type	Invoice #
45322190-AK	LapTop	INV10895

Invoice #	Vendor	Payment
INV10895	MyBuy	$104.56

David Jones
↓↓↓↓↓↓↓↓↓↓
David Jones

=

Dave Jones
↓↓↓↓ ↓↓↓↓↓
David Jones

≠

© 2009 IBM Corporation

What Computers Find Hard

Computer programs are natively **explicit, fast** and **exacting** in their calculation over numbers and symbols….But **Natural Language** is implicit, highly contextual, ambiguous and often imprecise.

Person	Birth Place
A. Einstein	ULM

Structured

Unstructured

- Where was X born?

  One day, from among his city views of Ulm, Otto chose a water color to send to Albert Einstein as a remembrance of Einstein´s birthplace.

Person	Organization
J. Welch	GE

- X ran this?

  If leadership is an art then surely Jack Welch has proved himself a master painter during his tenure at GE.

© 2009 IBM Corporation

Some Basic Jeopardy! Clues

- Category: ENDS IN "TH"
- This **fish** was thought to be extinct millions of years ago until one was found off South Africa in 1938
- Answer: **coelacanth**

The *type* of thing being asked for is often indicated but can go from specific to very vague

- Category: General Science
- When hit by electrons, a phosphor gives off electromagnetic energy in this **form**
- Answer: **light (or photons)**

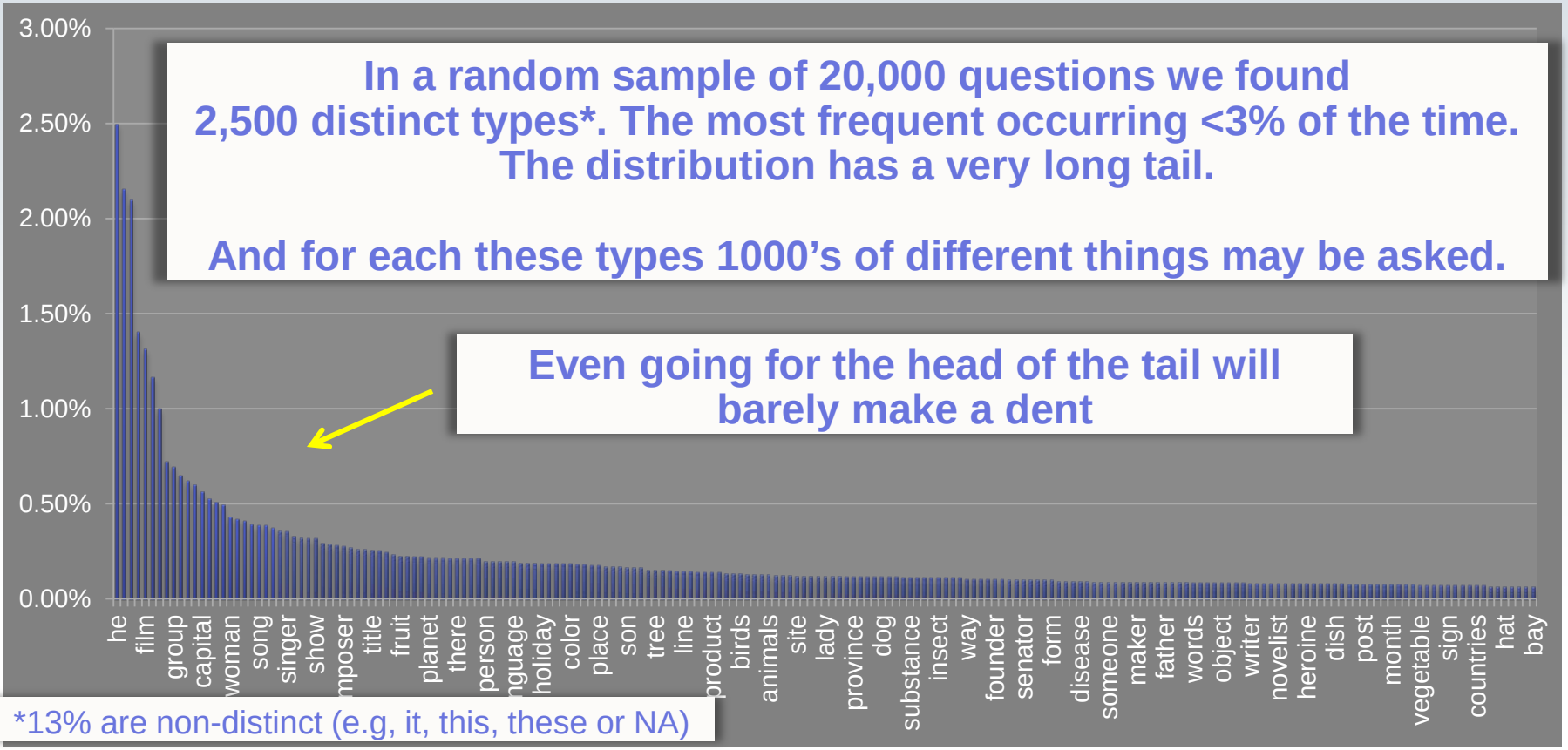
- Category: Lincoln Blogs
- Secy. Chase just submitted **this** to me for the third time--guess what, pal. This time I'm accepting **it**
- Answer: **his resignation**

© 2009 IBM Corporation

Broad Domain

We do NOT attempt to anticipate all questions and build databases.

We do NOT try to build a formal model of the world



In a random sample of 20,000 questions we found 2,500 distinct types*. The most frequent occurring <3% of the time. The distribution has a very long tail.

And for each these types 1000's of different things may be asked.

Even going for the head of the tail will barely make a dent

*13% are non-distinct (e.g, it, this, these or NA)

Our Focus is on reusable NLP technology for analyzing vast volumes of *as-is* text.
Structured sources (DBs and KBs) provide background knowledge for interpreting the text.

© 2009 IBM Corporation

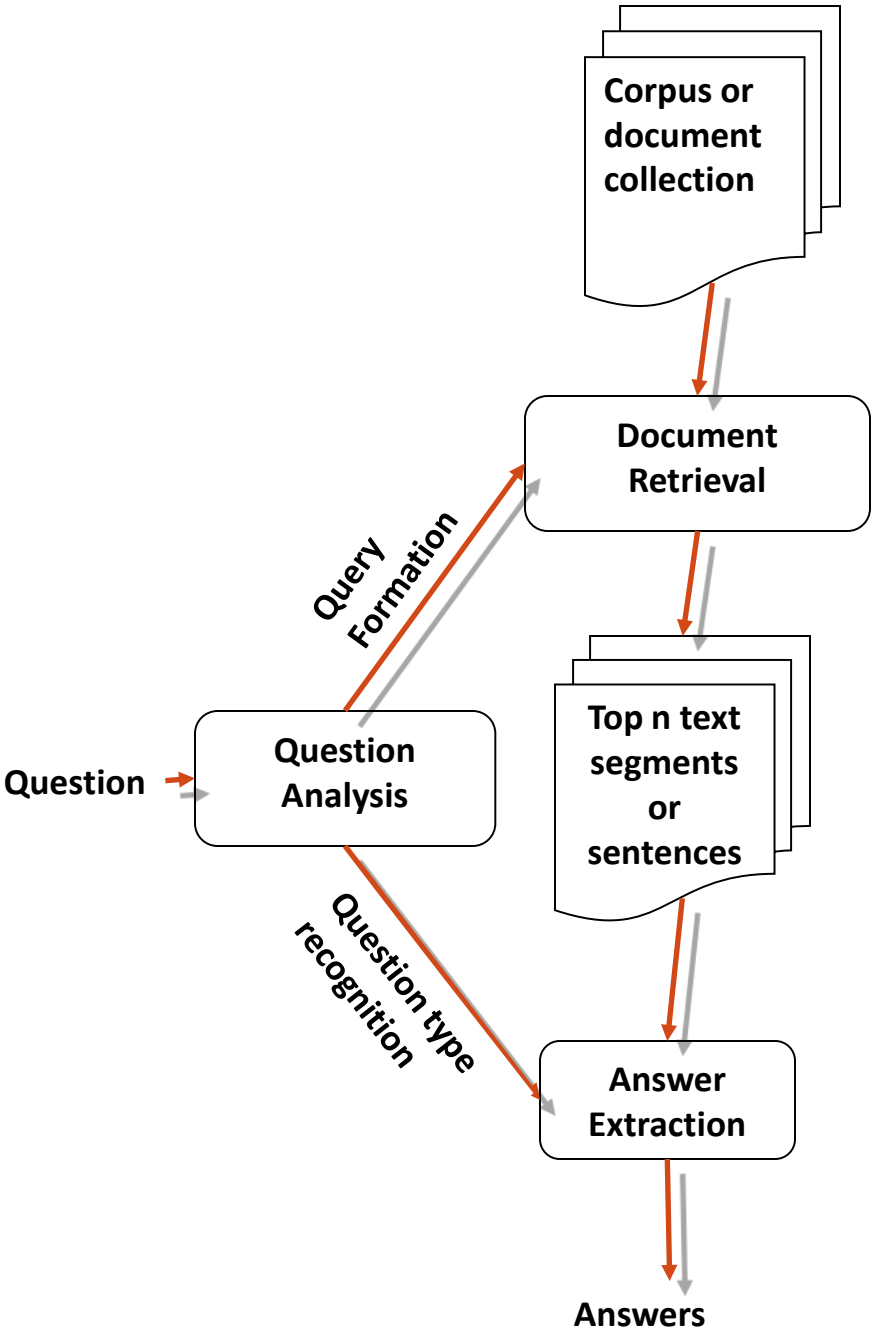
Generic Framework

The majority of current question answering systems designed to answer factoid questions consist of three distinct components:

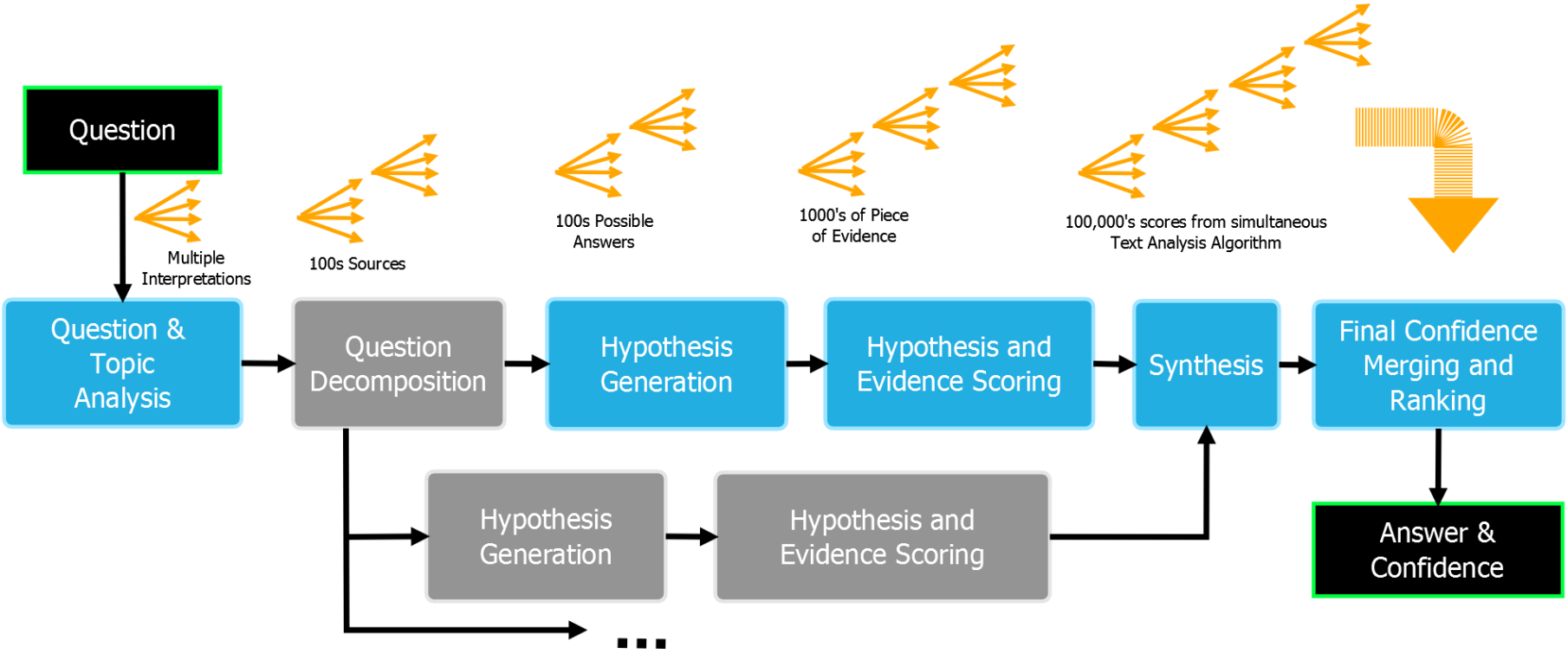
1)question analysis,

2)document or passage retrieval and finally

3)answer extraction.

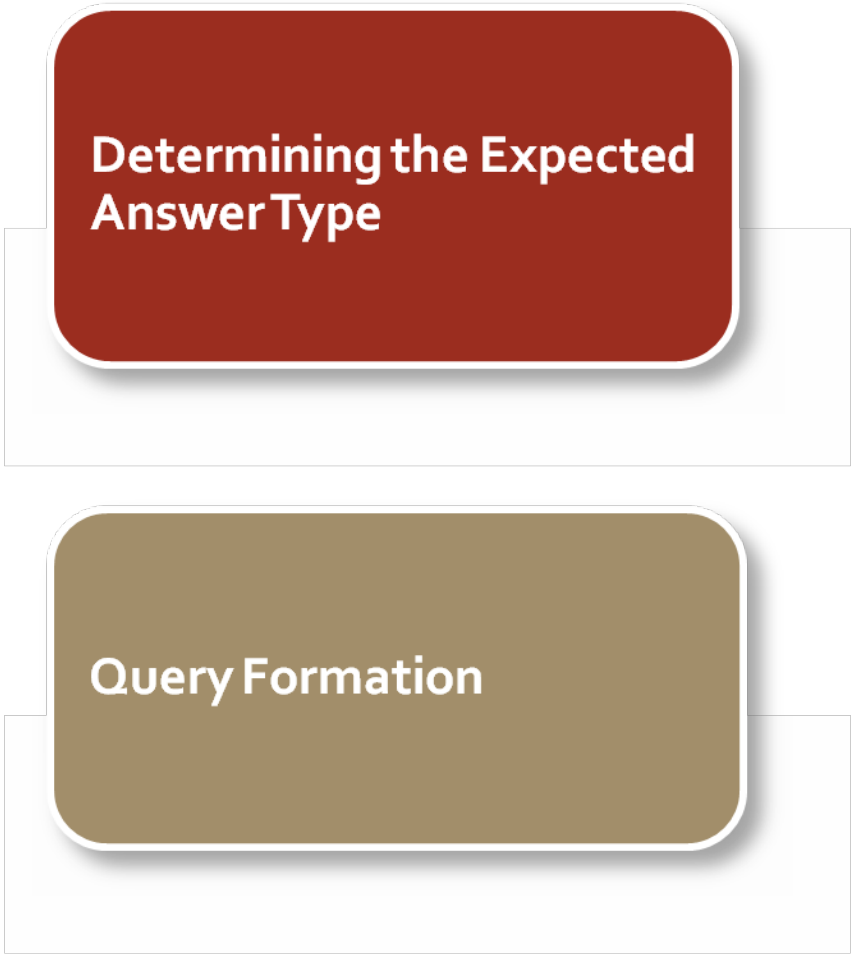


Basic Architecture



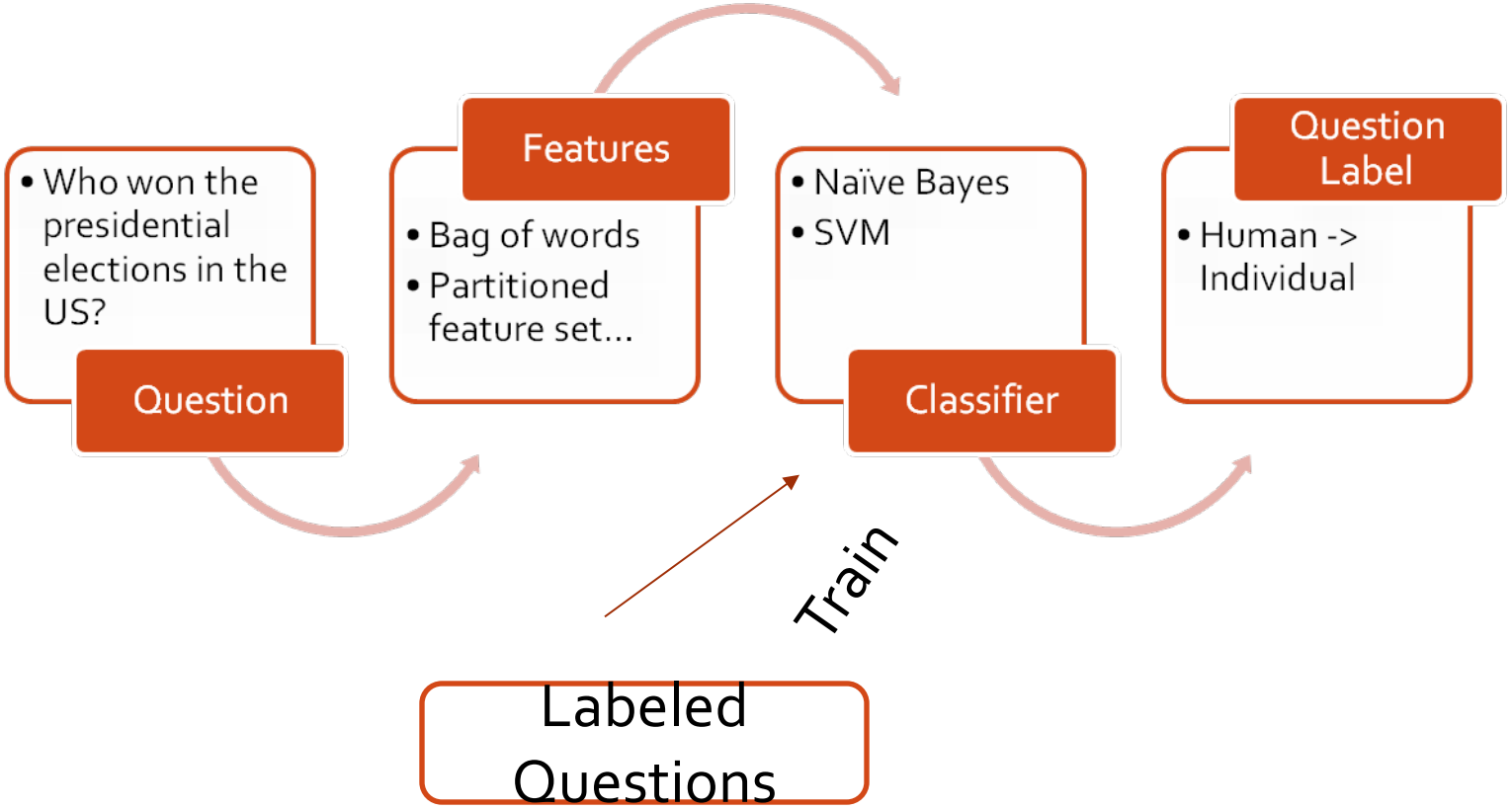
Question Analysis

As the first component in a QA system it could easily be argued that question analysis is the most important part. Any mistakes made at this stage are likely to render useless any further processing of a question.



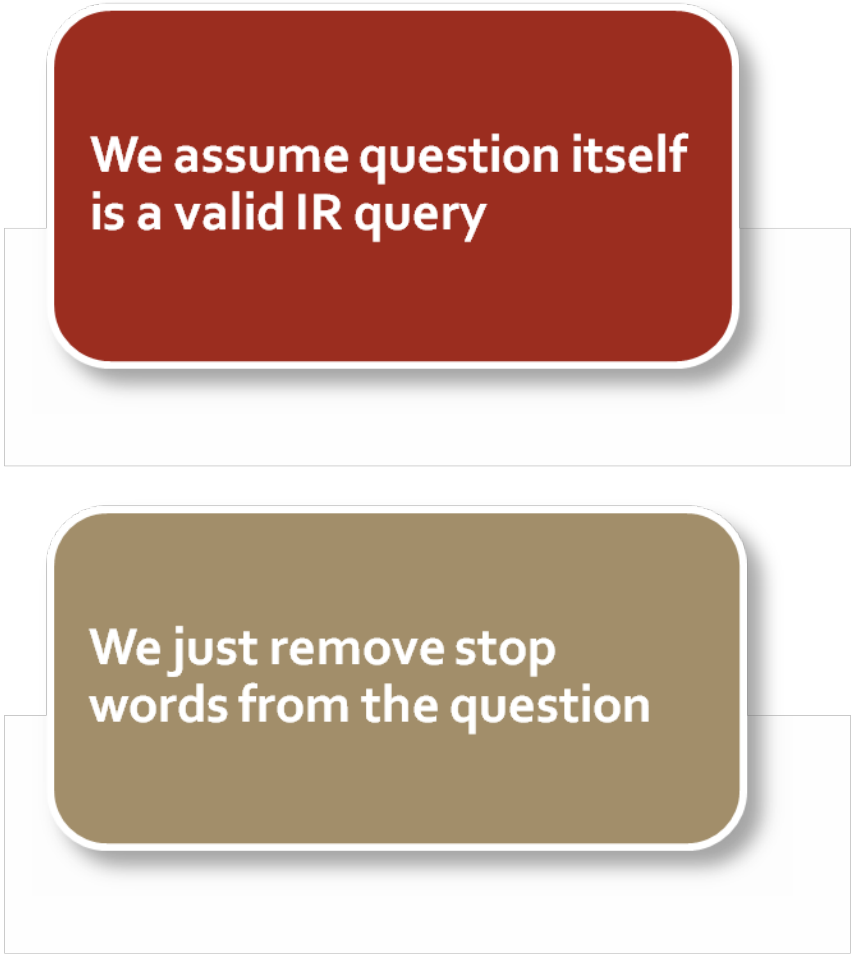
Determining the Expected Answer Type

Machine learning techniques to classify a question. We can train our system on thousands of tagged question corpus, Provided by cognitive computation group at the department of computer science, university of illinois at urbana-champaign to determine the expected answer.



Query Formation

The question analysis component of a QA system is usually responsible for formulating a query from a natural language questions to maximise the performance of the IR engine used by the document retrieval component of the QA system.



Database Access Schemata

- Who is the president of India?

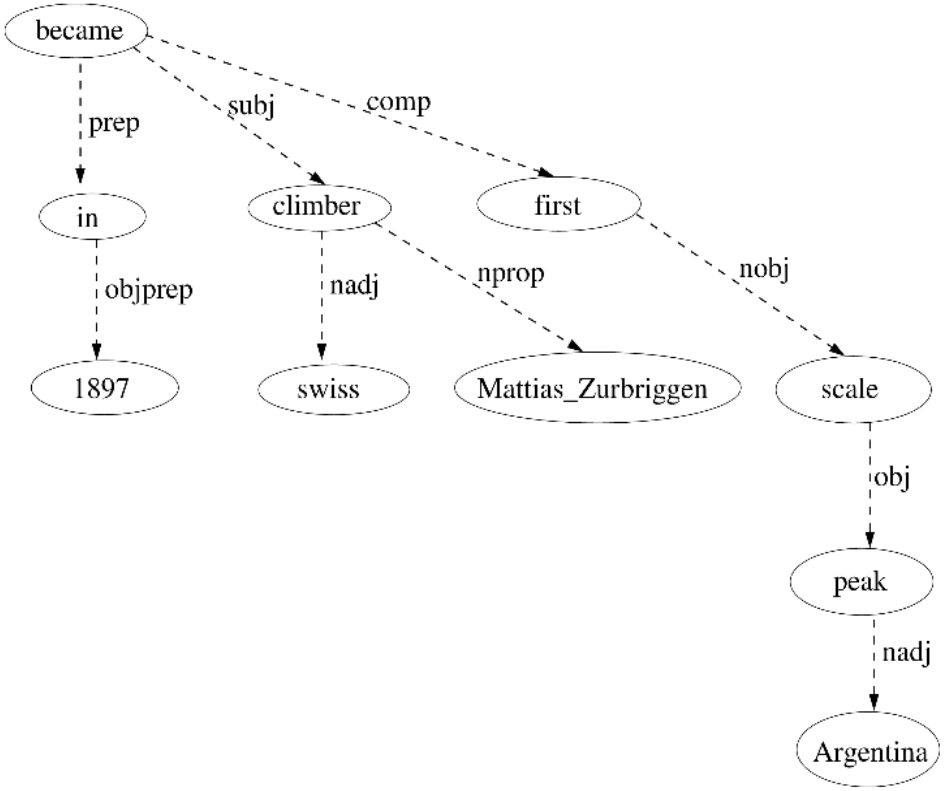
Access Schemata –

Search <> for name <> biography.com <> person – president <> place - India

How Watson works: Step 1 Analyzing the question

Category:
WORLD GEOGRAPHY

Clue:
In 1897 Swiss climber Matthias Zurbriggen became the first to scale this Argentinean peak.



Step 1 Watson dissects the clue to understand what it is asking for.

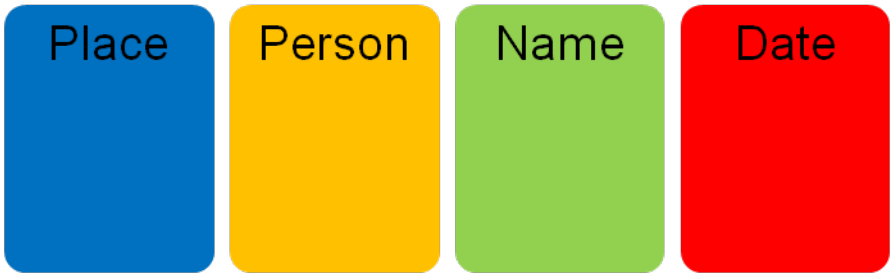
Watson tokenizes and parses the clue to identify the relationships between important words and find the focus of the clue, i.e. this Argentinean peak.

Document Retrieval

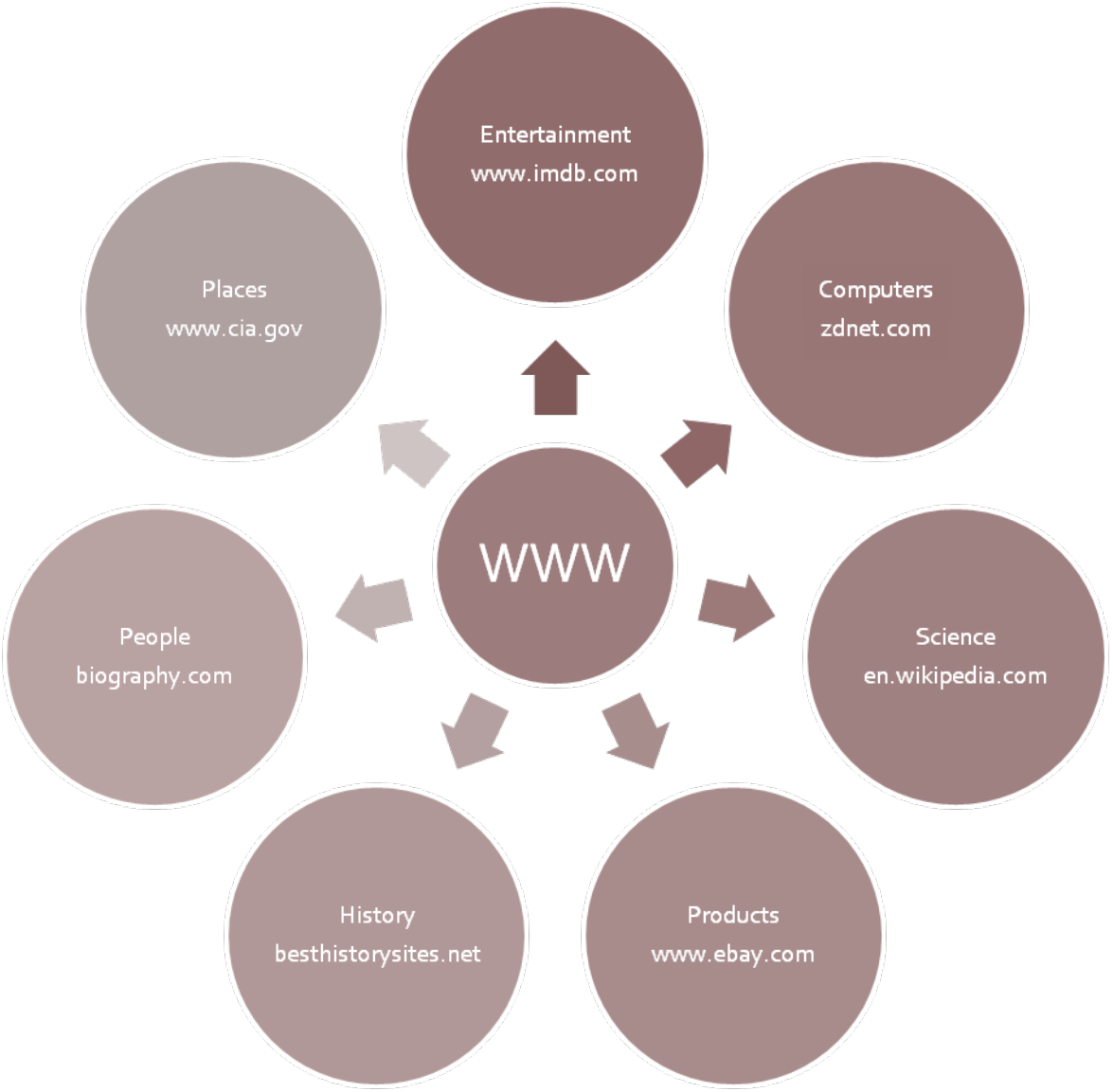
The text collection over which a QA system works tend to be so large that it is impossible to process whole of it to retrieve the answer. The task of the document retrieval module is to select a small set from the collection which can be practically handled in the later stages.

Knowledge Annotation

The Taj Mahal completed around 1648 is a mausoleum located in Agra, India, that was built under Mughal Emperor Shah Jahan in memory of his favourite wife, Mumtaz Mahal.



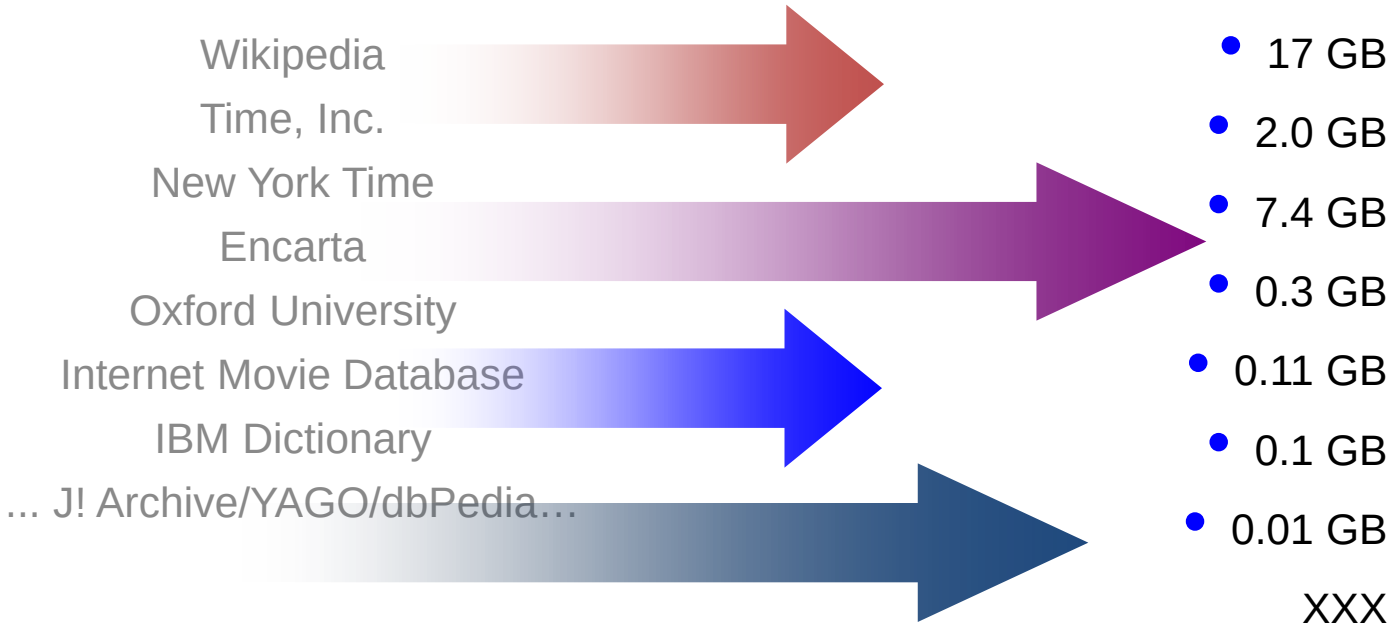
Pockets of structured and semi-structured knowledge



Where did it acquire knowledge?

Three types of knowledge

Domain Data (articles, books, documents)	**Training and test question sets w/answer keys**	**NLP Resources** (vocabularies, taxonomies, ontologies)



Total Raw Content	**70 GB**
Preprocessed Content	**500 GB**

How Watson works: Step 2 Search

Timeline of Climbing the Matterhorn

* August 25: H.R.H. the Duke of the Abruzzi made the ascent with Mr. A. F. Mummery and Dr. Norman Collie, and one porter, Pollinger, junior. According to Mummery the weather was threatening, and, the Prince climbing very well, they went exceedingly fast, so that their time was probably the quickest possible. They left the bivouac at the foot of the snow ridge at 3.40 a.m., and reached the summit at 9.50. A few days afterwards the first descent of the ridge was accomplished by Miss Bristow, with the guide Matthias Zurbriggen, of Macugnaga.

The first known ascent of Aconcagua was during an expedition was during an expedition led by Edward Fitz Gerald in the summer of 1897. Swiss climber Matthias Zurbriggen reached the summit alone on January 14 via today's Normal Route. A few days later Nicholas Lanti and Stuart Vines made the second ascent. These were the highest ascents in the world at that time. It's possible that the mountain had previously been climbed by Pre-Columbian Incans.

Step 2 Watson searches its content for text passages that relate to the clue.

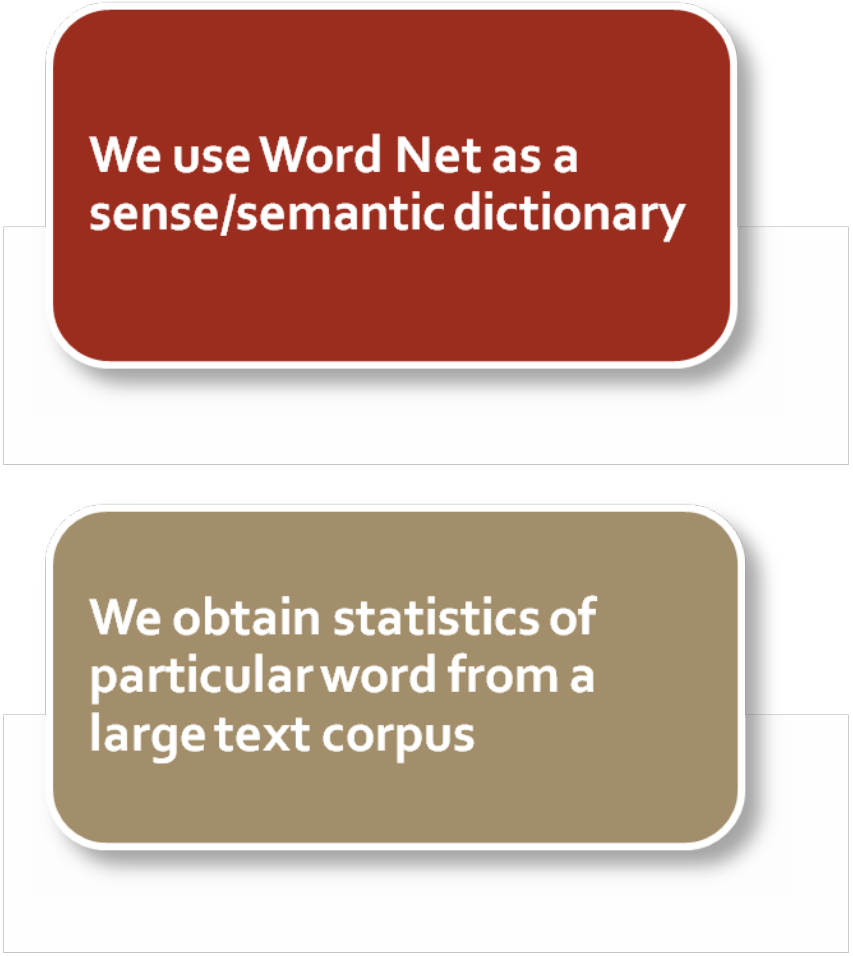
Using important terms from the clue, Watson performs a search over millions of documents to find relevant passages.

Answer Extraction

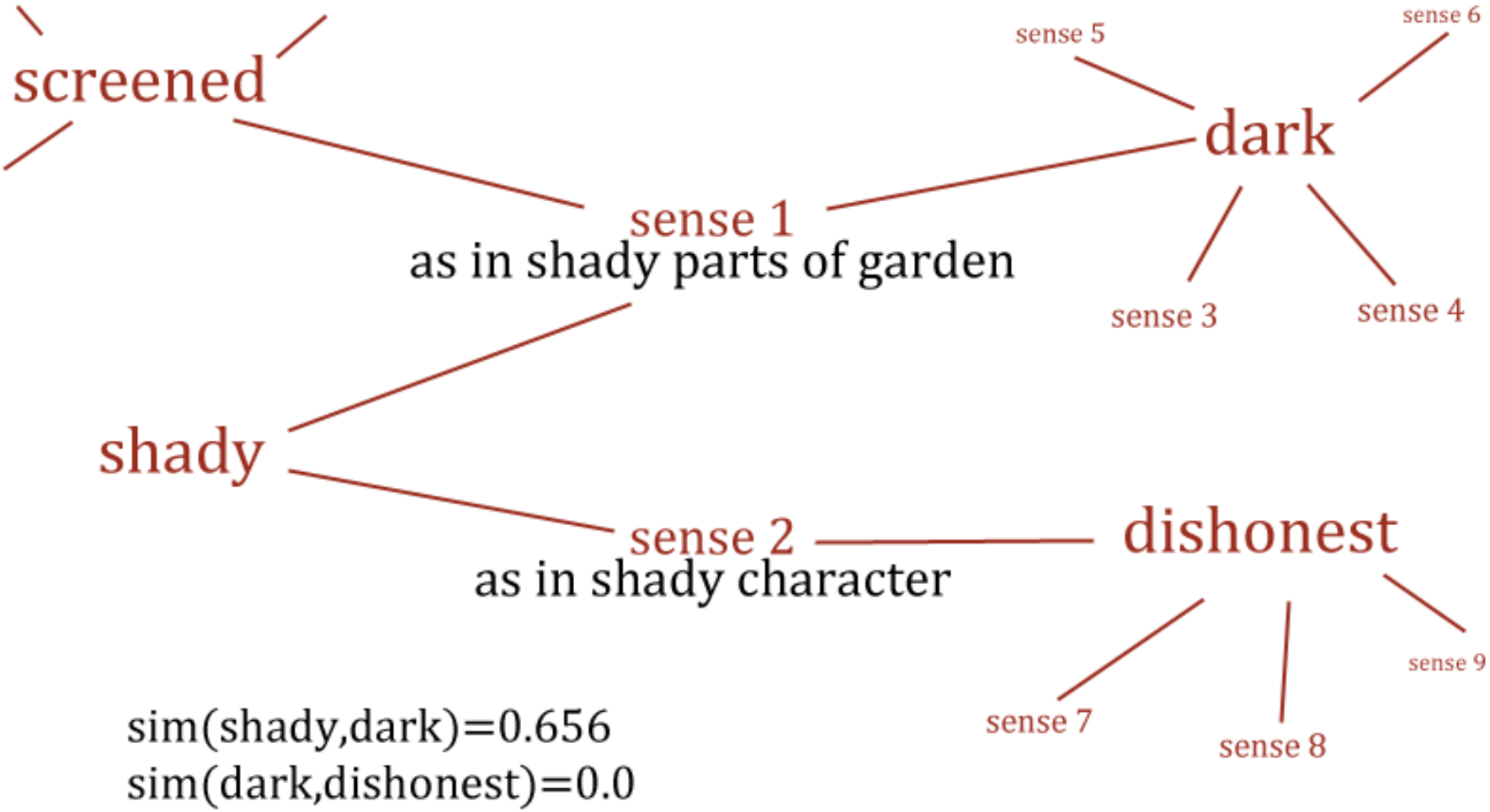
Is responsible for ranking the sentences and giving a relative probability estimate to each one. It also registers the frequency of each individual phrase chunk marked by the NE recognizer for a given question class at a given rank.

Sense/Semantic similarity

- We use statistics to compute information content value.
- We assign a probability to a concept in taxonomy based on the occurrence of target concept in a given corpus.



Word Net - Synsets



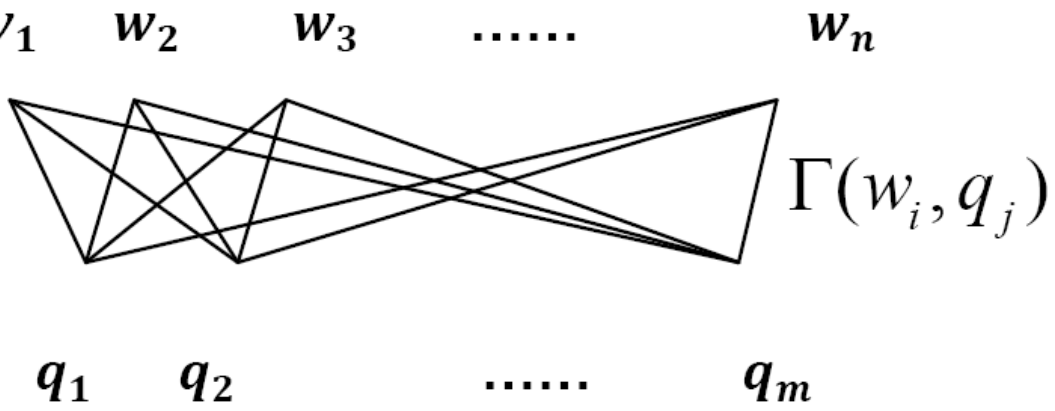
sim(shady,dark)=0.656
sim(dark,dishonest)=0.0

Sense Net Ranking Algorithm

The sentence as well as the query forms an ordered set of words. We then compute the sense network between every pair of words from query and sentence…

$$\Gamma(w_i, q_j) = \xi_{i,j}$$

$\xi_{i,j} \epsilon\ [0,1]$ is the value of sense/semantic similarity between $w_i \epsilon W$ and $q_j \epsilon Q$.



A sense network formed between a sentence and a query.

Exact Match Score

Given a sense network $\Gamma(w_i, q_j)$, we define the distance of a word w_i as

$$d(w_i) = i$$

$$d(q_j) = j$$

Word with maximum sense similarity with query word q_i is:

$$M(q_i) = w_j \mid j = \text{argmax}_j \xi_{j,i}$$

And the corresponding value of

$$\xi_{i,j} = V(q_i)$$

The exact match score is

$$E_{total} = \frac{\sum_i V(q_i)}{m}$$

Alignment Score

Let T = {ordered set of $M(q_i)\ \forall\ i\ \in [1, m]$} in increasing order of d(q). Function $\boldsymbol{\theta_i}$ is the distance of i^{th} element in T then the alignment score is

$$K_{total} = \frac{\sum_{i=1}^{m-1} sgn(\theta_{i+1} - \theta_i)}{m-1}$$

An alignment score of 1.0 signifies perfect alignment while a score of -1.0 signifies reverse order of occurrence.

Total Score

We define the following coefficients

$\mu = noise\ penalty\ coefficient$ $\psi = exact\ match\ coefficient$

$\lambda = sense\ similarity\ coefficient$ $\nu = order\ coefficient$

So the total score is a linear combination of individual scores

$$\boldsymbol{\eta = \psi \times E_{total} + \lambda \times S_{total} + \mu \times \delta_{total} + \nu \times k_{total}}$$

We fine tune the values of these coefficients to get maximum accuracy.

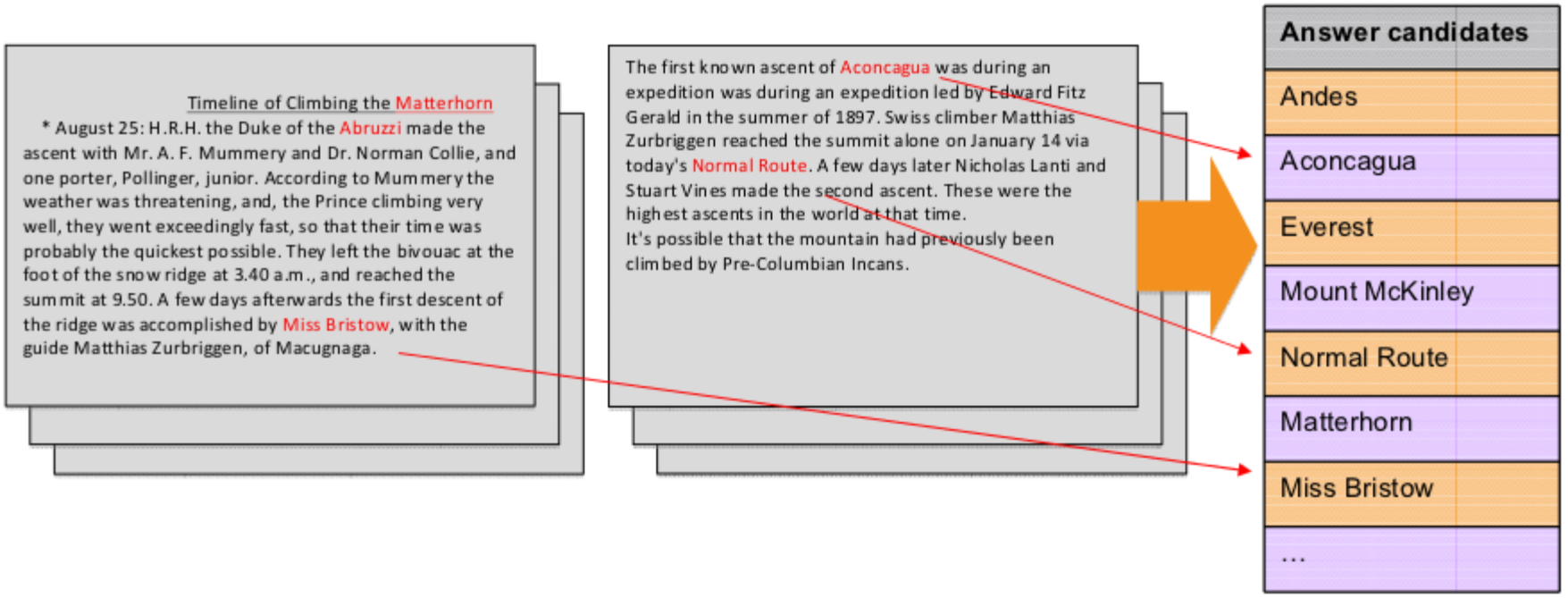
Answer Confidence Score

We take top *t* sentences and consider the plausible answers within them. If an answer appears with frequency *f* in sentence ranked *r then that* answer gets a confidence score -

$$C(ans) = \frac{1}{r}(1 + \ln(f))$$

all answers are sorted according to confidence score and top ϑ (=5 in our case) answers are returned along with corresponding sentence and URL

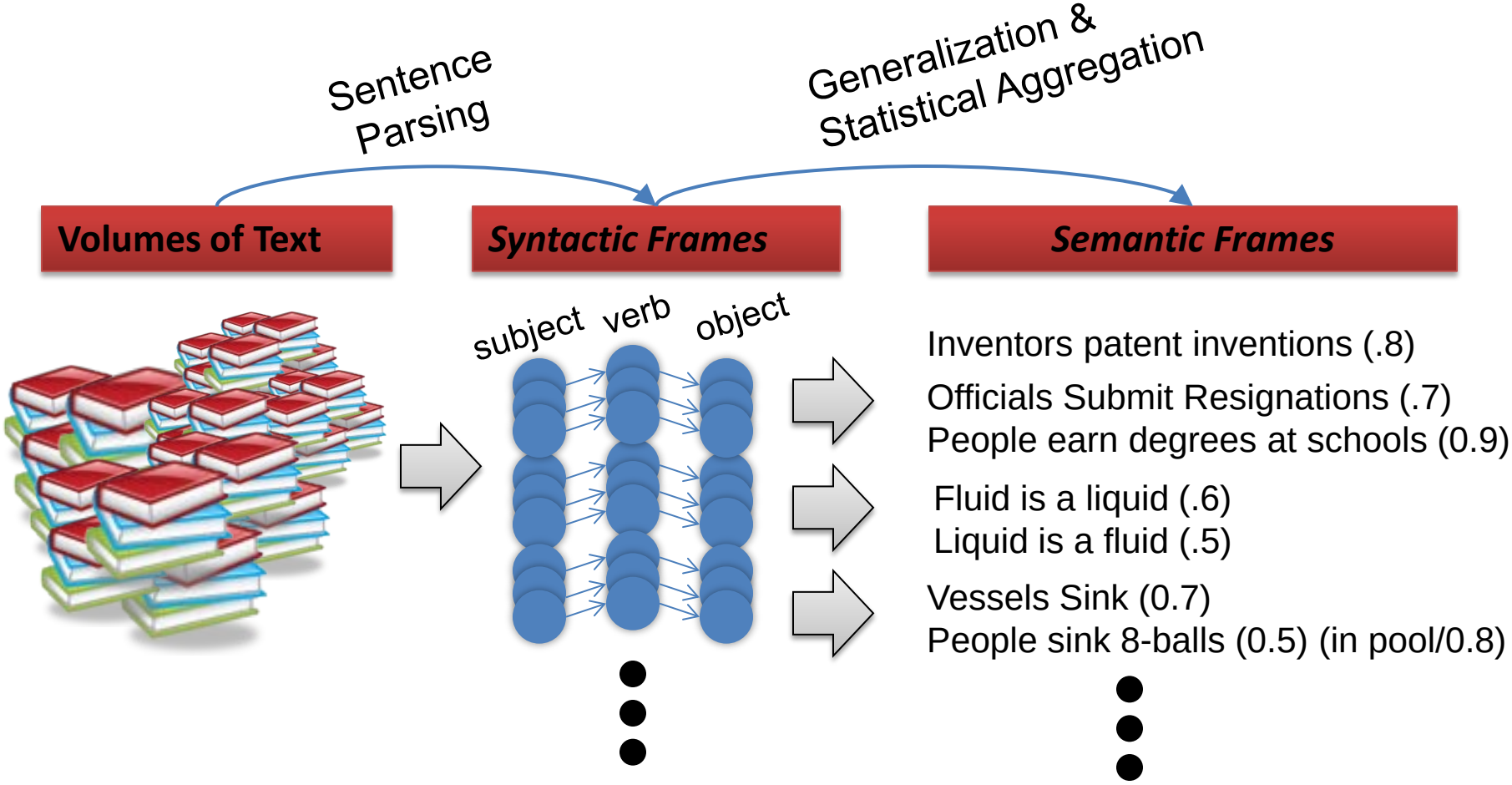
How Watson works: Step 3 Hypothesis & candidate generation



Step 3 Watson analyzes the text passages and generates possible "candidate answers".

Watson extracts important entities – so called "candidate answers" – from the documents. The focus is on coverage, which means that as much as possible is added (here, peaks, mountain ranges, people). At that stage, these are just possible answers to Watson.

Automatic Learning for “Reading”



© 2009 IBM Corporation

Evaluating Possibilities and Their Evidence

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

- *Organelle*
- *Vacuole*
- *Cytoplasm*
- *Plasma*
- *Mitochondria*
- *Blood …*

- Many candidate answers (CAs) are generated from many different searches
- Each possibility is evaluated according to different dimensions of evidence.
- Just One piece of evidence is if the CA is of the right type. In this case a "liquid".

Is("Cytoplasm", "liquid") = 0.2 ↑

Is("organelle", "liquid") = 0.1

Is("vacuole", "liquid") = 0.2

Is("plasma", "liquid") = 0.7

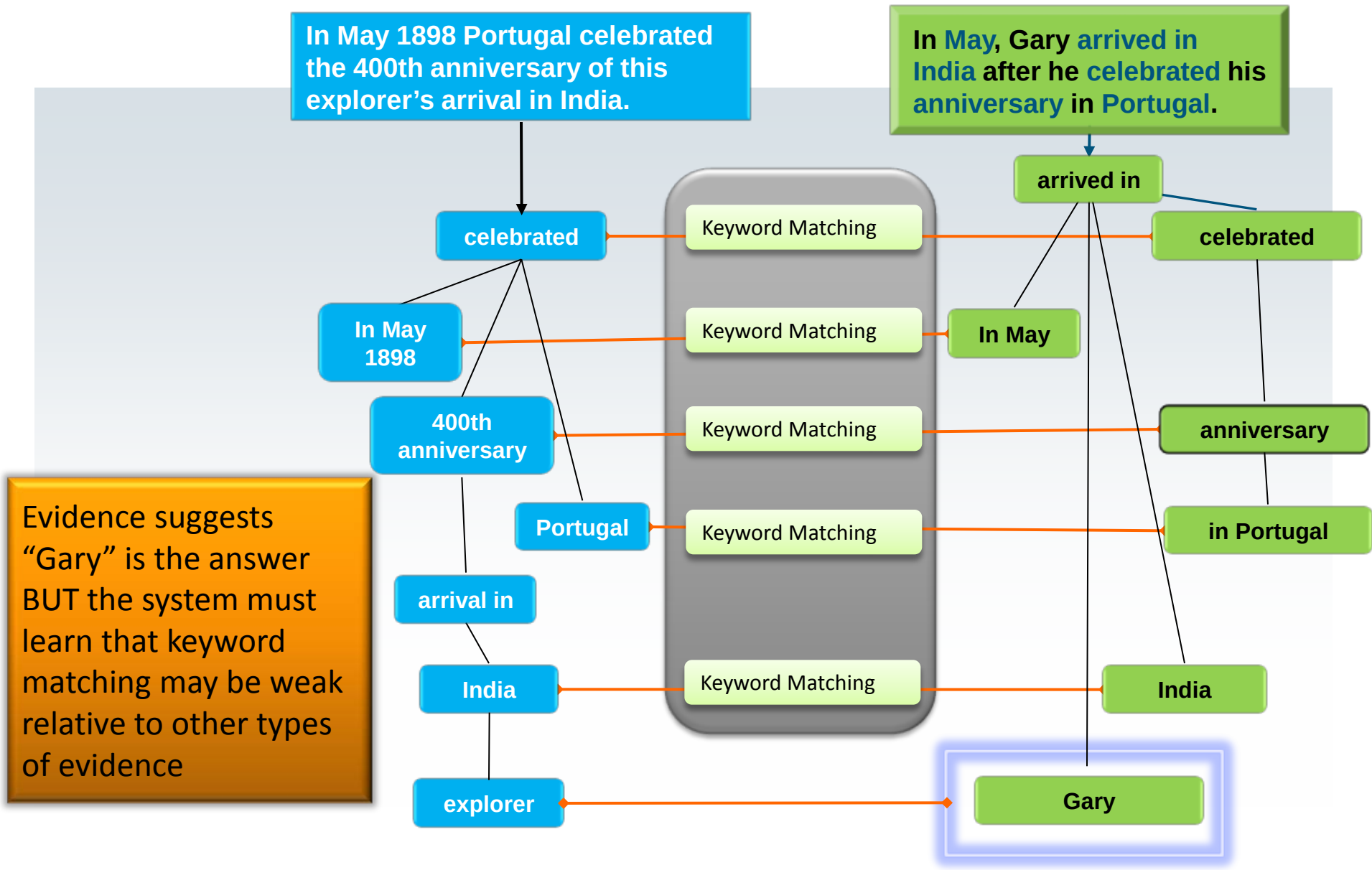
"Cytoplasm is a fluid *surrounding* the nucleus…"

Wordnet → Is_a(Fluid, Liquid) → ?

Learned → Is_a(Fluid, Liquid) → yes.

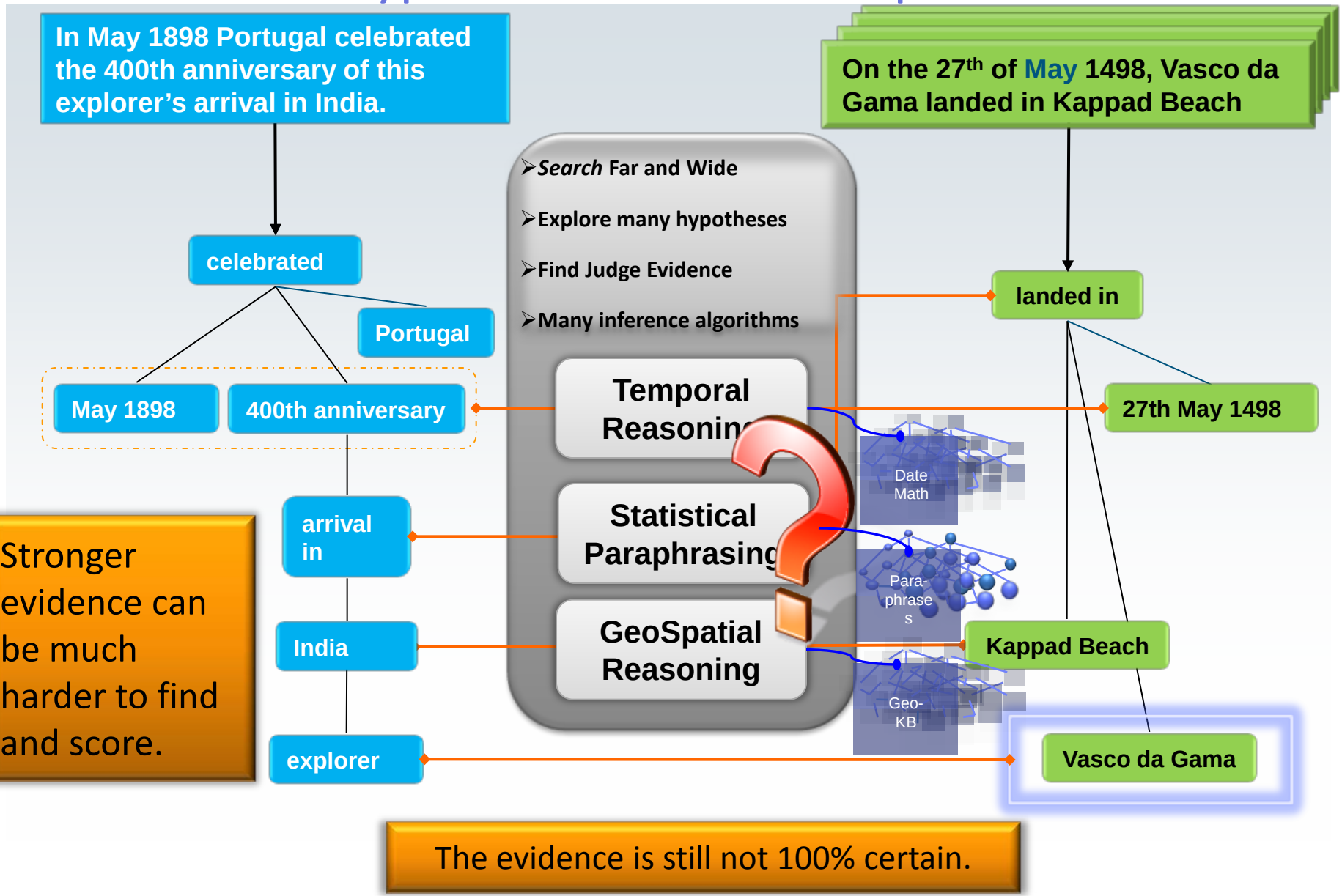
© 2009 IBM Corporation

Different Types of Evidence: Keyword Evidence



© 2009 IBM Corporation

Different Types of Evidence: Deeper Evidence

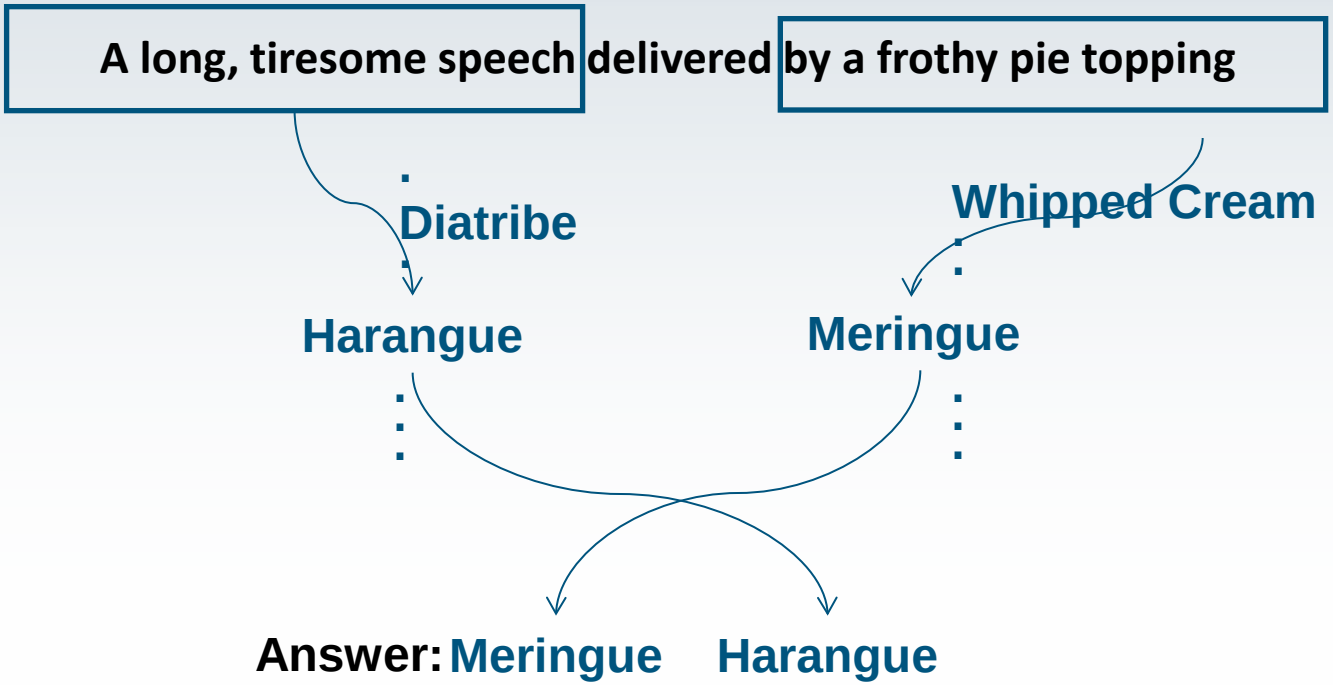


© 2009 IBM Corporation

Not ***Just*** for Fun

Category: Edible Rhyme Time

Some Questions require Decomposition and Synthesis

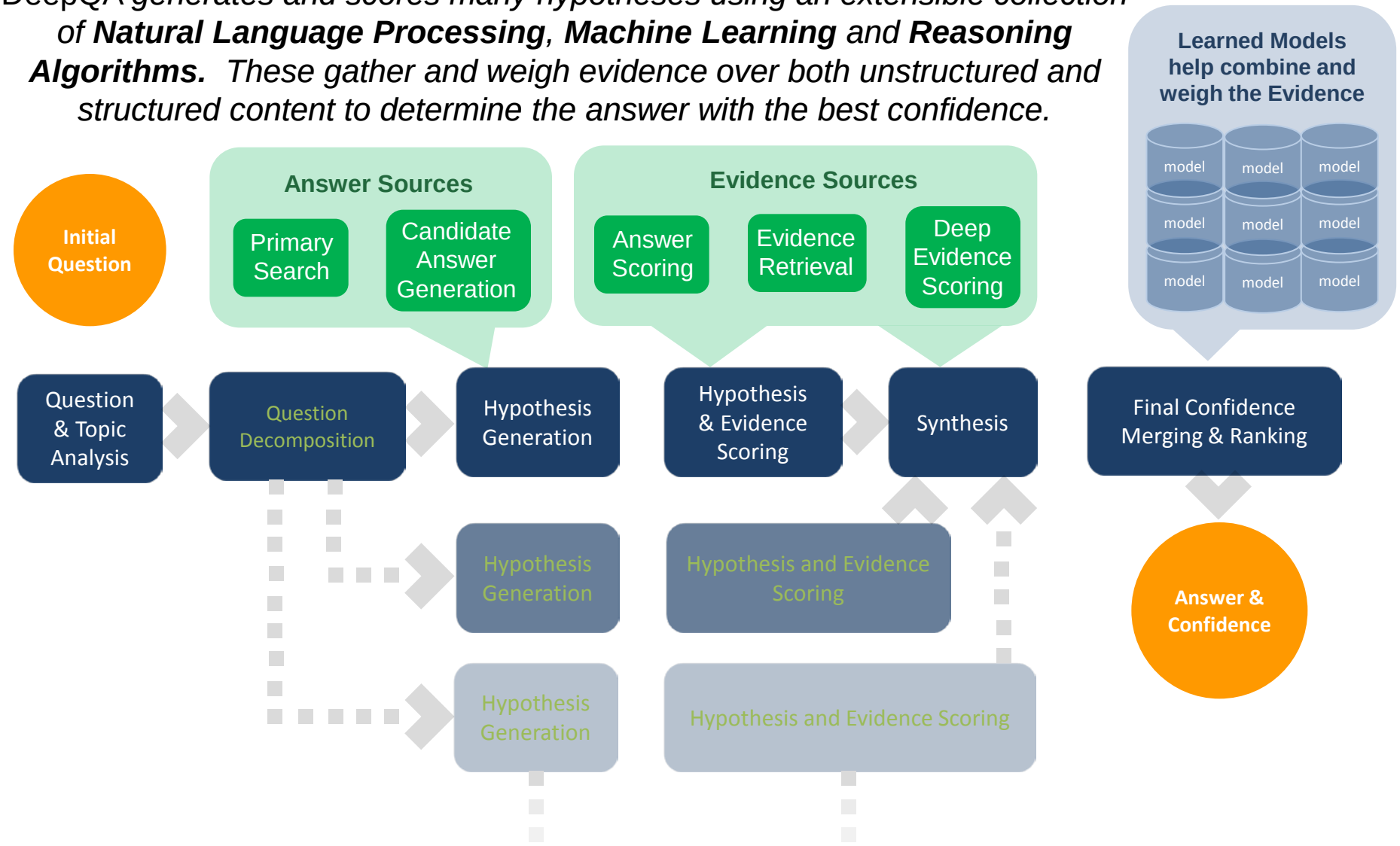


© 2009 IBM Corporation

DeepQA: the technology & architecture behind Watson:

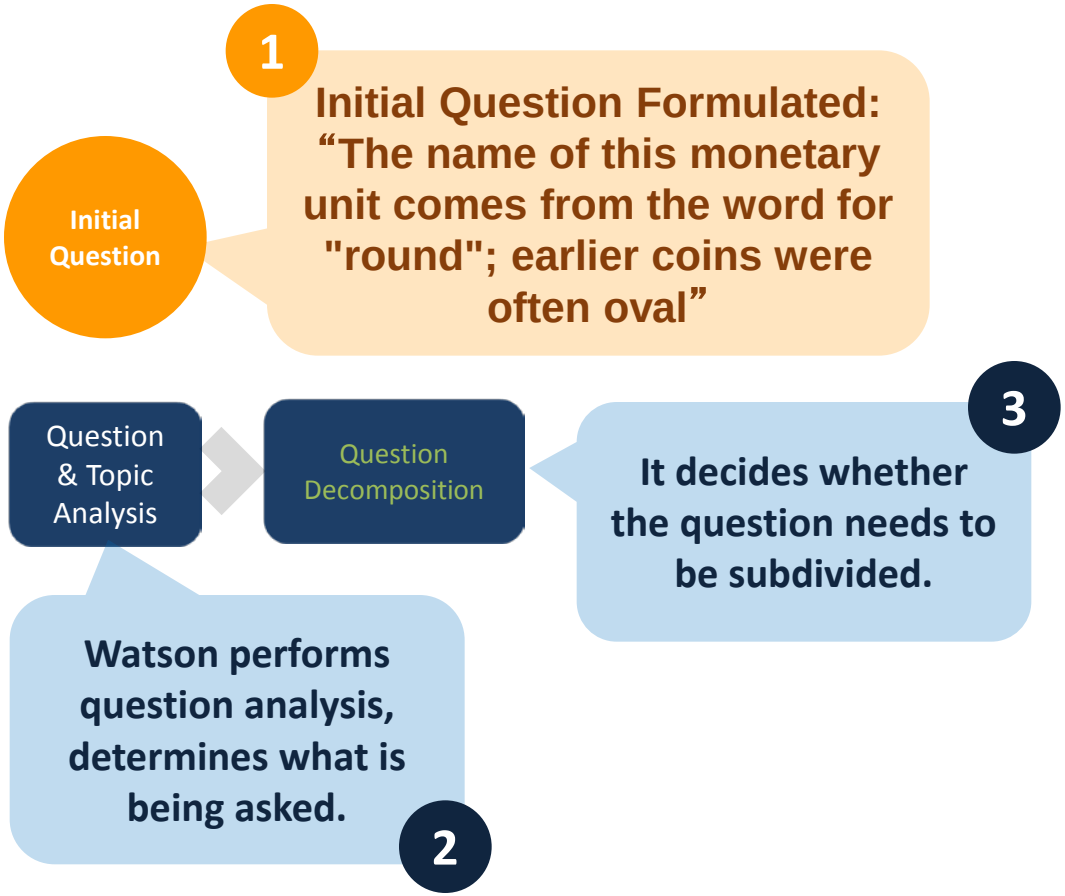
Massively Parallel Probabilistic Evidence-Based Architecture

DeepQA *generates and scores many hypotheses using an extensible collection of* ***Natural Language Processing***, ***Machine Learning*** *and* ***Reasoning Algorithms.*** *These gather and weigh evidence over both unstructured and structured content to determine the answer with the best confidence.*



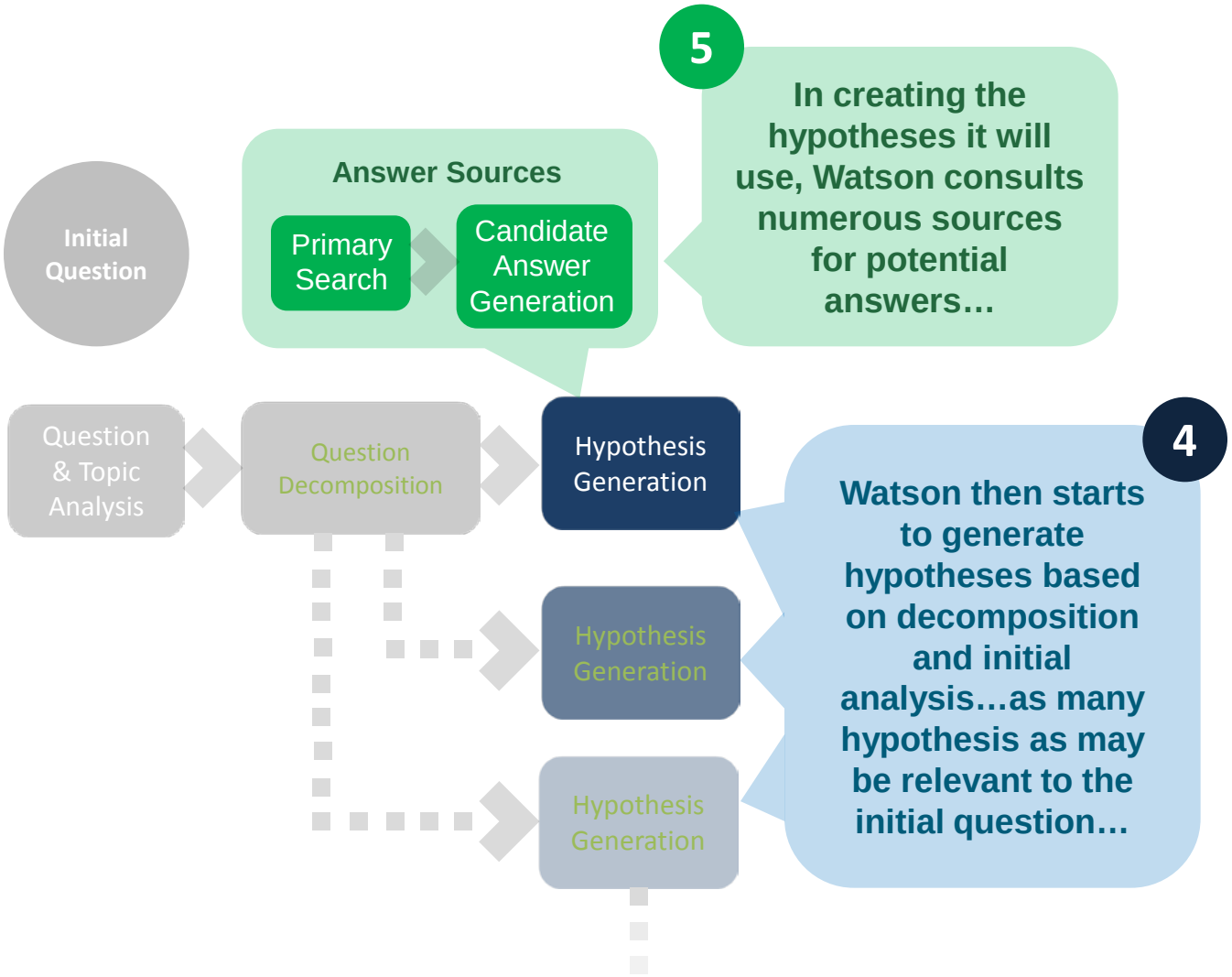
DeepQA: the technology & architecture behind Watson:

Massively Parallel Probabilistic Evidence-Based Architecture



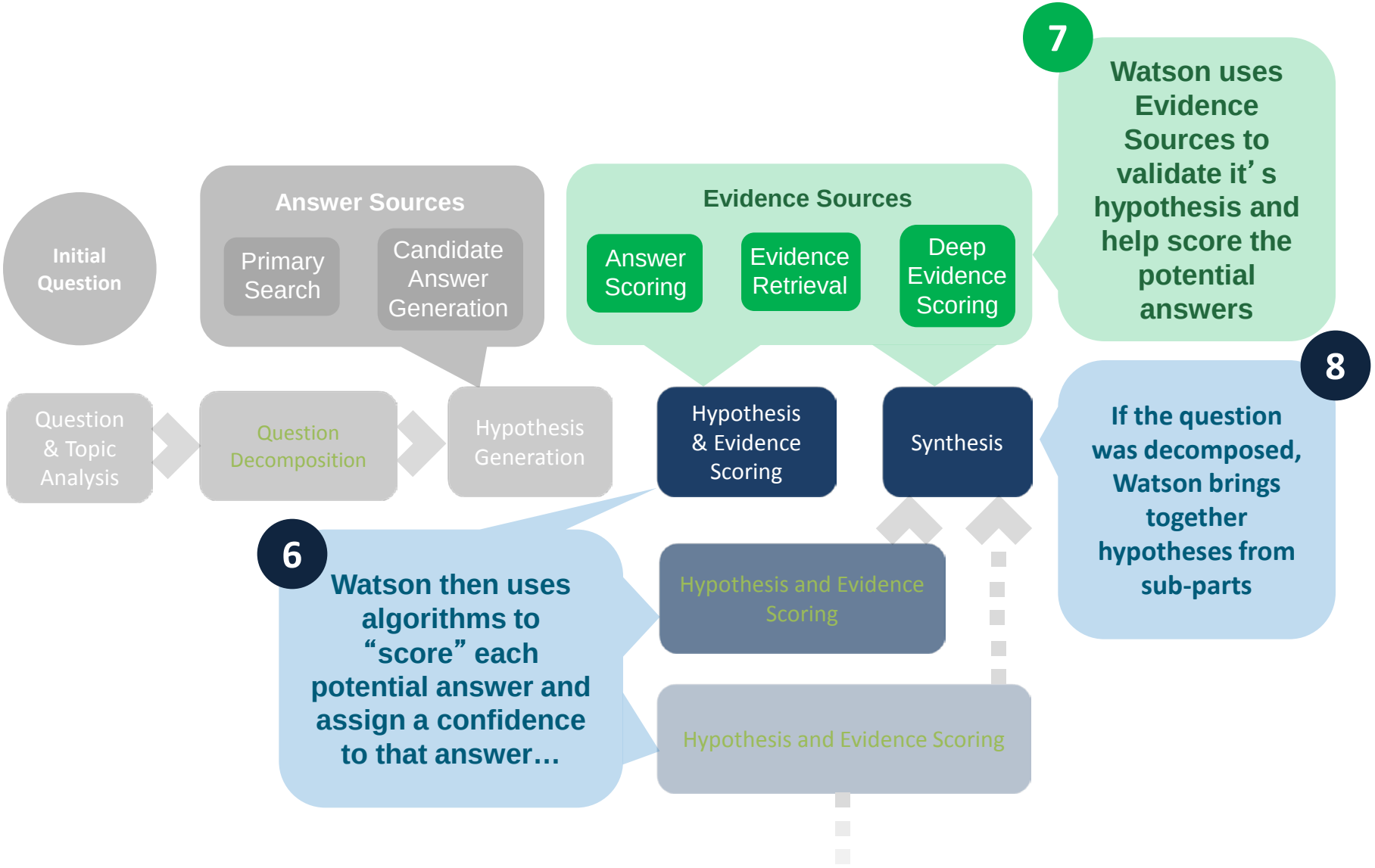
DeepQA: the technology & architecture behind Watson:

Massively Parallel Probabilistic Evidence-Based Architecture



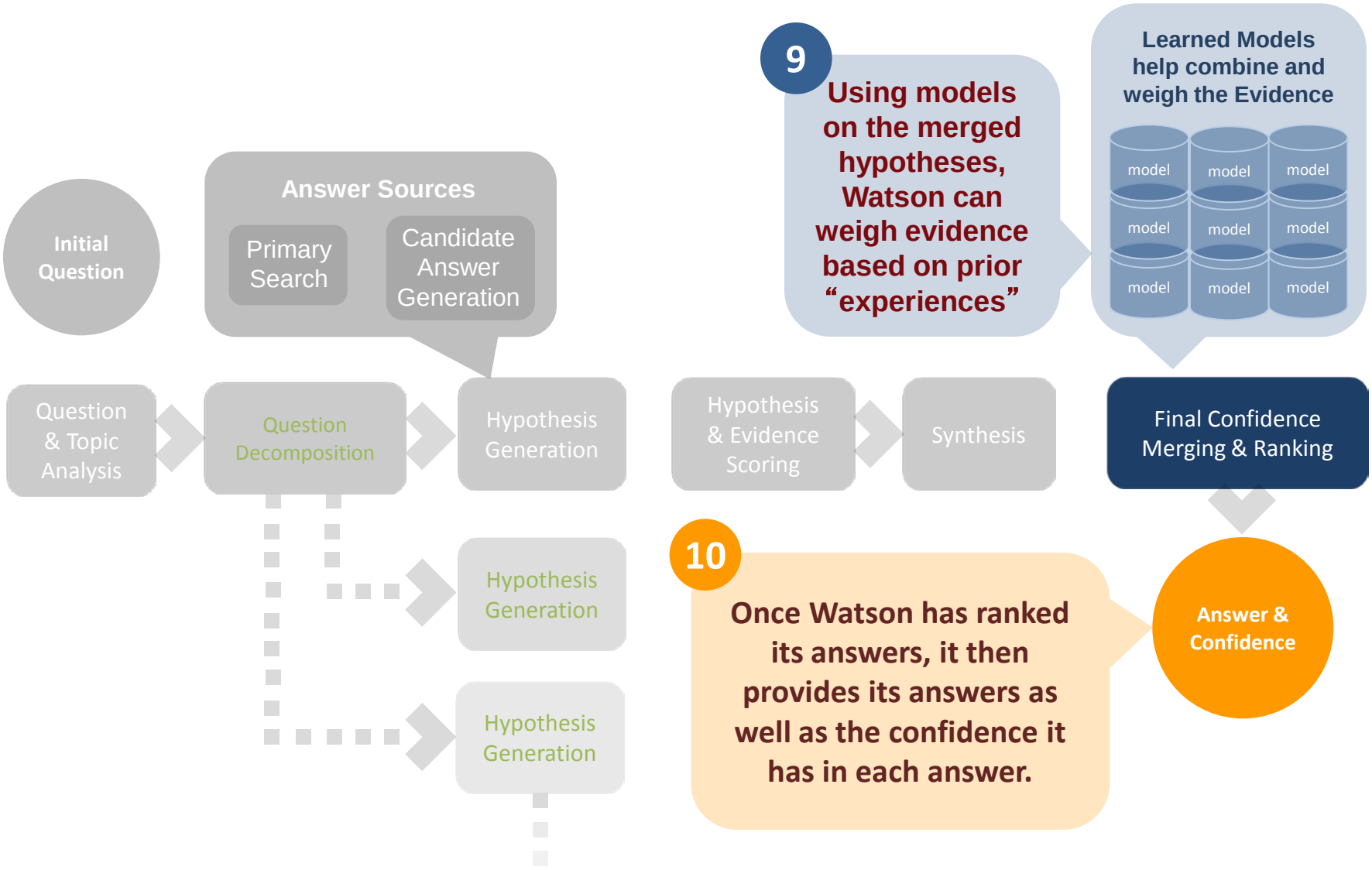
DeepQA: the technology & architecture behind Watson:

Massively Parallel Probabilistic Evidence-Based Architecture



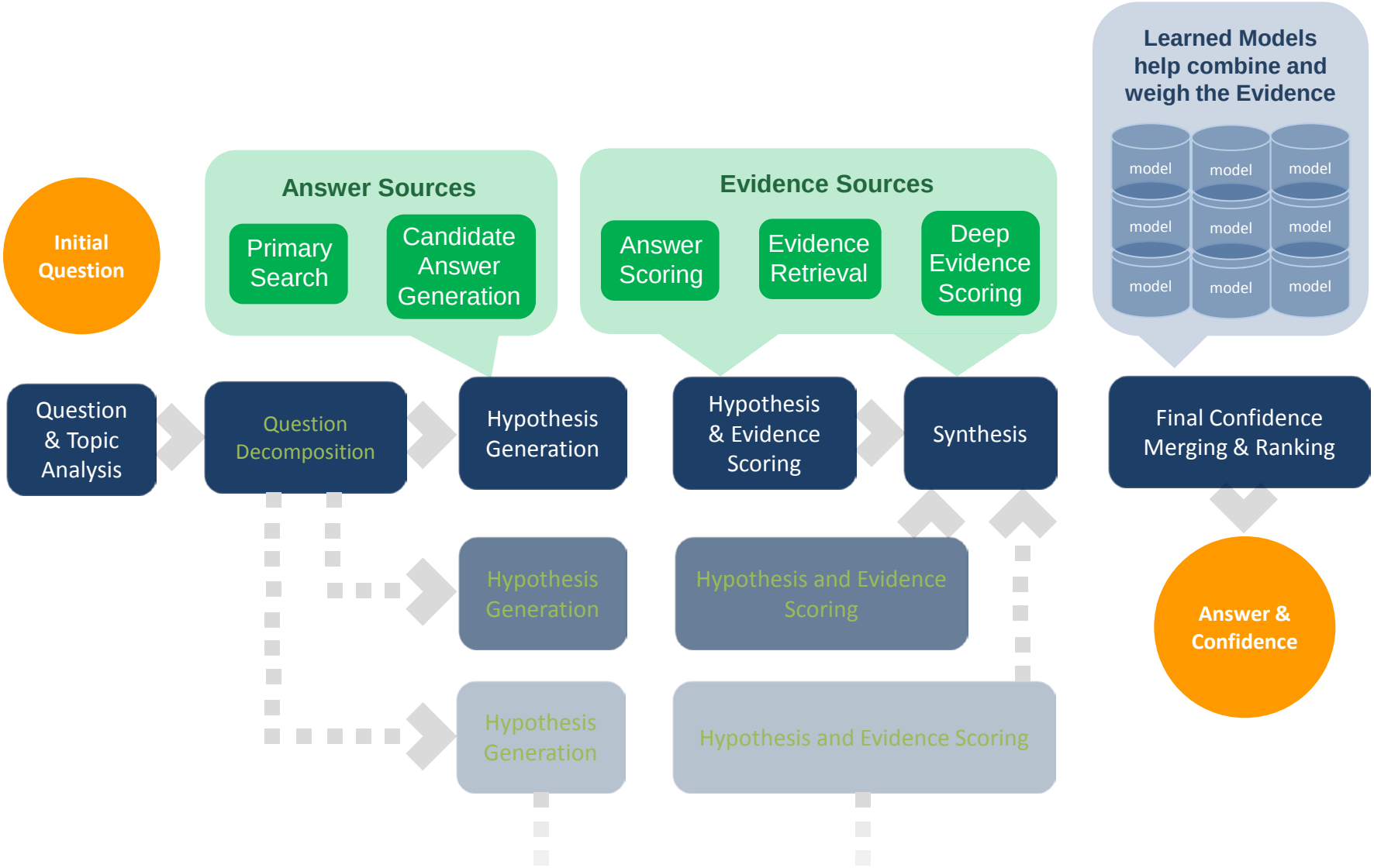
DeepQA: the technology & architecture behind Watson:

Massively Parallel Probabilistic Evidence-Based Architecture



DeepQA: the technology & architecture behind Watson:

Massively Parallel Probabilistic Evidence-Based Architecture



How we convert data into knowledge for Watson's use

Three types of knowledge

- **Domain Data** (articles, books, documents)
- **Training and test question sets w/answer keys**
- **NLP Resources** (vocabularies, taxonomies, ontologies)

Converted to Indices for search/passage lookup

Redirects extracted for disambiguation

Frame cuts generated with frequencies to determine likely context

Pseudo docs extracted for Candidate answer generation

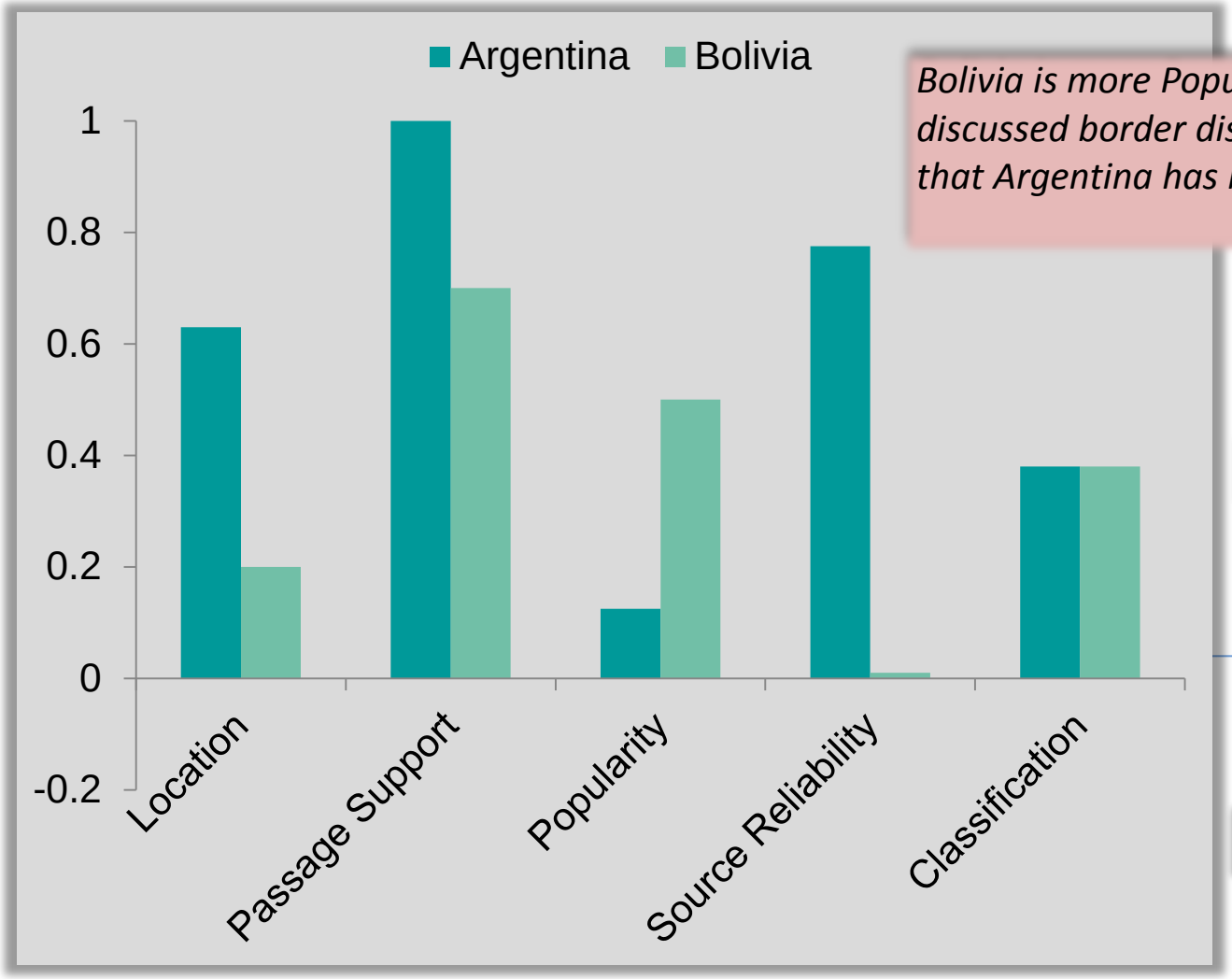
Used to create logistic regression model that Watson uses for merging scores

Named entity detection, relationship detection algorithms

Custom slot grammar parsers, prolog rules for semantic analysis

Grouping features to produce Evidence Profiles

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



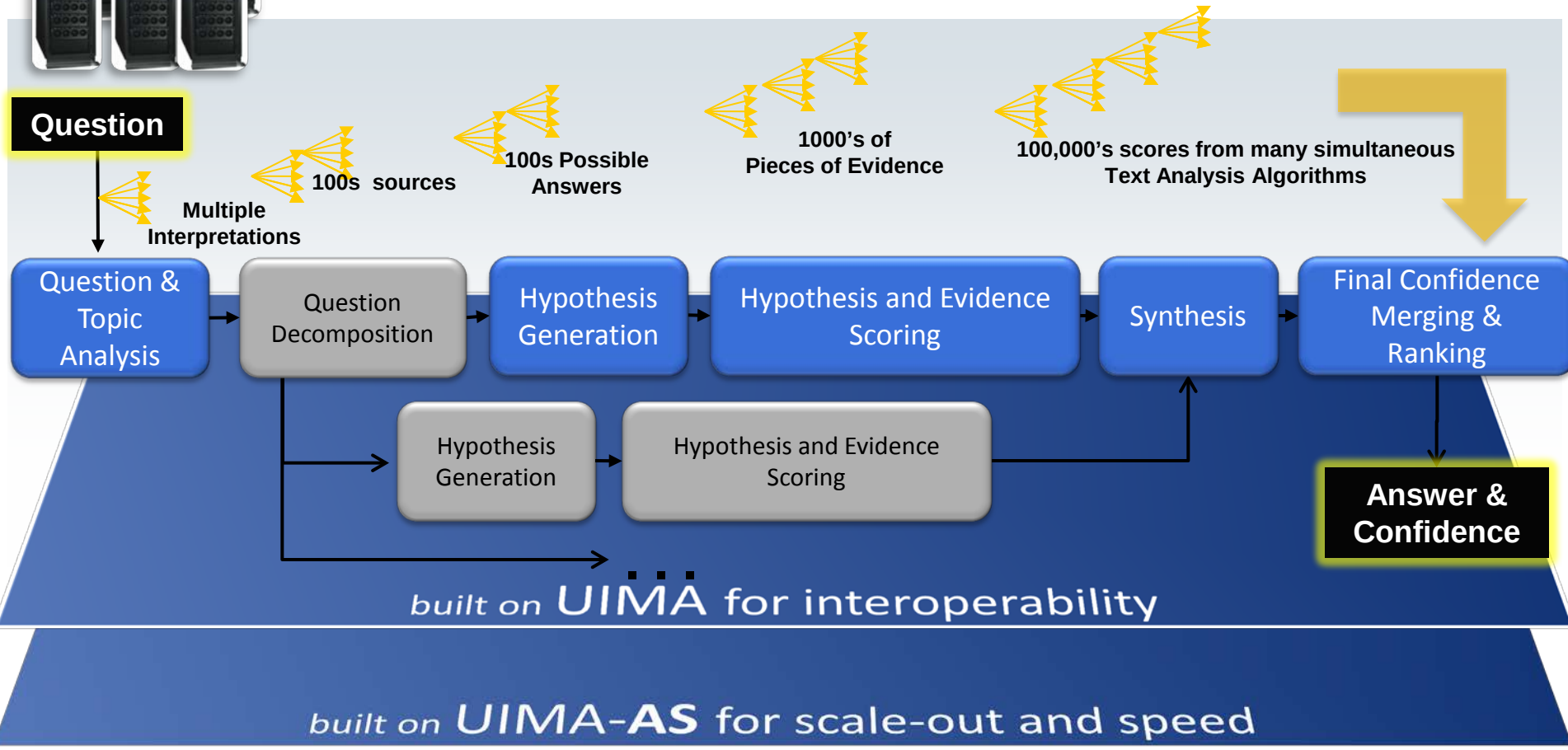
Bolivia is more Popular due to a commonly discussed border dispute. But Watson learns that Argentina has better evidence.

Positive Evidence

Negative Evidence

© 2009 IBM Corporation

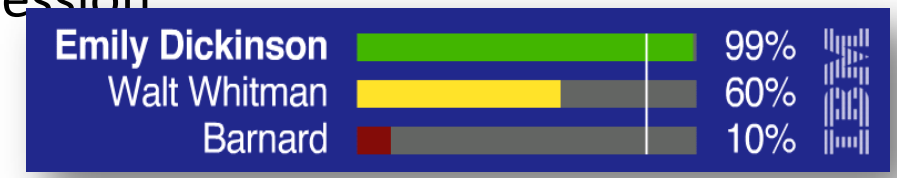
One Jeopardy! question can take **2 hours on a single 2.6Ghz Core**
Optimized & Scaled out on 2880-Core IBM workload optimized POWER7
HPC using UIMA-AS,
Watson answers in 2-6 seconds.

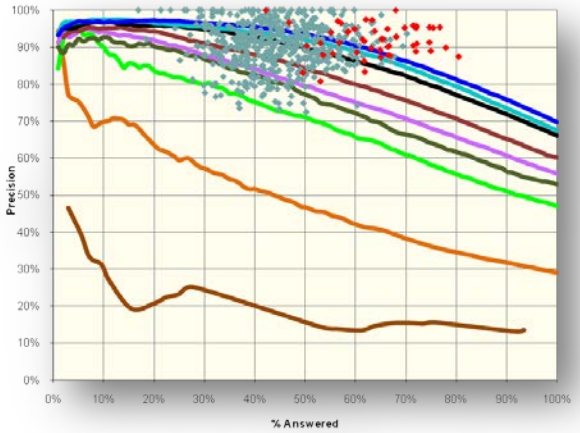


© 2009 IBM Corporation

Watson: Precision, Confidence & Speed

- **Deep Analytics –** We achieved champion-levels of *Precision* and *Confidence* over a huge variety of expression



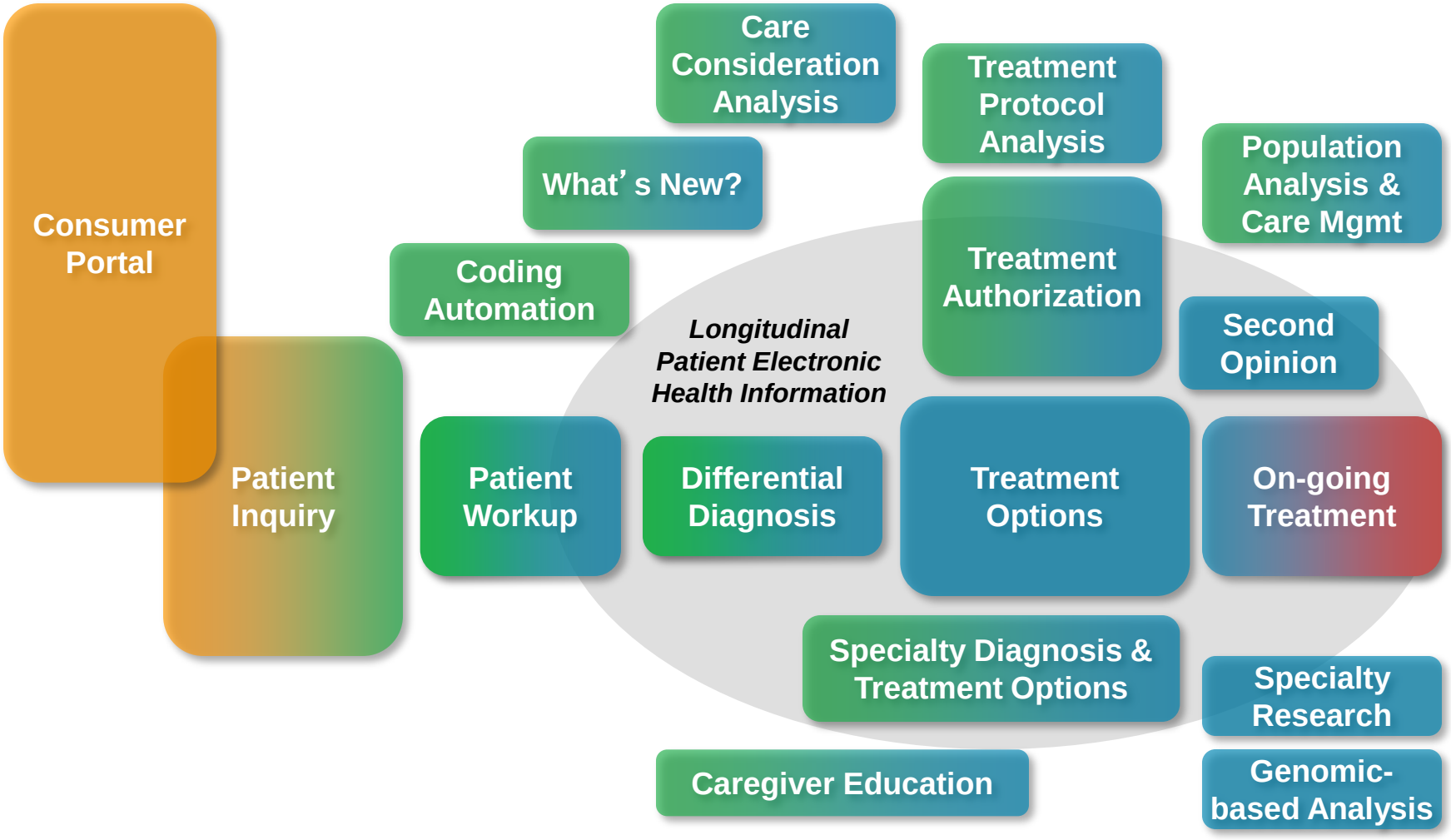


- **Speed** – By optimizing Watson's computation for Jeopardy! on **2,880 POWER7** processing cores we went **from 2 hours per question on a single CPU to an average of just 3 seconds** – fast enough to compete with the best.

- **Results –** in 55 real-time sparring against former **Tournament of Champion Players last year**, Watson put on a very competitive performance, winning 71%. In the final Exhibition Match against Ken Jennings and Brad Rutter, Watson won!

© 2009 IBM Corporation

Watson-enabled patient-centered healthcare solutions



Patient *Lay Caregiver…PA…* *Nurse Practitioner* *Physician*

Potential Business Applications

Healthcare / Life Sciences: Diagnostic Assistance, Evidenced-Based, Collaborative Medicine

Tech Support: Help-desk, Contact Centers

Enterprise Knowledge Management and Business Intelligence

Government: Improved Information Sharing and Security

© 2009 IBM Corporation

Our Study Path Forward for "Natural Language for Communication"

- Groundwork:
  - Review of probability: Ch. 13
  - Probabilistic reasoning over time: Ch. 15.1-15.3
  - Language models: Ch. 22.1
- Natural language for communication: Ch. 23