IBM Watson: a real NLP Application

Ruihong Huang

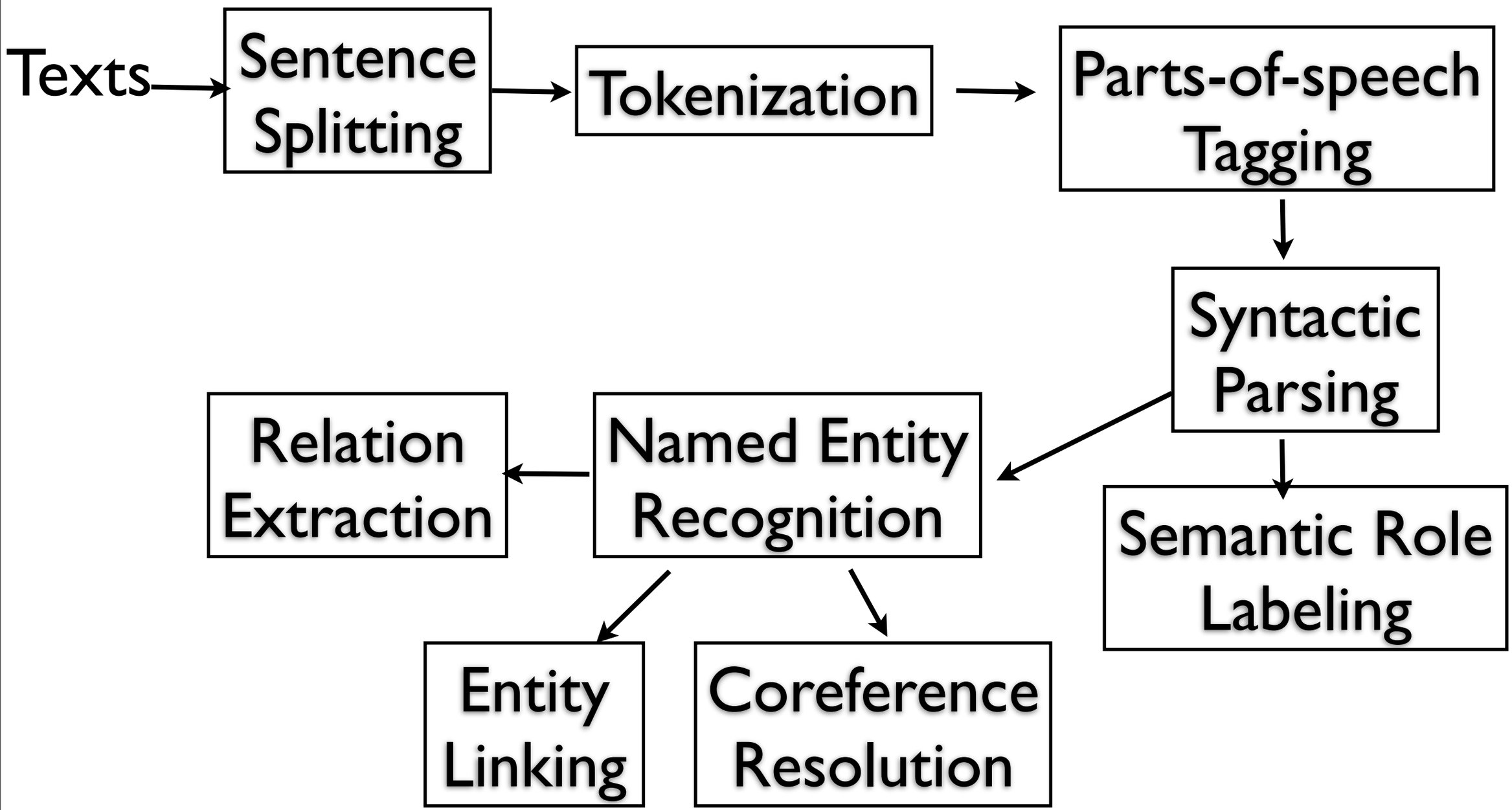
Assistant Professor

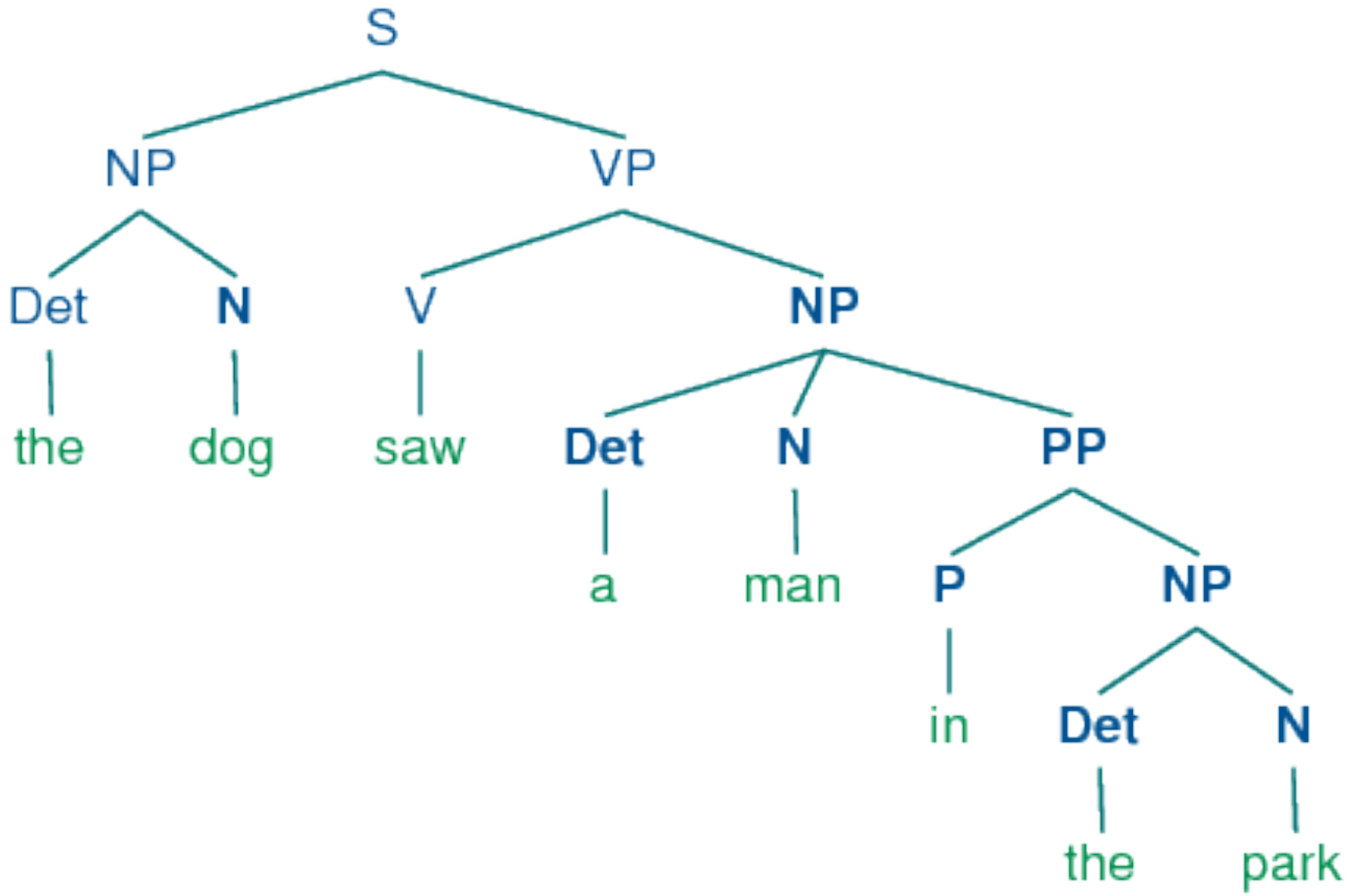
Computer Science and Engineering
Texas A&M University

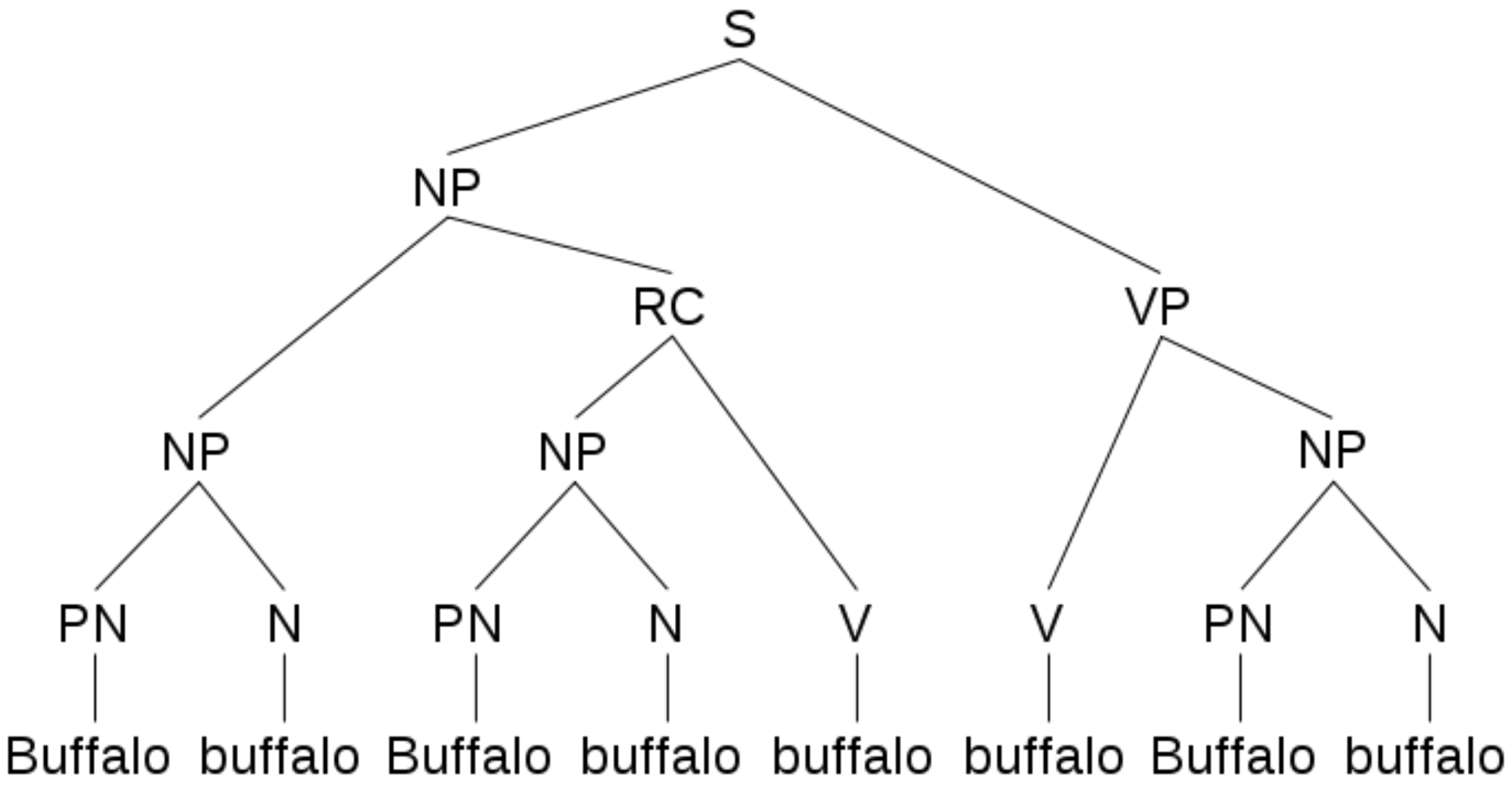
NLP Applications around us

- Email Spam Detection
- Apple Siri
- Google Translate
- Search Engines: google, yahoo
- .....

a Typical NLP Pipeline




S
NP
VP
Det
N
V
NP
the
dog
saw
Det
N
PP
a
man
P
NP
in
Det
N
the
park


S
NP
RC
VP
NP
NP
NP
PN
N
PN
N
V
V
PN
N
Buffalo
buffalo
Buffalo
buffalo
buffalo
buffalo
Buffalo
buffalo

Other Tasks

- Paraphrasing
- Textual Entailment
- Word Sense Disambiguation
- Semantic Parsing
- .....

Applications

- Text Summarization
- Machine Translation
- Text Generation
- Event Extraction
- Question Answering
- Dialogue Generation
- .....

Dan Jurafsky

Language Technology

making good progress

mostly solved

Spam detection

Let's go to Agra! ✓

Buy V1AGRA … ✗

Part-of-speech (POS) tagging

ADJ	ADJ	NOUN	VERB	ADV
Colorless	green	ideas	sleep	furiously.

Named entity recognition (NER)

PERSON	ORG	LOC

Einstein met with UN officials in Princeton

Sentiment analysis

Best roast chicken in San Francisco!

The waiter ignored us for 20 minutes.

Coreference resolution

Carter told Mubarak he shouldn't run again.

Word sense disambiguation (WSD)

I need new batteries for my ***mouse***.

Parsing

I can see Alcatraz from the window!

Machine translation (MT)

第13届上海国际电影节开幕...

The 13th Shanghai International Film Festival...

Information extraction (IE)

You're invited to our dinner party, Friday May 27 at 8:30

Party
May 27
add

still really hard

Question answering (QA)

Q. How effective is ibuprofen in reducing fever in patients with acute febrile illness?

Paraphrase

XYZ acquired ABC yesterday

ABC has been taken over by XYZ

Summarization

The Dow Jones is up

The S&P500 jumped

Housing prices rose

→ Economy is good

Dialog

Where is Citizen Kane playing in SF?

Castro Theatre at 7:30. Do you want a ticket?

Question Answering

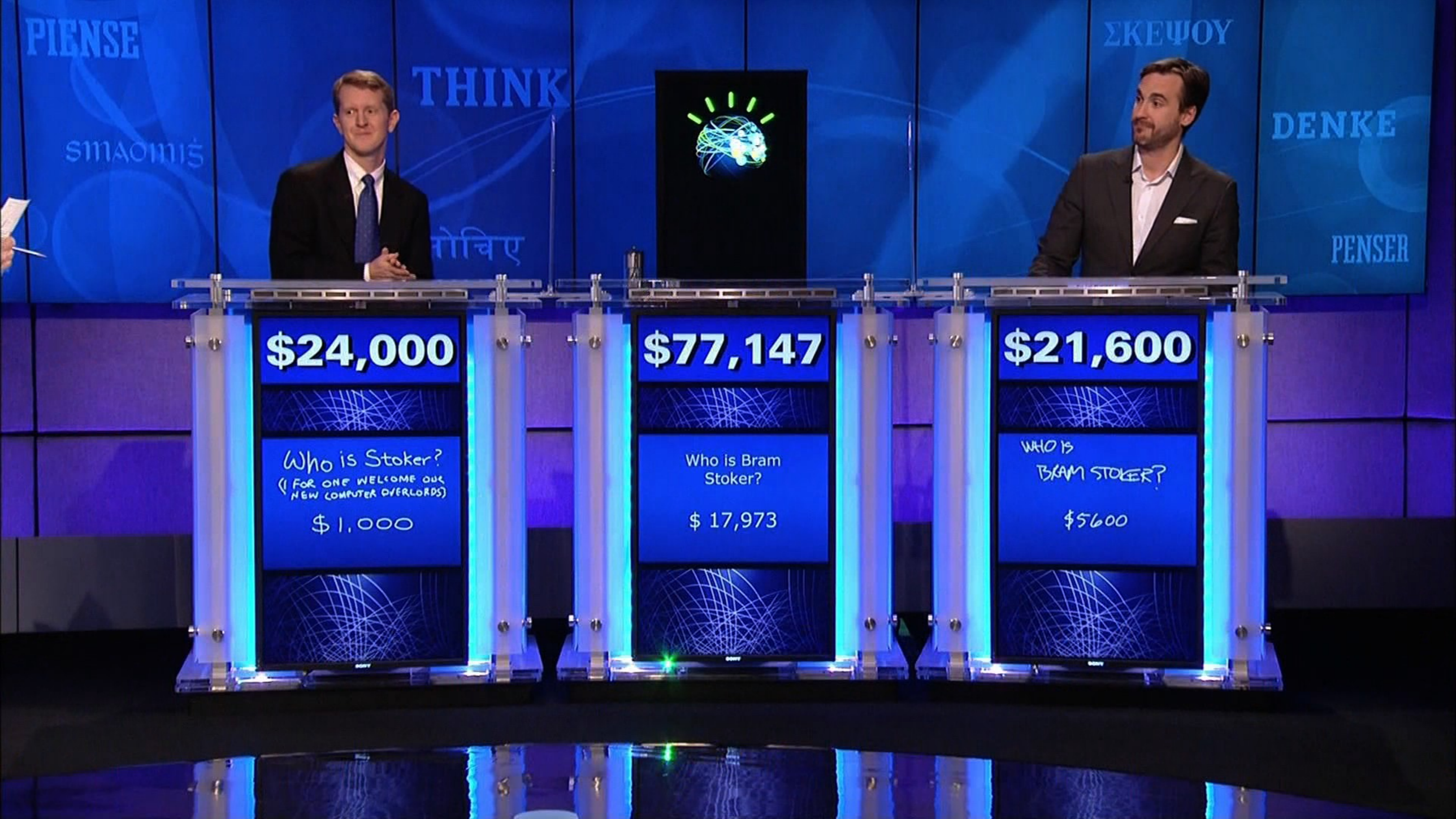
- Closed-domain: reading comprehension
- Open-domain: web-based

Question Answering

- Understanding questions
- Retrieving and processing relevant texts
- Answer Generation

Quiz

- How big is the system?
- Is Watson Online?
- open-domain or closed-domain?
- What NLP techniques used?


PIENSE
THINK
ΣΚΕΨΟΥ
DENKE
PENSER
$24,000
Who is Stoker?
(I FOR ONE WELCOME OUR NEW COMPUTER OVERLORDS)
$ 1,000
$77,147
Who is Bram Stoker?
$ 17,973
$21,600
WHO IS BRAM STOKER?
$5600

Real-Time Game Configuration
Used in Sparring and Exhibition Games

Technology	Classics	The Great Outdoors	Speak of the Dickens	Mind Your Manners	Before and 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

Human Player 1

Human Player 2

Jeopardy! Game Control System

Strategy & Text-to-Speech

Watson's Game Controller

Clue & Category

Answers & Confidences

Watson's QA Engine

400 Processes

90 IBM Power 750s

15 TB RAM

2 TB Disk

Clues, Scores & Other Game Data

Analysis of natural language content equivalent to 1 Million Books

Watson – a Workload Optimized System

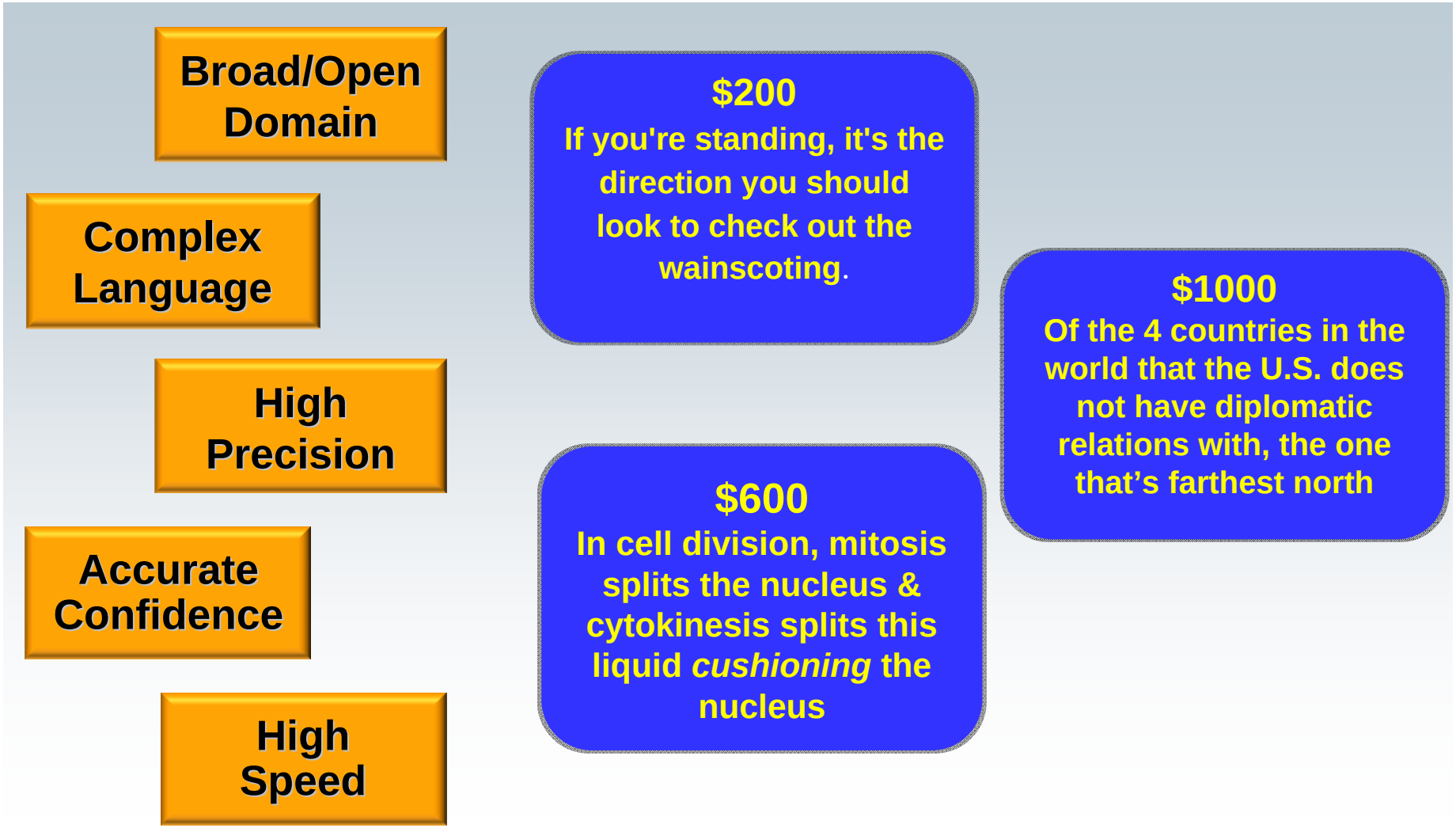
- 90 x IBM Power 750[1] servers
- 2880 POWER7 cores
- POWER7 3.55 GHz chip
- 500 GB per sec on-chip bandwidth
- 10 Gb Ethernet network
- 15 Terabytes of memory
- 20 Terabytes of disk, clustered
- Can operate at 80 Teraflops
- Runs IBM DeepQA software
- Scales out with and searches vast amounts of unstructured 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

The Jeopardy! Challenge: *A compelling and notable way to **drive** and **measure** the technology **of automatic Question Answering** along 5 Key Dimensions*



Real Language is Real Hard

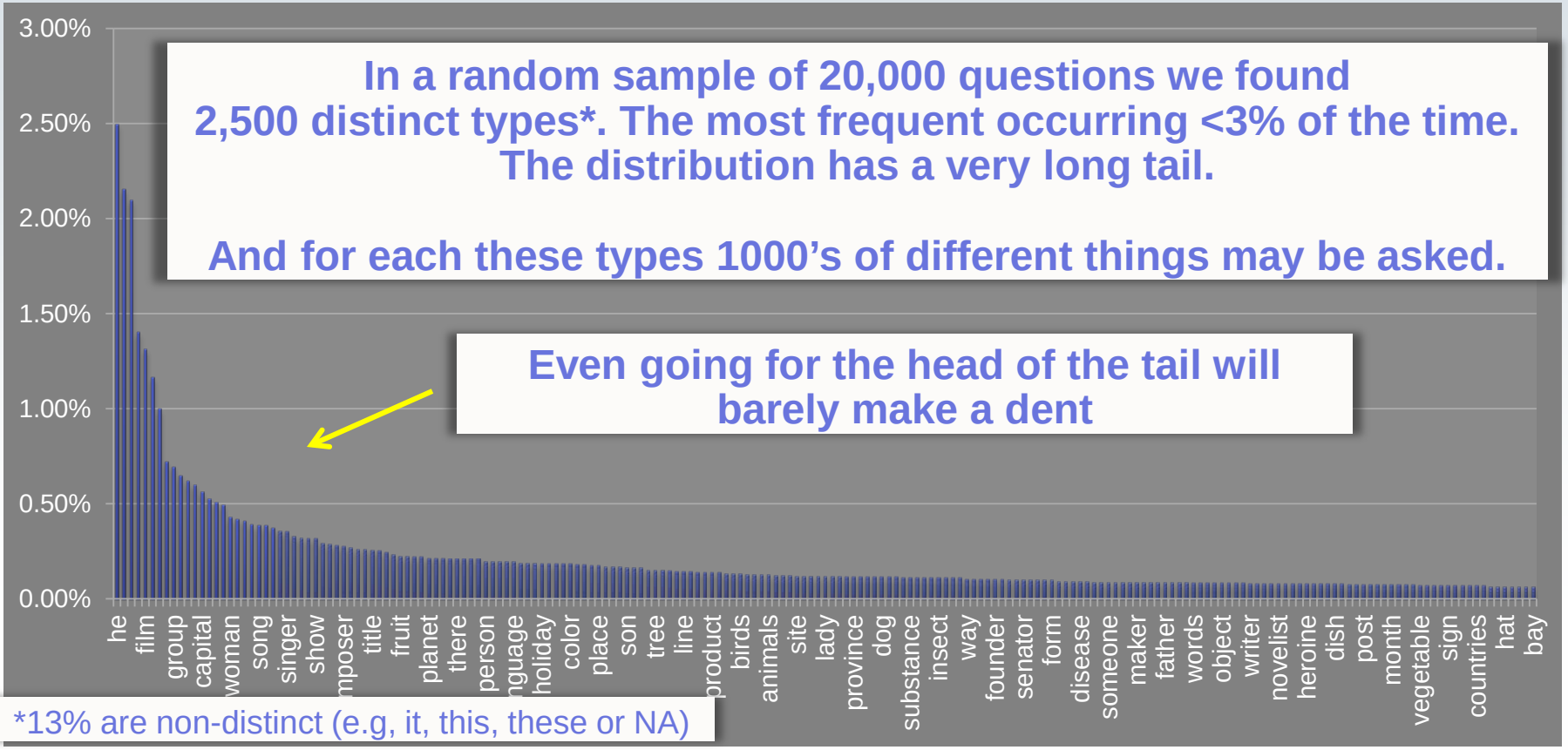
- Chess
  - A finite, mathematically well-defined search space
  - Limited number of moves and states
  - Grounded in **explicit**, **unambiguous** mathematical rules

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

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.

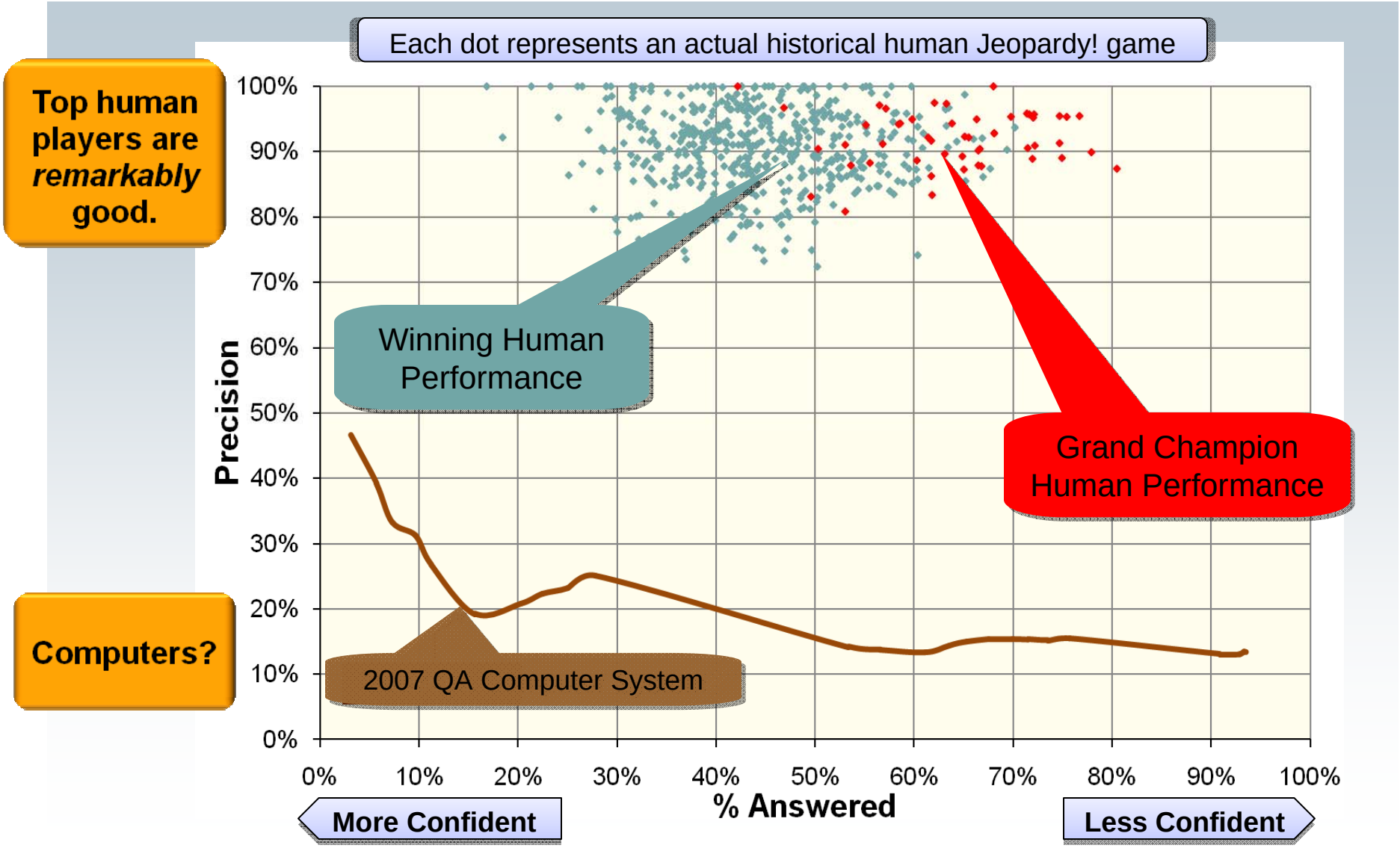
And for each these types 1000's of different things may be asked.

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.

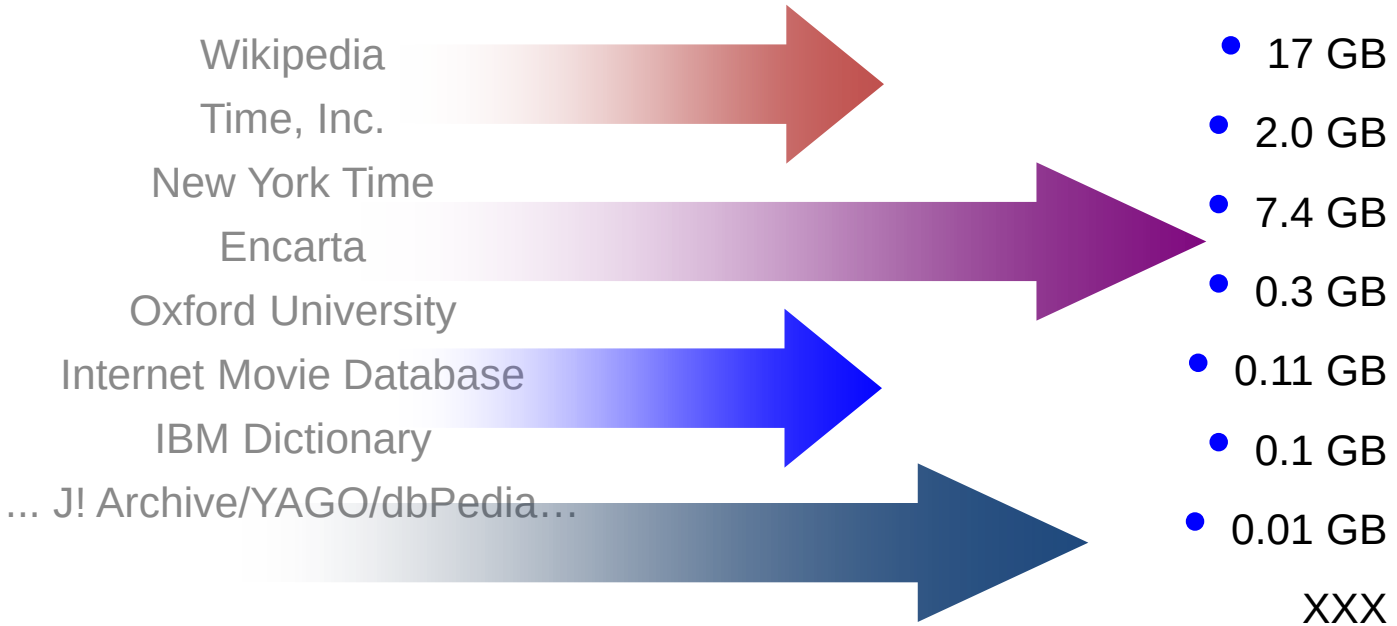
The Best Human Performance: *Our Analysis Reveals the **Winner's Cloud***



Where did it acquire knowledge?

Three types of knowledge

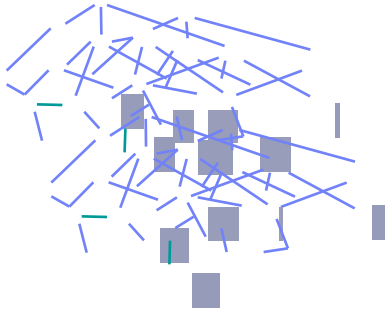
Domain Data (articles, books, documents)	**Training and test question sets w/answer keys**	**NLP Resources** (vocabularies, taxonomies, ontologies)



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

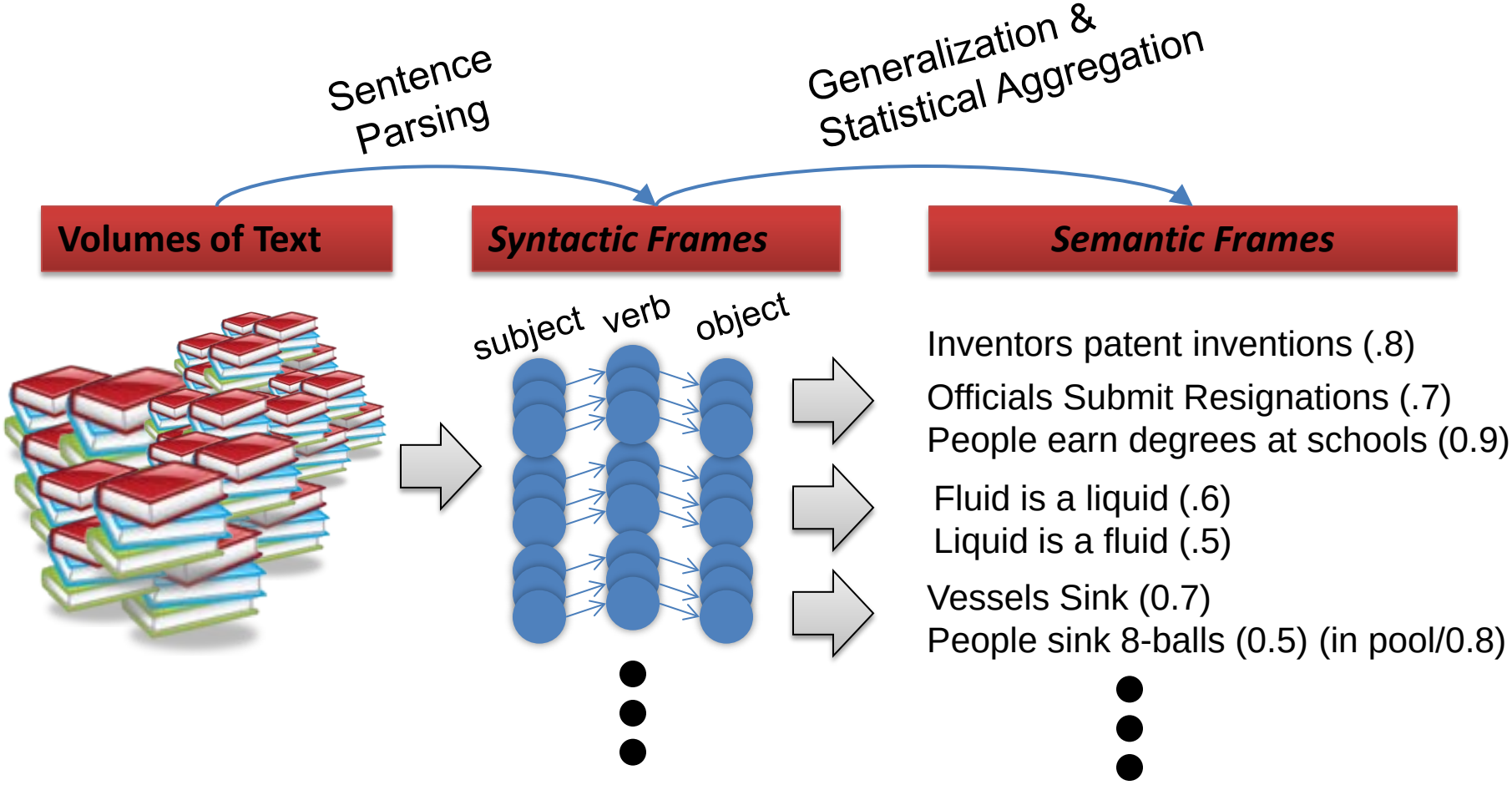
Watson's Knowledge for Jeopardy!

Watson has analyzed and stored the equivalent of about 1 million books (e.g., encyclopedias, dictionaries, news articles, reference texts, plays, etc)



Watson also uses structured sources such as WordNet and DBpedia

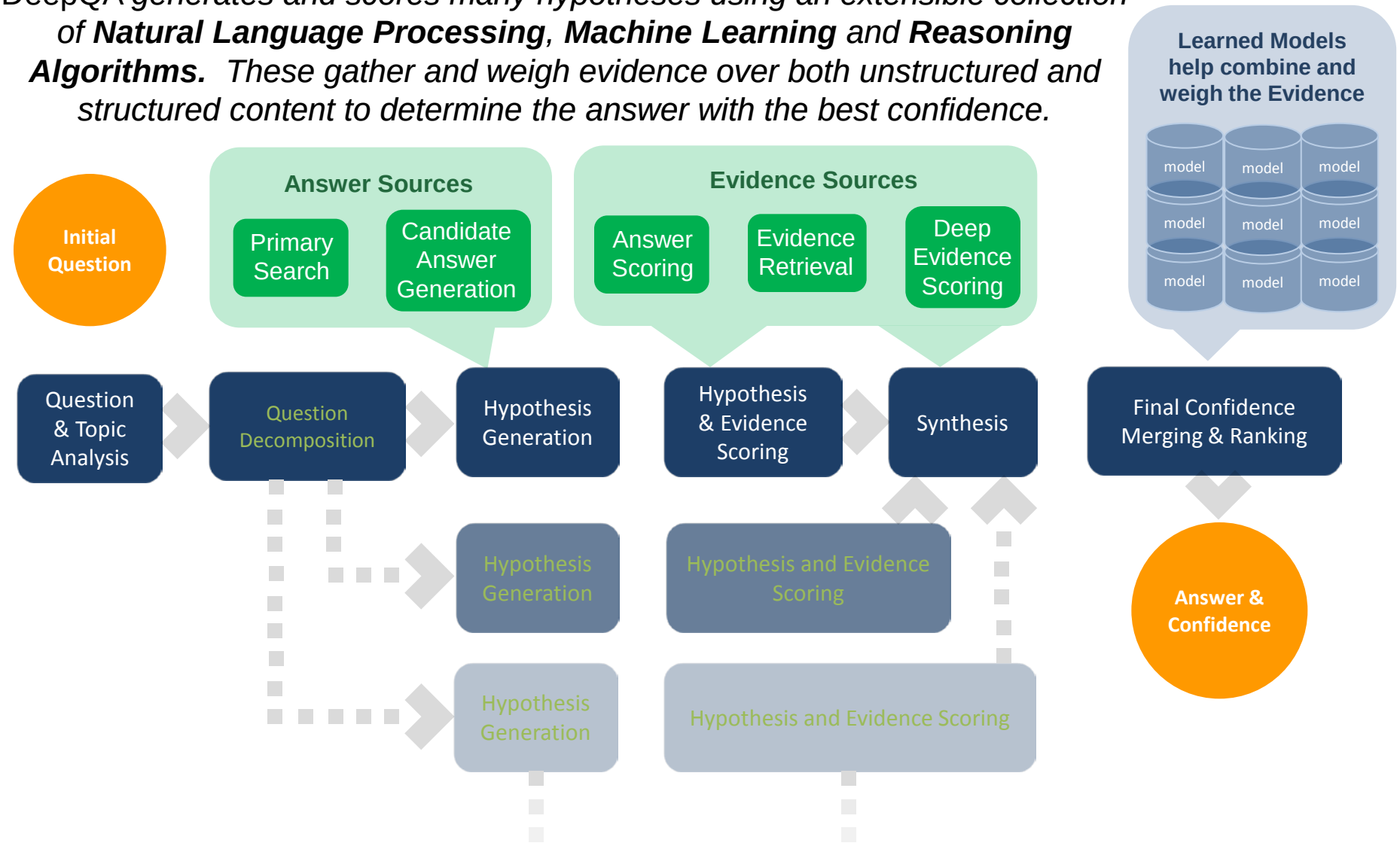
Automatic Learning by "Reading"



DeepQA: the technology & architecture behind Watson:

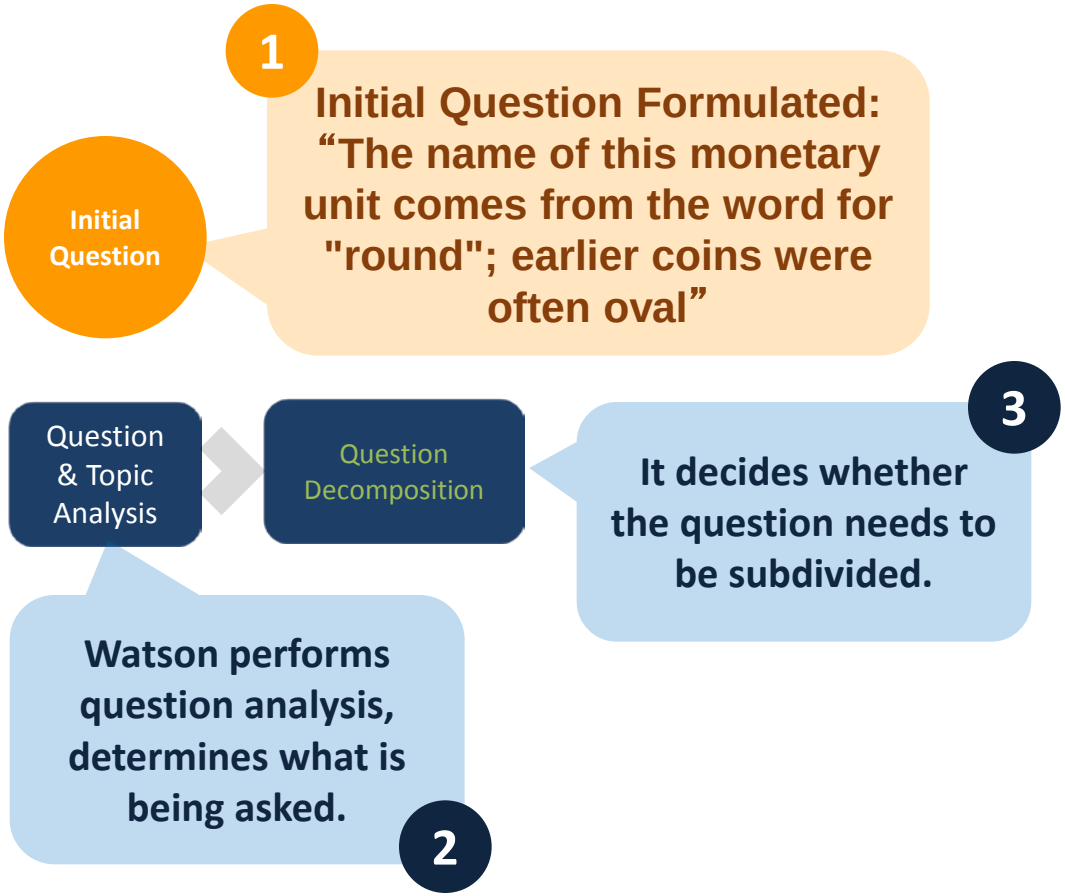
Massively Parallel Probabilistic Evidence-Based Architecture

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:

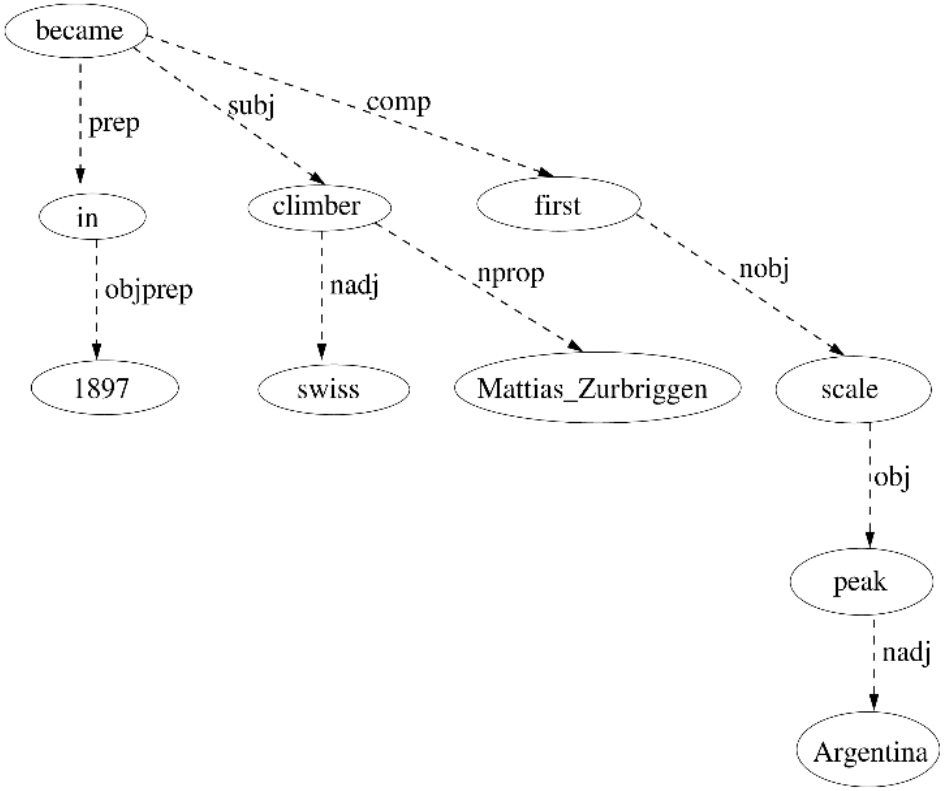
Massively Parallel Probabilistic Evidence-Based Architecture



Analyzing the question

Category:
WORLD GEOGRAPHY

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



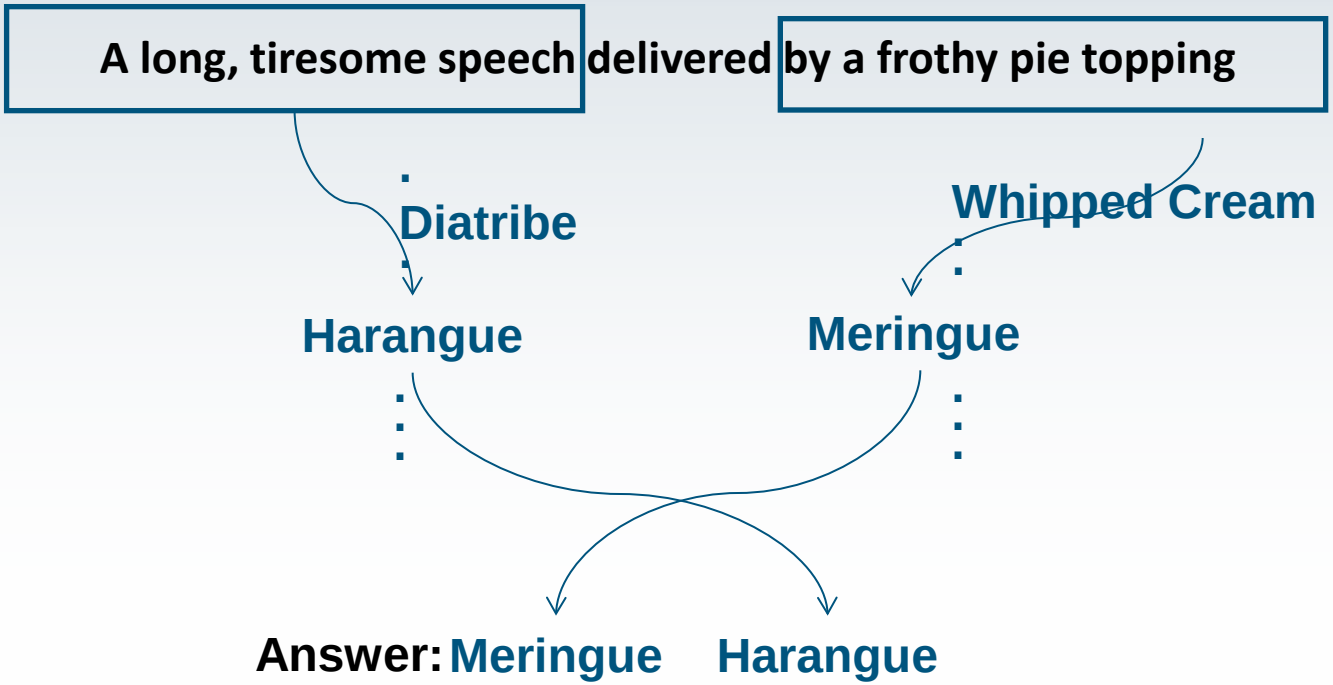
Step 1 Watson dissects the clue to understand what it is asking for.

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.

Not ***Just*** for Fun

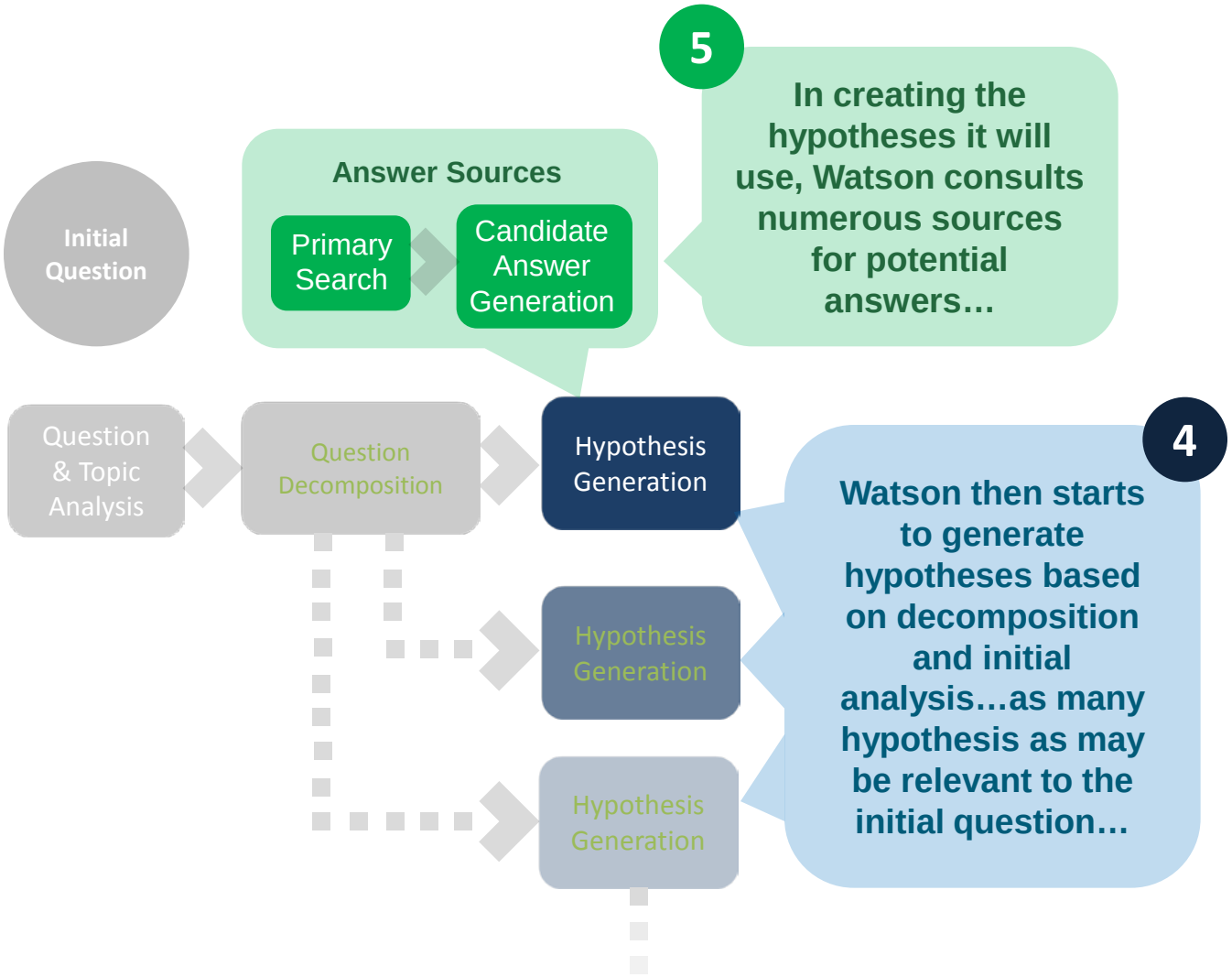
Category: Edible Rhyme Time

Some Questions require Decomposition and Synthesis



DeepQA: the technology & architecture behind Watson:

Massively Parallel Probabilistic Evidence-Based Architecture



Search

Timeline of Climbing the Matterhorn

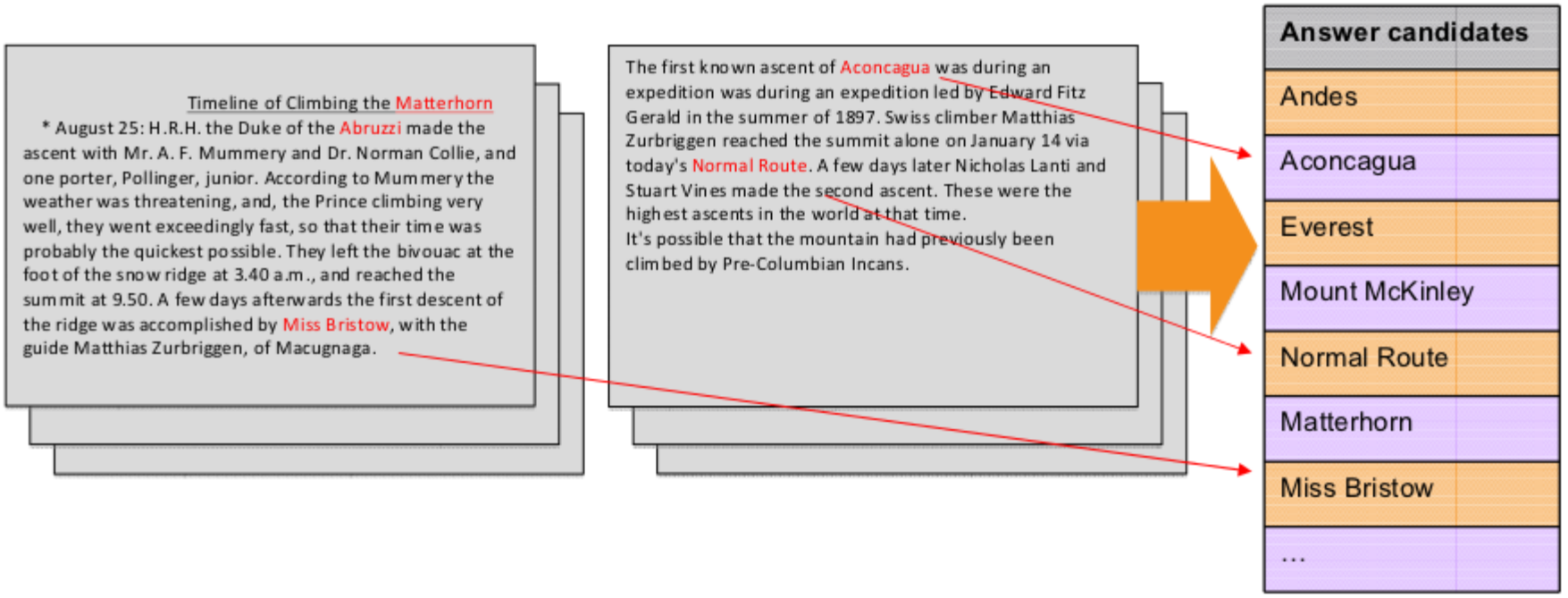
* August 25: H.R.H. the Duke of the Abruzzi made the ascent with Mr. A. F. Mummery and Dr. Norman Collie, and one porter, Pollinger, junior. According to Mummery the weather was threatening, and, the Prince climbing very well, they went exceedingly fast, so that their time was probably the quickest possible. 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.

Hypothesis & candidate generation

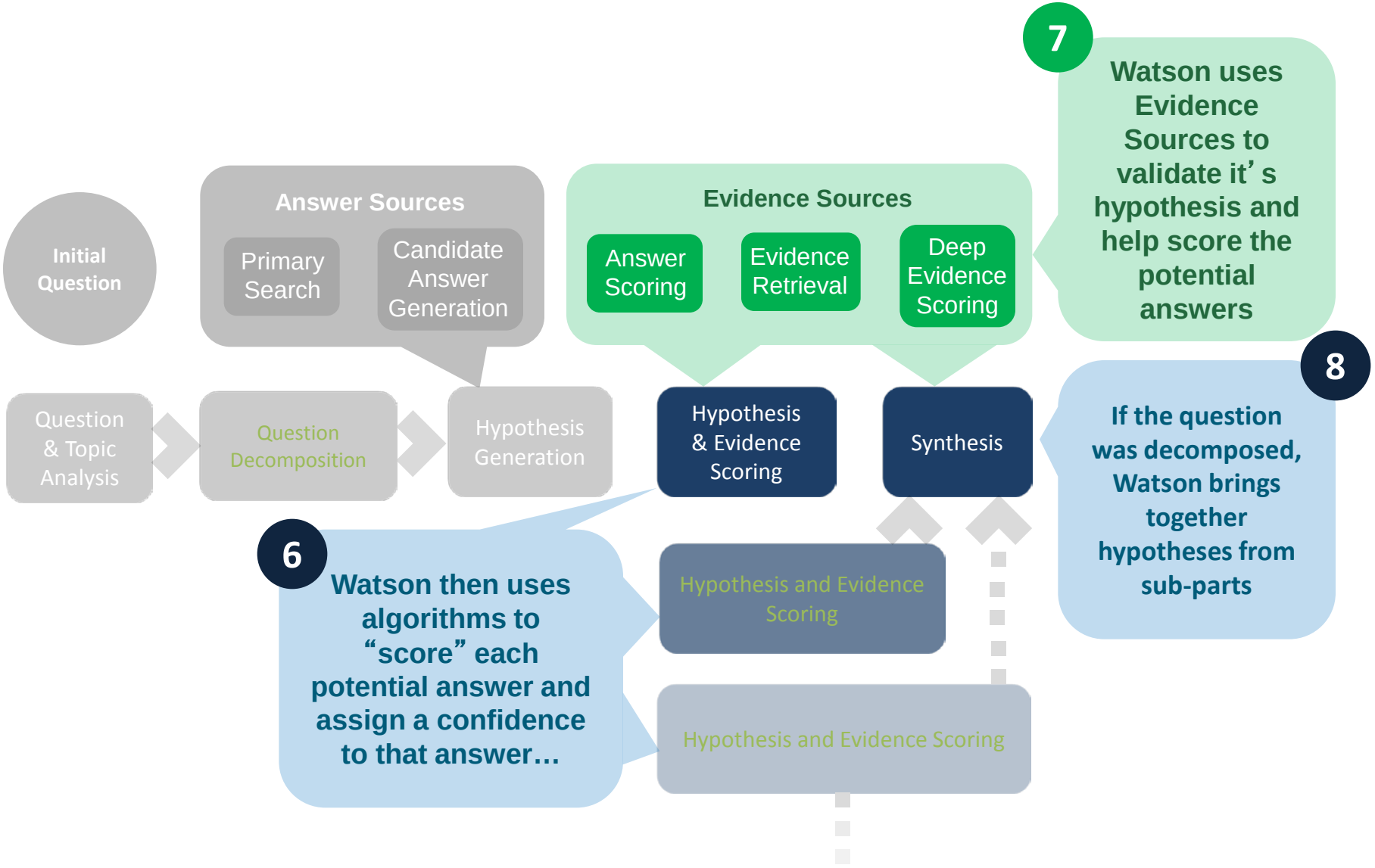


Step 3 Watson analyzes the text passages and generates possible "candidate answers".

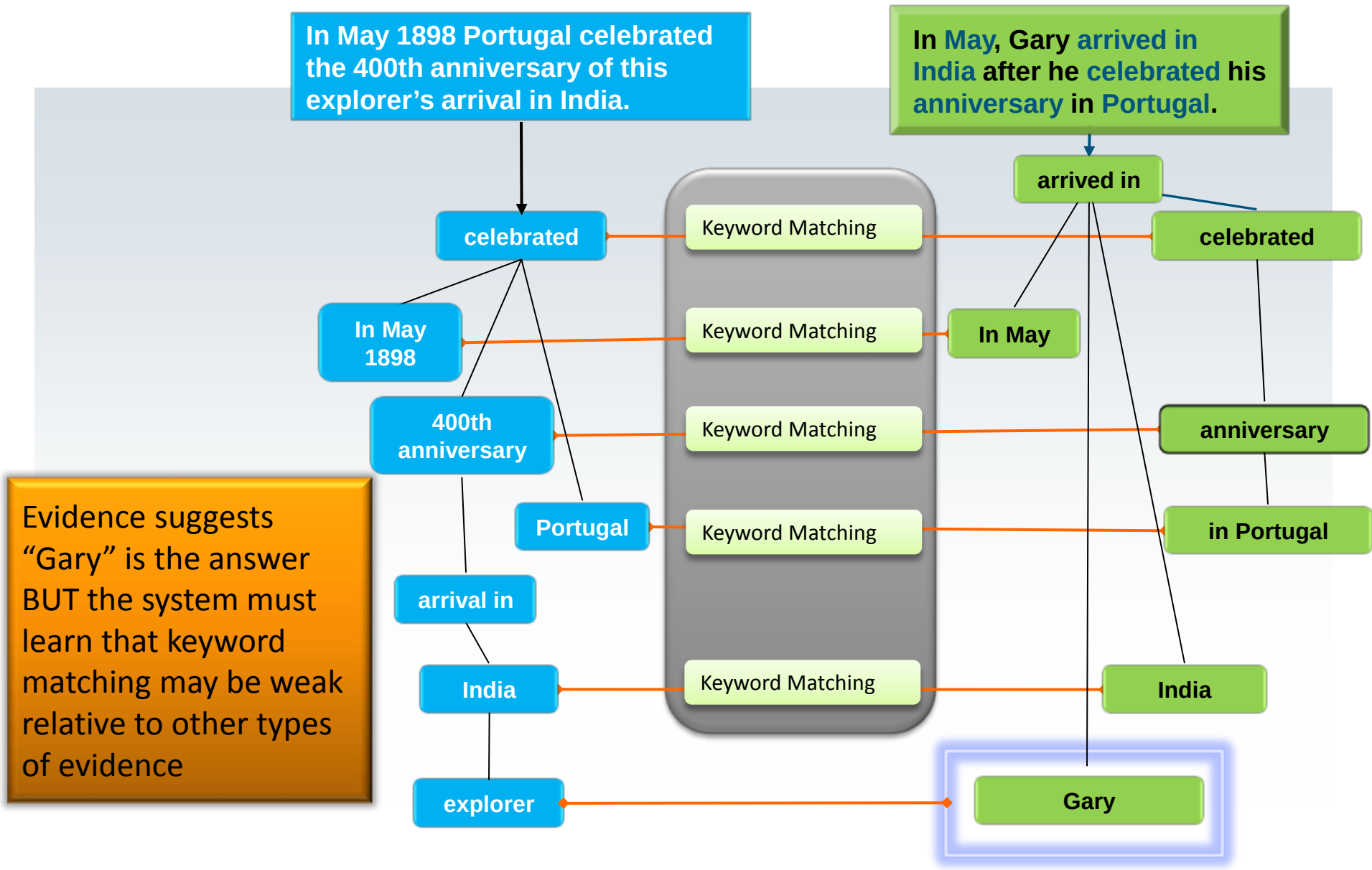
Watson extracts important entities – so called "candidate answers" – from the documents. The focus is on coverage, which means that as much as possible is added (here, peaks, mountain ranges, people). At that stage, these are just possible answers to Watson.

DeepQA: the technology & architecture behind Watson:

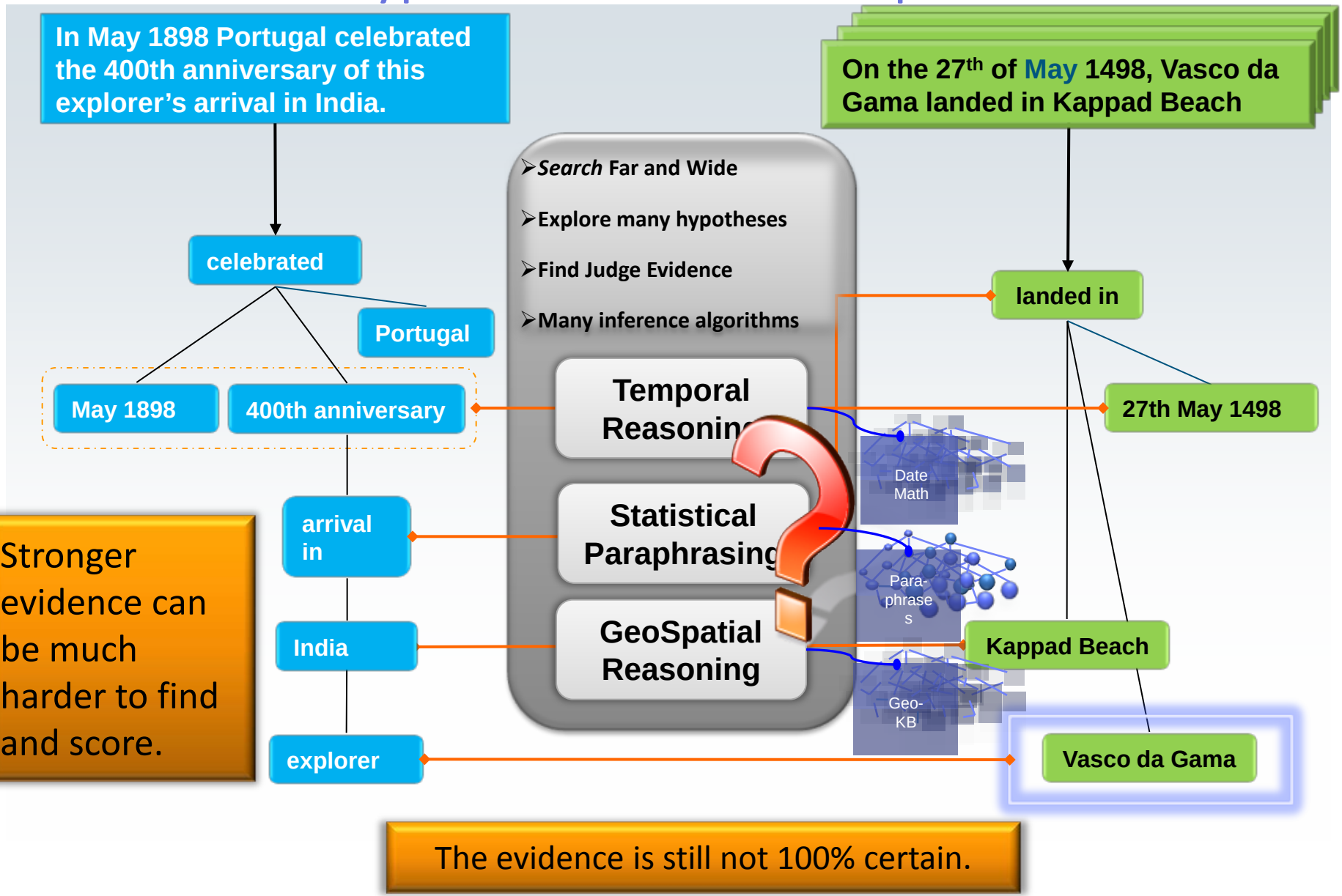
Massively Parallel Probabilistic Evidence-Based Architecture



Different Types of Evidence: Keyword Evidence

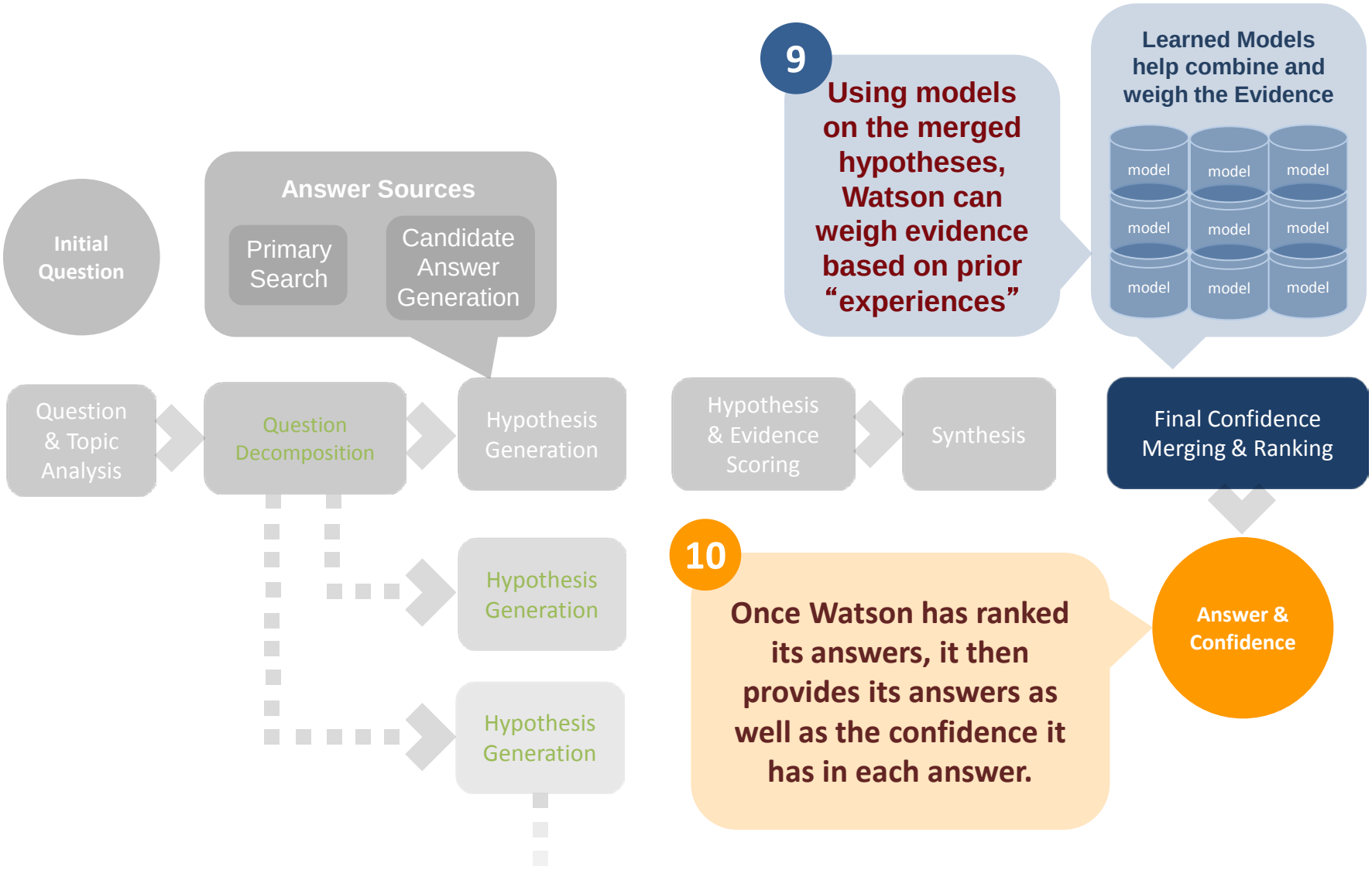


Different Types of Evidence: Deeper Evidence



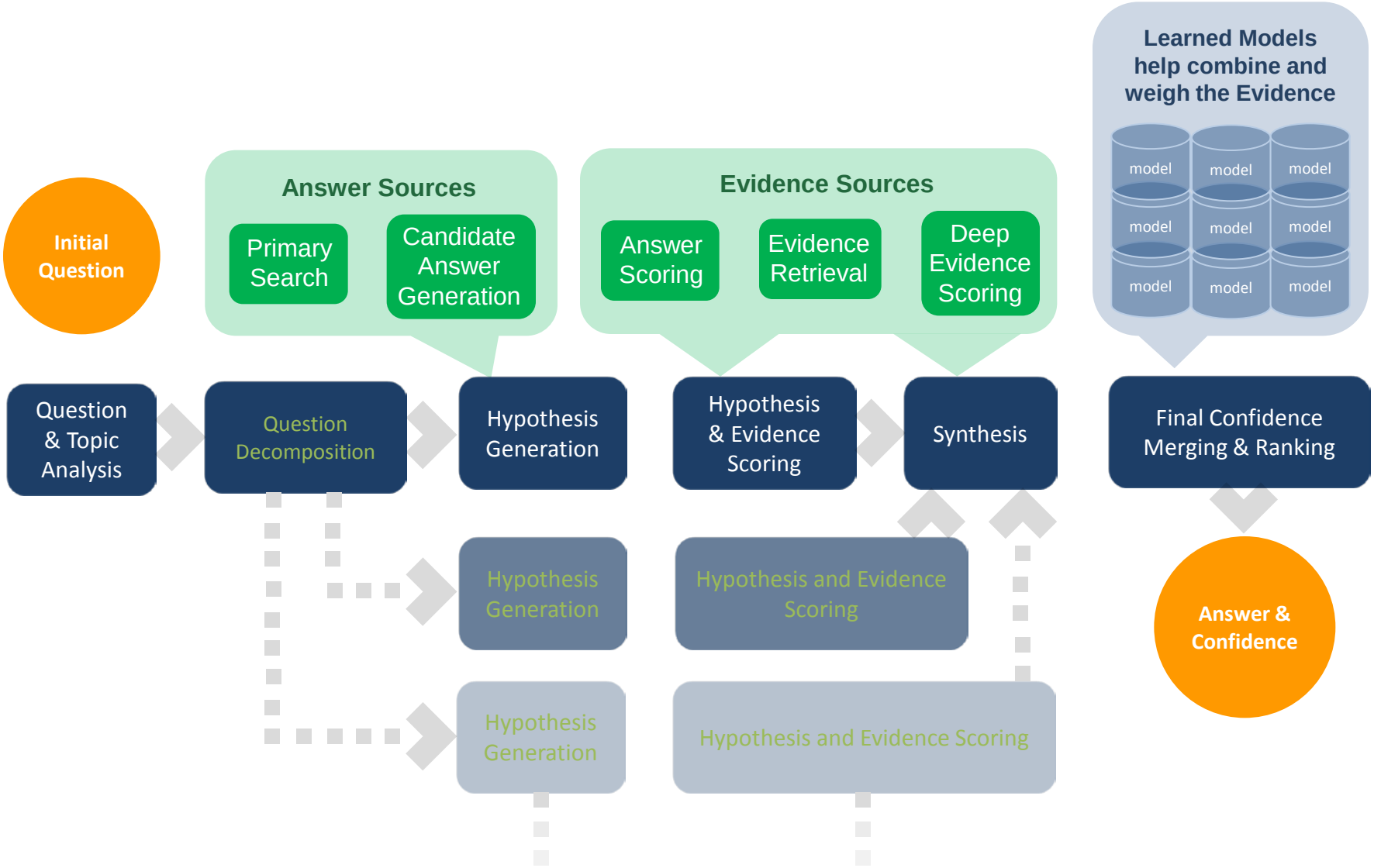
DeepQA: the technology & architecture behind Watson:

Massively Parallel Probabilistic Evidence-Based Architecture



DeepQA: the technology & architecture behind Watson:

Massively Parallel Probabilistic Evidence-Based Architecture



How Watson Processes a Question

IN 1698, THIS COMET DISCOVERER TOOK A SHIP CALLED THE PARAMOUR PINK ON THE FIRST PURELY SCIENTIFIC SEA VOYAGE

How Watson Processes a Question

IN 1698, THIS COMET DISCOVERER TOOK A SHIP CALLED THE PARAMOUR PINK ON THE FIRST PURELY SCIENTIFIC SEA VOYAGE

Question Analysis

Keywords: 1698, comet, paramour, pink, …
AnswerType(comet discoverer)
Date(1698)
Took(discoverer, ship)
Called(ship, Paramour Pink)
…

How Watson Processes a Question

IN 1698, THIS COMET DISCOVERER TOOK A SHIP CALLED THE PARAMOUR PINK ON THE FIRST PURELY SCIENTIFIC SEA VOYAGE

Question Analysis

Keywords: 1698, comet, paramour, pink, …
AnswerType(comet discoverer)
Date(1698)
Took(discoverer, ship)
Called(ship, Paramour Pink)
…

Primary Search

Primary Search Results

How Watson Processes a Question

IN 1698, THIS COMET DISCOVERER TOOK A SHIP CALLED THE PARAMOUR PINK ON THE FIRST PURELY SCIENTIFIC SEA VOYAGE

Question Analysis

Keywords: 1698, comet, paramour, pink, …
AnswerType(comet discoverer)
Date(1698)
Took(discoverer, ship)
Called(ship, Paramour Pink)
…

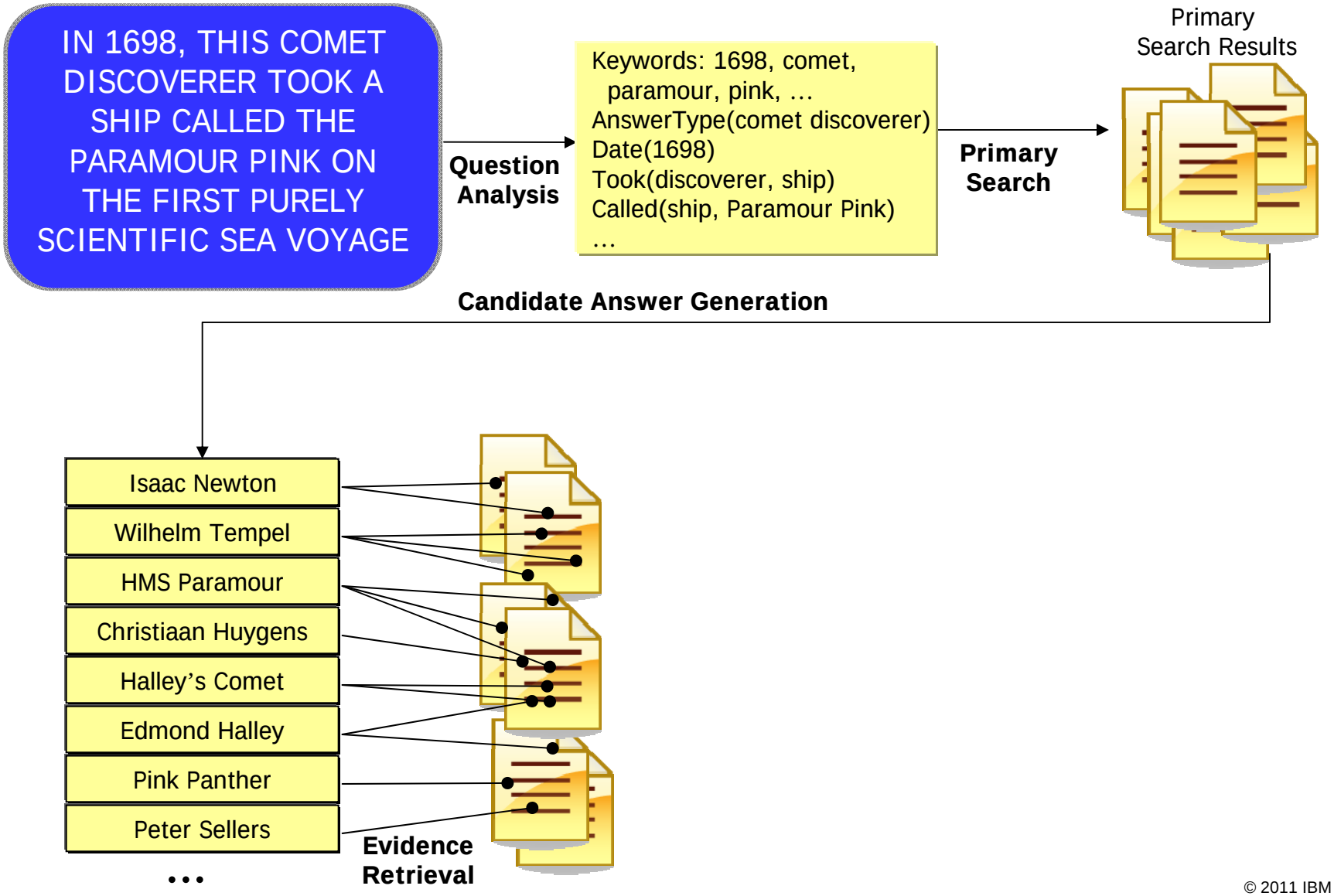
Primary Search

Primary Search Results

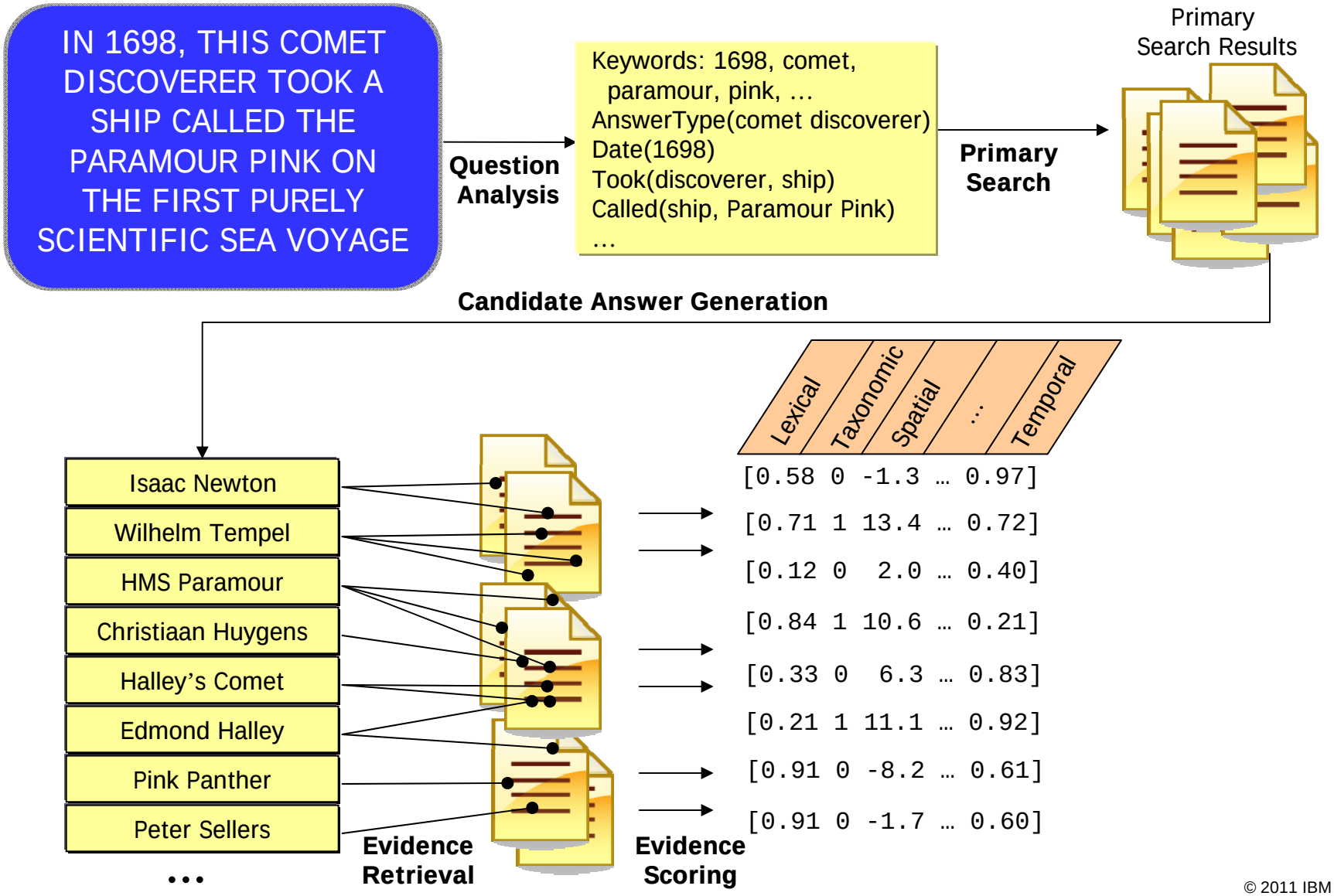
Candidate Answer Generation

- Isaac Newton
- Wilhelm Tempel
- HMS Paramour
- Christiaan Huygens
- Halley's Comet
- Edmond Halley
- Pink Panther
- Peter Sellers
- • • •

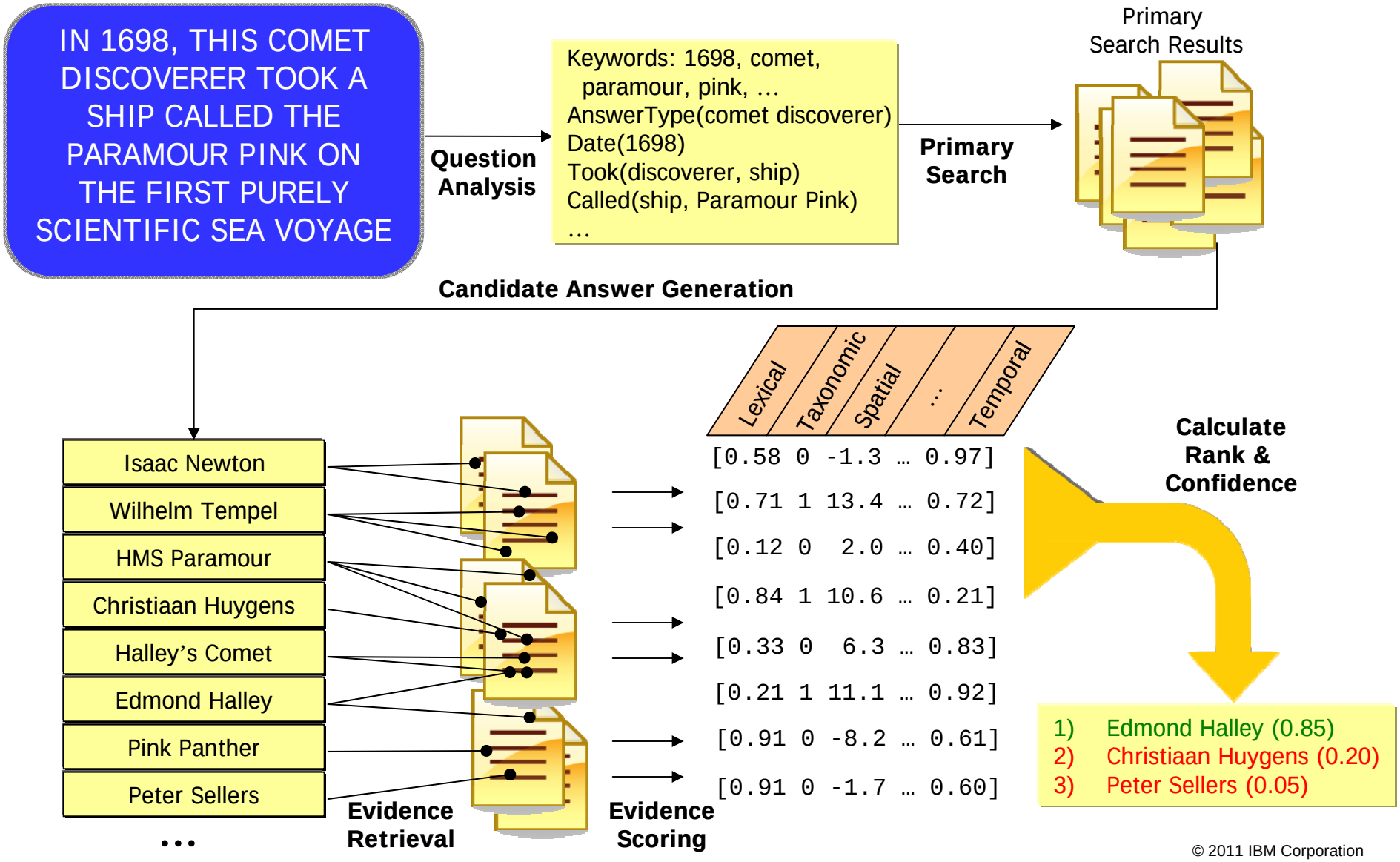
How Watson Processes a Question



How Watson Processes a Question

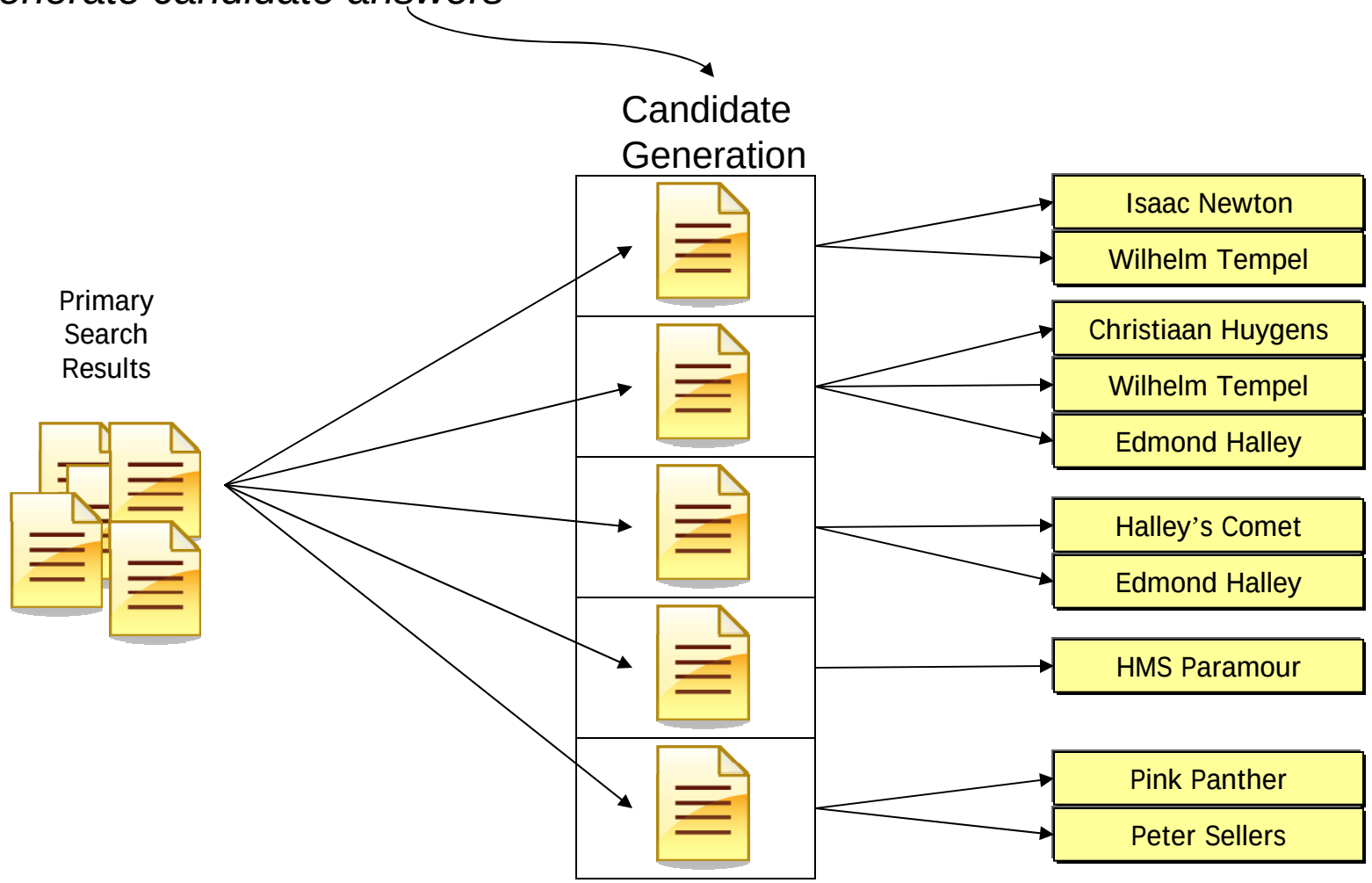


How Watson Processes a Question

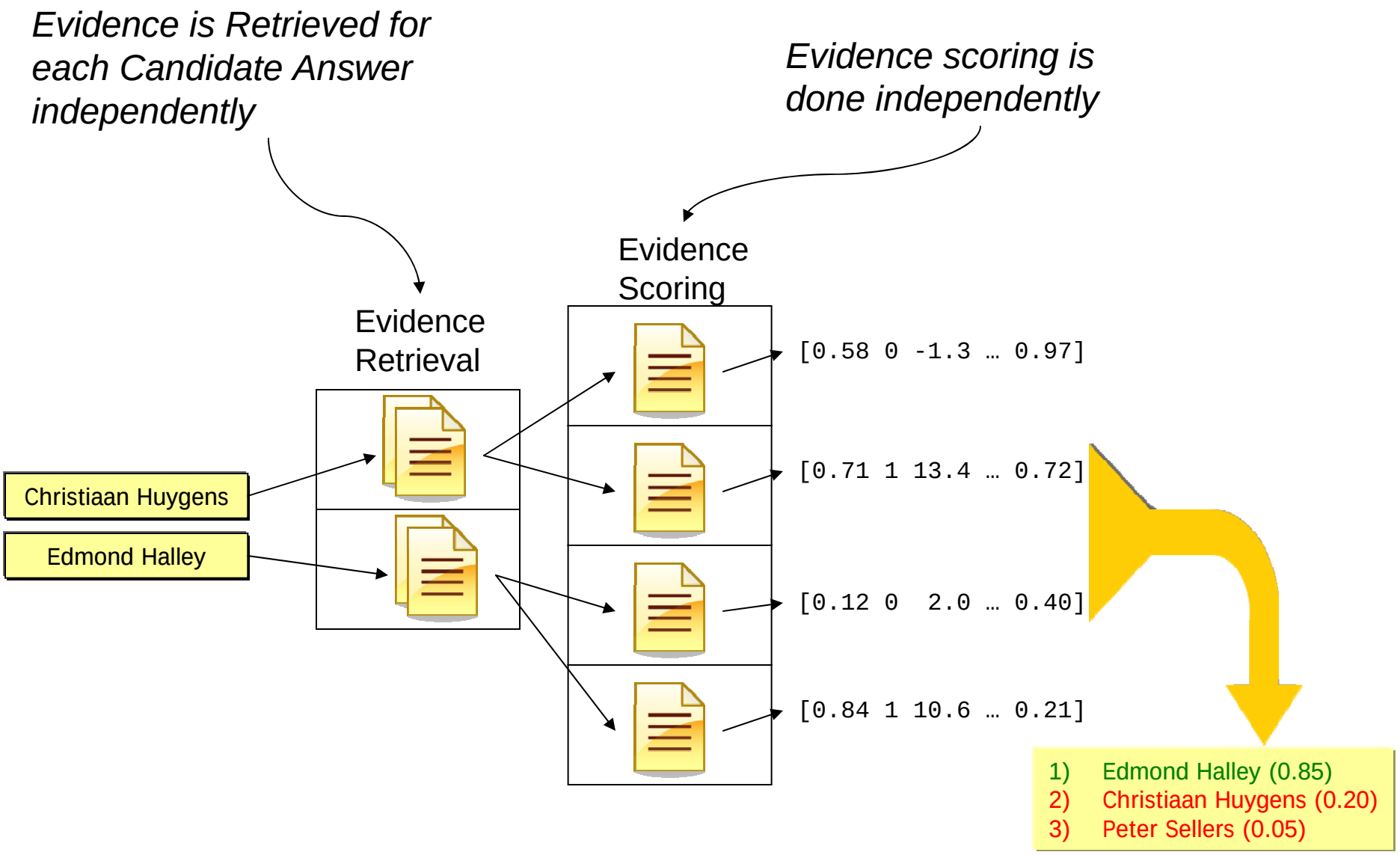


Opportunities for Parallel Processing

Primary Searches are independent, and each Search Result is analyzed independently to generate candidate answers

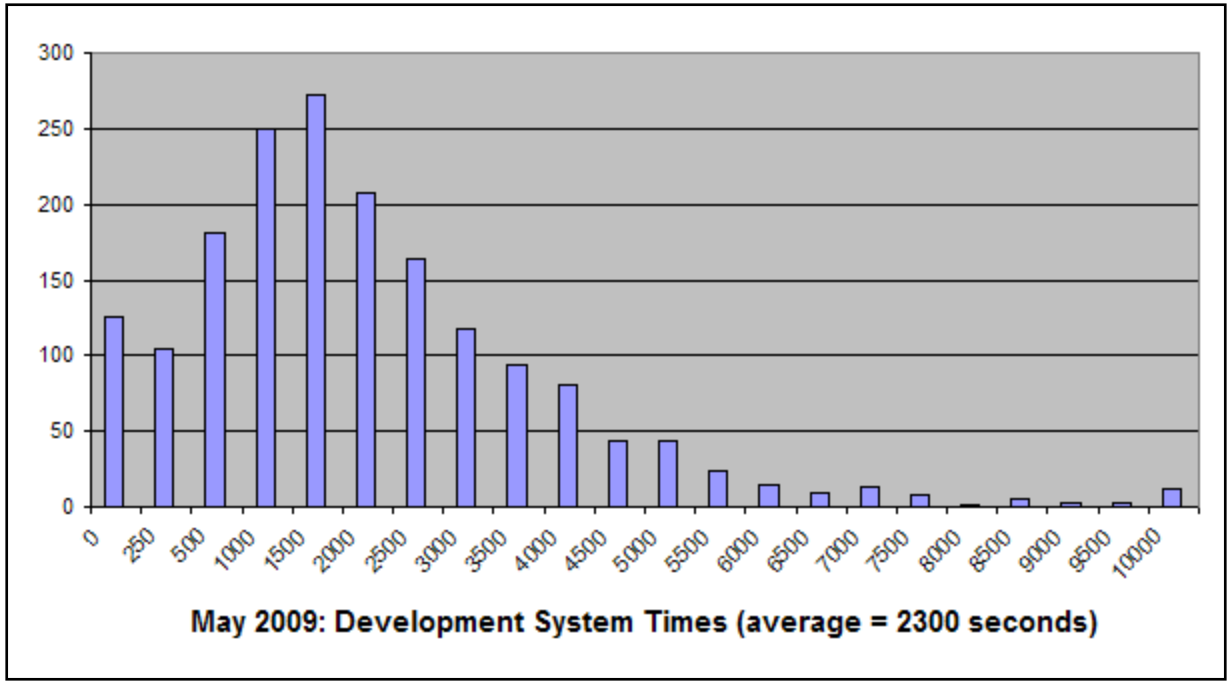


Opportunities for Parallel Processing



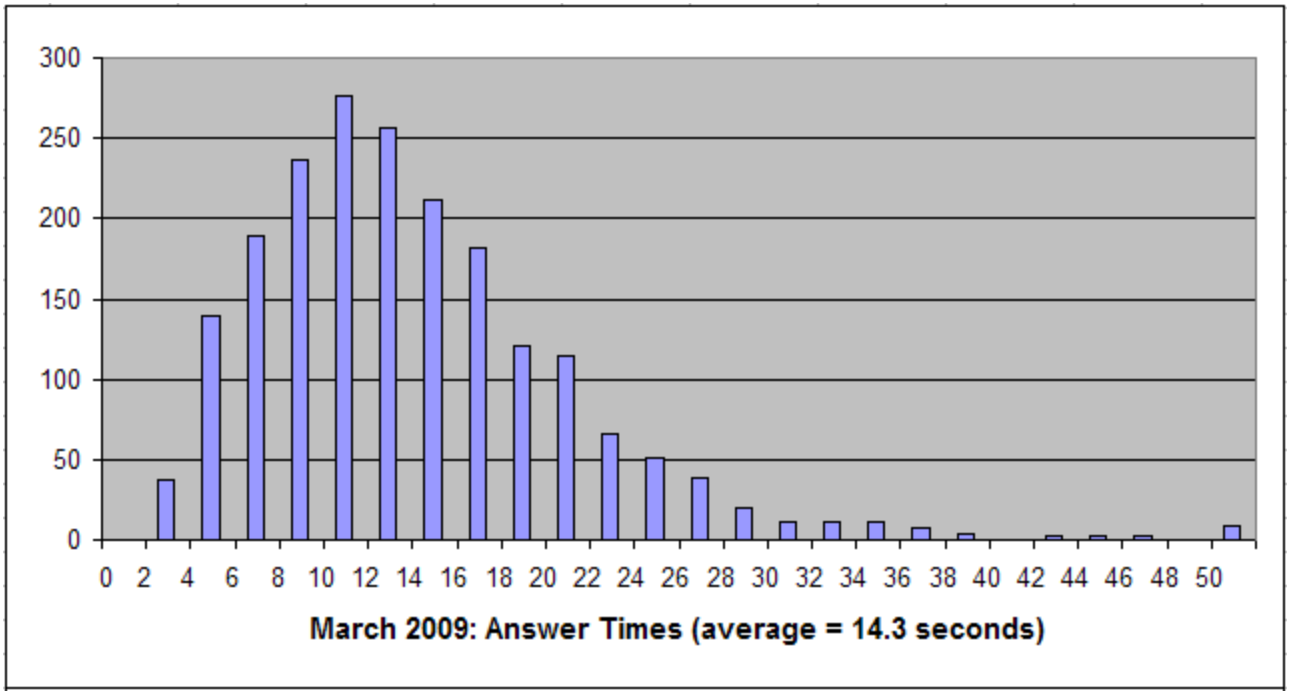
Development System Timing – Before Scale Out

- Single-threaded computation
- Search indexes on disk
- Remote Sesame server



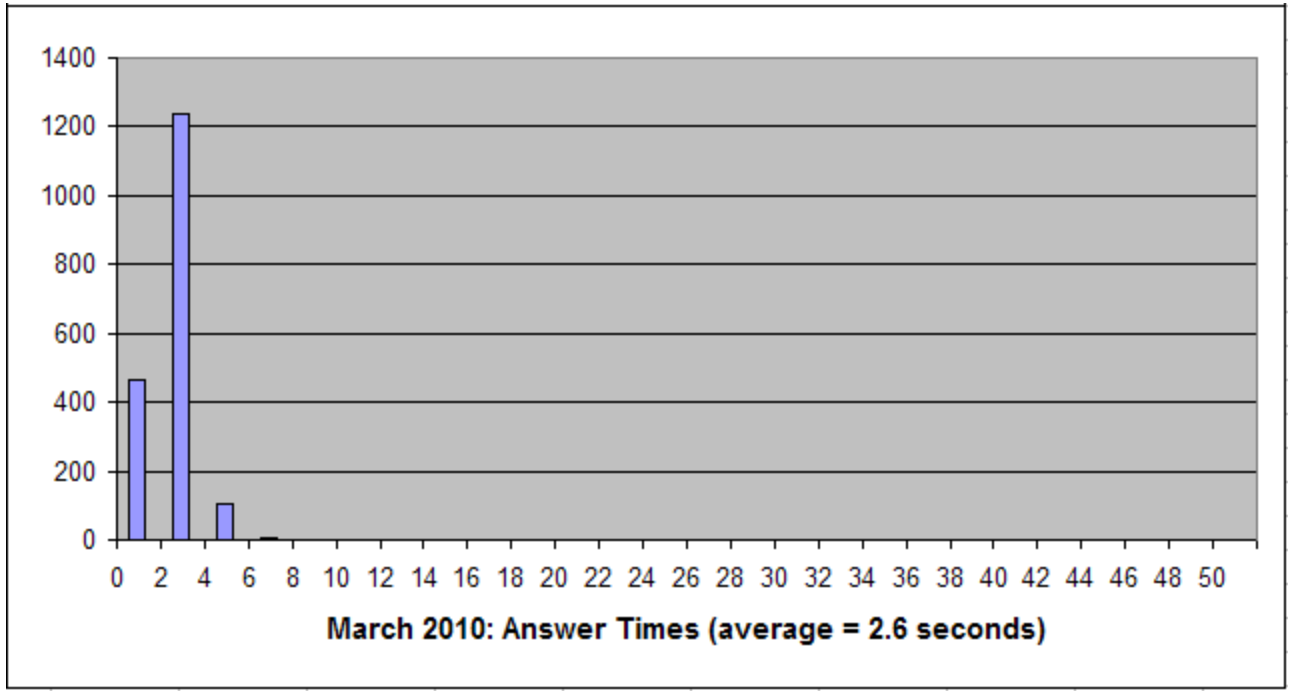
After first 8 months of Scaleout Work …

- Move everything into RAM
- Scale out components with UIMA-AS
- Distribute search

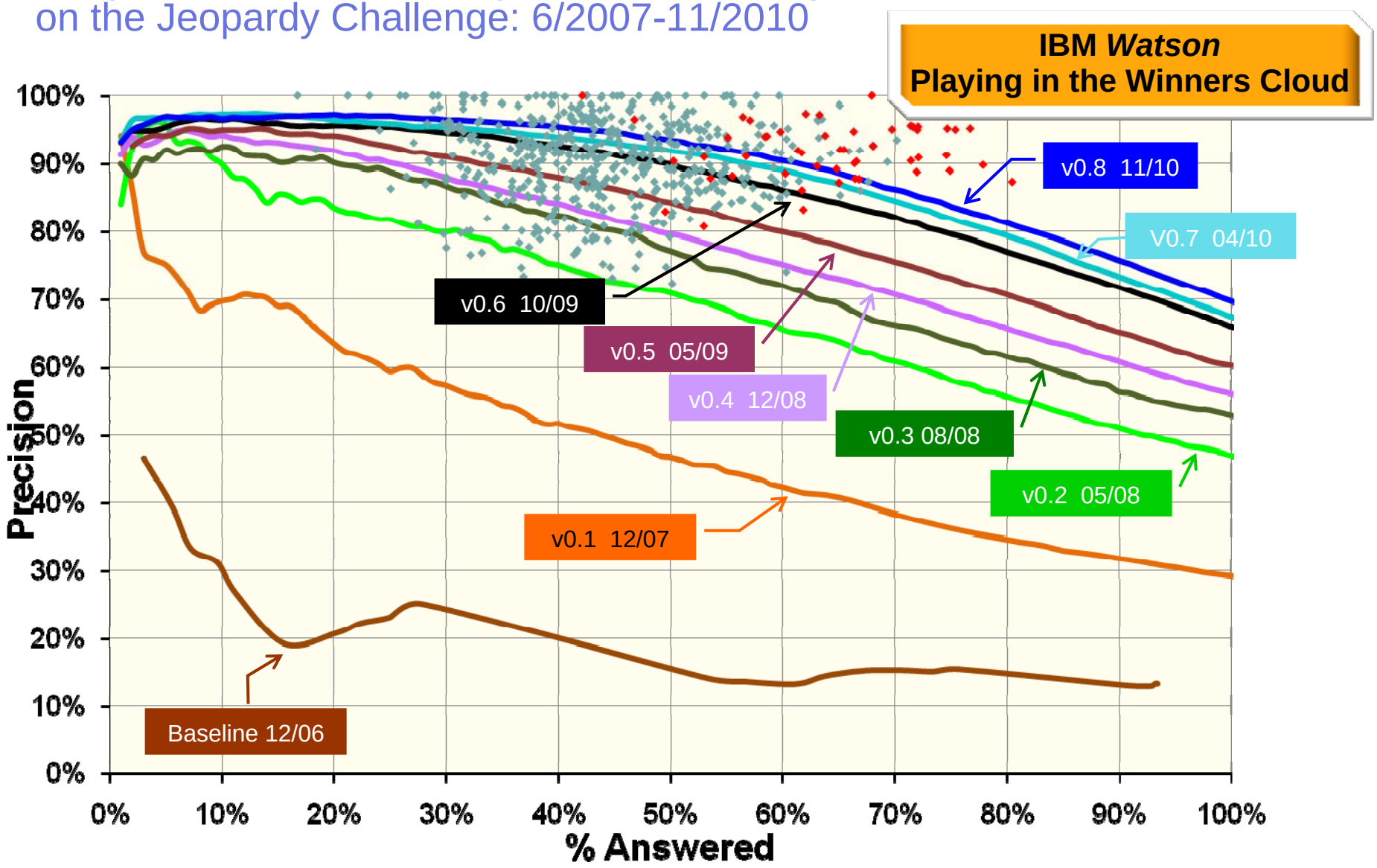


12 more months of Scaleout Work …

- Pre-compute deep NLP analysis of entire text corpus
- Hammer on every computation outlier
- Expand cluster



DeepQA: Incremental Progress in Answering Precision on the Jeopardy Challenge: 6/2007-11/2010



The Core Technical Team*

Researchers and Engineers in **NLP, ML, IR, KR&R and CL** at IBM Labs and a growing number of universities

PI: David Ferrucci

Systems & Speed
Eric Brown
Jerry Cwiklik
Pablo Duboue
Eddie Epstein
Tong Fin
Dan Gruhl
Bhavani Iyer
Adam Lally
Burn Lewis
Marshall Schor

Core Algorithms		
Eric Brown	Radu Florian	Dafna Sheinwald
Sugato Bagchi	David Gondek	Siddarth Patwardhan
Bran Boguraev	Aditya Kalyanpur	Kohichi Takeda
David Carmel	Hiroshi Kanayama	Yue Pan
Art Ciccolo	Adam Lally	John Prager
Jennifer Chu-Carroll	Tony Levas	Chris Welty
Bonaventura Coppola	Michael McCord	Wlodek Zadrozny
James Fan	Bill Murdock	Lei Zhang
David Ferrucci	Yuan Ni	
Achille Fokoue	Zhao Ming Qiu	

Strategy
David Gondek
Jon Lenchner
Gerry Tesauro
James Fan
John Prager

Speech
Andy Aaron
Raul Fernandez
Miroslav Novak
Andrew Rosenberg
Roberto Sicconi

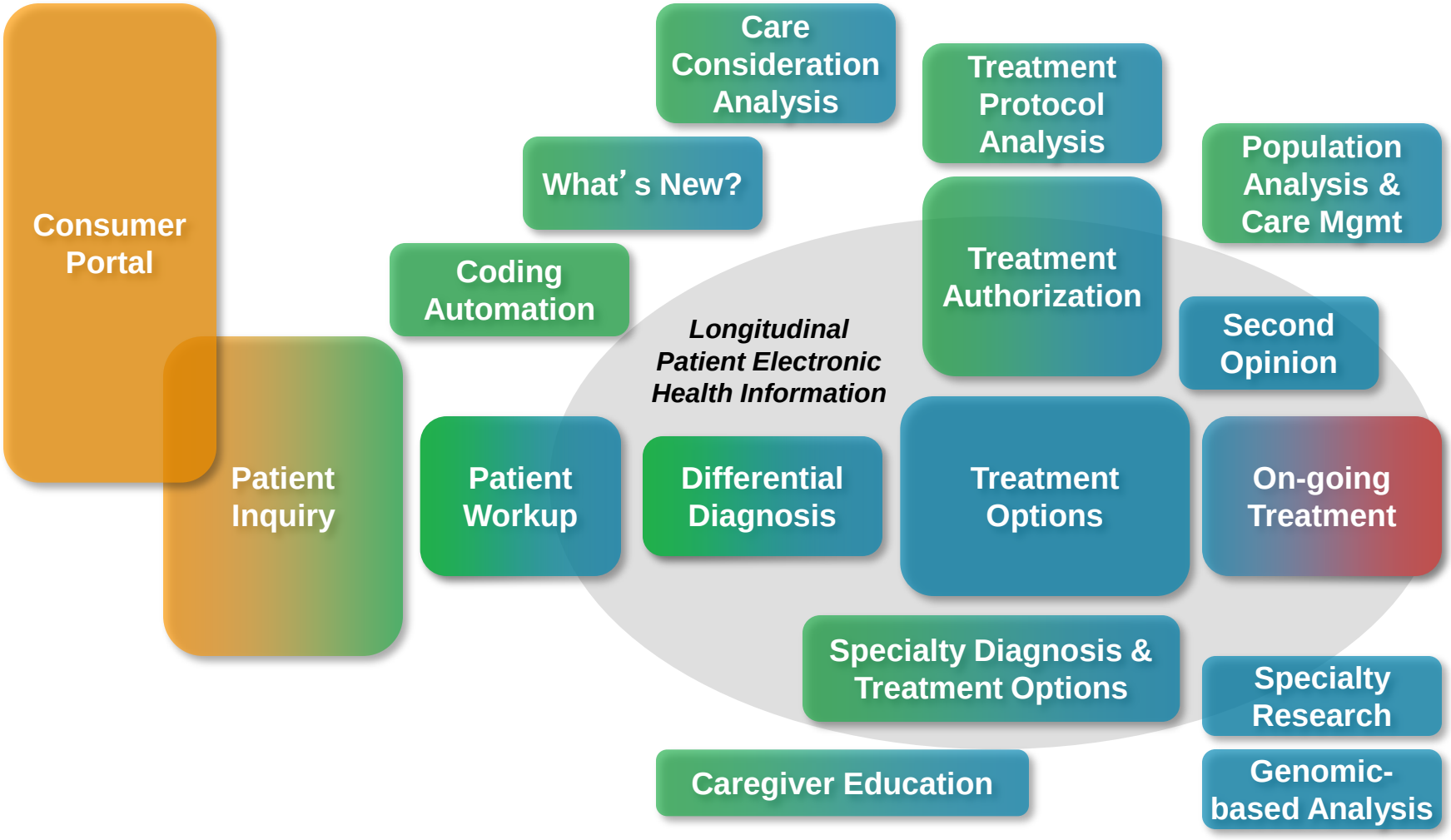
Data Annotation
Karen Ingraffea
Matt Mulholland

University Collaborations & Students		
Eric Nyberg (CMU)	James Allen (UMASS)	Andy Schlaikjer (CMU)
Nico Schlaefer (CMU)	Ed Hovy (USC)	Saurav Sahay (GT)
Manas Pathak (CMU)	Bruce Porter (UT)	Rutu Mulkar-Mehta (USC)
Chang Wang (UMASS)	Pallika Kanani (UMASS)	Doo Soon Kim (UT)
Hideki Shima (CMU)	Boris Katz (MIT)	
Barbara Cutler (RPI)	Alessandro Moschitti (Trento)	

There is a broader team that contributed to delivering Watson for the "Stage", to compete in Jeopardy Games

*NOT full-time Equivalents. Names listed if contributed some time to that part of project.

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