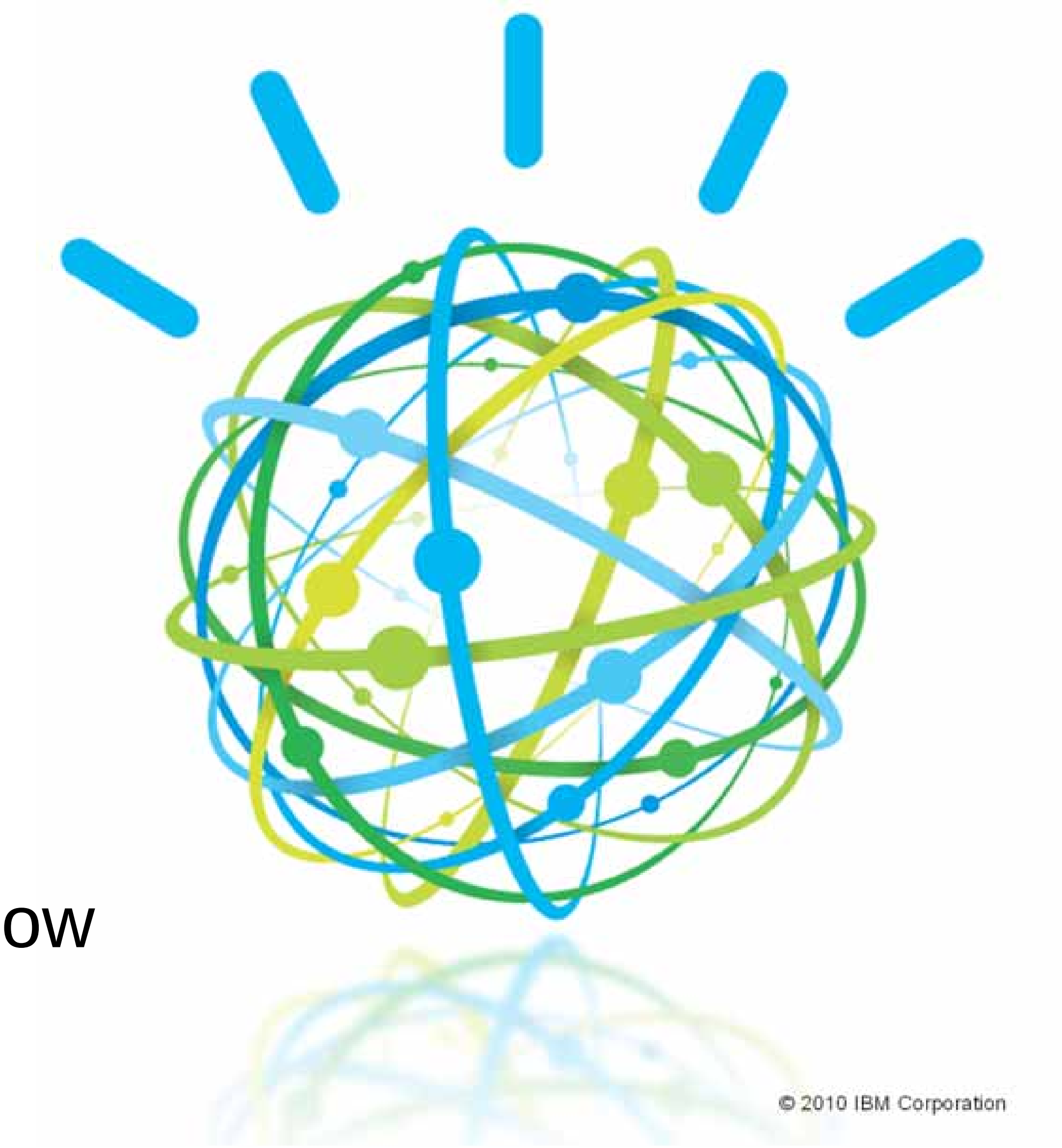
Building *Watson*



David Ferrucci, IBM Fellow

Principal Investigator

DeepQA@ IBM Research

A Grand Challenge Opportunity

- **Drive Important Scientific Advances**
  - – Envision new ways for computers to impact society & science
- **Be Relevant to IBM Customers**
  - – Enable better, faster decision making over unstructured and structured content
  - – *Business Intelligence, Knowledge Discovery and Management, Government, Compliance, Publishing, Legal, Healthcare, Product Support, etc.*
- **Capture the Broader Imagination**
  - – The Next *Deep Blue*

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

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.

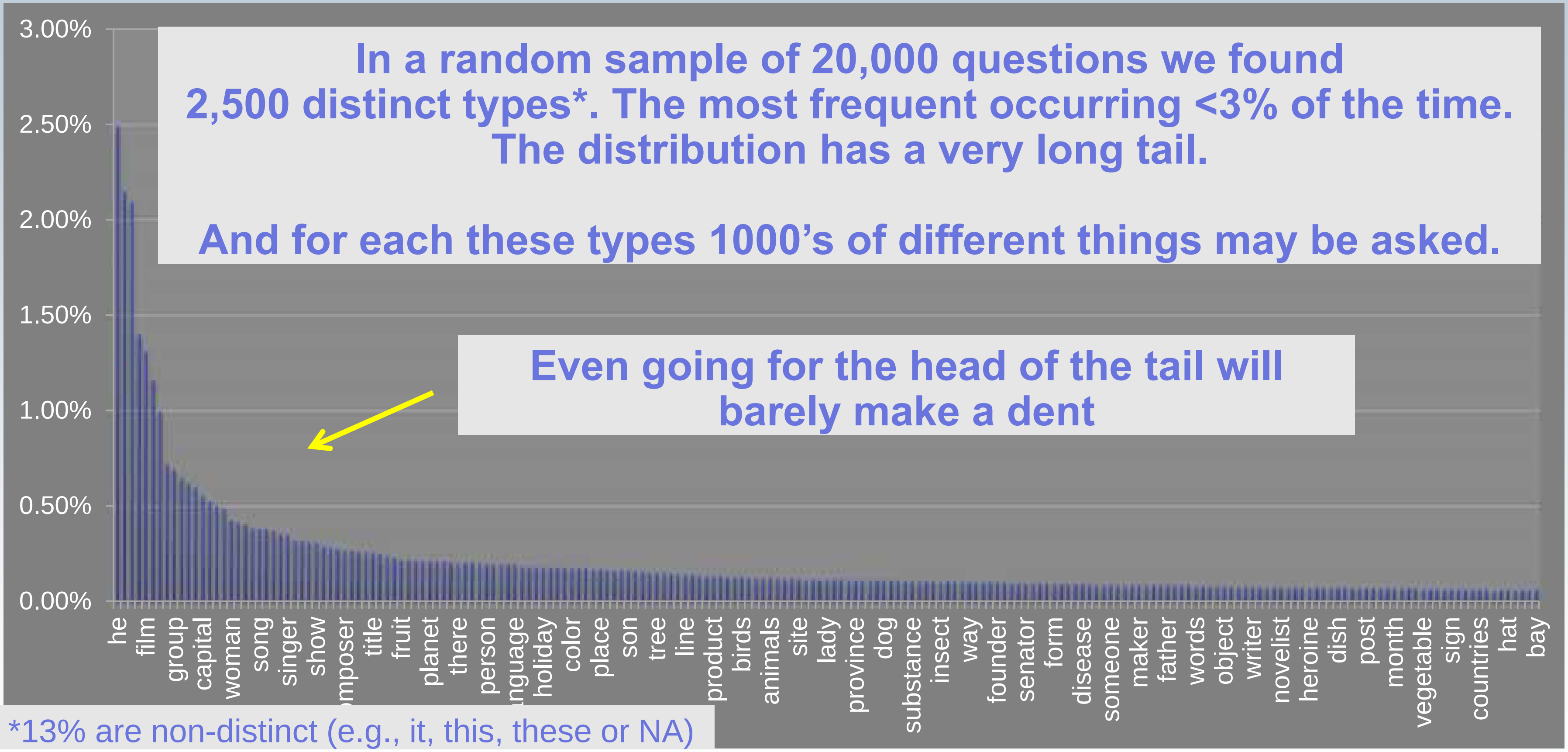
$1000
The first person mentioned by name in 'The Man in the Iron Mask' is this hero of a previous book by the same author.

$600
In cell division, mitosis splits the nucleus & cytokinesis splits this liquid *cushioning* the nucleus

$2000
Of the 4 countries in the world that the U.S. does not have diplomatic relations with, the one that's farthest north

Broad Domain

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



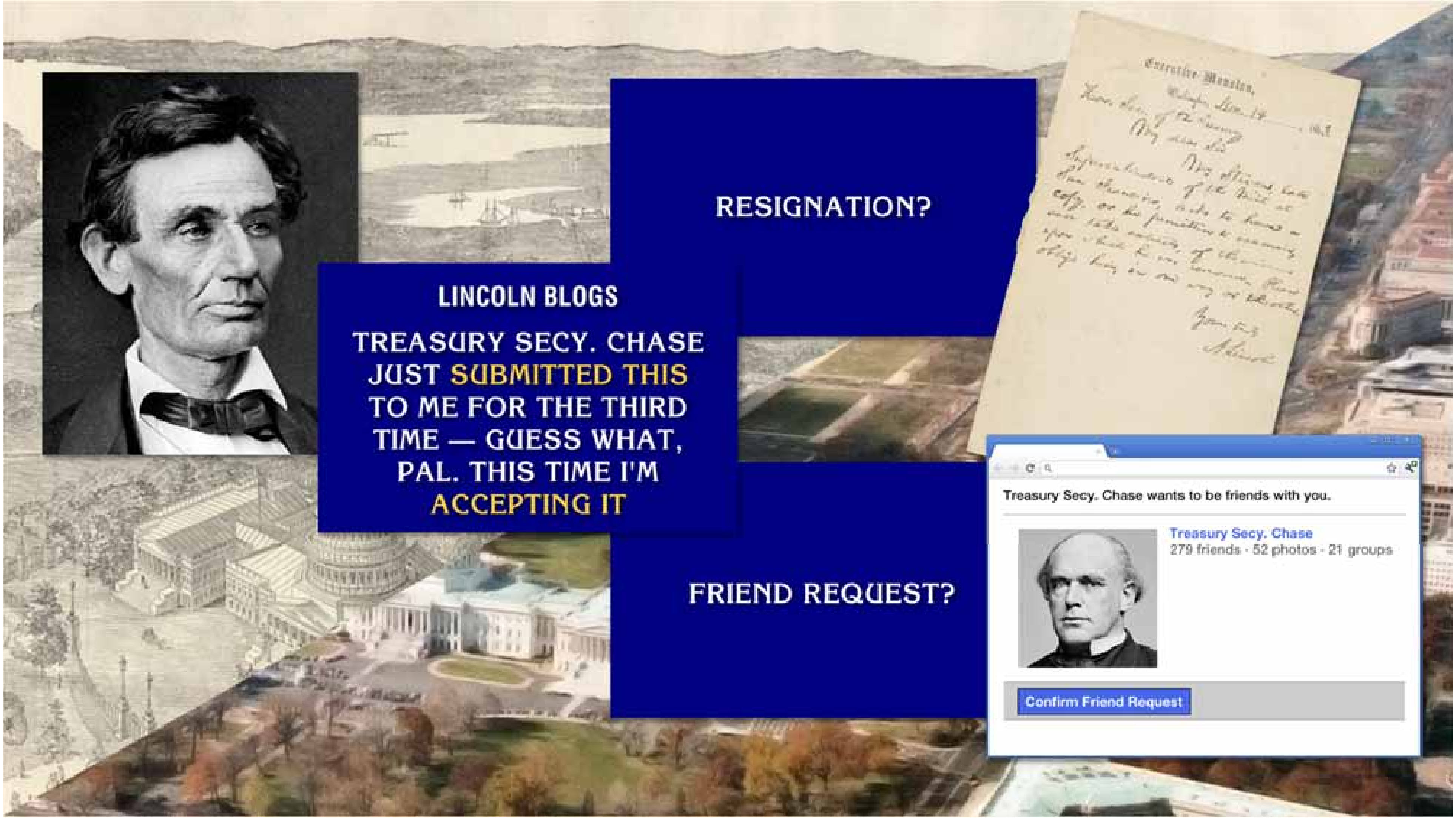
In a random sample of 20,000 questions we found 2,500 distinct types*. The most frequent occurring <3% of the time. The distribution has a very long tail.

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

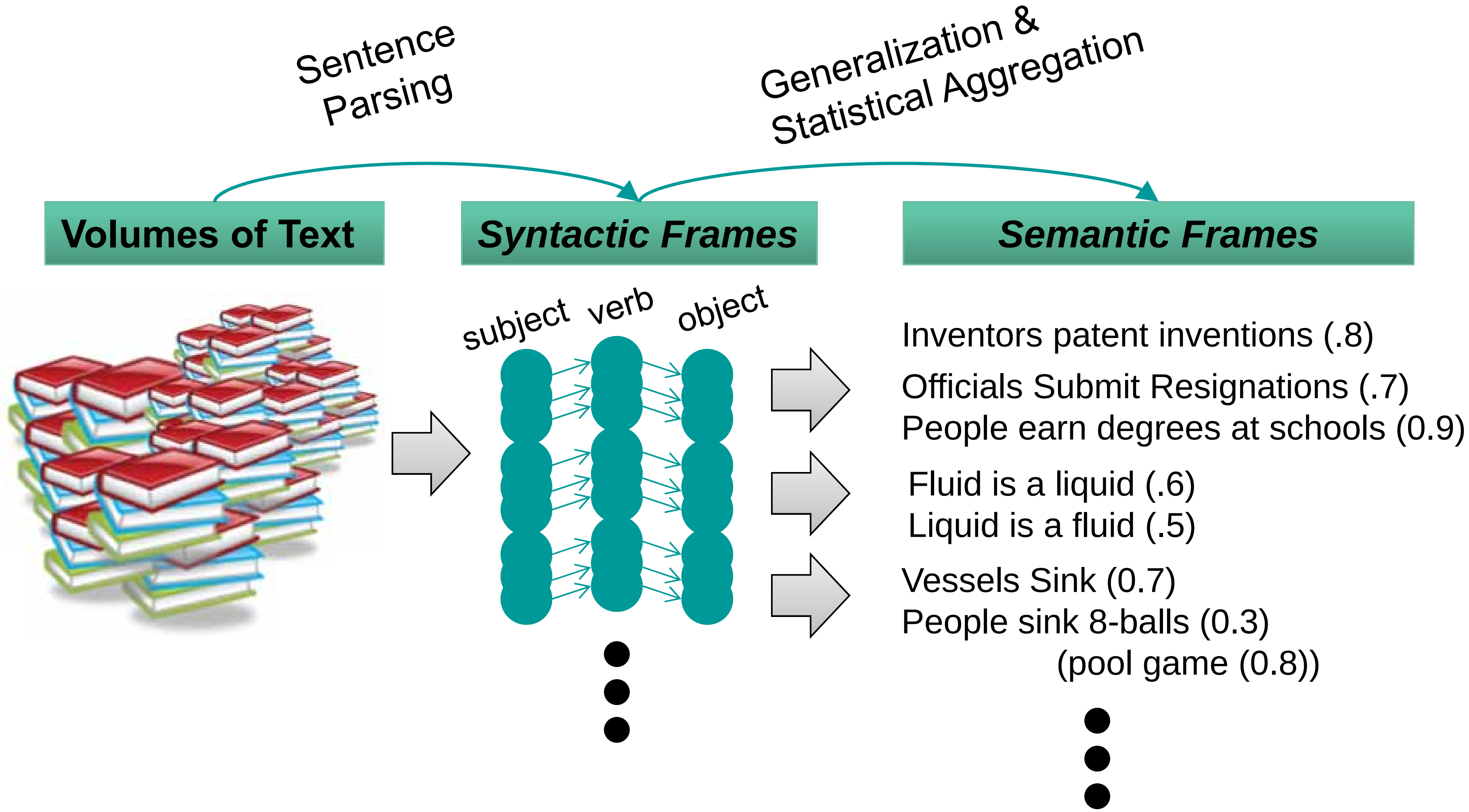
Even going for the head of the tail will barely make a dent

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

Our Focus is on reusable NLP technology for analyzing volumes of *as-is* text. Structured sources (DBs and KBs) are used to help interpret the text.


RESIGNATION?
LINCOLN BLOGS
TREASURY SECY. CHASE JUST SUBMITTED THIS TO ME FOR THE THIRD TIME — GUESS WHAT, PAL. THIS TIME I'M ACCEPTING IT
FRIEND REQUEST?
Executive Mansion,
Treasury Secy. Chase wants to be friends with you.
Treasury Secy. Chase
279 friends · 52 photos · 21 groups
Confirm Friend Request

Inducing Meaning



Sentence Parsing

Generalization & Statistical Aggregation

Volumes of Text

Syntactic Frames

Semantic Frames

subject verb object

Inventors patent inventions (.8)

Officials Submit Resignations (.7)
People earn degrees at schools (0.9)

Fluid is a liquid (.6)
Liquid is a fluid (.5)

Vessels Sink (0.7)
People sink 8-balls (0.3)
(pool game (0.8))

Generating Possibilities, Gathering and Scoring Evidence

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

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

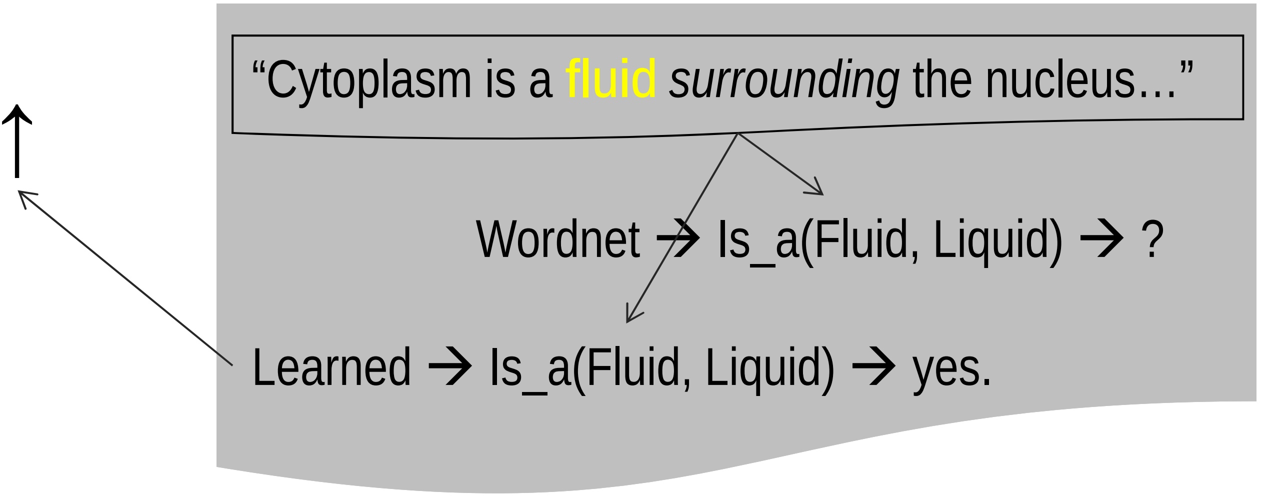
- Many candidate answers (CAs) are generated from many different searches
- Each possibility is evaluated according **to different dimensions of evidence.**
- **Just One** piece of evidence is if the CA is of the right type. In this case a "liquid".

Is("Cytoplasm", "liquid") = 0.2↑

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

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

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

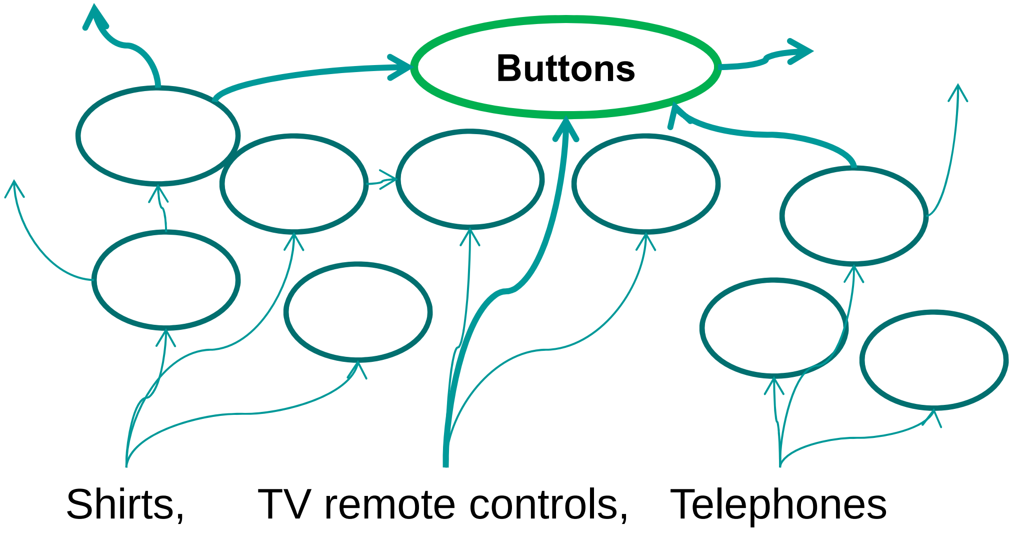


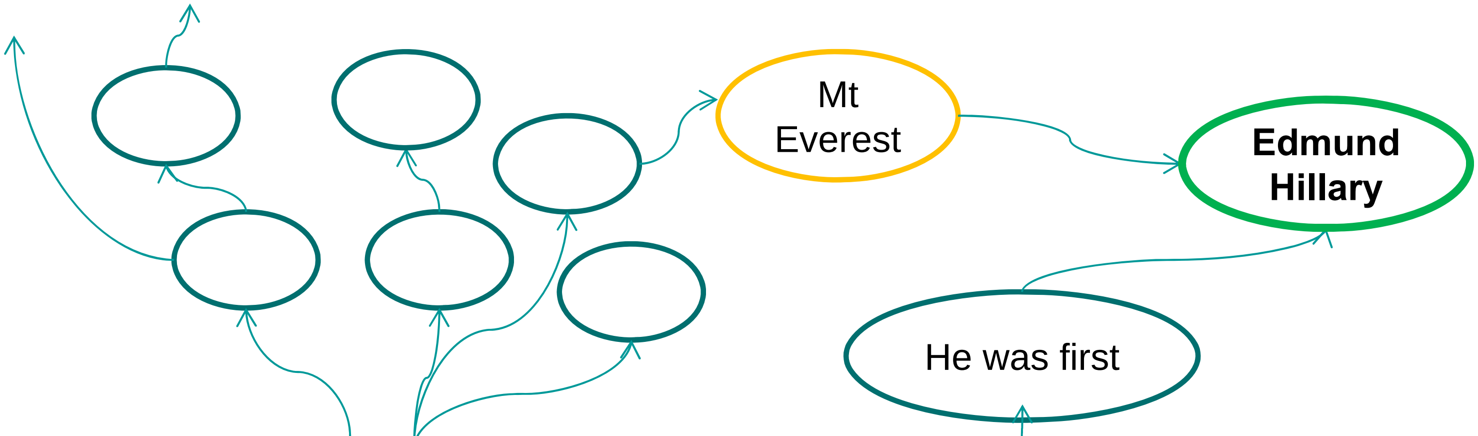
"Cytoplasm is a fluid *surrounding* the nucleus…"

Wordnet → Is_a(Fluid, Liquid) → ?

Learned → Is_a(Fluid, Liquid) → yes.

The Missing Link





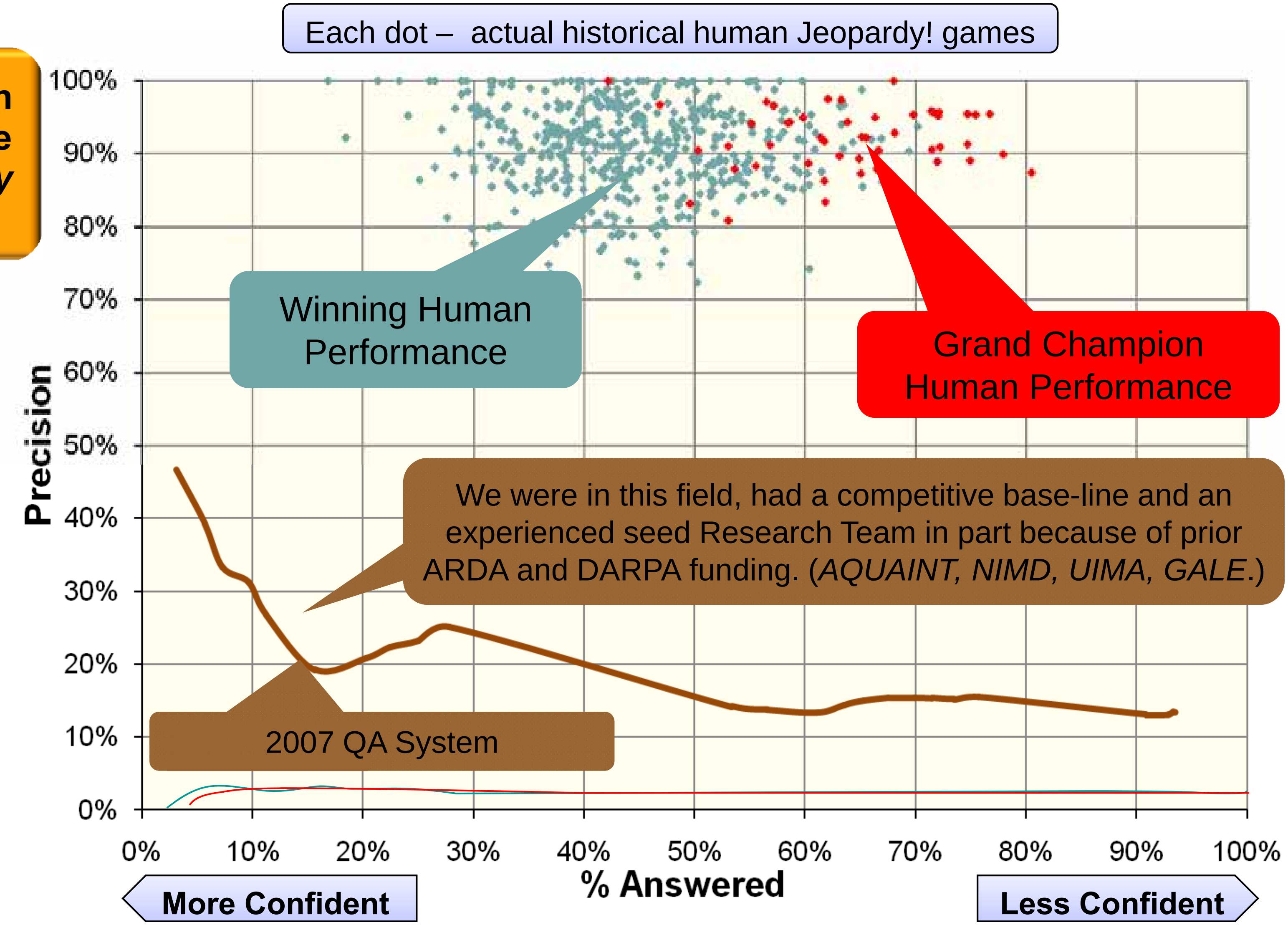
On hearing of the discovery of George Mallory's body, he told reporters he still thinks he was first.

The 1648 Peace of Westphalia ended a war that began on May 23 of this year.

What It Takes to compete against Top Human Jeopardy! Players

Our Analysis Reveals the ***Winner's Cloud***

Top human players are *remarkably* good.

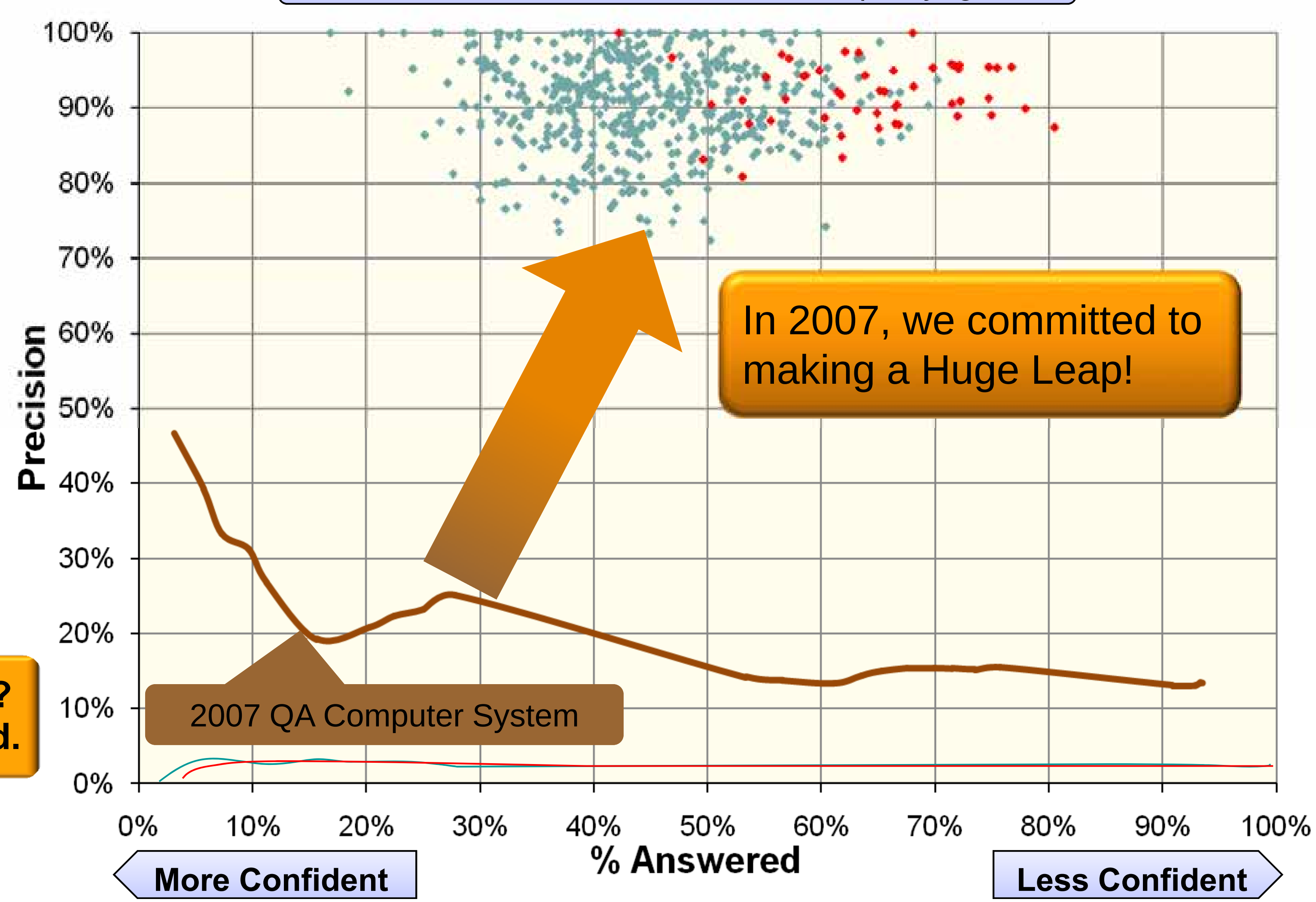


We were in this field, had a competitive base-line and an experienced seed Research Team in part because of prior ARDA and DARPA funding. (*AQUAINT, NIMD, UIMA, GALE*.)

What It Takes to compete against Top Human Jeopardy! Players

Our Analysis Reveals the ***Winner's Cloud***

Each dot – actual historical human Jeopardy! games



In 2007, we committed to making a Huge Leap!

2007 QA Computer System

Computers? Not So Good.

Enabling Technologies – The Time Was Right

Natural Knowledge

- Large volumes natural language electronic text (e.g., news, wikis, reference, web, etc.)
- Encodes knowledge and greater linguistic **contexts** to better resolve **intended meaning**

Semi-Structured Knowledge

- Large volumes of Thesauri, Dictionaries, Folksonomies, Lists, the Semantic Web, Linked Open Data
- Rapid, community-based construction
- Across many domains – Specialized and General

NLP (Text Analysis)

- Entity and Relation Detection
- Statistical NLP - Broader coverage, lower cost Information Extraction
- Statistical Paraphrasing -- Learn different ways to express same meaning

Compute Power

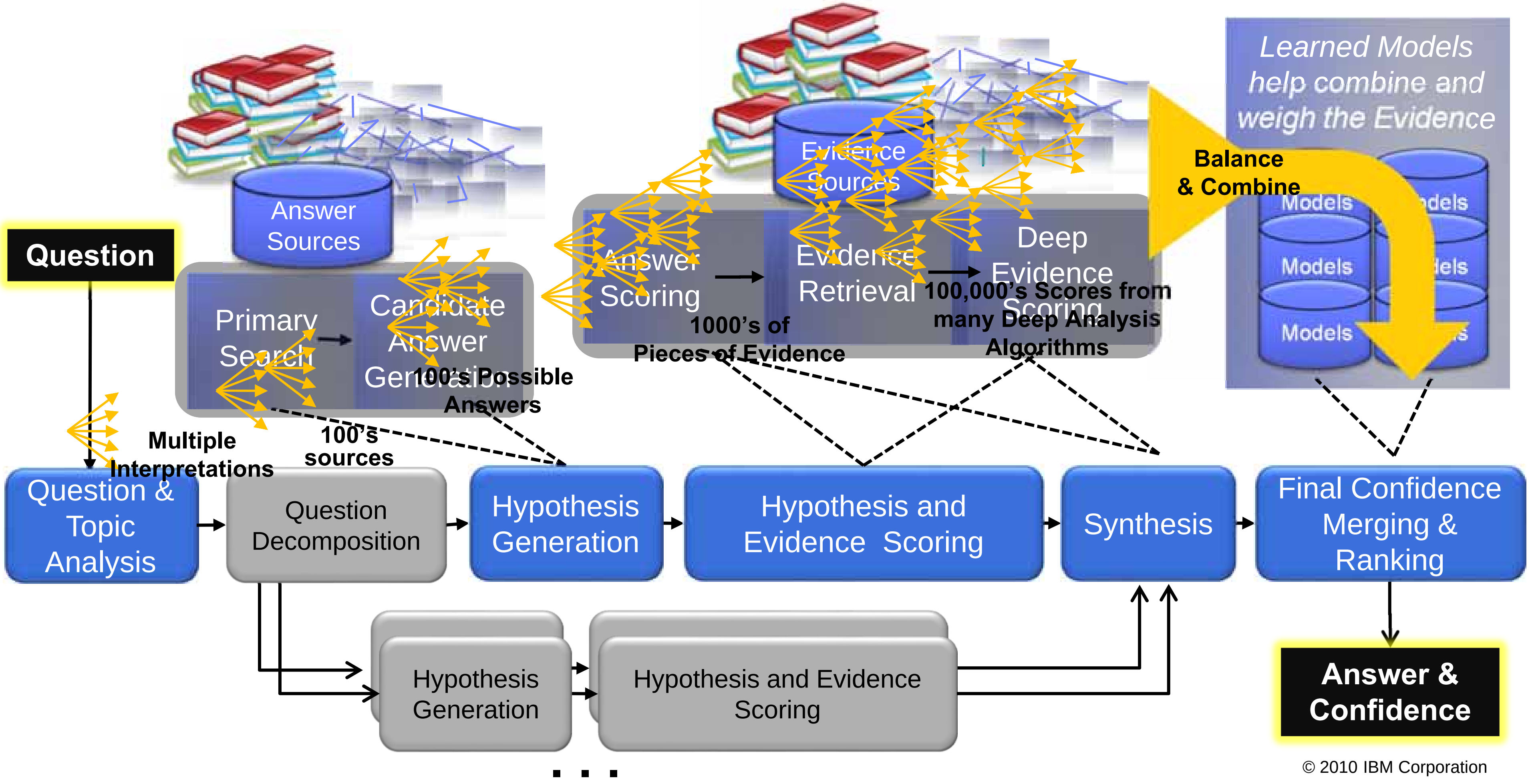
- Massive parallel compute power
- 1000s of compute cores working simultaneously
- TBs of globally addressable main memory

Key Assumptions

- Large Hand-Crafted Models won’t cut it
  - *Too Slow, Too Narrow, Too brittle, Too Bias*
  - *Need to acquire and analyze information from **As-Is Knowledge sources***
- Intelligence from the combination of many
  - *Consider **many hypotheses** . Reduce early biases.*
  - *Consider **many diverse algorithms** . No single one is perfect or complete.*
  - *Analyze evidence form different perspectives*
  - *Best combination is **continually learned** , tested and refined*
- *Massive Parallelism a Key Enabler*
  - *Pursue many competing independent hypotheses over large data*
  - *Efficiency will demand simultaneous threads of evidence evaluation*

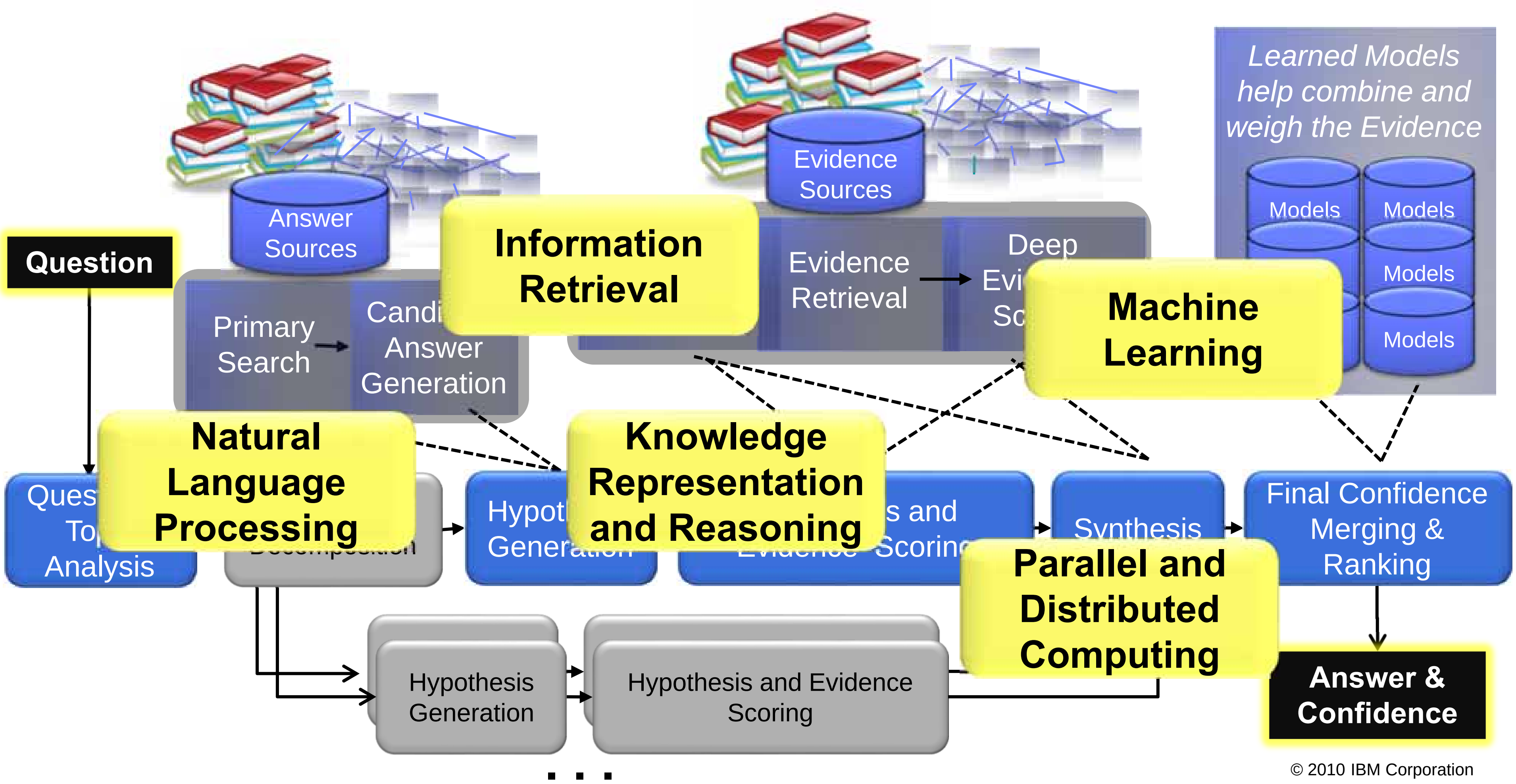
DeepQA: The architecture underlying Inside Watson

*Generates many hypotheses, **collects a wide range of evidence** and balances the combined confidences of **over 100 different analytics that analyze the evidence form different dimensions***



DeepQA: The architecture underlying Inside Watson

*Generates many hypotheses, **collects a wide range of evidence** and balances the combined confidences of **over 100 different analytics that analyze the evidence form different dimensions***



Rapid Innovation Methodology Emerged

- Goal-Oriented Metrics and Incremental Investments
  - – Identify a Target and Technical Approach
  - – Headroom Analysis: Estimate idea's potential impact on key metrics
  - – Balance long-term & short-term investments. Have the next priority ready. Be Agile.
- Extreme Collaboration
  - – Implemented "One Room" to optimize team work, communication and commitment
  - – Immediate access to the right "expert", spontaneous discussions, no good idea lost
- Disciplined Engineering and Evaluation (Regular Blind Data Experiments)
  - – Bi-weekly End-to-End Integration Runs & Evaluations (Large Compute Resources)
  - – >10 GBs of error analysis output made accessible via Web-Based Tool
  - – Positive impact on last run required to get into the next bi-weekly run

>8000 Documented experiments performed in 4 years

The *System* as a Market Place

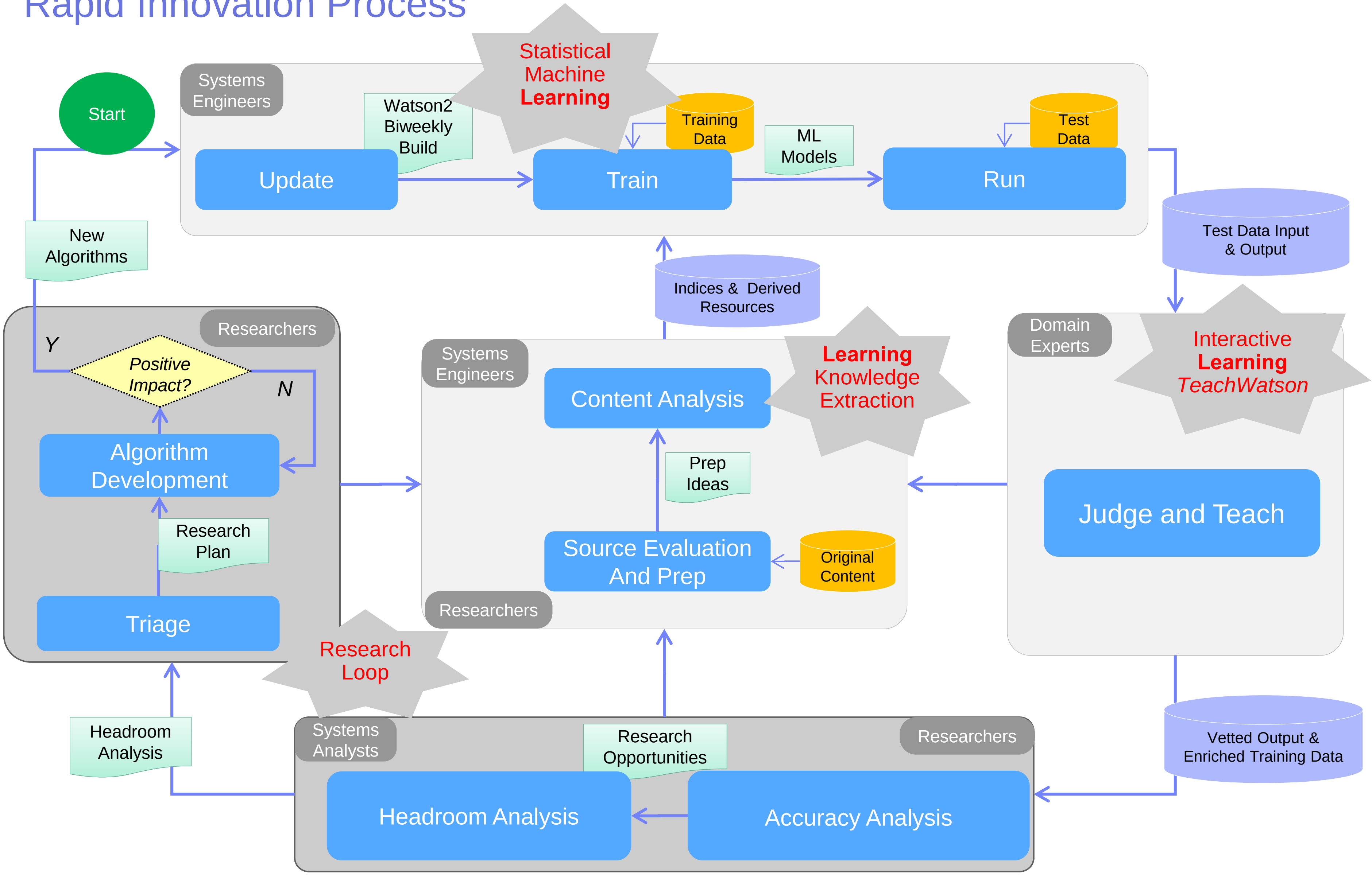
- Independent Component Results
  - Individual algorithm results can get you published
  - Examples: Word sense disambiguation, parsing, graph matching, co-reference, Text Entailments etc.
- But..
  - Integrated with many others your pet idea may be diminished
  - Integrated System Performance
  - The End-to-End system is enormously complex and its performance is empirical
  - Unpredictable interactions and hidden variables impact the net effect
- The System as an open “Market Place”
  - Your work must contribute in the context of all the other competing components
  - Algorithms need to stand up to simpler, cheaper in the context of the larger evolving system

Encouraged people to take a ***System-Wide view*.**

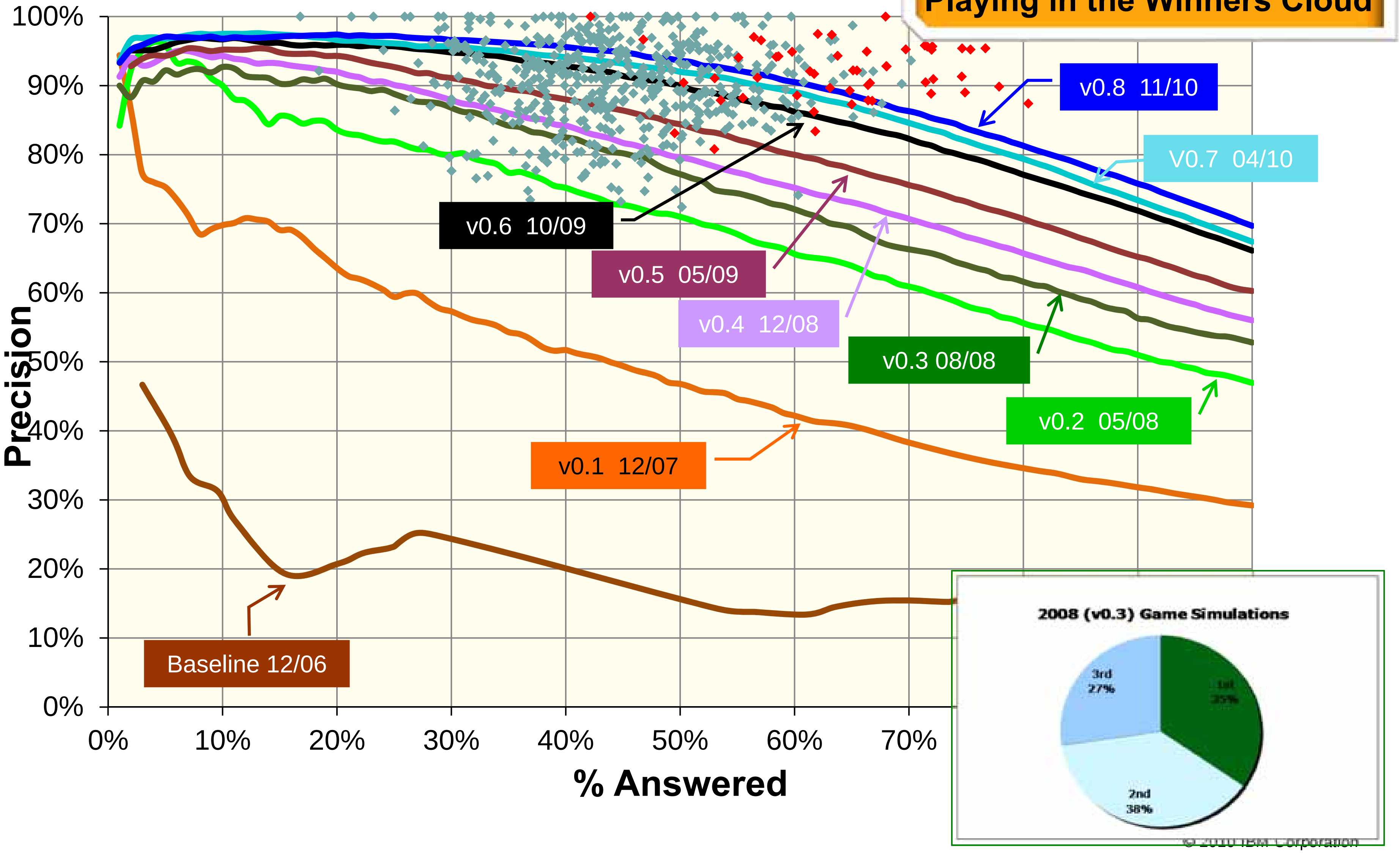
Tools to **isolate** and **measure** impact and cost.

An open architecture that supported rapid combination of components.

Rapid Innovation Process



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



IBM *Watson*
Playing in the Winners Cloud

Deployment Models

Development System

Easy to change/update High Experimental Throughput

2500 Questions on ~1500 Cores in a few hours

Algorithm & Data Migration

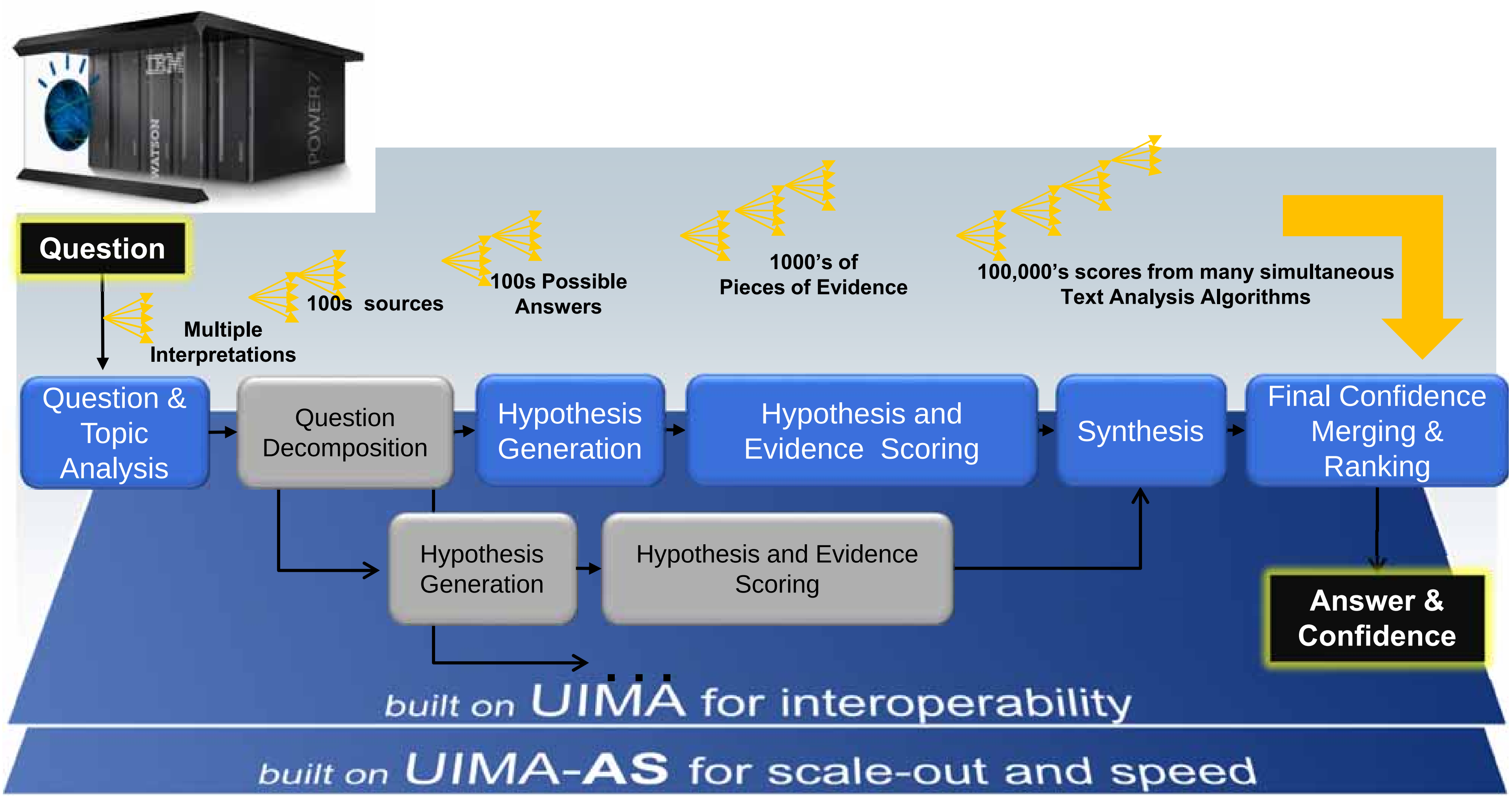
Software Bottlenecks

Production System

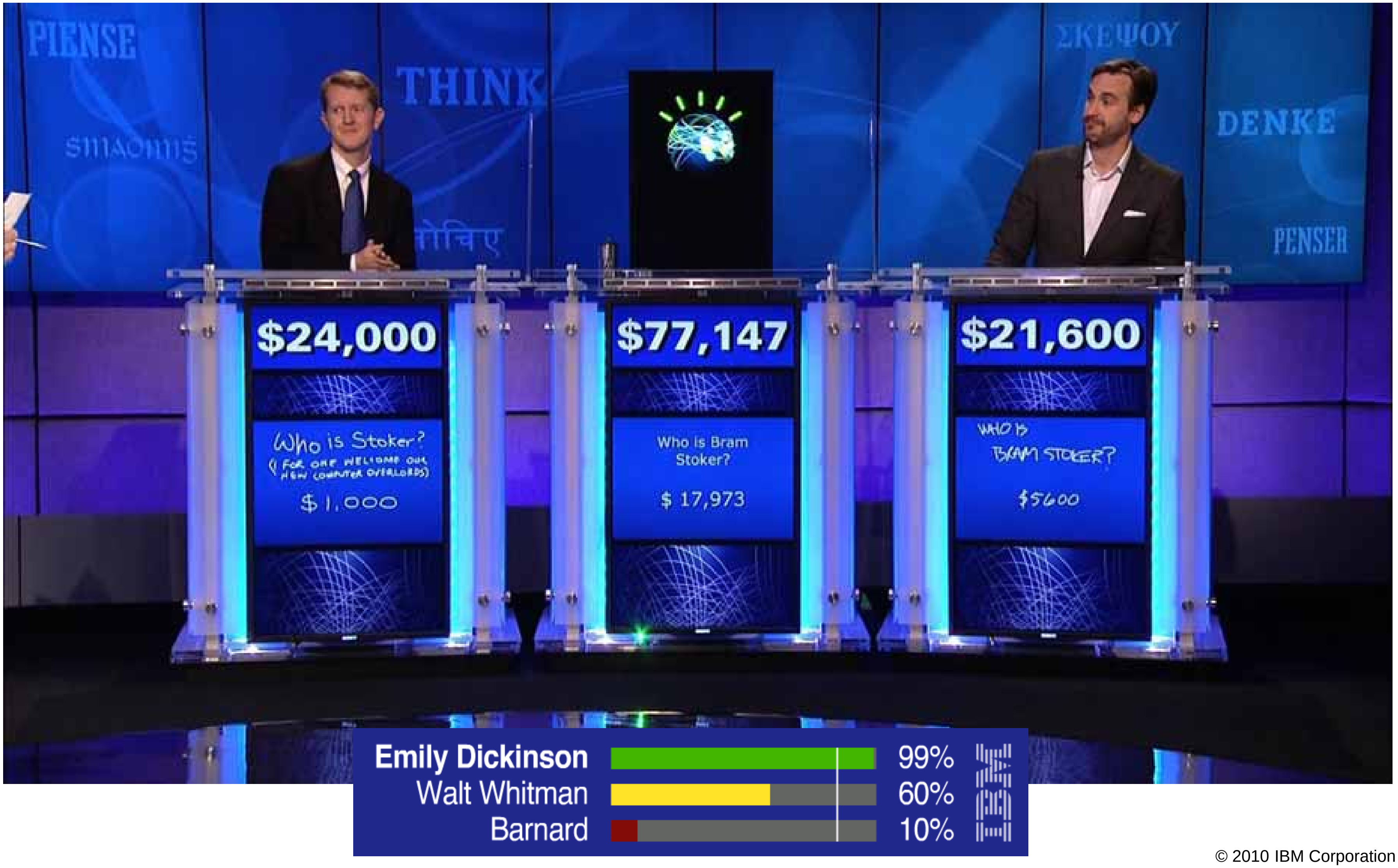
Low Latency, Dense Scale-Out

1 Question on 2880 Cores in a few seconds

Workload Optimization: One Jeopardy! question could take **2 hours on a** single 2.6Ghz Core Optimized & Scaled out on 2880-Core IBM Power750's using UIMA-AS, *Watson* is answering in 2-6 seconds.

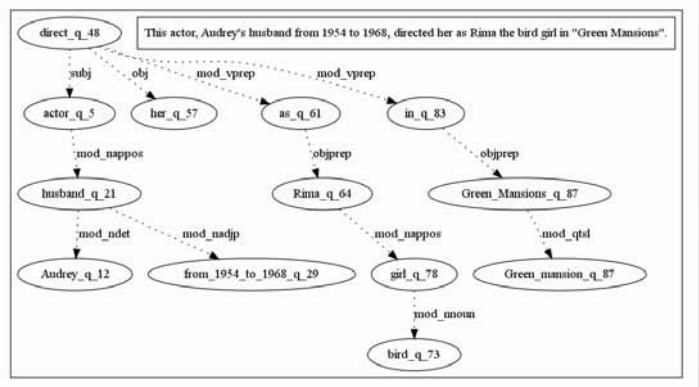

Question
Multiple Interpretations
100s sources
100s Possible Answers
1000's of Pieces of Evidence
100,000's scores from many simultaneous Text Analysis Algorithms
Question & Topic Analysis
Question Decomposition
Hypothesis Generation
Hypothesis and Evidence Scoring
Synthesis
Final Confidence Merging & Ranking
Hypothesis Generation
Hypothesis and Evidence Scoring
...
Answer & Confidence
built on UIMA for interoperability
built on UIMA-AS for scale-out and speed

With Precision, Accurate Confidence and Speed, the rest was History



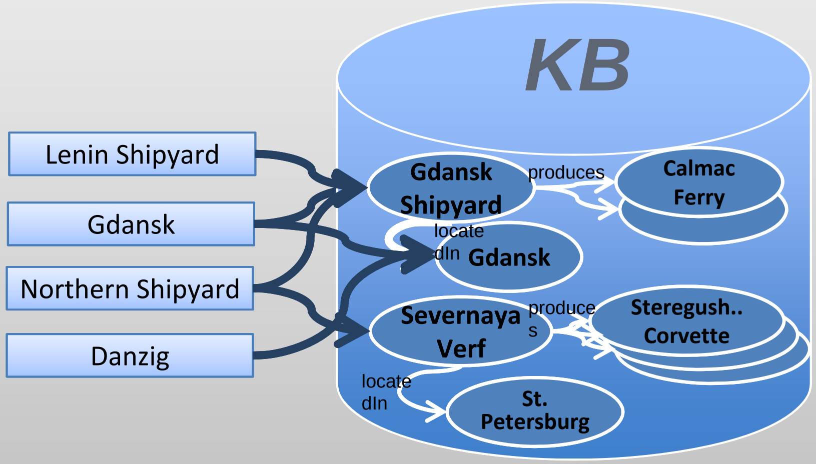
NLP Technology Highlights

Question Processing



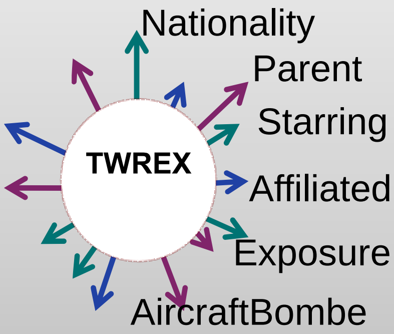
Dependency parse, Focus/LAT detection: 6000 rules, Decomp.
Eval on Jeopardy!: Parser: 92.4% acc, Stat LAT detection: 96.8%

KAFE: Knowledge From Extracted Content



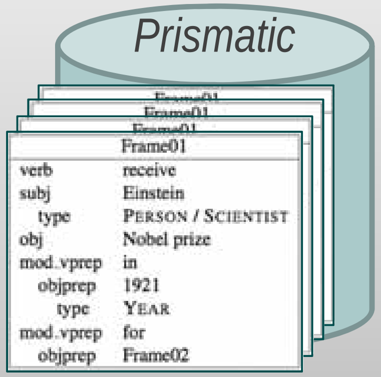
Entity Disambiguation, Entity Typing, Type Disambiguation
Eval on Wikipedia Disambig Task, F-score 92.5

Relation Extraction



Semantic Relation Repository 7000 rels
Eval on ACE: F-score 73.2 (leading score)

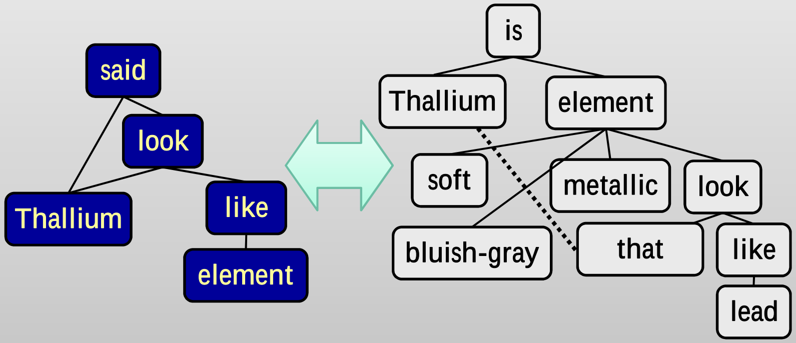
Linguistic Frame Extraction



Frame01	
verb	receive
subj	Einstein
type	PERSON / SCIENTIST
obj	Nobel prize
mod.vprep	in
objprep	1921
type	YEAR
mod.vprep	for
objprep	Frame02

>1 Billlon Frames. Mining from ClueWeb:
SVO/isa/etc. cuts,
Intensional/Extensional representation

Passage Matching Ensemble



Synonymy, Temporal/Geographic Reasoning, Linguistic Axioms
Eval on RTE 2010 Text Entailment: F-score 48.8 (leading score)

Potential Business Applications

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

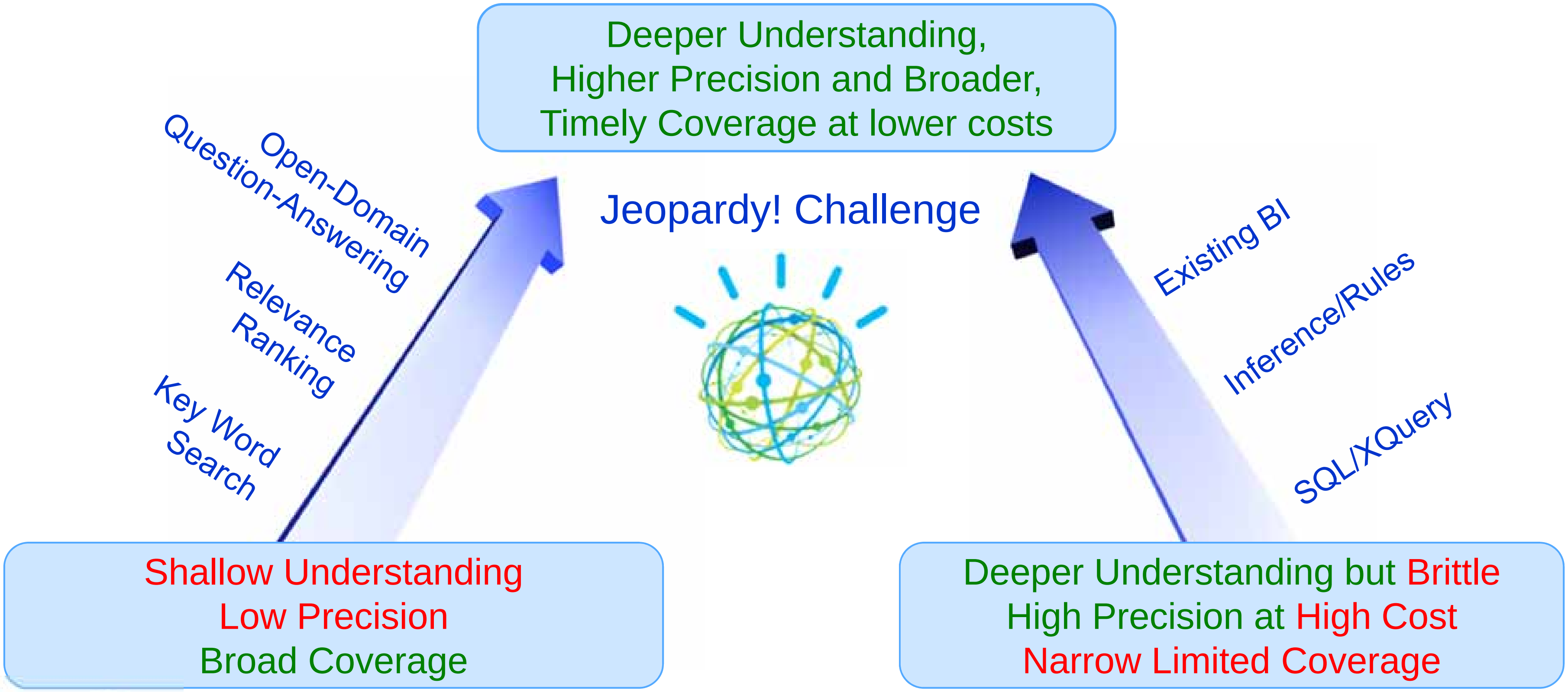
Tech Support: Help-desk, Contact Centers

Enterprise Knowledge Management and Business Intelligence

Government: Improved Information Sharing and Security

Watson’s Principal Value Proposition

Efficient decision support over unstructured (and structured) content



Deeper Understanding,
Higher Precision and Broader,
Timely Coverage at lower costs

Jeopardy! Challenge

Open-Domain
Question-Answering

Relevance
Ranking

Key Word
Search

Existing BI

Inference/Rules

SQL/XQuery

Shallow Understanding
Low Precision
Broad Coverage

Deeper Understanding but Brittle
High Precision at High Cost
Narrow Limited Coverage

Unstructured Data

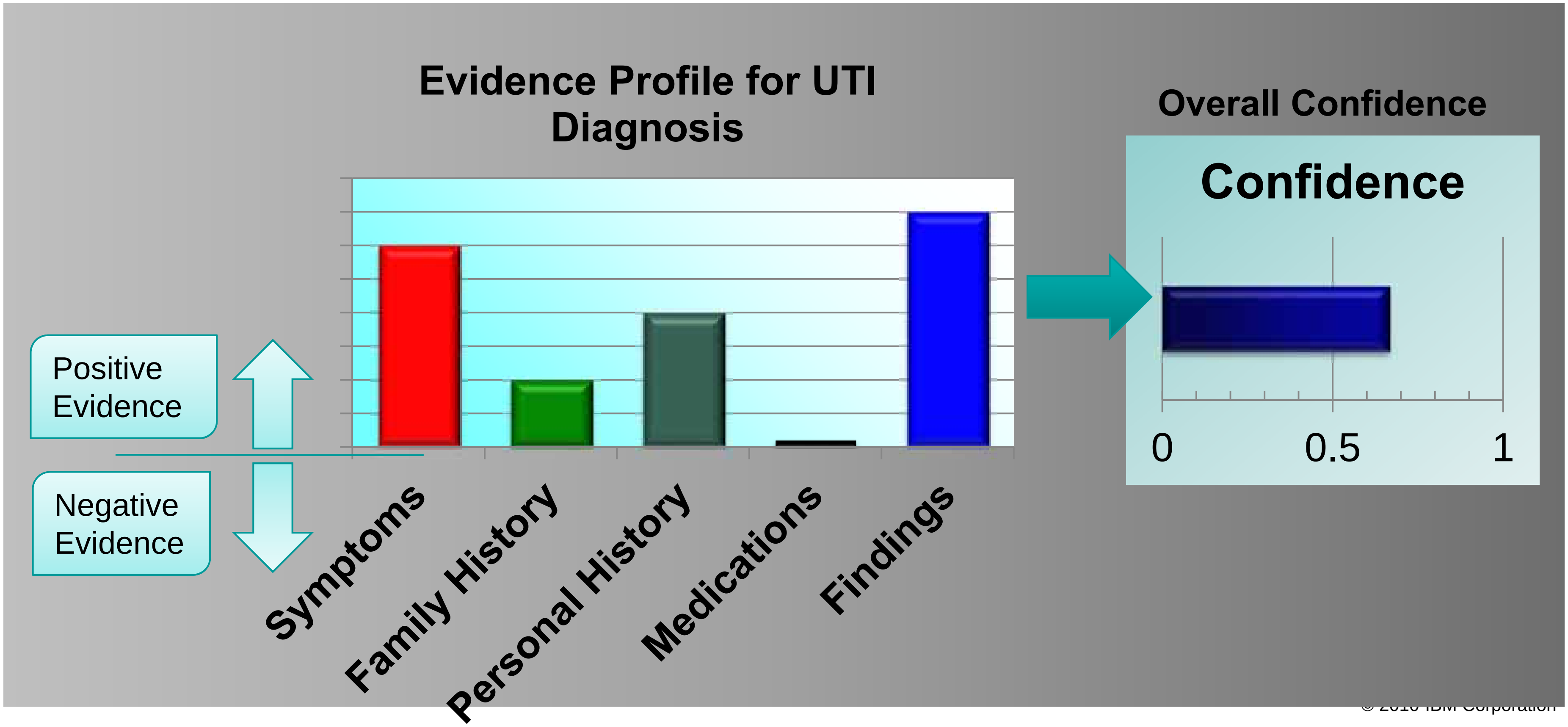
- Broad, rich in context
- Rapidly growing, current
- Invaluable yet under utilized

Structured Data

- Precise, explicit
- Narrow, expensive

Evidence Profiles from disparate data is a powerful idea

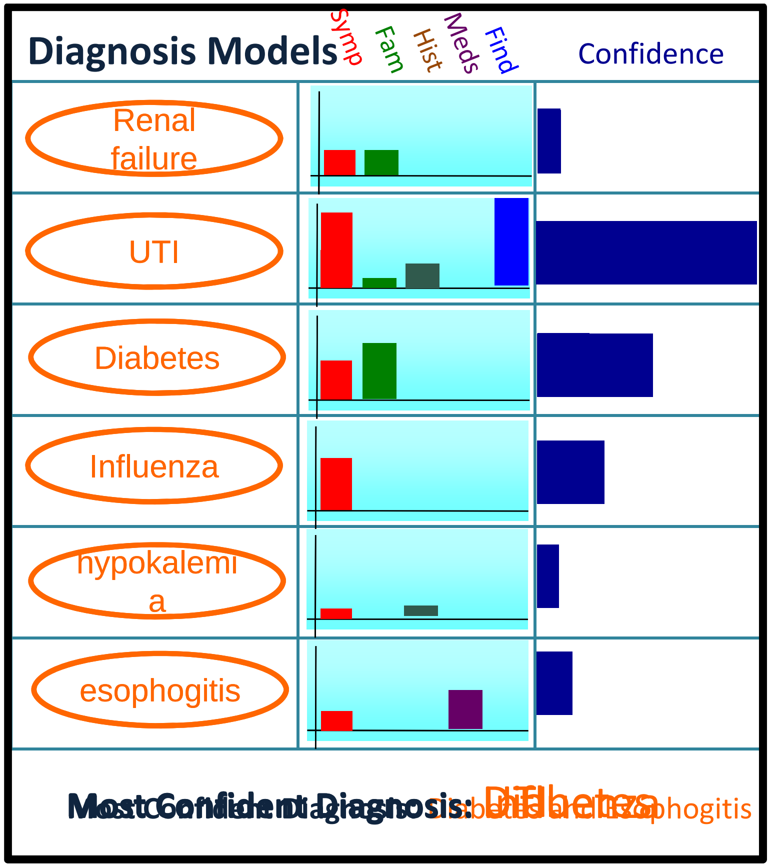
- Each dimension contributes to supporting or refuting hypotheses based on
  - **Strength of evidence**
  - **Importance of dimension for diagnosis** (learned from training data)
- Evidence dimensions are combined to produce an overall confidence



DeepQA in Continuous Evidence-Based Diagnostic Analysis

Considers and synthesizes a broad range of evidence improving quality, reducing cost

Huge Volumes of Texts, Journals, References, DBs etc.



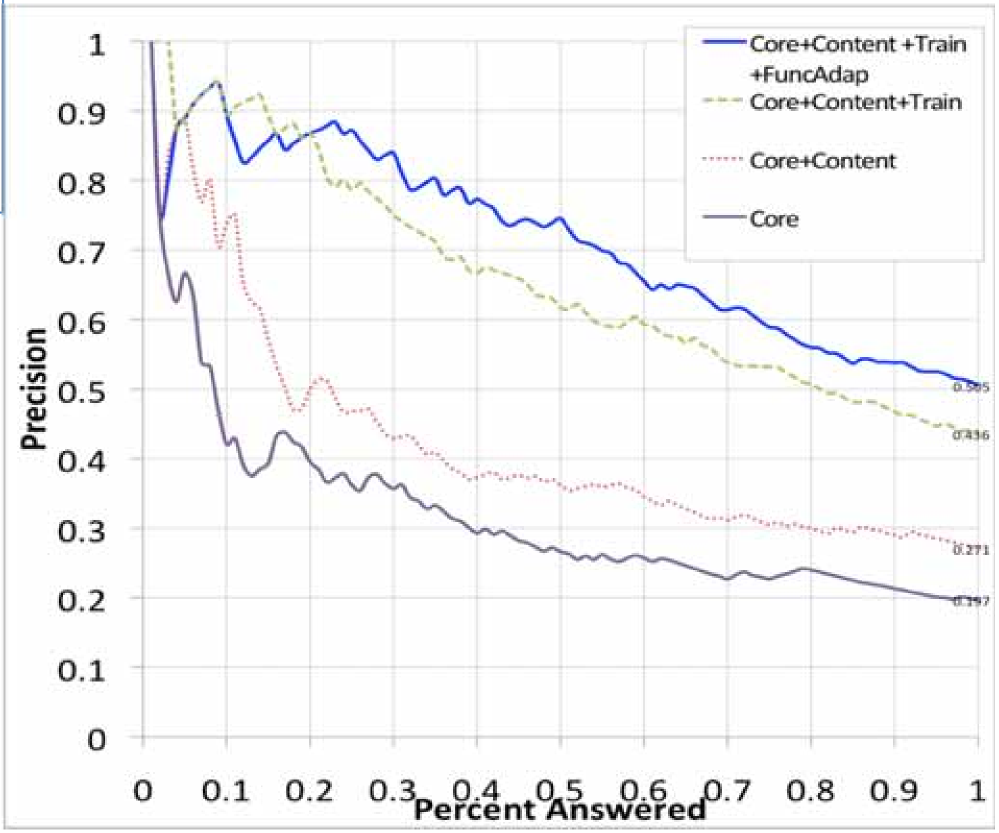
Medical Adaptation – The Beginning – Doctor’s Dilemma

- American College of Physician’s Doctors Dilemma Questions
- We ran Jeopardy system (out of the box) on 188 blind diagnosis questions

The syndrome characterized by narrowing of the extra-hepatic bile duct from mechanical compression by a gallstone impacted in the cystic duct

This inflammation is characterized by nasal mucosal atrophy and foul-smelling crusts in the nasal passages

Skin rash associated with Lyme Disease



THANK YOU

Taking Watson beyond Jeopardy!

Understanding

Specific Questions

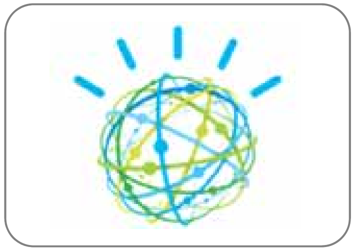
The type of murmur associated with this condition is harsh, systolic, and increases in intensity with Valsalva

From specific questions to rich, incomplete problem scenarios (e.g. EHR)

Rich Problem Scenarios

Interacting

Question-In/Answer-Out



Evidence analysis and look-ahead, drive interactive dialog to refine answers and evidence

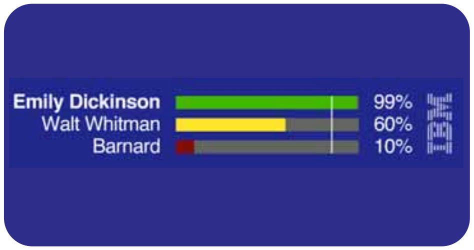
Input, Responses

Refined Answers, Follow-up Questions

Interactive Dialog Teach Watson

Explaining

Precise Answers & Accurate Confidences

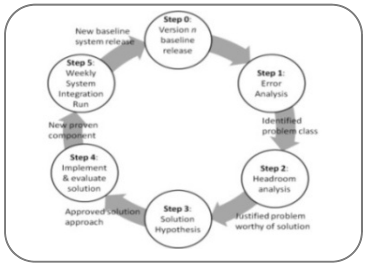


Move from quality answers to quality answers and evidence

Comparative Evidence Profiles

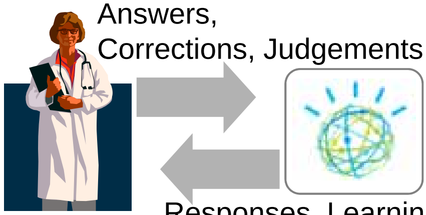
Learning

Batch Training Process


Step 0: Version n baseline release
Step 1: Error Analysis
Identified problem class
Step 2: Headroom analysis
Justified problem worthy of solution
Step 3: Solution Hypothesis
Approved solution approach
Step 4: Implement & evaluate solution
New proven component
Step 5: Weekly System Integration Run
New baseline system release

Scale domain learning and adaptation rate and efficiency

Answers, Corrections, Judgements



Responses, Learning Questions

Continuous Training & Learning Process