Building Watson Beyond Jeopardy

David Ferrucci, IBM Fellow
Principal Investigator
DeepQA@ IBM Research
What is Watson?

Dave: How do I get 8% return in this economy?
Watson: Don’t know…What is Toronto??????

Business Value

Inflated Expectation
Watson knows all things!

Watson can only play games

Disillusionment

No Way!

Enlightenment

Watson weighs many alternatives finds, compares valuable evidence quickly
Overview

- A start in Jeopardy!
  - Motivations and Challenges

- Watson’s Value Proposition
  - Real Application Potential

- Watson 2.0
  - What is next?
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
Want to Play Chess or Just Chat?

**Chess**
- A finite, mathematically well-defined search space
- Limited number of moves and states
- All the symbols are completely grounded in the mathematical rules of the game

**Human Language**
- Words by themselves have no meaning
- Only grounded in *human cognition*
- Words navigate, align and communicate an infinite space of intended meaning
- Computers can **not** ground words to human experiences to derive meaning
Hard Questions or Hard Answers or both? Depends on the Evidence.

- 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.

- X ran this?

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

(many different expressions, meaning highly dependent on context)

- food would “get stuck” when she was swallowing
- can cause food to move slowly in the esophagus.
- swallowing difficulty...
- food gets held-up...

- Abdomen Pain
- Urination Pain
- pneumaturia - bubbles in the urine
- Abdomen Pain exacerbated by exercise

- Flank Pain
- Kidney Pain
- Lower Back Pain

- Causation
- Chronology
- Terminology
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.
This actor, Audrey's husband from 1954 to 1968, directed her as Rima the bird girl in "Green Mansions".
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.

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)
LINCOLN BLOGS

TREASURY SECY. CHASE JUST SUBMITTED THIS TO ME FOR THE THIRD TIME — GUESS WHAT, PAL. THIS TIME I'M ACCEPTING IT

RESIGNATION?

FRIEND REQUEST?

Treasury Secy. Chase wants to be friends with you.

Confirm Friend Request
Inducing Meaning

Sentence Parsing

Volumes of Text ➔ Syntactic Frames ➔ Semantic Frames

Generalization & Statistical Aggregation

- 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 $\Rightarrow$ Is_a(Fluid, Liquid) $\Rightarrow$ ?

Learned $\Rightarrow$ Is_a(Fluid, Liquid) $\Rightarrow$ yes.
In May 1898 Portugal celebrated the 400th anniversary of this explorer’s arrival in India.

In May, Gary arrived in India after he celebrated his anniversary in Portugal.

This evidence suggests “Gary” is the answer BUT the system must learn that keyword matching may be weak relative to other types of evidence.
In May 1898 Portugal celebrated the 400th anniversary of this explorer’s arrival in India.

On the 27th of May 1498, Vasco da Gama landed in Kappad Beach.

This system must learn this is better evidence.

The evidence is still not 100% certain.
A long, tiresome speech delivered by a frothy pie topping

Diatribes

Harangues

Whipped Cream

Meringues

Answer: Meringue, Harangue

Category: Edible Rhyme Time
Divide and Conquer (Typical in Final Jeopardy!)

Lyndon B Johnson

In 1968 this man was U.S. president.

When "60 Minutes" premiered this man was U.S. president.

Must identify and solve sub-questions from different sources to answer the top level question

The DeepQA architecture attempts different decompositions and recursively applies the QA algorithms
The Missing Link

On hearing of the discovery of George Mallory's body, he told reporters he still thinks he was first.
What It Takes to compete against Top Human Jeopardy! Players
Our Analysis Reveals the Winner’s Cloud

Top human players are remarkably good.

Winning Human Performance

Grand Champion Human Performance

2007 QA System

More Confident

Less Confident

Each dot – actual historical human Jeopardy! games
What It Takes to compete against Top Human Jeopardy! Players

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

In 2007, we committed to making a Huge Leap!

Computers? Not So Good.

2007 QA Computer System

Winning Human Performance

Grand Champion Human Performance

Each dot – actual historical human Jeopardy! games

More Confident

Less Confident

% Answered
Some Ground 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 many diverse methods
  - 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 from 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

- **Extreme Collaboration**
  - Implemented “One Room” to optimize team work and communication
  - 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
Integrated System Performance was Job #1

- 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.
Growing Pains (Some Early Answers)

THE AMERICAN DREAM
Decades before Lincoln, Daniel Webster spoke of government "made for", "made by" & "answerable to" them.

NEW YORK TIMES HEADLINES
An exclamation point was warranted for the "end of" this! In 1918

MILESTONES
In 1994, 25 years after this event, 1 participant said, "For one crowning moment, we were creatures of the cosmic ocean"

THE QUEEN'S ENGLISH
Give a Brit a tinkle when you get into town & you've done this

FATHERLY NICKNAMES
This Frenchman was "The Father of Bacteriology"

© 2010 IBM Corporation
Central repository of Experiments

<table>
<thead>
<tr>
<th>Sel</th>
<th>Experiment</th>
<th>Experiment Date &amp; Time</th>
<th>Experiment Description</th>
<th>User</th>
</tr>
</thead>
<tbody>
<tr>
<td>7361</td>
<td>W34 Sparring Games 80-85 (Session2A, 9/10/10, with non-final answers)</td>
<td>Sep 13, 2010</td>
<td>jccencc</td>
<td></td>
</tr>
<tr>
<td>7357</td>
<td>W34 Sparring Games 80-85 (Session2A, 9/10/10)</td>
<td>Sep 13, 2010</td>
<td>jccencc</td>
<td></td>
</tr>
<tr>
<td>7274</td>
<td>W15-Temporal Ablation-T17 Ablate-AllTemporal (baseline: 8890)</td>
<td>Aug 27, 2010</td>
<td>spatward</td>
<td></td>
</tr>
<tr>
<td>7269</td>
<td>Weekly Run: Week 34, 2010 – Try60</td>
<td>Aug 27, 2010</td>
<td>adityakal</td>
<td></td>
</tr>
<tr>
<td>7267</td>
<td>swap W30 T14 (standardized, chai, missing, fixes)</td>
<td>Aug 27, 2010</td>
<td>adityakal</td>
<td></td>
</tr>
<tr>
<td>7264</td>
<td>swap W30 T9 (standardized, chai, missing, fixes)</td>
<td>Aug 26, 2010</td>
<td>adityakal</td>
<td></td>
</tr>
</tbody>
</table>
Experiment view: Aggregate Performance Stats


System Configuration

<table>
<thead>
<tr>
<th>Metric</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Question_count</td>
<td>517</td>
</tr>
<tr>
<td>Precision@70.0%</td>
<td>0.7238 [0.6777, 0.7698]</td>
</tr>
<tr>
<td>accuracy</td>
<td>0.5706 [0.5279, 0.6133]</td>
</tr>
<tr>
<td>answer_list_precision</td>
<td>0.0045</td>
</tr>
<tr>
<td>answer_rank_evidence</td>
<td>1</td>
</tr>
<tr>
<td>average_precision</td>
<td>0.8182</td>
</tr>
<tr>
<td>average_precision_lower_attempted</td>
<td>0.2998</td>
</tr>
<tr>
<td>average_precision_upper_attempted</td>
<td>0.8083</td>
</tr>
<tr>
<td>confidence_weighted_score</td>
<td>0.8244</td>
</tr>
<tr>
<td>correct_in_first</td>
<td>295</td>
</tr>
<tr>
<td>coryat_episode_available_count</td>
<td>517</td>
</tr>
<tr>
<td>coryat_episodes_in_place_1</td>
<td>0</td>
</tr>
<tr>
<td>coryat_episodes_in_place_2</td>
<td>0</td>
</tr>
<tr>
<td>coryat_episodes_in_place_3</td>
<td>24</td>
</tr>
<tr>
<td>coryat_episodes_in_place_4</td>
<td>493</td>
</tr>
<tr>
<td>coryat_pet_in_place_1</td>
<td>0</td>
</tr>
<tr>
<td>coryat_pet_in_place_1_tied</td>
<td>0</td>
</tr>
<tr>
<td>coryat_pet_in_place_2</td>
<td>0</td>
</tr>
<tr>
<td>coryat_pet_in_place_3</td>
<td>0</td>
</tr>
<tr>
<td>coryat_pet_in_place_4</td>
<td>0</td>
</tr>
<tr>
<td>coryat_pet_in_place_4_tied</td>
<td>0.0464</td>
</tr>
<tr>
<td>coryat_pet_in_place_5</td>
<td>0.9536</td>
</tr>
<tr>
<td>coryat_total_for_best_human</td>
<td>8771800</td>
</tr>
<tr>
<td>daily_double_accuracy</td>
<td>0</td>
</tr>
<tr>
<td>daily_double_right</td>
<td>0</td>
</tr>
<tr>
<td>daily_double_wrong</td>
<td>0</td>
</tr>
<tr>
<td>ID</td>
<td>Question</td>
</tr>
<tr>
<td>------</td>
<td>--------------------------------------------------------------------------</td>
</tr>
<tr>
<td>49560 (Ep. 1331)</td>
<td>SCIENTISTS: When Einstein won the 1921 Nobel Prize in physics, he was a naturalized citizen of this country</td>
</tr>
<tr>
<td>49621 (Ep. 1480)</td>
<td>MONARCHS: One of two English kings since William the Conqueror who were never crowned</td>
</tr>
<tr>
<td>49682 (Ep. 2352)</td>
<td>HISTORIC NAMES: Though he spent most of his life in Europe, he was governor of the Bahamas for most of World War II</td>
</tr>
<tr>
<td>49743 (Ep. 2682)</td>
<td>THE CABINET: James Wilson of Iowa, who headed this Department for 16 years, served longer than any other cabinet officer</td>
</tr>
<tr>
<td>49804 (Ep. 2826)</td>
<td>AMERICAN HISTORY: Some attribute these 1692 proceedings to the psychic effects of ergot poisoning</td>
</tr>
<tr>
<td>49865 (Ep. 2868)</td>
<td>RULERS: In 44 B.C. He was made dictator for life, a post abolished after he died the same year</td>
</tr>
<tr>
<td>49926 (Ep. 2953)</td>
<td>ENDANGERED SPECIES LIST: Added to the list in 1997 were the Quino checkerspot and Laguna Mountains Skipper, types of these</td>
</tr>
<tr>
<td>49987 (Ep. 2994)</td>
<td>TELEVISION: A 1997 episode of this series guest-starred Philip Michael Thomas and Tommy Chong</td>
</tr>
<tr>
<td>50048</td>
<td>FOOD &amp; DRINK HISTORY: Its original name meant &quot;bitter&quot;</td>
</tr>
</tbody>
</table>
### Experiments Question Detail

**Question:** AMERICAN HISTORY: Some attribute these 1692 proceedings to the psychotic effects of ergot poisoning

**Correct Answer Pattern (Acceptable) (W^I):(The Salem Witch Trials)(Salem witch trials)(Salem witch trial)(Salem trials)\end{quote}(W^I)


### Answer Selection

<table>
<thead>
<tr>
<th>Rank</th>
<th>Score [0.000:0.999]</th>
<th>Answer</th>
<th>Correct</th>
<th>An Ty Cor Aggregator SUM [0.000:4.839]</th>
<th>An Ty Cor Closed Lat [0.000:0.999]</th>
<th>An Ty Cor Gender Cor [0.000:0.997]</th>
<th>An Ty Cor Lexical [0.000:0.987]</th>
<th>An Ty Cor Lexical Verb [0.000:1.000]</th>
<th>An T Tyag [0.000]</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0.347</td>
<td>Ergotism</td>
<td>No</td>
<td>0.726</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.726</td>
</tr>
<tr>
<td>2</td>
<td>0.261</td>
<td>Salem witch trials</td>
<td>Yes</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.726</td>
</tr>
<tr>
<td>3</td>
<td>0.130</td>
<td>Trials</td>
<td>No</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
</tr>
<tr>
<td>4</td>
<td>0.062</td>
<td>symptoms</td>
<td>No</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
</tr>
<tr>
<td>5</td>
<td>0.006</td>
<td>witchcraft</td>
<td>No</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
</tr>
<tr>
<td>6</td>
<td>0.001</td>
<td>Witch-hunt</td>
<td>No</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
</tr>
<tr>
<td>7</td>
<td>0.001</td>
<td>George Savile</td>
<td>No</td>
<td>0.726</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.726</td>
</tr>
<tr>
<td>8</td>
<td>0.001</td>
<td>John Hay</td>
<td>No</td>
<td>0.726</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.726</td>
</tr>
<tr>
<td>9</td>
<td>0.001</td>
<td>the title</td>
<td>No</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
</tr>
<tr>
<td>10</td>
<td>0.001</td>
<td>release</td>
<td>No</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
</tr>
</tbody>
</table>
Individual Feature Engineering
The Machine Learning Problem in Watson

- **Instance:** (Question, Candidate-Answer) Pair
- **Features:** 200 features obtained from Scorers
  - + 400 Derived Features (i.e., Standardized and Imputed)
- **Label:** *True* or *False* for answer correctness
- **Training Set:** Single set collected over *all* questions
  - Typical Training Set: 11K Questions, 1M Instances
  - Final Game Training Set: 20K Questions, 2M Instances
- **Goal:** Produce a *confidence score* representing a probability of being correct and rank
Different question classes require weighing evidence differently

- Statistical (tree kernel) & rule-based methods classify questions
- Partitioned experts were trained for each question class
- When answering a question, question is routed based on 8 different classes

19th CENTURY PORTUGAL: In May 1898 Portugal celebrated the 400th anniversary of this explorer’s arrival in India.

2-PART WORDS: It's the name of the small hole in the sink, often just below the faucet, or what it may prevent

EDIBLE RHYME TIME: A long tiresome speech delivered by a frothy pie topping

FETAL ATTRACTION: From the Greek for "flat cake", this uterine wall organ connects to the fetus via the umbilical cord.
Final Ranking: Successive Classifier Phases

- **Filter**: light-weight features select candidates for deeper analysis

- **Base**: Initial ranking and confidence evaluation, specialized to main question types (600 Features)

- **Transfer Learning**: Transfer-learning for special, rare question types. Use base confidence as a feature. Include 6 target domain features.

- **Merge**: Merge and evaluate *equivalent hypotheses*. Distinct merge policies for each feature (e.g., max, min, average, decaying sum). Select most confident as canonical answer.

- **Elite**: Trains with all available features but on just top-ranked instances (i.e., *successive refinement*)

- **Evidence Diffusion**: *Shifts evidence across semantically related candidates* (e.g., author to book). Learns a how much to shift from related to target for each feature.

- **Multi-answers**: Learn how to combine evidence for assembled answers from individual answers (e.g., “snap, crackle, pop”)

---

Each classifier has access to all features as well as previous classifier scores and ranks.
Feature Interaction Engineering

Question#95506
IT'S A BIT CHILE TODAY: Chile shares its longest land border with this country

<table>
<thead>
<tr>
<th>Experiment Label</th>
<th>Diff Value Bar</th>
<th>Diff Value</th>
<th>Argentina</th>
<th>Bolivia</th>
<th>Headroom</th>
</tr>
</thead>
<tbody>
<tr>
<td>Final Score</td>
<td>0.332</td>
<td>0.121</td>
<td>0.332</td>
<td>0.121</td>
<td></td>
</tr>
<tr>
<td>Weighted Features Sum</td>
<td>0.194</td>
<td>0.572</td>
<td>0.194</td>
<td>0.572</td>
<td>0.876</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Feature Groups/Features</th>
<th>Argentina</th>
<th>Bolivia</th>
<th>Headroom</th>
</tr>
</thead>
<tbody>
<tr>
<td>SpatialDistance</td>
<td>0.020</td>
<td>0.028</td>
<td>n/a</td>
</tr>
<tr>
<td>SpatialDistance-Missing</td>
<td>0.224</td>
<td>0.224</td>
<td>n/a</td>
</tr>
<tr>
<td>SpatialDistance-Std</td>
<td>0.045</td>
<td>0.045</td>
<td>n/a</td>
</tr>
<tr>
<td>SpatialRelation-Std</td>
<td>0.590</td>
<td>0.578</td>
<td>0.012</td>
</tr>
<tr>
<td>IndirDocumentEngine_RANK</td>
<td>-0.389</td>
<td>-0.576</td>
<td>-0.208</td>
</tr>
<tr>
<td>IndirDocumentEngine.Score</td>
<td>-0.013</td>
<td>0.439</td>
<td>0.457</td>
</tr>
<tr>
<td>IndirDocument...e.Score-Std</td>
<td>0.267</td>
<td>0.410</td>
<td>0.143</td>
</tr>
<tr>
<td>FractionOfTitleCovered</td>
<td>0.290</td>
<td>1.173</td>
<td>0.873</td>
</tr>
<tr>
<td>FractionOfTitleCovered-Std</td>
<td>-0.227</td>
<td>-0.436</td>
<td>-0.209</td>
</tr>
<tr>
<td>DocTermMatchScore</td>
<td>0.003</td>
<td>0.006</td>
<td>0.006</td>
</tr>
<tr>
<td>DocTermMatchScore-Std</td>
<td>0.269</td>
<td>0.356</td>
<td>0.086</td>
</tr>
<tr>
<td>SkipBigramScore</td>
<td>-0.037</td>
<td>-0.243</td>
<td>-0.208</td>
</tr>
<tr>
<td>SkipBigramScore-Std</td>
<td>0.226</td>
<td>1.378</td>
<td>1.149</td>
</tr>
<tr>
<td>LACIS_Pass..ALLOW+DEEP</td>
<td>-0.099</td>
<td>0</td>
<td>0.089</td>
</tr>
<tr>
<td>LACIS_Pass..OW+DEEP-Std</td>
<td>-0.560</td>
<td>-0.018</td>
<td>0.542</td>
</tr>
<tr>
<td>TextualAlignment</td>
<td>-0.004</td>
<td>0.013</td>
<td>0.017</td>
</tr>
<tr>
<td>TextualAlignment-Std</td>
<td>-0.242</td>
<td>0.433</td>
<td>0.675</td>
</tr>
<tr>
<td>PassageTermMatch</td>
<td>-0.032</td>
<td>0.750</td>
<td>0.782</td>
</tr>
<tr>
<td>PassageTermMatch-Std</td>
<td>-0.056</td>
<td>0.760</td>
<td>0.816</td>
</tr>
<tr>
<td>PassageTermMatchInFocus</td>
<td>-0.039</td>
<td>-0.402</td>
<td>-0.363</td>
</tr>
<tr>
<td>PassageTermInfocus-Std</td>
<td>0.004</td>
<td>0.371</td>
<td>0.317</td>
</tr>
<tr>
<td>PassageTermInfocus-Std</td>
<td>-0.002</td>
<td>-0.276</td>
<td>-0.379</td>
</tr>
</tbody>
</table>
There’s a Bethel College and a Seminary in both cities. System is not weighing Geospatial evidence high enough.

These are Aggregate Feature Labels. For each of these shown here there are 5-20 finer grained features each corresponding to individual evidence scoring algorithms.

Humans may get this based on the pun to choose St. Paul since it’s a “holy” city. We are adding a Pun Scorer That will look for these sorts of Pun-like relationships.
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.
Evidence: Time, Popularity, Source, Classification etc.

**Clue:** You’ll find Bethel College and a Seminary in this “holy” Minnesota city.

There’s a Bethel College and a Seminary in both cities. System is not weighing location evidence high enough to give St. Paul the edge.
Evidence: Puns

Clue: You’ll find Bethel College and a Seminary in this “holy” Minnesota city.

Humans may get this based on the pun since St. Paul since is a “holy” city. We added a Pun Scorer that discovers and scores Pun relationships.

IBM Watson
Playing in the Winners Cloud
ML Enhancements

![Graph showing precision vs percent answered for different models: Base, Base+Std, Base+Std+Impute, Base+Std+Impute+Routes+Phases. Each line represents a different model, with the y-axis showing precision ranging from 0.65 to 1.0 and the x-axis showing percent answered ranging from 0 to 100. The graph highlights the performance improvements with the addition of more features.]
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.
Deployment Models

Development System
Easy to change/update
High Experimental Throughput

Production System
Low Latency, Dense Scale-Out

Algorithm & Data Migration

Software Bottlenecks

2500 Questions on ~1500 Cores in a few hours

1 Question on 2880 Cores in a few seconds
Game Strategy
Managing the Luck of the Draw

Risk Management
Modulate Aggressiveness

Dynamic Confidence

Direct Impact on Earnings
DDs & FJ

Betting

Learn at lower risk

Clue Selection

Prior Probabilities

Common Betting Patterns

Performance Relative to Competition

© 2010 IBM Corporation
With Precision, Accurate Confidence and Speed, the rest was History
Was Jeopardy! a True Technology Driver

Question Processing

<table>
<thead>
<tr>
<th>Relation</th>
<th>TWREX</th>
</tr>
</thead>
<tbody>
<tr>
<td>Nationality</td>
<td>Parent</td>
</tr>
<tr>
<td>Parent</td>
<td>Starring</td>
</tr>
<tr>
<td>Affiliated</td>
<td>Exposure</td>
</tr>
<tr>
<td>AircraftBomber</td>
<td></td>
</tr>
</tbody>
</table>

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

KAFE: Knowledge From Extracted Content

<table>
<thead>
<tr>
<th>KB</th>
</tr>
</thead>
<tbody>
<tr>
<td>Lenin Shipyard</td>
</tr>
<tr>
<td>Gdansk Shipyard</td>
</tr>
<tr>
<td>Northern Shipyard</td>
</tr>
<tr>
<td>Severnaya Verf</td>
</tr>
<tr>
<td>Danzig</td>
</tr>
<tr>
<td>Calmac Ferry</td>
</tr>
</tbody>
</table>

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

Relation Extraction

<table>
<thead>
<tr>
<th>TWREX</th>
</tr>
</thead>
<tbody>
<tr>
<td>AircraftBomber</td>
</tr>
</tbody>
</table>

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

Linguistic Frame Extraction

<table>
<thead>
<tr>
<th>Frame01</th>
</tr>
</thead>
<tbody>
<tr>
<td>verb</td>
</tr>
<tr>
<td>subj</td>
</tr>
<tr>
<td>type</td>
</tr>
<tr>
<td>obj</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Frame02</th>
</tr>
</thead>
<tbody>
<tr>
<td>verb</td>
</tr>
<tr>
<td>subj</td>
</tr>
<tr>
<td>type</td>
</tr>
<tr>
<td>obj</td>
</tr>
</tbody>
</table>

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

Passage Matching Ensemble

<table>
<thead>
<tr>
<th>Prismatic</th>
</tr>
</thead>
<tbody>
<tr>
<td>Thallium</td>
</tr>
<tr>
<td>soft</td>
</tr>
<tr>
<td>look</td>
</tr>
</tbody>
</table>

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
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

- Symptom
- Family History
- Patient History
- Medications
- Tests/Findings
- Notes/Hypotheses

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

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

Diagnosis Models

<table>
<thead>
<tr>
<th>Diagnosis</th>
<th>Sympt</th>
<th>Fam</th>
<th>Hist</th>
<th>Meds</th>
<th>Find</th>
<th>Confidence</th>
</tr>
</thead>
<tbody>
<tr>
<td>Renal failure</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>UTI</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Diabetes</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Influenza</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>hypokalemia</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>esophagitis</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Most Confident Diagnosis: Diabetes

Most Confident Diagnosis: Diabetes
It’s all about the evidence

Evidence Profile: Diagnosis

**Sources**

Textbook
- Uveitis specifically refers to inflammation of the middle layer of the eye, termed the "uvea" but in common usage may refer to any inflammatory process involving the interior of the eye.

Website
- Uveitis usually develops quickly and generally affects only one eye. Signs and symptoms may include any or all of the following, pain in the eye or brow region, worsened eye pain when exposed to bright light, and reddened eye, especially adjacent to the iris.

Medical Journal
- The most common form of uveitis is anterior uveitis, which involves inflammation in the front part of the eye. It is often called iritis because it usually only affects the iris, the colored part of the eye.

Textbook
- Uveitis is often marked by moderate to intense pain and usually involves impaired eyesight.

**Factors**

Present Factors
- Red eye
- Eye pain
- Eye inflammation
- Blurred vision
- Floating spots
- Sensitivity to light

Absent Factors
- Patient has low visibility
- Patient has loss of hearing
- Patient wears prescription lenses
- Patient has yellow eyes
- Patient has pus in eyes
- Patient is allergic to pollen
Rapid Innovation Process

**Start**
- Systems Engineers

**New Algorithms**
- Researchers

**Algorithm Development**
- Systems Engineers

**Triage**
- Researchers

**Research Loop**
- Headroom Analysis
- Systems Analysts

**Statistical Machine Learning**
- Watson2 Biweekly Build
- Training Data
- ML Models

**Run**
- Test Data

**Test Data Input & Output**
- Domain Experts

**Interactive Learning TeachWatson**
- Vetted Output & Enriched Training Data

**Judge and Teach**
- Original Content

**Source Evaluation And Prep**
- Prep Ideas

**Learning Knowledge Extraction**
- Content Analysis

**Y**
- Positive Impact?

**N**
- Systems Engineers

**Content Analysis**
- Systems Analysts

**Research Opportunities**
- Researchers

**Headroom Analysis**
- Accuracy Analysis

**Rapid Innovation Process**
**Taking Watson beyond Jeopardy!**

<table>
<thead>
<tr>
<th>Understanding</th>
<th>Interacting</th>
<th>Explaining</th>
<th>Learning</th>
</tr>
</thead>
<tbody>
<tr>
<td>Specific Questions</td>
<td>Question-In/Answer-Out</td>
<td>Precise Answers &amp; Accurate Confidences</td>
<td>Batch Training Process</td>
</tr>
</tbody>
</table>

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)

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

Move from quality answers to quality answers and evidence.

Scale domain learning and adaptation rate and efficiency.

**Rich Problem Scenarios**

**Interactive Dialog** Teach Watson

**Comparative Evidence Profiles**

**Continuous Training & Learning Process**

Answers, Corrections, Judgements

Responses, Learning Questions

**Entire Medical Record**

Input, Responses

Dialog

Refined Answers, Follow-up Questions

**Comparative Evidence Profiles**

**Continuous Training & Learning Process**
THANK YOU
TORONTO?
Categories are not as simple as they seem

Watson uses statistical machine learning to discover that Jeopardy! categories are only weak indicators of the answer type.

U.S. CITIES
St. Petersburg is home to Florida's annual tournament in this game popular on shipdecks (Shuffleboard)
Rochester, New York grew because of its location on this (the Erie Canal)

Country Clubs
From India, the shashpar was a multi-bladed version of this spiked club (a mace)
A French riot policeman may wield this, simply the French word for "stick" (a baton)

Authors
Archibald MacLeish based his verse play "J.B." on this book of the Bible (Job)
In 1928 Elie Wiesel was born in Sighet, a Transylvanian village in this country (Romania)
Toronto vs. Chicago

US CITIES
Its largest airport is named for a World War II hero; its second largest, for a World War II battle

Low because of weak evidence in content

Overall confidence was below threshold for both answers

<table>
<thead>
<tr>
<th>City</th>
<th>Confidence</th>
</tr>
</thead>
<tbody>
<tr>
<td>Toronto</td>
<td>14%</td>
</tr>
<tr>
<td>Chicago</td>
<td>11%</td>
</tr>
<tr>
<td>Omaha</td>
<td>10%</td>
</tr>
</tbody>
</table>

Low because being a **US City** is not a strong requirement simply based on Jeopardy! category
When being a **US City** is a more trusted requirement, **Chicago** is in first position but still weak, because Watson is not convinced about the WWII evidence it could find.

**US CITIES**

This **US City's** largest airport is named for a World War II hero; its second largest, for a World War II battle

---

**Chicago** 32%

**Toronto** 28%