Semantic Technologies in IBM Watson^{TM}

Professor: Alfio Massimiliano Gliozzo
TA: Or Biran
Hands On: Siddharth Patwardhan

Semantic Technologies in IBM Watson™
Lesson 1: The Jeopardy!™ grand challenge

Professor: Alfio Massimiliano Gliozzo
TA: Or Biran
Hands On: Siddharth Patwardhan

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 IBM Customers
  – 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.

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

IBM WATSON

Easy Questions?

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

≠

Hard Questions?

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.

Automatic Open-Domain Question Answering

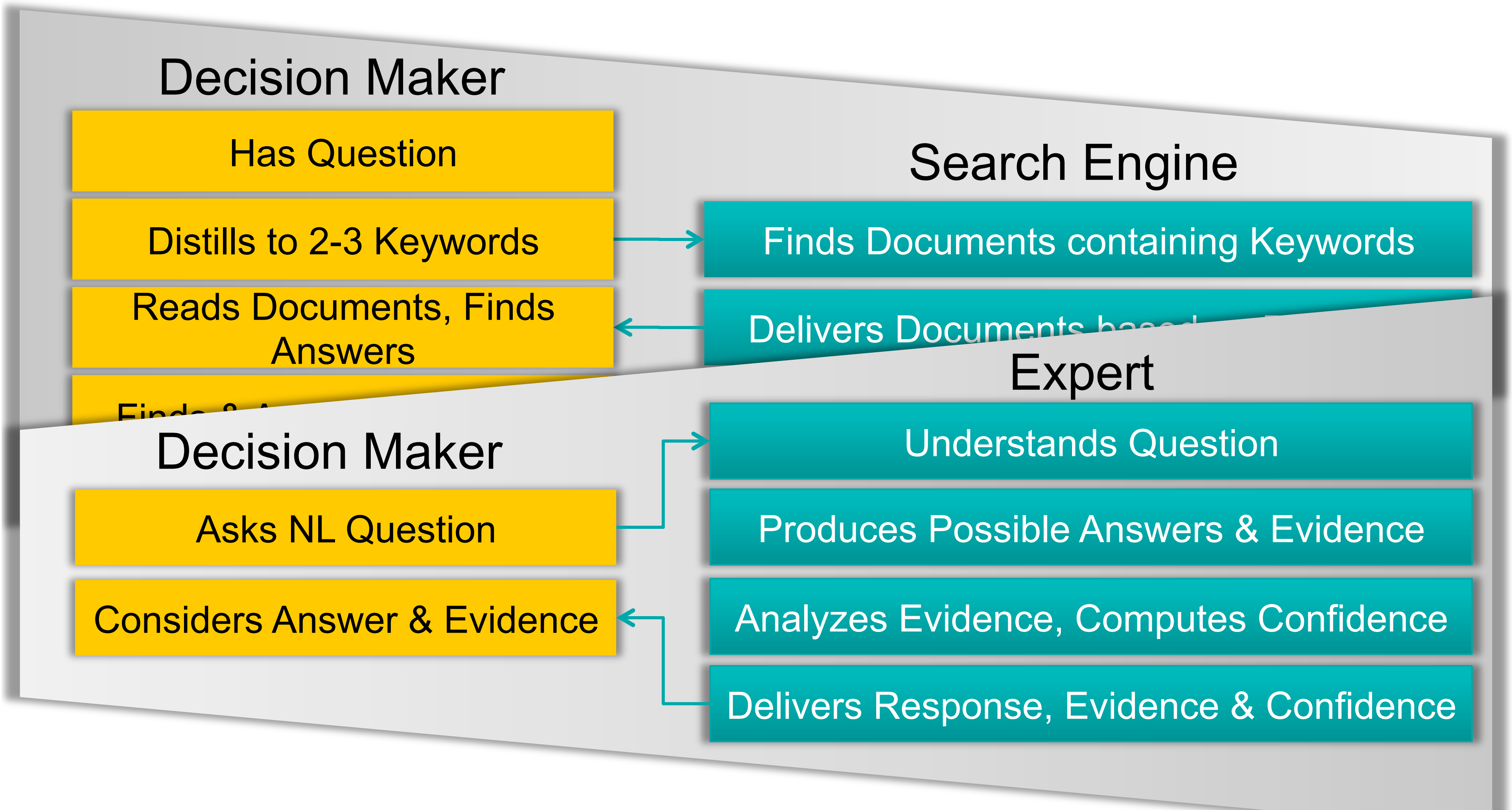
A Long-Standing Challenge in Artificial Intelligence to emulate human expertise

- Given
  - Rich **Natural Language Questions**
  - Over a **Broad Domain of Knowledge**

- Deliver
  - **Precise Answers:** Determine what is being asked & give precise response
  - **Accurate Confidences:** Determine likelihood answer is correct
  - **Consumable Justifications:** Explain why the answer is right
  - **Fast Response Time:** Precision & Confidence in <3 seconds

IBM WATSON

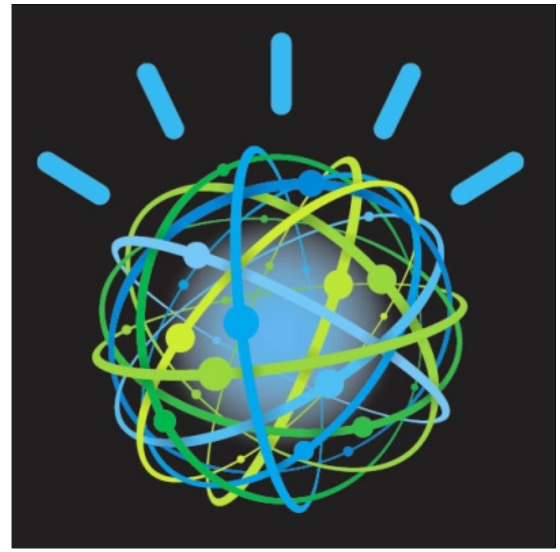
Informed Decision Making: Search vs. Expert Q&A



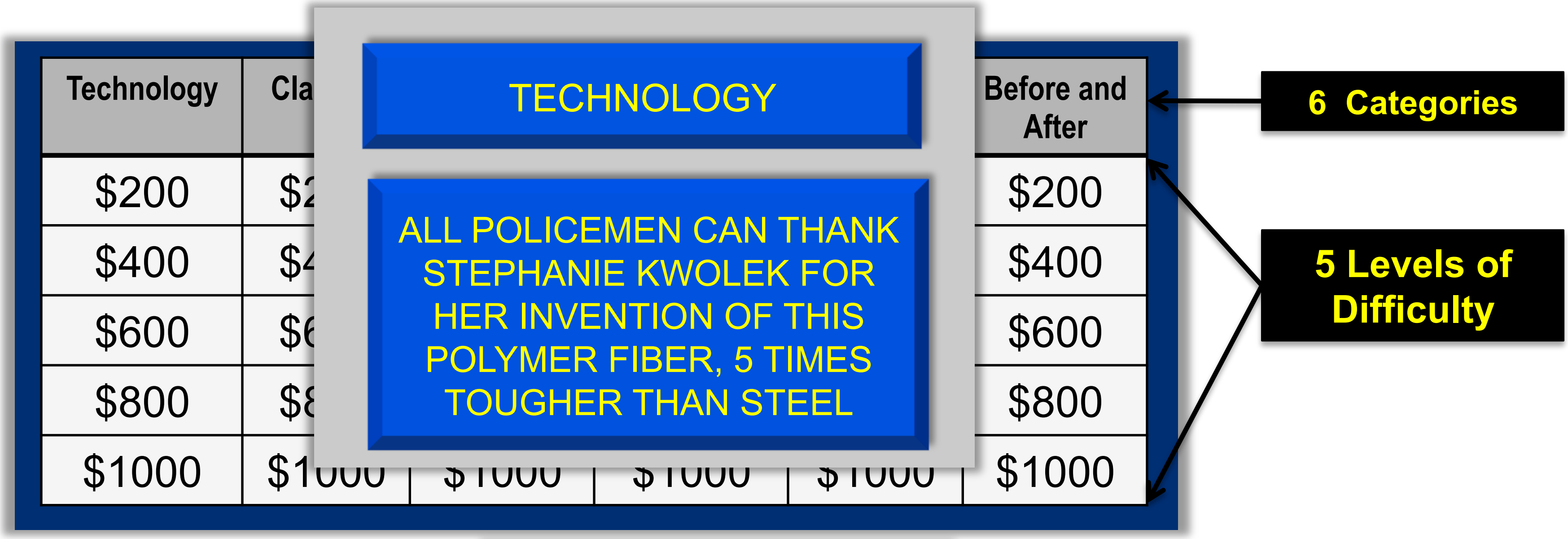
What is Jeopardy?

IBM WATSON

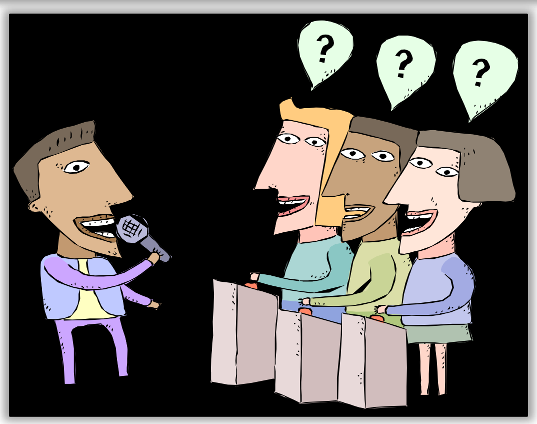
- ***Jeopardy!*** is an American quiz show
  - 1964 – Today
- **answer-and-question** format
  - contestants are presented with clues in the form of answers
  - must phrase their responses in question form.
- Example
  - Category: General Science
  - Clue: When hit by electrons, a phosphor gives off electromagnetic energy in this **form**
  - Answer: What is <u>light</u>?



Jeopardy! Game Play



- ❑ 1 of 3 Players Selects a Clue
- ❑ Host reads Clue out loud
- ❑ All Players compete to answer
- ❑ 1st to buzz-in gets to answer



- ❑ **IF correct**
  - ➢ earns $ value
  - ➢ selects Next Clue
- ❑ **IF wrong**
  - ➢ **loses** $ value
  - ➢ other players buzz again (rebounds)

❑ Two Rounds Per Game + Final Question

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.
What is down?

$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.
Who is D'Artagnan?

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

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

Some Basic Jeopardy! Clues

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

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

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

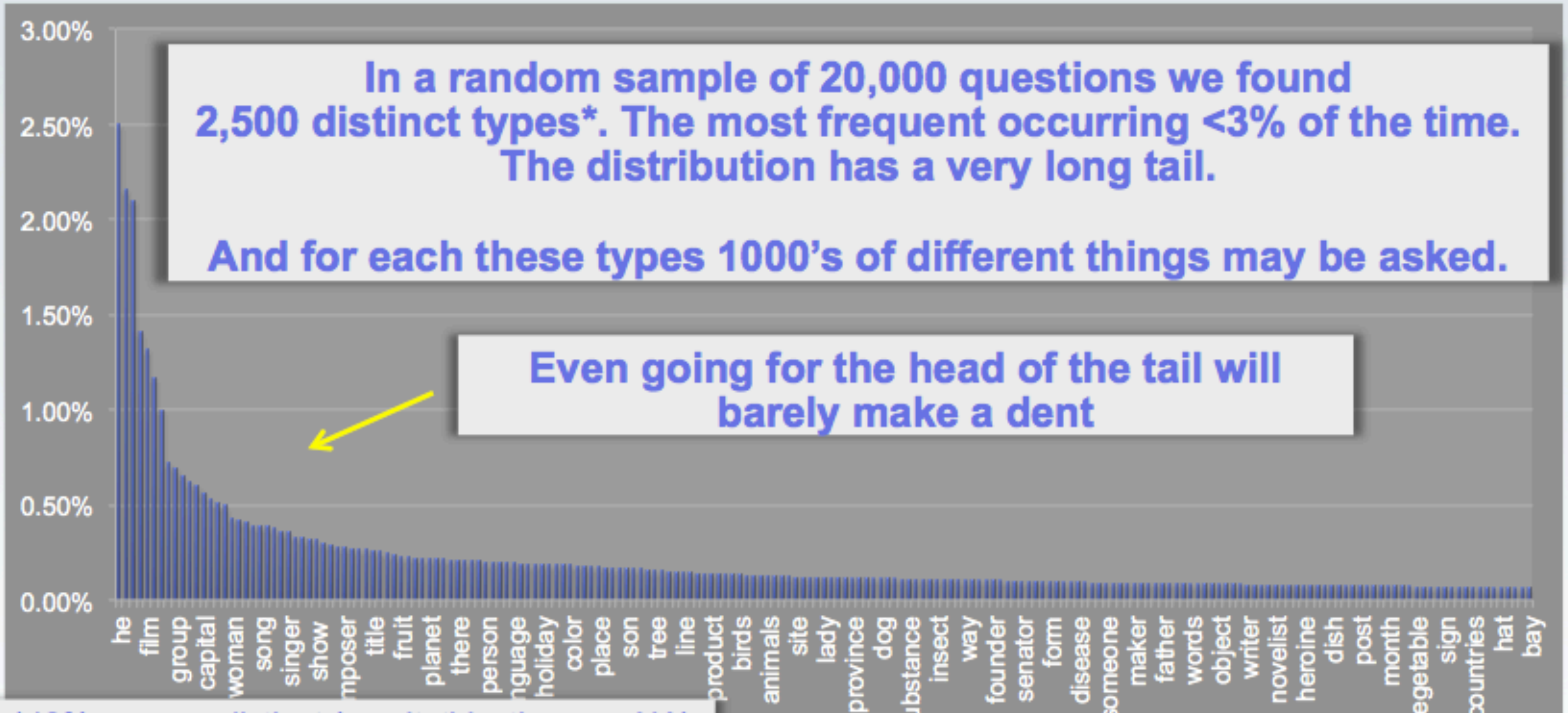
- Secy. Chase just submitted **this** to me for the third time--guess what, pal. This time I'm accepting **it**
- Category: Lincoln Blogs
- Answer: **his resignation**

IBM WATSON

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



*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.

Nested Decomposition Questions
(Typical in Final Jeopardy!)

IBM WATSON

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

Lyndon B Johnson

In 1968 this man was U.S. president.

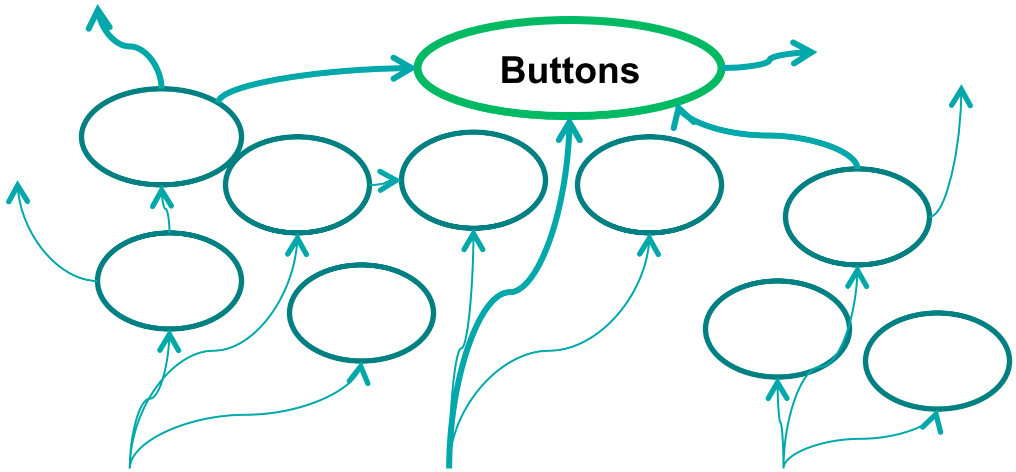
When "60 Minutes" premiered | this man was U.S. president.

?

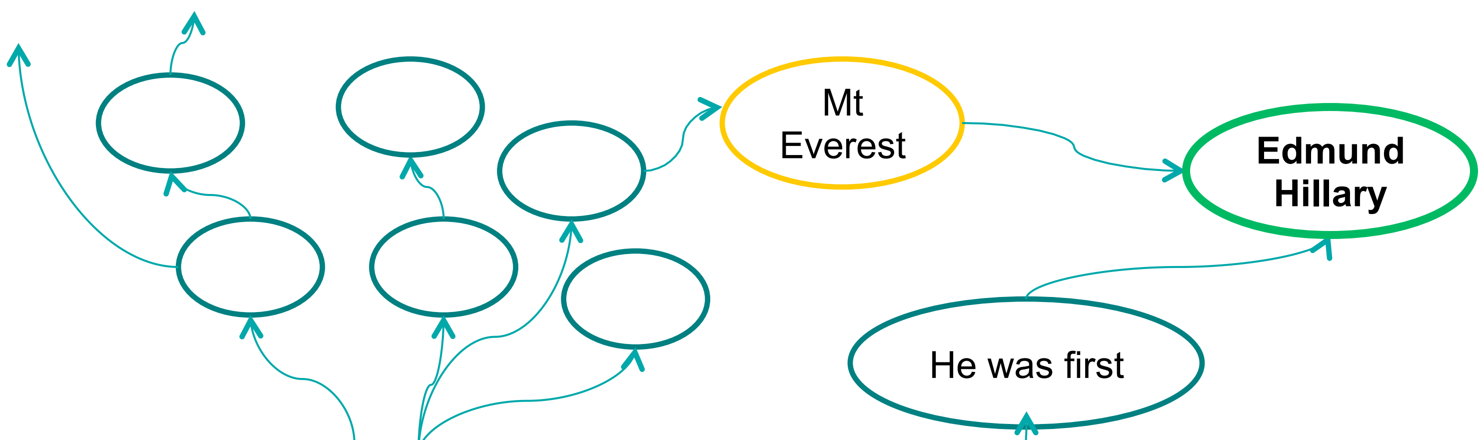
The DeepQA architecture attempts different *decompositions* and recursively applies the QA algorithms

Challenging Questions: Missing Links

IBM WATSON



Category: COMMON BONDS Shirts, TV remote controls, Telephones



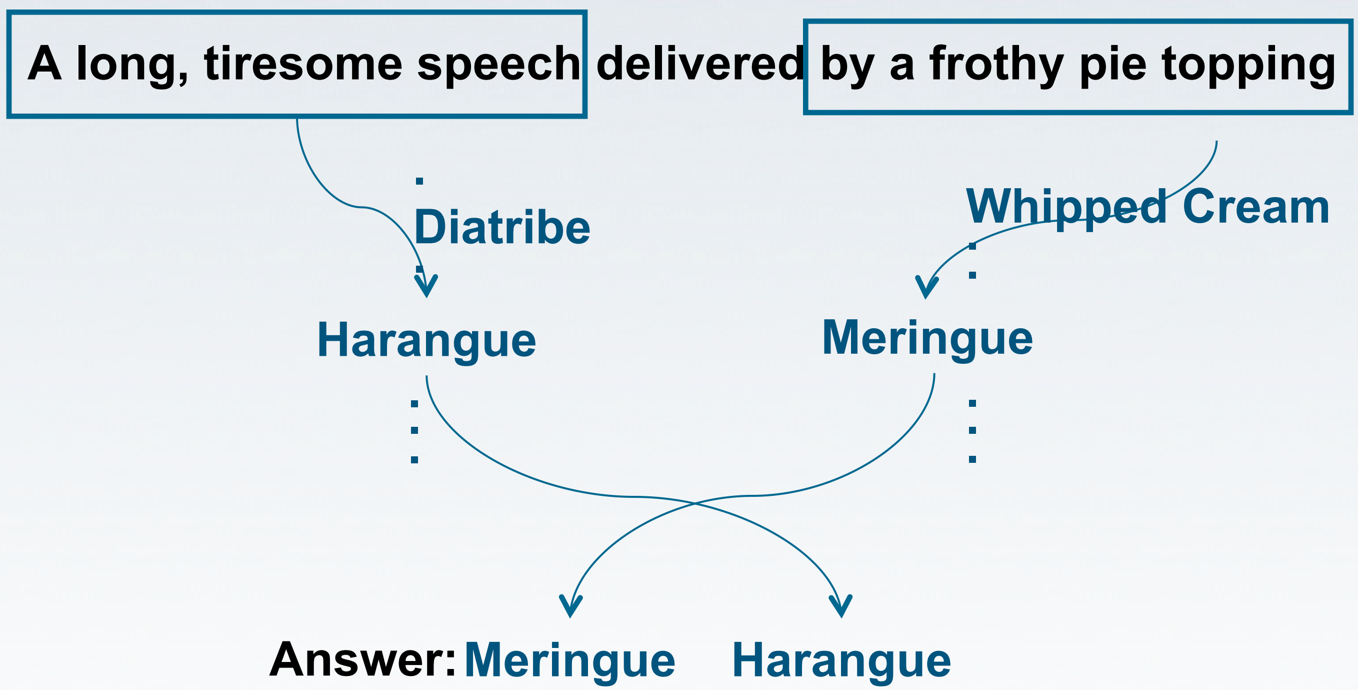
On hearing of the discovery of George Mallory's body, he told reporters he still thinks he was first.

IBM WATSON

Not *Just* for Fun

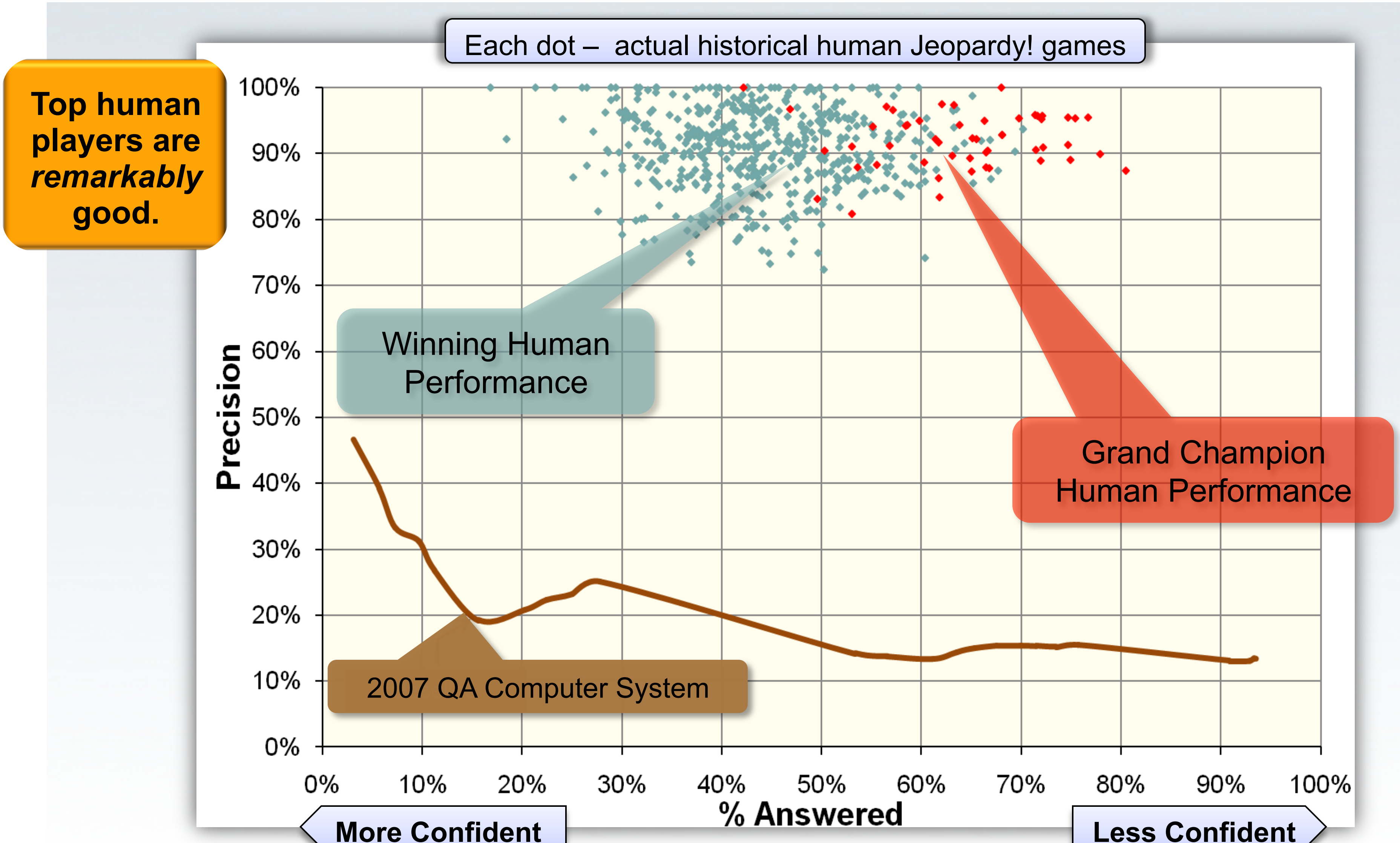
Some Questions require Decomposition and Synthesis

Category: Edible Rhyme Time



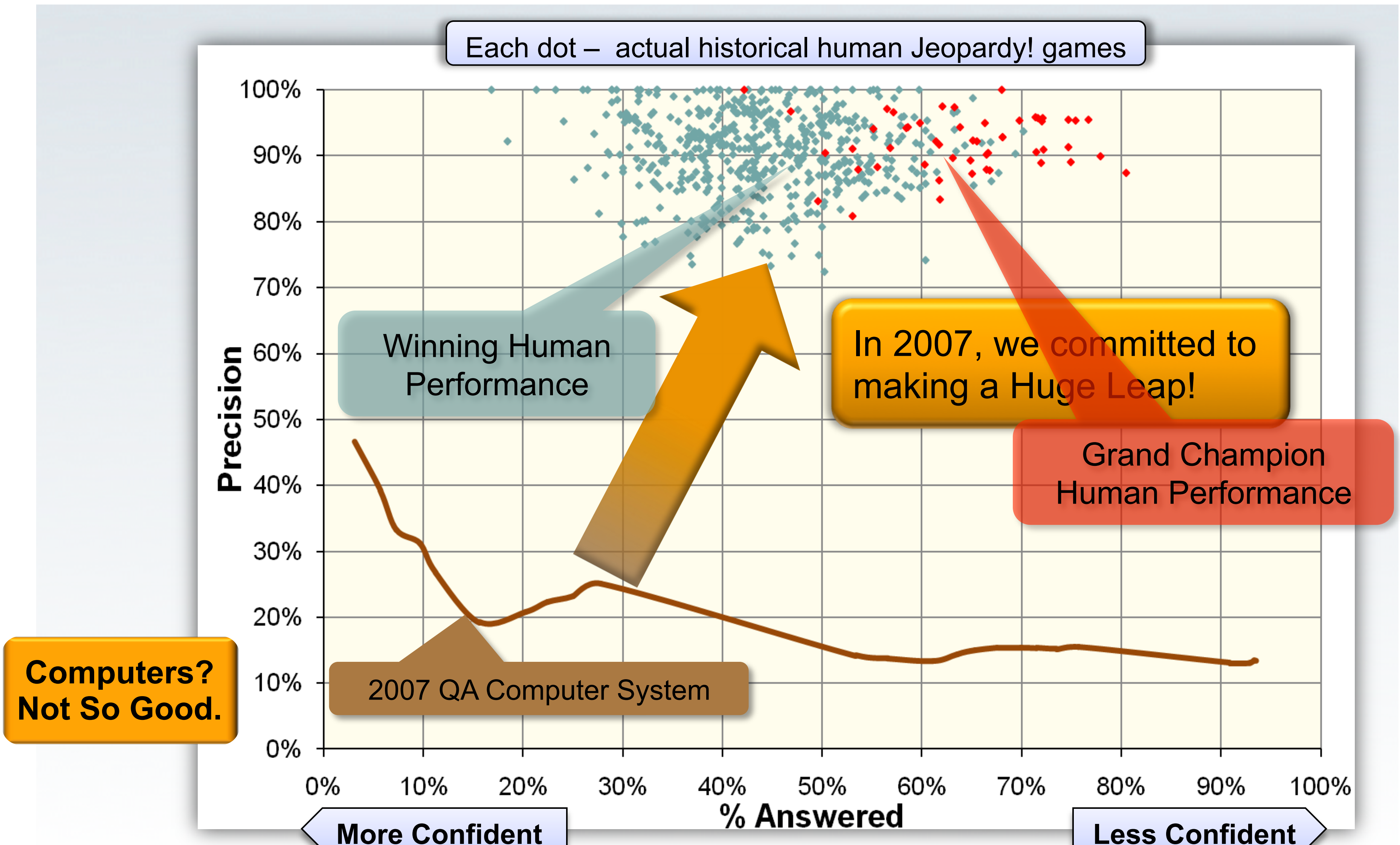
What It Takes to compete against Top Human Jeopardy! Players

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



What It Takes to compete against Top Human Jeopardy! Players

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

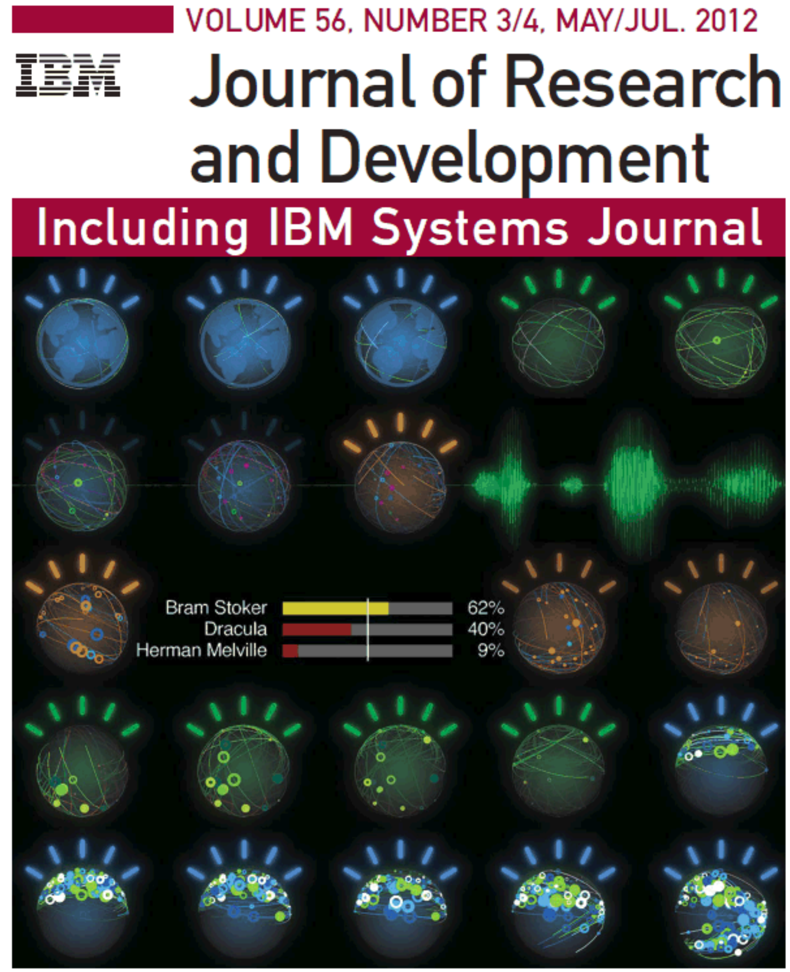


Goals of this course

- Learning about Watson and the Deep QA architecture
- Familiarize with the Watson technology (UIMA)
- Appreciate the role of semantic technology in information systems
- Perform a full research experience in the field of semantic technology

References

- Ferrucci et al., **Building Watson: An Overview of the DeepQA Project,** AI Magazine, 2010
- *Ferrucci et al.,* **Watson: Beyond Jeopardy!,** 2011 RC25270, to appear in Artificial Intelligence Journal.
- Deep QA publications website
  - http://researcher.ibm.com/view_grouppubs.php?grp=2099
- Videos on Watson
  - http://www-03.ibm.com/innovation/us/watson/index.html



This Is Watson

- http://ieeexplore.ieee.org/xpl/tocresult.jsp?reload=true&isnumber=6177717

Class Project

- 100% of grade
  - 30% midterm workshop
  - 70% final workshop
- Goal: a full research experience in the field of semantic technology
- In teams of up to 3-4 people.
  - Teams with at least one advanced grad student encouraged
- One of two types:
  - A research paper, focused on scientific contribution and evaluation
  - A software contribution

Class Project – Research Projects

- State of the art in a specific task (e.g. Word Sense Disambiguation)

- Identify an innovative approach (e.g. using Distributional Semantics)
- Implement system performing the task
  – Experimental prototype (No documentation)
- Evaluate the new approach
  – Find/develop an evaluation benchmark (e.g. Semeval)
  – Evaluate a baseline approach (e.g. most frequent sense)
  – Evaluate the proposed approach
- Present the research project at the student workshop
  – 10 – 20 minutes

- Scope similar to a workshop talk

Class Project – Software Contribution

- More like a software engineering project
  - Possibly using UIMA
- Can be a contribution/improvement of existing Open Source Projects
  - JoBimText
  - Open NLP
  - …
- Scope similar to a small open-source project
- Examples
  - Integrate a web crawler in the JobBimText project and train a Distributional Semantic model from downloaded content

Semantic Technologies in IBM Watson

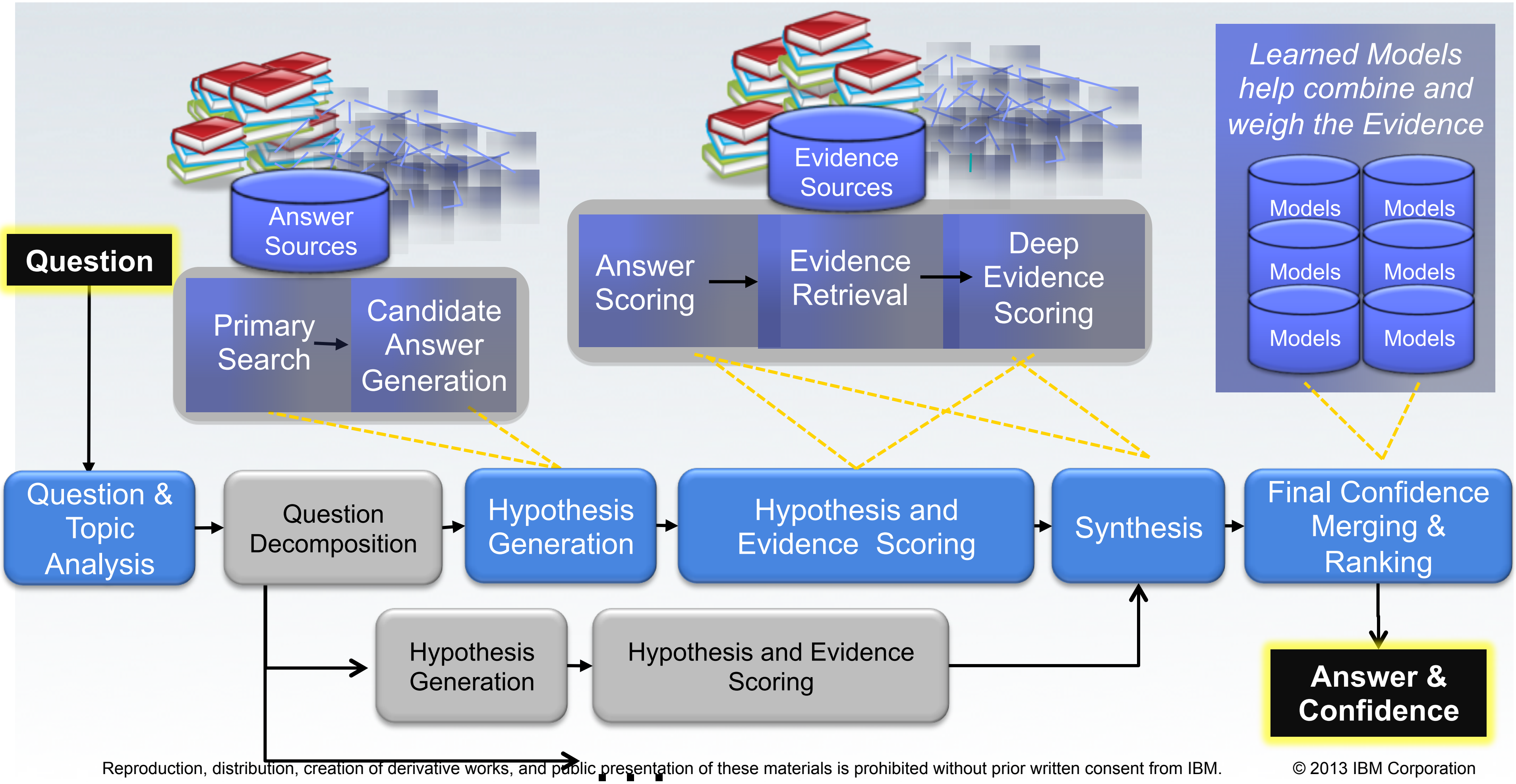
Lesson 2 – The Deep QA architecture (1/2)

Professor: Alfio Massimiliano Gliozzo
TA: Or Biran

DeepQA: The Technology Behind *Watson*

Massively Parallel Probabilistic Evidence-Based Architecture

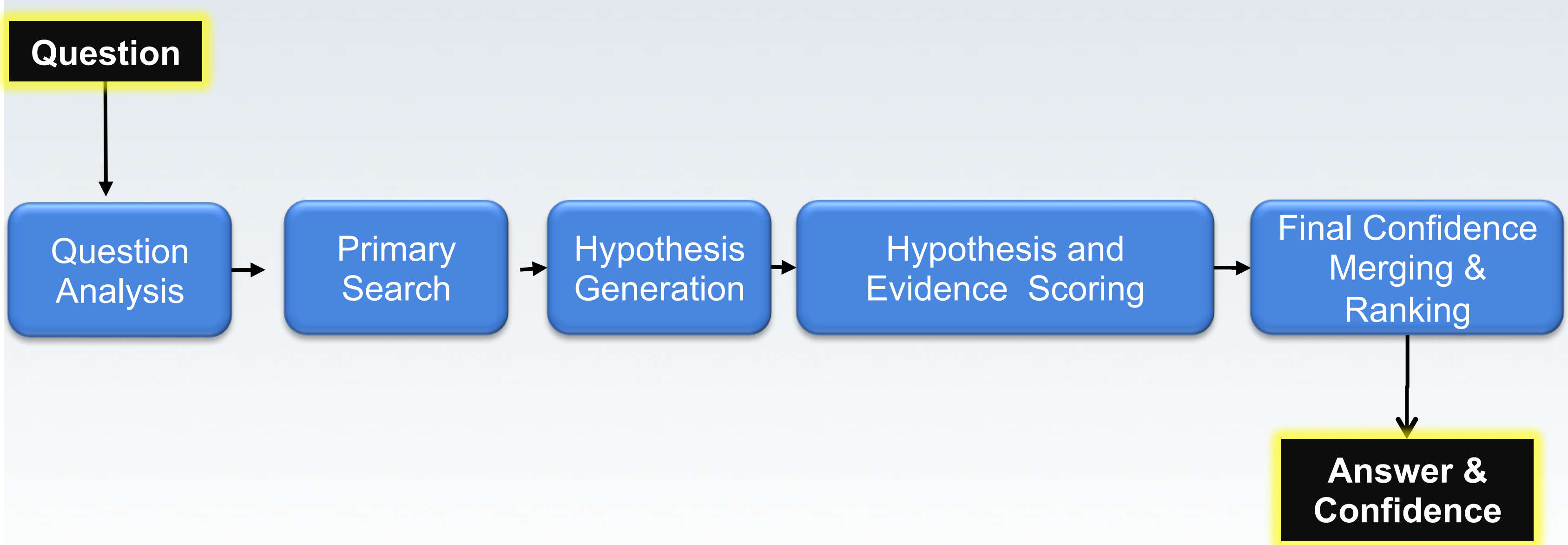
DeepQA *generates and scores many hypotheses using an extensible collection of* ***Natural Language Processing***, ***Machine Learning*** *and* ***Reasoning Algorithms.*** *These gather and weigh evidence over both unstructured and structured content to determine the answer with the best confidence.*



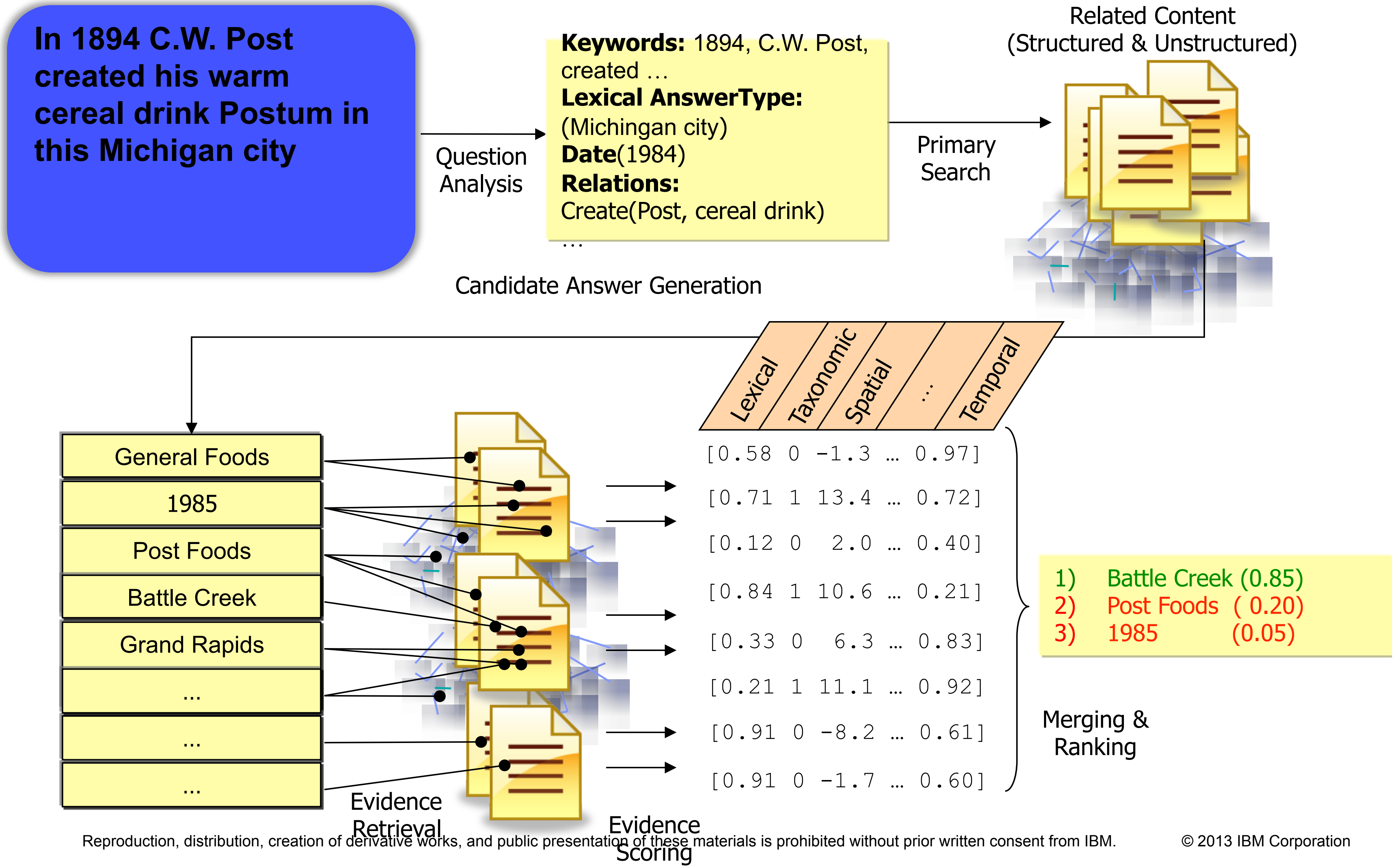
“Minimal” Deep QA Pipeline

Category: MICHIGAN MANIA

Clue: In 1894 C.W. Post created his warm cereal drink Postum in this Michigan city



Example Question



"Minimal" Deep QA Pipeline"

Category: MICHIGAN MANIA

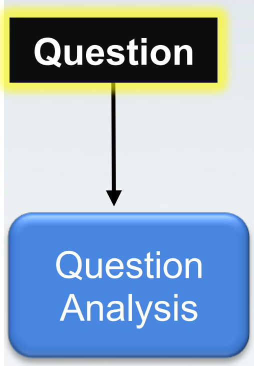
Clue: In 1894 C.W. Post created his warm cereal drink Postum in this Michigan city

Question

“Minimal” Deep QA Pipeline

Category: MICHIGAN MANIA

Clue: In 1894 C.W. Post created his warm cereal drink Postum in this Michigan city



Step 1: Question Analysis

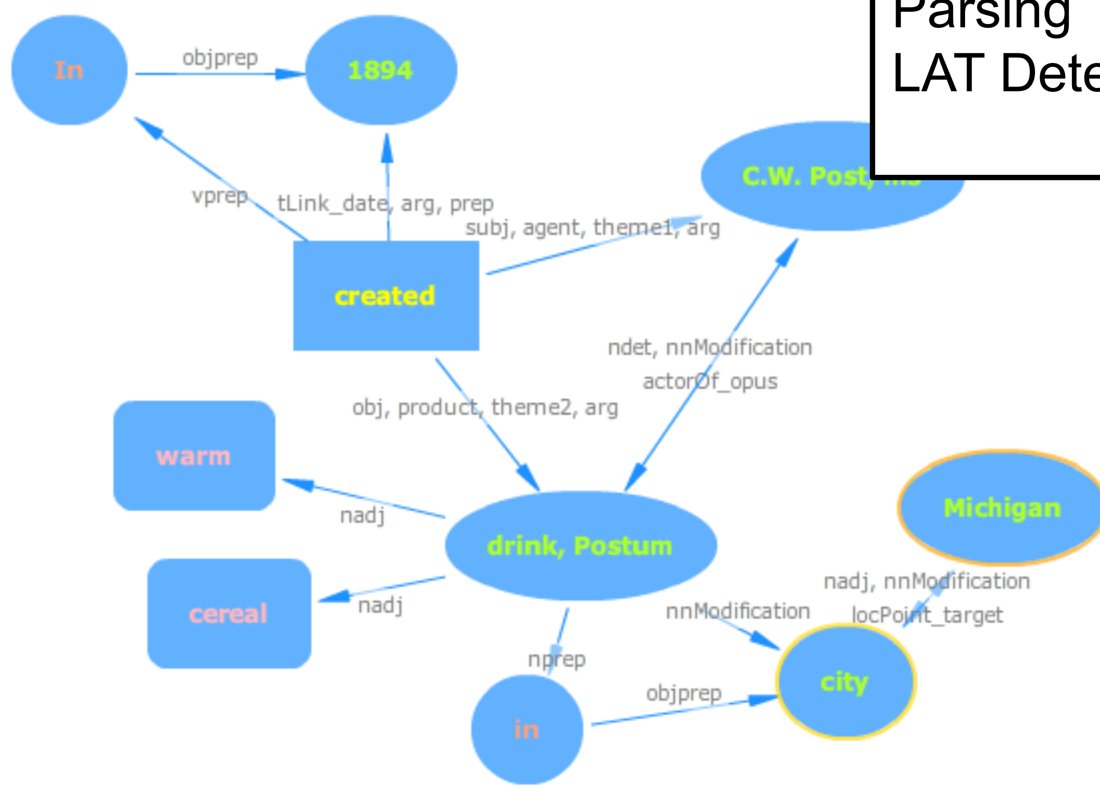
Category: MICHIGAN MANIA

Clue: In 1894 C.W. Post created his warm cereal drink Postum in this Michigan city

Parsing
LAT Detection

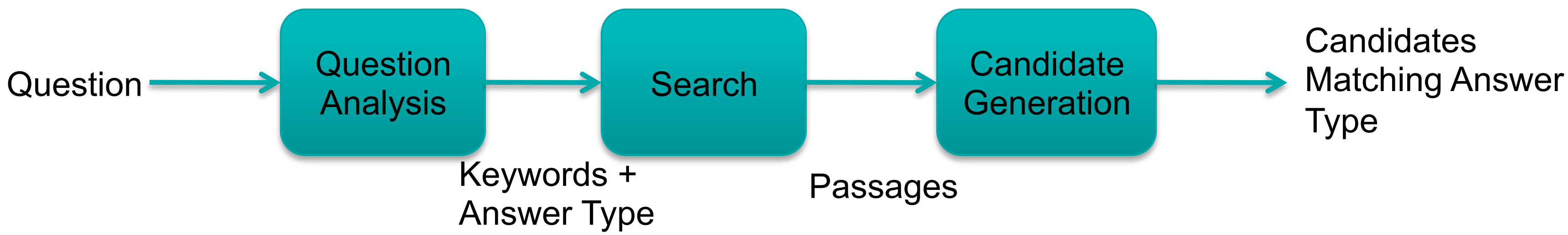
Focus: **this Michigan city**
LAT: Michingan city

Keywords
1894
C.W. Post
created
warm
cereal
drink,
Postum
Michigan
City



Brief History of Question Answering

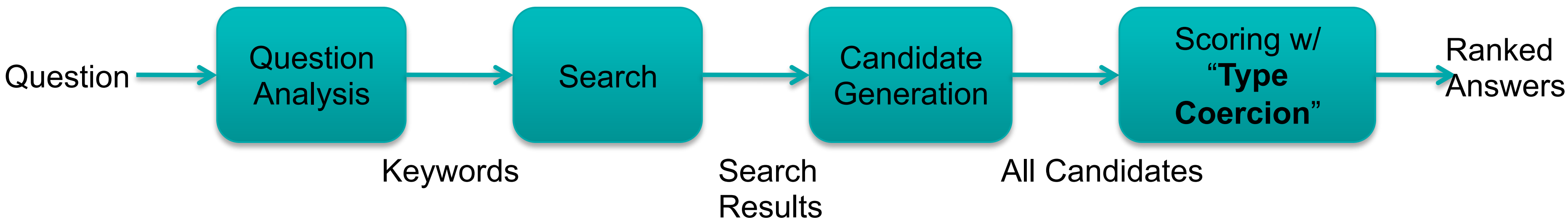
- Organized QA evaluation efforts since 1999: TREC, CLEF, & NTCIR
- Key focus in earlier years: "factoid questions"
  - What is the ***capital*** of Japan?
  - How ***high*** is Mount Everest?
  - Who was the 20th ***president*** of the United States?
- Canonical pipeline architecture for search and candidate generation [Prager et al, 2000, Moldovan et al, 2000, Clarke et al, 2002]



- Prerequisites for adopting a "type-and-generate" approach
  - Possible answer types must fall within a pre-defined type ontology
  - Must have effective strategies for
    - detecting answer type from question
    - identifying instances of given type in text

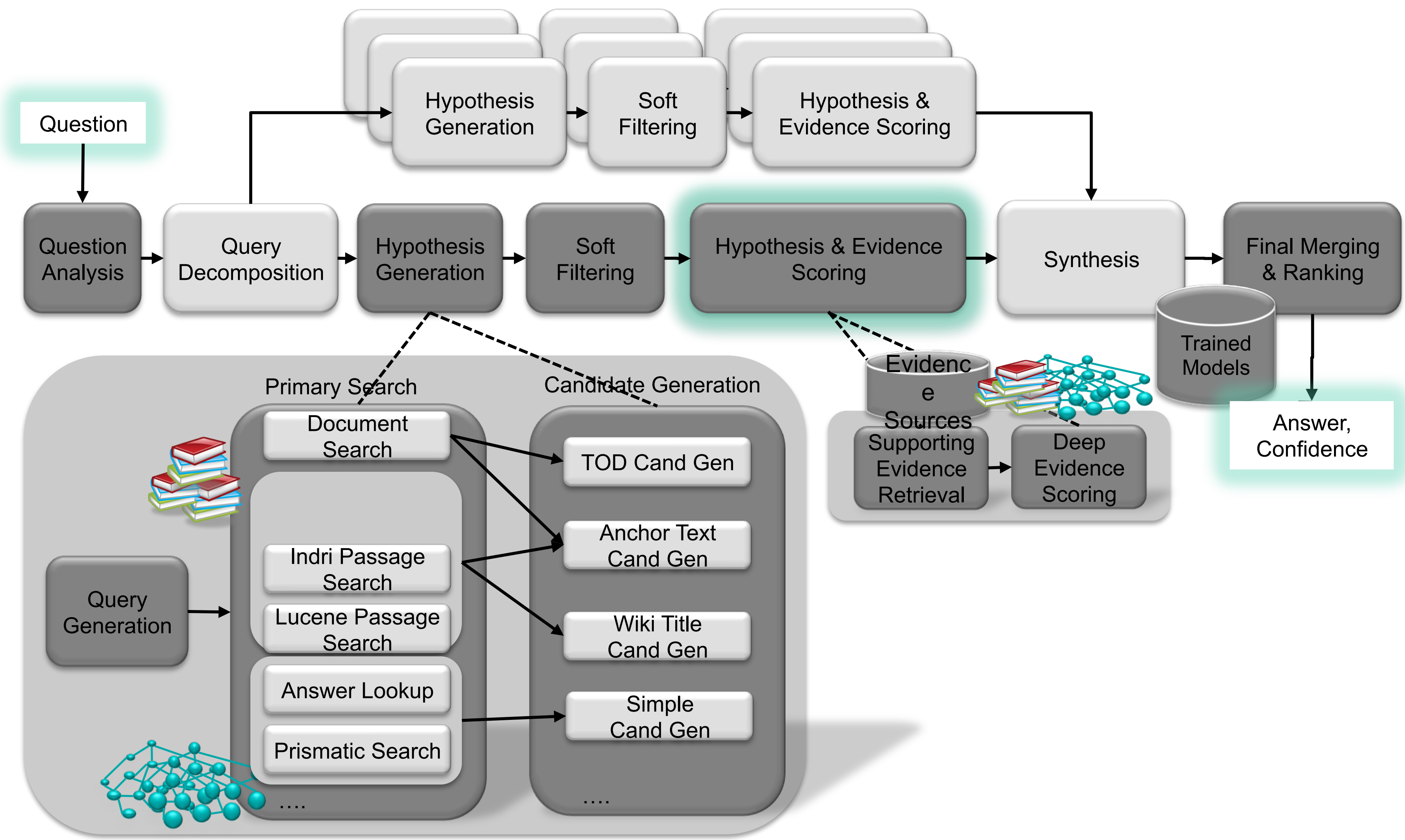
A “Generate-and-Type” Approach to Question Answering

- “Generate-and-Type” QA pipeline [Ferrucci et al, 2010]



- Design principles
  - Cast a wide net during the search phase
  - Generate all plausible candidates from search results
  - Compute type match during scoring
- Challenges addressed in this talk
  - Search strategies for high search recall
  - Candidate generation strategies for high candidate recall
  - How title-oriented linked documents can be leveraged to increase recall

Hypothesis Generation in Watson

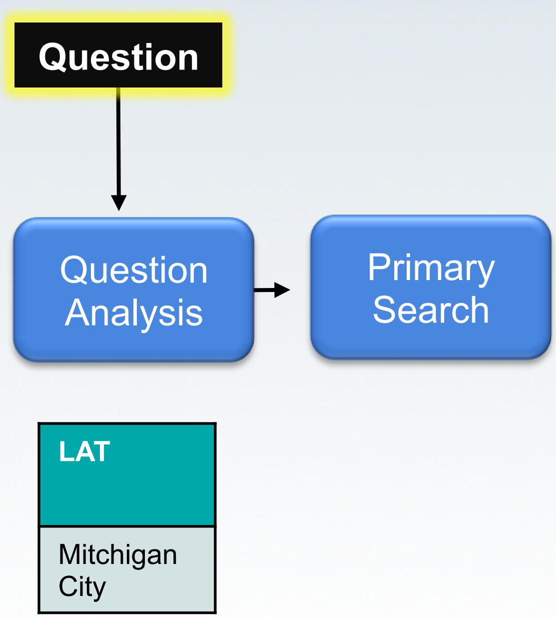


Jennifer Chu-Carroll and James Fan, *Leveraging Wikipedia Characteristics for Search and Candidate Generation in Question Answering*, AAAI 2011

"Minimal" Deep QA Pipeline

Category: MICHIGAN MANIA

Clue: In 1894 C.W. Post created his warm cereal drink Postum in this Michigan city



Step 2: Primary Search

Indri Passage Search

Lucene Passage Search

The keywords (**1894**, **C.W. Post, created, warm, cereal, drink, Postum, Michigan, city**) are used to search over millions of documents to find relevant hits. 55 documents are found, and 30 passages are found.

Document Search Results

Rank	Title
0	General Foods
1	Battle Creek
2	Post Foods
3	Will Keith Kellogg
4	Breakfast Cereal
5	John Harvey Kellogg
6	C. W. Post
7	Kellogg Company
8	Postum
	…

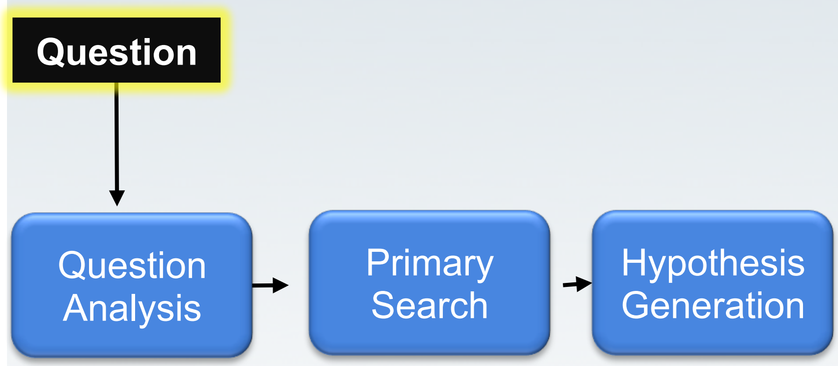
Passage Search Results

Rank	Passage
0	C.W. Post came to the Battle Creek sanitarium to cure his upset stomach. He later created Postum, a cereal-based coffee substitute
1	The caffeine-free beverage mix was created by The Postum Cereal Company founder C. W. Post in 1895 and produced and marketed by Postum Cereal Company as a healthful alternative to coffee
2	1895: In Battle Creek, Michigan, C.W. Post made the first POSTUM , a cereal beverage. Post created GRAPE-NUTS cereal in 1897, and POST TOASTIES corn flakes in 1908
3	1854 C. W. Post (Charles William) was born. He founded the Postum Cereal Co. in 1895 (renamed General Foods Corp. in 1922) to manufacture Postum cereal beverage
4	The company was incorporated in 1922, having developed from the earlier Postum Cereal Co. Ltd., founded by C.W. Post (1854-1914) in 1895 in Battle Creek, Mich. After a number of experiments, Post marketed his first product-the cereal beverage called Postum-in 1895
5	…

"Minimal" Deep QA Pipeline

Category: MICHIGAN MANIA

Clue: In 1894 C.W. Post created his warm cereal drink Postum in this Michigan city



LAT
Mitchigan City

Document Search Results

R	Title
0	General Foods
1	Battle Creek
2	Post Foods
3	Will Keith Kellogg

IBM WATSON

Step 3: Candidate Hypothesis Generation

Category: MICHIGAN MANIA

Clue: In 1894 C.W. Post created his warm cereal drink Postum in this

Candidate Answers (possible answers to the question) are identified in the search results. They are found by looking at document titles (including a variety of title variants and expansions) and possible answers in the text of the documents and passages, such as *named entities*, noun phrases, anchor text, dates, etc. The Candidate Answers are get their first evidence feature scores from their corresponding document search rank and passage search rank.

Candidate Answers	Evidence Feature Scores								
	Doc Rank	Pass Rank							
General Foods	0	1							
Post Foods	2	1							
Battle Creek	1	2							
Will Keith Kellogg	3								
Grand Rapids									
1895		0							

Content Types

Title Oriented Documents (TOD's)

- Documents:
  - Whose *Title* is a concept or named entity – (returned as an answer to a particular question) and
  - Whose *Body* is text that discusses the entity – (provides a rich search context for retrieving titles)
    - Examples are {*Web Pages, Encyclopedias, Dictionaries, etc…*}
  - We format/structure this content into documents we call TODs.
  - A *search index* is created for all TODs.
  - During Question Answering a search query is generated using "keywords" from the question
  - The *body* of these documents is searched - returns the document *title* which is a *candidate answer*

Passage Search Documents

- Documents
  - Standard documents such as journals, magazines, books, encyclopedias, web data.
  - A *Search Index* is created for this content
  - A *Search Query* is generated using "keywords" from the question
  - The *body* of these documents is searched - returns a passage that many contain a *candidate answer*

Structured Content

- Databases or Ontologies
  - Databases that can be searched
  - A search query (SQL, RDF, etc.) is generated from the question – returns a possible *candidate answer*
  - E.g. DBPedia, Freebase

Leveraging Title-Oriented Collections for Search

- Examined relationship between title of an answer-justifying document (AJD) and question/answer pair
- Identified three possible relationships
  - Title of AJD is the answer
    - This country singer was imprisoned for robbery and in 1972 was pardoned by Ronald Reagan.
    - the president of this country in 1995
    - obtain this object, with the motto "Non Sans Droict"

Every Man out of His Humour

Merle Haggard

From Wikipedia, the free encyclopedia

comedy written by English playwright Ben Jonson,

Aleksander Kwaśniewski

From Wikipedia, the free encyclopedia

(Redirected from Aleksander Kwasniewski)

Aleksander Kwaśniewski (Polish pronunciation: [alɛˈksandɛr kfaɕˈɲɛfski] (listen); born 15 November 1954) is a Polish politician who served as the President of Poland from 1995 to 2005. He was born in Białogard, and during communist rule he was active in the Socialist Union of Polish Students and was the Minister for Sport in the communist government in the 1980s. After the fall of communism he became a leader of the left-wing Social Democracy of the Republic of Poland, successor to the former ruling Polish United Workers Party, and a co-founder of the Democratic Left Alliance.

Kwaśniewski was democratically elected president in 1995, defeating the incumbent, Lech Wałęsa. He was re-elected to a second and final term as president in 2000 in a decisive first-

Merle Ronald Haggard (born April 6, 1937) is an American country music singer, guita

Kwaśniewski identifies himself as an atheist.[1][2][3][4]

first two songs Frizzell allowed Haggard to sing at the concert. The audience enjoyed Haggard and he began working on a full-time music career. After he had earned a local reputation, Haggard's money problems caught up with him. He was arrested for attempting to rob a Bakersfield tavern in 1957[3] and was sent to the San Quentin state prison for three years.

While in prison, Haggard ran a gambling and brewing racket from his cell. During a time of solitary confinement, he encountered an

Aleksander Kwaśniewski (Polish pronunciation: [alɛˈksandɛr kfaɕˈɲɛfski] (listen); born 15 November 1954) is a Polish politician who served as the President of Poland from 1995 to 2005

whose crest features a " boar without a head, rampant - A boar without a head, that's very rare!" and the mot

The motto of Shakespeare's recently-granted family coat of arms was *Non Sans Droit*, "not without right."

The motto of Shakespeare's recently-granted family coat of arms was *Non Sans Droit*,

Wikipedia Analysis

- High domain coverage: Wikipedia document titles cover
  - ~95% of Jeopardy! answers
  - ~98% of TREC answers (excluding numeric answers)
- Wikipedia contains “Title-Oriented Documents”
  - Document title is an entity/concept
  - Document content describes salient aspects of the title entity
  - Characteristics of general-purpose encyclopedic collections
- Wikipedia documents contain human generated meta data
  - Anchor texts represent salient concepts with respect to the document title
  - Characteristic of web and/or cross-referenced documents

Candidate Generation using Document Meta Data

- Generate “plausible” candidates from search results without type information
- Humans have intuitions about what “plausible” candidates are
  - Neapolitan pizzas are made with ingredients like San Marzano tomatoes, which grow on the volcanic plains south of Mount Vesuvius and Mozzarella di Bufala Campana, made with milk from water buffalo raised in the marshlands of Campania and Lazio
- Wikipedia documents contain meta data that represent salient concepts in a document
  - Anchor texts: links to other relevant documents
    - Mozzarella di Bufala Campana -> Buffalo Mozzarella
  - Redirects: links to current document
    - Neapolitan pizza -> pizza

Experimental Evaluation

- Corpus
  – August 2010 crawl of English Wikipedia,, indexed with Indri
- Data
  – Jeopardy!: randomly selected set of 2000 Jeopardy! questions
  – TREC: 575 non-numeric factoid questions from TRECs 11 & 12
- Evaluation metric
  – Binary recall (BR): % of questions where the correct answer is included in the result list
  – Search BR: % of questions where the search results contain the correct answer
  – Candidate BR: % of questions where the correct answer is produced as a candidate answer
- Objective: maximize search and candidate binary recall

Experimental Results

Search Result	Candidate Generation Strategy	Jeopardy! Binary Recall		TREC Binary Recall	
		Search	Candidate	Search	Candidate
Document	Doc Title + Anchor Text	66.5%	63.6%	58.3%	55.7%
Passage	Anchor Text	71.0%	64.8%	71.7%	64.4%
All	All	81.4%	75.3%	80.5%	73.0%

- End-to-end QA accuracy
  - Jeopardy!: 59.3%
  - TREC: 49.4%

Candidate Generation Evaluation

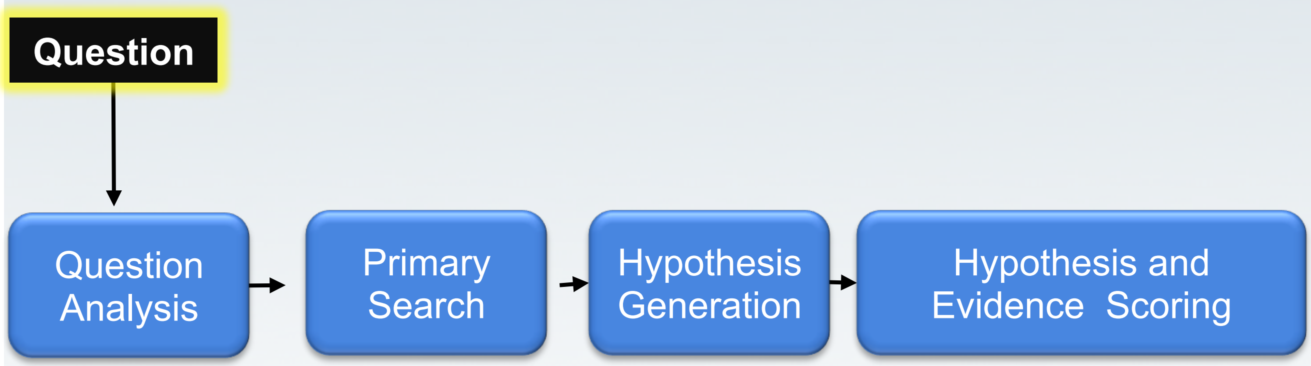
	Document Search	Passage Search	Answer Lookup	Prismatic Search	All
% Questions in Active Subset	99.07%	100.00%	13.64%	43.75%	100%
# Answers/ Active Question	90.55	162.47	15.11	11.57	216.53
Binary Recall*	74.43%	79.40%	3.53%	8.31%	87.17%
% Unique	7.24%	12.05%	0.03%	0.18%	-
Accuracy	62.65%	62.74%	2.18%	5.65%	71.32%

***Candidate binary recall :** percentage of questions for which the correct answer is generated as a candidate answer

"Minimal" Deep QA Pipeline

Category: MICHIGAN MANIA

Clue: In 1894 C.W. Post created his warm cereal drink Postum in this Michigan city



LAT
Mitchigan City

Document Search Results	
R	Title
0	General Foods
1	Battle Creek
2	Post Foods
3	Will Keith Kellogg

Candidate Answers
General Foods
Post Foods
Battle Creek

Step 4A: Hypothesis Scoring (Answer Scoring Features)

Category: MICHIGAN MANIA
Clue: In 1894 C.W. Post created his warm cereal drink Postum in this Michigan city

Other Answer Scorers can be applied depending on different relations or constraints detected in the question. For example, this question focus with modifiers is "Michigan city." Watson can detect this as a geospatial relation that indicates the correct answer must be a city spatially located within the sate of Michigan.

Tycor
Temporal
Spatial
Popularity
. . .

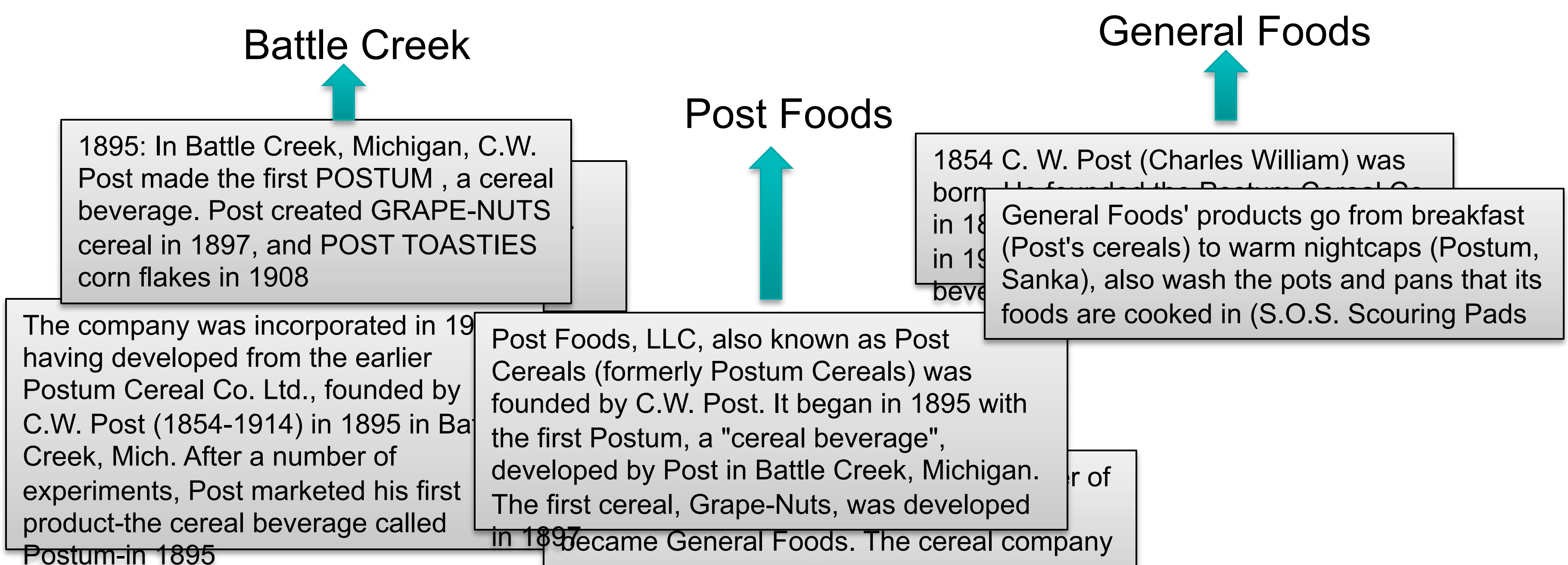
Candidate Answers	Evidence Feature Scores (Answer Scoring + Passage Scoring)								
	Doc Rank	Pass Rank	Ty Cor	Geo					
General Foods	0	1	0.1	0					
Post Foods	2	1	0.1	0					
Battle Creek	1	2	0.8	1					
Will Keith Kellogg	3		0.1	0					
Grand Rapids			0.9	1					
1895		0	0.0	0					

Step 4B: Hypothesis Scoring (Passage Scoring Features)

Category: MICHIGAN MANIA

Clue: In 1894 C.W. Post created his warm cereal drink Postum in this Michigan city

In Deep Evidence Scoring, Watson retrieves evidence for each candidate answer, then evaluates the evidence using a large number of deep evidence scoring analytics. The evidence for a candidate answer may come from the original document or passage where the candidate answer was generated, or it may come from an evidence retrieval search performed by taking the keyword search query from Step 2, replacing the focus terms with the candidate answer, and retrieving the relevant passages that are found. The passages, or “context” in which the candidate answer occurs are evaluated as evidence to support or refute the candidate answer as the correct answer for the question.



IBM WATSON

DeepQA developers work under the hood with many fine-grained features corresponding to many evidence discovery and scoring algorithms.

Question#95506

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

Exp Name / Answers	Final Score	Correct?
From week52		
Argentina	0.332	yes
Bolivia	0.121	no
Peru	0.078	no
Russia	0.056	no
Canada	0.026	no
South America	0.020	no
Zimbabwe	0.019	no
Argentine Republic	0.011	no
Andes	0.010	no
Kazakhstan	0.010	no
Morocco	0.009	no
The Andes Mountains	0.009	no
Andes Mountains	0.008	no
Uruguay	0.006	no
The Andes	0.005	no
Spain	0.005	no
Argentine	0.004	no
Republic of Chile	0.004	no
Atacama	0.002	no
Antarctica	0.002	no
Landlocked country	0.002	no
Pacific Ocean	0.002	no
United States	0.002	no
Argentina-Chile relations	0.002	no

Question#95506

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

Diff [Argentina , Bolivia]

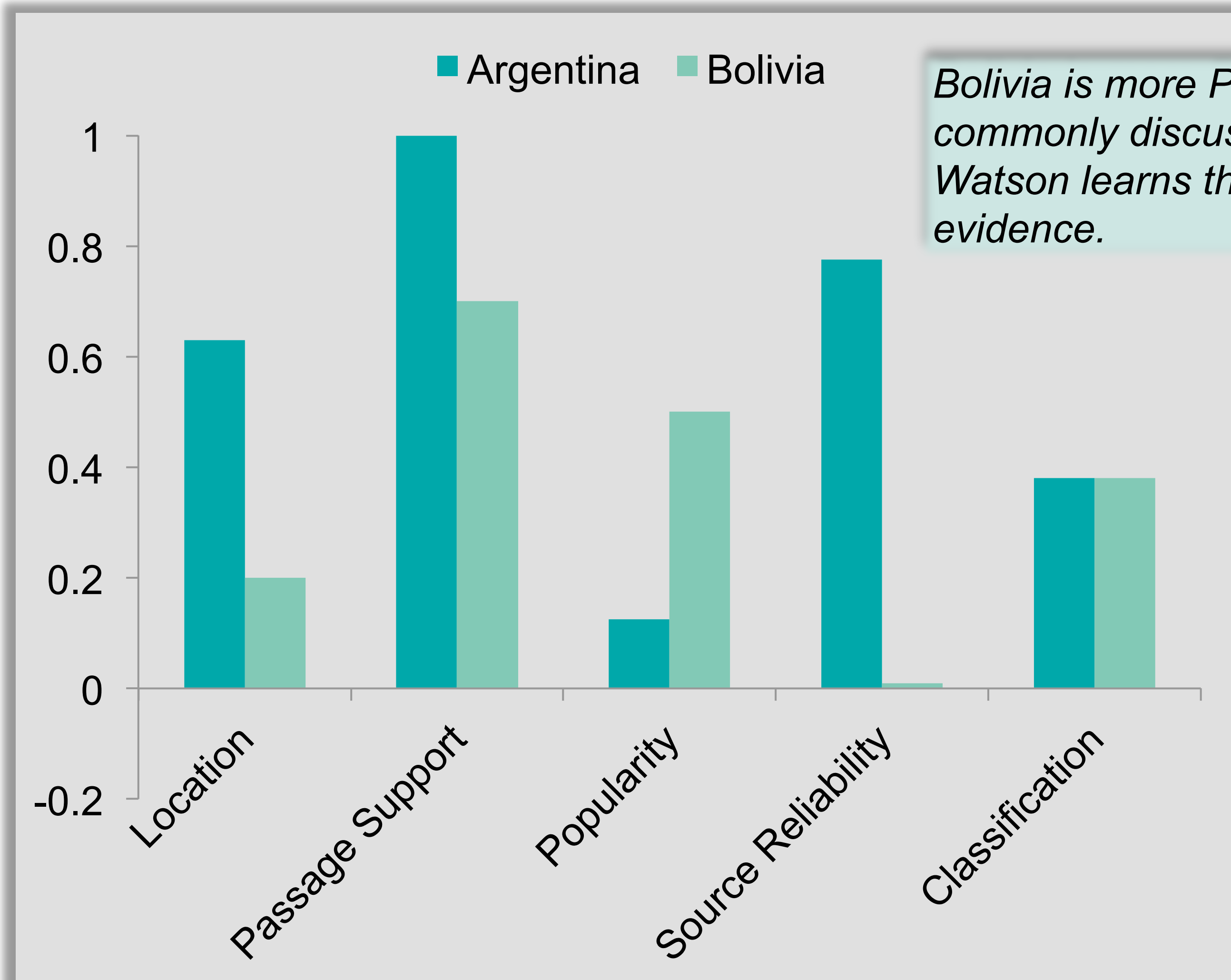
Compare "Argentina" and "Bolivia"

Feature Groups/Features	Diff Value Bar	Diff Value	Argentina	Bolivia	Headroom
Experiment Label			week52	week52	
Final Score			0.332	0.121	
Weighted Features Sum		-0.104	0.572	0.676	
SpatialDistance	Argentina Bolivia	-0.020	0.007	0.028	n/a
SpatialDistance-Missing		-0.224	0	0.224	n/a
SpatialDistance-Std		0.045	0.045	-0.000	n/a
SpatialRelationSat-Std		0.590	0.578	-0.013	n/a
IndriDocumentEngine_RANK		0.593	-0.137	-0.730	n/a
IndriDocumen...ne_RANK-Std		-0.369	-0.578	-0.208	n/a
IndriDocumentEngine_SCORE		-0.013	0.439	0.453	n/a
IndriDocume...e_SCORE-Std		0.267	0.410	0.143	n/a
FractionOfTitleCovered		0.300	1.173	0.873	n/a
FractionOfTitleCovered-Std		-0.227	-0.436	-0.209	n/a
DocTermMatchScore		0.003	0.008	0.006	n/a
DocTermMatchScore-Std		0.269	0.356	0.086	n/a
SkipBigramScore		-0.037	-0.243	-0.206	n/a
SkipBigramScore-Std		0.229	1.378	1.149	n/a
LFACS_Pass...ALLOW+DEEP		-0.089	0	0.089	n/a
LFACS_Passa...OW+DEEP-Std		-0.560	-0.018	0.543	n/a
TextualAlignment		-0.004	0.013	0.017	n/a
TextualAlignment-Std		-0.242	0.433	0.675	n/a
PassageTermMatch		-0.032	0.750	0.782	n/a
PassageTermMatc		-0.056	0.760	0.816	n/a

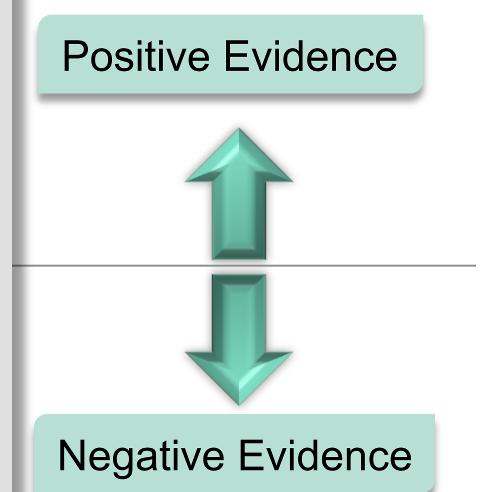
These fine-grained features are aggregated to produce higher-level Evidence Dimensions (Location, Temporal, Source Reliability, Popularity etc)

Grouping features to produce Evidence Profiles

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



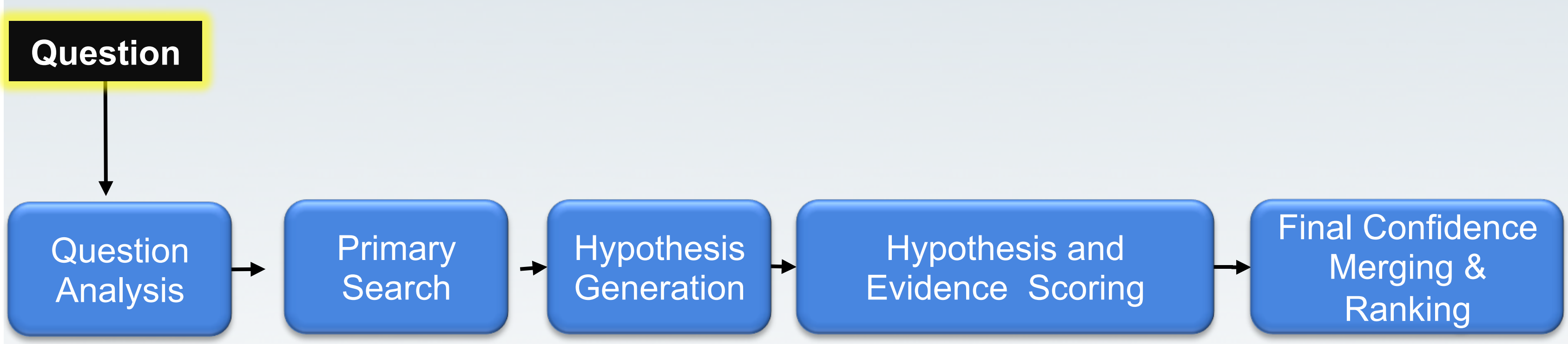
Bolivia is more Popular due to a commonly discussed border dispute. But Watson learns that Argentina has better evidence.



"Minimal" Deep QA Pipeline

Category: MICHIGAN MANIA

Clue: In 1894 C.W. Post created his warm cereal drink Postum in this Michigan city



LAT
Mitchigan City

Document Search Results	
R	**Title**
0	General Foods
1	Battle Creek
2	Post Foods
3	Will Keith Kellogg

Candidate Answers
General Foods
Post Foods
Battle Creek

Evidence Features	
Ty Cor	Geo
0.1	0
0.1	0
0.8	1

Step 6: Merging Candidate Answers and Scoring the Confidence

Category: MICHIGAN MANIA

Clue: In 1894 C.W. Post created his warm cereal drink Postum in this

In the final processing step, Watson detects variants of the same answer and merges their feature scores together. Watson then computes the final confidence scores for the candidate answers by applying a series of Machine Learning models that weight all of the feature scores to produce the final confidence scores.

Candidate Answers	Evidence Feature Scores						
	Doc Rank	Pass Rank	Ty Cor	Geo	LFACS	Term Match	Temp-oral
General Foods	0	1	0.1	0	0.2	22	1
Post Foods	2	1	0.1	0	0.4	41	1
Battle Creek	1	2	0.8	1	0.5	30	0.9
Will Keith Kellogg	3		0.1	0	0	23	0.5
Grand Rapids			0.9	1	0	10	0.5
1895		0	0.0	0	0	21	0.6

Machine Learning Model Application

Correct Answer

Final Answers	Confi-dence
Battle Creek	0.946
Post Foods	0.152
1895	0.040
Grand Rapids	0.033
General Foods	0.014

“Minimal” Deep QA Pipeline

Category: MICHIGAN MANIA

Clue: In 1894 C.W. Post created his warm cereal drink Postum in this Michigan city

Question

Battle Creek

Question Analysis → Primary Search → Hypothesis Generation → Hypothesis and Evidence Scoring → Final Confidence Merging & Ranking

LAT
Mitchigan City

Document Search Results	
R	**Title**
0	General Foods
1	Battle Creek
2	Post Foods
3	Will Keith Kellogg

Candidate Answers
General Foods
Post Foods
Battle Creek

Evidence Features	
Ty Cor	Geo
0.1	0
0.1	0
0.8	1

Final Answers	**Confi-dence**
Battle Creek	0.946
Post Foods	0.152
1895	0.040

Semantic Technologies in IBM Watson™

Lesson 3 – The Deep QA Architecture (2/2)

Professor: Alfio Massimiliano Gliozzo
TA: Or Biran

Outline

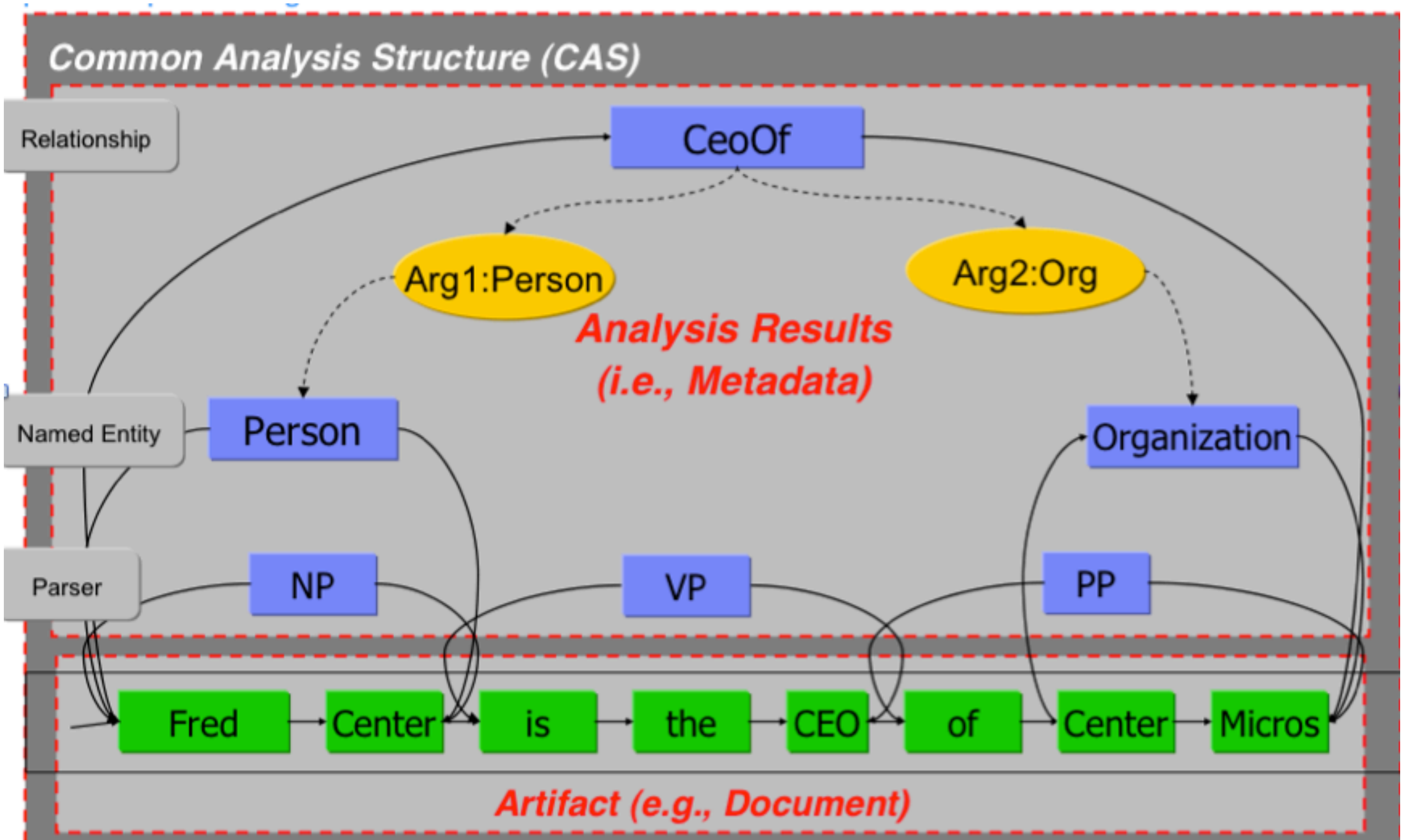
- Unstructured Information Management Architecture (UIMA)
- The Development Cycle
- Why Semantics?

Outline

- Unstructured Information Management Architecture (UIMA)
- The Development Cycle
- Why Semantics?

UIMA : Unstructured Information Management Architecture

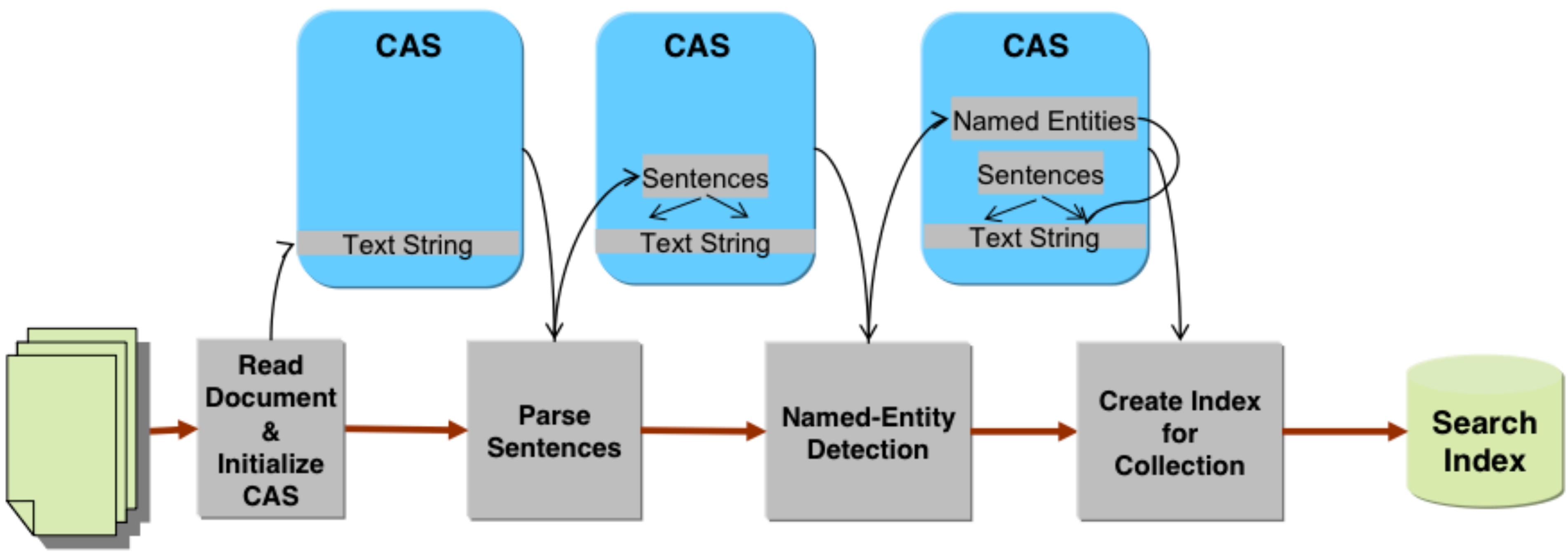
- **An integrating framework and architecture, supporting**
  - Interoperability of Analytics over unstructured data
  - Scaleout
  - Open Source Apache project



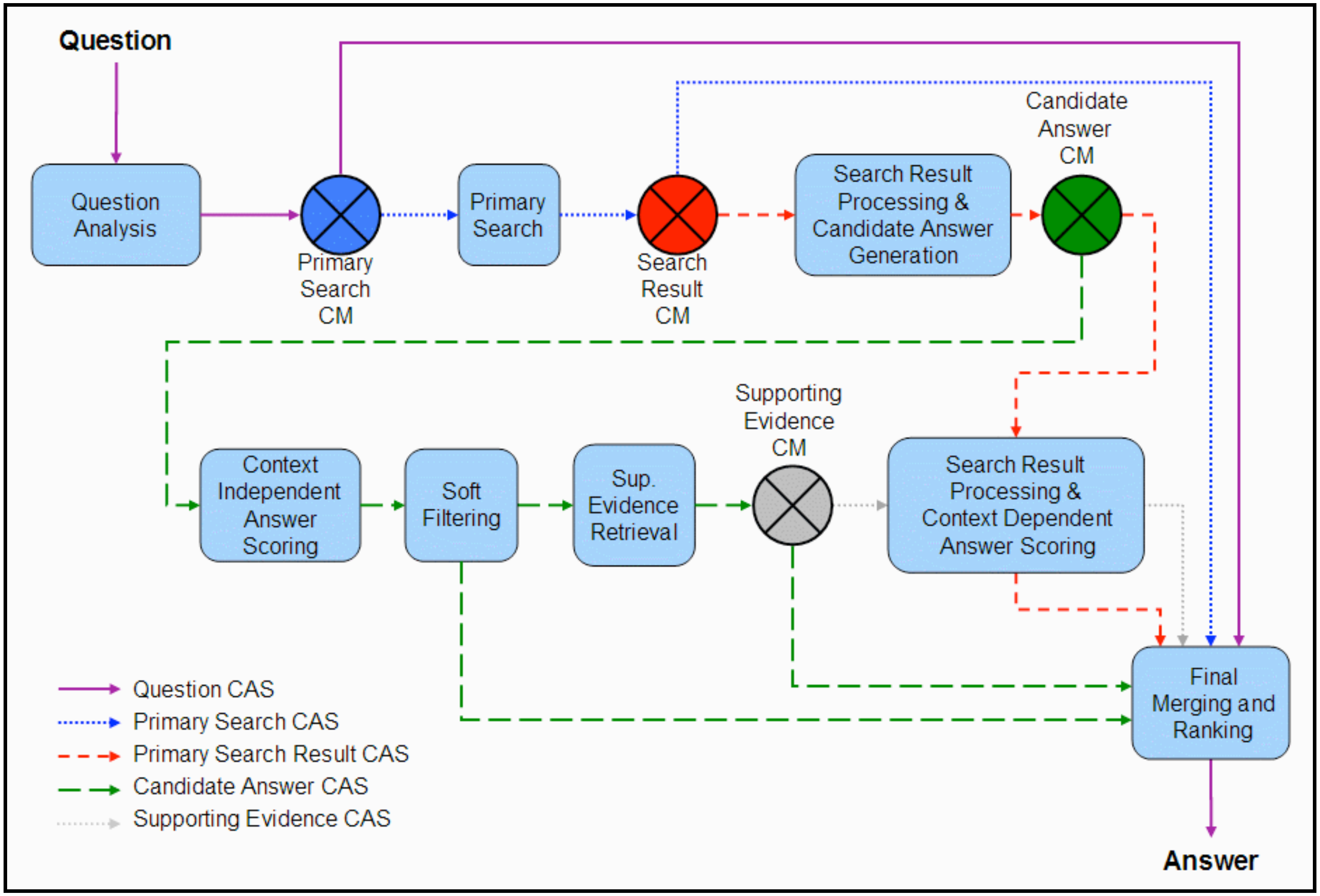
Basic Concepts

- **Type System:** A declarative data model that defines the **Types** (classes) and their **Features** (properties). An instantiated Type is called a **Feature Structure**.
- **Common Analysis Structure (CAS):** A unit of work in UIMA which is passed from one component to another. A CAS encapsulates Feature Structures which are organized into Views.
- **Annotator:** Receives an input CAS, does analysis of its contents and typically modifies it (e.g. adds, deletes or modifies Feature Structures).
- **Flow Controller:** Determines how CASes move among a collection of UIMA components.
- **CAS Multiplier:** create children of the input CAS containing just the data needed for specific subtasks
  - Allows concurrent processing
  - Minimizes the overhead of transferring data on other machines.

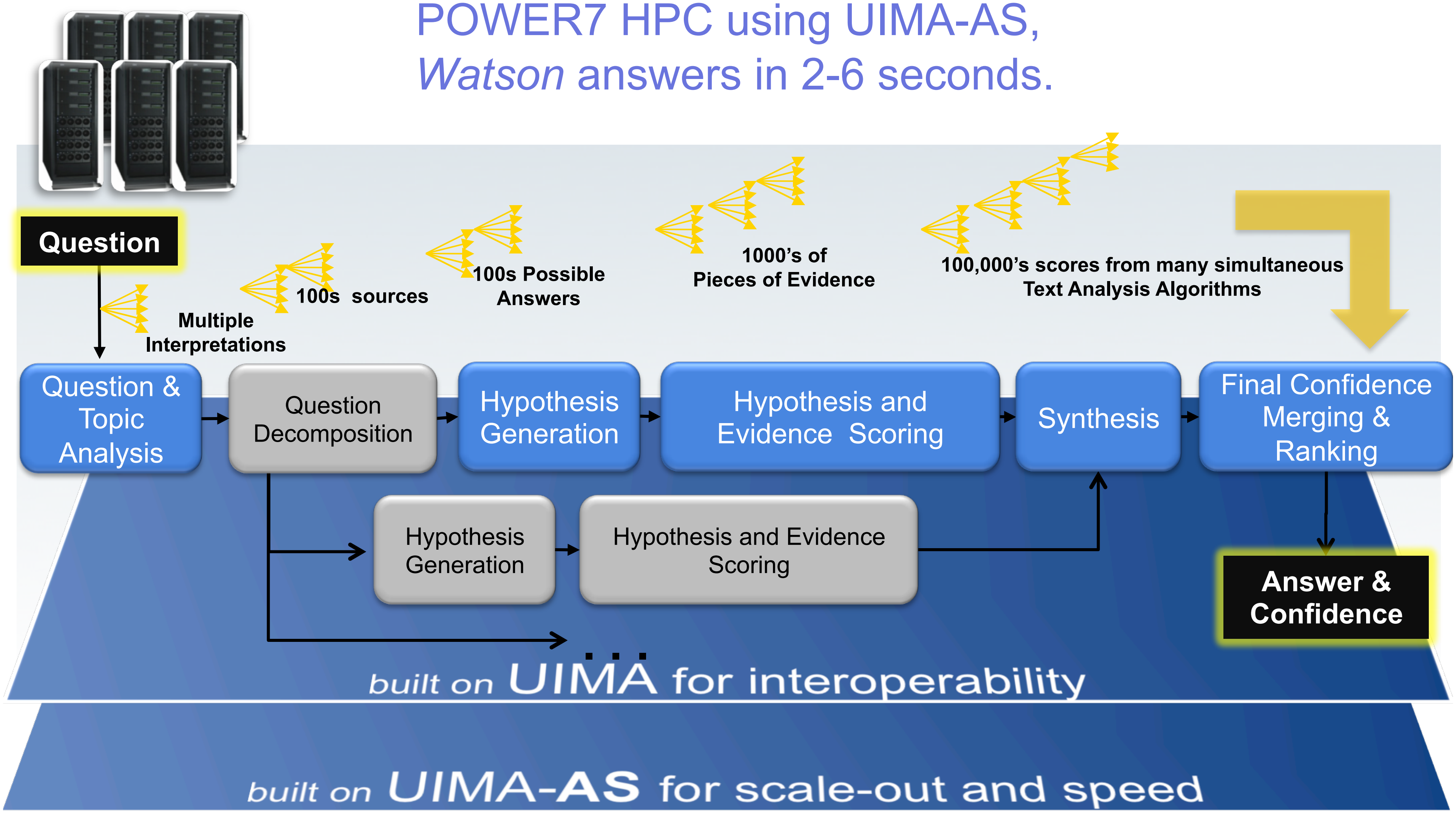
Simple UIMA Processing Pipeline



Deep QA UIMA architecture



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.



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

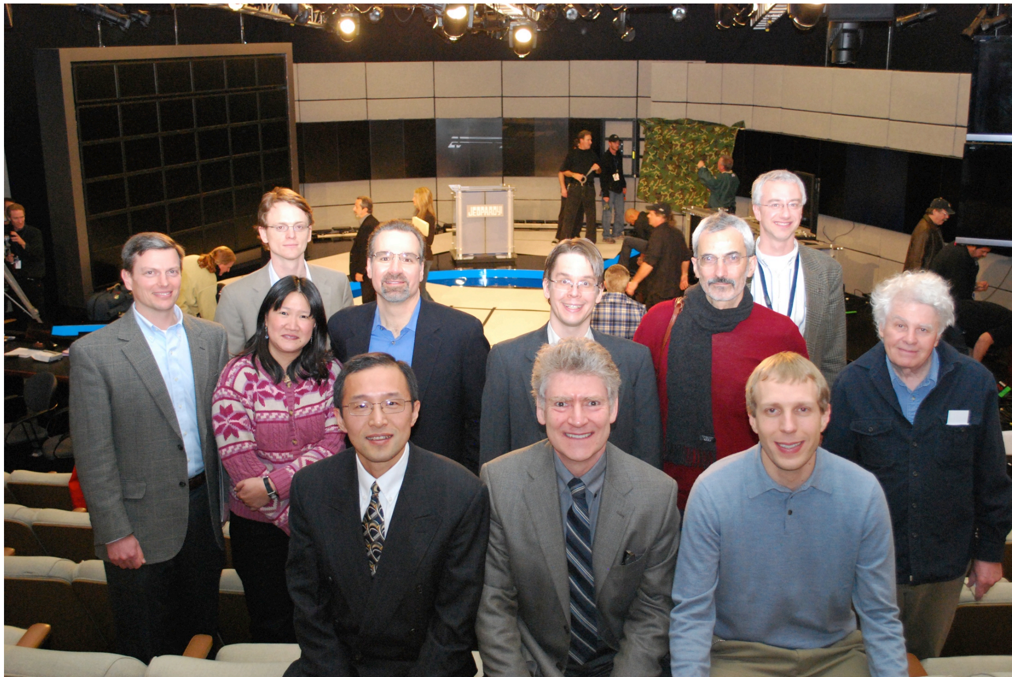
[1] Note that the Power 750 featuring POWER7 is a commercially available server that runs AIX, IBM i and Linux and has been in market since Feb 2010

Outline

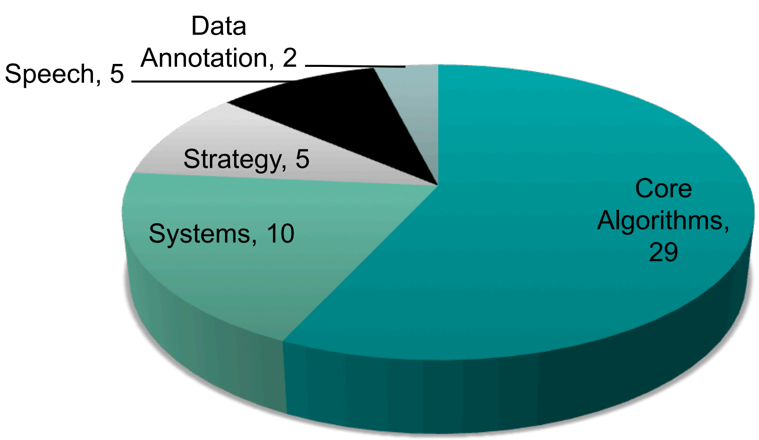
- Unstructured Information Management Architecture (UIMA)
- The Development Cycle
- Why Semantics?

The Core Team

IBM WATSON

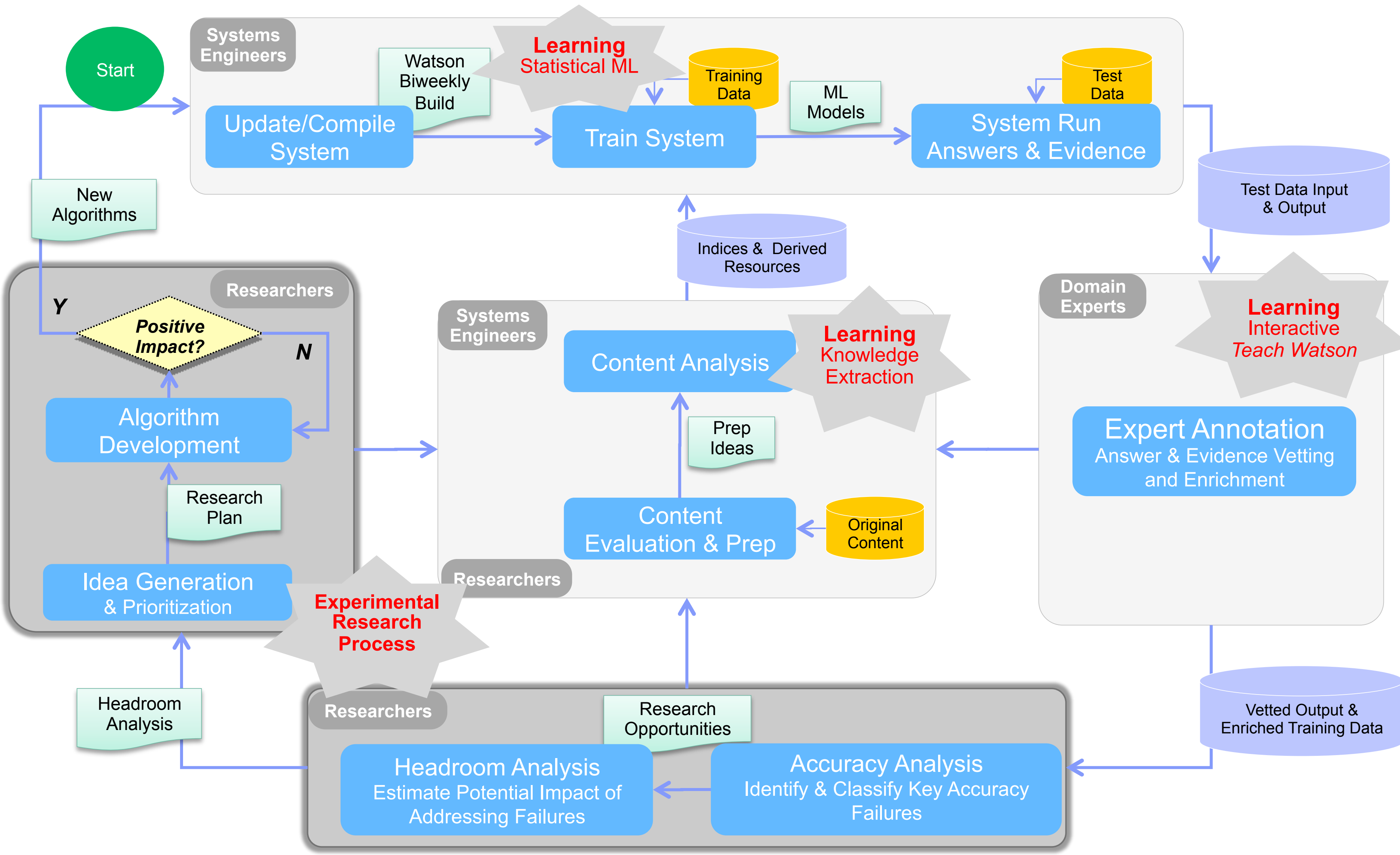


4 years
12 -> 30 people



AdaptWatson Process Step

Continuous Learning: Training and adapting Watson to new domains



Development Tools

- Data Repository
  - Data resources include question sets, corpora, indices, models, annotated data
  - Support for resource documentation, meta-data, and versioning
  - Web browsing or programmatic access via a URI
- Job Scheduling and Management Tool
  - Allocate Machines to Users' Jobs
  - Load Balance User's Job Across Machines
  - Report Errors
- Error Analysis Tool
  - Browse Results of an Experiment
  - Compare Results between Experiments
- Feature Analysis Tool
  - View "Evidence Profiles" for Answers
  - Find out which features contribute to an answer's (mis)ranking

WEAT: Watson Error Analysis Tool

WEAT: All Experiments - Mozilla Firefox

File Edit View History Bookmarks Tools Help

http://bluej558.watson.ibm.com:8080/bjea/report?reportName=experiments&expDate.lw=&expDate.uw=&expUser=&matchString=t20

Most Visited | IBM Business Transfor... | IBM Standard Softwar... | IT Help Central | Join World Community... | Windows Marketplace

WEAT: All Experiments

WEAT Home Views Upload Answer Key Editor

Session Logon Help

Private WEAT (bluej558)
(bluej580, bluej587, bluej559, bluej557)

All Experiments

(14 msecs; 1 concurrent requests)

Refine View

Experiments Date (MM-DD-YYYY) From

To

Experiments User

Description includes this string: t20

Check this box to display ALL Questions: ☐

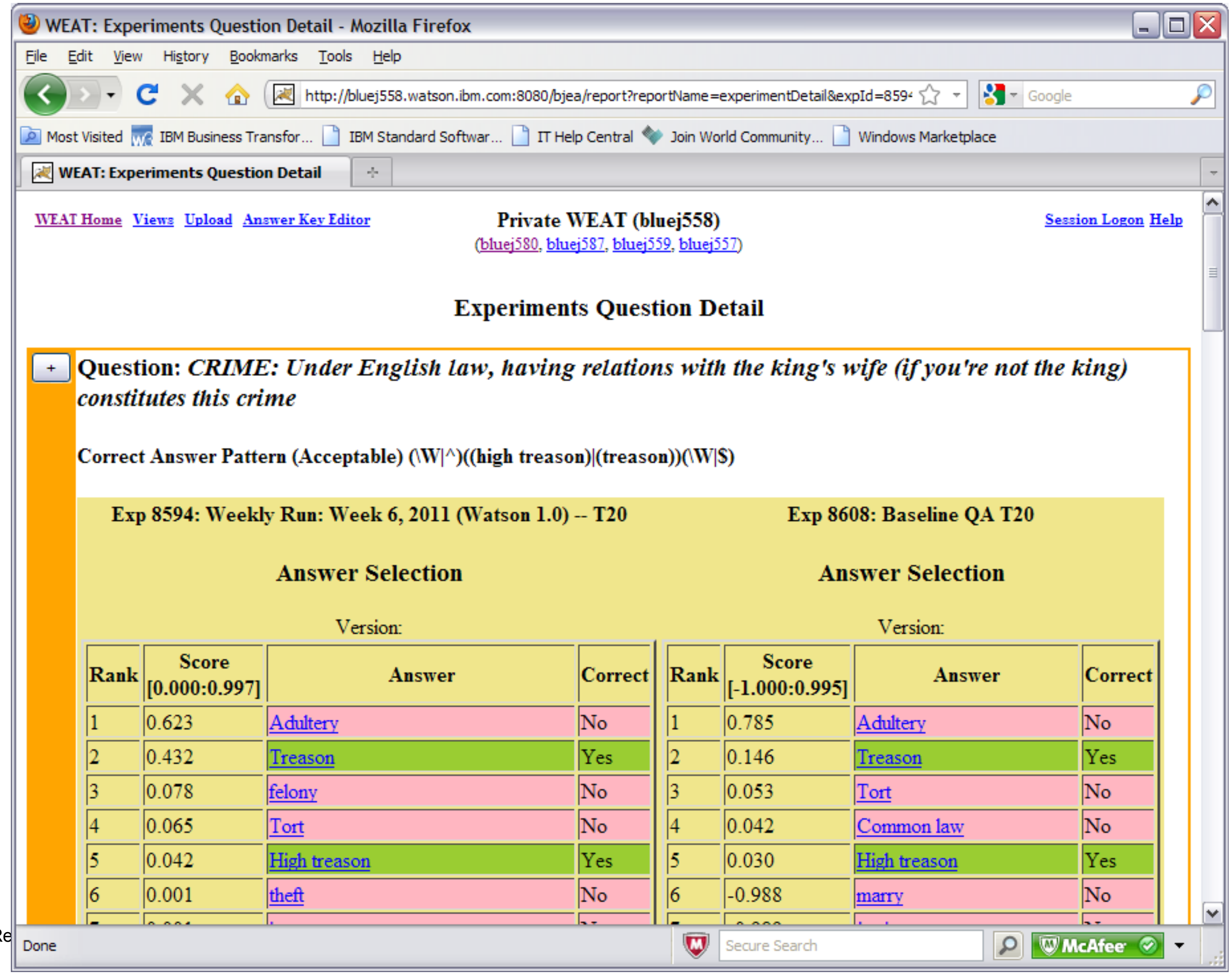
Refine View

Action: Compare Experiments | Perform Selected Action

Sel	Experiment	Experiment Date & Time	Experiment Description	User
☑	8618	Feb 28, 2011	Baseline QA with new ESG, T20	alally
☑	8608	Feb 22, 2011	Baseline QA T20	alally
☐	*8594s*	Feb 10, 2011	Weekly Run: Week 6, 2011 (Watson 1.0) -- T20	murdockj
☐	*8281s*	Dec 20, 2010	Weekly Run: Week 50, 2010 (ending on Dec 20, 2010) -- T20	murdockj
☐	*8123s*	Dec 6, 2010	Weekly Run: Week 48, 2010 (ending on Dec 6, 2010) -- T20	ewb
☐	8061	Nov 23, 2010	w44 T20 w/ alternative names max cap	fanj
☐	*8042s*	Nov 21, 2010	Weekly Run: Week 46, 2010 (ending on Nov 19, 2010) -- T20	jencc
☐	*7872s*	Nov 7, 2010	Weekly Run: Week 44, 2010 (ending on Nov 5, 2010) -- T20	murdockj
☐	*7481s*	Oct 9, 2010	Weekly Run: Week 40, 2010 (ending on Oct 8, 2010) -- T20	alally
☐	*7295s*	Aug 29, 2010	Weekly Run: Week 34, 2010 (ending on Aug 29, 2010) -- T20	ewb
☐	*7262s*	Aug 26, 2010	T20 metadata upload	jencc
☐	*7001s*	Jul 30, 2010	Weekly Run: Week 30, 2010 (ending on July 30, 2010) -- T20	murdockj

Done | Secure Search | McAfee

WEAT: View Answer Lists



WEAT: Experiments Question Detail - Mozilla Firefox

File Edit View History Bookmarks Tools Help

http://bluej558.watson.ibm.com:8080/bjea/report?reportName=experimentDetail&expId=859

Most Visited | IBM Business Transfor... | IBM Standard Softwar... | IT Help Central | Join World Community... | Windows Marketplace

WEAT: Experiments Question Detail

WEAT Home Views Upload Answer Key Editor

Private WEAT (bluej558)
(bluej580, bluej587, bluej559, bluej557)

Session Logon Help

Experiments Question Detail

Question: ***CRIME: Under English law, having relations with the king's wife (if you're not the king) constitutes this crime***

Correct Answer Pattern (Acceptable) (\W|^)((high treason)|(treason))(\W|$)

Exp 8594: Weekly Run: Week 6, 2011 (Watson 1.0) -- T20

Answer Selection

Version:

Rank	Score [0.000:0.997]	Answer	Correct
1	0.623	Adultery	No
2	0.432	Treason	Yes
3	0.078	felony	No
4	0.065	Tort	No
5	0.042	High treason	Yes
6	0.001	theft	No

Exp 8608: Baseline QA T20

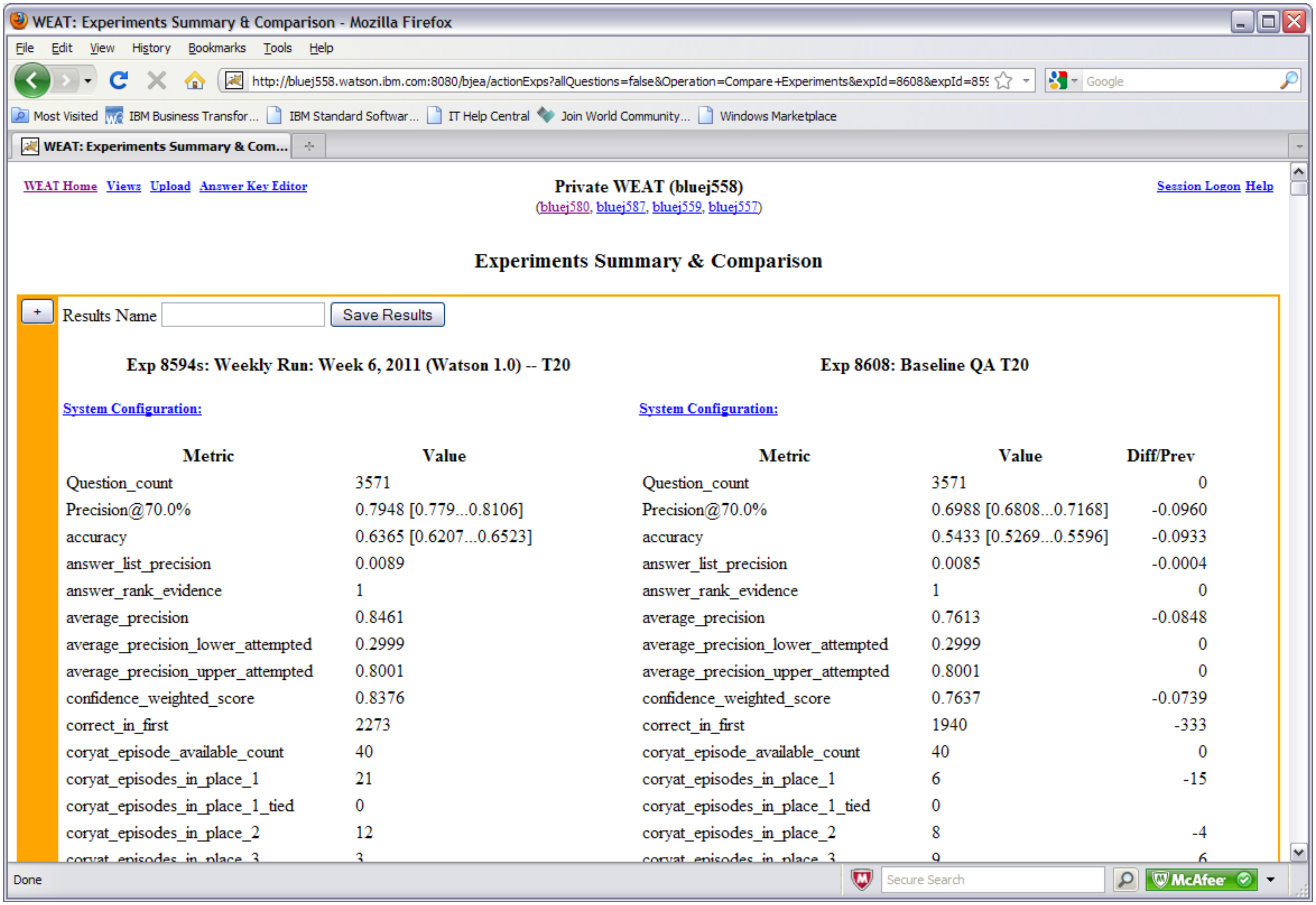
Answer Selection

Version:

Rank	Score [-1.000:0.995]	Answer	Correct
1	0.785	Adultery	No
2	0.146	Treason	Yes
3	0.053	Tort	No
4	0.042	Common law	No
5	0.030	High treason	Yes
6	-0.988	marry	No

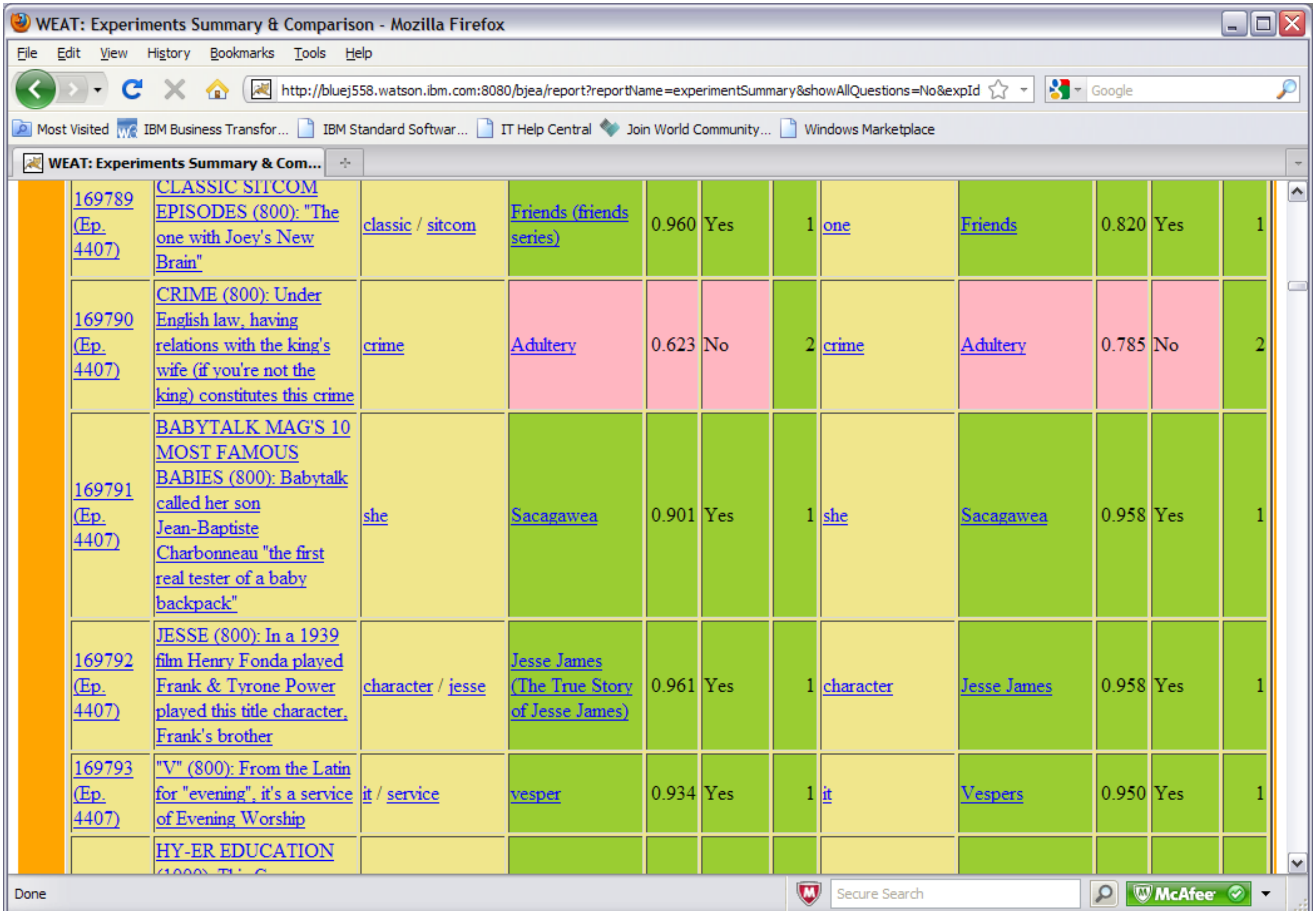
Done | Secure Search | McAfee

WEAT: Compare Experiments



Metric	Value	Metric	Value	Diff/Prev
Question_count	3571	Question_count	3571	0
Precision@70.0%	0.7948 [0.779...0.8106]	Precision@70.0%	0.6988 [0.6808...0.7168]	-0.0960
accuracy	0.6365 [0.6207...0.6523]	accuracy	0.5433 [0.5269...0.5596]	-0.0933
answer_list_precision	0.0089	answer_list_precision	0.0085	-0.0004
answer_rank_evidence	1	answer_rank_evidence	1	0
average_precision	0.8461	average_precision	0.7613	-0.0848
average_precision_lower_attempted	0.2999	average_precision_lower_attempted	0.2999	0
average_precision_upper_attempted	0.8001	average_precision_upper_attempted	0.8001	0
confidence_weighted_score	0.8376	confidence_weighted_score	0.7637	-0.0739
correct_in_first	2273	correct_in_first	1940	-333
coryat_episode_available_count	40	coryat_episode_available_count	40	0
coryat_episodes_in_place_1	21	coryat_episodes_in_place_1	6	-15
coryat_episodes_in_place_1_tied	0	coryat_episodes_in_place_1_tied	0	
coryat_episodes_in_place_2	12	coryat_episodes_in_place_2	8	-4
coryat_episodes_in_place_3	3	coryat_episodes_in_place_3	9	6

WEAT: Compare Experiments



169789 (Ep. 4407)	CLASSIC SITCOM EPISODES (800): "The one with Joey's New Brain"	classic / sitcom	Friends (friends series)	0.960	Yes	1	one	Friends	0.820	Yes	1
169790 (Ep. 4407)	CRIME (800): Under English law, having relations with the king's wife (if you're not the king) constitutes this crime	crime	Adultery	0.623	No	2	crime	Adultery	0.785	No	2
169791 (Ep. 4407)	BABYTALK MAG'S 10 MOST FAMOUS BABIES (800): Babytalk called her son Jean-Baptiste Charbonneau "the first real tester of a baby backpack"	she	Sacagawea	0.901	Yes	1	she	Sacagawea	0.958	Yes	1
169792 (Ep. 4407)	JESSE (800): In a 1939 film Henry Fonda played Frank & Tyrone Power played this title character, Frank's brother	character / jesse	Jesse James (The True Story of Jesse James)	0.961	Yes	1	character	Jesse James	0.958	Yes	1
169793 (Ep. 4407)	"V" (800): From the Latin for "evening", it's a service of Evening Worship	it / service	vesper	0.934	Yes	1	it	Vespers	0.950	Yes	1
	HY-ER EDUCATION										

Feature Analysis Tool

Question#508175

LITERARY CHARACTER APB: His victims include Charity Burbage, Mad Eye Moody & Severus Snape; he'd

Diff [Lord Voldemort , Harry Potter]

Compare "Lord Voldemort" and "Harry Potter"

Feature Groups/Features	Diff Value Bar	Diff Value	Lord ...mort	Harry Potter
Experiment Label			exhibi...swers	exhibi...swers
Selected Model			base	base
Final Score			0.444	0.622
Weighted Features Sum		-0.571	2.340	2.911
DOCUMENT_SUPPORT	Lord Voldemort Harry Potter	**-0.167**	**-0.797**	**-0.631**
GENERIC_SPECIFIC		**0.113**	**0.255**	**0.142**
PASSAGE_SUPPORT		**0.775**	**2.998**	**2.223**
POPULARITY		**-0.411**	**1.245**	**1.656**
SOURCE_RELIABILITY		**-0.585**	**-0.127**	**0.458**
TYPE_MATCH		**0.035**	**0.196**	**0.161**
WORD_ASSOCIATION		**-0.331**	**-1.430**	**-1.099**

Development cycle

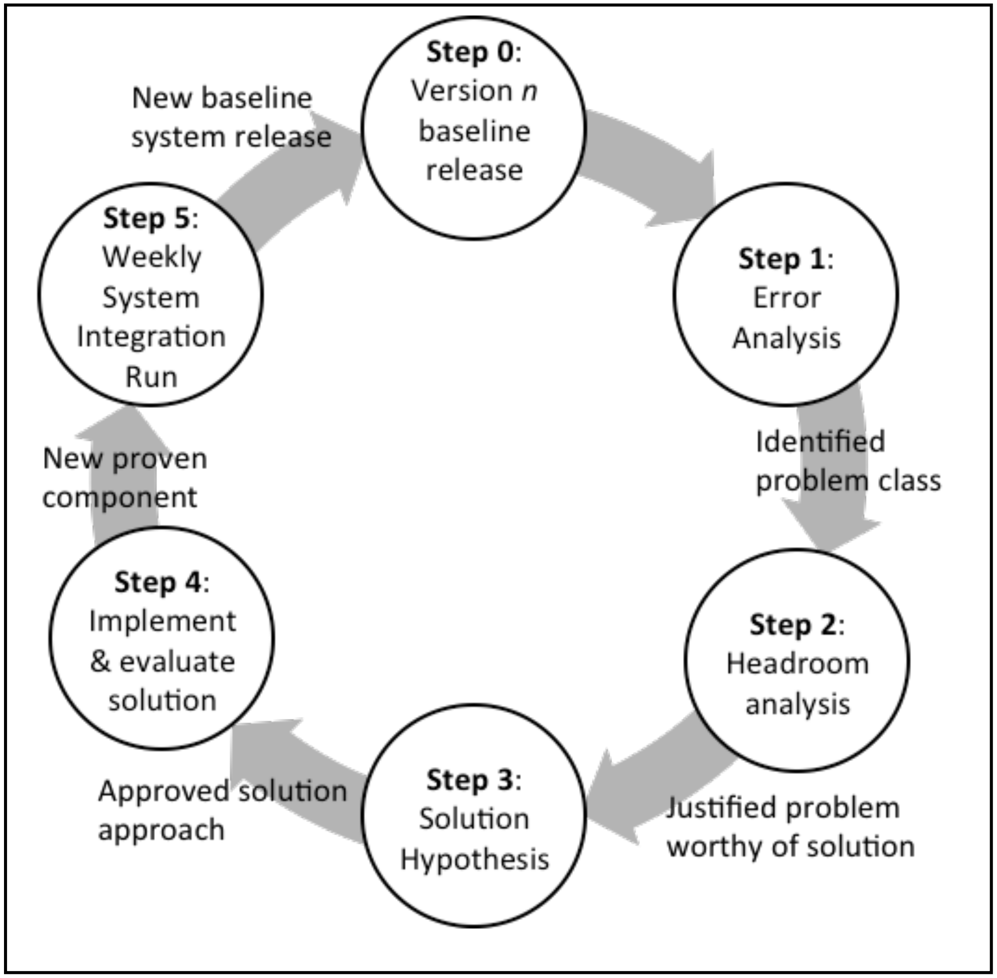


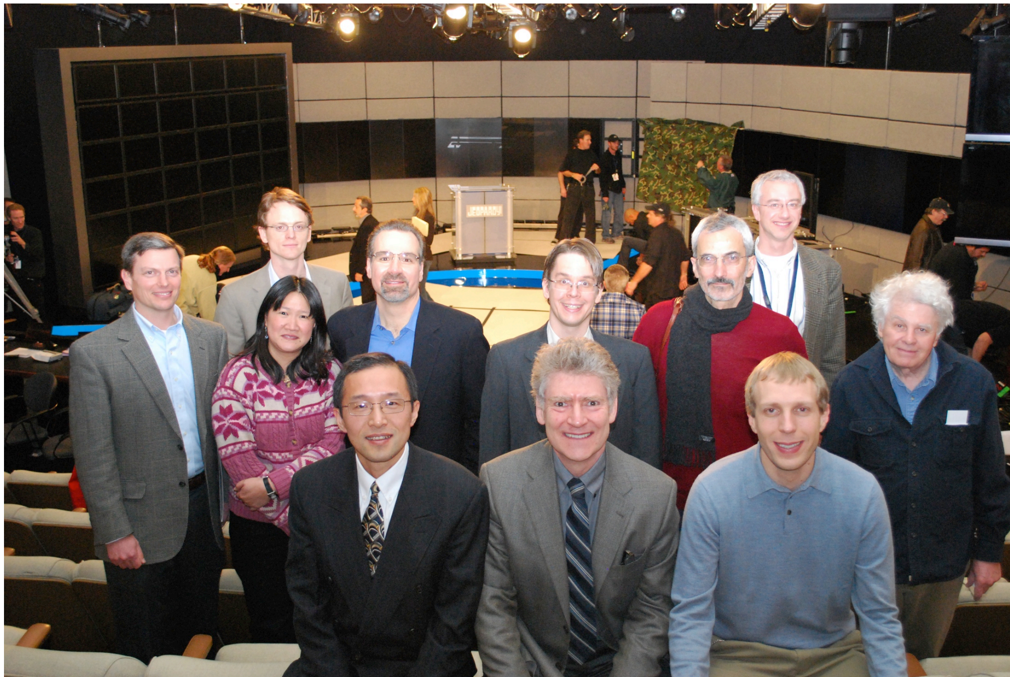
Figure 1 System development lifecycle.

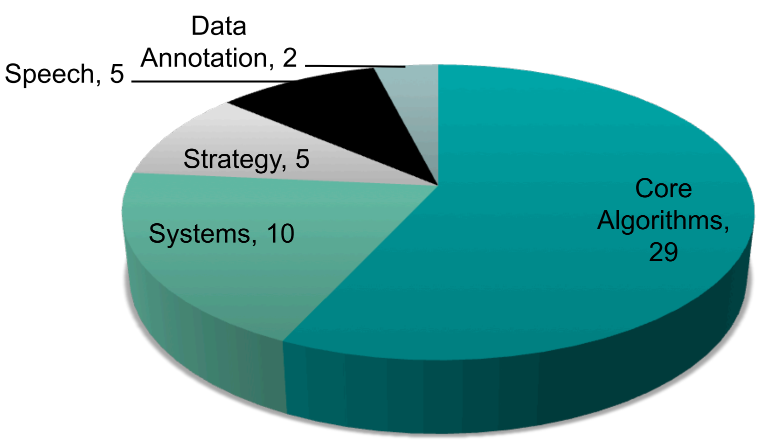
- **4 years**
- **12 -> 30 people**
- metrics-driven research
  - Component level
  - End-to-end system
- informative **error analysis**
- **High investment in Methodology and Tools**

Step1: Tycor is weak in President Names
Step2: Suppose Tycor is able to recognize all President Names in the system, what's the gain
Step 3: if step 2 is significant

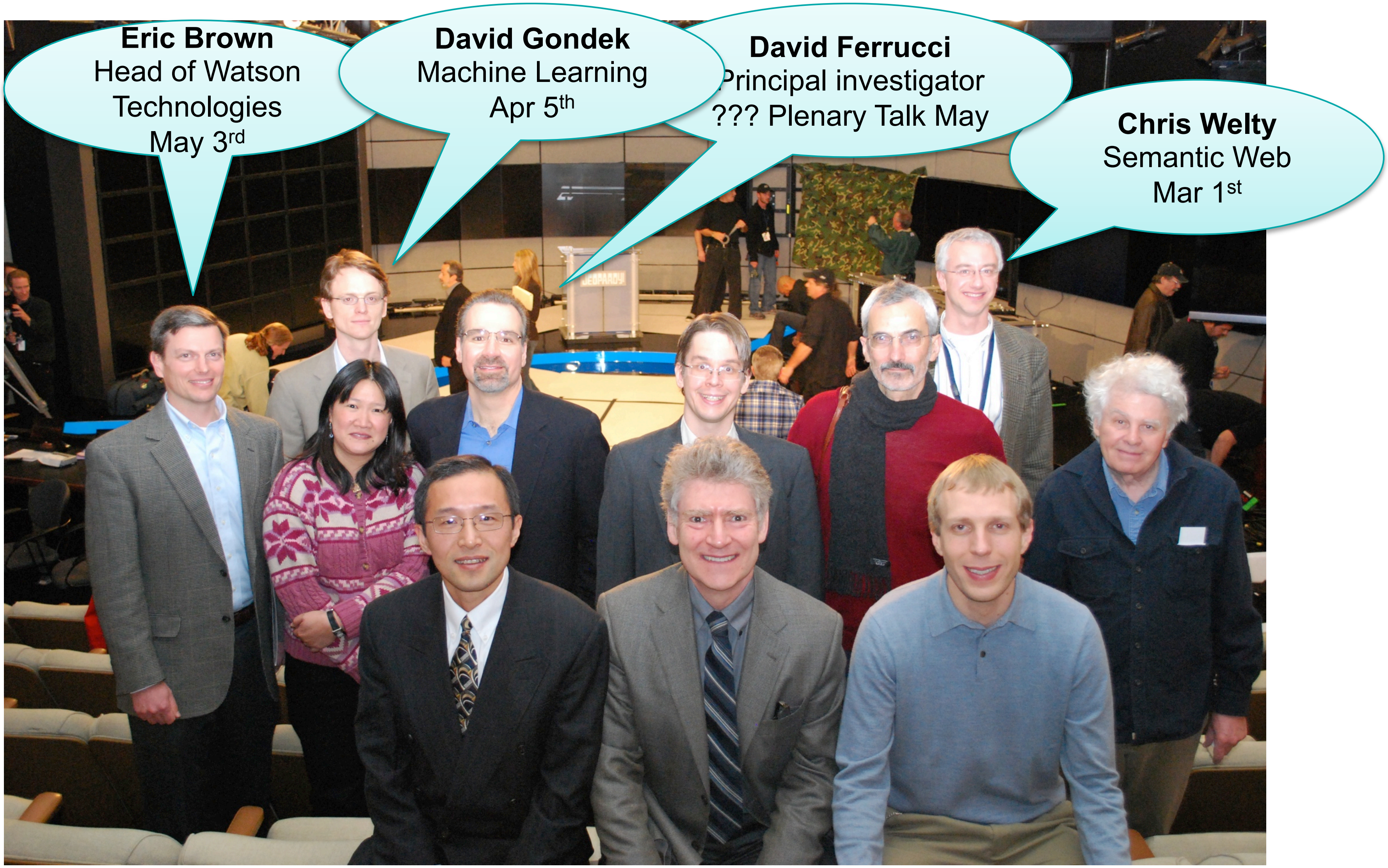
The Core Team

4 years
12 -> 30 people





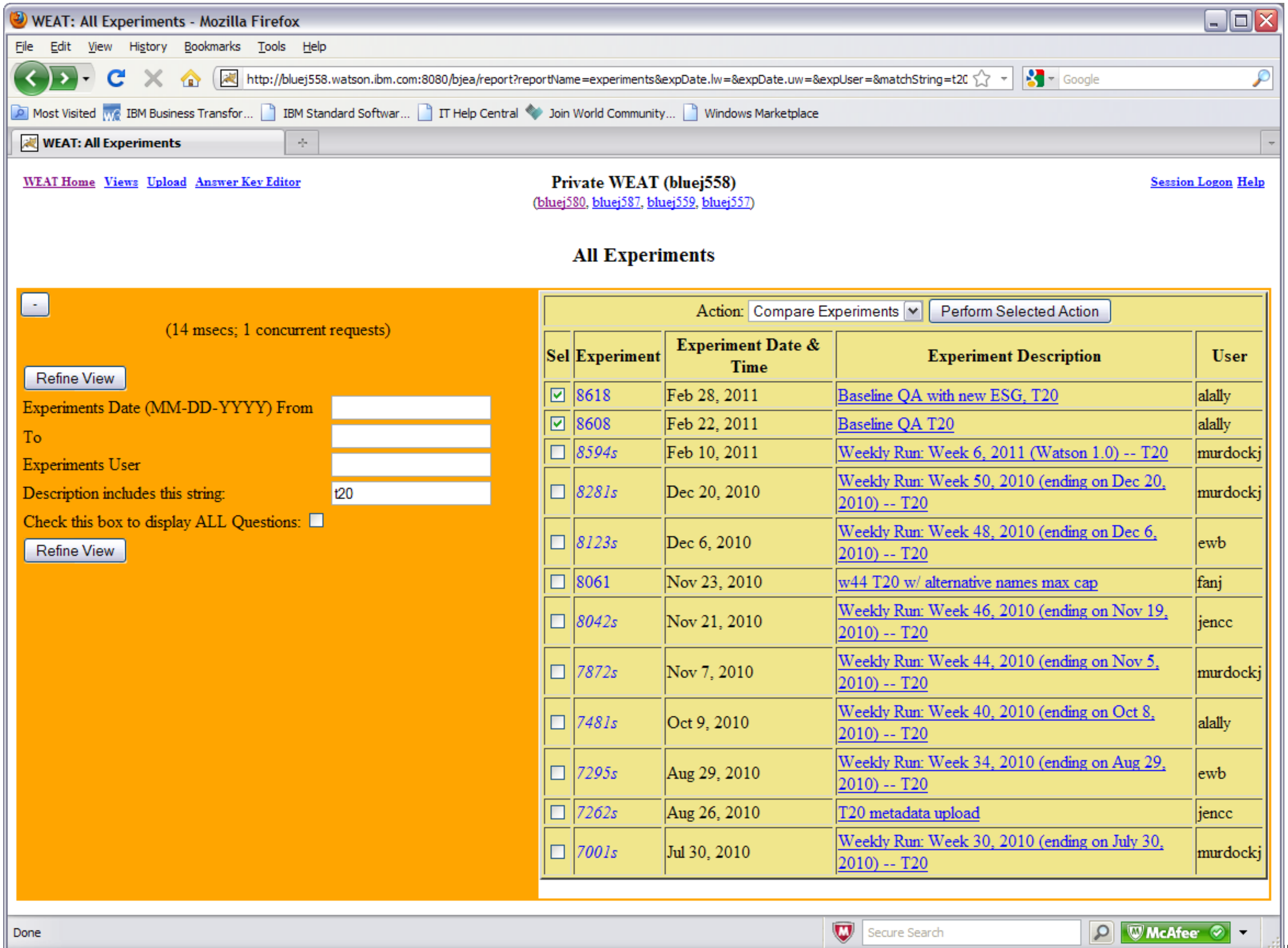
Guest Speakers



Development Tools

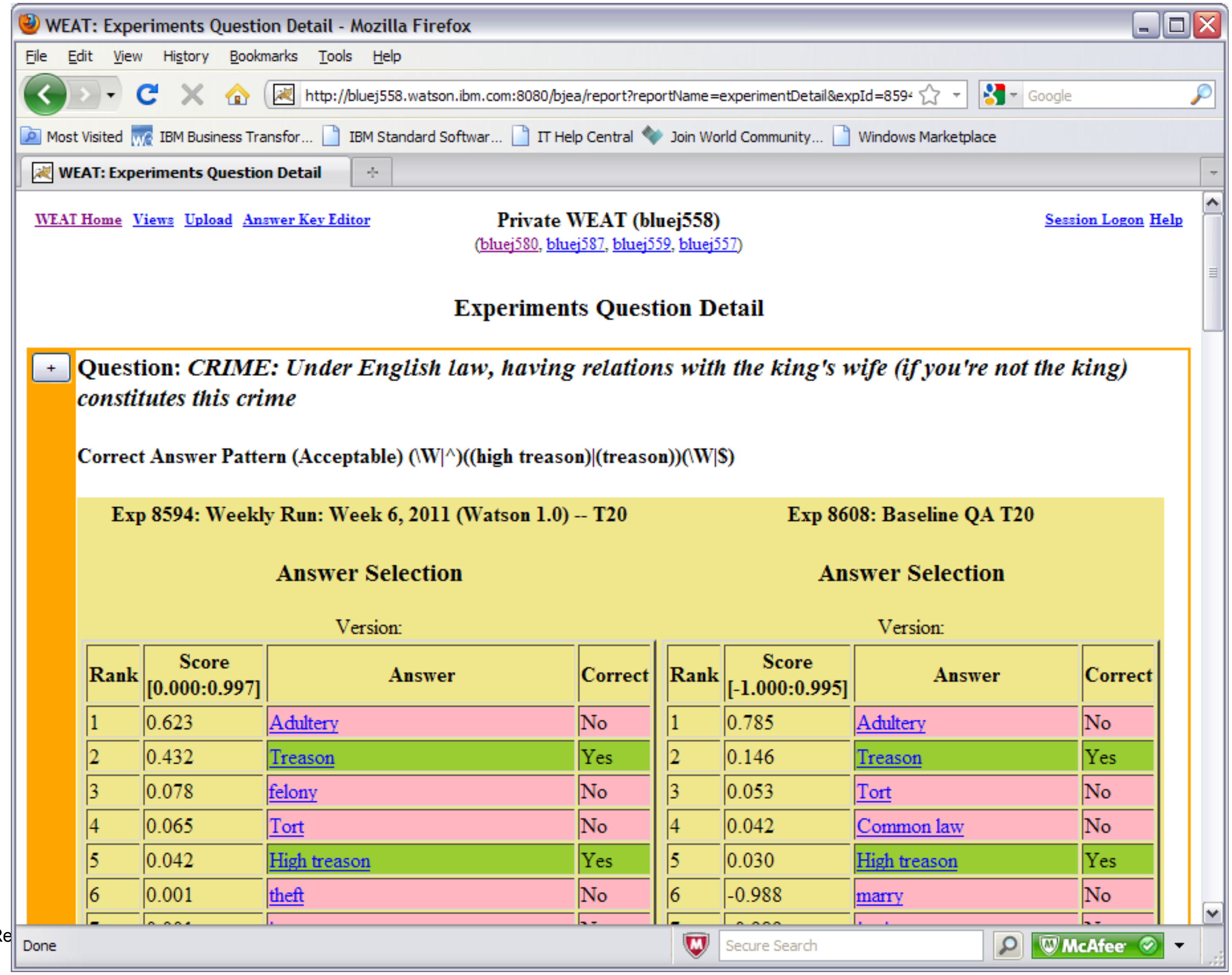
- Data Repository
  - Data resources include question sets, corpora, indices, models, annotated data
  - Support for resource documentation, meta-data, and versioning
  - Web browsing or programmatic access via a URI
- Job Scheduling and Management Tool
  - Allocate Machines to Users' Jobs
  - Load Balance User's Job Across Machines
  - Report Errors
- Error Analysis Tool
  - Browse Results of an Experiment
  - Compare Results between Experiments
- Feature Analysis Tool
  - View "Evidence Profiles" for Answers
  - Find out which features contribute to an answer's (mis)ranking

WEAT: Watson Error Analysis Tool



Sel	Experiment	Experiment Date & Time	Experiment Description	User
☑	8618	Feb 28, 2011	Baseline QA with new ESG, T20	alally
☑	8608	Feb 22, 2011	Baseline QA T20	alally
☐	8594s	Feb 10, 2011	Weekly Run: Week 6, 2011 (Watson 1.0) -- T20	murdockj
☐	8281s	Dec 20, 2010	Weekly Run: Week 50, 2010 (ending on Dec 20, 2010) -- T20	murdockj
☐	8123s	Dec 6, 2010	Weekly Run: Week 48, 2010 (ending on Dec 6, 2010) -- T20	ewb
☐	8061	Nov 23, 2010	w44 T20 w/ alternative names max cap	fanj
☐	8042s	Nov 21, 2010	Weekly Run: Week 46, 2010 (ending on Nov 19, 2010) -- T20	jencc
☐	7872s	Nov 7, 2010	Weekly Run: Week 44, 2010 (ending on Nov 5, 2010) -- T20	murdockj
☐	7481s	Oct 9, 2010	Weekly Run: Week 40, 2010 (ending on Oct 8, 2010) -- T20	alally
☐	7295s	Aug 29, 2010	Weekly Run: Week 34, 2010 (ending on Aug 29, 2010) -- T20	ewb
☐	7262s	Aug 26, 2010	T20 metadata upload	jencc
☐	7001s	Jul 30, 2010	Weekly Run: Week 30, 2010 (ending on July 30, 2010) -- T20	murdockj

WEAT: View Answer Lists



WEAT: Experiments Question Detail - Mozilla Firefox

File Edit View History Bookmarks Tools Help

http://bluej558.watson.ibm.com:8080/bjea/report?reportName=experimentDetail&expId=859

Most Visited | IBM Business Transfor... | IBM Standard Softwar... | IT Help Central | Join World Community... | Windows Marketplace

WEAT: Experiments Question Detail

WEAT Home Views Upload Answer Key Editor

Private WEAT (bluej558)

(bluej580, bluej587, bluej559, bluej557)

Session Logon Help

Experiments Question Detail

Question: ***CRIME: Under English law, having relations with the king's wife (if you're not the king) constitutes this crime***

Correct Answer Pattern (Acceptable) (\W|^)((high treason)|(treason))(\W|$)

Exp 8594: Weekly Run: Week 6, 2011 (Watson 1.0) -- T20

Answer Selection

Version:

Rank	Score [0.000:0.997]	Answer	Correct
1	0.623	Adultery	No
2	0.432	Treason	Yes
3	0.078	felony	No
4	0.065	Tort	No
5	0.042	High treason	Yes
6	0.001	theft	No

Exp 8608: Baseline QA T20

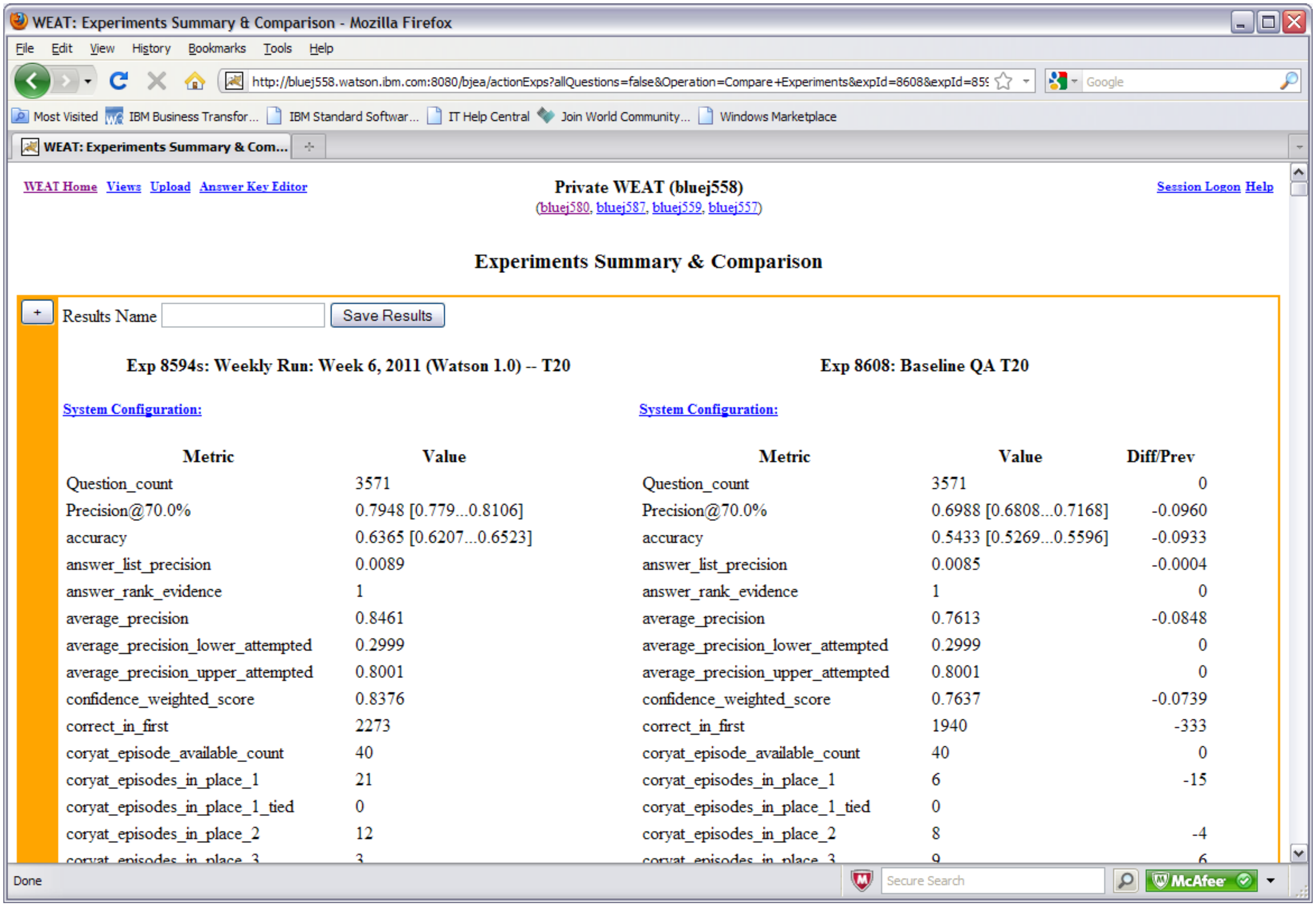
Answer Selection

Version:

Rank	Score [-1.000:0.995]	Answer	Correct
1	0.785	Adultery	No
2	0.146	Treason	Yes
3	0.053	Tort	No
4	0.042	Common law	No
5	0.030	High treason	Yes
6	-0.988	marry	No

Done | Secure Search | McAfee

WEAT: Compare Experiments



WEAT: Experiments Summary & Comparison - Mozilla Firefox

WEAT Home Views Upload Answer Key Editor — Session Logon Help

Private WEAT (bluej558)
(bluej580, bluej587, bluej559, bluej557)

Experiments Summary & Comparison

Results Name [ ] Save Results

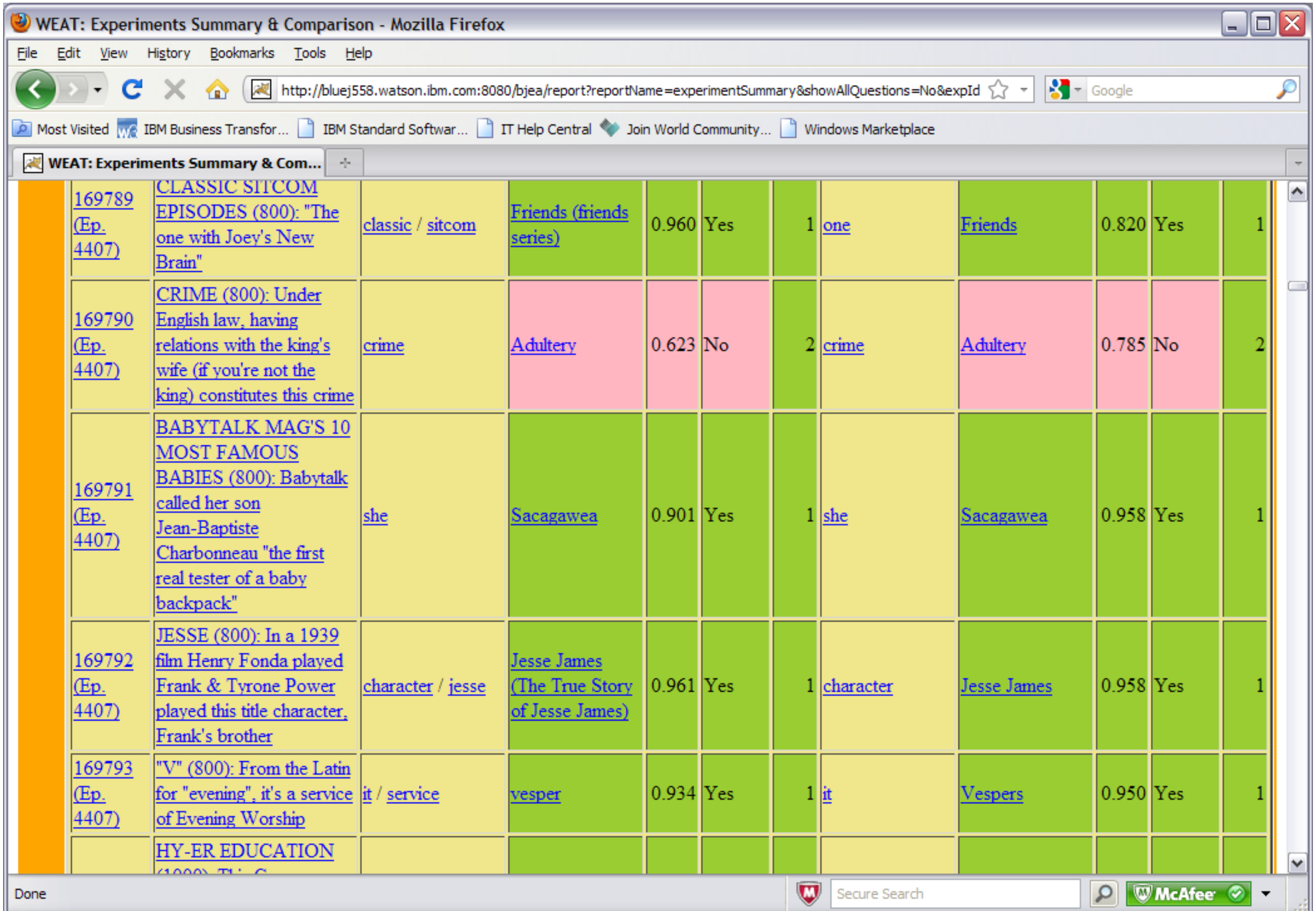
Exp 8594s: Weekly Run: Week 6, 2011 (Watson 1.0) -- T20 — **Exp 8608: Baseline QA T20**

System Configuration: — System Configuration:

Metric	Value	Metric	Value	Diff/Prev
Question_count	3571	Question_count	3571	0
Precision@70.0%	0.7948 [0.779...0.8106]	Precision@70.0%	0.6988 [0.6808...0.7168]	-0.0960
accuracy	0.6365 [0.6207...0.6523]	accuracy	0.5433 [0.5269...0.5596]	-0.0933
answer_list_precision	0.0089	answer_list_precision	0.0085	-0.0004
answer_rank_evidence	1	answer_rank_evidence	1	0
average_precision	0.8461	average_precision	0.7613	-0.0848
average_precision_lower_attempted	0.2999	average_precision_lower_attempted	0.2999	0
average_precision_upper_attempted	0.8001	average_precision_upper_attempted	0.8001	0
confidence_weighted_score	0.8376	confidence_weighted_score	0.7637	-0.0739
correct_in_first	2273	correct_in_first	1940	-333
coryat_episode_available_count	40	coryat_episode_available_count	40	0
coryat_episodes_in_place_1	21	coryat_episodes_in_place_1	6	-15
coryat_episodes_in_place_1_tied	0	coryat_episodes_in_place_1_tied	0	
coryat_episodes_in_place_2	12	coryat_episodes_in_place_2	8	-4
coryat_episodes_in_place_3	3	coryat_episodes_in_place_3	9	6

Done

WEAT: Compare Experiments



169789 (Ep. 4407)	CLASSIC SITCOM EPISODES (800): "The one with Joey's New Brain"	classic / sitcom	Friends (friends series)	0.960	Yes	1	one	Friends	0.820	Yes	1
169790 (Ep. 4407)	CRIME (800): Under English law, having relations with the king's wife (if you're not the king) constitutes this crime	crime	Adultery	0.623	No	2	crime	Adultery	0.785	No	2
169791 (Ep. 4407)	BABYTALK MAG'S 10 MOST FAMOUS BABIES (800): Babytalk called her son Jean-Baptiste Charbonneau "the first real tester of a baby backpack"	she	Sacagawea	0.901	Yes	1	she	Sacagawea	0.958	Yes	1
169792 (Ep. 4407)	JESSE (800): In a 1939 film Henry Fonda played Frank & Tyrone Power played this title character, Frank's brother	character / jesse	Jesse James (The True Story of Jesse James)	0.961	Yes	1	character	Jesse James	0.958	Yes	1
169793 (Ep. 4407)	"V" (800): From the Latin for "evening", it's a service of Evening Worship	it / service	vesper	0.934	Yes	1	it	Vespers	0.950	Yes	1
	HY-ER EDUCATION										

Feature Analysis Tool

Question#508175

LITERARY CHARACTER APB: His victims include Charity Burbage, Mad Eye Moody & Severus Snape; he'd

Diff [Lord Voldemort , Harry Potter]

Compare "Lord Voldemort" and "Harry Potter"

Feature Groups/Features	Diff Value Bar	Diff Value	Lord ...mort	Harry Potter
Experiment Label			exhibi...swers	exhibi...swers
Selected Model			base	base
Final Score			0.444	0.622
Weighted Features Sum		-0.571	2.340	2.911
DOCUMENT_SUPPORT	Lord Voldemort Harry Potter	-0.167	-0.797	-0.631
GENERIC_SPECIFIC		0.113	0.255	0.142
PASSAGE_SUPPORT		0.775	2.998	2.223
POPULARITY		-0.411	1.245	1.656
SOURCE_RELIABILITY		-0.585	-0.127	0.458
TYPE_MATCH		0.035	0.196	0.161
WORD_ASSOCIATION		-0.331	-1.430	-1.099

Development cycle

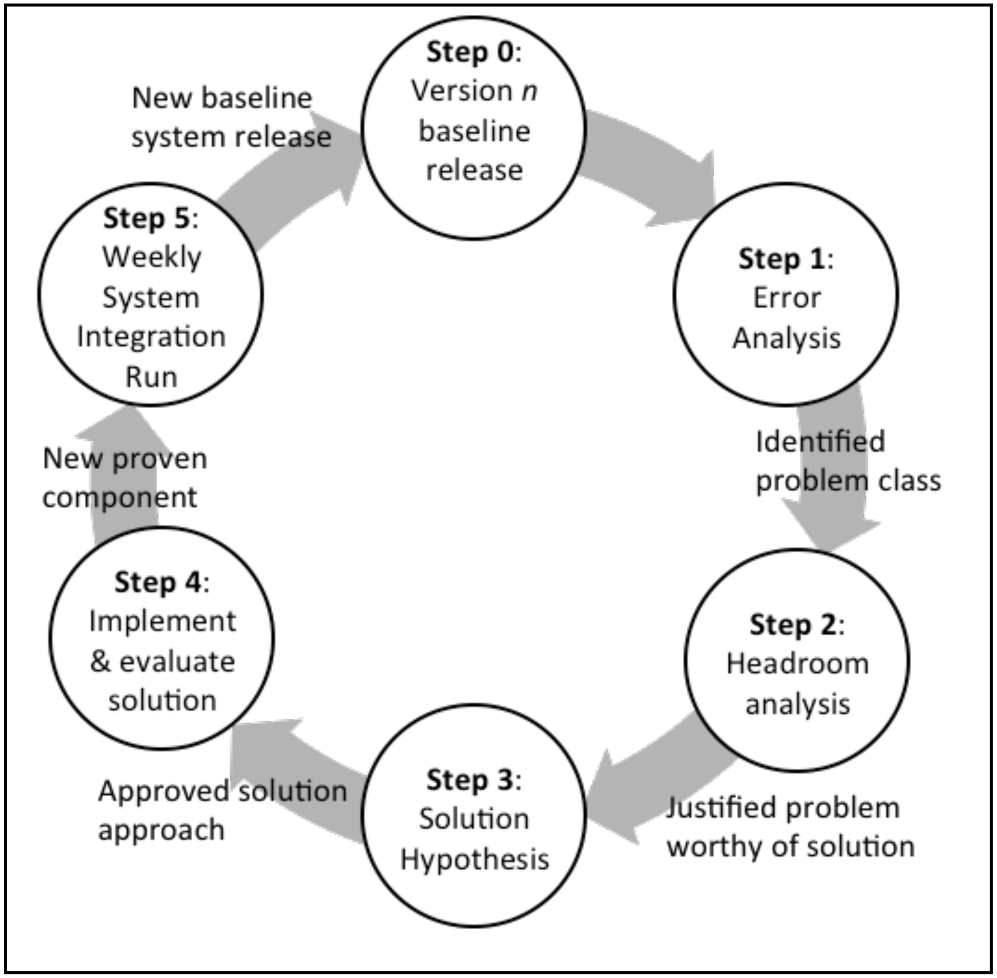
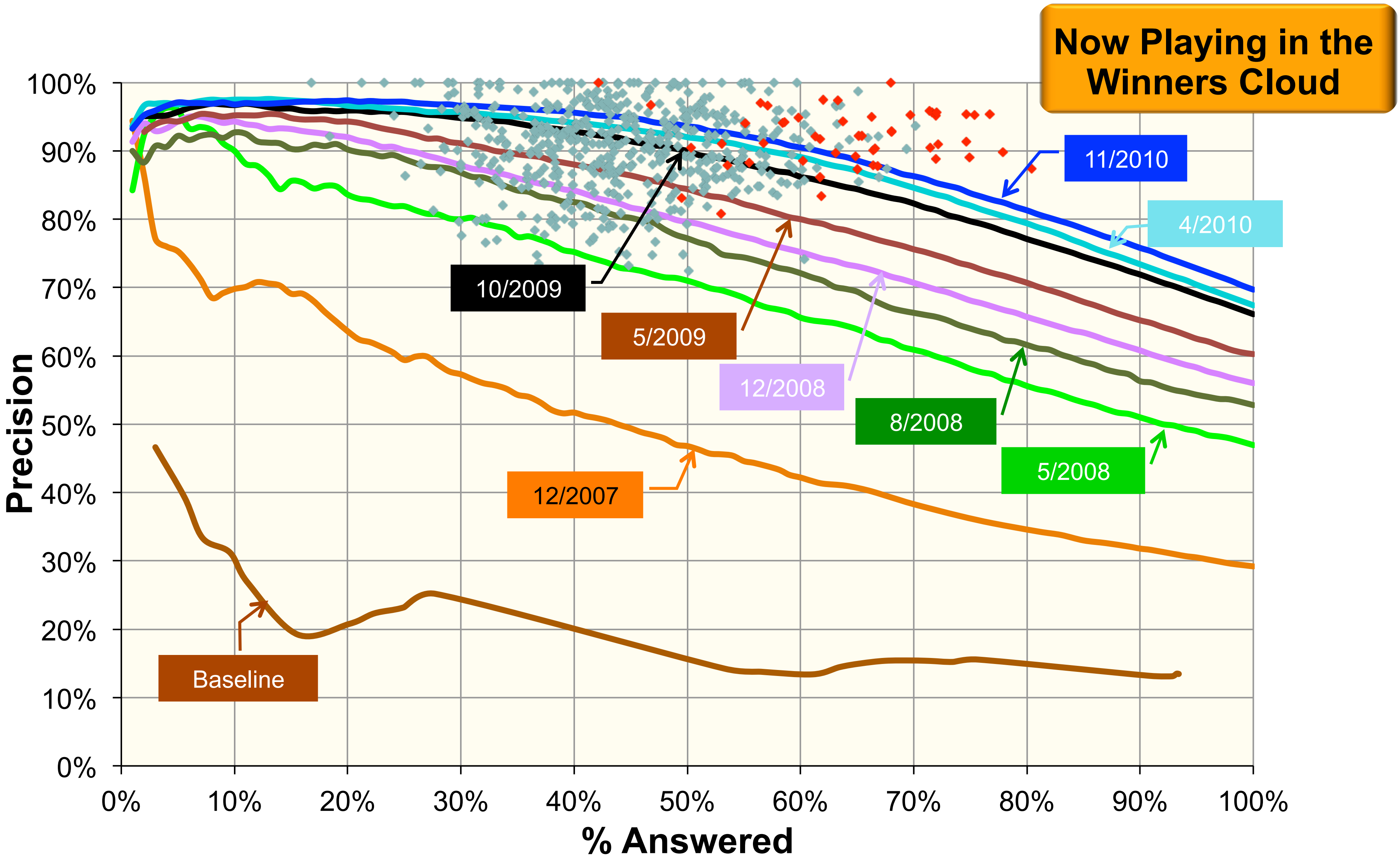


Figure 1 System development lifecycle.

- **4 years**
- **12 -> 30 people**
- metrics-driven research
  - –Component level
  - –End-to-end system
- informative **error analysis**
- **High investment in Methodology and Tools**

Step1: Tycor is weak in President Names
Step2: Suppose Tycor is able to recognize all President Names in the system, what's the gain
Step 3: if step 2 is significant

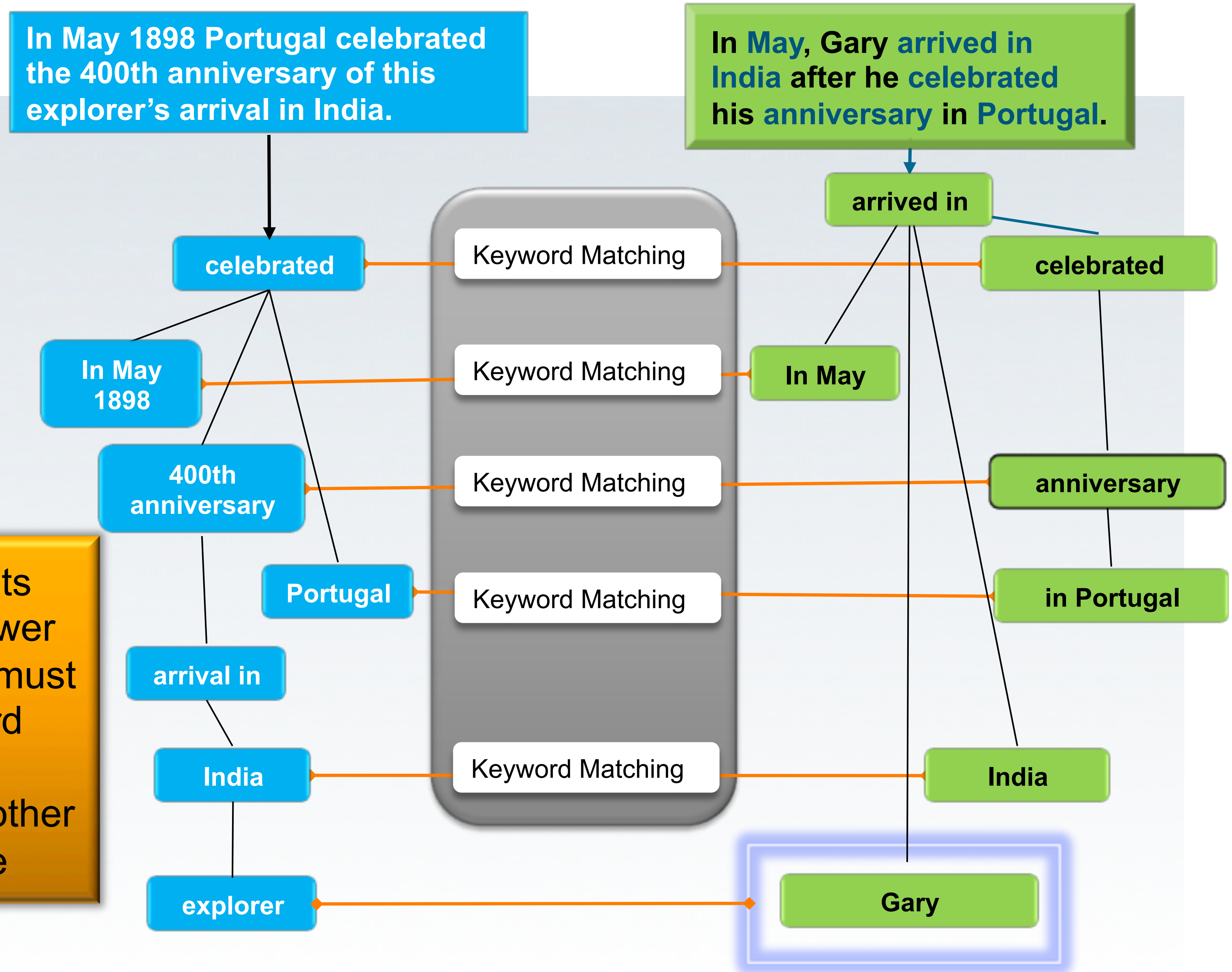
Compare Experiments



Outline

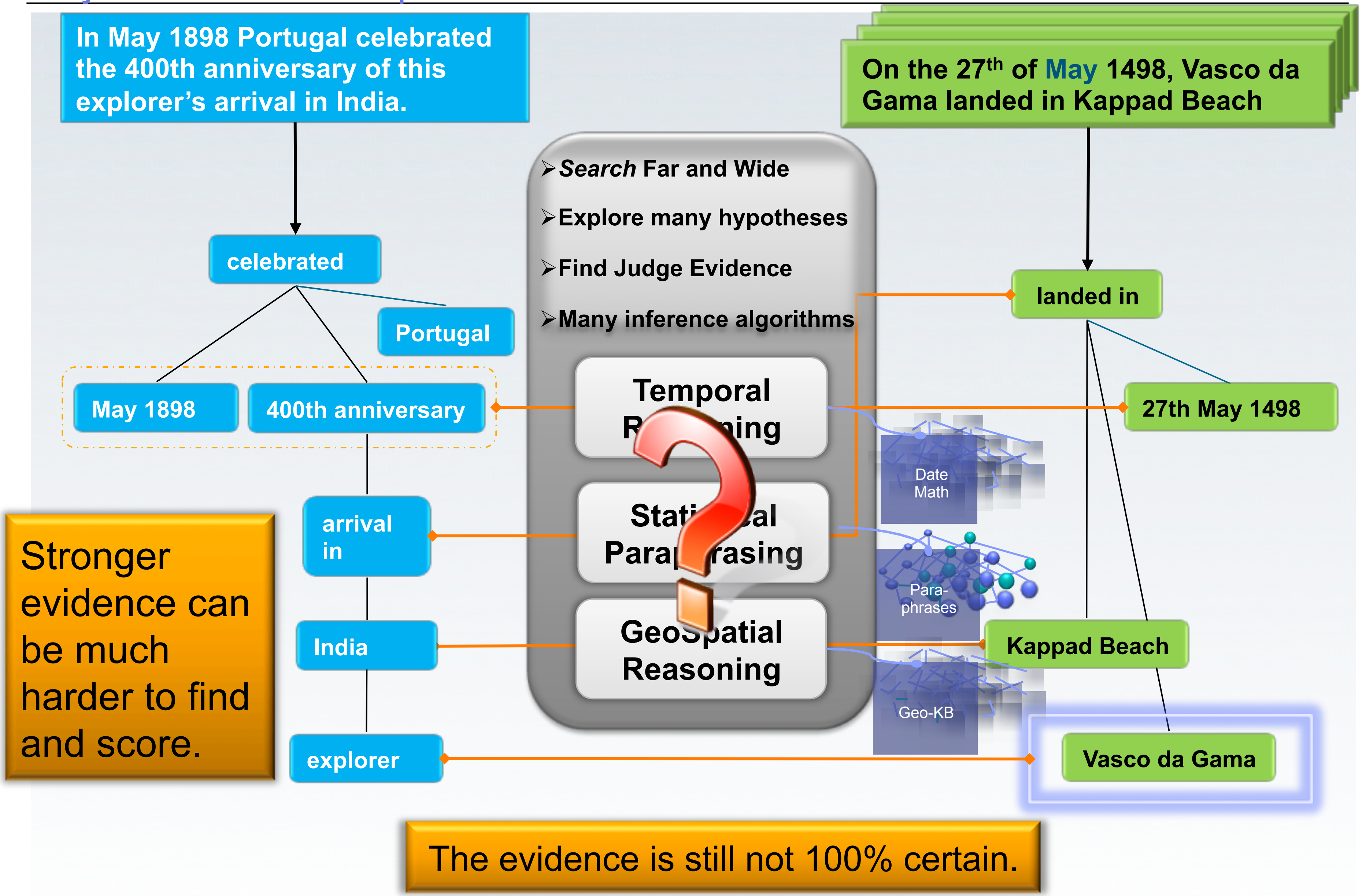
- More on Candidate Generation
- The Development Cycle
- Why Semantics?

Keyword Evidence



Evidence suggests "Gary" is the answer BUT the system must learn that keyword matching may be weak relative to other types of evidence

Why Semantics? Deeper Evidence

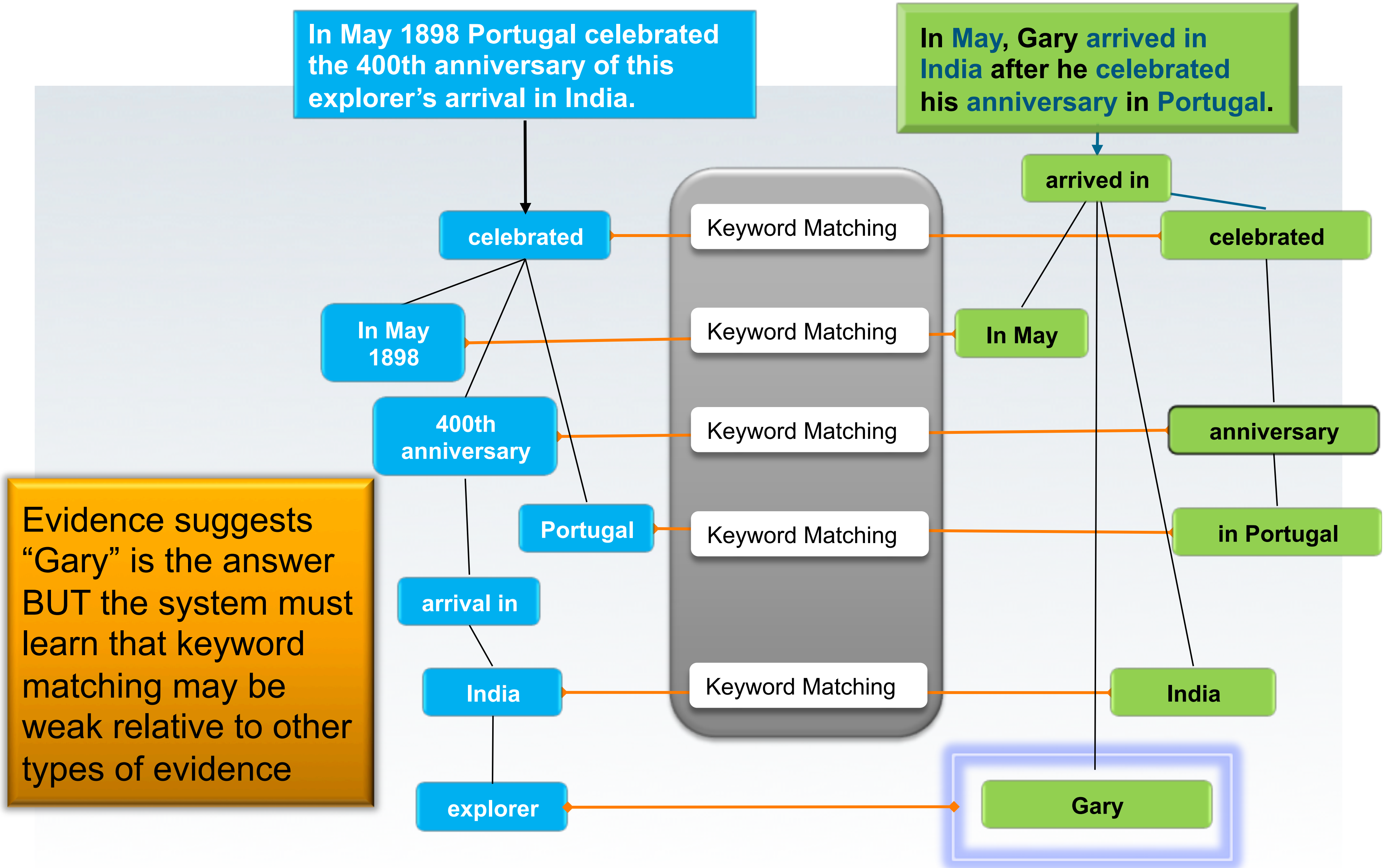


Semantic Technologies in IBM Watson™

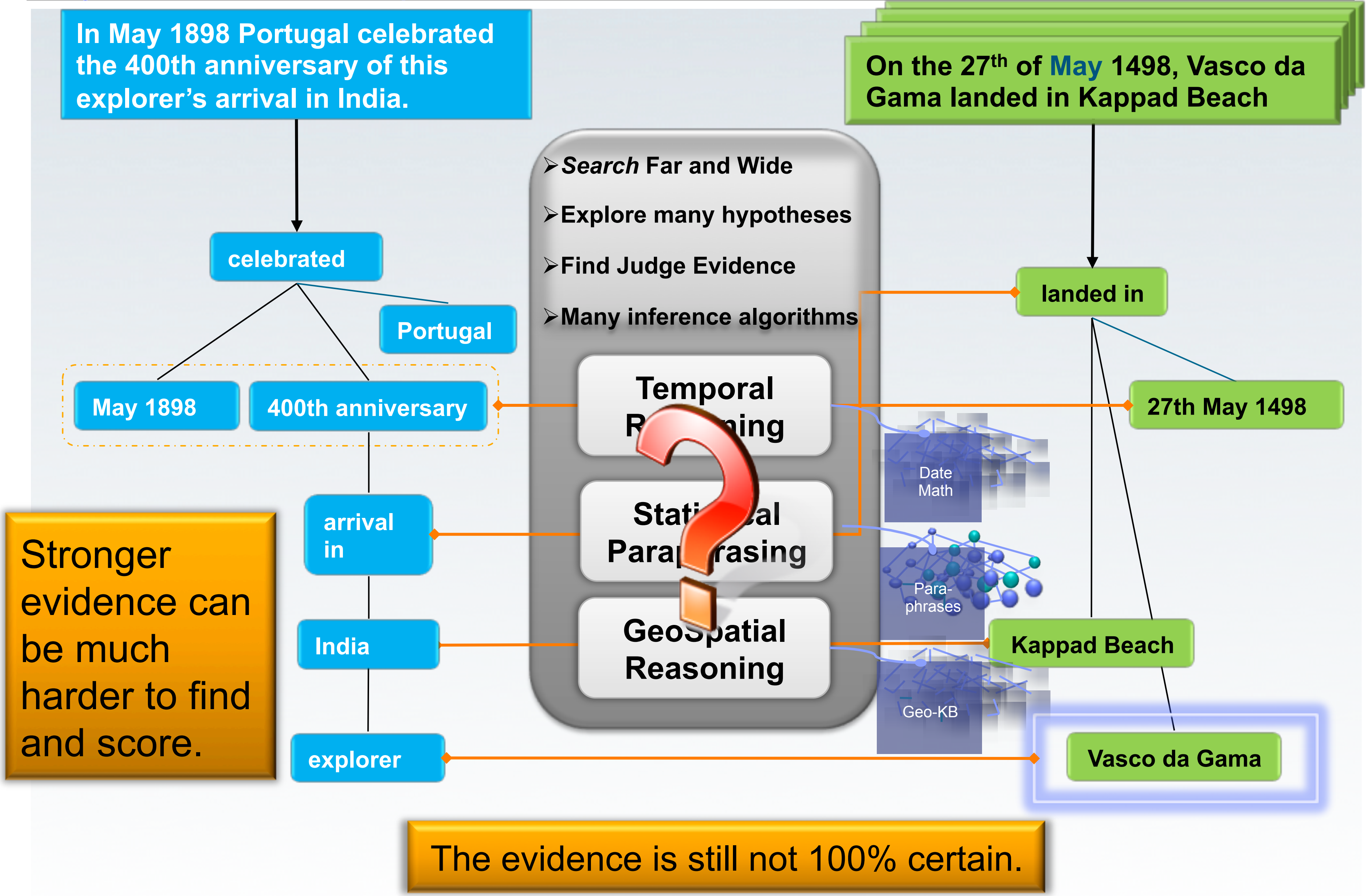
Lesson 4 – Natural Language Processing

Professor: Alfio Massimiliano Gliozzo
TA: Or Biran

Keyword Evidence



Why Semantics? Deeper Evidence



Outline

- The NLP Stack
- Question Analysis
- Passage Scoring

Outline

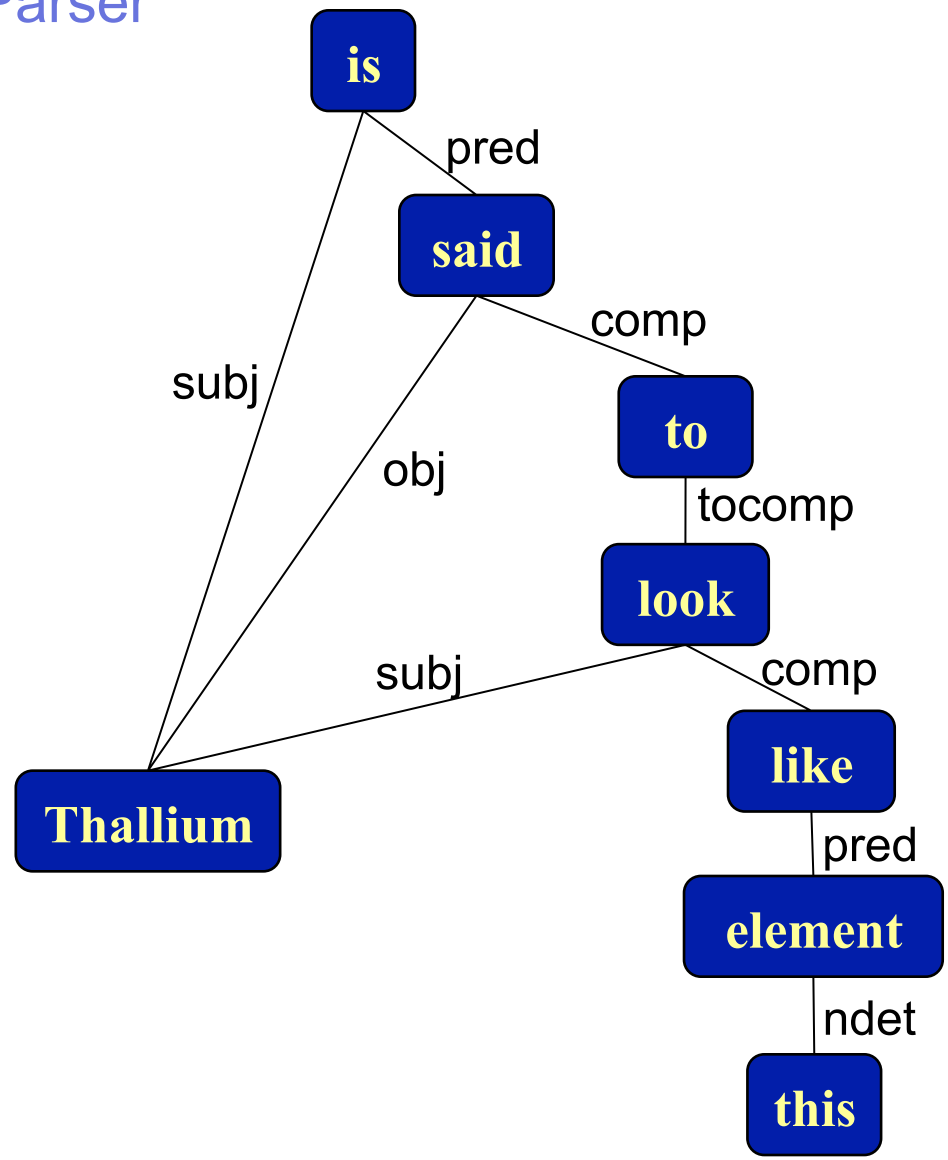
- The NLP Stack
- Question Analysis
- Passage Scoring

NLP Stack

- ESG: English Slot Grammar Parser
- Predicate Argument Structure (simplified parse)
- Rule-Based Named Entity Detection
- Intra-Paragraph Anaphora Resolution
- Temporal Normalization
- Temporal Arithmetic
- Pattern-Based Semantic Relation Detection
- Statistical Semantic Relation Detection

ESG: English Slot Grammar Parser

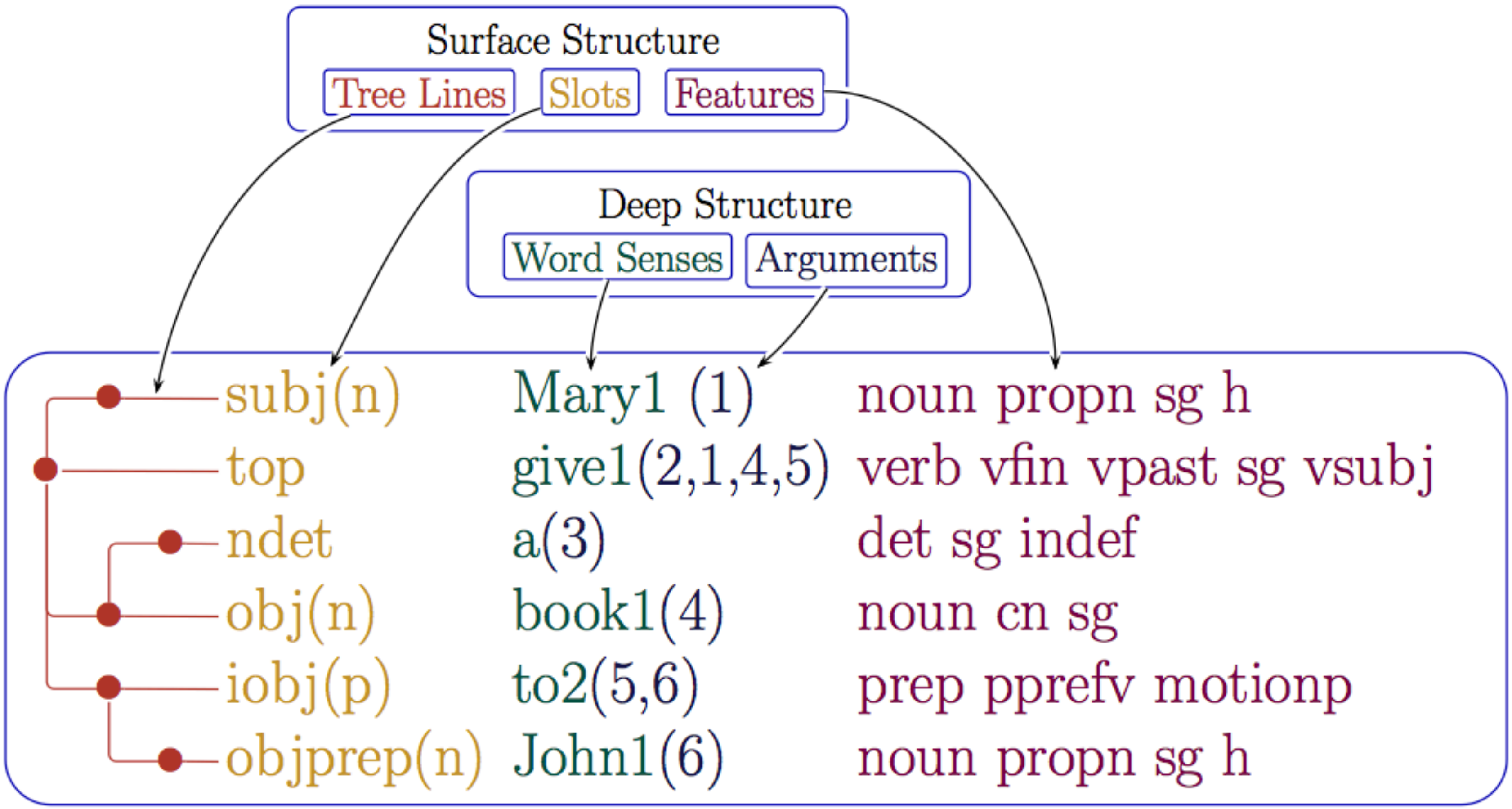
Thallium is said to look like this element



Extended Slot Grammar (ESG) Parser

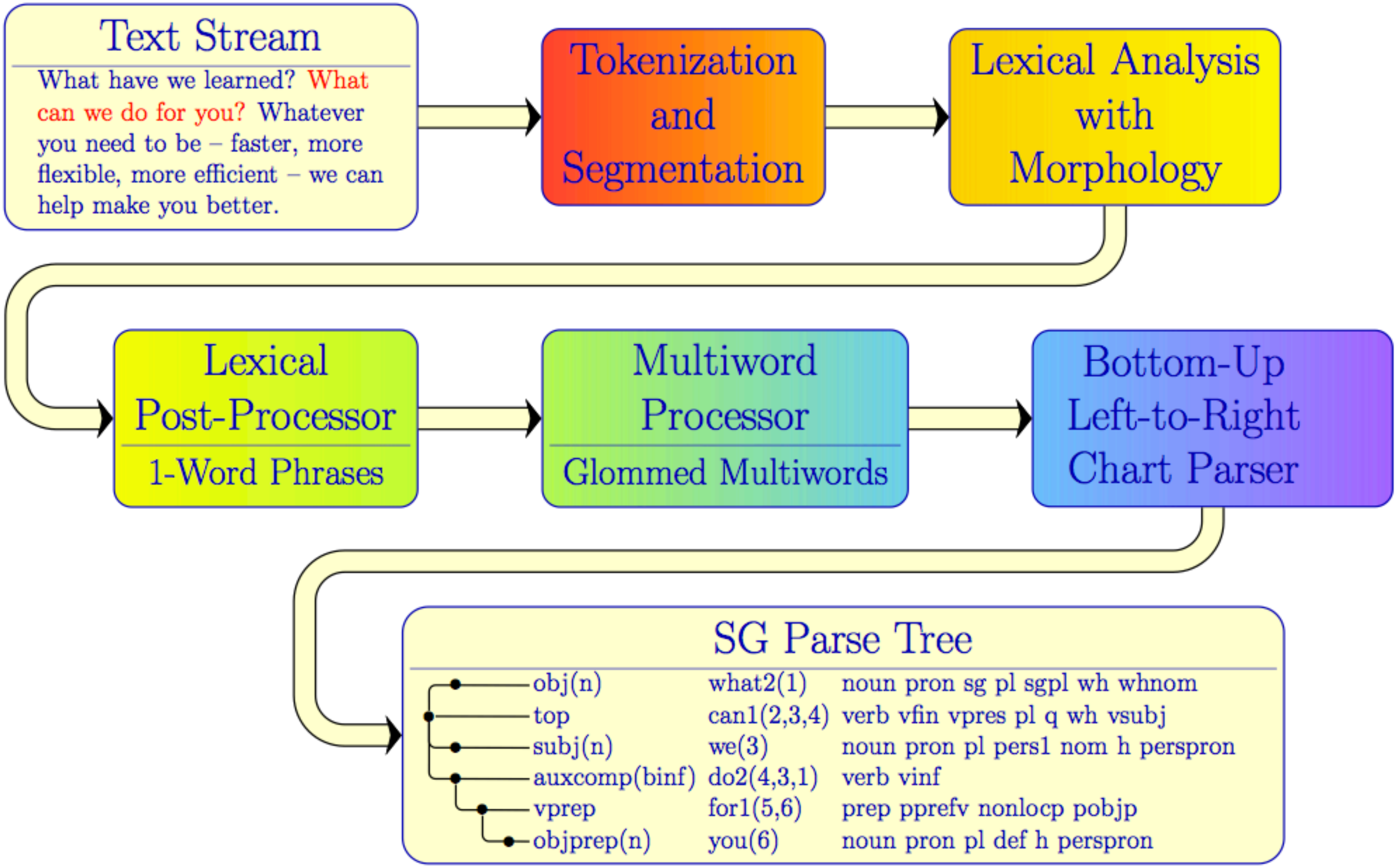
- The SG parser is a bottom-up, left-to-right chart parser.
- Rule-based (not statistical) – although it does use numerical scoring to arrive at most likely parses.
- SG can use multiple lexicons. There is always a main (base) lexicon, and there can be addendum lexicons
  - Easy domain adaptation
- The lexicons are indexed by citation forms (lemmas) of words. Morpholexical analysis.
- ESG base lexicon has about 88,000 entries, but many more word forms are recognized because of morphology and addendum lexicons.

Standard Slot Grammar Parse Display



Mary gave a book to John.

Pipeline for SG Parsing



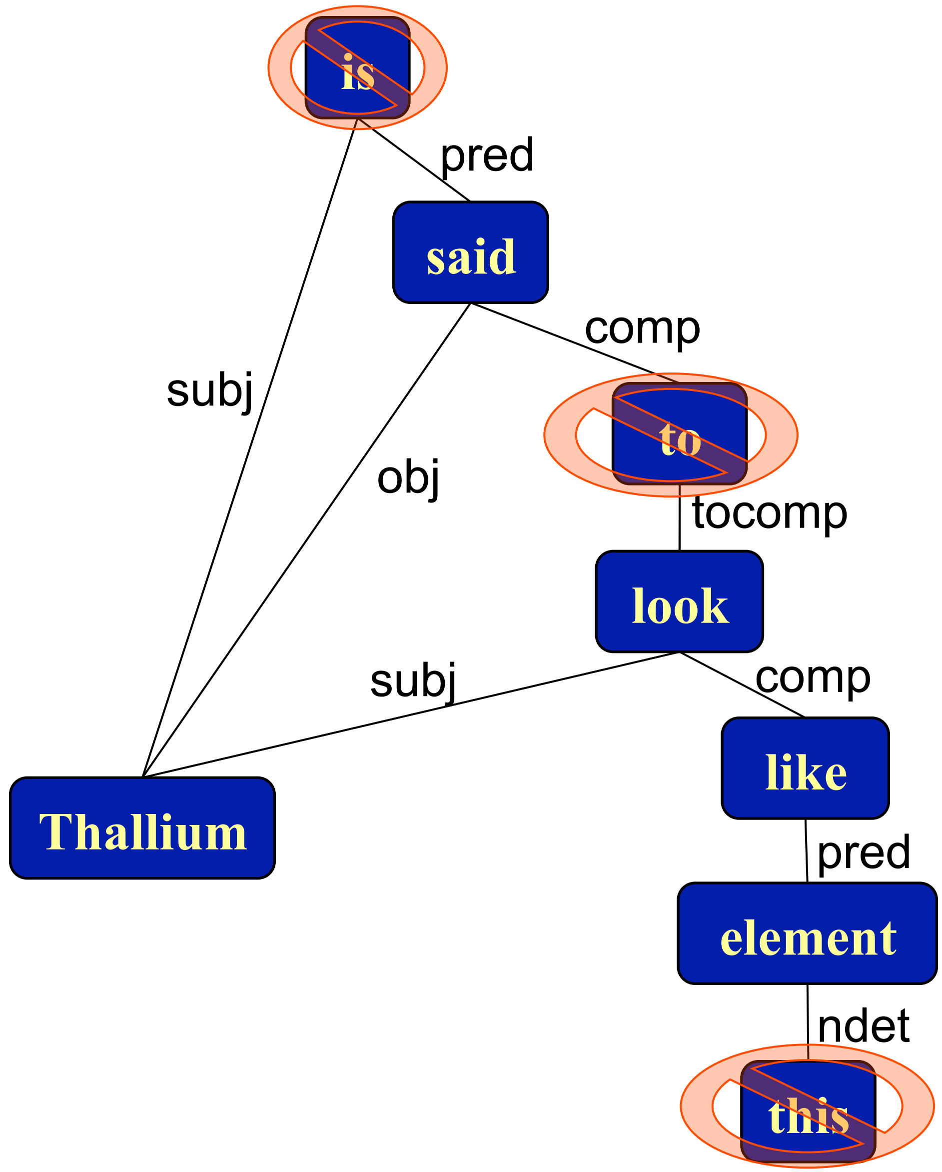
Parse of Jeopardy! Clue

This little bear of children's lit was named for the London locale in which he was found.

ndet	this(1)	det sg def
nadj	little(2,3)	adj erest
subj(n)	bear(3)	noun cn sg physobj anml
nprep	of(4,3,7)	prep pprefn nonlocp
nadj	child(5,u)	noun cn pl gen h relative
nposs	aposts(6)	special possmk
objprep(n)	lit(7)	noun cn sg discipline
top	be(8,3,9)	verb vpast sg auxv
pred(en)	name(9,u,3,10)	verb vpass
comp(p)	for(10,9,13)	prep nonlocp
ndet	the(11)	det sg def
nadj	London(12)	noun propn sg location cty govseat
objprep(n)	locale(13)	noun cn sg location
comp(lo)	in(14,18,15)	prep wh staticp
objprep(n)	which(15,13,u)	noun pron sg wh
subj(n)	he(16)	noun pron sg def nom m h perspron
nrel	be(17,16,18)	verb vpast sg wh auxv
pred(en)	find(18,u,16,14)	verb vpass

Predicate Argument Structure

Thallium is said to look like this element



Predicate Argument Structure: A layer of abstraction

Simplifies the ESG parse

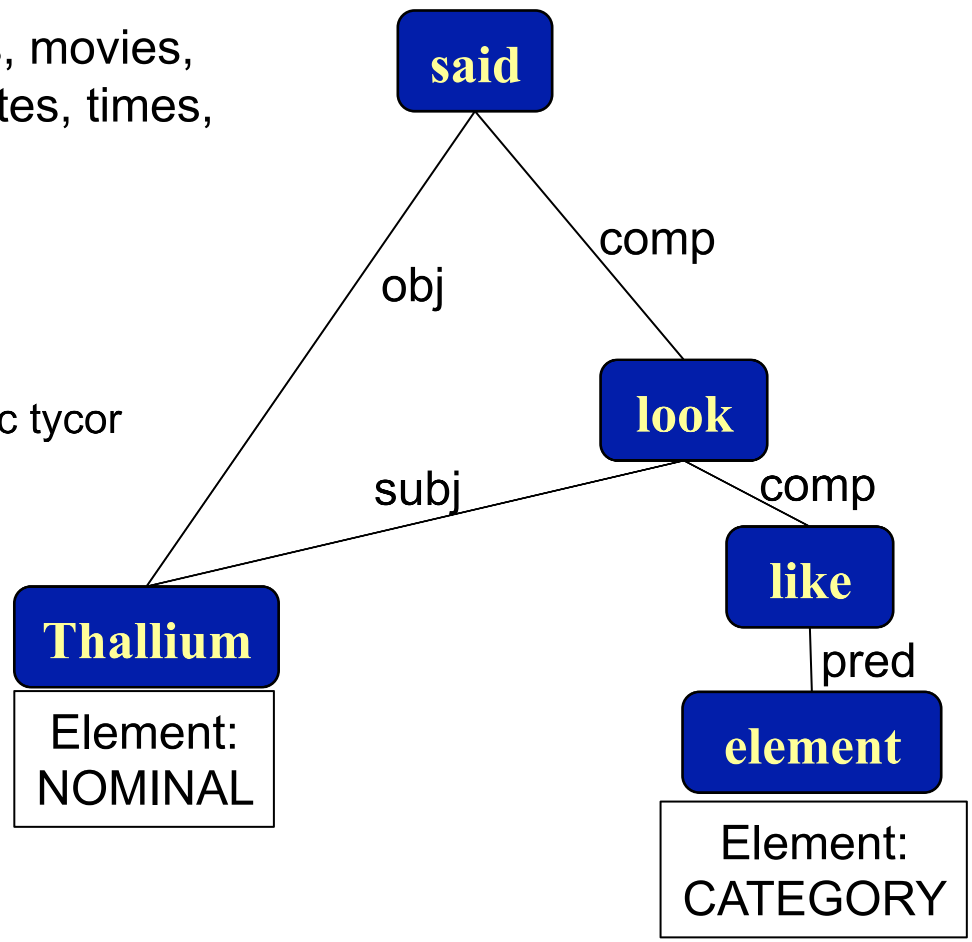
Does not add any new analysis

- Removes "unnecessary" nodes
- Simpler set of parts-of-speech
- Simplifies the encoding of lemma forms (no derivational morphology markers)
- Single dimension of links among parse nodes (no distinction between "deep" and "surface" structure)
- Both ESG and PAS have the usual suspects:
  - **noun**, **verb**, **adj** (adjective), **adv** (adverb), **prep** (preposition)
- ESG parse annotations have an uninflectedWord feature that corresponds roughly to the PAS lemmaForm feature

Named-Entity Detection

- Hundreds of entity types
- E.g., people, nations, cities, books, movies, weapons, musical instruments, dates, times, food, tools, elements
- Used in
  - Relation Detection
  - Answer lookup
  - Answer typing (positive in a specific tycor component)

Thallium is said to look like this element



- Extensible, Domain-Independent Named Entity Recognizer
- UIMA-based
  - Is itself a multi-annotator aggregate
- Primarily targets HUTT type system
  - About 400 types, about 170 of which in R2
- Recognition by lists, patterns, or both
  - Simple pattern file syntax, somewhat like regex

Relation Extraction

- Relation extraction: to classify the relation between two entity mentions into one of predefined relation classes
- Example:

  locatedAt? customerOf? employedBy?

  - "***The New Jersey Devils*** have signed ***Adam Larsson*** to a three-year, entry-level contract"
- Applications:
  - Information extraction
  - Database population
  - Machine reading
  - Question answering
  - Etc … …
- Challenges
  - Ambiguity: *Thomas Jefferson has signed the Declaration of Independence*
  - Context: … Nicole Kidman (1967)… vs. … Nicole Kidman (1990)…
  - Expressiveness of language:
    - IBM hired James, James started at IBM, James of IBM, …

Rule Based

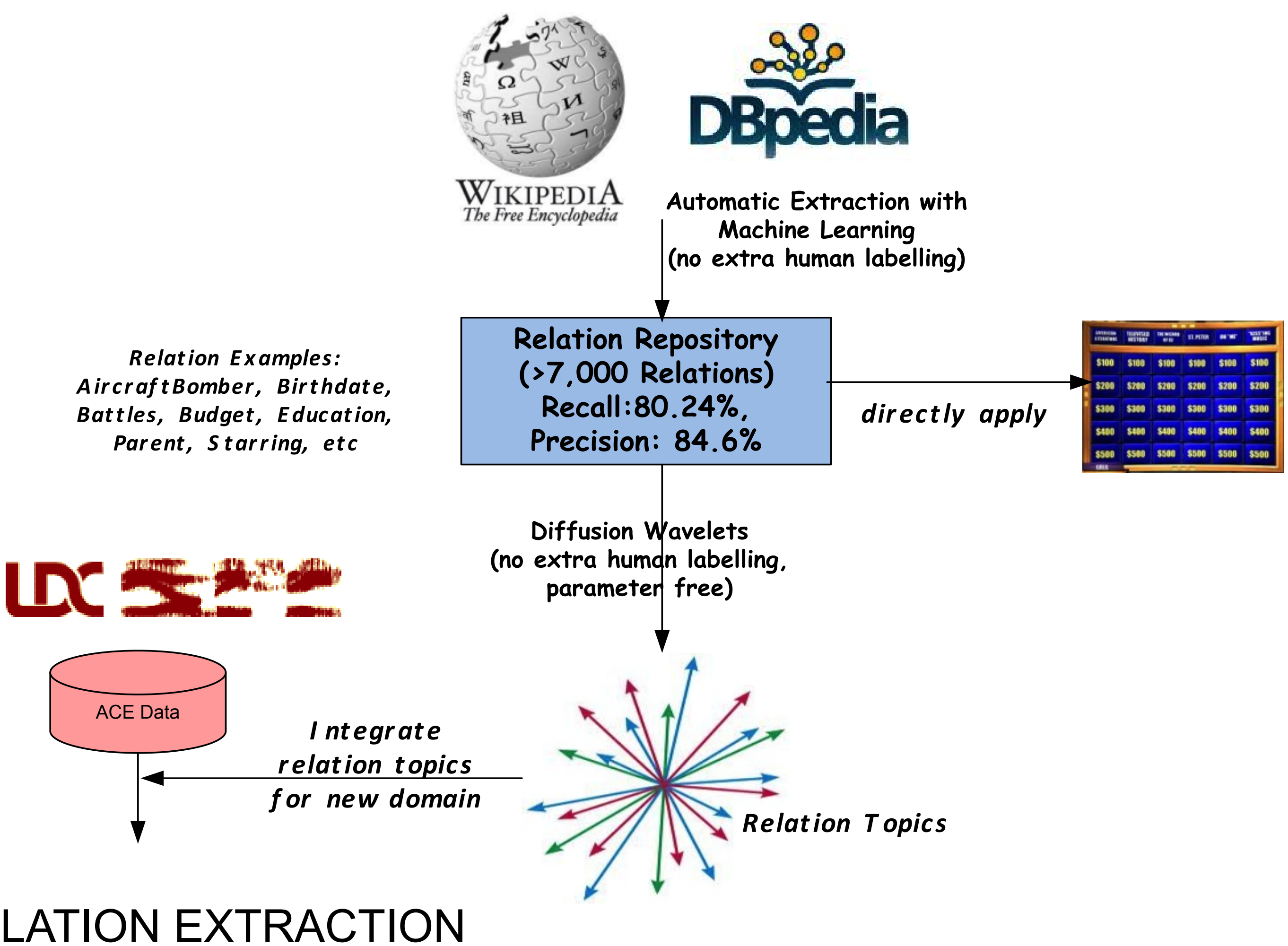
Prolog Rules

```
authorOf(Author,Composition) :-
  createVerb(Verb),
  subject(Verb,Author),
  author(Author),
  object(Verb,Composition),
  composition(Composition).

createVerb(Verb) :-
  partOfSpeech(Verb,verb),
  lemma(Verb,VerbLemma),
  member(VerbLemma,[“write”,”publish’
```

Inferred

```
author(1)
composition(3)
authorOf(1,3)
```

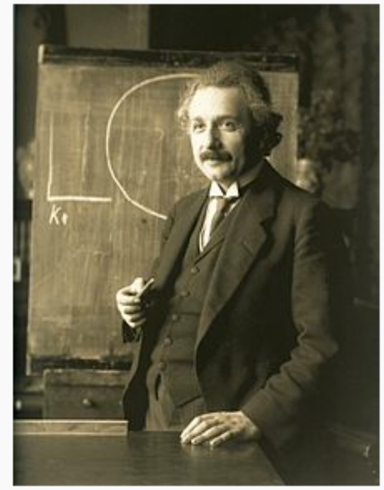
TWREX - Topicalized Wide Relation and Entity eXtraction



Distant supervision in TWREX

- Collecting training example from Wikipedia
- First occurrence heuristic on Dbpedia relations in Wikipedia,
  - For example, the Wikipedia page for "Albert Einstein" contains an infobox property "alma mater" with value "University of Zurich", and the first sentence mentioning the arguments is the following: "Einstein was awarded a PhD by the University of Zurich", which expresses the relation.

Albert Einstein



Albert Einstein in 1921

Born	14 March 1879 Ulm, Kingdom of Württemberg, German Empire
Died	18 April 1955 (aged 76) Princeton, New Jersey, United States
Residence	Germany, Italy, Switzerland, Austria, Belgium, United Kingdom, United States
Citizenship	Kingdom of Württemberg (1879–1896) Stateless (1896–1901) Switzerland (1901–1955) Austria–Hungary (1911–1912) German Empire (1914–1933) United States (1940–1955)

Early life and education

See also: Einstein family

Albert Einstein was born in Ulm, in the Kingdom of Württemberg in the German Empire on 14 March 1879.[10] His father was Hermann Einstein, a salesman and engineer. His mother was Pauline Einstein (née Koch). In 1880, the family moved to Munich, where his father and his uncle founded *Elektrotechnische Fabrik J. Einstein & Cie*, a company that manufactured electrical equipment based on direct current.[10]

Intra-Paragraph Anaphora Resolution (IPAR)

- Used in
  - Question Analysis to identify the Focus (and LATs)
  - Passage Scoring (LFACS, Skip Bigram)

Person: NAME

Person: PRONOUN

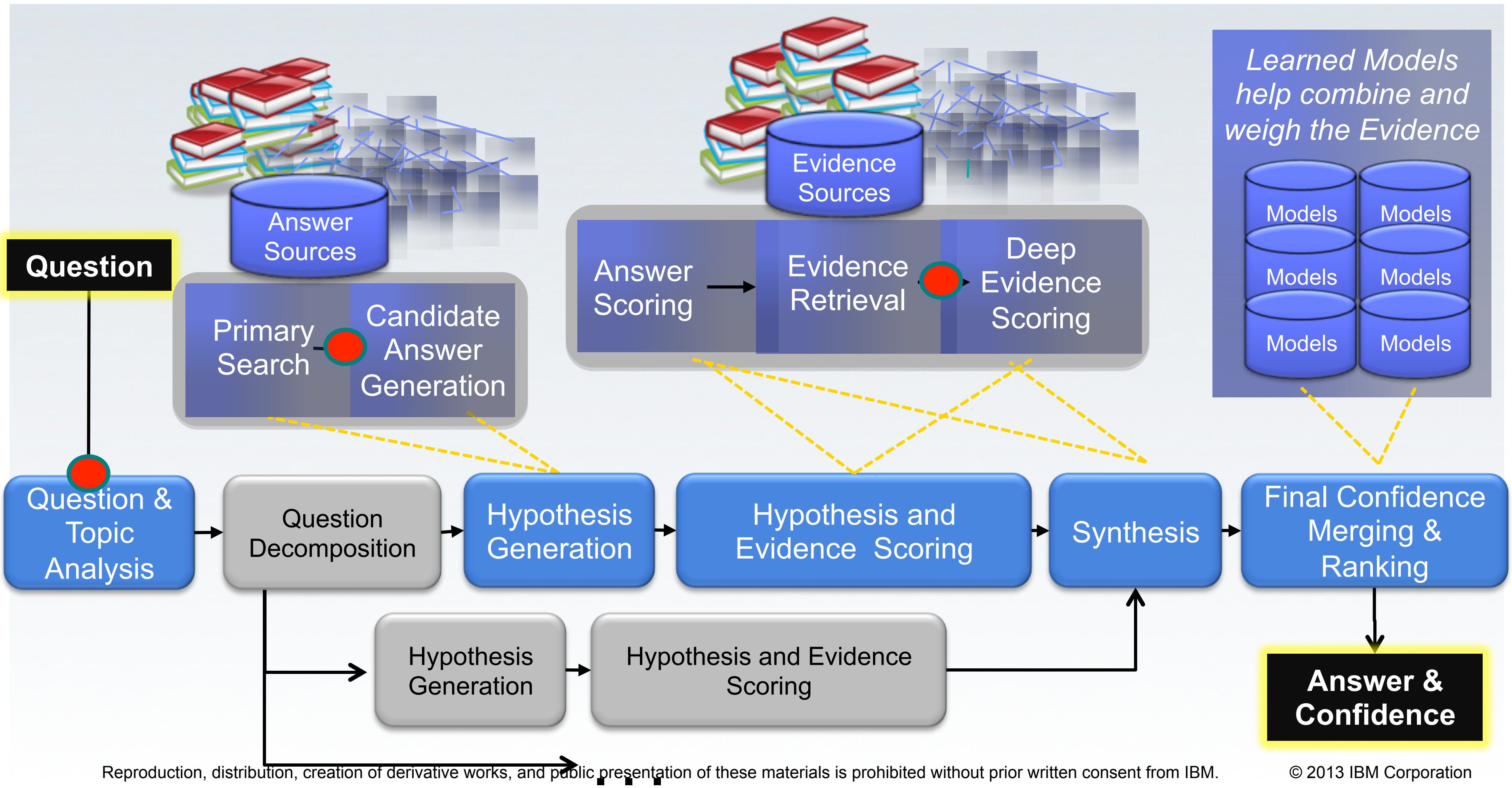
Vasco da Gama commanded a Portugese ship. He sailed more than 6000 miles over open ocean.

IPAR Operation

- Creates Co-reference chains between entities with compatible Hutt annotations
- Moves left-to-right within paragraph, at each Hutt entity (source) checks for compatibility with any other Hutt entity (target) to its left.
- Co-reference link made iff:
  - Type of target <= type of source
  - Gender of target = gender of source (if both known)
  - Number of target = number of source (if both known)
  - …

Where is the NLP stack run in the DeepQA pipeline?

- On questions, at the start of question analysis
- On primary search results, before candidate answer generation
- On supporting evidence, before deep evidence scoring



Outline

- The NLP Stack
- Question Analysis
- Passage Scoring

Question Analysis: what?

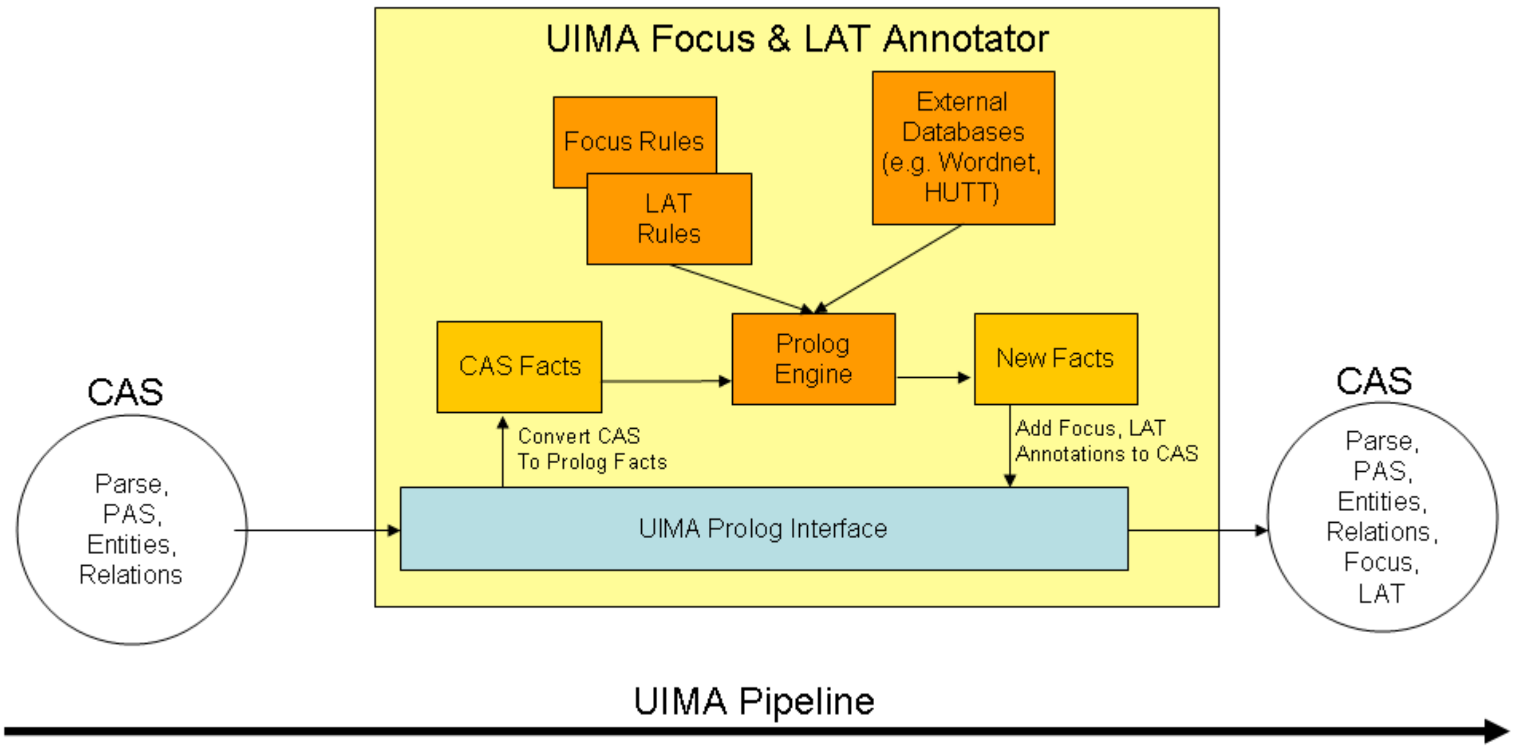
- *POETS & POETRY: He was a bank clerk in the Yukon before he published "Songs of a Sourdough" in 1907*
- *Identify*
  - *Focus:* part of the question that is a reference to the answer
    - *E.g. He*
  - *Lexical Answer Types:* terms in the question that indicate what type of entity is being asked for
    - "He," "clerk," and "poet"
  - *Question Classification:* Factoid (Most Jeopardy Questions), Definition, Multiple-Choice, Puzzle, Common-Bonds, Fill-in-the-blanks, and Abbreviation.
    - *Factoid*

Focus | LAT | FACTOID

Question Analysis

- Rule Based and statistical approaches over ESG Annotation
- Straightforward translation of CAS to Prolog facts
- *E.g. POETS & POETRY: He was a bank clerk in the Yukon before he published "Songs of a Sourdough" in 1907*

6000 prolog clauses



LAT detection rules

- A noun phrase with determiner “this” or “these”
  – THEATRE: A new play based on *this Sir Arthur Conan Doyle canine* ***classic*** opened on the London stage in 2007.
- “this” or “these” as a pronoun
  – ’88: In April 1988 Northwest became the first U.S. air carrier to ban ***this*** on all domestic flights
- One of the pronouns “he/she/his/her/him/hers”
  – OUT WEST: ***She*** joined Buffalo Bill Cody's Wild West Show after meeting him at the Cotton Expo in New Orleans
- One of the pronouns “it/they/them/its/their”
  – ME "FIRST"!: ***It*** forbids Congress from interfering with a citizen's freedom of religion, speech, assembly or petition
- The pronoun “one”
  – 12-LETTER WORDS (200): Leavenworth, established in 1895, is a federal ***one***
- When none of the above applies, the question may have no focus, as in:
  – MOVIE TITLE PAIRS: 1999: Jodie Foster & Chow Yun-Fat

Statistical approaches are used for LAT detection in combination with rules.

Question Classes

Looked at a sample of 500 questions and refined over time

Factoid is the default class

Some QC have different ML model

Some QC have different Candidate Generation

Some QC have different pipelines

QClass	Description	Example Questions (correct answer in parentheses)	Freq.
DEFINITION	A question that contains a definition of the answer	CONSTRUCTION: It can be the slope of a roof, or the gunk used to waterproof it. (pitch) CONSTRUCTION: The name of this large beam that supports the joists literally means "something that encircles". (a girder)	14.2%
CATEGORY-RELATION	The answer has a semantic relation to the question, where the relation is specified in the category.	FORMER STATE GOVERNORS: Nelson A. Rockefeller. (New York) COUNTRIES BY NEWSPAPER: Haaretz, Yedioth Ahronoth. (Israel)	7.2%
FITB	Fill-in-the-blank – question asks for completion of a phrase	COMPLETE IT: Attributed to Lincoln: "The ___ is stronger than the bullet". (ballot) SHAKESPEARE IN LOVE: "Not that I loved Caesar less", says Brutus, "but that I loved" this city "more" (Rome)	3.8%
ABBREVI-ATION	The answer is an expansion of an abbreviation in the question	MILITARY MATTERS: Abbreviated SAS, this elite British military unit is similar to the USA's Delta Force. (the Special Air Service)	2.9%
PUZZLE	A puzzle question - the answer requires derivation, synthesis, inference, etc.	BEFORE & AFTER: 13th Century Venetian traveler who's a Ralph Lauren short sleeve top with a collar. (Marco Polo shirt) THE HIGHEST-SCORING SCRABBLE WORD: Zoom, quiz or heaven. (quiz)	2.3%
ETYMOLO-GY	A question asking for an English word derived from a foreign word having a given meaning	ARE YOU A FOOD"E"?: From the Spanish for "to bake in pastry", it's South America's equivalent of a calzone. (an empanada)	1.9%
VERB	Question asks for a verb	THE NOT-SO-DEADLY SINS: To capitalize all text in an email is an abomination that signifies the person is doing this. (shouting)	1.5%
TRANS-LATION	A question asking for translation of a word or phrase from one language to another.	FRUITS IN FRENCH: Pomme. (apple)	1.1%
NUMBER	The answer is a number	YOU NEED TO CONVERT: One eighth of a circle equals this many degrees. (45)	1.0%
BOND	The question asks for what is in common between a set of entities	EDIBLE COMMON BONDS: Mung, snap, string. (bean)	0.7%
MULTIPLE-CHOICE	The question contains multiple possible answers from which to choose the correct answer.	THE SOUTHERNMOST CAPITAL CITY: Helsinki, Moscow, Bucharest. (Bucharest) OSCAR, GRAMMY OR BOTH: Mickey Rooney. (Oscar)	0.5%
DATE	A question asking for a date or year	THE TEENS: World War I ended in November of this year. (1918)	0.3%

Table 1: Question Classes

Question Analysis: Evaluation

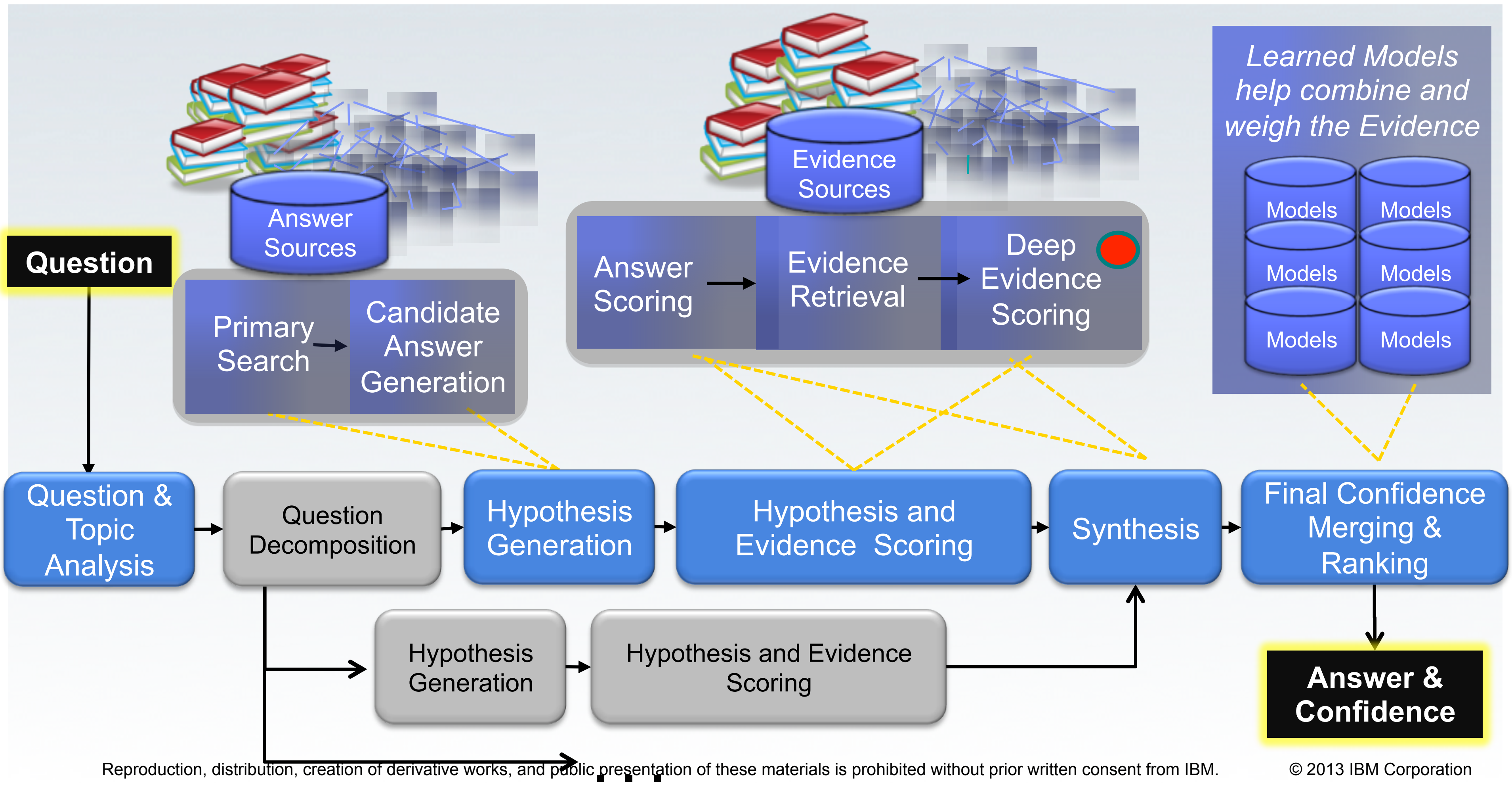
Component Level Evaluation	LAT Detection
Precision	0.829
Recall	0.766
F1	0.796
Per Question Recall	0.905

Question Classification	End To End Accuracy
No	68.1%
Yes	71.0%

Outline

- The NLP Stack
- Question Analysis
- Passage Scoring

Passage Scoring



Supporting Passage Retrieval (SPR)

Category: MICHIGAN MANIA

Clue: In 1894 C.W. Post created his warm cereal drink Postum in this Michigan city

In Deep Evidence Scoring, Watson retrieves evidence for each candidate answer, then evaluates the evidence using a large number of deep evidence scoring analytics. The evidence for a candidate answer may come from the original document or passage where the candidate answer was generated, or it may come from an evidence retrieval search performed by taking the keyword search query from Step 2, replacing the focus terms with the candidate answer, and retrieving the relevant passages that are found. The passages, or "context" in which the candidate answer occurs are evaluated as evidence to support or refute the candidate answer as the correct answer for the question.

Battle Creek

1895: In Battle Creek, Michigan, C.W. Post made the first POSTUM , a cereal beverage. Post created GRAPE-NUTS cereal in 1897, and POST TOASTIES corn flakes in 1908

The company was incorporated in 19
having developed from the earlier Postum Cereal Co. Ltd., founded by C.W. Post (1854-1914) in 1895 in Ba
Creek, Mich. After a number of experiments, Post marketed his first product-the cereal beverage called Postum-in 1895

Post Foods

Post Foods, LLC, also known as Post Cereals (formerly Postum Cereals) was founded by C.W. Post. It began in 1895 with the first Postum, a "cereal beverage", developed by Post in Battle Creek, Michigan. The first cereal, Grape-Nuts, was developed in 1897

r of
became General Foods. The cereal company unit was later sold off and is now Post Foods

General Foods

1854 C. W. Post (Charles William) was
born
in 18
in 19
beve

General Foods' products go from breakfast (Post's cereals) to warm nightcaps (Postum, Sanka), also wash the pots and pans that its foods are cooked in (S.O.S. Scouring Pads

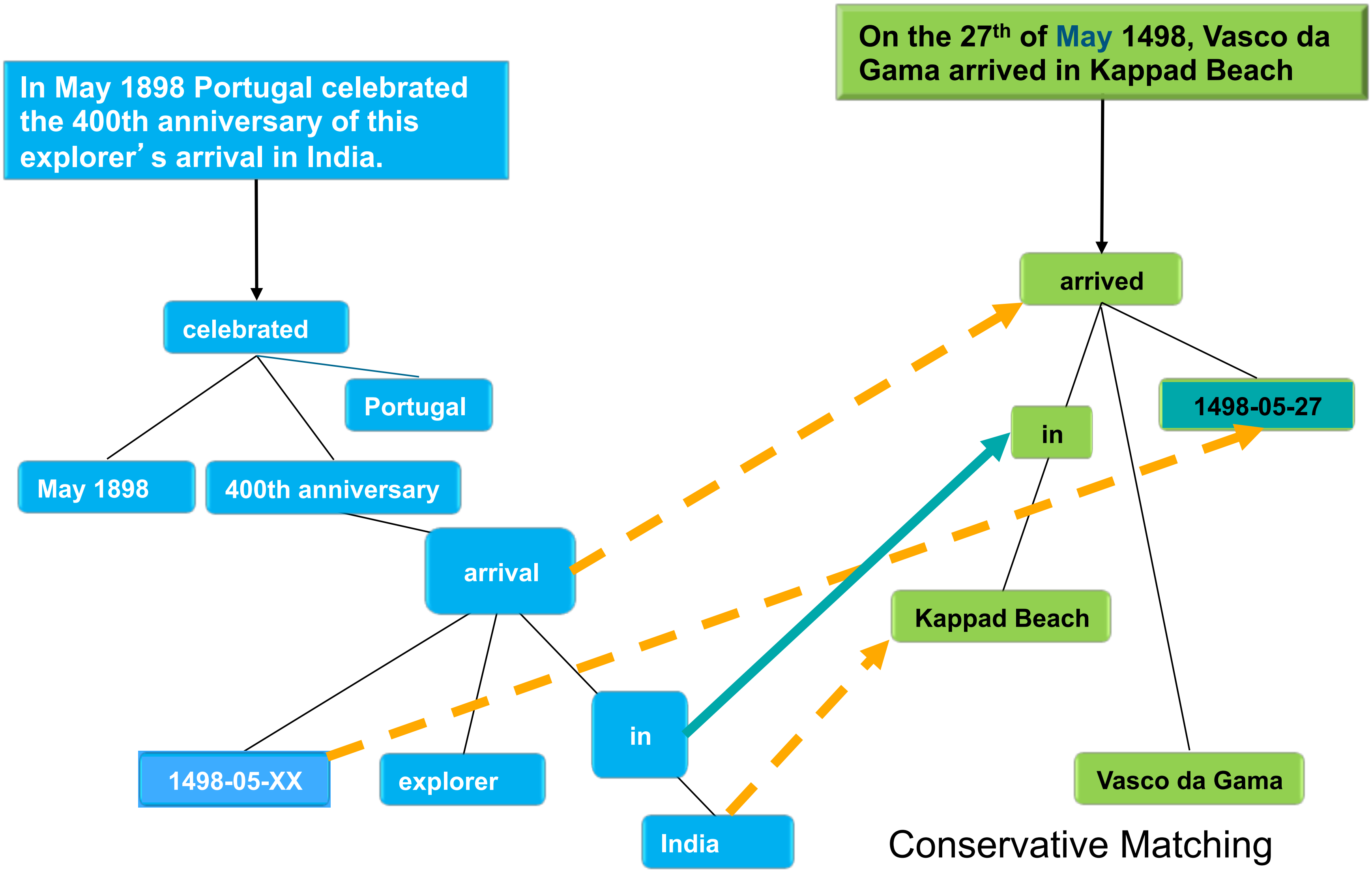
Passage scoring as a textual entailment problem

- In May 1898 Portugal celebrated the 400th anniversary of this explorer's arrival in India.
- In May 1898 Portugal celebrated the 400th anniversary of Vasco da Gama's arrival in India <- On the 27th of May 1498, Vasco da Gama landed in Kappad Beach
- In May 1898 Portugal celebrated the 400th anniversary of Gary's arrival in India <//- In May, Gary arrived in India after he celebrated his anniversary in Portugal.
- Textual Entailment is an open research issue
  – PASCAL Recognizing Textual Entailment Challenge (RTE-5) at TAC 2009
- State of the art approaches are still based on combining different similarity metrics
  – Kernel Methods
  – Edit distance
  – LSA similarity
  – Graph matching

Passage Scoring

- Apply multiple strategies to recognize textual entailment on SPR
- Passage Scoring Features
  - Passage Term Match
  - Textual Alignment
  - Skip Bigram
  - LFACS
  - LSA
  - String Kernel
- Feature are generated for each for each answer aggregating scores provided by passage scoring analytics
  - Average
  - Sum
  - Max
- Computationally expensive
  - 100 candidates per question = 2000 passages per question

Passage Term Match



Conservative Matching
Multiword, Exact String equality, any POS, Idf weighting on primary search corpus

Textual Alignment

IBM WATSON

In May 1898 Portugal celebrated the 400th anniversary of this explorer's arrival in India.

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

Not exactly, the same order, but pretty close, so Textual Alignment would give this a pretty good score, configured this way.

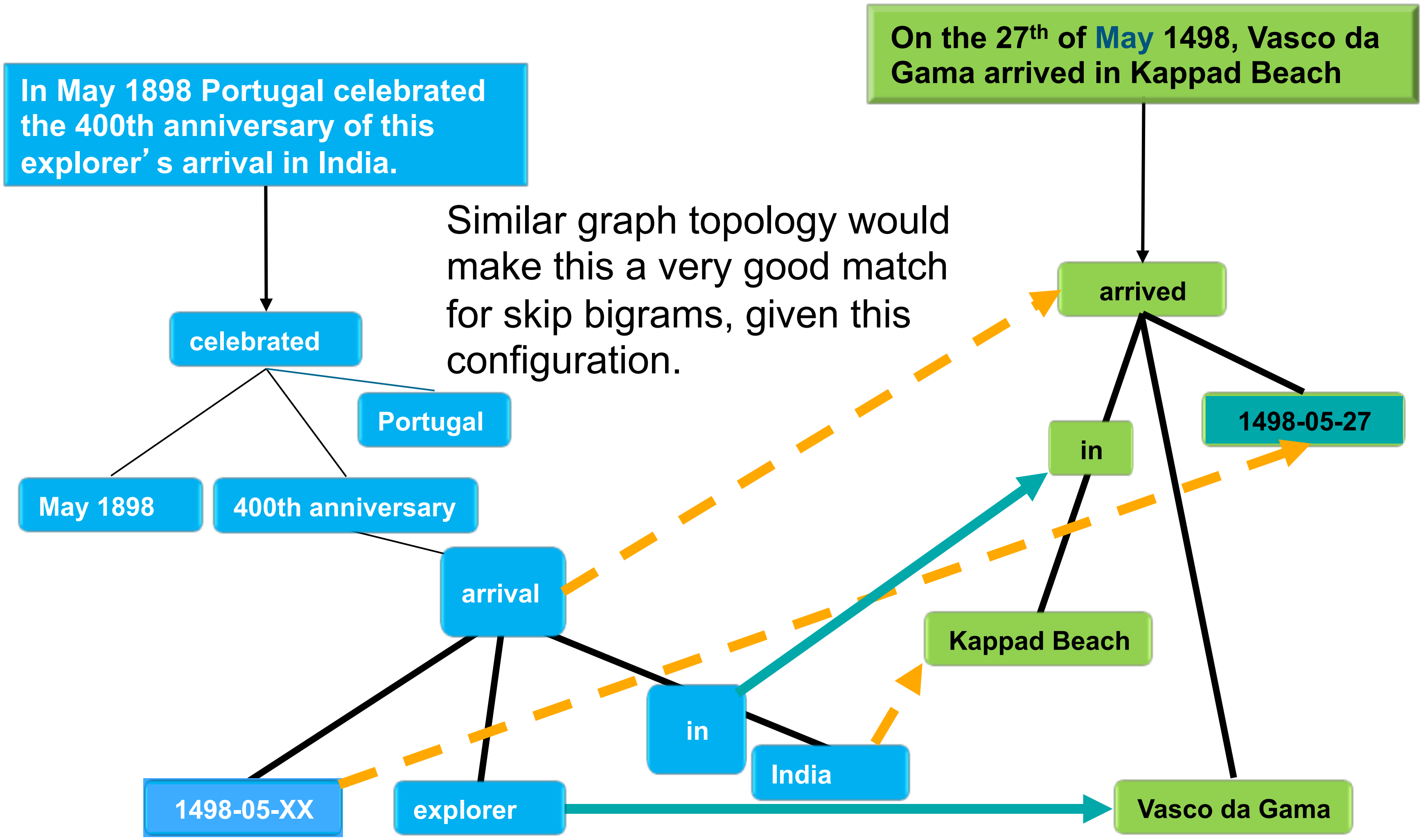
in
May 1898
Portugal
celebrated
400th anniversary
1498-05-XX
explorer
arrival
in
India

on
1498-05-27
Vasco da Gama
arrived
in
Kappad Beach

Strong Matching
Soft Matching

Order matters
Algorithm originally developed for gene sequences
Gap weighted score

Skip Bigram



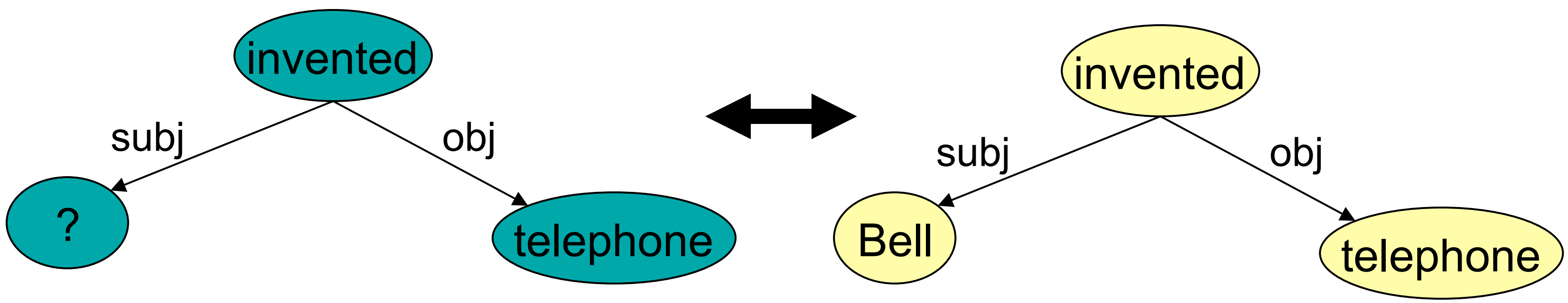
Similar graph topology would make this a very good match for skip bigrams, given this configuration.

Strong Matching
Soft Matching

Comparison with textual alignment
Dependency graph vs sequences
Gaps of length 1 vs Gap weighted

Logical Form Answer Candidate Scorer (LFACS)

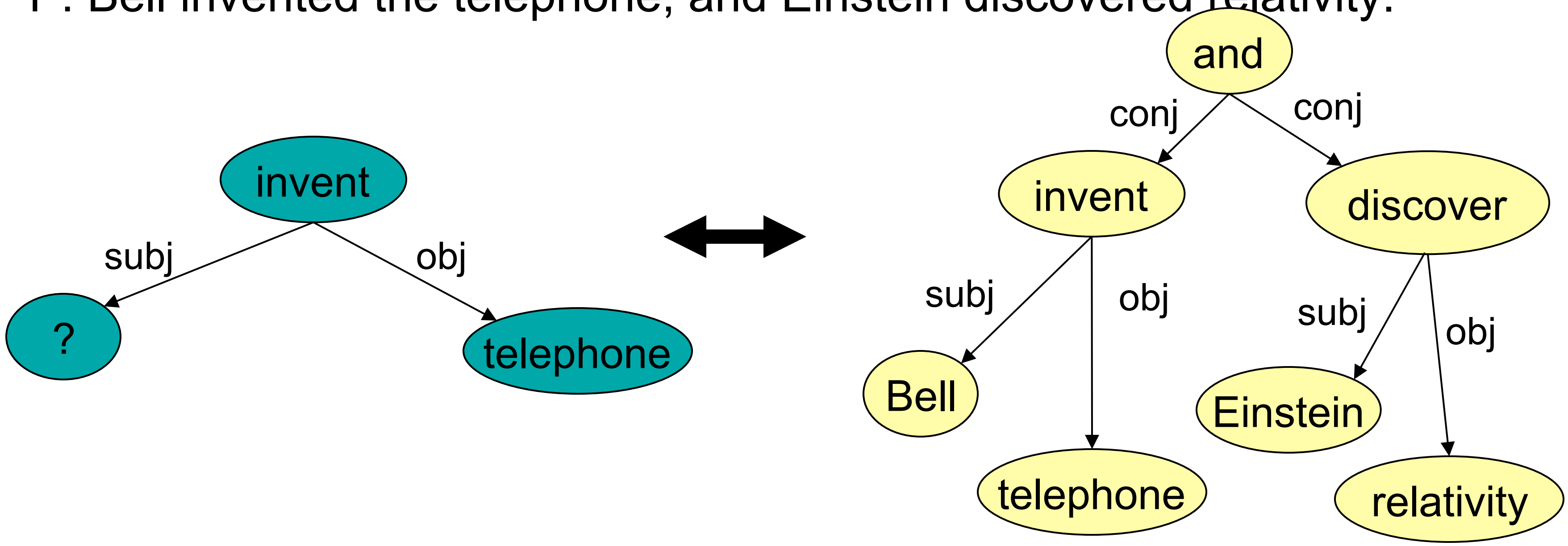
- LFACS tries to align a graph of the question to a graph of the passage:



- For complex domains (e.g., J!), there is virtually never a complete/perfect match.
- LFACS awards partial credit based on the extent to which it is able to align portions of the graph
- LFACS is part of a suite of four passage scoring algorithms (along with Passage Term Match, Textual Alignment, and Skip Bigram)

LFACS: Focus Centered Subgraph Matching

- LFACS aligns the focus to a specified candidate answer:
- Q: Who invented the telephone?
- P: Bell invented the telephone, and Einstein discovered relativity.

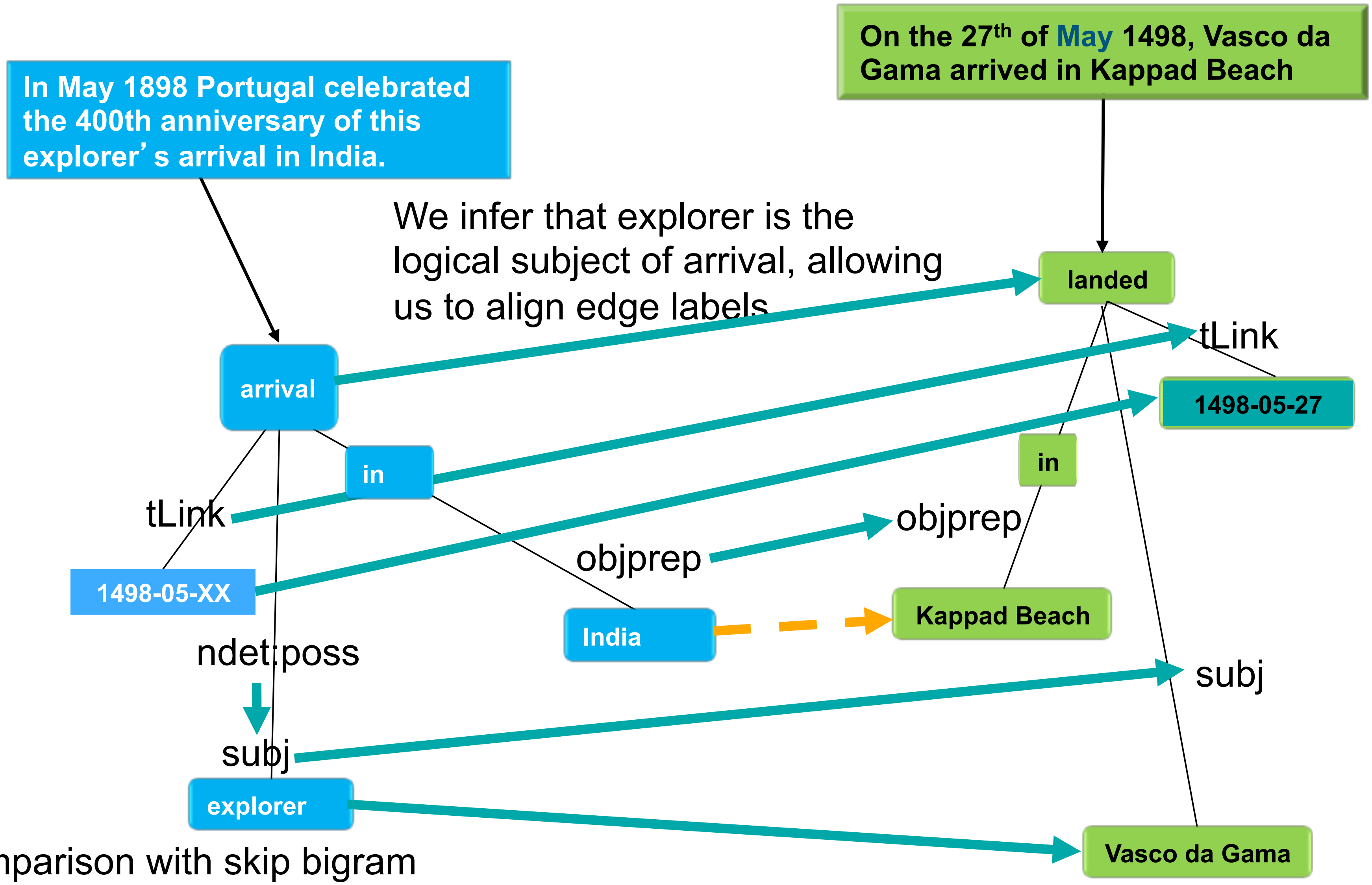


- Given this pair and the candidate answer "Bell", LFACS will give a high score (Bell is the subj of "invent" which has obj "telephone").
- Given this pair and the candidate answer "Einstein", LFACS will give a score of 0 because "Einstein" is not the subject of "invent"

How is the LFACS score computed

- Attempts to match the question graph to the passage graph with the restriction that the focus of the question align with the candidate the answer.
- Node matches are performed using a complex term matcher that can be configured with various matching resources, e.g., WordNet, Wikipedia redirects.
- Edge matches are performed using a simple edge matcher that has some sense of relations and subrelations.
- **LFACS score is the sum of the IDF values of the question nodes that matched some passage nodes, weighted by the degree of match**
- "Weighted by the degree of match" is a little complicated, because there are degree of match scores for edges and nodes and some nodes that match well are only connected via nodes that match poorly.

LFACS Example



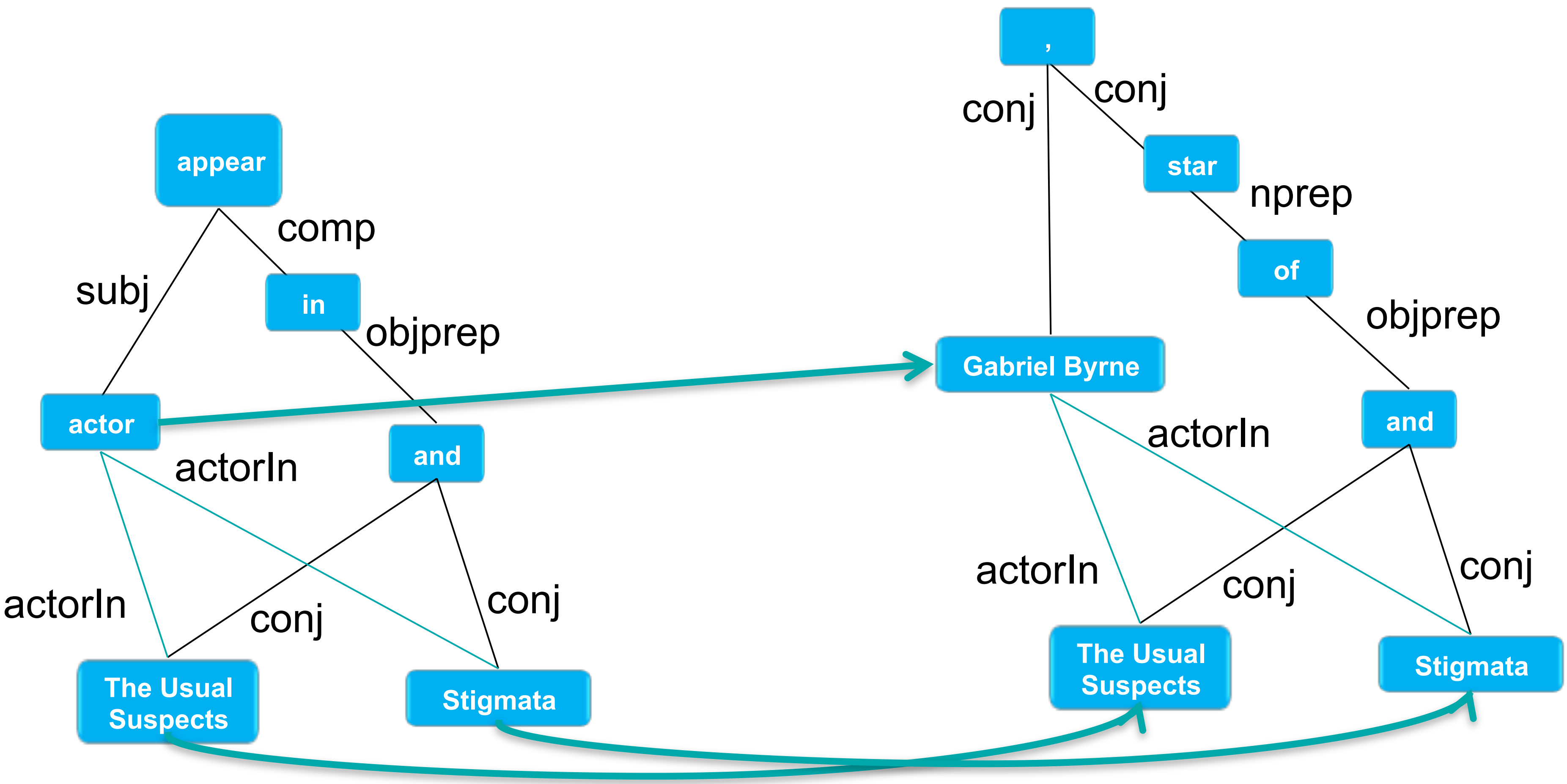
Comparison with skip bigram
Edges and nodes
No gaps
Focus is required

Strong Matching
Soft Matching

Semantic Relations in Logical Form Graphs

What actor appeared in "The Usual Suspects" and "Stigmata"?

Gabriel Byrne, star of "The Usual Suspects" and "Stigmata", …



Focus Centered Subgraph Matching is Precise but Brittle

- LFACS aligns the focus to a specified candidate answer:
- Q: Who invented the telephone?
- P: In later years, **Bell** described the **invention** of the **telephone** and linked it to his "dreaming place".
  - The passage doesn't say that Bell invented the telephone.
  - However, it is not a coincidence that the passage is talking about Bell, invention, and telephone.
  - It doesn't *prove* that Bell is the right answer, but it should be treated as *evidence* in favor of Bell being the right answer.
  - LFACS gives this passage a score of 0
- P: **Bell** is a famous **inventor**, best known for the **telephone**.
  - This passage does strongly imply that Bell invented the telephone.
  - However, "Bell" is still not the subject of the verb "invent" here. In fact there is no verb "invent"
  - LFACS gives this passage a score of 0
- P: **Bell invented** many devices including the **telephone**.
  - This passage states that Bell invented the telephone.
  - "Bell" is the subject of the verb "invent," but the object of "invent" is "devices"
  - LFACS gives this passage partial credit (for "invent" but not "telephone")

Kernels are similarity functions that can be applied to measure the similarity between two text

- Linear Kernel (BOW)
- Sequences (String Kernel, Word sequence kernel)
- Syntactic Structures (Tree Kernel)
- Similarity in a topic model (Domain Kernel, LSI)

A kernel is a function $K : \mathcal{X} \times \mathcal{X} -> \mathbb{R}$ such that

(18) $$K(x_i, x_j) = (def)\langle \mathbf{\Phi}(x_i), \mathbf{\Phi}(x_j)\rangle$$

where $\mathbf{\Phi} : \mathcal{X} -> \mathcal{K}$ is a feature mapping.

Kernel trick: equivalent (and more efficient) formulation in the instance space, avoid explicit feature mapping it counts the number of common subsequences of length p

(26) $$\Phi_u^p(s) = |\{\mathbf{i} : u = s(\mathbf{i})\}|, u \in \Sigma^p$$

(27) $$k_p(s,t) = \langle \Phi^p(s), \Phi^p(t) \rangle = \sum_{u \in \Sigma^p} \Phi_u^p(s) \Phi_u^p(t)$$

	c-a	c-r	a-r	c-t	a-t	c-u	u-t
Φ^2(car)	1	1	1	0	0	0	0
Φ^2(cat)	1	0	0	1	1	0	0
Φ^2(cut)	0	0	0	1	0	1	1

Gap Weighted Subsequence Kernel

It assigns different weights to sequences, according to how spread they are in the original strings

(28) $$\Phi_u^p(s) = \sum_{\mathbf{i}:u=s(\mathbf{i})} \lambda^{l(\mathbf{i})}, u \in \Sigma^p$$

$\lambda \in [0,1]$: When $\lambda = 1$ this kernel is equivalent to the fixed length subsequence kernel, if $\lambda -> 0$ it approximates the p-spectrum kernel

	c-a	c-r	a-r	c-t	a-t	c-u	u-t
ϕ(car)	λ^2	λ^3	λ^2	0	0	0	0
ϕ(cat)	λ^2	0	0	λ^3	λ^2	0	0
ϕ(cut)	0	0	0	λ^3	0	λ^2	λ^2

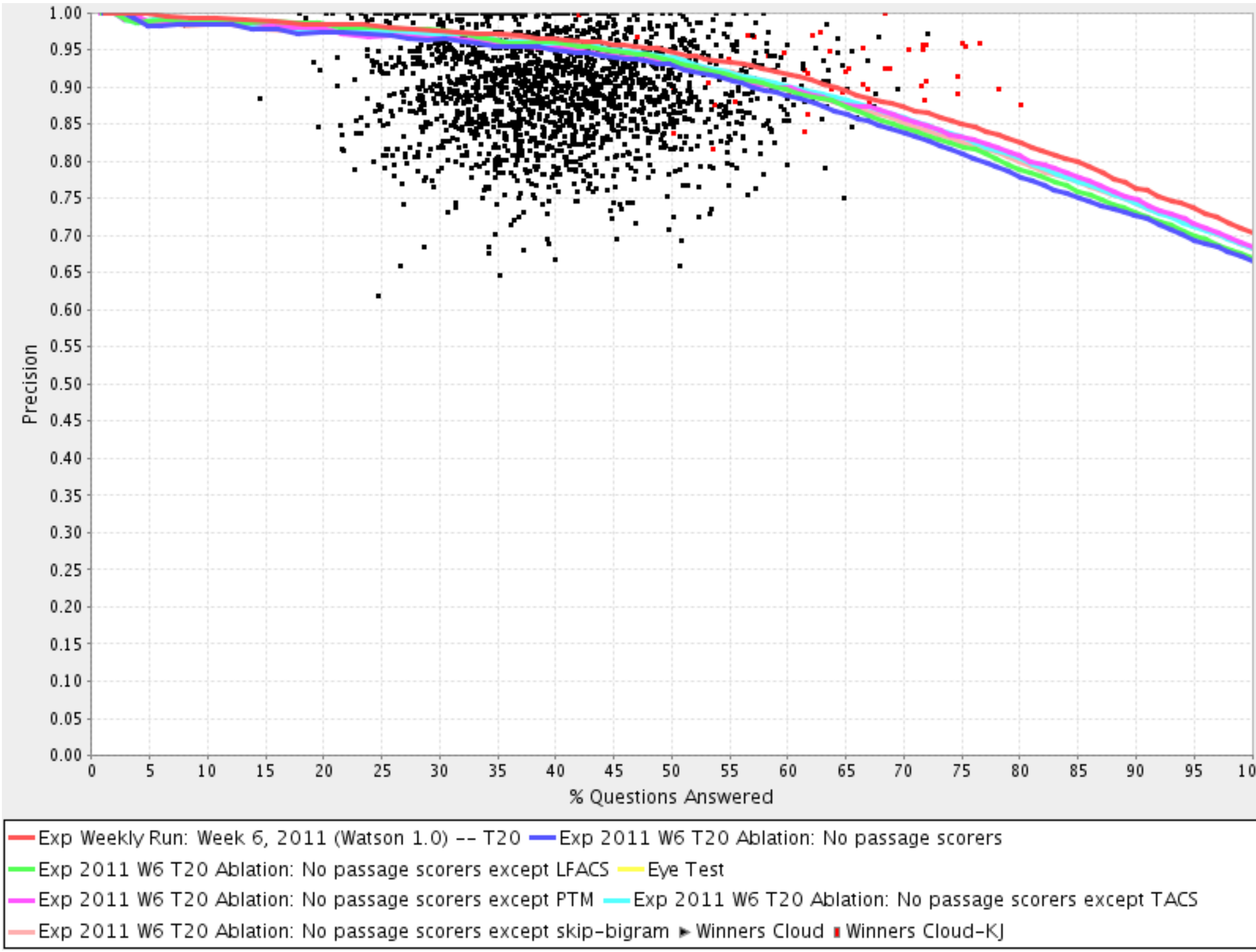
String Kernel Passage Scorer

- Measure the similarity between question and supporting passage where both focus and the candidate answer are replaced with a placeholder (FOCUS)
- Word Sequence Kernel (words are used instead of letters)
- Using ngrams of length 2 and 3
- Optimization lambda pruning (do not consider subsequences of span > k)
- Lemmatized forms
  - Q Who invented the telephone?
  - P1 **Bell invented** many devices including the **telephone**
- sim (FOCUS invent the telephone, FOCUS invent many device include the telephone)
  - Match (FOCUS invent _ telephone, FOCUS invent _ _ _ _ telephone)

LSA Passage Scorer

- Will be presented in Distributional Semantics Lesson

Passage Scoring evaluation



Outline

- The NLP Stack
- Question Analysis
- Passage Scoring

Semantic Technologies in IBM Watson^{TM}

Lesson 5 – Structured Knowledge in Watson

Professor: Alfio Massimiliano Gliozzo
TA: Or Biran

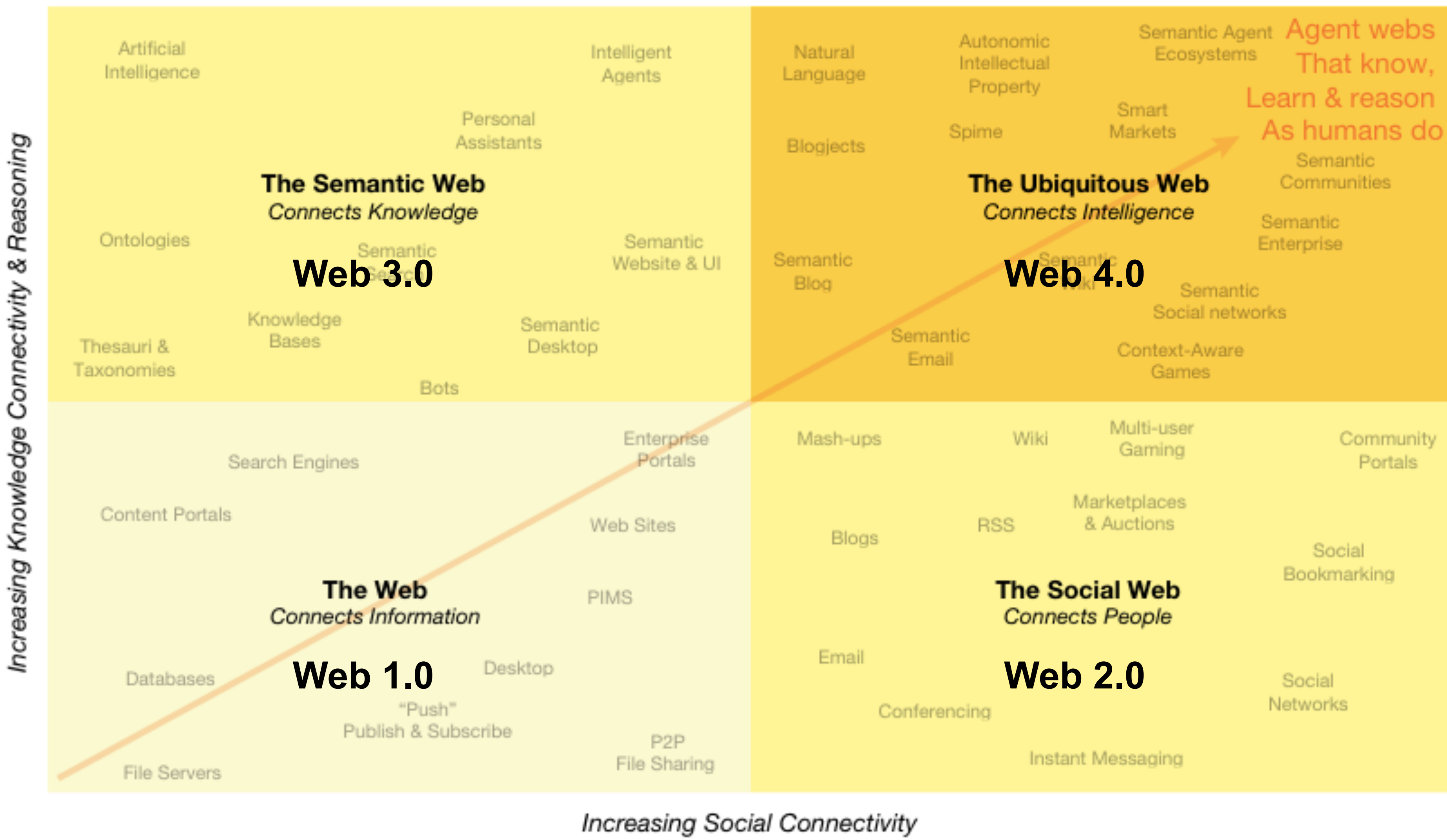
Outline

- Introduction to the Semantic Web
- Linking Language to Knowledge in Watson
- Type Coercion
- Temporal and Spatial Reasoning
- Prismatic

Outline

- Introduction to the Semantic Web
- Linking Language to Knowledge in Watson
- Type Coercion
- Temporal and Spatial Reasoning
- Prismatic

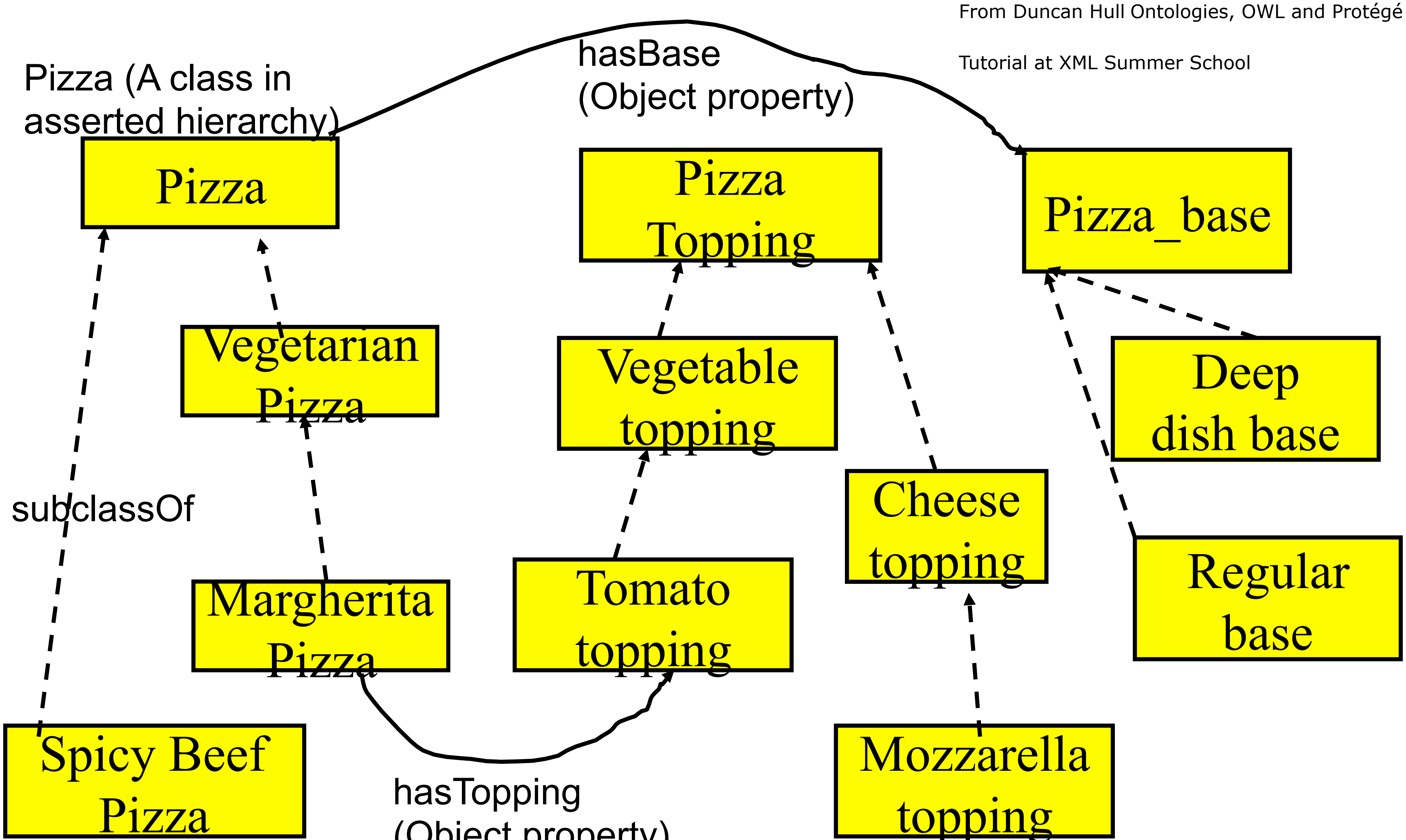
Evolution of the WEB:



Ontology in Computer science

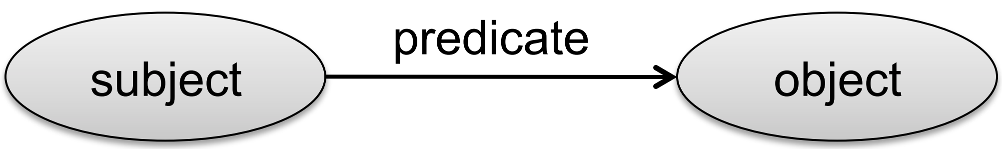
- "An ontology is a formal representation of a set of concepts within a domain and the relationships between those concepts. It is used to reason about the properties of that domain."
- Tbox vs Abox
  - Assertions about Classes (terminology)
    - represented in the Tbox
    - similar to a database schema
    - E.g. City capital-of Country
  - Assertions about Instances
    - Represented in the Abox
    - Similar to data in a database
    - Italy instance-of Country
    - Rome capital-of Italy
- Knowledge Base = Abox + Tbox

The Pizza Ontology



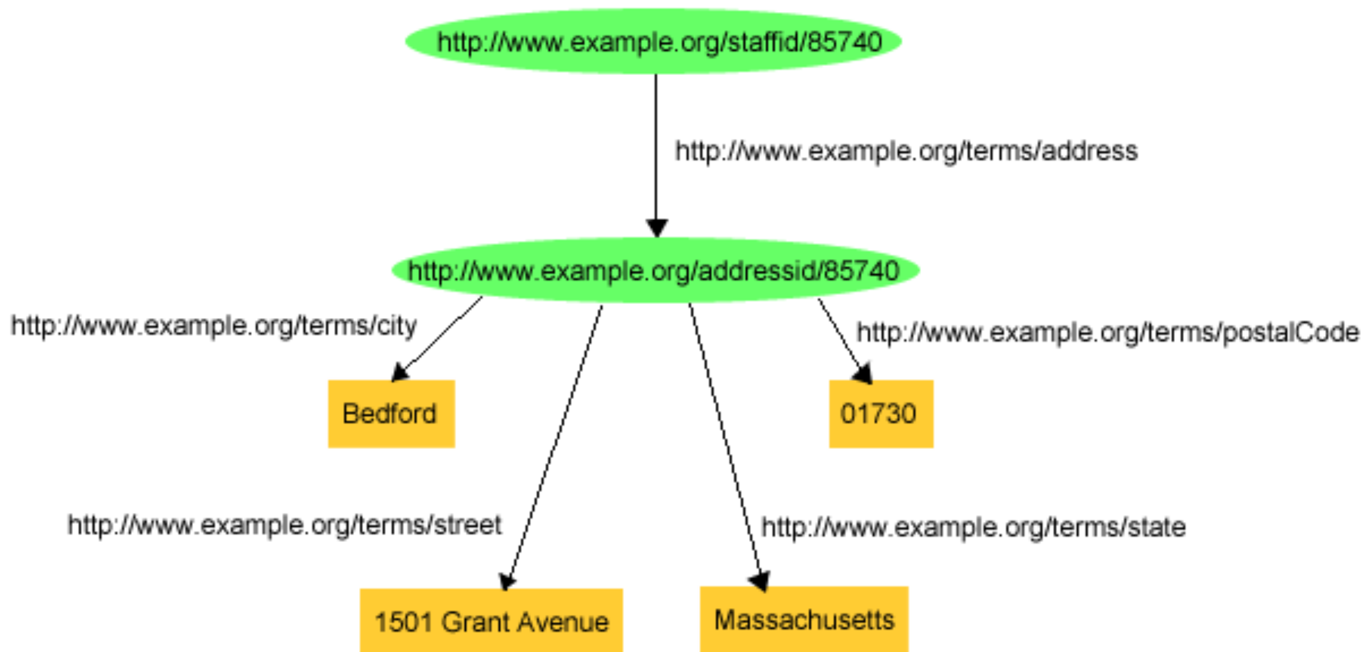
Resource Description Framework (RDF)

- The data model of the Semantic Web.
- A schema-less data model that features unambiguous identifiers and named relations between pairs of resources.
- A labeled, directed graph of relations between resources and literal values.
  - RDF graphs are collections of triples (assertions)
  - Triples are made up of a subject (node), a predicate (property), and an object (node)
  - The direction of the arc is significant: it always points toward the object.
  - Resources and relationships are named with Unique Resource Identifiers (URIs)



- The RDF Specification consists of a suite of W3C Recommendations published in 2004.
  - http://www.w3.org/TR/2004/REC-rdf-concepts-20040210/
- Ontology Web Language (OWL) is a specification of RDF to represent kowledge bases Abox and TBox

RDF Graph: example

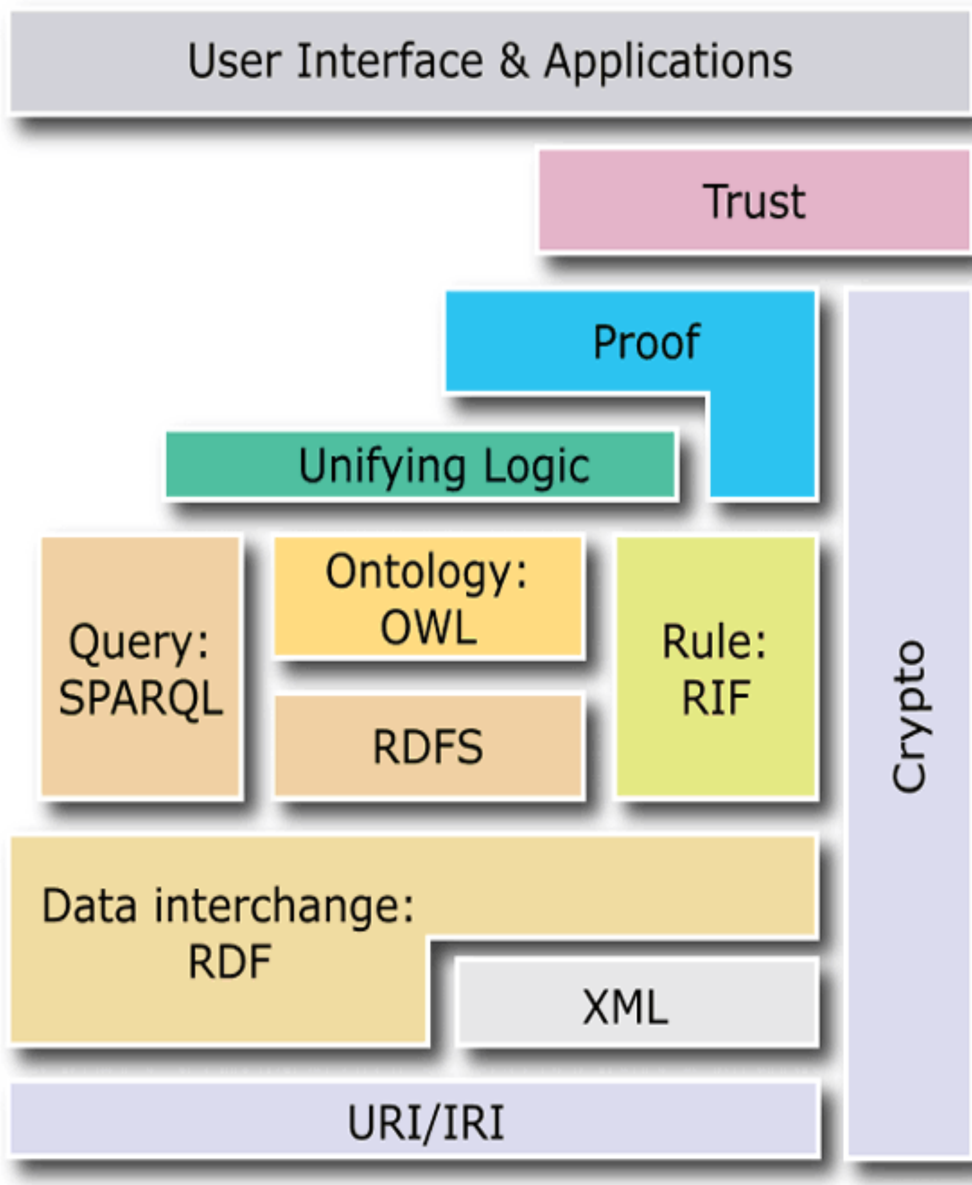


exstaff:85740	exterms:address	exaddressid:85740 .
exaddressid:85740	exterms:street	"1501 Grant Avenue" .
exaddressid:85740	exterms:city	"Bedford" .
exaddressid:85740	exterms:state	"Massachusetts" .
exaddressid:85740	exterms:postalCode	"01730" .

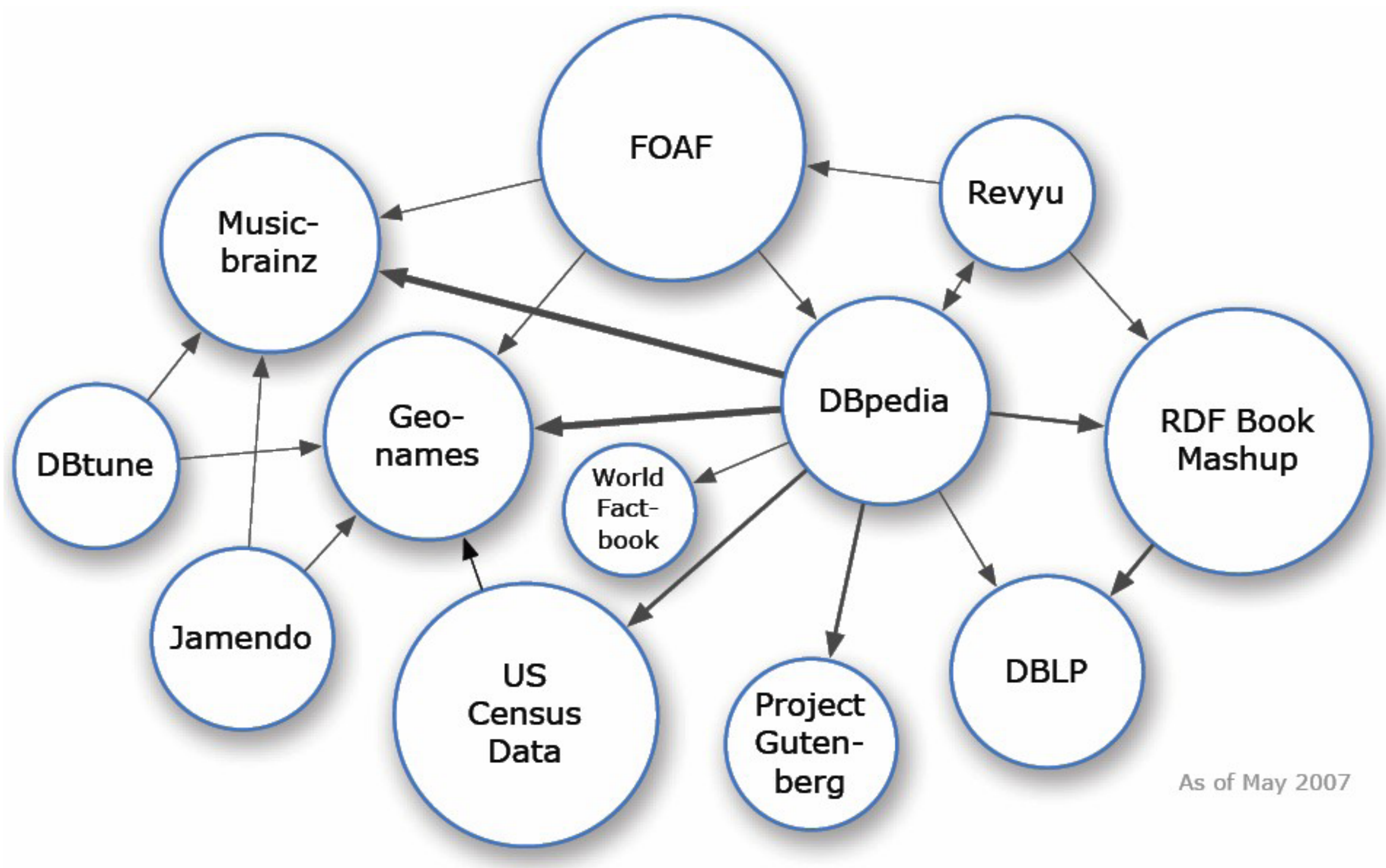
The Semantic Web Stack

- Augments the World Wide Web with machine-readable semantic data
- Integrated family of technology standards
  - Flexible data model
  - Expressive ontology language
  - Distributed query language

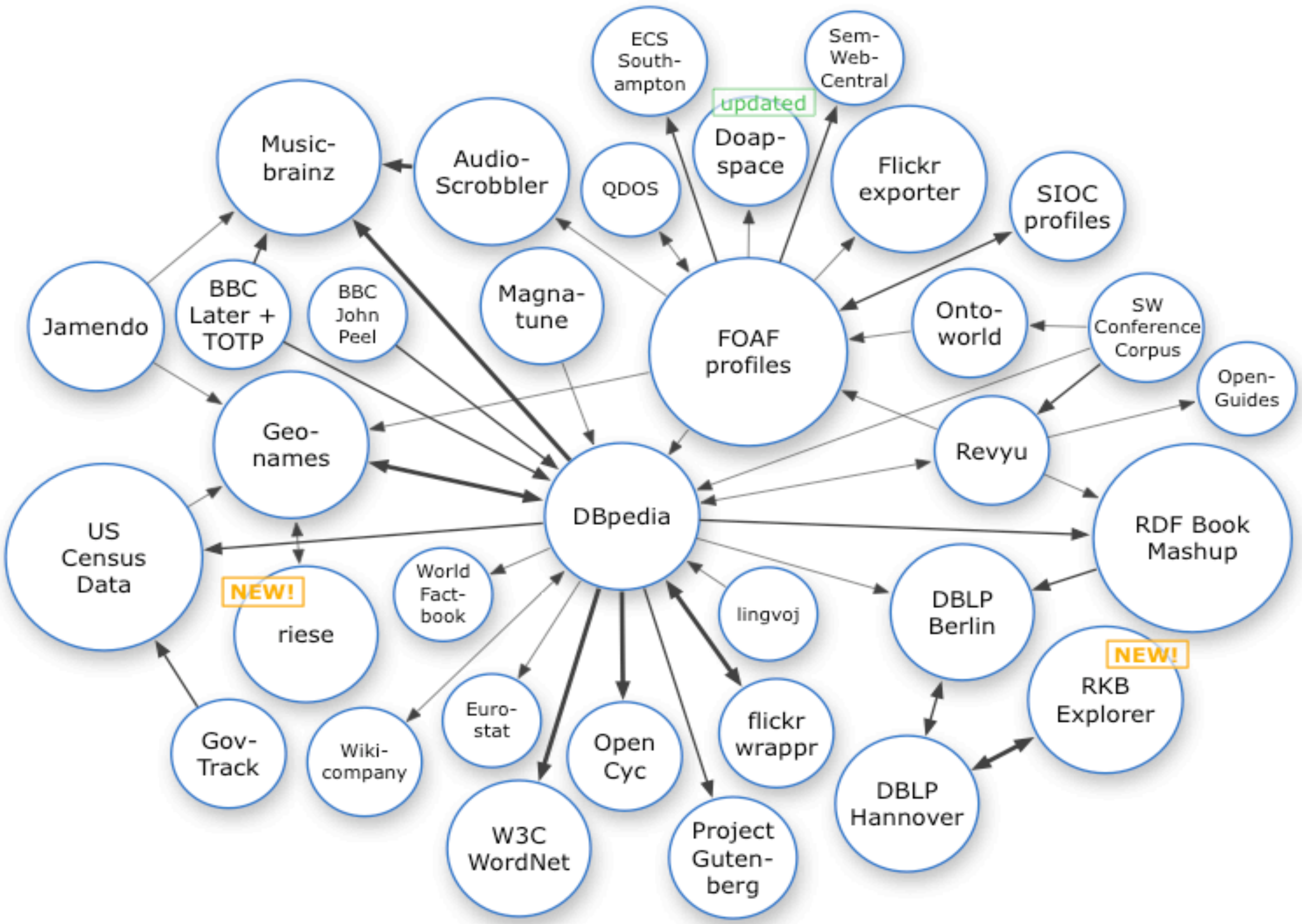
W3C Semantic Web



Linking Open Data (May 2007)

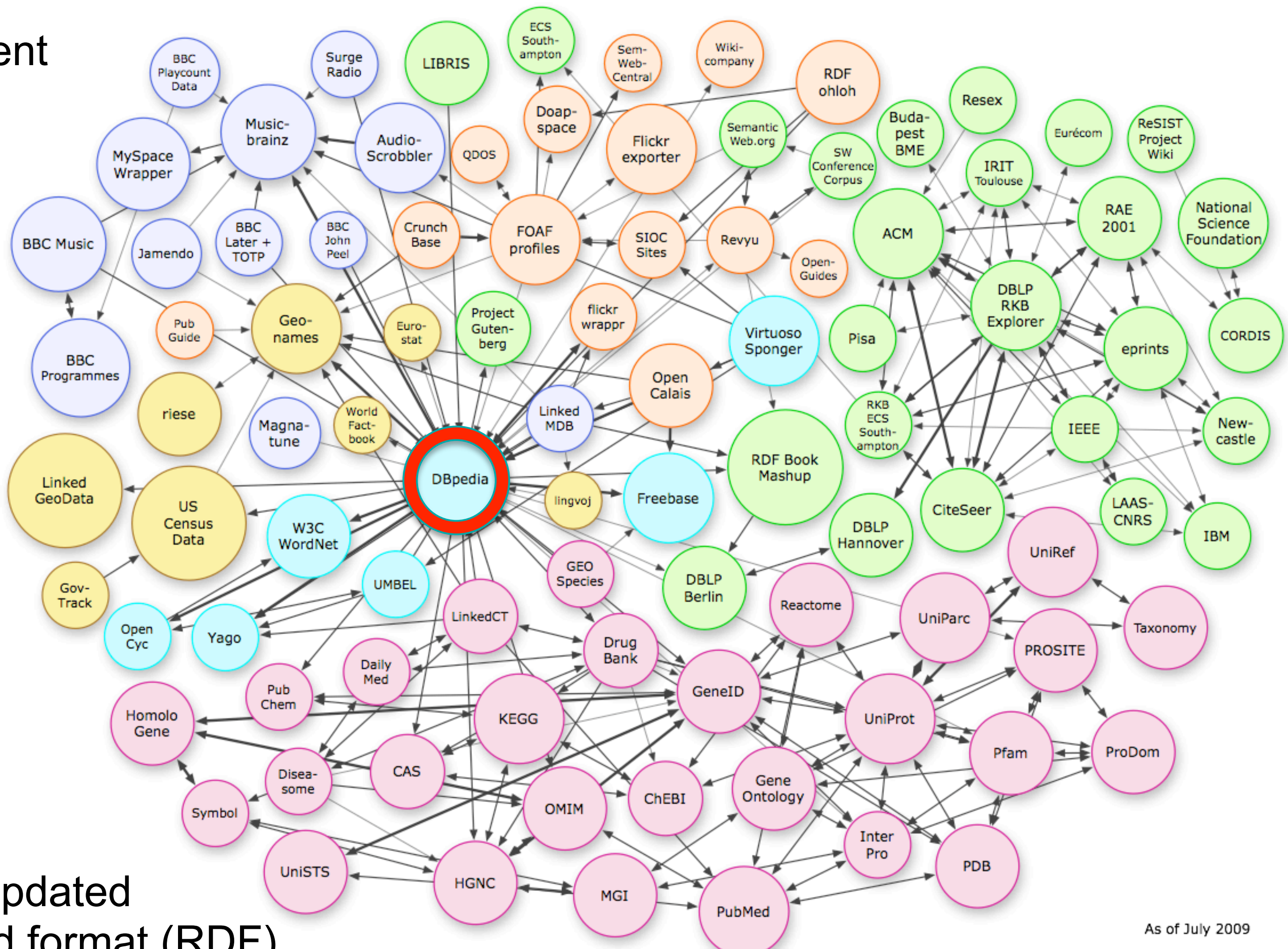


Linking Open Data (April 2008)



Linking Open Data (July 2009)
DBPedia is the conceptual hub …

- 100M hours of manual labor creating content



- Constantly updated
- Standardized format (RDF)

DBPedia is bridging text and knowledge

Wikipedia Article
http://en.wikipedia.org/wiki/Google

DBPedia Ontology
http://dbpedia.org/page/Google

The **Google headquarters**, the Googleplex, **is located in Mountain View, California**.

…

Google was founded by Larry Page and Sergey Brin while they were students at Stanford University and **the company was first incorporated as a privately held company on September 4, 1998.**

Property	Value
dbpedia-owl:assets	▪ 31768000000
dbpedia-owl:equity	▪ 28239000000
dbpedia-owl:foundationdate	▪ 1998-09-04 (xsd:date)
dbpedia-owl:foundationperson	▪ dbpedia:Larry_Page ▪ dbpedia:Sergey_Brin
dbpedia-owl:foundationplace	▪ dbpedia:California ▪ dbpedia:Stanford_University ▪ dbpedia:Menlo_Park%2C_California
dbpedia-owl:industry	▪ dbpedia:Computer_software ▪ dbpedia:Internet
dbpedia-owl:keyPerson	▪ dbpedia:Larry_Page ▪ dbpedia:Sergey_Brin ▪ dbpedia:CEO ▪ dbpedia:Eric_E._Schmidt ▪ dbpedia:Chairman ▪
dbpedia-owl:keyPersonPosition	▪ Eric E. Schmidt ▪ Larry Page ▪ Sergey Brin
dbpedia-owl:locationcity	▪ dbpedia:California ▪ dbpedia:Googleplex ▪ dbpedia:Mountain_View%2C_California

Text Enrichment

Your content enhanced!

Designing a computer that can process and understand natural language.

IBM is working to build a computing system that can understand and answer complex questions with enough precision and speed to compete against some of the best Jeopardy! contestants out there.

This challenge is much more than a game. Jeopardy! demands knowledge of a broad range of topics including history, literature, politics, film, pop culture and science. What's more, Jeopardy! clues involve irony, riddles, analyzing subtle meaning and other complexities at which humans excel and computers traditionally do not. This, along with the speed at which contestants have to answer, makes Jeopardy! an enormous challenge for computing systems.

Code-named "Watson" after IBM founder Thomas J. Watson, the IBM computing system is designed to rival the human mind's ability to understand the actual meaning behind words, distinguish between relevant and irrelevant content, and ultimately, demonstrate confidence to deliver precise final answers.

IN-TEXT LINKS

natural language | pop culture | Thomas J. Watson | Jeopardy! | IBM | computers | human mind | Watson | politics | humans

APPLY ALL

TAGS APPLY ALL

IBM | Jeopardy | Thomas J. Watson | Natural language | Watson | Popular culture | Mind | Hardware

IBM, Jeopardy, Thomas J. Watson, Natural language, Watson, Popular culture, Mind, Hardware

Content recommendations

Zemanta

REFINE UPDATE

MEDIA GALLERY

RELATED ARTICLES

IBM - Research: Jeopardy!
3 WEEKS AGO RESEARCH.IBM.COM

This supercomputer plans to beat humans on
3 WEEKS AGO DVICE.COM

IBM's in "Jeopardy"
3 WEEKS AGO TECHNOLOGIZER.COM

IBM's Watson is really smart, will try to prove it
3 WEEKS AGO ENGADGET.COM

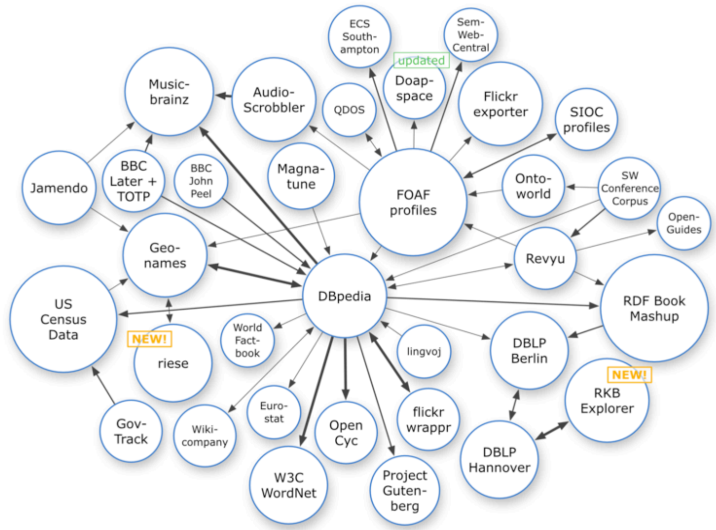
IBM's 'Watson' Takes on Jeopardy! You Can
3 WEEKS AGO SINGULARITYHUB.COM

Outline

- Introduction to the Semantic Web
- Linking Language to Knowledge in Watson
- Type Coercion
- Temporal and Spatial Reasoning
- Prismatic

Using Structured Evidence

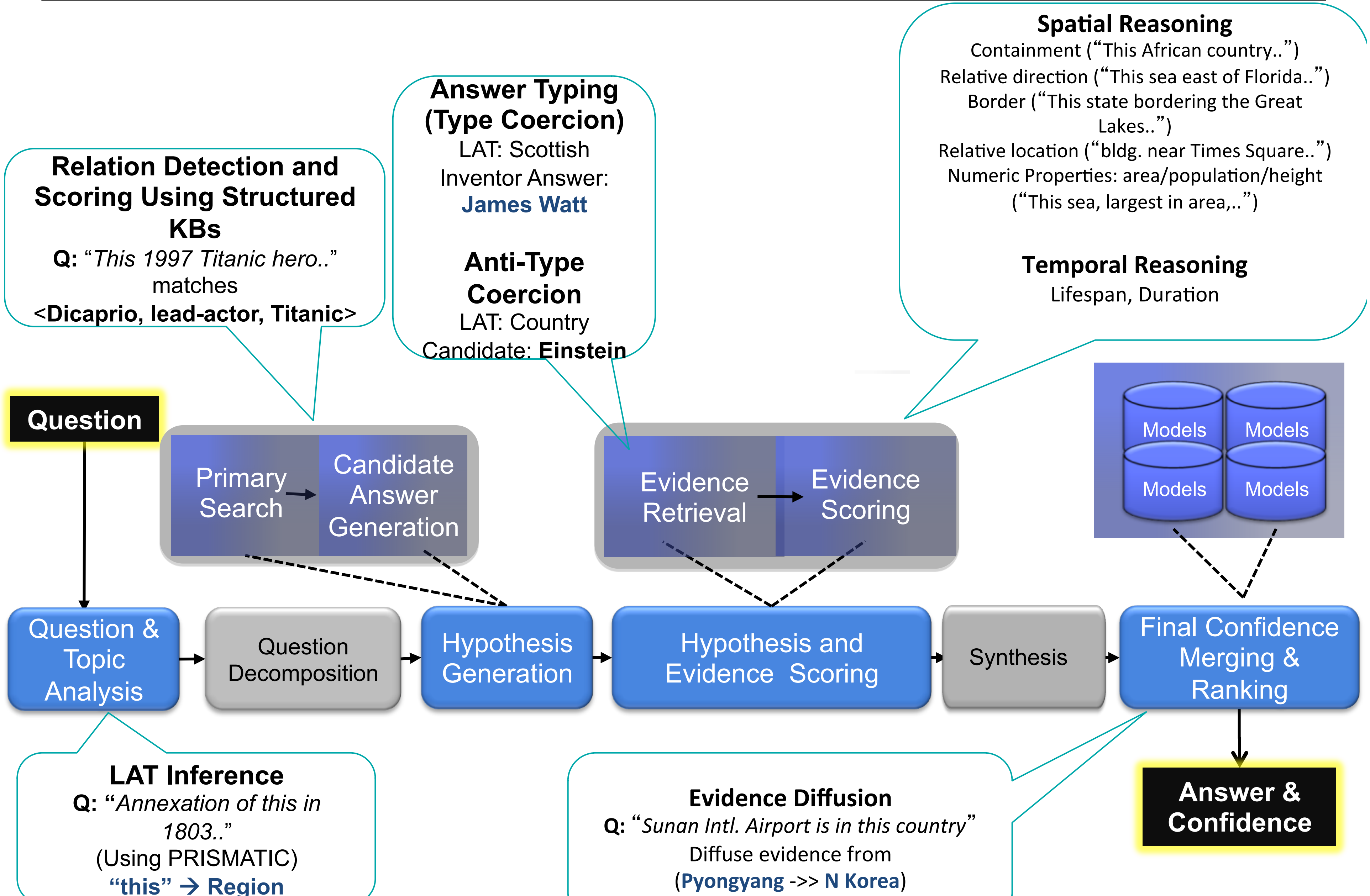
- Exploit wealth of freely available structured information
  - e.g. Linked Open Data (LOD)
  - Types, Relations, Links



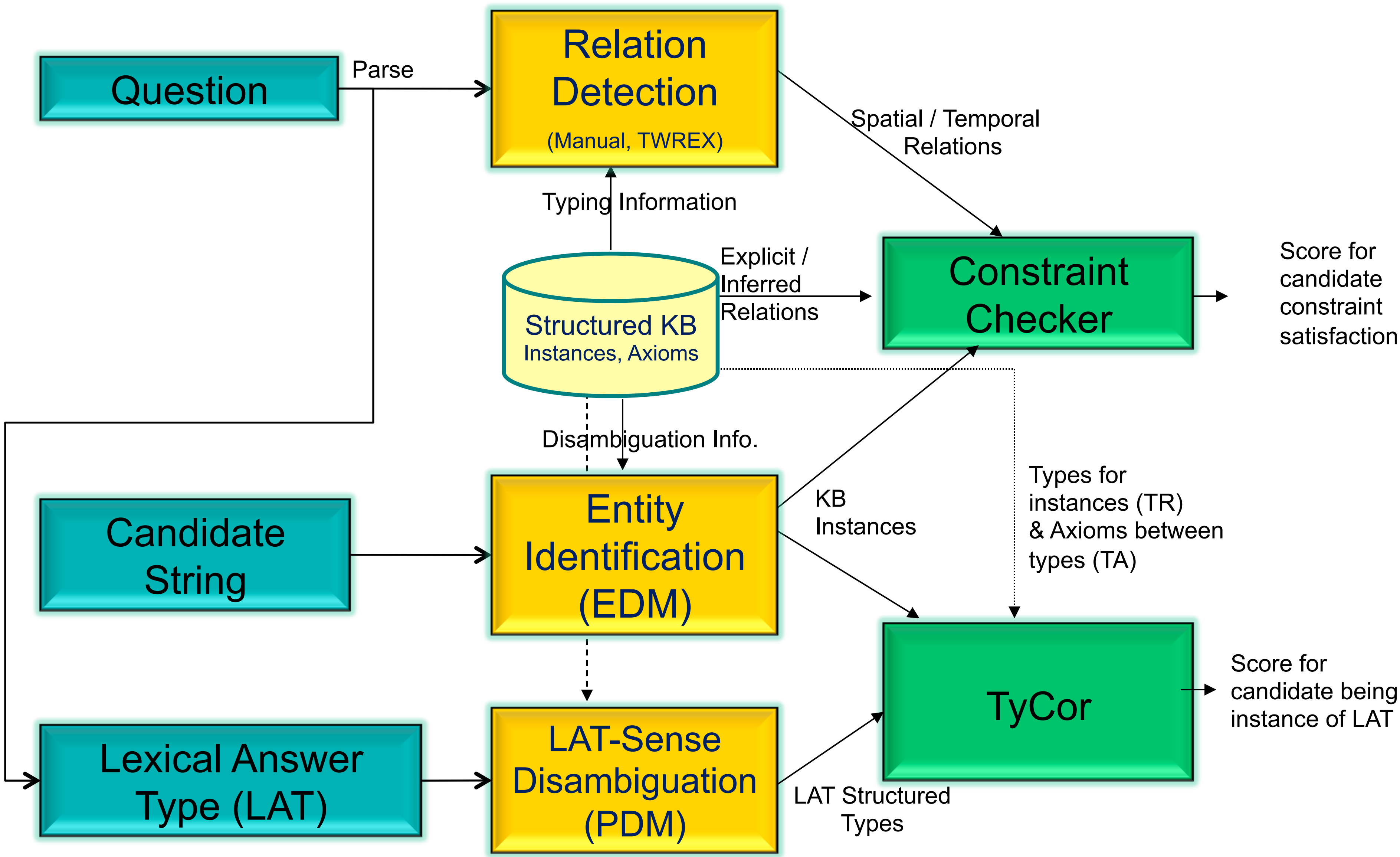
- Complement results from unstructured text analysis

- Useful for explanation data
  – Precise and reliable evidence
  (e.g. spatial / temporal constraint match)

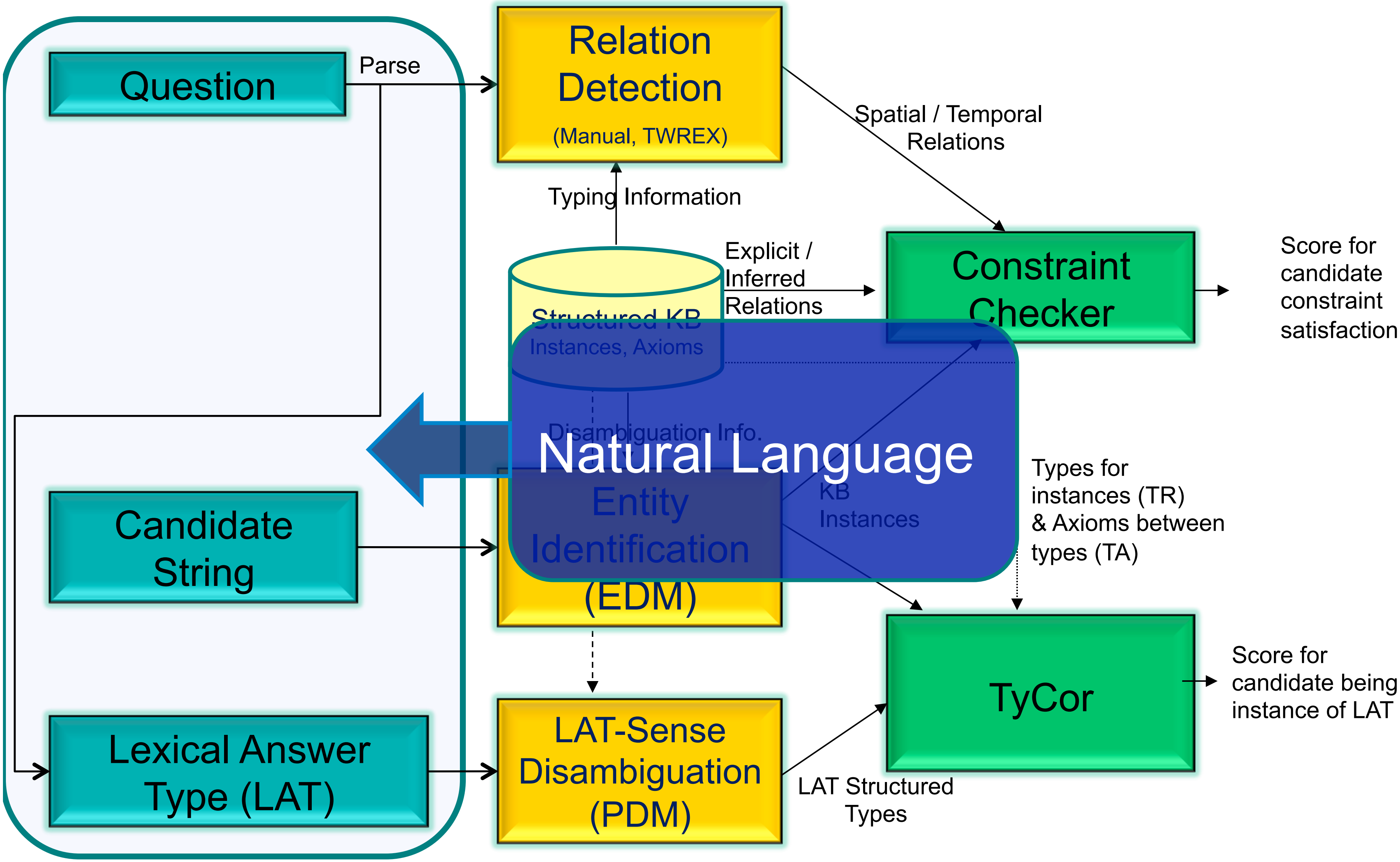
Using Structured Data and Inference in Watson



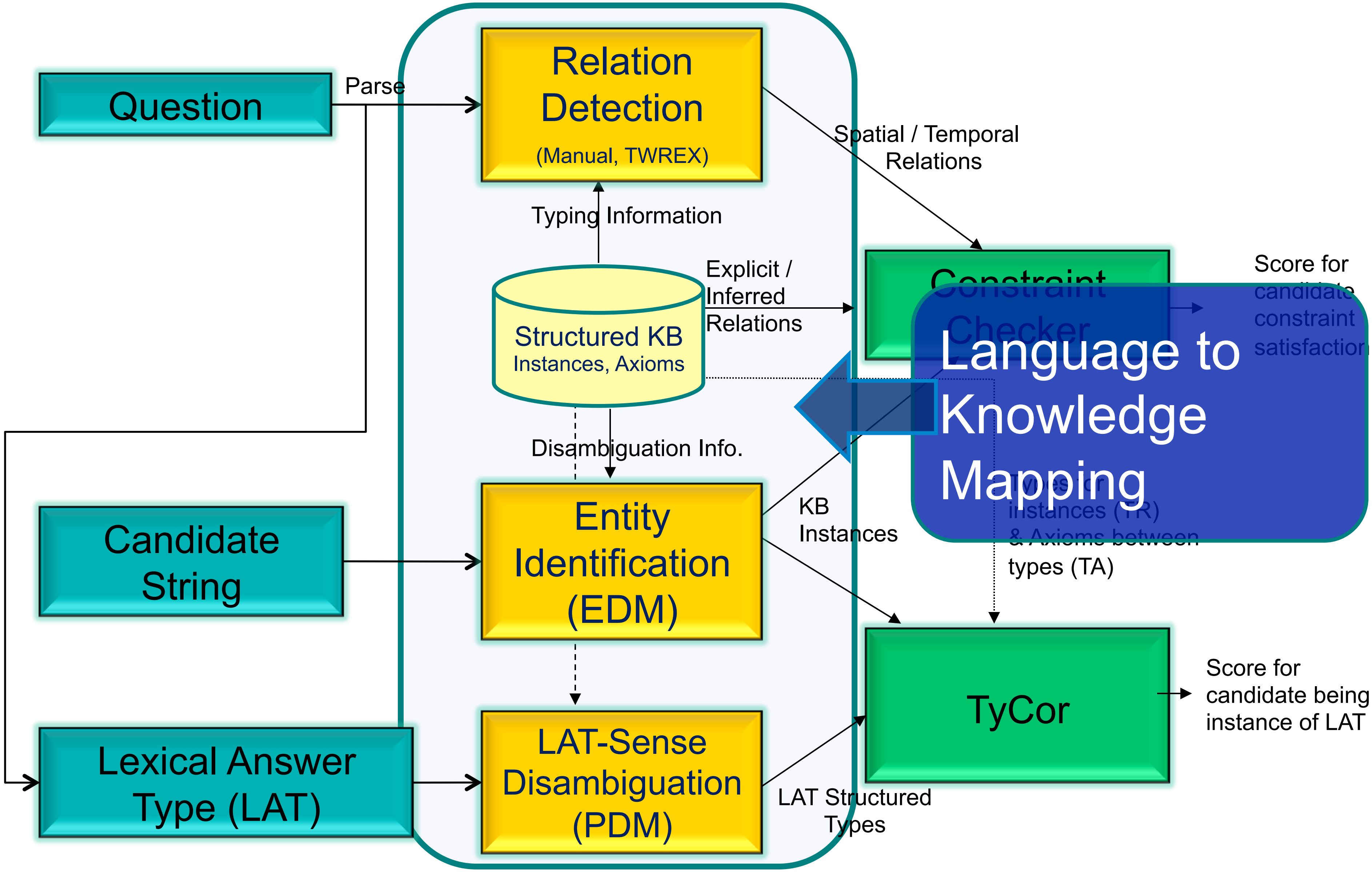
Structured Inference Architecture



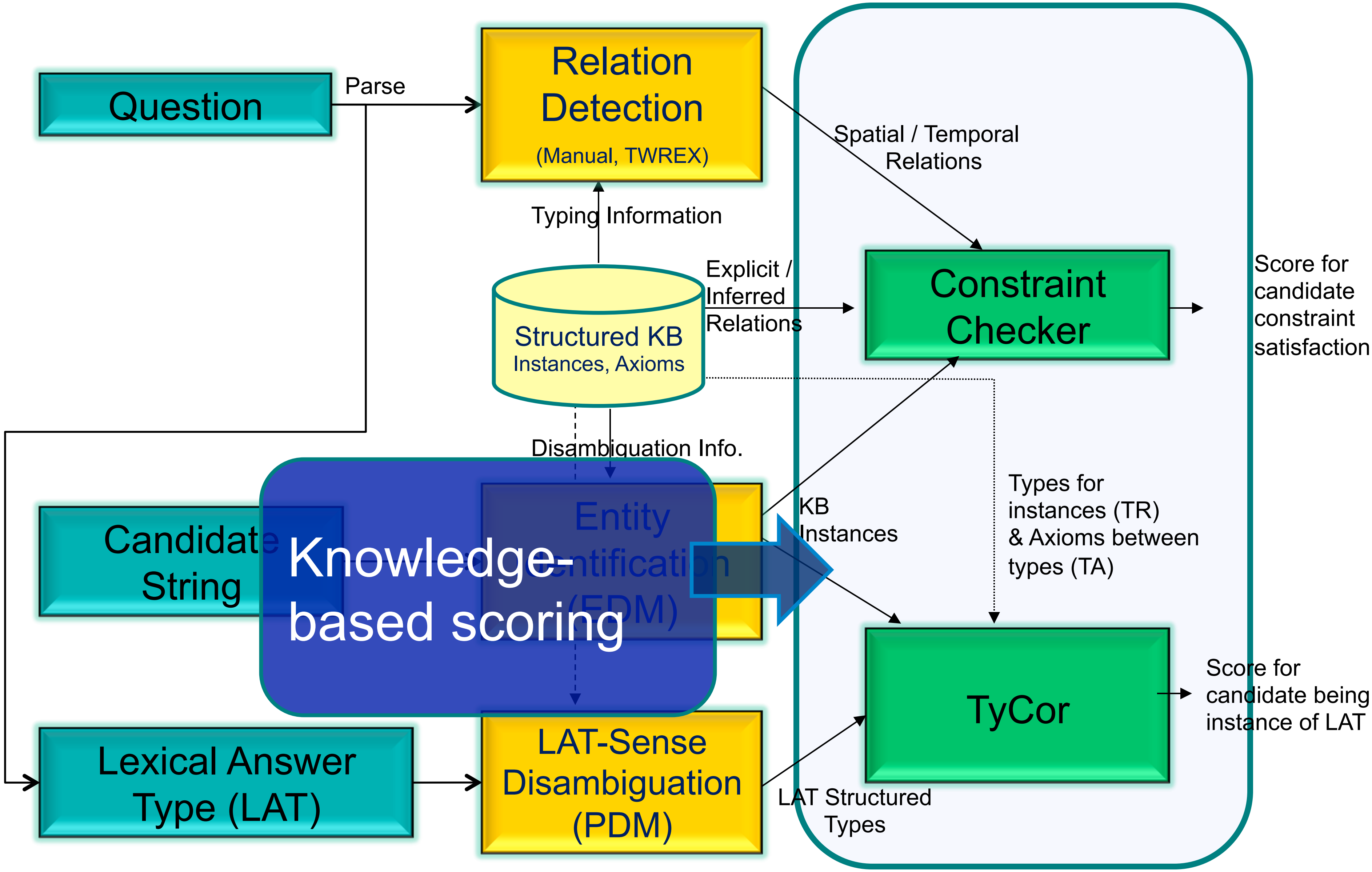
Structured Inference Architecture



Structured Inference Architecture

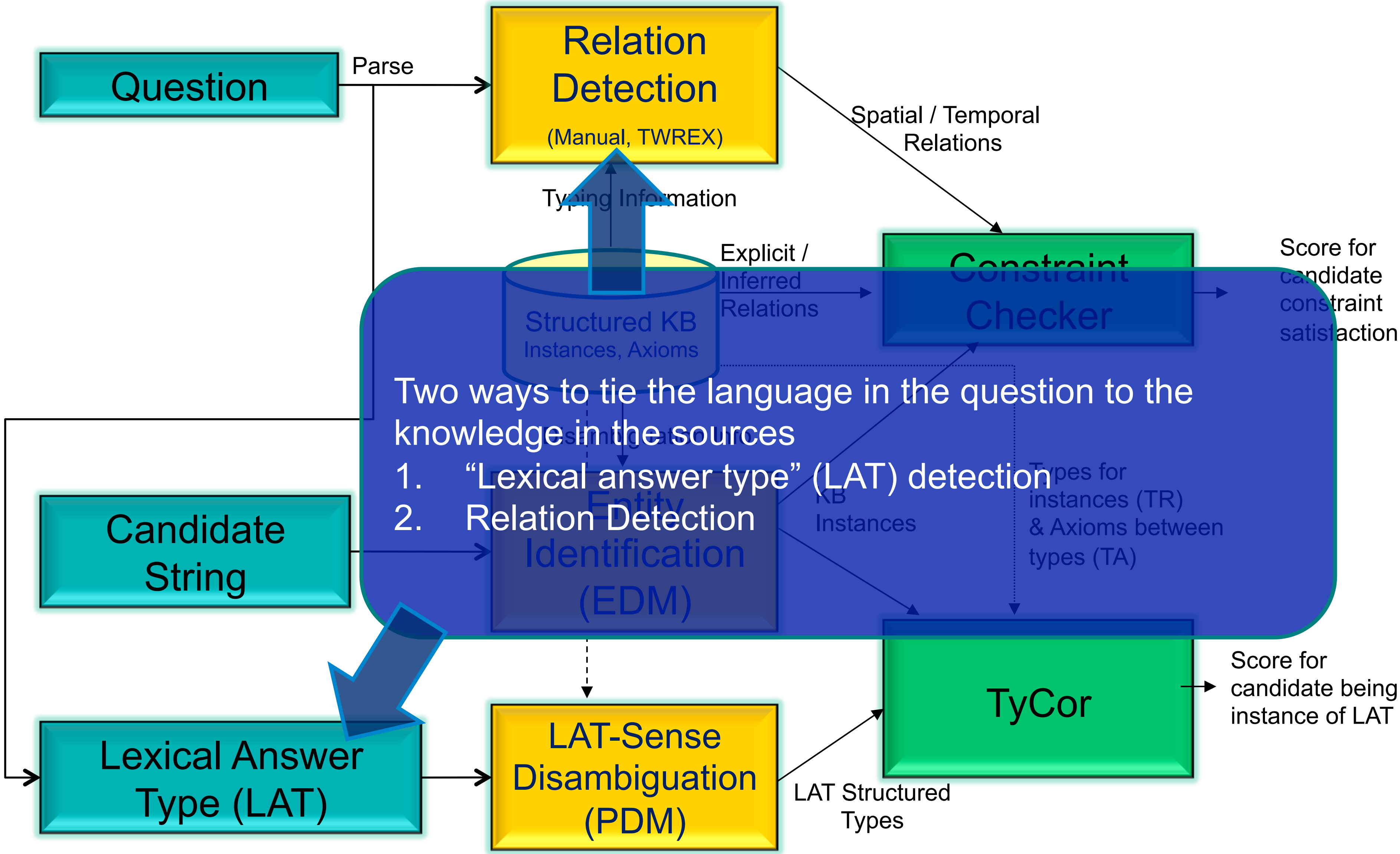


Structured Inference Architecture



IBM WATSON

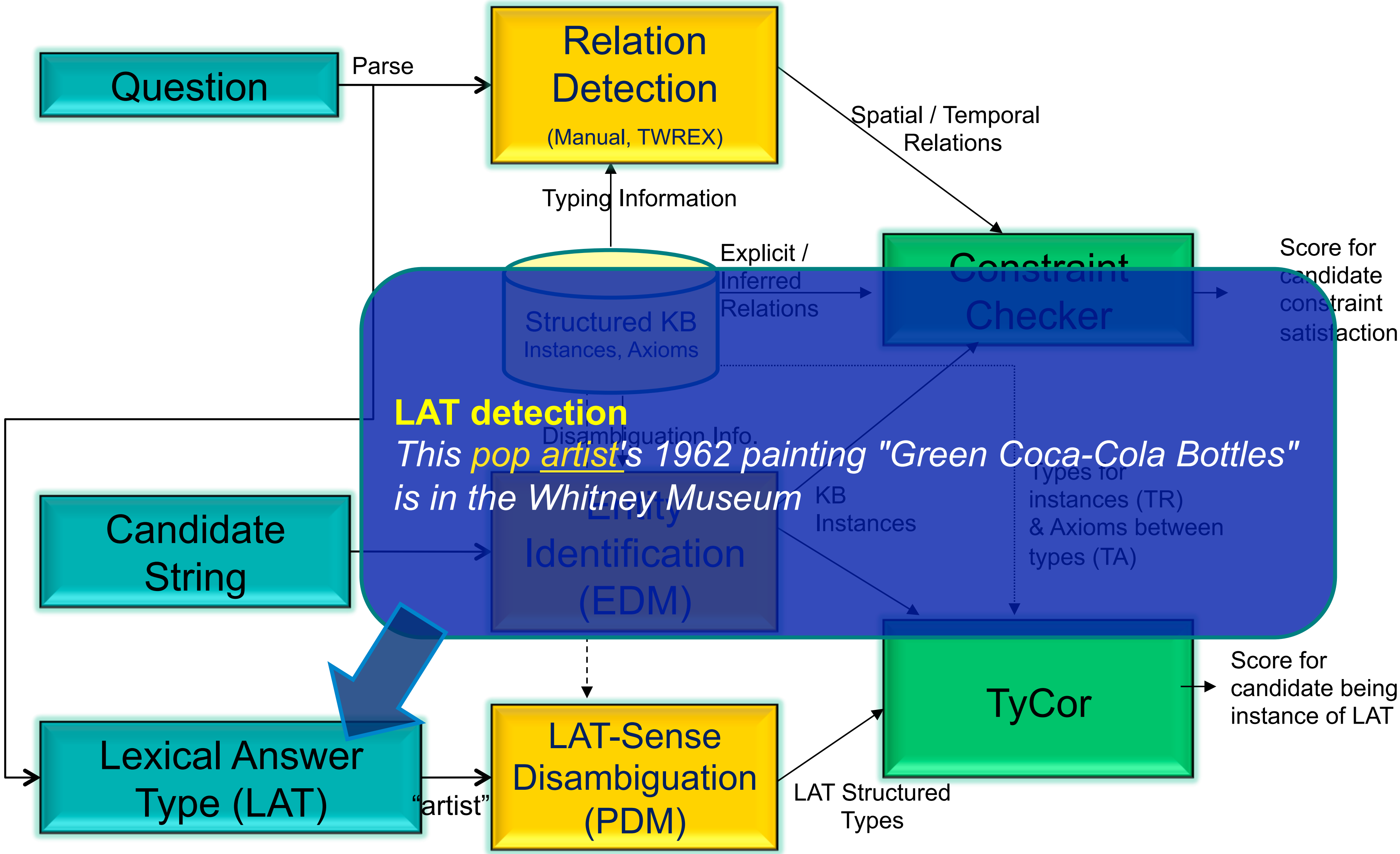
Structured Inference Architecture



Two ways to tie the language in the question to the knowledge in the sources

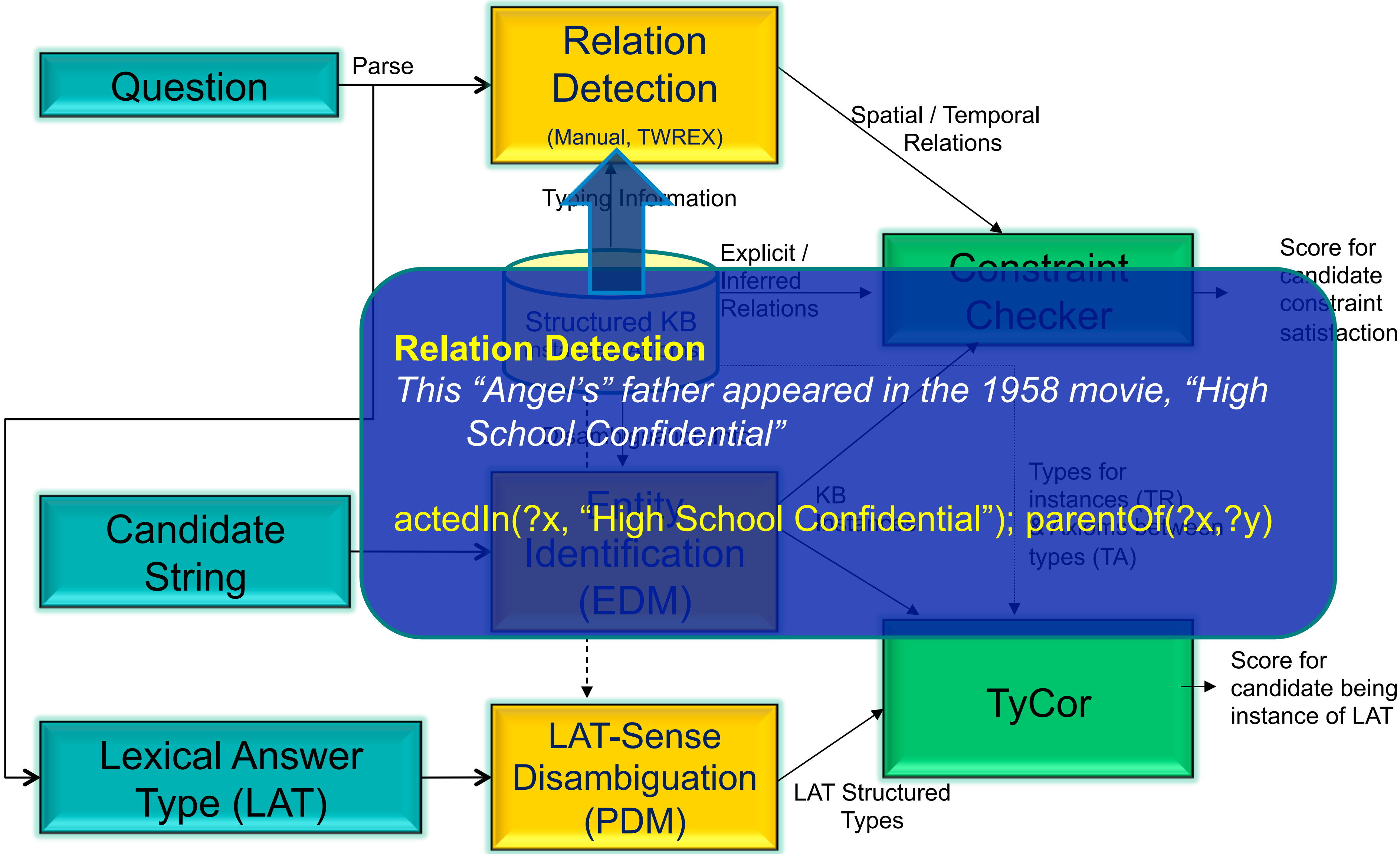
1. “Lexical answer type” (LAT) detection
2. Relation Detection

Structured Inference Architecture



IBM WATSON

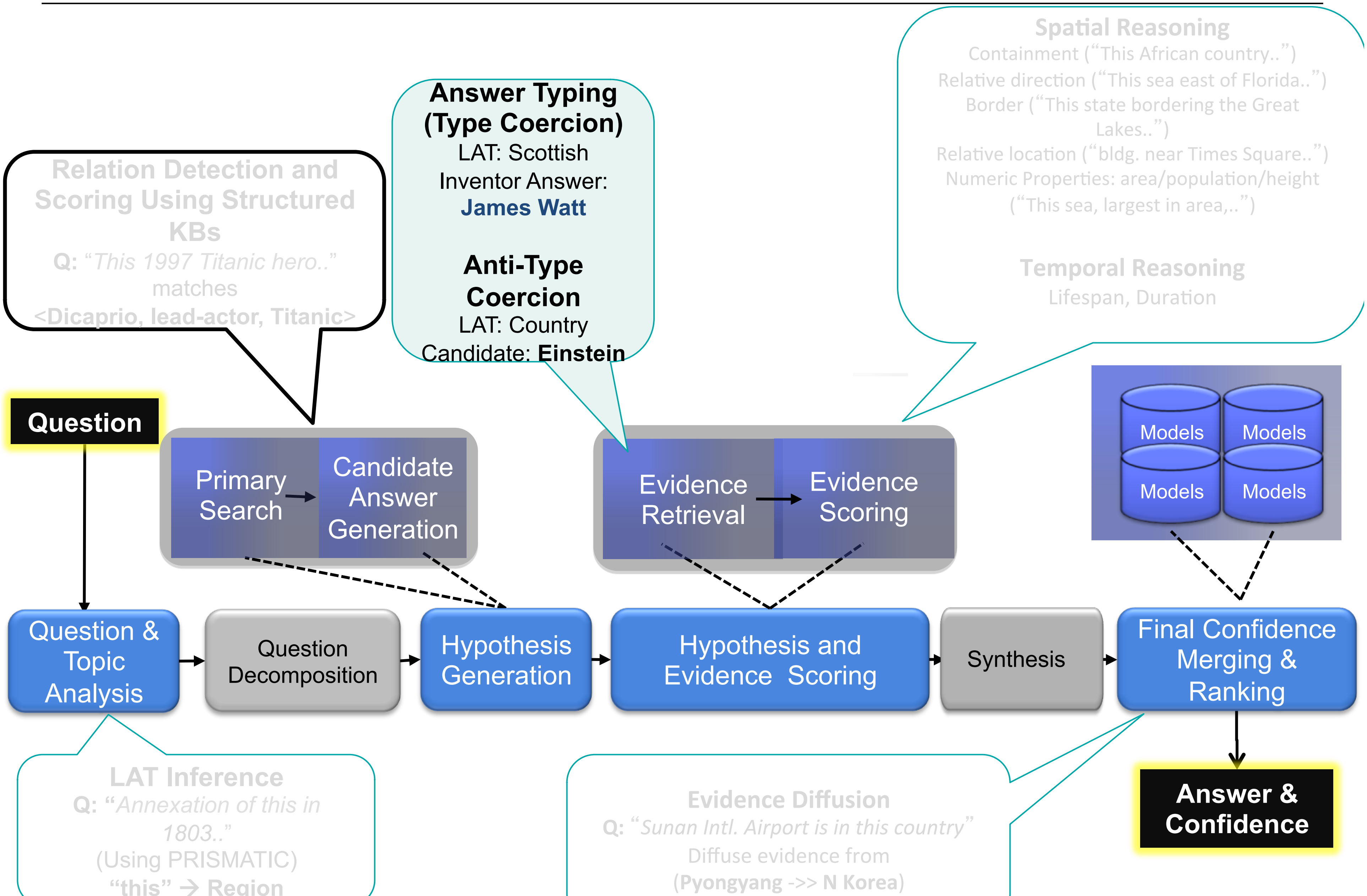
Structured Inference Architecture



Outline

- Introduction to the Semantic Web
- Linking Language to Knowledge in Watson
- Type Coercion
- Temporal and Spatial Reasoning
- Prismatic

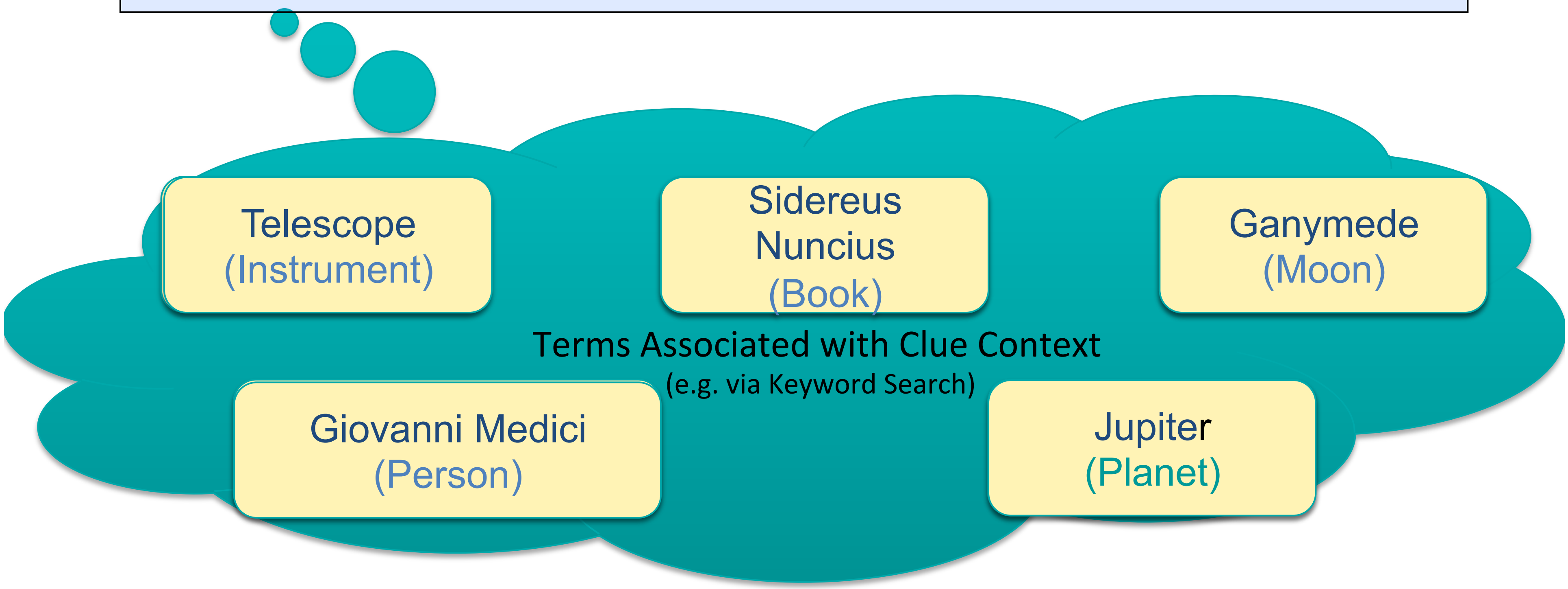
Using Structured Data and Inference in Watson



Role of Answer Typing in QA

- Type Information - a crucial hint to get the correct answer

ASTRONOMY: In 1610 Galileo named the moons of <u>this planet</u> for the Medici brothers



Closed Domain Type Checking

- Used in Traditional QA Systems

Based on "Type And Generate" Principle

- Focus on a pre-determined set of interesting types
  - People, Places, Organizations, Dates
- For these types, run Named Entity Recognizers (NER) over text corpus
  - People: {"Einstein", "Sir I. Newton"..}
  - Places: {"Germany", "UK"..}
  - Dates: {"1885", "3rd April 1715"..}
- At run-time, given a question, detect lexical answer type (LAT) and:
  - Generate candidates from pre-compiled list of LAT instances

Limitations

- Highly brittle – QA system breaks down if type not recognized
- Limited Coverage – need to enumerate all relevant types beforehand
- Dependent on quality of NERs used

Closed Domain Typing for Jeopardy!?

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

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

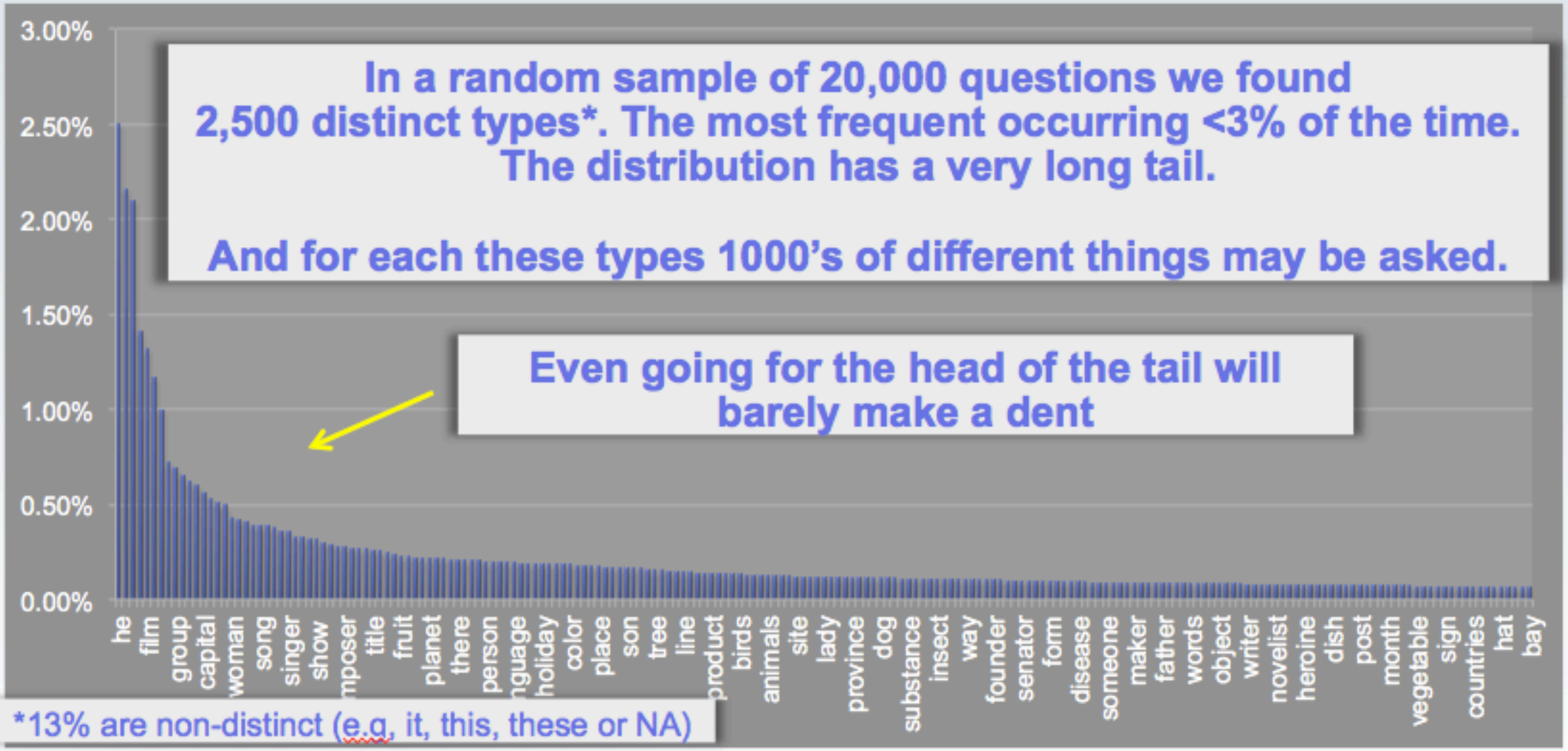
- Secy. Chase just submitted **this** to me for the third time--guess what, pal. This time I'm accepting **it**
- Category: Lincoln Blogs
- Answer: **his resignation**

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

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



*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.

Open Domain Type Coercion (TyCor)

- Approach taken in DeepQA
  - Based on "Generate-and-Type" Principle
    - Generate candidates without considering answer type (LAT)
    - Later check whether candidate can be *coerced* into LAT
    - Use a suite of Type-Coercion Algorithms
    - Use machine-learning to combine information from TyCors

- Advantages
  - More flexible as QA system does not break down if type is not detected or meaningful
  - Much wider type coverage possible using a variety of sources and analytics for TyCor

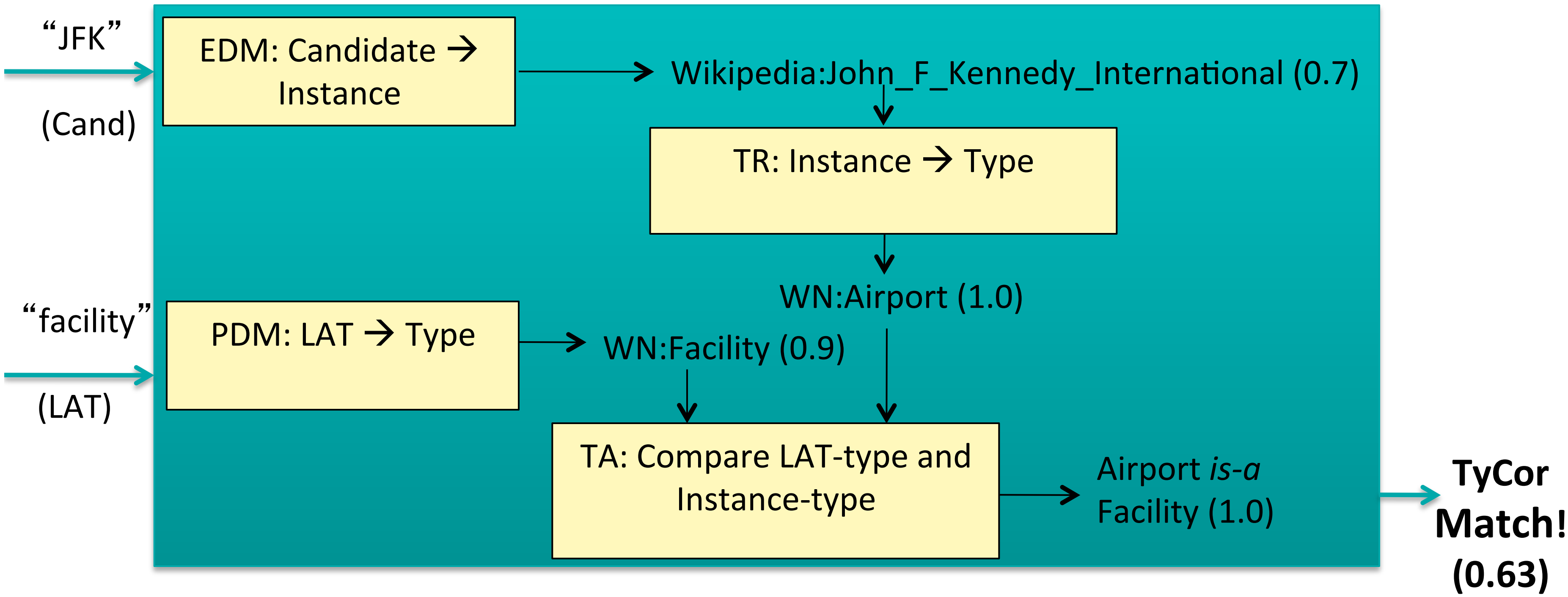
Open Domain Type Coercion (TyCor)

- Approach taken in DeepQA
  - Based on "Generate-and-Type" Principle
    - Generate candidates without considering answer type (LAT)
    - Later check whether candidate can be *coerced* into LAT
    - Use a suite of Type-Coercion Algorithms
    - Use machine-learning to combine information from TyCors

- Advantages
  - More flexible as QA system does not break down if type is not detected or meaningful
  - Much wider type coverage possible using a variety of sources and analytics for TyCor

Type Coercion (TYCOR) Framework

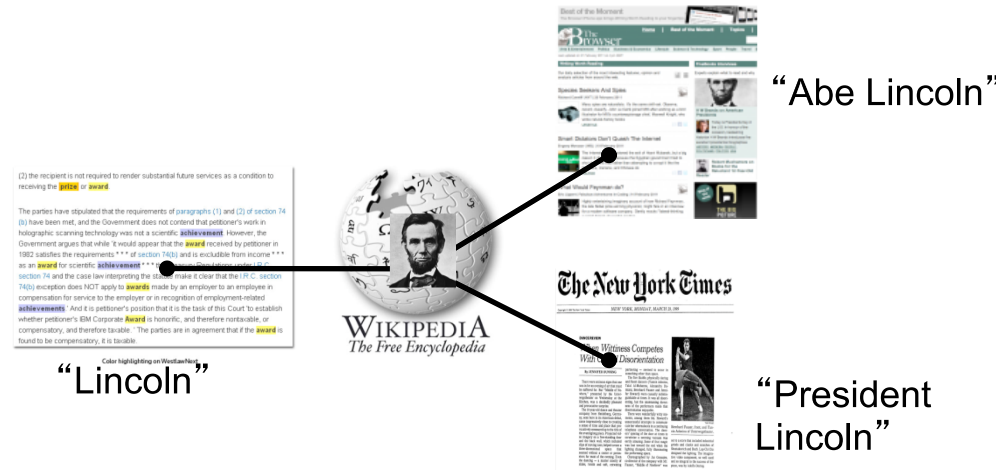
- **Problem**: Compute type match for candidate w.r.t. LAT
  – Both candidate and LAT expressed as Strings
– **4 Steps: EDM** (Entity Disambiguation and Matching)**, PDM** (Predicate Disambiguation and Matching)**, TR** (Type Retrieval)**, TA** (Type Alignment)



Entity Disambiguation and Matching (EDM)

Entity Disambiguation and Matching Problem

- Fundamental Task in NLP: Map entity string to meaningful reference



Issue 1:
Many different ways to refer to the same entity (spellings, aliases, nicknames, abbreviations)


World | All World Topics

Gaddafi? Kadafi? Qaddafi? What's the correct spelling?

You say, Gaddafi, we say Qaddafi. Other variations on the leader of Libya include "Gathafi," "Kadafi," and "Gadafy," creating an unholy mess for newspaper editors.

Issue 2:
Sense Disambiguation depends on context

Flight took off from JFK…	**JFK was assassinated…**	**Film critics loved JFK…**

More on EDM

Resources Used In DeepQA:

For Identification

- Wikipedia redirects (Myanmar ->> Burma)
- Synonyms / aliases extracted from text

  "*IBM's distinctive culture and product branding has given it the nickname Big Blue*"
- Anchor-Hyperlink Data

  "Newton Heath LYR Football Club" 16 links →
  "MUFC" 276 links →
  "Man Utd 1855 links →

  Manchester United FC Wikipedia Page

For Disambiguation

- Wikipedia Disambiguation Pages (wide coverage)
  - -~150K disambiguation pages in 2008
  - - E.g. "Java" has >50 different senses spanning >20 Distinct Types
- Do not need WordNet Synsets for Instances (poor coverage)

EDM in DeepQA

- Basic EDM Algorithm
  - **Input**: Entity String
  - **Output**: Set of Wikipedia URIs ranked along following criteria
    1. Exact Match to WP Page Title
    2. Redirect to WP Page
    3. {Disambiguations, Link Anchors, Mined Synonym Lists}

    Entities in (3) sorted by popularity

E.g. "Emerson" → <RW Emerson, 0.7>

<Emerson College, 0.2>

<Emerson Radio, 0.1>

Relevance

- Limitations – currently working on this

  Use context to disambiguate entities

  - Look at neighboring entities in the text ("IBM and Oracle..")

  -Look at relations involving entity ("Emerson wrote poem" -> Emerson: Person)

-Note: component level improvements reach a cost/effect saturation

Predicate Disambiguation and Matching (PDM)

(basically WSD)

– LAT: star

In the northern hemisphere, latitude is equal to the angle above the horizon of this star, Alpha Ursae Minoris	***This star of "The Practice" played Clint Eastwood's Secret Service partner in the film "In the Line of Fire"***

- Similar in principle to EDM
  - EDM – map named entity → instance
  - PDM – map generic noun → class/type

IBM WATSON

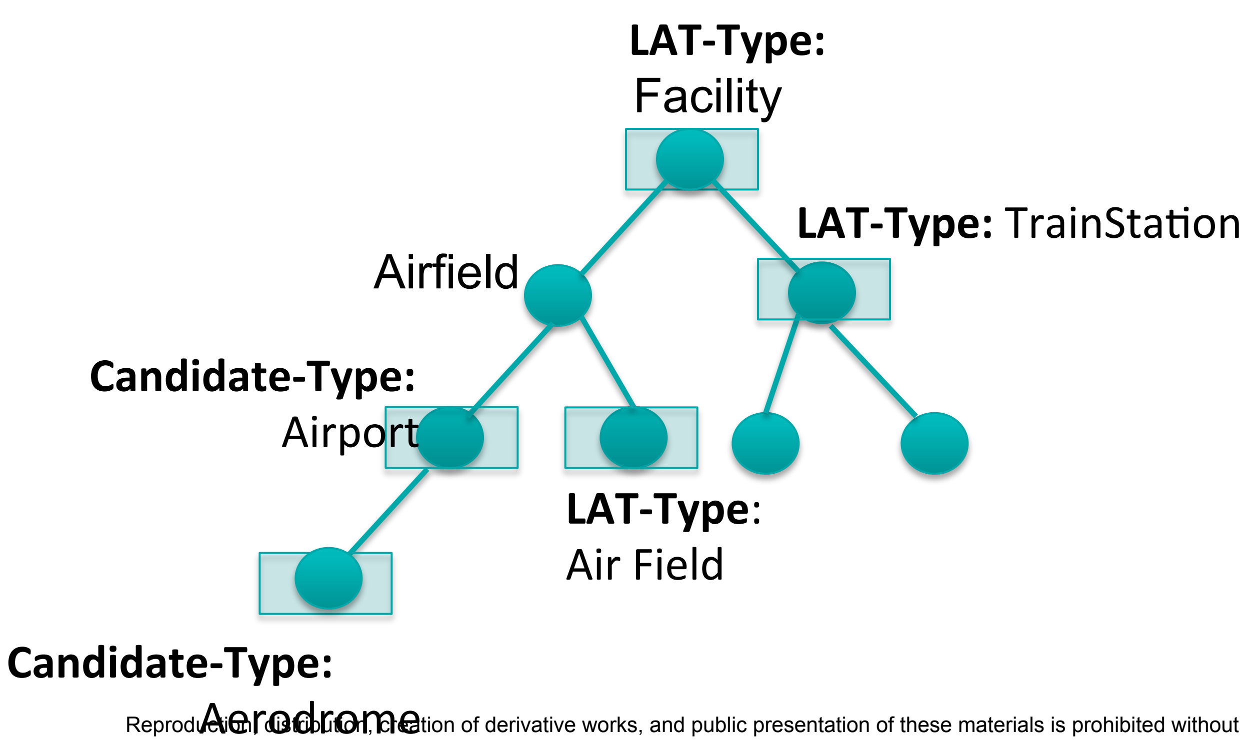
Type Recognition (TR)

- Obtain Types for Instances
- Taxonomies Used In DeepQA:
  - WordNet
  - Yago Ontology (from DBpedia)
  - Wikipedia Categories
  - Auto-Mined Types from Text (Prismatic)

RECALL PRECISION

Type Alignment (TA)

- Type Matching/Alignment Problem
  - Compare candidate types with LAT types
  - Produce a score depending degree of Match
- Various Types of Match Considered



Subclass Match (1.0)

Sibling Match (0.5)

Deep LCA Match (0.25)

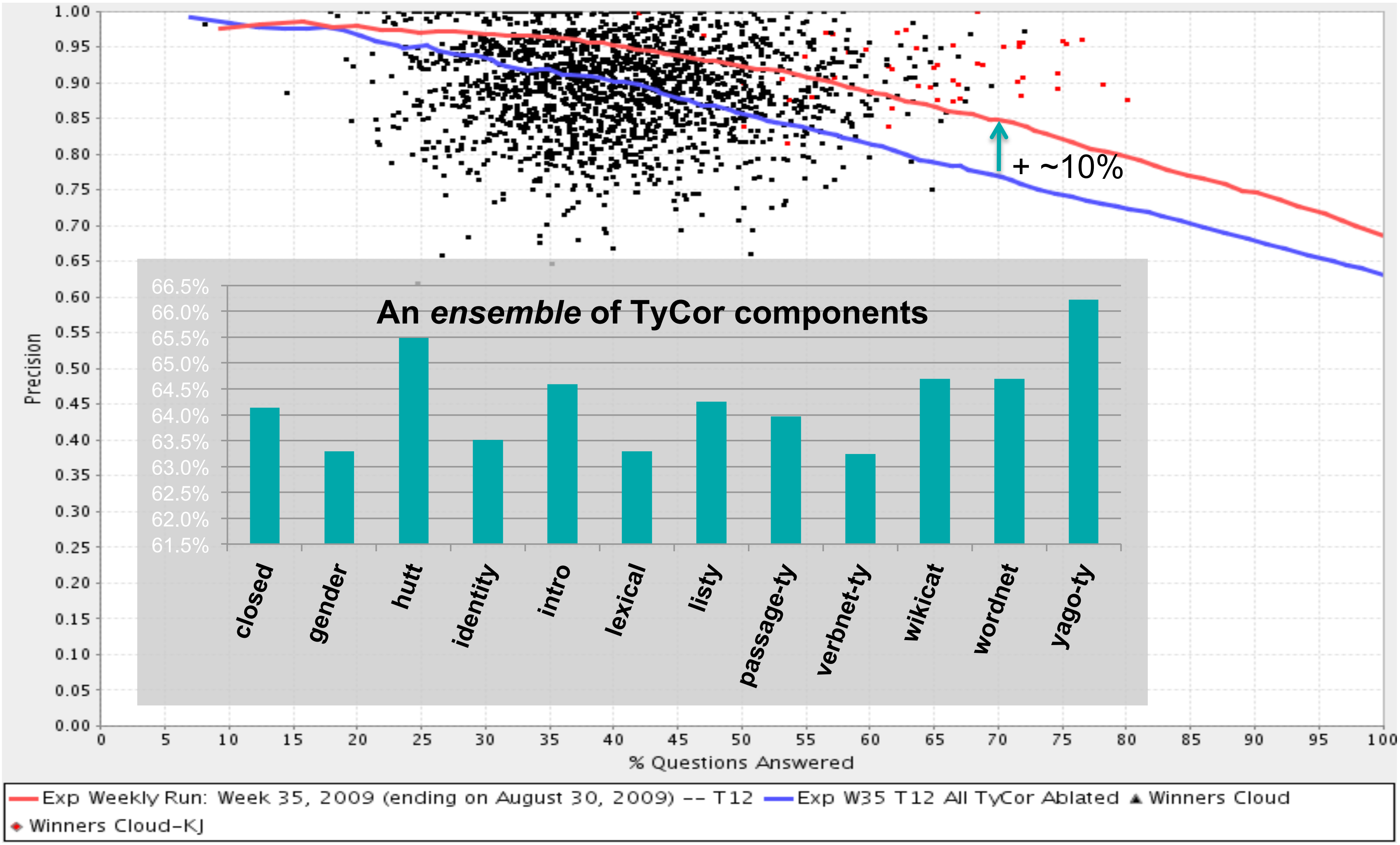
Disjoint Types (-1.0)

Putting it all together

- TyCor Score = EDM * TR * PDM * TA
  - Intermediate Failure
  - If any step fails, Tycor Score = 0
  - An-TyCor
  - When TA score is -1 (Disjoint Types) → AnTyCor Feature added to model
  - Strong negative signal against candidate
  - Helps rules out candidates of wrong type (e.g. LAT: Country, Candidate: Einstein)

- TyCor Algorithm Suite in DeepQA
  - -14 TyCors Developed
  - -Each TyCor score is a separate feature in model
  - -Model learns weights on diff. TyCors: balances/combines type information

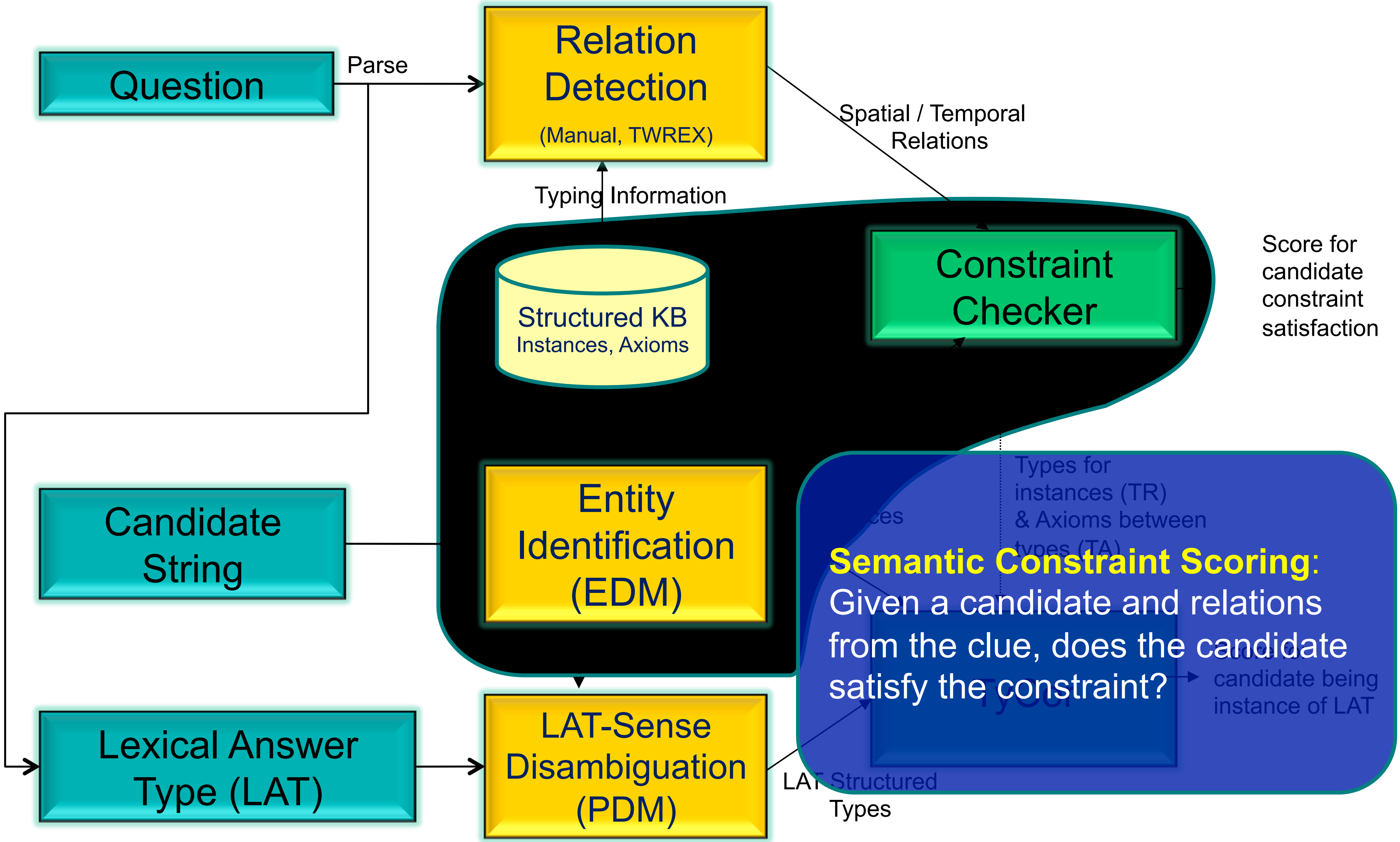
TyCor Impact in DeepQA



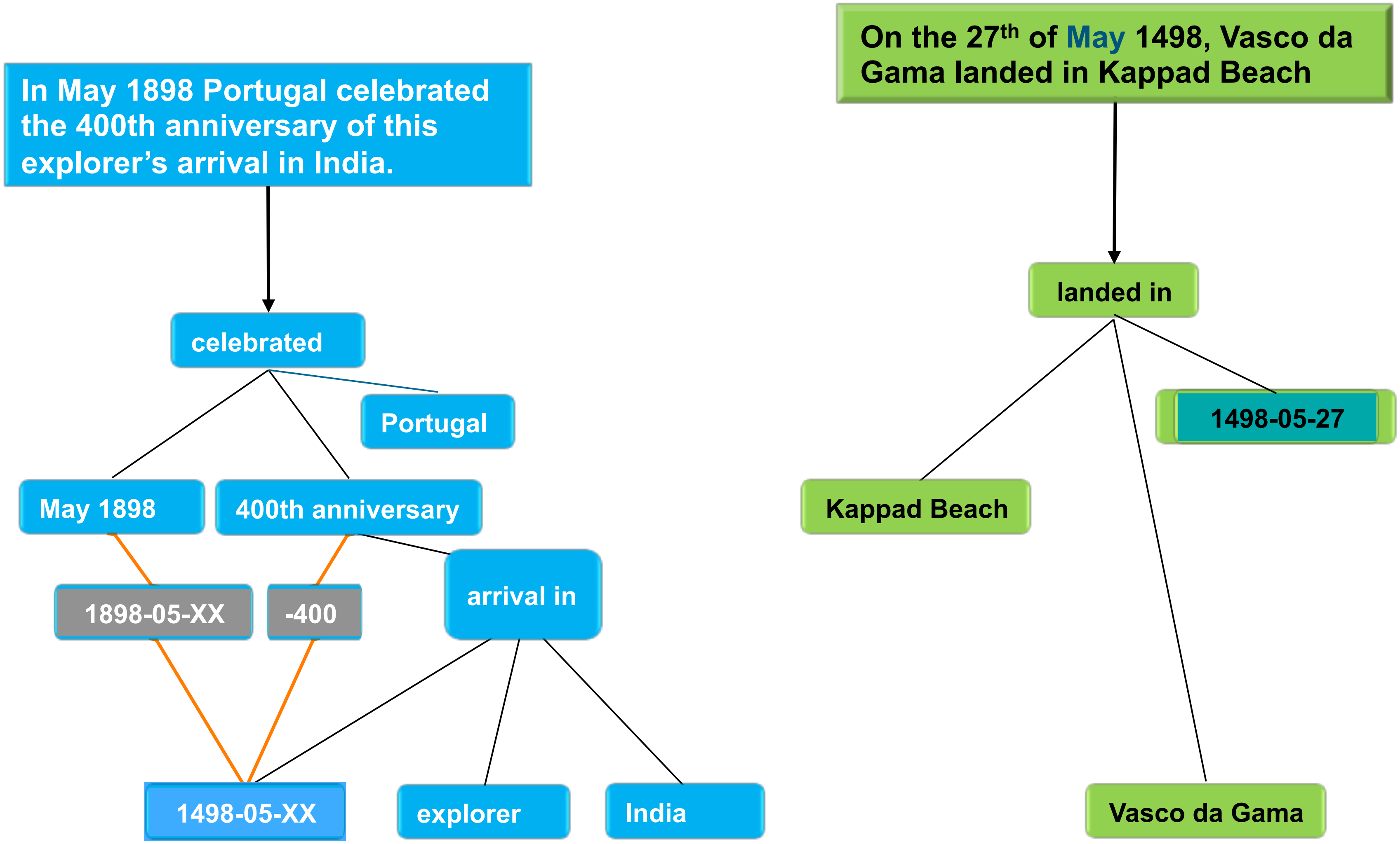
Outline

- Introduction to the Semantic Web
- Linking Language to Knowledge in Watson
- Type Coercion
- Temporal and Spatial Reasoning
- Prismatic

Structured Inference Architecture

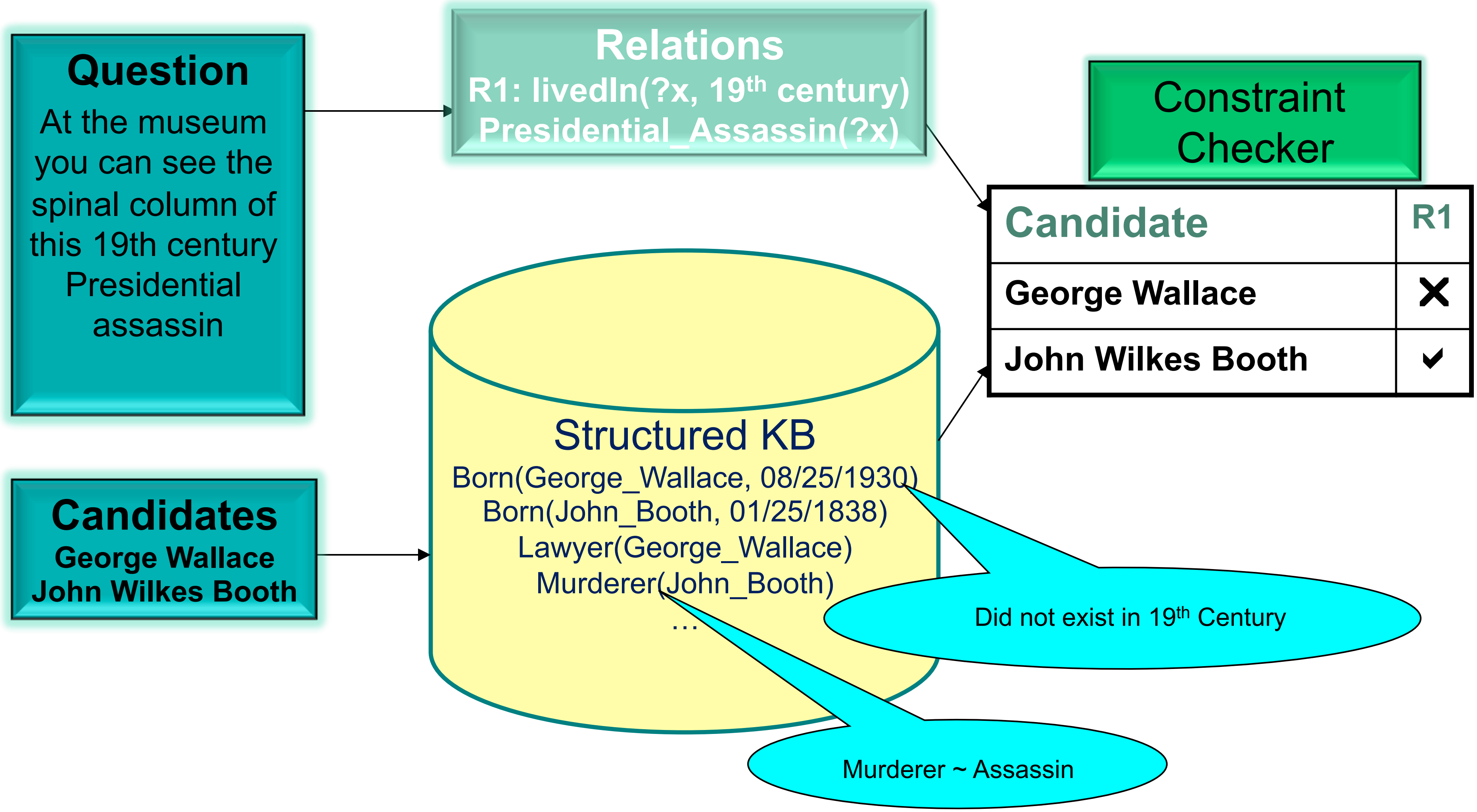


Temporal Normalization & Arithmetic

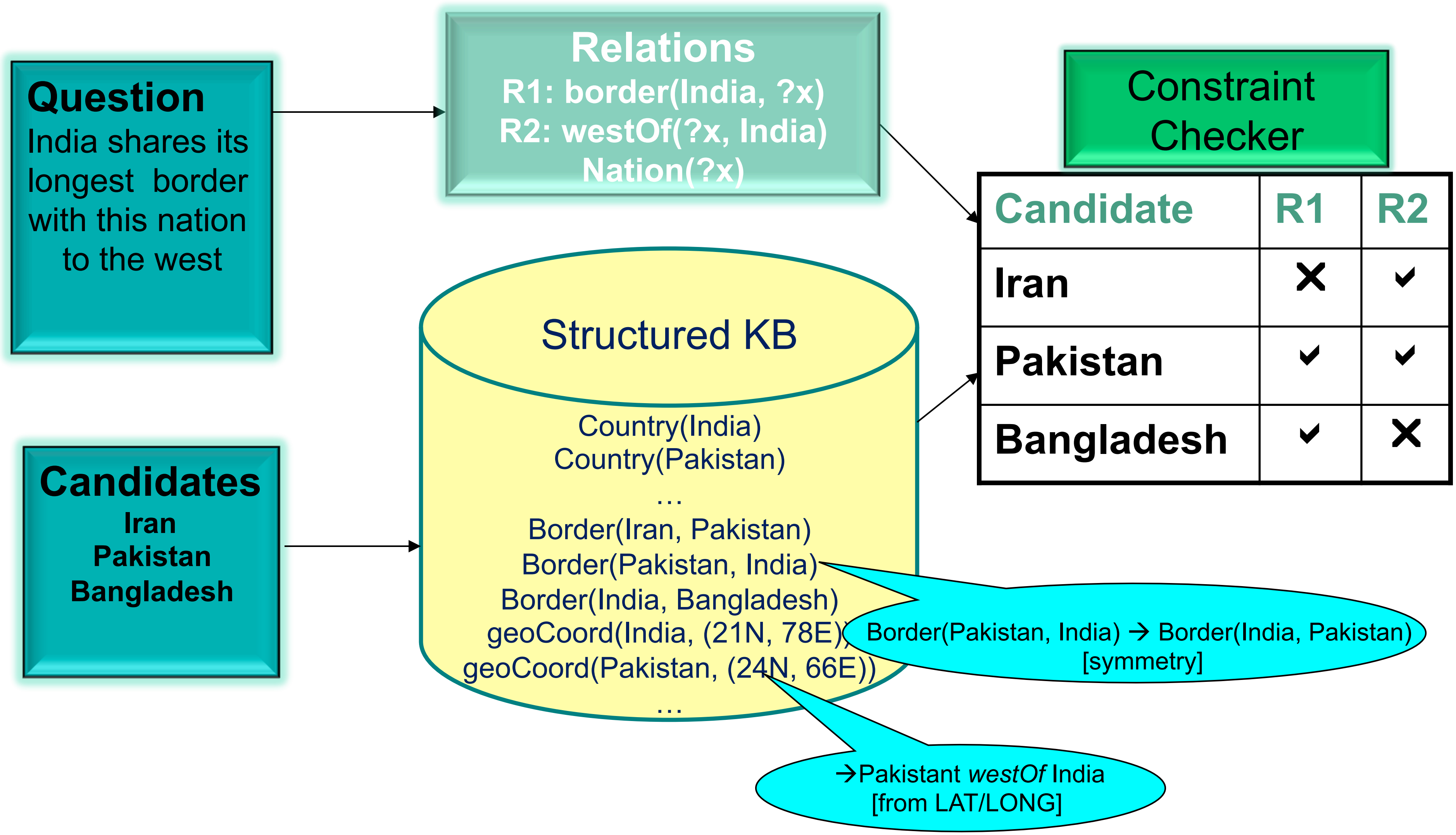


TimeStamps: TIMEX3 (normalized)
Rule Based anniversary detection system

Temporal Reasoning



Spatial Reasoning

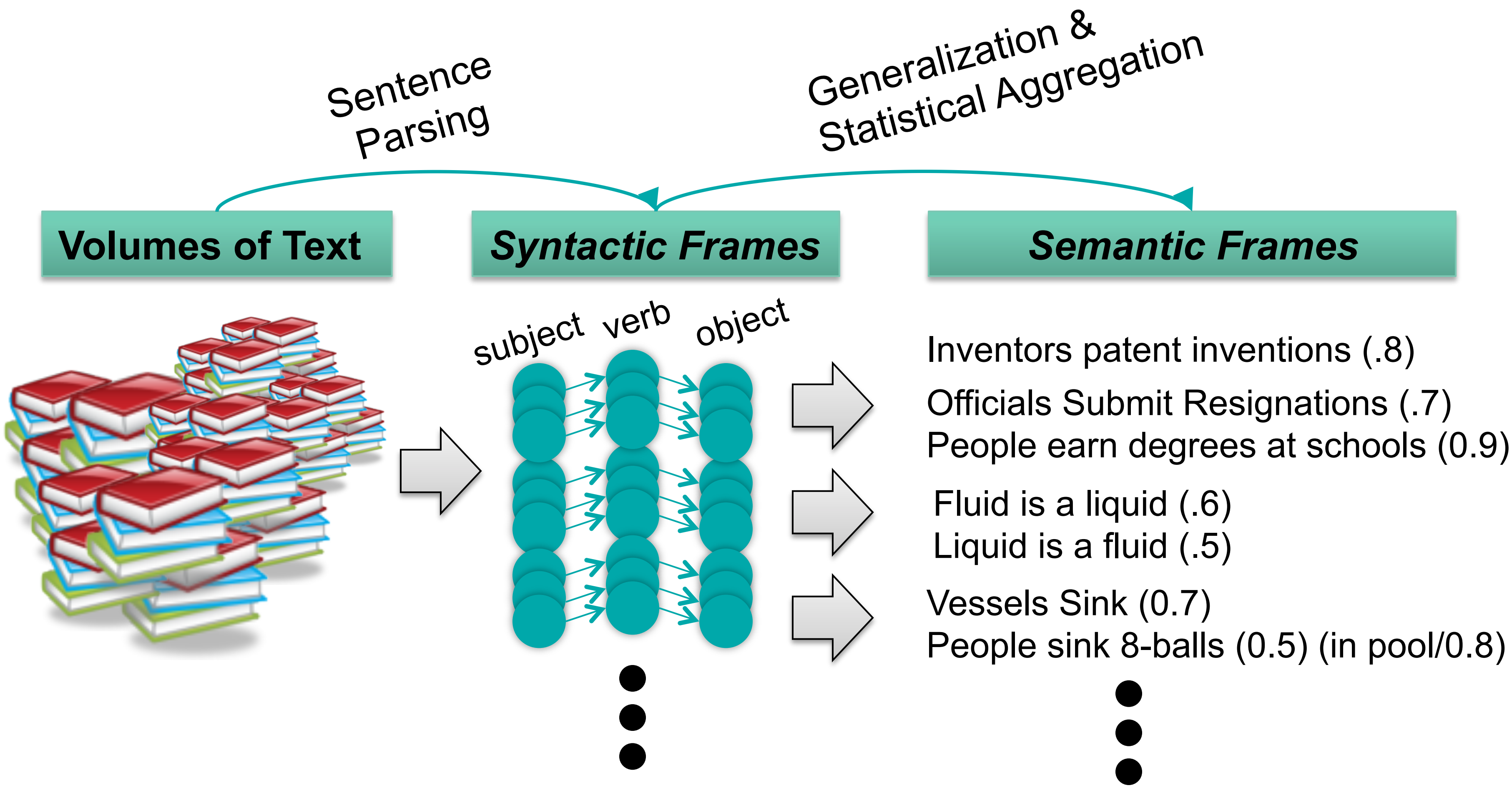


Candidate	R1	R2
Iran	✗	✓
Pakistan	✓	✓
Bangladesh	✓	✗

Outline

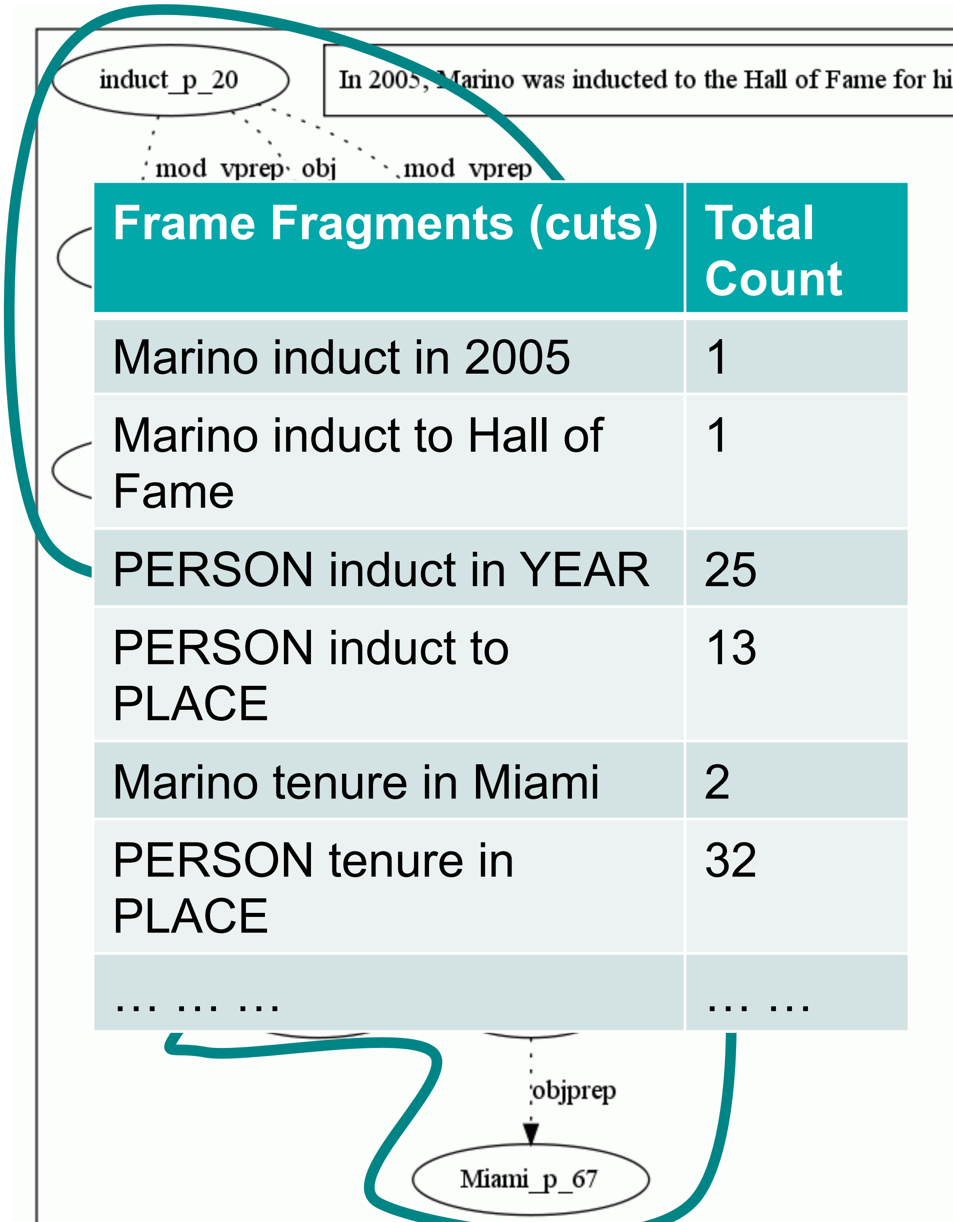
- Introduction to the Semantic Web
- Linking Language to Knowledge in Watson
- Type Coercion
- Temporal and Spatial Reasoning
- Prismatic

Learning by "Reading"



Motivations

- A large amount of unstructured sources
- A great variety of ways to express similar meanings:
- Can we explore the regularities and redundancy in large corpus?
  - ***Thomas Edison*** *was an American inventor, scientist, and businessman who developed many devices that greatly influenced life around the world, including the phonograph, the motion picture camera, and a long-lasting, practical electric* ***light bulb****.*
  - ***Thomas Edison*** *invented the* ***light bulb*** *and GE, a company that continues to bring science and commercial success together.*
  - ***Thomas Edison*** *invented the clear* ***light bulb*** *in 1879.*
  - *… …*



Frame Fragments (cuts)	Total Count
Marino induct in 2005	1
Marino induct to Hall of Fame	1
PERSON induct in YEAR	25
PERSON induct to PLACE	13
Marino tenure in Miami	2
PERSON tenure in PLACE	32
… … …	… …

Frame 1

- verb: *induct*
- obj: *Marino*
  - type: PERSON/PLAYER
- mod_vprep: *in*
- objprep: *2005*
  - type: YEAR
- mod_vprep: *to*
- objprep: *Hall of Fame*
  - type: PLACE

Frame 2

- noun: *tenure*
- mod_ndet: *Marino/his*
  - type: PERSON/PLAYER
- mod_nprep: *in*
- objprep: *Miami*
  - type: PLACE

Template Documentation

PRISMATIC Applications

- Named Entity Type Disambiguation:
  - *The Thunder outscored opponents by 8 points per game.*
  - Is *Thunder* a TEAM or a SOUND?
  - Hypothesis 1: Thunder is a TEAM
    - Frame: “TEAM outscore opponents” → 1182 occurrences
  - Hypothesis 2: Thunder is a SOUND
    - Frame “SOUND outscore opponents” → 0 occurrences

PRISMATIC Applications

- Coreference resolution:
  - *Dallas' last minute field goal at Lambeau field sent* ***it*** *to the NFC Championship game*
  - Is "*it*" Dallas or Lambeau field or field goal?
  - Hypothesis 1: "*it*" = Lambeau field
    - Frame: "send Lambeau field to game" → 0 occurrence in PRISMATIC
    - Frame: "send LOCATION to game" → 13 occurrence in PRISMATIC
  - Hypothesis 2: "it" = Dallas
    - Frame: "send Dallas to game" → 4 occurrences in PRISMATIC
    - Frame: "send TEAM to game" → 4711 occurrence in PRISMATIC
  - Hypothsis 3: "it" = field goal
    - Frame: "send field goal to game" → 0 occurrence in PRISMATIC
    - Frame: "send EVENT to game" → 0 occurrence in PRISMATIC

PRISMATIC Applications

- Relation discovery: what are the common relations between PERSON and TEAM , between TEAM and TEAM?

Frequency	Subj Type	Verb	Obj Type
1446	TEAM	play	TEAM
1204	TEAM	beat	TEAM
1115	TEAM	defeat	TEAM
556	TEAM	face	TEAM
196	TEAM	host	TEAM
192	TEAM	sweep	TEAM

Frequency	Subj Type	Verb	Obj Type
5168	PERSON	lead	TEAM
2586	PERSON	coach	TEAM
2367	PERSON	join	TEAM
2260	TEAM	sign	PERSON
2212	TEAM	draft	PERSON
2096	TEAM	trade	PERSON

Prismatic In Watson

- **Prismatic TyCor**
  - Determine Candidate Answer isa Lexical Answer Type: e.g. *vein* isa *vessel*
- **Prismatic Candidate Generation**
  - *Famous around the world, Morocco leather traditionally comes from this **farm animal***

Measure	Value
Corpus size	30 GB
Number of frames extracted	995 millions
Number of frames per sentence	1.4
Percent of named entities that are extracted as part of a frame	94.4%
Watson Improvement with Prismatic TyCor	+2.4% in accuracy

Outline

- Introduction to the Semantic Web
- Linking Language to Knowledge in Watson
- Type Coercion
- Temporal and Spatial Reasoning
- Prismatic

References

- Murdock, J. W.; Kalyanpur, A.; Welty, C.; Fan, J.; Ferrucci, D. A.; Gondek, D. C.; Zhang, L.; Kanayama, H., **Typing candidate answers using type coercion,** IBM Journal of Research and Development, Volume: 56 , Issue: 3.4
- Kalyanpur, A.; Boguraev, B. K.; Patwardhan, S.; Murdock, J. W.; Lally, A.; Welty, C.; Prager, J. M.; Coppola, B.; Fokoue-Nkoutche, A.; Zhang, L.; Pan, Y.; Qiu, Z. M., **Structured data and inference in DeepQA,** IBM Journal of Research and Development, Volume: 56 , Issue: 3.4
- Fan, J.; Kalyanpur, A.; Gondek, D. C.; Ferrucci, D. A. **Automatic knowledge extraction from documents,** IBM Journal of Research and Development, Volume: 56 , Issue: 3.4
- http://www.w3.org/RDF/
- http://www.w3.org/2004/OWL/

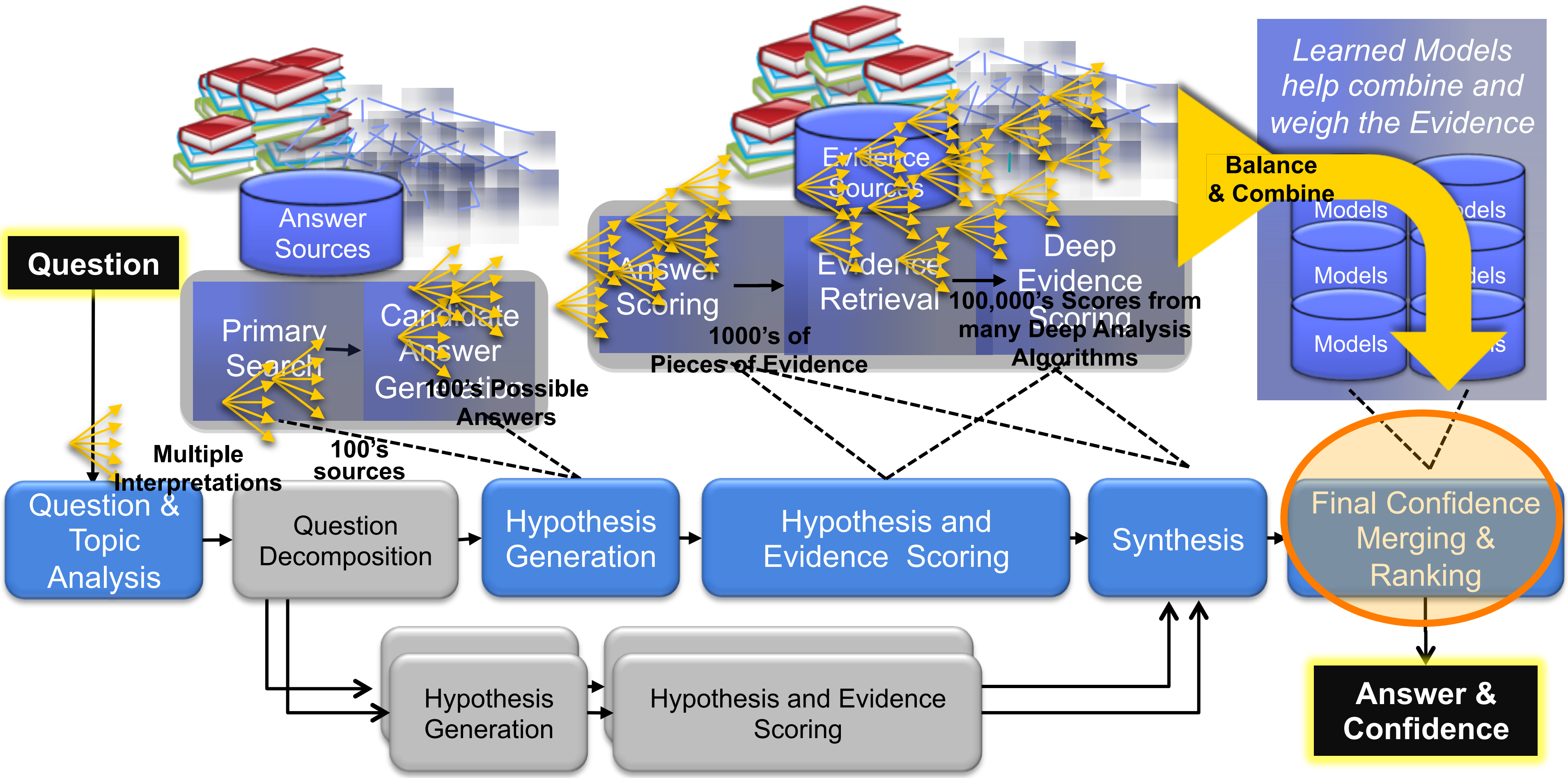
Semantic Technologies in IBM Watson™

Lesson 6 – Machine Learning in Watson

Guest Lecturer: David Gondek

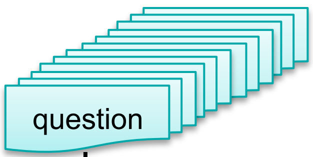
Where is ranking and confidence estimation in the pipeline?

- End of the pipeline
- Potential chokepoint/bottleneck (merges parallel computation streams)

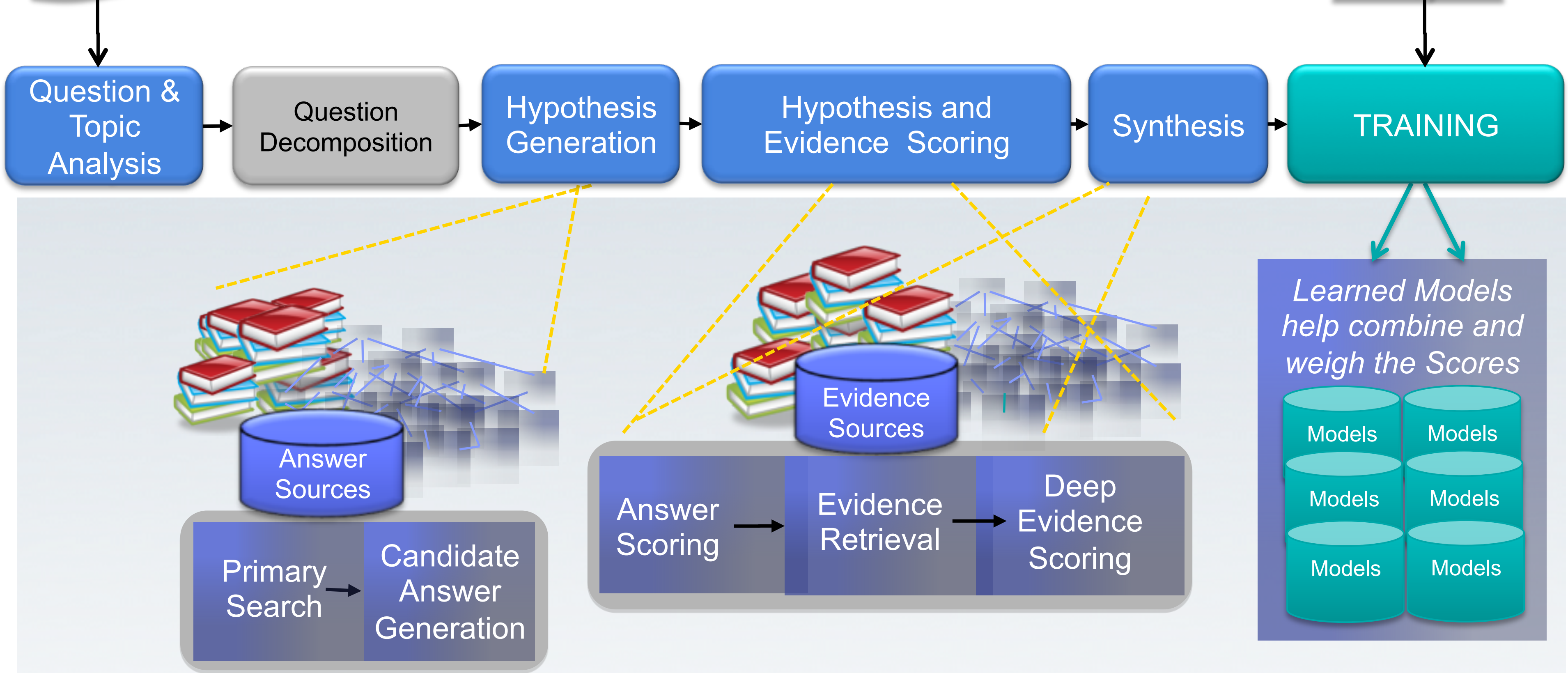


Train Time

Train-time



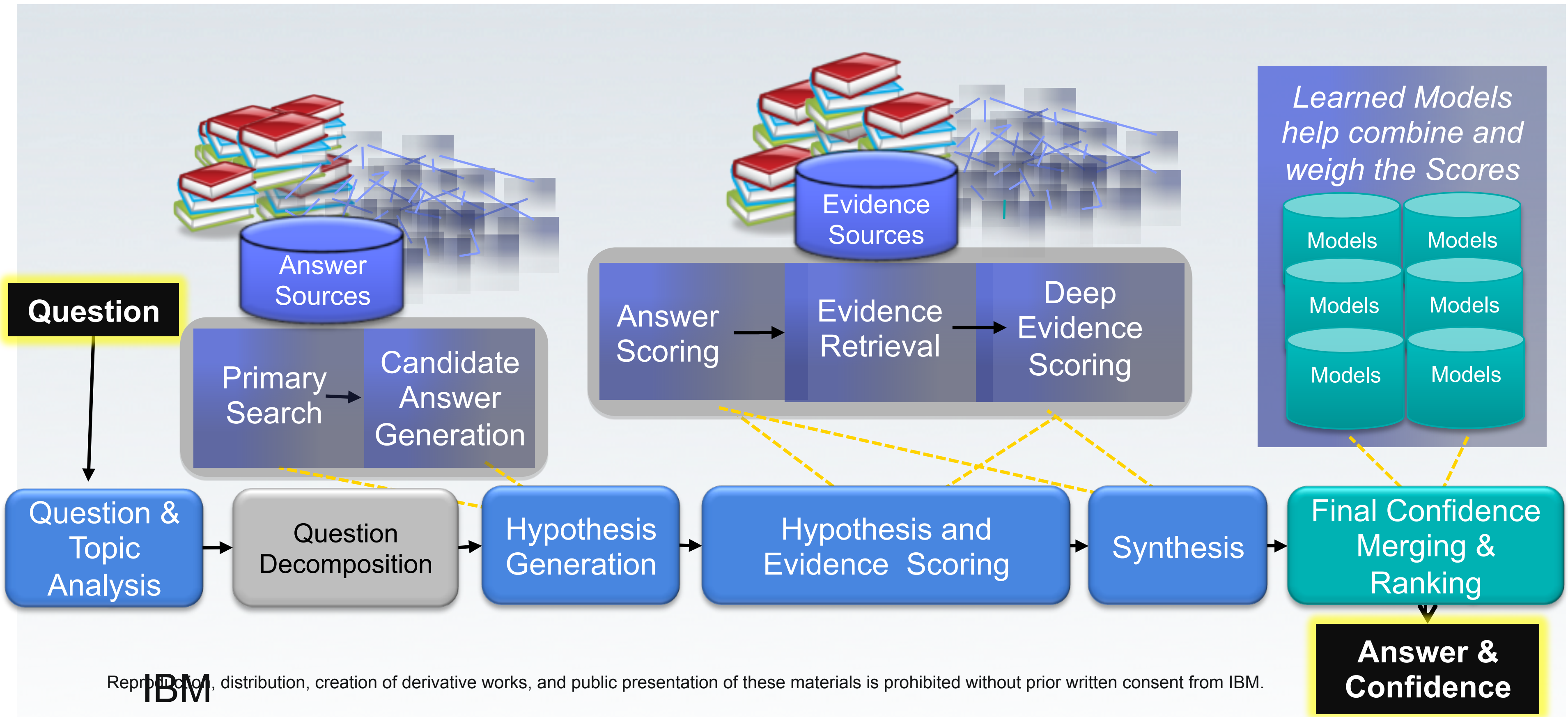
System runs all the algorithms on 10,000's of questions to generate features & answers and trains Machine Learning on them to produce a set of *models*



Game Time (Apply Time)

System runs all the algorithms on a *single question* and applies models to the features they produce to select an answer with a confidence

Game-time



Ranking Problem:

- Given
  - Training Instances (features) $X = \{x_0, x_1, \ldots x_N\}$
  - Training Labels $Y = \{y_0, y_1, \ldots y_N\}$
  - Where label represents correctness: $y_i = \{+1, -1\}$
- Develop a ranking technique to rank answers in order to maximize *accuracy* (percent of questions with correct answers in first place)
  - In practice we considered other metrics (incorporating confidence scores, etc)

- *Our Initial Solution*: We used logistic regression:
  - Rank according to y(x)
  - Use y(x) as confidence score

Logistic Regression

$$y(x) = \frac{1}{1 + e^{-(\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots)}}$$

Can you beat it?

Answer Merging

- Is it really a matrix?
- Multiple answers may be equivalent:

Candidate Answers	Evidence Feature Scores								
	Doc Rank	Pass Rank	Ty Cor	Geo	LFACS	Term Match	Temp-oral		
John F. Kennedy	0	1	0.1	0	0.2	22	1		
John F. Kennedy	2	1	0.1	0	0.4	41	1		
Kennedy	1	2	0.8	1	0.5	30	0.9		
J.F.K.	3		0.1	0	0	23	0.5		
JFK International Airport			0.9	1	0	10	0.5		
..		0	0.0	0	0	21	0.6		

Final Merger defines equivalency scorers and resources (e.g. Wikipedia redirects)
Answers identified as equivalent
What happens after merging?

Feature Merging

IBM WATSON

- Features for equivalent answers are merged to single value
- E.g. for all equivalent answers,
  - Use the minimum Doc Rank score
  - Use the maximum TyCor score
- Standard aggregation functions:
  - min, max, mean, decaying sum:

$$decay(p_0, \ldots, p_K) = \sum_{i=0}^{K} \frac{p_i}{2^i}$$

- How to pick merging policy?
  - Currently manually selected based on understanding of feature
- How to pick canonical form of merged answer?

Candidate Answers	Evidence Feature Scores								
	Doc Rank	Pas s Ran k	Ty Cor	Geo	LFAC S	Term Matc h	Temp-oral		
John F. Kennedy	0	1	0.1	0	0.2	22	1		
John F. Kennedy	2	1	0.1	0	0.4	41	1		
Kennedy	1	2	0.8	1	0.5	30	0.9		
J.F.K.	3		0.1	0	0	23	0.5		
John F. Kennedy International Airport			0.9	1	0	10	0.5		
..		0	0.0	0	0	21	0.6		

Candidate Answers	Evidence Feature Scores								
	Doc Rank	Pas s Ran k	Ty Cor	Geo	LFAC S	Term Matc h	Temp-oral		
John F. Kennedy	0	0	0.9	1	0.7	22	1		

Missing Value Imputation

- In Question Answering, scorers may not fire, e.g.
  - no Puzzle-Rank for Definition question candidates,
  - no Document rank if candidate generated from KB
  - no TWREX-Lookup (semantic relation) score if no relation is identified

- How to handle missing values?
- Default behavior: substitute with set mean
- More sophisticated missing value handling:

```
4,-7.192728,1,?,0,2.401191,…
6,-7.278166,1,?,0,4.590531,…
8,-7.351214,1,?,0,2.708208,…
?,?,1,?,1,2.241513,0.52262,…
?,?,1,?,1,0.000046,0,0,0,0,…
?,?,1,?,1,6.313291,1.272797,…
```

1. Impute to per-question mean
2. Allow user-defined rules to impute (use mean, max, 0, etc)
3. Expand `feature` into `feature-score`, `feature-missing` flag

?,?,1,?,1,2.241513,0.52262,…

↓

6,true,-7.2736,true,1,false,0,true,1,false,2.241513,false,0.52262,false,…

Feature-missing flag *allows reasonable imputation to be learned by the classifier*.
With `feature-missing` flag, selecting right imputation rule is not important [Lally]

Enhancement: Sparse Filter to avoid Overfitting

- In Question Answering, scorers may fire too rarely to provide statistically significant information
  - Suppose a scorer fires once on a correct answer
  - Classifier can learn that it is sufficient for correctness, give very high weight
  - (e.g. sparse SATTACK features led to nearly 5% drop on Useless questions)
- Existing solution: Implement a sufficient sample size filter
- Further research: Integrate with learning
  - Confidence-weighted learning, regularization, etc.

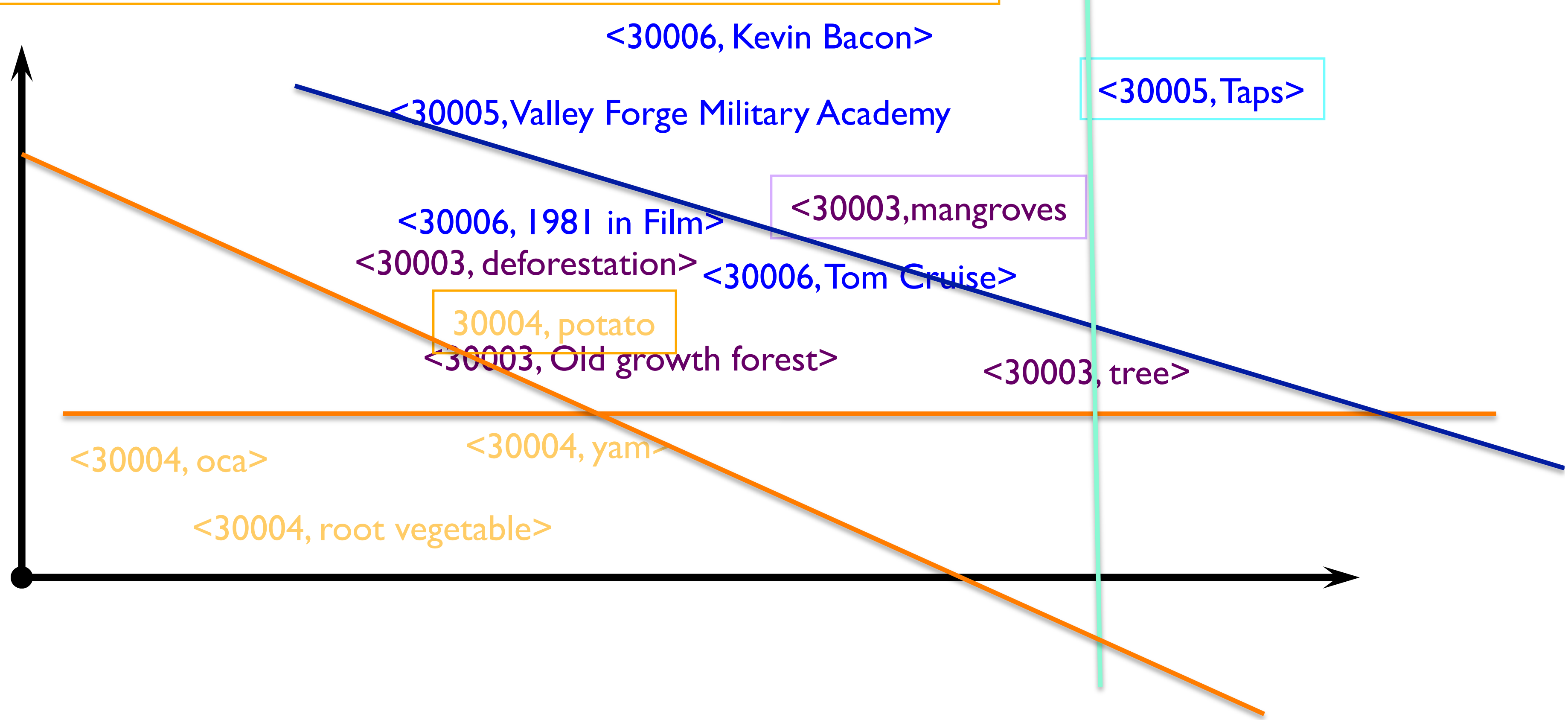
	label
1,-7.192728,1,?,0,2.401191,…,	true
0,-7.278166,1,?,0,4.590531,…,	false
?,-7.351214,1,?,0,2.708208,…,	false
?,?,1,?,1,2.241513,0.52262,…,	false
?,?,1,?,1,0.000046,0,0,0,0.2,…,	false
?,?,1,?,1,6.313291,1.27279,…,	false

Examples: Per Query variation

30003: Forests of this tree supported by stiltlike, above-ground roots grow in Florida’s coastal area

30006: Tom Cruise, Timothy Hutton, and Sean Penn played military cadets in this 1981 film

30005: In Dutch, these tuber vegetables are known as aardappelen

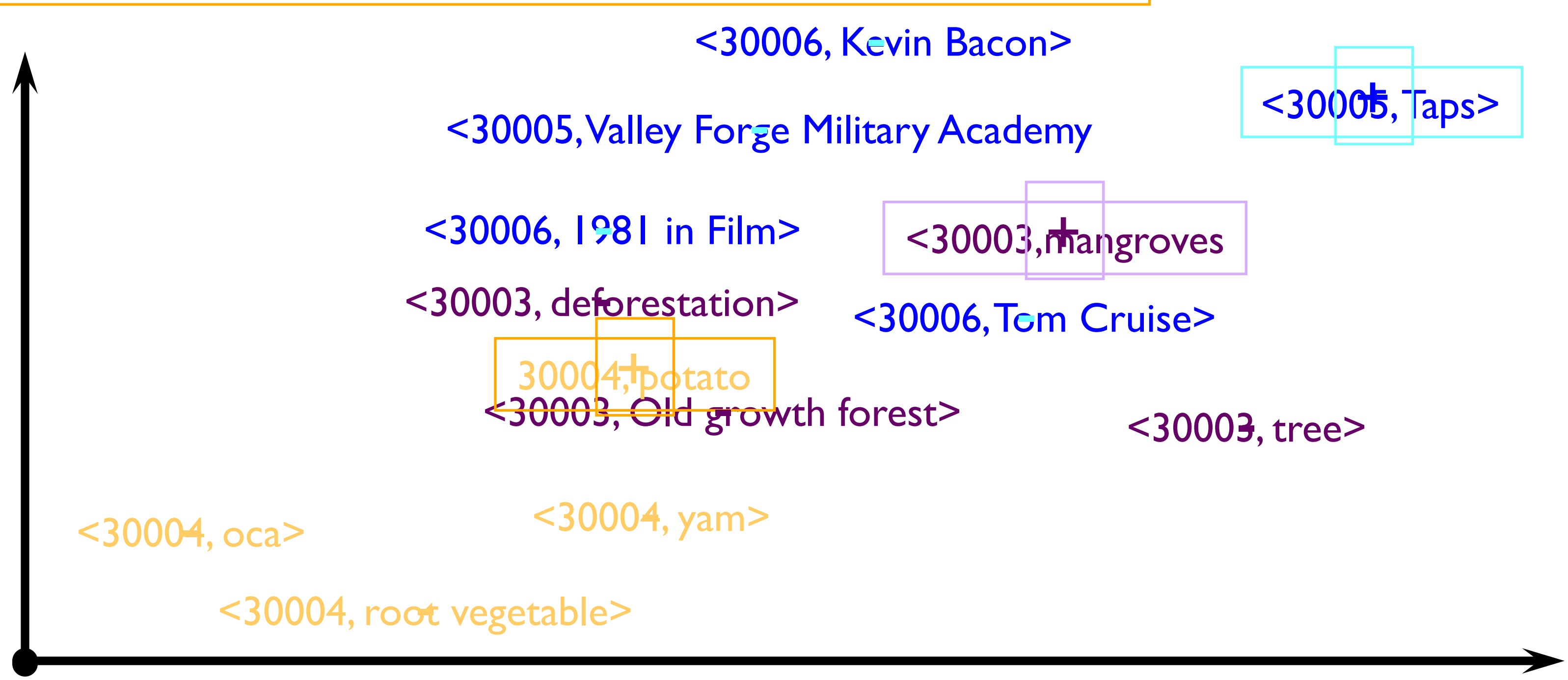


Examples

30003: Forests of this tree supported by stiltlike, above-ground roots grow in Florida's coastal area

30006: Tom Cruise, Timothy Hutton, and Sean Penn played military cadets in this 1981 film

30005: In Dutch, these tuber vegetables are known as aardappelen



30003 Candidate	Align	TyCor
Deforestation	0.6	0.5
Mangroves	0.7	0.7
Old growth forest	0.4	0.5
Tree	0.07	1.0

30006 Candidate	Align	TyCor
Kevin Bacon	1.2	0.7
Taps	1.1	1.0
Tom Cruise	.75	0.7
…		

30005 Candidate	Align	TyCor
Potato	0.5	0.5
Taps	0.3	0.5
Tom Cruise	0.3	0.1
Root vegetable	0.1	0.2

Machine Learning - Improvements

Per-Query Normalization

Normalized Feature Representation:

- Questions vary in difficulty
  - But we are training over all equally!
  - "Best" answer is relative *per question*

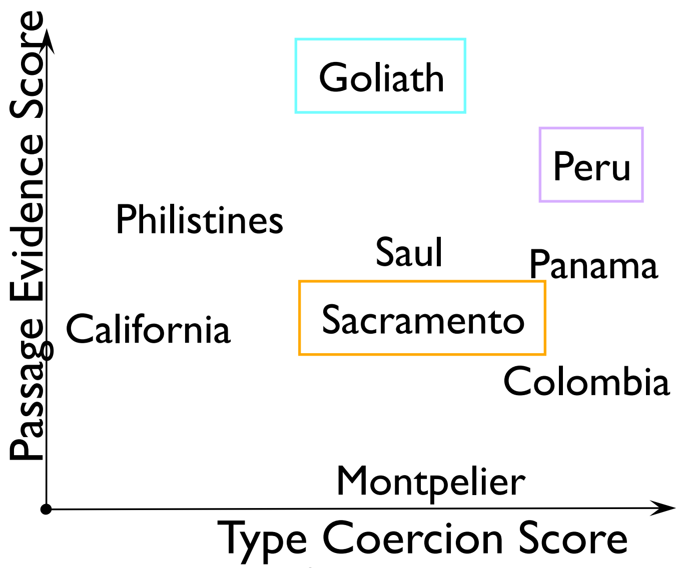
Standardize each candidate set per query: scale and center at 0. (Motivated by [Cao et al. 2006])

- *Objective can be shown mathematically similar to equivalent ranking objectives (e.g. RankSVM).*
- *Outperforms RankSVM [Hebrich 2000, Joachims 2002] by 20% relative*

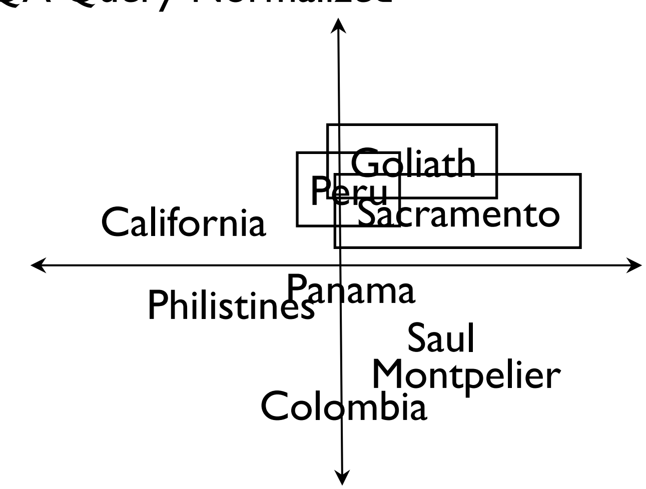
$$x_{ij}^{\text{std}} = \frac{x_{ij} - \mu_j}{\sigma_j}, \quad \mu_j = \frac{1}{|Q|}\sum_{k=1}^{|Q|} x_{kj},$$

$$\sigma_j = \sqrt{\frac{1}{|Q|}\sum_{k=1}^{|Q|}(x_{kj} - \mu_j)^2}.$$

Conventional



DeepQA Query-Normalized

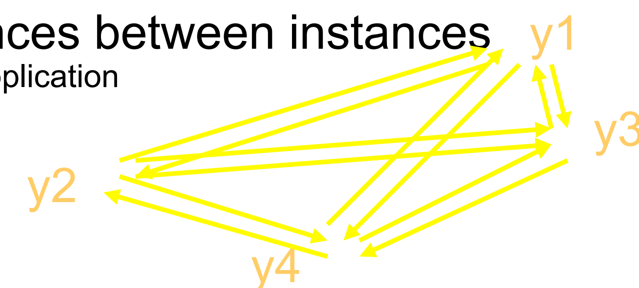


IBM WATSON

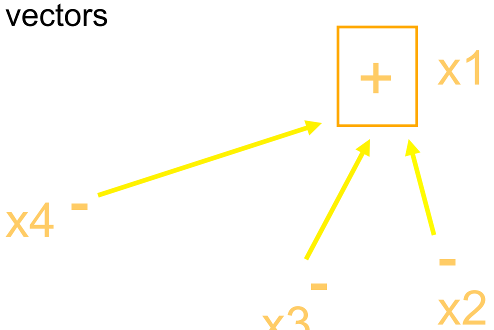
Rank-inspired Approaches: Difference Learning

- Rank Learning: train based on differences between instances

Application



Train on difference vectors



+ x1-x2: < 0.5,-0.1>

+ x1-x3: < 1.0, 0.8>

+ x1-x4: < 1.2, 0.6>

x2-x1: < -0.5,0.1>

.

.

- **Application**
  - Linear models (SVM, logistic) may be applied directly
  - More sophisticated models:
    - Calculate difference vectors
    - Apply model
    - Resolve (vote/tournament)
  - Did not show significant performance gains
  - Learning To Rank (w/ Melville et al showed gains)

Kernel Methods for passage scoring

Kernels are similarity functions that can be applied to measure the similarity between two text

- Linear Kernel
- Sequences (String Kernel, Word sequence kernel)
- Syntactic Structures (Tree Kernel)
- Similarity in a topic model (Domain Kernel, LSI)
- Polynomial, Radial Basis Function, etc.

A kernel is a function $K : \mathcal{X} \times \mathcal{X} -> \mathbb{R}$ such that

(18) $$K(x_i, x_j) = (def)\langle \mathbf{\Phi}(x_i), \mathbf{\Phi}(x_j)\rangle$$

where $\mathbf{\Phi} : \mathcal{X} -> \mathcal{K}$ is a feature mapping.

Experimented with different kernels for confidence estimation, ranking:
No significant gain.
SVM did not outperform logistic regression

Final Combination

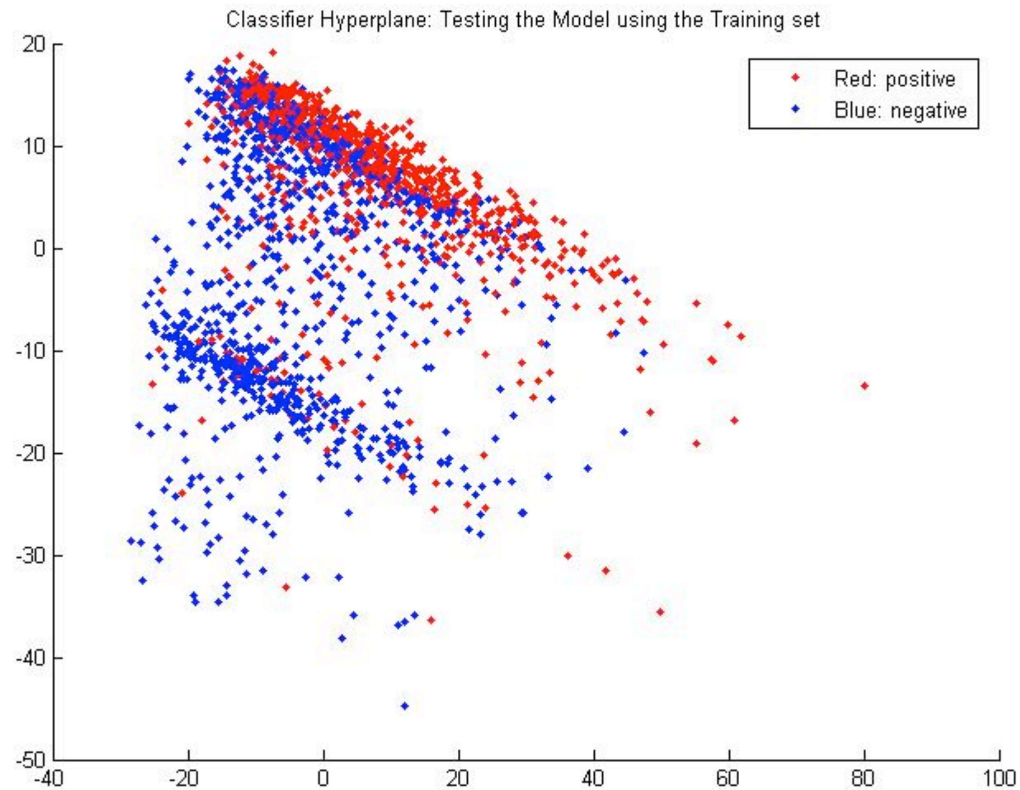
- Baseline: Logistic regression on base features
  - Score provides sort-order
  - Use logistic score as confidence
- Improvements:

	Improvement?
SVM	No
nonlinear kernels (rbf, poly, etc)	No
Neural Nets	No
Decision trees, LMTs	No
Class balancing (downsampling)	No
Cost-sensitive Learning	No
Feature Space Dimensionality Reduction	No
Feature Selection (Information Gain Ratio)	No
Separate rank/estimation models	No

	Improvement in accuracy at discovery	
Basic (baseline)	**+0.00%**	
Specialized Models	+1.0%	
Standardization	+3.2%	
Augmenting Basic with Standardized	+5.5%	
Ranking Objective (Rank-SVM)	+4.5%	
Imputation (with missing flags)	+1.7%	

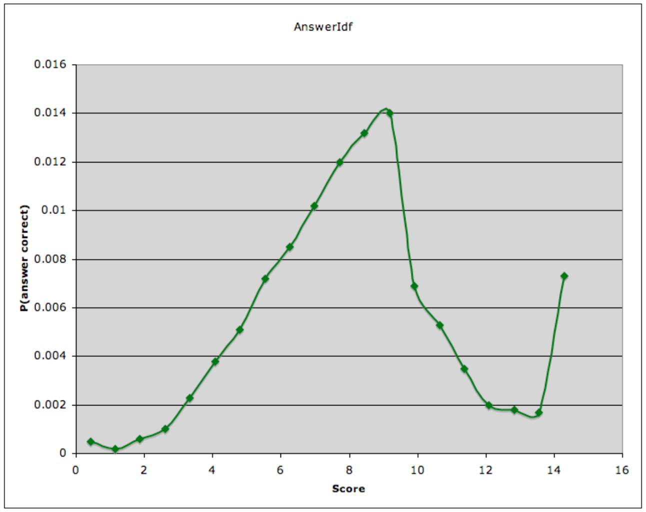
Class Imbalance

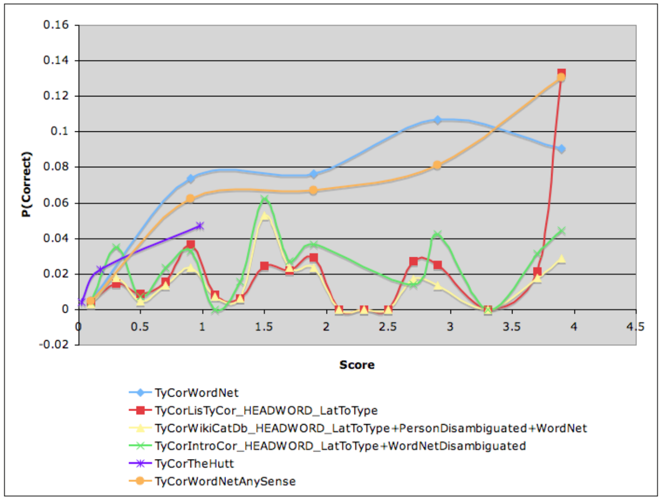
- 94-to-1 Negative to Positive instances
- Experimented with *cost-sensitive learning, instance weighting.*
- Ultimately best results in Jeopardy! set were using instance weighting of 0.5 for negative instances. (note this is not consistent for other data sets)



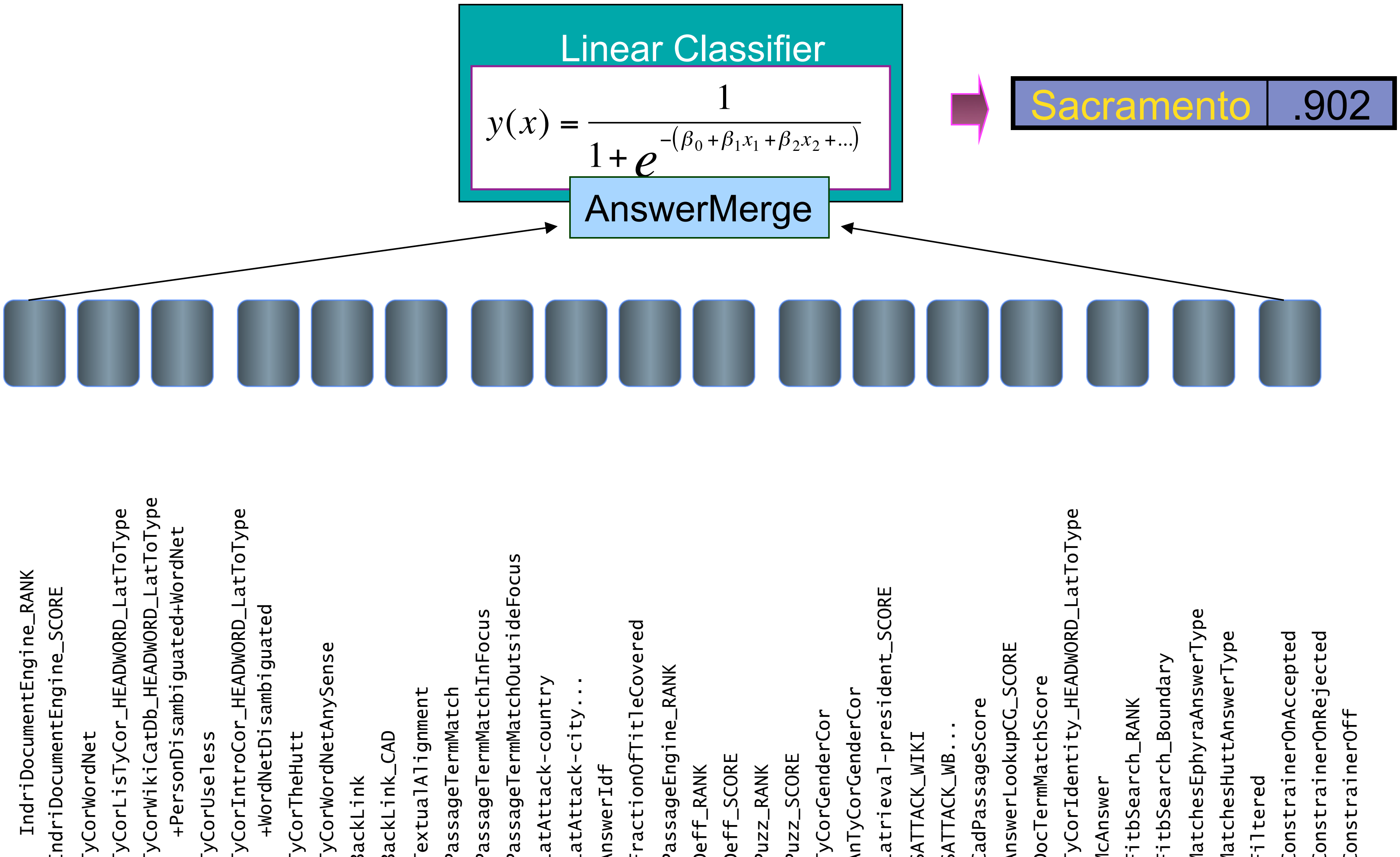
Feature Engineering

- How to handle non-monotonic
  - Non-linear Kernel?
  - Binning?
    - These exist for AnswerIdf [Lally]
  - Scaling to probability [Prager]
  - Internal improvements to scorer

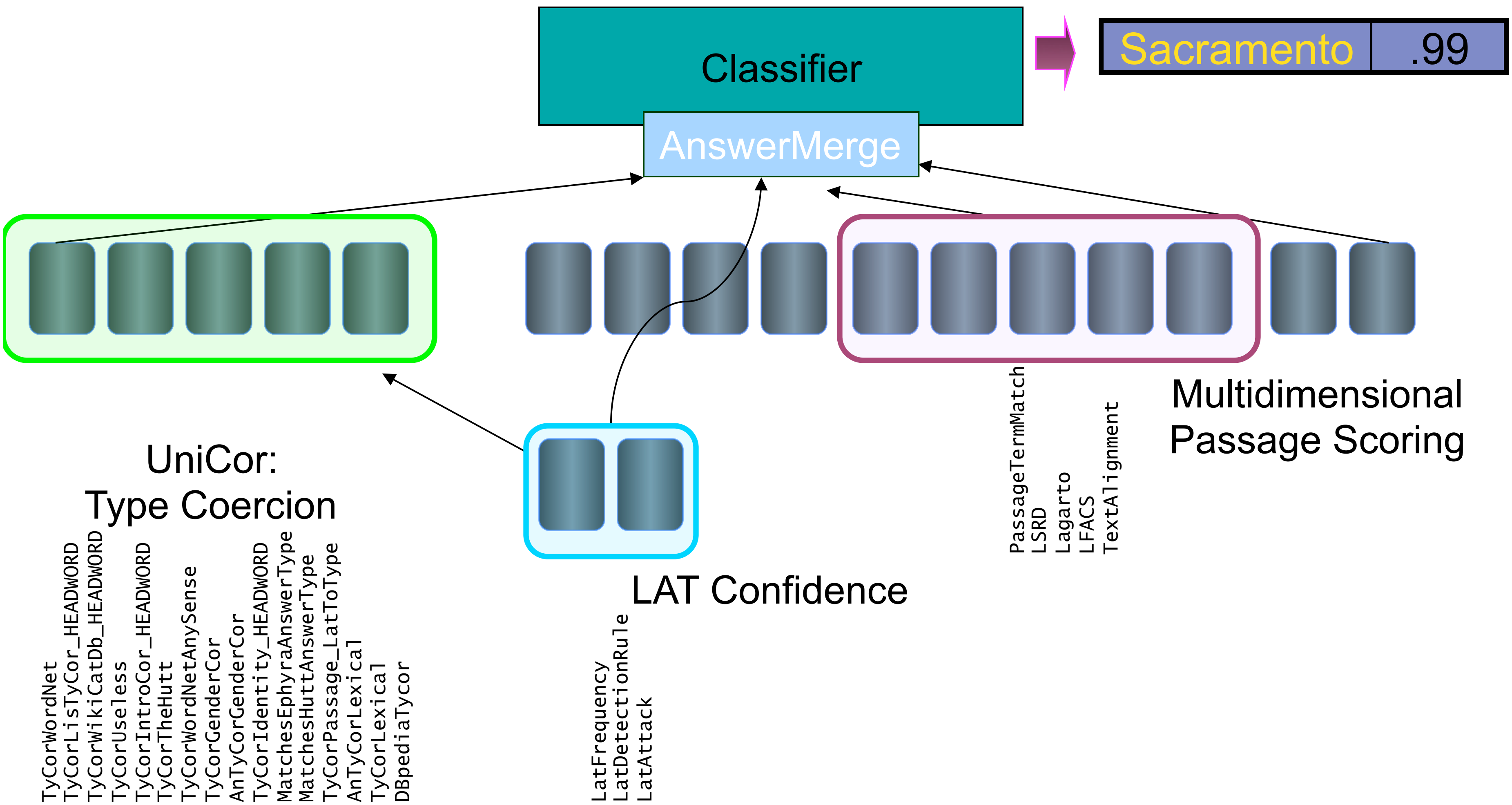




Stacking: Linear classifier over all components (flat)



Stacking: Hierarchical ML



Supporting Passage Retrieval (SPR)

IBM WATSON

Category: MICHIGAN MANIA

Clue: In 1894 C.W. Post created his warm cereal drink Postum in this Michigan city

In Deep Evidence Scoring, Watson retrieves evidence for each candidate answer, then evaluates the evidence using a large number of deep evidence scoring analytics. The evidence for a candidate answer may come from the original document or passage where the candidate answer was generated, or it may come from an evidence retrieval search performed by taking the keyword search query from Step 2, replacing the focus terms with the candidate answer, and retrieving the relevant passages that are found. The passages, or "context" in which the candidate answer occurs are evaluated as evidence to support or refute the candidate answer as the correct answer for the question.

Battle Creek

1895: In Battle Creek, Michigan, C.W. Post made the first POSTUM , a cereal beverage. Post created GRAPE-NUTS cereal in 1897, and POST TOASTIES corn flakes in 1908

The company was incorporated in 19 having developed from the earlier Postum Cereal Co. Ltd., founded by C.W. Post (1854-1914) in 1895 in Ba Creek, Mich. After a number of experiments, Post marketed his first product-the cereal beverage called Postum-in 1895

Post Foods

Post Foods, LLC, also known as Post Cereals (formerly Postum Cereals) was founded by C.W. Post. It began in 1895 with the first Postum, a "cereal beverage", developed by Post in Battle Creek, Michigan. The first cereal, Grape-Nuts, was developed in 1897

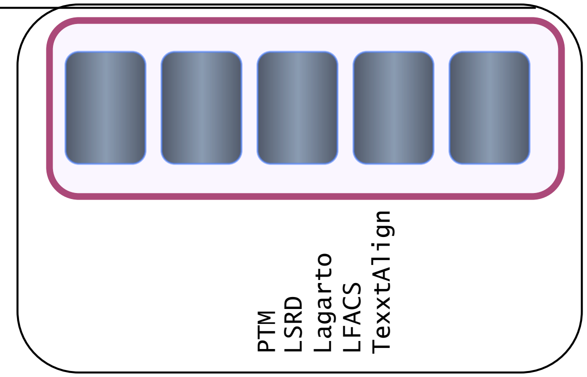
r of became General Foods. The cereal company unit was later sold off and is now Post Foods

General Foods

1854 C. W. Post (Charles William) was born in 18 in 19 beve

General Foods' products go from breakfast (Post's cereals) to warm nightcaps (Postum, Sanka), also wash the pots and pans that its foods are cooked in (S.O.S. Scouring Pads

Multiple Relations: Multidimensional Relation Scoring



- Question:
  - This ***Dublin***-born ***star*** ***of “The Usual Suspects”*** never set foot in America until he was 37.

Gabriel Byrne	Score	Dublin-born	never set foot in America until he was 37
	IMDB-LU	**1.0**	**?**
	Prismatic	**0.75**	**0.10**
	Lagarto	**0.63**	**0.52**
	TextAlign	**0.48**	**0.21**

Classifier

score

Gabriel Byrne started acting at the age of 29 and he went to ***America*** for the first time when he was 37

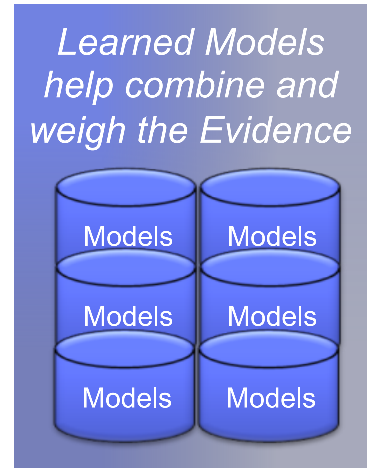
Mixture of Experts: Question Classes

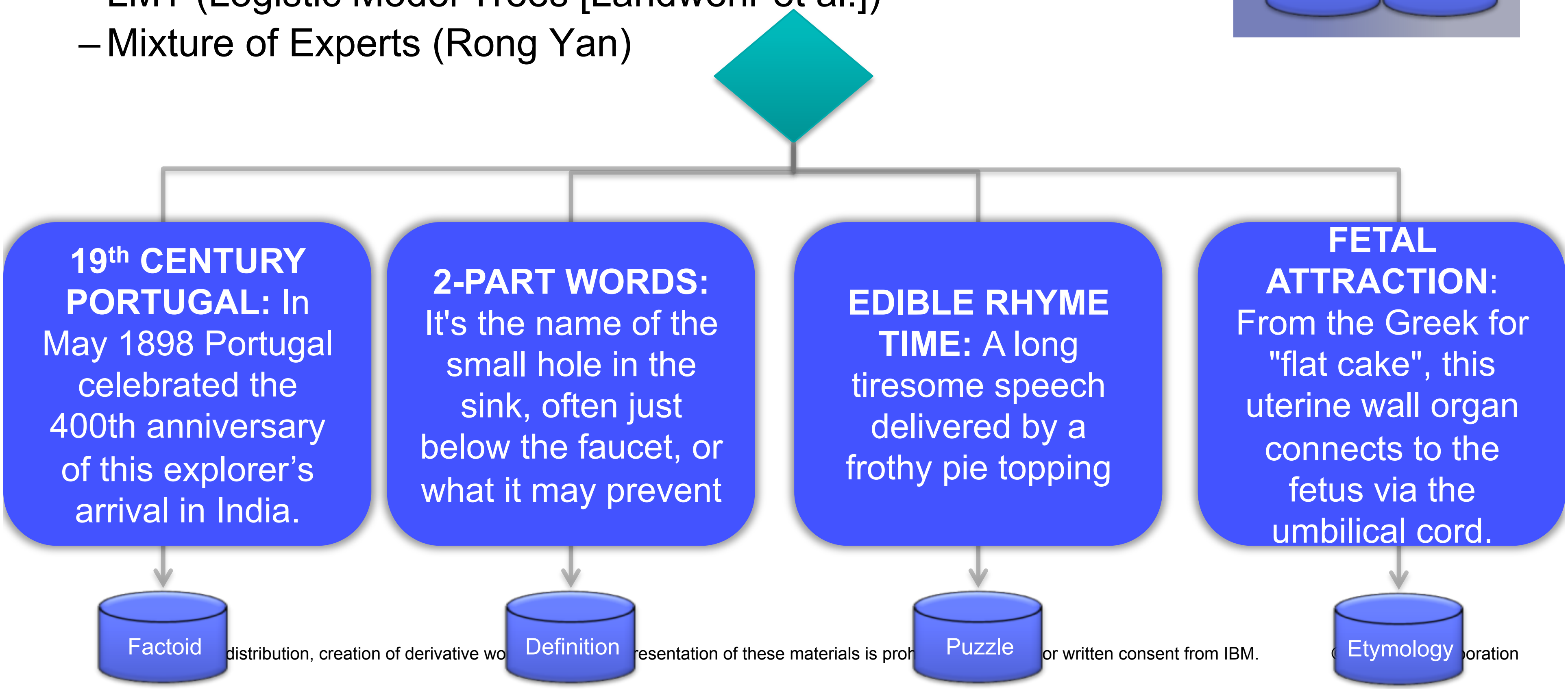
Different question classes weigh evidence differently

Statistical and rule-based classifiers identify question class

Partitioned Mixture of Experts trained for each question class

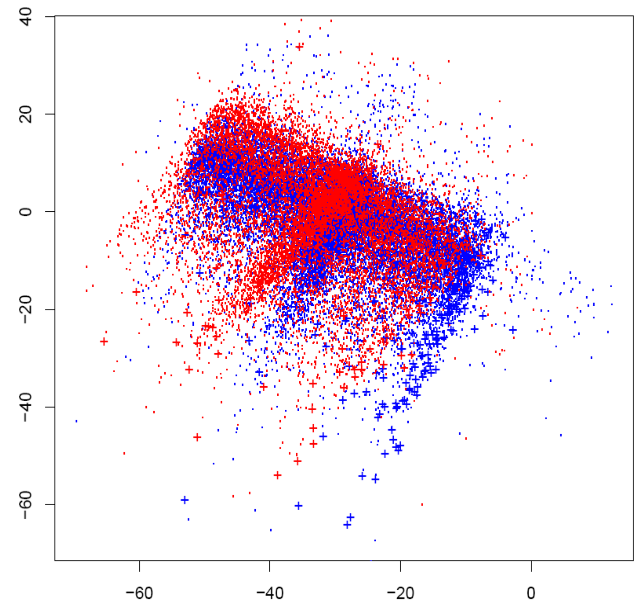
- Experimented with automatic techniques -
  - –Decision trees
  - –LMT (Logistic Model Trees [Landwehr et al.])
  - –Mixture of Experts (Rong Yan)





Transfer Learning

- We may have large amounts of general training data (e.g. 100,000 Jeopardy Questions)
- But a small amount of training data for a target class (e.g. 100 etymology questions)
- *Transfer Learning* uses a large amount of auxiliary data and attempts to transfer "useful" part to target task



Domain Adaptation in ML: Transfer Learning

- Visualization using top 2 principal components

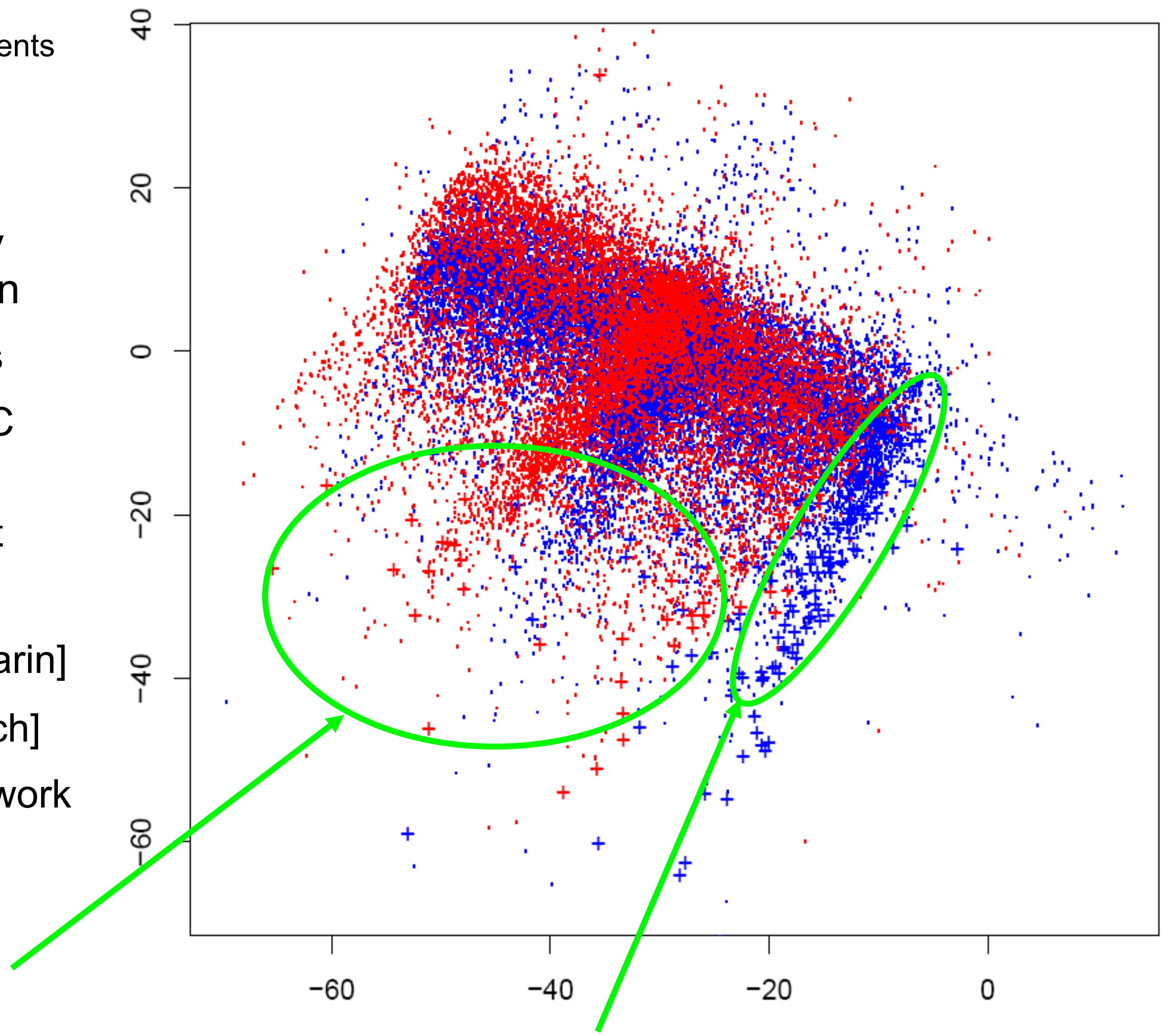
  blue Jeopardy

  red TREC

  Datasets share some similarity but are not same distribution

  - Slight skew in false instances
  - Positive instances from TREC have higher variance
  - Similar over all true/false split

- Adapt J! data in training with TREC
  - Migratory Logit [Liao, Xue, Carin]
  - Transfer SVM [Wu & Dietterich]
  - Semi-supervised CEM [joint work with CMU]



Improving Confidence by Rescaling: Producing probabilities from confidence scores

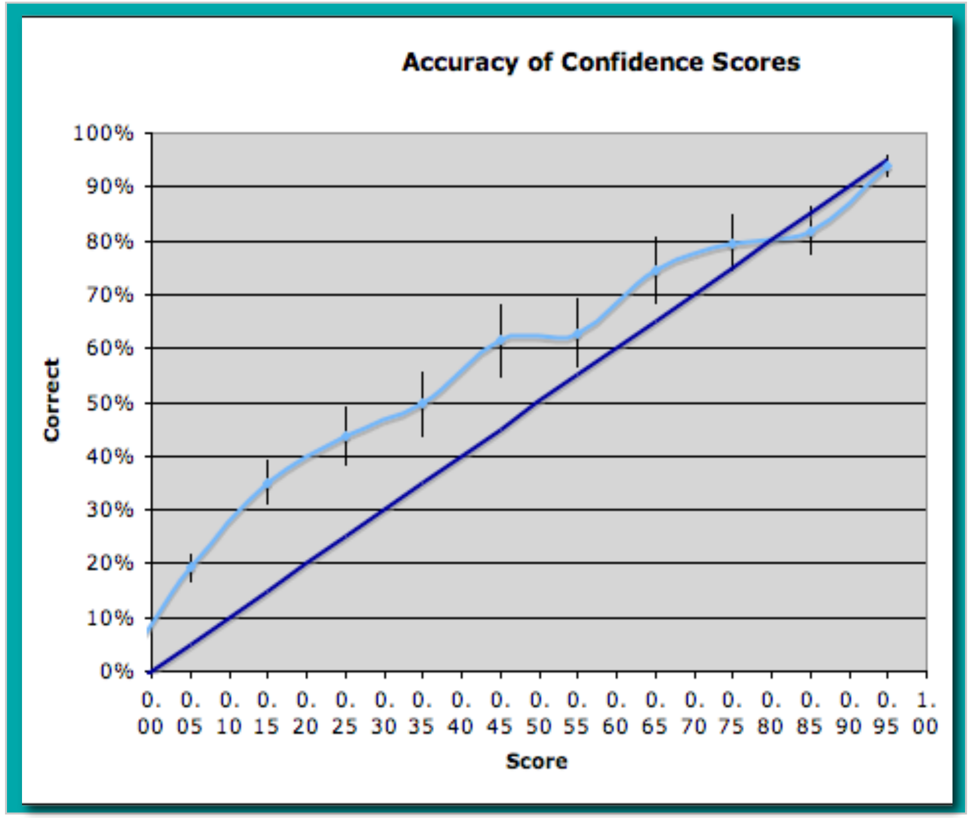
- Produce probabilities of correctness
  - Compare from multiple models
  - Compose answers probabilistically
- Accurate probabilities aid in combining answers:

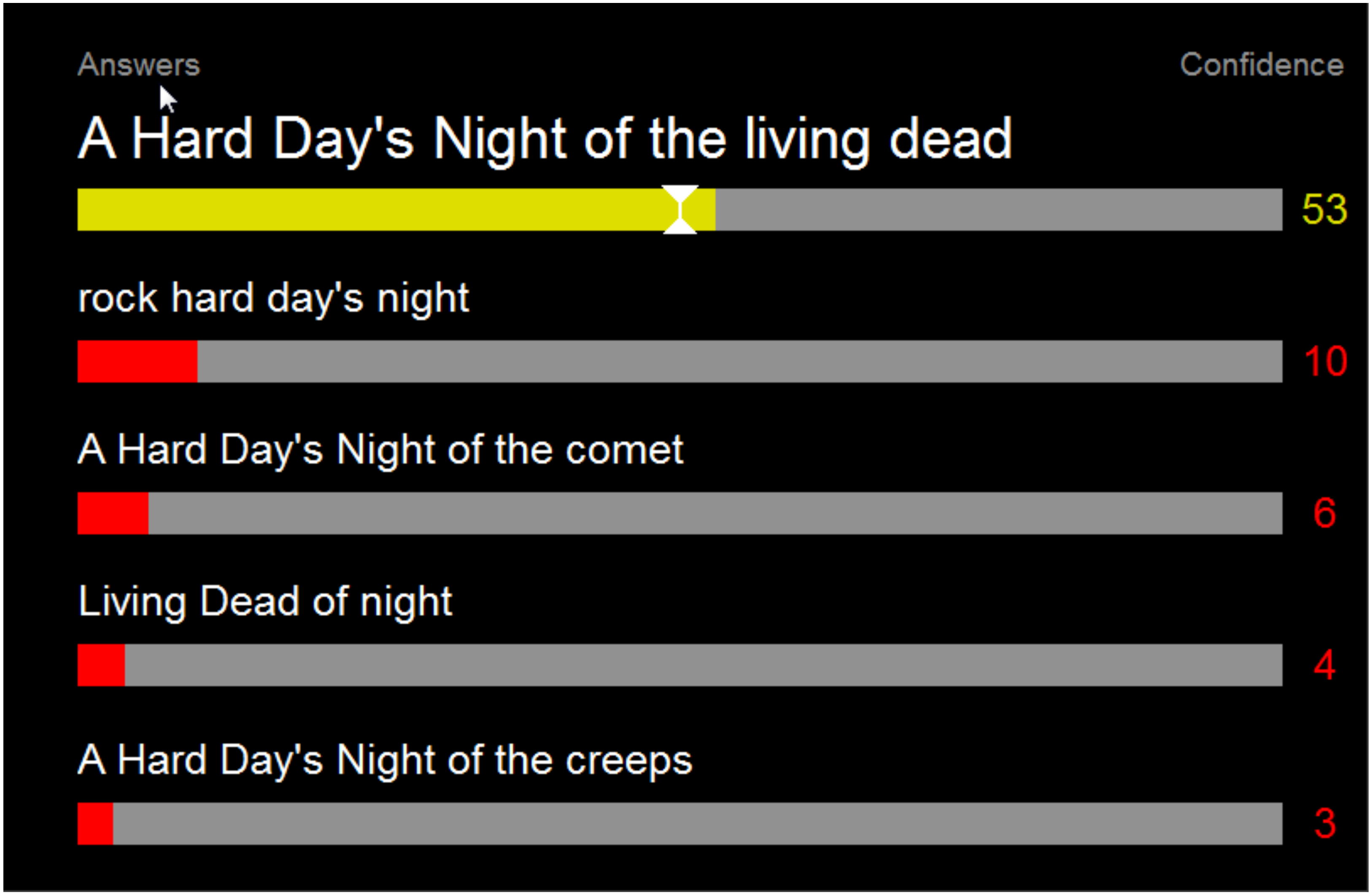
THIS AND THAT (1200): The 2 continents separated by the Dardanelles

Europe and Asia

INTERNATIONAL ROAD VEHICLE STICKERS (1600):2 West Indian islands, one republic: TT

Trinidad and Tobago



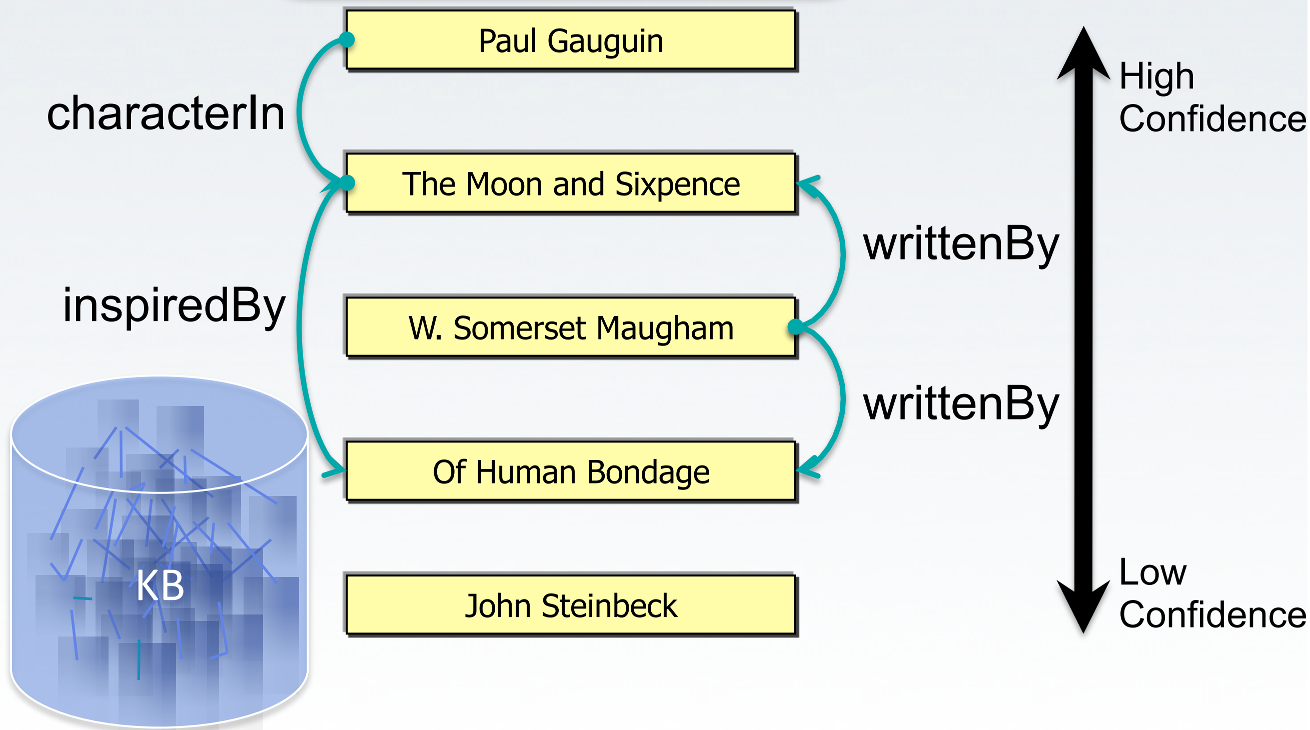

Answers
Confidence
A Hard Day's Night of the living dead
53
rock hard day's night
10
A Hard Day's Night of the comet
6
Living Dead of night
4
A Hard Day's Night of the creeps
3

IBM WATSON

Answers are related: Evidence Diffusion

Candidate Answers are not independent.

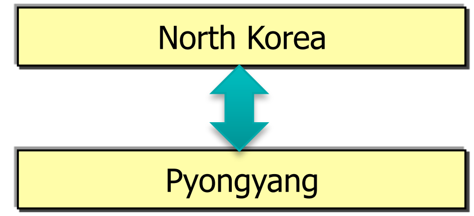
20TH CENTURY NOVELISTS: A critic said that a character of his, "Yearning for the moon...never saw the sixpence at his feet"; he made that into a title



By sharing evidence based on the relationships between candidate answers we raise the score of the right answer

Evidence Diffusion

WORLD TRAVEL: If you want to visit this country, you can fly into Sunan International Airport or . . . or not visit this country



If two answers (source and target) meet following conditions:

1. The target meets the expected answer type (in this example, "country").
2. There is a semantic relation between the two candidates (in this case, *located-in*).
3. The transitivity of the relation allows for meaningful diffusion given the question.

KB

- Append transferred features x'_m to target answer
- Train a new classifier with original and transferred features

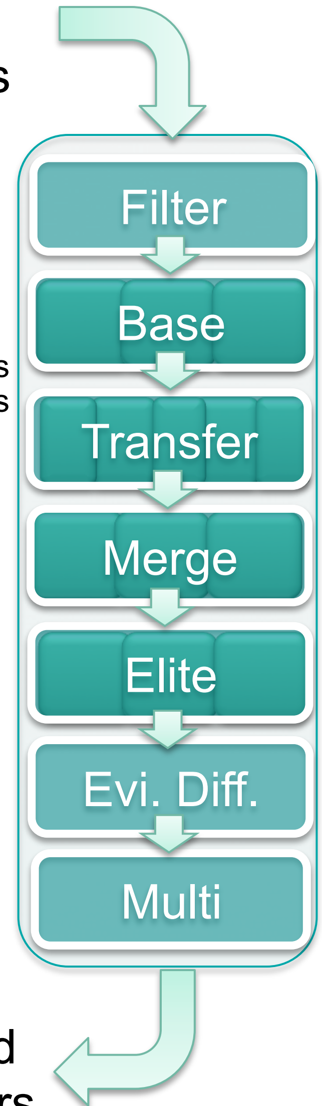
$$
\begin{aligned}
f'(x) &= \gamma_0 + \sum_m \gamma_m x'_m + \gamma_{M+1} f(x) \\
&= \gamma_0 + \sum_m \gamma_m x'_m + \gamma_{M+1}\left(\beta_0 + \sum_m \beta_m x_m\right), \\
&= \gamma_0 + \gamma_{M+1}\beta_0 + \sum_m \gamma_m x'_m + \gamma_{M+1}\beta_m x_m,
\end{aligned}
$$

Final Ranking: Classifier Phases

- **Final Ranking consists of successive application of a pipeline of classifiers.**
  - **Filter**: use light-weight features to select hypotheses for deeper analysis
  - **Base**: Initial ranking and confidence evaluation, specialized to main question types
  - **Transfer Learning:** Transfer-learning for special, rare question types
  - **Merge:** Merge equivalent hypotheses
  - **Elite:** Successive Refinement
  - **Evidence Diffusion:** Merge supporting hypotheses
  - **Multi-answers:** Join hypotheses for multiple-entity questions

Answer Features

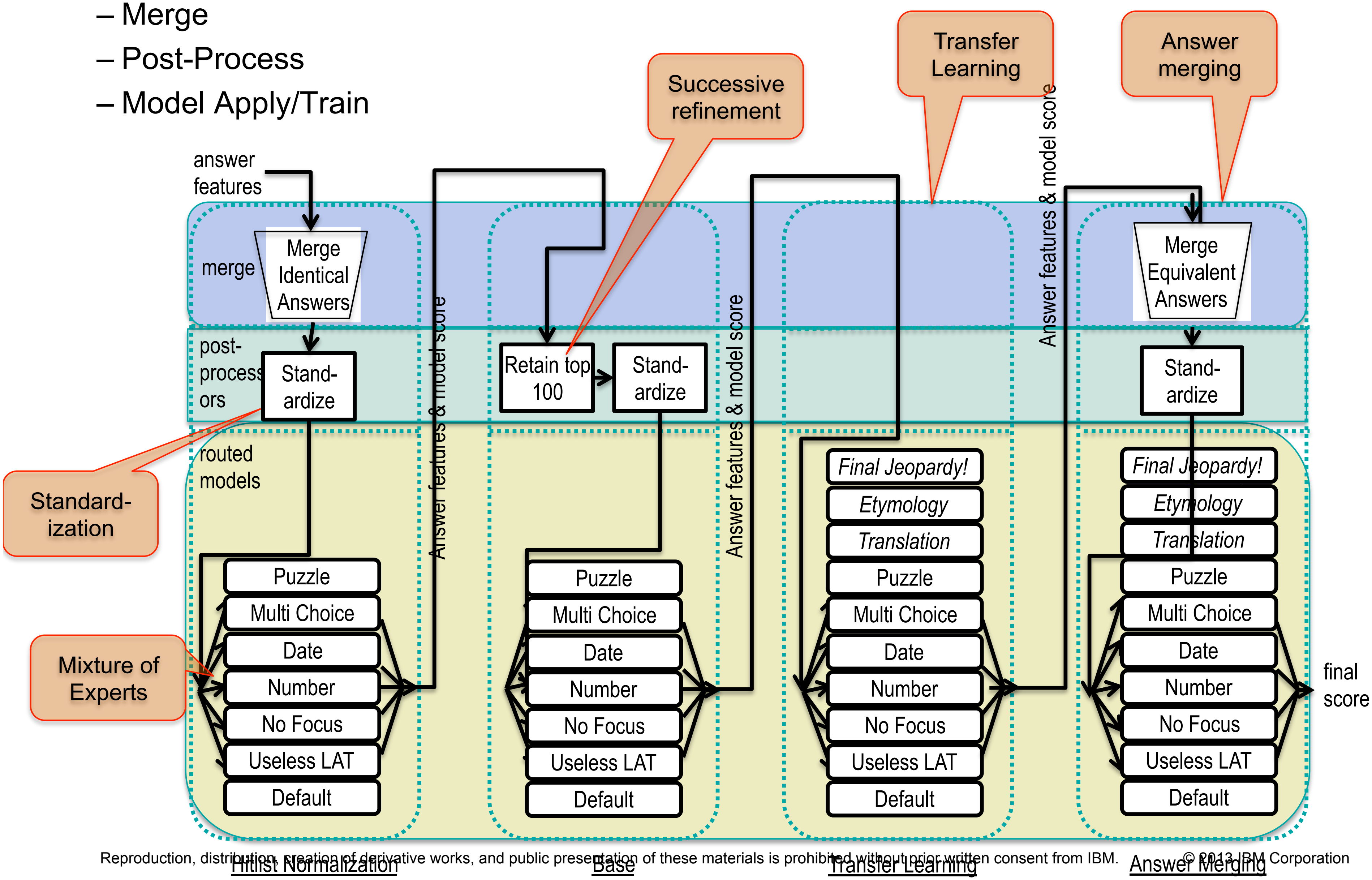
Each classifier has access to all features as well as previous classifier scores and ranks.



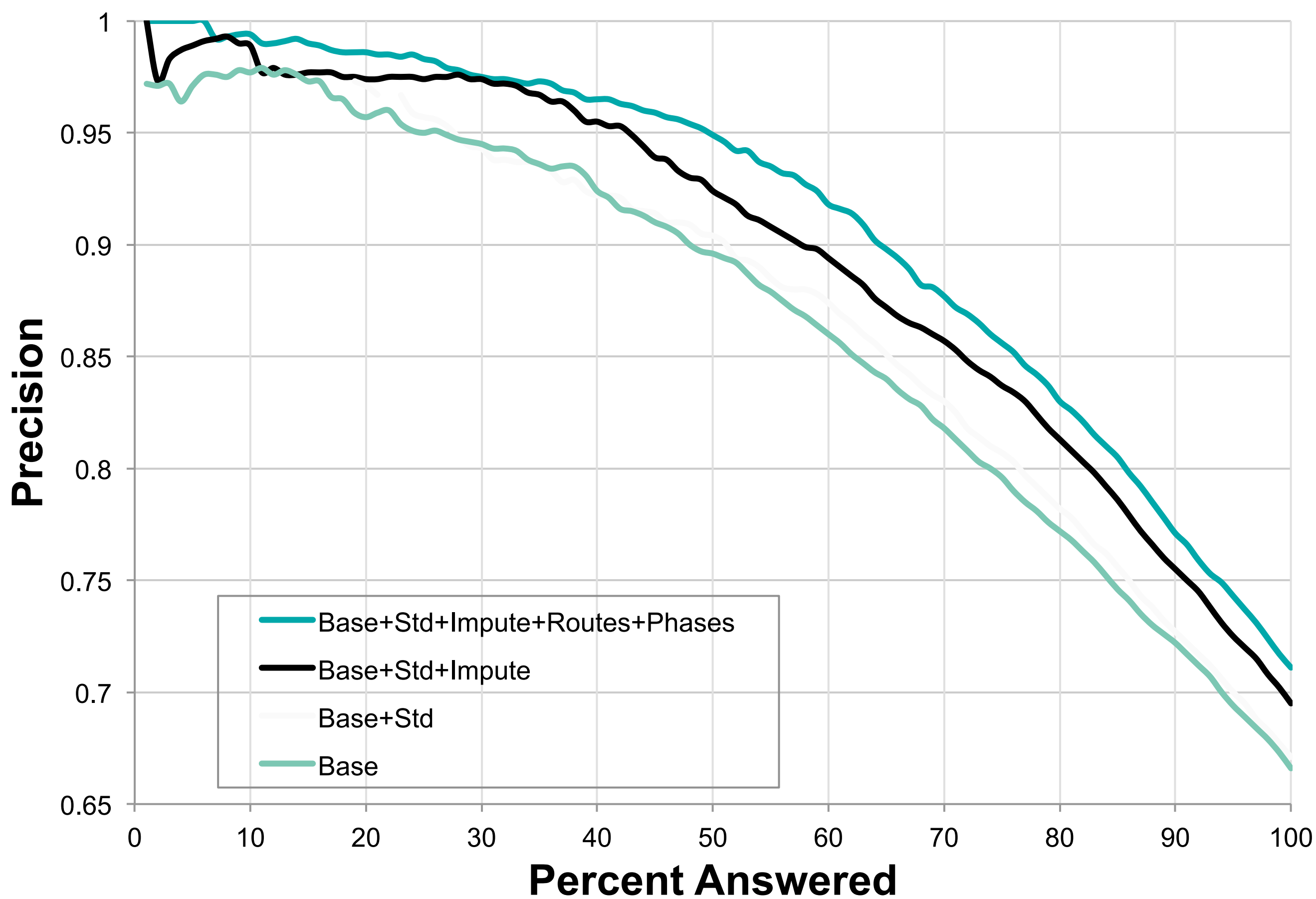
Ranked Answers

Phase-based Architecture

- Architecture to implement techniques
  - Merge
  - Post-Process
  - Model Apply/Train



DeepQA ML Enhancements



Working with a large, constantly adapting system

- –Results may be inconsistent week-to-week (e.g. Multiple Models)
- –Scorers are always changing (10,000s of experiments; each experiment requires training)
- –Models changing : Meta-learning methodology
- –Scorers added based on how they aid existing default ML technique
- –General scorers mixed in with more specialized: scorers don't always fire
- –Nondeterministic features
- –Sources changing, other features changing, *answer keys* changing

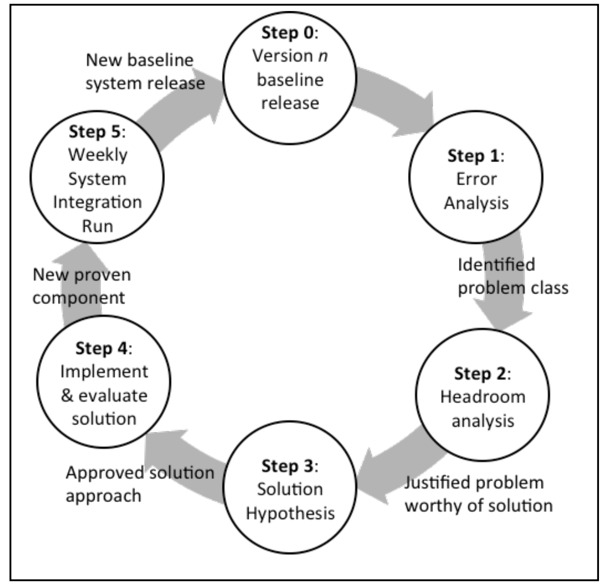


Figure 1 System development lifecycle.

Very different from setting where you have a ML expert to review each experiment

Machine Learning must be as automatic as possible!

Semantic Technologies in IBM Watson™

Lesson 7 – UIMA in Watson

Guest Lecturer: Siddharth Patwardhan
TA: Or Biran

Objectives

- At the completion of our two sessions, you should be able to:
  - Explain need for UIMA in Watson
  - Develop and run a UIMA application
  - Explain UIMA architecture in Watson
  - Setup, run and explain the UIMA framework for the class project

UIMA: What, Why, How

- What is UIMA: history and overview
- Why UIMA: why are we spending two lectures on this!
- How do I... : build a UIMA application from scratch?

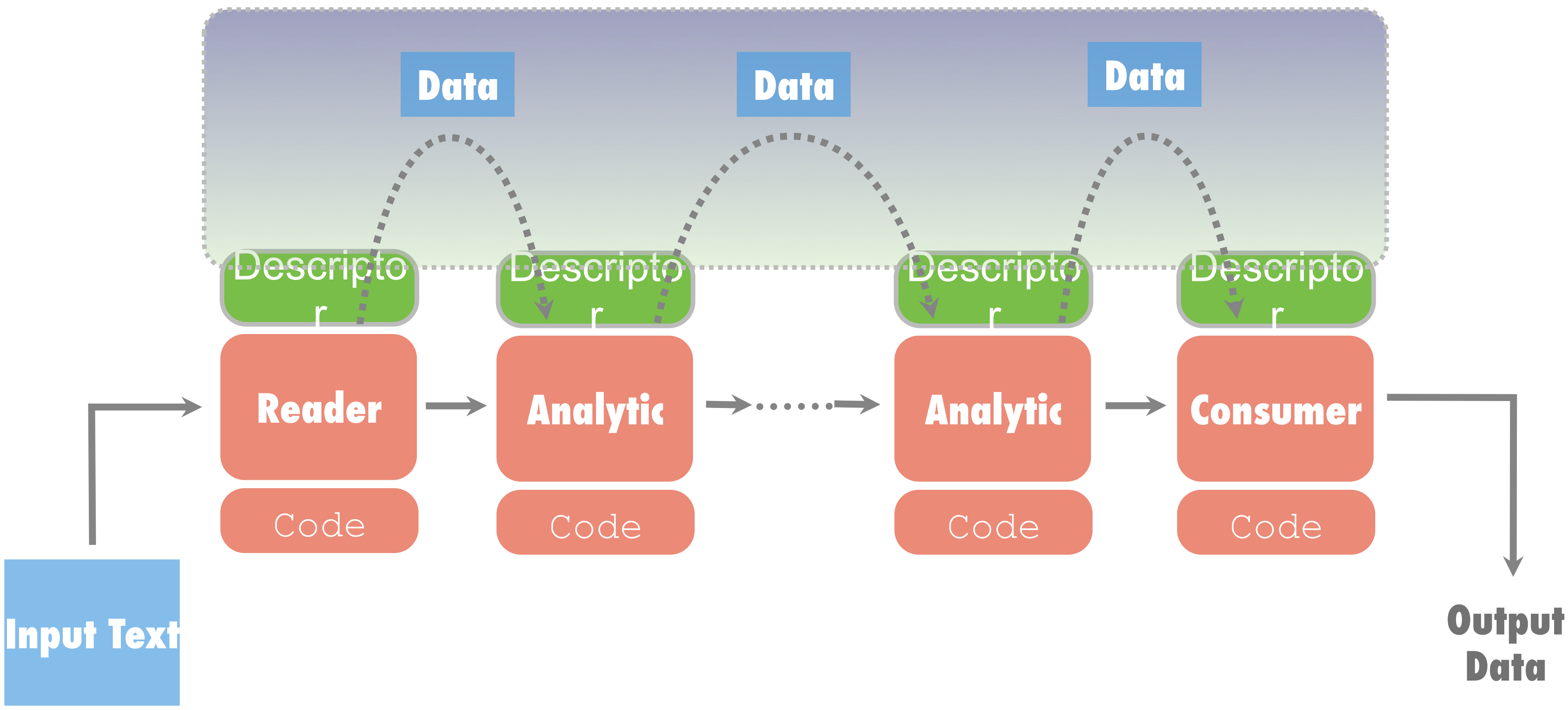
What is UIMA

- UIMA: Unstructured Information Management Architecture
- Developed at IBM, now Apache open source project
- Actively developed by a community of open source contributors
- http://uima.apache.org

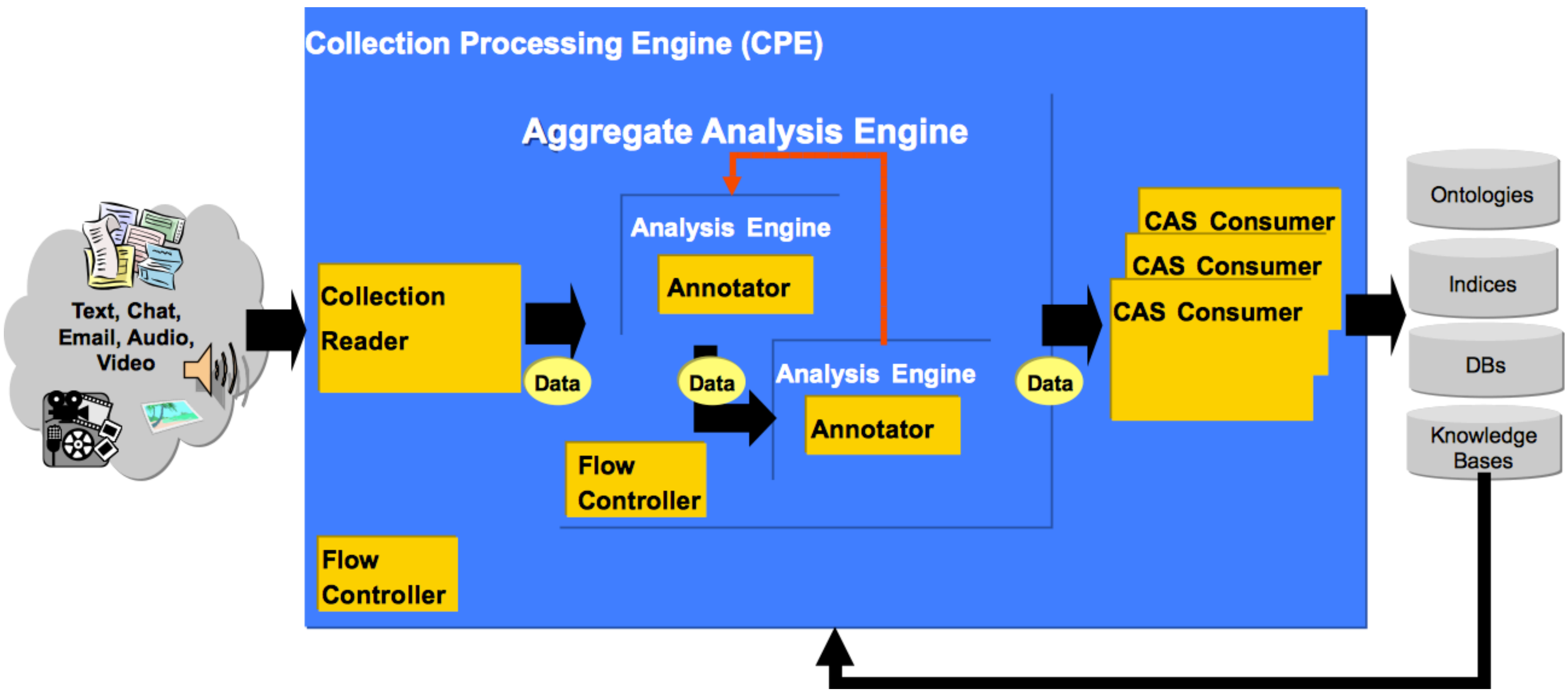
What is UIMA

- A specification, framework and SDK
- Provides a “plugin” approach for unstructured information management
  - Manages data structs / book-keeping
  - Easy to add / remove components
- Provides many development tools

The UIMA Framework



The UIMA Framework



Why UIMA

Building complex systems is (maybe) like...

Why UIMA

Building complex systems...

- Many different data representations
- Format/schema conversion costs
- Large overheads managing resources
- Steep learning curves

Why UIMA

- Why did we use UIMA in Watson:
  - "Plugin" approach makes development easy
  - Managed component interaction enables rapid development
  - Common data format fosters collaboration
  - Support for massive scaleout and speed

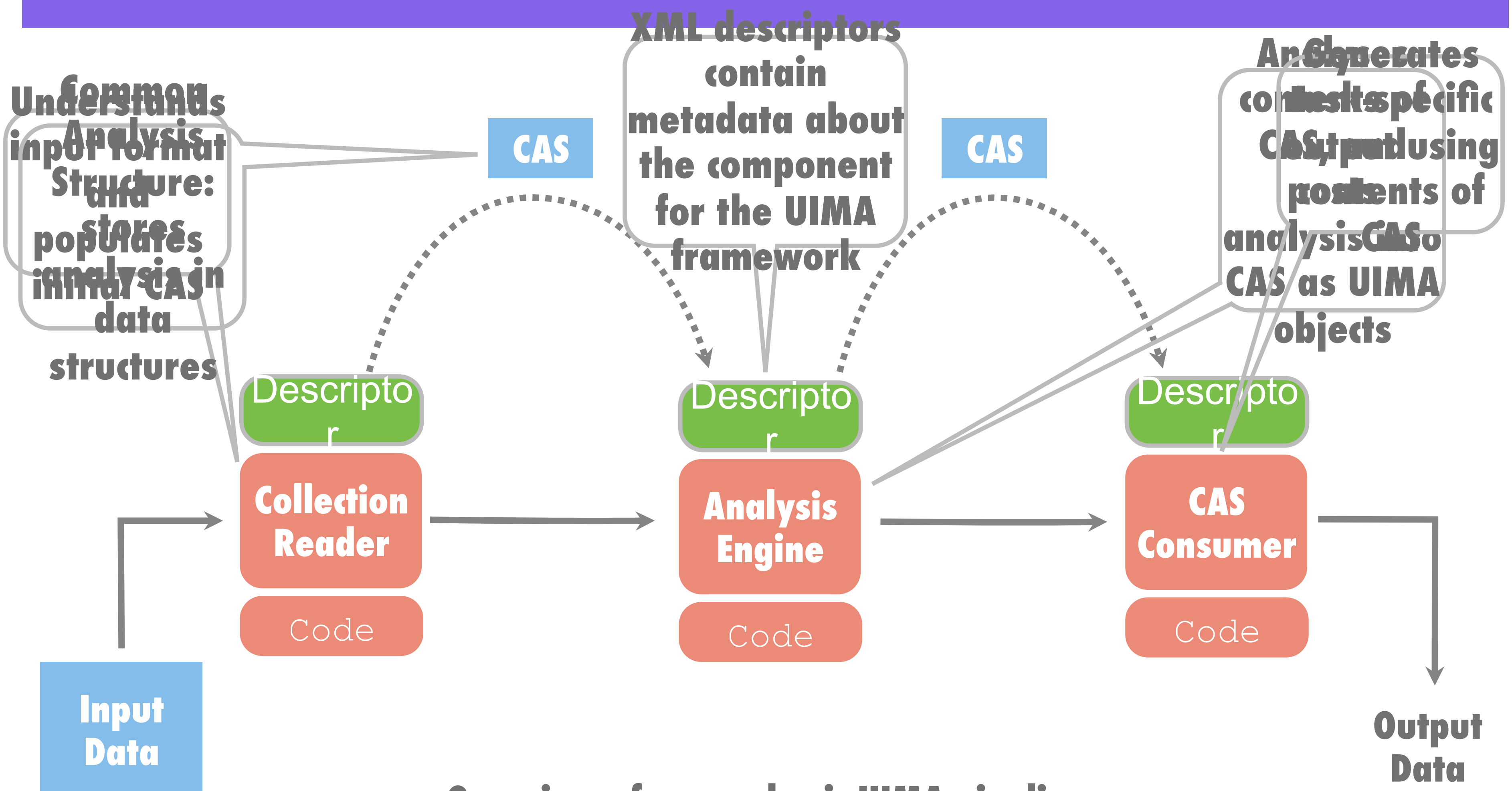
***Of course, having the original developers of UIMA on the Watson team factored into the decision too!**

Creating a UIMA App

Creating a UIMA App

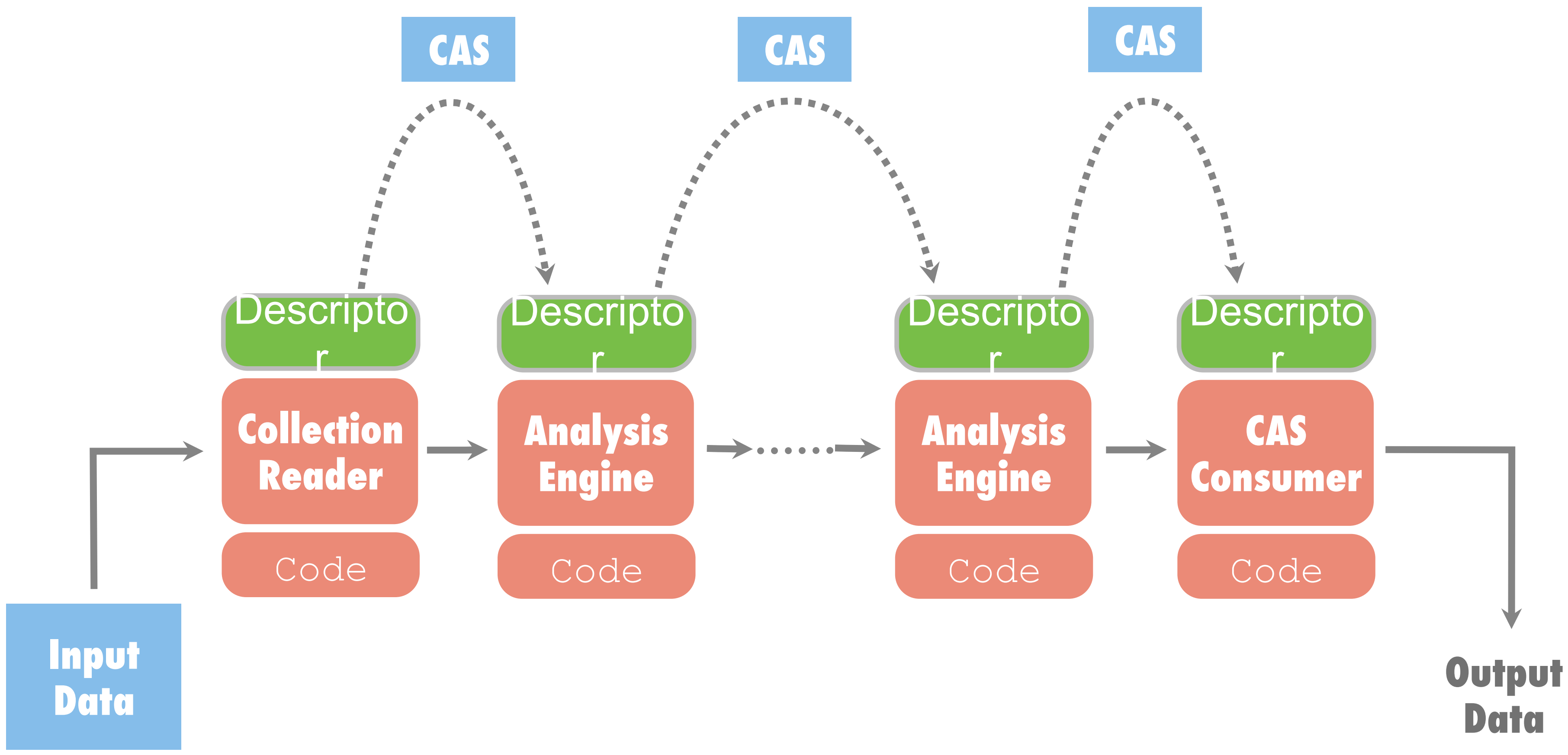
- Initial prep-work: UIMA SDK & Eclipse
- Defining the data types
- Writing the component code
- Defining the component properties
- Reading input and writing output
- Creating and running the pipeline

Building a UIMA App

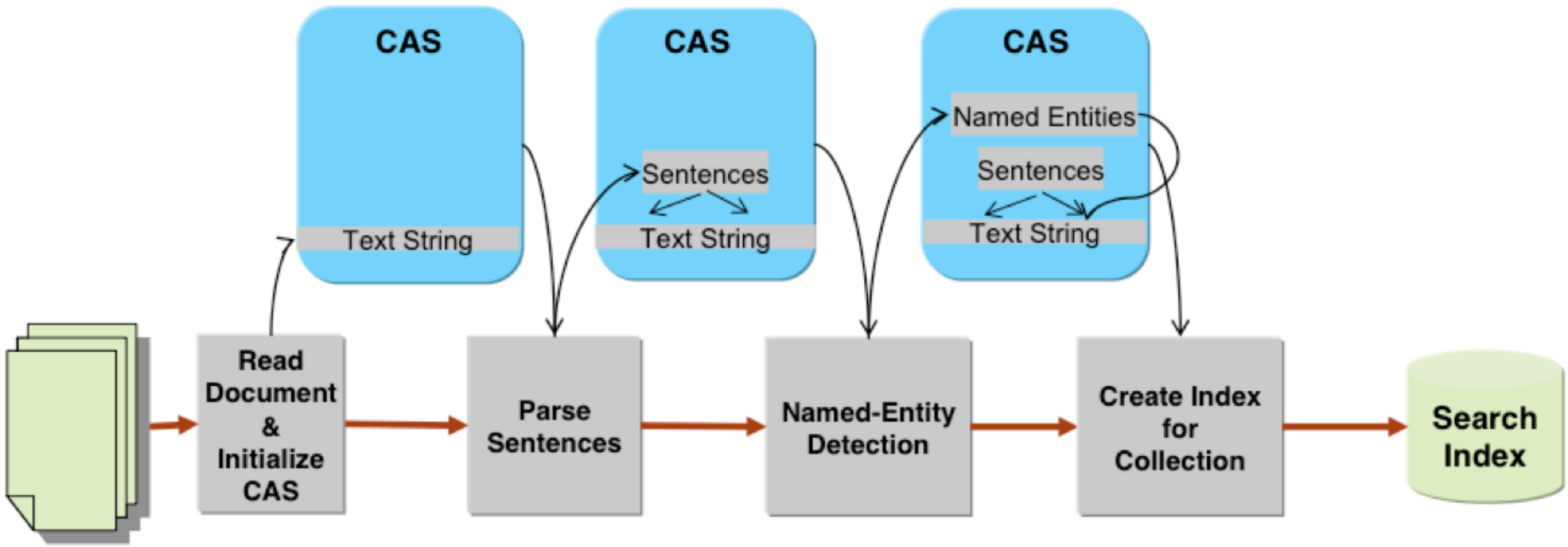


Overview of a very basic UIMA pipeline

Building a UIMA App

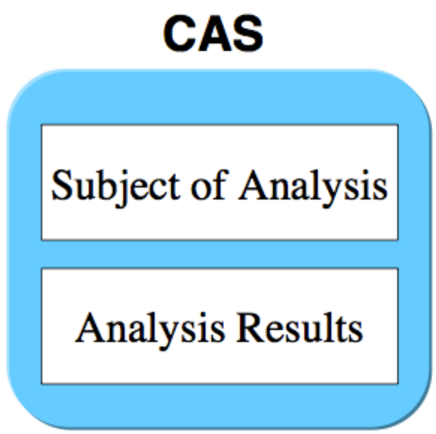


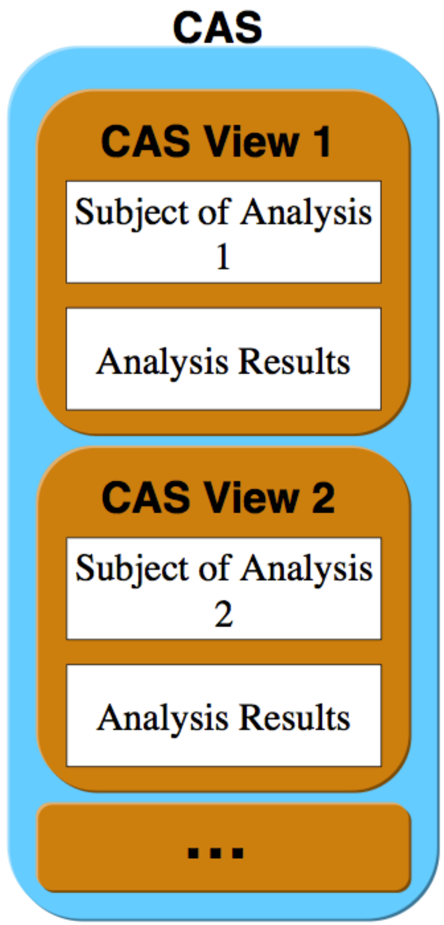
An Example



CAS
Text String
CAS
Sentences
Text String
CAS
Named Entities
Sentences
Text String
Read Document & Initialize CAS
Parse Sentences
Named-Entity Detection
Create Index for Collection
Search Index

CASes, Views and SofAs

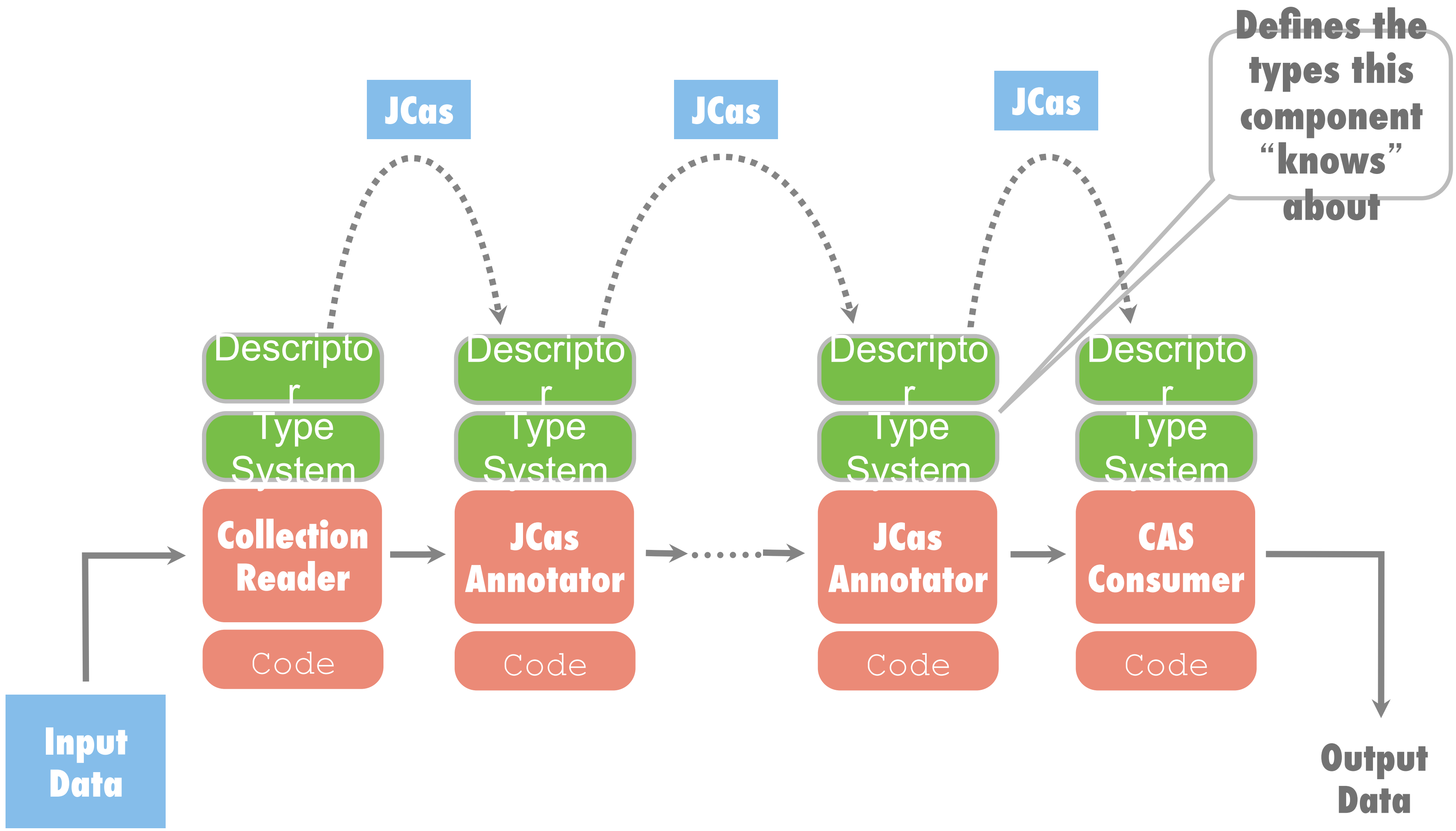




UIMA Java SDK

- UIMA SDK provides:
  - JCas object represents a CAS instance
  - interfaces for Collection Reader, Analysis Engine, CAS Consumer
- Application developer:
  - defines data types used in component
  - writes component code & descriptor

The UIMA Type System



Defining UIMA Types

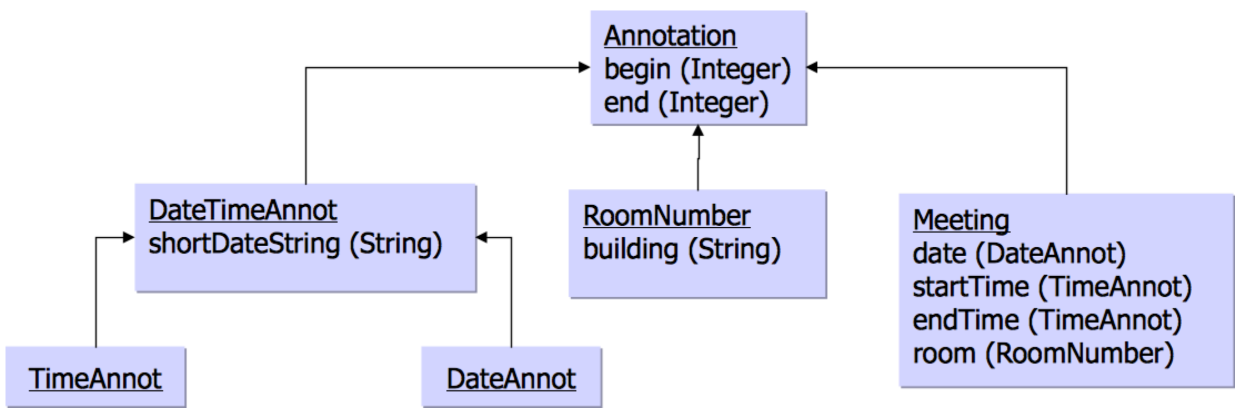
- Create XML description of types
- Generate Java classes for the types

```
:ypeDescription>
 <name>
   org.apache.uima.tutorial.RoomNumber
 </name>
 <supertypeName>
   uima.tcas.Annotation
 </supertypeName>
 <features>
   <featureDescription>
     <name>building</name>
     <description>
        Building containing this
        room
     </description>
     <rangeTypeName>
       uima.cas.String
     </rangeTypeName>
  </featureDescription>
 </features>
'typeDescription>
```

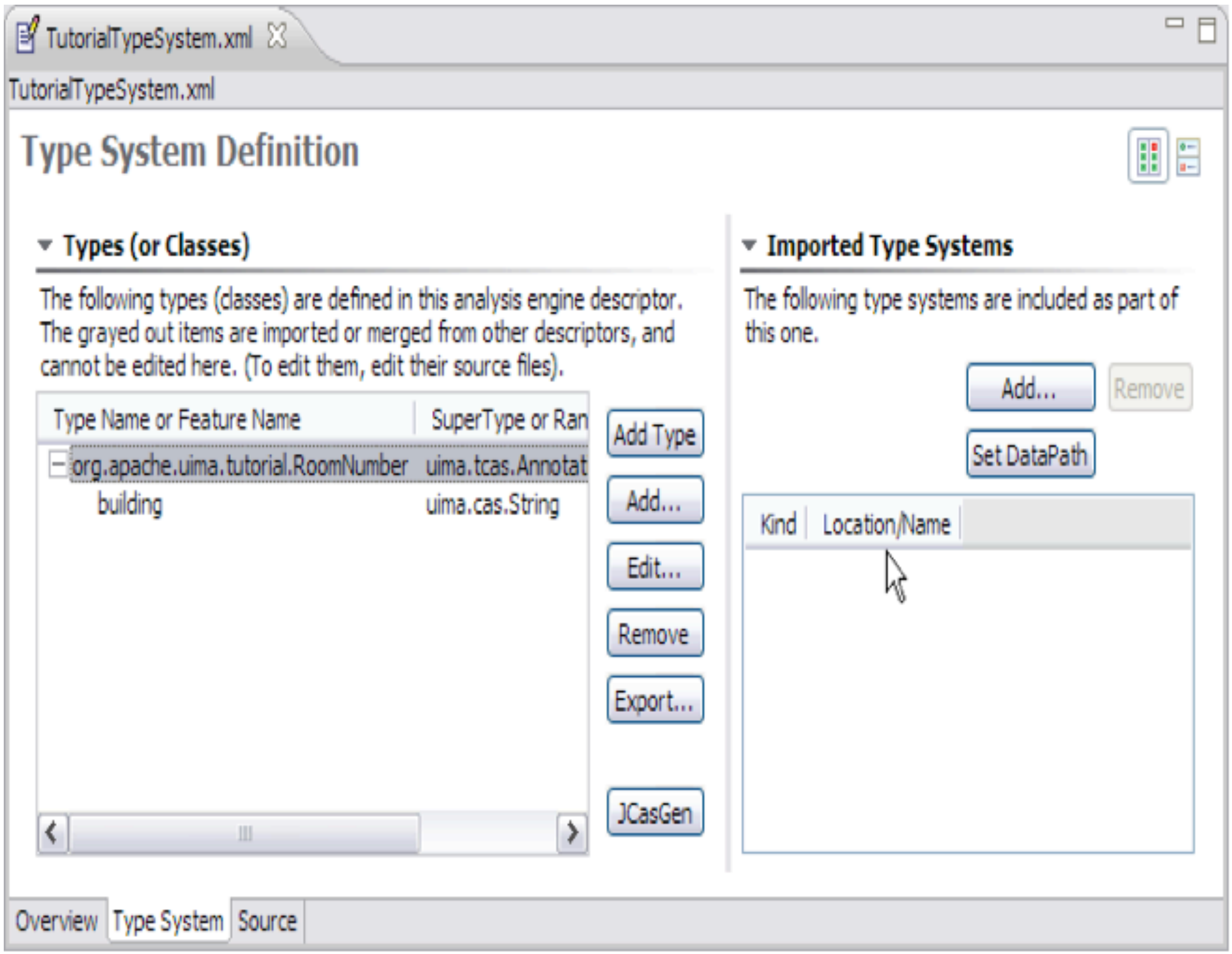
Defining UIMA Types

- UIMA types have inheritance just like java classes

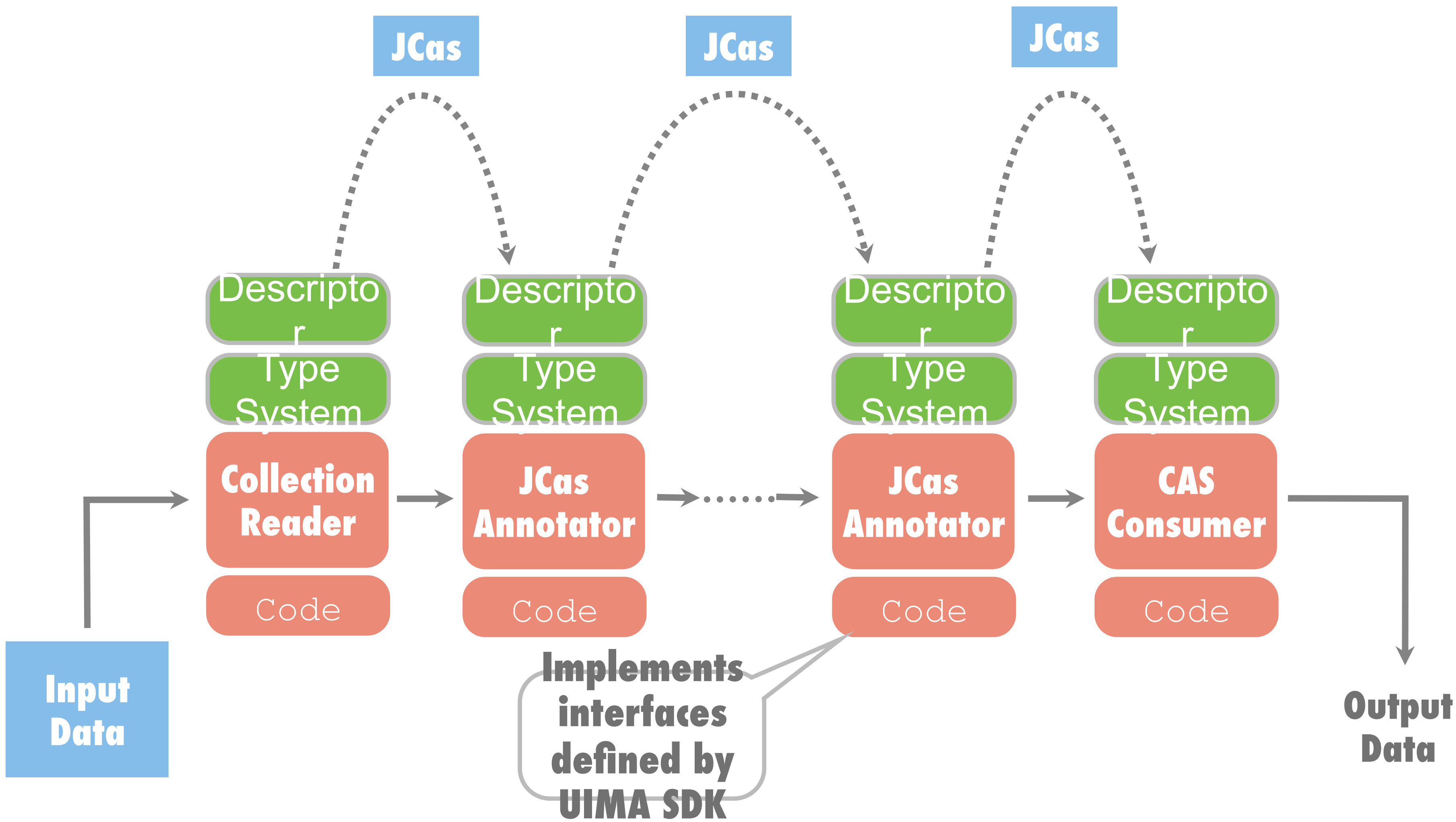
- UIMA types contain "features", similar to java class fields



Defining UIMA Types



Writing the Code

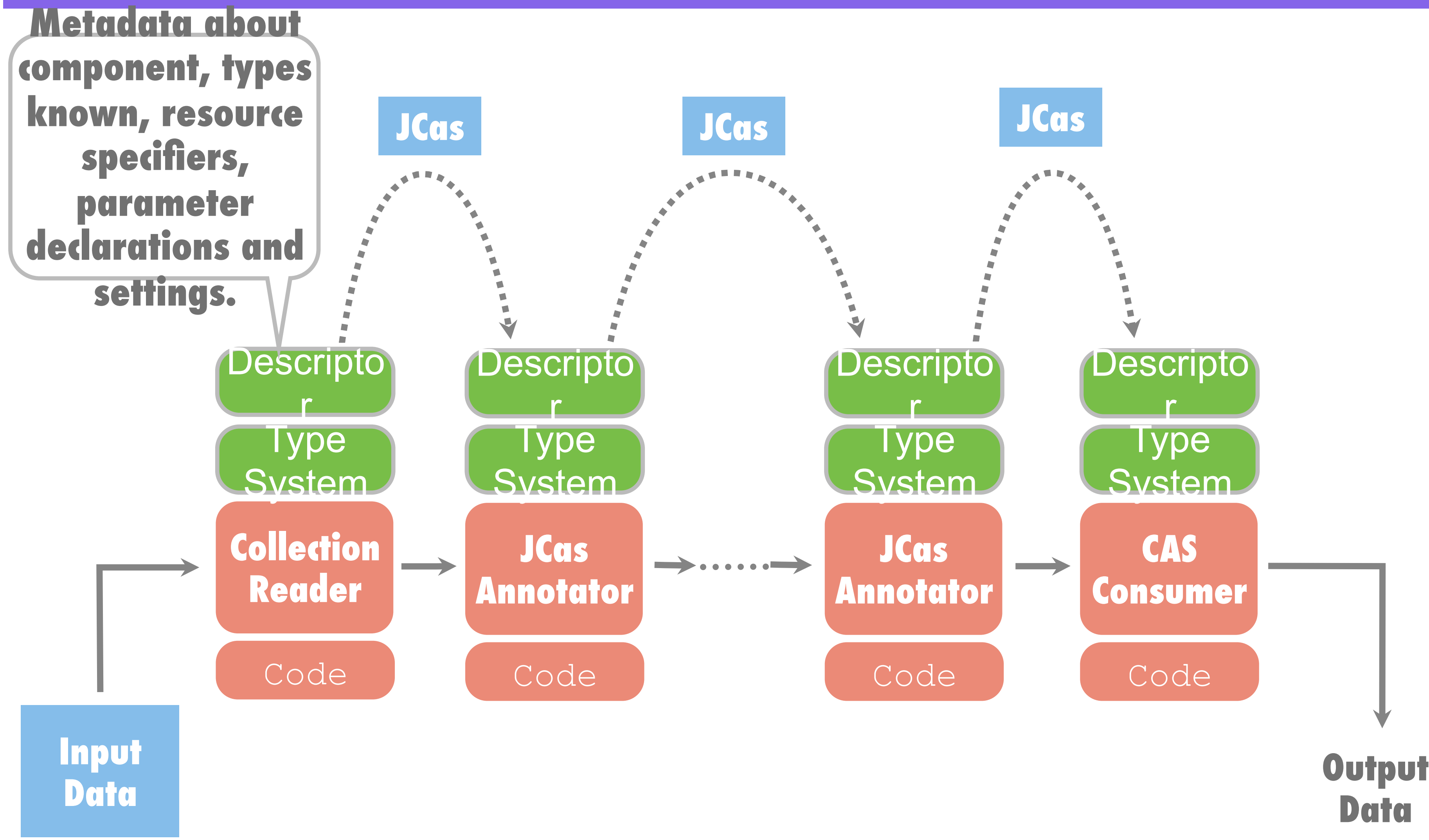


Writing the Code

Extend the Java abstract classes defined in the UIMA SDK:

```
public class Tokenizer extends JCasAnnotator_ImplBase
{

   @Override
   public void initialize(UimaContext aContext) throws
ResourceInitializationException
   { ... }

   @Override
   public void process(JCas cas) throws
AnalysisEngineProcessException
   { ... }

   @Override
   public void collectionProcessComplete() throws
AnalysisEngineProcessException
   { ... }
```

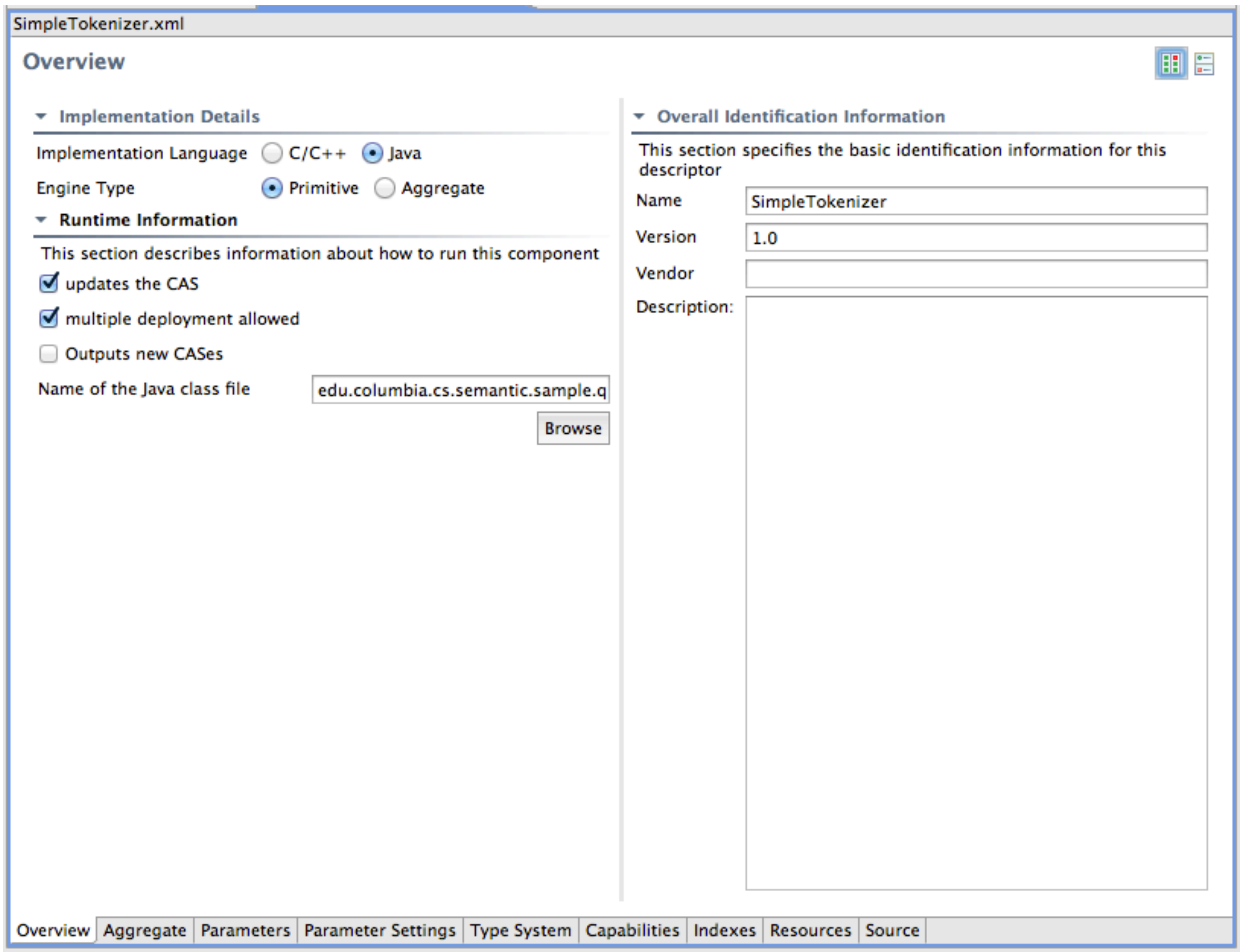
Component Descriptor



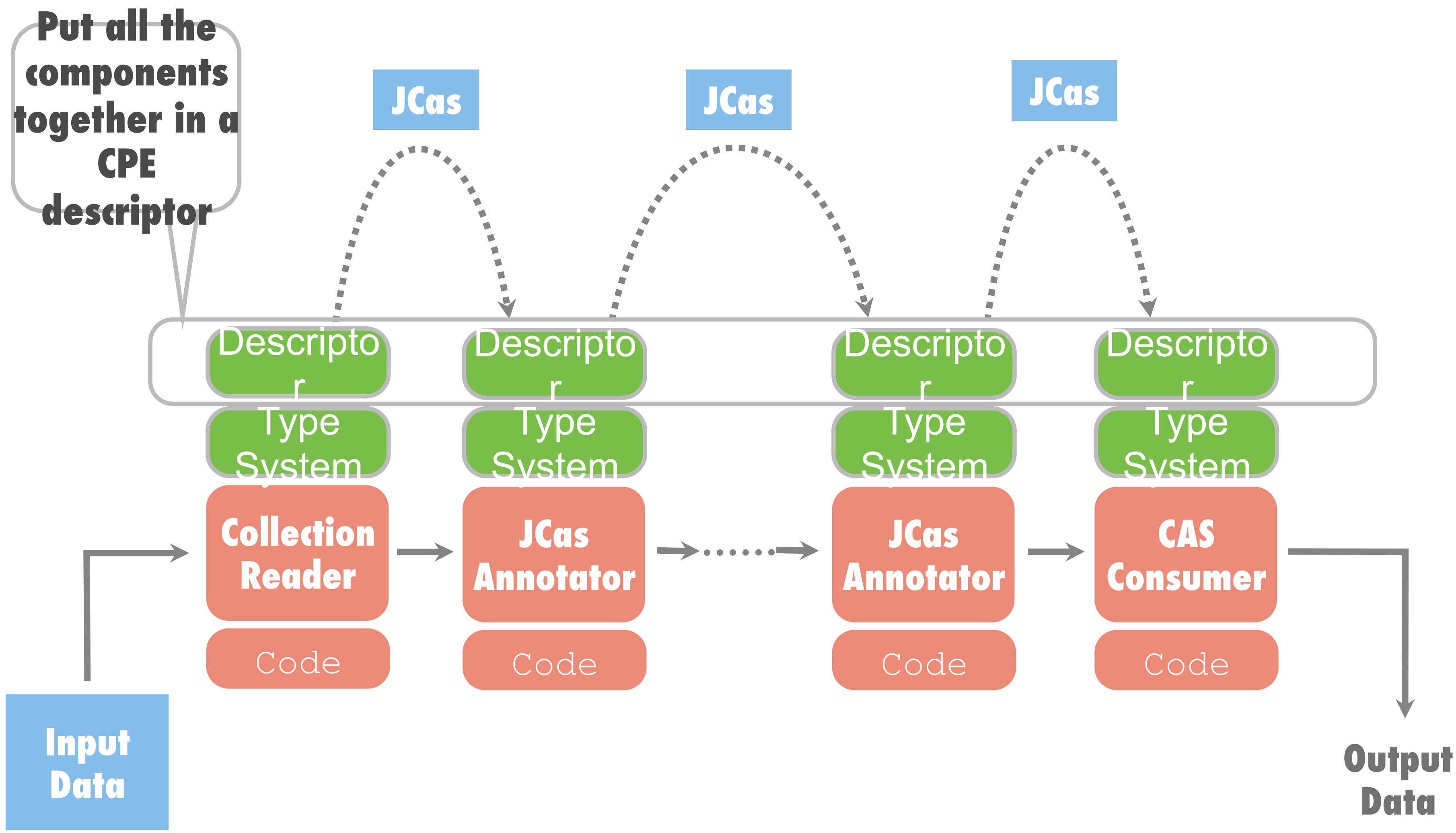
Component Descriptor

- Contains metadata information about component: name, description, pointer to java class, type system
- “Bridges” UIMA framework and the component code
- Contains parameter settings and resource specifiers

Component Descriptor



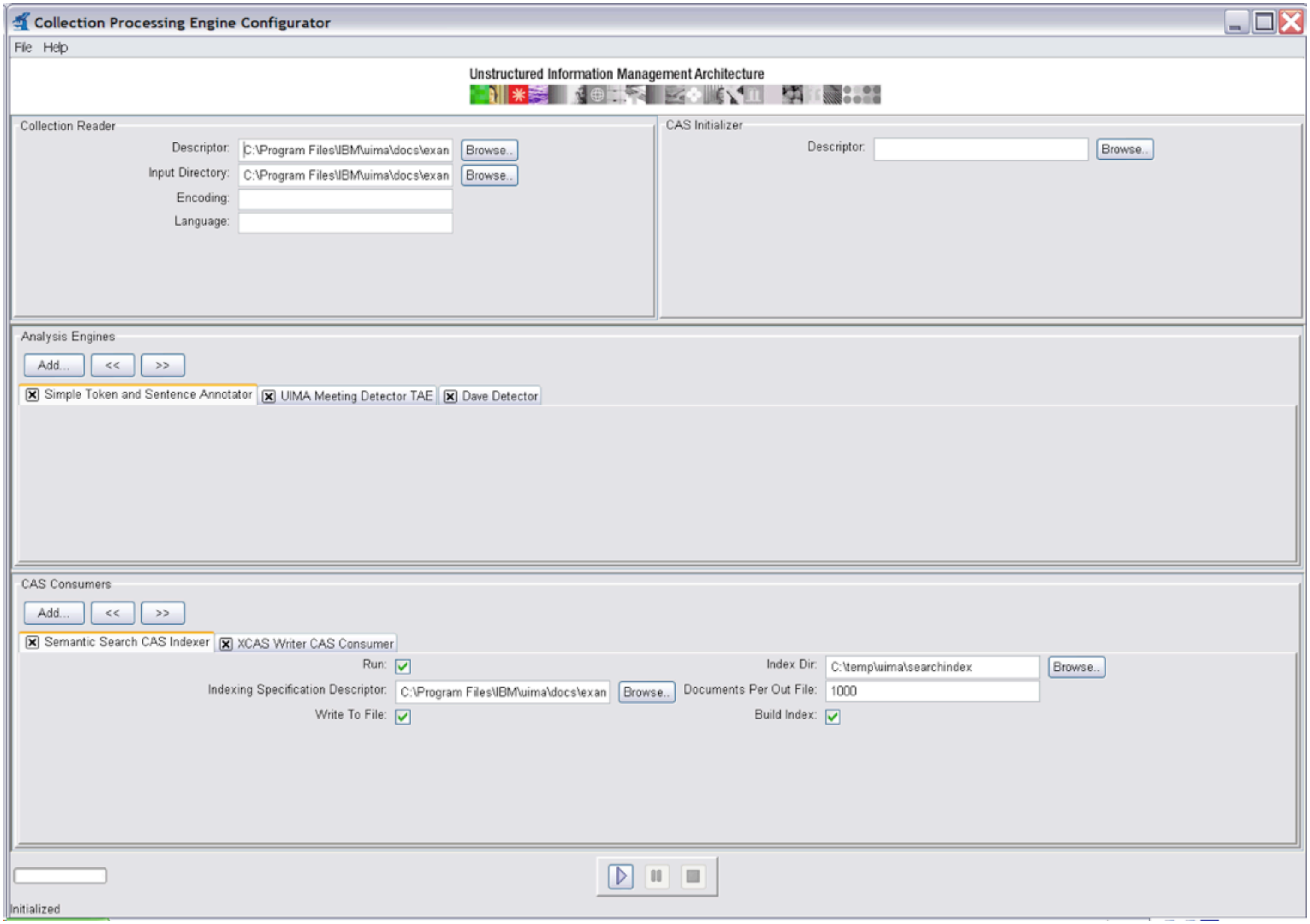
Putting it all Together



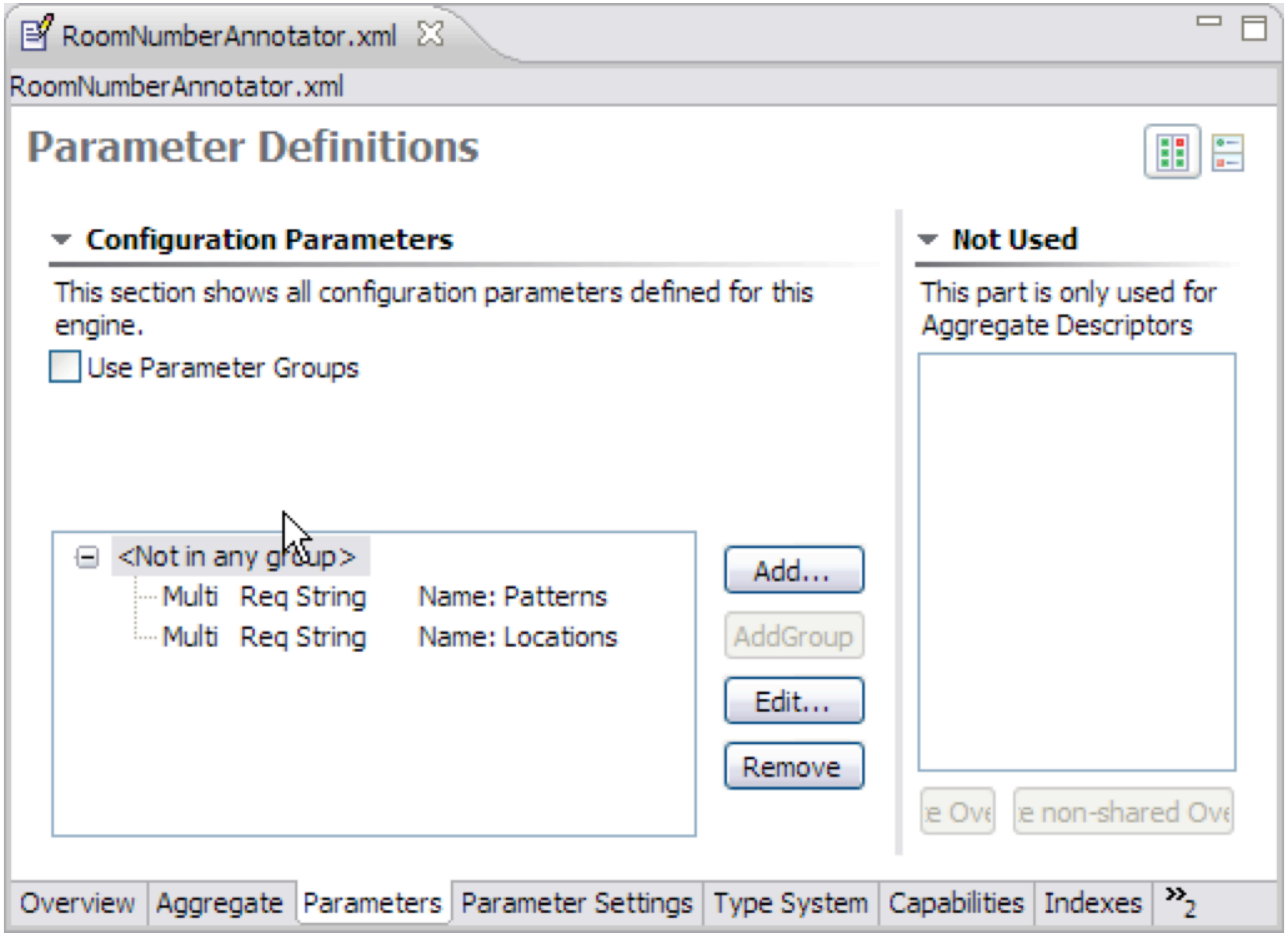
Putting it all Together

- CPE descriptor (XML) describes a sequence of UIMA components:
  - Collection Reader
  - Analysis Engines
  - CAS Consumers
- Invoke the UIMA framework with the runCPE script or the CPE GUI

Putting it all Together



“Parameterizing” AEs



“Parameterizing” AEs

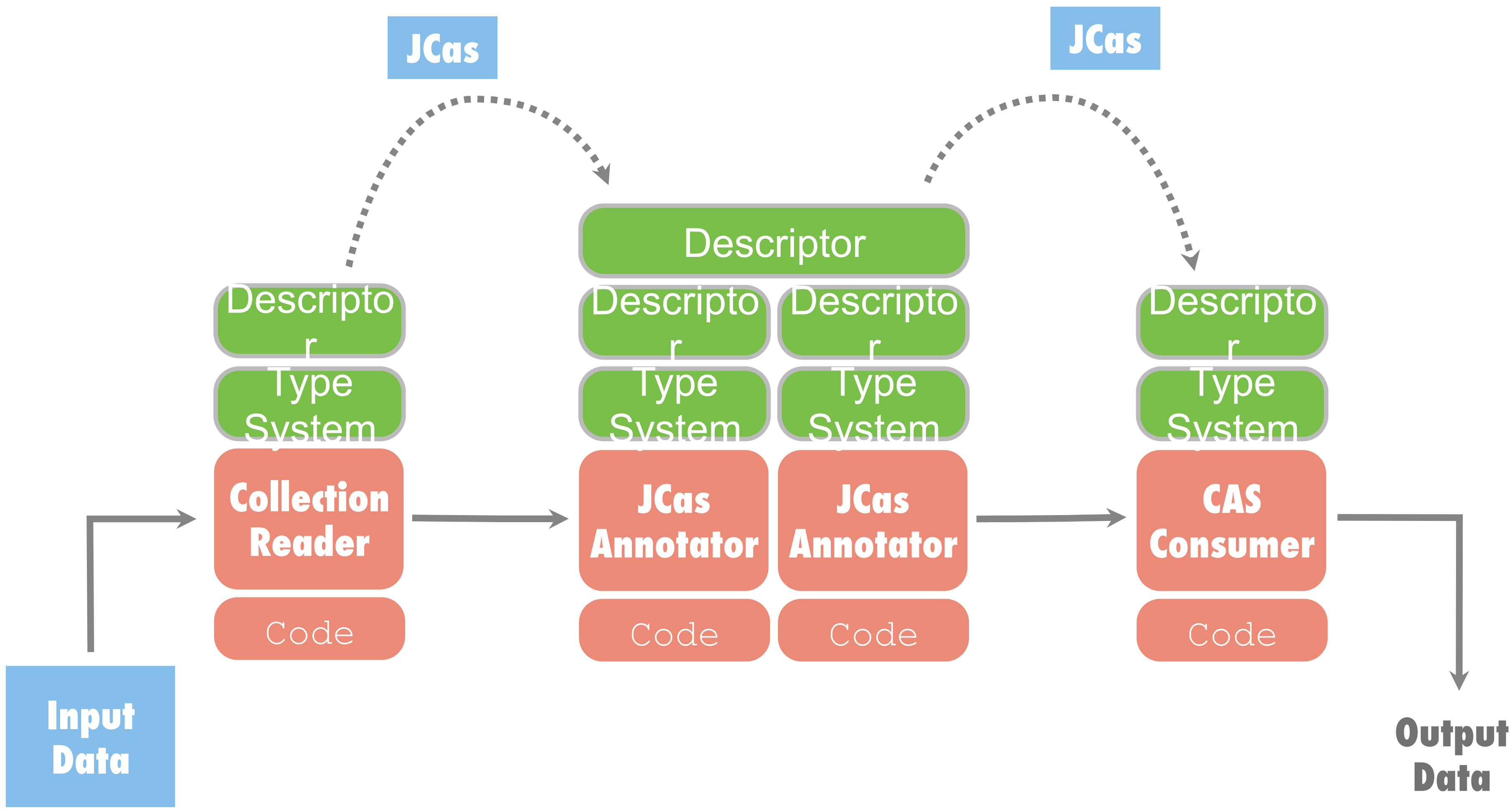
Parameters get passed to Java code through UimaContext in initialize()

```
public class Tokenizer extends JCasAnnotator_ImplBase
{

   @Override
   public void initialize(UimaContext aContext) throws
ResourceInitializationException
   {
      String[] paramSetting = (String[])
aContext.getConfigParameterValue("Patterns");
   }

   ...

}
```

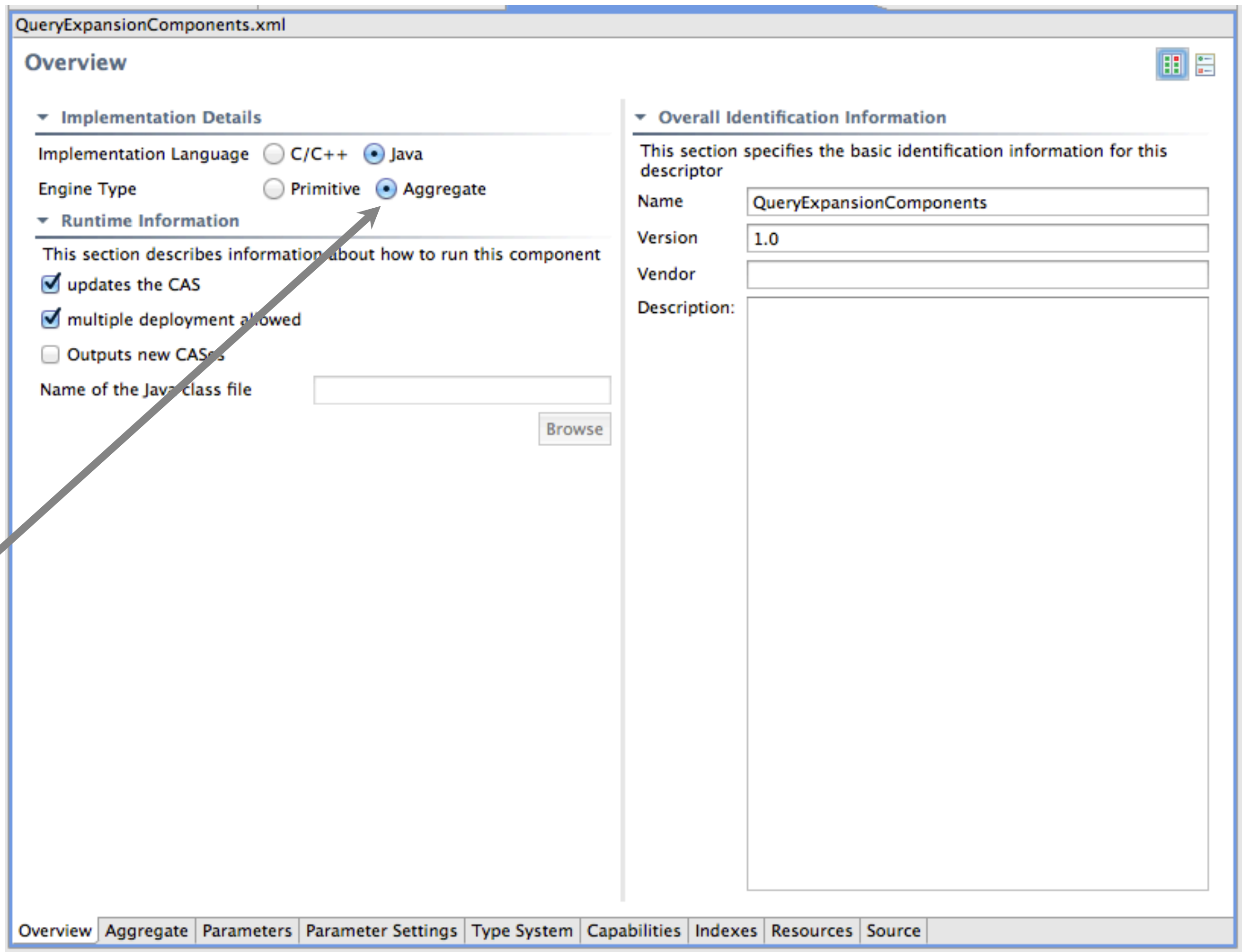
“Aggregating” AEs

- Multiple Analysis Engines can be combined into single one by:
  - creating new “aggregate” Analysis Engine descriptor
  - adding delegate Analysis Engines to the descriptor

“Aggregating” AEs

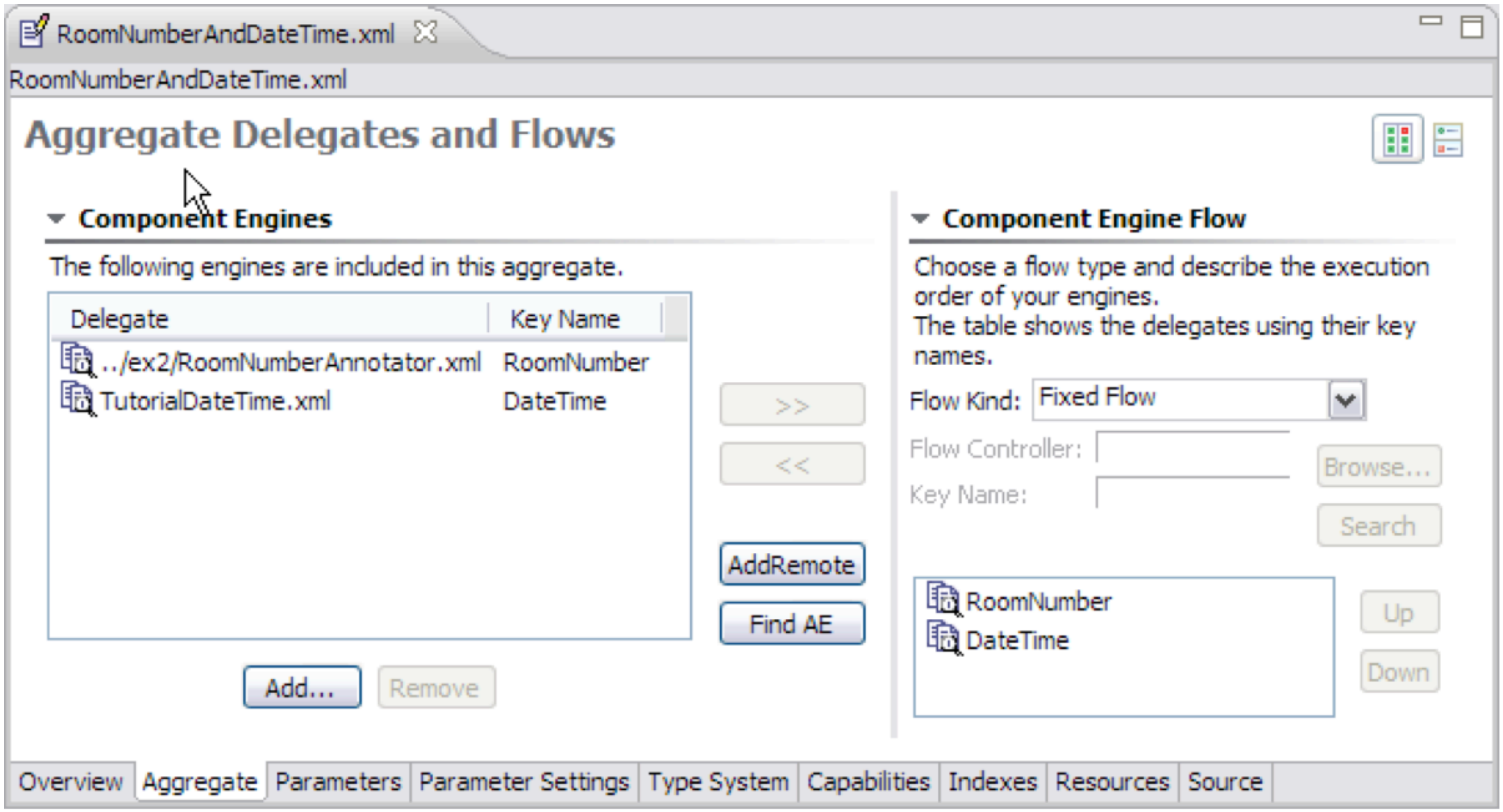


“Aggregating” AEs



“aggregate” flag is set

"Aggregating" AEs

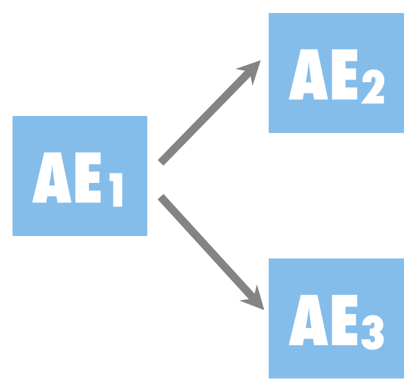


Advanced Topics

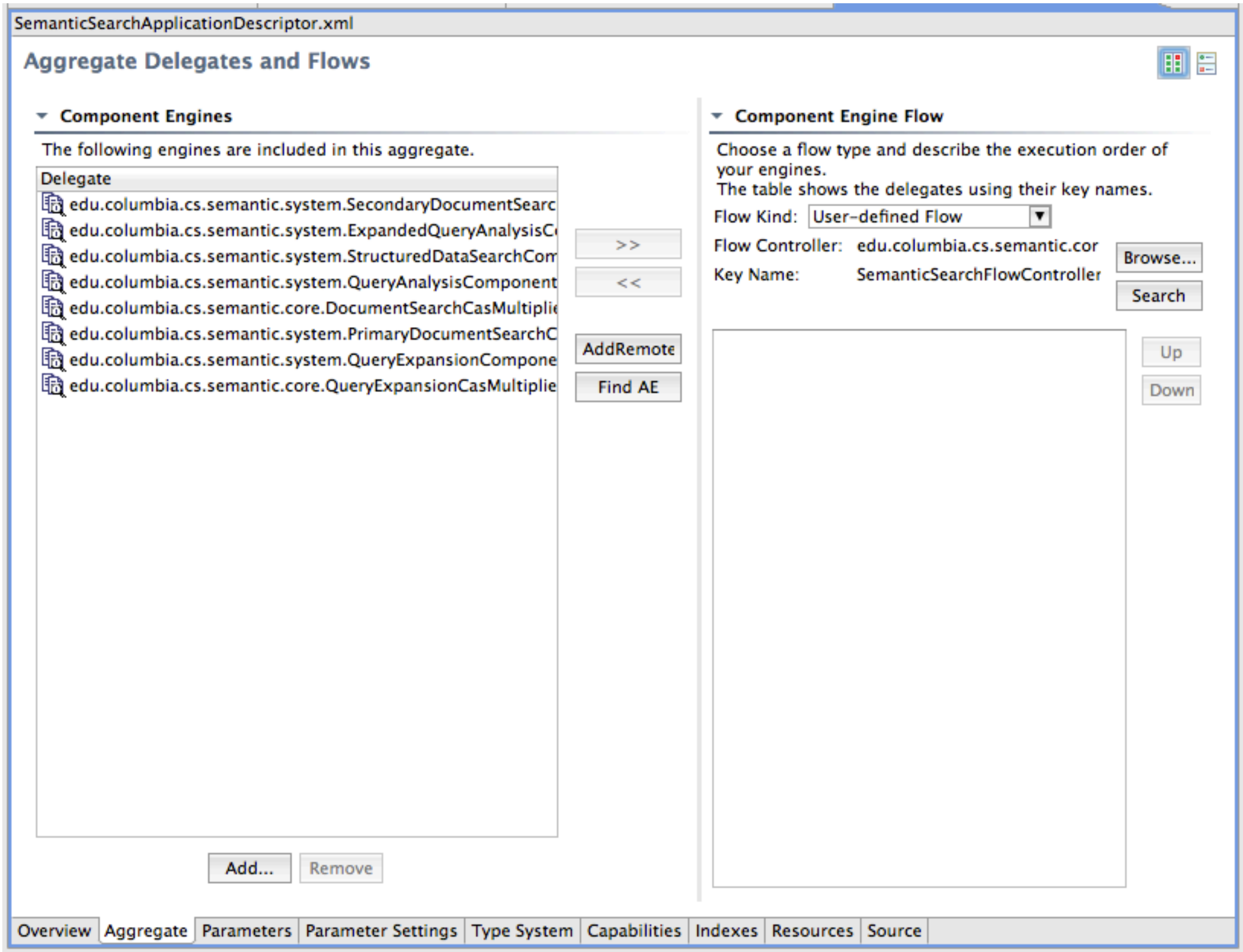
- Flow controllers
- CAS multipliers
- UIMA-AS

Flow Controllers

- "Fixed Flow" is the default for an aggregate AE
- Can be replaced with a user-defined flow, using a Flow Controller

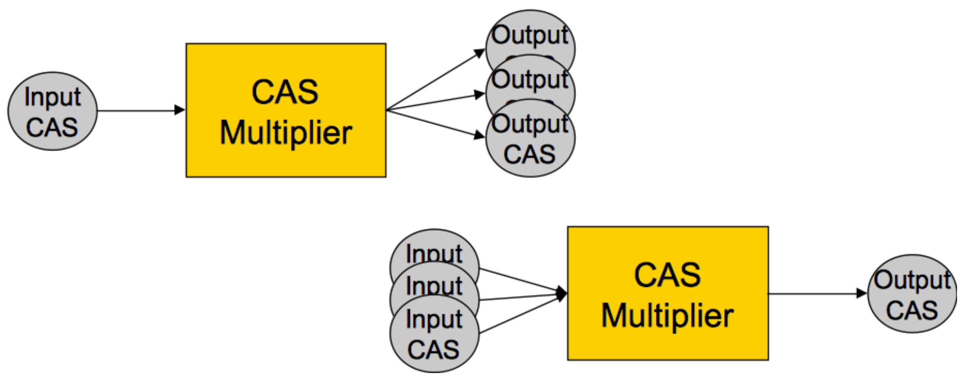


Flow Controllers

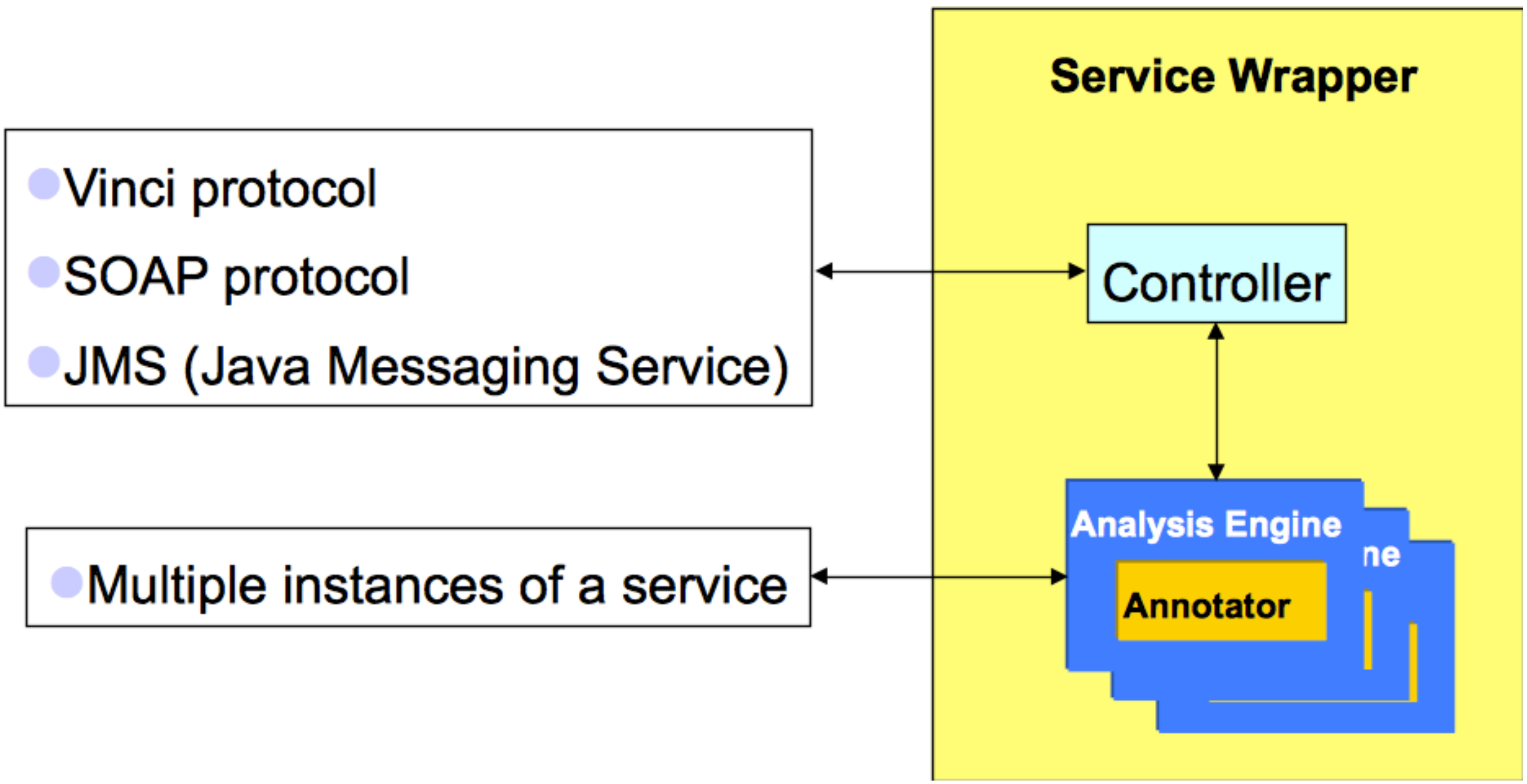


CAS Multipliers

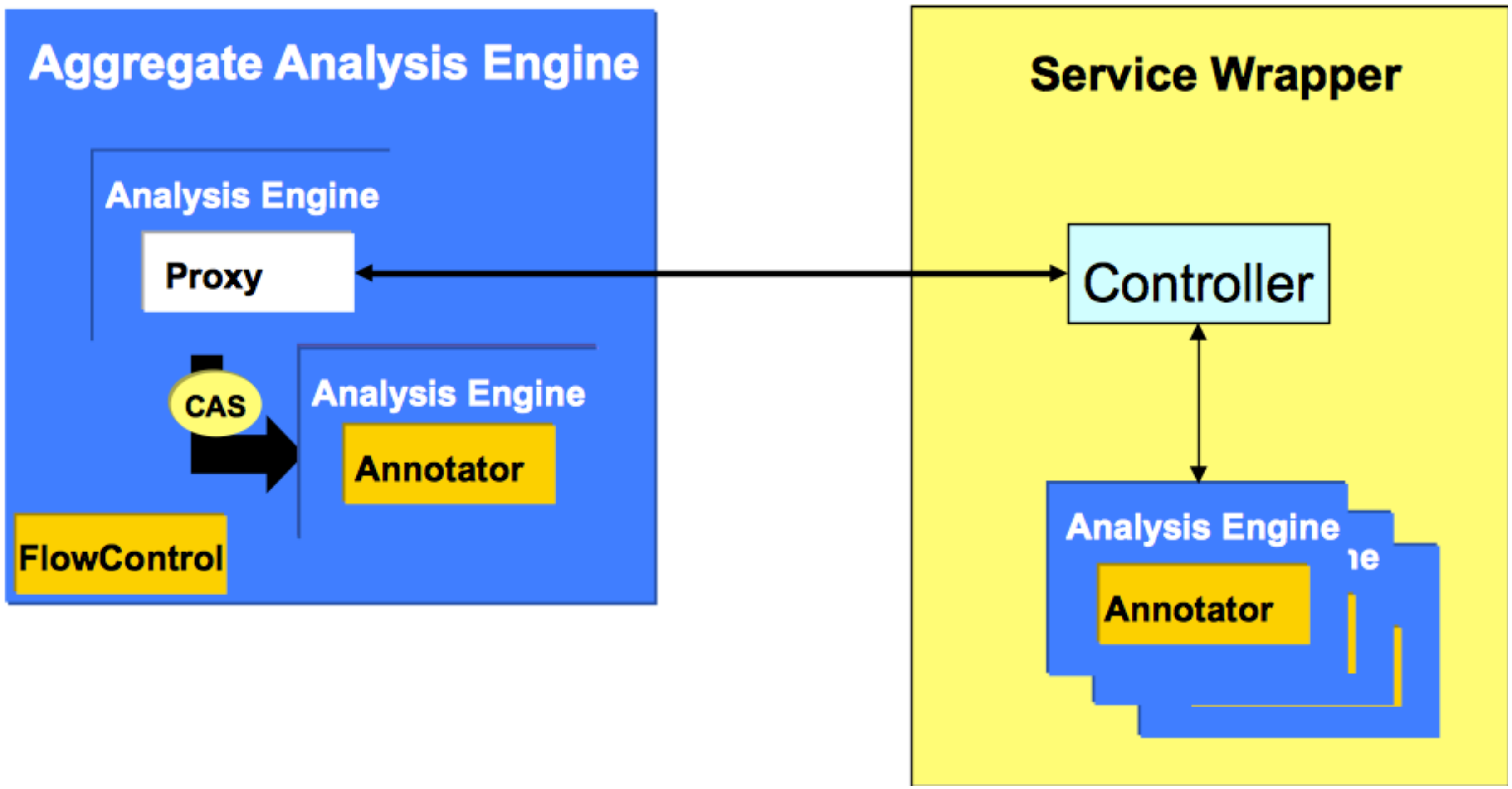
- Component that can produce new CASes
- Takes a "trigger" CAS as input
- Produces arbitrary number of CASes
- Anywhere in flow



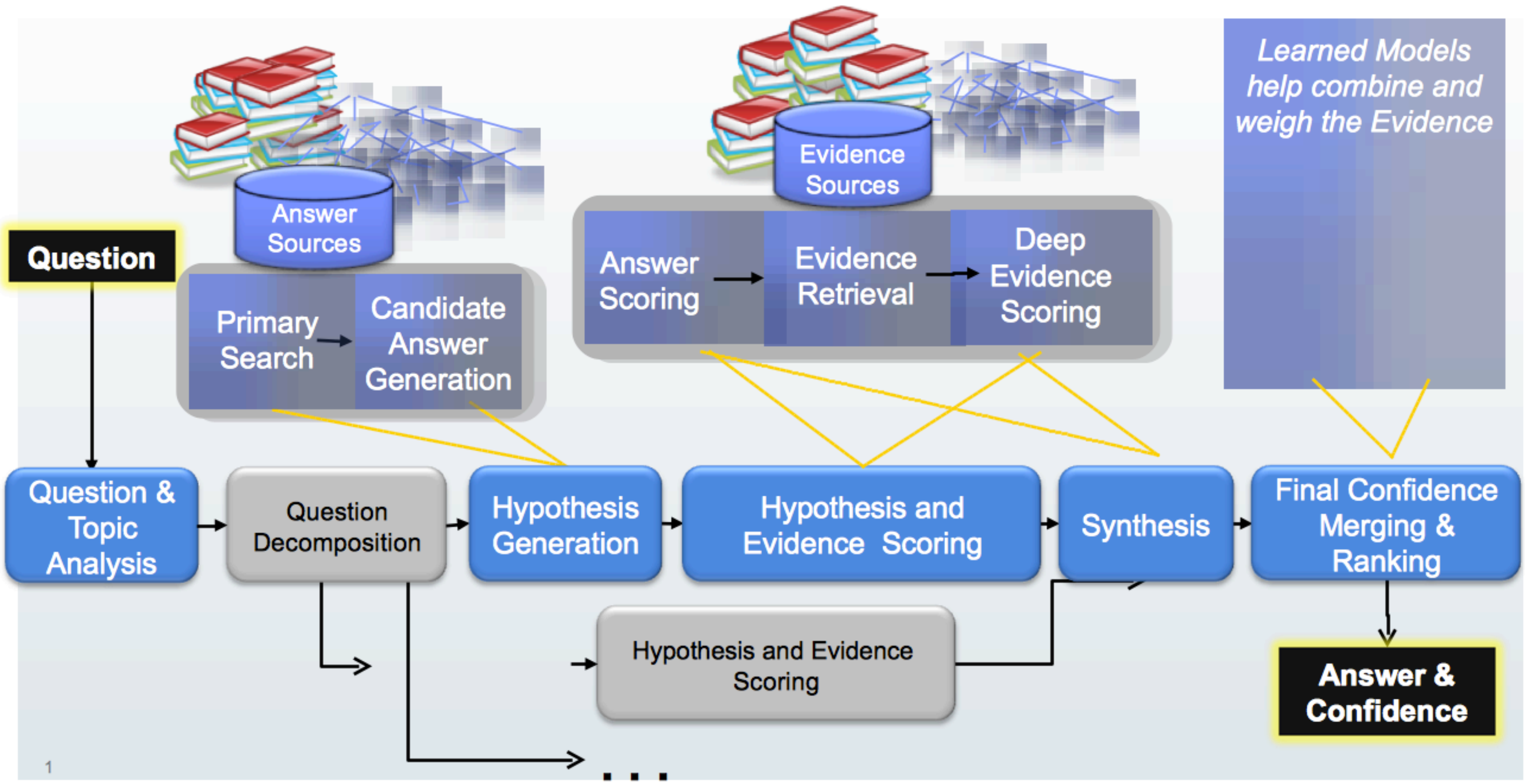
UIMA-AS


Vinci protocol
SOAP protocol
JMS (Java Messaging Service)
Multiple instances of a service
Service Wrapper
Controller
Analysis Engine
Annotator

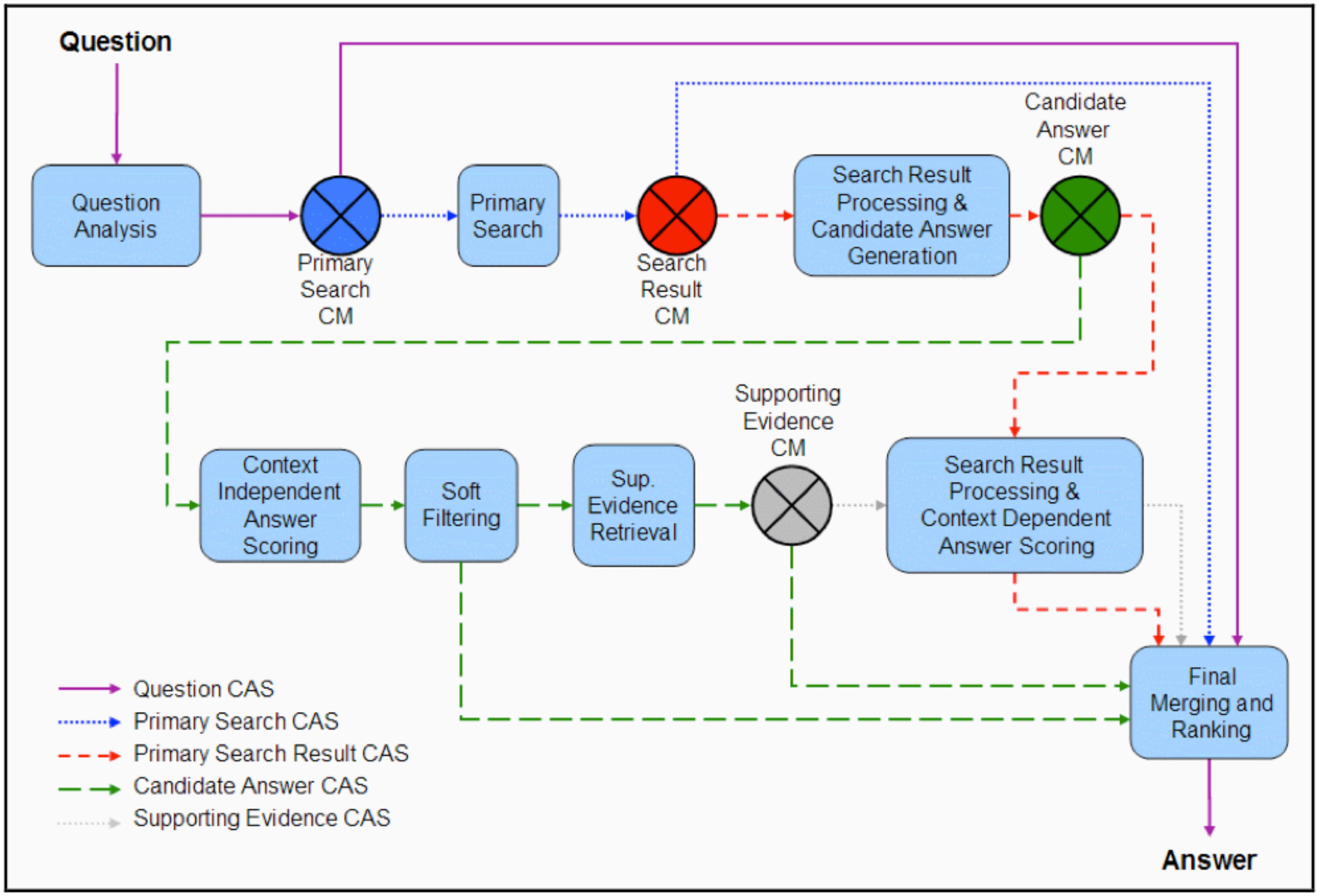
UIMA-AS


Aggregate Analysis Engine
Analysis Engine
Proxy
CAS
Analysis Engine
Annotator
FlowControl
Service Wrapper
Controller
Analysis Engine
ne
Annotator

UIMA in Watson



UIMA in Watson



Class Project Setup

- Install Eclipse
- Download UIMA SDK
- Get the framework code for class project: Or will send details

References

- http://uima.apache.org/documentation.html
  - Tutorial and Users' Guides
  - Overview and Setup
- E. Epstein, M. Schor, B. Iyer, A. Lally, E. Brown, J. Cwiklik. Making Watson Fast. IBM Journal of Research and Development. Volume 56 (3-4). 2012.

Semantic Technologies in IBM Watson^{TM}

Lesson 8 – UIMA hands on

Guest Lecturer: Siddharth Patwardhan
TA: Or Biran

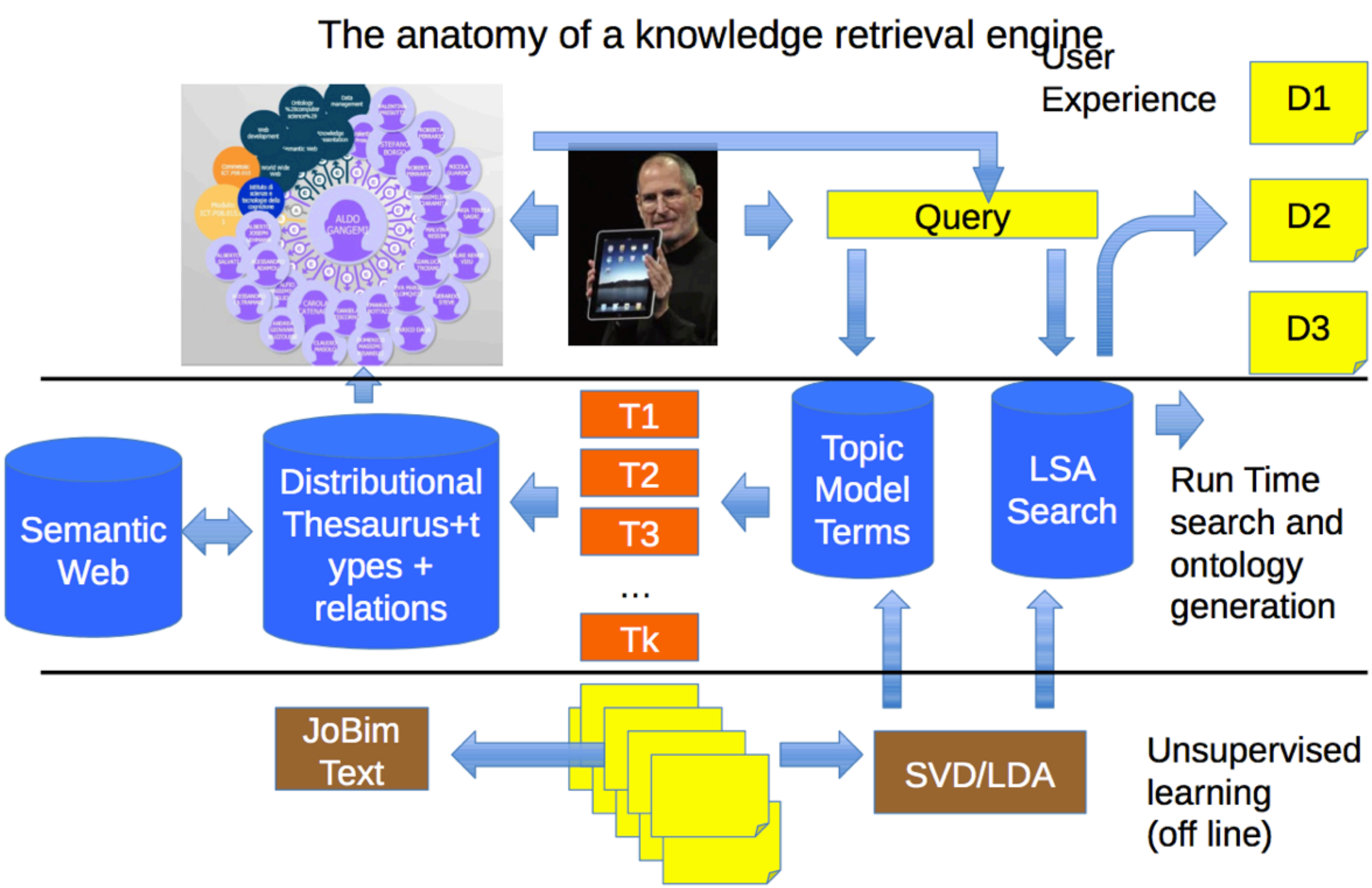
Objectives

- At the completion of this class, you should be able to:
  - Explain the architecture of the class project pipeline
  - Download and install the source code
  - Develop and add a UIMA component to the pipeline

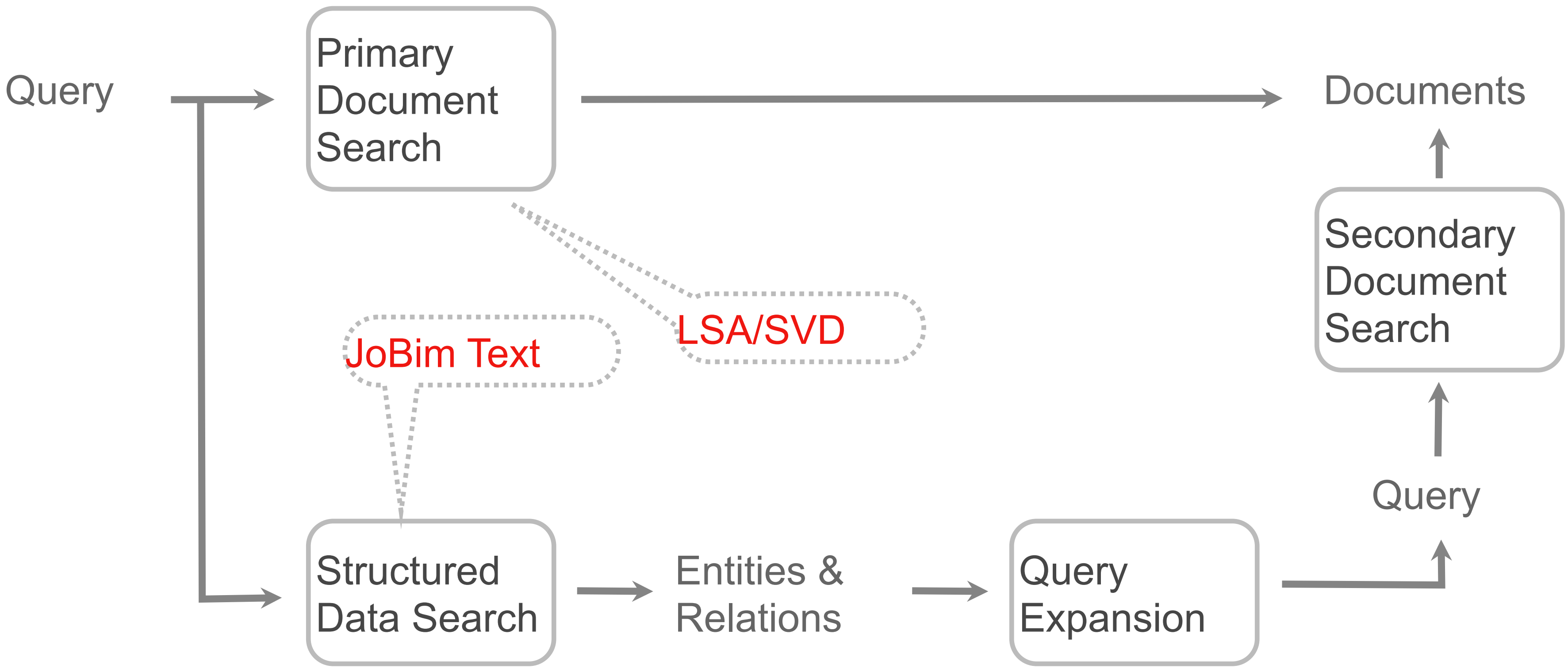
Outline

- Overview of project framework
- Setup Eclipse and UIMA
- Download, compile and run code
- Code “orientation”
- Creating a new component
- Exploring the output

Framework Overview



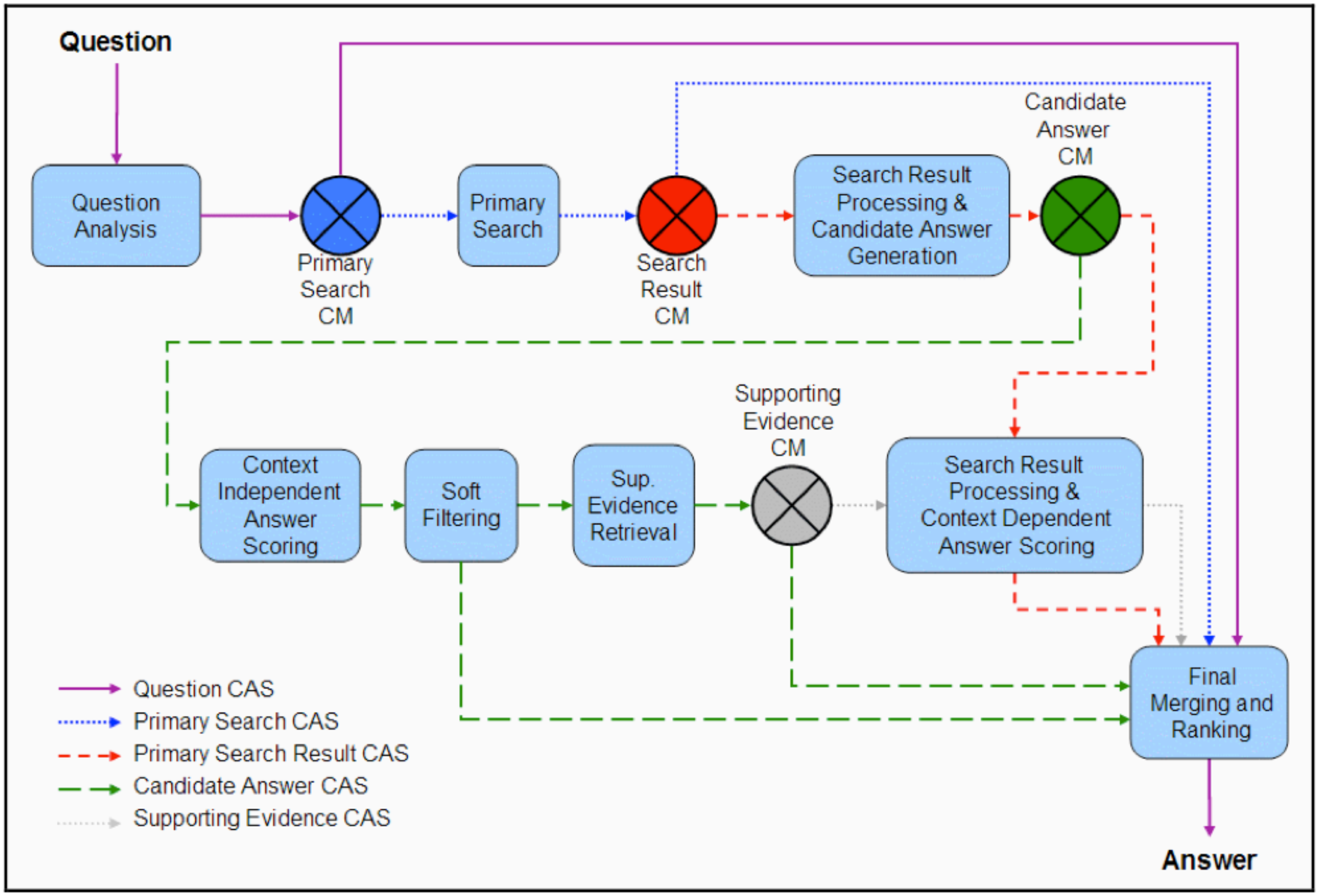
Framework Overview



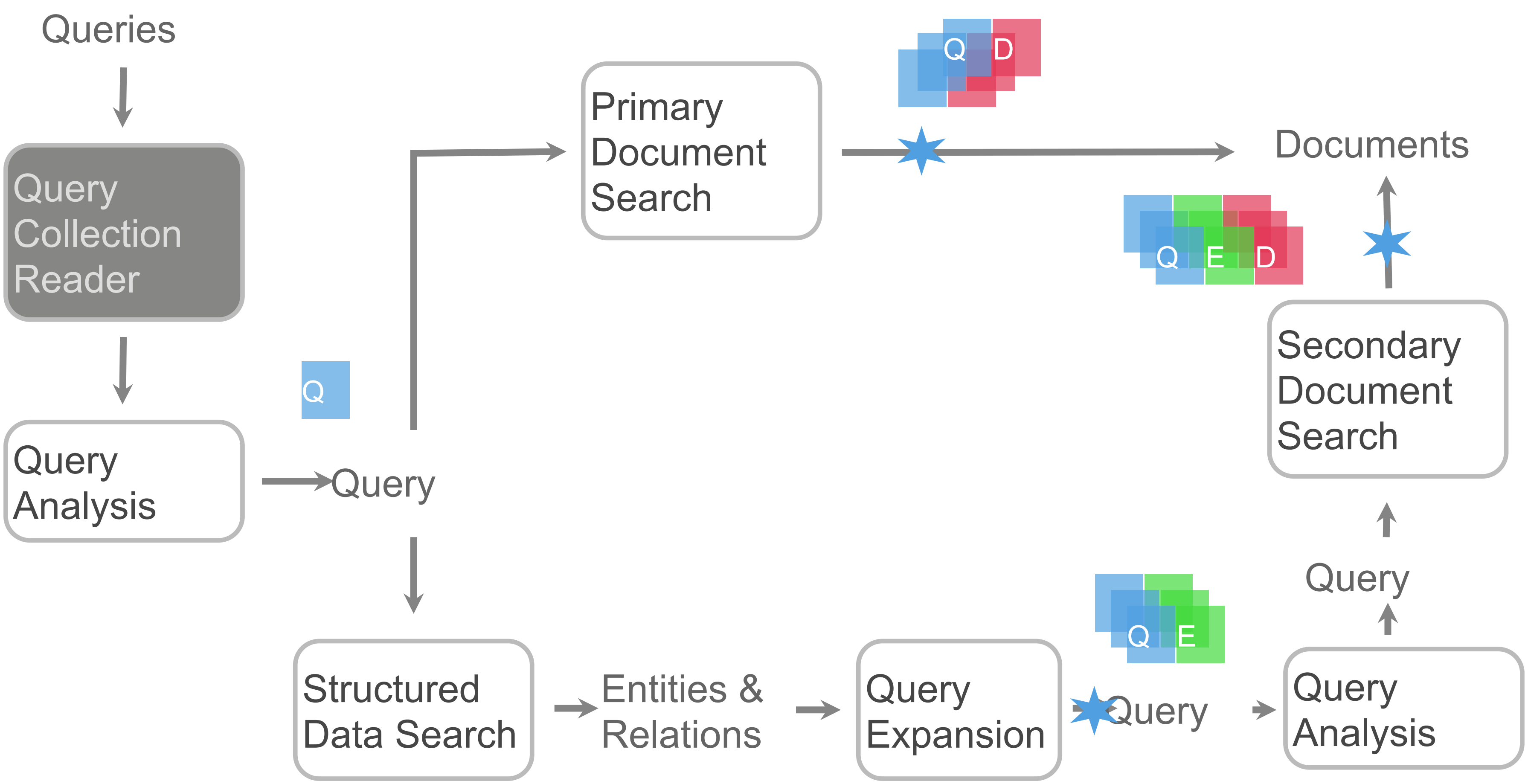
Architectural Design

- Follow approach similar to Watson
  - Segment the pipeline into high-level modules
  - Multiple CASes flow in parallel through pipeline
  - Components independently analyze CASes at various stages of pipeline

UIMA in Watson



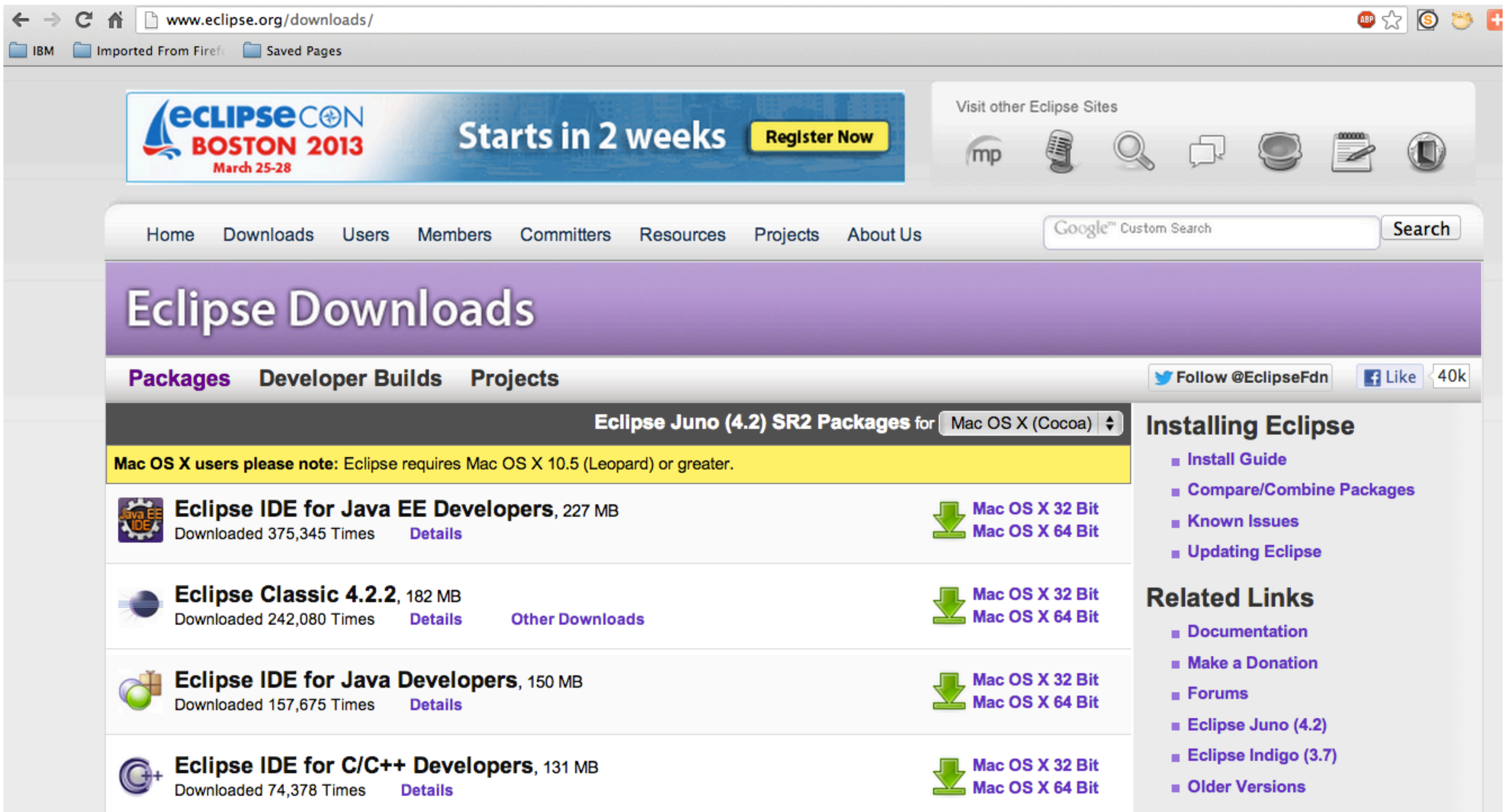
UIMA in Class Project



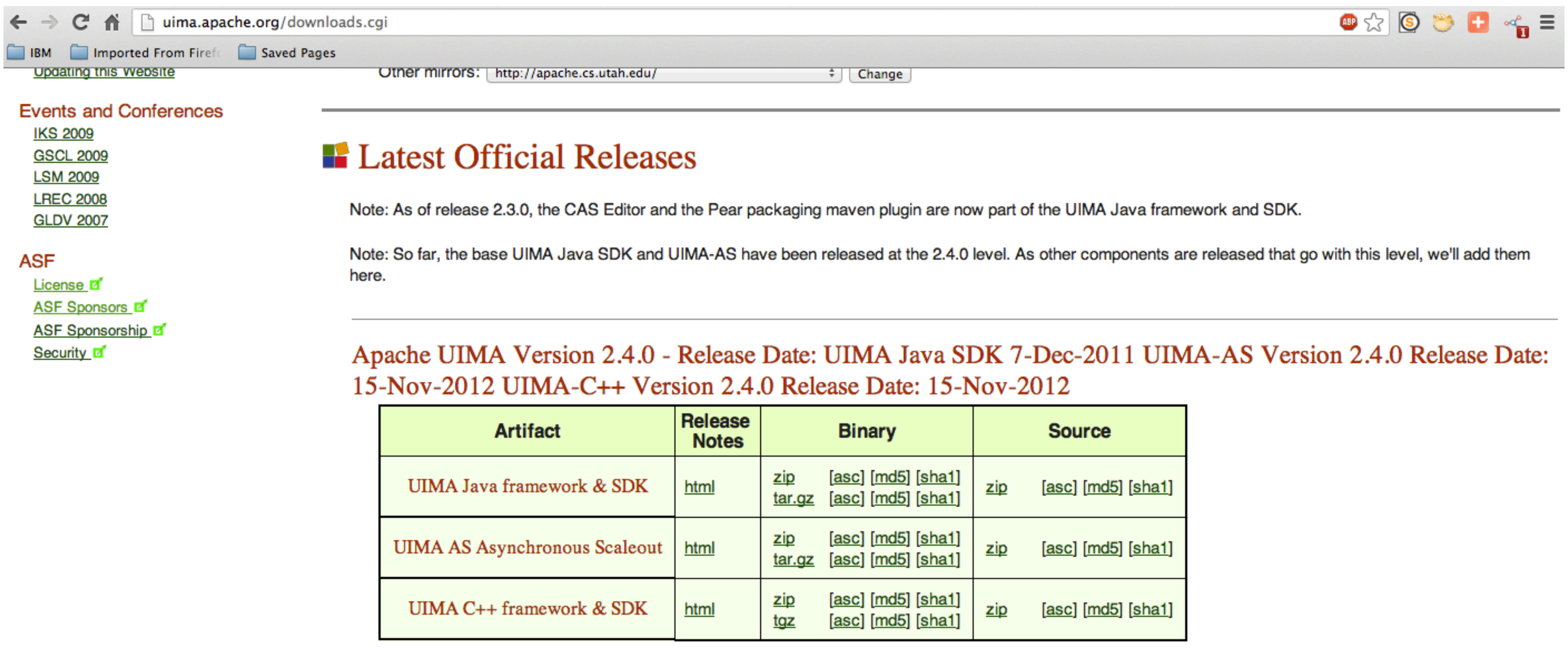
Setup Eclipse and UIMA

- Get Eclipse
- Download UIMA
- Install UIMA plugins

Get Eclipse


www.eclipse.org/downloads/
IBM
Saved Pages
ECLIPSECON BOSTON 2013
March 25-28
Starts in 2 weeks
Register Now
Visit other Eclipse Sites
Home
Downloads
Users
Members
Committers
Resources
Projects
About Us
Search
Eclipse Downloads
Packages
Developer Builds
Projects
Follow @EclipseFdn
Like
40k
Eclipse Juno (4.2) SR2 Packages for Mac OS X (Cocoa)
Mac OS X users please note: Eclipse requires Mac OS X 10.5 (Leopard) or greater.
Eclipse IDE for Java EE Developers, 227 MB
Downloaded 375,345 Times
Details
Mac OS X 32 Bit
Mac OS X 64 Bit
Eclipse Classic 4.2.2, 182 MB
Downloaded 242,080 Times
Details
Other Downloads
Mac OS X 32 Bit
Mac OS X 64 Bit
Eclipse IDE for Java Developers, 150 MB
Downloaded 157,675 Times
Details
Mac OS X 32 Bit
Mac OS X 64 Bit
Eclipse IDE for C/C++ Developers, 131 MB
Downloaded 74,378 Times
Details
Mac OS X 32 Bit
Mac OS X 64 Bit
Installing Eclipse
Install Guide
Compare/Combine Packages
Known Issues
Updating Eclipse
Related Links
Documentation
Make a Donation
Forums
Eclipse Juno (4.2)
Eclipse Indigo (3.7)
Older Versions

Get UIMA



uima.apache.org/downloads.cgi

IBM | Imported From Firef | Saved Pages

Updating this website

Other mirrors: http://apache.cs.utah.edu/ Change

Events and Conferences

- IKS 2009
- GSCL 2009
- LSM 2009
- LREC 2008
- GLDV 2007

ASF

- License
- ASF Sponsors
- ASF Sponsorship
- Security

Latest Official Releases

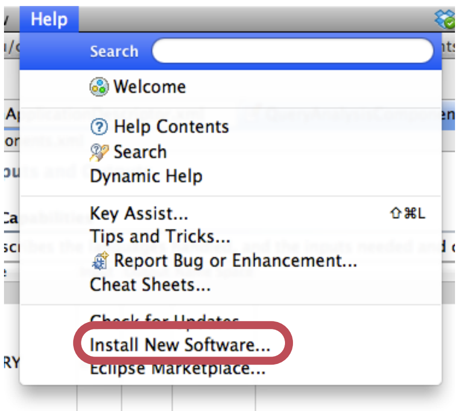
Note: As of release 2.3.0, the CAS Editor and the Pear packaging maven plugin are now part of the UIMA Java framework and SDK.

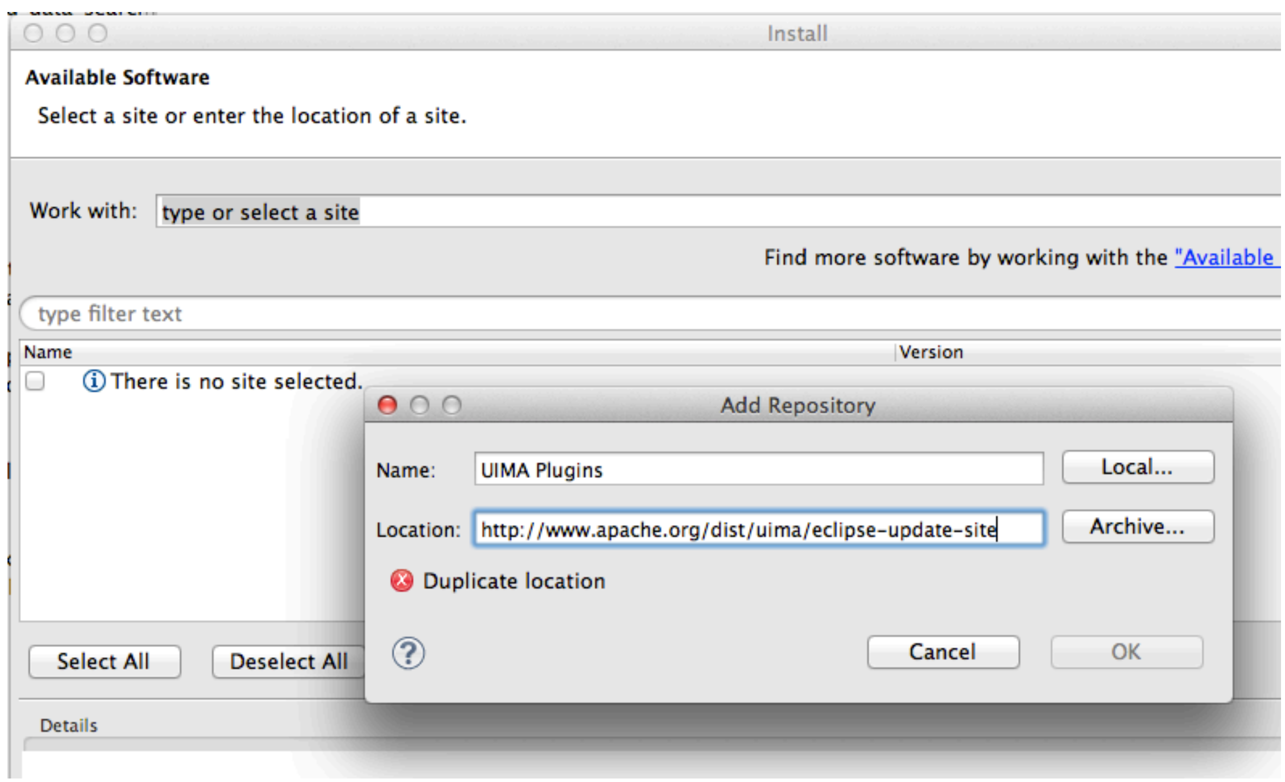
Note: So far, the base UIMA Java SDK and UIMA-AS have been released at the 2.4.0 level. As other components are released that go with this level, we'll add them here.

Apache UIMA Version 2.4.0 - Release Date: UIMA Java SDK 7-Dec-2011 UIMA-AS Version 2.4.0 Release Date: 15-Nov-2012 UIMA-C++ Version 2.4.0 Release Date: 15-Nov-2012

Artifact	Release Notes	Binary	Source
UIMA Java framework & SDK	html	zip [asc] [md5] [sha1] tar.gz [asc] [md5] [sha1]	zip [asc] [md5] [sha1]
UIMA AS Asynchronous Scaleout	html	zip [asc] [md5] [sha1] tar.gz [asc] [md5] [sha1]	zip [asc] [md5] [sha1]
UIMA C++ framework & SDK	html	zip [asc] [md5] [sha1] tgz [asc] [md5] [sha1]	zip [asc] [md5] [sha1]

UIMA Eclipse Plugins





Source Code Setup

- Download
- Compile
- Run
- View output

Get the Source Code

- http://www.cs.columbia.edu/~orb/class_project.zip

Setup the Workspace

- Create new workspace
- Import code into workspace
- Create run configuration: Driver
- Run with input data
- Create run configuration: Annotation Viewer
- Inspect output CASes

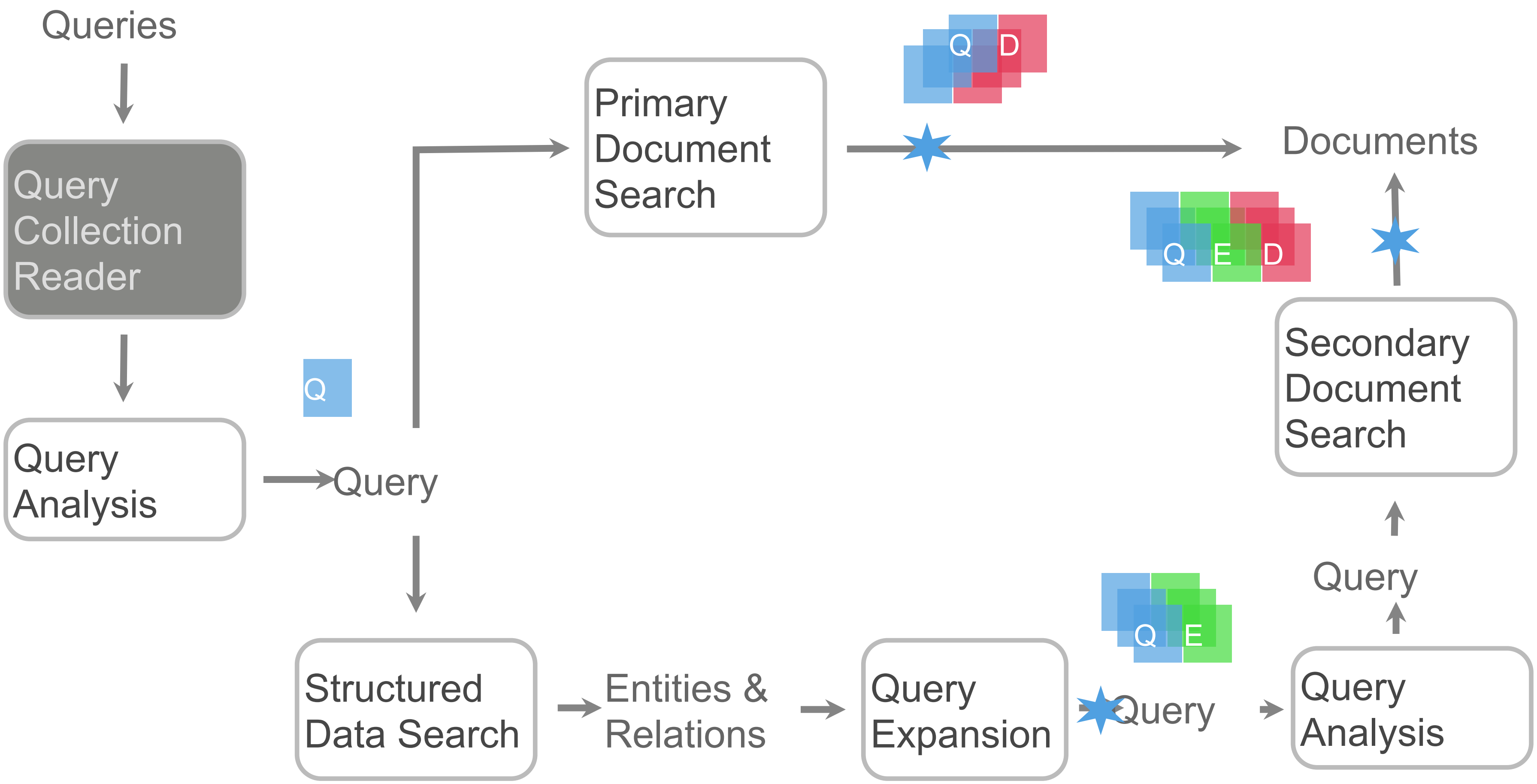
Code Orientation

- Top-level organization: src, lib, etc.
- Source code organization
- Descriptor organization

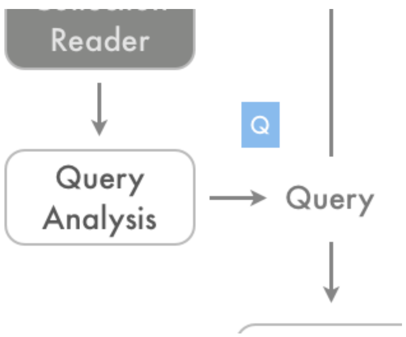
Developing a Component

- Design your component
- Define types
- Create type system
- Implement the source code
- Create descriptors
- Plug into the framework, and run

Stages of Pipeline



QueryAnalysis

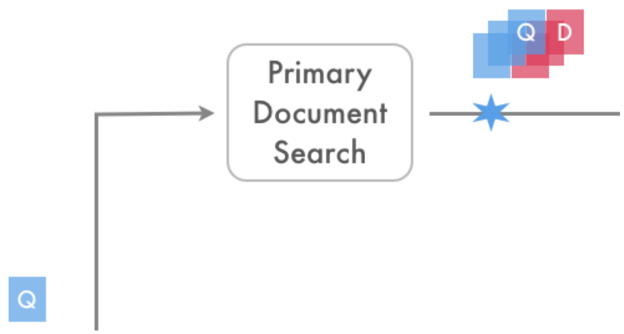


Input Views: QUERY

Input Annotations:
QueryString

Output Annotations:
-unspecified-

PrimaryDocumentSearch

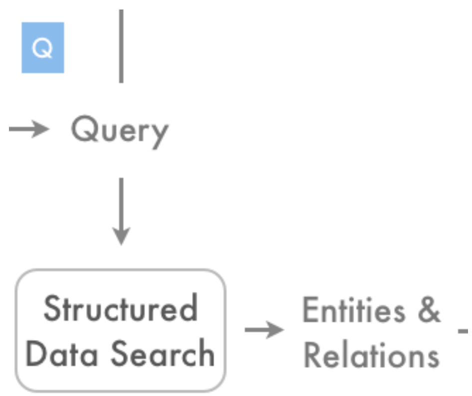


Input Views: QUERY

Input Annotations:
QueryString

Output Annotations:
SearchHitList
SearchHit

StructuredSearch



Input Views: QUERY

Input Annotations:
QueryString

Output Annotations:
StructuredLookupResult

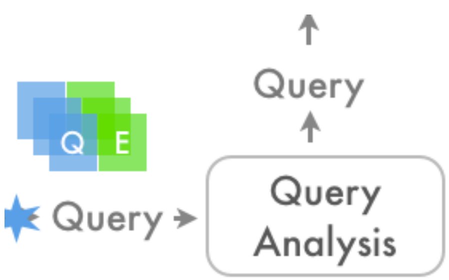
QueryExpansion

Input Views: QUERY

Input Annotations:
QueryString
StructuredLookupResult

Output Annotations:
DerivedQuery

ExpandedQueryAnalysis

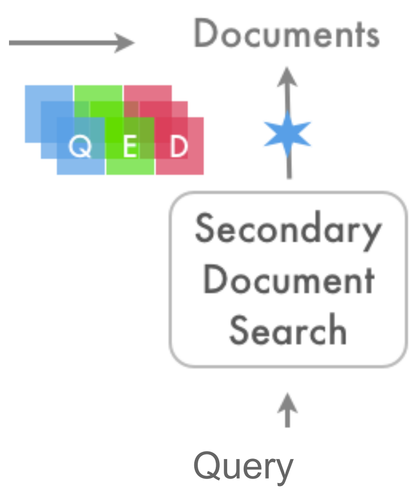


Input Views: QUERY, EXPANDED

Input Annotations:
QueryString(EXPANDED)

Output Annotations:
-unspecified-

SecondarySearch



Input Views: QUERY, EXPANDED

Input Annotations (EXPANDED):
QueryString

Output Annotations(EXPANDED):
SearchHitList
SearchHit

Example Component

- Identify person and organization names
  - simple dictionary lookup strategy
  - can possibly train a model
  - use off-the-shelf components
- Perform primary document search using person/org names as query

Implementation Strategy

- Query Analysis: detect person and organization names
- Primary Document Search: new search component uses detected named entities in queries

New UIMA Types

- Define new types (on paper):
  - single type for named entities, or
  - separate person and org types
- Extend project type system
- Generate Java classes using JCasGen

Named Entity Detector

- Implement the code for NE detector
  - create package
  - implement Java class
- Create descriptor for NE detector
- Add component to appropriate stage in project pipeline

Primary NE Search

- Implement the code for NE search
  - create package
  - implement Java class
- Create descriptor for component
- Add component to appropriate stage in project pipeline

Inspect the Output

- Observe different types of CASes
- Open CASes in Annotation Viewer
  - Inspect annotations and features
  - Observe different views
- Open CASes in CAS Visual Debugger
  - Inspect non-annotation UIMA structures

Off-the-Shelf Components

- Can easily use off-the-shelf UIMA code
- Example use case: apply OpenNLP to query analysis

OpenNLP Example

1. Download OpenNLP libraries
2. Download the trained models
3. Copy jar files into workspace (into the ‘lib’ project)
4. Add jars to the library path
5. Create new package under ‘descriptors’
6. Copy OpenNLP descriptors into the package
7. Add one or more OpenNLP components to QueryAnalysisComponents
a. SofA-map to QUERY view
b. ensure multiple deployment is not allowed
8. Ensure that resource bindings are appropriately set (to location of models)
9. Run the pipeline!
10. ???
11. Profit?

Summary

- We learned about…
  - Architecture of class project pipeline
  - Source code organization
  - Designing and developing a UIMA component for the project

References

- http://www.eclipse.org/documentation
- http://uima.apache.org/documentation.html
  - Tutorial and Users' Guides
  - Overview and Setup

- E. Epstein, M. Schor, B. Iyer, A. Lally, E. Brown, J. Cwiklik. Making Watson Fast. IBM Journal of Research and Development. Volume 56 (3-4). 2012.

Semantic Technologies in IBM Watson™

Lesson 9 – Domain Adaptation

Professor: Alfio Massimiliano Gliozzo
TA: Or Biran

Outline

- Domain Adaptation Methodology
- Content Adaptation
- Training Adaptation
- Functional Adaptation

Outline

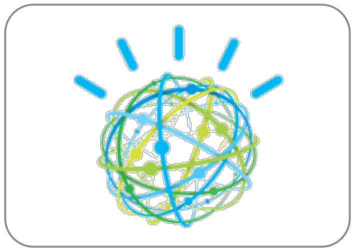
- Domain Adaptation Methodology
- Content Adaptation
- Training Adaptation
- Functional Adaptation

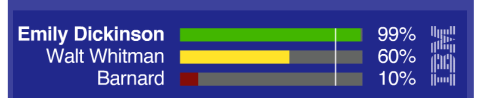
IBM WATSON

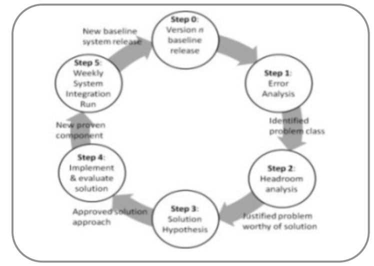
Taking Watson beyond Jeopardy!

Understanding	Interacting	Explaining	Learning
Specific Questions	Question-In/Answer-Out	Precise Answers & Accurate Confidences	Batch Training Process

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





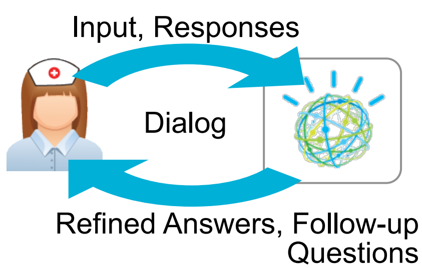


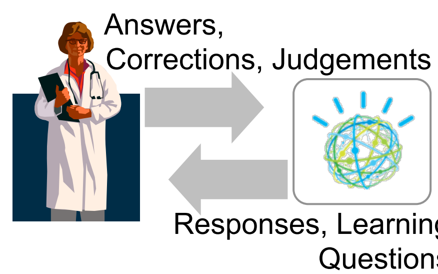
From specific questions to rich, incomplete problem scenarios (e.g. EHR)

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

Medical Adaptation – The Beginning – Doctor's Dilemma

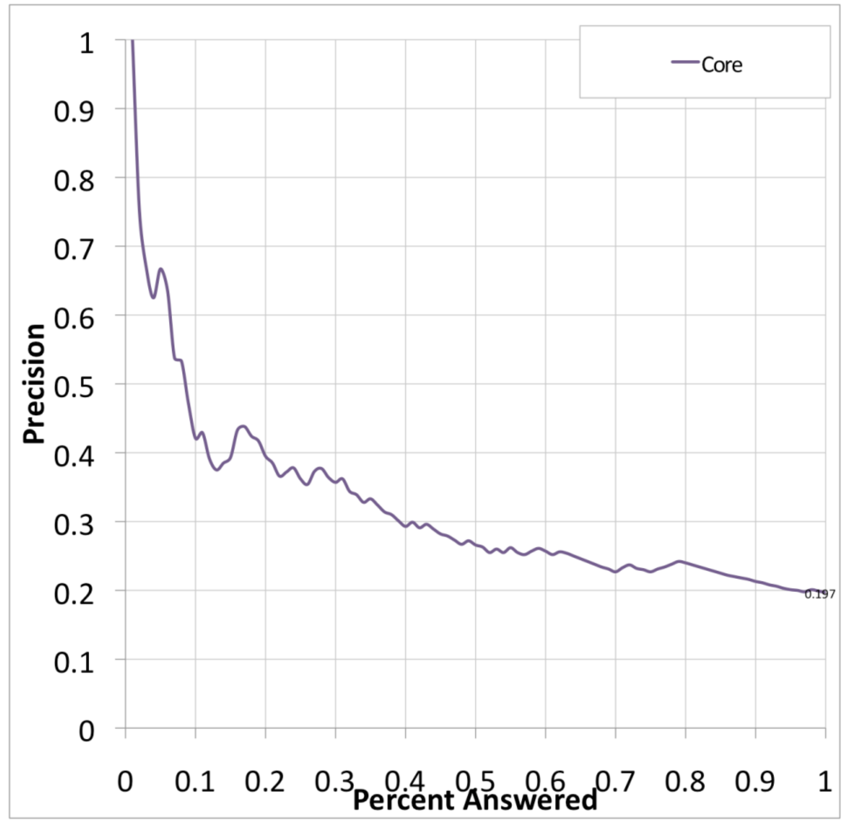
- American College of Physician's Doctors Dilemma Questions
  - Jeopardy-like questions for medical residents
  - Question has "single answer"
  - All and only relevant information given
- We ran Jeopardy system (out of the box) on 188 unseen diagnosis questions

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

Accuracy: 20%



Development cycle

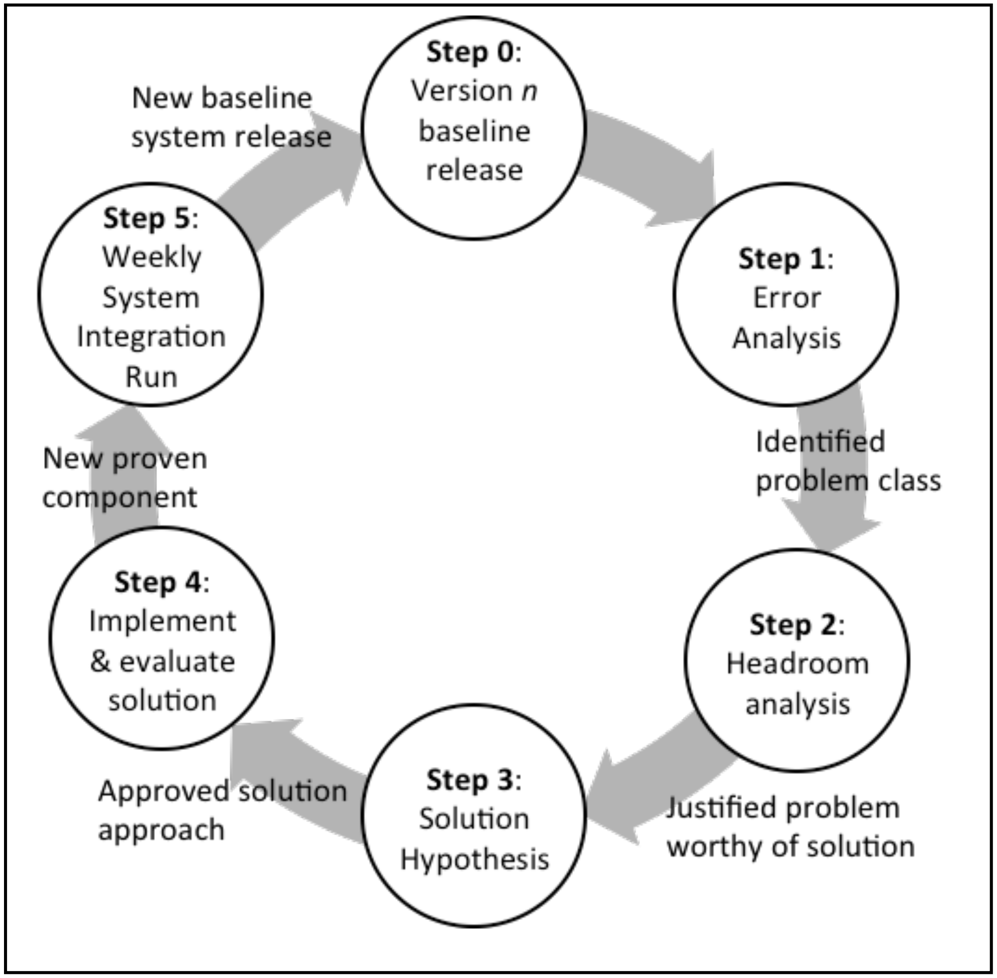


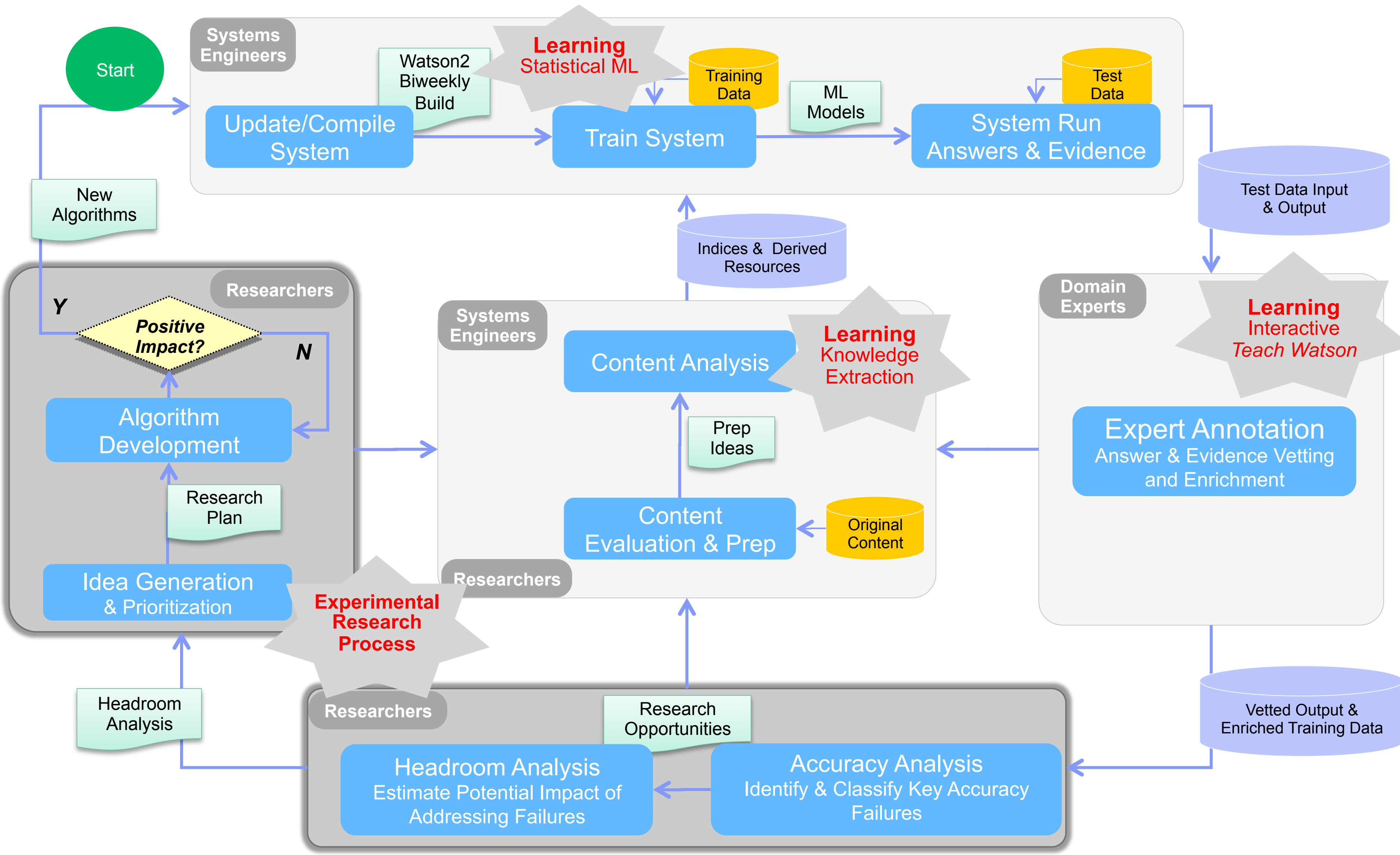
Figure 1 System development lifecycle.

- **4 years**
- **12 -> 30 people**
- metrics-driven research
  - Component level
  - End-to-end system
- informative **error analysis**
- **High investment in Methodology and Tools**

Step1: Tycor is weak in President Names
Step2: Suppose Tycor is able to recognize all President Names in the system, what's the gain
Step 3: if step 2 is significant

IBM WATSON

Domain Adaptation: Training and adapting Watson to new domains



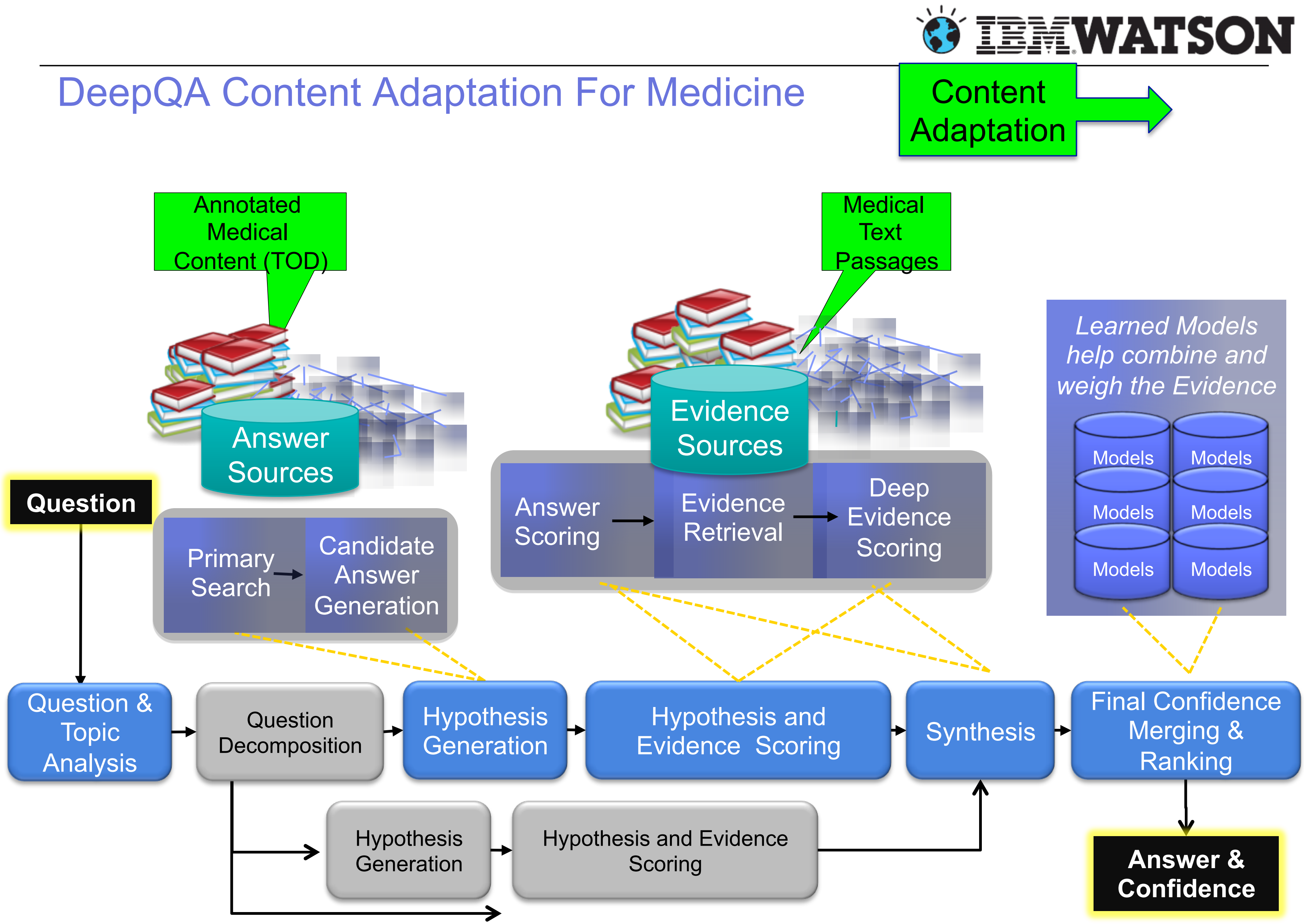
Adaptation: What do we have in a new domain?

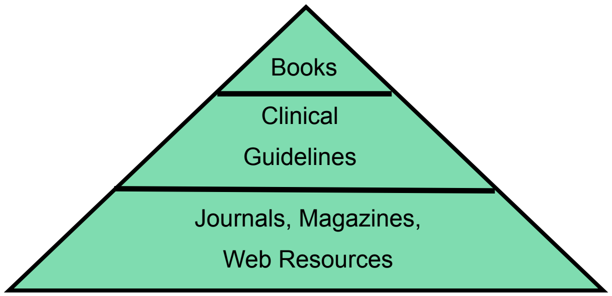
Content Adaptation

New Text Content

Structure and ingest text content

DeepQA Content Adaptation For Medicine





Content Forms of Evidence

Internal Medicine:	Harrison's, Cecil's, ACP, MERCK
Medical Dictionaries:	Stedman's, Taber's, Jablonski's
General Resource:	National Guidelines Clearinghouse, UMLS
Specific Resources:	American Heart Ass., National Heart Lung & Blood
Journals, Magazines:	NEJM, JAMA, BJM, Nature Medicine, Lancet
Web Resources:	NIH, NIM, Wikipedia, WWW

Unstructured Sources

When **frequent urination** is accompanied by **fever**, an urgent need to urinate, and **pain or discomfort** in the abdomen, you may have a urinary tract infection.

Cranberries

Many women get Urinary Tract Infections (UTI), also known as bladder infections. It is said that if you drink cranberry juice when you feel a UTI coming on that the UTI may go away.

Urinary Tract Infection

The most common symptoms of a bladder infection are **burning** with urination, **frequency of urination**, an urge to urinate, **without** vaginal discharge or significant pain. [4] An upper urinary tract infection or pyelonephritis may additionally present with flank pain and a **fever**. Healthy women have an average of 5 days of symptoms.[4] The symptoms of urinary tract infections may vary with age and the part of the urinary system that was affected. In young children, urinary tract infection symptoms may include **diarrhea**, loss of appetite, **nausea** and **vomiting**, fever and **excessive crying** that cannot be resolved by typical measures. [5] 6]

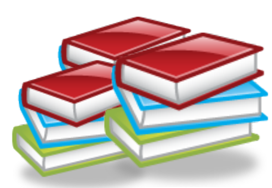
Cystitis

Symptoms:
- Sudden onset
- Dysuria (painful **urination**)
- Nocturia
- **Low back pain**
- Pneumaturia

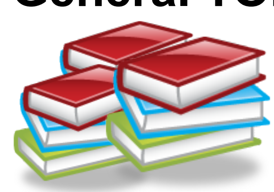
Diagnosis:
- Urianalysis
- Urine culture

Content Organization

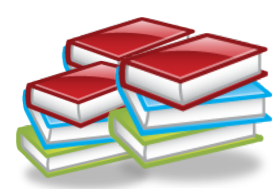
Medical TODs



General TODs



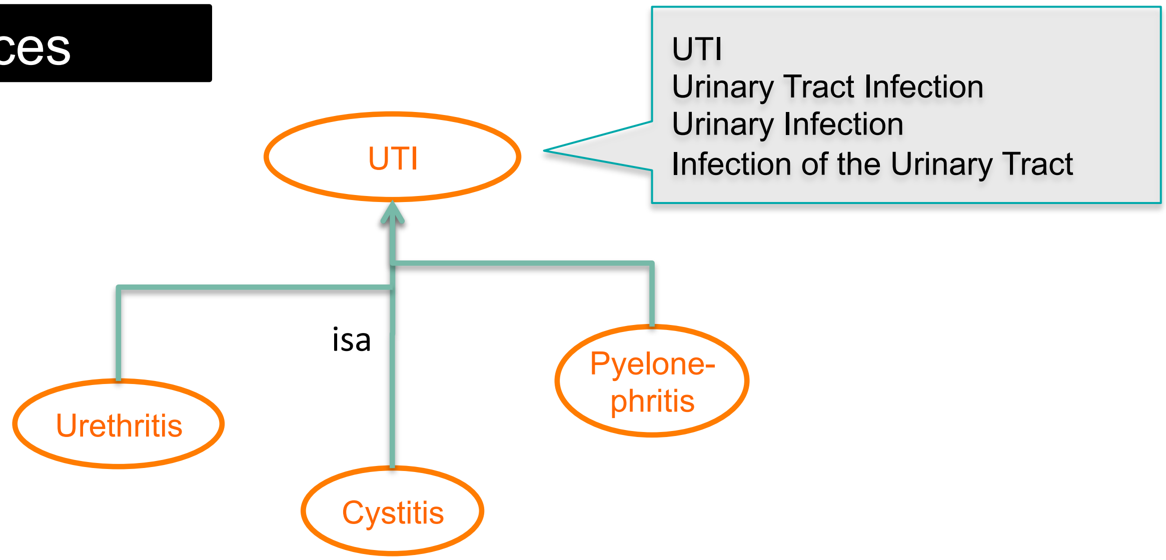
Non TODs



Content
Forms of Evidence

Structured Sources

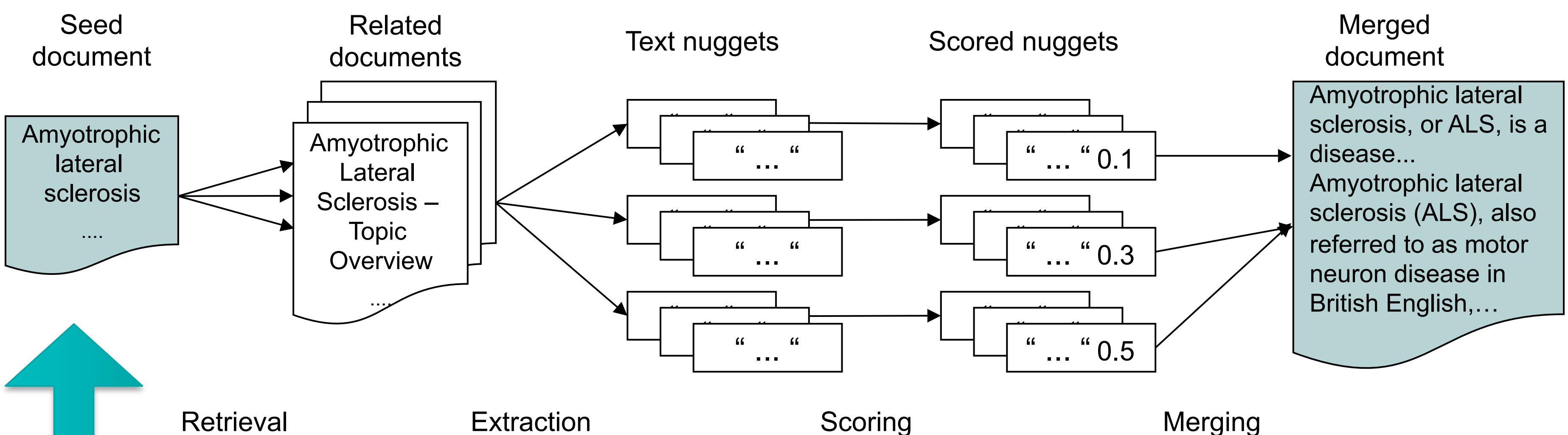
NLM UMLS
(Unified Medical
Language System)



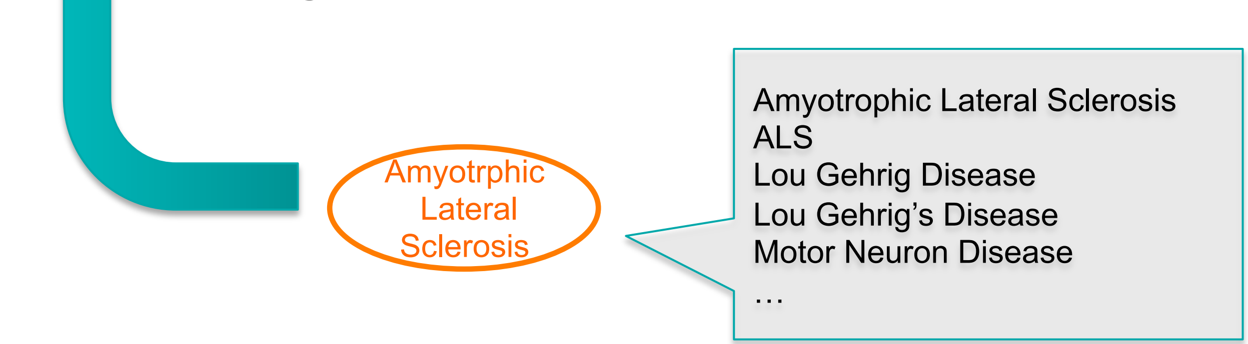
Hypothesis Generation

A harsh, late-peaking, basal murmur radiating to the carotid arteries suggests aortic stenosis; a systolic murmur that increases with the Valsalva maneuver and disappears with squatting suggests hypertrophic cardiomyopathy. A systolic click followed by a flowing systolic murmur that moves closer to the 1st heart sound on standing suggests mitral valve prolapse (suggesting the cause is an arrhythmia).

Corpus Expansion: Focused Augmentation of Existing Content

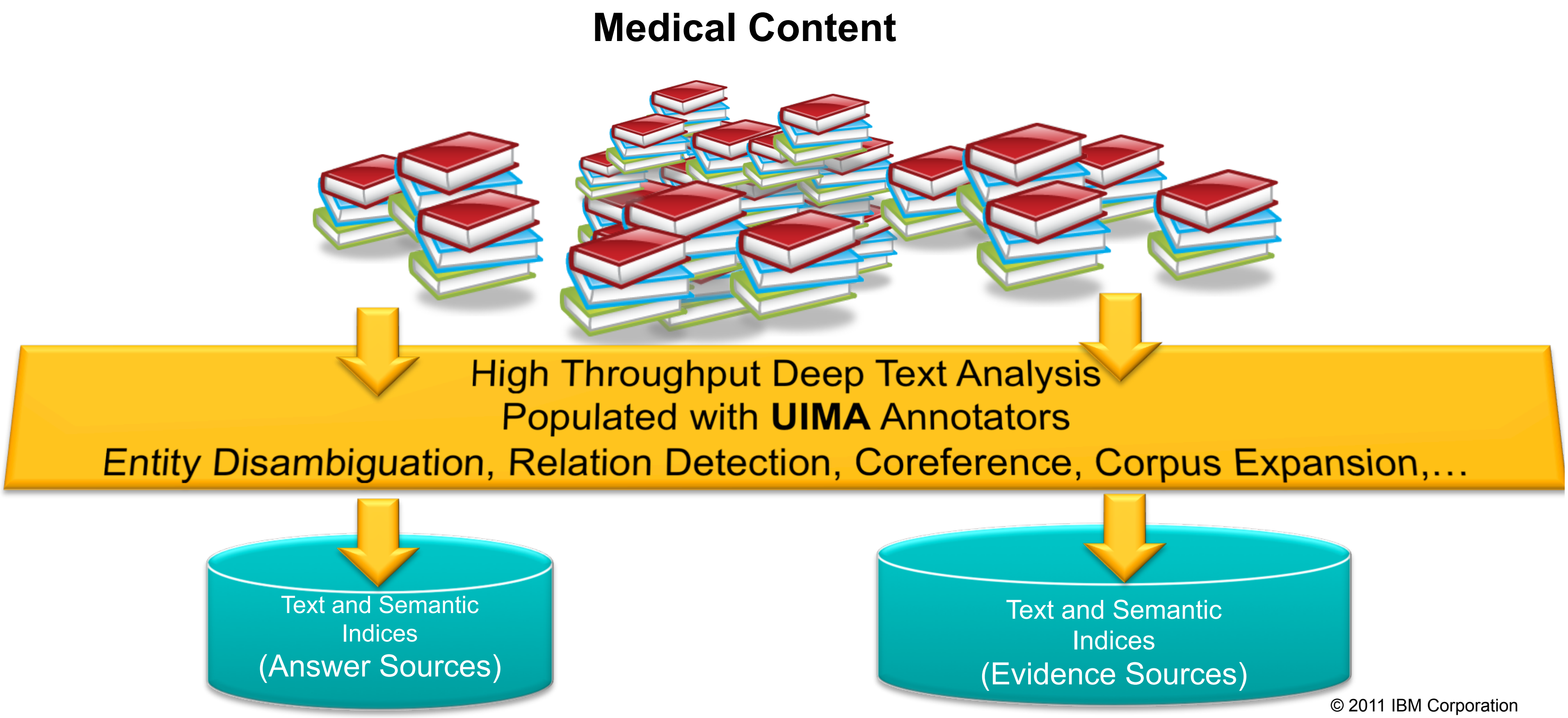


Using Structured Sources in Corpus Expansion



Before Runtime: Content Ingestion

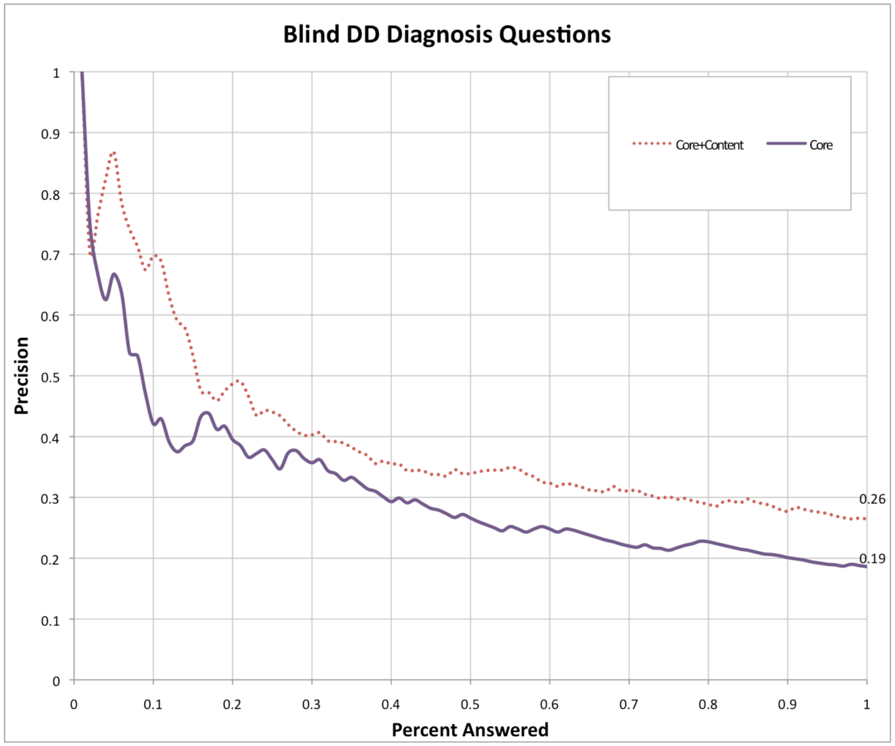
- Content Ingestion
  - Format and preprocessing (XMLparsing, tokenization, special characters)
  - Structuring (Organizing by condition / treatment / etc.)
  - Extending (Corpus Expansion)



Medical Adaptation – Content (Feb 2012 results)

Accuracy: 26%

- Content Ingestion
  – Format and preprocessing (XML parsing, tokenization, special characters)
  – Structuring (Organizing by condition)
  – Extending (Corpus Expansion)
- Performance graph excludes
  – New content acquired since 2/2012
  – Hypothesis generation using structured medical source
- Work in progress
  – Medical focused corpus expansion
  – Seedless corpus expansion



Adaptation: What do we have in a new domain?

Content Adaptation

New Text Content

Structure and ingest text content

Training Adaptation

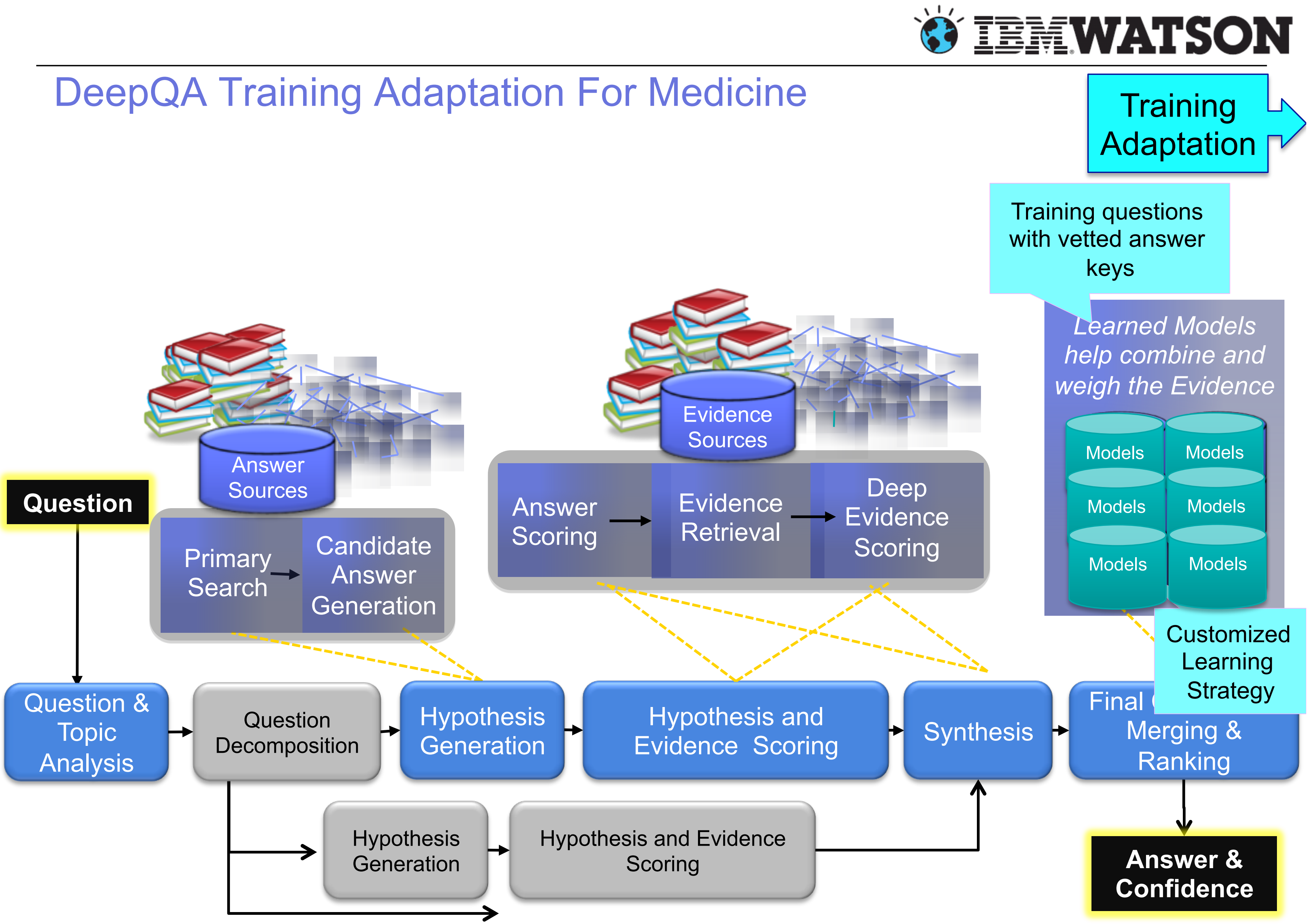
New "Questions"

Train the system on target scenarios

58-year-old woman presenting to her primary care physician after several days of dizziness, anorexia, dry mouth, increased
She had also had a f
would "get stuck" wh
reported no pain in h
no cough, shortness
Her family history inc
in her mother, Grave

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

DeepQA Training Adaptation For Medicine



IBM WATSON

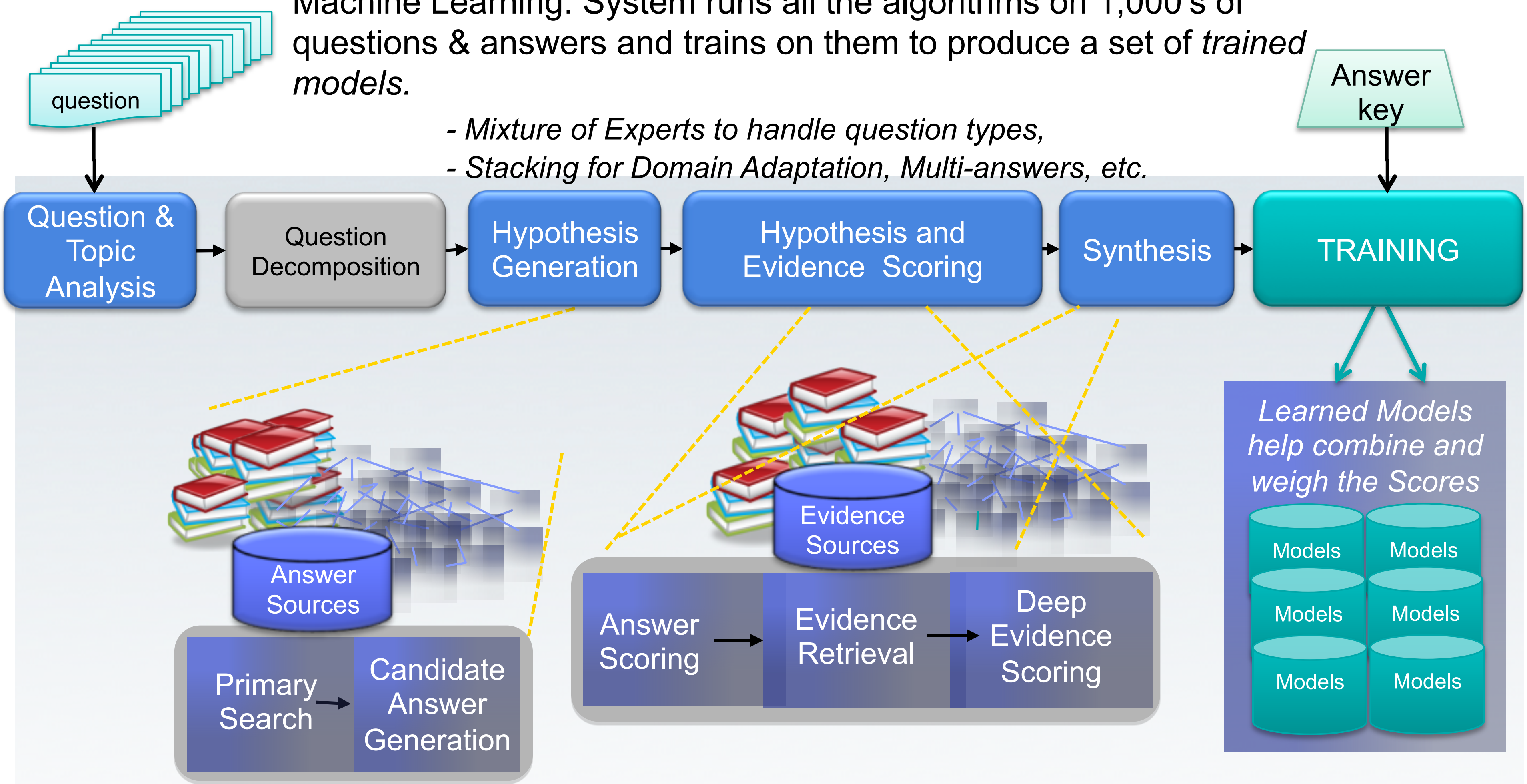
Before Runtime: Machine Learning

Train-time

Machine Learning: System runs all the algorithms on 1,000's of questions & answers and trains on them to produce a set of *trained models.*

- Mixture of Experts to handle question types,

- Stacking for Domain Adaptation, Multi-answers, etc.

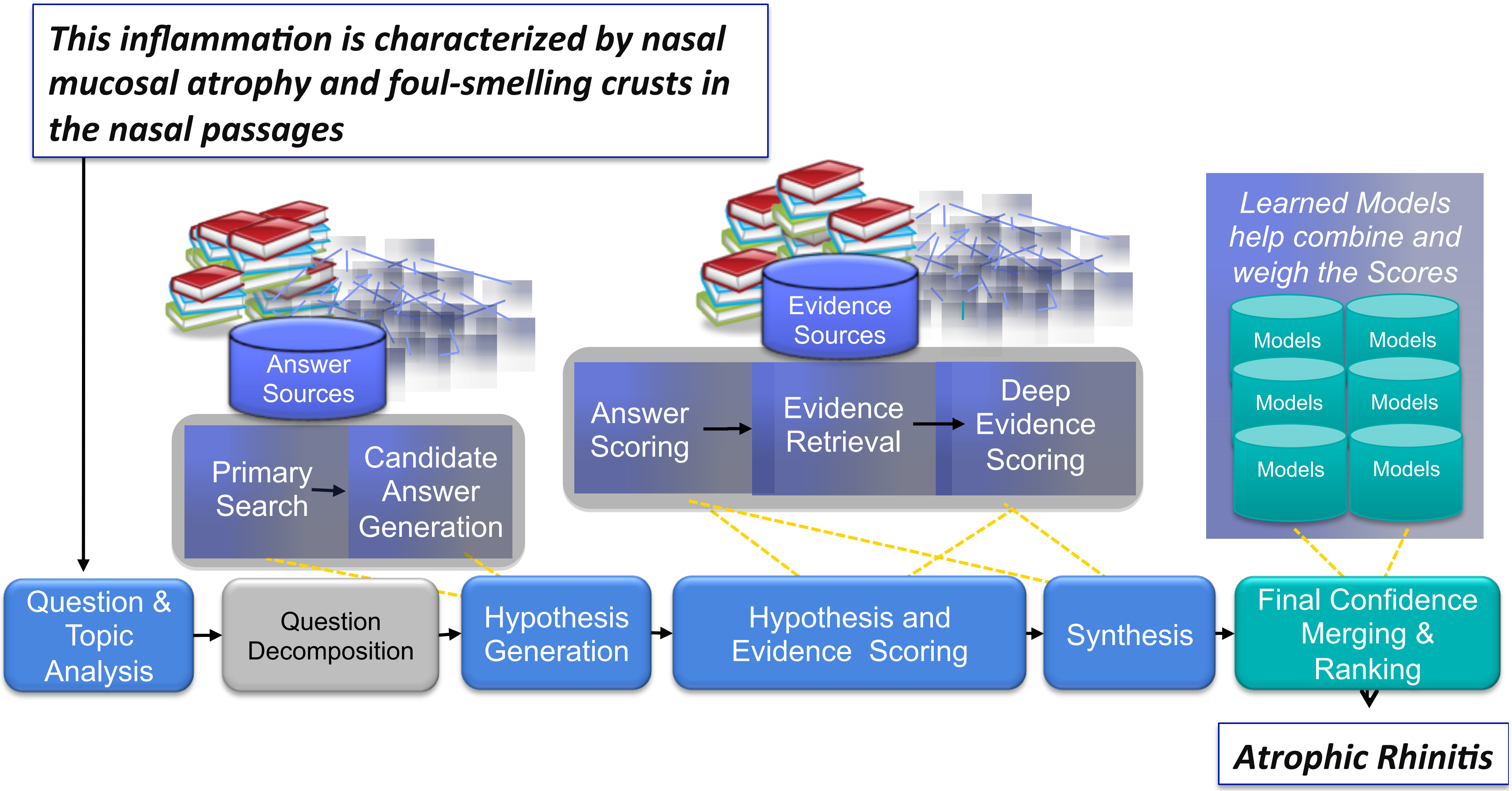


In 4 years DeepQA has trained for over 10,000 recorded experiments

Runtime Time

Apply-time

System runs all the algorithms on a *single question* and applies models to the features they produce to select an answer with a confidence



Doctors Dilemma Answer Vetting

Watson finds more acceptable answers

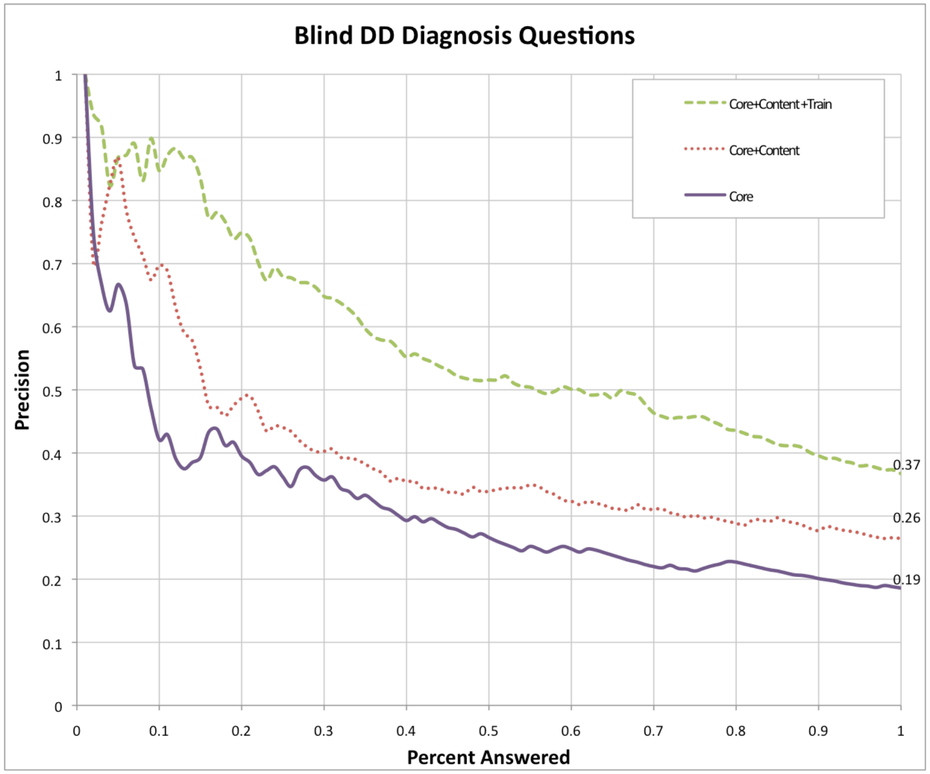
Condition characterized by hypertension and unprovoked hypokalemia *(Answer: Apparent mineralocorticoid excess syndrome)*		
0.777	Hyperaldosteronism	~~Incorrect~~ Acceptable
0.511	Primary aldosteronism	~~Incorrect~~ Acceptable
0.277	Liddle's syndrome	~~Incorrect~~ Acceptable
0.069	Metabolic Alkalosis	No
0.055	Bartter syndrome	No
0.049	aldosterone	No
0.028	Apparent mineralocorticoid excess syndrome	Correct

In non-STEMI, addition of this antithrombotic agent is superior to heparin alone *(Answer: IIb / IIa receptor inhibitor)*		
0.254	Bivalirudin	Incorrect
0.168	IIb receptor inhibitor	Correct
0.125	Warfarin	Incorrect
0.123	Clopidogrel	~~Incorrect~~ Correct
0.076	Aspirin	Incorrect
0.042	IIa receptor inhibitor	Correct
0.039	tirofiban	~~Incorrect~~ Correct
0.019	LMWH	Incorrect

Medical Adaptation – Training

Accuracy: 37 %

- Trained on 1322 DD questions
- Tuned ML technique…
- And had medical experts vet the top 50 answers for 1322 training questions (66,100 answers vetted)



Adaptation: What do we have in a new domain?

Content Adaptation

New Text Content

Structure and ingest text content

Training Adaptation

New "Questions"

Train the system on target scenarios

58-year-old woman presenting to her primary care
physician after several days of dizziness, anorexia,
dry mouth, increased
She had also had a
would "get stuck" wh
reported no pain in h
no cough, shortness
Her family history inc
in her mother, Grave

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

Functional Adaptation

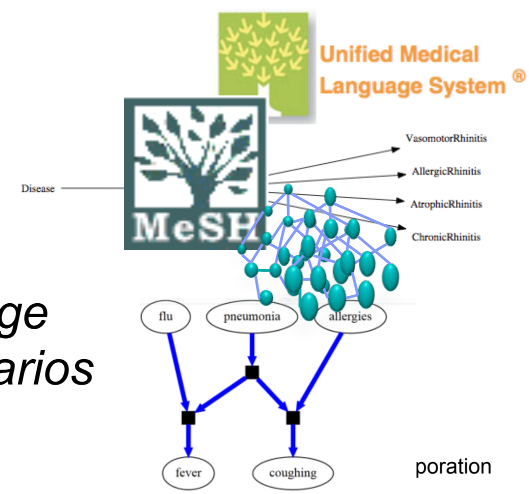
New Concepts / Reasoning / Discourse

Enhance the functional capabilities with domain-specific

- Concepts: entities, relations from domain modeling

- Reasoning: domain axioms and background knowledge

- Discourse: algorithms for domain text / problem scenarios



Functional Adaptation

IBM WATSON

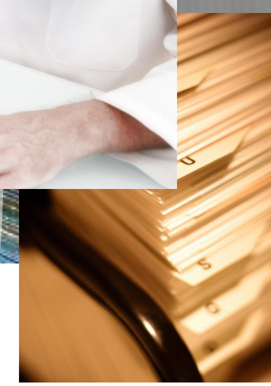
Information Extraction: Identifying important concepts in text

Which medications have been used for neuropathic pain for this patient?

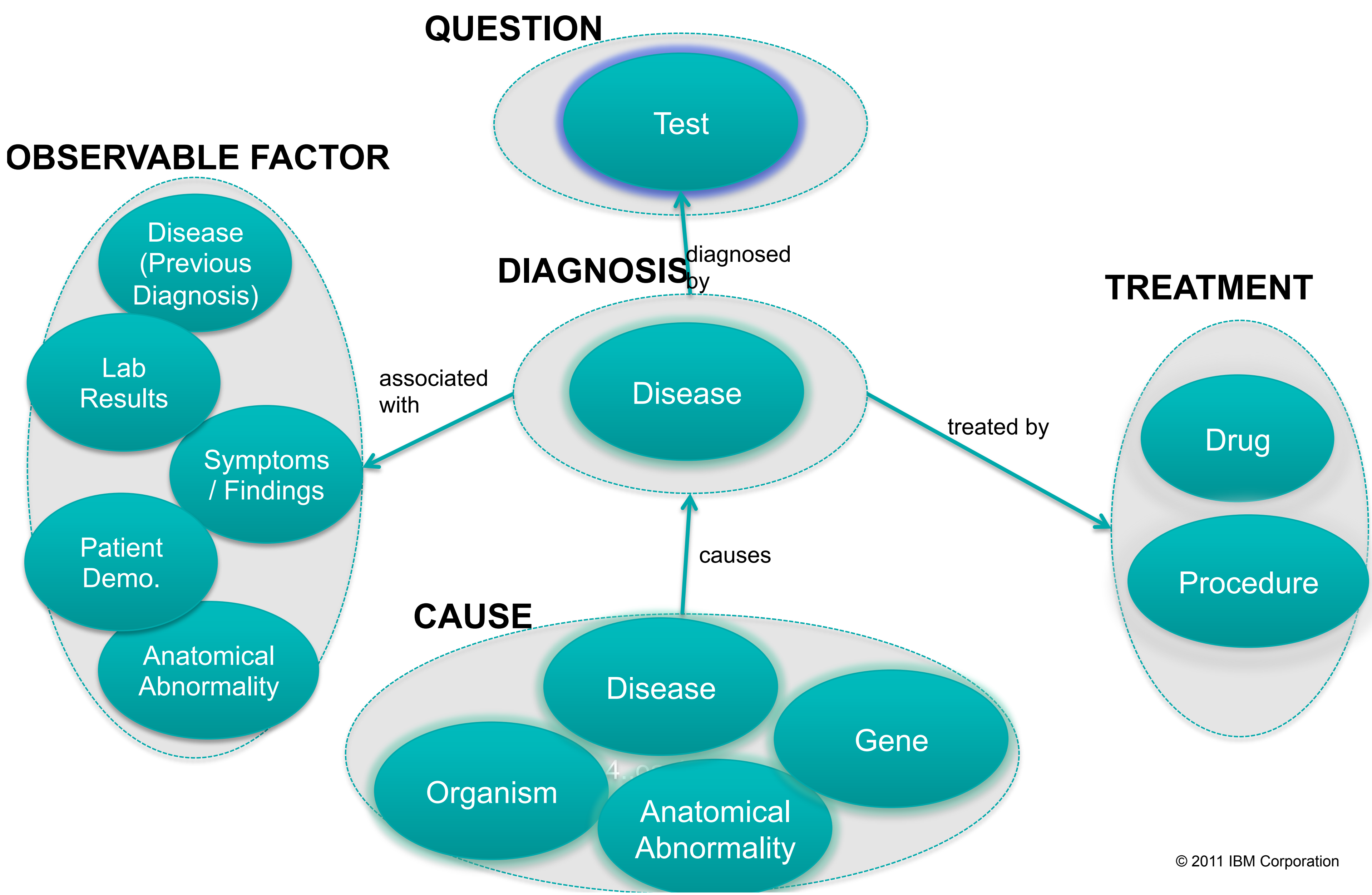
Information Extraction: what's there?

Semantic Analysis: how does it relate to what you need?

- Matching / Textual Entailment
- Background Knowledge

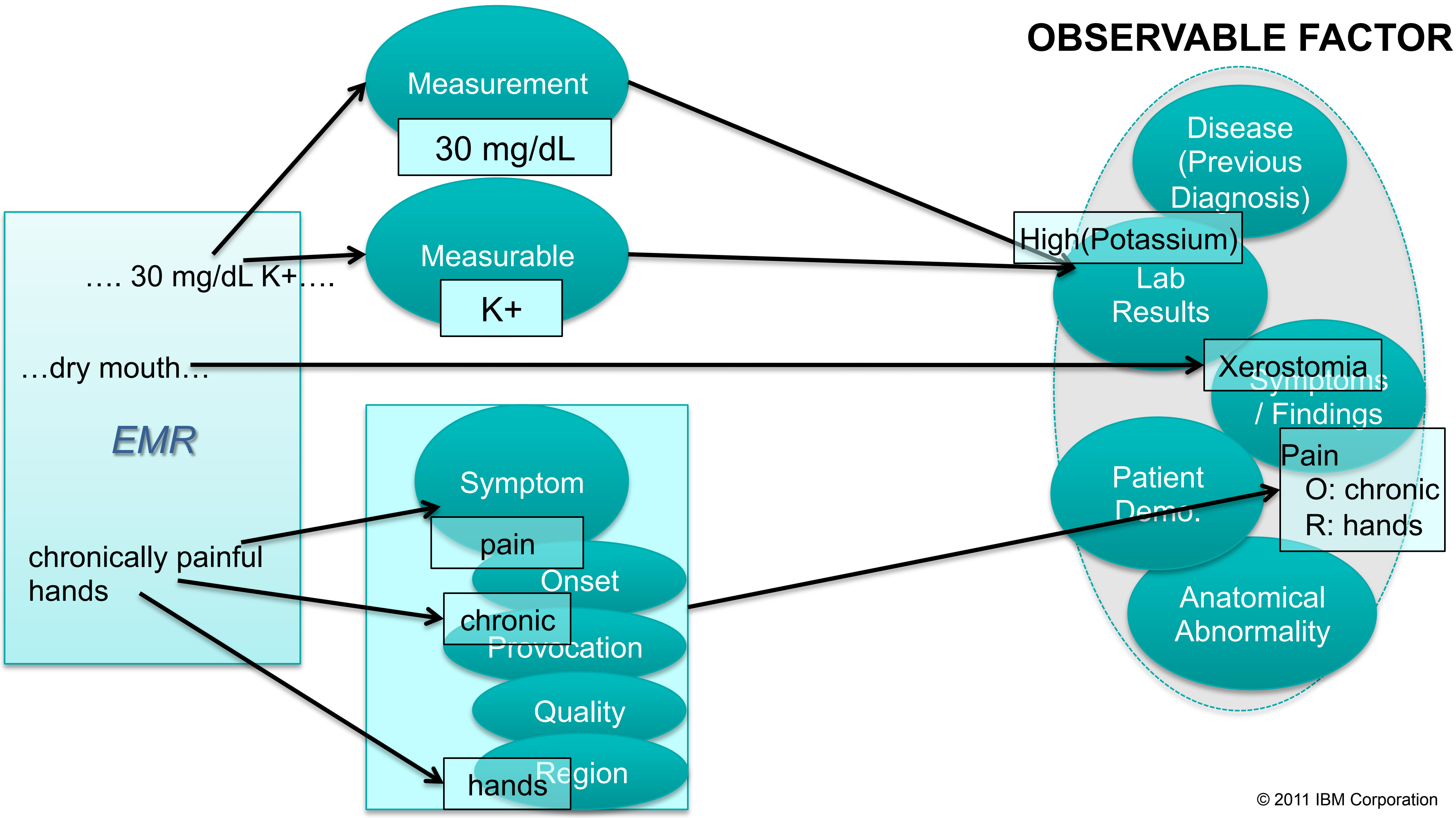


High Level Domain Model



Mapping from language to knowledge

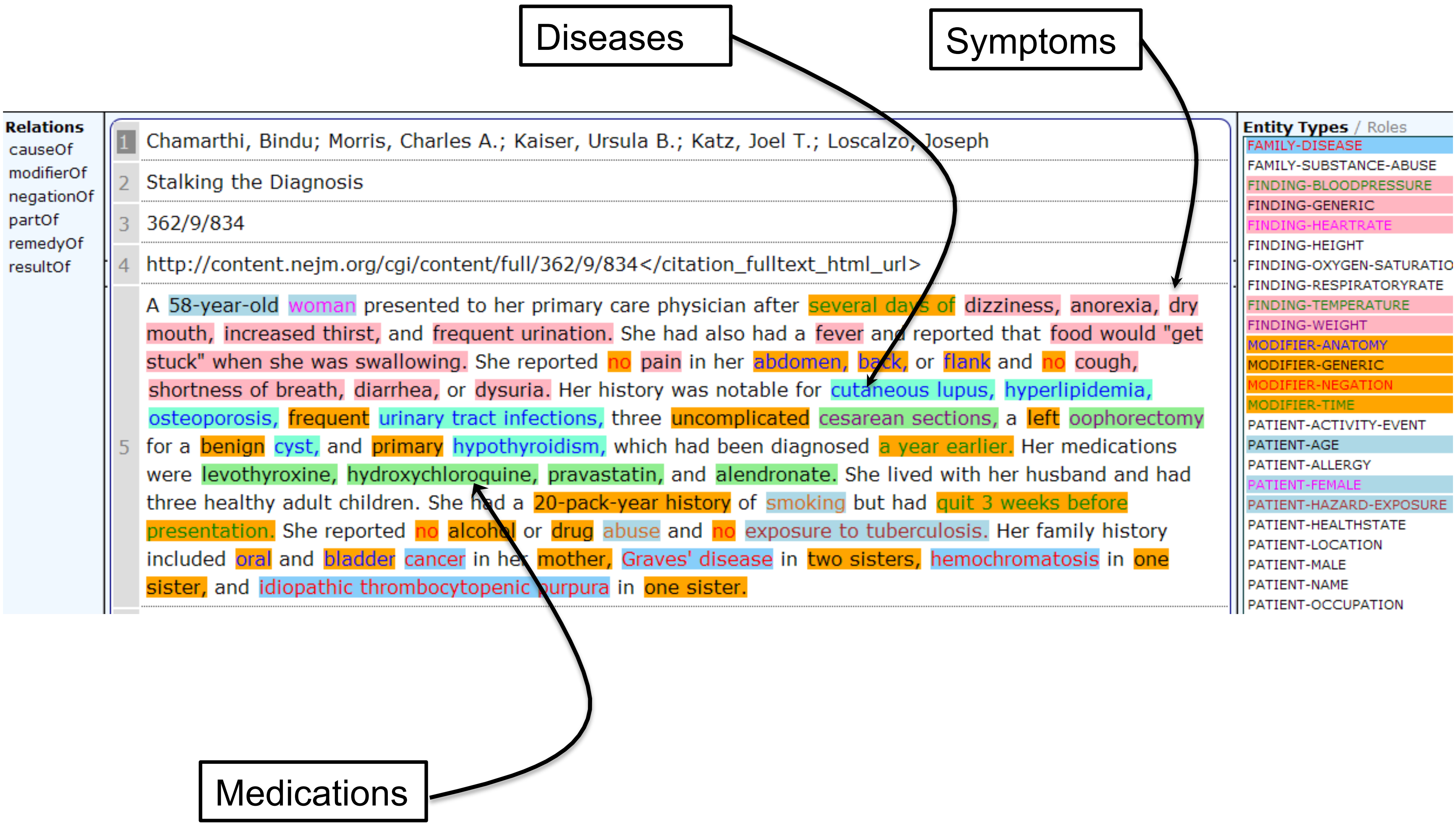
Beyond simple keywords…



Identifying Signs or Symptoms

Patient presenting with **polyuria, polydipsia, weight loss x 2wks**	Straightforward, duration (**x 2wks**)
Head NC/AT, PERRLA, EOMI, anicteric, MMM, no oral lesions	Abbreviations, negation
syncope unlikely to be of neurogenic origin; no witnessed **seizures**;	"unlikely to be neurogenic", "no witnessed"= not present, but not fully negated
EKG normal, placed patient on tele to monitor for **dysrhythmias**	**dysrhythmias** in text but not observed or negated
Interval slight enlargement of previously seen **mediastinal lymph nodes.**	Must associate **enlargement** with **mediastinal lymph nodes**, previously noted
She has **pain with straight leg raise** , **left** greater than **right** and 5/5 strength in right quads compared to 4/5 on the left	Have to associate **pain** with **straight leg raise**, **left** greater
Findings compatible with **chronic interstitial lung disease.**	Class of symptoms, "compatible with"
No findings consistent with **ischemic colitis.**	Class of symptoms, "consistent with"

NEJM Medical Concept Annotations



Extraction Headroom on 1700 NEJM Findings

Sample of UMLS Misses
parasternal heave
bilateral breath sounds,
tonsillar pillars were red
jugular veins were distended
hands and feet were warm, dry, and nonedematous.
hemorrhages in the nail beds of both hands and the left foot
cardiac examination revealed a sustained point of maximal impulse, a regular rate and rhythm, and the absence of a murmur, rub, or gallop.
site of entry of a right subclavian dialysis catheter appeared unremarkable.

- Percent of gold standard Finding+Regions which do **NOT fall under *some CUI by UMLS: 76.6%***
- Percent of gold standard Findings which do **NOT fall under a *Finding CUI: 36.7%***
- Percent of gold standard Regions which do **NOT fall under a *Region CUI: 43.9%***

Parse-Based Entity Detection

IBM WATSON

- Use syntactic parses to map phrases to known medical concepts

	Additional concepts w/ parse
heart sounds were distant	Distant heart sounds
pain in the abdomen	Abdominal pain
nasal and respiratory tract chondritis	Nasal chondritis
both wrists and the left knee were tender	Tender wrists, Tender knee
leg pain and swelling	leg swelling
developmental language delay	developmental delay, language delay

- Use syntactic parses to hypothesize unknown medical concepts

Acute gastrointestinal pain	Likely a type of “gastrointestinal pain”
Generalized onset seizure disorder	Likely a type of “seizure disorder”

Lab Tests

His electrolytes are **normal**	Text description
The patient's liver function tests are **all out of whack**	Text description
Albumin, serum **3.5-5.5 g/dL (35-55 g/L)**	Range (two units)
His AST 349, ALT 186, alk-phos 138	abbreviation
Oxygen saturation **95% or greater**	“or” greater than
Red Blood Cell Count (RBC) Normal Value: Male: **4.3-6.2x106/ul**	Range, conditional
Hemoglobin threshold (g/dl) Children **11.5** Teens **12.0** Women, pregnant **11.0**	Conditional (implied greater than)
The patient's BUN and creatinine yesterday were **23 and 2.1 respectively**	“respectively”
H&H of **17.7 and 51.6**	Abbreviation, respectively

Negation

No **breathing**	“No”
Generally not **itchy**.	“Generally”
non-viral hepatitis	“non-”, class
the absence of the **rash**	“absence of”
No CT evidence of **intra-abdominal** or **intra-pelvic adenopathy, masses** or **obstruction**	Negation distributes over “**or**”
Breathing problems, including **no breathing**, **shortness of breath**, or **rapid breathing**.	Negation does not distribute over “**or**”
Hearing loss, with or without **dizziness**	negation or presence
Aneurysms usually cause no **symptoms** unless they rupture and cause **bleeding** into the brain	“unless”
acute bronchitis has cleared	“has cleared”
ALS does not affect the senses (**sight, smell, taste, hearing**, **touch**), **bladder** or **bowel function**, or **a person's ability to think**	Classes of symptoms

Coreference

With aortic stenosis, the murmur is systolic, beginning after S1 and ending at or before aortic valve closure. It's harsh and grating, medium-pitched, and crescendo-decrescendo	It = murmur
The diagnosis is confirmed with positive findings on Finkelstein's test. To perform this test, the patient places the thumb…	This test = Finkelstein's test
Elevation in the serum total bilirubin and alkaline phosphatase concentrations are not common in uncomplicated cholecystitis, since biliary obstruction is limited to the gallbladder; if present, they should raise concerns about complicating conditions such as…	They = elev in serum total bilirubin concentration + elev in alkaline phosphatase concentration
Collagenous colitis and lymphocytic colitis are distinguished by the presence or absence of a thickened subepithelial collagen layer. The cause of microscopic colitis syndrome is uncertain.	Microscopic colitis = collagenous colitis + lymphocytic colitis
A 5-year-old boy has papular and pustular lesions on his face. A serous honey-colored fluid exudes from the lesions. A Gram stain of the pus reveals many neutrophils and Gram-positive cocci in chains. The organism is non-	The lesion = lesions The pus = fluid The organism = cocci

Specificity

Cardiology (400): Most common congenital abnormality associated with coarctation of the aorta	abnormality	heart defect (Congenital heart defect)	0.182

Question: ***CARDIOLOGY: Most common congenital abnormality associated with coarctation of the aorta***

Correct Answer Pattern (Acceptable) (\W|^)((Bicuspid aortic valve))(\W|$)

Exp 8599: DD Week51 DD base model (w/SPR)

Answer Selection

Version:

Rank	Score [0.000:1.000]	Answer	Correct
1	0.182	heart defect	No
2	0.172	Patent ductus arteriosus	No
3	0.148	Turner syndrome	No
4	0.104	Bicuspid aortic valve	Yes
5	0.059	Congenital heart disease	No
6	0.048	Ventricular septal defect	No
7	0.032	Atrial septal defect	No
8	0.022	Double aortic arch	No
9	0.020	monosomy	No
10	0.018	malformations	No

Functional Adaptation

Semantic Analysis: Matching equivalent/related concepts

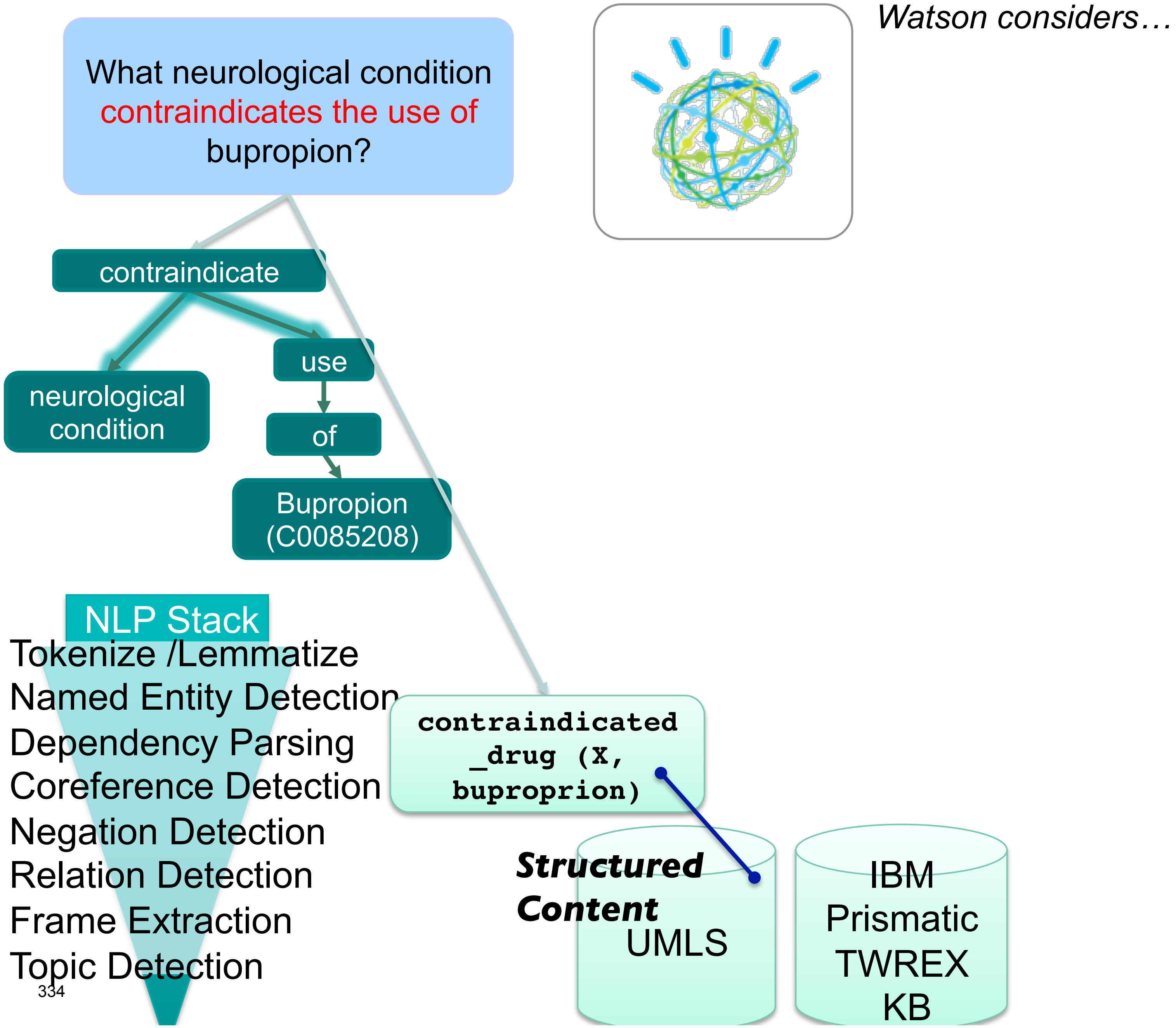
Which medications have been used for neuropathic pain for this patient?

Information Extraction: what's there?

Semantic Analysis: how does it relate to what you need?

- Matching
- Background Knowledge

Watson uses an ensemble of NLP techniques justify an answer



IBM WATSON

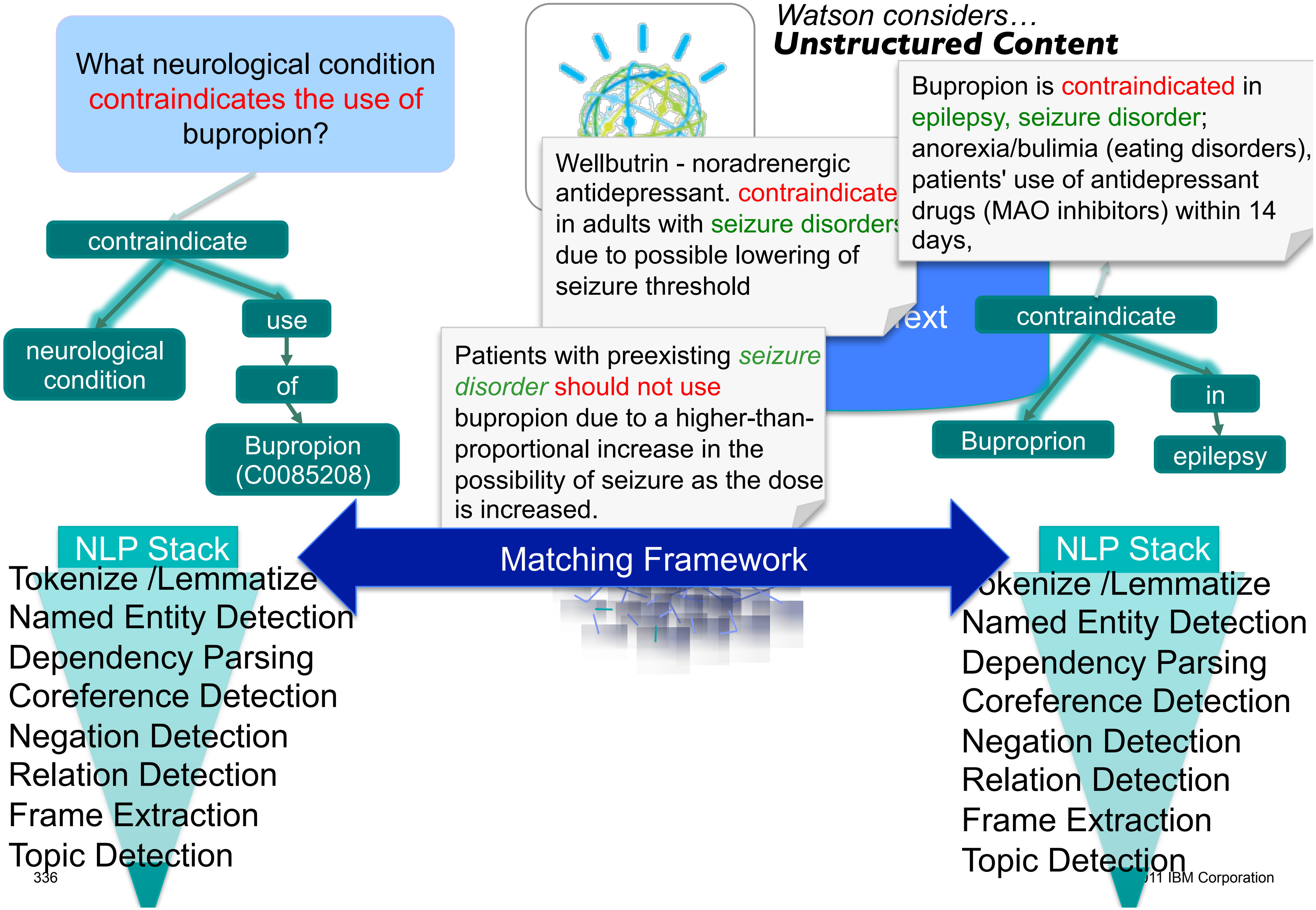
Relation Inventory

Structured Content

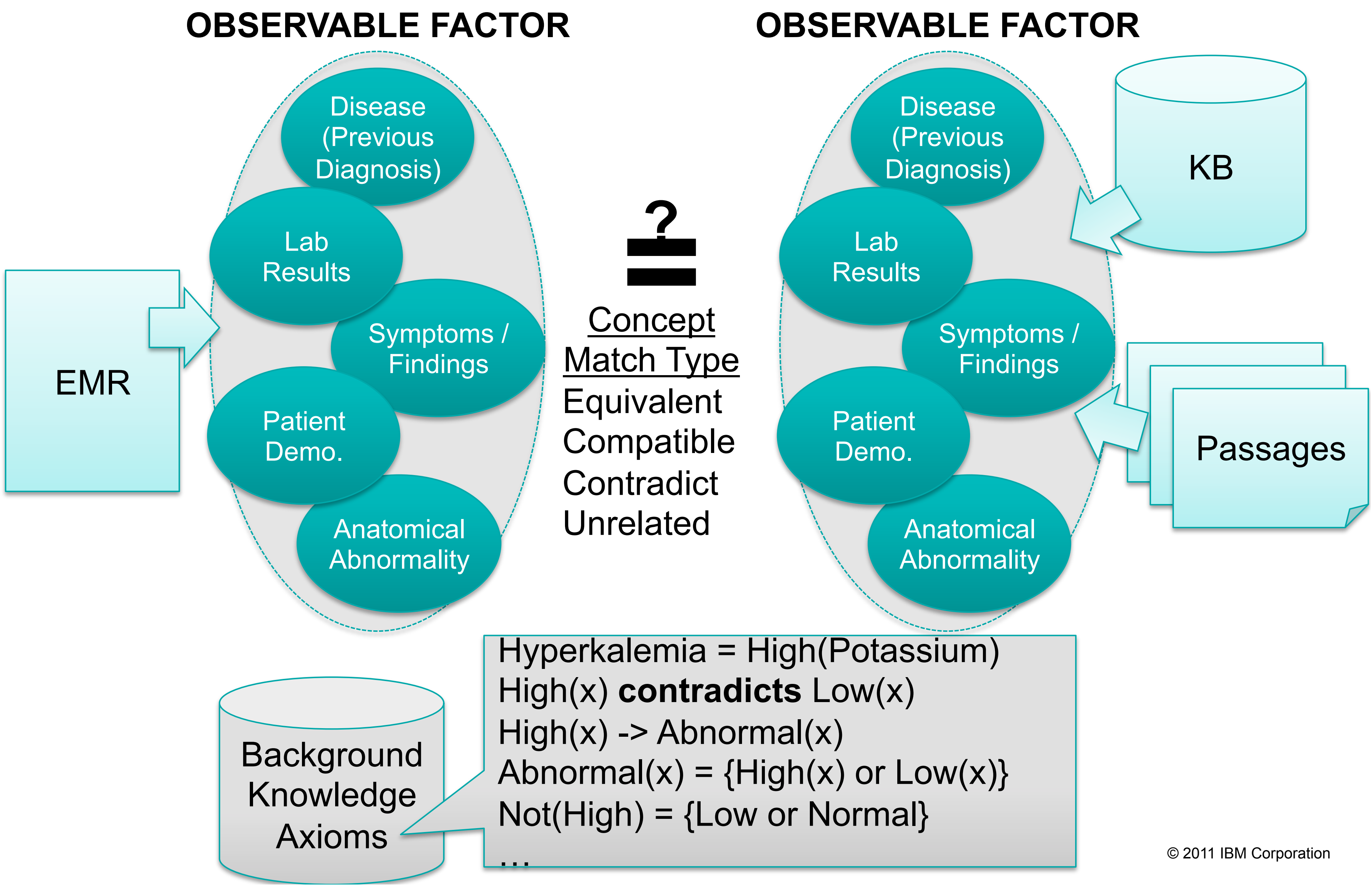
UMLS

*6% of diseases have a **treatment** identified in UMLS*

Relation	KB	Count	IBM Automatic Extension
isa (hyponomy)	MeSH / SNOMED	1345388	Hyponymy rules (Prismatic)
partOf (meronymy)	UMLS	19436	TWREX
disease_has_finding	UMLS	17540	TWREX, SKB
may_treat	UMLS	43780	TWREX, SKB
causes	UMLS	3569	TWREX, SKB
prevented by	UMLS	5823	TWREX, SKB
diagnosed_by	UMLS	937	TWREX, SKB
location_of	UMLS	15277	TWREX, SKB
contraindicated	UMLS	71018	TWREX, SKB


What neurological condition contraindicates the use of bupropion?
contraindicate
neurological condition
use
of
Bupropion (C0085208)
Watson considers…
Unstructured Content
Bupropion is contraindicated in epilepsy, seizure disorder; anorexia/bulimia (eating disorders), patients' use of antidepressant drugs (MAO inhibitors) within 14 days,
Wellbutrin - noradrenergic antidepressant. contraindicate in adults with seizure disorders due to possible lowering of seizure threshold
Patients with preexisting seizure disorder should not use bupropion due to a higher-than-proportional increase in the possibility of seizure as the dose is increased.
ext
contraindicate
in
Buproprion
epilepsy
NLP Stack
Tokenize /Lemmatize
Named Entity Detection
Dependency Parsing
Coreference Detection
Negation Detection
Relation Detection
Frame Extraction
Topic Detection
Matching Framework
NLP Stack
okenize /Lemmatize
Named Entity Detection
Dependency Parsing
Coreference Detection
Negation Detection
Relation Detection
Frame Extraction
Topic Detection

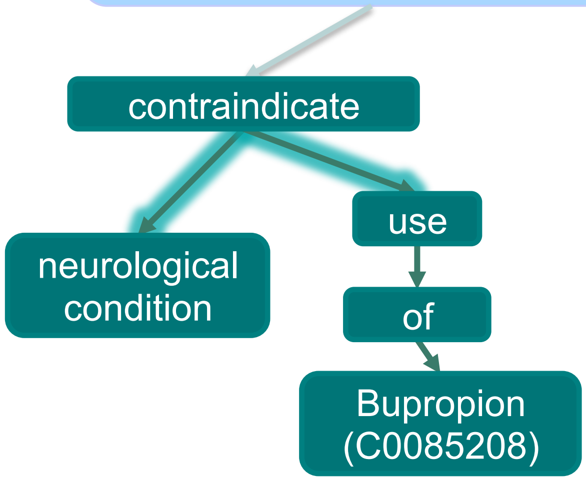
“Factor Matching”



Watson uses an ensemble of NLP techniques justify an answer

IBM WATSON

What neurological condition contraindicates the use of bupropion?



NLP Stack

Tokenize /Lemmatize
Named Entity Detection
Dependency Parsing
Coreference Detection
Negation Detection
Relation Detection
Frame Extraction
Topic Detection

Need to consider the type (“**neurological condition**”) of the answer for possible candidates:

- Epilepsy
- Seizure disorder
- Anorexia
- Bulimia

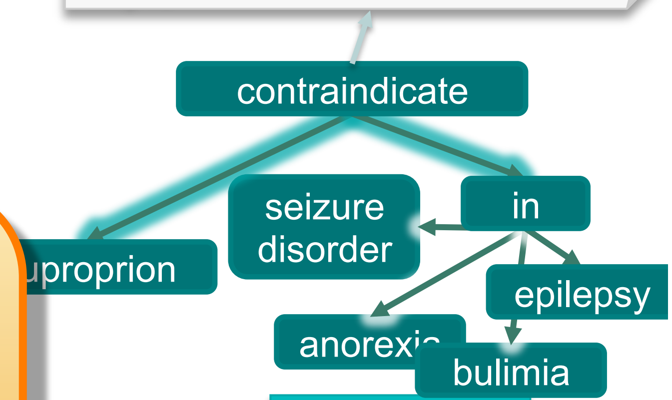
Structured Content

UMLS

Prismatic mined KB

Watson considers…
Unstructured Content

Bupropion is contraindicated in epilepsy, seizure disorder; anorexia/bulimia (eating disorders), patients' use of antidepressant drugs (MAO inhibitors) within 14 days,

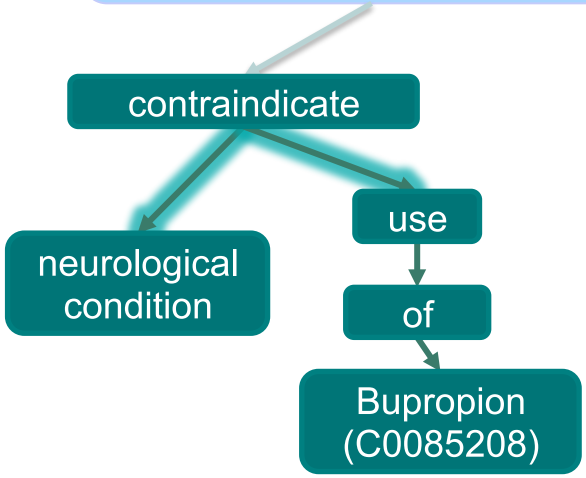


NLP Stack

Tokenize /Lemmatize
Named Entity Detection
Dependency Parsing
Coreference Detection
Negation Detection
Relation Detection
Frame Extraction
Topic Detection

Watson uses an ensemble of NLP techniques justify an answer

What neurological condition contraindicates the use of bupropion?



Watson considers…
Unstructured Content

Wellbutrin - noradrenergic antidepressant. contraindicated in adults with seizure disorders due to possible lowering of seizure threshold

Use background medical knowledge (**Wellbutrin** is a brand name of **bupropion**)

NLP Stack

- Tokenize /Lemmatize
- Named Entity Detection
- Dependency Parsing
- Coreference Detection
- Negation Detection
- Relation Detection
- Frame Extraction
- Topic Detection

Structured Content

UMLS

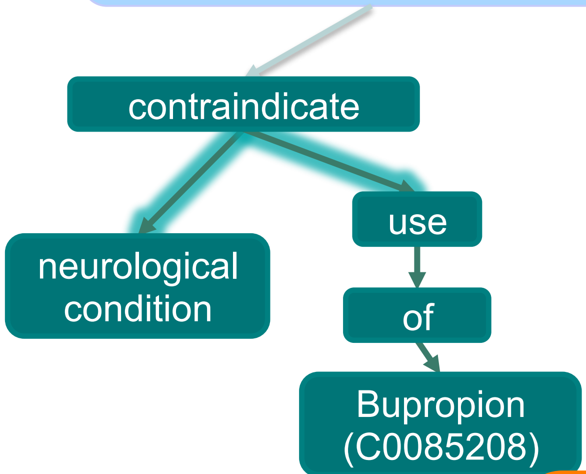
Prismatic mined KB

NLP Stack

- Tokenize /Lemmatize
- Named Entity Detection
- Dependency Parsing
- Coreference Detection
- Negation Detection
- Relation Detection
- Frame Extraction
- Topic Detection

Watson uses an ensemble of NLP techniques justify an answer

What neurological condition contraindicates the use of bupropion?



Patients with preexisting *seizure disorder* should not use bupropion due to a higher-than-proportional increase in the possibility of seizure as the dose is increased.

Consider paraphrases in medical language:
(should not use = contraindicate)

NLP Stack
- Tokenize /Lemmatize
- Named Entity Detection
- Dependency Parsing
- Coreference Detection
- Negation Detection
- Relation Detection
- Frame Extraction
- Topic Detection

Structured Content

UMLS

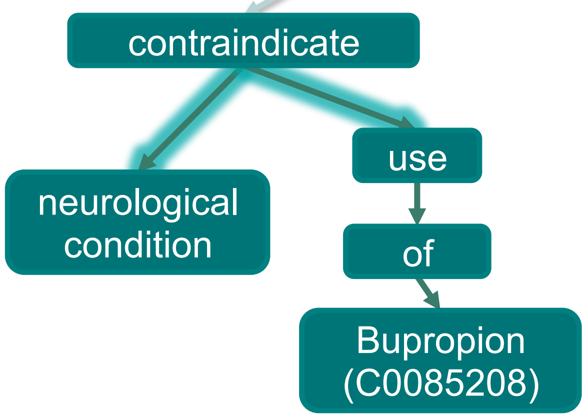
Prismatic mined KB

NLP Stack
- Tokenize /Lemmatize
- Named Entity Detection
- Dependency Parsing
- Coreference Detection
- Negation Detection
- Relation Detection
- Frame Extraction
- Topic Detection

Watson uses an ensemble of NLP techniques justify an answer

IBM WATSON

What neurological condition contraindicates the use of bupropion?



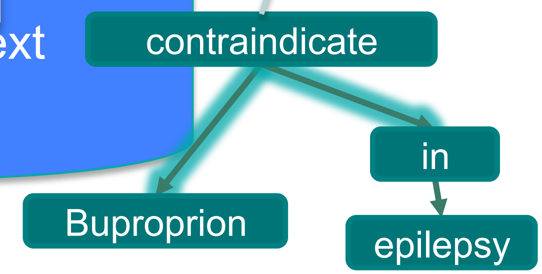
Watson considers…

Unstructured Content

Bupropion is contraindicated in epilepsy, seizure disorder; anorexia/bulimia (eating disorders), patients' use of antidepressant drugs (MAO inhibitors) within 14 days,

Wellbutrin - noradrenergic antidepressant. contraindicate in adults with seizure disorders due to possible lowering of seizure threshold

Patients with preexisting *seizure disorder* should not use bupropion due to a higher-than-proportional increase in the possibility of seizure as the dose is increased.



Matching Framework

NLP Stack

Tokenize /Lemmatize
Named Entity Detection
Dependency Parsing
Coreference Detection
Negation Detection
Relation Detection
Frame Extraction
Topic Detection

NLP Stack

Tokenize /Lemmatize
Named Entity Detection
Dependency Parsing
Coreference Detection
Negation Detection
Relation Detection
Frame Extraction
Topic Detection

Passage Justification

Neurological condition that contraindicates the use of bupropion

Patients with preexisting seizure disorder **should not use** bupropion due to a higher-than-proportional increase in the possibility of seizure as the dose is increased

The most common first manifestation of bladder cancer

Because patients typically present with symptoms suggestive of urinary tract infection, interstitial cystitis, or prostatitis, the diagnosis of bladder cancer is often delayed. The most common presenting symptom is hematuria, which is frequently macroscopic.

Attacks of Meniere's disease are precipitated by this dietary indiscretion

For treating Meniere's disease, treatment includes use of antibiotics and dietary changes. A low salt diet might also help in alleviating the symptoms of tinnitus and Meniere's disease to some extent

Drug treatment for nerve gas exposure that reactivates acetylcholinesterase

Some of the nerve agents attack and destroy acetylcholinesterase by phosphorylation, so the action of acetylcholine becomes prolonged, pralidoxime (2-PAM) is the cure for organophosphate poisoning because it can cleave this phosphorylation

Functional Adaptation

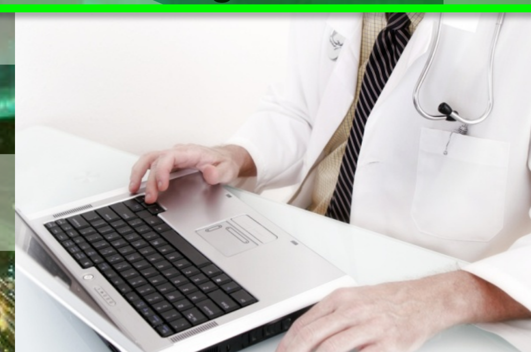
Semantic Analysis: Leveraging Background Knowledge

Which medications have been used for neuropathic pain for this patient?

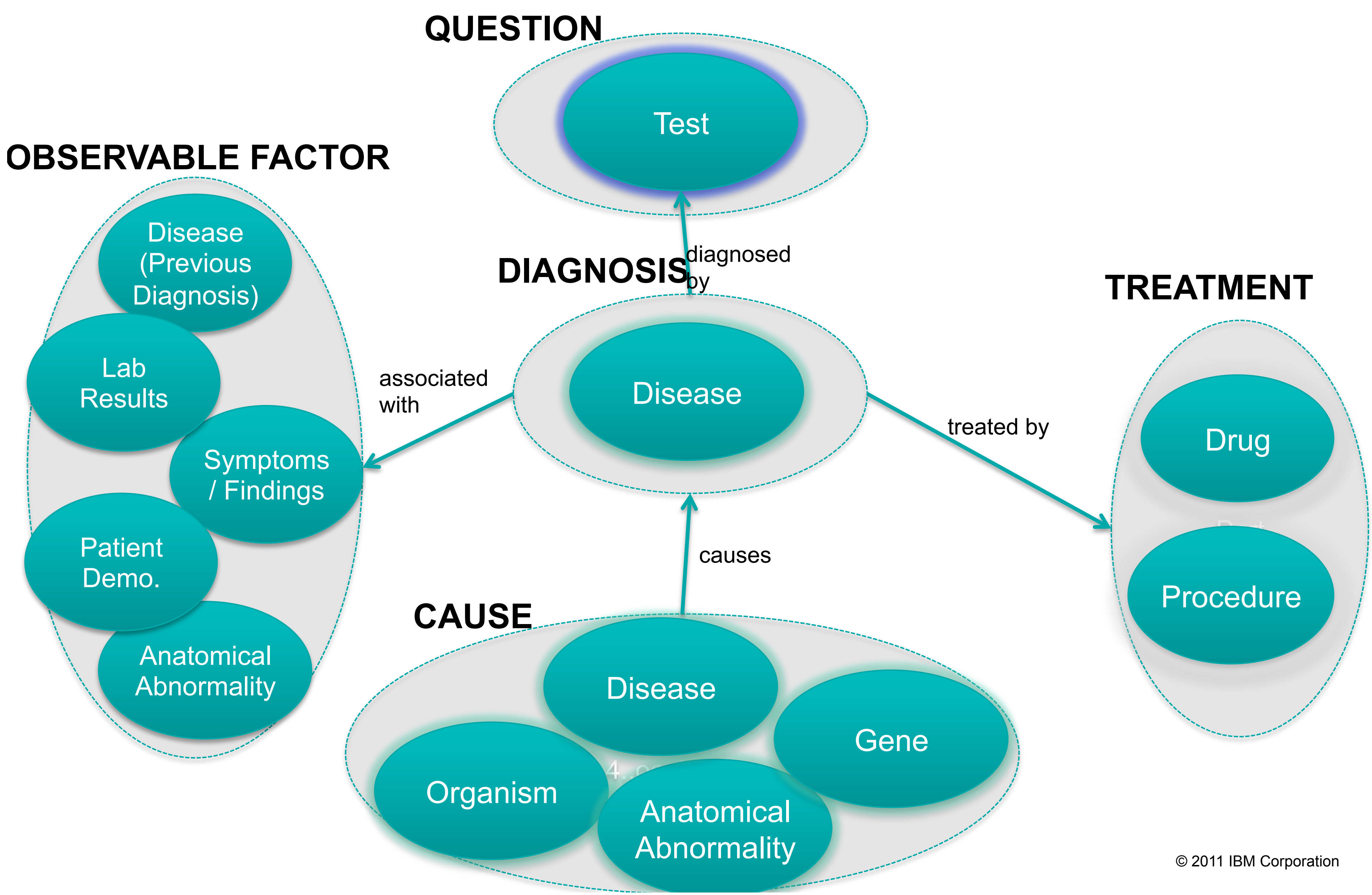
Information Extraction: what's there?

Semantic Analysis: how does it relate to what you need?

- Matching / Textual Entailment
- Background Knowledge



High Level Domain Model



Relation Inventory

Structured Content
UMLS

*6% of diseases have a **treatment** identified in UMLS*

Relation	KB	Count	IBM Automatic Extension
isa (hyponomy)	MeSH / SNOMED	1345388	Hyponymy rules (Prismatic)
partOf (meronymy)	UMLS	19436	TWREX
disease_has_finding	UMLS	17540	TWREX, SKB
may_treat	UMLS	43780	TWREX, SKB
causes	UMLS	3569	TWREX, SKB
prevented by	UMLS	5823	TWREX, SKB
diagnosed_by	UMLS	937	TWREX, SKB
location_of	UMLS	15277	TWREX, SKB
contraindicated	UMLS	71018	TWREX, SKB

Collecting *may_treat:* Running relation detection over NLM Medline article abstracts

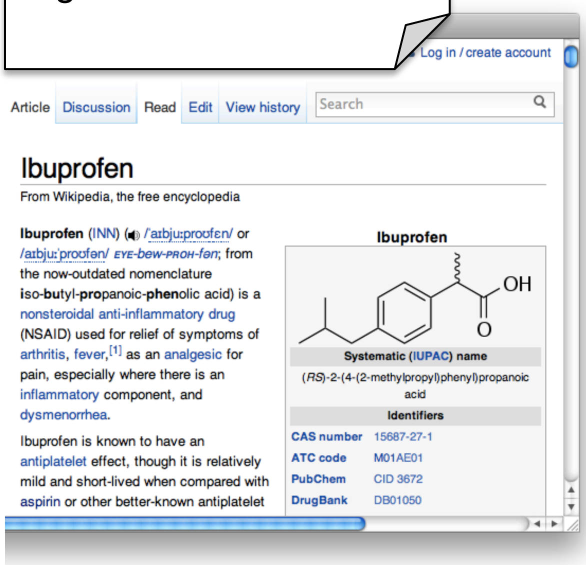
Extractions for **may_treat**

Fosinopril thus offers an effective and well tolerated option for the treatment of hypertension in adult and elderly patients, including those with renal or hepatic impairment		**may_treat(hypertension,Fosinopril)**
Fosinopril represents a clinically useful agent for the treatment of hypertension		**may_treat(hypertension,Fosinopril)**
He received treatment for this congestive cardiac failure and hypertension with enalapril, nifedipine and fursemide		**may_treat(hypertension,fursemide)**
These results suggest that guanabenz is safe and effective for initial and sole therapy of hypertension		**may_treat(hypertension,guanabenz)**
Since eight of the nine patients had mild hypertension, they may have responded maximally to the lower guanabenz doses, precluding larger decreases with the 24 and 32 mg doses		**may_treat(hypertension,guanabenz)**
In light of the efficacy without severe side effects, guanadrel may be an agent for step II therapy of hypertension		**may_treat(hypertension,guanadrel)**
Guanadrel was found to be an effective antihypertensive agent for all levels of hypertension		**may_treat(hypertension,Guanadrel)**
Our conclusion is that guanadrel is an effective, well-tolerated medication for treatment of hypertension in the elderly		**may_treat(hypertension,guanadrel)**

Inducing Meaning: is_a relations:

As with other NSAIDs, ibuprofen may be useful in the treatment of severe orthostatic hypotension

Lasix (furosemide), a diuretic, and ibuprofen, an NSAID, can be taken together



Rule-based relation detector identifies hyponymy relations in text

Frame01	
subj	Ibuprofen
type	NSAID

Frame02	
subj	Lasix
type	diuretic

Frame03	
subj	NSAID
type	drug

Ibuprofen isa **NSAID**

Lasix isa **diuretic**

NSAID isa drug

.

.

Tycor / Typing

	LAT	Answer
Patient initiated maneuver to differentiate thyroid from neck mass	maneuver	Swallowing
Relative leukocyte alkaline phosphatase level in leukomoid reaction	level	High
Treatment strategy responsible for reducing the incidence of HIV-associated Kaposi's sarcoma	strategy	HAART
The most common 2 sites for lumbar disk herniations	sites	L5-S1
Most common anatomical location of Wegener's granulomatosis	location	Upper Airway
Changes in LDL concentration in patients treated for subclinical hypothyroidism (normal T4, elevated TSH)	changes	Reduction in LDL / Improved LDL
Most common manifestation of Multiple Endocrine Neoplasia type 1 (MEN-1)	manifestation	hyperparathyroidism

LSA 2.0: Factors related to suicide

- UMLS-enhanced LSA can identify related factors

"Suicide"

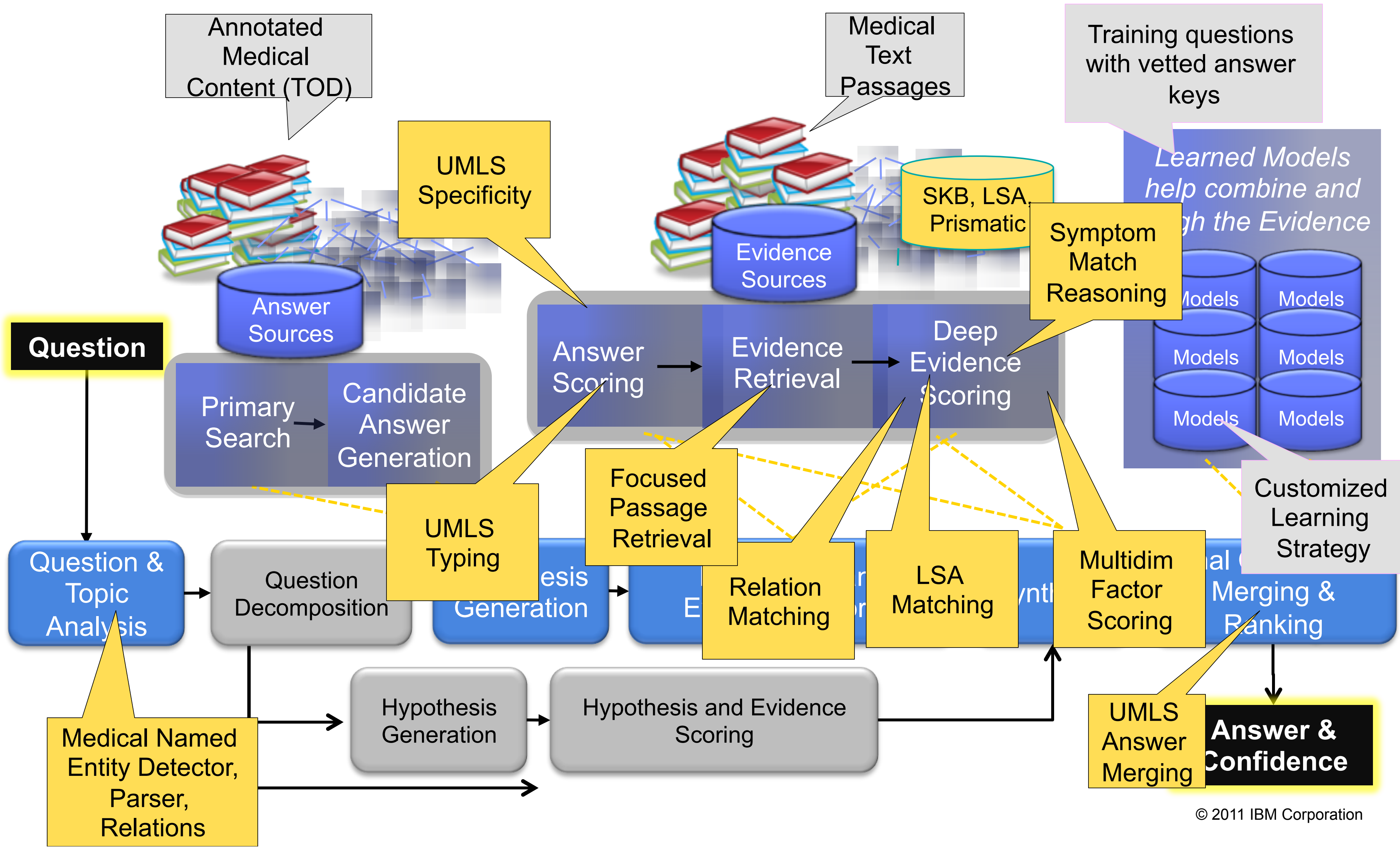
DANGER OF HARM TO SELF	0.94843552	Feeling hopeless	0.69763276
Depressive Symptoms	0.85787663	CYCLOTHYMIC REACTION	0.6956163
marked mood shift	0.83171128	Mental health counselor	0.6916423
loss of interest in activity	0.83171128	Demoralization	0.68469489
Other mood affective disorders	0.80852182	Ability to maintain self-esteem	0.67854127
Mood Disorders	0.80852182	Normal mood	0.67817024
Bipolar affective disorder, current episode manic	0.79134531	Despondency	0.67736145
Depressive disorder NEC in SNOMEDCT	0.78274978	Other and unspecified episodic mood disorder	0.67540516
change in self-esteem	0.77332301	Loss of interest	0.67413379
(Depression: [episode, unspecified] or [NOS (&	0.76803559	Suicidal	0.67144792
Self Esteem	0.72473412	pleasurable emotion	0.67024476
self-esteem as an AODC	0.7247341	Mood (psychological function)	0.67023923
AODE on self-esteem	0.7247341	Mood:-:Point in time:^Patient:-	0.66983514
		Suicidal behavior	0.6680896
		Adjustment disorder with depressed mood	0.6555974
		Depression aggravated	0.6528632
		Coping with Chronic Illness Topics	0.64542571
		Mental Health and Behavior	0.6454257
		Recurrent depression	0.64434724
		Other specified episodic mood disorder	0.64310002
		Melancholia	0.64063775
		Mild recurrent major depression	0.63696897

Learning NLP Axioms and Common Sense Knowledge

Question Text	Passage Text	Learned Axiom
Murmur associated with this condition is harsh, systolic, diamond-shaped, and increases in intensity with Valsalva	A systolic murmur that increases with the valsalva maneuver and disappears with squatting suggests hypertrophic cardiomyopathy	X suggests Y => X associated with Y
Class of drugs causing regression of polyposis in familial adenomatous polyposis	NSAIDs have been shown to induce adenoma regression in patients with familial adenomatous polyposis	X has been shown to induce Y => X causes Y
Intravenous treatment for cyanide poisoning	Antidotes for cyanide poisoning include amyl nitrate, sodium nitrate, and intravenous sodium thiosulfate.	Antidotes for X include Y => Y is treatment for X
Syndrome characterized by narrowing of the extra-hepatic bile duct from mechanical compression by a gallstone impacted in the cystic duct	Mirizzi's syndrome, a rare condition in which a gallstone impacting the cystic duct obstructs the common bile duct by edema and extrinsic compression	X obstructs Y => narrowing of Y by X
Preferred corrective treatment for acute episodes of angioedema in patients with hereditary angioedema	For acute episodes of angioedema in hereditary angioedema, adminster intravenous, purified, nanofiltered C1-INH concentrate as first-line therapy	For X, administer Y as first-line therapy => Y is preferred treatment for X

DeepQA Functional Adaptation For Medicine

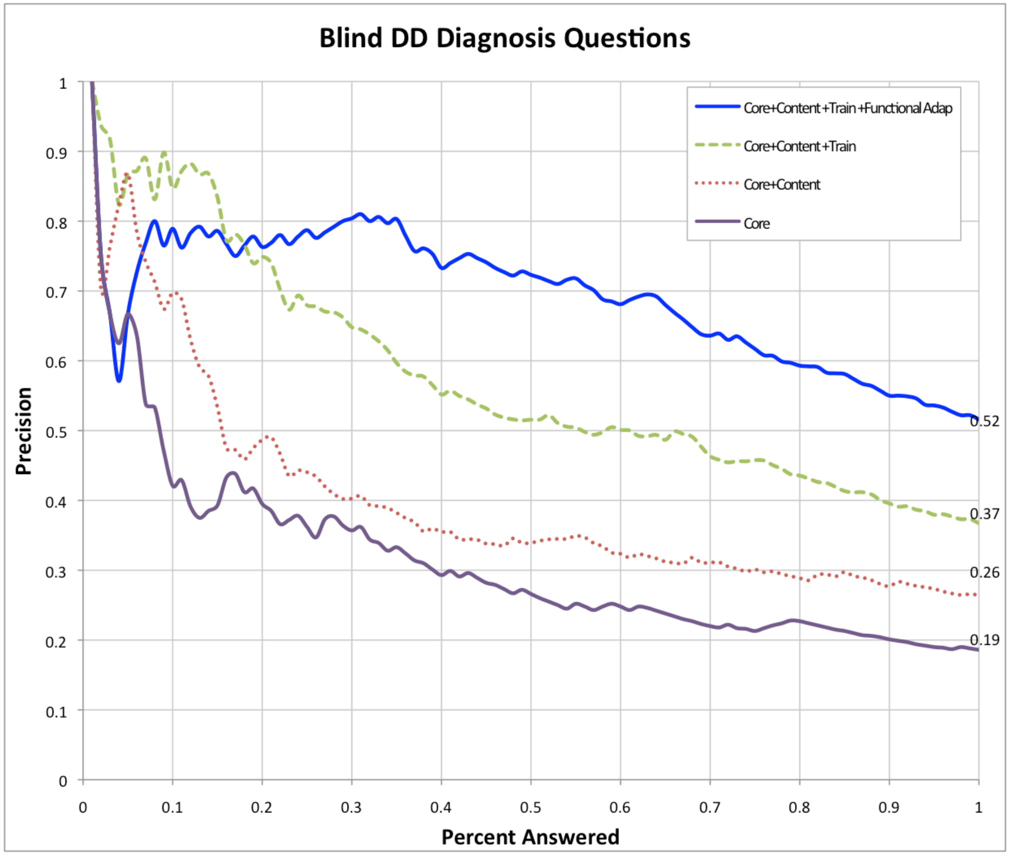
Content Adaptation → Training Adaptation → Functional Adaptation



Medical Adaptation – Functional Adaptation

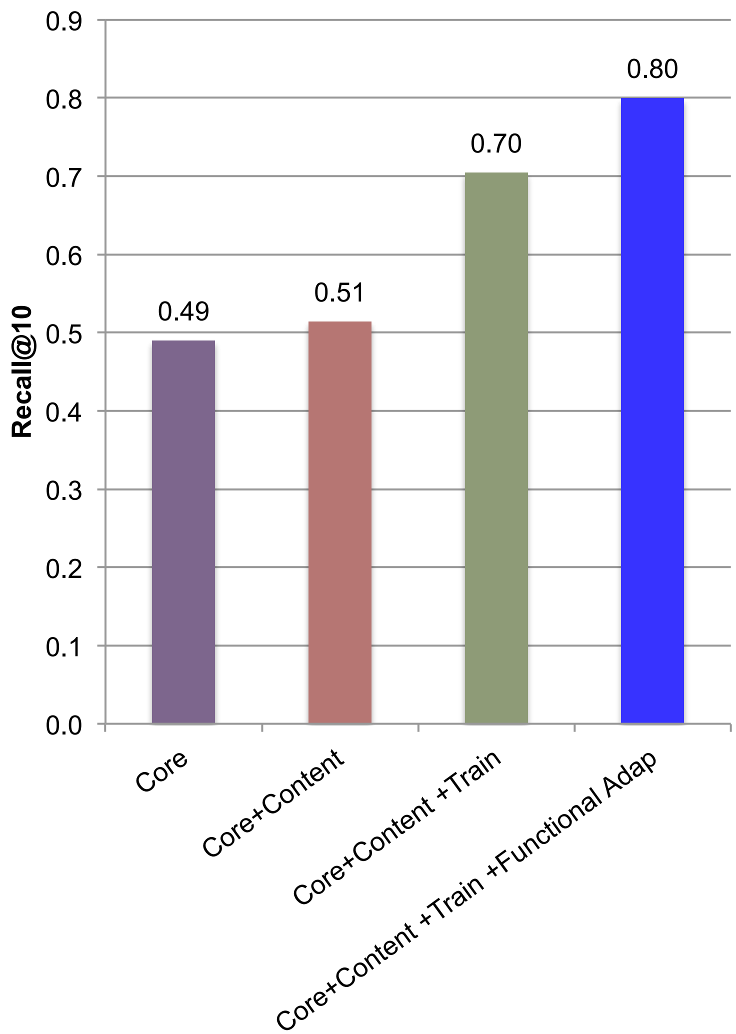
- NLP stack
  - Named Entity Detection
  - Relation Detection
  - Coreference
- Medical Reasoning
  - Symptom matching / contradiction
  - Negation
  - Anatomical Reasoning
- Mined Resources
  - Prismatic linguistic patterns
  - Symptom Association KB
  - LSA

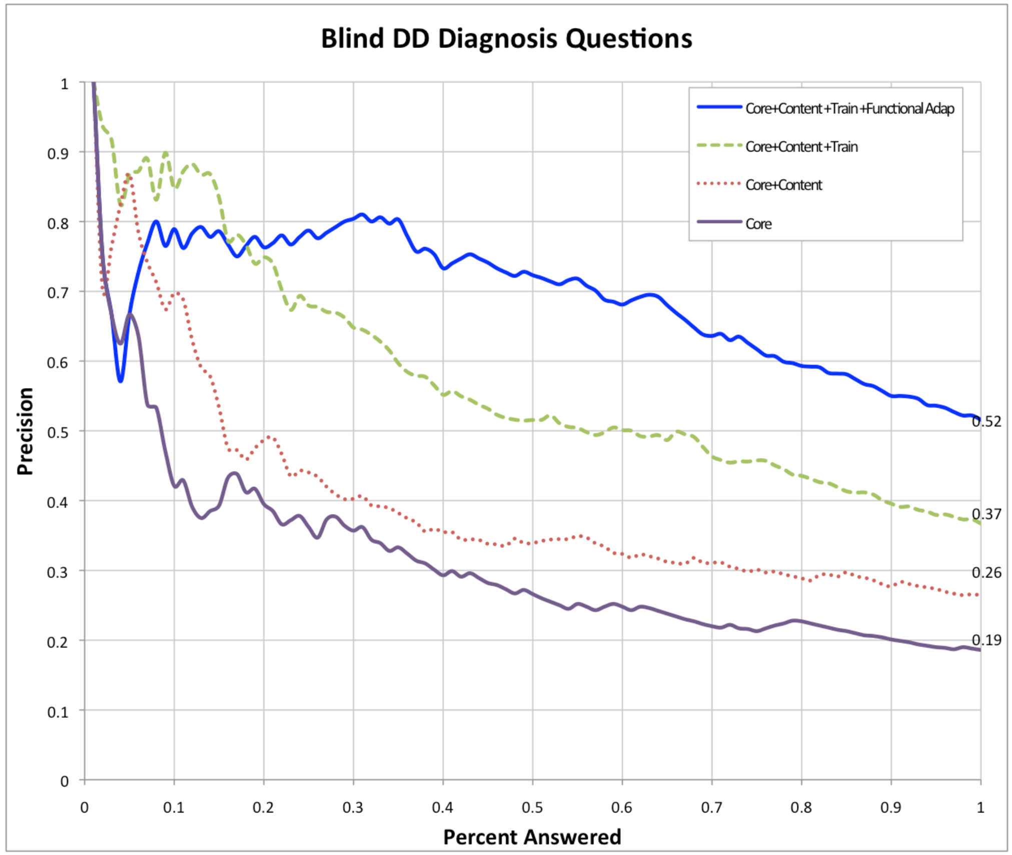
Accuracy: 52%



Medical Adaptation - Results

Accuracy: 52%





Domain Adaptation: lesson learnt

- Functional Adaptation is difficult!

- Requires:
  - Deeply skilled research team across all the key disciplines (ML, NLP, IR, KR)
  - Domain Experts (Doctors) for annotation/vetting and design reasoning strategies
  - Collaboration between the two groups !
  - Background knowledge for new domains (e.g. UMLS) and analytics exploiting that
  - Rigorous methodological discipline (e.g. blind test!)
  - …
- Future Challenge: Scalable and cost effective functional adaptation process
  - Unsupervised Domain Adaptation Process for
    - Semantic Search (Topic Models – LSA, query expansion by distributional thesauri)
    - Tycor (Taxonomy induction)
    - Passage Scoring (Statistical Paraphrasing)
    - EDM/PDM(Sense Induction/ Unsupervised WSD)
    - …
  - Using the same analytics across domains

- We call it Distributional Semantics!

Outline

- Domain Adaptation Methodology
- Content Adaptation
- Training Adaptation
- Functional Adaptation

Semantic Technologies in IBM Watson™

Lesson 10 – Distributional Semantics

Professor: Alfio Massimiliano Gliozzo
TA: Or Biran

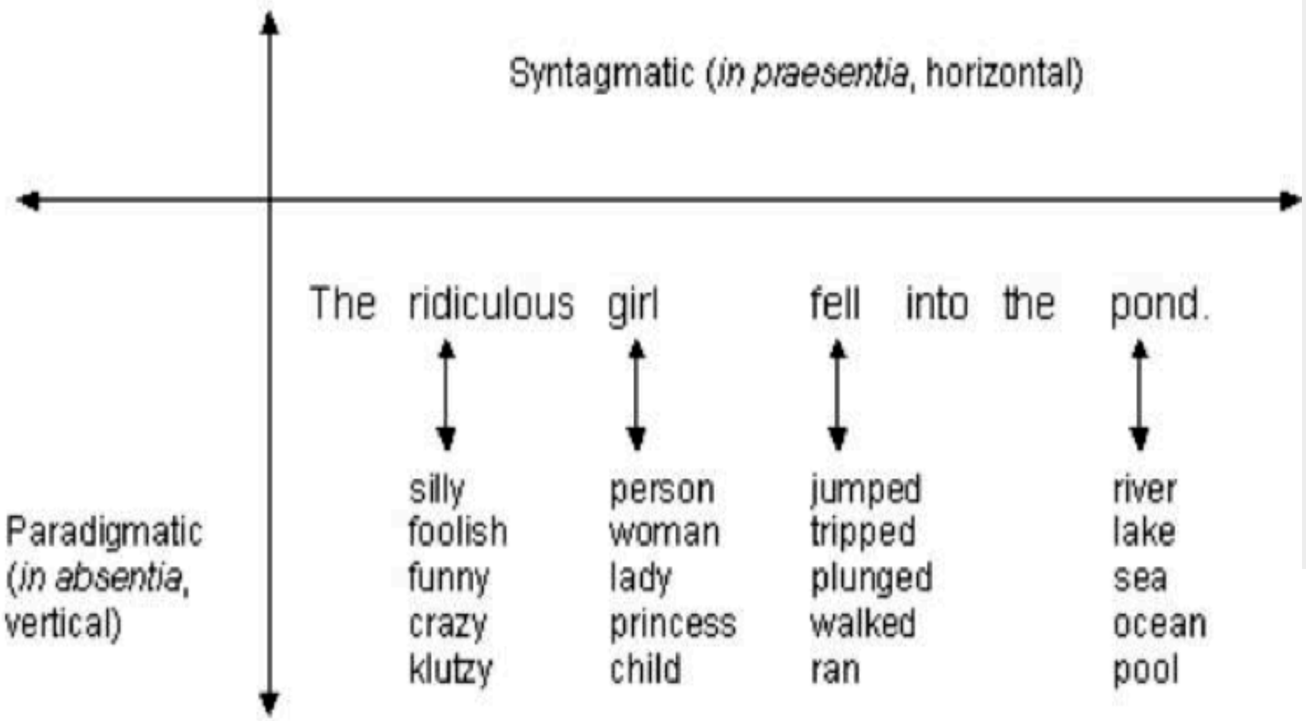
Outline

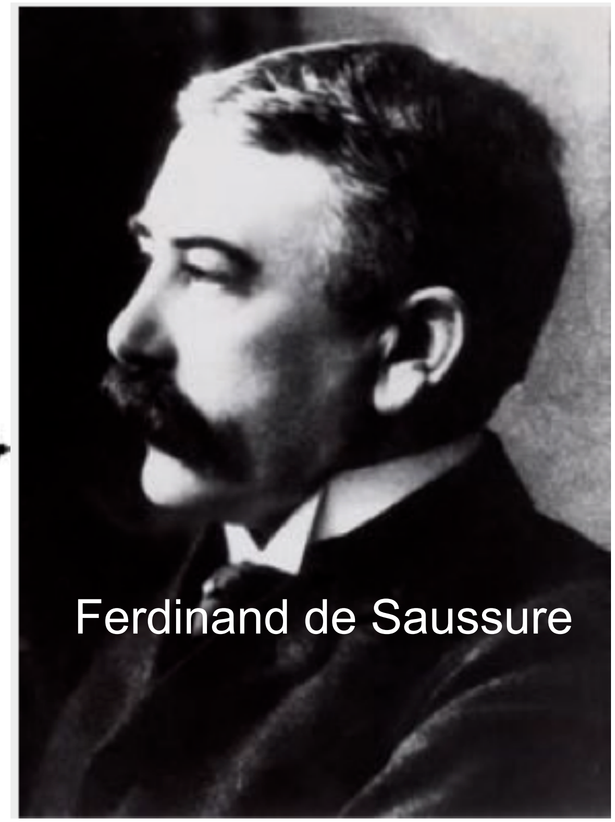
- Semantic Domains and LSA
- Semantic Domains in NLP
- LSA in Watson
- Distributional Semantics
- Inducing Meaning and Linking to Knowledge
- JoBimText and its application to Watson

Outline

- Semantic Domains and LSA
- Semantic Domains in NLP
- LSA in Watson
- Distributional Semantics
- Inducing Meaning and Linking to Knowledge
- JoBimText and its application to Watson

Syntagmatic vs. Paradigmatic Relations



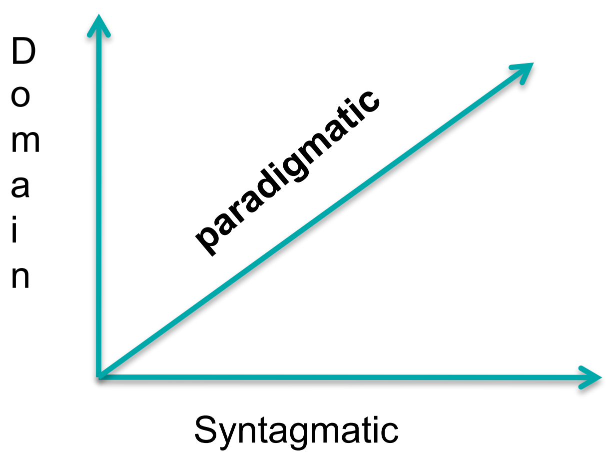


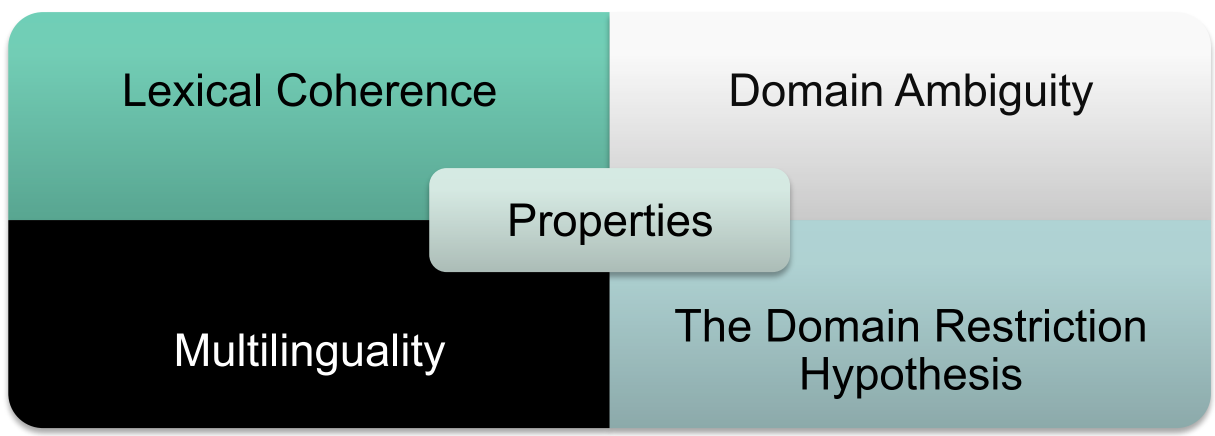
http://courses.nus.edu.sg/course/elltankw/history/Vocab/B.htm

- Syntagmatic Relations: syntactic constraints in the context
- Paradigmatic Relations: associations, semantic constraints

Semantic Domains

"a set of <u>concepts</u> that covers a whole spectrum of phenomena, a <u>domain</u>" (Trier, 1931)





Domain Ambiguity
(Magnini, Strapparava, Pezzulo, Gliozzo, JNLE 2001)

Chair

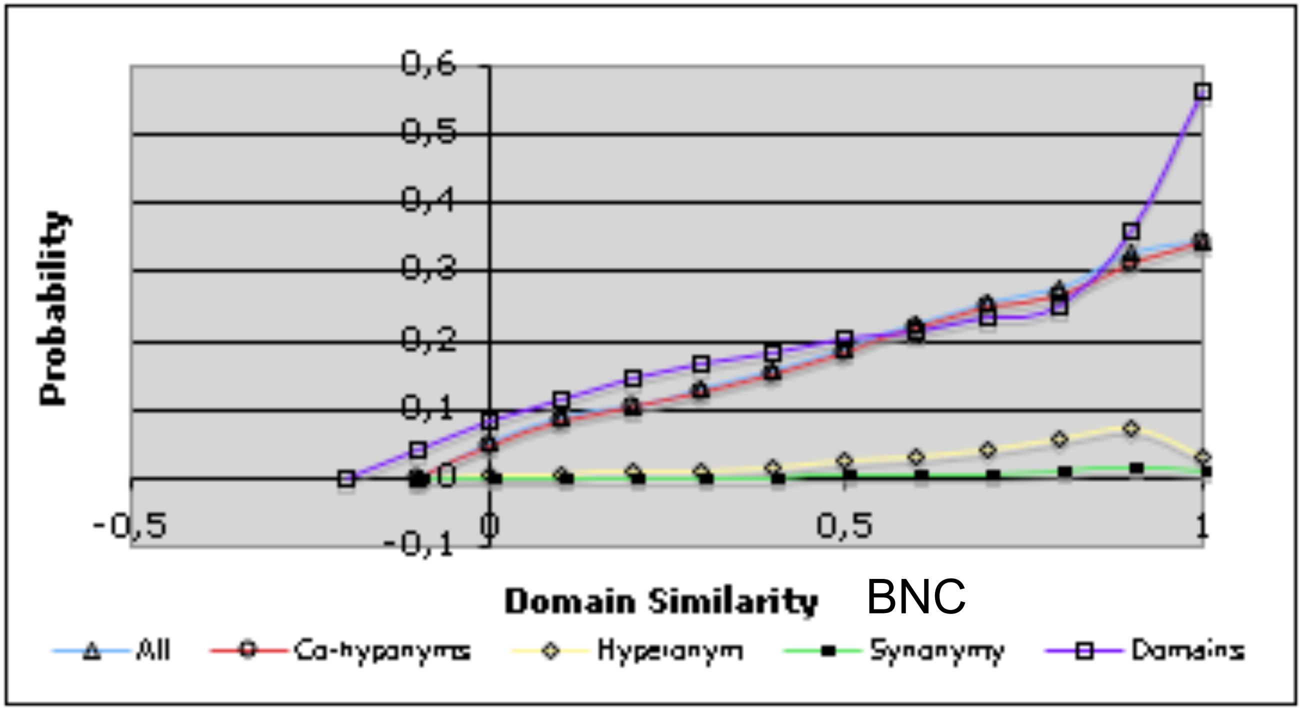
1. [**Furniture**] chair -- (a seat for one person)
2. [**University**] professorship, chair -- (the position of professor)
3. [**Administration**] president, chairman, chairwoman, chair, chairperson
4. [**Law**] electric chair, chair, death chair, hot seat

Source: WordNet Domains (Magnini and Cavaglià, 1999)

The Domain Restriction Hypothesis

Semantic Relations hold mainly among terms in the same domain, while those belonging to different domains are unrelated

W o r d n e t



BNC

e.g. laptop - PC (ok) soccer – classical music (bad)

Associations among terms and Semantic Domains

	MEDICINE	COMPUTER_SCIENCE
HIV	1	0
AIDS	1	0
virus	0.5	0.5
laptop	0	1

(domain-)Ambiguity : a term belongs to many domains
(domain-)Variability: words in the same domain are semantically related

distributional hypothesis

Words with similar meanings tend to occur in similar context

Latent Semantic Analysis: Acquiring Domain Models by SVD

IBM WATSON

M =

	d_1	d_2	d_3	d_4	d_5	d_6
shuttle	1	0	1	0	0	0
astronaut	0	1	0	0	0	0
moon	1	1	0	0	0	0
car	1	0	0	1	1	0
truck	0	0	0	1	0	1

$$\mathbf{M} = \mathbf{K}_{t \times s} \mathbf{S}_{s \times s} \mathbf{D}^T_{s \times N}$$

concepts; terms: txs X sxs X documents; concepts: sxN

K =

	dim_1	dim_2	dim_3	dim_4	dim_5
shuttle	-0.44	-0.30	0.57	0.58	0.25
astronaut	-0.13	-0.33	-0.59	0.00	0.73
moon	-0.48	-0.51	-0.37	0.00	-0.61
car	-0.70	0.35	0.15	-0.58	0.16
truck	-.0.26	0.65	-0.41	0.58	-0.09

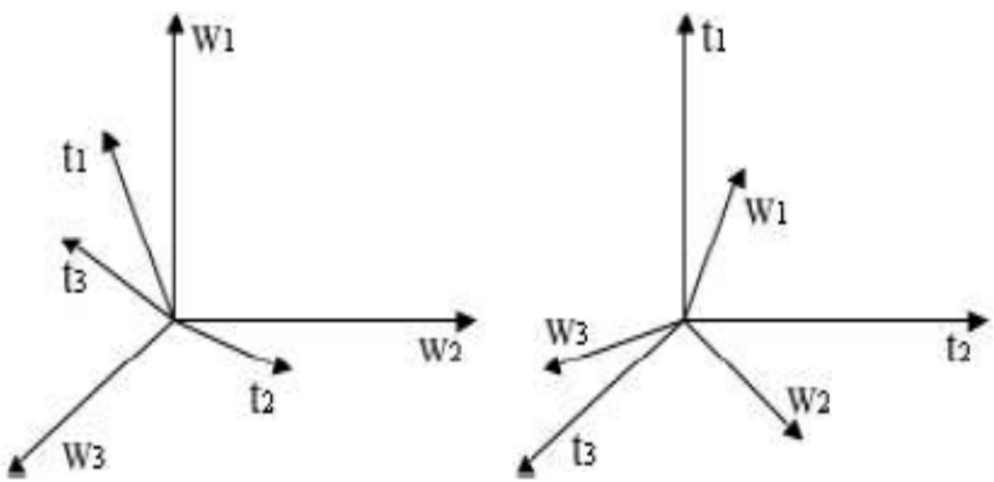
S =

2.16	0.00	0.00	0.00	0.00
0.00	1.59	0.00	0.00	0.00
0.00	0.00	1.28	0.00	0.00
0.00	0.00	0.00	1.00	0.00
0.00	0.00	0.00	0.00	0.39

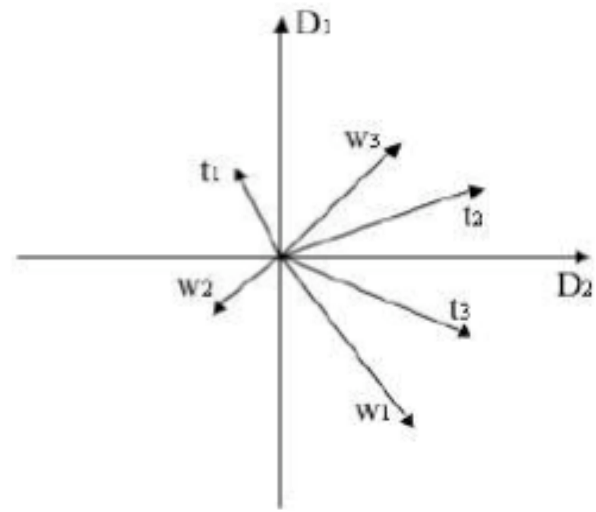
Term Vectors are mapped into a lower dimensional Space

Alternative dimensionality reduction: Random Indexing, LDA

Similarity in the Domain Space: terms and texts are projected into the same space



Classical VSMs

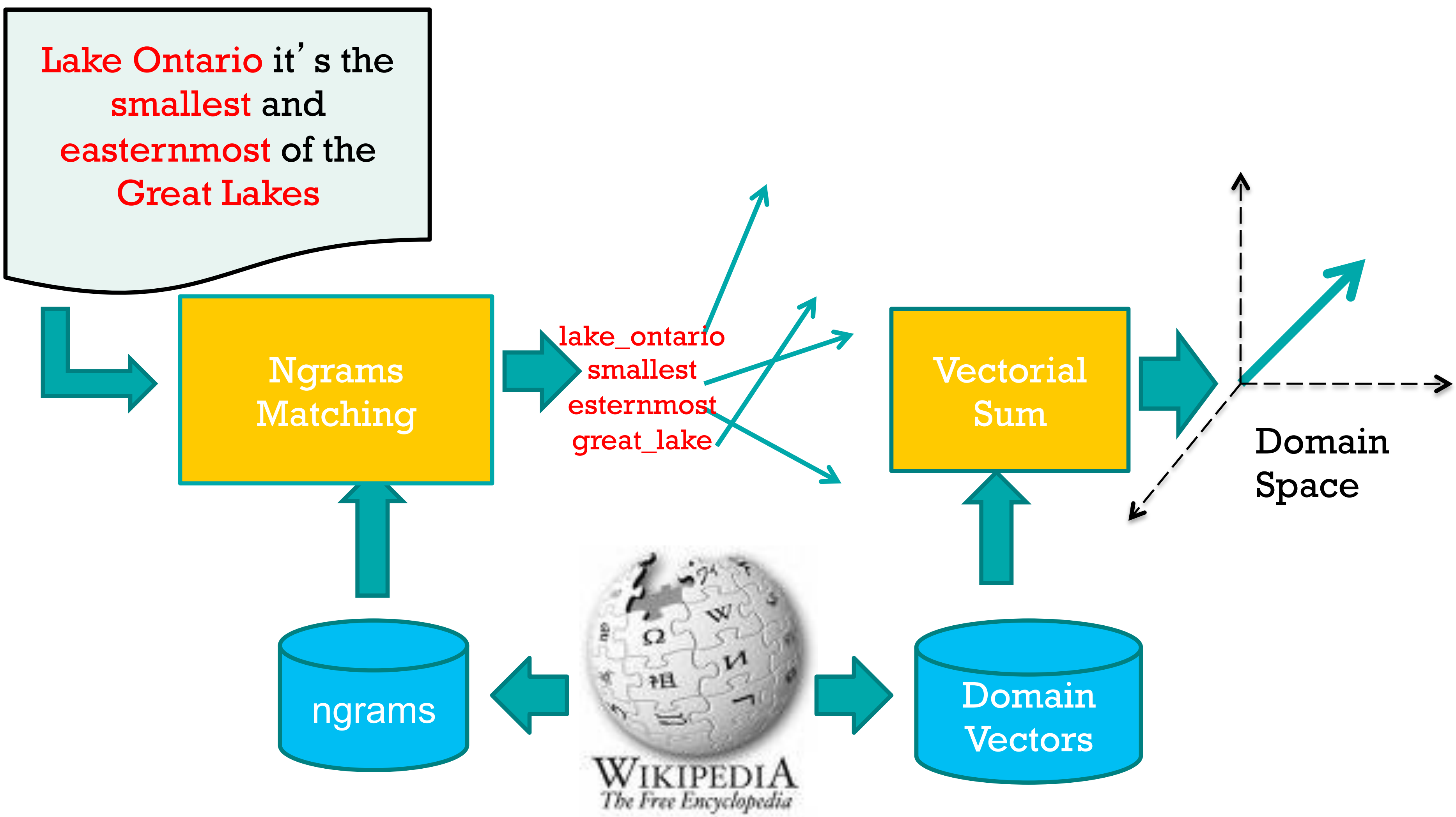


Domain Space

Similarity among terms and documents is estimated by cosine
(Dot product of unitary vectors is more efficient)

Representing text by Pseudo Vectors

IBM WATSON



Outline

- Semantic Domains and LSA
- Semantic Domains in NLP
- LSA in Watson
- Distributional Semantics
- Inducing Meaning and Linking to Knowledge
- JoBimText and its application to Watson

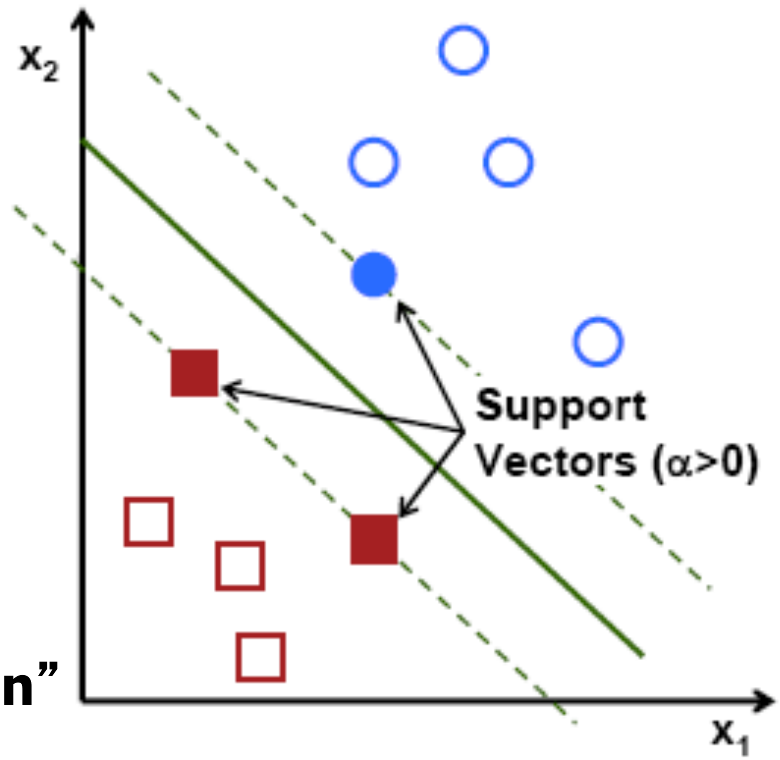
The Domain Kernel

IBM WATSON

Support Vector Machines

Similarity with support vectors is evaluated by the kernel function

$$f(x) = \sum_{i=1}^{T} y_i \lambda_i F(x_i) \cdot F(x) + \lambda_0 = $$
$$= \sum_{i=1}^{T} y_i \lambda_i K(x_i, x) + \lambda_0$$



external knowledge can be "plugged in"

$$\mathcal{D}(\vec{t_j}) = \vec{t_j}(\mathbf{I}^{\mathbf{IDF}}\mathbf{D}) = \vec{t'_j}$$

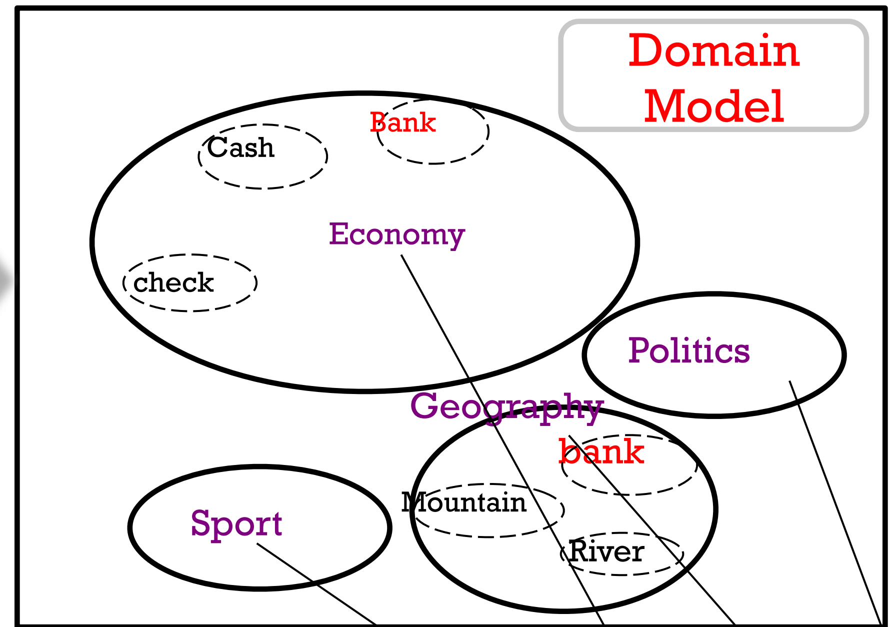
$$\mathbf{D} =$$

	MEDICINE	COMPUTER_SCIENCE
HIV	1	0
AIDS	1	0
virus	0.5	0.5
laptop	0	1

Domain Models for Semi Supervised Learning

IBM WATSON

Large Corpus → Dimensionality reduction →



Unsupervised Learning

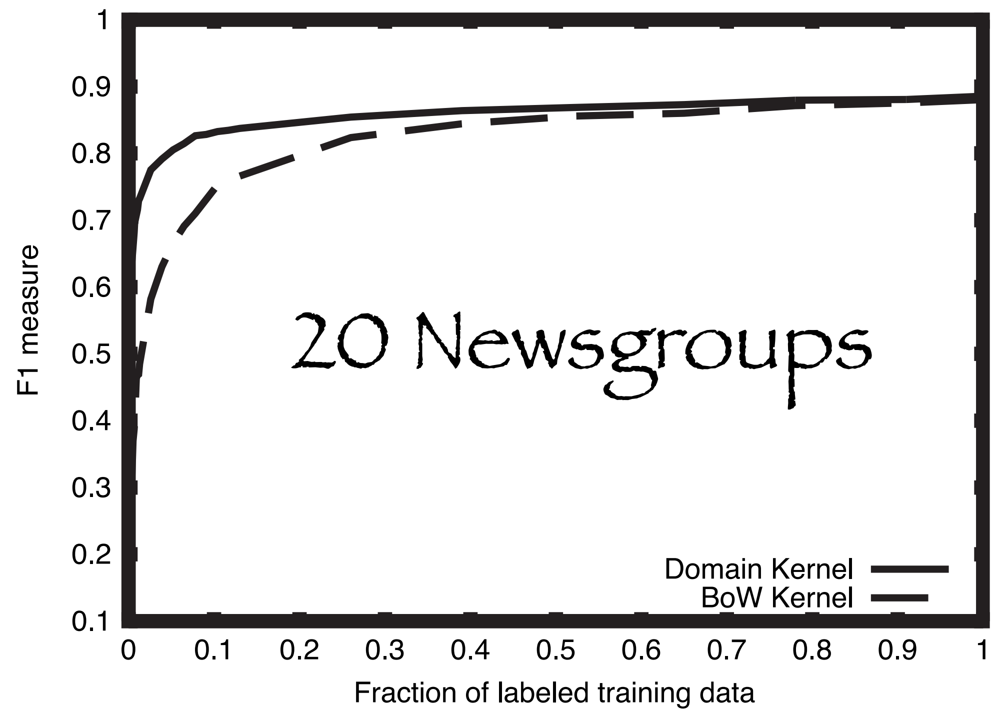
Supervised Learning

Labeled text Small Corpus → Feature Extraction (Pseudo Vectors) →

	SPO	ECO	GEO	POL
Text_1	0,3	0,1	0,001	0
Text_2	0,002	0,2	0.01	0.5
…	…	…	…	…
Text_n	0,27	0,02	0,009	0,005

Domain Kernels for Text Categorization
(Gliozzo and Strapparava, CoNLL 2005)

- Supervised TC
  - Labeled documents for each category are provided
- Semi Supervised Learning
  - Usupervised learning from the full training set
  - Supervised learning from labeled data
- Support Vector Machines



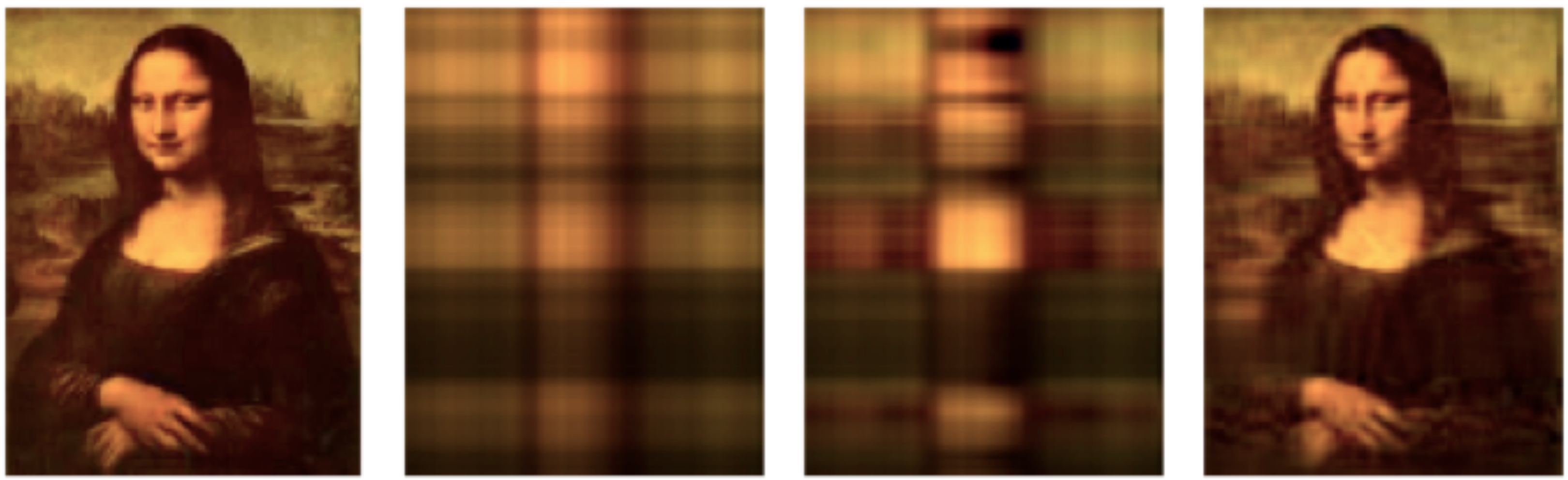
F1	*Domain Kernel*	*Bow Kernel*	*Ratio*
.54	**14**	267	5%
.84	**146**	1380	10%
.90	**668**	6680	10%

Table 3: Number of training examples needed by K_D and K_{BoW} to reach the same micro-F1 on the *Reuters* task

F1	*Domain Kernel*	*Bow Kernel*	*Ratio*
.50	**30**	500	6%
.70	**98**	1182	8%
.85	**2272**	7879	29%

Table 4: Number of training examples needed by K_D and K_{BoW} to reach the same micro-F1 on the *20newsgroups* task

Dimensionality reduction and learning theory



The original | 1 singular value | 2 singular values | 16 singular values

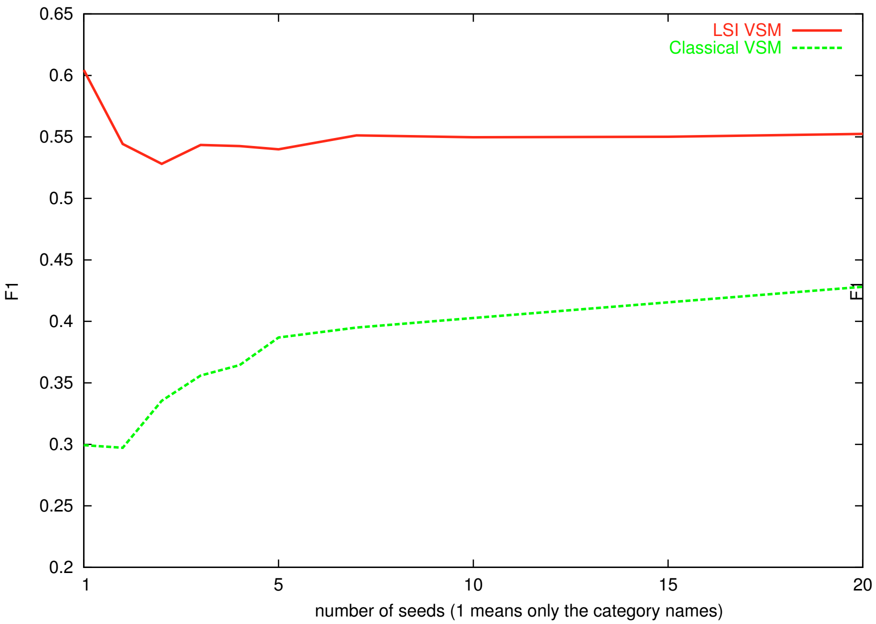
$$m > \frac{1}{\epsilon}(4\log_2 \frac{2}{\delta} + 8VC(H)\log_2 \frac{13}{\epsilon})$$

Text Categorization Bootstrapping
(Gliozzo, Strapparava, Dagan, ACM TSLP 6(1), 2009)

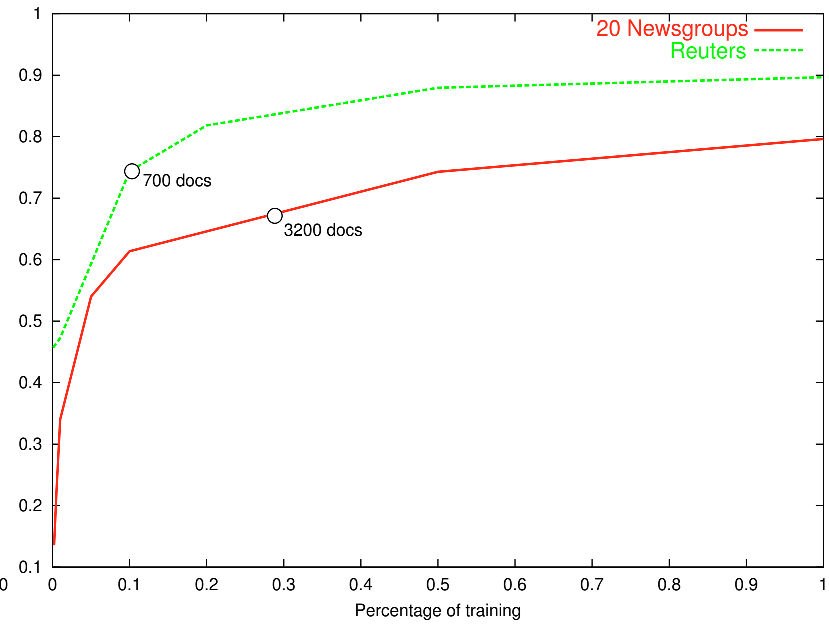
- Categories are described by seed terms (only category names) instead of labeled documents.
  – Step 1: Sets of documents for any category are retrieved
  – Step 2: Training on Docs from step 1

20newsgroups		*Reuters-10*	
Categories	*Seeds*	*Categories*	*Seeds*
alt.atheism	atheism#n	acq	acquisition#n
comp.graphics	graphics#n	corn	corn#n
comp.os.ms-windows.misc	microsoft#n windows#n	crude	crude#n
comp.sys.ibm.pc.hardware	ibm#n pc#n	earn	earn#v
comp.sys.mac.hardware	mac#n	grain	grain#n
comp.windows.x	x-windows#n	interest	interest#n
misc.forsale	sale#n	money-fx	money#n
rec.autos	car#n	ship	ship#n
rec.motorcycles	motorcycle#n	trade	trade#n
rec.sport.baseball	baseball#n	wheat	wheat#n
rec.sport.hockey	hockey#n		
sci.crypt	cryptography#n		
sci.electronics	electronics#n		
sci.med	medicine#n		
sci.space	space#n		
soc.religion.christian	christian#n christian#a		
talk.politics.guns	gun#n		
talk.politics.mideast	mideast#n		
talk.politics.misc	politics#n		
talk.religion.misc	religion#n		

Bootstrapp



Comparison with Supervised



	Our	*IDs per cat.*	*Liu et al.*	*IDs per cat.*
REC	0.94	1	0.95	5
TALK	0.80	1	0.80	20
SCI	0.92	1	0.93	20
COMP	0.81	1	0.71	15

Domain Kernels for WSD

IBM WATSON

(Gliozzo, Giuliano, Strapparava, ACL 2005)

- Syntagmatic Kernel
  - the bank is on the corner of Nassau and Witherspoon [building]
  - that bank holds the mortgage o my home [institution]
  - Word Sequence Kernels: evaluate similarity by counting the common(sparse) subsequences of words/POS in a local window
  - l("I go very quickly to the old school")> l("He goes to school")

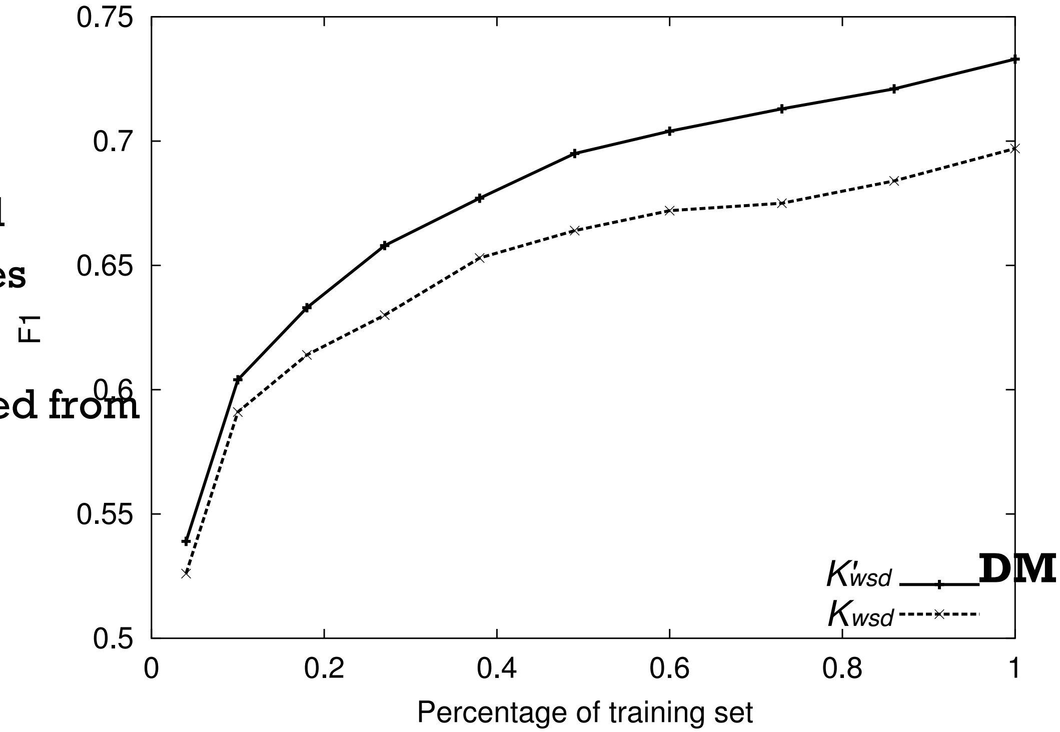
$$K_n(s,t) = \sum_{u \in \Sigma^n} \sum_{\mathbf{i}:s[\mathbf{i}]=u} \sum_{\mathbf{j}:t[\mathbf{j}]=u} \lambda^{l(\mathbf{i})+l(\mathbf{j})}$$

- Domain Kernel
  - LSA on BNC

$$K_C(x_i, x_j) = \sum_{l=1}^{n} \frac{K_l(x_i, x_j)}{\sqrt{K_l(x_j, x_j) K_l(x_i, x_i)}}$$

- Kernel Combination

- Lexical Sample Task:
- Word Expert approach
  - One classifier per word
  - Word Senses are classes
- Semi supervised:
  - Domain Models acquired from unlabeled data
  - BNC for English



	#w	pol	# train	# test	# unlal
Catalan	27	3.11	4469	2253	23935
English	57	6.47	7860	3944	-
Italian	45	6.30	5145	2439	74788
Spanish	46	3.30	8430	4195	61252

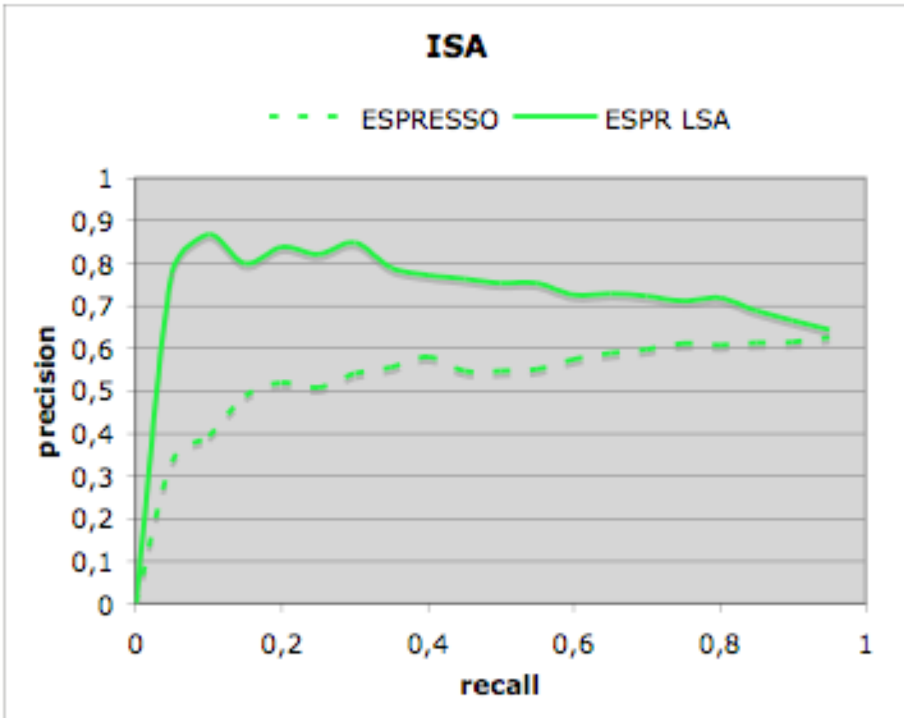
	MF	Agreement	BEST	K_{wsd}	K'_{wsd}	DM+	BEST+
English	55.2	67.3	72.9	69.7	**73.3**	3.6	0.4
Catalan	66.3	93.1	85.2	85.2	**89.0**	3.8	3.8
Italian	18.0	89.0	53.1	53.1	**61.3**	8.2	8.2
Spanish	67.7	85.3	84.2	84.2	**88.2**	4.0	4.0

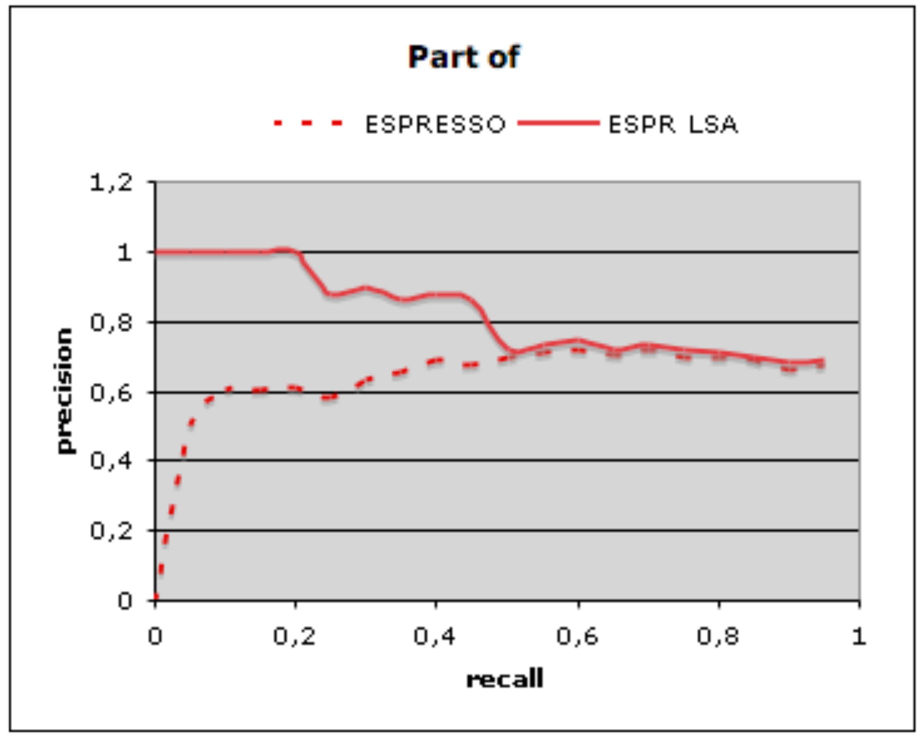
Domain Driven Relation Extraction
(Gliozzo, Pennacchiotti, Pantel, HLT/NAACL 2007)

- Re-ranking pattern based relation extraction system domain similarity of extracted pairs
- We used ESPRESSO (Pantel and Pennacchiotti,ACL 2006)
- Benchmark: college chemistry textbook (Brown et al. 2003)

Relation	ESP-	ESP - LSA
X is-a Y	Aluminum ; metal nitride_ion ; strong_Br heat_flow ; calorimeter complete_ionic_equation ; spectator	F ; electronegative_atoms O ; electronegative_atoms NaCN ; cyanide_salt NaCN ; cyanide_salts
X part-of Y	elements ; compound composition ; substance blocks ; tripeptide elements ; sodium_chloride	amino_acid_building_blocks ; tripeptide acid_building_blocks ; tripeptide powdered_zinc_metal ; battery building_blocks ; tripeptide
X react Y	hydrazine ; water magnesium_metal ; hydrochloric_acid magnesium ; oxygen magnesium_metal ; acid	magnesium_metal ; elemental_oxygen nitrogen ; ammonia sodium_metal ; chloride carbon_dioxide ; methane
X produce Y	bromine ; bromide oxygen ; oxide common_fuels ; dioxide kidneys ; stones	high_voltage ; voltage reactions ; reactions dr_jekyll ; hyde yellow_pigments ; green_pigment

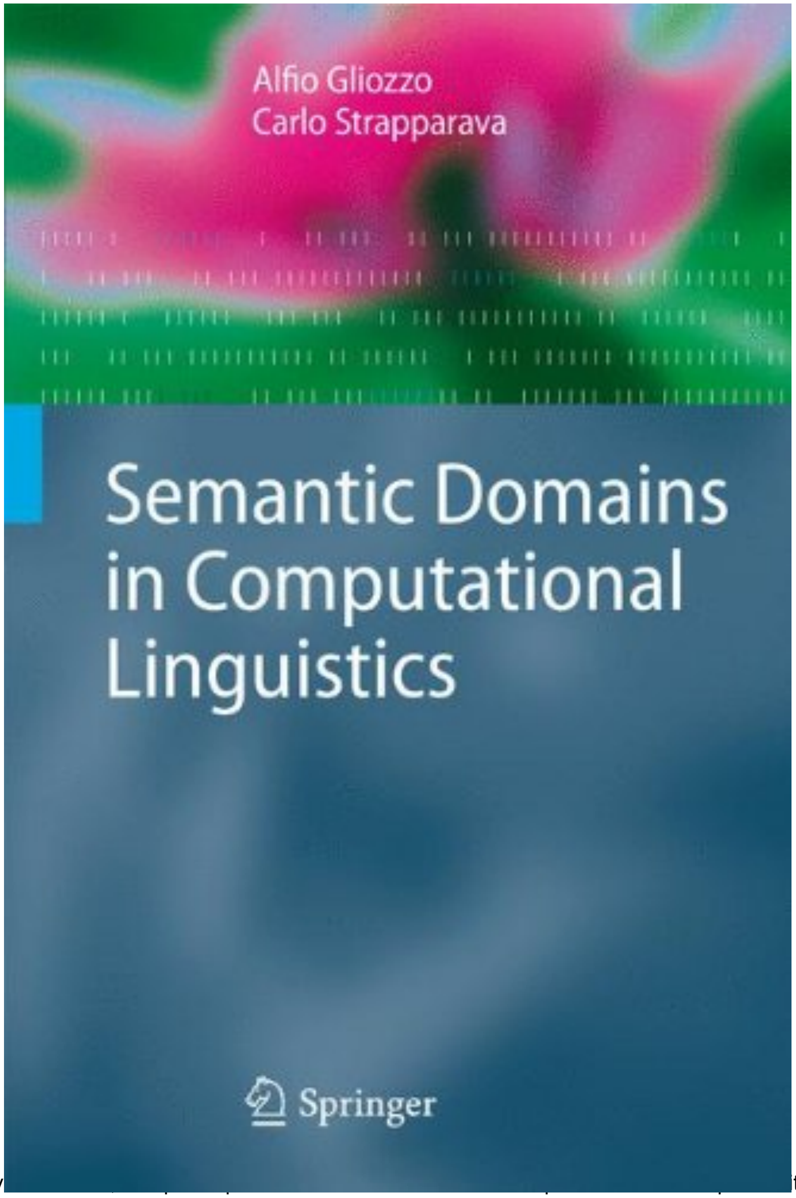
Domain Driven Relation Extraction
(Gliozzo, Pennacchiotti, Pantel, HLT/NAACL 2007)





	ESP-	ESP-LSA	
is-a	0.54	**0.75**	(+0.21)
part-of	0.65	**0.82**	(+0.17)
react	0.75	**0.82**	(+0.07)
produce	0.55	**0.62**	(+0.07)

References


Alfio Gliozzo
Carlo Strapparava
Semantic Domains in Computational Linguistics
Springer

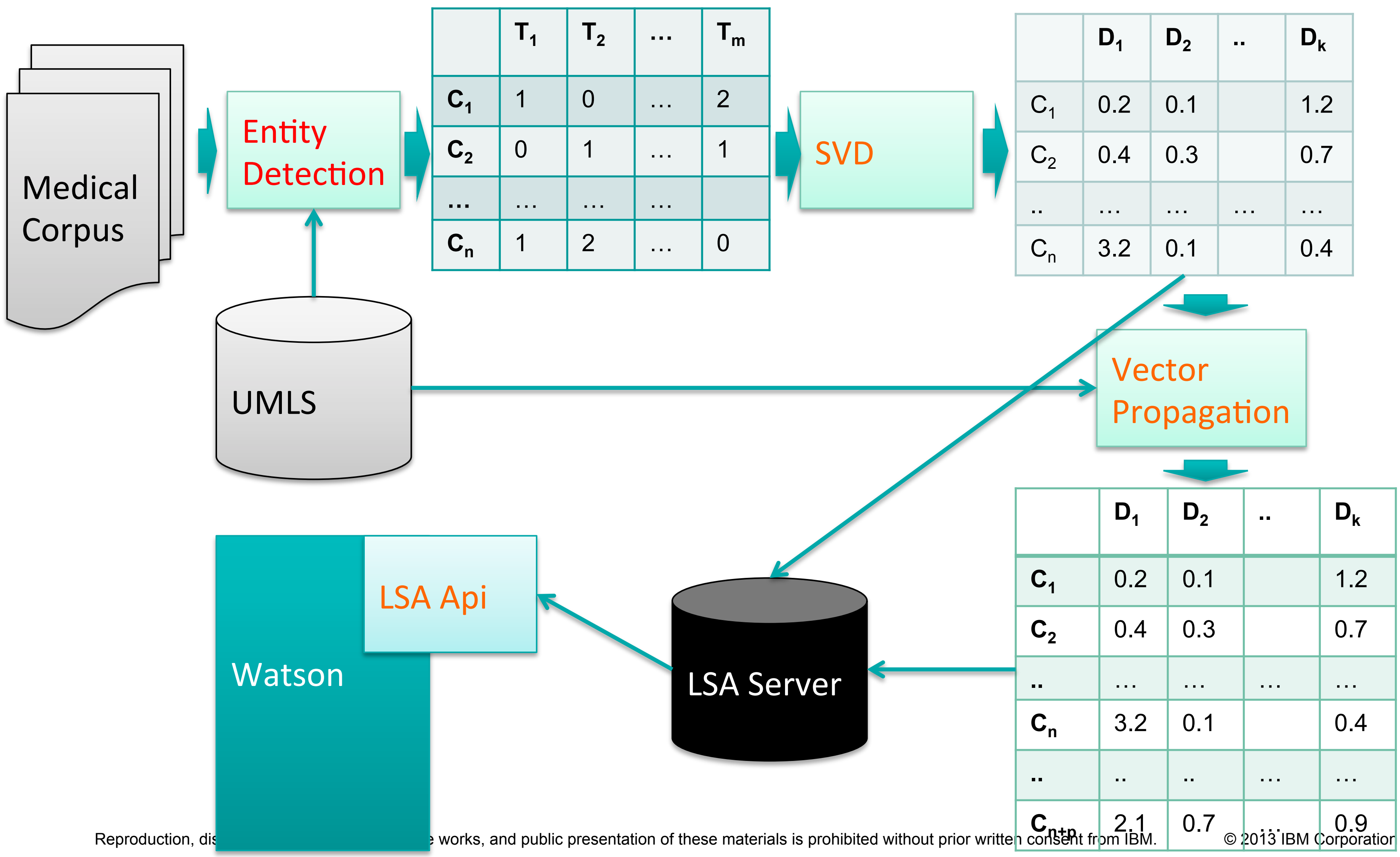
Outline

- Semantic Domains and LSA
- Semantic Domains in NLP
- LSA in Watson
- Distributional Semantics
- Inducing Meaning and Linking to Knowledge
- JoBimText and its application to Watson

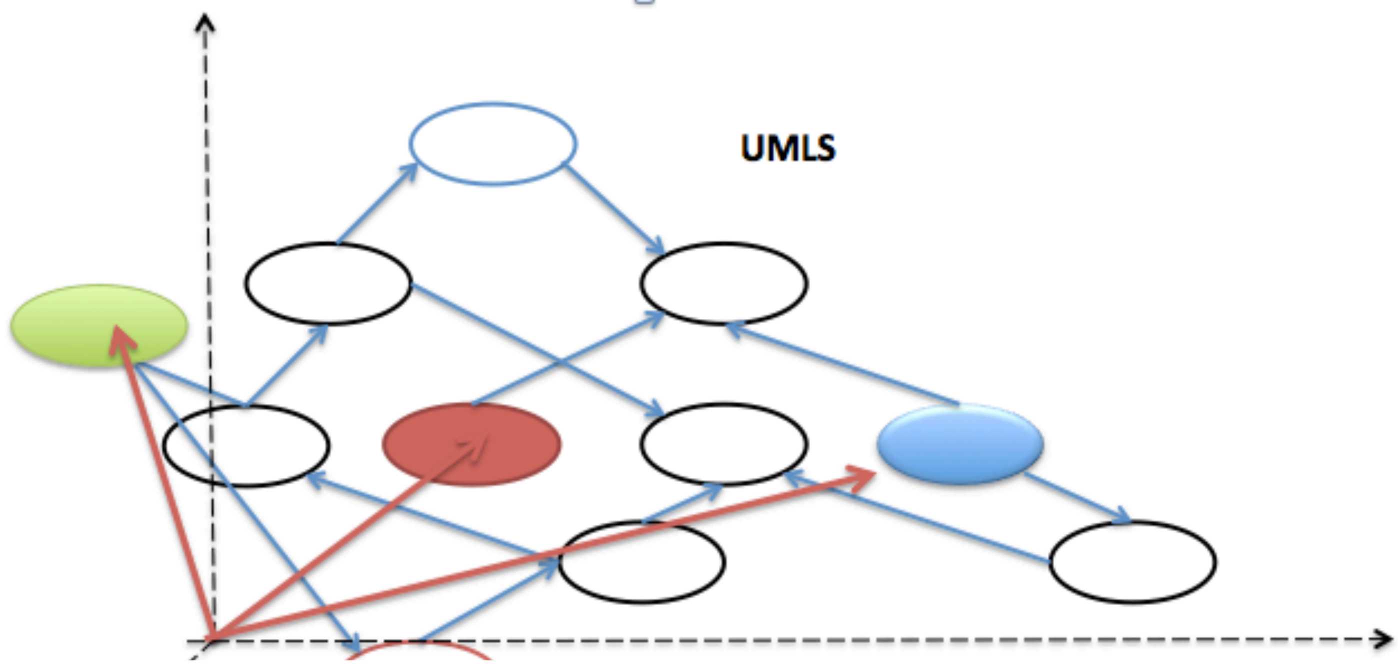
Distributional Semantics: Early Steps (2010 – 2012)

IBM WATSON

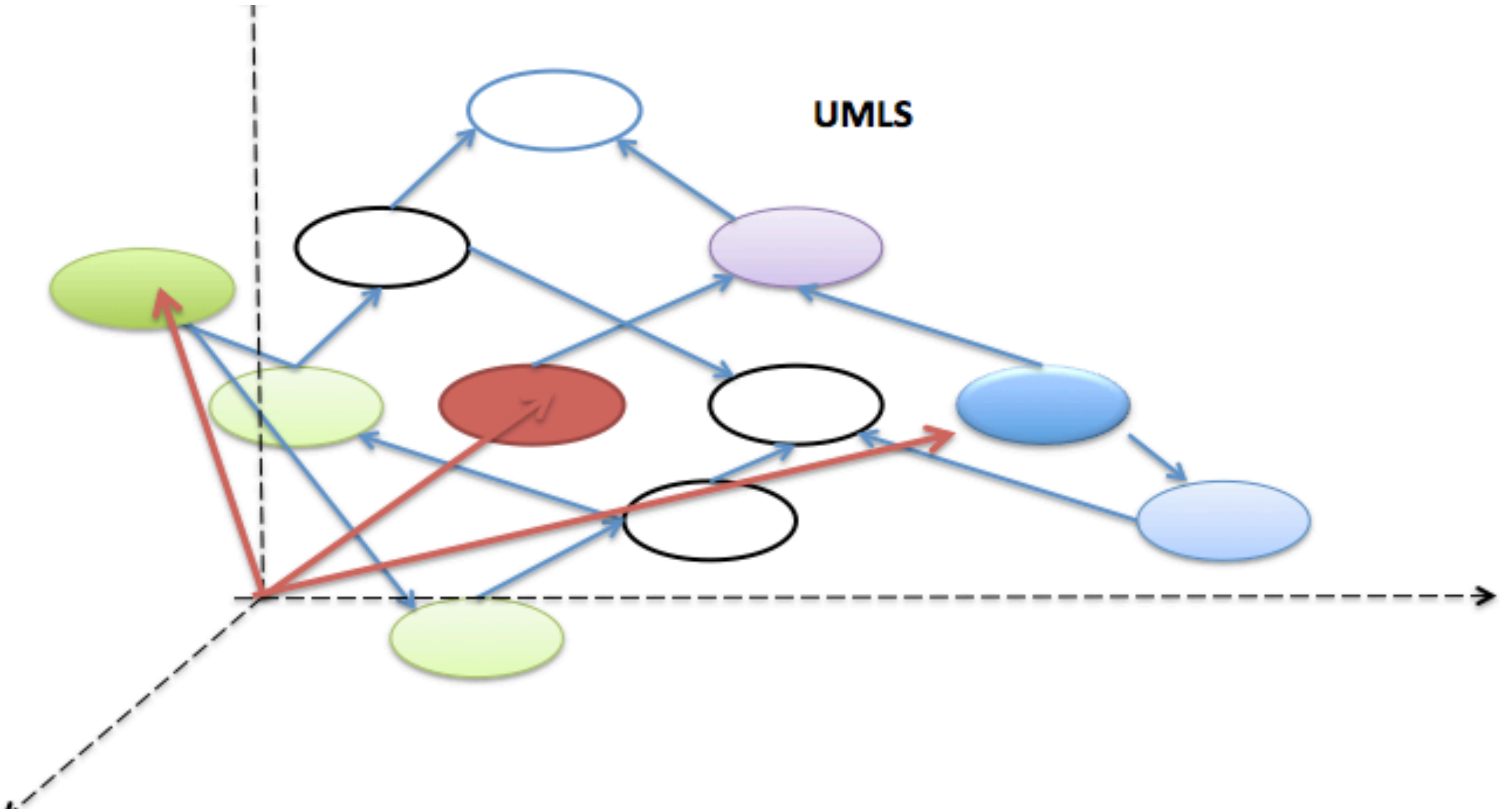
Latent Semantic Analysis (2.0)



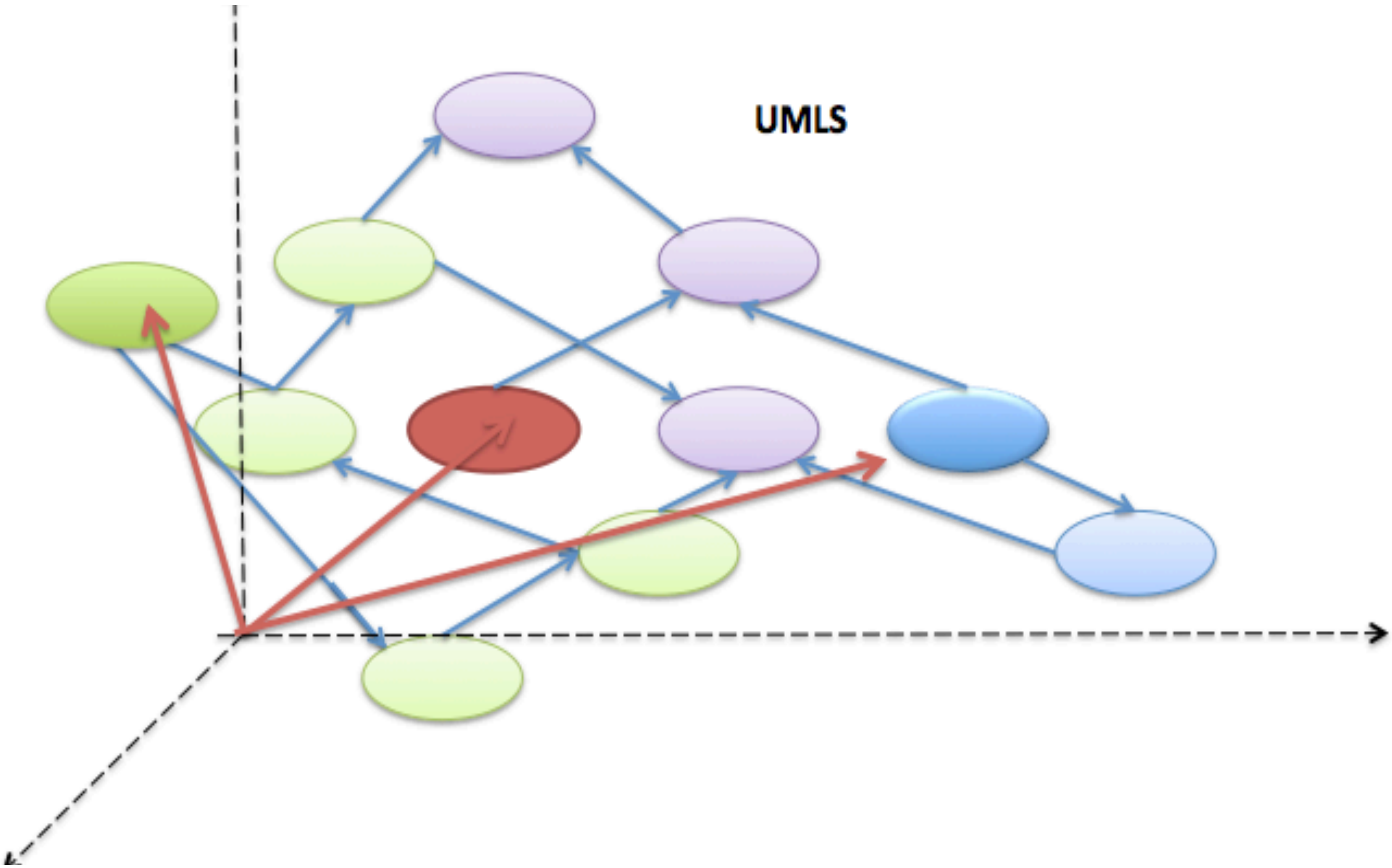
LSA 2.0: Vector Propagation over the Network



Vector Propagation 2/3



Vector Propagation 3/3

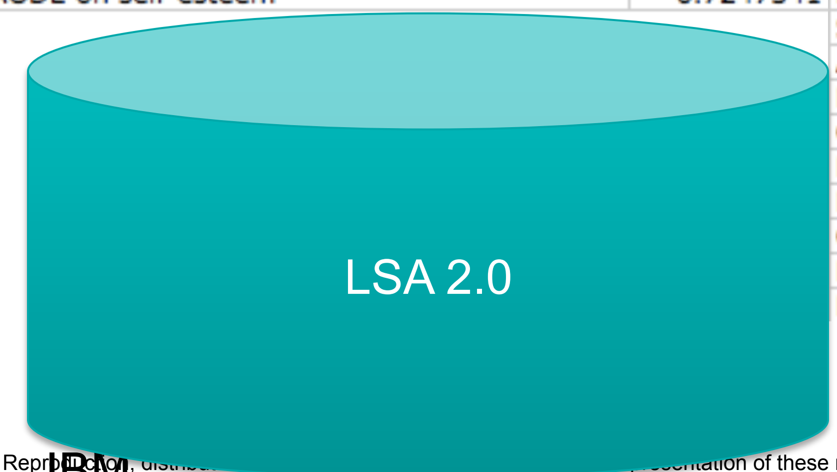


LSA 2.0: Factors related to suicide

IBM WATSON

"Suicide"

DANGER OF HARM TO SELF	0.94843552	Feeling hopeless	0.69763276
Depressive Symptoms	0.85787663	CYCLOTHYMIC REACTION	0.6956163
marked mood shift	0.83171128	Mental health counselor	0.6916423
loss of interest in activity	0.83171128	Demoralization	0.68469489
Other mood affective disorders	0.80852182	Ability to maintain self-esteem	0.67854127
Mood Disorders	0.80852182	Normal mood	0.67817024
Bipolar affective disorder, current episode manic	0.79134531	Despondency	0.67736145
Depressive disorder NEC in SNOMEDCT	0.78274978	Other and unspecified episodic mood disorder	0.67540516
change in self-esteem	0.77332301	Loss of interest	0.67413379
(Depression: [episode, unspecified] or [NOS (&	0.76803559	Suicidal	0.67144792
Self Esteem	0.72473412	pleasurable emotion	0.67024476
self-esteem as an AODC	0.7247341	Mood (psychological function)	0.67023923
AODE on self-esteem	0.7247341	Mood:-:Point in time:^Patient:-	0.66983514
		Suicidal behavior	0.6680896
		Adjustment disorder with depressed mood	0.6555974
		Depression aggravated	0.6528632
		Coping with Chronic Illness Topics	0.64542571
		Mental Health and Behavior	0.6454257
		Recurrent depression	0.64434724
		Other specified episodic mood disorder	0.64310002
		Melancholia	0.64063775
		Mild recurrent major depression	0.63696897



- Used as an Answer Scorer
  - –similarity between the answers and the questions
- Could be used as a Passage Scorer
  - –Similarity between passages containing the candidate answer and questions

	Accuracy	Precision@70
LSA (CUI)	+0.66%	+0.47%
LSA 2.0 (first 3)	+1.13% (4.46%)	+1.58%(5.229%

- average_precision: +1.62% (1.215%)
- confidence_weighted_score: +1.5% (0.669%)
- recall@005: 1.5 (0.895%)
- recall@020: 1.3 (2.281%)

Limitations of LSA

- SVD algorithm does not scale to WEB size
  - max 1M terms x3M documents
- Represents topicality but not paradigmatic relations (e.g. isa, synonymy, …)
- It compresses information rather than expanding it
  - Tells you that terms are similar without explaining why
- Need for something more precise but still learnable from large corpora
  - Near-synonym
  - Taxonomy Induction
  - Sense Clustering
- Need for something in the middle (Dave Ferrucci’s request)
  - Trainable from large corpora without supervision
  - More fine grained meaning representation (near synonymy vs. topicality)
  - Stronger semantic relations (may_treat vs. associated_with)

Outline

- Semantic Domains and LSA
- Semantic Domains in NLP
- LSA in Watson
- Distributional Semantics
- Inducing Meaning and Linking to Knowledge
- JoBimText and its application to Watson

The Distributional Semantics Paradigm

- The challenge: Fully Unsupervised Computational Semantics
  - Input: few Gigabytes of raw text in a specific domain
  - Output: Semantic Analyzer having the following capabilities
    - Term/Text Similarity beyond Keyword Matching
    - WSD, Lexical Substitution
    - Matching: terms, relations
    - Linking text to knowledge bases
  - Radical Approach:
    - Mining (clustering) big data
    - No Rules, No labeled data
- Making it scalable (Hadoop)
  - More text = more hardware = same time
  - Fast semantic parsing
  - Web size Distributional Semantics to capture background knowledge

Distributional Hypothesis and Structuralism

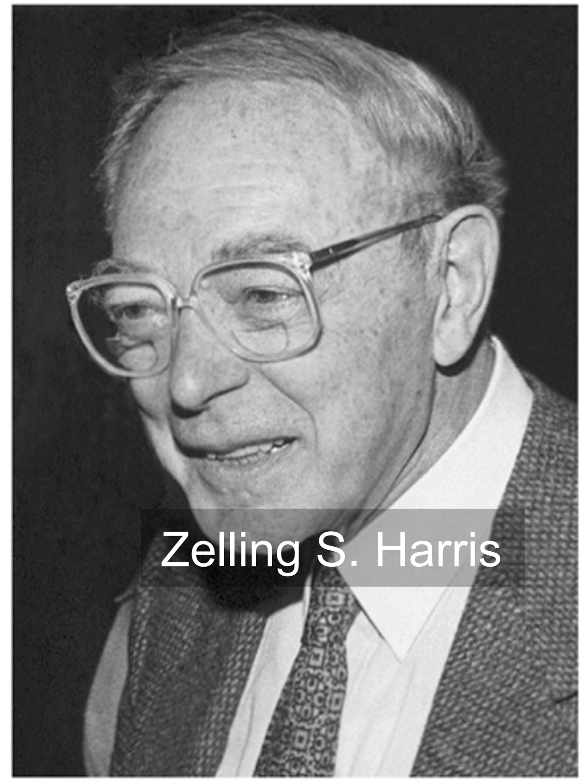
The **Distributional Hypothesis** in linguistics is the theory that words that occur in similar contexts tend to have similar meanings (paradigmatic relations).

The Distributional Hypothesis is the basis for Distributional Semantics.

It states that the meaning of a word can be defined in terms of its context (properties).

- other words in the same sentence/document (bag of words)
- words in the immediate neighbors
- words along dependency paths
- Predicate Argument Structure
- Frame

➔ any process that builds a structure on sentences can be used as a source for properties

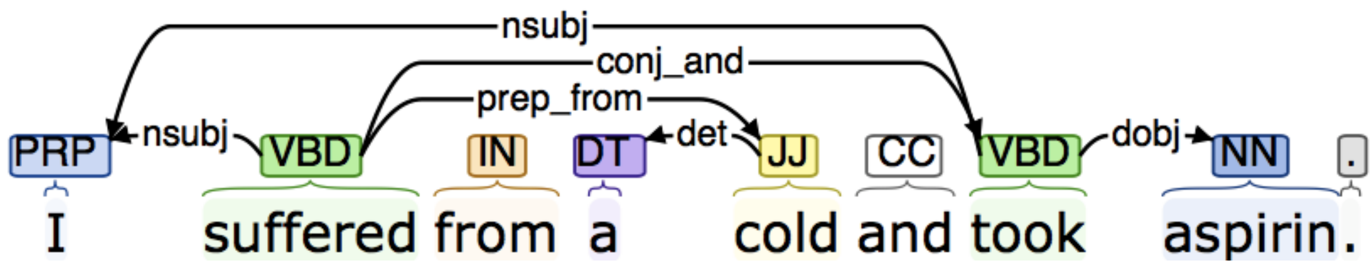


Z. Harris. (1954). Distributional Structure. Word 10 (2/3)

IBM WATSON

The @@ operation: JoBimPairs for Syntax Based Distributional Similarity

SENTENCE:



Dependency Parser:

nsubj(suffered, I); nsubj(took, I); root(ROOT, suffered); det(cold, a); prep_from(suffered, cold); conj_and(suffered, took); dobj(took, aspirin)

WORD-dependency PAIRS:

suffered	nsubj(@@, I)	1
took	nsubj(@@, I)	1
cold	det(@@, a)	1
suffered	prep_from(@@, cold)	1
suffered	conj_and(@@, took)	1
took	dobj(@@, aspirin)	1

I	nsubj(suffered, @@)	1
I	nsubj(took, @@)	1
a	det(cold, @@)	1
cold	prep_from(suffered, @@)	1
took	conj_and(suffered, @@)	1
aspirin	dobj(took, @@)	1

The @@ operation:
JoBim Pairs for ngram based distributional similarity

SENTENCE:

I suffered from a cold and took aspirin.

Ngrams

w-1(suffered, I); w-2(from, I); w+1(suffered, from), w+2(suffered, a);

Word-ngram PAIRS:

suffered	w-1(@@, I)	1
from	w-2(@@, I)	1
suffered	w+1(@@, from)	1
suffered	w+2(@@, a)	1

Jo	Bim	
I	w-1(suffered, @@)	1
I	w-2 (from, @@)	1
from	w+1(suffered, @@)	1
a	w+2(suffered, @@)	1

The @@ operation:
JoBim Pairs for Topic Similarity (LSA)

SENTENCES:

1) I suffered from a cold and took aspirin.

2) Antibiotics are more effective than aspirin

Word-document pairs

doc(suffered,1); doc(cold, 1); doc(aspirin, 1);

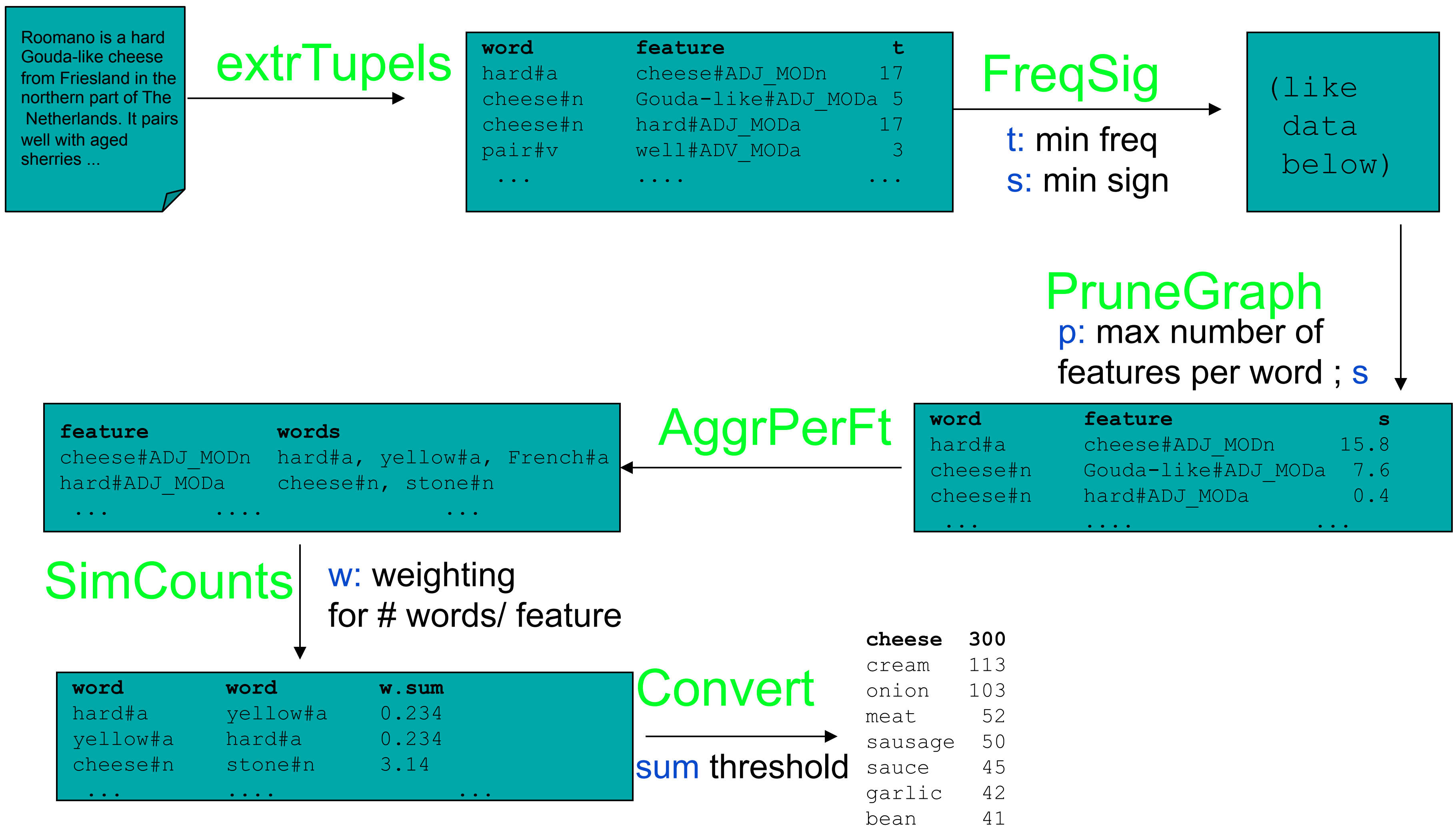
doc(antibiotics,2); doc(aspirin,2);

W Jo **FEATU** Bim **AIRS:**

suffered	doc(@@, 1)	1
cold	doc(@@, 1)	1
aspirin	doc(@@, 1)	1
antibiotics	doc(@@, 2)	1
aspirin	doc(@@,2)	1

1	doc(suffered, @@)	1
1	doc(cold, @@)	1
1	doc(aspirin, @@)	1
2	doc(antibiotics, @@)	1
2	doc(aspirin,@@)	1

Distributional Thesaurus with MapReduce steps



DT entry “paper#NN” with contexts

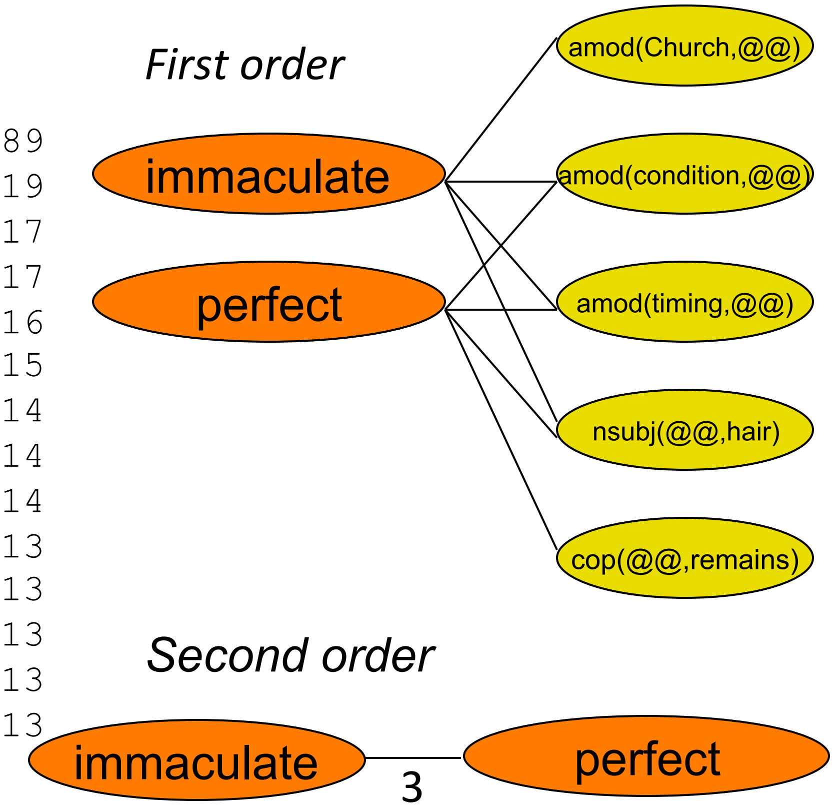
paper#NN	s	common contexts
newspaper#NN	45	told#VBD#-dobj column#NN#-prep_in local#JJ#amod editor#NN#-poss edition#NN#-prep_of editor#NN#-prep_of hometown#NN#nn industry#NN#-nn clips#NNS#-nn shredded#JJ#amod pick#VB#-dobj news#NNP#appos daily#JJ#amod writes#VBZ#-nsubj write#VB#-prep_for wrote#VBD#-prep_for wrote#VBD#-prep_in wrapped#VBN#-prep_in reading#VBG#-prep_in reading#VBG#-dobj read#VBD#-prep_in read#VBD#-dobj read#VBP#-prep_in read#VB#-dobj read#VB#-prep_in record#NN#prep_of article#NN#-prep_in reports#VBZ#-nsubj reported#VBD#-nsubj printed#VBN#amod printed#VBD#-nsubj printed#VBN#-prep_in published#VBN#-prep_in published#VBN#partmod published#VBD#-nsubj sunday#NNP#nn section#NN#-prep_of school#NN#nn saw#VBD#-prep_in ad#NN#-prep_in copy#NN#-prep_of page#NN#-prep_of pages#NNS#-prep_of morning#NN#nn story#NN#-prep_in
book#NN	33	recent#JJ#amod read#VB#-dobj read#VBD#-dobj reading#VBG#-dobj edition#NN#-prep_of printed#VBN#amod industry#NN#-nn described#VBN#-prep_in writing#VBG#-dobj wrote#VBD#-prep_in wrote#VBD#rcmod write#VB#-dobj written#VBN#rcmod written#VBN#-dobj wrote#VBD#-dobj pick#VB#-dobj photo#NN#nn co-author#NN#-prep_of co-authored#VBN#-dobj section#NN#-prep_of published#VBN#-dobj published#VBN#-nsubjpass published#VBD#-dobj published#VBN#partmod copy#NN#-prep_of buying#VBG#-dobj buy#VB#-dobj author#NN#-prep_of bag#NN#-nn bags#NNS#-nn page#NN#-prep_of pages#NNS#-prep_of titled#VBN#partmod
article#NN	28	authors#NNS#-prep_of original#JJ#amod notes#VBZ#-nsubj published#VBN#-dobj published#VBD#-dobj published#VBN#-nsubjpass published#VBN#partmod write#VB#-dobj wrote#VBD#rcmod wrote#VBD#-prep_in written#VBN#rcmod wrote#VBD#-dobj written#VBN#-dobj writing#VBG#-dobj reported#VBD#-nsubj describing#VBG#partmod described#VBN#-prep_in copy#NN#-prep_of said#VBD#-prep_in recent#JJ#amod read#VB#-dobj read#VB#-prep_in read#VBD#-dobj read#VBD#-prep_in reading#VBG#-dobj author#NN#-prep_of titled#VBN#partmod lancet#NNP#nn
magazine#NN	26	editor#NN#-poss editor#NN#-prep_of edition#NN#-prep_of industry#NN#-nn copy#NN#-prep_of article#NN#-prep_in ad#NN#-prep_in published#VBD#-nsubj published#VBN#partmod published#VBN#-prep_in page#NN#-prep_of pages#NNS#-prep_of story#NN#-prep_in buy#VB#-dobj wrote#VBD#-prep_in wrote#VBD#-prep_for printed#VBN#-prep_in printed#VBN#amod reading#VBG#-dobj read#VBD#-prep_in read#VB#-dobj reported#VBD#-nsubj reports#VBZ#-nsubj column#NN#-prep_for glossy#JJ#amod told#VBD#-dobj
plastic#NN	24	wrapped#VBD#-prep_in wrapped#VBN#-prep_in wood#NN#conj_and sheet#NN#-prep_of sheets#NNS#-prep_of shredded#JJ#amod bits#NNS#-prep_of paper#NN#-conj_and paper#NN#conj_and cardboard#NN#-conj_and cardboard#NN#conj_and pieces#NNS#-prep_of piece#NN#-prep_of rolls#NNS#-prep_of bags#NNS#-nn bags#NN#-nn bag#NN#-nn recycled#JJ#amod cups#NNS#-nn made#VBN#-prep_from white#JJ#amod glossy#JJ#amod glass#NN#-conj_and glass#NN#conj_and
metal#NN	23	bits#NNS#-prep_of made#VBN#-prep_from work#NN#-nn wood#NN#conj_and scrap#NN#nn paper#NN#-conj_and piece#NN#-prep_of pile#NN#-prep_of pieces#NNS#-prep_of plastic#NN#conj_and plastic#NN#-conj_and plastic#NN#conj_or plate#NN#-nn plates#NNS#-nn recycled#JJ#amod clip#NN#-nn products#NNS#-nn put#VBD#-prep_to put#VB#-prep_to glass#NN#-conj_and glass#NN#conj_and tons#NNS#-prep_of white#JJ#amod

Distributional Thesaurus (DT)

- Computed from distributional similarity statistics
- **Entry** for a **target** word consists of a ranked list of neighbors

meeting	
meeting	288
meetings	102
hearing	89
session	68
conference	62
summit	51
forum	46
workshop	46
hearings	46
ceremony	45
sessions	41
briefing	40
event	40
convention	38
gathering	36
...	

articulate	
articulate	89
explain	19
understand	17
communicate	17
defend	16
establish	15
deliver	14
evaluate	14
adjust	14
manage	13
speak	13
change	13
answer	13
maintain	13
...	



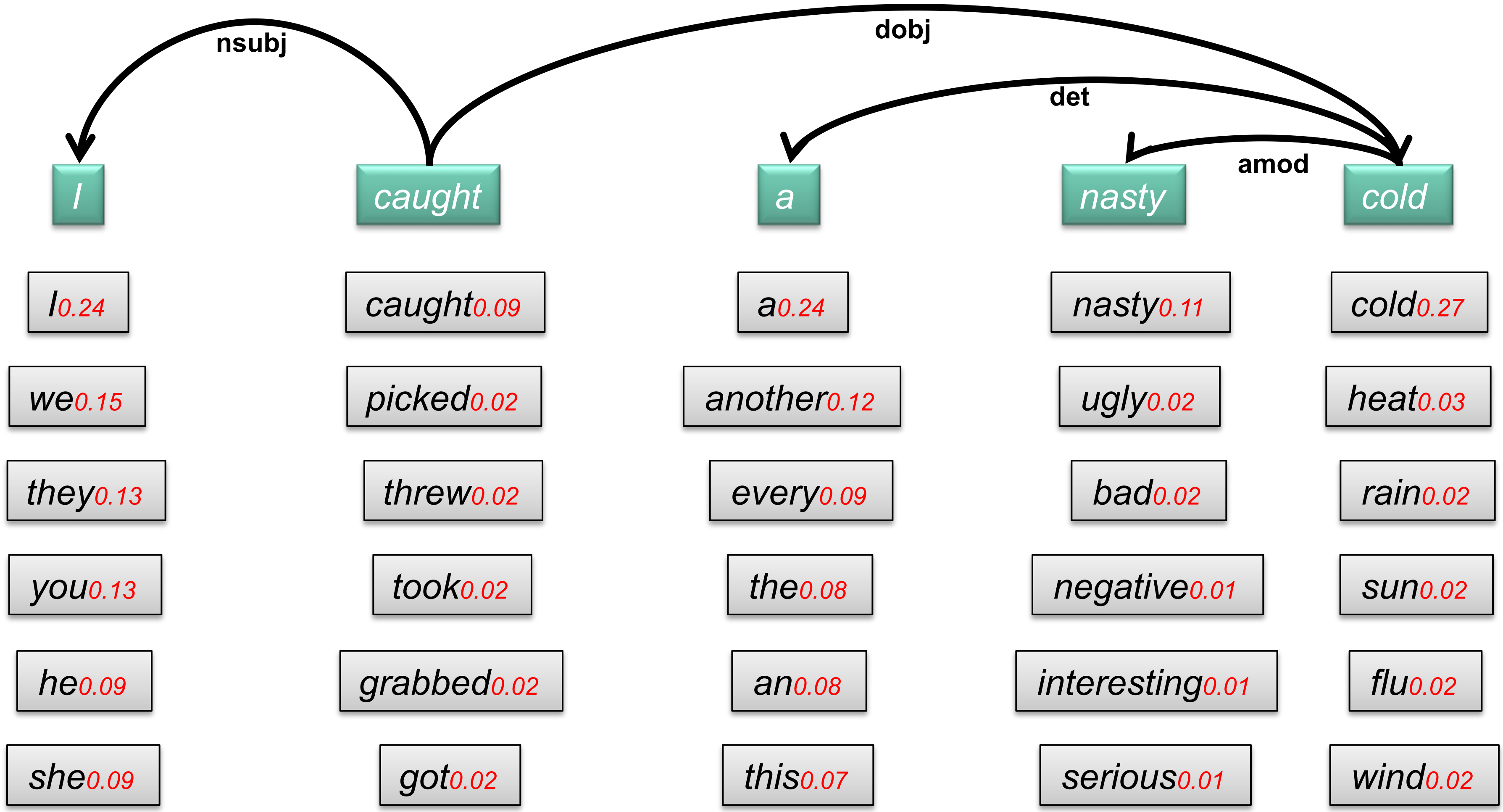
From Distributional to Contextual Thesaurus

- Distributional Thesaurus (DT)
  - computed from global statistics
  - ranked list of most similar terms, mix of senses
  - words that have similar contexts, but will not fit in all contexts of the target
- Contextual Thesaurus (CT)
  - use DT as a source for similar terms
  - re-rank entries according to how they fit **in a given context**
  - correct senses should be ranked higher

Why not use (only) a lexical resource for word sense disambiguation/expansion?

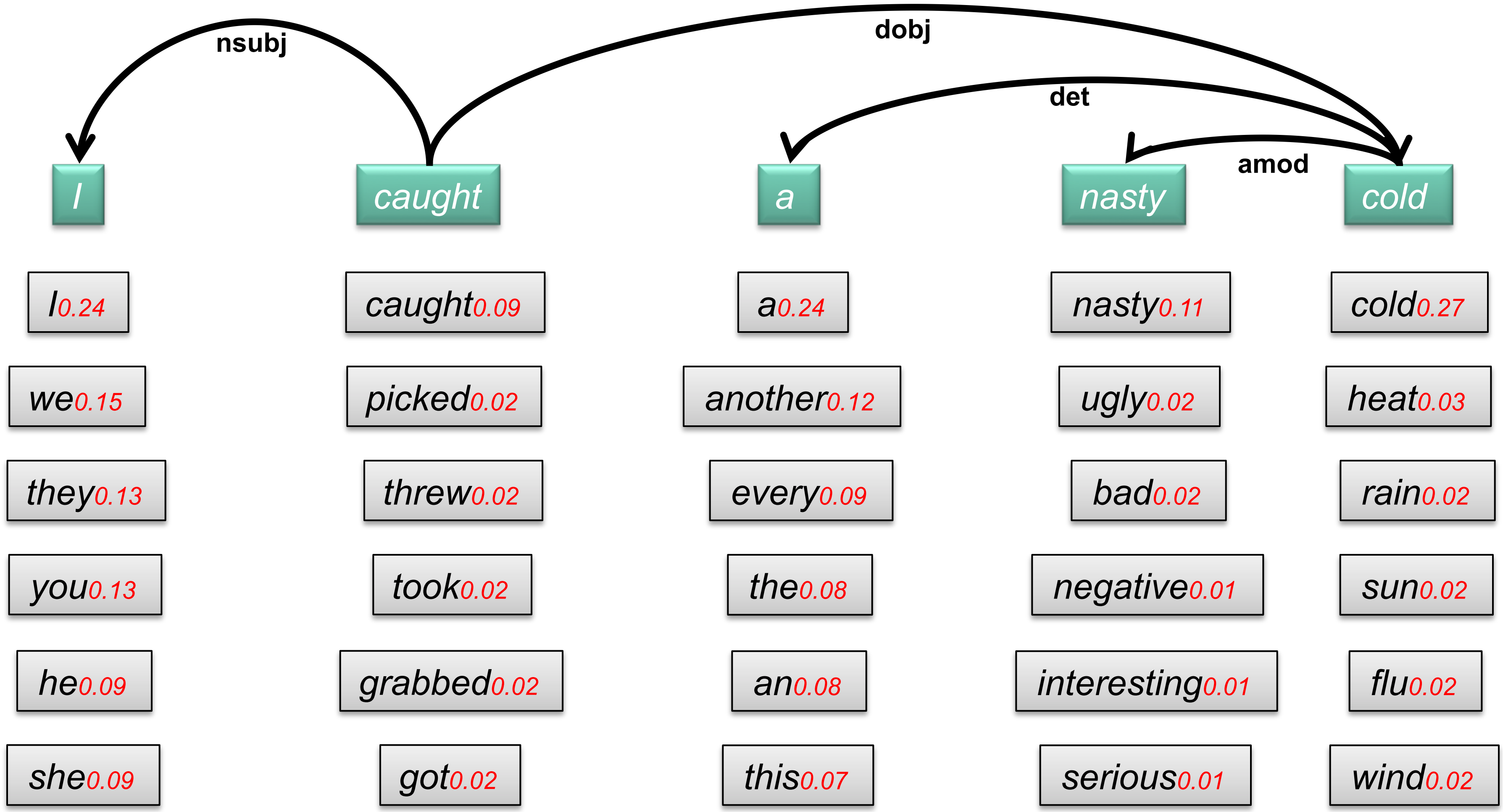
- all resources leak
- we want to be agnostic of sense distinctions
- data-driven, language- and domain-independent

Parse and obtain priors



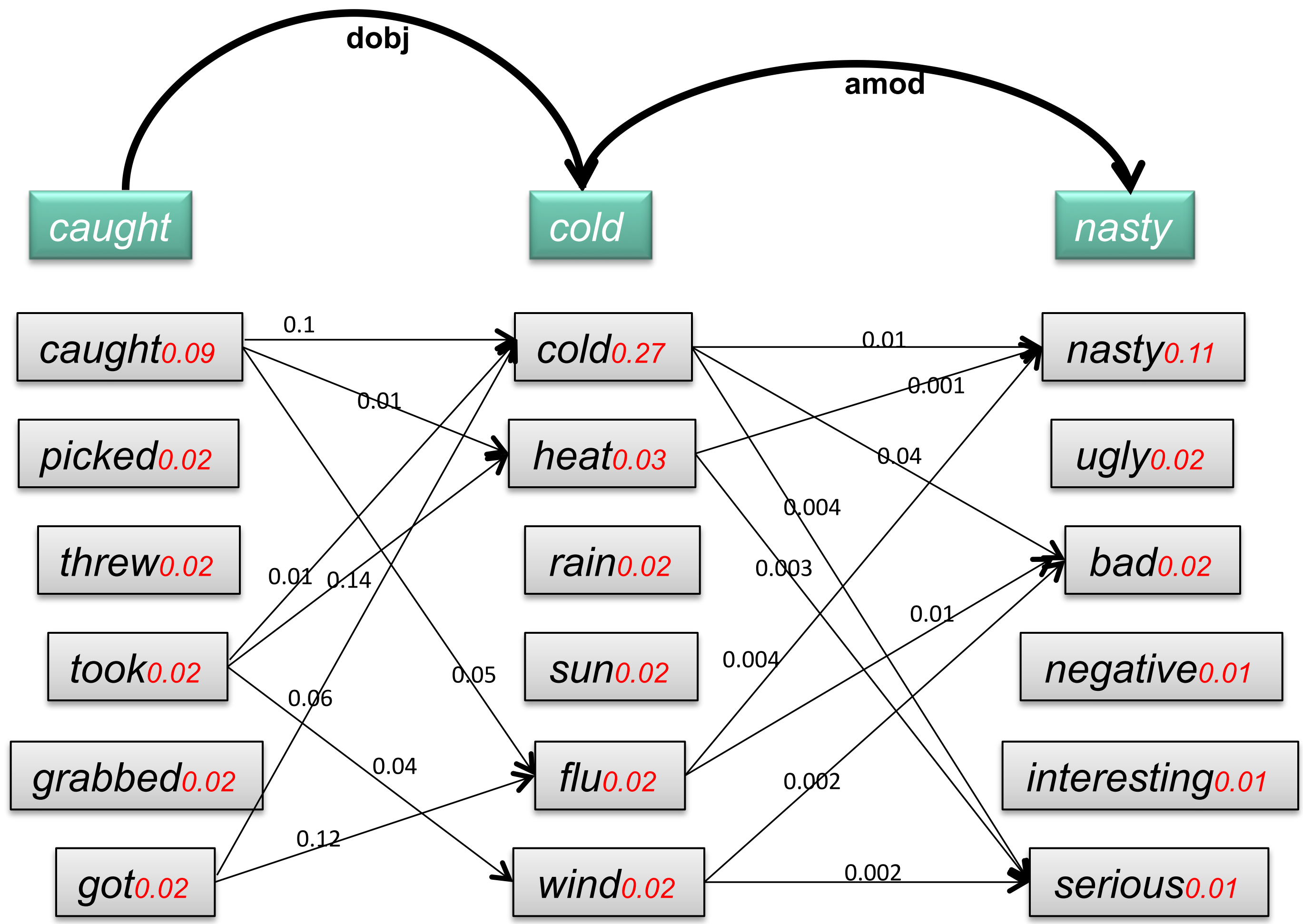
IBM WATSON

Reorder (to simplify the example)



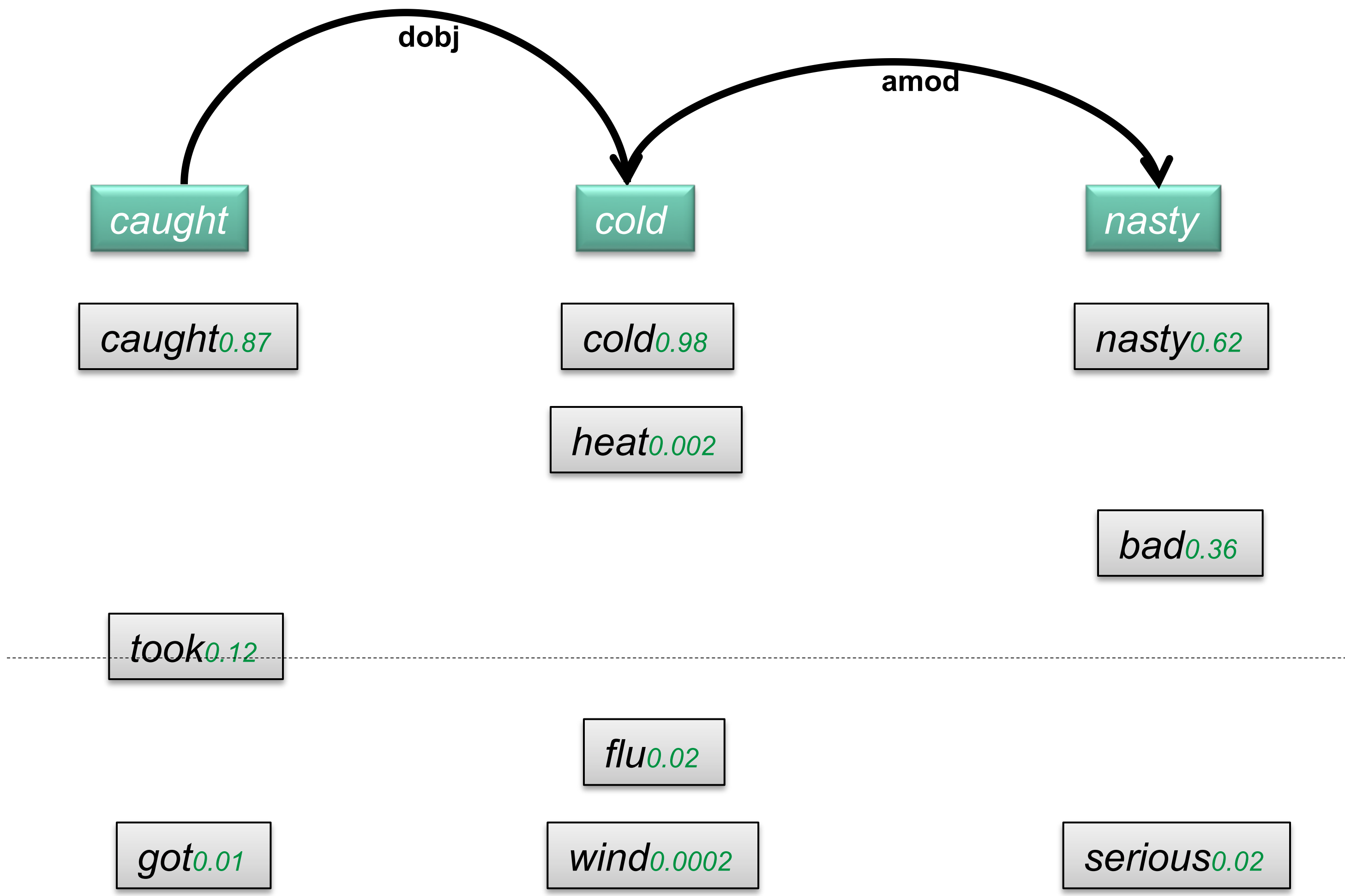
Conditional Probabilities and Priors

- Run Viterbi algorithm over all possible paths
- Sum path probabilities per word



Representation:
probability distribution over words

- Re-rank by sum of path distributions
- cut on threshold



POS-tagging with two-dimensional text

- POS-tagging is hard for unknown words, which have not been observed in the training
- Schema: replace the unknown words with the most similar known word from an n-gram DT
- tag new sentence with standard POS tagger

	JJ	JJ	NN	
Renting out an	unfurnished	one-bedroom	triplex	in San Francisco
	empty	two-bedroom	duplex	
		three-bedroom		
		two-room		

OOV acc. PTB, TreeTagger

1-dimensional: 38.2%

2-dimensional: 76.4%

Simplified Lesk (SL)

A patient **fell** over a **stack** of <u>magazines</u> in an aisle at a physiotherapist practice.

patient	fell	stack	aisle	physiotherapist	practice
customer	rose	pile	field	physician	session
student	dropped	copy	hill	attorney	game
individual	climbed	lots	line	psychiatrist	camp
person	increased	dozens	river	scholar	workouts
mother	slipped	array	stairs	engineer	training
user	declined	collection	road	journalist	meeting
passenger	tumbled	amount	hall	contractor	work
..	surged	ton	driveway	…	…
	…	…	…		

Zero word overlap

WordNet: S: (n) <u>magazine</u> (product consisting of a paperback periodic publication as a physical object) "tripped over a **pile** of <u>magazines</u>"

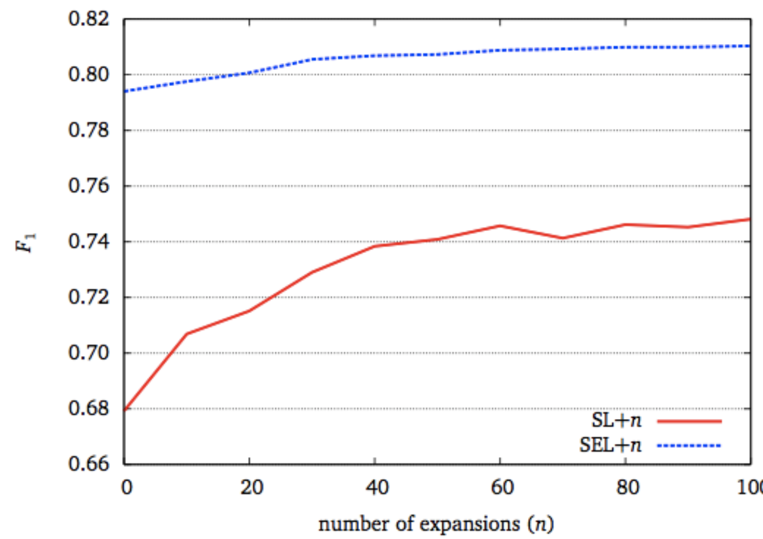
tripped	pile
jumped	stack
woke	tons
turned	piece
drove	heap
walked	collection
blew	bag
put	loads
fell	mountain
..	..

Tristan Miller, Chris Biemann, Torsten Zesch, Iryna Gurevych (2012): Using Distributional Similarity for Lexical Expansion in Knowledge-based Word Sense Disambiguation. Proceedings of COLING-12, Mumbai, India

Using Expansions in WSD: Evaluation

	part of speech				
system	adj.	noun	adv.	verb	all
MFS baseline	84.25	77.44	87.50	75.30	78.89
random baseline	68.54	61.96	69.15	52.81	61.28
SL+0	75.32	69.71	69.75	59.46	67.92
SL+100	82.18	76.31	78.85	66.07	74.81
SEL+0	87.19	81.52	74.87	**72.26**	79.40
SEL+100	**88.40**	**83.45**	80.29	72.25	**81.03**
TKB-UO	78.73	70.76	74.04	62.61	70.21
MII+ref	82.04	80.05	**82.21**	70.73	78.14
WN++-DC	—	79.4	—	—	—

Table 2: Results (F_1) on the SemEval-2007 corpus by part of speech

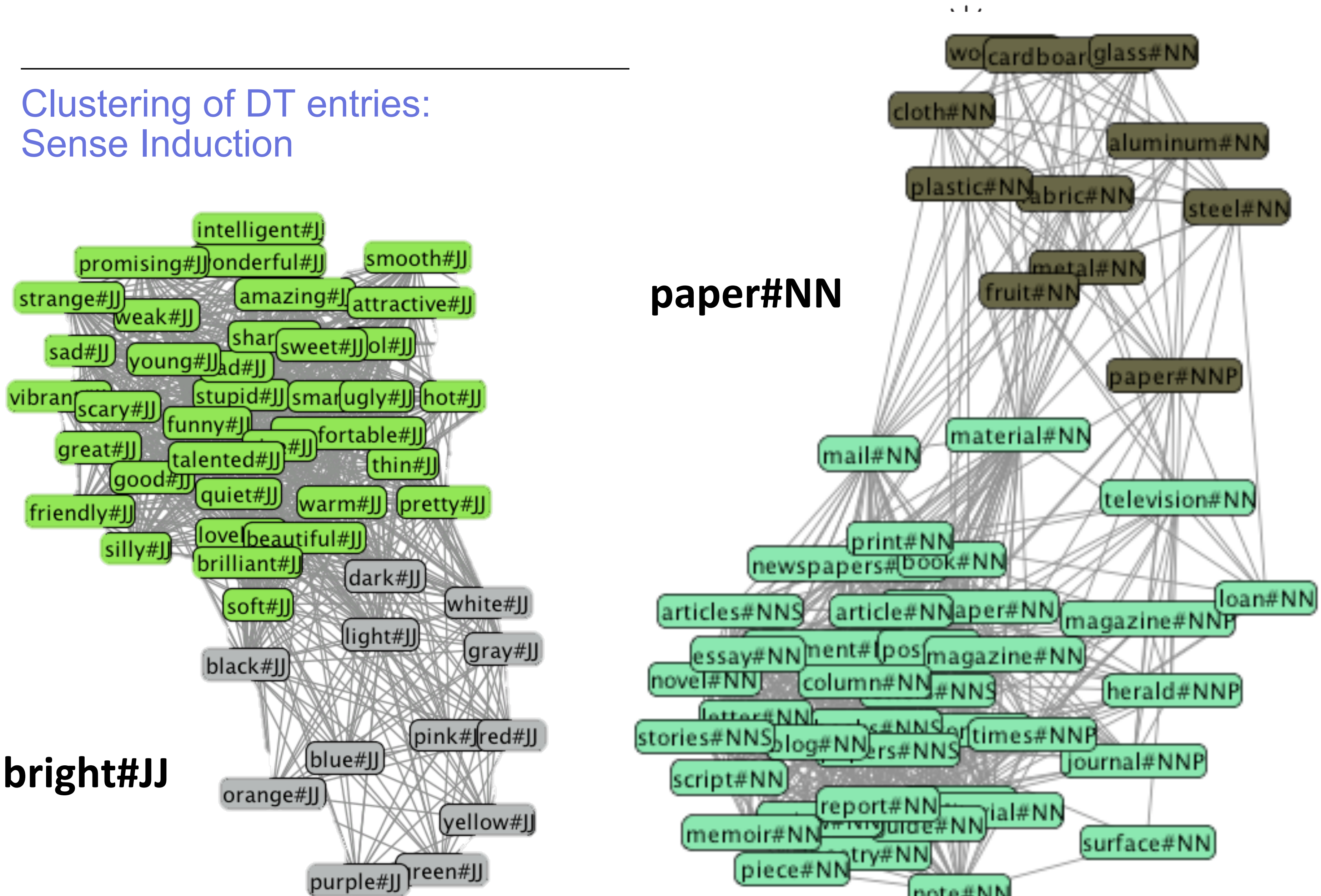


Simplified Lesk (SL)
Simplified Extended Lesk (SEL)

Outline

- Semantic Domains and LSA
- Semantic Domains in NLP
- LSA in Watson
- Distributional Semantics
- Inducing Meaning and Linking to Knowledge
- JoBimText and its application to Watson

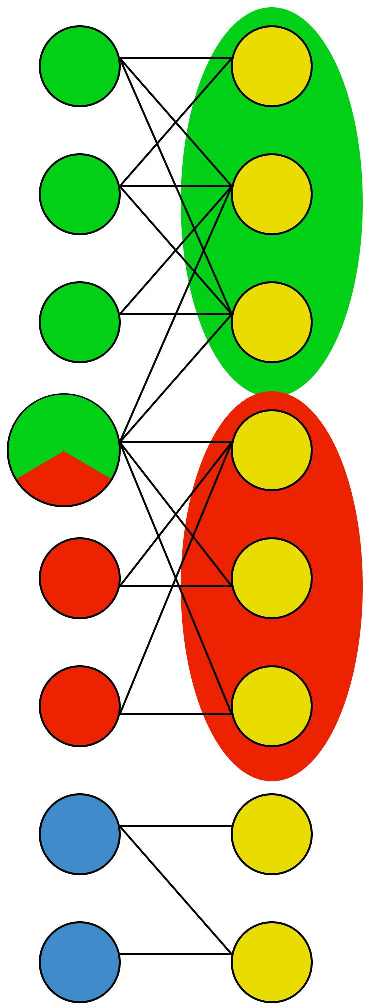
Clustering of DT entries: Sense Induction



C. Biemann (2006): Chinese Whispers - an Efficient Graph Clustering Algorithm and its Application to Natural Language Processing Problems. Proceedings of the HLT-NAACL-06 Workshop on Textgraphs-06, New York, USA.

Ambiguity

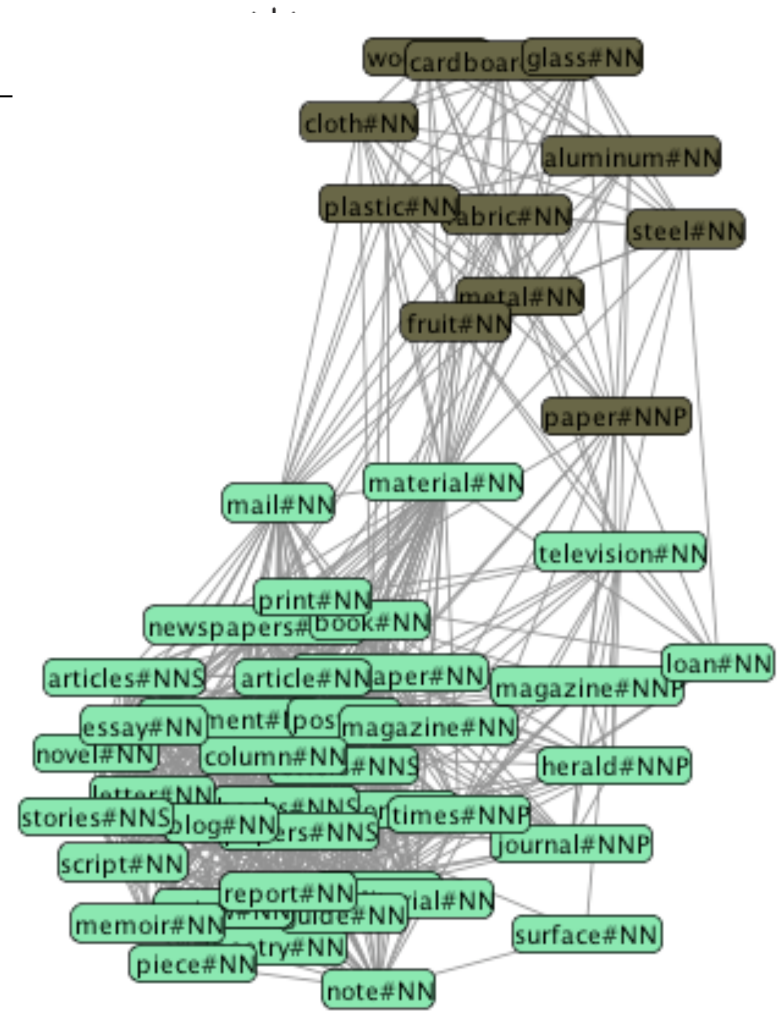
- Ambiguous items have several senses: connect to different clusters
- Estimation of sense priors



<u>Applications</u>:

- disambiguation pages
- result diversity
- refinement options
- lexical expansion

Features for Disambiguation



paper 0 (newspaper)

read#VB#-dobj	45
reading#VBG#-dobj	45
write#VB#-dobj	38
read#VBD#-dobj	37
writing#VBG#-dobj	36
wrote#VBD#-dobj	34
original#JJ#amod	27
wrote#VBD#-prep_in	26
recent#JJ#amod	26
published#VBN#partmod	25
written#VBN#-dobj	23
published#VBN#-nsubjpass	20
published#VBD#-dobj	19
copy#NN#-prep_of	18
said#VBD#-prep_in	18
author#NN#-prep_of	17
pages#NNS#-prep_of	16
told#VBD#-dobj	15
buy#VB#-dobj	14
published#VBN#-prep_in	14
page#NN#-prep_of	14

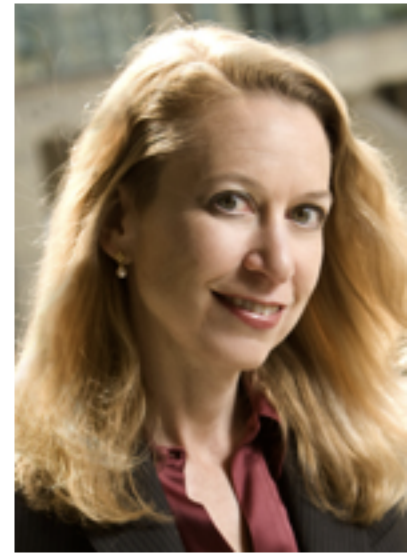
paper 1 (material)

piece#NN#-prep_of	21
pieces#NNS#-prep_of	17
made#VBN#-prep_from	13
bags#NNS#-nn	11
white#JJ#amod	9
paper#NN#-conj_and	9
glass#NN#-conj_and	9
products#NNS#-nn	9
industry#NN#-nn	8
plastic#NN#conj_and	8
plastic#NN#-conj_and	8
bits#NNS#-prep_of	8
bag#NN#-nn	8
plastic#NN#conj_or	8
sheet#NN#-prep_of	7
recycled#JJ#amod	7
tons#NNS#-prep_of	7
glass#NN#conj_and	7
buy#VB#-dobj	6
plates#NNS#-nn	6
pile#NN#-prep_of	6

These are shared by **paper** and the cluster members.

Disambiguation: find features in context.
I am reading an original paper on the **recycled** paper **industry** .

Cluster Labeling with IS-A Relations



- Run Hearst IS-A patterns (e.g. `NP such as NP, NP and NP`) on a large collection of text and store (noisy) IS-A pairs with their frequency, if above a threshold
- Activate hypernyms from cluster entries

Typical regular polysemy in the medical domain:

influenza#0 viral gastroenteritis, bird flu, pulmonary anthrax, h1n1, tularaemia, west nile fever, mumps, influenza a, herpes zoster, respiratory infection, uri, phn, chicken pox, …

influenza#1 trivalent influenza vaccine, antiviral, influenza vaccine, amantadine, peramivir, chemoprophylaxis, influenza vaccination, vaccine, poultry, chickenpox vaccine, …

Hearst, M. Automatic Acquisition of Hyponyms from Large Text Corpora, Proceedings of the Fourteenth International Conference on Computational Linguistics, Nantes, France, July 1992.

Per-Cluster IS-A Pattern Counts

viral gastroenteritis ISA gastroenteritis	555
viral gastroenteritis ISA illness	27
viral gastroenteritis ISA flu	24
viral gastroenteritis ISA stomach flu	24
viral gastroenteritis ISA infection	18
viral gastroenteritis ISA cause	15
viral gastroenteritis ISA The Basics	14

bird flu ISA flu	829
bird flu ISA avian influenza	42
bird flu ISA infection	42
bird flu ISA influenza	27
bird flu ISA pandemic	27
bird flu ISA avian flu	16
bird flu ISA vaccine	15

chicken pox ISA pox	2390
chicken pox ISA virus	93
chicken pox ISA disease	92
chicken pox ISA infection	76
chicken pox ISA viral infection	69
chicken pox ISA symptom	43
chicken pox ISA illness	38

influenza#0 viral gastroenteritis, bird flu, pulmonary anthrax, h1n1, tularaemia, west nile fever, mumps, influenza a, herpes zoster, respiratory infection, uri, phn, chicken pox, …

influenza#1 trivalent influenza vaccine, antiviral, influenza vaccine, amantadine, peramivir, chemoprophylaxis, influenza vaccination, vaccine, poultry, chickenpox vaccine, …

peramivir ISA inhibitor	10
peramivir ISA option	4
peramivir ISA neuraminidase inhibitor	3

antiviral ISA treatment	8
antiviral ISA medication	6
antiviral ISA drug	5

influenza vaccine ISA vaccine	1532
influenza vaccine ISA flumist	10
influenza vaccine ISA intranasal	10
influenza vaccine ISA 2010-XX-XX	8
influenza vaccine ISA LAIV	8
influenza vaccine ISA table	8
influenza vaccine ISA contrast	6

- Sum counts of ISA hypernym per cluster
- Multiply by number of times it was found by the cluster members

Work done in the MRP project, Chris & Alfio Sept 2012

Cluster Labeling with IS-A Patterns

influenza#0:

infection(3310937) disease(1748000) virus(817950) cause(783692) symptom(578480) fever(375228) condition(209022) illness(192675) complication(161158) influenza(155469)

viral gastroenteritis, bird flu, pulmonary anthrax, h1n1, tularaemia, west nile fever, mumps, influenza a, herpes zoster, respiratory infection, uri, phn, chicken pox, human papillomavirus, sars, shingle, upper respiratory tract infection, varicella infection, hepatitis a, omsk hemorrhagic fever, respiratory viral infection, monkeypox, tonsillitis, acute respiratory disease, †, h1n1 influenza, h1n1 virus, rubeola, avian flu, vhf, rotavirus infection, pneumonia, upper respiratory infection, rabies, a/h1n1, varicella, grippe, bacterial pneumonia, croup, acute bronchitis, avian flu virus, rhinovirus, primary coccidioidomycosis, respiratory tract disease, rubella, influenza a virus, cold, contagious disease, [illegible]on, scarlet fever, infectious disease, hib, poliomyelitis, rmsf, giardiasis, siv, virus, swine influenza virus, dengue fever, h3n2, primary herpes simplex, acute infection, primary hiv infection, infectious mononucleosis, swine influenza, hepatitis b, sepsis, varivax, influenza epidemic, yellow fever, secondary bacterial pneumonia, chickenpox, cholera, bird flu virus, q fever, flu symptom, whooping cough, stomach flu, influenza pneumonia, neonatal herpes simplex virus infection, pneumococcal infection, adenoviral infection, respiratory disease, mmr, distemper, egg allergy, strain, chikungunya, swine flu, reye's syndrome, non-a, viral infection, parainfluenza, cold symptom, sinus infection, anthrax, adenovirus infection, atypical measle, malaria, poliovirus, vee, hiv, h5n1, human immunodeficiency virus, viral shedding, hepatitis b infection, hpv, eye infection, tetanus, rhinotracheitis, human rabies, coccidioidomycosis, human influenza, fetal damage, adenovirus, hbv, rotavirus, ili, hepatitis, genital herpes simplex, reactive lymphocytosis, neonatal herpes simplex, sore throat, pertussis, mono, bcg, vaccinia, hantavirus, postherpetic neuralgia, illness, respiratory syncytial virus infection, superinfection, hepatitis e, zoster, dengue, h1n1 flu, strep throat, mononucleosis, guillain-barré, influenza virus, yf, japanese encephalitis, acute respiratory infection, typhoid fever, common cold, tick fever, polio, herpangina, pneumococcal pneumonia, cowpox, viral pneumonia, jev, complication, flu, human flu, measle, chikv, symptomatic aortic stenosis, bronchitis, mumps orchitis, virus infection, avian influenza, lyme disease, fhv, hav, shigellosis, meningococcal disease, kyasanur forest disease, smallpox, meningococcal infection, sinusitis

influenza#1:

vaccine(38566) drug(9990) agent(5004) vaccination(3960) treatment(2796) inhibitor(1504) medication(792) oseltamivir(752) medicine(448) zanamivir(396)

trivalent influenza vaccine, antiviral, influenza vaccine, amantadine, peramivir, [illegible], influenza vaccination, vaccine, [illegible] vaccine, live attenuated vaccine, neuraminidase inhibitor, oseltamivir, inactivated vaccine, oscillococcinum, tamiflu, relenza, antiviral drug, pneumococcal vaccine, rimantadine, zanamivir, shot, antiviral agent, flumist, flu vaccine, laiv, influenza virus vaccine

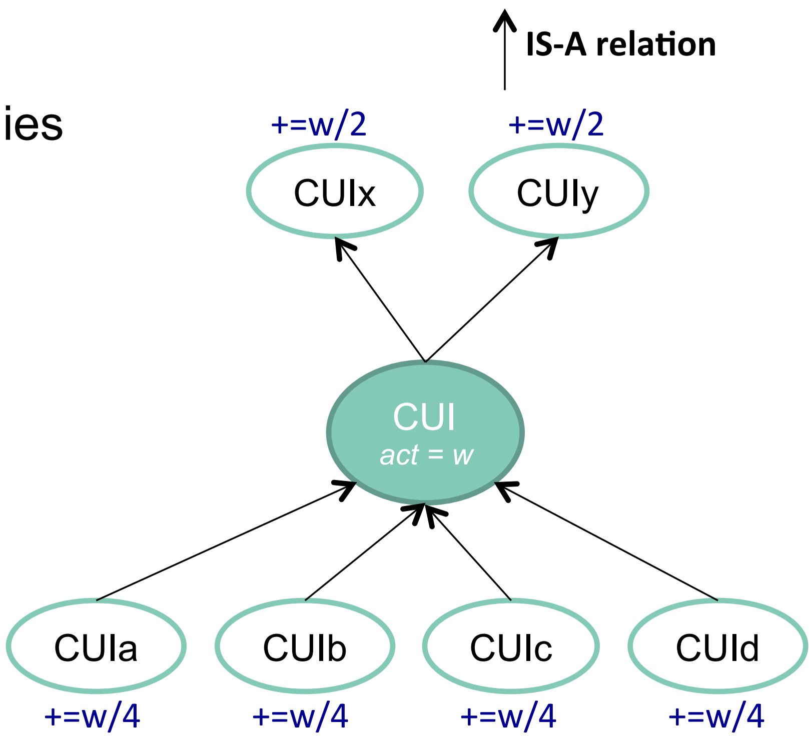
Connecting Clusters to the UMLS Ontology

Differences between IS-A hierarchies and ontologies

- UMLS is more fine-grained than IS-A patterns
- UMLS is less noisy
- UMLS has (almost) no frequency information

Approach

- activate all CUIs in cluster
- REPEAT k times:
  - spread activation downwards from all nodes
  - spread activation upwards from all nodes



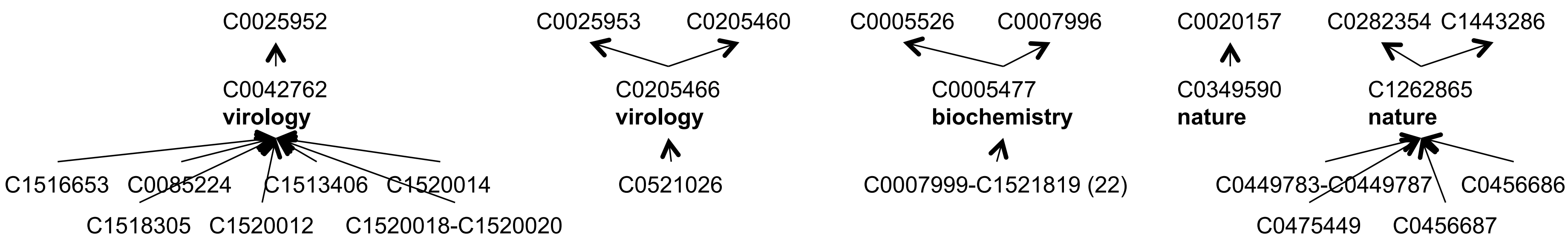
Result: Activations of target CUIs and most activated other CUIs.

Work done in the MRP project, Chris & Alfio, Sept 2012

Connecting Clusters to the UMLS Ontology example

“anatomy” CUIs:

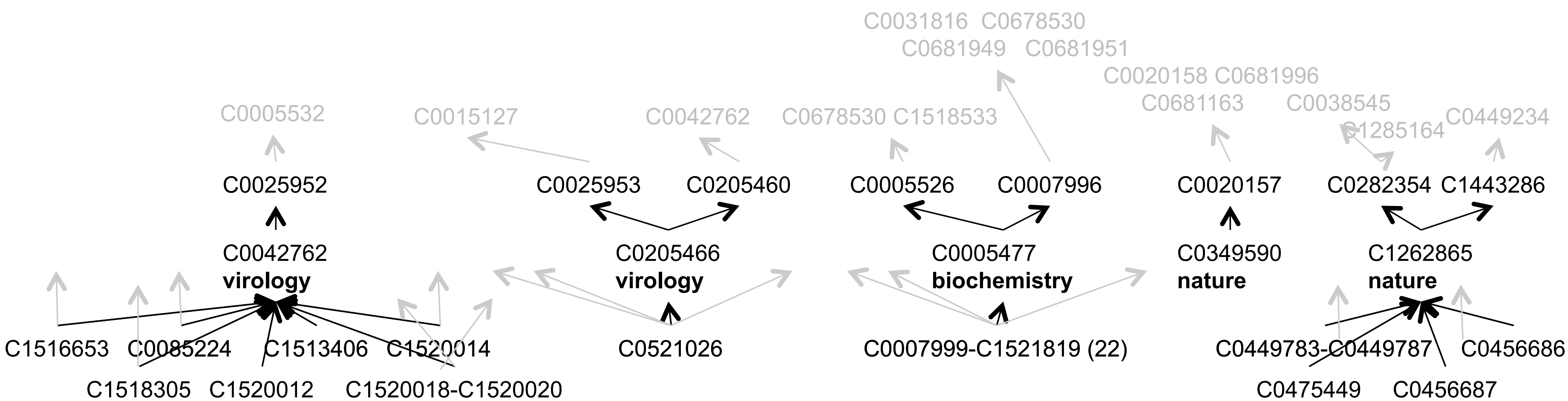
C0002808: Science of Anatomy
C0700276: Anatomic structures
C1384516: anatomy aspects



Cluster 0: virology, biochemistry, nature

Connecting Clusters to the UMLS Ontology example

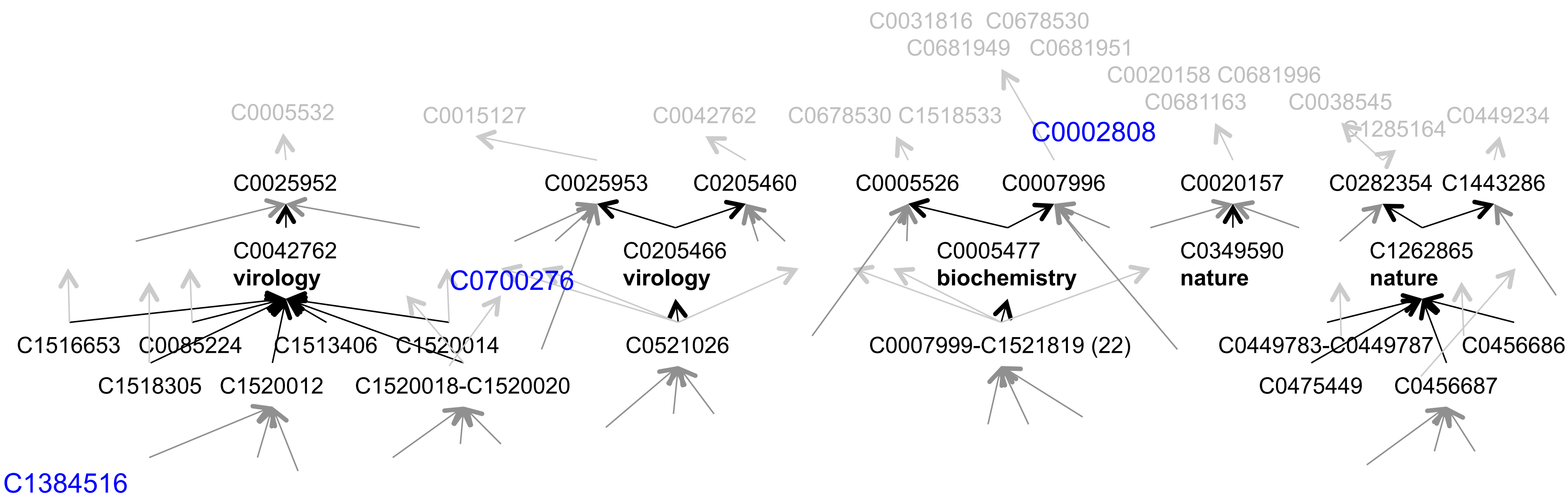
"anatomy" CUIs:
C0002808: Science of Anatomy
C0700276: Anatomic structures
C1384516: anatomy aspects



Cluster 0: virology, biochemistry, nature

Connecting Clusters to the UMLS Ontology example

“anatomy” CUIs:
C0002808: Science of Anatomy
C0700276: Anatomic structures
C1384516: anatomy aspects



Collect activations of
C0002808
C0700276
C1384516

Cluster 0: virology, biochemistry, nature

Anatomy example

"anatomy" CUIs:
C0002808: Science of Anatomy
C0700276: Anatomic structures
C1384516: anatomy aspects

anatomy#0 C0002808 0.175

virology, biochemistry, nature

C0005477: Biochemistry
C0007996: Science of Chemistry
C0005526: Life Science
C0025952: Microbiology

anatomy#1 C0700276 5.720 C0002808 0.286

sized, annular, thigh bone, gluteal, outer, inflamed, flexible, infarcted, involved, fibrous, affected, uninfect, latin =, injured, image2 =, palmar, stenotic, hyperemic, hyperechoic, vulnerable, tibial, removed, contralateral, metatarsophalangeal, inguinal, slender, nearby, fifth, edematous, entire, lined, intact, dorsal, caudal, anterior, larg … upper, medial, delicate, swollen, one or more, proximal, hypoplastic, branchfrom =, hypoechoic, movable, visible, diseased, exposed, irritated, adjacent, undeveloped, supporting, torn, supernumerary, underdeveloped, shaped, weak, translucent, internal, tortuous, inner, scaphoid, avascular, bulky, non-ossifying, pertinent, thin, small, right and left, heterotopic, fragile, curved, accessory, lateral, collagenous, ipsilateral, neighboring, elongated, thick, femoral, … popliteal, carpal, vital, malformed, flaccid, membranous, normal, erythematous, contiguous, lobulated, distant, leaky, deep, nontender, immobile, deformed, damaged, cordlike, incompetent, flat, germinal, anterior abdominal wall, tubelike, pubic, ulnar, straight, soft, dilated, distal, caption2 =, attached, redundant, cadaveric, striated, floppy, congested, posterolateral, twisted, lingual, uninvolved

C0013604: Edema
C1264758: (inactive concept)
C1274013: (duplicate concept)
C1514495: Property or Attribute
C1514623: Modifier / Qualifier
C0282354: Descriptor
C0015811: Bone structure of femur

anatomy#2 C0700276 16.987 C0002808 0.237 C1384516 0.143

cartilaginous, embryologic, surgical, cranial, hilar, bony, capsular, sonographic, extranodal, synovial, anatomical, parietal, irritative, palatal, intrathoracic, intraabdominal, pineal, underlying, genital, tubular, sternal, extrinsic, dural, skeletal, pharyngeal, reproductive, orofacial, paraspinal, ectopic, uterine, sacral, mandibular, anomalous, tarsal, osseous, penile, hepatobiliary, trabecular, anorectal, endocardial, perianal, temporal, central, bronchial, precise, subglottic, epithelial, optimum, malar, scrotal, periarticular, glottic, afferent, ligamentous, extrapulmonary, extraskeletal, … epiphyseal, particular, cortical, temporomandibular, retroperitoneal, fascial, anatomic, aberrant, supraglottic, articular, reticular

C0037303: Cranium Bone
C0521324: Skeletal
C0950133: OMIM (database)
C0229962: Anatomic part

Outline

- Semantic Domains and LSA
- Semantic Domains in NLP
- LSA in Watson
- Distributional Semantics
- Inducing Meaning and Linking to Knowledge
- JoBimText and its application to Watson

JoBimText

Linking Language to Knowledge with Distributional Semantics

- www.jobimtext.org
- Open Source Software
  - Apache License
  - SourceForge
- Contributors
  - TU Darmstadt, Germany, FG Language Technology
    - **Chris Biemann** (**Bim**) , Martin Riedl
  - IBM T.J. Watson Research - Watson Technologies
    - **Alfio Gliozzo** (**Jo**) , Michael Glass, Bonaventura Coppola
- What's there
  - Scalable Distributional Similarity (Hadoop)
  - UIMA based text processing implementing @@ operation on different languages/NLP
  - Fast and Scalable Knowledge Management
  - Sense Clustering, WSD, lexical substitution, Thesauri induction, Paraphrasing, Entity Linking, …
  - Machine Learning: CRF, Chinese Whisper Clustering, …

UIMA™ Unstructured Information Management Architecture
An Apache Project

hadoop

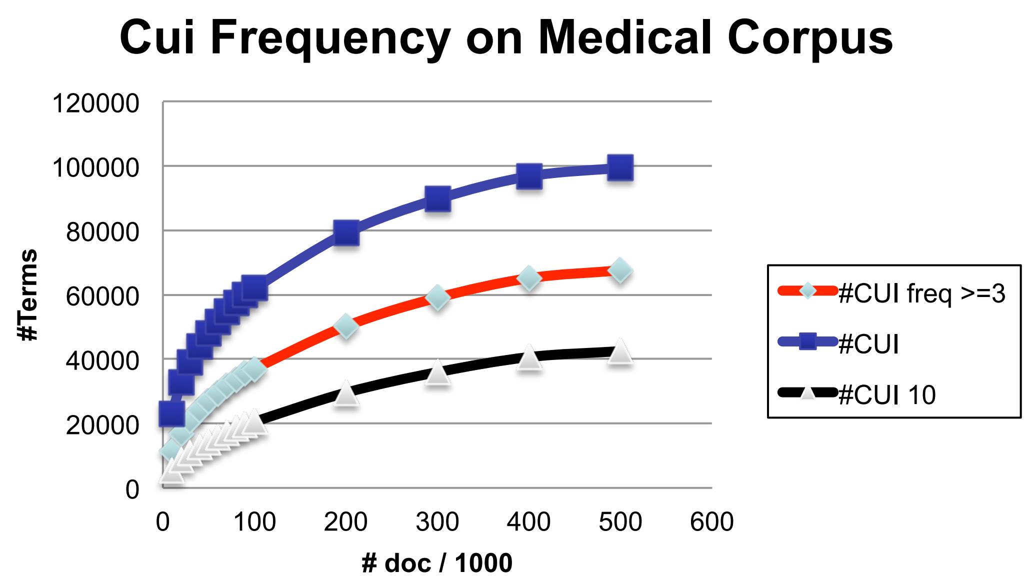
Operationalize JOBim Text in the Medical Domain

- Input:
  - Watson Medical Corpus
    - ~ 2 Gigabytes of text
    - UMLS
- Preprocessing:
  - Medical Extended Slot Grammar (ESG) Parser
    - Dependency Parser
    - Medical Adaptation of the Jeopardy Parser
  - TWREX
    - Relation Extraction system adapted to UMLS relations
- @@ system:
  - Terms are represented by
    - syntactic dependencies
    - TWREX relations
- Unsupervised learning on a Small Hadoop Cluster
- Watson Analytics for Answer Scoring, Matching, Passage Scoring
- Demo

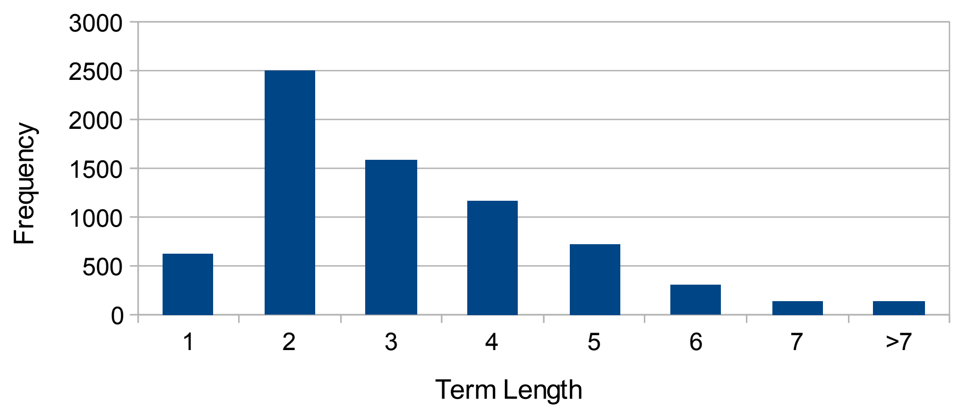
Operationalize JOBim Text in the Medical Domain

IBM WATSON

- For a real application we need multiwords
  - Most domain specific concepts are multiwords
- Large Corpora
  - Multiwords are usually not very frequent
- Semantic Relations provide more informative properties
  - Relation Between Different arguments of PAS



Distribution of Term Length



SENTENCE:

I suffered from a common cold and took aspirin.

Dependency Parser:

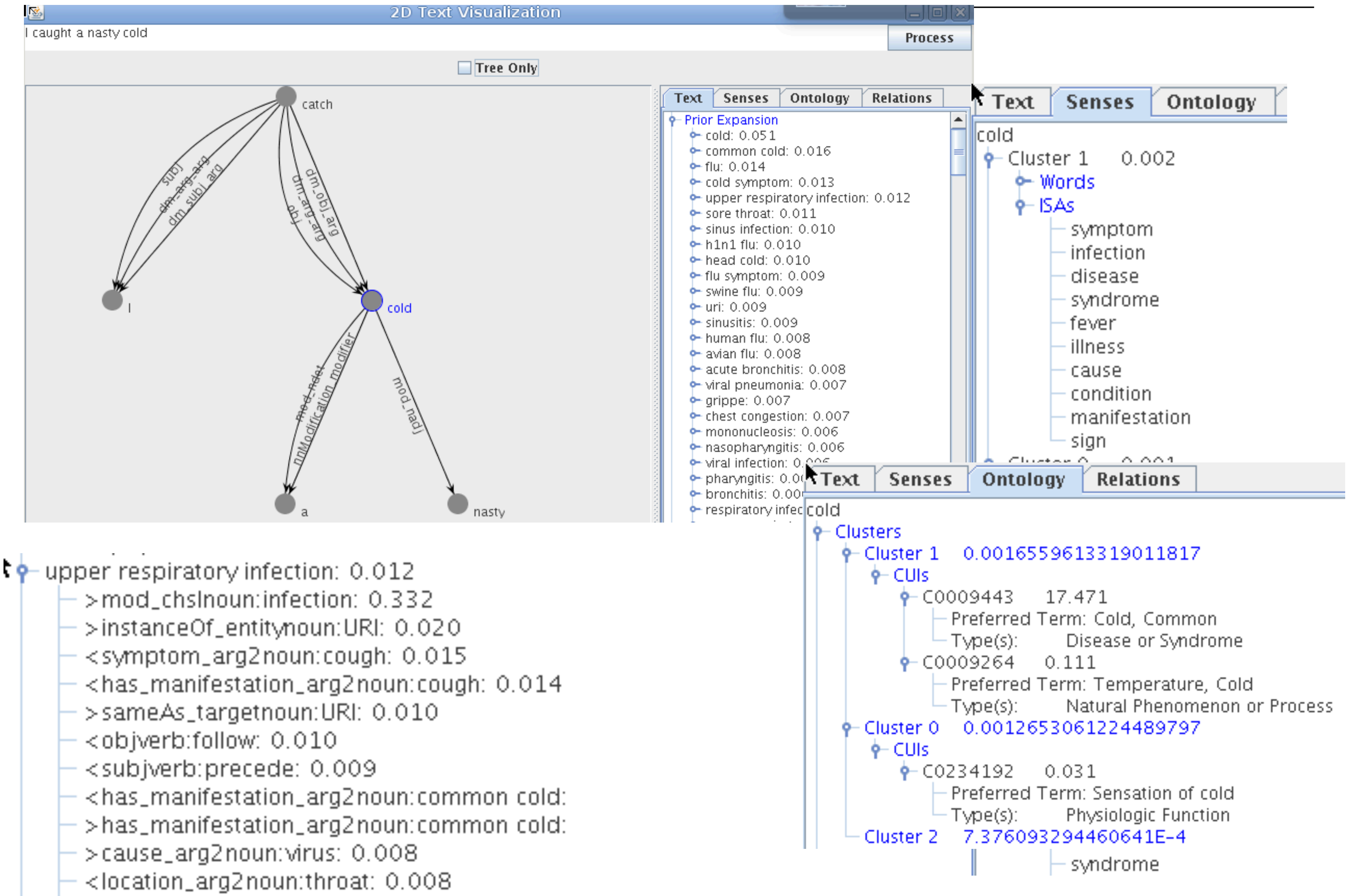
nsubj(suffered, I); nsubj(took, I); root(ROOT, suffered); det(cold, a); prep_from(suffered, common_cold); conj_and(suffered, took); dobj(took, aspirin); may_treat(aspirin, common_cold)

TERMS-Properties PAIRS:

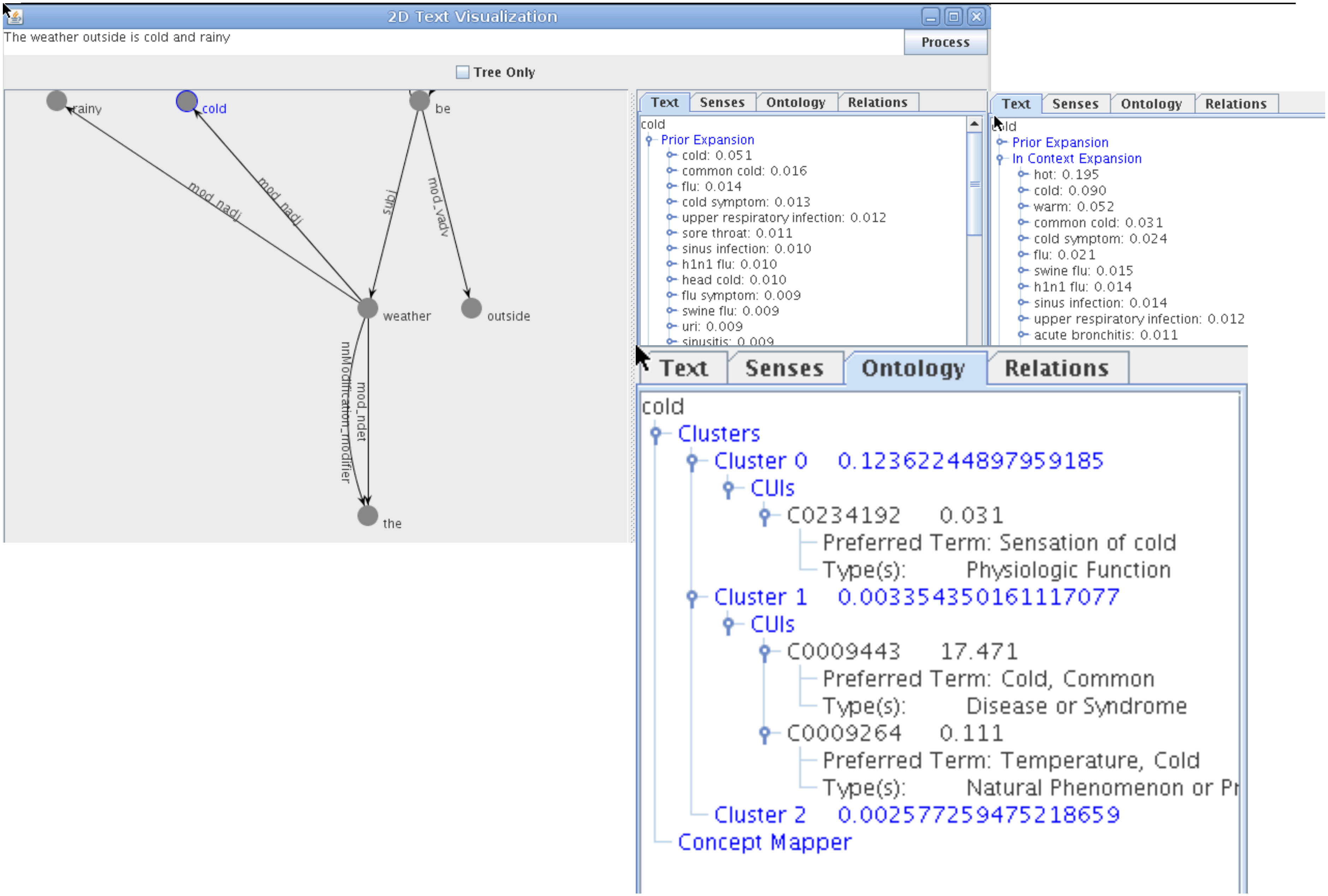
suffered	nsubj(@@, I)	1
took	nsubj(@@, I)	1
common_cold	det(@@, a)	1
suffered	prep_from(@@, **common_cold**)	1
suffered	conj_and(@@, took)	1
aspirin	**may_treat**(@@, **common_cold)**	1

Jo	Bim	
I	nsubj(suffered, @@)	1
I	nsubj(took, @@)	1
a	det(**common_cold**, @@)	1
common_cold	prep_from(suffered, @@)	1
Took	conj_and(suffered, @@)	1
common_cold	**may_treat**(aspirin, @@)	1

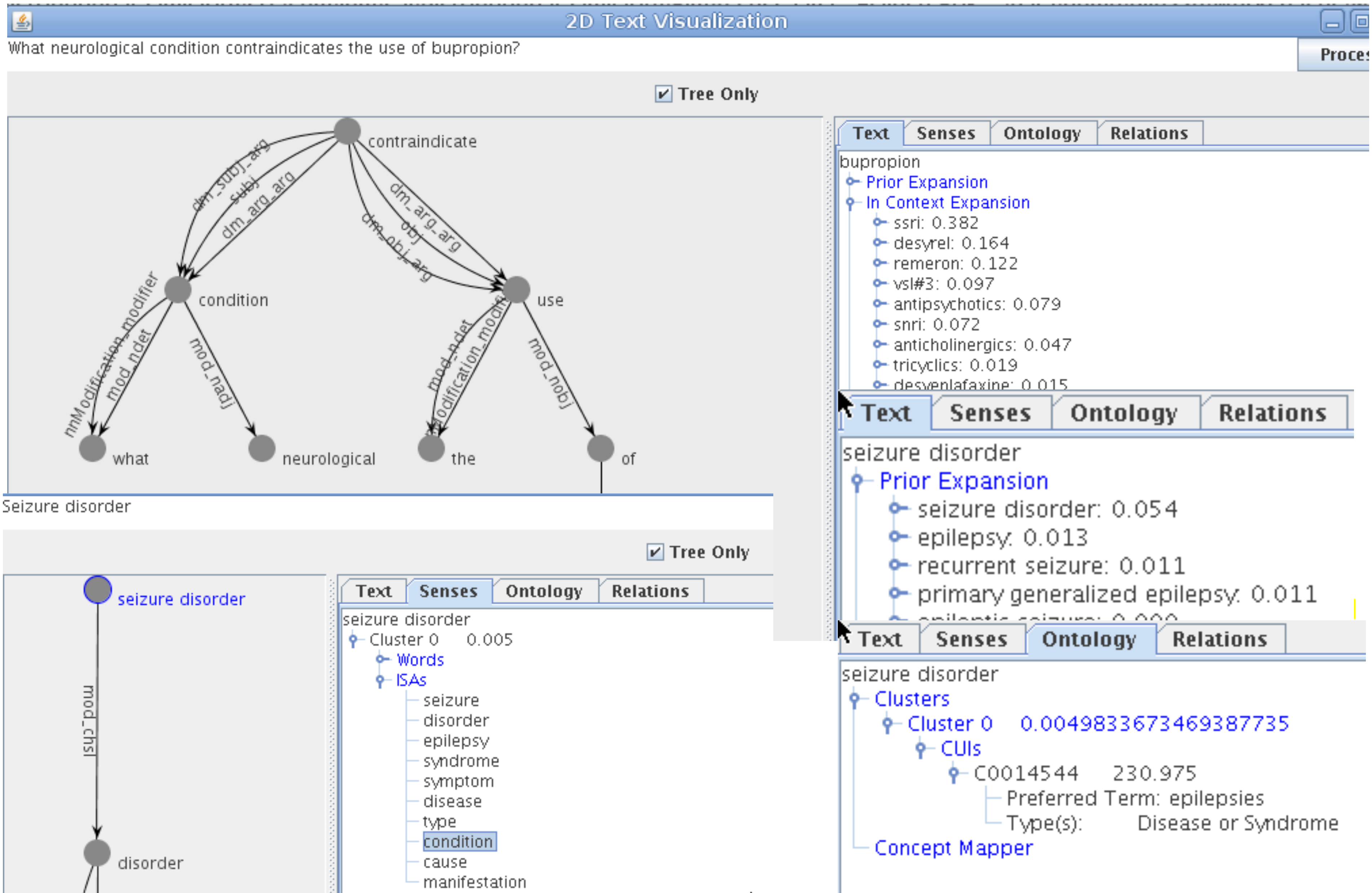
JoBimText Demo



JoBimText Demo



JoBimText Demo: Question Processing



Outline

- Semantic Domains and LSA
- Semantic Domains in NLP
- LSA in Watson
- Distributional Semantics
- Inducing Meaning and Linking to Knowledge
- JoBimText and its application to Watson

Semantic Technologies in IBM Watson™

Lesson 11 – Strategy for Jeopardy!

Guest Lecturer: David Gondek
TA: Or Biran

Strategy

Four Aspects of Jeopardy! Strategy

1) Daily Double wager
   blunders cost ~5-10%

2) Final Jeopardy wager
   blunders cost ~10-20%

3) Clue selection
   no DD seeking costs ~10%

4) Confidence threshold for buzz-in
   blunders cost ~5-10%

Approach to Modeling Jeopardy!

- Modeling J! in great detail would be vastly more difficult than classic board games (chess, checkers, backgammon, Go, etc.)
  - huge amount of imperfect information
  - modeling range of contestant knowledge across many categories would be highly challenging
  - modeling distributions of categories, clues would be equally challenging
- We resort to extreme simplification: models average over all contestants, categories, clues!
  - average stochastic process models for regular clues, DDs, FJ, clue selection
  - keep in mind, we really want to model **humans vs. Watson**, not humans vs. humans

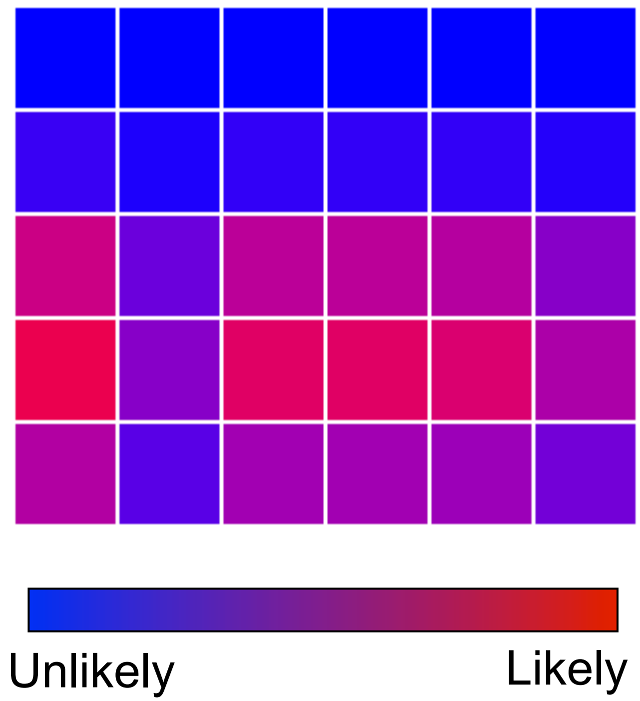
Components of a J! Game Simulator

IBM WATSON

- We obtained detailed historical records of thousands of past J! episodes from **J! Archive (www.j-archive.com)**
- Model **Daily Double placement**
- Model **Human Contestant** performance profile
  - How often they attempt to buzz in
  - How often they are right/wrong when they win the buzz
  - Accuracy and betting patterns on Daily Doubles
  - Accuracy and betting patterns in Final Jeopardy!
- We built three different human models:
  - **"Average Contestant"** model: average over all regular J! episodes (ex-College, Teen, Celebrity ToC games)
  - **"Champion"** model: All-time top 100 player stats
  - **"Grand Champion"** model: All-time top 10 player stats

Modeling Daily Double Placement

- **Statistics over 9k DDs** (3k Round1, 6k Round2):
  - (Widely known) DDs most frequent in the high-value rows (third, fourth, fifth) with harder clues
    - Row frequencies published on J! Archive
  - (Previously unknown) Some columns are more likely than others to have a DD!
    - First column most likely to have a DD
    - Second column least likely to have a DD

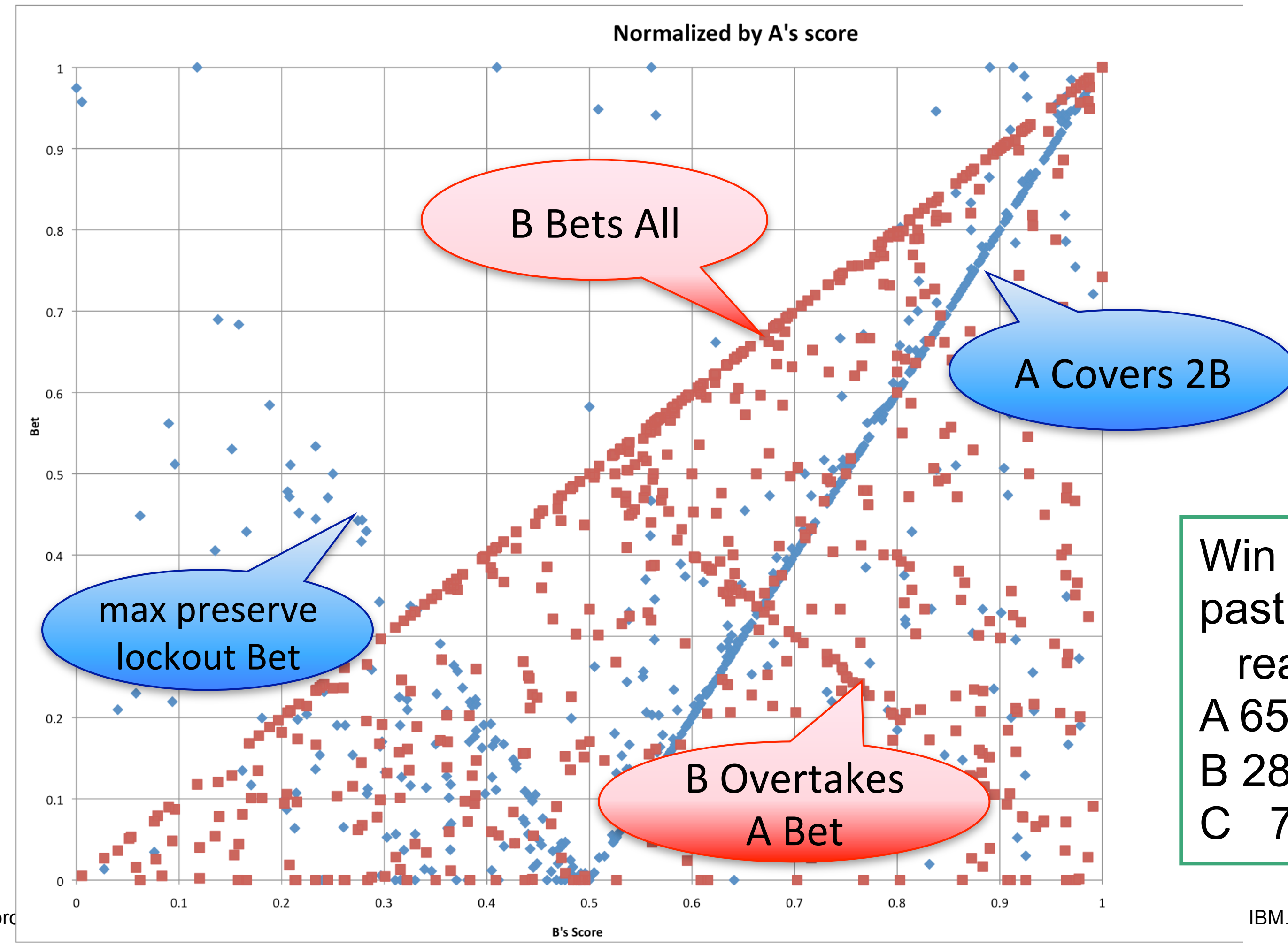


- row-column frequencies used to randomly place DDs in simulated games; Watson uses them as Bayesian prior

Modeling Human Final Jeopardy!

Average FJ accuracy ≈ **50%**   FJ accuracy correlation ≈ **0.3**

Bets depend on score positioning: 1st place (**"A"**), 2nd place (**"B"**), 3rd place (**"C"**)



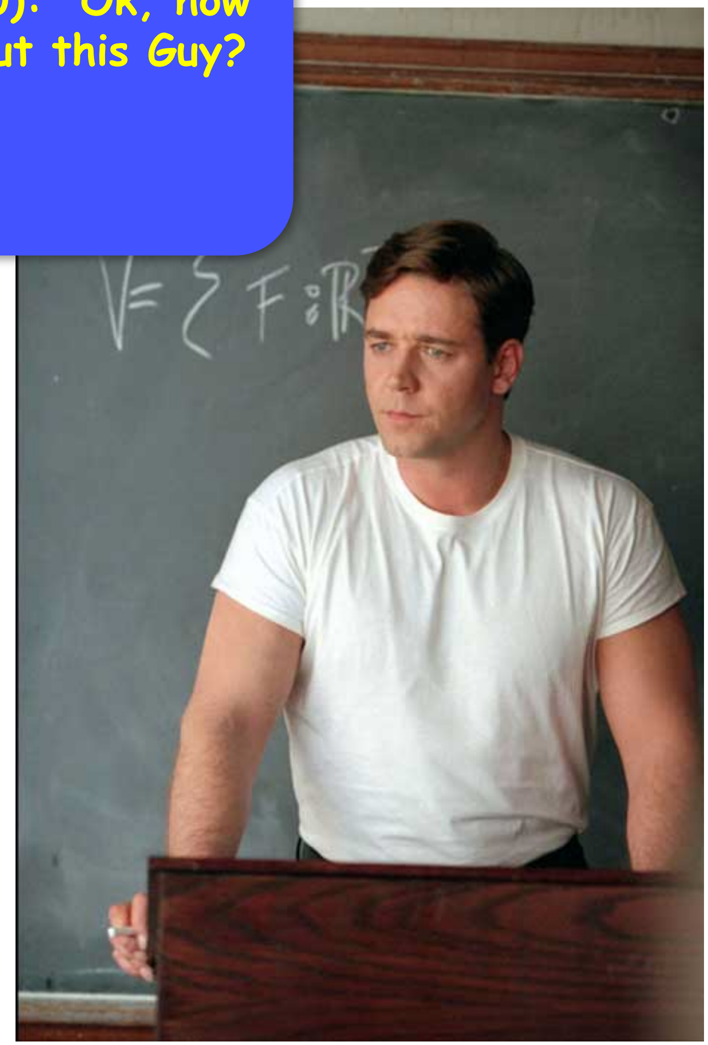
Win rates in 2092 past episodes:

	real	model
A	65.3%	64.8%
B	28.2%	28.1%
C	7.5%	7.4%

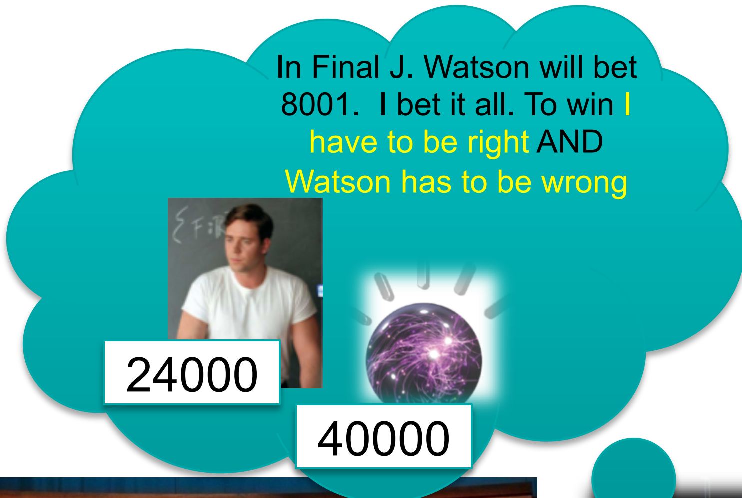
Game Theory: Modeling Equilibria

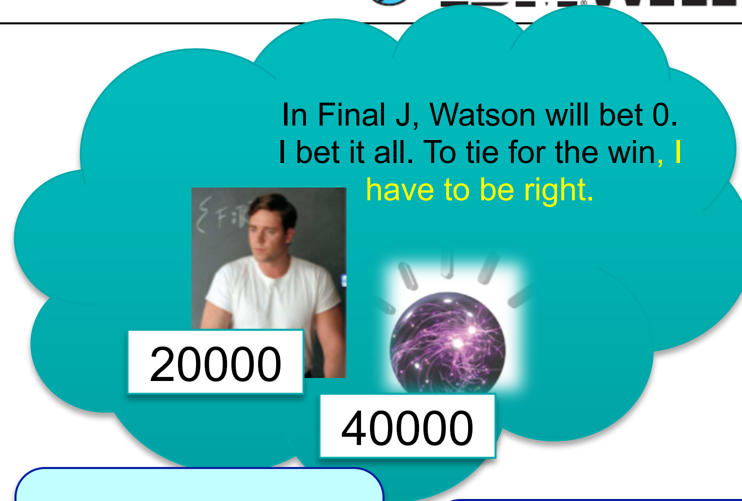
GAME THEORY (700): Who's This Guy?

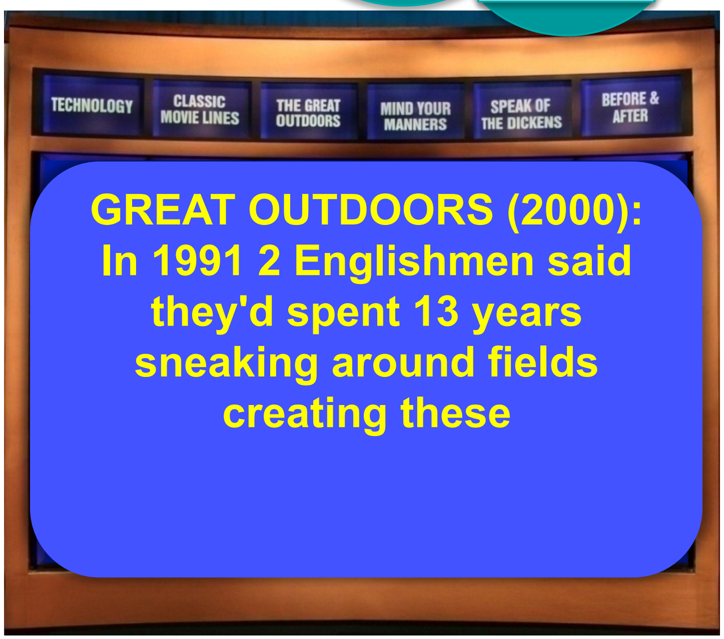
GAME THEORY (700): Ok, how about this Guy?

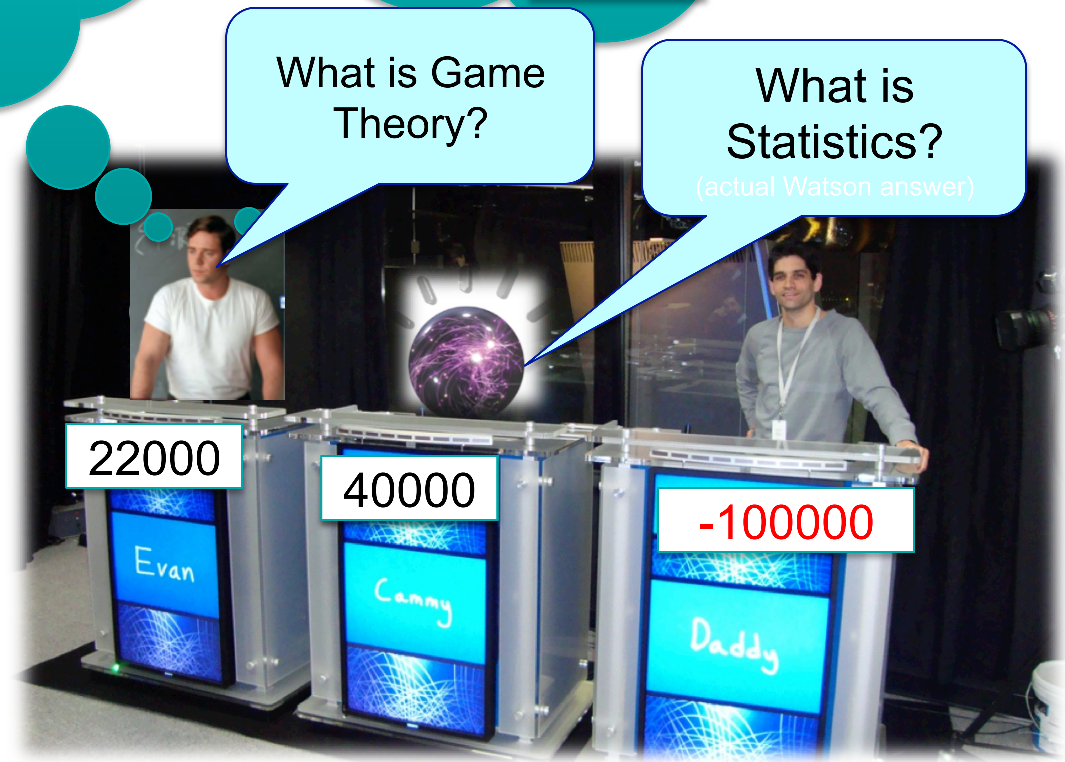


The Jeopardy Challenge: Lock Tie



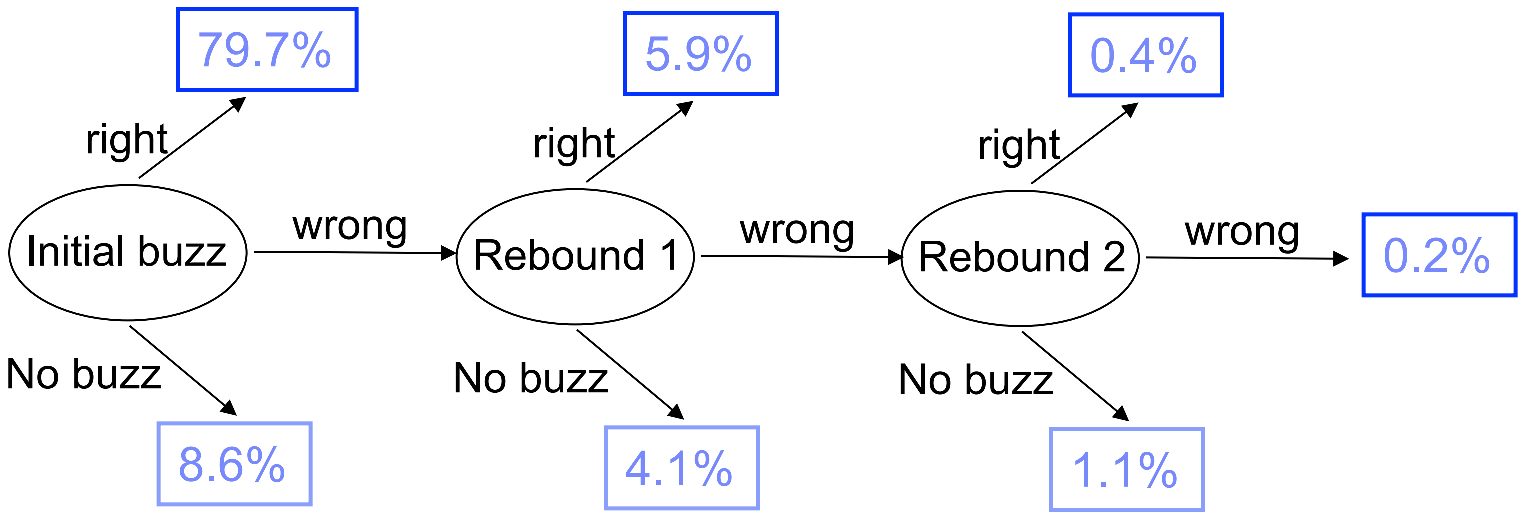






Stochastic Process Model of Regular Clues

Statistics over 150K regular-clue (no DD) outcomes:



Derive avg. contestant precision / buzz rate model:

precision **p = 0.88**

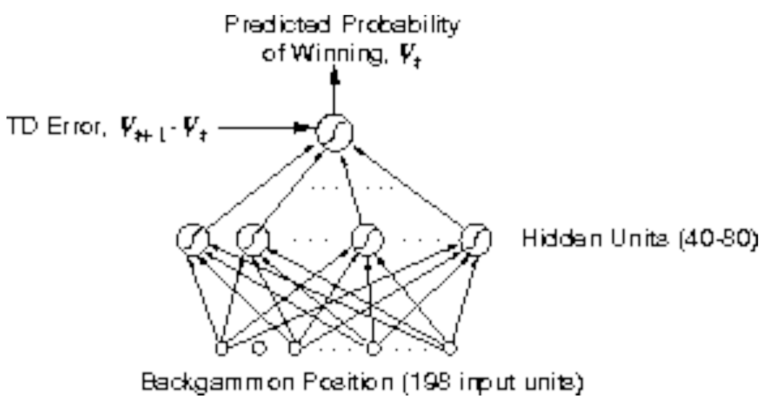
right/wrong correlation ρ_p **= 0.2** (known from rebound stats)

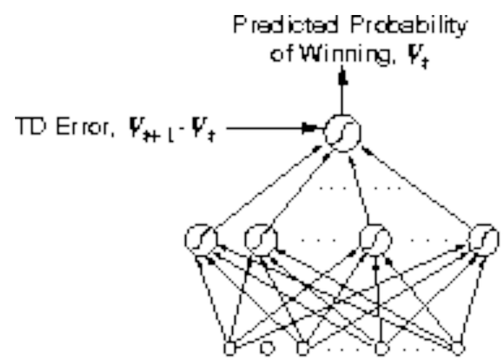
buzz attempt rate **b = 0.61**

buzz/no-buzz correlation ρ_b **= 0.2**

Watson's Daily Double Betting Strategy

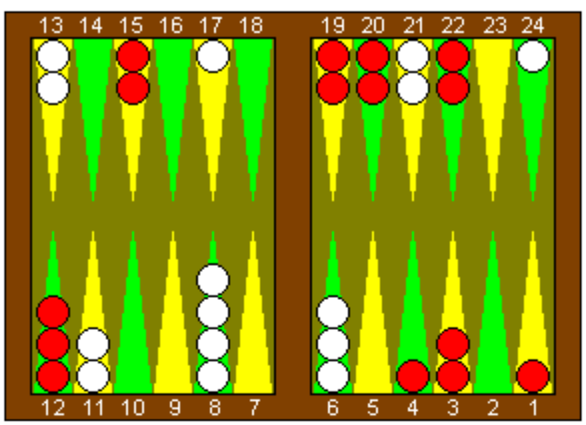
- Train an Aritifical Neural Net over millions of simulated games pitting Watson vs. two simulated human opponents
- Use TD(λ) reinforcement learning algorithm just as in TD-Gammon ☺





20 hidden units

Jeopardy game state (23 input units)



Simplified List of "State" Variables

- Scores of three players
- Round (SingleJ, DoubleJ, FinalJ)
- # of remaining clues
- total $ value of remaining clues
- # of remaining Daily Doubles
- player with control of board

(Watson also has in-category confidence, from right/wrong answers to previous clues in the category.)

Computing Optimal DD Bet

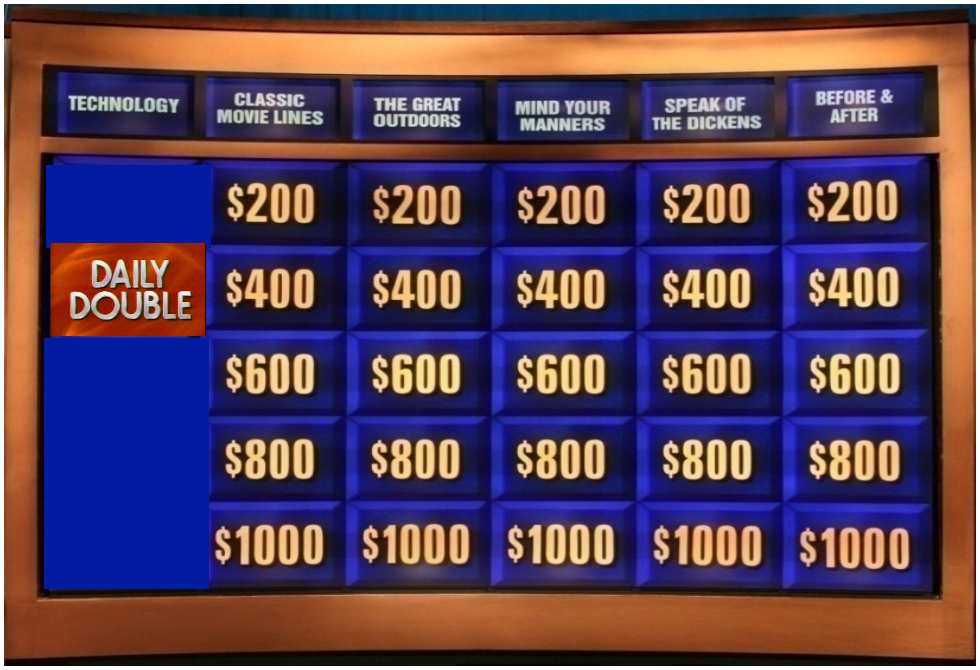
- Let **s** = Watson's score and **V = V(s, …)** = NN output = Watson's win prob in the current game state.
- "Equity" (expected payoff) of a bet:

  E(bet) = conf * V(s+bet) + (1-conf) * V(s-bet)

  where **conf =** Watson's in-category confidence
- Best **risk-neutral** bet maximizes E(bet)
- **Risk mitigation:**
  - Penalize bets with high volatility (std. deviation)
  - Prohibit bets that entail "too much" downside risk
  - significantly reduces risk, only costs 0.3% equity

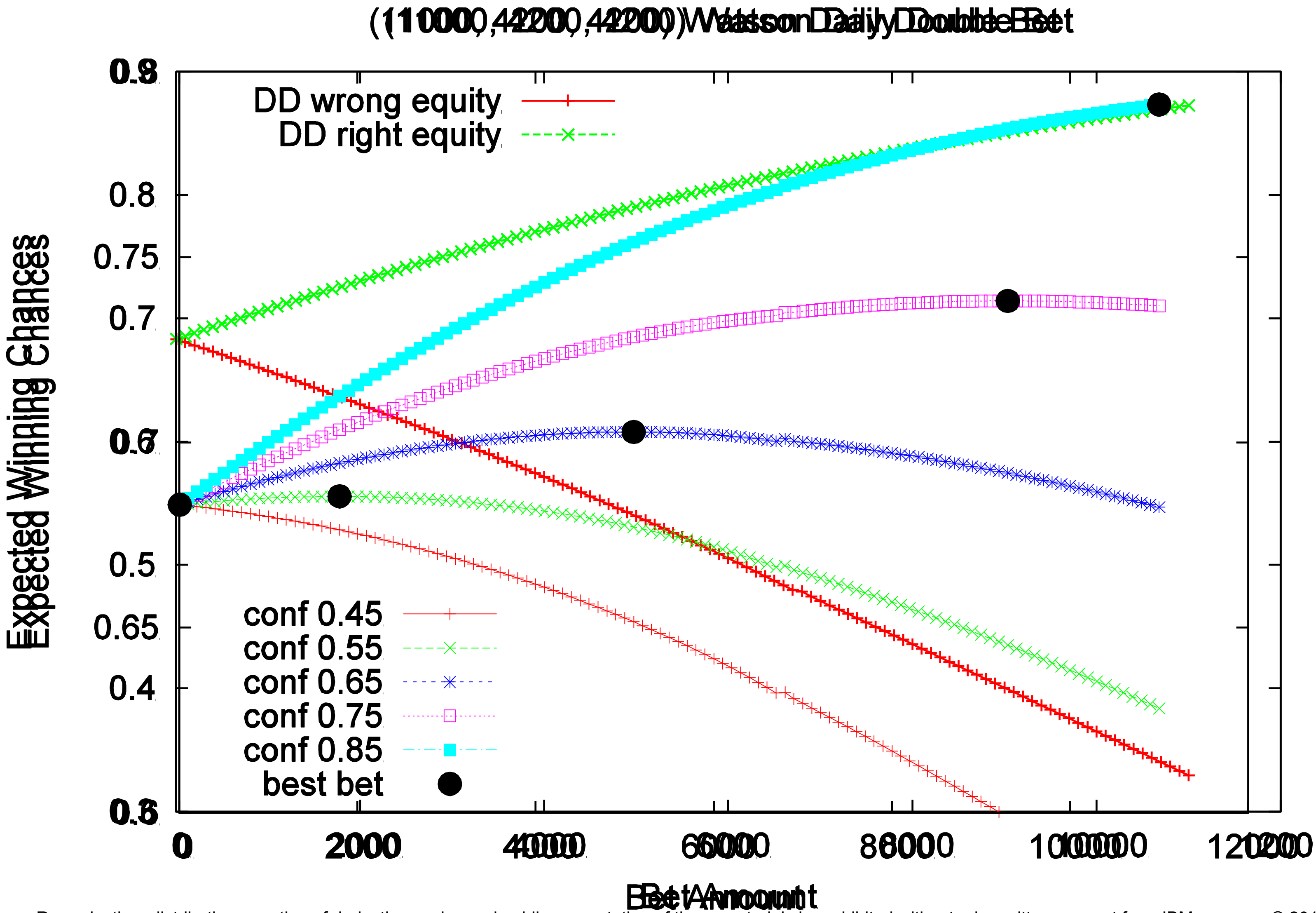
Illustrative Example of NN DD betting

- Start of DoubleJ, Watson ran the column and then found the first DD. Watson leads (11000, 4200, 4200).



- Watson bet = $6700 (!)

NNDD Analysis



Watson's Final Jeopardy! Betting Strategy

- Live Best-Response to Randomized Human FJ Model:
  - Analytic probabilities of the eight possible right/wrong triples
  - Draw ~10k samples of human bet pairs
  - For each legal Watson bet, compute prob (Watson wins) given the bet pair and the right/wrong probs

- Can extract logical betting rules from Best-Response output:

```
IF (B [illegible] 2A/3) AND (B < 2C) {
        IF( (2C-B) [illegible] (3B-2A) )  THEN BET = 2C-B
        ELSE BET = B
}
```

Win rates in 2092 historic FJs:

	human	BR
A	65.3%	67.0%
B	28.2%	34.4%
C	7.5%	10.4%

Watson's Clue Selection Strategy

- Systematically studied trade-offs between three considerations:
  - 1. Finding Daily Doubles:

    -- use historical DD locations as Bayesian prior

    -- combine prior probs with revealed clue evidence using Bayes' rule to obtain posterior DD probs
  - 2. Keeping control of the board:

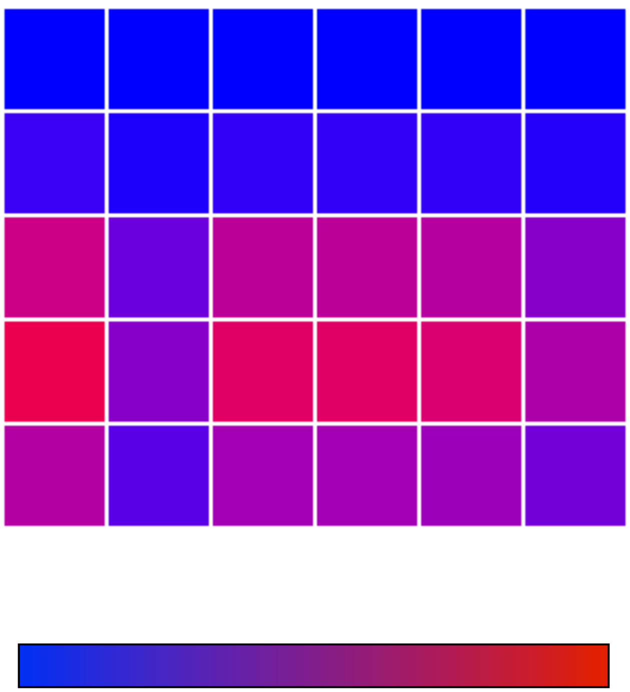
    -- tend to stay in categories where Watson is doing well
  - 3. Learning the "gist" of a category from revealed clues

    -- tend to pick low-value clues, to do better on high-value clues

- Resulting strategy that maximizes simulation win rate:
  - If DDs left, ~90% #1 and ~10% #2
  - If no DDs left, 100% #3.

Daily Double Seeking Animation

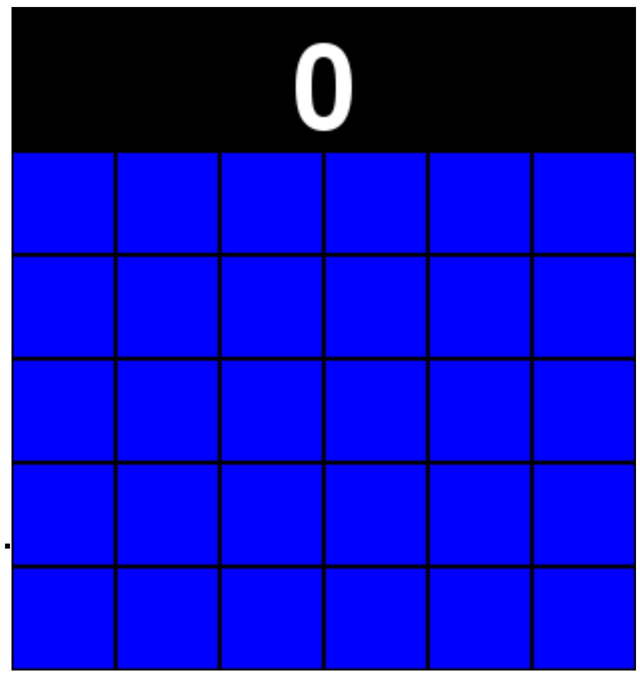
Watson



Unlikely

Likely

Watson finds DDs in 65% of the time it takes humans.



Square Selection: Jeopardy

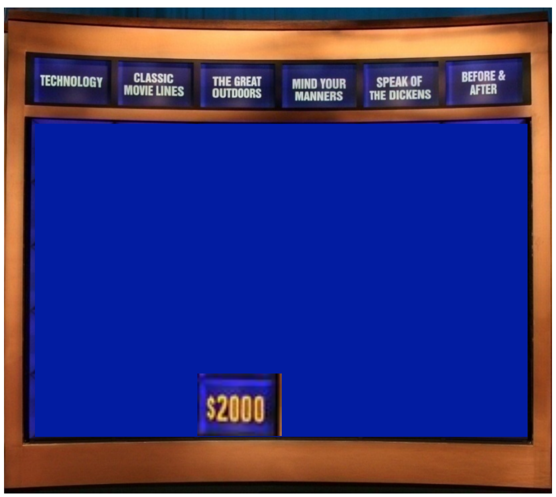
It is somewhat unintuitive, but A DD is much more likely to exist behind the square in the 4th row and 5th column than in the 4th row and 6th column.

Square Selection: Jeopardy

It is somewhat unintuitive, but A DD is much more likely to exist behind the square in the 4th row and 5th column than in the 4th row and 6th column.

Elements of Strategy: Buzz-in Threshold

- In most typical situations the buzz-in threshold is set to 0.5
- At the end of a game,



– When Watson is close to mathematically locking out its opponents it will shift its threshold to try to achieve or retain lock-out.

Ken	**Watson**	**Brad**
12,000	25,000	4000

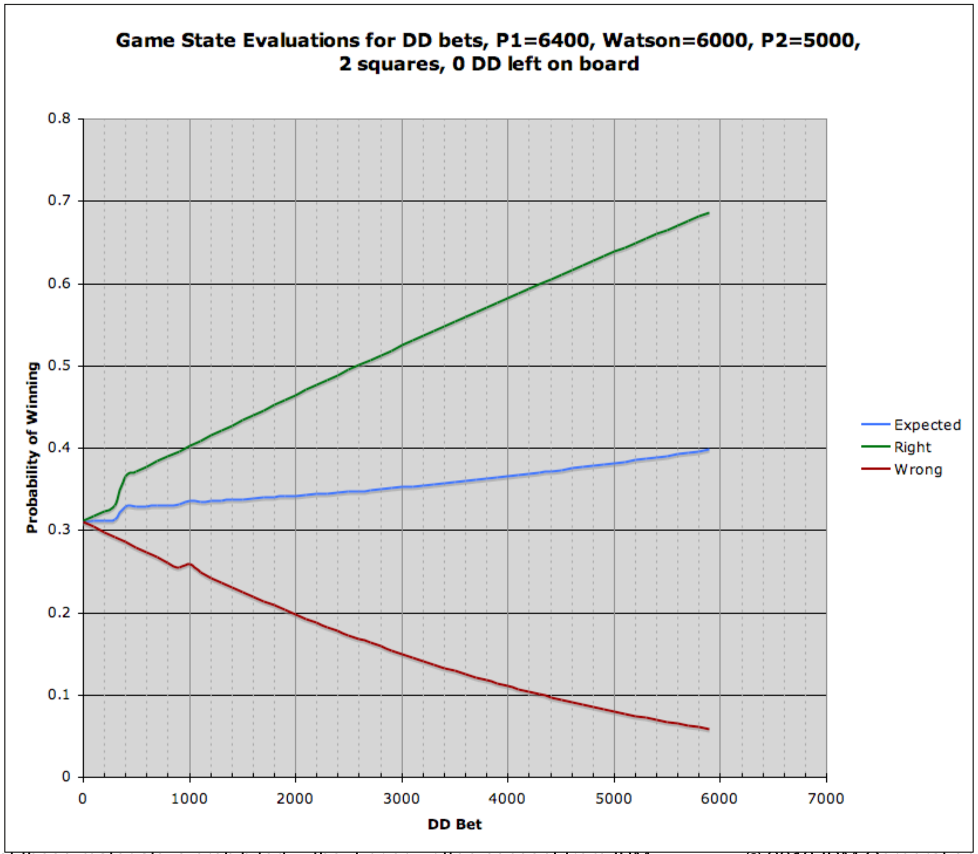
– If Watson is close to **getting** locked out, it will generally lower the threshold

Ken	**Watson**	**Brad**
25,000	12,000	4000

- Desperation situations that got me interested in Strategy!
- Humans tend to give up when they fall far behind, not adopting the appropriate amount of desperation – e.g. more freely buzzing in, since it is the only hope

Elements of Strategy: Daily Double Betting

- Keys to Watson's Daily Double betting strategy:
  - Very accurate estimate of what its chance of winning, given any **game state**
  - Estimate of the probability of Watson getting the clue right in the category



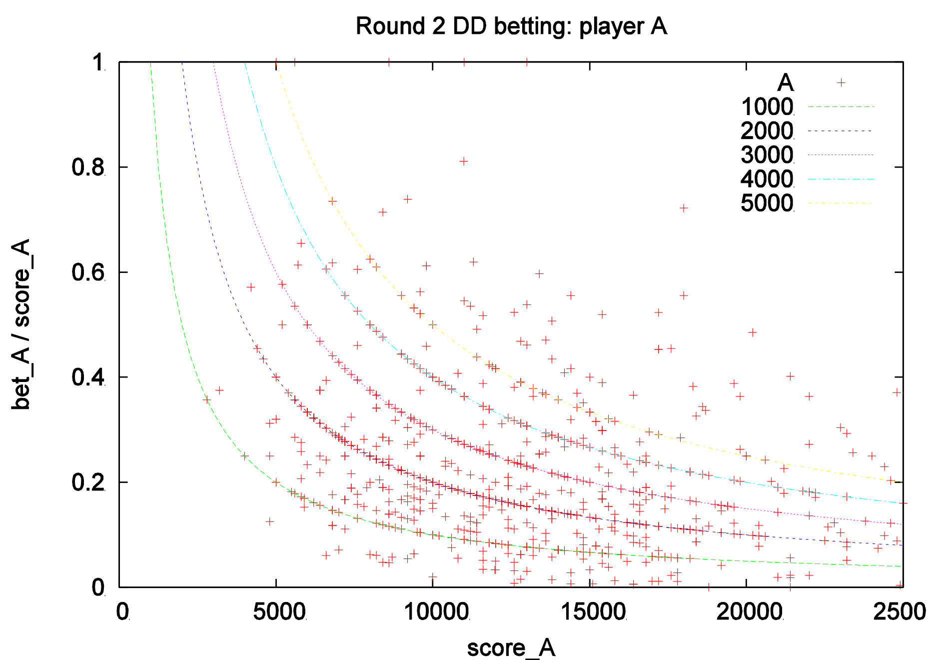
- In 2-game Tournament, 1st Final Jeopardy is similar to DD betting, with complication of modeling opponents behavior

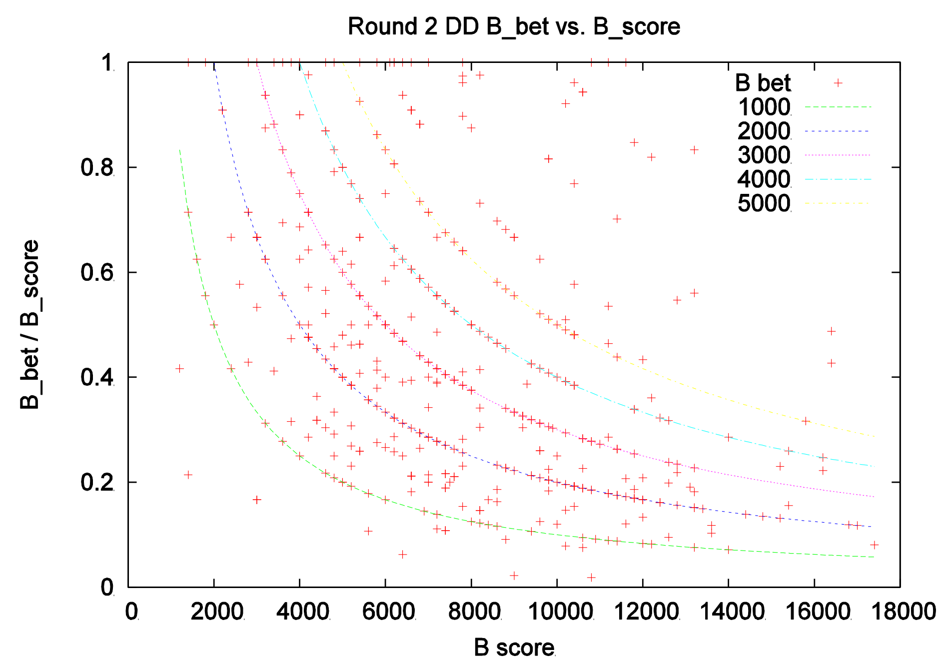
IBM WATSON

Modeling Human DD Accuracy/Betting

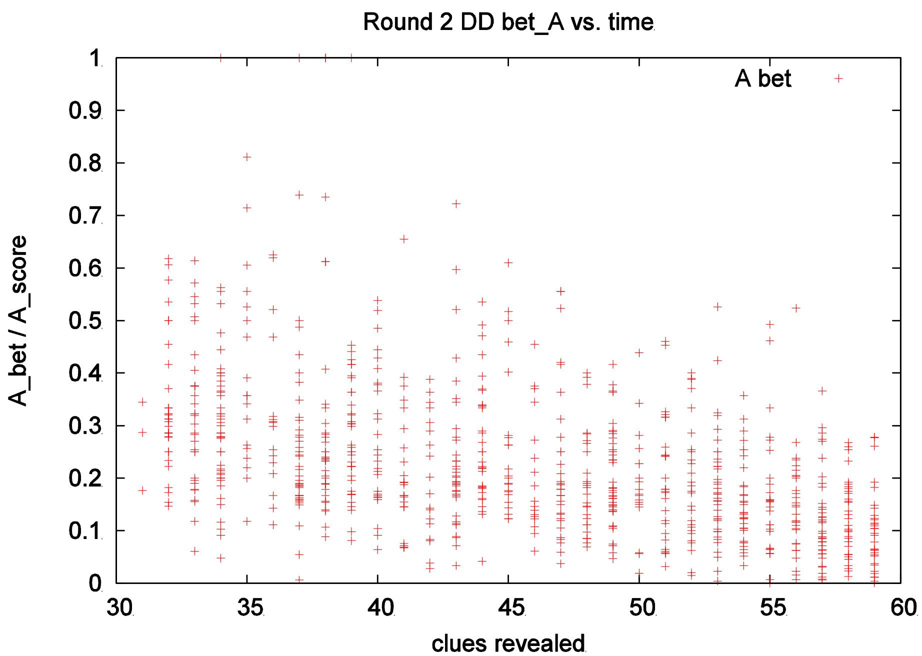
Average DD Accuracy [illegible] **64%**

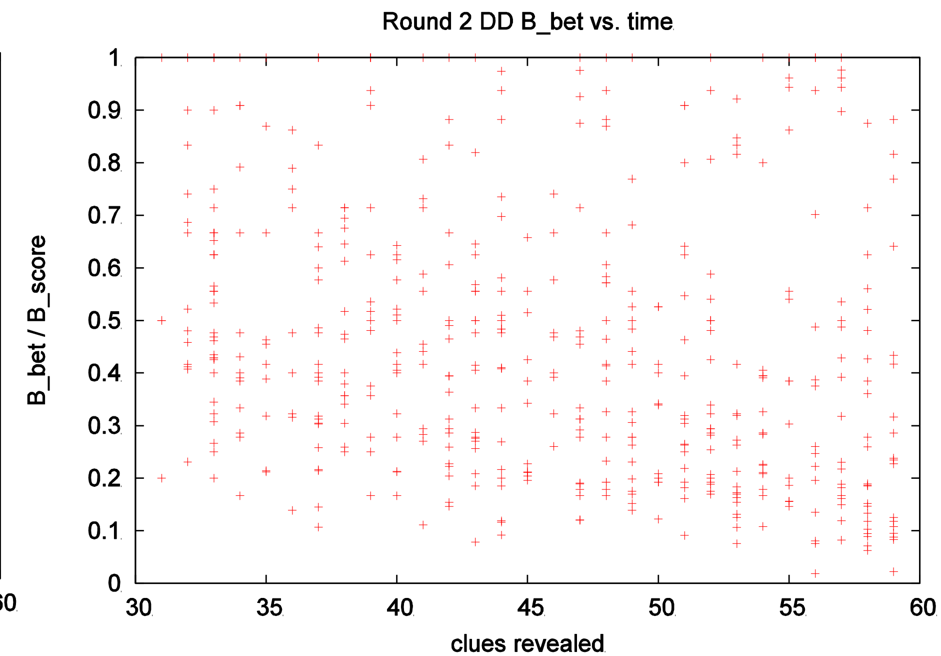
- Main observations of human DD betting behavior:
  - tend to bet round numbers, rarely [illegible] 5000
  - leader (A) becomes very conservative toward end of Double J!
  - B and C also become conservative, but to a much lesser degree





Betting vs. Time Remaining





- When playing vs. Watson, strong players quickly learn to bet DDs very aggressively!

Modeling Human Clue Selection

- Mostly boring tendencies:
  - top-to-bottom within a category (strong)
  - left-to-right across categories (weaker)
  - Stronger players may jump around the board ("Forrest bounce") and hunt DDs in the lower rows

- When playing vs. Watson, strong players quickly adopt DD seeking behavior