

2005 Semantic Web Conference - Galway, Ireland

Semantic Acceleration or "The Practical Web"

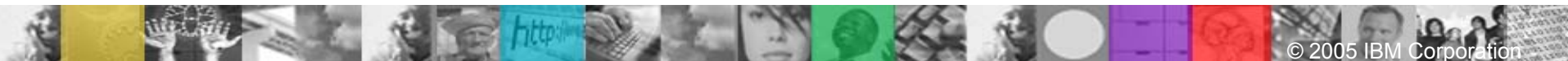
(Due credit for this talk to Chris Welty and the UIMA team)

***Dr. Alfred Spector
Chief Technology Officer
IBM Software Group***



Outline

- **Innovation and Semantics**
- **Semantic Web**
- **The Challenge**
- **The Opportunity**
- **Unstructured Information Management Architecture:
UIMA (!)**
- **Connections to the Semantic Web**
- **Successes to date**
- **Opportunities and the IBM Innovation Grants**
- **Conclusions and Summary**



Semantic Acceleration or “The Practical Web”

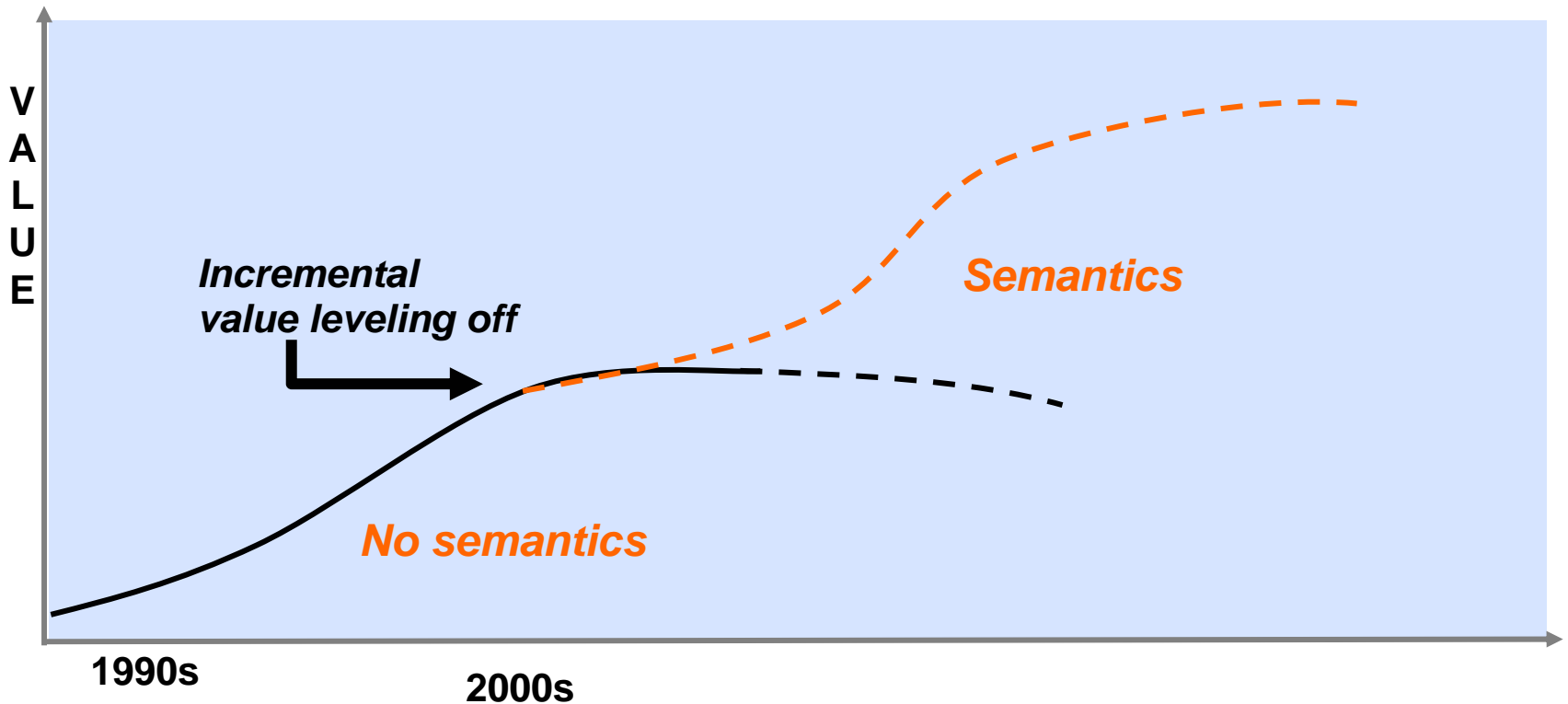
Abstract

The Semantic Web envisions a future where applications (computer programs) can make sense and therefore more productive use of all the information on the web by assigning common "meaning" to the millions of terms and phrases used in billions of documents. AI and knowledge representation must rise to the occasion and work with decentralized representations, imprecision and incompleteness. Standard web-based representations are an essential enabler and we have made good progress in their design. But we still rely on humans to assign semantics and here there is a big leap of faith: The World Wide Web has grown at startling rates because humans are prolific at producing enormous volumes of unstructured information, that is, information without explicit semantics; on the other hand navigating this mass of information has proven to be both possible and profitable to the point that there is a \$6 B search advertising industry. It's is not practical to expect the same will automatically happen for semantically enriched content. And yet we need semantics to better leverage the huge value on the web.

The Practical Web is about confronting this challenge. Its about realizing that we will need to automate the assignment of semantics to unstructured content to ultimately realize the vision of the Semantic Web. If well done the results will be synergistic with the motors of web expansion: user value and commercial value.



Information Semantics will Drive Greatly Increased Value ...in Virtually Every Domain.



Semantic Web

The **Semantic Web** provides a common framework that allows **data** to be shared and reused across application, enterprise, and community boundaries. It is a collaborative effort led by W3C with participation from a large number of researchers and industrial partners. It is based on the Resource Description Framework ([RDF](#)), which integrates a variety of applications using XML for syntax and URIs for naming.

"The Semantic Web is an extension of the current web in which information is given well-defined meaning, better enabling computers and people to work in cooperation." -- *Tim Berners-Lee, James Hendler, Ora Lassila, [The Semantic Web](#), Scientific American, May 2001*

From the W3C Semantic Web home: <http://www.w3.org/2001/sw/>



Unstructured versus Structured Information: *What does it mean?*

Structured Information:

Semantics of information captured in DB schema

Name	Occupation	Organization	Age	Office Location
Jones	Engineer	IBM	29	San Francisco
Carbonell	Professor	CMU CSD	39	The Burgh.
Brown	CEO	Textract	42	New York

Unstructured Information:

Semantics inherent in usage and context

The associated press reported today that, Ms Jones, an Engineer at IBM has been recently spotted at the Summit meeting in Zurich,.....At, 29, Ms. Brown, Is the youngest CEO at the Summit,...



Wherefrom the Semantics?

- **Some will be manually created**
- **Some web content generated from existing databases**
 - Structured, but semantics often hidden
 - Still may requires efforts to harmonize, extend, declare, & expose
- **However, most web and enterprise data contains only latent structure**
 - Manual markup hard –perhaps even impossible– to scale
 - Therefore, automated and semi-automated methods required



Analytics: The Promise and the Challenge

- Independently developed
- From an increasing # of sources

- Different technologies & interfaces
- Highly specialized & fine grained

Analysis Capabilities

- Language, Speaker Identifiers
- Tokenizers
- Part of Speech Detectors
- Document Structure Detectors
- Parsers, Translators
- Named-Entity Detectors
- Face Recognizers
- Relationship Detectors
- Classifiers ...

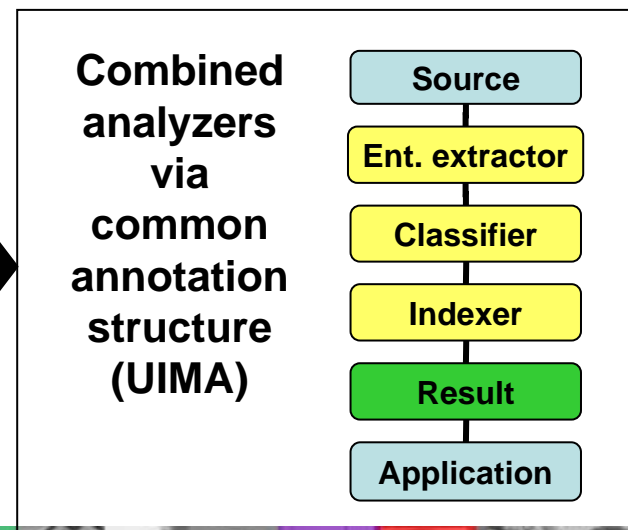
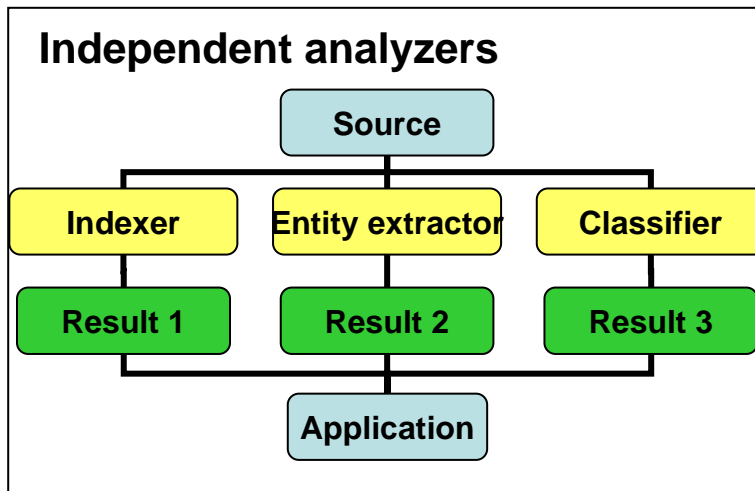
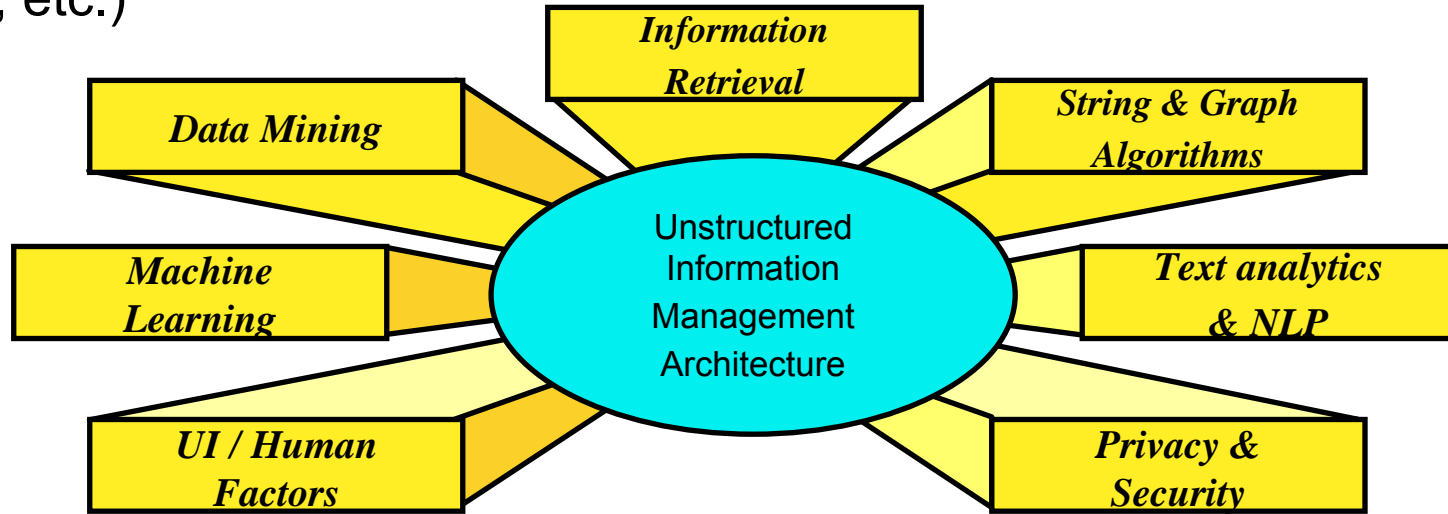
Capability Specializations

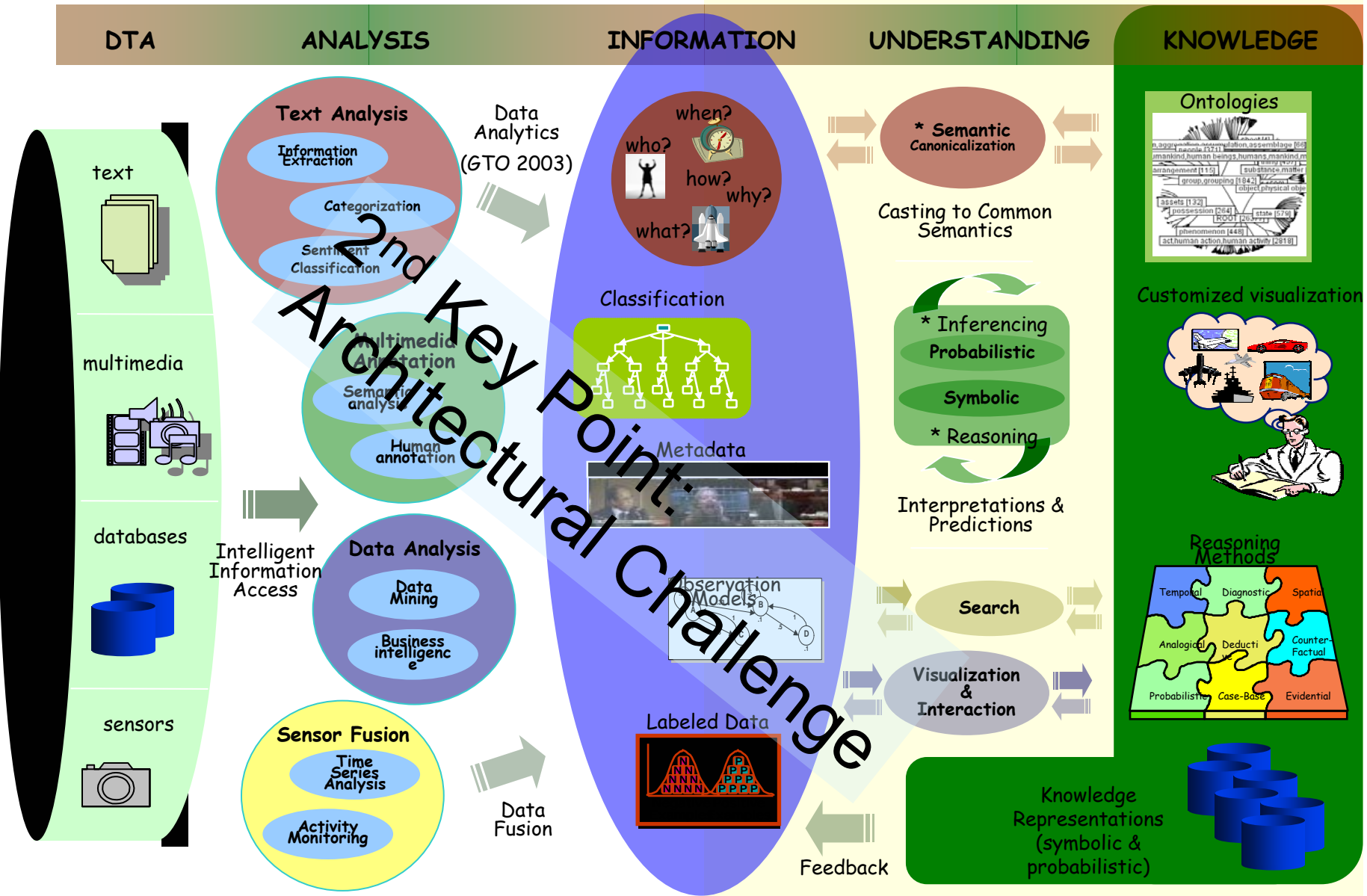
- Modality
- Human Language
- Domain of Interest
- Source: Style and Format
- Input/Output Semantics
- Privacy/Security
- Precision/Recall Tradeoffs
- Performance/Precision Tradeoffs...

The right analysis for the job will likely be a best-of-breed combination integrating across many dimensions.



Key point: The **combination hypothesis**: If intimately integrated, various KM technologies will provide higher quality results (accuracy, recall, etc.)





UIMA: The Project

▪ Start

- IBM Research, Watson and Worldwide beginning 2001
- An internal project to accelerate Research and Technology Transfer
- *And to bring order out of our own chaos ☺*

▪ Focus

- Text and multi-modal analysis integration and component reuse in support of Information and Knowledge Management products and solutions

▪ Requirements

- Text, video and speech analysis
- Advanced (concept/semantic) search and knowledge representation and reasoning

▪ Architecture

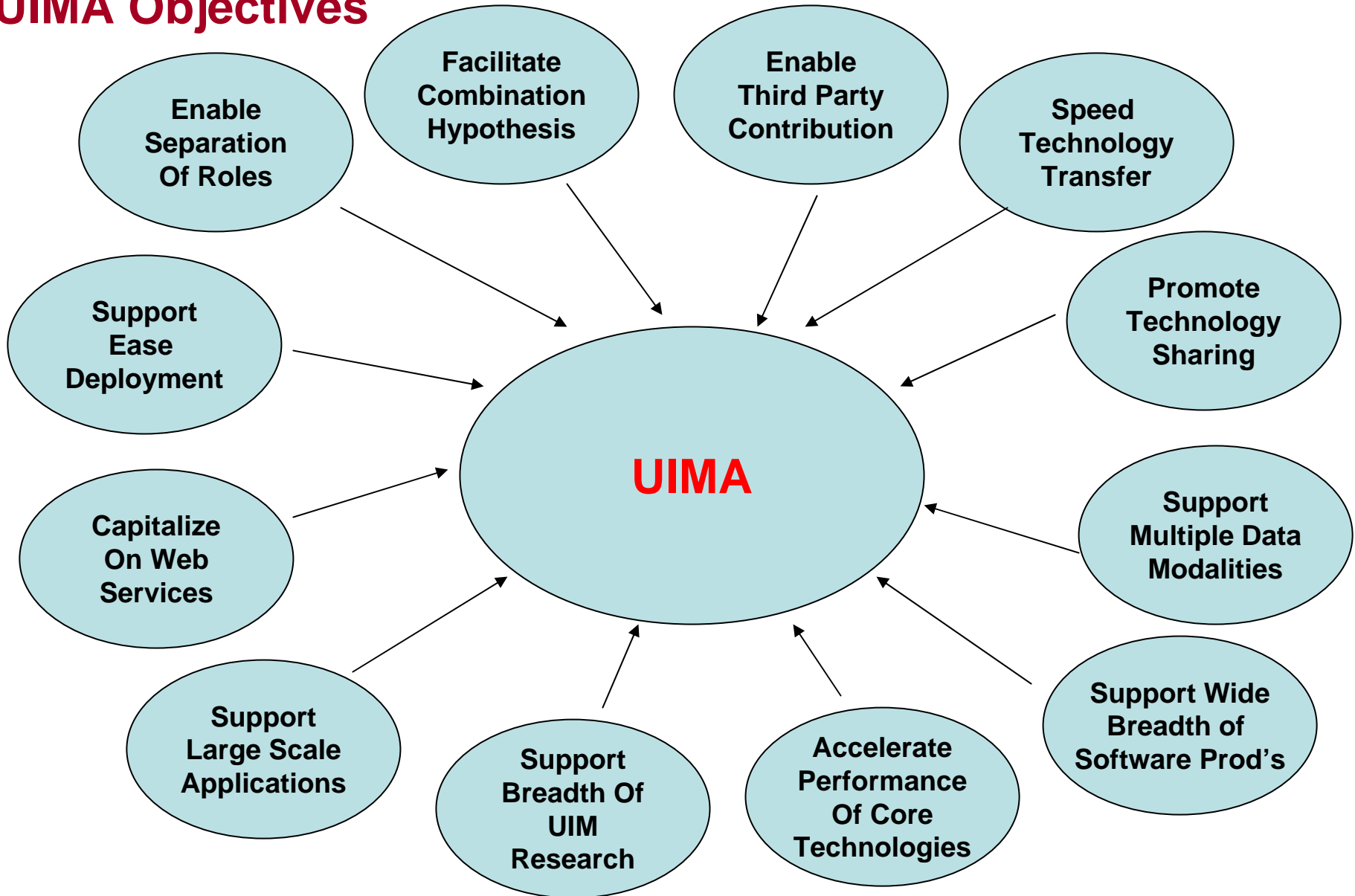
- Informed by TIPSTER, Catalyst, Atlas, GATE, TAF, Talent, WebFountain
- Modern software engineering approaches

▪ Individuals involved

- David Ferrucci, Arthur Ciccolo, Andrei Broder, and many more



UIMA Objectives



UIMA Defined

- **Architecture for composing analytics that extract knowledge from unstructured sources & integrate results with structured information**
 - Interfaces, Data Representation Schemes, Design Patterns
- **Principal Architectural Commitments**
 - Common representation scheme
 - Common component engine interfaces (task and domain-independent)
 - Common component metadata
 - Pluggable Workflow
 - Pluggable Transports
 - Embeddable
- **Independent of but interoperable with**
 - Specific data models
 - Specific algorithms
 - Specific Language-level or domain-level concepts or tools
 - Specific workflows or workflow engines
 - Specific Back-end Systems (DB, Search Engine, KB Interfaces)

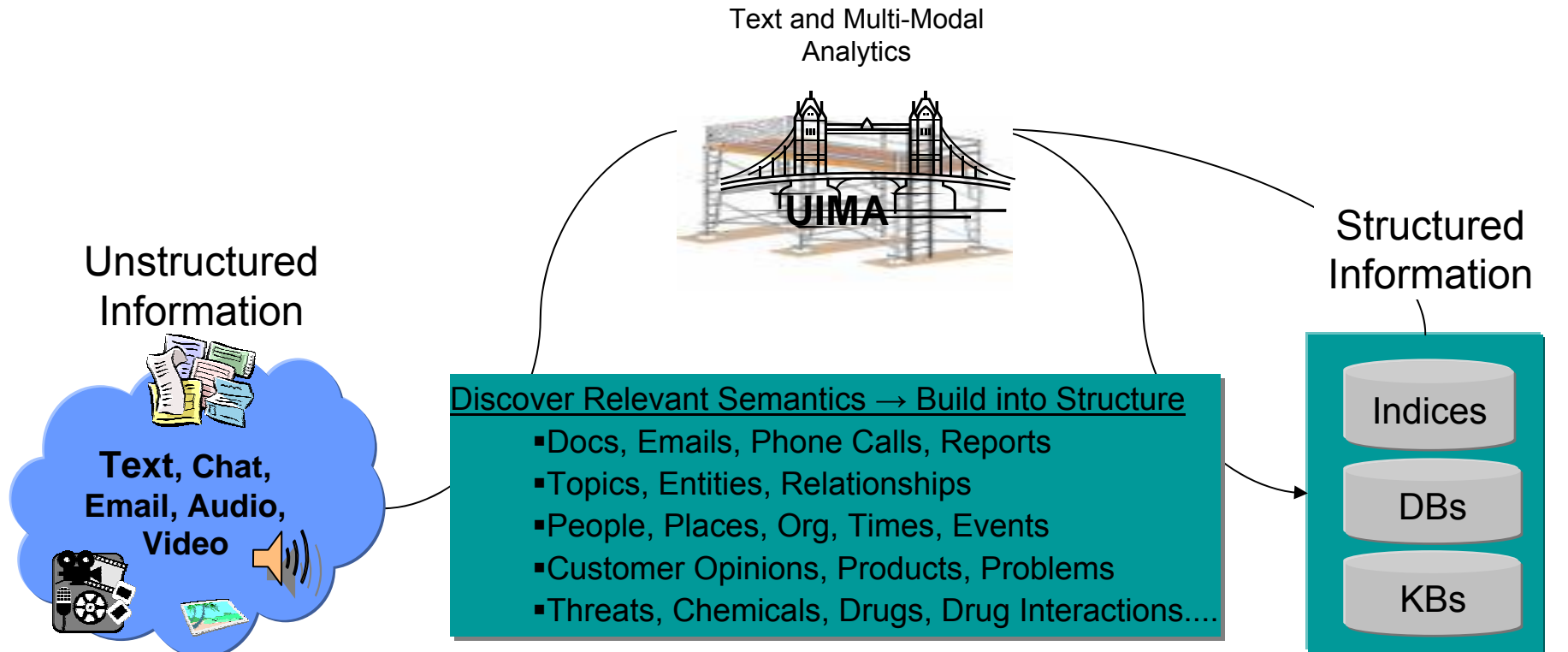


UIMA: The Software

- **Supports UIMA-compliant development, composition & deployment**
- **Java and C++ framework implementations**
 - Analytics in other languages possible through service-based interfaces
- **Support for co-located and service-oriented deployments**
- **Support for specialized APIs to common data representation**
- **UIMA SDK (Software Development Kit)**
 - Stand-alone Java Install
 - Freely Available from IBM alphaWorks
 - Includes Tutorial and Development-Level Utilities and Tooling
 - Ships with a “Semantic Search” Engine and CAS Indexer
 - Core framework goes open-source by end of 2005



Analytics Bridge the Unstructured & Structured Worlds



- **High-Value**
- **Most Current**
- **Fastest Growing (80% of Corporate Data)**

...**BUT** ...

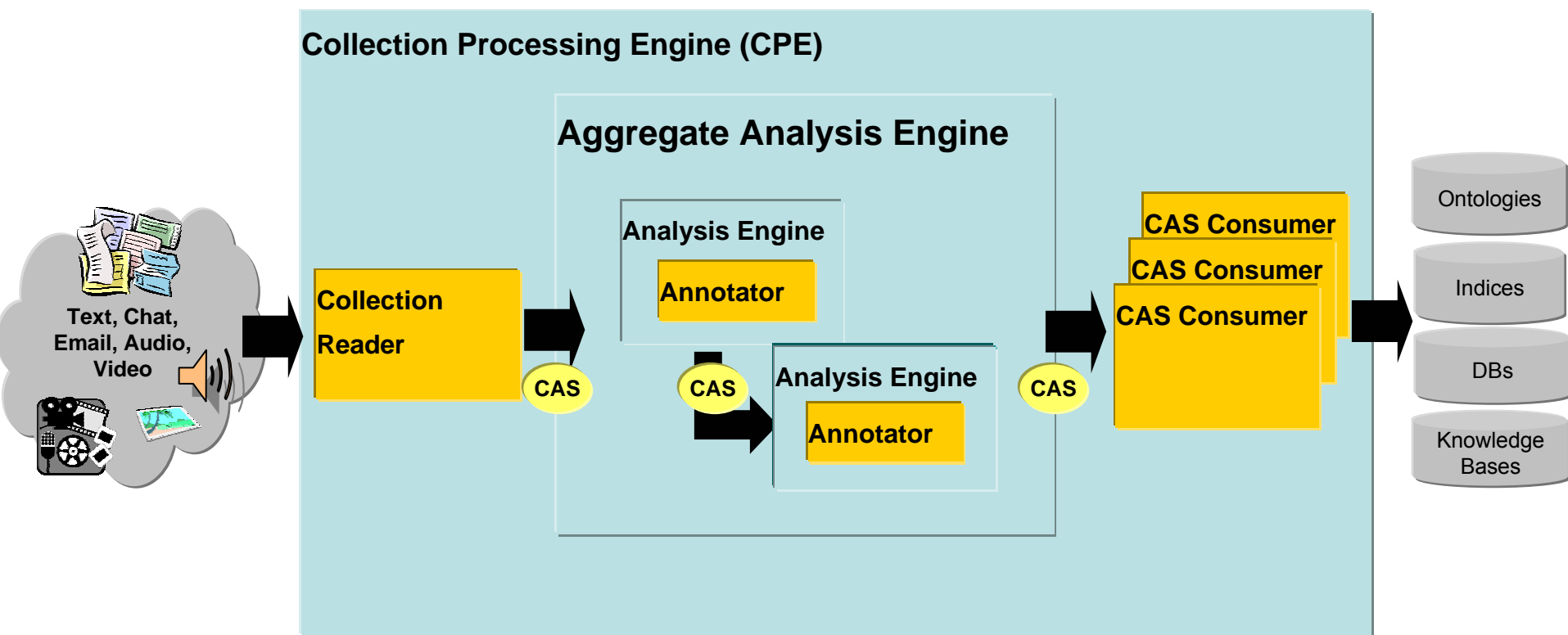
- **Buried in Huge Volumes (Noise)**
- **Implicit Semantics**
- **Inefficient Search**

- **Explicit Semantics**
- **Efficient Search**
- **Focused Content**

...**BUT**...

- **Slow Growing**
- **Narrow Coverage**
- **Less Current/Relevant**

UIMA High-Level Analytic Component Architecture



UIMA Annotation Viewer

Report Date 10 March 2003. Slick business dealings keep local olive oil importer out of the pits. Robert Crane was recognized by local business leaders for his skill at leading the Gorman Food Importers Inc. to strong profits while others are struggling. Mr. Crane, owner of Gorman Food Importers Inc., has consistently been able to produce exceptional results, while still keeping a focus on his employees. Gorman Food Importers Inc. has been in business since 1970 and specializes in food imports from the Middle East, including olive oil and figs. Gorman Food Importers Inc. is headquartered in NYC, and their warehouse is located in Paramus, NJ. The company employs 659 people in the two locations. Robert Crane can be reached at 608-703-2317.

Click In Text to See Annotation Detail

- Organization ("Gorman Food Importers Inc.")
 - begin = 185
 - end = 211
 - componentId = ACE
 - mentionType = NAME
- Organization ("Gorman Food Importers Inc.")
 - begin = 185
 - end = 211
 - componentId = IBMEAnnotator
 - mentionType = NAME

Legend

- | | | | | |
|--|--|--|--|--|
| <input checked="" type="checkbox"/> Person | <input checked="" type="checkbox"/> Facility | <input checked="" type="checkbox"/> GPE | <input checked="" type="checkbox"/> Organization | <input checked="" type="checkbox"/> Location |
| <input checked="" type="checkbox"/> GeneralStaff | <input checked="" type="checkbox"/> BasedIn | <input checked="" type="checkbox"/> Management | | |

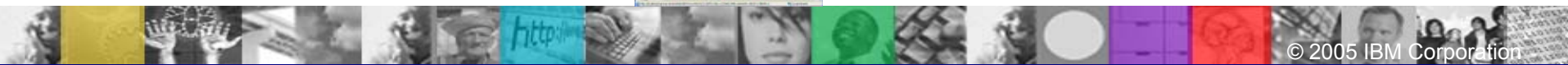
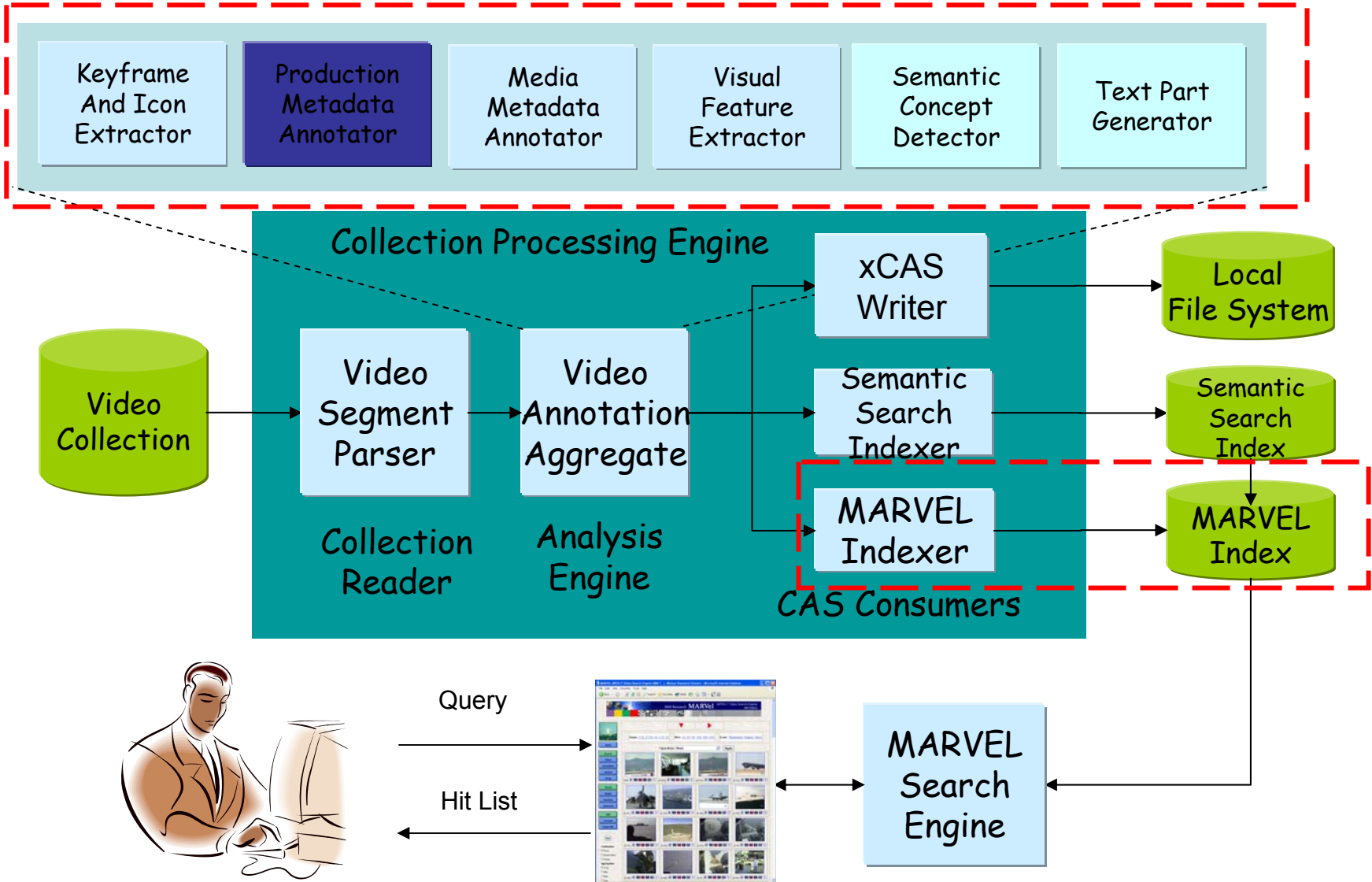
A CAS

- Analyzed by a combination of Analysis Engines
- Semantic Entities & Relations Represented
- Highlighted here in a GUI



UIMA Pipeline for Video Concept Detection & Indexing

Video segments about Basketball,...skiing, vehicles



UIMA Pipeline Video segment

Keyframe
And Icon
Extractor



MARVEL MPEG-7 Video Search Engine (IBM T. J. Watson Research Center) - Microsoft Internet Explorer

File Edit View Favorites Tools Help

Back Search Favorites Media

IBM Research **MARVEL** MPEG-7 Video Search Engine 2005 Edition

Zoom: 0.5x, 0.75x, 1x, 1.5x, 2x Hits: 12, 24, 48, 100, 200, 500 Icons: Thumbnails, Frames, Shots

Operation: None Apply

Local intranet

http://localhost/cgi-marvel-bin/WebCBR03.exe?8.COLS=28TD=08L=12548.COMB=18AIGCA=38CP=238DM=2



UIMA Hierarchy

- JediTutorial
 - Java classes of JediTutorial
 - Primitive Analysis Engines
 - Aggregate Analysis Engines
 - Collection Processing Engines
 - Type Systems
 - CAS Consumers
 - CAS Initializers
 - Collection Readers
 - Corpora
 - ap2000
 - tutorial
- UIMAWC

```

public class RoomNumberAnnotator
    extends Annotator_ImplBase
    implements ITextAnnotator {

```

Outline

- com.ibm.uima_examples.me
 - import declarations
 - RoomNumberAnnotator
 - mPatterns : Pattern[]
 - mBuildings : String[]
 - process(JCas, ResultSpe
 - initialize(AnnotatorConte
 - typeSystemInit(TypeSy
 - reconfigure()
 - destroy()

Overview

Implementation Details

Implementation Language: C/C++ Java

Engine Type: primitive aggre

Information

Name: Room Number Annotator

Version: 1.0

Vendor: IBM

Description: An example annotator that searches for room numbers in the IBM Watson research buildings.

Context Menu

- New
- Go Into
- Open in New Window
- Open Type Hierarchy F4
- Copy Ctrl+C
- Paste Ctrl+V
- Delete Delete
- Source Alt+Shift+S
- Refactor Alt+Shift+T
- Import...
- Export...
- Refresh F5
- Close Project
- Add UIMA Nature
- Generate PEAR file
- Run
- Debug
- Team
- Compare With
- Restore from Local History...
- Properties Alt+Enter

Aggregate Parameters Parameter Settings Type System Capabilities Indexes Source

Analysis Result Viewer

Results

shows the annotations in a CAS

9:00AM - 5:00PM in GN-K35

in the FROST team will be here to talk to us about FROST.

The JEDII Advanced Topics Tutorial

15: 9:00AM - 5:30PM in Hawthorne 18-F53

will introduce some new UIMA concepts and walk the student through hands-on

Click In Text to See Annotation Detail

- Meeting ("September 15: 9:00AM - 5:30PM in Hawthorne")
 - begin = 2681
 - end = 2732
 - room = RoomNumber ("18-F53")
 - date = DateAnnot ("September 15")
 - startTime = TimeAnnot ("9:00AM")
 - endTime = TimeAnnot ("5:30PM")

WordAnnot Meeting DateAnnot TimeAnnot

Select All Deselect All



Search - Microsoft Internet Explorer


File Edit View Favorites Tools Help

Back Forward Stop Refresh Home Search Favorites

Address <http://sith.watson.ibm.com/newsith/default.htm>

Component Search: [Advanced Search](#)

UIMA Component Library

 The available components include analysis engines, CAS processors (analysis engines, collection readers and CAS initializers) and CAS consumers (indexers, printers, etc.). They can be search by using search field above based on description or can be browsed based on the categories.

CAS Processors

[Entity Detectors](#) [Relation Detectors](#) [Tokenizers, Parsers](#)

Collection Readers

[XML Collection Reader](#) [DB2 Collection Reader](#)

[CAS Consumers](#) [Indexers, Filewriters](#) [Domain Specific A](#) [Life Science, Marke](#) [OpenNLP Analytic](#)

Component Search - Microsoft Internet Explorer













File Edit View Favorites Tools Help

Back Forward Stop Refresh Home Search Favorites Media

Address http://sith.watson.ibm.com/newsith/entity_detectors.htm

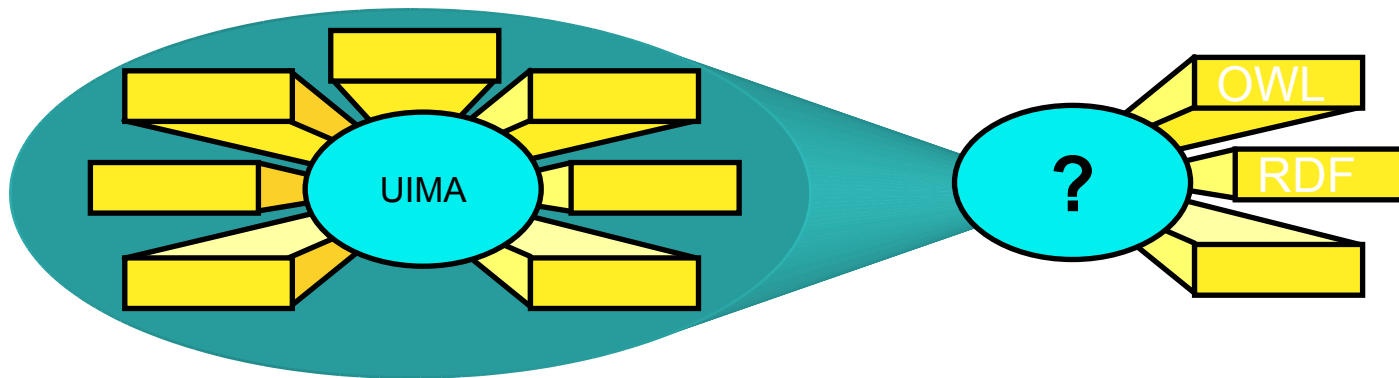
Component Search: [Advanced Search](#)

Entity Detectors

Analytics	Date Added	Rating	Downloads
<p>JTALENT</p> <p>Annotates Document Structure, Syntax, Multi-Word Terms, Named Entities, Cardinals, Money, and Dates</p> <p>OS: Windows (all) File Size: 25.18MB Owner: James Cooper ComponentID: com.ibm.uima.JTalent</p>	02/17/2005	 (23 votes)	 PEAR  XML Descriptor  Type tree 10 downloads
<p>JRESPORATOR</p> <p>A pure Java reimplement of the RESPORATOR (RESPONse GeneRATOR) system originally developed in C++ by John Prager. Builds upon JTalent and detects over 100 semantic classes.</p> <p>OS: Windows (all) File Size: 42.29MB Owner: Adam Lally ComponentID: com.ibm.uima.JResporator</p>	02/17/2005	 (23 votes)	 PEAR  XML Descriptor  Type tree 10 downloads
<p>PERSONTITLE</p> <p>An example annotator that discovers Person Titles in text and classifies them into three categories - Civilian (e.g. Mr., Mrs.), Military (e.g. Lt., Col.), and government (e.g. Gov., Sen.)</p> <p>OS: Windows (all) File Size: 3KB</p>	02/17/2005	 (23 votes)	 PEAR  XML Descriptor  Type tree

More on Relationship to Semantic Web: The Return of the Combination Hypothesis

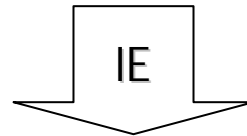
If intimately integrated, various KM technologies will provide higher quality results (accuracy, recall, etc.)



- **Can this be generalized to combination of UIMA & Semantic Web?**
 - Can we combine annotators and formal ontologies and reasoners to accelerate the population of the semantic web?
 - What would “higher quality” mean in this context? How would it be measured?

From UIMA Analytics to the Semantic Web

“13 delegates from Turkey arrived today.”



“13 delegates from <country>Turkey</country> arrived today.”



<country rdf:id=“Turkey” />

Easy!!!

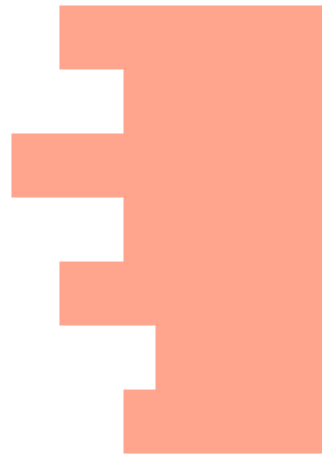


UIMA and Semantic Web Technologies

Text Analytics



Semantic Web

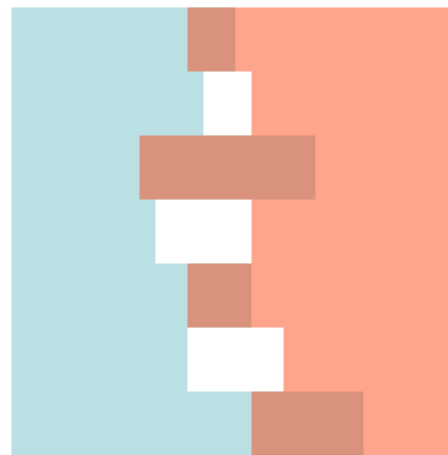


Relations
Entities vs. Mentions
Precision
Recall
Explainability
Brittleness
Scale



UIMA and Semantic Web Technologies

Text Analytics vs. Semantic Web



- Relations
- Entities vs. Mentions
- Precision
- Recall
- Explainability
- Brittleness
- Scale

Very interesting and fruitful work to be done!

Adoption: UIMA within IBM

▪ IBM Research Labs developing UIMA compliant Analysis Engines

- Deep and Shallow Parsing
- Categorization
- Summarization
- Semantic Class Detection
- POS, English/Chinese/Japanese NE
- Classifier Trainers
- Machine Translation
- Video and Speech Analytics
- BioInformatics, etc.

▪ Some products:

- Portal
- Omnifind

▪ [IBM Internal Component Repository](#)

- [80+ Analysis Components and 23+ UIMA-based systems/solutions](#)



Adoption: Outside of IBM

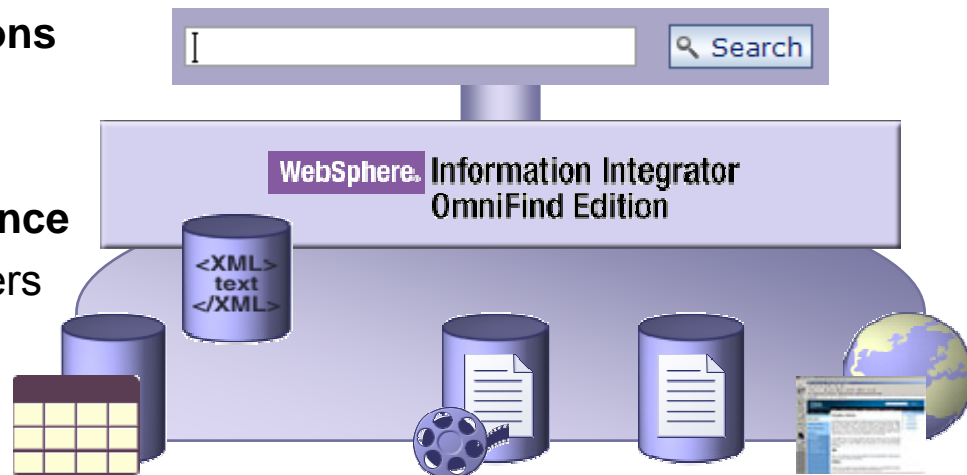
- **First Version of UIMA SDK Released on AlphaWorks Dec 2004**
 - 2,600+ Downloads as of September 2005
 - Open Source Announcement met with broad industry interest
- **Mayo Clinic – an early adopter**
- **UIMA Working Group driven by DARPA and IBM**
 - Small initial group of academics & researchers to evaluate & provide feedback
 - Stanford, Carnegie Mellon, Columbia, UMASS
 - BBN, MITRE, SAIC (Object Sciences)
- **DHS/National Labs - Threat Assessment Project**
- **DARPA/ITPO GALE Project (Speech-to-text, Translation, Distillation)**
- **TC-STAR Speech-to-Speech Project**
- **Third party development of UIMA compliant analytics**
 - GATE Interoperability Layer (University of Sheffield)
 - OpenNLP Components UIMAfied (Tokeniser, Parser, POS, NE, Sent Chunker)
 - Components from UIMA working group members
 - Endorsement by 16+ software companies



Enterprise Search Middleware - Omnifind

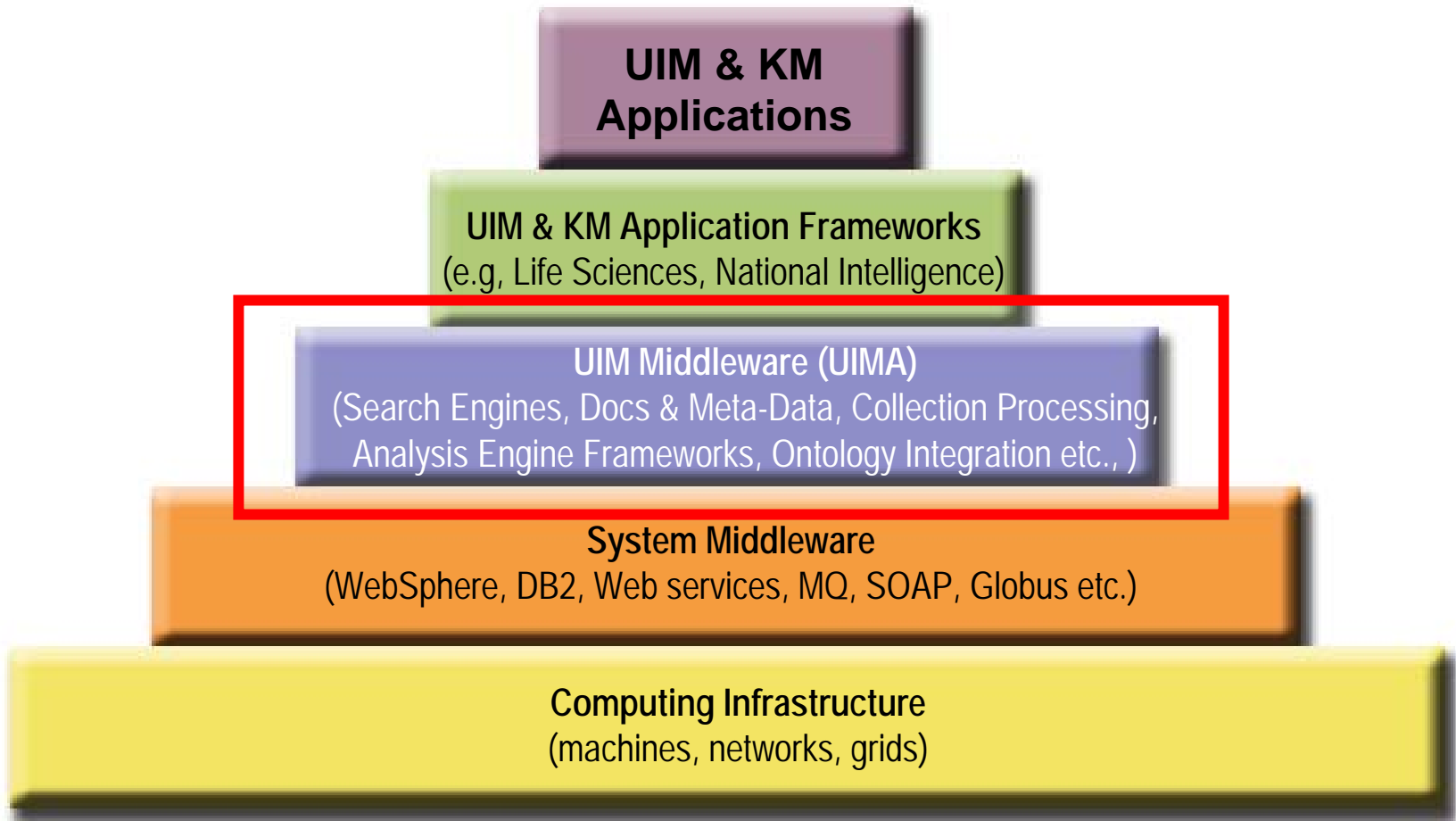
*Unstructured data in the Enterprise forces Innovation in Search Engine
Differentiated value based on Unstructured Information Mgmt. Architecture (UIMA)*

- **Delivers the best results with sub-second response**
 - Sophisticated relevancy algorithms for corporate content
- **Scales for large collections or enterprises**
 - 500K documents and above
 - 1000s of concurrent users
- **Fits easily into enterprise applications**
 - Java APIs
 - Document level security
- **Eases administration and maintenance**
 - Analysis features all under-the-covers



HTTP/HTTPS, News groups (NNTP), File systems, Domino databases, MS Exchange public folders, DB2 Content Mgr, DB UDB, Informix, Oracle Documentum & FileNet via integrated WebSphere II Content Edition

Where does UIMA fit in to the Business World?



ISVs Supporting UIMA and OmniFind

Deliver content to platform for analysis

Provide components text analytics

Provide applications that leverage text analysis and semantic search



Some UIMA Links

- [UIMA Homepage](#) at IBM Research
- Download [UIMA SDK](#) from IBM alphaWorks site
- IBM Systems Journal Article: [Building an example application with the Unstructured Information Management Architecture](#)
- Open Source Press Release: [IBM to Open Source Technology for Analysis of Unstructured Information](#)
 - Related Press
 - [Volume Analytics: IBM's UIMA - and Why You Should Care, DMReview](#)
 - [EE Times](#), [Computerworld](#), [InformationWeek](#), [Computerwire](#), [Database Trends and Applications](#), [SearchCRM.com](#), [BizReports](#), [Information Today](#), [ZDnet](#), [Slashdot.com](#), [ebizQ](#), [CRM Today](#), [CXOtoday](#), [Ovum](#), [WebProNews](#), [Marketing Vox](#)
 - In addition for folks in Ireland, contact [Elaine Stephen@ie.ibm.com](mailto:Elaine.Stephen@ie.ibm.com), Director of IBM Dublin lab to find out about our text analytics in Dublin and for career opportunities there!



IBM Innovation Awards

- Unstructured Information Management Architecture (UIMA) Innovation Award for 2006
- Background: The UIMA framework separates the hard work of advancing the state-of-the-art in Natural Language Processing and more generally algorithms for unstructured information (text, audio, video) analysis... Curriculum and Research.
- Grant size: \$10,000. - \$30,000.
- Objective: Proposals are sought in this area, in the porting of significant analysis algorithms to the UIMA framework, and in the use of UIMA to support knowledge acquisition for the semantic web.
- Online submission will open on January 26, 2006
- Information will be posted at: www.ibm.com/university
- Key dates:
 - **January 26, 2006** **Online submission opens.**
 - **February 17, 2006** **Evaluation begins for proposals rec'd by this date.**
 - **February 28, 2006** **Deadline for submitting a proposal.**
 - **April 28, 2006** **Award winners notified via email & postal mail.**



Conclusions

- **Semantic processing of unstructured information seems exceedingly useful**
- **The semantic processing will be based on many forms of analytics, developed by many – yet operating together.**
- **The combination of these analytics will result in higher accuracy analytics: a.k.a. the *Combination Hypothesis is true.***
- **UIMA provides very valuable engineering support for this**
- **IBM intends to Open Source UIMA shortly to facilitate adoption**
- **We think UIMA will be of value to the semantic web**
- **It seems to us there is valuable research to be done here**
- **IBM will make available Innovation Grants available to catalyze efforts in this important area**



Thank
YOU

