Architecting Knowledge Middleware

WWW 2002, Honolulu, May 9, 2002

Alfred Z. Spector
Vice President, Services and Software
IBM Research Division

aspector@us.ibm.com

*Thomas J. Watson Research Center
PO Box 218
Yorktown Heights, NY 10598*

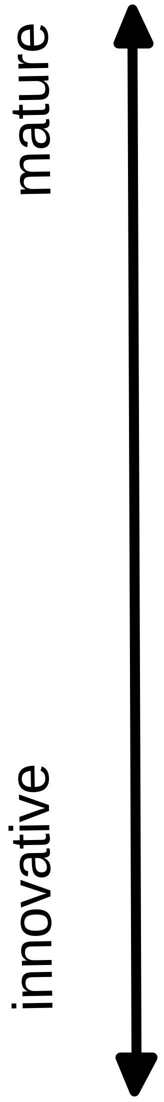
Motivation

- The early 90's Web was elegant & simple
- However, our high aspirations require new technologies, in particular, *for text analysis*

Thesis: (1) Breadth of requirements, (2) technological complexity & (3) sufficient maturity of core component technologies argue for a cohesive, integrated architecture (Knowledge Middleware Architecture)

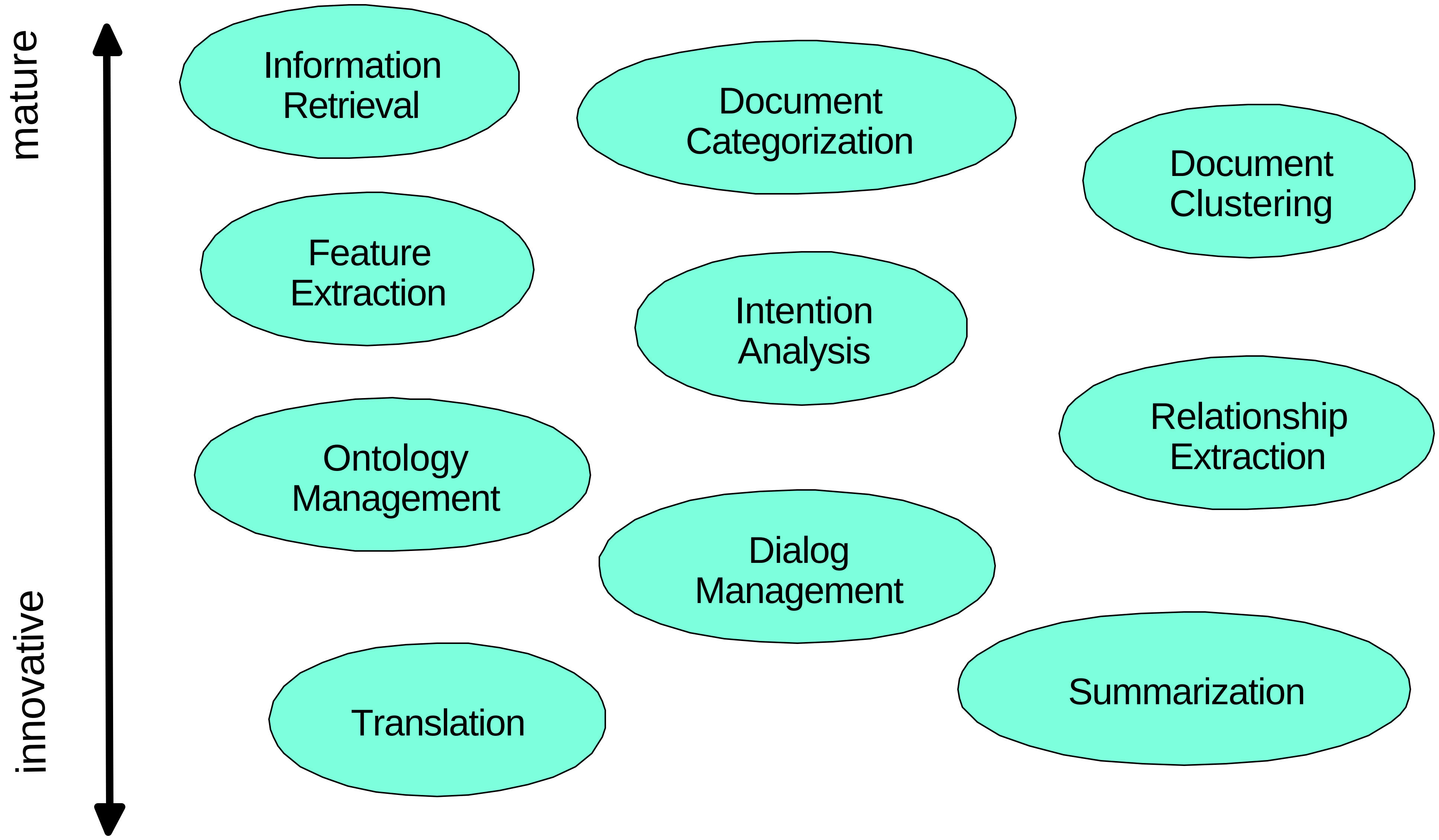
- The Web of the future may not be simple, but it can remain elegant and robust and be highly functional

Many KM/Text Analysis Technologies & Applications



Task	Technology	Capability	Applications
Find	Information Retrieval	Match query to documents	KM and Data Management in portals, mail applications, help systems.
Organize	Document categorization	Assign documents to predefined classes	CRM E-mail routing; portals; mail filing
Organize	Document clustering	Discover groups of similar documents	Text mining
Discover	Feature extraction	Discover names and terms, e.g. company names, dates, custom features	Base of many text mining approaches
Discover	Intention analysis	Discovers value statements in text	CRM - mine help desk reports; BI discovery of negative product evaluations
Discover	Relationship extraction	Measures or discovers how terms are related in text	KM Fuzzy hyperlinking; DeB ontology building
Organize	Ontology management	Encodes semantic knowledge and provides inferencing capability.	DeB, AIM, Data Mgt. Resolve semantics of XML data
Find	Dialog management	Manage sustained interaction with a user, using natural language & a domain model	e-commerce. Natural language assistants to find information and carry out tasks using written or spoken instructions
Find	Summarization	Generates a compact readable representation of a document	Portals, search engines, pervasive-device access to documents
Discover	Translation	Statistical, syntactic, and semantically augmented conversion	Multi-lingual product literature, Web page conversion, Enterprise Globalization

KM/Text Analysis Technologies: Another View



Note:

Text Analysis *technologies & techniques* seldom work well together

Argument Outline

Toward... Architected Knowledge Middleware

1. The technical and economic imperative
2. The challenges
3. The practicality
4. The benefits

1a. The Technical Imperative

- We have a complex problem
- We have core technologies
- Historic trends are toward integrating technologies
- No single approach can succeed well enough: Combination analysis necessary
- And together, we need to mitigate complexity

Explosive Growth in Size & Heterogeneity

- The ***amount*** of accessible data
  - growing to petabytes online
- The ***sources*** of data
  - Web, intranets, extranets, subscription services
- The ***types*** of data
  - structured, semi-structured, and unstructured
- The **formats** of data
  - text, html, pdf, gif, jpeg, etc.

Trends in Heterogeneity of Data

- Despite heterogeneity, users would like seamless use of all kinds of information
  - Parametric & Text
  - Multilingual
  - without syntax/protocol differences
- And they want good results!
- XML will play a very large role
  - Structured data, when annotated with semantics, context and explanation -> XML
  - Textual data, when tagged with semantics and/or syntax, and associated with numerical data and metadata -> XML
  - One XML document might not be quite enough when multiple annotators are involved
    - might need different views of same document, different tokenizations, etc.
- However, XML by itself doesn't make life easy

Example:

- Problem: misspelled product names (e.g., "thinpad") result in lost sales
- Solution:
  - Collect *Failing* e-store queries
  - Locate words that sound like product names (e.g., “ThinkPad”) in a product name context
- Result: Index “thinpad” under all locations for “ThinkPad” or add it to the spelling dictionary
- To make this work: need to combine:
  - two annotators: sounds_like and named_entity
  - search engine

Core Technology is Available

There is an enormous amount of technology in the fields of Information Retrieval, Text Analysis, and NLP.

They will of necessity become key means of going beyond semi-structured information management as enabled by manual XML markup, to the management of unstructured information such as free text.

Example: Search evolving from its I/R roots

- Web Search adds metadata search, some statistical processing, and massive scalability to basic search
- Discovery adds text analysis, to handle semantics and context, as well as structured, unstructured and semi-structured (XML) data

	1st Generation: **Information Retrieval**	**2nd Generation:** **Web-based Search**	**3rd Generation:** **Discovery (Text Mining)**
User:	Trained specialist	Everyone	Everyone and software agents
Scope:	Small, closed collections	WWW	Structured, semi-structured and unstructured information
Technology:	Pattern/string matching with weights on importance	Pattern/string matching and hyperlink analysis for relevance ranking + categorization	Linguistic, advanced statistical, & semantic processing

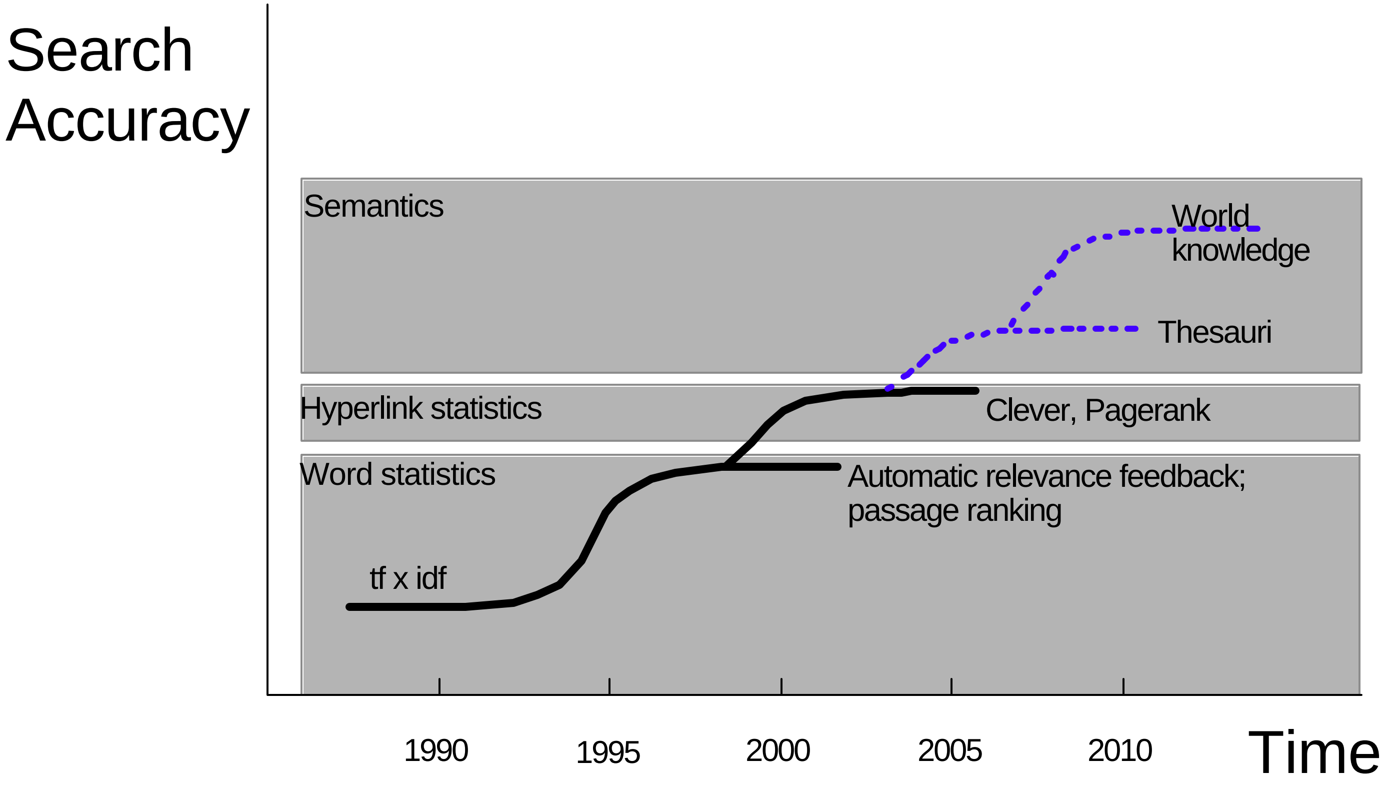
1960 - 1993

1994 - 1999

2000+

Evolution of Text Search

Bibliographic text search, using word statistics alone, is approaching a limit of accuracy. Further improvements in text search will be enabled by the use of semantic resources, used to disambiguate query terms, and to support limited inferencing over the domain of the search.



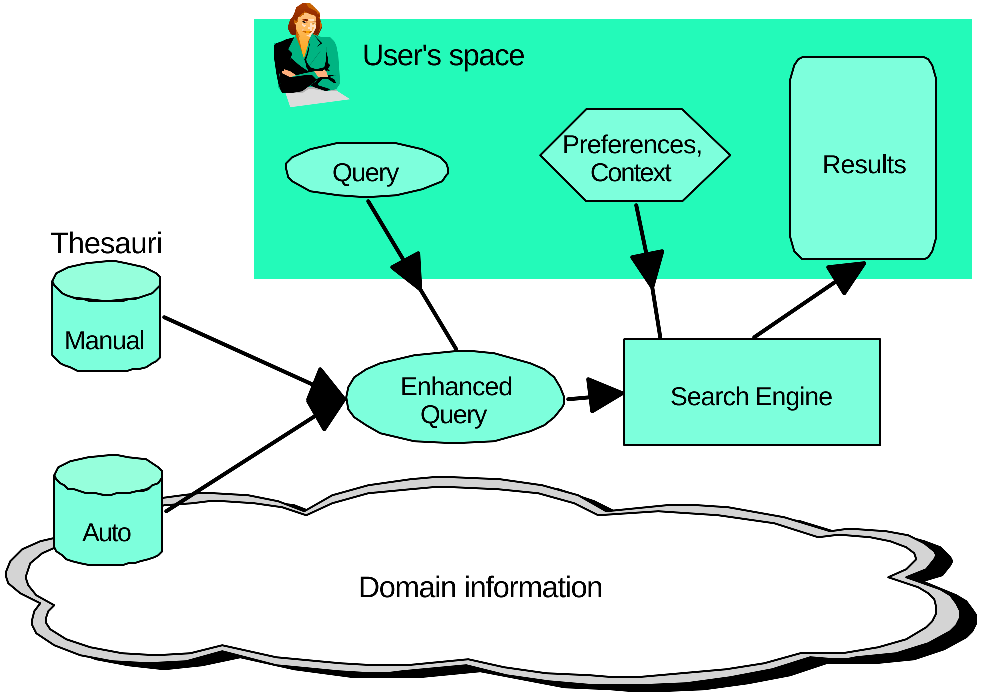
Adding Knowledge to Search

Knowledge about the user

- Personalized user information
  - Geographic information
  - Bandwidth
  - Language, ...
- The context of the search
  - task
  - business process
  - previous queries, ...

Knowledge about the domain

- Knowledge bases
- Information about the meaning of words, derived from Thesauri which may be hand-built or -- increasingly -- constructed automatically
- Geographic information, ...



Adding Ontologies in Search

- First stage: use ontology to further disambiguate query terms, or to enhance the query
- Second stage: Perhaps, change search method to use conceptual graph structures
  - Initial applications begin in narrow domains
  - and broaden over time

Greater Accuracy & Experience Via Combination

- If *combined,* various technologies will provide higher quality results (accuracy, recall, etc.) and will prove necessary
- They will also provide more *modes of interaction*
- Analogy is drug combination therapy; e.g., in Tuberculosis and AIDS triple drug therapy

  (See, "The Forgotten Plague", Frank Ryan, Little Brown, 1993)

  I argue that a combination of Information Retrieval, Grammatical, Statistical, Advanced Statistical, and Semantic technologies will prove needed to achieve quality (e.g., accuracy, recall) requirements

- And the technology is generalizable to many problem areas

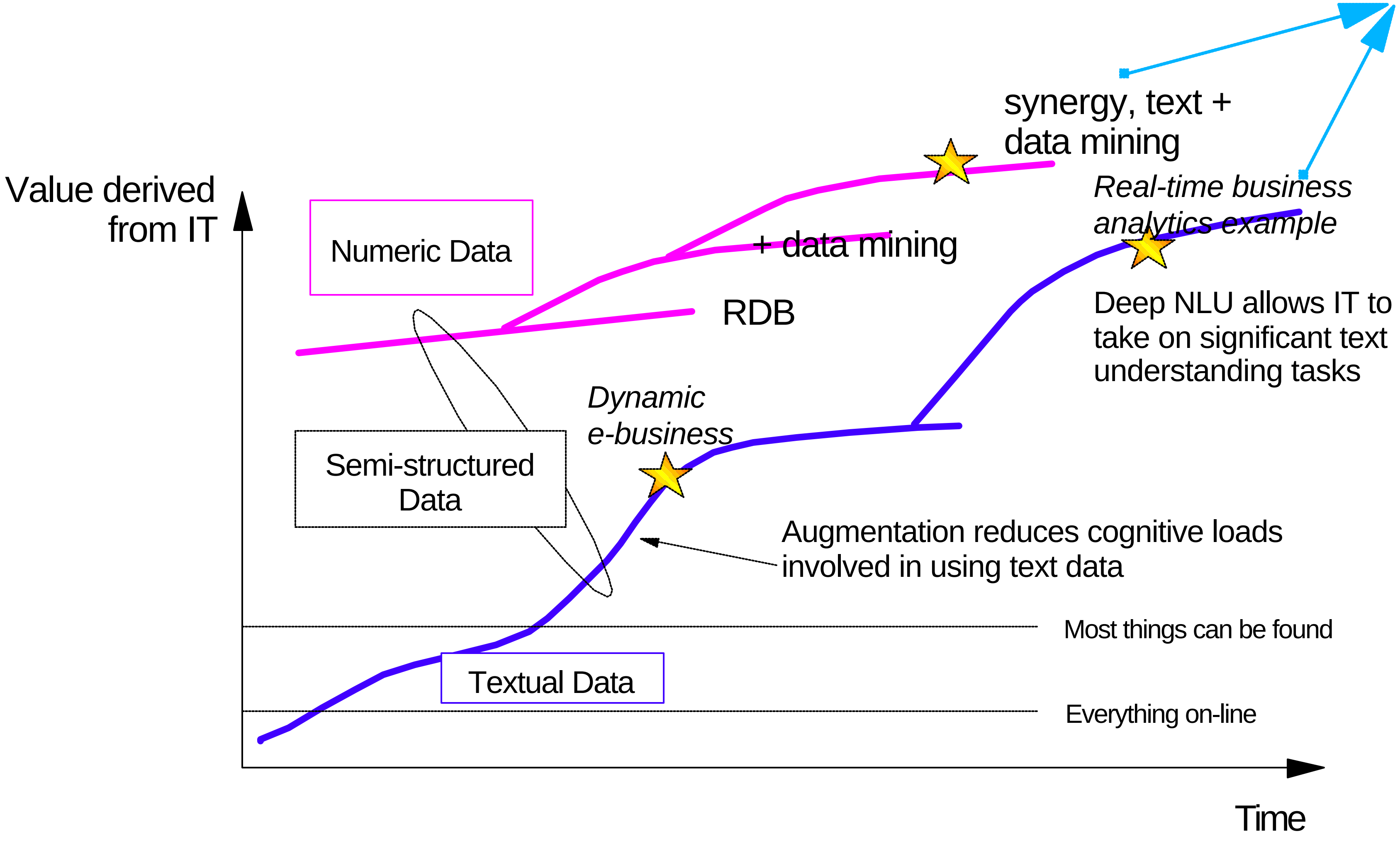
Reduced Complexity and Work

- The KM and Text Analysis technologies are getting complex
- They need to share common structures and processing algorithms
- Without sharing, the cost of developing systems will grow too high and systems will be unwieldy

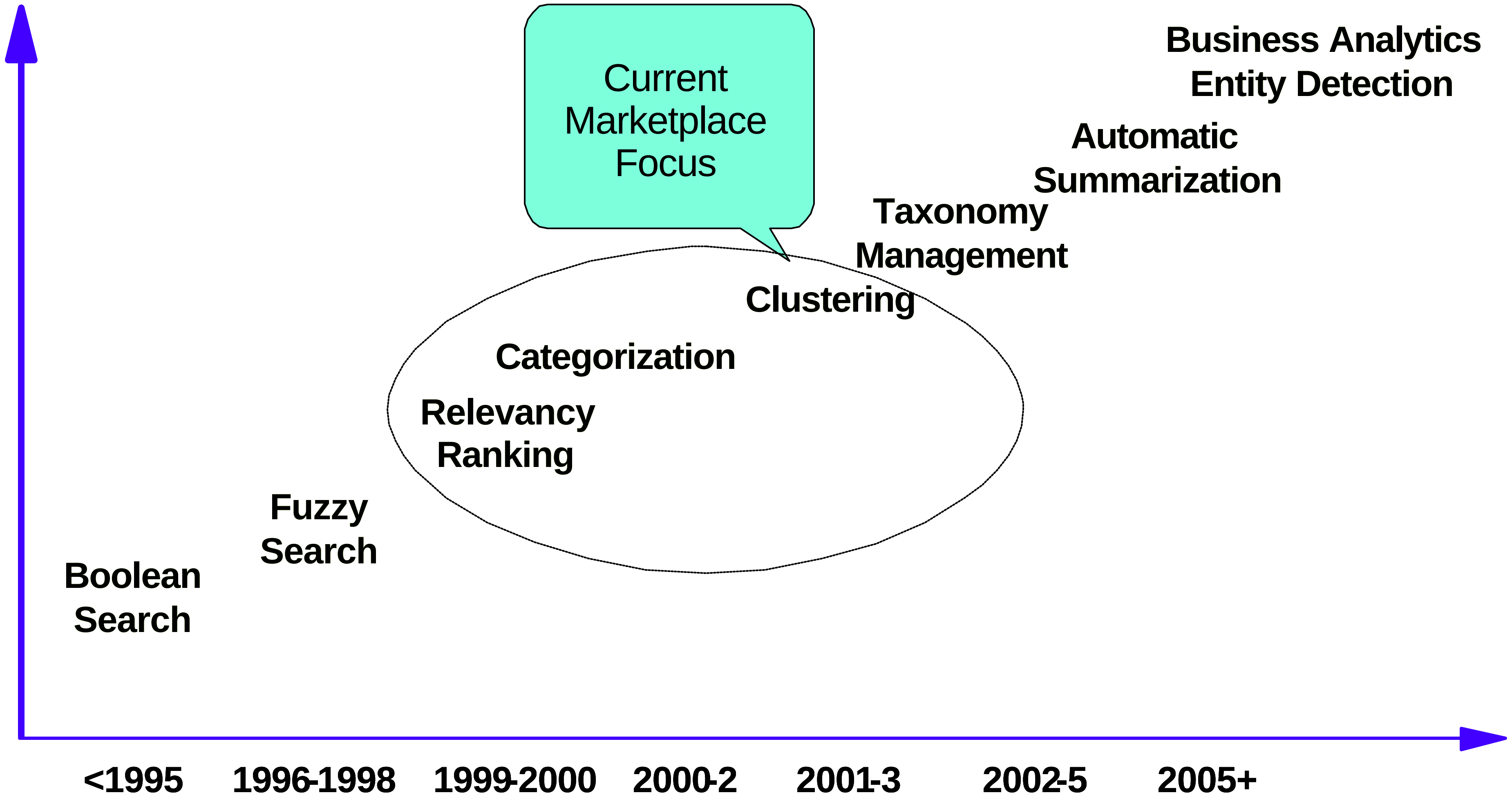
The Database Analogy

- The relational model was/is elegant
- Internally, the RDBMS has become a wonderous merge of many technologies:
  - Low-level search
  - Storage
  - Synchronization
  - Recovery
  - Security
  - Optimization
  - and more
- They work nearly seamlessly to implement the relational concept...

1b. Economic Imperative



Escalating Demands for Search & Text Analysis



Current Marketplace Focus

Boolean Search

Fuzzy Search

Relevancy Ranking

Categorization

Clustering

Taxonomy Management

Automatic Summarization

Business Analytics
Entity Detection

<1995 1996-1998 1999-2000 2000-2 2001-3 2002-5 2005+

Knowledge Middleware in IBM Products

- IBM EIP/II
  - v7: Search, categorization, summarization
  - v8: clustering, extraction
- Lotus Knowledge Discovery System
  - Knowledge map building
- IBM WebSphere Business Components
  - Text Analyzer (text classification)
- IBM Global Services Offerings
  - Business Intelligence, Life Sciences, Knowledge Management ...

Outline

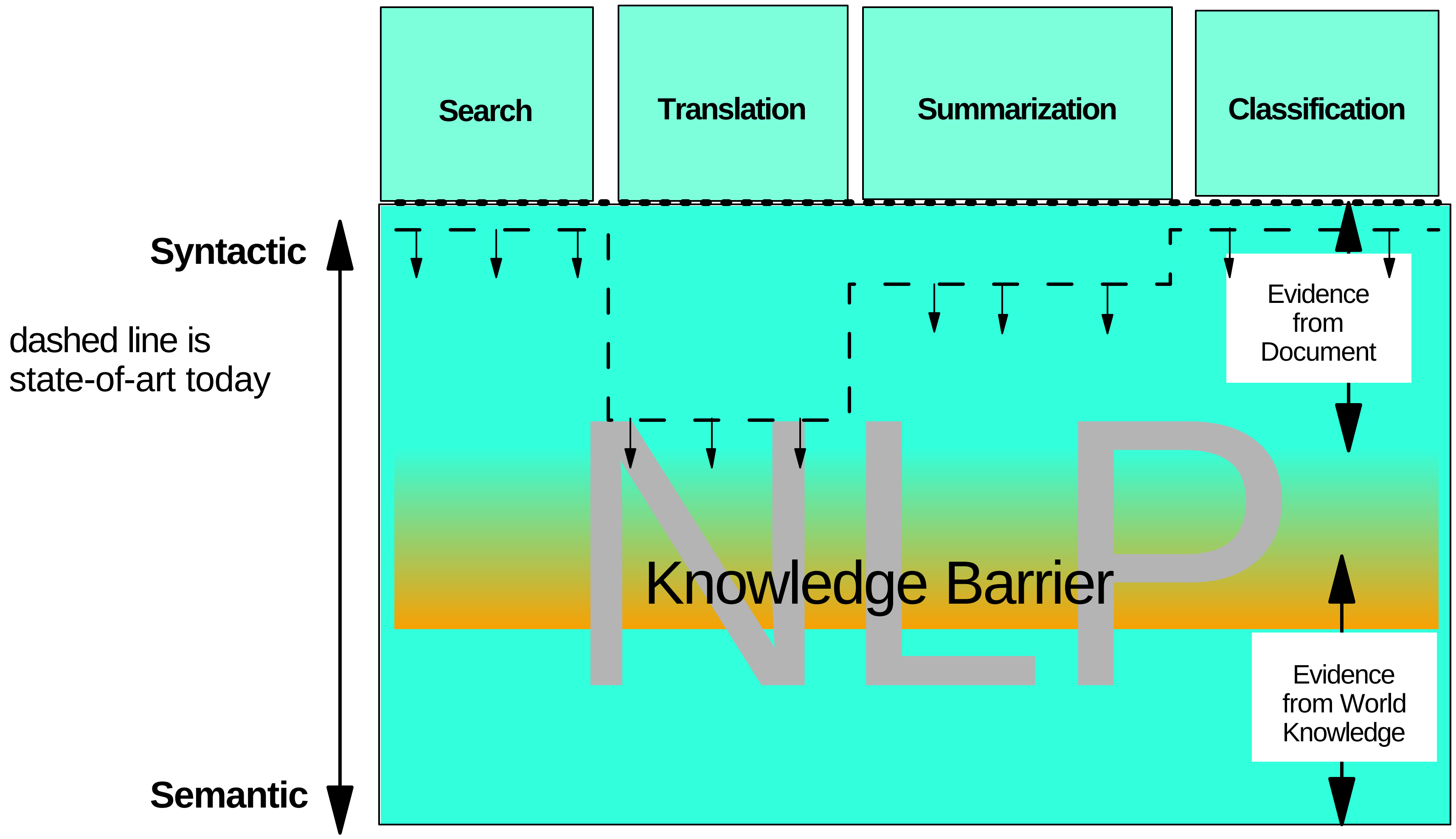
Argument:

1. The economic and technical imperative
2. The challenges
3. The practicality
4. The benefits

2. The Challenges

- Crossing the semantic barrier
- Integration
- Scientific domain division vs. technological integration

A Central Challenge:
Crossing The Semantic Barrier



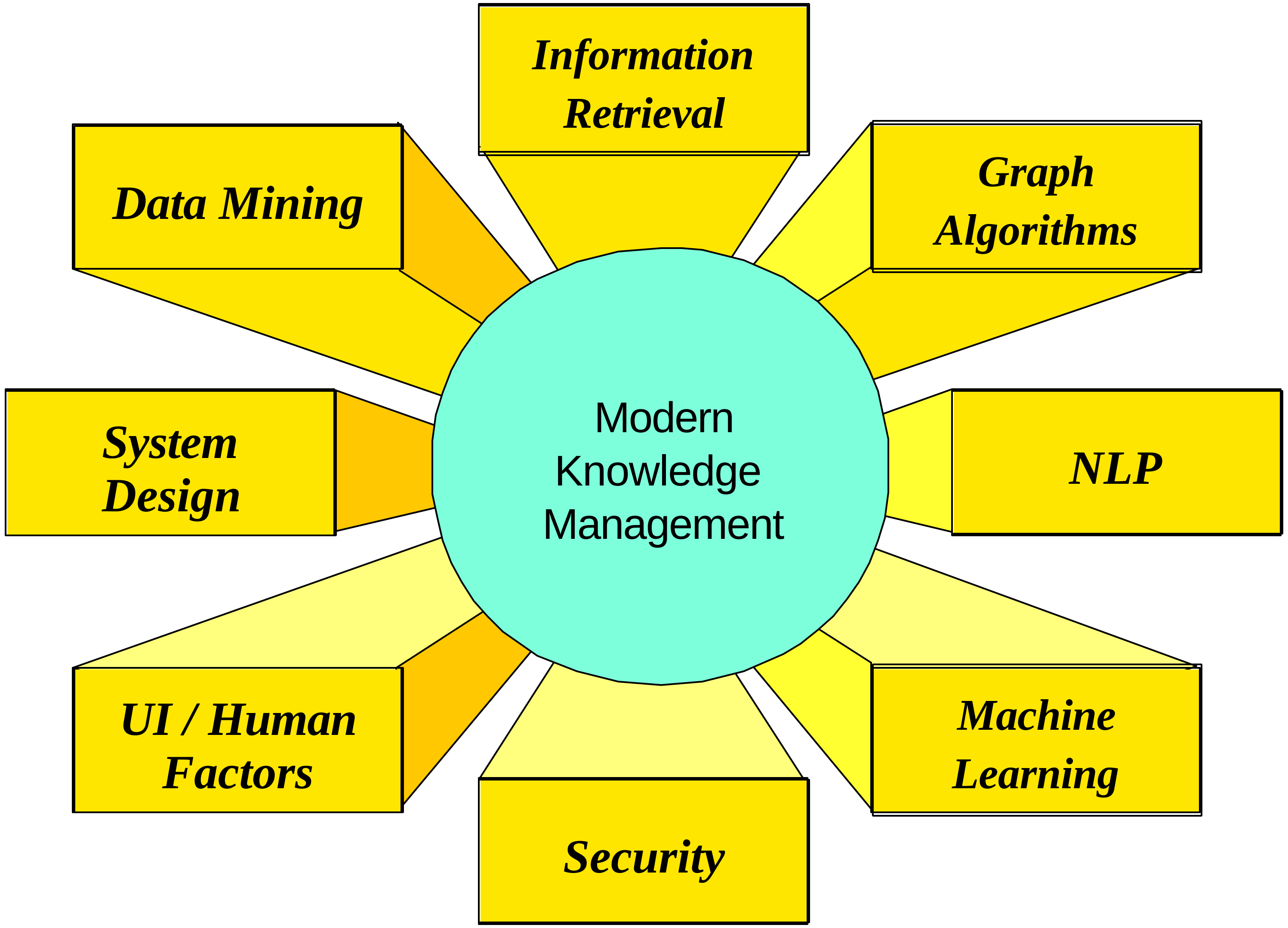
Integrational

- Do we know how to integrate so many technologies?
  - Many are so different
  - They've been implemented so differently in the past
- Do we even know the right storage structure for information?

Organizational

- Text analysis technical/NLP communities fragment by:
  - Intended Application
  - Approach (e.g., grammatical, I/R, statistical, ...)
- Advanced motivational techniques required to induce technical teams to work together on a Knowledge Middleware Architecture

Challenge: Scientific Domain Division vs. Application Integration



Outline

Argument:

1. The economic and technical imperative
2. The challenges
3. The practicality
4. The benefits

3. Practicality

- A realistic integration example
- Architectural progress at IBM

Customer Claim Mining (IBM Tokyo Res. Lab)

Analysis of inquiry records from PC Help Center

- ❑ discover and analyze trends and patterns
  - discover product failures in their early stages
  - discover customer behavior patterns

Trend Discovery and Analysis

- ● Find query with largest increase

  From mid June, queries relating to Windows 98 increase sharply.

- ● Analyze cause of increase

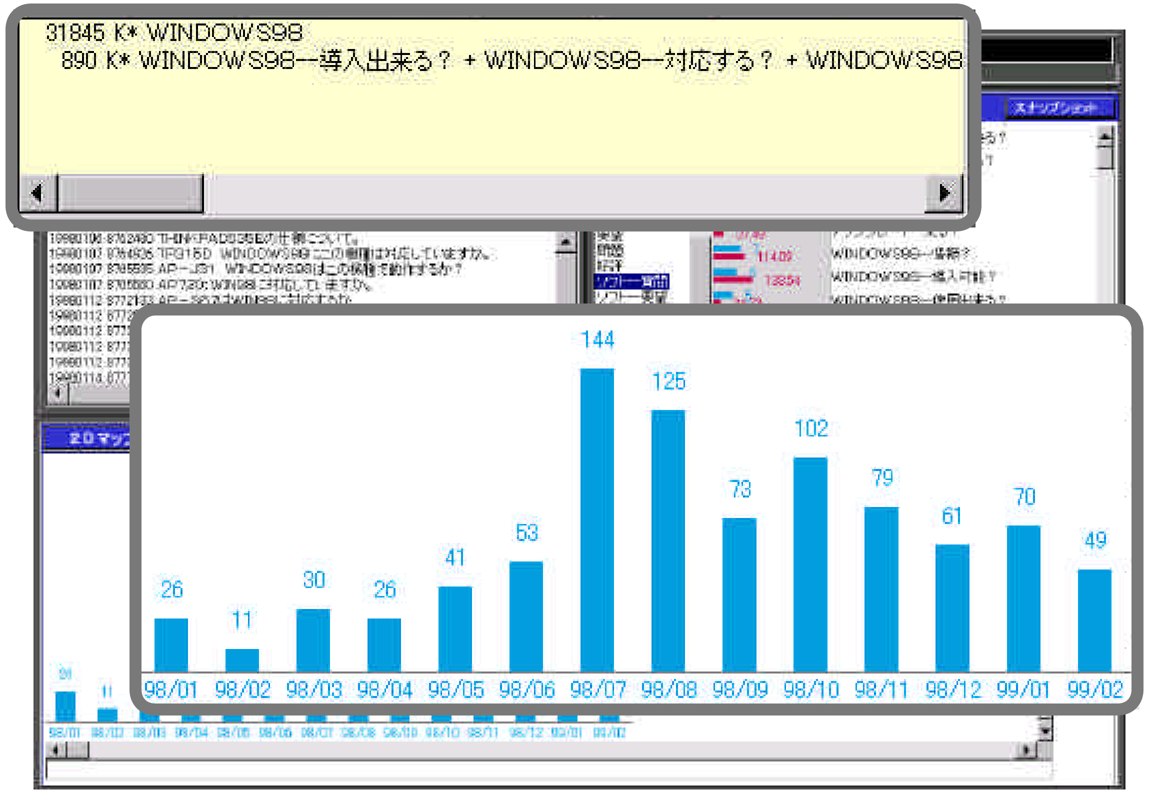
  Sudden increase in July attributed to queries relating to Windows 98 installation.

- ➙ Action

  Place a list of Windows 98 compatible machines on web page

- ● Result

  Windows 98 installation queries declined after August.



31845 K* WINDOWS98
890 K* WINDOWS98—導入出来る？ + WINDOWS98—対応する？ + WINDOWS98

TAKMI: Text Analysis and Knowledge MIning

Technology for Mining in TAKMI

- Trend Analysis --- *Analysis of changes in time sequences*
  - Topic Extraction
    - analyzes changes of topics and extracts their patterns
  - Trend Analysis
    - analyzes patterns of increase/decrease of concepts
- Feature Analysis --- *Analysis of remarkable features of concepts/facts*
  - Singularity Analysis
    - extracts concepts strongly associated to a set of data
  - 2D Association Analysis
    - detects remarkable features of a concept in comparison with other concepts in the same category
- Relationship Analysis --- *Analysis of relations among concepts/facts*
  - Analysis of numerical ranges with concepts
    - associates concepts with numerical ranges such as problem/call taker with call duration
  - FAQ generation
    - associates facts (predicate-argument pairs) with other facts
- etc.

IBM Research's Knowledge Middleware Architecture (UIMA)

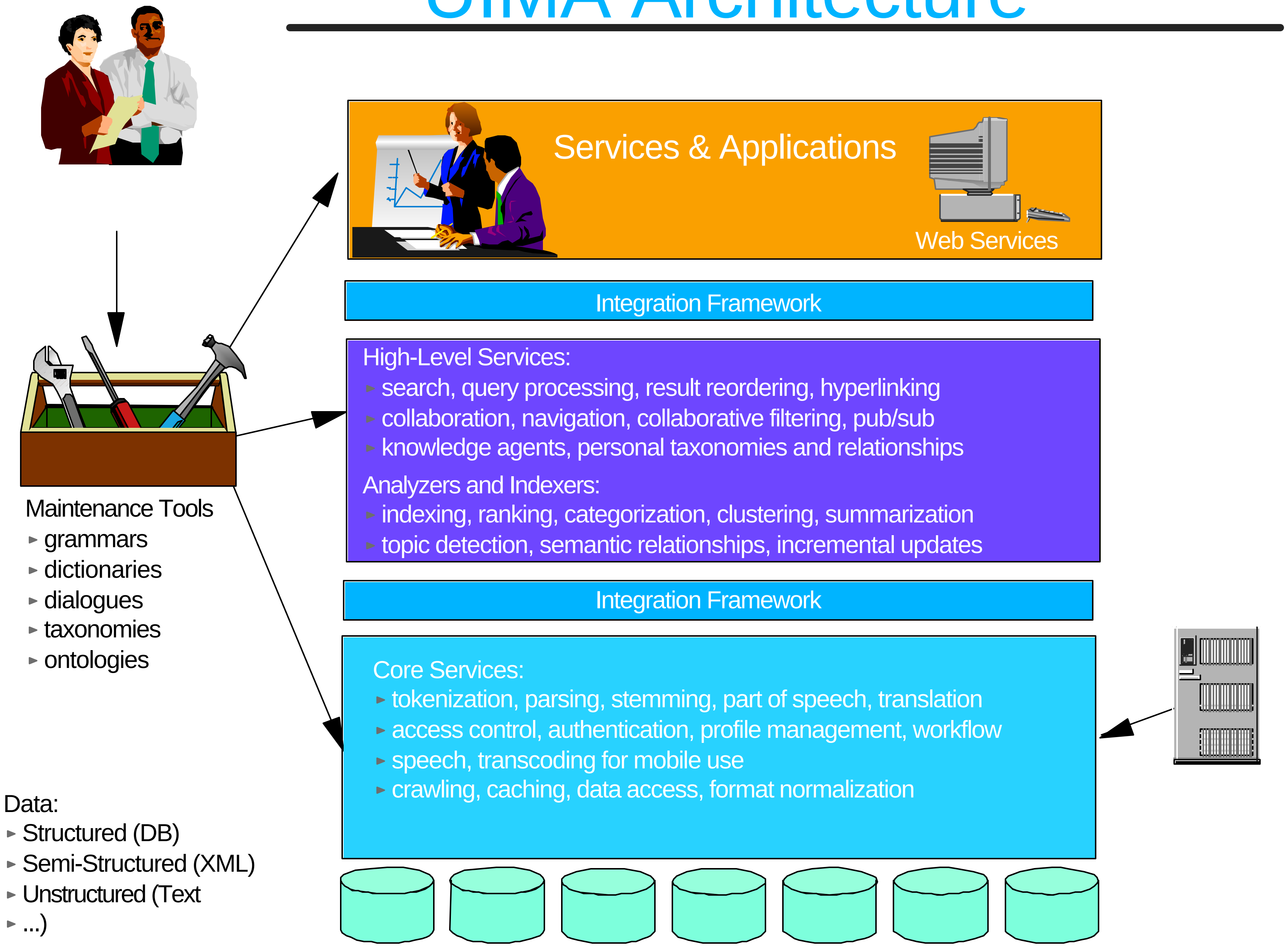
Traditional approaches to building UIM applications are "algorithmic centric", resulting in tightly integrated vertical applications, whose design is dominated by concerns of computational load.

A new approach for providing NLP functionality is evolving which recognizes the inherent need for flexibility and exploits todays extrodinary MIPS, storage, and networking capacity.

IBM KM Architecture: UIMA Project

- Provide a common framework for the integration of UIM technologies
  - Common Annotation System (Abstract Data Structure)
- Flexible & Adaptable (Service Oriented Architecture):
  - uses XML standards to support dynamic binding of services and distributed (multiagent) implementations (RDF, WSDL, WSFL...)
  - supports "persistent binding" to avoid dynamic binding overhead for batch, single agent processes
  - both tightly- and loosely-coupled variants
  - toolkit / library, not monolithic system
    - accommodates variety of applications and separates programming tasks that require distinct skills
- Seamless integration of:
  - structured, semi-structured, and unstructured data
  - human agents and computer agents

UIMA Architecture



Services & Applications

Web Services

Integration Framework

High-Level Services:
- search, query processing, result reordering, hyperlinking
- collaboration, navigation, collaborative filtering, pub/sub
- knowledge agents, personal taxonomies and relationships

Analyzers and Indexers:
- indexing, ranking, categorization, clustering, summarization
- topic detection, semantic relationships, incremental updates

Integration Framework

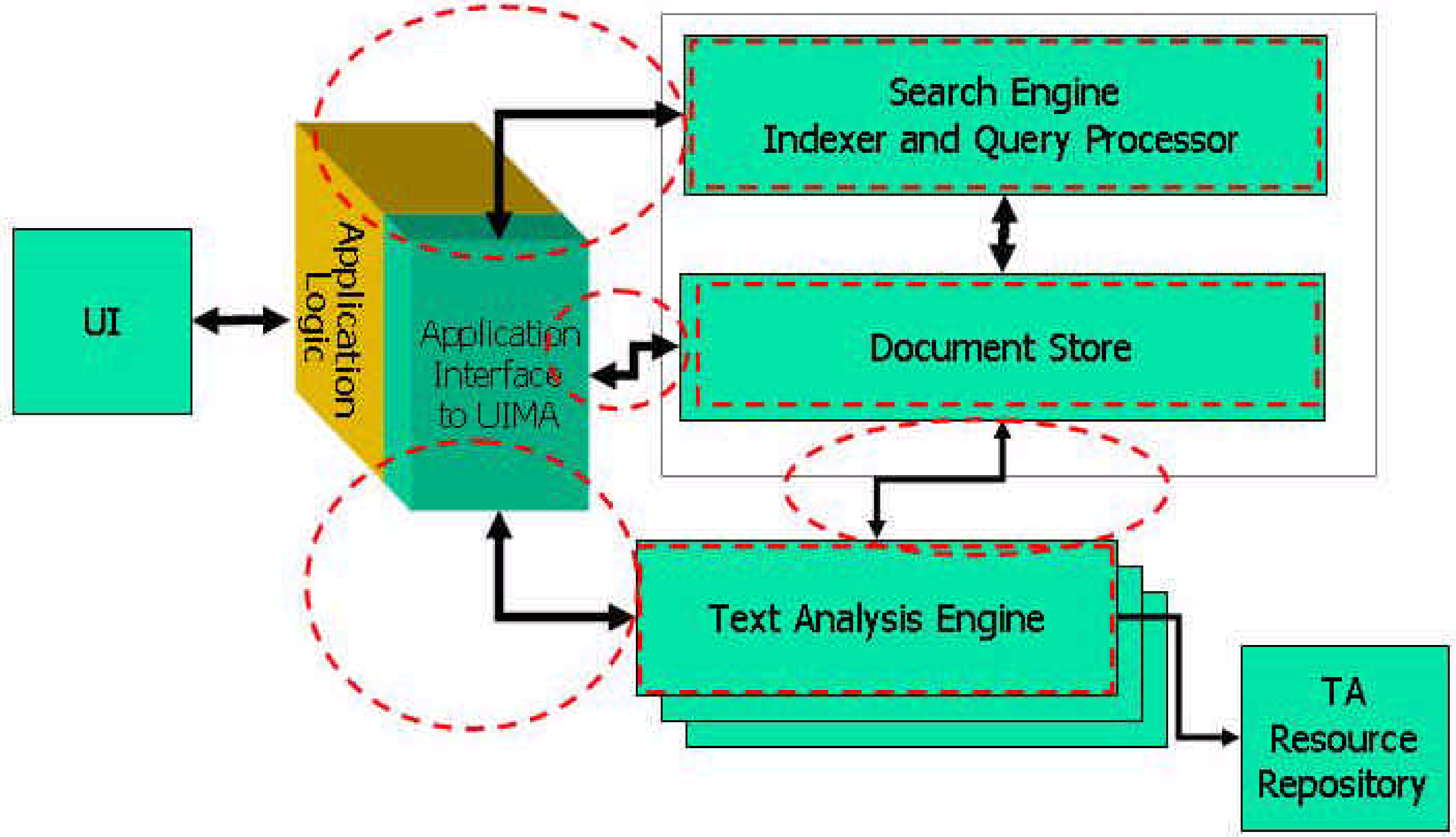
Core Services:
- tokenization, parsing, stemming, part of speech, translation
- access control, authentication, profile management, workflow
- speech, transcoding for mobile use
- crawling, caching, data access, format normalization

Maintenance Tools
- grammars
- dictionaries
- dialogues
- taxonomies
- ontologies

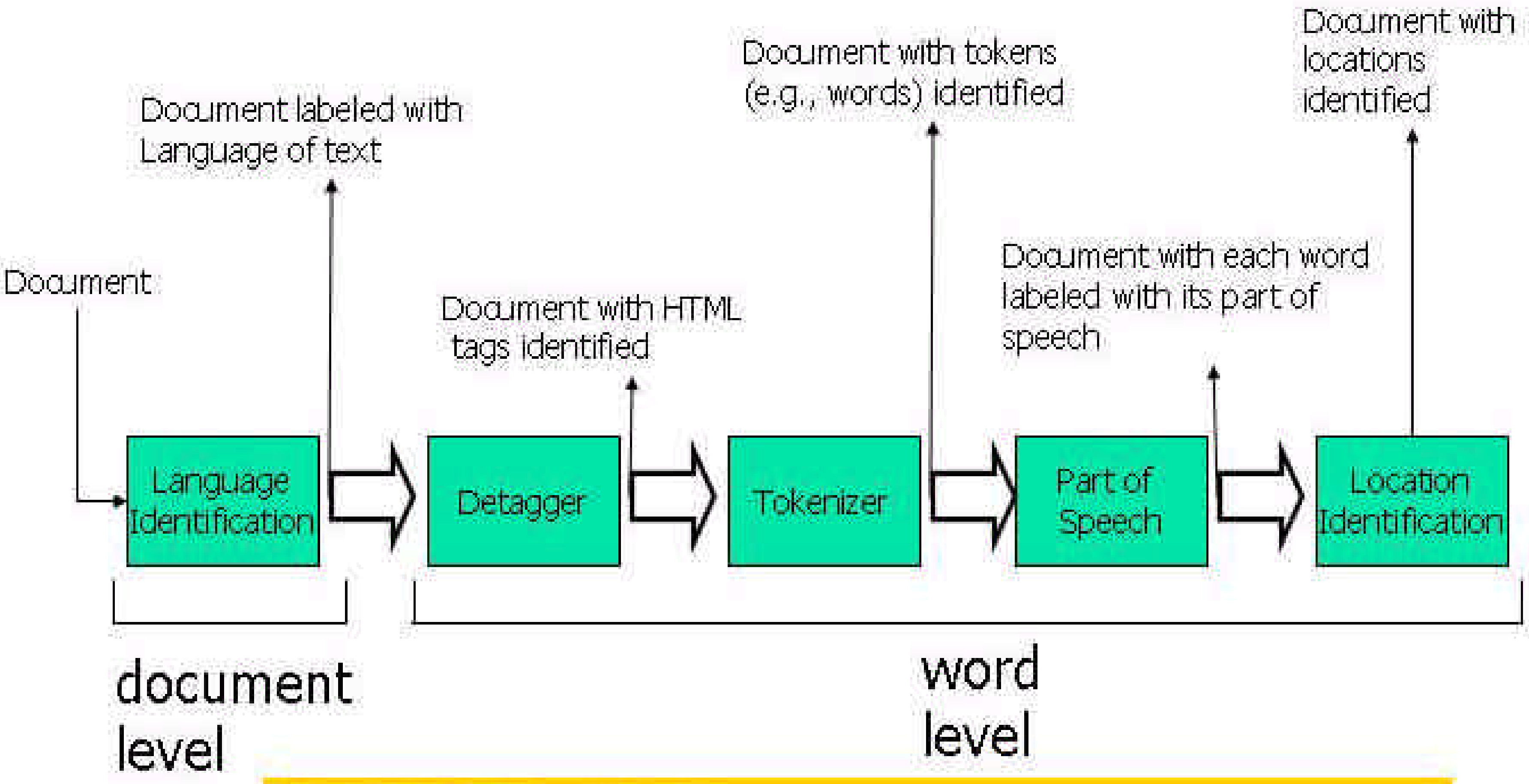
Data:
- Structured (DB)
- Semi-Structured (XML)
- Unstructured (Text
- ...)

Structured - Semi-structured - Unstructured

Top-Level Architecture



Simple Text Analysis Application: Location Identification



Document

Document labeled with Language of text

Document with HTML tags identified

Document with tokens (e.g., words) identified

Document with each word labeled with its part of speech

Document with locations identified

Language Identification

Detagger

Tokenizer

Part of Speech

Location Identification

document level

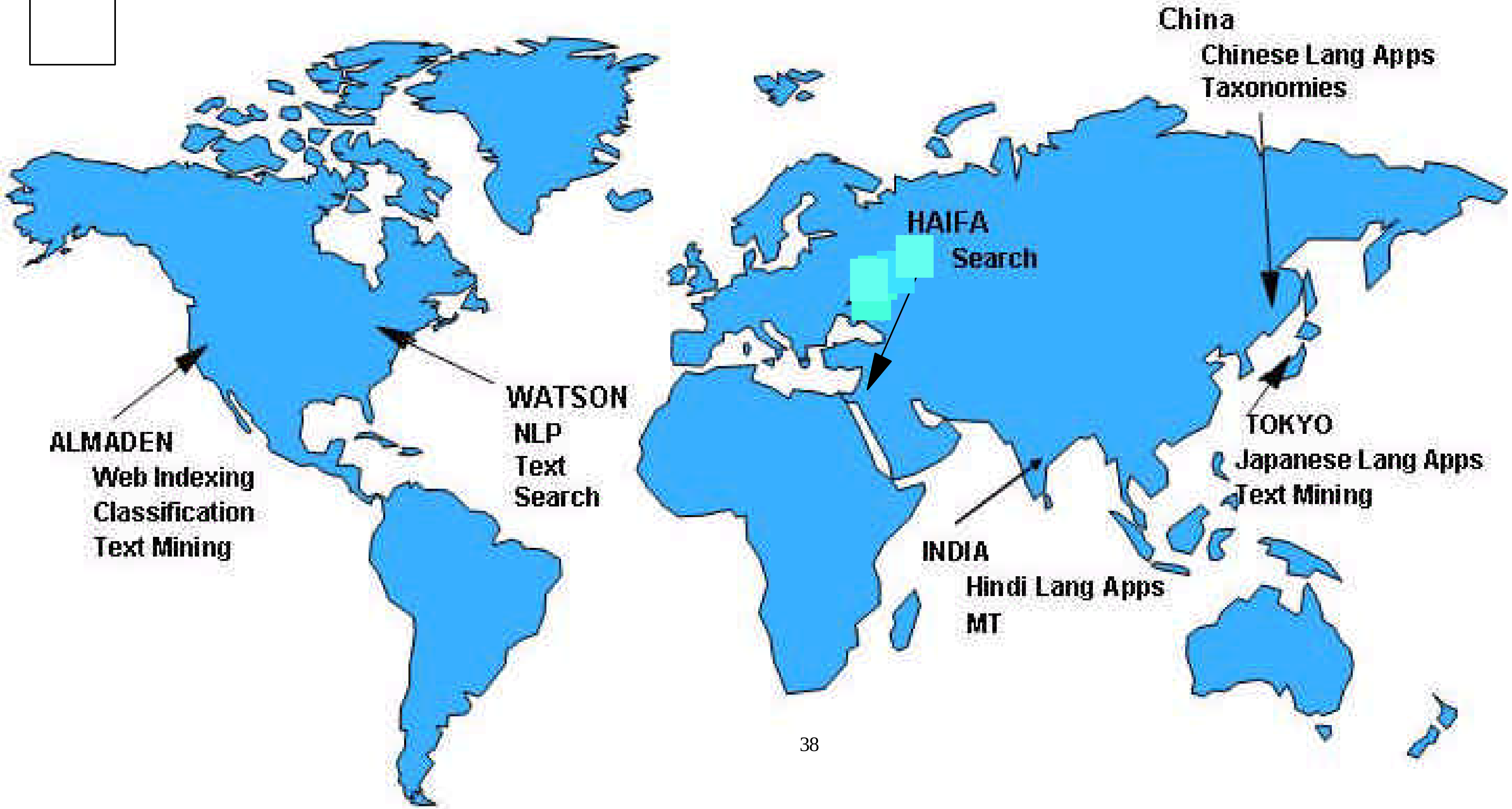
word level

Annotators: Analyze,
Recognize &
Label specific semantic content for next consumer

Status

- Design being completed
- Growing list of components
- Distribution ?

IBM Research Worldwide UIMA Investment



Summary & Benefits

- We argue for the benefits of a common KM Architecture
  - Provides a common facility for accessing/creating & importing/ exporting annotations of documents using multiple views
  - Enables coordination of a set of annotators based on common syntax & semantics
  - Has reduced duplication of common functions
  - Enables same annotators to be used in different architectural variants
- Upside:
  - Combination Analysis improve standard KM functions
  - Pooling of approaches & talent will yield GREAT results
  - Supports creation of Semantic Webs
  - Permits great progress on Computer Science NLP Grand Challenge!
- Downsides:
  - Any standardization imposes some constraints

Thank you for listening.