IBM Software Group

IBM

e business software

Unstructured Information Management

Why information is so hard to find these days and what's being done about it.

D.J McCloskey
IBM LanguageWare

Introduction

- Ability to mangage information is a crucial enabler of progress
  - ancestors
    - invented language and writing systems
    - enable paradigmatic shifts in civilization
  - We are at the next juncture
    - Almost victims of our success
    - Ability to capture and share information

The ultimate resource

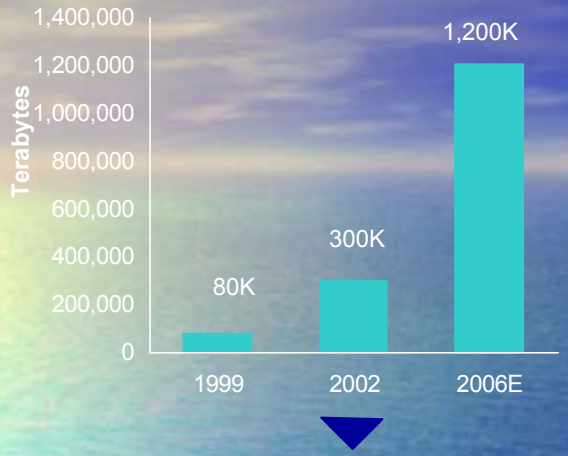
- The internet
  - Petabytes of info
  - Every topic/Every level of detail
- The intranet
  - Petabytes
  - rich indepth
  - current info
  - domain focused

How Much is a Lot…

- GenBank
  - 22 billion sequences (as of 2002)
  - grown exponentially since 1982
  - **220 GB** (each sequence 10 bytes)
- 138 Terabytes Newspapers
- 39 Tb Books
- 52 Tb Mass Market periodicals
- 14 Petabytes of TV Series
- 20 Petabytes of X-rays
- Almost 2 exabytes on hard disk… per year… original
- Flows…. Telephone – 17.3 exabytes!!!! In one year 2002…

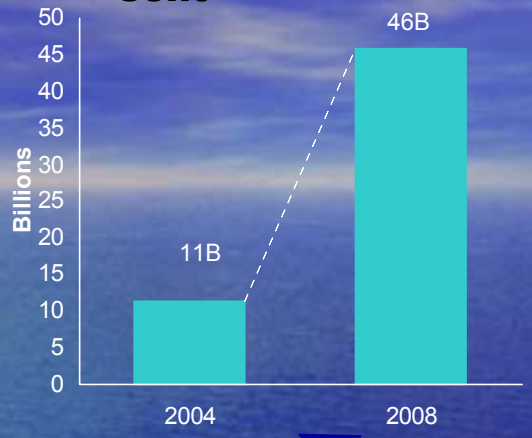
Source: How Much Information Oct 27, 2003, Berkeley Report

Worldwide annual amount of email messages sent[1]



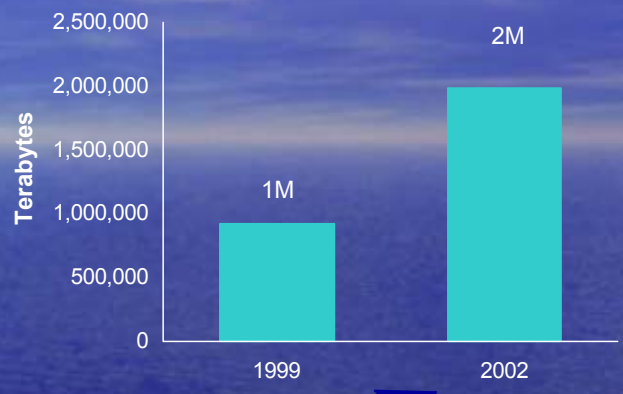
- The current average size of an email box is estimated between 26MB and 50MB and expected to increase at a 14% CAGR
- The average corporate end-user sends 5MB of content per day
- Unsolicited email (spam), commercial notifications and news alerts account for one-third of today's email load and are forecast to comprise nearly half of the traffic four years from now

Worldwide number of daily Instant Messages sent[2]



- IM is used in 85% of all enterprises in North America (includes either personal or business sanctioned)
- There are approximately 831 million active IM accounts:
  - 768 million (92%) are public IM network accounts
  - 63 million (8%) are enterprise-only

Worldwide annual production of original content on Hard Drives[3]



- Production on Hard Disks had a CAGR of 29% from 1999 to 2002 driven by compliance needs and a general increase in the volume of documents stored
- The enterprise videoconferencing industry is growing at 25 percent

Office workers spend an average of 9.5 hours each week searching for information

Source: How Much Information 2003, Berkeley Report, Radicati Group, IDC Research, Merrill Lynch, Wainhouse research, Mercer Analysis

1. Assumptions: Average email size: 59KB
2. # of IM's are user-to-user and defined as each time something is sent
3. Includes information stored on both hard drives and servers

It gets worse…

- It is a multilinugal world!.
- More content in non-english now (>75%)

- It gets worse still…
- Not just text
  - Images
  - Sound
  - Video

- Worse again..
- Unstructured info tends to store most uptodate info too, newspapers meeting minutes

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

Structured Information:

Semantics of information captured in DB schema

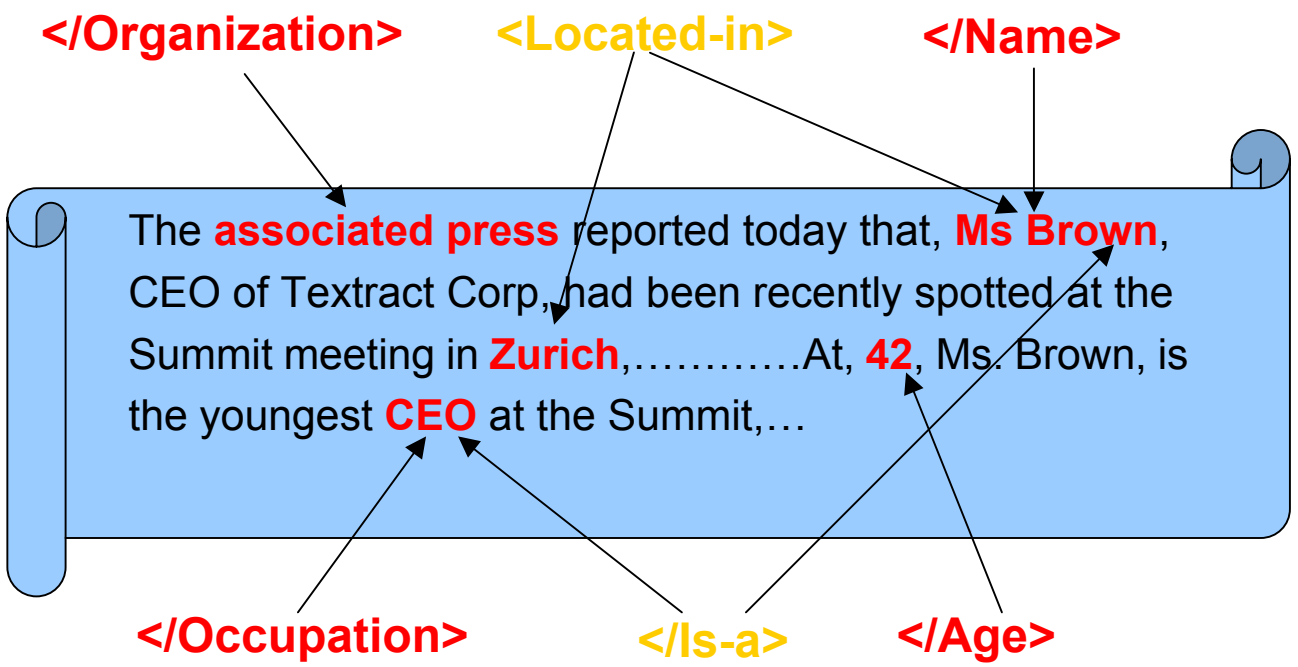
Name	Occupation	Organization	Age	Office Location
Jones	Engineer	IBM	29	San Francisco
Smith	Journalist	AP	32	Boston
Brown	CEO	Textract	42	New York

Unstructured Information:

Semantics inherent in usage and context

The associated press reported today that, Ms Brown, CEO of Textract Corp, had been recently spotted at the Summit meeting in Zurich,……At, 42, Ms. Brown, Is the youngest CEO at the Summit,…

UIM: *Adding structure to unstructured information through metadata created by automated text analysis*



Analysis Engines (pattern recognition techniques) discover Semantics and do Automatic Tagging, allowing us to process text as we would structured data.

IBM

What's being done about it?..

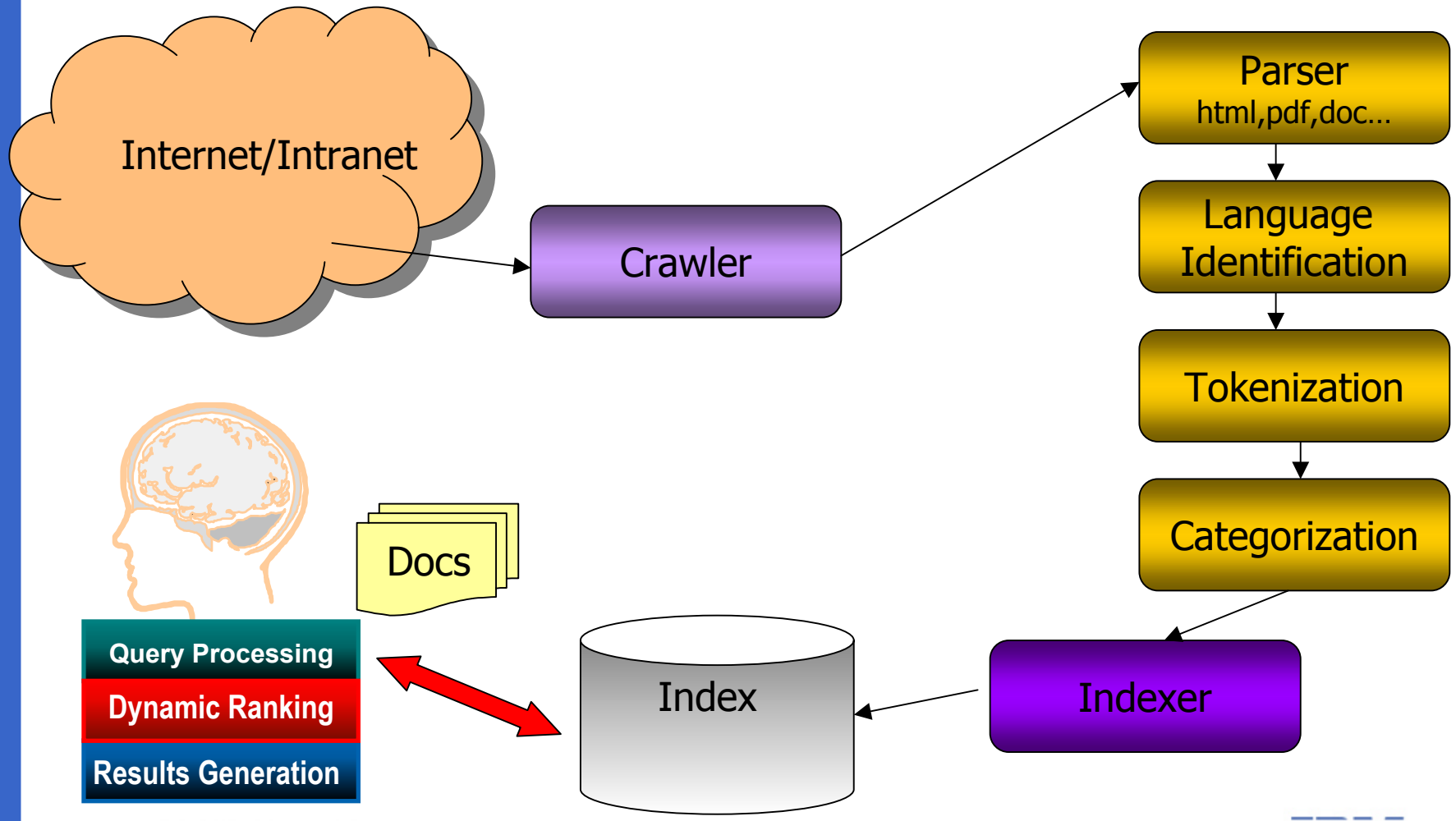
- What tools do we have to manage all this information
  - IR (Information Retrieval/Search)
    - Ubiquitous
  - IE (Information Extraction)
    - Increasingly borrowed by IR
  - Text Analytics
    - Underpins the solution
    - Computer frameworks
  - Exploiting the domain
    - Taxonomies/Ontologies
    - Divide and Conquer

Information Retrieval (IR)

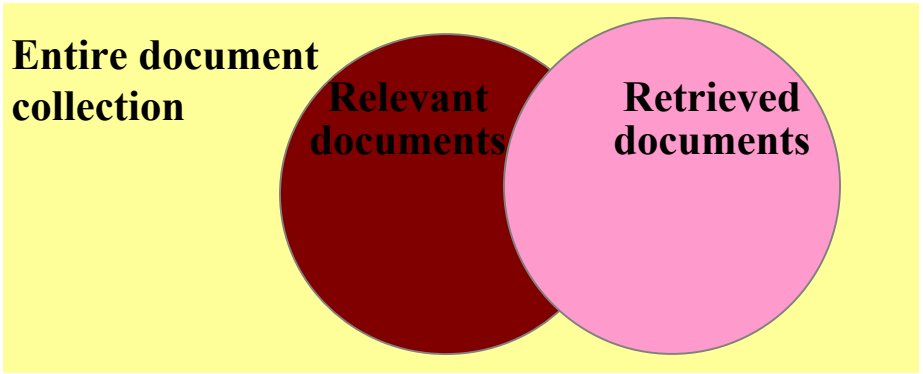
- Process of determining the relevant documents from a collection of documents, based on a query presented by the user.

- String matching
  - Query = “Hotels” => the results must contain “Hotels”
    - what about documents with “Hotel” ?
- Stemming
  - Uses the “Stem” of the word for increased recall.
    - So “Hotels” -> “Hotel” (i.e. the “s” gets chopped off)
    - therefore we find a broader set of documents.
    - Not so good for Geese -> Goose though
- Morphological Lemma
  - This is the linguistic “Stem”
  - Geese -> Goose now, “went –> go” etc…
- N-grams – works pretty well

IR Processing flow



Effectiveness measurements



	retrieved	not retrieved
irrelevant	retrieved & irrelevant	Not retrieved & irrelevant
relevant	retrieved & relevant	not retrieved but relevant

$$recall = \frac{Number\ of\ relevant\ documents\ retrieved}{Total\ number\ of\ relevant\ documents} \quad \textbf{=R}$$

$$precision = \frac{Number\ of\ relevant\ documents\ retrieved}{Total\ number\ of\ documents\ retrieved} \quad \textbf{=P}$$

Effectiveness measurements

- Precision @ n
  - No. of rel docs in first n docs
- Relevance
  - Statistical methods predominate
  - Linguistic understanding now understood to be crucial
- F-Measure: Harmonic mean of R and P

$$F = \frac{2PR}{P+R} = \frac{2}{\frac{1}{R}+\frac{1}{P}}$$

- TREC

General Opinions of Market Analysts

Delphi Group June ’04 IR Survey

- 62% express dissatisfaction or extreme dissatisfaction with their corporate IR
- 50% recognize that their organization needs to put greater emphasis on utilizing a navigable taxonomy to improve information discovery

IDC

- Linguistic capabilities embedded into broader range of applications
  - Provide more accurate information finding and concept level understanding

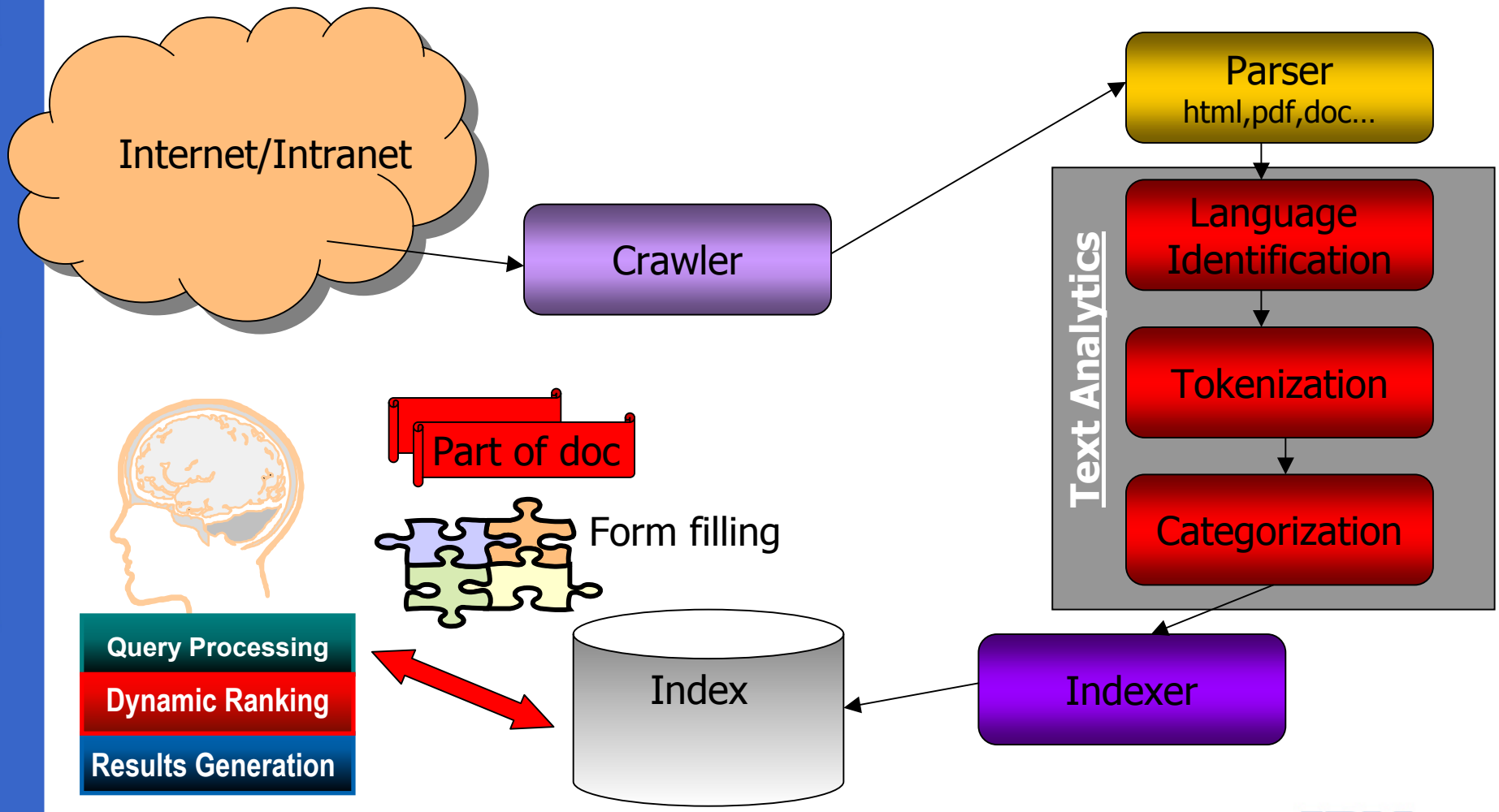
Recap – where are we?

- Somewhat relevant documents are returned to an information seeker.
- Data volume
  - meant it was ok to rely on the human to review
  - Not anymore
    - "why is information so hard to find these days" - because largely this is were we are in the cycle
- All the major players are now chasing the holy grail – "understanding of content"
  - Next stop on route – IE (not the browser!)

Information Extraction

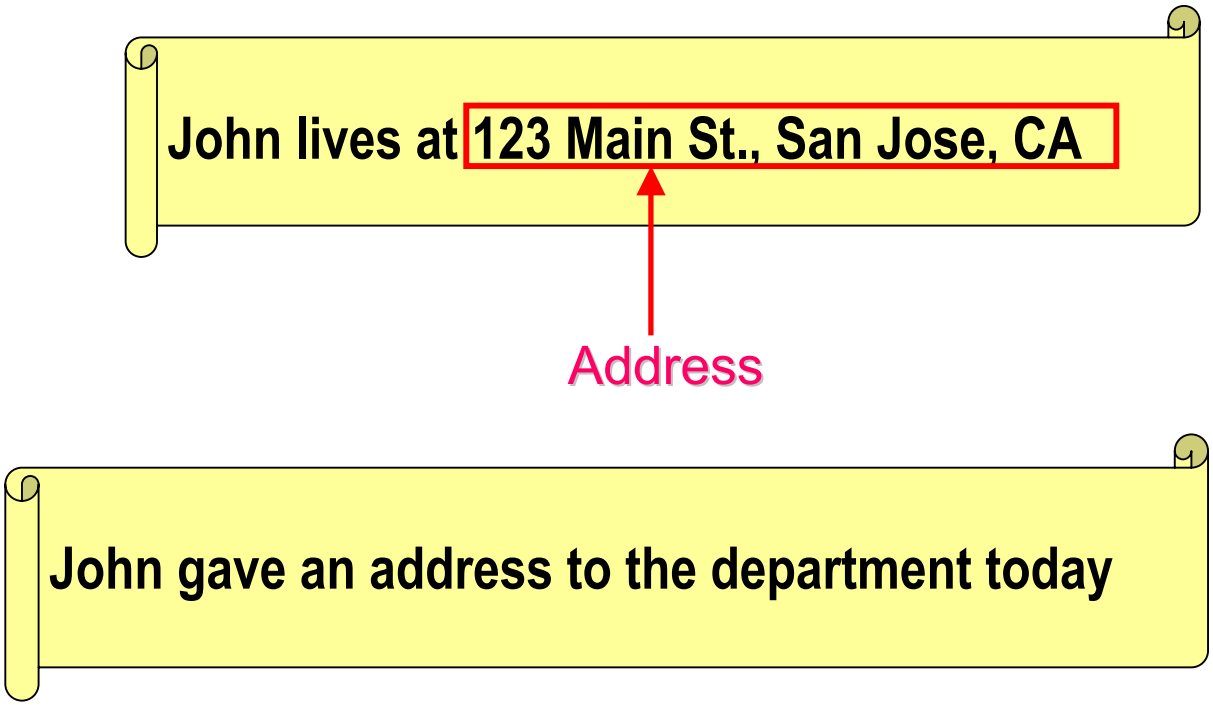
- The next step to “understanding” the text
- Returns relevant pieces of info – not documents
- Question answering systems
- Form filling
- Text Analytics…
- Rich Annotation
  - Identify “entities” in a document
  - Identify relationships between “entities”
  - Application of grammar, context (local, global)
  - normalization of concepts
  - disambiguation

IE Processing flow



IBM

The importance of semantic information

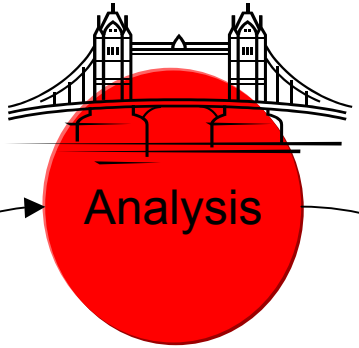


Query: *john's address*

Today, search engines will return the second document, not the first

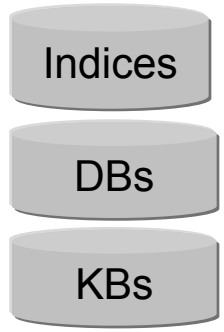
Analysis is the Bridge between the Unstructured and Structured worlds. And UIMA is the middleware that helps you build it.

Identify items of interest in unstructured data & induce structure



Unstructured Information

Structured Information



Inefficient Search

- People, Places, Org, Events
- Times, Topics, Opinions, Relationships
- Transfers, Phone Calls, Complaints
- Threats, Plots, etc.

Efficient Search

What is Text Analytics?

- Makes sense out of unstructured text
  - Segmentation/tokenization
    - Lexical Units: Multi-word units, words, subwords, …
  - Normalization
    - Relate different forms of conceptually similar units
    - Organizations ~ organisation, 三菱 ~ Mitsubishi, acquisition ~ take-over, IBM ~ International Business Machines, …
  - Annotation
    - Provides information about each lexical unit
    - E.g. Part-of-Speech

Text Analytics

- IR Objective: extract “words” and normal forms from text.
  - It is easy to extract **strings** from unstructured text
    - However, similarity between strings does not mean similarity in meaning
- IE requires deeper “understanding” of the text
  - Grammar
  - Concepts and relations
- Lets start by seeing how easy it is to get the **words** from a piece of text…
  - Sounds easy does it?
  - Use whitespace and punctuation you say?

Text Analytics: Tokenization

第10条
すべての人は、自己の権利及び義務並びに自己に対する刑事責任が決定されるに当たって、独立の公平な裁判所による公平な公開の審理を受けることについて完全に平等の権利を有する。

合 有 的 所 有 权。㈡ 任 何 人 的 财 产 不 得 任 意 剥 夺。 第 十 八 条
人 人 有 思 想、 良 心

It was only the tip of the iceberg and was the reason he went A.W.O.L. Dr. Frost said.

Arbeitsvertragsrechtsanpassungsgesetz ist der ...

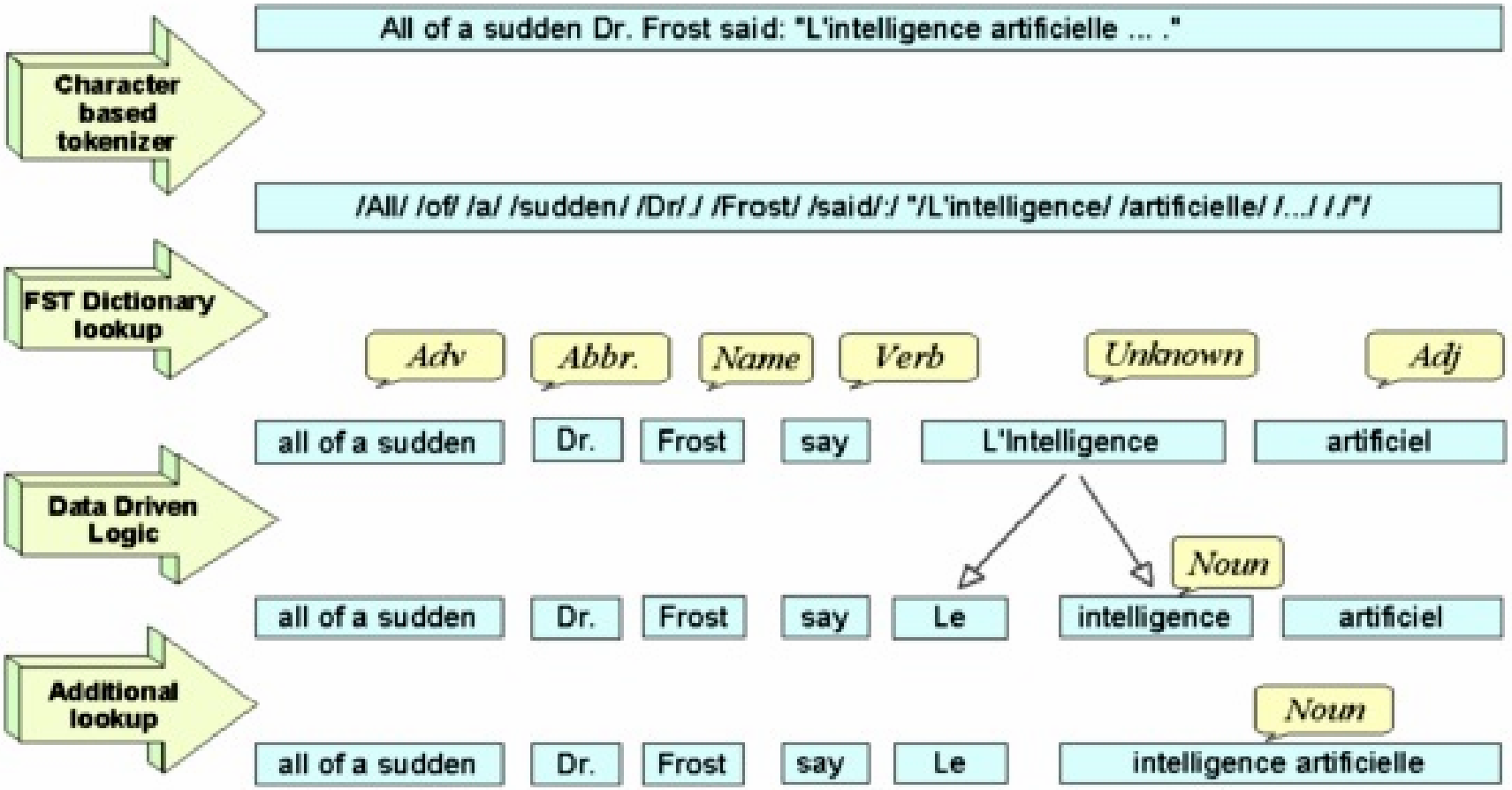
Fédération des producteurs de **pommes** de **terre** du Québec, FPPTQ.

4-Benzoxazolecarboxylic acid, 5-(methylamino)-2-((3,9,11-trimethyl-8-(1-methyl-2-oxo-2-(1H-pyrrol-2-yl)ethyl)-1,7-dioxaspiro(5.5)undec-2-yl)methyl)-, (6S-(6alpha(2S*,3S*),8beta(R*),9beta,11alpha))- => calcimycin

Text Analytics: Tokenization

- Text chunking into paragraphs, sentences, words, subwords
- Lexical Analysis
- To find **lexical units** – *atomic carriers of the information* - is the real challenge
  - This requires language understanding
    - Discriminating between word senses
    - Understanding the context within which the word is used
    - Handling word variations
      - Morphological, orthographical and derivational
- Identifying expressions such as
  - Dates/Times
  - Numbers
  - E-mail addresses
  - Chemicals

Layered analysis…



IBM

Lexical Analysis

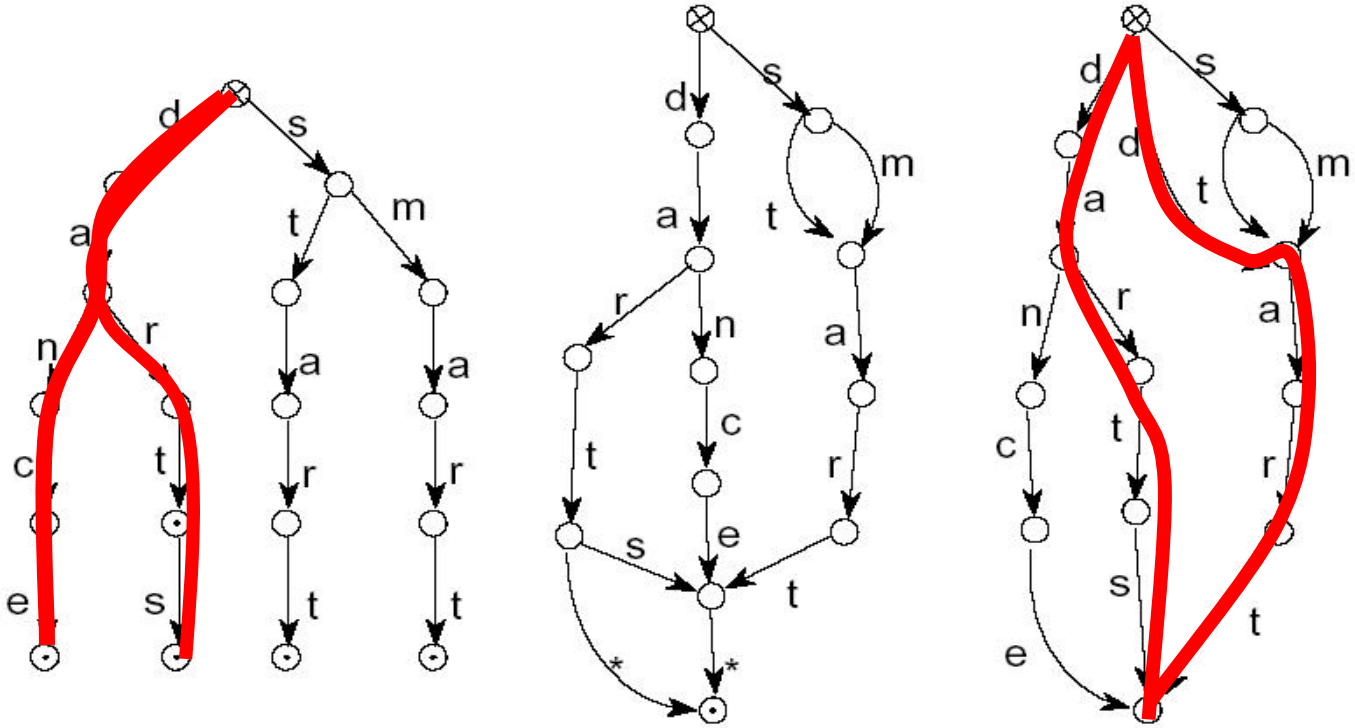
- More precise units with richer set of annotation
- Simple lexical units – internal structure is better to be ignored
  - Acquisition, takeover, the tip of the iceberg, hotdesk,
- Complex lexical units – internal structure is of semantic importance
  - Compounds
    - thom.thum@de.ibm.com
    - Weihnachtsbaum (Christmas tree)
    - N-Acetyl-Muramyl-L-Alanyl-D-Glutamic-alpha-Amide
  - Multi-word Units
    - Unstructured Information Management Architecture
    - Acetylmuramyl Alanyl Isoglutamine

Statistics vs. Linguistics

- Statistical methods
  - Learned models
  - Depend on training set, existence of corpora
- Computational Linguistics
  - Explicit models
  - Slow? – to develop new languages maybe.
- Combination of both
  - No reason why not.

Finite-State Devices for Morphology

Nodes represent states, arcs (labeled by characters) represent the transitions. Paths correspond to orthographic words.



State transition diagrams for a sample (from left to right): a) letter tree; b) deterministic directed acyclic FSA; c) non-deterministic directed acyclic FSA (borrowed from [Sgarbas et al. 2000]). Paths that correspond to the words ***dance, darts, dart*** are highlighted.

Normalization

- Normalization of lexical units
  - This means that different orthographic and/or morphological variants of the same, or similar, lexical units can be mapped to one form
  - Normalizations can be of varying conceptual or semantic depth
    - Character → Böblingen ~ Boeblingen ~ Boblingen
    - Case → Company ~ company
    - Language or script → 三菱 ~ Mitsubishi
    - Morphological → Bought ~ buy, geese ~ goose
    - Derivation → Retrovirus ~ virus, endometrium ~ endometrial ~ endometriosis, organizational ~ organization ~ organizations
    - Semantic → Muramyl Dipeptide ~ Acetylmuramyl Alanyl Isoglutamine, acquisition ~ take-over
- FST allows normalization “on-the-fly”

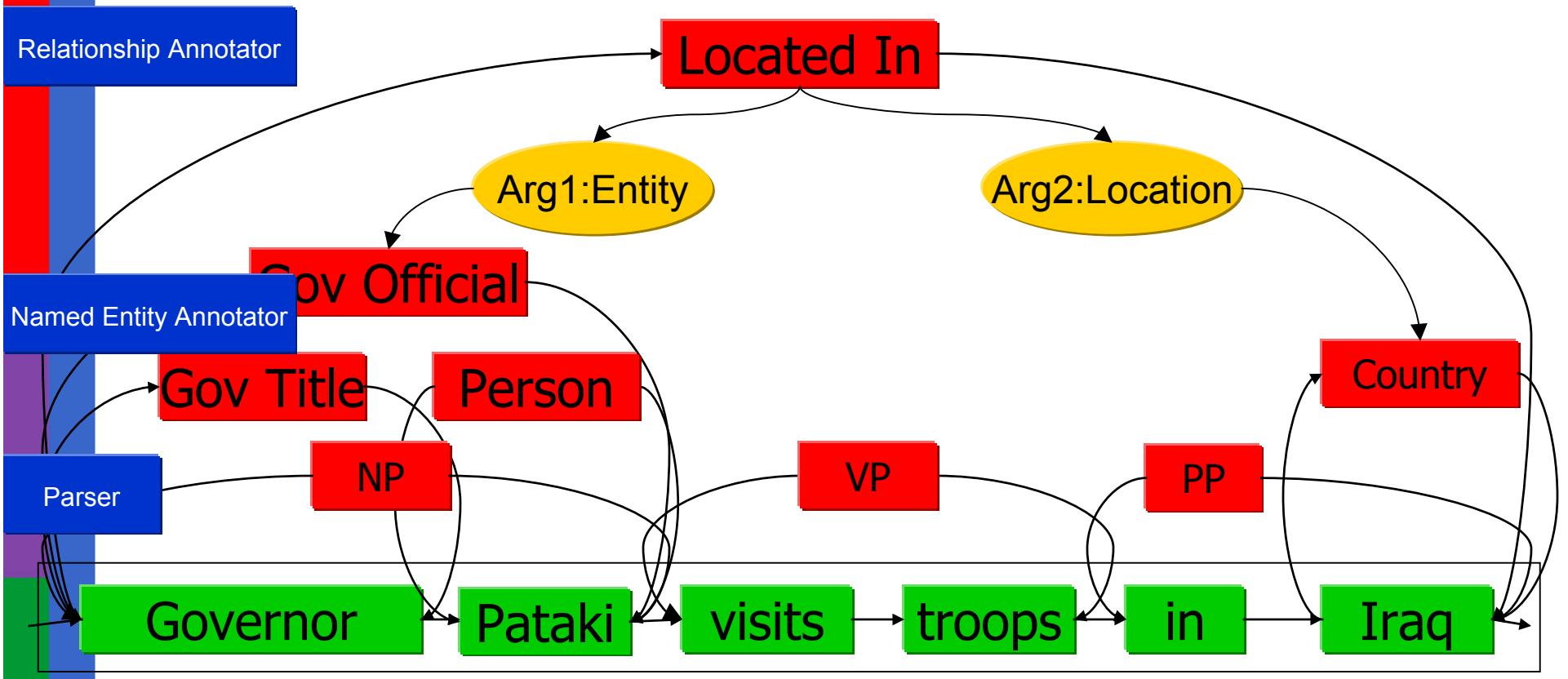
Linguistic Phenomena

Phenomena	Example
Compounding	Russian (Nuclear/thermonuclear) – термоядерное, термо-ядерных, термоядерных, термоядерного, …
Clitics	Arabic – red highlights “my” and “our”, blue highlights changed shape because of clitics. Book كتاب My Book كتابي Books كتب Our Books كتبنا
Inflection	Korean (to learn) – formal/informal inflections 배웁니다, 배웠습니다, 배우겠습니다 배워요, 배웠어요, 배우겠어요
Language Variants & standards	German – Swiss, National, pre-reform, reform English – UK, US,
Character Variants	Russian – e ~ ё, German – ß ~ ss, …
Synonyms	English – acquisition ~ take-over Arabic (lion) – د ~ ضرغام ليث سبع
Spelling Errors	English – in ~ m, Russian – шиш ~ шши (OCR)

UIMA: Unstructured Information Management Architecture

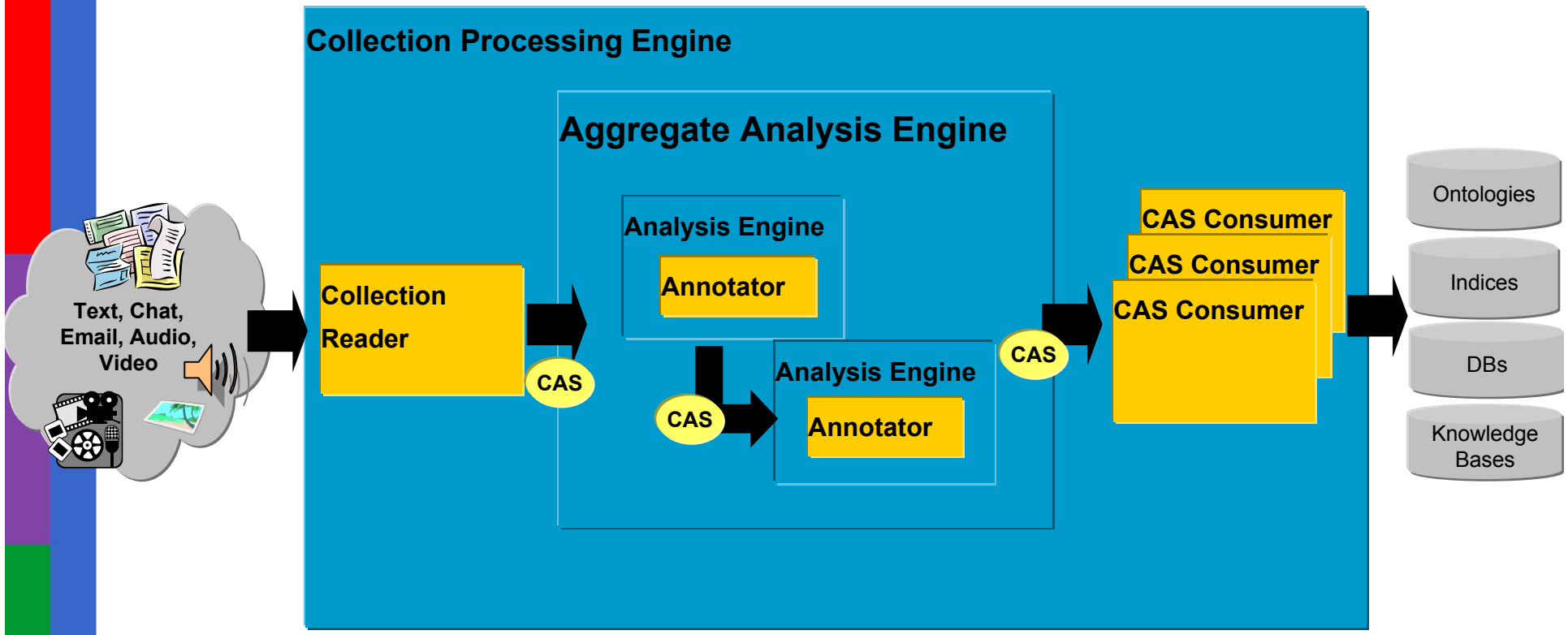
- IBM's Industry strength framework for TA
- Available generally
  - http://www.alphaworks.ibm.com/tech/uima
- Provides the software framework
- Annotators
- Collection Processing engines
- Configurable & Extensible
- Enables collaboration and reuse of assets

Annotators iterate over a document to discover new annotations based on existing ones and updates the **Common Analysis Structure** (CAS).



Component Aggregation and Encapsulation in UIMA

From document analysis to collection processing



IBM

Taxonomies, Ontologies and Domain Knowledge

- Visualizing the information space
  - Given a query, which hits only two documents because it uses a term which is not the most common way to express the concept
  - Tree navigation in a Taxonomy
  - Ontology has relationships between links, so expansion(and narrowing) of the query can be done with a view of the relationships between concepts.
  - The MeSH tree position for Calcimycin is D03.438.221.173 which has the supercategories:
    - Heterocyclic Compounds [D03]
      - Heterocyclic Compounds, 2-Ring [D03.438]
        - » Benzoxazoles [D03.438.221]
- Domain knowledge can be exploited.
  - Narrower scope
  - Usually more precise terminology and morphological usage
  - Taxonomy gives a framework for ambiguity resolution

Summary

- Why is information so hard to find?...
  - Lots of it!
  - Complicated by
    - myriad of languages
    - variety of formats
- What's being done about it?...
  - Applying
    - Linguistic knowledge through high speed lexicons
    - Conceptual representation (Ontologies)
    - All done under industry strength component frameworks
  - Involving humans in the process
    - Content author in metadata creation
    - Information seeker exposing knowledge structures
    - Information seeker feedback on relevance