

# Interactive Information Extraction

## getting results out of documents

David Milward  
Linguamatics Ltd

[david.milward@linguamatics.com](mailto:david.milward@linguamatics.com)  
[www.linguamatics.com](http://www.linguamatics.com)

# Overview

- Linguamatics
- Information overload
- IR/IE/QA
- Interactive Information Extraction

# Linguamatics

- Founded 2001 in Cambridge UK by team from SRI International
- Business: B2B, software & IP licensing to product manufacturers & service providers
- Main product areas:
  - Information Extraction
  - Flexible Spoke Dialogue Systems for automation and control

# Information Overload

- Vast amount of free-text available electronically:
  - Internet: technical articles, web pages, newsgroups
  - Intranet: company memos, minutes, procedure documents, personnel records
  - Local: email, notes
- Increasingly difficult to locate desired information:
  - Internet: search engines
  - Intranet: search engines (hopefully!)
  - Local: find, grep
- These methods are coarse and almost always generate too many spurious results

# What might you want to extract?

- Expertise information
  - Who in our organisation knows about Java
- Media analysis
  - What are the bad reports on Product Y
- Company analysis
  - Which companies are involved with Mr.X
- Bioinformatics research
  - Which protein interacts with which protein

# Getting the Information Out

- Information Retrieval
- Question Answering
- Information Extraction
- Interactive Information Extraction

# Information Retrieval (IR)

- e.g. Google, Altavista
  - Input: keywords
  - Output: relevance ordered list of documents containing (some of) those keywords.
- Advantages:
  - Quick, easy to use
- Disadvantages:
  - Limited kinds of queries: “tell me about X Y Z”
  - Output isn’t formatted appropriately for later processing
  - Keyword search is coarse grained – may want to impose extra constraints

# Question-Answering

- e.g Ask Jeeves
  - Input: natural language queries
  - Output: relevance ordered sentences extracted from source documents
- Advantages:
  - intuitive input language
  - focussed output
- Disadvantages:
  - only suitable for certain types of query:
    - questions with a single answer e.g. “who is..?” “what is..?”
    - but not “list the kings of Spain between 1800 and 1902”, “which proteins interact with other proteins?”
  - Often falls-back to straight IR

# Information Extraction (IE)

- uses linguistic analysis to enable
  - more precise search
  - results presented in a structured format e.g.
    - Person = John Smith
    - Company = Esso
    - Position = chairman
- provides all answers rather than the best match

# Information Extraction (2)

- Advantages
  - more specific queries than IR using linguistic information
  - results not documents
- Disadvantages
  - harder to use (skilled users only)
  - operates in batch mode
  - most useful for repetitive task

# Interactive Information Extraction

- Provides the benefits of IE with IR usability
  - Interactive querying by non-experts (like IR)
  - Rich query terms (like IE)
  - Structured output ready to use (like IE)
- Uses
  - one off queries
  - iterative development of advanced queries e.g. for batch use

# Linguistic information in querying

- same sentence
  - just as easy to understand as within 5 words
  - tends to suggest a real relationship between the words
- morphology
  - interacts/interacted, bind / bound
  - suggests alternatives which might otherwise be missed
- sorts
  - identify kinds of things e.g. company names, names of people, locations, dates, times, prices
- grammatical information
  - joint venture with **IBM** vs. joint venture with IBM competitor **Dell**
  - avoids spurious hits

# IIE Details

- Search for sorts (people, companies, prices etc), in the same sentence, phrase, precedence reln. etc.
- Seamlessly move from IR to IE style queries
  - “*Smith Ltd*” and “*director*” in the same document
  - “*Smith Ltd*” and “*director*” in the same sentence
  - *Person, Position, Company* in the same sentence
  - *Person* followed by *Position* followed by “*of*” followed by *Company* e.g.

John Smith director of Smith Ltd.
- Users get a feel for the accuracy of the system similar to IR
  - linguistically motivated constraints used if helpful, just like using extra keyword constraints

# Example Applications (1)

- Includes those for IE e.g.
  - Mining large journal databases for e.g. protein interactions (BioInformatics)
  - Expertise information from company documents, or from CVs on file
  - Media analysis/clippings service
  - Customer/potential customer/competitor analysis

# Example Applications (2)

- One off searches where might not program an IE engine e.g.
  - Finding companies working in a particular country
  - Finding out who has worked with Mr Jones and also has experience of Java
  - Looking at how proteins interact with a particular catalyst

# Example query (1)

The screenshot displays the 'Interactive Information Extraction' application window. The main window has a menu bar with 'File', 'New', 'View', and 'Query'. Inside, a 'sentence' window contains a 'phrase' window. The 'phrase' window is composed of four smaller windows: 'Prot 1' (containing 'protein'), 'interaction' (containing 'interacts'), 'prep' (containing 'with'), and 'Prot 2' (containing 'protein').

Interactive Information Extraction

File New View Query

sentence

phrase

Prot 1  
protein

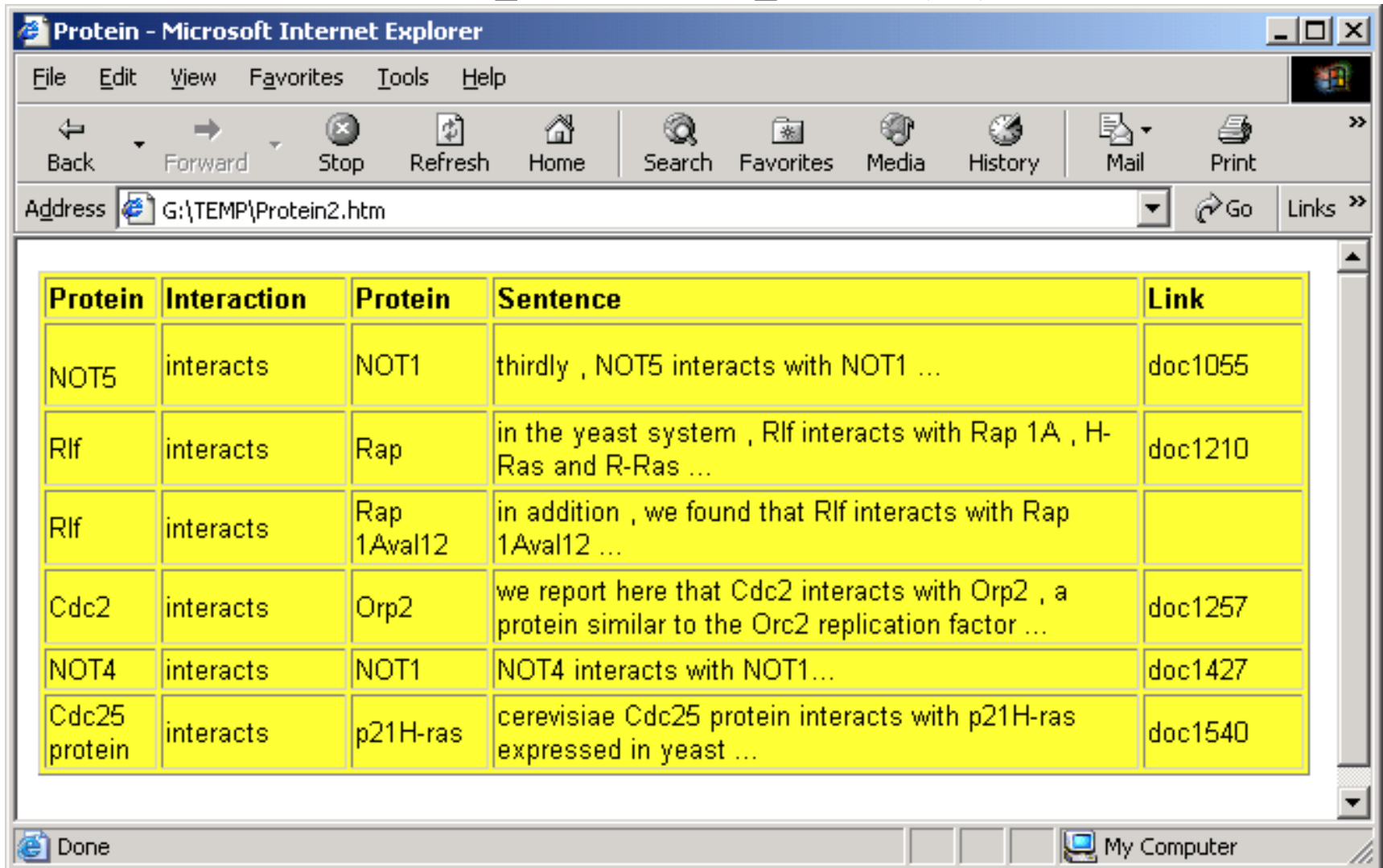
interaction  
interacts

prep  
with

Prot 2  
protein

**Linguamatics** Speech and Language Technology [www.linguamatics.com](http://www.linguamatics.com)

# Example output (1)



The screenshot shows a Microsoft Internet Explorer window titled "Protein - Microsoft Internet Explorer". The address bar displays "G:\TEMP\Protein2.htm". The main content area contains a table with the following data:

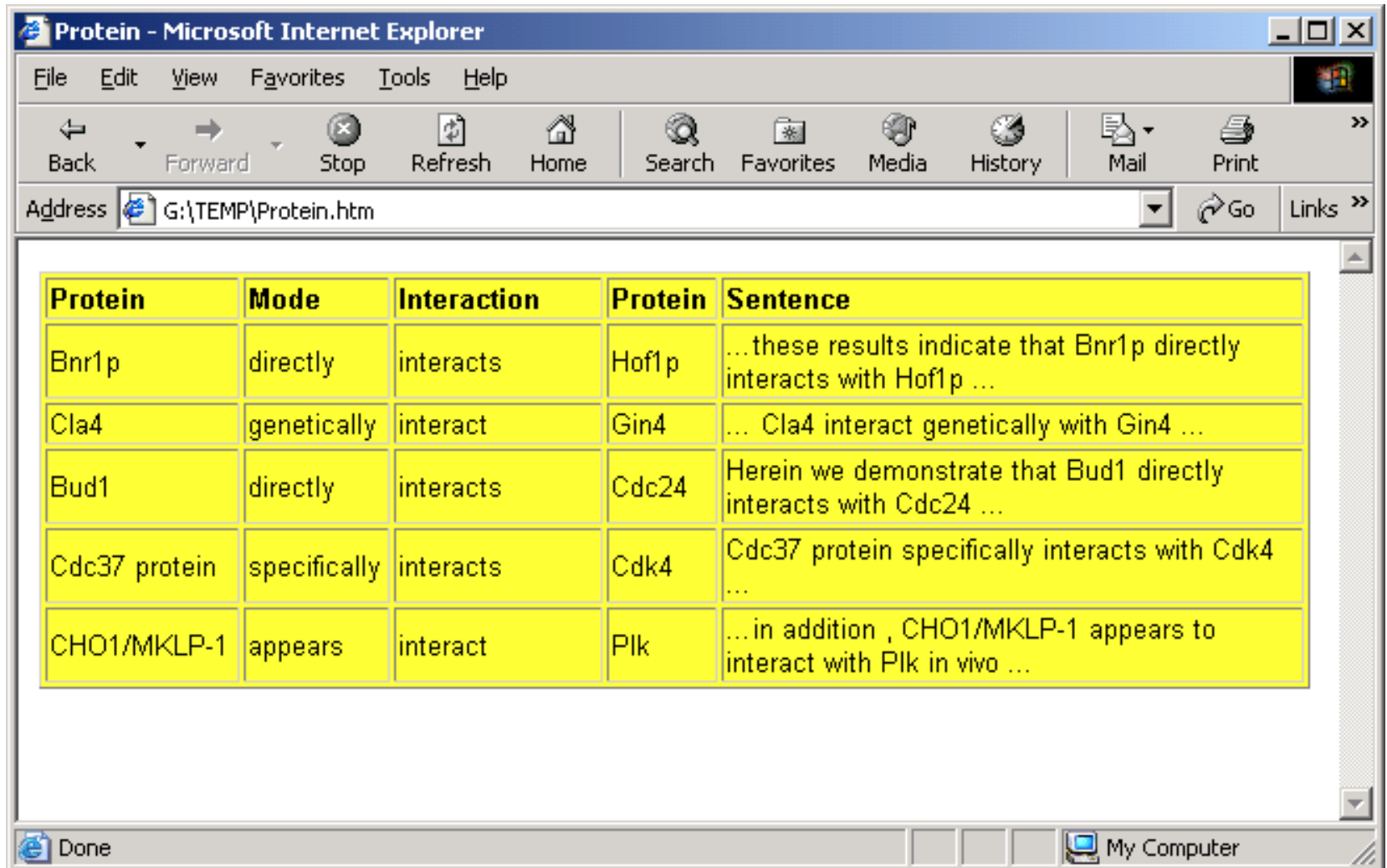
Protein	Interaction	Protein	Sentence	Link
NOT5	interacts	NOT1	thirdly , NOT5 interacts with NOT1 ...	doc1055
Rlf	interacts	Rap	in the yeast system , Rlf interacts with Rap 1A , H-Ras and R-Ras ...	doc1210
Rlf	interacts	Rap 1Aval12	in addition , we found that Rlf interacts with Rap 1Aval12 ...	
Cdc2	interacts	Orp2	we report here that Cdc2 interacts with Orp2 , a protein similar to the Orc2 replication factor ...	doc1257
NOT4	interacts	NOT1	NOT4 interacts with NOT1...	doc1427
Cdc25 protein	interacts	p21H-ras	cerevisiae Cdc25 protein interacts with p21H-ras expressed in yeast ...	doc1540

# Example query (2)

The screenshot displays the 'Interactive Information Extraction' software interface. The main window is titled 'sentence' and contains a 'phrase' window. Inside the 'phrase' window, there are four sub-windows: 'Prot 1' (containing 'protein'), 'VP' (containing 'adv' with 'ADVP'), 'prep' (containing 'to'), and 'Prot 2' (containing 'protein'). The 'VP' window is further nested, containing an 'interaction' window which in turn contains an 'interact+' window with the text 'interact'. The 'prep' window also contains a 'to' window with the text 'to'. The 'interact+' and 'to' windows are highlighted in yellow. The interface includes a menu bar with 'File', 'New', 'View', and 'Query' options. At the bottom of the window, the Linguamatics logo and website information are visible.

**Linguamatics** Speech and Language Technology [www.linguamatics.com](http://www.linguamatics.com)

# Example output (2)



The screenshot shows a Microsoft Internet Explorer window titled "Protein - Microsoft Internet Explorer". The address bar displays "G:\TEMP\Protein.htm". The main content area contains a table with five rows of protein interaction data. The table has five columns: Protein, Mode, Interaction, Protein, and Sentence. The data is as follows:

Protein	Mode	Interaction	Protein	Sentence
Bnr1p	directly	interacts	Hof1p	...these results indicate that Bnr1p directly interacts with Hof1p ...
Cla4	genetically	interact	Gin4	... Cla4 interact genetically with Gin4 ...
Bud1	directly	interacts	Cdc24	Herein we demonstrate that Bud1 directly interacts with Cdc24 ...
Cdc37 protein	specifically	interacts	Cdk4	Cdc37 protein specifically interacts with Cdk4 ...
CHO1/MKLP-1	appears	interact	Plk	...in addition , CHO1/MKLP-1 appears to interact with Plk in vivo ...

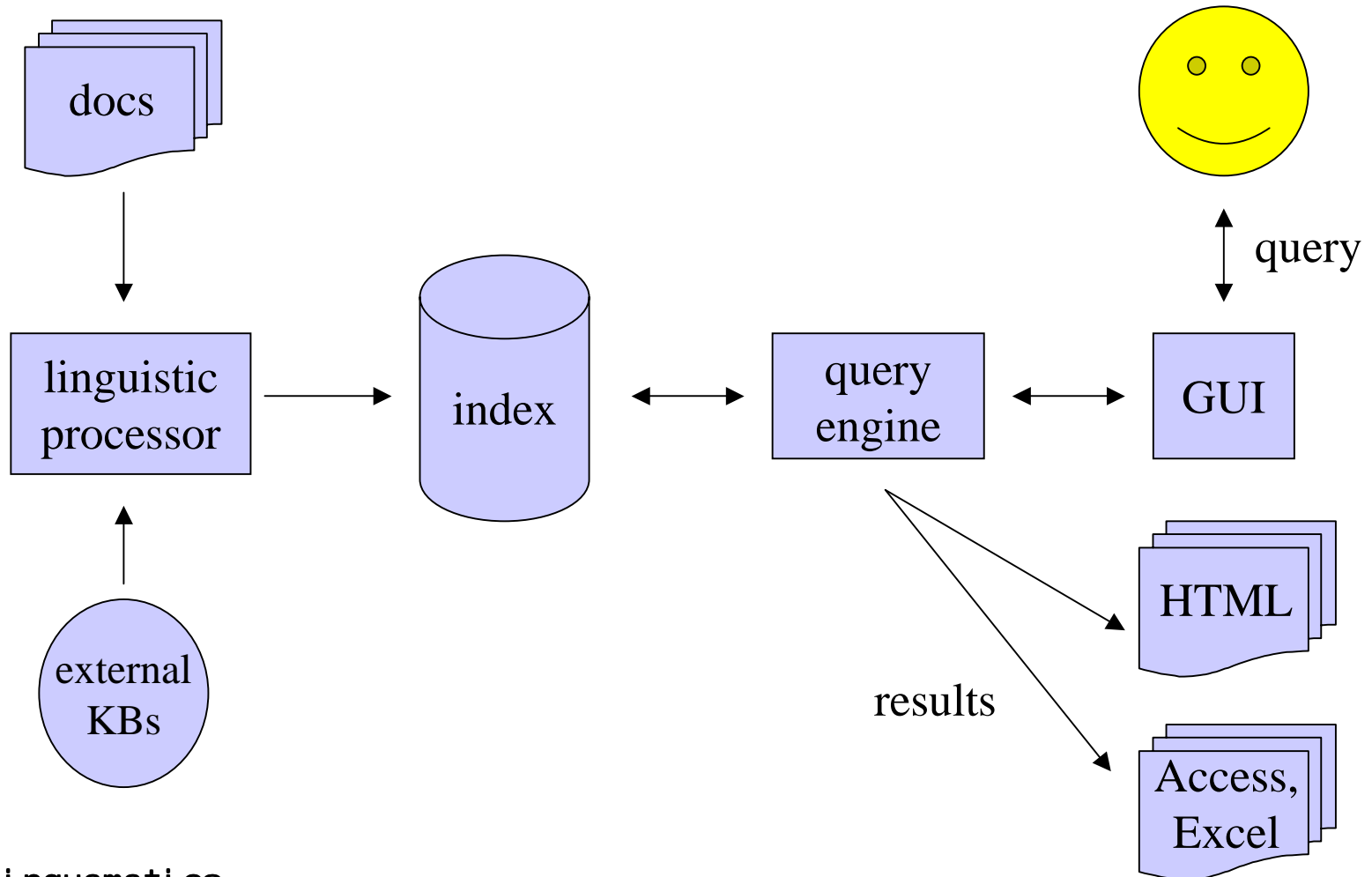
# How Do We Do This?

- Linguistic analysis
  - part of speech information (noun, verb) using statistical methods
  - morphology (*bind, bound; interacts, interacted*)
- Use of knowledge sources
  - identify entities: people, places, times, dates, proteins...
- Don't discard useful information!
  - positional information
  - keep all words (unlike most IR)
- Choose what information to output and in what format

# How can we do querying in real time?

- Proprietary indexing scheme for linguistic structure
- Efficient querying algorithm on indexed structures
- documents undergo linguistic pre-processing and indexing once only
  - on first access, or when edited

# Separate Indexing and Querying



# Modes of Use

- **Interactive** for querying over document corpus/emails/minutes/intranet etc.
- **Batch** for generating large databases to be queried with standard DB tools
  - Both modes use the same graphical interface to construct queries

# Extracting further information from batch queries

Canonicalised output format enables DB queries

- indirect relationships

- e.g.

- If

- Mr Jones is chairman of XYZ Corp.

- XYZ Corp. in a joint venture with ABC Corp.

- then

- Mr Jones and ABC Corp. related through XYZ Corp.

- statistics

- e.g. 30% of reports praised the new washing machine

# IIE Summary

- Interactive extraction of information from unstructured text
- Suitable for non-specialists
  - no need to be a programmer or a linguist
- Unique blend of IE and IR
  - interactive query, IR style
  - more powerful IE-style queries and output