

# Introduction to Text Mining

## Module 1: NLP Foundations



## Course summary

- This course introduces the topic of text mining, and is split into 4 modules
  - Introduction to Text Mining: NLP foundations
  - Semantic Annotation, Ontology-based Information Extraction and Semantic Indexing
  - Opinion Mining
  - Applications and Evaluation

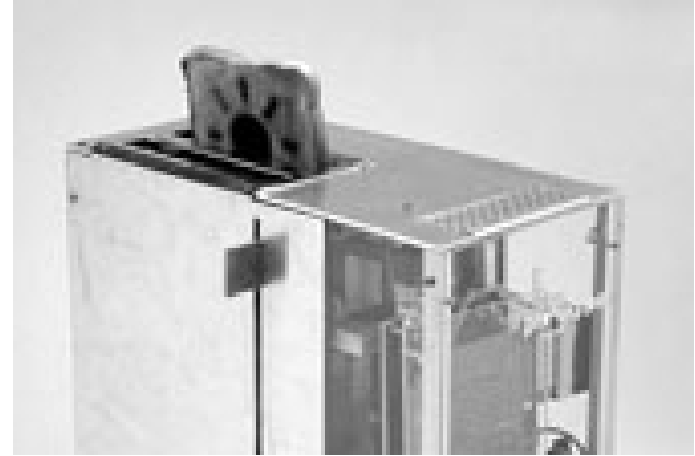
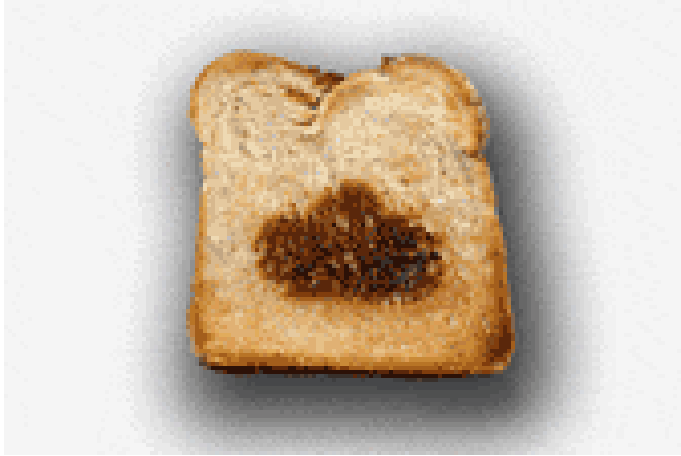
## Topics in this module

- What is Text Mining and why do we need it?
- NLP foundations
- GATE and other text mining tools

Oddly enough, people have successfully combined information and toast...

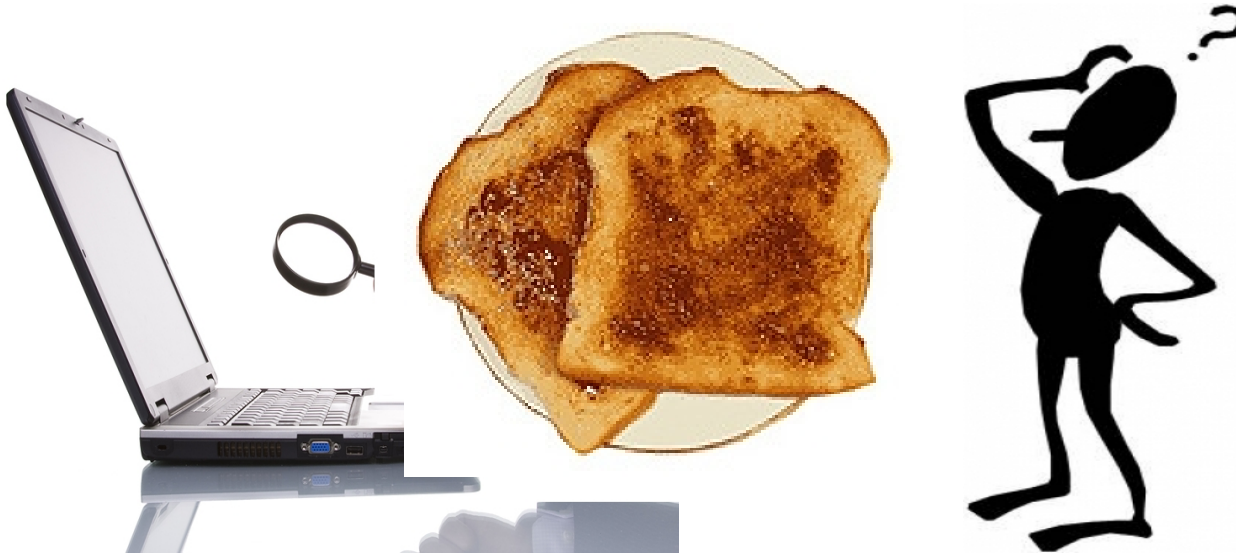


## The weather-forecasting toaster is 10 years old



- This weather-forecasting toaster, connected to a phone point, was designed in 2001 by a PhD student
- It accessed the MetOffice website via a modem inside the toaster and translated the information into a 1, 2 or 3 for rain, cloud or sun
- The relevant symbol was then branded into the toast in the last few seconds of toasting

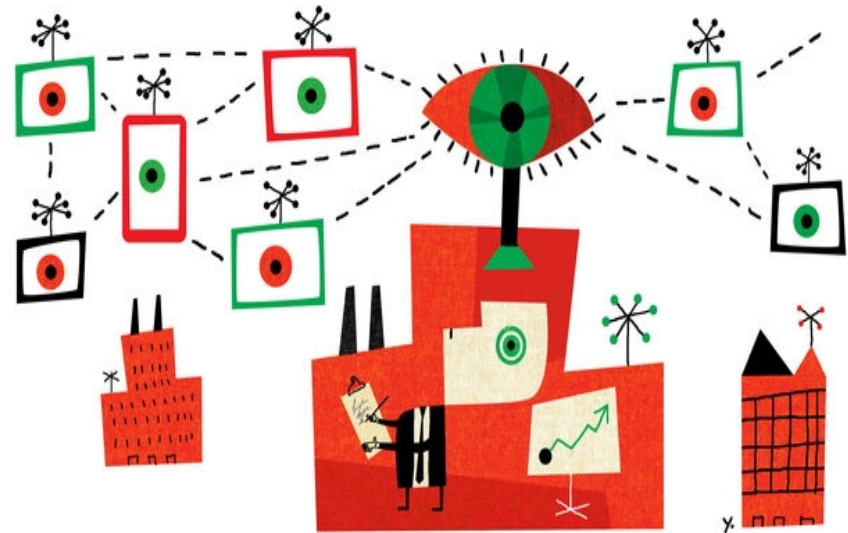
# With tools such as these, why do we need text mining?



- It turns out that toast isn't actually a very good medium for finding information

## Information overload or filter failure?

- We all know that there's masses of information available online these days, and it's constantly growing
- You often hear people talk about “information overload”
- But as Clay Shirky says, the real problem is not the amount of information, but our inability to filter it correctly
- You can hear his talk on this topic here:  
<http://bit.ly/oWJTNZ>



## Value vs volume

- Trying to get specific information out of the huge volumes of data on the internet can be an overwhelming task
- If you have content that is sufficiently high value or low volume, then you can use sophisticated methods to help people find, browse or abstract over that content
- These methods include building symbolic models (taxonomic, logical, conceptual, semantic...) of the subject matter and annotating content with references to those models
- **Technological** success: the expressivity of the modelling languages and the quality of the annotation algorithms, and of the indexing, search and browsing tools
- **Social** success: the level of expertise and the quantity of time and effort deployed by the people building models and extraction patterns (or creating training data and running learning algorithms), or tuning indices or user interfaces



# It is difficult to access unstructured information efficiently

Text mining tools can help you:

- Save time and money on **management of text** and data from multiple sources
- Find **hidden links** scattered across huge volumes of diverse information
- Integrate **structured data** from variety of sources
- **Interlink** text and data
- **Collect information** and extract new facts

## What is text mining?

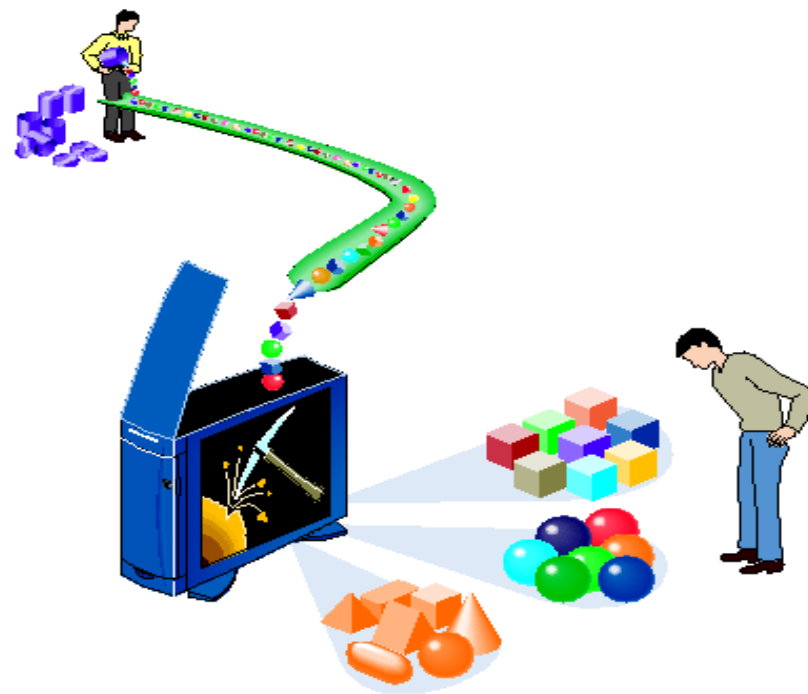
- Text Mining is the automatic discovery of new, previously unknown information, by automatically extracting information from different textual resources.
- A key element is the linking together of the extracted information together to form new facts or new hypotheses to be explored further
- In text mining, the goal is to discover previously unknown information, i.e. something that no one yet knows

## Text Mining is not Data Mining

Data mining is about using analytical techniques to find interesting patterns from large structured databases

Examples:

- using consumer purchasing patterns to predict which products to place close together on shelves in supermarkets
- analysing spending patterns on credit cards to detect fraudulent card use.



## Text Mining is not Web Search

- Text mining is also different from traditional web search.
- In search, the user is typically looking for something that is already known and has been written by someone else.
- The problem lies in sifting through all the material that currently isn't relevant to your needs, in order to find the information that is.
- The solution often lies in better ways to ask the right question
- You can't ask Google to tell you about all the speeches made by Tony Blair about foot and mouth disease while he was Prime Minister, or all documents in which a person born in Sheffield is quoted as saying something.

## Text Mining is not (quite) Information Extraction

- There are programs that can, with reasonable accuracy, extract information from text such as names of people, organisations, locations and so on, and find relations between them (e.g. John works for the BBC)
- Information Extraction (IE) is about getting facts out of unstructured information
- It's a major component of text mining, but it doesn't tell the full story
- In a criminal investigation, finding the facts (names of witnesses, alibis for the night of the murder) are like the IE component
- Text mining is about making the deductions as to who could or could not have committed the murder



# Ghostbusters?



# Sherlock Holmes?





Or the police?



# Threat tracking: extraction and linking of facts

The screenshot shows the Threat Tracker application window. On the left, a sidebar contains several sections: 'Huda Ammash 5 Hearts', 'Saved Searches' (with a search for 'Huda Salih Mahdi Ammash'), 'Entities' (with '070 Huda Ammash' circled), 'SubTrackers', and 'Documents'. A yellow callout box points to the '070' in the Entities section, stating 'Count of how many documents system found'. The main content area has tabs for 'Excerpt', 'Documents', and 'Concordance'. The 'Excerpt' tab is active, displaying a lead paragraph: 'Displaying lead paragraphs from documents mentioning Huda Ammash. Displaying 1 to 10 of approximately 70 documents.' The number '70' is circled. Below this is a document entry titled '2.22 - Mrs Anthrax is 18th leader caught - Original - Add to Tracker'. The text of the document is partially visible, with several words highlighted in yellow: 'Mrs Anthrax', 'Huda Salih Mahdi Ammash', 'US', 'Iraqi', 'five of hearts', 'her', and 'weapons of mass destruction'. A yellow callout box points to the ellipsis '...' in the document text, stating: '... indicates sentences have been removed because they don't mention the Entity'. Another yellow callout box points to the word 'her' in the document text, stating: 'Mentions of the Entity Hudda Ammash found by ThreatTrackers'. The browser window title is 'Threat Tracker - Alias I, Inc - Microsoft Internet Explorer'.

# Text mining stages

- Document selection and filtering (IR techniques)
- Document pre-processing (NLP techniques)
- Document processing (NLP / ML / statistical techniques)

# Stages of document processing

- Document selection involves identification and retrieval of potentially relevant documents from a large set (e.g. the web) in order to reduce the search space. Standard or semantically-enhanced IR techniques can be used for this.
- Document pre-processing involves cleaning and preparing the documents, e.g. removal of extraneous information, error correction, spelling normalisation, tokenisation, POS tagging, etc.
- Document processing generally consists of information extraction (NER, relation/event recognition etc.) and potentially opinion mining.

## NLP components for text mining

- A text mining system is usually built up from a number of different NLP components
- First, you need some Information Extraction tools (the police) to do the donkey work of getting all the relevant pieces of information and facts.
- Then you need some tools to apply the reasoning (Sherlock Holmes) to the facts, e.g. opinion mining, information aggregation, semantic technologies, dynamics analysis and so on.
- GATE is an example of a tool for text mining which allows you to combine all the necessary NLP components
- First, we'll look at IE.

## IE for Document Access

- With traditional query engines, getting the facts can be hard and slow
  - Where has the Queen visited in the last year?
  - Which airports are currently closed due to the volcanic ash?
- Which search terms would you use to get these?
- How can you specify you want to see someone's home page?
- IE returns information in a structured way
- IR returns documents containing the relevant information somewhere

# Approaches to Information Extraction

## Knowledge Engineering

- rule based
- developed by experienced language engineers
- make use of human intuition
- easier to understand results
- development could be very time consuming
- some changes may be hard to accommodate

## Learning Systems

- use statistics or other machine learning
- developers do not need LE expertise
- requires large amounts of annotated training data
- some changes may require re-annotation of the entire training corpus

# Information Extraction with GATE



## GATE: the Swiss Army Knife of NLP

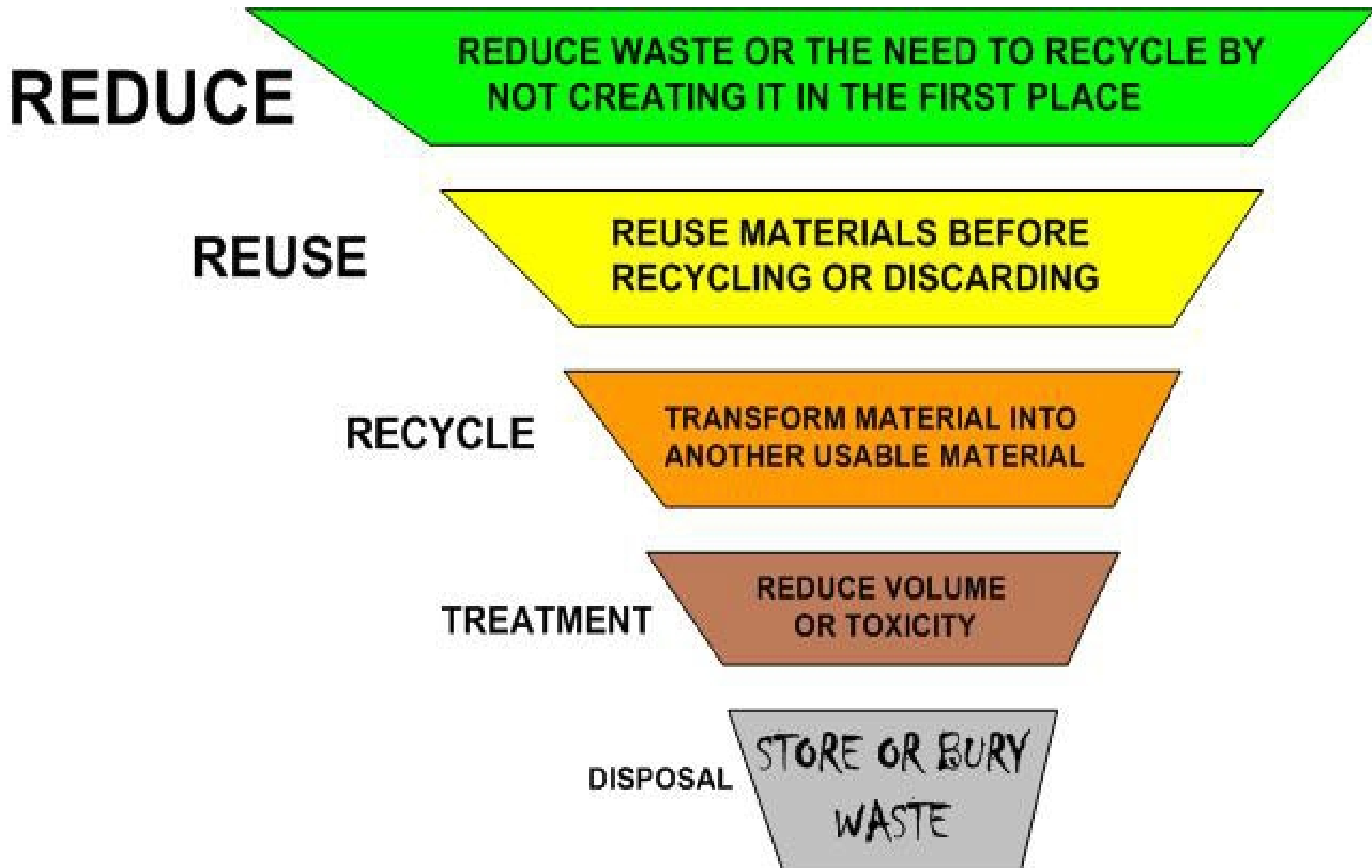
- An attachment for almost every eventuality
- Some are hard to prise open
- Some are useful, but you might have to put up with a bit of clunkiness
- Some will only be useful once in a lifetime, but you're glad to have them just in case.
- There are many imitations, but nothing like the real thing.



## History of GATE

- **early 1990s**: you want me to write that all over again?
- **1995-7**: first GATE (and "large-scale IE") project
- **1996**: GATE 1: Tcl/Tk, Perl, C++, ...
- **2002**: release of completely rewritten version 2, 100% Java
- **2011**: mature ecosystem with established community
  - Tens of thousands of research users
  - 25,000 downloads per year
  - commercial users getting serious

# Text mining tools need to be eco-friendly



## What exactly is GATE?

- An **architecture**: a macro-level organisational picture for HLT software systems
- A **framework**: for programmers, GATE is an object-oriented class library that implements the architecture.
- A **development environment**: for language engineers, computational linguists et al, a graphical development environment.
- A **community** of users and contributors

## Architectural principles

- Non-prescriptive, theory neutral (strength and weakness)
- Re-use, interoperation, not reimplementation (e.g. diverse XML support, integration of Protégé, Jena, Yale...)
- (Almost) everything is a component, and component sets are user-extendable
- (Almost) all operations are available both from API and GUI

## In short...

GATE includes:

- **components** for language processing, e.g. parsers, machine learning tools, stemmers, IR tools, IE components for various languages...
- tools for **visualising** and **manipulating** text, annotations, ontologies, parse trees, etc.
- **various information extraction** tools
- **evaluation** and **benchmarking** tools

## GATE components

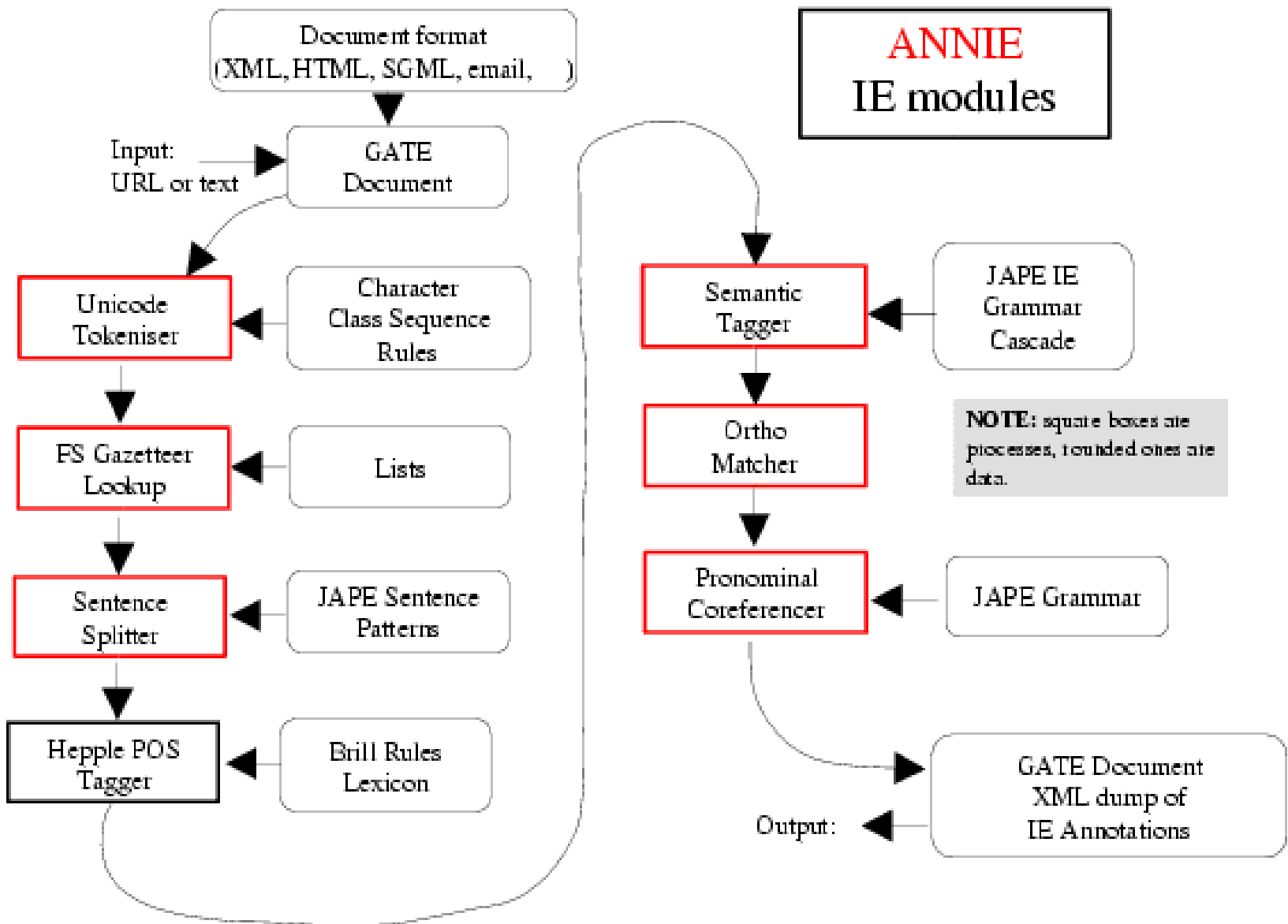
- **Language Resources** (LRs), e.g. lexicons, corpora, ontologies
- **Processing Resources** (PRs), e.g. parsers, generators, taggers
- **Visual Resources** (VRs), i.e. visualisation and editing components
- Algorithms are separated from the data, which means:
  - the two can be developed independently by users with different expertise.
  - alternative resources of one type can be used without affecting the other, e.g. a different visual resource can be used with the same language resource

# ANNIE

- **ANNIE** is GATE's rule-based IE system
- It uses the language engineering approach (though we also have tools in GATE for ML)
- Distributed as part of GATE
- Uses a finite-state pattern-action rule language, JAPE
- ANNIE contains a reusable and easily extendable set of components:
  - generic preprocessing components for tokenisation, sentence splitting etc
  - components for performing NE on general open domain text



# ANNIE Modules



# Named Entity Recognition: the cornerstone of IE

Traditionally, NE is the identification of proper names in texts, and their classification into a set of predefined categories of interest

- Person
- Organisation (companies, government organisations, committees, etc)
- Location (cities, countries, rivers, etc)
- Date and time expressions

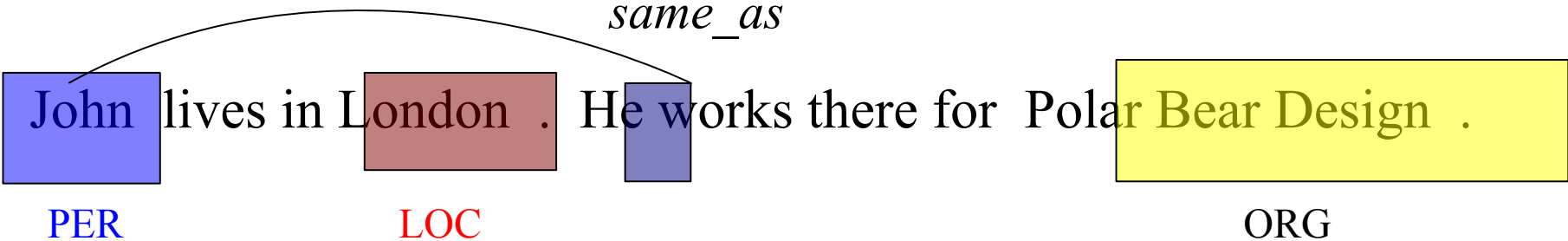
Various other types are frequently added, as appropriate to the application, e.g. newspapers, ships, monetary amounts, percentages etc.

## Basic NE Recognition

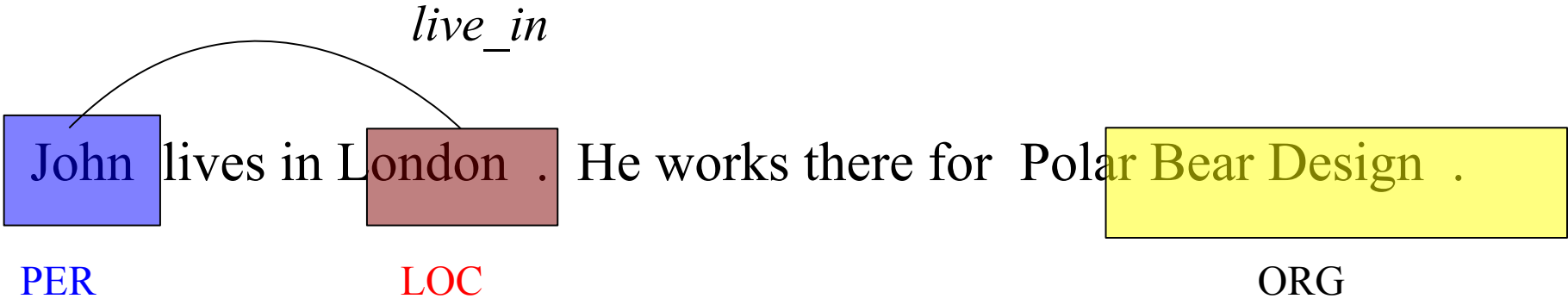
John lives in London . He works there for Polar Bear Design .

PERSON LOCATION ORGANISATION

# Co-reference




# Relations







## Hands-on: ANNIE (1)

- Start GATE by double clicking on the icon
- Because ANNIE is a ready-made application, we can just load it directly from the menu
- Click the  icon from the top GATE menu OR Select File → Load ANNIE system
- Select “with defaults”
- Create a new corpus (right click on Language Resources → New → GATE corpus and click “OK”)
- Populate the corpus (right click on the corpus, and use the file chooser in Directory URL to select the news-texts directory. Click “Open” and then “OK”)



## Hands-on: ANNIE (2)

- Double click on ANNIE and select “Run this application”
- This will now run the application on all the documents in your corpus
- Double click on a document to open it
- Select “Annotation Sets” and “Annotations” from the top tabs
- Click on the top arrow above “Original markups” in the right pane
- 
-

## Inspecting the results

- You should see a mixture of Named Entity annotations (Person, Location etc) and some other linguistic annotations (Token, Sentence etc) in the Default annotation set
- Click on the name of an annotation in the right hand pane that you want to view (you can select multiple annotations)
- All annotations of that type will be highlighted in the main text window
- Hover over the annotation in the text to get a popup box with more info

## Let's look at the PRs

- Each PR in the ANNIE pipeline creates some new annotations, or modifies existing ones
- Document Reset → removes annotations
- Tokeniser → Token annotations
- Sentence Splitter → Sentence, Split annotations
- Gazetteer → Lookup annotations
- POS tagger → adds category features to Token annotations
- JAPE transducer → Date, Person, Location, Organisation, Money, Percent annotations
- Orthomatcher → adds match features to NE annotations

# Tokeniser

- Tokenisation based on Unicode classes
- Declarative token specification language
- Produces Token and SpaceToken annotations with features orthography and kind
- Length and string features are also produced
- Rule for a lowercase word with initial uppercase letter

```
"UPPERCASE_LETTER" LOWERCASE_LETTER"* >  
  Token; orthography=upperInitial; kind=word
```

# Document with Tokens

Annotation Sets Annotations List Annotations Stack Class Co-reference Editor Instance Text 🔍 ▼

Union Appeals For Talks To End BA Strike

Skip to navigation | Skip to content |  
 Home | Contact Us | News Search;  
 HubPage  
 Airwise News  
 Airport Guide  
 Airwise Travel  
 Search  
 Union Appeals For Talks To End BA Strike  
 March 22, 2010

Union leaders on Sunday called for talks with British Airways bosses to end strike action by cabin crew that has led to the cancellation of hundreds of flights and disrupted travel plans for thousands of passengers.

Type	Features
Token	{ category=NNP, kind=word, length=5, orth=upperInitial, string=Union}
Token	{ category=NNPS, kind=word, length=7, orth=upperInitial, string=Appeals}
Token	{ category=IN, kind=word, length=3, orth=upperInitial, string=For}
Token	{ category=NNS, kind=word, length=5, orth=upperInitial, string=Talks}
Token	{ category=TO, kind=word, length=2, orth=upperInitial, string=To}

- Date
- FirstPerson
- JobTitle
- Location
- Lookup
- Money
- Organization
- Percent
- Person
- Sentence
- SpaceToken
- Split
- Title
- Token
- Unknown
- ▶ Original markups

## ANNIE English Tokeniser

- The English Tokeniser is a slightly enhanced version of the Unicode tokeniser
- It comprises an additional JAPE transducer which adapts the generic tokeniser output for the POS tagger requirements
- It converts constructs involving apostrophes into more sensible combinations
  - don't → do + n't
  - you've → you + 've

## Sentence Splitter

- The default splitter finds sentences based on Tokens
- Creates Sentence annotations and Split annotations on the sentence delimiters
- Uses a gazetteer of abbreviations etc. and a set of JAPE grammars which find sentence delimiters and then annotate sentences and splits

# Document with Sentences

Annotation Sets Annotations List Annotations Stack Class Co-reference Editor Instance Text

the opposition conservatives, ahead in opinion polls, have been turning up the pressure on Labour over its links to Unite, saying the government had failed to take action quickly enough because it did not want to alienate its financial backers.

"We deplore the strike, and the prime minister and the transport secretary have said that absolutely clearly," Foreign Secretary David Miliband told Sky News.

"The way to resolve these disputes is through negotiation, it is damaging for the company, it is damaging for the crews and it is damaging for the country."

The dispute arose because BA, which has 12,000 cabin crew, wants to save an annual GBP£62.5 million pounds (USD\$95 million) to help cope with a fall in demand, volatile fuel prices and increased competition from low-cost carriers.

A spokesman said there was no estimate yet as to how much the industrial action would cost the company.

Type	Features
Sentence	{}
Sentence	{}
Sentence	{}
Sentence	{}
Sentence	{}

- Date
- FirstPerson
- JobTitle
- Location
- Lookup
- Money
- Organization
- Percent
- Person
- Sentence
- SpaceToken
- Split
- Title
- Token
- Unknown
- Original markups



## Sentence splitter variants

- An alternate set of rules can be loaded with the regular sentence splitter
- The main difference is the way it handles new lines
- In some cases, you might want a new line to signal a new sentence, e.g. addresses
- In other cases, you might not, e.g. in emails that have been split by the email program
- A regular expression Java-based splitter is also available, called RegEx Sentence Splitter, which is sometimes faster
- This handles new lines in the same way as the default sentence splitter

## POS tagger

- ANNIE POS tagger is a Java implementation of Brill's transformation based tagger
- Previously known as **Hepple Tagger** (you may find references to this and to **heptag**)
- Trained on WSJ, uses Penn Treebank tagset
- Default ruleset and lexicon can be modified manually (with a little deciphering)
- Adds category feature to Token annotations
- Requires Tokeniser and Sentence Splitter to be run first

## Morphological analyser

- Not an integral part of ANNIE, but can be found in the Tools plugin as an “added extra”
- Flex based rules: can be modified by the user (instructions in the User Guide)
- Generates “root” feature on Token annotations
- Requires Tokeniser to be run first
- Requires POS tagger to be run first if the considerPOSTag parameter is set to true

# Gazetteers

- Gazetteers are plain text files containing lists of names (e.g rivers, cities, people, ...)
- The lists are compiled into Finite State Machines
- Each gazetteer has an index file listing all the lists, plus features of each list (majorType, minorType and language)
- Lists can be modified using the Gazetteer Editor
- Gazetteers generate Lookup annotations with relevant features corresponding to the list matched
- Lookup annotations are used primarily by the NE transducer
- Various different kinds of gazetteer are available, which enable more flexible matching, e.g on the root form of a word

# Gazetteer editor

File Options Tools Help

Messages 1269258352.html... ANNIE ANNIE Gazetteer

airport.lst New List

List name	Major	Minor	Language
charities.lst	organization		
city.lst	location	city	
city_cap.lst	location	city	
company.lst	organization	company	
company_cap.lst	organization	company	
country.lst	location	country	
country_abbrev.lst	location	country_abbrev	
country_adj.lst	country_adj		
country_cap.lst	location	country	
currency_prefix.lst	currency_unit	pre_amount	
currency_unit.lst	currency_unit	post_amount	
date_key.lst	date_key		
date_unit.lst	date_unit		
day.lst	date	day	
day_cap.lst	date	day	
department.lst	organization	departmen	

New Entry Add Cols

Value
Aaccra
Aalborg
Aarhus
Ababa
Abadan
Abakan
Aberdeen
Abha
Abi Dhabi
Abidjan
Abilene
Abu
Abu Dhabi
Abuja
Acapulco

Filter: 1993 entries

Gazetteer Initialisation Parameters Gazetteer Editor

definition file entries

entries for selected list

## The ANNIE gazetteer

- The ANNIE gazetteer has about 60,000 entries arranged in 80 lists
- Each list reflects a certain category, e.g. airports, cities, first names etc.
- List entries might be entities or parts of entities, or they may contain contextual information (e.g. job titles often indicate people)

## NE transducer

- Gazetteers can be used to find terms that suggest entities
- However, the entries can often be ambiguous
  - “May Jones” vs “May 2010” vs “May I be excused?”
  - “Mr Parkinson” vs “Parkinson's Disease”
  - “General Motors” vs. “General Smith”
- Handcrafted grammars are used to define patterns over the Lookups and other annotations
- These patterns can help disambiguate, and they can combine different annotations, e.g. Dates can be comprised of day + number + month

## Named Entity Grammars

- Hand-coded rules written in JAPE applied to annotations to identify NEs
- Phases run sequentially and constitute a cascade of FSTs over annotations
- Annotations from format analysis, tokeniser, splitter, POS tagger, morphological analysis, gazetteer etc.
- Because phases are sequential, annotations can be built up over a period of phases, as new information is gleaned
- Standard named entities: persons, locations, organisations, dates, addresses, money
- Basic NE grammars can be adapted for new applications, domains and languages



## Using co-reference

- Different expressions may refer to the same entity
- Orthographic co-reference module (orthomatcher) matches proper names and their variants in a document
- [Mr Smith] and [John Smith] will be matched as the same person
- [International Business Machines Ltd.] will match [IBM]

## Orthomatcher PR

- Performs co-reference resolution based on orthographical information of entities
- Produces a list of annotation ids that form a co-reference chain
- List of such lists stored as a document feature named “MatchesAnnots”
- Improves results by assigning entity type to previously unclassified names, based on relations with classified entities
- May not reclassify already classified entities
- Classification of unknown entities very useful for surnames which match a full name, or abbreviations, e.g. “Bonfield” <Unknown> will match “Sir Peter Bonfield” <Person>
- A pronominal PR is also available

# Coreference editor

The screenshot displays the 'Core-reference Editor' window. The main text area contains four paragraphs with several words highlighted in colored boxes: 'National Air Traffic Services' (red), 'London' (blue), 'Nats' (red), 'UK' (yellow), and 'March' (pink). To the right, the 'Co-reference Data' panel shows a list of entities with checkboxes, all of which are checked. The entities are: 'National Air Traffic Services' (red), 'Airline Group' (green), 'UK' (yellow), 'London' (blue), and 'March' (pink). Red lines connect the highlighted words in the text to their corresponding entries in the list.

Annotation Sets   Annotations List   Annotations Stack   Class   Co-reference Editor   Instance   T

Sets : Default

Types : Organization   Show

Co-reference Data

Default

- National Air Traffic Services
- Airline Group
- UK
- London
- March

Document Editor   Initialisation Parameters

## Identity Resolution issues

### Same Person Name different Entity

- P1) **Antony John** was born in 1960 in Gilfach Goch, a mining town in the Rhondda Valley in Wales. He moved to Canada in 1970 where the woodlands and seasons of Southwestern Ontario provided a new experience for the young naturalist...
- P2) **Antony John** - Managing Director. After working for National Westminster Bank for six years, in 1986, Antony established a private financial service practice. For 10 years he worked as a Director of Hill Samuel Asset Management and between 1999 and 2003 he was an Executive Director at the private Swiss bank, Lombard Odier Darier Hentsch. Antony joined IMS in 2003 as a Partner. Antony's PA is Heidi Beasley...

# Identity Resolution

Same company name, different company

- C1) Operating in the market where knowledge processes meet software development, **Metaware** can support organizations in their attempts to become more competitive. Metaware combines its knowledge of company processes and information technology in its services and software. By using intranet and workflow applications, Metaware offers solutions for quality control, document management, knowledge management, complaints management, and continuous improvement.
- C2) **Metaware S.r.l.** is a small but highly technical software house specialized in engineering software and systems solutions based on internet and distributed systems technology. Metaware has participated in a number of RTD cooperative projects and has a consolidated partnership relationship with Engineering.

# Approaches to Identity Resolution

## Text based approach

- clustering informed by semantic analysis and summarization
- extract sentences containing entity of interest and create a summary
- extract semantic information from summaries and create term vectors for clustering
- apply clustering to the set of vectors
- performance around 80%

# Approaches to Identity Resolution

## Ontology-based approach

- define rules for each class in the ontology
- rules combine different similarity criteria using a weighting mechanism:
- compare alias name (“Alcoa” vs “Alcoa Inc.”)
- compare location (Scotland is in the UK)
- select candidate instances from ontology
- compare target instance to each candidate
- evaluation of merging information extracted from company profiles:
- performance ~ 89%

## Creating your own NE applications

- Tools such as GATE allow you to mix and match different components in the pipeline as you want
- For example, you can change the POS tagger to process text in a different language, or add a syntactic parser to the application
- GATE, for example, allows conditional processing so that you can set up an application to use different resources automatically for a multilingual corpus, depending on which language the document is in



## Other NLP Toolkits

- UIMA
- OpenCalais
- Lingpipe

All integrated into  
GATE as plugins



"I used to feel the same way you do about no-good bums, sir, but I finally decided, if you can't beat 'em, join 'em!"

## UIMA

- UIMA is an NL engineering platform developed by IBM
- Shares some functionality with GATE, but is complementary in most respects.
- Interoperability layer has been developed to allow UIMA applications to be run within GATE, and vice versa, in order to combine elements of both.
- Emphasis is on architectural support, including asynchronous scaleout (deploying many copies of an application in parallel)
- Much narrower range of resources provided than GATE

<http://incubator.apache.org/uima/>

## OpenCalais

- Web service for semantic annotation of text.
- The user submits a document to the web service, which returns entity and relations annotations in RDF, JSON or some other format.
- Typically, users integrate OpenCalais annotation of their web pages to provide additional links and 'semantic functionality'.
- OpenCalais annotates both relations and entities, although the GATE plugin only supports entities.

<http://www.opencalais.com>

## LingPipe

- Provides set of IE and data mining tools largely ML-based. Has a set of models trained for particular tasks/corpora.
- Limited ontology support: can connect entities found to databases and ontologies
- Advantage: ML models can suggest more than one output, ranked by confidence. The user can choose number of suggestions generated.
- Disadvantage: ML models only apply to specific tasks and domains.

<http://alias-i.com/lingpipe/index.html>

# Commercial users of text mining products

## Typical commercial **uses**:

- dynamic search and indexing of repositories
- finding relations between elements in distributed repositories
- aggregating information from different text sources
- populating repositories
- fact finding from distributed knowledge sources

## Typical **users**:

- Pharmaceuticals, news, intelligence (business, competitor, government, etc.), manufacturing, telecommunications

- Home
- About us
- Medicines
- Research
- Careers
- Responsibility
- Partnering
- Media
- Investors



## LIFE INSPIRING IDEAS

[READ MORE](#) 

Our business is focused on turning good ideas into innovative, effective medicines that make a real difference in important areas of healthcare.



### Our medicines



- ▶ Arimidex
- ▶ Crestor
- ▶ Nexium

### Our research

In the fight against human disease, we focus on six disease areas where we believe our skills and experience can make the most difference:

- ▶ Cancer
- ▶ Cardiovascular
- ▶ Gastrointestinal
- ▶ Infection
- ▶ Neuroscience

### Our responsibility



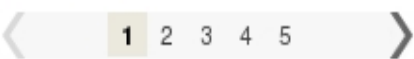
- ▶ Animal research
- ▶ Clinical trials
- ▶ Safety of medicines

### Latest news

01 June 2009

▶ AstraZeneca and Merck & Co., Inc. Form Pioneering Collaboration to Investigate Novel Combination Anticancer Regimen

▶ more news





## Roche Product Portfolio

# Focus on unsolved medical problems

*Our aim is to develop new and improved drugs, and diagnostic tests and services that offer significant benefits over existing options.*



### Explore our Portfolio

Products A-Z ▶

Diseases ▶

Products for Researchers ▶

Solutions for Diagnostics ▶

### Pharmaceuticals



Roche has brought many highly effective drugs onto the market and is a world leader in innovative cancer drugs. Other areas include viral infections, metabolic, central nervous system disorders and inflammatory diseases. [> More](#)

### Solutions for Diagnostics



As the world leader in in-vitro diagnostics, we supply a wide range of rapid, reliable instruments and tests for disease screening and diagnosis in laboratories, at the point of care, and for patient self-management. [> More](#)

### Products for Researchers



Roche Applied Science supplies a broad array of instruments and highly specific reagents and test kits for use in the diverse research market. The portfolio is especially strong in genomics and proteomics. [> More](#)





Use consumer friendly channels

Industries

Solutions

- + Telecommunications & Media
- + Retail & Consumer Goods
- + Financial Services & Insurance
- + Transportation & Travel
- + Leisure & Entertainment

Fizzback is an **on-demand solution** that drives **customer engagement** at the **point-of-experience**

[» more](#)

Spotlight

Resources

Latest News

[» more](#)



**T-Mobile selects Fizzback**

T-Mobile UK has partnered with Fizzback to drive satisfaction across all touch-points...[read more](#)



**Case Studies**



**Video Testimonials**

**T-Mobile to present at ECEW**

T-Mobile will present how Fizzback has allowed them to drive the customer...[more](#)



# GENERIC



Home

Niche Sectors



Careers

About Us

Contact Us



## Business Intelligence

Business Intelligence offers a filter to identify the most useful data to the management team within a business...

As the universe continuously expands, so the demand for specialist skills increases: without the right experience and talent, no business...

Home

About Us

Product

Careers

News

## Revolutionising recruitment practices for the HR & Staffing sector

Find. Analyse. Connect.  
With Insight 3.0

Enjoy sophisticated access to information and market intelligence on companies who advertise online

Insight 3.0 gives you a detailed understanding of the online job advertising activity of hundreds of thousands of UK employers. It enables you to monitor and assess the recruitment needs and business activity of clients, prospects and competitors. It also allows you to audit your own recruitment and identify the most appropriate media for your ads



simple ingenuity

recruitment news

register

.....  
D.O.B.     
[forgotten your details?](#) [Log in](#)

Garlik's DataPatrol helps people take control of their personal information and protect themselves against identity theft and financial fraud...



[DataPatrol for businesses](#)

[DataPatrol for individuals](#)

**Latest news** [all news](#)

 [DataPatrol service review](#)  
5/15/09

 [Garlik Secures Further Funding in Battle Against Cybercrime](#)  
4/23/09

**Welcome to Garlik**

Garlik are leading technology innovators and identity experts set up by the founders of the online banks Egg and First Direct. With our range of products and services we aim to give individuals power over the use of their personal

**Financial fraud soars**



Economic crisis fuels new cybercrime wave. Check out our UK cybercrime report.



## Homepage



Tailored text, pictures and video for use on any platform

## Contact Us >

**PRESS ASSOCIATION Sport**

New name, same unrivalled coverage

### Wire Service

All the day's breaking news stories in words and pictures.

### Digital

Content to power websites, widgets, mobile services, digital display screens and much more.

### Images

**Log in** to see the latest news, sport and showbiz pictures, plus an archive of over 15 million images.

### Video

News, sport and entertainment coverage from around the UK available as footage, clips and packages.

### Specialist News

Tailored news feeds for corporate websites, in-depth news from Westminster and detailed financial coverage.

### Pages

Newspaper and magazine production services from individual pages and supplements to entire cover-to-cover solutions. Bespoke services available.

### PR Services

Media monitoring and training to press release distribution and broadcast consultancy - PR services from a journalistic point of view.

### Business Information

Essential news and information services for public and private sector organisations, Government departments, financiers and PR companies.

## Enhancing the Independent's website with video

The Independent has launched an enhanced UK video news offering following an innovative distribution deal with Press Association and Octopus Media Technology. Read the full [video deal report](#).

## 2012 Olympic updates



Providing fail-safe information and publishing solutions for more than 200 years

Part of the Williams Lea Group

[Home](#)

[About TSO](#)

[Solutions](#)

[Testimonials](#)

[Insights](#)

[Press Office](#)

[My Account](#)

[A-Z Index](#)

[Contact Us](#)

## Information and Publishing Solutions

TSO (The Stationery Office) provides expert management of all the printed and digital information organisations share with their internal and external audiences. Applying proven information management methods and experience, we deliver transformational solutions and rapid results.

We have been providing fail-safe information management and publishing solutions to private and public sector organisations for more than 200 years.

Find out how our [information and publishing solutions](#) can reduce costs, increase efficiency and improve interaction with citizens and stakeholders.

[Contact us for more information](#)

### Bookshop

Order and purchase any UK book in print from TSO's online bookshop

[Enter \[tsoshop.co.uk\]\(http://tsoshop.co.uk\)](#)

### Latest News

The London Gazette used to create Dick Turpin e-fit . . . more



## More about text mining applications

- In Module 4, we'll spend more time looking at some real-life text mining applications
- In the next module, we'll look at how to add semantics to the applications