# Text Mining: Expanding into Drug Regulatory Submissions and Surveillance

J. Dixon[$], D. Milward[£], A. Winter[£] & P. Milligan[£]

## Introduction

Text mining is the process of deriving high-quality knowledge from unstructured text. Text analytics, particularly natural language processing, identifies relevant facts and relationships, and extracts them in a structured form for review and faster analysis; and connects these facts together in new ways, to synthesize knowledge, and create actionable insights.

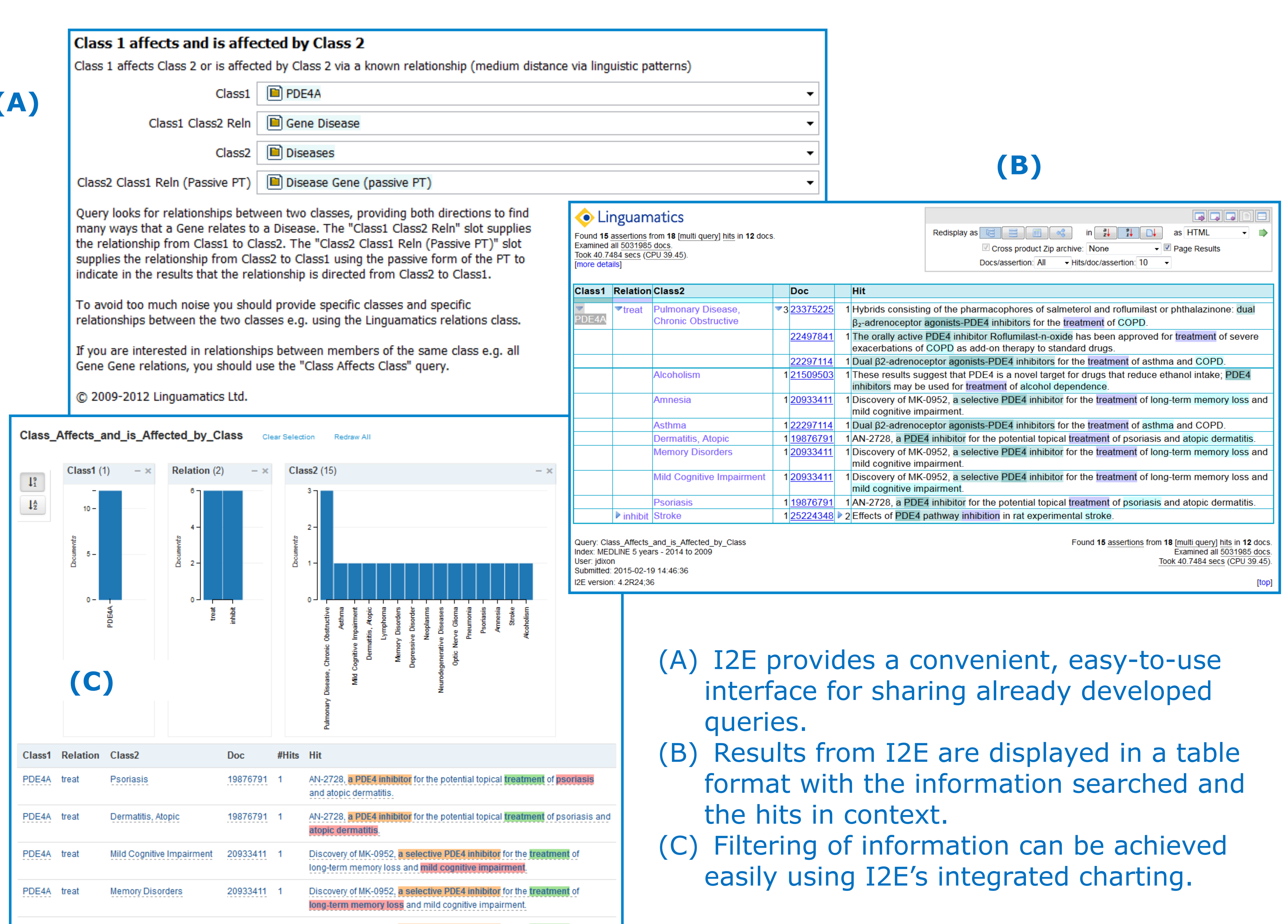
For years, pharmaceutical companies have been using text mining tools to extract key information from scientific literature, patents, and other data sources where valuable information is buried in unstructured text. Text analytics has been used for a wide range of applications across the drug discovery & development lifecycle, such as target prioritization, biomarker discovery, clinical trial analytics, mechanisms of action (MOA), or drug repurposing.

Recently, more and more pharmaceutical companies are using text mining later in the drug development pipeline and beyond. For example, text analytics can assist in quality control of regulatory document submission – used to cross-check MedDRA coding, references to tables, decimal place errors, and discrepancies between the summary document and source documents. This requires the use of advanced processing to extract information from tables in PDF documents as well as natural language processing to analyze the free text.

In this poster, we would like to highlight three use cases within the drug development pipeline and post-marketing activities where text mining with Linguamatics I2E* has enhanced and/or sped up the auditing, approval and market surveillance processes.

## Early to Late Stage Drug Discovery

Using text mining we can discover potential new drug repurposing targets or MOAs. Utilizing queries that come standard with I2E we are easily able to use publically available ontologies to identify new targets and then display them in a cluster table format or in a bar chart.




(A) I2E provides a convenient, easy-to-use interface for sharing already developed queries.
(B) Results from I2E are displayed in a table format with the information searched and the hits in context.
(C) Filtering of information can be achieved easily using I2E's integrated charting.

### Tables

**Table 1: Most Frequently Reported Medical Conditions (≥5% in Any Treatment Group)**

| Study | 2000 Pooled Studies | | 2003 Pooled Study | | 2006 Pooled Studies | |
|---|---|---|---|---|---|---|
| Total Number Subjects | Rx N=997 | Pbo N=927 | Rx N=1021 | Pbo N=956 | Rx N=1040 | Pbo N=930 |
| | Number (%) of Subjects | | | | | |
| Cardiac disorders | 70 (7.0) | 32 (3..5) | 108 (10.6) | 101 (10.6) | 214 (20.6) | 178 (19.1) |
| Angina pectoris | 4 (0.4) | 5 (0.5) | 74 (7.2) | 71 (7.4) | 79 (7.6) | 79 (8.5) |
| Gastrointestinal disorders | 391 (39.2) | 261 (28.2) | 60 (5.9) | 58 (6.1) | 689 (66.3) | 472 (50.8) |
| Dyspepsia | 174 (17.5) | 120 (12.9) | 3 (0.3) | 2 (0.2) | 219 (22.1) | 155 (16.7) |
| GERD | 83 (8.3) | 52 (5.6) | 30 (2.9) | 27 (2.8%) | 173 (16.6) | 115 (12.4) |
| Immune system disorders | 503 (50.5) | 301 (32.5) | 37 (3.6) | 23 (2.4) | 759 (73.0) | 449 (48.3) |
| Drug hypersensitivity | 103 (10.3) | 78 (8.4) | 18 (1.8) | 9 (0.9) | 148 (14.2) | 109 (11.7) |
| Seasonal allergies | 295 (29.6) | 197 (21.3) | 16 (1.6) | 4 (0.4) | 455 (43.8) | 272 (29.2) |
| Metabolic/ nutritional disorders | 253 (25.4) | 165 (17.8) | 194 (19.0) | 212 (22.2) | 757 (72.8) | 581 (62.5) |
| Dyslipidemia | 1 (0.1) | 0 (0) | 15 (1.5) | 19 (2.0) | 43 (4.1) | 23 (2.5) |
| Hypercholesterolaemia | 65 (6.5) | 50 (5.4) | 88 (8.6) | 103 (10.8) | 234 (22.5) | 218 (23.4) |
| Hyperlipidaemia | 147 (14.7) | 79 (8.5) | 56 (5.5) | 66 (6.9) | 289 (27.8) | 183 (19.7) |
| Type 2 diabetes mellitus | 16 (1.6) | 10 (1.1) | 38 (3.7) | 52 (5.4) | 82 (7.9) | 85 (9.1) |
| Musculoskeletal/ connective tissue disorders | 616 (61.8) | 341 (36.8) | 55 (5.4) | 46 (4.8) | 877 (84.3) | 512 (55.1) |
| Back pain | 166 (16.6) | 96 (10.4) | 12 (1.2) | 10 (1.0) | 239 (23) | 135 (14.5) |
| Osteoarthritis | 102 (10.2) | 57 (6.6) | 12 (1.2) | 11 (1.2) | 155 (14.9) | 85 (9.1) |
| Nervous system disorders | 628 (63.0) | 409 (44.1) | 28 (2.7) | 19 (2.0) | 845 (81.3) | 531 (57.1) |
| Headache | 413 (41.4) | 280 (30.2) | 9 (0.9) | 7 (0.7) | 499 (48.0) | 326 (35.1) |
| Psychiatric disorders | 137 (13.7) | 81 (8.7) | 14 (1.4) | 15 (1.6) | 243 (23.4) | 156 (16.8) |
| Insomnia | 84 (8.4) | 47 (5.1) | 9 (0.9) | 8 (0.8) | 146 (14.0) | 94 (10.1) |

### Documents

Commonly reported conditions included Seasonal allergies, Back pain, and Hypercholesterolaemia. The majority of AEs were considered treatment related in all cohorts and the relationship between treatment groups and between cohorts was similar to that observed for all-causality AEs. Permanent discontinuations were reported at higher rates in the Rx groups than in the placebo groups in the 3 pooled cohorts. The majority of AEs leading to permanent discontinuation were considered treatment related in both treatment groups in all cohorts. The single most frequently reported event was headache, which was reported in approximately 40% of Rx subjects and 20% of placebo subjects in the 2000 Pooled cohort. Other AEs reported across all cohorts at rates greater in Rx subjects than placebo subjects included Seasonal allergies and Insomnia (2000 8.4% vs 5.4%, 2003 0.9% vs 0.8%, 2006 14.0% vs 10.1%; Rx vs placebo respectively).

## Submission QC in Late Stage Drug Development

Using advanced text mining processing, we are able to identify inconsistencies within FDA submission documents, across tables and textual parts of the reports.

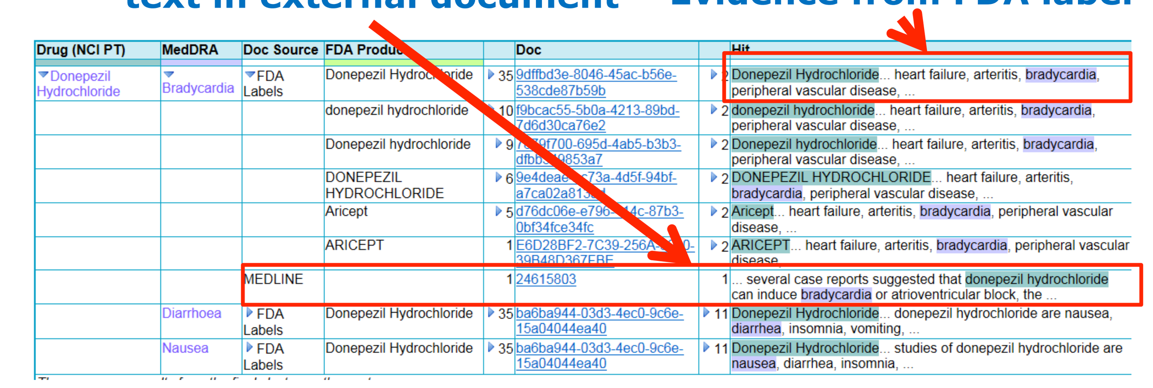
**Key**

Incorrect formatting: doubled period, incorrect number of decimal places, addition of percent sign
Incorrect calculation: number of patients divided by total number does not agree with percent term
Incorrect threshold: presence of row does not agree with table title
Text-Table inconsistency: numbers in the table do not agree with numbers in the accompanying text

## Post-Marketing Surveillance

By comparing literature or literature abstracts with FDA labels, I2E is able to help identify adverse events (AEs) that are not identified in FDA labels or help justify why a given AE is identified in an FDA label.

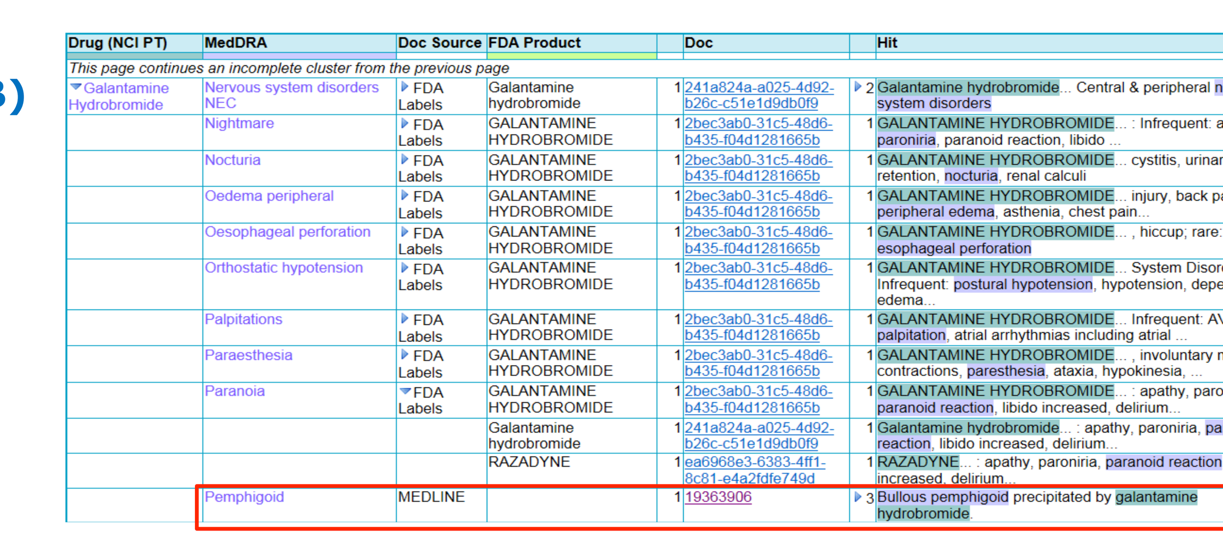
**(A)** Evidence from unstructured text in external document — Evidence from FDA label



| Drug (NCI PT) | MedDRA | Doc Source | FDA Product | Doc | Hit |
|---|---|---|---|---|---|
| Donepezil Hydrochloride | Bradycardia | FDA Labels | Donepezil Hydrochloride | 35 9dffbd3e-8046-45ac-b56e-538cde87b59b | 2 Donepezil Hydrochloride... heart failure, arteritis, bradycardia, peripheral vascular disease, ... |
| | | | donepezil hydrochloride | 10 f9bcac55-5b0a-4213-89bd-7d6d30ca76e2 | 2 donepezil hydrochloride... heart failure, arteritis, bradycardia, peripheral vascular disease, ... |
| | | | Donepezil hydrochloride | 9 [illegible]f700-695d-4ab5-b3b3-dfb[illegible]0853a7 | 2 Donepezil hydrochloride... heart failure, arteritis, bradycardia, peripheral vascular disease, ... |
| | | | DONEPEZIL HYDROCHLORIDE | 6 9e4deae[illegible]73a-4d5f-94bf-a7ca02a81[illegible] | 2 DONEPEZIL HYDROCHLORIDE... heart failure, arteritis, bradycardia, peripheral vascular disease, ... |
| | | | Aricept | 5 d76dc06e-e796-[illegible]4c-87b3-0bf34fce34fc | 2 Aricept... heart failure, arteritis, bradycardia, peripheral vascular disease, ... |
| | | | ARICEPT | 1 E6D28BF2-7C39-256A-[illegible]0-39B48D367EBE | 2 ARICEPT... heart failure, arteritis, bradycardia, peripheral vascular disease |
| | | MEDLINE | | 1 24615803 | 1 ... several case reports suggested that donepezil hydrochloride can induce bradycardia or atrioventricular block, the ... |
| | Diarrhoea | FDA Labels | Donepezil Hydrochloride | 35 ba6ba944-03d3-4ec0-9c6e-15a04044ea40 | 11 Donepezil Hydrochloride... donepezil hydrochloride are nausea, diarrhea, insomnia, vomiting, ... |
| | Nausea | FDA Labels | Donepezil Hydrochloride | 35 ba6ba944-03d3-4ec0-9c6e-15a04044ea40 | 11 Donepezil Hydrochloride... studies of donepezil hydrochloride are nausea, diarrhea, insomnia, ... |

(A) MEDLINE provides justification for Brachychardia in FDA Label of Donepezil Hydrochloride.

**(B)**



| Drug (NCI PT) | MedDRA | Doc Source | FDA Product | Doc | Hit |
|---|---|---|---|---|---|
| This page continues an incomplete cluster from the previous page | | | | | |
| Galantamine Hydrobromide | Nervous system disorders NEC | FDA Labels | Galantamine hydrobromide | 1 241a824a-a025-4d92-b26c-c51e1d9db0f9 | 2 Galantamine hydrobromide... Central & peripheral nervous system disorders |
| | Nightmare | FDA Labels | GALANTAMINE HYDROBROMIDE | 1 2bec3ab0-31c5-48d6-b435-f04d1281665b | 1 GALANTAMINE HYDROBROMIDE... : Infrequent: apathy, paroniria, paranoid reaction, libido ... |
| | Nocturia | FDA Labels | GALANTAMINE HYDROBROMIDE | 1 2bec3ab0-31c5-48d6-b435-f04d1281665b | 1 GALANTAMINE HYDROBROMIDE... cystitis, urinary retention, nocturia, renal calculi |
| | Oedema peripheral | FDA Labels | GALANTAMINE HYDROBROMIDE | 1 2bec3ab0-31c5-48d6-b435-f04d1281665b | 1 GALANTAMINE HYDROBROMIDE... injury, back pain, peripheral edema, asthenia, chest pain... |
| | Oesophageal perforation | FDA Labels | GALANTAMINE HYDROBROMIDE | 1 2bec3ab0-31c5-48d6-b435-f04d1281665b | 1 GALANTAMINE HYDROBROMIDE... , hiccup; rare: esophageal perforation |
| | Orthostatic hypotension | FDA Labels | GALANTAMINE HYDROBROMIDE | 1 2bec3ab0-31c5-48d6-b435-f04d1281665b | 1 GALANTAMINE HYDROBROMIDE... System Disorders: Infrequent: postural hypotension, hypotension, dependent edema... |
| | Palpitations | FDA Labels | GALANTAMINE HYDROBROMIDE | 1 2bec3ab0-31c5-48d6-b435-f04d1281665b | 1 GALANTAMINE HYDROBROMIDE... Infrequent: AV block, palpitation, atrial arrhythmias including atrial ... |
| | Paraesthesia | FDA Labels | GALANTAMINE HYDROBROMIDE | 1 2bec3ab0-31c5-48d6-b435-f04d1281665b | 1 GALANTAMINE HYDROBROMIDE... , involuntary muscle contractions, paresthesia, ataxia, hypokinesia, ... |
| | Paranoia | FDA Labels | GALANTAMINE HYDROBROMIDE | 1 2bec3ab0-31c5-48d6-b435-f04d1281665b | 1 GALANTAMINE HYDROBROMIDE... : apathy, paroniria, paranoid reaction, libido increased, delirium... |
| | | | Galantamine hydrobromide | 1 241a824a-a025-4d92-b26c-c51e1d9db0f9 | 1 Galantamine hydrobromide... : apathy, paroniria, paranoid reaction, libido increased, delirium... |
| | | | RAZADYNE | 1 ea6968e3-6383-4ff1-8c81-e4a2fdfe749d | 1 RAZADYNE... : apathy, paroniria, paranoid reaction, libido increased, delirium... |
| | Pemphigoid | MEDLINE | | 1 19363906 | 3 Bullous pemphigoid precipitated by galantamine hydrobromide. |

(B) Discovery of an adverse event in an external document which is not listed in any of the FDA labels containing Galantamine Hydrobromide (Pemphigoid is only found in MEDLINE).

## Conclusions

These three use cases demonstrate the value of I2E's text analytics solution to very different challenges.

- Structuring information from the scientific literature for drug – gene – disease relationships can provide an excellent substrate for drug repurposing or MOA decisions.
- Using linguistics for quality assurance of submission documents can save countless hours or weeks of tedious manual checking, and potentially prevent a re-submission request.
- The ability to search and extract key facts from drug labels, and compare these facts with those from scientific literature or elsewhere, provides high quality decision support for medical affairs teams.

With the ever-expanding volume of free text, from scientific literature to patents to electronic medical records, the need for ways to handle these data increases daily, and advanced text analytics is one tool that can help with this challenge.

*Linguamatics' agile natural language processing (NLP) text mining software, I2E, provides rapid knowledge discovery from unstructured text. Using I2E, knowledge can be extracted from a wide range of unstructured and/or semi-structured content sources. This knowledge can then be used to answer high-value questions in real time.

[$]Linguamatics Ltd, 1900 West Park Dr, Suite 280, Westborough, MA 02466, USA
[£]Linguamatics Ltd, 324 Cambridge Science Park, Milton Rd, Cambridge, CB4 0WG, UK

Website: www.linguamatics.com
Email: info@linguamatics.com

Linguamatics