Unstructured Text in Big Data
The Elephant in the Room

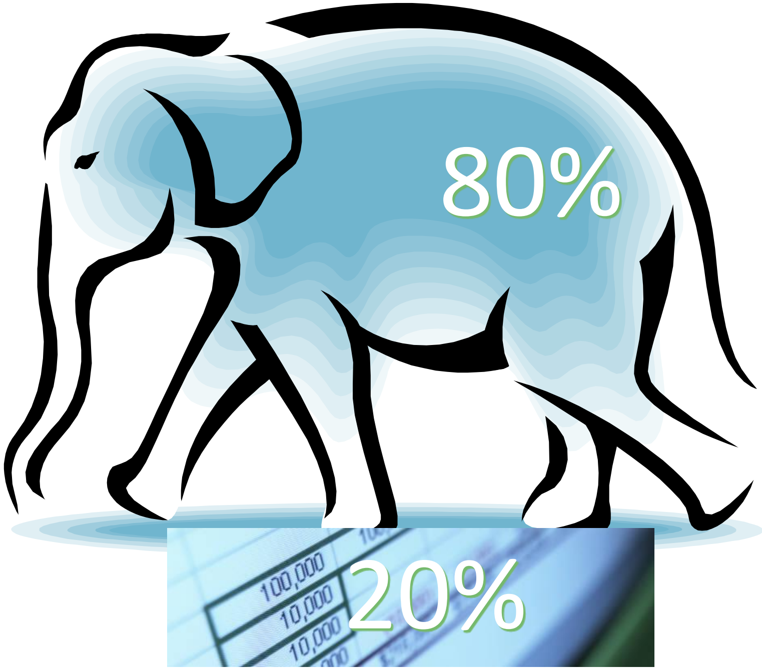
David Milward

ICIC, October 2013

Linguamatics

Unstructured Big Data

- Big Data
  - Volume, Variety, Velocity
- Estimated 80% of all data is unstructured
  - Need to be able to make decisions based on this data
- Ever increasing amount of data makes it harder and harder to filter by hand
  - Need more automation



From Unstructured Data to Structured

MEDLINE
U.S. National Library of Medicine
ClinicalTrials.gov
A service of the U.S. National Institutes of Health
European Patent Office
FDA
WIPO
PDF
http://
XML

Unstructured or Semi-Structured Content Sources

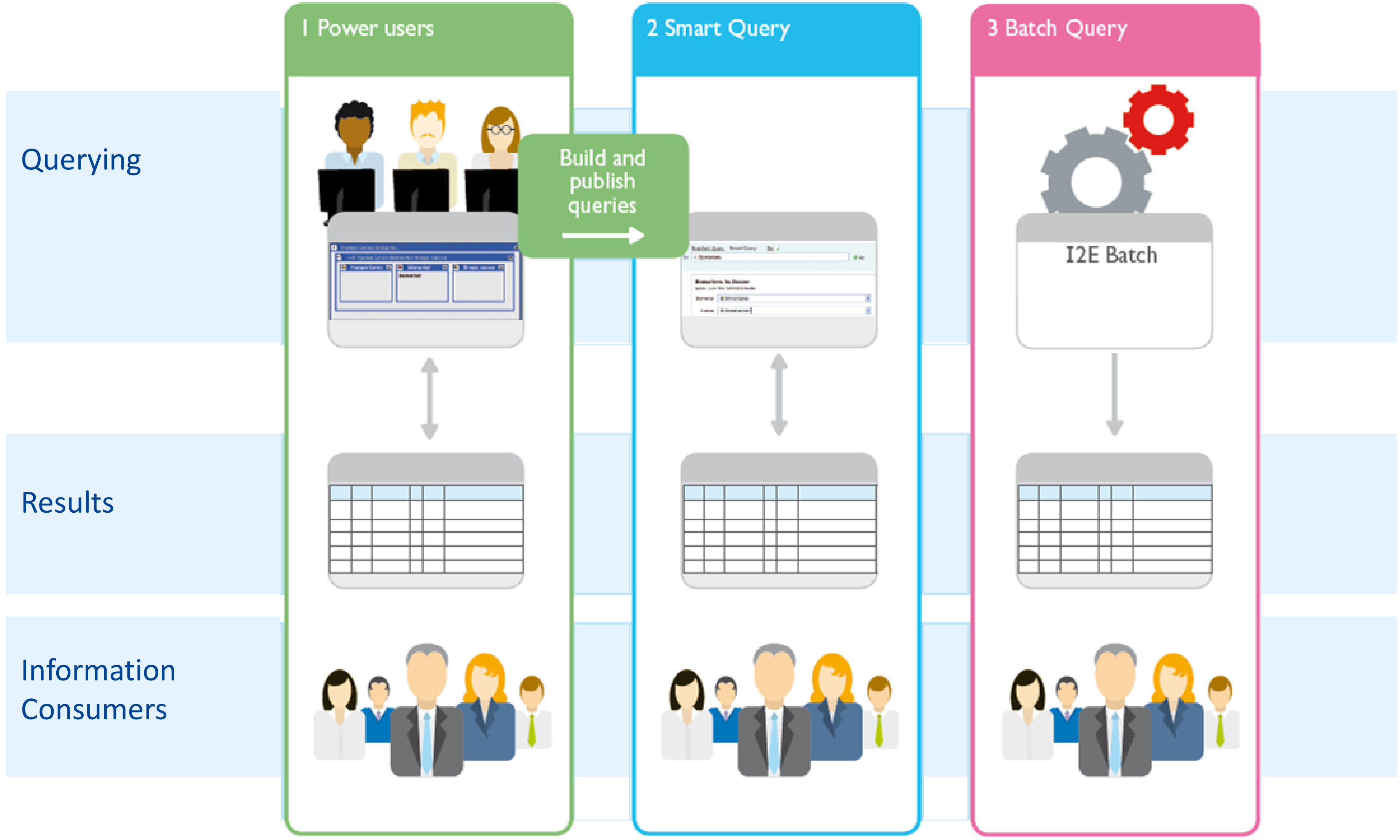
Identify concepts and relationships

Structured, relational data
Normalized concept identifiers

Traditional text mining works in batch mode

Requires each extraction to be programmed in, or learnt from a large amount of annotated material

I2E Agile Text Mining



Querying

Results

Information Consumers

1 Power users

Build and publish queries

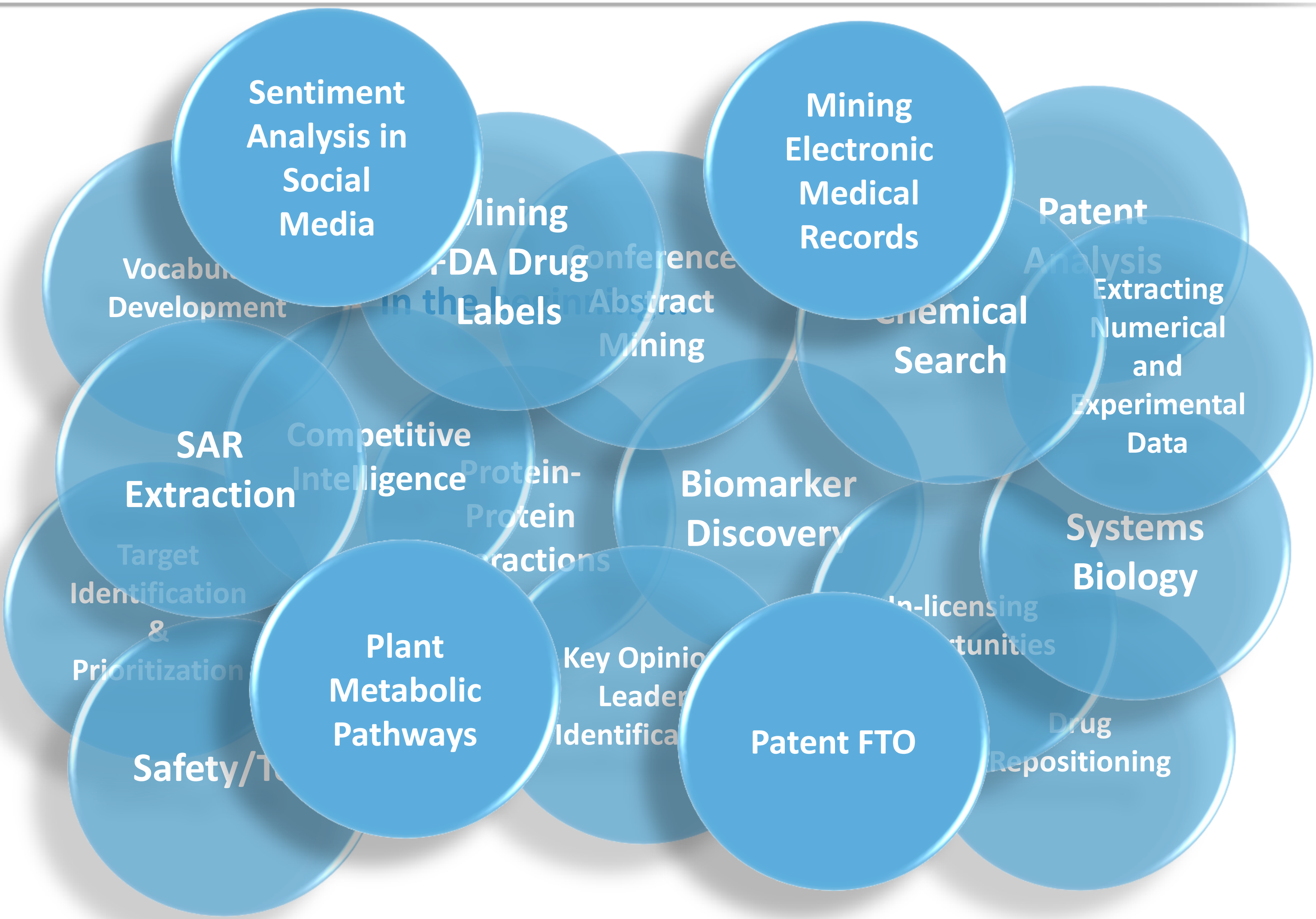
2 Smart Query

3 Batch Query

I2E Batch

Linguamatics

Increasing Range Of Applications



Sentiment Analysis in Social Media

Vocabu
Development

lining
FDA Drug
in the
Labels

Conference
Abstract
Mining

Mining Electronic Medical Records

Patent
Analysis

hemical
Search

Extracting
lumerical
and
xperimental
Data

SAR
Extraction

Competitive
Intelligence

Protein-
Protein
ractions

Biomarker
Discovery

Systems
Biology

Target
Identification
&
Prioritization

Safety/T

Plant
Metabolic
Pathways

Key Opinio
Leader
Identifica

Patent FTO

n-licensing
tunities

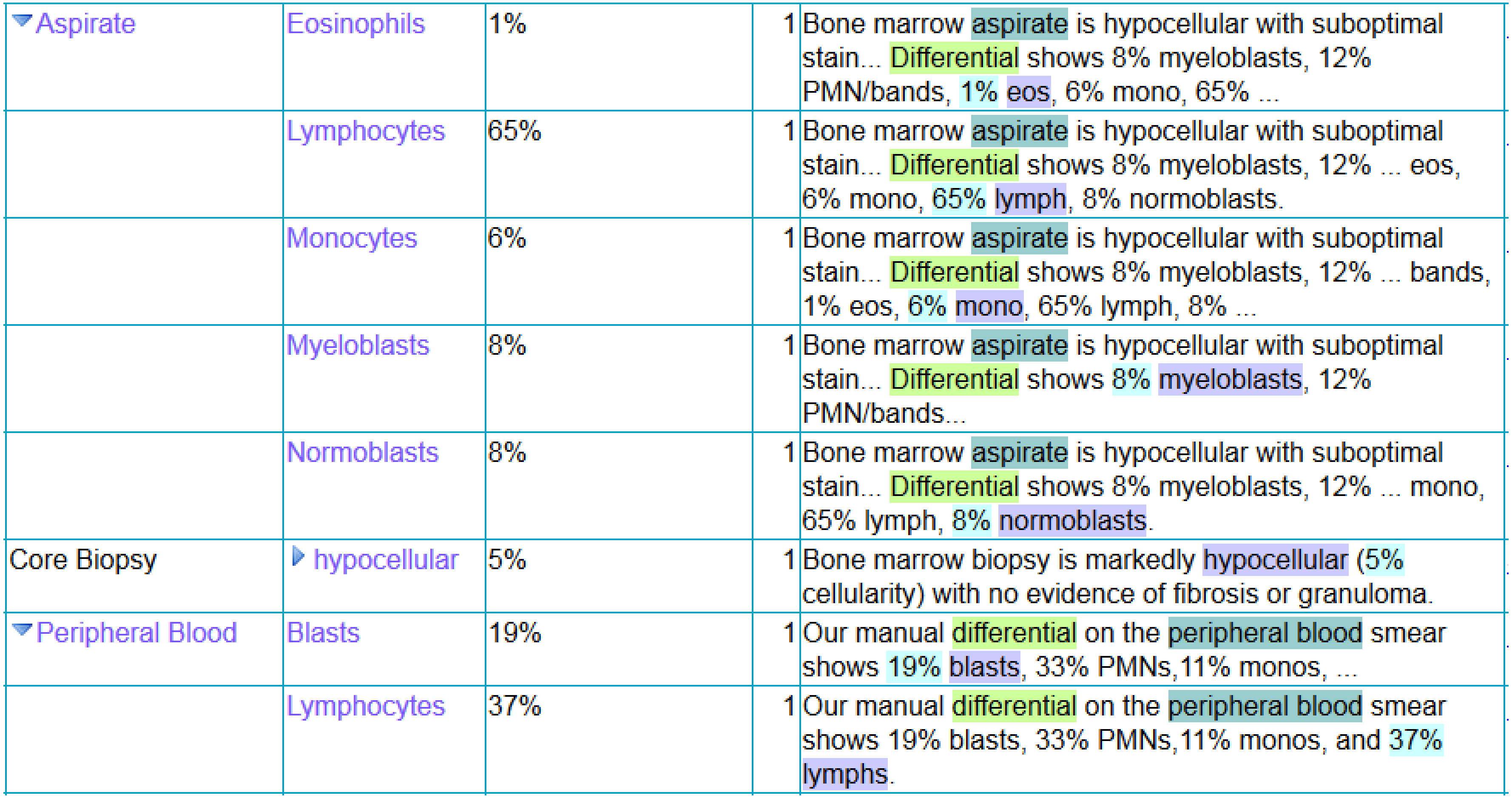
Drug
Repositioning

The Database Approach

- Can we predetermine what is required?
- One approach is to try to extract everything … but in practice have to make assumptions, whether by hand or automatically
  - Do we include negative statements?
  - Do we include speculation?
  - What context do we include?
- Need a good idea of how the data will be used to know what needs to be extracted
- There are cases where this works well e.g.
  - have useful structured data and want to add to it from free text
  - have new records structured, but want to bring in information from legacy free text

Extracting for a Relational Database

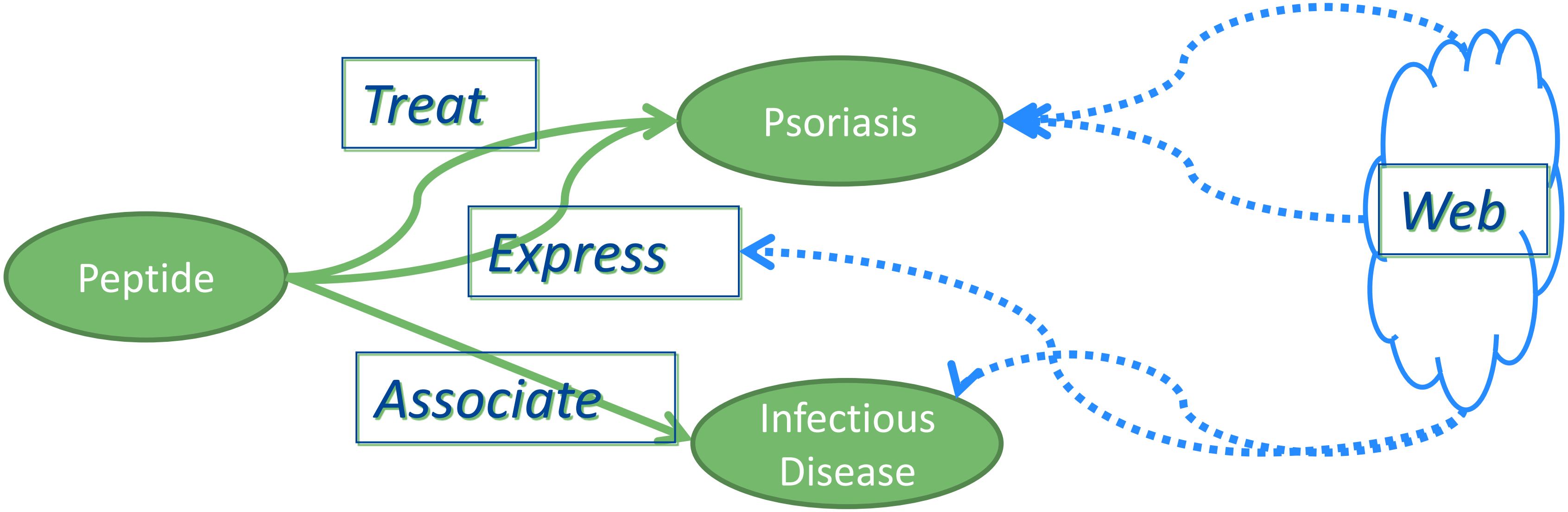
- Extracting all relevant information from pathology reports



▼ Aspirate	Eosinophils	1%	1	Bone marrow aspirate is hypocellular with suboptimal stain… Differential shows 8% myeloblasts, 12% PMN/bands, 1% eos, 6% mono, 65% …
	Lymphocytes	65%	1	Bone marrow aspirate is hypocellular with suboptimal stain… Differential shows 8% myeloblasts, 12% … eos, 6% mono, 65% lymph, 8% normoblasts.
	Monocytes	6%	1	Bone marrow aspirate is hypocellular with suboptimal stain… Differential shows 8% myeloblasts, 12% … bands, 1% eos, 6% mono, 65% lymph, 8% …
	Myeloblasts	8%	1	Bone marrow aspirate is hypocellular with suboptimal stain… Differential shows 8% myeloblasts, 12% PMN/bands…
	Normoblasts	8%	1	Bone marrow aspirate is hypocellular with suboptimal stain… Differential shows 8% myeloblasts, 12% … mono, 65% lymph, 8% normoblasts.
Core Biopsy	▶ hypocellular	5%	1	Bone marrow biopsy is markedly hypocellular (5% cellularity) with no evidence of fibrosis or granuloma.
▼ Peripheral Blood	Blasts	19%	1	Our manual differential on the peripheral blood smear shows 19% blasts, 33% PMNs,11% monos, …
	Lymphocytes	37%	1	Our manual differential on the peripheral blood smear shows 19% blasts, 33% PMNs,11% monos, and 37% lymphs.

Extracting for a Semantic Store

- Output as a set of triples
- Because all the parts have URIs, the parts are *web-addressable*



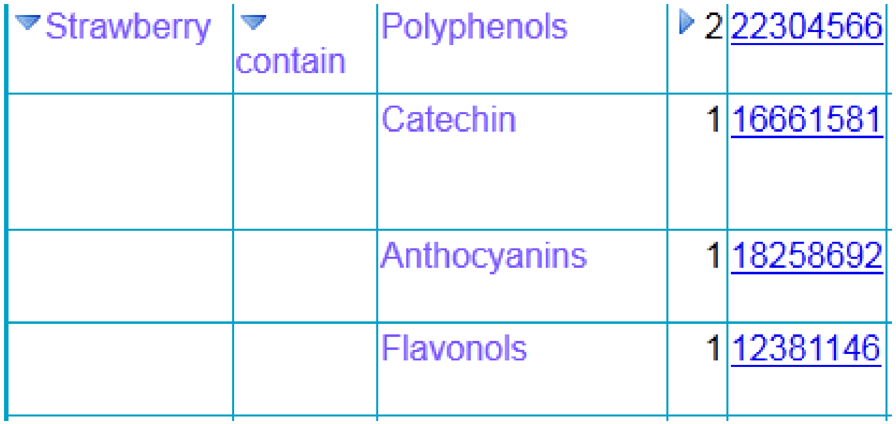
Linguamatics

The Ad Hoc Approach

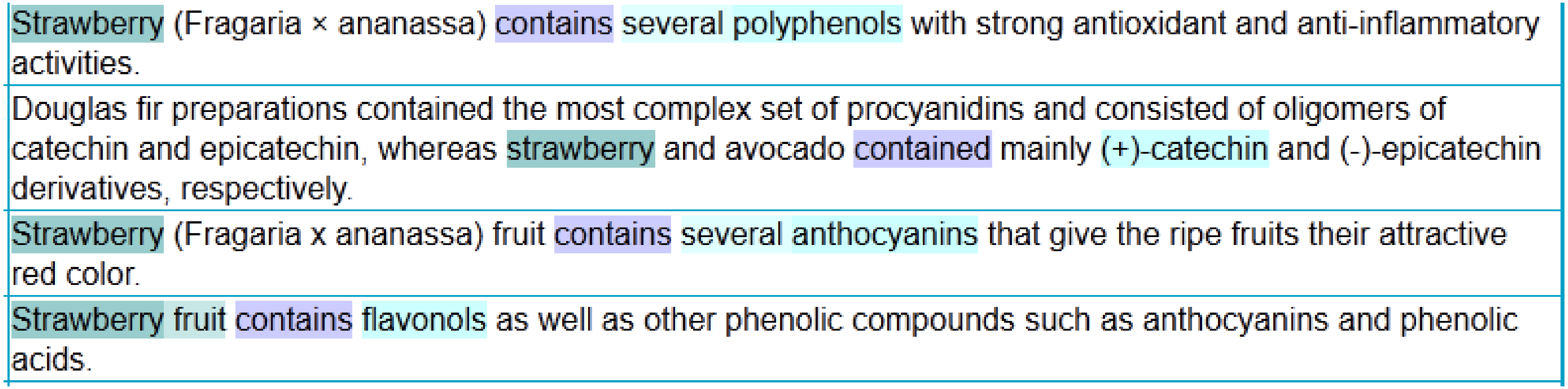
- No need to build a database
- Regard text mining queries as ways to create a database on the fly
- Keep all the unstructured free text, and query for what you need when you need it
- Even index structured data with text mining to link information together
- Used very successfully by knowledge professionals:
  - Just like a search engine, there are thousands of questions people want to ask of the data
  - You can't predict them all beforehand

Ad Hoc Querying

- Find relationships that you want for any particular information request
- Treat the text mining as an “agile” database to answer thousands of different questions e.g. metabolites from fruit



▼Strawberry	▼ contain	Polyphenols	▶ 2	22304566
		Catechin	1	16661581
		Anthocyanins	1	18258692
		Flavonols	1	12381146



Strawberry (Fragaria × ananassa) contains several polyphenols with strong antioxidant and anti-inflammatory activities.

Douglas fir preparations contained the most complex set of procyanidins and consisted of oligomers of catechin and epicatechin, whereas strawberry and avocado contained mainly (+)-catechin and (-)-epicatechin derivatives, respectively.

Strawberry (Fragaria x ananassa) fruit contains several anthocyanins that give the ripe fruits their attractive red color.

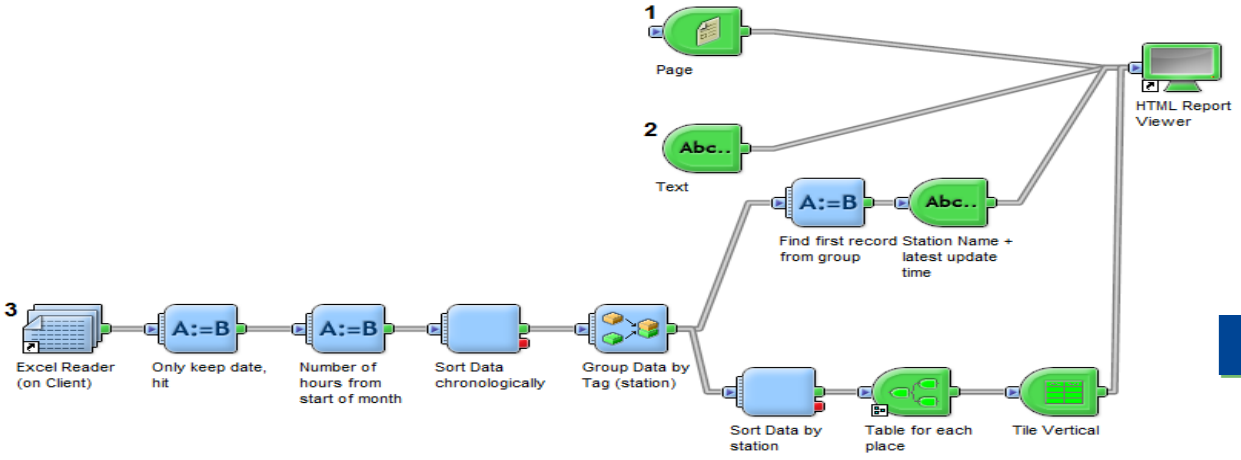
Strawberry fruit contains flavonols as well as other phenolic compounds such as anthocyanins and phenolic acids.

Do We Need More?

- I2E Agile Text Mining has opened up text mining to the end user, but it is still mainly used by knowledge professionals
- Given the importance of unstructured big data, can we bring the benefits of text mining to an even wider audience?

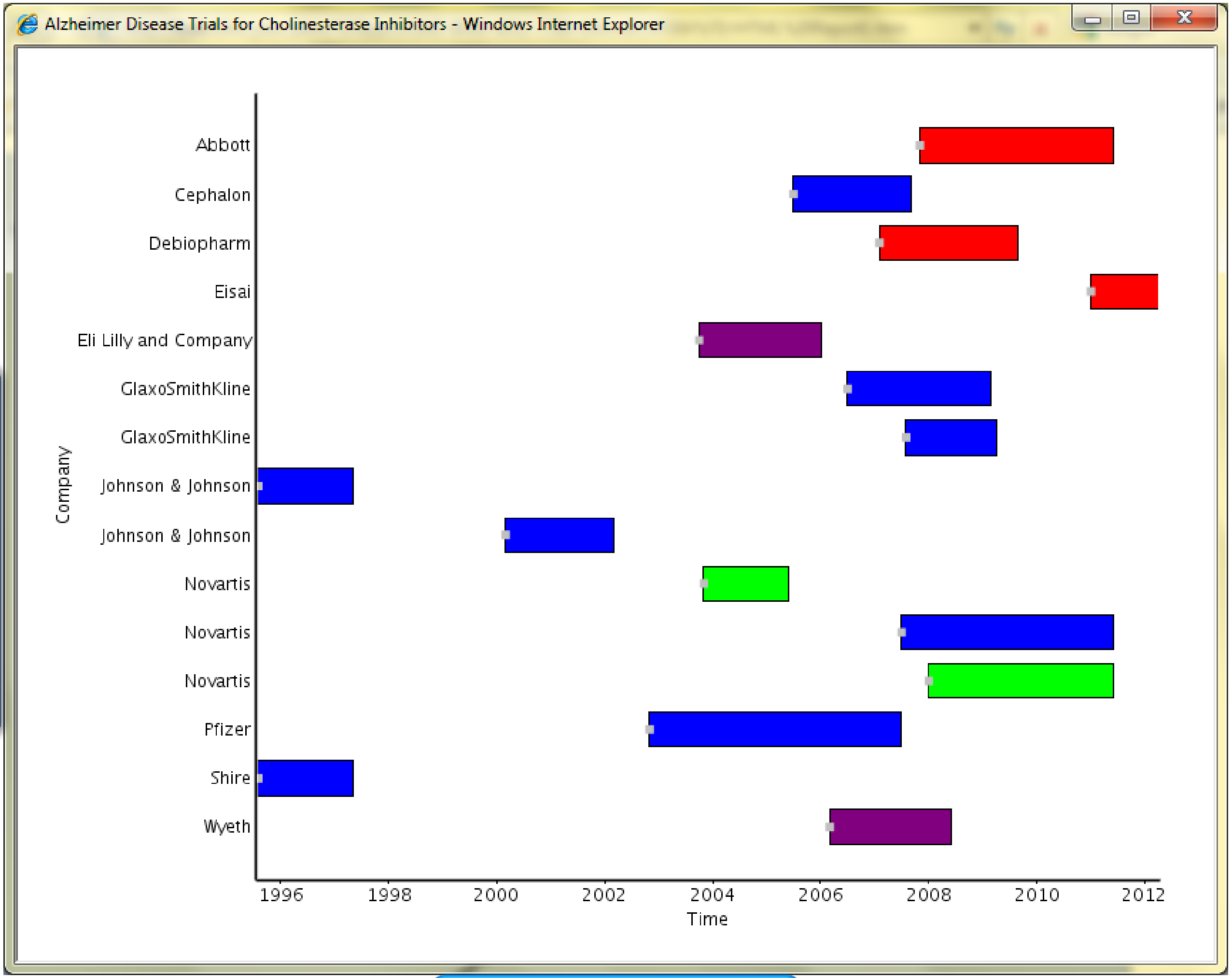
Workflow Automation

- Successful queries converted into regular workflows using Pipeline Pilot or KNIME
- Real-time workflows for up-to-date dashboards, alerts
- Further analysis, visualization, integration with other data


1
Page
2
Abc..
Text
HTML Report Viewer
A:=B
Find first record from group
Abc..
Station Name + latest update time
3
Excel Reader (on Client)
A:=B
Only keep date, hit
A:=B
Number of hours from start of month
Sort Data chronologically
Group Data by Tag (station)
Sort Data by station
Table for each place
Tile Vertical

Pipeline Pilot

Clinical Trials Analysis from I2E Text Mining Results



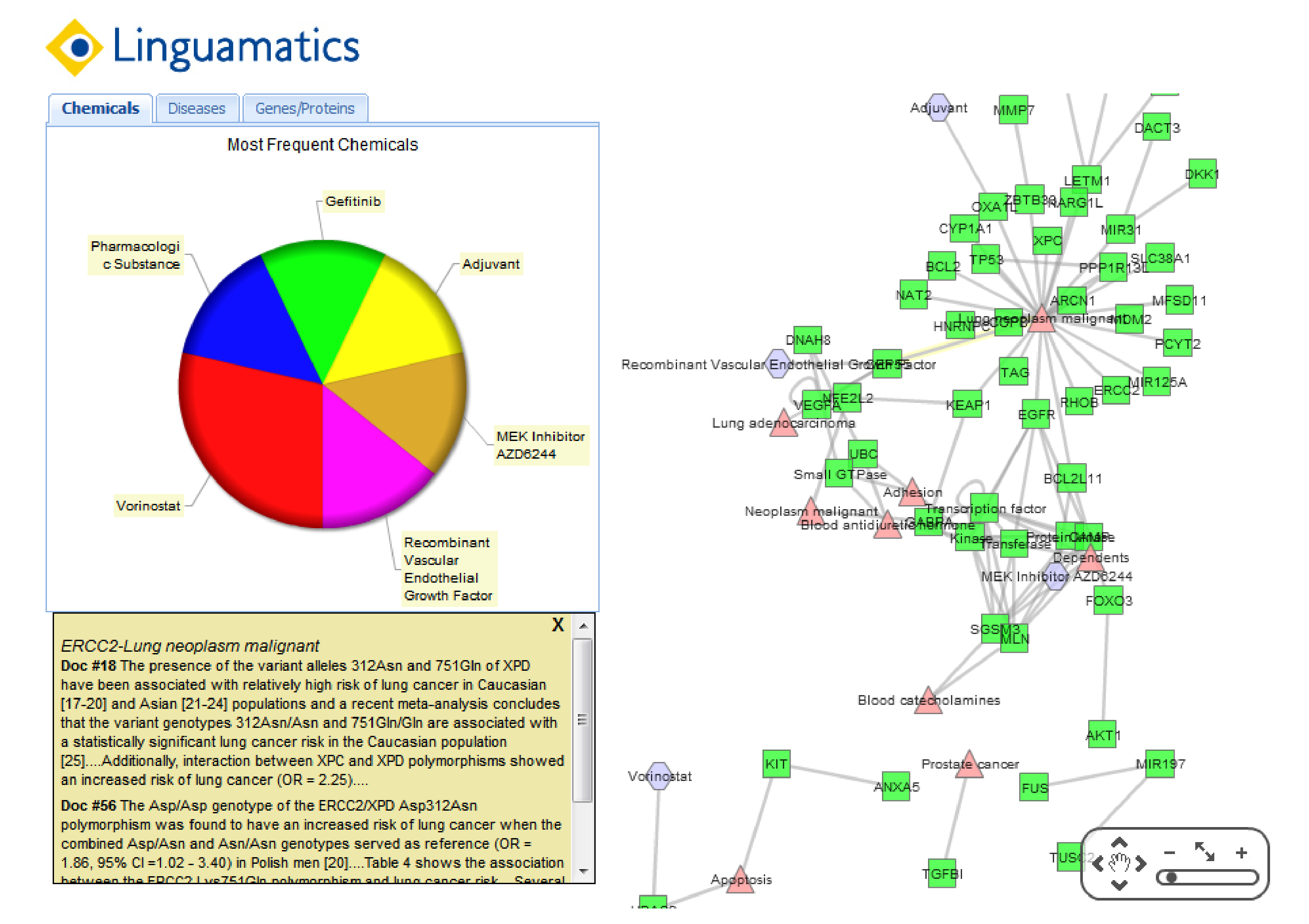
Sponsors

Dates

Colours indicate Phase

Using Pipeline Pilot visualization

Embedding I2E within Web Apps



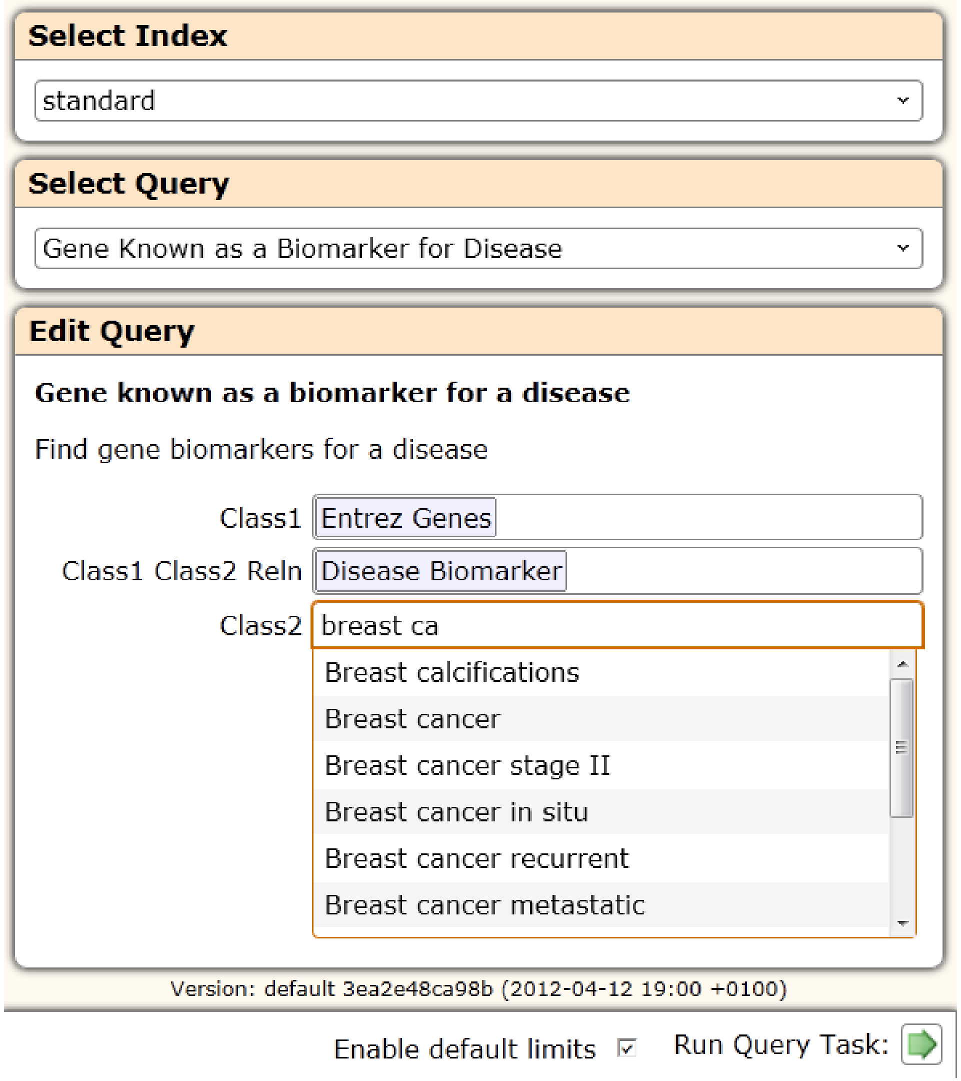
ERCC2-Lung neoplasm malignant

Doc #18 The presence of the variant alleles 312Asn and 751Gln of XPD have been associated with relatively high risk of lung cancer in Caucasian [17-20] and Asian [21-24] populations and a recent meta-analysis concludes that the variant genotypes 312Asn/Asn and 751Gln/Gln are associated with a statistically significant lung cancer risk in the Caucasian population [25]....Additionally, interaction between XPC and XPD polymorphisms showed an increased risk of lung cancer (OR = 2.25)....

Doc #56 The Asp/Asp genotype of the ERCC2/XPD Asp312Asn polymorphism was found to have an increased risk of lung cancer when the combined Asp/Asn and Asn/Asn genotypes served as reference (OR = 1.86, 95% CI =1.02 - 3.40) in Polish men [20]....Table 4 shows the association between the ERCC2 Lys751Gln polymorphism and lung cancer risk....Several

Form-Based Querying

- Create web interfaces connecting to I2E server
- Information professionals develop sophisticated queries
- Queries are parameterized to allow end users to customize
- Javascript allows portability to mobile devices such as iPads



Select Index

standard

Select Query

Gene Known as a Biomarker for Disease

Edit Query

Gene known as a biomarker for a disease

Find gene biomarkers for a disease

Class1	Entrez Genes
Class1 Class2 Reln	Disease Biomarker
Class2	breast ca

Breast calcifications
Breast cancer
Breast cancer stage II
Breast cancer in situ
Breast cancer recurrent
Breast cancer metastatic

Version: default 3ea2e48ca98b (2012-04-12 19:00 +0100)

Enable default limits ☑ Run Query Task:

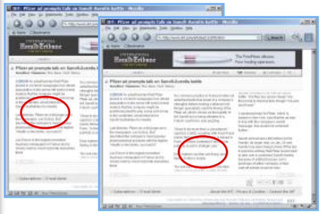
Linguamatics

Enhancing Enterprise Search

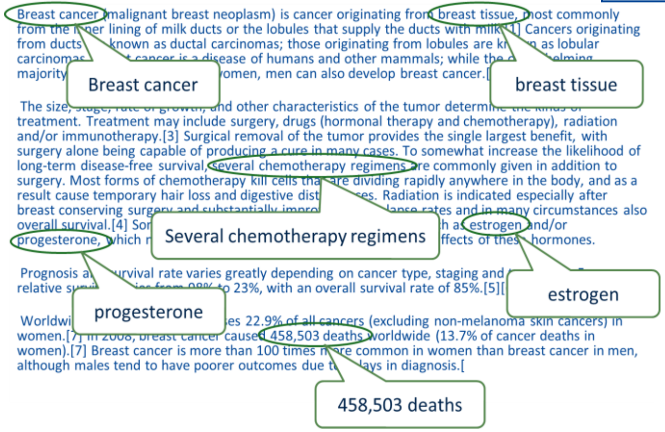
- Use text mining to annotate concepts to feed into a search engine to:
  - provide concept search
  - provide concepts for facets
  - provide intelligent thumbnails for documents, pulling out key information

Semantic annotation of documents

Value



Keywords



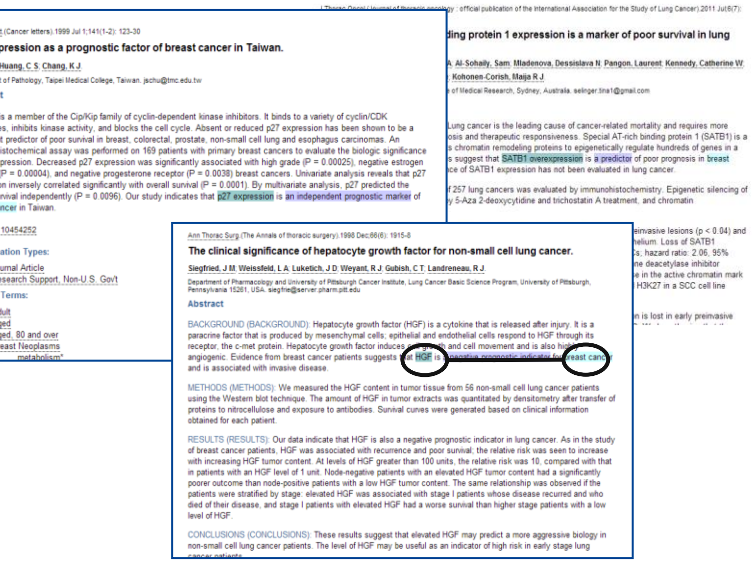
Breast cancer (malignant breast neoplasm) is cancer originating from breast tissue, most commonly from the inner lining of milk ducts or the lobules that supply the ducts with milk.[1] Cancers originating from ducts known as ductal carcinomas; those originating from lobules are kn n as lobular carcinomas t cancer is a disease of humans and other mammals; while the o elming majority omen, men can also develop breast cancer.[

The size, and other characteristics of the tumor determi treatment. Treatment may include surgery, drugs (hormonal therapy and chemotherapy), radiation and/or immunotherapy.[3] Surgical removal of the tumor provides the single largest benefit, with surgery alone being capable of producing a cure in many cases. To somewhat increase the likelihood of long-term disease-free survival, several chemotherapy regimens are commonly given in addition to surgery. Most forms of chemotherapy kill cells tha are dividing rapidly anywhere in the body, and as a result cause temporary hair loss and digestive dist ces. Radiation is indicated especially after breast conserving surg lapse rates and in many circumstances also overall survival.[4] Sor h as estrogen and/or progesterone, which r ffects of thes hormones.

Prognosis a urvival rate varies greatly depending on cancer type, staging and relative su to 23%, with an overall survival rate of 85%.[5][

Worldwi es 22.9% of all cancers (excluding non-melanoma skin cancers) in women.[7] caused 458,503 deaths worldwide (13.7% of cancer deaths in women).[7] Breast cancer is more than 100 times more common in women than breast cancer in men, although males tend to have poorer outcomes due t ays in diagnosis.[

Concepts



Cancer Lett.(Cancer letters).1999 Jul 1;141(1-2): 123-30

p27 expression as a prognostic factor of breast cancer in Taiwan.

Chu, J S; Huang, C S; Chang, K J.

Department of Pathology, Taipei Medical College, Taiwan. jschu@tmc.edu.tw

Abstract

p27Kip1 is a member of the Cip/Kip family of cyclin-dependent kinase inhibitors. It binds to a variety of cyclin/CDK complexes, inhibits kinase activity, and blocks the cell cycle. Absent or reduced p27 expression has been shown to be a significant predictor of poor survival in breast, colorectal, prostate, non-small cell lung and esophagus carcinomas. An immunohistochemical assay was performed on 169 patients with primary breast cancers to evaluate the biologic significance of p27 expression. Decreased p27 expression was significantly associated with high grade (P = 0.00025), negative estrogen receptor (P = 0.00004), and negative progesterone receptor (P = 0.0038) breast cancers. Univariate analysis reveals that p27 expression inversely correlated significantly with overall survival (P = 0.0001). By multivariate analysis, p27 predicted the overall survival independently (P = 0.0096). Our study indicates that p27 expression is an independent prognostic marker of breast cancer in Taiwan.

PMID: 10454252

Publication Types:

Journal Article
Research Support, Non-U.S. Gov't

MeSH Terms:

Adult
Aged
Aged, 80 and over
Breast Neoplasms
metabolism*

J Thorac Oncol (Journal of thoracic oncology : official publication of the International Association for the Study of Lung Cancer).2011 Jul;6(7):

ding protein 1 expression is a marker of poor survival in lung

A; Al-Sohaily, Sam; Mladenova, Dessislava N; Pangon, Laurent; Kennedy, Catherine W;
; Kohonen-Corish, Maija R J

e of Medical Research, Sydney, Australia. selinger.tina1@gmail.com

Lung cancer is the leading cause of cancer-related mortality and requires more
osis and therapeutic responsiveness. Special AT-rich binding protein 1 (SATB1) is a
s chromatin remodeling proteins to epigenetically regulate hundreds of genes in a
s suggest that SATB1 overexpression is a predictor of poor prognosis in breast
nce of SATB1 expression has not been evaluated in lung cancer.

f 257 lung cancers was evaluated by immunohistochemistry. Epigenetic silencing of
y 5-Aza 2-deoxycytidine and trichostatin A treatment, and chromatin

einvasive lesions (p < 0.04) and
helium. Loss of SATB1
Cs; hazard ratio: 2.06, 95%
ne deacetylase inhibitor
e in the active chromatin mark
H3K27 in a SCC cell line

n is lost in early preinvasive

Ann Thorac Surg.(The Annals of thoracic surgery).1998 Dec;66(6): 1915-8

The clinical significance of hepatocyte growth factor for non-small cell lung cancer.

Siegfried, J M; Weissfeld, L A; Luketich, J D; Weyant, R J; Gubish, C T; Landreneau, R J.

Department of Pharmacology and University of Pittsburgh Cancer Institute, Lung Cancer Basic Science Program, University of Pittsburgh, Pennsylvania 15261, USA. siegfrie@server.pharm.pitt.edu

Abstract

BACKGROUND (BACKGROUND): Hepatocyte growth factor (HGF) is a cytokine that is released after injury. It is a paracrine factor that is produced by mesenchymal cells; epithelial and endothelial cells respond to HGF through its receptor, the c-met protein. Hepatocyte growth factor induces cell growth and cell movement and is also highly angiogenic. Evidence from breast cancer patients suggests that HGF is a negative prognostic indicator for breast cancer and is associated with invasive disease.

METHODS (METHODS): We measured the HGF content in tumor tissue from 56 non-small cell lung cancer patients using the Western blot technique. The amount of HGF in tumor extracts was quantitated by densitometry after transfer of proteins to nitrocellulose and exposure to antibodies. Survival curves were generated based on clinical information obtained for each patient.

RESULTS (RESULTS): Our data indicate that HGF is also a negative prognostic indicator in lung cancer. As in the study of breast cancer patients, HGF was associated with recurrence and poor survival; the relative risk was seen to increase with increasing HGF tumor content. At levels of HGF greater than 100 units, the relative risk was 10, compared with that in patients with an HGF level of 1 unit. Node-negative patients with an elevated HGF tumor content had a significantly poorer outcome than node-positive patients with a low HGF tumor content. The same relationship was observed if the patients were stratified by stage: elevated HGF was associated with stage I patients whose disease recurred and who died of their disease, and stage I patients with elevated HGF had a worse survival than higher stage patients with a low level of HGF.

CONCLUSIONS (CONCLUSIONS): These results suggest that elevated HGF may predict a more aggressive biology in non-small cell lung cancer patients. The level of HGF may be useful as an indicator of high risk in early stage lung cancer patients.

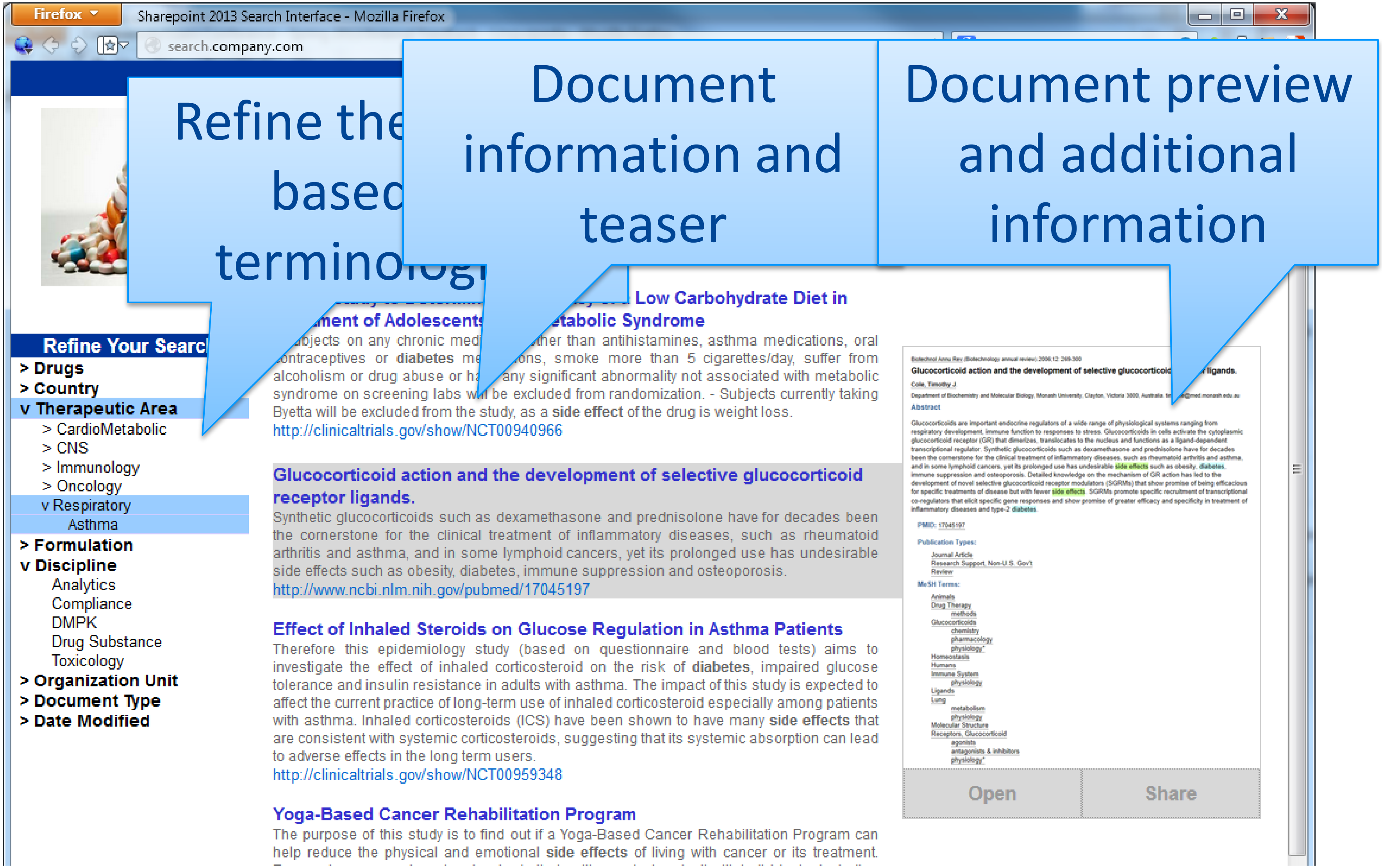
Relationships

NLP-based Text Mining

Search Engine

Linguamatics

Provide Concepts as Facets for Enterprise Search



Refine the based terminolog

Document information and teaser

Document preview and additional information

Refine Your Searc

- > Drugs
- > Country
- v Therapeutic Area
  - > CardioMetabolic
  - > CNS
  - > Immunology
  - > Oncology
  - v Respiratory
    - Asthma
- > Formulation
- v Discipline
  - Analytics
  - Compliance
  - DMPK
  - Drug Substance
  - Toxicology
- > Organization Unit
- > Document Type
- > Date Modified

Low Carbohydrate Diet in ment of Adolescents etabolic Syndrome

ubjects on any chronic medi ther than antihistamines, asthma medications, oral contraceptives or **diabetes** me ons, smoke more than 5 cigarettes/day, suffer from alcoholism or drug abuse or ha any significant abnormality not associated with metabolic syndrome on screening labs wi be excluded from randomization. - Subjects currently taking Byetta will be excluded from the study, as a **side effect** of the drug is weight loss.
http://clinicaltrials.gov/show/NCT00940966

Glucocorticoid action and the development of selective glucocorticoid receptor ligands.

Synthetic glucocorticoids such as dexamethasone and prednisolone have for decades been the cornerstone for the clinical treatment of inflammatory diseases, such as rheumatoid arthritis and asthma, and in some lymphoid cancers, yet its prolonged use has undesirable side effects such as obesity, diabetes, immune suppression and osteoporosis.
http://www.ncbi.nlm.nih.gov/pubmed/17045197

Effect of Inhaled Steroids on Glucose Regulation in Asthma Patients

Therefore this epidemiology study (based on questionnaire and blood tests) aims to investigate the effect of inhaled corticosteroid on the risk of **diabetes**, impaired glucose tolerance and insulin resistance in adults with asthma. The impact of this study is expected to affect the current practice of long-term use of inhaled corticosteroid especially among patients with asthma. Inhaled corticosteroids (ICS) have been shown to have many **side effects** that are consistent with systemic corticosteroids, suggesting that its systemic absorption can lead to adverse effects in the long term users.
http://clinicaltrials.gov/show/NCT00959348

Yoga-Based Cancer Rehabilitation Program

The purpose of this study is to find out if a Yoga-Based Cancer Rehabilitation Program can help reduce the physical and emotional **side effects** of living with cancer or its treatment.

Biotechnol Annu Rev (Biotechnology annual review) 2006;12: 269-300

Glucocorticoid action and the development of selective glucocorticoid ligands.

Cole, Timothy J

Department of Biochemistry and Molecular Biology, Monash University, Clayton, Victoria 3800, Australia. ti @med.monash.edu.au

Abstract

Glucocorticoids are important endocrine regulators of a wide range of physiological systems ranging from respiratory development, immune function to responses to stress. Glucocorticoids in cells activate the cytoplasmic glucocorticoid receptor (GR) that dimerizes, translocates to the nucleus and functions as a ligand-dependent transcriptional regulator. Synthetic glucocorticoids such as dexamethasone and prednisolone have for decades been the cornerstone for the clinical treatment of inflammatory diseases, such as rheumatoid arthritis and asthma, and in some lymphoid cancers, yet its prolonged use has undesirable side effects such as obesity, diabetes, immune suppression and osteoporosis. Detailed knowledge on the mechanism of GR action has led to the development of novel selective glucocorticoid receptor modulators (SGRMs) that show promise of being efficacious for specific treatments of disease but with fewer side effects. SGRMs promote specific recruitment of transcriptional co-regulators that elicit specific gene responses and show promise of greater efficacy and specificity in treatment of inflammatory diseases and type-2 diabetes.

PMID: 17045197

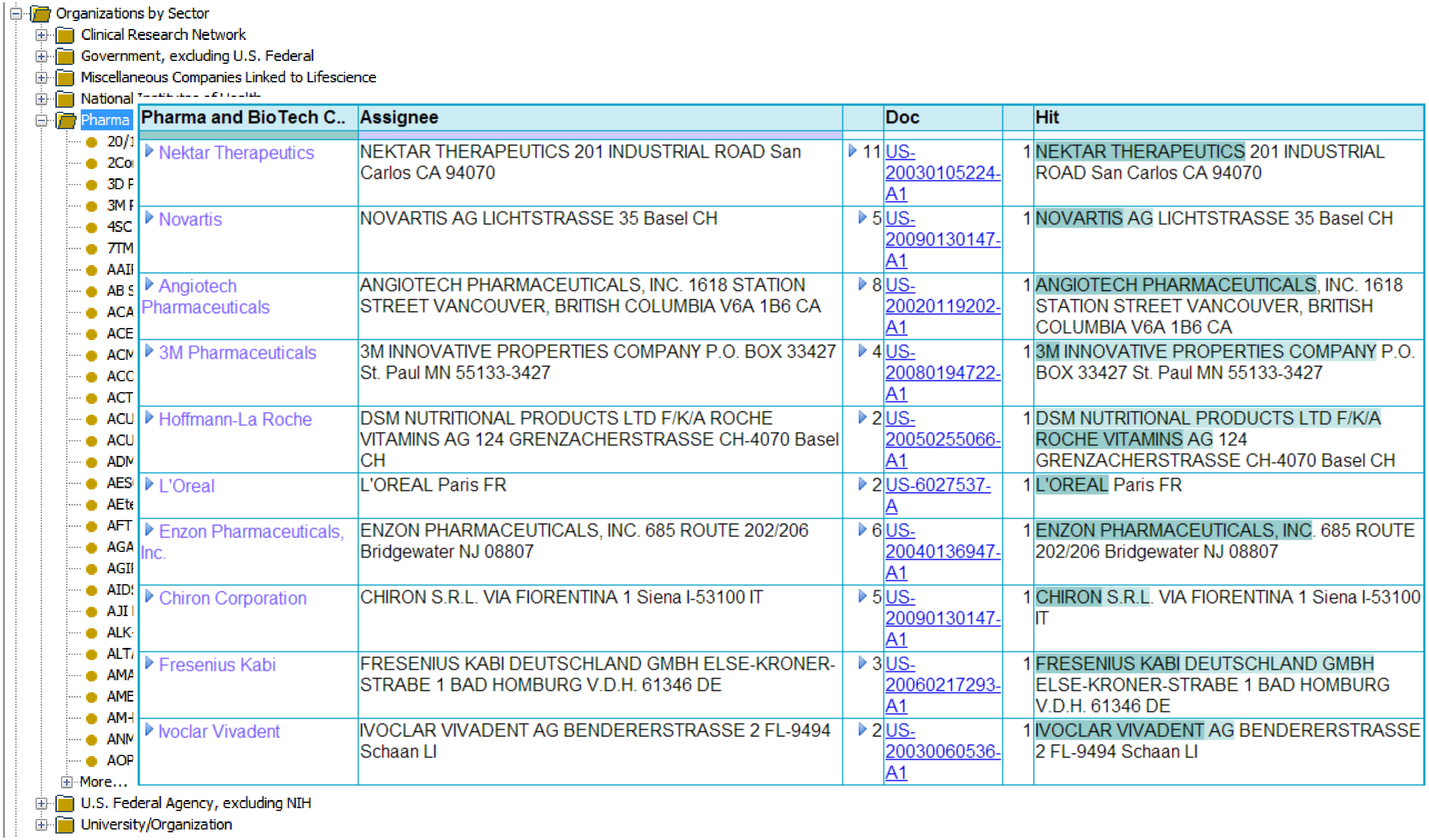
Publication Types:

Journal Article
Research Support, Non-U.S. Gov't
Review

MeSH Terms:

Animals
Drug Therapy
  methods
Glucocorticoids
  chemistry
  pharmacology
  physiology*
Homeostasis
Humans
Immune System
  physiology
Ligands
Lung
  metabolism
  physiology
Molecular Structure
Receptors, Glucocorticoid
  agonists
  antagonists & inhibitors
  physiology*

Dictionary Matching - Companies

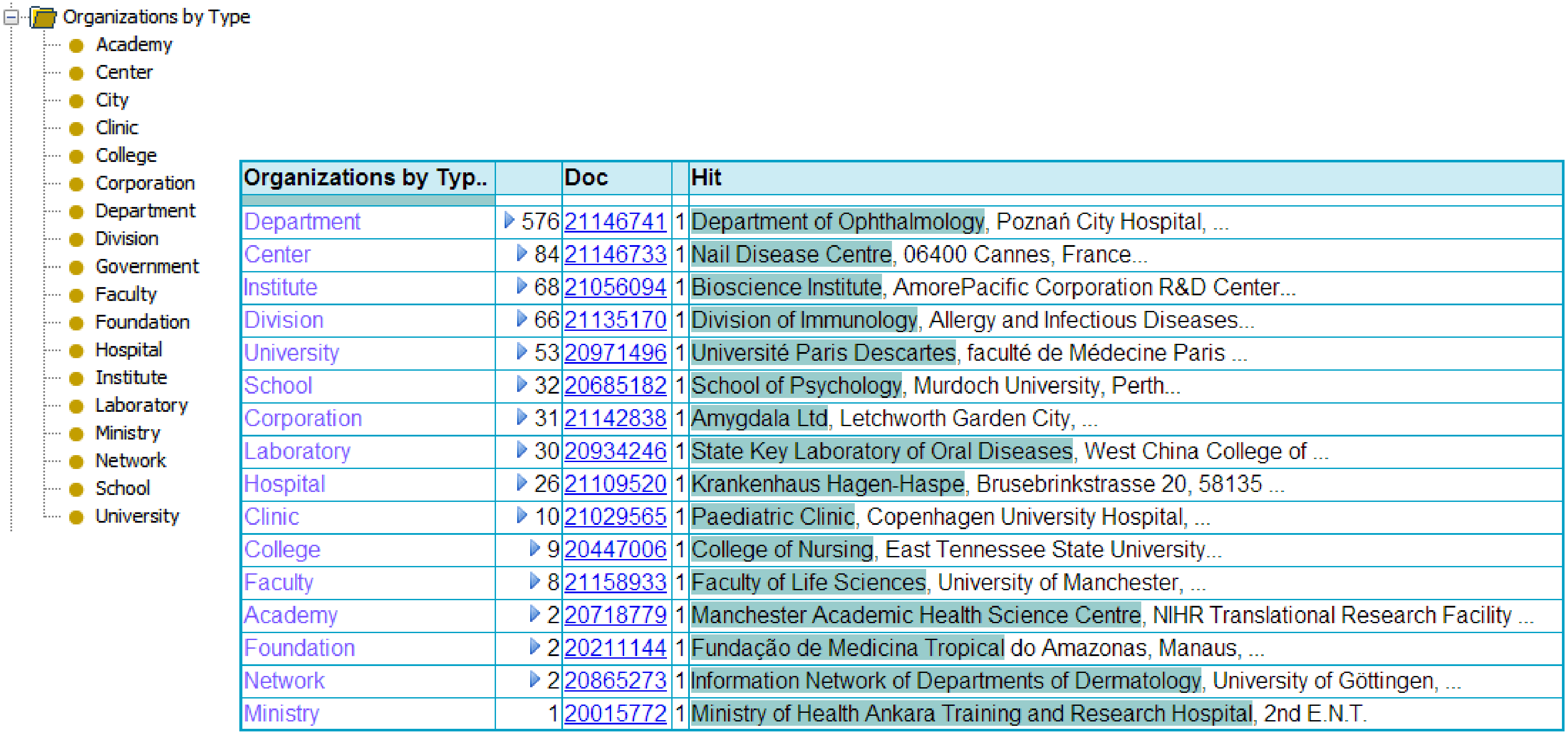


- Organizations by Sector
  - Clinical Research Network
  - Government, excluding U.S. Federal
  - Miscellaneous Companies Linked to Lifescience
  - National
  - Pharma
    - 20/1
    - 2Co
    - 3D F
    - 3M F
    - 4SC
    - 7TM
    - AAII
    - AB S
    - ACA
    - ACE
    - ACM
    - ACC
    - ACT
    - ACU
    - ACU
    - ADM
    - AES
    - AEte
    - AFT
    - AGA
    - AGII
    - AIDS
    - AJI
    - ALK
    - ALT
    - AMA
    - AME
    - AM-
    - ANM
    - AOP
    - More...
  - U.S. Federal Agency, excluding NIH
  - University/Organization

Pharma and BioTech C..	Assignee		Doc		Hit
Nektar Therapeutics	NEKTAR THERAPEUTICS 201 INDUSTRIAL ROAD San Carlos CA 94070	11	US-20030105224-A1	1	NEKTAR THERAPEUTICS 201 INDUSTRIAL ROAD San Carlos CA 94070
Novartis	NOVARTIS AG LICHTSTRASSE 35 Basel CH	5	US-20090130147-A1	1	NOVARTIS AG LICHTSTRASSE 35 Basel CH
Angiotech Pharmaceuticals	ANGIOTECH PHARMACEUTICALS, INC. 1618 STATION STREET VANCOUVER, BRITISH COLUMBIA V6A 1B6 CA	8	US-20020119202-A1	1	ANGIOTECH PHARMACEUTICALS, INC. 1618 STATION STREET VANCOUVER, BRITISH COLUMBIA V6A 1B6 CA
3M Pharmaceuticals	3M INNOVATIVE PROPERTIES COMPANY P.O. BOX 33427 St. Paul MN 55133-3427	4	US-20080194722-A1	1	3M INNOVATIVE PROPERTIES COMPANY P.O. BOX 33427 St. Paul MN 55133-3427
Hoffmann-La Roche	DSM NUTRITIONAL PRODUCTS LTD F/K/A ROCHE VITAMINS AG 124 GRENZACHERSTRASSE CH-4070 Basel CH	2	US-20050255066-A1	1	DSM NUTRITIONAL PRODUCTS LTD F/K/A ROCHE VITAMINS AG 124 GRENZACHERSTRASSE CH-4070 Basel CH
L'Oreal	L'OREAL Paris FR	2	US-6027537-A	1	L'OREAL Paris FR
Enzon Pharmaceuticals, Inc.	ENZON PHARMACEUTICALS, INC. 685 ROUTE 202/206 Bridgewater NJ 08807	6	US-20040136947-A1	1	ENZON PHARMACEUTICALS, INC. 685 ROUTE 202/206 Bridgewater NJ 08807
Chiron Corporation	CHIRON S.R.L. VIA FIORENTINA 1 Siena I-53100 IT	5	US-20090130147-A1	1	CHIRON S.R.L. VIA FIORENTINA 1 Siena I-53100 IT
Fresenius Kabi	FRESENIUS KABI DEUTSCHLAND GMBH ELSE-KRONER-STRABE 1 BAD HOMBURG V.D.H. 61346 DE	3	US-20060217293-A1	1	FRESENIUS KABI DEUTSCHLAND GMBH ELSE-KRONER-STRABE 1 BAD HOMBURG V.D.H. 61346 DE
Ivoclar Vivadent	IVOCLAR VIVADENT AG BENDERERSTRASSE 2 FL-9494 Schaan LI	2	US-20030060536-A1	1	IVOCLAR VIVADENT AG BENDERERSTRASSE 2 FL-9494 Schaan LI

Linguamatics

Pattern Matching - Institutions

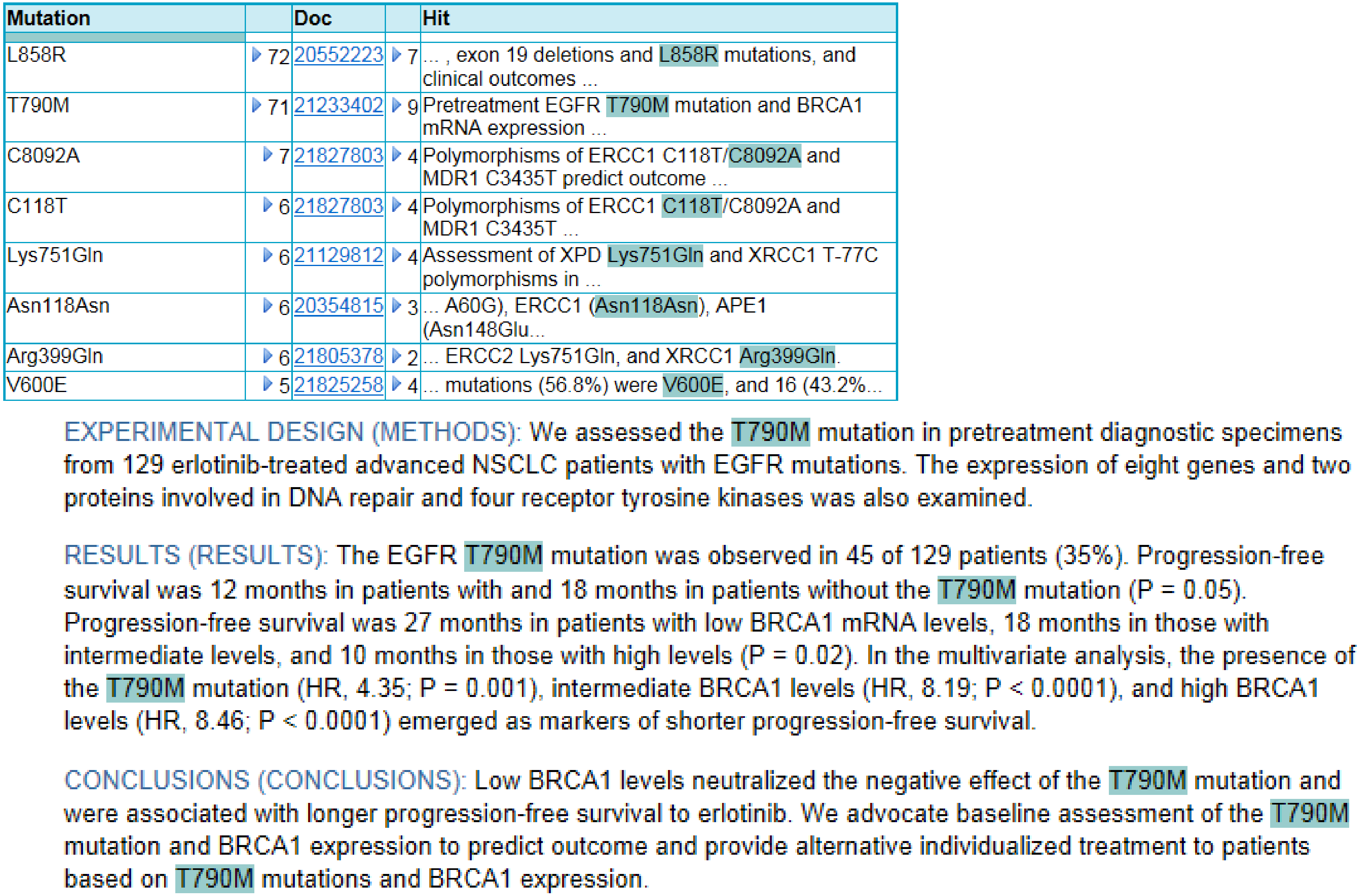


- Organizations by Type
  - ● Academy
  - ● Center
  - ● City
  - ● Clinic
  - ● College
  - ● Corporation
  - ● Department
  - ● Division
  - ● Government
  - ● Faculty
  - ● Foundation
  - ● Hospital
  - ● Institute
  - ● Laboratory
  - ● Ministry
  - ● Network
  - ● School
  - ● University

Organizations by Typ..		Doc		Hit
Department	▶ 576	21146741	1	Department of Ophthalmology, Poznań City Hospital, ...
Center	▶ 84	21146733	1	Nail Disease Centre, 06400 Cannes, France...
Institute	▶ 68	21056094	1	Bioscience Institute, AmorePacific Corporation R&D Center...
Division	▶ 66	21135170	1	Division of Immunology, Allergy and Infectious Diseases...
University	▶ 53	20971496	1	Université Paris Descartes, faculté de Médecine Paris ...
School	▶ 32	20685182	1	School of Psychology, Murdoch University, Perth...
Corporation	▶ 31	21142838	1	Amygdala Ltd, Letchworth Garden City, ...
Laboratory	▶ 30	20934246	1	State Key Laboratory of Oral Diseases, West China College of ...
Hospital	▶ 26	21109520	1	Krankenhaus Hagen-Haspe, Brusebrinkstrasse 20, 58135 ...
Clinic	▶ 10	21029565	1	Paediatric Clinic, Copenhagen University Hospital, ...
College	▶ 9	20447006	1	College of Nursing, East Tennessee State University...
Faculty	▶ 8	21158933	1	Faculty of Life Sciences, University of Manchester, ...
Academy	▶ 2	20718779	1	Manchester Academic Health Science Centre, NIHR Translational Research Facility ...
Foundation	▶ 2	20211144	1	Fundação de Medicina Tropical do Amazonas, Manaus, ...
Network	▶ 2	20865273	1	Information Network of Departments of Dermatology, University of Göttingen, ...
Ministry	1	20015772	1	Ministry of Health Ankara Training and Research Hospital, 2nd E.N.T.

- Not a fixed list: can find previously unknown institutions

Pattern Matching - Mutations



Mutation		Doc		Hit
L858R	▶ 72	20552223	▶ 7	... , exon 19 deletions and L858R mutations, and clinical outcomes ...
T790M	▶ 71	21233402	▶ 9	Pretreatment EGFR T790M mutation and BRCA1 mRNA expression ...
C8092A	▶ 7	21827803	▶ 4	Polymorphisms of ERCC1 C118T/C8092A and MDR1 C3435T predict outcome ...
C118T	▶ 6	21827803	▶ 4	Polymorphisms of ERCC1 C118T/C8092A and MDR1 C3435T ...
Lys751Gln	▶ 6	21129812	▶ 4	Assessment of XPD Lys751Gln and XRCC1 T-77C polymorphisms in ...
Asn118Asn	▶ 6	20354815	▶ 3	... A60G), ERCC1 (Asn118Asn), APE1 (Asn148Glu...
Arg399Gln	▶ 6	21805378	▶ 2	... ERCC2 Lys751Gln, and XRCC1 Arg399Gln.
V600E	▶ 5	21825258	▶ 4	... mutations (56.8%) were V600E, and 16 (43.2%...

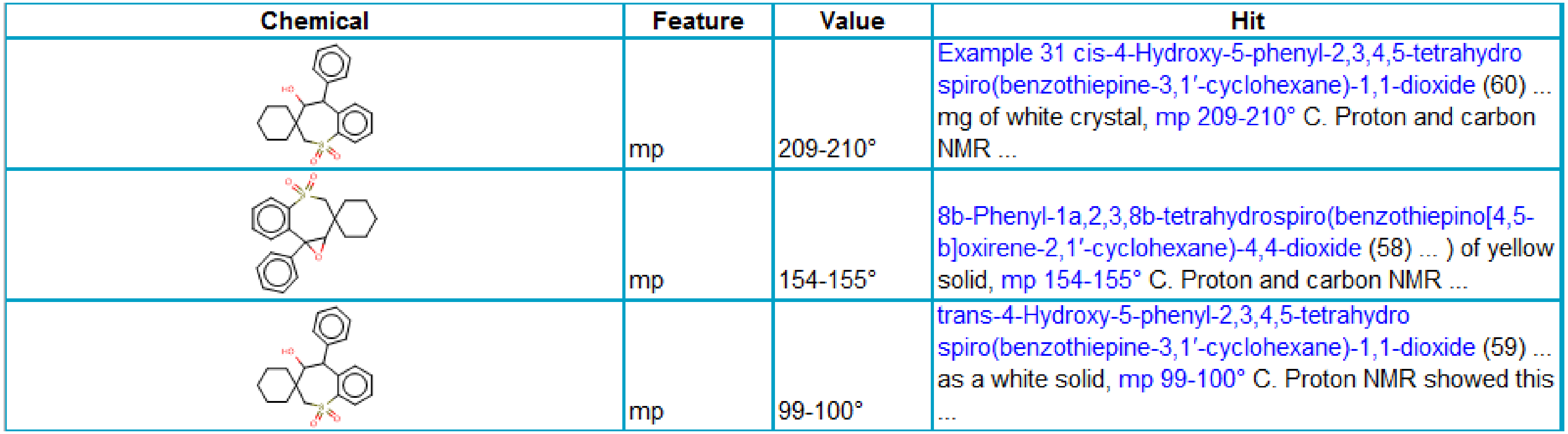
EXPERIMENTAL DESIGN (METHODS): We assessed the T790M mutation in pretreatment diagnostic specimens from 129 erlotinib-treated advanced NSCLC patients with EGFR mutations. The expression of eight genes and two proteins involved in DNA repair and four receptor tyrosine kinases was also examined.

RESULTS (RESULTS): The EGFR T790M mutation was observed in 45 of 129 patients (35%). Progression-free survival was 12 months in patients with and 18 months in patients without the T790M mutation (P = 0.05). Progression-free survival was 27 months in patients with low BRCA1 mRNA levels, 18 months in those with intermediate levels, and 10 months in those with high levels (P = 0.02). In the multivariate analysis, the presence of the T790M mutation (HR, 4.35; P = 0.001), intermediate BRCA1 levels (HR, 8.19; P < 0.0001), and high BRCA1 levels (HR, 8.46; P < 0.0001) emerged as markers of shorter progression-free survival.

CONCLUSIONS (CONCLUSIONS): Low BRCA1 levels neutralized the negative effect of the T790M mutation and were associated with longer progression-free survival to erlotinib. We advocate baseline assessment of the T790M mutation and BRCA1 expression to predict outcome and provide alternative individualized treatment to patients based on T790M mutations and BRCA1 expression.

Pattern Matching - Chemicals

- Assign structure to novel chemicals
- Distinguish exemplified compounds
- Distinguish compounds given properties

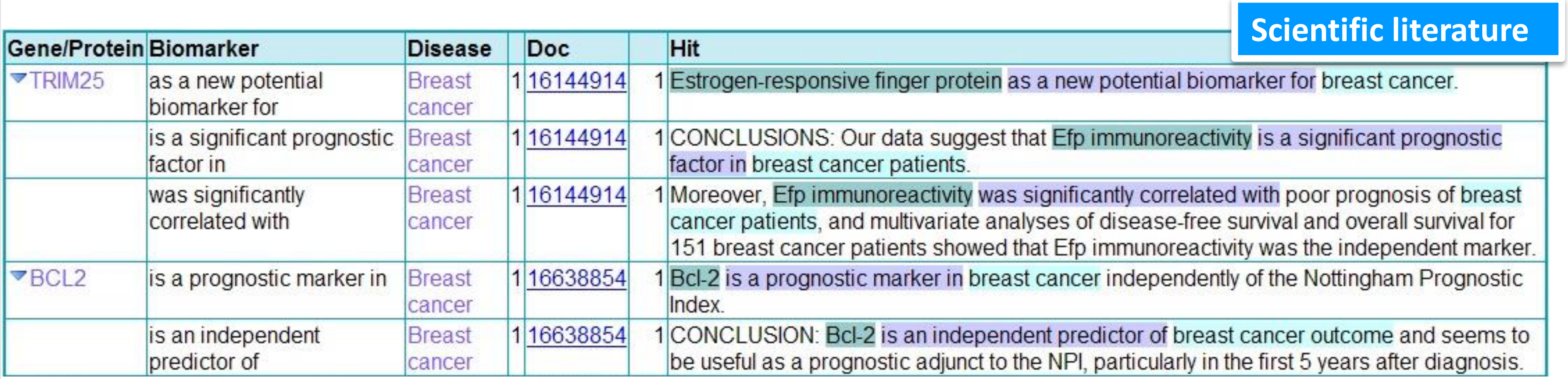


Chemical	Feature	Value	Hit
	mp	209-210°	Example 31 cis-4-Hydroxy-5-phenyl-2,3,4,5-tetrahydro spiro(benzothiepine-3,1′-cyclohexane)-1,1-dioxide (60) ... mg of white crystal, mp 209-210° C. Proton and carbon NMR ...
	mp	154-155°	8b-Phenyl-1a,2,3,8b-tetrahydrospiro(benzothiepino[4,5-b]oxirene-2,1′-cyclohexane)-4,4-dioxide (58) ...) of yellow solid, mp 154-155° C. Proton and carbon NMR ...
	mp	99-100°	trans-4-Hydroxy-5-phenyl-2,3,4,5-tetrahydro spiro(benzothiepine-3,1′-cyclohexane)-1,1-dioxide (59) ... as a white solid, mp 99-100° C. Proton NMR showed this ...

Whatever the Content...

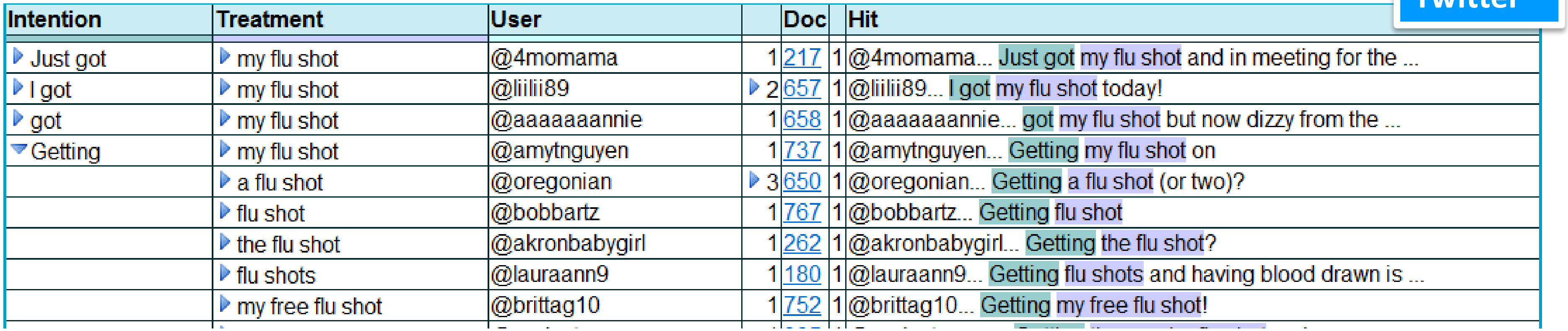
- ... I2E can mine and extract with precision

Scientific literature



Gene/Protein	Biomarker	Disease		Doc		Hit
▼TRIM25	as a new potential biomarker for	Breast cancer	1	16144914	1	Estrogen-responsive finger protein as a new potential biomarker for breast cancer.
	is a significant prognostic factor in	Breast cancer	1	16144914	1	CONCLUSIONS: Our data suggest that Efp immunoreactivity is a significant prognostic factor in breast cancer patients.
	was significantly correlated with	Breast cancer	1	16144914	1	Moreover, Efp immunoreactivity was significantly correlated with poor prognosis of breast cancer patients, and multivariate analyses of disease-free survival and overall survival for 151 breast cancer patients showed that Efp immunoreactivity was the independent marker.
▼BCL2	is a prognostic marker in	Breast cancer	1	16638854	1	Bcl-2 is a prognostic marker in breast cancer independently of the Nottingham Prognostic Index.
	is an independent predictor of	Breast cancer	1	16638854	1	CONCLUSION: Bcl-2 is an independent predictor of breast cancer outcome and seems to be useful as a prognostic adjunct to the NPI, particularly in the first 5 years after diagnosis.

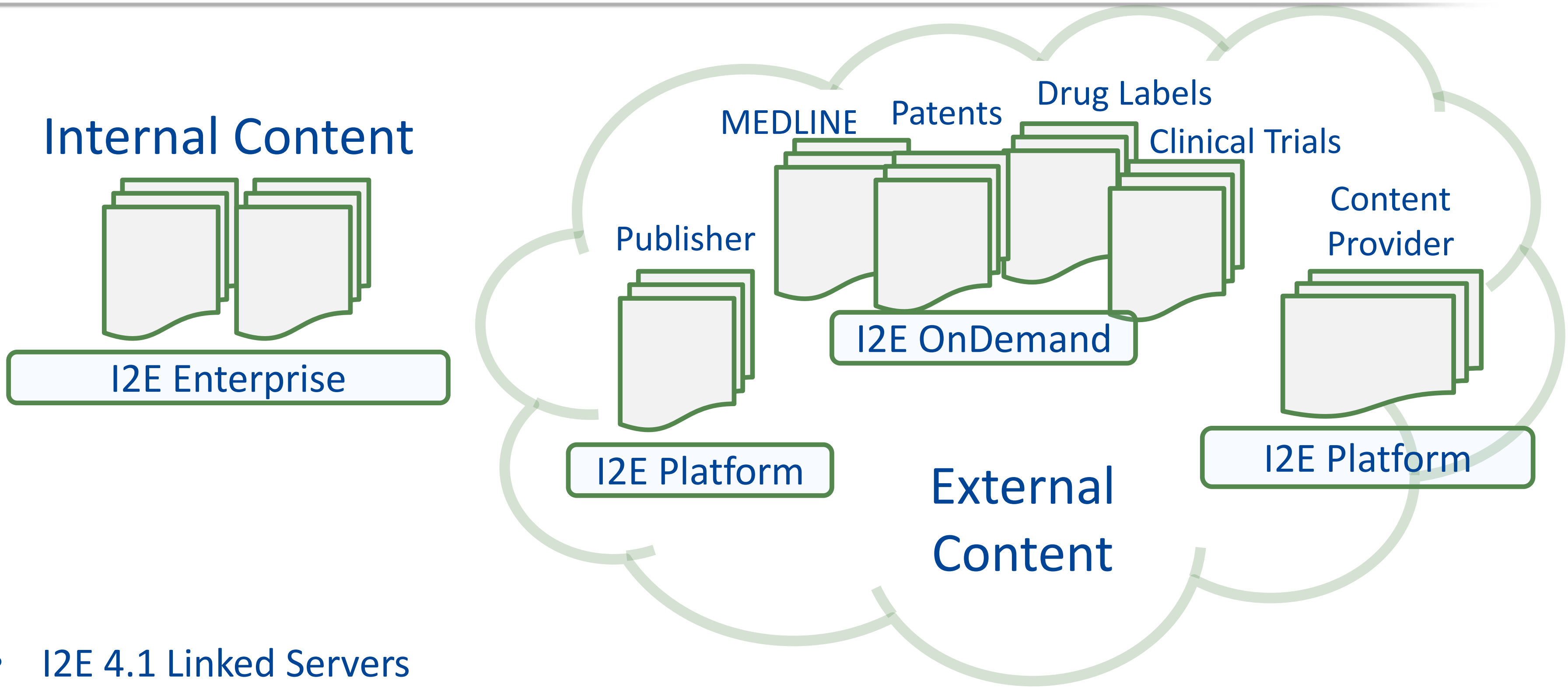
Twitter



Intention	Treatment	User		Doc		Hit
▶ Just got	▶ my flu shot	@4momama	1	217	1	@4momama... Just got my flu shot and in meeting for the ...
▶ I got	▶ my flu shot	@liilii89	▶ 2	657	1	@liilii89... I got my flu shot today!
▶ got	▶ my flu shot	@aaaaaaannie	1	658	1	@aaaaaaannie... got my flu shot but now dizzy from the ...
▼Getting	▶ my flu shot	@amytnguyen	1	737	1	@amytnguyen... Getting my flu shot on
	▶ a flu shot	@oregonian	▶ 3	650	1	@oregonian... Getting a flu shot (or two)?
	▶ flu shot	@bobbartz	1	767	1	@bobbartz... Getting flu shot
	▶ the flu shot	@akronbabygirl	1	262	1	@akronbabygirl... Getting the flu shot?
	▶ flu shots	@lauraann9	1	180	1	@lauraann9... Getting flu shots and having blood drawn is ...
	▶ my free flu shot	@brittag10	1	752	1	@brittag10... Getting my free flu shot!

Linguamatics

…. and Wherever



- I2E 4.1 Linked Servers
- Users can query local information or information on the cloud
- Proprietary content can reside within Enterprise
- Standard sources e.g. patents can be hosted on the cloud, and shared by all users
- Data from content providers can reside on their sites

Bringing the Elephant Down to Size

- Data warehouses for data that you know you need
- Embedding text mining in other applications e.g. Enterprise Search to provide benefits of semantic search
- Building specialized new interfaces to provide self-service applications for end-users
- Continuing role for ad hoc text mining
  - text mining can ask almost any question of the data
  - you can never predict every question