Making Knowledge Discoverable: The Role of Agile Text Mining

David Milward

Linguamatics

II-SDV, Nice, April 2012

Linguamatics

Overview

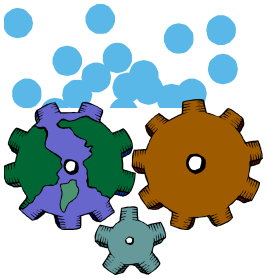
- Search vs. Text Mining
- Agile Text Mining
  - Linguamatics I2E
- Relationship Extraction: Text Mining + Search
- Finding the Most Relevant Documents: Search + Text Mining
- Accelerating a Search Strategy
- Example Results from
  - multiple documents
  - within complex documents
- Reproducible Workflows

Search vs. Text Mining

News Feeds | Scientific Literature | Patents | Internal Reports | Clinical Trials

Search Engine

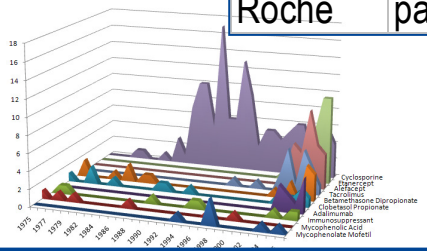
Filter to find most relevant documents, then read



company	activity	company
Sanofi	bid	Aventis
Roche	partner	Antisoma

Text Mining

Manipulate the text to discover what is there



Natural Language Processing (NLP) to understand meaning

Statistics to provide trends

Linguamatics

Agile Text Mining

- *Text mining* provides ability to **discover**
  - but typically queries have to be programmed in, and processing is slow
- *Search* provides ability to **filter** quickly to relevant documents
  - but poor at answering open questions e.g. “what are biomarkers for breast cancer”
- Combine *text mining* with *search* to **discover** within **specific contexts** e.g.

What | is a risk factor for diabetes

Discover what is available
Smoking
alcohol intake
Diet
Excess adiposity
Heavy alcohol consumption
heavy alcohol intake
lower birth weight

Filter to the context of interest
is a risk factor for type 2 diabetes
is a potential risk factor for incident diabetes
is a potential risk factor for newly diagnosed type 2 diabetes mellitus
is a potentially modifiable risk factor for diabetes
is a risk factor for diabetes
is a risk factor for diabetes mellitus
is an established risk factor for diabetes

Technology Adoption

- Linguamatics I2E first adopted in pharma/biotech, including 9 of top 10
- Used to answer a wide range of questions e.g.
  - *Dose durations of follow-on clinical trials*
  - *Therapeutic usages of recombinant proteins*
  - *Cofactors for Nuclear Receptors*
- Wide range of application areas across the drug pipeline e.g.
  - Target Prioritization, Safety/Toxicity, Clinical Trial Design
  - Competitive Intelligence, Marketing
- Are the Life Sciences special?
  - Particularly knowledge intensive, so high demand
  - A lot of complex, ambiguous terminology, balanced by good resources
- I2E is a generic platform
  - Now being used in chemicals, consumer products, health …

Example Data Sources

Competitive Intelligence, R&D, Marketing

- Scientific Literature
  - Abstracts e.g. MEDLINE
  - Full text journal articles e.g. via Quosa
- Patents
- News
- Conference Abstracts
- Internal documents
- Social Media
  - Twitter

Clinical Trial Design, Safety, Relative Efficacy of Drugs

- Clinical Trials
- FDA Drug Label Inserts
- Electronic Health Records

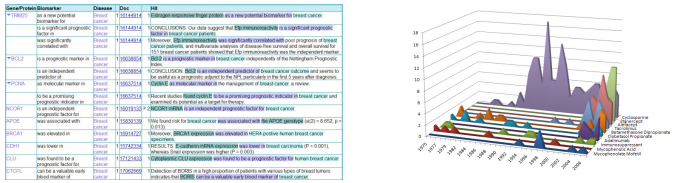
Cloud Based Service

- MEDLINE (20 million abstracts)
- Patent Full Text
  - USPTO, WIPO, EPO …
- *Clinical Trials …*

Linguamatics I2E: An Agile Text Mining Platform

Decision Support

Structured Results (Actionable Information)



Agile NLP Querying

- Ad hoc queries
- Smart queries
- Multi queries
- Batch queries
- NLP
- Class, concept
- Regular Expression
- Chemical Structure

Indexing

Flexible, Highly Scalable Rich Indexes

Ontologies

Domain knowledge – ontologies, thesauri, dictionaries

Documents

HTML XML PDF Adobe

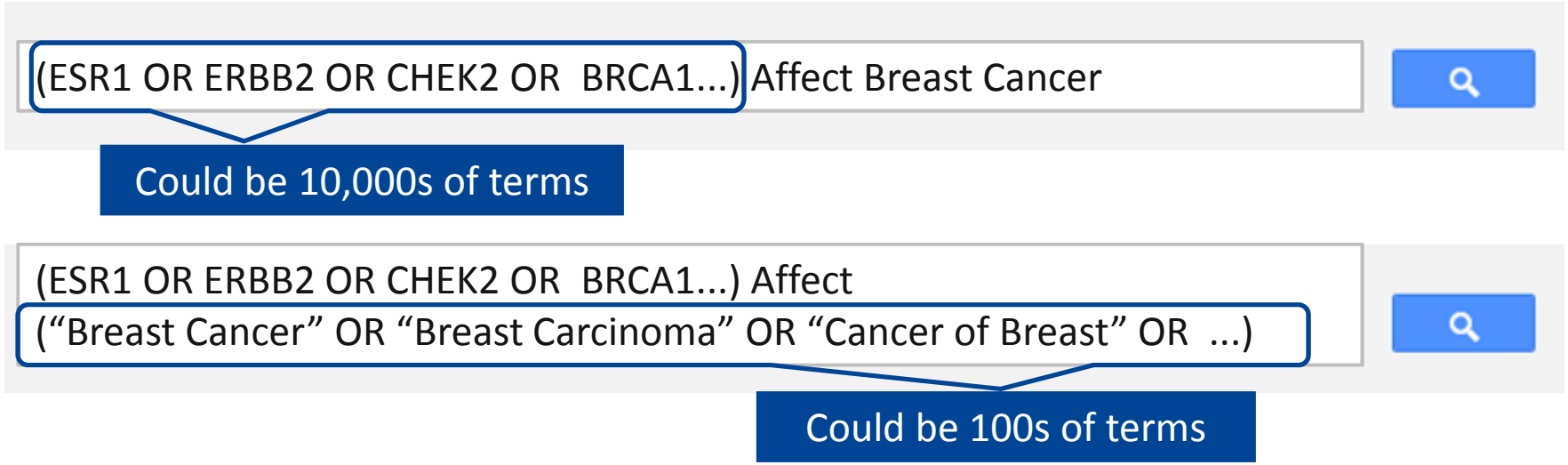
Internal/External Sources

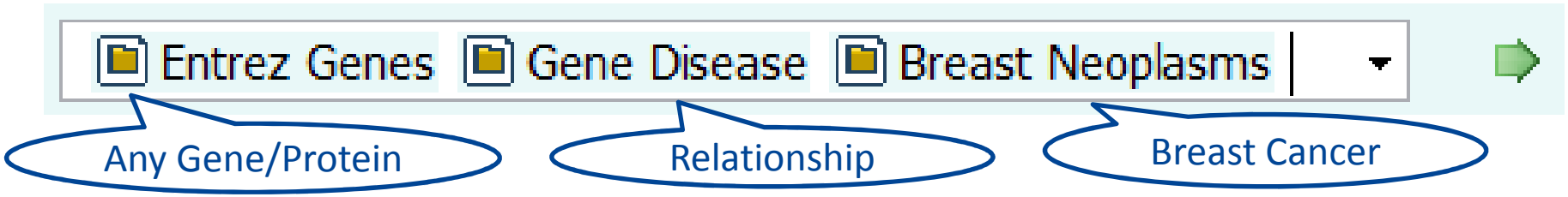
Linguamatics

Using I2E to Extract Relationships

Improve Recall Using Terminologies

Which genes are known to affect breast cancer?





HER2 oncoprotein plays an essential role in breast cancer growth and differentiation.

Linguamatics

Extracting Relationships using NLP: Biomarkers

Gene (from Entrez)

Complex linguistic relationship

Disease (from MedDRA)

Link to source document

Relevant sentence extracted with terms highlighted

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.
PCNA	as molecular marker in	Breast cancer	1	16637514	1	Cyclin E as molecular marker in the management of breast cancer: a review.
	to be a promising prognostic indicator in	Breast cancer	1	16637514	1	Recent studies found cyclin E to be a promising prognostic indicator in breast cancer and examined its potential as a target for therapy.
NCOR1	is an independent prognostic factor for	Breast cancer	1	16019133	2	NCOR1 mRNA is an independent prognostic factor for breast cancer.
APOE	was associated with	Breast cancer	1	15830139	1	We found risk for breast cancer was associated with the APOE genotype (xi(2) = 8.652, p = 0.013).
BRCA1	was elevated in	Breast cancer	1	16914727	1	Moreover, BRCA1 expression was elevated in HER4-postive human breast cancer specimens.
CDH1	was lower in	Breast cancer	1	15742334	1	RESULTS: E-cadherin mRNA expression was lower in breast carcinoma (P = 0.001), whereas Snail expression was higher (P = 0.003).
CLU	was found to be a prognostic factor for	Breast cancer	1	17121433	1	Cytoplasmic CLU expression was found to be a prognostic factor for human breast cancer.
CTCFL	can be a valuable early blood marker of	Breast cancer	1	17062669	1	Detection of BORIS in a high proportion of patients with various types of breast tumors indicates that BORIS can be a valuable early blood marker of breast cancer.

Extracting Numerical Data in Context: Safety

Compound

Potential safety issues

In this organ

At this dosage

Pharmacologic Substance	Safety issues	Tissue	Dosage		Doc		Hit
Cyclosporine	Safety issues	Kidney	15 mg/kg/day	7	17497475	22	... of hyperbaric oxygen on cyclosporine-induced nephrotoxicity and oxidative stress in rats... : a control group, a cyclosporine group (15 mg/kg/day intraperitoneally for 14 days)...
Mycophenolate Mofetil	Safety issues	Kidney	2 g/day	4	16152998	7	Uncommon side effect of MMF in renal transplant recipients. ... Patients used MMF 2 g/day.
Amphotericin B	Safety issues	Kidney	0.4 mg/kg/day	2	15761070	8	Low nephrotoxicity of an effective amphotericin B ... for 10 consecutive days with 0.4 mg/kg/day AMB in the form of traditional ...
Everolimus	Safety issues	Kidney	1.5 mg/day	3	16041270	3	Concerns over nephrotoxicity led to a protocol amendment ... loss to follow-up) were everolimus 1.5 mg/day, 33.7% (65/...
Gentamicin	Safety issues	Kidney	100 mg/kg/day	2	14748758	6	... 4-hydroxy tempo) on gentamicin-induced nephrotoxicity in rats. ... The rats were given gentamicin (100 mg/kg/day, i.p., once a ...
Alcohol	Safety issues	Liver	20 g/day	1	15553597	7	Insulin sensitivity and hepatic steatosis in obese subjects with ... analyzed 86 obese patients whose alcohol intake was less than 20 g/day and who showed no signs ...
Lamivudine	Safety issues	Liver	100 mg/day	5	17283489	1	Lamivudine (100 mg/day) was continued throughout the ... Liver Transplantation adverse effects
Sirolimus	Safety issues	Kidney	1 mg/day	2	16364861	2	... primary immunosuppressant in calcineurin inhibitor-induced nephrotoxicity. ... Sirolimus was started at 1 mg/day with titration over 2 weeks ...
		Liver	2 mg/day	1	15899725	1	... of sirolimus, participants received sirolimus 2 mg/day for at least 7 days... dropped out because of trimethoprim-sulfamethoxazole-related hepatotoxicity.
Atorvastatin	Safety issues	Liver	10 mg/day	3	17473378	3	... a potential role in statin-related adverse events, and withdrawal of ... in patients developing myotoxicity or liver toxicity. ... Twenty-six patients with hypercholesterolemia received atorvastatin at 10 mg/day for 3 months.
			80 mg/day	1	16731999	1	... that intensive lipid-lowering therapy with atorvastatin 80 mg/day provides significant clinical benefit beyond ... in the rates of treatment-related adverse events and persistent elevations in liver enzymes.
		Kidney	10 mg/d	1	17889157	2	...) were treated posttransplantation with atorvastatin (10 mg/d) for 12 weeks without ... Kidney Transplantation adverse effects immunology physiology

Standardizing/Clustering to Save Review

- Go directly to answers, e.g. find all the genes associated with a specific disease
- Highlighted evidence and link to the document
- Save time in review:
  - Rather than reading all 470 documents for ERBB2, just read enough to check relationship exists
  - Can then concentrate on the longer tail of less well-known genes/proteins
- Customer reports of order of magnitude speed-up

Entrez Genes		Doc
ERBB2	470	15870086
BRCA1	248	15564800
BRCA2	160	15986445
PGR	129	15272277
EGFR	101	16280056
VEGFA	76	15897560
TP53	73	15583825
INS	68	15805581
CCND1	51	16140974
AKT1	50	16619501

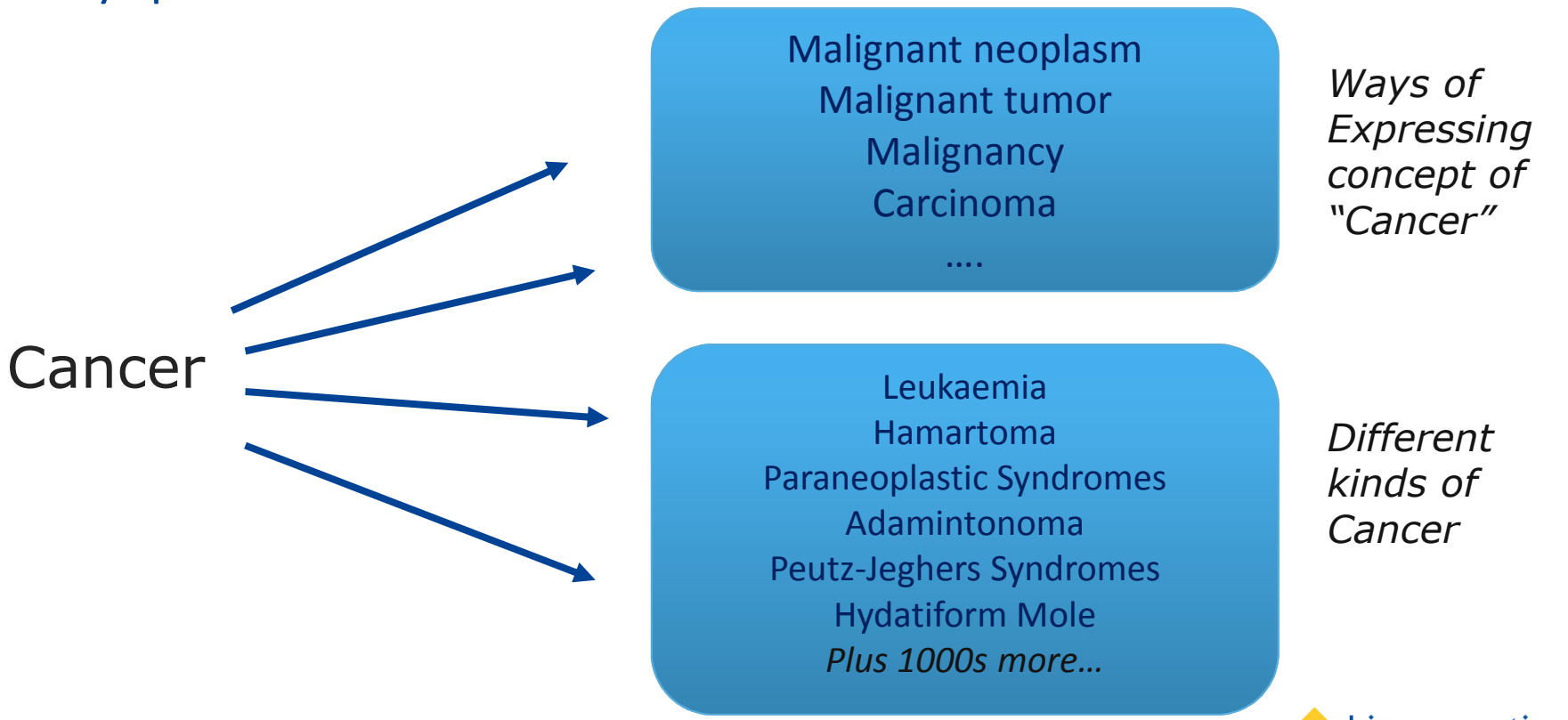
Finding the Most Relevant Documents

Finding Relevant Documents

- More comprehensive results
  - Terminologies
  - Regular Expressions
    - Precise expressions e.g. for miRNA (simplified)
      - **let-?\d+.***
      - **mirn?a?-?\d+.***
  - High throughput searches
    - lists of 500 genes, chemicals etc.
  - Chemical substructure and similarity searching
- Reduced noise in results
  - Use of linguistics rather than distance e.g. n words
  - Regions
    - abstract, methods, claims, claim, table
  - Local negation e.g. “dead” for death, but not “dead time”

More Comprehensive Results: Terminologies

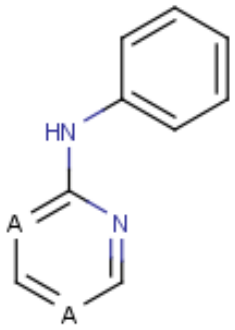
- Re-use terminologies with 10s of thousands of concepts, and 100s of thousands of synonyms
- If we are interested in genes associated with cancer, we don't just want synonyms of "cancer" e.g. "malignant neoplasm", but also any specific cancer



Linguamatics

Chemical Text Mining

- Ability to efficiently answer precise queries e.g.
  - What chemicals with this substructure act as inhibitors



Class1	Relation	Class2		Doc		Hit
imatinib	inhibit	ABL1	2	16303243	1	Imatinib, an inhibitor of BCR-ABL tyrosine kinase, also inhibits BCRP-mediated drug transport.
imatinib methanesulfonate	inhibit	ABL1	3	15803362	1	BACKGROUND: Imatinib mesylate is a potent inhibitor of Abl, KIT, and PDGFR tyrosine kinases.
gefitinib	inhibit	EGFR	2	15692759	2	The clinical benefit and safety of the EGFR tyrosine kinase inhibitor gefitinib ('Iressa')1 was evaluated in this Phase II, multicentre study of patients with taxane and anthracycline pretreated, metastatic breast cancer.
lapatinib	inhibit	ERBB2	3	16452223	2	Alternatively, inhibition of ErbB2 signaling using lapatinib (GW572016), a reversible small -molecule inhibitor of ErbB1/ErbB2 tyrosine kinases, at pharmacologically relevant concentrations, leads to marked inhibition of survivin protein with subsequent apoptosis.

Accelerating a Search Strategy

Contextual Landscaping

Looking for words before or after the word of interest

a chocolate	48
and chocolate	37
the chocolate	35
of chocolate	35
chocolate or	31
chocolate and	29
or chocolate	25
as chocolate	25
dark chocolate	20
chocolate is	18
milk chocolate	17
in chocolate	16
The chocolate	15
chocolate products	14
hot chocolate	14
chocolate bars	13

Linguistics to reduce noise, to find types of chocolate

dark chocolate	20
milk chocolate	17
chocolate products	14
hot chocolate	14
chocolate bars	13
chocolate milk	13
chocolate bar	12
white chocolate	12
chocolate liquor	9
chocolate product	9
chocolate production	8
chocolate mass	7
chocolate flavoring	7
chocolate flavor	7
chocolate chips	6
molten chocolate	6

Results from 100K USPTO Patents

Which Regions Does this Term Appear In

- In search we can often restrict to a particular region of a document
- In text mining we can first check all the regions that the term appears in
  - E.g. look for the regions that IC50 appears in
- See what you want and what you might lose
  - We then have better evidence to restrict to particular regions

Description (EN)	1446	US-20110212940-A1	1	RELATED APPLICATIONS This application is a divisional application of copending U.S.... inflection point (EC50IP or IC 50 IP) ... slope (EC50 slope or IC50 slope). ...
Claims (EN)	37	US-20110251266-A1	1	1 . An aptamer that binds to NGF and inhibits binding of NGF and an NGF receptor.... a 50% inhibitory concentration (IC50) of not more than ... a 50% inhibitory concentration (IC50) of not more than ...
Abstract (EN)	4	US-20110245180-A1	1	Chemokine signaling is important in neuropathic pain, with microglial cells expressing CCR2 playing a well established key role.... antagonism for both CCR2 (IC 50 4.2 pM) and CCR5 (IC 50 0.18 pM) in monocyte ...

Results from 100K 2011 Patents

What IPC Codes are Assigned to a Company's Patents

- Find most frequent codes in patents where a company or set of companies (e.g. pharma) is the Applicant or Assignee

IPCR	Group	Subgroup	Organization		Doc		Hit
A61K	31	506	Abbott	5	US-20120003291-A1	3	ABBOTT CARDIOVASCULAR SYSTEMS
C07D	403	12	Abbott	3	US-8101754-B2	3	ABBOTT GMBH & CO KG
A61P	25	00	Abbott	6	US-20120022103-A1	3	ABBOTT LAB
A61M	25	00	Abbott	4	US-8100856-B2	3	ABBOTT LAB VASCULAR ENTPR LTD
G01N	33	53	Abbott	3	US-20120020972-A1	3	ABBOTT LAB
C12N	5	16	Abbott	3	US-20120020972-A1	3	ABBOTT LAB
A61F	2	06	Abbott	5	US-8088157-B2	3	ABBOTT LAB VASCULAR ENTPR LTD
C07K	14	00	Abbott	2	US-8101565-B2	3	ABBOTT CARDIOVASCULAR SYSTEMS
C07C	311	14	Abbott	2	US-20120040947-A1	2	ABBOTT GMBH & CO.
A61B	17	04	Abbott	2	US-8100923-B2	3	ABBOTT LAB
B65D	43	10	Abbott	1	US-20120000929-A1	3	ABBOTT LAB
C12P	21	00	Abbott	1	US-20110318349-A1	3	ABBOTT LAB
C12Q	1	68	Abbott	4	US-8092998-B2	3	ABBOTT LAB
C07F	9	28	Abbott	1	US-20120022252-A1	3	ABBOTT LAB

Results from recent USPTO Patents

Summarizing Results from Multiple Documents

for Efficient Review and Integration

Linguamatics

Semantics: 1000s of ways of saying the same thing

- Find and extract patterns to find a particular concept
- For example: how do we distinguish people:
  - willing to get a vaccine
  - not willing
- If we can partition the two populations we can then see how they are influenced

Concept of getting a vaccine

Just got	my flu shot
I got	a flu shot
got	the flu shot
Getting	my first flu shot
Got	my seasonal flu shot
just got	flu shot
I just got	my Flu shot
i got	a Flu shot
Had	a flu vaccine
getting	a flu vaccine shot
I'm getting	an H1N1 shot

Examples from Twitter

Semantics: Standardized Identifiers for Concepts

- Terminologies can be used to link individual synonyms to a semantic concept
- Concept is uniquely identified by the ontology and the node identifier (and, typically, by the preferred term)

[PT] Cyclosporine	[SNID] Cyclosporine	Synonym		Doc		Hit
Cyclosporine	nci.C406	cyclosporine	960	7822521	20	Low-dose short-term cyclosporine versus etretinate in psoriasis: ...
		cyclosporin	336	2456356	17	Levels of cyclosporin in epidermis of treated psoriasis ...
		csa	120	2451427	12	... action of cyclosporin A (CsA) in psoriasis and other ...
		ciclosporin	104	17986302	10	Methotrexate vs. ciclosporin in psoriasis: effectiveness, ...
		neoral	30	9374930	11	Neoral--new cyclosporin for old...
		sandimmun	19	9764154	6	... of cyclosporin, Neoral and Sandimmun, in patients with severe ...
		sandimmune	6	8786246	5	... formulations of cyclosporine A (Sandimmune and Neoral; Sandoz Pharmaceuticals...

- This allows better clustering of results, better statistics, and allows us to connect results from text mining with other databases, or the semantic web

Semantics: Directed Relationships

- Extract the same relationship, even if it is expressed very differently e.g.
  - SRC phosphorylates EGFR
  - phosphorylation of EGFR by SRC
  - EGFR is phosphorylated by SRC
- Establish the direction of the relationship
  - SRC phosphorylates EGFR, not EGFR phosphorylates SRC

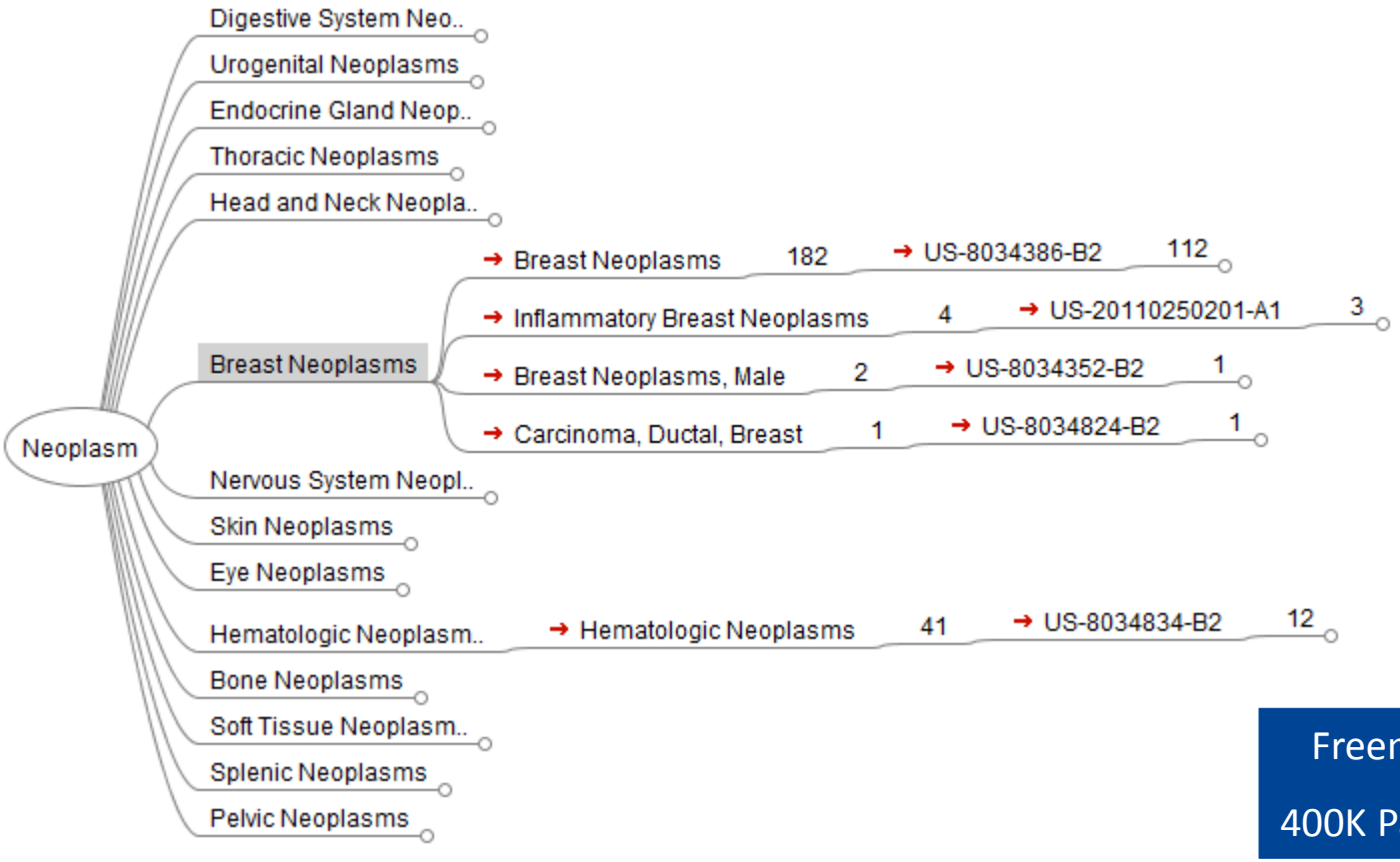
SRC	phosphorylate	EGFR	Co-incubation of purified c-Src and EGFR results in phosphorylation of Tyr 845 in vitro, indicating that c-Src can directly phosphorylate EGFR on Tyr 845.	14750954.txt
SRC	phosphorylate	EGFR	the C-terminal autophosphorylation domain of EGFR was extensively phosphorylated by c-src and EGFR kinase activities in vitro as determined by electrospay ionization mass spectrometry.	8845374.txt
SRC	phosphorylate	EGFR	the phosphorylation of EGFR by c-src was studied in order to evaluate the molecular basis for this interaction.	8845374.txt

Finding Information Across Multiple Documents

Linguamatics

Document Categorization

- Categorizing Patents according to Disease Area
  - Linguistics and ontologies provide better context, greater insight



Freemind

400K Patents

Linguamatics

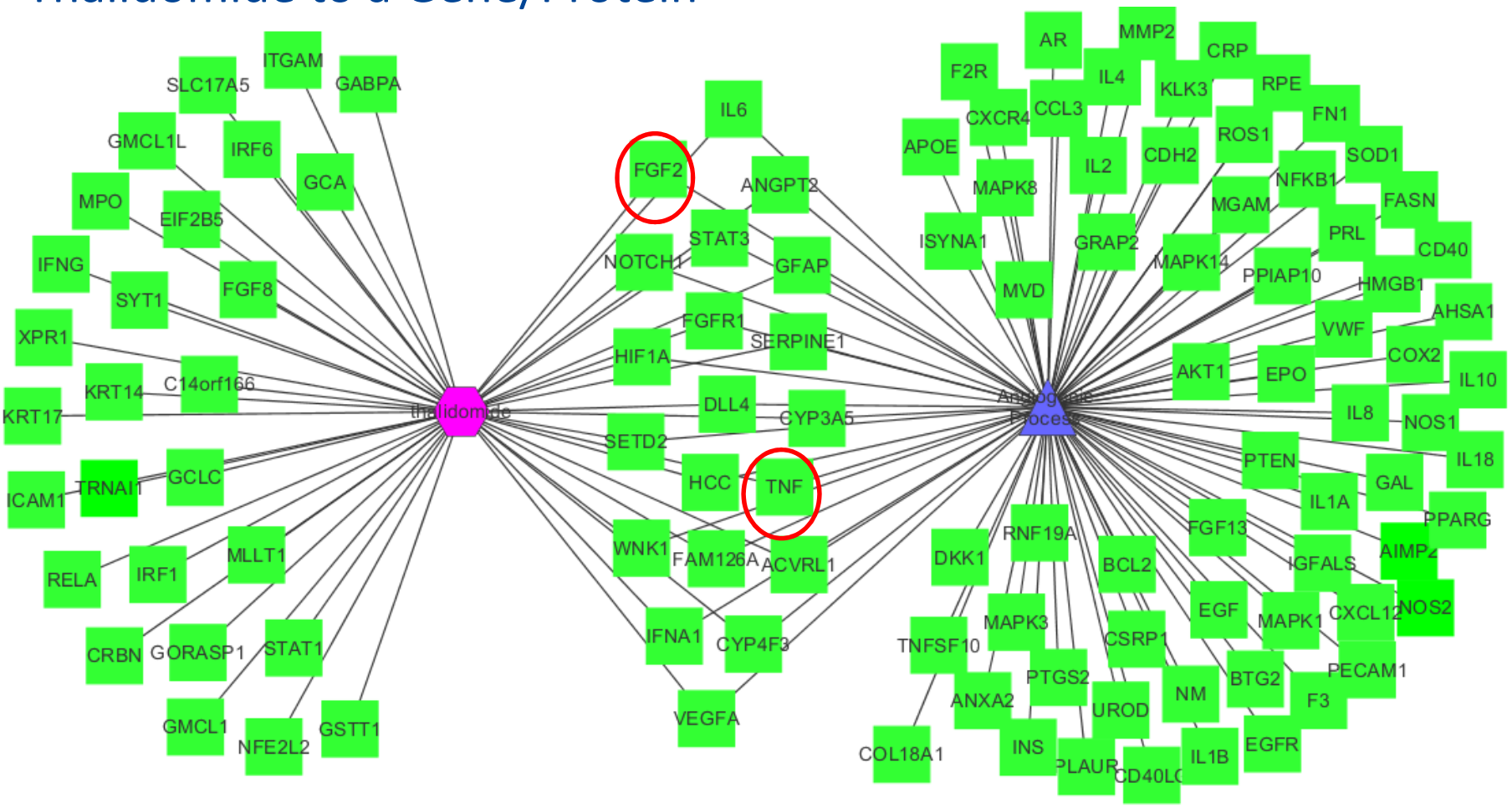
Tell me everything about X

- Search provides most relevant documents mentioning X
- Text mining can summarize distinct properties of X by clustering facts extracted from all documents

Pharmacologic Substance	Relation	Entity		Doc		Hit	qID
Cyclosporine	treats	Psoriasis	81	1401311	3	Cyclosporine therapy for psoriasis: a cell cycle-derived dosing schedule.	7
	dosage	5 mg/kg/day	15	1552052	3	Alterations in renal function in psoriasis patients treated with cyclosporine, 5 mg/kg/day.	13
	causes	nephrotoxic	2	11210296	1	Methotrexate treatment can lead to bone marrow suppression and hepatotoxicity, and cyclosporine can cause nephrotoxicity.	2
	inhibit	Growth factor	2	8884530	1	FK506 and cyclosporin A inhibit growth factor-stimulated human keratinocyte proliferation by blocking cells in the G0/G1 phases of the cell cycle.	1
	affect	PGP	1	15788683	1	CONCLUSIONS: Cyclosporin A modulates Pgp, MRP-1, BCRP, and LRP, and this broad-spectrum activity may contribute to its clinical efficacy.	1
	IC50	3.0 ng/ml	1	9585801	1	We classified these patients into two groups on the basis of their PBMC sensitivity to cyclosporine with use of the median cyclosporine IC50 (3.0 ng/ml) of these patients as the cutoff point.	12

Linking Knowledge: Indirect Relationships

Thalidomide to a Gene/Protein



Visualization via Cytoscape

A Gene/Protein to Angiogenesis

Linguamatics

Within a Document

Linking Within Complex Patent Documents

- Linking information in one part of a patent to another e.g.
  - Finding compounds with a particular substructure where a value is reported

EXAMPLE 12

2-(R)-2-(2-Amino-2-methylpropionylamino)-3-(2,4-difluorophenyl)methoxy propionic acid N-[5-(4-chlorophenyl)-3,3-dimethyl-1,1-dioxo-2,3-dihydroisothiazol-4-ylmethyl]-N-ethylamide Hydrochloride (Z=2,4-F_2)

...

Example 5	0.6
Example 7	1.7
Example 12	8.0
Example 15	2.1

Chemical	Inhibition Metric	Value	Units		Doc
2-(R)-2-(2-Amino-2-methylpropionylamino)-3-(2,4-difluorophenyl)methoxy propionic acid N-[5-(4-chlorophenyl)-3,3-dimethyl-1,1-dioxo-2,3-dihydroisothiazol-4-ylmethyl]-N-ethylamide Hydrochloride (Z=2,4-F 2)	EC50	8.0	nM	2	US-7396846-B2

From One Claim to the Next

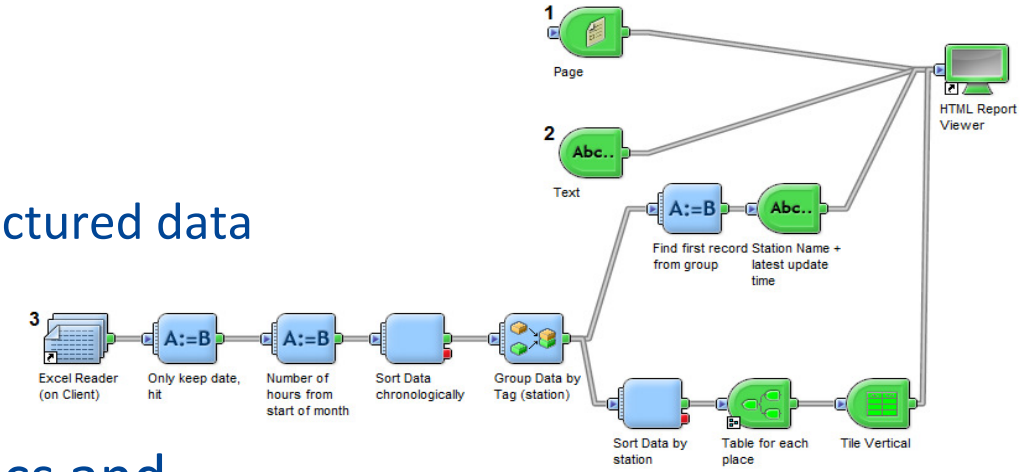
- For information in claims, often want to work back along the chain of claims

7582287	14. The method of claim 12 , wherein the chemotherapeutic agent is selected from the group consisting of bleomycin, chlorambucil, epirubicin, 5-fluorouracil, ifosfamide, mitomycin, methotrexate, vincristine, cisplatin and vinblastine.	12. The method of claim 10 , wherein the cervical cancer cells are treated with radiation in conjunction with the polypeptide.	10. The method of claim 5 , wherein the polypeptide comprises amino acid residues 1-176 of SEQ ID NO:2.	5. A method for inhibiting the growth and or proliferation of cervical cancer cells comprising bringing a polypeptide comprising amino acid residues 33-176 of SEQ ID NO:2 into contact with the cervical cancer cells.	1	7582287	1	... claim 12 , wherein the chemotherapeutic agent is selected from the ...
	7. The method of claim 5 , wherein the cervical cancer cells are treated with one or more chemotherapeutic agents in conjunction with the polypeptide.	5. A method for inhibiting the growth and or proliferation of cervical cancer cells comprising bringing a polypeptide comprising amino acid residues 33-176 of SEQ ID NO:2 into contact with the cervical cancer cells.			1	7582287	1	... treated with one or more chemotherapeutic agents in conjunction with the ...

Reproducible Workflows

Automation

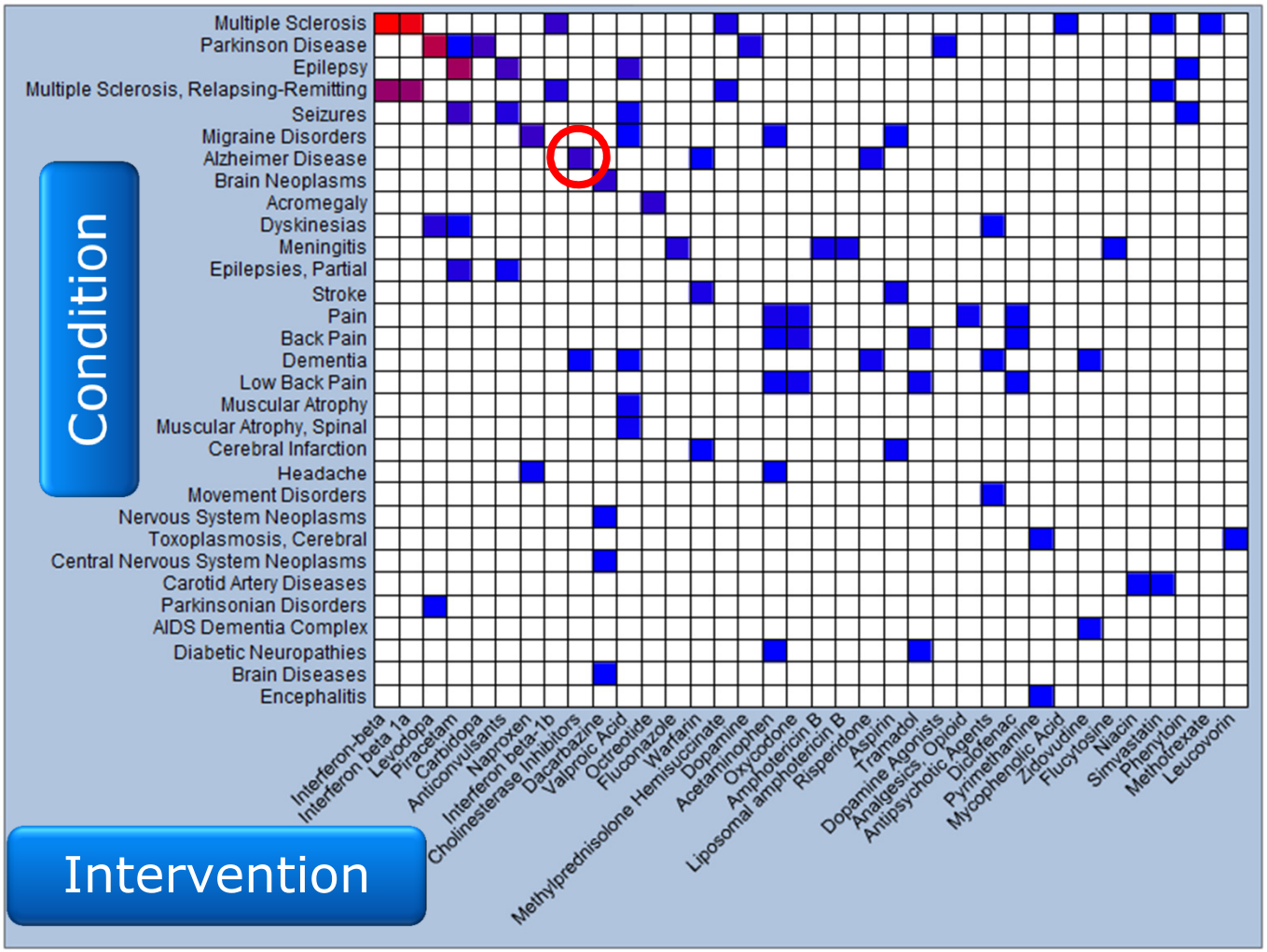
- Queries can be run regularly as documents get updated
- Integrate text mining with workflow tools for
  - up-to-date dashboards
  - alerts
  - integration with other structured data sources
  - web portals
- Provide a range of analytics and visualization



Pipeline Pilot

Linguamatics

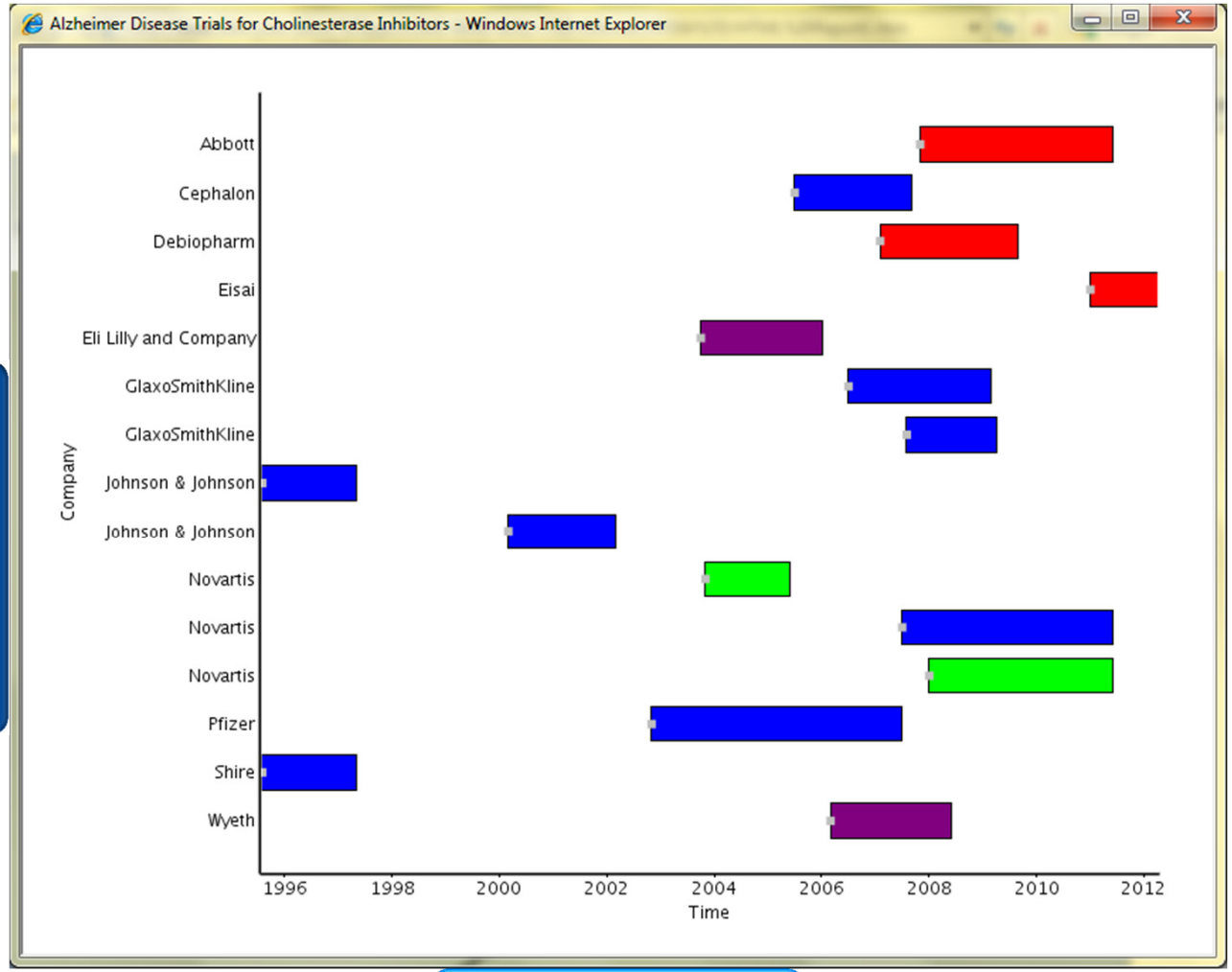
Clinical Trials Analysis from I2E Text Mining Results



Using Pipeline Pilot Visualization

Linguamatics

Clinical Trials Analysis from I2E Text Mining Results



Sponsors

Colours indicate Phase

Dates

Using Pipeline Pilot visualization

Linguamatics

Conclusions

- Agile Text Mining provides query power and flexibility
  - Can address the long tail of less predictable questions
  - Allows “queries of arbitrary complexity” combining
    - NLP, ontologies, regular expressions, regions, numerical expressions, chemical substructure/similarity, disambiguation
  - Uses the data itself to
    - inform search strategies
    - build terminologies
  - Provides flexible, structured output to
    - allow integration with existing structured data
    - fit into existing workflows
- This enables very wide application of text mining: wherever there is a need to search, extract, or categorize information