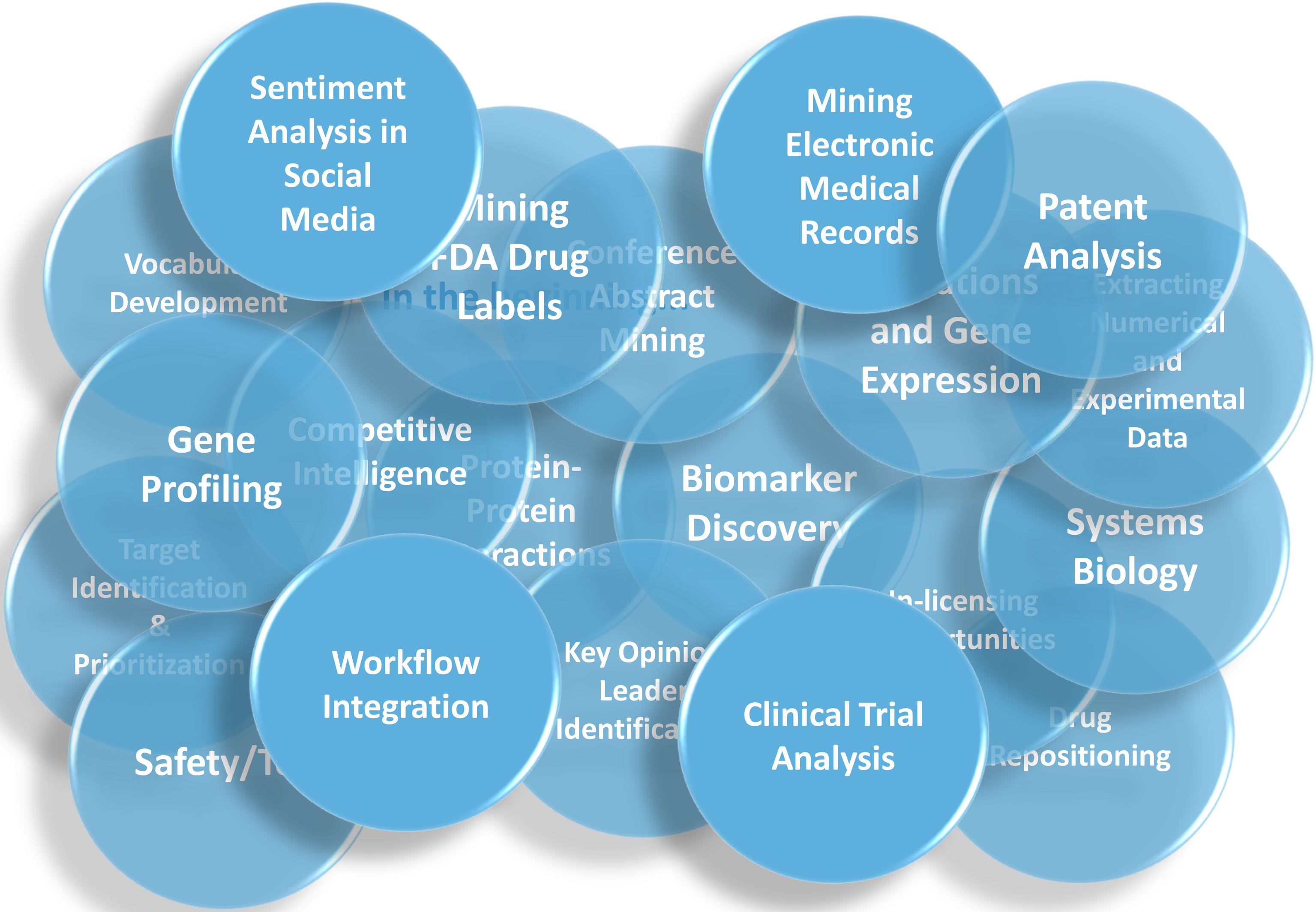
Synthesizing Knowledge by Exploiting Diverse Data Sources: From Microblogs to Patents

David Milward

ICIC, October 2011

Linguamatics

Range Of Application Areas



Sentiment Analysis in Social Media

Mining FDA Drug Labels

Conference Abstract Mining

Mining Electronic Medical Records

Patent Analysis

Vocabulary Development

tions and Gene Expression

Extracting Numerical and Experimental Data

Gene Profiling

Competitive Intelligence

Protein-Protein Interactions

Biomarker Discovery

Systems Biology

Target Identification & Prioritization

In-licensing Opportunities

Workflow Integration

Key Opinion Leader Identification

Clinical Trial Analysis

Drug Repositioning

Safety/T

Linguamatics

Example Data Sources

- External Data
  - Social Media
  - Scientific Literature
  - Conference Abstracts
  - Patents
  - Clinical Trials
- Internal Data
  - SharePoint repositories
  - Excel spreadsheets
  - Project reports
  - Electronic Health Records

Mining Social Media

To provide actionable information from public sentiment

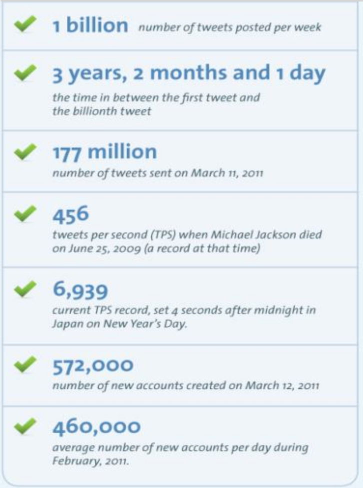
What is being said about our brands? How are the messages being relayed?

Who are the key opinion leaders?

Which sites are influencing our consumers?

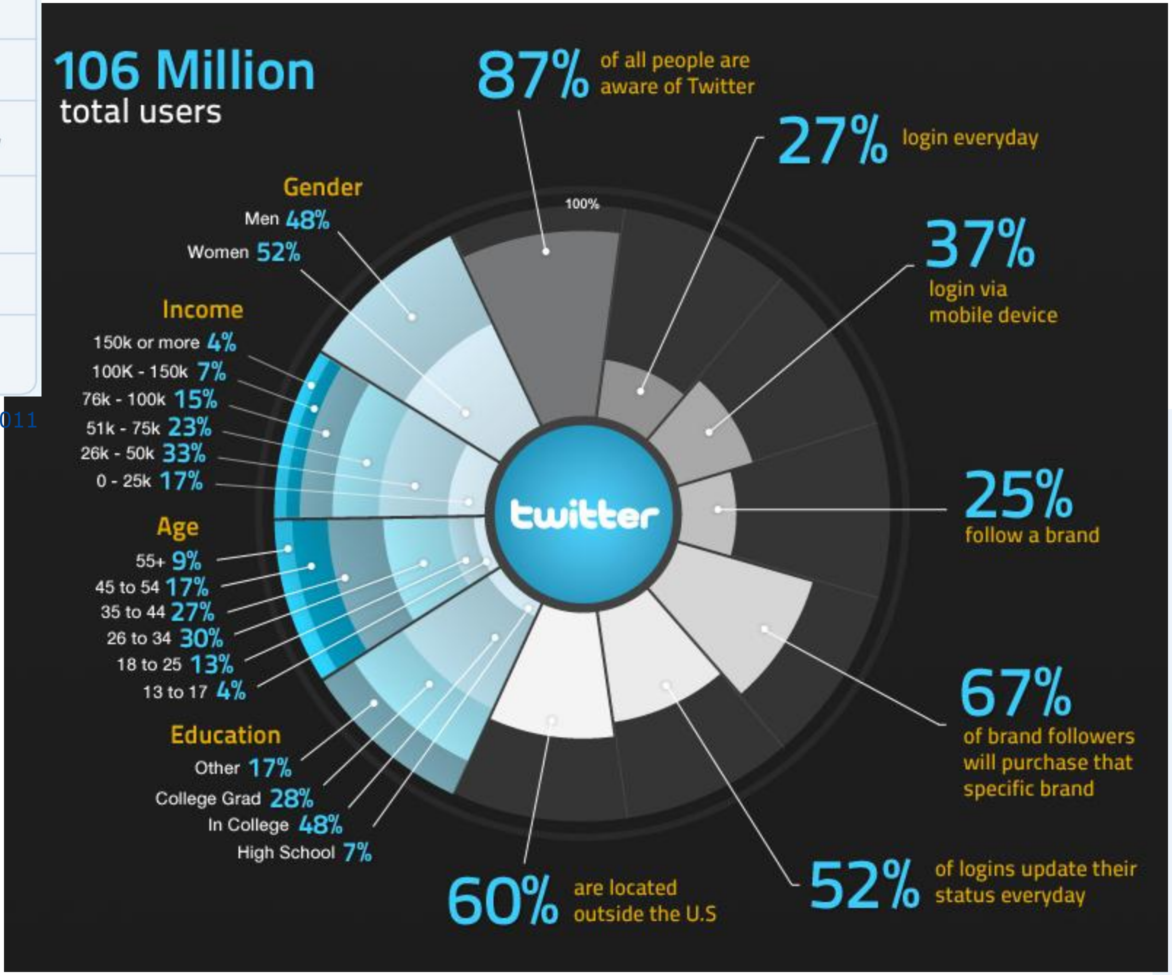
What are people saying about us and our competitors?

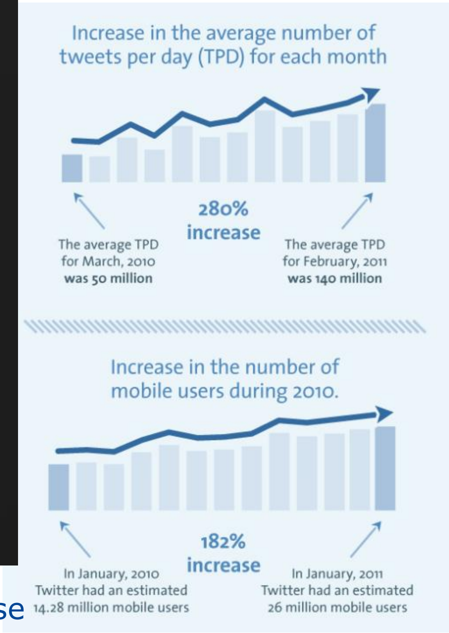
Microblogs: Twitter



- ✓ **1 billion** *number of tweets posted per week*
- ✓ **3 years, 2 months and 1 day** *the time in between the first tweet and the billionth tweet*
- ✓ **177 million** *number of tweets sent on March 11, 2011*
- ✓ **456** *tweets per second (TPS) when Michael Jackson died on June 25, 2009 (a record at that time)*
- ✓ **6,939** *current TPS record, set 4 seconds after midnight in Japan on New Year's Day.*
- ✓ **572,000** *number of new accounts created on March 12, 2011*
- ✓ **460,000** *average number of new accounts per day during February, 2011.*

* twitter.com -March 2011

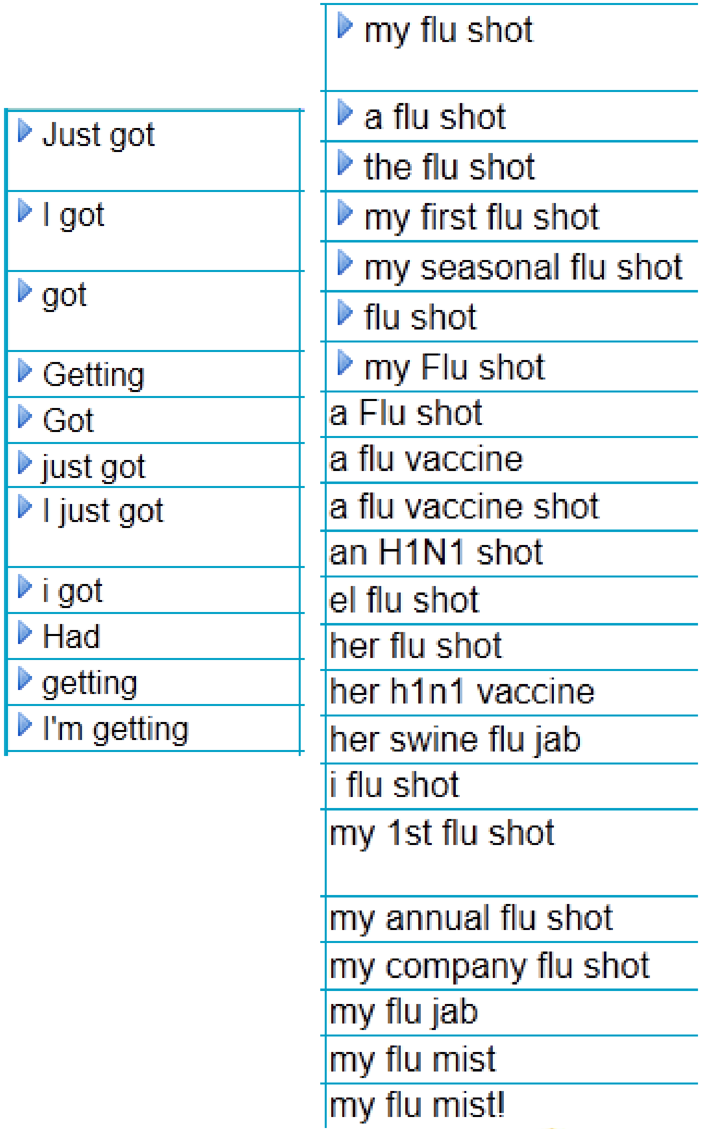




©Digital Surgeons 2010, All stats are based in U.S unless specified otherwise

Natural Language Processing (NLP)

- Find and extract patterns, not just keywords
- Capturing the 1000s of ways people say the same thing
- For example: how do people express that they are going to get a flu vaccine vs. not get one?

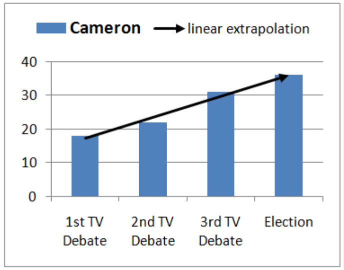


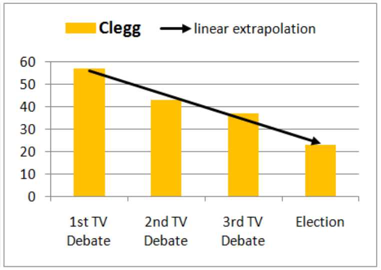
▸ Just got
▸ I got
▸ got
▸ Getting
▸ Got
▸ just got
▸ I just got
▸ i got
▸ Had
▸ getting
▸ I'm getting

▸ my flu shot
▸ a flu shot
▸ the flu shot
▸ my first flu shot
▸ my seasonal flu shot
▸ flu shot
▸ my Flu shot
a Flu shot
a flu vaccine
a flu vaccine shot
an H1N1 shot
el flu shot
her flu shot
her h1n1 vaccine
her swine flu jab
i flu shot
my 1st flu shot
my annual flu shot
my company flu shot
my flu jab
my flu mist
my flu mist!

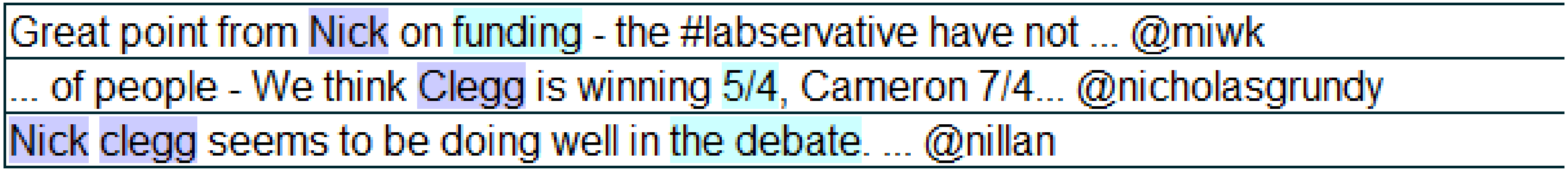
Refining Sentiment via Context: UK Elections

- Accurate prediction of UK Election Results based on sentiment trends





- What we counted as positive

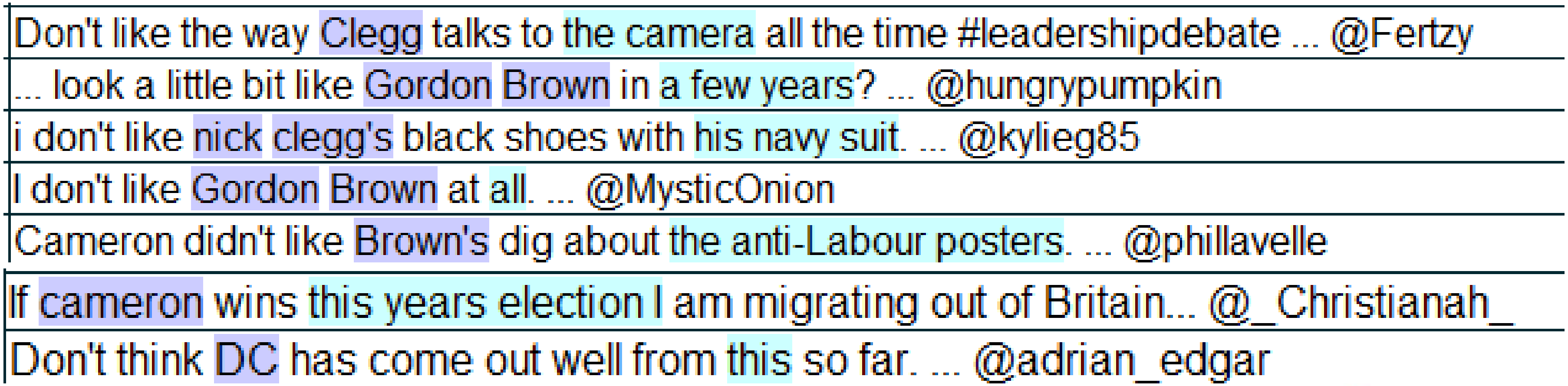


Great point from Nick on funding - the #labservative have not ... @miwk

... of people - We think Clegg is winning 5/4, Cameron 7/4... @nicholasgrundy

Nick clegg seems to be doing well in the debate. ... @nillan

- What we did not count as positive



Don't like the way Clegg talks to the camera all the time #leadershipdebate ... @Fertzy

... look a little bit like Gordon Brown in a few years? ... @hungrypumpkin

i don't like nick clegg's black shoes with his navy suit. ... @kylieg85

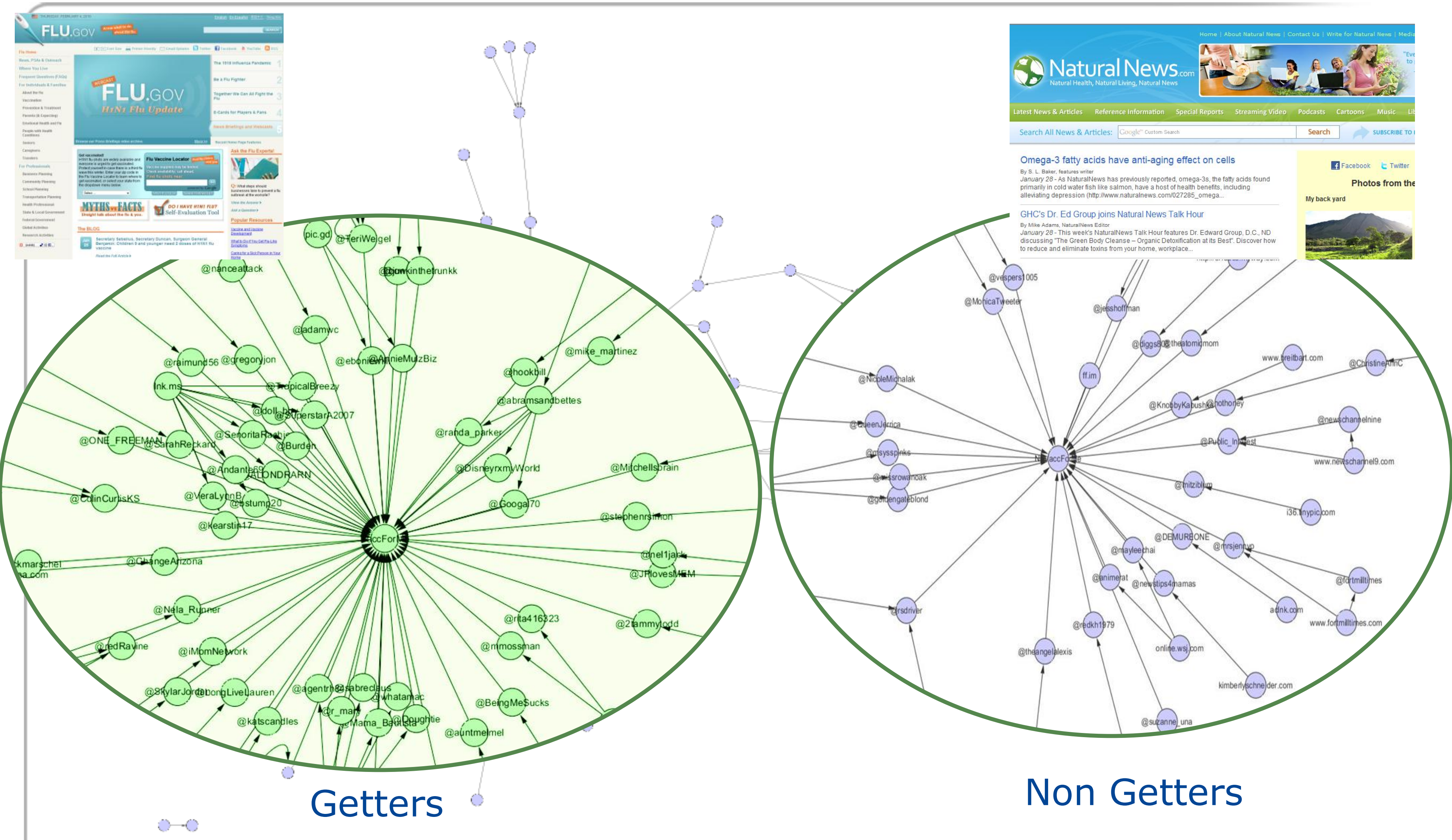
I don't like Gordon Brown at all. ... @MysticOnion

Cameron didn't like Brown's dig about the anti-Labour posters. ... @phillavelle

If cameron wins this years election I am migrating out of Britain... @_Christianah_

Don't think DC has come out well from this so far. ... @adrian_edgar

Who will get a vaccine, and finding their influences



FLU.GOV Know what to do about the flu

Flu Home
News, PSAs & Outreach
Where You Live
Frequent Questions (FAQs)
For Individuals & Families

WEBCAST FLU.GOV H1N1 Flu Update

The 1918 Influenza Pandemic 1
Be a Flu Fighter 2
Together We Can All Fight the Flu 3
E-Cards for Players & Fans 4
News Briefings and Webcasts 5

Flu Vaccine Locator

MYTHS vs FACTS Straight talk about the flu & you.

DO I HAVE H1N1 FLU? Self-Evaluation Tool

Ask the Flu Experts!

Popular Resources

The BLOG

Secretary Sebelius, Secretary Duncan, Surgeon General Benjamin: Children 9 and younger need 2 doses of H1N1 flu vaccine

Read the Full Article

Home | About Natural News | Contact Us | Write for Natural News | Media

NaturalNews.com Natural Health, Natural Living, Natural News

Latest News & Articles | Reference Information | Special Reports | Streaming Video | Podcasts | Cartoons | Music

Search All News & Articles: Google Custom Search Search SUBSCRIBE TO

Omega-3 fatty acids have anti-aging effect on cells
By S. L. Baker, features writer
January 28 - As NaturalNews has previously reported, omega-3s, the fatty acids found primarily in cold water fish like salmon, have a host of health benefits, including alleviating depression (http://www.naturalnews.com/027285_omega...

GHC's Dr. Ed Group joins Natural News Talk Hour
By Mike Adams, NaturalNews Editor
January 28 - This week's NaturalNews Talk Hour features Dr. Edward Group, D.C., ND discussing "The Green Body Cleanse – Organic Detoxification at its Best". Discover how to reduce and eliminate toxins from your home, workplace...

Facebook Twitter

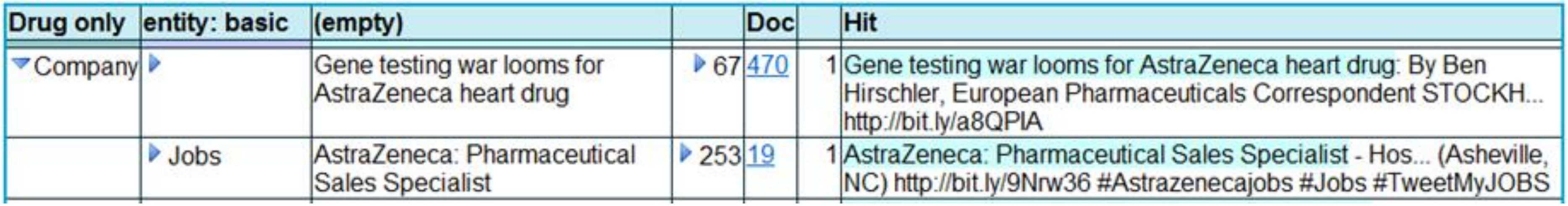
Photos from the

My back yard

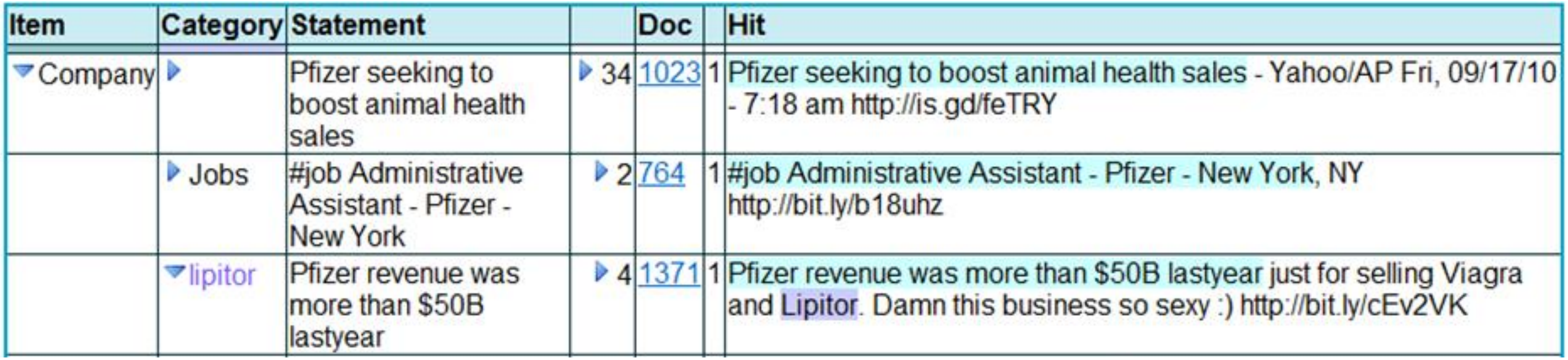
Getters

Non Getters

Categorizing by Topic and by Statement



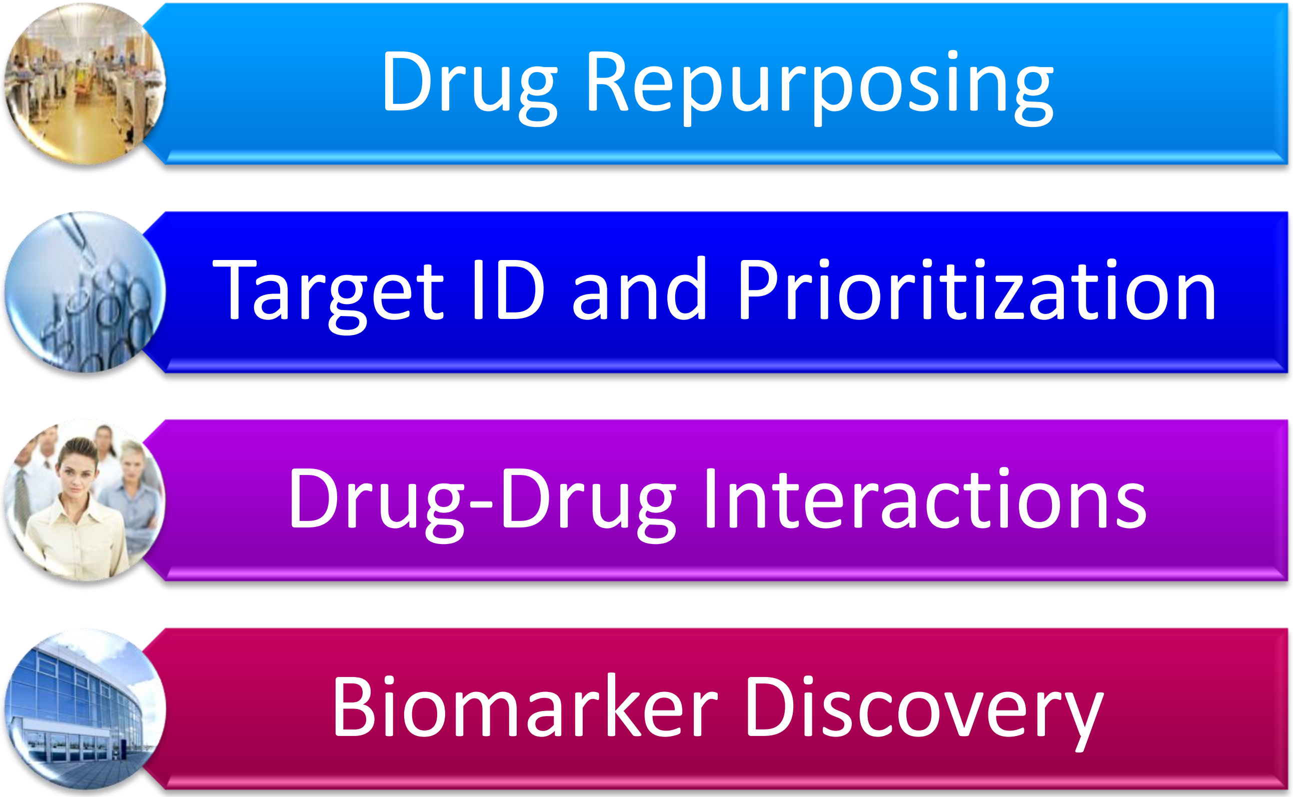
Drug only	entity: basic	(empty)		Doc		Hit
▼Company	▶	Gene testing war looms for AstraZeneca heart drug	▶ 67	470	1	Gene testing war looms for AstraZeneca heart drug: By Ben Hirschler, European Pharmaceuticals Correspondent STOCKH... http://bit.ly/a8QPIA
	▶ Jobs	AstraZeneca: Pharmaceutical Sales Specialist	▶ 253	19	1	AstraZeneca: Pharmaceutical Sales Specialist - Hos... (Asheville, NC) http://bit.ly/9Nrw36 #Astrazenecajobs #Jobs #TweetMyJOBS



Item	Category	Statement		Doc		Hit
▼Company	▶	Pfizer seeking to boost animal health sales	▶ 34	1023	1	Pfizer seeking to boost animal health sales - Yahoo/AP Fri, 09/17/10 - 7:18 am http://is.gd/feTRY
	▶ Jobs	#job Administrative Assistant - Pfizer - New York	▶ 2	764	1	#job Administrative Assistant - Pfizer - New York, NY http://bit.ly/b18uhz
	▼lipitor	Pfizer revenue was more than $50B lastyear	▶ 4	1371	1	Pfizer revenue was more than $50B lastyear just for selling Viagra and Lipitor. Damn this business so sexy :) http://bit.ly/cEv2VK

- Many Tweets or newsfeed statements are duplicates, but comparing character-by-character does not show this
- Here we are categorizing by topic e.g. jobs or drug, and using NLP to cluster by the messages to allow faster reviewing

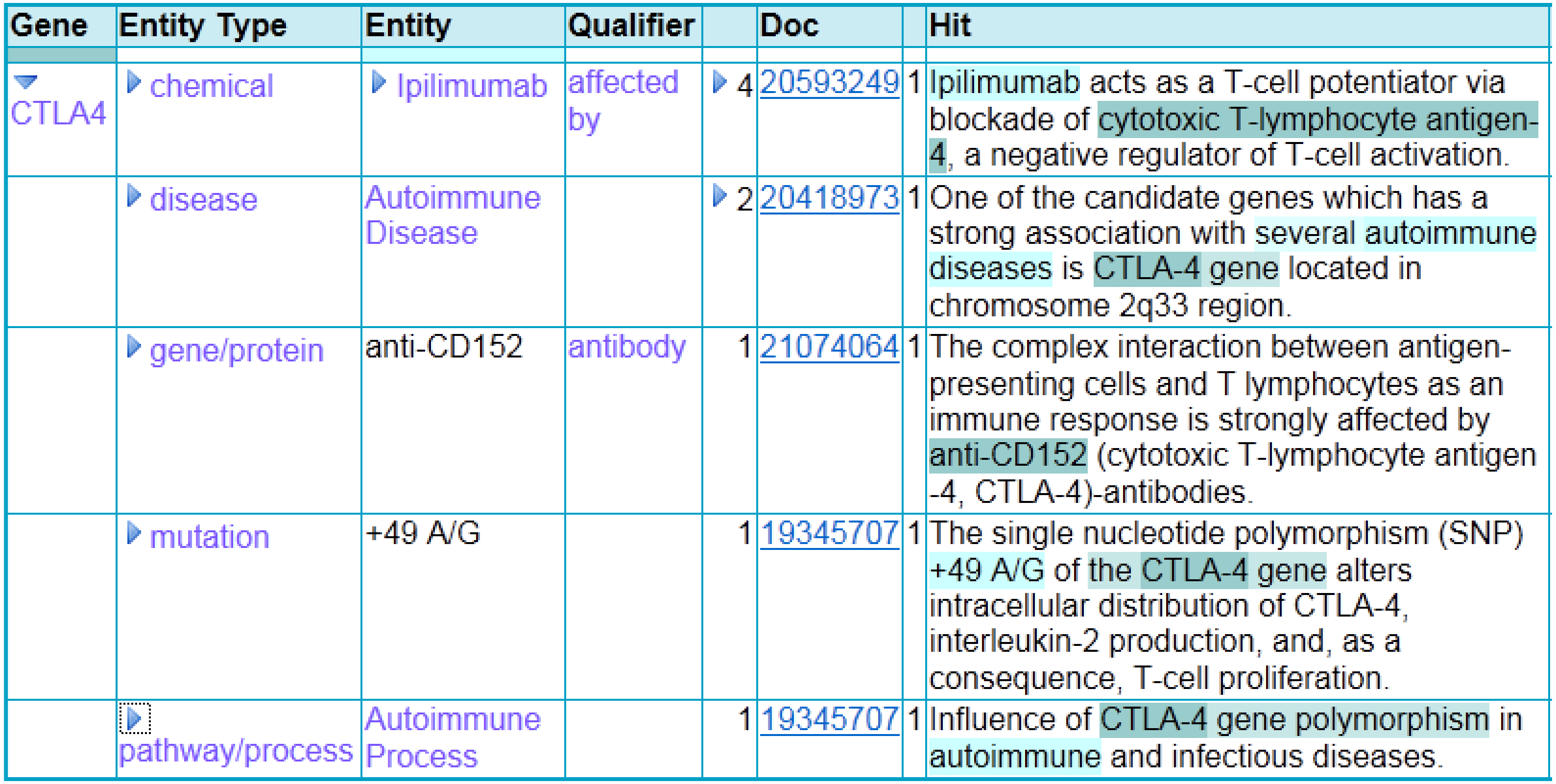
Mining MEDLINE



- Drug Repurposing
- Target ID and Prioritization
- Drug-Drug Interactions
- Biomarker Discovery

Gene Profile

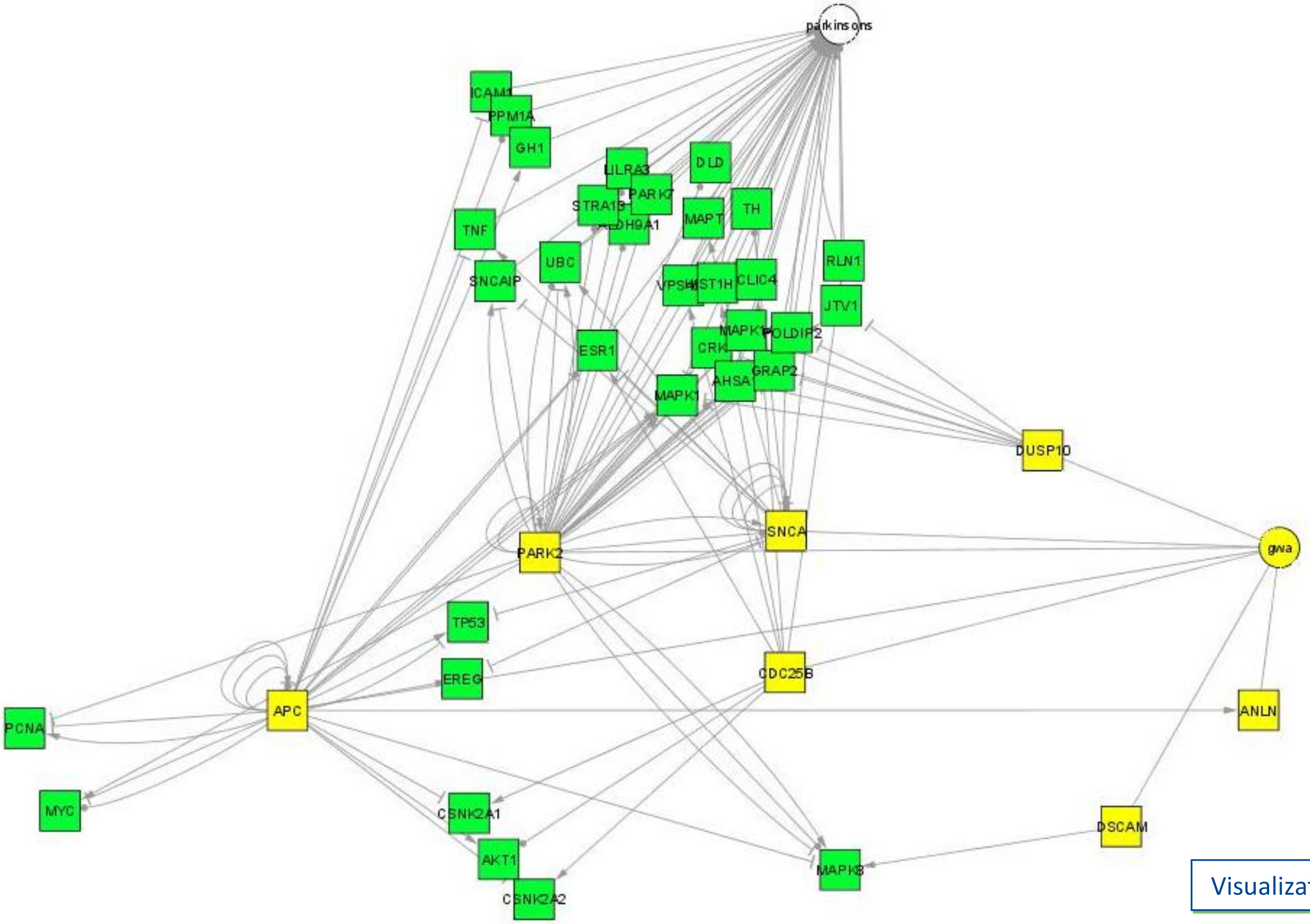
- Gathers together information relevant for a specific gene from multiple documents



Gene	Entity Type	Entity	Qualifier		Doc		Hit
▼ CTLA4	▶ chemical	▶ Ipilimumab	affected by	▶ 4	20593249	1	Ipilimumab acts as a T-cell potentiator via blockade of cytotoxic T-lymphocyte antigen-4, a negative regulator of T-cell activation.
	▶ disease	Autoimmune Disease		▶ 2	20418973	1	One of the candidate genes which has a strong association with several autoimmune diseases is CTLA-4 gene located in chromosome 2q33 region.
	▶ gene/protein	anti-CD152	antibody	1	21074064	1	The complex interaction between antigen-presenting cells and T lymphocytes as an immune response is strongly affected by anti-CD152 (cytotoxic T-lymphocyte antigen -4, CTLA-4)-antibodies.
	▶ mutation	+49 A/G		1	19345707	1	The single nucleotide polymorphism (SNP) +49 A/G of the CTLA-4 gene alters intracellular distribution of CTLA-4, interleukin-2 production, and, as a consequence, T-cell proliferation.
	▶ pathway/process	Autoimmune Process		1	19345707	1	Influence of CTLA-4 gene polymorphism in autoimmune and infectious diseases.

Linguamatics

Relating Genes to Disease



Visualization via Cytoscape

Mining Conference Proceedings

To provide actionable information about competitors or activities of potential acquisition targets

Identify industry leaders

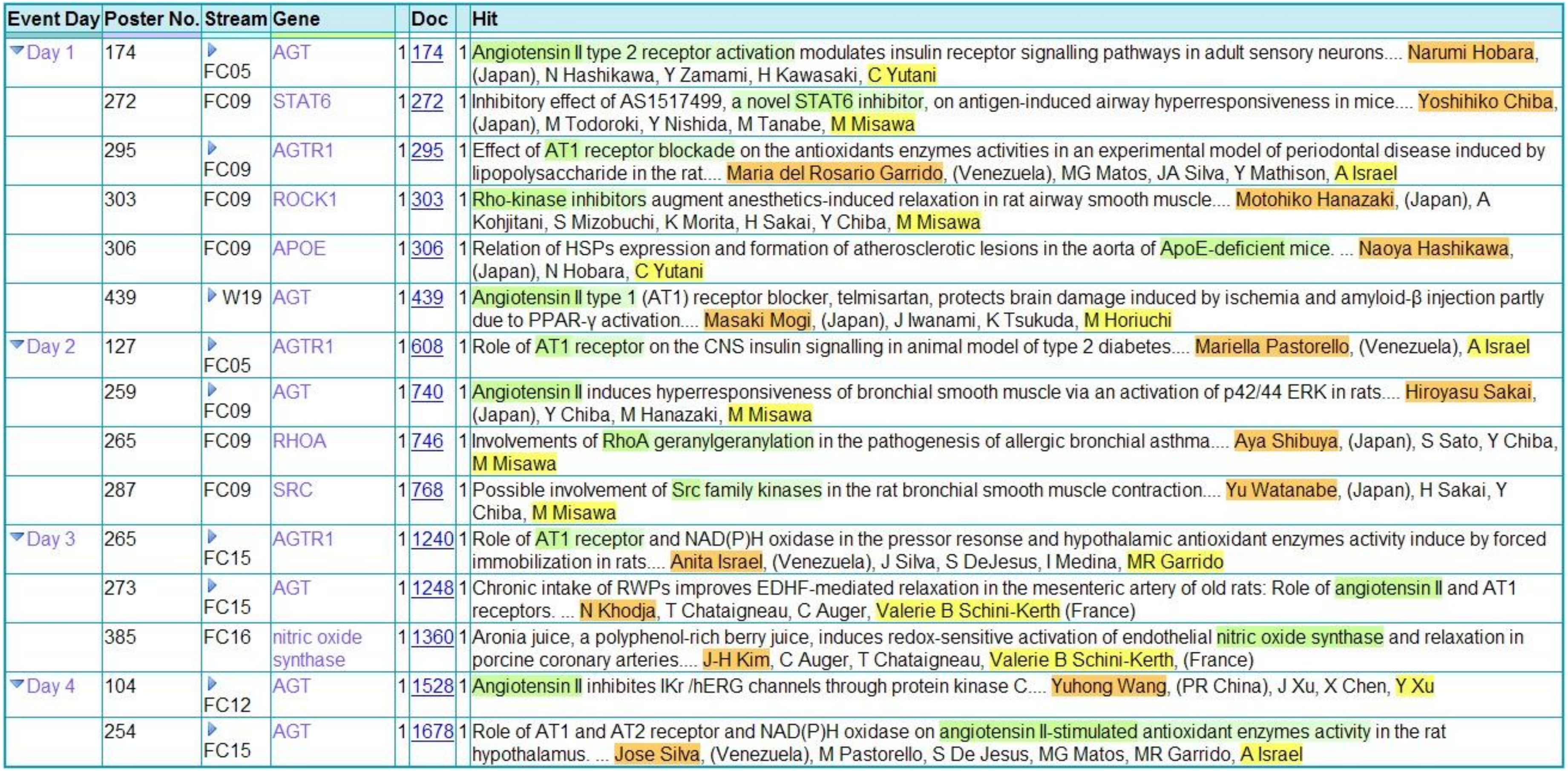
Gather information for internal customers

Identify new/novel areas of research

Find other companies / researchers working in the same area

Conference Proceedings

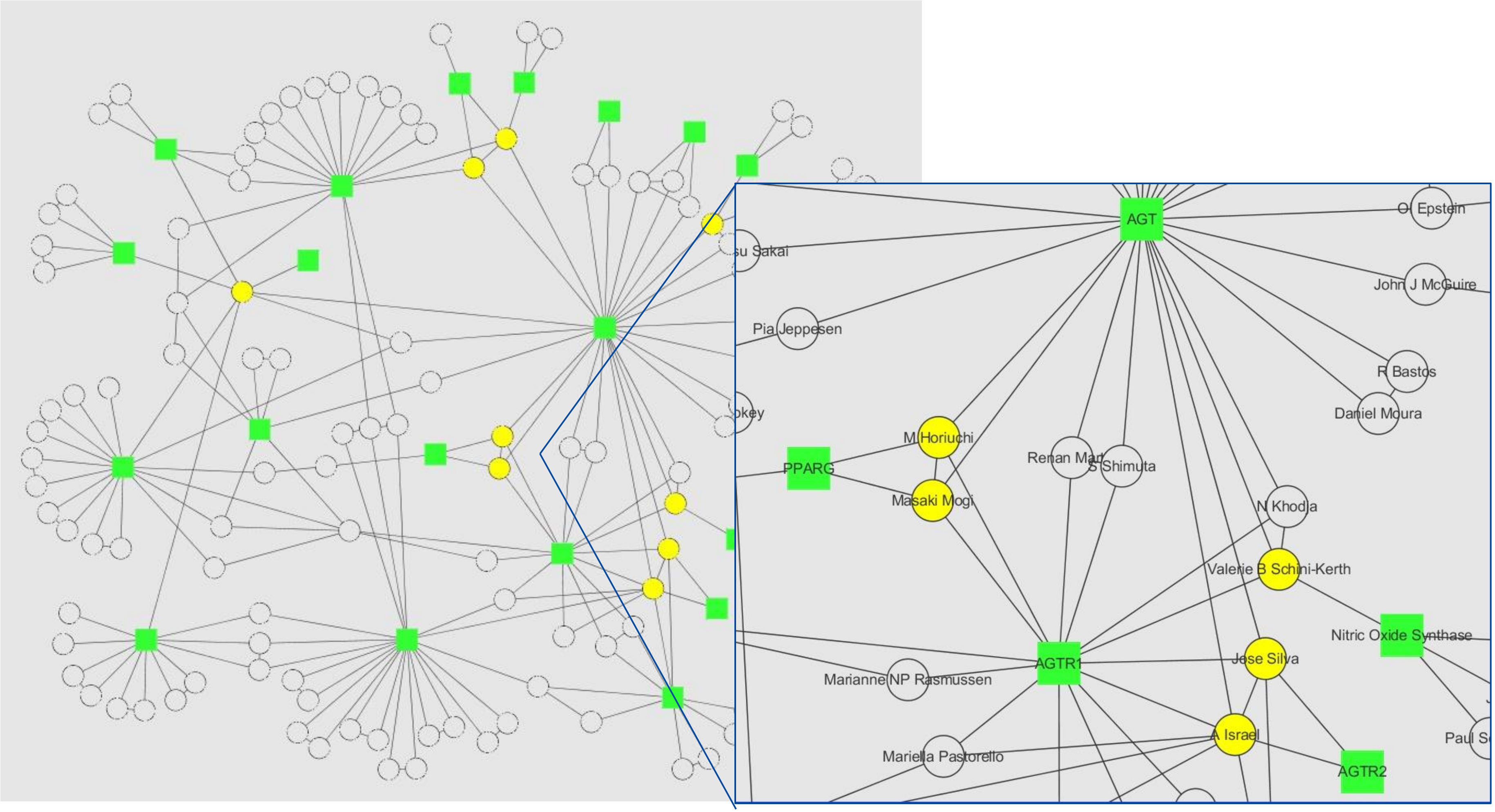
- Creating an itinerary for poster viewing



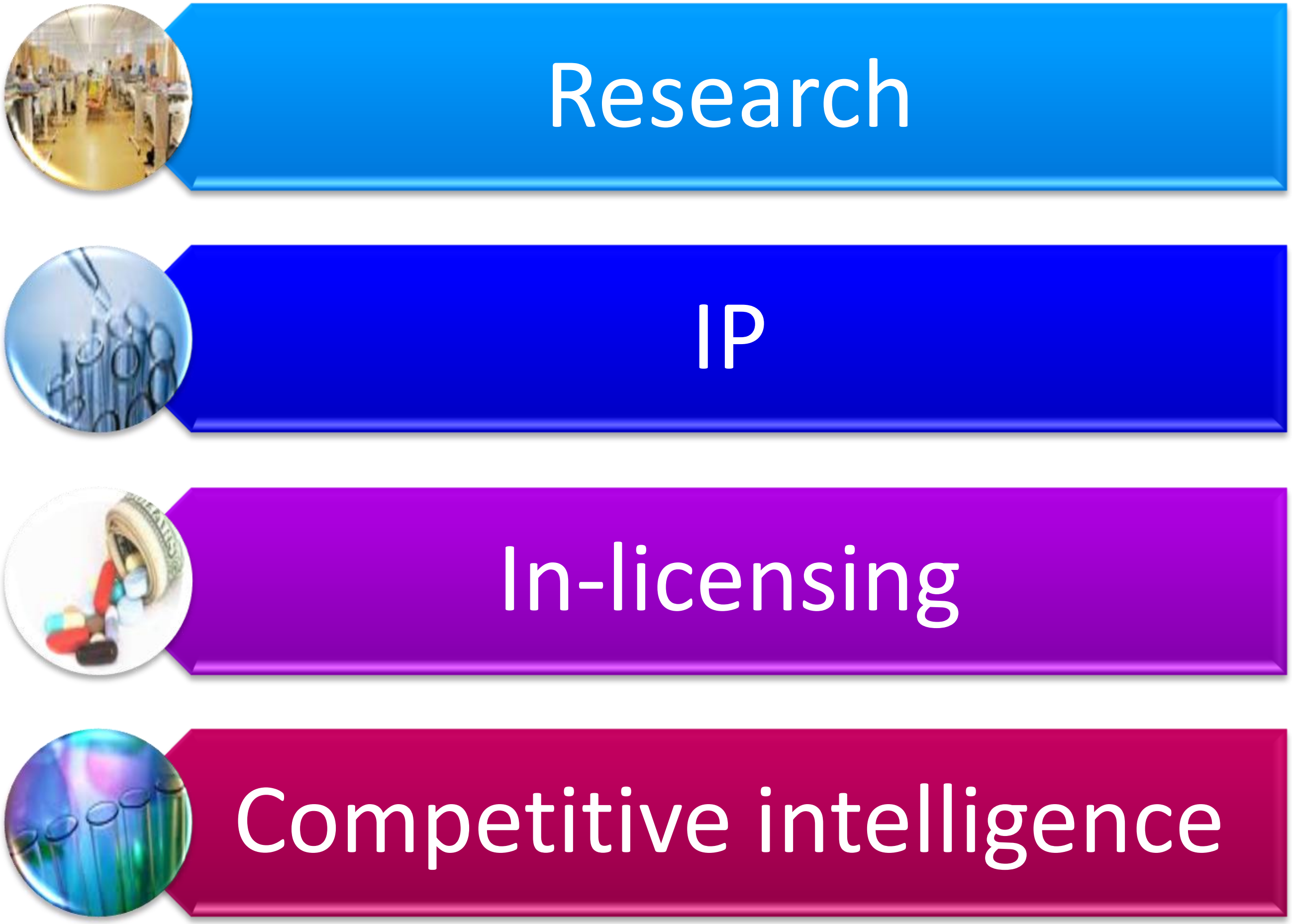
Event Day	Poster No.	Stream	Gene		Doc		Hit
▼Day 1	174	▶ FC05	AGT	1	174	1	Angiotensin II type 2 receptor activation modulates insulin receptor signalling pathways in adult sensory neurons.... Narumi Hobara, (Japan), N Hashikawa, Y Zamami, H Kawasaki, C Yutani
	272	FC09	STAT6	1	272	1	Inhibitory effect of AS1517499, a novel STAT6 inhibitor, on antigen-induced airway hyperresponsiveness in mice.... Yoshihiko Chiba, (Japan), M Todoroki, Y Nishida, M Tanabe, M Misawa
	295	▶ FC09	AGTR1	1	295	1	Effect of AT1 receptor blockade on the antioxidants enzymes activities in an experimental model of periodontal disease induced by lipopolysaccharide in the rat.... Maria del Rosario Garrido, (Venezuela), MG Matos, JA Silva, Y Mathison, A Israel
	303	FC09	ROCK1	1	303	1	Rho-kinase inhibitors augment anesthetics-induced relaxation in rat airway smooth muscle.... Motohiko Hanazaki, (Japan), A Kohjitani, S Mizobuchi, K Morita, H Sakai, Y Chiba, M Misawa
	306	FC09	APOE	1	306	1	Relation of HSPs expression and formation of atherosclerotic lesions in the aorta of ApoE-deficient mice. ... Naoya Hashikawa, (Japan), N Hobara, C Yutani
	439	▶ W19	AGT	1	439	1	Angiotensin II type 1 (AT1) receptor blocker, telmisartan, protects brain damage induced by ischemia and amyloid-β injection partly due to PPAR-γ activation.... Masaki Mogi, (Japan), J Iwanami, K Tsukuda, M Horiuchi
▼Day 2	127	▶ FC05	AGTR1	1	608	1	Role of AT1 receptor on the CNS insulin signalling in animal model of type 2 diabetes.... Mariella Pastorello, (Venezuela), A Israel
	259	▶ FC09	AGT	1	740	1	Angiotensin II induces hyperresponsiveness of bronchial smooth muscle via an activation of p42/44 ERK in rats.... Hiroyasu Sakai, (Japan), Y Chiba, M Hanazaki, M Misawa
	265	FC09	RHOA	1	746	1	Involvements of RhoA geranylgeranylation in the pathogenesis of allergic bronchial asthma.... Aya Shibuya, (Japan), S Sato, Y Chiba, M Misawa
	287	FC09	SRC	1	768	1	Possible involvement of Src family kinases in the rat bronchial smooth muscle contraction.... Yu Watanabe, (Japan), H Sakai, Y Chiba, M Misawa
▼Day 3	265	▶ FC15	AGTR1	1	1240	1	Role of AT1 receptor and NAD(P)H oxidase in the pressor resonse and hypothalamic antioxidant enzymes activity induce by forced immobilization in rats.... Anita Israel, (Venezuela), J Silva, S DeJesus, I Medina, MR Garrido
	273	▶ FC15	AGT	1	1248	1	Chronic intake of RWPs improves EDHF-mediated relaxation in the mesenteric artery of old rats: Role of angiotensin II and AT1 receptors. ... N Khodja, T Chataigneau, C Auger, Valerie B Schini-Kerth (France)
	385	FC16	nitric oxide synthase	1	1360	1	Aronia juice, a polyphenol-rich berry juice, induces redox-sensitive activation of endothelial nitric oxide synthase and relaxation in porcine coronary arteries.... J-H Kim, C Auger, T Chataigneau, Valerie B Schini-Kerth, (France)
▼Day 4	104	▶ FC12	AGT	1	1528	1	Angiotensin II inhibites IKr /hERG channels through protein kinase C.... Yuhong Wang, (PR China), J Xu, X Chen, Y Xu
	254	▶ FC15	AGT	1	1678	1	Role of AT1 and AT2 receptor and NAD(P)H oxidase on angiotensin II-stimulated antioxidant enzymes activity in the rat hypothalamus. ... Jose Silva, (Venezuela), M Pastorello, S De Jesus, MG Matos, MR Garrido, A Israel

Linguamatics

Conference Proceedings – Presenter Network

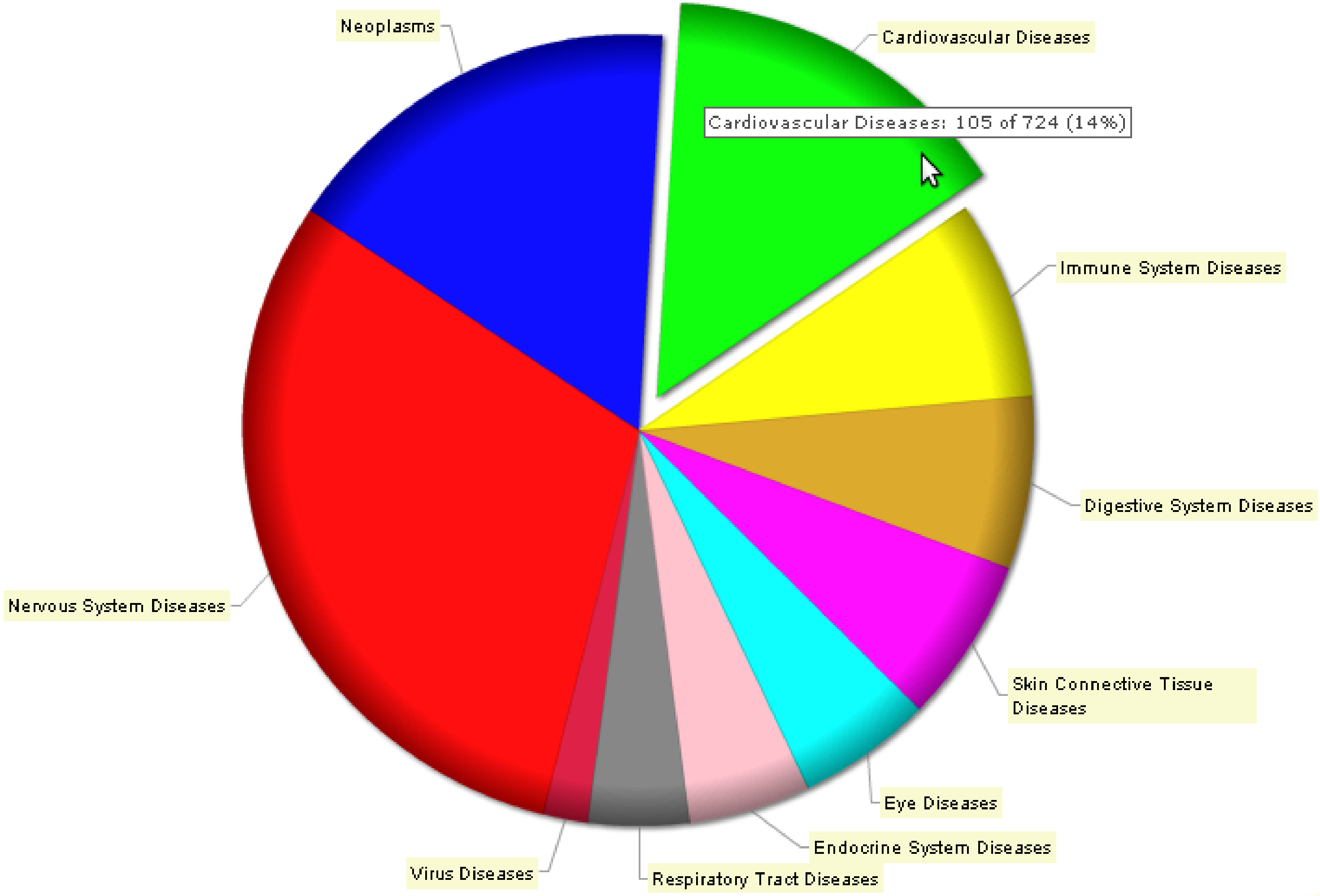


Mining Patent literature

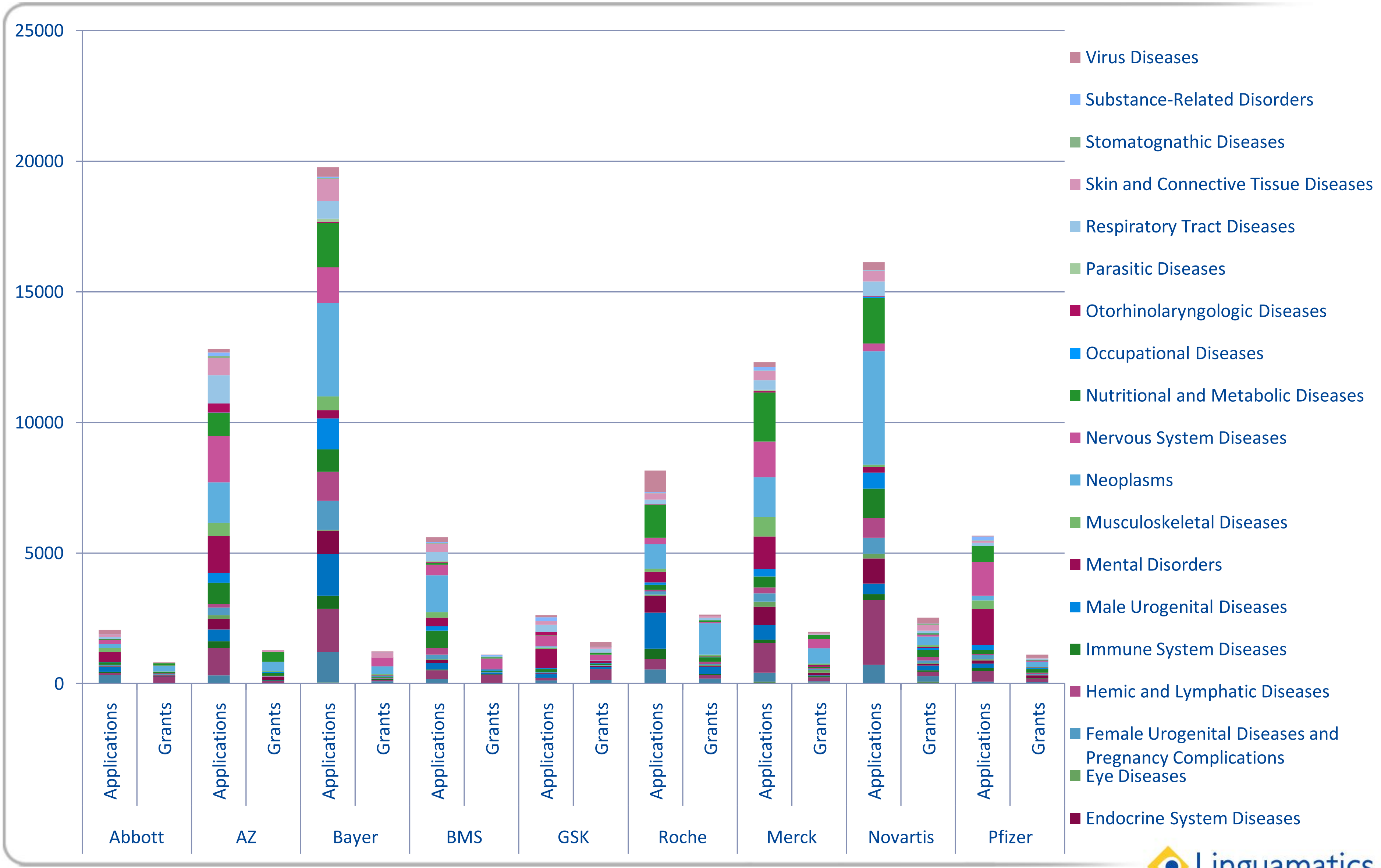


- Research
- IP
- In-licensing
- Competitive intelligence

Patent Grants Distribution by Disease (Sample)

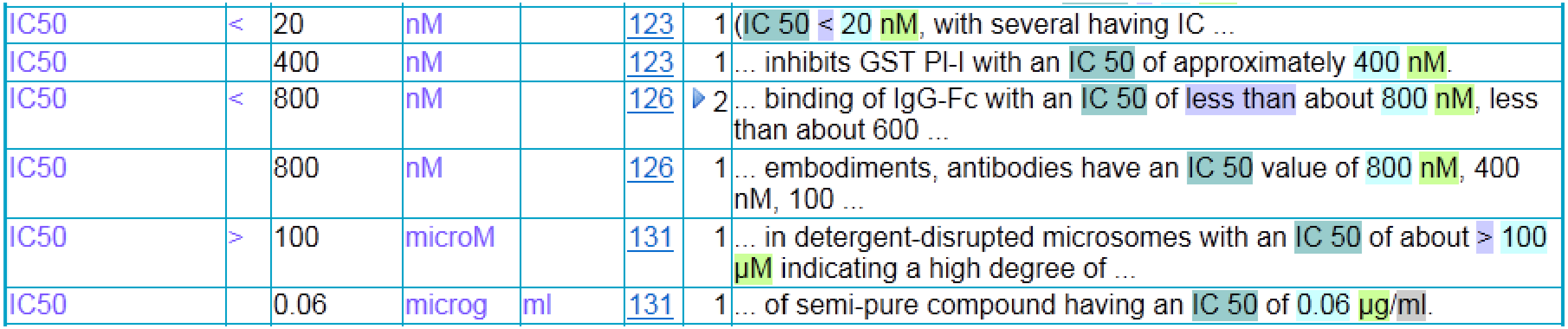


Patent Applications and Grants: Companies vs. Diseases

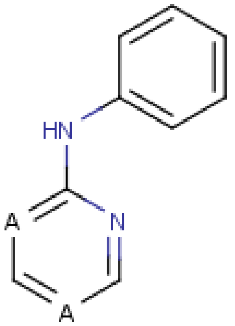


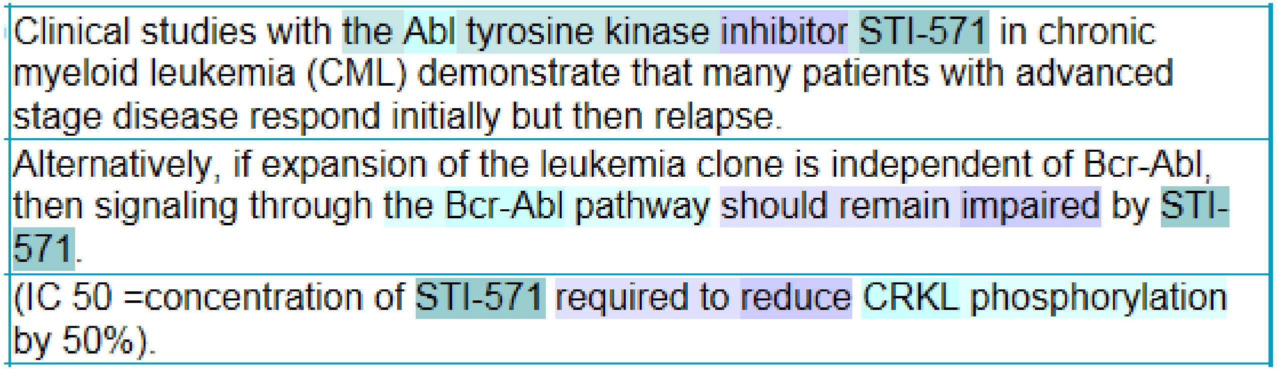
Linguamatics

Finding Inhibition Information in Patents



IC50	<	20	nM		123	1	(IC 50 < 20 nM, with several having IC ...
IC50		400	nM		123	1	... inhibits GST PI-I with an IC 50 of approximately 400 nM.
IC50	<	800	nM		126	▶ 2	... binding of IgG-Fc with an IC 50 of less than about 800 nM, less than about 600 ...
IC50		800	nM		126	1	... embodiments, antibodies have an IC 50 value of 800 nM, 400 nM, 100 ...
IC50	>	100	microM		131	1	... in detergent-disrupted microsomes with an IC 50 of about > 100 µM indicating a high degree of ...
IC50		0.06	microg	ml	131	1	... of semi-pure compound having an IC 50 of 0.06 µg/ml.





Clinical studies with the Abl tyrosine kinase inhibitor STI-571 in chronic myeloid leukemia (CML) demonstrate that many patients with advanced stage disease respond initially but then relapse.

Alternatively, if expansion of the leukemia clone is independent of Bcr-Abl, then signaling through the Bcr-Abl pathway should remain impaired by STI-571.

(IC 50 =concentration of STI-571 required to reduce CRKL phosphorylation by 50%).

Mining Clinical Trial Data

To provide actionable information about worldwide clinical development activities

Select trial sites more effectively

Find precedents for study design

Monitor progress of competitors' trials

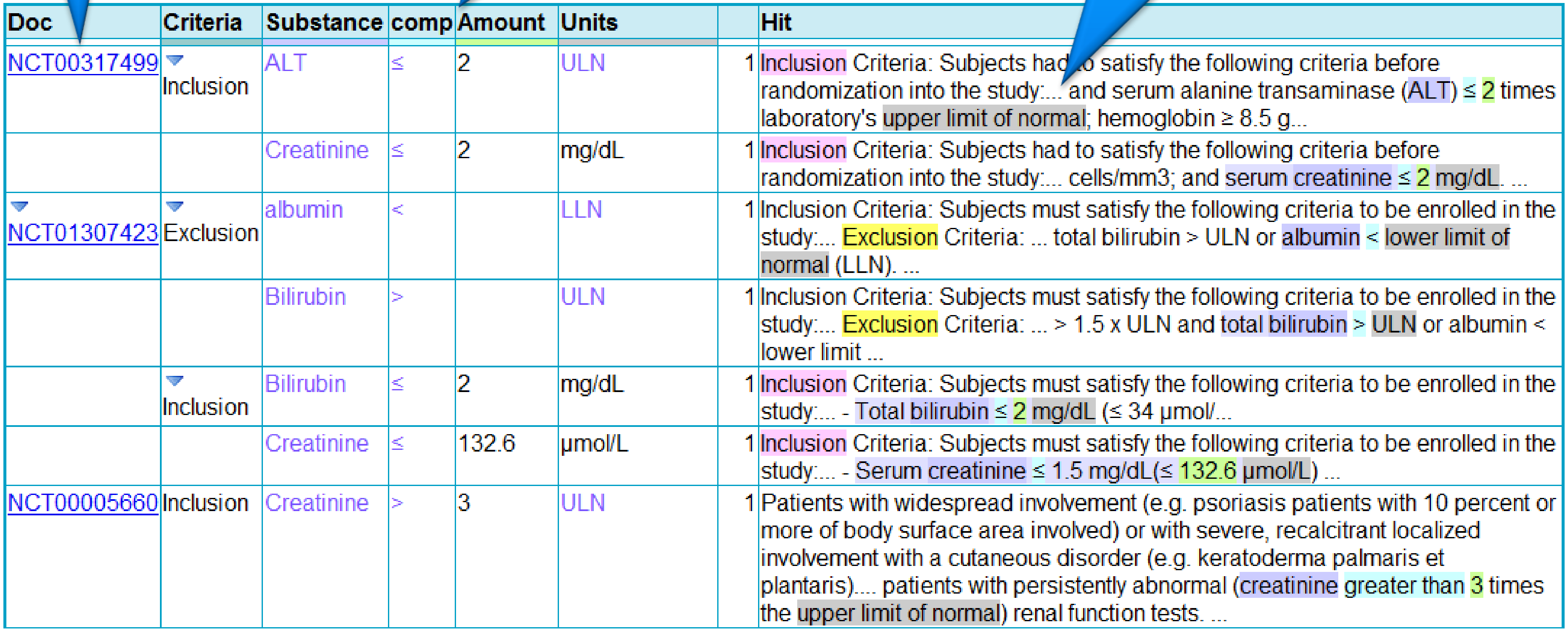
Find other companies running clinical trials in the same therapeutic area

Clinical Trial Analysis

NCT Number

Standardization of data

Hits in context



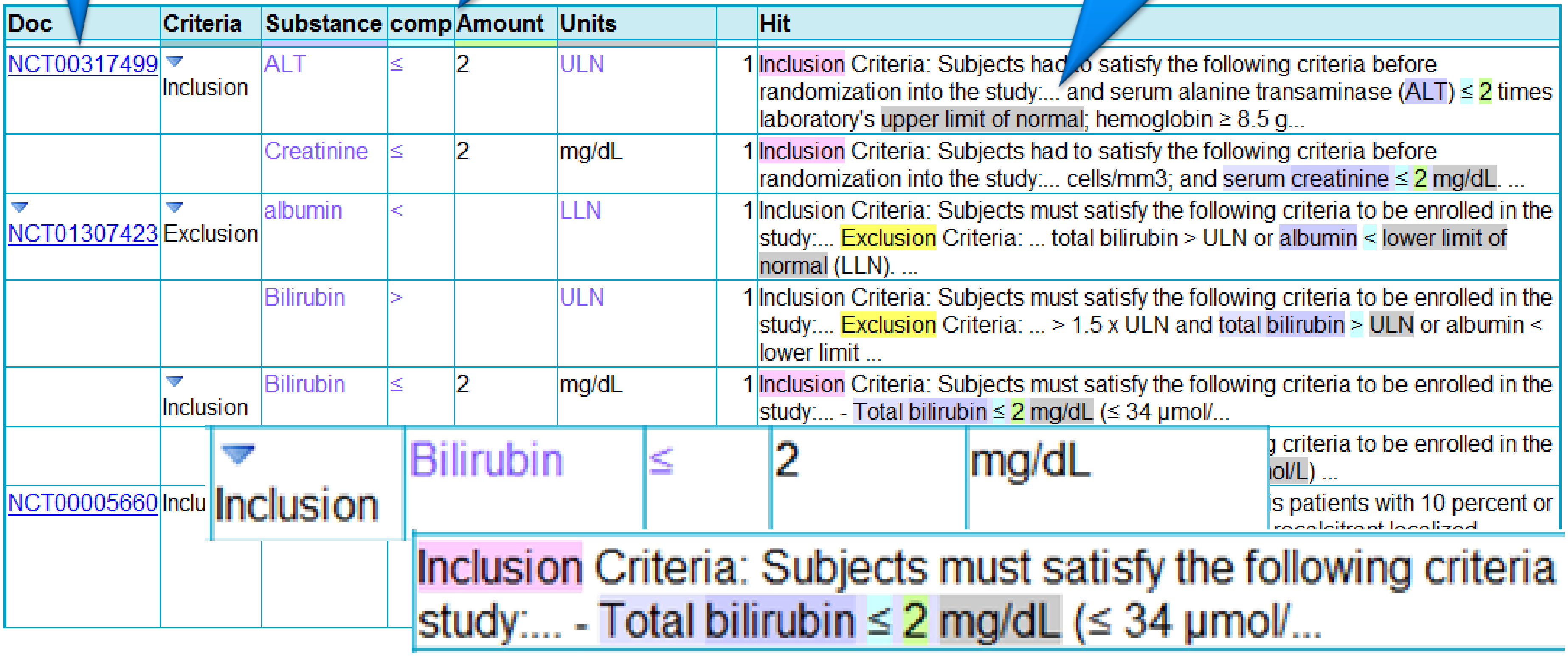
Doc	Criteria	Substance	comp	Amount	Units		Hit
NCT00317499	▼ Inclusion	ALT	≤	2	ULN	1	Inclusion Criteria: Subjects had to satisfy the following criteria before randomization into the study:... and serum alanine transaminase (ALT) ≤ 2 times laboratory's upper limit of normal; hemoglobin ≥ 8.5 g...
		Creatinine	≤	2	mg/dL	1	Inclusion Criteria: Subjects had to satisfy the following criteria before randomization into the study:... cells/mm3; and serum creatinine ≤ 2 mg/dL. ...
▼ NCT01307423	▼ Exclusion	albumin	<		LLN	1	Inclusion Criteria: Subjects must satisfy the following criteria to be enrolled in the study:... Exclusion Criteria: ... total bilirubin > ULN or albumin < lower limit of normal (LLN). ...
		Bilirubin	>		ULN	1	Inclusion Criteria: Subjects must satisfy the following criteria to be enrolled in the study:... Exclusion Criteria: ... > 1.5 x ULN and total bilirubin > ULN or albumin < lower limit ...
	▼ Inclusion	Bilirubin	≤	2	mg/dL	1	Inclusion Criteria: Subjects must satisfy the following criteria to be enrolled in the study:... - Total bilirubin ≤ 2 mg/dL (≤ 34 µmol/...
		Creatinine	≤	132.6	µmol/L	1	Inclusion Criteria: Subjects must satisfy the following criteria to be enrolled in the study:... - Serum creatinine ≤ 1.5 mg/dL(≤ 132.6 µmol/L) ...
NCT00005660	Inclusion	Creatinine	>	3	ULN	1	Patients with widespread involvement (e.g. psoriasis patients with 10 percent or more of body surface area involved) or with severe, recalcitrant localized involvement with a cutaneous disorder (e.g. keratoderma palmaris et plantaris).... patients with persistently abnormal (creatinine greater than 3 times the upper limit of normal) renal function tests. ...

Clinical Trial Analysis

NCT Number

Standardization of data

Hits in context

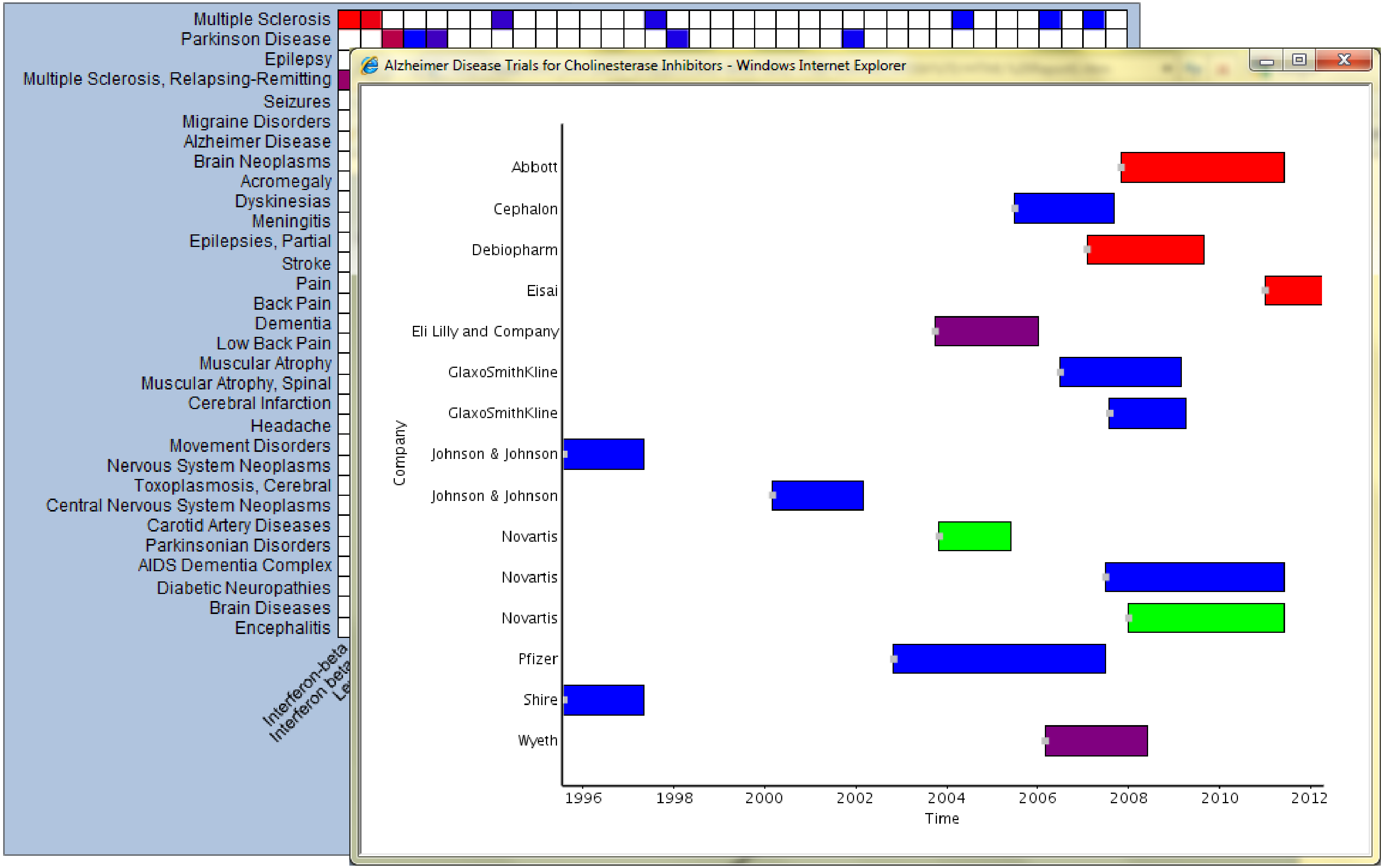


Doc	Criteria	Substance	comp	Amount	Units		Hit
NCT00317499	Inclusion	ALT	≤	2	ULN	1	Inclusion Criteria: Subjects had to satisfy the following criteria before randomization into the study:... and serum alanine transaminase (ALT) ≤ 2 times laboratory's upper limit of normal; hemoglobin ≥ 8.5 g...
		Creatinine	≤	2	mg/dL	1	Inclusion Criteria: Subjects had to satisfy the following criteria before randomization into the study:... cells/mm3; and serum creatinine ≤ 2 mg/dL. ...
NCT01307423	Exclusion	albumin	<		LLN	1	Inclusion Criteria: Subjects must satisfy the following criteria to be enrolled in the study:... Exclusion Criteria: ... total bilirubin > ULN or albumin < lower limit of normal (LLN). ...
		Bilirubin	>		ULN	1	Inclusion Criteria: Subjects must satisfy the following criteria to be enrolled in the study:... Exclusion Criteria: ... > 1.5 x ULN and total bilirubin > ULN or albumin < lower limit ...
	Inclusion	Bilirubin	≤	2	mg/dL	1	Inclusion Criteria: Subjects must satisfy the following criteria to be enrolled in the study:... - Total bilirubin ≤ 2 mg/dL (≤ 34 µmol/...
							g criteria to be enrolled in the ol/L) ...
NCT00005660	Inclu						s patients with 10 percent or recalcitrant localized

Inclusion	Bilirubin	≤	2	mg/dL

Inclusion Criteria: Subjects must satisfy the following criteria study:... - Total bilirubin ≤ 2 mg/dL (≤ 34 µmol/...

Workflow Integration + Clinical Trial Analysis



Mining Electronic Health Records

Diagnosis prediction from reports

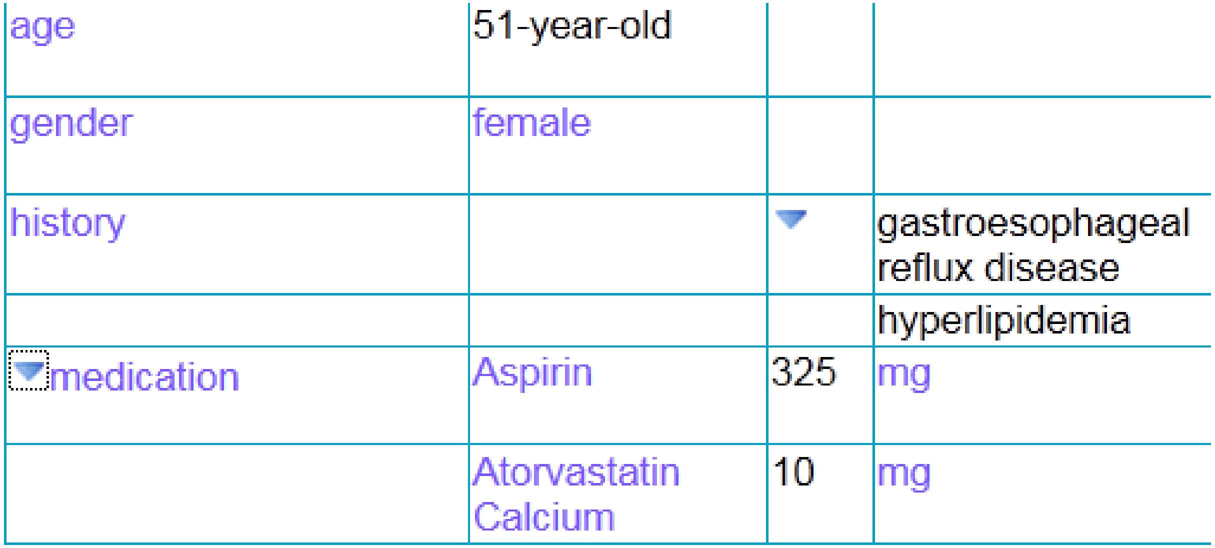
Translational medicine

Analysis of treatment outcomes

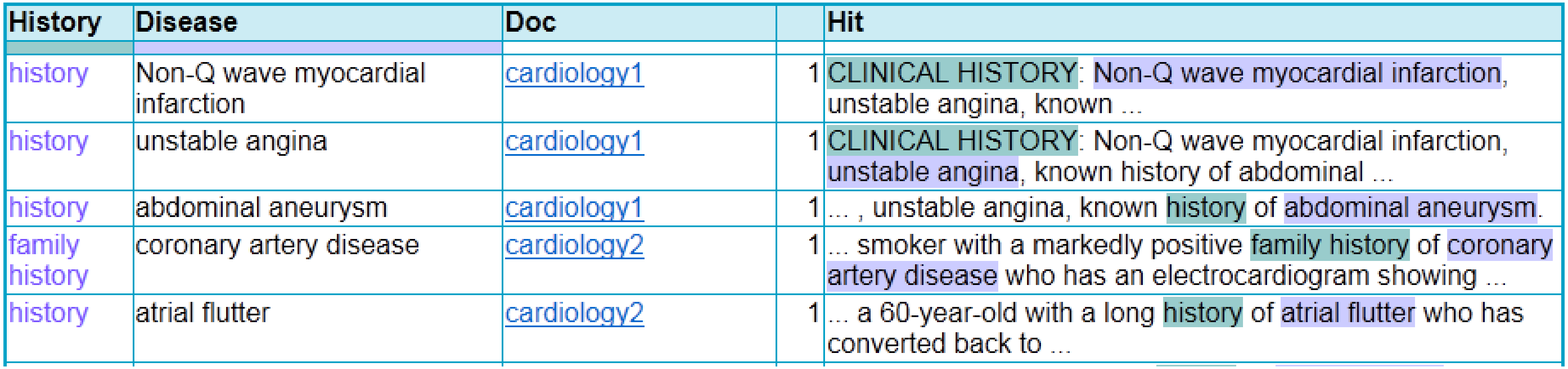
Monitoring of adherence to protocols

Electronic Health Records

- Identify patient information
- Need to consider context within the record, as well as the immediate sentence



age	51-year-old		
gender	female		
history		▼	gastroesophageal reflux disease
			hyperlipidemia
▼medication	Aspirin	325	mg
	Atorvastatin Calcium	10	mg



History	Disease	Doc		Hit
history	Non-Q wave myocardial infarction	cardiology1	1	CLINICAL HISTORY: Non-Q wave myocardial infarction, unstable angina, known ...
history	unstable angina	cardiology1	1	CLINICAL HISTORY: Non-Q wave myocardial infarction, unstable angina, known history of abdominal ...
history	abdominal aneurysm	cardiology1	1	... , unstable angina, known history of abdominal aneurysm.
family history	coronary artery disease	cardiology2	1	... smoker with a markedly positive family history of coronary artery disease who has an electrocardiogram showing ...
history	atrial flutter	cardiology2	1	... a 60-year-old with a long history of atrial flutter who has converted back to ...

Mining Multiple Sources

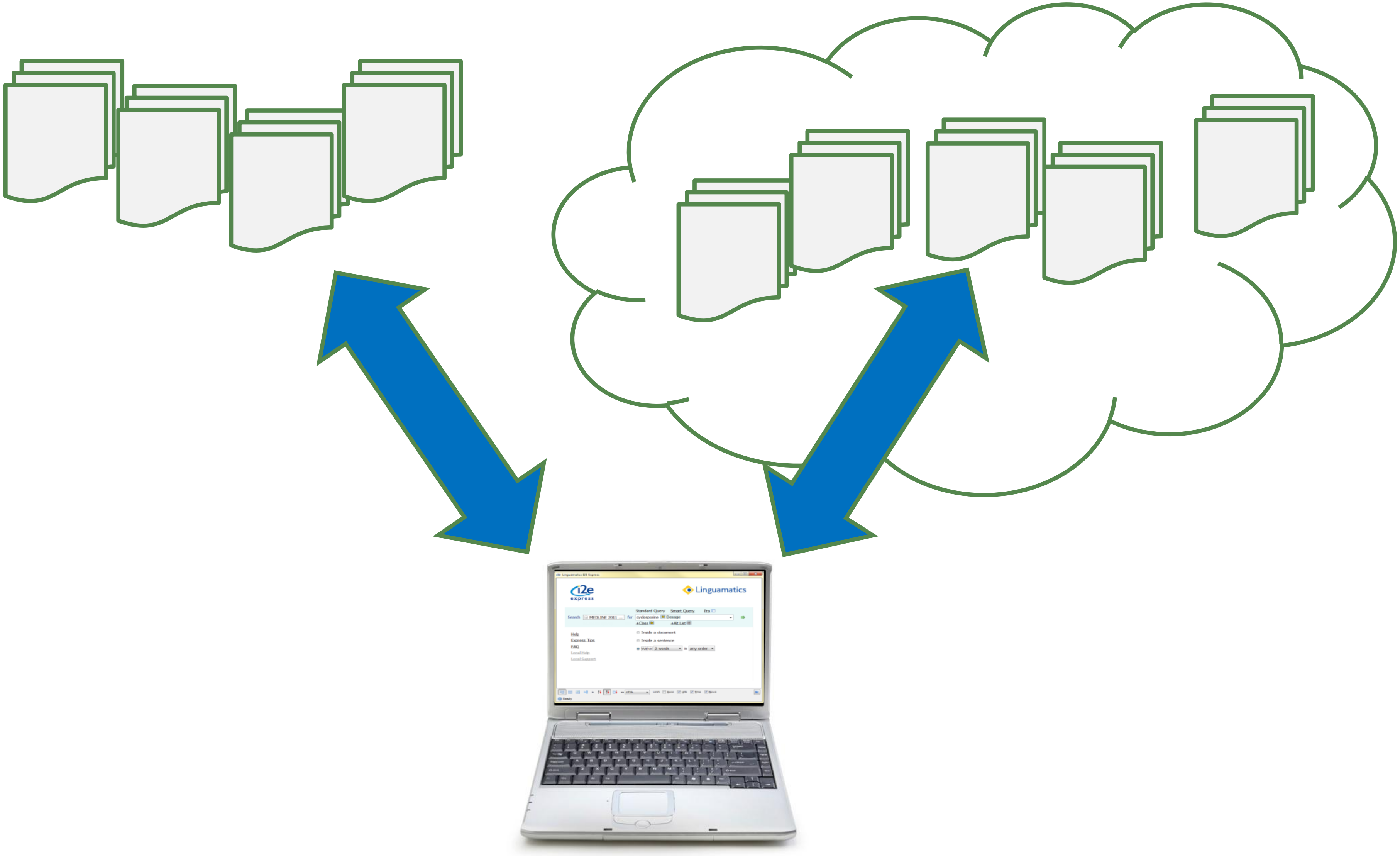
Merge information from all possible sources

Find more detailed information from a second source

Connect information across different structured and unstructured sources

Compare trends in different data sources

Delivering Diverse Data Sources



Linguamatics I2E Express

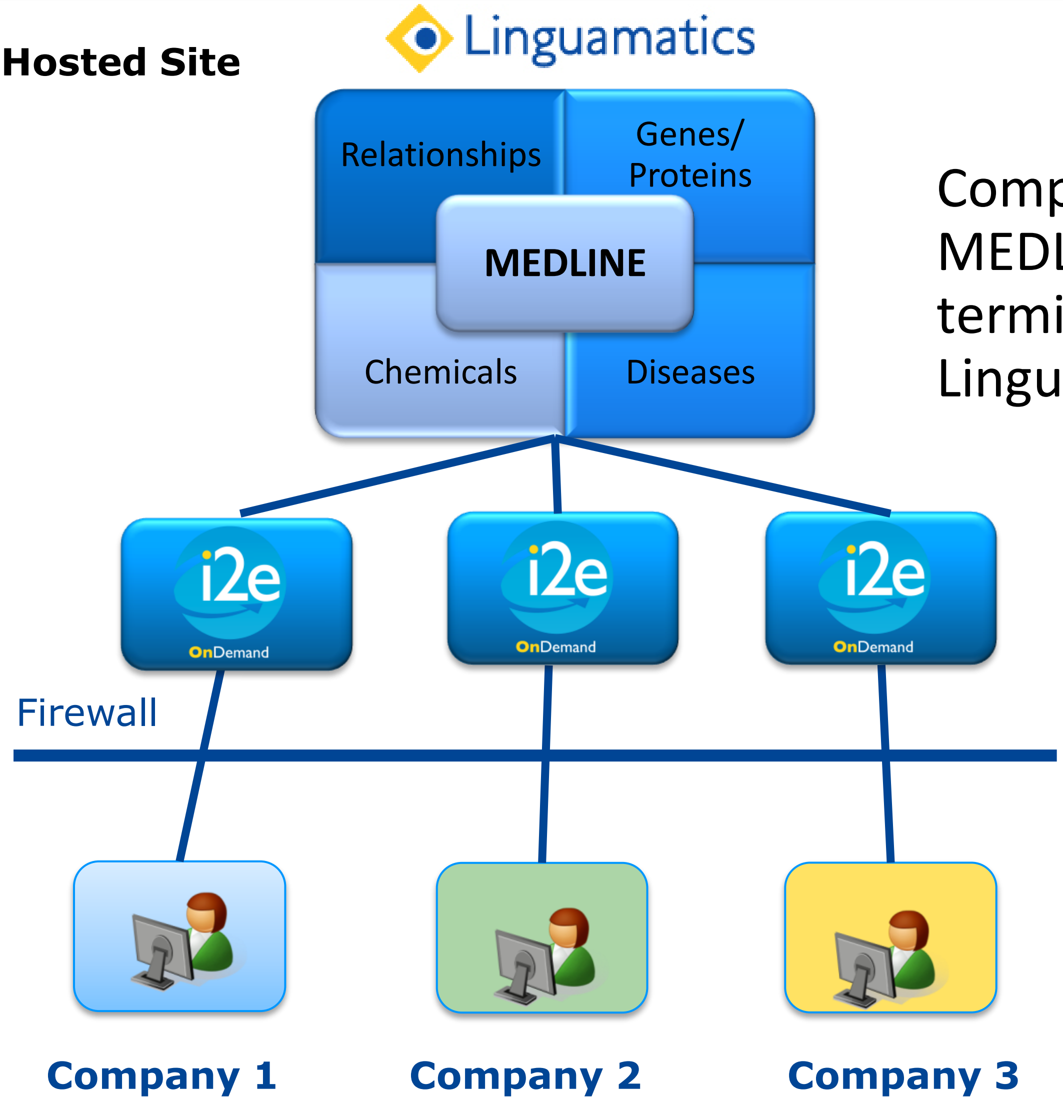
Search MEDLINE 2011 ... for cyclosporine Dosage

○ Inside a document

○ Inside a sentence

● Within 2 words in any order

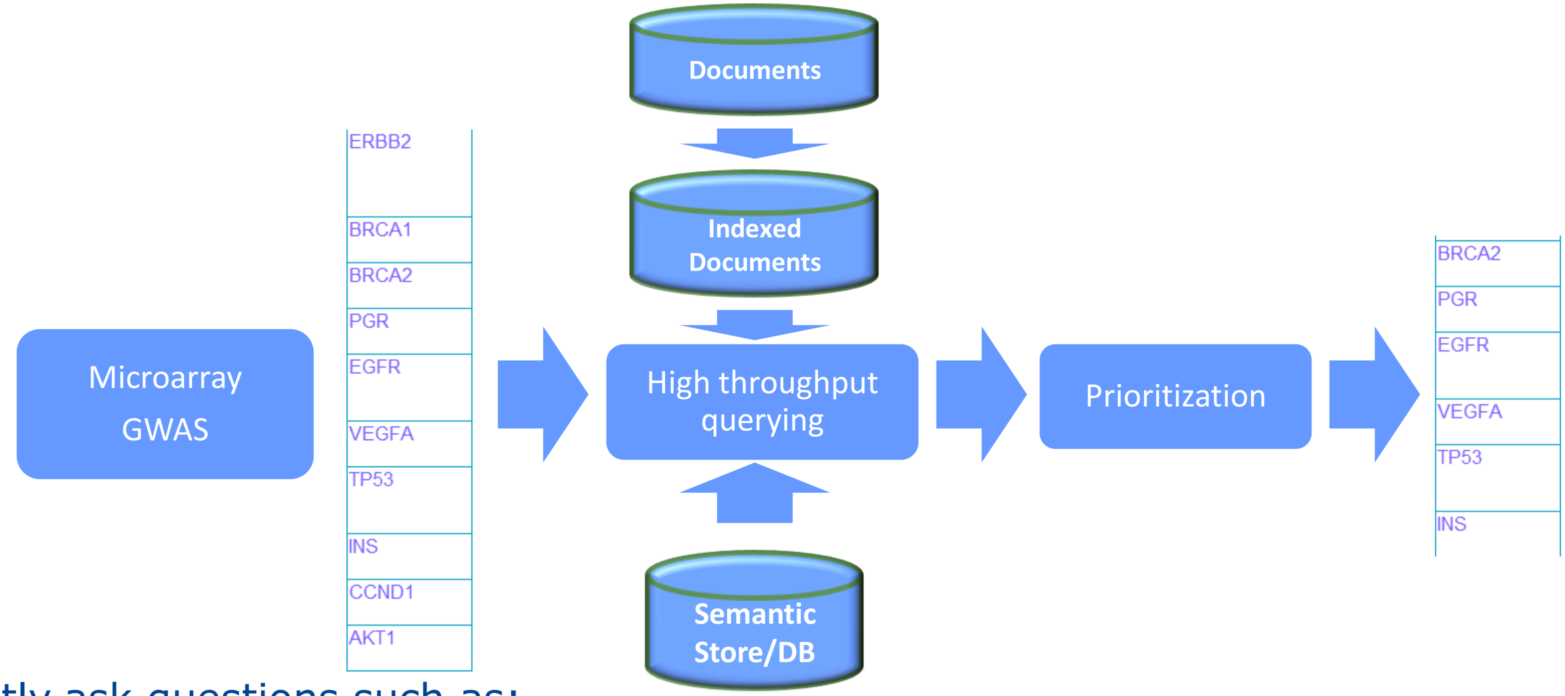
I2E OnDemand



Complete and up-to-date MEDLINE index with built in terminologies managed by Linguamatics

Structured and Unstructured Data Workflows

- Successful queries converted into regular workflows for up-to-date dashboards, alerting
- Structured resources (tabular/semantic web) augmented with information from literature



Directly ask questions such as:

- *Which genes from microarray experiment (or GWAS) show evidence of association with a disease*

Linguamatics

Conclusions

- Text Mining allows exploitation of diverse data sources e.g.
  - Scientific Literature
  - Patents
  - Social Media
- Wide range of applications, from R&D to Marketing
- Results from each source are valuable in themselves, but can also combine
  - Structured and unstructured data
  - Data from different unstructured sources