BI and the "Unstructured Data" Challenge

Seth Grimes

Alta Plana Corporation

301-270-0795 -- http://altaplana.com

The Data Warehousing Institute

Washington DC chapter

May 9, 2008

Alta Plana

Introduction

Seth Grimes –

Principal Consultant with Alta Plana Corporation.

Contributing Editor, *IntelligentEnterprise.com*.

Channel Expert, *B-Eye-Network.com*.

Founding Chair, Text Analytics Summit (Boston, June 16-17).

TDWI Instructor – T.A. Course (San Diego, August).

Disclaimer: *I am not paid to promote any vendor.*

Key Message -- #1

If you are not analyzing text, you're missing opportunity...

360° views

Single version of the truth

or running unacceptable risk...

Industries such as travel and hospitality and retail live and die on customer experience. – *Clarabridge CEO Sid Banerjee*

** in many applications/businesses but not all.

This is the "Unstructured Data" challenge

Key Message -- #2

Text analytics can add lift to your BI initiatives...

Organizations embracing text analytics all report having an epiphany moment when they suddenly knew more than before." – *Philip Russom, the Data Warehousing Institute*

And it can do a lot more.

Text Analytics is an answer to the "Unstructured Data" challenge

Key Message -- #3

You may need to expand your view of what BI is about.

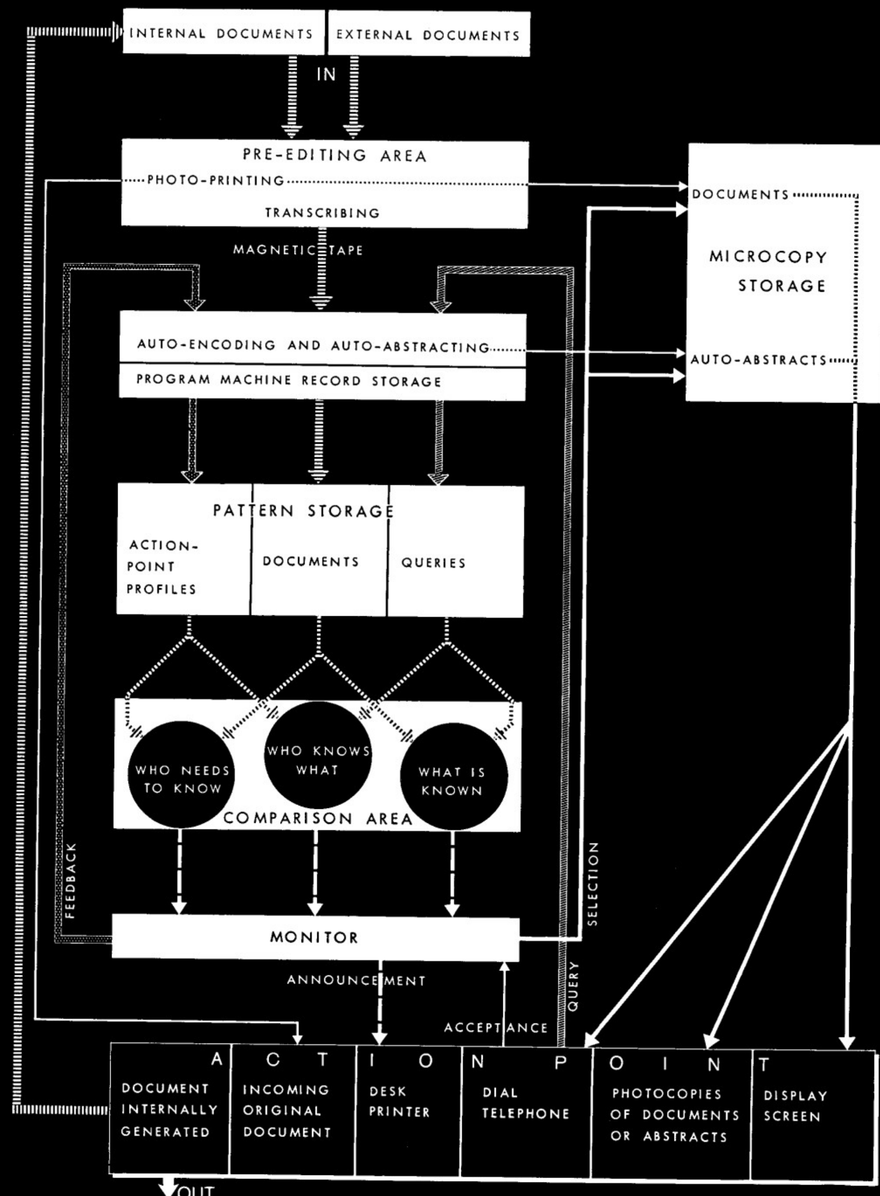


Figure 1 **A Business Intelligence System**

Key Message -- #3

In this paper, business is a collection of activities carried on for whatever purpose, be it science, technology, commerce, industry, law, government, defense, et cetera. The communication facility serving the conduct of a business (in the broad sense) may be referred to as an intelligence system. The notion of intelligence is also defined here, in a more general sense, as "the ability to apprehend the interrelationships of presented facts in such a way as to guide action towards a desired goal."

– *Hans Peter Luhn,* A Business Intelligence System, *IBM Journal, October 1958*

The "Unstructured Data" Challenge

"The bulk of information value is perceived as coming from data in relational tables. The reason is that data that is structured is easy to mine and analyze."

– *Prabhakar Raghavan, Yahoo Research, former CTO of enterprise-search vendor Verity (now part of Autonomy)*

Yet 80% of enterprise information is in "unstructured" form (IDC, others). The value equation is out of balance: it reflects actuality rather than potential.

The "Unstructured Data" Challenge

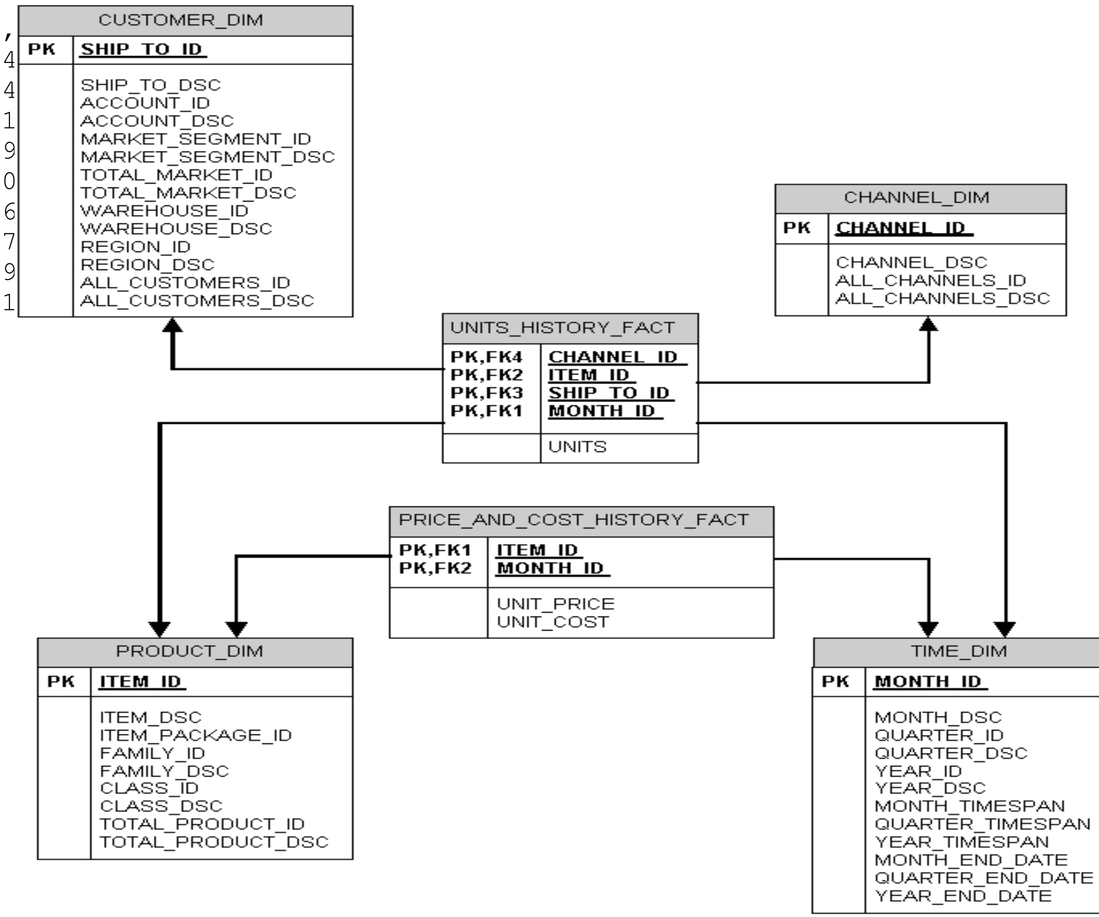
Traditional BI feeds off:

```
"SUMLEV","STATE","COUNTY","STNAME","CTYNAME","YEAR","POPESTIMATE",
50,19,1,"Iowa","Adair County",1,8243,4036,4207,446,225,221,994,509
50,19,1,"Iowa","Adair County",2,8243,4036,4207,446,225,221,994,509
50,19,1,"Iowa","Adair County",3,8212,4020,4192,442,222,220,987,505
50,19,1,"Iowa","Adair County",4,8095,3967,4128,432,208,224,935,488
50,19,1,"Iowa","Adair County",5,8003,3924,4079,405,186,219,928,495
50,19,1,"Iowa","Adair County",6,7961,3892,4069,384,183,201,907,472
50,19,1,"Iowa","Adair County",7,7875,3855,4020,366,179,187,871,454
50,19,1,"Iowa","Adair County",8,7795,3817,3978,343,162,181,841,439
50,19,1,"Iowa","Adair County",9,7714,3777,3937,338,159,179,805,417
```

The "Unstructured Data" Challenge

Traditional BI feeds off:

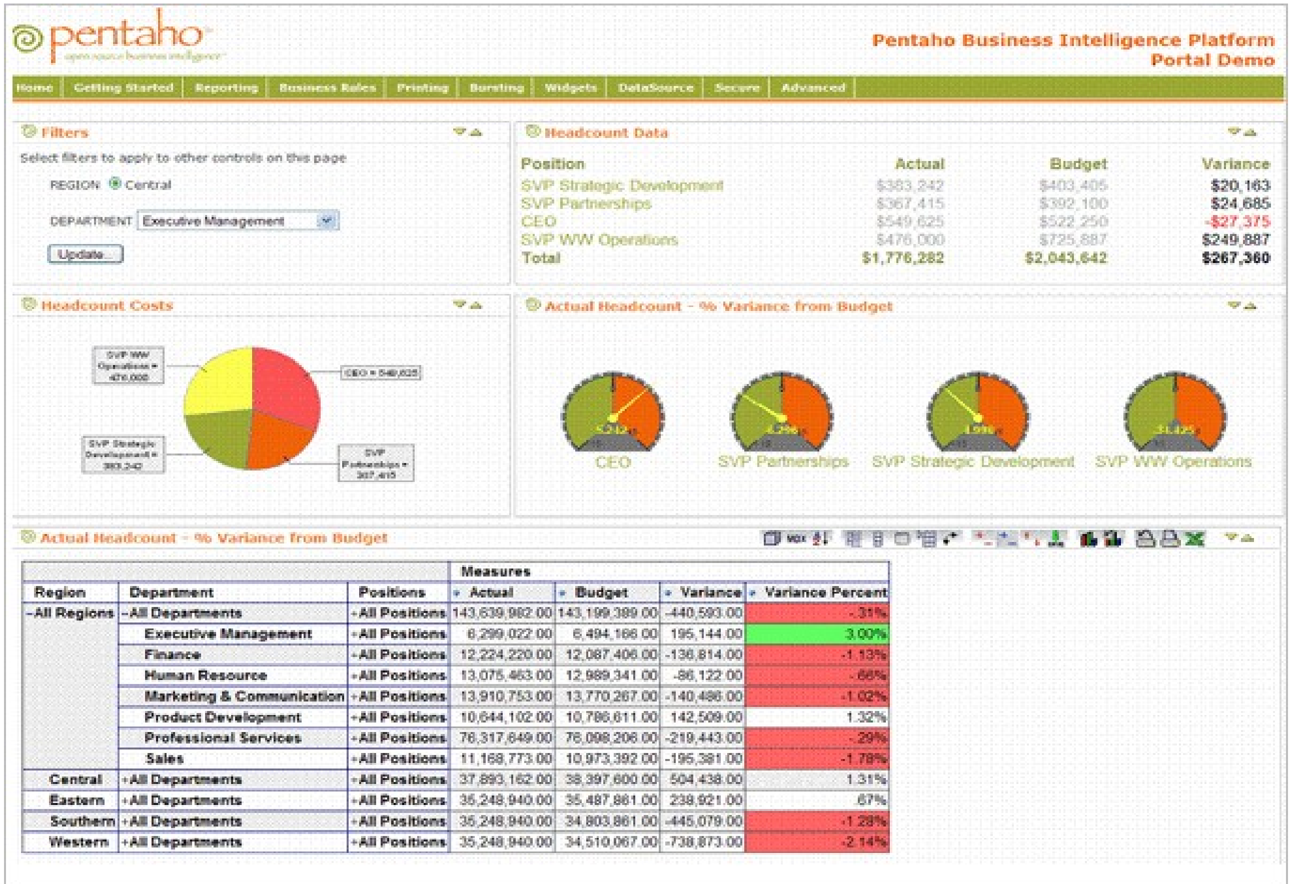
```
"SUMLEV","STATE","COUNTY","STNAME",
50,19,1,"Iowa","Adair County",1,824
50,19,1,"Iowa","Adair County",2,824
50,19,1,"Iowa","Adair County",3,821
50,19,1,"Iowa","Adair County",4,809
50,19,1,"Iowa","Adair County",5,800
50,19,1,"Iowa","Adair County",6,796
50,19,1,"Iowa","Adair County",7,787
50,19,1,"Iowa","Adair County",8,779
50,19,1,"Iowa","Adair County",9,771
```



It runs off:

The "Unstructured Data" Challenge

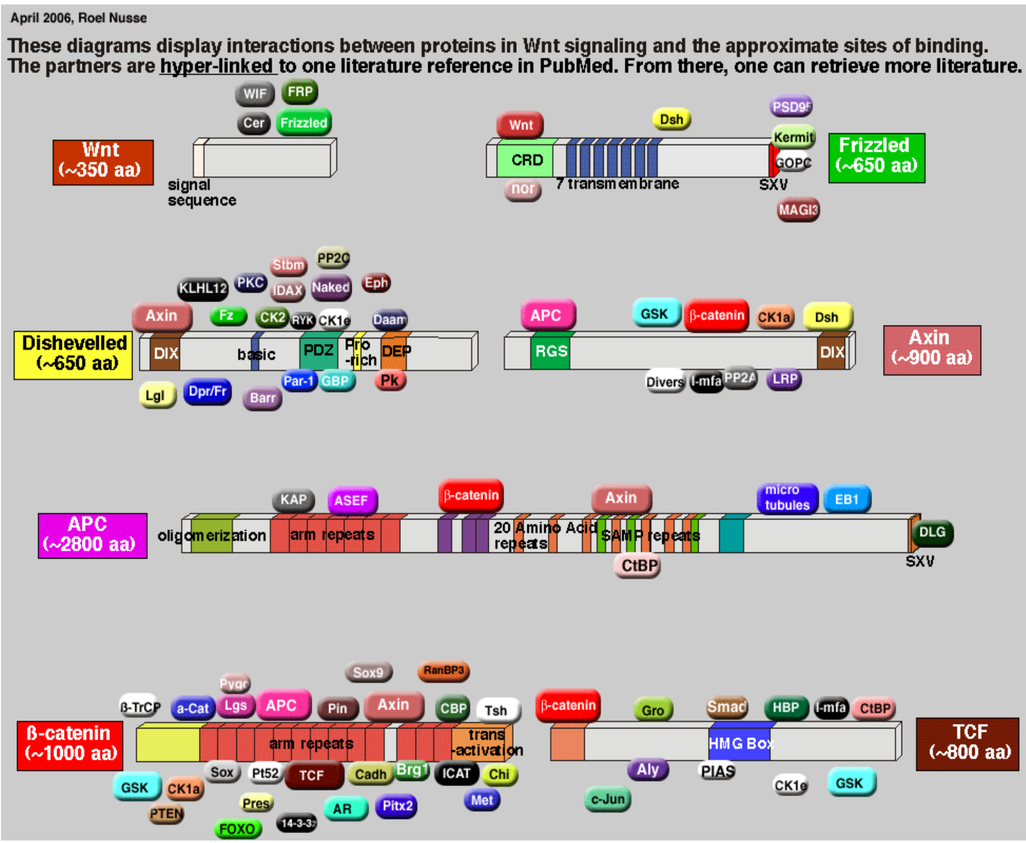
Traditional BI produces:



http://www.pentaho.com/products/dashboards/

The “Unstructured Data” Challenge

Some information doesn’t come from a data file.



www.stanford.edu/%7ernusse/wntwindow.html

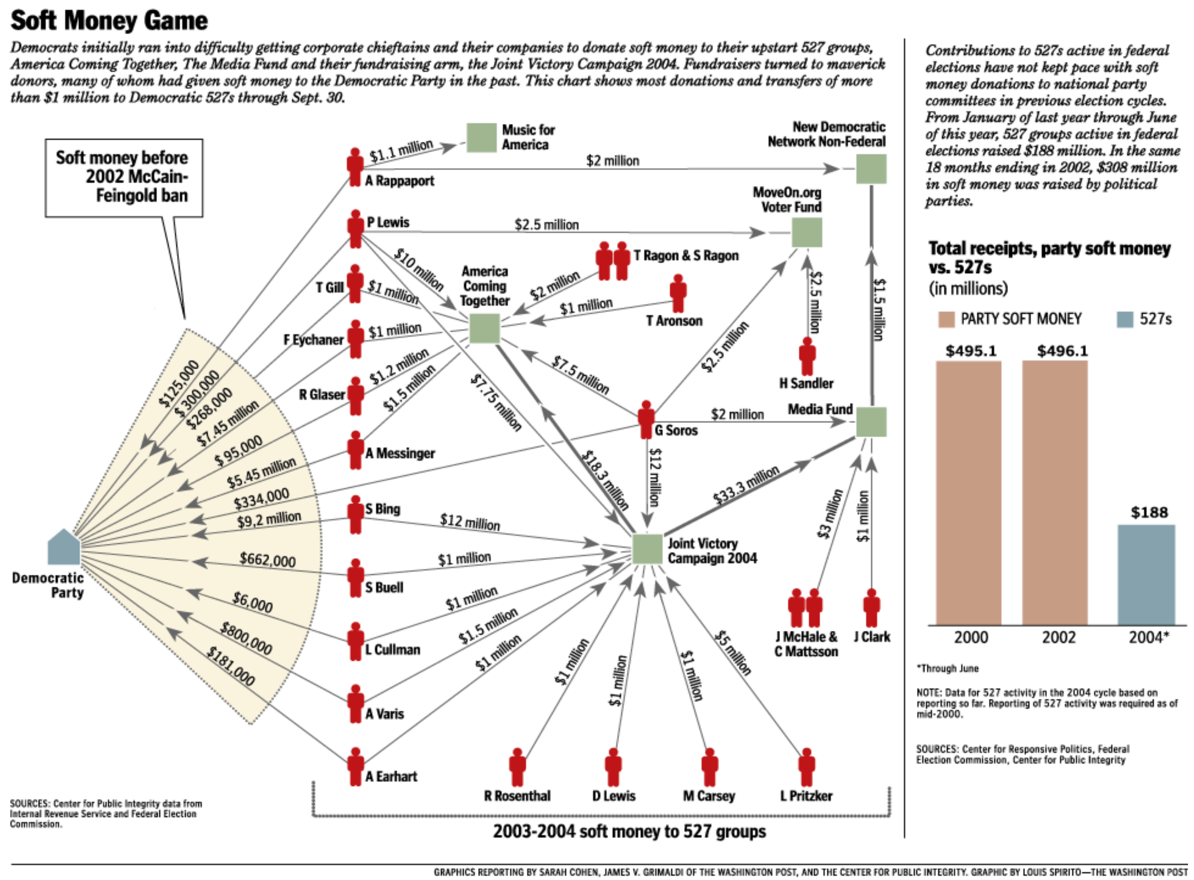
Axin and Frat1 interact with dvl and GSK, bridging Dvl to GSK in Wnt-mediated regulation of LEF-1.

Wnt proteins transduce their signals through dishevelled (Dvl) proteins to inhibit glycogen synthase kinase 3beta (GSK), leading to the accumulation of cytosolic beta-catenin and activation of TCF/LEF-1 transcription factors. To understand the mechanism by which Dvl acts through GSK to regulate LEF-1, we investigated the roles of Axin and Frat1 in Wnt-mediated activation of LEF-1 in mammalian cells. We found that Dvl interacts with Axin and with Frat1, both of which interact with GSK. Similarly, the Frat1 homolog GBP binds Xenopus Dishevelled in an interaction that requires GSK. We also found that Dvl, Axin and GSK can form a ternary complex bridged by Axin, and that Frat1 can be recruited into this complex probably by Dvl. The observation that the Dvl-binding domain of either Frat1 or Axin was able to inhibit Wnt-1-induced LEF-1 activation suggests that the interactions between Dvl and Axin and between Dvl and Frat may be important for this signaling pathway. Furthermore, Wnt-1 appeared to promote the disintegration of the Frat1-Dvl-GSK-Axin complex, resulting in the dissociation of GSK from Axin. Thus, formation of the quaternary complex may be an important step in Wnt signaling, by which Dvl recruits Frat1, leading to Frat1-mediated dissociation of GSK from Axin.

www.ncbi.nlm.nih.gov/entrez/query.fcgi?db=PubMed&cmd=Retrieve&list_uids=10428961&dopt=Abstract

The "Unstructured Data" Challenge

Some is best shown as other than a dashboard.



www.washingtonpost.com/wp-srv/politics/daily/graphics/527Diagram_101704.html

Key Message -- #3

So what's BI – the 1958 definition and today's?

In this paper, business is a collection of activities carried on for whatever purpose, be it science, technology, commerce, industry, law, government, defense, et cetera. The communication facility serving the conduct of a business (in the broad sense) may be referred to as an intelligence system. The notion of intelligence is also defined here, in a more general sense, as "the ability to apprehend the **interrelationships of presented facts** in such a way as **to guide action towards a desired goal**."

– *Hans Peter Luhn,* A Business Intelligence System, *IBM Journal, October 1958*

The "Unstructured Data" Challenge

Consider:

- E-mail, news & blog articles, forum postings, and other social media.
- Contact-center notes and transcripts.
- Surveys, feedback forms, warranty claims.
- And every kind of corporate documents imaginable.

These sources may contain "traditional" data.

- The Dow fell 46.58, or 0.42 percent, to 11,002.14. The Standard & Poor's 500 index fell 1.44, or 0.11 percent, to 1,263.85, and the Nasdaq composite gained 6.84, or 0.32 percent, to 2,162.78.

Search

Search is not the answer. I don't (usually) want to find a document; I want to find a fact, the answer to a question:

- What was the population of Paris in 1848?
- What's the best price for new laptop that I'll use for business trips and around the office?
- What do people think of the *Iron Man* movie?
- Who are the top 4 sales people for each product line, region, and quarter for the last two years?

Search

Q&A may involve hidden knowledge:

What was the population of Paris in 1848?

Concepts and complexity:

What's the best price for new laptop that I'll use for business trips and around the office?

Opinion:

What do people think of the *Iron Man* movie?

Calculation and structuring:

Who were the top 4 sales people for each product line, region, and quarter for the last two years?

Search

Search involves –

- Words & phrases: search terms & natural language.
- Qualifiers: include/exclude, and/or, not, etc.

Answers involve –

- Entities: names, e-mail addresses, phone numbers
- Concepts: abstractions of entities.
- Facts and relationships.
- Abstract attributes, e.g., "expensive," "comfortable"
- Opinions, sentiments: attitudinal data.
- ... and sometimes BI objects.

Search

Search is not enough.

Search helps you find things you already know about. It doesn't help you ***discover*** *things you're unaware of.*

Search results often lack ***relevance****.*

Search finds documents, not ***knowledge****.*

Search finds information, but it doesn't enhance your analyses.

Text Mining

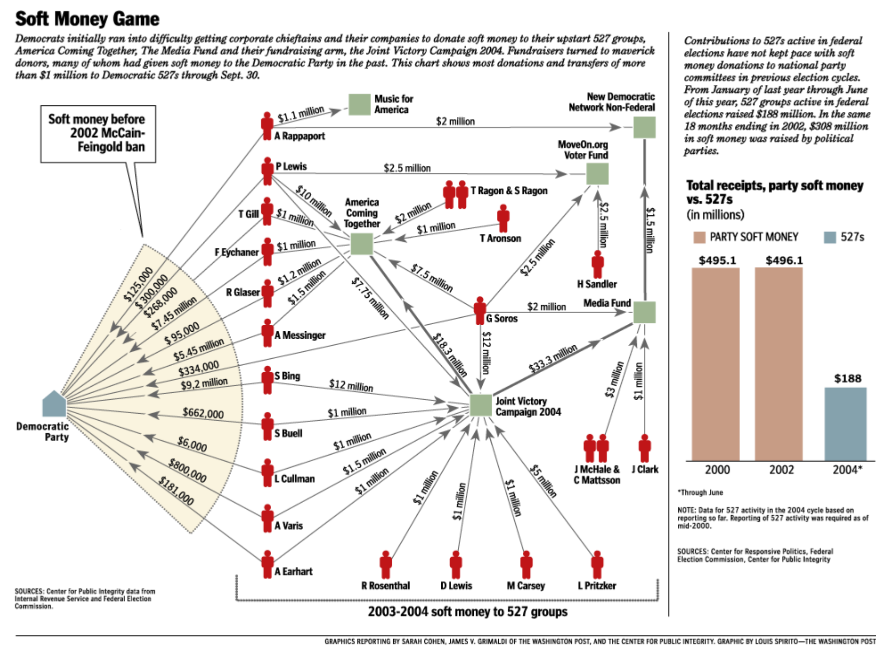
	Search/Query (goal-oriented)	Discovery (opportunistic)
Fielded Data	Data Retrieval	Data Mining
Documents	Information Retrieval	Text Mining

Based on Je Wei Liang, *www.database.cis.nctu.edu.tw/seminars/2003F/TWM/slides/p.ppt*

Text Mining

Text Mining = Data Mining of textual sources.

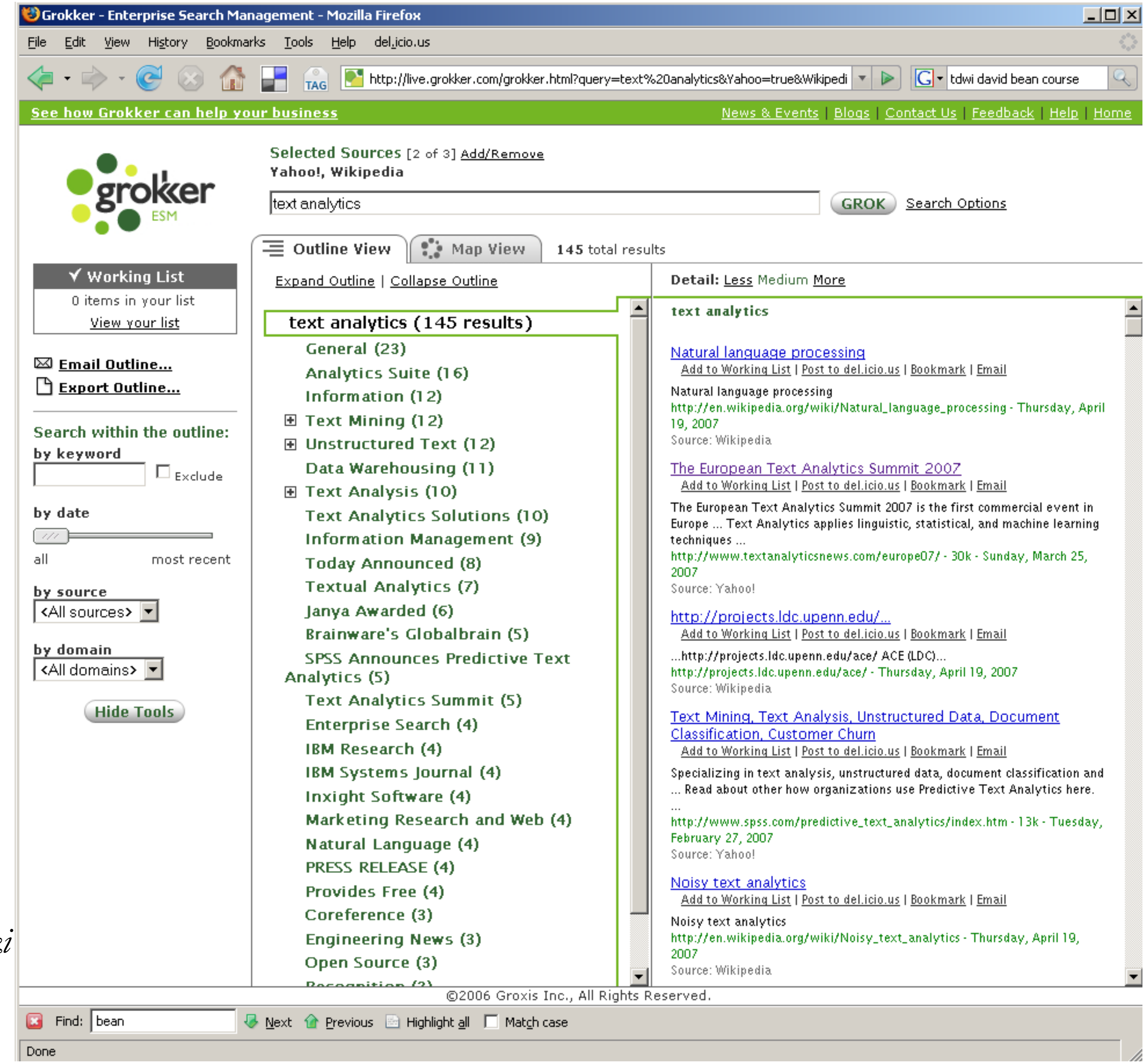
- Clustering and classification.
- Link Analysis.
- Prediction.
- Association rules.
- ~~Regression.~~
- ~~Forecasting.~~

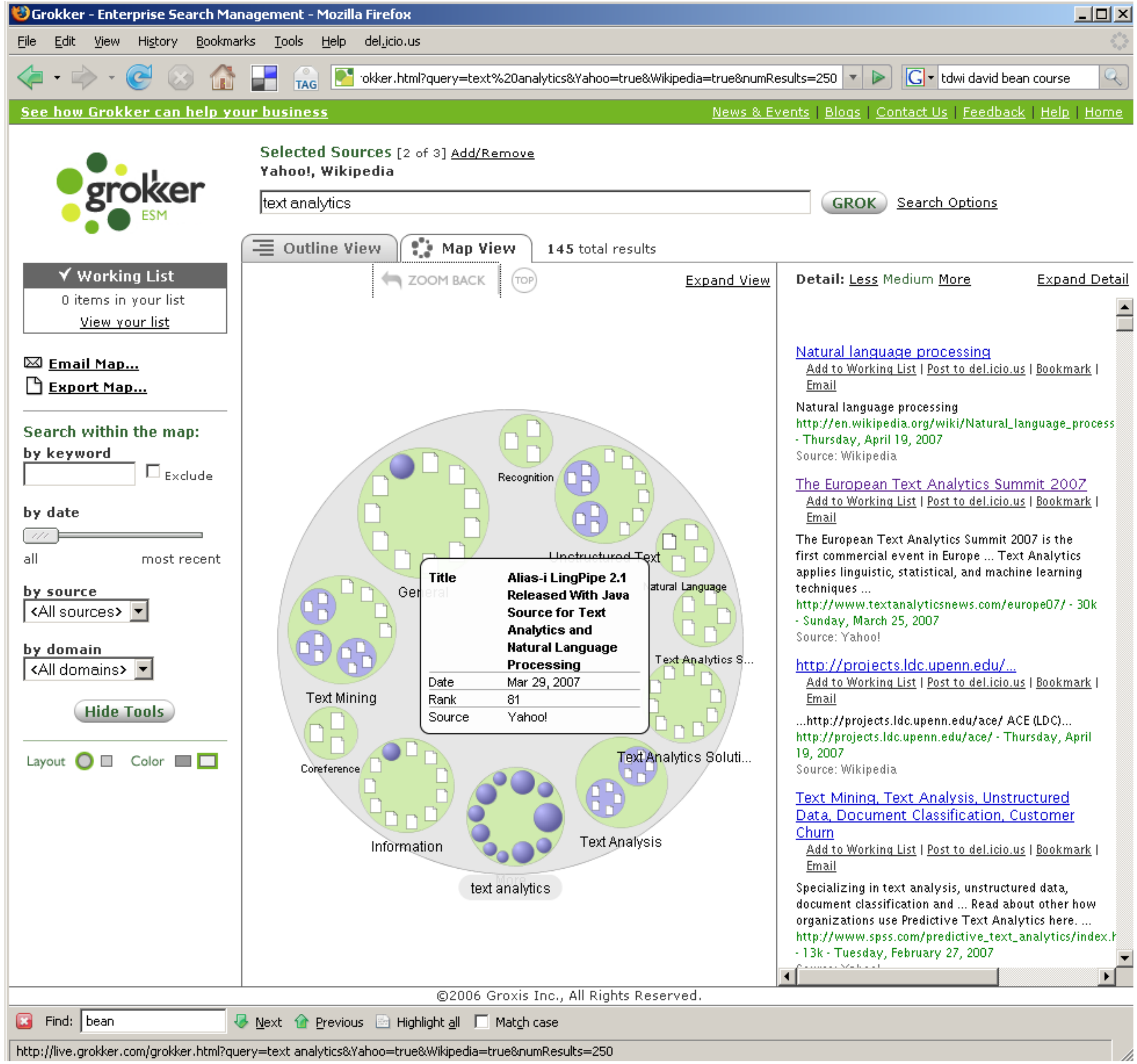


Text Mining = Knowledge Discovery in Text.

Dynamic, clustered search results from Grokker …

live.grokker.com/grokker.html?query=text%20analytics&Yahoo=true&Wikipedia=true&numResults=250

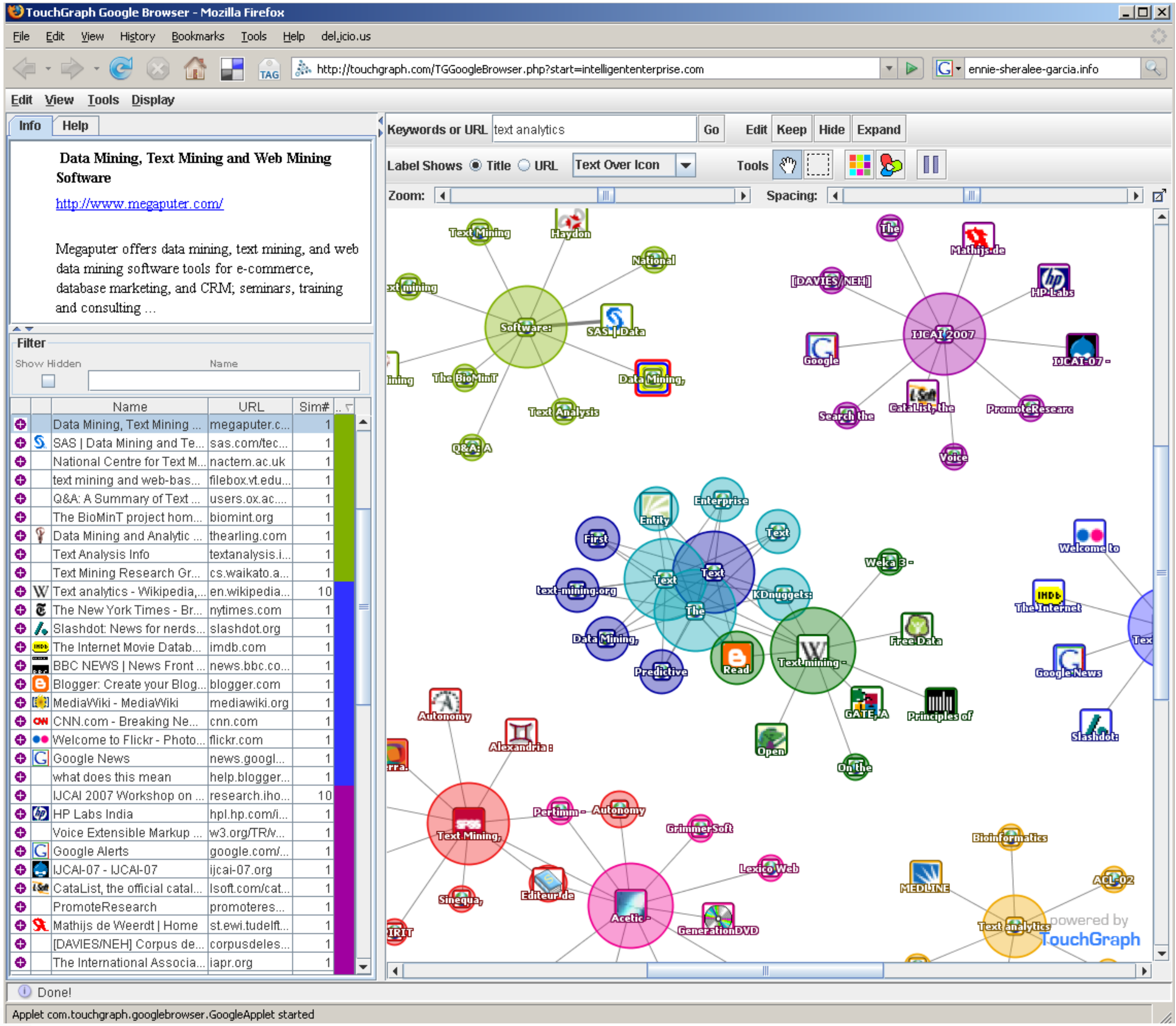


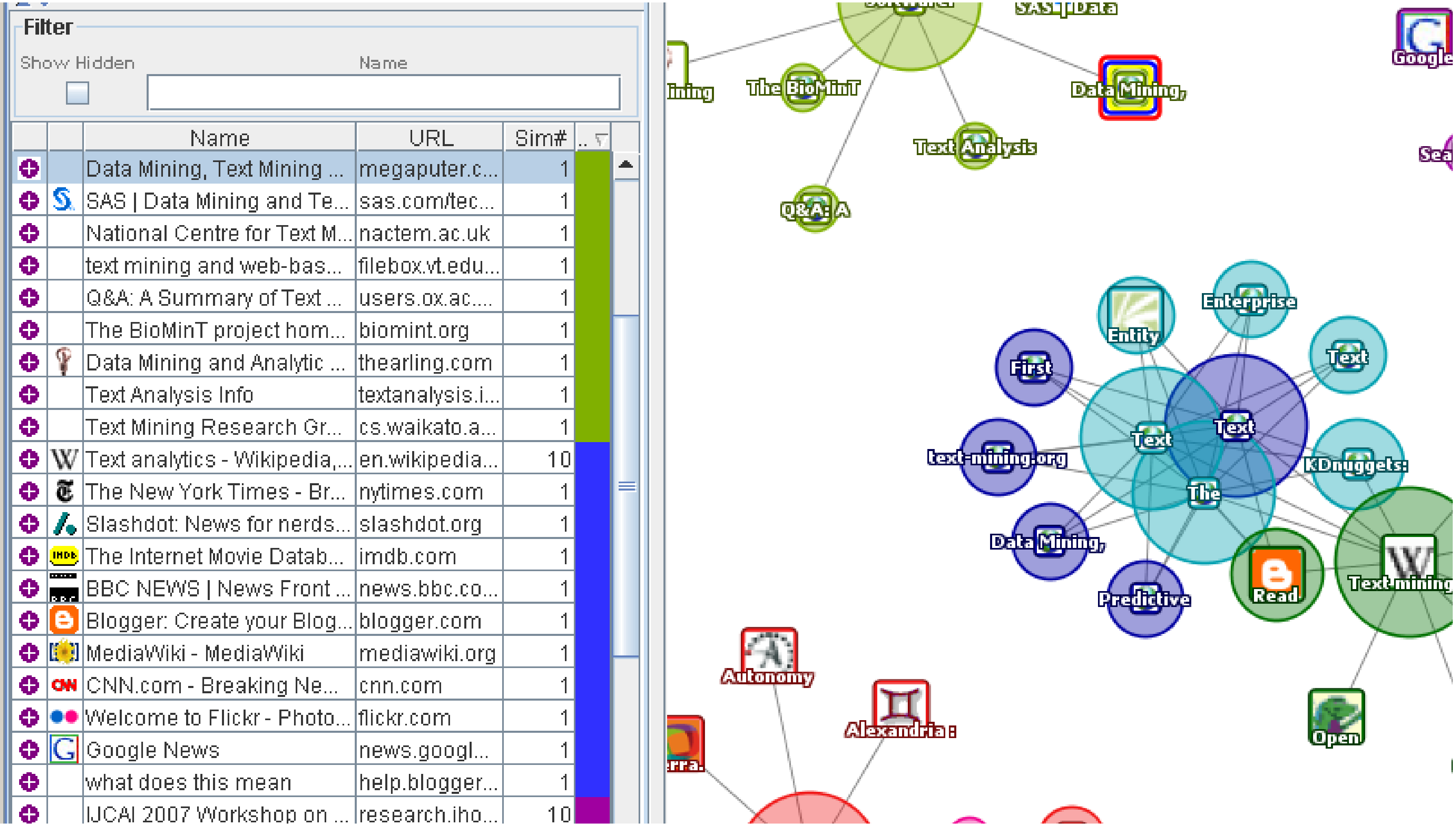


…with a zoomable display

A dynamic network viz.: the Touch-Graph Google-Browser applet

touchgraph.com/TGGoogleBrowser.php?start=text%20analytics




Filter
Show Hidden
Name
Name
URL
Sim#
Data Mining, Text Mining ... megaputer.c... 1
SAS | Data Mining and Te... sas.com/tec... 1
National Centre for Text M... nactem.ac.uk 1
text mining and web-bas... filebox.vt.edu... 1
Q&A: A Summary of Text ... users.ox.ac.... 1
The BioMinT project hom... biomint.org 1
Data Mining and Analytic ... thearling.com 1
Text Analysis Info textanalysis.i... 1
Text Mining Research Gr... cs.waikato.a... 1
Text analytics - Wikipedia,... en.wikipedia... 10
The New York Times - Br... nytimes.com 1
Slashdot: News for nerds... slashdot.org 1
The Internet Movie Datab... imdb.com 1
BBC NEWS | News Front ... news.bbc.co... 1
Blogger: Create your Blog... blogger.com 1
MediaWiki - MediaWiki mediawiki.org 1
CNN.com - Breaking Ne... cnn.com 1
Welcome to Flickr - Photo... flickr.com 1
Google News news.googl... 1
what does this mean help.blogger... 1
IJCAI 2007 Workshop on ... research.iho... 10
The BioMinT
Data Mining,
Text Analysis
Q&A: A
Google
Entity
Enterprise
First
Text
Text
Text
text-mining.org
The
KDnuggets:
Data Mining,
Predictive
Read
Text-mining
Autonomy
Alexandria :
Open

Alta Plana

Text Analytics

Text (and media?) mining **automates** what researchers, writers, scholars,… and all the rest of us have been doing for years. Text mining –

Applies linguistic and/or statistical techniques to extract concepts and patterns that can be applied to categorize and classify documents, audio, video, images.

Transforms "unstructured" information into data for application of traditional analysis techniques via modelling.

Unlocks meaning and relationships in large volumes of information that was previously unprocessable by computer.

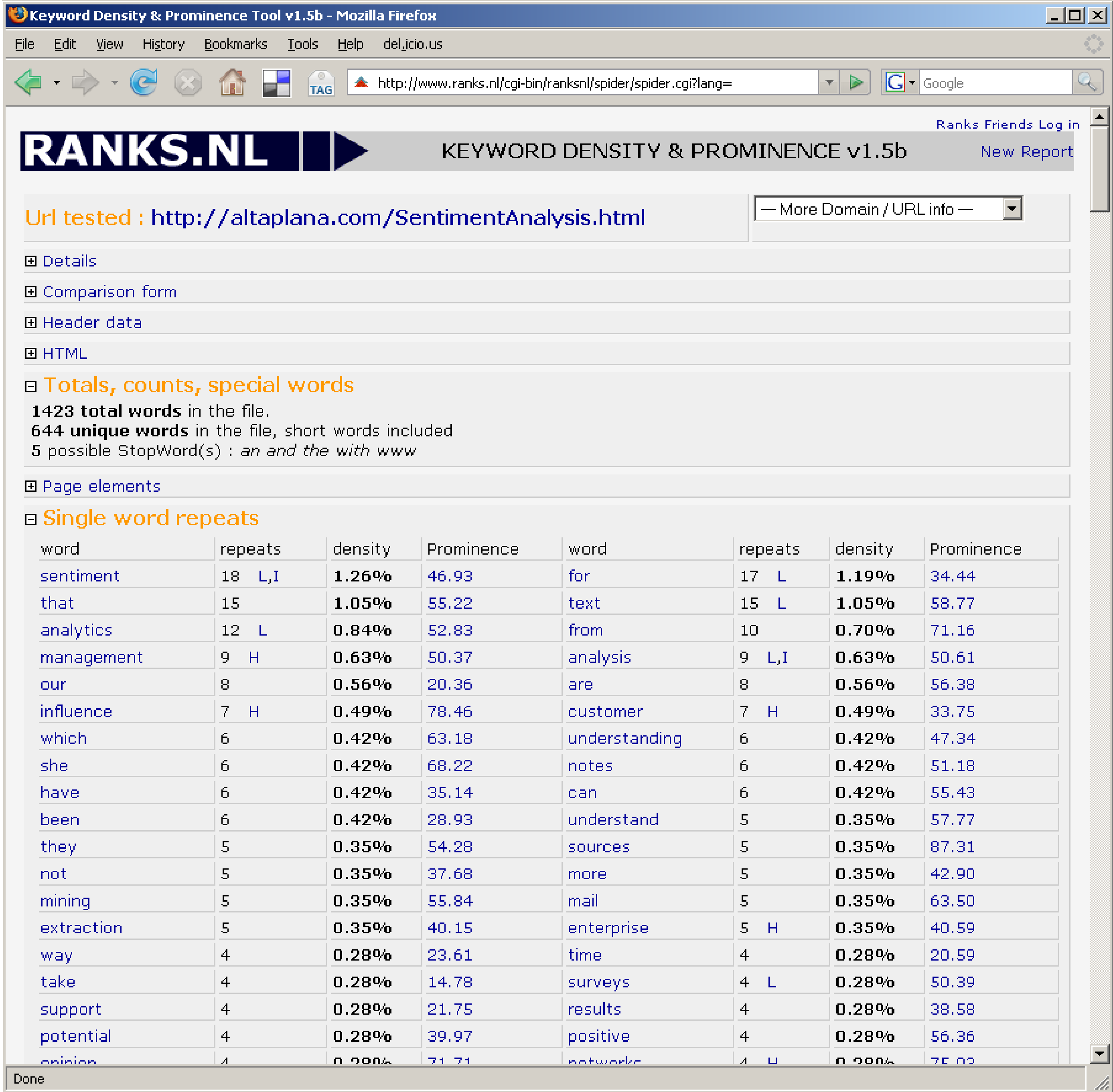
Text Analytics

To digress… Is text really unstructured?

No! If it were, you wouldn't be able to understand this sentence.

Text is instead ***unmodelled****.*

We'll look for that inherent structure, but first, we'll do a lexical analysis of a text file...



Keyword Density & Prominence Tool v1.5b - Mozilla Firefox

File Edit View History Bookmarks Tools Help del.icio.us

http://www.ranks.nl/cgi-bin/ranksnl/spider/spider.cgi?lang=

Ranks Friends Log in

RANKS.NL KEYWORD DENSITY & PROMINENCE v1.5b New Report

Url tested : http://altaplana.com/SentimentAnalysis.html

— More Domain / URL info —

⊞ Details

⊞ Comparison form

⊞ Header data

⊞ HTML

⊟ Totals, counts, special words

1423 total words in the file.
644 unique words in the file, short words included
5 possible StopWord(s) : *an and the with www*

⊞ Page elements

⊟ Single word repeats

word	repeats	density	Prominence	word	repeats	density	Prominence
sentiment	18 L,I	**1.26%**	46.93	for	17 L	**1.19%**	34.44
that	15	**1.05%**	55.22	text	15 L	**1.05%**	58.77
analytics	12 L	**0.84%**	52.83	from	10	**0.70%**	71.16
management	9 H	**0.63%**	50.37	analysis	9 L,I	**0.63%**	50.61
our	8	**0.56%**	20.36	are	8	**0.56%**	56.38
influence	7 H	**0.49%**	78.46	customer	7 H	**0.49%**	33.75
which	6	**0.42%**	63.18	understanding	6	**0.42%**	47.34
she	6	**0.42%**	68.22	notes	6	**0.42%**	51.18
have	6	**0.42%**	35.14	can	6	**0.42%**	55.43
been	6	**0.42%**	28.93	understand	5	**0.35%**	57.77
they	5	**0.35%**	54.28	sources	5	**0.35%**	87.31
not	5	**0.35%**	37.68	more	5	**0.35%**	42.90
mining	5	**0.35%**	55.84	mail	5	**0.35%**	63.50
extraction	5	**0.35%**	40.15	enterprise	5 H	**0.35%**	40.59
way	4	**0.28%**	23.61	time	4	**0.28%**	20.59
take	4	**0.28%**	14.78	surveys	4 L	**0.28%**	50.39
support	4	**0.28%**	21.75	results	4	**0.28%**	38.58
potential	4	**0.28%**	39.97	positive	4	**0.28%**	56.36
opinion	4	**0.28%**	71.71	networks	4 H	**0.28%**	75.03

Done

Keyword Density & Prominence Tool v1.5b - Mozilla Firefox

File Edit View History Bookmarks Tools Help del.icio.us

http://www.ranks.nl/cgi-bin/ranksnl/spider/spider.cgi?lang=

Phrase repeats

Total 2 word phrases : 102 - Total Repeats : 246

phrase	repeats	density	Prominence
text analytics	9	**1.26 %**	58.87
of the	6	**0.84 %**	46.49
and the	4	**0.56 %**	48.45
e mail	4	**0.56 %**	62.86
from sources	4	**0.56 %**	88.12
influence networks	4 H	**0.56 %**	76.00
notes and	4	**0.56 %**	52.11
of text	4	**0.56 %**	52.37
to the	4	**0.56 %**	60.17
to understand	4	**0.56 %**	63.55
by the	3	**0.42 %**	34.65
call center	3	**0.42 %**	68.96
can be	3	**0.42 %**	81.68
customer experience	3 H	**0.42 %**	52.99
enterprise feedback	3 H	**0.42 %**	52.73
experience management	3 H	**0.42 %**	52.92
feedback management	3 H	**0.42 %**	52.66
in the	3	**0.42 %**	41.79
of opinion	3	**0.42 %**	69.97
real time	3	**0.42 %**	17.01
seek to	3	**0.42 %**	28.58
sentiment analysis	3 L,I	**0.42 %**	69.52
sentiment extraction	3	**0.42 %**	37.29
the results	3	**0.42 %**	33.45
triggered by	3	**0.42 %**	26.00
a decision	2	**0.28 %**	20.41
a new	2	**0.28 %**	65.21
analytics can	2	**0.28 %**	97.15
analytics vendor	2	**0.28 %**	55.02
analyze attitudinal	2	**0.28 %**	96.66
and analyze	2	**0.28 %**	96.73
and other	2	**0.28 %**	37.70

Total 3 word phrases : 45 - Total Repeats : 93

phrase	repeats	density	Prominence
customer experience management	3 H	**0.63 %**	52.99
enterprise feedback management	3 H	**0.63 %**	52.73
of text analytics	3	**0.63 %**	46.78
analytics can be	2	**0.42 %**	97.15
analyze attitudinal information	2	**0.42 %**	96.66
and analyze attitudinal	2	**0.42 %**	96.73
and survey responses	2	**0.42 %**	95.54
applied to extract	2	**0.42 %**	96.94
articles blog postings	2	**0.42 %**	96.10
as articles blog	2	**0.42 %**	96.17
as varied as	2	**0.42 %**	96.31
attitudinal information from	2	**0.42 %**	96.59
be applied to	2	**0.42 %**	97.01
blog postings e	2	**0.42 %**	96.03
call center notes	2	**0.42 %**	95.75
can be applied	2	**0.42 %**	97.08
center notes and	2	**0.42 %**	95.68
ceo of text	2	**0.42 %**	55.24
cries for help	2	**0.42 %**	7.70
e mail call	2	**0.42 %**	95.89
experience management enterprise	2 H	**0.42 %**	62.65
extract and analyze	2	**0.42 %**	96.80
focus on applications	2	**0.42 %**	97.96
from linguamatics to	2	**0.42 %**	81.52
from sources as	2	**0.42 %**	96.45
information from sources	2	**0.42 %**	96.52
mail call center	2	**0.42 %**	95.82
management enterprise feedback	2 H	**0.42 %**	62.58
notes and survey	2	**0.42 %**	95.61
of opinion leadership	2	**0.42 %**	80.43
online consumer forums	2	**0.42 %**	55.90
postings e mail	2	**0.42 %**	95.96

Done

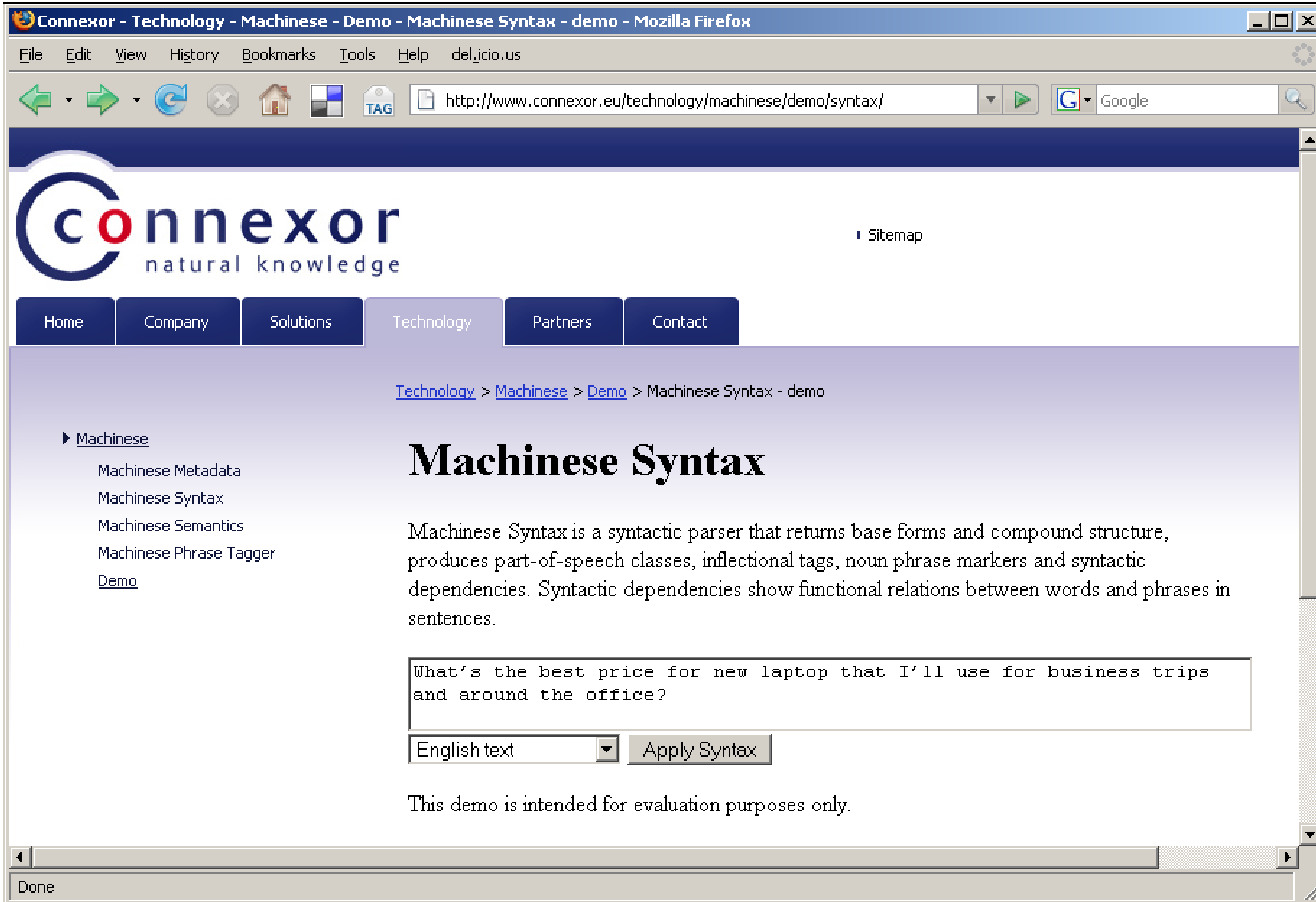
Text Analytics

Lesson: "Structure" may not matter.

Shallow parsing and statistical analysis can be enough to arrive at the *Whatness* of a text, for instance, to support classification. (But that's not BI.)

It can help you get at meaning, for instance, by studying cooccurrence of terms.

Now a syntactic analysis of a bit of text, a sentence...


Connexor - Technology - Machinese - Demo - Machinese Syntax - demo - Mozilla Firefox
File Edit View History Bookmarks Tools Help del.icio.us
http://www.connexor.eu/technology/machinese/demo/syntax/
Google
connexor
natural knowledge
Sitemap
Home Company Solutions Technology Partners Contact
Technology > Machinese > Demo > Machinese Syntax - demo
Machinese
Machinese Metadata
Machinese Syntax
Machinese Semantics
Machinese Phrase Tagger
Demo
Machinese Syntax
Machinese Syntax is a syntactic parser that returns base forms and compound structure, produces part-of-speech classes, inflectional tags, noun phrase markers and syntactic dependencies. Syntactic dependencies show functional relations between words and phrases in sentences.
What’s the best price for new laptop that I’ll use for business trips and around the office?
English text
Apply Syntax
This demo is intended for evaluation purposes only.
Done


Connexor - Technology - Machinese - Demo - Machinese Syntax - demo - Mozilla Firefox
File Edit View History Bookmarks Tools Help del.icio.us
http://www.connexor.eu/technology/machinese/demo/syntax/
Google
connexor
natural knowledge
Sitemap
Home
Company
Solutions
Technology
Partners
Contact
Technology > Machinese > Demo > Machinese Syntax - demo
Machinese
Machinese Metadata
Machinese Syntax
Machinese Semantics
Machinese Phrase Tagger
Demo
Analysis of Machinese Syntax for English:
root
main:
's
subj:
comp:
What
price
attr:
det:
mod:
best
the
for
pcomp:
laptop
attr:
mod:
new
use
v-ch:
obj:
ha:
'll
that
for
subj:
cc:
cc:
around
I
trips
and
pcomp:
attr:
office
business
det:
the
Note: The Connexor Machinese demos are intended for evaluation purposes only.
Connexor Oy, Helsinki Business and Science Park. Finland. info@connexor.com
© Connexor Oy. Powered by ToimiSait
Applet Dtree started

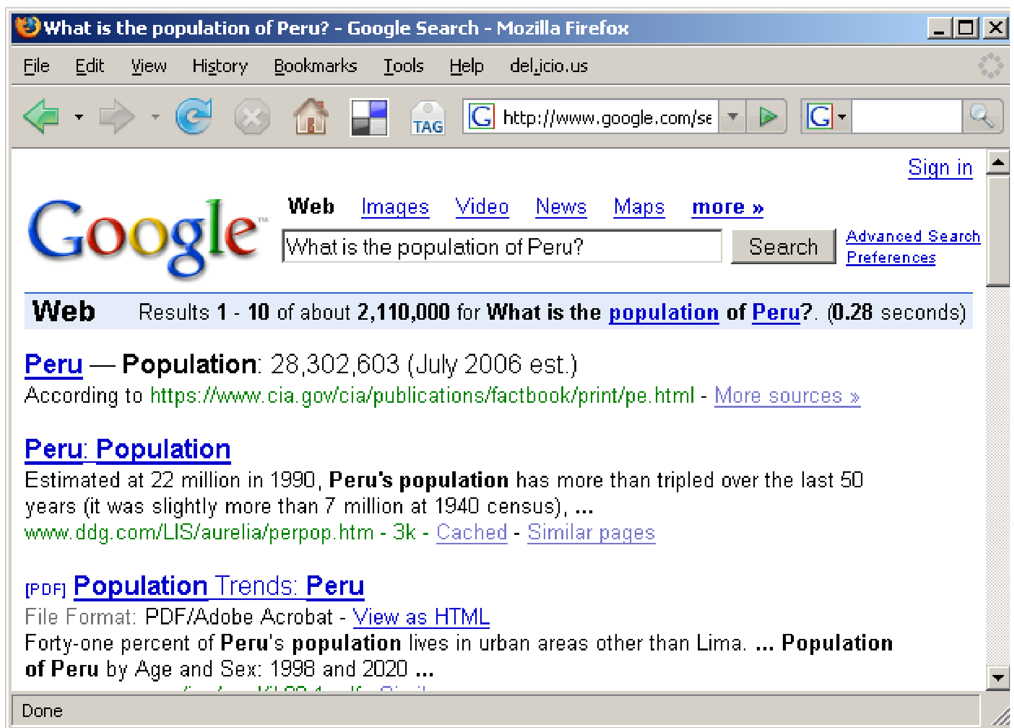
connexor
natural knowledge

English Machinese Phrase Tagger 4.6 analysis:

Text	Baseform	Phrase syntax and part-of-speech
What	what	nominal head, pro-nominal
's	be	main verb, indicative present
the	the	premodifier, determiner
best	good	premodifier, superlative adjective, noun phrase begins
price	price	nominal head, noun, noun phrase continues
for	for	postmodifier, preposition, noun phrase continues
new	new	premodifier, adjective, noun phrase continues
laptop	lap top	nominal head, noun, noun phrase ends
that	that	nominal head, pro-nominal
I	I	nominal head, pro-nominal
'll	will	auxiliary verb, indicative present
use	use	main verb, infinitive
for	for	preposed marker, preposition
business	business	premodifier, noun, noun phrase begins
trips	trip	nominal head, plural noun, noun phrase ends

Text Analytics

So the form may be unstructured but the content isn't. Text analytics – unified analytics – should present findings that suit the information and the user.



Text Analytics

Typical steps in text analytics include –

Retrieve documents for analysis.

Create a categorization/taxonomy from the extracts or acquire and apply a domain-specific taxonomy.

Apply statistical techniques to classify documents, look for patterns such as associations and clusters.

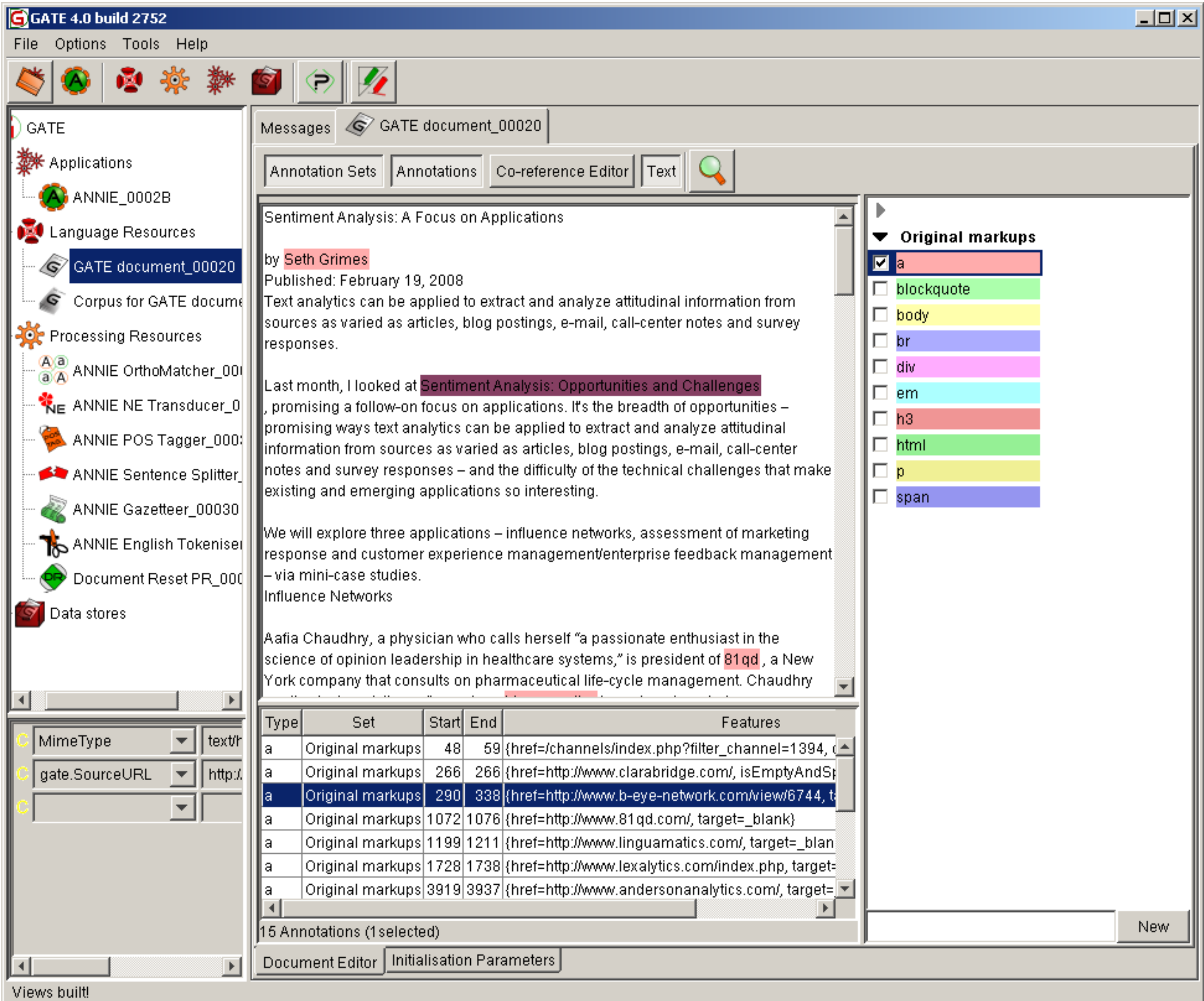
Apply statistical &/ linguistic &/ structural techniques to **identify, tag, and extract** entities, concepts, relationships, and events (features) within document sets.

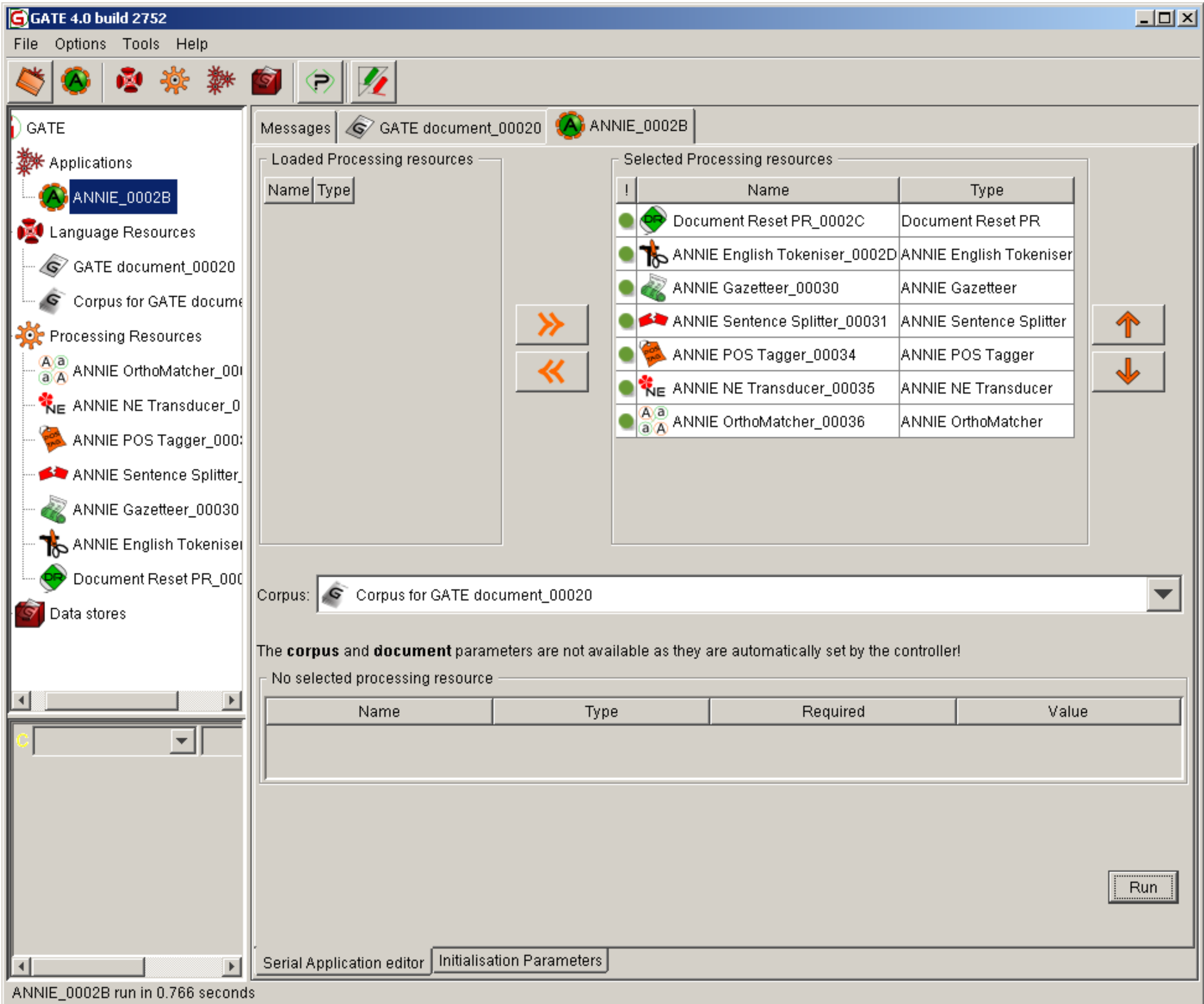
- tagging = text augmentation

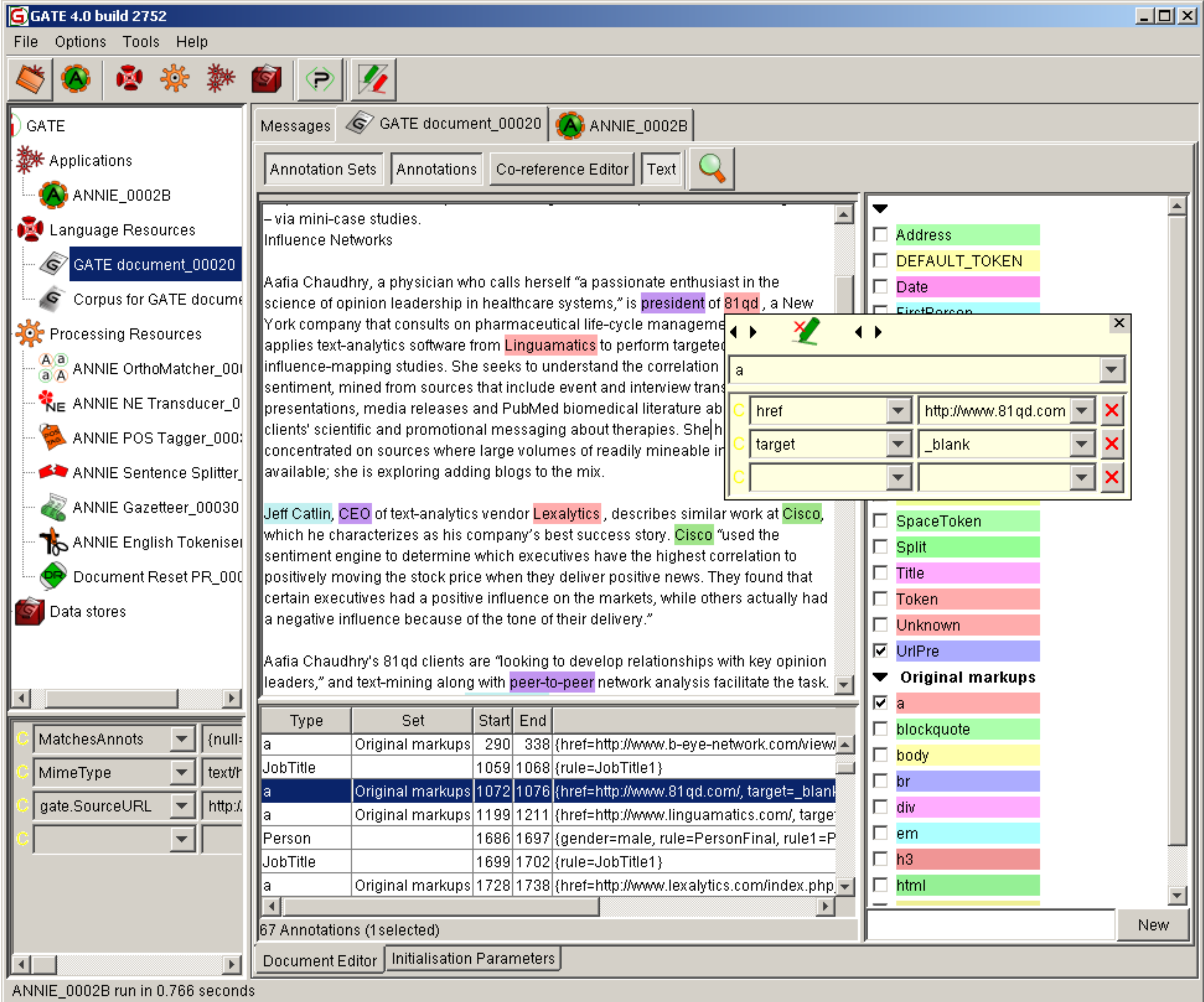
Information Extraction

For "traditional" BI on text, key in on extracting information to databases.

- Entities and concepts (features) are like dimensions in a standard BI model. Both classes of object are hierarchically organized and have attributes.
- We can have both discovered and predetermined classifications (taxonomies) of text features.
- Text-sourced information is **very** high dimensionality.


GATE 4.0 build 2752
File Options Tools Help
GATE
Applications
ANNIE_0002B
Language Resources
GATE document_00020
Corpus for GATE docume
Processing Resources
ANNIE OrthoMatcher_00
ANNIE NE Transducer_0
ANNIE POS Tagger_000
ANNIE Sentence Splitter_
ANNIE Gazetteer_00030
ANNIE English Tokenise
Document Reset PR_000
Data stores
MimeType
text/h
gate.SourceURL
http:/
Messages
GATE document_00020
Annotation Sets
Annotations
Co-reference Editor
Text
Sentiment Analysis: A Focus on Applications
by Seth Grimes
Published: February 19, 2008
Text analytics can be applied to extract and analyze attitudinal information from sources as varied as articles, blog postings, e-mail, call-center notes and survey responses.
Last month, I looked at Sentiment Analysis: Opportunities and Challenges, promising a follow-on focus on applications. It's the breadth of opportunities – promising ways text analytics can be applied to extract and analyze attitudinal information from sources as varied as articles, blog postings, e-mail, call-center notes and survey responses – and the difficulty of the technical challenges that make existing and emerging applications so interesting.
We will explore three applications – influence networks, assessment of marketing response and customer experience management/enterprise feedback management – via mini-case studies.
Influence Networks
Aafia Chaudhry, a physician who calls herself "a passionate enthusiast in the science of opinion leadership in healthcare systems," is president of 81qd, a New York company that consults on pharmaceutical life-cycle management. Chaudhry
Original markups
a
blockquote
body
br
div
em
h3
html
p
span
Type | Set | Start | End | Features
a | Original markups | 48 | 59 | {href=/channels/index.php?filter_channel=1394,
a | Original markups | 266 | 266 | {href=http://www.clarabridge.com/, isEmptyAndSp
a | Original markups | 290 | 338 | {href=http://www.b-eye-network.com/view/6744, t
a | Original markups | 1072 | 1076 | {href=http://www.81qd.com/, target=_blank}
a | Original markups | 1199 | 1211 | {href=http://www.linguamatics.com/, target=_blan
a | Original markups | 1728 | 1738 | {href=http://www.lexalytics.com/index.php, target=
a | Original markups | 3919 | 3937 | {href=http://www.andersonanalytics.com/, target=
15 Annotations (1 selected)
New
Document Editor
Initialisation Parameters
Views built!

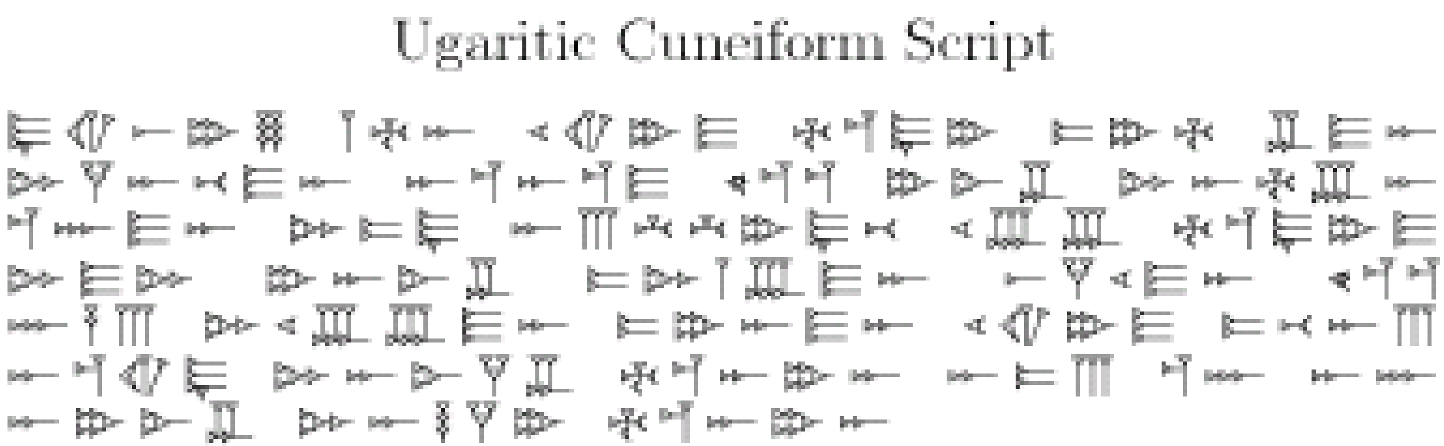

GATE 4.0 build 2752
File Options Tools Help
GATE
Applications
ANNIE_0002B
Language Resources
GATE document_00020
Corpus for GATE docume
Processing Resources
ANNIE OrthoMatcher_00
ANNIE NE Transducer_0
ANNIE POS Tagger_000
ANNIE Sentence Splitter_
ANNIE Gazetteer_00030
ANNIE English Tokenise
Document Reset PR_000
Data stores
Messages
GATE document_00020
ANNIE_0002B
Loaded Processing resources
Name Type
Selected Processing resources
! Name Type
Document Reset PR_0002C Document Reset PR
ANNIE English Tokeniser_0002D ANNIE English Tokeniser
ANNIE Gazetteer_00030 ANNIE Gazetteer
ANNIE Sentence Splitter_00031 ANNIE Sentence Splitter
ANNIE POS Tagger_00034 ANNIE POS Tagger
ANNIE NE Transducer_00035 ANNIE NE Transducer
ANNIE OrthoMatcher_00036 ANNIE OrthoMatcher
Corpus: Corpus for GATE document_00020
The corpus and document parameters are not available as they are automatically set by the controller!
No selected processing resource
Name Type Required Value
Run
Serial Application editor
Initialisation Parameters
ANNIE_0002B run in 0.766 seconds


GATE 4.0 build 2752
File Options Tools Help
GATE
Applications
ANNIE_0002B
Language Resources
GATE document_00020
Corpus for GATE docume
Processing Resources
ANNIE OrthoMatcher_00
ANNIE NE Transducer_0
ANNIE POS Tagger_000
ANNIE Sentence Splitter
ANNIE Gazetteer_00030
ANNIE English Tokeniser
Document Reset PR_000
Data stores
MatchesAnnots {null
MimeType text/
gate.SourceURL http:/
Messages
GATE document_00020
ANNIE_0002B
Annotation Sets
Annotations
Co-reference Editor
Text
– via mini-case studies.
Influence Networks
Aafia Chaudhry, a physician who calls herself "a passionate enthusiast in the science of opinion leadership in healthcare systems," is president of 81qd, a New York company that consults on pharmaceutical life-cycle manageme applies text-analytics software from Linguamatics to perform targete influence-mapping studies. She seeks to understand the correlation sentiment, mined from sources that include event and interview trans presentations, media releases and PubMed biomedical literature ab clients' scientific and promotional messaging about therapies. She h concentrated on sources where large volumes of readily mineable i available; she is exploring adding blogs to the mix.
Jeff Catlin, CEO of text-analytics vendor Lexalytics, describes similar work at Cisco, which he characterizes as his company's best success story. Cisco "used the sentiment engine to determine which executives have the highest correlation to positively moving the stock price when they deliver positive news. They found that certain executives had a positive influence on the markets, while others actually had a negative influence because of the tone of their delivery."
Aafia Chaudhry's 81qd clients are "looking to develop relationships with key opinion leaders," and text-mining along with peer-to-peer network analysis facilitate the task.
a
href http://www.81qd.com
target _blank
Address
DEFAULT_TOKEN
Date
SpaceToken
Split
Title
Token
Unknown
UrlPre
Original markups
a
blockquote
body
br
div
em
h3
html
New
Type Set Start End
a Original markups 290 338 {href=http://www.b-eye-network.com/view/
JobTitle 1059 1068 {rule=JobTitle1}
a Original markups 1072 1076 {href=http://www.81qd.com/, target=_blank
a Original markups 1199 1211 {href=http://www.linguamatics.com/, targe
Person 1686 1697 {gender=male, rule=PersonFinal, rule1=P
JobTitle 1699 1702 {rule=JobTitle1}
a Original markups 1728 1738 {href=http://www.lexalytics.com/index.php
67 Annotations (1 selected)
Document Editor
Initialisation Parameters
ANNIE_0002B run in 0.766 seconds

Information Extraction

Syntactic/linguistic analysis is key to semantic understanding and difficult stuff like sentiment.

Regular expressions and term co-occurrence, also simple statistical signatures, are not enough.



Information Extraction

Consider –

The Dow fell 46.58, or 0.42 percent, to 11,002.14. The Standard & Poor's 500 index fell 1.44, or 0.11 percent, to 1,263.85, and the Nasdaq composite gained 6.84, or 0.32 percent, to 2,162.78.

The Dow gained 46.58, or 0.42 percent, to 11,002.14. The Standard & Poor's 500 index fell 1.44, or 0.11 percent, to 1,263.85, and the Nasdaq composite fell 6.84, or 0.32 percent, to 2,162.78.

Example from Luca Scagliarini, Expert System.

The bag/vector of words approach falls short.

Information Extraction

We want concepts and not just entities.

What concepts are found in these similar examples?

Smaller cars generally get better gas mileage than larger cars.

Some larger hybrids consume less fuel than some smaller vehicles with standard gasoline engines.

Ford is an American automobile manufacturer and Nissan is Japanese.

Information Extraction

What concepts are found in these domain-related statements?

> Smaller cars generally get better gas mileage than larger cars.
>
> Some larger hybrids/hybrids consume less fuel than some smaller vehicles with standard gasoline engines.
>
> Ford is an American automobile manufacturer and Nissan is Japanese.

Vehicle is a *concept* with *conceptual* size and energy consumption attributes and a *conceptual* engine type. Energy consumption itself has a relative measure. Nationality is another concept. What's Ford?

Information Extraction

What's Ford? –

"Ford is an American automobile manufacturer…"

- A president?
- A company that both makes and sells cars and other stuff?
- A person who founded a car company?
- A shallow place you cross a river?

Ford is an entity whose meaning a) is contextually derived; b) may be disambiguated, and c) is more than what is plainly read in our source text.

Information Extraction

Let's look at an e-mail message –

Date: Sun, 13 Mar 2005 19:58:39 -0500

From: Adam L. Buchsbaum <alb@research.att.com>

To: Seth Grimes <grimes@altaplana.com>

Subject: Re: Papers on analysis on streaming data

seth, you should contact divesh srivastava, divesh@research.att.com regarding at&t labs data streaming technology.

adam

Information Extraction

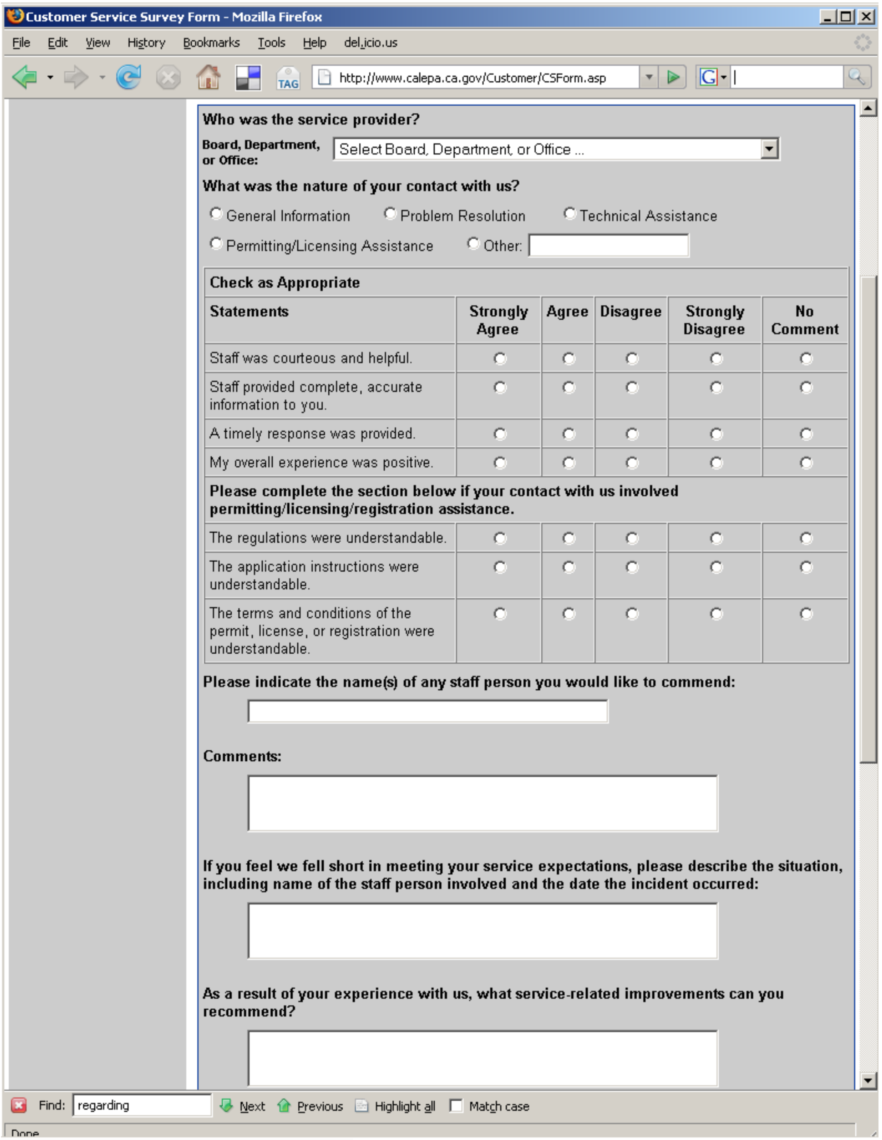
An e-mail message is "semi-structured."

- Semi=half. What's "structured" and what's not?
- Is augmentation/tagging and entity extraction enough?
- What categorization might you create from that example message?
- If we extracted all the entities to a database, what could you do with them?

From semi-structured text, it's especially easy to extract metadata.

There are many forms of s-s information...

Example: Survey



Customer Service Survey Form - Mozilla Firefox

File Edit View History Bookmarks Tools Help del.icio.us

http://www.calepa.ca.gov/Customer/CSForm.asp

Who was the service provider?

Board, Department, or Office: Select Board, Department, or Office ...

What was the nature of your contact with us?

General Information · Problem Resolution · Technical Assistance · Permitting/Licensing Assistance · Other:

Check as Appropriate

Statements	Strongly Agree	Agree	Disagree	Strongly Disagree	No Comment
Staff was courteous and helpful.	○	○	○	○	○
Staff provided complete, accurate information to you.	○	○	○	○	○
A timely response was provided.	○	○	○	○	○
My overall experience was positive.	○	○	○	○	○
Please complete the section below if your contact with us involved permitting/licensing/registration assistance.					
The regulations were understandable.	○	○	○	○	○
The application instructions were understandable.	○	○	○	○	○
The terms and conditions of the permit, license, or registration were understandable.	○	○	○	○	○

Please indicate the name(s) of any staff person you would like to commend:

Comments:

If you feel we fell short in meeting your service expectations, please describe the situation, including name of the staff person involved and the date the incident occurred:

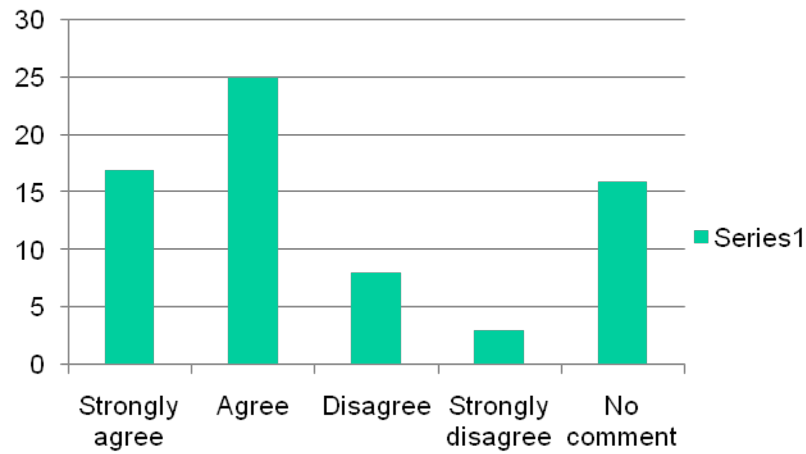
As a result of your experience with us, what service-related improvements can you recommend?

Find: regarding Next Previous Highlight all Match case

Done

Example: Survey

In analyzing surveys, we typically look at frequencies and distributions:



There may be fields that indicate what product/service/person the coded rating applies to.

Comments may be linked to coded ratings.

Example: Survey

The respondent is invited to explain his/her attitude:

My overall experience was positive.					
Please complete the section below if your contact with us involved permitting/licensing/registration assistance.					
The regulations were understandable.					
The application instructions were understandable.					
The terms and conditions of the permit, license, or registration were understandable.					

Please indicate the name(s) of any staff person you would like to commend:

Comments:

If you feel we fell short in meeting your service expectations, please describe the situation, including name of the staff person involved and the date the incident occurred:

Example: Survey

A survey of this type, like an e-mail message, is "semi-structured."

- Exploit what is structured in interpreting and using the free text.
- Generally, textual source information doesn't come in without *some* form of envelope, of metadata that describes the information and its provenance.
- It's still hard to automate interpretation of the free text, that is, to do more than count words and note cooccurrence. Sentiment extraction comes into play.

Unified Analytics

Text analytics is good for…

Creating machine-exploitable models in/of information stores that were previously resistant to machine understanding,

Exploiting discovered or predefined structures to detect patterns: categories, linkages, etc.,

Applying the derived patterns to classify and support other automated processing according to document-extracted concepts and to establish relationships, and

Boosting traditional BI to create unified, 360° analytics.

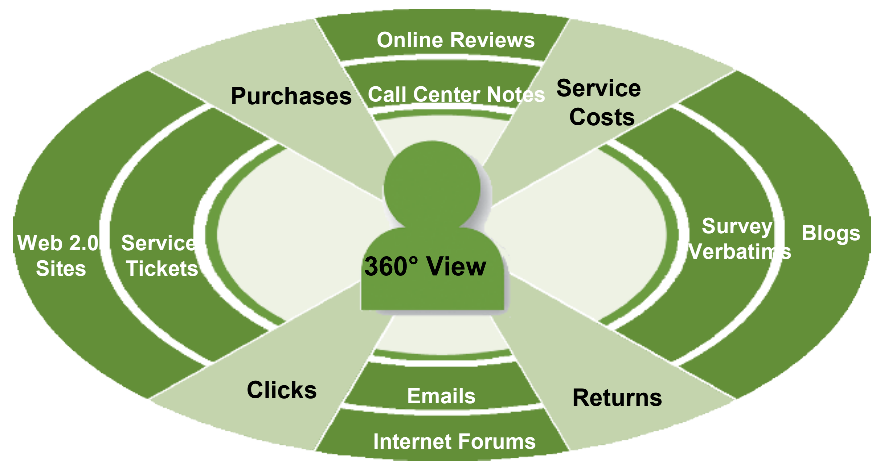
Unified Analytics

Why integrate analytics?

- 360º views.
- Single version of the truth.

Discussion questions:

- What's interoperability?
- What's integration?
- What's federation?
- How/what can you integrate?



Clarabridge's version: text + data

Alta Plana

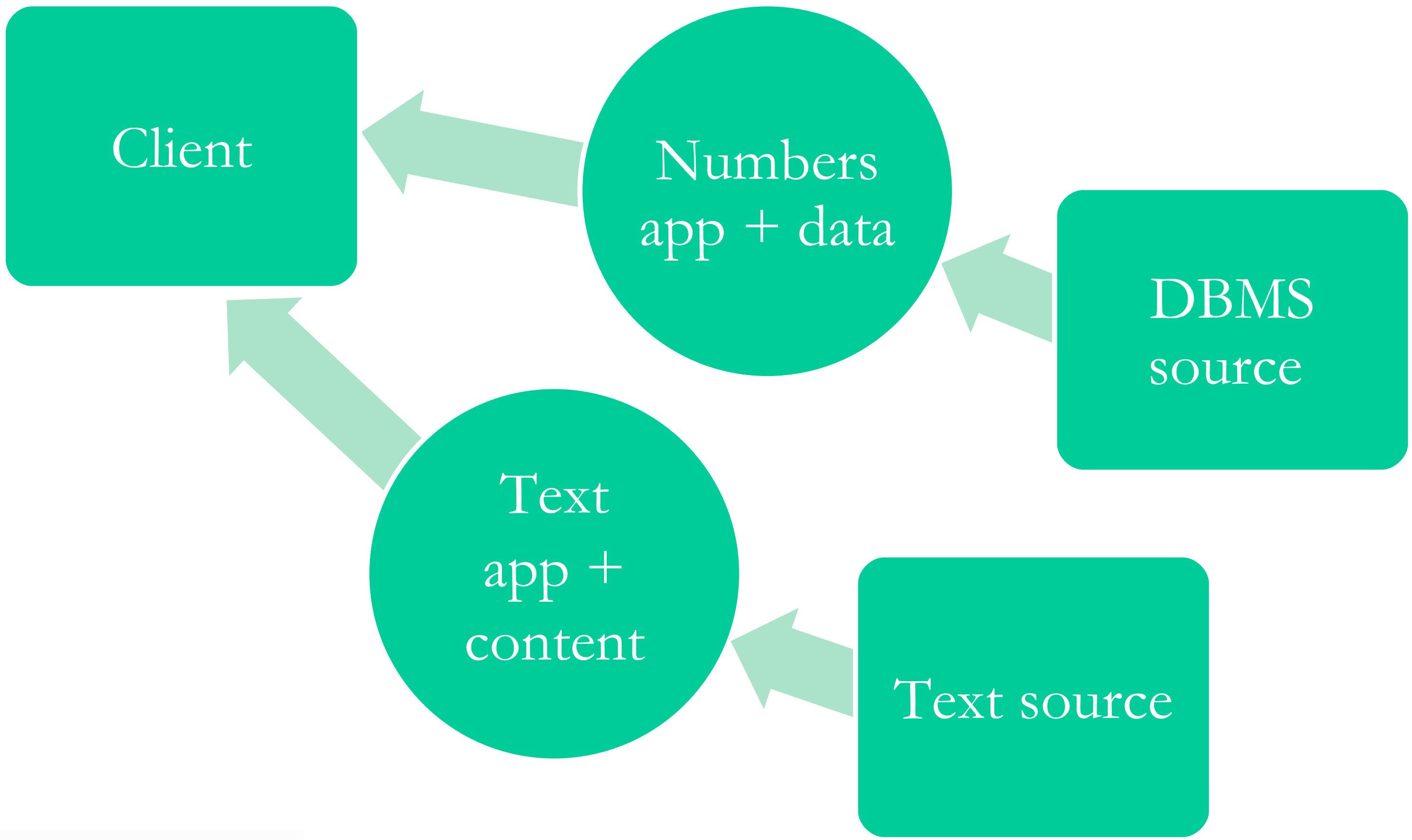
Unified Analytics

How/what can you integrate?

- Components, via some form of API or framework.
- Data, via defined, commonly understood formats and meanings.
  - What's the latter form of integration called?
- Business processes.
- Other resources including project teams.
- Standards play a major role.

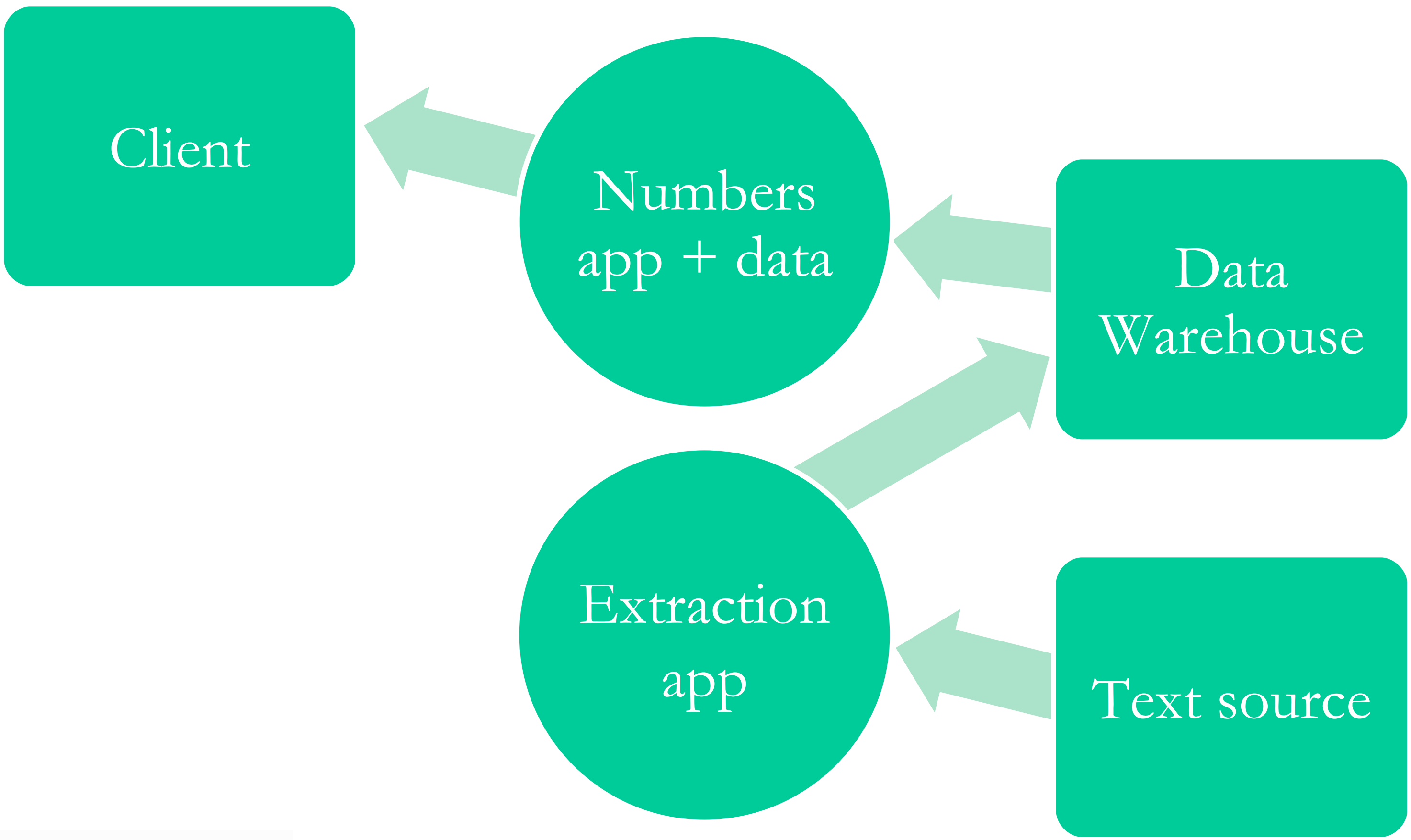
Unified Analytics

Unintegrated applications: not of interest here.



Unified Analytics

Information extraction and loading.



Unified Analytics

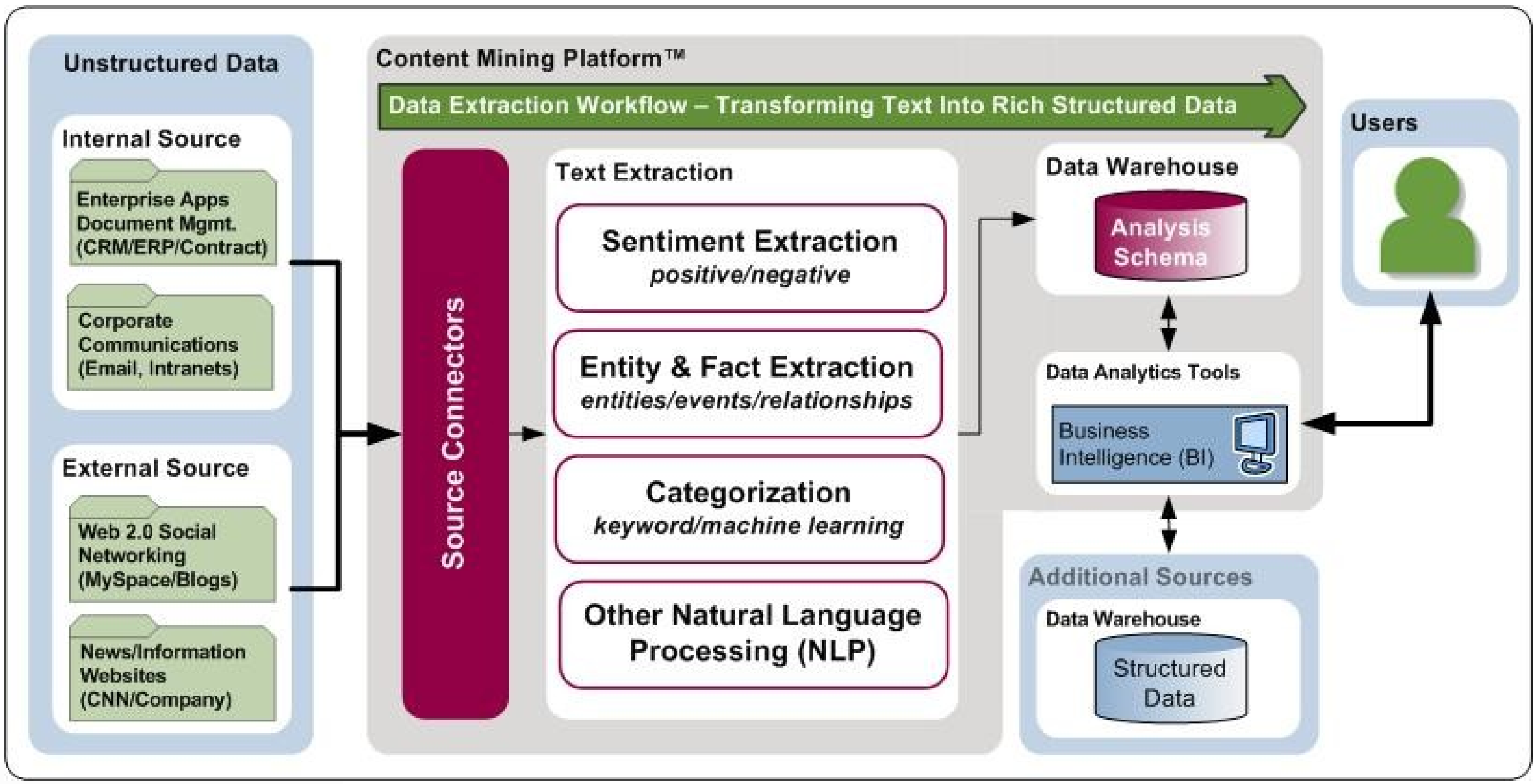
In information extraction for unified analytics, we do –

- Information retrieval, that is, locate source documents of interest.
- Identify relevant entities, concepts, and relationships.
- Extract them to appropriate DBMS structures.

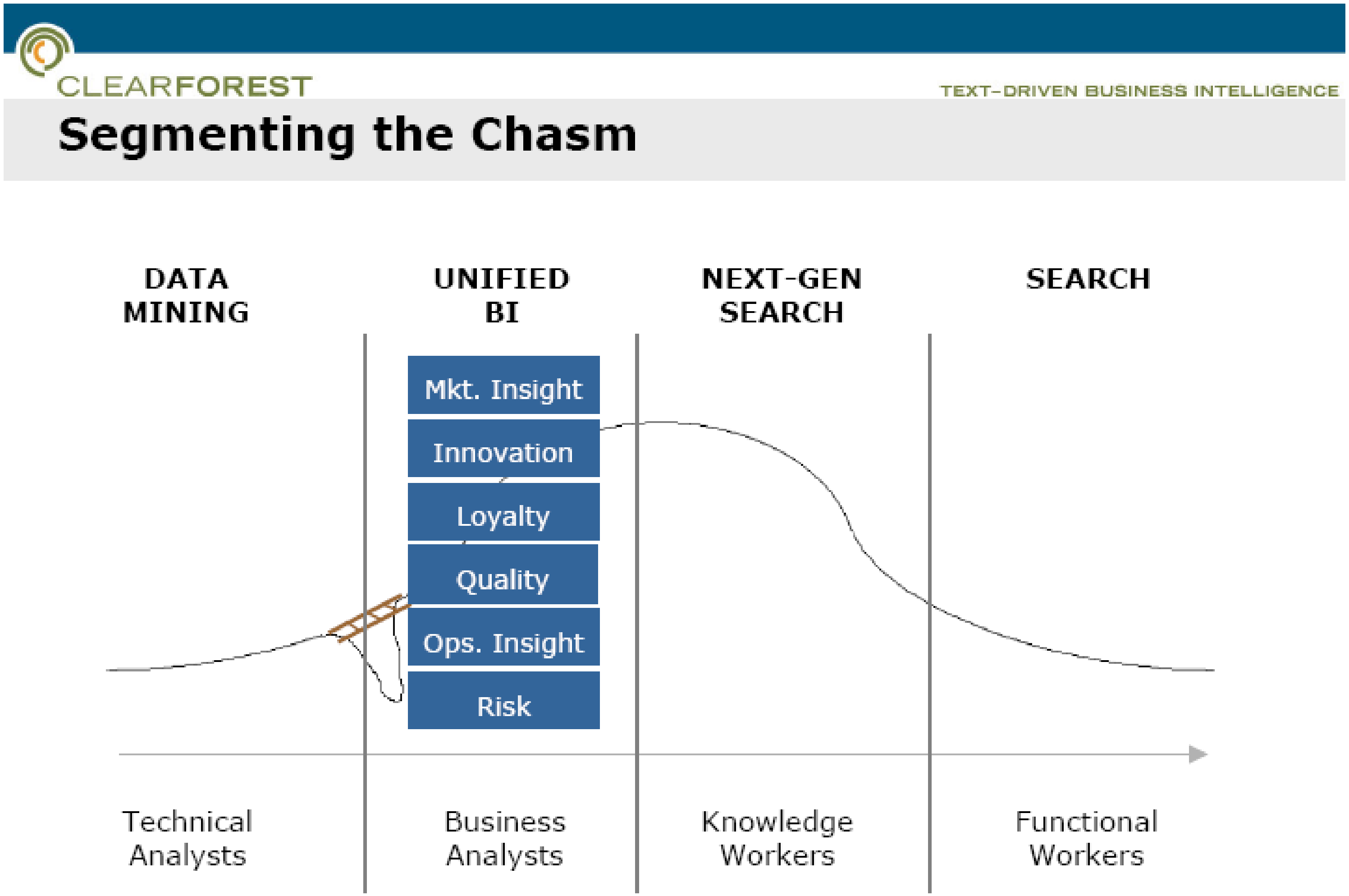
We need strong **semantic integration** that associates information that originated in disparate sources.

Unified Analytics

Clarabridge's Content Mining Platform implements this architecture –



Unified Analytics



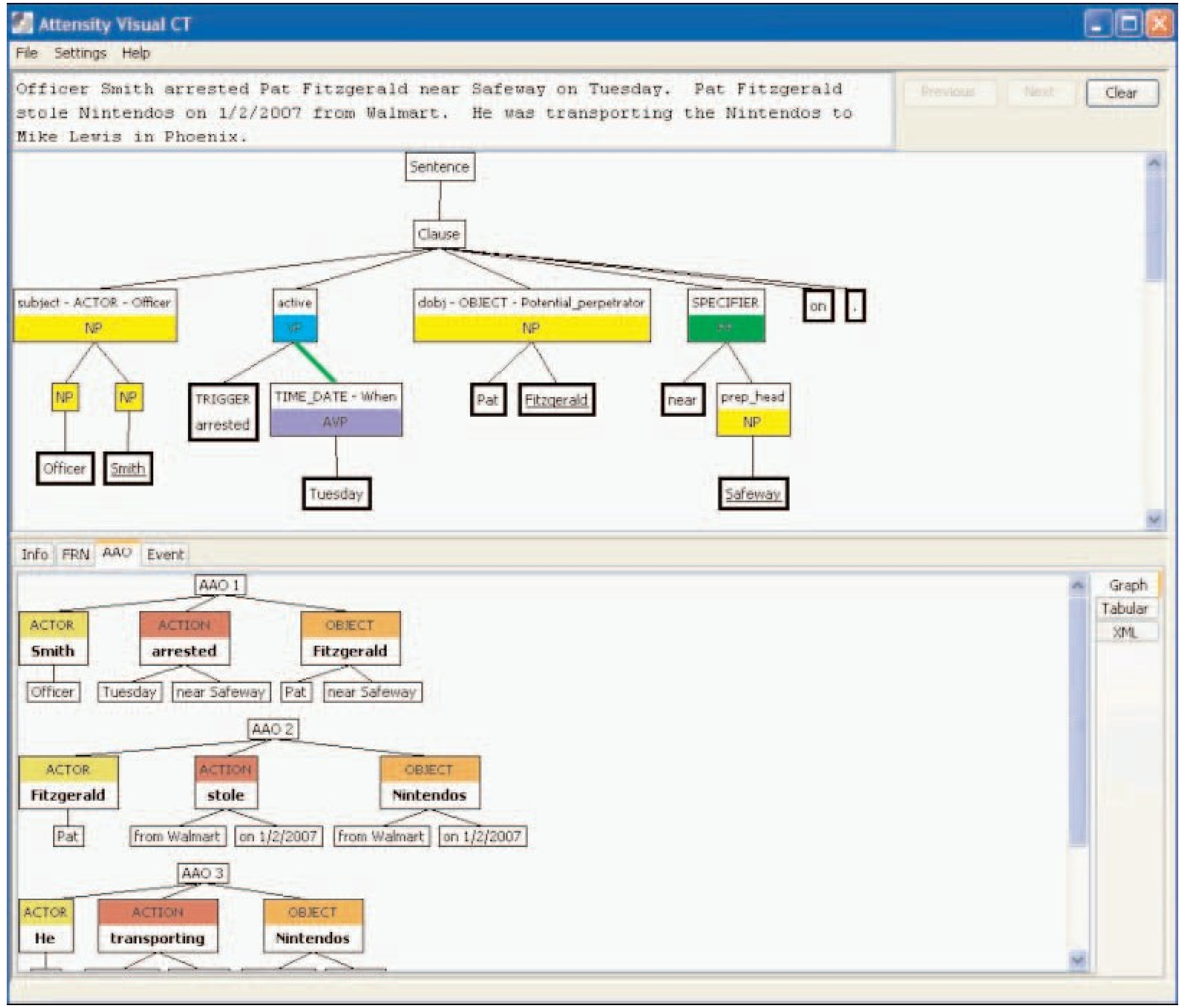
Applications

Law enforcement.

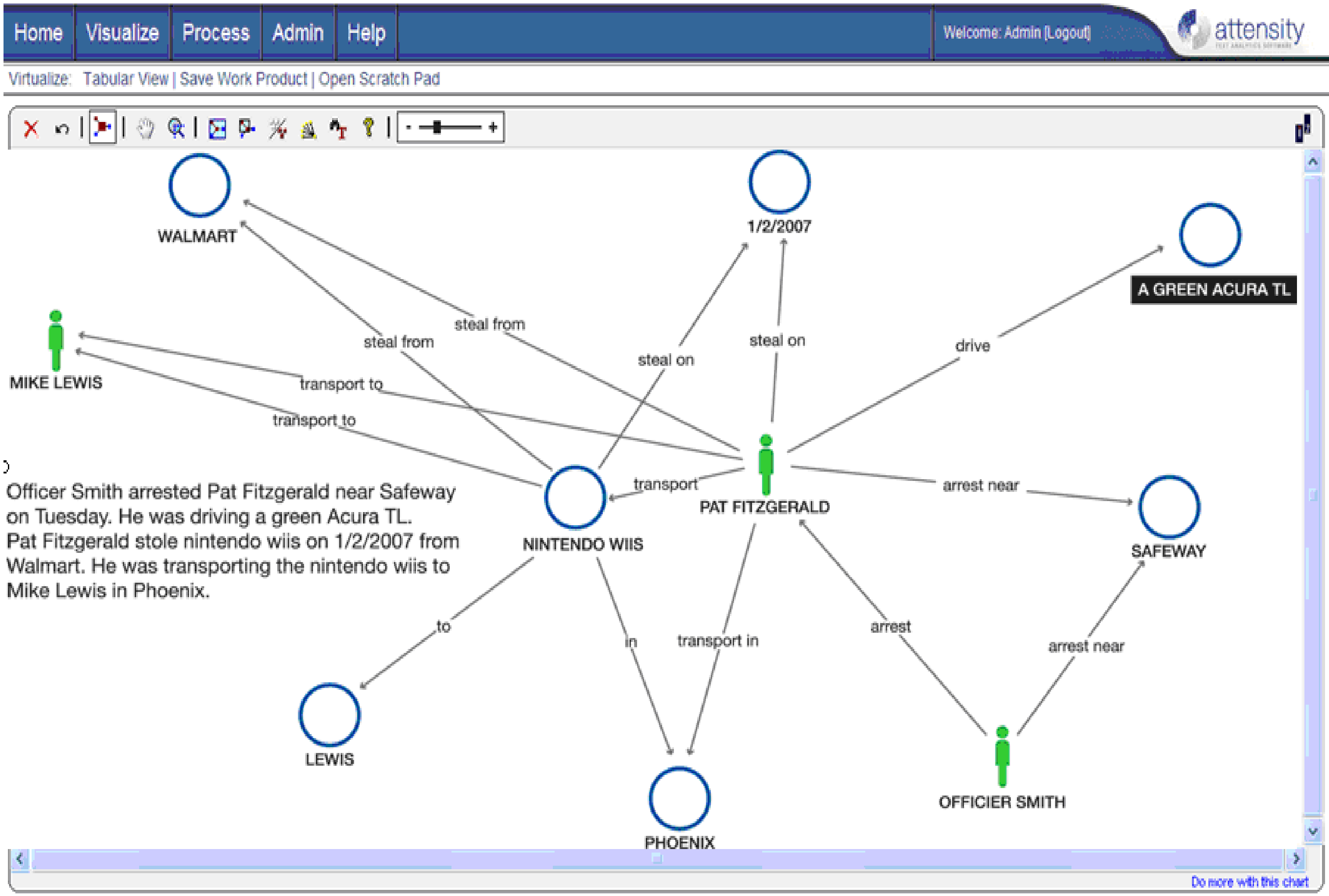
- Sources: case files, crime reports, incident and victimization databases, legal documents
- Targets: crime patterns, criminal investigation, networks

Applications

An Attensity law-enforcement example – NLP to identify roles and relationships.



Applications



Applications

Customer Relationship Management (CRM)

- Sources: customer e-mail, letters, contact centers
- Targets: product and service quality issues, product management, contact routing and CRM automation

Finance and compliance

- Sources: financial & news reports, corporate filings & documents, trading records
- Targets: insider trading, reporting irregularities, money laundering and illegal transactions, pricing anomalies

Applications

Health Care Case Management

- Sources: clinical research databases, patient records, insurance filings, regulations
- Targets: enhance diagnosis and treatment, promote quality of service, increase utilization, control costs

Intelligence and counter-terrorism

- Sources: news and investigative reports, communications intercepts, documents
- Targets: organization associations and networks, behavioral/attack patterns, strategy development

Questions?

Discussion?

Thanks!

Seth Grimes

Alta Plana Corporation

301-270-0795 – *http://altaplana.com*