Text Analytics for BI/DW Practitioners

Seth Grimes

Alta Plana Corporation

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

The Data Warehousing Institute

San Diego

August 20, 2008

Introduction

Seth Grimes –

Principal Consultant with Alta Plana Corporation.

Contributing Editor, *IntelligentEnterprise.com*.

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

Founding Chair, Text Analytics Summit, *textanalyticsnews.com*.

Instructor, The Data Warehousing Institute, *tdwi.org*.

I am not paid to promote any vendor.

Perspectives

Perspective #1: You're a business analyst or other "end user."

You have lots of text, and you want an automated way to deal with it.

Perspective #2: You work in IT.

You support end users who have lots of text.

Perspective #3: Other?

You just want to learn about text analytics.

Perspectives

Perspective #1a, 2a: Extending analysis.

You want to extend an existing business intelligence (BI) / data-mining initiative to encompass information from textual sources.

Perspective #1b, 2b: New to analysis.

You don't do traditional data analysis (yet).

Perspectives

What do people do with electronic documents?

1. Publish, Manage, and Archive.
2. Index and Search.
3. Categorize and Classify according to *metadata* & contents.
4. Information Extraction.

For textual documents, text analytics enhances #2 and enables #3 & #4.

Text analytics can be automated or interactive.

Agenda

Introduction (done).

The "Unstructured Data" Challenge.

Text Technologies.

Examples & Applications.

Best Practices.

The Market.

Key Message -- #1

If you are not analyzing text – if you're analyzing only transactional information – you're missing opportunity or incurring risk...

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

This is the "Unstructured Data" challenge

Key Message -- #2

Text analytics can boost business results...

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

...via established BI / data-mining programs, or independently.

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

Key Message -- #3

Some folks may need to expand their views of what BI and business analytics are about.

Others can do text analytics without worrying about BI.

Let's deal with text-BI first. Here are an image and a quotation from a 1958 paper introducing BI as a method for processing documents and extracting knowledge...

Document input and processing

Knowledge handling is key

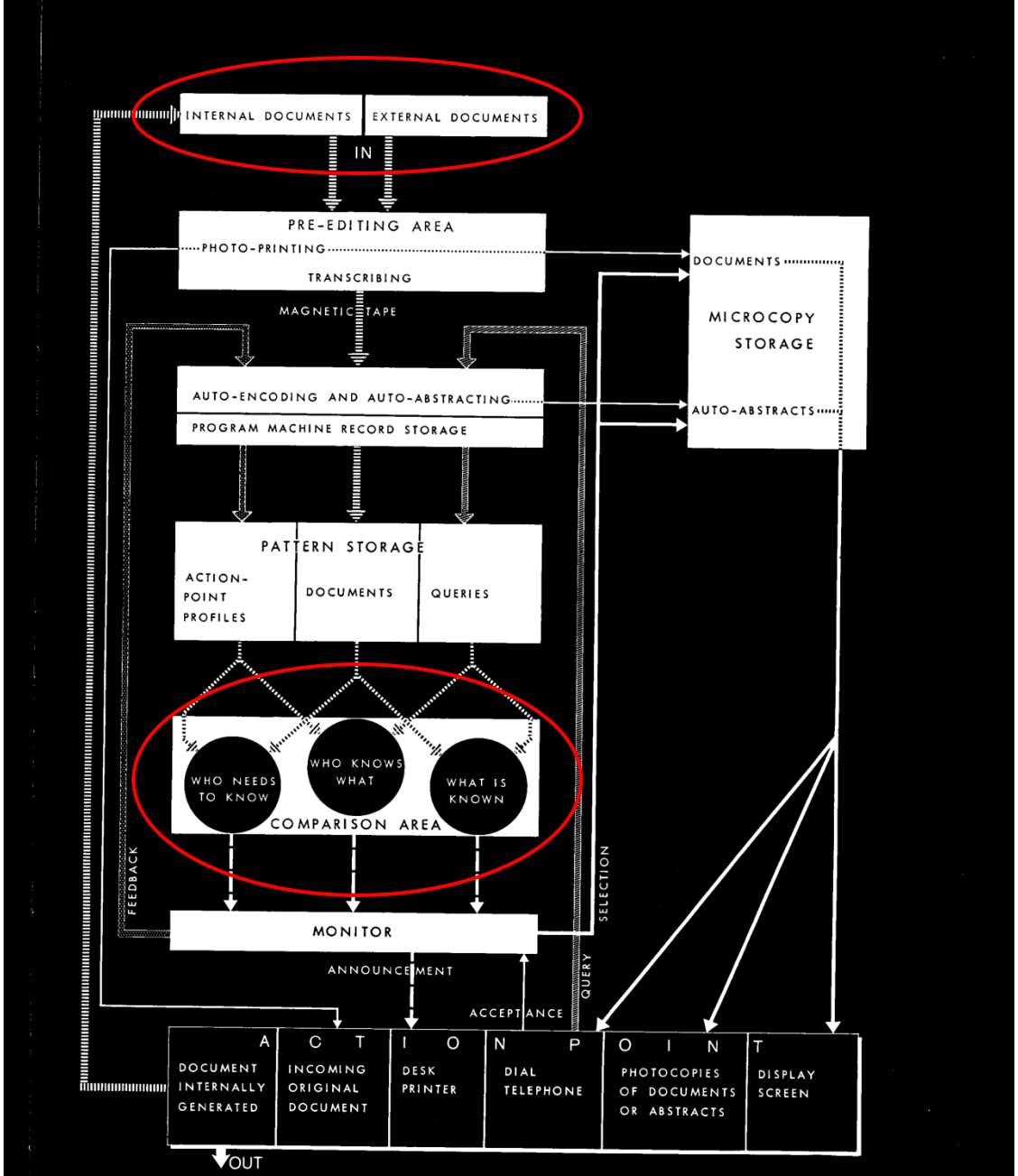


Figure 1 **A Business Intelligence System**

Text-BI: *Back to the Future*

What is business intelligence (BI)? A 1958 definition, based on processing documents:

> 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 notion of intelligence** is also defined here... 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*

Why was BI redefined as work on DBs?

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)*

BI operates on data in relational tables that originated in transactional systems.

Yet it's a truism that 80% of enterprise information is in "unstructured" form.

The "Unstructured Data" Challenge

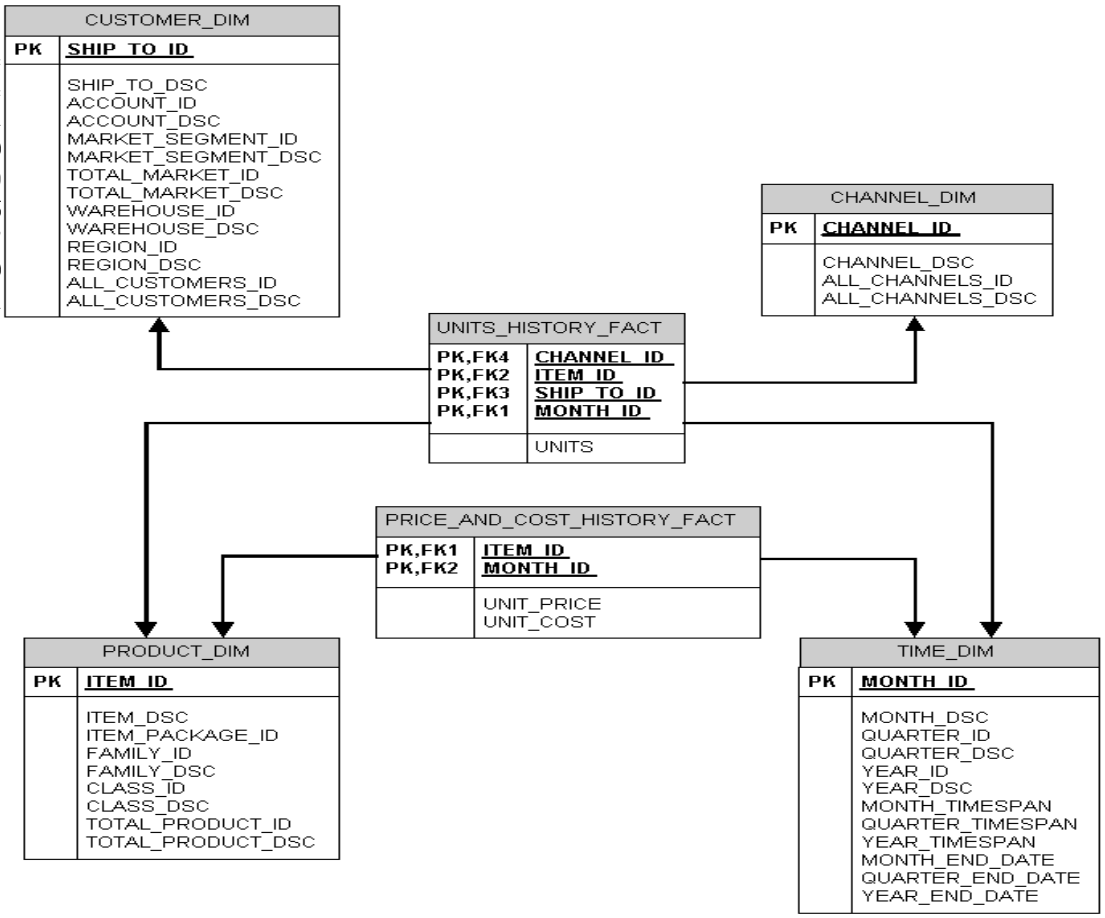
Traditional BI feeds off:

```
"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:

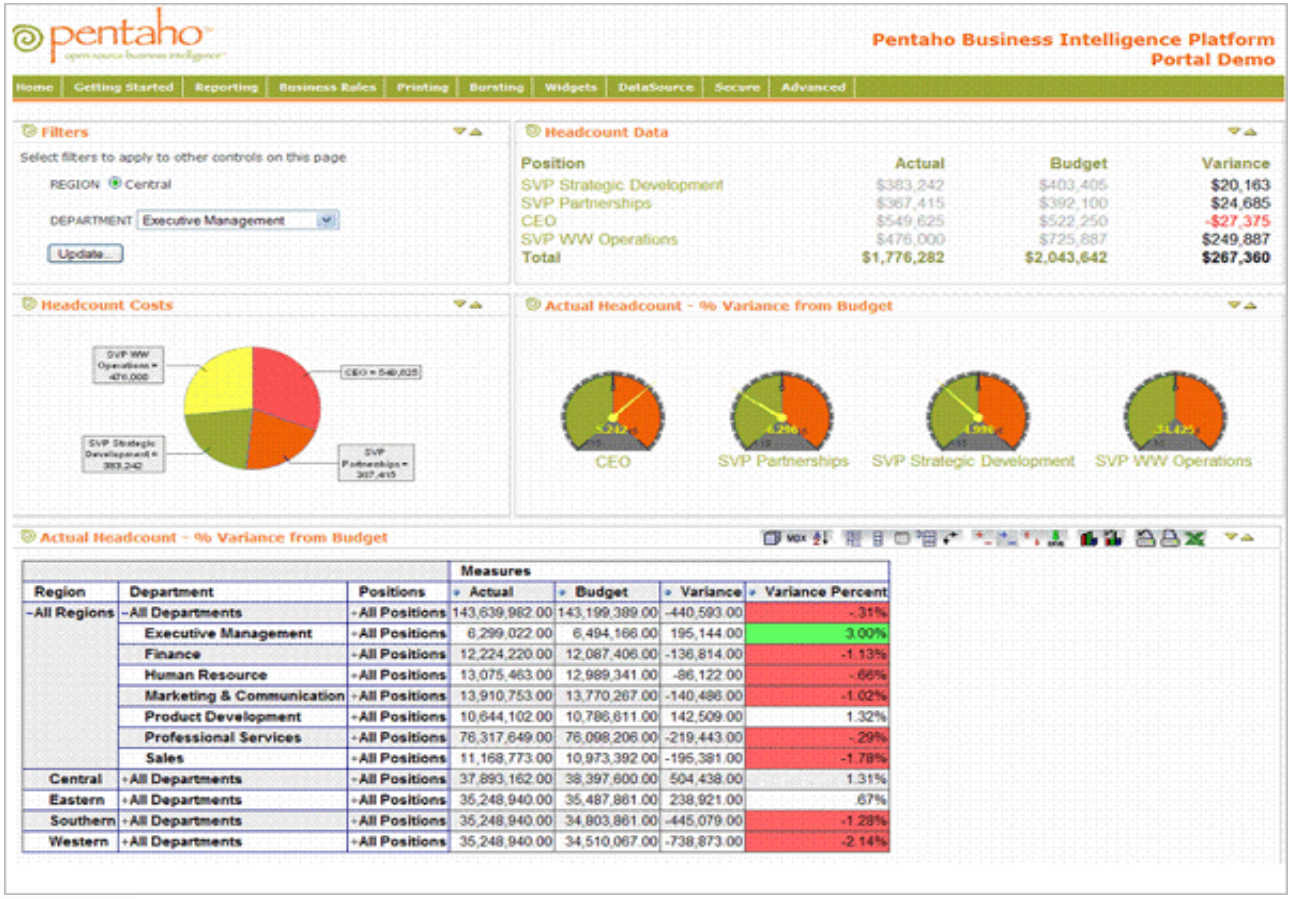
```
"SUMLEV","STATE","COUNTY","STNAME",
50,19,1,"Iowa","Adair County",1,824
50,19,1,"Iowa","Adair County",2,824
50,19,1,"Iowa","Adair County",3,821
50,19,1,"Iowa","Adair County",4,809
50,19,1,"Iowa","Adair County",5,800
50,19,1,"Iowa","Adair County",6,796
50,19,1,"Iowa","Adair County",7,787
50,19,1,"Iowa","Adair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



It runs off:

The "Unstructured Data" Challenge

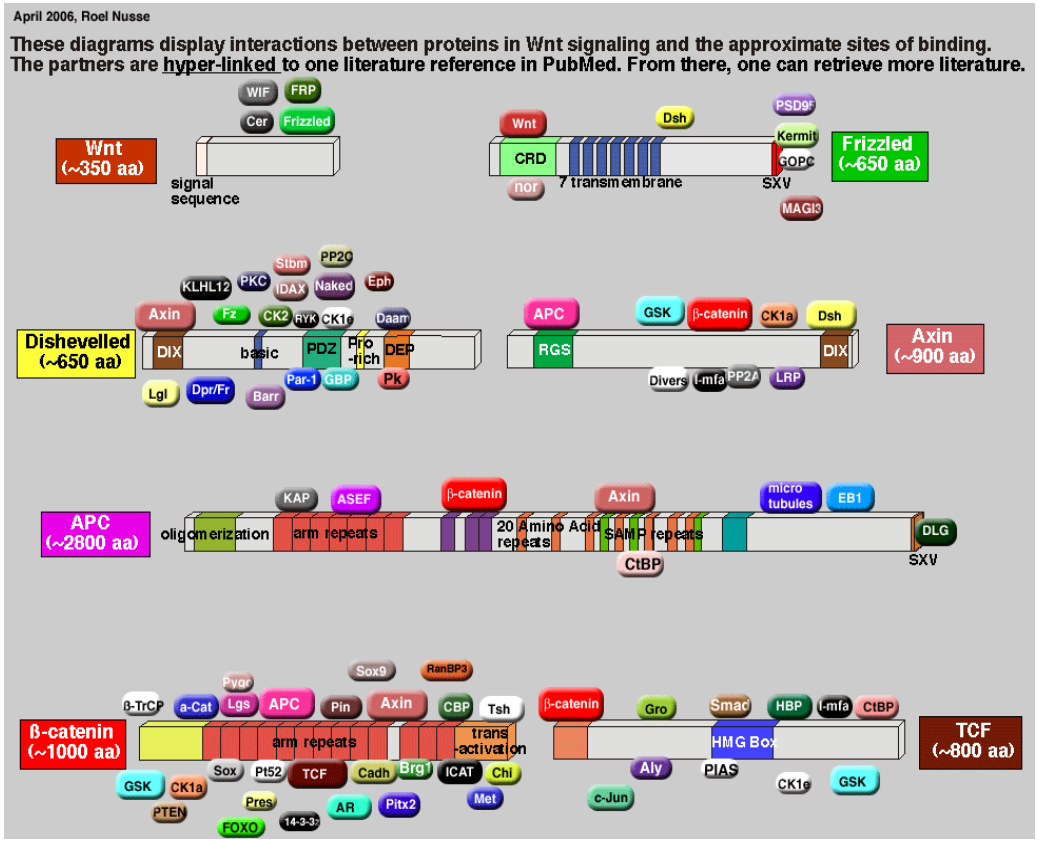
Traditional BI produces:



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

What do you do when your source information looks like this?

When you walk in the foyer of the hotel it seems quite inviting but the room was very basis and smelt very badly of stale cigarette smoke, it would have been nice to be asked if we wanted a non smoking room, I know the room was very cheap but I found this very off putting to have to sleep with the smell, and it was to cold to leave the window open. Excellent location for restaurants and bars

Overall I would never sell/buy a Motorola V3 unless it is demanded. My life would be way better without this phone being around (I am being 100% serious) Motorola should pay me directly for all the problems I have had with these phones. :-(

The "Unstructured Data" Challenge

Consider again –

- The purpose of intelligence is to "guide action towards a desired goal." (Luhn)
- "Industries such as travel and hospitality and retail live and die on customer experience." (Banerjee)

Exercise...

The "Unstructured Data" Challenge

Rereading the first text, what 1) goals and 2) useful information do you see? 3) How might you "structure" that information?

When you walk in the foyer of the hotel it seems quite inviting but the room was very basis and smelt very badly of stale cigarette smoke, it would have been nice to be asked if we wanted a non smoking room, I know the room was very cheap but I found this very off putting to have to sleep with the smell, and it was to cold to leave the window open. Excellent location for restaurants and bars

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.

The "Unstructured Data" Challenge

Exercise: Organize the information in this paragraph into a table –

> 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.

What is the primary key? Is there a derived field?

Is there another way to represent this data?

The "Unstructured Data" Challenge

How about a (fabricated) XML representation –

```
<?xml version="1.0"?>
<rdf:RDF xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#"
  xmlns:market="http://www.markets.fake/market#">
    <rdf:Description rdf:about="http://www.dow.fake/market/Dow">
     <market:market>Dow</market:market>
     <market:change>-46.58</market:change>
     <market:index>11002.14</market:index>
     <market:date>20061201</market:date>
    </rdf:Description>
    <rdf:Description rdf:about="http://www.SandP.fake/market/SandP">
     <market:market>Standard &amp;amp; Poor&amp;apos;s</market:market>
     <market:change>-1.44</market:change>
     <market:index>1263.85</market:index>
     <market:date>20061201</market:date>
    </rdf:Description>
. . .
</rdf:RDF>
```

Who can tell us about the mark-up here?

Search

So there's data and other interesting information in text. How do we get at it?

Search is not the answer. It returns documents.

Analysts want facts, answers to questions.

And what if you're unsure what question to ask?

All the same, let's think about searches and answers...

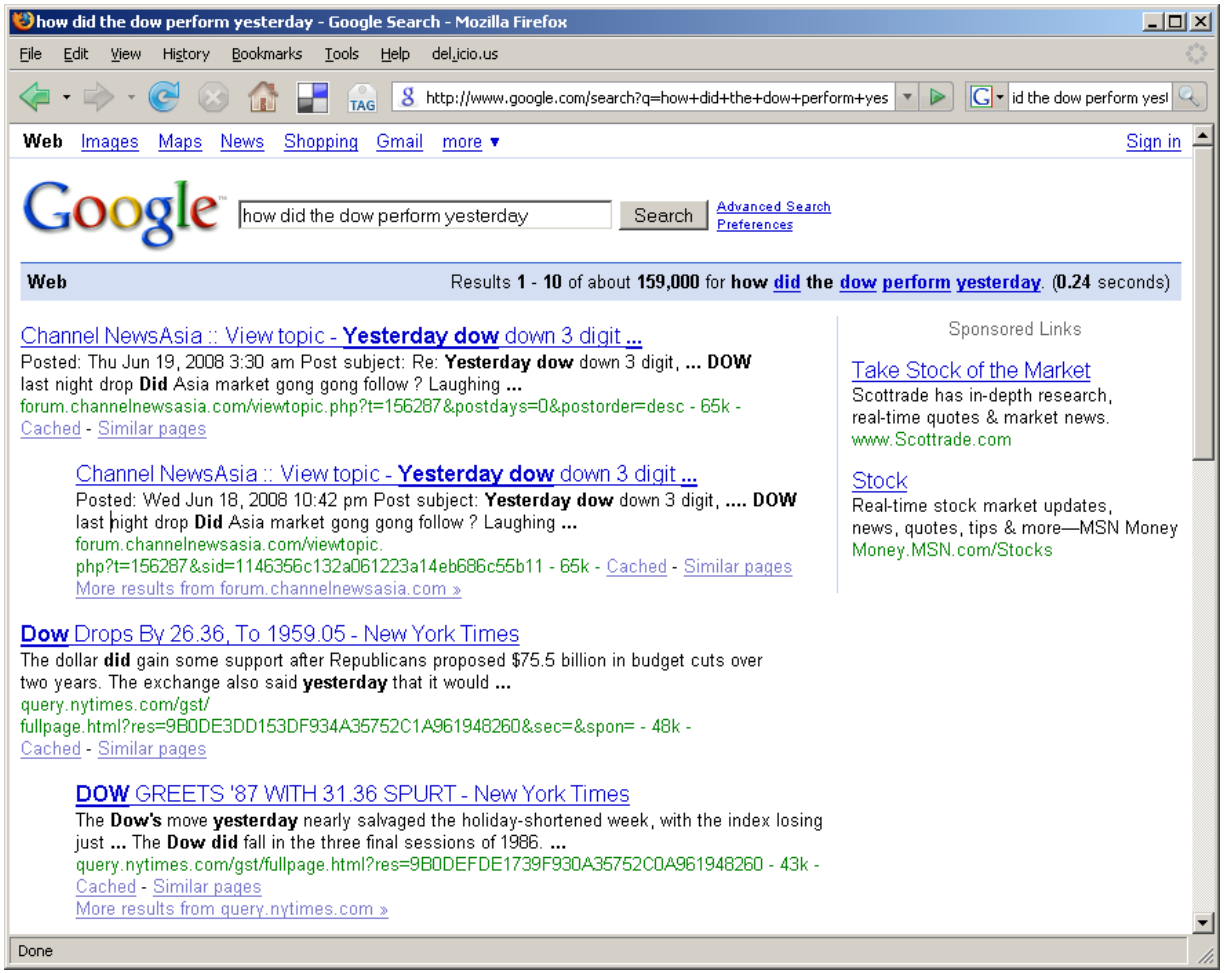
Search

Relevance?

Concepts?

Articles from a forum site

Articles from 1987



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

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

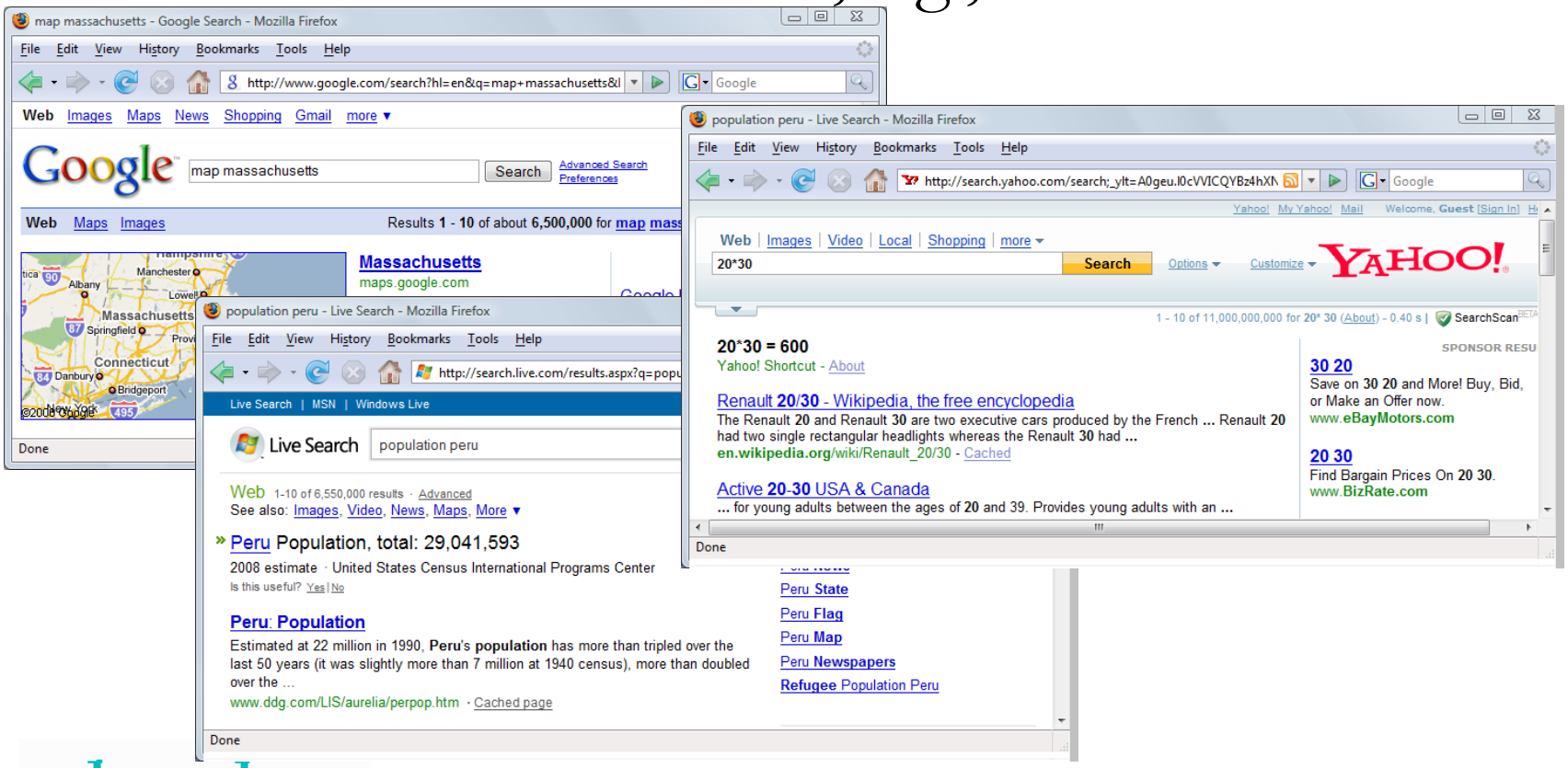
Search is not enough.

- *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 doesn't enable **unified analytics** that links data from textual and transactional sources.*

Text analytics can make search better...

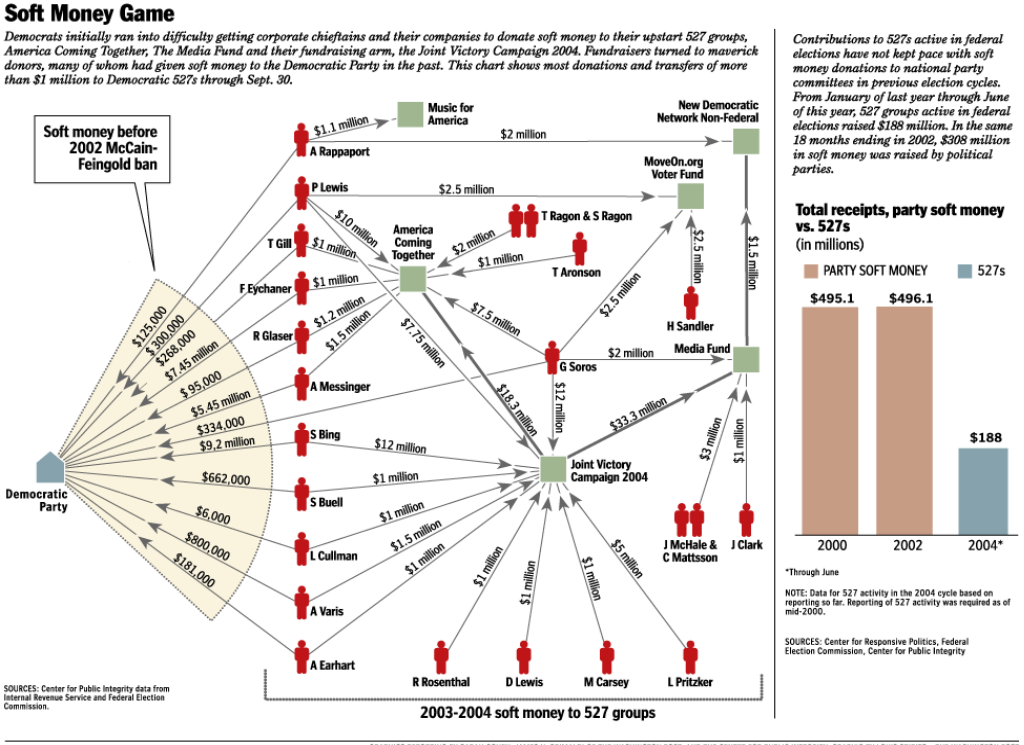
Beyond Search: Analysis

Text analytics enables results that suit the information and the user, e.g., answers –



Beyond Search: Analysis

Now on to knowledge discovery, to discerning ***interrelationships of presented facts...***



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

Beyond Search: Analysis

Exercise: Association rules.

Do you have a child or children living at home?

Do you live in an apartment or a house?

Are there statistically significant correlations?

What rule can we derive?

	Yes	No
Apt.		
House		

Exercise: Link analysis, discovery & search.

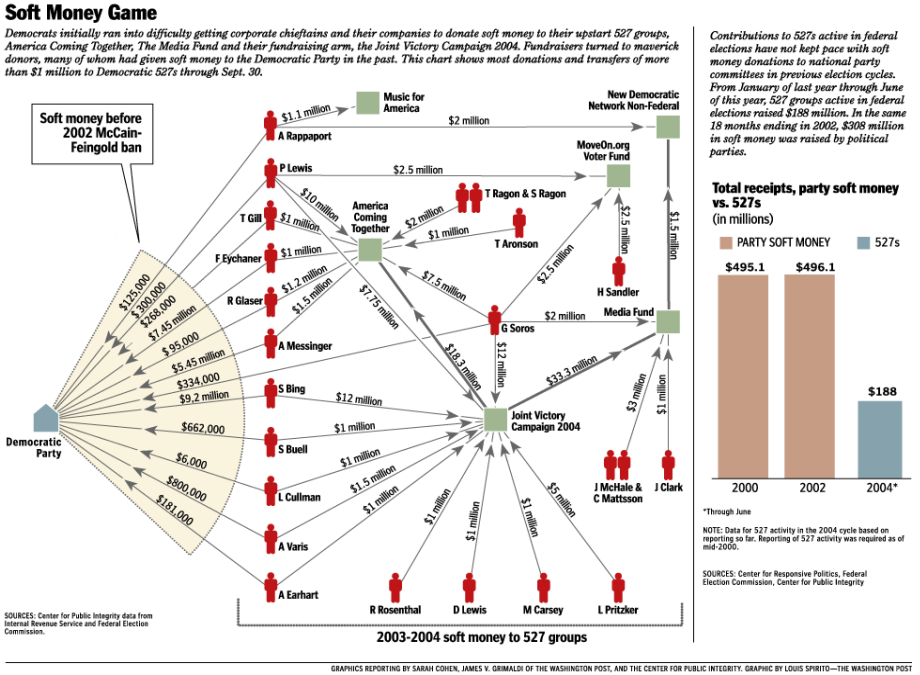
Find something you and a person next to you have in common, e.g., school attended, the industry you work in, favorite sport, other.

(Next step is mine.)

Beyond Search: Analysis

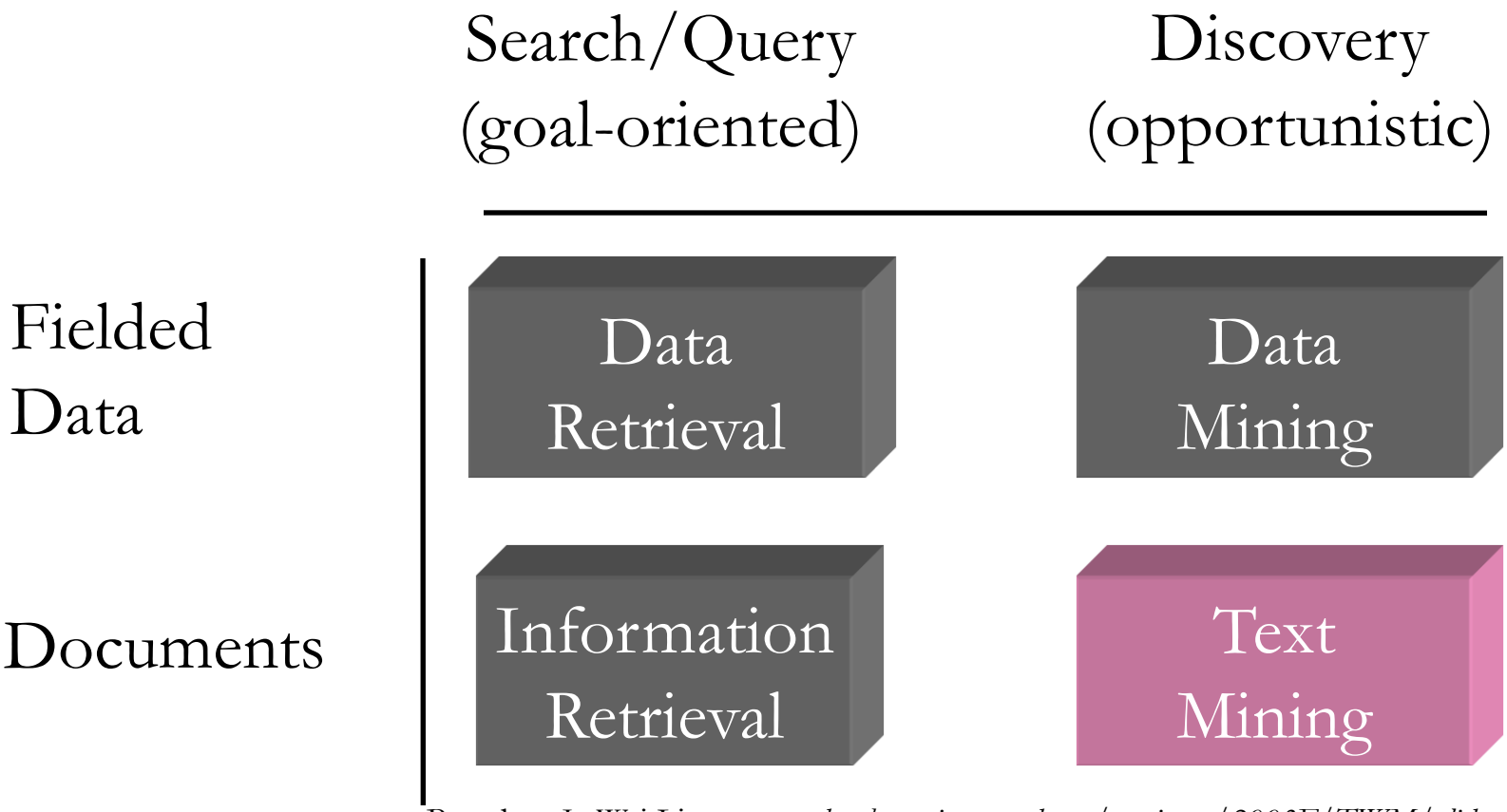
Text Mining = Data Mining of textual sources.

- Clustering and Classification.
- Link Analysis.
- Association Rules.
- Predictive Modelling.
- ~~Regression.~~
- ~~Forecasting.~~



Text Mining = Knowledge Discovery in Text.

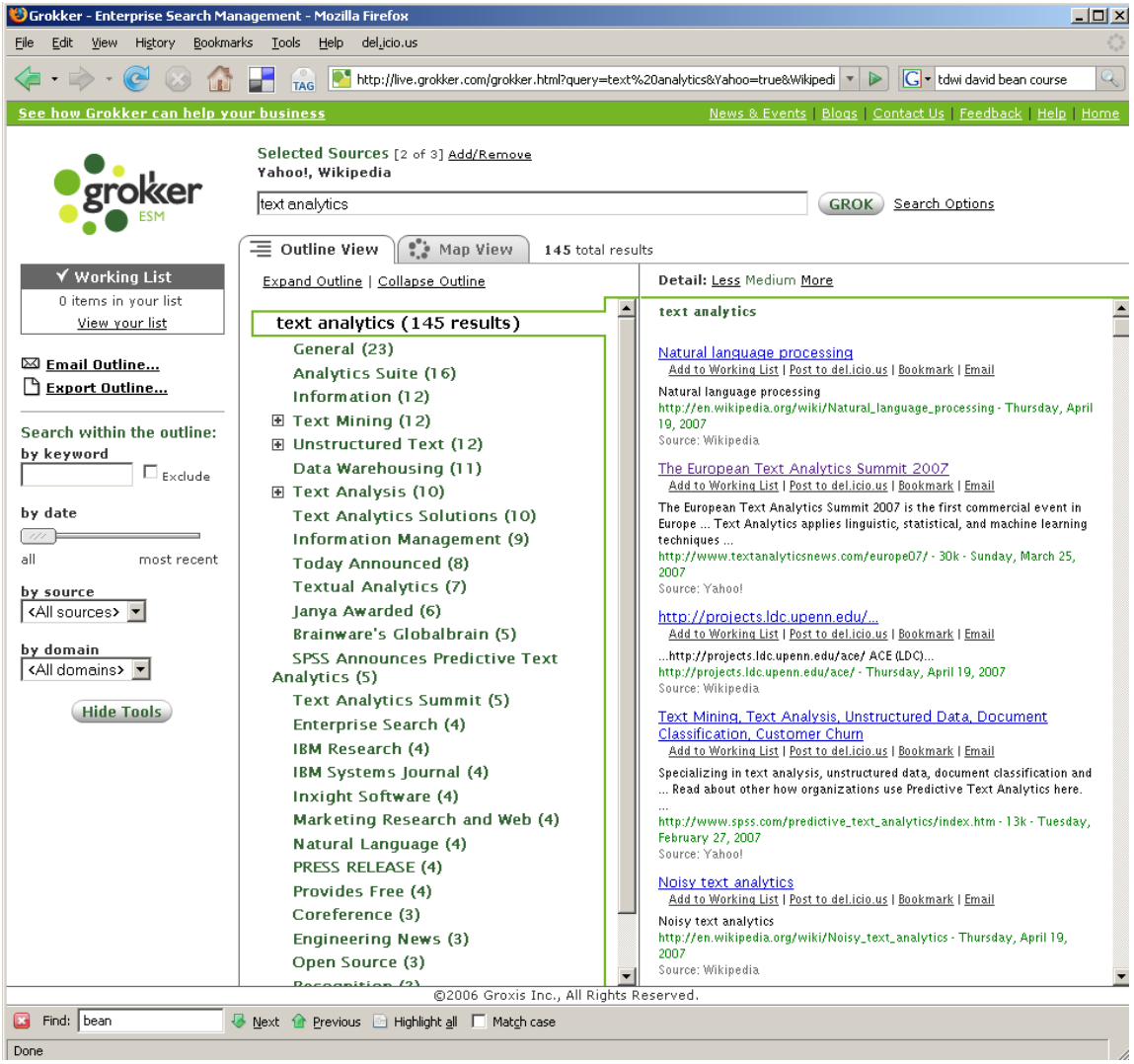
Text Mining



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

Presentation of search results can be enhanced by discovery.

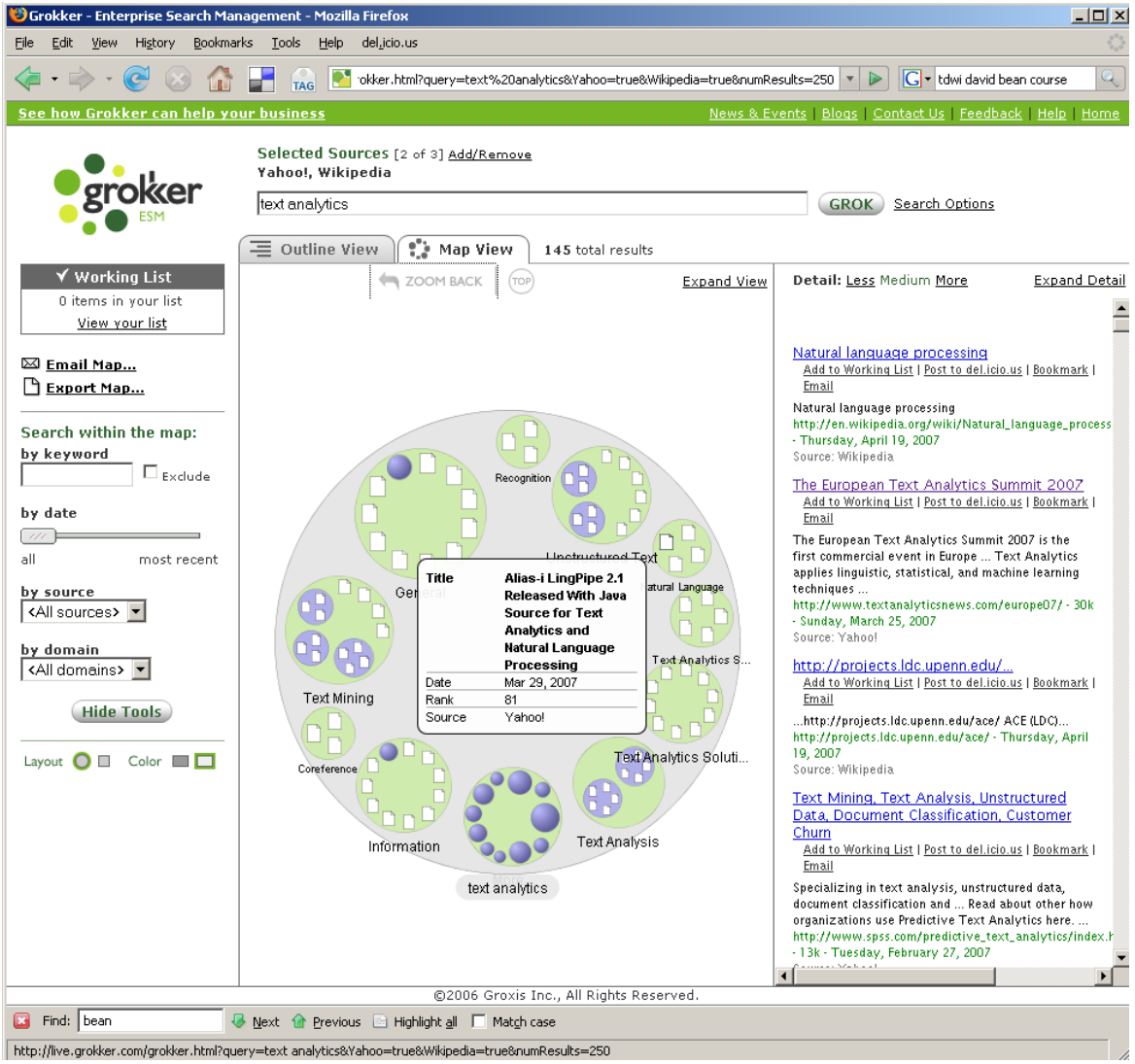
This slide and the next show dynamic, clustered search results from Grokker…



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

…with a zoomable display.

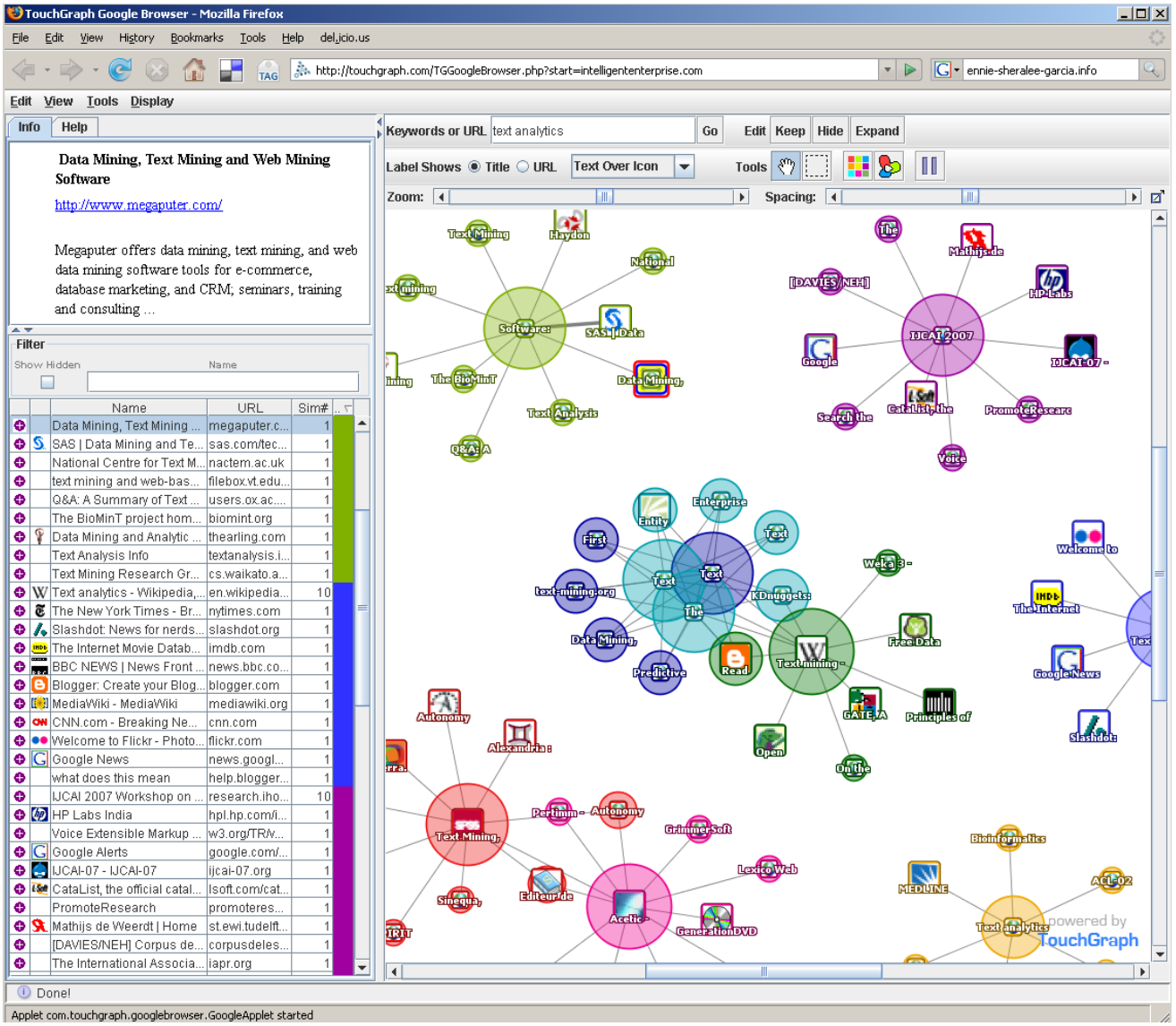
Clustering here utilizes statistical (text) data mining techniques to identifying cohesive groupings of retrieved documents.



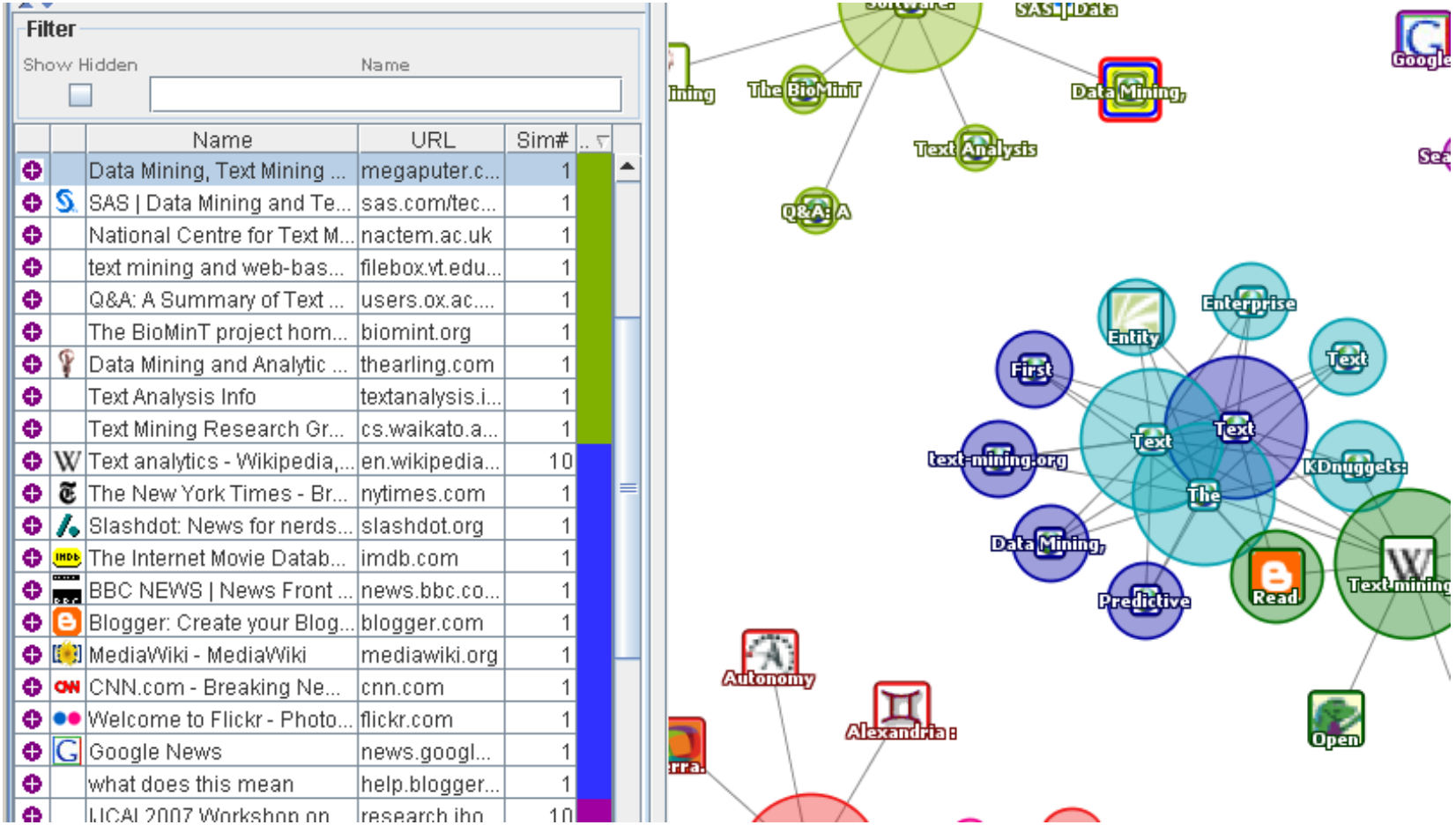
More results clustering...

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

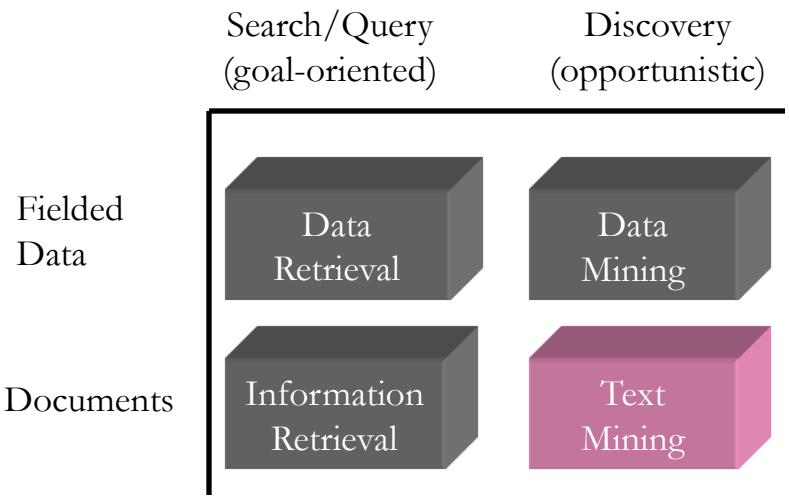
touchgraph.com/TGGoogleBrowser.php?start=text%20analytics

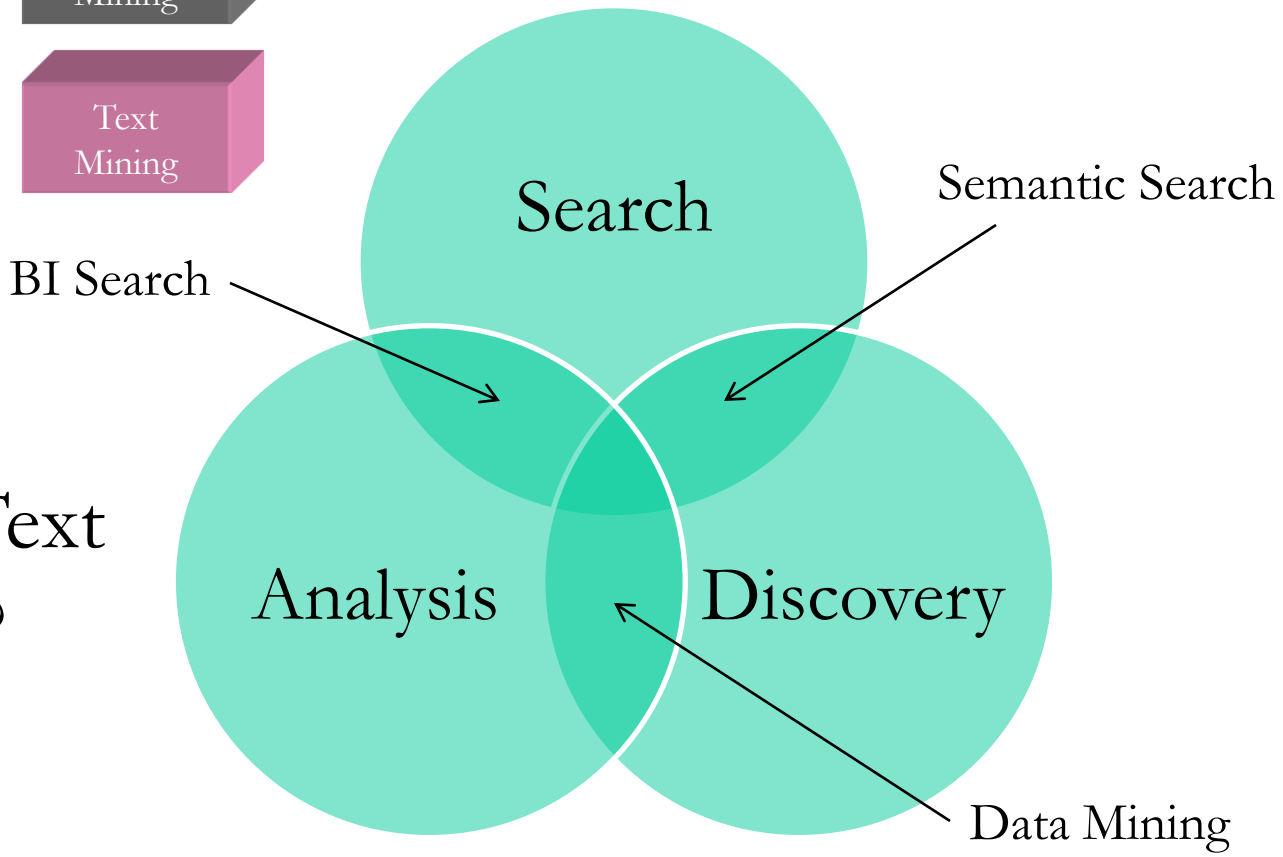


Beyond Search: Analysis



Text Analytics





Where's Text Analytics?

Alta Plana

Text Analytics

So text analytics enhances results of search, a.k.a. Information Retrieval (IR).

- It recognizes patterns and "named entities" in search queries to enable basic question answering.
- It recognizes patterns in search results to enable clustering and classification of results.

We want to get beyond IR to Information Extraction (IE).

First, *time out* to summarize and provide some definitions...

Glossary

Text analytics automates what researchers, writers, scholars, and all the rest of us have been doing for years. Text analytics –

- ***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.*
- ***Unlocks meaning and relationships*** *in large volumes of information that were previously unprocessable by computer.*

Glossary

Text Analytics is perhaps a superset of ***Text Mining.***

Information Extraction (IE) involves pulling features – entities & their attributes, facts, relationships, etc. – out of textual sources.

Entity: Typically a name (person, place, organization, etc.) or a patterned composite (phone number, e-mail address).

Concept: An abstract entity or collection of entities.

Co-reference: Multiple expressions that describe the same thing.

Fact: A relationship between two entities.

Sentiment: A valuation at the entity or higher level.

Opinion: A fact that involves a sentiment.

Glossary

Semantics: A fancy word for meaning, as distinct from ***Syntax,*** which is structuring.

Natural Language Processing (NLP): Computers hear humans.

Parsing: Evaluating the contents of a document.

Tokenization: Identification of distinct elements within a text.

Stemming/Lemmatization: Reducing variants of word bases created by conjugation, declension, case, pluralization, etc.

Tagging: Wrapping XML tags around distinct text elements, a.k.a. ***text augmentation***.

POS Tagging: Specifically identifying parts of speech.

Glossary

Categorization: Specification of ways like items can be grouped.

Clustering: Creating categories according to statistical criteria.

Taxonomy: An exhaustive, hierarchical categorization of entities and concepts, either specified or generated by clustering.

Classification: Assigning an item to a category, perhaps using a taxonomy.

Ontology : In practice, a classification of a set of items in a way that represents knowledge, e.g., Assigning an item to a category, perhaps using a taxonomy.

> *A rose is a flower. A deer is an animal. A sparrow is a bird. Russia is our fatherland. Death is inevitable.*
>
> -- *P. Smirnovskii*, A Textbook of Russian Grammar

Glossary

Precision: The proportion of decisions (e.g., classifications) that are correct.

Recall: The proportion of actual correct decisions (e.g., classifications) relative to the total number of correct decisions.

Find the even numbers:

9 17 12 4 1 6 2 20 7 3 8 10

Exercise: What is my Precision? What is my Recall?

Accuracy: How well an IE or IR task has been performed, computed as an ***F-score*** weighting ***Precision*** & ***Recall***, typically:

```
f = 2*(precision * recall) / (precision + recall)
```

Text Analytics

Typical steps in text analytics include –

Retrieve documents for analysis.

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

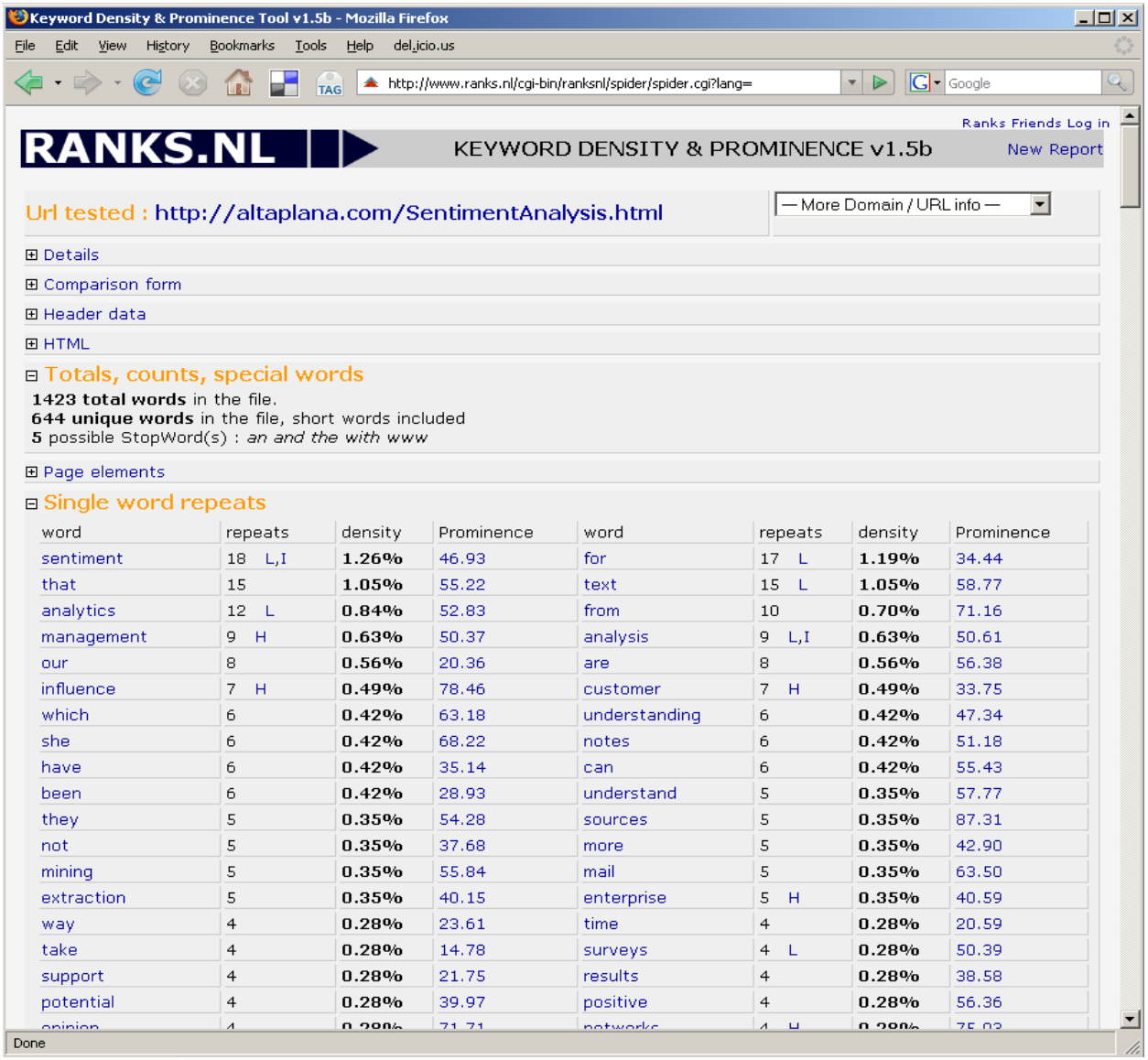
Apply statistical pattern-matching & similarity techniques to **classify** documents and organize extracted features according to a specified or generated categorization / taxonomy.

– via a *pipeline* of statistical & linguistic steps.

Text Analytics

So text analytics looks for structure that is inherent in documents, the textual source materials. Let's look at some of the steps.

First, we'll do a lexical analysis of a text file, essentially a basic statistical analysis of the words and multi-word terms...



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

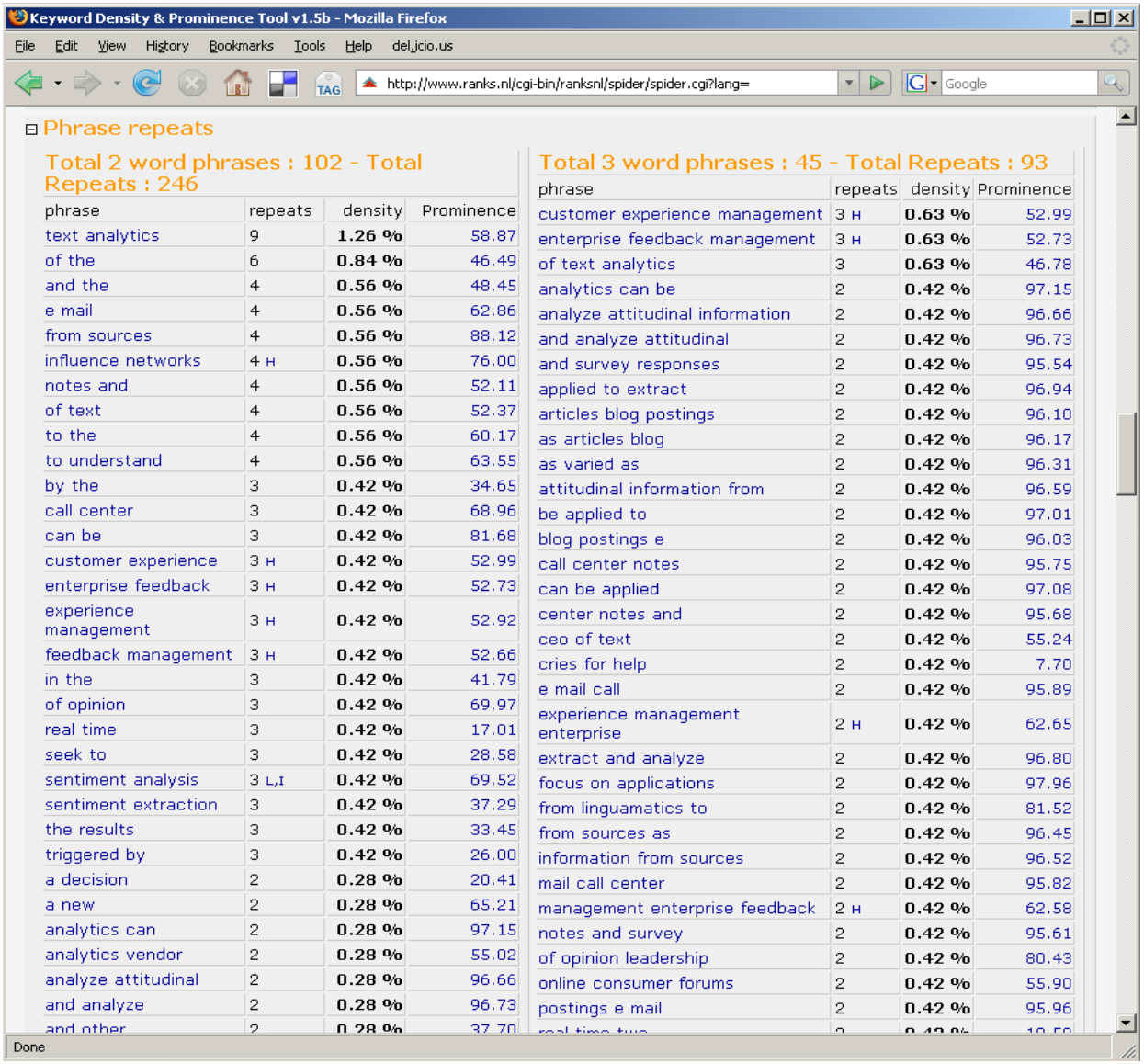
1423 total words in the file.
644 unique words in the file, short words included
5 possible StopWord(s) : *an and the with www*

⊞ 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



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
real time two	2	**0.42 %**	18.58

Text Analytics

Those "tri-grams" are pretty good at describing the *Whatness* of the source text.

Lesson: "Structure" may not matter.

- Shallow parsing and statistical analysis can be enough, for instance, to support classification. (But that's not BI.)
- It can help you get at meaning, for instance, by studying co-occurrence of terms.
- Yet something is missing. What? (Hint: It's defined on p. 36.)

Statistical pattern matching – the bag/vector of words approach – may fall short.

The Need for Linguistics

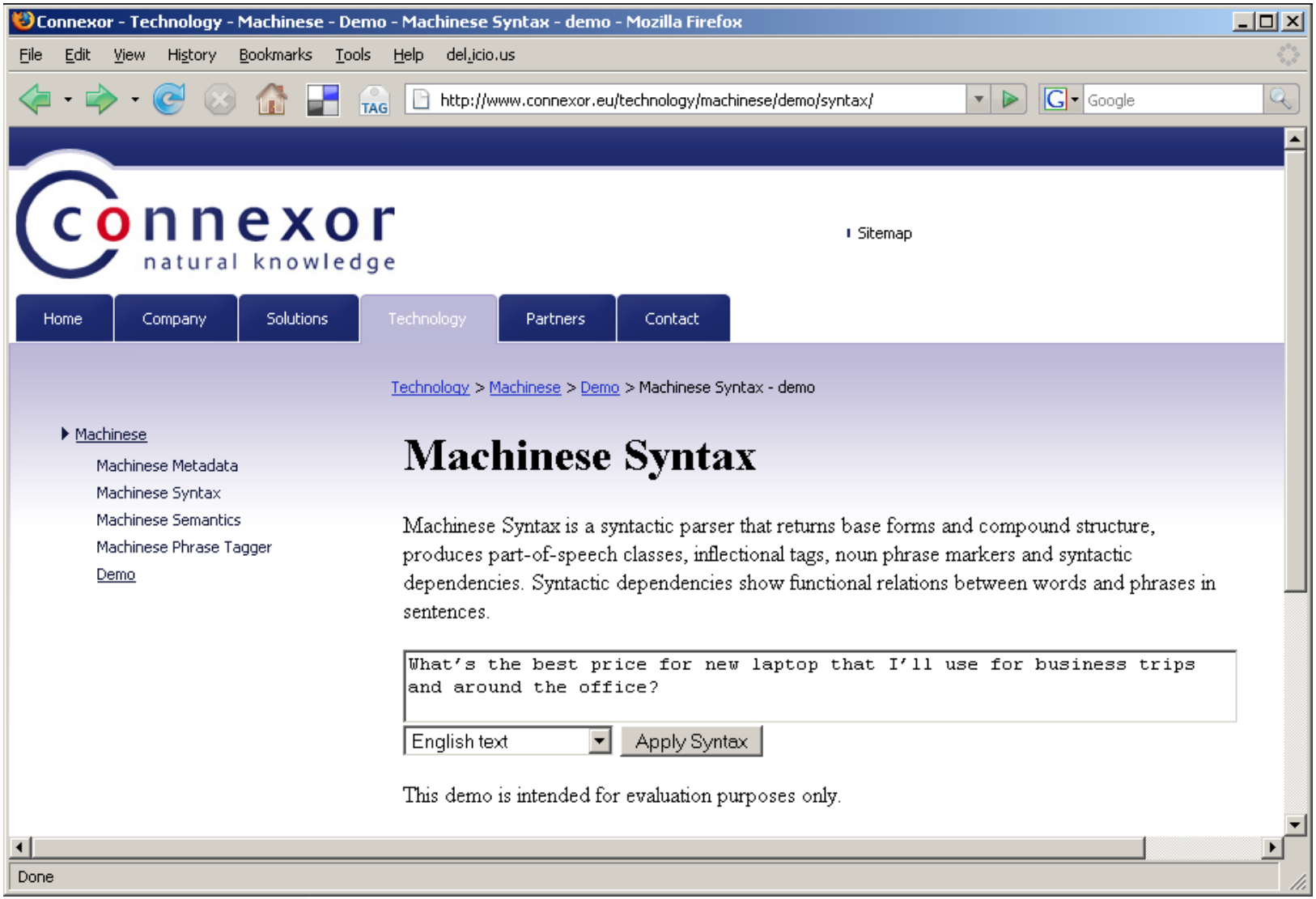
Consider –

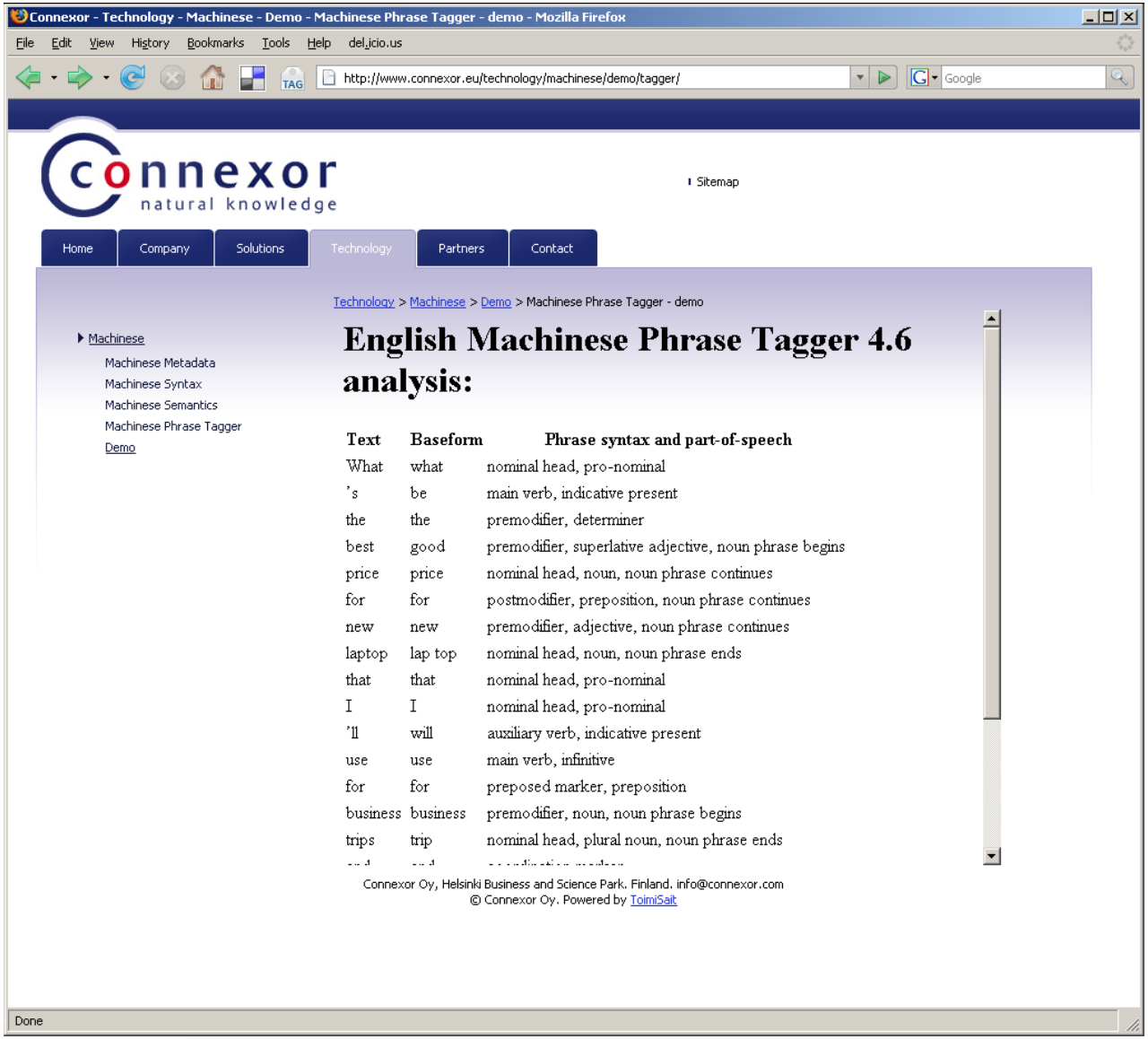
The Dow *fell* 46.58, or 0.42 percent, to 11,002.14. The Standard & Poor's 500 index fell 1.44, or 0.11 percent, to 1,263.85, and the Nasdaq composite *gained* 6.84, or 0.32 percent, to 2,162.78.

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.

Let's try syntactic analysis of a bit of text...

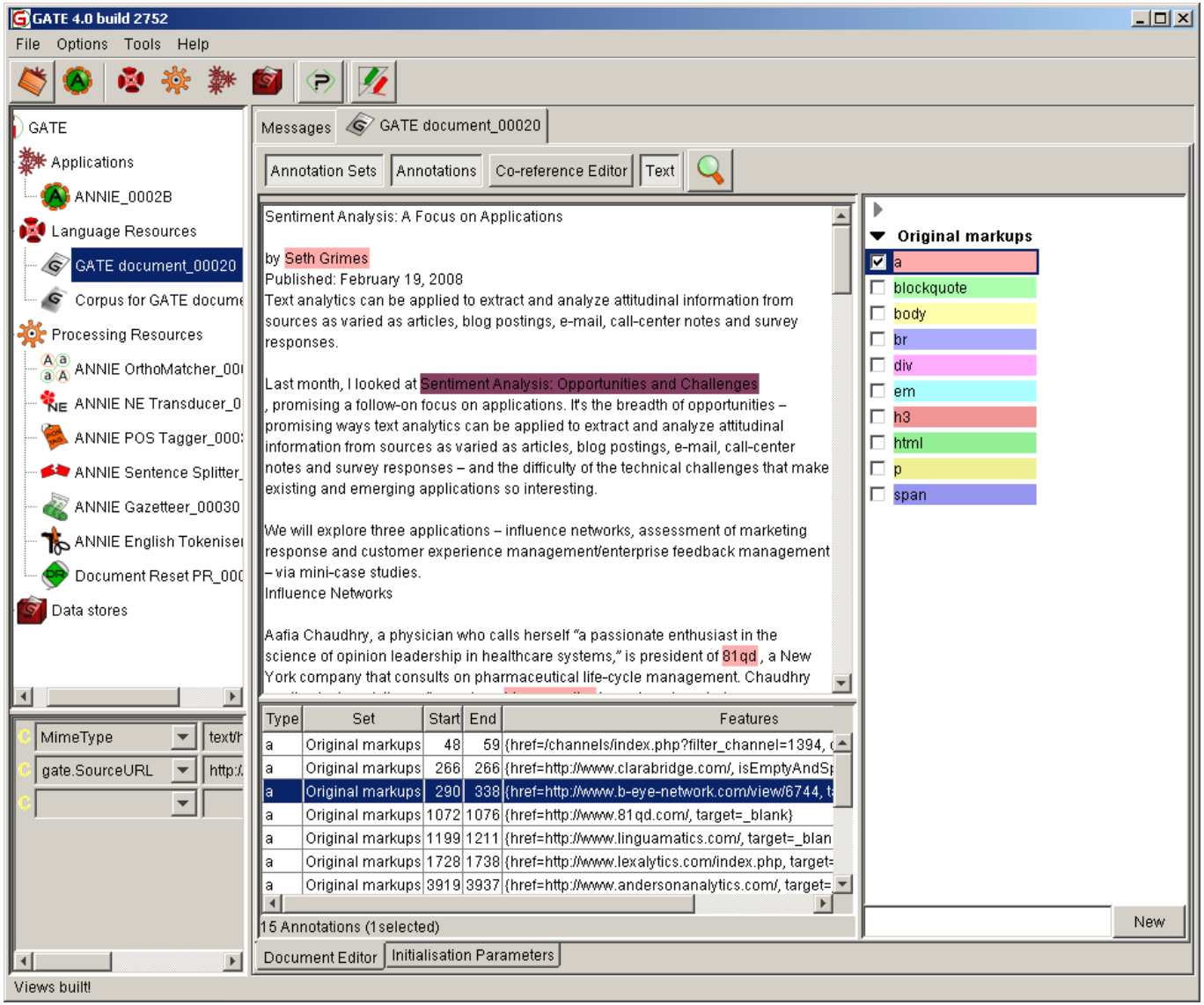


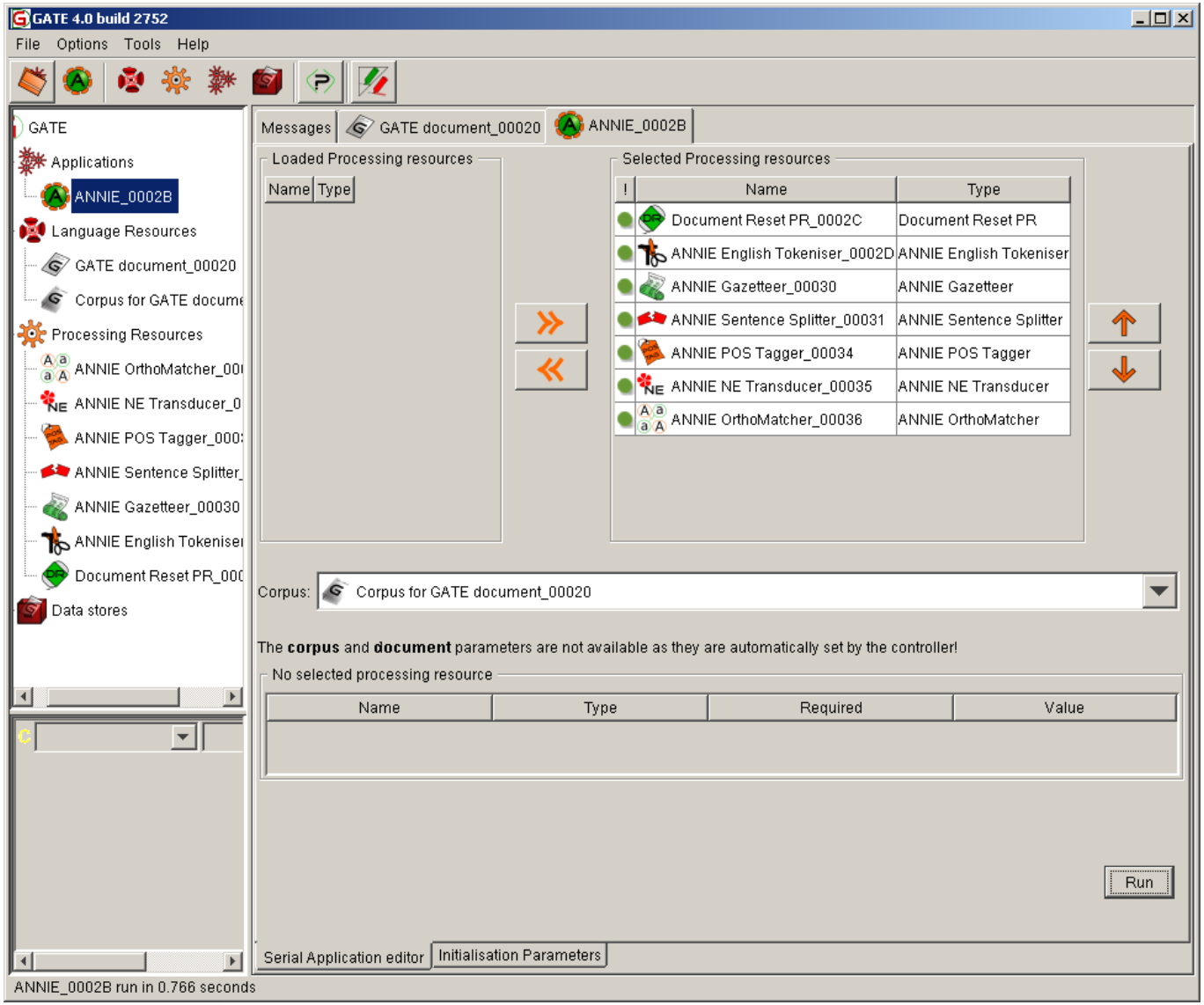


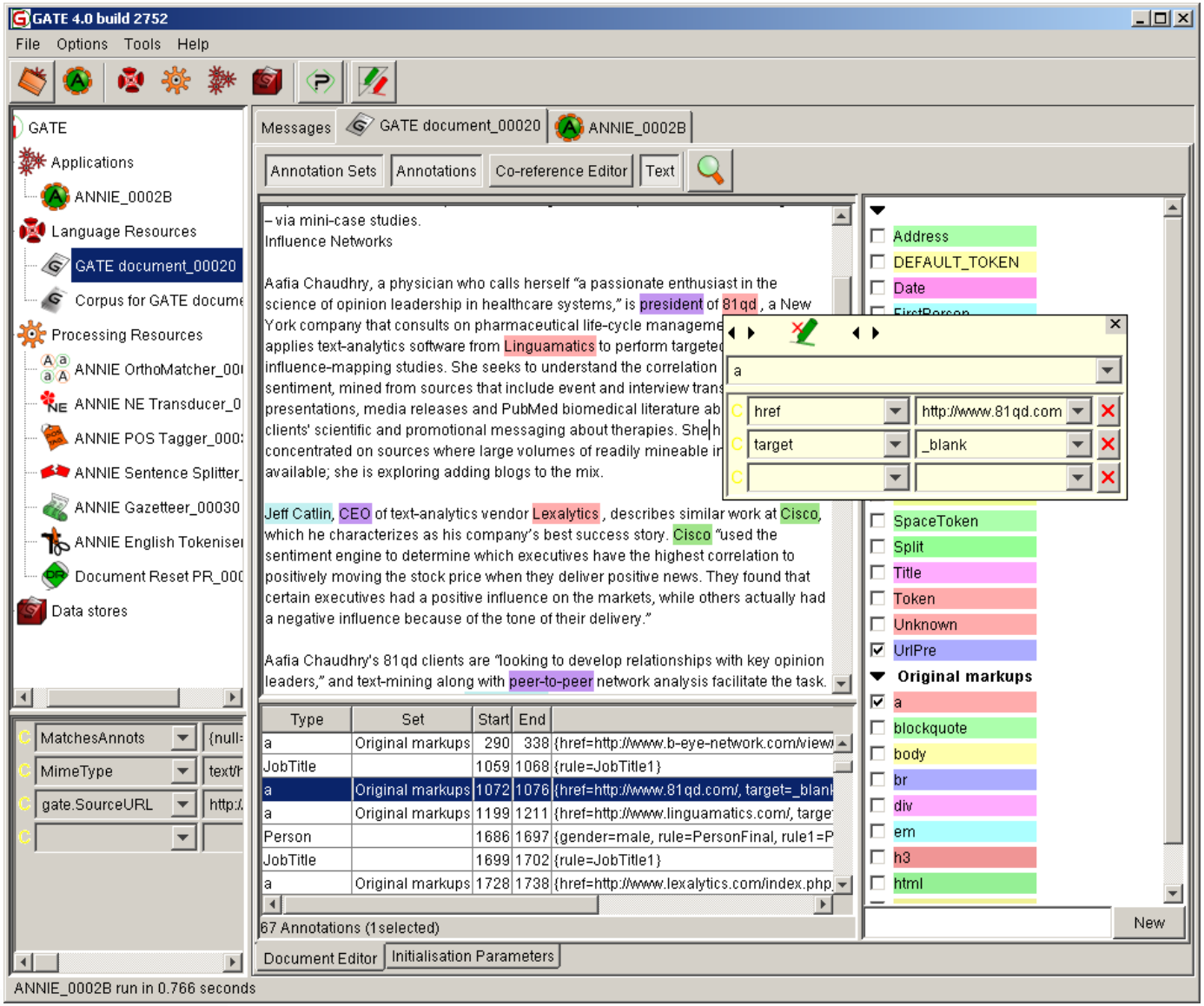
Information Extraction

When we understand, for instance, parts of speech – <subject> <verb> <object> – we're in a position to discern facts and relationships.

Let's see text augmentation (tagging) in action. We'll use GATE, an open-source tool...







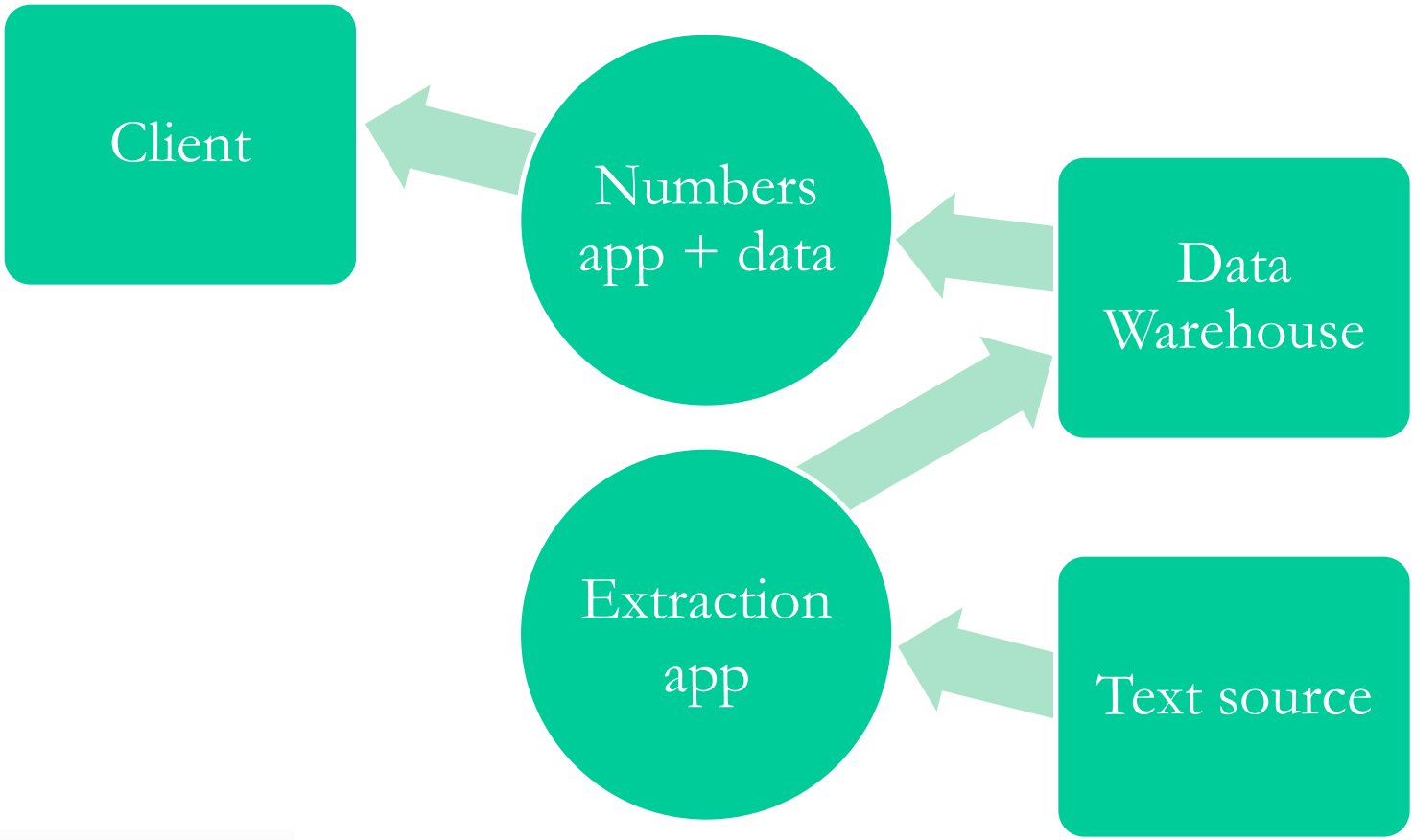
Information Extraction

For content analysis, 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.

Information Extraction

Data integration via information extraction.



Information Extraction

XML-annotated text is an intermediate format.

```
<?xml version='1.0' encoding='windows-1252'?>
<GateDocument>
<!-- The document's features-->

<GateDocumentFeatures>
<Feature>
      <Name className="java.lang.String">MimeType</Name>
      <Value className="java.lang.String">text/html</Value>
</Feature>
<Feature>
      <Name className="java.lang.String">gate.SourceURL</Name>
      <Value className="java.lang.String">http://altaplana.com/SentimentAnalysis.html</Value>
</Feature>
</GateDocumentFeatures>
<!-- The document content area with serialized nodes -->

<TextWithNodes><Node id="0" />Sentiment<Node id="9" /> <Node id="10" />Analysis<Node id="18" />:<Node
id="19" /> <Node id="20" />A<Node id="21" /> <Node id="22" />Focus<Node id="27" /> <Node id="28"
/>on<Node id="30" /> <Node id="31" />Applications<Node id="43" />
<Node id="44" />
<Node id="45" />by<Node id="47" /> <Node id="48" />Seth<Node id="52" /> <Node id="53" />Grimes<Node
id="59" />
<Node id="60" />Published<Node id="69" />:<Node id="70" /> <Node id="71" />February<Node id="79" />
<Node id="80" />19<Node id="82" />,<Node id="83" /> <Node id="84" />2008<Node id="88" />
<Node id="89" />Text<Node id="93" /> <Node id="94" />analytics<Node id="103" />
                                                                    <material cut>
</TextWithNodes>
```

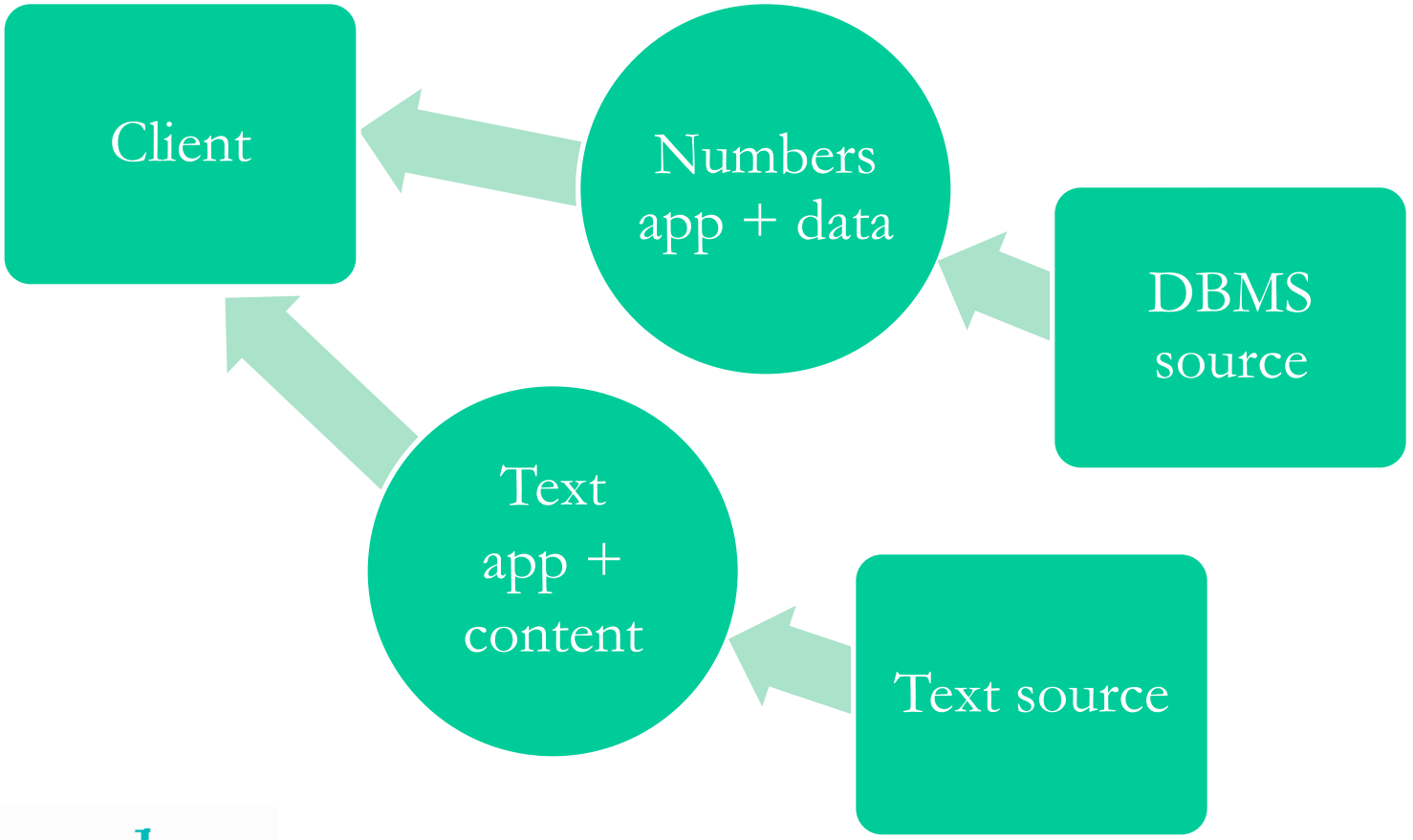
Information Extraction

XML-annotated text...

```
<!-- The default annotation set -->
<AnnotationSet>
                                                                    <material cut>
<Annotation Id="67" Type="Token" StartNode="48" EndNode="52">
      <Feature>
            <Name className="java.lang.String">length</Name>
            <Value className="java.lang.String">4</Value>
      </Feature>
      <Feature>
            <Name className="java.lang.String">category</Name>
            <Value className="java.lang.String">NNP</Value>
      </Feature>
      <Feature>
            <Name className="java.lang.String">orth</Name>
            <Value className="java.lang.String">upperInitial</Value>
      </Feature>
      <Feature>
            <Name className="java.lang.String">kind</Name>
            <Value className="java.lang.String">word</Value>
      </Feature>
      <Feature>
            <Name className="java.lang.String">string</Name>
            <Value className="java.lang.String">Seth</Value>
      </Feature>
</Annotation>
                                                                    <material cut>
</AnnotationSet>
</GateDocument>
```

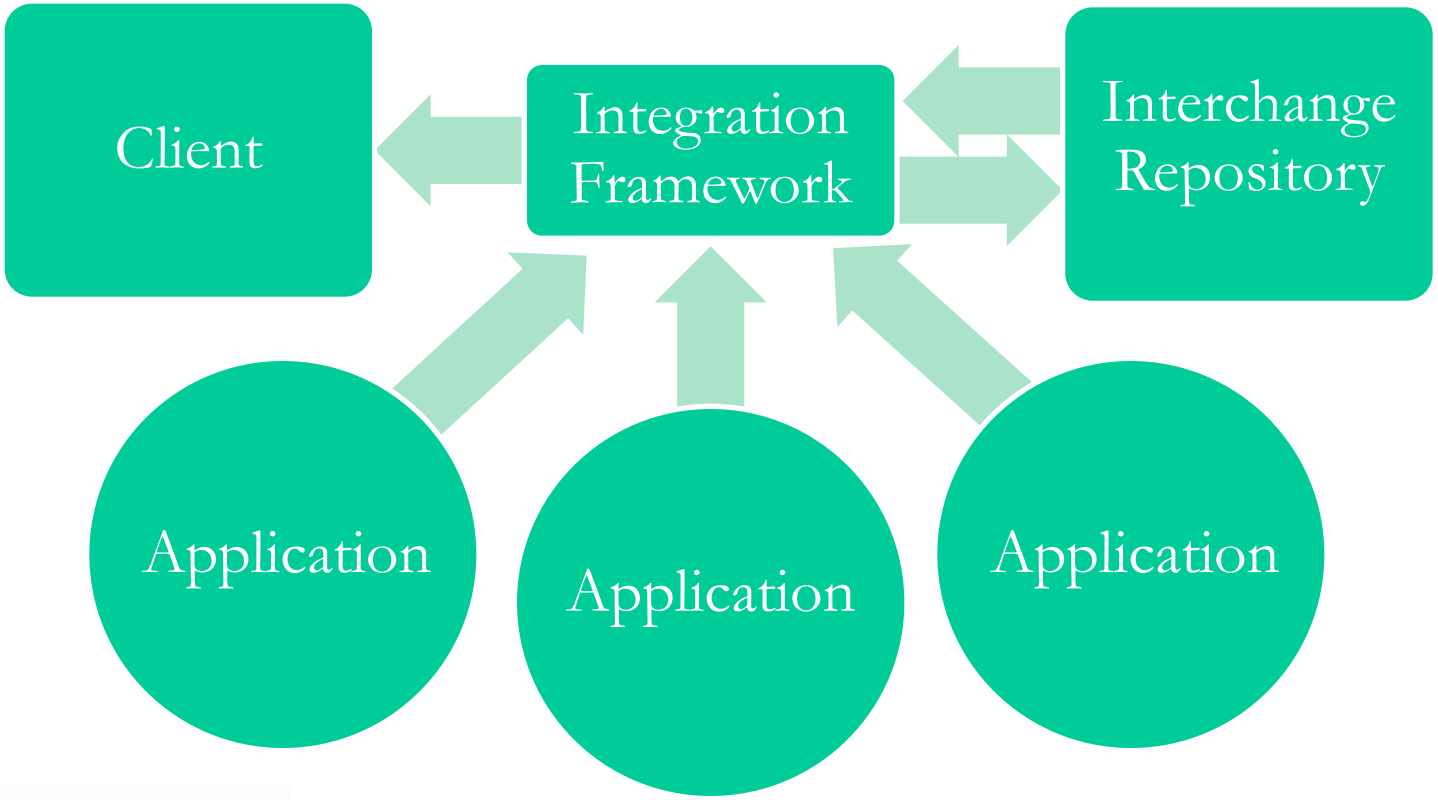
Other Integration Models

Application integration works in some instances.



Integration models

Another model of application/component integration.



Example: E-mail

For Text-BI/DW, we're most interested in IE.

What else can we extract? 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

Example: E-mail

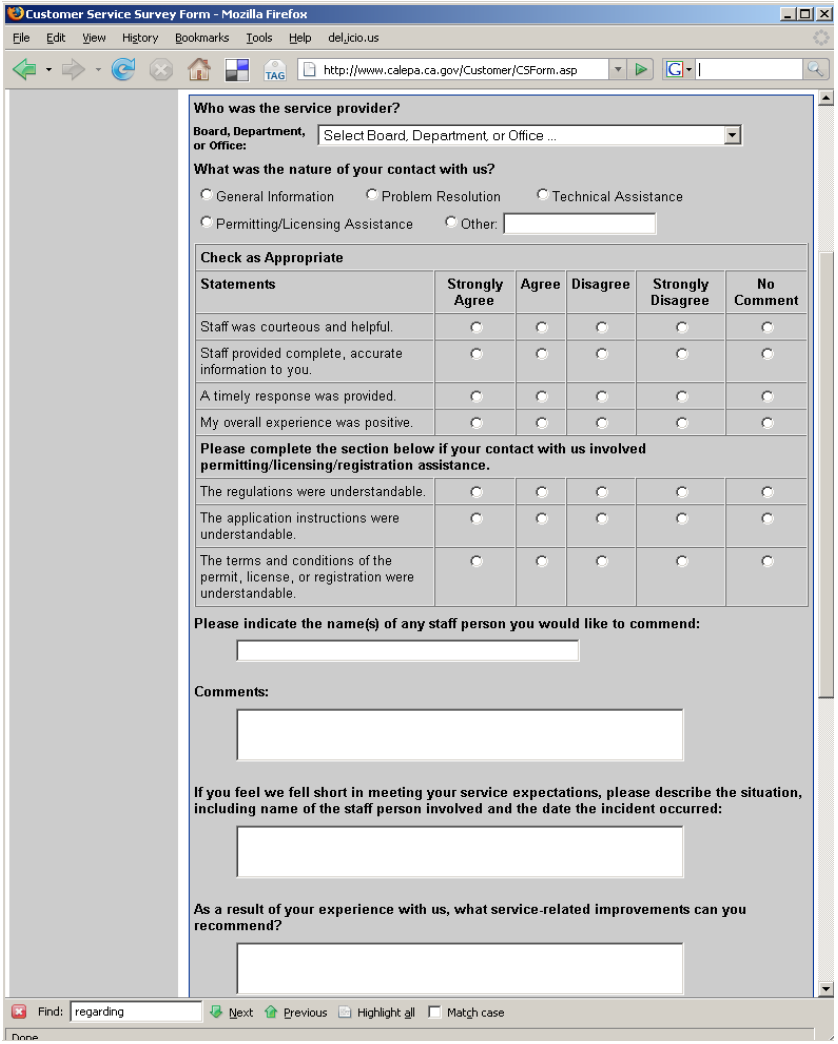
An e-mail message is "semi-structured."

- 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?

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

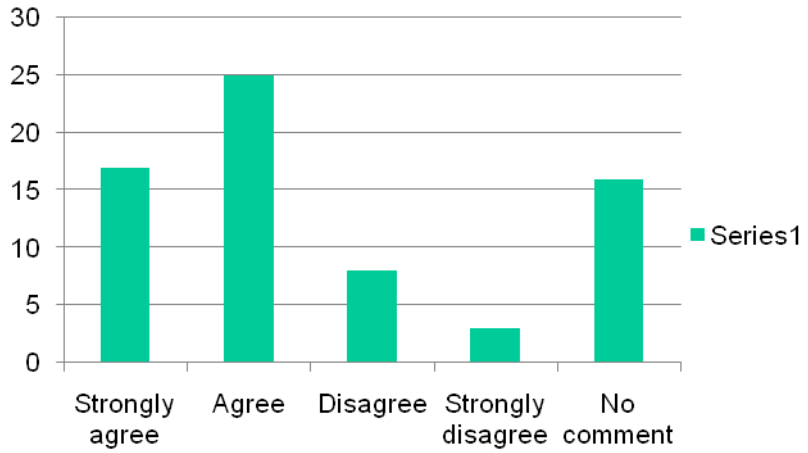
There are many forms of s-s information...

Example: Survey



Example: Survey

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

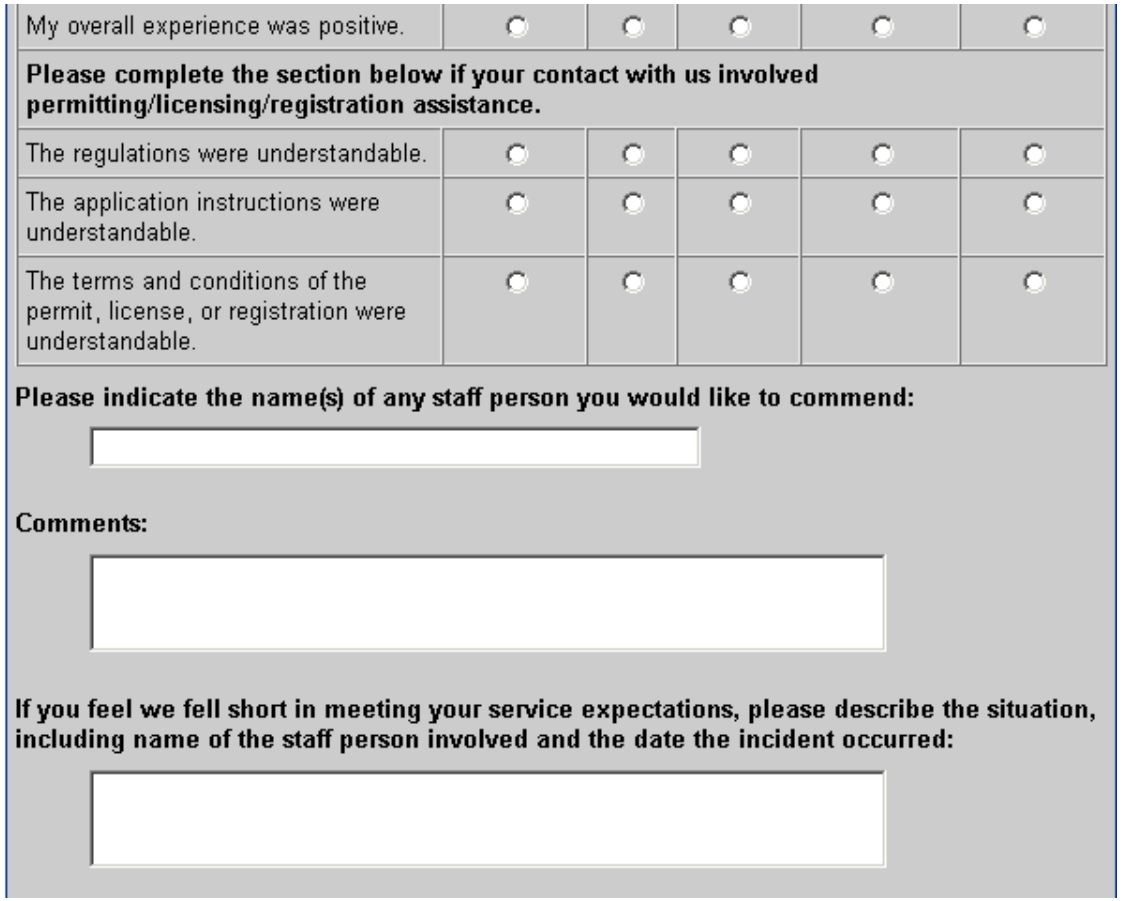


There may be fields that indicate what product/service/person the coded rating applies to.

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.
- Use the *metadata* that describes the information and its provenance.
- Sentiment extraction comes into play for Voice of the Customer / Customer Experience Management applications.

Sentiment Extraction

Sentiment (opinion) extraction –

- Applications include:
  - Reputation management.
  - Competitive intelligence.
  - Quality improvement.
  - Trend spotting.
- Sources include:
  - Wikis, blogs, forums, and newsgroups.
  - Media stories and product reviews.
  - Contact-center notes and transcripts.
  - Customer feedback via Web-site forms and e-mail.
  - Survey verbatims.

Sentiment Extraction

We need to –

- Identify and access candidate sources.
- Extract sentiment to databases.
- Correlate expressed sentiment to measures such as:
  - Sales by product, location, time, etc.
  - Defects by part, circumstances, etc.
- And information such as –
  - Customer information and customer's transactions.
- Correlation depends on semantic agreement: are we talking about the same things?

Example: Attitudinal Data

Exercise: Identify the attitudinal information in this excerpt from Dell's IdeaStorm.com –

"Dell really... REALLY need to stop overcharging... and when i say overcharing... i mean atleast double what you would pay to pick up the ram yourself."

What Sentiment is expressed?

Subject?

Polarity?

Intensity?

Opinion?

Applications

Text Analytics is applied in many domains –

- Life sciences.
- Intelligence and law enforcement.
- E-discovery (legal) and compliance.
- Publishing and information services.
- Insurance and financial services.
- Voice of the Customer (sales, marketing, and product) applications.

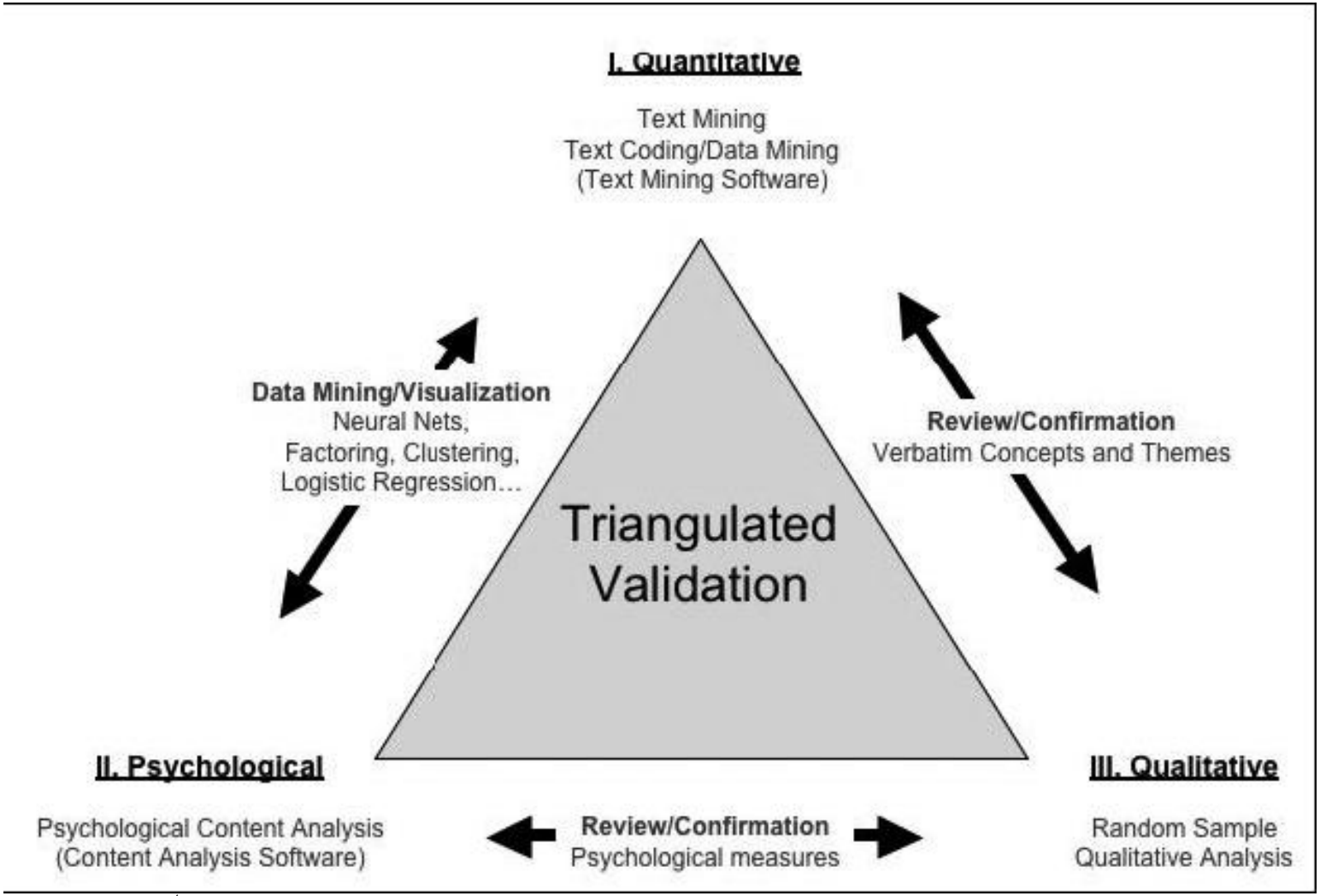
Voice of the Customer

You see the value of the voice of the customer, the need to understand –

- the totality of stakeholder needs and opinions, whether explicitly stated or indirectly implied.
- individual views and collective, market thinking.
- customers wherever and however they express themselves.

Our thesis is that customer voices are most frequently expressed in text.

Voice of the Customer



www.andersonanalytics.com/index.php?mact=News,cntnt01,getfile,1&cntnt01filename=SCIP0208TomAndersonArticle.pdf&cntnt01returnid=46&page=46

Voice of the Customer

Additional concepts and tools apply...

"**Net Promoter** is a discipline by which companies profitably grow by focusing on their customers."

"One simple question - **Would you recommend us to a friend or colleague?** - allows companies to track promoters and detractors and produces a clear measure of an organization's performance through its customers' eyes."

-- *http://www.netpromoter.com/netpromoter/index.php*

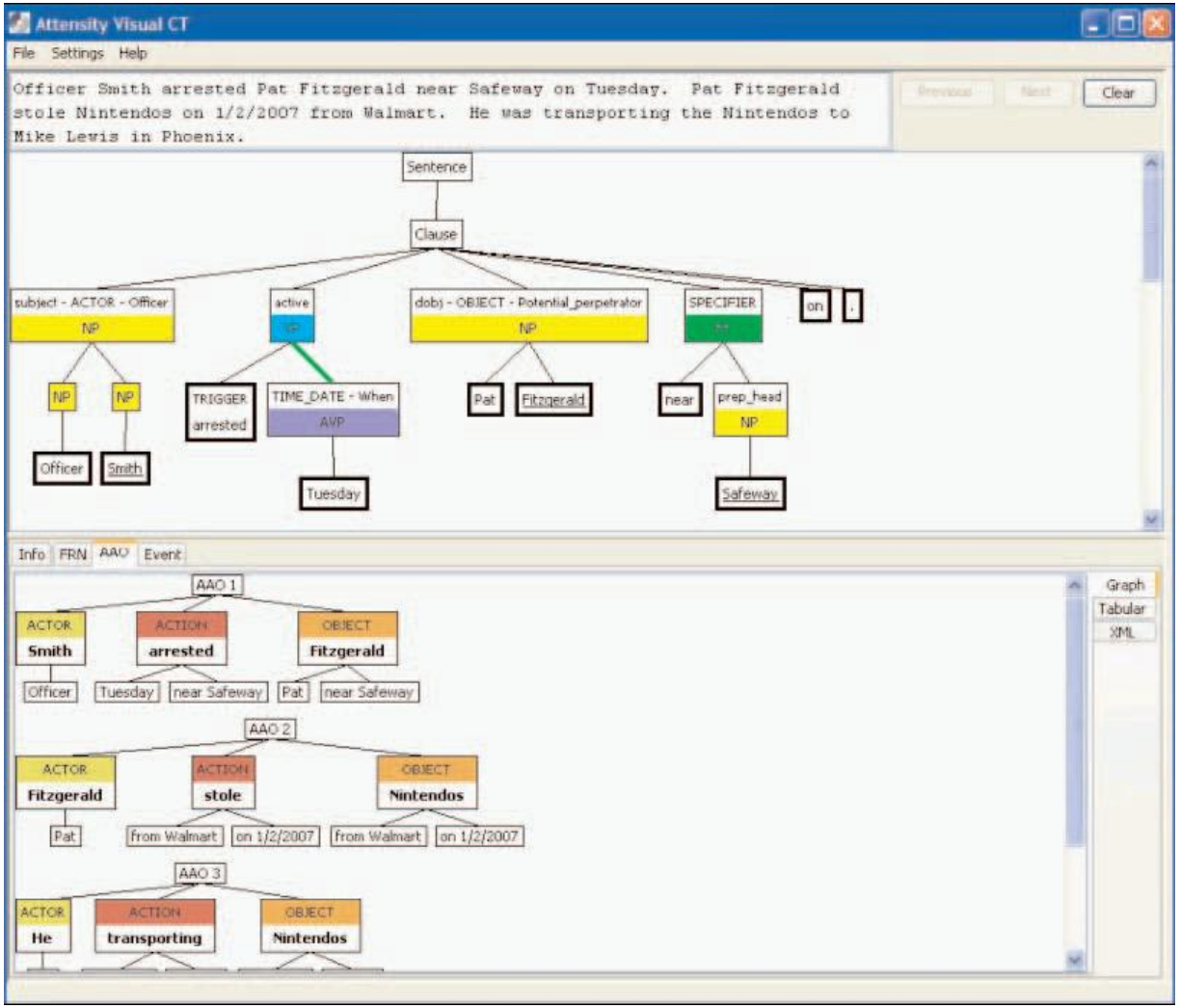
Applications

Take law enforcement as an example–

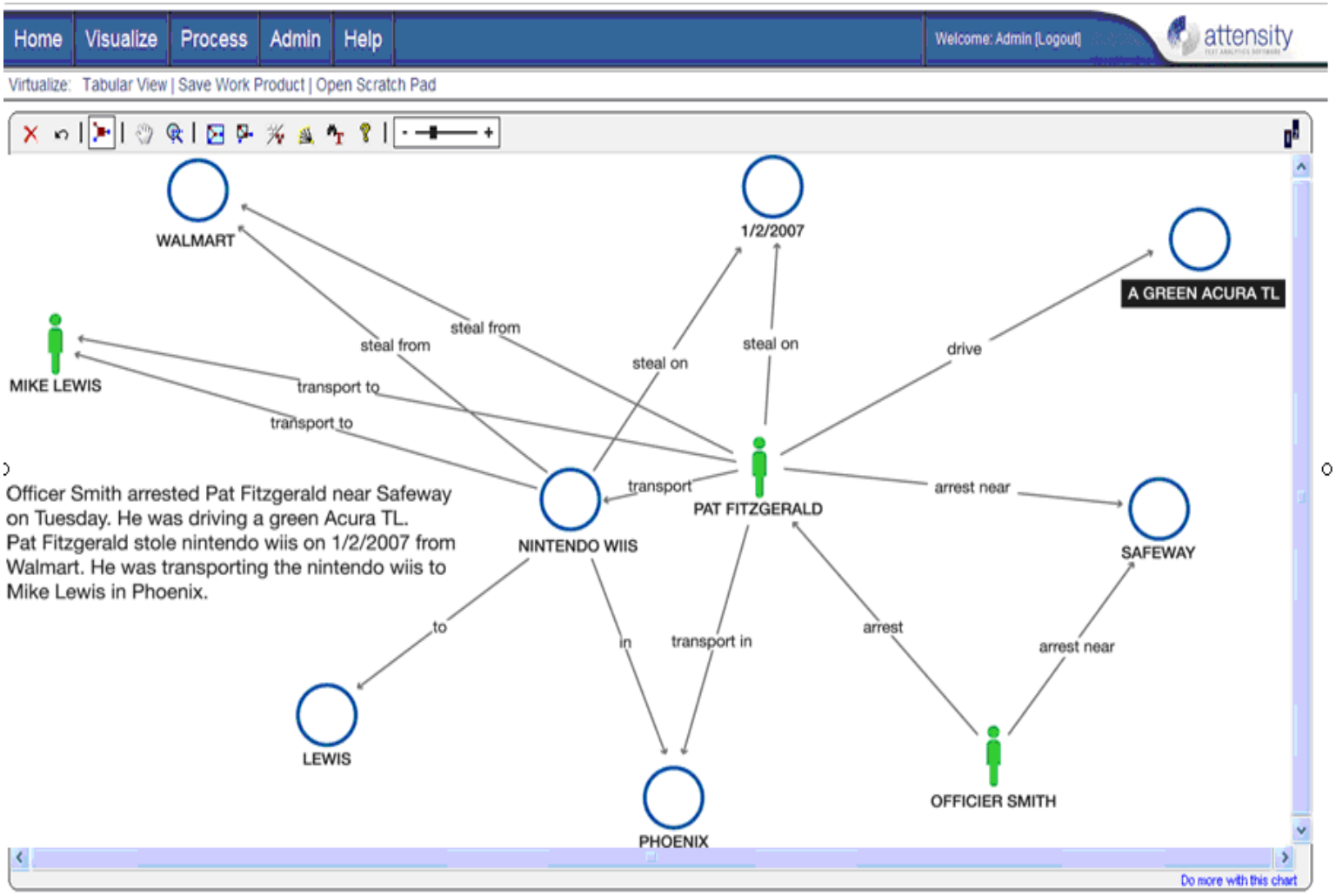
- Sources: case files, crime reports, incident and victimization databases, legal documents
- Targets: crime patterns, criminal investigation, networks

Applications

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



Applications



Applications

Customer Relationship Management (CRM)

- Sources: customer e-mail, letters, call 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

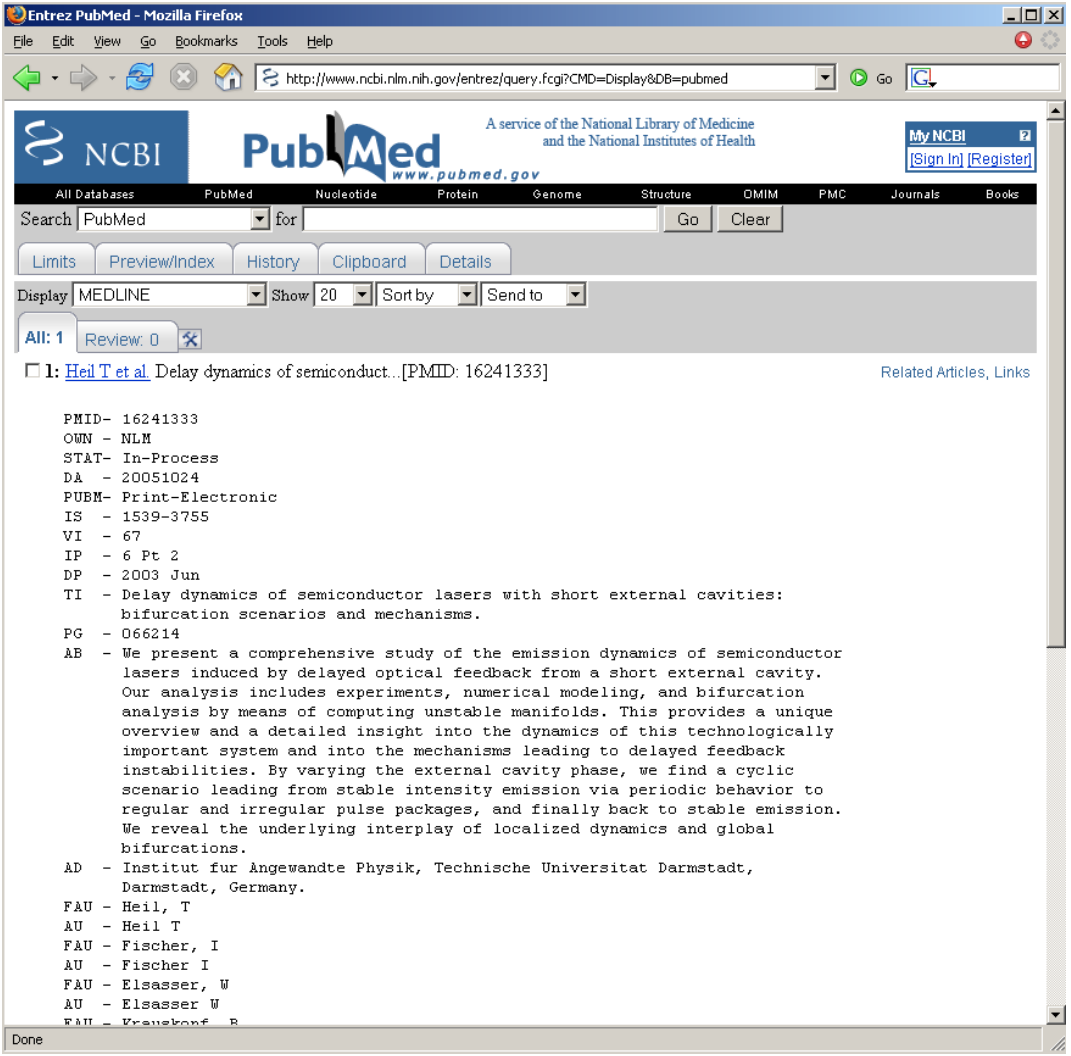
- Sources: clinical research databases, patient records, insurance filings, regulations
- Targets: enhance diagnosis and treatment, promote quality of service, increase utilization, control costs

Intelligence and counter-terrorism

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

Case study: IBM's MedTAKMI

MEDLINE from the National Center for Biotechnology Information hosts links to many widely used information sources such as the PubMed database of 15 million biomedical journal abstracts. Visit *www.ncbi.nlm.nih.gov*.



Alta Plana

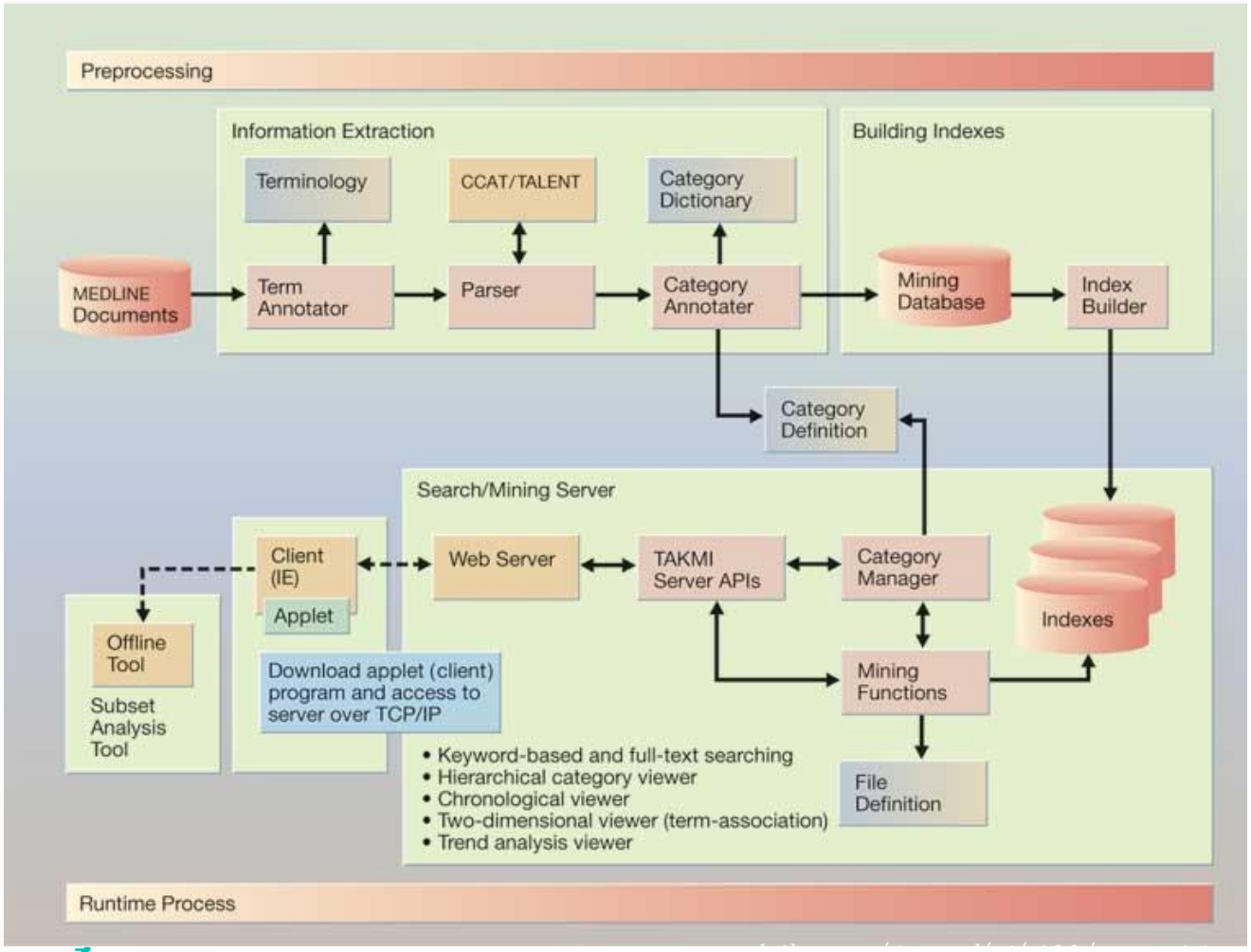
Case study: IBM's MedTAKMI

MedTAKMI = Text Analysis and Knowledge MIning for Biomedical Documents (*ibm.com*):

- An extension of the TAKMI (Text Analysis and Knowledge MIning) system originally developed for text mining in CRM applications.
- Goal is to extract relationships among biomedical entities (e.g. proteins and genes), from patterns such as "A inhibits B" and "A activates B," where A and B represent specific entities.
- Work starts with a "syntactic parser" that identifies entities and basic binary (a noun and a verb) and ternary (two nouns and a verb) relationships.

Case study: IBM's MedTAKMI

Figure 1 MedTAKMI architecture



Case study: IBM's MedTAKMI

Figure 7 Two-dimensional view of proteins vs proteins (mouse-related) for a document set that contains the term "p53"

2D Map

Save Copy Print

cross category: Mus musculus
vertical category: Mus musculus
Cross Axle: By Frequecy / By Alphabet

Compound
Amino Acid
Organ
+Dry Lab Methods
+FUNCTION
+cellular component
+gene from Gene Ontology
+Root of LocusLink phenotype
+MeSH Minor (Tree)
+MeSH Major (Tree)
Minor MeSH
Major MeSH
+Protein By Species
Age Tags
Country
Data Completed
Publication Data
URL Full-Text
Last Revision Date
Organization Name
Prime Name of Substance
Personal Name as Subject
Publication Type
Subset
Secondary Source Identifier
Source
Summary For Patient In
Publication Status
Journal Title Abbreviation
Check Tags
Transliterated/Vernacular Title
Update In
Update Of
URL Summary
+PROPERTY
Minor Qualifier
Major Qualifier
+SACCHARIDE
+SMART Domain
Affiliation
Author
Data Created

	protein	tumor suppressor	breast cancer	Bc12-associated X protein	myelocytomatosis oncogene	transcription factor
protein	632 (100.0%)	57 (9.01%)	42 (6.64%)	60 (9.49%)	24 (3.79%)	27 (4.27%)
tumor suppressor	57 (38.51%)	148 (100.0%)	4 (2.7%)	8 (5.4%)	6 (4.05%)	10 (6.75%)
breast cancer	42 (35.9%)	4 (3.41%)	117 (100.0%)	2 (1.7%)	8(6.83%)	3 (2.56%)
Bcl2-associated X-prot	60 (59.41%)	8 (7.92%)	2 (1.98%)	101 (100.0%)	4 (3.96%)	3 (2.97%)
myelocytomatosis onco	24 (38.1%)	6 (9.52%)	8 (12.7%)	4 (6.34%)	63 (100.0%)	2 (3.17%)
transcription factor	27 (45.0%)	10 (16.66%)	3 (5.0%)	3 (5.0%)	2 (3.33%)	60 (100.0%)
G elongation factor	28 (53.85%)	3 (5.76%)	1 (1.92%)	19 (36.54%)	1 (1.92%)	1 (1.92%)
proliferative cell nuclea	19 (37.25%)	1 (1.96%)	3 (5.88%)	2 (3.92%)	3 (5.88%)	1 (1.96%)
period	18 (37.5%)	2 (4.16%)	4 (8.33%)	2 (4.16%)	3 (6.25%)	0 (0.0%)
enhancer of rudimentar	17 (42.5%)	2 (5.0%)	23 (57.5%)	0 (0.0%)	1 (2.5%)	3 (7.5%)
epiregulin	17 (42.5%)	2 (5.0%)	23 (57.5%)	0 (0.0%)	1 (2.5%)	3 (7.5%)
progesterone receptor	13 (32.5%)	0 (0.0%)	20 (50.0%)	0 (0.0%)	1 (2.5%)	0 (0.0%)
Harvey rat sarcoma viru	12 (32.42%)	5 (13.51%)	0 (0.0%)	1 (2.7%)	9 (24.32%)	1 (2.7%)
tumor necrosis factor	11 (33.33%)	3 (9.09%)	0 (0.0%)	4 (12.12%)	1 (3.03%)	2 (6.06%)
epidermal growth factor	11 (33.33%)	2 (6.06%)	6 (18.18%)	0 (0.0%)	3 (9.09%)	0 (0.0%)
vascular endothelial gro	9 (29.03%)	0 (0.0%)	1 (3.22%)	1 (3.22%)	1 (3.22%)	1 (3.22%)
vitamin D receptor	10 (33.33%)	0 (0.0%)	2 (6.66%)	1 (3.33%)	1 (3.33%)	1 (3.33%)
estrogen receptor	8 (28.57%)	0 (0.0%)	16 (57.14%)	0 (0.0%)	1 (3.57%)	0 (0.0%)
protein-L-isoaspartate	7 (25.92%)	0 (0.0%)	2 (7.4%)	0 (0.0%)	3 (11.11%)	1 (3.7%)
MAS1 oncogene	9 (33.33%)	0 (0.0%)	11 (40.74%)	0 (0.0%)	1 (3.7%)	0 (0.0%)
protein kinase	7 (25.92%)	1 (3.7%)	0 (0.0%)	4 (14.81%)	1 (3.7%)	1 (3.7%)
cornichon	8 (30.76%)	1 (3.84%)	3 (11.54%)	0 (0.0%)	0 (0.0%)	1 (3.84%)

Case study: IBM's MedTAKMI

Entity extraction here is recognition of gene, protein, and chemical names from biomedical text based on a domain dictionary with two million entities.

Categories are constructed from public ontologies.

Figure 11 2D map analysis correlating LocusLink phenotypes (vertical axis) and signaling proteins (horizontal axis) in 1051 papers

	S_TKc	STYKc	TyrKc	HATPase_c	HisKA	HR1	SAM	ITAM
Leukemia	9 (5.23%)	9 (5.23%)	9 (5.23%)	3 (1.74%)	2 (1.16%)	2 (1.16%)	2 (1.16%)	1 (0.58%)
HMG-CoA lyase deficiency	7 (10.29%)	7 (10.29%)	7 (10.29%)	3 (4.41%)	2 (2.94%)	3 (4.41%)	2 (2.94%)	0 (0.0%)
Hepatic lipase deficiency	7 (10.29%)	7 (10.29%)	7 (10.29%)	3 (4.41%)	2 (2.94%)	3 (4.41%)	2 (2.94%)	0 (0.0%)
Miller-Dieker lissencephaly syndrome	2 (5.4%)	2 (5.4%)	2 (5.4%)	1 (2.7%)	1 (2.7%)	0 (0.0%)	1 (2.7%)	0 (0.0%)
Colorectal cancer	0 (0.0%)	0 (0.0%)	0 (0.0%)	2 (6.06%)	1 (3.03%)	0 (0.0%)	0 (0.0%)	1 (3.03%)
Lupus erythematosus	1 (3.57%)	1 (3.57%)	1 (3.57%)	3 (10.71%)	3 (10.71%)	0 (0.0%)	0 (0.0%)	0 (0.0%)
Osteosarcoma	0 (0.0%)	0 (0.0%)	0 (0.0%)	1 (3.7%)	0 (0.0%)	0 (0.0%)	0 (0.0%)	1 (3.7%)
Histiocytoma	0 (0.0%)	0 (0.0%)	0 (0.0%)	1 (3.7%)	0 (0.0%)	0 (0.0%)	0 (0.0%)	1 (3.7%)
Li-Fraumeni syndrome	0 (0.0%)	0 (0.0%)	0 (0.0%)	1 (3.7%)	0 (0.0%)	0 (0.0%)	0 (0.0%)	1 (3.7%)

www.research.ibm.com/journal/sj/433/uramoto.html

Getting Started: Options

Deployment Options –

- Traditional software installation.
- SaaS / Hosted.
- Outsourced to a service bureau.
- Web Services APIs.
- VOC text analytics embedded in line-of-business applications.

Consider evaluation and implementation Best Practices. What did my study discover?

Getting Started: Analytical Styles

Unified Analytics –

- Approaches build on familiar BI and predictive-analytics tools and approaches...
- Adding data and text mining...
- Extracting entities, facts, sentiment, etc....
- Relying on semantic integration...
- ...for true, 360º enterprise views.

Getting Started: Vendors

Let's categorized vendors by analytic style –

- Text-BI.
- Text-integrated predictive analytics/data mining.
- Focused on Voice of the Customer/Customer Experience Management.
- Toolkits.

Research Study

The study centered on lessons solicited from:

- individuals with experience applying Voice of the Customer (VoC) text analytics to real-world business problems at their organizations.
- a number of vendor representatives and industry analysts.

It looked at Best Practices defined as –

- generalized principles, techniques, and methodologies derived from theory, academic and industrial research, direct experience, and customer stories.

Research Study

For the study,

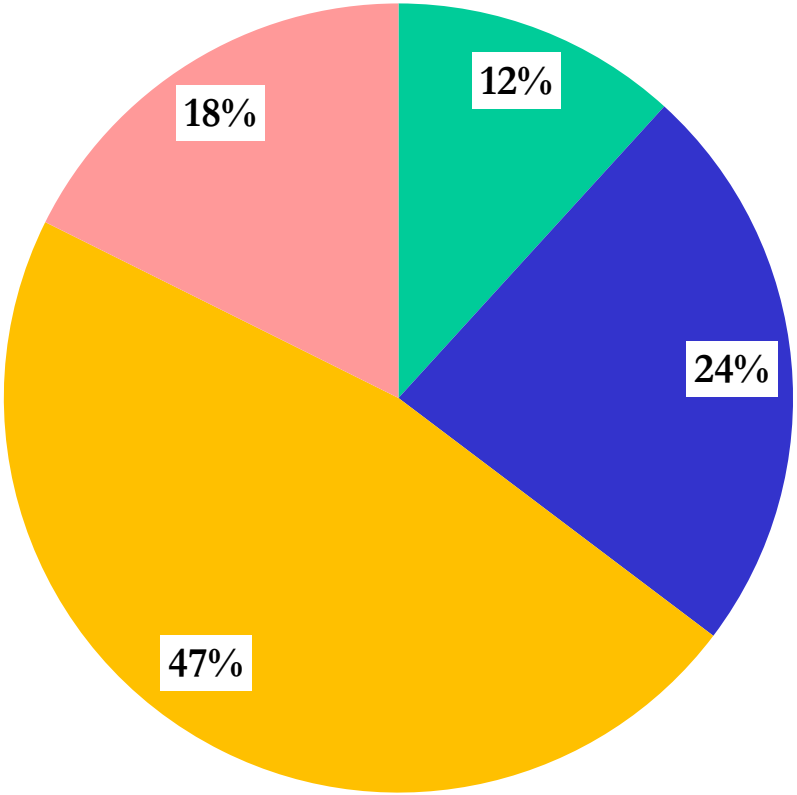
- I had free-form discussions with five industry analysts, six end users, and a variety of vendor executives.
- I additionally conducted a formal, small-sample survey of VoC text-analytics practitioners – end users and consultants – that attracted 26 valid responses.

Length of Respondent Experience

The plurality of respondents have been using text analytics for Voice of the Customer work for two years or more!

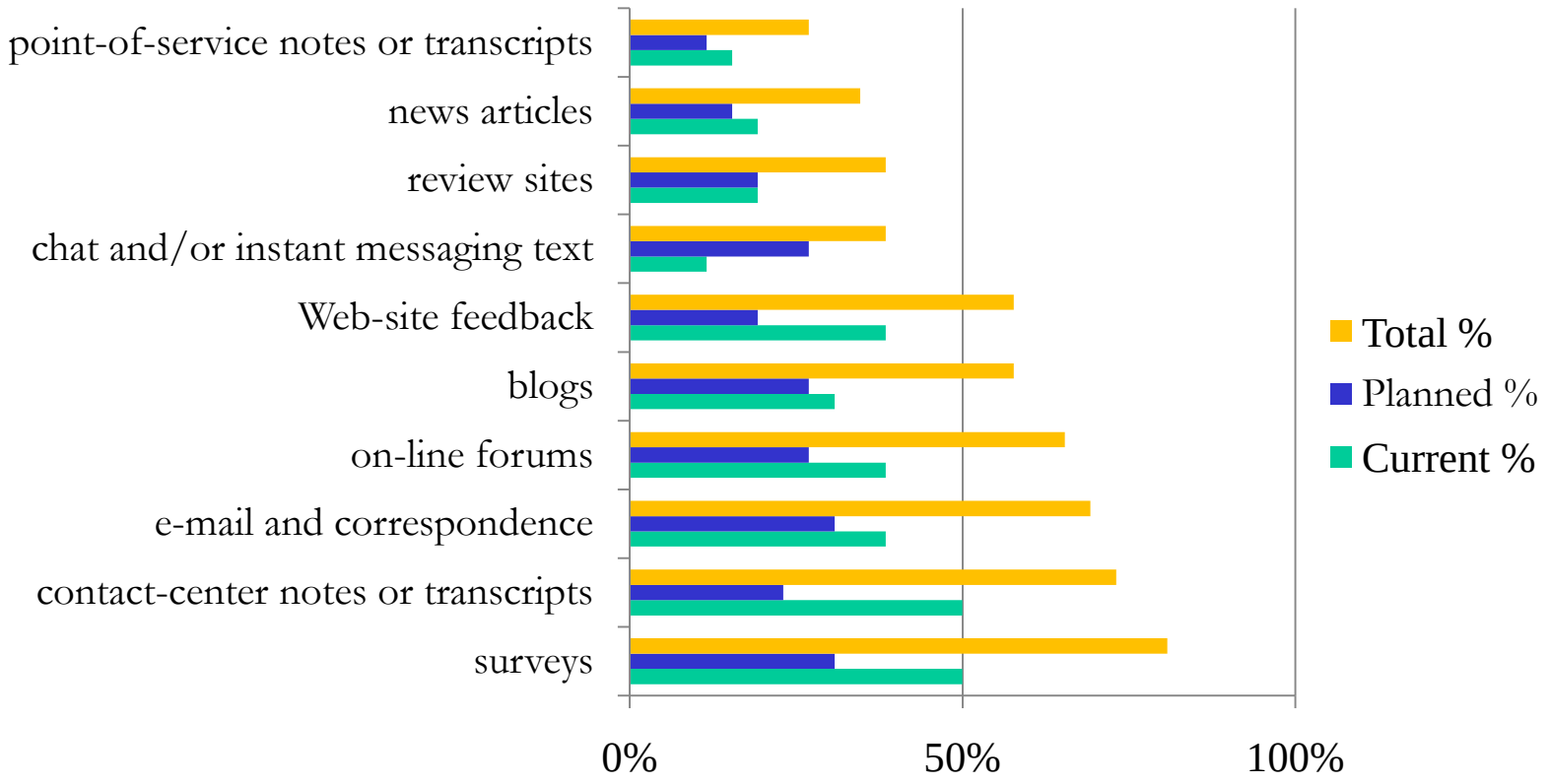
Length of Experience	Response Percent
still evaluating/not using	31%
less than 6 months	15%
6 months to less than one year	8%
one year to less than two years	12%
two years or more	**35%**

Satisfaction with VOC Text Analytics

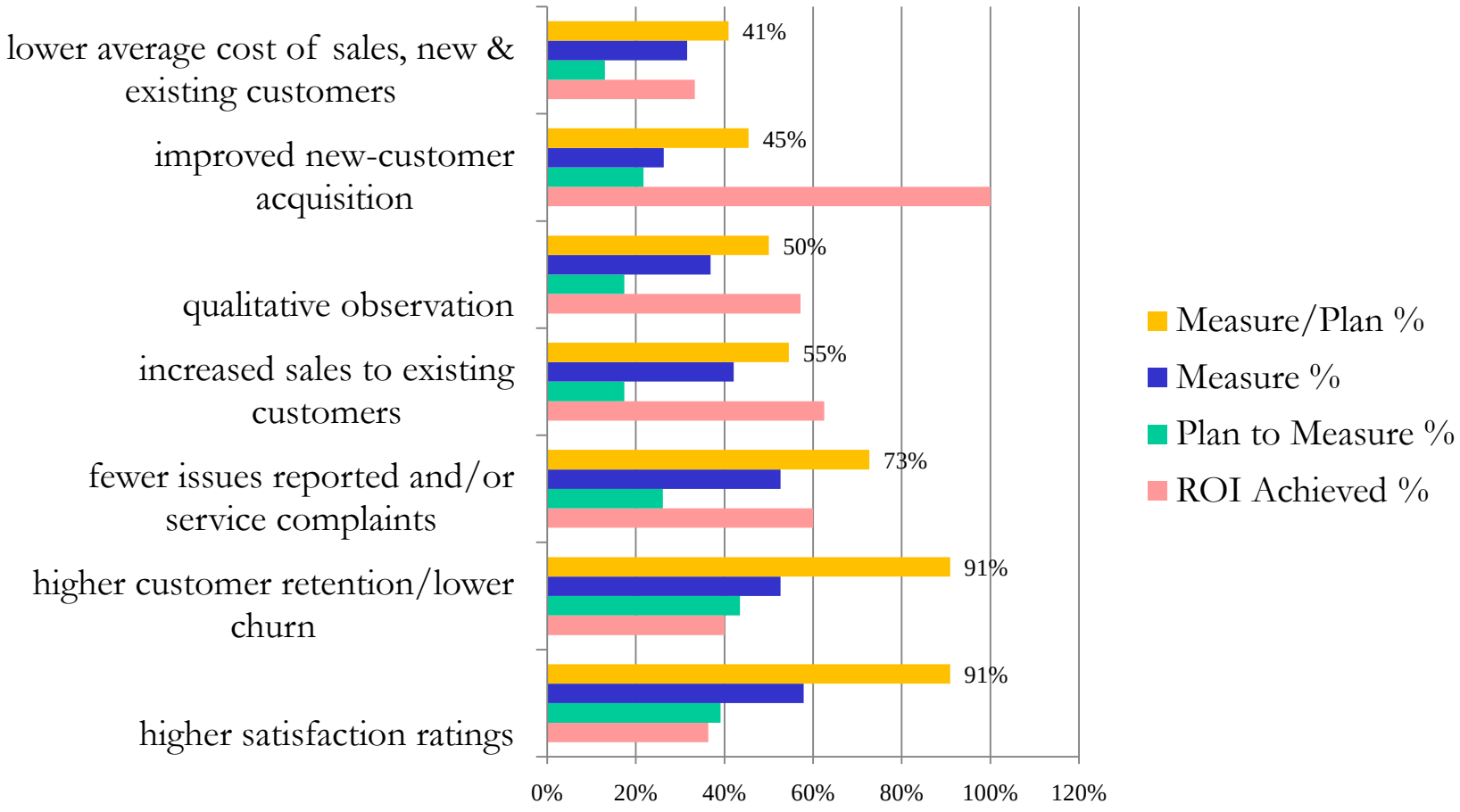


There were no responses below Neutral.

Information Analyzed



ROI Measured, Planned & Achieved



Solution Providers

What should a prospect look for?

Response	Percent
deep sentiment/opinion extraction	80%
ability to use specialized dictionaries or taxonomies	76%
broad information extraction capability	60%
adaptation for particular sectors, e.g., hospitality, retail, health care, communications	56%
predictive-analytics integration	48%
BI (business intelligence) integration	48%
support for multiple languages	48%
ability to create custom workflows	32%
low cost	32%
hosted or "as a service" option	32%
specialized VoC analysis interface	24%

Advice to End Users

Advice responses lend themselves to clustering in four categories:

1. Business Goals
2. Evaluation/Pilot
3. Implementation/Start-up
4. Operational Principles

Subsequent slides present the more-interesting responses classified in these categories…

Business Goals

Start with the always-useful advice –

- Clearly define initiatives based on strategic objectives.

A number of items address making the case –

- Unless there is a taste for innovation in the organization, it is hard to fund a project.
- You should understand your business. Are they ready?
- Is your business ready to take action on the data?
- Understand that you're going to run into opposition.
- Try to impress on the people you're talking to that there's urgency for change.
- You need an in-house team championing the effort.

Evaluation/Pilot

Pilot with open source tools to learn.

Study the TA industry. Develop detailed documentation of your needs

Do the homework. Go through benchmarking 3-4 tools in the domain. Look at extra services: initiation, end-user training.

Do a proof of concept evaluation.

Understand the problem you're solving and the audience you're solving it for.

You want to look for a tool that you can use across different constituencies.

Adherence to UIMA – the Unstructured Information Management Architecture – is important.

Implementation/Start-up

The largest cluster of best-practices/advice responses. Some see a progressive approach –

- Think big, start small, use an incremental approach.
- Start in an area where value can be created quickly and expand from there.
- Win fast and win often.
- Those quick wins mean a lot. It took us a while to understand the technology.
- Don't go after the biggest problem first.
- Collect requirements from all potential internal customers and ensure they are included when building out reporting.
- Provide simple access to data and analysis.

Implementation/Start-up

One advisor suggests looking for **services** while two emphasize proper **training** and one stresses "Have the **right tool** in place."

Two focus on staff:

- Dedicate someone to the job.
- I'll emphasize getting a core team together that knows the technology and can support it in the organization.

Two focus on start-up time:

- Don't underestimate the time/effort needed to build libraries.
- It takes time to develop the screening rules to make text analytics really useful.

Implementation/Start-up

One user advises –

> Keep the project off executive radar until you know if it is meaningful in your business.

While another explains –

> I didn't tell anybody what I did. I just did it.

Operational Principles

There were three (inconsistent) accuracy points –

- Users should demand sensitivity of analysis.
- Don't expect the high levels of accuracy we experience in the quant side.
- Accuracy is less critical to VOC than most uses of Text Analytics (80%+ good enough).

Operational Principles

Other points –

- Text analytics is not push button. It takes work.
- Be open to possibilities. VOC text analytics can be transformative in a whole variety of ways.
- Don't try to mix initiatives, look for synergies and relations but keep scope within reach
- Integrate both structured and unstructured data.
- [Forget ROI.] We need to get a Return on Time [spent on manual processes and on implementing automated text analytics].

Market View

Let's look at key attributes of the text analytics market –

- I estimate the overall software market at $250 million/year, growing at over 25% annually. IDC and Hurwitz analysts agree.
- There's been significant consolidation, which will continue.
- There's been tremendous emergence of new solution focused companies.
- Community has been picking up steam. Open source shouldn't be far behind.

Market View: Vendor Consolidation

Market View: Community

Alta Plana

To Learn More…

Visit my Business Intelligence Network expert channel, *http://www.b-eye-network.com/sethgrimes*

You can find my VOC Text Analytics report there.

Get in touch:

Seth Grimes

Alta Plana Corporation

grimes@altaplana.com

+1 301-270-0795

Thanks!