Text Analytics for BI/DW Practitioners

Seth Grimes

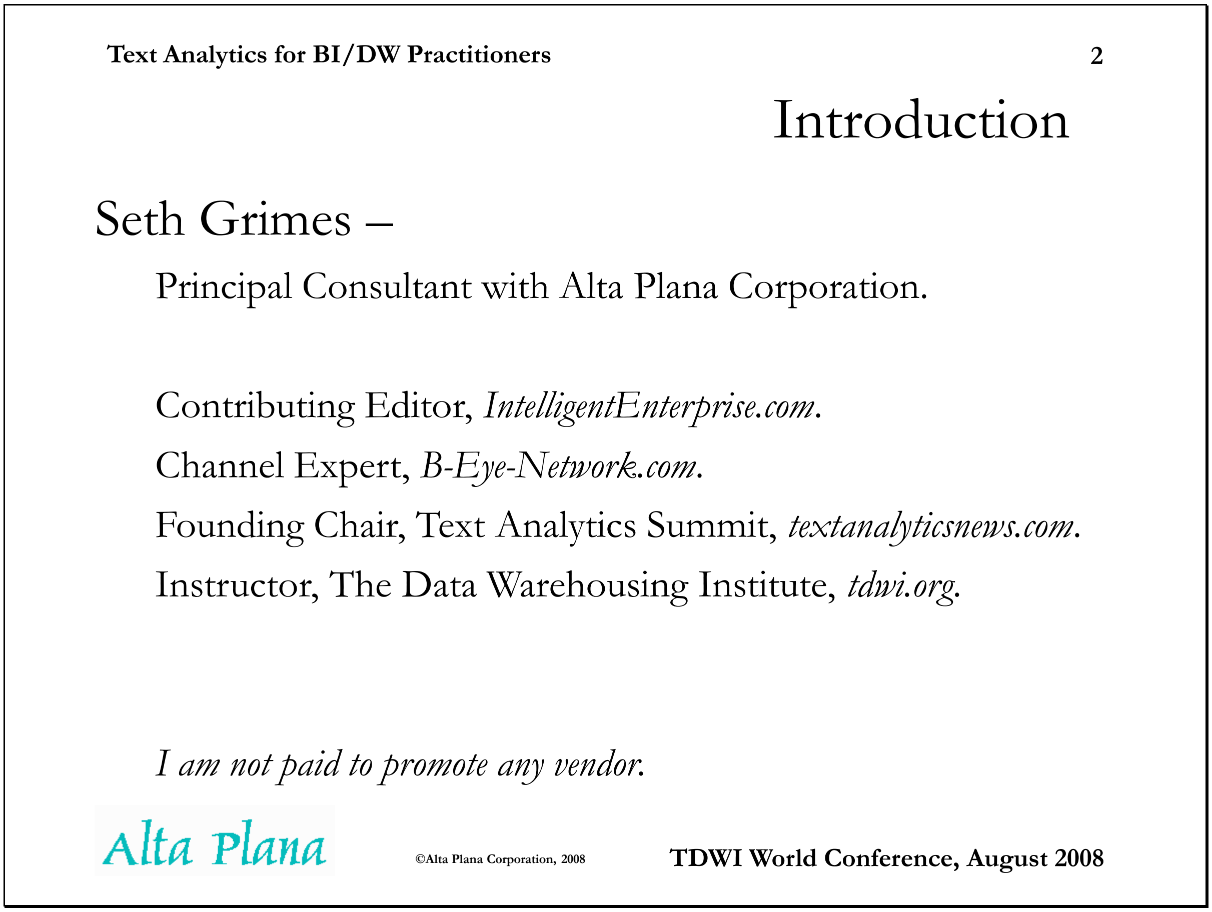
Alta Plana Corporation

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?

- Publish, Manage, and Archive.
- Index and Search.
- Categorize and Classify according to *metadata* & contents.
- 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

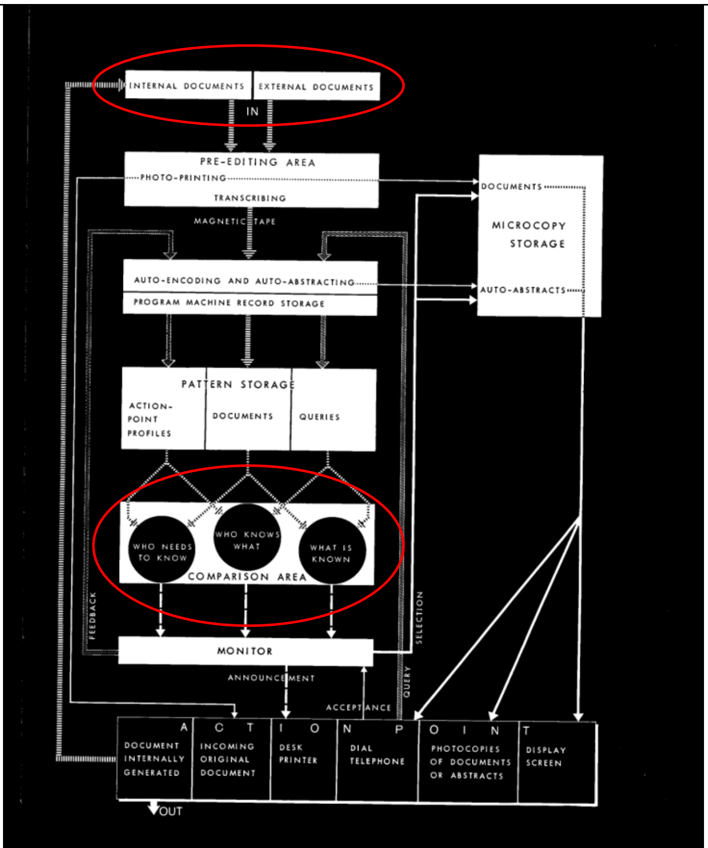
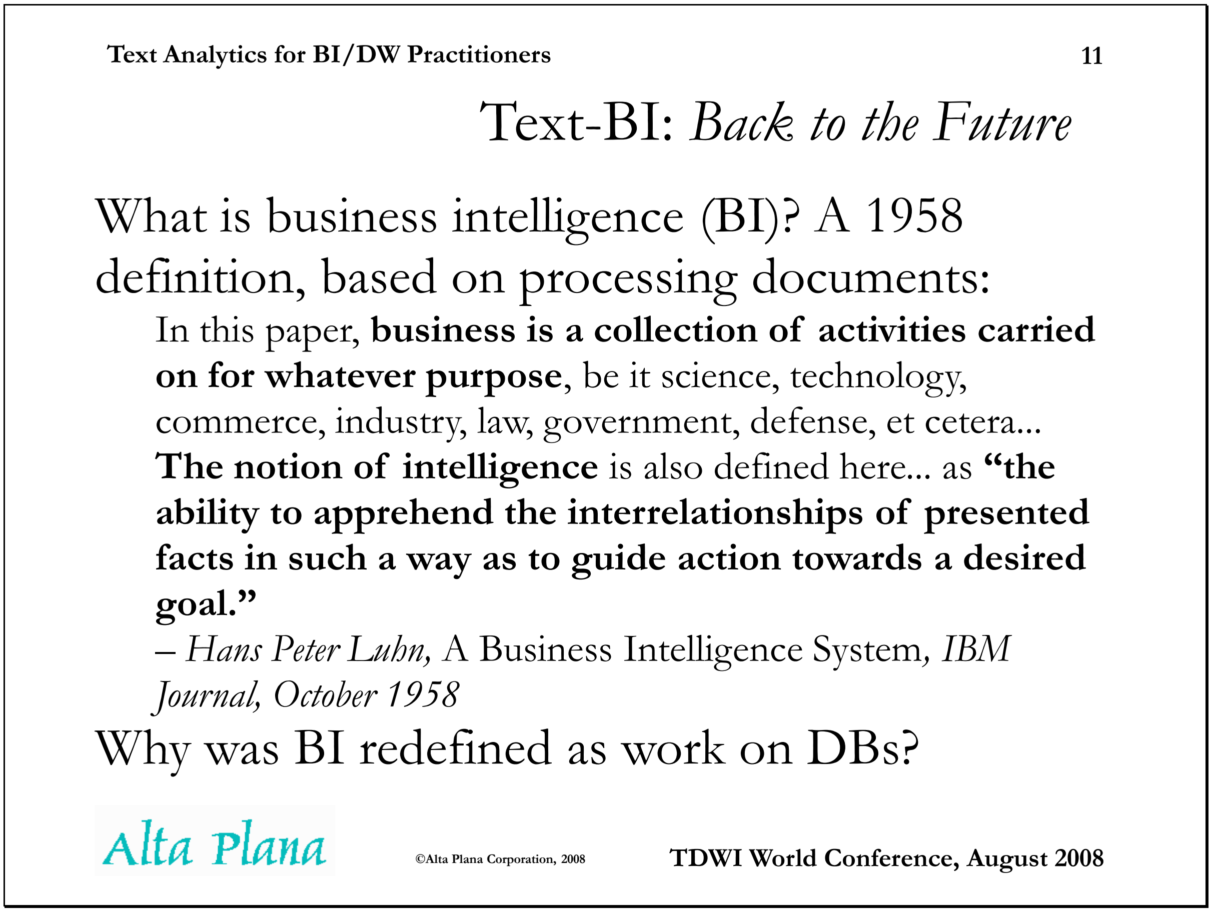


Figure 1 **A Business Intelligence System**

This diagram from H.P. Luhn's 1958 business intelligence paper shows the role of document input and processing, but the centerpiece is, essentially, knowledge management: questions What is Known?, Who Knows What?, Who Needs to Know?



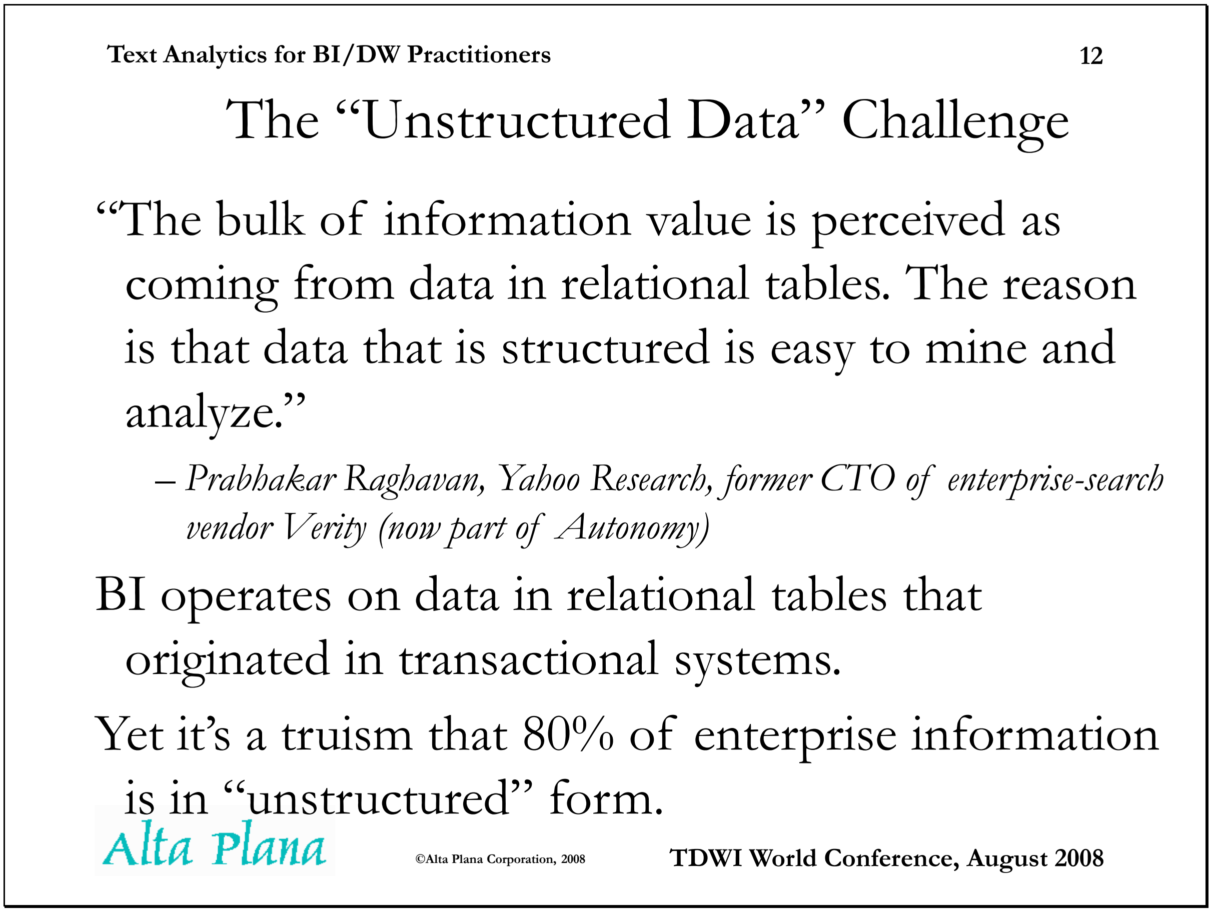
Text-BI: *Back to the Future*

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 center-piece of Luhn's conception of business intelligence is a focus on interrelationships – facts – action.



The "Unstructured Data" Challenge

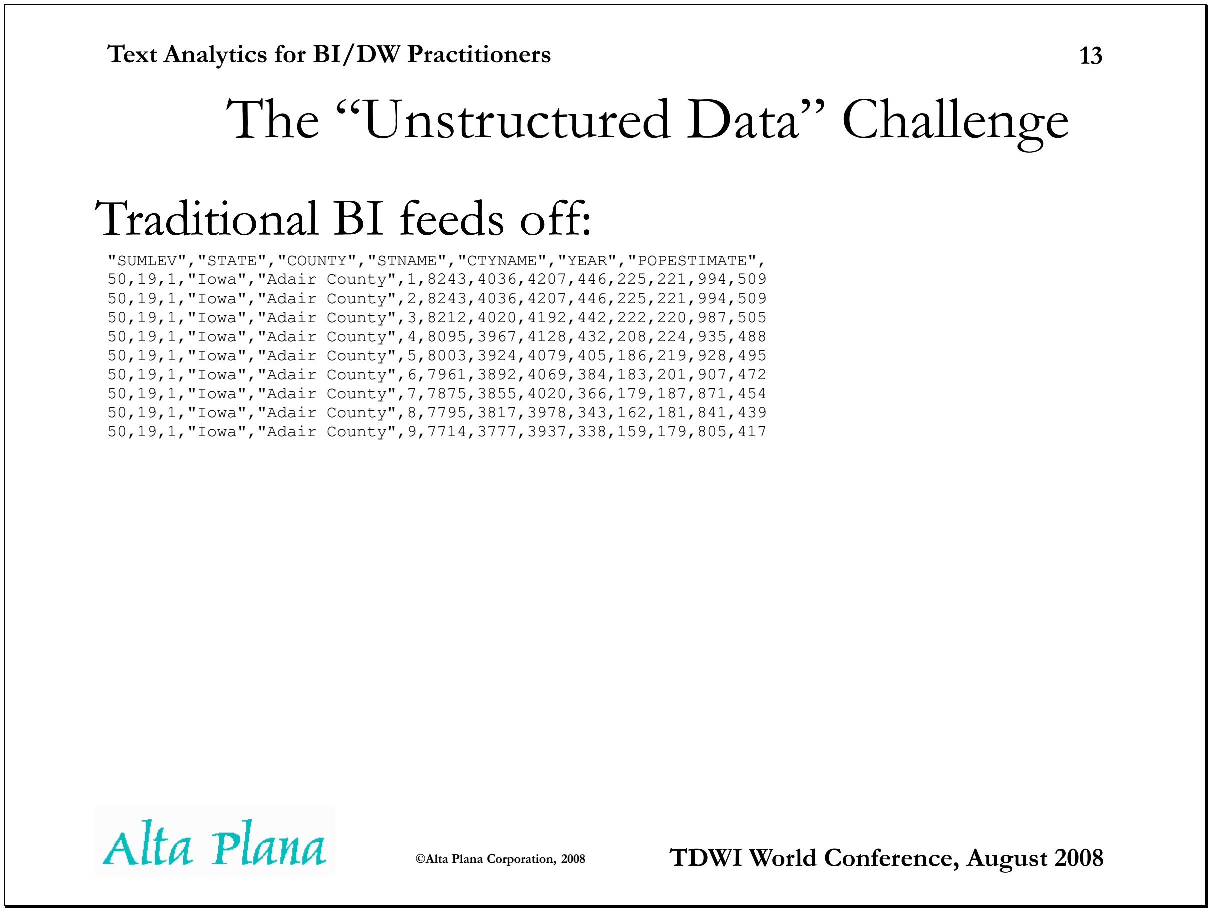
"The bulk of information value is perceived as coming from data in relational tables. The reason is that data that is structured is easy to mine and analyze."

– *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.

That 80% figure – sometimes you hear more, sometimes less. When TDWI's Philip Russom polled the TDWI community, he got something like 65%: but that community is data-warehousing folks. Sometimes you hear higher figures, 85% or even more.



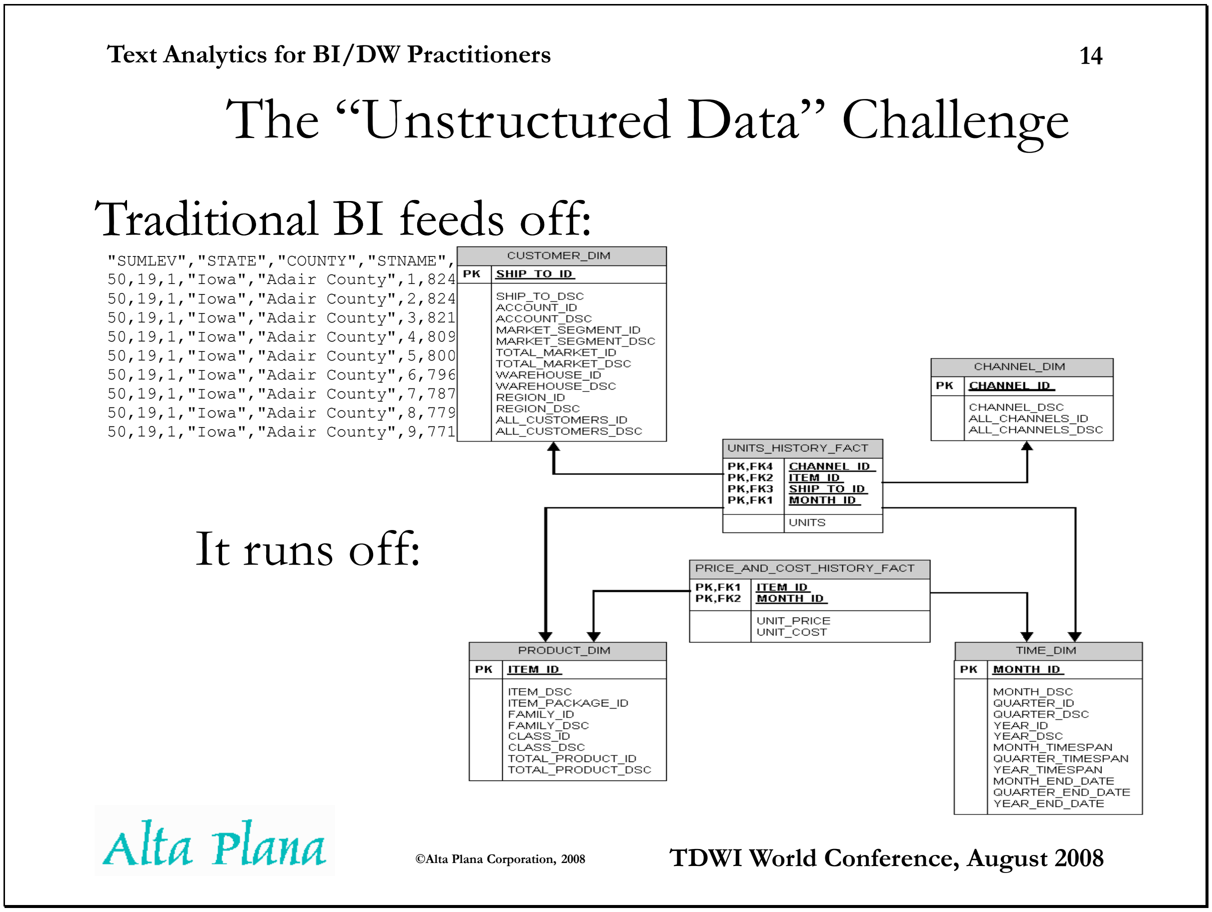
The “Unstructured Data” Challenge

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
```

This source material is a CSV (formatted) file. The first row consists of column labels, i.e., variable names. The first few variables are dimensions. They are followed by measures.

The data we might read from a transactional database would convey pretty much the same information, just organized differently.



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:

CUSTOMER_DIM	
PK	SHIP_TO_ID
	SHIP_TO_DSC ACCOUNT_ID ACCOUNT_DSC MARKET_SEGMENT_ID MARKET_SEGMENT_DSC TOTAL_MARKET_ID TOTAL_MARKET_DSC WAREHOUSE_ID WAREHOUSE_DSC REGION_ID REGION_DSC ALL_CUSTOMERS_ID ALL_CUSTOMERS_DSC

CHANNEL_DIM	
PK	CHANNEL_ID
	CHANNEL_DSC ALL_CHANNELS_ID ALL_CHANNELS_DSC

UNITS_HISTORY_FACT	
PK,FK4 PK,FK2 PK,FK3 PK,FK1	CHANNEL_ID ITEM_ID SHIP_TO_ID MONTH_ID
	UNITS

PRICE_AND_COST_HISTORY_FACT	
PK,FK1 PK,FK2	ITEM_ID MONTH_ID
	UNIT_PRICE UNIT_COST

PRODUCT_DIM	
PK	ITEM_ID
	ITEM_DSC ITEM_PACKAGE_ID FAMILY_ID FAMILY_DSC CLASS_ID CLASS_DSC TOTAL_PRODUCT_ID TOTAL_PRODUCT_DSC

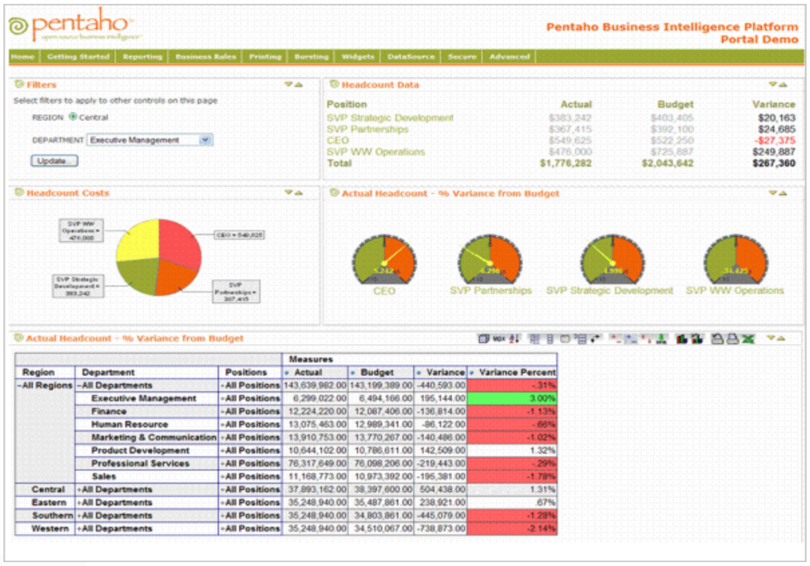
TIME_DIM	
PK	MONTH_ID
	MONTH_DSC QUARTER_ID QUARTER_DSC YEAR_ID YEAR_DSC MONTH_TIMESPAN QUARTER_TIMESPAN YEAR_TIMESPAN MONTH_END_DATE QUARTER_END_DATE YEAR_END_DATE

Alta Plana

BI typically runs off a star schema. In this example, we have two fact tables indexed via a foreign-key relationship with four dimension tables. This dimensional data structure is designed for data-analysis applications.

The "Unstructured Data" Challenge

Traditional BI produces:


pentaho
open source business intelligence
Pentaho Business Intelligence Platform
Portal Demo
Home
Getting Started
Reporting
Business Rules
Printing
Bursting
Widgets
DataSource
Secure
Advanced
Filters
Headcount Data
Headcount Costs
SVP WW Operations = 476,000
CEO = 549,625
SVP Strategic Development = 383,242
SVP Partnerships = 367,415
Actual Headcount - % Variance from Budget
5.242
6.296
4.996
34.425
CEO
SVP Partnerships
SVP Strategic Development
SVP WW Operations
Actual Headcount - % Variance from Budget

Select filters to apply to other controls on this page

REGION ● Central

DEPARTMENT Executive Management

Update...

Position	Actual	Budget	Variance
SVP Strategic Development	$383,242	$403,405	$20,163
SVP Partnerships	$367,415	$392,100	$24,685
CEO	$549,625	$522,250	-$27,375
SVP WW Operations	$476,000	$725,887	$249,887
Total	$1,776,282	$2,043,642	$267,360

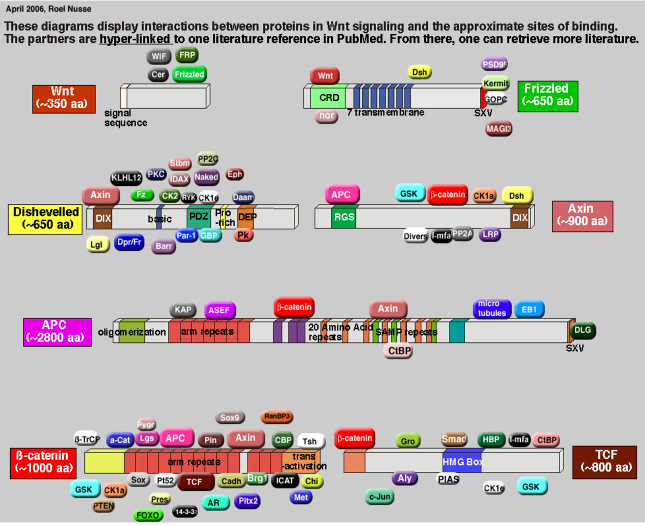
			Measures			
Region	Department	Positions	Actual	Budget	Variance	Variance Percent
-All Regions	-All Departments	+All Positions	143,639,982.00	143,199,389.00	-440,593.00	-.31%
	Executive Management	+All Positions	6,299,022.00	6,494,166.00	195,144.00	3.00%
	Finance	+All Positions	12,224,220.00	12,087,406.00	-136,814.00	-1.13%
	Human Resource	+All Positions	13,075,463.00	12,989,341.00	-86,122.00	-.66%
	Marketing & Communication	+All Positions	13,910,753.00	13,770,267.00	-140,486.00	-1.02%
	Product Development	+All Positions	10,644,102.00	10,786,611.00	142,509.00	1.32%
	Professional Services	+All Positions	76,317,649.00	76,098,206.00	-219,443.00	-.29%
	Sales	+All Positions	11,168,773.00	10,973,392.00	-195,381.00	-1.78%
Central	+All Departments	+All Positions	37,893,162.00	38,397,600.00	504,438.00	1.31%
Eastern	+All Departments	+All Positions	35,248,940.00	35,487,861.00	238,921.00	.67%
Southern	+All Departments	+All Positions	35,248,940.00	34,803,861.00	-445,079.00	-1.28%
Western	+All Departments	+All Positions	35,248,940.00	34,510,067.00	-738,873.00	-2.14%

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

This dashboard/portal nicely captures common BI display elements: 1) dashboard speedometer widgets that show status of key performance indicators (KPIs), 2) a typical graphic, in this case, a pie chart, 3) a multi-measure, multi-dimensional pivot table that can be explored interactively, 4) a more conventional, report-style tabular data presentation, and 5) a parameterization mechanism, a mean of choosing a value from a pick-list that is then used to decide what is display in other interface elements.

The “Unstructured Data” Challenge

Some information doesn’t come from a data file.



April 2006, Roel Nusse

These diagrams display interactions between proteins in Wnt signaling and the approximate sites of binding. The partners are hyper-linked to one literature reference in PubMed. From there, one can retrieve more literature.

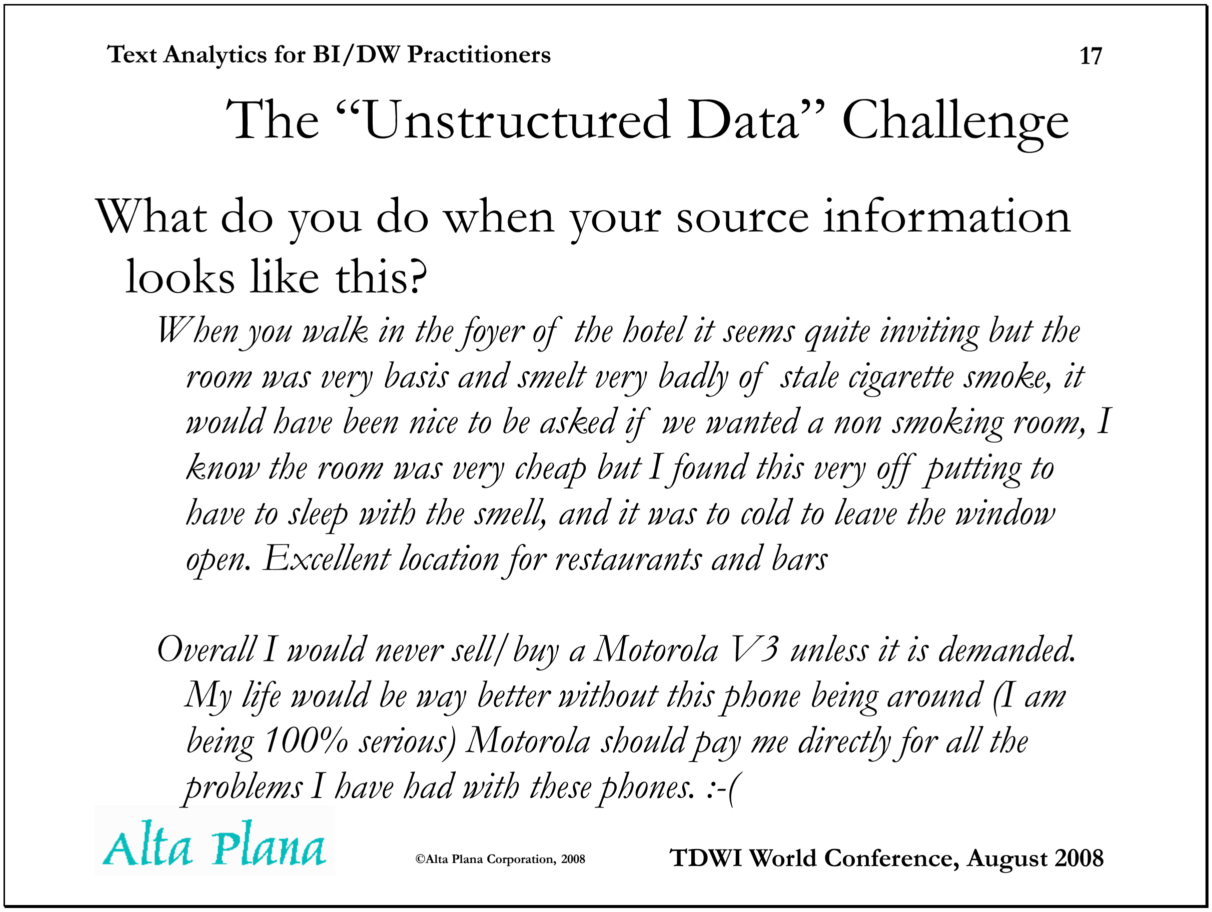
www.stanford.edu/%7ernusse/wntwindow.html

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

This depicts genetic markets on a genome. The text is from an associated abstract. This type of life-science information is mined for protein-interaction data by researchers working in drug-discovery processes.



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

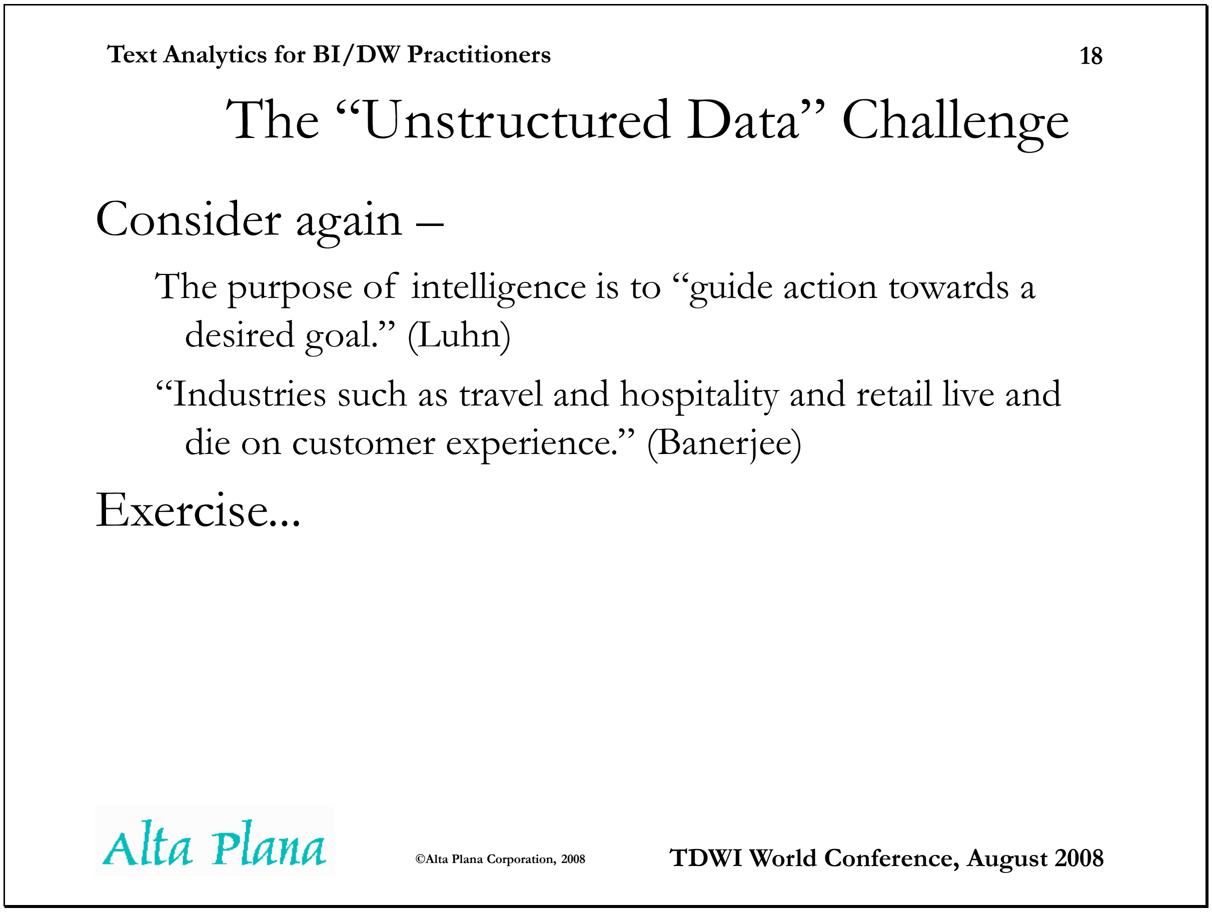
Overall I would never sell/ buy a Motorola V3 unless it is demanded. My life would be way better without this phone being around (I am being 100% serious) Motorola should pay me directly for all the problems I have had with these phones. :-(

These are excerpts of longer comments posted in on-line forums, reprinted here verbatim.

What would you do if you were a customer-support manager and had to deal with these messages?

What about if you needed to make something of 40 such messages?

What about 40 new messages each day, posted on a variety of on-line forums, blogs, and review sites and transmitted to your company by e-mail and via comment forms and surveys?



The "Unstructured Data" Challenge

Consider again –

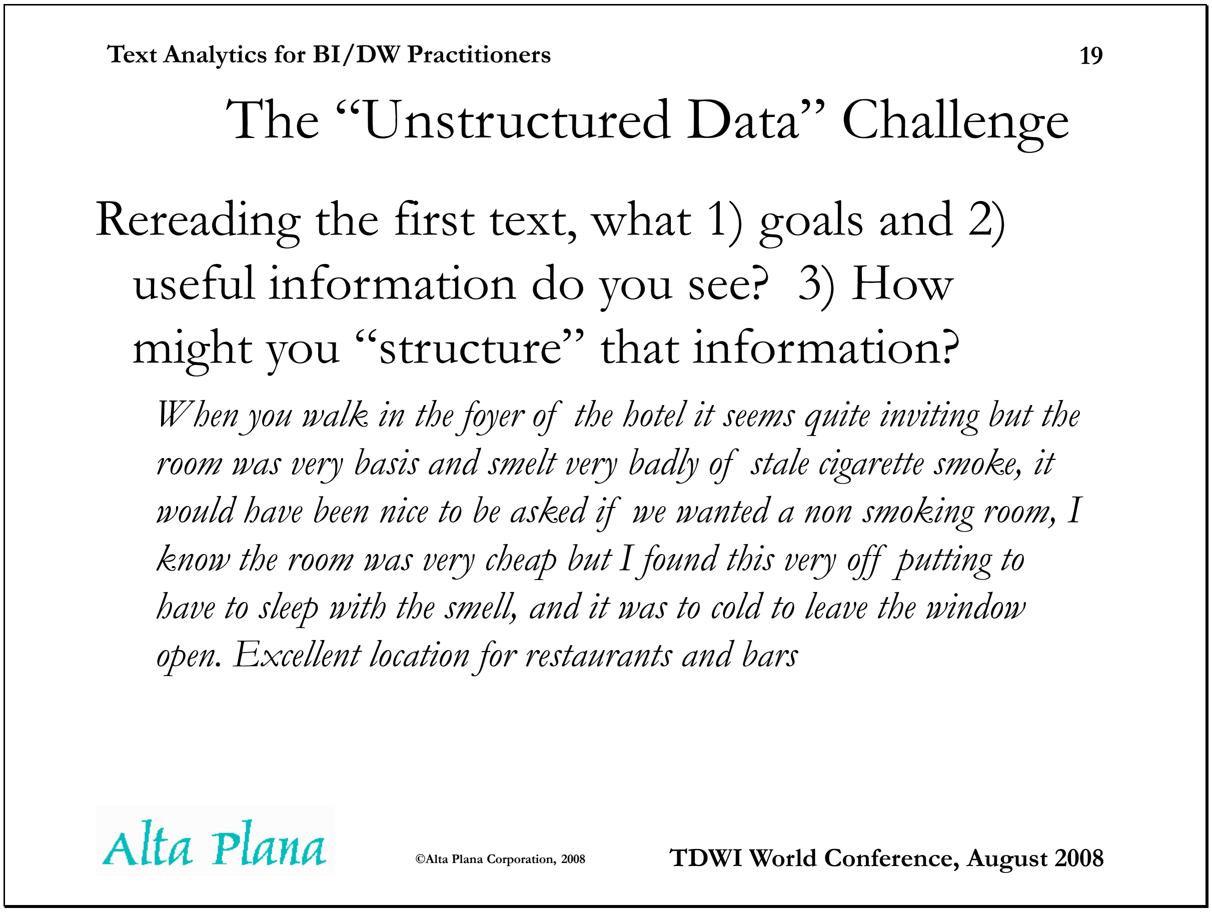
- The purpose of intelligence is to "guide action towards a desired goal." (Luhn)
- "Industries such as travel and hospitality and retail live and die on customer experience." (Banerjee)

Exercise...

Focus on action.

Focus on business drivers.

Find information (and analysis methods) that can support decisions in the name of business drivers.

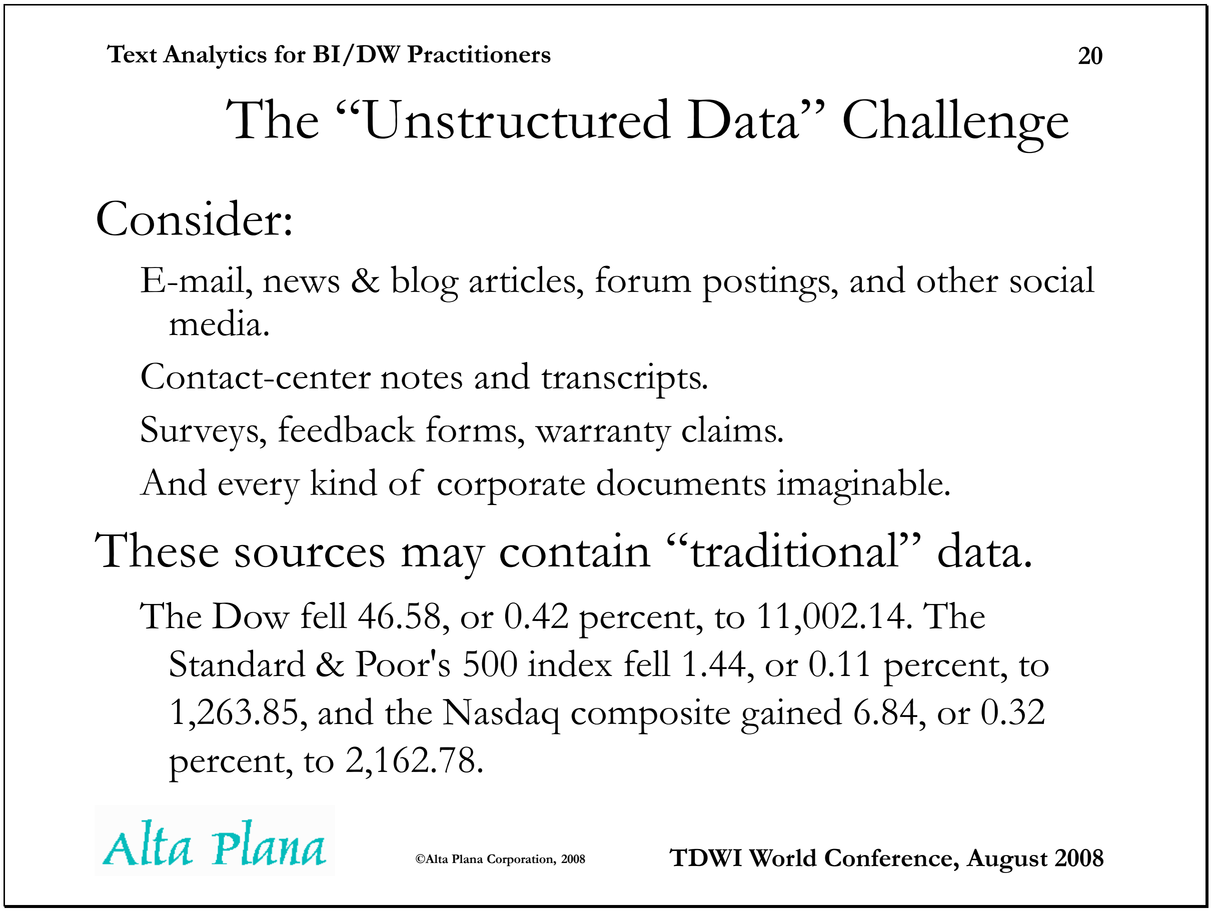


The “Unstructured Data” Challenge

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 goals we’re looking for here are business goals. Useful information is information that can be turned into data to contribute in some way toward analysis that help us derive actions that can help us reach those goals. Structuring information here means, essentially, putting it in a form, with enough descriptive semantics (meaning), that will make it susceptible to analytical methods.



The “Unstructured Data” Challenge

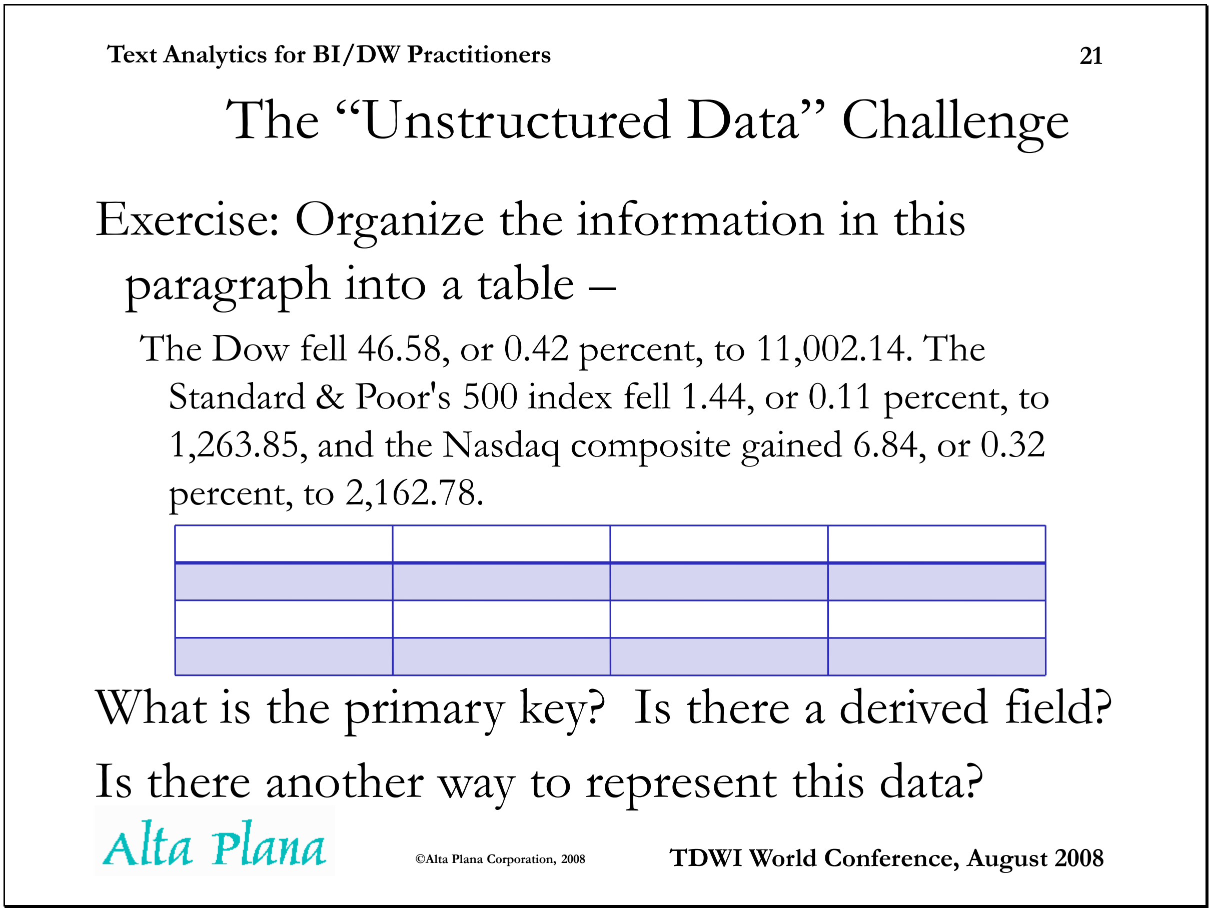
Consider:

- 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” sources listed here are those that appertain to CRM and Voice of the Customer type applications. There are other sources, not listed, that would be used in other application domains such as life science, intelligence and law enforcement, insurance, compliance and electronic (legal) discovery, etc.



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.

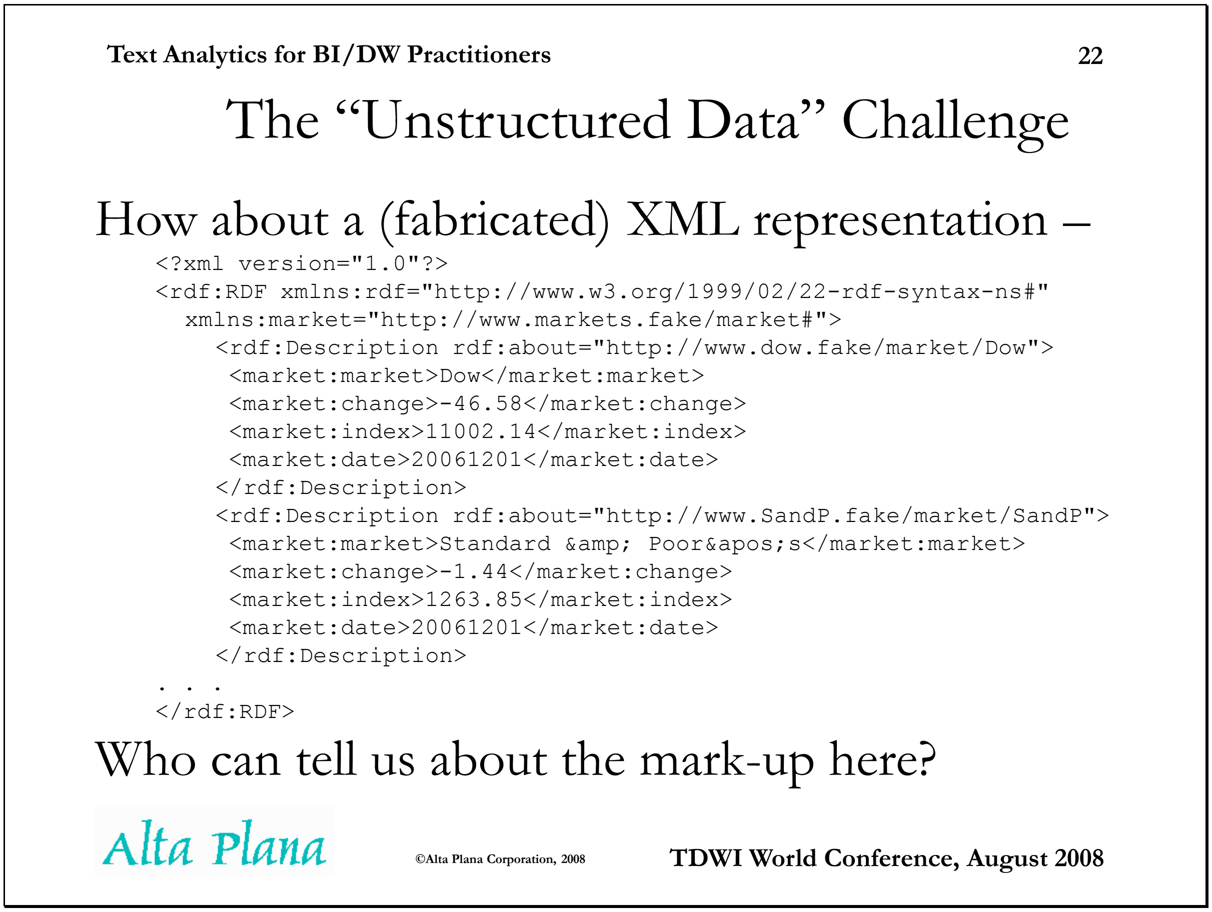
What is the primary key? Is there a derived field?

Is there another way to represent this data?

A table is the basic data representation in a relational model.

The primary key is a table column that contains a unique value for each row of the table. Key fields in general are used to “join” tables.

The value of a derived field may be computed via a formula from the value(s) of one or more other fields.



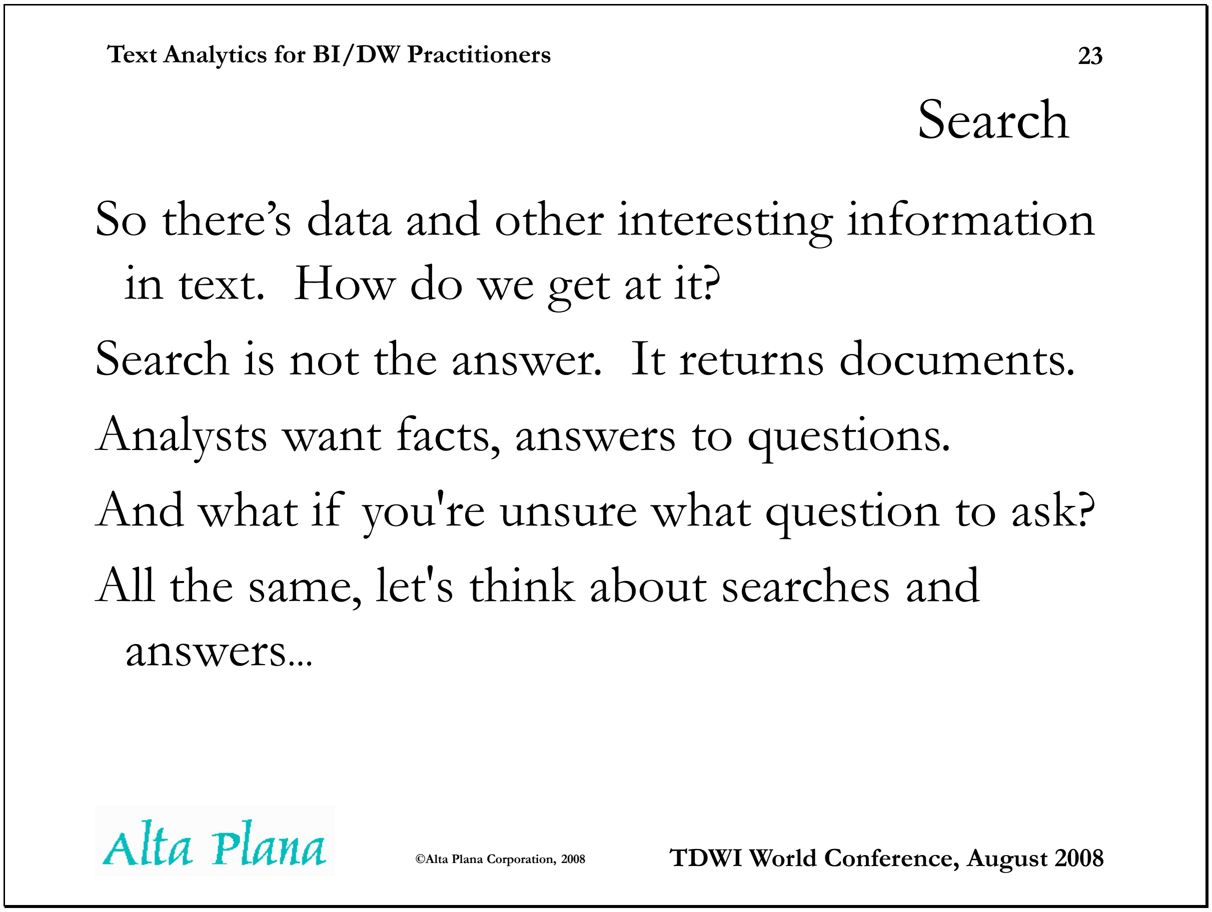
The “Unstructured Data” Challenge

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; Poor&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?

XML is the Extensible Markup Language, a mechanism for hierarchically structuring information via a tag schema defined in some name space.



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…

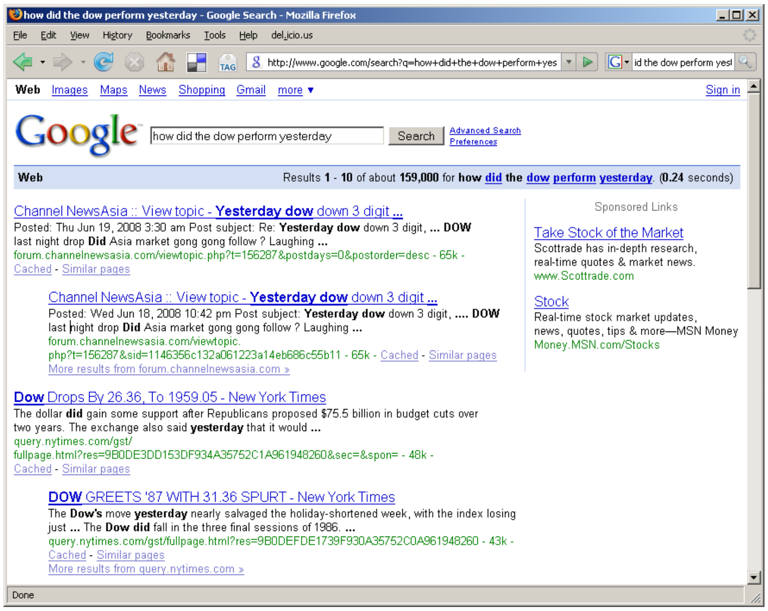
Search

Relevance?

Concepts?

Articles from a forum site

Articles from 1987


how did the dow perform yesterday - Google Search - Mozilla Firefox
File Edit View History Bookmarks Tools Help del.icio.us
TAG
http://www.google.com/search?q=how+did+the+dow+perform+yes
id the dow perform yest
Web Images Maps News Shopping Gmail more ▼
Sign in
Google
how did the dow perform yesterday
Search
Advanced Search
Preferences
Done

Web Results **1** - **10** of about **159,000** for **how did the dow perform yesterday**. (**0.24** seconds)

Channel NewsAsia :: View topic - **Yesterday dow** down 3 digit **...**
Posted: Thu Jun 19, 2008 3:30 am Post subject: Re: **Yesterday dow** down 3 digit, **... DOW** last night drop **Did** Asia market gong gong follow ? Laughing **...**
forum.channelnewsasia.com/viewtopic.php?t=156287&postdays=0&postorder=desc - 65k - Cached - Similar pages

Channel NewsAsia :: View topic - **Yesterday dow** down 3 digit **...**
Posted: Wed Jun 18, 2008 10:42 pm Post subject: **Yesterday dow** down 3 digit, **.... DOW** last night drop **Did** Asia market gong gong follow ? Laughing **...**
forum.channelnewsasia.com/viewtopic.php?t=156287&sid=1146356c132a061223a14eb686c55b11 - 65k - Cached - Similar pages
More results from forum.channelnewsasia.com »

Dow Drops By 26.36, To 1959.05 - New York Times
The dollar **did** gain some support after Republicans proposed $75.5 billion in budget cuts over two years. The exchange also said **yesterday** that it would **...**
query.nytimes.com/gst/fullpage.html?res=9B0DE3DD153DF934A35752C1A961948260&sec=&spon= - 48k - Cached - Similar pages

DOW GREETS '87 WITH 31.36 SPURT - New York Times
The **Dow's** move **yesterday** nearly salvaged the holiday-shortened week, with the index losing just **...** The **Dow did** fall in the three final sessions of 1986. **...**
query.nytimes.com/gst/fullpage.html?res=9B0DEFDE1739F930A35752C0A961948260 - 43k - Cached - Similar pages
More results from query.nytimes.com »

Sponsored Links

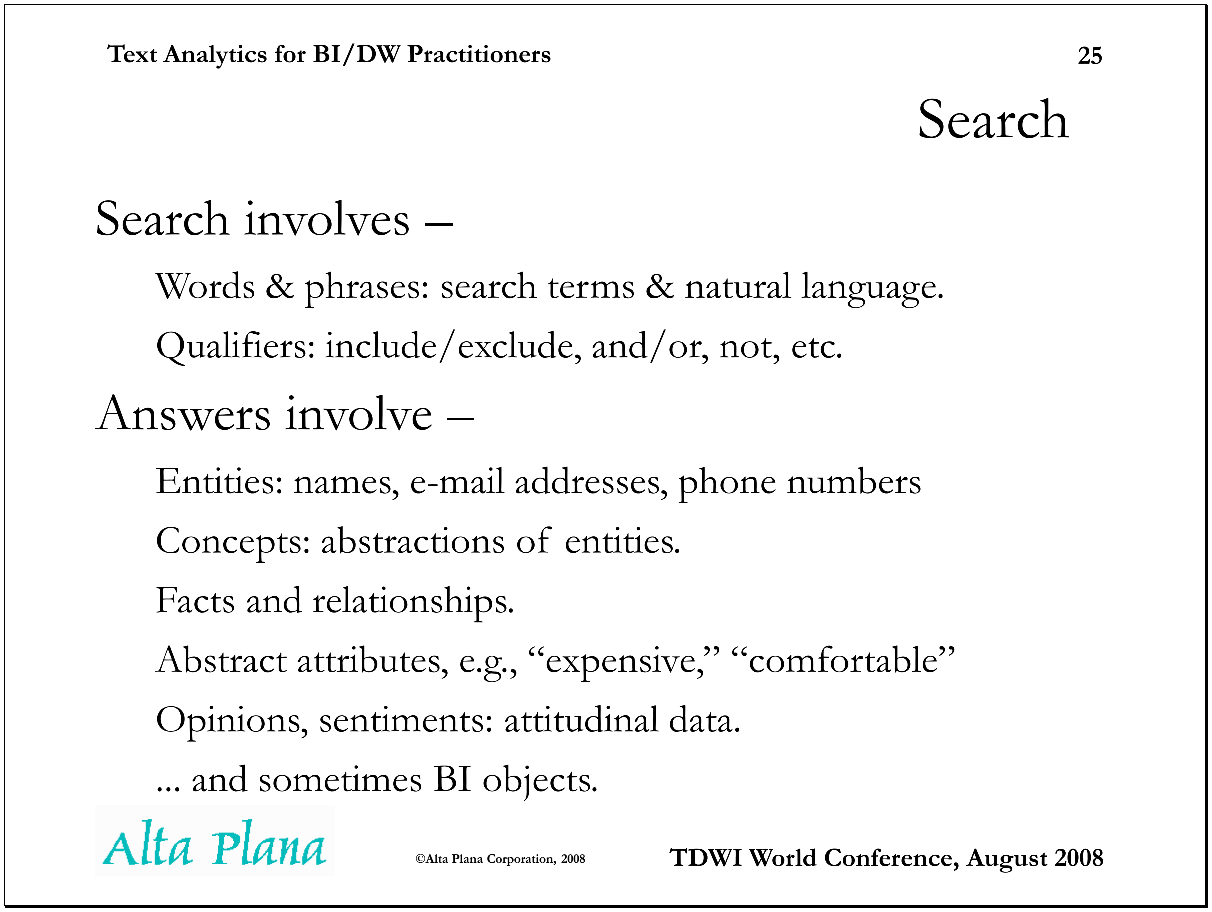
Take Stock of the Market
Scottrade has in-depth research, real-time quotes & market news.
www.Scottrade.com

Stock
Real-time stock market updates, news, quotes, tips & more—MSN Money
Money.MSN.com/Stocks

Google does not successfully answer the straightforward question posed, How did the Dow perform yesterday.

“Yesterday” is a concept. It should not be treated here as a keyword.

Similarly, the search-text entered should be recognizable as a question and not as a string of keywords.



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.

By a BI object, I mean an existing report or chart or cube or bit of analysis output. Optimally, the concept of a BI object would extend to data presentation objects that are generated dynamically. If we had that, we’d have natural-language query.

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?

This about that last example. The "answer" I'm looking for is not a single number. What form/structure does it take?

Search

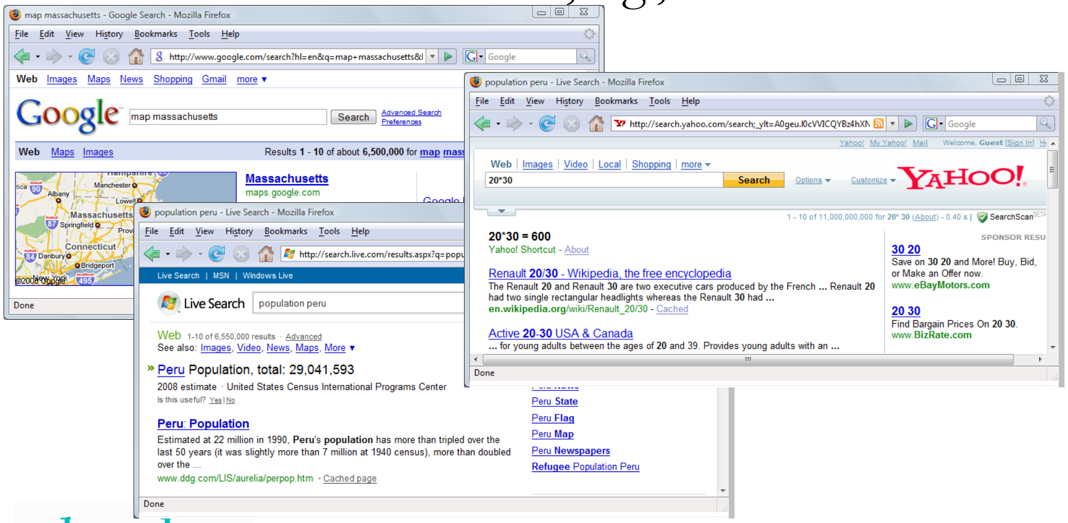
Search is not enough.

- *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 –


map massachusetts - Google Search - Mozilla Firefox
File Edit View History Bookmarks Tools Help
http://www.google.com/search?hl=en&q=map+massachusetts&l
Google
Web Images Maps News Shopping Gmail more ▾
map massachusetts
Search
Advanced Search
Preferences
Web Maps Images
Hampshire
Manchester
Albany
Lowell
Massachusetts
Springfield
Provi
Connecticut
Danbury
Bridgeport
New York
©2008 Google
Done
population peru - Live Search - Mozilla Firefox
File Edit View History Bookmarks Tools Help
http://search.live.com/results.aspx?q=popu
Live Search | MSN | Windows Live
Live Search
population peru
Done
population peru - Live Search - Mozilla Firefox
File Edit View History Bookmarks Tools Help
http://search.yahoo.com/search;_ylt=A0geu.l0cVVICQYBz4hXN
Google
Yahoo! My Yahoo! Mail Welcome, Guest [Sign In]
Web Images Video Local Shopping more ▾
20*30
Search
Options ▾
Customize ▾
YAHOO!
SearchScan BETA
SPONSOR RESU
Done

Results **1** - **10** of about **6,500,000** for **map mass**

Massachusetts
maps.google.com

Web 1-10 of 6,550,000 results · Advanced
See also: Images, Video, News, Maps, More ▾

» Peru Population, total: 29,041,593
2008 estimate · United States Census International Programs Center
Is this useful? Yes | No

Peru: Population
Estimated at 22 million in 1990, **Peru**'s **population** has more than tripled over the last 50 years (it was slightly more than 7 million at 1940 census), more than doubled over the ...
www.ddg.com/LIS/aurelia/perpop.htm · Cached page

Peru **State**
Peru **Flag**
Peru **Map**
Peru **Newspapers**
Refugee Population Peru

1 - 10 of 11,000,000,000 for **20* 30** (About) - 0.40 s

20*30 = 600
Yahoo! Shortcut - About

Renault **20/30** - Wikipedia, the free encyclopedia
The Renault **20** and Renault **30** are two executive cars produced by the French ... Renault **20** had two single rectangular headlights whereas the Renault **30** had ...
en.wikipedia.org/wiki/Renault_20/30 - Cached

Active **20-30** USA & Canada
... for young adults between the ages of **20** and 39. Provides young adults with an ...

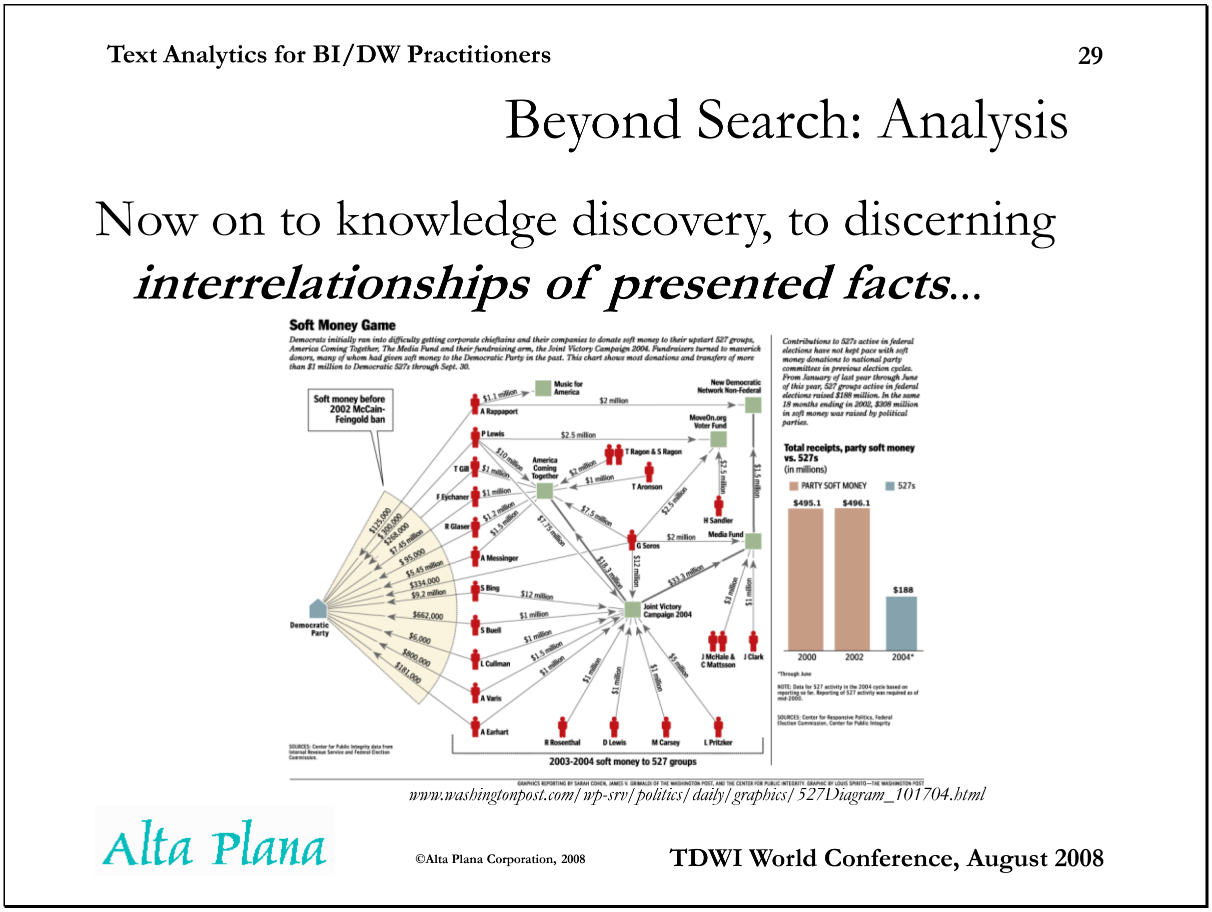
30 20
Save on **30 20** and More! Buy, Bid, or Make an Offer now.
www.**eBayMotors.com**

20 30
Find Bargain Prices On **20 30**.
www.**BizRate.com**

Alta Plana

Even though Google is way out in the lead on the Web, it and Yahoo and Live search have roughly equivalent capabilities where question-answering is concerned.

One can infer that they can combine dictionaries and pattern-matching to detect at least some questions (as opposed to search terms) sent their way.



Beyond Search: Analysis

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

Soft Money Game

Democrats initially ran into difficulty getting corporate chieftains and their companies to donate soft money to their upstart 527 groups, America Coming Together, The Media Fund and their fundraising arm, the Joint Victory Campaign 2004. Fundraisers turned to maverick donors, many of whom had given soft money to the Democratic Party in the past. This chart shows most donations and transfers of more than $1 million to Democratic 527s through Sept. 30.

SOURCES: Center for Public Integrity data from Internal Revenue Service and Federal Election Commission.

Contributions to 527s active in federal elections have not kept pace with soft money donations to national party committees in previous election cycles. From January of last year through June of this year, 527 groups active in federal elections raised $188 million. In the same 18 months ending in 2002, $308 million in soft money was raised by political parties.

Total receipts, party soft money vs. 527s
(in millions)

*Through June

NOTE: Data for 527 activity in the 2004 cycle based on reporting so far. Reporting of 527 activity was required as of mid-2000.

SOURCES: Center for Responsive Politics, Federal Election Commission, Center for Public Integrity

GRAPHICS REPORTING BY SARAH COHEN, JAMES V. GRIMALDI OF THE WASHINGTON POST, AND THE CENTER FOR PUBLIC INTEGRITY. GRAPHIC BY LOUIS SPIRITO—THE WASHINGTON POST

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

Alta Plana

“Interrelationships of presented facts” comes from the Luhn paper. This Washington Post campaign-donor illustration presents facts in a number of forms and does interrelate them. Is this information “actionable”? It depends, I suppose, on who you are and what actions you’re responsible for.

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.

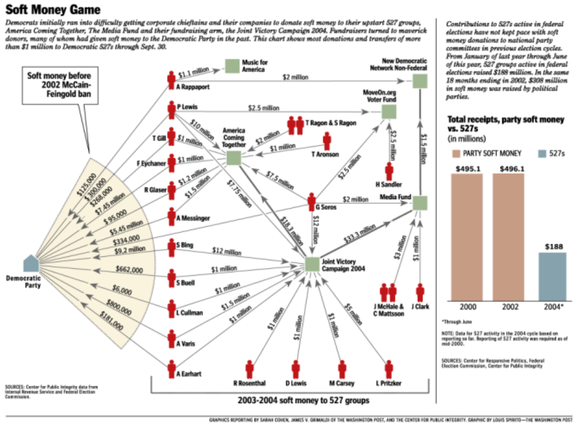
Find something you and a person next to you have in common, e.g., school attended, the industry you work in, favorite sport, other.

(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.~~



Soft Money Game

Democrats initially ran into difficulty getting corporate chieftains and their companies to donate soft money to their upstart 527 groups, America Coming Together, The Media Fund and their fundraising arm, the Joint Victory Campaign 2004. Fundraisers turned to maverick donors, many of whom had given soft money to the Democratic Party in the past. This chart shows most donations and transfers of more than $1 million to Democratic 527s through Sept. 30.

Contributions to 527s active in federal elections have not kept pace with soft money donations to national party committees in previous election cycles. From January of last year through June of this year, 527 groups active in federal elections raised $188 million. In the same 18 months ending in 2002, $308 million in soft money was raised by political parties.

SOURCES: Center for Public Integrity data from Internal Revenue Service and Federal Election Commission.

*Through June

NOTE: Data for 527 activity in the 2004 cycle based on reporting so far. Reporting of 527 activity was required as of mid-2000.

SOURCES: Center for Responsive Politics, Federal Election Commission, Center for Public Integrity

GRAPHICS REPORTING BY SARAH COHEN, JAMES V. GRIMALDI OF THE WASHINGTON POST, AND THE CENTER FOR PUBLIC INTEGRITY. GRAPHIC BY LOUIS SPIRITO—THE WASHINGTON POST

Text Mining = Knowledge Discovery in Text.

I present here four basic data-mining techniques. Two others are struck out because they really wouldn't apply to textual information (even if they could apply to numerical information extracted from textual sources).

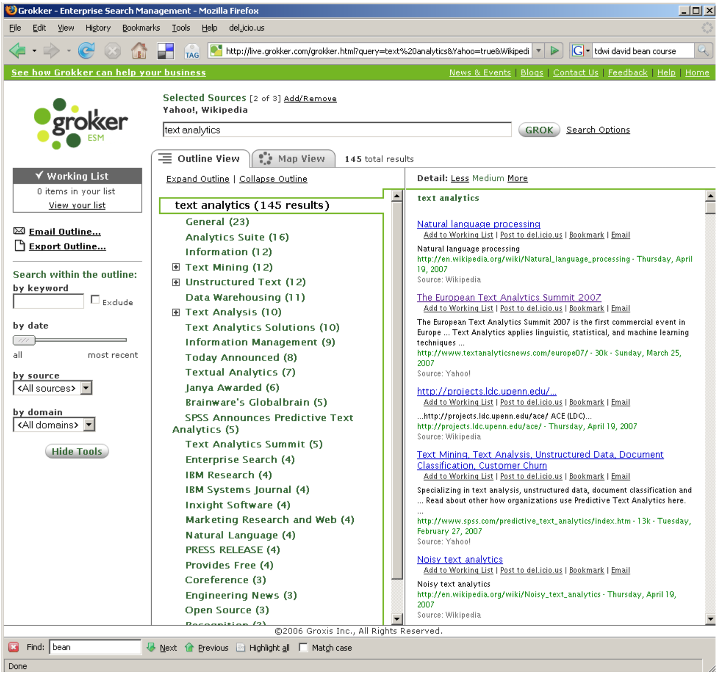
Text Mining

	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*

Presentation of search results can be enhanced by discovery.

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


Grokker - Enterprise Search Management - Mozilla Firefox
File Edit View History Bookmarks Tools Help del.icio.us
TAG
See how Grokker can help your business
News & Events | Blogs | Contact Us | Feedback | Help | Home
grokker ESM
GROK
Search Options
Outline View
Map View
Expand Outline | Collapse Outline
Email Outline...
Export Outline...
Hide Tools
Next
Previous
Highlight all
Add to Working List | Post to del.icio.us | Bookmark | Email

http://live.grokker.com/grokker.html?query=text%20analytics&Yahoo=true&Wikipedi

tdwi david bean course

Selected Sources [2 of 3] Add/Remove
Yahoo!, Wikipedia

text analytics

145 total results

✓ Working List
0 items in your list
View your list

Search within the outline:
by keyword □ Exclude
by date
all most recent
by source <All sources>
by domain <All domains>

text analytics (145 results)
- General (23)
- Analytics Suite (16)
- Information (12)
- ⊞ Text Mining (12)
- ⊞ Unstructured Text (12)
- Data Warehousing (11)
- ⊞ Text Analysis (10)
- Text Analytics Solutions (10)
- Information Management (9)
- Today Announced (8)
- Textual Analytics (7)
- Janya Awarded (6)
- Brainware's Globalbrain (5)
- SPSS Announces Predictive Text Analytics (5)
- Text Analytics Summit (5)
- Enterprise Search (4)
- IBM Research (4)
- IBM Systems Journal (4)
- Inxight Software (4)
- Marketing Research and Web (4)
- Natural Language (4)
- PRESS RELEASE (4)
- Provides Free (4)
- Coreference (3)
- Engineering News (3)
- Open Source (3)
- Recognition (3)

Detail: Less Medium More

text analytics

Natural language processing
Add to Working List | Post to del.icio.us | Bookmark | Email
Natural language processing
http://en.wikipedia.org/wiki/Natural_language_processing - Thursday, April 19, 2007
Source: Wikipedia

The European Text Analytics Summit 2007
Add to Working List | Post to del.icio.us | Bookmark | Email
The European Text Analytics Summit 2007 is the first commercial event in Europe ... Text Analytics applies linguistic, statistical, and machine learning techniques ...
http://www.textanalyticsnews.com/europe07/ - 30k - Sunday, March 25, 2007
Source: Yahoo!

http://projects.ldc.upenn.edu/...
Add to Working List | Post to del.icio.us | Bookmark | Email
...http://projects.ldc.upenn.edu/ace/ ACE (LDC)...
http://projects.ldc.upenn.edu/ace/ - Thursday, April 19, 2007
Source: Wikipedia

Text Mining, Text Analysis, Unstructured Data, Document Classification, Customer Churn
Add to Working List | Post to del.icio.us | Bookmark | Email
Specializing in text analysis, unstructured data, document classification and ... Read about other how organizations use Predictive Text Analytics here. ...
http://www.spss.com/predictive_text_analytics/index.htm - 13k - Tuesday, February 27, 2007
Source: Yahoo!

Noisy text analytics
Add to Working List | Post to del.icio.us | Bookmark | Email
Noisy text analytics
http://en.wikipedia.org/wiki/Noisy_text_analytics - Thursday, April 19, 2007
Source: Wikipedia

©2006 Groxis Inc., All Rights Reserved.

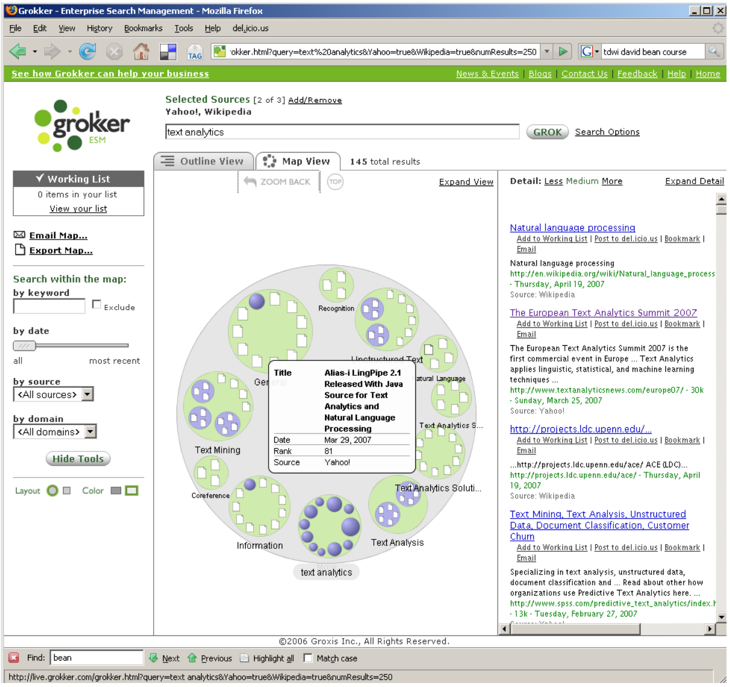
Find: bean □ Match case

Done

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.


Grokker - Enterprise Search Management - Mozilla Firefox

okker.html?query=text%20analytics&Yahoo=true&Wikipedia=true&numResults=250

tdwi david bean course

Selected Sources [2 of 3] Add/Remove
Yahoo!, Wikipedia

text analytics

145 total results

Working List
0 items in your list
View your list

Search within the map:
by keyword ☐ Exclude
by date
all most recent
by source <All sources>
by domain <All domains>

Title	Alias-i LingPipe 2.1 Released With Java Source for Text Analytics and Natural Language Processing
Date	Mar 29, 2007
Rank	81
Source	Yahoo!

Detail: Less Medium More

Natural language processing
Add to Working List | Post to del.icio.us | Bookmark | Email
Natural language processing
http://en.wikipedia.org/wiki/Natural_language_process - Thursday, April 19, 2007
Source: Wikipedia

The European Text Analytics Summit 2007
Add to Working List | Post to del.icio.us | Bookmark | Email
The European Text Analytics Summit 2007 is the first commercial event in Europe ... Text Analytics applies linguistic, statistical, and machine learning techniques ...
http://www.textanalyticsnews.com/europe07/ - 30k - Sunday, March 25, 2007
Source: Yahoo!

http://projects.ldc.upenn.edu/...
Add to Working List | Post to del.icio.us | Bookmark | Email
...http://projects.ldc.upenn.edu/ace/ ACE (LDC)...
http://projects.ldc.upenn.edu/ace/ - Thursday, April 19, 2007
Source: Wikipedia

Text Mining, Text Analysis, Unstructured Data, Document Classification, Customer Churn
Add to Working List | Post to del.icio.us | Bookmark | Email
Specializing in text analysis, unstructured data, document classification and ... Read about other how organizations use Predictive Text Analytics here. ...
http://www.spss.com/predictive_text_analytics/index.h - 13k - Tuesday, February 27, 2007

©2006 Groxis Inc., All Rights Reserved.

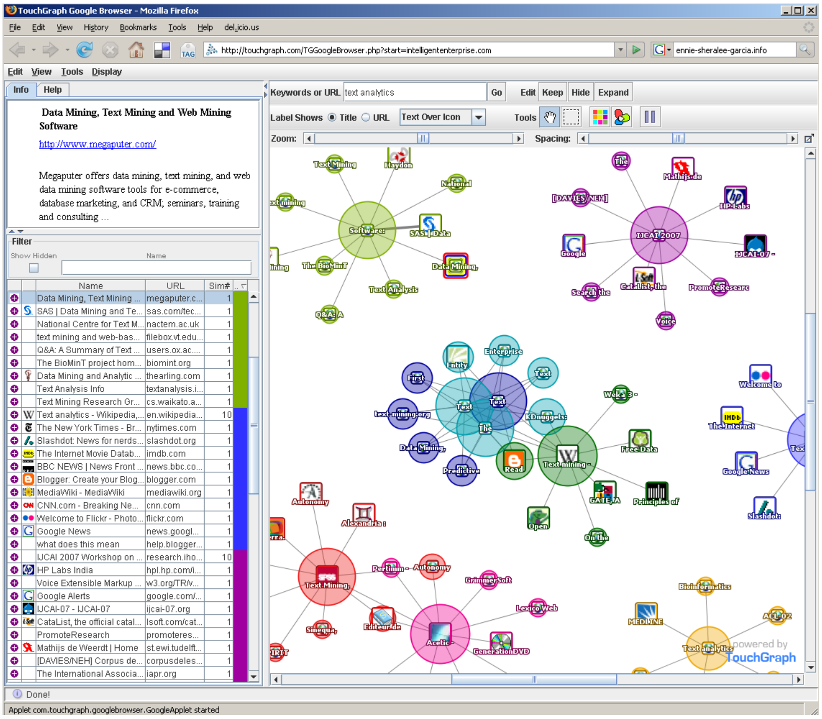
Find: bean

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

More results clustering...

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

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



TouchGraph Google Browser - Mozilla Firefox

http://touchgraph.com/TGGoogleBrowser.php?start=intelligententerprise.com

Keywords or URL: text analytics

Label Shows ● Title ○ URL   Text Over Icon

Data Mining, Text Mining and Web Mining Software

http://www.megaputer.com/

Megaputer offers data mining, text mining, and web data mining software tools for e-commerce, database marketing, and CRM; seminars, training and consulting ...

Filter

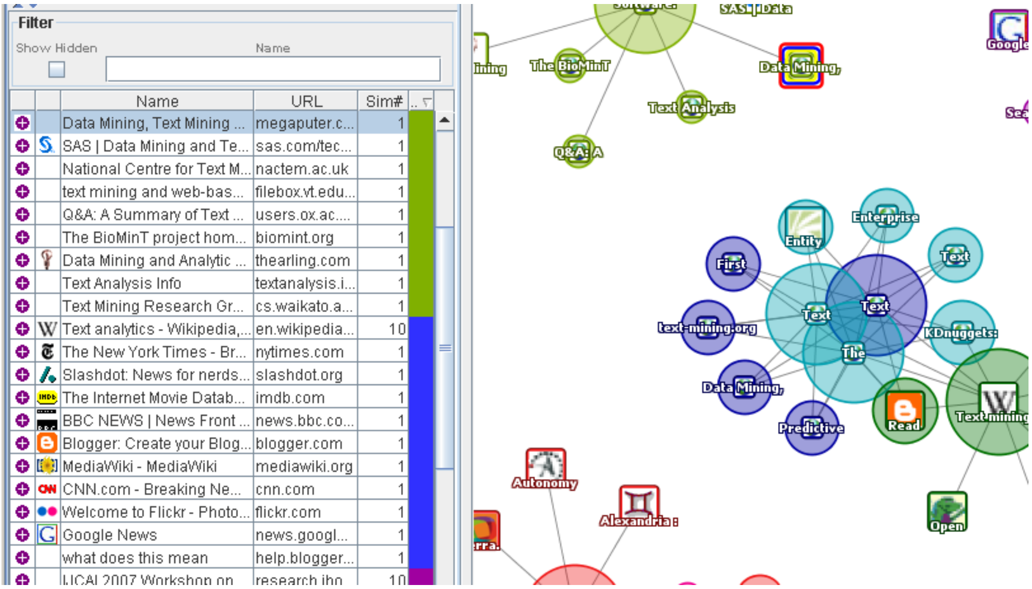
Show Hidden □   Name

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
HP Labs India	hpl.hp.com/i...	1
Voice Extensible Markup ...	w3.org/TR/v...	1
Google Alerts	google.com/...	1
IJCAI-07 - IJCAI-07	ijcai-07.org	1
CataList, the official catal...	lsoft.com/cat...	1
PromoteResearch	promoteres...	1
Mathijs de Weerdt | Home	st.ewi.tudelft...	1
[DAVIES/NEH] Corpus de...	corpusdeles...	1
The International Associa...	iapr.org	1

Done!

Applet com.touchgraph.googlebrowser.GoogleApplet started

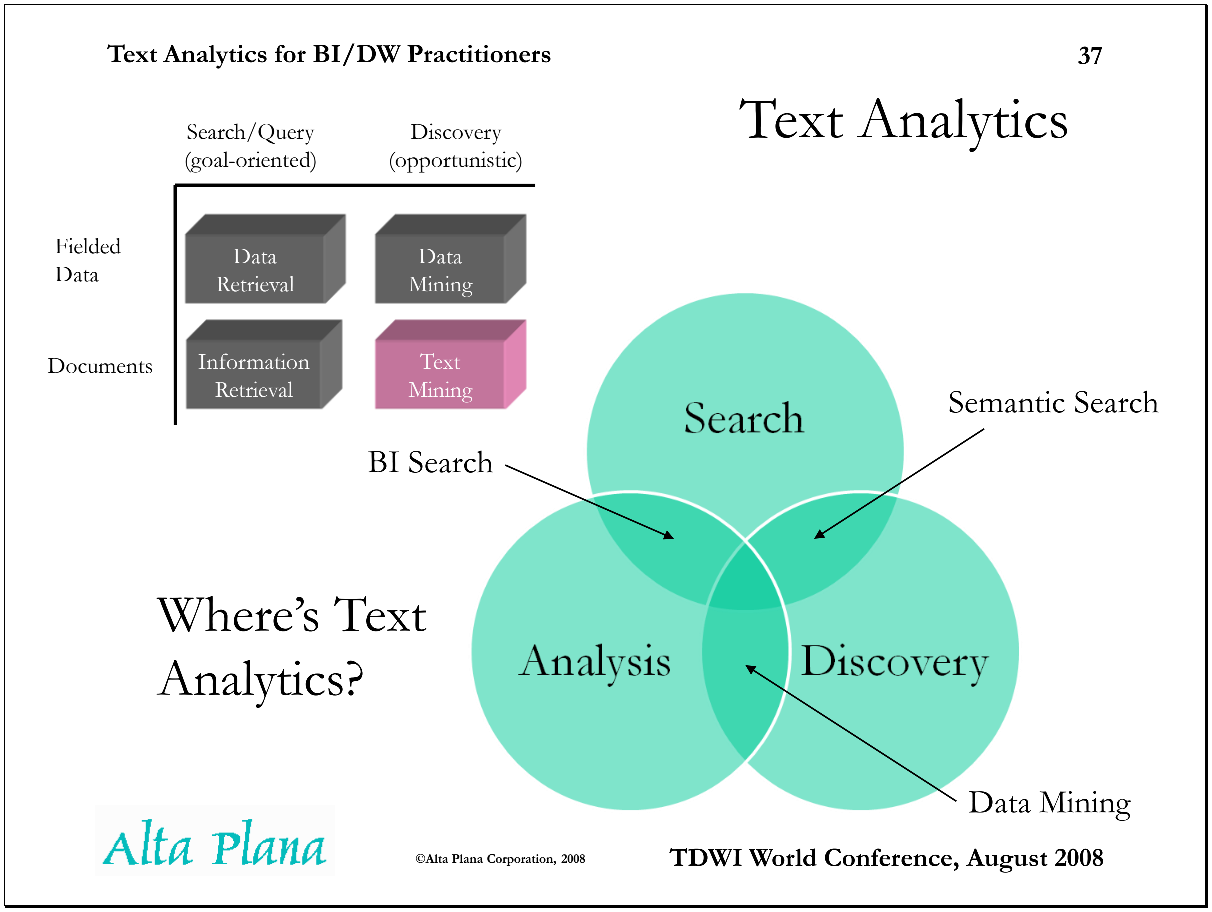
Beyond Search: Analysis



Filter

Show Hidden □ Name

		Name	URL	Sim#	..
⊕		Data Mining, Text Mining ...	megaputer.c...	1	🟩
⊕	S	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	🟩
⊕	W	Text analytics - Wikipedia,...	en.wikipedia...	10	🟦
⊕	𝕿	The New York Times - Br...	nytimes.com	1	🟦
⊕	/.	Slashdot: News for nerds...	slashdot.org	1	🟦
⊕	IMDb	The Internet Movie Datab...	imdb.com	1	🟦
⊕	BBC	BBC NEWS | News Front ...	news.bbc.co...	1	🟦
⊕	B	Blogger: Create your Blog...	blogger.com	1	🟦
⊕		MediaWiki - MediaWiki	mediawiki.org	1	🟦
⊕	CNN	CNN.com - Breaking Ne...	cnn.com	1	🟦
⊕	●●	Welcome to Flickr - Photo...	flickr.com	1	🟦
⊕	G	Google News	news.googl...	1	🟦
⊕		what does this mean	help.blogger...	1	🟦
⊕		IJCAI 2007 Workshop on	research.ib	10	🟪



Text Analytics

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

Where's Text Analytics?

Search

Analysis

Discovery

BI Search

Semantic Search

Data Mining

We earlier used a diagram that showed the relationship between search and discovery and operations on fielded data and on free-text documents. We will take those two methods, search and discovery, and add a third, analysis to the picture. In the intersection of search and analysis we have BI search and in the intersection of search and discovery we have semantic search. Text analytics effectively sits in the intersection of three circles.

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.

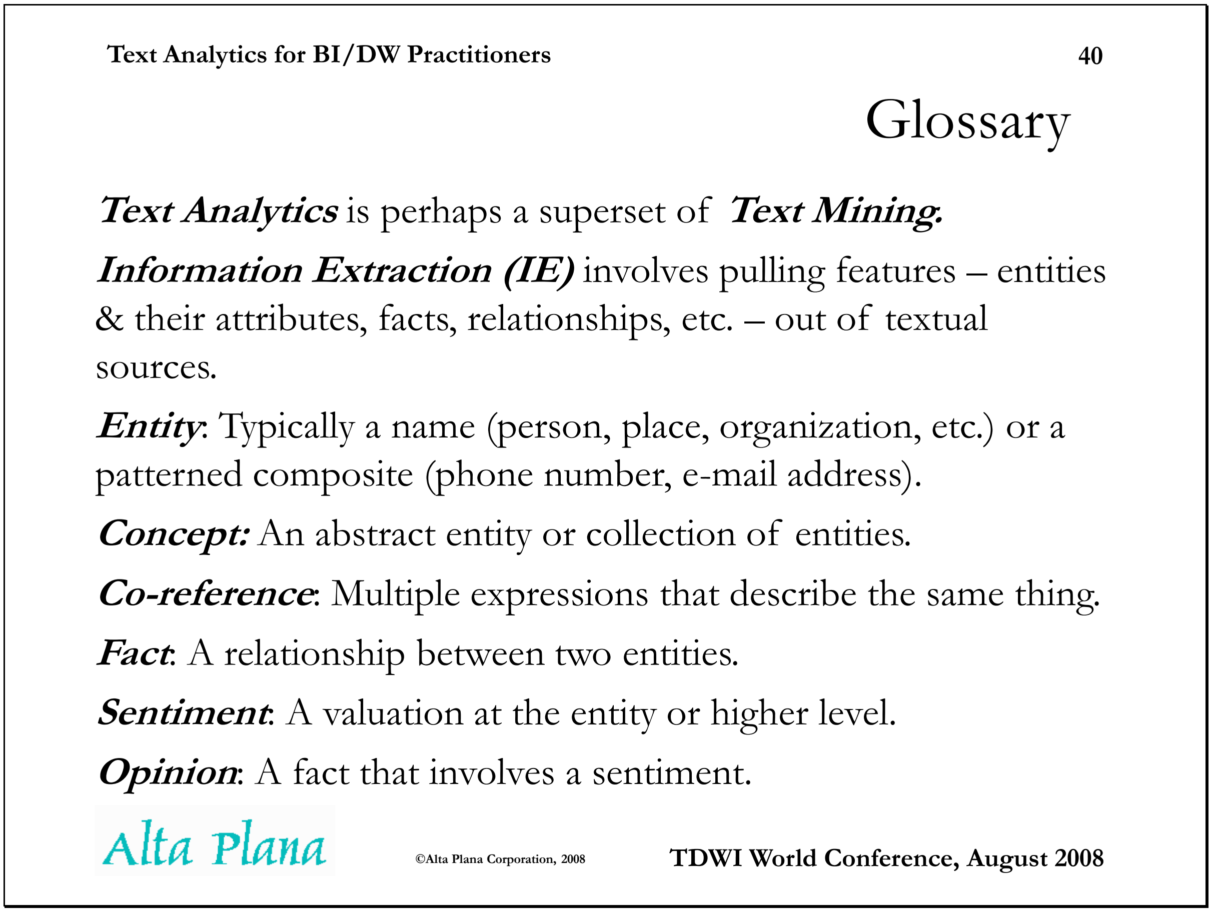
We want to get beyond IR to Information Extraction (IE).

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.

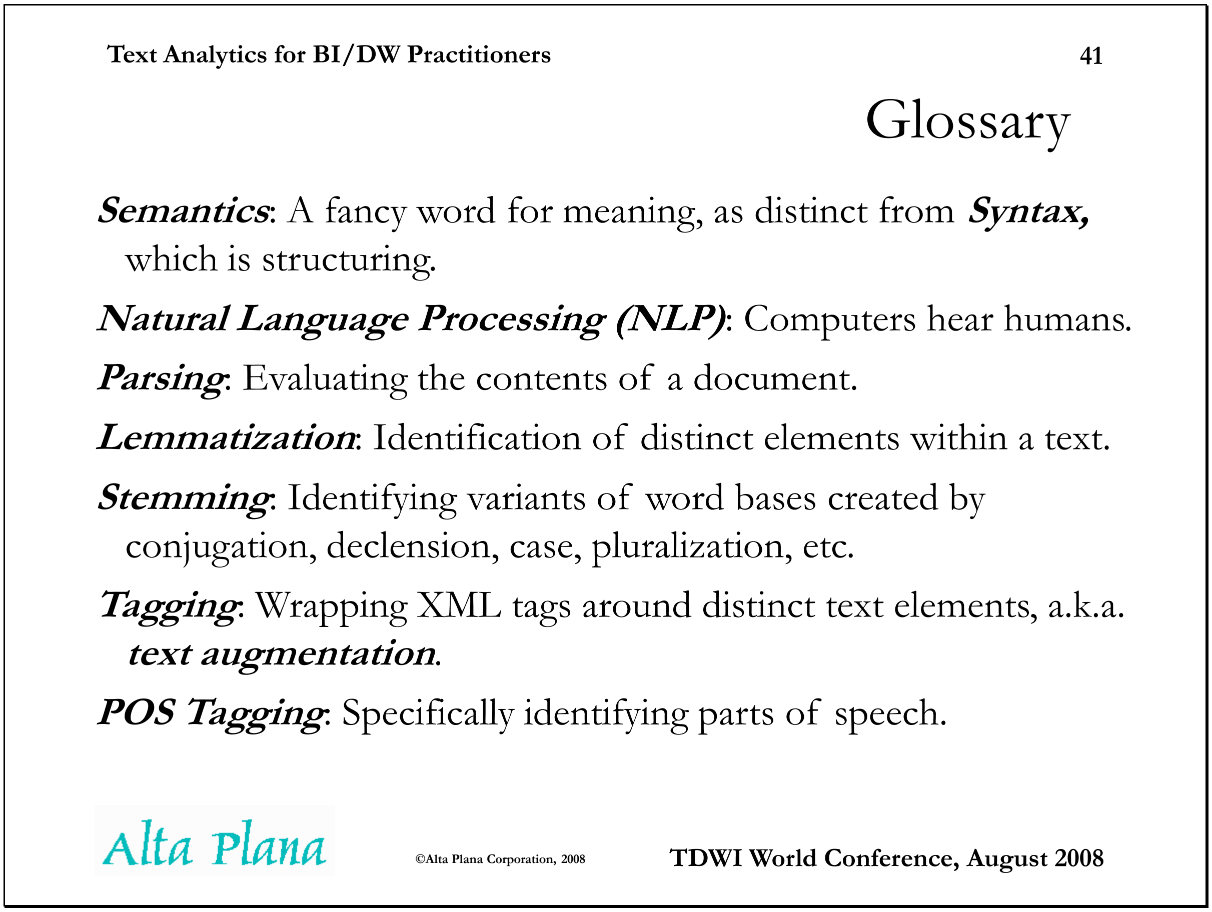
Co-reference: Multiple expressions that describe the same thing.

Fact: A relationship between two entities.

Sentiment: A valuation at the entity or higher level.

Opinion: A fact that involves a sentiment.

Information Extraction and a variety of terms deal with the features – the content – of textual documents.



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.

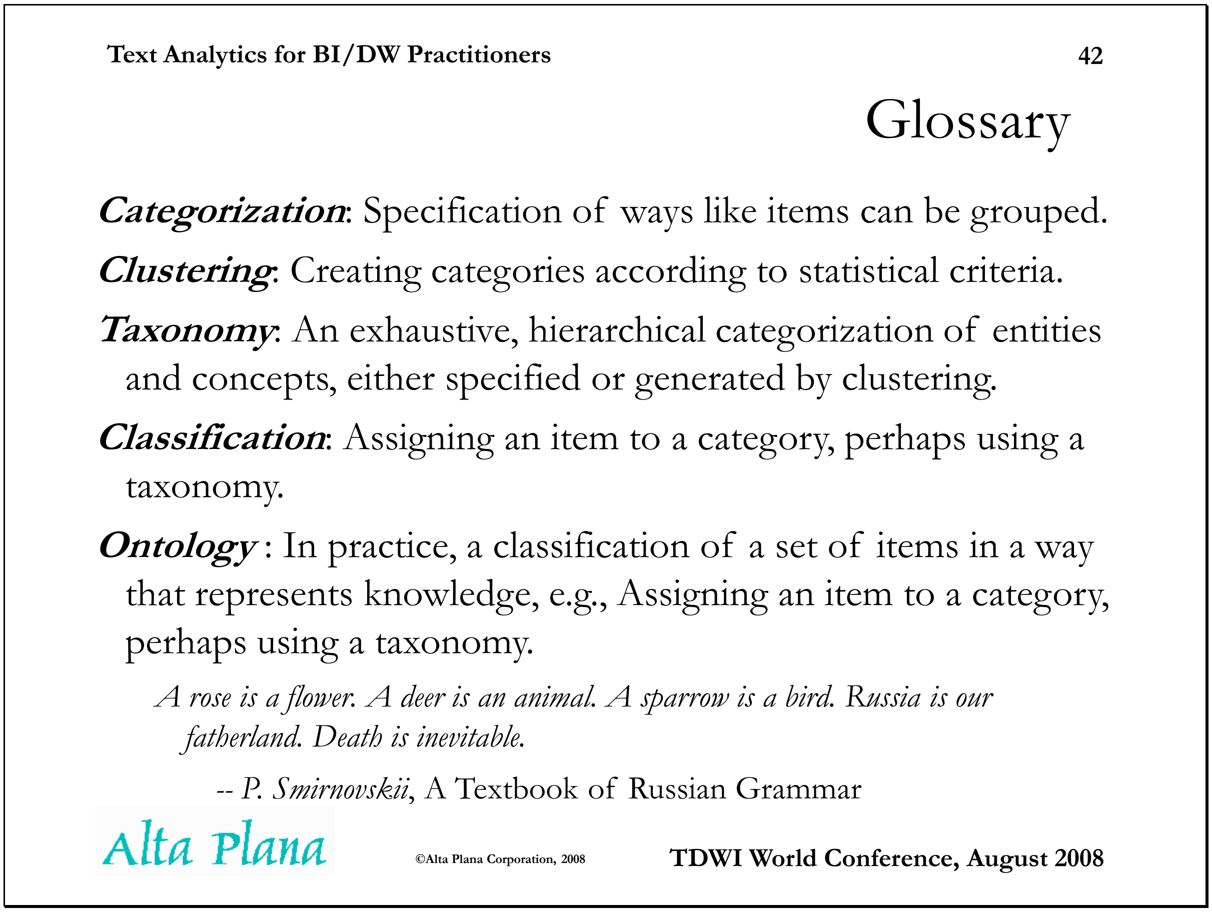
Lemmatization: Identification of distinct elements within a text.

Stemming: Identifying 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.

Text analytics – Natural Language Processing – includes a number of transformational steps that identify and contextualize – provide meaning for – features.



Glossary

Categorization: Specification of ways like items can be grouped.

Clustering: Creating categories according to statistical criteria.

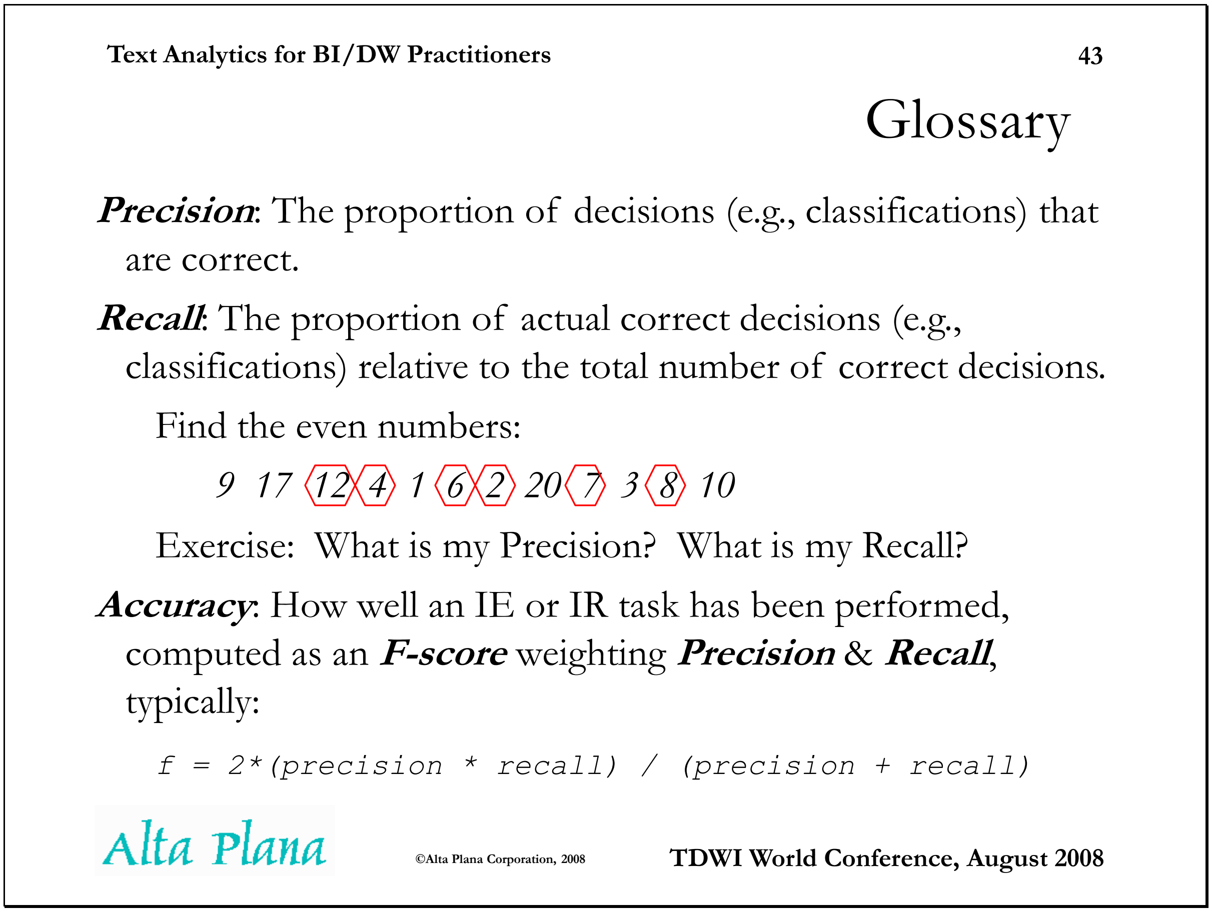
Taxonomy: An exhaustive, hierarchical categorization of entities and concepts, either specified or generated by clustering.

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

Numbers without context are meaningless; for numbers, we typically rely on the column header and the row key value (serving as a record ID) to provide context within an RDBMS table devoted to a particular topic. Features derived from text – and integral documents – are analogously contextualized by creating some form of classification scheme (a.k.a. categorization) and then classifying them in that scheme.



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:

$9\ \ 17\ \boxed{12}\boxed{4}\ 1\ \boxed{6}\boxed{2}\ 20\boxed{7}\ 3\ \boxed{8}\ 10$

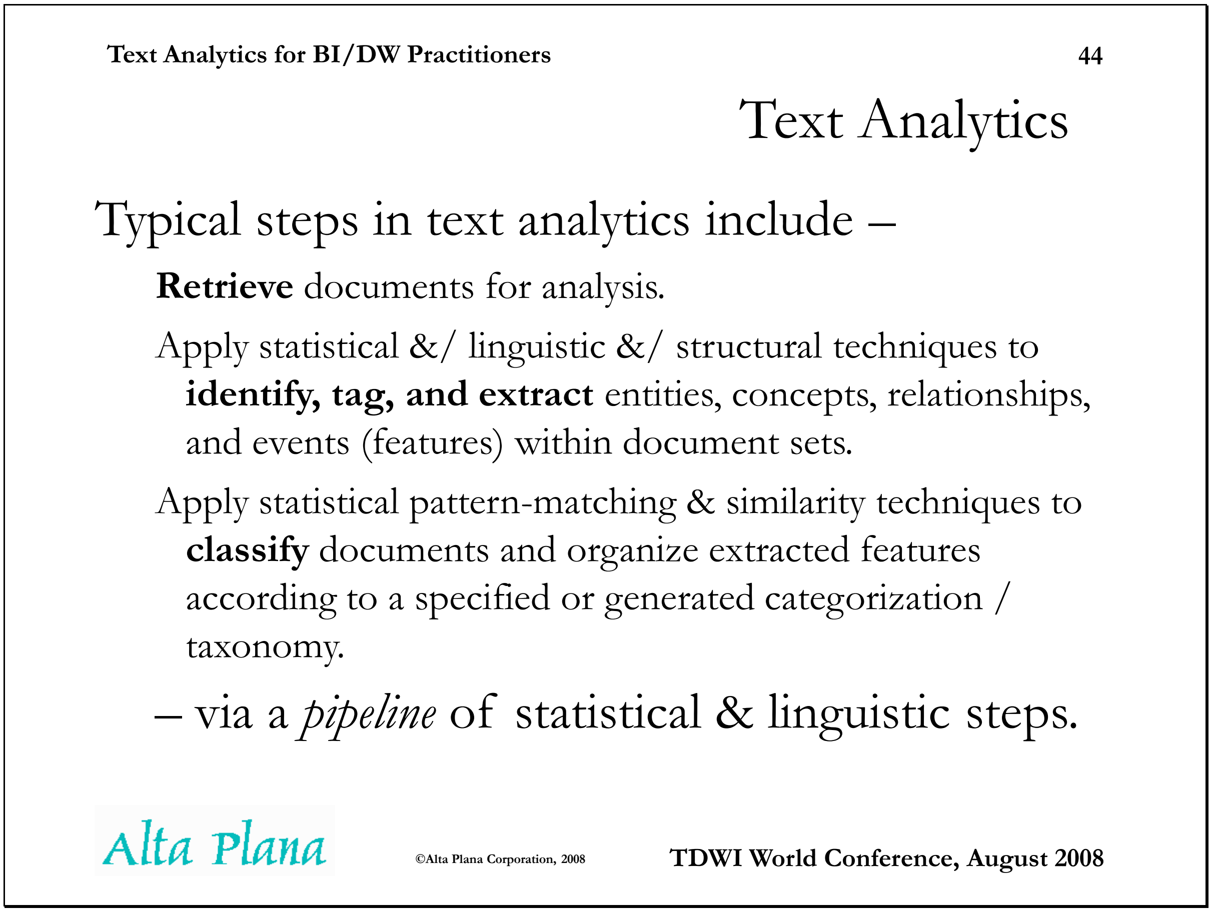
Exercise: What is my Precision? What is my Recall?

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)
```

Alta Plana

Accuracy in classification, information extraction, and other operations on documents and text-derived features is analogous but not identical to the concept of data quality in the data warehousing world.



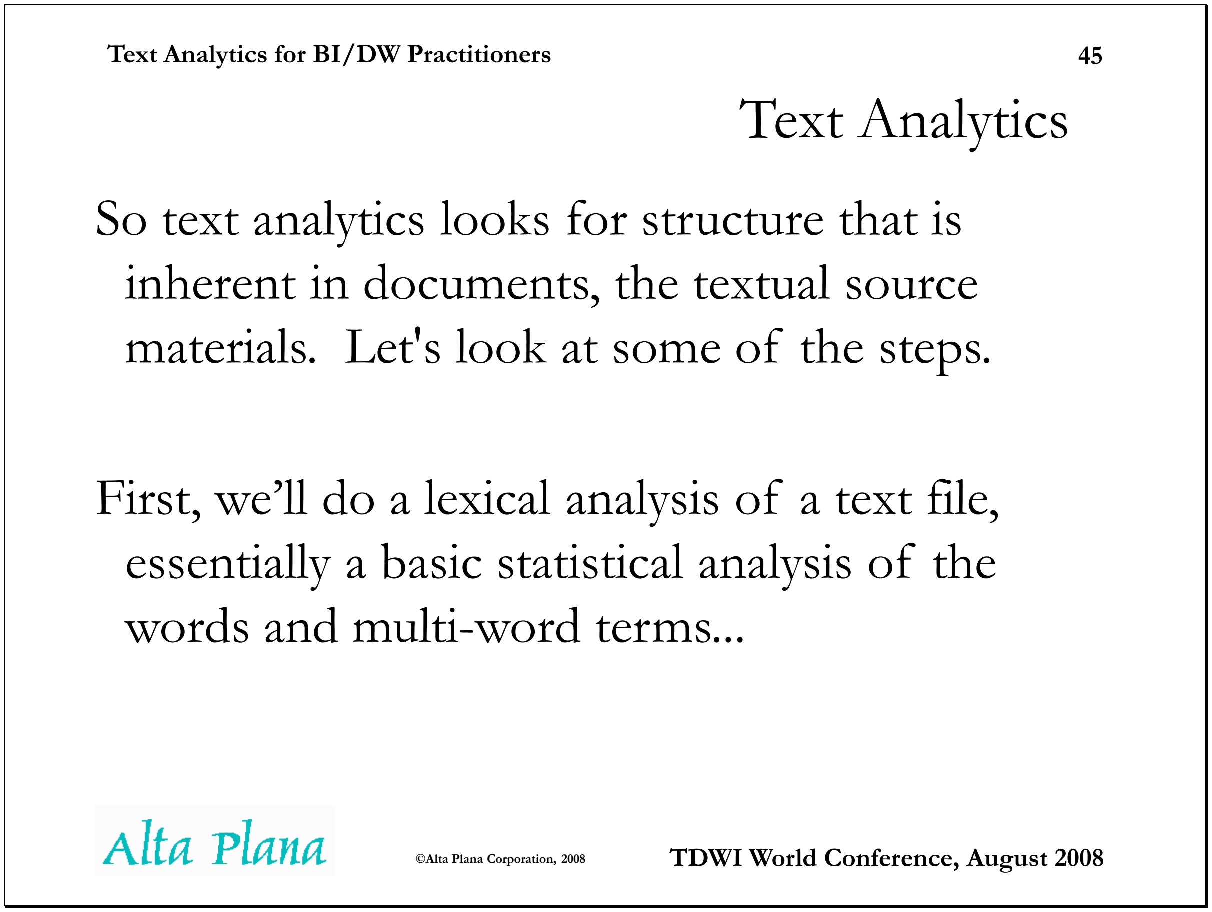
Text Analytics

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.

We here reduce the text-analytics pipeline to three basic elements.

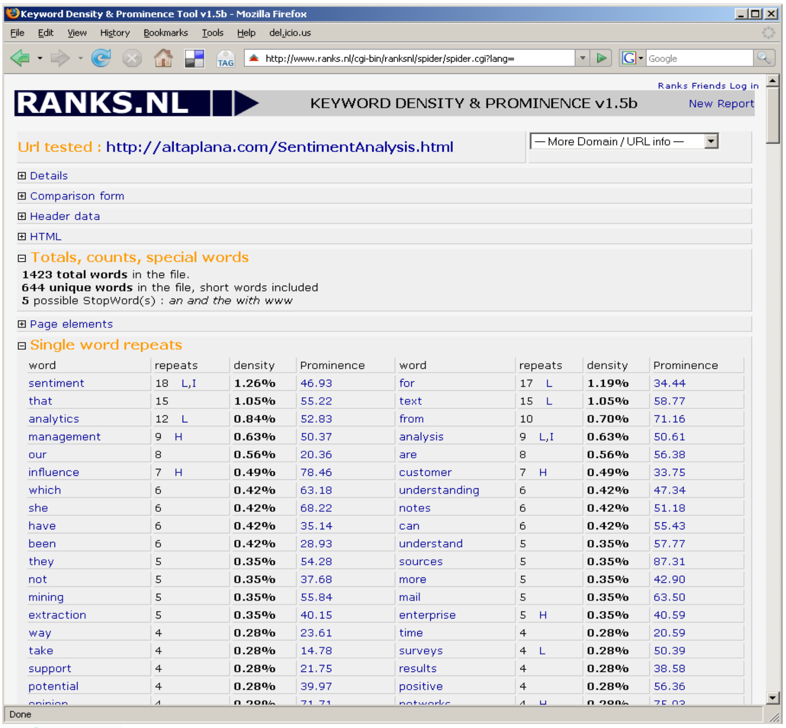


Text Analytics

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...

“Lexical” is defined by Merriam-Webster as “of or relating to words or the vocabulary of a language as distinguished from its grammar and construction.”


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=
Google
Ranks Friends Log in
RANKS.NL
KEYWORD DENSITY & PROMINENCE v1.5b
New Report
— More Domain / URL info —
⊞ Details
⊞ Comparison form
⊞ Header data
⊞ HTML
⊞ Page elements
Done

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

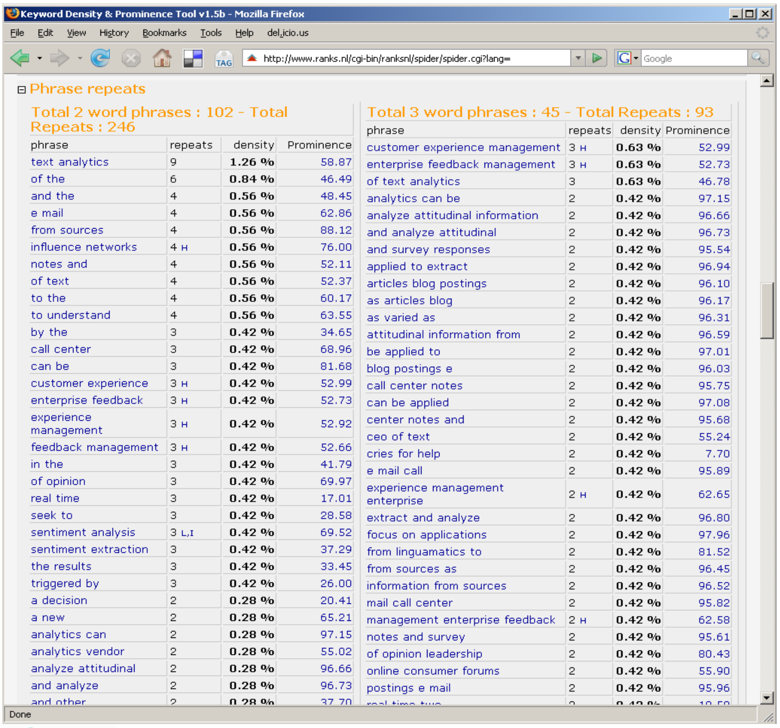
⊟ Totals, counts, special words

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

⊟ 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

This site, Ranks.nl, is on the public Web. It was designed for Search Engine Optimization (SEO), to support efforts to make Web pages findable and highly ranked by Google and the rest. It looks at a variety of the properties of pages. We'll focus on the properties of interest here, namely, the occurrence, frequency, and "weight" of 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=
Google
Done

Phrase repeats

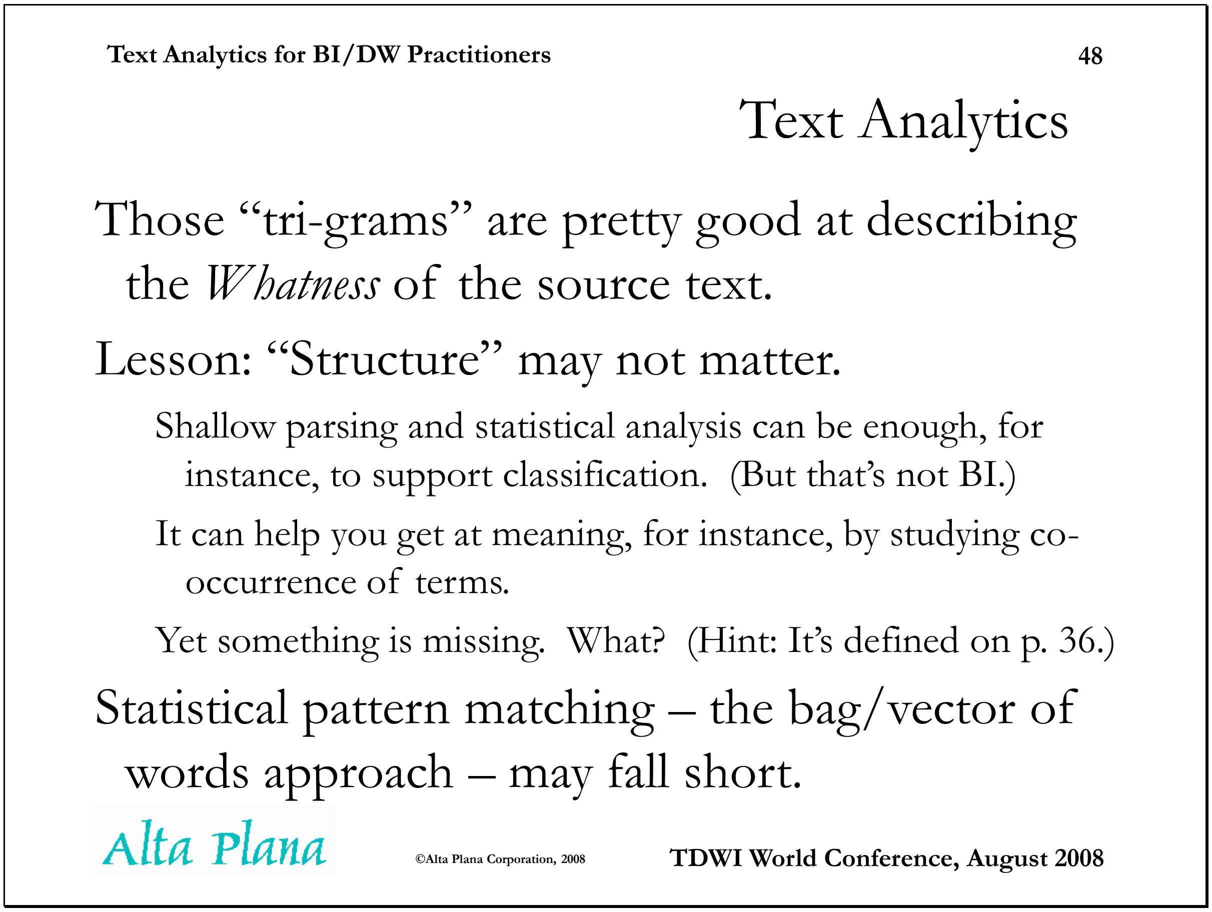
Total 2 word phrases : 102 - Total Repeats : 246

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 %	10.50

Scrolling down the results page, we look at “bi-grams” and “tri-grams,” 2- and 3-consecutive-word terms.



Text Analytics

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

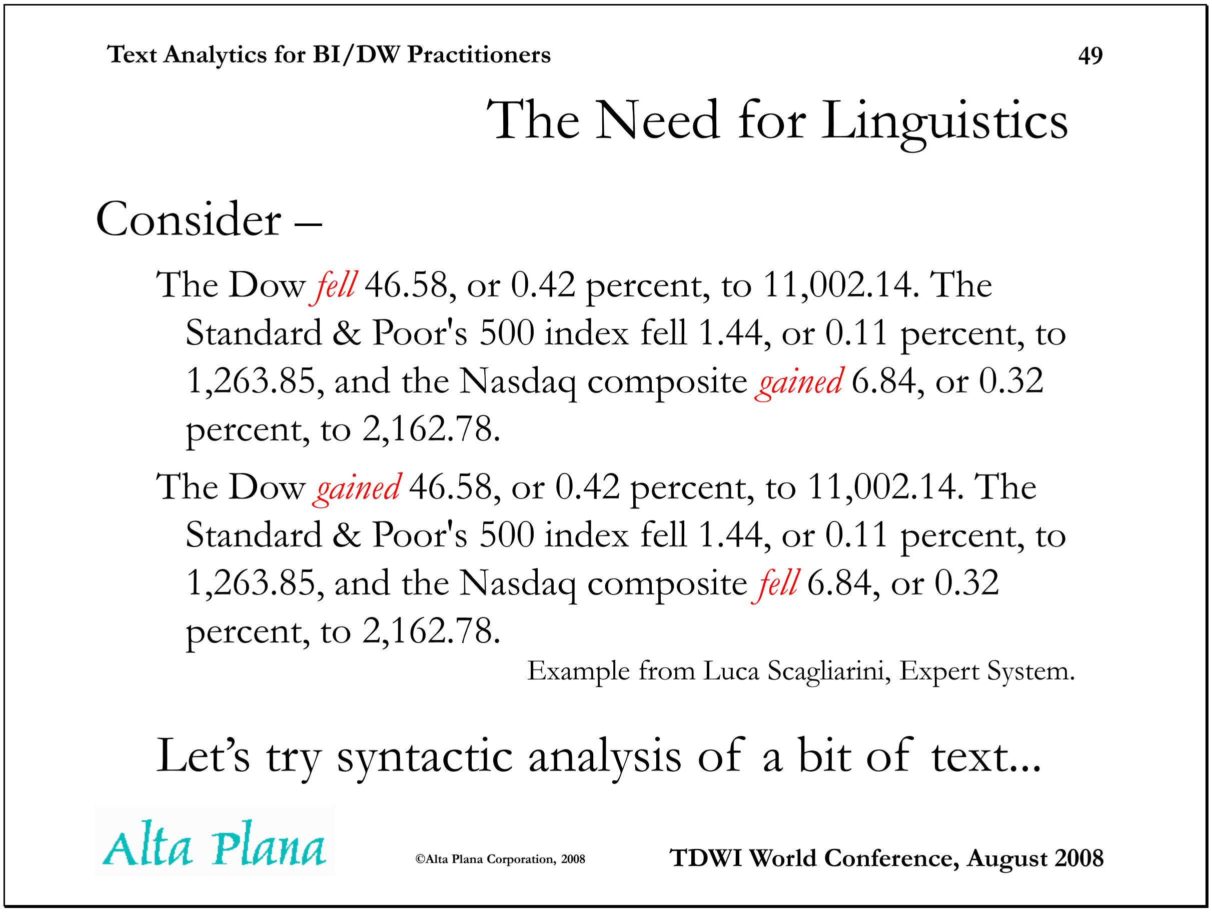
Lesson: “Structure” may not matter.

- 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.

By “whatness,” I mean the primary subjects/topics/themes of the submitted document.

The computer can come up with this information *without* really grasping or being able to process the “meaning” of a document or its content in any significant way.



The Need for Linguistics

Consider –

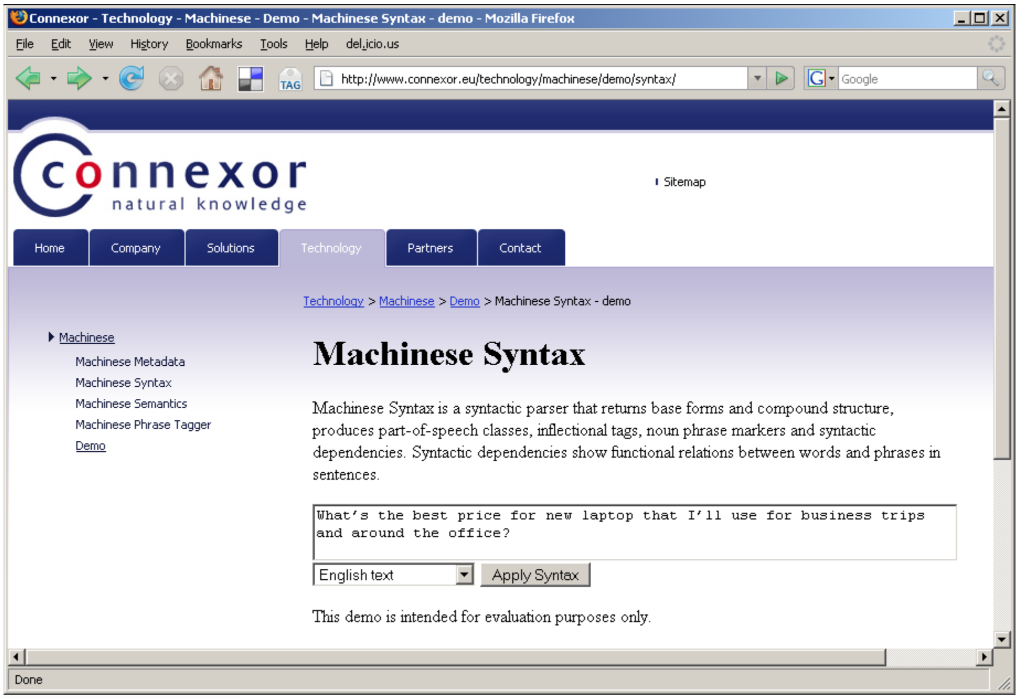
The Dow *fell* 46.58, or 0.42 percent, to 11,002.14. The Standard & Poor's 500 index fell 1.44, or 0.11 percent, to 1,263.85, and the Nasdaq composite *gained* 6.84, or 0.32 percent, to 2,162.78.

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…

The two paragraphs shown here are identical so far as their word contents are concerns. The position of two words is switched from one paragraph to the other, changing the information content. If we consider on the word content, we'll miss that meaning change. We need to see *facts* -- <subject> <predicate> <object> triplets – by seeing parts of speech and grammatical syntax.



Technology > Machinese > Demo > Machinese Syntax - demo

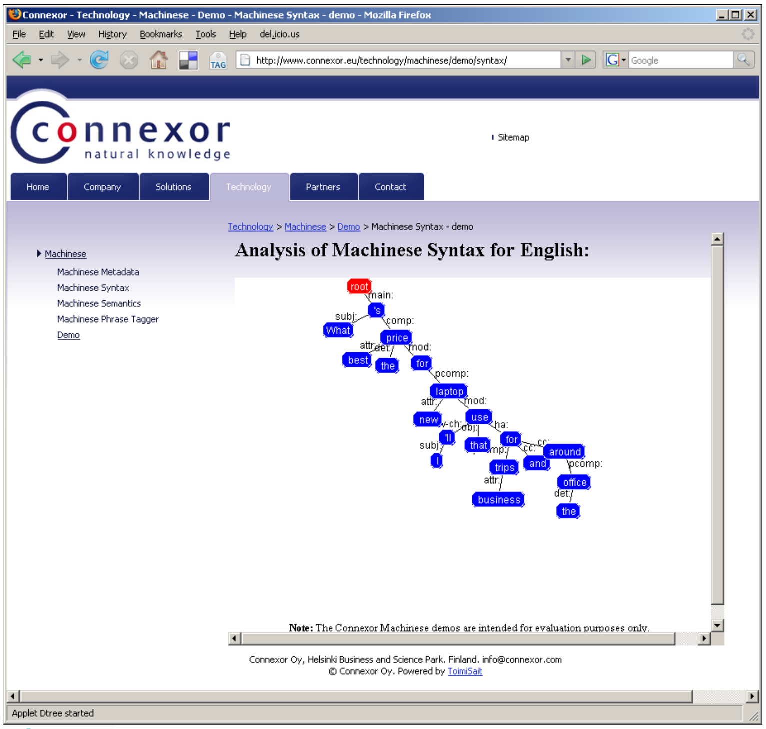
Machinese Syntax

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?
```

This demo is intended for evaluation purposes only.

This is another site on the public Web, one that demonstrates parsing and syntactic analysis of free text. The purpose here is to demonstrate what computer software is capable of doing, what many of us were taught to do in elementary school…



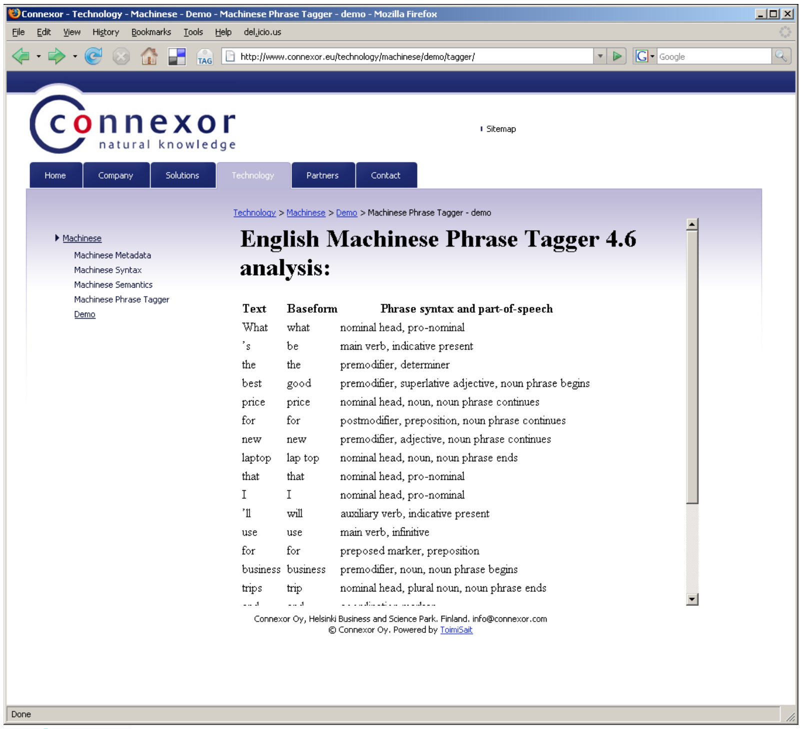
Technology > Machinese > Demo > Machinese Syntax - demo

Analysis of Machinese Syntax for English:

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

The software creates a hierarchical, tree-structured sentence diagram.


Connexor - Technology - Machinese - Demo - Machinese Phrase Tagger - demo - Mozilla Firefox
File Edit View History Bookmarks Tools Help del.icio.us
http://www.connexor.eu/technology/machinese/demo/tagger/
Google
connexor
natural knowledge
Sitemap
Home
Company
Solutions
Technology
Partners
Contact
Machinese
Machinese Metadata
Machinese Syntax
Machinese Semantics
Machinese Phrase Tagger
Demo
Done

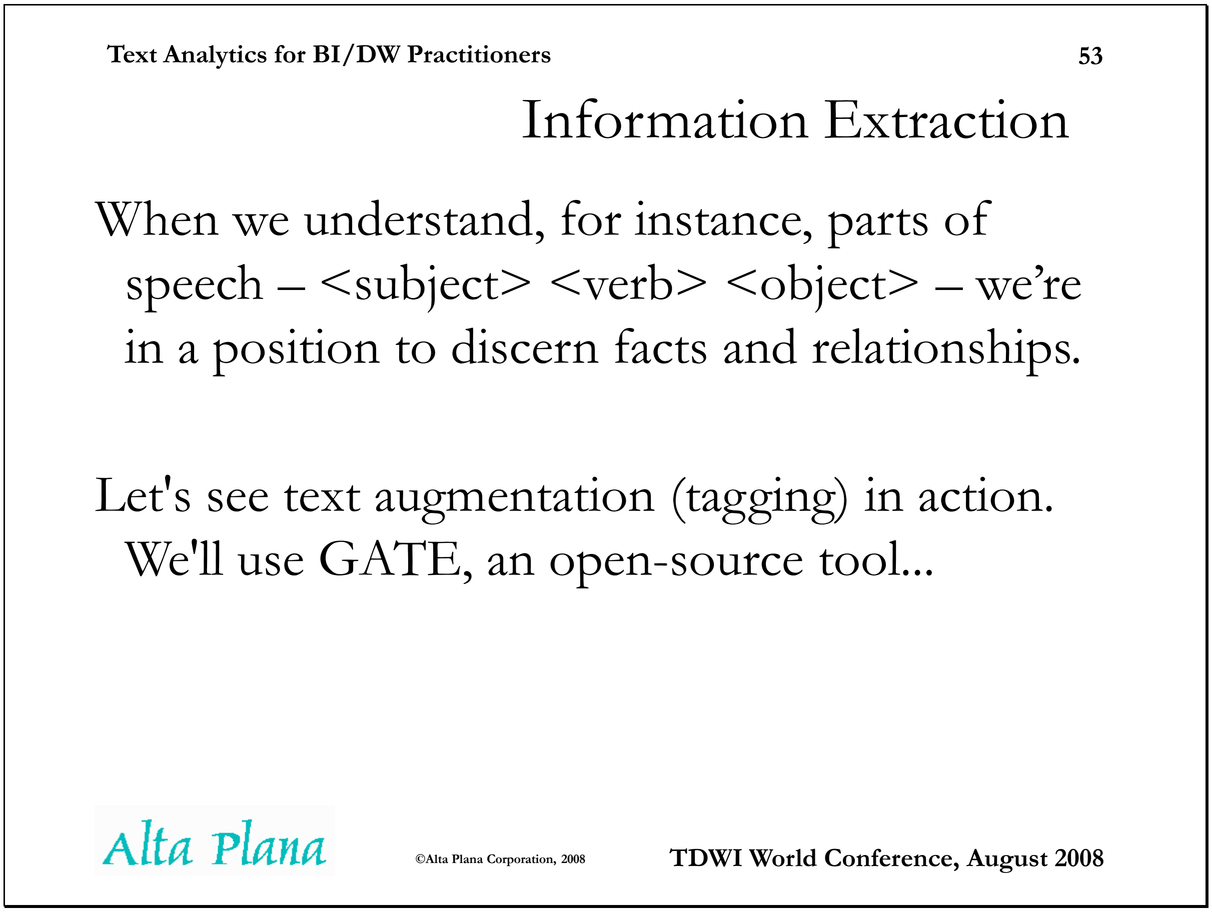
Technology > Machinese > Demo > Machinese Phrase Tagger - demo

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

Connexor Oy, Helsinki Business and Science Park. Finland. info@connexor.com
© Connexor Oy. Powered by ToimiSait

In an alternative view, we see that the software identifies structural elements of the text – e.g., clauses – as well as parts of speech.

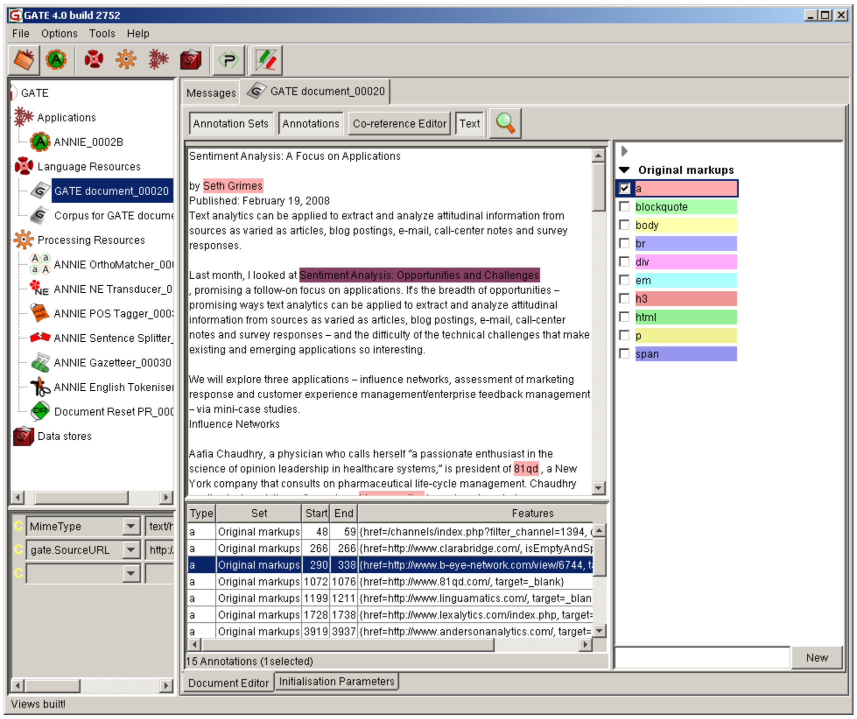


Information Extraction

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...

The <subject> <verb> <object> triplet is something we see in RDF and in representations of knowledge in ontologies: X is/does Y. It, with the elaboration provides by attributes and modifiers, is the basic pattern for a fact or a relationship, so that if we can discern that pattern in source materials, we can detect those facts and relationships.



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

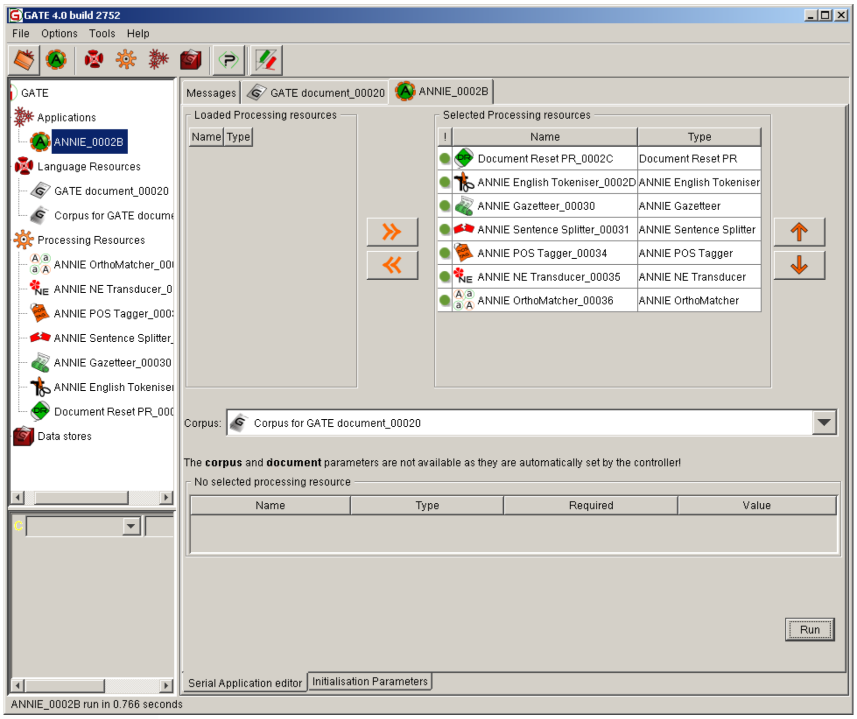
Aafia Chaudhry, a physician who calls herself "a passionate enthusiast in the science of opinion leadership in healthcare systems," is president of 81qd , a New York company that consults on pharmaceutical life-cycle management. Chaudhry

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, ta
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)

Here's the Gate interface operating out-of-the-box, with defaults. Simply choose an auto-loaded processing pipeline as a Processing Resource, set up a corpus (document set) with a single document, and apply the pipeline to the source. Actually, we haven't done that yet in this screen. Even without running the pipeline, Gate recognizes HTML (and other) pre-existing mark-up tags in the source file.

These are many, many additional processing options.



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

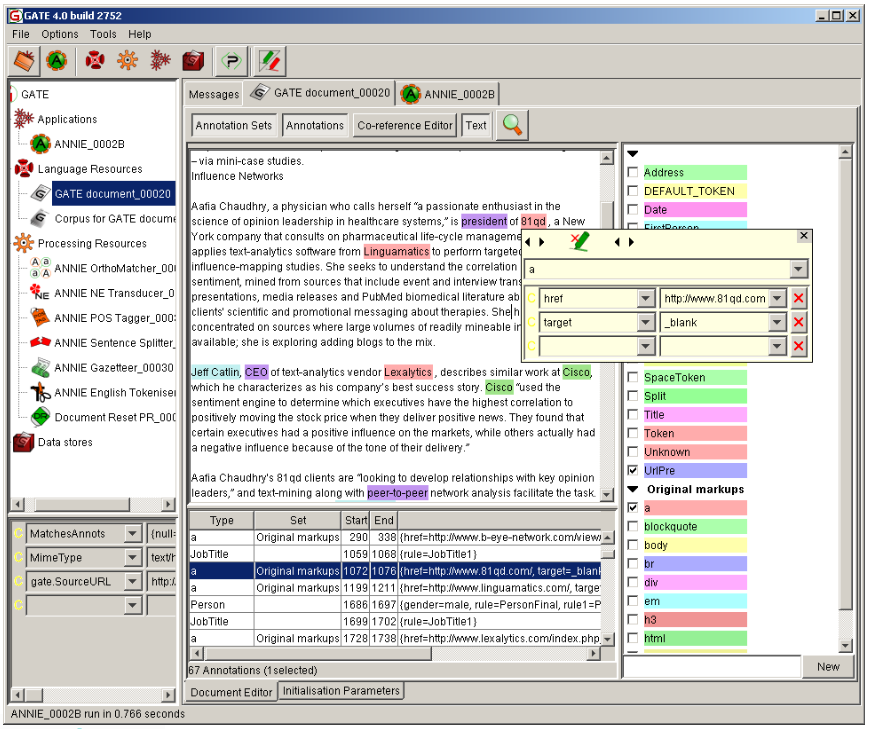
The **corpus** and **document** parameters are not available as they are automatically set by the controller!

No selected processing resource

Name	Type	Required	Value

ANNIE_0002B run in 0.766 seconds

This screen displays an NLP pipeline available in Gate out of the box.



MatchesAnnots {null=
MimeType text/h
gate.SourceURL http:/

– 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 targeted
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 in
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.

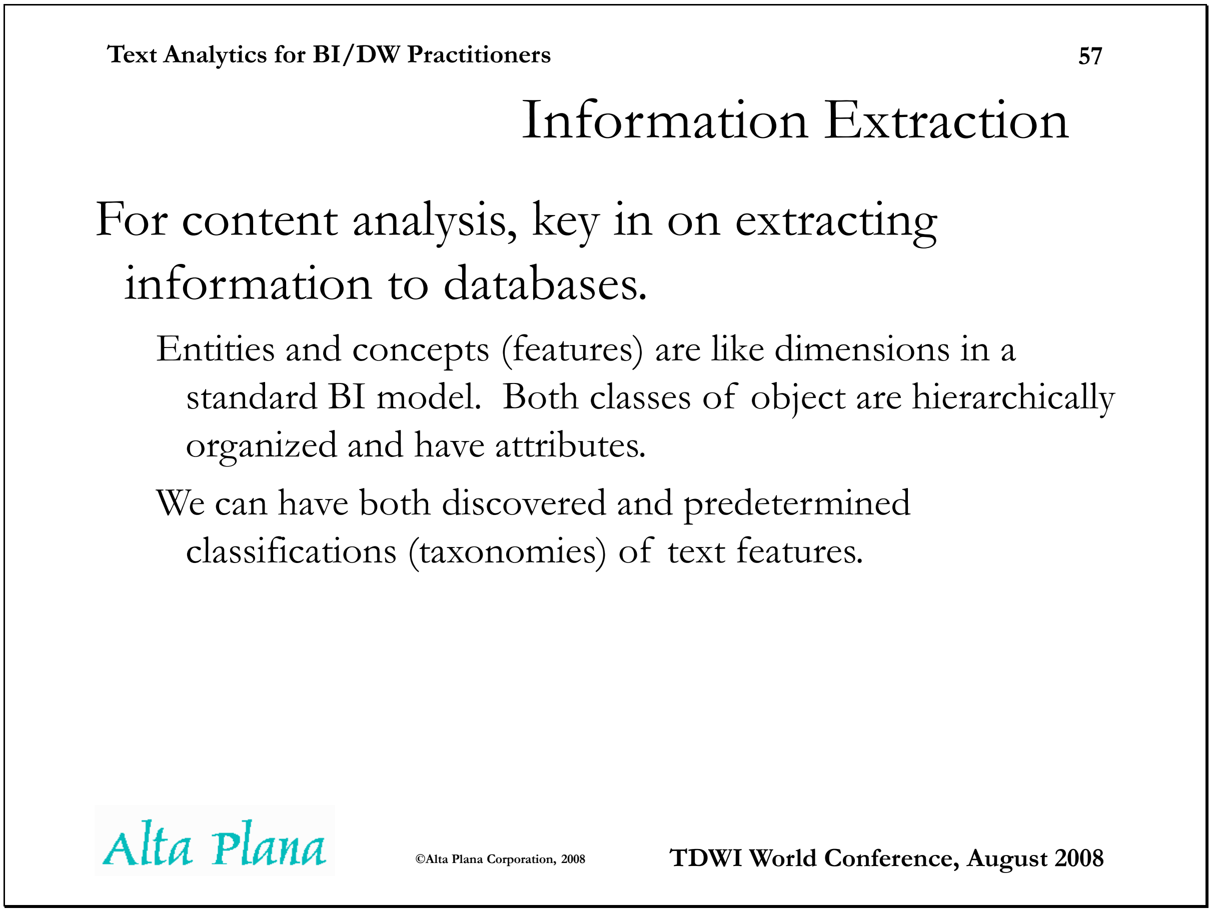
a
href http://www.81qd.com
target _blank

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)

ANNIE_0002B run in 0.766 seconds

Gate results after the NLP annotation (text augmentation) pipeline is executed on the source document.

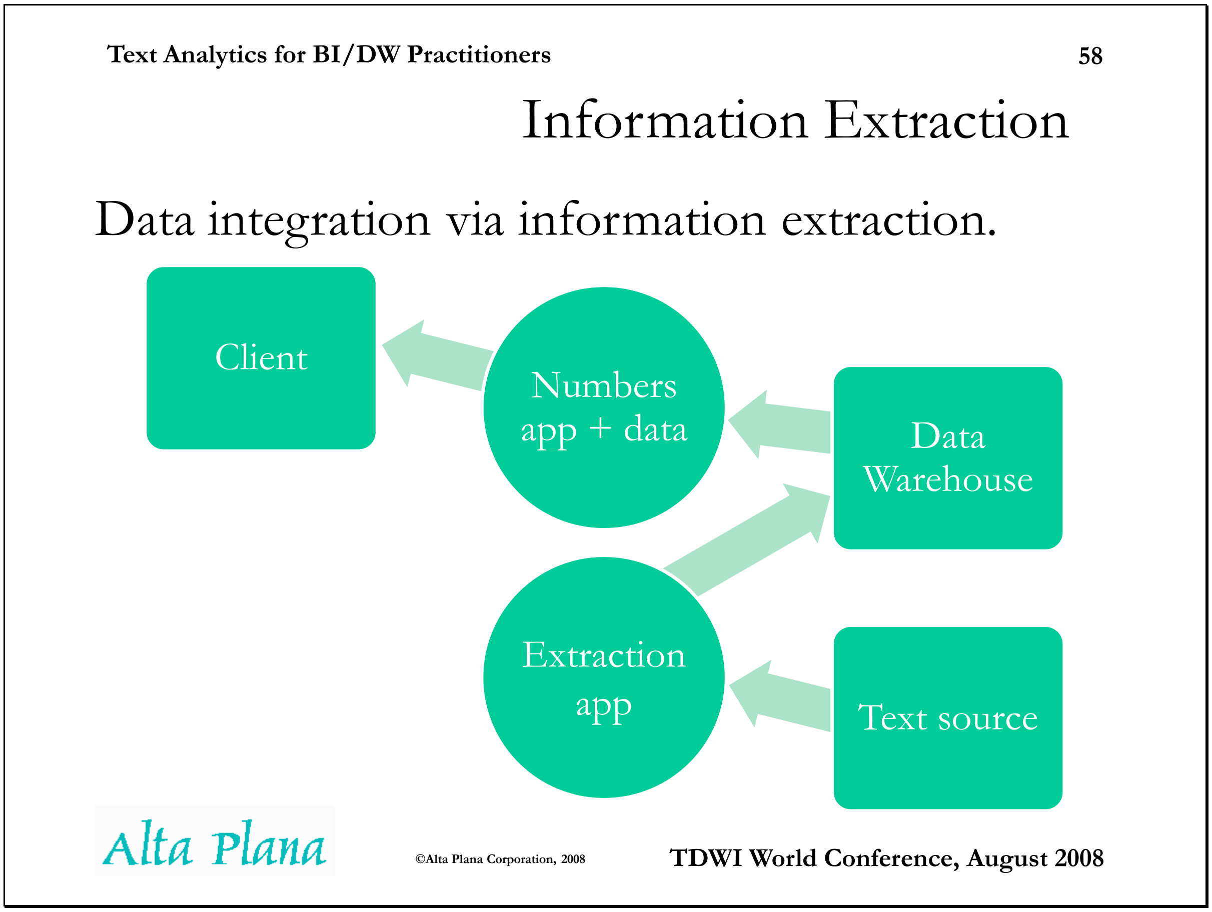


Information Extraction

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.

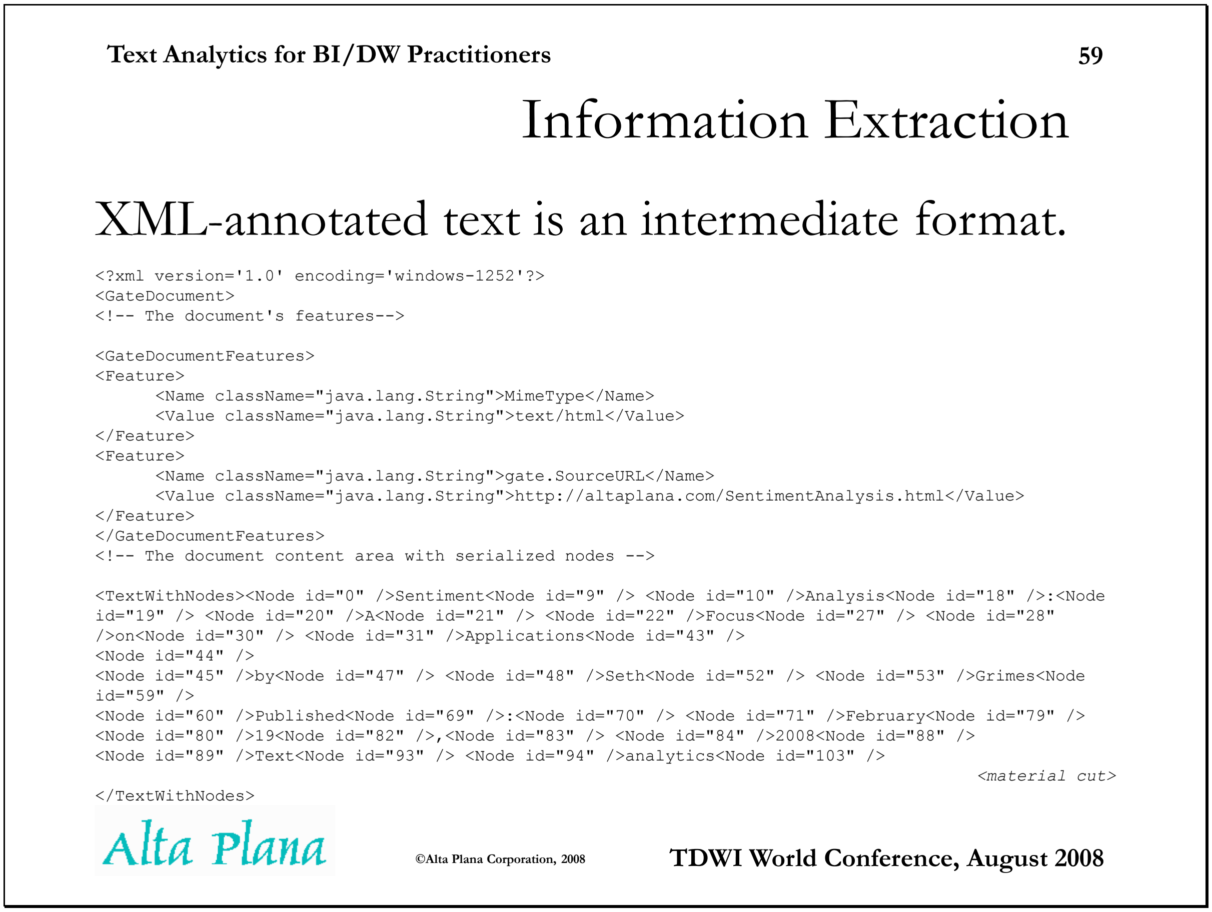
The standard BI model is a dimensional model where fact tables that contain measure variables are related to dimension tables via a foreign-key relationship.



Information Extraction

Data integration via information extraction.

One data integration model involved IE from textual sources to a data warehouse for integrated/unified analysis of text-sourced information and data that originated in transactional systems.



Information Extraction

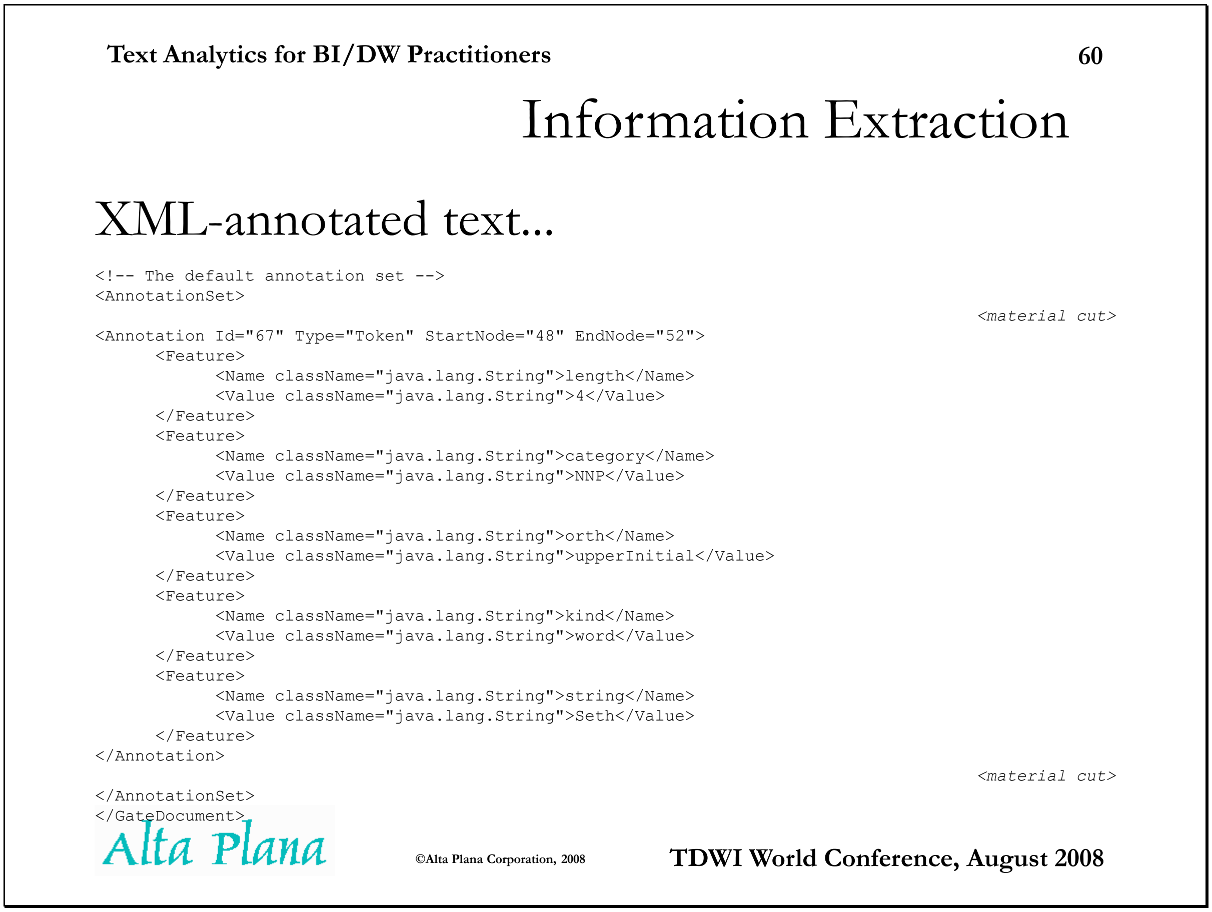
XML-annotated text is an intermediate format.

```
<?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>
```

Textual information marked up with XML can be considered an intermediate format, that is, a step between the raw text and database tables. Certain XML documents are essentially equivalent to a database (or a DB record) however, which means that in certain scenarios, the XML documents are a final form and not just an intermediate form.

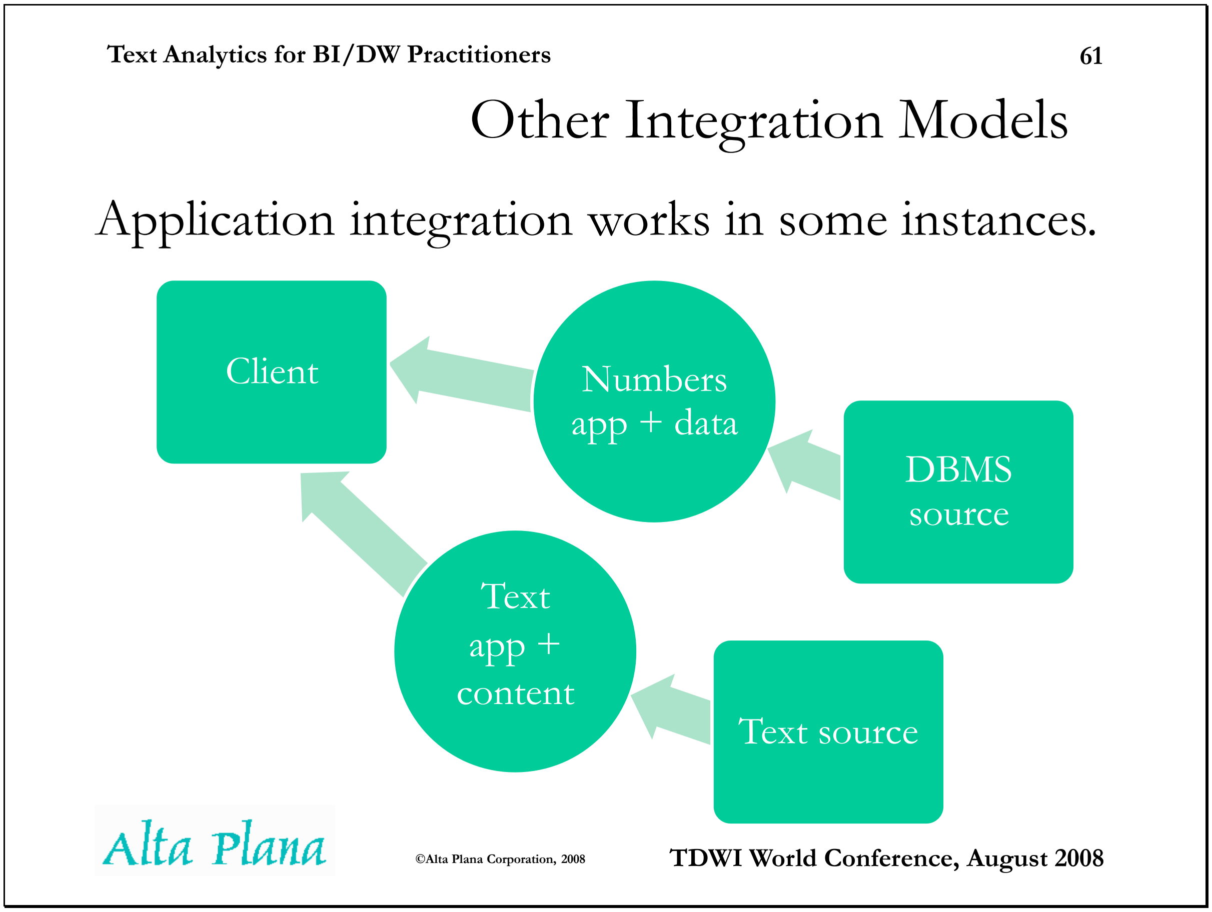


Information Extraction

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>
```

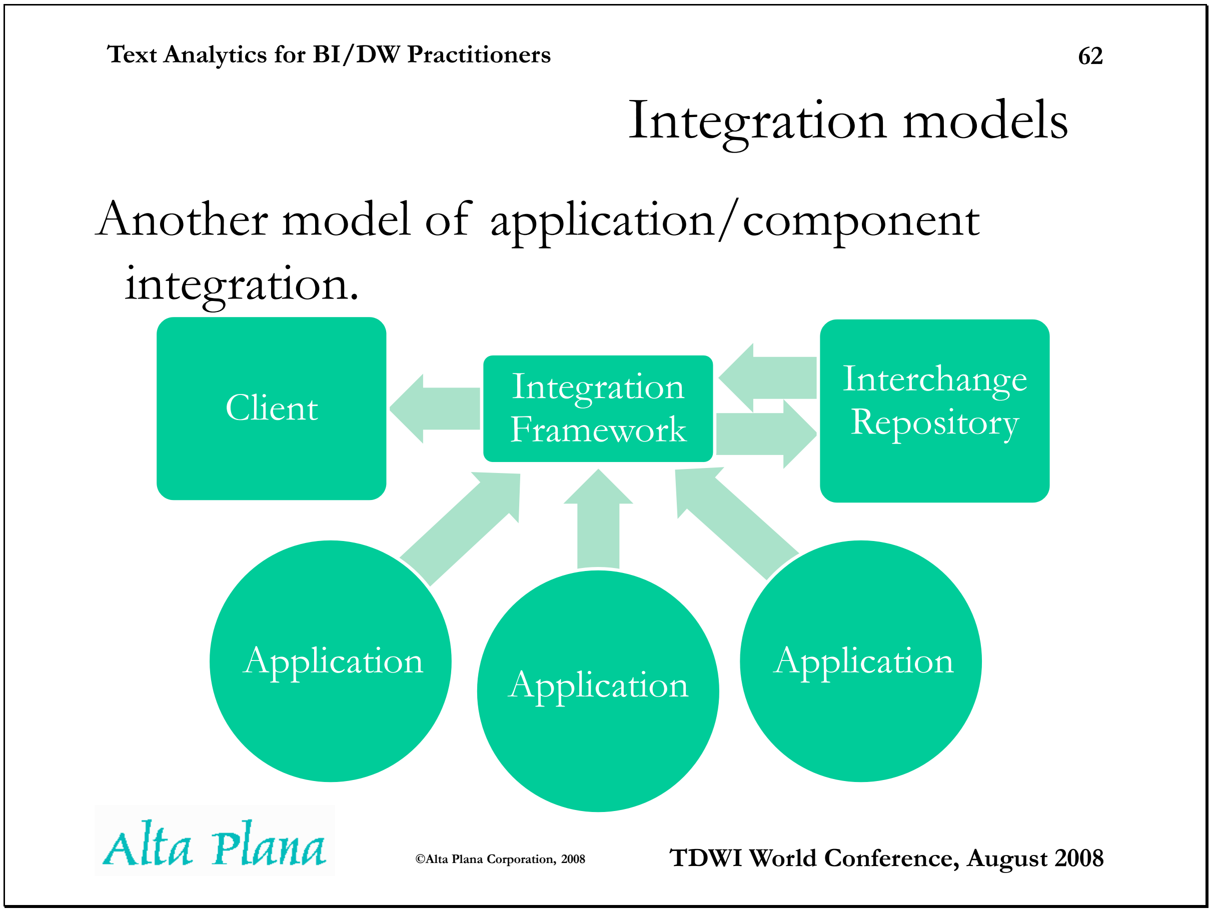
There are a variety of text-annotation XML schemas. This example uses Gate’s schema but there are others.



Other Integration Models

Application integration works in some instances.

Unintegrated applications, where all that’s brought together is perhaps a display, are not of great interest. Interfaces – and they may include dashboards and portals and other trendy stuff – that gather information from different sources without a linkage mechanism provide little added value beyond the value of the individual elements.



Integration models

Another model of application/component integration.

Client

Integration Framework

Interchange Repository

Application

Application

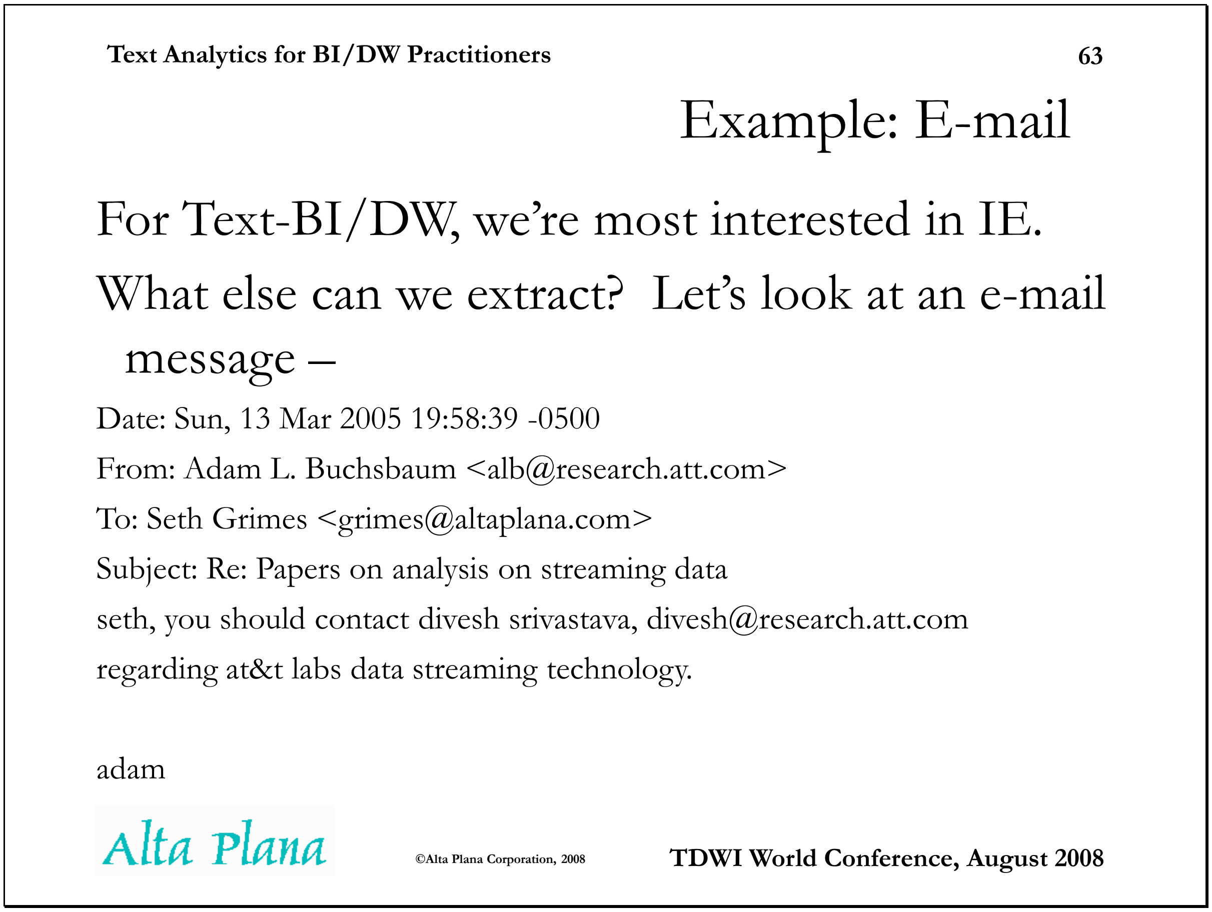
Application

Alta Plana

Another form of integration worth note is framework based, where disparate applications use standardized access methods and interchange structures to interoperate in a mid-way coupled form. Frameworks of this sort typically include components that provide commonly used services.

The most notable frameworks are bigger than the analytics world: Java Enterprise Edition (Java EE), Microsoft's .Net, the Eclipse universal tool environment and rich client platform. There are literally dozens if not hundreds of frameworks with multiple frameworks associated with a given programming or software tool.

Service platforms are based on frameworks. Examples include Microsoft's Sharepoint, Apache Tomcat, IBM WebSphere, and Red Hat JBoss. There are many others.



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>

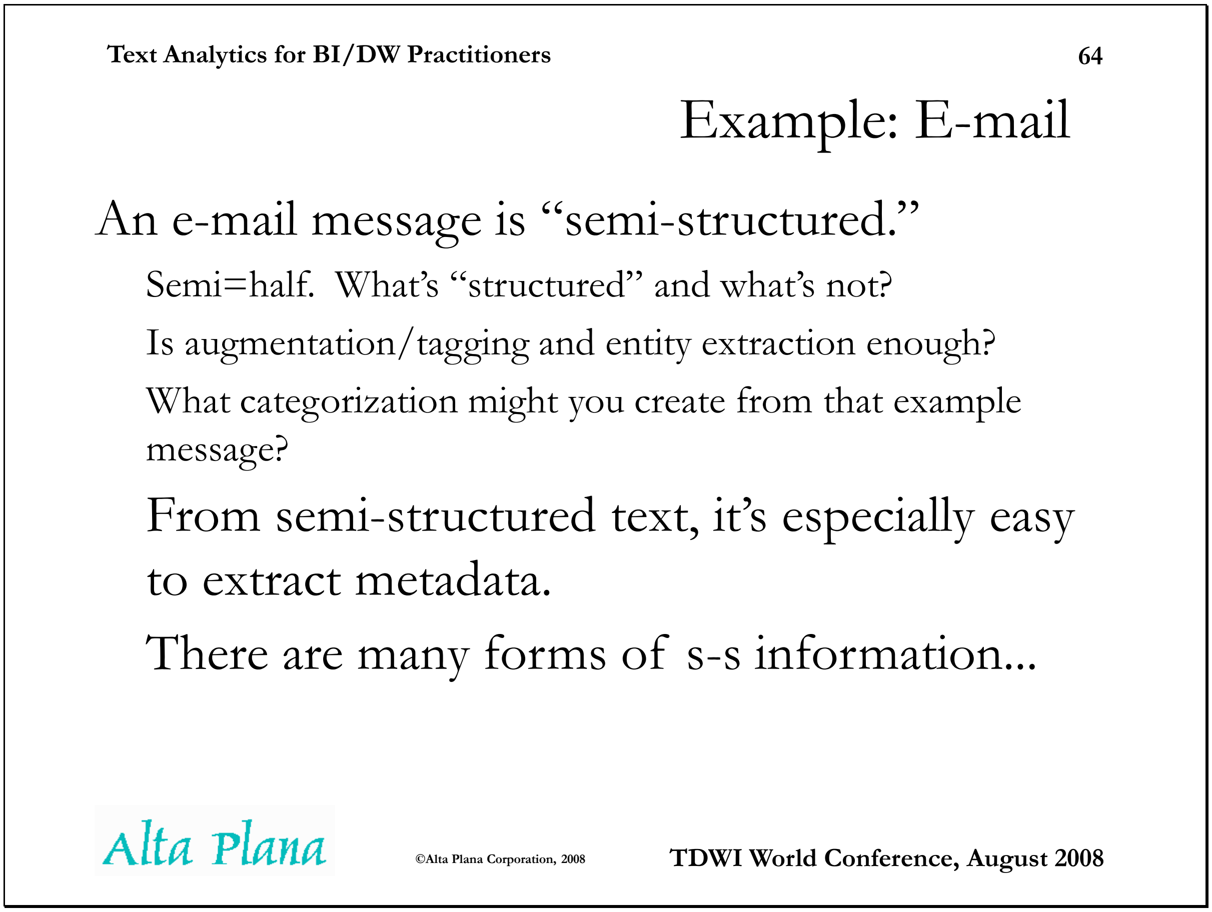
To: Seth Grimes <grimes@altaplana.com>

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

An e-mail message includes “header” fields that are used to describe and e-mail message including its attachments and the path by which it was transmitted from sender to receiver. The header fields provide, in effect, metadata for message handling. Most header fields are not, by default, displayed by e-mail reader programs.



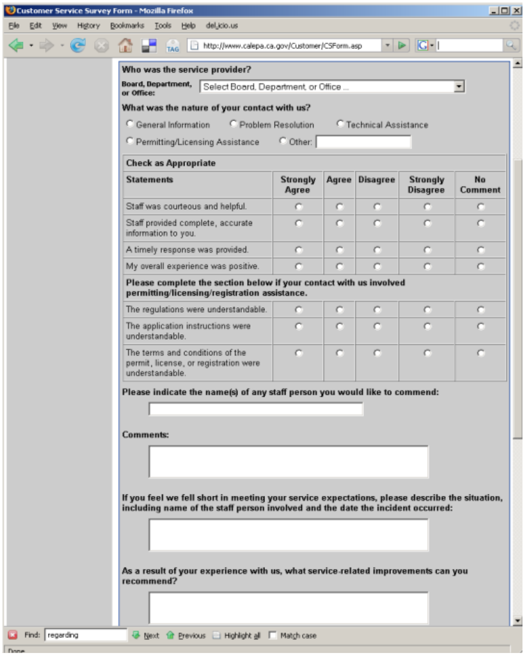
Example: E-mail

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...

Semi-structured can actually mean two similar, but different things: 1) Part of the document is “structured” and part not, and 2) The whole document is partly “structured.” Let’s leave it at that.

Example: Survey


Customer Service Survey Form - Mozilla Firefox
File Edit View History Bookmarks Tools Help del.icio.us
TAG
http://www.calepa.ca.gov/Customer/CSForm.asp
Find: regarding
Next Previous Highlight all Match case
Done

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.	○	○	○	○	○

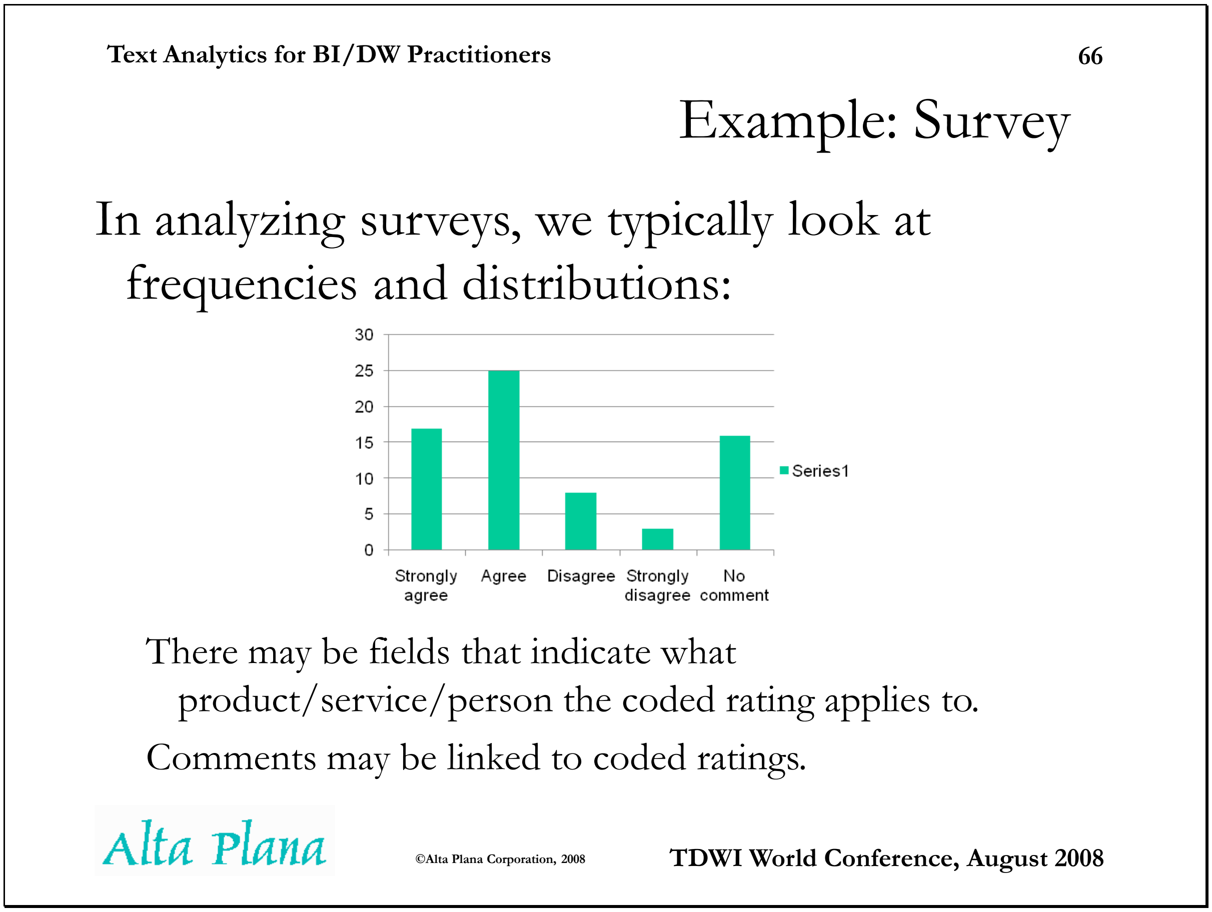
Please indicate the name(s) of any staff person you would like to commend:

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?

Survey responses are semi-structured in the sense that certain questions have responses that are restricted to some pre-determined set while other questions allow freely chosen responses.



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.

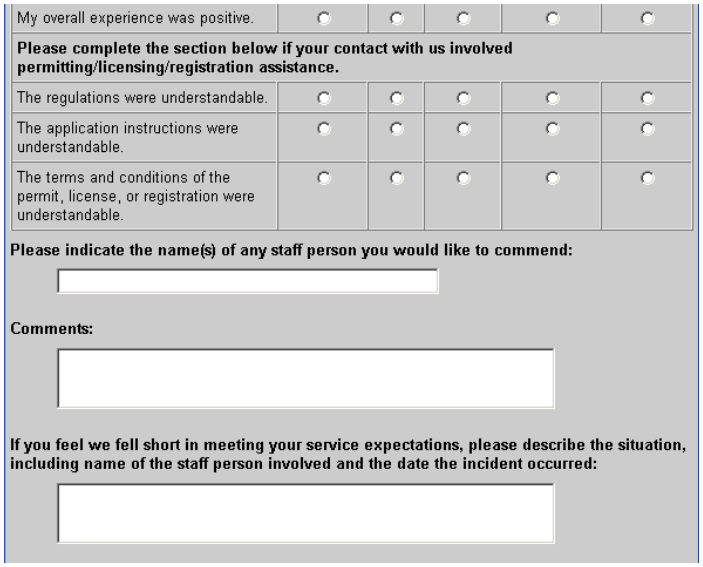
Frequencies and distributions apply, of course, to categorical and numerical responses.

Once you categorize free-text responses, you can compute frequency distributions for those categories.

What happens when you can apply multiple categorization to a given free-text response question? Is this similar to what happens when you have a multi-response, restricted-answer-set question?

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:

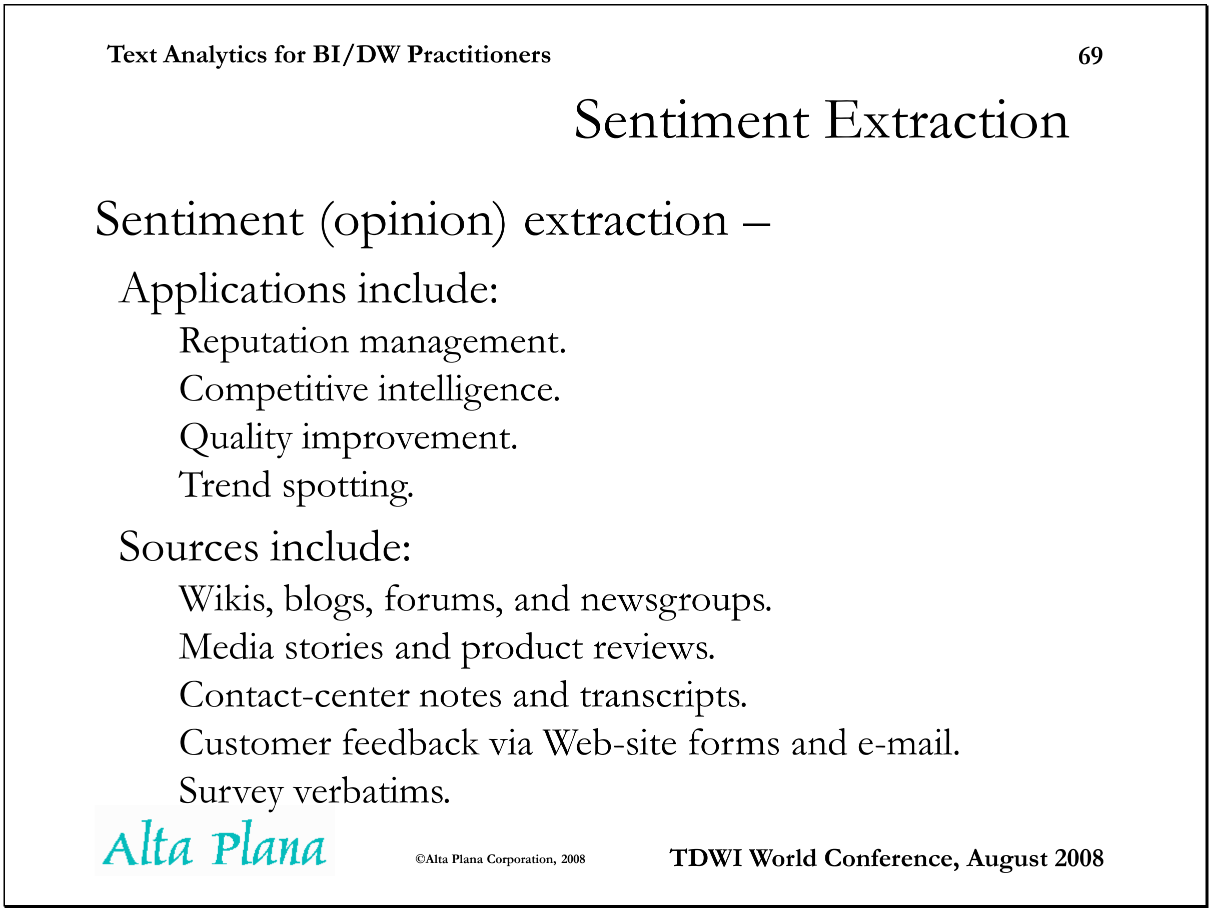
If you feel we fell short in meeting your service expectations, please describe the situation, including name of the staff person involved and the date the incident occurred:

Just a close-up. Do the coded-response questions here give you a head start in analyzing the free-response questions?

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.

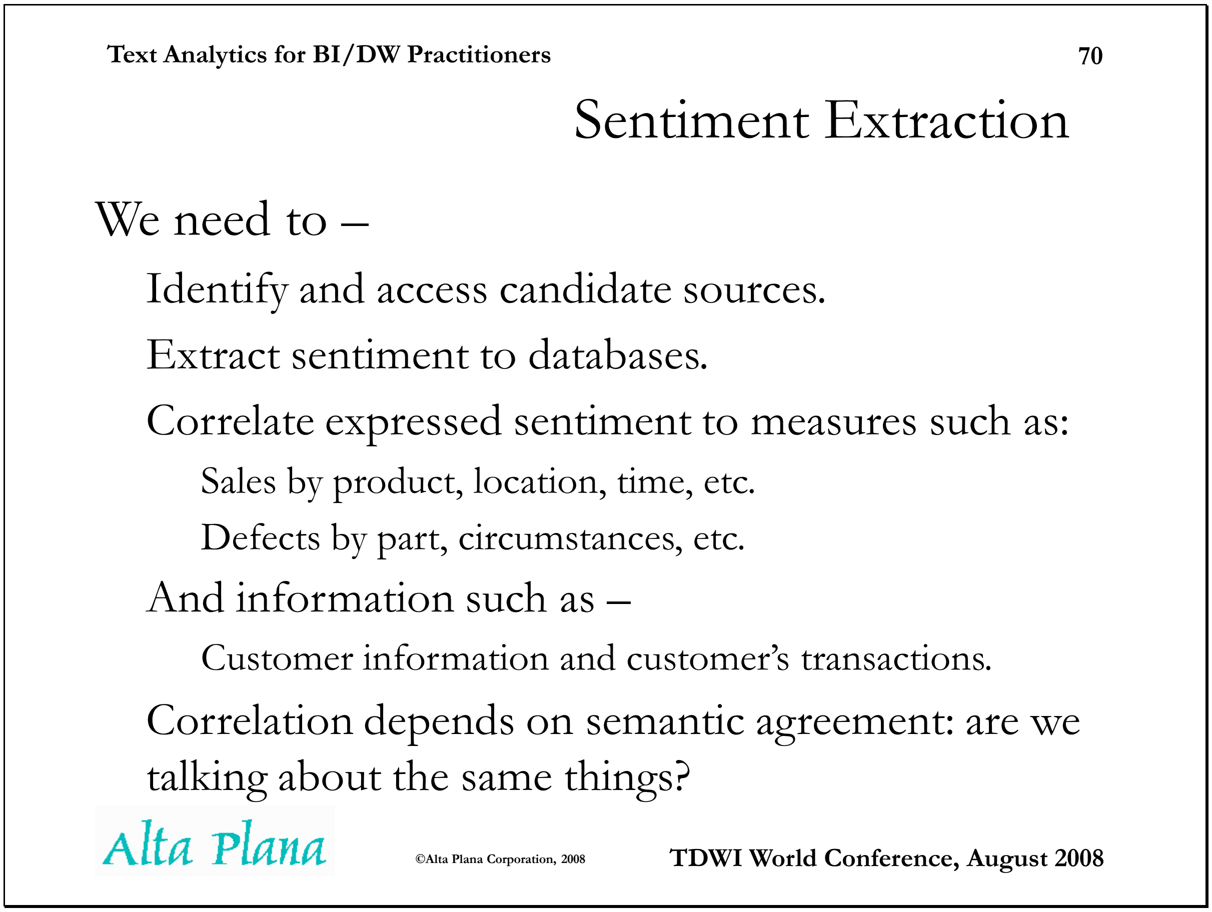
Sources include:

- 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 is important for Voice of the Customer and similar applications.

Opinions are essentially facts (assertions) that involve sentiment.

Sentiment and opinion together constitute attitudinal information.



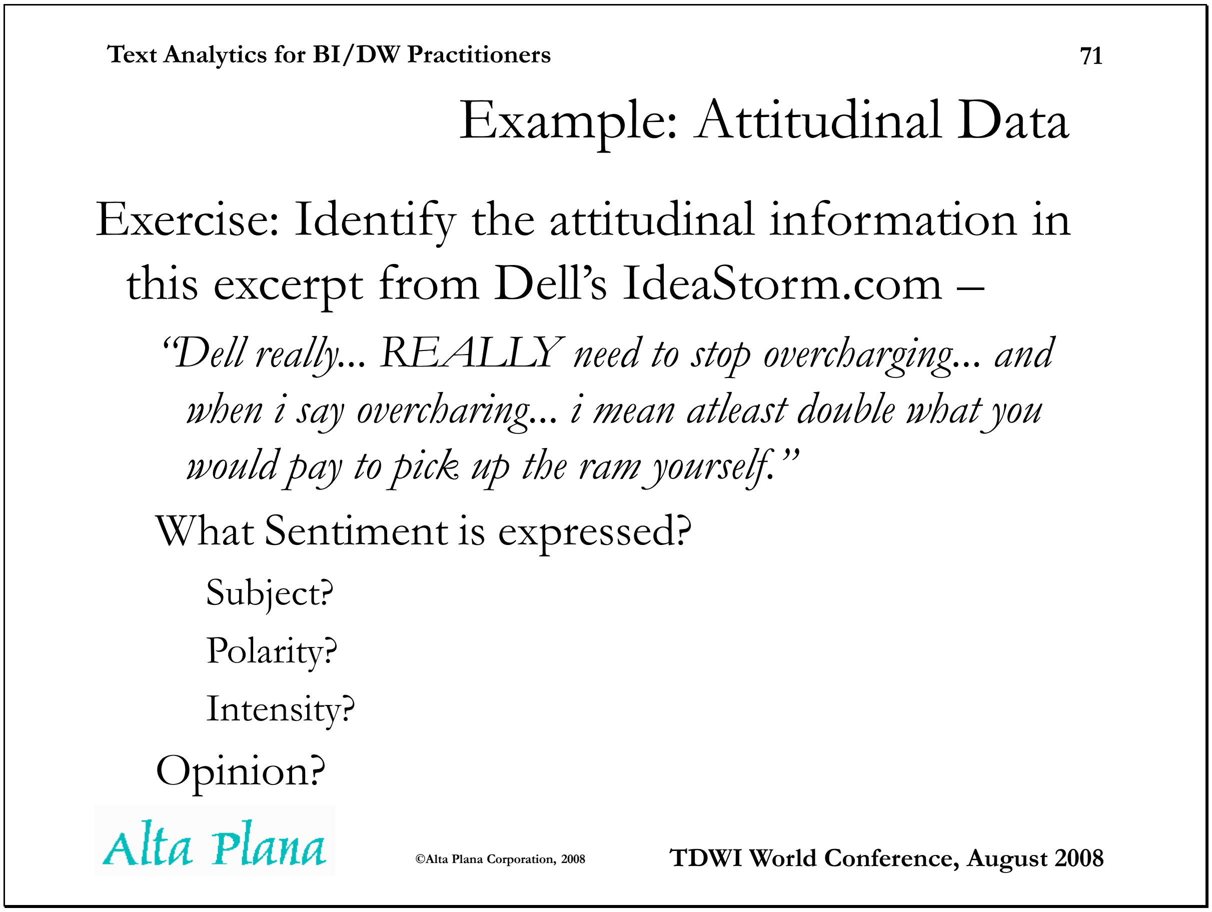
Sentiment Extraction

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?

Alta Plana

The sentiment analysis pipeline really doesn't differ much from the general information-extraction pipeline as applied for analysis of factual information.



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?

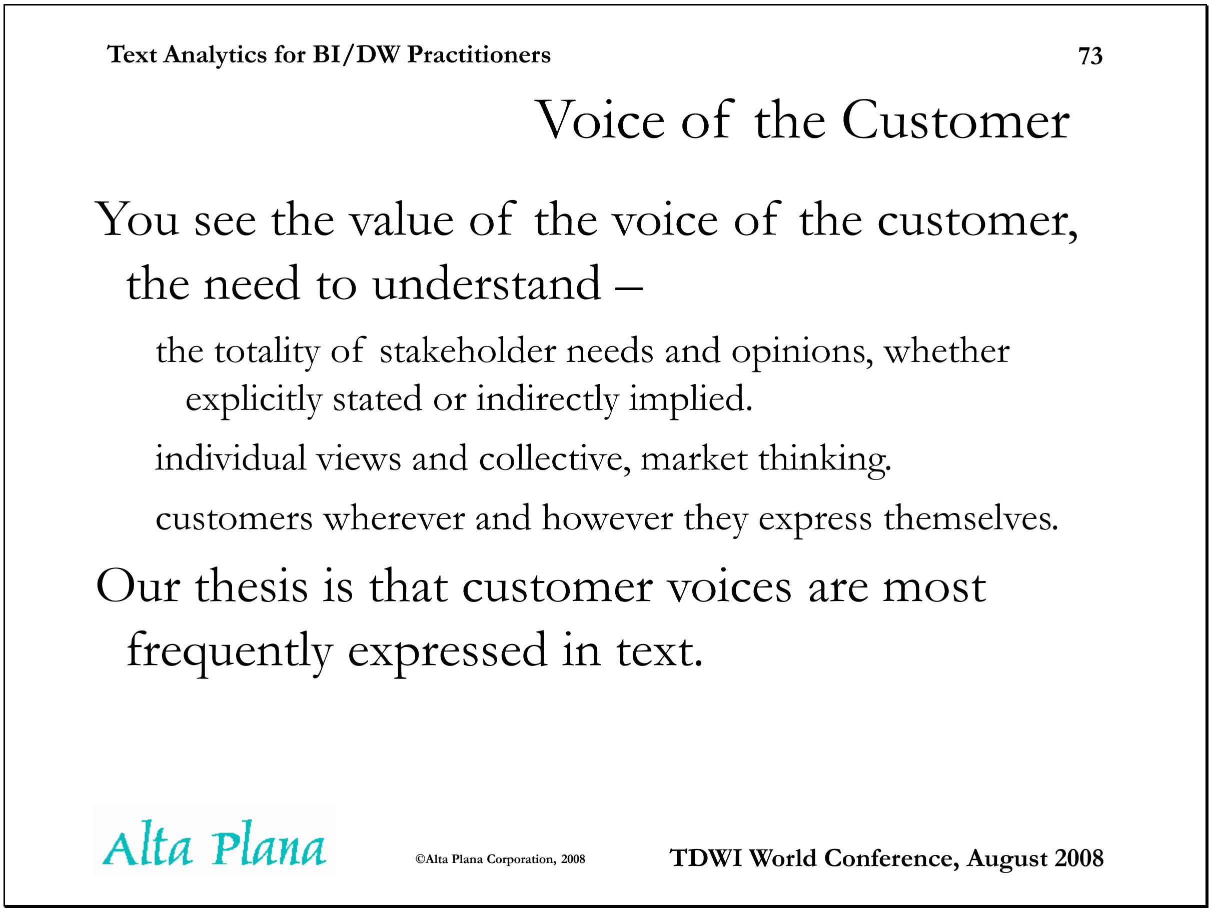
Opinion?

There are many issues to deal with when we encounter authentic stakeholder voices: irregular capitalization and punctuation, incorrect spelling, fractured grammar, unconventional usage, etc., and that's not even getting into idiom, sarcasm, irony, and the like. This quotation illustrates many of the challenges.

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.



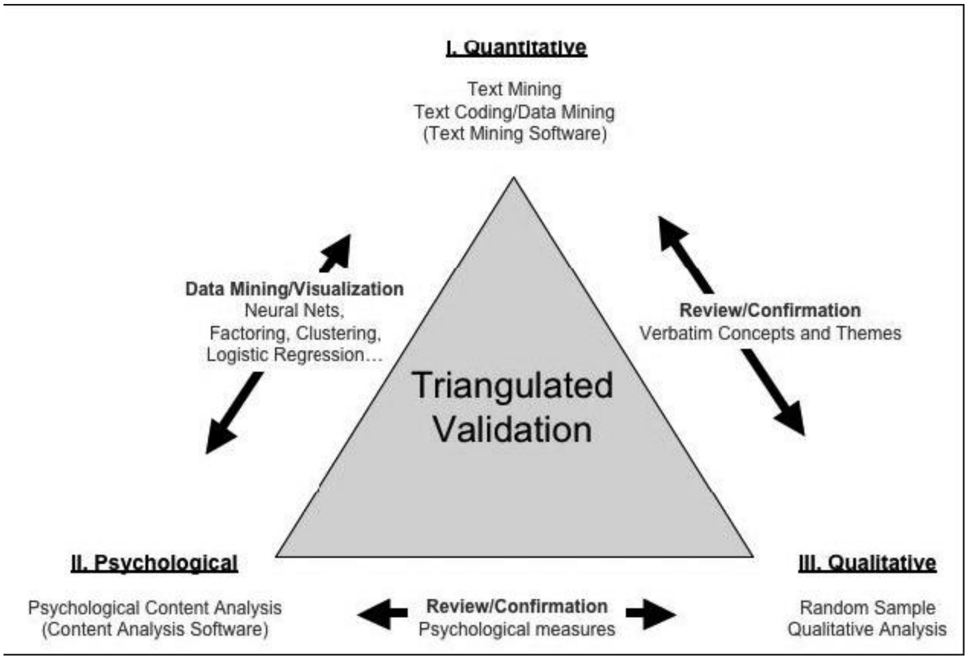
Voice of the Customer

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



I. Quantitative

Text Mining
Text Coding/Data Mining
(Text Mining Software)

Data Mining/Visualization
Neural Nets,
Factoring, Clustering,
Logistic Regression…

Review/Confirmation
Verbatim Concepts and Themes

Triangulated Validation

II. Psychological

Psychological Content Analysis
(Content Analysis Software)

Review/Confirmation
Psychological measures

III. Qualitative

Random Sample
Qualitative Analysis

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

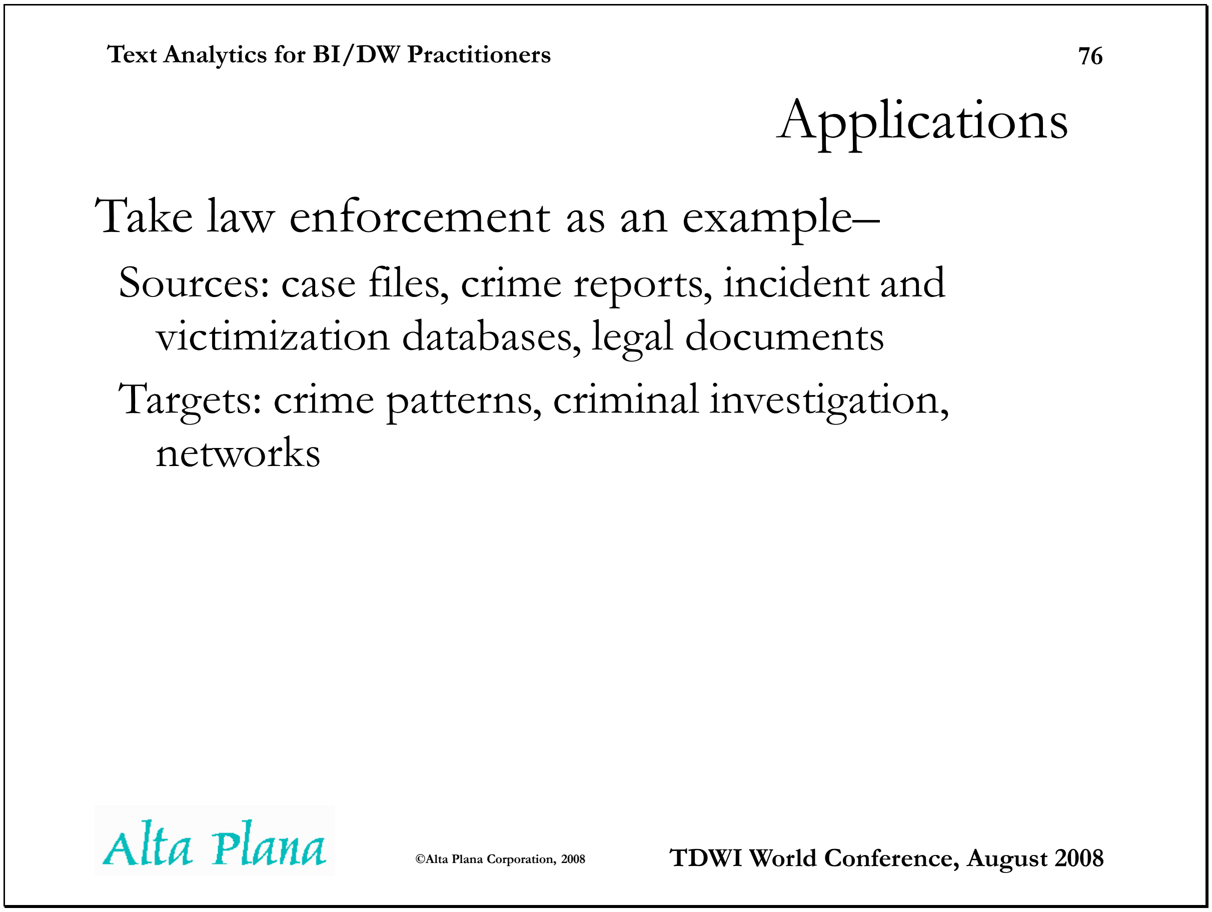
Voice of the Customer

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

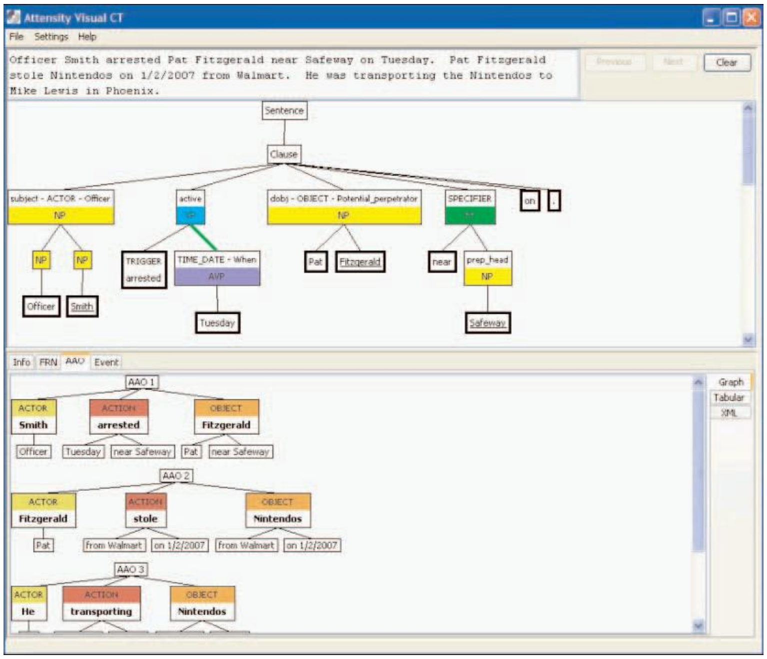
Take law enforcement as an example–

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

Law enforcement generates and consumes a lot of textual information in free-text and semi-structured forms as well as data in traditional, fielded databases. Materials are mined for names, relationships, events, location, etc., with information extracted to database systems. It's then the analyst's turn to search for patterns and predictive rules.

Applications

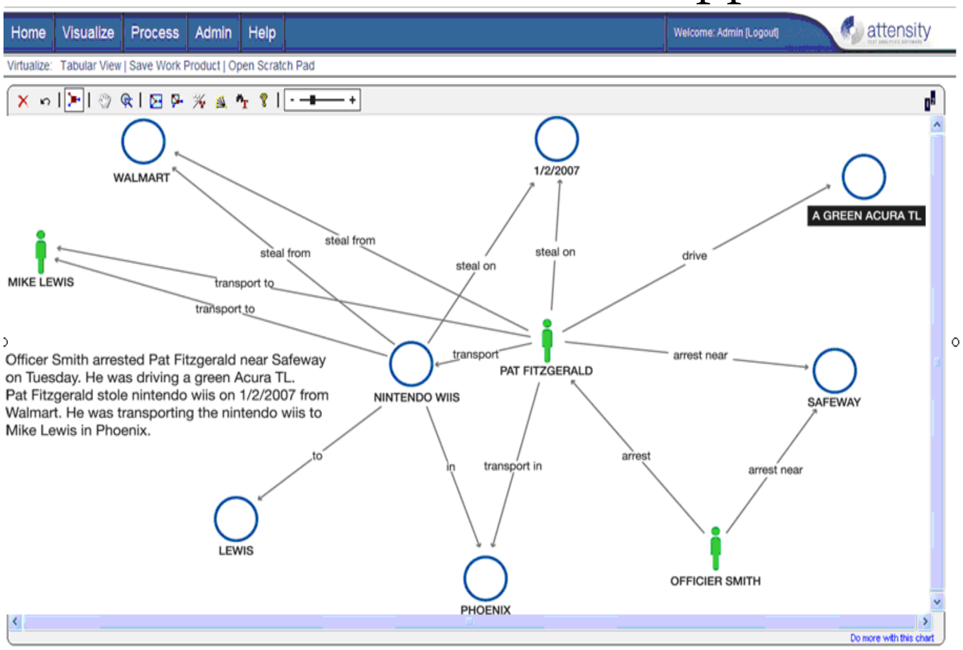
An Attensity law-enforcement example – NLP to identify roles and relationships.



```
Officer Smith arrested Pat Fitzgerald near Safeway on Tuesday.  Pat Fitzgerald
stole Nintendos on 1/2/2007 from Walmart.  He was transporting the Nintendos to
Mike Lewis in Phoenix.
```

Alta Plana

Applications



Officer Smith arrested Pat Fitzgerald near Safeway on Tuesday. He was driving a green Acura TL. Pat Fitzgerald stole nintendo wiis on 1/2/2007 from Walmart. He was transporting the nintendo wiis to Mike Lewis in Phoenix.

From the extracted roles and relationships information, we generate a link-analysis graph.

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

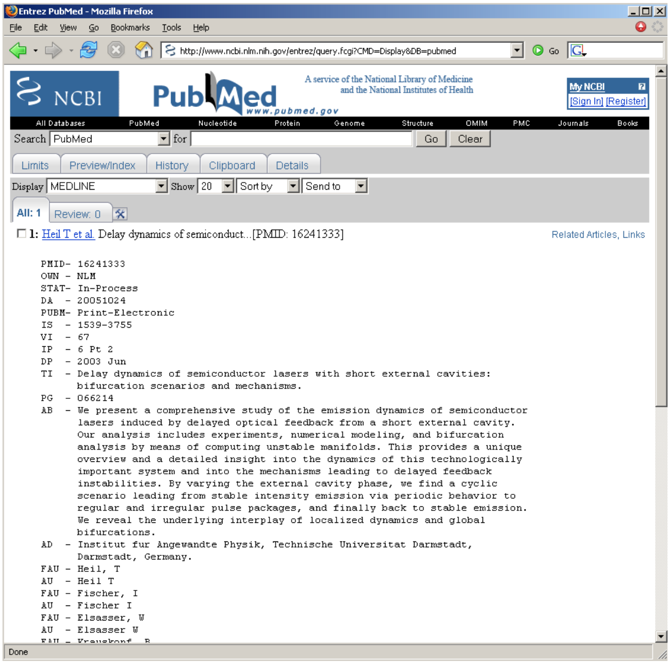
- Sources: clinical research databases, patient records, insurance filings, regulations
- Targets: enhance diagnosis and treatment, promote quality of service, increase utilization, control costs

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.*



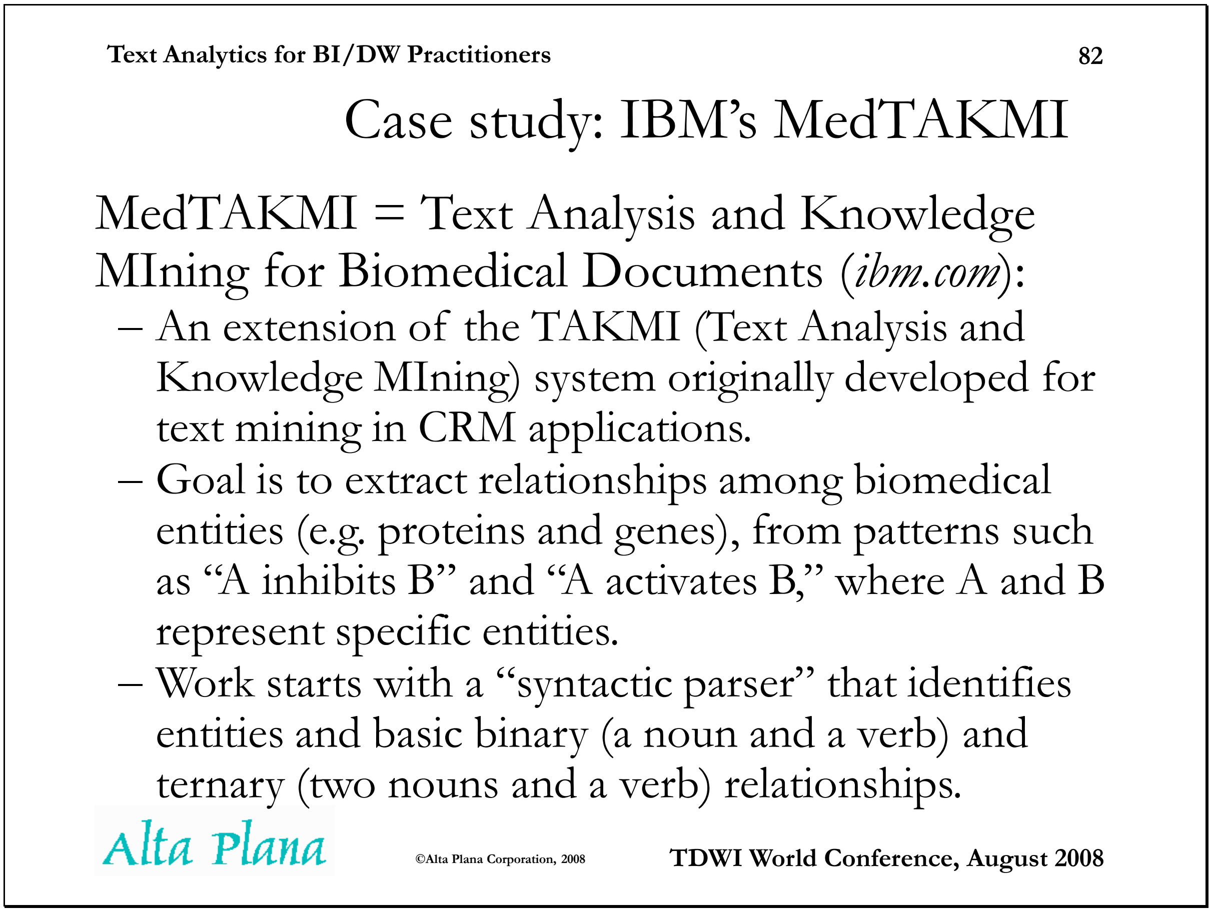
□ 1: Heil T et al. Delay dynamics of semiconduct...[PMID: 16241333]

```
PMID- 16241333
OWN - NLM
STAT- In-Process
DA  - 20051024
PUBM- Print-Electronic
IS  - 1539-3755
VI  - 67
IP  - 6 Pt 2
DP  - 2003 Jun
TI  - Delay dynamics of semiconductor lasers with short external cavities:
      bifurcation scenarios and mechanisms.
PG  - 066214
AB  - We present a comprehensive study of the emission dynamics of semiconductor
      lasers induced by delayed optical feedback from a short external cavity.
      Our analysis includes experiments, numerical modeling, and bifurcation
      analysis by means of computing unstable manifolds. This provides a unique
      overview and a detailed insight into the dynamics of this technologically
      important system and into the mechanisms leading to delayed feedback
      instabilities. By varying the external cavity phase, we find a cyclic
      scenario leading from stable intensity emission via periodic behavior to
      regular and irregular pulse packages, and finally back to stable emission.
      We reveal the underlying interplay of localized dynamics and global
      bifurcations.
AD  - Institut fur Angewandte Physik, Technische Universitat Darmstadt,
      Darmstadt, Germany.
FAU - Heil, T
AU  - Heil T
FAU - Fischer, I
AU  - Fischer I
FAU - Elsasser, W
AU  - Elsasser W
FAU - Krauskopf, B
```

Alta Plana

Here’s another page of “semi-structured” text, in this instance from a large, searchable database of life sciences literature abstracts. You see here many labeled fields including on labeled AB for abstract.

Life sciences researchers very often mine Medline and similar content databases containing literature and abstracts. They are look for protein interactions and the like in abstracts as part of lead generation in the pharmaceutical drug discovery process. Text analytics is far cheaper than clinical trials.



Case study: IBM's MedTAKMI

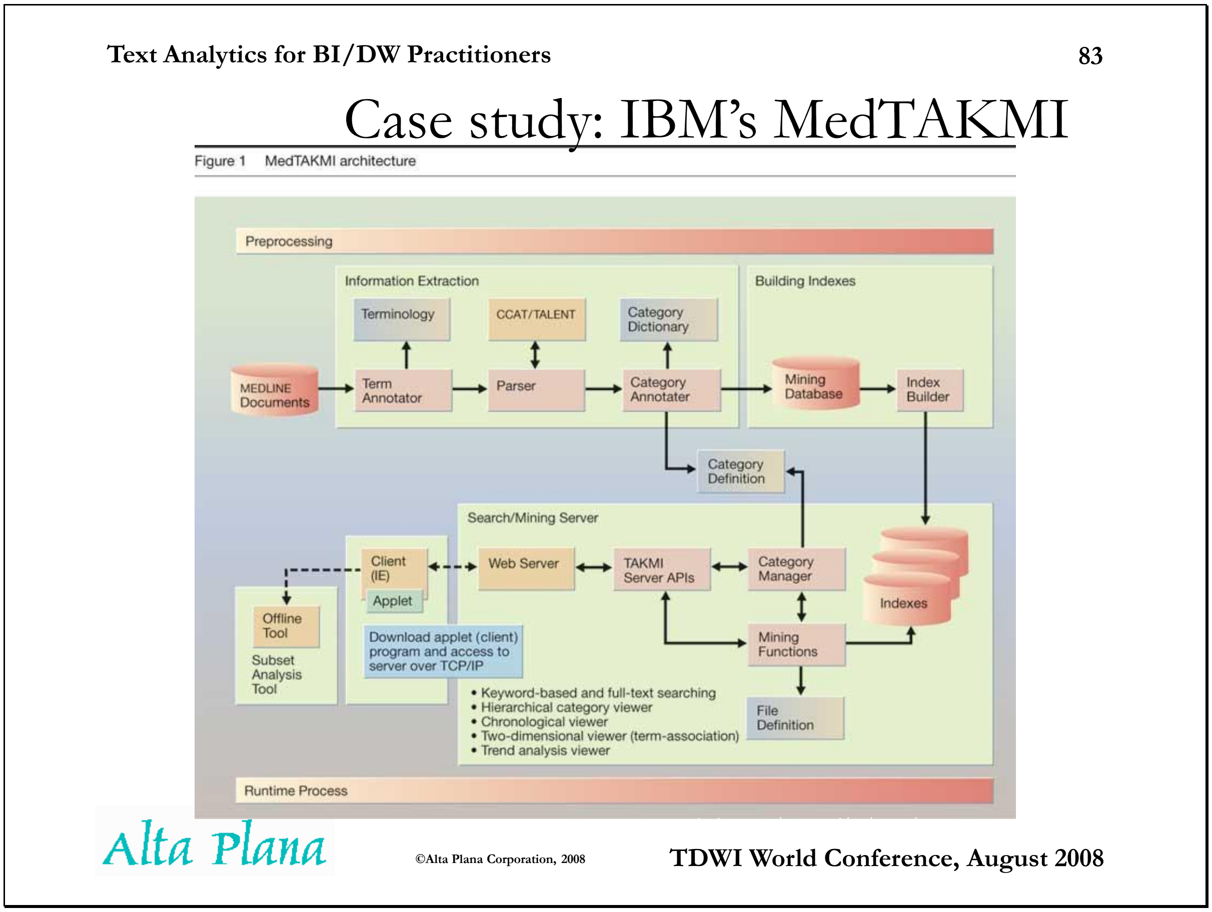
MedTAKMI = Text Analysis and Knowledge MIning for Biomedical Documents (*ibm.com*):

- 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.

IBM's MedTAKMI is a great case study for application of text analytics to mine Medline. We do some shallow linguistics that identifies parts of speech and relationships indicated by syntax.

Case study: IBM's MedTAKMI

Figure 1 MedTAKMI architecture



Here's an overall MedTAKMI process that runs from Information Retrieval through annotation and categorization to mining that generates indexes for semantic processing.

Case study: IBM's MedTAKMI

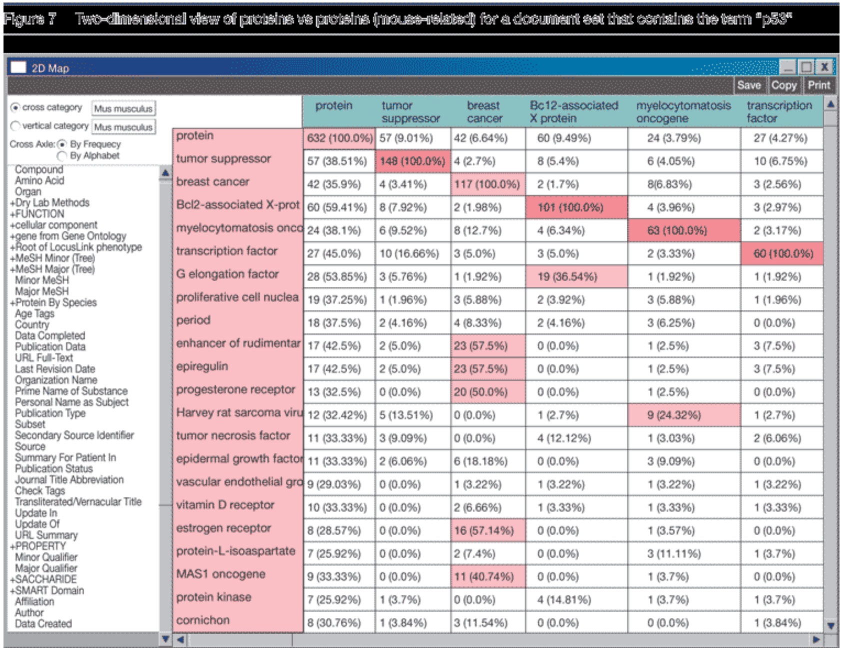


Figure 7 Two-dimensional view of proteins vs proteins (mouse-related) for a document set that contains the term "p53"

● cross category Mus musculus

○ vertical category Mus musculus

Cross Axle: ● By Frequecy ○ By Alphabet

	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%)

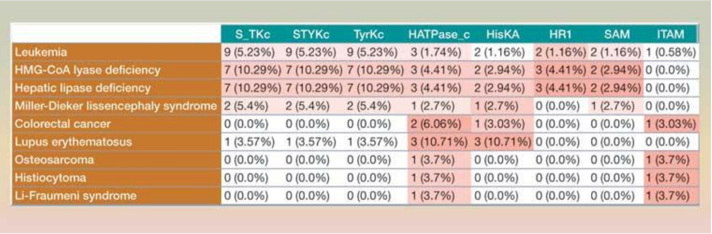
Here we display term counts and frequencies of cross-interactivity…

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

… and distill down to highly suggestive relationships.

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

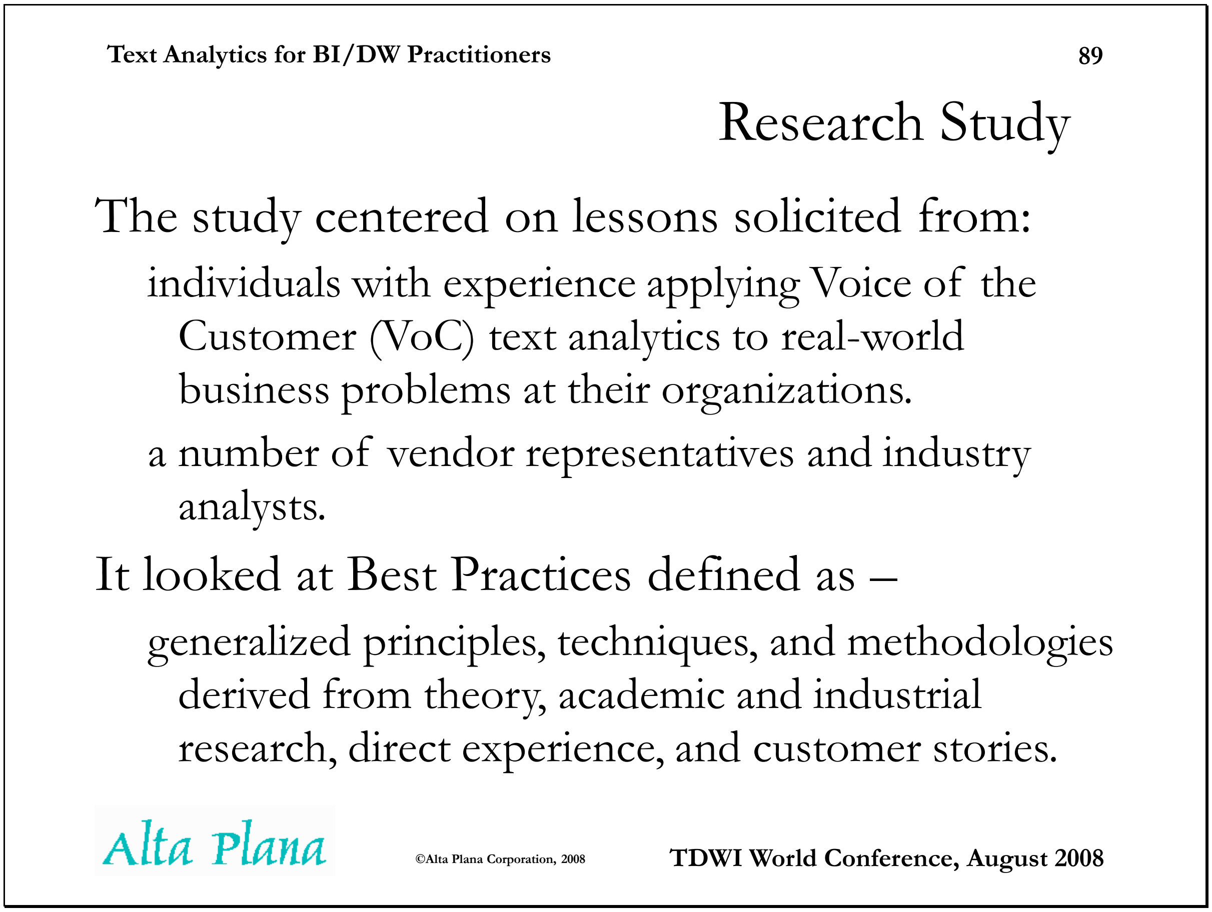
Unified Analytics –

- 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, 360^{o} 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.

I conducted my research study work in late-spring/early-summer 2008 in cooperation with the Business Intelligence Network. The report was sponsored by Business Objects, IBM, and SPSS although it was editorially independent.

Research Study

For the study,

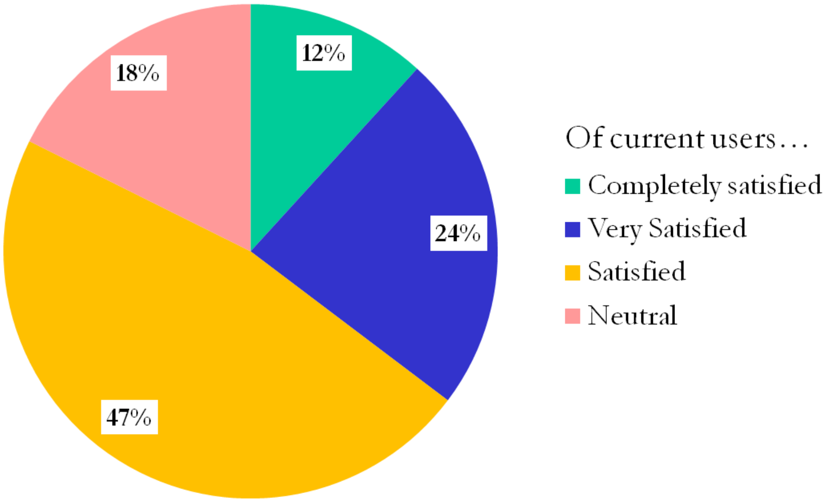
- I had free-form discussions with five industry analysts, six end users, and a variety of vendor executives.
- I additionally conducted a formal, small-sample survey of VoC text-analytics practitioners – end users and consultants – that attracted 26 valid responses.

Length of Respondent Experience

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

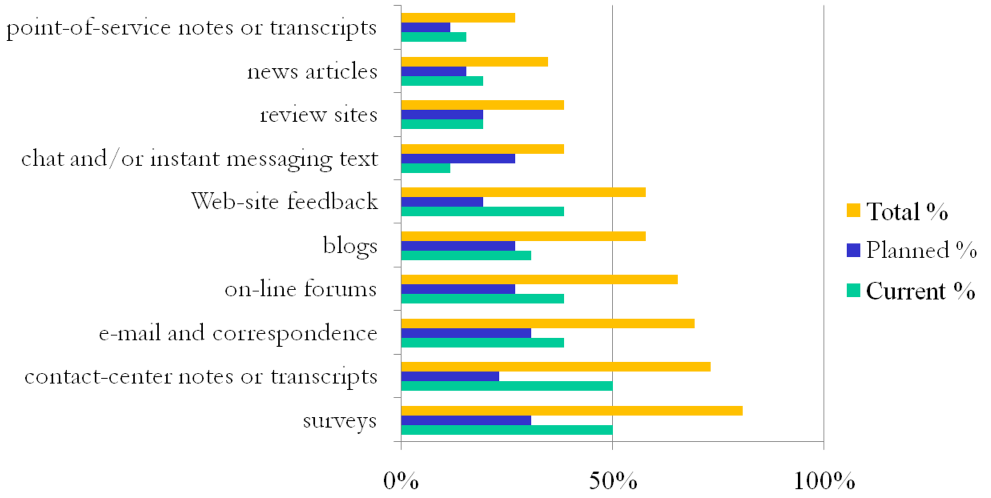
Length of Experience	Response Percent
still evaluating/not using	31%
less than 6 months	15%
6 months to less than one year	8%
one year to less than two years	12%
two years or more	**35%**

Satisfaction with VOC Text Analytics

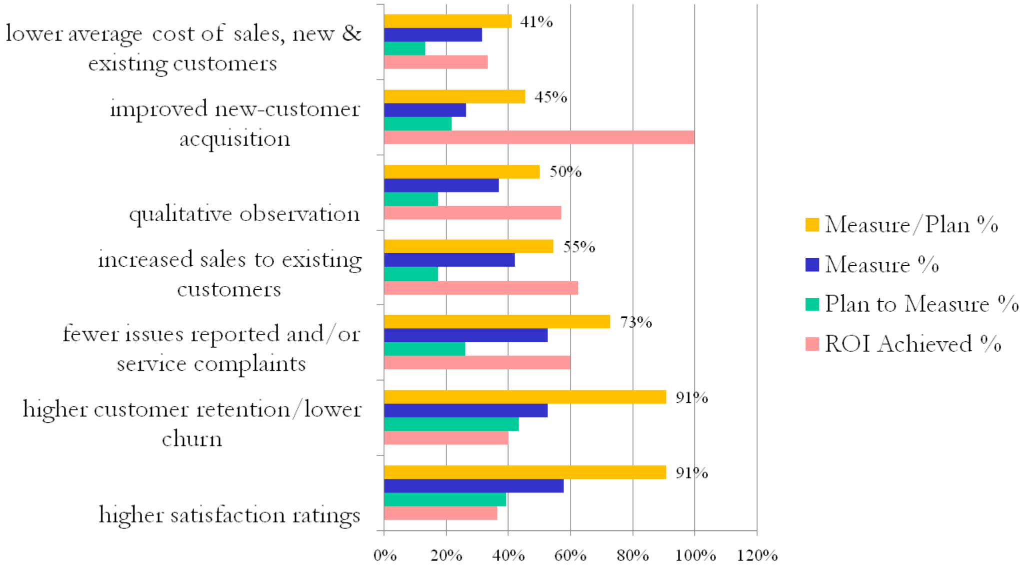


There were no responses below Neutral.

Information Analyzed



ROI Measured, Planned & Achieved



Alta Plana

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…

Alta Plana

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

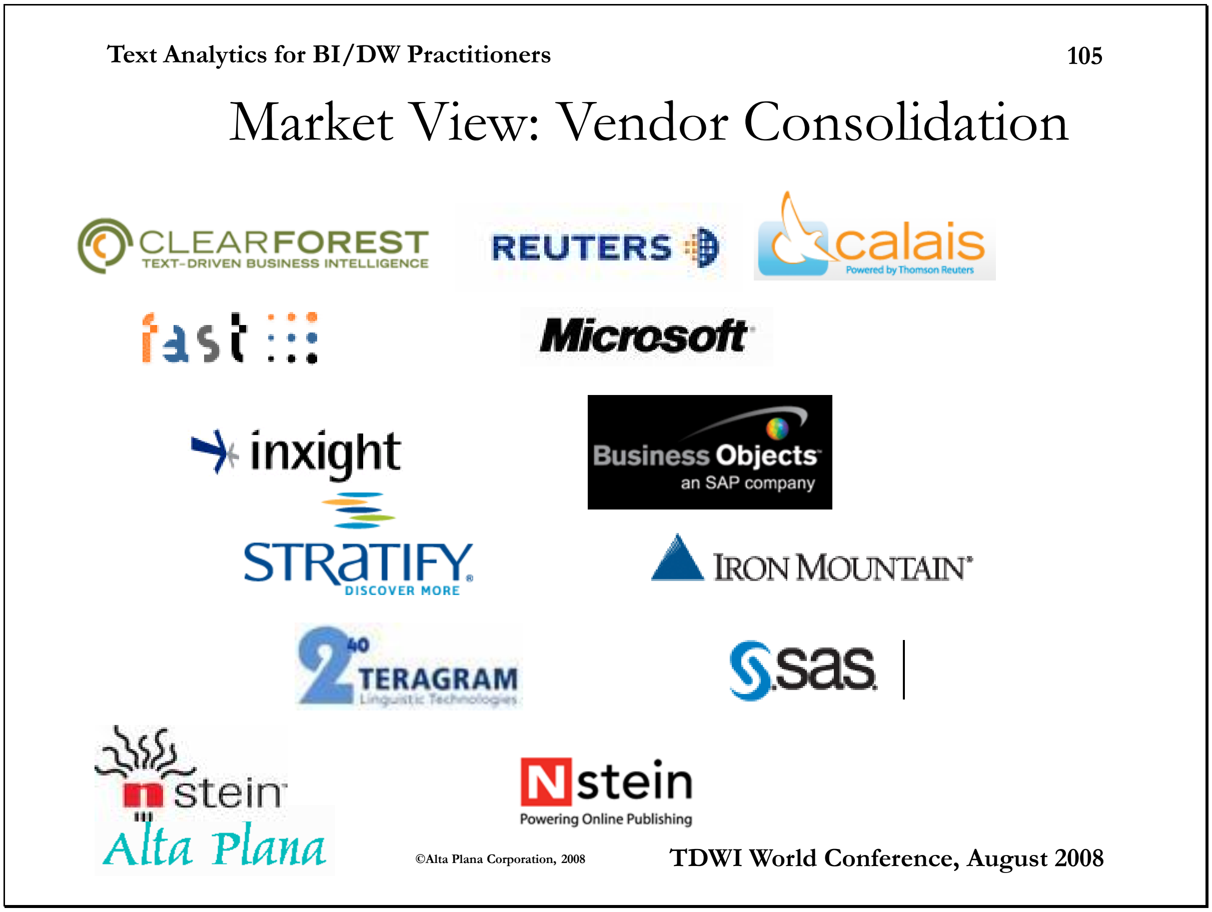
Integrate both structured and unstructured data.

[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.



Text Analytics for BI/DW Practitioners 105

Market View: Vendor Consolidation

©Alta Plana Corporation, 2008

TDWI World Conference, August 2008

There has been quite a bit of vendor consolidation in the last 18 months. I'd characterize most of it as larger companies looking to acquire text capabilities that would complement their existing capabilities.

The Nstein pseudo-consolidation is the transformation of the company from a technology focus to a media & publishing focus.

Market View: Community

4th Annual
Text Analytics Summit
YAHOO!
text analytics
revealing unstructured data
facebook.
LinkedIn.
text analytics
revealing unstructured data
Alta Plana

There’s been significant development of (non-academic) user forums and communities.

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!