

Integrated Analytics: Text and Data

Seth Grimes

Alta Plana Corporation

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

The Data Warehousing Institute

World Conference, Spring 2007

May 17, 2007

Alta Plana

Introduction

Seth Grimes --

Principal Consultant with Alta Plana Corporation near Washington DC.

Contributing Editor and *Breakthrough Analysis* columnist/blogger, Intelligent Enterprise magazine, *IntelligentEnterprise.com*.

Founding Chair, Text Analytics Summits (Boston 2005-7, Amsterdam 2007).

Morning Agenda

Information as an enterprise asset

Discerning structure; extracting information

Semantic search, Search-BI

Afternoon Agenda

Text-data integration

Business applications

Market survey

Implementation

Morning Agenda

Information as an enterprise asset

Solutions for numerical data.

“Content” solutions.

Discerning structure; extracting information

Semantic search, Search-BI

Enterprise information assets

Why? – Efficiency! Effectiveness! Profitability!

You know the drill:

360° views.

Single version of the truth.

One-to-one marketing/service delivery.

24/7.

and the (business oriented) goals:

Customer acquisition, retention & service.

Up-sell, cross-sell -- better, faster, cheaper.

Enterprise information assets

We're in the midst of a information explosion:

E-mail, Web sites, blogs, Wikis, ...

Documents, on-line books.

Images, video, and audio.

Tracking and geolocation data (RFID, GPS).

Transactional data.

Much of this information is in databases.

Yet enterprises realize that valuable knowledge is locked in the “unstructured” forms.

Enterprise information assets

“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 search vendor Verity

Enterprise information assets

The traditional analytics stack:

Operational systems – ODS.

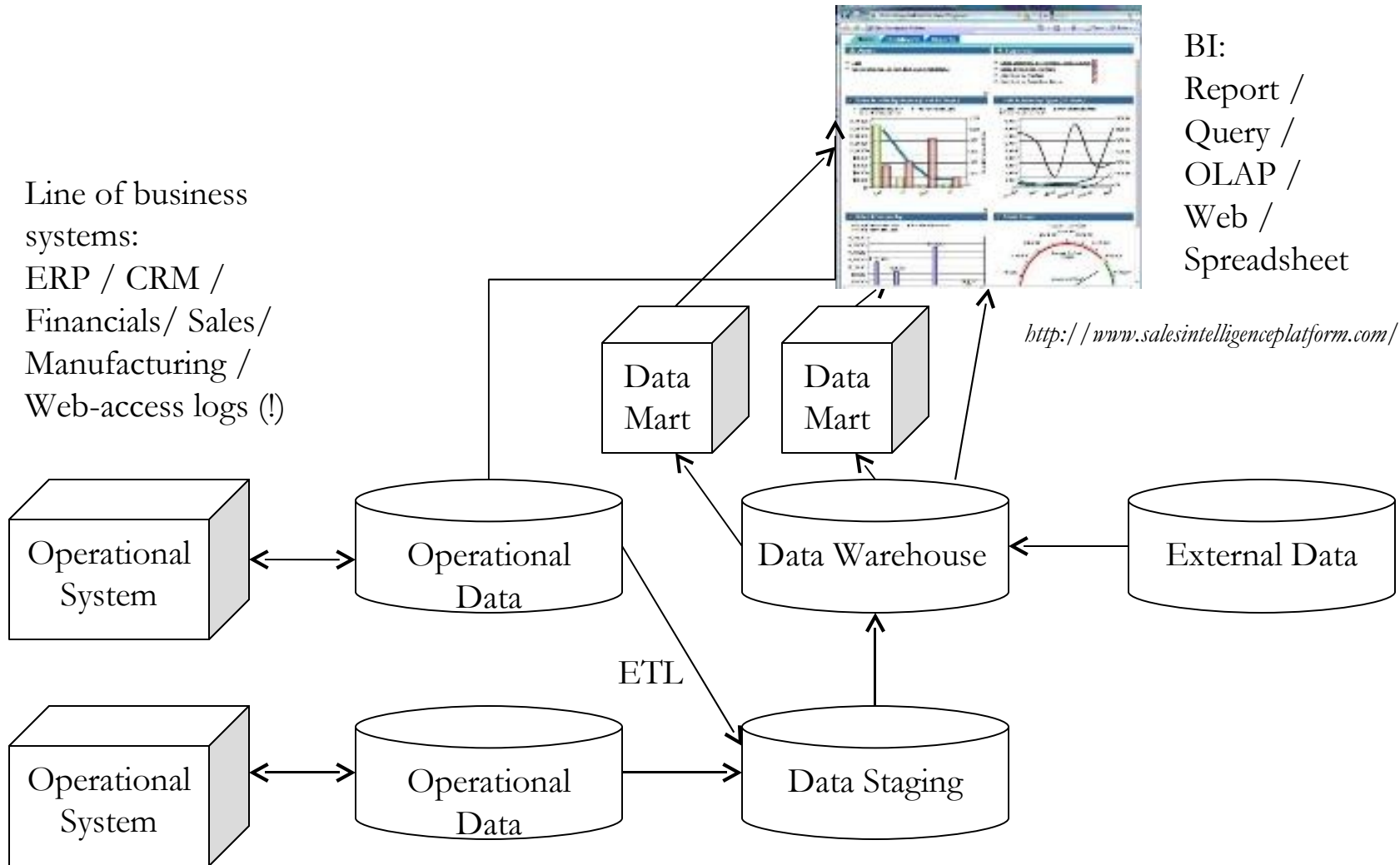
Data warehouse.

Data marts.

Reporting, OLAP, data mining, visualization.

Dashboards, portals, and spreadsheets.

Line of business systems:
ERP / CRM /
Financials/ Sales/
Manufacturing /
Web-access logs (!)



BI:
Report /
Query /
OLAP /
Web /
Spreadsheet

<http://www.salesintelligenceplatform.com/>

Enterprise information assets

Trends:

Operational/real-time/embedded BI.

Enterprise Application Integration.

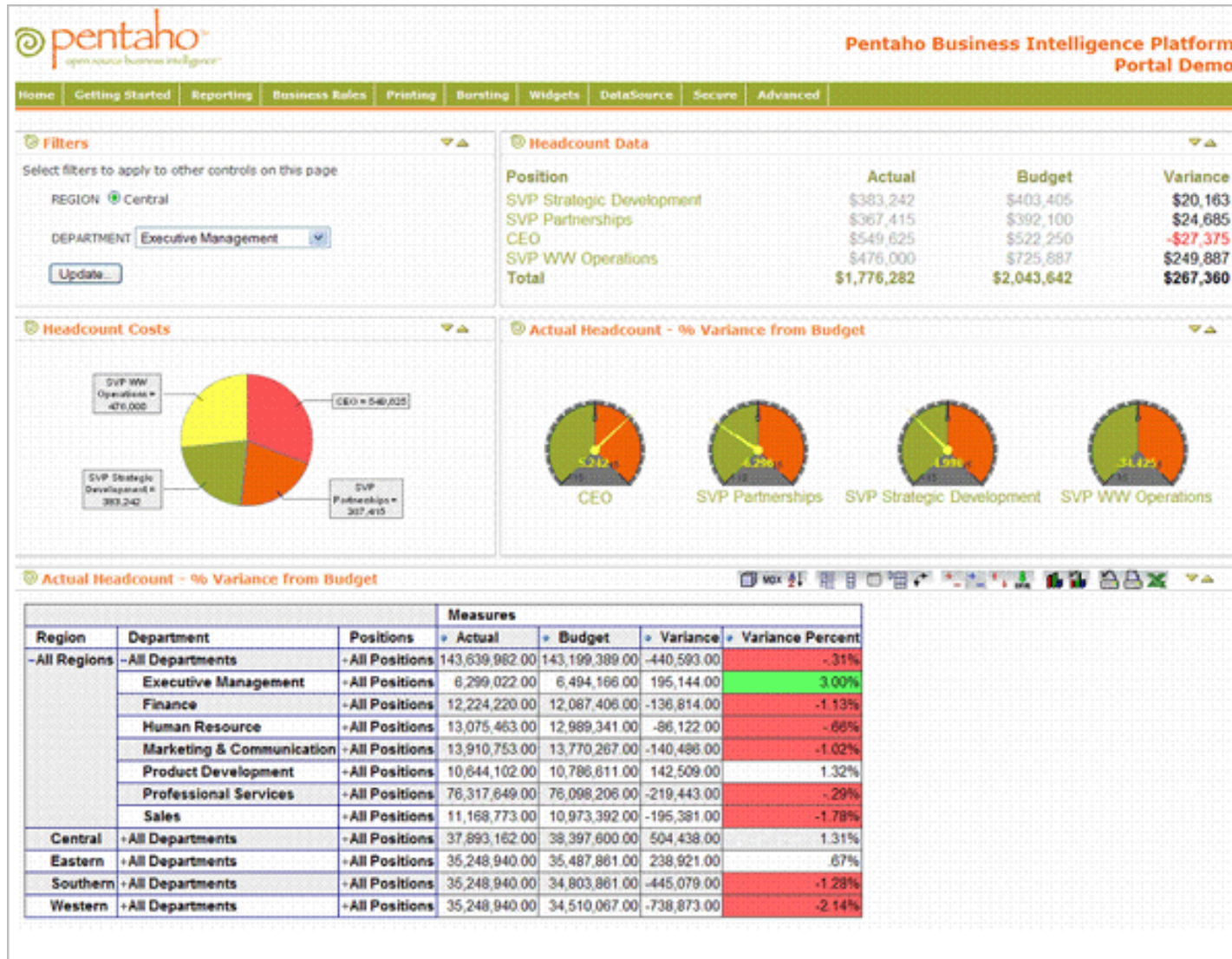
Portals/dashboards/mobile devices.

On-demand/SaaS/outsourced.

The mainstream is moving “up the stack.”

Vendor consolidation/emergence.

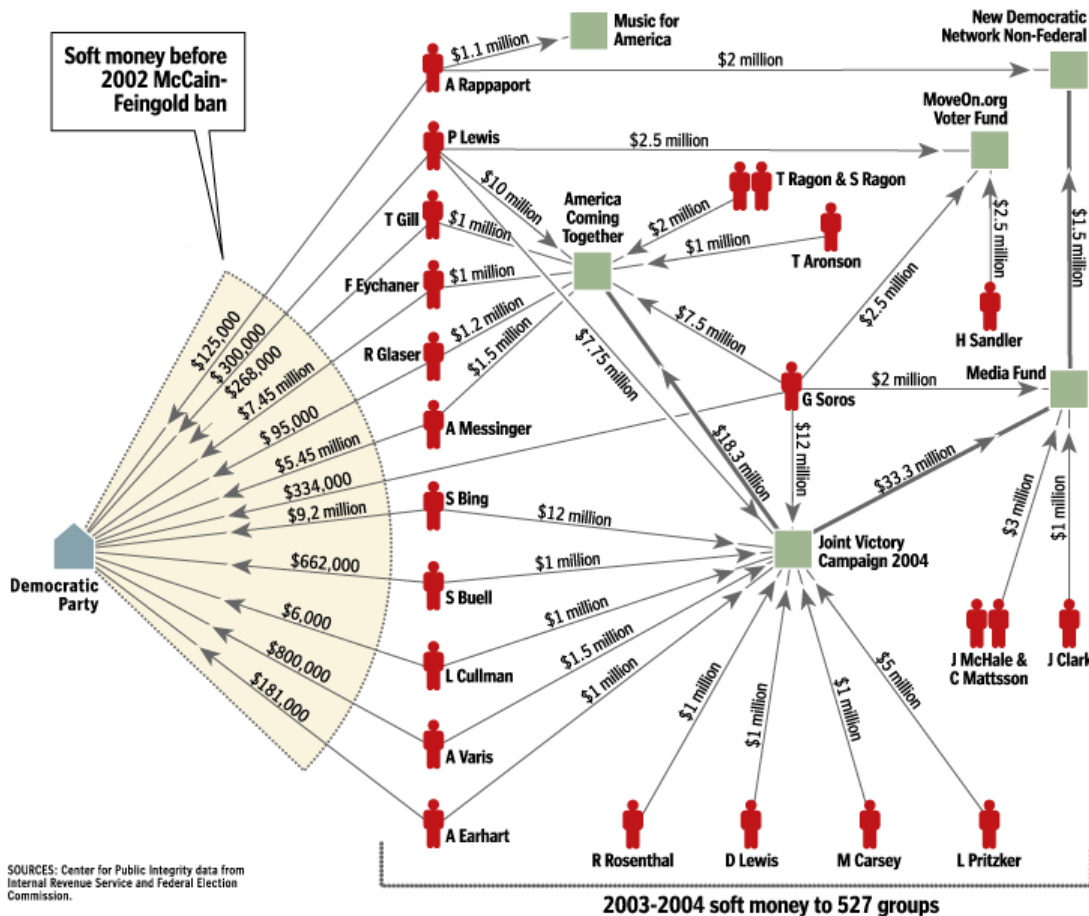
Open source platforms and tools.



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

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.

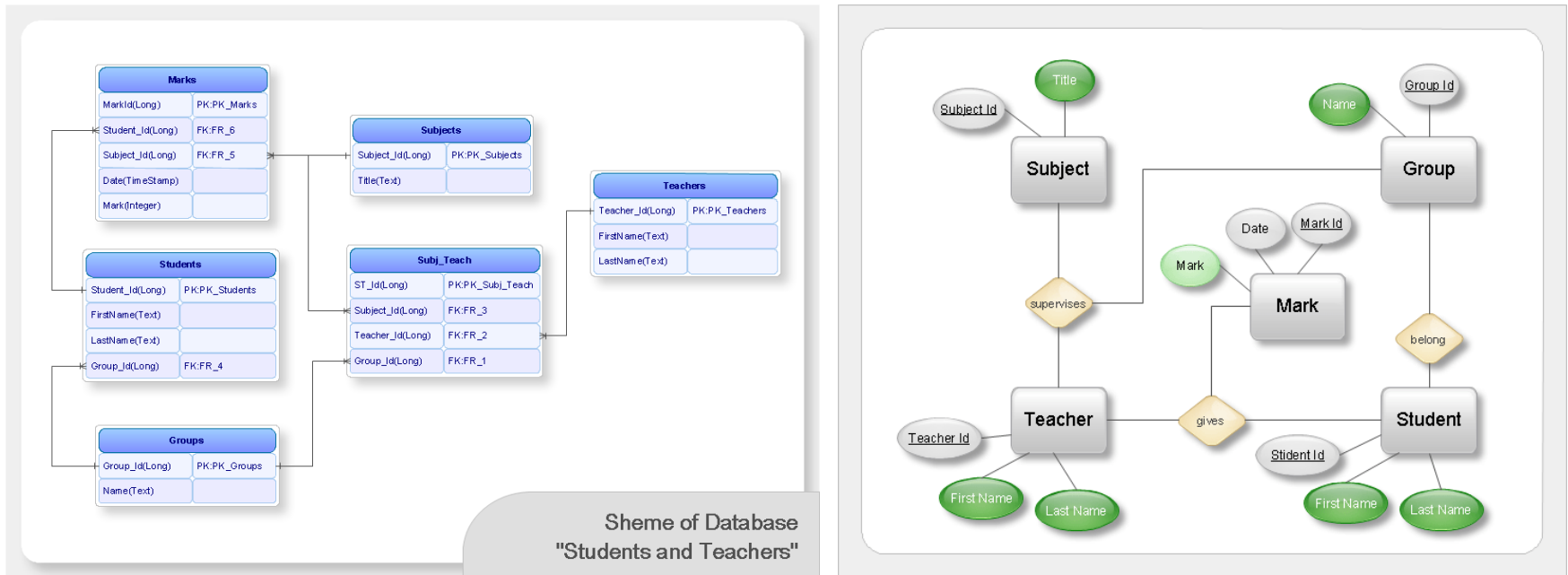
SOURCES: Center for Public Integrity data from Internal Revenue Service and Federal Election Commission.

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

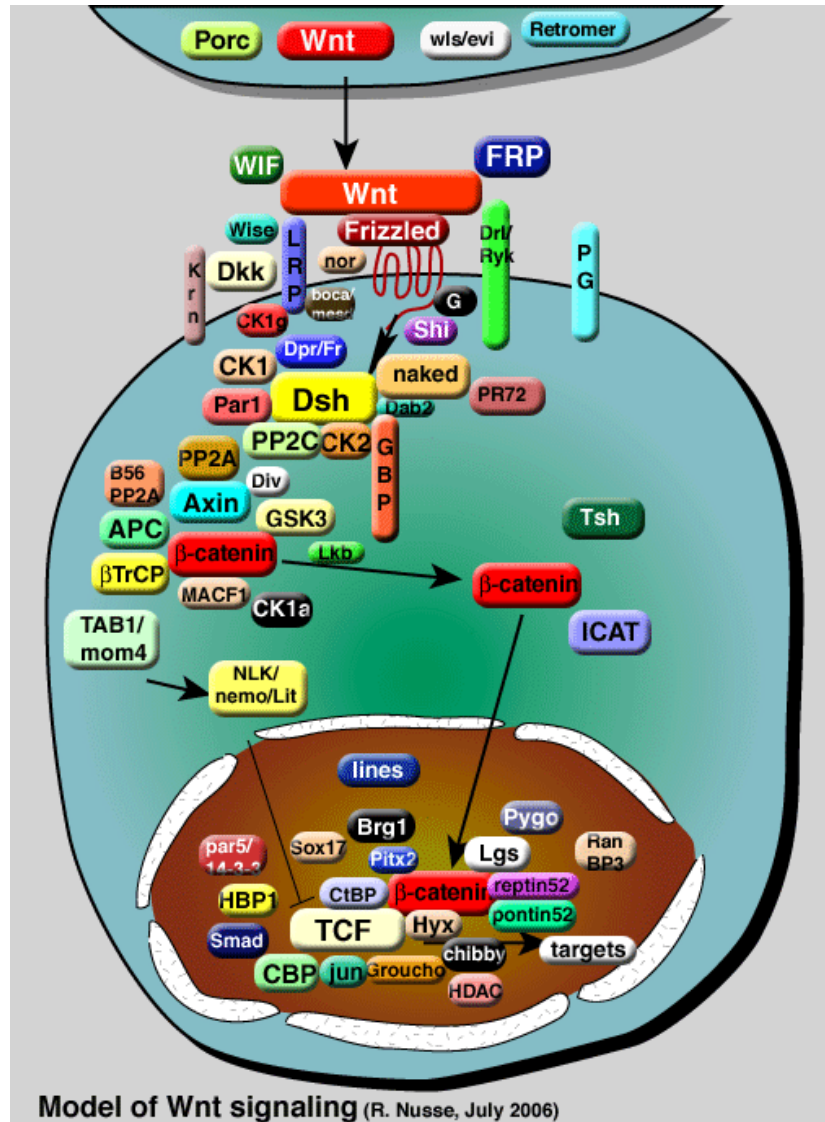
www.washingtonpost.com/wp-srv/politics/daily/graphics/527Diagram_101704.html

Enterprise information assets

But information doesn't come in visual, presentation form. It doesn't come this way either:



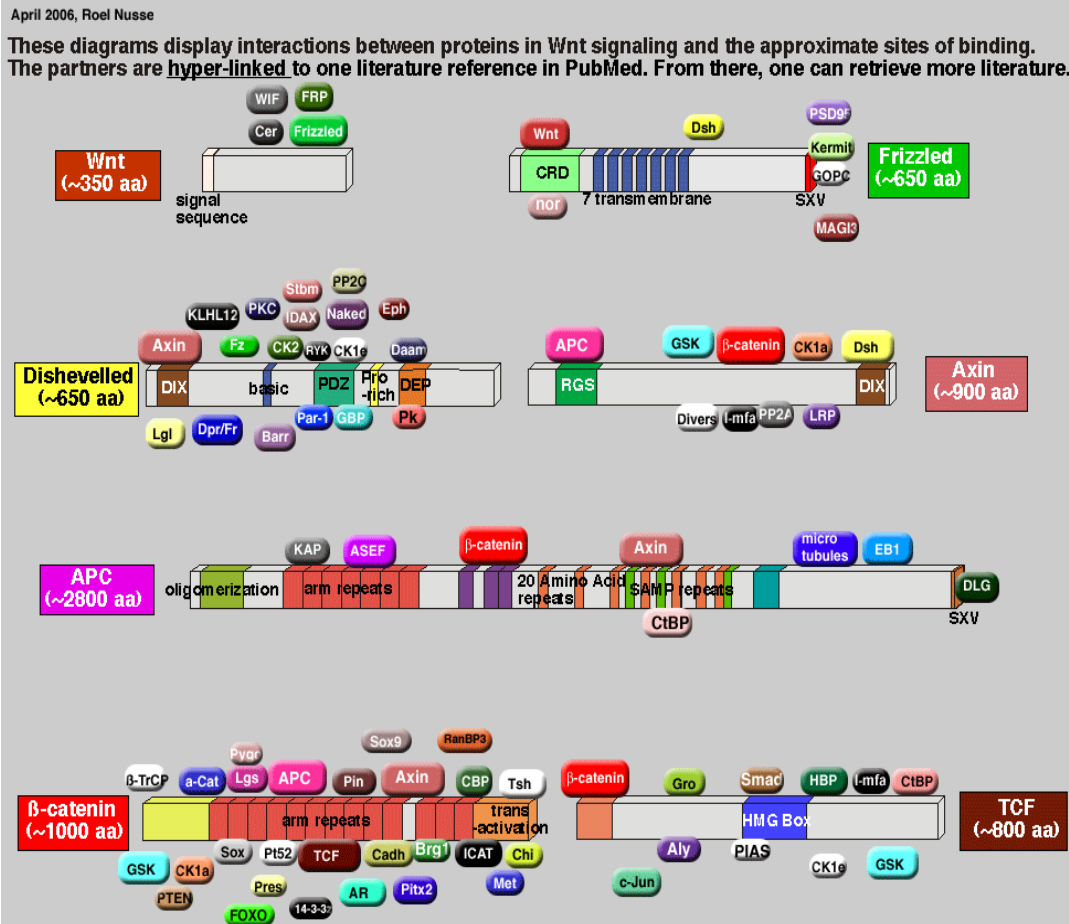
www.conceptdraw.com/en/products/cd5/ap_er_diagram.php



“The *Wnt* signaling pathway describes a complex network of proteins most well known for their roles in embryogenesis and cancer, but also involved in normal physiological processes in adult animals.”

-- D. C. Lie, S. A. Colamarino, H. J. Song, L. Desire, H. Mira, A. Consiglio, E. S. Lein, S. Jessberger, H. Lansford, A. R. Dearie and F. H. Gage (2005) "Wnt signalling regulates adult hippocampal neurogenesis" in *Nature* Volume 437, pages 1370-1375. Entrez PubMed 16251967

<http://www.stanford.edu/%7ernusse/wntwindow.html>



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

Enterprise information assets

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

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

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

“Content” technologies

So we need to store/access information, understand how to use it, and deliver it to “users.”

Consider Content Management, Knowledge Management, and Portals?

Why manage content?

What is enterprise knowledge?

How do you get from content to knowledge?

How do you present/represent knowledge?

“Content” technologies

Content Management System are analogous to data warehouses.

What about Web, fileserver, e-mail, desktop, and other non-CMS managed documents?

How far should we extend the CM concept?

How do we access and retrieve “content”?

CMSes are inherently limited, as are constrained access methods.

So we have search...

Search

Thus the orb he roamed

With narrow search, and with inspection deep

Considered every creature.

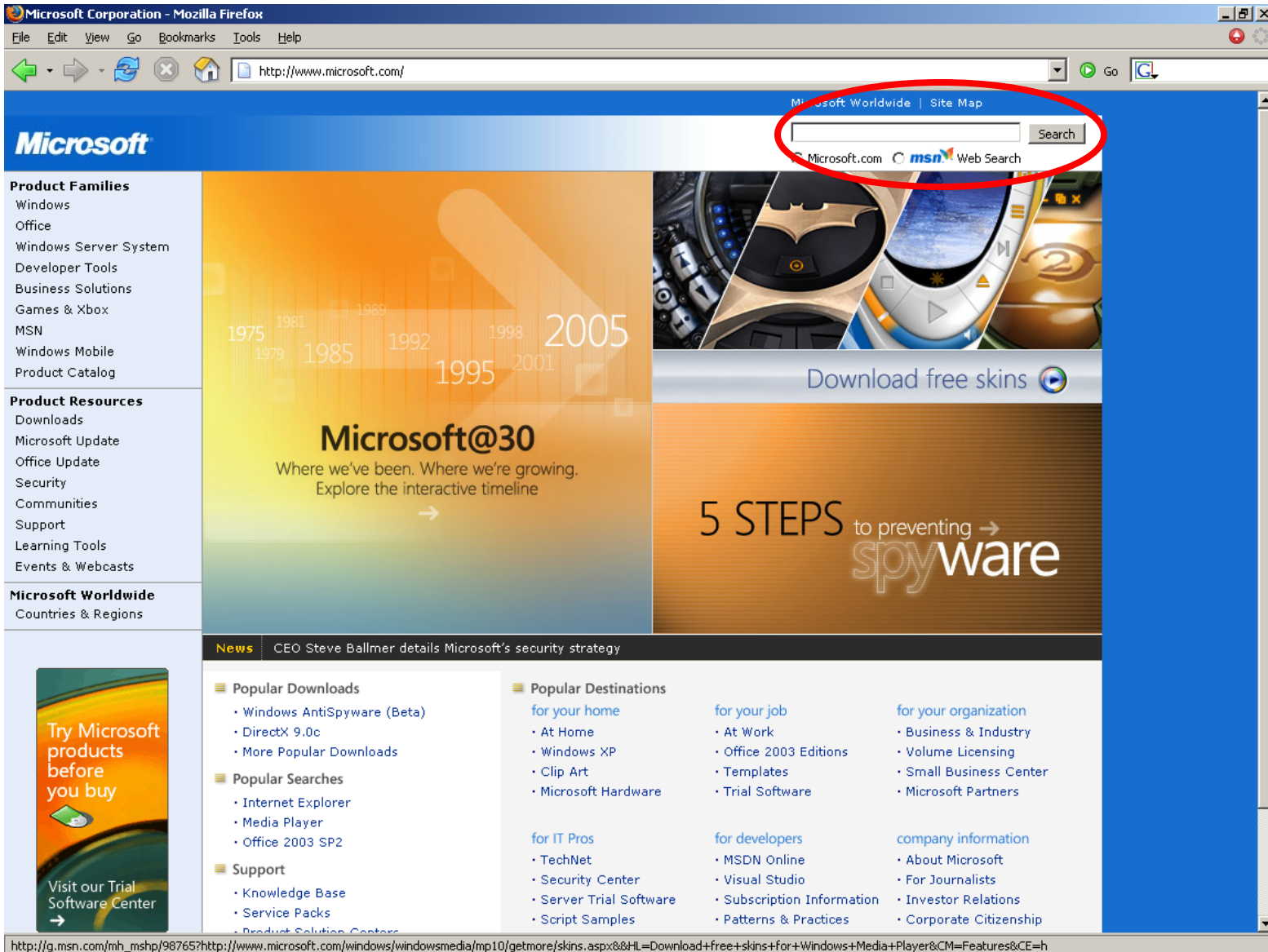
-- John Milton, Paradise Lost

Search is omnipresent.

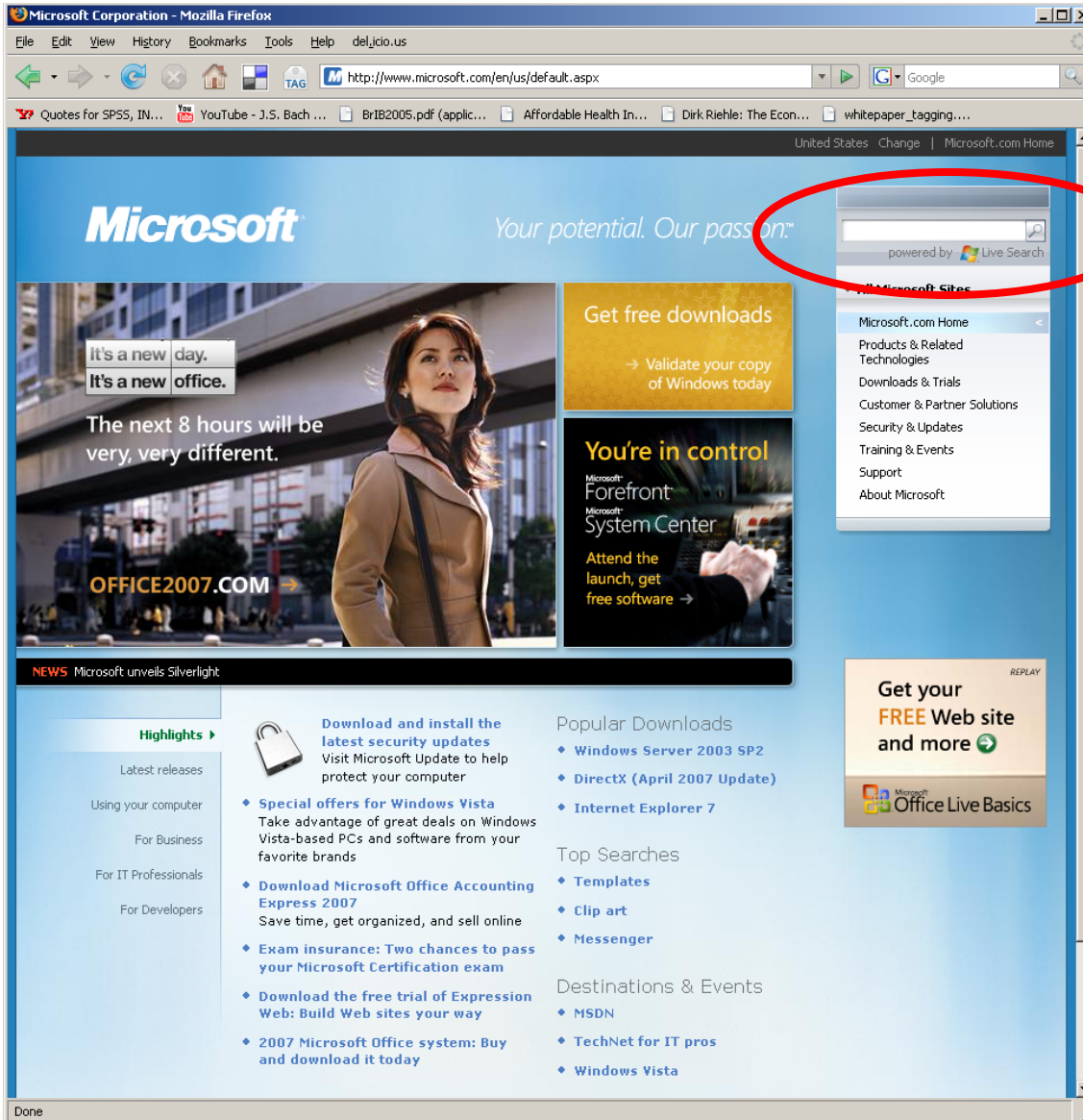
Is search enough?

Search





2005



... and 2007:
search is part
of a failure of
design.

Discussion 1: What's wrong with these designs?



Discussion 2: What's was the point of discussion 1?

Alta Plana

Search

How do we search? According to the Delphi Group:

Word or phrase: 47%

Iterative queries: 26%

Describing an idea or concept: 12%

Browsing topics: 10%

Metadata search: 5%

Our primary interest is in the bottom three.

Search

Is search is a failure of design? Maybe not.

Things are not where they “should” be.

There are too many of them.

There are too many kinds of things.

Kinds = meanings, uses, forms.

They change too fast.

They are not easy to categorize.

It’s hard to determine their relevance.

Other forms for UI are too cumbersome.

Search

What's a “thing” anyway? What do we want to find?

Text

Images

Video

Audio

Numeric data

Fielded data

Records, that is, from a database

Search

But that's not really it. I don't really (usually) want to find a particular document; I want to find a fact, the answer to a question.

- What was the population of Rome in 1848?
- What's the best price for new laptop that I'll use for business trips and around the office?
- What do people think of Bill Frist?

Search

For now, to look for answers, we use –

- Words & phrases: search terms.
- Entities: names, e-mail addresses, phone numbers
- Concepts: abstractions of entities.
- Abstract attributes derived from data and terms, e.g., “expensive,” “comfortable,” “dangerous”
- Qualifiers: include/exclude, and/or, not, etc.

These are simplifications of complex patterns.

They are proxies for what we really want: relationships, connections.

Discovery

Search (obviously) has value but it isn't 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**.*

Text mining is a variant of knowledge discovery.

Ramana Rao, CTO of Inxight, says that Search is people chasing documents while Text Mining is documents chasing people.

Discovery

Search/Query
(goal-oriented)

Discovery
(opportunistic)

Fielded
Data



Documents



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

... and analytics

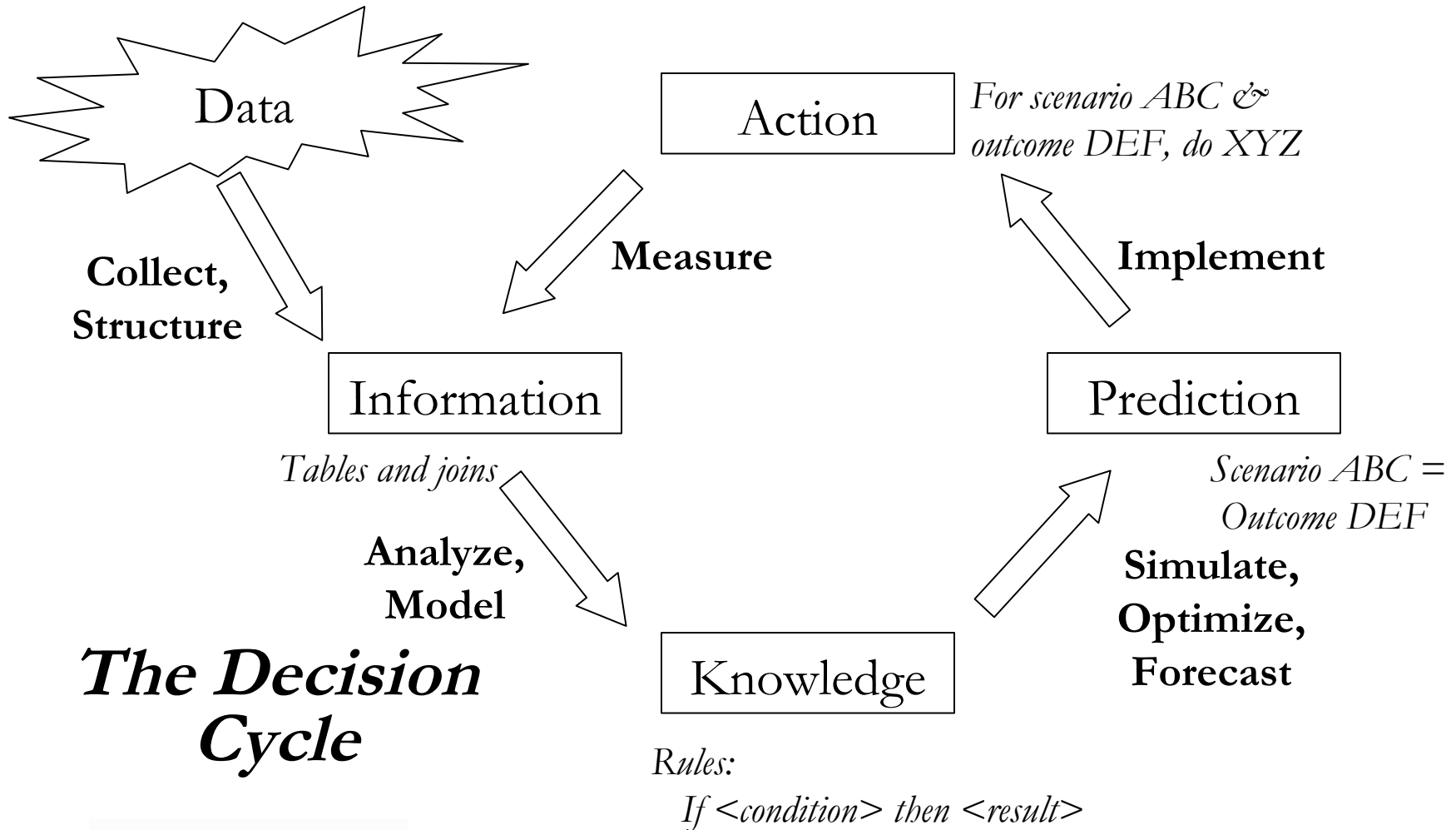
Text mining is part of a renewed focus on numbers and analytical magic --

- measuring
- modeling, forecasting, and predicting
- optimizing

-- and a valuation of knowledge.

We need to understand and manage knowledge-related business processes and align them with business goals.

... and analytics



... and analytics

Laying the groundwork:

- Statistics!

Breakthrough:

- Data warehousing: structuring data for analysis
- Spreadsheets
- Content and Knowledge Management and Search

Bread & butter analytics:

- Business Intelligence: OLAP, query & reporting

Revival:

- Advanced analytics: reintroduction of statistics

... and analytics

Data Mining comprises a number of statistically rooted techniques for automated detection of patterns and relationships:

Segmentation and clustering -- finding and applying the characteristics (dimensions) that best group data

Link analysis -- detecting association patterns and rules

Predictive modeling -- classification and scoring

Predictive modeling -- regression for anomaly detection and forecasting

Predictive modeling completes the decision cycle.

... and analytics

Text Mining = Data Mining of textual sources.

Text Mining = Knowledge Discovery in Text.

Text mining is a form of analytics.

But recognize that the overall endeavor involves more steps than does data mining. What we're really after is *text analytics*.

Morning Agenda

Information as an enterprise asset

Discerning structure; extracting information

Semantic search, Search-BI

Text analytics

Text (and media?) mining **automates** what researchers, writers, scholars, ... and all the rest of us have been doing for years. Text mining – *applies linguistic and/ or statistical techniques to extract concepts and patterns that can be applied to categorize and classify documents, audio, video, images.*

transforms “unstructured” information into data for application of traditional analysis techniques via modeling.

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

Text analytics

But to digress...

Is text really unstructured?

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

*Text is instead **unmodeled**.*

Is the Web, which is a document collection, unstructured? What about a library?

No! The Web is structured via links, a library via a catalog.

Does **Search** exploit the structure inherent in documents or the Web?

1) No if keyword based. 2) Somewhat as links imply relevance.



Text analytics

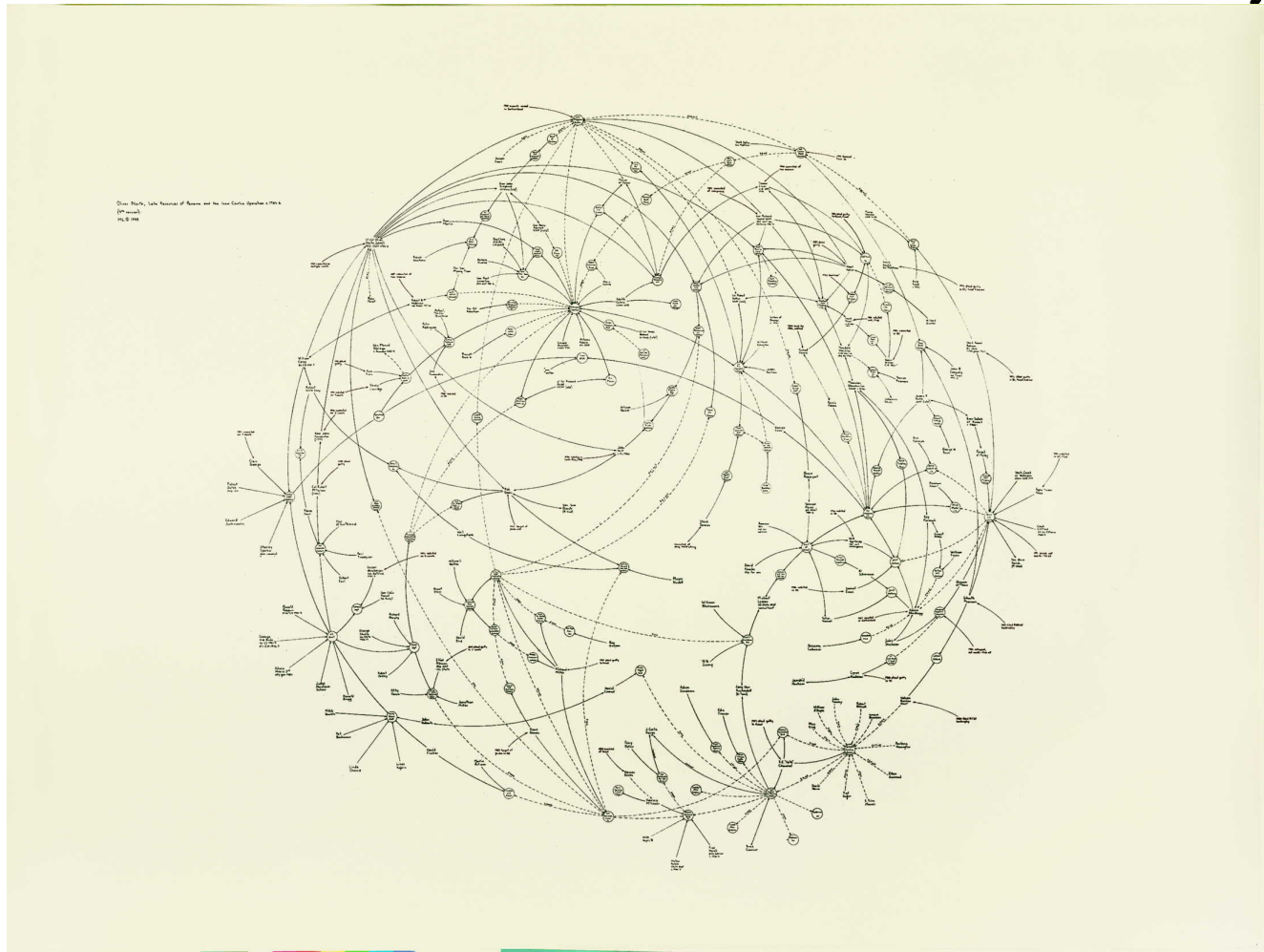
So the form may be unstructured but the content isn't. That's an important distinction.

How do we create structure in order to mine textual sources?

And what do we do with the information we've mined, with the knowledge we've discovered?

Again, both elements are in the realm of analytics. Consider the second first...

Text analytics



Mark Lombardi: Oliver North & Iran-Contra

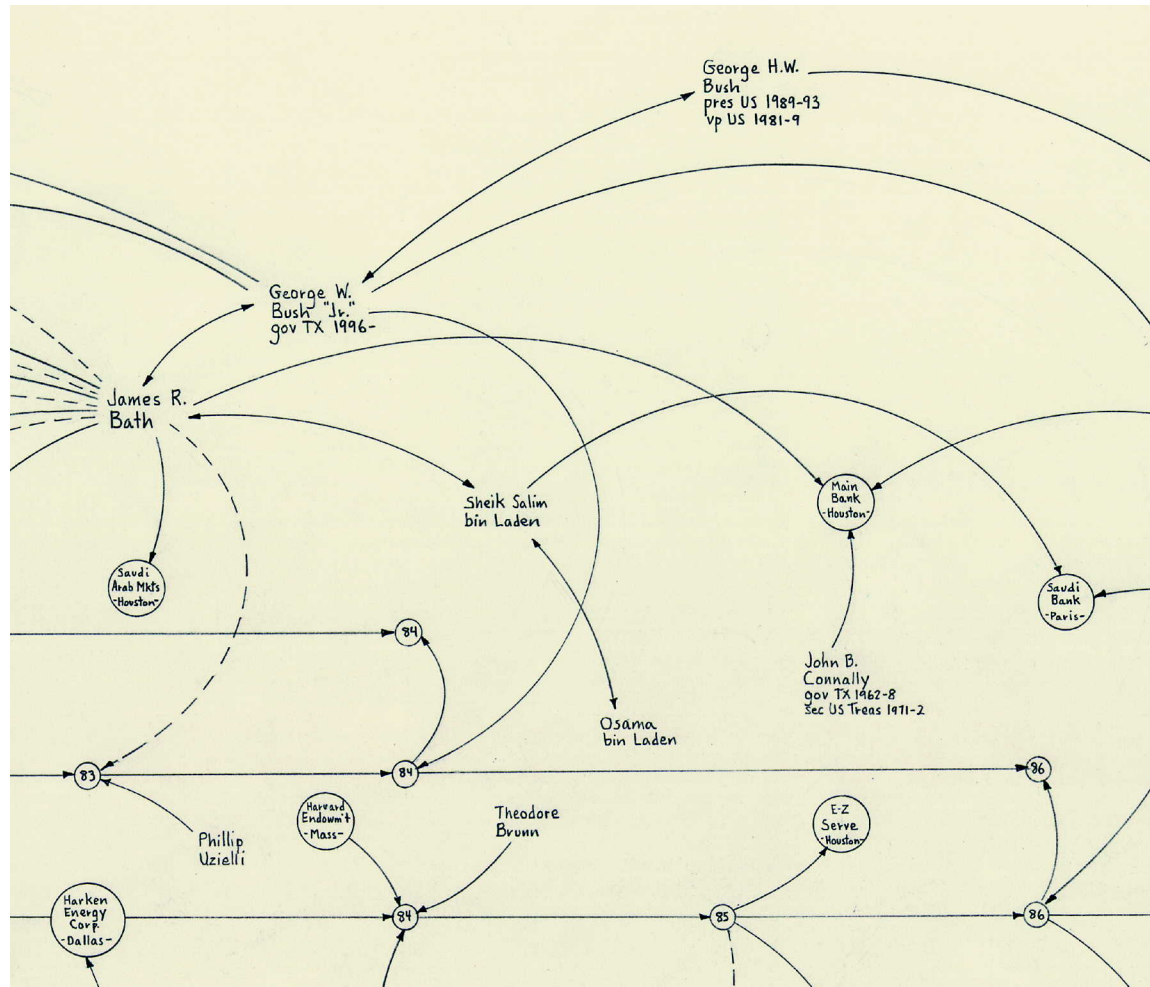
Text analytics

Mark Lombardi is an investigative reporter's Conceptual artist. His subject is conspiracy and scandal, his method is to "follow the money." His pursuit results in big airy line drawings that exemplify Conceptual art's propensity for diagrams, masses of information and showing how the world works.... To keep facts and sources straight, he created a handwritten database that now includes around 12,000 3-by-5-inch cards.

-- Roberta Smith in the *New York Times*, Dec. 25, 1998

Images on previous and next slides are courtesy of Pierogi, pierogi2000.com

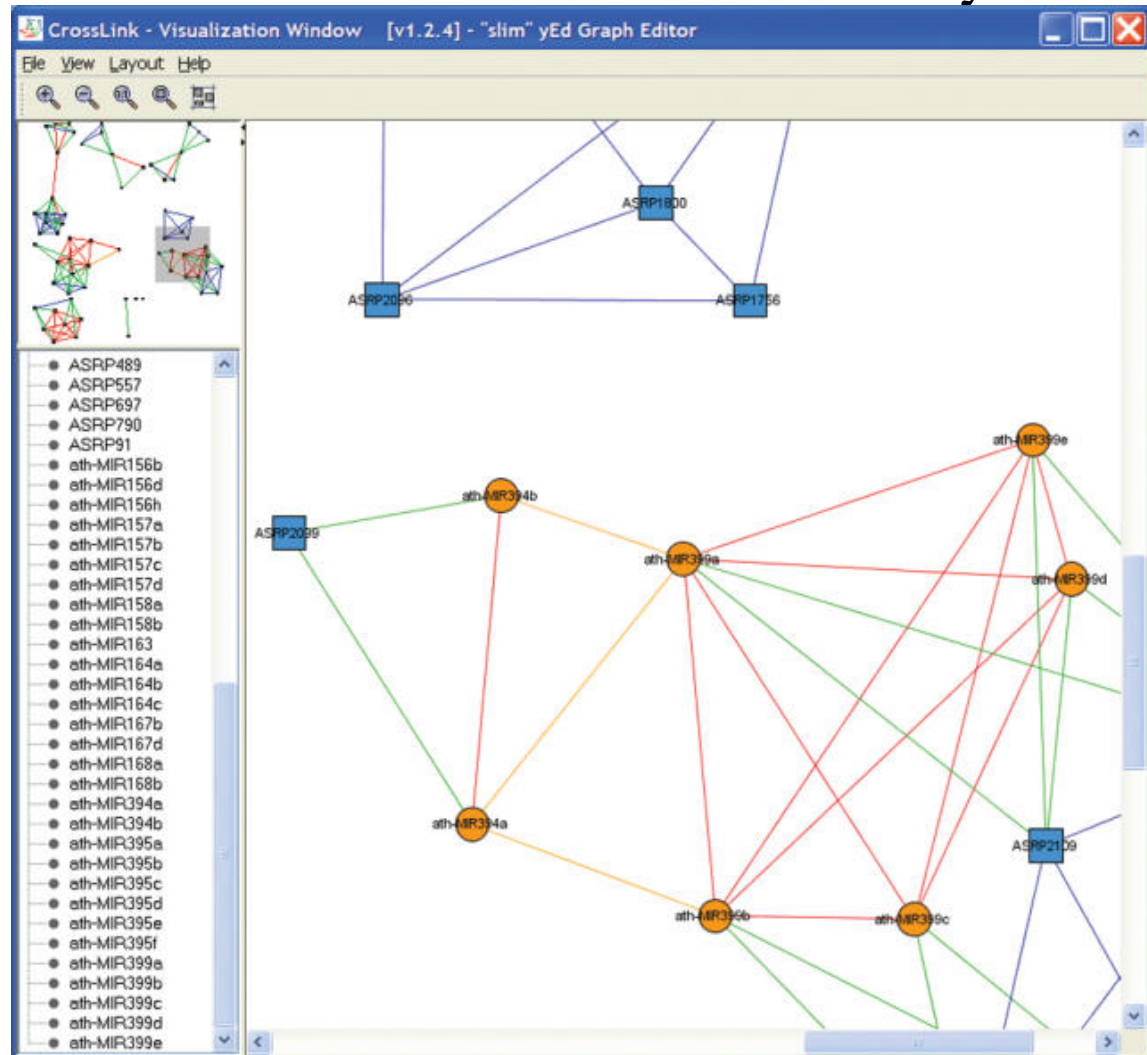
Text analytics



Mark Lombardi: george w. bush, harken energy... (detail)

Text analytics

Now we can automate the visualizations.



Text analytics

Typical steps in text analytics include --

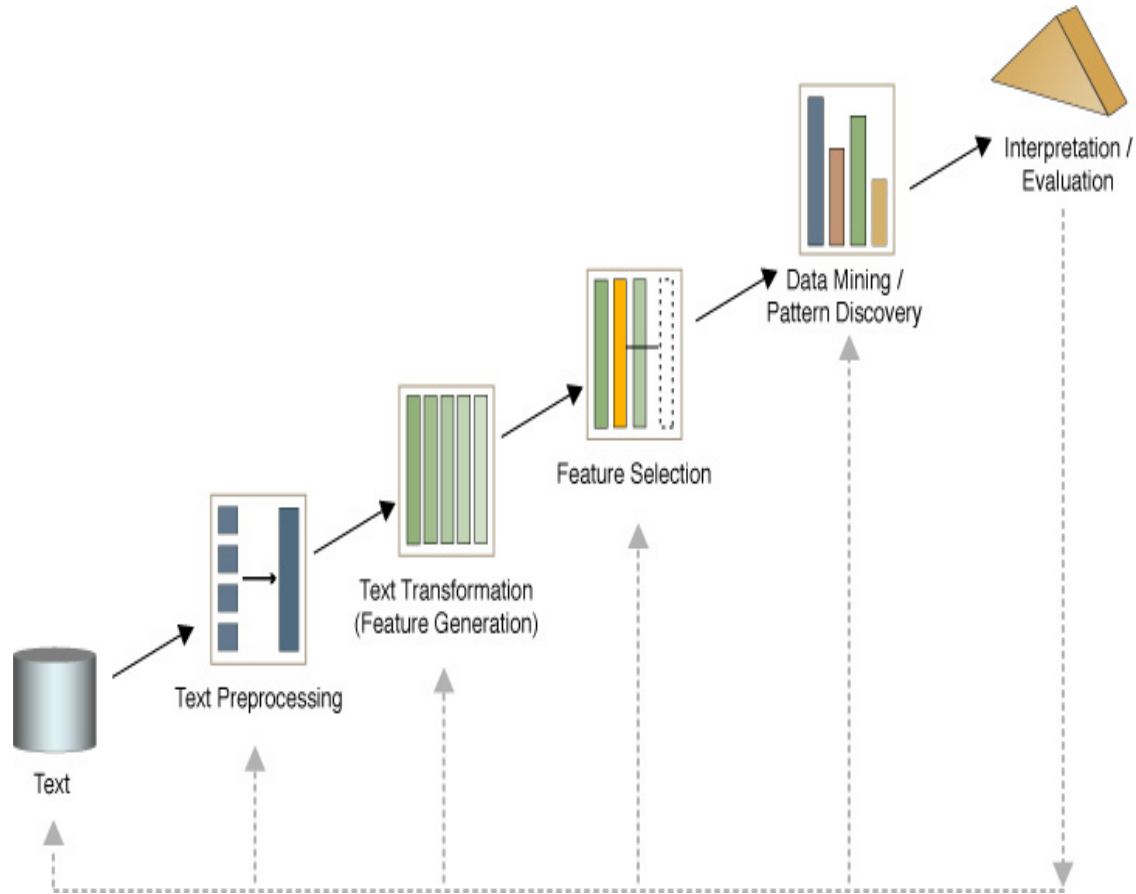
- Retrieve documents for analysis.
- Apply statistical &/ linguistic &/ structural techniques to identify, tag, and extract entities, concepts, relationships, and events within document sets.
 - tagging = text augmentation
- Create a categorization/taxonomy from the extracts or acquire and apply a domain-specific taxonomy.
- Apply statistical techniques to classify documents, look for patterns such as associations and clusters.

Foundations

Mid-way between parsing and classifying, there's categorizing: generating taxonomies.

- From Wiki: “Taxonomy is probably the most familiar kind of organization or classification scheme used in ContentManagement. The basic simple taxonomy is a hierarchy (or tree) with a top element (or root), depending on your preference for TopDown or BottomUp design. Nodes (or branch points) are names for things (objects) or concepts.”
- (Wiki itself is an Ontology, a knowledge representation, no? And Lombardi's work?)

Text analytics



Process schematic: Je Wei Liang, www.database.cis.nctu.edu.tw/seminars/2003F/TWM/slides/p.ppt

Augmentation

Text augmentation exercise (1) – What entities do you see here?

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

Augmentation

Text augmentation (2):

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

[We have not marked every discernable feature.]

Alta Plana

Augmentation

Text augmentation (3) – tagging:

Date: <datetime><day>Sun</day>, <dom>13</dom>
<mon>Mar</mon> <year>2005</year> <time>19:58:39 -0500<
</time></datetime>

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

To: <name>Seth Grimes</name>
<email>grimes@altaplana.com</email>

Subject: Re: Papers on analysis on streaming data

<name>seth</name>, you should contact <name>divesh srivastava
</name>, <email>divesh@research.att.com</email>

regarding <company>**at&t labs**</company> data streaming technology.

<name>adam</name>

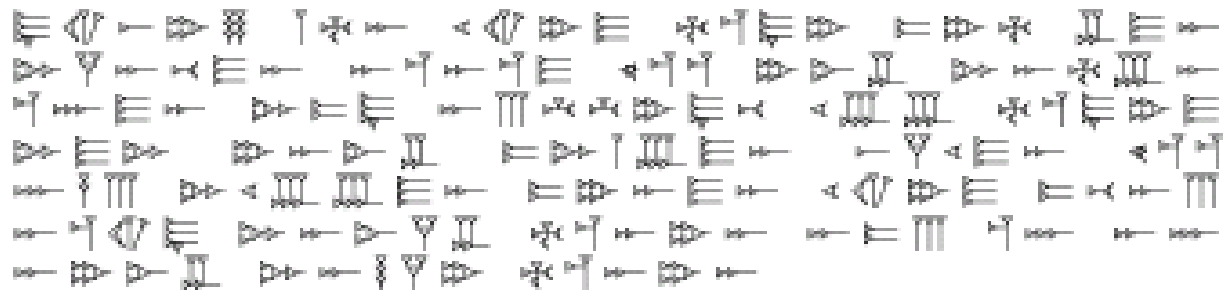
NLP

This is shallow parsing. What about (semantic) relationships, e.g.,

“you should contact name divesh srivastava regarding at&t labs data streaming technology”

Co-occurrence, also simple statistical signatures, are not enough.

Ugaritic Cuneiform Script



NLP

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.

NLP

Natural Language Processing (NLP) –

Seeks to understand:

Speaker: *Intention, Generation, Synthesis*

Hearer: *Perception, Analysis, Disambiguation, Incorporation*

Is multi-stage, implemented as a pipeline:

Lexical / syntactic analysis: *word morphology, part of speech, sentence/phrase/clause boundary.*

Semantic interpretation: *term normalization, disambiguation.*

We want to infer and extract latent semantics.

Material drawn from Hagit Shatkay, Queen's University.

NLP

We want concepts and not just entities.

What concepts are found in these similar examples?

Smaller cars generally get better gas mileage than larger cars.

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

Ford is an American automobile manufacturer and Nissan is Japanese.

NLP

What concepts are found in these domain-related statements?

Smaller cars generally get better gas mileage than larger cars.

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

Ford is an American automobile manufacturer and Nissan is Japanese.

Vehicle is a *concept* with *conceptual* size and energy consumption attributes and a *conceptual* engine type.

Energy consumption itself has a relative measure.

Nationality is another concept. What's Ford?

Alta Plana

NLP

What's Ford? –

“Ford is an American automobile manufacturer...”

2.A president?

3.A company that both makes and sells cars and other stuff?

4.A shallow place you cross a river?

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

Foundations

Let's go back to that 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

Foundations

An e-mail message is “semi-structured.”

Semi=half. What’s “structured” and what’s not?

Is augmentation/tagging and entity extraction enough?

What categorization might you create from that example message?

If we extracted all the entities to a database, what could you do with them?

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.

Entrez PubMed - Mozilla Firefox
 http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?CMD=Display&DB=pubmed

NCBI PubMed
 A service of the National Library of Medicine and the National Institutes of Health
 My NCBI Sign In Register

All Databases PubMed Nucleotide Protein Genome Structure OMIM PMC Journals Books

Search PubMed for Go Clear

Limits Preview/Index History Clipboard Details

Display MEDLINE Show 20 Sort by Send to

All: 1 Review: 0

1: Heil T et al. Delay dynamics of semiconduct... [PMID: 16241333] Related Articles, Links

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 für Angewandte Physik, Technische Universität Darmstadt, Darmstadt, Germany.
 FAU - Heil, T
 AU - Heil T
 FAU - Fischer, I
 AU - Fischer I
 FAU - Elsasser, W
 AU - Elsasser W
 FAU - Kreuzkopf, B

Done

Alta Plana

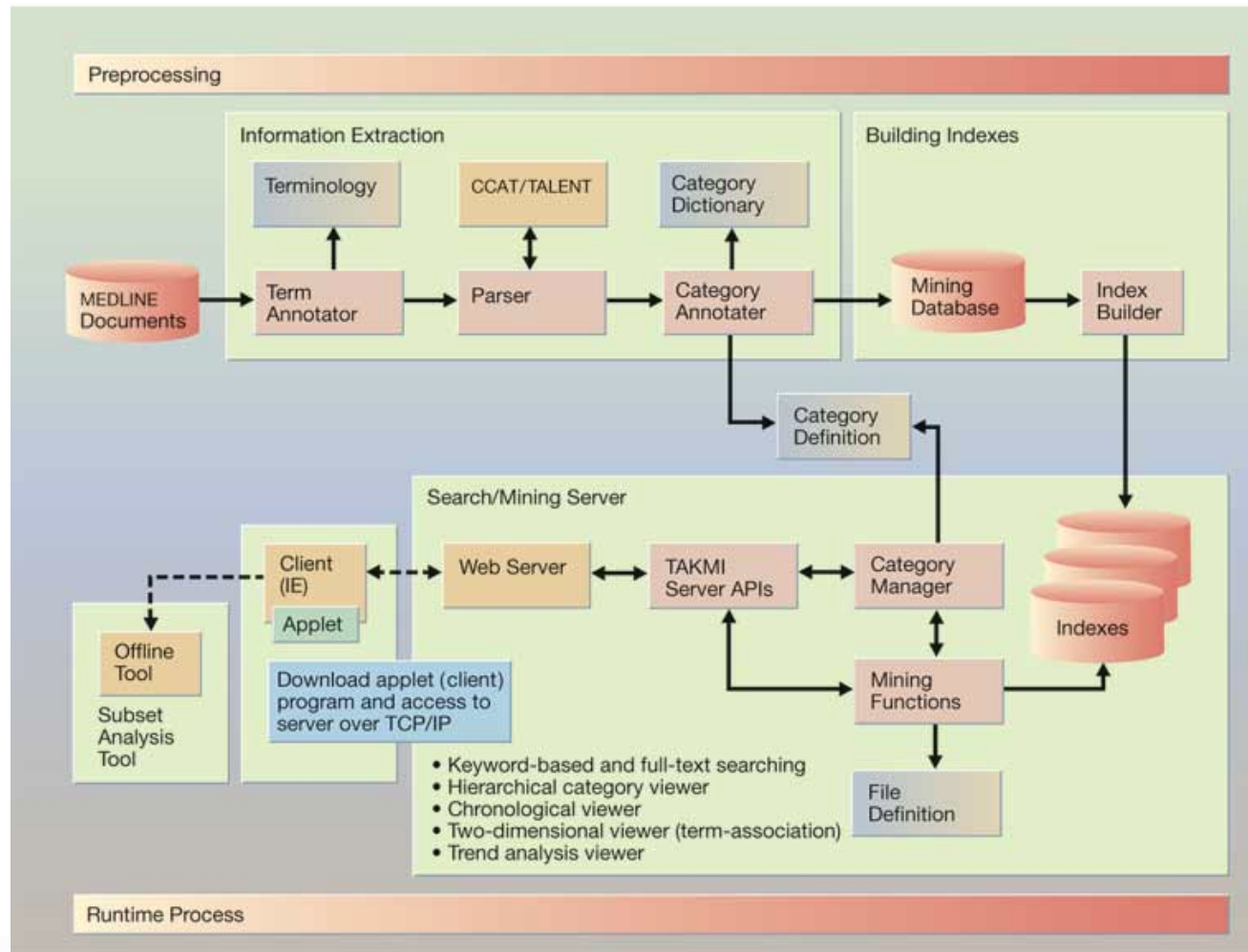
Case study: IBM's MedTAKMI

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

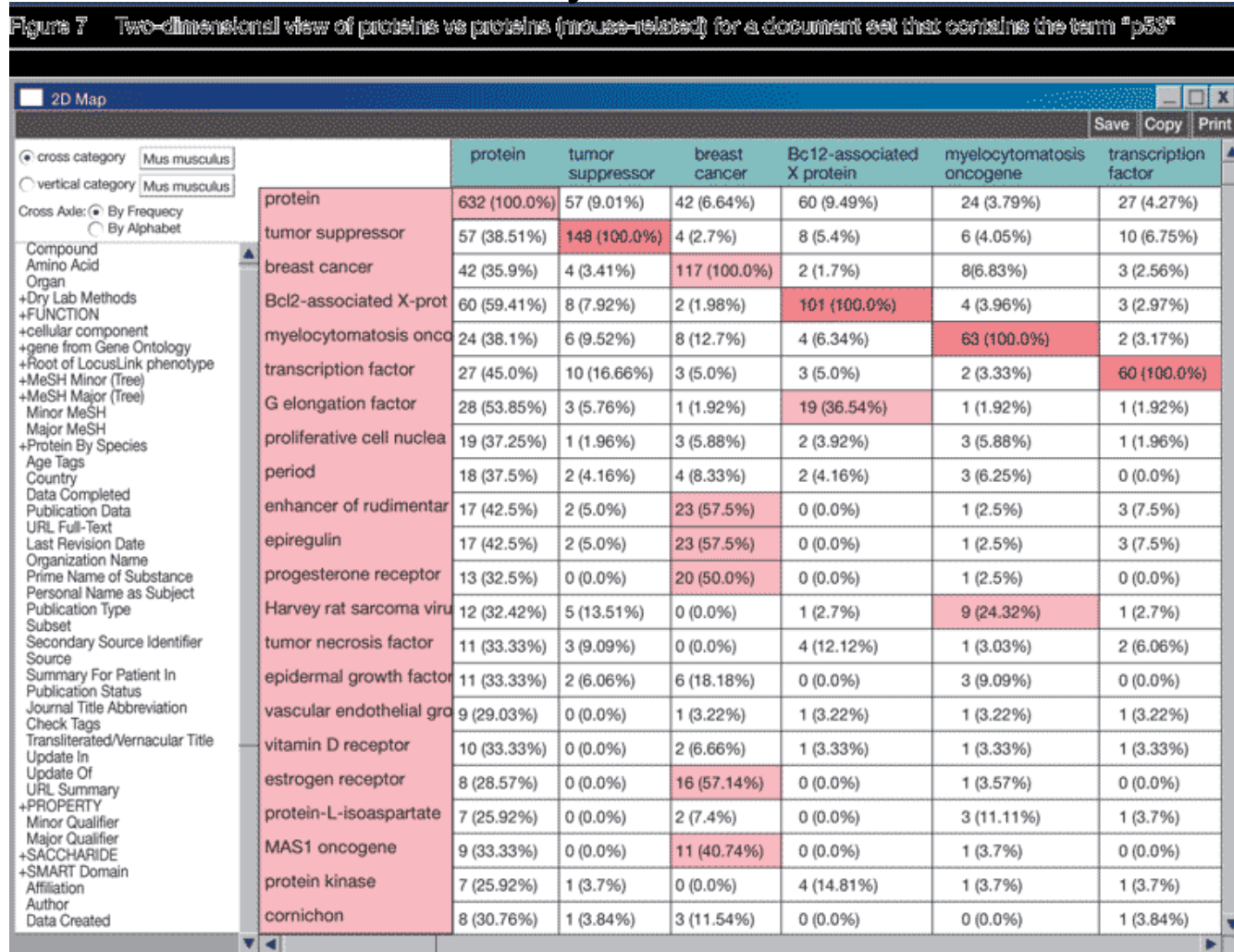
- An extension of the TAKMI (Text Analysis and Knowledge Mining) system originally developed for text mining in CRM applications.
- Goal is to extract relationships among biomedical entities (e.g. proteins and genes), from patterns such as “A inhibits B” and “A activates B,” where A and B represent specific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

Case study: IBM's MedTAKMI

Figure 1 MedTAKMI architecture



Case study: IBM's MedTAKMI



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

Text mining

Text mining's strengths are in...

- creating machine-exploitable models in/of information stores that were previously resistant to machine understanding, turning human communications into data,
- exploiting discovered or predefined structures to detect patterns: categories, linkages, etc., and
- applying the derived patterns to classify and support other automated processing according to document-extracted concepts and to establish relationships.

Grand challenge

There's a tradition in many scientific fields of creating Grand Challenges that set a research agenda for years to come. www.darpa.mil/grandchallenge/



“The Defense Advanced Research Projects Agency (DARPA) will hold its third Grand Challenge competition on November 3, 2007.

“The DARPA Urban Challenge features autonomous ground vehicles conducting simulated military supply missions in a mock urban area. Safe operation in traffic is essential to U.S. military plans to use autonomous ground vehicles to conduct important missions.”

Grand challenge

Ronen Feldman's Grand Challenge:

“Text mining systems that will be able to pass standard reading comprehension tests such as SAT, GRE, GMAT, etc.”

Entails improved entity recognition and relation extraction, 98+⁰% precision and 95+⁰% recall.

Should work in any domain; totally autonomous and require no human intervention.

Analyze huge corpuses and come up with “truly interesting findings.”

Grand challenge

One might add to the list the need to –

Deal with real-world information sources and conditions: the ability to mine noisy materials such as call-center notes, survey responses, e-mail, etc.

Assess and weigh the correctness of responses and formulate a single, contextually best answer.

Grand challenge

The Turing Test is a conversation that could serve as the basis for an even better grand challenge. Note NLP interpretation points –

Speaker: *Intention, Generation, Synthesis*

Hearer: *Perception, Analysis, Disambiguation, Incorporation*



www.turing.org.uk/turing/scrapbook/run.html

A conversation

Text analytics – integrated analytics – should present findings in a way that suits the information and the user.

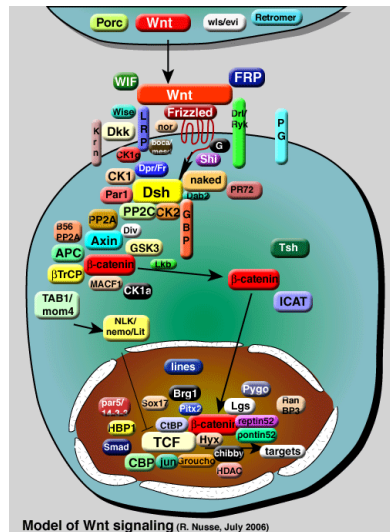
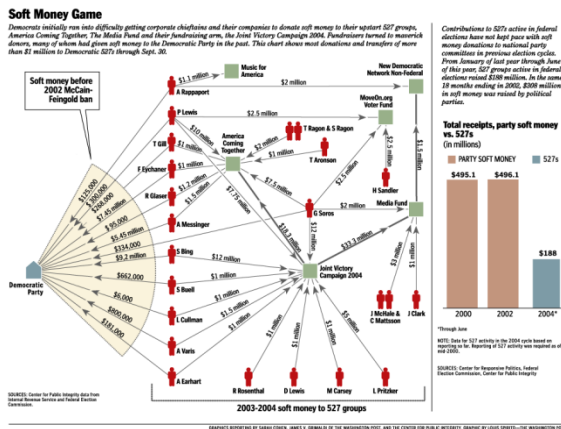
Searchers don't really want hit lists but rather answers to questions.



A conversation

What do business analysts and researchers really want, that is, by way of presentation of information?

Integration, domain awareness, usefulness!



Text analytics

We'll come back later to the question how we can do more analytics – how can we apply familiar BI techniques – to knowledge derived / extracted from text.

A last thought for now:

Entities and concepts are analogous to dimensions in a standard BI model. Both classes of object are hierarchically organized and have attributes.

Morning Agenda

Information as an enterprise asset

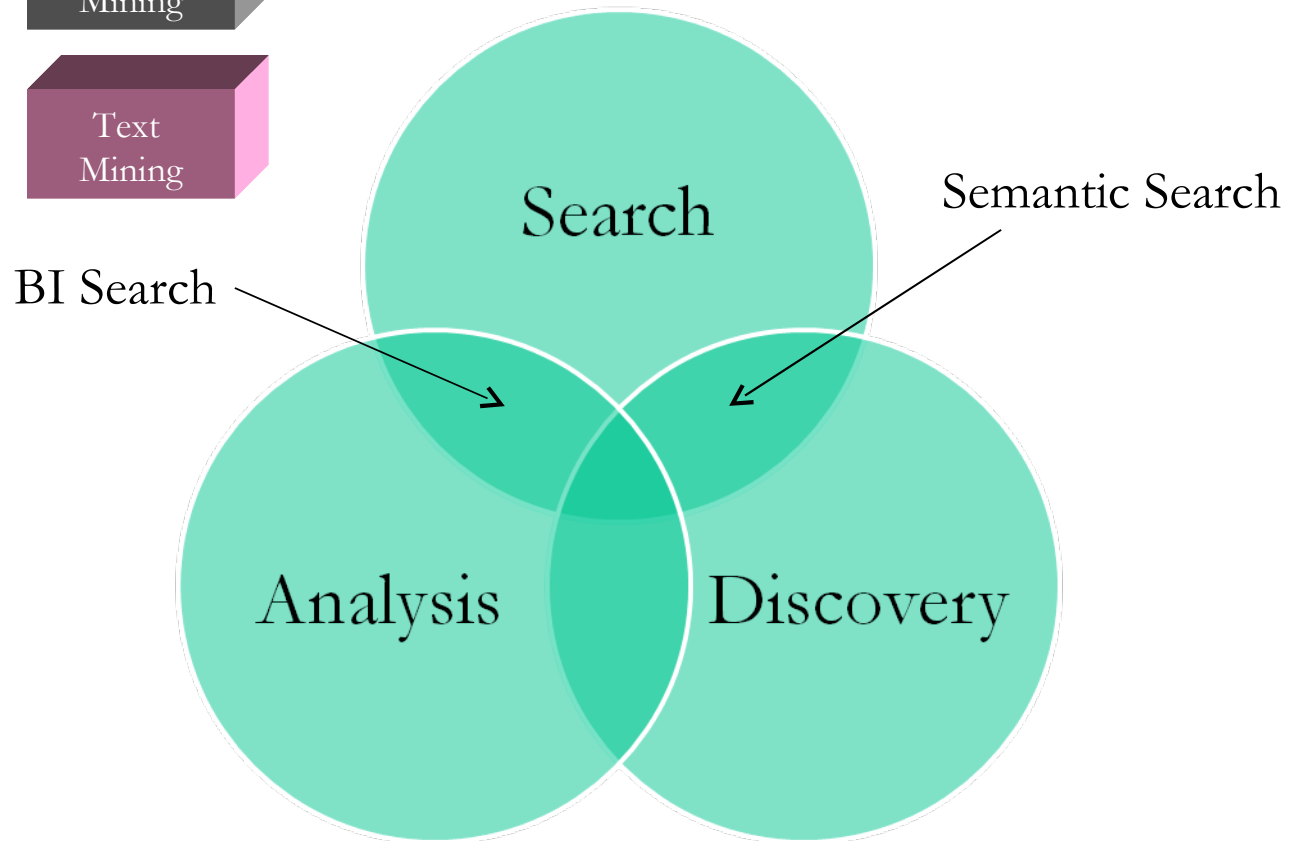
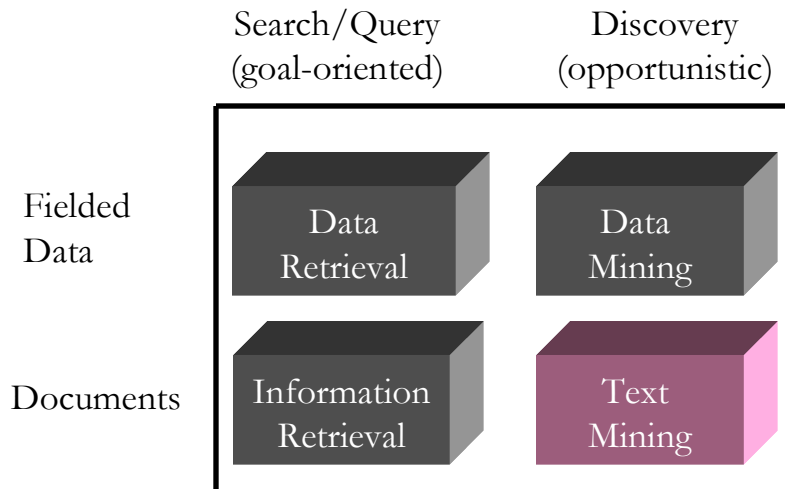
Discerning structure; extracting information

Semantic search, Search-BI

Improving search

Tapping search for BI

Improving search



Improving search

How can we improve search?

Allow more expressive queries.

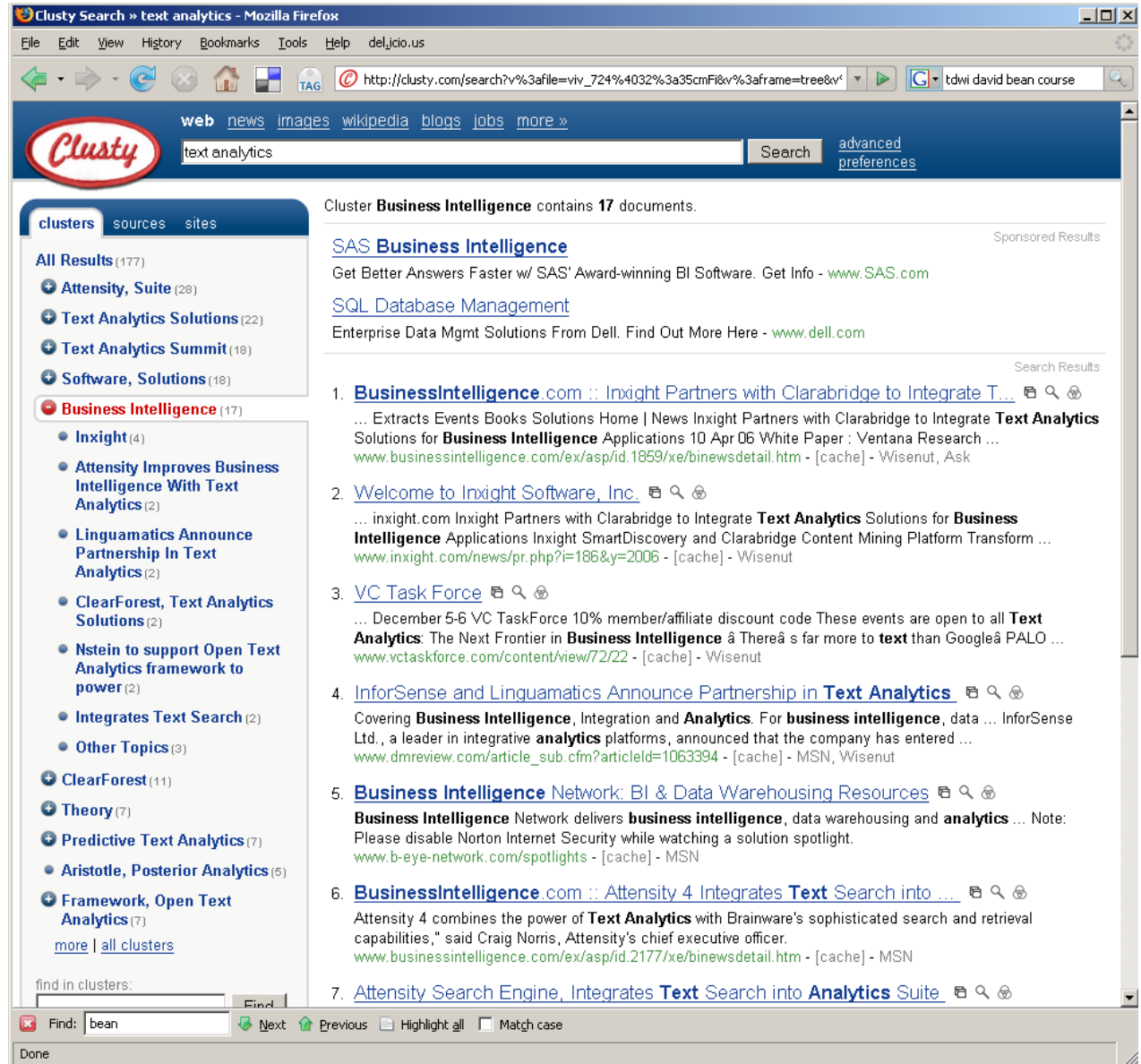
Improve relevance and usability of results .

Widen access to new information sources and formats.

Let's tackle results first.

Clustering using categories dynamically generated from a couple of hundred search hits...

Clustered
metasearch
results via
Clusty.com:
useful, but
ugh!



Dynamic,
clustered
search
results
from
Grokker

...

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

...with a zoomable display

The screenshot shows the Grokker Enterprise Search Management interface in Mozilla Firefox. The browser address bar shows the URL: `http://live.grokker.com/grokker.html?query=text%20analytics&Yahoo=true&Wikipedia=true&numResults=250`. The page title is "Grokker - Enterprise Search Management - Mozilla Firefox".

The interface features a search bar with the query "text analytics" and a "GROK" search button. Below the search bar, there are navigation tabs for "Outline View" and "Map View", with "Map View" selected. The map view displays 145 total results in a circular, zoomable layout. A tooltip is visible over a result, showing the following information:

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!

The map view also includes a "Zoom Back" button and a "TOP" button. On the right side, there is a "Detail" section with links for "Less", "Medium", and "More". The detail section shows the following information:

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/> - 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

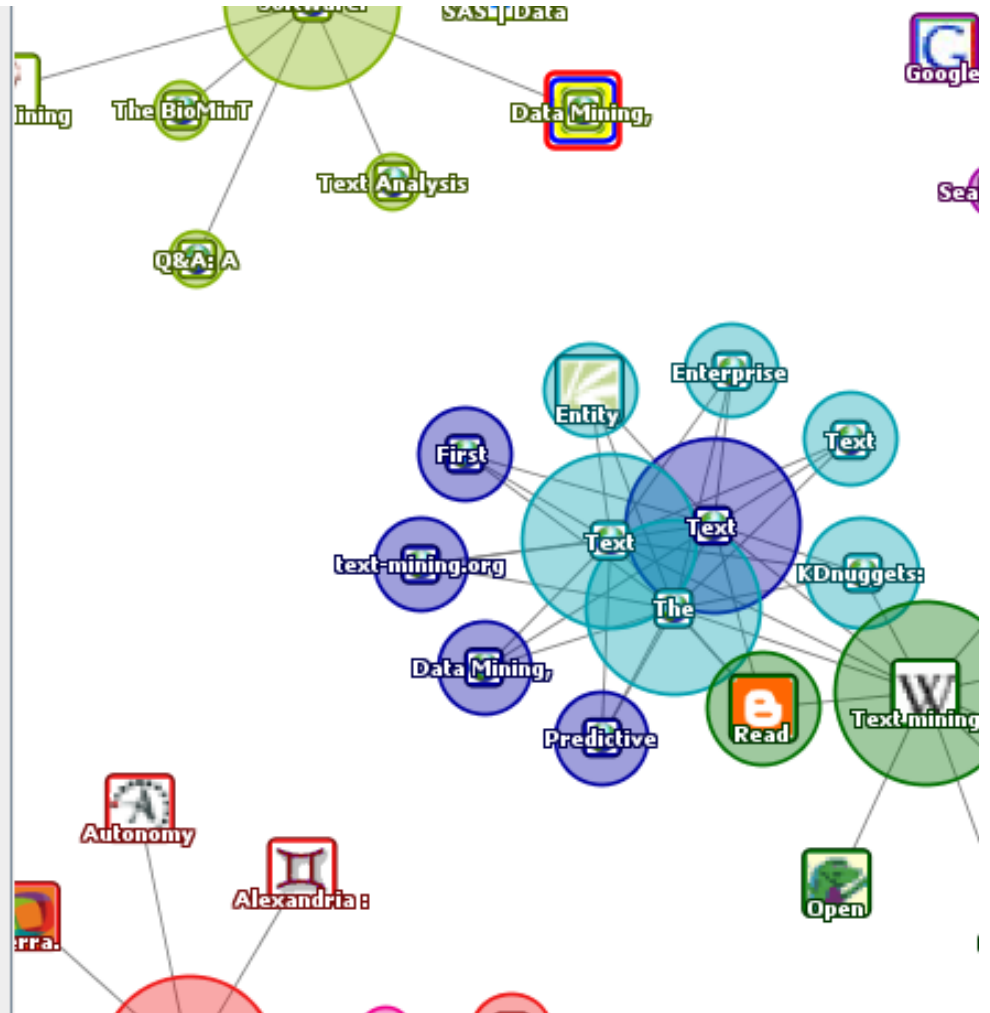
Text Mining, Text Analysis, Unstructured Data, Document Classification, Customer Churn
Add to Working List | Post to del.icio.us | Bookmark | Email
Specializing in text analysis, unstructured data, document classification and ... Read about other how organizations use Predictive Text Analytics here.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

At the bottom of the interface, there is a search bar with the text "Find: bean" and buttons for "Next", "Previous", "Highlight all", and "Match case". The footer of the page includes the copyright notice "©2006 Groxis Inc., All Rights Reserved." and the URL `http://live.grokker.com/grokker.html?query=text analytics&Yahoo=true&Wikipedia=true&numResults=250`.

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
+	W Text analytics - Wikipedia,...	en.wikipedia...	10
+	The New York Times - Br...	nytimes.com	1
+	Slashdot: News for nerds...	slashdot.org	1
+	IMDb 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



Improved search results

A variety of techniques make graphical complexity manageable...

- Zooming.
- Partitioning.
- “Fisheye” or hyperbolic: context-preserving non-linear display.
- Hiding/collapsing.
- Dimensionality reduction.
- Filtering.

Improved search results

To “make the connection,” we exploit

- Top-down (predetermined) or bottom-up (extracted, discovered) structure.
- Display options (layouts) that suit the data and the analytical goals.
- Dynamic, interactive data exploration.
- Complexity management.
- Embedded analytical algorithms.

Expressive queries

We've just seen screens of post-search clustering.
What if we could search on concepts rather than just term expressions?

Concepts may employ a knowledge classification, that is, a taxonomy.

A search on “cars” might return “Ford” indexed items.

What if we could use relational concepts in query expressions?

Larger/smaller, expensive/cheap.

New sources

Natural language processing gives the possibility of inferring concepts from search *queries*.

Explicitly: searches become queries.

Consider a search on Ford Mustang for semantics.

Each word helps disambiguate the other.

What if we could use search to access BI objects?

What's a BI object?

Hint: What are typical BI interfaces and data-delivery mechanisms?

How do you normally get to existing BI objects?

New sources

What's a BI object? In order of fixedness to dynamic potential –

Document.

Report.

Spreadsheet.

Pivot table/cube: the cross-product of hierarchical dimensions.

Database.

New sources

How do you normally get to existing BI objects?

Catalogs.

Portals.

Menu systems.

It would be nice to identify reports, pivot tables, and database tables with search on...

Dimensions/variables.

Value labels.

Data values.

Search-BI

Leading vendors have announced initiatives in the last year:

Cognos Go!, Business Objects, Information Builders, SAS, others: all have strategy to provide searchability via enterprise search tools from Autonomy, FAST &/ Google.

FAST Adaptive Information Warehouse (AIW) would even attempt to obviate traditional BI.

Data Cleansing Solution for integrated access.

(Corporate) Radar dashboards, charts, scorecards.

Search-BI

Earlier I said that search and information retrieval are not enough...

FAST AIW takes us in the right direction, integrated analytics –

Text and numerical sources.

A start at application embedding.

Morning Agenda

Information as an enterprise asset

Discerning structure; extracting information

Semantic search, Search-BI

Questions?

Discussion?

Afternoon Agenda

Text-data integration

Business applications

Market survey

Implementation

IT integration

Why integrate analytics?

360° views.

Single version of the truth.

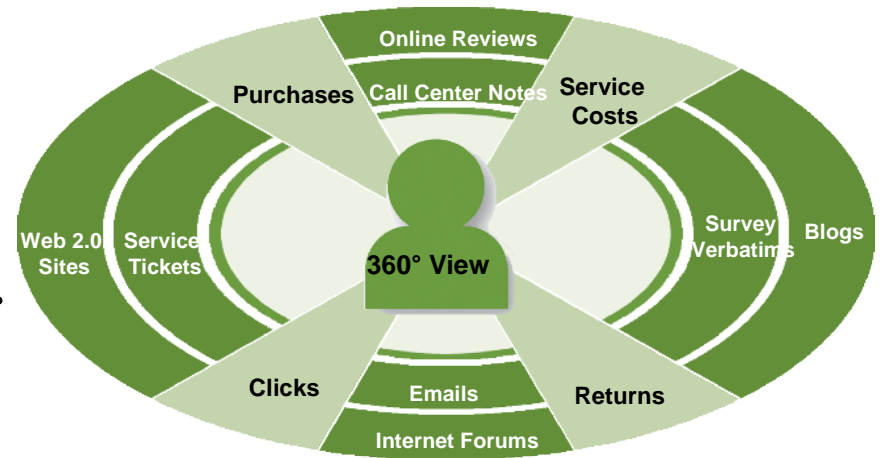
Discussion questions:

What's interoperability?

What's integration?

What's federation?

How/what can you integrate?



Clarabridge's version: text + data

IT integration

My definitions:

Interoperability is making systems work together toward a larger goal.

Integration goes beyond interoperation to make the “facilities” of one system available to another.

Integration is not necessarily bi-directional.

With federation, a controlling process distributes work across multiple providers.

The integrated whole must be better in some way than the sum of the parts.

IT integration

How/what can you integrate?

Components, via some form of API or framework.

Data, via defined, commonly understood formats and meanings.

What's the latter form of integration called?

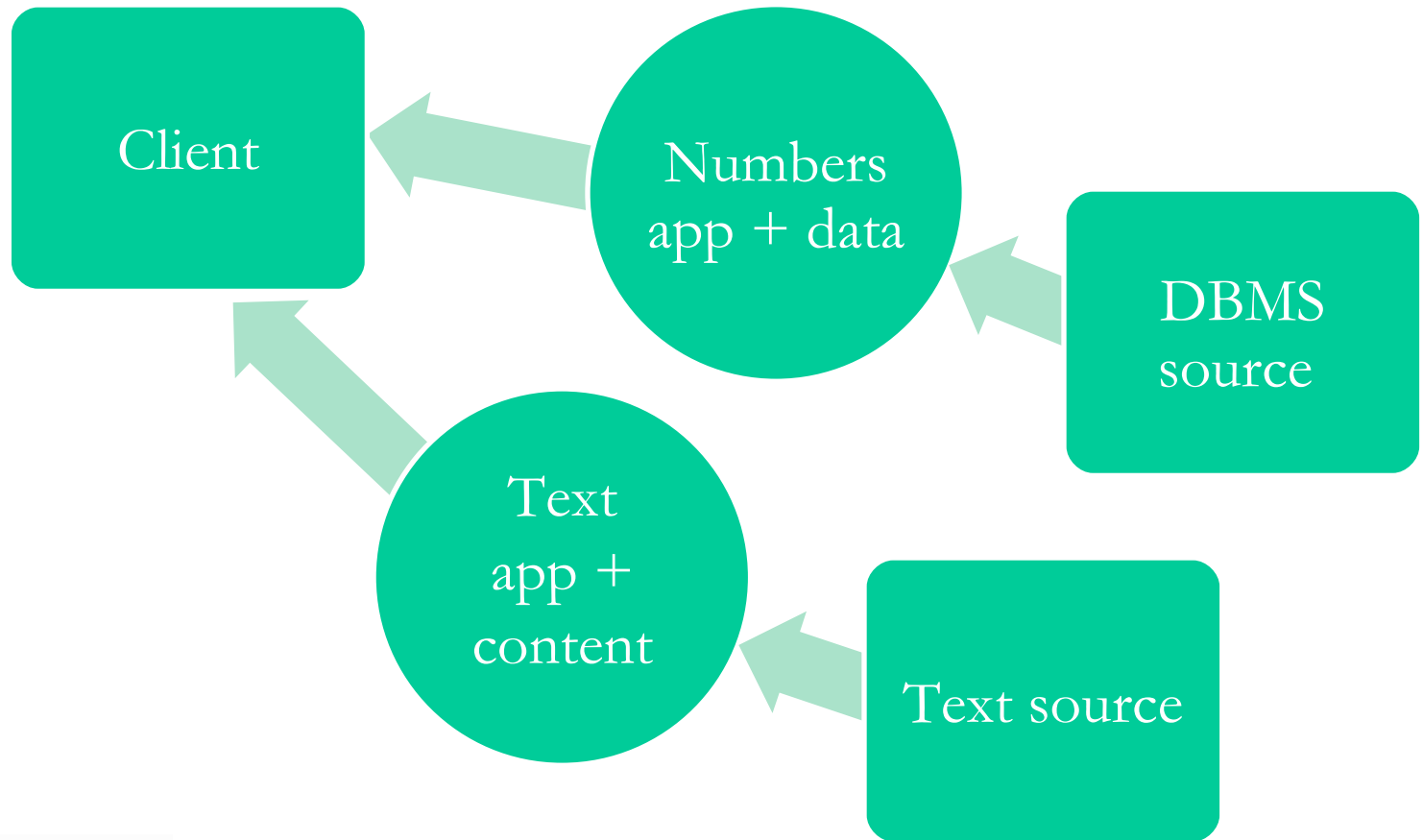
Business processes.

Other resources including project teams.

Standards play a major role.

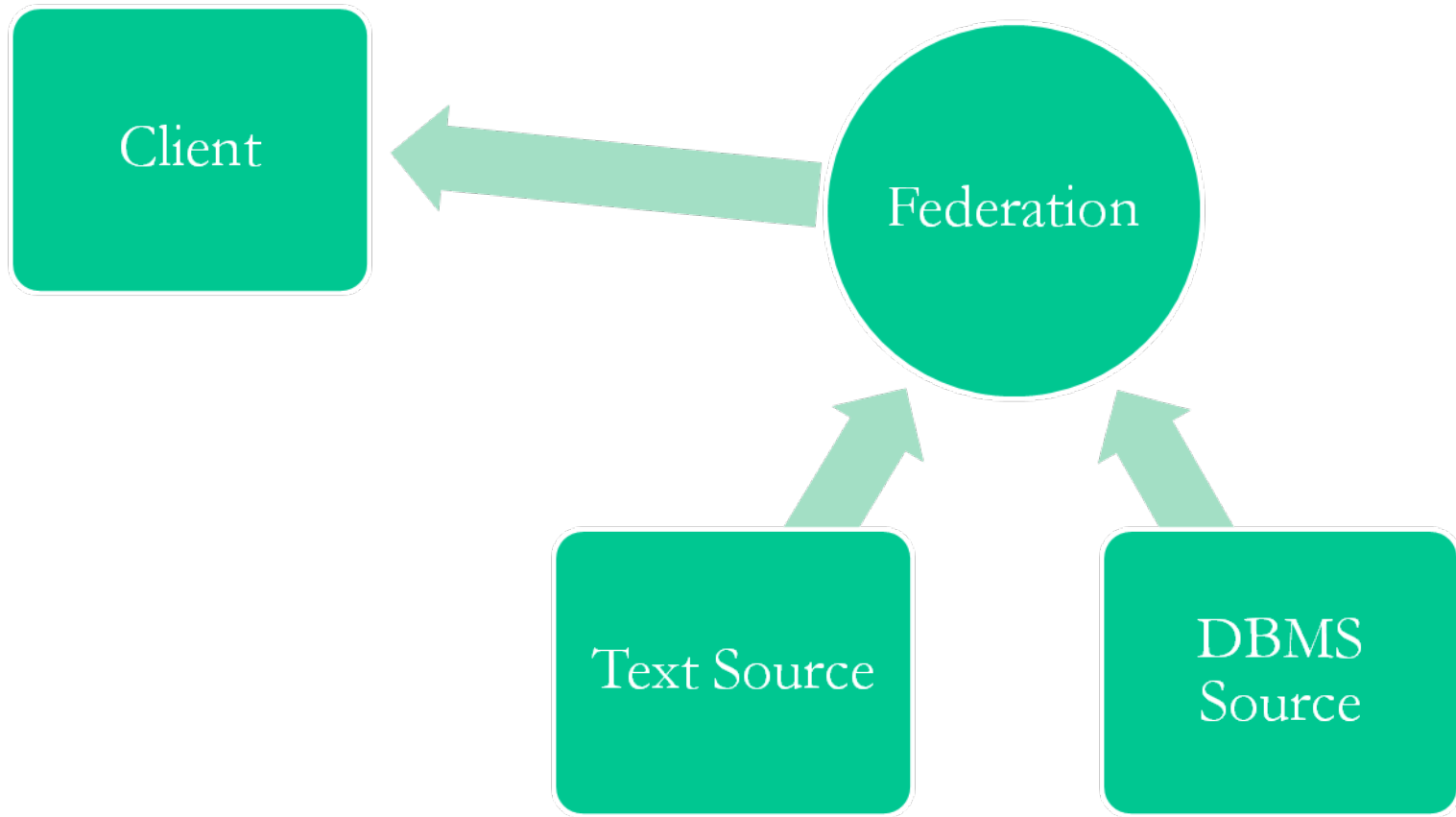
Integration models

Unintegrated applications: not of interest here.



Integration models

Integrated, federated application



IT integration

Object-relational databases can provide a variety of federation.

Single queries can tap both conventional data and free text that is stored as a smart binary object.

Smart = text-aware functions, indexing are built in.

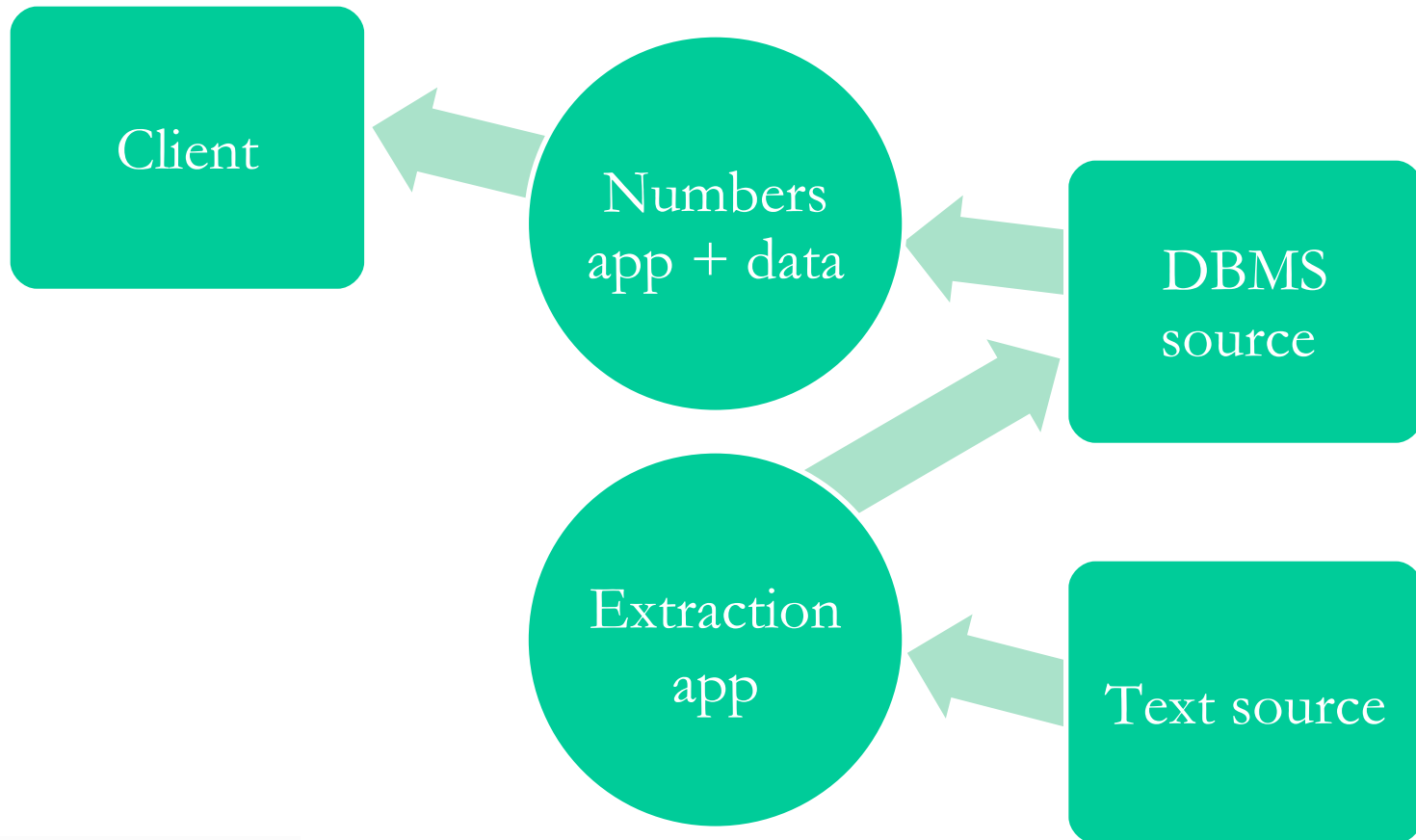
Here we necessarily have some degree of semantic integration, namely –

- the text is held in some field

- the text co-occurs with the other data in each record, that is, text and data are linked.

Integration models

Information extraction and loading.



IT integration

In information extraction, we do –

Information retrieval, that is, locate source documents of interest.

Identify relevant entities, concepts, and relationships.

Extract them to appropriate DBMS structures.

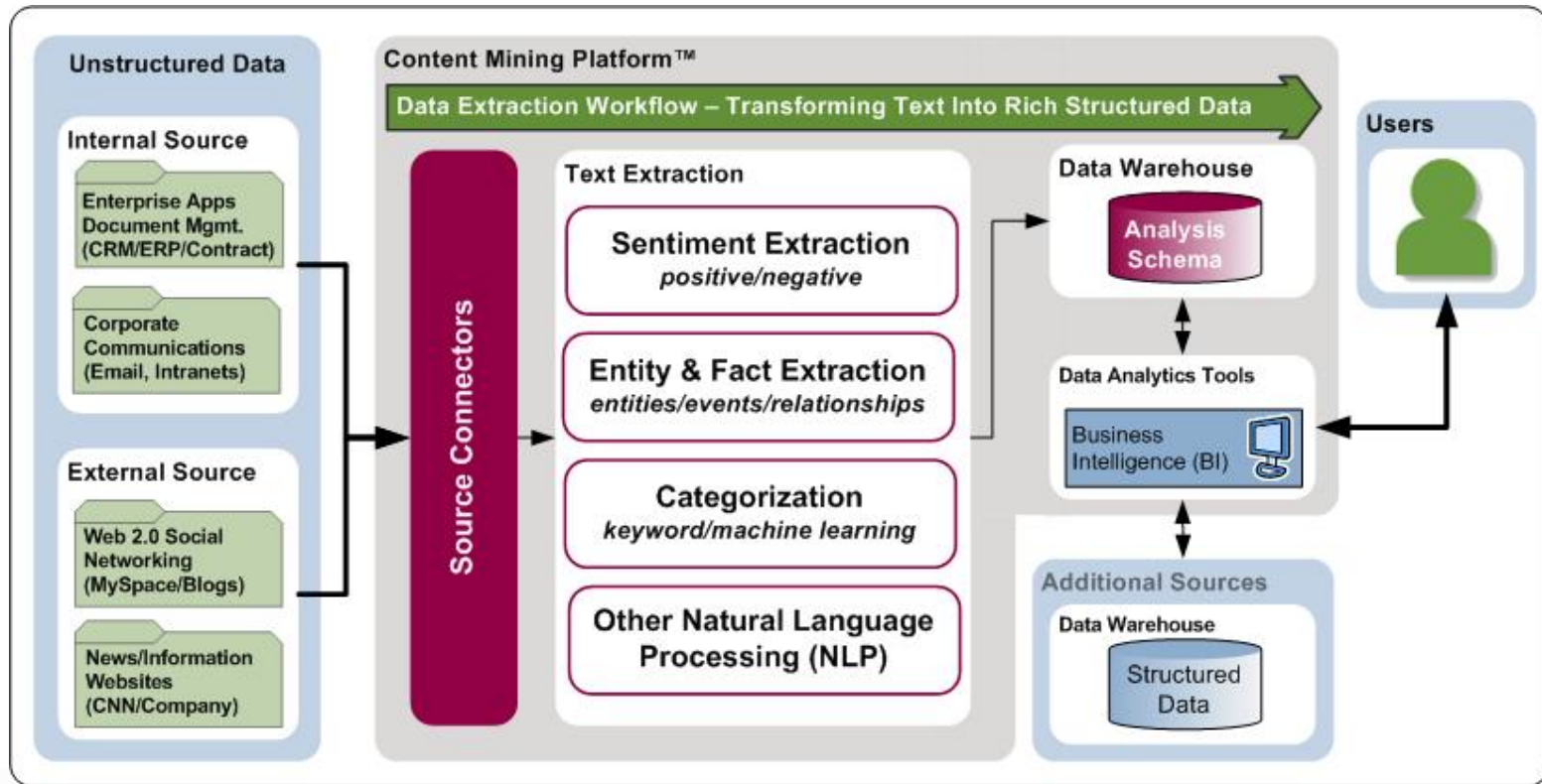
Here, again we necessarily have semantic integration, namely –

the extracted information is held in a field

the extracted information co-occurs with the other data in each record, that is, EI and data are linked.

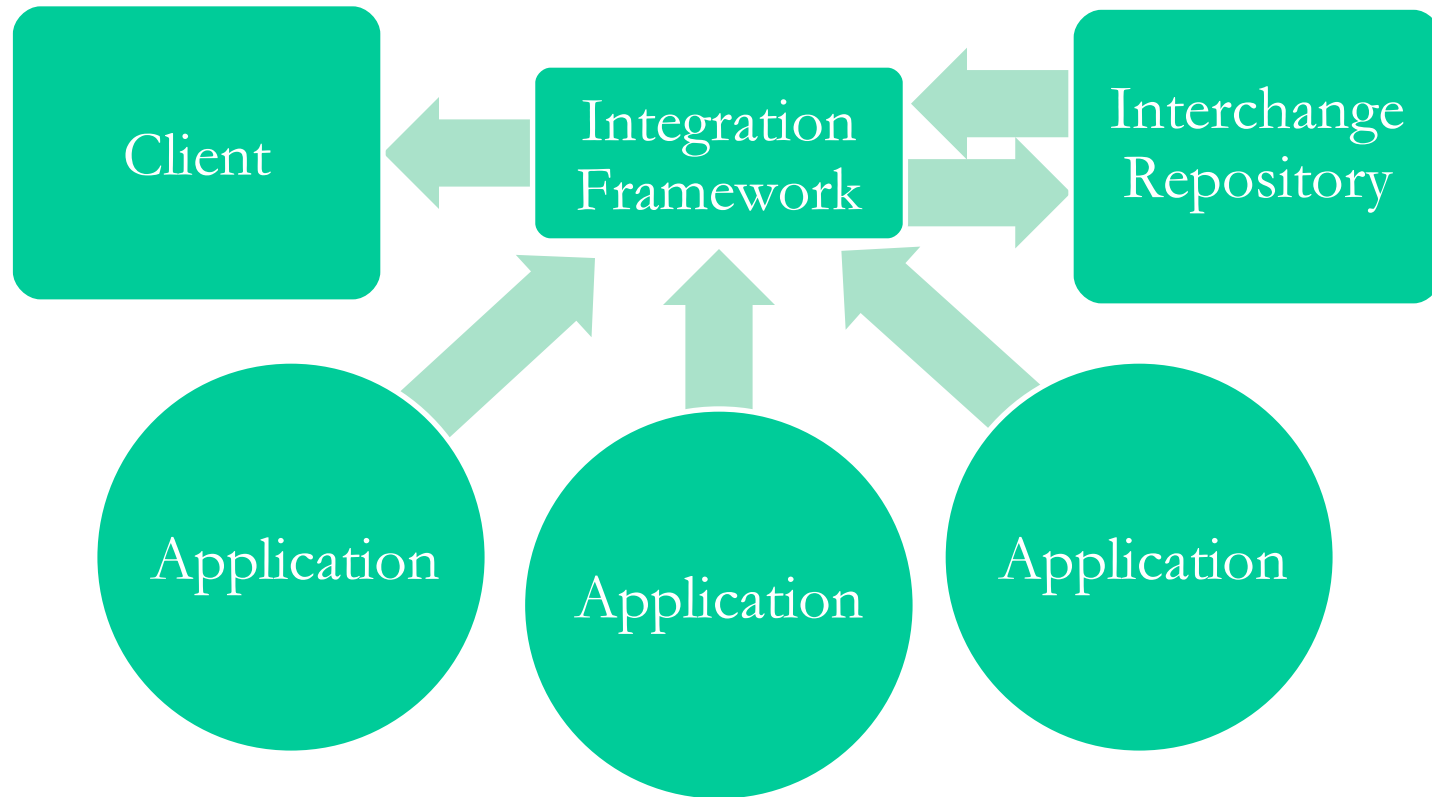
IT integration

Clarabridge's Content Mining Platform implements this architecture –



Integration models

Framework integration.



UIMA

The Unstructured Information Management Architecture is an integration framework created by IBM, then released to open source.

UIMA is an architectural and software framework that supports creation, discovery, composition, and deployment of a broad range of analysis capabilities and the linking of them to structured information services, such as databases or search engines. The UIMA framework provides a run-time environment in which developers can plug in and run their UIMA component implementations, along with other independently-developed components, and with which they can build and deploy UIM applications. The framework is not specific to any IDE or platform.

UIMA

Analysis Engine (AE) building blocks.

Algorithms are packaged within components that are called Annotators.

Collection Processing Architecture for analysis of sets of documents.

Common Analysis Structure (CAS) for results representation and sharing.

CAS consumers may use results for search-engine indexing or extract them to an RDBMS.

UIMA

Resources –

Now an Apache Incubator project.

incubator.apache.org/uima/

Specification development now governed by an OASIS Technical Committee.

www.oasis-open.org/committees/uima/

IBM provides a UIMA Java SDK at AlphaWorks.

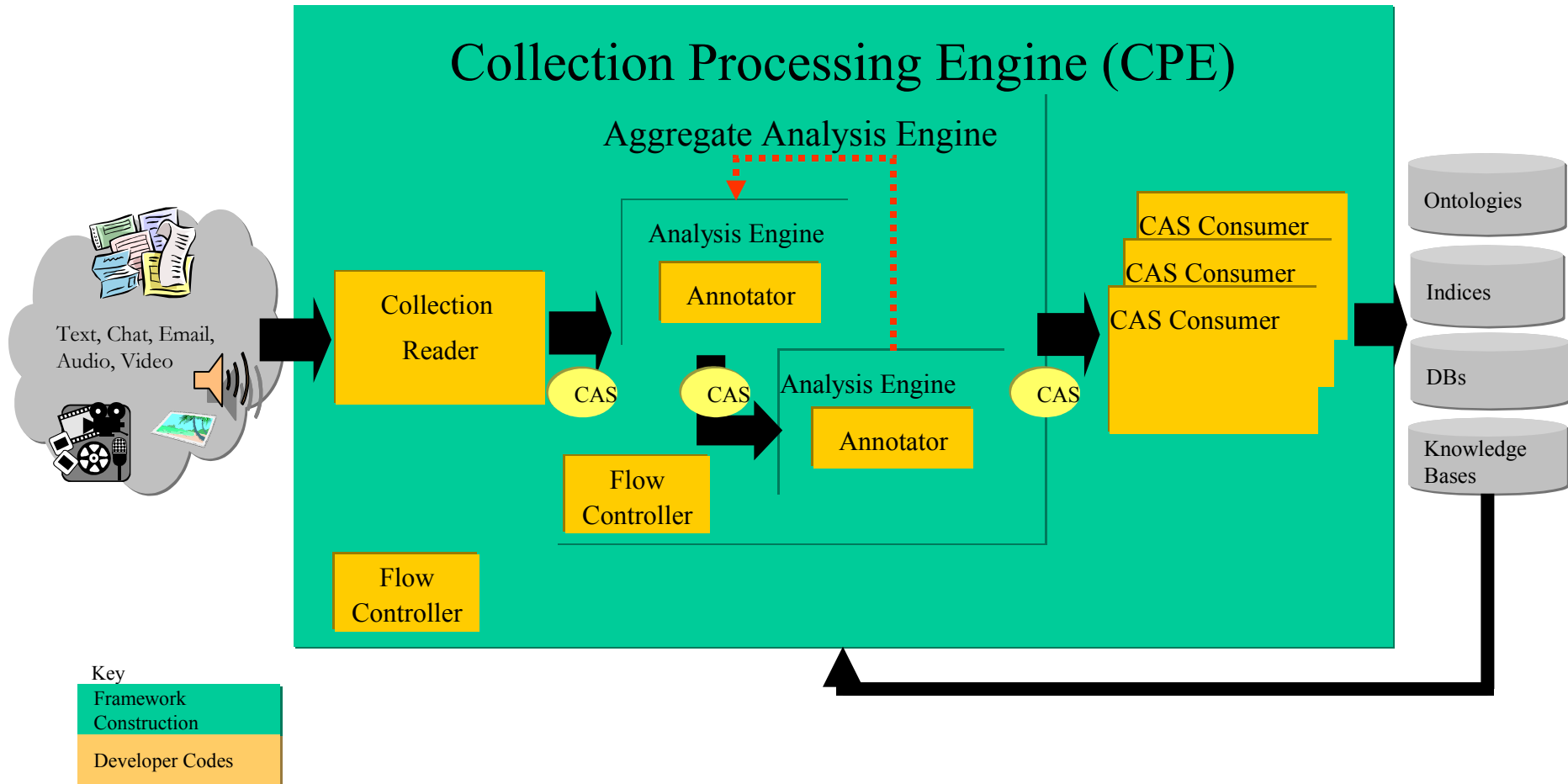
www.alphaworks.ibm.com/tech/uima/

GCC, C++, Python, Perl, TCL bindings.

CMU maintains a component repository.

uima.lti.cs.cmu.edu:8080/UCR/Welcome.do

UIMA



UIMA Component Architecture (from IBM)

UIMA

Commercialization –

IBM WebSphere Information Integrator OmniFind Edition is a platform for commercial use.

Nstein provides a dozen annotators.

Computer Business Review (Nov 2006) states –

BI and text analysis software vendors, including ClearForest, Cognos, SAS Institute, Factiva, and NStein Technologies, have endorsed UIMA in their products.

Commercially available UIMA-compliant products are currently freely available from IBM, Attensity, ClearForest, Temis and Nstein.

Afternoon Agenda

Text-data integration

Business applications

Survey analysis

Reputation management

Market survey

Implementation

Investigation vs. description


The MedTAKMI example was investigative in nature, trying to find a needle in a haystack.

This is a data mining function.

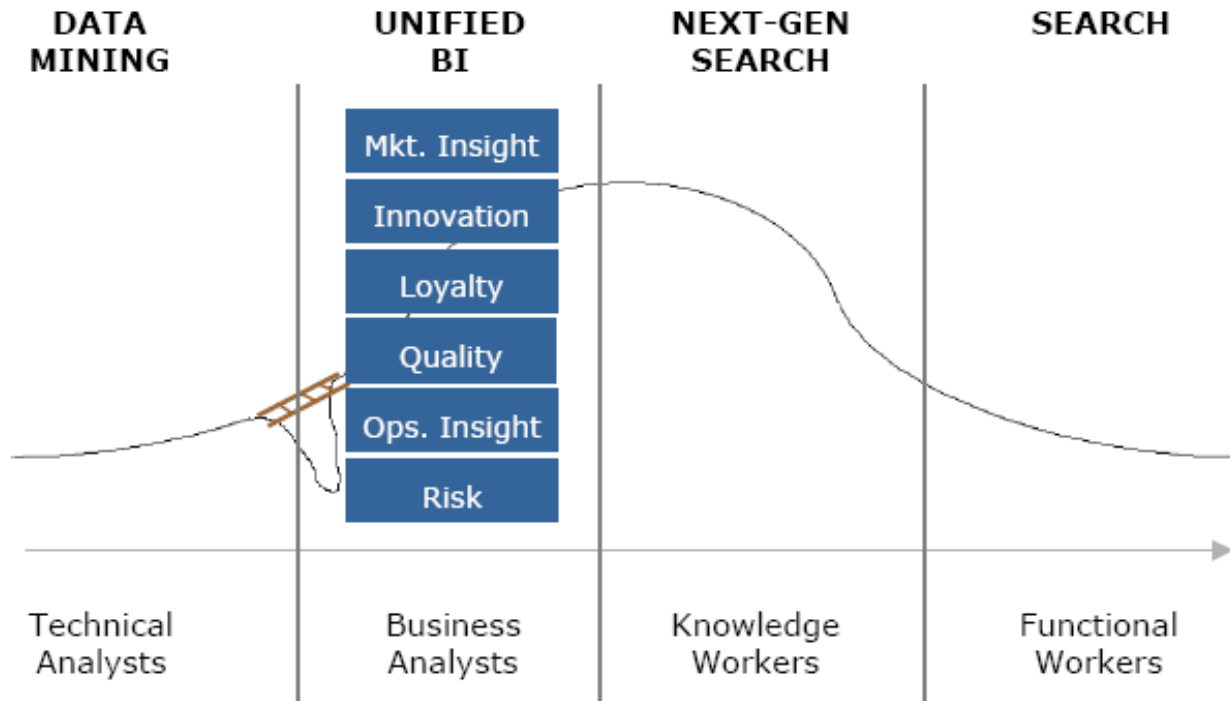
Many business users, by contrast, are interested in broad numbers and in trends.

This is a BI function.

The gap

 CLEARFOREST TEXT-DRIVEN BUSINESS INTELLIGENCE

Segmenting the Chasm



Survey

Customer Service Survey Form - Mozilla Firefox
 http://www.calepa.ca.gov/Customer/CSForm.asp

Who was the service provider?
 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.	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Staff provided complete, accurate information to you.	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
A timely response was provided.	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
My overall experience was positive.	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>

Please complete the section below if your contact with us involved permitting/licensing/registration assistance.

The regulations were understandable.	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
The application instructions were understandable.	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
The terms and conditions of the permit, license, or registration were understandable.	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>

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

Comments:

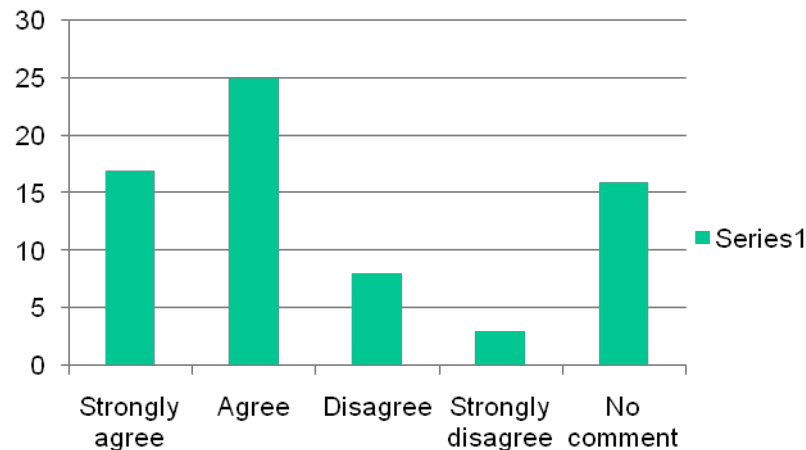
If you feel we fell short in meeting your service expectations, please describe the situation, including name of the staff person involved and the date the incident occurred:

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

Find: regarding Next Previous Highlight all Match case

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 also will be linked to coded ratings.

Survey

The respondent is invited to explain his/her attitude:

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

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:

Survey

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

Exploit what is structured in interpreting and using the free text.

Generally, textual source information doesn't come in without *some* form of envelope, of metadata that describes the information and its provenance.

It's still hard to automate interpretation of the free text, that is, to do more than count words and note cooccurrence. Sentiment extraction comes into play.

Sentiment extraction

Technology: sentiment (opinion) extraction –

The applications are:

Reputation management.

Competitive intelligence.

Quality improvement.

Trend spotting.

Sources include:

Wikis, blogs, forums, and newsgroups.

Product reviews on the Web.

Call-center notes.

Customer feedback via Web-site forms and e-mail.

Media.

Sentiment extraction

What's hard about extracting sentiment?

The way opinion is expressed, for instance explicitly –

Direct statements: I like it.

Subjective evaluative language: It is good.

-- and implicitly, involving sarcasm, irony, idiom and other deeper cultural referents –

It's really jumped the shark. (cultural referent)

It's great if you like dead batteries. (irony/sarcasm)

I give this two thumbs up.

www.kamalnigam.com/papers/metric-EAAT04.pdf

Sentiment extraction

Typical approaches involve:

- Domain-specific subject identification.

- Use of a lexicon and pattern database.

- Subject-sentiment association, typically via natural-language analysis.

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.

Correlation depends on semantic agreement: are we talking about the same thing?

Applications

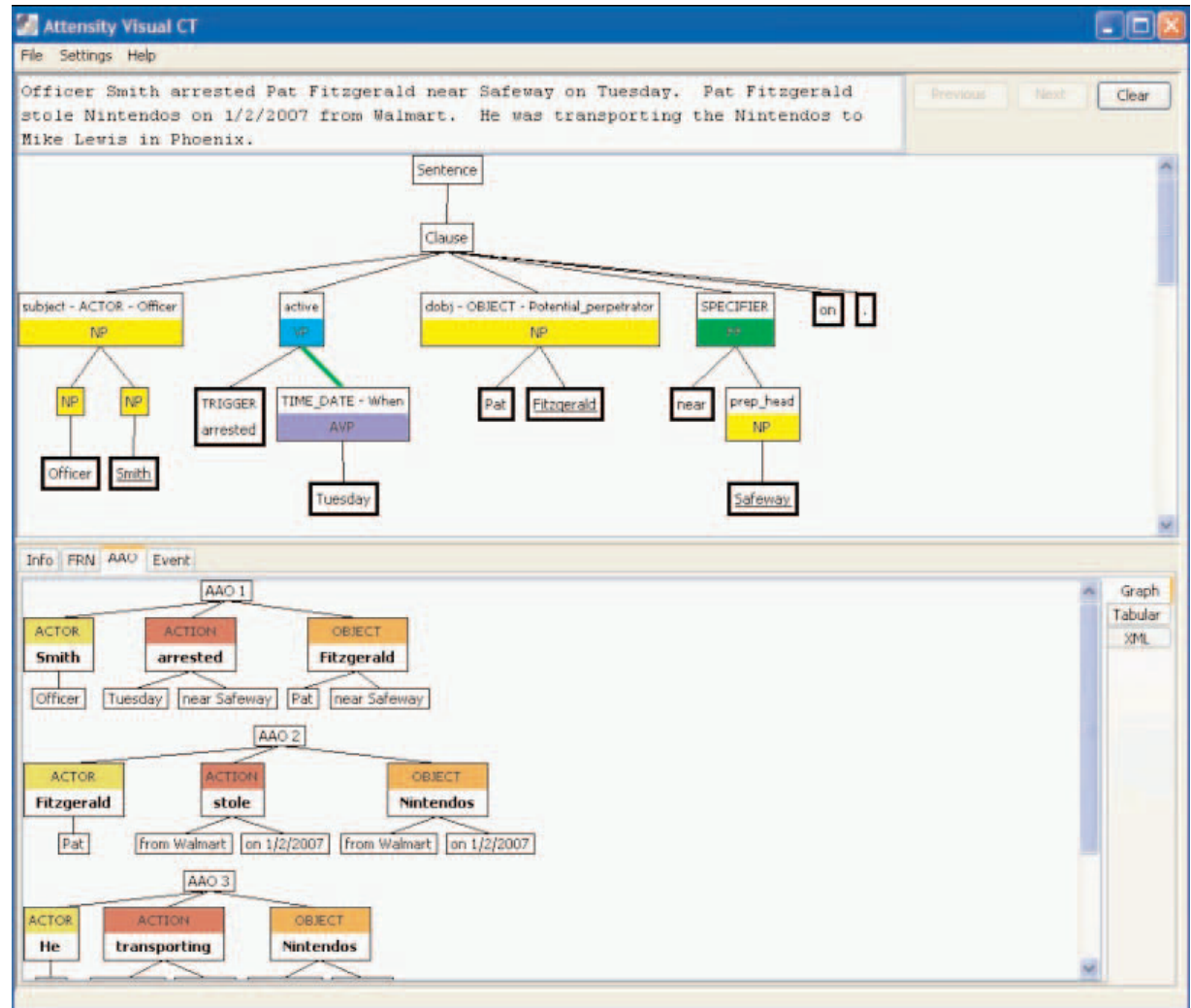
Take law enforcement as an example—

Sources: case files, crime reports, incident and victimization databases, legal documents

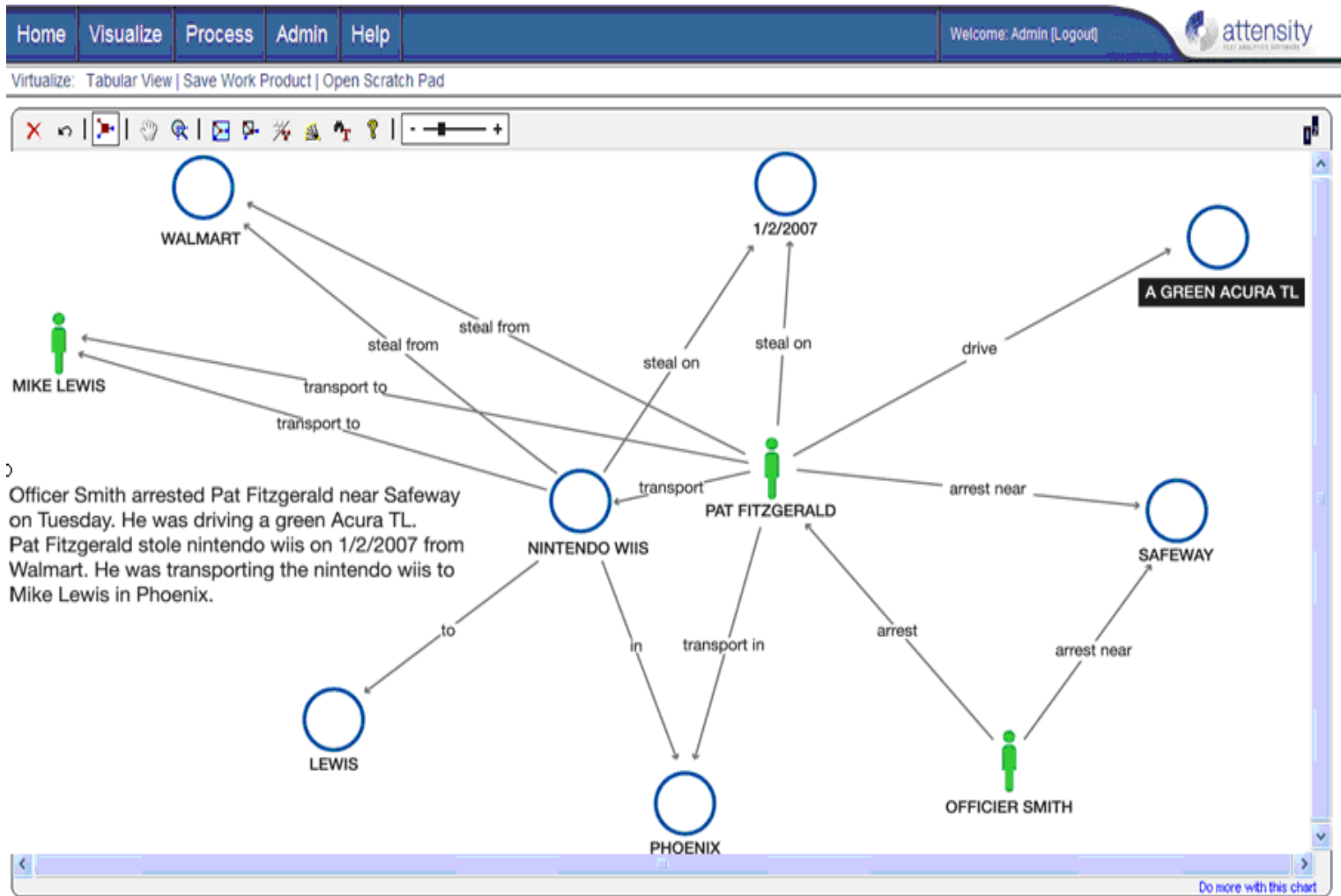
Targets: crime patterns, criminal investigation, networks

Applications

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



Applications



Other 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

Other 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

Afternoon Agenda

Text-data integration

Business applications

Market survey

Classification

Vendors

Implementation

Product positioning

Let's look at products according to a number of dimensions:

- Capabilities.
- Domain of application.
- Integration with other tools.
- Vendor background and strengths.

Product positioning

Decision factors may include:

- Language support.
- Entity understanding.
- Business-domain experience/expertise.
- Usage approach: Do you need a solution, an application, or software tools with an SDK?
- Applications integration with BI, CRM, SFA, and other systems.
- Implementation requirements including needed consulting.

Product positioning

Vendor origins/positioning:

- Enterprise search.
- Content/knowledge management.
- Pure-play text mining.
- Integrated analytics: text mining, data mining, statistics.
- Information provider.
- Open source.

The material on vendors that follows is not a recommendation.

and I do not have a paid relationship with any vendor named.

Enterprise search

Autonomy

- Strong with non-text media.
- IDOL (Intelligent Data Operating Layer) is core platform for processing “unstructured” information.

Convera

- Struggling.

Endeca

- Guided navigation.

Exalead

- One:enterprise product unifies access.

Enterprise search

FAST

- Adaptive Information Warehouse (AIW).
- Data Cleansing Solution
 - “Linguistic analysis and fuzzy matching to cleanse multiple disparate repositories into a cleansed master index.”
- Radar
 - Browser based BI framework complete with charting, dashboards and scorecards.

Enterprise search

Google Enterprise OneBox

- Search appliance.
- BI-vendor alliances with Business Objects, Cognos, Groxis, Hyperion, Information Builders, SAS.
- Text analytics alliance with Inxight.
 - Data mining.
 - Entity, concept, and relationship extraction.

Content and knowledge management

Selected vendors:

- EMC Documentum
- Interwoven
- Mark Logic
 - Search & discovery solution.

“Until the content has been integrated, its true value can't be realized. Consider the content sources below, which are likely in your content library or in content you've licensed...”



- Open Text



Text mining

Inxight

- Spin-off of Xerox Palo Alto Research Center (PARC).
- SmartDiscovery metadata management.
- Categorizer and summarizer.
- Linguistic analysis (LinguistX); 32 languages.
- Entity extraction (Thing Finder), and visualization.
- LinguistX is widely licensed: Business Objects, Factiva, Google, Oracle, QL2, SAS.

Text mining

TEMIS

- XeLDA multilingual linguistic engine and XTS terminology suite from Xerox Research Centre Europe.
 - Insight Discoverer Extractor.
 - Luxid analysis & visualization module.
 - Clustering and categorizing modules.
 - Skill cartridges for life sciences, competitive intelligence, publishing, and related areas.
- ... a profile very similar to Inxight's.

Text mining

Pure-play text-mining vendors

- NetOwl (SRA)
- Nstein
 - Provides UIMA fact, topic & sentiment annotators.
 - IBM alliance for OmniFind text-analytics enrichment.
- SAIC
 - TeraText (InQuirion acquisition), national security focus.

Vendors

Computational linguistics orientation

– Basis Technologies

- Rosette Linguistics Platform.
- Initial specialization in Asian languages, added Middle Eastern, then European via a TEMIS partnership.
- Lucene integration.

– Teragram

- Linguistics suite for European & Middle Eastern languages; limited set of Asian language tools.
- OEM linguistics provider.

– (Inxight)

Alta Plana

Vendors

Integrated analytics

– Attensity

- Business Objects Open Search Initiative (OSI); resell Crystal xCelcius.
- Cognos partnership and integration.
- IBM resells. Company works with former Ascential unit to extend ETL (WebSphere Data Integration Suite) and with DB2 unit to create a Unified Data Warehouse.
- Teradata resells. Optimized to run on Teradata platform.

Vendors

Integrated analytics

– Clarabridge

- Principals come from a BI background. Strategic partnerships with Business Objects, Cognos, Endeca, MicroStrategy, Oracle.
- Populates analysis-ready data warehouse structures.

Vendors

Integrated analytics

– ClearForest

- Business Objects Open Search Initiative (OSI).
- Cognos, Endeca, Information Builders partner.
- ClearPath BI methodology.
- Domain targeted packaged analytics.
- Interesting set of Semantic Web Services; supports mashups.

Data miners

IBM

- iPhrase (recent acquisition)
- OmniFind (enterprise search)

Insightful

Megaputer

SAS

* SPSS

- Text mining for Surveys

Media miners

Attentio, Market Sentinel, Nielsen Buzzmetrics,
Onalytica

- Sentiment extraction from social media.

Factiva

- News and business information.

TextWise

- Contextual advertising.

Open source

Carrot2 (results clustering).

GATE (information extraction).

- Provides a framework similar in ways to UIMA.

Apache Lucene (search-engine library).

R (data mining and statistical programming).

Weka (data mining).

Afternoon Agenda

Text-data integration

Business applications

Market survey

Implementation

Implementation agenda

We've talked about technology. Now let's address implementation

- Focus first on requirements, on business goals.
- Understand “best practices” including enterprise architecture plus your organization's work practices, business processes, and data.
- Categorize and then evaluate products.
- Implement.

Business goals and requirements

Business goals imply...

Functional and nonfunctional requirements...

And are met by a combination of

- Technological and other resources.
- Information inputs.
- Transformative business processes.
- Business-focused results presentation.

Business goals and requirements

Business goals and resources –

- What problems do you want to solve?
- What sources are being or can be tapped?
- What outcomes do you expect?

How will you measure results and ROI?

Best practices, etc.

You must understand –

- What the technology can do: applications.
- Industry “best practices”: we’ll look at case studies.
- Your own organization’s work practices.

Look for methodologies such as –

- ClearForest’s ClearPath.
- CRISP-DM 2, the Cross Industry Standard Process for Data Mining, spearheaded by SPSS and Teradata.
 - Currently undergoing modernization from version 1.

Enterprise architecture

Typically constructed with a layered set of reference models.

- E.g., for Federal EA, performance, business, service-component, data, and technical models.
- Systems are implemented in conformance with the models.
- The approaches include process focus and standards compliance.

The goal is to enhance performance, lower risk, control costs, allow for growth.

Integration approach

Integration questions include –

- Mechanisms for information & applications integration, e.g., with content-management systems, with desktop or server file systems, with e-mail systems, with the Web or intranets.
- Operational integration, that is, will capabilities be offered through existing or new user interfaces?
- Is an exclusive focus on text and similar “unstructured” information sufficient?
- How will you apply text mining in conjunction with BI and data mining techniques?

Business processes

On to processes...

- How must you alter information acquisition and management processes to access and exploit untapped sources?
 - Note digital rights management (DRM) implications.
- What will your staff do differently?
- What are operational support requirements?

Product evaluation

You'll want to –

- Review technology positioning, case studies, and vendor solutions.

- Create and execute an evaluation and assessment plan with short-listed products.

- Proof of concept is important.

A pilot can help.

- Reveal solution gaps.

- Provide good cost figures.

- Set expectations.

Product evaluation

One Enterprise Architecture principle is minimization of the number of vendors.

Does your data mining, BI, search, or content management vendor have a suitable solution (direct of OEM) or exploitable industry alliances?

Do candidate products offer a full set of solution components?

Search/information retrieval.

Text mining/information extraction.

Appropriate, integrated analytics.

Questions?

Discussion?

Thanks!