# The Word on Text Mining

Seth Grimes

Alta Plana Corporation

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

Portals, Collaboration, and Content Management

April 14, 2005

Alta Plana

# Introduction

Seth Grimes --

- Principal Consultant with Alta Plana Corporation.
- Contributing Editor and decision-support columist, Intelligent Enterprise magazine, *IntelligentEnterprise.com*.

This presentation …

- is available on-line at *altaplana.net/TheWord.pdf*.

# Agenda

Enterprise "imperatives"

Analytical responses

The statistical advantage

Text mining foundations

Integrated analytics

Text mining applications

Implementation roadmap

# Enterprise "imperatives"

Efficiency! Effectiveness! Profitability!

You know the drill:

- 360° views
- Single version of the truth
- One-to-one marketing
- 24/7

and the goals:

- customer acquisition & retention
- up-sell, cross-sell -- better, faster, cheaper

# Enterprise "imperatives"

These imperatives translate into a process focus, into --

- understanding and managing business processes
- aligning business process with goals

and a focus on numbers and analytical magic --

- measuring
- modeling, forecasting, and predicting
- optimizing

and a valuation of knowledge.

# Analytical responses

Laying the groundwork:

- Statistics!

Breakthrough:

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

Bread & butter analytics:

- Business Intelligence: OLAP, query & reporting

Revival:

- Advanced analytics: reintroduction of statistics

# The statistical advantage

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*

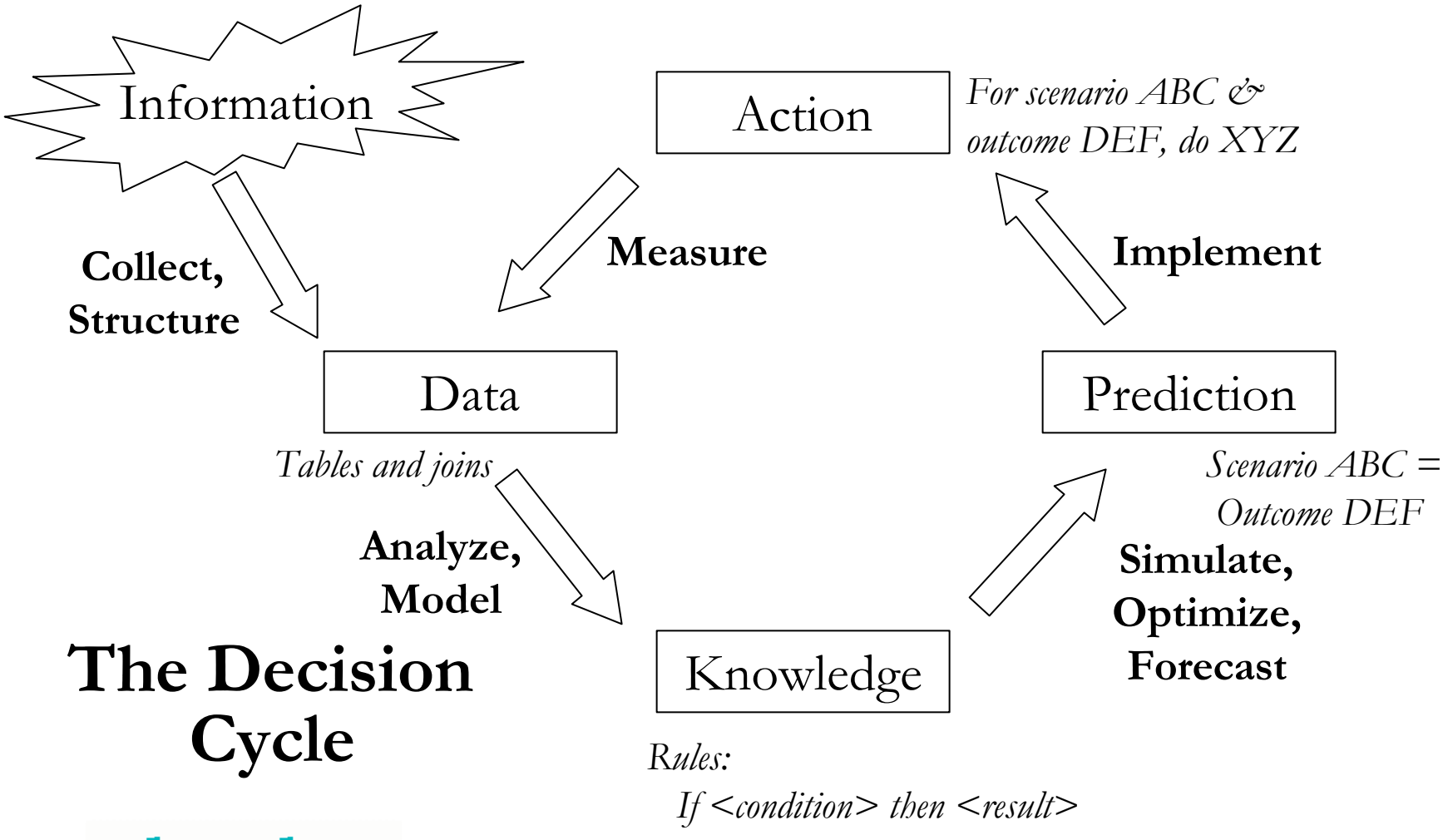
*Link analysis -- detecting association patterns and rules*

*Predictive modeling -- classification and scoring*

*Predictive modeling -- regression for anomaly detection and forecasting*

Predictive modeling completes the decision cycle.

# The statistical advantage




**The Decision Cycle**

Alta Plana

# Text mining foundations

We're in the midst of a information explosion:

- E-mail and other documents
- Images, video, and audio
- Tracking and geolocation data (RFID, GPS)
- Transactional data

Most of this information is "unstructured" or semi-structured -- 80% according to IDC -- and enterprises realize that valuable knowledge is locked in these forms.

Text mining is a variant of knowledge discovery.

# Text mining foundations

## Why isn't Search 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**.

## What about 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?*

# Text mining foundations

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 pattens 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 mining foundations

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 structured inherent in documents or the Web?

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

# Text mining foundations

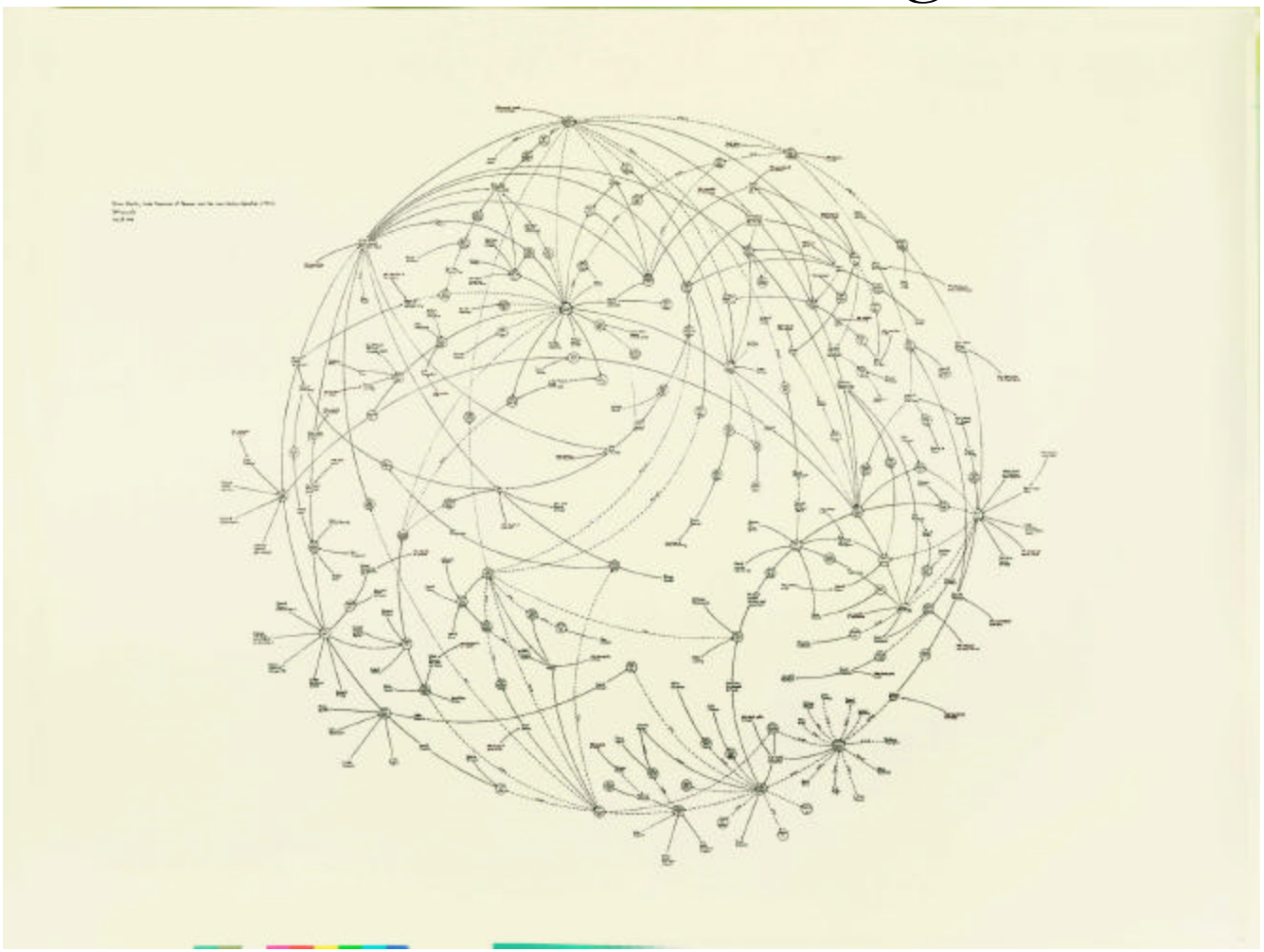
## The old way -- an illustration:

*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

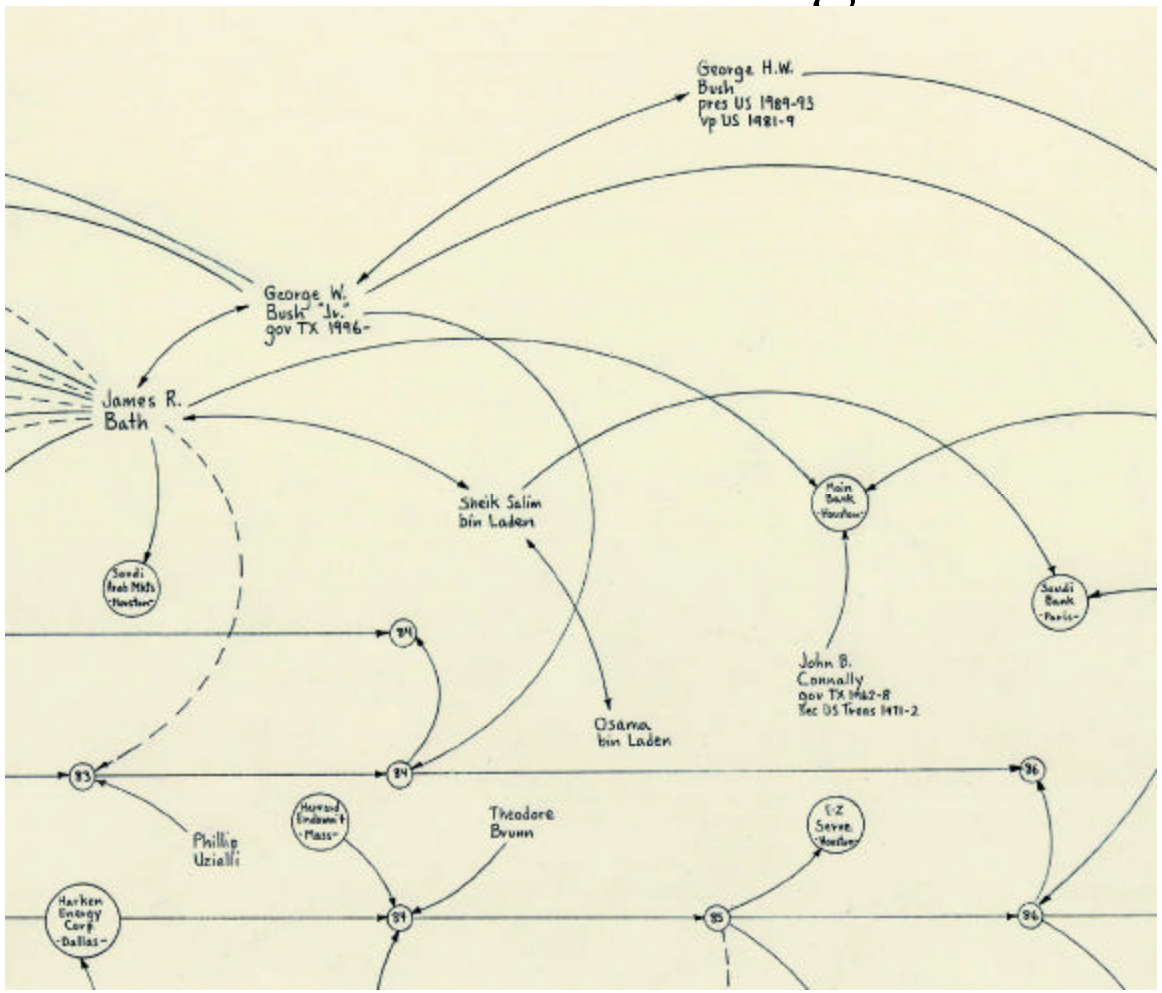
Images on next slides are courtesy of Pierogi, pierogi2000.com

# Text mining foundations



Mark Lombardi: Oliver North & Iran-Contra

# Text mining foundations



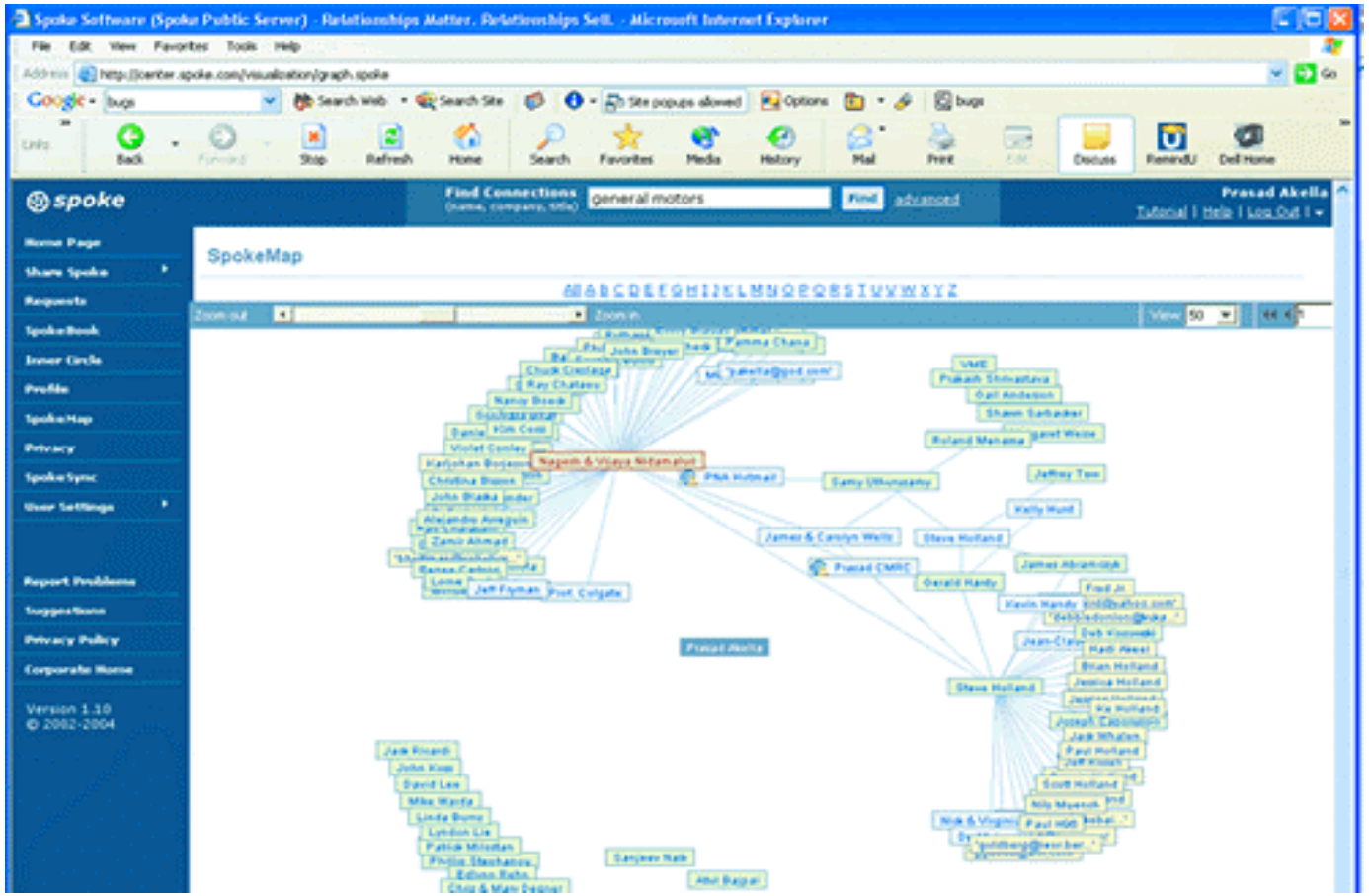

Mark Lombardi: george w. bush, harken energy... (detail)

Alta Plana

# Text mining foundations

Similar to what you see in a SpokeMap...

based on relation-ships mined from e-mail. (Visit spoke.-com)




Alta Plana

# Text mining foundations

Typical steps in text mining include --

- Apply statistical &/ linguistic &/ structural techniques to identify, tag, and extract entities, concepts, relationships, and events within document sets.
- Create a categorization/taxonomy from the extracts.
- Apply statistical techniques to classify documents.

Think about the Spoke example:

- E-mail addresses are readily identified and tagged.
- Relationships are inferred from frequency & type of exchange.

# Text mining foundations

## Entity extraction example (1):

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

# Text mining foundations

## Entity extraction example (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

# Text mining foundations

## Entity (& relationship) extraction example (3) :

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

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

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

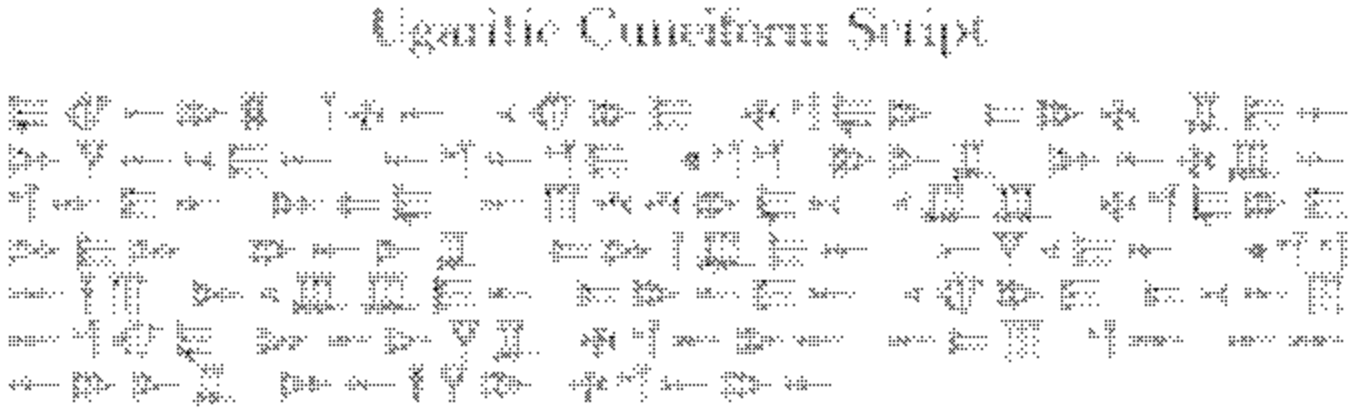
<name>adam </name>

# Text mining foundations

These e-mail messages might be termed "semi-structured." What about a document marked up with XML?

*Does Structure = Syntax? What about Meaning (Semantics)?*

What about this text? --



# Text mining foundations

Mid-way between tagging 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 mining foundations

Text mining's strengths are in…

- creating machine-exploitable models in/of information stores that were previously resistent 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.

# Integrated analytics

Is an exclusive focus on text and similar "unstructured" information sufficient?

Remember those imperatives...

- 360° views
- Single version of the truth
- Efficiency

Text mining can be applied in conjunction with business intelligence and data mining techniques.

Text mining can be integrated into operations.

# Text mining applications

## Case study: IBM's MedTAKMI (Text Analysis and Knowledge MIning for Biomedical Documents -- (drawn from ibm.com) --

- 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.
- 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.

# Text mining 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

# Text mining 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

# Text mining applications

## Law Enforcement

Sources: case files, crime reports, legal documents

Targets: crime patterns, criminal investigation

## The list of potential applications is long…

- Legal discovery and strategy
- Patents
- Recruitment
- Reputation Management
- Survey Analysis

# Implementation roadmap

Focus first on requirements:

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

Think next about enterprise architecture:

- Where is the information now, e.g., in Content Management systems, on desktop or server file systems, in e-mail systems, on the Web or intranet?
- How will you integrate text analytics into operations, that is, via what existing or new user interfaces?
- How will you measure results and ROI?

# Implementation roadmap

On to processes:

- How must you alter your information acquisition and management processes to access and exploit untapped sources?
- What will your staff do differently?
- What are operational support requirements?

Then review case studies and vendor solutions.

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

A pilot and a phased implementation help...

Questions?

Discussion?

Thanks!

# The Word on Text Mining

Seth Grimes

Alta Plana Corporation

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

Portals, Collaboration, and Content Management

April 14, 2005

Alta Plana