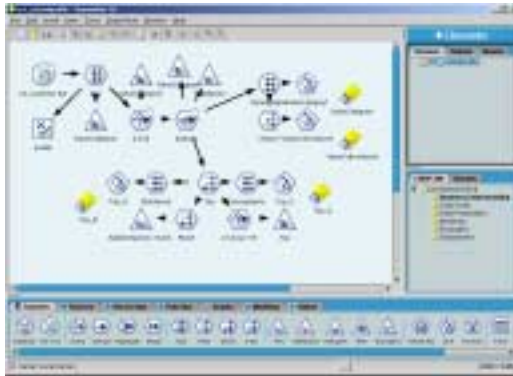




Data Mining and Text Mining

A marriage made in heaven!

Tom Khabaza



SPSS



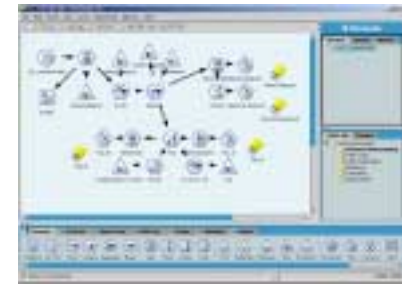
 **Clementine**



Data Mining & Text Mining

A marriage made in heaven

- ▼ Data mining & Clementine
- ▼ Text mining – A data miner's view
- ▼ Requirement for integration
- ▼ First steps towards integration
- ▼ A new integration
- ▼ A bright future





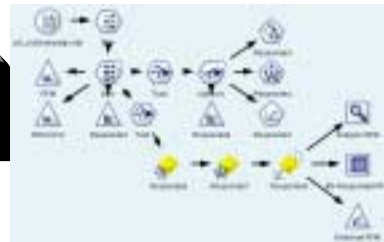
What is Data Mining?

Finding patterns in your data
which you can use
to do your business better

Critical
factors:



Business
Knowledge



Comprehensive
Facilities



Deployment
of Results



Applications of Data Mining

▼ Customer Relationship Management (CRM)



Who are our best customers?
Can we get more like that?
What/why do they buy?
Why do they leave?

▼ eCRM – Web-mining

How do they behave?



▼ Fraud detection



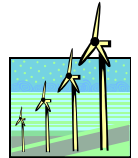
▼ Crime analysis



▼ Money laundering detection

▼ Network intrusion detection

▼ Wind turbine maintenance



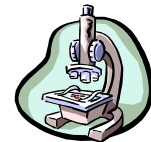
▼ Industrial process optimisation & QA

▼ Environmental management / conservation



▼ Drug discovery

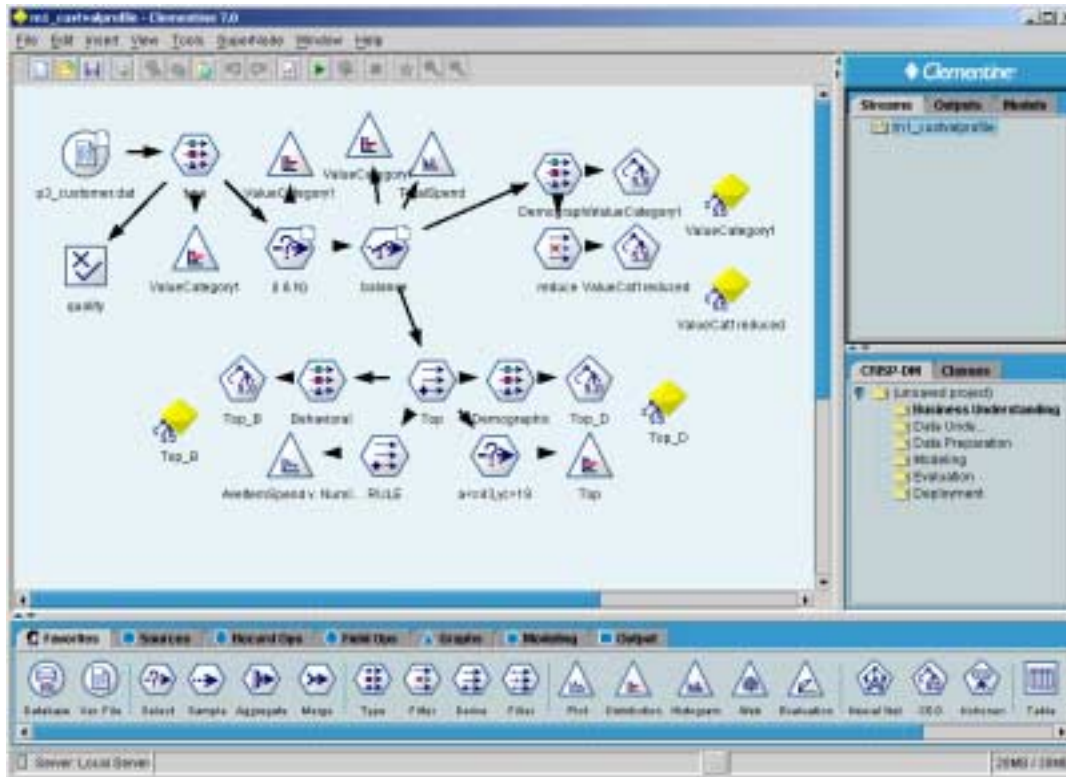
▼ Medical research



▼ Food authentication



Clementine data mining system

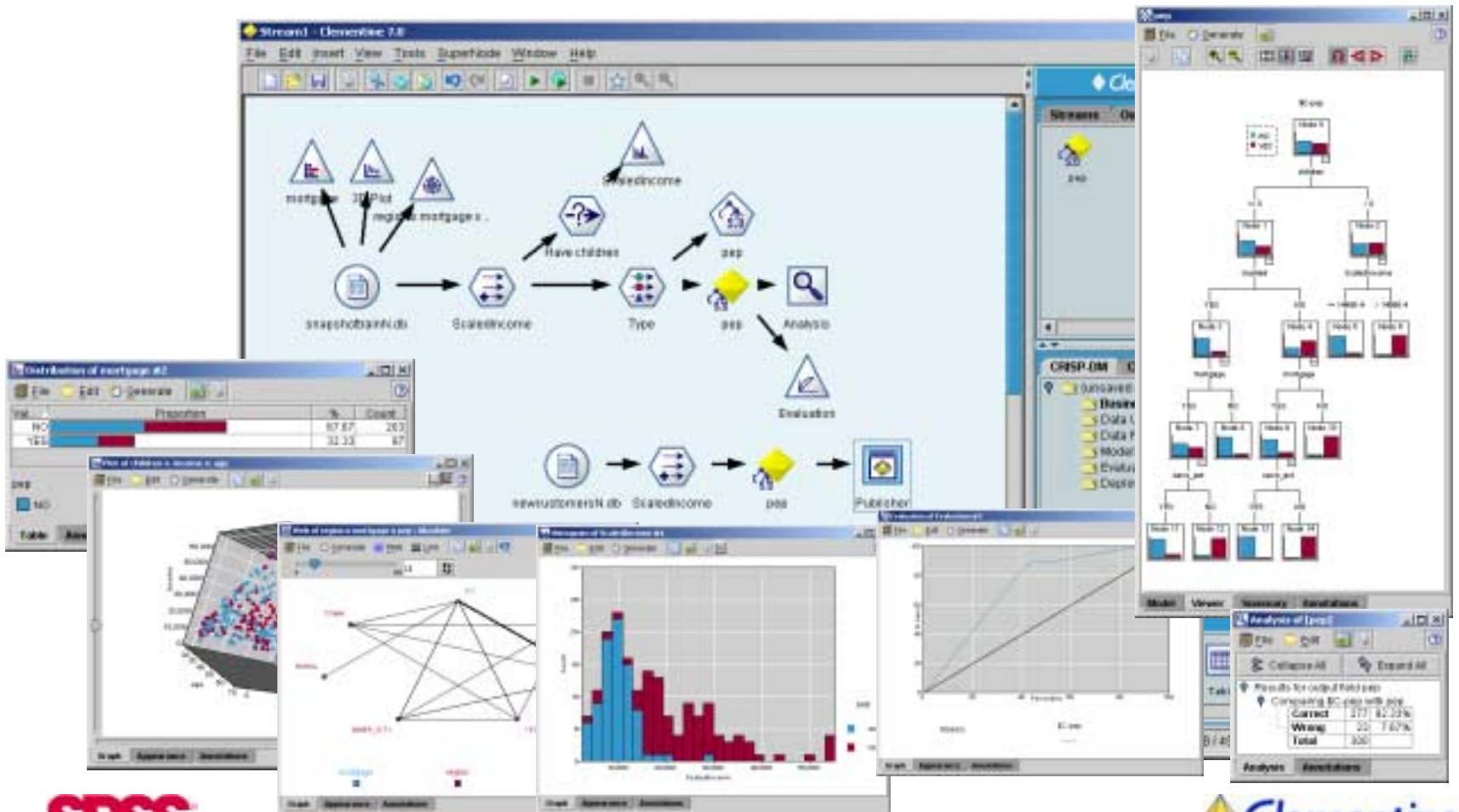


Comprehensive
Interactive
Problem-oriented

Enables the analyst
to “engage” with
the data



What is data mining like?



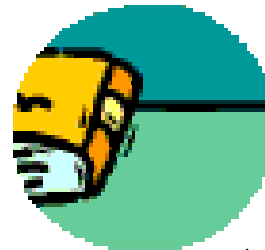
SPSS

Clementine



Text Mining – a data miner's view

- ▼ Data mining addresses problems through data
- ▼ Information extraction derives structured data from free text documents
- ▼ A natural match
- ▼ Find the concepts using information extraction then find the patterns using data mining



Name	Age	Income	Married	City	Country	Profession	Value	Last Purchase	City	Source
F. Blois	25	250k	Single	Yes	M/C	5	23.5	34	0	L1
I. Smit	37	330k	Married	Yes	VIS	3	123	102	2	L2
J. Dow	45	400k	Divorced	No	VIS	12	15.2	48	1	L1

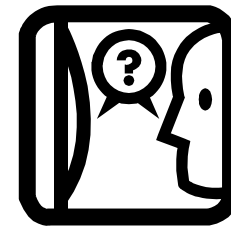
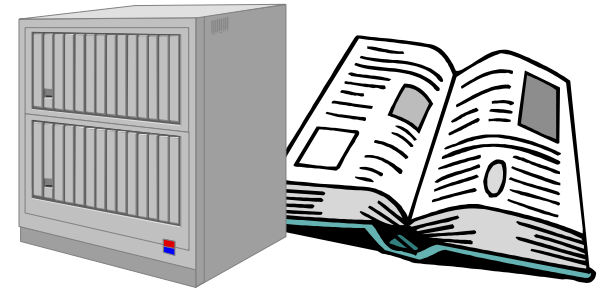




Data Mining / Text Mining Integration

The Need

- ▼ A large proportion of available data is free-text
- ▼ Classical data mining addresses problems only through structured data
- ▼ But...
The requirement to analyse structured and free-text data co-exist in the same application
Often in the same database
- ▼ Data mining suppliers get a constant stream of requests for free-text analysis

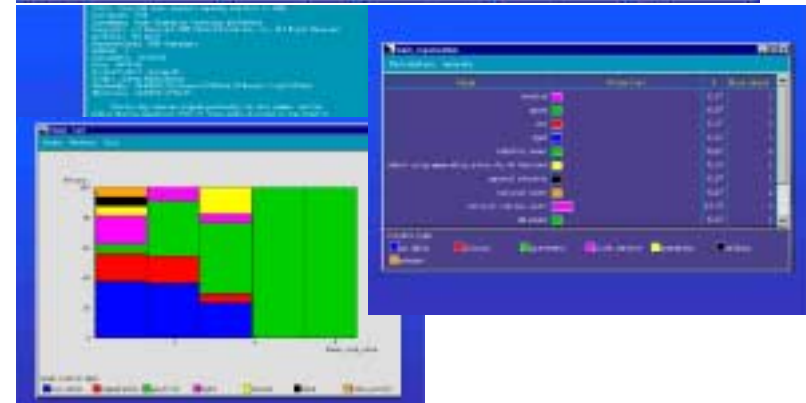
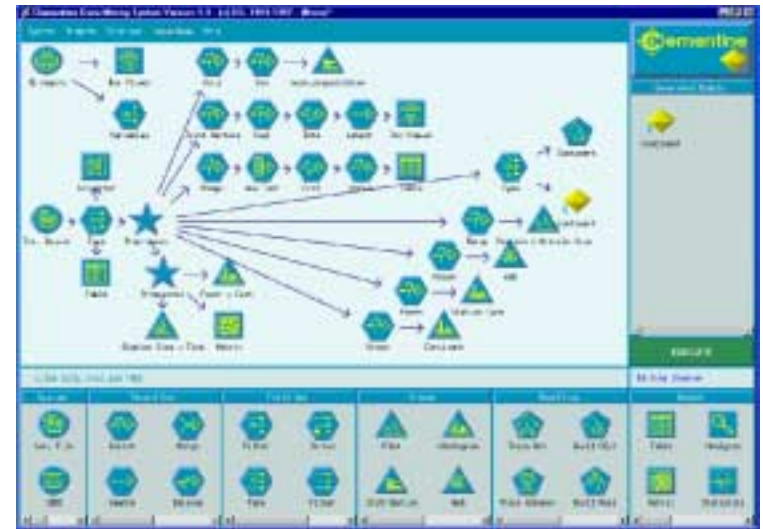




First steps towards integration

- ▼ An initial “demonstration”
- ▼ Power industry news items / competitive intelligence
- ▼ Integrated a purpose-built information extraction engine (from Brighton U. ITRI) with Clementine
- ▼ Highly successful, useful results

But: very domain-specific
So expensive to replicate





A New Integration

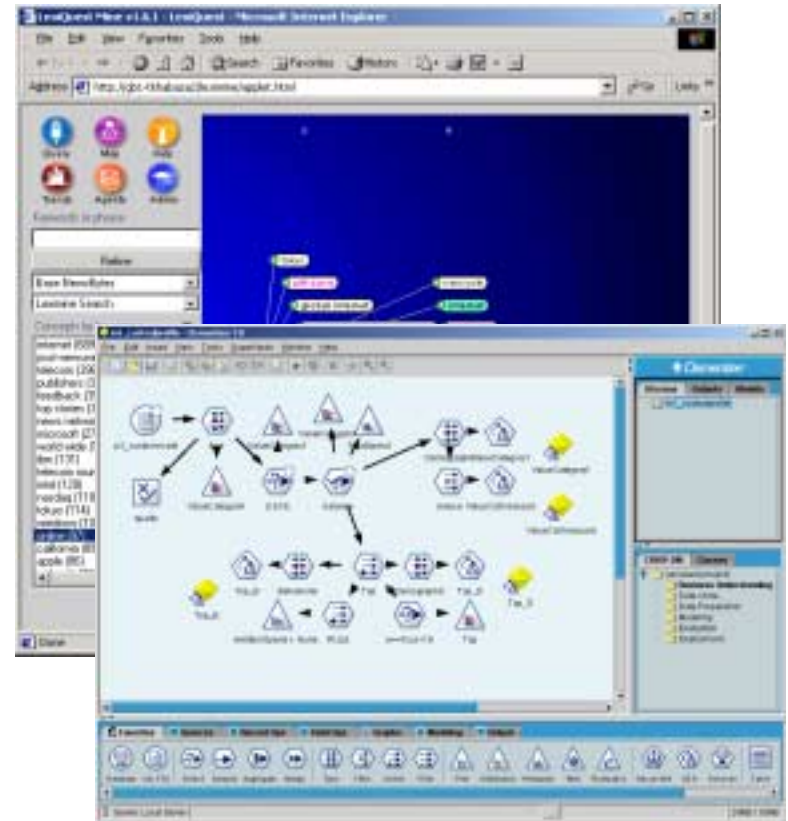
- ▼ LexiQuest text mining company acquired by SPSS in February 2002
- ▼ LexiQuest text mining tools
 - Mine – text mining
 - Categorize – document classifier
 - Information retrieval tools





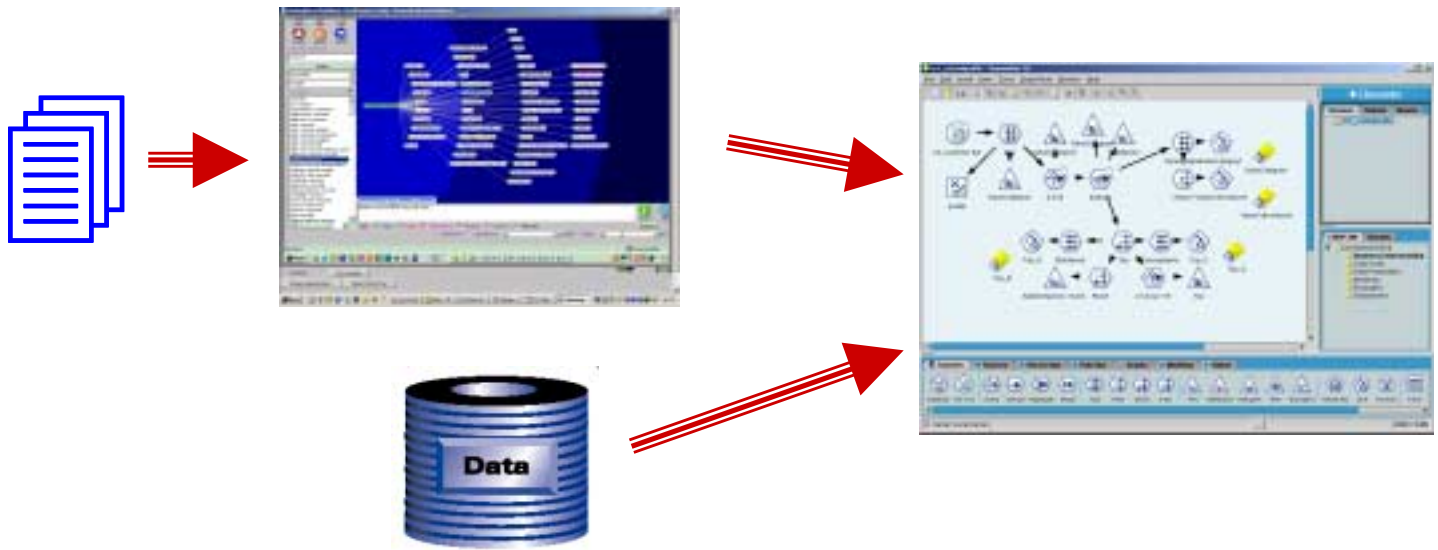
LexiQuest Mine Integration with Clementine

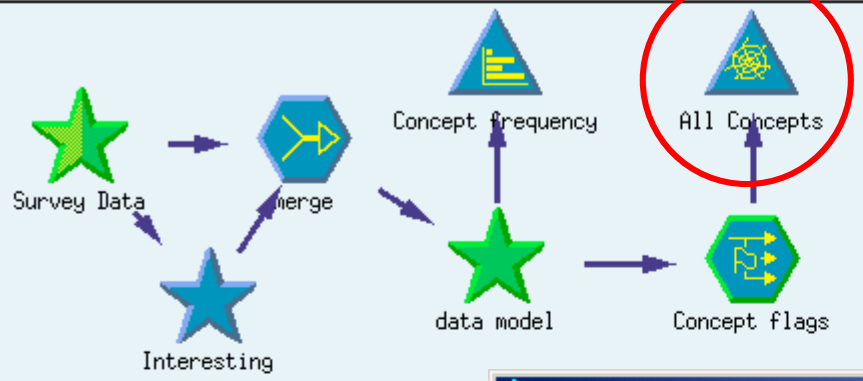
- ▼ LexiQuest's underlying "extraction engine"
 - ▼ Integrated with Clementine
 - ▼ The free-text miner's dream:
 - Extract the concepts
 - Find the patterns
- As simple as that!



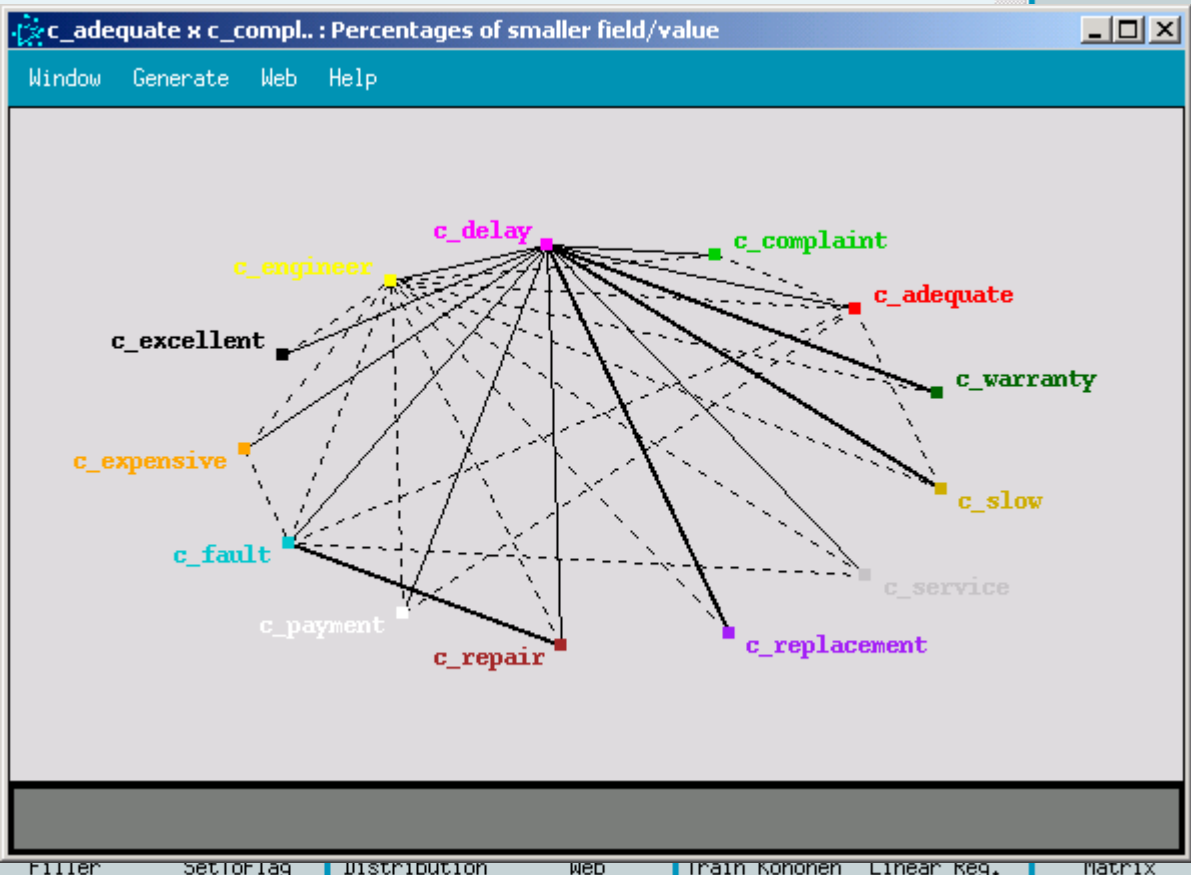


Text Mining Integrated into Clementine





Generated Models



Sources

- Var. File
- ODBC

Record Ops

- Sort
- Aggregate
- Distinct
- Append

EXECUTE

Output

- Analysis
- Statistics



Data Mining and Text Mining

A Bright Future



- ▼ Structure data and free-text data are no longer separate domains
- ▼ The fluency of exploration and discovery provided by data mining for structured data is now available for free text data and for combinations of the two.
- ▼ This will revolutionise CRM, fraud detection, crime investigation, competitive intelligence,