



KE5205 TEXT MINING 2017

INTRODUCTION TO TEXT ANALYTICS TOOLS & SOLUTIONS FOR TEXT ANALYTICS

Leong Mun Kew
Institute of Systems Science
National University of Singapore

email: munkew@nus.edu.sg

© 2017 National University of Singapore. The contents contained in this document may not be reproduced in any form or by any means, without the written permission of ISS, NUS, other than for the purpose for which it has been supplied.



Objectives of this module

At the end of this module, you can:

- **Describe the difference between data mining and text mining**
- **List the 5 basic use cases for text mining and provide examples relevant to real business usage**



Outline for these modules

- **Setting the stage**
- **What is text mining?**
- **What can text mining do?**
 - The 5 Basic Use Cases of text mining
- **Tools & solutions for text mining**
- **Workshop Assessment & Discussion**



WHAT IS TEXT MINING



Exercise

What is Data Mining?



What is Data Mining?

the process of

- From Wikipedia:

- The overall goal of the data mining process is to extract information from a data set and transform it into an understandable structure for further use.
- The actual data mining task is the automatic or semi-automatic analysis of large quantities of data to extract previously unknown interesting patterns such as groups of data records (cluster analysis), unusual records (anomaly detection) and dependencies (association rule mining)



What is Data Mining?

A
the outcome of

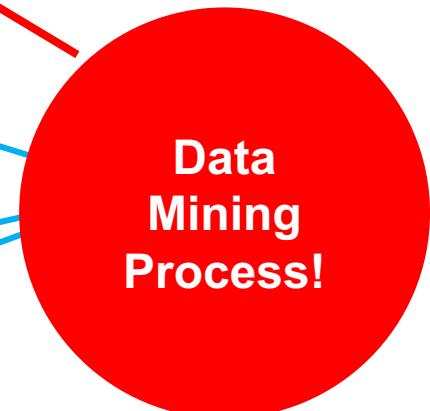
- From a business perspective:
 - Data mining is the transformation of structured **data** into **answers** to **business questions**
 - If you don't have a business context...
 - Then data mining is an academic exercise
 - If you don't have a business question...
 - Then data mining is a waste of time
 - If you don't have data...
 - Then data mining is really easy, but really useless



What is Data Mining?

the outcome of

- From a business perspective:
 - Data mining is the transformation of structured **data** into **answers** to **business questions**
 - If you don't have a **business context**...
 - Then data mining is an academic exercise
 - If you don't have a **business question**.
 - Then data mining is a waste of time
 - If you don't have **data**...
 - Then data mining is really easy, but really useless





What is Data Mining?

the outcome of

- From a business perspective:
 - Data mining is the transformation of structured **data** into **answers** to **business questions**
 - If you don't have a **business context**...
 - Then data mining is an academic exercise
 - If you don't have a **business question**.
 - Then data mining is a waste of time
 - If you don't have **data**...
 - Then data mining is really easy, but really useless



the right answer



Easy Exercise

What is the Outcome of Text Mining?



What is the Outcome of Text Mining?

- Obvious answer?
 - Text mining is the task of transforming **unstructured text data** into answers to business questions



So, what is Text Mining?

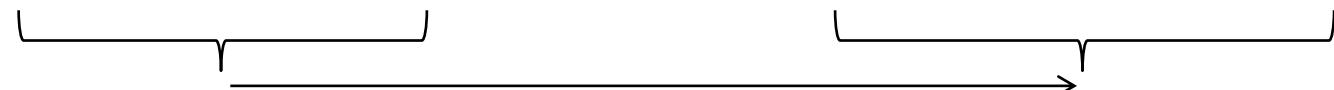
- What do you learn in “text mining”?
 - Text mining is the task of transforming unstructured text data into structured numerical data so that automatic algorithms can be applied to large document databases
 - Converting text to numbers requires the use of techniques for handling text at the individual word/character level to semi-structured documents to unstructured documents to document databases



Unstructured to structured

Cust ID	Date:time	Model	Comments
00010	20121203:2201	8560	Doesn't work – may have dropped. Out of warranty. Sent to svc.
00023	20121203:1034	8850	Cannot roam. System is enabled. Reset settings on phone. Done.
00025	20121203:1640	2338	No sound. Rebooted many times. Sent to svc. 3 months old.
01003	20121203:1030	6000-1	Bought 2 weeks back. Gift. No receipt. Wants to upgrade. Sent to svc.
20456	20121203:1025	6000-1	Out of space. 4GB uSD. Set default save to uSD for songs. Done.

Cust ID	Date:time	Model	Svc	Closed	Issue	... etc
00010	20121203:2201	8560	1	1	99	
00023	20121203:1034	8850	0	1	45	
00025	20121203:1640	2338	1	1	12	
01003	20121203:1030	6000-1	1	1	99	
20456	20121203:1025	6000-1	0	1	28	



Text Mining converts
unstructured text fields into one
or more columns of easily
processed numeric data



Why is text mining so tough?

- Feature extraction is necessary (and not easy!)
 - Need background knowledge and resources
- Documents represented by very many features
 - Short fat databases
 - Features that are significant may not be intuitive
- Patterns supported by small number of documents can be significant
- Very large numbers of patterns
 - Which patterns are significant in what context and domain?
 - Training data with outcomes to prune patterns
 - Interactive exploration also useful



WHAT CAN TEXT MINING DO?

THE 5 BASIC USE CASES OF TEXT MINING



Why text mining?

- **Data mining works**
 - Most information in the world is not in structured data form
 - The information in text needs to be unlocked
- Text is being **created in digital format** and available
 - Formal documents: word processing
 - Semi-structured text documents: patents, websites, ...
 - Informal text: email, social media, sms, tweets, ...
- Analyzing text, by itself or in conjunction with data, provides **better outcomes** for business decisions

Text Analytics Software, What Is It and Why is It Worth \$1.8 Billion?

InternetTimeMachine

Subscribe

127 videos ▾



35,744



Uploaded by InternetTimeMachine on Sep 24, 2010

<http://www.TheInternetTimeMachine.com> looks at text analytics software and why IBM bought Netezza for \$1.7 billion recently. Text mining or text digging has been around for years, so why is it so valuable now? Check out this video on data mining to see..

46 likes, 1 dislike

From: <http://www.youtube.com/watch?v=sTDWJBebwNY>



What can text mining do?

5 basic Use Cases:

- 1. Extract “meaning” from unstructured text**
- 2. Automatically put text into categories**
- 3. Improve accuracy in predictive modeling or unsupervised learning**
- 4. Identify specific or similar/relevant documents**
- 5. Extract specific information from the text**



1. Extract “meaning” from unstructured text

- Extract answers from large corpus of small documents or small corpus or large documents that is not doable by human eye
- Sentiment analysis
 - What are my customers saying about me?
 - What are the areas of concern to a target group?
 - Analyzing open-ended responses to survey questions
- Trending themes in a stream of text
 - Insurance claims trends, warranty claims analysis
- Summarizing text
 - Gisting – main theme of text documents/websites
 - Automatic keyword extraction



Overview -- Analyzing Twitter data with IBM BigSheets

IBMetInfo



Subscribe

28 videos ▾

Curt Hall	Curtsiphone	Sat Sep 11 17:08:03 +0000 2010
Martin Richard	Marlen1929	Fri Sep 10 19:55:06 +0000 2010
????? Bieber	RachSmiles4JB	Wed Sep 15 22:34:43 +0000 2010
Curt Hall	Curtsiphone	Sat Sep 11 17:08:04 +0000 2010
KickPost	KickPost	Fri Sep 10 19:55:06 +0000 2010
Leonidas Koustimpis	leonbis2000	Wed Sep 15 22:34:43 +0000 2010
Curt Hall	Curtsiphone	Sat Sep 11 17:08:04 +0000 2010
Black.Mamba	MsLadyJoycelynn	Fri Sep 10 19:55:06 +0000 2010
Jennifer ?	jenn4sgb13	Wed Sep 15 22:34:43 +0000 2010
Tweets Espana	espana_es	Sat Sep 11 17:08:03 +0000 2010
Hairulnizar	rullysmully	Fri Sep 10 19:55:06 +0000 2010
J.As	R_Angel_9	Wed Sep 15 22:34:43 +0000 2010
iPhone?????? ??	iphone_akashi	Sat Sep 11 17:08:04 +0000 2010



24 Hour Twitte... 36

▶ 2:46 / 5:24



2,193



Uploaded by [IBMetInfo](#) on Oct 31, 2010

This demonstration shows how IBM BigSheets can be used to find buyer sentiment in Twitter data. This is a shortened version of the demo. You can see the full step-by-step version at <http://www.youtube.com/watch?v=Jqq66INIQ0U>

8 likes, 0 dislikes

From: <http://www.youtube.com/watch?v=PSq7hZ0shLs>

Things to Note

This does the hard work

A	B	C	D
id	name	Sheet Name:	created_at
628	????	Sheet1	0 19:55:06 +0000 2010
029	waldheins	LW - Sentiment Analysis	15 22:34:43 +0000 2010
126	Curt Hall	(This is a Languageware UDF for sentiment analysis.)	1 17:08:03 +0000 2010
352	Alan Phillips		0 19:55:05 +0000 2010
165	Bryan Hammond		15 22:34:43 +0000 2010
355	Curt Hall		1 17:08:03 +0000 2010
765	Martin Richard		0 19:55:06 +0000 2010
394	????? Bieber		15 22:34:43 +0000 2010
624	Curt Hall		1 17:08:04 +0000 2010
155	KickPost		0 19:55:06 +0000 2010
709	Leonidas Koustimpis		15 22:34:45 +0000 2010
855	Curt Hall		1 17:08:04 +0000 2010
244	Black.Mamba		0 19:55:06 +0000 2010
290	Jennifer ?		15 22:34:46 +0000 2010
679	Tweets Espana		1 17:08:04 +0000 2010
341	Hairulnizar		0 19:55:06 +0000 2010
785	J.As	R Angel 9	Wed Sep 15 22:34:46 +0000 2010

New Sheet: Macro

Sheet Name: Sheet1

LW - Sentiment Analysis

This is a Languageware UDF for sentiment analysis.

Fill in parameters:

content*

text

type*

com.ibm.DictTrigger
com.ibm.Watchlist
com.ibm.NegativeIndicator
com.ibm.PositiveIndicator
com.ibm.QuestionIndicator
com.ibm.TwitterID
com.ibm.URL

Parameters Carry Over

✓ ✘



Things to Note

Result Data:

Ready

	type	name	screen_name	
1	com.ibm.en.PositiveIndicator	Special	Marlen1929	@m...
2	com.ibm.en.PositiveIndicator	fantastic	Curtsiphone	Mak...
3	com.ibm.en.PositiveIndicator	glorious	sharding	@Ke...
4	com.ibm.en.PositiveIndicator	Amazing	thaibisz	Top...
5	com.ibm.en.PositiveIndicator	like	MagsJB	My a...
6	com.ibm.en.PositiveIndicator	Cool	Cellphonez	Ipho...
7	com.ibm.en.PositiveIndicator	best	THE_Efram	@H...
8	com.ibm.en.PositiveIndicator	Quick	Leesa19043	@il...
9	com.ibm.en.PositiveIndicator	First	paladigarisbiz	(HO...
10	com.ibm.en.PositiveIndicator	wow	sjbuchanan007	@Ar...
11	com.ibm.en.PositiveIndicator	Fast	Leesa19043	@il...
12	com.ibm.en.PositiveIndicator	great	grattonboy	I'm ...
13	com.ibm.en.NegativeIndicator	doubt	gadgetinn	App...
14	com.ibm.en.NegativeIndicator	hurt	msluvmylife	Girl...
15	com.ibm.en.PositiveIndicator	like	POSTMODERNISM_	Rec...
16	com.ibm.en.PositiveIndicator	like	gadgetinn	App...

Decides sentiment based on lists of built-in keywords



What was the Business Context?

What was the Business Question/Need?

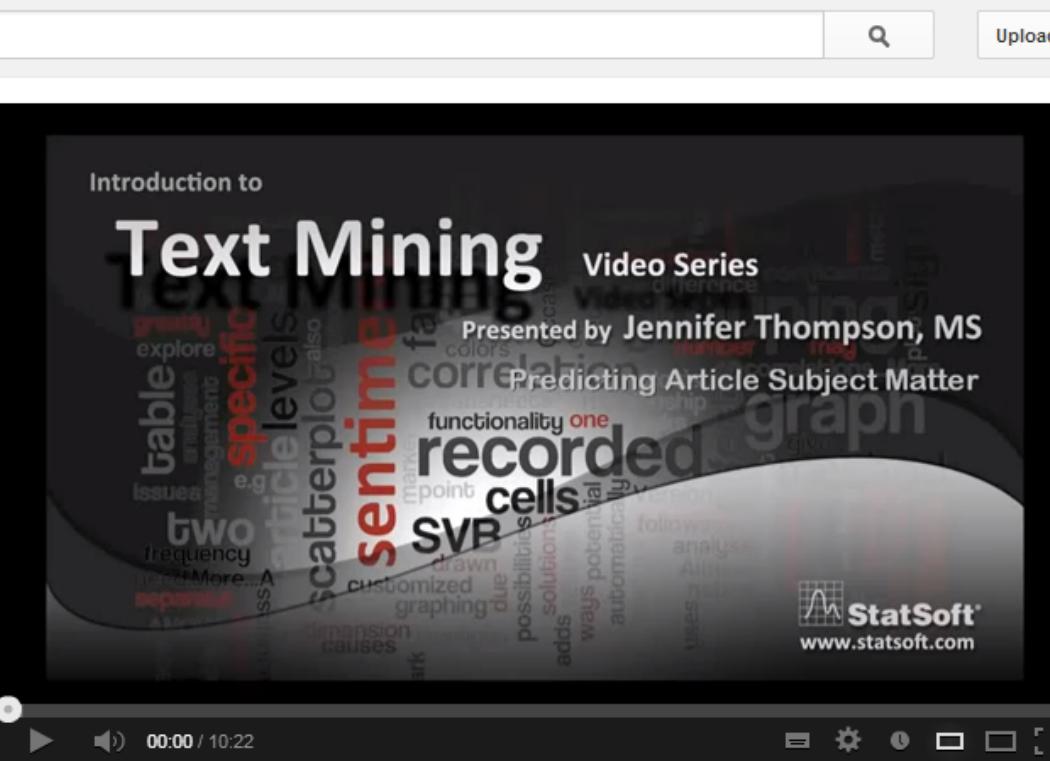
What was the data that was used?

What was the answer that was obtained?

What advantage did text mining provide in this case?

2. Automatically put text into categories

- **Classification – assigning one or more predefined categories to a text document, for subsequent processing**
- **Automatic actions based on category**
 - Email routing, spam filtering
 - News filtering
- **Identifying anomalies based on text descriptions**
 - Fraud detection, normally flag for human intervention



Text Mining Series - Automatically Classify Text Documents



StatSoft · 102 videos



Subscribe

1,629

3,402

13

1



Like



About

Share

Add to



Uploaded on 27 Oct 2011

In this case study, there is a need to automatically classify text documents based on their content. Currently, the text articles are manually read and acted upon. Our goal is to automate as much as possible with a predictive model sorting the text files. Articles related to financial earnings should be flagged for review and sent to the appropriate individuals. In this video, we explore how STATISTICA Text Miner can be used to explore and index the text.

From: <http://www.youtube.com/watch?v=Q5K3gyQJkC0>

Predicting Article Subject Matter

- Project Goal
 - To automatically classify articles as either related to financial earning or not

- Project Plan
 - Using 5,000 expertly classified articles from

Reuter, index the text and build predictive models that will classify new articles

Clear Business Objective

Standard Oil Co. to manage BP's
gasoline in B.R.
<BPD>, which is
called BP/Stat
under the over-

Texas Comptroller filed an application to create the largest network would likely be discounted.

BankAmerica Corp. is not under pressure to act quickly on its proposed equity offering and would do well to delay it because of the stock's recent poor performance, banking analysts said. Some analysts say they have recommended that BankAmerica delay its up to \$1 billion equity offering, which has yet to be approved by the company's board of directors. Investors, BankAmerica's stock fell this week, along with other banking issues, on the news that Brazil had suspended interest payments on a large portion of its foreign debt. The stock closed at \$12.12, down 7 1/8, the afternoon after falling 1 1/2 points the previous week on concern that the market would settle with the immediate threat of the First International Ratings Co. takeover bid gone, BankAmerica's underwriters to sell the securities into a market that will be nervous on bank stocks in the near term. BankAmerica filed a registration statement with the Securities and Exchange Commission on Feb. 21. It must wait 180 days before it can begin selling. The First International ratings agency's takeovers of several U.S. banks, a spokesman said and S&P approval is taking longer than expected and market conditions must now be reevaluated. "The circumstances at the time will determine what we do," said Andrew J. Lerner, BankAmerica's chairman and President for Financial Institutions. He said he had asked if it was possible to postpone the offer indefinitely after it receives S&P approval. "I'd put it off as long as they conceivably could," said Lawrence Cohn, analyst with Merrill Lynch, Pierce, Fenner and Smith. Cohn said the longer BankAmerica is able to defer the offering, the better it will be for its financial institutions. Although BankAmerica has not yet specified the type of deal it would offer, most analysts believed a convertible preferred stock would encompass at least part of it. Such an offering at a depressed stock price would mean a lower conversion price and a lower value for the stock held by existing shareholders. Joseph Aman, analyst with Joseph L. Gray, however, believes that while they believe the Brazilian debt problems will continue to hang over the banking industry through the quarter, the initial stock reaction is likely to ease over the coming weeks. Nevertheless, BankAmerica, which holds about 10 percent of Brazil's banking system, is facing a difficult situation. Brazil's interest rate is now 200 percent on the dollar, and as much as 250 million dls if Brazil pays no interest for a year, said Joseph Aman, analyst with Bres, Wilcox and Co. He noted, however, that any potential losses would not show up in the first quarter, and that the company would have to live with more losses. The BankAmerica bid fails to service its debt, the analysts said and they expect the debt will be written down, similar to way Mexico's debt was, minimizing losses to the banking industry.

Data has “ground truth” established by experts



What was the Business Context?

What was the Business Question/Need?

What was the data that was used?

What was the answer that was obtained?

What advantage did text mining provide in this case?



3. Improve predictive accuracy in predictive modeling or unsupervised learning

- Use text mining to improve data mining results (“Lift”)
- Changing text to numbers to work with data mining
 - Build a data matrix based on word/phrase counts
 - Compute various indices based on those matrices
 - Merge indices, counts with structured data for mining
- Predicting insurance fraud from claims processing notes
- Using dictionaries to control vocabulary, reduce variance

Text Mining Series: Predicting Fraudulent Claims

StatSoft



Subscribe

88 videos

Introduction to

Text Mining

Video Series

Presented by Jennifer Thompson, MS

Predicting Fraudulent Claims

 StatSoft
www.statsoft.com

0:16 / 6:01



1,165

Uploaded by StatSoft on Nov 15, 2011

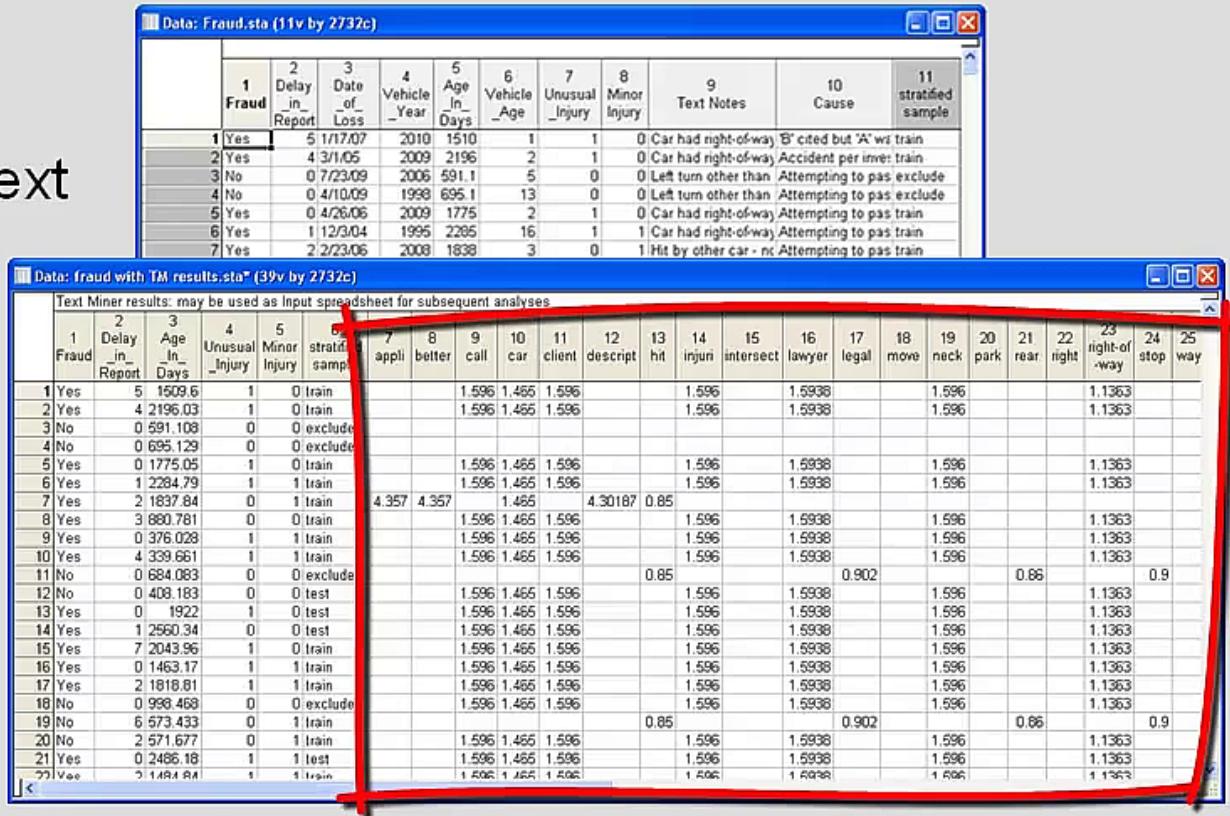
5 likes, 0 dislikes

In this case study, fraud detection models are built using the structured variables and provide a good predictive model, finding fraudulent claims. Then with the aid of STATISTICA Text Miner, the notes for each claim were indexed and the results were added to the predictive variable pool. Predictive models built with the added text mining results gave a 10% improvement in finding fraudulent claims.

From: <http://www.youtube.com/watch?v=OIQpm8qTog4>

Predicting Fraudulent Claims

- Can predictability of fraudulent claims be improved by adding Text Mining results?
- Variables for analysis include
 - Delay in report
 - Policy age
 - Unusual injury
 - Minor injury
 - Text notes



The image shows two adjacent Microsoft Excel spreadsheets. The left spreadsheet, titled 'Data: Fraud.sta (11v by 2732c)', contains raw data with columns labeled 1 through 11. The right spreadsheet, titled 'Data: fraud with TM results.sta* (39v by 2732c)', shows the same data converted into numerical values, with columns labeled 1 through 25. A red box highlights the first few rows of both spreadsheets to illustrate the transformation process.

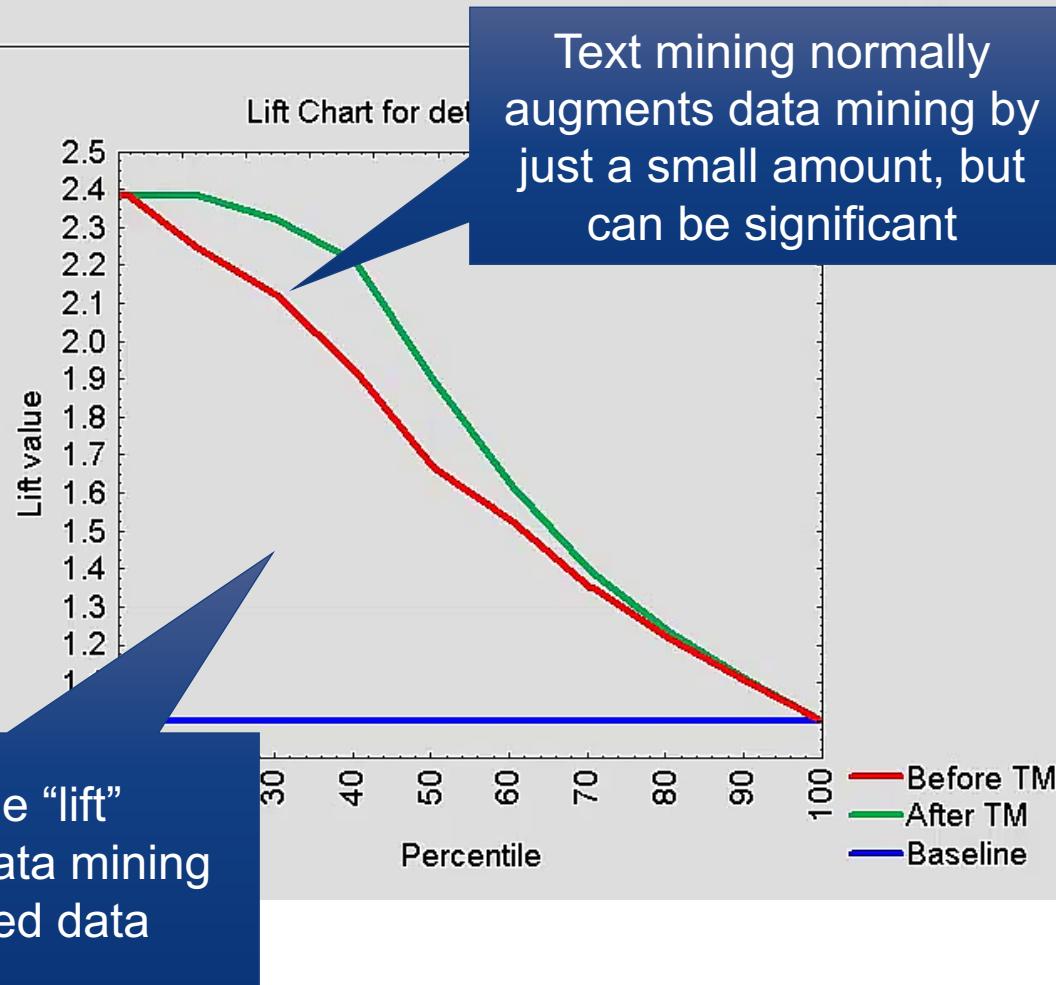
	1 Fraud	2 Delay_in_Report	3 Date_of_Loss	4 Vehicle_Year	5 Age_In_Days	6 Vehicle_Age	7 Unusual_Injury	8 Minor_Injury	9 Text_Notes	10 Cause	11 stratified_sample
1	Yes	5	1/17/07	2010	1510	1	1	0	Car had right-of-way, 'B' cited but 'A' was train		
2	Yes	4	3/1/05	2009	2196	2	1	0	Car had right-of-way, Accident per inv: train		
3	No	0	7/23/09	2006	591.1	5	0	0	Left turn other than Attempting to pass, exclude		
4	No	0	4/10/09	1998	695.1	13	0	0	Left turn other than Attempting to pass, exclude		
5	Yes	0	4/26/08	2009	1775	2	1	0	Car had right-of-way, Attempting to pass train		
6	Yes	1	12/3/04	1995	2285	16	1	1	Car had right-of-way, Attempting to pass train		
7	Yes	2	2/23/06	2008	1838	3	0	1	Hit by other car - no Attempting to pass train		

	1 Fraud	2 Delay_in_Report	3 Age_In_Days	4 Unusual_Injury	5 Minor_Injury	6 stratified_sample	7 appli	8 belter	9 call	10 car	11 client	12 descript	13 hit	14 injuri	15 intersect	16 lawyer	17 legal	18 move	19 neck	20 park	21 rear	22 right	23 right-of-way	24 stop	25 way
1	Yes	5	1509.6	1	0	train				1.596	1.465	1.596		1.596	1.5938	1.596							1.1363		
2	Yes	4	2196.03	1	0	train				1.596	1.465	1.596		1.596	1.5938	1.596							1.1363		
3	No	0	591.108	0	0	exclude																			
4	No	0	695.129	0	0	exclude																			
5	Yes	0	1775.05	1	0	train				1.596	1.465	1.596		1.596	1.5938	1.596							1.1363		
6	Yes	1	2284.79	1	1	train				1.596	1.465	1.596		1.596	1.5938	1.596							1.1363		
7	Yes	2	1837.84	0	1	train	4.357	4.357		1.465		4.30187	0.85												
8	Yes	3	880.781	0	0	train				1.596	1.465	1.596		1.596	1.5938	1.596							1.1363		
9	Yes	0	376.028	1	1	train				1.596	1.465	1.596		1.596	1.5938	1.596							1.1363		
10	Yes	4	339.661	1	1	train				1.596	1.465	1.596		1.596	1.5938	1.596							1.1363		
11	No	0	684.083	0	0	exclude								0.85			0.902							0.9	
12	No	0	408.183	0	0	test				1.596	1.465	1.596		1.596	1.5938	1.596							1.1363		
13	Yes	0	1922	1	0	test				1.596	1.465	1.596		1.596	1.5938	1.596							1.1363		
14	Yes	1	2560.34	0	0	test				1.596	1.465	1.596		1.596	1.5938	1.596							1.1363		
15	Yes	7	2043.96	1	0	train				1.596	1.465	1.596		1.596	1.5938	1.596							1.1363		
16	Yes	0	1463.17	1	1	train				1.596	1.465	1.596		1.596	1.5938	1.596							1.1363		
17	Yes	2	1818.81	1	1	train				1.596	1.465	1.596		1.596	1.5938	1.596							1.1363		
18	No	0	998.468	0	0	exclude				1.596	1.465	1.596		1.596	1.5938	1.596							1.1363		
19	No	6	573.433	0	1	train							0.85			0.902							0.9		
20	No	2	571.677	0	1	train				1.596	1.465	1.596		1.596	1.5938	1.596							1.1363		
21	Yes	0	2486.18	1	1	test				1.596	1.465	1.596		1.596	1.5938	1.596							1.1363		
22	Yes	3	1484.84	1	1	train				1.596	1.465	1.596		1.596	1.5938	1.596							1.1363		

Text converted into numbers

Predicting Fraudulent Claims – Project Steps

1. Build predictive models using the structured data
2. Index the text notes for accident claims
3. Build predictive models using the structured data and text mining results
4. Compare model performance





What was the Business Context?

What was the Business Question/Need?

What was the data that was used?

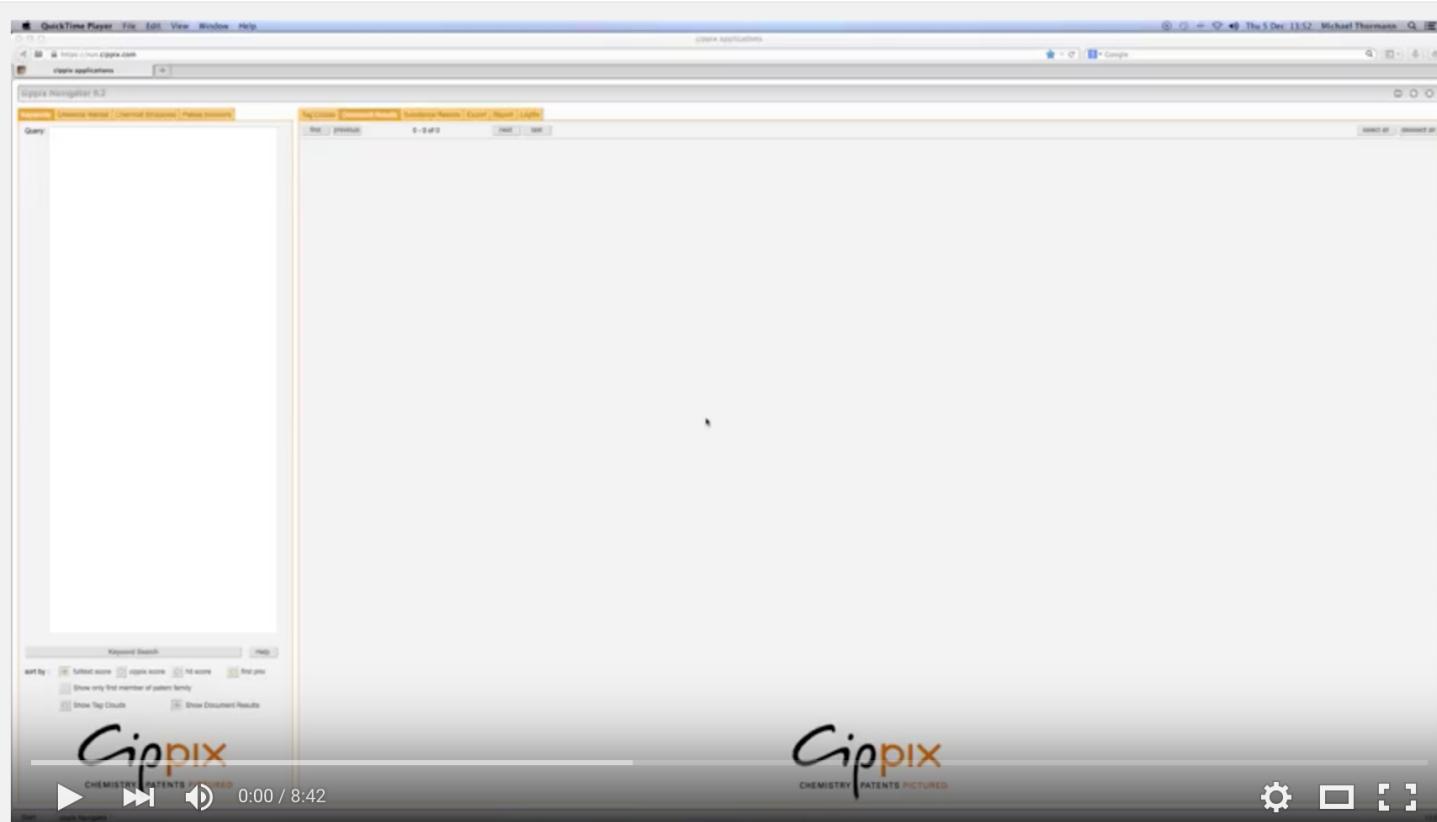
What was the answer that was obtained?

What advantage did text mining provide in this case?



4. Identify specific or similar/relevant documents

- **Document searching – given a specific documents, identify other documents in the corpus which are similar and relevant**
- **Create a pool of similar/linked documents for analysis**
 - Patent search, primary research
 - Forensic investigations into text
- **Web search**



Cippix Tutorial: How to search for similar documents



Mestrelab Research

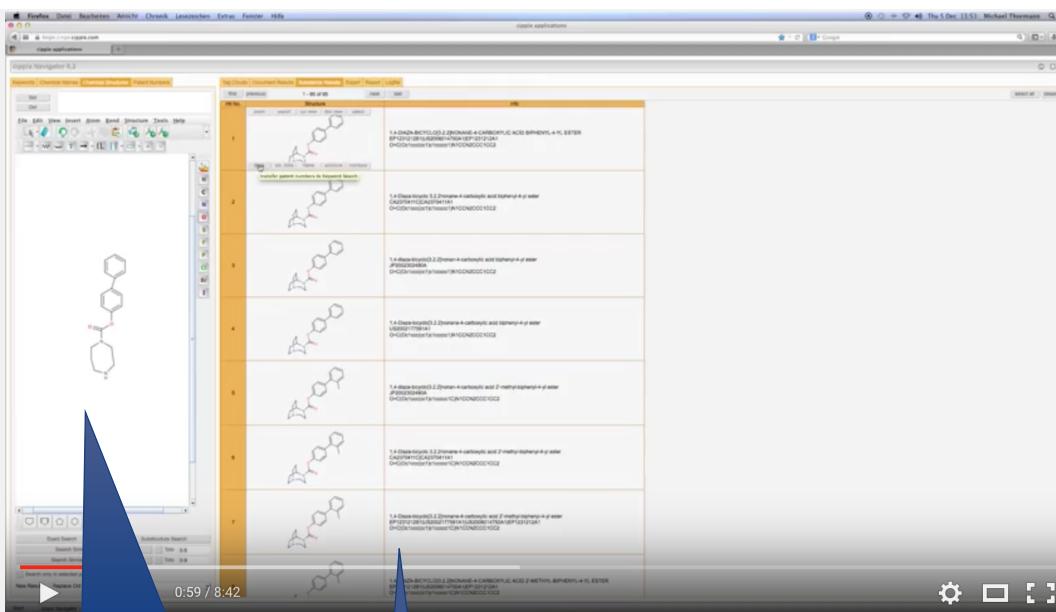
142

32 views

More

0 0

<https://www.youtube.com/watch?v=evLDjHQzMRU>



Chemical
Substructure
query

Matching
Documents

Things to Note

Things to Note

Chemical Substructure query

Keyword Search

Matching Documents

Matching Patent



The screenshot displays a patent search interface with three main sections. On the left, a video player shows a chemical structure being input into a search field. In the center, a list of matching documents is shown, each with a thumbnail image of the chemical structure and a brief description. A blue callout box labeled "Keyword Search" points to the search bar. On the right, a detailed view of a single matching patent is shown, featuring a large chemical structure, a tag cloud of keywords, and several smaller chemical structures. A blue callout box labeled "Matching Patent" points to the detailed patent view. Another blue callout box labeled "Matching Documents" points to the list of results in the middle section.

Things to Note

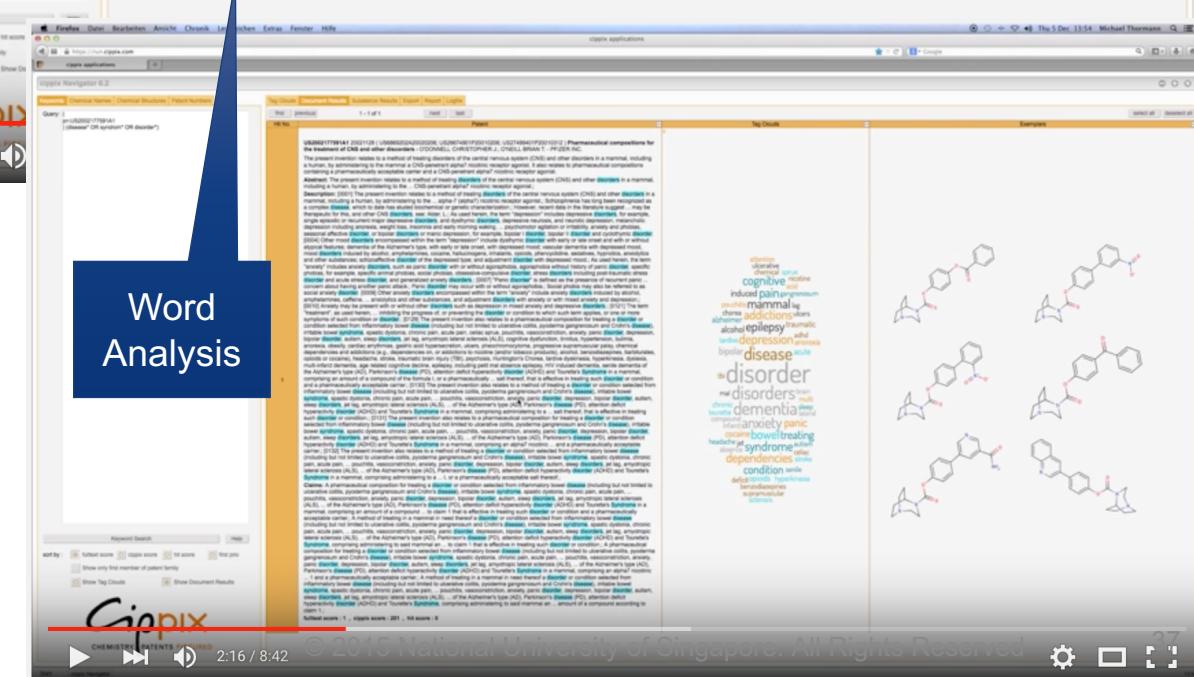
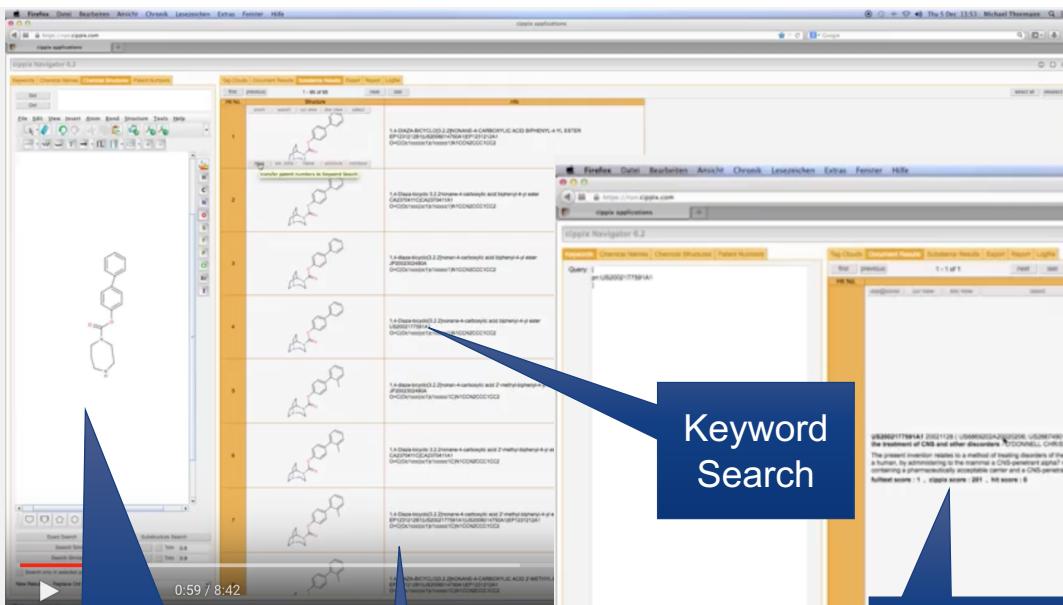
Chemical
Substructure
query

Matching
Documents

Keyword
Search

Matching
Patent

Word
Analysis



Things to Note

Chemical
Substructure
query

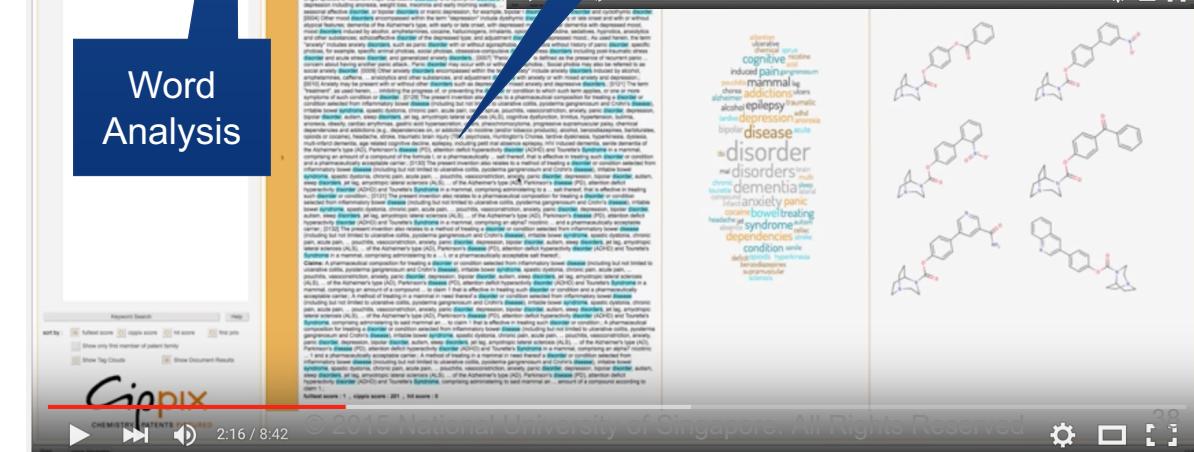
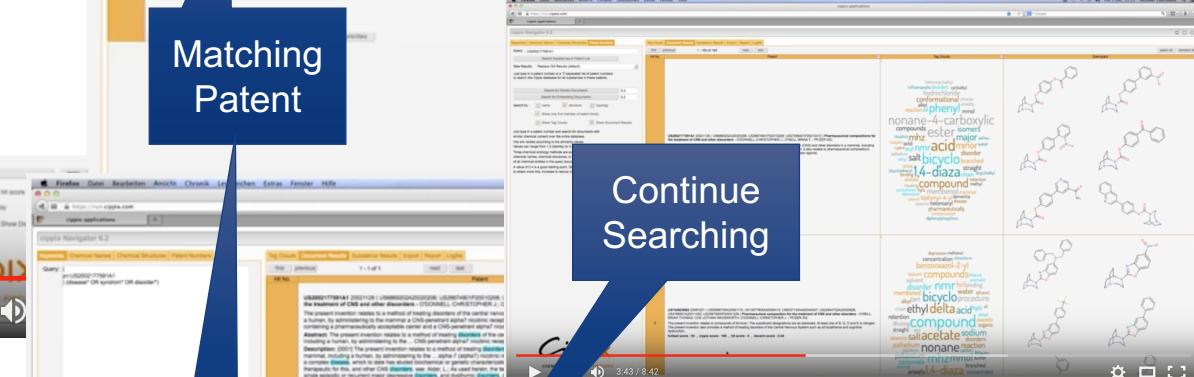
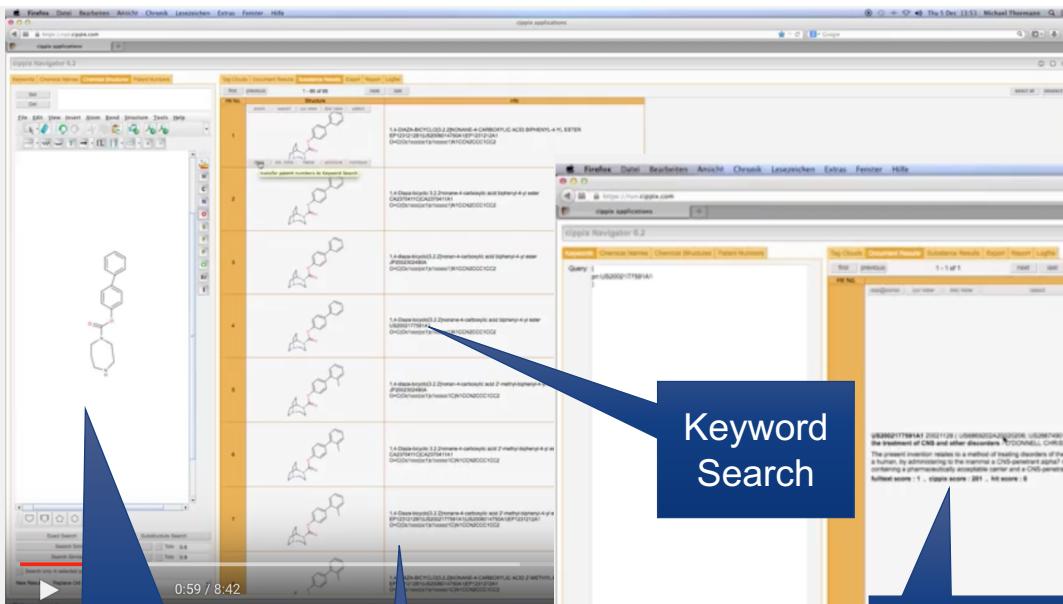
Matching
Documents

Keyword
Search

Matching
Patent

Continue
Searching

Word
Analysis





What was the Business Context?

What was the Business Question/Need?

What was the data that was used?

What was the answer that was obtained?

What advantage did text mining provide in this case?



5. Extract specific information from the text

- There are many answers in text documents. The problem is given a question, how to get the answer, not just the document. The task is called “question answering” (QA)
- At a more basic level, identify and extract “named entities” from documents and corpora
- Automatic QA
 - Compare interest rates at banks for best deal
 - Automated help desk and FAQs
- Name Entity Extraction (NER)
 - Dates, money sums, organizations, stock symbols, etc.

How IBM's Watson supercomputer wins at Jeopardy, with IBM's Dave Gondek



Subscribe

447 videos



Like



Dislike



Share



113,383

Uploaded by [engadget](#) on Jan 13, 2011

How IBM's Watson supercomputer wins at Jeopardy, with IBM's Dave Gondek.

343 likes, 5 dislikes

<http://www.engadget.com/2011/01/13/ibms-watson-supercomputer-destroys-all-hum...>

From: http://www.youtube.com/watch?v=d_yXV22O6n4

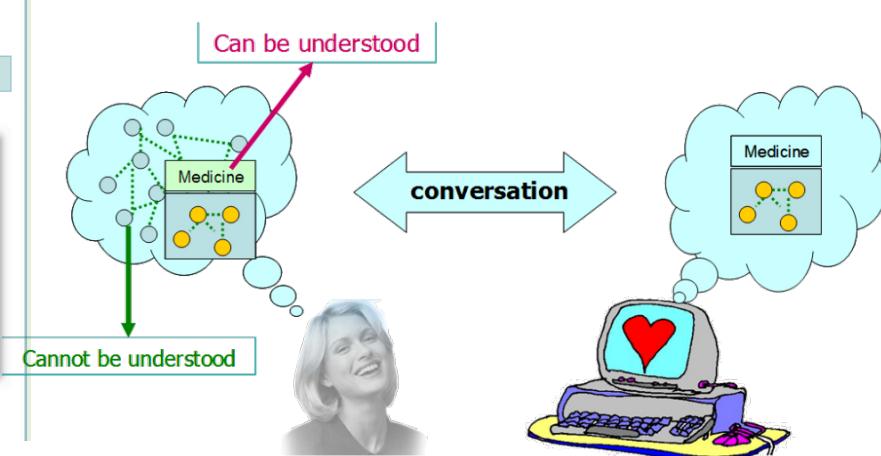
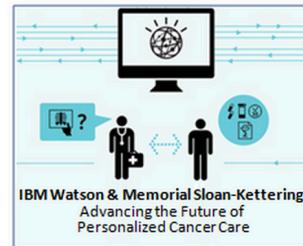
High Tech Advancing Future of Personalized Cancer Care

01/19/2013

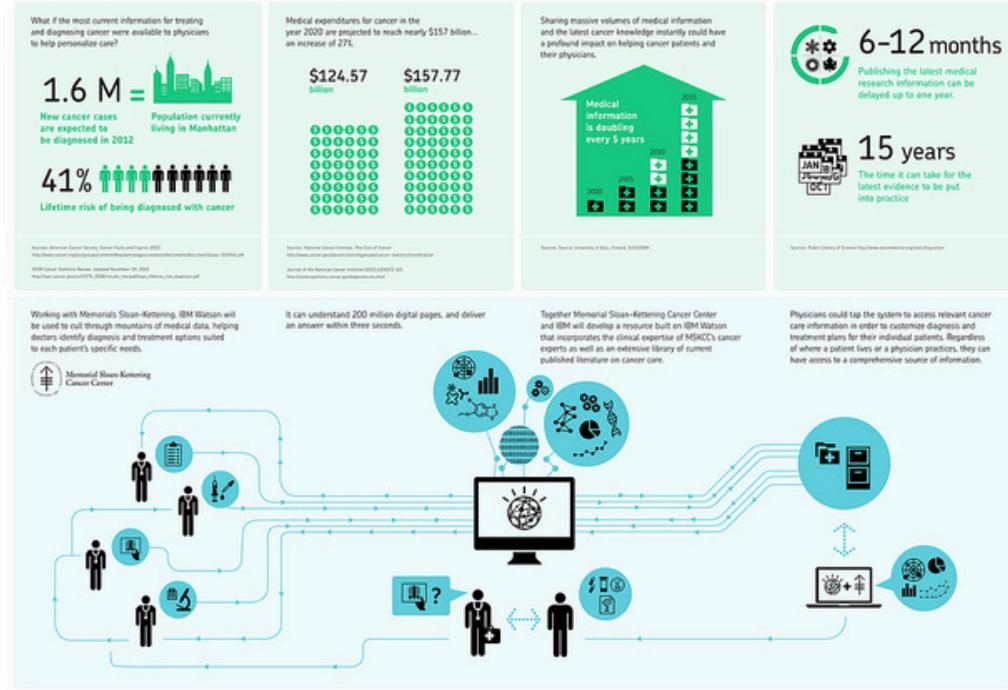
Memorial Sloan-Kettering Cancer Center, IBM to Collaborate in Applying Watson Technology to Help Oncologists

IBM Watson combined with MSKCC's clinical knowledge will help physicians access and integrate latest science and knowledge

New York City – 22 Mar 2012: Memorial Sloan-Kettering Cancer Center and IBM have agreed to collaborate on the development of a powerful tool built upon IBM Watson in order to provide medical professionals with improved access to current and comprehensive cancer data and practices. The resulting decision support tool will help doctors everywhere create individualized cancer diagnostic and treatment recommendations for their patients based on current evidence.



Memorial Sloan Kettering & IBM Watson: Advancing the Future of Personalized Cancer Care





IBM Watson Demo Oncology Diagnosis and Treatment 2 min.



kuresurem

Subscribe

72

1,315

Add to

Share

More

0

0

Published on 14 Aug 2013

The IBM Watson Cancer Diagnosis and Treatment Adviser demo was created in close collaboration with Memorial Sloan Kettering, one of the world's preeminent cancer treatment and research institutions. The demo scenario follows the interactions of a hypothetical oncologist and patient as they move through consultations, tests, treatment options, patient preferences and pre-authorization. It showcases IBM Watson's

<https://www.youtube.com/watch?v=T7M1Dgyaapw>



What was the Business Context?

What was the Business Question/Need?

What was the data that was used?

What was the answer that was obtained?

What advantage did text mining provide in this case?



TOOLS & SOLUTIONS FOR TEXT MINING



6 Phases of CRISP-DM

Cross-Industry Standard Process for Data Mining

- **Business Understanding:** This initial phase focuses on understanding the project objectives and requirements from a business perspective, and then converting this knowledge into a data mining problem definition, and a preliminary plan designed to achieve the objectives.
- **Data Understanding:** The data understanding phase starts with an initial data collection and proceeds with activities in order to get familiar with the data, to identify data quality problems, to discover first insights into the data, or to detect interesting subsets to form hypotheses for hidden information.
- **Data Preparation:** The data preparation phase covers all activities to construct the final dataset (data that will be fed into the modeling tool(s)) from the initial raw data. Data preparation tasks are likely to be performed multiple times, and not in any prescribed order. Tasks include table, record, and attribute selection as well as transformation and cleaning of data for modeling tools.
- **Modeling:** In this phase, various modeling techniques are selected and applied, and their parameters are calibrated to optimal values. Typically, there are several techniques for the same data mining problem type. Some techniques have specific requirements on the form of data. Therefore, stepping back to the data preparation phase is often needed.
- **Evaluation:** At this stage in the project you have built a model (or models) that appear to have high quality, from a data analysis perspective. Before proceeding to final deployment of the model, it is important to more thoroughly evaluate the model, and review the steps executed to construct the model, to be certain it properly achieves the business objectives. A key objective is to determine if there is some important business issue that has not been sufficiently considered. At the end of this phase, a decision on the use of the data mining results should be reached.
- **Deployment:** Creation of the model is generally not the end of the project. Even if the purpose of the model is to increase knowledge of the data, the knowledge gained will need to be organized and presented in a way that the customer can use it. Depending on the requirements, the deployment phase can be as simple as generating a report or as complex as implementing a repeatable data mining process. In many cases it will be the customer, not the data analyst, who will carry out the deployment steps. However, even if the analyst will not carry out the deployment effort it is important for the customer to understand up front the actions which will need to be carried out in order to actually make use of the created models.

From: http://en.wikipedia.org/wiki/Cross_Industry_Standard_Process_for_Data_Mining



CRISP-DM Process Diagram

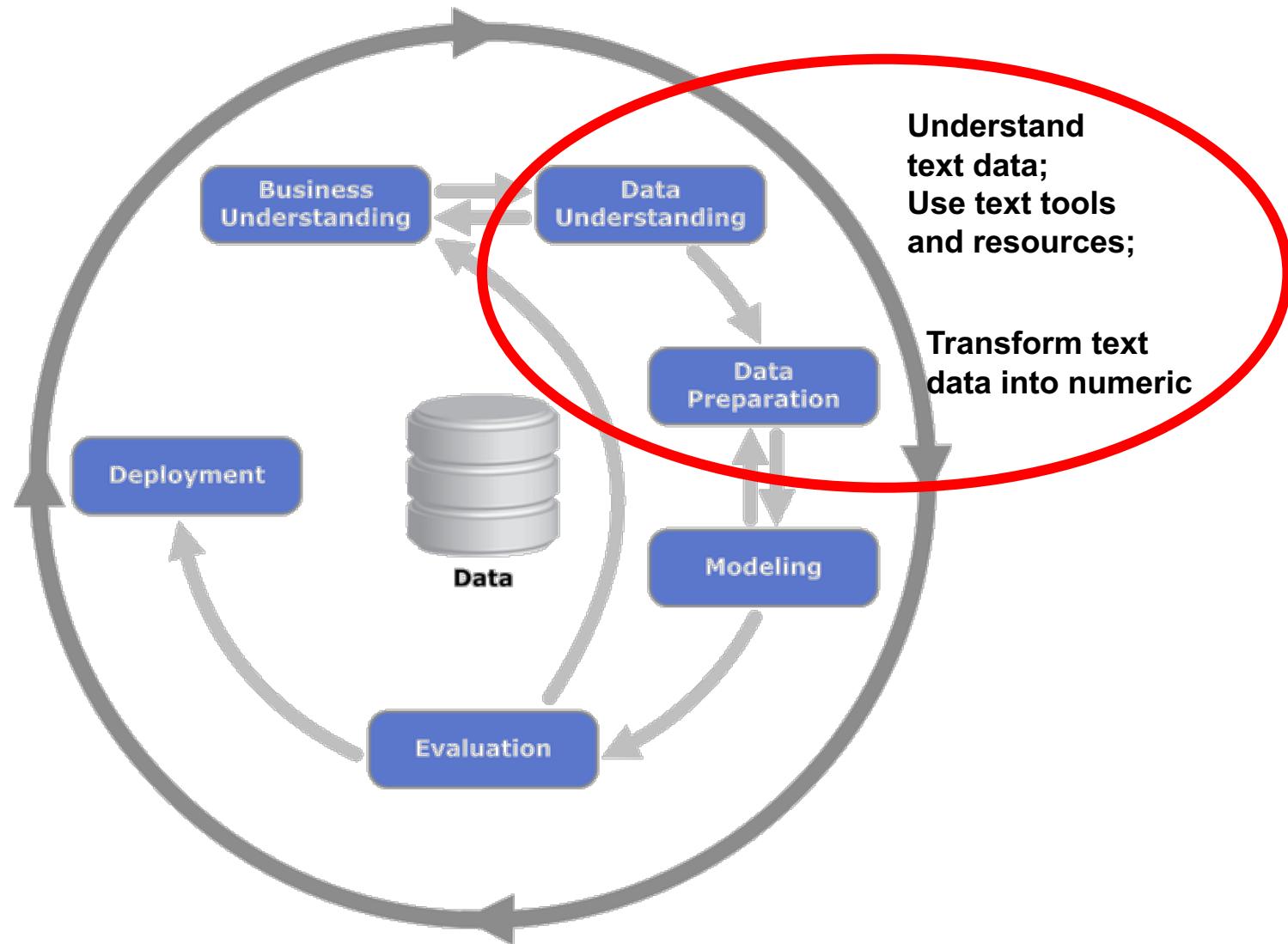
Note the Arrows – iteration & sequence



From: http://en.wikipedia.org/wiki/Cross_Industry_Standard_Process_for_Data_Mining



CRISP-DM Process with Text Mining

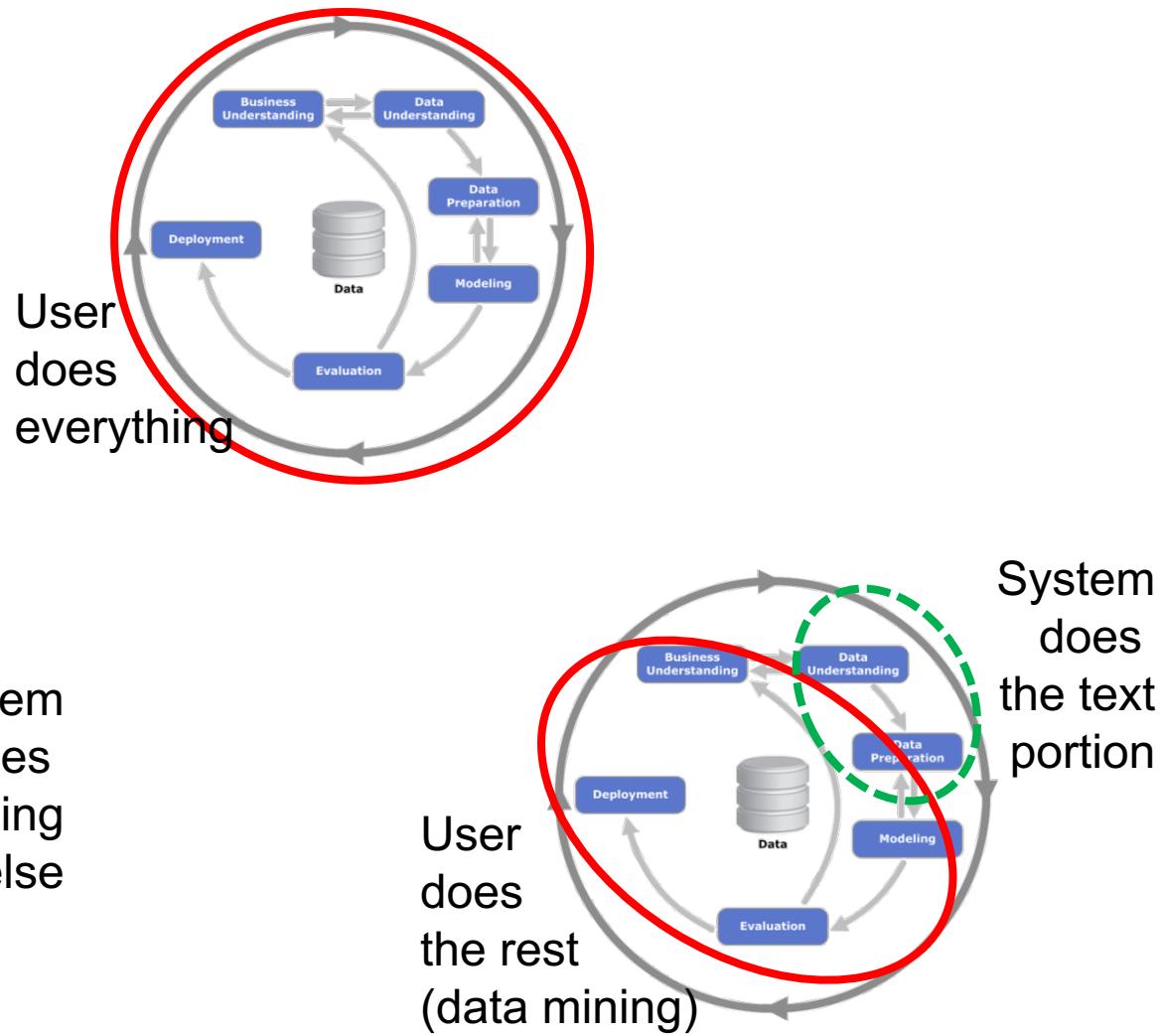




Factors to decide on the solution

- **What are your business outcomes?**
 - Numbers? Or “Actionable Insights”
- **How much skill do you have?**
 - In the process, in the tool, in the linguistic ability
- **How much time do you have?**
 - Skill + time = capacity to improve the linguistic extraction models
 - No time OR no skill = stay with the defaults
- **How often are you going to do this?**
 - Ad hoc usage? Unique data?
 - Regular runs over similar data?
- **What's your budget?**

Some possible models...





Factors to decide on the tool(s)

- **How general/flexible a tool do you want?**
 - Does one thing only (e.g., text survey) or very flexible?
 - Domain specific (dictionaries) or open domain?
 - Monolingual, true-multilingual, or translated multilingual?
- **How much help do you need?**
 - Open source tools are “free”, but you learn for yourself
 - Vendor tools are expensive, but you get training and support
 - Is the support local or overseas?
 - Courses can be tool specific, or tool independent
- **Do you need to do it yourself?**
 - If you know what you need, can you outsource the how?

What is the Business of my organisation?

What is the Business Need?

Do we have data that can be used? Who understands it?

What answer do we want? What action will result from that?



Reference & Resources

- Chris Manning & Hinrich Schutze, *Foundations of Statistical Natural Language Processing*, MIT Press, 1999
- NLP resources: <http://nlp.stanford.edu/links/statnlp.html>
- Christopher Potts (Stanford University), Sentiment Symposium Tutorial,
<http://sentiment.christopherpotts.net/index.html>
- John Elder, Gary Miner, Bob Nisbet. *Practical Text Mining and Statistical Analysis for non-Structured Text Data Applications*, Academic Press, 2012
- Roger Bilisoly. *Practical Text Mining with PERL*, John Wiley & Sons, 2008