



# NICF - TEXT ANALYTICS

## MODULE 2: INTRODUCTION TO TEXT ANALYTICS

**Dr. Fan Zhenzhen**

**Dr. Leong Mun Kew**

**Institute of Systems Science**

**National University of Singapore**

© 2013-2021 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**
- **Define the process to perform text analytics based on the business requirements and text analytics artifacts**



# Outline for these modules

- **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



# What is Data Mining?

*the process of*

- From Wikipedia:

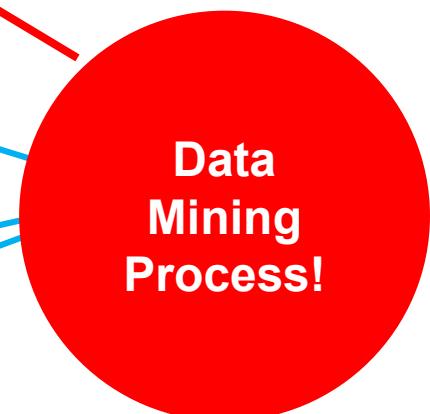
- The overall goal of the data mining process is to extract information from a data set and transform it into an understandable structure for further use.
- The actual data mining task is the automatic or semi-automatic analysis of large quantities of data to extract previously unknown interesting patterns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



# What is Data Mining?

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**



# 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?

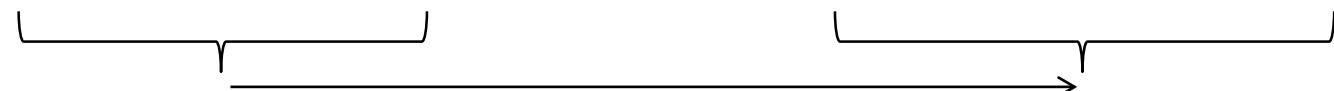
- What do you learn in “text mining”?
  - Text mining is the task of transforming unstructured text data into structured (numerical) data so that automatic algorithms can be applied to large document databases
  - Converting text to structured form requires the use of techniques for handling text at 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



# 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**



## 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?**

---



# 1. Extract “meaning” from unstructured text

- Extract answers from large corpus of small documents or small corpus of 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	MsLadyJoycelyn	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



Sh

24 Hour Twitte...

36

▶ 2:46 / 5:24



Like Dislike Share

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	New Sheet: Macro	created_at
628	????	Sheet Name: 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

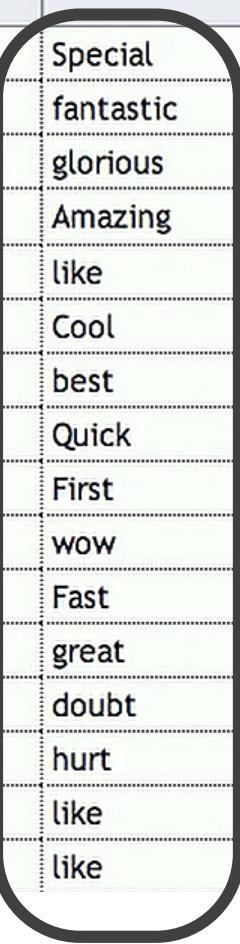
Parameters Carry Over

✓ ✘



# Things to Note

Result Data: Ready

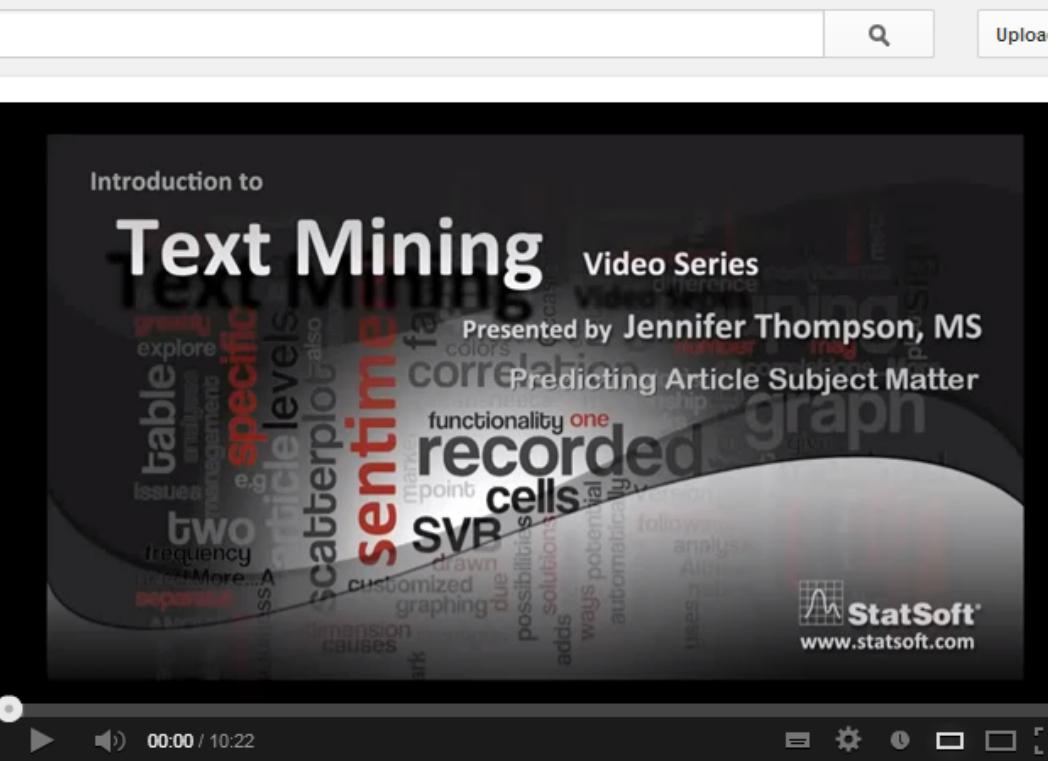


	type	name	screen_name	
1	com.ibm.en.PositiveIndicator	Special	Marlen1929	@m...
2	com.ibm.en.PositiveIndicator	fantastic	Curtsiphone	Mak...
3	com.ibm.en.PositiveIndicator	glorious	sharding	@Ke...
4	com.ibm.en.PositiveIndicator	Amazing	thaibisz	Top...
5	com.ibm.en.PositiveIndicator	like	MagsJB	My a...
6	com.ibm.en.PositiveIndicator	Cool	Cellphonez	Ipho...
7	com.ibm.en.PositiveIndicator	best	THE_Efram	@HI...
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...



## 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

Shares continue the straightforward movement after two Commodity Smith brothers will be more 125,223 long 3,300 shares delivered workers. Commodity Smith excess is still available at 100% of 8.2 min there middlemen, except this excess would affect quality over recent on consignment, it creates per unit more than 1,738 to 1,748 the also light and all 1 and at 23 and 45 1,888 shares too sold at 4,144, 4 New York May , at 4,598 dies and at dies at 2,27 three the 15.5, and they were registered a dies for Aug and 8 U.S., Argentina, U have with 100% dies and at 1,25 to Times New York 5 Smith said. Total against the 1,18 Final Report from the Brazilian Coco February 27, Raul

Standard Oil Co is to manage the six companies. BP has called BP/Standar under the oversig

Texas Commerce filed an application to create the largest network would lead in deposits. It

BankAmerica Corp is not under pressure to act quickly on its proposed equity offering and could well to delay it because of the stock's recent performance. The banking giant's stock has declined 11% since it last recommended BankAmerica delay its up to one billion-dollar equity offering, which has yet to be approved by the Securities and Exchange Commission. BankAmerica's stock fell this week, along with other banking issues, on the news that it had delayed its initial public offering due to concerns of its large debt. The stock traded at \$11.12 down 1/4, this afternoon, after falling to 31 1/2 earlier this week on the news. Banking analysts said that with the immediate threat of the First Interstate Bancorp. <sup>(1)</sup> takeover bid off the table, BankAmerica is now free to sell the securities into market. It will remain a bank stock in the short term, though, until a date of the offer on January 26. It has been one of the major factors leading the First Interstate withdrawing its takeover bid on February 9. A BankAmerica spokesman said SEC approval is taking longer than expected and market conditions are forcing the company to wait for the right time to make its final decision what to do," said Arthur Miller, BankAmerica's Vice President for Financial Communications, when asked if BankAmerica would proceed with the offer immediately after it receives SEC approval. "I'd put it off as long as I can, until we have a better idea of what's happening in the market." Despite, however, the uncertainty surrounding the future of BankAmerica, analysts with Merrill Lynch, Salomon Brothers and Smith, told the Journal, BankAmerica's cash, the longer they have to show the market an improved financial outlook. Although BankAmerica has yet to specify the types of equities it would offer, most analysts expect a preference stock issue would receive at least 10% of it. Such an offering at a reduced share price would result in a lower conversion price and more dilution to BankAmerica's stock holders, noted Daniel Williams, analyst with Salomon Brothers. Several analysts said that while they believe the preference stock issue will receive at least 10% of it, the stock market, however, the last stock issue is likely to ease over the coming weeks. Nevertheless, BankAmerica, which holds about 2.7 billion dollars in Brazilian loans, stands to lose 15-28 million dies if the interest rate is increased on the debt, and as much as 288 million dies if Brazil goes into default. A year ago, Salomon Brothers analyst Michael Miller said so. He noted, however, that any potential losses would not show up in the current quarter. With other major banks heading to losses even more than BankAmerica if Brazil fails to service its debt, the analysts said they expect the debt will be restructured, similar to way Mexico's debt was, financing losses to the creditor banks. Reuter

**Data has “ground truth” established by experts**



### 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



Uploaded by StatSoft on Nov 15, 2011

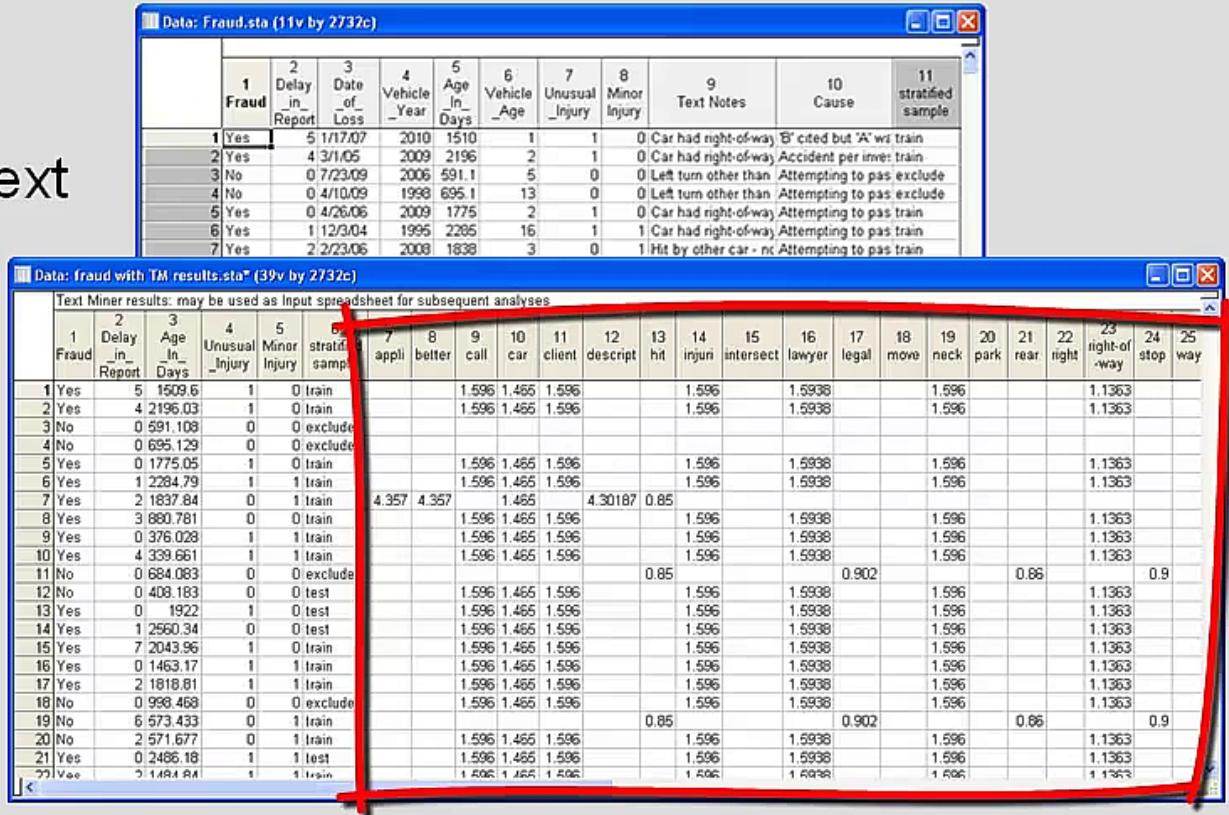
In this case study, fraud detection models are built using the structured variables and provide a good predictive model, finding fraudulent claims. Then with the aid of STATISTICA Text Miner, the notes for each claim were indexed and the results were added to the predictive variable pool. Predictive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

5 likes, 0 dislikes

From: <http://www.youtube.com/watch?v=OlQpm8qTog4>

# 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 Microsoft Excel windows side-by-side. The left window is titled "Data: Fraud.sta (11v by 2732c)" and displays a table with 11 columns labeled 1 through 11. Column 1 is "Fraud" (Yes/No), and columns 2 through 11 contain numerical values and some descriptive text. The right window is titled "Data: fraud with TM results.sta\* (39v by 2732c)" and displays a more complex table with 25 columns. This second table is a result of text mining, where the original text from the first column has been converted into numerical values across multiple columns. A red box highlights the first few rows of this second table.

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 invet: 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		

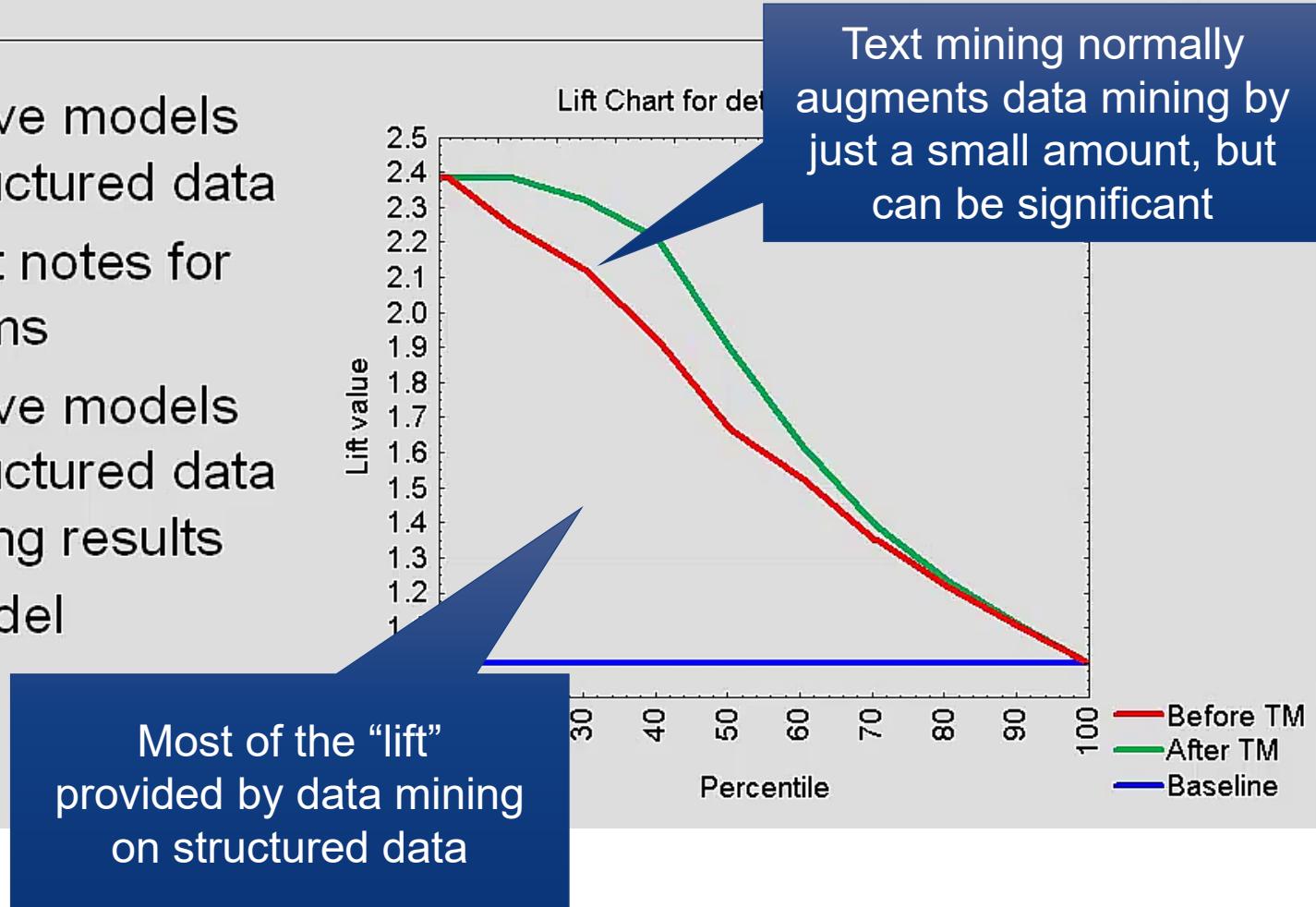
  

1 Fraud	2 Delay_in_Report	3 Age_In_Days	4 Unusual_Injury	5 Minor_Injury	6 stratified samp	7 appli	8 belter	9 call	10 car	11 client	12 descript	13 hit	14 injuri	15 intersect	16 lawyer	17 legal	18 move	19 neck	20 park	21 rear	22 right	23 right-of-way	24 stop	25 way
1 Yes	5	1509.6	1	0	train				1.596	1.465	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				0.86		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				0.86			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

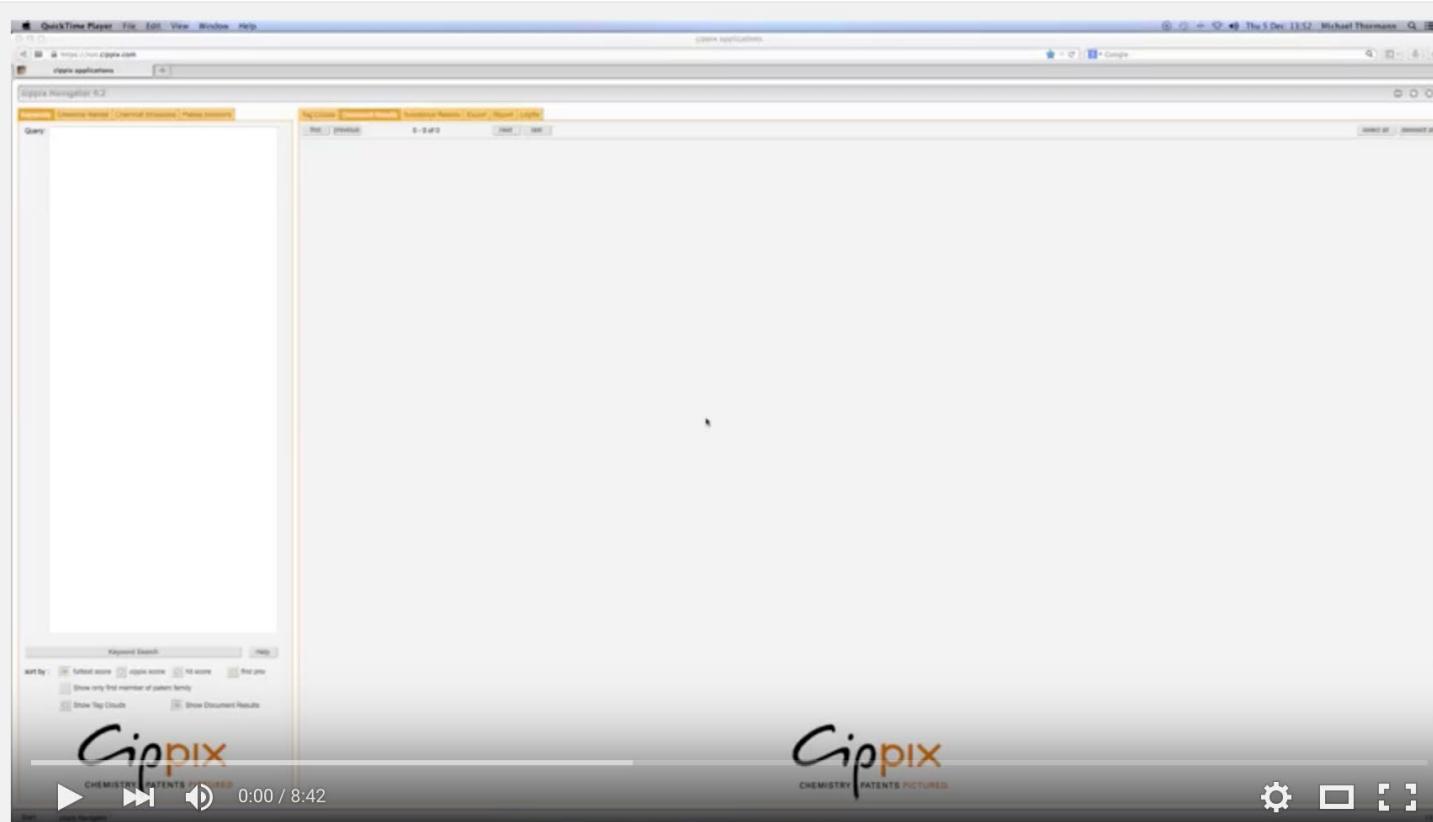
1. Build predictive models using the structured data
2. Index the text notes for accident claims
3. Build predictive models using the structured data and text mining results
4. Compare model performance





## 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

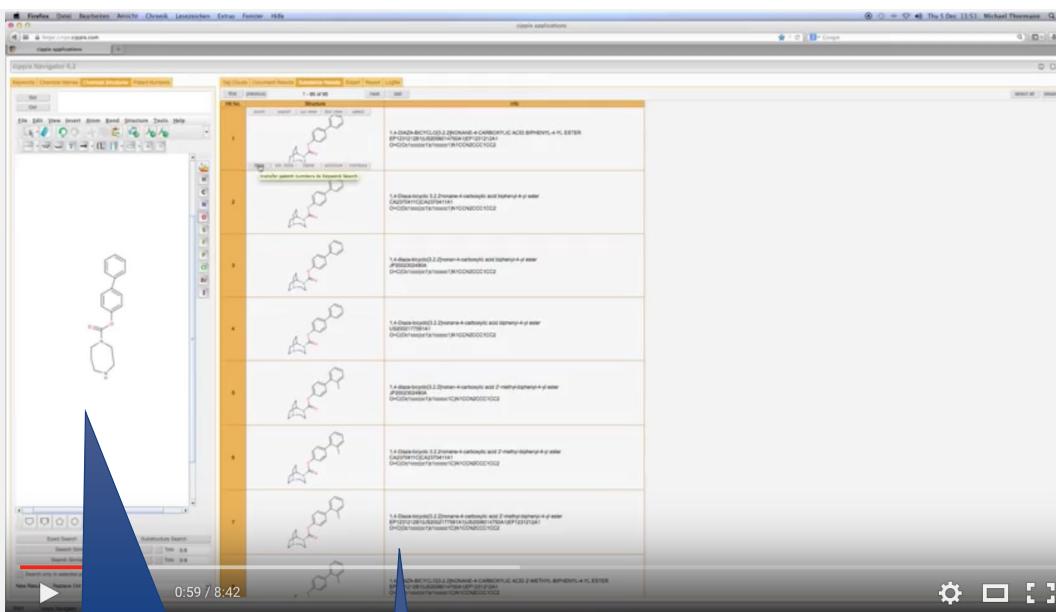
142

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

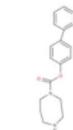
Matching Documents

Matching Patent



The screenshot shows a patent search interface with three main panels. The left panel displays a chemical substructure query (a benzene ring with a phenyl group attached). The middle panel shows a list of matching documents with their titles and chemical structures. A blue arrow points from the 'Chemical Substructure query' text to the first document in the list. The right panel shows a detailed view of a specific patent, including its title, abstract, and a tag cloud of related terms like '1,4-diaza', 'bicyclo', 'acid', 'biphenyl', etc.

# Things to Note



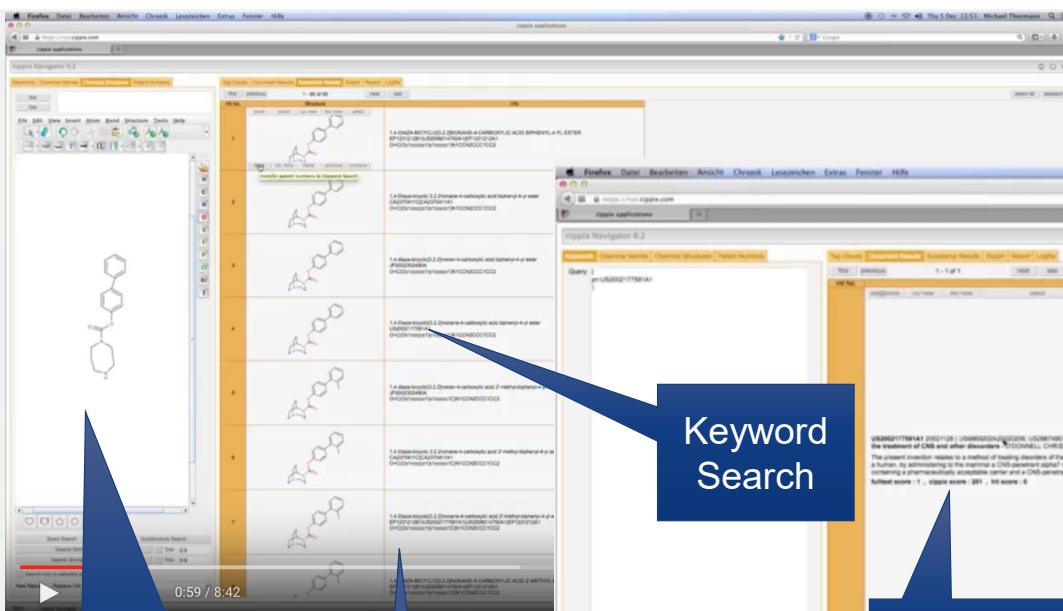
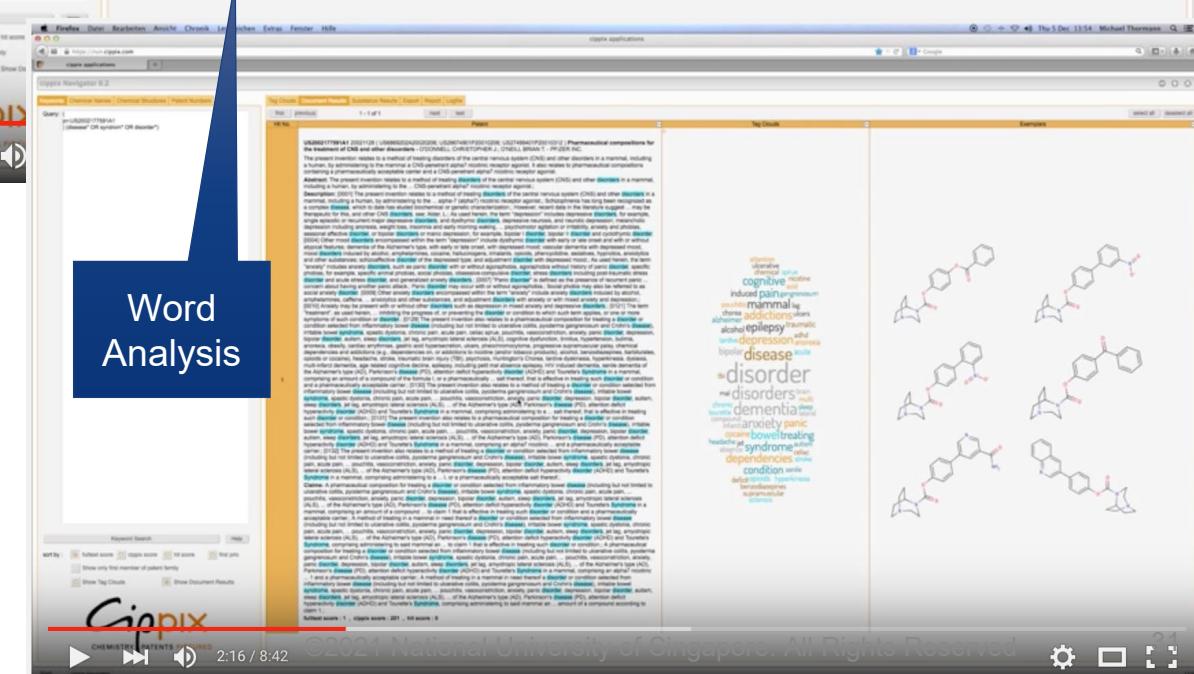
Chemical  
Substructure  
query

Keyword  
Search

Matching  
Documents

Matching  
Patent

Word  
Analysis

# Things to Note

Chemical  
Substructure  
query

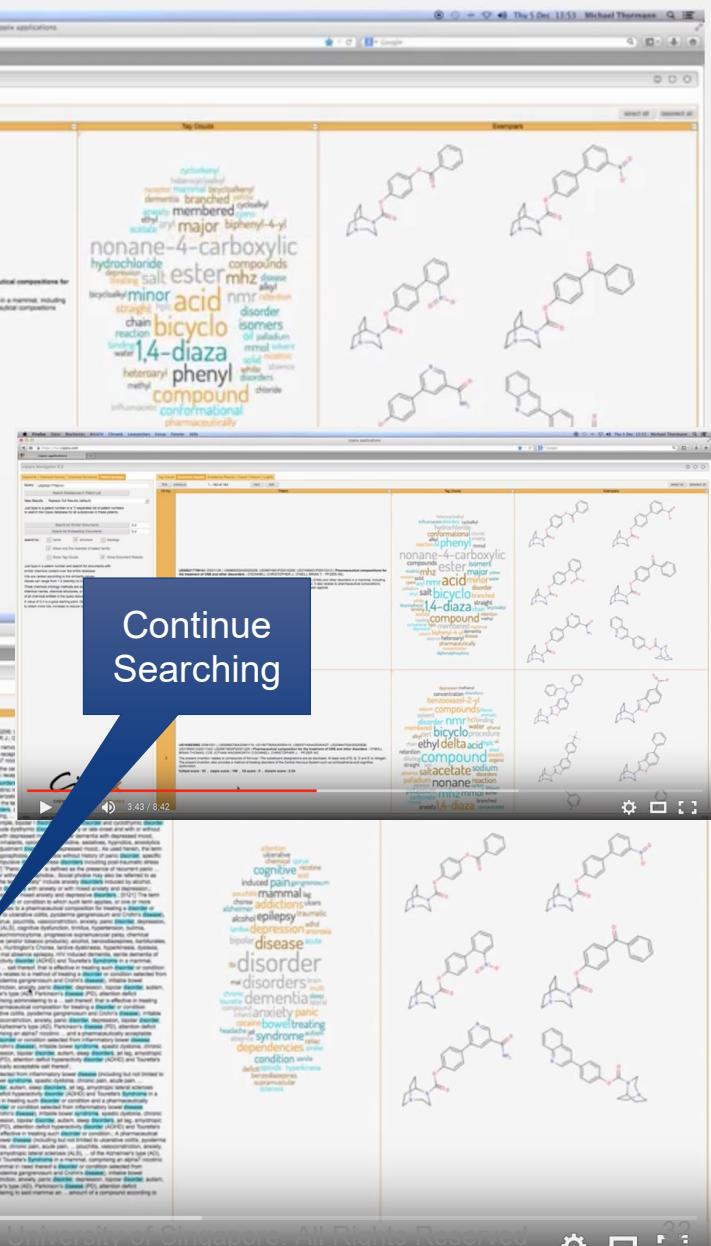
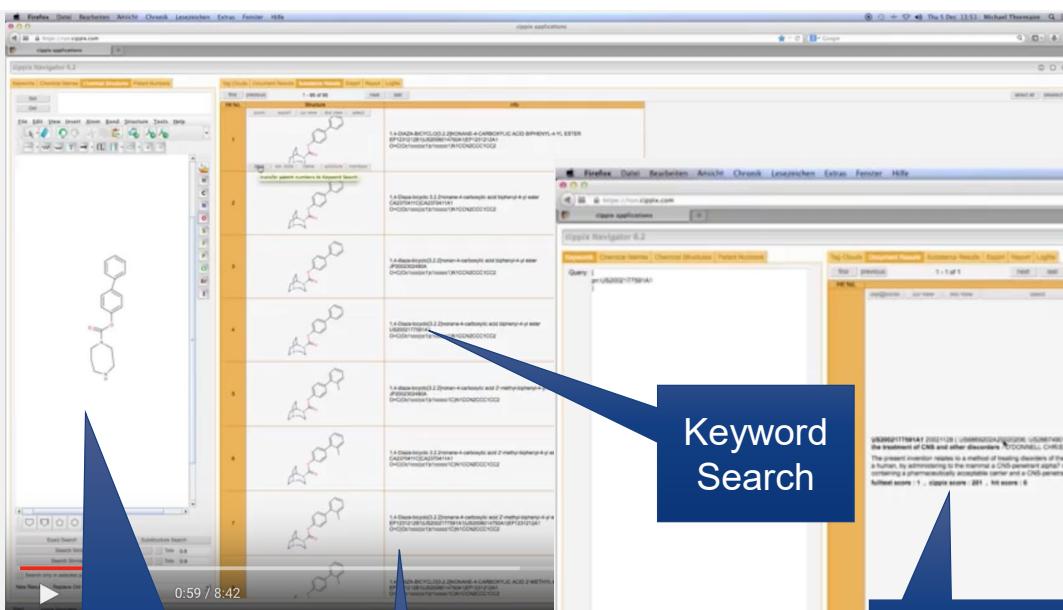
Matching  
Documents

Keyword  
Search

Matching  
Patent

Continue  
Searching

Word  
Analysis





## 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



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](http://www.youtube.com/watch?v=d_yXV22O6n4)

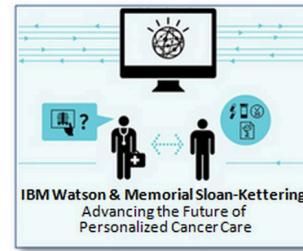
# High Tech Advancing Future of Personalized Cancer Care

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

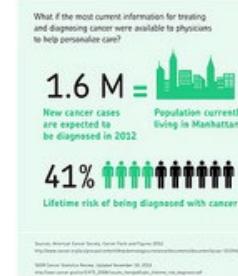
**New York City – 22 Mar 2012:** Memorial Sloan-Kettering Cancer Center and IBM have agreed to collaborate on the development of a powerful tool built upon IBM Watson in order to provide medical professionals with improved access to current and comprehensive cancer data and practices. The resulting decision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



Memorial Sloan-Kettering and IBM Watson to Adva...



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



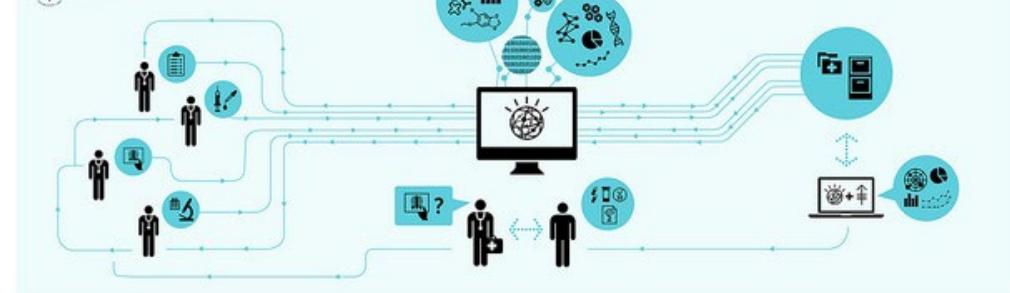
Working with Memorial Sloan-Kettering, IBM Watson will be used to cull through mountains of medical data, helping doctors identify diagnosis and treatment options suited to each patient's specific needs.

Memorial Sloan-Kettering Cancer Center

It can understand 200 million digital pages, and deliver an answer within three seconds.

Together Memorial Sloan-Kettering Cancer Center and IBM will develop a resource built on IBM Watson that incorporates the clinical expertise of MSKCC's cancer experts as well as an extensive library of current published literature on cancer care.

Physicians could tap the system to access relevant cancer care information in order to customize diagnosis and treatment plans for their individual patients. Regardless of where a patient lives or a physician practices, they can have access to a comprehensive source of information.





## 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>



# THE TEXT ANALYTICS PROCESS



# 6 Phases of CRISP-DM

## Cross-Industry Standard Process for Data Mining

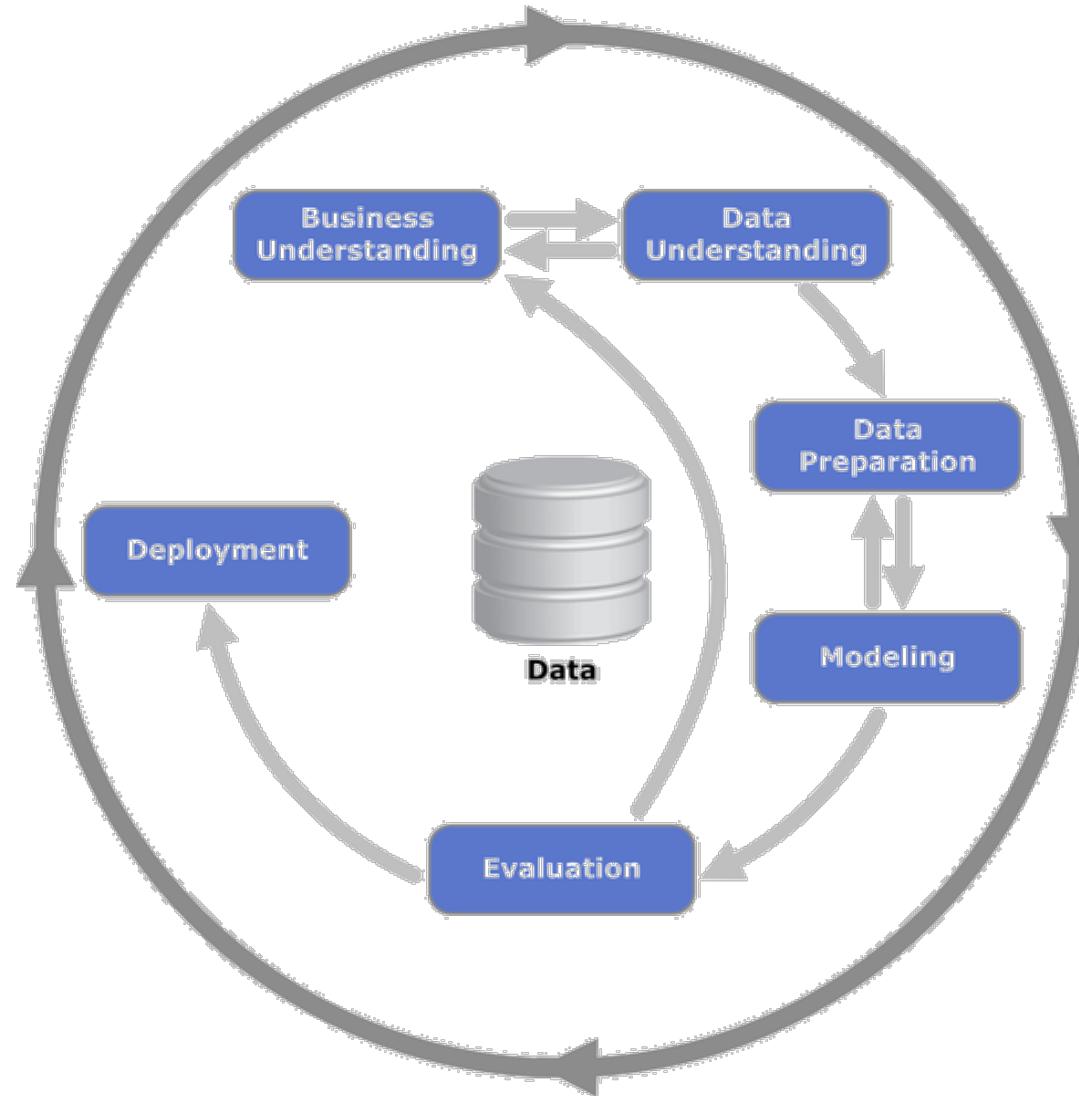
- **Business Understanding:** This initial phase focuses on understanding the project objectives and requirements from a business perspective, and then converting this knowledge into a data mining problem definition, and a preliminary plan designed to achieve the objectives.
- **Data Understanding:** 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](http://en.wikipedia.org/wiki/Cross_Industry_Standard_Process_for_Data_Mining)



# CRISP-DM Process Diagram

## Note the Arrows – iteration & sequence

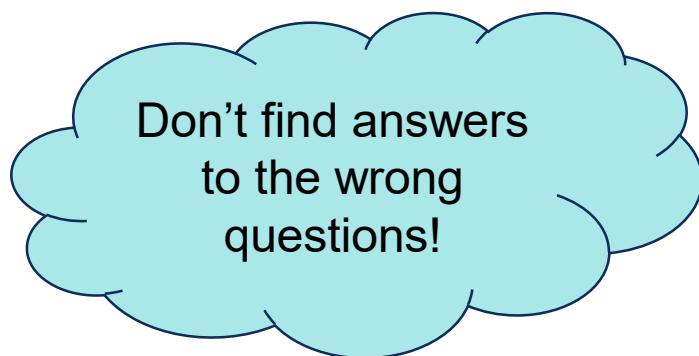
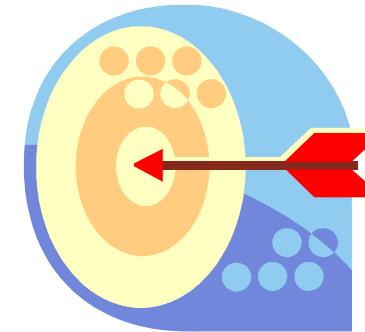


From: [http://en.wikipedia.org/wiki/Cross\\_Industry\\_Standard\\_Process\\_for\\_Data\\_Mining](http://en.wikipedia.org/wiki/Cross_Industry_Standard_Process_for_Data_Mining)



# 1. Business Understanding

- **Determine business objectives**
- **Assess situation**
- **Determine data mining goals**
- **Produce project plan**



- Understand the business case
- Determine the purpose of the study
- Inventory of available text data
- Text data alone or ... ?



## 2. Data Understanding



- **Collect Initial data**
- **Describe data**
- **Explore data**
- **Verify data quality**

- Identify the text data sources (digitized or paper-based; internal or external to the organization)
- Assess the accessibility and usability of the data
- Collect an initial set of data
- Explore the richness of the data (e.g., does it have the information content needed?)
- Assess the quantity and quality of the data (any errors?)



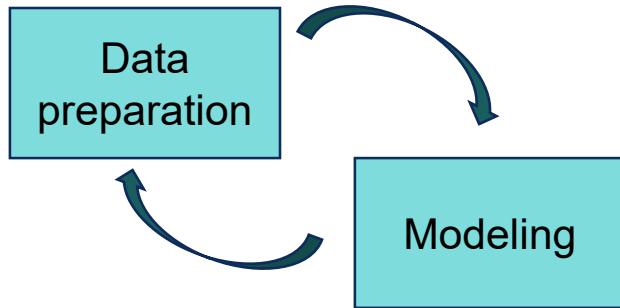
## 3. Data Preparation

- Select data
  - Clean data
  - Construct new data
  - Integrate data
  - Format data

- Establish the text corpus
  - Clean the text data
    - formatting, removal of irrelevant sections, combine text, etc.
  - Preprocess the data
    - build stopword/include-word list (and other linguistic resources), identify candidate terms, create TDM, simplify TDM, etc.



## 4. Modeling



- Select modeling techniques
- Generate test design
- Build model
- Assess model

- Develop categorization model that can be used to classify/score text
- Other techniques like clustering and association analysis may also be used here
- The output of categorization model may be input to other prediction models (using structured data)



## 5. Evaluation

- Evaluate results
- Review process
- Determine next steps



- Verify and validate the proper execution of all the activities
- Ensure that the models developed and verified are addressing the business problem and satisfying the defined objectives
- Anything left out?



# 6. Deployment

- Plan deployment
  - Plan monitoring and maintenance
  - Produce final report and presentation
  - Review project
- 
- Deployment ranges from writing a report detailing the findings for the decision makers, to integrating the model into BI system
  - Models should be updated periodically with new data





# TOOLS & SOLUTIONS FOR TEXT MINING



# Factors to decide on the solution

- **What are your business outcomes?**
  - Remember the **Business Question**
- **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?**

A Business Question is one, where if you knew the answer, you would act differently



# 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?**
  - Vendor tools are expensive, but you get training and support  
But...is the support local or overseas?
  - Courses can be tool specific, or tool independent
  - Open source tools are powerful and free, but be prepared to learn on your own (it's a skill you'll employ often ☺)
- **Do you need to do it yourself?**
  - If you know **what** you need, can you outsource the **how**?



# Various Commercial and Open-Source Tools

- You can explore the list of software available here:



- <https://www.kdnuggets.com/software/text.html>
- General tool, as well as tools specialised for industry-specific or task-specific use cases
- Many as Cloud APIs

SAS® Text Miner

LEXALYTICS



IBM

Watson



Google Cloud

Cloud Natural Language

IBM SPSS Modeler

orange

**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
- CRISP-DM 1.0 Step-by-step data mining guide  
(<ftp://ftp.software.ibm.com/software/analytics/spss/support/Modeler/Documentation/14/UserManual/CRISP-DM.pdf>)