

Introduction to Text Mining

Hongning Wang

CS@UVa

What is “Text Mining”?

- “Text mining, also referred to as **text data mining**, roughly equivalent to text analytics, refers to the process of deriving high-quality information from text.” - wikipedia
- “Another way to view text data mining is as a process of **exploratory** data analysis that leads to **heretofore unknown** information, or to answers for questions for which the answer is not currently known.” - Hearst, 1999

Two different definitions of mining

- Goal-oriented (effectiveness driven)
 - Any process that generates useful results that are non-obvious is called “mining”.
 - Keywords: “**useful**” + “**non-obvious**”
 - Data isn’t necessarily massive
- Method-oriented (efficiency driven)
 - Any process that involves extracting information from massive data is called “mining”
 - Keywords: “**massive**” + “**pattern**”
 - Patterns aren’t necessarily useful

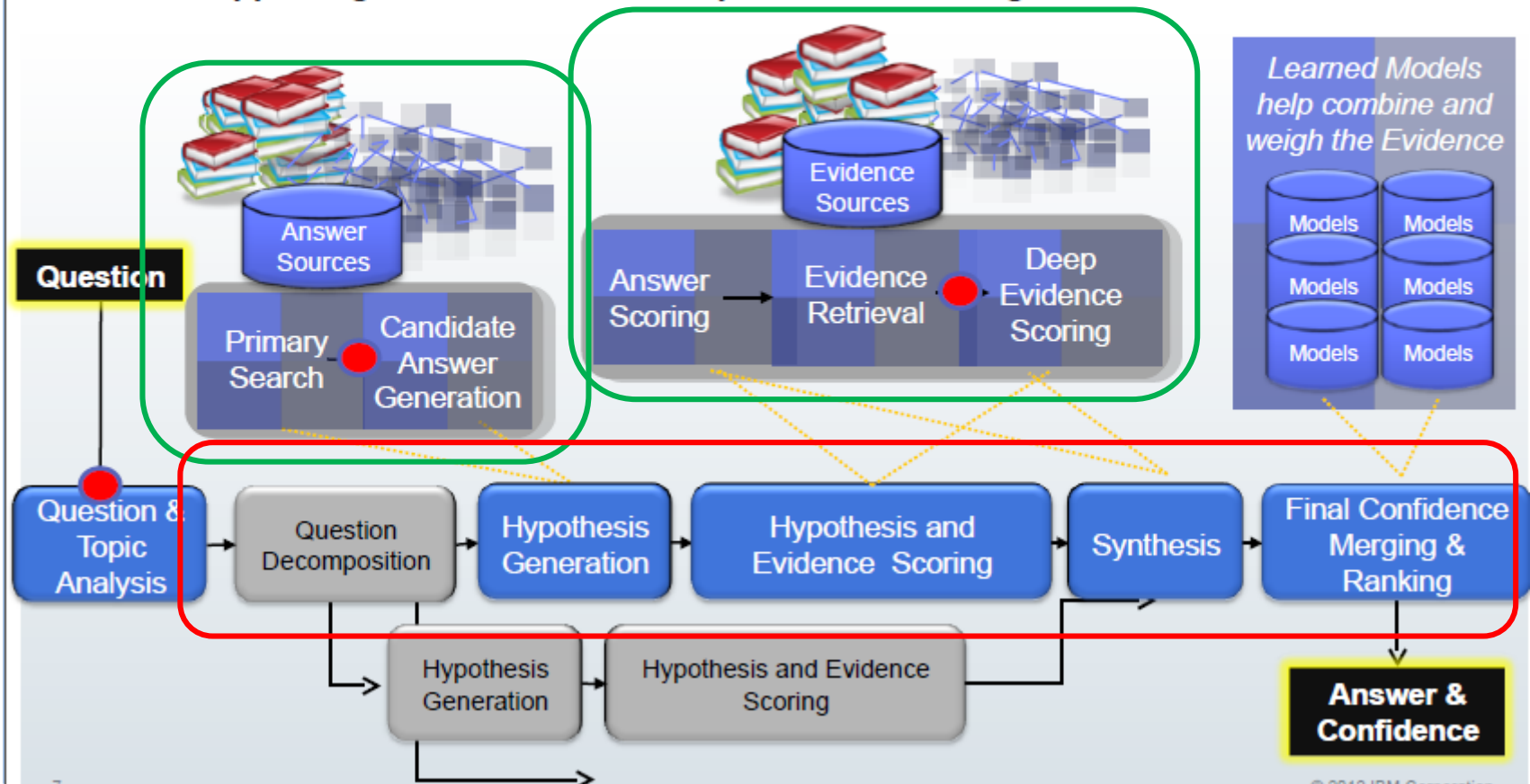
Knowledge discovery from text data

- IBM's Watson wins at Jeopardy! - 2011



An overview of Watson

- On questions, at the start of question analysis
- On primary search results, before candidate answer generation
- On supporting evidence, before deep evidence scoring



What is inside Watson?

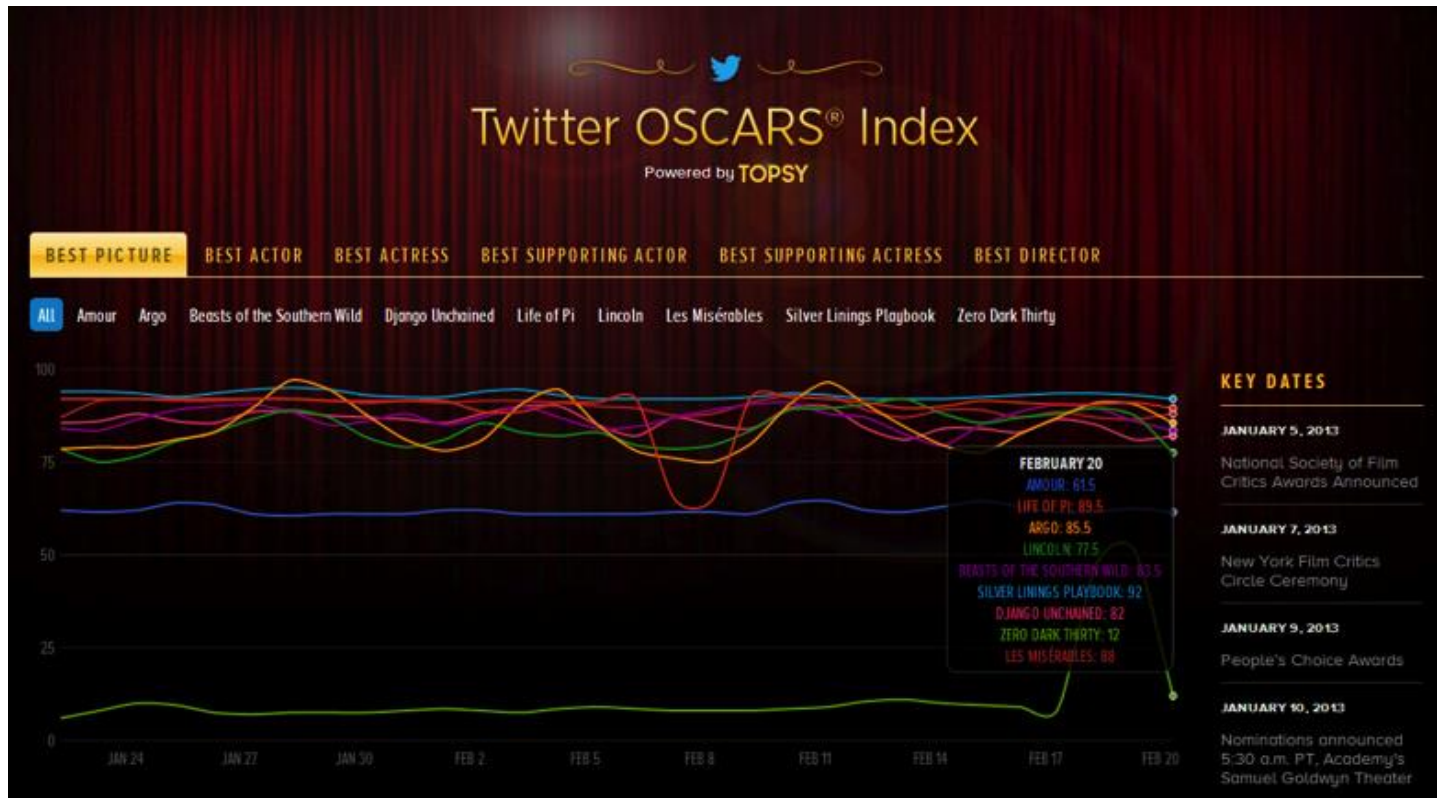
- *“Watson had access to 200 million pages of structured and unstructured content consuming four terabytes of disk storage including the full text of Wikipedia” – PC World*
- *“The sources of information for Watson include encyclopedias, dictionaries, thesauri, newswire articles, and literary works. Watson also used databases, taxonomies, and ontologies. Specifically, DBPedia, WordNet, and Yago were used.” – AI Magazine*

What is inside Watson?

- DeepQA system
 - *“Watson's main innovation was not in the creation of a new algorithm for this operation but rather its ability to **quickly** execute hundreds of proven language analysis algorithms simultaneously to find the correct answer.”* – New York Times
 - [The DeepQA Research Team](#)

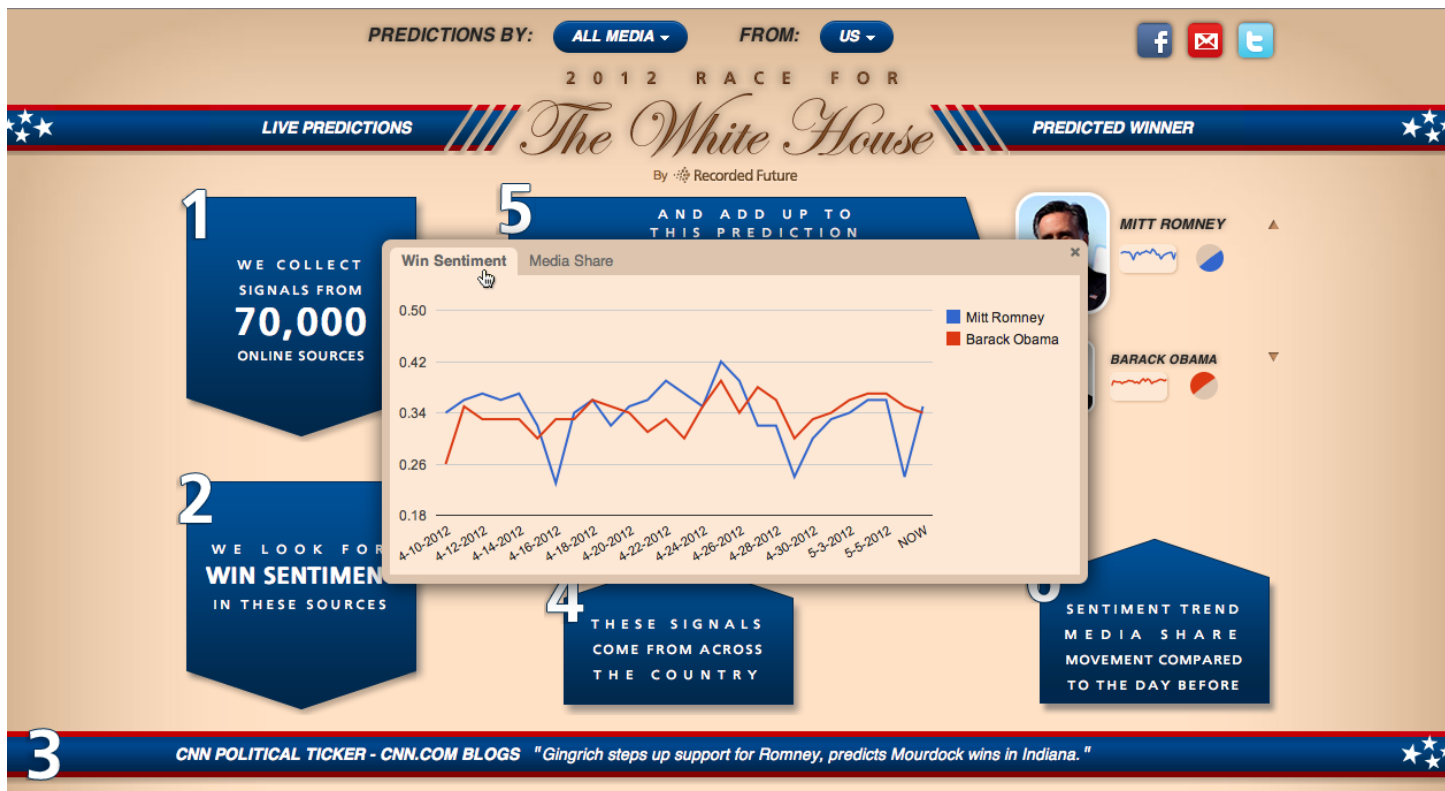
Text mining around us

- Sentiment analysis



Text mining around us

- Sentiment analysis



- Document summarization



Text mining around us

- Document summarization

The image shows a Bing search results page for the query 'text mining'. The search bar at the top contains the text 'text mining' and a magnifying glass icon. Below the search bar, there are tabs for 'Web', 'Images', 'Videos', 'Maps', 'News', and 'More'. The 'Web' tab is selected. The search results show 19,200,000 results, filtered by 'Any time'. The first result is from Wikipedia, titled 'Text mining - Wikipedia, the free encyclopedia'. The snippet for this result is highlighted with a red box. Below it are two more results: 'Text Mining (Big Data, Unstructured Data)' from statsoft.com and 'Text Mining' from academic.research.microsoft.com. On the right side of the page, there is a 'Text mining' knowledge panel, also with a red box around its snippet. Below the panel are 'Related people' and 'People also search for' sections. At the bottom right, there are 'Related searches' for 'Text Analysis Software' and 'Text Analytics'.

bing text mining

Web Images Videos Maps News More

19,200,000 RESULTS Any time ▾

Text mining - Wikipedia, the free encyclopedia
en.wikipedia.org/wiki/Text_mining ▾
Text mining, also referred to as text data mining, roughly equivalent to text analytics, refers to the process of deriving high-quality information from text. High ...
[Text mining and text ...](#) · [History](#) · [Text analysis processes](#) · [Applications](#)

Text Mining (Big Data, Unstructured Data)
www.statsoft.com/Textbook/Text-Mining ▾
Text Mining Introductory Overview. The purpose of Text Mining is to process unstructured (textual) information, extract meaningful numeric indices from the text, ...

Text Mining
academic.research.microsoft.com/Keyword/41731/text-mining ▾
Text mining is defined as knowledge discovery in large text collections. It detects interesting patterns such as clusters, associations, deviations, similarities, and ...

What is **text mining** (text analytics)? - Definition from ...
searchbusinessanalytics.techtarget.com/definition/text-mining ▾
Text mining is the analysis of data contained in natural language text. The application of text mining techniques to solve business problems is called text analytics.

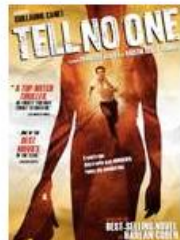
Text mining
Text mining, also referred to as text data mining, roughly equivalent to text analytics, refers to the process of deriving high-quality information from text. High-quality information is typically derived through the devising of patterns and trends through means such as statistical pattern learning. Text mining usually involves the process of struct... +
en.wikipedia.org
Related people: Jun'ichi Tsujii · Alfonso Valencia · Tomoko Ohta · Carol Friedman · Michael Bery · Hsinchun Chen
People also search for: Sentiment analysis · Natural language processing · Web mining · Analytics · Cluster analysis +
Data from: Wikipedia · [Freebase](#)
[Feedback](#)

Related searches
[Text Analysis Software](#)
[Text Analytics](#)

Text mining around us

- Movie recommendation

FOREIGN SUGGESTIONS (about 104) [See all >](#)



Tell No One

Because you enjoyed:
Memento
Syriana
Children of Men



Let the Right One In

Because you enjoyed:
Seven Samurai
This Is Spinal Tap
The Big Lebowski



I've Loved You So Long

Because you enjoyed:
The Queen
Syriana
Good Night, and Good Luck

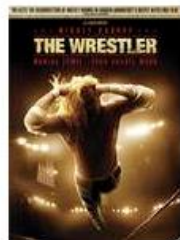


Downfall

Because you enjoyed:
Das Boot
The Killing Fields
Seven Samurai



DRAMA SUGGESTIONS (about 82) [See all >](#)



The Wrestler

Because you enjoyed:
Sin City
Reservoir Dogs
The Big Lebowski



The Visitor

Because you enjoyed:
Gandhi
The Motorcycle Diaries
The Queen



Brick

Because you enjoyed:
The Big Lebowski
Rushmore
Fight Club



The Pianist

Because you enjoyed:
Amadeus
The Killing Fields
Empire of the Sun



Text mining around us

- Restaurant/hotel recommendation

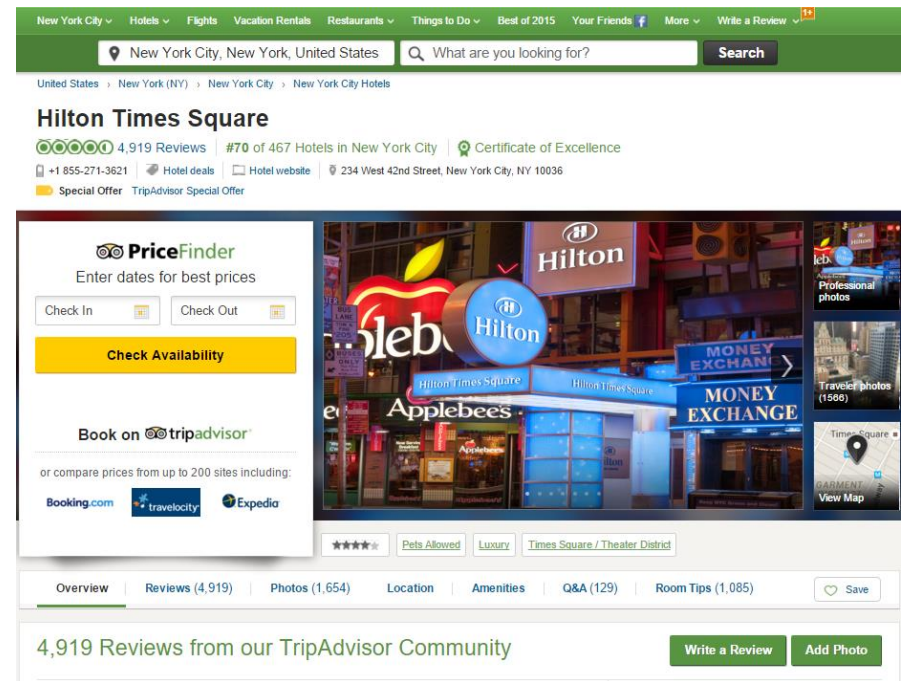


The image shows a Yelp page for Bodo's Bagels. The header includes the Yelp logo, a search bar with the text "Find tacos, cheap dinner, Max's", and a location filter set to "Near Charlottesville, VA". The page title is "Bodo's Bagels" with a 4.5-star rating and 186 reviews. Below the title is a map showing the location at 1418 Emmet St N, Charlottesville, VA 22903. To the right of the map are three photos of bagels: one with cream cheese, one with turkey, lettuce, and pesto, and one with a meaty center. Below the photos are three reviews. The first review says "Almost any combination of bagel, cream cheese or spread or sandwich you could dream of you can find at Bodos." in 38 reviews. The second review says "A few favorite items would include the Everything bagel with the Deli Egg which has a tasty meaty center encased in steaming hot eggs." in 4 reviews. The third review says "There's a reason why Bobo's has been in business since well before I was a UVA." in 10 reviews. At the bottom, there is a "Recommended Reviews" section with a search bar and a "Sort by Highest Rated" button.

Bodo's Bagels
186 reviews
\$ · Bagels, Breakfast & Brunch, Sandwiches
1418 Emmet St N
Charlottesville, VA 22903
(434) 977-9598
Message the business
bodosbagels.com

Reviews:
"Almost any combination of bagel, cream cheese or spread or sandwich you could dream of you can find at Bodos." in 38 reviews
\$0.60 Cream Cheese
"A few favorite items would include the Everything bagel with the Deli Egg which has a tasty meaty center encased in steaming hot eggs." in 4 reviews
"There's a reason why Bobo's has been in business since well before I was a UVA." in 10 reviews

Hours:
Mon 6:30 am - 8:00 pm
Tue 6:30 am - 8:00 pm
Wed 6:30 am - 8:00 pm
Thu 6:30 am - 8:00 pm
Fri 6:30 am - 8:00 pm
Sat 7:00 am - 8:00 pm
Sun 8:00 am - 4:00 pm



The image shows a TripAdvisor page for Hilton Times Square. The header includes the TripAdvisor logo, a search bar with the text "What are you looking for?", and a location filter set to "New York City, New York, United States". The page title is "Hilton Times Square" with a 4.5-star rating and 4,919 reviews. Below the title is a map showing the location at 234 West 42nd Street, New York City, NY 10036. To the right of the map are three photos of the hotel: one showing the hotel entrance, one showing the hotel lobby, and one showing the hotel room. Below the photos are three reviews. The first review says "Hilton Times Square is a great hotel with a great location." in 10 reviews. The second review says "Hilton Times Square is a great hotel with a great location." in 10 reviews. The third review says "Hilton Times Square is a great hotel with a great location." in 10 reviews. At the bottom, there is a "Recommended Reviews" section with a search bar and a "Sort by Highest Rated" button.

Hilton Times Square
4,919 Reviews | #70 of 467 Hotels in New York City | Certificate of Excellence
+1 855-271-3621 | Hotel deals | Hotel website | 234 West 42nd Street, New York City, NY 10036
Special Offer | TripAdvisor Special Offer

PriceFinder
Enter dates for best prices
Check In | Check Out
Check Availability


Book on TripAdvisor
or compare prices from up to 200 sites including:
Booking.com | travelocity | Expedia

Reviews:
"Hilton Times Square is a great hotel with a great location." in 10 reviews
"Hilton Times Square is a great hotel with a great location." in 10 reviews
"Hilton Times Square is a great hotel with a great location." in 10 reviews

Text mining around us


- News recommendation

[All Stories](#) [News](#) [Entertainment](#) [Sports](#) [Business](#) [More](#) ▼




Flying high: Airstream can't keep up with demand
JACKSON CENTER, Ohio (AP) — Bob Wheeler still gets the question sometimes when people find out he runs the company that builds those shiny aluminum campers: "Airstreams? They still make those?"
[Associated Press](#)

North Korea's Internet down again. US spooks at work?
North Korea's web connection to the rest of the world — always sketchy and limited at best — went on the blink again Saturday. Most North Koreans wouldn't have noticed, of course. But
[Christian Science Monitor](#) 45 mins ago



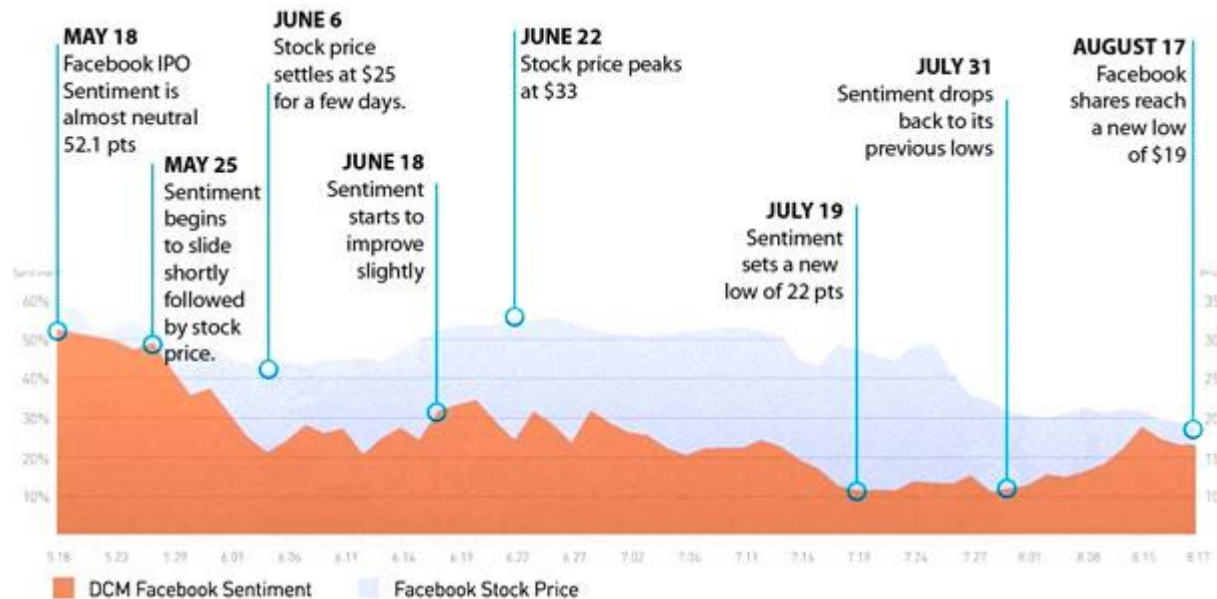
Wisconsin man keeps 40-year-old Christmas tree up until son returns
By Brendan O'Brien (Reuters) - A Wisconsin man will refuse for about the 40th time to partake in the annual after-holiday chore of putting Christmas
[Reuters](#)



Navy Helicopter Drone Completes First Round of Testing
Imagine trying to land a remote-controlled helicopter on top of a motorboat that's speeding across a lake. Navy pilots recently had to contend with just such a scenario as they tested the U.S. military's newest drone, the MQ-8C
[LiveScience.com](#)

Text mining around us

- Text analytics in financial services



Text mining around us

- Text analytics in healthcare

REQUEST FOR MEDICAL/DENTAL RECORDS		DATE
1. PATIENT (Last Name - First Name - Middle Name)		December 20, 1989
2. NATIONAL PERSONNEL RECORDS CENTER (Military Personnel Records) 9700 Pappe Boulevard St. Louis, Missouri 63132		RCMP File
3. TO:	4. SERVICE NO.(S)	5. GRADE OR RATE
Commander V.A. Air Force Hospital Scott AFB, Texas		A 2/c
6. VA CLAIM NUMBER		
7. ORGANIZATION AND PLACE OF TREATMENT	8. DATES OF TREATMENT (mm/dd)	9. DISEASE OR INJURY
Your Hospital	1-23-61 to 3-28-61	Kidney operation
10. RECORDS REQUESTED <input type="checkbox"/> CLINICAL <input type="checkbox"/> OUTPATIENT <input type="checkbox"/> HEALTH RECORD <input type="checkbox"/> DENTAL RECORD <input type="checkbox"/> X-RAY <input checked="" type="checkbox"/> MEDICAL REPORT CARDS, EMERGENCY MEDICAL TAGS, FIELD MEDICAL CARDS OTHERS (See remarks)	11. REMARKS Forward records to address in item 13, below	
12. SIGNATURE <i>[Signature]</i> DATE 12/17/89 NATIONAL PERSONNEL RECORDS CENTER (MILITARY) ST. LOUIS, MO 63132 RTY <i>B. Day</i>		
13. TO: VARD 1000 Liberty Avenue Pittsburgh, PA 15222		
14. ACTION TAKEN <input type="checkbox"/> AVAILABLE RECORDS ENCLOSED <input type="checkbox"/> NO RECORDS ON FILE		
15. ENCLOSURES (Number of) <input type="checkbox"/> CLINICAL <input type="checkbox"/> OUTPATIENT <input type="checkbox"/> HEALTH RECORD <input type="checkbox"/> DENTAL RECORD <input type="checkbox"/> X-RAY <input type="checkbox"/> MEDICAL REPORT CARDS, EMERGENCY MEDICAL TAGS, FIELD MEDICAL CARDS OTHERS (See remarks)		
16. REMARKS		
17. DATE		
18. SIGNATURE		

NATIONAL ARCHIVES AND RECORDS ADMINISTRATION RA FORM 13042-a (9-85)

WebMD-moderated WebMD® Heart Disease Community

Home
Discussions
Tips
Resources
About This Community
Staying Informed
My Watchlist
Related Men's Health Communities
All Communities
Community FAQs
Crisis Assistance

Stay Informed with Newsletters

Sign up for the Heart Health newsletter and keep up with all the latest news, treatments, and research with WebMD.

☐ I have read and agree to WebMD's Privacy Policy.
Enter Email Address

Sign Up

What's Happening Now

See All Discussions | Tips | Resources

Post Now

Search This Community

GO

Popular Discussions



11 surprising ways to prevent a heart attack

<http://www.foxnews.com/health/2016/01/18/11-surprising-ways-to-prevent-a-heart-attack/>
Chances are you're still riding the New Year's high and you're motivated and committed to eating healthy...

Posted by cardiostarus1

Was this Helpful?

2 of 2 found this Resource helpful

0 Replies

Report This



Reply: Angiogram

Consult with an interventional cardiologist and bring the disc of the angiogram video with you.

Posted by cardiostarus1

3 Replies

INCLUDES EXPERT CONTENT

Reply: Internal Bleeding after heart cath

Could be that there isn't enough in it for the lawyers. My husband lost his leg because a NP who was supposed...

Posted by loveRandy

16 Replies

INCLUDES EXPERT CONTENT

Reply: Trouble Breathing

You need to consult with a doctor. If you don't have the money to pay for it, use the internet to find the...

Posted by smacmill

1 Reply

Report This

Helpful Tips

HOW TO EAT FOR A HEALTHY HEART?

1. Eat food less in fat, much less saturated and trans-fat 2. More servings of fruits and vegetables considering its variety daily and ... More

Was this Helpful?

1 of 1 found this helpful

tip for the pain.

Post a Tip | See All

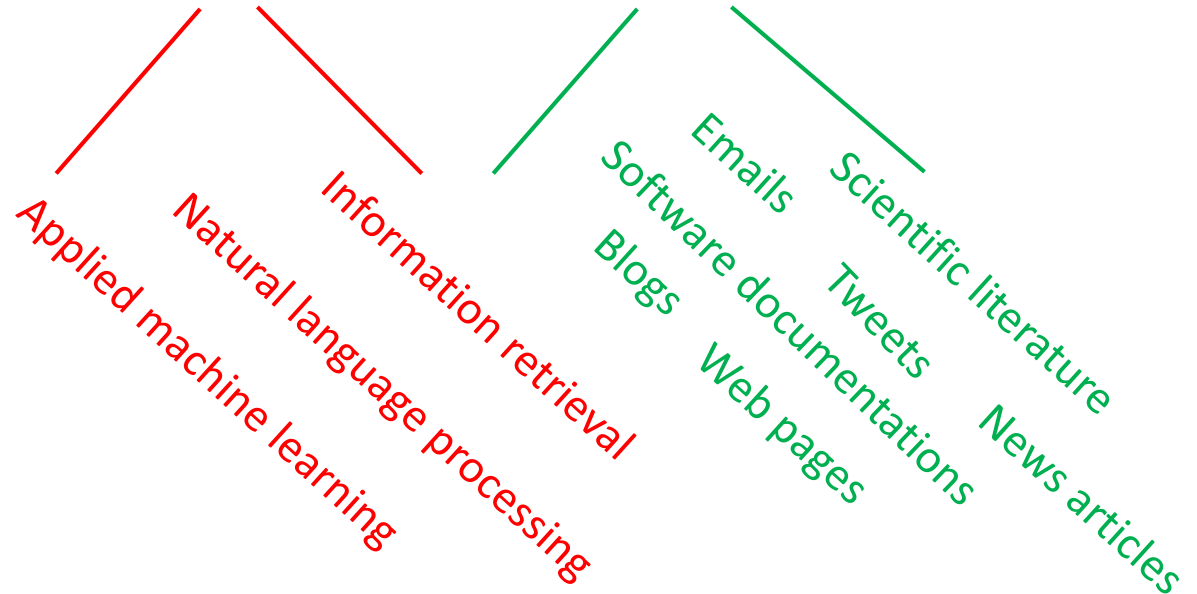
Helpful Resources

- Super-safe iodine may save mil...
- Eating More Fruit Cuts Heart D...
- Heart Attack Treatment: Timing...
- Can heart attack damage be rev...
- Causes of Panic Attacks

Post a Resource | See All

How to perform text mining?

- As computer scientists, we view it as
 - Text Mining = Data Mining + Text Data



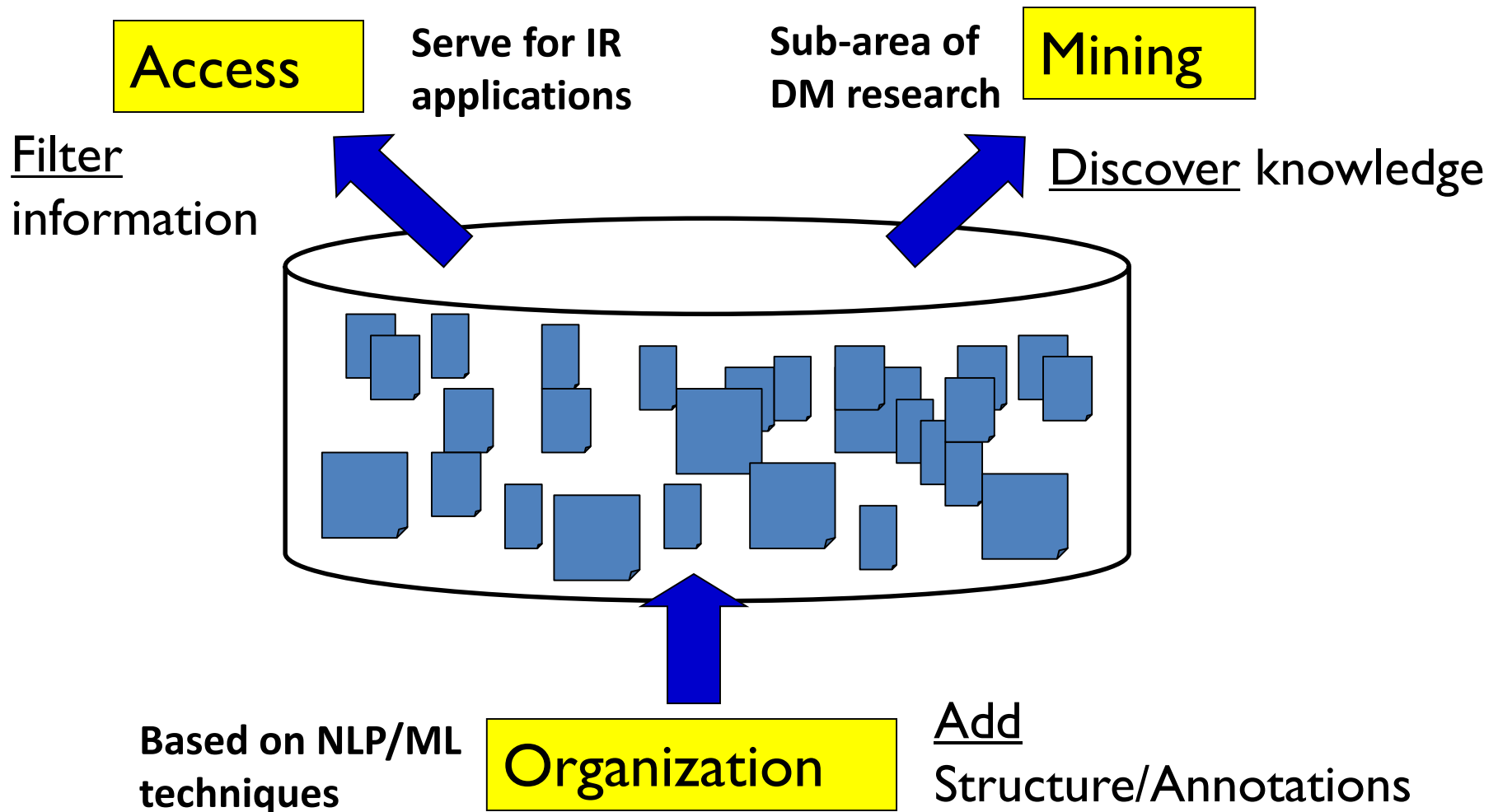
Text mining v.s. NLP, IR, DM...

- How does it relate to data mining in general?
- How does it relate to computational linguistics?
- How does it relate to information retrieval?

	Finding Patterns	Finding “Nuggets”	
		Novel	Non-Novel
Non-textual data	General data-mining	Exploratory analysis	Database queries
Textual data	Comp Ling		Information retrieval

Text Mining

Text mining in general

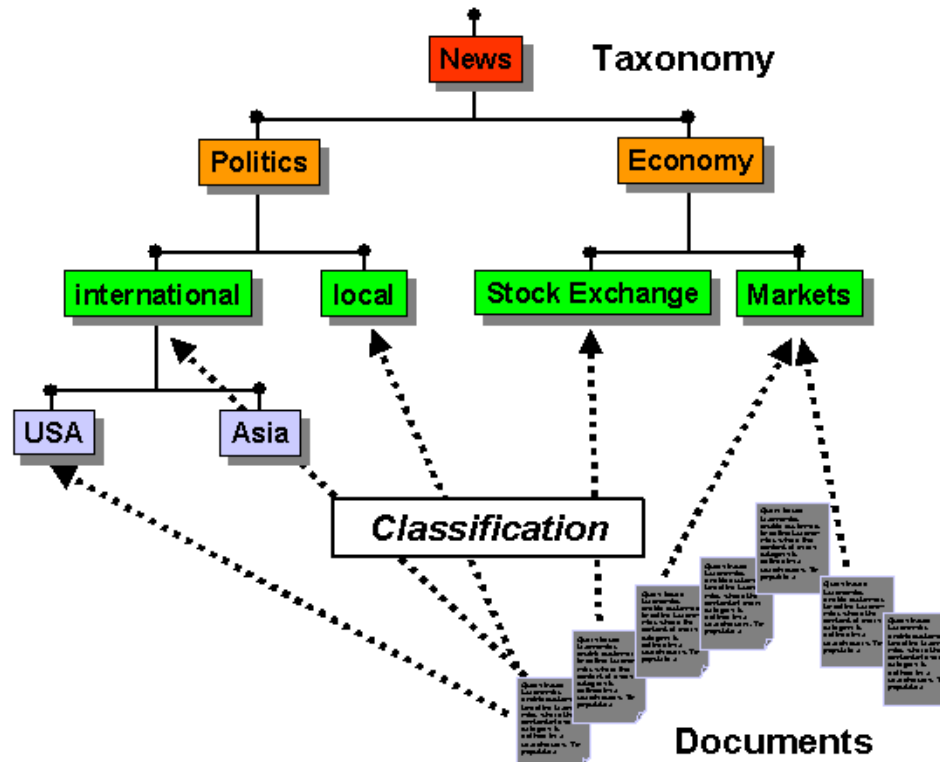


Challenges in text mining

- Data collection is “free text”
 - Data is not well-organized
 - Semi-structured or unstructured
 - Natural language text contains ambiguities on many levels
 - Lexical, syntactic, semantic, and pragmatic
 - Learning techniques for processing text typically need annotated training examples
 - Expensive to acquire at scale
- What to mine?

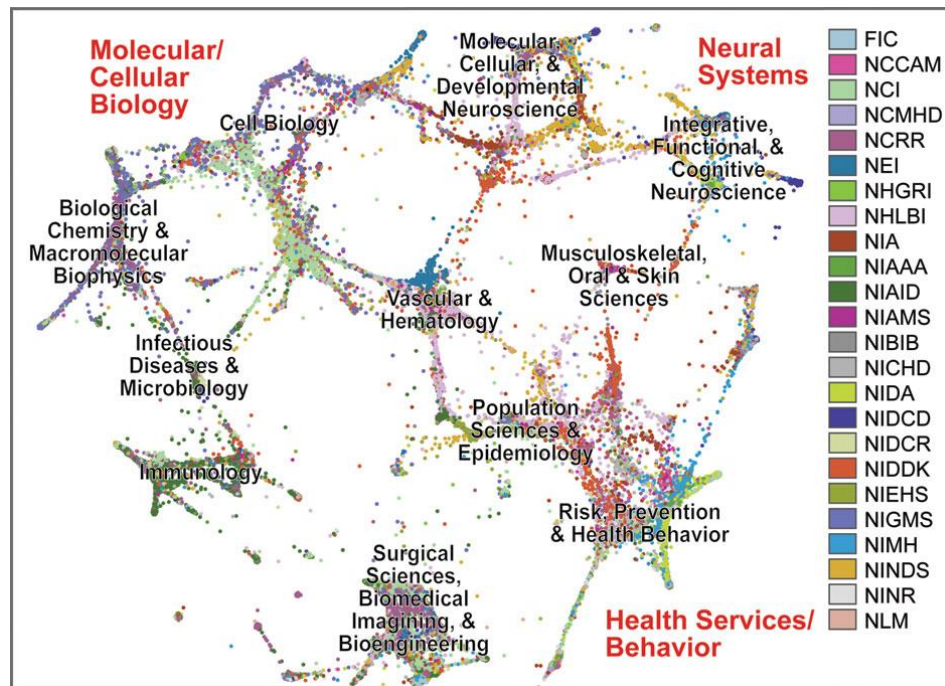
Text mining problems we will solve

- Document categorization
 - Adding structures to the text corpus



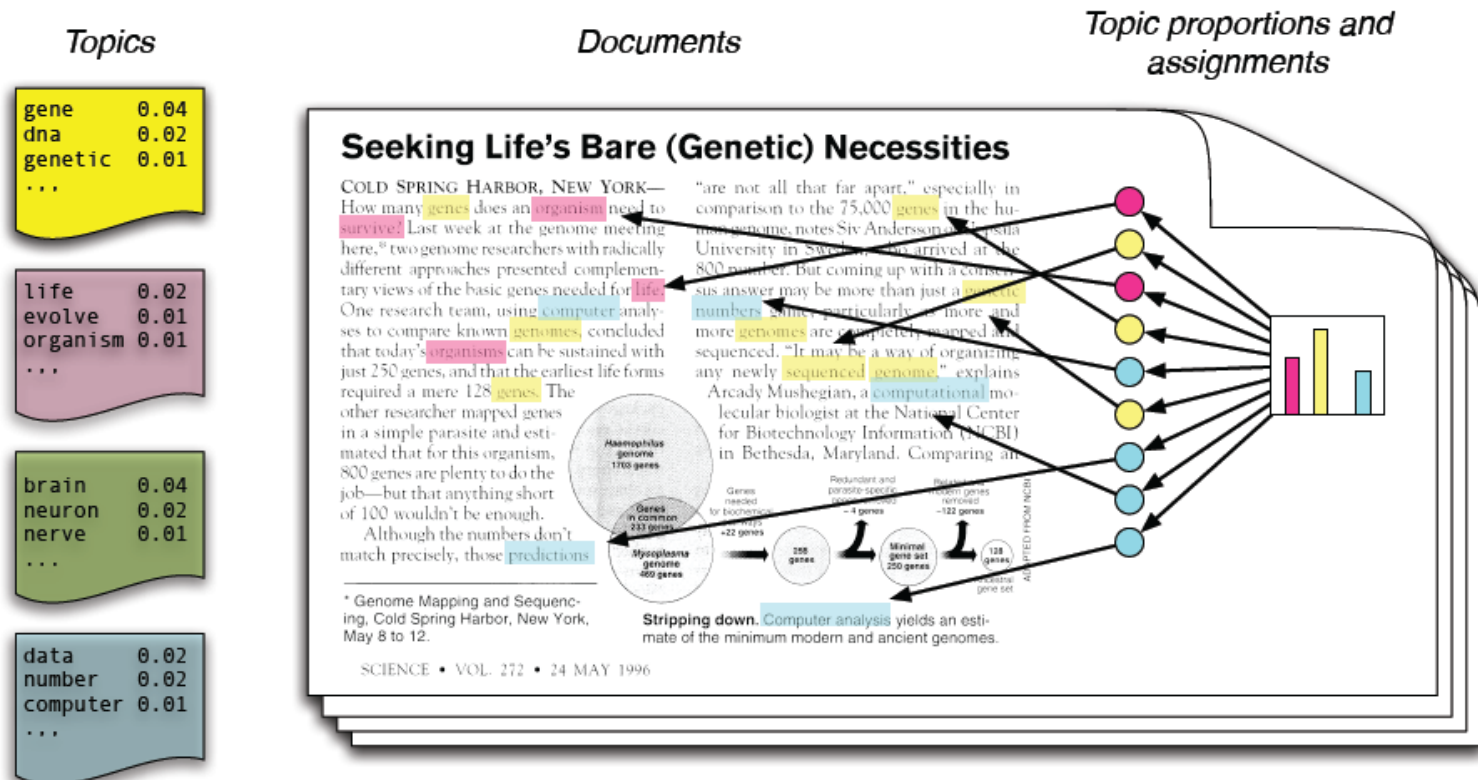
Text mining problems we will solve

- Text clustering
 - Identifying structures in the text corpus



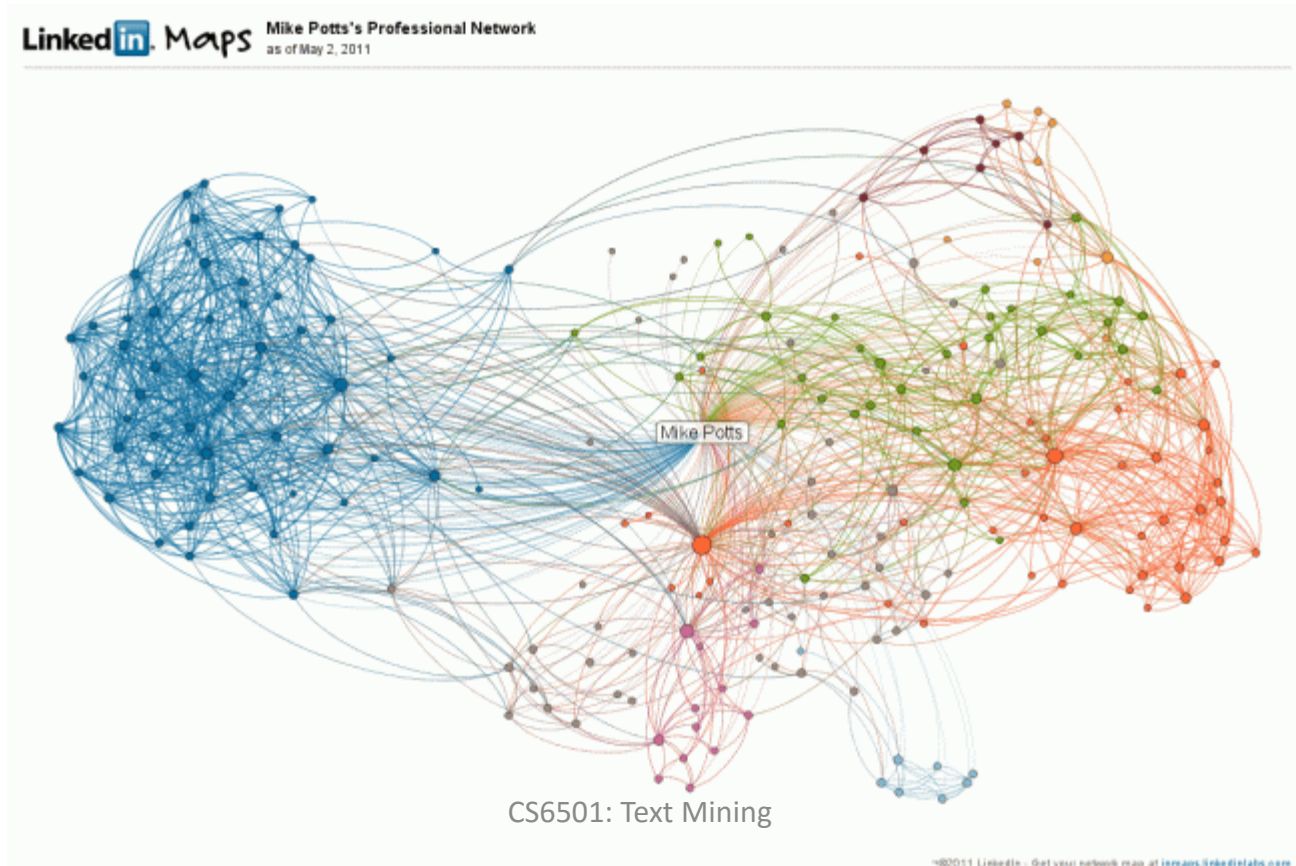
Text mining problems we will solve

- Topic modeling
 - Identifying structures in the text corpus



Text mining problems we will solve

- Social media and network analysis
 - Exploring additional structure in the text corpus



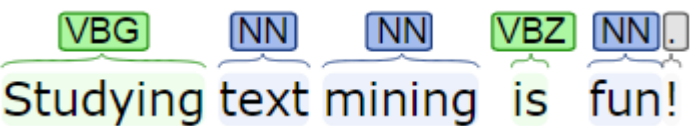
We will also briefly cover

- Natural language processing pipeline

- Tokenization

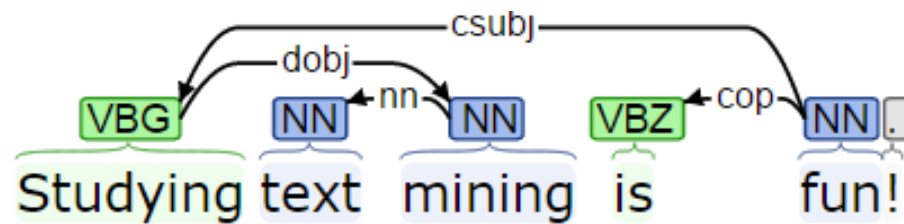
- “Studying text mining is fun!” -> “studying” + “text” + “mining” + “is” + “fun” + “!”

- Part-of-speech tagging

- “Studying text mining is fun!” -> 

- Dependency parsing

- “Studying text mining is fun!” ->



We will also briefly cover

- Machine learning techniques
 - Supervised methods
 - Naïve Bayes, k Nearest Neighbors, Logistic Regression
 - Unsupervised methods
 - K-Means, hierarchical clustering, topic models
 - Semi-supervised methods
 - Expectation Maximization

Text mining in the era of Big Data

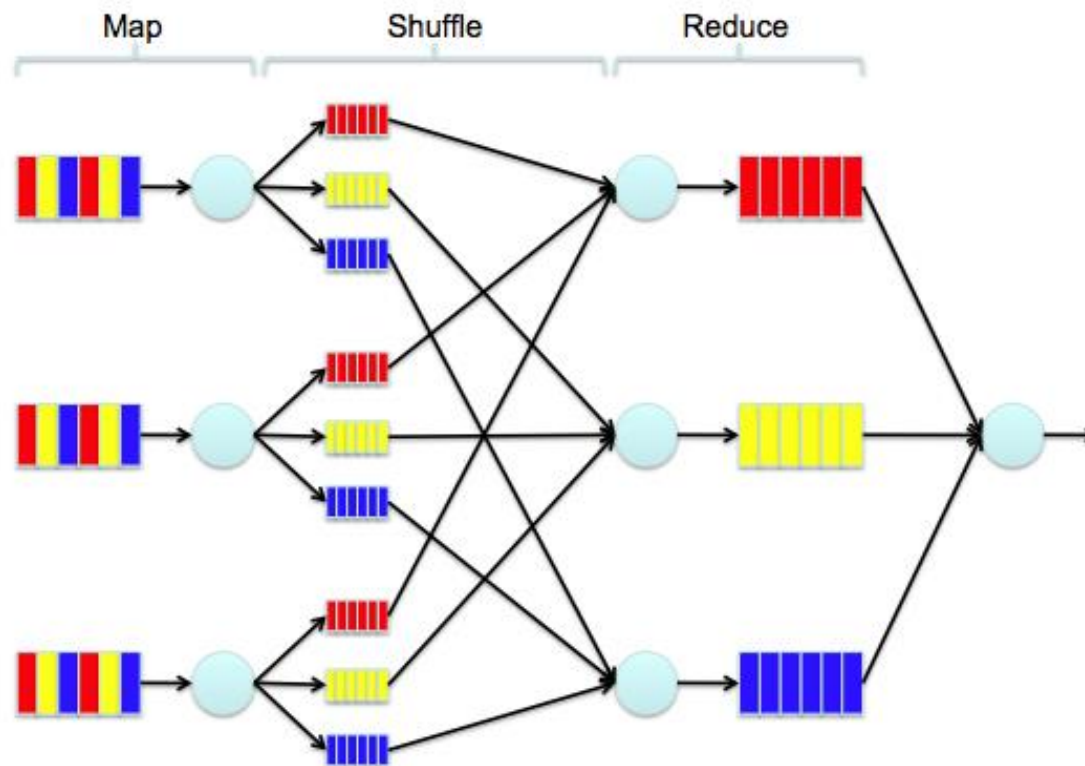
- Huge in size
 - Google processes 5.13B queries/day (2013)
 - Twitter receives 340M tweets/day (2012)
 - Facebook has 2.5 PB of user data + 15 TB/day (1/2009)
 - eBay has 6.5 PB of user data + 50 TB/day (1/2009)
- 80% data is unstructured (IBM, 2010)

640K ought to be enough for anybody.



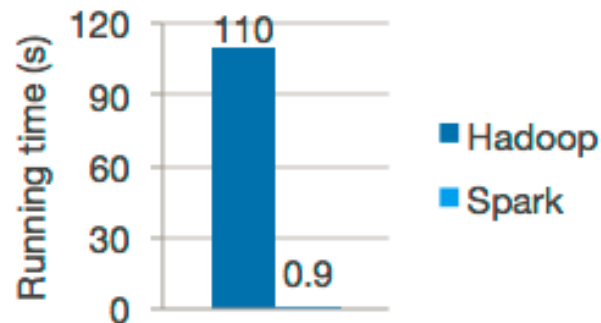
Scalability is crucial

- Large scale text processing techniques
 - MapReduce framework



State-of-the-art solutions

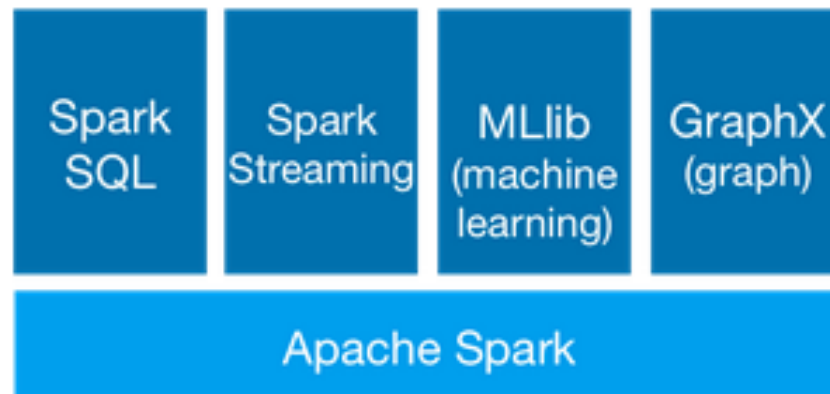
- Apache Spark (spark.apache.org)
 - In-memory MapReduce
 - Specialized for machine learning algorithms
 - Speed
 - 100x faster than Hadoop MapReduce in memory, or 10x faster on disk.



Logistic regression in Hadoop and Spark

State-of-the-art solutions

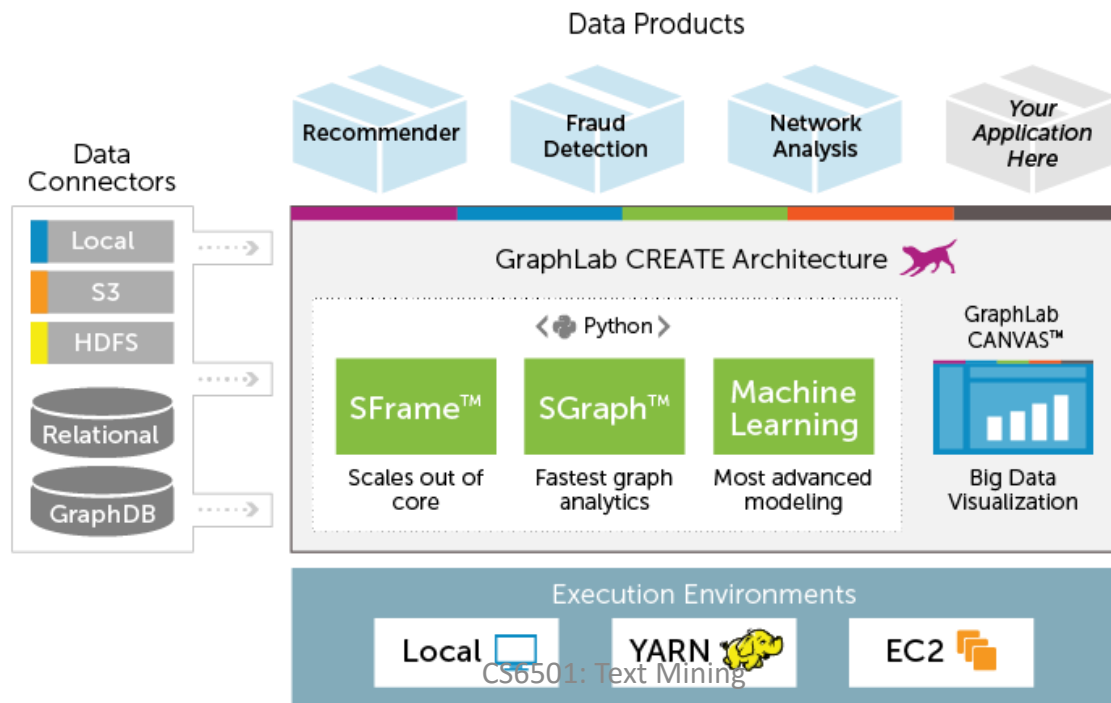
- Apache Spark (spark.apache.org)
 - In-memory MapReduce
 - Specialized for machine learning algorithms
 - Generality
 - Combine SQL, streaming, and complex analytics



State-of-the-art solutions

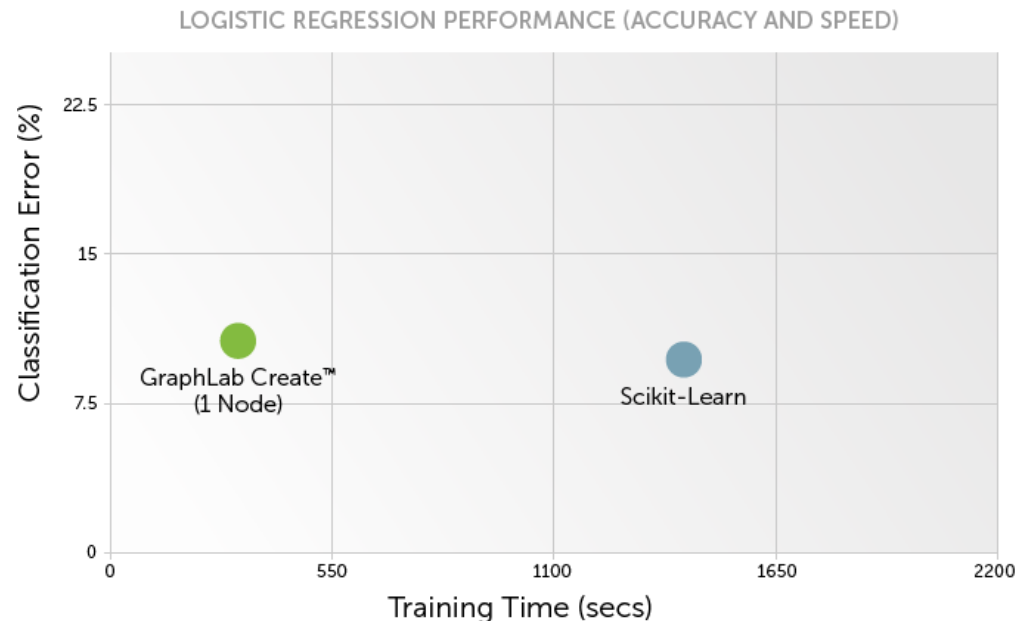
- GraphLab (graphlab.com)
 - Graph-based, high performance, distributed computation framework

ARCHITECTURE

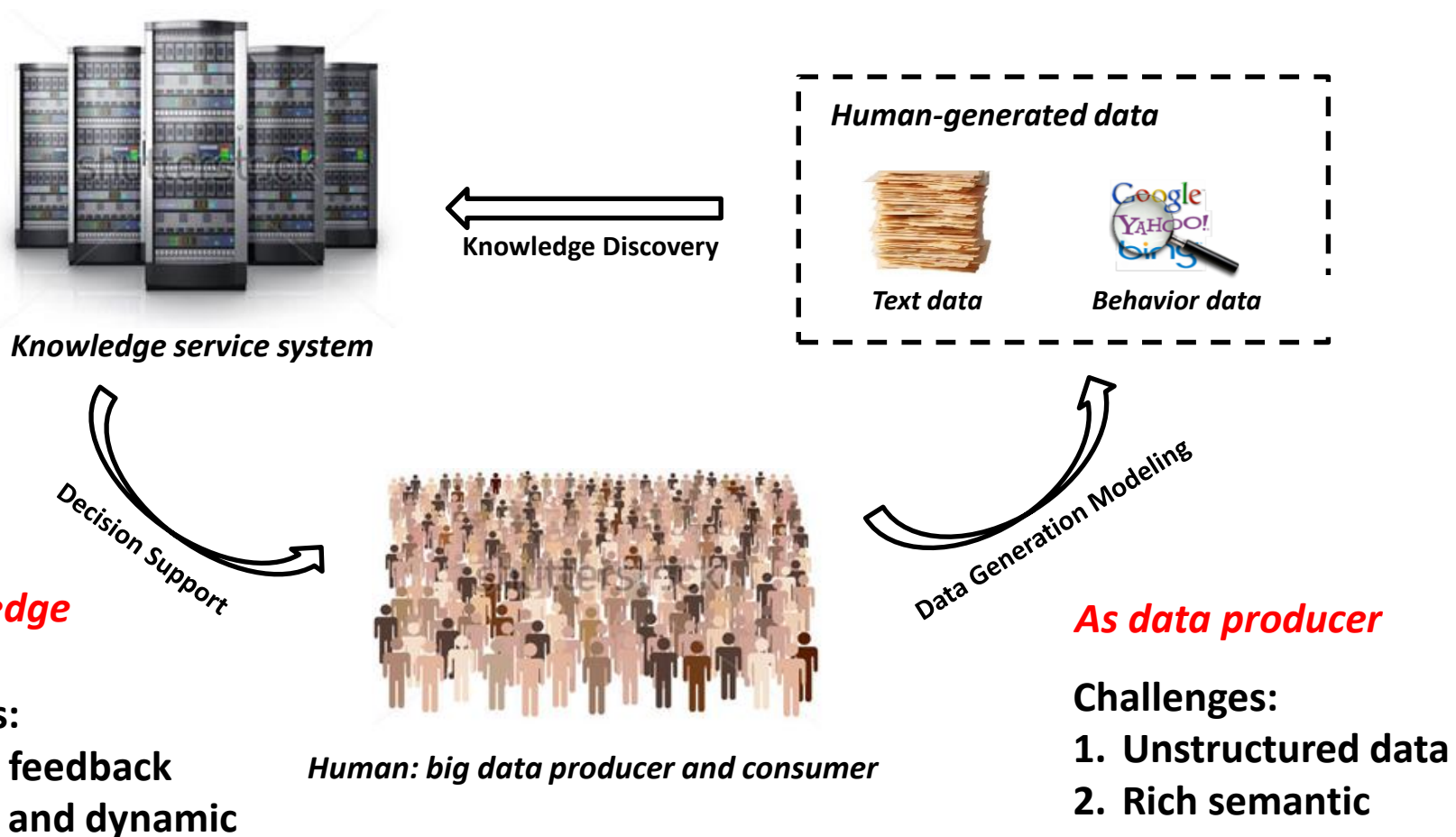


State-of-the-art solutions

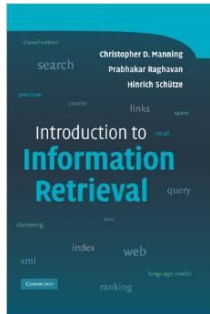
- GraphLab (graphlab.com)
 - Specialized for sparse data with local dependencies for iterative algorithms



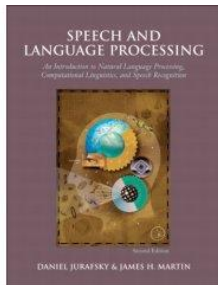
Text mining in the era of Big Data



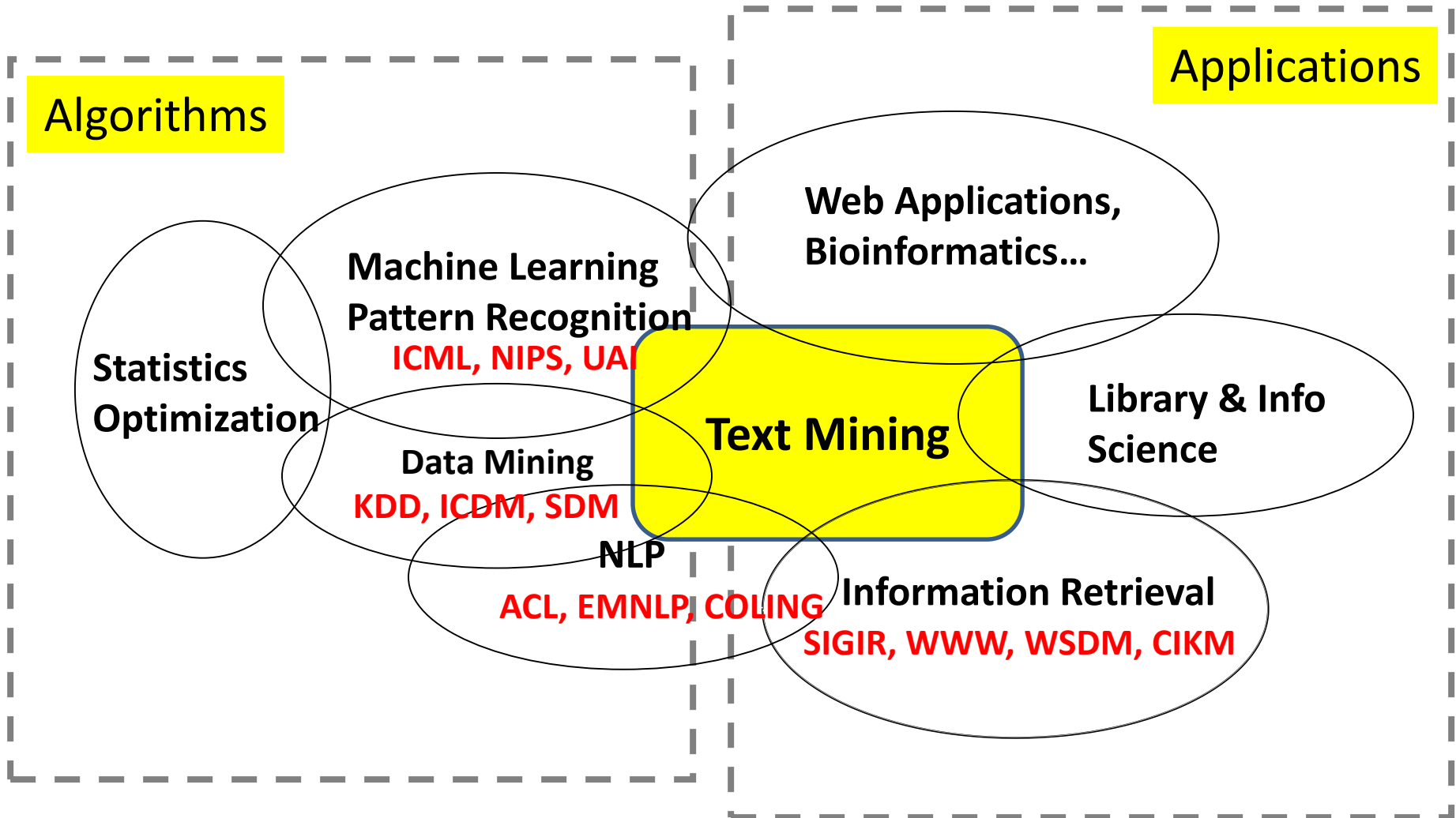
Text books



- ***Introduction to Information Retrieval.*** Christopher D. Manning, Prabhakar Raghavan, and Hinrich Schuetze, Cambridge University Press, 2007.
- ***Speech and Language Processing.*** Daniel Jurafsky and James H. Martin, Pearson Education, 2000.
- ***Mining Text Data.*** Charu C. Aggarwal and ChengXiang Zhai, Springer, 2012.



What to read?



- Find more on course website for resource

Welcome to the class of “Text Mining”!

