

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

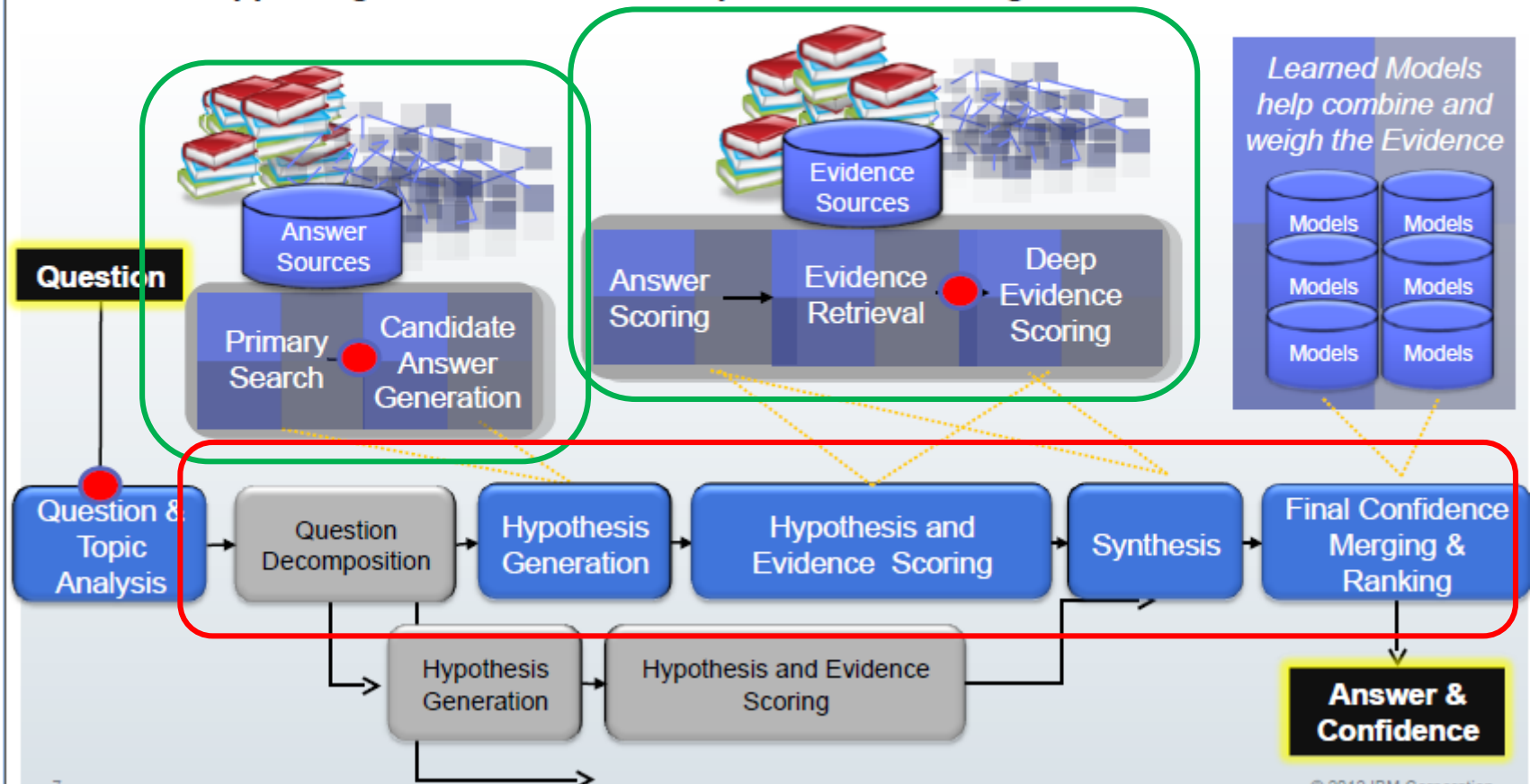
Knowledge discovery from text data

- 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?

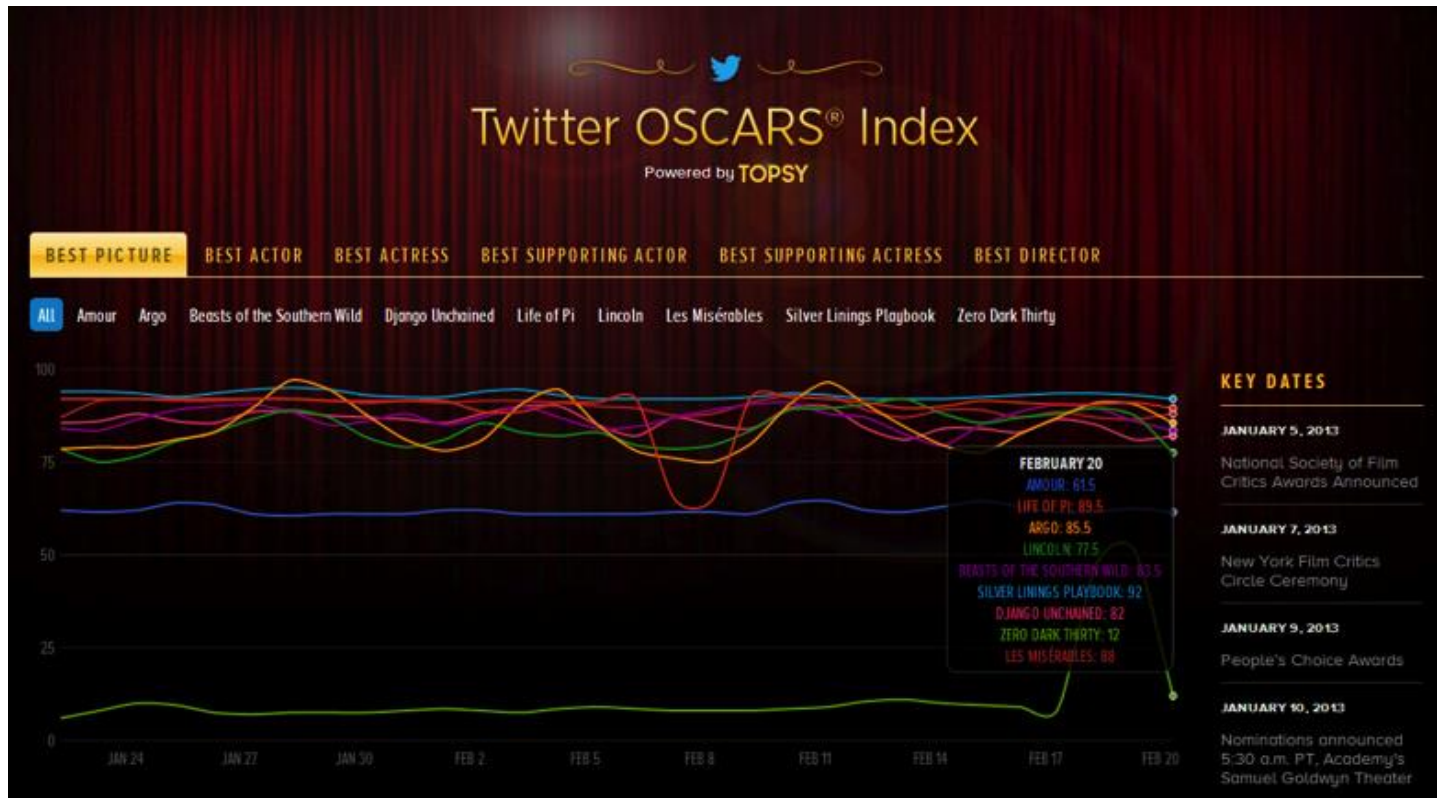
- *“Watson had access to 200 million pages of structured and unstructured content consuming four terabytes of disk storage including the full text of Wikipedia” – PC World*
- *“The sources of information for Watson include encyclopedias, dictionaries, thesauri, newswire articles, and literary works. 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](#)

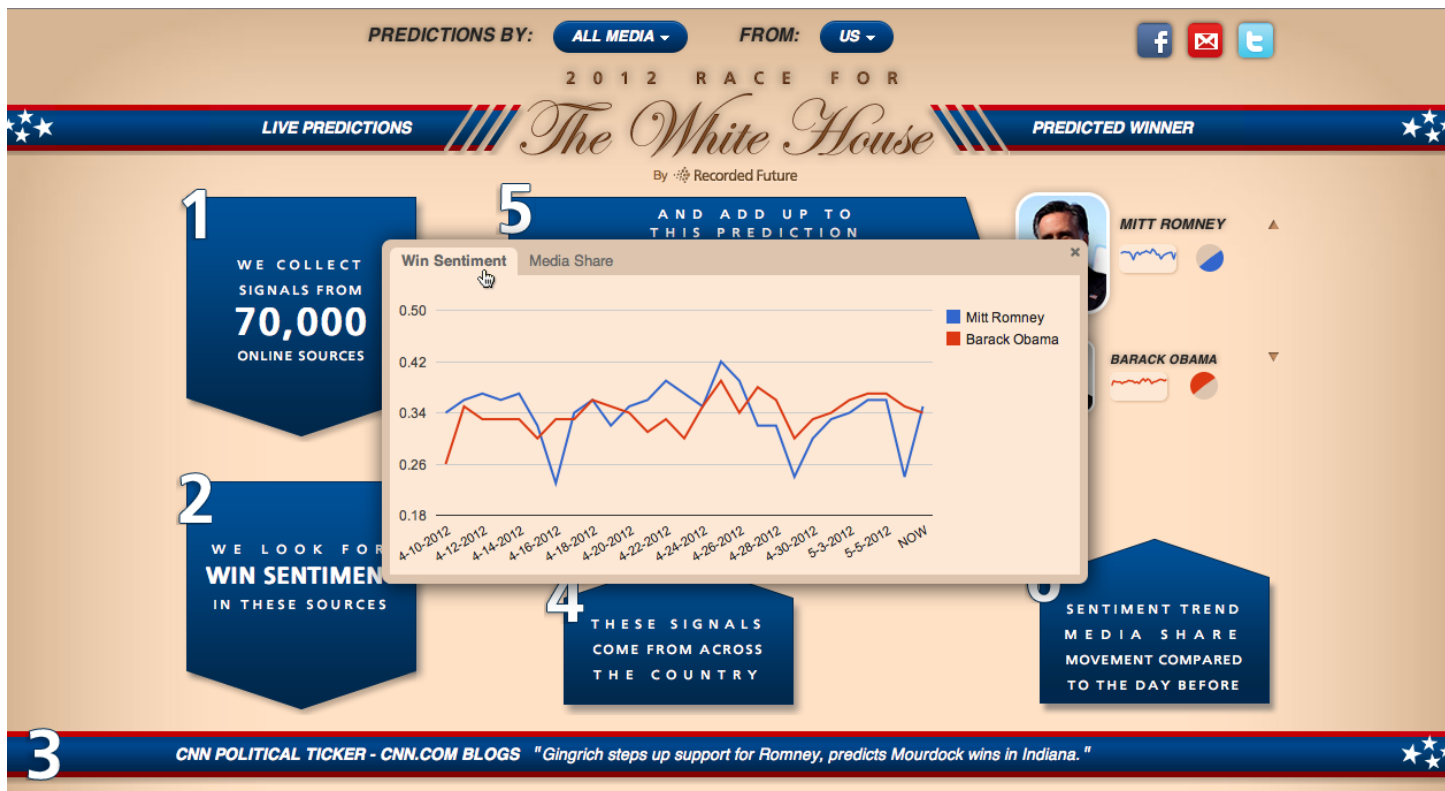
Text mining around us

- 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', with 'Web' being the selected tab. The search results are displayed below the tabs, showing 19,200,000 results. 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 is a result from statsoft.com titled 'Text Mining (Big Data, Unstructured Data)', also with a red box around its snippet. The third result is from academic.research.microsoft.com titled 'Text Mining', with a red box around its snippet. The fourth result is from searchbusinessanalytics.techtarget.com titled 'What is text mining (text analytics)? - Definition from ...', with a red box around its snippet. On the right side of the page, there is a 'Text mining' knowledge panel. It contains a definition of text mining, a list of related people, a list of related topics, and a list of related searches. The definition and the list of related people are highlighted with red boxes.

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.

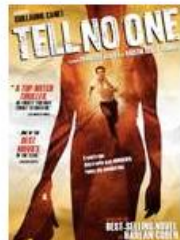
Text mining
Text mining, also referred to as text data mining, roughly equivalent to text analytics, refers to the process of deriving high-quality information from text. High-quality information is typically derived through the devising of patterns and trends through means such as statistical pattern 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

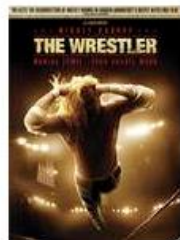


Downfall

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


- 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



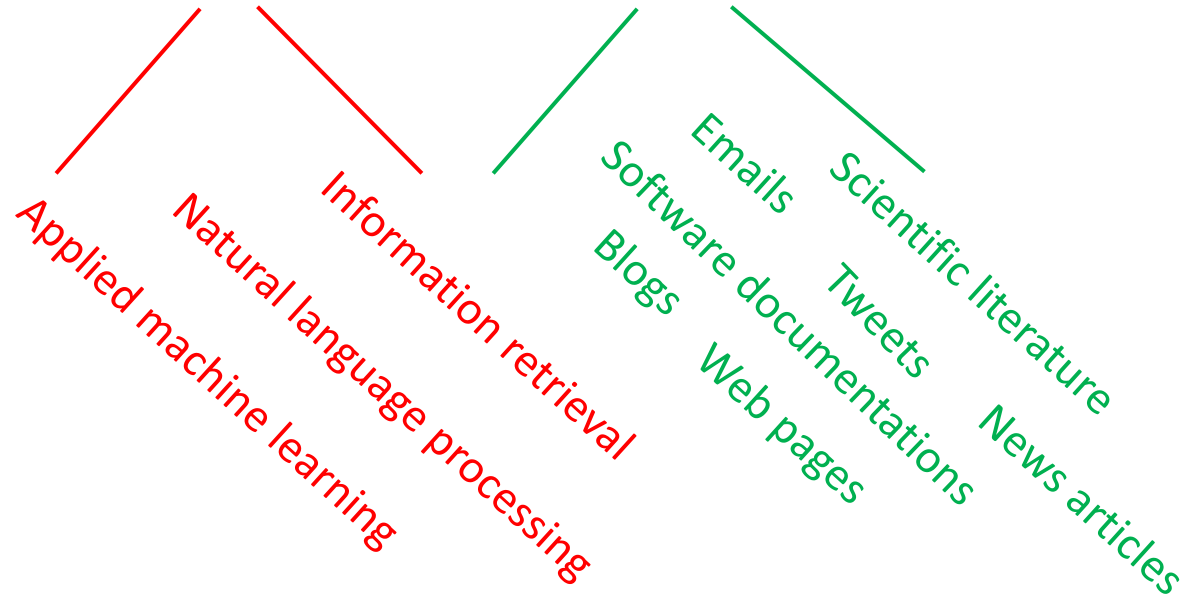
Wisconsin man keeps 40-year-old Christmas tree up until son returns
By Brendan O'Brien (Reuters) - A Wisconsin man will refuse for about the 40th time to partake in the annual after-holiday chore of putting Christmas
[Reuters](#)



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](#)

How to perform text mining?

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



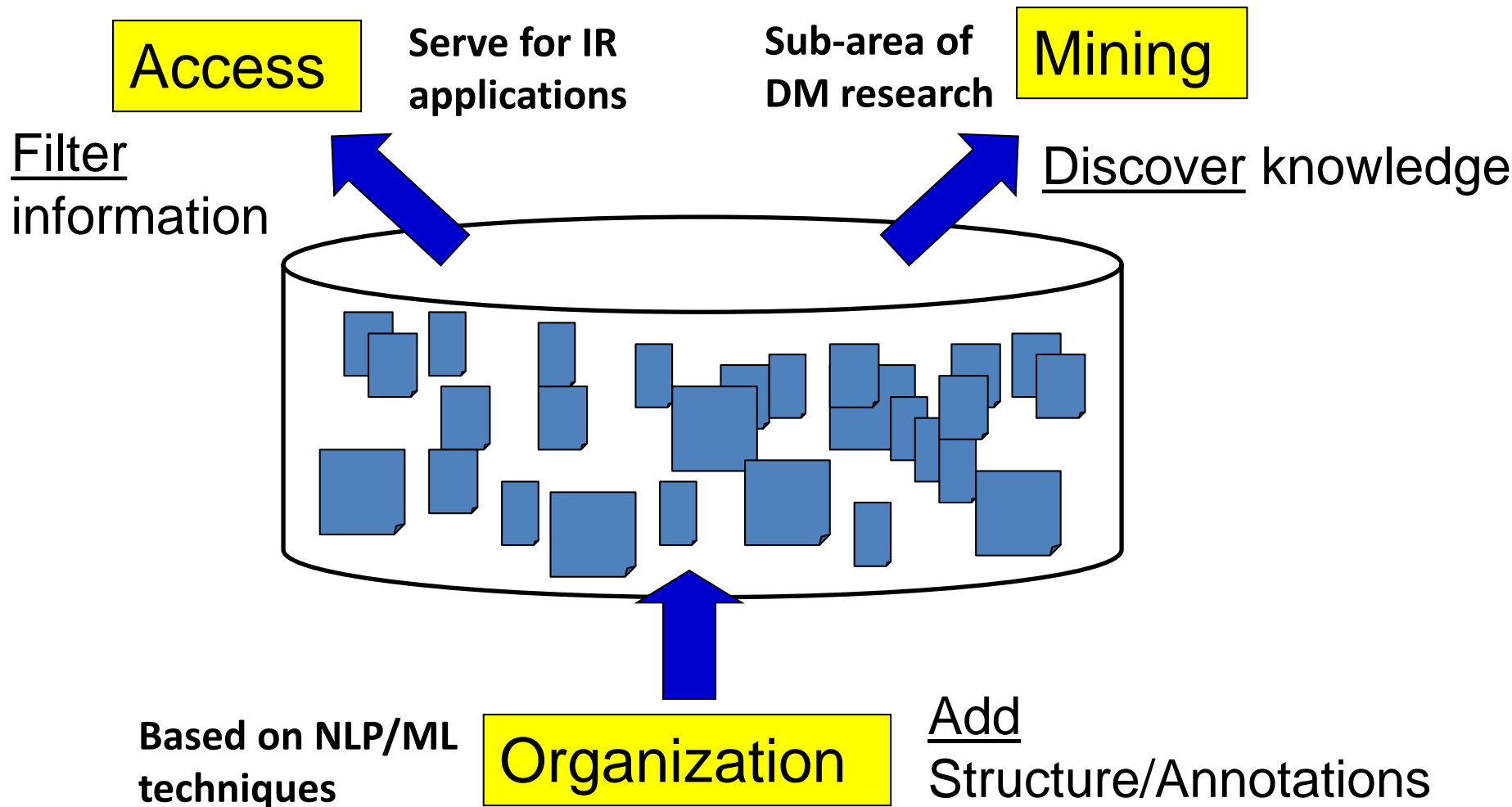
Text mining v.s. NLP, IR, DM...

- 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

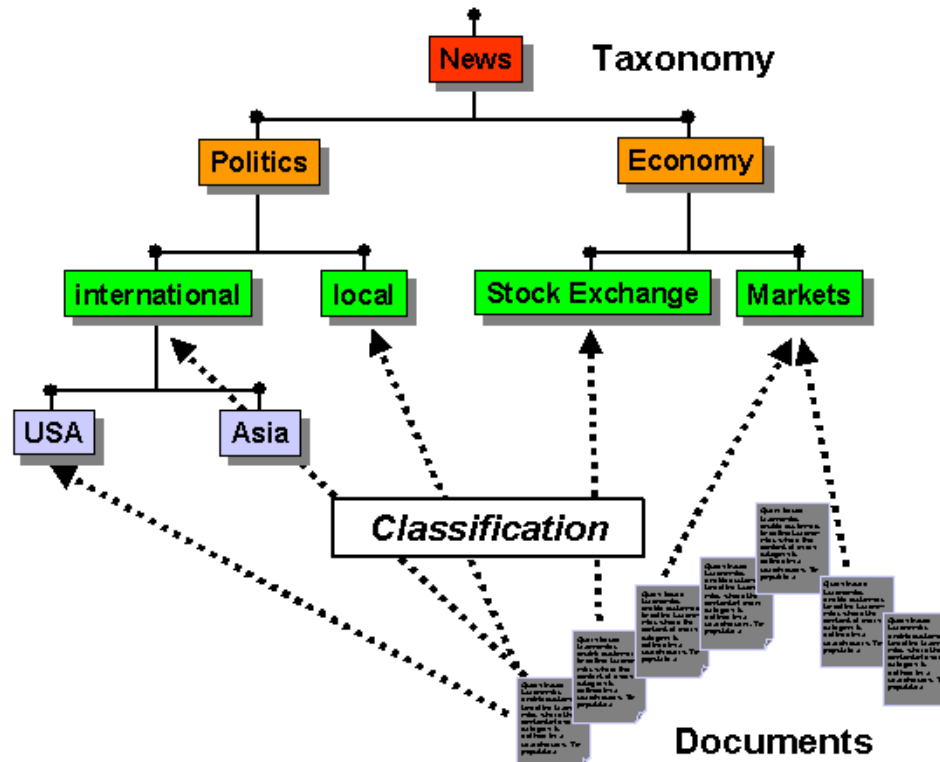


Challenges in text mining

- 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?

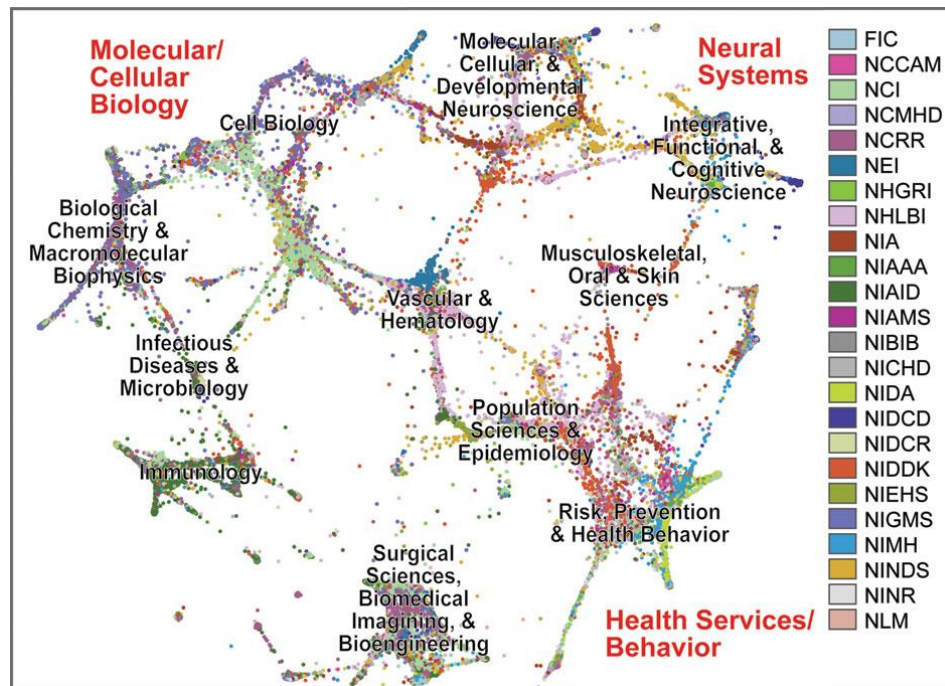
Text mining problems we will solve

- Document categorization
 - Adding structures to the text corpus



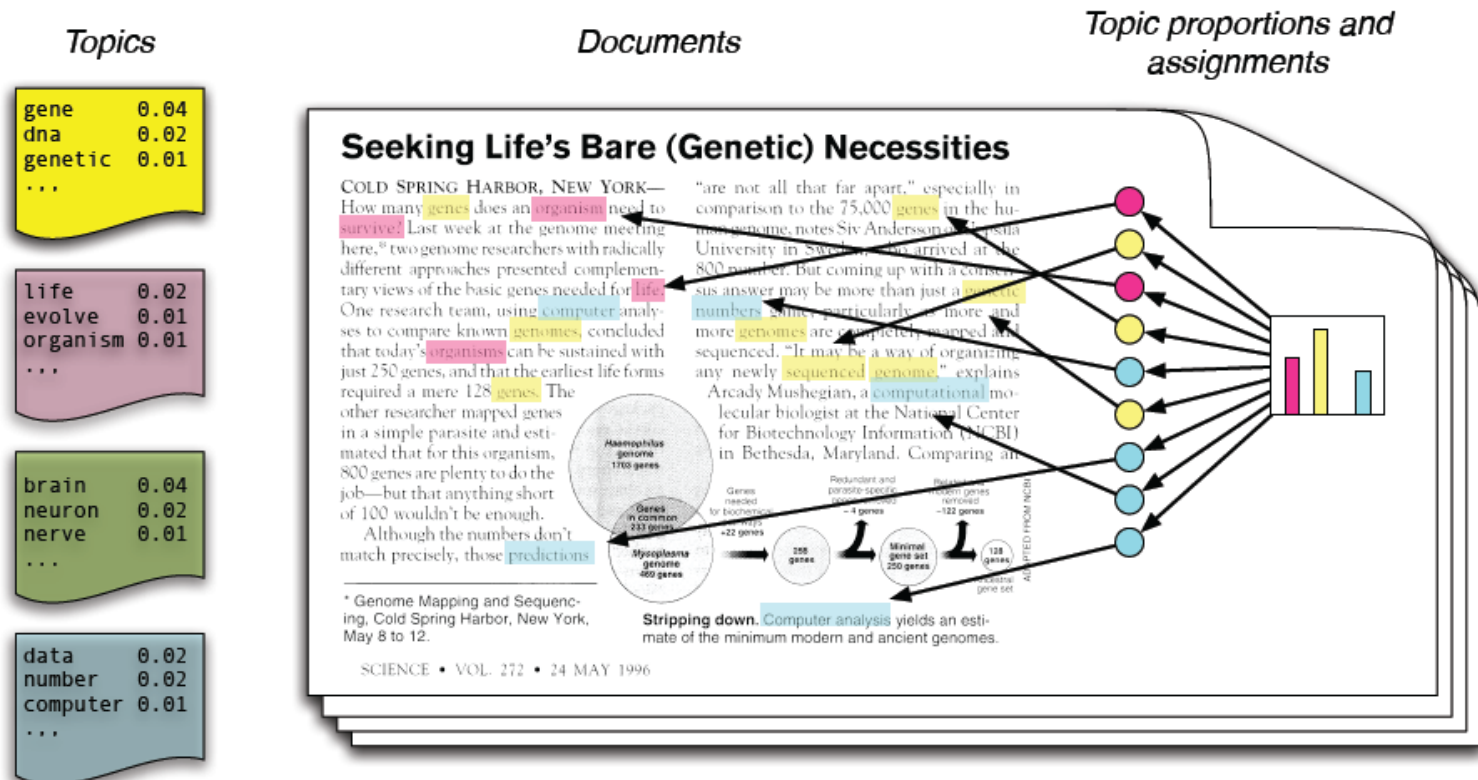
Text mining problems we will solve

- Text clustering
 - Identifying structures in the text corpus



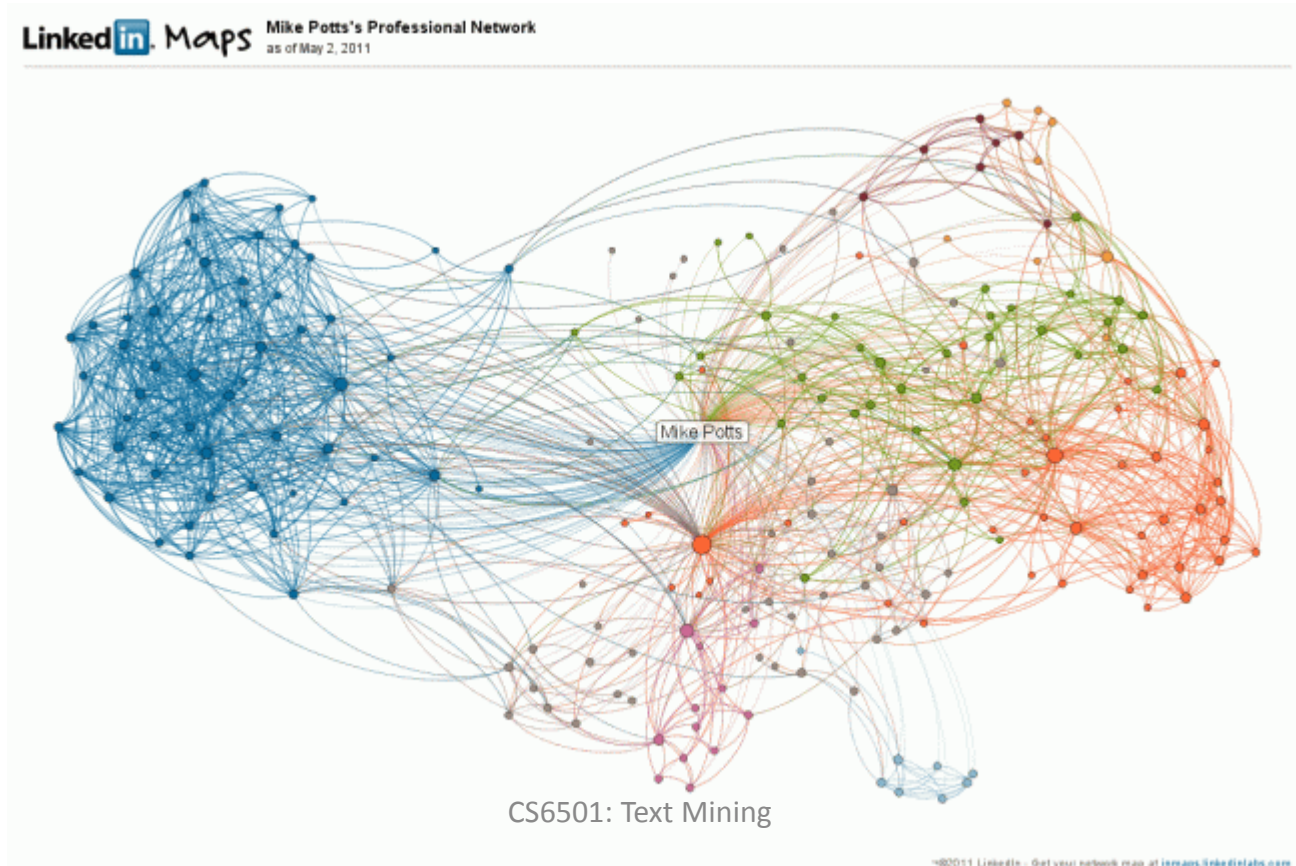
Text mining problems we will solve

- 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



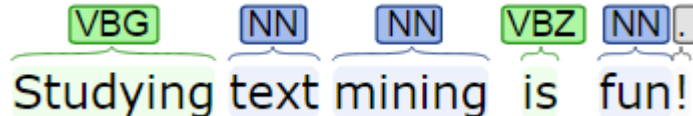
We will also briefly cover

- Natural language processing pipeline

- Tokenization

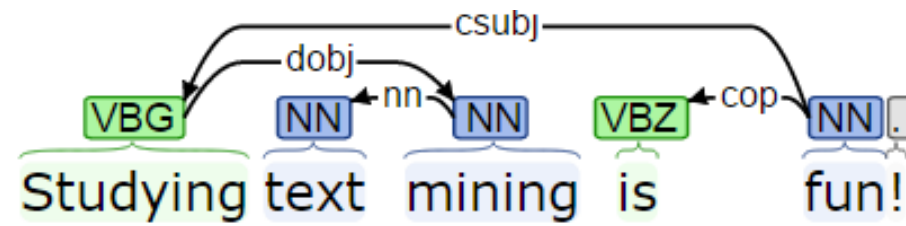
- “Studying text mining is fun!” -> “studying” + “text” + “mining” + “is” + “fun” + “!”

- 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
 - Semi-supervised methods
 - Expectation Maximization

Text mining in the era of Big Data

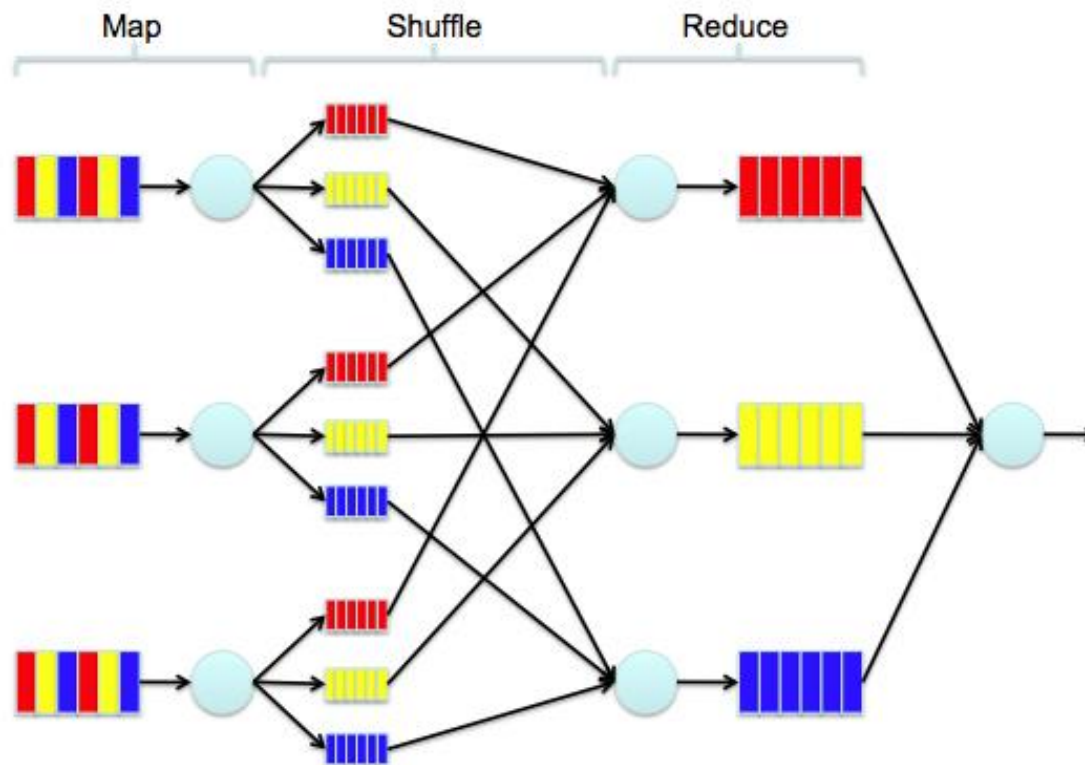
- Huge in size
 - Google processes 5.13B queries/day (2013)
 - Twitter receives 340M tweets/day (2012)
 - Facebook has 2.5 PB of user data + 15 TB/day (1/2009)
 - eBay has 6.5 PB of user data + 50 TB/day (1/2009)
- 80% data is unstructured (IBM, 2010)

640K ought to be
enough for anybody.



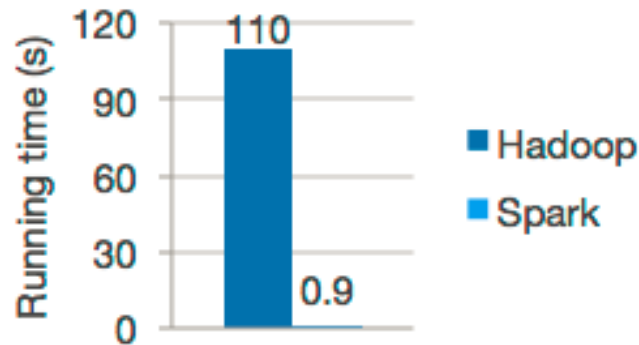
Scalability is crucial

- Large scale text processing techniques
 - MapReduce framework



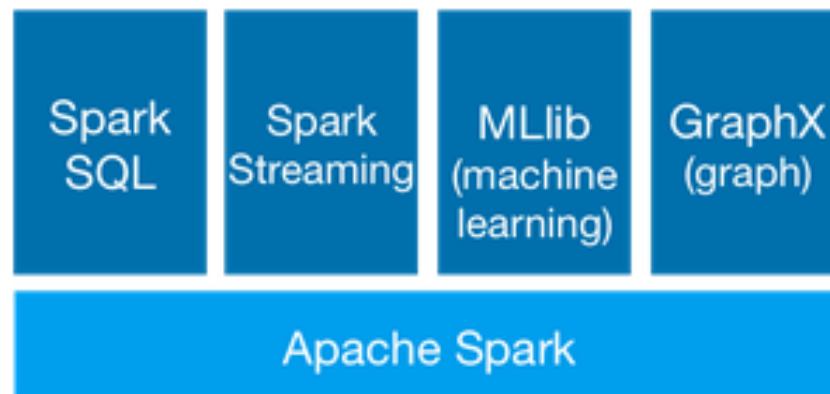
State-of-the-art solutions

- 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.



State-of-the-art solutions

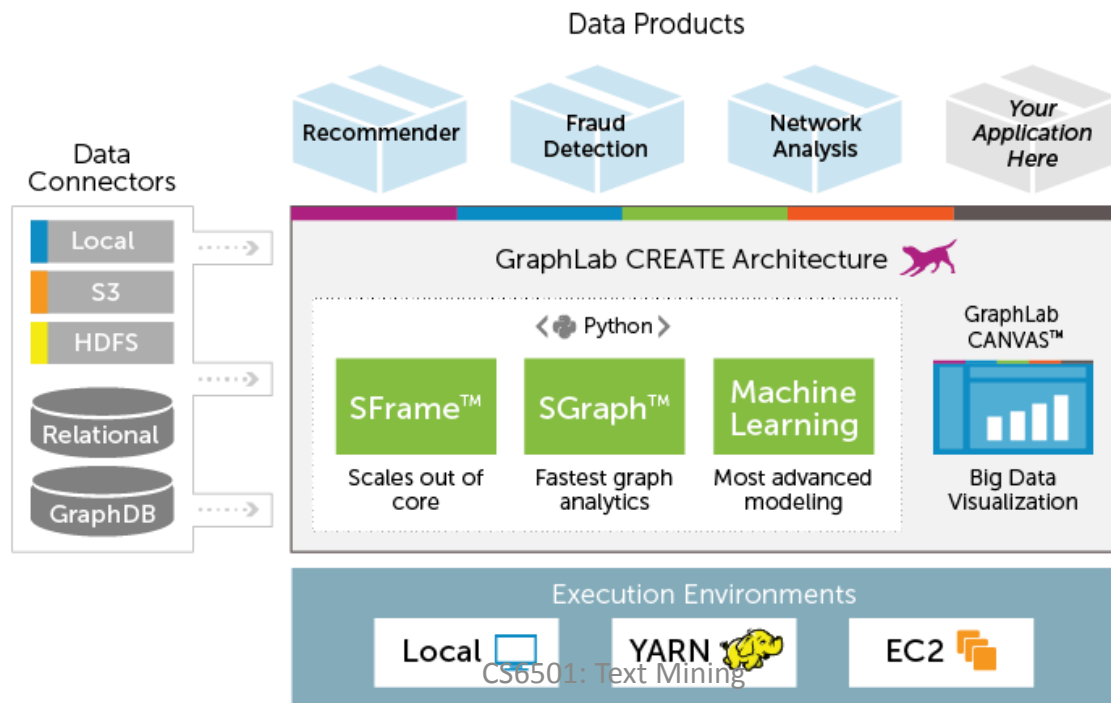
- Apache Spark (spark.apache.org)
 - In-memory MapReduce
 - Specialized for machine learning algorithms
 - Generality
 - Combine SQL, streaming, and complex analytics



State-of-the-art solutions

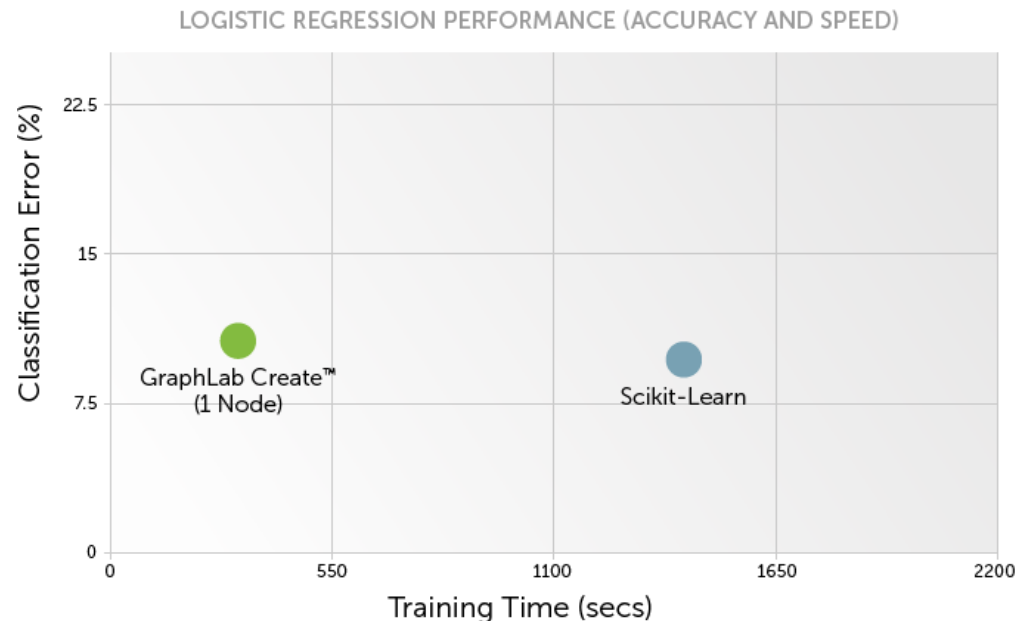
- GraphLab (graphlab.com)
 - Graph-based, high performance, distributed computation framework

ARCHITECTURE

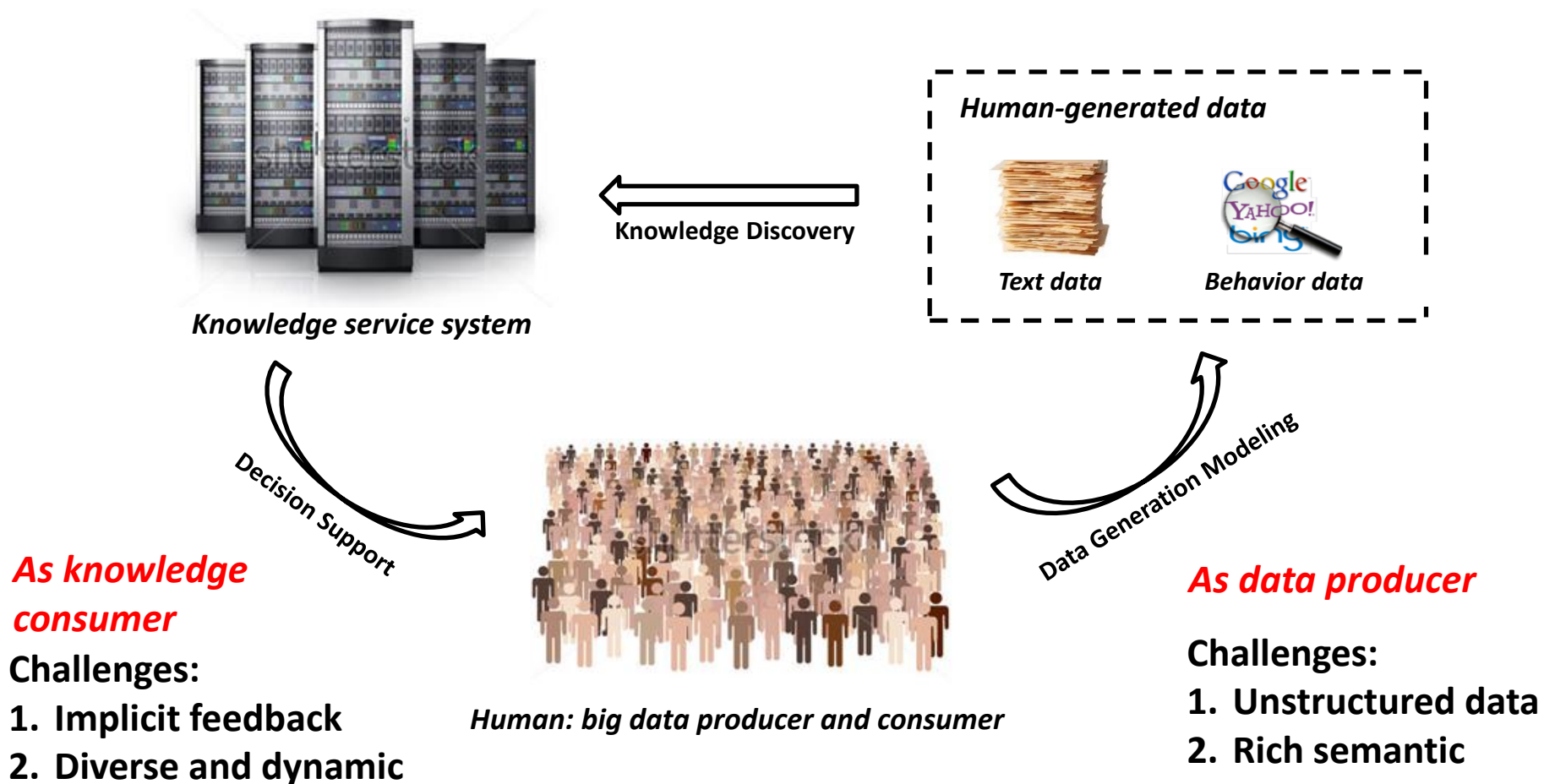


State-of-the-art solutions

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



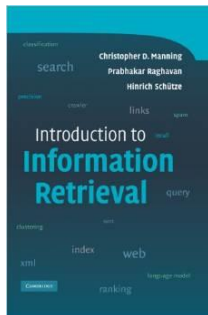
Text mining in the era of Big Data



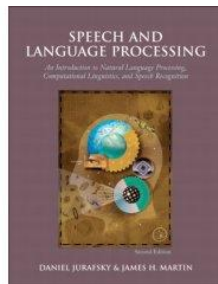
Text books



- ***Mining Text Data***. Charu C. Aggarwal and ChengXiang Zhai, Springer, 2012.

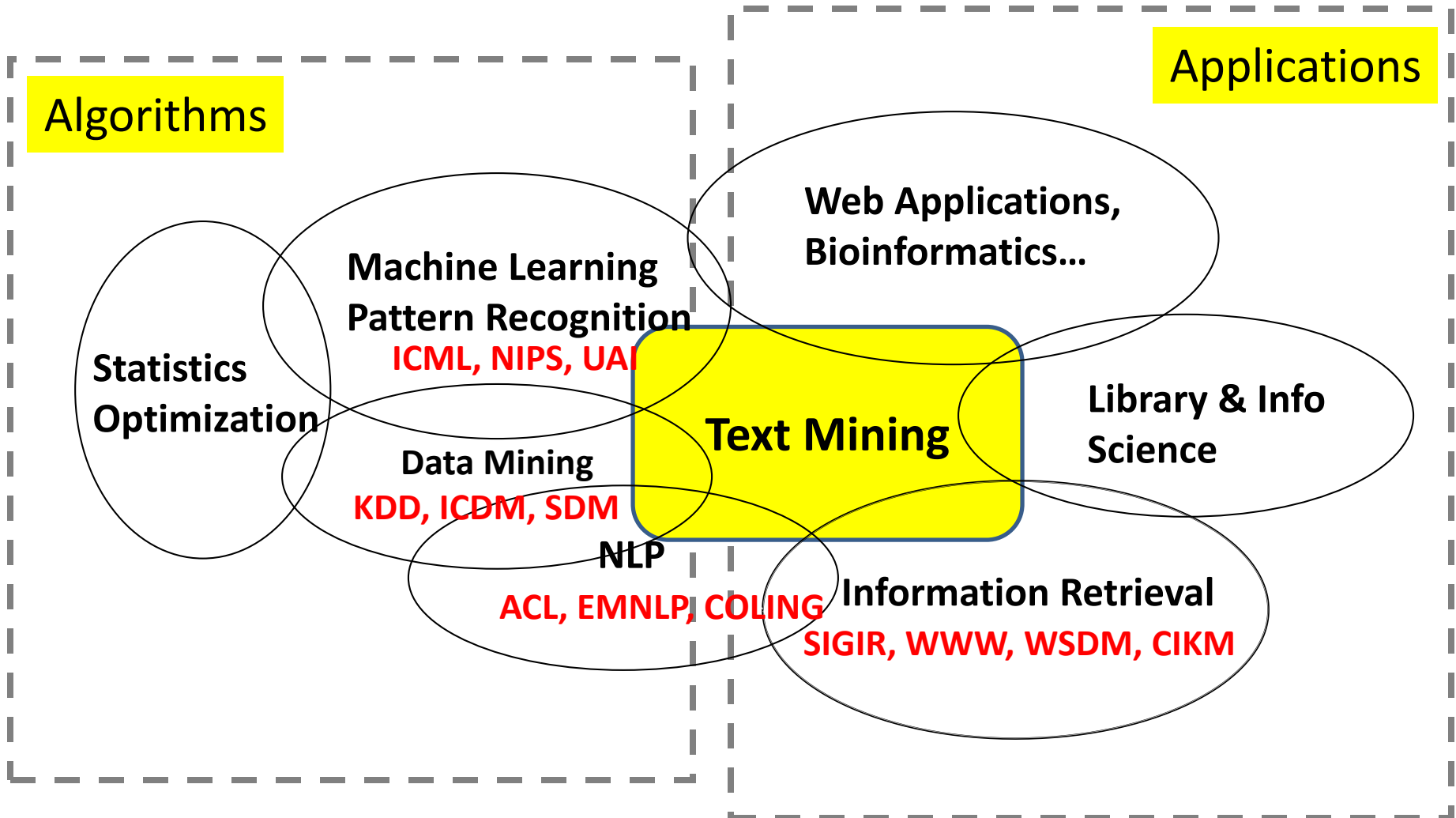


- ***Introduction to Information Retrieval***. Christopher D. Manning, Prabhakar Raghavan, and Hinrich Schuetze, Cambridge University Press, 2007.



- ***Speech and Language Processing***. Daniel Jurafsky and James H. Martin, Pearson Education, 2000.

What to read?



- Find more on course website for resource

Welcome to the class of “Text Mining”!

