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

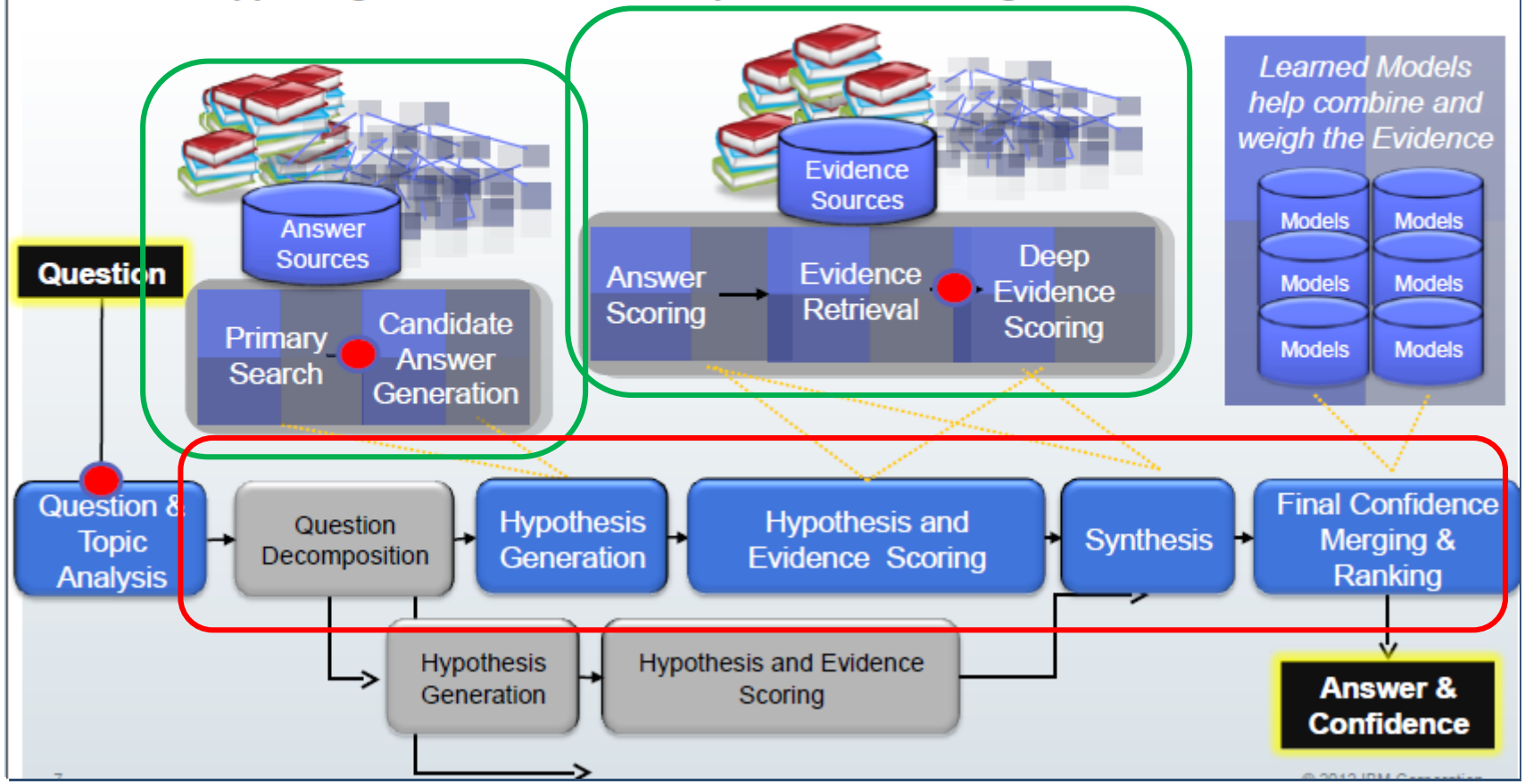
- Goal-oriented (effectiveness driven)
  - Any process that generates useful results that are non-obvious is called “mining”.
  - Keywords: “**useful**” + “**non-obvious**”
  - Data isn’t necessarily massive
- Method-oriented (efficiency driven)
  - Any process that involves extracting information from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

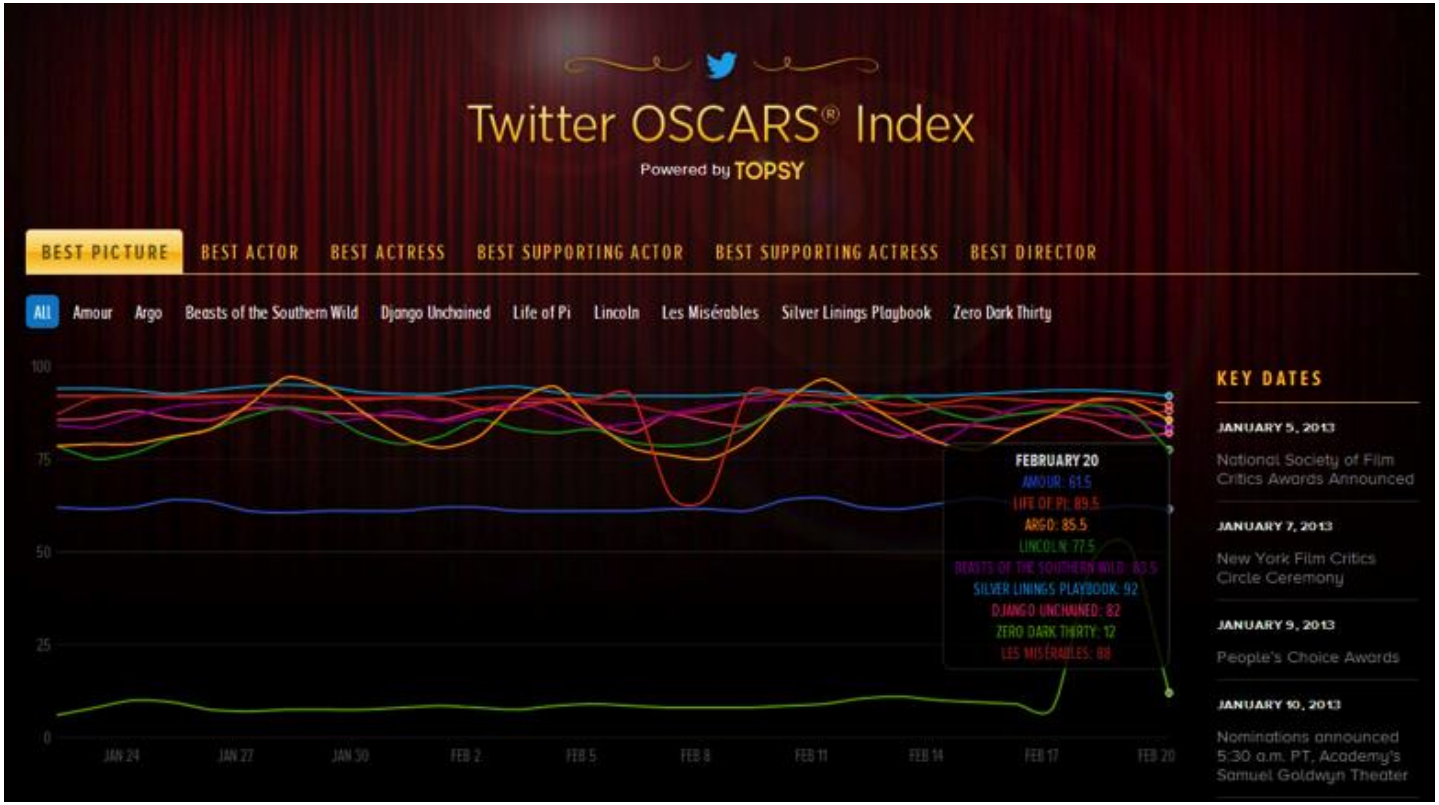
- *"Watson had access to <u>200 million pages</u> of structured and unstructured content consuming four terabytes of disk storage including the full text of Wikipedia" – PC World*
- *"The sources of information for Watson include encyclopedias, dictionaries, thesauri, newswire articles, and literary 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

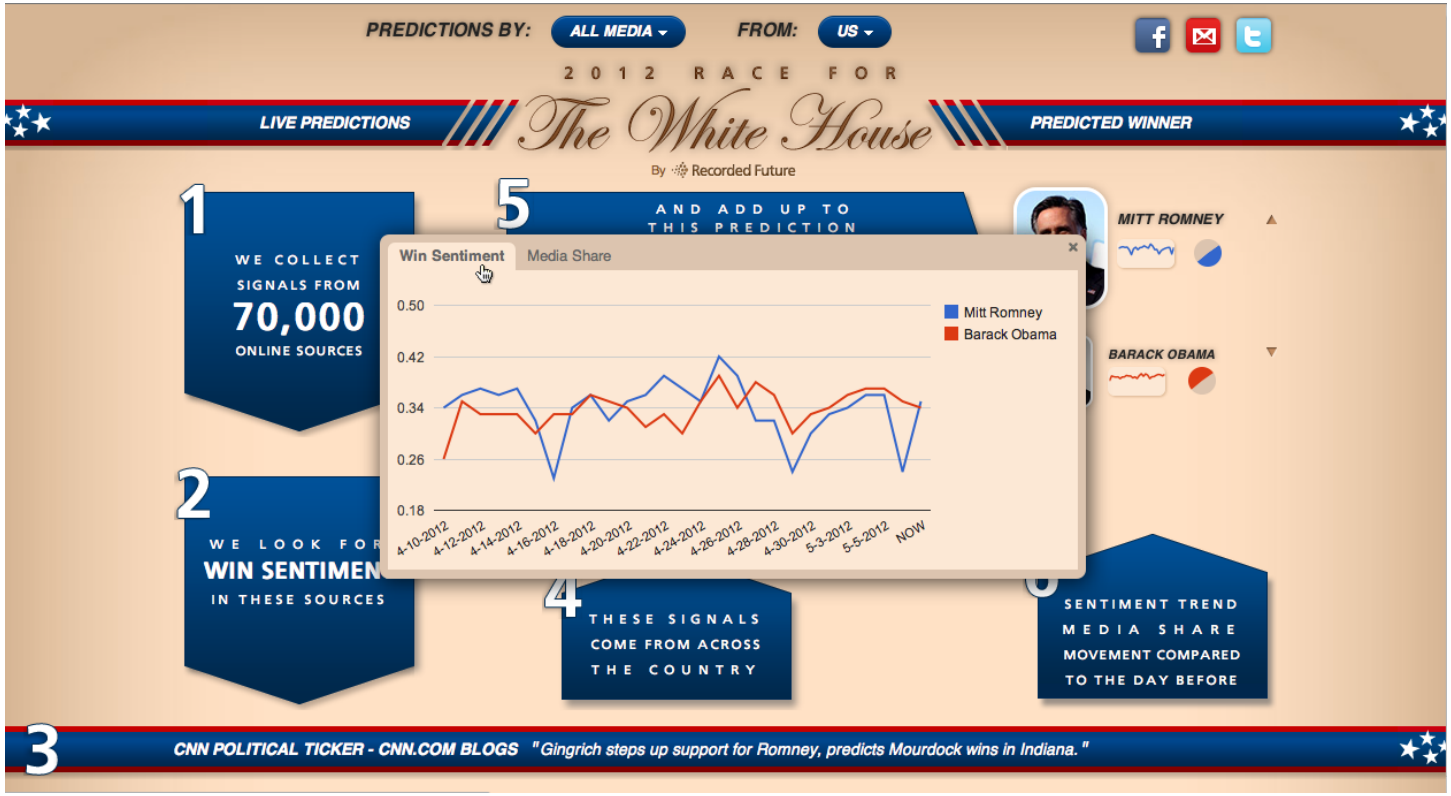
Text mining around us

- Sentiment analysis



Text mining around us

- Sentiment analysis

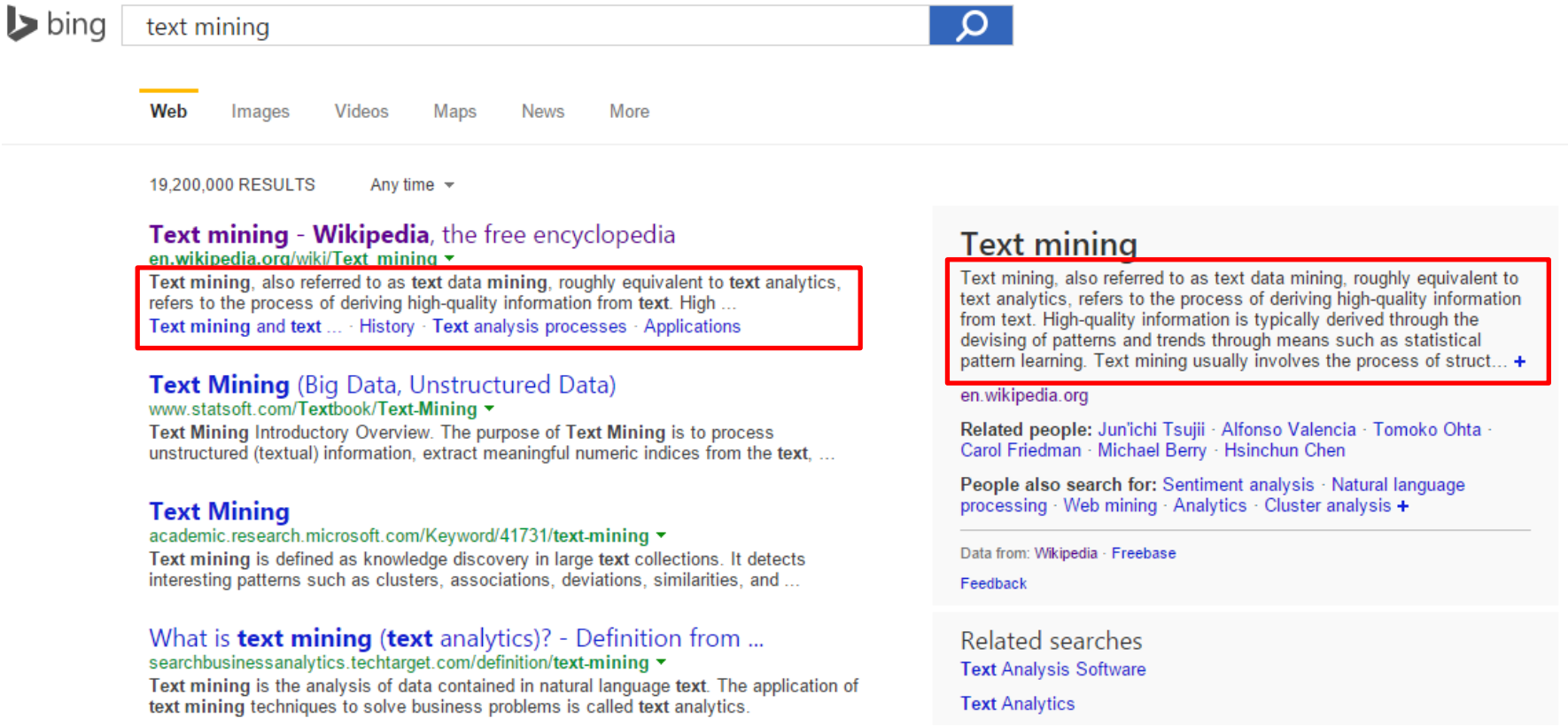


Text mining around us

- Document summarization

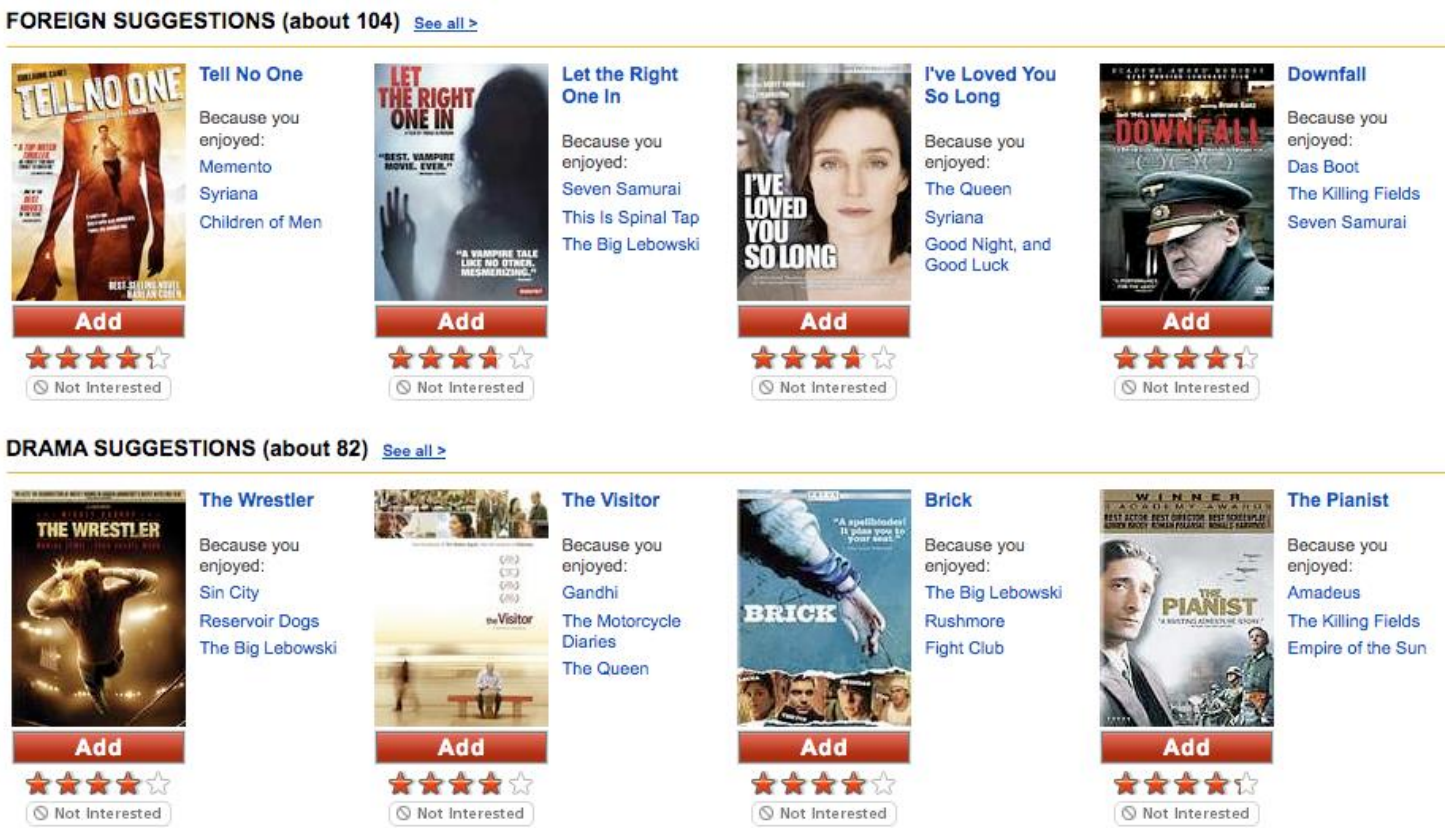
Text mining around us

- Document summarization



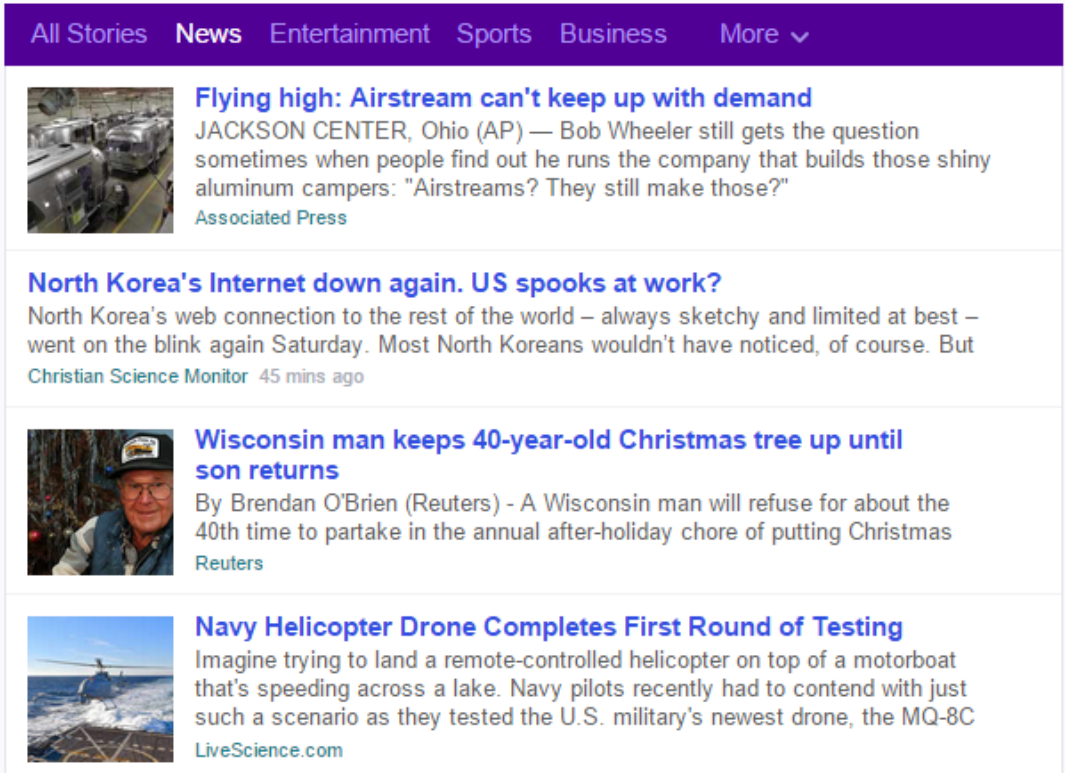
Text mining around us

- Movie recommendation



Text mining around us

- News recommendation



How to perform text mining?

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

Applied machine learning
Natural language processing
Information retrieval

Emails
Software documentations
Blogs
Web pages
Tweets
Scientific literature
News articles

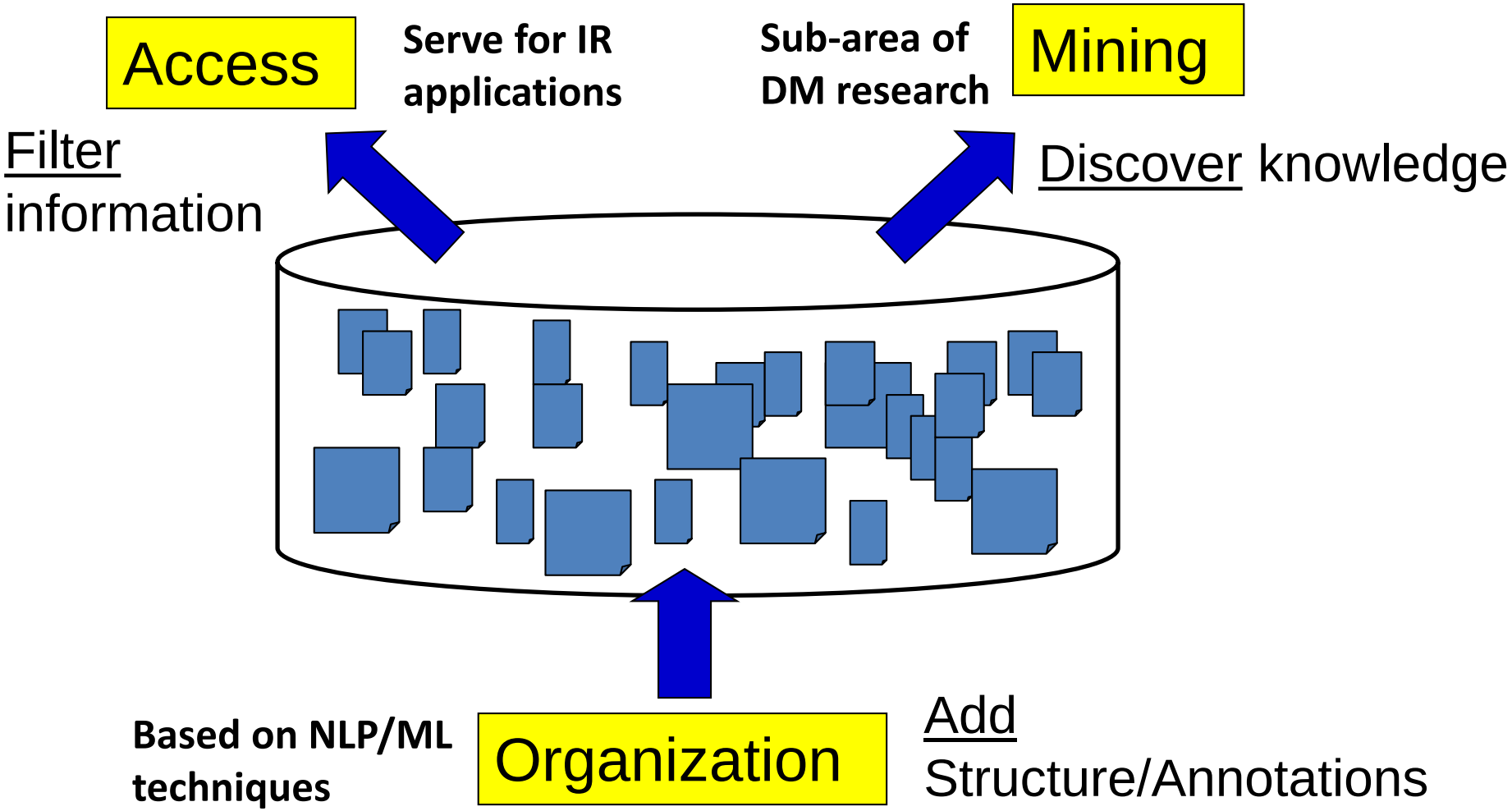
Text mining v.s. NLP, IR, DM…

- 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	Database queries
Textual data	Comp Ling	ysis	Information retrieval

Text Mining

Text mining in general

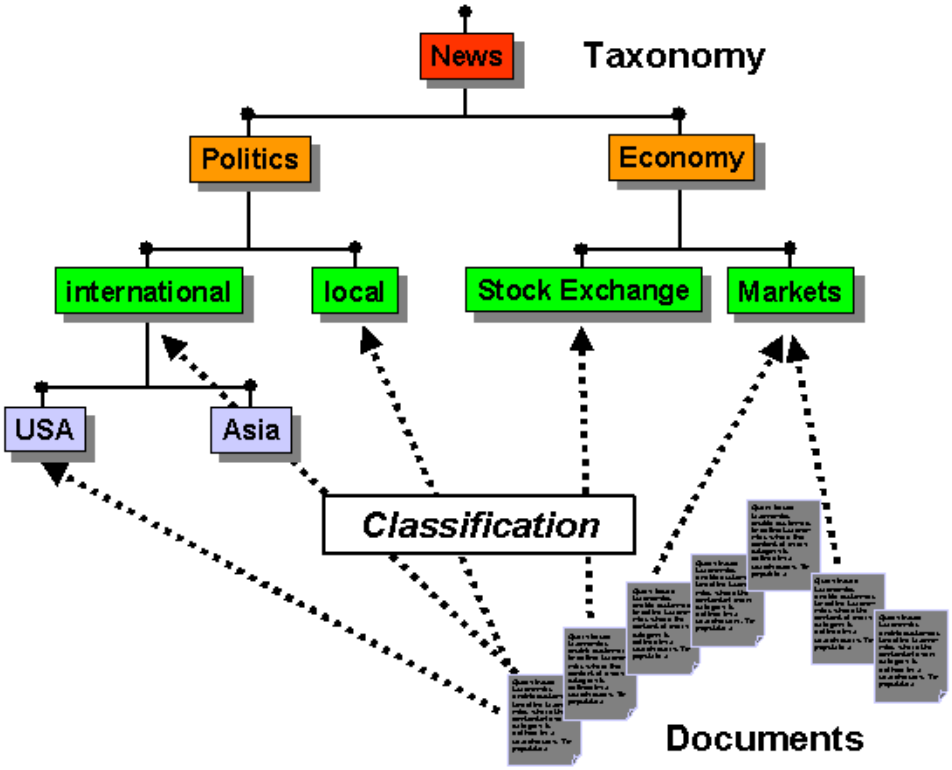


Challenges in text mining

- 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?

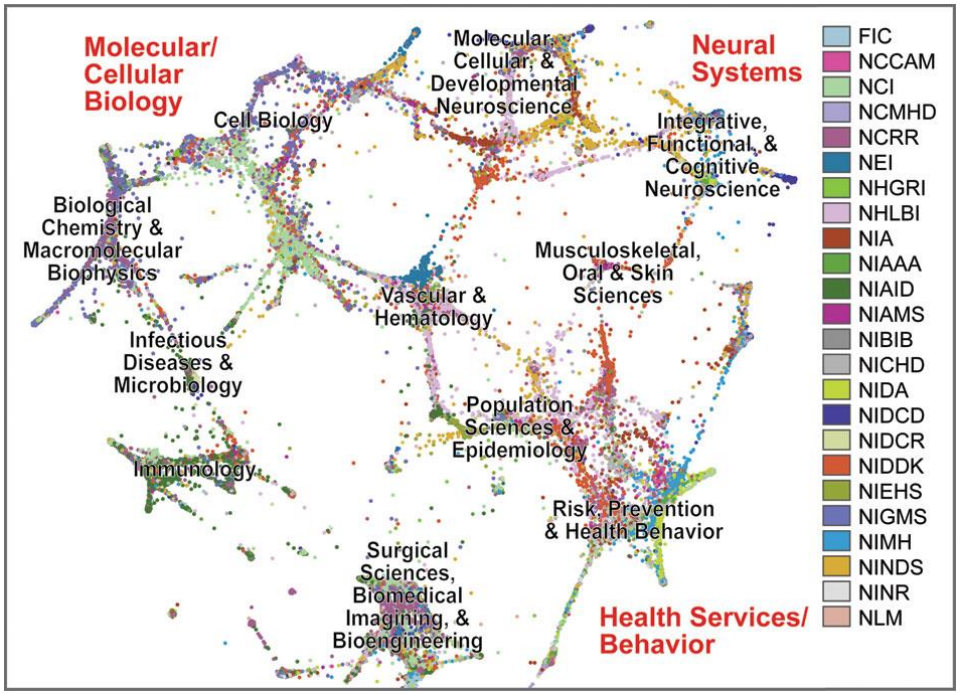
Text mining problems we will solve

- Document categorization
  - Adding structures to the text corpus



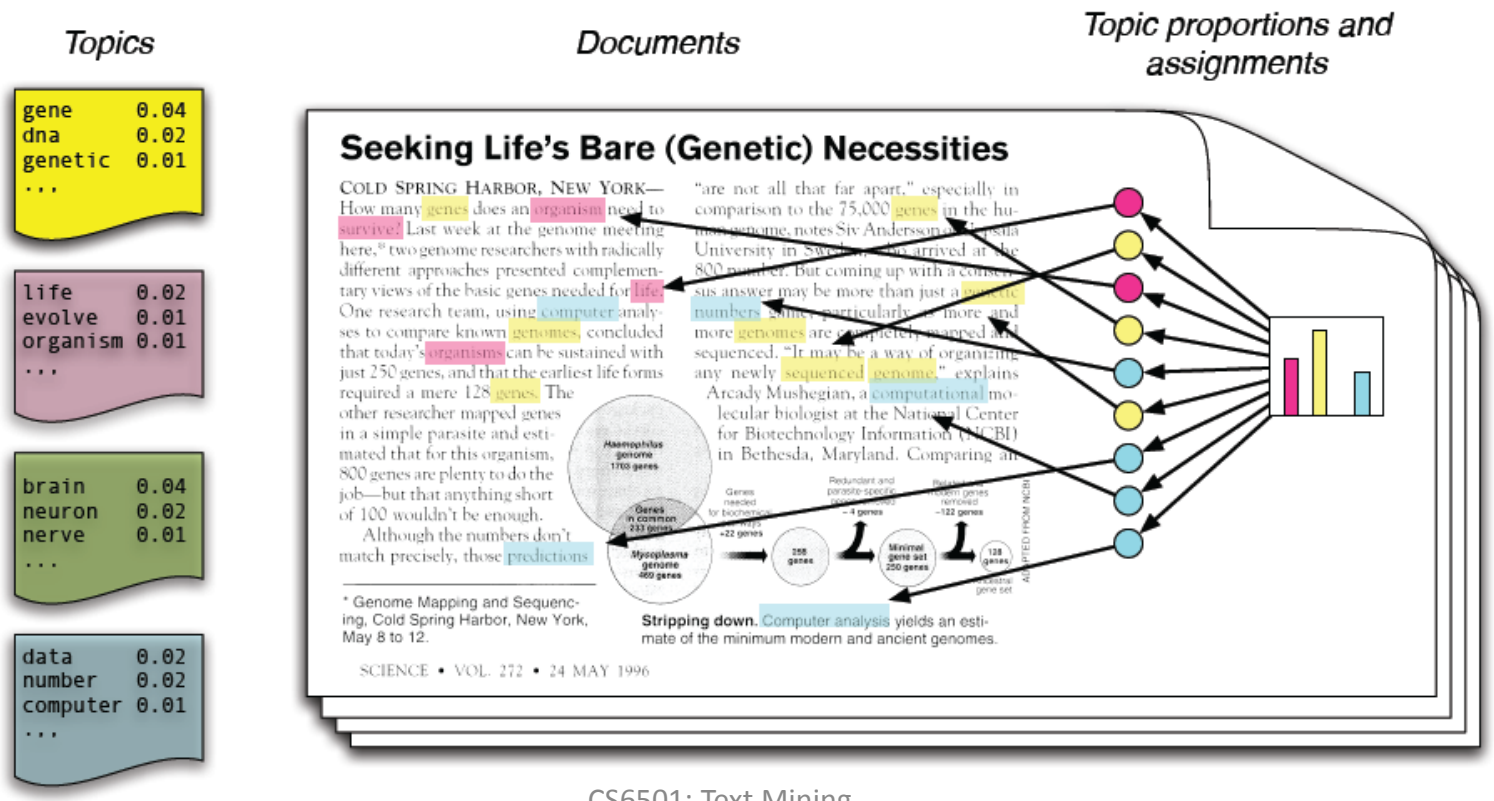
Text mining problems we will solve

- Text clustering
  - Identifying structures in the text corpus



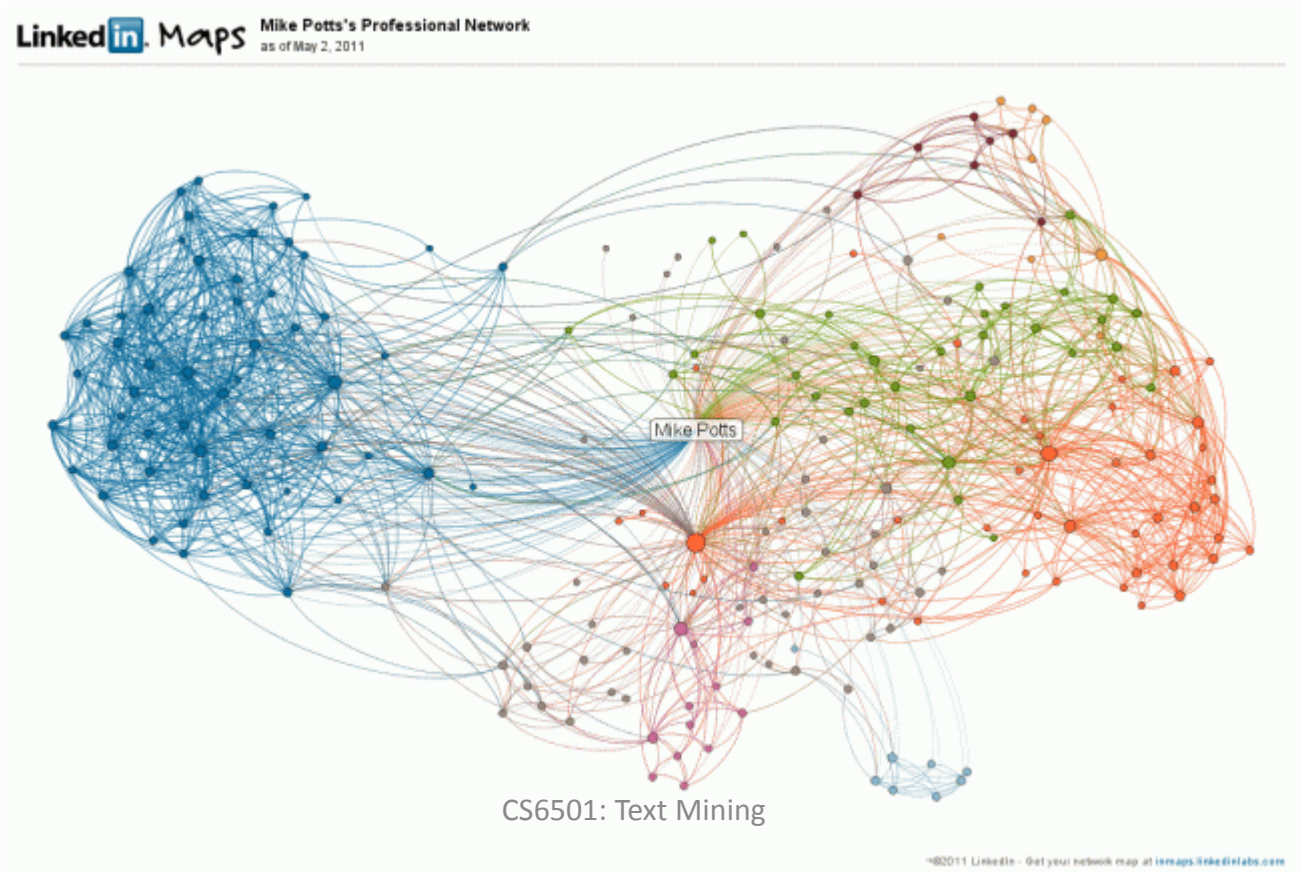
Text mining problems we will solve

- Topic modeling
  - Identifying structures in the text corpus

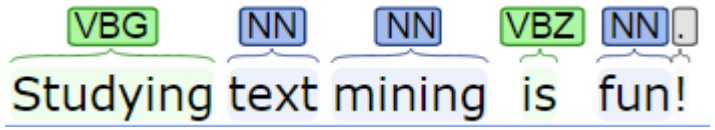


Text mining problems we will solve

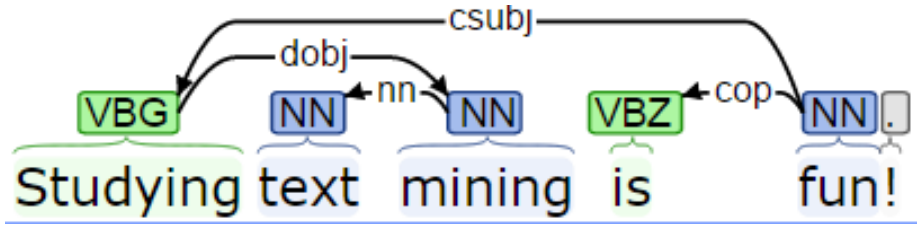
- Social media and network analysis
  - Exploring additional structure in the text corpus



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

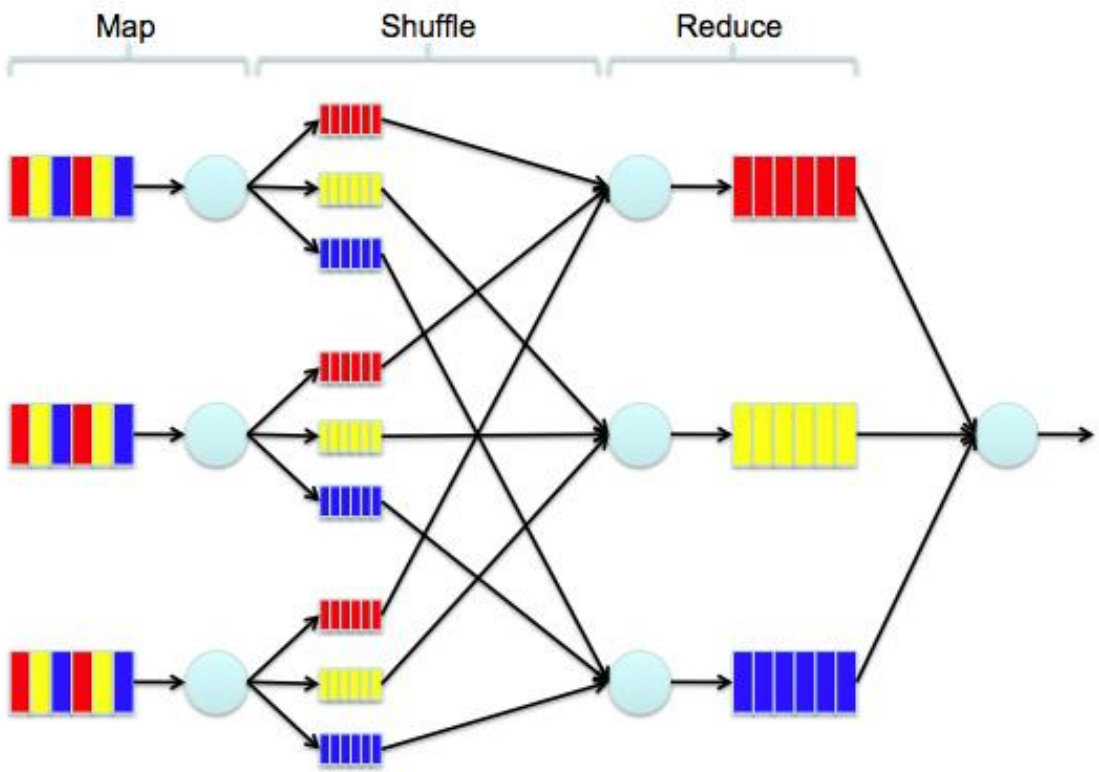
- Machine learning techniques
  - Supervised methods
    - Naïve Bayes, k Nearest Neighbors, Logistic Regression
  - Unsupervised methods
    - K-Means, hierarchical clustering
  - Semi-supervised methods
    - Expectation Maximization

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 (4/2009)
  - eBay has 6.5 PB of user data + 50 TB/da
- 80% data is unstructured (IBM, 2010)

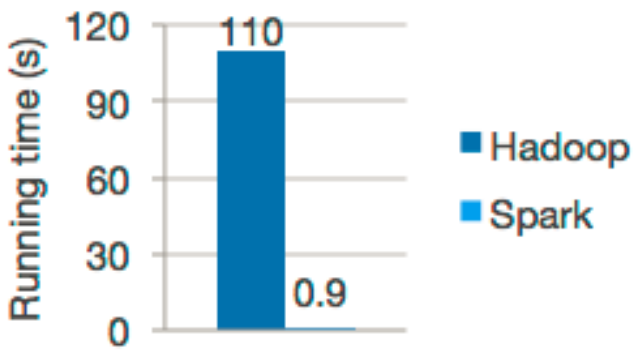
Scalability is crucial

- Large scale text processing techniques
  - MapReduce framework



State-of-the-art solutions

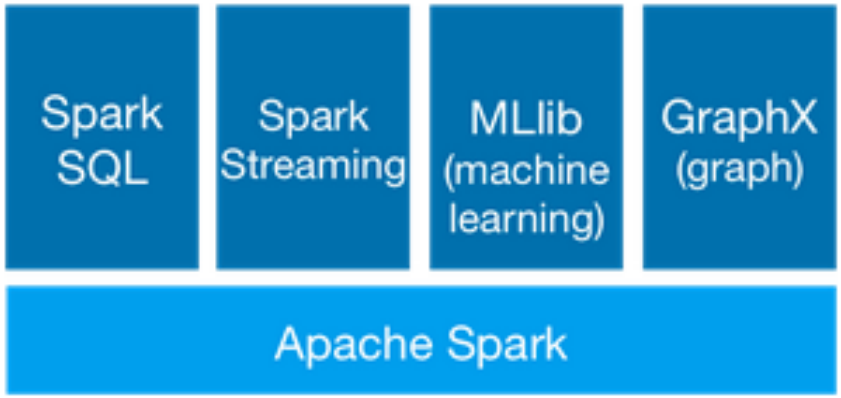
- Apache Spark (spark.apache.org)
  - In-memory MapReduce
    - Specialized for machine learning algorithms
  - Speed
    - 100x faster than Hadoop MapReduce in memory, or 10x faster on disk.



Logistic regression in Hadoop and Spark

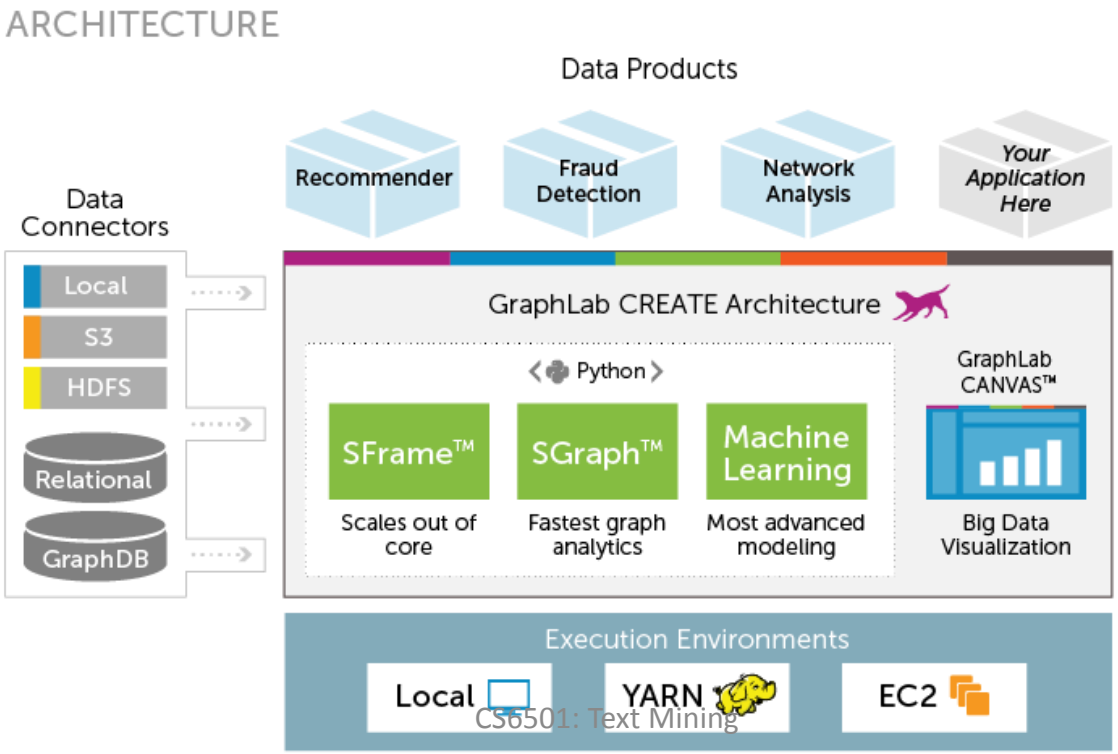
State-of-the-art solutions

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



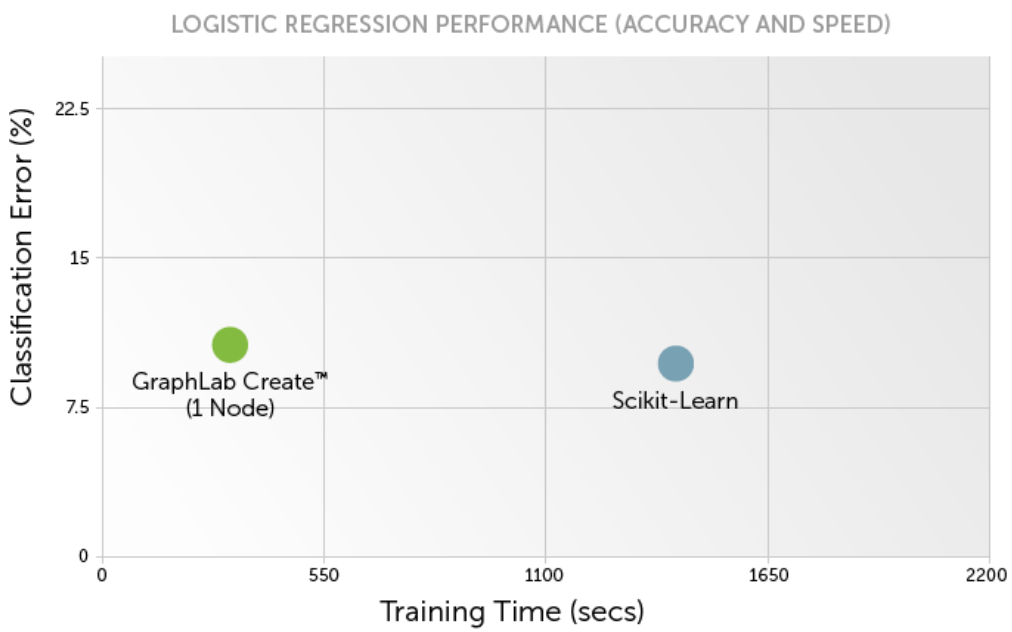
State-of-the-art solutions

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



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

Knowledge service system

Knowledge Discovery

Human-generated data

Text data

Behavior data

Decision Support

Human: big data producer and consumer

Data Generation Modeling

As knowledge consumer

Challenges:

1. **Implicit feedback**
2. **Diverse and dynamic**

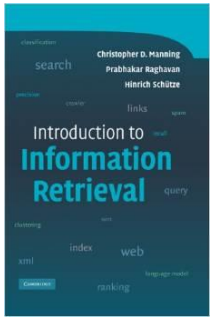
As data producer

Challenges:

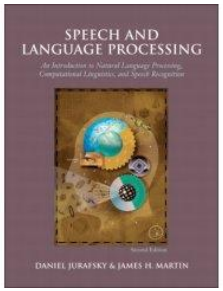
1. **Unstructured data**
2. **Rich semantic**

Text books

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

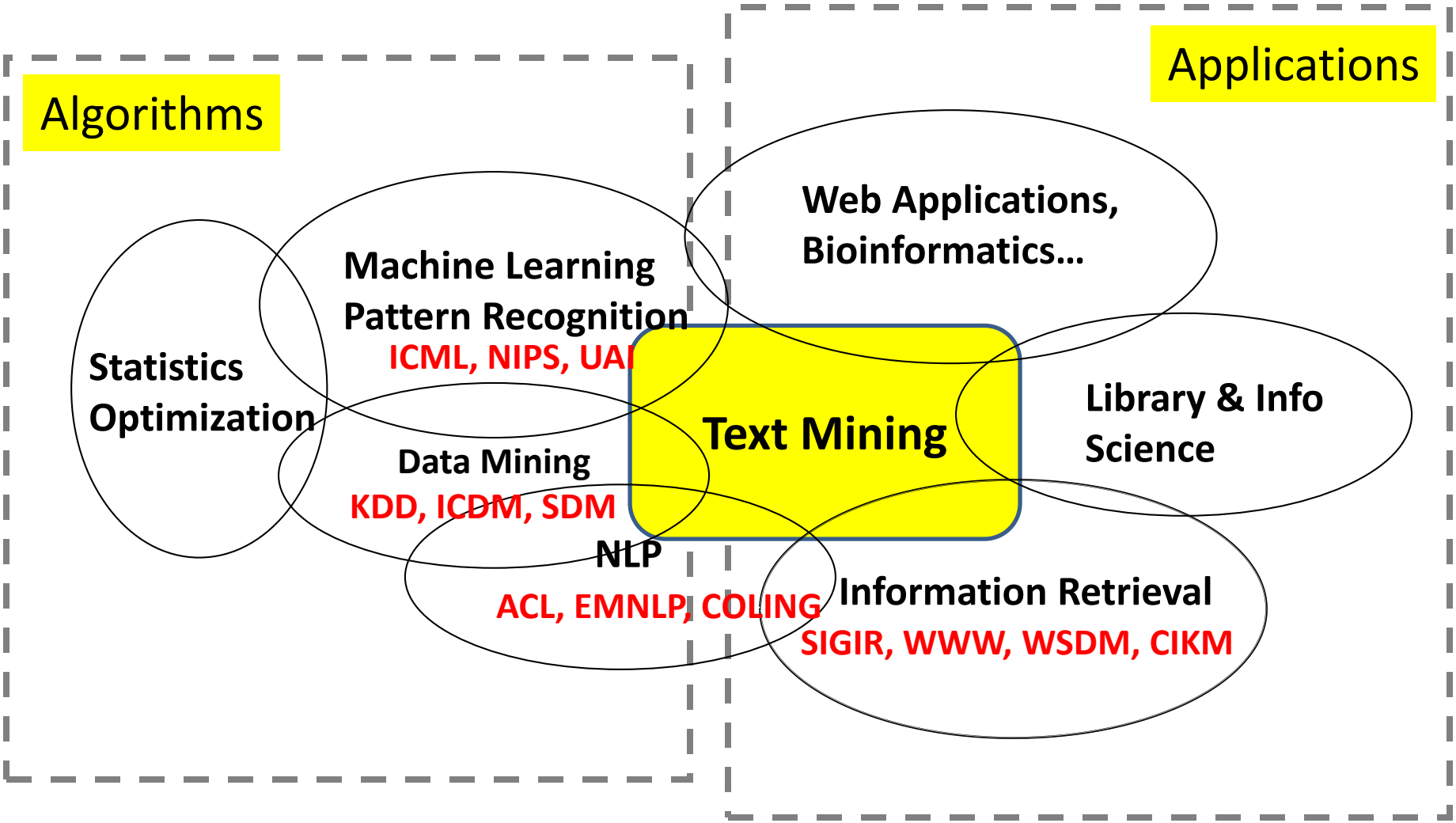


- ***Introduction to Information Retrieval***. Christopher D. Manning, Prabhakar Raghavan, and Hinrich Schuetze, Cambridge University Press, 2007.



- ***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”!